ENCYCLOPEDIA OF NEUROSCIENCE

EDITOR-IN-CHIEF
LARRY R. SQUIRE

ACADEMIC PRESS

EDITOR IN CHIEF

Larry R. Squire

Departments of Psychiatry, Neurosciences, and Psychology
University of California, San Diego and VA Medical Center
San Diego, CA
USA

SENIOR EDITORS

Thomas D. Albright

Systems Neurobiology Laboratories
Salk Institute for Biological Studies
La Jolla, CA
USA

Floyd E. Bloom

Molecular and Integrative Neuroscience Department
The Scripps Research Institute
La Jolla, CA
USA

Fred H. Gage

Laboratory of Genetics
Salk Institute for Biological Studies
La Jolla, CA
USA

Nicholas C. Spitzer

Neurobiology Section
University of California
La Jolla, CA
USA

ASSOCIATE EDITORS

Carrolee Barlow
BrainCells, Inc.
San Diego, CA
USA

M. Flint Beal
Department of Neurology and Neuroscience
Weill Medical College of Cornell University
New York, NY
USA

Darwin K. Berg
Department of Biology
University of California
La Jolla, CA
USA

Anders Bjorklund
Department of Neurobiology
Lund University
Lund
Sweden

M. Christian Brown
Eaton-Peabody Lab
Massachusetts Eye and Ear Infirmary
Boston, MA
USA

Geoffrey Burnstock
Autonomic Neuroscience Centre
Royal Free and University College Medical School
London
UK

John H. Byrne
University of Texas Medical School
Houston, TX
USA

Edward M. Callaway
Systems Neurobiology Laboratory
Salk Institute for Biological Studies
San Diego, CA
USA

William A. Catterall
Department of Pharmacology
University of Washington
Seattle, WA
USA

Verne S. Caviness
Department of Neurology
Massachusetts General Hospital
Boston, MA
USA

John E. Dowling
Department of Neuroscience
Harvard University
Cambridge, MA
USA

James Eberwine
Department of Pharmacology
University of Pennsylvania
Philadelphia, PA
USA

Howard Eichenbaum
Center for Memory and Brain
Boston University
Boston, MA
USA

Caleb E. Finch
Department of Neurogerontology
University of Southern California
Los Angeles, CA
USA

Charles D. Gilbert
The Rockefeller University
New York, NY
USA

Yukiko Goda
MRC Cell Biology Unit
University College London
London
UK

Lawrence S. B. Goldstein
Howard Hughes Medical Institute
UCSD School of Medicine
La Jolla, CA
USA

Antony W. Goodwin
Department of Anatomy and Cell Biology
University of Melbourne
Victoria
Australia

Tomas Hokfelt
Department of Neuroscience
Karolinska Institute
Stockholm
Sweden

Leslie Iversen
Department of Pharmacology
University of Oxford
Oxford
UK

Eugene M. Johnson
Department of Neurology
Washington University School of Medicine
St. Louis, MO
USA

Edward G. Jones
Center for Neuroscience
University of California
Davis, CA
USA

Jon H. Kaas
Department of Psychology
Vanderbilt University
Nashville, TN
USA

Helmut Kettenmann
Max-Delbrück Center for Molecular Medicine
Berlin
Germany

Christopher R. Kintner
Salk Institute for Biological Studies
La Jolla, CA
USA

Keith R. Kluender
Department of Psychology
University of Wisconsin
Madison, WI
USA

Richard J. Krauzlis
Salk Institute for Biological Studies
La Jolla, CA
USA

William B. Kristan
Division of Biological Sciences
University of California San Diego
La Jolla, CA
USA

Joseph E. LeDoux
Center for Neural Science
New York University
New York, NY
USA

Greg E. Lemke
Molecular Neurobiology Laboratory
Salk Institute for Biological Studies
San Diego, CA
USA

David A. Lewis
Departments of Psychiatry and Neuroscience
University of Pittsburgh
Pittsburgh, PA
USA

Pierre J. Magistretti
Center for Psychiatric Neuroscience
Prilly
Switzerland

Robert C. Malenka
Department of Psychiatry and Behavioral Sciences
Stanford University School of Medicine
Palo Alto, CA
USA

Peter R. Marler
Department of Neurobiology, Physiology and Behavior
University of California
Davis, CA
USA

Bruce S. McEwen
The Rockefeller University
New York, NY
USA

Earl K. Miller
The Picower Institute for Learning and Memory and
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA
USA

Paul E. Sawchenko
Salk Institute for Biological Studies
La Jolla, CA
USA

Wolfram Schultz
Department of Physiology, Development and
Neuroscience
Cambridge University
Cambridge
UK

Terrence J. Sejnowski
Computational Neurobiology Lab
Salk Institute for Biological Studies
La Jolla, CA
USA

Gordon M. Shepherd
Department of Neurobiology
Yale University
New Haven, CT
USA

Clarke R. Slater
Institute of Neuroscience
Newcastle University
Newcastle upon Tyne
UK

Craig E. L. Stark
Department of Neurobiology and Behavior
University of California
Irvine, CA
USA

Robert Stickgold
Psychiatry Medical Department
Harvard Medical School
Boston, MA
USA

Peter L. Strick
Pittsburgh Veterans Affairs Medical Center and
Department of Neurobiology
University of Pittsburgh
Pittsburgh, PA
USA

Edward M. Stricker
Department of Neuroscience
University of Pittsburgh
Pittsburgh, PA
USA

Thomas C. Südhof
Stanford University of Medicine
Palo Alto, CA
USA

Fred W. Turek
Center for Sleep and Circadian Biology
Northwestern University
Evanston, IL
USA

Stephen C. Woods
Department of Psychiatry
University of Cincinnati
Cincinnati, OH
USA

FOREWORD

What is an encyclopedia? The term is derived from two Greek words: *enkuklios*, which means cyclical, and *paideia*, which means education. In the early sixteenth century, copyists of Latin manuscripts combined the two words into a Latin designation which comes to us in English with the same spelling and with the meaning 'general course of instruction' (The American Heritage Dictionary, 2000, Boston: Houghton Mifflin, p. 589). The original Greek provides a dynamic connotation, an effort to approach knowledge in a probing, integrated fashion. In Latin and then English, the term has generally been applied to reference volumes addressing broad areas of knowledge.

The Greek emphasis on creative approaches to information is particularly apt for the neurosciences. The term 'neuroscience' is of remarkably young vintage. Francis Schmitt, in his foreword to the first edition of this encyclopedia, related his process of concatenating scientists from disparate fields into an invisible college whose workshops attacked the brain's most recalcitrant puzzles. He dubbed the organization the Neuroscience Research Program (NRP). The early NRP 'associates' included giants such as Melvin Calvin, whose Nobel Prize honored his work on photosynthesis; the Nobel laureate physical chemist Manfred Eigen; the biochemist Albert Lehninger; and Marshall Nirenberg, then a young molecular biologist in the throes of breaking the genetic code.

When Schmitt established the NRP, neuroscience as an integrated endeavor hardly existed. The invention of the microelectrode was permitting neurophysiologists to record from single cells, but characterizing them biochemically was impossible. A major step forward was the emergence in the mid-1960s of the Falck-Hillarp fluorescence microscopic techniques, which permitted selective visualization of catecholamine and serotonin neurons. Mapping the aminergic pathways soon led to collaborative efforts of behaviorists, who could examine the consequences of selective lesions, and biochemists, who no longer were relegated to monitoring neurotransmitters in homogenates of the whole brain. Advances in the 1970s in receptor biochemistry, their localization by autoradiography, and neuropeptide immunohistochemistry further enhanced discourse among neurophysiologists, neuroanatomists, neurochemists, and neuropharmacologists. In the past two decades, the tools of molecular biology have furthered the dialogue.

The explosion of the neurosciences can also be documented through the chronicles of the Society for Neuroscience (SFN). The SFN was founded in 1970 with Vernon Mountcastle as the first elected president, and its inaugural annual meeting in Washington, DC hosted a few hundred researchers. When I served as president in 1980, SFN numbered 7000 members. One of my key tasks was to combat attacks on the *raison d'être* of the society. Some argued, "We have too many scientists in this organization. Let's split into two societies, the 'Wets' and the 'Drys.'" Instead, to emphasize the integrated nature of the field, we launched the *Journal of Neuroscience*. Also, we argued that growth might plateau and that careful meeting organization would prevent individuals from getting 'lost in a crowd.' My prediction about a plateau was off the mark. As of this writing, May 2007, SFN numbers about 38 000 active members, with up to 35 000 people attending each annual meeting, dwarfing any other biomedical research society.

Most would agree that neuroscience is the most integrated scientific discipline. As such, the concept of an encyclopedia, in the original Greek sense of a circle of learning, is notably appropriate yet immensely challenging. The current edition, like earlier ones, succeeds by careful attention to organization and, most importantly, to the selection of the finest researchers as Associate Editors for individual topics. The Associate Editors are all seasoned veterans yet active researchers

whose vision remains at the forefront of their field. All areas of importance are covered, from 'soup to nuts.' Emphasis is elegantly balanced between molecular and systems neuroscience.

In this era of rapid advances, one can question whether an encyclopedia, comprising a snapshot in time, serves a meaningful function. Might not all the information in such an enterprise be obsolete soon after publication? An effective encyclopedia, exemplified in these volumes, integrates disparate areas in a lucid, reader-friendly format. Such a publication can be provocative and invigorating to the most sophisticated professionals. At the same time, the entries are presented in such an inviting fashion that the encyclopedia serves for novices as the ideal entrée into the world of the nervous system.

Solomon H. Snyder
Distinguished Service Professor,
Johns Hopkins University

PREFACE

During the second half of the twentieth century, the study of the nervous system moved from a peripheral position within the biological and psychological sciences to become an interdisciplinary field called neuroscience. The new discipline brought biochemists, cell biologists, anatomists, physiologists, psychologists, neurologists, and psychiatrists – scientists and clinicians from diverse backgrounds, all drawn to the promise and excitement of studying the brain. They aimed to discover the mechanisms of neuronal function, elucidate the neural substrates of behavior and cognition, and learn about the diseases of the nervous system. The development of the discipline was catalyzed in 1969 by the formation of the Society for Neuroscience, which now has nearly 37 000 members. The first academic training programs for neuroscience were established in medical schools (the Department of Neurosciences at the University of California, San Diego in 1965 and the Department of Neurobiology at Harvard University in 1966). The first undergraduate training programs in neuroscience were established in 1972 at Amherst College and at Oberlin College, alma mater of Nobelist Roger Sperry and three Past-Presidents of the Society for Neuroscience. Today, there are more than 300 neuroscience departments and programs around the world.

The *Encyclopedia of Neuroscience* is intended to catalog and explicate the rich, diverse subject matter of the discipline and to facilitate communication among its subspecialties. It is meant to be an authoritative source of information for all areas of neuroscience. It will hopefully make neuroscience more accessible to a wide range of readers, from students making their first acquaintance with the field to general readers seeking information about specific topics. It should also serve as a useful reference for working neuroscientists and be useful as well to undergraduate and graduate students in neuroscience training programs, teachers in the life sciences, clinicians, and science writers.

The inaugural edition of this encyclopedia, which was the first comprehensive reference work for the field, was published in 1987 under the able leadership of George Adelman. It included some 700 entries and appeared in two volumes. The second edition, edited by George Adelman and Barry Smith, appeared in 1999 in two volumes and included more than 800 entries. A CD-ROM version of this edition was also published. A third edition, published only in electronic form, appeared in 2004.

The *Encyclopedia of Neuroscience* includes nearly 1500 entries. The full work will be published online at Science Direct, which can be accessed with subscription at www.sciencedirect.com. To assemble the entries, the Senior Editors identified 46 major areas of the discipline and then invited 46 Associate Editors, all experts in their field, to survey the content of neuroscience within each of these areas. Each Associate Editor then invited 30 to 40 authors to prepare articles on specific topics, with the objective of obtaining complete coverage for each area. Many of the authors are the recognized leaders in their field. The result is a compendium of expert articles representing the current world of neuroscience – the most important research, the most powerful tools, and the most promising applications.

Most of the entries are self-contained reviews that can be read as independent articles. Extensive cross-listing at the conclusion of each entry directs readers to articles on related topics. The principal organization of the *Encyclopedia* lies in the alphabetically arranged list of entries. In addition, the comprehensive subject classification will help readers find related topics and appreciate the structure of the discipline.

While no single reference work in neuroscience can claim to include every notable idea and fact about the

brain, the Senior Editors hope that these volumes provide a summary of contemporary neuroscience that is both comprehensive and instructive. Neuroscience is still a developing field, but the *Encyclopedia* will have succeeded if it conveys the considerable promise that neuroscience offers for conquering the diseases that affect the nervous system and for understanding the brain, the mind, and ourselves.

The Senior Editors are grateful to the Developmental Editors at Elsevier, Michael Bevan, Joanna De Souza and Richard Berryman, and the Editorial Assistants, Caroline Phipps, Afandi Mohamed and Nicky Carter, for capably managing the formidable task of assembling and organizing the contents of the Encyclopedia. Andrew Lowe and Laura Jackson, the Project Managers, diligently brought the project through its several stages of production.

Larry R. Squire
Editor-in-Chief

Acetylcholine Neurotransmission in CNS

M-M Mesulam, Northwestern University, Chicago, IL, USA

© 2009 Published by Elsevier Ltd.

Introduction

Acetylcholine (ACh) was the first neurotransmitter to be discovered. It is synthesized by choline acetyltransferase (ChAT) from choline and acetate as precursors. Its action on cholinergic receptors is terminated through hydrolysis by acetylcholinesterase (AChE). Neurons and axons that contain ChAT and that use ACh for neurotransmission are designated 'cholinergic,' and the term 'cholinoceptive' is used to designate neurons that respond to ACh through muscarinic or nicotinic receptors.

Cholinergic innervation is ubiquitous at all levels of the neuraxis but also shows precise patterns of origin and distribution. The vast majority of cholinergic innervation of the thalamus arises from the pedunculopontine and laterodorsal tegmental nuclei (also known as Ch5 and Ch6); the vast majority of cholinergic innervation for the caudate and putamen arises from intrinsic cholinergic neurons within the striatum; and the vast majority of cholinergic innervation for the cerebral cortex, hippocampus, and amygdala arises from the basal forebrain.

The human cerebral cortex has no cholinergic neurons but receives massive cholinergic innervation. Nearly all of it arises from four partially overlapping basal forebrain cell groups where cholinergic and noncholinergic neurons are intermingled. The Ch1–Ch4 nomenclature was introduced to designate the cholinergic (i.e., ChAT-containing) neurons within these four cell groups. According to this nomenclature, Ch1 designates the cholinergic cells associated predominantly with the medial septal nucleus; Ch2, those associated with the vertical nucleus of the diagonal band; Ch3, those associated with the horizontal limb of the diagonal band nucleus; and Ch4, those associated with the nucleus basalis of Meynert. Tracer experiments in laboratory animals have shown that Ch1 and Ch2 provide the major cholinergic innervation for the hippocampal complex, Ch3 for the olfactory bulb, and Ch4 for the rest of the cerebral cortex and the amygdala. The term 'nucleus basalis' can be used to designate the cholinergic as well as noncholinergic components in this nucleus, whereas the more restrictive Ch4 designation is reserved for its cholinergic neurons.

ACh Neurotransmission

Anatomy of the Nucleus Basalis

The nucleus basalis is a phylogenetically progressive nucleus which displays its greatest differentiation in the cetacean and human brains. The human nucleus basalis extends from the level of the olfactory tubercle to that of the posterior amygdala, spanning a distance of 13–14 mm in the anteroposterior axis and attaining a mediolateral width of 18 mm within the substantia innominata (subcommissural gray). It contains approximately 200 000 neurons in each hemisphere. In the human brain, Ch4 neurons express ChAT, the vesicular ACh transporter, AChE, calbindin-d28k, the high affinity nerve growth factor receptor (NGFr) trkA, and the low affinity p75 NGFr. A minority of Ch4 neurons are NGFr-negative and, at least in the rat, project preferentially to the amygdala. The nucleus basalis also contains a complex mosaic of noncholinergic neurons that are nicotinamide adenine dinucleotide phosphate diaphorase (NADPHd)-positive, γ-aminobutyric acid (GABA)ergic, peptidergic, or tyrosine hydroxylase-positive. Based on its location, connectivity, and cellular morphology, the nucleus basalis has been conceptualized as a site of confluence for the limbic system and the ascending reticular activating system.

Inputs and Neurotransmitter Circuitry of the Nucleus Basalis

Although the primate nucleus basalis projects to the entire cerebral cortex, it receives major cortical projections from only the limbic and paralimbic regions of the brain, including the amygdala. The cortical inputs are mostly glutamatergic but can also be GABAergic. The primate nucleus basalis receives dopamine, norepinephrine, and serotonin inputs, presumably all arising in the brain stem.

In the monkey, cholinergic synapses onto cholinergic Ch4 neurons are quite frequent and take the form of large asymmetrical synapses. The precise source of the cholinergic input to the nucleus basalis is unknown but could include collaterals from cholinergic Ch1–Ch4 neurons of the basal forebrain or ascending projections from the Ch5–Ch6 group of pontomesencephalic cholinergic nuclei. In the rat and monkey, the m2 muscarinic receptor subtype is prominently expressed by only a third of Ch4 neurons. Therefore, the often cited assumptions that all Ch4 neurons express m2 and that m2 is a universal presynaptic marker of cortical cholinergic innervation need to be modified.

Topography and Distribution of Cortical Cholinergic Projections

The density of cholinergic axons is higher in the more superficial layers of the cerebral cortex, suggesting that the axons which enter the cortex from the underlying gray matter undergo branching as they course toward the pial surface, an interpretation which is supported by physiological evidence. The density of cholinergic innervation is lower within unimodal and heteromodal association areas than in paralimbic areas of the brain. Core limbic areas such as the amygdala and hippocampus contain the highest densities of cholinergic innervation.

Trajectory of Cholinergic Pathways from the Nucleus Basalis to the Cerebral Cortex

Two bundles of cholinergic fibers extend from the nucleus basalis to the cerebral cortex and amygdala in the human brain. A medial pathway joins the white matter of the gyrus rectus, curves around the rostrum of the corpus callosum to enter the cingulum, and merges with fibers of the lateral pathway within the occipital lobe. It supplies the parolfactory, cingulate, pericingulate, and retrosplenial cortices in the medial parts of the cerebral hemispheres. The lateral pathway is subdivided into a capsular division traveling in the white matter of the external capsule and a perisylvian division traveling within the claustrum. Branches of the perisylvian division supply the fronto-parietal operculum, insula, and superior temporal gyrus. Branches of the capsular division innervate the remaining parts of the frontal, parietal, and temporal neocortex.

Cholinergic Synapses in the Cerebral Cortex

In the rat and cat cerebral cortex, ChAT-immunoreactive fibers are almost exclusively unmyelinated, display numerous varicosities, and make mostly symmetrical synapses on the perikaryon, dendritic shaft, and spines of pyramidal as well as nonpyramidal neurons. In the temporal cortex of the human brain, most ChAT-positive terminals form identifiable synaptic specializations. These are usually quite small and symmetric and located predominantly on the dendritic shafts and spines of pyramidal neurons.

Cholinergic Receptors in the Cerebral Cortex

The ACh released by cholinergic axons exerts its influence on the cerebral cortex by interacting with neurons expressing muscarinic and nicotinic receptors. In mammals, the m1 subtype is the most common species of cholinergic receptor in the cerebral cortex. Muscarinic receptor subtypes are found not only at traditional postsynaptic cholinoceptive sites but also at additional locations where they can function as postsynaptic modulators of noncholinergic transmission or presynaptic autoreceptors that influence the release of ACh.

Cholinoceptive Neurons of the Cerebral Cortex

Incoming cholinergic axons innervate at least three kinds of cortical neurons: glutamatergic neurons, GABAergic interneurons, and NADPHd-positive infracortical neurons. The one marker most closely associated with cholinoceptive neurons is AChE, the enzyme which terminates cholinergic neurotransmission through the rapid hydrolysis of ACh into acetate and choline. Although the vast majority of cholinoceptive neurons probably express some AChE, only a subset yields an AChE-rich histochemical staining pattern. In the cerebral cortex of the adult rat, the AChE-rich cytochemical pattern is limited to a few polymorphic neurons. The situation is dramatically different in the human cerebral cortex, which contains a dense network of AChE-rich cortical neurons, especially in layers III and V of premotor and sensory association cortex. These neurons are mostly pyramidal in shape and are likely to represent a glutamatergic contingent of cholinoceptive neurons. The cerebral cortex also contains GABAergic cholinoceptive interneurons that respond to ACh through muscarinic receptors. As shown by Freund and Meskenaite, these neurons may receive additional inputs from the GABAergic projection neurons of the nucleus basalis.

Functionality of Cortical Cholinergic Innervation

Although the primate Ch4 projects to all cortical areas, it receives major inputs almost exclusively from limbic and paralimbic regions of the cerebral cortex. This creates a unique circuitry through which cholinergic projections can modulate neural activity in all cortical areas in ways that reflect the prevailing influence of the limbic system. Experiments in rodents and cats have shown that the principal action of cholinergic stimulation at m1 receptor sites is to induce a relatively prolonged reduction of potassium conductance so as to make the cholinoceptive neuron more responsive to other incoming excitatory inputs. A sensory event that activates Ch4 would thus be expected to have a stronger impact on cortical circuitry and to attract more robust neural encoding. These aspects of receptor physiology and the preferential relationship to the limbic system suggest that the attentional modulation of behaviorally relevant events and their preferential encoding in memory are two cognitive faculties most likely to be under the influence of cortical cholinergic innervation.

Memory and Cortical Cholinergic Innervation

The characteristically intense cholinergic innervation of limbic areas, including the hippocampus and entorhinal cortex, had promoted the opinion that cortical cholinergic neurotransmission plays a crucial role in memory function. In rodents and monkeys, damaging the cholinergic neurons in the basal forebrain causes learning and memory impairments in some experiments but not in all. Turchi and colleagues showed that selective immunotoxic lesioning of cholinergic axons in the parahippocampal–rhinal region of the monkey brain leads to distinct memory deficits in the delayed nonmatching-to-sample task. The deficit was as severe as that obtained by lesioning the parahippocampal–rhinal cortex, suggesting that the mnemonic function of this region is dependent on its cholinergic innervation. Humans respond to the muscarinic antagonist scopolamine with a memory impairment somewhat similar to that seen in Alzheimer's disease (AD). The role of ACh in hippocampal long-term potentiation, its ability to suppress proactive interference, and its participation in memory consolidation may provide some of the cellular mechanisms that mediate its importance for memory function.

Attentional State and Cortical Cholinergic Innervation

There is considerable agreement concerning the importance of cortical cholinergic innervation for various aspects of attention, including the setting of signal-to-noise ratios during information processing and the online holding of information. The nucleus basalis projects to all cortical areas while receiving its cortical inputs almost exclusively from components of the limbic system, leading to an arrangement that promotes the selective release of cortical ACh in response to events that are of limbic relevance. Cortical cholinergic innervation is thus in a position to preferentially enhance the cortical impact of events that are of emotional and motivational significance. In keeping with this formulation, neurons of the nucleus basalis in the monkey are sensitive to novel and motivationally relevant sensory events.

Cortical ACh and Plasticity

Neuroplasticity is a life-long process that mediates the structural and functional reaction of dendrites, axons, and synapses to experience, attrition, and injury. One of the most interesting functional correlates of cortical ACh is its role in mediating neuroplasticity. In an experiment by Baskerville and colleagues, all whiskers except for D2 and D3 were trimmed in rats. This led to a pairing between the D2 and D3 barrel fields so that the D2 neurons in the cerebral cortex started to show a greater responsivity to stimulation of D3 than to stimulation of the adjacent D1, which had been trimmed. This pairing, indicative of experience-induced synaptic plasticity, could not be obtained in rats with selective immunotoxic lesions of the cholinergic neurons in the nucleus basalis. Furthermore, Kilgard and Merzenich showed that pairing auditory stimuli with the electrical stimulation of the nucleus basalis in adult rats caused a long-lasting reorganization of primary auditory cortex so that the area optimally responsive to the paired tone expanded substantially. This plasticity was also abolished following the selective immunotoxic destruction of cholinergic nucleus basalis neurons. Based on these observations, it appears that cortical cholinergic denervation can undermine the learning-dependent reorganization of cortical circuitry and perhaps also the ability of the brain to keep itself in good repair in response to attrition and injury.

Cortical Cholinergic Innervation and AD

AD is associated with a prominent loss of cortical (but not thalamic or striatal) cholinergic innervation (Figure 1). The modest therapeutic effect of cholinesterase inhibitors such as galantamine, donepezil, and rivastigmine in AD is attributed to their cholinomimetic effects. The initial reports of cortical cholinergic denervation and cell loss in the nucleus basalis were based on patients who had advanced AD. Recent evidence shows that neurofibrillary degeneration of the Ch4–nucleus basalis complex starts in the course of normal aging and that it becomes more pronounced even in the prodromal stages of AD (also known as mild cognitive impairment (MCI)). By early stages of AD, this cholinergic cytopathology is already quite advanced. There is a significant inverse correlation between the intensity of neurofibrillary degeneration in the nucleus basalis and performance in delayed memory tests throughout the age–MCI–AD continuum.

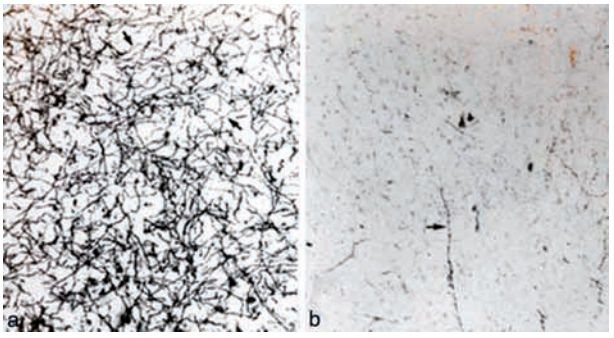

Figure 1 Cholinergic fibers (arrows) of the temporal neocortex in a cognitively normal 70-year-old (a) and in a patient of similar age with Alzheimer's disease (b). Magnification 100×.

The presence of cytopathology in the nucleus basalis of cognitively normal older people may initially raise the possibility that it may have no functional relevance. Such an inference would be based on the erroneous assumption that 'normal for age' reflects a stability of cognitive function. However, a person in the age range of 70–100 may have experienced a considerable decline of performance from a former baseline and still remain within the normal range for age. The nucleus basalis cytopathology observed in cognitively normal people may have contributed to the emergence of these age-related changes. The substantially more severe nucleus basalis lesion of AD may contribute to the emergence of the dementia by interfering with cholinergic neurotransmission and also by eroding the capacity for neuroplasticity to the point that the brain cannot keep itself in good repair.

Numerous hypotheses have been advanced to explain why the pathway from the basal forebrain to the cerebral cortex is so vulnerable to AD. Biosynthetic bottlenecks in the production of ACh, special vulnerabilities to amyloid, perturbation of intracellular calcium homeostasis, and impairments in the axonal transport of growth factors have been invoked as possible explanations, but proof is lacking. The answer may lie in the anatomical affiliations of the Ch4–nucleus basalis complex. This cell group, together with the hippocampus, amygdala, and entorhinal cortex, is part of an uninterrupted band of core limbic areas. These are the regions of the brain that attract the greatest neurofibrillary degeneration and related cell death in the course of aging and AD, perhaps because they sustain an unusually high neuroplasticity burden throughout the life span. The severe and selective loss of cortical cholinergic innervation may thus reflect the location and connectivity of the nucleus basalis rather than its relationship to cholinergic neurotransmission.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Cholinergic Pathways in CNS; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Nicotinic Acetylcholine Receptors.

Further Reading

Baskerville KA, Schweitzer JB, and Herron P (1997) Effects of cholinergic depletion on experience-dependent plasticity in the cortex of the rat. *Neuroscience* 80: 1159–1169.

Berger-Sweeney J, Heckers S, Mesulam M-M, Wiley RG, Lappi DA, and Sharma M (1994) Differential effects upon spatial navigation of immunotoxin-induced cholinergic lesions of the medial septal area and nucleus basalis magnocellularis. *Journal of Neuroscience* 14: 4507–4519.

De Long MR (1971) Activity of pallidal neurons during movement. *Journal of Neurophysiology* 34: 414–427.

De Rosa E, Desmond JE, Anderson AK, Pfefferbaum A, and Sullivan EV (2004) The human basal forebrain integrates the old and the new. *Neuron* 41: 825–837.

Freund TF and Meskenaite V (1992) g-Amynobutyric acid-containing basal forebrain neurons innervate inhibitory interneurons in the neocortex. *Proceedings of the National Academy of Sciences of the United States of America* 89: 738–742.

Gorry JD (1963) Studies on the comparative anatomy of the ganglion basale of Meynert. *Acta Anatomica* 55: 51–104.

Heckers S, Ohtake T, Wiley RG, Lappi DA, Geula C, and Mesulam M-M (1994) Complete and selective cholinergic denervation of rat neocortex and hippocampus but not amygdala by an immunotoxin agoinst the p75 NGF receptor. *Journal of Neuroscience* 14: 1271–1289.

Kilgard MP and Merzenich MM (1998) Cortical map reorganization enabled by nucleus basalis activity. *Science* 279: 1714–1718.

McCormick DA and Prince DA (1985) Two types of muscarinic responses to acetylcholine in mammalian cortical neurons. *Proceedings of the National Academy of Sciences of the United States of America* 82: 6344–6348.

Mesulam M-M (1996) The systems-level organization of cholinergic innervation in the cerebral cortex and its alterations in Alzheimer's disease. *Progress in Brain Research* 109: 285–298.

Mesulam M-M and Geula C (1991) Acetylcholinesterase-rich neurons of the human cerebral cortex: Cytoarchitectonic and ontogenetic patterns of distribution. *Journal of Comparative Neurology* 306: 193–220.

Mesulam M-M, Shaw P, Mash D, and Weintraub S (2004) Cholinergic nucleus basalis tauopathy emerges early in the aging-MCI-AD continuum. *Annals of Neurology* 55: 815–828.

Ridley RM, Murray TK, Johnson JA, and Baker HF (1986) Learning impairment following lesion of the basal nucleus of Meynert in the marmoset: Modification by cholinergic drugs. *Brain Research* 376: 108–116.

Sarter M and Bruno JP (2000) Cortical cholinergic inputs mediating arousal, attentional processing and dreaming: Differential afferent regulation of the basal forebrain by telencephalic and brainstem afferents. *Neuroscience* 95: 933–952.

Turchi J, Saunders RC, and Mishkin M (2005) Effects of cholinergic deafferentation of the rhinal cortex on visual recognition memory in monkeys. *Proceedings of the National Academy of Sciences of the United States of America* 102: 2158–2161.

Wilson FAW and Rolls ET (1990) Neuronal responses related to novelty and familiarity of visual stimuli in the substantia innominata, diagonal band of Broca and periventricular region of the primate basal forebrain. *Experimental Brain Research* 80: 104–120.

Záborszky L, Gaykema RP, Swanson DJ, and Cullinan WE (1997) Cortical input to the basal forebrain. *Neuroscience* 79: 1051–1078.

Acetylcholinesterase

P Taylor, S Camp, and Z Radić, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The turnover number of acetylcholinesterase (AChE) approaches 1.5×10^4 s^{-1}, making it one of the most efficient enzymes known. Rapidity of catalysis of released acetylcholine in a submillisecond time frame is essential in the skeletal neuromuscular junction to allow the next volley of released acetylcholine to trigger a postsynaptic excitatory potential. AChE (EC 3.1.1.7) is distinguished from butyrylcholinesterase (BChE; EC 3.1.1.8) by its catalytic selectivity for acetylcholine over butyrylcholine hydrolysis. AChE is typically synthesized in nerve, muscle, and certain hematopoietic cells. In excitable tissues, AChE expression is regulated by tissue-specific development, and the enzyme is localized on the extracellular surface of both nerve and muscle. The AChE found in the neuromuscular junction of skeletal muscle is synthesized by the muscle rather than the nerve cell. AChE inhibitors are employed in the treatment of Alzheimer's disease, myasthenia gravis, glaucoma, smooth muscle atony, and assorted disorders of autonomic nervous system functions.

BChE is synthesized largely in the liver and is exported into the plasma. It displays a wider substrate range for catalysis than does AChE and is thought to play a primary role in the metabolism of dietary esters, perhaps only in selected species. Genetic polymorphisms that preclude the expression of BChE in humans yield no apparent phenotype, unless drugs in doses requiring BChE catalysis for rapid degradation are given. BChE has no apparent role in the nervous system, and its protective role only becomes manifest in AChE-deleted transgenic mice.

Genes Encoding the Cholinesterases

To date, about 150 DNA sequences encoding cholinesterases, extending from nematodes and insects to humans, have been reported. Although in mammals a single AChE subserves all cholinergic nervous system function, this is not the case in lower species, such as nematodes and certain insects, in which multiple AChE species exist and influence nervous system function. In *Drosophila*, a single cholinesterase is found; its structure shows features of both AChE and BChE.

Distinct, but single, genes in mammals encode AChE and BChE, yielding gene products with ~50% amino acid identity. AChE exists in multiple molecular species that arise from three alternative splice options in the gene, not found in BChE. The AChE gene contains three constitutive exons within the open reading frame, starting in exon 2; in the following two exons, three options exist: reading through exon 4, splicing to exon 5, and splicing to exon 6 (Figure 1). These options yield a soluble monomer, glycophospholipid-linked dimers, or a tetrameric form of AChE, respectively. The monomeric subunits of all of these forms, although differing in their very C-termini, contain identical sequences forming the catalytic and inhibitor binding sites. The monomeric unit is glycosylated at three or four N-linked sites and has a peptide molecular weight of ~62 000 (Figure 2).

Exon 6 encodes a C-terminal region that contains a proline-attachment domain, enabling it, when fully assembled as a tetramer, to associate with proline-rich sequences of two structural subunits. The disulfide-linked assembly with the first structural subunit yields a heteromeric tetramer of catalytic subunits with the single structural subunit (Figure 2). This structural subunit contains the requisite hydrophobicity to be attached to the outer leaflet of the cell membrane. Assembly with a second type of structural subunit yields a more complex heteromeric structure composed typically of three AChE tetramers associated with a collagen-containing unit. This subunit also has a proline-rich sequence at its N-terminus for association and attachment through a disulfide linkage. The C-terminal portion of the collagen-containing subunit appears to associate with perlecan and perhaps other extracellular synaptic proteins. Each subunit of the triple-helical collagen tail associates with a tetramer. Accordingly, the assembled species approaches a molecular mass of 10^6 Da. The membrane-associated heteromeric species is the predominant form in brain, while the collagen-containing heteromer is localized to the basal lamina of the neuromuscular junction.

Regulation of AChE gene expression occurs at multiple levels to account for the great enhancement in expression associated with muscle and nerve differentiation. Accordingly, transcriptional activation and mRNA stabilization have been described in muscle differentiation. The primary promoter region of the *AChE* gene is selectively regulated in skeletal muscle, in which an intron between exons 1 and 2 serves to interact with the promoter to enhance expression during the myoblast-to-myotube conversion. This 250-bp intronic region of the gene, containing several regulatory elements, is well conserved between mammalian species. Its importance in controlling

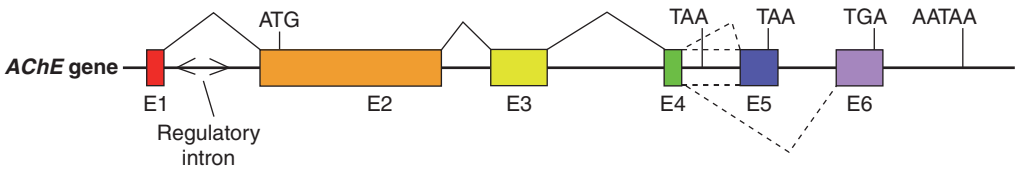

Figure 1 Acetylcholinesterase (AChE) gene structure. The gene is compacted into approximately 7.5 kbp with six closely spaced exons (E1–E6). Although activity from a more distal 5′ promoter has been reported, the primary promoter is immediately 5′ of the cap site of the start of transcription. Its activity is regulated by a 250-bp region in the intron between exons 1 and 2. Exon 2 encodes the start methionine, and the region 3′ of the start site in exon 2, exon 3, and exon 4 encodes the invariant region of the multiple spliced forms of AChE. Following exon 4, three alternatives are possible: (1) reading through exon 4 into the intron, giving rise to a soluble monomeric form of AChE; (2) splicing to exon 5, giving rise, after processing, to a glycophospholipid-linked dimer; and (3) splicing to exon 6, yielding a sequence predisposed to attaching to proline rich sequences in the structural subunits.

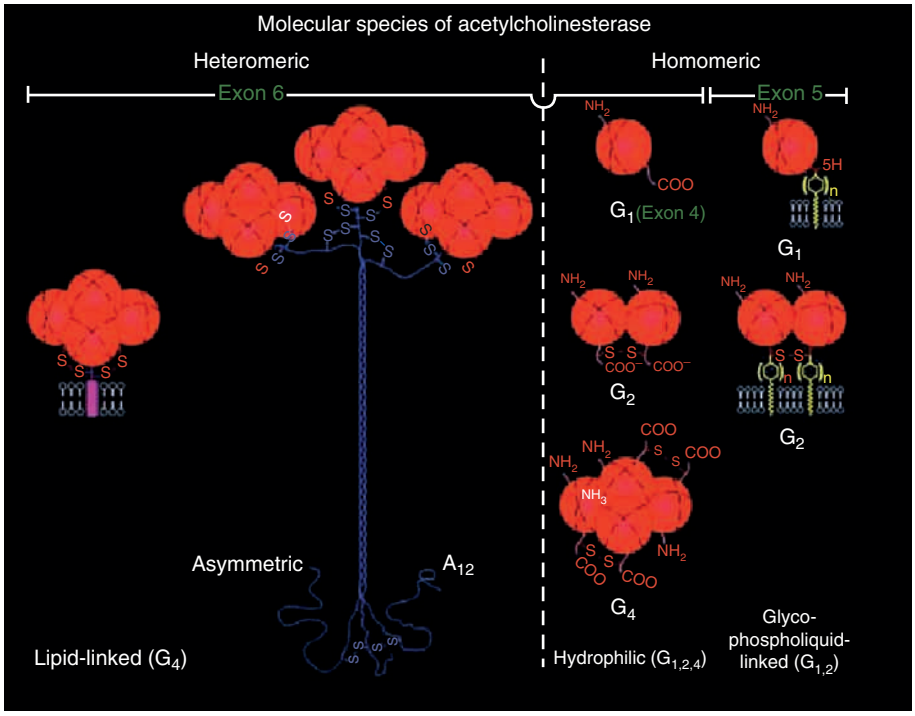

Figure 2 Molecular forms of acetylcholinesterase (AChE). The molecular species arising from alternative mRNA processing of the gene in Figure 1 and the subunit assembly are shown. Depending on the splice alternative, AChE can exist in multiple forms of assembly (red), ranging from homomeric monomers, glycophospholipid (yellow) attached dimers and tetramers, to heteromeric species formed by disulfide attachment to a hydrophobic, lipid-linked subunit (fushia) or a subunit with a collagen-containing sequence (blue).

expression *in vivo* has been documented in transgenic mice in which selective regions of the *AChE* gene have been deleted.

Structure of the Cholinesterases

About 100 structures of cholinesterases as apo forms or complexed or conjugated with various ligands have been deposited in the Protein Data Bank. From these structures the molecular bases for efficient catalysis and inhibitor selectivity are beginning to be uncovered. The active center of the enzyme is centrosymmetric to the subunit and lies at the bottom of a narrow gorge, 18–20 Å in depth and lined with aromatic residues (Figure 3). Hydrogen bonding between Asp334, His447, and Ser203 (the catalytic triad) renders the serine an effective nucleophile. The side chains of Phe295 and Phe297 outline the space restricted acyl pocket in AChE, and contain smaller side chains for BChE. Aromatic residues of Trp86, Tyr337, and Try338 surround the choline subsite, forming a π orbital nest for the choline moiety. At the rim of the gorge lies a peripheral site, distinguished by an aromatic patch, where allosteric inhibitors, such as propidium, and the peptidic snake toxin fasciculin bind. Dimers of AChE are stabilized through a four-helix bundle. An amphiphilic, tryptophan-containing sequence, encoded only by exon 6, stabilizes the tetrameric form.

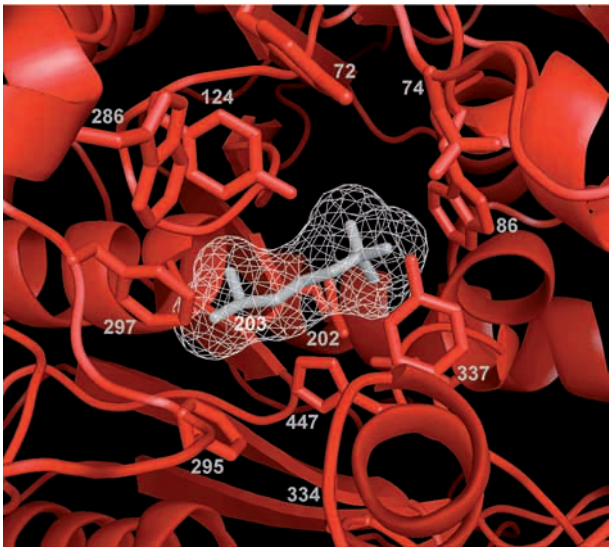

Figure 3 Structure of the active center of mouse acetylcholinesterase. Shown are residue side chains that constitute the catalytic triad (Ser203, His447, and Glu334), the acyl pocket (Phe295 and Phe297), the choline subsite (Trp86 and Try337, among other residues), and the peripheral site near the gorge entrance (Trp286, Try72, and Try124). Asp74, residing midway down the gorge, may have a gating or orienting role for entering cationic substrates.

Inhibitor Interactions

The primary AChE inhibitors used clinically or of toxicologic consequence associate with the active center. These include reversibly bound inhibitors, such as edrophonium and tacrine, that associate with the choline subsite; carbamoylating agents, such as physostigmine and neostigmine, that carbamylate transiently the active center serine; the alkyl phosphate agents, such as diisopropyl fluorophosphate (DFP), echothiophate, and diazinon, and nerve agents such as sarin and soman, that phosphorylate or phosphonylate the serine, forming a stable bond. The alkyl phosphates that form a dimethoxyphosphoryl or diethoxyphosphoryl conjugates with the serine are primarily used as pesticides, whereas the methylphosphonylating agents, such as sarin, have been insidiously employed in chemical terrorism in isolated areas of the globe.

Reactivating Agents and Antidotes

Oximes containing a pyridinium aldoxime moiety serve as nucleophiles for reactivating the alkyl phosphate conjugates with the enzyme and are used as antidotes. Their quaternary structure limits their capacity to reactivate alkyl phosphate-inhibited enzyme in the central nervous system. Atropine and benzodiazepines are also used as antidotes to control excessive muscarinic stimulation and convulsions.

See also: Acetylcholine Neurotransmission in CNS; Acetylcholinesterase Inhibitors and Alzheimer's Disease; Basal Ganglia: Acetylcholine Interactions and Behavior; Cholinergic System; Cholinergic Pathways in CNS; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Neuromuscular Junction (NMJ): Acetylcholinesterases; Neutrotransmission and Neuromodulation: Acetylcholine; Nicotinic Acetylcholine Receptors.

Further Reading

Bourne Y, Taylor P, and Marchot P (1995) Acetylcholinesterase inhibition by fasciculin: Crystal structure of the complex. *Cell* 83: 493–506.

Cygler M, Schrag JD, Sussman JL, et al. (1993) Relationship between sequence conservation and three-dimensional structure in a large family of esterases, lipases, and related proteins. *Protein Science* 2: 366–382.

Dvir H, Harel M, Bon S, et al. (2004) The synaptic acetylcholinesterase tetramer assembles around a polyproline II helix. *EMBO Journal* 23: 4394–4405.

Giacobini E (ed.) (2000) *Cholinesterases and Cholinesterase Inhibitors*. London: Martin Dunitz Publishers.

Giacobini E (ed.) (2003) *Butyrylcholinesterase – Its Function and Inhibitors*. London: Martin Dunitz Publishers.

Hotelier T, Renault L, Cousin X, et al. (2004) ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins. *Nucleic Acids Research* 32: D145.

Li Y, Camp S, and Taylor P (1993) Tissue-specific expression and alternative mRNA processing of the mammalian acetylcholinesterase gene. *Journal of Biological Chemistry* 268: 5790–5797.

Massoulie J (2002) The origin of the molecular diversity and functional anchoring of cholinesterases. *Neurosignals* 11: 130–143.

Quinn DM (1987) Acetylcholinesterase: Enzyme structure, reaction dynamics, and virtual transition states. *Chemical Reviews* 87: 955–979.

Radić Z and Taylor P (2001) Interaction kinetics of reversible inhibitors and substrates with acetylcholinesterase and its fasciculin 2 complex. *Journal of Biological Chemistry* 276: 4622–4633.

Radić Z and Taylor P (2006) Structure and function of cholinesterases. In: Gupta R (ed.) *Toxicology of Organophosphate and Carbamate Compounds*, pp. 161–186. Amsterdam: Elsevier.

Sussman JL, Harel M, Frolow F, et al. (1993) Atomic structure of acetylcholinesterase from *Torpedo californica*: A prototypic acetylcholine-binding protein. *Science* 253: 872–879.

Taylor P (2006) Acetylcholinesterase agents. In: Brunton LL (ed.) *Goodman & Gilman's Pharmacological Basis of Therapeutics*, 11th edn., pp. 201–216. New York: McGraw-Hill.

Relevant Website

http://bioweb.ensam.inra.fr – ESTHER Database (analysis of protein and nucleic acid sequences homologous to cholinesterases).

Acetylcholinesterase Inhibitors and Alzheimer's Disease

H Kaduszkiewicz and H van den Bussche,
University Medical Center Hamburg-Eppendorf,
Hamburg, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The age specific prevalence of Alzheimer's disease (AD) doubles every 5 years of life, affecting 1.2% of those aged 65–69, 2.8% of those aged 70–74, 6.0% of those aged 75–79, etc. The disease causes irreversible memory loss, behavioral and cognitive decline, personality changes, and a decreasing ability to cope with everyday life. Amyloid plaques, neurofibrillary tangles, and a loss of neurons are the typical cellular pathology observed in the central nervous system of AD patients, though these changes are not pathognomonic. In the 1970s, postmortem examinations of the brains of patients with largely end-stage AD also reported markedly reduced activity of choline acetyltransferase, an acetylcholine-synthesizing enzyme. The cholinergic deficits were reported to correlate with cognitive deficits and disease severity. They were also linked to neuropsychiatric symptoms of AD, when medial frontal and limbic cholinergic deficits were identified. The discovery of widespread loss of cholinergic innervation triggered development of cholinesterase inhibitors, which aim to raise acetylcholine levels in the brain by blocking the enzymes metabolizing this molecule. Donepezil, rivastigmine, and galantamine are the three acetylcholinesterase (AChE) inhibitors licensed for the treatment of mild to moderately severe AD. Their characteristics are listed in Table 1.

Preclinical Testing

Originally, preclinical testing of the AChE inhibitors was promising. For example, donepezil at doses of 2.5 mg kg^{-1} showed maximum increases of the extracellular acetylcholine concentration in the hippocampus of rats of 499% of the prelevel at about 1.5 h after administration. The time courses of brain AChE inhibition with donepezil correlated strongly with the extracellular acetylcholine-increasing action. Donepezil demonstrated efficacy in tests of reference memory in animals, but had less consistent activity in tests of working memory. AChE activity was also inhibited in samples of postmortem human brain, fresh brain cortex biopsies, and human erythrocytes. In addition, rapid, sustained, and dose-dependent inhibition of cerebral spinal fluid AChE was demonstrated in AD patients.

Some findings suggested a neuroprotective potential of AChE inhibitors. For example, donepezil was found to attenuate A-beta(25-35)-induced toxicity in rat pheochromocytoma PC12 cells in clinically relevant concentrations. Additional hope came from findings of special action of galantamine directly interacting with nicotinic acetylcholine receptors. These allosterically potentiating ligands sensitize nicotinic receptors by increasing the probability of channel opening induced by acetylcholine and nicotinic agonists and by slowing down receptor desensitization.

Clinical Efficacy

Despite the promising preclinical findings, the evaluation of clinical efficacy did not reveal striking effects. In fact, the study results are interpreted so diversely that 10 years after licensure, the clinical efficacy of AChE inhibitors is still controversial.

In this article, the manifold discussions on the study results on AChE inhibitors will be presented, as they are crucial for the estimation of clinical efficacy. There is no direct way to deduce a clinical benefit from study results. In fact, there is still a long way to go to accomplish this, namely:

1. Selection of relevant studies
2. Identification of the methodological quality of the studies
3. Assessment of the relevance of the methodological quality for the results
4. Discussion of the external validity of the trials
5. Determination of the clinical relevance of effects and side effects measured in the trials
6. Comprehension of the clinical experience of patients, caregivers, and health personnel

Not until these steps are taken is an estimation of clinical benefit possible.

Selection of Relevant Studies

In order to assess the efficacy of cholinesterase inhibitors, there is wide consensus that the analysis of randomized controlled double blind trials is the first choice. In these studies the agent should have been tested against placebo, and clinical parameters should have been used as outcome measures. To date, 18 randomized controlled trials (RCTs) on donepezil, five on rivastigmine, and seven on galantamine have been published. The majority of these studies included patients with mild to moderate stages of AD; the minority focused on patients in the severe stage.

Table 1 Characteristics of the three licensed AChE inhibitors for the treatment of AD patients

	Donepezil hydrochloride (Aricept®)	*Galantamine hydrobromide (Reminyl®)*	*Rivastigmine tartrate (Exelon®)*
Active principle	Reversible and highly centrally selective inhibitor of acetylcholinesterase (AChE) with minimal effects on butrylcholinesterase (BuChE)	Dual effect on the cholinergic system: selective competitive and reversible inhibitor of AchE and allosterical modulator of presynaptic nicotinic acetylcholine receptors (nAChR)[a]	'Pseudoirreversible'[b] centrally selective inhibitor (focus on cortex and hippocampus) of AChE and BuChE
Molecule type	Piperidine-based molecule	Tertiary alkaloid	Carbamate type
$t_{1/2}$	Approximately 70 h	7–8 h	0.6–2 h
Bioavailability	100%; 96% bound to plasma proteins	88%; 18% bound to plasma proteins	36%; 40% bound to plasma proteins
T_{max}	4 h	1 h	1 h
Steady state	After 2–3 weeks	After 2–3 days	
Administration	Once daily	Twice daily	Twice daily
Metabolism	One active metabolite (6-O-desmethyldonepezil), hepatic metabolism by CYP 450 isoenzymes 3A4 and 2D6 and glucuronidation	Hepatic metabolism by CYP 450 isoenzymes 2D6 and 3A4	95% metabolism by AChE (decarbamylation); no involvement of the CYP 450 system
Elimination	Excreted in urine and hepatic metabolism	>90% renal elimination of metabolites	Renal elimination of metabolites (>90% within 24 h)

[a]*In vitro*, galantamine binds to a site on nAChR that is different from the binding site of the natural agonist, acetylcholine (described as allosteric, meaning 'other site'). When galantamine and acetylcholine bind simultaneously to nAChR, the response of these receptors to acetylcholine is amplified.
[b]Pseudoirreversible means that AChE is carbamylated and after some hours hydrolyzed, that is, AChE regenerates without new synthesis.

Mostly, cognitive performance was assessed by means of the Alzheimer's Disease Assessment Scale – Cognitive subscale (ADAS-Cog), which ranges from 0 to 70 points, 0 indicating no impairment and 70 points indicating final dementia. On average, the mean difference between the groups receiving AChE inhibitors and placebo was 3 points, which was statistically significant in favor of AChE inhibitors. This corresponds to an effect size of circa 0.5 standard deviations. In addition to the ADAS-Cog scale, the CIBIC-plus scale (Clinician's Interview Based Impression of Change with caregiver input) was often used. This semistructured interview, performed first with the caregiver and then with the patient, provides a global clinical assessment of change. The change between baseline and endpoint assessment is marked on a 7-point Likert-like scale, where 4 represents no change compared with baseline, 1–3, different degrees of improvement, and 5–7, degrees of deterioration. On this scale, the mean difference between the active treatment and the placebo group amounted to some 0.5 points in favor of donepezil.

Concerning the activities of daily living, various instruments were used in the studies. For example, in the rivastigmine and galantamine studies, improvements in the range of 3 points on the DAD (Disability Assessment for Dementia Scale, total range 0–100) and the PDS (Progressive Deterioration Scale, total range 0–100) were seen.

Identification of the Methodological Quality of the Studies

For a comprehensive assessment of methodological quality, several questions have to be answered. As most important aspects the following are considered:

- *Randomization/allocation concealment*. Was the randomization of the patient to the treatment or placebo groups really by chance? And did neither patients nor investigators know who was randomized to which group?
- *Blinding*. Were the investigators assessing the patients' performance reliably blinded to the treatment of the patients? Did they not know about the side effects the patients experienced?
- *Handling of dropouts*. Which strategies were used to include results of those patients into the analyses who dropped out of the study?
- *Violation of the intention-to-treat-principle*. Was the intention-to-treat-principle (i.e., inclusion of all randomized patients into endpoint analyses) violated, and if yes, to what extent?

The methodological quality of the studies altogether can be judged as moderate or even partly insufficient. In the first place, this is due to the fact that in most of the studies the intention-to-treat-principle was violated. Also, in more than half of the studies the dropout rates in at least one of the treatment arms amounted to 20% or more. Finally, precise

information on randomization, allocation concealment, and blinding was mostly missing in the study reports.

Assessment of the Relevance of the Methodological Quality for the Results

After having identified the methodological shortcomings of the studies, the crucial question is whether these shortcomings might have biased the results substantially. To date, this question is controversial.

Two state institutions of health technology assessment have lately extensively evaluated the trials on AChE inhibitors. Both the British National Institute for Health and Clinical Excellence (NICE) and the German Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Economy in Health Care) (IQWiG) confirmed the methodological shortcomings of the trials but did not assume that these shortcomings biased the results substantially. Because a definite estimation of bias caused by methodological shortcomings can only be given when original data are available, the assumption of NICE and IQWiG – who did not have the data – is not conclusive.

Another question concerns the method used to account for dropouts. In most of the studies the LOCF ('last observation carried forward') method was used: when a patient drops out of the study prematurely, the last test results available are used as if they were results of endpoint assessment. In the case of a progressively deteriorating disease like AD, the use of LOCF can lead to better results in groups with higher dropout rates and earlier dropout time points. As in the majority of the AChE inhibitor trials, more patients dropped out of the verum groups; this may have led to better results in comparison to placebo. However, this theoretical criticism can again only be substantiated on the basis of original data. Unfortunately, the pharmaceutical companies do not disclose these data.

Discussion of the External Validity of the Trials

In the majority of the RCTs, patients with other-severe 'not controlled' medical or psychiatric diseases, insulin-dependent diabetes mellitus, other endocrinological dysfunction, asthma, or obstructive pulmonary disease were excluded. A Canadian research group investigated to what extent a cohort of 6424 older adults who were newly dispensed donepezil in Ontario was represented in the RCTs on cholinesterase inhibitors. They found that 51% to 78% of the patients would not have been included in the trials. This means that, strictly speaking, the validity of the results of these trials is limited to some half of the patients.

A second point concerning the external validity of the trials is their short duration, as most of the RCTs do not exceed 6 months. In consequence, there is hardly any RCT-based evidence for treatment lasting for more than 6 months.

Determination of the Clinical Relevance of Effects and Side Effects as Measured in the Trials

What does a mean group difference of 3 points on a 70 points cognitive scale mean? The answer is unclear. A common answer given is that a US Food and Drug Administration (FDA) panel defined a change of 4 points on the ADAS-Cog as clinically relevant. But irrespective of the scientific basis of this consensus definition, these 4 points apply to changes in individuals and not in groups. The answer on the question of clinical relevance could be clearer if the distribution of changes had been published. Assuming a normal distribution, IQWiG calculated that the difference in success rates between the cholinesterase inhibitor and placebo groups amounts to 13–16% if success is defined as an individual improvement of $\geq$4 points. This means that at least six to eight patients have to be treated with the active drug in order to see success in cognition in one patient. Concerning activities of daily living, the number needed to treat amounts to 13–14. But these calculations have to be interpreted with caution, as they are based on the assumption of a normal distribution of the results.

The leading question concerning clinical relevance is the impact of treatment on life quality of patients and caregivers. Does the one out of six to eight patients who scores 4 points better on the ADAS-Cog experience any benefit in everyday life? There is a consensus that the data from the RCTs are not sufficient (either results are inconsistent or life quality was not evaluated) to ascertain a benefit concerning life quality of patients and caregivers. The same applies to institutionalization and a desirable disease-modifying effect: There is no evidence from RCTs that cholinesterase inhibitors delay institutionalization of patients or that they modify disease progression.

The question of side effects is also not controversial. AChE inhibitors have side effects: nausea, vomiting, diarrhea, weight loss, and agitation are typical. They occur in 5–10% of patients taking donepezil, 5–20% of patients taking galantamine, and 10–40% of patients taking rivastigmine. They can occur just in the titration phase of the drug, but can also persist and are the main reason for dropout from the

trials and for discontinuation of therapy in clinical practice.

Comprehension of the Clinical Experience of Patients, Caregivers, and Health Personnel

Clinicians state that they see a clear benefit in 5% to 15% of their patients, but it has not been possible to identify these responders before treatment. Here again, hypotheses could be deduced from the studies that already exist, but the pharmaceutical companies do not make the original data available.

Another point is that some clinicians or caregivers point out that they see improvement in areas that are difficult to measure in standardized tests. For example, in their opinion the patient appears to be better regulated. Such an observation is in fact difficult to measure and to quantify. One of the latest studies on galantamine tried to overcome this dilemma and measured treatment outcomes by means of the Goal Attainment Scaling (GAS). The GAS is a personalized outcome measure in which people set goals according to their own needs and define improved or worsened states in their own words. After 16 weeks, at the end of double blind treatment, clinicians reported statistically significant changes on the GAS, whereas patients/caregivers did not. Even though the data do not support the assumption that the GAS is more sensitive in detecting treatment effects than other commonly used instruments in cholinesterase inhibitor trials, this instrument highlights the views of patients and caregivers as to whether treatment is seen as meaningful. Further research on this subject is needed.

Consequences of Unclear Clinical Benefit

Reflecting the controversial estimation of clinical benefit, national regulations concerning prescription and treatment documentation differ greatly. For example, in Germany anybody with mild to moderate AD can be treated with cholinesterase inhibitors as long as the treating doctor prescribes the drug, without any further restrictions. In many countries (e.g., Belgium, France, the UK, Australia), in contrast, there are clear rules concerning initiation, evaluation, and termination of therapy. In the Netherlands, donepezil and galantamine (and memantine) are not covered by the statutory health insurance – only rivastigmine is – on the basis of a European Union-wide admission procedure. In Belgium, a patient registry is to be established to follow these patients. Data from such registries could be used to analyze prescription patterns and effects in the great variety of patients treated under real life conditions.

Imprecise Cost Estimates

Costs for the drugs in Germany amount to $1600 to $2000 per year. Not included are all other costs incurred, for example, costs of doctor consultations to get the prescription or costs of diagnosis and treatment of side effects.

The data basis for cost estimates in relation to efficacy is extremely arguable. In their final appraisal document, the NICE reports amounts of $39 000 to $260 000 per quality adjusted life year gained. A huge range is also reported for the costs of gaining an additional year in a nonsevere state: $2200 up to $26 000. All these cost calculations are difficult to follow.

Summary

- The appropriateness of including RCTs only in assessments of efficacy of cholinesterase inhibitors is questioned, but there is a broad consensus that RCTs constitute the basis.
- The published RCTs on cholinesterase inhibitors have methodological shortcomings.
- The relevance of the methodological shortcomings is valued differently.
- Without original data it is not possible to prove whether the methodological shortcomings substantially biased the results or not.
- If the methodological shortcomings are ignored, the study results show small differences between the cholinesterase inhibitor and placebo groups. The clinical relevance of these differences is controversial.
- There is no evidence from RCTs concerning benefits on life quality of patients and caregivers and delay of institutionalization.
- Due to broad exclusion criteria and the duration of the trials mostly not exceeding 6 months, the external validity of the trials is rather low.
- Cholinesterase inhibitors have side effects.
- No disease-modifying effect has been able to be measured.
- Trial data from the industry should be accessible for research purposes.
- Further research directed at improved pharmacological therapy for patients with AD is needed.

See also: Acetylcholine Neurotransmission in CNS; Acetylcholinesterase; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Basal Forebrain and Memory; Cholinergic System; Cholinergic Pathways in CNS; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Nicotinic Acetylcholine

Receptors; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Burns A and O'Brien J on behalf of the BAP Dementia Consensus Group (2006) Clinical practice with anti-dementia drugs: A consensus statement from British Association for Psychopharmacology. *Journal of Psychopharmacology* 20(6): 732–755.

Gill SS, Bronskill SE, Mamdani M, et al. (2004) Representation of patients with dementia in clinical trials of donepezil. *Canadian Journal of Clinical Pharmacology* 11(2): e274–e285.

IQWiG (2006) Cholinesterasehemmer bei Alzheimer Demenz [Cholinesterase inhibitors in Alzheimer's disease]. Vorbericht A05/19-A. Köln: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG).

Kaduszkiewicz H, Zimmermann T, Beck-Bornholdt H-P, and van den Bussche H (2005) Cholinesterase inhibitors for patients with Alzheimer's disease: Systematic review of randomized clinical trials. *British Medical Journal* 331: 321–327.

National Institute for Health and Clinical Excellence (NICE) (2006) Alzheimer's disease – Donepezil, rivastigmine, galantamine and memantine (review) – Final appraisal document. 26 May 2006.

Rockwood K, Fay S, Song X, MacKnight C, Gorman M, and Video-Imaging Synthesis of Treating Alzheimer's Disease (VISTA) Investigators (2006) Attainment of treatment goals by people with Alzheimer's disease receiving galantamine: A randomized controlled trial. *Canadian Medical Association Journal* 174(8): 1099–1105.

Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines

S Halpain, B Calabrese, and L Dehmelt, The Scripps Research Institute, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The range and complexity of cell morphologies found in neurons are largely defined by the structure and dynamic organization of the underlying neuronal cytoskeleton. Neurons contain three major cytoskeleton types: microtubules, intermediate filaments, and filamentous actin. Here we review the properties and roles of actin in early and late neuronal development.

In the brain, neurons develop within an intricate, layered, three-dimensional matrix, which makes it difficult to analyze the detailed organization of the neuronal cytoskeleton. Therefore, many studies on actin organization during neuromorphogenesis have been performed using dissociated neuronal cultures grown on glass coverslips. In such a simplified model system, neurons first display a quasi-symmetrical shape. Actin is concentrated in the cell periphery in one or more flattened, veil-like structures, usually referred to as lamellipodia. During neurite initiation these lamellipodia divide into smaller lamellipodia, which then slowly move outward, trailed by a condensed, microtubule-rich shaft. If the shaft reaches significant length, the whole neuronal protrusion is called a neurite, and the lamellipodium at the neurite tip is called a growth cone (Figure 1, left).

In one of the classical neuronal model systems – primary hippocampal neurons – one of the neurites will grow significantly longer and accumulate certain markers, such as antigenicity to the tau-1 antibody. This neurite usually continues to grow rapidly and will become the future axon. During this stage, the remaining smaller neurites are called minor neurites, and these eventually mature and become the future dendrites.

As the dendritic arbor develops, many filopodia rapidly protrude and retract from the dendritic shaft, with a lifetime of minutes. Over a protracted period of time, the number of filopodia declines, as the number of synaptic contacts and dendritic spines increase. Dendritic spines are the postsynaptic receptive regions of most excitatory synapses. They consist of a bulbous head connected to the dendritic shaft by a narrow neck (Figure 1, right). Normally each spine is connected across the synapse to a specialized region of an axon called a presynaptic bouton or nerve terminal. It is thought that the shape of spines is critical for the normal function of such synapses and essential for the normal development of neuronal networks.

In addition to being one of the most abundant cellular proteins, actin is also one of the most versatile building blocks in cells. By itself, monomeric actin (G-actin) is a simple globular protein with intrinsic adenosine triphosphatase (ATPase) activity. If ATP is bound, G-actin can polymerize into a linear, two-stranded helical filament (F-actin). The intrinsic ATPase activity converts ATP-actin monomers to ADP-actin. The two ends of an actin filament have distinct biochemical activities: the (+) end has a higher affinity for monomeric actin and therefore elongates faster under standard conditions. The (−) end has lower monomer-binding activity and elongates slower than the (+) end does. At equilibrium, this difference of affinities leads to a process termed 'treadmilling,' which is characterized by net monomer addition at the (+) end and net depolymerization at the (−) end. Filament behavior is regulated by a variety of actin-regulating proteins that either bind to the (+) end, the (−) end, or along the sides of filaments.

The overall organization of actin filaments in cells can be controlled by regulating the local rates of *de novo* filament formation (nucleation), addition of monomers to existing filaments (polymerization), removal of monomers from filament ends (depolymerization), breaking of intact filaments (severing), sliding of individual filaments, and physical cross-linking of filaments to other cellular structures. Throughout this article, we discuss various examples of how precise tuning of such activities within neurons leads to the specific organization and dynamic properties of key neuronal structures.

Early Development

Actin Nucleation and Control of Actin Polymerization Rates in Growth Cones

The 'dendritic nucleation model' is a popular model for how actin filaments might maintain and control protrusive cell behavior at the plasma membrane. In this case the term 'dendritic' refers not to neurites but to the branched organization of actin filaments that defines this particular mode of actin assembly. The dendritic nucleation model is often thought to be generally applicable; however, most data that led to its proposal were derived from nonneuronal cells, notably fibroblasts. In a nutshell, this model proposes that rapid nucleation and ATP-dependent actin

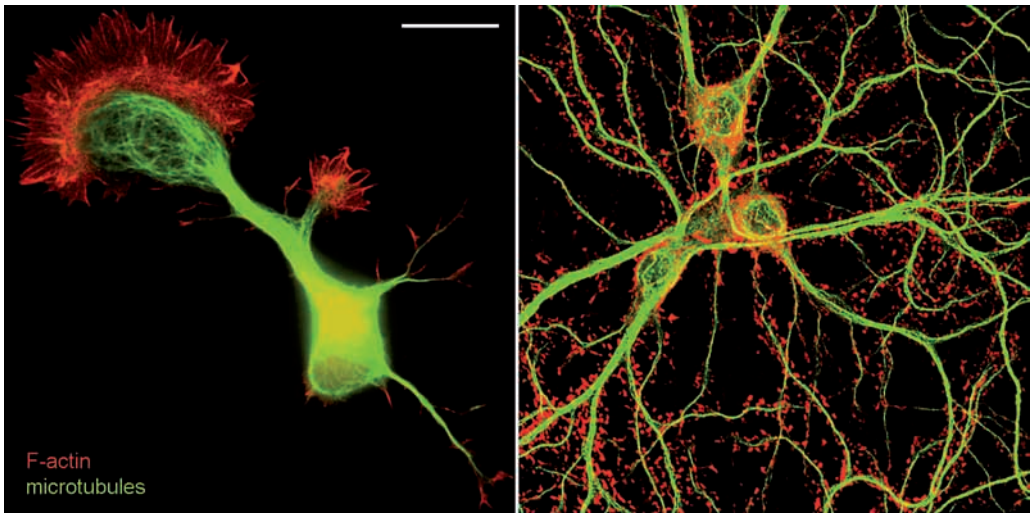

Figure 1 Localization of filamentous actin (F-actin) in rat cultured hippocampal neurons. Left: hippocampal neuron (1 day after plating) stained for F-actin (red) and neuronal tubulin (green). Right: hippocampal neurons (3 weeks after plating) stained for F-actin (red) and MAP2 (a microtubule-associated protein specific to dendrites; green). Note that F-actin is enriched in small clusters along the dendrites, which correspond to dendritic spines. Scale bar = 20 μm.

polymerization at the leading edge of motile cells are sufficient to drive membrane protrusion. To be effective, the turnover of actin must be very high, in order to maintain a constant supply of monomers to drive protrusion. Therefore, depolymerization must be ensured shortly after the initial polymerization. The zones of polymerization and depolymerization are separated by a zone of active myosin motor- and actin polymerization-driven flux. In this zone, the level of adhesion is thought to be important to regulate the level of protrusion or slippage – a process first suggested in the 'clutch hypothesis' (see below).

Some key aspects of this model are likely to apply to growth cones as well. However, there are also some very significant differences. Most notably, the dendritic nucleation model proposes that actin filament nucleation at the leading edge is driven by a actin-related protein (ARP) assembly termed the ARP2/3 complex, which binds to the sides of existing filaments and primes the nucleation of a new filament in a characteristic 60° angle off the mother filament. In fibroblasts, this mechanism leads to the formation of an intricate network of interdigitated filaments of 'dendritic' (branched) appearance. However, such dendritic networks are not observed in neuronal growth cones, suggesting that the way in which actin filaments initially form differs significantly between neuronal and nonneuronal systems.

It is believed that in addition to the ARP2/3 complex, another class of actin-binding proteins called formins plays an important role in actin nucleation. In the case of growth cones, formin-mediated actin nucleation might be dominant. In fact, whereas ARP2/3-induced nucleation is necessary for fibroblast protrusion, in growth cones, this nucleation complex is a negative regulator of neurite protrusion, suggesting a distinct role for ARP2/3 in the growth cone. Interestingly, ARP2/3-mediated cell protrusions are typically characterized by broad, flat lamellipodia, usually found at the leading edge of fibroblasts, whereas formin-based cell protrusions are characterized by actin spike structures called filopodia, which in turn are frequently formed in growth cones (**Figures 1 (left) and 2(a)**).

Apart from this obvious difference in the actin nucleation mechanism, other aspects of actin polymerization regulation appear to be similar in fibroblasts and growth cones. In both systems, for example, rapid polymerization at the leading edge, promoted by actin bound to profilin, is thought to drive membrane protrusion. β-Thymosin is present in both neuronal and nonneuronal systems and is known to sequester actin monomers, thereby limiting polymerization rates. Finally, depolymerization is thought to be the rate-limiting step in actin turnover control (see the section titled 'Control of actin depolymerization and turnover in early neuromorphogenesis') and thus is thought to be a key modulator of polymerization rates as well.

Substrate Adhesion and Generation of Traction in Axon Outgrowth and Guidance

Once actin is polymerized at the leading edge of the growth cone, it enters a zone of rapid retrograde flow usually referred to as the lamellipodium. According to the 'clutch hypothesis' proposed by Mitchison and

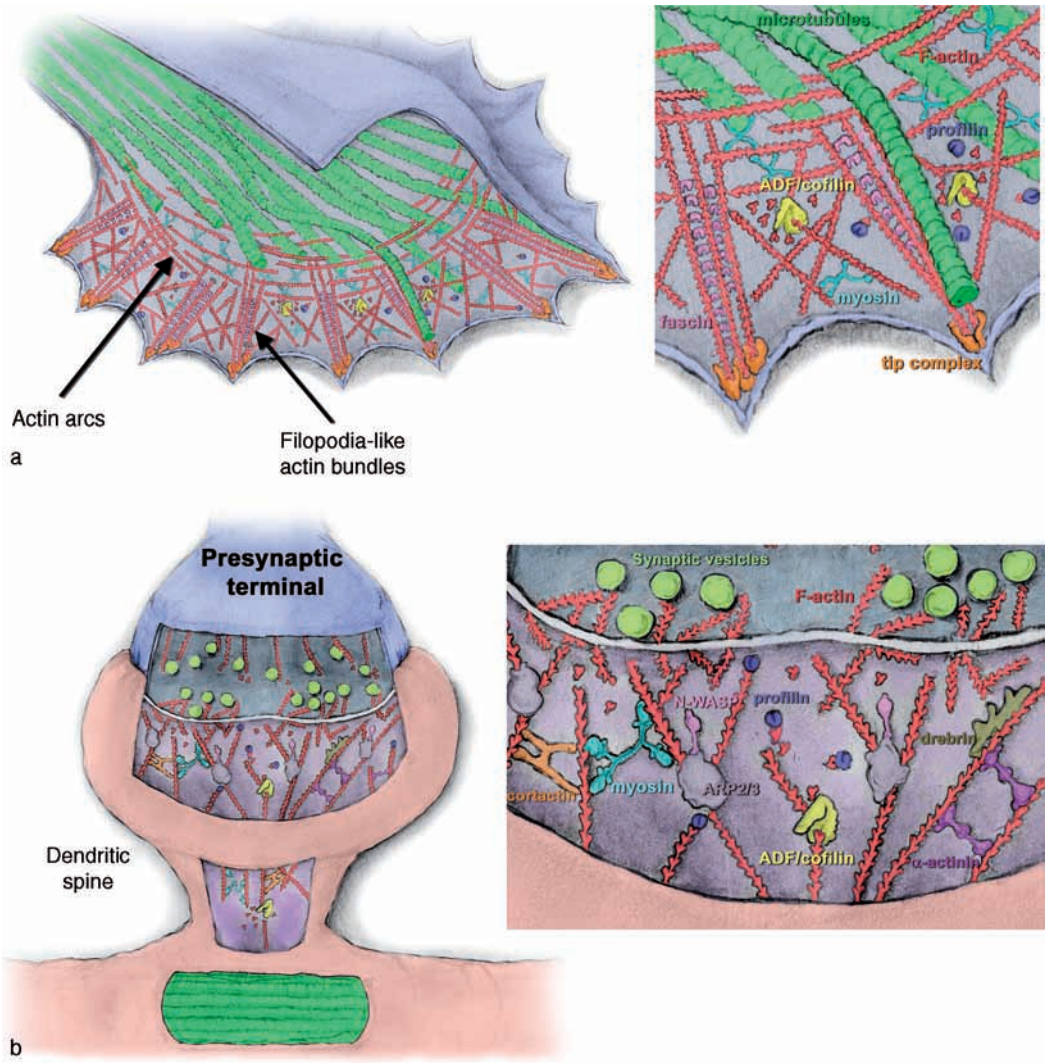

Figure 2 Cytoskeletal organization in the growth cone and the synapse. (a) Within the growth cone, actin is concentrated in the peripheral lamella. Actin-binding proteins, such as profilin, cofilin, and myosin, are thought to play a role in regulating actin dynamics and/or organizing actin into specific structures. (b) Actin filaments are distributed throughout the presynaptic nerve terminal, preferentially concentrated around synaptic vesicles. However, mature nerve terminals contain much less actin compared to either growth cones or dendritic spines. Actin regulatory molecules (such as cofilin and profilin) control the extent and rate of actin polymerization, and, ultimately, dendritic spine shape or growth cone motility. ADF; actin-depolymerizing factor; N-WASP, neural Wiskott–Aldrich syndrome protein; ARP2/3, actin-related protein complex 2/3. Drawings by James Lim; colorization by Barbara Calabrese and Leif Dehmelt.

Kirschner, this retrograde flow acts as a motor for cell protrusion. By analogy to a car engine clutch, the retrograde flow is constitutively active, even in resting growth cones. This retrograde flow does not produce any forward movement unless it is linked to the substratum by formation of cell adhesions. Thus, the temporal regulation of the extent of adhesion or slippage is thought to control the speed of growth cone progression.

The giant growth cones of the sea snail *Aplysia* have been instrumental in revealing how local manipulations of adhesion affect growth cone behavior and protrusion. Such studies revealed that adhesion to a bead coated with the signaling molecule ApCAM (*Aplysia* cell adhesion molecule) is not sufficient by itself to initiate a full response of the growth cone. However, if the coated bead is mechanically restrained to allow the growth cone to pull on it, Src kinase is locally activated at the bead interaction site. This is followed by microtubule targeting toward the bead, modulation of retrograde flow, growth cone advance, and axonal turning behavior.

Retrograde flow is driven both by forces generated distally from actin polymerization at the leading edge and proximally by myosin motor activity within the lamellipodium (Figure 2). There is still some

controversy about the roles of myosin II and myosin I in growth cones. These motors have been suggested either to participate in the organization and maintenance of structural integrity of the lamellipodium or in the generation of the retrograde flow. It is clear, however, that myosin II is mostly localized to the proximal part of the growth cone, where it co-localizes with contractile filamentous structures called actin arcs. These structures accumulate at the lateral sides of the growth cone and are thought to compress microtubules into parallel arrays. Furthermore, actin arcs are also thought to mediate contractile responses during neurite retraction.

Roles for Actin–Microtubule Interactions in Early Neurite Development

Several laboratories have reported physical, rigid interactions between microtubules and filamentous actin within cells. This physical coupling might have several nonexclusive roles: (1) As mentioned earlier, actin arcs might shape the distribution of microtubules within growth cones and might even be involved in the formation and/or maintenance of the tight microtubule bundles typically found in neurons. (2) Similar to the clutch hypothesis, physical linkage of actin filaments to a stable scaffold of microtubules might promote protrusion of the leading cell edge. (3) Microtubules might be involved in signaling cross-talk mediated by regulatory molecules found on the two cytoskeletons. (4) Dynein-generated forces on microtubule arrays are suggested to counteract actomyosin-driven neurite retraction. Interestingly, microtubules also appear to have an instructive role in growth cone turning, as local pharmacological manipulations of microtubule dynamics are sufficient to drive growth cone turning behavior.

Several potential microtubule-actin cross-linkers or signaling mediators have been identified, including the coronin-like protein pod-1, microtubule-associated protein MAP2c, the dynein–dynactin complex, and the IQGAP1/Rac1 complex. These proteins have been implicated in several stages of early neurite development, including neurite initiation, neurite growth, and axon pathfinding.

Control of Actin Depolymerization and Turnover in Early Neuromorphogenesis

The dynamic behavior of actin filaments is not only dependent on the rates of polymerization, nucleation, and motor-dependent filament sliding, but is also dependent on a constant supply of active monomers that can be incorporated at the filament ends. Thus, depolymerization is required for rapid actin turnover, and biochemical studies suggest that this might be the rate-limiting step in the cycle of actin dynamics. Therefore, actin-depolymerizing factors (ADFs), such as cofilin and ADF, play important roles as regulators of actin dynamics.

ADF/cofilin binds filamentous actin and increases its depolymerization rate by altering the helical twist of the filament. The activity of ADF/cofilin is controlled by multiple mechanisms, including phosphorylation of serine 3 by LIM and TES protein kinases, dephosphorylation by the protein phosphatase 'slingshot' and 'chronophin' and binding of 14-3-3 proteins. Furthermore, another actin-binding protein called tropomyosin protects filaments against ADF/cofilin and thereby adds another level of depolymerization rate regulation.

Consistent with the idea that ADF/cofilin regulates a rate-limiting step during actin turnover, overexpression leads to an increase in axonal outgrowth, while inhibition of its activity using dominant-negative mutants or by overexpression of an inactivating kinase slows axon outgrowth.

Late Development

Actin Organization and Function during the Development of Synapses

The actin cytoskeleton is the main structural component of both pre- and postsynaptic terminals. Here we focus on excitatory synapses, where the function of the actin cytoskeleton is best understood.

The formation of synapses in the vertebrate central nervous system is a complex process that occurs over a protracted period of time. It is generally the case that dendritic filopodia are the structural precursors of dendritic spines during synaptogenesis, although direct emergence of new spines from the dendrite shaft has also been observed. Dendritic filopodia are long, thin protrusions characterized by a highly transient and motile behavior. In these protrusions F-actin is sparse until pre- and postsynaptic transmembrane cell adhesion proteins, such as the cadherin–catenin complex, link adjacent synapse precursors. The F-actin accumulates within early filopodial synaptic contacts, where it recruits synaptic signaling molecules. Simultaneously, the dynamic dendritic filopodia begin to morphologically change into intermediate structures sometimes called 'protospines.' These intermediate structures have also been referred to as cluster-type filopodia, or synaptic filopodia, and they can eventually mature into more stable dendritic spines. Only a fraction of the filopodia that emerge from a dendrite persist to become dendritic spines.

Mature, postsynaptic dendritic spines are highly enriched in F-actin compared to the dendritic shaft (Figure 1, right). Based on measurements of green

fluorescent protein (GFP)–actin fluorescence recovery after photobleaching in cultured hippocampal neurons, only approximately 5% of the total actin in spines is relatively stable, while the majority of the actin almost completely turns over in a 2 min period. Unlike F-actin, G-actin is distributed homogeneously throughout axonal and dendritic processes in hippocampal neurons, which indicates that its main role at the synapse is to maintain the synaptic F-actin pool. The characteristic shape of dendritic spine heads is highly dependent on actin dynamics and the fraction of polymerized actin. Disruption of filamentous actin by the toxin latrunculin A transforms dendritic spines into filopodia-like processes.

F-actin is found within the presynaptic nerve terminal and is preferentially concentrated around synaptic vesicles (Figure 2(b)). In resting conditions, approximately 30% of actin is in a polymerized state, with a turnover halftime of 20 s. During electrical activity, actin is further polymerized and recruited around vesicles from nearby axonal regions, leading many groups to propose a propulsive role for actin, either in maintaining the vesicle cluster or in guiding vesicle recycling. However, the precise role for actin in nerve terminals remains controversial, and may differ depending on the nature of the synapse. Although actin plays an active role in synaptic vesicle endocytosis in some systems, recent evidence indicates that in mammalian central synapses it passively inhibits vesicle fusion at the active zone, and helps sequester synaptic vesicles in the reserve pool. Synaptic vesicle trafficking from the reserve pool to the readily releasable pool can occur independently of actin. Thus, actin may serve primarily as a scaffold for other regulatory proteins involved in vesicle trafficking within the terminal. The recruitment and retention of presynaptic matrix proteins, such as bassoon and piccolo, also are dependent on actin.

The higher content of F-actin in dendritic spines compared to nerve terminals correlates with more dynamic shape changes. Nevertheless, the pre- and postsynaptic cytomatrix components are strongly interconnected, and therefore their movements are physically coupled in established synapses. Engulfment of spinules (membrane extensions emerging from the head of big mushroom spines) by presynaptic axons is a process known as transendocytosis and suggests retrograde signaling or coordinated remodeling of pre- and postsynaptic membranes, which is likely modulated by actin.

Actin Dynamics and Spine Morphogenesis

In recent years, imaging techniques both *in vivo* and *in vitro* have revealed that a majority of mature dendritic spines are stable over months in the adult brain. They rarely form and disappear; instead, the dendritic spine head is constantly changing shape (morphing). This rapid spine motility is caused by regulated polymerization and depolymerization of actin, based on studies using drugs and cDNAs that perturb actin polymerization.

The Rho family of GTPases is one common denominator controlling spine morphology. Multiple signaling pathways converge on the Rho family of small GTPases (particularly RhoA, Rac1, and Cdc42), which ultimately converge on the actin cytoskeleton to regulate spine morphology and dynamics. Generally speaking, RhoA inhibits, whereas Rac and Cdc42 promote, the growth and/or stability of dendritic spines.

An expanding number of guanine nucleotide exchange factors (GEFs) and GTPase-activating proteins (GAPs) control these Rho family GTPases. For example, kalirin-7, a GEF for Rac1, mediates effects of the transmembrane signaling molecule EphB on spine maturation by activating Rac1 and its downstream effector p21-activated kinase (PAK). PAK1 and PAK3 are Rac effectors that promote formation and/or growth of spines.

PAK also stimulates LIM kinase (LIMK), which phosphorylates and inactivates ADF/cofilin, an actin-depolymerizing protein that mediates reorganization of the actin cytoskeleton. Recently, cofilin has been detected by immunoelectron microscopy to concentrate in the periphery of the spine head and within the postsynaptic density. Mice expressing dominant-negative PAK in the forebrain showed reduced spine density in cortical neurons, impaired long-term synaptic plasticity, and reduced memory consolidation. PAK activity is markedly reduced in Alzheimer's disease due to a reduction in phospho-PAK, which in turn results in an increased cofilin activity and downstream loss of the spine actin-regulatory protein drebrin.

Insulin receptor substrate p53 (IRSp53), an adaptor protein that connects Rac1 to the Wiskott–Aldrich syndrome protein WAVE2, is implicated in filopodia and lamellipodia formation in nonneural cells. IRSp53 is enriched in the postsynaptic density (PSD) by binding to Shank and PSD-95 family scaffolds. Overexpression of IRSp53 increases spine density, whereas RNA interference (RNAi) or dominant-negative inhibition of IRSp53 causes reduction of spine density and size. The Abl-interactor (Abi) adaptor proteins, so-named because they bind to the Abl tyrosine kinases, also play a part in Rac GTPase signaling and actin regulation. Abi2 is highly expressed in brain; and Abi2-deficient mice show reduced spine density and a decrease in the relative proportion of mushroom spines, which is thought to be associated with deficits in memory.

It is worth noting that components of Rho GTPase signaling pathways are highly represented among the genes so far identified in hereditary forms of human nonsyndromic mental retardation. These mutant genes include X-linked genes, such as PAK3 (see earlier) and oligophrenin, which is a Rho GAP. RNAi suppression of oligophrenin decreases spine length, mimicking the effect of active RhoA.

Numerous scaffold proteins and actin-binding proteins are concentrated in spines and are known to control spine morphogenesis (Figure 2(b)). Many of these are regulated by the small GTPases described earlier. Overexpression of α-actinin 2, a protein linking *N*-methyl-D-aspartate (NMDA) receptors to the actin cytoskeleton, increases the length and density of dendritic protrusions in cultured hippocampal neurons. Clustering of the actin-binding protein drebrin occurs early in development, supporting its involvement in spinogenesis and synaptogenesis. Cortactin is an activator of the ARP2/3 actin nucleation machinery that interacts with the postsynaptic density scaffold Shank. RNAi knockdown of cortactin results in depletion of dendritic spines, whereas overexpression of cortactin causes enlargement of spines.

Myosin II is also enriched in postsynaptic density preparations and acts as a molecular motor that produces tension on actin filaments. Inhibition of myosin IIB destabilizes mushroom spines, suggesting that the structure and function of spines are regulated by an actin-based motor in addition to actin polymerization.

Effects of Synaptic Activity on Actin Polymerization in Dendritic Spines

Synaptic plasticity is associated with changes in actin polymerization and spine morphology. Long-term potentiation (LTP) is associated with formation of new spines, enlargement of existing spines, and a shift of the actin turnover equilibrium toward polymerization of F-actin. By contrast, long-term depression (LTD) is associated with shrinkage and/or retraction of spines due to an increase in actin depolymerization. In addition, synaptic activity is thought to play an important role in determining whether a synapse will be stabilized or eliminated.

During induction of LTP, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors are rapidly recruited to dendritic spines through Ca^{2+}-permeable NMDA receptor activation. In response to NMDA receptor activation, profilin, a promoter of F-actin assembly, is targeted to spine heads, while cortactin redistributes from spines to the dendritic shaft. Distinct but overlapping signaling pathways are involved in LTD and spine shrinkage. During induction of LTD, NMDA receptor-mediated calcium influx triggers phosphatase 2B (calcineurin) activity. Calcineurin in turn activates a downstream phosphatase 'slingshot' and promotes actin severing and depolymerization through ADF/cofilin.

Synaptic activity can also affect the relationship between F-actin and the plasma membrane of dendritic spines. Myristoylated alanine-rich protein kinase C substrate (MARCKS), a major target of protein kinase C (PKC), might be involved in local coupling of activity-dependent calcium entry with regulation of actin assembly–disassembly cycles. During synaptic activity, when calcium levels increase and stimulate PKC, MARCKS detaches from the plasma membrane, thereby freeing phosphatidylinositol 4,5-bisphosphate to interact with actin-binding proteins such as profilin and ARP2/3, which in turn can locally alter actin dynamics and membrane trafficking.

Comparing Growth Cones, Dendritic Spines, and Nerve Terminals

One might be tempted to generalize the role of actin in growth cones, synaptic terminals, and spines, and introductions to this topic often try to point out similarities. However, the current state of our knowledge does not allow an in-depth comparison. Nonetheless, certain parellels exist (Figures 2(a) and 2(b)). In growth cones, a relatively stable, less dynamic F-actin population is found toward the growth cone center, which is thought to be involved in contractile behavior. Highly dynamic F-actin is concentrated in the periphery of the growth cone, where it is thought to mediate membrane protrusion, resulting in motile filopodia and lamellipodia. Similarly, a stable pool of F-actin is present in the dendritic spine core, while a dynamic pool lies peripherally in its shell, which exhibits highly dynamic, sometime protrusive behavior.

In growth cones, the organization and dynamics of the underlying actin structures have been analyzed using both high-contrast electron microscopy and dynamic speckle microscopy. These analyses converge in a detailed working model of growth cone behavior. In spines, defining the detailed actin organization and dynamics has been more difficult. Ultrastructural analysis of filament polarity suggests that most actin filaments are oriented with their barbed end pointing toward the head of the spine protrusion, similar to the preferred orientation of actin filaments toward the growth cone leading edge. One notable difference is that spines generally lack the dense parallel organization of F-actin that characterizes growth cone filopodia. However, at times, rapidly growing, elongated filopodia-like

structures are observed on spine heads. The purpose of such filopodial-like protrusions emerging from spine heads remains unclear. In growth cones, and along developing dendrites, such filopodial protrusions are thought to extend the structure's sensory radius to probe the environment for guidance cues or neurotransmitters. Speckle microscopic analysis of actin dynamics has not been performed in spines; however, indirect data on spine actin dynamics have been derived from photobleaching recovery experiments (see the section titled 'Actin organization and function during the development of synapses'). These experiments reveal a highly dynamic F-actin population, which may be roughly comparable to the rapidly turning over F-actin of growth cone lamellipodia.

While there might be some similarities between actin behavior in growth cones and spines, the role and organization of actin in nerve terminals appear to be distinct. F-Actin is distributed throughout the presynaptic nerve terminals and is preferentially concentrated around synaptic vesicles. However, mature nerve terminals do not exhibit significant protrusive behavior and contain much less actin compared to either growth cones or spines. In nerve terminals, a specific role for actin remains unclear.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal and Dendritic Identity and Structure: Control of; Cytoskeletal Interactions in the Neuron; Cytoskeleton in Plasticity; Dendritic Spine History; Glutamate Regulation of Dendritic Spine Form and Function; Growth Cones; LIM Kinase and Actin Regulation of Spines; Microtubules: Organization and Function in Neurons; Spine Plasticity.

Further Reading

Calabrese B, Wilson MS, and Halpain S (2006) Development and regulation of dendritic spine synapses. *Physiology (Bethesda)* 21: 38–47.

Carlisle HJ and Kennedy MB (2005) Spine architecture and synaptic plasticity. *Trends in Neurosciences* 28: 182–187.

Dehmelt L and Halpain S (2004) Actin and microtubules in neurite initiation: Are MAPs the missing link? *Journal of Neurobiology* 58: 18–33.

Dent EW and Gertler FB (2003) Cytoskeletal dynamics and transport in growth cone motility and axon guidance. *Neuron* 40: 209–227.

Ethell IM and Pasquale EB (2005) Molecular mechanisms of dendritic spine development and remodeling. *Progress in Neurobiology* 75: 161–205.

Lippman J and Dunaevsky A (2005) Dendritic spine morphogenesis and plasticity. *Journal of Neurobiology* 64: 47–57.

Luo L (2002) Actin cytoskeleton regulation in neuronal morphogenesis and structural plasticity. *Annual Review of Cell and Developmental Biology* 18: 601–635.

Oertner TG and Matus A (2005) Calcium regulation of actin dynamics in dendritic spines. *Cell Calcium* 37: 477–482.

Pollard TD and Borisy GG (2003) Cellular motility driven by assembly and disassembly of actin filaments. *Cell* 112: 453–465.

Sarmiere PD and Bamburg JR (2004) Regulation of the neuronal actin cytoskeleton by ADF/cofilin. *Journal of Neurobiology* 58: 103–117.

Strasser GA, Rahim NA, VanderWaal KE, et al. (2004) Arp2/3 is a negative regulator of growth cone translocation. *Neuron* 43: 81–94.

Suter DM, Schaefer AW, and Forscher P (2004) Microtubule dynamics are necessary for SRC family kinase-dependent growth cone steering. *Current Biology* 14: 1194–1199.

Zhang W and Benson DL (2002) Developmentally regulated changes in cellular compartmentation and synaptic distribution of actin in hippocampal neurons. *Journal of Neuroscience Research* 69: 427–436.

Zhang XF, Schaefer AW, Burnette DT, et al. (2003) Rho-dependent contractile responses in the neuronal growth cone are independent of classical peripheral retrograde actin flow. *Neuron* 40: 931–944.

Action Potential Initiation and Conduction in Axons

J H Caldwell, University of Colorado, Aurora, CO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Transmission of electrical signals in biological systems operates under constraints imposed by the materials used to build the conducting pathways. The conduction medium in nerves is a membrane-enclosed dilute salt solution rather than a metal wire, for example, as in macroscopic physical systems. Axons have been compared for decades to undersea cables that are surrounded by salt water. However, the axon relies on a salt solution both inside and outside the axon. Since the resistivity of mammalian physiological saline is about 1 ohm-meter and that of copper is about 1.7×10^{-8} ohm-meter, the ease of current flow in 1 mm of an axon (a typical length constant in a large-diameter axon) is equivalent to that of a wire that is 170 km long. The difficulty of passive propagation of signals in an axon is compounded by two factors. First, the signal is carried by cell membranes that have a capacitance ($\sim$1 μF cm^{-2}) that needs to be charged to propagate a voltage change. Second, the membrane is inherently leaky due to a requirement for ion channels that are open to set the negative resting membrane potential. Thus, passive conduction of a signal in neurons is effective over distances of only a few millimeters. For this reason, it does not matter whether you are a worm or a whale: transmission of signals in axons requires a booster mechanism to replenish the decrementing electrical signal. The analogy between undersea cables and axons has other similarities and divergences. Undersea cables use repeater or booster stations that amplify the analog signal and send it on. The nervous system also uses a booster system but transforms the analog signal into a digital one (i.e., the continuously varying voltage signal becomes either 'on' or 'off'). This encoding is a trade-off that neurons accept for the sake of long-distance communication. The amplitude of the graded signal is transformed into the timing and frequency of action potentials.

Ionic Basis of the Action Potential in Axons

Sodium and Potassium Channels

The fundamental ionic mechanism of a propagating, regenerative increase in sodium conductance was described by Hodgkin and Huxley for the squid giant axon, and this process holds in both the central nervous system (CNS) and peripheral nervous system (PNS) of both vertebrates and invertebrates. The action potential is generated by the opening and subsequent inactivation of voltage-gated sodium channels and, with a slight delay, the opening of voltage-gated potassium channels. This ionic interplay of opening and closing of sodium and potassium channels is present in such diverse phyla that it must have evolved hundreds of millions of years ago. This conservation of the molecular foundation of the action potential has additional complexities that are not yet completely understood. In mammals there are 10 voltage-gated sodium channels (Nav1 and Nav2 families), and many of these are expressed in neurons and localized in axons. In addition, there are at least 20 voltage-gated potassium channels in vertebrates (Kv1 to Kv4 and KCNQ families). Thus, the molecular species and combinations of sodium and potassium channels utilized for conduction are potentially quite large. This suggests that characteristics of different axon types (e.g., frequency of firing and ability to maintain conduction) are dependent on the isoforms of sodium and potassium channels present in the axon.

Sodium–Potassium Pump

Although the immediate basis of the action potential is the activation of voltage-gated sodium and potassium channels, long-term support of the action potential requires sodium–potassium pumps to maintain the concentration gradients. It has been often emphasized that if the sodium pump is blocked (e.g., with ouabain) in a squid axon, hundreds of thousands of action potentials can still be generated because few sodium and potassium ions cross the membrane for each action potential. The squid axon illustrates the fact that the energy for ion flux (the concentration gradient of sodium and potassium between the inside and outside of the axon) has already been established by the pump. However, the squid giant axon is 1 mm in diameter, and when considering small-diameter axons that are less than a micrometer in diameter (e.g., pain and temperature fibers in the periphery or parallel fibers in the cerebellar cortex), sodium and potassium fluxes during each action potential are significant. For these small-diameter axons, the ability to maintain action potential firing is highly dependent on sodium pump activity.

Structural and Functional Differences of Vertebrate Axons

Axons fall into two major categories depending on the structure of the glial cells that envelop them. The first category is the unmyelinated axon, which describes all invertebrate axons and small axons of vertebrates, typically axons with a diameter below 1 μm. The unmyelinated axon is usually loosely surrounded by a glial cell or in some cases, such as parallel fibers in the cerebellar molecular layer, is not covered by a glial cell. The speed of conduction of an action potential in an unmyelinated axon is proportional to the square root of the axon diameter. Thus, invertebrates have large-diameter axons for signals that need to be propagated rapidly. The squid giant axon, which is part of a circuit used for rapid propulsion in the escape response, is as large as 1 mm in diameter, conducts at $25\,\mathrm{m\,s^{-1}}$ (at 25 °C), and is formed by the fusion of axons of many neurons.

Vertebrates have evolved an alternative strategy for increasing the speed of action potential conduction. Vertebrate axons larger than about 1 μm are tightly wrapped by many layers of the glial cell, creating the second category, the myelinated axon. Myelination occurs in a repeating pattern, with long wrapped regions (internodes that are up to 1–2 mm in length) interrupted by a very short bare region (the node of Ranvier, 1–2 μm in length). In the PNS the glial cell is a Schwann cell. Each Schwann cell can envelop many unmyelinated axons, but when myelination occurs, one Schwann cell is devoted to the formation of one myelinated internode. Myelin in the CNS is formed by oligodendrocytes, and one oligodendrocyte sends out tens of processes, each one forming an internode on a different axon. Functionally, the myelin acts as an insulator, by reducing the leak of current through the membrane in the internodal regions. The myelin also reduces the effective capacitance of the internodal region, which in turn reduces the capacitive current required to charge the internodal membrane. Consequently, current travels rapidly with little loss in the internodal regions to the node of Ranvier, where the current is boosted or regenerated by voltage-gated sodium channels concentrated at the node (described later). Conduction velocity in myelinated axons is proportional to the axon diameter, and the general rule of thumb is that for axons with an outside diameter greater than 11 μm, the speed of conduction, in meters per second, is about six times the axon diameter, in micrometers. For smaller axons the proportionality factor is 4.5. An axon 20 μm in diameter, which is one of the largest in the mammalian nervous system, conducts at $120\,\mathrm{m\,s^{-1}}$, about five times faster than the squid axon, even though it is 50 times smaller. Thus, myelination not only increases speed of conduction, but also does this with an economy of space. This concept of economizing the volume used for conduction is invoked to explain why we have many more unmyelinated axons than myelinated ones: more information can be carried in a given volume with small, unmyelinated axons. In the mammalian nervous system, pain and temperature information is carried by small, unmyelinated axons in the PNS, and the molecular layer of the cerebellum in the CNS is densely packed with parallel fibers that are unmyelinated axons of granule cells.

Initiation of the Action Potential

Stimulation Required for Electrogenesis

A rapid membrane depolarization is necessary to open voltage-gated sodium channels and start the action potential. This depolarization can be achieved artificially by inserting a microelectrode into an axon and injecting current, by extracellularly stimulating an axon with an electrode, or even by mechanically hitting the nerve – for example, when we hit our 'funny bone' (the ulnar nerve near the elbow). Naturally produced depolarizations fall into two categories. For most neurons, synaptic input to the dendrites and cell body provides the required depolarization. Sensory neurons, such as stretch receptors in muscle or cutaneous receptors in the skin, propagate action potentials centrally; a sensory signal in the periphery creates a depolarization called the receptor or generator potential that opens sodium channels to produce action potentials.

Site of Initiation

Unmyelinated axons The simplest treatment of a neuron separates it into three regions: soma, dendrite, and axon. The elementary concept of a passive dendritic tree that simply receives excitatory and inhibitory synaptic inputs and sends these inputs to the cell body, where they are summated, is now known to be an oversimplification. Many neurons have voltage-dependent sodium and calcium channels in dendrites. However, only in rare cases can dendrites initiate action potentials that are propagated orthodromically to the soma. Thus, the integration of excitatory and inhibitory potentials takes place at the soma. The soma contains a variety of voltage-gated channels, but sodium channel density is highest at the axon hillock, an enlargement of the axon at the point it leaves the cell body (Figure 1). For example, sodium channel density is over sevenfold higher on the initial segment of a neurite (presumptive nascent axon),

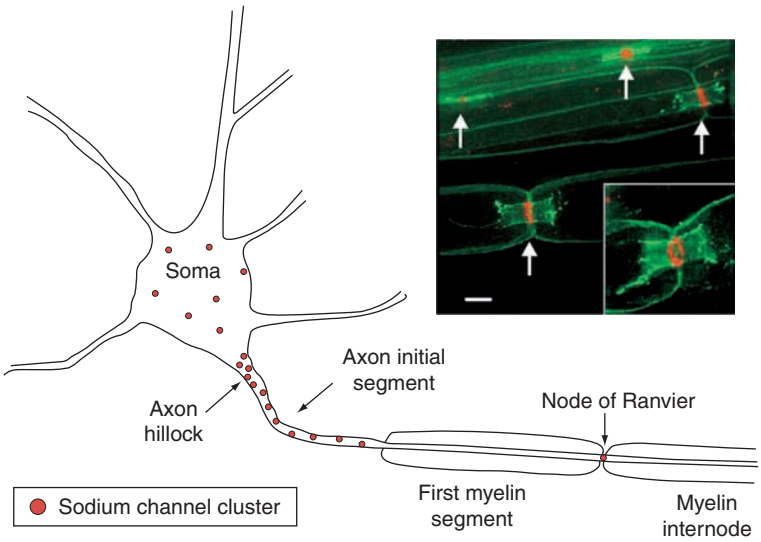

Figure 1 Sites of action potential initiation and sodium channel clustering. Action potentials begin at the point of lowest threshold, which is a function of the balance between sodium, potassium, and leak currents. In general, the action potential originates where sodium channels are clustered at high density, and for both unmyelinated and myelinated axons this is usually at the axon hillock. In myelinated neurons initiation can also occur in the initial segment or at the first node of Ranvier, where sodium channels are highly concentrated. Inset: Four nodes of Ranvier (arrows) in a teased sciatic nerve with nodal labeling by an anti-Nav1.6 antibody (red) and paranodal labeling with an anti-Caspr antibody (green). The node shown at higher power within the inset is tilted such that the labeling of the entire circumference of the nodal membrane can be seen. Scale bar = 5 μm (3.3 μm for the tilted node). (Inset) Reproduced from Caldwell JH, Schaller KL, Lasher RS, et al. (2000) Sodium channel Na(v)1.6 is localized at nodes of Ranvier, dendrites, and synapses. *Proceedings of the National Academy of Sciences of the United States of America* 97: 5616–5620.

compared to the soma in cultured spinal cord neurons. The higher density at the axon hillock has been confirmed with immunolabeling of neurons in many regions of the CNS. The main consequence of the increased sodium channel density at the axon hillock is that the threshold for action potential initiation is lowest there. Thus, for unmyelinated axons, action potential initiation takes place at the axon hillock.

Myelinated axons The site of action potential origination in myelinated axons was also shown to be the axon hillock about 50 years ago. More recently, with improvements in the ability to measure voltage changes optically and with patch clamp electrodes, the precise location of the origin of the action potential has been identified. Action potentials can originate not only at the axon hillock, but also in the axon initial segment, 30–40 μm from the soma and close to the first myelinated segment. In some neurons the action potential even originates at the first node of Ranvier, where sodium channels are highly concentrated (Figure 1). For both myelinated and unmyelinated axons, once the action potential begins in the axon, it not only propagates orthodromically toward the nerve terminals but also propagates antidromically, back into the soma and dendrites.

Conduction in Unmyelinated and Myelinated Axons

Unmyelinated Axons

Before considering a propagating action potential, it is useful to understand the currents that underlie a stationary action potential. It is possible to control the membrane potential experimentally along the length of an axon. In this case, a short stimulus current can be applied to bring the entire length of membrane to threshold, and the whole axon subsequently and simultaneously undergoes an action potential. Once the applied stimulus is over, total membrane current in this artificial situation is zero. Since membrane current is the sum of capacitive and ionic currents, ionic and capacitive currents are equal and opposite. Ionic currents through sodium and potassium channels simply change the membrane potential by charging membrane capacitance. For a propagating action potential, the relationship between ionic and capacitive currents is more complex (Figure 2). The reason for this complexity is that once the action potential is initiated, either by transduction of a sensory stimulus or by summation of postsynaptic potentials, sodium current in the active region not only depolarizes the active region further, but also provides depolarizing

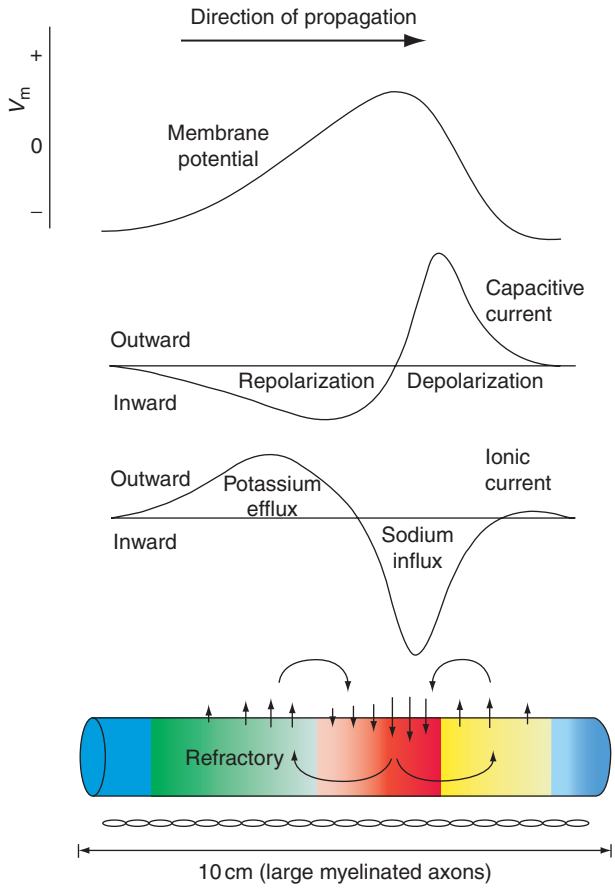

Figure 2 Ionic and capacitive currents that underlie a propagating action potential. The membrane potential change of an action potential propagating from left to right is illustrated at the top. Capacitive and ionic currents are schematically shown below the membrane potential and are drawn to illustrate the major relationships. Since capacitive current is proportional to the first derivative of the membrane potential, peak capacitive currents occur at the maximum slopes of depolarization and repolarization, and capacitive current is zero at the peak of the action potential. Total membrane current (the sum of capacitive and ionic currents) is proportional to the second derivative of the membrane potential and is zero at the maxima of the capacitive current where total current changes between inward and outward (yellow to red boundary and red to green boundary in the schematic of the axon at the bottom). For an action potential propagating at a constant velocity, the scale bar at the bottom can be thought of as time at a fixed point on the axon (typical action potential duration is 1 ms) or as distance over which the action potential is occurring at one instant of time (10 cm for an action potential with a duration of 1 ms and conduction velocity of 100 m s^{-1}). A series of myelinated nodes is shown above the scale bar to illustrate that many nodes participate at any given time. For the fastest conducting axons, there would be five times as many nodes as are illustrated here.

current to the adjacent region of the axon at the leading edge of the action potential to bring it beyond threshold (Figure 2). The depolarization at the leading edge of the action potential is primarily a capacitive current until threshold is reached. Current also spreads longitudinally behind the action potential, but an action potential is not created in the retrograde or backward direction because of the residual changes in the state of potassium and sodium channels. Potassium channels are still activated and are holding the membrane potential near the resting potential while sodium channels in this region are still recovering from the depolarization; they are inactivated, and this part of the axon is temporarily refractory to action potential generation.

The conduction velocity of unmyelinated axons depends on how much current is injected into the axon by the sodium channels, how far the current can spread longitudinally, and how quickly the adjacent membrane can be brought to threshold. The amount of current depends on the density of sodium channels. Since more sodium channels provide more current, one might think that an increase in channel density will always increase conduction velocity. This proportionality is valid only for low to moderate sodium channel densities because the channels act as dipoles (the source of their voltage sensitivity) and add additional capacitance to the membrane. The time required to charge the membrane is the product of the specific membrane resistance and capacitance. At very high channel density, the effect of the added capacitance outweighs the additional current provided because it takes longer to charge the membrane and conduction velocity is decreased. Typical sodium channel density of unmyelinated axons is 50–500 channels μm^{-2}, with potassium channel density about tenfold lower. An especially low density (2–3 channels μm^{-2}) has been reported in garfish olfactory nerve and neonatal rat optic nerve.

Channel subtypes In general, all neurons express multiple subtypes of sodium and potassium channels. All channel subtypes in the Nav1 sodium channel family have the basic features described by Hodgkin and Huxley over 50 years ago; they are activated by depolarization, with subsequent inactivation that is removed when the membrane is repolarized. Ion selectivity seems to be the same for all the subtypes, but the details of voltage dependence, the kinetics of opening and closing, and the modulation of these gating properties vary from one subtype to another. The multiplicity of potassium channel subtypes is much greater than that of sodium channels, and their properties are also more variable.

There is evidence that neurons use different channel subtypes in subcellular regions of the cell. This would allow the cell to fine-tune the excitability of the cell in different regions. The expression and targeting of different sodium and potassium channel

proteins to the unmyelinated axon remain an active area of research. Some sodium channel subtypes in the PNS, such as Nav1.8 and Nav1.9, seem to be predominantly expressed in small dorsal root ganglion neurons and are targeted to unmyelinated axons. Parallel fibers in the cerebellar cortex utilize Nav1.6.

Myelinated Axons

The current available for depolarizing the next axonal segment to threshold is dependent on the loss of current through the membrane (its leakiness) and the decrease due to capacitive current required to charge the membrane and change the membrane potential. The number of wraps of the axon by the myelin and the length of the myelin internode have important electrical consequences for both the ionic and the capacitive currents. Since the extracellular fluid between each wrap is squeezed to a negligible volume and since the cytoplasm is also squeezed out of the glial wraps, the axonal membrane is essentially increased in thickness by the myelin membranes. The number of wraps by myelin varies from a low of about 10 to as many as 150, with each wrap consisting of a pair of membranes. For example, a large myelinated axon with 150 wraps will decrease ionic current loss through the internodal membrane by a factor of 300. Because capacitance is inversely related to the distance between the charged surfaces (in this case the thickness of a membrane), capacitance and capacitive current will also be reduced 300-fold. The current thus moves rapidly in the internode, with little loss through the membrane or in charging the membrane, essentially jumping from one node to the next. This is described as saltatory conduction. Each node acts as a booster station to ensure propagation to the next node, and to accomplish this regeneration of the signal, sodium channels are highly concentrated at each node (2000–3000 channels μm^{-2}) and are about 100-fold lower in density in the internodal membrane.

Increasing the internodal distance increases the speed of conduction because the current is jumping farther. However, there is an optimal internodal length. If internodal distances were to become very large, conduction velocity is predicted to decrease. This decrease in velocity is due to the loss of current in the internodal region, slowing the rate of rise of depolarization at the next node. Internodal distances are found to be about 100-fold greater than the axon diameter (in agreement with internodal distances predicted to optimize the conduction speed) and range from a few hundred micrometers to 1–2 mm.

The mental image of an action potential occupying a single node and hopping from one node to the next is a common misconception. Although the action potential is jumping from one node to the next at the leading edge, many nodes are simultaneously participating. The extent of axons actively involved in the action potential is dependent on the speed of conduction. The fastest conducting myelinated fibers have a speed of 100–120 $m\,s^{-1}$. If the action potential duration is 1 ms, an action potential traveling at 100 $m\,s^{-1}$ will, at a given instant of time, occupy 10 cm of the axon, or approximately 100 nodes, since the internodal length is on the order of 1 mm for large-diameter axons (Figure 2).

A measure of the reliability of conduction is called the safety factor, which is defined as the current in excess of that required to reach threshold and maintain propagation. A safety factor of 2 means that the current generated by the sodium channels is twice the minimum needed for conduction. Axons have a safety factor of about 5, and this excess is important because it speeds conduction (allowing the membrane to reach threshold faster) and provides the extra current needed at branch points. Axons branch hundreds of times, each branch imposing an increased load on the current provided by the upstream axonal membrane. If several branches occur close together, conduction can fail in some branches, especially during high-frequency firing. For similar reasons, additional current is also needed at the synaptic terminal where additional membrane must be depolarized. The internodal distances in motor axons decrease as the synaptic terminal is approached, and in some cases are as short as 10–20 μm. The effect of decreasing internodal distance is to concentrate nodes of Ranvier near the synaptic terminal, to provide the necessary current for terminal depolarization. It is not known if the terminals have sodium channels, since immunolabeling with antibodies specific for voltage-gated sodium channels have failed to show this.

Channel subtypes Many subtypes of sodium channels can be targeted to nodes of Ranvier, and during development several subtypes are found at neonatal nodes. In the adult mammal almost all nodes of Ranvier in the PNS and CNS contain predominantly one subtype, Nav1.6. The switch between neonatal and adult subtypes at the node coincides with the formation of compact myelin. Three types of potassium channels have been identified pharmacologically (inward rectifier, slow outward rectifier, and 4-aminopyridine-sensitive channels) and attributed primarily to internodal membrane. Subtype-specific antibodies have shown that Kv1.1, Kv1.2, and Kv1.4 are present in the internodal region, with the highest density in the juxtaparanodal region at the boundary

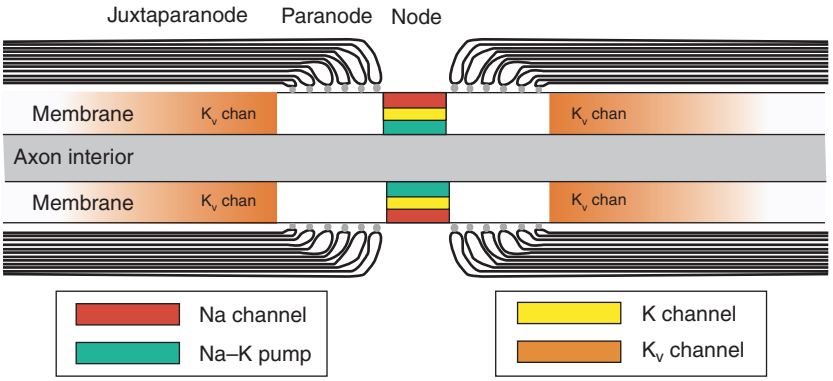

Figure 3 Ion channels and pumps concentrated in the vicinity of the node of Ranvier. Saltatory conduction in myelinated nerves is dependent on a high concentration of sodium channels at the node of Ranvier, to provide the inward current needed for depolarization of the next node. Repolarization is accomplished not only by inactivation of the sodium channels but also by a high resting potassium conductance at the node (KCNQ potassium channels) and by voltage-gated potassium channels excluded from the node and concentrated in the juxtaparanodal region near the node. Sodium–potassium pumps are concentrated at the node to maintain the concentration gradients. The thickness of the membrane (5 nm) is highly exaggerated relative to the axon diameter (>1 μm).

of the paranode. In addition, there is a high resting potassium conductance at some nodes, and a high concentration of KCNQ2 and possibly KCNQ3 potassium channels is co-extensive with the high concentration of sodium channels at the node. These separate highly aggregated clusters of channels are illustrated in Figure 3. As mentioned earlier, the flux of sodium and potassium ions needed to charge and discharge the membrane for each action potential is small, but maintenance of the ion gradients is dependent on sodium–potassium pumps. These pumps are highly concentrated in the nodal membrane.

Summary

The essential features of action potential initiation and propagation were determined over 50 years ago. Research into the electrical excitability of neurons is, however, far from moribund. Recent advances in molecular biology have revealed a multiplicity of sodium and potassium channel subtypes in neurons. Subtle changes in the activation, inactivation, and kinetics of voltage-gated sodium and potassium channels are predicted to have large effects on action potential threshold and rate of firing. The subcellular placement of specific isoforms and the modulation of these isoforms are also critical parameters of neuronal excitability. Many basic questions at the cellular and subcellular level remain. What is the lifetime of these channels in different regions, such as the axon hillock, initial segment, or node of Ranvier? Are there intracellular pools of the channels that can be rapidly inserted to provide plasticity at the level of conduction? What are the signals that target channels to the nodal region and adjacent paranodal and juxtaparanodal regions of myelinated axons? How does the cell achieve a balance between channel synthesis and degradation in the cell body, as well as insertion and retrieval at specific sites such as the node of Ranvier? How is the distribution of channels established during development? Maintenance of electrical excitability during adulthood is a process of continual remodeling that requires constant feedback with signals from target cells and interactions with glial cells. These molecular signals and interactions remain unknown.

See also: Demyelinating Diseases; Demyelination and Demyelinating Antibodies; Ion Channel Localization in Axons; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Schwann Cells and Axon Relationship; Sodium Channels; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Ariyasu RG, Nichol JA, and Ellisman MH (1985) Localization of sodium/potassium adenosine triphosphatase in multiple cell types of the murine nervous system with antibodies raised against the enzyme from kidney. *Journal of Neuroscience* 5: 2581–2596.

Baker M, Bostock H, Grafe P, et al. (1987) Function and distribution of three types of rectifying channel in rat spinal root myelinated axons. *Journal of Physiology* 383: 45–67.

Boiko T, Van Wart A, Caldwell JH, et al. (2003) Functional specialization of the axon initial segment by isoform-specific sodium channel targeting. *Journal of Neuroscience* 23: 2306–2313.

Caldwell JH, Schaller KL, Lasher RS, et al. (2000) Sodium channel Na(v)1.6 is localized at nodes of Ranvier, dendrites, and synapses. *Proceedings of the National Academy of Sciences of the United States of America* 97: 5616–5620.

Catterall WA (1981) Localization of sodium channels in cultured neural cells. *Journal of Neuroscience* 1: 777–783.

Colbert CM and Johnston D (1996) Axonal action-potential initiation and Na^+ channel densities in the soma and axon initial segment of subicular pyramidal neurons. *Journal of Neuroscience* 16: 6676–6686.

Hodgkin AL (1975) The optimum density of sodium channels in an unmyelinated nerve. *Philosophical Transactions of the Royal Society of London* 270: 297–300.

Jack JJB, Noble D, and Tsien RW (1975) *Electric Current Flow in Excitable Cells*. London: Oxford University Press.

Nicholls JG, Martin AR, Wallace BG, et al. (2001) *From Neuron to Brain*, 4th edn. Sunderland, MA: Sinauer Associates.

Palmer LM and Stuart GJ (2006) Site of action potential initiation in layer 5 pyramidal neurons. *Journal of Neuroscience* 26: 1854–1863.

Quick DC, Kennedy WR, and Donaldson L (1979) Dimensions of myelinated nerve fibers near the motor and sensory terminals in cat tenuissimus muscles. *Neuroscience* 4: 1089–1096.

Active Perception

D H Ballard, The University of Texas at Austin, Austin, TX, USA

© 2009 Published by Elsevier Ltd.

Perception Is Active Perception

What makes perception 'active'? Our everyday perception seems to require some engagement of attention on our part. We can be unaware of an ant crawling on the surface of our skin and at another time be conscious of its presence. Furthermore, this consciousness has an aim: we need to intercept the ant's path to get rid of it. Such is the nature of active perception. The awareness of these events has a neural trace in the brain's cortex. In binocular rivalry, separate images are presented to each eye, but the brain 'chooses' one to see on a momentary basis having an interval of a few seconds. Neurons in the brain's cortex respond to the perceived stimulus with the effects being more pronounced in higher cortical areas. In another example, microstimulation of the medial temporal (MT) cortical motion area affects a monkey's report of stimulus direction of motion. Thus, clearly signals related to perception are evident in the brain.

One or both of two additional constraints influence active perception. The first is the addition of a task agenda. The requirements of a specific task almost always allow shortcuts in the computation of the requisite information. In following a car on a highway, the observer only need lock the gaze on the rear of the car. Once this is done, the looming reflex provides a reliable measure of the change in following distance. The second constraint is the addition of the use of the body's mobility. Changing the body's posture or motion in a constrained trajectory can also lead to dramatic computational cost reductions, as any tennis player is acutely aware.

Early Vision

Early theoretical thinking in active perception has been inevitably influenced by our phenomenological experience of seeing a coherent three-dimensional scene. Although we are admonished not to think of the job of visual computation as producing a 'picture in the head,' all work in vision is heavily influenced by the seminal work of David Marr, who proposed that the job of vision was a form of this Cartesian idea. Marr proposed that the brain, particularly the cortex, constructs a layered series of representations that gradually progress to a more physical description of the 'outside' world. Evidence for such a progression is very solid, with the most well-known example being the MT visual cortical area, in which the representation of motion is transformed from a retinotopic representation of photometric change to a retinotopic image of local velocities, the latter termed 'optic flow.'

Active Perception Avoids Marr's 'Ill-Posed Problem'

Inspired by the discoveries of retinotopic maps of visual properties of increasing abstraction, early theoretical work focused on computing such representations. Initially, motor systems were seen as divorced from this effort, but the upshot was that the computational problems were very difficult. Marr characterized them as ill-posed problems because all the methodologies seemed to require iterating over the visual array of information. However, motor systems integration has largely revised that view because the ability to move the head, hand, and eyes provides important constraints that vastly simplify the computation required.

Consider the computation of optic flow, which is an instantaneous measurement of retinal velocity. Each point x, y in the optic flow 'image' indexes a velocity $u(x, y)$, $v(x, y)$. Fundamental conservation considerations lead to the constraint that the image velocity is the source of the spatial and temporal photometric variations. This can be codified as the flow equation that relates the two components of the optic flow velocity to the derivatives of the time-varying image $I(x, y, t)$ as shown:

$$I_x u + I_y v = I_t$$

where I_x, I_y, and I_t represent partial derivatives of the image that can be approximated with standard Gabor filter ('simple cell') measurements. Because these derivatives can be measured, the result at every point is one equation in two unknowns. This indeterminacy caused Marr to refer to this is an ill-posed problem. Without a further constraint this problem has to be solved with costly iterations over the image array.

However, consider the case of a static scene and further note that the motion of the head is constrained by the use of fixation to lock the gaze onto a specific point in the scene. In this case, active perception provides another constraint that greatly simplifies this problem. The actual direction of motion is known, providing another equation that now makes the problem well posed and easily solved for each retinotopic neighborhood. Furthermore, in this

special case, the velocities are now correlated with depth so that a depth cue is obtained as a side effect. Kinetic depth is extremely fundamental because it is the first of several depth cues obtained during infant development.

Dynamic Frames of Reference

The key to understanding how these calculations are simplified is to realize that the primate visual system has a special capability with regard to its ability to fixate. Five independent stabilization systems – the vestibulo-occlular reflex, optokenetic nystagmus, saccades, vergence, and pursuit systems – allow it to lock its fixation point on a specific location in space. For humans, this ability is especially important because of the very expanded spatial resolution of photoreceptors at the fixation point. In one degree of visual area, characterized as the width of a thumb held at arm's length, the spatial resolution is approximately 100 times greater than the periphery. Thus, the visual system has evolved to rely on the point of fixation for high-precision visual calculations.

It sometimes takes more than a moment's reflection to realize that the point of fixation is not in the body. Thus, the fundamental computations of vision, as well as audition, are about a point that is not only distal in space but also dynamically changes from moment to moment. A way to remember this is to recall the coding for depth using binocular disparity. In striate and subsequent cortical areas, depth is coded in a relative system. Neurons have receptive fields that are tuned to positive, negative, and zero or near-zero disparities. Zero disparity means that the cell's best response occurs inputs from the left and right eyes match, and this is of course at the distal point of fixation. It is estimated that humans make more than 150 000 such fixations every day. Laboratory tests show that each such fixation has a specific purpose.

The interaction of increased resolution with active perception was first appreciated in studies of reading. When reading text on a computer terminal, the text outside of approximately 13 characters around the fixation point can be changed almost arbitrarily without consequence to reading speed, as long as the characters under each momentary fixation are correct. However, this result is not specialized for reading but, as subsequent work has shown, is likely to be a general property of the perceptual system.

Modeling Active Perception with Robots

The importance of active perception led to motorized camera systems as a basic input device for models. The very first such device was built by a team led by Ruzena Bajczy at the University of Pennsylvania. The first system with real-time image processing capability was built at the University of Rochester. **Figure 1** shows the combined system for studying hand–eye coordination. In order to provide the degrees of freedom of a head, the binocular camera system is mounted on a Puma industrial robot arm. It coordinates the movements of a robot hand that is mounted on a similar arm.

The development of robot active perception systems was a breakthrough because it allowed the computations in active perception to be studied directly. One of the first findings was that although the general problem of perceiving a scene at some level of detail in its entirety was very computationally expensive, there were subproblems of the general problem that could be done very cheaply. **Figures 2(a)** and **2(c)** show examples of the classical dichotomy of looking for a specific colored object in a scene when the colors are known in advance together with the complementary problem of having a close look at a colored object and trying to identify that object from a library of objects. These kinds of results can be related to the organization of cortex, which typically has separate cortical sensory maps devoted to object identity and object location. The former are located in temporal cortex and the latter are located in dorsal cortex.

Experiments with active perceiving robot systems showed that each of these problems could be done very cheaply. Work has continued along these lines and now incorporates probabilistic methods that are discussed later.

The Quantal Nature of Data Acquisition

One of the most remarkable findings of active perception is its inherently quantal nature. Since 'active' denotes a specific agenda, the perceiver almost always needs very specific information from each fixation. Thus, the larger percept of a scene may be replaced by a question: 'Is the cup I am filling with tea full yet?' This requires monitoring the rising level of liquid in a cup. This computation in turn can be modeled with simple template matching, which is vastly simpler than the general-purpose computation that would be required to create a three-dimensional model of the cup and its immediate venue.

The quantal nature of active perception is very dramatically revealed in sports in which the speed of travel of the ball places high demands on rapid perception. Interestingly, many such sports are handled identically by the perceptual system. Consider the example of cricket, in which the ball bounces on the cricket pitch at approximately 90 mph. When

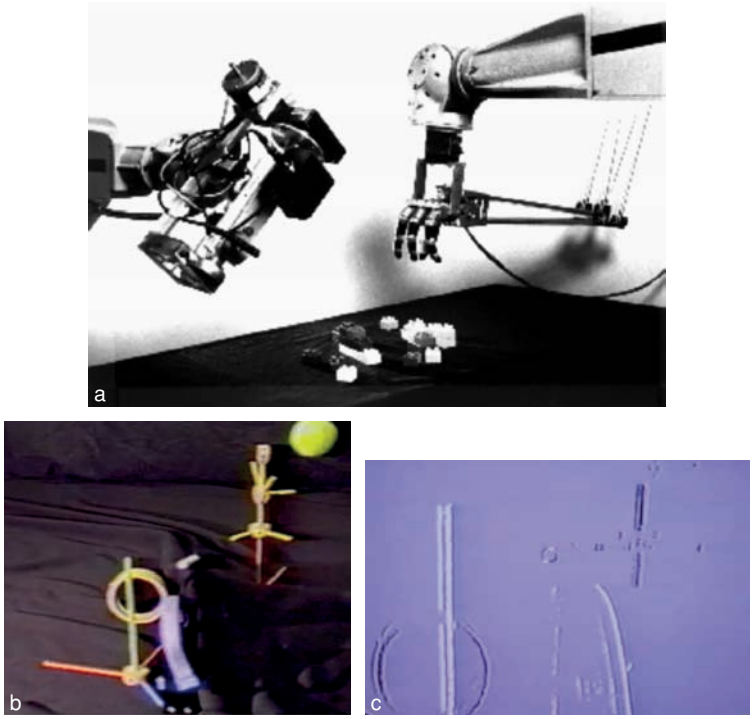

Figure 1 (a) Binocular camera vision system combined with a MIT-Utah robot hand. Motors drive the cameras in pitch and yaw at near saccadic speeds. (b) Table scene containing simple toy objects. (c) Kinetic depth image obtained in real time by moving the head horizontally in a frontoparallel plane. Points in the scene in front of the fixation point are coded as white and those behind in black.

Figure 2 (a) A 'what' task solved by focusing on a single object and matching its colors against a color database. (b) A 'where' task: a floor strewn with colored shirts, one of which needs to be found. (c) The location of its colors can be established by looking for localized clusters of them in the image.

analyzing the pitch, the batsman first fixates the bowler's hand at the release point and then saccades to the estimated bounce point. At the bounce point, the computation required is to estimate the path of the ball, on which the rest of the batsman's motion is determined, based on extensive training from similar pitches. This strategy is mirrored in table tennis, in which players fixate the bounce point on the table,

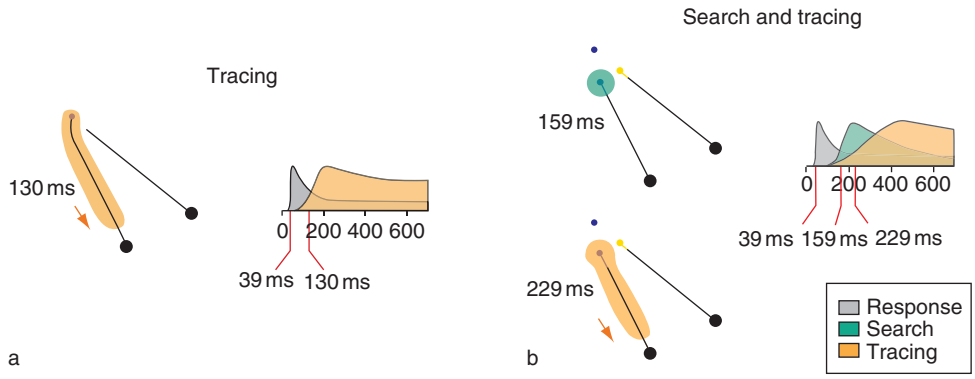

Figure 3 (a) A monkey is trained to fixate a small spot on an otherwise blank CRT screen. Subsequently, two lines appear and the monkey must fixate the end of the line connected to the fixation spot. (b) A modified version of the task in (a), where now the monkey must fixate the line that has a stub of the same color as the fixation spot.

and in squash, in which players fixate the bounce point on the front wall. However, in each of these cases, information gleaned at the fixation point keys an elaborate motor response.

Neural evidence of the quantal nature of active perception comes from experiments such as those by Pieter Roelfsema, in which monkeys fixate a small spot on a blank screen. While they hold fixation, two lines that are oriented radially with respect to the spot appear. One is connected to the fixation spot and the monkey's task is to make a saccadic eye movement to its other terminal. The experiment takes advantage of the fact that in the absence of cues, it takes approximately 250–300 ms to program the saccade. In addition to this protocol, the experimenters monitor cells in striate cortex with receptive fields that are oriented coincident to the line's orientation. The crucial finding is that the firing rate of the cell is increased at a specific point in time, approximately 130 ms after the target onset. The interpretation is that the monkey uses a specific visual routine to trace along the connected line to find its terminal. In a modification of this protocol, a monkey has to make a saccade to the end of the line that has a stub which is the same color as the fixation spot. Under the visual routines hypothesis, the tracing should now start later in time. In fact, the elevated onset in firing now starts at 229 ms (Figure 3). Experiments such as these provide corroborating evidence for active perception and suggest that it may consist of elemental operations that can be combined to solve more complicated tasks.

Bayes' Rule

The need for visual routines is compelling, but at the same time they introduce a modeling complexity. One is led to the same conclusions as Shimon Ullman and Steve Kosslyn, two of the original visual routine proponents: there must be some library of standard operations that can be composed into short programs that solve specific tasks. How such a library gets established through learning and development is very much an open question, but a very helpful way of characterizing the computations of a routine abstractly has emerged via Bayesian probability theory. The reader will be very familiar with the following version of Bayes' rule:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

Within a constant of proportionality we can write this as

$$P(M|D) \propto P(D|M)P(M)$$

Let us interpret this using the example of anticipating the cricket ball bounce location on the cricket pitch. The batsman needs to know exactly where the ball will bounce, and to do this he or she has to integrate two sources of information. In this case, $P(M)$ models the probability distribution of likely locations of the ball based on numerous prior pitches. The term $P(D|M)$ is data on the location of the ball estimated by the batter for a particular pitch. Multiplication of these two distributions provides the Bayesian estimate for the likely bounce point of the ball. This kind of Bayesian formulation has been repeated in numerous experiments, suggesting that it is a general mechanism that the brain uses to combine perceptual estimates with previous experience.

Virtual Environments

Virtual environments have proven to be a breakthrough for the study of active perception because

they allow specific features to be present or absent during an active perception task. Typically, subjects can wear a binocular head-mounted color display to see scenes that are rich in visual cues. Magnetic and inertial-based sensors on the head-mounted display (HMD) allow for the scene to be updated at frame rates, thus providing a sense of immersion in a three-dimensional world. Because the subject can execute different tasks based on the visual display contents, the motor system provides a new and reliable arbiter of what was perceived.

For example, consider the block-copying task used by Triesch et al. Subjects perform the task of picking up blocks in a virtual environment. Subjects have a sense of touch delivered through a virtual force feedback device. The device shown in Figure 4 provides a realistic sensation of force through opposed kinematic links that can be back-driven with cable-linked electric motors based on the position of two fingers, typically the finger and thumb used in a pinch grip.

Experimental subjects are instructed through visual cues in the virtual reality environment to pick up one of several blocks based on a feature such as color and place it on a conveyor belt based on a second, possibly the same, feature. The key manipulation is that on a small percentage of the trials the features of the block change between the pickup and put down. Subjects are told to discard the block if they notice any such changes.

The key observation is that on missed color changes, where the subjects behave as if nothing happened, the block is sorted by its original color rather than its new color. This decision is made even when the subjects have fixated the block after the color change. The implication is that, consistent with neural data, the acquisition of the color is done in a quantal manner. When the color is registered (represented) early on in the experiment, later there is no need to check it in the neural protocol.

Tasks with Sequential Steps

Simple tasks such as staying on a sidewalk require only the measurement of the sidewalk edge. The history of the traverse is not needed. However, more complicated behaviors, such as making a sandwich, require more state. If you want to put peanut butter on a slice of bread, you must be holding the knife. These causal dependencies can be represented in a graphical network. Such a network is a suitable tool for this class of problems because it uses easily observable evidence to update or infer the probabilistic distribution of the underlying random variables. A Bayesian graphical network represents the causalities with a directed acyclic graph; its nodes denote variables and edges denote causal relations that can be annotated with conditional probability distributions. In this active perception case, the state of the agent is dynamically changing, and the observations are updated throughout the task execution process, so one must specify the temporal evolution of the network. Figure 5 illustrates the two-slice representation of a dynamic Bayesian network. Shaded nodes are observed; the others are hidden. Probability distribution matrices determine causalities, represented by straight arrows. The state of the lowest hidden node is determined by its prior distribution in the first time/slice, and thereafter it is jointly determined by its previous state and the transition matrix, as denoted by the curved arrow. The two-slice representation can be easily unrolled to address behaviors with arbitrary numbers of slices. At each moment, the observed sensory data (gray nodes), along with its history, are used to compute the probability of the hidden nodes being in certain states:

$$P(S^{t\in[1,T]}) = P(S^1)P(O^1|S^1)\prod_{t=2}^{T} P(S^t|S^{t-1})P(O^t|S^t)$$

Although this equation might look complex, it is simply an expression for the probability of being in the

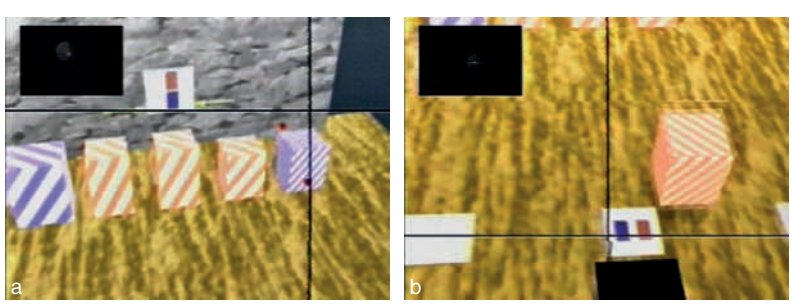

Figure 4 (a) The view from a subject's helmet in a block manipulation task. A subject is about to pick up a blue block using virtual force fingertips indicated by the red dots on the screen. (b) After moving it closer, a cue on the table instructs that a blue block goes on the conveyor belt to the left, and a red block goes on the conveyor belt to the right. If the block changes color, as it has in this case, it is to be discarded, but instead the subject follows the instructions for a blue block.

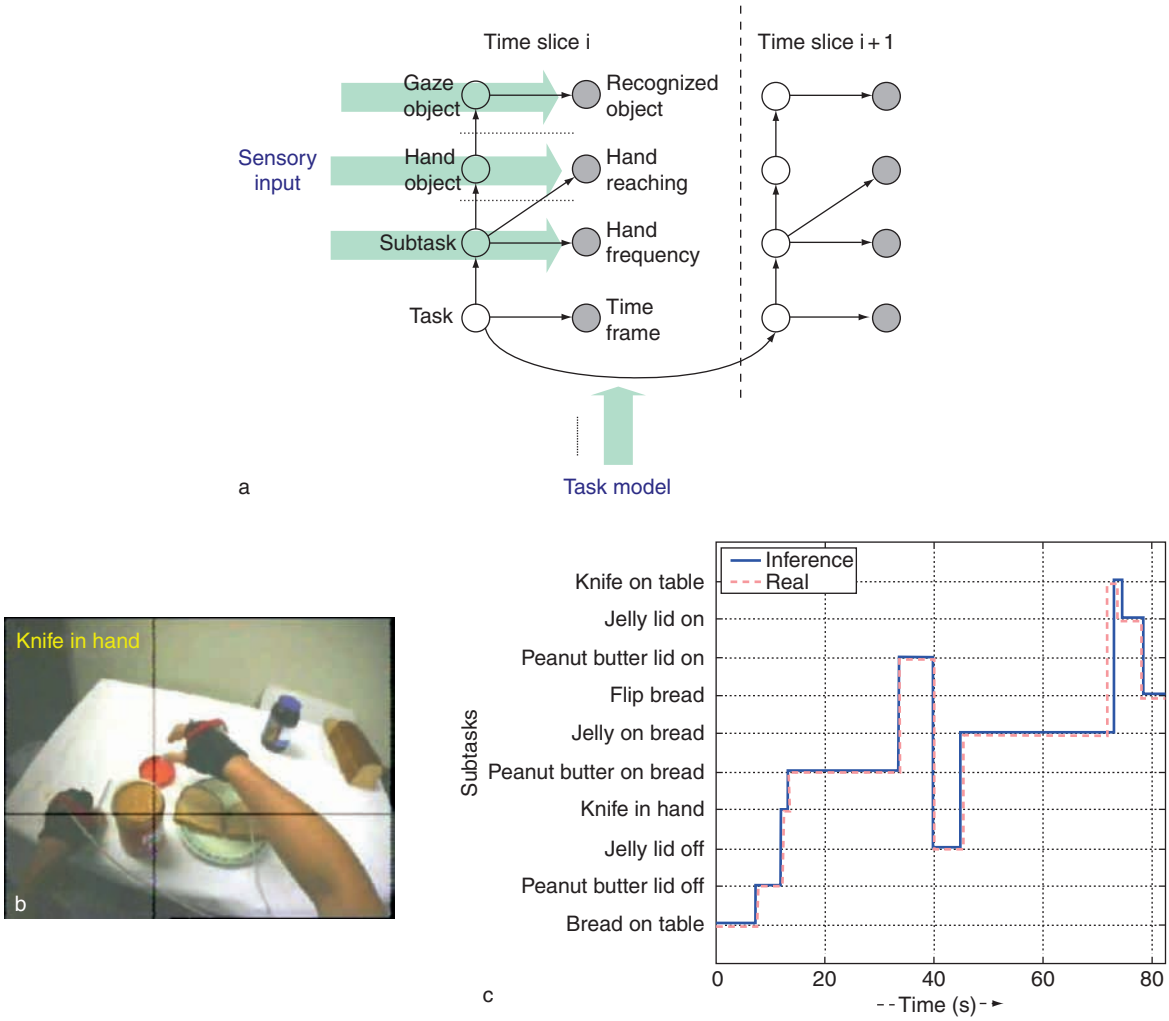

Figure 5 (a) The basic structure of the dynamic Bayesian network (DBN) used to model sandwich making. Visual and hand measurements provide input to the shaded nodes. The sequencing probabilities between subtasks are provided from a task model that in turn is based on human subject data. (b) A frame in the video of a human subject in the process of making a sandwich showing that the DBN has correctly identified the subtask as 'knife-in-hand.' (c) A trace of the entire sandwich-making process showing perfect subtask recognition by the DBN. This trace uses the entire data set in an off-line mode. In the online mode in which the classifications have to be made on the partial data sets, small errors can be made.

states $\{S^t, t = 1, \ldots, T\}$ and seeing observations $\{O^t, t = 1, \ldots, T\}$ from those states.

A behavior-recognition algorithm computes the states of each hidden node S^tat time t that maximize the probability of observing the given sensory data:

$$S^t = \arg\max_s P(S^t = S|O^{[1,T]})$$

Conclusion

This introduction to active perception has traced the development of its key ideas. The notion that physical descriptions of the world are represented in visual maps in the cortex heralded the beginning of understanding how visual information is represented, but early static models were very computationally inefficient. Active perception research, driven by advances in robotics technology, provided a more behavioral framework that led to dramatic computational simplifications. Two other contributions to active perception are the Bayesian revolution and virtual reality technology. The former has demonstrated that the perception of almost all visual information is governed by prior information stored as probability distributions. Complicated sequential tasks can use graphical structures to organize complex probabilistic dependencies. Virtual reality technology provides a way of probing these dependencies because the virtual sensory world can be manipulated on a timescale that is sufficiently fast to interact with the brain's management of such information.

See also: Bayesian Cortical Models; Bayesian Models of Motor Control; Binocular Rivalry; Contextual Interactions in Visual Perception; Ideal Observer Theory; Spatial Transformations for Eye–Hand Coordination; Statistical Analysis of Visual Perception.

Further Reading

Aloimonos J, Weiss I, and Bandyopadhyay A (1988) Active vision. *International Journal of Computer Vision* 1(4): 333–356.

Ballard DH (1991) Animate vision. *Artificial Intelligence* 48(1): 57–86.

Bishop C (2006) *Pattern Recognition and Machine Learning*. New York: Springer.

Britten KH, Shadlen MN, Newsome WT, and Movshon JA (1992) The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience* 12: 4745–4765.

Droll JA, Hayhoe MM, Triesch J, and Sullivan B (2003) Task relevance of object features modulates the content of visual working memory. *Vision Research* 3(9).

Gray WD (ed.) (2007) *Integrated Models of Cognitive Systems*. Oxford: Oxford University Press.

Jordan MI (ed.) (1998) *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Land M and Hayhoe M (2003) In what ways do eye movements contribute to everyday activities? *Vision Research* 41: 3559–3566.

Landy MS, Maloney LT, and Pavel M (eds.) (1995) *Exploratory Vision: The Active Eye*. New York: Springer.

Marr D (1982) *Vision*. New York: Freeman.

Pearl J (2000) *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge University Press.

Roelfsema PR, Khayat PS, and Spekreijse H (2003) Subtask sequencing in the primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 100: 5467–5472.

Triesch J, Ballard D, Hayhoe M, and Sullivan B (2003) What you see is what you need. *Journal of Vision* 3: 86–94.

Weiss Y, Simoncelli EP, and Adelson EH (2002) Motion illusions as optimal percepts. *Nature Neuroscience* 5: 598–604.

Active Zone

P S Kaeser, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Synapses are highly specialized contacts between nerve cells that transmit signals from the presynaptic neuron to the postsynaptic cell. They are composed of the presynaptic terminal, the synaptic cleft, and the postsynaptic membrane. The terminal of the presynaptic neuron translates the arriving action potential into a chemical signal, which in turn stimulates a response on the postsynaptic cell. In brief, the action potential induces opening of presynaptic voltage-gated calcium channels, and the local increase in calcium in the presynaptic terminal promotes fusion of neurotransmitter filled synaptic vesicles at the presynaptic membrane on a millisecond timescale. Fusing vesicles release their content into the synaptic cleft, and the neurotransmitters induce a response in the target cell upon binding to postsynaptic receptors.

This brief overview of synaptic transmission demonstrates the asymmetric nature of synapses, where both synaptic compartments – namely the presynaptic terminal and the postsynaptic membrane – contain highly specialized components. In principle, the presynaptic terminal is an organ of membrane trafficking and the postsynaptic compartment is a signal transduction machinery. The apparatus of postsynaptic reception is the postsynaptic membrane with the receptors and the postsynaptic density (PSD), which consists of signaling proteins, scaffolding proteins, and cytoskeletal proteins that enable the postsynaptic cell to sense the presynaptic release of neurotransmitters and respond appropriately. The fusion of synaptic vesicles at the presynaptic membrane is under tight spatial and temporal control. Synaptic vesicle fusion only occurs at highly specialized hot spots called active zones (AZs).

Definition of Active Zones and Functional Participation in the Synaptic Vesicle Cycle

Originally, AZs were described morphologically in seminal studies by Couteaux and Pecot-Dechavassine and Akert et al., and they appear as electron dense material that is tightly attached to the presynaptic membrane (Figure 1). Functionally, AZs are defined as sites where neurotransmitters are released into the synaptic cleft, and biochemically, AZs are composed of a network of insoluble proteins, containing AZ-specific and other proteins.

In the presynaptic nerve terminal, synaptic vesicles undergo a series of events in order to release their content into the synaptic cleft before being recycled and refilled. Investigators divide the synaptic vesicle cycle into nine steps. Synaptic vesicles are filled with neurotransmitters (step 1), and the vesicles then cluster around the AZ (step 2). They are docked to the AZ (step 3), where they gothrough an ATP-dependent priming reaction (step 4) that enables them to undergo fast, calcium-dependent fusion pore opening (step 5) which is initiated by the arriving action potential in the presynaptic nerve terminal. Upon releasing neurotransmitters into the synaptic cleft, three pathways have been proposed: Vesicles might remain in the docked stage (step 6, also referred to as 'kiss-and-stay'), or they might be undocked and locally refilled (step 7, called 'kiss-and-run'). Alternatively, they may be recycled through clathrin-dependent endocytosis of synaptic vesicles (step 8), which leads to either direct refilling or refilling after being passed through the endosome (step 9). It is important to note that only approximately 10–20% of all action potentials that arrive in a nerve terminal induce exocytosis of synaptic vesicles.

AZs participate in multiple steps within the synaptic vesicle cycle. Their core function is to bring synaptic vesicles in close proximity to the presynaptic plasma membrane and to presynaptic calcium channels in order to enable the synapse for fast, calcium-dependent fusion. In addition, AZs participate in a regulatory manner by modifying the probability of neurotransmitter release. This involvement in synaptic plasticity is mediated through AZ-specific proteins, and in principle it can be achieved by changing the number of either docked or primed vesicles. The molecular events that occur within the AZ in order to dock and prime synaptic vesicles and to change synaptic strength are under intense investigation, and aspects of these events are discussed after a more in-depth examination of the structure of AZs.

Morphology of the Active Zone

The fine structure of AZs is diverse among different species, and it is also variable among different kinds of synapses. Nevertheless, all AZs share common features. In electron microscopic images, they appear as a dense structure that is tightly attached to the presynaptic membrane (Figure 1). The electron dense nature of AZs reflects the very high protein content, and they consist of a network of highly

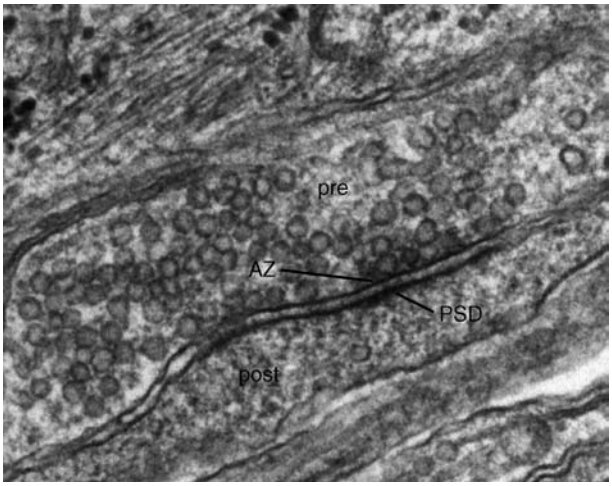

Figure 1 Electron micrograph of a central nervous synapse. Synaptic vesicles tether in the presynaptic terminal (pre) around the active zone (AZ), which consists of electron dense material that is tightly associated with the presynaptic membrane. The postsynaptic density (PSD) is precisely opposed to the AZ at the postsynaptic cell (post). Courtesy of Dr. Xinran Liu, University of Texas Southwestern Medical Center, Dallas, TX.

insoluble proteins that interact with each other and with multiple cellular proteins. AZs are precisely opposed to PSDs, and they are separated from PSDs by the presynaptic plasma membrane, the synaptic cleft, and the neurotransmitter receptor containing postsynaptic plasma membrane. This architecture of precise opposition of the area of release and sensing of neurotransmitters is crucial because it minimizes the distance that neurotransmitters have to diffuse before they bind to the postsynaptic receptors. In mammalian synapses, the variability of this geometric assembly of the AZ and the PSD correlates with the synapse type. Whereas excitatory, glutamatergic synapses are typically asymmetric with a PSD that is thicker than the AZ, inhibitory GABAergic and glycinergic synapses appear more symmetric.

In typical central nervous system mammalian synapses, electron microscopy has revealed that AZs consist of a hexagonal grid, and synaptic vesicles are embedded in depressions of the grid close to the presynaptic membrane (Figures 2(a) and 2(b)). The electron

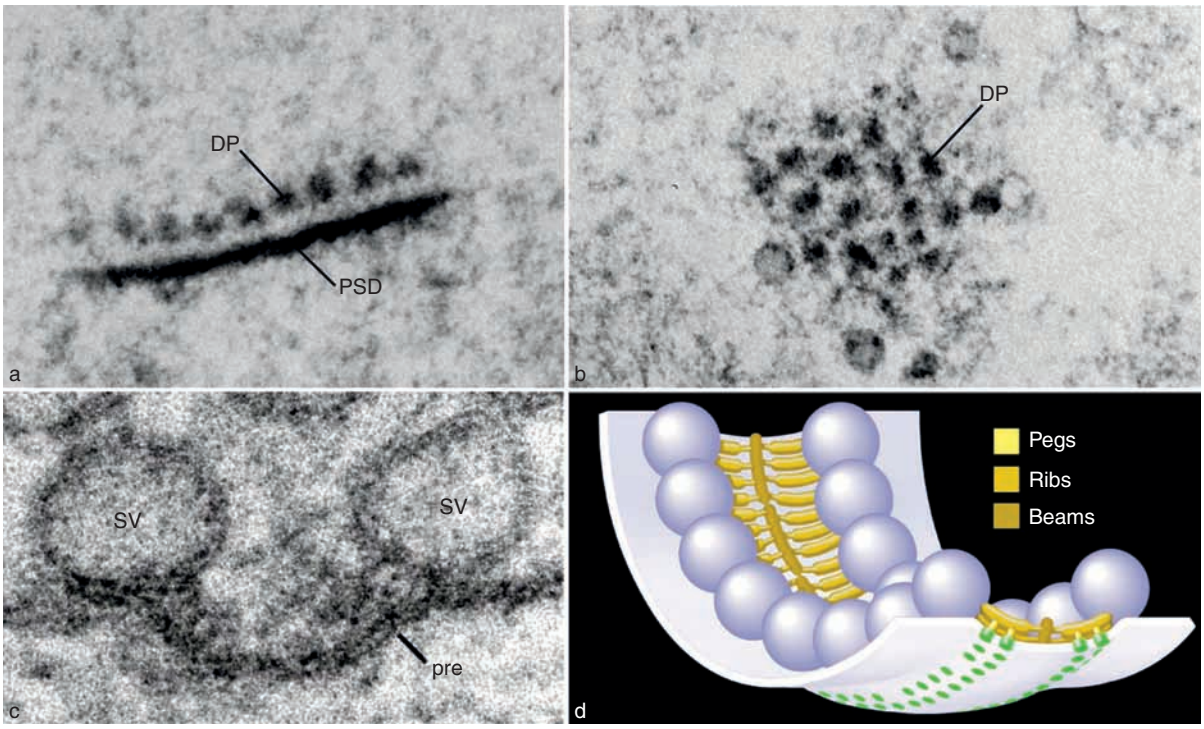

Figure 2 Fine structure of active zones (AZs) of a mouse central nervous system synapse and the frog neuromuscular junction. (a and b) Phosphotungstic acid (PTA) staining of a central nervous synapse in cross section (a) and as top view from the presynaptic terminal (b). PTA selectively visualizes the proteinacious contents of AZs and postsynaptic densities (PSDs); the synaptic plasma membrane is not stained. Whereas the PSD appears as a regular, dense band, the AZ forms dense projections (DPs) that protrude approximately 50 nm into the presynaptic terminal. The DPs are assembled in a regular, hexagonal grid and are connected with thin fibrils. This regular assortment allows the synaptic vesicles to closely approach the presynaptic plasma membrane in the depressions of the grid. Courtesy of Dr. Xinran Liu, University of Texas Southwestern Medical Center, Dallas, TX. (c) Cross section and a model (d) of a frog neuromuscular junction based on conventional two-dimensional electron microscopy. The AZ material is located between the synaptic vesicles (SVs) forming a regular structure with a central beam, ribs that connect the central beam with the SVs, and pegs that connect the ribs with the synaptic plasma membrane containing the calcium channels. pre, presynaptic membrane. Reprinted by permission from Macmillan Publishers Ltd: *Nature* (Harlow ML, Ress D, Stoschek A, Marshall RM, McMahan UJ (2001) The architecture of active zone material at the frog's neuromuscular junction. *Nature* 409: 479–484), copyright 2001.

dense protrusions, called dense projections, are relatively small compared to other synapses; they extend for approximately 50 nm from the presynaptic plasma membrane into the presynaptic terminal, and they are connected to each other by thin fibrils. Each dense projection has six immediate neighbors, and they are evenly spaced with a distance of approximately 50–100 nm. This highly organized architecture leaves slots for docking of synaptic vesicles in close proximity to the membrane.

Harlow et al. analyzed the fine structure of the AZ at frog neuromuscular junction (NMJ) in detail (**Figures 2(c) and 2(d)**). At these synapses, the vesicles line up at both sides along an elongated ridge at the presynaptic terminals of motor neurons. Each synaptic vesicle is connected via ribs to a central beam, and each rib contains two connections (pegs) to the presynaptic plasma membrane. It is assumed that these pegs consist of macromolecules that connect the synaptic vesicle anchoring ribs with presynaptic calcium channels. Although this study leaves the molecular identities of ribs, pegs, and beams open, it clearly shows the highly specialized organization of presynaptic AZs.

The size and morphology of dense projections vary greatly between different synapses and different species. In contrast to the rib- and beamlike structure at the frog NMJ, the *Caenorhabditis elegans* NMJ forms plaque-like AZs, and *Drosophila* NMJs contain T-shaped projections called 'T-bars.' At so-called ribbon synapses in the rat retina, long, prominent ribbonlike protrusions extend into the cytoplasm and tether synaptic vesicles around them. Many molecular and morphological studies have been performed on ribbon synapses because they are relatively easy to access and visualize. Another sensory synapse with a very specialized architecture is the hair cell in the auditory pathway, where ovoid or spherical dense projections are surrounded by a halo of synaptic vesicles. Both the photoreceptor synapse and the ribbon synapse are functionally very different because they do not respond to incoming action potentials but, rather, to sensory stimuli that produce graded receptor potentials. Their distinct function presumably led to a highly specialized morphology.

Some investigators subdivide AZs morphologically into three compartments: the presynaptic plasma membrane, the cytomatrix immediately attached to the membrane (also referred to as 'CAZ' – cytomatrix of the active zone), and the electron dense protrusions into the cytoplasm, called 'dense projections' (also referred to as membrane thickenings or AZM – active zone material). Although this subdivision is morphologically correct, AZs are functional units, and many proteins involved in their architecture do not respect these morphological borders.

Molecular Components of the Active Zone

AZs consist of a dense, insoluble protein network that is tightly attached to the presynaptic plasma membrane. Although several protein/gene families have been described to be involved in the formation, maintenance, and function of AZs, the molecular characterization has only just begun. Many interactions of proteins involved in AZs have been illustrated, but most of them are poorly understood in terms of their functional importance. In addition, the fact that most AZ proteins are large and highly insoluble makes biochemical characterization difficult, and the results of binding studies that have identified new interactions *in vitro* are difficult to confirm *in vivo*. Despite these difficulties, investigators have made significant progress in identifying molecules that participate in AZs. Gene knockdown/out studies in *Drosophila*, *C. elegans*, and mice have provided fascinating insights into potential AZ functions.

In principle, AZs consist of two classes of molecules: non-AZ-specific and AZ-specific (or enriched) components. Among the non-AZ-specific members are five functional classes of proteins that are involved in different processes. First, there are proteins that are directly involved in synaptic vesicle fusion and associated with the presynaptic plasma membrane, such as the SNAREs SNAP25 and syntaxin, and also Munc18, a protein associated with the presynaptic release machinery that is required for synaptic transmission. Second, although the link between the AZ and the cytoskeleton is only poorly understood, it is clear that cytoskeletal proteins such as actin, spectrin, myosin, and tubulin participate in the structural backbone of the AZ. Third, synaptic scaffolding molecules are tightly associated with both AZ and PSD, and presynaptically CASK, Mints, Velis, and SAP97s have been identified. Fourth, as mentioned previously, the core function of the AZ is to bring together synaptic vesicles and calcium channels at the presynaptic terminal. Thus, voltage-gated calcium channels are an integral part of the AZ, but the molecular link between the intracellular portions of the channel subunits and the dense protein mesh of the AZ is only poorly understood. Finally, cell adhesion molecules play a crucial role during synaptic development and in mature synapses. Several molecules have been shown to reach from the AZ through the presynaptic plasma membrane into the synaptic cleft to form either homophilic or heterophilic interactions with proteins in the synaptic cleft or with proteins that reach through the postsynaptic membrane into the PSD. Such proteins are integrins, cadherins, neurexins, and SynCAMs. Several fascinating studies have shown how these transmembrane molecules induce synapse

formation, and neurexins with their postsynaptic interaction partner neuroligin have been suggested to be crucially involved in the pathogenesis of neurodevelopmental defects such as autism.

Six families of AZ-specific proteins have been identified (Figure 3): Munc13s, Rab3A interacting molecules (RIMs), RIM-binding proteins (RBPs or RIM-BPs), α-Liprins, ELKS (proteins rich in E, L, K, and S), and Piccolo/Bassoon. They all consist of large multidomain proteins that are encoded by multiple genes and that are involved in protein–protein interactions. Figure 3 demonstrates that these proteins interact with each other in multiple ways to form the dense material that can be observed in electron microscopic images. Table 1 provides an overview of the members of these proteins, their evolutionary conservation, and their involvement in AZ function. Next, the functional and biochemical involvement of the AZ-specific proteins is discussed.

Munc13s were initially identified in *C. elegans* as phorbol ester (β-PE)/diacylglycerol (DAG)-binding proteins encoded by the unc13 gene, and the phenotype in unc13 mutant worms suggested that these proteins were involved in neurotransmitter release. Since then, data obtained from *C. elegans*, *Drosophila*, and mouse models have suggested that synaptic transmission is crucially dependent on Munc13-mediated vesicle priming. In mammals, the four Munc13 genes (Munc13-1 to -4) produce three brain-specific Munc13 isoforms (Munc13-1, bMunc13-2, and Munc13-3), the ubiquitous Munc13-2, and the nonneuronal Munc13-4 (Table 1). Munc13s consist of a conserved C-terminal region that contains a C1 domain and two C2 domains flanking the Mun homology domain. The Mun domain is sufficient to rescue the defective synaptic transmission in synapses lacking Munc13. In addition, Munc13-1 and ubMunc13-2 have an N-terminal domain that binds to the RIM zinc finger and calmodulin and that contains a C2 domain (Figure 3). It is interesting to note that different synapses have been shown to express different Munc13 isoforms, and the expression pattern of these isoforms determines parameters of short-term synaptic plasticity. The central C1 domain of Munc13 binds β-PE/DAG, and this interaction mediates β-PE/DAG-induced augmentation in the hippocampus. The conserved C-terminal region of Munc13s has been suggested to bind to the synaptic plasma membrane SNARE syntaxin, two proteins that have been implicated in vesicular trafficking/exocytosis (DOC2α and msec7-1 ARF-GEF), and a novel spectrin (spectrin β-spIIIΣ) that potentially links the AZ to the cytoskeleton. Functional experiments in *C. elegans* and *in vitro* binding studies have proposed a molecular mechanism by which Munc13s promote the open confirmation of syntaxin, which in turn results in a loose assembly of the SNARE complex, bringing synaptic vesicles into a fusion competent state. However, this model, which involves a direct interaction between Munc13 and syntaxin, remains controversial because nuclear magnetic resonance measurements

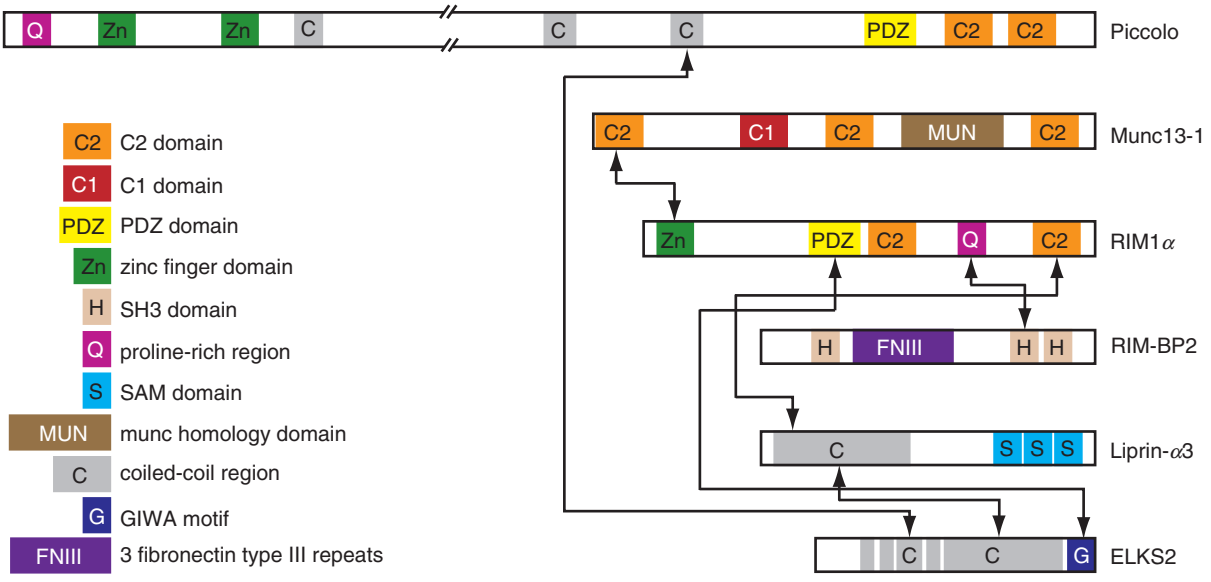

Figure 3 Schematic overview showing a representative member of each family of active zone (AZ) proteins. Interactions among AZ proteins are indicated with black arrows. All AZ-specific proteins are large proteins that contain multiple domains. Together, they form an insoluble protein scaffold that brings together synaptic vesicles, the presynaptic plasma membrane, and voltage-gated calcium channels at the presynaptic nerve terminal.

Table 1 Active zone proteins

Protein family	*Mammalian protein isoforms*[a]	*Drosophila/C. elegans homologs*	*Proposed functions*
Munc13s	Munc13-1 ubMunc13-2, bMunc13-2 Munc13-3 Munc13-4	Dunc13/unc-13	SV priming, scaffolding, synaptic plasticity, cytosceleton anchoring
RIMs	RIM1α RIM2α, RIM2β, RIM2γ RIM3γ RIM4γ	RIM-PA/unc-10	SV docking and priming, short- and long-term plasticity, scaffolding, possibly channel anchoring
ELKS	ELKS1A, ELKS1B ELKS2	*bruchpilot* (*brp*)?/ELKS	Scaffolding, NF-κB signaling
Piccolo/ Bassoon	Piccolo/Aczonin Bassoon	Not identified to date	Scaffolding, anchoring, membrane trafficking, endocytosis
α-Liprins	Liprin-α1 Liprin-α2 Liprin-α3 Liprin-α4	Dliprin/syd-2	Scaffolding, receptor anchoring, membrane trafficking
RIM-BPs	RIM-BP1 RIM-BP2	Not identified to date	Scaffolding, channel anchoring

[a]Each gene is represented on one line. For example, Munc13s are encoded by four mammalian genes, and the Munc13-2 gene produces two protein isoforms (ubMunc13-2 and bMunc13-2).

have not confirmed a direct interaction between the two proteins.

The N-terminal RIM zinc finger forms a tripartite complex with Munc13 and Rab3A, a small GTPase that is located on synaptic vesicles. This biochemical interaction is important because it links Munc13s and RIMs directly to synaptic vesicles. A detailed morphological analysis of *C. elegans* has shown that RIM is necessary for connecting synaptic vesicles to dense projections but not for bringing them in close proximity to the synaptic plasma membrane. In addition, the RIM C2B domain has been suggested to bind to the calcium sensor synaptotagmin 1 *in vitro*, providing another potential link between AZs and synaptic vesicles. RIMs contain a number of domains and interaction motifs that are involved in scaffolding at AZs by binding to multiple other AZ proteins (Table 1). With their central PDZ domain, RIMs interact with ELKS, and with the C-terminal C2B domain they bind to α-Liprins (Figure 3). RIMs are a conserved gene family, with RIM-PA in *Drosophila*, unc10 in *C. elegans*, and four genes (RIM1–4) in mammals. The RIM2 gene has additional internal promoters to produce shorter isoforms (RIM2β and RIM2γ), and multiple sites of alternative splicing enhance the variety of RIM1 and RIM2. The absence of RIM1α in a mouse model leads to a large deficit in multiple parameters of synaptic plasticity, including a reduction of release probability at excitatory synapses and a lack of presynaptic long-term potentiation. This functional deficit is also reflected in the fact that these animals show severe shortcomings in various tests for learning and memory. Electrophysiological and behavioral analysis of the RIM1α mutant mice demonstrates how AZs are critically involved in modulating synaptic transmission, being an essential component for the brain's capacity to adapt to environmental inputs. When RIM1α and RIM2α, two of the six major RIM isoforms, are absent, the mice show a drastic calcium-dependent release deficit at the NMJ, and they die immediately after birth. Thus, RIMs are critical not only for modulating the plasticity of synapses but also for normal synaptic release. RIM proteins contain a proline-rich region in the linker between the central and the C-terminal C2 domains that binds to RIM-BPs.

RBPs are produced by two genes – RIM-BP1 (also referred to as PRAX-1) and RIM-BP2 (Table 1) – and contain three Sarc homology 3 (SH3) domains and three fibronectin type III repeats (Figure 3). No homologs have been identified in either *C. elegans* or *Drosophila*. An *in vitro* study showed that RIM-BPs might bind to voltage-gated calcium channels with their SH3 domains, and thus they can potentially link RIMs and, indirectly, all other AZ-specific proteins to presynaptic calcium channels. In addition,

one study suggested a direct interaction between RIMs and calcium channels. Although these data are very interesting, they have not been confirmed *in vivo*, and functional data have only been provided for the RIM-BP–calcium channel interaction using a neurosecretory PC12 cell line. The *in vivo* existence of the potential calcium channel–RIM-RIM-BP complex remains to be elucidated.

Liprins have been identified as a family of proteins that interact with LAR transmembrane tyrosine phosphatases, and they consist of two subfamilies – Liprin-α and Liprin-β. Liprin-β proteins are not brain specific, and Liprin-αs are encoded by four genes (1–4), of which Liprin-α2 and Liprin-α3 are brain specific (Table 1). Mammalian Liprin-α function has only been addressed *in vitro*, but it has been found that Liprin-α3 forms a complex at the AZ with ELKS and RIM through its N-terminal coiled-coil regions (Figure 3). In *C. elegans*, it has been convincingly shown that the absence of the Liprin homolog syd-2 disrupts the regular AZ structure, and that syd-2-deficient synapses have a defect in neurotransmitter release but a normal amount of synaptic vesicles. It has been concluded that syd-2 is involved in the structural assembly of AZs, and that it probably functions as an intracellular anchor of LAR transmembrane tyrosine phosphatase signaling in synaptic junctions. The structural involvement of Liprins was later confirmed by analysis of Dliprin-deficient flies, which showed a deficit in synaptic morphogenesis. Importantly, Liprins were the first AZ proteins to provide a molecular link between the presynaptic plasma membrane and the electron dense material at the AZ through their interaction with LAR transmembrane tyrosine phosphatases. In addition, Liprin-αs have been shown to bind to GIT (which in turn binds to Piccolo) and the motor protein KIF1a via coiled-coil domains, to CASK and Liprin-β via the SAM domain, and to the scaffolding protein GRIP with its C-terminal PDZ binding motif, supporting the function as a scaffolding protein at AZs.

ELKS proteins were originally described as proteins that bind to the IKK complex and take part in NF-κB signaling. Later, they were identified in mammals as Rab6-interacting proteins (Rab6IP2) that are involved in Golgi trafficking, and they were also described as AZ proteins called CAST. They were abbreviated as ERCs (ELKS/Rab6IP2s/CAST). A genetic analysis revealed that mammals have two genes, ELKS1 and ELKS2. The ELKS1 gene produces two main isoforms – a ubiquitously expressed ELKS1A and the AZ protein ELKS1B, which is alternatively spliced at the C-terminus and binds to the RIM PDZ domain. ELKS2 also has the PDZ binding motif at its C-terminus (Figure 3) and binds to RIM. ELKS is conserved in *C. elegans*, and Liprin-α's function in synapse assembly requires ELKS in the nematode NMJ. In *Drosophila*, there is only a distant homolog to ELKS, which has a partially conserved N-terminus but a nonhomologous C-terminus that resembles cytoskeletal proteins and lacks the RIM-interacting C-terminal motif. Investigators called this protein *bruchpilot* (*brp*; 'crash pilot') due to the unstable flight that these flies demonstrate when *brp* is absent. *brp* null mutant analysis suggested that T-bars, the dense projections at the neuromuscular synapse of *Drosophila*, are absent, calcium channel density is reduced, and synaptic release is deficient. Whether ELKS is equally important at mammalian synapses has to be investigated.

Piccolo (also called Aczonin) and Bassoon are the largest AZ proteins (530 and 420 kDa, respectively), and they are exceptional because they are not conserved in *C. elegans* or flies (Table 1). They contain multiple domains, and 10 regions of homology between Piccolo and Bassoon have been summarized as Piccolo–Bassoon homology domains. The large multidomain structure that includes many protein interaction domains make it likely that Piccolo and Bassoon act as AZ scaffolding proteins. They have the potential to act as calcium sensors with their unusual C2 domains, and they are potentially targeted to the presynaptic plasma membrane via N-terminal myristoylation. Interestingly, in a mouse mutant model that expresses a truncated Bassoon which lacks the central region anchoring Bassoon to the AZ, the dense projections at ribbon synapses are freely floating in presynaptic terminals and synaptic transmission is impaired in these synapses. The delivery of these large proteins to AZs depends on an intact Golgi apparatus, in line with the hypothesis that presynaptic AZs might be preassembled as a complex in the Golgi apparatus and then transported in so-called Piccolo–Bassoon transport vesicles to the presynaptic terminal during synaptogenesis.

Conclusion

Presynaptic AZs consist of a highly specialized network of proteins that acts in neurotransmitter release. Although there is wide variability in the morphology of AZs in different synapses and among various species, their function and molecular architecture are largely conserved. We have only begun to understand how the various proteins present in AZs work together, and crucial issues such as the tight membrane association, the link to the cytoskeleton, the docking mechanism of synaptic vesicles, and the clustering of calcium channels at these hot spots of neurotransmitter release are only partially

understood. A fascinating question that has not been addressed regards the use-dependent structural plasticity of AZs and how it is involved in processes of memory formation and learning.

See also: Exocytosis: Ca^{2+}-Sensitivity; Liprins, ELKS, and RIM-BP Proteins; Munc13 and Associated Molecules; Piccolo and Bassoon; Presynaptic Development and Active Zones; Rab3A Interacting Molecules (RIMs); Synaptic Plasticity: Short-Term Mechanisms; Synaptic Vesicles.

Further Reading

Akert K, Moor H, and Pfenninger K (1971) Synaptic fine structure. *Advances in Cytopharmacology* 1: 273–290.

Augustin I, Rosenmund C, Sudhof TC, and Brose N (1999) Munc13-1 is essential for fusion competence of glutamatergic synaptic vesicles. *Nature* 400: 457–461.

Couteaux R and Pecot-Dechavassine M (1970) Synaptic vesicles and pouches at the level of 'active zones' of the neuromuscular junction. *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences* 271: 2346–2349.

Harlow ML, Ress D, Stoschek A, Marshall RM, and McMahan UJ (2001) The architecture of active zone material at the frog's neuromuscular junction. *Nature* 409: 479–484.

Phillips GR, Huang JK, Wang Y, et al. (2001) The presynaptic particle web: Ultrastructure, composition, dissolution, and reconstitution. *Neuron* 32: 63–77.

Schoch S and Gundelfinger ED (2006) Molecular organization of the presynaptic active zone. *Cell and Tissue Research* 326: 379–391.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

tom Dieck S, Sanmarti-Vila L, Langnaese K, et al. (1998) Bassoon, a novel zinc-finger CAG/glutamine-repeat protein selectively localized at the active zone of presynaptic nerve terminals. *Journal of Cell Biology* 142: 499–509.

Wang Y, Liu X, Biederer T, and Sudhof TC (2002) A family of RIM-binding proteins regulated by alternative splicing: Implications for the genesis of synaptic active zones. *Proceedings of the National Academy of Sciences of the United States of America* 99: 14464–14469.

Zhai RG and Bellen HJ (2004) The architecture of the active zone in the presynaptic nerve terminal. *Physiology* 19: 262–270.

Zhen M and Jin Y (1999) The liprin protein SYD-2 regulates the differentiation of presynaptic termini in *C elegans*. *Nature* 401: 371–375.

Activity in Visual Development

M B Feller, University of California at Berkeley, Berkeley, CA, USA
E S Ruthazer, McGill University – Montreal Neurological Institute, Montréal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The development of neuronal circuits involves initial coarse wiring under the guidance of molecular cues and the refinement of connections through mechanisms that are governed by patterned spontaneous activity and sensory experience. This fine-tuning of neuronal circuits is evident in the development of sensory maps in the brain. Sensory maps are organized layouts of neurons in which cells that prefer specific stimulus features are found in close physical proximity. The term map can apply either to the central representation of the sensory periphery, as in the case of retinotopic or other topographic maps, or to the orderly representation of higher order stimulus features such as orientation or interaural time difference.

Neural activity can be either permissive or instructive for circuit formation. When serving a permissive role, the presence of activity acts as a switch to regulate other downstream signaling events. In this case, neural activity essentially converts neurons from one state to another. An example of this is the ability to delay the onset of the critical period for plasticity by dark-rearing animals. Even brief visual experience can activate critical period plasticity in such animals. In contrast, an instructive role for activity is where the specific levels or patterns of neuronal firing carry information that allows different neurons to be distinguished from one another exclusively on the basis of these activity patterns. It is often difficult to prove unambiguously that activity plays an instructive role in a system because the simplest experiments involving blocking activity cannot distinguish between instructive and permissive roles. Nonetheless, there is compelling evidence from experiments in which activity patterns but not levels are altered, such as the ocular dominance shift in strabismic animals or the requirement for retinal waves of activity in the refinement of the retinocollicular map, that activity can also play an instructive role.

Activity in the developing visual system is not limited to that driven by visual experience. Throughout the visual system, in the thalamus, colliculus, and cortex, visual experience appears to be important in the maintenance of functional maps, but early development in these structures actually precedes vision. In this latter case, internally generated patterns of spontaneous activity play a key role in circuit refinement.

Sources of Activity in the Developing Visual System

In mammals, retinal ganglion cell (RGC) axons reach their targets before mammals' eyes open, and even before they have functional photoreceptors. In these early stages of development, in the absence of light responses, retinas are spontaneously active. These early activity patterns are called 'retinal waves' because they propagate across the ganglion cell layer, correlating the firing of tens to hundreds of RGCs. The synaptic circuits that mediate retinal waves are transient, with retinal waves disappearing as light responses are first developing. In contrast, in lower vertebrates, such as turtles, frogs, and chicks, there is an extended time during which waves and visual responses overlap. The synaptic mechanisms underlying retinal waves are described elsewhere in this encyclopedia.

An interesting feature of retinal waves is that they coincide with the period of development when visual responses are first detected in the retina. In mice, light-evoked responses have been detected as early as P10, which is 3 or 4 days before eye opening. In ferrets, which are born at approximately the same developmental stage as mice but have an elongated developmental period lasting 4 weeks until eye opening, light-driven responses are detectable in the dorsal lateral geniculate nucleus of the thalamus (dLGN) and visual cortex as much as 14 days before eye opening. Several experiments in mice, rats, and ferrets indicate that both spontaneous and light-evoked activity are detected in visual cortex before eye opening, indicating that both may influence developmental events. Visual deprivation by dark rearing, even when the eyelids are closed, alters the refinement of circuits within the retina and dLGN. Similarly, dark rearing and/or pharmacological manipulations of spontaneous activity have distinct influences on the development of RGCs in turtle retina, which have an extended period of light-evoked activity and retinal waves.

Once the eyes open, vision improves quickly, as determined by several measures. There is an immediate and steady increase in acuity and contrast sensitivity and a more gradual increase in spectral sensitivity. Neurons tuned to several features of the visual scene can be detected at eye opening, including ocular dominance and orientation.

Another source of neural activity that may be critically important for development is the activity patterns that are intrinsic to local circuits. Even in the absence of sensory input, there is a tremendous amount of spontaneous activity, some of which can be highly patterned. For example, spindle waves, which are fast oscillations in the cortical field potential generated by thalamocortical circuits, persist when the eyes are removed.

Retinotopic Maps in Tectum/Superior Colliculus

The two primary targets of RGCs in the brain are the superior colliculus (SC) and the LGN of the thalamus. In these targets, RGCs establish an arrangement of connections in target fields, termed a retinotopic map, that reflects the spatial arrangement of the RGCs in the retina, and eye-specific maps with inputs from the two retinas layering in neighboring but nonoverlapping regions.

The precise retinotopic and eye-specific targeting of RGCs axons observed in the adult emerges from initially diffuse and overlapping projections, prior to visual experience. There is a clear role for both neural activity and molecular factors, such as the ephrins and their corresponding receptors, for the establishment of these maps, although the relative importance of the two throughout the process of axon targeting and refinement is the subject of ongoing research. This article reviews the evidence that activity plays a role in the establishment of retinotopic maps.

The degree of retinotopic mapping can be assayed by different techniques. Most studies have relied on small focal injections of anterograde tracers (e.g., DiI) into the retina to visualize the axonal arbors in the SC/tectum, which is referred to as the termination zone. In addition, retinotopic maps have also been assayed by the spatial distribution of RGCs that are labeled by focal injections of retrograde tracers into the SC. Third, physiological measures in *in vitro* slices containing the optic tract and SC can assay the number of functional retinal inputs onto individual SC neurons. Last, *in vivo* physiological measurements of receptive sizes of SC/tectum neurons reveal the physiological consequences of topographic refinement.

The first preparations used for establishing a role for activity in retinotopic map formation were frogs and fish. These species have two advantages. First, topographic refinement occurs throughout life. The retina is constantly adding new cells at its periphery, whereas the tectum grows from the caudal end. Consequently, the retinal projections must constantly shift in order to maintain a retinotopic map. Second, in these species, RGC axons regenerate after injury, and hence maps can be studied while reforming in this more adult stage. Blockade of activity during either development or regeneration does not affect the course topography of projections but does profoundly affect the development of fine topography – the projections that mediate the fine point-to-point connectivity between RGCs and tectal neurons. These classic experiments led to the generation of a major dogma in developmental neuroscience that molecular cues mediate the development of course maps, whereas activity is important for the establishment of fine topography.

There has been growing evidence for activity also playing a role in the refinement of maps in mammalian systems. One fundamental difference between refinement in mammals and in frogs and fish is the location of axon branches that undergo refinement. In frogs and fish, refinement is mediated by small-scale changes in higher order axon branching emerging from the tip of the RGC axon. In contrast, during the development of retinocollicular maps in mammals (and similar to retinotectal maps in chicks), RGC axons overshoot their targets in the A–P axis. Branching in the appropriate retinotopic location occurs along the RGC axonal shaft, at sites anterior to the growth cone. Then, in what appears to be a distinct process, the overshooting axon and, in some cases, entire axonal branches are eliminated.

Pharmacological blockade in mammals leads to small, although significant, effects on the final level of retinotopy. To demonstrate a role for correlated retinal activity, mice that lack β_2-containing nicotinic acetylcholine receptors (nAChRs), which exhibit a pattern of retinal activity in which RGCs spike in a seemingly random pattern with little correlation between the spike trains of neighboring RGCs, have been examined. β_2-nAChR−/− mice exhibit less retinotopic refinement than wild-type mice. The absence of retinal waves in β_2-nAChR−/− mice is correlated with the irregular refinement of retinotopic maps despite the presence of approximately normal levels of activity in individual RGCs. Similar results were obtained with intraocular injections of nAChR antagonists, indicating that disruption of retinal waves can prevent the retinotopic refinement of retinocollicular projections.

Eye-Specific Maps in the Lateral Geniculate Nucleus

RGC axons project to the dLGN of the thalamus terminating in regions that are organized topographically and are segregated into eye-specific layers (i.e.,

projections from one eye end in regions spatially distinct from those of the other eye). When RGC projections from the eyes first grow into the dLGN, they are partially intermixed. The eye-specific layers then emerge gradually as the termination fields of the eyes segregate into regions containing either ipsilateral or contralateral retinal projections. This process is known to be activity dependent since intracranial infusion of TTX, a blocker of voltage-activated sodium channels, prevents segregation. Moreover, experiments have revealed that the activity driving this segregation comes from the retina since prolonged desynchronization of spontaneous retinal activity by parmacological disruption of nAChR activation in the eye also prevents layer formation. Blocking spontaneous activity in a single eye also significantly alters the distribution of RGC axons, indicating that competition from the two eyes is critical for the formation of eye-specific layers.

A wide array of transgenic mice and pharmacological manipulations have been used to gain insights into the mechanisms that mediate the formation of eye-specific layers. A few studies have taken advantage of the fact that the cellular basis of retinal waves switches from one mediated by nAChR to one mediated by ionotropic glutamate receptors to transiently block retinal waves. If nAChR-mediated retinal waves are eliminated during the initial period of refinement, either by pharmacological manipulation or by using β_2-nAChr−/− mice, eye-specific layers fail to form. However, even in the absence of the initial establishment of layers, RGC axons segregate into local eye-specific regions, with ipsilaterally regions segregated into small islands within the contralateral region. Thus, axons segregate without forming distinct eye-specific layers, indicating that eye-specific segregation and layer formation are separable processes that may occur through different mechanisms.

Retinal activity is essential not only for the establishment but also for the maintenance of eye-specific layers. In ferrets, intraocular injections of APB block glutamate-mediated waves after layers have been established and it has been shown to result in desegregation. In no-b-wave (*nob*) mice, RGCs fire in very frequent synchronous bursts that desegregate after layers have been established.

Whether retinal waves provide an instructive or permissive signal for driving eye-specific segregation is controversial. β_2-nAChR−/− mice do not form eye-specific layers, but pharmacological and genetic manipulations that significantly disrupt nearest neighbor correlations by increasing the uncorrelated firing of RGCs do not prevent layers from forming. The resolution of this controversy may rely on gaining insights into what aspect of the highly correlated activity is critical or driving refinement. Information required for activity-dependent segregation might be encoded in the slow periodic firing generated in individual neurons by waves. These periodic bursts of action potentials lead to substantial increases of intracellular calcium concentration in the participating neurons. There is growing evidence that periodic changes in intracellular calcium occurring on the order of minutes can profoundly influence a variety of intracellular processes. Thus, the periodicity of circuit activation may be tuned to the periodicity of intracellular signaling required to ensure the normal maturation of neurons in the retina or the segregation of retinal inputs in the dLGN.

Similarly, important information might be encoded in the spatial pattern of the activity. Synchronous activation of cells contains no distinct spatial information regarding the relative positions of cells involved in each event. However, the propagating activity seen in the retina synchronizes the activity of subsets of cells, thereby encoding their relative positions. A single retinal wave synchronizes firing of cells along a wavefront with a particular orientation on the retina, generating an activity pattern that might be used to establish orientation selectivity in visual cortical neurons. Activity patterns averaged over a large number of waves would lose orientation information but would maintain highly correlated firing among neighboring neurons, thus providing information that might be used to establish topographic projections.

The resolution of the question of whether the retinal waves are instructive or permissive for map refinement will rely on better targeted disruptions of spontaneous retinal activity based on a deeper understanding of the cellular mechanisms underlying plasticity.

Ocular Dominance Column Formation and Plasticity

The segregation of retinal inputs from each eye into eye-specific layers in the dLGN sets the stage for further segregation of eye-specific inputs in the thalamocortical projection to primary visual cortex. Transneuronal labeling studies in which radioactive amino acid or other anterograde neuronal traces that can jump synapses such as wheat germ agglutinin–horseradish peroxidase (WGA-HRP) are injected into one eye reveal a high degree of segregation of thalamocortical afferents into ocular dominance columns (ODCs) in layer 4 of the visual cortex of carnivores and certain primates, including humans.

Note, however, that eye-specific segregation is not evident in the visual cortices of all mammalian species: mice, rats, and even highly visual animals such as squirrels lack ODCs. Regardless of whether a species has segregated ODCs in the binocular zone of its visual cortex, binocular responsiveness (along with orientation and direction selectivity) of neurons is an important emergent property of the cortex not found at earlier levels of the visual system in normal adult animals.

This binocularity has proven to be a remarkably useful tool for studying cortical developmental plasticity. The pioneering work of Hubel and Wiesel revealed that deprivation of visual information through one eye by eyelid suture or image defocusing, known as monocular deprivation (MD), results in a dramatic shift of the responsiveness of cortical neurons to favor the nondeprived eye. These changes in the response properties of individual neurons are accompanied in most cases by a corresponding loss of visual acuity through the deprived eye, or amblyopia. The ocular dominance (OD) shift in response to MD is particularly powerful during a limited critical period in development, although evidence suggests that some degree of shift is possible even in adults. During the critical period, MD for as little as 1 day leads initially to a reduction of responsiveness to the deprived eye followed by an enhancement of the response driven by the nondeprived eye.

These changes are stronger and more rapid in extragranular (outside layer 4) layers of visual cortex, suggesting that plasticity of local cortical circuitry may guide the process. Nonetheless, the MD shift does propagate to cells in layer 4 and ultimately back to the dLGN. With less than 1 week of MD in the cat, the axonal arbors of thalamocortical neurons representing the deprived eye shrink, whereas those representing the nondeprived eye expand. Bulk transneuronal labeling from the eyes, as well as physiological assays, also reveals a gross shrinkage of deprived eye ODCs and a corresponding expansion of nondeprived eye columns. This propagation of OD plasticity back to successively earlier stages of the visual processing stream suggests the existence of retrograde messengers. This is further supported by numerous experiments in which blockade of spiking activity or of synaptic transmission through *N*-methyl-D-aspartate (NMDA) receptors in cortical neurons results in a disruption of the OD plasticity of thalamic afferents.

In addition to MD, misalignment of the eyes, or strabismus, during the critical period can also lead to a shift in the distribution of OD responses of cortical neurons, including the selective loss of binocular responses accompanied by a sharpening of ODC borders. This observation, together with the fact that comparable periods of binocular deprivation do not result in significant loss of visual responsiveness, argues that the equilibrium of inputs representing the two eyes is the result of a developmental competitive process.

This raises the question of whether a competitive process such as ocular dominance plasticity could be responsible for the initial segregation of ODCs. This question remains unresolved. Computer simulations, as well as the finding that RGC axons spontaneously segregate in an activity-dependent manner into ODC-like stripes in the optic tectum of fish and amphibia, indicate that the information contained in the spontaneous firing of retinal ganglion cells should be sufficient in principle to drive segregation into ODCs without the need for a molecular scaffold. On the other hand, ODCs are evident soon after thalamocortical innervation prior to the onset of the critical period for MD effects, and they do not appear to be disrupted by monocular enucleation at this early time. The presence of spontaneous activity within the already segregated dLGN at this stage, however, does not permit activity-dependent segregation of thalamocortical inputs to be ruled out at this point.

Molecular Mechanisms of Plasticity

It has become increasingly clear that developmental plasticity in the visual system is not mediated by a single mechanism. For example, in the dLGN of the ferret, segregation of eye-specific layers does not appear to require NMDA receptor (NMDAR) activation, whereas the segregation of inputs from on-center and off-center RGCs into sublaminae in the dLGN is prevented by application of NMDAR antagonists. In the visual cortex, NMDARs appear to play a key role in ocular dominance plasticity because pharmacological blockade or genetic knock-down of cortical NMDARs prevents the shift of OD in response to MD. The fact that current through NMDARs in response to presynaptic glutamate release is blocked by Mg^{2+} ions except when relieved by concurrent depolarization of the postsynaptic neuron allows NMDARs to serve as molecular detectors of correlated pre- and postsynaptic firing. Consistent with the role of NMDARs in the induction of synaptic plasticity, mutant mice deficient in the alpha isoform of Ca^{2+}/calmodulin-dependent protein kinase (CaMKII) or the serine/threonine phosphatase calcineurin, required for NMDAR-mediated long-term potentiation and long-term depression, respectively, lack normal OD plasticity.

Long-lasting neuronal plasticity generally requires protein synthesis. This is also true for OD plasticity

because application of cyclohexamide to visual cortex (but not LGN) to inhibit protein synthesis prevents the OD shift. The identities of the gene products required for OD plasticity have not been revealed. However, at least two key regulators of gene transcription, the extracellular signal-related kinase (ERK) and cyclic adenosine monophosphate response element-binding (CREB) transcription factors, which have both been implicated in long-term synaptic plasticity, are required to produce an OD shift. These protein synthesis-dependent pathways may be important for long-lasting structural plasticity, such as axonal arbor and dendritic spine remodeling in visual cortex.

Structural plasticity involves both the assembly of new connections and the dismantling of existing connections. Existing connections may be stabilized by interactions with the extracellular matrix and through cell–cell adhesion and signaling. Consistent with this model, activity of the serine protease tissue plasminogen activator has been shown to facilitate the OD shift during the critical period, leading to dendritic spine remodeling. After the critical period, a large degree of plasticity can be restored under conditions that reduce signaling by outgrowth inhibitory molecules such as chondroitin sulfate proteoglycans or the myelin inhibitor receptor NogoR.

These myriad molecular signaling cascades all nonetheless share a requirement for discriminable differences between the patterned neural activity in the two eyes. The ability of cortical neurons to detect these differences appears to rely critically on the balance between excitation and inhibition in the circuit. Evidence points to the developmental maturation of inhibitory circuitry, GABAergic basket cells in particular, as a key event for initiating the critical period for OD plasticity. The critical period in mice is opened precociously by augmenting the immature endogenous inhibitory circuitry with administration of the GABA-A receptor partial agonist diazepam. The excitatory–inhibitory balance may be regulated in part by the activity-regulated expression of brain-derived neurotrophic factor (BDNF). Such activity-dependent control of the susceptibility to undergo plastic changes, known as 'meta-plasticity,' is a prediction of the influential Bienenstock–Cooper–Munro (BCM) model for neuronal plasticity. The BCM model posits that activity levels determine a sliding threshold of input strengths required for synaptic modification, above which synaptic strengthening occurs and below which synapses are weakened. Alternatively, an important role for inhibition may be to help sharpen the temporal precision of firing of postsynaptic neurons in response to sensory inputs. Spike timing-dependent plasticity, in which synaptic changes depend critically on whether a postsynaptic neuron fires before or after its presynaptic partner, may be facilitated by maintenance of an appropriate excitatory–inhibitory balance.

The increasing availability of useful transgenic mouse models for the study of activity-dependent developmental plasticity ensures that many more candidate genes will be found to participate in this process.

See also: Adult Cortical Plasticity; Spontaneous Patterned Activity in Developing Neural Circuits; Vision: Light and Dark Adaptation; Visual Associative Memory; Visual Deprivation; Visual Development.

Further Reading

Demas J, Sagdullaev BT, Green E, et al. (2006) Failure to maintain eye-specific segregation in nob, a mutant with abnormally patterned retinal activity. *Neuron* 50: 247–259.

Grubb MS and Thompson ID (2004) The influence of early experience on the development of sensory systems. *Current Opinion in Neurobiology* 14: 503–512.

Hensch TK (2004) Critical period regulation. *Annual Review of Neuroscience* 27: 549–579.

Hofer SB, Mrsic-Flogel TD, Bonhoeffer T, and Hübener M (2006) Lifelong learning: Ocular dominance plasticity in mouse visual cortex. *Current Opinion in Neurobiology* 16(4): 451–459.

Huberman AD, Wang GY, Liets LC, Collins OA, Chapman B, and Chalupa LM (2003) Eye-specific retinogeniculate segregation independent of normal neuronal activity. *Science* 300: 994–998.

McLaughlin T, Torborg CL, Feller MB, and O'Leary DD (2003) Retinotopic map refinement requires spontaneous retinal waves during a brief critical period of development. *Neuron* 40: 1147–1160.

Pfeiffenberger C, Yamada J, and Feldheim DA (2006) Ephrin-As and patterned retinal activity act together in the development of topographic maps in the primary visual system. *Journal of Neuroscience* 26(50): 12873–12884.

Ruthazer ES and Cline HT (2004) Insights into activity-dependent map formation from the retinotectal system: A middle-of-the-brain perspective. *Journal of Neurobiology* 59: 134–146.

Torborg CL and Feller MB (2005) Spontaneous patterned retinal activity and the refinement of retinal projections. *Progress in Neurobiology* 76: 213–235.

Wiesel T (1981) *The Postnatal Development of the Visual Cortex and the Influence of Environment,* Nobel lecture.

Activity-Dependent Metabolism in Glia and Neurons

K A Kasischke, University of Rochester Medical Center, Rochester, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The extraordinary density and cellular diversity of the brain can possibly be recognized as a solution to an optimization problem in which the complex and energy-demanding tasks of this organ must be organized within the spatial constraints imposed by oxygen and substrate delivery through the vascular system and diffusion. The intimate spatial and functional interactions between the two most prominent cell types of the central nervous system – neurons and astrocytes – epitomize the division of metabolic and signaling pathways between fundamentally differing cells and provide both a key and a challenge for our understanding of this organ.

The historical view of energy metabolism in the human brain was centered on the neuron as the foremost energy consumer, with its energy needs exclusively met by oxidative metabolism of glucose. However, during the past few decades conclusive evidence has emerged that glial cells participate directly in neuronal energy metabolism in such a way that important metabolic pathways are strictly compartmentalized between neurons and their adjacent glial partners, particularly astrocytes. For instance, synthesis of glutamine in the central nervous system (CNS) occurs exclusively in astrocytes. Consequently, neurons depend on glutamine released by astrocytes to produce the neurotransmitters glutamate and γ-aminobutyric acid (GABA). In contrast, the enzyme glutaminase and the rate-limiting enzyme in the biosynthesis of the inhibitory neurotransmitter GABA, glutamic acid carboxylase, are present preferentially in neuronal cells.

The complex and highly dynamic metabolic interactions between neurons and astrocytes in the resting and in the activated brain have become a major and contested field in contemporary neuroscience. Our current knowledge of the metabolic compartmentalization and activity-dependent metabolism in glia and neurons of the brain relies primarily on biochemical studies on brain tissue slices and cultured neural cells using radioactively labeled substrates. This knowledge has been further expanded by the noninvasive measurement of metabolic fluxes in biochemical pathways in the intact brain. These remarkable insights have been made possible by the successful application of magnetic resonance spectroscopy (MRS) to study energy metabolism in both experimental animals and humans. However, none of the established methodologies can provide both the spatial and the temporal resolution to directly investigate cellular energy metabolism in the intact, living brain. Thus, fundamental questions relating to the spatial and temporal organization of the metabolic pathways in neurons and glial cells remain unresolved. For example, current measurements of metabolic fluxes in neurons and astrocytes depend critically on assumptions regarding the size and turnover rate of substrate pools on which there is no universal agreement. Furthermore, there is no general consensus on which substrates in addition to blood glucose the activated brain may utilize at rest and upon activation. Lactate may provide an alternative or even preferred oxidative substrate for neurons, and astrocytes readily metabolize additional substrates such as glycogen, acetate, and glutamate.

The Glutamate–Glutamine Cycle as the Axis of Metabolic Neuron–Glia Coupling

The field of metabolic interaction between astrocytes and neurons began to emerge when it was recognized in the 1950s and 1960s that the metabolism of glutamate was apparently divided between two distinct compartments within the brain. These two metabolic compartments were characterized in groundbreaking biochemical studies using radioactively labeled substrates on cortical tissue slices which identified a 'large,' presumably neuronal pool that contains the majority of glutamate and a 'small,' presumably glial pool that contains the majority of glutamine and in addition a small glutamate pool. These findings laid the foundation for the proposal of the glutamate–glutamine cycle by Benjamin and Quastel in 1975, with "flow of glutamate from neurons to glia, its conversion there into glutamine, and the return of glutamine to neurons and its conversion there to glutamate to complete the cycle." The absence of glutamine synthetase in neurons, endothelial cells, and choroid epithelium and its restricted localization in glial cells corroborated the glial origin of the small pool. The subsequent finding that glutamine synthetase is exclusively localized in astrocytes defined a specific metabolic role for these nonneuronal cells in glutamatergic neurotransmission. The unique metabolic position of astrocytes was further substantiated by the discovery that pyruvate carboxylase is another astrocyte-specific enzyme. The exclusive

presence of pyruvate carboxylase implies a distinctive role of oxidative metabolism in astrocytes in CO_2 fixation and anaplerotic metabolism, specifically the replenishment of citric acid cycle intermediates as precursors for amino acid neurotransmitters.

The first step of the glutamate–glutamine cycle is the uptake of synaptically released glutamate. Glutamate uptake occurs through plasma membrane-bound glutamate transporters. Although glutamate transporters are ubiquitously expressed in the CNS, their compartmentalized and cell-specific distribution provides another remarkable example of differentiation between the neuronal and glial compartments. The glutamate transporter subtype GLAST (EAAT1) is ubiquitously expressed, GLT1 (EAAT2) is highly specific for astrocytes, and EAAC (EAAT3) is predominately neuronal. It is believed that the functional role of neuronal transporters for glutamate uptake is minor; in contrast, the glial compartment has a high capacity for glutamate uptake. Experimental evidence implies astrocytes as the principal uptake site for synaptically released glutamate, even though their intrinsic transport rate is relatively slow. Glutamate transporters are highly expressed on the glial outer cell membranes, thus providing a high-affinity, high-capacity binding site for extracellular glutamate which effectively serves as a glutamate buffer. Electrophysiological measurements have shown that extracellular glutamate remains elevated for several milliseconds in the extrasynaptic space and activates astrocytic transporters even at considerable distance from the glutamate release site. Consequently, in mice with suppressed or missing transporter activity, failure of astrocytic glutamate uptake causes elevated glutamate levels, hyperexcitability, and excitotoxicity, as well as neurodegeneration.

Glutamate uptake directly links neuronal activity and astrocytic energy metabolism because cell culture studies have demonstrated immediate activation of both glycolytic and oxidative energy metabolism in astrocytes upon exposure to glutamate. Glutamate uptake is an energy-requiring process because it depends on the Na^+/K^+-ATPase to clear the three Na^+ ions from the intracellular space that are cotransported with each glutamate molecule. Intra-astrocytic glutamate can then be converted to glutamine by glutamine synthetase or be oxidized by the Krebs cycle, in which it is converted to the intermediary metabolites malate or oxaloacetate to yield diverse products such as lactate, aspartate, glutamine, and glutamate. Glutamate oxidation in astrocytes is a conceptually important pathway in cellular energy metabolism because it provides a significant source of oxidatively generated ATP that does not utilize glucose.

Although the concept of the glutamate–glutamine cycle was formulated early, the actual fraction of the total CNS glutamate involved in this pathway was not known. It was initially assumed that two functionally distinct glutamate pools existed – one small transmitter pool and one large metabolic pool. However, early MRS measurements in humans predicted a high rate of glutamate and glutamine labeling from ^{13}C-glucose, arguing against the existence of substantial glutamate and glutamine pools which do not participate in the cycle. It is current knowledge that the glutamate pool is uniform and that the glutamine–glutamate cycle is highly active in both rodents and humans.

Nuclear Magnetic Resonance Spectroscopy of the Glutamate–Glutamine Cycle and Resulting Quantitative Models of Neuronal–Glial Energetics

Our current understanding of the glutamate–glutamine cycle provides the fundament for the quantitative measurement of neuronal and astroglial energy metabolism and of glutamatergic and GABAergic neurotransmission by MRS studies. MRS is often referred to as nuclear magnetic resonance (NMR) spectroscopy. MRS allows the measurement of concentrations and synthesis rates of certain labeled molecules, such as glucose, glutamate, aspartate, GABA, ammonia, acetate, or lactate, in defined regions in the intact, living brain with high chemical specificity. These molecules need to be labeled with an MRS observable nucleus, typically ^1H, ^{13}C, or ^{15}N, and are applied intravenously to the experimental subject. The appearance and disappearance of these labels over time are then observed by MRS measurements. The obtained spectra are quantitative and even sensitive to the position of the labeled nucleus within the molecule, which provides an exceptional window to measure the activity of intracellular pathways *in vivo*.

Original MRS studies in anesthetized rodents reported an approximately 250% increase in the Krebs cycle flux during sustained sensory activation of the rat brain. The important implication is that oxidative glucose metabolism provides the required energy not only under resting but also under activated conditions. Further studies using functional imaging modalities such as positron emission tomography (PET) and magnetic resonance imaging (MRI) have consistently reported increases in regional cerebral oxygen metabolism during activation, albeit with much smaller amplitudes. MRS studies in anesthetized rats have shown that cerebral glucose metabolism in the Krebs

cycle and glutamatergic activity are linearly coupled with a slope close to 1 over a wide range of neural activities from deep anesthesia to shallow anesthesia closely approaching the awake, resting state. The quantitative coupling between glucose metabolism and glutamatergic neurotransmission with a 1:1 stoichiometry has also been reported in humans. The far-reaching conclusions from these MRS studies are that the majority (up to 80%) of oxidative energy production in the CNS is devoted to supporting glutamatergic neurotransmission and that oxidative glucose metabolism provides a quantitative measure of synaptic glutamate release. The localization of activity-dependent glucose uptake in the CNS to the projection zones of activated functional pathways in the neuropil provides independent support for this model.

In extension of these studies, a strictly quantitative model of neuronal–glial coupling that integrates the astrocyte–neuron lactate shuttle hypothesis was proposed. In this model, two ATPs are consumed to fuel the glutamate–glutamine cycle. The first ATP is utilized by the Na^+/K^+-ATPase for the restoration of the membrane potential after co-uptake of Na^+ with glutamate, and the second ATP is used for the synthesis of glutamine. As a consequence of the predicted 1:1 stoichiometry, one molecule of glucose would be oxidized for each molecule of synaptically secreted glutamate. The two ATPs needed to fuel the glutamate–glutamine cycle could be readily supplied by glycolytic production of two ATPs from the same glucose. Lactate as the product of astrocytic glycolysis would then be secreted in the extracellular space to be taken up and completely oxidized by neurons. An important implication of this model is the possibility of precisely quantifying glutamatergic neurotransmission as the very essence of functional brain activity in functional neuroimaging studies. Regarding the 1:1 stoichiometry coupling between glutamatergic neurotransmission and neuronal–glial glucose oxidation, a major revision has been put forward. It is still upheld that glial lactate is the major substrate for neuronal oxidation but with the fraction approximately 30% less than initially predicted. Glial cells produce at least 8% and neurons produce at least 88% of total oxidative ATP. Approximately 26% of the total glucose is taken up by neurons, leaving the majority of glucose for the glial compartment.

A caveat for the quantitative interpretation of glutamate–glutamine cycling as a measure of astrocytic energetics is that the previously mentioned oxidation of glutamate as an additional significant cellular energy source needs to be considered. As discussed later, astrocytes are likely to be highly oxidative cells, and their energetics may be underestimated if quantified by glutamate–glutamine cycling alone.

Nonlinear Coupling between Glucose and Oxygen Utilization in the Activated Brain

Further complicating the quantitative interpretation of energy metabolism in neurons and glial cells, the coupling between glucose utilization and oxygen consumption during activation is not linear. The apparent deviation from the long-held dogma that the brain is an entirely oxidative organ has been the subject of long-lasting discussions. The controversy was initiated by a landmark PET study that reported increases in glucose utilization in significant excess of oxygen consumption during focal neural activity. The proposal that the energy demands of the activated brain are fueled by glycolysis instead of oxidative metabolism marked a paradigm shift in the field of brain energy metabolism. Numerous studies reported increased lactate levels in the activated brain using differing methodologies, further substantiating this view. The fact that lactate accumulation occurs during a period of augmented oxygen delivery to the brain argued against the interpretation that increased lactate in the activated brain is the consequence of hypoxia. Moreover, the elevated oxyhemoglobin levels and decreased deoxyhemoglobin levels as a consequence of increased blood flow in the activated areas provide the basis for functional magnetic resonance imaging (fMRI) using the blood oxygen level-dependent signal.

Although it is now accepted that the ratio of the regional cerebral oxygen to glucose metabolism (oxygen–glucose index) decreases upon focal activation of the brain, a coherent and validated model of the underlying changes in neuronal and glial energy metabolism is still lacking. For example, the decrease in the oxygen–glucose index during activation is not consistent with the concept of linear coupling between glutamatergic neurotransmission and is at odds with the previously mentioned measurements of significant and sustained elevations of oxidative metabolism following sensory stimulation. A theoretical model to reconcile these contractions predicts that a fraction of the glucose taken up by astrocytes is cycled through the glial glycogen pool to supply rapid glial energy for clearance of glutamate while the remaining glucose is processed into lactate to support oxidative metabolism in neurons. In this scenario, both a transient decrease in the oxygen–glucose index and the increased production of lactate are accounted for, whereas the linear coupling between glutamate–glutamine cycling and oxidative metabolism in neurons remains preserved. An alternative explanation for the transient uncoupling of oxygen to glucose metabolism is that increased glial flux through pyruvate carboxylase removes car-

bons from oxidative glucose metabolism to supply anaplerotic metabolism in the glial Krebs cycle.

The Astrocyte–Neuron Lactate Shuttle Hypothesis

The astrocyte–neuron lactate shuttle hypothesis by Pellerin and Magistretti provides a possible mechanism for the quantitative coupling of glutamatergic metabolism to glucose utilization in the CNS and also an explanation for the apparent uncoupling of glucose and oxidative utilization in the activated brain. In its original form, this model predicted that uptake of synaptically released glutamate triggers glycolytic production of lactate in astrocytes. The lactate is released and immediately taken up by surrounding neurons to fuel activity-dependent oxidative metabolism. The lactate shuttle model was primarily based on the experimental observation that glutamate uptake activates glycolytic metabolism in astrocytes with significant accumulation of lactate in the culture medium. This hypothesis is indirectly supported by the cell-specific distribution of lactate dehydrogenase isoforms, which were interpreted as facilitating the proposed directed transport of lactate from astrocytes to neurons. In its extreme form (discussed previously), the model proposed a strict and linear coupling between synaptic glutamate release and astrocytic glycolysis with export of lactate for complete oxidation in neurons.

However, the astrocyte–neuron lactate shuttle hypothesis did not specifically address the role of glycolysis in neurons and of oxidative metabolism in astrocytes and was at odds with PET measurements of cerebral oxygen consumption and theoretical considerations which suggested that neurons increase their oxidative metabolism in parallel with an increase in pyruvate, implying that glycolysis in neurons, not in astrocytes, determines the kinetics of the metabolic response to activation. Furthermore, a stringent explanation of why glutamate uptake in astrocytes should be strictly coupled and exclusively fueled by astrocytic glycolysis was lacking and, most important, the possibility that glutamate oxidation may easily provide the required energy was not excluded. Numerous NMR spectroscopy studies have provided substantial evidence that a major portion of the energy for glutamine synthesis in astrocytes is indeed derived from oxidative metabolism and not glycolysis. Finally, the metabolic consequences of glutamate uptake in astrocytes are not restricted to activation of glycolysis since numerous studies have established that glutamate is capable of increasing the oxidative rate in astrocytes as well.

Based on these and other findings, the astrocyte–neuron lactate model has been the subject of considerable debate and criticism. In an amendment of their model, Pellerin and Magistretti loosened the temporal coupling between initial glycolysis in astrocytes and successive oxidative metabolism in neurons and proposed that both pathways could occur in arbitrary sequence in activated brain areas. Observations of activity-dependent fluctuations of the fluorescent coenzyme nicotinamide adenine dinucleotide (NADH) are indeed consistent with spatiotemporal partitioning of oxidative and glycolytic metabolism between astrocytes and neurons, albeit in a reversed temporal sequence with early oxidative metabolism in neurons followed by subsequent activation of glycolysis in astrocytes (**Figure 1**).

Further experimental support for the astrocyte–neuron lactate shuttle hypothesis derives from NMR spectroscopy studies which demonstrated that activity-dependent glycolytic lactate production occurs primarily in a nonneuronal compartment, strongly implicating astrocytes as the primary lactate source. Independent observations that glutamate increases membrane transport of glucose in cultured astrocytes and mediates inhibition of glucose transport in cultured hippocampal neurons further sustain this model. In the latest amendment of the lactate shuttle model, it has been proposed that astrocytic lactate is used as an oxidative substrate for neurons even under resting conditions.

In summary, there is little doubt that astrocytes can release lactate and that neurons can utilize external lactate as an oxidative substrate. However, conclusive proof for the astrocyte–neuron lactate shuttle hypothesis requires the direct experimental observation of a directed transfer of lactate from astrocytes to neighboring neurons.

Subcellular Differentiation of Energy Metabolism within Neurons and Astrocytes

The compartmentation and differentiation of energy metabolism extends even within neuronal and glial cells. Potentially highly relevant results for the understanding of activity-dependent metabolism in glial cells and neurons can be found in older histological literature on enzyme localization. For example, most of the glycolytic activity in neurons is localized within the perikarya and proximal dendrites. In contrast, the distal segments of neurons typically contain numerous mitochondria, with the highest densities in the smaller dendritic branches and axon terminals. The subcellular distribution of cytochrome oxidase activity follows an analogous pattern. An electron microscopy study found consistent and substantial staining for hexokinase in astroglia and neurons, with the remarkable

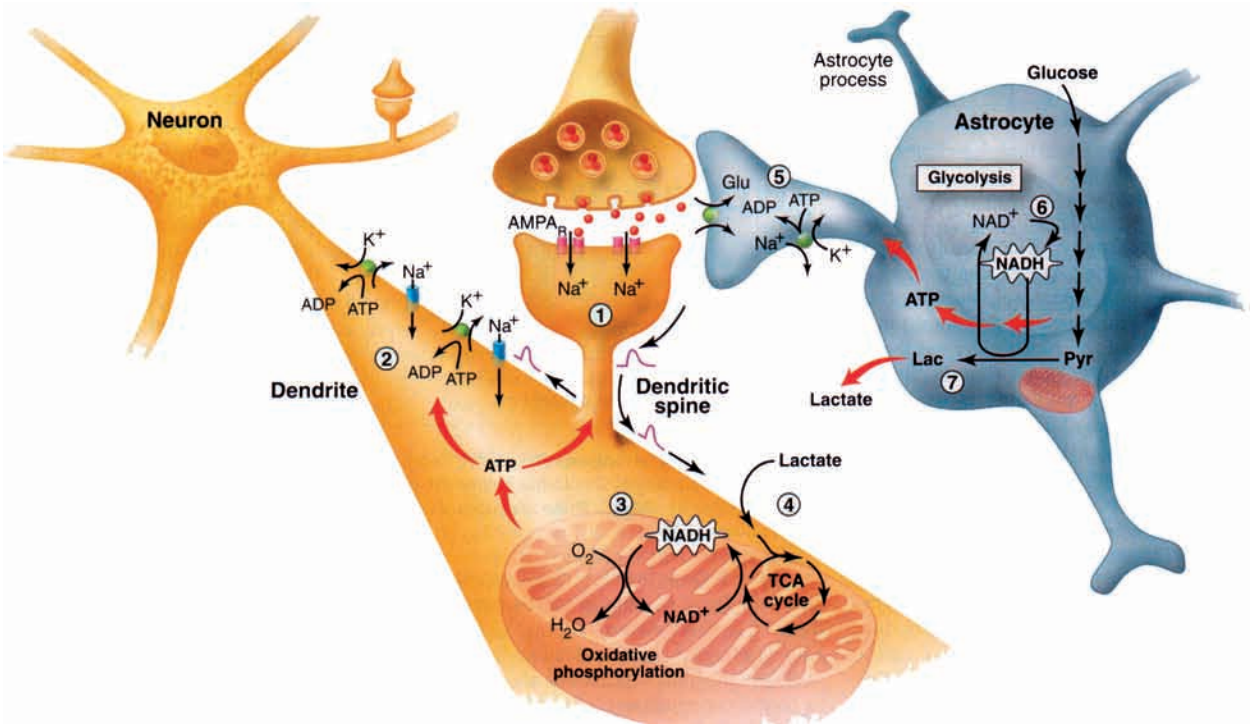

Figure 1 Brain energetics in the limelight. Separate activation of oxidative phosphorylation (respiration) in neurons (brown) and glycolysis in astrocytes (gray), as revealed by two-photon fluorescence imaging of NADH (4). (1) Stimulation of excitatory (glutamatergic) neurons activates postsynaptic AMPA receptors and induces an excitatory postsynaptic potential (EPSP) in the dendritic spine of the neuron. (2) The depolarization propagates from the dendritic spine to the dendrite, where it may cause further opening of voltage-gated sodium channels and activation of the Na^+/K^+-ATPase, leading to an increased demand for energy (ATP). (3) In response, oxidative phosphorylation is rapidly activated, causing a decrease in mitochondrial NADH content (the so-called 'dip' in the fluorescent signal). (4) Recovery of mitochondrial NADH in dendrites is accomplished by stimulation of the TCA cycle, fueled largely by lactate from the extracellular pool. (5) In parallel, but delayed in time, glutamate reuptake in astrocytes (gray) activates the glial Na^+/K^+-ATPase. (6) The increased energy demand leads to a strong enhancement of glycolysis in the cytoplasm of astrocytes, as indicated by the large increase in cytosolic NADH fluorescence (the so-called 'overshoot'). (7) To maintain the high glycolytic flux, NAD^+ must be regenerated via the conversion of pyruvate to lactate through the activity of the enzyme lactate dehydrogenase. Release of lactate into the extracellular space not only replenishes the extracellular pool but also may sustain the late phase of neuronal activation. *In vivo*, glucose is delivered from the blood to both the extracellular space and astrocytes (via astrocytic protrusions called end-feet that are in close contact with the blood vessel wall). $AMPA_R$, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors; GLU, glutamate; LAC, lactate; PYR, pyruvate; TCA, tricarboxylic acid. Reprinted from Pellerin L and Sutliff K, in Pellerin L and Magistretti PJ (2004) Let there be (NADH) light. *Science* 305: 50–52, with permission from AAAS.

exception that no hexokinase activity was found in the terminal segments of both Purkinje and granule cell dendrites. The corollary is that these dendritic terminals apparently do not depend on local metabolism of glucose to satisfy their energy needs. Another interesting finding is the localized presence of glycolytic enzymes in the postsynaptic density of the dendritic spines, with the implication that postsynaptic energy metabolism in the vicinity of the postsynaptic density may be fueled by glycolytic ATP production.

An important question concerns the number, density, and activity of mitochondria in astrocytes and neurons (Figure 2). A widely held misconception is that astrocytes contain substantially fewer mitochondria than neurons. However, even in the nineteenth century the presence of 'gliosomes' – large, dense granules in astrocyte processes – was noted by light microscopists. Later, these gliosomes were identified as mitochondria. An electron microscopy study reported that neurons and glia have approximately the same number of mitochondria, with the crystal volume fraction or crystal packing density being considerably higher in neurons (17.3%) than in astrocytes (11%). Numerous studies have found that the density of mitochondria is approximately equal within astrocytes and neurons, with the important implication that astrocytes are highly oxidative cells. These reports are in good agreement with MRS studies that report high glial Krebs cycle fluxes which account for approximately 30% of the total Krebs cycle activity, a percentage similar to the volume fraction of astrocytes in the CNS.

A striking subcellular differentiation between glial cells and neurons is the exclusive presence of glycogen

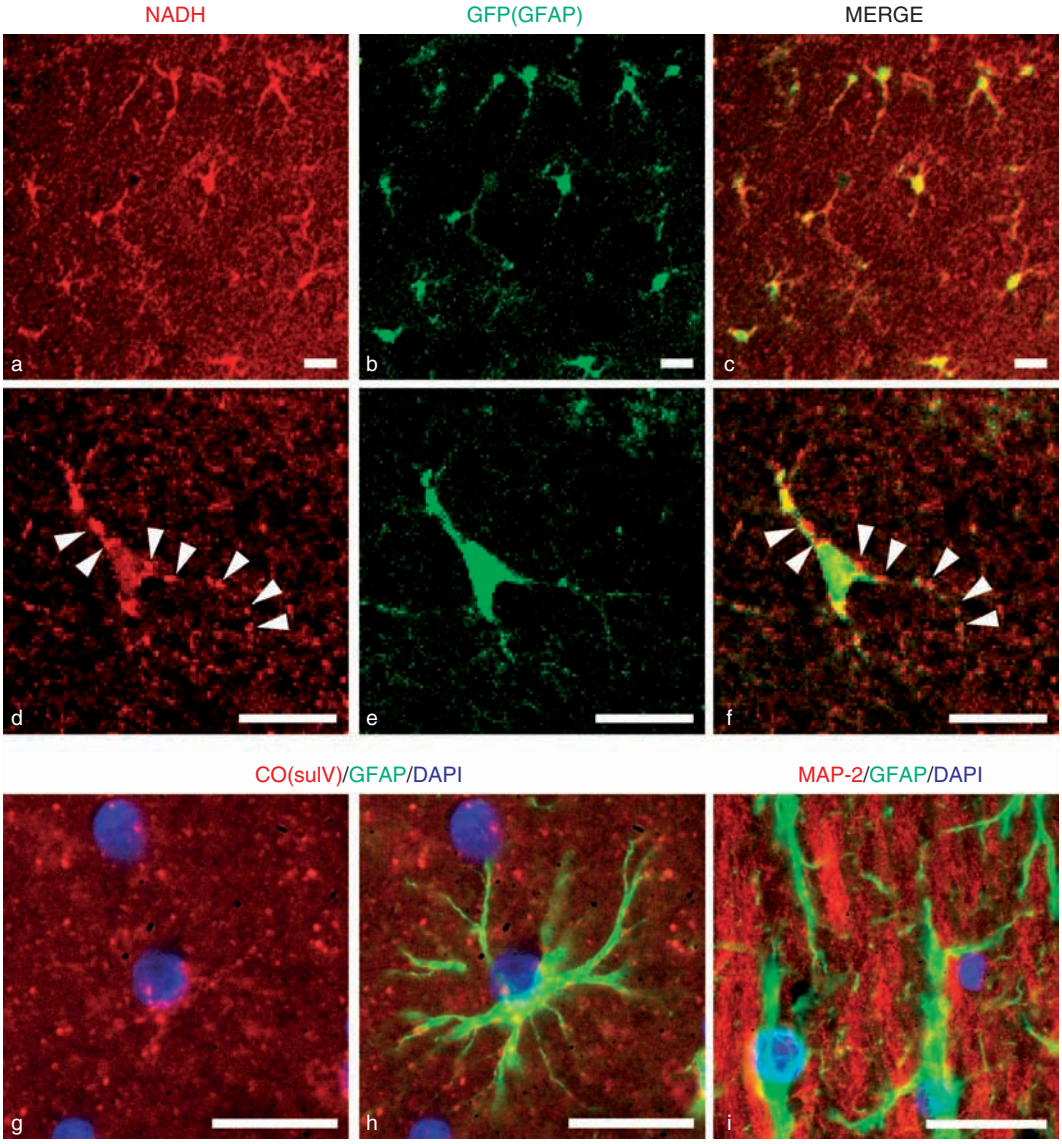

Figure 2 Astrocytes exhibit a characteristic NADH fluorescence pattern. (a–c) Astrocytes can be identified by means of their characteristic NADH fluorescence pattern as demonstrated by the 1:1 correlation of the disseminated, intrinsically bright fluorescent cells in the neuropil of CA1 (stratum radiatum) with green fluorescent protein (GFP)-expressing astrocytes. (d–f) A GFP-expressing astrocyte shown at high magnification with confluent cytoplasmic NADH fluorescence and granular mitochondrial NADH fluorescence (arrowheads) along its processes. (g) Distribution of mitochondria (cytochrome oxidase subunit IV immunolabeling) in the neuropil of CA1 (stratum radiatum). (h) Additional glial fibrillary acidic protein (GFAP) labeling of the same section shows that the concentration of mitochondria in astrocytes is not elevated in comparison with the surrounding neuronal neuropil. (i) Astrocytes (GFAP immunolabeling) and dendrites (MAP-2 immunolabeling) are the two dominant compartments in the neuropil of the stratum radiatum of CA1. Nuclei were counterstained with DAPI in (g–i). Scale bars = 20 μm. Reprinted from Kasischke KA, Vishwasrao HD, Fisher PJ, Zipfel WR, and Webb WW (2004) Neural activity triggers neuronal oxidative metabolism followed by astrocytic glycolysis. *Science* 305: 99–103, with permission from AAAS.

in astrocytes. The greatest accumulation of astrocytic glycogen is found in areas of high synaptic activity and in close proximity to neuronal perikarya. Consequently, it was proposed that astrocytic glycogen levels may be influenced by changes in neuronal activity and that astrocytic glycogen may even be considered as a store for lactate rather than for glucose. The unique morphology of astrocytes further suggests a special necessity and role for glycolysis and glycogenolysis, simply because the fine astrocyte processes and especially their filopodial and lamellipodial extensions are too narrow for mitochondria. The contribution of mitochondrial ATP provided by diffusion into this exceptionally narrow compartment is uncertain. A special role for glycolysis in astrocytes is further supported by the distribution of the reduced

coenzyme NADH, with the cytoplasmic levels of NADH in astrocytes apparently substantially higher than in the surrounding neuronal neuropil, indicating the possibility of higher glycolytic capacities in astrocytes than in neurons (Figure 2).

Activity-Dependent Metabolic Transitions in Neurons and Astrocytes

Principal questions related to activity-dependent metabolism in neurons and glia remain unresolved. Prominent examples are the role of cerebral glycogen metabolism, conclusive evidence for the cellular localization of lactate production and utilization or efflux, and the exact extent and function of oxidative metabolism in the glial compartment. A fundamental limitation has been the lack of spatial resolution of current functional neuroimaging and spectroscopy techniques to directly resolve the intertwined neuronal and glial compartments in the intact living brain.

Another important consideration is that many reported measurements have been obtained under steady-state conditions, and not much is known about the complex dynamic changes which occur in the relatively short transitional period between the resting state and the fully activated state. The methodological challenges are significant because metabolic transitions typically occur on a millisecond to second timescale and are difficult to capture with the relatively low temporal resolution offered by MRS, PET, and fMRI. Further complicating their interpretation, many reported metabolic transitions are biphasic and actually the sum of overlapping changes in utilization and delivery of a substrate at spatially close, but functionally and anatomically distinct, locations. Examples of such biphasic transitions include intrinsic imaging and spectroscopy of blood oxygenation, the measurement of interstitial tissue oxygen, interstitial glucose fluctuations, lactate dynamics, and NADH fluorescence fluctuations in electrically activated hippocampal tissue slices.

The consideration of metabolic transitions implies that coupling between neuronal and glial metabolism in the activated brain is more complex than that under steady-state conditions. A redox switch hypothesis has been proposed which may explain how the activity state of the brain directly determines the flux in cellular energy pathways. Here, the cytoplasmic redox state of the NADH/NAD^+–lactate/pyruvate couple in neurons and astrocytes acts as a metabolic switch, favoring either glucose or lactate as the preferred oxidative substrate for neurons.

Answers to difficult experimental questions can be expected from the successful application of novel methodologies. Perhaps most prominently among them, multiphoton microscopy has emerged as a popular and powerful tool providing many opportunities for the direct cellular resolution of neurometabolic and neurovascular coupling. For example, two-photon imaging and spectroscopy of NADH fluctuations in hippocampal tissue slices has provided evidence for enhanced glycolytic capacities in astrocytes and enabled the resolution of activity-dependent oxidative and presumably glycolytic responses in astrocytic and neuronal processes. Based on the observed NADH transitions, a model for neurometabolic coupling has been proposed in which early oxidative metabolism in neurons is eventually sustained by late activation of glycolysis in astrocytes. Similar observations were obtained using two-photon calcium imaging in the mouse barrel cortex *in vivo*, with delayed astrocytic calcium responses following prolonged whisker stimulation. In both exemplary cases, the observed astrocytic calcium and NADH responses prevailed after pharmacological blockade of postsynaptic activity, strongly implying a direct activation of astrocytic signaling and metabolism by synaptically released glutamate.

It will be exciting to determine whether microscopic functional imaging of the astrocytic and neuronal compartments in intact animals will provide a useful complement to established macroscopic imaging and spectroscopy technologies, and whether microscopic and macroscopic imaging approaches in combination with molecular and physiological techniques will ultimately yield a unifying model for the exceedingly complex metabolic interactions between neurons and astrocytes.

See also: Glial Glutamate Transporters: Electrophysiology; Glial Glutamate Transporters; Glial Glycogen Metabolism; Glial Glutamate and GABA Metabolism; Glial Energy Metabolism: A NMR Spectroscopy Perspective; Glutamate Receptor Organization: Ultrastructural Insights.

Further Reading

Balazs R, Machiyama Y, Hammond VJ, Julian T, and Richter D (1970) The operation of the gamma-aminobutyrate pathway of the tricarboxylic acid cycle in brain tissue *in vitro*. *Biochemical Journal* 116: 445–461.

Garcia-Espinosa MA, Rodrigues TB, Sierra A, et al. (2004) Cerebral glucose metabolism and the glutamine cycle as detected by *in vivo* and *in vitro* ^{13}C NMR spectroscopy. *Neurochemistry International* 45: 297–303.

Gjedde A, Marrett S, and Vafee M (2002) Oxidative and nonoxidative metabolism of excited neurons and astrocytes. *Journal of Cerebral Blood Flow and Metabolism* 22: 1–14.

Gruetter R (2004) *In vivo* ^{13}C-NMR studies of compartmentalized cerebral carbohydrate metabolism. *Neurochemistry International* 41: 143–154.

Hertz L (2003) Intercellular metabolic compartmentation in the brain: Past, present, and future. *Neurochemistry International* 45: 285–296.

Hertz L, Peng L, and Dienel GA (2007) Energy metabolism in astrocytes: High rate of oxidative metabolism and spatiotemporal dependence on glycolysis/glycogenolysis. *Journal of Cerebral Blood Flow and Metabolism* 27: 219–249.

Hyder F, Patel AB, Gjedde A, Rothman DL, Behar KL, and Shulman RG (2006) Neuronal–glial glucose oxidation and glutamatergic–GABAergic function. *Journal of Cerebral Blood Flow and Metabolism* 26: 865–877.

Kasischke KA, Vishwasrao HD, Fisher PJ, Zipfel WR, and Webb WW (2004) Neural activity triggers neuronal oxidative metabolism followed by astrocytic glycolysis. *Science* 305: 99–103.

Magistretti PJ, Pellerin L, Rothman DL, and Shulman RG (1999) Energy on demand. *Science* 283: 496–497.

Norenberg MD and Martinez-Hernandez A (1979) Fine structural localization of glutamine synthetase in astrocytes of rat brain. *Brain Research* 161: 303–310.

Patel HJ and Balazs R (1970) Manifestation of metabolic compartmentation during the maturation of the rat brain. *Journal of Neurochemistry* 17: 955–971.

Pellerin L and Magistretti PJ (1994) Glutamate uptake into astrocytes stimulates aerobic glycolysis: A mechanism coupling neuronal activity to glucose utilization. *Proceedings of the National Academy of Sciences of the United States of America* 91: 10625–10629.

Pellerin L and Magistretti PJ (2004) Let there be (NADH) light. *Science* 305: 50–52.

Rothman DL, Sibson NR, Hyder F, Shen J, Behar K, and Shulman RG (1999) *In vivo* nuclear magnetic resonance spectroscopy studies of the relationship between the glutamate–glutamine cycle and functional neuroenergetics. *Philosophical Transactions of the Royal Society of London, Series B* 354: 1165–1177.

Sibson NR, Dhankhar A, Mason GF, Rothman DL, Behar KL, and Shulman RG (1998) Stoichiometric coupling of brain glucose metabolism and glutamatergic neuronal activity. *Proceedings of the National Academy of Sciences of the United States of America* 95: 316–321.

Sonnewald U, Westergaard N, and Schousboe A (1997) Glutamate transport and metabolism in astrocytes. *Glia* 21: 56–63.

Activity-Dependent Regulation of Glucose Transporters

L F Barros, Centro de Estudios Científicos, Valdivia, Chile

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The energetic needs of the mammalian brain are provided almost exclusively by glucose, which is metabolized 10 times faster in the brain than in the rest of the body. Other energy sources, such as fatty acids, ketone bodies, and lactate, play minor roles in adulthood but are important during early embryonic development and lactation and under special circumstances (e.g., during strenuous exercise). After entering the intermediary metabolism at hexokinase, glucose is oxidized to CO_2 by the successive action of glycolysis and oxidative phosphorylation. In the brain the process is 90% efficient, with the expense of nearly six O_2 per glucose molecule. The free energy extracted from the sugar is spent mostly by neuronal cation pumps to restore ion gradients dissipated by postsynaptic currents and action potentials, thus a good correlation exists between neuronal signaling and local glucose usage. Like glucose, isotope-labeled tracer glucose analogs are preferentially metabolized in active regions, an observation that has permitted functional mapping of the brain by autoradiography and PET scanning.

Temporal Fluctuations in Local Glucose Demand

Throughout day and night, during mental exertion or rest, the overall metabolic demand of the brain remains fairly constant; however, local metabolic demand varies. For example, a striatal neuron jumping from electric silence to a firing rate of 20 Hz will increase its rate of ATP hydrolysis by a factor of 15. Because neurons do not store glycogen or other high-energy molecules, the new energy demand must be provided by glucose and/or lactate, the only readily available fuels present in the neuronal cytosol. The concentrations of glucose and lactate, about 1 mM, will sustain a 15-fold increase in demand for no longer than tens of seconds, implying that sustained neuronal activity must somehow increase the delivery of glucose/lactate to active regions. To examine how this may be achieved, it is useful to consider the architecture of the neuropil, the tissue where most ATP is spent. In the neuropil, dendrites and axons are surrounded by fine astrocytic processes, each astrocyte housing about 10^4 synapses. Astrocytes are in turn fixed to a capillary by means of their end feet, thin processes interconnected by glucose-permeable gap junctions. This complement of capillary, neighboring astrocytes, and their cargo of neuronal processes and synapses is termed the gliovascular unit (Figure 1). Due to the slowness of diffusion over long distances relative to the rapid fluxes involved, neuronal processes are energetically isolated from their soma. Instead, the fuel must be supplied locally, which defines the gliovascular unit as the metabolic unit of the brain. At variance with neurons, which can be fueled by either glucose or lactate, the gliovascular unit's only energy substrate is glucose, for lactate is not normally present at significant concentrations in the blood.

Although the firing rate of many neuronal types is known to be highly variable, it is not equally clear how variable is the metabolic demand of the gliovascular unit. The latter will depend on an unknown factor, which is the functional homogeneity of the neuronal processes in a given unit. For example, if all processes were to increase their metabolic demand by 15-fold, the whole unit will require an increase in glucose delivery of similar magnitude, the metabolic demand of endothelium and astrocyte being relatively smaller. But if a significant fraction of the processes in the unit belongs to other circuits, which do not become activated, or perhaps decrease their electrical activity, the maximum metabolic burden on the unit will be correspondingly lower. A further mechanism of metabolic load distribution is provided by glucose and lactate present in neighboring gliovascular units, their efficacy depending on how permeable is the interphase between units. In summary, sustained neuronal activity must be accompanied by enhanced flux of fuels into neurons and enhanced glucose flux into the tissue. Thanks to the spatial buffering provided by the gliovascular unit, the large increases in fuel delivery demanded by single processes need not be fully translated into increased glucose delivery from blood to brain. Studies utilizing autoradiography and positron emission tomography (PET) scanning, techniques that measure the uptake of glucose analogs, have reported maximum activity-dependent increases in brain metabolism of around twofold. Because these techniques at best average the behavior of thousands of gliovascular units over 20 min or longer, this value should be considered as a lower estimate of the actual degree of stimulation. As already discussed, the highest estimate is provided by the increase in metabolic flux in single neurons (e.g., 15-fold). Somewhere in between these two extremes lies the increase in glucose flux into the gliovascular unit, its precise

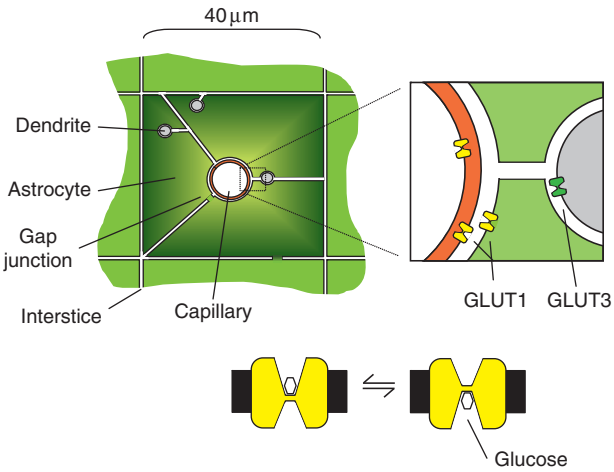

Figure 1 The gliovascular unit is the metabolic unit of the brain neuropil. Glucose diffuses into the gliovascular unit through facilitative transporters, present in endothelium and glia (GLUT1) and in neurons (GLUT3). GLUT transporters alternate their sugar binding site between the extracellular space and the cytosol, a process which is favored by the presence of the sugar.

quantification awaiting the development of new metabolic measurement techniques, capable of resolving at the level of micrometers and seconds.

Rate Limitation and Control of Glucose Flux into Brain Cells

Except for a few layers of cells lining the surface of the organ and the ventricles, the brain parenchyma is fed by blood-borne glucose. Overall glucose extraction by the organ is only 10%, which is a lower estimate of local extraction than occurs in active zones. Once inside brain cells, glucose enters the intermediary metabolic pathway at hexokinase, which is expressed in both astrocytes and neurons. By moving glucose beyond the permeability barriers imposed by the endothelial blood–brain barrier and the plasma membrane of neurons and astrocytes, hexokinase generates the concentration gradient that drives the sugar into the brain. This sole stimulation by hexokinase, by making the gradient steeper, may in principle augment flux. However, measurements of brain glucose concentration with microprobes and nuclear magnetic resonance (NMR) spectroscopy show that the gradient is nearly maximal even under resting conditions, leaving little room for further increase during activation. In an extreme case, in the striatum of the resting rat, glucose drops from 8 mM inside the capillary to 0.35 mM in the interstice, so hexokinase stimulation, however intense, may only enhance the consumption rate of glucose by less than 3%. In other regions of the brain, where glucose levels are about 1 mM, maximum flux stimulation would be around 15%. In addition to the gradient, the other factor that determines flux is permeability. But an isolated increase in permeability is not sufficient to modify flux either, because hexokinase, with its K_m of 50 μM, is nearly saturated. The result of this particular balance between permeability and metabolism is that the increase in glucose consumption demanded by neuronal activity requires commensurate stimulation of hexokinase and glucose permeability. Both glucose transporters (GLUTs) and hexokinase exert control over the rate of glucose consumption by the brain.

Mechanisms of Glucose Permeation into the Brain

Glucose is a hydrophilic molecule, thus its rate of diffusion across lipid bilayers is negligible. Glucose permeation into most mammalian cells is mediated by facilitative transporters of the GLUT family, of which 14 members have been described to date. Most of what is known about the function of GLUTs was learned from kinetic studies of the glucose transporter GLUT1, which is abundantly expressed in readily available erythrocytes. GLUT1 has a single hexose-binding site that becomes exposed alternatively to the extracellular space and the cytosol (Figure 1), a process that occurs 200 times per second in the absence of substrate and 10 times faster when glucose is bound to the site. The binding site is never exposed simultaneously to both sides of the plasma membrane, a key kinetic property that separates transporters from ion channels, which have a 'see-through' pore during the open state. GLUT carriers, however, do resemble channels insofar as they are unable to generate concentration gradients. The cloning of the *glut1* gene in 1985 revealed a highly hydrophobic 492-amino-acid protein, predicted to span the plasma membrane 12 times, with both ends and a large central loop oriented toward the cytosol. Thanks to biochemical work and mutagenesis, the sugar binding site has been mapped in detail, but the long-sought mechanism of catalysis remains elusive. Obtaining well-diffracting crystals of GLUT proteins has proved difficult; meanwhile a three-dimensional (3-D) structure of GLUT1 has been proposed based on the crystal structure of other transporters of the same superfamily. The transporter is visualized as having conic vestibules at each side. A large conformational change is thought to inundate the binding site with the water phase in alternating fashion, from one vestibule to the other.

The main GLUT isoform in the brain capillary endothelium is GLUT1, which according to immunogold studies is present at similar densities at both luminal and abluminal membranes, with a third pool

in the cytoplasm. Endothelial GLUT1 is a kinetically symmetric carrier, displaying K_m values of zero-trans influx and efflux of about 5 mM. The concentration of glucose in the brain interstitium is 10–20% of that in the blood, indicating that 80–90% of the overall flux resistance is endothelial. In astrocytes, the main glucose transporter is also GLUT1, which is conspicuously abundant in capillary-ensheathing end feet. Astrocytic GLUT1 is not glycosylated, as is the case for endothelial and erythrocytic GLUT1, a structural difference without apparent functional consequences. In neurons the main isoform is GLUT3, similar to GLUT1 in substrate specificity and affinity, but with a higher turnover rate. In microglia, the main isoform expressed is GLUT5, which has a high affinity for fructose. The low-affinity transporter GLUT2, found in liver and pancreatic islet cells, is also present in the hypothalamus and in tanycytes, astroglial cells that line the ventricles, where they may participate in gluco-sensing by the central nervous system. GLUT4, the carrier responsible for insulin-regulatable glucose transport in muscle cells and adipocytes, is also expressed in neurons scattered across the brain cortex. Several other isoforms of the GLUT carrier, including the intraneuronal GLUT8, are expressed at low levels throughout the brain; their physiological roles have yet to be determined. Of note, GLUTs can also transport dehydroascorbate, the oxidized form of vitamin C.

Acute Regulation of Glucose Transport

Neuronal activity causes an acute increase in local energy demand, which requires stimulation of glucose phosphorylation and a concomitant stimulation of glucose permeability. A possible molecular mechanism that may contribute to the increase in permeability was suggested by the observation that GLUT1 in cultured astrocytes is stimulated by the excitatory neurotransmitter glutamate, a phenomenon that develops in seconds, is reversible, and is not present in other cell types expressing GLUT1. The effect of glutamate is mediated by the Na^+–glutamate cotransporter and explained by a modification in transport capacity (V_{max}), but it is not yet known whether it results from translocation of GLUT1 to the plasma membrane or from an increase in the intrinsic activity of surface-resident transporters. Cultured neurons also react to glutamate reversibly and in seconds but, in contrast to that observed in astrocytes, the effect is an inhibition of hexose uptake. The neuronal inhibition is mediated by α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors, and, again, it remains to be demonstrated whether this inhibition responds to changes in surface transporter number or intrinsic activity. Both the stimulation of astrocytic GLUT1 and the inhibition of neuronal GLUT3 by glutamate are mediated by Na^+, which also commands the activation of astrocytic glycolysis by glutamate and participates in the waves of glycolytic activation that spread between adjacent gliovascular units, making Na^+ a key signal for the control of glucose distribution and usage in the brain. If these phenomena, observed in cultured cells, were to occur *in situ*, it would mean that glutamate redirects glucose toward astrocytes when fuel is badly needed in neurons. A possible solution for this paradox is provided by the astrocyte-to-neuron lactate-shuttle hypothesis (ANLSH), a model proposed by Pellerin and Magistretti that suggests astrocytic lactate, and not glucose, is the preferential substrate for active neurons. The proposed participation of glucose transporters in ANLSH is illustrated in Figures 2 and 3. In response to synaptic glutamate, astrocytes rapidly activate GLUT1 and the glycolytic machinery, generating lactate that diffuses to neurons through monocarboxylate transporters (MCTs). The simultaneous inhibition of neuronal GLUT3 by glutamate, mediated by AMPA receptors, slows down neuronal glycolysis, releasing NAD^+ at glyceraldehyde 3-phosphate dehydrogenase (GAPDH). NAD^+ is thus made available for lactate dehydrogenase (LDH), which drives lactate into the neuron. The energy cycle is completed when the ATP generated by neuronal mitochondria from pyruvate is used by the Na^+/K^+-ATPase to restore the ion gradients dissipated during electrical activity.

Given that the larger fraction of the total glucose gradient extends across the endothelium, endothelial GLUT1 is predicted to be acutely upregulated during neuronal activity – otherwise, glucose flux would not increase sufficiently during hexokinase stimulation. The detection of this phenomenon, which is expected to be highly local, is beyond current technology. However, a sizable increase in the V_{max} for glucose transport in whole brain is observed after only 3 min of seizures in the rat. A potential mechanism for the stimulation is suggested by the pool of GLUT1 inside endothelial cells, which may translocate to the cell surface and increase endothelial glucose permeability in response to neuronal activity, much in the way that GLUT4 translocates in muscle cells in response to insulin or contraction.

Slower Regulation of Glucose Transport

In addition to the acute response to glutamate, which is specific to astrocytes, slower stimulations of GLUT1 can be elicited in most cell types. Hormones, kinase

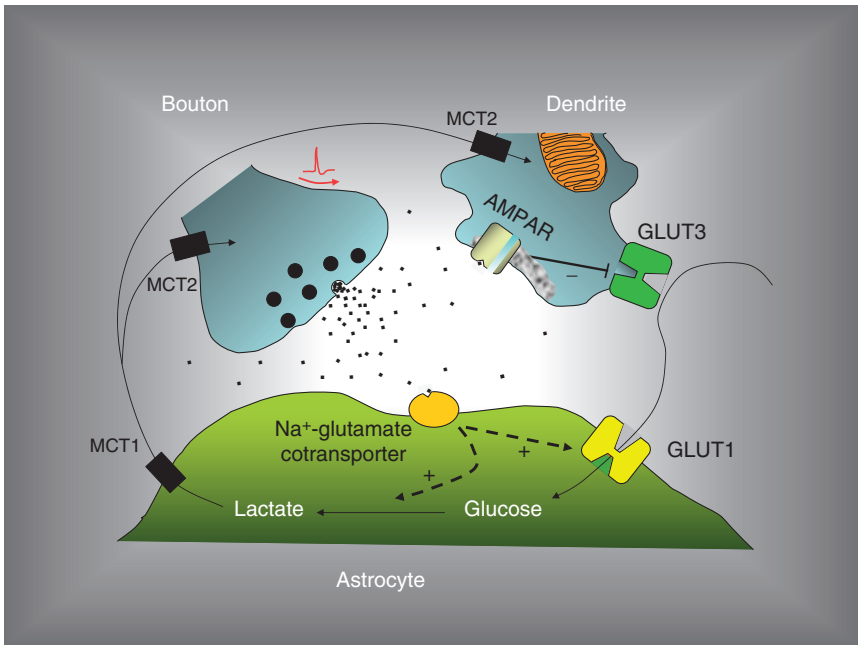

Figure 2 Glutamate released from synaptic boutons activates α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) in neurons and Na^+–glutamate cotransporters in astrocytes, leading to stimulation of the astrocytic glucose transporter GLUT1 and inhibition of the neuronal glucose transporter GLUT3. Glutamate also activates astrocytic glycolysis via the Na^+–glutamate cotransporter. The result is a surge in lactate, which diffuses through monocarboxylate transporters (MCTs) into neurons, where oxidation is completed.

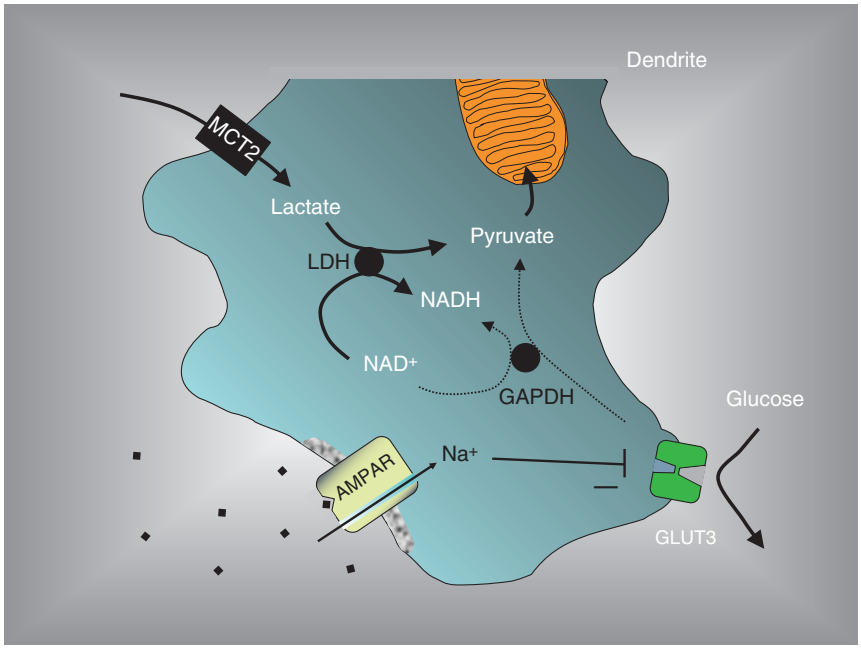

Figure 3 Glutamate opens ionotropic α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs), allowing the influx of Na^+ into postsynaptic spines and dendrites. Na^+, by an unknown mechanism, inhibits the glucose transporter GLUT3, decreasing the rate of glycolysis in neurons. The reduced flux through glyceraldehyde 3-phosphate dehydrogenase (GAPDH) releases NAD^+, which becomes available for lactate dehydrogenase (LDH) to deal with the lactate wave from astrocytes. Pyruvate is then metabolized in the mitochondria.

activators, hypoxia, metabolic poisons, viral infection, osmotic stress, and calcium ionophores all trigger a stereotyped biphasic response. In the first hour, the V_{max} for transport increases to two- to threefold. If the stimulus continues, a second phase that lasts 12–24 h takes the V_{max} to tenfold the basal value. The first phase of the response is due to translocation or activation of preexisting carriers. The delayed phase results from an increase in the total number of carriers, brought about by transactivation of the *glut1* gene. The same kind of adaptive phenomenon appears to be present in the brain, where increased GLUT1 expression has been reported in response to seizures, hypoxia, ischemia, and hypoglycemia. In contrast, the behavior of neuronal GLUT3 expression under metabolic demands is less consistent. Reportedly, GLUT1 expression, but not GLUT3 expression, increases in the hippocampus during learning.

Choice of Functional Probes for the Study of Glucose Transporters

Glucose transport in cell populations is often studied with isotope-labeled hexoses, both in culture and *in vivo*. After entering cells, labeled glucose is rapidly metabolized to lactate and CO_2. Because both products can leave the cell very rapidly, glucose uptake is not a reliable measure of glucose transport. 3-O-Methyl-D-glucose (3MG) is a good nonmetabolized substrate for GLUTs, but in brain cells it is transported so quickly that accurate measurement of initial rates at room temperature or higher is not easy. 2-Deoxyglucose (DG) is transported with high efficiency and then phosphorylated by hexokinase but does not proceed further. Although DG is by far the most popular substrate used to characterize GLUTs in the brain and elsewhere, DG data must be interpreted with caution, for depending on the times chosen for assay and the relative number of GLUTs and hexokinase, DG uptake rates may reflect the phosphorylation step rather than the transport step. Contrary to common wisdom, a linear time course of uptake should not be considered proof of the contrary. The transport of sugars into single cells can also be studied by fluorescence microscopy. A fluorescent derivative of DG, 6-NBDG, is a hexose that binds GLUTs with apparent affinity similar to that of glucose but translocates with very low probability, a useful property which allows real-time monitoring of uptake rates. 2-NBDG behaves similarly but it is phosphorylated by hexokinase. As with DG, whether changes in 2-NBDG uptake reflect modulation of transport or phosphorylation must be determined under each experimental condition. Finally, GLUTs can be studied by following osmotically obliged water movements in calcein-loaded cells using confocal microscopy. This technique works best for cells endowed with a high-capacity, low-affinity transporter and provides direct estimates of V_{max} values.

See also: Activity-Dependent Metabolism in Glia and Neurons; Glial Energy Metabolism: Overview; Glial Glutamate Transporters: Electrophysiology; Glial Glutamate Transporters; Glial Glycogen Metabolism; Noradrenaline.

Further Reading

Attwell D and Laughlin SB (2001) An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism* 21: 1133–1145.

Barros LF, Bittner CX, Loaiza A, et al. (2007) A quantitative overview of glucose dynamics in the gliovascular unit. *Glia* 55(12): 1222–1237.

Barros LF, Porras OH, and Bittner CX (2005) Why glucose transport in the brain matters for PET. *Trends in Neurosciences* 28: 117–119.

Loaiza A, Porras OH, and Barros LF (2003) Glutamate triggers rapid glucose transport stimulation in astrocytes as evidenced by real-time confocal microscopy. *Journal of Neuroscience* 23: 7337–7342.

Pellerin L and Magistretti PJ (1994) Glutamate uptake into astrocytes stimulates aerobic glycolysis: A mechanism coupling neuronal activity to glucose utilization. *Proceedings of the National Academy of Sciences of the United States of America* 91: 10625–10629.

Pellerin L and Magistretti PJ (2004) Neuroenergetics: Calling upon astrocytes to satisfy hungry neurons. *Neuroscientist* 10: 53–62.

Porras OH, Loaiza A, and Barros LF (2004) Glutamate mediates acute glucose transport inhibition in hippocampal neurons. *Journal of Neuroscience* 24: 9669–9673.

Vannucci SJ, Maher F, and Simpson IA (1997) Glucose transporter proteins in brain: Delivery of glucose to neurons and glia. *Glia* 21: 2–21.

Activity-Dependent Remodeling of Presynaptic Boutons

C H Bailey and E R Kandel, College of Physicians and Surgeons of Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Activity-dependent changes in the structure of the synapse accompany various forms of learning and memory. This is particularly well documented at the presynaptic connections of identified sensory neurons in the marine invertebrate *Aplysia californica*. Long-term sensitization of the gill-withdrawal reflex in *Aplysia* has been extensively studied in this respect and is associated with the growth of new synaptic connections between the sensory neurons and their postsynaptic target neurons. Recent *in vitro* studies of the sensory-to-motor neuron synapse reconstituted in dissociated cell culture have begun to define the mechanisms that underlie these presynaptic structural changes and their functional contribution to the different temporal phases of long-term memory. Insights provided by these studies in *Aplysia* suggest that the presynaptic remodeling and growth of new synapses induced by learning in the adult brain may reutilize some of the cellular and molecular mechanisms important for *de novo* synapse formation during development.

Learning-Induced Remodeling and Growth of Sensory Neuron Presynaptic Varicosities during Long-Term Sensitization in *Aplysia*

The structural mechanisms contributing to implicit memory storage have been most extensively studied for the gill-withdrawal reflex of *Aplysia*. As is the case with other types of defensive reflexes, several different forms of implicit learning can modify the gill-withdrawal reflex. We focus here on 'sensitization,' an elementary form of nonassociative learning by which an animal learns about the properties of a single noxious stimulus. The animal learns to strengthen its defensive reflexes and to respond vigorously to a variety of previously neutral stimuli after it has been exposed to a potentially threatening stimulus. In *Aplysia,* sensitization of the gill-withdrawal reflex can be induced by a strong stimulus applied to the tail. This activates facilitatory interneurons that synapse on identified sensory neurons to strengthen the synaptic connection between the sensory neurons and their target motor neurons. The behavioral memory for sensitization of the gill-withdrawal reflex is graded and retention is proportional to the number of training trials. A single stimulus to the tail gives rise to short-term sensitization lasting minutes to hours, while repetition of this stimulus produces long-term sensitization that can last for days or weeks.

Short- and long-term sensitizations lead to enhanced synaptic transmission at the monosynaptic connection between identified mechanoreceptor sensory neurons and motor neurons. This monosynaptic pathway can be reconstituted in dissociated cell culture in which serotonin (5-hydroxytryptamine (5-HT)), a modulatory neurotransmitter normally released by sensitizing stimuli, can substitute for the tail shock used during behavioral training in the intact animal. In parallel to behavioral sensitization, a single application of 5-HT produces short-term changes in synaptic effectiveness, whereas five spaced applications given over a period of 1.5 h produce long-term changes lasting 1 or more days.

Biophysical studies of this monosynaptic connection suggest that both the similarities and the differences in short- and long-term memory reflect, at least in part, intrinsic cellular mechanisms of the nerve cells participating in memory storage. Thus, studies of the connections between sensory and motor neurons in both the intact animal and in cells in culture indicate that the short-term changes are surprisingly similar to the long-term changes. A component of the increase in synaptic strength observed during both the short- and long-term changes is due, in each case, to enhanced release of transmitter by the sensory neuron, accompanied by an increase in the excitability of the sensory neuron, which is attributable to the depression of specific sets of potassium channels.

Despite this phenotypic similarity, the short-term changes differ fundamentally from the long-term changes in two important ways. First, the short-term change involves only covalent modification of preexisting proteins and an alteration of preexisting connections. Neither short-term behavioral sensitization in the animal nor short-term facilitation in dissociated cell culture requires ongoing macromolecular synthesis: the short-term change is not blocked by inhibitors of transcription or translation. By contrast, these inhibitors selectively block the induction of the long-term changes both in the semi-intact animal and in primary cell culture.

Second, the long-term but not the short-term process involves a structural change (see later). Long-term sensitization training is associated with the growth of new synaptic connections by the sensory

neurons onto their follower cells. This synaptic growth can be induced in the intact ganglion by the intracellular injection of cyclic adenosine monophosphate (cAMP), a second messenger activated by 5-HT, and can be reconstituted in sensory–motor neuron co-cultures by repeated presentations of 5-HT.

In the early 1980s, studies in *Aplysia* began to explore the morphological basis of the synaptic plasticity that may underlie the transition from short-term to long-term memory. By combining selective intracellular labeling techniques with the analysis of serial thin sections and transmission electron microscopy, complete reconstructions of unequivocally identified sensory neuron synapses were quantitatively analyzed from both control and behaviorally modified animals. The storage of long-term memory for sensitization (lasting several weeks) was accompanied by a family of distinct structural changes at identified sensory neuron synapses. These changes reflected a learning-induced remodeling of the functional architecture of presynaptic sensory neuron varicosities (boutons) at two different levels of synaptic organization: (1) alterations in focal regions of membrane specialization of the synapse that mediate transmitter release (the number, size, and vesicle complement of sensory neuron active zones were larger in sensitized animals than in controls) and (2) a parallel but more pronounced and widespread effect involving modulation of the total number of presynaptic varicosities per sensory neuron. Sensory neurons from long-term sensitized animals exhibited a two-fold increase in the total number of synaptic varicosities, as well as an enlargement in the size of each neuron's axonal arbor. The duration of the increases in active zone and varicosity number, which persisted unchanged for at least 1 week and were only partially reversed at the end of 3 weeks, paralleled the behavioral time course of memory, indicating that only changes in the number of sensory neuron synapses contribute to the retention of long-term sensitization.

These studies in *Aplysia* demonstrated that clear structural changes could accompany long-term behavioral modifications and indicated, for the first time, that learning-induced changes could be detected at the level of identified synaptic connections known to be critically involved in the behavior. Results from these initial studies of structural synaptic plasticity also provided evidence for an intriguing notion – that active zones are plastic rather than immutable components of the synapse and that even elementary forms of learning can alter the organization and number of transmitter release sites in the presynaptic compartment to modulate the functional expression of synaptic connections. In addition, complete reconstructions of sensory neuron varicosities in control animals revealed that approximately 60% of these presynaptic terminals lacked an active zone, suggesting the possibility of nascent or silent synapses in the adult brain. (The extent to which learning and memory can convert these immature synapses into mature and functionally competent synaptic connections is discussed in the following section.) Finally, these studies indicated that the growth of new sensory neuron synapses may represent the final and perhaps most stable phase of long-term memory storage, and suggested that the stability of the long-term process may be achieved, at least in part, because of the relative stability of synaptic structure.

This long-lasting growth of new synaptic connections between sensory neurons and their follower cells (both interneurons and motor neurons) during long-term sensitization can be reconstituted in dissociated sensory–motor neuron co-cultures by repeated applications of 5-HT. In culture, the structural change can be correlated with the long-term (24–72 h) enhancement in synaptic effectiveness and depends upon the presence of an appropriate target cell similar to the synapse formation that occurs during development.

Long-Term Facilitation Is Associated with Presynaptic Activation of Silent Varicosities and Growth of New Functional Synaptic Varicosities

These earlier studies did not examine whether the increase in synaptic strength during long-term sensitization resulted from the conversion of preexisting but nonfunctional (silent) synapses to active synapses, or from the addition of newly formed functional synapses, or both. To address these issues, more recent *in vitro* studies of the sensory-to-motor neuron synapse in culture have monitored both functional and structural changes simultaneously, so as to follow remodeling at the same specific synaptic varicosities continuously over time and to examine the functional contribution of these presynaptic structural changes to the different time-dependent phases of long-term facilitation (LTF). Toward that end, time-lapse confocal imaging of individual presynaptic varicosities of sensory neurons was combined with different fluorescent markers: the whole cell marker Alexa-594, and two presynaptic marker proteins, synaptophysin-e-GFP, which monitors changes in the distribution of synaptic vesicles within individual varicosities, and synapto-PHluorin (synPH), a monitor of active transmitter release sites. Repeated pulses of 5-HT were found to induce two temporally, morphologically, and molecularly distinct classes of presynaptic changes: (1) a rapid activation of silent presynaptic

terminals through the filling of preexisting empty varicosities with synaptic vesicles, which requires translation but not transcription, and (2) a generation of new synaptic varicosities which occurs more slowly and requires both transcription and translation. The enrichment of preexisting but empty varicosities with synaptophysin is completed within 3–6 h, parallels intermediate-term facilitation, and accounts for approximately 32% of the newly activated synapses evident at 24 h. By contrast, the new sensory neuron varicosities, which account for 68% of the newly activated synapses at 24 h, do not form until 12–18 h after exposure to five pulses of 5-HT. The rapid activation of silent presynaptic terminals suggests that in addition to its role in LTF, this modification of preexisting synapses may also contribute to the intermediate phase of synaptic plasticity and memory storage (Figure 1).

In this study, a reduced 5-HT protocol was used to induce selectively facilitation in the intermediate-term time domain without inducing LTF. Isolated intermediate-term facilitation was also accompanied by the redistribution and clustering of synaptic vesicle proteins into empty sensory neuron varicosities at 0.5 and 3 h, similar to what occurred when intermediate-term facilitation and LTF were recruited together. However, the presynaptic structural changes induced by the reduced 5-HT protocol differed from those induced by long-term training in at least two ways. First, there was no growth of new sensory neuron varicosities in the isolated intermediate phase. Second, unlike the filling of preexisting empty varicosities during the intermediate-term phase induced by the long-term protocol, the newly filled varicosities did not persist for 24 h and were unaffected by inhibitors of protein synthesis, suggesting that the structural remodeling induced by the reduced 5-HT protocol involved only a simple rearrangement of preexisting synaptic components. This may reflect a fundamental difference in the molecular mechanisms recruited by the two 5-HT protocols. Although both protocols induce intermediate-term facilitation, the long-term protocol may activate additional molecular events (including the machinery for translational activation) required to set up the long-term phase, perhaps by stabilizing the intermediate phase. At present it is not known how the covalent modifications that lead to the rearrangement of preexisting synaptic proteins at empty varicosities are converted by the long-term protocol to a more stable, protein-synthesis-dependent process.

In most model systems used to study long-lasting forms of synaptic plasticity, the functional contribution of the structural changes that accompany memory storage remains largely unknown. In particular, one would like to know if changes in the number or structure of synaptic connections induced by learning are functionally effective and capable of contributing to the storage of long-term memory. The synPH imaging of the sensory to motor neuron synapse in *Aplysia* represents an important step in addressing this issue. Previous studies have shown that specific 5-HT protocols or experimental manipulation in *Aplysia* can induce LTF at 24 h without the formation of new varicosities. How might such an increase in synaptic strength persist for 24 h in the absence of synaptic growth? Clearly, additional modifications

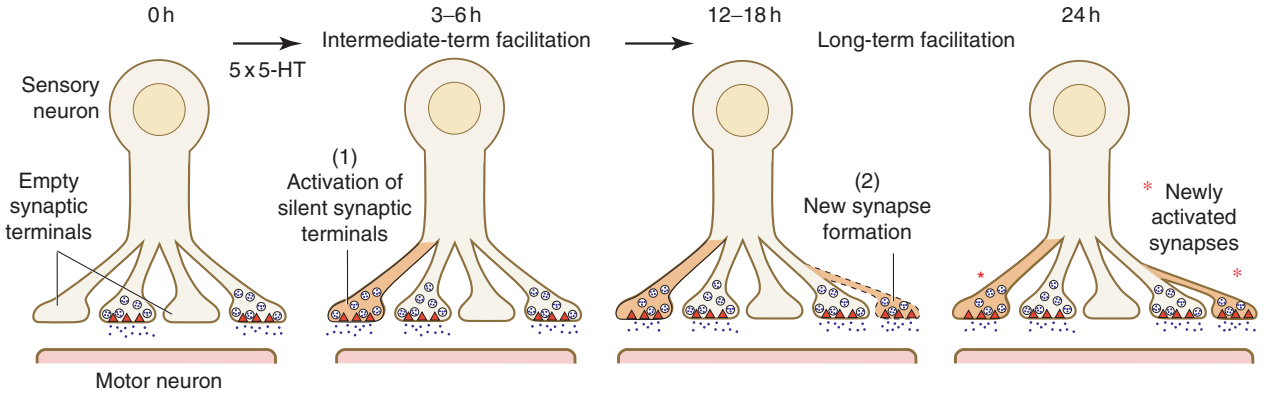

Figure 1 Time course and functional contribution of two distinct presynaptic structural changes associated with intermediate- and long-term facilitation in *Aplysia*. Repeated pulses of 5-HT in sensory to motor neuron co-cultures trigger two distinct presynaptic structural changes: (1) the rapid clustering of synaptic vesicles to preexisting silent sensory neuron varicosities (3–6 h) and (2) the slower generation of new sensory neuron synaptic varicosities (12–18 h). The resultant newly filled and newly formed varicosities are functionally competent (capable of evoked transmitter release) and contribute to the synaptic enhancement that underlies LTF. The rapid filling and activation of silent presynaptic terminals at 3 h suggest that, in addition to having a role in LTF, this modification of preexisting varicosities may also contribute to the intermediate phase of synaptic plasticity. Red triangles represent transmitter release sites (active zones).

of preexisting connections, including the activation of previously silent synapses previously outlined, may play an important role in the initial phases of synaptic maintenance and highlight the fact that there are likely to be multiple types of structural mechanisms that can contribute to LTF at 24 h.

Of the two classes of presynaptic structural plasticity induced by 5-HT in culture, synaptic growth appears to contribute more to the synaptic enhancement present at 24 h than does the activation of preexisting silent synapses. It will be of interest to see if the functional contribution by newly formed synapses increases with time when the growth process is more fully developed and memory storage is likely to be more stable. This would be consistent with the earlier studies in the intact animal outlined previously, which have shown that only the increases in the number of sensory neuron varicosities and active zones persist for several weeks in parallel with the behavioral duration of the memory, as well as more recent work in culture which has demonstrated that synaptic growth plays a more prominent role in the expression of the later phases of LTF.

5-HT-Induced Regulation of the Presynaptic Actin Network Is a Nodal Point for Learning-Related Synapse Remodeling and Growth

The 5-HT-induced enrichment of synaptic vesicle proteins and concomitant recruitment of active zone components in both preexisting and newly formed sensory neuron synapses during LTF involve an activity-dependent rearrangement of the actin cytoskeleton. How do repeated applications of 5-HT lead to reorganization of the actin cytoskeleton? The balance between actin polymerization and depolymerization is tightly regulated by extracellular signaling molecules, many of which act through the Rho family of GTPases. In *Aplysia*, the application of toxin B, a general inhibitor of the Rho family, blocks 5-HT-induced LTF, as well as growth of new synapses in sensory–motor neuron co-cultures. Moreover, repeated pulses of 5-HT selectively induce the spatial and temporal regulation of the activity of one of the small GTPases, Cdc42, leading to a rearrangement of the presynaptic actin network, followed by the assembly, insertion, and functional maturation of active transmitter release sites at sensory neuron varicosities. The 5-HT activation of ApCdc42 is dependent on signaling through the phosphatidylinositol 3-kinase (PI3K) and phospholipase C (PLC) pathways and, in turn, ApCdc42 activates the downstream effectors PAK (p21–Cdc42/Rac-activated kinase) and neuronal Wiskott–Aldrich syndrome protein (N-WASP).

The activation of ApCdc42 in sensory neurons leads to the outgrowth of filopodia from presynaptic varicosities. Interestingly, 5-HT stimulation by itself naturally induces filopodia, which is dependent on the activation of ApCdc42. Filopodia have been proposed to be a morphological precursor of dendritic spines in the mammalian nervous system, and this process may be regulated by neuronal activity. The 5-HT-induced activation of Cdc42 in *Aplysia* triggers not only the formation of filopodia but also the molecular maturation of new transmitter release sites. A major synaptic vesicle protein, synaptophysin, accumulates at the tips of 5-HT-induced filopodia, some of which then give rise to new varicosities. These observations support the following ideas: (1) filopodia represent one of the morphological precursors for the growth of new presynaptic varicosities during learning-related synaptic plasticity and (2) the formation of filopodia and initial assembly of the presynaptic compartment can be induced by the activation of Cdc42. Thus, 5-HT-induced regulation of the Cdc42 signaling pathways and the consequent reorganization of the presynaptic actin network appear to be a part of the initial molecular cascade required for the growth of new sensory neuron varicosities associated with the storage of long-term memory.

Activity-Dependent Modulation of Cell Adhesion Molecules and the Initiation of Learning-Related Presynaptic Growth

Studies in both higher invertebrates and mammals have shown that at critical developmental stages the refinement of synaptic connections, both their growth and regression, is determined by an activity-dependent process involving the modulation of cell adhesion molecules. This in turn has suggested that, in addition to their role in the development of the brain, cell adhesion molecules may also participate in the remodeling of synaptic architecture during learning-related synaptic plasticity. Some of the first evidence for a role of cell adhesion molecules during learning and memory came from studies of an immunoglobulin-related cell adhesion molecule in *Aplysia*, designated ApCAM, which is homologous to NCAM in vertebrates and Fasciclin II in *Drosophila*. 5-HT-induced synaptic growth in sensory–motor neuron co-cultures is associated with a decrease in ApCAMs on the surface membrane of the sensory neuron. This downregulation is particularly prominent at sites at which the processes of the sensory neurons contact

one another and is achieved by the protein-synthesis-dependent activation of a coordinated program of clathrin-mediated endocytosis, leading to the internalization and apparent degradation of ApCAM. *Aplysia* neurons express two isoforms of ApCAM – a membrane form and a phosphoinositol-linked form. Only the transmembrane isoform is internalized following exposure to 5-HT. This internalization can be blocked by overexpression of transmembrane ApCAM with a point mutation in the two mitogen-activated protein kinase (MAPK) phosphorylation consensus sites, as well as by injection of a specific MAPK antagonist into sensory neurons, suggesting that activation of the MAPK pathway is important for the internalization of ApCAMs and may represent one of the initial and perhaps permissive stages of learning-related synaptic remodeling and growth in *Aplysia*. Furthermore, overexpression of the transmembrane isoform, but not the glycosyl phosphatidylinositol (GPI)-linked isoform of ApCAM, blocks both LTF and 5-HT-induced synaptic growth. Long-term facilitation can also be blocked by overexpression of the cytoplasmic tail portion of ApCAM, designed to bind proteins such as MAP kinase p42. Taken together, these findings confirm that the extracellular domain of transmembrane ApCAM has an inhibitory function that needs to be neutralized by internalization to induce LTF and synaptic growth and that the cytoplasmic tail provides an interactive platform for both signal transduction and the internalization machinery.

The Presynaptic Remodeling and Growth of New Synapses Induced by Learning in the Adult Brain May Reutilize Mechanisms That Govern *De Novo* Synapse Formation during Development

As already discussed, long-term facilitation of the sensory to motor neuron synapse is accompanied by two temporally and morphologically distinct classes of presynaptic structural change: the rapid activation of silent preexisting varicosities by filling with synaptic vesicles and the slower growth of new functional varicosities. These findings, the first to be made on individually identified presynaptic varicosities, suggest that the duration of the changes in synaptic effectiveness that accompany memory storage may be reflected by the differential regulation of two fundamentally disparate forms of presynaptic compartment: (1) nascent (empty) silent varicosities that can be rapidly and reversibly remodeled into active transmitter release sites and (2) mature, more stable and functionally competent varicosities that, following long-term training, may undergo a process of fission to form new, stable synaptic contacts. What are the cellular and molecular mechanisms responsible for these two distinct classes of learning-related presynaptic structural change?

Some potential insights into this question are suggested by recent live imaging studies of developing synapses in the mammalian central nervous system (CNS). In *Aplysia*, the 5-HT-induced enrichment and subsequent activation of preexisting silent varicosities occur very rapidly – initial changes in the recruitment of synaptic vesicle proteins to empty presynaptic varicosities can be detected at 30 min after 5-HT training. This short delay is similar to what has been reported for the establishment of functional transmitter release sites in cultured hippocampal neurons and suggests that organization of the presynaptic compartment could be achieved by a rapid recruitment of preassembled active zone components to the sites of cell contact.

One attractive hypothesis that has been proposed to account for this rapid differentiation of the presynaptic compartment is based on a unitary model of active zone assembly. Here proteins that comprise the cytoskeletal matrix of the active zone (CAZ) are packaged into transport vesicles for delivery and fusion with the plasma membrane at nascent synaptic contacts. These precursor vesicles contain multiple active zone components, including proteins Piccolo and Bassoon – hence the name Piccolo/Bassoon transport vesicles (PTVs) – as well as other CAZ scaffolding molecules implicated in synaptic vesicle exocytosis, such as Rim, Munc13s, Munc18–1, syntaxin, Snap25, N-type calcium channels, α-liprin, and ERC/CAST, but typically not synaptic vesicle proteins. The assembly of each new active zone appears to be preceded by the recruitment and fusion of integer multiples of PTVs – typically two to five.

This model of developing synapses in the mammalian CNS provides a molecular mechanism of active zone assembly that is consistent with the rapid remodeling and presynaptic activation of preexisting empty sensory neuron varicosities induced by 5-HT in *Aplysia* culture. Moreover, the apparent heterogeneity in the content of these mobile preassembled packets – some of which contain synaptic vesicle proteins and the others which contain components required for assembly of the active zone – could explain why more than half of the sensory neuron varicosities enriched in synaptophysin following 5-HT treatment are not functional. It is likely the maturation of transmitter release sites is not yet complete in these varicosities and that varicosities which are enriched only in synaptic vesicle proteins, but lack a fully differentiated active zone, would not be functionally competent.

How are these modular transport packets induced by learning for presynaptic assembly targeted to empty sensory neuron varicosities? In *Aplysia* culture, only a specific subset of presynaptic varicosities becomes enriched in synaptic vesicles and activated, whereas others do not. Since repeated 5-HT treatment leads to a differential enrichment only in specific sensory neuron varicosities, proteins produced locally at each varicosity might contribute to this structural alteration. Alternatively, these protein components might not be synthesized within or nearby each varicosity, but might be transported and captured at specific varicosities that have been 'tagged' following 5-HT stimulation. Thus, the 5-HT-induced recruitment of synaptic vesicle and active zone proteins at a specific subset of sensory neuron varicosities, and the subsequent functional activation of these previously silent synapses, may be one of the local structural consequences of long-term synapse-specific plasticity. It should be noted that although considerable turnover of synapses and concomitant renewal of active zone proteins characterize the mature nervous system, PTVs are exceedingly rare when compared to developing synapses. It will be of interest to see if this developmental mechanism for presynaptic differentiation can be induced by learning and memory in the adult brain and if the learning-related recruitment of active zone proteins and subsequent assembly of transmitter release sites at empty sensory neuron varicosities are associated with an increase in the frequency of these modular precursor transport vesicles.

The second general class of learning-related presynaptic structural change associated with LTF in *Aplysia* is the 5-HT-induced formation of new sensory neuron varicosities. In culture, this increase in synapse number appears to be accomplished, at least in part, by the division or splitting of preexisting varicosities. Based on time-lapse confocal imaging, the apparent sequence of structural plasticity that gives rise to new varicosities involves the 5-HT-induced recruitment of synaptic vesicle proteins and components of the active zone to a preexisting varicosity, leading to both an enrichment of these presynaptic constituents and to an overall increase in the size of that varicosity. This remodeling and growth of preexisting varicosities ultimately result in the generation of new varicosities that are also enriched in synaptic vesicles and active zone proteins.

Precisely how these new varicosities come into being is still not clear. The close spatial association of preexisting and newly formed sensory neuron varicosities in time-lapse images, including the occasional confluence of both presynaptic compartments, suggests a process involving either a physical division of the original varicosity or a budding off of some components of the transmitter release site and associated synaptic vesicle cluster of that varicosity. Aspects of this reorganization of the presynaptic compartment that precedes learning-related synaptic growth in *Aplysia* have been reported at developing synapses in mammals. For example, recent imaging studies of the early stages of synapse formation have shown that presynaptic sites formed immediately after initial contact of axonal and dendritic processes are highly unstable. Moreover, even apparent mature presynaptic sites are relatively unstable, as occasionally 'orphan release sites' break off from fully formed boutons and either migrate to adjacent presynaptic sites or participate in the formation of completely new ones.

Combined, these observations indicate that differentiation of the presynaptic compartment, either induced by learning in the mature nervous system or as a mechanistic step during development, is a highly dynamic and rapid process that can recruit both preexisting proteins as well as preassembled synaptic components. These findings also suggest that there may be an upper limit to the size of an active zone and associated synaptic vesicle cluster that a neuron can construct. Any increase beyond this limit may cause the entire transmitter release apparatus to become unstable, perhaps because of some inherent metabolic or thermodynamic constraints. The final result is that smaller units of active zone material and cognate synaptic vesicle clusters can bud from the established site and translocate along the axon, where, under appropriate conditions, they may participate in the assembly of new presynaptic sites.

The instability of nascent presynaptic compartments during the early stages of neuronal differentiation is characterized by the dispersion of mobile packets of synaptic vesicles, synaptic vesicle precursors, and active zone precursors and a renewal of their migration once the transient pre- and postsynaptic contacts break up. As the neurons mature, an increasing proportion of these initial contacts develop into more stable and functionally competent presynaptic terminals. Results in the *Aplysia* sensory-to-motor neuron culture preparation indicate that these mature synaptic contacts can be selectively destabilized in an activity-dependent fashion during learning and memory.

How does the long-term process stabilize the transformation of labile, nascent presynaptic compartments into more mature and functionally competent release sites? Conversely, how are mature and fully functional presynaptic compartments destabilized to give rise to the formation of new release sites? Both the stabilization of presynaptic assembly and the

recruitment of destabilizing factors that may remove molecular inhibitory constraints and permit fission are likely to require some regulatory interaction with the postsynaptic neuron. Several studies have now suggested a potential role of the postsynaptic neuron in modulating the 5-HT-induced structural changes observed at presynaptic terminals in *Aplysia*. This interaction between postsynaptic and presynaptic neurons is critical for the formation and maturation of synapses during development, and may play a key role in both the pre- and postsynaptic expression of the structural plasticity associated with learning and memory storage. Moreover, since the two classes of 5-HT-induced changes in presynaptic structure that accompany LTF in *Aplysia* culture share their postsynaptic counterpart, there must also be transsynaptic signals, both anterograde and retrograde, to coordinate and regulate the learning-induced structural remodeling in an ongoing manner.

Obvious candidates for the molecules involved in triggering and stabilizing the presynaptic differentiation associated with long-term memory are those involved in synaptogenesis during development. With the extensive investigation in recent years of the developing neuromuscular junction and central synapses of both vertebrates and invertebrates, the list of such molecules has increased considerably. Once neuritic outgrowth and axonal pathfinding are completed and the incipient axon–target interactions occur, signaling molecules begin to engage in bidirectional communication to coordinate the differentiation of pre- and postsynaptic membrane specializations. Several cell surface molecules have been implicated in regulating differentiation of the presynaptic compartment during synaptogenesis, including WNT-7a, SynCAM, and the fibroblast growth factor (FGF) 22; these molecules might mediate this process through their adhesion and signaling capabilities. Some of these transsynaptic signaling systems employ heterophilic interactions that, in principle, could introduce an element of directionality required for the coordinated functional differentiation of the pre- and postsynaptic compartment. Among the transsynaptic molecular candidates, the β-neurexin–neuroligin interaction is particularly intriguing.

Neurexins are presynaptic transmembrane proteins present in many variants. Neurexins associate with synaptic vesicles by interaction with presynaptic scaffolding proteins such as CASK and Mints, which are present in the cytomatrix of the active zone and direct binding to synaptotagmin. Neuroligins are postsynaptic transmembrane proteins and bind to the PDZ domains of PSD-95, a scaffolding protein in the postsynaptic compartment of excitatory synapses. Thus, the β-neurexin–neuroligin interaction may act as a transsynaptic bridge, bringing synaptic vesicles into alignment with neurotransmitter receptor–ion channel complexes in the postsynaptic density. This hypothesis gained support when neuroligin expressed in nonneuronal cells was shown to cluster synaptic vesicles in contacting glutamatergic axons. Moreover, antibody-induced clustering of recombinant neurexin directly induced the co-clustering of synaptic vesicles.

Although details of neuroligin-induced presynaptic differentiation are not completely understood, aspects of the proposed consequences of this transsynaptic interaction during development might provide additional molecular insights into the mechanisms that underlie the two learning-related presynaptic changes in *Aplysia*. Neuroligin activity depends on the lateral clustering of individual neuroligin molecules that appear to induce the clustering of neurexin in the presynaptic membrane. These lateral clusters of neurexin may, in turn, activate signaling cascades that promote presynaptic assembly by recruiting scaffolding molecules such as CASK, Mint, and Veli, which interact directly with the PDZ-binding motifs in the cytoplasmic tail of neurexin.

Since CASK also interacts with proteins that regulate the actin–spectrin cytoskeleton, additional neurexin molecules and/or other synaptic components could be inserted into this presynaptic scaffold at the cytomatrix of the active zone, perhaps facilitating the 5-HT-induced growth and stabilization of the newly formed release sites at empty sensory neuron varicosities. Conversely, activity-dependent stimulation that gives rise to learning and memory could lead to a selective and synapse-specific interruption of the adhesive and/or signaling capabilities of the neuroligin–neurexin transsynaptic interaction at a subset of mature sensory neuron varicosities. This learning-induced alteration might, in turn, destabilize the cytoskeleton matrix of the presynaptic compartment – a transient and permissive step that would allow units of active zone material and associated synaptic vesicles to break off and to participate in the establishment of new presynaptic sites.

Conclusions

The morphological correspondence between studies in higher invertebrates and mammals suggests that learning resembles a process of neuronal growth and differentiation across a broad segment of the animal kingdom. Many of the mechanisms and signaling interactions utilized for experience-dependent remodeling of the synapse and the growth of new synaptic connections appear to share features in common with those that govern synaptogenesis. Recent studies of

the synaptic growth that accompanies LTF in *Aplysia* have begun to characterize the sequence of cellular and molecular events responsible for both the initiation and persistence of these presynaptic structural changes. This, in turn, has revealed that molecules and mechanisms important for synapse formation during the development of the nervous system may be reutilized in the adult for the purposes of synaptic plasticity and memory storage.

See also: Active Zone; Cell Adhesion Molecules at Synapses; Endocytosis and Presynaptic Scaffolds; Learning and Memory in Invertebrates: *Aplysia*; Presynaptic Facilitation; Presynaptic Inhibition; Presynaptic Development: Functional and Morphological Organization; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Bailey CH and Kandel ER (1993) Structural changes accompanying memory storage. *Annual Review of Physiology* 55: 397–426.

Bailey CH, Kandel ER, and Si K (2004) The persistence of long-term memory: A molecular approach to self-sustaining changes in learning-induced synaptic growth. *Neuron* 44: 49–57.

Goda Y and Davis GW (2003) Mechanisms of synapse assembly and disassembly. *Neuron* 40: 243–264.

Huntley GW, Benson DL, and Colman DR (2002) Structural remodeling of the synapse in response to physiological activity. *Cell* 108: 1–4.

Kandel ER (2001) The molecular biology of memory storage: A dialogue between genes and synapses. *Science* 294: 1030–1038.

Lamprecht R and LeDoux J (2004) Structural plasticity and memory. *Nature Reviews Neuroscience* 5: 45–54.

Scheiffele P (2003) Cell–cell signaling during synapse formation in the CNS. *Annual Review of Neuroscience* 26: 485–508.

Ziv NE and Garner CG (2004) Cellular mechanisms of presynaptic assembly. *Nature Reviews Neuroscience* 5: 385–399.

Addiction: Neurobiological Mechanism

G F Koob, The Scripps Research Institute, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Neurocircuitry of Drug Reward, Dependence, and 'Craving'

Substance dependence can be defined as a chronically relapsing disorder characterized by (1) compulsion to seek and take the drug, (2) loss of control in limiting intake, and (3) the emergence of a negative emotional state (e.g., dysphoria, anxiety, and irritability) when access to the drug is prevented (defined here as withdrawal). Addiction and substance dependence (as currently defined by the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn.) are used interchangeably throughout this article to refer to a final stage of a usage process that moves from drug use to abuse to addiction. The term dependence has two meanings: (1) to describe an acute withdrawal syndrome and (2) to describe a syndrome in which a subject meets the criteria for substance dependence. Other terms that need to be defined include 'reward,' which connotes a reinforcing stimulus with positive hedonic valence, and 'anti-reward.' Anti-reward is a concept based on the hypothesis that there are brain systems in place to limit reward and is represented at the neurocircuit level by a between-system neuroadaptation to activation of the reward system.

Clinically, the occasional but limited use of a drug with the potential for abuse or dependence (the second meaning given previously) is distinct from escalated drug use and the emergence of a chronic drug-addicted state. An important goal of current neurobiological research is to understand the neuropharmacological and neuroadaptive mechanisms within specific neurocircuits that mediate the transition from occasional, controlled drug use to the loss of behavioral control over drug-seeking and drug-taking that defines chronic addiction.

Addiction has been conceptualized as a chronic relapsing disorder with roots in both impulsivity and compulsivity and neurobiological mechanisms that change as an individual moves from one domain to the other. Subjects with impulse control disorders experience an increasing sense of tension or arousal before committing an impulsive act; pleasure, gratification, or relief at the time of committing the act; and finally regret, self-reproach, or guilt following the act. In contrast, individuals with compulsive disorders experience anxiety and stress before committing a compulsive repetitive behavior and then relief from the stress by performing the compulsive behavior. In addiction, drug-taking behavior progresses from impulsivity to compulsivity in a three-stage cycle: binge/intoxication, withdrawal/negative affect, and preoccupation/anticipation. As individuals move from an impulsive to a compulsive disorder, the drive for the drug-taking behavior shifts from positive to negative reinforcement.

Much of the progress in understanding the mechanisms of addiction has derived from the study of animal models of addiction. Although no animal model of addiction fully emulates the human condition, animal models do permit investigation of specific elements of the process of drug addiction. Such elements can be defined by models of different systems, models of psychological constructs such as positive and negative reinforcement, and models of different stages of the addiction cycle. Although much focus in animal studies has been on the synaptic sites and molecular mechanisms in the nervous system on which drugs with dependence potential act initially to produce their positive reinforcing effects, new animal models of components of the negative reinforcing effects of dependence, the transition to dependence, and propensity to relapse have been developed and are beginning to be used to explore how the nervous system adapts to drug use. The neurobiological mechanisms of addiction that are involved in various stages of the addiction cycle have a specific focus on certain brain circuits and the neurochemical changes associated with those circuits during the transition from drug taking to drug addiction and how those changes persist in the vulnerability to relapse.

A key element of drug addiction is how the brain reward system changes with the development of addiction, and one must understand the neurobiological bases for acute drug reward to understand how the reward systems change with the development of addiction. A principle focus of research on the neurobiology of the positive reinforcing effects of drugs with dependence potential has been on the activation of the circuitry related to the origins and terminals of the mesocorticolimbic dopamine system. There is compelling evidence for a critical role of this system in drug reward associated with psychostimulant drugs, and there is evidence that all major drugs of abuse activate this system as measured either by increased extracellular levels of dopamine in the terminal areas (e.g., the nucleus accumbens) or by activation of the firing of neurons in the ventral tegmental area. However, although selective neurotoxin-induced lesions of the mesolimbic dopamine system do block cocaine,

amphetamine, and nicotine self-administration, rats continue to self-administer heroin and alcohol in the absence of the mesocorticolimbic dopamine system, and place preference studies show robust place preferences to morphine and nicotine in the presence of major dopamine receptor blockade. Together these results suggest that activation of the mesolimbic dopamine system is a component of drug seeking in general but only critical for the rewarding effects of stimulant drugs.

Specific components of the basal forebrain associated with the amygdala also have been identified with drug reward. One hypothetical construct, the extended amygdala, includes not only the central nucleus of the amygdala (CeA) but also the bed nucleus of the stria terminalis (BNST) and a transition zone in the medial subregion of the nucleus accumbens (shell of the nucleus accumbens), and these regions share certain cytoarchitectural and circuitry similarities (Figure 1). As the neural circuits for the reinforcing effects of drugs with dependence potential have evolved, the role of neurotransmitters/neuromodulators also has evolved, and multiple neurotransmitter systems have been identified to have a role in the acute reinforcing effects of drugs of abuse in these basal forebrain areas: mesolimbic dopamine, opioid peptide, γ-aminobutyric acid (GABA), glutamate, endocannabinoids, and serotonin.

The neural substrates and neuropharmacological mechanisms for the negative motivational effects of

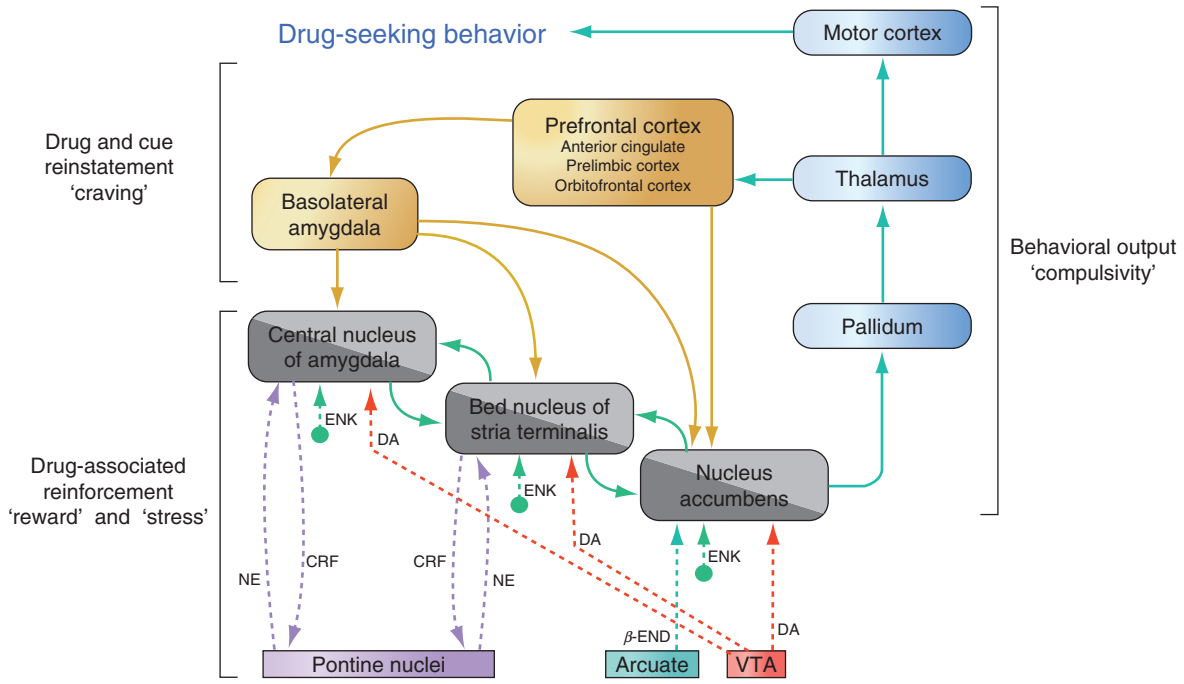

Figure 1 Key common neurocircuitry elements in drug-seeking behavior of addiction. Three major circuits that underlie addiction can be distilled from the literature. A drug-reinforcement circuit ('reward' and 'stress') is composed of the extended amygdala, including the CeA, the bed nucleus of the stria terminalis, and the transition zone in the shell of the nucleus accumbens. Multiple modulator neurotransmitters are hypothesized, including dopamine and opioid peptides for reward and corticotropin-releasing factor and norepinephrine for stress. The extended amygdala is hypothesized to mediate integration of rewarding stimuli or stimuli with positive incentive salience and aversive stimuli or stimuli with negative aversive salience. During acute intoxication, valence is weighted on processing rewarding stimuli, and during the development of dependence aversive stimuli come to dominate function. A drug- and cue-induced reinstatement ('craving') neurocircuit is composed of the prefrontal (anterior cingulate, prelimbic, and orbitofrontal) cortex and basolateral amygdala, with a primary role hypothesized for the basolateral amygdala in cue-induced craving and a primary role for the medial prefrontal cortex in drug-induced craving, based on animal studies. Human imaging studies have shown an important role for the orbitofrontal cortex in craving. A drug-seeking ('compulsive') circuit is composed of the nucleus accumbens, ventral pallidum, thalamus, and orbitofrontal cortex. The nucleus accumbens has long been hypothesized to have a role in translating motivation to action and forms an interface between the reward functions of the extended amygdala and the motor functions of the ventral striatal–ventral pallidal–thalamic–cortical loops. The striatal–pallidal–thalamic loops reciprocally move from prefrontal cortex to orbitofrontal cortex to motor cortex, ultimately leading to drug-seeking behavior. Note that for the sake of simplicity, other structures are not included, such as the hippocampus (which presumably mediates context-specific learning, including that associated with drug actions). Also note that dopamine and norepinephrine both have widespread innervation of cortical regions and may modulate function relevant to drug addiction in those structures. DA, dopamine; ENK, enkephalin; CRF, corticotropin-releasing factor; NE, norepinephrine; β-END, β-endorphin. Reproduced from Koob GF and Le Moal M (2006) *Neurobiology of Addiction*. London: Academic Press, with permission from Elsevier.

drug withdrawal may involve disruption of the same neural systems implicated in the positive reinforcing effects of drugs. Measures of brain reward function during acute abstinence from all major drugs with dependence potential have revealed increases in brain reward thresholds as measured by direct brain stimulation reward. These increases in reward thresholds may reflect decreases in the activity of reward neurotransmitter systems in the midbrain and forebrain implicated in the positive reinforcing effects of drugs and as such represent a 'within-system' neuroadaptation. Examples of such changes at the neurochemical level include decreases in dopaminergic and serotonergic transmission in the nucleus accumbens during drug withdrawal as measured by *in vivo* microdialysis, increased sensitivity of opioid receptor transduction mechanisms in the nucleus accumbens during opiate withdrawal, and decreased GABAergic transmission. One also sees differential regional changes in nicotine receptor function during nicotine withdrawal. The decreases in reward neurotransmitters have been hypothesized to contribute significantly to the negative motivational state associated with acute drug abstinence and also the long-term biochemical changes that contribute to the clinical syndrome of protracted abstinence and vulnerability to relapse.

Different neurochemical systems involved in arousal and stress modulation also may be engaged within the neurocircuitry of the brain stress systems in an attempt to overcome the chronic presence of the perturbing drug and to restore normal function despite the presence of drug. Glutamate, an excitatory neurotransmitter, has been implicated in neuroadaptation to repeated exposure to drugs of abuse in two major domains. First, glutamate hyperactivity in the basal forebrain has been linked to the hyperexcitability associated with ethanol withdrawal, and this hyperexcitability has been observed in slices of the hippocampus, nucleus accumbens, and amygdala. This hyperexcitability is linked to the protracted abstinence state in alcohol dependence and is hypothesized to be a neural substrate for the anti-relapse effects of acamprosate, a medication for the treatment of alcoholism. Second, glutamate neuroplasticity has been implicated in cocaine-induced reinstatement where increased glutamate release combined with reduced basal glutamate function in the prefrontal cortex-to-nucleus accumbens core pathway has been hypothesized to explain increased glutamate release in response to repeated cocaine administration. In both models, increased activation of glutamatergic function contributes to increased drug seeking in addiction.

Chronic administration of drugs with dependence potential also dysregulates both the hypothalamic–pituitary–adrenal axis and the brain stress system mediated by corticotropin-releasing factor (CRF). Common responses include an activated pituitary adrenal stress response and elevated adrenocorticotropic hormone and corticosteroids and an activated brain stress response with activated amygdala CRF during acute withdrawal from all major drugs of abuse. Acute withdrawal from drugs of abuse may also increase the release of norepinephrine in the BNST and decrease levels of neuropeptide Y (NPY) in the central and medial nuclei of the amygdala (**Figure 1**).

These results suggest not only a change in function of neurotransmitters associated with the acute reinforcing effects of drugs (dopamine, opioid peptides, serotonin, and GABA) during the development of dependence but also recruitment of the brain arousal and stress systems (glutamate, CRF, and norepinephrine) and dysregulation of the NPY brain antistress system. These changes would represent a 'between-system' neuroadaptation. Thus, reward mechanisms in dependence are compromised by disruption of neurochemical systems involved in processing natural rewards and by recruitment of the anti-reward systems that represent neuroadaptation to the chronic exposure of the brain reward neurocircuitry to drugs of abuse.

The neuroanatomical entity termed the extended amygdala thus may represent a common anatomical substrate for acute drug reward and a common neuroanatomical substrate for the negative effects on reward function produced by stress that help drive compulsive drug administration. The extended amygdala receives numerous afferents from limbic structures such as the basolateral amygdala and hippocampus, and it sends efferents to the medial part of the ventral pallidum and a large projection to the lateral hypothalamus, thus further defining the specific brain areas that interface classical limbic (emotional) structures with the extrapyramidal motor system.

Animal models of craving involve the use of drug-primed reinstatement, cue-induced reinstatement, or stress-induced reinstatement in animals that have acquired drug self-administration and then have been subjected to extinction of responding for the drug. Most evidence from animal studies suggests that drug-induced reinstatement is localized to a medial prefrontal cortex/nucleus accumbens/ventral pallidum circuit mediated by the neurotransmitter glutamate. In contrast, neuropharmacological and neurobiological studies using animal models for cue-induced reinstatement involve the basolateral amygdala as a critical substrate with a possible feed-forward mechanism through the prefrontal cortex system involved in drug-induced reinstatement.

Stress-induced reinstatement of drug-related responding in animal models appears to depend on the activation of both CRF and norepinephrine in elements of the extended amygdala (CeA and BNST).

In summary, three neurobiological circuits have been identified that have heuristic value for the study of the neurobiological changes associated with the development and persistence of drug dependence. The acute reinforcing effects of drugs of abuse that comprise the binge/intoxication stage of the addiction cycle most likely involve actions localized to a nucleus accumbens–amygdala reward system, dopamine inputs from the ventral tegmental area, local opioid peptide circuits, and opioid peptide inputs in the arcuate nucleus of the hypothalamus. In contrast, the symptoms of acute withdrawal important for addiction, such as dysphoria and increased anxiety associated with the withdrawal/negative affect stage, most likely involve decreases in function of the extended amygdala reward system and recruitment of brain stress neurocircuitry. The preoccupation/anticipation (or craving) stage involves key afferent projections to the extended amygdala and nucleus accumbens, specifically the prefrontal cortex (for drug-induced reinstatement) and the basolateral amygdala (for cue-induced reinstatement). Compulsive drug-seeking behavior is hypothesized to be driven by ventral striatal–ventral pallidal–thalamic–cortical loops (Figure 1).

Molecular and Cellular Targets within the Brain Circuits Associated with Addiction

Cellular mechanisms within the neurocircuitry of the basal forebrain have been focused on changes in patterns of neuronal firing in freely moving animals, glutamate-dependent plasticity in the nucleus accumbens, and GABAergic plasticity in the CeA. In freely moving animals, extracellular recordings have revealed a neuronal population in the nucleus accumbens that exhibits phasic excitatory responses that are time locked to drug-related events and, with repeated intravenous self-administration, decreases in background firing, suggesting a net enhancement of drug-related signals. This differential inhibition of drug reward-related firing and background firing may represent a form of filtering that could narrow the ensemble of neurons in the nucleus accumbens to those that mediate the strengthening of associations.

Long-term potentiation at excitatory synapses has also been observed in nucleus accumbens and ventral tegmental area slices. Increases in synaptic strength in ventral tegmental area slices have been reported even after a single *in vivo* administration of cocaine. These changes in synaptic strength have been hypothesized to share some of the same mechanisms with long-term potentiation, such as a dependence on glutamatergic neurotransmission.

Cellular mechanisms in the amgydala involving both glutamate and GABA have been identified as the basis for changes in reward system neuroadaptation to excessive alcohol exposure. Acute alcohol markedly decreased both *N*-methyl-D-aspartate (NMDA) and non-NMDA-mediated excitatory postsynaptic currents in CeA neurons from naive control rats. Early (on the order of hours) and long-term (weeks) withdrawal from chronic alcohol treatment significantly increase the sensitivity of CeA NMDA responses to acute alcohol. In contrast, acute ethanol significantly increased evoked $GABA_A$ receptor-mediated inhibitory postsynaptic currents in a majority of neurons in CeA slices via increased GABA release in the CeA. Also, chronic ethanol treatment did not change the effect of acute ethanol in augmenting inhibitory postsynaptic current sizes and the frequency of mini-postsynaptic currents, suggesting a lack of tolerance for the presynaptic effects of ethanol. Studies have also shown a key role for CRF in driving alcohol-induced GABA release in the CeA. Thus, CRF and GABA interactions may represent a cellular substrate underlying the neuroadaptations associated with the development of an anti-reward construct in dependence.

Acknowledging that all drugs of abuse share some common neurocircuitry actions, namely inhibition of medium spiny neurons in the ventral striatum and interneurons in the CeA, the search at the molecular level has led to examining mechanisms for these cellular changes. Repeated perturbation of intracellular signal transduction pathways may cause changes in neuronal function and/or changes in nuclear function and altered rates of transcription of particular target genes. Altered expression of such genes would lead to presumably long-term altered activity of the neurons where such changes occur and ultimately to changes in neural circuits in which those neurons operate.

Two transcription factors in particular have been implicated in the plasticity associated with addiction: cyclic adenosine monophosphate (cAMP) response element binding protein (CREB) and ΔFosB. CREB regulates the transcription of genes that contain a cAMP response element site within the regulatory regions and can be found ubiquitously in genes expressed in the central nervous system such as those encoding neuropeptides, synthetic enzymes for neurotransmitters, signaling proteins, and other transcription factors. CREB can be phosphorylated by protein kinase A and by protein kinases regulated by growth factors, placing it at a point of convergence

for several intracellular messenger pathways that can regulate the expression of genes.

Much work in the addiction field has shown that activation of CREB in the nucleus accumbens, one part of the brain reward circuit, is a consequence of chronic exposure to opiates, cocaine, and alcohol and deactivation in the CeA, another part of the reward circuit. The activation of CREB in the nucleus accumbens with psychostimulant drugs is linked to the symptoms of psychostimulant withdrawal, possibly through induction of the opioid peptide dynorphin which binds to κ opioid receptors and has been hypothesized to represent a mechanism of motivational tolerance and dependence. These molecular adaptations decrease an individual's sensitivity to the rewarding effects of subsequent drug exposures (tolerance) and impair the reward pathway (dependence) so that after removal of the drug the individual is left in an amotivational, dysphoric, or depressed-like state.

In contrast, decreased CREB phosphorylation has been observed in the CeA during alcohol withdrawal and has been linked to decreased NPY function and consequently increased anxiety-like responses associated with acute alcohol withdrawal. These molecular changes may occur simultaneously and point to transduction mechanisms that could produce neurochemical changes in the neurocircuits outlined previously and are critical for breaks with reward homeostasis in addiction.

These transcription factors can change gene expression and produce long-term changes in protein expression and, as a result, neuronal function. Whereas acute administration of drugs of abuse can cause a rapid (within hours) activation of members of the Fos protein family, such as c-*fos*, FosB, Fra-1, and Fra-2 in the nucleus accumbens, other transcription factors, isoforms of ΔFosB, accumulate over longer periods of time (days) with repeated drug administration. Animals with activated ΔFosB have exaggerated sensitivity to the rewarding effects of drugs of abuse, and ΔFosB may be a sustained molecular 'switch' that helps to initiate and maintain a state of addiction.

Genetic and molecular genetic animal models have provided some convergence of data to support the neuropharmacological substrates identified in neurocircuitry studies. High alcohol-preferring rats have been bred that show high voluntary consumption of alcohol, increased anxiety-like responses, and numerous neuropharmacological phenotypes such as decreased dopaminergic activity and decreased NPY activity. In an alcohol-preferring and -nonpreferring cross, a quantitative trait locus was identified on chromosome 4, a region to which the gene for NPY has been mapped. In the inbred preferring and nonpreferring quantitative trait loci analyses, loci on chromosomes 3, 4, and 8 have been identified which correspond to loci near the genes for the dopamine D_2 and serotonin $5HT_{1B}$ receptors.

Advances in molecular biology have led to the ability to systematically inactivate the genes that control the expression of proteins that make up receptors or neurotransmitter/neuromodulators in the central nervous system using the gene knockout and transgenic knockin approaches. Although these approaches do not guarantee that these genes are the ones that convey vulnerability in the human population, they provide viable candidates for exploring the genetic basis of endophenotypes associated with addiction.

Notable positive results with opioids with gene knockout studies in mice have focused on knockout of the μ opioid receptor. Opiate (morphine) reinforcement as measured by conditioned place preference or self-administration is absent in μ knockout mice, and there is no development of somatic signs of dependence to morphine in these mice. Indeed, all morphine effects tested, including analgesia, hyperlocomotion, respiratory depression, and inhibition of gastrointestinal transit, are abolished in μ knockout mice. Knockout of the μ opioid receptor also decreases nicotine reward, cannabinoid reward, and alcohol drinking in mice.

Knockout and neuropharmacological studies have implicated numerous neurotransmitter systems in ethanol preference, including opioid, dopamine, GABA, and serotonin. Novel modulatory effects on ethanol preference have been suggested by protein kinase and G-protein channel knockout studies.

Selective deletion of the genes for expression of different dopamine receptor subtypes and the dopamine transporter has revealed significant effects to challenges with psychomotor stimulants. Dopamine D_1 receptor knockout mice show no response to D_1 agonists or antagonists and show a blunted response to the locomotor-activating effects of cocaine and amphetamine. D_1 knockout mice are also impaired in their acquisition of intravenous cocaine self-administration compared to wild-type mice. D_2 knockout mice have severe motor deficits and blunted responses to psychostimulants and opiates, but the effects on psychostimulant reward are less consistent. Dopamine transporter knockout mice are not only dramatically hyperactive but also show a blunted response to psychostimulants. Although developmental factors must be taken into account for the compensatory effect of deleting any one or a combination of genes, it is clear that D_1 and D_2 receptors and the dopamine transporter play important roles in the actions of psychomotor stimulants.

Brain Imaging Circuits Involved in Human Addiction

Brain imaging studies using positron emission tomography with ligands for measuring oxygen utilization or glucose metabolism or using magnetic resonance imaging techniques are providing dramatic insights into the neurocircuitry changes in the human brain associated with the development and maintenance and even vulnerability to addiction. These imaging results bear a striking resemblance to the neurocircuitry identified by human studies. During acute intoxication with alcohol, nicotine, and cocaine, there is an activation of the orbitofrontal cortex, prefrontal cortex, anterior cingulate, extended amygdala, and ventral striatum. This activation is often accompanied by an increase in availability of the neurotransmitter dopamine. During acute and chronic withdrawal, there is a reversal of these changes with decreases in metabolic activity, particularly in the orbitofrontal cortex, prefrontal cortex, and anterior cingulate, and decreases in basal dopamine activity as measured by decreased D_2 receptors in the ventral striatum and prefrontal cortex. With limited studies, cue-induced reinstatement appears to involve a reactivation of these circuits much like acute intoxication. Two strongly represented markers for active substance dependence in humans across drugs of different neuropharmacological actions are decreases in prefrontal cortex metabolic activity and decreases in brain dopamine D_2 receptors that are hypothesized to reflect decreases in brain dopamine function.

Conclusions

Much progress in neurobiology has provided a heuristic neurocircuitry framework with which to identify the neurobiological and neuroadaptive mechanisms involved in the development of drug addiction. The brain reward system implicated in the development of addiction is composed of key elements of the basal forebrain, including the nucleus accumbens and amygdala and their connections. Neuropharmacological studies in animal models of addiction have provided evidence for the dysregulation of specific neurochemical mechanisms in specific brain reward systems (dopamine, opioid peptides, and GABA). There is also recruitment of brain arousal and stress systems (glutamate, CRF, and norepinephrine) and dysregulation of brain anti-stress systems (NPY) that provide the negative motivational state associated with drug abstinence. Additional neurobiological and neurochemical systems have been implicated in animal models of relapse, with the prefrontal cortex and basolateral amygdala (and glutamate systems therein) being implicated in drug- and cue-induced relapse, respectively. The brain stress systems in the extended amygdala are directly implicated in stress-induced relapse. The changes in reward and stress systems are hypothesized to remain dysregulated and outside of a homeostatic state, and combined with compromised executive function they convey the vulnerability for development of dependence and relapse in addiction. Genetic studies in animals suggest roles for the genes encoding the neurochemical elements involved in the brain reward (dopamine and opioid peptide) and stress (NPY) systems in the vulnerability to addiction, and molecular studies have identified transduction and transcription factors that may mediate the dependence-induced reward dysregulation (CREB) and chronic vulnerability changes (ΔFosB) in neurocircuitry associated with the development and maintenance of addiction. Human imaging studies reveal similar neurocircuits involved in acute intoxication, chronic drug dependence, and vulnerability to relapse.

See also: Drug Addiction: Behavioral Neurophysiology; Drug Addiction: Behavioral Pharmacology of Drug Addiction in Rats; Drug Addiction: Cellular Mechanisms; Drug Addiction: Neuroimaging; Drugs Addiction: Actions; Transcription and Reward Systems.

Further Reading

Contet C, Kieffer BL, and Befort K (2004) Mu opioid receptor: A gateway to drug addiction. *Current Opinion in Neurobiology* 14: 370–378.

Crabbe JC, Phillips TJ, Harris RA, Arends MA, and Koob GF (2006) Alcohol-related genes: Contributions from studies with genetically engineered mice. *Addiction Biology* 11: 195–269.

Everitt BJ and Wolf ME (2002) Psychomotor stimulant addiction: A neural systems perspective. *Journal of Neuroscience* 22: 3312–3320. (Erratum: 22(16), 1a.)

Gardner EL and Vorel SR (1998) Cannabinoid transmission and reward-related events. *Neurobiology of Disease* 5: 502–533.

Goldstein RZ and Volkow ND (2002) Drug addiction and its underlying neurobiological basis: Neuroimaging evidence for the involvement of the frontal cortex. *American Journal of Psychiatry* 159: 1642–1652.

Heimer L and Alheid G (1991) Piecing together the puzzle of basal forebrain anatomy. *Advances in Experimental Medicine and Biology* 295: 1–42.

Heinrichs SC and Koob GF (2004) Corticotropin-releasing factor in brain: A role in activation, arousal, and affect regulation. *Journal of Pharmacology and Experimental Therapeutics* 311: 427–440.

Koob GF, Bartfai T, and Roberts AJ (2001) The use of molecular genetic approaches in the neuropharmacology of corticotropin-releasing factor. *International Journal of Comparative Psychology* 14: 90–110.

Koob GF and Le Moal M (2001) Drug addiction, dysregulation of reward, and allostasis. *Neuropsychopharmacology* 24: 97–129.

Koob GF and Le Moal M (2006) *Neurobiology of Addiction*. London: Academic Press.

McFarland K and Kalivas PW (2001) The circuitry mediating cocaine-induced reinstatement of drug-seeking behavior. *Journal of Neuroscience* 21: 8655–8663.

Nestler EJ (2004) Historical review: Molecular and cellular mechanisms of opiate and cocaine addiction. *Trends in Pharmacological Sciences* 25: 210–218.

Pandey SC (2004) The gene transcription factor cyclic AMP-responsive element binding protein: Role in positive and negative affective states of alcohol addiction. *Pharmacology and Therapeutics* 104: 47–58.

Shalev U, Grimm JW, and Shaham Y (2002) Neurobiology of relapse to heroin and cocaine seeking: A review. *Pharmacological Reviews* 54: 1–42.

Siggins GR, Roberto M, and Nie Z (2005) The tipsy terminal: Presynaptic effects of ethanol. *Pharmacology and Therapeutics* 107: 80–98.

Adenosine

K A Jacobson and Z-G Gao, National Institutes of Health, Bethesda, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The behavioral stimulant effects of the alkylxanthine caffeine and related xanthines are well known. The mechanism of these effects at moderate xanthine doses is now known to be through antagonism of one or more subtypes of adenosine receptors (ARs), rather than through inhibition of phosphodiesterases or stimulation of calcium release, which occur only at higher concentrations of caffeine. At very high doses, caffeine depresses locomotor activity, which does not appear to be related to ARs. Adenosine, an endogenous agonist of the ARs, acts as a local modulator of the action of various neurotransmitters, including biogenic amines and excitatory amino acids. In fact, adenosine modulates the release of many, if not most, neurotransmitters, both in the brain and in the peripheral nervous system.

The ARs activated by extracellular adenosine are classified as four subtypes: A_1, A_{2A}, A_{2B}, and A_3 (Figure 1). A_{2A} and A_{2B} ARs are relatively close in sequence identity (59%, for the human homologs), as are A_1 and A_3 ARs (49%, for the human homologs). Typically, adenosine has an affinity of 10–30 nM at the high-affinity binding sites of the A_1 and A_{2A} ARs. A_{2B} AR has the lowest affinity (high micromolar) of all of the subtypes for adenosine, and its affinity at the A_3 AR is intermediate. The classical second messengers associated with these subtypes are stimulation of adenylate cyclase through G_s proteins for the A_{2A} and A_{2B} subtypes and inhibition of adenylate cyclase through G_i proteins for the A_1 and A_3 subtypes. In recent years, it has become apparent that other effector mechanisms are important in the physiological actions of adenosine, such as phosphoinositide 3-kinase and mitogen-activated protein kinases (MAPKs), which may be activated through the β and γ subunits of the G-proteins.

In general, the role of adenosine in the brain and peripherally is protective: that is, when released or produced in response to organ stress or tissue damage, adenosine tends to increase the ratio of oxygen supply to oxygen demand. It also suppresses various cytotoxic processes, such as cytokine-induced apoptosis. In the brain, both neuronal and glial cell functions are regulated by adenosine. In the autonomic and enteric nervous systems, adenosine has been shown to modulate neurotransmitter release and the interaction of neurotransmitters with receptors. ARs are critical in controlling bladder function via a neurological mechanism. In the intestine, adenosine also plays an essential role in modulating the function of neurons located in the myenteric plexuses between muscle coats and submucous plexus. In the peripheral immune system, adenosine has been shown to 'put the brakes' on excessive inflammation and to promote tissue protection against ischemic damage. In the cardiovascular system, adenosine promotes vasodilation, vascular integrity, and angiogenesis, and counteracts the lethal effects of prolonged ischemia on cardiac myocytes. Nearly every cell type in the body expresses one or more of the subtypes of ARs, indicating the fundamental nature of adenosine as a cytoprotective mediator in both the peripheral and the central nervous systems.

Medicinal chemists have succeeded in designing and synthesizing for the ARs numerous selective and potent ligands that act competitively with adenosine at the orthosteric site of the receptors. Highly selective agonists and antagonists of the A_1, A_{2A}, and A_3 ARs have been reported both as research tools and as experimental therapeutic agents, but for the A_{2B} ARs only selective antagonists are currently known. Some of these selective agents are excluded from the central nervous system when administered peripherally, due to lack of penetration of the intact blood–brain barrier. Thus, these are important pharmacological probes for separating the central and peripheral actions of adenosine. It should be noted that nucleosides, such as adenosine and its analogs, by virtue of the hydrophilic 9-ribose moiety, tend to enter the brain only to a small degree, if at all. Most of the adenosine antagonists are nonnucleoside heterocyclic derivatives, which tend to be more hydrophobic than are the adenosine agonists. Therefore, a given adenosine agonist cannot be assumed to enter the brain in sufficient amounts to exogenously stimulate one of the ARs, unless there is supportive evidence for its action in the central nervous system. The half-life of adenosine in peripheral circulation can be as short as 1 s, and therefore its peripheral administration is unlikely to affect the concentration of extracellular adenosine in the brain. In contrast, a significant number of adenosine antagonists enter the brain when administered peripherally. The simple xanthines, caffeine and theophylline, readily permeate the blood–brain barrier.

Envisioned therapeutic applications of AR agonists and antagonists include treatment of inflammatory bowel diseases, arrhythmia, bladder dysfunction, ischemia, neurodegenerative diseases (such as Parkinson's

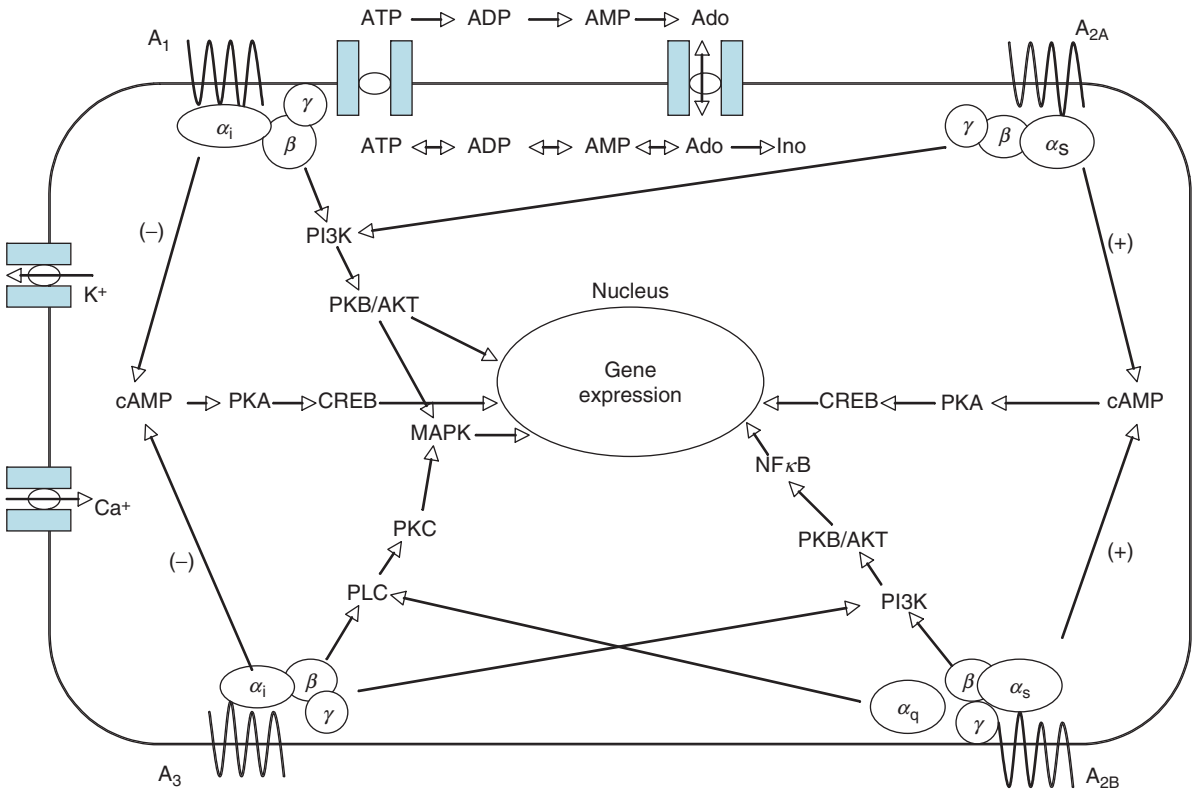

Figure 1 Signaling pathways associated with the four subtypes (A_1, A_{2A}, A_{2B}, and A_3) of adenosine receptors. ATP, adenosine triphosphate; ADP, adenosine diphosphate; AMP, adenosine monophosphate; Adp, adenosine; Ino, inosine; CREB, cyclic AMP response element-binding protein; MAPK, mitogen-activated protein kinase; NFκB, nuclear factor-kappa B; PI3K, phosphatidylinositol 3-kinase; PKA, protein kinase A; PKB, protein kinase B; PKC, protein kinase C; PLC, phospholipase C.

disease and Alzheimer's disease), pain, psychiatric disorders, and sleep disorders. The use of both genetic deletion of the receptor (now each of the subtypes has been deleted in mice strains) and selective agonists/antagonists has contributed toward these concepts.

Sources of Adenosine, Its Transport Mechanisms, and Metabolism

Adenosine production in the synapse is not through vesicular release in response to nerve firing, as is the case for classical neurotransmitters. Rather, adenosine acts as a local autacoid, the release of which increases upon stress to an organ or tissue. Most cells in culture and *in situ* produce adenosine and release it extracellularly, which tends to influence the outcome of pharmacological studies if not properly controlled. The levels of extracellular adenosine in a given tissue or organ may vary widely, depending on stress factors present, leading to highly variable basal levels of stimulation of the ARs by endogenous adenosine.

The source of extracellular adenosine may be both from inside the cell, where it is present in millimolar concentrations, and from the breakdown of extracellular adenine nucleotides, which activate signaling pathways through their own superfamily of receptors (P2Y metabotropic and P2X inotropic receptors). Several nucleotide precursors of adenosine, such as ATP and adenosine diphosphate (ADP), act at P2 receptors, but the monophosphate, AMP, which does not act at any of the P2 receptors, is weakly active as an agonist of ARs. The level of extracellular adenosine may be as low as ~20 nM in the resting brain and as high as 100 μM in severe ischemic conditions in the brain or elsewhere in the body. As a hydrophilic small molecule, adenosine does not diffuse across the plasma membrane freely; rather, it may pass through an equilibrative transporter such as the ENT1 nucleoside transporter. Levels of extracellular adenosine may also rise as a result of the enzymatic hydrolysis of extracellular adenine nucleotides by widely occurring ectonucleotidases or as a result of cell lysis. Ectonucleotidases, which are also ubiquitously expressed on the cell surface but with characteristic distribution patterns, cleave the phosphate moieties of adenine nucleoside 5′-phosphate derivatives, to culminate in the formation of adenosine. There are many classes

of ectonucleotidases; however, the most relevant species in the family of ectonucleoside triphosphate dihydrolases (E-NTPDases) are apyrase (NTPDase1, which converts ATP and ADP to AMP) and NTPDase2 (which converts ATP to ADP). A separate enzyme ecto-5′-nucleotidase (CD73) converts AMP to adenosine. CD73 is characteristically found on the surface of astrocytes but not of neurons.

Unlike classical neurotransmitters, adenosine does not have a rapid synaptic uptake system (as for the biogenic amines), or a rapid chemical inactivation system (as for acetylcholine). Rather, adenosine is metabolized extracellularly by widespread enzymes adenosine kinase (AK; to produce AMP) and adenosine deaminase (to produce inosine) and therefore is inactivated with respect to the ARs. The metabolite inosine may have its own biological actions. For example, there is evidence that inosine activates the A_3 AR.

Inhibition of the metabolism of extracellular adenosine or its uptake proteins is being explored for therapeutic purposes. AK inhibitors have been proposed for the treatment of pain and seizure; however, the promising clinical development of these efficacious compounds has been discontinued due to their toxicity.

Receptor Structure, Signaling Pathways, and AR Regulation

Receptor Structure

All of the ARs are G-protein-coupled receptors (GPCRs), which structurally consist of seven transmembrane helices connected by extracellular and cytoplasmic loops. There is a strong sequence homology between A_1 and A_3 ARs and between A_{2A} and A_{2B} ARs. The neurotransmitter receptors that are closest in primary sequence to the ARs are the biogenic amine receptors. Each of the ARs has been modeled both in overall three-dimensional structure and in the nature of molecular recognition in the putative ligand binding site, based on homology to bovine rhodopsin or by other methods, such as the recently developed threading assembly refinement method. Essential residues for the binding of agonists and antagonists and for the activation of the receptors have been defined through modeling and mutagenesis. Use of the molecular models to design *de novo* new ligands has not yet met with extensive success. However, confidence in the receptor models and in their predictive abilities has increased gradually since the first models were reported in the early 1990s. The neoceptor approach to reengineering GPCRs in general, and ARs specifically, has validated binding site hypotheses for adenosine at its receptors.

The binding and activation steps of receptor action have been dissected computationally, although not yet in a global fashion, largely due to the lack of a suitable template for the activated conformation. The conformational dynamics of the activation of the A_3 AR have been approximated with respect to isolated portions of the receptor.

Signaling Pathways

Two AR subtypes, A_1 and A_3, couple through G_i to the inhibition of adenylate cyclase, while the other two subtypes, A_{2A} and A_{2B}, stimulate adenylate cyclase through G_s or G_{olf} (for A_{2A}). The A_{2B} AR is also coupled to activation of phospholipase C through G_q. Furthermore, each of these receptors may couple through the β and γ subunits of the G-proteins to other effector systems, including ion channels and phospholipases. ARs have been found to couple to MAPKs in a variety of circumstances, leading to effects on differentiation, proliferation, and cell death. The A_3 AR can reduce apoptosis through the activation of Akt. ARs can couple to ion channels; for example, activation of the A_1 AR can induce the influx of calcium ions or the efflux of potassium ions.

Cross-talk occurs between ARs and other receptors. For example, an otherwise subthreshold concentration of acetylcholine (ACh), as might be present in the Alzheimer's brain, still produces a strong calcium signal when the A_1 AR is costimulated. Cross-talk occurs with the striatal dopamine receptor system, in which a direct physical association (dimerization) occurs between A_{2A} and D_2 receptors, and between other subtypes. A_{2A} AR stimulation also transactivates the neuroprotective Trk neurotrophin receptor pathway, even in the absence of neurotrophic growth factor. Recently, heterodimers of adenosine A_1 receptors and either $P2Y_1$ or $P2Y_2$ nucleotide receptors have been characterized pharmacologically.

Regulation

Similar to the function and regulation of other GPCRs, both activation and desensitization of the ARs occur after agonist binding. Interaction of the activated ARs with the G-proteins leads to second-messenger generation and classical physiological responses. Interaction of the activated ARs with G-protein receptor-coupled kinases (GRKs) leads to their phosphorylation. Downregulation of ARs should be considered in both the basic pharmacological studies and with respect to the possible therapeutic application of agonists. AR responses desensitize rapidly, and this phenomenon is associated with receptor downregulation, internalization, and degradation. Mutagenesis has been applied to analyze the molecular basis for the differences in the kinetics

of the desensitization response displayed by various AR subtypes. The most rapid downregulation among the AR subtypes is generally seen with the A_3 AR, due to phosphorylation by GRKs.

Novel and Definitive Ligands

Potent and selective AR antagonists have been prepared for all four AR subtypes, and selective agonists are known for three subtypes. Thus, numerous pharmacological tools are available for *in vitro* and *in vivo* use. Potent and selective A_{2B} AR agonists are yet to be reported, although several research groups have identified lead compounds. Molecular modeling of the ARs and ligand docking have provided insights into the putative binding sites of all of the subtypes, which has aided in ligand design.

Adenosine Agonists

Until recently, nearly all AR agonists have been purine nucleoside derivatives. With the synthesis of pyridine-3,5-dicarbonitrile derivatives, there is now an example of a nonnucleoside chemical class that fully activates ARs. Medicinal chemists have extensively explored the structure–activity relationships of adenosine derivatives as agonists of the ARs (Figure 2). In general, for the adenine moiety of adenosine, modifications at the N^6 position have led to selectivity for the A_1 AR and the A_3 AR, and modifications at the 2 position, especially with ethers, secondary amines, and alkynes, have led to selectivity for the A_{2A} AR. Commonly used A_1 AR agonists that are N^6-cycloalkyl derivatives are cyclopentyladenosine (CPA), its 2-chloro analog CCPA, and endo-norbornyl adenosine (*S*(−)-ENBA) (Figure 2(a)). At the human ARs, *S*(−)-ENBA is more highly A_1 AR-selective than are CPA and CCPA. Adenosine amine congener (ADAC) is an amine-functionalized congener designed for covalent coupling to carriers with the retention of A_1 AR activity. The A_{2A} AR-enhancing effects of substitution at the 2 position are further boosted with a uronamido substitution at the 5′ position of adenosine (Figure 2(b)). For example, the 5′-*N*-ethyl derivatives CGS21680 and ATL-146e are both selective in binding to the rat A_{2A} AR and less selective at the human subtypes. When the 5′-*N*-alkyluronamide group alone is present, high affinity at the A_{2A} AR, but not selectivity, is typically observed, as with the nonselective agonist 5'-*N*-ethylcarboxamide adenosine (NECA). The N^6-substituted A_1 AR agonists NNC-21-0136 and GR79236 and the doubly substituted selodenoson have been clinical candidates. A_1 AR agonists are of interest for use in treating cardiac arrhythmias (for which adenosine itself, under the name Adenocard, is in widespread use). AMP579 is a carbocyclic nucleoside that activates both A_{2A} and A_1 ARs. A_{2A} AR agonists such as ATL-146e, binodeson, CVT-3146, and MRE0094 have been clinical candidates. Such agonists are of interest for use as vasodilatory agents in cardiac imaging and in suppressing inflammation.

Numerous exceptions to the preceding generalizations concerning selectivity of specific substitution sites on adenosine have been found. For example, the mono-N^6-substituted derivative of adenosine, N^6-[2-(3,5-dimethoxyphenyl)-2-(2-methylphenyl)-ethyl] adenosine (DPMA), is potent and selective at the rat A_{2A} AR. At the human ARs, the A_{2A} AR and A_1 AR affinities of DPMA are roughly equal. N^6-Arylmethyl derivatives of adenosine, such as N^6-benzyl derivatives, tend toward selectivity for the A_3 AR. MRS3997, an adenosine derivative that is monosubstituted at the 2 position with a (6-bromotryptophol)ether moiety, is a promising lead for the A_{2B} AR. It is to be noted that MRS3997 fully activates human A_{2A} and A_{2B} ARs equipotently.

The A_3 agonists (Figure 2(c)) IB-MECA and Cl-IB-MECA are widely used as selective agonists of the A_3 AR, although yet more selective agents are now known (e.g., the conformationally locked (*N*) methanocarba derivative MRS3558 and the 4′-thio nucleoside LJ568). CP-608039 is a selective A_3 agonist that was developed for cardioprotection.

Adenosine Antagonists

The structures of A_3 antagonists are typically more chemically diverse than are the classical xanthine antagonists of the A_1 and A_2 receptors (Figure 3). The dihydropyridine derivatives MRS1191 and MRS1334 and the pyridylquinazoline derivative VUF5574 are potent, selective A_3 AR antagonists in the human, but are weak or inactive at the rat A_3 AR. Nevertheless, MRS1191 has been used successfully in murine species. MRS1220 is very potent at the human but not at the rat or the mouse receptor. The pyridine derivative MRS1523 and the nucleosides MRS1292, however, are selective A_3 AR antagonists in both rats and humans. Selective A_3 AR antagonists, such as the aforementioned compounds and the heterocyclic derivatives OT-7999 and MRE-3008-F20 and the adenine derivative MRS3777, are in general of interest for the treatment of cancer, stroke, inflammation, and glaucoma.

Variations in the relative efficacy of nucleosides, depending on structure, have been noted. This phenomenon is especially pronounced for the A_3 AR, at which changes on the adenine moiety (N^6 and 2 positions) and ribose moiety can either reduce efficacy to the point of pure antagonism (i.e., combination of 2-Cl and N^6-(3-iodobenzyl)) or can guarantee

a robust, nearly full activation of the A_3 AR (i.e., 5′-uronamide).

Although the classical AR antagonists are xanthines derivatives such as caffeine and theophylline, today an enormous diversity of heterocyclic structures have been reported as AR antagonists. The micromolar affinity of the naturally occurring antagonists has been greatly exceeded with the introduction of selective

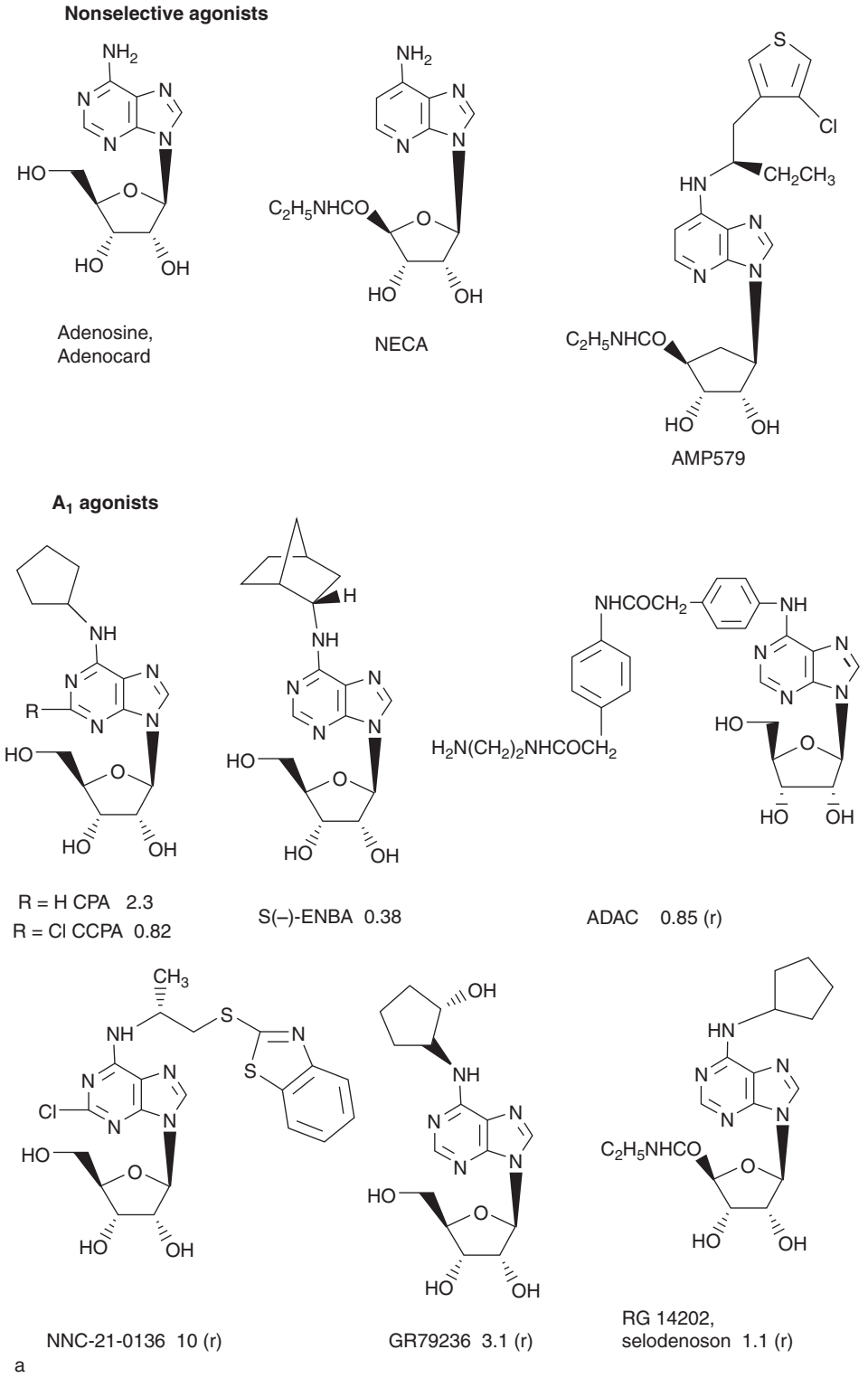

Figure 2 Continued

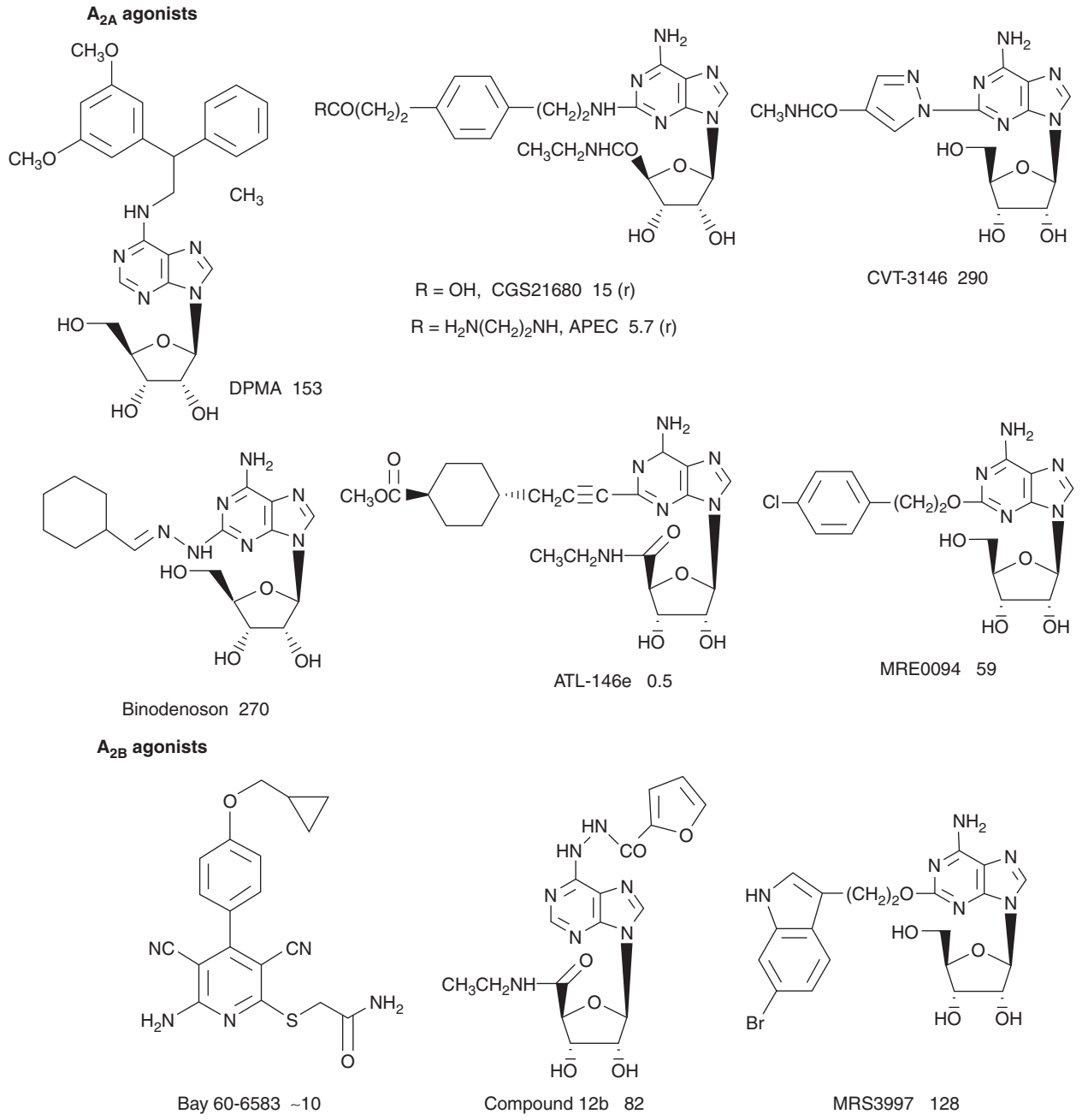

b

Figure 2 Continued

antagonists of subnanomolar affinity. High selectivity of xanthines at the A_1 (e.g., the epoxide derivative BG 9719 and the more selective BG9928; KW3902), A_{2A} (e.g., KW6002), and A_{2B} (e.g., MRS1706 and MRS1754) ARs has been achieved (Figure 3). Although xanthines that are selective for the A_3 AR are not available, certain cyclized derivatives of xanthines, such as PSB-11, are A_3 AR selective. In general, modifications of the xanthine scaffold at the 8 position with aryl or cycloalkyl groups has led to selectivity for the A_1 AR, and modifications at the same position with alkenes (specifically, styryl groups) has led to selectivity for the A_{2A} AR. 8-Cyclopentyl-1, 3-dipropylxanthine (DPCPX) is highly A_1 AR selective in the rat but less A_1 AR selective among the human AR subtypes. The 8-styrylxanthine derivatives KW6002, CSC, and MSX are moderately potent A_{2A} AR antagonists. Some 8-styrylxanthine derivatives,

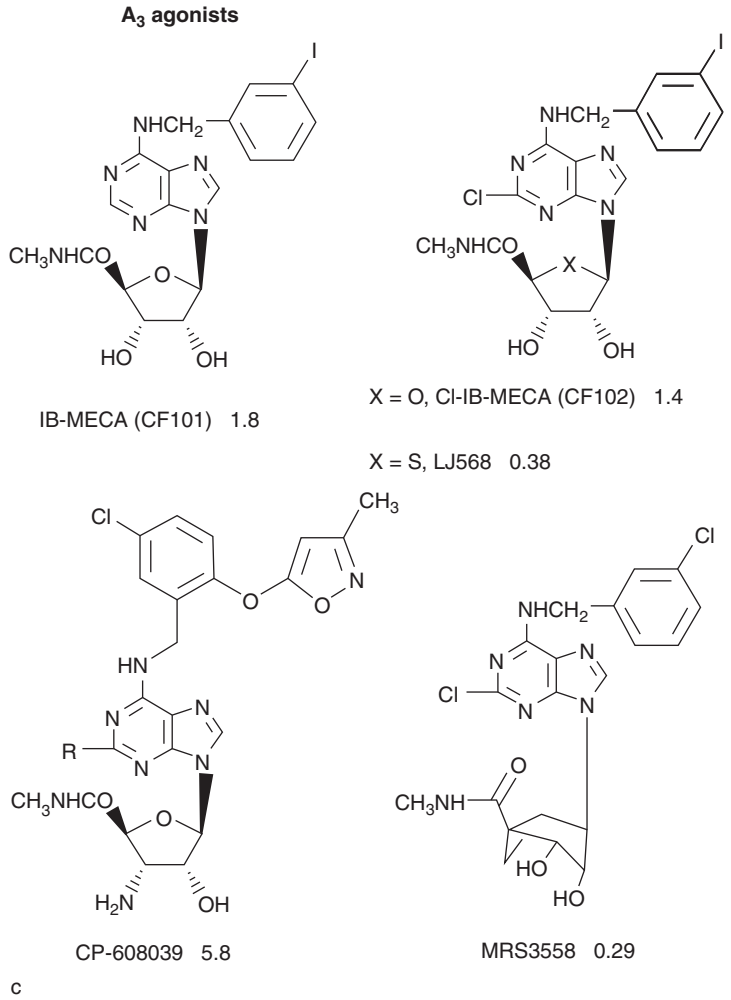

Figure 2 Structures of (a) nonselective and A_1, (b) A_{2A} and A_{2B}, and (c) A_3 adenosine receptor agonist probes used as pharmacological tools, and in some cases as clinical candidates. The K_i values (nM) in binding to the appropriate human adenosine receptor are indicated after the name (all are human adenosine receptors, unless indicated 'R' for rat). Values indicated for the A_{2B} adenosine receptor are functional EC_{50} values. Bay 60-6583 is a nonnucleoside that potently and selectively activates the A_{2B} adenosine receptor. MRS3997 is a mixed agonist for the A_{2A} and A_{2B} adenosine receptors.

especially CSC, have been discovered to inhibit monoamine oxidase-B as well. A persistent problem in the use of xanthine derivatives as AR antagonists of the A_1 AR is their interaction at the A_{2B} AR. Use of adenine derivatives, such as the inverse agonist WRC-0571, provides A_1 AR-selective antagonists having low affinity at the A_{2B} AR.

The triazolotriazine ZM241,385 and the pyrazolotriazolopyrimidines SCH-58261 and SCH442,416 are highly potent and selective A_{2A} AR antagonists. SCH442,416 displays >23 000-fold selectivity for the human A_{2A} AR (K_i, 0.048 nM), in comparison to human A_1 AR, and an $IC_{50} > 10\,\mu M$ at the A_{2B} and A_3 ARs. A_{2A} AR antagonists, such as the xanthine KW6002 and the nonxanthines SCH58261, SCH442416, VER6947, and VER7835, are of interest for use in treating Parkinson's disease.

There is a marked species dependence of antagonist affinity at the A_3 AR. Therefore, commonly used antagonists must be treated with caution in species other than humans. In general, one must be cognizant of potential species differences for both AR agonists and antagonists.

Radioligands

Radioligands commonly used for the ARs are A_1 agonist [^{3}H]CCPA, antagonist [^{3}H]DPCPX, A_{2A} agonist [^{3}H]CGS21680, antagonist [^{3}H]ZM241,385 or [^{3}H]SCH58261, A_3 agonist [^{125}I]I-AB-MECA, and antagonist [^{3}H]PSB-11. Ligands for *in vivo* positron emission tomographic (PET) imaging of A_1 and A_{2A} ARs have been developed. For example, the xanthine [^{18}F]CPFPX and the nonxanthine [^{11}C]FR194921 have been developed as centrally active PET tracers

for imaging the A_1 AR in the brain. Potent fluorescent ligands have been reported for A_1, A_{2A}, and A_3 ARs.

Allosteric Modulation

In addition to directly acting AR agonists and antagonists, allosteric modulators of agonist action are under consideration for disease treatment. Such modulators, either positive enhancers or negative allosteric inhibitors, might have advantages over the directly acting (orthosteric) receptor ligands. The action of the allosteric compounds would depend on the presence of a high local concentration of

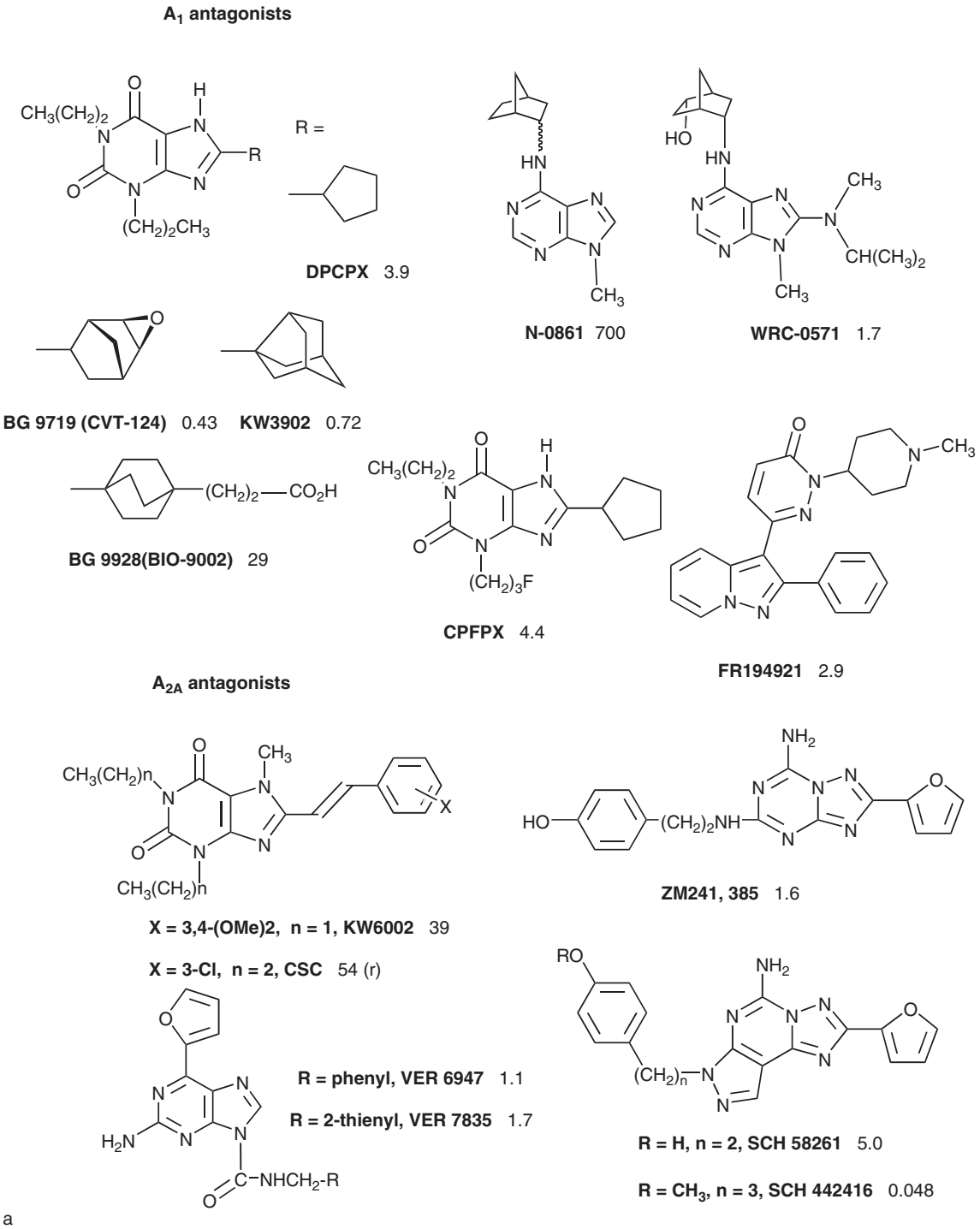

a

Figure 3 Continued

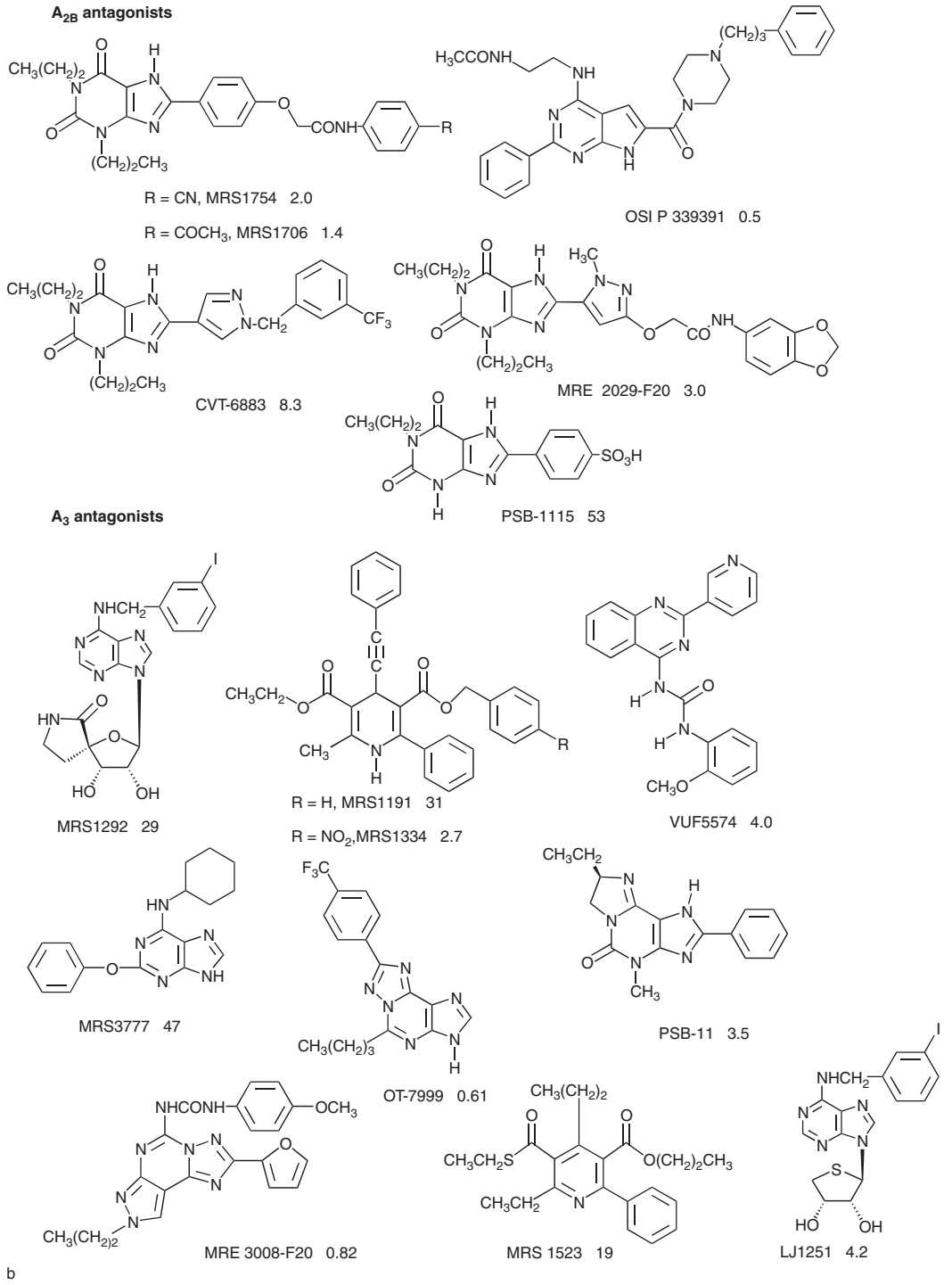

Figure 3 Structures of selected (a) A_1 and A_{2A} and (b) A_{2B} and A_3 adenosine receptor antagonist probes used as pharmacological tools, and in some cases as clinical candidates. The K_i values (nM) in binding to the appropriate human adenosine receptor are indicated after the name (all are human adenosine receptors, unless indicated 'R' for rat).

adenosine, which often occurs in response to a pathological condition. In some cases (dependent on tissue, receptor subtype, and other conditions), one would wish to boost the adenosine effect, and therefore an allosteric enhancer would be useful; in other cases, the elevated adenosine may be detrimental, in which case one would want to apply a negative modulator. Positive allosteric modulators have been explored for the A_1 (benzoylthiophene derivatives) and A_3 (imidazoquinoline derivatives) AR subtypes.

Role of ARs in Autonomic Nervous System Disorders

Distribution

ARs are widely distributed in the autonomic and enteric nervous systems. Distribution of neural ARs in the human intestine has been investigated. Messenger RNAs (mRNAs) of subtypes of AR are differentially expressed in neural and nonneural layers of the jejunum, ileum, colon, and cecum. The A_1 AR is expressed in jejunal myenteric neurons and colonic submucosal neurons. The A_{2A} AR is also found in other neurons, but A_{2B} AR immunoreactivity is more prominent than that of the A_{2A} AR in myenteric neurons, nerve fibers, and glia. The A_3 AR largely occurs in substance P-positive jejunal submucosal neurons and less in vasoactive intestinal peptide (VIP) neurons.

The AR that mediates the relaxation of bladder strips induced by AR agonists such as CGS21680 and NECA has been classified as A_{2A}; however, the highest mRNA levels are found for A_{2B} transcripts in the bladder. However, reports of tissue distribution based on mRNA levels might not correspond to protein levels. Thus, quantitative determination with a potent and selective A_{2B} AR radioligand may be necessary. Western blot analysis shows that all four ARs are expressed in the uroepithelium. A_1 ARs are prominently localized to the apical membrane of the umbrella cell layer, whereas A_{2A}, A_{2B}, and A_3 ARs are localized intracellularly or on the basolateral membrane of umbrella cells and the plasma membrane of the underlying cell layers.

Various subtypes of ARs have also been detected in the heart, blood vessels, kidney, and other organs throughout the body. A_1, A_{2A}, and A_{2B} ARs are all expressed on normal human airway smooth muscle cells, and both A_1 and A_{2A} ARs are expressed in vagal pulmonary C fibers. A_1 and A_{2A} ARs are highly expressed in gastric mucosa. In the brain, the A_1 AR is widely expressed in almost all areas. Pre- and postsynaptic activation of the A_1 AR inhibits synaptic transmission, in part by suppressing the release of excitatory transmitters. The A_{2A} AR is less widely expressed, with greatest density in the striatum, nucleus accumbens, and olfactory tubercle. The A_{2B} and A_3 ARs are expressed at low density in most brain regions, and are implicated in purinergic signaling in neuronal–glial interactions.

Functions of ARs in the Autonomic and Enteric Systems

The enteric nervous system contains several hundred million neurons located in the myenteric plexuses between muscle coats and submucous plexus. ARs in the enteric nervous system are critical for the control of motor and secretomotor functions. Adenosine is known to suppress intestinal motility by activating putative neural A_1 ARs in the small intestine. A_{2A} and A_{2B} ARs in the myenteric neurons were also suggested to contribute to effects of adenosine on motility. ARs in circular muscle may contribute to the postjunctional actions of adenosine on motility. Adenosine directly modulates intestinal tone in the rat by causing relaxation via A_{2B} ARs or contraction via A_1 ARs in longitudinal muscle cells. Electrophysiological studies in rodents provided evidence for pre- and postsynaptic A_1 AR-mediated inhibition of slow synaptic transmission and presynaptic inhibition of fast synaptic transmission. A_{2B} AR gene expression products are widely expressed in the mucosa of the human intestinal tract, where they are postulated to be involved in the pathophysiology of diarrheal diseases. Both A_{2A} and A_3 ARs have been found in mucosal tissues, suggesting that they influence secretion and/or absorption in the human intestine. The ARs are involved in neuroplastic changes occurring in inflamed gut. The A_{2A} AR modulates the activity of colonic excitatory cholinergic nerves via facilitatory control on inhibitory nitrergic pathways, and such a regulatory function is enhanced in the presence of bowel inflammation.

Activation of the A_{2A} AR inhibits stress-induced gastric inflammation and damage. Thus, selective A_{2A} AR agonists may be useful for preventing ulcers and gastric inflammation. Both A_1 and A_{2A} ARs are involved in gastrin release. The modification of AR expression by changes in intraluminal acidity may represent a novel regulatory feedback mechanism to control gastric acid secretion.

The discharge of urine from bladder is controlled by the nervous system. The causes of bladder dysfunction related to the nervous system include multiple sclerosis, traumatic or developmental brain or spinal cord injury, or Parkinson's disease. Although it is generally agreed that ACh acting on smooth muscle muscarinic receptors is the primary neurologic mechanism controlling bladder emptying, neural stimulation of the bladder is only partially inhibited in many cases by the muscarinic receptor antagonist atropine. The atropine-resistant component of parasympathetic contraction was later found to be ATP sensitive. There is plenty of evidence to suggest that ARs also play an important role in bladder function. It has been shown that adenosine-evoked membrane hyperpolarization and relaxation of bladder smooth muscle is mediated by A_{2A} AR-mediated activation of K_{ATP} channels via adenylate cyclase and elevation

of cAMP. High mRNA levels of A_{2B} transcript are also found in the bladder.

Adenosine was found to be released from the uroepithelium, which was potentiated tenfold by stretching the tissue. It is generally accepted that the sensation of bladder fullness is relayed through the mucosal layer by afferent nerves, which are activated by the release of neurotransmitters such as ATP from the urothelium when it is stretched as the bladder fills. It is suggested that adenosine reduces the force of nerve-mediated contractions by acting predominantly at the presynaptic sites at the nerve muscle junction via the A_1 AR.

The autonomic nervous system, a complex and self-organized entity, plays a key role in regulating cardiovascular function. In the heart, a number of intrinsic nerves in the atrial and intra-atrial septum have been shown to release ATP, ACh, 5-hydroxytryptamine, and other neurotransmitters. Adenosine may regulate both the release and the interaction with their receptors of these neurotransmitters. Adenosine-mediated myocardial protection has been suggested to be through a neurogenic pathway. The infarct-reducing effect of intravenous adenosine in intact rats was blocked with both the ganglionic blocker hexamethonium and the nitric oxide synthase inhibitor N^{ω}-nitro-L-arginine. The A_1 AR was determined to be the primary subtype involved in the modulation of norepinephrine release from cardiac nerve terminals using isolated rat hearts.

Vascular tone in blood vessels is controlled by perivascular nerves and the endothelial cells. Adenosine acts directly on smooth muscles as well as modulates the release of neurotransmitters, such as ACh and ATP. A study of neurotransmitter release following electrical depolarization of nerve endings from the rat mesenteric artery suggests that activation of the presynaptic A_{2A} AR and A_3 AR modulates neurotransmission by inhibiting the release of norepinephrine but not neuropeptide Y. The A_1 ARs, not P2 receptors, inhibit prejunctionally sympathetic neurotransmission in the hamster mesenteric arterial bed. Activation of both A_1 AR and A_{2A} AR is required to attenuate neurogenic coronary constriction due to sympathetic stimulation.

In the respiratory system, adenosine has been shown to stimulate vagal pulmonary C fiber terminals through activation of the A_1 AR, and subsequently cause

Table 1 Exploration of the role of adenosine receptors in disorders of the nervous system

Condition/system	*Model*[a]	*Subtypes implicated*	*Relationship*
Aggression	A_1 KO	A_1	Increased aggressiveness
	A_{2A} KO	A_{2A}	Increased aggressiveness
Alzheimers's disease	β-Amyloid	A_{2A}	Reduced neurotoxicity by antagonists
Anxiety	A_{2A} KO	A_{2A}	Increased anxiety
Cardiac arrhythmias	Various	A_1	Agonist protects
Depression	Swim	A_{2A}	Antagonist improves
	A_3 KO	A_3	Increased despair-like behavior
Epilepsy	Various	A_1	Agonist protects in some models
Huntington's disease	3-NP	A_1	Agonist protects striatal damage
	3-NP	A_{2A}	Antagonist protects striatal damage
Memory	A_1 KO	A_1	No change
Neurodegeneration	Ischemia	A_1	Agonist protects
	Ischemia	A_{2A}	Antagonist protects
	Ischemia	A_3	Chronic agonist protects
	A_1 KO	A_1	No effect in adults, benefits newborns
	A_{2A} KO	A_{2A}	Detrimental in newborns, benefits adults
	A_3 KO	A_3	Detrimental in adults
Pain	Constriction injury	A_1	Agonist protects
	Formalin	A_{2A}	Agonist protects
	A_1 KO	A_1	Increased thermal nociception
	A_{2A} KO	A_{2A}	Lowers thermal nociception
Parkinson's disease	6-OHDA	A_{2A}	Antagonist protects
	MPTP	A_{2A}	Antagonist protects
	Haloperidol	A_{2A}	Decreased catalepsy by antagonists
Pulmonary inflammation	ADA KO mice	A_{2B}	Antagonist protects
Sleep disorders	A_{2A} KO	A_{2A}	Agonist lost effect

[a]ADA, adenosine deaminase; KO, knockout; 6-OHDA, 6-hydroxydopamine; MPTP, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine; 3-NP, 3-nitropropionic acid.

bronchoconstriction, which was significantly attenuated by A_1 AR antagonists. Recent evidence suggests that both A_1 AR and A_{2A} AR are involved in the activation of vagal sensory C fibers in the lung. The A_1 AR has also been suggested to be directly involved in the mobilization of calcium in human bronchial smooth muscle cells. The A_{2B} AR is known to mediate mast cell degranulation in large animals and humans, and A_{2B} AR antagonists are of potential clinical applications for asthma.

A_1 AR agonists reduce pain signaling in the spinal cord, where the receptors are highly expressed. In humans, infusion of adenosine in the spinal cord is effective in decreasing postoperative pain. Recent studies suggest that A_1 ARs might be more important in chronic pain than in acute pain. The A_1 AR agonists are being evaluated in phase II clinical trials for the treatment of pain and migraine. An A_1 AR-selective allosteric enhancer is also used in clinical trials as a treatment for neuropathic pain. Peripherally administered A_{2A} AR agonists have an antinociceptive effect.

In summary, in the peripheral nervous systems, ARs are related to gut inflammation, rheumatoid arthritis, ischemia, Crohn's disease, constipation, bladder disorders, and a variety of other conditions. In the brain, ARs appear to regulate important functions in cerebral ischemia, dementia, Parkinson's disease and other neurodegenerative diseases, pain, sleep disorders, anxiety, and schizophrenia (Table 1).

Genetic Deletion of ARs

Deletion of each of the four subtypes has been carried out, and the resulting single-AR knockout (KO) mice are viable and not highly impaired in function. The pharmacological profile indicates that the analgesic effect of adenosine is mediated by the A_1 AR, and analgesia is lost in mice in which the A_1 AR has been genetically eliminated. Genetic KO of the A_1 AR in mice removes the discriminative-stimulus effects, but not the arousal effect, of caffeine, and increases anxiety and hyperalgesia. Study of A_{2A} AR KO mice reveals functional interaction between the spinal opioid receptors and peripheral ARs. A_1 AR KO mice demonstrate a decreased thermal pain threshold, whereas A_{2A} AR null mice demonstrate an increased threshold to noxious heat stimulation, supporting an A_1 AR-mediated inhibitory and an A_{2A} AR-mediated excitatory effect on pain transduction pathways. KO of the A_{2A} AR eliminates the arousal effect of caffeine. Genetic KO of the A_{2A} AR also suggests a link to increased anxiety and protects against damaging effects of ischemia and the striatal toxin 3-nitropropionic acid. Genetic KO of the A_3 AR leads to increased neuronal damage in a model of carbon monoxide-induced brain injury. Neutrophils lacking A_3 ARs show correct directionality but diminished speed of chemotaxis. Although studies on A_{2B} ARs KO mice have been reported, the importance of A_{2B} ARs in the brain still awaits future investigation.

In conclusion, adenosine is a ubiquitous neuromodulator of many functions, by activating one or more of the widely distributed AR subtypes. Selective AR agonists and antagonists have been developed, some of which are in advance stages of clinical trials for therapeutic applications. The use of both knockout animals and selective drugs has contributed toward elucidation of the physiological role of adenosine and the signaling pathways involved.

See also: Adenosine Receptor Mediated Functions; Autonomic Disorders; Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Nervous System; Neuromodulation; Pharmacology of Sleep: Adenosine; Stimulant and Wake-Promoting Substances.

Further Reading

Burnstock G (2007) Physiology and pathophysiology of purinergic neurotransmission. *Physiological Reviews* 87: 659–797.

Christofi FL, Zhang H, Yu JG, et al. (2001) Differential gene expression of adenosine A1, A2a, A2b, and A3 receptors in the human enteric nervous system. *Journal of Comparative Neurology* 439: 46–64.

Costanzi S, Ivanov AA, Tikhonova IG, et al. (2007) Structure and function of G protein-coupled receptors studied using sequence analysis, molecular modelling, and receptor engineering: Adenosine receptors. In: *Frontiers in Drug Design and Discovery,* vol. 3. Hilversum, The Netherlands: Bentham Science Publishers, Inc.

Fields RD and Burnstock G (2006) Purinergic signalling in neuron-glia interactions. *Nature Reviews Neuroscience* 7: 423–436.

Franco R, Casado V, Mallol J, et al. (2006) The two-state dimer receptor model: A general model for receptor dimers. *Molecular Pharmacology* 69: 1905–1912.

Fredholm BB, IJzerman AP, Jacobson KA, et al. (2001) International Union of Pharmacology. XXV. Nomenclature and classification of adenosine receptors. *Pharmacological Reviews* 53: 527–552.

Fredholm BB, Chen JF, Masino SA, et al. (2005) Actions of adenosine at its receptors in the CNS: Insights from knockouts and drugs. *Annual Review of Pharmacology and Toxicology* 45: 385–412.

Jacobson KA and Gao ZG (2006) Adenosine receptors as therapeutic targets. *Nature Reviews Drug Discovery* 5: 247–264.

Linden J (2005) Adenosine in tissue protection and tissue regeneration. *Molecular Pharmacology* 67: 1385–1387.

McGaraughty S, Cowart M, Jarvis MF, et al. (2005) Anticonvulsant and antinociceptive actions of novel adenosine kinase inhibitors. *Current Topics in Medicinal Chemistry* 5: 43–58.

Moro S, Gao ZG, Jacobson KA, et al. (2006) Progress in pursuit of therapeutic adenosine receptor antagonists. *Medicinal Research Reviews* 26: 131–159.

Ruggieri MR Sr. (2006) Mechanisms of disease: role of purinergic signaling in the pathophysiology of bladder dysfunction. *Nature Clinical Practice Urology* 3: 206–215.

Schulte G and Fredholm BB (2003) Signalling from adenosine receptors to mitogen-activated protein kinases. *Cell Signaling* 15: 813–827.

Yan L, Burbiel JC, Maass A, et al. (2003) Adenosine receptor agonists: From basic medicinal chemistry to clinical development. *Expert Opinion in Emerging Drugs* 8: 537–576.

Yu L, Huang ZL, Mariani J, et al. (2004) Selective inactivation or reconstitution of adenosine A_{2A} receptors in bone marrow cells reveals their significant contribution to the development of ischemic brain injury. *Nature Medicine* 10: 1081–1087.

Adenosine Receptor Mediated Functions

R W Greene, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The purine adenosine (Ado) forms the backbone of adenosine triphosphate (ATP), the cell's primary energy currency, and accordingly, is well situated to act as a sensor for a cell's metabolic state. Both neurons and cardiac cells express membrane receptors for adenosine as well as for equilibrative adenosine transporters, so that changes in intracellular adenosine metabolism may be reflected extracellularly, and this in turn may exert a paracrine-like effect through activation of Ado receptors.

Extracellular Ado concentration is increased by metabolic stressors that increase the ratio of metabolite demand to metabolite availability (Figure 1). Increased electrical activity due to electrical stimulation, increased extracellular potassium, or increased extracellular glutamate (or related agonists) all increase metabolite demand and extracellular Ado. A reduction of oxygen (hypoxia or ischemia), a reduction of glucose, or inhibition of the Kreb's cycle all reduce metabolite availability and increase Ado.

In the mammalian central nervous system, the most prevalently expressed membrane surface Ado receptor (the A1AdoR) exerts an inhibitory effect on excitability. Taken together with the extracellular Ado correlation with metabolism, the primary components of a negative feedback system to maintain the homeostasis between the metabolic state of a neuron (or cardiac cell) and its electrical excitability are in place. Thus, in response to a metabolic stress of local cells, whenever there is an increase of metabolite demand relative to metabolite availability, Ado may be released into the extracellular environment to act on A1AdoRs and inhibit the excitability of these same cells. The reduced excitability results in a reduced metabolic demand, restoring normal metabolic state. This Ado-mediated homeostasis has profound effects on synaptic function, and accordingly on circuit systems and ultimately on behavioral levels, as discussed in the following sections.

Ado Metabolism and Extracellular Ado Concentration

Extracellular Ado derives from a number of sources. ATP is released from synaptic vesicles from both presynaptic terminals and from glia in an activity-dependent manner (Figure 2). The ATP is rapidly broken down by ectonucleotidases to form Ado in the extracellular medium, where it has been shown to both reduce postsynaptic excitability and inhibit presynaptic release of neurotransmitter. Other sources include cyclic adenosine monophosphate (cAMP) that may be directly transported from the intracellular to extracellular space and, especially, intracellular Ado. Intracellular Ado is maintained in equilibrium with extracellular Ado by equilibrative, facilitated transport. A slowing of intracellular Ado metabolism can increase extracellular Ado by direct transport of Ado to the outside in extreme or pathological conditions, when intracellular concentration exceeds that on the outside (for either neurons or glia or both), and in less-extreme or in physiological conditions, by a reduction of the flux from the extracellular to intracellular space.

Under physiological conditions, intracellular Ado concentration is maintained primarily by the low-capacity (but high-affinity) enzyme adenosine kinase (AK). This enzyme requires the binding of Ado, AMP, ADP, and ATP as it phosphorylates Ado to AMP by the dephosphorylation of ATP to ADP. This enzyme maintains a ratio of about 10 000:1 between ATP, ADP, AMP, and Ado, consistent with an extremely small change in ATP being reflected as a large percentage change in Ado that will in turn be reflected extracellularly. In addition, this enzyme undergoes substrate inhibition, suggesting that as the intracellular concentration of Ado increases, AK will be inhibited, leading to a still more rapid increase in concentration. This mechanism provides a threshold for Ado production rate that, once exceeded, would be expected to greatly reduce the excitability of the neuronal tissue, within the locale of the production. Finally, at a high intracellular concentration of Ado, as might be expected during ischemia, hypoglycemia, or epileptic activity, the capacity of AK is exceeded and low-affinity, high-capacity pathways become active to catabolize Ado, including the enzymes adenosine deaminase and *S*-adenosine methionine hydrolase.

Adenosine Receptor Activation

Extracellular Ado modulates neuronal excitability through activation of the A1 Ado receptor (A1R) or by activation of the A2a receptor. The expression of the A2a receptor is limited primarily to the striatum, including the nucleus accumbens, while the expression of the A1R is widespread throughout the central nervous system (CNS). Two other Ado receptors – the

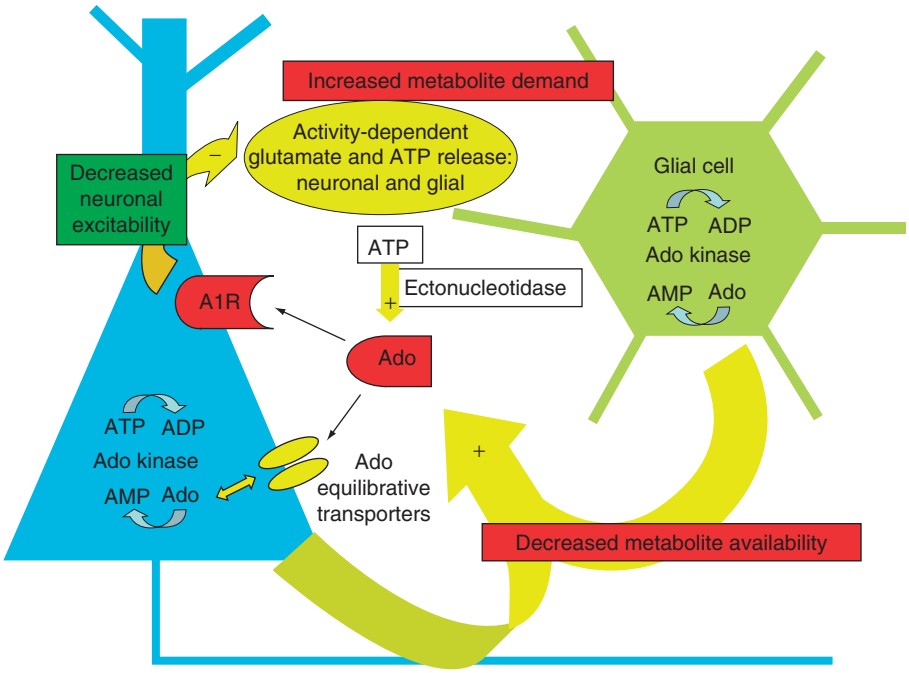

Figure 1 Metabolite demand and metabolite availability ratio, depicting extracellular adenosine (Ado) and the A1 Ado receptor (A1R).

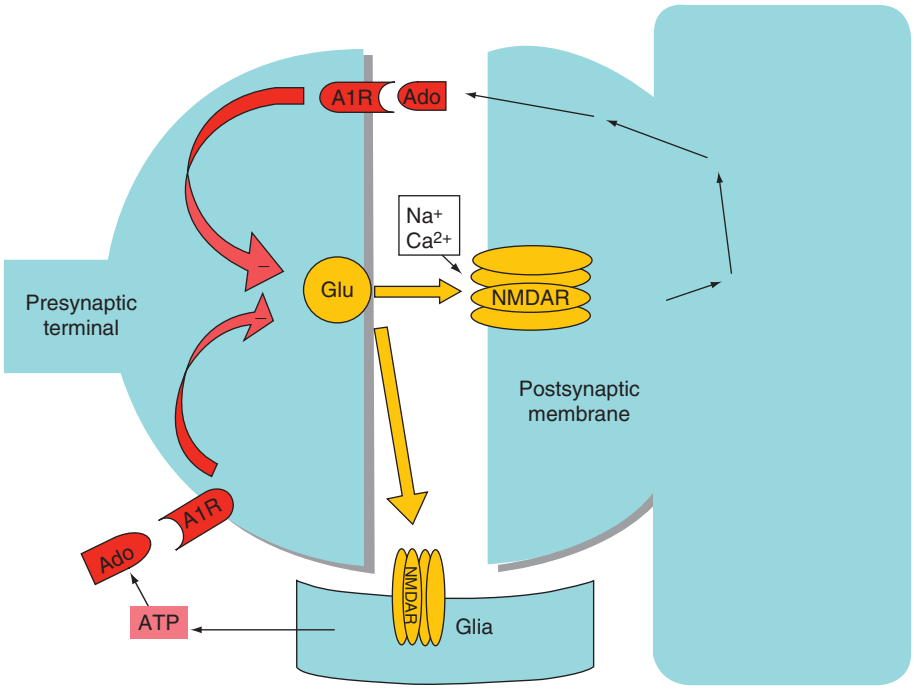

Figure 2 Adenosine (Ado) metabolism. Adenosine triphosphate (ATP) is broken down to form Ado in the extracellular medium, where it has been shown to both reduce postsynaptic excitability and inhibit presynaptic release of neurotransmitter. A1R, A1 Ado receptor; NMDAR, *N*-methyl-D-aspartate receptor.

A2b and A3 receptors – have been cloned; these are expressed in the CNS, although at considerably lower levels than the A1R is. The physiological actions of the A2b and the A3 receptors are not known. All of the receptors are metabotropic and coupled to G-proteins as follows: A1R is primarily coupled to G_i, A2a and A2b receptors are primarily coupled to G_s, and the A3 receptor is primarily coupled to G_{i3}/G_q.

The A2a receptor activation suppresses motor striatal-dependent motor activity and A2a knockout mice have increased motor activity. The electrophysiological mechanism responsible for A2a-dependent motor activity modulation is not known. Evidence from *in vitro* slice intracellular recordings suggests both enhancement and inhibition of GABAergic synaptic transmission. The enhancement, observed in the globus pallidus, is mediated by an A2a-induced cAMP activation of protein kinase A (PKA) to increase miniature synaptic current frequency, suggestive of a presynaptic effect.

A1 Receptor Electrophysiological Postsynaptic Effects

The best-characterized and most significant electrophysiological effects of Ado are mediated by A1 receptors on both the pre- and postsynaptic membranes that are found throughout the CNS. The A1Rs have the highest affinity for Ado (<100 nM), compared to A2a affinity that is $\sim 2 \times$ less sensitive and A2b and A3 receptors that are more than an order of magnitude less sensitive still.

The postsynaptic effects include (1) an increase in the G-protein-dependent, inwardly rectifying potassium current (GIRK); (2) an increase in the calcium-dependent, potassium-mediated, slow after-hyperpolarization; (3) a decrease in hyperpolarization-activated current (I_h); and (4) a decrease in voltage-gated calcium currents. All four effects, but especially the GIRK current increase, may reduce the responsiveness of the neuron. The GIRK current increase from A1R activation at high concentration of Ado (>100 nM) can render a neuron completely unresponsive to excitatory input and it desensitizes only very slowly, if at all. The increase in GIRK current is less sensitive than are the other Ado effects and thus is more likely to manifest in conditions of extreme or pathological metabolic stress, when Ado release is particularly high.

The increase in GIRK conductance coupled with a decrease in I_h facilitates single-cell oscillation of a 'burst' and a 'pause' in the frequency range of 0.5–4.5 Hz. This is the same frequency that predominates during slow-wave sleep (this is also called the delta frequency range) and imparts the name 'slow wave' to this CNS state. The oscillation is dependent on the low-threshold calcium current (I_T) that provides the depolarizing drive engendering a burst of action potentials during the burst phase of the oscillation. The depolarization results in an inactivation of the I_T (it turns off and cannot be turned back on until the inactivation is removed by hyperpolarization) and a deactivation of I_h (I_h simply turns off and will turn back on when the membrane potential falls within the range of activation for this current). As a result, the membrane potential hyperpolarizes beginning the 'pause' phase of the oscillation. With hyperpolarization, inactivation of I_T is removed (so that with depolarization it may become active, to cause a burst) and I_h is activated, causing the cell to depolarize, until threshold for the I_T- dependent burst is reached and the oscillation is repeated. The GIRK channel facilitates this effect by enhancing the hyperpolarization and, due to its inwardly rectifying property, the conductance is less with depolarization, allowing the burst to readily occur. Ado, acting through A1Rs, increases GIRK channel conductance and decreases I_h, both effects resulting in more hyperpolarization with greater removal of I_T inactivation and a more robust ensuing burst response. If many thalamocortical neurons are influenced by the A1R activation, then it is likely that a more powerful delta oscillation will be observed on the cortical electroencephalogram (EEG) during slow-wave sleep.

A1 Receptor Presynaptic Effects

A1 presynaptic receptors cause presynaptic inhibition. The presynaptic inhibition has been observed mostly with excitatory glutamatergic terminals in the cortex. However, Ado-mediated presynaptic inhibition is well documented for GABAergic terminals as well in the hypothalamus and brain stem tegmentum and most likely in other areas. Presynaptic inhibition of both excitatory and inhibitory inputs amounts to a relative functional deafferentation, predisposing the neuron to its own voltage-gated channel activity that can (as, for example, with thalamic neurons) result in slow oscillations. This may be important for behavioral-state-related activity, as discussed later.

Presynaptic inhibition mediated by A1R activation occurs in both the periphery and the CNS. At the neuromuscular junction, there is evidence for a role of reduced calcium ion flux across the membrane in the Ado-mediated presynaptic inhibition. At CNS excitatory glutamate synapses, a decreased calcium flux may also have a role, but this has yet to be convincingly demonstrated. However, A1R activation causes a well-characterized decrease in synaptic release in the CNS, due to fewer vesicles being released from the pool of vesicles used in action potential-dependent synaptic vesicle release. This decrease in vesicular release from the readily releasable pool is independent of calcium flux across the presynaptic membrane since osmotically evoked release (i.e., independent of calcium flux) is reduced by A1R activation. This effect is very sensitive to extracellular Ado and exerts an inhibitory Ado tone under physiological conditions.

Adenosine-Mediated Homeostasis of Synaptic Glutamate Release

Activation of *N*-methyl-D-aspartate (NMDA) receptors in cortical slices results in an increase of Ado release. Of particular note, this release is sufficient to exert electrophysiological inhibitory effects on glutamate-mediated synaptic release, mediated by A1Rs. It has been recently observed that synaptically released glutamate can, through NMDA receptor activation, cause Ado release that, in turn, activates presynaptic A1Rs on glutamatergic terminals. A complete negative feedback circuit is thus formed, since these presynaptic A1Rs inhibit glutamate synaptic vesicle release of vesicles from the readily releasable pool. (Consistent with inhibition of release from the readily releasable pool, the A1R inhibition also reduces action potential-evoked glutamate excitatory postsynaptic currents.) Thus, synaptically released glutamate activates NMDA receptors that cause an increase in extracellular Ado that feeds back onto presynaptic A1Rs to inhibit the release of glutamate. This negative feedback control exerts a tonic restraint on synaptic glutamate release under conditions that are comparable to those of *in vivo* synaptic release. Following an increase in tonic synaptic activity (the experimental manipulation employed was a small hypertonic stimulus in the presence of tetrodotoxin while monitoring miniature excitatory postsynaptic current frequency as a measure of vesicular release; see Figure 3), the activity returns to its former level over ~25–30 min.

The return to the original level of synaptic activity requires NMDA receptor activation, an increase in extracellular Ado, and A1R activation. The source of the extracellular Ado could be neuronal, glial, or both. The mechanism of the Ado release is also unclear, as it may result from a direct release of Ado, from ectonucleotidase-mediated breakdown of extracellular ATP (for example, released from synaptic vesicles), or from slowed reuptake of Ado or some combination of these mechanisms.

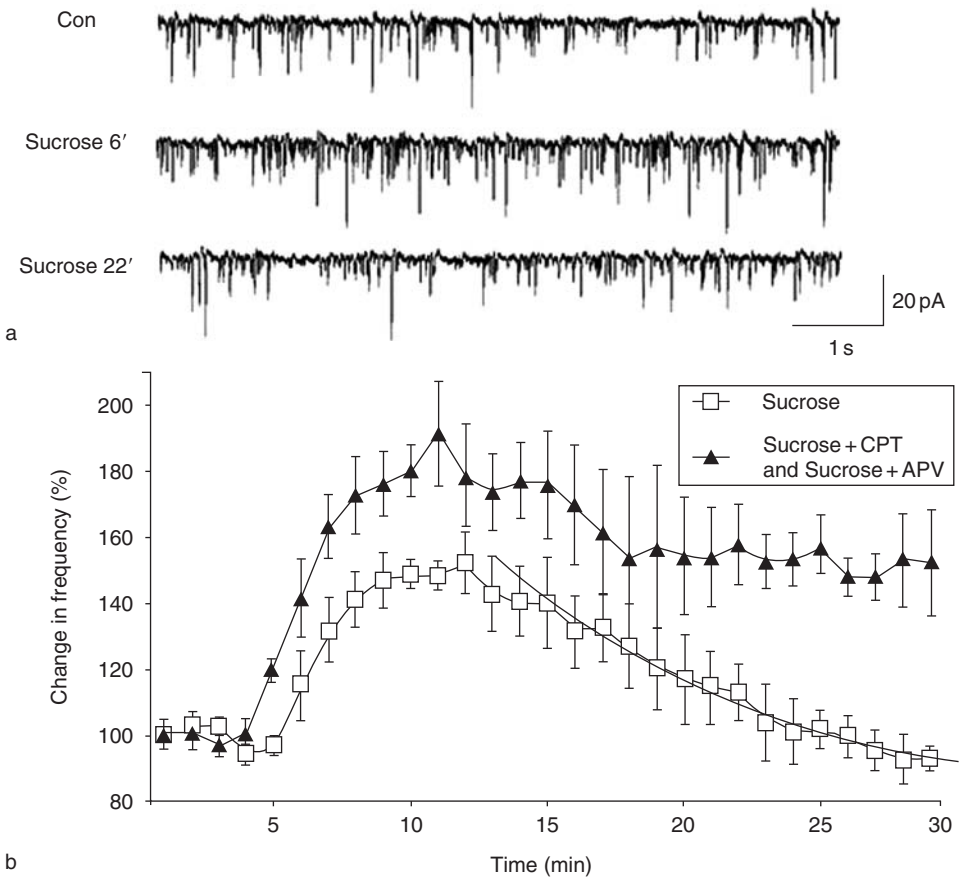

Figure 3 (a) The frequency of miniature excitatory postsynaptic currents mediated by glutamate in the presence of tetrodotoxin increases in hypertonic medium and then (over the next 25–30 min) spontaneously decreases back to the original baseline frequency. The decay occurs despite the continued presence of hypertonic medium. (b) The decay is blocked by an antagonist of the A1 receptor, by catabolism of adenosine on the addition of adenosine deaminase, or by antagonists of the *N*-methyl-D-aspartate receptor. APV, 2-amino-5-phosphonovaleric acid; CON, control; CPT, 8-cyclopenthyl-1,3-dimethylxanthine.

Adenosine and Glia

Glial cells receive glutamatergic synaptic input that can activate calcium-permeable α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) postsynaptic glial receptors. For example, in the hippocampus, Schaffer collateral input from the CA3 field pyramidal neurons to the CA1 field innervates not only pyramidal neurons and interneurons, but glia as well. In addition, these glial cells have vesicles that release their contents, including ATP, in a calcium-dependent manner. It was recently demonstrated that ATP released from glial vesicles can be broken down to Ado by ectonucleotidase and can reduce neuronal excitability in the hippocampus; this, in turn, reduces the likelihood of long-term potentiation of excitatory synaptic input to pyramidal cells. It is conceivable that the metabolic state of the hippocampal neural tissue determines the concentration of Ado in the extracellular medium derived from glial sources and, thus, affects hippocampal-dependent learning and memory. Indeed, when rodents explore a new environment, local Ado concentration, assessed by microdialysis, increases, suggesting a link with hippocampal activity and information processing. A similar correlation has been observed with motor activity and striatal Ado concentration.

Adenosine and Slow-Wave Sleep

A necessary condition for the transition from wake to sleep is the relative quiescence of the two cholinergic thalamocortical activating centers in the brain stem and basal forebrain. Ado, released locally, exerts an inhibitory tonus on the neurons of these two thalamocortical activating centers that is capable of reducing their excitability and action potential firing rates. The Ado exerts its inhibitory effect through activation of both presynaptic and postsynaptic A1Rs as described earlier. An increase in extracellular Ado in either or both of the two cholinergic arousal centers, resulting from either a local increase in metabolic demand (for example, increased neuronal activity) or a decrease in metabolite availability, can facilitate a transition from waking to sleep. Thus local metabolic factors may be linked through Ado to a change of behavioral state. Indeed, a local increase in extracellular Ado is sufficient to facilitate the transition from waking to sleep, as demonstrated by the sleep-inducing effects of local microdialysis of Ado into the brain stem or the basal forebrain cholinergic arousal centers.

The rebound sleep response to sleep deprivation suggests a homeostatic control of sleep, with characteristic dynamics that have been modeled as 'Process S' by Borbely and colleagues. After 6 h or more of enforced waking, a well-characterized rebound sleep ensues in mammals, including C57/BL mice. The neurophysiological mechanism(s) and substrates of this process are unknown, although recent evidence indicates that it is under genetic control. Synchronized, electrical, neuronal oscillations in the delta wave frequency range of 0.5–4.5 Hz can be quantified as delta power using Fourier analysis of the surface EEG wave form, and the delta power during the nonrapid eye movement sleep (NREM) period following sleep deprivation is proportional to the prior sleep loss. Further, delta power decreases during an NREM episode with a time course that is independent of the circadian phase. Together, these findings suggest a sleep homeostatic process that is closely associated with delta power.

Sleep deprivation elicits an increase in the extracellular level of Ado, and when sleep is allowed the Ado level returns toward baseline in the basal forebrain, where changes in extracellular Ado may be reasonably assessed by currently available microdialysis techniques. Thus, it may be predicted that a localized disruption of normal A1R-mediated inhibition would dampen the normal homeostatic response (both the rebound NREM duration and total delta power increase) to prolonged waking.

The observed increase in Ado with prolonged waking suggests a state-related control of Ado levels in the cholinergic arousal centers. Changes in Ado concentration are considered to regulate neurotransmission in the brain, especially with respect to glutamatergic neurotransmission, as described previously for the negative feedback circuit mediated by Ado at glutamatergic synapses that drive neurons in the cholinergic arousal centers. It has now been demonstrated that NMDA receptor (NMDAR) activation is sufficient to exert an Ado-mediated inhibition in the brain stem cholinergic arousal center to affect sleep/wake states in a manner that is closely related to glutamate-mediated afferent input. Glutamate-mediated input to the brain stem arousal center is greatly reduced during NREM sleep as evidenced by the decreased neuronal activity during this state, consistent with a negative feedback role (not necessarily an exclusive one) for NMDAR-induced adenosine-mediated inhibition in sleep/wake homeostasis (Figure 4). Accordingly, with prolonged glutamate activity and NMDAR activation of waking, extracellular Ado concentration will increase, reducing the glutamate-mediated drive onto the arousal centers, facilitating the transition from waking to sleep. Thus, an Ado-mediated negative feedback loop to maintain homeostasis of tonic glutamate-mediated transmission and homeostasis between metabolic and electrophysiological states, localized to the arousal centers, can have a global effect, as the widely divergent cholinergic (and possibly

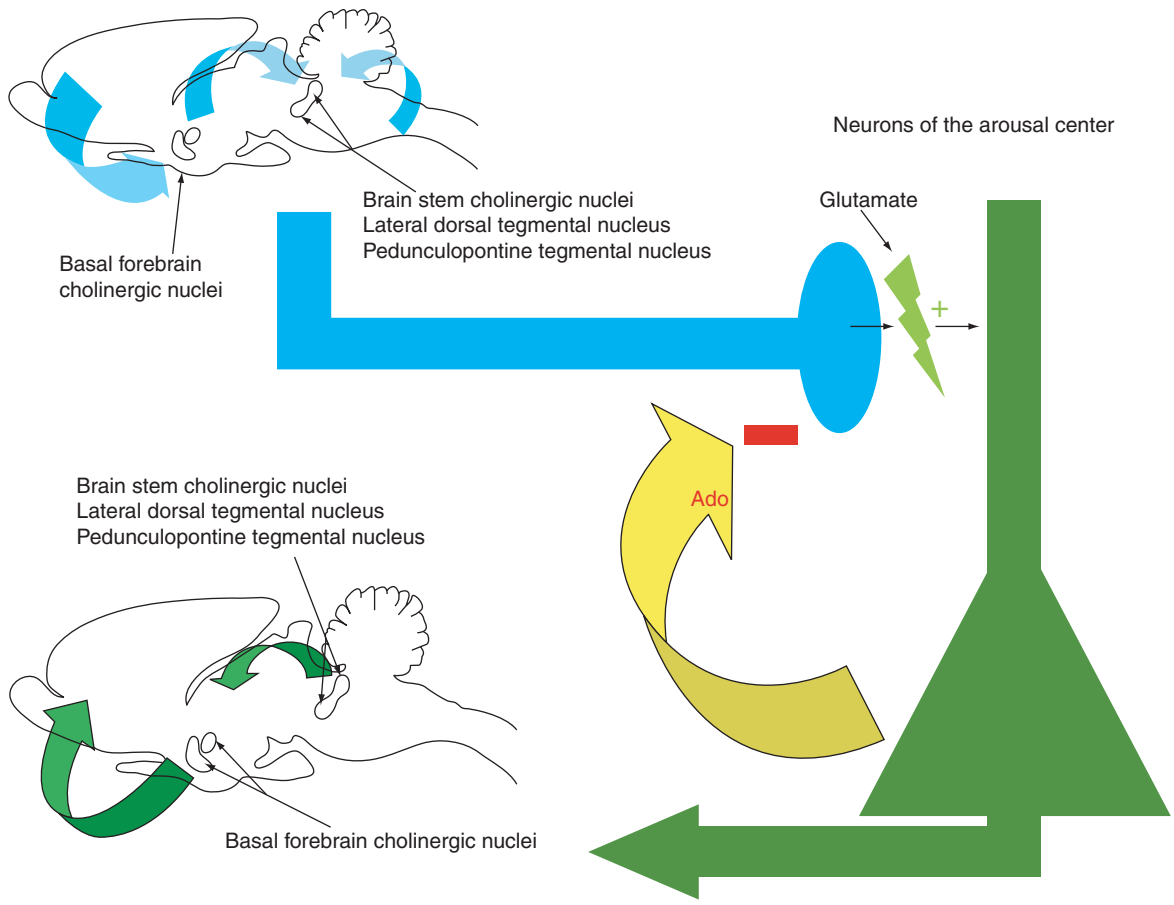

Figure 4 Neurons of the brain stem and forebrain arousal centers: from local to global effects, depicting the negative feedback circuit mediated by adenosine (Ado) at glutamatergic synapses.

gabaergic as well) tone exerted by these centers is diminished by Ado-dependent influence. Additionally at the target sites of the cholinergic centers' tone, especially the thalamocortical systems, extracellular Ado may be increased, as determined by the degree of metabolic demand placed on the particular system by the waking activity. In regions where the Ado is high, synchronous oscillatory activity in the delta frequency range (0.5–4.5 Hz) may be enhanced by Ado. During periods of prolonged wakefulness the increase in Ado may be widespread and may account, at least in part, for the observed rebound delta power observed in the surface EEG following periods of sleep deprivation. The burst–pause neuronal activity underlying the delta wave generation must be associated with an oscillation of intracellular calcium concentration, with the potential for eliciting a sleep functionally relevant change in intracellular metabolism. In summary, Ado might act at two sites to facilitate slow-wave sleep, as follows: first, at the cholinergic arousal centers, to reduce the modulatory influence of these centers on the rest of the CNS, and, second, at their target sites, to facilitate delta frequency oscillations.

See also: Adenosine Triphosphate (ATP); Adenosine Triphosphate (ATP) as a Neurotransmitter; NMDA Receptors and Development; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation; Pharmacology of Sleep: Adenosine.

Further Reading

Brambilla D, Chapman D, and Greene R (2005) Adenosine mediation of presynaptic feedback inhibition of glutamate release. *Neuron* 46: 275–283.

Dunwiddie TV and Masino SA (2001) The role and regulation of adenosine in the central nervous system. *Annual Review of Neuroscience* 24: 31–55.

Greene RW and Haas HL (1991) The electrophysiology of adenosine in the mammalian central nervous system. *Progress in Neurobiology* 36: 329–341.

Manzoni OJ, Manabe T, and Nicoll RA (1994) Release of adenosine by activation of NMDA receptors in the hippocampus. *Science* 265: 2098–2101.

Pascual O, Casper KB, Kubera C, et al. (2005) Astrocytic purinergic signaling coordinates synaptic networks. *Science* 310: 113–116.

Porkka-Heiskanen T, Strecker RE, Thakkar M, et al. (1997) Adenosine: A mediator of the sleep-inducing effects of prolonged wakefulness. *Science* 276: 1265–1268.

Rainnie DG, Grunze HC, McCarley RW, et al. (1994) Adenosine inhibition of mesopontine cholinergic neurons: Implications for EEG arousal. *Science* 263: 689–692.

Adenosine Triphosphate (ATP)

G Burnstock, Royal Free and University College School of Medicine, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Early History

Drury and Szent-Györgyi, in 1929, were the first to demonstrate the potent extracellular actions of adenosine 5′-triphosphate (ATP) and adenosine on the heart and coronary blood vessels. In 1948, Emmelin and Feldberg demonstrated that intravenous injection of ATP into cats caused complex effects that affected both peripheral and central mechanisms. Injection of ATP into the lateral ventricle produced muscular weakness, ataxia, and a tendency of the cat to sleep. Application of ATP to various regions of the brain produced biochemical or electrophysiological changes. Holton presented in 1959 the first hint of a transmitter role for ATP in the nervous system by demonstrating the release of ATP during antidromic stimulation of sensory nerves supplying the ear artery. Buchthal and Folkow recognized a physiological role for ATP at the neuromuscular junction in 1948, finding that acetylcholine (ACh)-evoked contraction of skeletal muscle fibers was potentiated by exposure to ATP. Presynaptic modulation of ACh release from the neuromuscular junction by purines in the rat was also reported.

The existence of nonadrenergic, noncholinergic (NANC) neurotransmission in the gut and bladder was established in the mid-1960s. Several years later, after many experiments, Burnstock and his colleagues published a study that suggested that the NANC transmitter in the guinea pig taenia coli and stomach, rabbit ileum, frog stomach, and turkey gizzard was ATP. The experimental evidence included mimicry of the NANC nerve-mediated response by ATP; measurement of release of ATP during stimulation of NANC nerves with luciferin-luciferase luminometry; histochemical labeling of subpopulations of neurons in the gut with quinacrine, a fluorescent dye known to selectively label high levels of ATP bound to peptides; the later demonstration that the slowly degradable analogue of ATP, α,β-methylene ATP (α,β-meATP), which produces selective desensitization of the ATP receptor, blocked the responses to NANC nerve stimulation. Soon after, evidence was presented for ATP as the neurotransmitter for NANC excitatory nerves in the urinary bladder. The term 'purinergic' was proposed in a short letter to *Nature* in 1971, and the evidence for purinergic transmission in a wide variety of systems was presented in *Pharmacological Reviews* in 1972 (Figure 1). This concept met with considerable resistance for many years. This was partly because it was felt that ATP was established as an intracellular energy source involved in various metabolic cycles and that such a ubiquitous molecule was unlikely to be involved in extracellular signaling. However, ATP was one of the biological molecules to first appear, and therefore it is not really surprising that it should have been utilized for extracellular, as well as intracellular, purposes early in evolution. The fact that potent ectoATPases were described in most tissues in the early literature was also a strong indication for the extracellular actions of ATP. Purinergic neurotransmission is now generally accepted, and a volume of *Seminars in Neuroscience* was devoted to purinergic neurotransmission in 1996.

Purinergic Cotransmission

Another concept that has had a significant influence on our understanding of purinergic transmission was that of cotransmission. Burnstock wrote a commentary in *Neuroscience* in 1976 titled, "Do some nerves release more than one transmitter?" This position challenged the single-neurotransmitter concept, which became known as 'Dale's Principle,' even though Dale himself never defined it as such. The commentary was based on hints about cotransmission in the early literature describing both vertebrate and invertebrate neurotransmission and more specifically, with respect to purinergic cotransmission, on the surprising discovery in 1971 that ATP was released from sympathetic nerves supplying the taenia coli as well as from NANC inhibitory nerves. The excitatory junction potentials (EJPs) recorded in the vas deferens were blocked by α,β-meATP, a selective desensitizer of P2X receptors (Figures 2(a) and 2(b)). This finding clearly supported the earlier demonstration of sympathetic cotransmission in the vas deferens in the laboratory of Dave Westfall, following an earlier report of sympathetic cotransmission in the cat nictitating membrane. Purinergic cotransmission was later described in the rat tail artery and in the rabbit saphenous artery. Noradrenaline (NA) and ATP are now well established as cotransmitters in sympathetic nerves (see Figure 3 (a)), although the proportions vary in different tissues and species, during development and aging, and in different pathophysiological conditions.

ACh and ATP are cotransmitters in parasympathetic nerves supplying the urinary bladder. Subpopulations of sensory nerves have been shown to utilize ATP in addition to substance P and calcitonin gene-related peptide; it seems likely that ATP cooperates

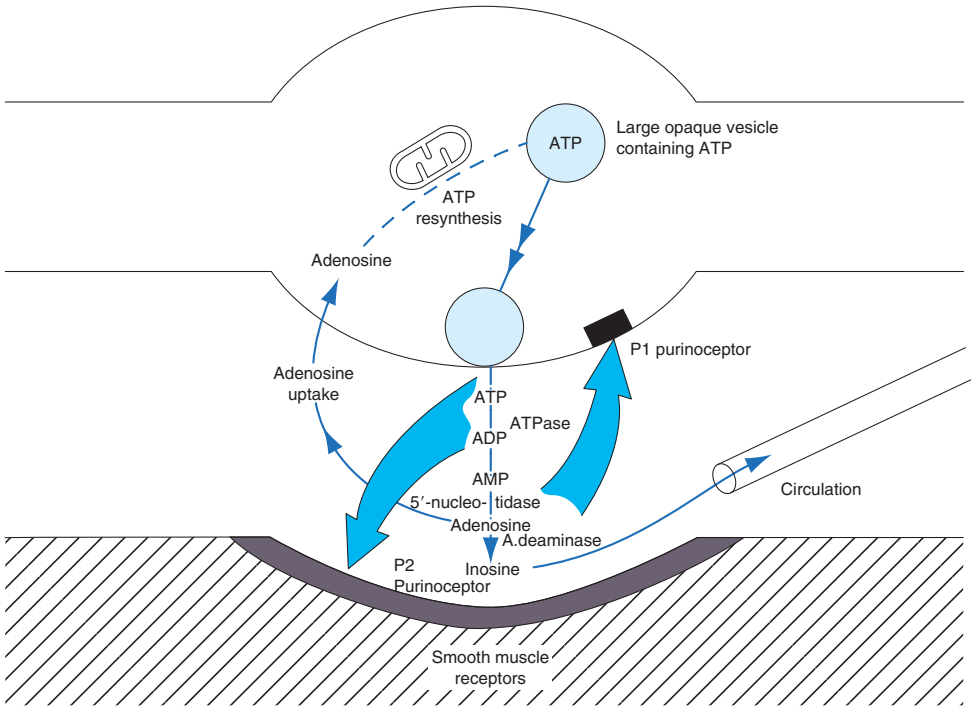

Figure 1 Purinergic neuromuscular transmission depicting the synthesis, storage, release, and inactivation of adenosine 5′-triphosphate (ATP). ATP, stored in vesicles in nerve varicosities, is released by exocytosis to act on postjunctional P2 purinoceptors on smooth muscle. ATP is broken down extracellularly by ATPases and 5′-nucleotidase to adenosine, which is taken up by varicosities to be resynthesized and restored in vesicles. Adenosine acts prejunctionally on P1 purinoceptors to modulate transmitter release. If adenosine is broken down further by adenosine deaminase to inosine, it is removed by the circulation. Adapted from Burnstock G (1972) Purinergic nerves. *Pharmacological Reviews* 24: 509–581, with permission from the American Society for Pharmacology and Experimental Therapeutics.

with these peptides in axon reflex activity. ATP, vasoactive intestinal polypeptide and nitric oxide (NO) are cotransmitters in NANC inhibitory nerves, but that they vary considerably in proportion in different regions of the gut. More recently, ATP has been shown to be a cotransmitter with NA, 5-hydroxytryptamine, glutamate, dopamine and γ-aminobutyric acid (GABA) in the central nervous system (CNS) (see Figure 3(b)). ATP and NA act synergistically to release vasopressin and oxytocin, which is consistent with ATP cotransmission in the hypothalamus. ATP, in addition to glutamate, is involved in long-term potentiation in hippocampal CA1 neurons associated with learning and memory.

Receptors for Purines and Pyrimidines

Implicit in the purinergic neurotransmission hypothesis was the presence of purinoceptors. A basis for distinguishing two types of purinoceptor, identified as P1 and P2 for adenosine and ATP/adenosine diphosphate (ADP), respectively, was recognized in 1978. This helped resolve some of the ambiguities in earlier reports, which were complicated by the breakdown of ATP to adenosine by ectoenzymes so that some of the actions of ATP were directly on P2 receptors, whereas others were due to indirect action via P1 receptors.

At about the same time, two subtypes of P1 (adenosine) receptor were recognized, but it was not until 1985 that a pharmacological basis for distinguishing two types of P2 receptors (P2X and P2Y) was proposed. A year later, two further P2 receptor subtypes were named, a P2T receptor selective for ADP on platelets and a P2Z receptor on macrophages. Further subtypes followed, perhaps the most important being the P2U receptor, which could recognize pyrimidines such as uridine 5′ triphosphate (UTP) in addition to ATP. However, to provide a more manageable framework for newly identified nucleotide receptors, Abbracchio and Burnstock proposed in 1994 that purinoceptors should belong to two major families: a P2X family of ligand-gated ion channel receptors and a P2Y family of G-protein-coupled receptors. This was based on studies of transduction mechanisms and the cloning of nucleotide receptors: P2Y receptors were cloned first, in 1993, and a year later P2X receptors were cloned. This nomenclature has been widely adopted, and currently seven P2X subtypes and eight P2Y

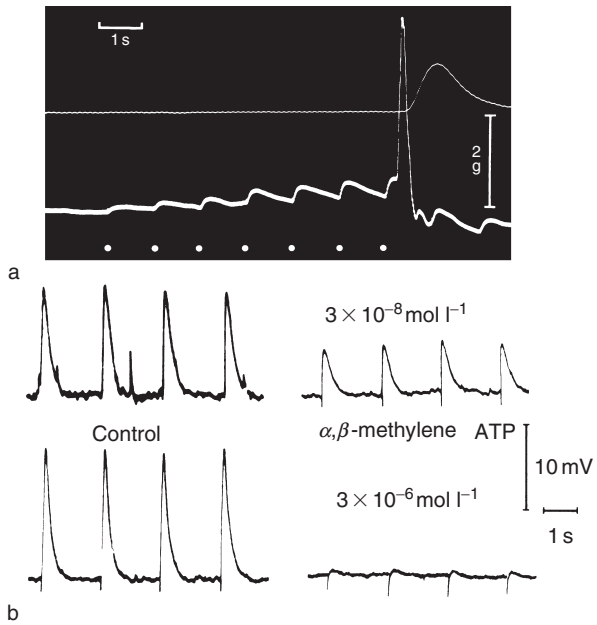

Figure 2 (a) EJPs in response to repetitive stimulation of adrenergic nerves (white dots) in the guinea pig vas deferens. The upper trace records the tension, the lower trace the electrical activity of the muscle recorded extracellularly by the sucrose gap method. Note both summation and facilitation of successive junction potentials. At a critical depolarization threshold, an action potential is initiated which results in contraction. (b) The effect of various concentrations of α,β-methylene ATP (α,β-meATP) on EJPs recorded from guinea pig vas deferens (intracellular recordings). The control responses to stimulation of the motor nerves at 0.5 Hz are shown on the left. After at least 10 min in the continuous presence of the indicated concentration of α,β-meATP, EJPs were recorded using the same stimulation parameters. (a) Reproduced from Burnstock G and Costa M (eds.) (1975) Adrenergic neuroeffector transmission. In *Adrenergic Neurones: Their Organisation, Function and Development in the Peripheral Nervous System*, pp. 51–106. London: Chapman and Hall, with permission from Springer Science and Business Media. (b) Reproduced from Sneddon P and Burnstock G (1984) Inhibition of excitatory junction potentials in guinea-pig vas deferens by α,β-methylene-ATP: Further evidence for ATP and noradrenaline as cotransmitters. *European Journal of Pharmacology* 100: 85–90, with permission from Elsevier.

receptor subtypes are recognized. Four subtypes of P1 receptor have been cloned and characterized.

P2X receptors in general mediate fast neurotransmission but are sometimes located prejunctionally to mediate increase in release of cotransmitters, for example, glutamate in terminals of primary afferent neurons in the spinal cord. $P2X_3$ and $P2X_{2/3}$ receptors are prominent in sensory neurons and are involved in nociception. $P2X_7$ receptors are involved in cell death.

P2Y receptors are particularly involved in prejunctional inhibitory modulation of transmitter release, as well as long-term (trophic) events such as cell pro liferation. $P2Y_1$ receptors are widespread in many regions of the brain, while the $P2Y_2$ receptors have been localized on pyramidal neurons in the hippocampus and prefrontal cortex, on supraoptic magnocellular neurosecretory neurons in the hypothalamus, and on neurons in the dorsal horn of the spinal cord. In addition, mRNA but not protein has been reported for $P2Y_4$ and $P2Y_6$ receptor subtypes in the cerebellum and hippocampus, while $P2Y_{12}$ receptor mRNA has also been described in the cerebellum and $P2Y_{13}$ in the cortex. In the periphery, $P2Y_{1,2,4,6}$ receptors have been described on subpopulations of sympathetic neurons, $P2Y_2$ and $P2Y_4$ receptors in intracardiac ganglia, $P2Y_1$ and $P2Y_2$ receptors on sensory neurons (although $P2Y_4$ and $P2Y_6$ mRNA have also been reported) while $P2Y_1$ receptors appear to be the dominant subtype on enteric neurons. $P2Y_{1,2,4,6}$ functional receptors have been described on astrocytes in the CNS and also on microglia, where functional $P2Y_{12}$ receptors have also been identified. $P2Y_1$ and $P2Y_2$ receptors have been located in Schwann cells and oligodendrocytes, where functional $P2Y_{12}$ receptors also appear to be present. $P2Y_2$ (and/or $P2Y_4$) receptors are expressed on enteric glial cells. There is also emerging evidence for P2Y receptors on stem cells.

ATP Release and Degradation

There is clear evidence for exocytotic vesicular release of ATP from nerves, and the concentration of nucleotides in vesicles is claimed to be up to 1000 mmol l^{-1}. It was generally assumed that the main source of ATP acting on purinoceptors was damaged or dying cells. However, it is now recognized that ATP release from many cells is a physiological or pathophysiological response to mechanical stress, hypoxia, inflammation, and some agonists. There is debate, however, about the ATP transport mechanisms involved. There is compelling evidence for exocytotic release from endothelial and urothelial cells, osteoblasts, astrocytes, mast, and chromaffin cells, but other transport mechanisms have also been proposed, including ATP binding cassette transporters, connexin hemichannels, and plasmalemmal voltage-dependent anion channels.

Much is now known about the ectonucleotidases that break down ATP released from neurons and non-neuronal cells. Several enzyme families are involved: ecto-nucleoside triphosphate diphosphohydrolases (E-NTPDases), of which NTPDase1, 2, 3, and 8 are extracellular; ectonucleotide pyrophosphatase of 3 subtypes; alkaline phosphatases; ecto-5′-nucleotidase; and ecto-nucleoside diphosphokinase. NTPDase1 hydrolyzes ATP directly to adenosine monophosphate (AMP) and UTP to uridine diphosphate (UDP), while NTPDase2 hydrolyzes ATP to ADP and 5′-nucleotidase AMP to adenosine.

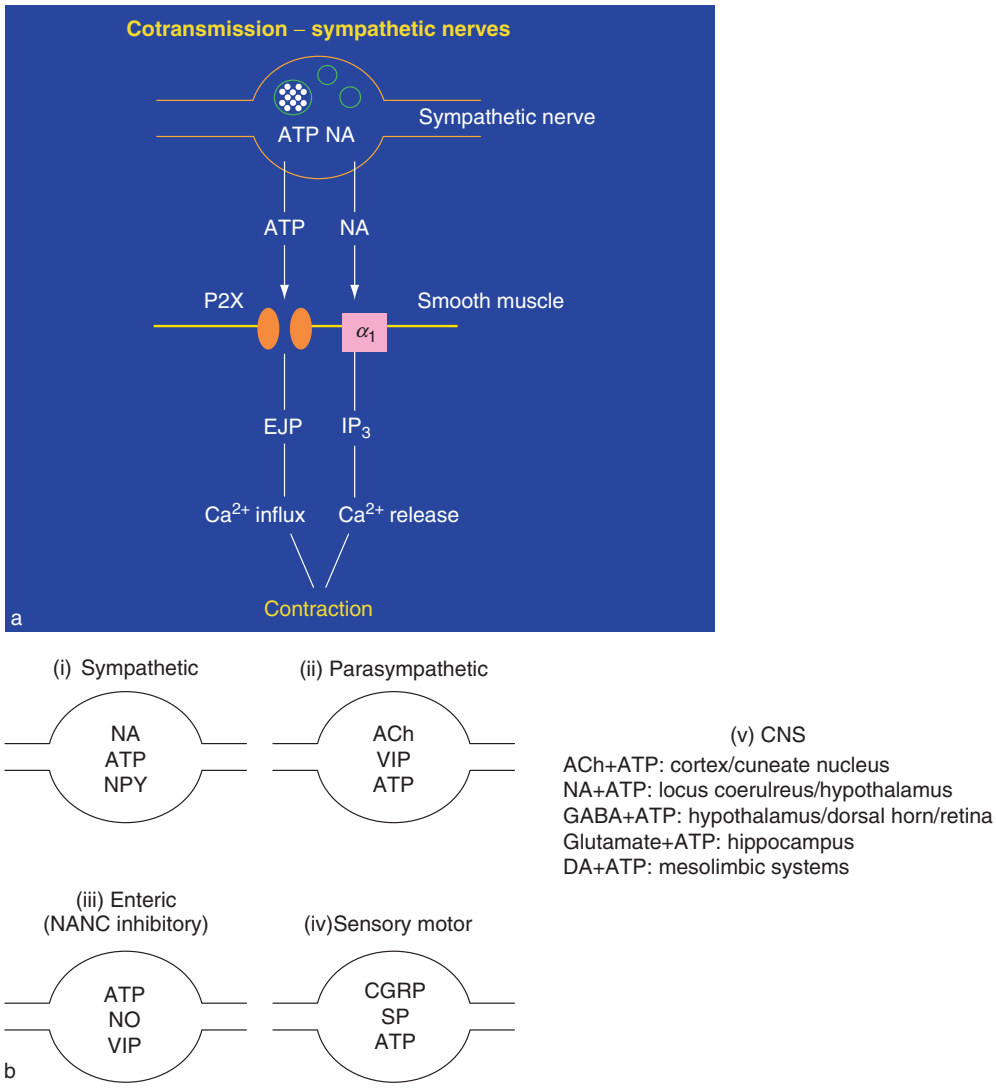

Figure 3 (a) Cotransmission in sympathetic nerves. Adenosine 5′-triphosphate (ATP) and noradrenaline (NA) from terminal varicosities of sympathetic nerves can be released together. With NA acting via the postjunctional α_1-adrenoceptor to release cytosolic Ca^{2+}, and with ATP acting via $P2X_1$-gated ion channels to elicit Ca^{2+} influx, both contribute to the subsequent response (contraction). IP_3 is inositol triphosphate, EJP is excitatory junction potential. (b) Schematic diagram of the principal cotransmitters with ATP in the nervous system. Nerve terminal varicosities of (i) sympathetic, (ii) parasympathetic, (iii) enteric (NANC inhibitory), (iv) sensory-motor neurons, and (v) central nervous system (CNS). (a) Adapted from Kennedy C, McLaren GJ, Westfall TD, and Sneddon P (1996) ATP as a co-transmitter with noradrenaline in sympathetic nerves – Function and fate. In: Chadwick DJ and Goode J (eds.) *P2 Purinoceptors: Localization, Function and Transduction Mechanisms*, pp. 223–235. Chichester: John Wiley and Sons, with permission from John Wiley & Sons. (b) Reproduced from Burnstock G (2007) Physiology and pathophysiology of purinergic neurotransmission. *Physiological Reviews* 87: 659–797, with permission from The American Physiological Society.

Physiology of Purinergic Neurotransmission

Purinergic signaling appears to be a primitive system that is involved in many nonneuronal and neuronal mechanisms, in both short-term and long-term (trophic) events, including exocrine and endocrine secretion, immune responses, inflammation, mechanosensory transduction, platelet aggregation, endothelial-mediated vasodilatation and in cell proliferation, differentiation, migration, and death in development and regeneration.

The first clear evidence for nerve–nerve purinergic synaptic transmission was published in 1992. Synaptic potentials in the coeliac ganglion and in the medial habenula in the brain were reversibly antagonized by the antitrypanosomal agent suramin. Since then, many articles have described either the distribution of various P2 receptor subtypes in the brain and

spinal cord or electrophysiological studies of the effects of purines in brain slices, isolated neurons, and glial cells. Synaptic transmission has also been demonstrated in the myenteric plexus and in various sensory, sympathetic, parasympathetic, and pelvic ganglia. Adenosine produced by the ectoenzymatic breakdown of ATP acts through presynaptic P1 receptors to inhibit the release of excitatory neurotransmitters in both the peripheral and the central nervous systems. Purinergic signaling is also implicated in higher order cognitive functions, including learning and memory in the prefrontal cortex.

CNS Control of Autonomic Function

Functional interactions seem likely to occur between purinergic and nitrergic neurotransmitter systems; these interactions might be important for the regulation of hormone secretion and body temperature at the hypothalamic level and for cardiovascular and respiratory control at the level of the brain stem. The nucleus tractus solitarius (NTS) is a major integrative center of the brain stem involved in reflex control of the cardiovascular system; stimulation of P2X receptors in the NTS evokes hypotension. P2X receptors expressed in neurons in the trigeminal mesencephalic nucleus might be involved in the processing of proprioceptive information.

Neuron–Glia Interactions

ATP is an extracellular signaling molecule between neurons and glial cells. ATP released from astrocytes might be important in triggering cellular responses to trauma and ischemia by initiating and maintaining reactive astrogliosis, which involves striking changes in the proliferation and morphology of astrocytes and microglia. Some of the responses to ATP released during brain injury are neuroprotective, but at higher concentrations, ATP contributes to the pathophysiology initiated after trauma. Multiple P2X and P2Y receptor subtypes are expressed by astrocytes, oligodendrocytes, and microglia. ATP and basic fibroblast growth factor (bFGF) signals merge at the mitogen-activated protein kinase cascade, which underlies the synergistic interactions of ATP and bFGF in astrocytes. ATP can activate $P2X_7$ receptors in astrocytes to release glutamate, GABA, and ATP, which regulate the excitability of neurons.

Microglia, immune cells of the CNS, are also activated by purines and pyrimidines to release inflammatory cytokines such as interleukins 1β (IL-1β) and IL-6 and tumor necrosis factor α. Thus, although microglia might play an important role against infection in the CNS, overstimulation of this immune reaction might accelerate the neuronal damage caused by ischemia, trauma, or neurodegenerative diseases. $P2X_4$ receptors induced in spinal microglia gate tactile allodynia after nerve injury. $P2X_7$ receptors mediate superoxide production in primary microglia and are upregulated in a transgenic mouse model of Alzheimer's disease, particularly around β-amyloid plaques.

Purine Transmitter and Receptor Plasticity

The purinergic neurotransmission field is expanding rapidly; there is increasing interest in the physiology and pathophysiology of this neurosignaling system, and therapeutic interventions are being explored. The autonomic nervous system shows marked plasticity: that is, the expression of cotransmitters and receptors shows dramatic changes during development and aging, in nerves that remain after trauma or surgery, and in disease conditions. There are several examples where the purinergic component of cotransmission is increased in pathological conditions. The parasympathetic purinergic nerve-mediated component of contraction of the human bladder is increased to 40% in pathophysiological conditions such as interstitial cystitis, outflow obstruction, idiopathic instability, and also some types of neurogenic bladder. ATP also has a significantly greater cotransmitter role in sympathetic nerves supplying hypertensive compared to normotensive blood vessels. Upregulation of $P2X_1$ and $P2Y_2$ receptor mRNA in hearts of rats with congestive heart failure has been reported, and there is a dramatic increase in expression of $P2X_7$ receptors in the kidney glomerulus in diabetes and hypertension.

Neuroprotection

In the brain, purinergic signaling is involved in nervous tissue remodeling following trauma, stroke, ischemia, or neurodegenerative disorders. The hippocampus of chronic epileptic rats shows abnormal responses to ATP associated with increased expression of $P2X_7$ receptors. Neuronal injury releases fibroblast growth factor, epidermal growth factor, and platelet-derived growth factor. In combination with these growth factors, ATP can stimulate astrocyte proliferation, contributing to the process of reactive astrogliosis and to hypertrophic and hyperplasic responses. P2Y receptor antagonists have been proposed as potential neuroprotective agents in the cortex, hippocampus, and cerebellum. Blockade of

A_{2A} (P1) receptors antagonizes tremor in Parkinson's disease. ATP–$MgCl_2$ is being explored for the treatment of spinal cord injuries.

Dual Purinergic Neural and Endothelial Control of Vascular Tone and Angiogenesis

ATP and adenosine are much involved in the mechanisms underlying local control of vessel tone in addition to cell migration, proliferation, and death during angiogenesis, atherosclerosis, and restenosis following angioplasty. ATP, released as a cotransmitter from sympathetic nerves, constricts vascular smooth muscle via P2X receptors, whereas ATP released from sensory-motor nerves during 'axon reflex' activity dilates vessels via P2Y receptors. Furthermore, ATP released from endothelial cells during changes in flow (shear stress) or hypoxia acts on P2Y receptors in endothelial cells to release NO, resulting in relaxation (Figure 4). Adenosine, following breakdown of extracellular ATP, produces vasodilatation via smooth muscle P1 receptors.

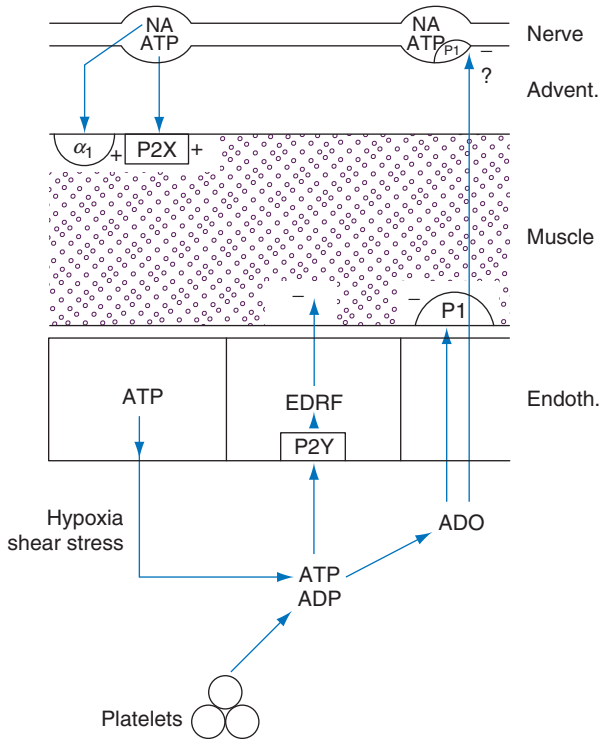

Figure 4 A schematic representation of the interactions of adenosine 5′-triphosphate (ATP) released from perivascular nerves and from the endothelium (Endoth.). ATP is released from endothelial cells during hypoxia to act on endothelial P2Y receptors, leading to the production of endothelium-derived relaxing factor (EDRF), nitric oxide (NO), and subsequent vasodilation (–). In contrast, ATP released as a cotransmitter with noradrenaline (NA) from perivascular sympathetic nerves at the adventitia (Advent.)–muscle border produces vasoconstriction (+) via P2X receptors on the muscle cells. Adenosine (ADO), resulting from rapid breakdown of ATP by ectoenzymes, produces vasodilation by direct action on the muscle via P1 receptors and acts on the perivascular nerve terminal varicosities to inhibit transmitter release. Reproduced from Burnstock G (1987) Local control of blood pressure by purines. *Blood Vessels* 24: 156–160, with permission from S. Karger AG, Basel.

Pain and Purinergic Mechanosensory Transduction

The involvement of ATP in the initiation of pain was recognized first in 1966 and later in 1977 using human skin blisters. A major advance was made when the $P2X_3$ ionotropic receptor was cloned in 1995 and shown later to be predominantly localized in the subpopulation of small nociceptive sensory nerves that label with isolectin B4 in dorsal root ganglia whose central projections terminate in inner lamina II of the dorsal horn. A unifying purinergic hypothesis for the initiation of pain was proposed in 1996 with ATP acting via $P2X_3$ and $P2X_{2/3}$ receptors associated with causalgia, reflex sympathetic dystrophy, angina, migraine, and pelvic and cancer pain. This has been followed by an increasing number of published reports expanding on this concept for acute, inflammatory, neuropathic, and visceral pain. $P2Y_1$ receptors have also been demonstrated in a subpopulation of sensory neurons that colocalized with $P2X_3$ receptors.

A hypothesis was proposed that purinergic mechanosensory transduction occurred in visceral tubes and sacs, including ureter, bladder, and gut, where ATP, released from epithelial cells during distension, acted on $P2X_3$ homomultimeric and $P2X_{2/3}$ heteromultimeric receptors on subepithelial sensory nerves, initiating impulses in sensory pathways to pain centers in the CNS (Figure 5(a)). Subsequent studies of bladder, ureter, gut, tongue, and tooth pulp have produced evidence in support of this hypothesis. $P2X_3$ knockout mice were used to show that ATP released from urothelial cells during distension of the bladder act on $P2X_3$ receptors on subepithelial sensory nerves to initiate both nociceptive and bladder voiding reflex activities. In the distal colon, ATP released during moderate distension acts on $P2X_3$ receptors on low-threshold intrinsic subepithelial sensory neurons to influence peristalsis, whereas high-threshold extrinsic subepithelial sensory fibers respond to severe distension to initiate pain (see Figure 5(b)).

ATP is also a neurotransmitter released from the spinal cord terminals of primary afferent sensory nerves to act at synapses in the central pain pathway. Using transverse spinal cord slices from postnatal rats, excitatory postsynaptic currents have been shown to be

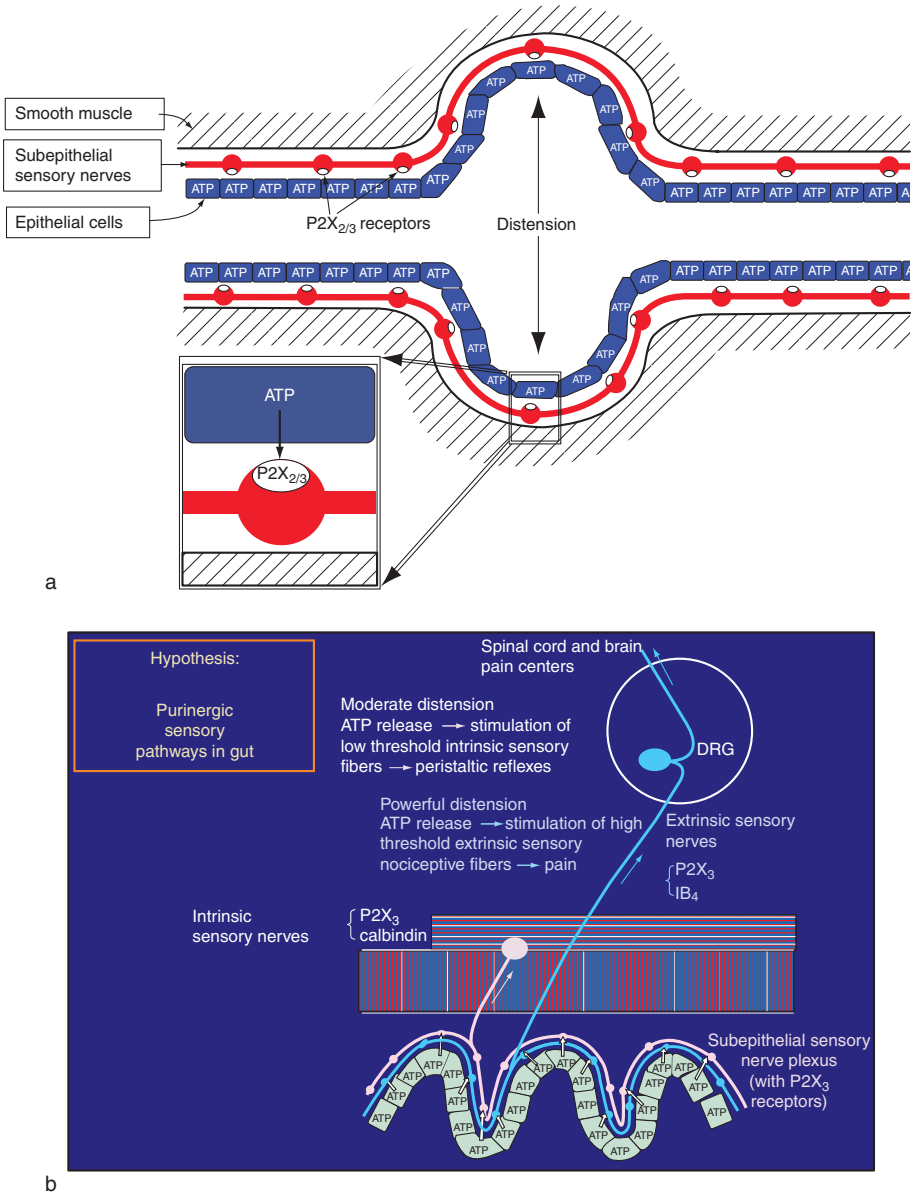

Figure 5 (a) Schematic representation of the hypothesis for purinergic mechanosensory transduction in tubes (e.g., ureter, vagina, salivary, and bile ducts and gut) and sacs (e.g., urinary and gall bladders and lung). It is proposed that distension leads to the release of adenosine 5′ triphosphate (ATP) from the epithelium lining the tube or sac, which then acts on $P2X_{2/3}$ receptors on subepithelial sensory nerves to convey sensory (nociceptive) information to the central nervous system (CNS). (b) Schematic of a novel hypothesis about purinergic mechanosensory transduction in the gut. It is proposed that ATP released from mucosal epithelial cells during moderate distension acts preferentially on $P2X_3$ receptors on low-threshold subepithelial intrinsic sensory nerve fibers (labeled with calbindin), contributing to peristaltic reflexes. ATP released during extreme distension also acts on $P2X_3$ receptors on high-threshold extrinsic sensory nerve fibers (labeled with isolectin B_4 (IB_4)) that send messages via the dorsal root ganglia (DRG) to pain centers in the CNS. (a) Adapted from Burnstock G (1999) Release of vasoactive substances from endothelial cells by shear stress and purinergic mechanosensory transduction. *Journal of Anatomy* 194: 335–342, with permission from Blackwell Publishing. (b) Adapted from Burnstock G (2001) Expanding field of purinergic signaling. *Drug Development Research* 52: 1–10, with permission of Wiley-Liss, Inc.

mediated by P2X receptors activated by synaptically released ATP, in a subpopulation of less than 5% of the neurons in lamina II, a region known to receive major input from nociceptive primary afferents.

There is an urgent need for selective $P2X_3$ and $P2X_{2/3}$ receptor antagonists that do not degrade *in vivo*. Pyridoxal-phosphate-6-azophenyl-2′, 4′-disulphonic acid is a nonselective P2 receptor antagonist but has the advantage that it dissociates about 100 to 10 000 times more slowly than other known antagonists. The trinitrophenyl-substituted nucleotide TNP-ATP is a selective and very potent antagonist at both

$P2X_3$ and $P2X_{2/3}$ receptors. 5-((3-Phenoxybenzyl) [(1*S*)-1,2,3,4-tetrahydro-1-naphthalenyl]amino carbonyl)-1,2,4-benzenetricarboxylic acid (A-317491) is a potent and selective nonnucleotide antagonist of $P2X_3$ and $P2X_{2/3}$ receptors, and it reduces chronic inflammatory and neuropathic pain in the rat. Antisense oligonucleotides have been used to downregulate the $P2X_3$ receptor, and in models of neuropathic (partial sciatic nerve ligation) and inflammatory (complete Freund's adjuvant) pain, inhibition of the development of mechanical hyperalgesia was observed within 2 days of treatment. $P2X_3$ double-stranded-short interfering RNA also relieves chronic neuropathic pain and opens up new avenues for therapeutic pain strategies in humans. Tetramethylpyrazine, a traditional Chinese medicine used as an analgesic for dysmenorrhoea, is claimed to be a P2X receptor antagonist, and it inhibited significantly the first phase of nociceptive behavior induced by 5% formalin and attenuated slightly the second phase in the rat hindpaw pain model. Antagonists to P2 receptors are also beginning to be explored in relation to cancer pain.

Special Senses

Eye

$P2X_2$ and $P2X_3$ receptor mRNA is present in the retina and receptor protein expressed in retinal ganglion cells. $P2X_3$ receptors are also present on Müller cells, which release ATP during Ca^{2+} wave propagation. ATP, acting via both P2X and P2Y receptors, modulates retinal neurotransmission, affecting retinal blood flow and intraocular pressure. Topical application of diadenosine tetraphosphate has been proposed for the lowering of intraocular pressure in glaucoma. The formation of $P2X_7$ receptor pores and apoptosis is enhanced in retinal microvessels early in the course of experimental diabetes, suggesting that purinergic vasotoxicity might have a role in microvascular cell death, a feature of diabetic retinopathy. The possibility has been raised that alterations in sympathetic nerves might underlie some of the complications observed in diabetic retinopathy; ATP is well established as a cotransmitter in sympathetic nerves, raising the potential for P2 receptor antagonists in glaucoma. $P2Y_2$ receptor activation increases salt, water, and mucus secretion and thus represents a potential treatment for dry eye conditions.

Ear

Both P2X and P2Y receptors have been identified in the vestibular system. ATP regulates fluid homeostasis, cochlear blood flow, hearing sensitivity, and development and thus might be useful in the treatment of Ménières disease, tinnitus, and sensorineural deafness. ATP, acting via P2Y receptors, depresses sound-evoked gross compound action potentials in the auditory nerve and the distortion product otoacoustic emission, the latter being a measure of the active process of the outer hair cells. P2X splice variants are found on the endolymphatic surface of the cochlear endothelium, an area associated with sound transduction. Sustained loud noise produces an upregulation of $P2X_2$ receptors in the cochlear, particularly at the site of outer hair cell sound transduction. $P2X_2$ receptor expression is also increased in spiral ganglion neurons, indicating that extracellular ATP acts as a modulator of auditory neurotransmission that is adaptive and dependent on the noise level. Excessive noise can irreversibly damage hair cell stereocilia, leading to deafness. Data have been presented that release of ATP from damaged hair cells is required for Ca^{2+} wave propagation through the support cells of organ of Corti, involving P2Y receptors, and this might constitute the fundamental mechanism to signal the occurrence of hair cell damage.

Nasal Organs

The olfactory epithelium and vomeronasal organs contain olfactory receptor neurons that express $P2X_2$, $P2X_3$, and $P2X_{2/3}$ receptors. It is suggested that the neighboring epithelial supporting cells or the olfactory neurons themselves can release ATP in response to noxious stimuli, acting on P2X receptors as an endogenous modulator of odor sensitivity. Enhanced sensitivity to odors was observed in the presence of P2 antagonists, suggesting that low-level endogenous ATP normally reduces odor responsiveness. It has been suggested that the predominantly suppressive effect of ATP on odor sensitivity could be involved in reduced odor sensitivity that occurs during acute exposure to noxious fumes and might be a novel neuroprotective mechanism.

See also: Adenosine; Adenosine Triphosphate (ATP) as a Neurotransmitter; Adenosine Receptor Mediated Functions; Cotransmission; Pharmacology of Sleep: Adenosine; Purinergic Receptors; Purines and Purinoceptors: Molecular Biology Overview.

Further Reading

Abbracchio MP and Burnstock G (1994) Purinoceptors: Are there families of P2X and P2Y purinoceptors? *Pharmacology and Therapeutics* 64: 445–475.

Abbracchio MP and Burnstock G (1998) Purinergic signalling: Pathophysiological roles. *Japanese Journal of Pharmacology* 78: 113–145.

Bodin P and Burnstock G (2001) Purinergic signalling: ATP release. *Neurochemical Research* 26: 959–969.

Burnstock G (1972) Purinergic nerves. *Pharmacological Reviews* 24: 509–581.

Burnstock G (1987) Local control of blood pressure by purines. *Blood Vessels* 24: 156–160.

Burnstock G (1999) Release of vasoactive substances from endothelial cells by shear stress and purinergic mechanosensory transduction. *Journal of Anatomy* 194: 335–342.

Burnstock G (2001) Expanding field of purinergic signaling. *Drug Development Research* 52: 1–10.

Burnstock G (2001) Purine-mediated signalling in pain and visceral perception. *Trends in Pharmacological Sciences* 22: 182–188.

Burnstock G (2002) Purinergic signalling and vascular cell proliferation and death. *Arteriosclerosis, Thrombosis and Vascular Biology* 22: 364–373.

Burnstock G (2001) Purinergic signalling in gut. In: Abbracchio MP and Williams M (eds.) *Handbook of Experimental Pharmacology, Vol. 151: Purinergic and Pyrimidinergic Signalling II: Cardiovascular, Respiratory, Immune, Metabolic and Gastrointestinal Tract Function*, pp. 141–238. Berlin: Springer.

Burnstock G (2001) Purinergic signalling in lower urinary tract. In: Abbracchio MP and Williams M (eds.) *Handbook of Experimental Pharmacology, Vol. 151: Purinergic and Pyrimidinergic Signalling I: Molecular, Nervous and Urinogenitary System Function*, pp. 423–515. Berlin: Springer.

Burnstock G (2003) Purinergic receptors in the nervous system. In: Schwiebert EM (ed.) *Current Topics in Membranes, Vol. 54: Purinergic Receptors and Signalling*, pp. 307–368. San Diego, CA: Academic Press.

Burnstock G (2004) Cotransmission. *Current Opinion in Pharmacology* 4: 47–52.

Burnstock G (2006) Pathophysiology and therapeutic potential of purinergic signalling. *Pharmacological Reviews* 58: 58–86.

Burnstock G (2007) Physiology and pathophysiology of purinergic neurotransmission. *Physiological Reviews* 87: 659–797.

Burnstock G and Costa M (eds.) (1975) Adrenergic neuroeffector transmission. In *Adrenergic Neurones: Their Organisation, Function and Development in the Peripheral Nervous System*, pp. 51–106. London: Chapman and Hall.

Burnstock G and Knight G (2004) Cellular distribution and functions of P2 receptor subtypes in different systems. *International Review of Cytology* 240: 31–304.

Dunn PM, Zhong Y, and Burnstock G (2001) P2X receptors in peripheral neurones. *Progress in Neurobiology* 65: 107–134.

James G and Butt AM (2002) P2Y and P2X purinoceptor mediated Ca^{2+} signalling in glial cell pathology in the central nervous system. *European Journal of Pharmacology* 447: 247–260.

Kennedy C, McLaren GJ, Westfall TD, and Sneddon P (1996) ATP as a co-transmitter with noradrenaline in sympathetic nerves – Function and fate. In: Chadwick DJ and Goode J (eds.) *P2 Purinoceptors: Localization, Function and Transduction Mechanisms*, pp. 223–235.

Sneddon P and Burnstock G (1984) Inhibition of excitatory junction potentials in guinea-pig vas deferens by α, β-methylene-ATP: Further evidence for ATP and noradrenaline as cotransmitters. *European Journal of Pharmacology* 100: 85–90.

Zimmermann H (2001) Ectonucleotidases: Some recent developments and a note on nomenclature. *Drug Development Research* 52: 44–56.

Adenosine Triphosphate (ATP) as a Neurotransmitter

A Verkhratsky, The University of Manchester, Manchester, UK
O Krishtal, Bogomoletz Institute of Physiology, Kiev, Ukraine

© 2009 Elsevier Ltd. All rights reserved.

Historic Remarks

The molecule of adenosine 5′-triphosphate, ATP, was discovered in 1929 by Karl Lohman in Heidelberg and by Cyrus Hartwell Fiske and Yellapragada SubbaRow at Harvard. In the same year, the role for purines and ATP as extracellular signaling molecules was also suggested by Drury and Szent-Györgyi, who found that purines exert a potent negative chronotropic effect on the heart and trigger dilatation of coronary vessels. The signaling function of ATP in peripheral tissues was subsequently confirmed by numerous experiments.

In 1959 Pamela Horton made a seminal discovery that ATP can be released from nerves upon electrical stimulation. In the 1960s nonadrenergic, noncholinergic inhibitory transmission was described in the taenia coli and stomach, activity that subsequently was demonstrated to be mediated by ATP. These experimental observations led Geoffrey Burnstock to suggest a distinct type of 'purinergic' neurotransmission. Meanwhile, the first indications of excitatory action of extracellular ATP on central neurons were obtained in experiments on neurons from the nucleus cuneatus. Further indications that ATP may act as a neurotransmitter in neuronal–neuronal synapses were obtained by Thomas White, who noted the release of ATP from brain synaptosomes treated by veratridine or high extracellular K^+; he subsequently found depolarization-induced ATP release from synaptosomes isolated from cortex and striatum and also demonstrated that in central nervous tissue, ATP was not co-released with acetylcholine (ACh).

Investigations of purinergic nerves, primarily led by Burnstock, resulted in discovery of specific family of purinoreceptors, which are classified as two fundamentally distinct classes, the P1 (A1) receptors, sensitive to adenosine, and the P2 receptors, activated by ATP and its analogs. The P2 receptors are represented by the ionotropic P2X and metabotropic P2Y receptors. The P2X receptors are ligand-gated cationic channels, which are permeable to Na^+, K^+, and Ca^{2+}. Seven P2X receptor subunits ($P2X_1$–$P2X_7$) encoded by distinct genes have been identified. These subunits may form homo- or heteromeric receptors, with each functional receptor containing at least three monomers. The P2Y receptors are classical seven-transmembrane-domain metabotropic receptors coupled to G-proteins. These receptors are represented by at least 14 subtypes. Activation of neuronal P2Y receptors regulates K^+ channels, potentiates high-voltage-activated Ca^{2+} channels, and triggers inositol 1,4,5-trisphosphate (IP_3)-mediated release of Ca^{2+} from endoplasmic reticulum Ca^{2+} stores.

ATP-Mediated Transmission in the Peripheral Nervous System

Peripheral Nerves

ATP acts as a nerve-to-effector transmitter in many tissues. As a rule, ATP is co-released with other neurotransmitters – for example, with noradrenalin (NA) or ACh. This co-release is well documented for the sympathetic nervous system. ATP mediates neuroeffector transmission between the hypogastric nerve and vas deferens, being in fact the main mediator triggering the excitatory junction potential and initiating contraction of smooth muscle. Both ATP and NA, co-released from the same terminals, are instrumental in sympathetic control of vascular tone. The relative fraction and hence the relative functional importance of purinergic component will vary among different types of vessels, the most significant activity occurring in the arterioles of the mesentery and submucosal plexus of the intestine; here ATP acts as a main excitatory transmitter, whereas NA, which is co-released with ATP, modulates the release of the latter. In the parasympathetic system, ATP acts as an excitatory cotransmitter together with ACh in the terminal innervating the urinary bladder. Similarly, ATP is co-released with ACh from motor nerve terminals. In these terminals ATP mostly potentiates the excitatory action of ACh, acting through postjunctional metabotropic receptors. In the neuromuscular junction, ATP also appears to be a source of adenosine, which, by activating prejunctional P1(A1) receptors, regulates release of ACh. Notably, however, P2X receptors are found in immature skeletal muscle cells, and several P2X receptors subunits have been identified in developing myocytes. Therefore, P2X-mediated excitatory transmission is likely operative during yet unknown developmental steps.

Sensory Neurons

ATP-induced currents, mediated through opening of ionotropic P2X receptors, were initially recorded from neurons acutely isolated from nodose, vestibular, trigeminal, and dorsal root ganglia (DRG). These currents were concentration dependent (with K_D for ATP ~5 μM) and were carried mostly by sodium ions, as the substitution of the latter by tris(hydroxymethyl)aminomethane ($Tris^+$), tetraethylammonium (TEA^+), or $choline^+$ markedly suppressed the ATP-induced responses (Figure 1). This initial observation was confirmed and extended by many groups, and purinoreceptor-mediated currents in sensory neurons have been characterized in minute detail. Analysis of mRNA expression reveals the presence of transcripts for $P2X_{1-6}$ receptors, with $P2X_3$ mRNA showing the highest level. Similarly, immunohistochemical analysis of sensory neurons in DRG shows the appearance of $P2X_{1-6}$ subunits, with clear predominance of the $P2X_3$ subunit and significant expression of the $P2X_2$ subunit. Both $P2X_3$ and $P2X_2$ subunits are mostly confined to small and medium-size sensory neurons putatively executing nociceptive functions. Expression of various P2X subunits is reflected by heterogeneity of ATP-induced currents evoked in sensory neurons. Some authors suggest the participation of P2X receptors in the functioning of the synapses between sensory neurons and the neurons of dorsal horn. Sensory neurons also express metabotropic $P2Y_1$ receptors, which mediate long-lasting depolarization.

Autonomic Neurons

Sympathetic neurons express $P2X_{1-6}$ purinoreceptor subunits, with a particularly high presence of $P2X_2$ subunits. External application of ATP to cultured celiac neurons triggers depolarization and inward ion currents, which are blocked by suramin and reactive blue 2. Synaptic currents recorded from these neurons display similar properties, suggesting the involvement of P2X receptors. In parasympathetic neurons, ATP has an excitatory action mediated through ATP-induced currents, which have been recorded

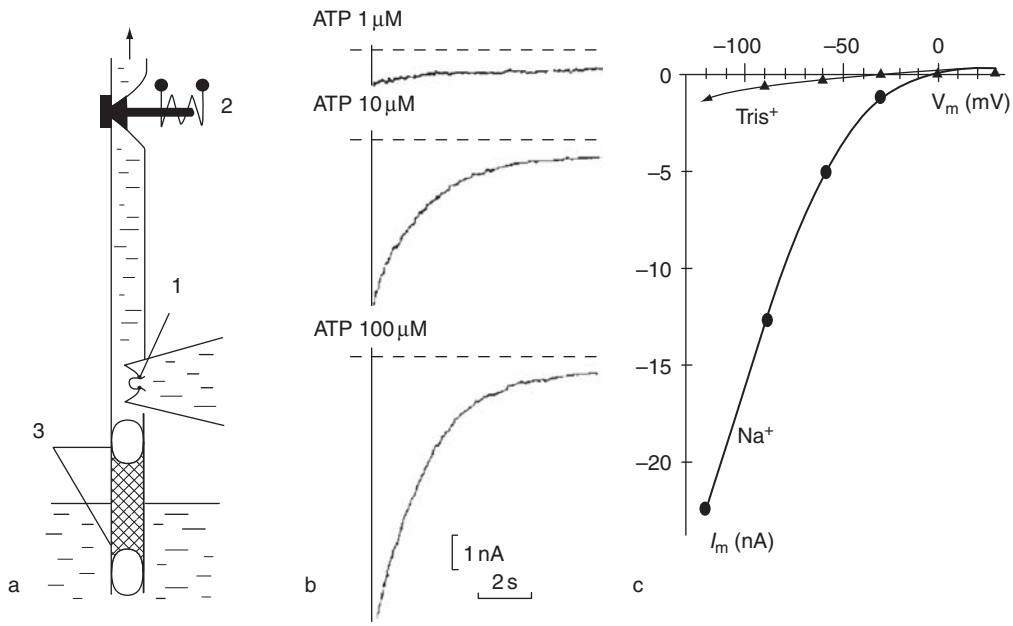

Figure 1 First recordings of ATP-induced membrane currents in acutely isolated sensory neurons. (a) The 'square pulse' application technique used for rapid application of ATP to internally dialyzed sensory neurons. The tip of the micropipette (with the cell; 1) is inserted into a plastic tube. An invagination in the tip of the micropipette prevents the cell from damage. The lower end of the tube can be exposed to different external solutions or to air. Suction applied to the upper end of the tube (arrow) is controlled by an electromagnetic valve (2). A preprogrammed sequence of current pulses applied to the valve allows a column of test solution to form in the tube; the test solution is separated from the normal solution by air bubbles (3). Another sequence of pulses exposes the cell to the test solution for the desired period of time by rapid displacement of the column along the tube. The electrical recording is unharmed, since there is a thin layer of saline between the air bubbles and the walls of the tube. (b) ATP-activated inward current (ordinate) elicited by application of different concentrations of ATP as indicated on the graph. (c) Voltage dependence of the peaks of ATP-activated current. ATP (5 μM) was applied in normal saline (●) and in an external solution in which Na was substituted with tetraethylammonium (▲). Holding potential was −90 mV throughout. Reproduced from Krishtal OA, Marchenko SM, and Pidoplichko VI (1983) Receptor for ATP in the membrane of mammalian sensory neurones. *Neuroscience Letters* 35: 41–45, with permission.

from variety of neurons isolated from different ganglia (e.g., intramural ganglia from the guinea pig urinary bladder, chick ciliary ganglia, and rabbit vesical parasympathetic ganglia). These ionotropic purinergic responses are mostly mediated through receptors constructed from $P2X_{2/3}$ subunits. Finally, a majority of myenteric neurons from various species demonstrate both P2X current responses and P2Y/IP_3-mediated Ca^{2+} signals.

ATP-Mediated Synaptic Transmission in the Central Nervous System

P2X-Mediated Synaptic Currents

All types of P2X receptors ($P2X_{1-7}$) and many P2Y receptors ($P2Y_1$, $P2Y_6$, $P2Y_{11}$, $P2Y_{12}$, $P2Y_{13}$, and $P2Y_{14}$) are expressed in the brain. Several types of P2X subunits can be co-expressed in the same neuron, and this expression can be quite heterogeneous even within the same brain region. As a consequence, ATP-mediated currents recorded from central nervous system (CNS) neurons display heterogeneous kinetics and sensitivity to pharmacological modulators (Figure 2).

In the CNS, the ATP-induced excitatory responses were initially demonstrated in cultured dorsal horn neurons. Ten year later, the ATP-mediated synaptic transmission was discovered in brain slices of medial habenula, where both excitatory postsynaptic currents and spontaneous 'miniature' postsynaptic currents were identified (Figure 3(a)). Subsequently, ATP-mediated synaptic responses were found in other brain areas, including spinal cord, hippocampus, and

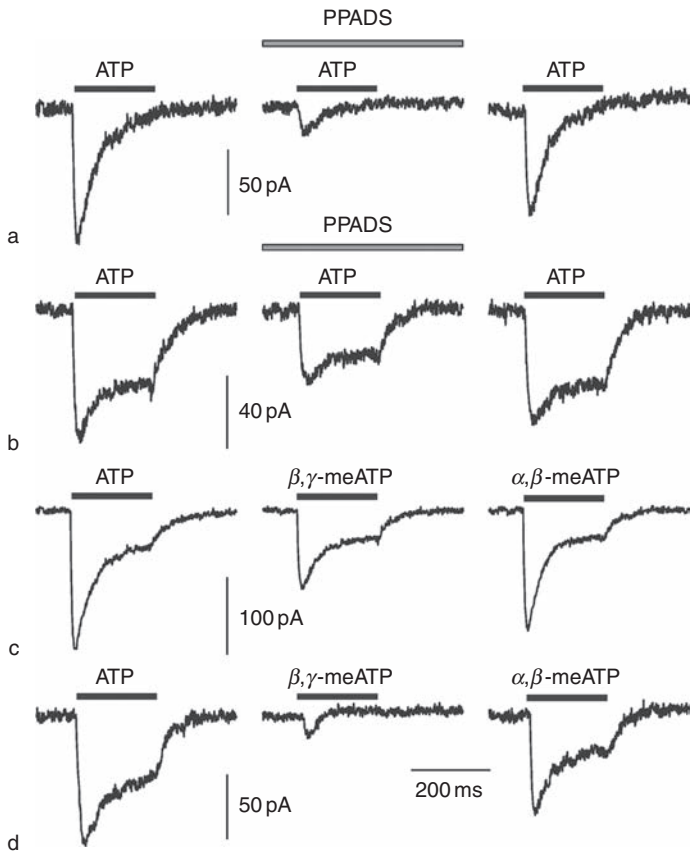

Figure 2 Heterogeneity of P2X receptor-mediated currents in the isolated pyramidal neocortical layer V neurons. (a, b) Examples of inward currents evoked by short (200 ms) applications of ATP (20 μM). The traces (from left to right) represent control, response to ATP under action of pyridoxal phosphate-6-azophenyl-2′,4′-disulfonic acid (PPADS; 30 μM), and response to ATP after washout of drug. (c, d) Examples of currents evoked by application of various agonists of the P2X receptor. The traces (from left to right) represent response to ATP (20 μM), response to β,γ-methylene-ATP (25 μM), and response to α,β-methylene-ATP (20 μM). Groups of traces (a, b; and c, d) were recorded from different cells. Recordings were made with a 5 min time interval between applications at the holding potential of –80 mV. Reproduced from Pankratov Y, Lalo U, Krishtal O, et al. (2003) P2X receptor-mediated excitatory synaptic currents in somatosensory cortex. *Molecular and Cellular Neuroscience* 24: 842–849, with permission.

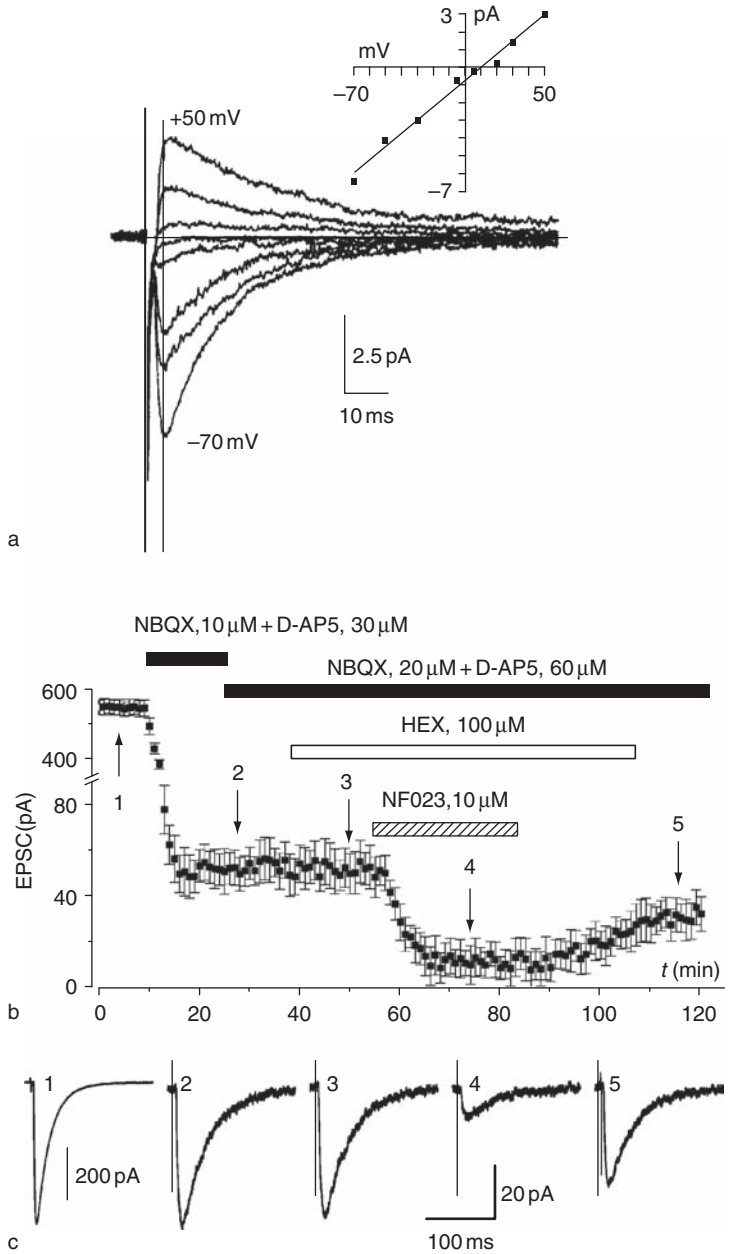

Figure 3 ATP-mediated synaptic transmission in CNS neurons. (a) Voltage dependence of ATP-mediated synaptic currents, recorded from a 24-day-old male rat medial habenula neuron in Krebs solution containing 5 mM Ca^{2+}. Stimulus pulses (13 V, 200 μs) were delivered at 2 Hz, and 53–215 currents were recorded before the holding potential was stepped to the next voltage. The traces are averages of all responses (including failures). The peaks of the currents shown are plotted in the inset and fitted to a straight line using a linear regression. Peak current was measured from the single exponential fit of each current at the time indicated by the vertical line. The membrane potentials shown in this plot are −70, −50, −30, −5, +5, +20, +30, and +50 mV. Note that the voltage dependence of peak current amplitude is well fitted by a straight line, with clear outward currents at positive holding potentials. (b) Dissection of the ATP-mediated component of excitatory postsynaptic currents (EPSCs) in pyramidal neurons of layer II of the somatosensory cortex. (Top) Changes in amplitude of the EPSC following bath application of the glutamatergic antagonists 2,3-dioxo-6-nitro-1,2,3,4-tetrahydrobenzo[*f*]quinoxaline-7-sulfonamide (NBQX) and D-(−)-2-amino-5-phosphopentanoic acid (D-AP5), the cholinergic antagonist hexamethonium (HEX), and the P_{2X} receptor antagonist NF023 at various time points (1–5) . Each point represents the mean ± SD for six sequential trials (holding potential, −80 mV; stimulation frequency, 0.1 Hz). (Bottom) Examples of the EPSCs (average of six traces) recorded at moments 1–5 as indicated on the upper graph. (a) Reproduced from Edwards FA, Robertson SJ, and Gibb AJ (1997) Properties of ATP receptor-mediated synaptic transmission in the rat medial habenula. *Neuropharmacology* 36: 1253–1268, with permission. (b) Reproduced from Pankratov Y, Lalo U, Krishtal O, et al. (2002) Ionotropic P2X purinoreceptors mediate synaptic transmission in rat pyramidal neurones of layer II/III of somato-sensory cortex. *Journal of Physiology* 542: 529–536, with permission.

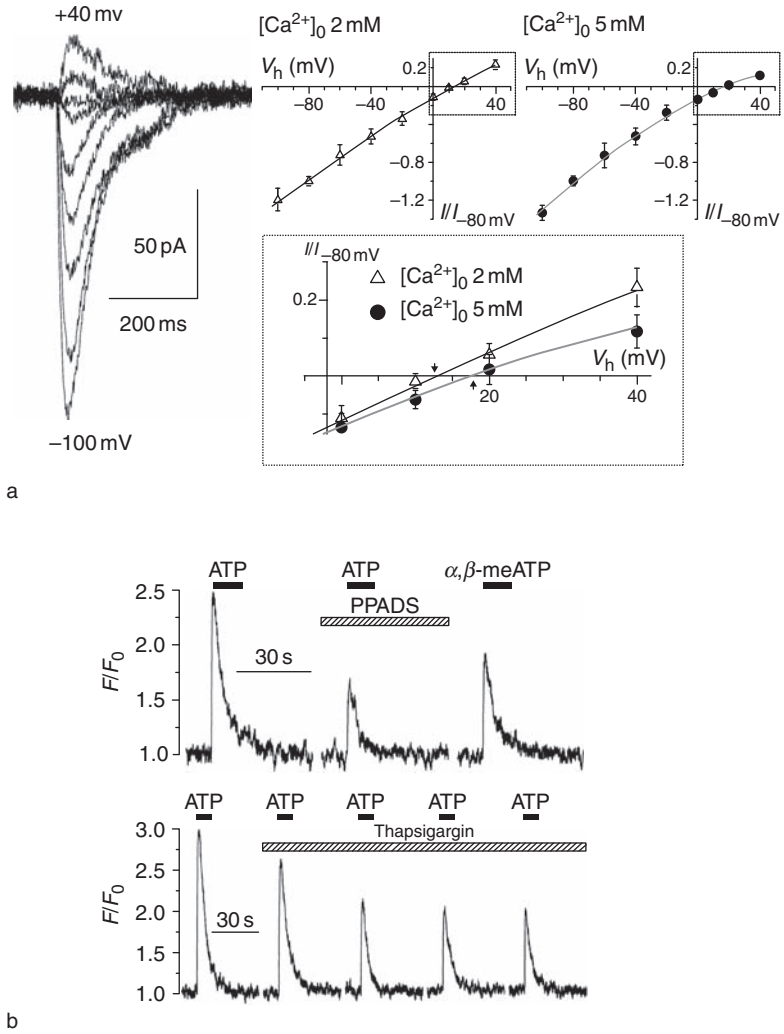

Figure 4 Calcium permeability of P2X receptors in acutely isolated pyramidal cortical neurons. (a) Left panel shows examples of inward currents induced in the pyramidal neurons of somatosensory cortex layer II/III by the application of ATP (10 μM) at membrane potentials ranging from −100 to +40 mV. Shown on the right are voltage–current relationships for the ATP-induced currents measured at different extracellular calcium concentrations. Amplitudes of currents were normalized to the maximal value measured at −80 mV. Each point is the mean ± SD for seven cells. Lines represent the cubic polynomial fit. Lower inset demonstrates the superposition of voltage–current relationships in the vicinity of reversal potential. (b) ATP-induced $[Ca^{2+}]_i$ transients in CA1 pyramidal neurons. Examples in the upper graph show $[Ca^{2+}]_i$ transients induced in the pyramidal cells by fast application of ATP (100 μM) and α,β-methylene-ATP (100 μM) to the hippocampal slice in the control and after bath application of 20 μM pyridoxal phosphate-6-azophenyl-2′,4′-disulfonic acid (PPADS). All traces were recorded in the same cell at 5 min intervals. Note the substantial amplitude of the response to α,β-methylene-ATP, which is attributable only to the calcium influx via ionotropic ATP receptors. Examples in the lower graph show $[Ca^{2+}]_i$ transients induced by repetitive fast application of ATP (100 μM) in the presence of 1 μM thapsigargin at 5 min time intervals. It is worth noting that the amplitudes of the third and the following transients, which are attributable to entry of the extracellular calcium via ionotropic ATP receptors, comprise ∼40% of the initial response, which represents the combined activity of P2X and P2Y purinoreceptors. (a) Reproduced from Pankratov Y, Lalo U, Krishtal O, et al. (2002) Ionotropic P2X purinoreceptors mediate synaptic transmission in rat pyramidal neurones of layer II/III of somato-sensory cortex. *Journal of Physiology* 542: 529–536, with permission. (b) Reproduced from Pankratov YV, Lalo UV, and Krishtal OA (2002) Role for P2X receptors in long-term potentiation. *Journal of Neuroscience* 22: 8363–8369, with permission.

cortex. The purinergic nature of the synaptic responses in central neurons was usually confirmed based on neuronal sensitivity to P2X receptors inhibitors or by desensitization induced following incubation with α,β-methylene-ATP (α,β-meATP). An example of the analysis of evoked postsynaptic currents mediated by P2X receptors is shown in Figure 3(b).

As a rule, P2X-mediated synaptic currents are not very large, rarely exceeding 50–100 pA in amplitude and representing 5–15% of the synaptic currents

mediated by glutamate. Nonetheless, ATP-mediated synaptic transmission can have functional importance, especially considering that activation of postsynaptic P2X receptors can provide a route for Ca^{2+} entry at resting membrane potentials. This is in contrast to another main source for Ca^{2+} entry in the postsynaptic density, that mediated by *N*-methyl-D-aspartate (NMDA) receptors; activation of NMDA

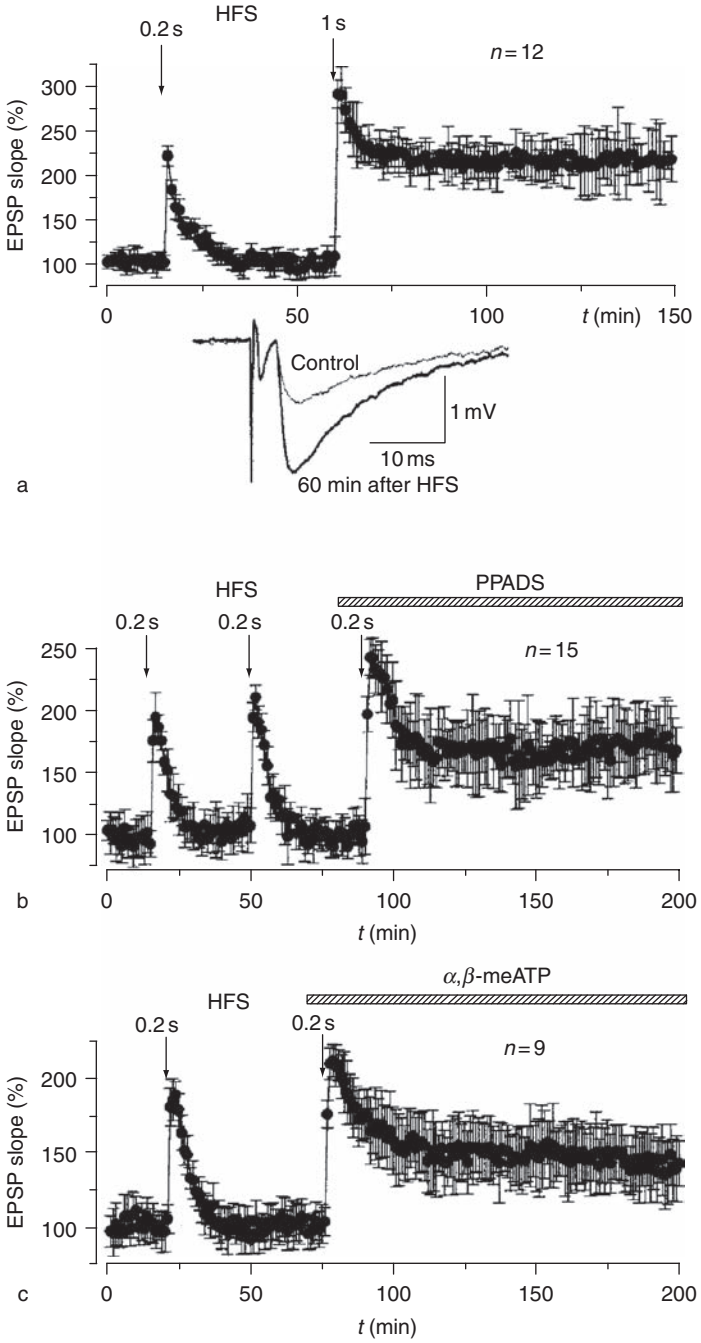

Figure 5 Inhibition of P2X receptors facilitates long-term potentiation initiation in the CA1 region of the hippocampus: changes in CA1 field potentials induced by 100 Hz high-frequency stimulation (HFS) delivered to the Schaffer collateral in the control and after inhibition of the ATP receptors. (a) The time course of the potentiation evoked by 0.2 and 1 s of 100 Hz stimulation trains in control conditions. Examples of field excitatory postsynaptic potentials (EPSPs) recorded before and 60 min after the 1-s-long HFS train are indicated in the inset. Each trace represents the average of ten EPSPs. (b, c)The time course and magnitude of the potentiation evoked by the 0.2 s HFS train in control and after bath application of 20 μM pyridoxal phosphate-6-azophenyl-2′,4′-disulfonic acid (PPADS) and α,β-methylene-ATP, respectively. Baseline stimulation frequency is 0.08 Hz. Reproduced from Pankratov YV, Lalo UV, and Krishtal OA (2002) Role for P2X receptors in long-term potentiation. *Journal of Neuroscience* 22: 8363–8369, with permission.

receptors requires cell predepolarization, which removes the Mg^{2+} block. Functional P2X receptors expressed in the CNS display substantial Ca^{2+} permeability; Ca^{2+} permeability, relative to monovalent cations, can range between 2 and 12 (Figure 4 (a)). Application of ATP triggers substantial Ca^{2+} signals in central neurons; the signals are mediated by Ca^{2+} entry through P2X receptors/voltage-gated Ca^{2+} channels and by P2Y-mediated Ca^{2+} release from intracellular stores (Figure 4(b)). Ca^{2+} signals induced by stimulation of P2 receptors are functionally important; for example, these signals may modulate NMDA receptors and facilitate long-term potentiation in CA1 hippocampal neurons (Figure 5).

Mechanisms of ATP Release

The overall concentration of ATP in the brain is rather high, varying between 2 and 4 mM in different regions. The cellular membranes are not permeable to ATP, therefore ATP-mediated intercellular signaling requires specialized pathways for ATP release. There are several of these pathways, including both exocytotic release from presynaptic vesicles and non-exocytotic release through plasmalemmal pores.

Vesicular ATP release is widespread throughout the peripheral and CNSs. ATP can be accumulated and stored in synaptic vesicles together with other neurotransmitters and then both can be co-released. This co-release of ATP with other neurotransmitters is abundant in the peripheral nervous system (PNS); in the CNS, ATP can be co-released with γ-aminobutyric acid (GABA; as seen in cultured neurons from spinal cord and lateral hypothalamus) or with glutamate (as seen in hippocampal organotypic slices). Early studies of ATP release from synaptosomal preparations also indicated that some fraction of ATP could be co-released with ACh or NA, but this co-release is probably of minor importance. Alternatively, ATP may be released on its own from specific ATP-containing vesicles, which can be present in specific ATP-containing terminals or in shared glutamatergic terminals (Figure 6).

Nonvesicular ATP release may occur through gap junction hemichannels or via volume-sensitive chloride channels, or even through $P2X_7$ receptors, known to form large transmembrane pores following stimulation. These mechanisms are predominantly associated with glial cells.

ATP Mediates Neuronal–Glial Signaling

Purinergic transmission is particularly important for neuronal–glial integration in the CNS. All types of

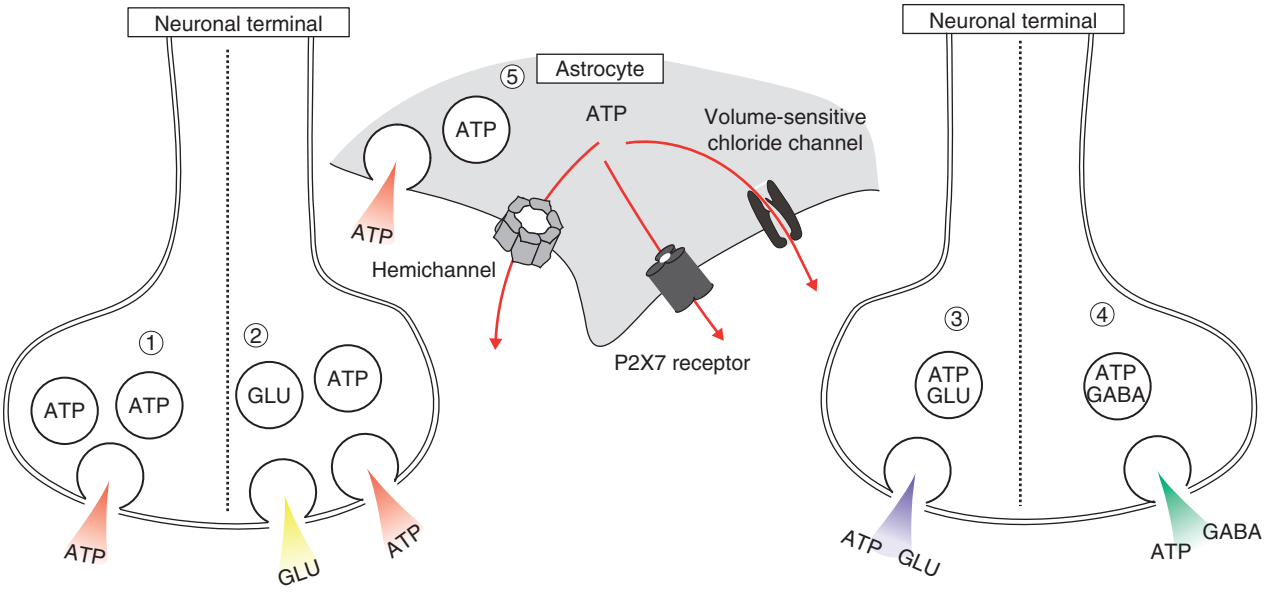

Figure 6 Adenosine triphosphate (ATP) release pathways in the CNS, where ATP can be stored and released from the synaptic terminals in several distinct fashions. (1) ATP can act as a sole transmitter (e.g., in terminals from medial habenula). (2) Vesicles containing ATP can coexist with vesicles containing glutamate (GLU) in the same terminals; release of ATP occurs simultaneously with but independently from the release of glutamate (this mechanism is, for instance, operative in the hippocampus). (3) In principle, ATP can be co-stored and co-released with glutamate or other neurotransmitters, although this is not supported by direct experimental evidence from *in situ* preparations. (4) ATP can be co-stored and co-released with γ-aminobutyric acid (GABA; e.g., in cultured spinal or lateral hypothalamic neurons). (5) ATP can be also released from glial cells via regulated exocytosis or through plasmalemmal channels.

glia (i.e., the astrocytes, oligodendrocytes, and microglia in the CNS and Schwann cells in the PNS) express functional ATP receptors. Application of exogenous ATP triggers Ca^{2+} responses in both cultured glia and glial cells in *in situ* preparations. These Ca^{2+} signals are usually driven through the metabotropic route, mediated by P2Y receptors and IP_3-induced Ca^{2+} release from intracellular stores. The functional role of P2X receptors in macroglia is much less clear: although P2X-mediated ion currents are detected in cultured astrocytes, ATP-mediated ionotropic responses *in situ* remain mostly uncharacterized. Reverse transcriptase–polymerase chain reaction (RT-PCR) revealed expression of mRNA for all types ($P2X_{1-7}$) of receptors in astroglial cells in nucleus accumbens and in all receptors save $P2X_6$ in astrocytes from cortex. Functional expression of P2X receptors was found in oligodendrocytes from the optic nerve.

ATP acts as a mediator not only in neuronal–glial communications but also in glial–glial and potentially glial–neuronal signaling. ATP is generally acknowledged as a 'glio' transmitter, and regulated release of ATP from astrocytes has been observed in several glial preparations. ATP released by the glial cells is instrumental for generation and maintenance of propagating Ca^{2+} waves, which are the substrate for glial excitability and long-range signaling. The ATP released from glial cells can also signal to neurons either by activating neuronal P2 receptors, or, following degradation to adenosine, by stimulating neuronal A1 receptors.

ATP-mediated signaling is particularly important for pathological reactions of glial cells. Astroglial cells express $P2X_7$ receptors, which can be activated following massive release of ATP upon brain injury; stimulation of P2X7 receptors may be involved in initiation of reactive gliosis. Similarly, massive release of ATP can act as a specific signal for activation of microglia; microglial cells express an extended repertoire of P2X receptors (which can be substantially modified in the course of microglial activation) and P2Y receptors linked to intracellular Ca^{2+} signaling. The P2 receptors control motility of microglial processes, and upon overstimulation they trigger rapid movement of these processes toward the site of injury.

Conclusions

ATP released from neuronal terminals and from astroglial cells mediates signaling within neuronal–glial networks. Effects of ATP are diverse and are determined by idiosyncrasies of ionotropic (P2X) and metabotropic (P2Y) receptors variably expressed in target cells. In pathological conditions, ATP may act as an important signal, triggering and controlling defensive reactions of microglia and reactive astrogliosis.

See also: Adenosine; Adenosine Triphosphate (ATP); Adenosine Receptor Mediated Functions; Calcium Waves: Purinergic Regulation; P2X Receptors; Pharmacology of Sleep: Adenosine; Presynaptic Receptor Signaling; Purinergic Receptors; Purines and Purinoceptors: Molecular Biology Overview; Synaptic Plasticity: Short-Term Mechanisms.

Further Reading

Abbracchio MP, Burnstock G, Boeynaems JM, et al. (2006) International Union of Pharmacology LVIII: Update on the P2Y G protein-coupled nucleotide receptors: From molecular mechanisms and pathophysiology to therapy. *Pharmacological Reviews* 58: 281–341.

Burnstock G (1972) Purinergic nerves. *Pharmacological Reviews* 24: 509–581.

Burnstock G (2003) Purinergic receptors in the nervous system. *Current Topics in Membranes* 54: 307–368.

Burnstock G, Campbell G, Satchell D, et al. (1970) Evidence that adenosine triphosphate or a related nucleotide is the transmitter substance released by non-adrenergic inhibitory nerves in the gut. *British Journal of Pharmacology* 40: 668–688.

Collo G, North RA, Kawashima E, et al. (1996) Cloning of P2X5 and P2X6 receptors and the distribution and properties of an extended family of ATP-gated ion channels. *Journal of Neuroscience* 16: 2495–2507.

Drury AN and Szent-Györgyi A (1929) The physiological activity of adenine compounds with special reference to their action upon mammalian heart. *Journal of Physiology (London)* 68: 213–237.

Edwards FA, Gibb AJ, and Colquhoun D (1992) ATP receptor-mediated synaptic currents in the central nervous system. *Nature* 359: 144–147.

Edwards FA, Robertson SJ, and Gibb AJ (1997) Properties of ATP receptor-mediated synaptic transmission in the rat medial habenula. *Neuropharmacology* 36: 1253–1268.

Fields RD and Burnstock G (2006) Purinergic signalling in neuron–glia interactions. *Nature Reviews Neuroscience* 7: 423–436.

Fiske CH and SubbaRow Y (1929) Phosphorous compounds of. muscle and liver. *Science* 70: 381–382.

Holton P (1959) The liberation of adenosine triphosphate on antidromic stimulation of sensory nerves. *Journal of Physiology* 145: 494–504.

Jahr CE and Jessell TM (1983) ATP excites a subpopulation of rat dorsal horn neurones. *Nature* 304: 730–733.

Khakh BS, Burnstock G, Kennedy C, et al. (2001) International union of pharmacology. XXIV. Current status of the nomenclature and properties of P2X receptors and their subunits. *Pharmacological Reviews* 53: 107–118.

Krishtal OA, Marchenko SM, and Pidoplichko VI (1983) Receptor for ATP in the membrane of mammalian sensory neurones. *Neuroscience Letters* 35: 41–45.

Lohmann K (1929) On the pyrophosphate fraction in muscle. *Naturwissenschaften* 17: 624–625.

North RA and Verkhratsky A (2006) Purinergic transmission in the central nervous system. *Pflugers Archives* 452: 479–485.

Pankratov Y, Lalo U, Krishtal O, et al. (2002) Ionotropic P2X purinoreceptors mediate synaptic transmission in rat pyramidal neurones of layer II/III of somato-sensory cortex. *Journal of Physiology* 542: 529–536.

Pankratov YV, Lalo UV, and Krishtal OA (2002) Role for P2X receptors in long-term potentiation. *Journal of Neuroscience* 22: 8363–8369.

Pankratov Y, Lalo U, Krishtal O, et al. (2003) P2X receptor-mediated excitatory synaptic currents in somatosensory cortex. *Molecular and Cellular Neuroscience* 24: 842–849.

Pankratov Y, Lalo U, Verkhratsky A, et al. (2006) Vesicular release of ATP at central synapses. *Pflugers Archives* 452: 589–597.

Pascual O, Casper KB, Kubera C, et al. (2005) Astrocytic purinergic signaling coordinates synaptic networks. *Science* 310: 113–116.

Silinsky EM and Gerzanich V (1993) On the excitatory effects of ATP and its role as a neurotransmitter in coeliac neurons of the guinea-pig. *Journal of Physiology* 464: 197–212.

Suadicani SO, Brosnan CF, and Scemes E (2006) $P2X_7$ receptors mediate ATP release and amplification of astrocytic intercellular Ca^{2+} signaling. *Journal of Neuroscience* 26: 1378–1385.

White TD (1978) Release of ATP from a synaptosomal preparation by elevated extracellular K^+ and by veratridine. *Journal of Neurochemistry* 30: 329–336.

Adolescent Brain Development and the Risk of Psychiatric Disorders

J N Giedd, National Institute of Mental Health, Bethesda, MD, USA

© 2009 Published by Elsevier Ltd.

Introduction

Adolescence, the transition from childhood to adulthood, usually occurs in the second decade of life and is characterized by tumultuous changes in the body, brain, and behavior. The behavioral changes, famously characterized by GS Hall in 1904 as a time of *sturm und drang* (storm and stress), have been a prominent theme in art and literature. Across cultures and millennia, adolescence has been noted as a time of heightened passions, mood volatility, sensation seeking, risk taking, and conflict with parents.

Although most teens successfully navigate the transition from being dependent upon a caregiver to self-sufficiency, adolescence is a time of peak onset for the emergence of several classes of neuropsychiatric illnesses including anxiety, bipolar disorder, depression, eating disorder, psychosis, and substance abuse.

Advances in neuroimaging technology have created an unprecedented opportunity to explore the neurobiological substrates of these social, cognitive, behavioral, and pathological changes. In this article, we summarize adolescent brain changes, discuss the possible relationship to the emergence of psychopathology, and outline possible directions for future research.

Neurobiological Changes of Adolescence

Postmortem Data

Human In a series of postmortem studies conducted over the past 25 years, the lab of Peter Huttenlocher has examined over 50 autopsy brains spanning an age range from 28-week-old fetuses to 90-year-old adults. This work has provided seminal information about the characteristic of synapses across the life span, although only four of the specimens are from subjects in the adolescent age range. The investigations have targeted three different brain areas (frontal lobes, occipital, and auditory cortex) allowing assessment of regional variation. In the frontal cortex, synaptic density is similar to adult levels at birth, rises to 150–200% of adult levels during early childhood, and then declines from ages 7 to 16 years. The rate of decline for frontal synaptic density was noted to be protracted compared to auditory and occipital areas which stabilized at 11–12 years.

Primate Because of the paucity of human adolescent postmortem data, much of the speculation regarding adolescent brain changes has been inferred from studies of nonhuman primates. Studies of macaque monkeys indicate a striking decrease in cortical synapses during puberty with a loss of up to 30 000 synapses per second. This leads to a 50% reduction in average number of synaptic contacts per neuron compared to the prepubertal state.

Neurotransmitter Systems

Animal studies also demonstrate dynamic changes in neurotransmitter systems throughout childhood and adolescence. Dopamine receptors, like synapses, follow a developmental trajectory of low levels at birth, maximum density during prepuberty, followed by selective pruning during adolescence. The pattern is nuanced by sex and brain region. For instance, in rat striatum prepubertal dopamine receptor overproduction and adolescent pruning is much more pronounced in males than females and concurrent prepubertal D1 receptor overproduction in the nucleus accumbens is not followed by adolescent pruning. Consistent with the relatively late maturation of the prefrontal cortex in humans, D1 receptor pruning in rat prefrontal cortex does not begin until early adulthood, but then the pruning proceeds more robustly than in the striatum. The changing balance between cortical and subcortical dopamine activity may underlie the motivational, appetitive, and novelty seeking changes observed during adolescence.

Serotonin systems, which reciprocally interact with dopamine systems to modulate many behaviors, also display dynamic changes during adolescence with a study of male human postmortem brains demonstrating dramatically reduced 5HT1a binding during adolescence.

Receptor bindings for excitatory glutamate and inhibitory GABA neurotransmitters are higher during prepuberty than adulthood in rats and in nonhuman primates. In macaque monkeys, biochemical markers in a subtype of GABA neurons called chandelier cells, which connect to initial axon segments of multiple pyramidal cells to regulate pyramidal cell firing patterns within cortical columns, also follow the peak at puberty decline during adolescence pattern. Biochemical markers in another subtype of GABA neurons called wide arbor cells, which regulate firing of pyramidal cells in neighboring columns, increase steadily throughout from infancy to adulthood. Distribution of postsynaptic GABA receptor subtypes also undergoes

substantial changes throughout pediatric development making this neurotransmitter system involved in working memory and other cognitive functions particularly volatile during adolescence.

Although developmental trajectories of these various neurotransmitters do not overlap precisely, the general pattern of increases during childhood followed by selective decreases during adolescence seems to hold. The relationship between adolescent cognitive and behavioral changes and these dynamically changing and interacting neurotransmitter systems is only beginning to be elucidated.

Structural Neuroimaging

Most of the information regarding brain anatomy in adolescents comes from magnetic resonance imaging (MRI) studies. MRI is particularly well suited for *in vivo* pediatric studies, because unlike conventional X-rays and computerized axial tomography it does not use ionizing radiation. This allows not only the initial scanning of children and adolescent but also of repeated scans over time.

The spatial resolution of MRI scans is constantly improving and greater resolution can be purchased with the currency of time in the scanner, but in current studies the usual size of the smallest volume with a single MRI signal value (i.e., voxel) is approximately 1 ml. This is worth noting with respect to interpretation of the MRI findings, because within a given 1 ml voxel there may be millions of neurons and trillions of synaptic connections. For example, one unit of the mouse cerebral cortex contains about 30% axons, 30% dendrites, 14% cell bodies, 12% dendritic spines, and 9% glia. However, given that a single MR voxel of cortical gray-matter in the human brain contains a mixture of cell types, it is impossible, at this point, to attribute the observed effects to any single cellular compartment.

White matter volumes increase throughout childhood and adolescence, although rates of increase vary by region and age. This increase in volume is thought to be due to ongoing myelination, the wrapping of an insulating material around axons by oligodendrocytes, which serves to increase greatly the speed of neuronal signal transmission and facilitates the integration of distributed neural networks.

An MR technology called diffusion tensor imaging (DTI), which enables assessment of directionality of fiber tracts, is being used to characterize further white matter development during childhood and adolescence. Pediatric DTI studies have begun to explore relationships between white matter development and cognitive capacities.

Gray matter volumes follow a distinctly different developmental path following an inverted U shaped trajectory. Gray matter volume trajectories vary by region but peak during adolescence for frontal, temporal, and parietal lobes. Particularly late to reach adult levels of cortical thickness are areas in the prefrontal cortex and posterior superior temporal region. Although specifics of the relationships between the anatomic changes and cognitive or emotional changes have not been well elucidated, discussions regarding the implications of the relatively late anatomical maturation of the prefrontal cortex (a key component of circuitry involved in judgment, decision making, and impulse control) have prominently entered social, educational, and judicial realms.

An important, but unresolved, question is the extent to which these gray matter volume changes reflect intracortical myelination versus dendritic and axonal arborization/pruning (the number and extent of neuronal branches). Although arborization/pruning and myelination processes both occur, understanding their relative contributions to the gray matter changes has implications for the conceptualization of brain development in health and illness, the design of future studies, and ultimately as a guide to interventions.

Functional Neuroimaging

Electroencephalography Several large sample electroencephalographic (EEG) studies have been conducted on children and adolescents and consistently demonstrate a cyclic reduction in EEG power during adolescence, particularly for the frontal lobes. The reduction in EEG power continues into the 20s although the rate of reduction is greater in adolescence. EEG studies also indicate profound sleep changes during adolescence with a 50% reduction of deep (stage 4) sleep and a 75% reduction in the peak amplitude of delta waves. EEG signals are thought to be generated by spatially coherent synaptic activity and to possibly reflect changes in synaptic density.

Positron emission tomography Positron emission tomography studies, because of the use of ionizing radiation, have been sparse for children and adolescents. The most cited data are from a study of 29 subjects with epilepsy ranging in age from 5 days to 15 years. Glucose utilization was 30% below adult levels at birth rising to 200% of adult values by the age of 4 years. Glucose utilization remained relatively stable from ages 4 to 10 years and then slowly declined to adult levels during adolescence.

Functional magnetic resonance imaging Functional magnetic resonance imaging (fMRI), which capitalizes on different magnetic properties of oxygenated versus deoxygeneated hemoglobin to assess indirectly blood flow (assumed to be related to neuronal activity), is increasingly being used in pediatric investigations of typically and atypically developing populations.

Although larger sample sizes and longitudinal studies will strengthen the certainty of the results, studies assessing language, social cognition, reward and motivational systems, and executive functions such as working memory and response inhibition have all demonstrated different activation patterns among children and adults. Consistent with the structural MRI findings, associative areas display greater developmental changes than sensorimotor during this developmental period.

A general pattern emerging from these studies is for neural activation during tasks to become less diffuse from childhood to adulthood. This pattern seen over several years in typical development seems to follow that seen for acute learning in adults as task relevant cerebral areas develop enhanced activation while other areas become less active.

Summary of Adolescent Brain Changes

Human postmortem data, animal studies, and neuroimaging all suggest that the tumultuous changes in behavior and cognition are paralleled by equally tumultuous changes in brain anatomy and physiology. The basic process of overproduction followed by selective/competitive elimination that shapes the developing nervous system *in utero* seems to continue to refine the central nervous system throughout adolescent development as prepubertal peaks in the number of synapses, glucose utilization, and neurotransmitter receptors are reduced during the teenage years. Concomitantly, aided by ongoing myelination, distributed brain modules become more and more integrated facilitating greater associative interactions and prefrontal cortex modulation of behavior.

Relationship to Psychopathology

Results of the National Comorbidity Survey Replication study, which entailed in-person household assessments of over 9000 people representative of the United States population (conducted from February 2001 to April 2003), indicate that the peak age of onset for having any mental health disorder is 14 years. Anxiety disorders, bipolar disorder, depression, eating disorder, psychosis, and substance abuse all most commonly emerge during adolescence. A hypothesis is that the emergence of certain psychopathology is related to anomalies or exaggerations of typical adolescent maturation processes.

Schizophrenia

The peak age of onset for schizophrenia is 18 years for males and 25 years for females. A relationship between this age of onset and changes in typical development was first proposed in 1982 by Irvin Feinberg. Based on his interpretation of the dramatic decreases in delta sleep of healthy adolescents as reflecting robust synaptic pruning, Feinberg postulated that schizophrenia may be a consequence of an exaggeration of typical synaptic elimination. Subsequently, several lines of evidence have lent support to this 'exaggeration of typical adolescent changes' hypothesis for schizophrenia. In addition to the exaggerated reductions in adolescent delta sleep, membrane phospholipids, prefrontal metabolism, and frontal cortical gray matter changes are all consistent with an exaggeration of changes seen in typical development. In a rare but phenomenologically similar to adult onset group of subjects with childhood onset schizophrenia (onset prior to age 12 years), the typical frontal gray matter reduction seen in subjects from ages 12 to 17 years was exaggerated fourfold in frontal areas. Direct evidence of a decrease in the number of synapses in schizophrenia, possibly impairing neuronal plasticity, comes from postmortem studies indicating decreased density of synaptic spines, reduction in neuropil, and decreased expression of the synaptic marker synaptophysin.

This notion of an exaggeration of typical adolescent changes is meant to be a further consideration in the neurodevelopmental hypothesis, not a comprehensive account of the etiology of schizophrenia. An increase in prenatal and perinatal adverse events and subtle cognitive, motor, and behavioral anomalies during childhood years before illness onset, all support earlier developmental disturbances underlying the abnormal maturational events during adolescence.

Although there is sparse data for other conditions of adolescent onset, conceptually the notion of anomalies in the dynamic processes of pruning or connectivity may extend to other disorders as well.

Substance Abuse

Across a wide array of mammalian species, adolescents exhibit increased risk taking, novelty seeking, and a greater valuation of social factors. These characteristics not only foster independence from the natal family unit but also increase the risk for harmful behaviors such as substance abuse. The most commonly abused substance is alcohol, and adolescent neurobiology may make them particularly prone to addiction. Evidence from rats indicates that adolescents are relatively insensitive to factors that may limit alcohol intake such as developing motor impairment, getting a 'hangover,' or becoming sedated (perhaps related to immaturity of the developing γ-aminobutyric acid (GABA)-A receptor systems).

Furthermore, the adolescent rat hippocampus is unusually susceptible to ethanol-induced inhibition of long-term potentiation making them more sensitive to the memory-impairing effect of alcohol. The

mechanism for this effect appears to be largely mediated via alcohol's effect on *N*-methyl-D-aspartate (NMDA) receptors and juvenile-greater-than-adult impact occurs at the single cell level, is not confined to the hippocampus, and occurs at alcohol concentrations as low as from a single drink (5 mM).

Morphometric studies of humans suggest that some neural alterations predispose to risk, whereas others may be the result of the abuse. For instance, the right amygdala is smaller in youths with a family history of alcohol abuse even prior to the onset of problem drinking, whereas hippocampal volumes are reduced only after a history of alcohol use.

Increases in risk taking, novelty seeking, and priority on social factors along with a relative insensitivity to intake reducing factors and a greater sensitivity to ethanol-induced effects on hippocampal and neurotransmitter systems combine to make adolescence a particularly vulnerable time to develop substance abuse.

Affective and Anxiety Disorders

The most dramatic behavioral changes during adolescence are in the social realm with the emergence of sexuality and a shift from family to peer influences. This shift is reflected in a greatly heightened emotional responsiveness to social stimuli during adolescence. Nelson et al. have proposed a three-part model of social information processing involving an infancy/early childhood maturing node for detection of socially relevant environmental cues; an early adolescent maturing affective node which ascribes emotional significance to the cues; and a late adolescent/early adulthood maturing cognitive-regulatory node which inhibits prepotent responses and directs goal-directed behavior. The affective node (which includes limbic structures such as the amygdala, ventral striatum, septum, and hypothalamus) is preferentially sensitive to hormonal effects, and its maturation before the cognitive-regulatory node combined with the increasing importance of social factors may underlie the adolescent surge and gender specificity in the prevalence of anxiety and depressive disorders.

Structural MRI studies of depressed and anxious adolescents report structural anomalies in the superior temporal gyrus, ventral prefrontal cortex, and amygdala and an fMRI study of depressed and anxious adolescents reported anomalous amygdala responsive to social stimuli. Neuroimaging studies support this developmental lag between the affective and cognitive-regulatory nodes. In an fMRI face viewing study, adults, but not adolescents, were able to selectively engage the orbitofrontal cortex when asked to switch from an emotional assessment (i.e., how afraid does it make you feel?) to a nonemotional (i.e., how wide is the nose?).

Increasingly powerful emotional responses to social stimuli, abrupt changes in motivation and reward systems, and a developmental lag in the neural systems necessary to regulate and inhibit highly motivated behaviors in a goal directed manner may underlie the onset of anxiety and depressive disorders during adolescence.

Discussion

The relationship between typical adolescent brain changes and onset of psychopathology is not a unitary phenomenon, but an underlying theme may be conceptualized as 'moving parts get broken.' Many aspects of the complicated and reciprocally interconnected development trajectories may go awry predisposing people to the wide variety of disorders noted from epidemiological studies. Anxiety disorders, bipolar disorder, depression, eating disorder, psychosis, and substance abuse all most commonly emerge during adolescence. Other disorders such as autism or attention deficit hyperactivity disorder (ADHD) have earlier onsets, whereas Alzheimer's disease has a much later onset. The specificity of which disorders have an adolescent onset may offer clues as to the mechanisms involved. A greater understanding of the relationship between specific adolescent changes and the specific cognitive, behavioral, and emotional consequences may provide insight into preventive or treatment interventions.

The remaining section offers considerations for future research exploring the connection between typical adolescent brain changes and psychopathology.

Future Directions

Longitudinal Studies

An emerging theme from pediatric neuroimaging studies is that the path of brain development is often as important as the end point. For example, IQ is predicted by the developmental trajectory of cortical thickness, not by the adult size. Large individual variability in brain anatomy and function call for longitudinal study designs to capture the nuances of heterochronous developmental curves.

The first phases of longitudinal studies have mapped developmental trajectories for typical development, and less so for some psychiatric illnesses. The next phases should go beyond simply mapping brain growth and begin to discern the influences, for good or ill, on those trajectories.

Genetics

A common initial approach to assessing influences is to discern the relative effects of genetic versus nongenetic factors. This is best addressed through comparisons of monozygotic and dyzygotic twins. Results from an ongoing pediatric longitudinal neuroimaging project at the Child Psychiatry Branch of the National Institute of Mental Health indicate significant age by heritability interactions with gray matter heritability generally decreasing with age and white matter heritability generally increasing with age.

Heritability by age interactions may be related to timing of gene expression, which in turn may relate to timing of onset of illness. Postmortem human and animal studies indicate that developmental genes have diverse effects at various stages of brain development. Ongoing studies of specific gene effects on brain maturation may help to sharpen our understanding of brain development mechanisms and provide insight into the etiologies of various pathologies. Genetics may also provide biologically relevant subtypes of neuropsychiatric disorders which are obscured in current diagnostic schemes.

Sex Differences

The marked sex differences in age of onset, prevalence, and symptomatology for nearly every neuropsychiatric disorder may provide important clues as to pathophysiology. The most obvious outward physical manifestations of puberty are caused by changing levels of hormones. Perhaps this has contributed to the tendency to attribute all of the cognitive and behavioral changes of adolescence to 'raging hormones' as well. However, the relationship between hormones, brain, and behavior is complex, reciprocal, and poorly understood. Steroid hormones affect neuronal activity and morphology throughout development. Most neurons have receptors for adrenal and gonadal hormones receptors which when activated affect neurotransmitter function. Short-term effects are mediated by membrane receptors, whereas long-term effects alter gene expression via intraneuronal or nuclear receptors. Conversely, the dramatic hormonal changes of puberty are triggered by alterations in excitatory and inhibitory inputs to gonadotropin-releasing hormone neurons in the pituitary. Behaviorally, hormonal effects drive aggression and sexual interest but the impact on impulse control, logical problem solving, or other cognitive tasks has not been well established.

Social and cultural factors for boys and girls are profoundly different, and the relationship of these differences to manifest pathology should be explored. In the biological realm, sex differences likely stem directly from different genes on the X or Y chromosomes or indirectly through the effects of different hormone levels. Studies of subjects with sex chromosome variations (e.g., XO, XXY, XXYY, XXX, and XXXXY) or anomalous hormone levels (e.g., congenital adrenal hyperplasia, androgen insensitivity syndrome, and familial male precocious puberty) will be useful to sort out the relative contributions of gene and hormone effects. For instance, males with an extra X chromosome (XXY or Klinefelter's syndrome) have a high incidence of language disorders, ADHD, and social skills deficits that are reflected in cortical thickness changes consistent with reports in the literature for XY subjects with those disorders. Girls with 'congenital adrenal hyperplasia' (characterized by high levels of intrauterine testosterone) have an entirely different pattern of brain morphometric findings indicating differential effects of sex chromosomes and hormones.

Multimodal Imaging

Combining multiple imaging modalities on the same individuals, such as structural MRI, fMRI, DTI, magnetization transfer imaging, EEG, and magnetoencephalography (MEG), will synergistically enhance our ability to interpret the signals for each of the modalities. Being able to simultaneously examine interindividual variation from cellular to macroscopic levels will be instrumental in bridging gaps between genes, brain, and behavior.

Translational Research

Postmortem studies of animals that have been imaged will be needed to clarify the nature of changes driving the MRI findings. Of immediate relevance will be discerning the degree to which cortical gray matter changes as detected via MRI are related to arborization/pruning of neurons or to encroachment of white matter on the inner cortical border.

Integration with Social and Educational Science

Studies of adolescent behavior and decision making will need to be integrated better with social and educational science. Laboratory studies of teens using hypothetical situations in calm environments without peer influence may have little relevance for understanding real world decision making which is often in the context of intense physical or emotion arousal, conflicting priorities, and in the presence of peers.

Despite the shared goal of guiding people through the adolescent years safely and optimally prepared for the adult world little rigorous research has been conducted across the neuroscience and education science disciplines.

Summary

Adolescence is a time of tumultuous neurobiological and behavioral change. These changes are usually healthy and optimize the brain for the challenges ahead, but may also confer a vulnerability to certain types of psychopathology. The instruments and technologies to elucidate the relationship between specific neurobiological maturational processes and specific normative or pathologic changes are already in place. Applying these tools to understand when and how deviations from typical development occur may lead to improved treatments for disorders affecting a substantial number of people.

See also: Anxiety Disorders; Cognitive Dysfunction in Psychiatric Disorders; Diffusion Tensor Imaging (DTI); Dopamine – CNS Pathways and Neurophysiology; Electroencephalography (EEG); fMRI: BOLD Contrast; Gamma-Aminobutyric Acid (GABA); Genomics of Brain Aging: Twin Studies; Glutamate; Neuroimaging; Positron Emission Tomography (PET); Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Stress Response: Sex Differences; Stress, Sex and Adolescent Nicotine Response; Substance Abuse and Dependence.

Further Reading

Casey BJ, Tottenham N, Liston C, and Durston S (2005) Imaging the developing brain: What have we learned about cognitive development? *Trends in Cognitive Sciences* 9: 104–110.

Feinberg I (1982) Schizophrenia: Caused by a fault in programmed synaptic elimination during adolescence? *Journal of Psychiatric Research* 17: 319–334.

Giedd JN, Blumenthal J, Jeffries NO, et al. (1999) Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience* 2: 861–863.

Giedd JN, Clasen LS, Lenroot R, et al. (2006) Puberty-related influences on brain development. *Molecular and Cellular Endocrinology* 254–255: 154–162.

Huttenlocher PR and Dabholkar AS (1997) Regional differences in synaptogenesis in human cerebral cortex. *The Journal of Comparative Neurology* 387: 167–178.

Keshavan MS, Anderson S, and Pettegrew JW (1994) Is schizophrenia due to excessive synaptic pruning in the prefrontal cortex? The Feinberg hypothesis revisited. *Journal of Psychiatric Research* 28: 239–265.

Kessler RC, Berglund P, Demler O, et al. (2005) Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry* 62: 593–602.

Lewis DA (1997) Development of the prefrontal cortex during adolescence: Insights into vulnerable neural circuits in schizophrenia. *Neuropsychopharmacology* 16: 385–398.

Luna B and Sweeney JA (2004) The emergence of collaborative brain function: FMRI studies of the development of response inhibition. *Annals of the New York Academy of Sciences* 1021: 296–309.

Paus T (2005) Mapping brain maturation and cognitive development during adolescence. *Trends in Cognitive Sciences* 9: 60–68.

Rakic P, Bourgeois JP, and Goldman-Rakic PS (1994) Synaptic development of the cerebral cortex: Implications for learning, memory, and mental illness. *Progress in Brain Research* 102: 227–243.

Shaw P, Greenstein D, Lerch J, et al. (2006) Intellectual ability and cortical development in children and adolescents. *Nature* 440: 676–679.

Spear LP (2000) The adolescent brain and age-related behavioral manifestations. *Neuroscience and Biobehavioral Reviews* 24: 417–463.

Steinberg L (2005) Cognitive and affective development in adolescence. *Trends in Cognitive Sciences* 9: 69–74.

Thatcher RW (1991) Maturation of the human frontal lobes: Physiological evidence for staging. *Developmental Neuropsychology* 7: 397–419.

Adrenal Steroids: Biphasic Effects on Neurons

M Joëls and H Karst, University of Amsterdam, Amsterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Stress activates the sympathoadrenomedullar system as well as the hypothalamic–pituitary–adrenal axis. As a consequence of the latter, large amounts of corticosteroid hormones are released from the adrenal glands. In addition to the stress-related release, there is also a circadian rhythmicity in the circulating hormone levels.

Corticosteroid hormones – cortisol in humans and corticosterone in most rodents – not only reach peripheral organs but also pass the blood–brain barrier, thus entering the brain. Within the brain, corticosteroid hormones bind to discretely localized intracellular receptors. Two receptors have been recognized: the high-affinity mineralocorticoid receptor (MR) and the lower affinity glucocorticoid receptor (GR). In view of the receptor affinities and circulating hormone levels, it is generally thought that in brain under rest, mainly MRs are activated. These are primarily localized in limbic regions, such as the hippocampus, lateral septum, and central amygdala, and in motor nuclei in the brain stem. During stress and at the circadian peak, GRs also become substantially activated, next to MRs. The GRs are much more ubiquitously distributed in brain.

Activated corticosteroid receptors translocate to the nucleus where they modulate the transcription of responsive genes. This means that corticosterone, through its gene-mediated actions, exerts a slow but persistent control over the cellular protein content, including molecules that are involved in cellular excitability and synaptic transmission. In addition to these slow actions, corticosteroid hormones can also affect neurotransmission through rapid, nongenomic pathways, but these are not considered here.

During the past decade, the effects of selective MR and GR activation on electrical properties of neurons have been mainly examined in the CA1 region of the hippocampus, where both receptors are co-localized. It was observed that MR and GR often induce opposite effects on a given parameter, resulting in biphasic effects of adrenal steroid hormones on neuronal properties.

Corticosteroid Hormones Affect Some but Not All Properties

Of the potential targets for corticosteroid hormone actions, some are extremely relevant to neurons. For instance, steroid modulation of ion channel function would largely affect the excitability and potentially the viability of neurons. Likewise, influences on neurotransmitter receptors and the signaling cascades linked to these receptors would have profound consequences for the information transfer in essential brain regions.

It was found that corticosterone affects some but not all of these properties in the CA1 region of the hippocampus. For instance, voltage-dependent calcium currents were found to be very sensitive to corticosteroid receptor occupation. By contrast, most of the tested potassium and sodium currents are far less sensitive to changes in the level of corticosterone. Of the G-protein-coupled receptors, particularly the serotonin$_{1A}$ (5-HT$_{1A}$) receptor is modulated by corticosteroid hormones, although to some extent other receptors (e.g., for acetylcholine or noradrenaline) are also changed in their function. The modulation of ionotropic receptors is more complex. Here, rapid nongenomic effects seem more prominent. However, a slow enhancement of glutamate receptor-mediated responses by corticosterone or stress has been described. It should be noted that corticosteroid effects on other endpoints (e.g., related to metabolism or cell structure (not discussed here)) will also indirectly affect the efficacy of synaptic transmission, including transmission mediated by glutamate.

Not only cell properties but also phenomena that depend on network function, such as long-term potentiation (LTP) and long-term depression (LTD), strongly depend on the circulating levels of corticosterone.

Biphasic Effects on Neural Properties

In the CA1 hippocampal area, conditions that result in predominant MR activation often induce effects opposite to those where both receptor types are activated. This was first demonstrated for calcium currents and cell properties that are regulated by the intracellular calcium level, such as adaptation of cell firing frequency. Under rest, with predominant MR occupation, influx through high-voltage activated calcium channels is limited. Hence, little adaptation in firing frequency is seen so that excitatory input is effectively transmitted. After stress (resulting in additional GR occupation), calcium influx slowly increases and adaptation of cell firing frequency becomes strong. In the absence of corticosterone, calcium influx through voltage activated channel is also large, as is the firing frequency adaptation. Overall, the dependence of calcium influx through voltage-dependent

channels and related phenomena on the dose of corticosterone displays a U shape (Figure 1).

This dose–response relationship was subsequently also found for other properties of CA1 cells, most notably the responses mediated via 5-HT_{1A} receptors. These receptors are linked through a G-protein to inwardly rectifying potassium channels. Activation of the 5-HT_{1A} receptor therefore leads to opening of potassium channels and thus hyperpolarization of the membrane. Both in the absence of corticosterone and with high doses of the hormone, responses via 5-HT_{1A} receptors are large. By contrast, moderately low hormone levels lead to small 5-HT_{1A} receptor-mediated responses.

At the network level, U-shaped dose dependencies can also be observed for corticosteroid actions in the CA1 hippocampal area. Many studies have reported optimal induction of LTP in animals with low to moderate circulating levels of corticosterone. Both in the absence of the hormone and after considerable stress, it appears to be more difficult to induce LTP, while at the same time the occurrence of LTD is facilitated. It has been postulated that stress and corticosteroid hormones prime the circuits involved in LTP formation such that the threshold for subsequent potentiation is elevated. This may relate to behavioral observations that learning after a period of stress (unrelated to the learning situation) impairs memory

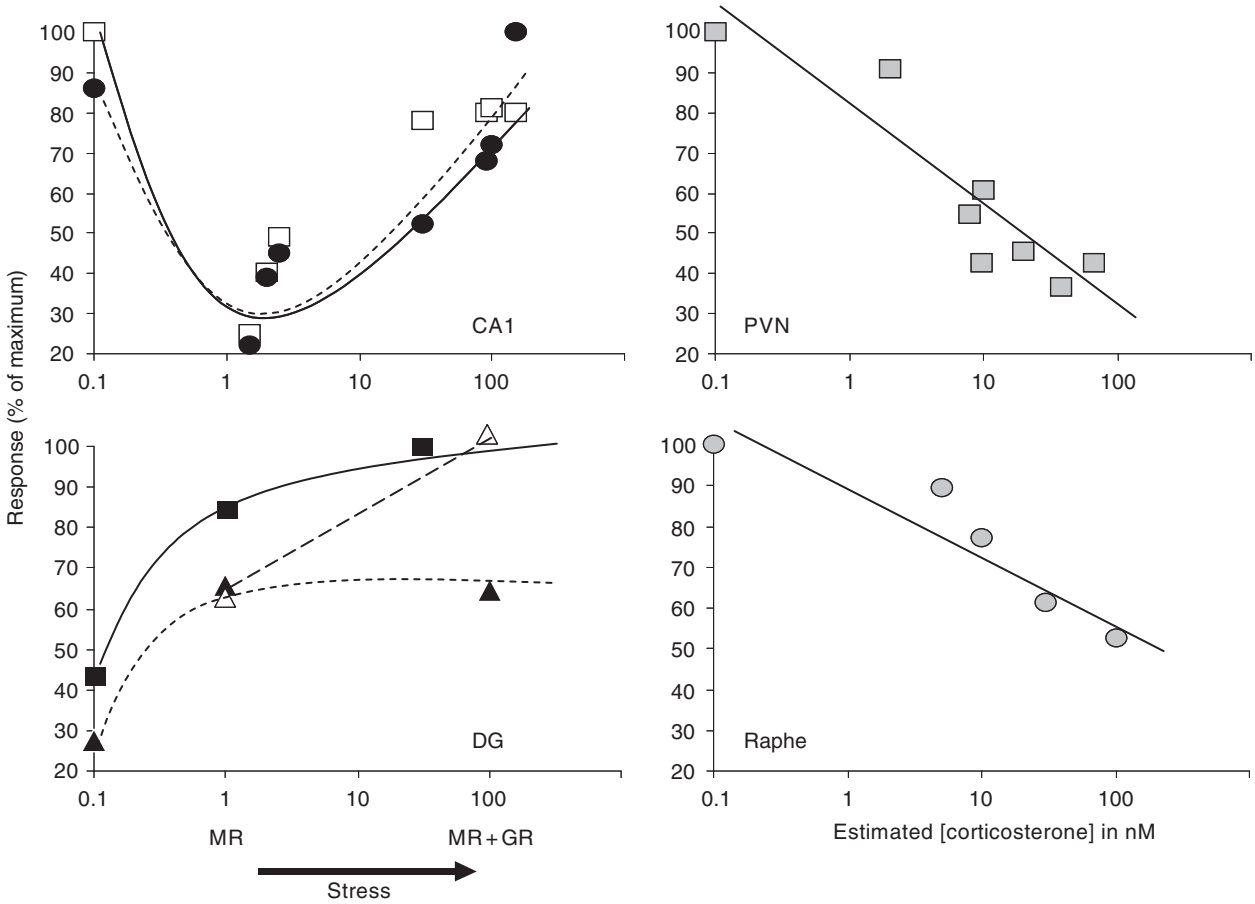

Figure 1 Adrenal corticosteroids cause different effects on neurons, depending on the concentration of the hormone that circulates. For instance, under rest (a situation that is associated with predominant mineralocorticoid receptor (MR) activation), calcium currents in CA1 neurons are small, whereas their amplitude increases several hours after stress (open squares). In the absence of the hormone, such as occurs after adrenalectomy, calcium currents are also large, resulting in a U-shaped dose dependency. A similar dose dependency has been described for responses of CA1 neurons to serotonin (solid circles). The differential responses are linked to the activation of two receptor types, with different affinity for corticosterone. However, U-shaped dose dependencies are not always observed. This can be due to the presence of only one receptor type, as illustrated for GABAergic responses in the paraventricular nucleus (PVN) or serotonergic responses in the dorsal raphe. In the dentate gyrus (DG), the dose–response relationship for glutamate-mediated field (squares) or single cell (solid triangles) responses also seems to be different from that of the CA1 region, despite the presence of MRs and glucocorticoid receptors (GRs) in both areas; after chronic stress (open triangles), GRs are responsive again. Adapted from Joëls M (2006) Corticosteroid effects in brain: U shape it. *Trends in Pharmacological Sciences* 27: 244–250, with permission from Elsevier.

formation. It should be noted, however, that stress which forms an intrinsic part of and occurs concurrently with the event to be remembered helps to consolidate this information.

Two Receptor Types Contribute to the Biphasic Responses in CA1 Neurons

With the use of selective corticosteroid receptor agonists and antagonists it was shown that the dose dependency is coupled to specific receptor occupation ratios. Collectively, the data show that a situation of predominant MR activation in the CA1 region, such as occurs under rest at the circadian trough, promotes ongoing activity and synaptic plasticity, yet at the same time permits only limited calcium influx through voltage-dependent channels. This is important because it allows cells to transfer excitatory information but protects them against calcium-dependent delayed damage. The reverse seems to be true for GR activation (always, of course, against a background of already activated MRs), as happens after stress. In the short term (i.e., over the course of several hours), GR activation will slowly suppress activity in the CA1 area. This is in line with the adaptational character of the hormone because this effect allows CA1 cells to normalize their activity after a temporary period of 'arousal.' However, GR activation also allows more calcium ions to flow into the cells. This may become detrimental when GR activation concurs with a period of heavy excitation (e.g., during an epileptic insult). In agreement, animal models have shown that hippocampal epileptogenesis and ischemic damage is exacerbated by exposure to high corticosteroid levels.

Do Corticosteroids Always Induce Biphasic Responses?

The U-shaped dose dependency is typical for the CA1 area but not for all regions in the brain. This is partly due to the differential distribution of corticosteroid receptors. Most neurons do express GRs, but the expression levels for MR are usually very low, with the exception of some limbic areas. Hence, cellular effects in most areas will be determined by one receptor type only, the GR. Although theoretically biphasic responses can be established through a single receptor system, this does not seem to be the case for neuronal effects of adrenal steroids. This is exemplified by responses measured in the paraventricular nucleus of the hypothalamus and the dorsal raphe, two areas in which GRs prevail. In these areas, a linear dose dependency is observed (**Figure 1**).

However, even in cells that do express both MRs and GRs, the dose dependency is not always U-shaped. For instance, glutamate receptor-mediated responses in the dentate gyrus are enhanced by MR activation (compared to the situation in which corticosterone is absent), but additional GR activation does not lead to noticeable changes.

Several factors could explain why, in this case, biphasic responses are not observed. First, the bioavailability of corticosterone may exhibit local differences. This could not only be due to regional differences in the expression of converting enzymes, such as 11-β-steroid dehydrogenase, but also to differences in transporters which determine the amount of corticosteroids transported over the plasma membrane.

Second, many variants of MR as well as GR have been described. Some of these variants have reduced transcriptional activity. Local differences in the expression of these variants could explain why high concentrations of corticosterone have little impact on cell function.

Finally, transcriptional activity of corticosteroid receptors to a considerable extent depends on protein–protein interactions with cofactors and other transcription factors. The cellular content of these additional molecules therefore also determines the efficacy with which GRs (and MRs) affect neuronal function. Because the cellular content of all of these factors can display large regional differences, it is not difficult to understand why biphasic responses do not always occur, even in the presence of the two receptor systems.

See also: Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Gene Therapy and Protection from Stress-Induced Brain Damage; Stress and Neural Involvement in Metabolism; Stress and Cognition; Stress and Suicide; Stress and Vulnerability to Brain Damage; Stress, Dopamine, and Puberty; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness.

Further Reading

De Kloet ER, Joëls M, and Holsboer F (2005) Stress and the brain: From adaptation to disease. *Nature Reviews Neuroscience* 6: 463–475.

Joëls M (1997) Steroid hormones and excitability in the mammalian brain. *Frontiers in Neuroendocrinology* 18: 2–48.

Joëls M (2006) Corticosteroid effects in brain: U shape it. *Trends in Pharmacological Sciences* 27: 244–250.

Joëls M, Pu Z, Wiegert O, Oitzl MS, and Krugers HJ (2006) Learning under stress: How does it work? *Trends in Cognitive Sciences* 10: 152–158.

Kim JJ and Diamond DM (2002) The stressed hippocampus, synaptic plasticity and lost memories. *Nature Reviews Neuroscience* 3: 453–462.

Kim JJ and Yoon KS (1998) Stress: Metaplastic effects in the hippocampus. *Trends in Neurosciences* 21: 505–509.

Lambert JJ, Belelli D, Peden DR, Vardy AW, and Peters JA (2003) Neurosteroid modulation of $GABA_A$ receptors. *Progress in Neurobiology* 71: 67–80.

Rose JD (2000) Corticosteroid actions from neuronal membrane to behavior: Neurophysiological mechanisms underlying rapid behavioral effects of corticosterone. *Biochemistry and Cell Biology* 78: 307–315.

Adrenergic Receptors

J P Hieble, Wayne, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The adrenoceptors (adrenergic receptors) mediate the diverse effects of the neurotransmitters of the sympathetic nervous system, norepinephrine and epinephrine, at virtually all sites throughout the body. During the past century, the adrenoceptors have been extensively studied by a variety of functional and molecular techniques, and they have been progressively subdivided. Currently, nine adrenoceptors have been cloned, and they have been divided into three major categories – the α_1-adrenoceptors, α_2-adrenoceptors, and β-adrenoceptors. Although there are pharmacological data that suggest the existence of additional adrenoceptor subtypes, it now seems likely that this results from multiple 'affinity states' of a particular adrenoceptor subtype, rather than from a new adrenoceptor protein yet to be cloned. Although the specific receptors involved have not all been identified, norepinephrine-releasing neurons within the central nervous system (CNS) play an important role in psychiatric disorders such as major depression and attention-deficit/hyperactivity disorder (ADHD).

The adrenoceptors belong to the large family of seven transmembrane-spanning, G-protein-coupled receptors, and there is 32–54% amino acid identity between the individual adrenoceptor proteins (Figure 1). The most likely sites for interaction of the catecholamines with conserved structural elements of the adrenoceptors have been identified.

α_1-Adrenoceptors

α_1-Adrenoceptors are widely distributed in both central and peripheral sites, and they are activated either by norepinephrine released from sympathetic nerve terminals or by epinephrine released from the adrenal medulla and some central adrenergic neurons. Peripheral α_1-adrenoceptors mediate a variety of functions, including contraction of smooth muscle, cardiac stimulation, cellular proliferation/apoptosis, and activation of hepatic gluconeogenesis and glycogenolysis. α_1-Adrenoceptors are widely distributed within the CNS, where their activation generally results in depolarization and increased neuronal firing. Most of the peripheral actions of α_1-adrenoceptors are mediated through the inositol phosphate pathway, although there is evidence for activation of adenylyl cyclase within the CNS.

Three distinct α_1-adrenoceptor proteins have been cloned; after some confusion in nomenclature, it has been established that these three recombinant α_1-adrenoceptors correspond to the pharmacologically defined α_{1A}, α_{1B}, and α_{1D} adrenoceptors in native tissues. Selective antagonists are available for each of the α_1-adrenoceptor subtypes. Multiple splice variants of the α_{1A}-adrenoceptor have been identified; however, they appear to have identical pharmacological characteristics. The contraction of several vascular and urogenital tissues has distinct pharmacology from other α_{1A}-mediated responses. This was thought to implicate an additional α_1-adrenoceptor, designated as the α_{1L}-adrenoceptor. This receptor was never cloned, and it now seems likely that the α_{1A}-adrenoceptor may have either α_{1A} or α_{1L} pharmacology, depending on the particular assay, tissue, or experimental conditions employed.

α_1-Adrenoceptor subtypes can be differentially distributed in certain tissues. This differential distribution can reflect either expression of receptor protein/message or pharmacological characteristics in functional assays. For example, the predominant α_1-adrenoceptor found in rat spleen is the α_{1B} subtype, and the functional pharmacological profile of α_1-adrenoceptor agonists (i.e., contraction) in this tissue has α_{1B}-adrenoceptor characteristics, as determined by the use of subtype selective antagonists. Although all of the α_1-adrenoceptor subtypes are present in most blood vessels, the pharmacological characteristics of agonist-induced contraction can vary, such that the responses may be α_{1A} (human microvessels and rat caudal artery), α_{1B} (rabbit or canine aorta), or α_{1D} (rat aorta).

Knockout and overexpression experiments have been reported only for the α_{1B}-adrenoceptor subtype. Knockout of the α_{1B}-adrenoceptor results in impaired vascular responsiveness to phenylephrine, demonstrated both in isolated blood vessels and through measurement of blood pressure responses in anesthetized mice. Cardiac-specific overexpression of the α_{1B}-adrenoceptor results in an impairment of cardiac function without cardiomyocyte hypertrophy.

The α_1-adrenoceptor subtypes are differentially distributed within the CNS. Studies in rat brain have shown α_{1B}-adrenoceptor mRNA to be concentrated in cerebral cortex, thalamus, raphe nuclei, cranial and spinal motor neurons, as well as the pineal gland. mRNA for the α_{1D}-adrenoceptor is localized to the olfactory bulb, cerebral cortex, hippocampus, dentate gyrus, reticular thalamic nucleus, motor neurons, and the inferior olivary complex. In the thalamus, the α_{1B}- and α_{1D}-adrenoceptors have a complimentary

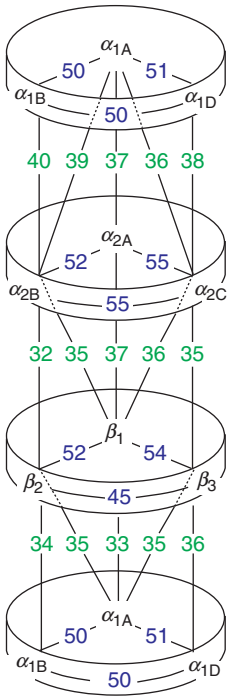

Figure 1 Structural similarity, based on overall amino acid identity, between the nine adrenoceptors. Based on this similarity, the adrenoceptors can be divided into three groups (α_1, α_2, and β), with each group further divided into three subtypes. Structural similarity is higher (45–55% identity) within either of the three groups than between individual members of different groups (33–40% identity).

distribution. Highest levels of α_{1A}-adrenoceptor mRNA are found in regions of the olfactory system, hypothalamic nuclei, and in regions of the brain stem and spinal cord related to motor function. Much less information is available on the distribution of α_1-adrenoceptors in the human brain. Although α_1-adrenoceptors are found in neocortex and dentate gyrus, their distribution differs substantially from that observed in the rat.

Centrally acting α_1-adrenoceptor agonists often potentiate motor stimulation or arousal produced by other pharmacological or behavioral challenges and will increase vigilance when administered alone. α_1-Adrenoceptor agonists can reverse sedation induced by norepinephrine depletion produced by a dopamine β-hydroxylase inhibitor or the cognitive impairment produced by a neurotoxin selective for noradrenergic neurons. Likewise, they can reverse cataplexy in a genetic canine model of narcolepsy. Based on studies in rats and primates, modafinil, currently marketed for the treatment of narcolepsy, may act through the activation of central α_1-adrenoceptors, although based on its structure and peripheral pharmacological effects, this drug would not be expected to be an α-adrenoceptor agonist.

Activation of central α_1-adrenoceptors, presumably in the hypothalamus, stimulates the secretion of several pituitary hormones in human volunteers. A centrally active α_1-adrenoceptor agonist, SDZ NVI-085 (naphtoxazine), improved the performance of patients with Parkinson's disease in some, but not all, tests for attentional deficits. SDZ NVI-085 has shown activity in many of the animal assays for central α_1-adrenoceptor activation.

The functional roles of each α_1-adrenoceptor subtype within the CNS are not completely defined. It has been proposed that both α_{1B}- and α_{2A}-adrenoceptors are present on cell bodies of presympathetic ganglionic neurons in the intermediolateral cell column of the thoracic spinal cord, mediating an excitatory and inhibitory action, respectively. This reciprocal action is consistent with the observation that central α_1-adrenoceptor stimulation will attenuate the sedative actions of an α_2-adrenoceptor agonist in the rat. Mice lacking the α_{1B}-adrenoceptor show different behavioral effects from wild-type mice, in which enhanced reactivity to new situations is observed. Learning behavior in a water maze was impaired, although the ability to escape onto a visible platform was unaffected.

Overexpression of either wild-type or constitutively active α_{1B}-adrenoceptors results in granulovacular neurodegeneration, beginning in the areas of the brain expressing highest levels of the α_{1b}-adrenoceptor (cortex, cerebellum, and hypothalamus) and progressing to other brain areas with age. These mice showed symptoms consistent with neuronal damage, including parkinsonian hindlimb symptoms associated with paralysis and tremor. Grand mal seizures were also observed by 12 months of age. The severity of the seizure disorder was proportional to the level of α_{1B}-adrenoceptor activation, with mice expressing constitutively active receptors being most affected. α_{1B}-Adrenoceptor blockade with terazosin partially reversed the behavioral pathology. However, knockout of the norepinephrine uptake transporter, which results in an increase in central noradrenergic activity, decreases the vulnerability of mice to seizures induced by several convulsant agents.

Correlation of affinity for recombinant α_1-adrenoceptors with their activity in the canine narcolepsy model suggests that the ability of α_1-adrenoceptor agonists to reverse cataplexy in this model results from α_{1B}-adrenoceptor activation. The presence of α_1-adrenoceptors in layers of human hippocampus having a concentration of glutaminergic synapses suggests the possibility for noradrenergic modulation of glutaminergic transmission. Activation of α_1-adrenoceptors in the nucleus tractus solitarius is involved in control of vagal activity to the stomach.

α_2-Adrenoceptors

α_2-Adrenoceptors are also widely distributed throughout the body. Perhaps the most extensively characterized action is the prejunctionally mediated inhibition of neurotransmitter release from many peripheral and central neurons. The demonstration that this effect had a different pharmacologic profile from actions mediated by postjunctional α-adrenoceptors, such as vascular contraction, originally led to the division of α-adrenoceptors into the α_1- and α_2-subtypes. Activation of prejunctional α_2-autoreceptors on sympathetic neurons results in a sympatholytic action. α_2-Adrenoceptors are also present at postjunctional sites, where they mediate actions such as smooth muscle contraction, platelet aggregation, and inhibition of insulin secretion. Many responses produced by α_2-adrenoceptor agonists are mediated through the inhibition of adenylyl cyclase as a consequence of interaction of the agonist–receptor complex with G_i, although other second messengers remain to be characterized.

Three α_2-adrenoceptor proteins have been cloned. These recombinant receptors result in four discrete pharmacological profiles since the α_{2A}-adrenoceptor appears to exist as species orthologs, with those of human, pig, and rabbit having a profile designated as α_{2A}, whereas those of rat, mouse, guinea pig, and cow exhibit pharmacologic profiles designated as α_{2D}. α_{2A}- and α_{2D}-adrenoceptor-mediated responses can be differentiated by the low sensitivity of the α_{2D}-adrenoceptor to blockade by the commonly used antagonists yohimbine and rauwolscine.

Studies in the rat demonstrate the existence of mRNA for all three α_2-adrenoceptor subtypes to be present in the CNS. α_{2A}-Adrenoceptor mRNA is most widely distributed, being found in the cerebral cortex, locus coeruleus, amygdala, hypothalamic paraventricular nucleus, nucleus tractus solitarii, ventrolateral reticular formation, spinal cord, and dorsal root ganglia. Message for the α_{2B}-adrenoceptor is found almost exclusively in the thalamus. The α_{2C}-adrenoceptor is found in olfactory bulb, islands of Calleja, cerebral and cerebellar cortex, hippocampal formation, and dorsal root ganglia. In human brain, radioligand binding assays demonstrate the presence of α_{2A}-adrenoceptor protein in frontal cortex, cerebellum, and hippocampal formation, with another subtype (α_{2B} or α_{2C}) predominant in neostriatum.

α_2-Adrenoceptors within the CNS have long been known to be involved in the antihypertensive action of clonidine and other α_2-adrenoceptor agonists. Although the involvement of a nonadrenergic imidazoline receptor in this action has been postulated, the failure of α_2-adrenoceptor agonists to produce a sympatholytic action in mice where the α_{2A}-adrenoceptor has been knocked out or mutated supports functional experiments in rats and rabbits suggesting an α_2-adrenoceptor-mediated action, at least for systemically administered agonists. In addition to inhibition of sympathetic outflow, the anesthetic and analgesic actions of α_2-adrenoceptor agonists appear to be mediated by the α_{2A}-adrenoceptor. The hypertensive action of α_2-adrenoceptor agonists likely results from activation of vascular α_{2B}-adrenoceptors. Consistent with the apparent role of the α_{2B}-adrenoceptor in the maintenance of vascular tone, mice lacking this receptor subtype failed to develop salt-induced hypertension. Hence, a selective α_{2A}-adrenoceptor agonist may be preferable as a centrally active drug, although sedative and antihypertensive actions can apparently not be dissociated by subtype selectivity. The α_{2A}- or α_{2D}-adrenoceptor subtype (depending on species) appears to be responsible for many other α_2-adrenoceptor-mediated responses, including the major component of prejunctional modulation of sympathetic neurotransmission.

Knockout of the α_{2C}-adrenoceptor has no apparent cardiovascular effect. However, whereas knockout or mutation of the α_{2A}-adrenoceptor subtype produces only partial attenuation of α_2-adrenoceptor-mediated inhibition of transmitter release, elimination of both α_{2A}- and α_{2C}-adrenoceptors results in complete loss of prejunctional modulation of adrenergic neurotransmission and induces pathologic effects related to excess adrenergic tone. Hence, the α_{2C}-adrenoceptor can also participate in prejunctional modulation of neurotransmission. The α_{2C}-adrenoceptor may also participate in cold-induced augmentation of α-adrenoceptor-mediated vasoconstriction.

α_{2C}-Adrenoceptors appear to be involved in the startle reflex and isolation-induced aggression since knockout of α_{2C}-adrenoceptors in mice resulted in an enhanced startle response and shortened attack latency, whereas overexpression of this receptor produced the opposite effect on both parameters. Similarly, the hypothermic response to an α_2-adrenoceptor agonist was attenuated in α_{2C}-adrenoceptor knockout mice, and this response was accentuated by α_{2c}-adrenoceptor overexpression. Behavioral despair in the forced swimming test was enhanced or inhibited by α_{2C}-adrenoceptor overexpression or knockout, respectively. Overexpression of the α_{2C}-adrenoceptor impairs the performance of mice in a water maze.

Despite clinical trials for several indications, no molecule specifically designed to block the α_2-adrenoceptor has been successfully developed as a drug, although α_2-adrenoceptor antagonist activity may play an important role in antidepressants such as mirtazepine.

In addition to their use as antihypertensives, selective centrally acting α_2-adrenoceptor agonists are used as adjuncts to general anesthesia in the treatment of opiate withdrawal, ADHD, and Tourette's syndrome. Their ability to alleviate many of the symptoms of opiate withdrawal most likely results from the ability of both opiate agonists and α_2-adrenoceptor agonists to inhibit locus coeruleus firing, although an action at spinal centers may also contribute. Based on data in a rat model, activation of α_2-adenoceptors may contribute to the activity of methylphenidate in ADHD.

β-Adrenoceptors

Important physiological consequences of β-adrenoceptor activation include stimulation of cardiac rate and force; relaxation of vascular, urogenital, and bronchial smooth muscle; stimulation of renin secretion from the juxtaglomerular apparatus; stimulation of insulin and glucagon secretion from the endocrine pancreas; stimulation of glycogenolysis in liver and skeletal muscle; and stimulation of lipolysis in the adipocyte. Prejunctional β-adrenoceptors are present on some central and peripheral nerve terminals, where their activation results in facilitation of stimulation-evoked neurotransmitter release. However, in contrast to the prejunctional α_2-adrenoceptors, these prejunctional receptors do not appear to have major physiologic significance. Most β-adrenoceptor-mediated actions involve stimulation of adenylyl cyclase through the interaction of the agonist–receptor complex with G_s.

Three β-adrenoceptor proteins have been cloned, and the characteristics of these recombinant receptors correspond with those of the three well-characterized β-adrenoceptors on native tissues, designated as β_1, β_2, and β_3. Many useful pharmacological tools are available for β-adrenoceptor characterization. These include agonists capable of selectively activating β_1-, β_2-, or β_3-adrenoceptors, as well as antagonists selective for each of the three subtypes. Although multiple β-adrenoceptor subtypes can participate in the cardiac stimulation produced by the catecholamines, gene knockout experiments demonstrate that the β_1-adrenoceptor subtype plays the major role. Renin release also appears to be mediated by the β_1-adrenoceptor. Bronchodilation is mediated primarily by the β_2-adrenoceptor. The β_3-adrenoceptor is responsible for lipolysis in white adipose tissue and thermogenesis in the brown adipose tissue found in rodents. Species differences appear to be important for the β_3-adrenoceptor since several selective β_3-adrenoceptor agonists can activate rodent, but not human, β_3-adrenoceptors.

In rat brain, mRNA for the β_1-adrenoceptor is widely distributed, whereas message for the β_2-adrenoceptor is concentrated in olfactory bulb, hippocampal formation, piriform cortex, and cerebellar cortex. The β_3-adrenoceptor does not appear to be present in the CNS. Radioligand assays using dissected rat brain have shown that β_1-adrenoceptors are typically associated with forebrain structures (cerebral cortex, striatum, and hippocampus), whereas β_2-adrenoceptors are most dense in the cerebellum. The β-adrenoceptors of rat cerebral cortex may be localized primarily on glial cells.

Central β-adrenoceptors appear to play a role in movement disorders. β-Adrenoceptor antagonists reduce tremor in MPTP-treated monkeys. Studies in human brain show dense β_1-adrenoceptor binding in striatum, which decreases in the late stages of Huntington's chorea. It is well established that chronic treatment with tricyclic antidepressants produces downregulation of central β-adrenoceptors, presumably as a consequence of increased synaptic levels of the catecholamine neurotransmitters as a result of blockade of neuronal reuptake. β-Adrenoceptors may be involved in the antinociceptive activity of the tricyclic antidepressants since β-adrenoceptor antagonists can reverse this activity. Other functions attributed to central β-adrenoceptors include adaptation to stress, memory and learning, control of respiration, and glial proliferation.

Knockouts of β_1-, β_2, and β_3-adrenoceptors have been reported. Mice lacking both β_1- and β_2-adrenoceptors have also been prepared. None of these mice have phenotypic changes reflecting deficits in CNS function, although detailed behavioral studies have not been reported. Knockout of the β_1-adrenoceptor has a marked effect on embryonic viability, although the few homozygous animals surviving appear normal. This effect on viability is not observed with knockout of the β_2- or β_3-adrenoceptor subtypes. Mice lacking the β_1 adrenoceptor fail to respond to the inotropic action of β-adrenoceptor agonists, confirming the importance of this subtype in the control of cardiac contractility. However, maximal exercise capacity is not reduced. Mice lacking the β_2-adrenoceptor have a normal response to exogenous β-adrenoceptor agonists and have even greater exercise capacity than wild-type mice. However, these animals become hypertensive during exercise and have a lower respiratory exchange ratio, suggesting influences of the β_2-adrenoceptor on energy metabolism. Mice lacking both β_1- and β_2-adrenoceptors have normal basal cardiovascular parameters and normal exercise capacity, although the ability of exercise or the administration of exogenous agonists to increase heart rate is blunted. Mice lacking the β_3-adrenoceptor show mild increases in body fat stores and do not show metabolic responses to a selective β_3-adrenoceptor

agonist. In several cases, an increased responsiveness to one of the remaining β-adrenoceptors is observed when one of the subtypes is knocked out. Thus, there is increased β_1-adrenoceptor responsiveness in β_3-adrenoceptor knockout animals and enhanced β_3-adrenoceptor responsiveness in β_1/β_2 knockout animals. These observations support the concept that many physiological functions can be mediated by multiple β-adrenoceptor subtypes.

Selective overexpression of β_2-adrenoceptors in airway epithelium or smooth muscle results in decreased sensitivity to methacholine-induced bronchoconstriction and an increased sensitivity to a β-adrenoceptor agonists. Cardiac-directed overexpression of the human β_1-adrenoceptor results in the accumulation of fibrous tissue between cardiac myocytes, myocyte hypertrophy, and myofibrilar disarray. These changes result in cardiac dysfunction in older animals. Overexpression of human β_2-adrenoceptors in mouse heart enhances basal cardiac function due to constitutive activity of the expressed receptors. Fibrotic cardiomyopathy is observed in mice expressing high levels of the β_2-adrenoceptor, with the severity and rate of onset being dependent on the level of receptor expression. Overexpression of cardiac β_2-adrenoceptors exacerbates functional deterioration following pressure overload as a result of aortic stenosis.

Several β-adrenoceptor antagonists have also been evaluated for a variety of CNS applications; perhaps the most convincing results are from the use of these agents in treating anxiety of varying etiology, although it is not clear whether a central or peripheral site of action is involved.

Conclusions

Agonists or antagonists at central or peripheral adrenoceptors have several important therapeutic applications; most of these drugs interact with multiple adrenoceptor subtypes. Newly developed pharmacological tools will help to identify the functional role of specific subtypes. Although an action at multiple adrenoceptors may in some cases be beneficial, drugs interacting selectively with adrenoceptor subtypes could result in either an improvement in therapeutic index or the identification of new therapeutic opportunities.

See also: Brain Adrenergic Neurons; Norepinephrine: CNS Pathways and Neurophysiology; Norepinephrine: Adrenergic Receptors; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress.

Further Reading

Arnsten AFT and Dudley AG (2005) Methylphenidate improves prefrontal cortical cognitive function through α_2-adrenoceptor and dopamine D1 receptor actions: Relevance to therapeutic effects in attention deficit hyperactivity disorder. *Behavioral and Brain Functions* 1: 2.

Bjorklund M, Sirvio J, Riekkinen M, et al. (2000) Overexpression of alpha2C-adrenoceptors impairs water maze navigation. *Neuroscience* 95: 481–847.

Chotani MA, Flavahan SA, Mitra S, Daunt D, and Flavahan NA (2000) Silent α_{2C}-adrenergic receptors enable cold-induced vasoconstriction in cutaneous arteries. *American Journal of Physiology* 278: H1075–H1083.

Grijalba B, Callado LF, Javier-Meana J, Garcia-Sevilla JA, and Pazos A (1996) Alpha-2 adrenoceptor subtypes in human brain. A pharmacological delineation of [^{3}H] RX-821002 binding to membranes and tissue sections. *European Journal of Pharmacology* 310: 83–93.

Hein L, Altman JD, and Kobilka BK (1999) Two functionally distinct α-adrenergic receptors regulate sympathetic neurotransmission. *Nature* 402: 181–184.

Hermann GE, Nasse JS, and Rogers RC (2005) α-1 Adrenergic input to solitary nucleus neurons: Calcium oscillations, excitation and gastric reflex control. *Journal of Physiology* 562: 553–568.

Hieble JP, Bondinell WE, and Ruffolo RR, Jr. (1995) α- and β-Adrenoceptors: From the gene to the clinic. 1: Molecular biology and adrenoceptor subclassification. *Journal of Medicinal Chemistry* 38: 3415–3444.

Koshimizu TA, Tanoue A, Hirasawa A, Yamauchi J, and Tsujimoto G (2003) Recent advances in α_1-adrenoceptor pharmacology. *Pharmacology and Therapeutics* 98: 235–244.

Neumeister A, Charney DS, Belfer I, et al. (2005) Sympathoneural and adrenomedullary functional effects of α_{2C}-adrenoreceptor gene polymorphism in healthy humans. *Pharmacogenetics and Genomics* 15: 143–149.

Nicholas AP, Hökfelt T, and Pieribone VA (1996) The distribution and significance of CNS adrenoceptors examined with *in situ* hybridization. *Trends in Pharmacological Sciences* 17: 245–255.

Ruffolo RR, Jr., Bondinell WE, and Hieble JP (1995) α- and β-Adrenoceptors: From the gene to the clinic. 2: Structure–activity relationships and therapeutic application. *Journal of Medicinal Chemistry* 38: 3681–3716.

Sallinen J, Haapalinna A, MacDonald E, et al. (1999) Genetic alteration of the α_2-adrenoceptor subtype c in mice affects the development of behavioral despair and stress-induced changes in plasma corticosterone levels. *Molecular Psychiatry* 4: 443–452.

Spreng M, Cotecchia S, and Schenk F (2001) A behavioral study of alpha-1b adrenergic receptor knockout mice: Increased reaction to novelty and selectively reduced learning capacities. *Neurobiology of Learning & Memory* 75: 214–229.

Waeber C, Rigo M, Chinaglia G, Probst A, and Palacios JM (1991) Beta-adrenergic receptor subtypes in the basal ganglia of patients with Huntington's chorea and Parkinson's disease. *Synapse* 8: 270–280.

Zujic MJ, Sands S, Ros SA, et al. (2000) Overexpression of the α_{1B}-adrenergic receptor causes apoptotic neurodegeneration: Multiple system atrophy. *Nature Medicine* 6: 1388–1394.

Adult Cortical Plasticity

U T Eysel, Ruhr-University Bochum, Bochum, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Plasticity of the brain is a lifelong phenomenon. In general, cortical plasticity describes changes of neuronal wiring and function in the cerebral cortex in response to new challenges of external or internal origin. External stimuli are environmental conditions that impose new tasks on the system. Internal stimuli can be normal developmental demands or pathological lesions in the widest sense – structural, biochemical, or genetic – leading to severe functional impairment. Extrinsic as well as intrinsic neuronal activity can shape wiring and strength of connectivity. Developmental plasticity, as opposed to adult plasticity, has a stronger potential and can include axonal and dendritic growth at a larger scale. Many factors that guide normal development of the brain (e.g., growth factors, netrins) are downregulated after the critical periods of development. In response to lesions in adulthood, some factors can eventually be upregulated to again allow more extensive neuronal plasticity. In that case, neuronal plasticity is the response of the system in an attempt to repair or compensate loss of function. In development as well as in adult cortical plasticity, neuronal activity can change functional connections; however, the effectiveness of synaptic plasticity is usually reduced following postnatal development and critical periods. In the case of adult cortical plasticity, functional and structural adaptations remain, as a rule, spatially localized and restricted to the level of axonal terminals and synapses. New functional connections can be formed by increase of synaptic efficacy in a preexisting network. The principles of reorganization and remapping are common features of the adult auditory, somatosensory, and visual cortices. This article, however, focuses on plasticity in the visual system.

Transition from Visual Cortical Development into Adult Cortical Plasticity

Some principles of adult sensory cortex plasticity can be understood by comparing development in early postnatal life and reorganization after injury of sensory cortical maps. Sensory cortices are characterized by topographic maps that are arranged according to the neighborhood principle. Cells that are close in the sensory periphery are also close in the cortical representations, be it sound frequency in the auditory, body surface in the somatosensory, or retinal surface and visual space in the visual system. The representational maps remain stable throughout adulthood once they have formed through dependence on use in early postnatal life. Accordingly, the postnatal development of the visual cortex is characterized by use-dependent sculpturing of the necessary, correct connections. Characteristic cortical maps of retinal topography and stimulus specificity are formed. At the end of the critical periods, the functional connections are stabilized, and the programs that prevailed during this period, such as axon growth and guidance, removal of incorrect connections, strengthening of optimal, and weakening nonoptimal inputs (synaptic plasticity) are modified in a way to foster stability without sacrificing flexibility in adulthood. Finally, the ability to change synaptic weights by long-term potentiation (LTP) or long-term depression (LTD) persists throughout adult life into senescence. This is exemplified by the lifelong ability for perceptual learning. Use, disuse, and lesions are the strong stimuli for plasticity and reorganization in the adult visual cortex. For example, when challenged by central or peripheral lesions, the visual system displays characteristic patterns of remodeling. While the observed subcortical changes remain limited, more widespread reorganization is seen in the cortex. In the early phase, cortical reorganization appears more restricted and based on functional modifications of cortical connections governed by subtle changes in the balance of excitation and inhibition. In later stages, more extended reprogramming of connections, based on long-term changes in synaptic connectivity (LTP, LTD), is observed and can ultimately involve anatomical reorganization including local growth mechanisms.

There are many different types of adult cortical plasticity. One type is purely use-dependent and closely related to perceptual learning as observed in humans. This plasticity is operating in everyday life when cellular properties are modified in response to repetitive use in demanding tasks. This use-dependent plasticity is closely related to 'synaptic learning' based on mechanisms like LTP and LTD. Other types of visual cortical plasticity are observed after lesions; they are less specific and generally lead to changes in single cell properties and reorganization of cortical representations (maps). The following sections focus on adult use-dependent plasticity and two different types of lesion-induced plasticity in the visual cortex.

Use-Dependent Adult Visual Cortical Plasticity

Fast associative plasticity of visual cortical cells has been shown *in vitro* with electrical stimulation as well as *in vivo* by pairing of natural stimuli with artificial depolarization of single cells. The *in vitro* experiments were performed with intracellular recordings in brain slices of adult rats and cats and showed that adult visual cortical cells are capable of synaptic plasticity. High-frequency repetitive stimulation of specific inputs results in LTP, which is expressed in long-lasting increases of synaptic potentials.

To induce changes in cellular responses *in vivo* on a fast timescale, natural sensory stimuli have been paired with electrical or pharmacological stimulation of cells in the visual cortex of anesthetized animals. Properties such as orientation specificity (e.g., vertical vs. horizontal) or ocular dominance (left vs. right) were shifted toward the property reinforced by repetitive pairing with a facilitating stimulus during supervised learning. These experiments provide *in vivo* evidence that the modifiability of mature visual cortical connections follows the Hebbian rule, which requires both pre- and postsynaptic activity within a defined time window. While many experiments used electrical or pharmacological stimuli to facilitate synaptic plasticity, is was also shown in experiments with anesthetized animals that purely natural repetitive stimulation can induce long-lasting changes of visual cortex cell properties such as receptive field (RF) size and substructure. Visual costimulation of the central parts of an excitatory RF with an unresponsive region located just outside was applied to elicit 'associative' synaptic changes. With this paradigm visual cortical cells did specifically 'learn' to respond to an originally subthreshold region outside their excitatory RF or to change the subfield composition within their RF.

Recently, sophisticated *in vivo* studies have shed additional light on the possible properties and mechanisms of use-dependent adult visual cortical plasticity by examining short-term changes of orientation specificity induced by adaptation to certain orientations. Repulsive shifts of preferred orientation were observed that involved time-dependent processes of depression as well as enhancement. This kind of short-term plasticity was dependent on the location of cells within the layout of the orientation preference map of the visual cortex that is characterized by large coherent orientation domains and 'pinwheel centers' or singularities, where all orientations meet within a small circumscribed region. In the vicinity of singularities, the capacity for adaptive changes in response to oriented gratings is much stronger than in orientation domains. It has further turned out that certain forms of short-term plasticity in the adult visual cortex are stimulus timing-dependent. In respective experiments, plasticity of orientation selectivity *in vivo* depended on stimulus timing within a ± 40 ms window like the crucial parameters for spike timing-dependent plasticity. This suggests that underlying mechanisms are modifications of synaptic weights of intracortical connections.

Beyond the single cell level, use-dependent adult visual cortex plasticity can be shown in whole cortical networks. When optical imaging of intrinsic signals was combined with localized electrical intracortical microstimulation, profound changes in orientation preference maps were observed in the adult cat visual cortex. After a few hours of high-frequency focal electrical stimulation, the cortical representation of the orientation represented at the stimulation site increased because surrounding cells with previously different orientation preferences shifted toward the stimulated orientation at sites as distant as 4 mm. These widespread changes did not involve attention, special training, or reinforcement; they were purely dependent on high-frequency synchronized activation of the cortical network.

When monkeys are trained for several months to identify the orientation of a small grating, they show a marked increase in performance. Single cell recordings from primary visual cortex (V1) of these monkeys after successful training reveal significantly increased orientation sensitivity of only those cells and orientations that represented the trained orientation. This proves that use-dependent single cell plasticity in the adult visual cortex can be linked directly to behavior and perceptual learning.

Lesion-Induced Cortical Plasticity

Retinal lesions remove afferent input to thalamus and cortex. Cortical lesions destroy cortical cells that have a dual function, being targets for topographic retinal projections via the thalamus on one hand, and being the origin of intracortical connections on the other. These completely different types of lesion, retinal versus cortical, pose very different problems for reorganization and restitution of function to the visual system. In the case of retinal lesions, the option is to reinnervate intact cortical cells that have lost their sensory input; however, the topographically correct fibers are no longer available due to retinal destruction. In the case of cortical lesions, the retinal afferent information is still completely on hand, but the cortical target cells are lost. Accordingly, the options for reorganization are different, and so is the outcome of functional reorganization: after retinal lesions, a blind cortical region is filled in with retinal topography redundantly spreading across the cortical surface and resulting in many cortical cells newly

representing the same retinal topography, whereas after cortical lesions, the surviving cells in the surround receive new suprathreshold information from geniculocortical afferents and acquire additional topographical representations from the retina. While the former fills in a cortical region of still existing but blind cells with new but topographically redundant activity, the latter enables true functional improvement by recovering previously lost afferent information for further cortical processing. The lesion-induced changes can be roughly subdivided into acute, subacute, and chronic effects. The pathology of cell death and functional suppression that is predominant in the acute phase (first day postlesion) is followed by events of neuronal plasticity in the subacute (second day to 1 week postlesion) and chronic phases (weeks to months after a lesion). It is interesting that the different types of lesions seem to trigger similar effects that lead to receptive field plasticity, topographical remapping, and a limited recovery of function.

Plasticity in the Primary Visual Cortex after Cortical Lesions

Focal lesions of the visual cortex lead to a local visual field loss (scotoma). The size of the scotoma mirrors cell death and functional depression of cortical neurons. However, the dimension of functional loss following cortical lesions can diminish in both animals and humans with time and especially when specific training is performed. A possible basis for restoration of function is a reactivation of 'silent' cells in the only partially damaged, functionally suppressed surround of the lesion (the penumbra) by strengthening of subthreshold synaptic inputs. Another possibility is the increase in size of receptive fields of intact cells at the border of the lesions, that is, an additional representation of lost parts of the visual field by cells that also maintain their original topographical representation of space and thereby enable the perception of previously lost parts of the visual field. It is interesting that in the first weeks after a lesion in the visual cortex, an increase of excitability is observed in its surround. This increased excitability seems to support the theory of increased plasticity of cells surrounding the lesion. Changes of receptive field size that do not occur spontaneously in the first days following a lesion develop months later close to the border of lesions in the cat visual cortex.

Faster changes in receptive field size can be experimentally induced by repetitive visual stimuli. This can be achieved even in neurons of the normal adult cat visual cortex. Repetitive stimulation associated with high response rates in the neurons under study can lead to a significant and stimulus-specific widening of receptive fields. The observed effects last from 20 min to several hours and are reminiscent of effects similar to short-term potentiation and LTP. Such LTP-like mechanisms might be enhanced because of locally increased excitability in the surround of visual cortical lesions and might be associated with the RF enlargements observed in the surround of cortical lesions. LTP is the best candidate to explain cellular learning and memory and can be studied in brain slices *in vitro* following high-frequency electrical stimulation of inputs. LTP is characterized by strengthened synaptic transmission over a span from more than 1 h to days and weeks. LTP was first described in the mammalian hippocampus and can also be observed in the visual cortex of mice, rats, and cats. During early postnatal development, when synaptic plasticity is highly expressed – as in the visual cortex during the critical period – very effective LTP is observed. In fact, lesion-induced reorganization after lesions in the visual cortex is associated with increased LTP, as was shown by increased synaptic plasticity in slices of the lesioned rat visual cortex *in vitro*.

Reorganization of the Primary Visual Cortex Following Retinal Lesions

Retinal lesions trigger chronic reorganization in the afferent visual pathway. This follows the interruption of visual signal flow from the destroyed retinal area to the visual cortex that locally deprives cells in the thalamus (lateral geniculate nucleus) and the primary visual cortex of their normal input. The excitability of cells in the visual thalamic relay nucleus (dorsal lateral geniculate nucleus, dLGN) that are primarily deafferented by a monocular photocoagulator lesion of the retina changes in a characteristic way: initially the cells show a complete loss of visual drive and a significant decrease in spontaneous activity; then spontaneous activity increases in deafferented cells, and some cells inside the deafferented region regain visual input originating from the retina directly adjacent to the lesion. Accordingly the RFs of these newly connected cells change their retinotopy and shift to a new position located at the border of the retinal lesion. This reorganization of the retinotopic map leads to a partial filling in of the scotoma on the subcortical level and is associated with a lateral spread of excitation in the dLGN of up to 300 μm. With the same type of lesion, a much more extended reorganization can be observed in the visual cortex. Apart from the longer distances involved, the reorganization was exactly of the same kind as observed in the dLGN: originally silenced cells regained visual input that originated from the border of the retinal lesion, a retinal region that did not excite these cells

before. The same kind of plasticity was found in cat and monkey visual cortex. The reorganization gradually fills in cortical regions that were depleted of retinal inputs with new excitability over distances up to 5 mm. This distance conforms to the range of lateral horizontal connection in the visual cortex. This chronic retinotopic remapping is preceded by fast changes in visual RF properties that follow acute disuse induced by retinal lesions and are elicited as well when selective surround stimulation is applied to artificial retinal scotomata.

The cortical region affected by loss of input after retinal lesions constitutes an anatomically defined lesion projection zone (LPZ, Figure 1(a)). A characteristic imbalance of input occurs in and outside the LPZ as cells with normal afferent input are situated in direct proximity to visually silenced cells. In addition, excitation is upregulated and inhibition downregulated inside the LPZ. This situation leads to changes in cortical cell RF size and topography when the active cells gain influence on their inactive neighbors through the intracortical horizontal fiber systems. The filling in follows a characteristic spatiotemporal pattern in which hyperactive and visually hyperexcitable cells are first found close to the order of normal cortex in the LPZ, then progressively deeper inside the LPZ (Figure 1(b)). The long-range horizontal fibers are the most probable basis of a low-level perceptual filling in of retinal scotomata. In fact, artificial scotomata (produced in normal people by stabilized retinal images) take considerable time to fill in whereas scotomata are filled in immediately in patients with chronic

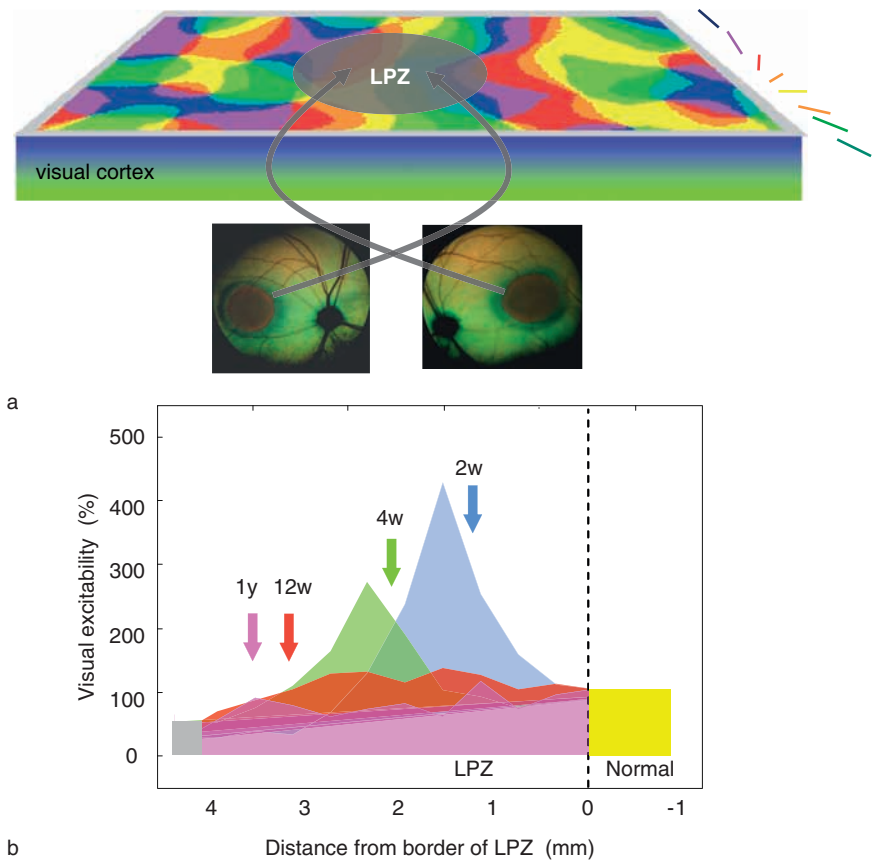

Figure 1 Spatiotemporal pattern of filling-in of the lesion projection zone (LPZ) in the cat visual cortex. (a) Schematic representation of the projection from the left and right eye retinae (with central 10° diameter photocoagulator lesions) to the primary visual cortex. The resulting binocular LPZ is surrounded by normal cortex as shown on top of a topographic map with color-coded orientation preferences. (b) Visual excitability inside the LPZ as a function of distance from the border between LPZ and normal cortex. Normal activity outside the LPZ is shown in yellow; the nonreorganized, and after 1 year visually nonresponsive center of the LPZ is shown in gray. Visual excitability in general moves gradually with time into the LPZ and is associated in the first weeks with a peak of significant hyperexcitability at the front of reorganization. Arrows indicate the first, well defined receptive fields plottable at the different time points. (b) Adapted from Giannikopoulos DV and Eysel UT (2006) Dynamics and specificity of cortical map reorganization after retinal lesions. *Proceedings of the National Academy of Sciences of the United States of America* 103: 10805–10810.

retinal lesions. This observation further supports the proposition that cortical connectivity is profoundly and permanently changed after chronic lesions. This proposition is corroborated by the above-mentioned experimental evidence that shows primary cortical reorganization taking place on the basis of functional changes in synaptic weights of preexisting connections but followed in a final step by anatomical stabilization associated with terminal sprouting, as observed more than 8 months after retinal lesions in the adult cat visual cortex.

It is interesting that when state-of-the-art high-resolution functional magnetic resonance imaging (fMRI) was applied in the visual cortex of adult monkeys, no changes of the visuotopic projection were observed. This seems to indicate that fMRI, a method prone to predominantly detect population responses and presynaptic activity, does not reflect the subtle changes found in suprathreshold performance of individual neurons. On the other hand, the lesion-induced remapping detected on the single-cell level shows close correlation with the psychophysical observations in humans with retinal lesions, indicating the functional relevance of cortical reorganization and filling in.

Mechanisms of Lesion-Induced Cortical Plasticity

Hyperactivity after Cortical and Retinal Lesions

Suppression and facilitation of activity was observed at different distances from the border of heat lesions in the visual cortex. Subnormal activity was seen in a region of <1 mm around the lesion whereas hyperactivity prevails at 1–2.5 mm from its border; at more distant positions (>2.5 mm) the activity was normal. This hyperactivity after heat lesions was already present 1 day after the lesion, became maximal after 3–7 days, and was still visible 30 days postlesion. One possible mechanism for the increase of activity in the surround of lesions is an imbalance between γ-aminobutyricacidergic (GABAergic) inhibition and glutamatergic excitation. This is supported by experimental evidence: in the immediate surround of neocortical lesions, inhibitory transmission is downregulated and a widespread reduction of GABA receptors is observed; then 1–5 days postlesion and up to 2 mm from the border of the lesions, the fast $GABA_A$- and late $GABA_B$-induced inhibitory postsynaptic potentials show reduction in amplitudes and peak currents.

At the same time, the glutamatergic excitation is increased, as substantiated by higher amplitudes and longer durations of *N*-methyl-D-aspartate (NMDA) receptor-mediated excitatory postsynaptic potentials. In addition, excitatory field potentials were significantly larger than normal. These enlarged potentials were NMDA-dependent as shown by blockade with d-amino-phosphonovaleric acid.

It is interesting that also after retinal lesions, a hyperactive zone was found within the LPZ in the visual cortex (**Figure 1(b)**), and an increase in glutamate immunoreactivity was observed at the same time. Hyperactivity and hyperexcitability appear to be a common feature associated with postlesion functional rewiring in the adult visual cortex following deafferentation.

The GABAergic and glutamatergic systems have been investigated with immunhistochemical methods 2 weeks after homonymous lesions in the retina (central 10°) and reveal a characteristic pattern of transmitter system immunoreactivity (IR) inside and in the surround of the deafferented region (LPZ). The glutamic acid decarboxylase (GAD) IR was downregulated inside the LPZ where the number of positive profiles was extremely reduced in the neuropil while the GAD IR in the cell somata remained rather unchanged. IR of the excitatory neurotransmitter glutamate revealed glutamate-positive cells in cortical layers II to VI of area 17. The retinal lesions caused a clear reduction (by 15–26%) in the number of glutamate-immunoreactive cells in the supra- and infragranular layers of the cortical LPZ, compared with normal cortex. Furthermore, close to the border of the LPZ inside the deafferented region, glutamate IR displayed a sharp increase (by 50–100%) throughout layers II to VI of area 17. This glutamate peak was largest in layer VI and had a width of 600–800 μm.

Both the changes in GAD and glutamate IR diminished with time. When the central and the peripheral portions of area 17 were compared in cats with postlesion survival times of more than 12 weeks, no significant differences in the number of glutamate IR-positive cells and in the GAD IR of the neuropil were observed. Obviously the initial imbalance of the excitatory and the inhibitory immunhistochemical reactions returns to normal once the functional reorganization is completed.

Enhanced LTP-Like Effects

Pharmacological studies in brain slices have shown that reduced inhibition and/or increased excitation facilitates the expression of LTP. This situation is mimicked in the visual cortex following lesions in which a decreased strength of inhibition and an increased excitation are observed. Accordingly, electrophysiological

recordings in slices of rats with cortical lesions revealed a significantly elevated level of LTP at synapses from neurons located at a defined distance of 1–4 mm from the border of the lesion and up to 1 week following induction of the injury.

This enhanced synaptic plasticity is accompanied by changes in the intracellular calcium concentration $(Ca^{2+})_i$. Both the neuronal resting calcium concentration and the stimulus-evoked calcium influx are moderately increased at the border of the lesion. The origin of this increase of intracellular calcium was investigated by means of specific neuronal calcium permeation blockers for ionotropic NMDA and AMPA receptors, revealing an increase of Ca^{2+} influx mediated through both of these glutamate receptor types postlesion. While a previously present calcium permeability is just increased in the case of the NMDA receptors, there is a fundamental lesion-induced change in the functional properties of the ionotropic AMPA receptors. AMPA receptors are known to be calcium-impermeable under physiological conditions in rat neurons after postnatal day 15. However, after cortical lesions in animals more than 23 days old, a neuronal calcium influx was measured in the presence of pharmacological blockers of all known sources of intraneuronal calcium influx. This calcium influx was blocked by an antagonist for AMPA. This unusual calcium permeability mediated by AMPA receptors can be explained by a lesion-induced change in the specific composition of the AMPA receptor protein subunits: a reduction in the expression of the glutamate receptor 2 subunit, which is normally observed only in the young postnatal brain, is also observed after lesions in the adult.

A similar reversion to more juvenile patterns is observed for the NMDA receptors. Cortical lesions change the relative expression levels of the messenger RNA for the ionotropic NMDA receptor subunits NR2A and NR2B. Experimental data suggest an increased relative expression of the NR2B subunit of the NMDA receptor at the border of the lesion (due to downregulation of the amount of NR2A subunits). It is interesting that this relatively high amount of NMDA receptors containing the NR2B subunit has also been described in the early postnatal cortex in the phase when enhanced synaptic plasticity (LTP) is observed. Following cortical lesions the true relevance of NR2B subunits for the lesion-induced facilitation of LTP has been clearly shown: blockade of NR2B subunit-containing receptors with ifenprodil reduces the increased LTP in lesioned rat cortex to control values.

Calcium imaging with Fura2-AM (a fluorescent probe) combined with LTP induction can be utilized to record the stimulus-evoked and field potential-correlated calcium influx in the lesion-treated cortex. In fact, significantly increased calcium levels are observed during the induction of LTP by theta burst stimulation as well as 55 min thereafter. This increase of the stimulus-correlated calcium influx postlesion might explain the expression of facilitated LTP as elevated intracellular calcium levels directly correlate with the strength of synaptic LTP.

Growth Factors and Morphological Correlates of RF Plasticity

The neurotrophins brain-derived neurotrophic factor (BDNF), neurotrophin-3, nerve growth factor, and the insulin-like growth factor IGF-1 have been found elevated in the visual cortex as early as 3 days after binocular retinal lesions. The related neurotrophin receptors were elevated as well. Furthermore, increased transcription levels of calcium/calmodulin-dependent kinase II, microtubule-associated protein 2, and synapsins are observed in the area undergoing cortical reorganization. There is increased neuronal activity in regions with elevated BDNF expression. Thus, BDNF appears directly linked to the activity-dependent early unmasking of existing connections on one hand, and as a growth factor, it might also represent a link to the late morphological changes involving axonal sprouting.

See also: Activity in Visual Development; BDNF in Synaptic Plasticity and Memory; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Hebbian Plasticity; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Metaplasticity; Neuronal Plasticity after Cortical Damage; Pain and Plasticity; Perceptual Learning and Sensory Plasticity; Plasticity and Activity-Dependent Regulation of Gene Expression; Plasticity of Intrinsic Excitability; Somatosensory Plasticity; Spike-Timing Dependent Plasticity (STDP); Synapse Formation: Competition and the Role of Activity; Synaptic Plasticity: Neuronal Sprouting; Synaptic Plasticity: Short-Term Mechanisms; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging.

Further Reading

Arckens L, Schweigart G, Qu Y, et al. (2000) Cooperative changes in GABA, glutamate and activity levels: The missing link in cortical plasticity. *European Journal of Neuroscience* 12: 4222–4232.

Chino YM, Smith EL III, Kaas JH, Sasaki Y, and Cheng H (1995) Receptive-field properties of deafferentated visual cortical neurons after topographic map reorganization in adult cats. *Journal of Neuroscience* 15: 2417–2433.

Darian-Smith C and Gilbert CD (1994) Axonal sprouting accompanies functional reorganization in adult cat striate cortex. *Nature* 368: 737–740.

Dreher B (2006) Reprogramming of striate and extrastriate visual cortices following retinal lesions. In: Lomber SG and Eggermont JJ (eds.) *Reprogramming the Cerebral Cortex*, pp. 3–46. Oxford, UK: Oxford University Press.

Eysel UT (2002) Plasticity of receptive fields on early stages of the adult visual system. In: Fahle M and Poggio T (eds.) *Perceptual Learning*, pp. 43–65. Cambridge, MA: MIT Press.

Eysel UT and Mittmann T (2006) Remodeling of cortical connections and enhanced long-term potentiation after lesions of the visual cortex. In: Lomber SG and Eggermont JJ (eds.) *Reprogramming the Cerebral Cortex*, pp. 61–71. Oxford, UK: Oxford University Press.

Giannikopoulos DV and Eysel UT (2006) Dynamics and specificity of cortical map reorganization after retinal lesions. *Proceedings of the National Academy of Sciences of the United States of America* 103: 10805–10810.

Gilbert CD (1998) Adult cortical dynamics. *Physiological Reviews* 78: 467–485.

Gilbert CD and Wiesel TN (1992) Receptive field dynamics in adult primary visual cortex. *Nature* 356: 150–152.

Kaas JH, Collins CE, and Chino YM (2003) The reactivation and reorganization of retinotopic maps in visual cortex of adult mammals after retinal and cortical lesions. In: Pessoa L and De Weerd P (eds.) *Filling-In: From Perceptual Completion to Cortical Reorganization*, pp. 187–206. Oxford, UK: Oxford University Press.

Kaas JH, Krubitzer LA, Chino YM, Langston AL, Polley EH, and Blair N (1990) Reorganization of retinotopic cortical maps in adult mammals after lesions of the retina. *Science* 248: 229–231.

Obata S, Obata J, Das A, and Gilbert CD (1999) Molecular correlates of topographic reorganization in primary visual cortex following retinal lesions. *Cerebral Cortex* 9: 238–248.

Smirnakis SM, Brewer AA, Schmid MC, et al. (2005) Lack of long-term cortical reorganization after macaque retinal lesions. *Nature* 435: 300–307.

Zur D and Ullman S (2003) Filling-in of retinal scotomas. *Vision Research* 43: 971–982.

Aggression: Hormonal Basis

N G Simon, Lehigh University, Bethlehem, PA, USA
C F Ferris, University of Massachusetts Medical School, Worcester, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Characterization of the hormonal contribution to mammalian aggressive behavior and analogous emotional states in humans has followed a split path. Advances in the understanding of the molecular, cellular, and biochemical processes that mediate hormonal effects have enabled increasingly sophisticated models of behavioral regulation in animal models. Studies in humans, however, continue to face methodological limitations that limit progress in regard to cellular, systemic, and state-dependent processes, even as genome-based technologies identify polymorphisms that may impact on human aggression and aversive emotional states. Fortunately, trends in human research that recognize specific subtypes of aggression (e.g., hostility, impulsivity, and dominance) have provided a conceptual basis for a stronger link between animal models and human states. This more refined framework should facilitate defining the neuroendocrine contribution to pathological aggression, violent behavior, and aversive emotions.

Offensive aggression in males and females can serve as a useful model to elucidate hormonal influences on human behavior. In males, the central role of testosterone (T) is well-established. The hormone–aggression relationship in females, although more complex, involves dehydroepiandrosterone (DHEA), the most abundant hormone in humans and an androgenic neurosteroid. By defining regulatory pathways for normal, sex-typical aggression, a comparative basis for identifying changes in hormonal function that contribute to pathological aggression and aversive emotional states can be developed.

Hormonal Regulation: Shared Features in Males and Females

The neuromodulator hypothesis provides an overarching concept that bridges hormonal influences on aggressive behavior in both males and females. The model posits that hormones influence aggression by modulating neurochemical function. One strength of this hypothesis is that it bridges basic and clinical considerations related to hormone function, aggression, and emotional states. By considering hormonal processes that regulate aggression in adulthood and how these processes interact with representative neurochemical systems (e.g., serotonin (5HT), vasopressin (AVP), and γ-aminobutyric acid (GABA)) that are central to the expression of aggressive behavior, the potential utility of the neuromodulator hypothesis can be illustrated.

Metabolic processes critically influence hormonal effects in target cells and represent a second shared feature between neuroendocrine regulatory pathways in males and females. In males, aromatization and 5α-reduction of T profoundly influence both the development and expression of regulatory pathways. Defining the modulatory effect of DHEA on female-typical aggression parallels circumstances in the male in that multiple metabolites also may be involved. The biosynthesis and metabolism of DHEA has been described in the periphery, but it is not well characterized in the central nervous system (CNS) (Figure 1). The 3β-hydroxysteroid dehydrogenase (3β-HSD), hydroxysteroid sulfotransferase (HST), steroid sulfatase (SST), and CYP7B pathways potentially merit attention in relation to female typical aggression. The formation of androstenedione (AE) in response to 3β-HSD activity can increase the availability of more potent androgens, the relative activity of HST and SST determines the potential contributions of DHEA sulfate (DHEA-S) versus DHEA, and CYP7B family activity catalyzes the formation of 7α- and 7β-hydroxy DHEA in the brain. The contribution of these metabolites to female-typical aggression may involve both genomic and nongenomic effects. The demonstration of the direct androgenic effects of DHEA and the suggestion that more potent androgens are formed from DHEA establish a genomic component to the mechanism of action of DHEA. In regard to nongenomic effects of DHEA, the formation of the sulfate versus DHEA itself, and their differential potencies as modulators of $GABA_A$receptor function, further demonstrates the importance of metabolic processes.

Hormonal Regulation in the Adult CNS

Males

The ability of T to facilitate the display of intermale aggressive behavior is widely recognized. Research in adult males has shifted to a mechanistic emphasis to enable the characterization of cell/molecular processes underlying functional T-mediated pathways.

Hormonal substrates T can promote aggression in males through four distinct pathways (Table 1): androgen-sensitive, which responds to T itself or its 5α-reduced metabolite, dihydro testosterone (DHT);

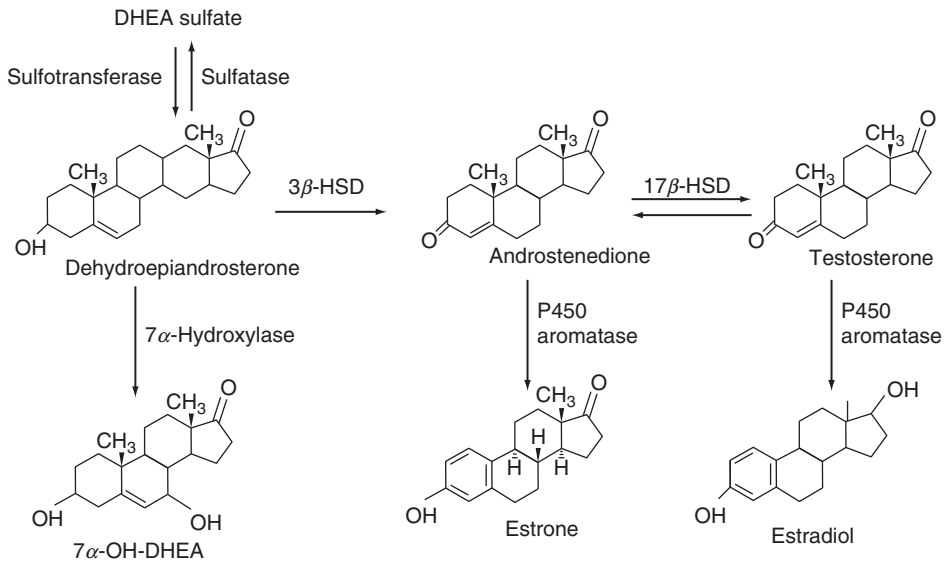

Figure 1 A summary of the various routes of DHEA metabolism in target tissues. Three pathways have been identified with DHEA as the initial substrate: (1) to DHEA sulfate, a reversible path involving hydroxysteroid sulfotransferase and steroid sulfatase, (2) to 7α- or 7β-hydroxy DHEA, which involves the CYP7B enzyme family, and (3) to androstenedione, which involves 3β-hydroxysteroid dehydrogenase (3β-HSD) and provides the possibility for the formation of more potent androgens and estrogens. DHEA, dehydroepiandrosterone. Reproduced from Simon NG, Mo Q, Hu S, Garippa C, and Lu S (2006) Hormonal pathways regulating intermale and interfemale aggression. *International Review of Neurobiology* 73: 99–123, with permission from Elsevier.

Table 1 Hormonal pathways in the adult male brain that facilitate the display of offensive intermale aggression[a]

Genotype	*Estrogen-sensitive*	*Androgen-sensitive*	*Synergistic (estrogen + androgen)*	*Direct T-mediated*
CF-1	+ + +	+ +	−	−
CFW	+ + +	−	−	−
CD-1	−	+ +	+ + +	−
C57BL/6J	−	−	−	+

[a]Plus (+) and minus (−) signs indicate relative sensitivity or insensitivity. Reprinted from Simon NG (2002) Hormonal processes in the development and expression of aggressive behavior. In: Pfaff DA, Etgen AM, Fahrbach SE, and Rubin RT (eds.) *Hormones, Brain, and Behavior,* pp. 339–392. San Diego: Academic Press.

estrogen-sensitive, which uses E_2 derived by aromatization; synergistic or combined, in which both the androgenic and estrogenic metabolites of T are used to promote behavioral expression; and direct T-mediated, which uses T itself. The full range of steroid-sensitive systems is not necessarily present in every male because genotype is the major determinant of the functional pathway(s). The most common system uses E_2, which supports a key role for aromatization and estrogen signaling in aggression. In males, all pathways share the feature of high sensitivity.

Neural steroid receptors

Androgen receptor. Major regions exhibiting abundant androgen receptors (ARs) in mammals include the bed nucleus of the stria terminalis (BNST), lateral septum (LS), medial preoptic area (MPOA), basal hypothalamus, and medial amygdala (MAMYG), all of which are part of the neuroanatomical substrate for conspecific aggression. These descriptive findings help define functional circuitry, but do not elucidate how AR regulation contributes to behavioral expression. The fact that AR protein level is controlled through a common, autoregulated mechanism in both male and female neural tissue exemplifies this point. This finding indicates that the rapid increase in AR protein level seen in response to androgen is not sufficient to produce parallel changes in behavioral responsiveness. An example of this point is that the activation of male-typical aggression in ovariectomized females requires 16–21 days of androgen treatment, whereas AR level increases dramatically within 24 h. Increased cellular AR content alters the transcription of other androgen-regulated genes, which over time leads to the expression of aggression. The extended time course required to induce

male-like aggression in females thus may be a useful model for defining androgen-dependent circuitry and the role of trophic factors in the hormone-dependent elaboration and maintenance of regulatory pathways.

Estrogen receptor. There are two estrogen receptor (ER) subtypes, α and β. ERα appears to play a primary role based on knockout studies. Offensive attacks were rarely displayed by ER knockout (ERKO) males, whereas wild-type and heterozygous males showed significantly greater attack durations. ERβ seems to exert a more subtle, modulatory influence on estrogen signaling and male-typical aggression. Several potential ways that ERβ could effect aggressive behavior have been proposed based on behavioral changes in males fed diets containing high or low levels of soy phytoestrogens for 15 months. Specifically, males on a high-soy diet showed extreme agonistic behavior. Soy phytoestrogens preferentially bind to ERβ and are less active than E_2–ERα complexes in promoting gene expression. The behavioral and molecular results indicate that in normal males perturbations in estrogen signaling can lead to alterations in aggression. Potential mechanisms underlying the more extreme agonism potentially include decreased 5HT function in the dorsal raphe, where ERβ is the sole subtype thus far detected in primates, or enhanced ERα activity as a result of decreased negative modulation of ERα:ERβ heterodimers. These concepts are supported by *in vitro* assays, available immunochemical findings, and results with βERKO mice.

Neurochemical targets Aggression is a complex social behavior that involves virtually every known neurochemical system. In males, however, the roles of 5HT and AVP are widely recognized, as are hormonal influences on their function. The interactions between hormones and these two systems can be used to illustrate the neuromodulator hypothesis.

Hormonal modulation of serotonin function Lower serotonergic tone is associated with increased aggression, whereas enhanced serotonergic function reduces the expression of aggressive behavior. These relationships have been demonstrated in numerous mammalian species, including humans. Although there are different 5HT receptors, agonists with selective affinity for the 5HT1 subtype, particularly $5HT_{1A}$ and $5HT_{1B}$, specifically and selectively reduce offensive intermale aggression.

Testicular hormones most likely influence aggression by altering serotonin function at $5HT_{1A}$ and/or $5HT_{1B}$ sites in brain regions that either constitute, or project to, the neuroanatomical substrate for intermale aggression. To establish that 5HT function at these receptors is affected by gonadal hormones requires demonstrating that androgens or estrogens differentially affect the ability of $5HT_{1A}$, $5HT_{1B}$, or combined agonist treatments to alter the display of offensive intermale aggression. If this is shown, it then becomes critical to identify neuronal populations in which these effects are produced and the mechanisms through which androgens or estrogens influence $5HT_{1A}$ or $5HT_{1B}$ receptor function in these regions. Unfortunately, few studies have directly addressed this question in the specific context of intermale aggression, which represents a gap in the field. The effects of systemic treatment with serotonergic 1A and 1B agonists seem to be determined, at least in part, by the hormonal environment. If estrogens were present (as the aromatized metabolite of T), the ability of 1A or 1B agonists to inhibit offensive aggression was restricted. When aggression was promoted by a direct androgenic treatment like DHT, however, $5HT_{1A}$ and $5HT_{1B}$ agonists were highly effective in decreasing the expression of offensive behaviors.

Microinjection studies point to the underlying complexity associated with elucidating neuroendocrine modulation of 5HT function. For example, the introduction of selective $5HT_{1A}$ or $5HT_{1B}$ agonists into the LS in the presence of diethylstilbetrol (DES), a potent, specifically acting estrogen, had essentially no effect on behavior. When gonadectomized males were implanted with DHT, however, aggressive behavior was decreased with 1B-agonist microinjection alone or in combination with the 1A agonist 8-OH-DPAT. At the level of the LS, then, these observations suggest that an androgen-sensitive pathway can be attenuated by the action of serotonin at 1B receptor sites. In the MPOA, observed effects were robust. Significantly reduced aggression was seen in the presence of either androgen or estrogen with $5HT_{1A}$- or $5HT_{1B}$-agonist microinjections. The MPOA may thus be a major integrative site for T–5HT interactions in the regulation of T-dependent aggression.

The alteration by gonadal hormones of the ability of serotonergic 1A or 1B agents to effect T-dependent intermale aggression supports the neuromodulator hypothesis. Examples of comparable hormone–neurostransmitter interactions are more numerous in other systems, particularly in regard to reproductive behavior, anxiety, and mood disorders.

The findings of steroidal enhancement or repression of the ability of $5HT_{1A}$ and $5HT_{1B}$ agonists to attenuate offensive intermale aggression raise several potential mechanisms that require attention. Estrogens, for example, can alter $5HT_{1A}$ gene expression or influence ligand availability through effects on synthetic or degradative processes. Another possibility for establishing a direct effect on $5HT_{1A}$ gene

function would be the identification of a functional estrogen responsive element (ERE) in the promoter region of the $5HT_{1A}$ receptor gene. Interestingly, both mouse and human $5HT_{1A}$ receptor genes contain an ERE. The spacer element represents a difference between the postulated motifs and the consensus sequence (i.e., 5 nucleotide (nt) rather than 3). However, nonconsensus EREs with different spacer lengths are responsive to estrogenic regulation. For example, the salmon gonadotropin-releasing hormone (GnRH) and BDNF genes have ERE motifs with either 8- or 9-nt spacers and can bind activated estrogen receptors *in vitro*.

Arginine vasopressin The structural and functional integrity of specific aspects of the AVP system in rats, hamsters, and other species is dependent on gonadal steroids. Direct evidence for a systematic relationship among AVP, aggression, and testosterone has been obtained from over 15 years of studies with male Syrian hamster (*Mesocricetus auratus*). Related investigations in rats, mice, voles, and other species, including humans, have broadened our understanding of this interrelationship by defining the ways that social structure interacts with the AVP system and by identifying the cellular and molecular processes involved in hormonal modulation of AVP synthesis and V1a receptor.

AVP can act on multiple brain areas to facilitate aggression through binding to V1a and, perhaps, V1b receptors. Microinjection of AVP into the anterior or ventrolateral hypothalamus of resident hamsters significantly increases biting attacks on intruders. Prairie voles show a dose-dependent increase in aggression toward intruders following intracerebroventricular (ICV) administration of AVP. Results from rats and humans show a positive correlation between aggression and AVP levels in cerebrospinal fluid (CSF). In hamsters, microinjections of Manning compound, a V1a antagonist, into the anterior hypothalamus causes a dose-dependent inhibition of aggression by a resident male toward an intruder and blocks aggression associated with the development of dominant-subordinate relationships. In male prairie voles, ICV injection of V1a receptor antagonist reduces aggression toward male intruders. SRX251, an orally active V1a antagonist with high affinity for the human receptor, causes a significant dose-dependent reduction in latency to bite and number of bites by a resident hamster toward an intruder. Although studies to date have focused on the blockade of the V1a receptor in the regulation of aggression, evidence from knockout mice suggests that the V1b receptor also may be involved. In support of a role for V1b receptor in aggression are results showing that treatment with a selective V1b antagonist, SSR149415, reduces offensive aggression of resident male hamsters.

T is essential for the modulation of aggression by AVP. Microinjections of AVP into the ventrolateral hypothalamus (VLH) did not activate offensive aggression in castrated male hamsters, whereas attack behavior was induced if animals were treated with T implants. The facilitative effect of T on the AVP system in hamsters appears to primarily involve maintenance of V1a receptor populations in the anterior and ventrolateral hypothalamus. Whether this is accomplished via a direct effect on V1a receptor gene expression is unknown because the promoter region for the hamster V1a gene has not yet been sequenced.

The androgenic and estrogenic metabolites of T differentially contribute to steroidal modulation of the AVP system. Estrogen treatment alone after castration partially restored AVP immunoreactivity, DHT alone was ineffective, and combined treatment produced the greatest effect. A direct effect on AVP gene expression appears to be an important mechanism because AVP mRNA also increases in response to circulating gonadal steroid levels. The recent identification of a putative androgen responsive element (ARE), ERE, and AP1 elements in the rat AVP gene promoter region strongly supports enhanced AVP gene transcription as a component of steroidal modulation of the AVP system in rats. Hormonal modulation of the AVP system in rats thus may be achieved through a mechanism different from that in hamsters, in which regulation of the V1a receptor appears to be the primary target for steroidal effects.

It seems reasonable to conclude that T modulates the AVP system. This can be achieved through different pathways, including regulatory effects on the AVP or V1a receptor genes. A specific role for estrogen, the product of aromatization, is indicated, but how the androgenic metabolites contribute has not yet been elucidated. Social structure and seasonality must be overlaid on these effects because they impose boundaries on the extent to which AVP contributes to offensive aggression. Recent studies suggest a potential molecular basis for these effects, at least in voles. Characterizing these interrelationships appears to represent one of the important steps needed for developing an integrative neurobiology of offensive aggression.

Females

The hormonal contribution to female-typical aggression appears to primarily involve DHEA, the most abundant steroid in humans and a neuroactive neurosteroid synthesized in the brain of humans and other mammals. DHEA, which is intrinsically androgenic and can be metabolized to more potent androgens (**Figure 1**), inhibits female-typical aggression when

administered chronically. DHEA is a modulator of $GABA_A$ receptor function, and GABA is a principal neurochemical regulator of interfemale aggression. Collectively, these findings raise the possibility that the effects of DHEA on female-typical aggression are exerted through multiple mechanisms. Defining the relationship among DHEA, aggression, and aversive emotional states is a major issue confronting the field.

DHEA: Anti-aggressive mechanism of action The prevailing view concerning the modulation of female-typical aggression by DHEA is that it reduces bioavailable pregnenolone sulfate (Preg-S), a potent negative modulator of the $GABA_A$ receptor, which in turn enhances GABA function and the inhibition of offensive aggression. DHEA also can act at membrane sites rapidly to alter receptor conformation (a nongenomic effect) and effect longer-term processes directly or through neurosteroid metabolites (a genomic effect), which could then alter membrane receptor function.

The $GABA_A$ receptor has a pentameric structure with multiple binding sites, suggesting a number of possibilities for modulation by DHEA. Potential targets for DHEA include sites that bind GABA, the benzodiazepines, the Cl^- ionophore, barbiturates, and a still unidentified neurosteroid binding site.

The finding of direct androgenic effects of DHEA adds another level of complexity to the underlying mechanisms. DHEA exhibits characteristics of typical androgenic compounds, including upregulation of AR protein expression, conferring AR transcriptional activity and inducing dose-dependent increases in limbic system AR content. The androgenic activity of DHEA was confirmed by the observation that DHEA induced intracellular translocation of AR-green fluorescent protein (GFP) and the formation of nuclear clusters. When COS-7 cells transfected with an AR-GFP expression vector were treated with 10^{-7} M DHEA for 24 h, AR-GFP protein translocated from the cytoplasm into the nucleus and led to the formation of punctate fluorescent foci. Another potential mechanistic component tied to genomic effects of DHEA or its metabolites is an alteration of $GABA_A$ subunit structure, which potentially can influence the extent of modulation (Figure 2).

The combination of cell surface and direct androgenic effects of DHEA and its metabolites represents a cross-talk cellular signaling system potentially linked to its anti-aggressive effect. Defining the interrelationship among DHEA, its androgenic effects, the subunit structure of $GABA_A$ receptor, and attendant changes in function to elaborate the anti-aggressive mechanism of action of this neurosteroid and how it modulates the expression of female-typical aggression is a major challenge in the field.

Neuroanatomical Substrates

Characterizing the endocrine neuroanatomy of intermale and interfemale aggression requires elucidating critical sites where the modulatory effects of hormones are exerted on neurochemical function. The focus of the current discussion is on regions where these effects are produced directly in target cells, that is, where co-localization of steroid receptors and key neurotransmitters has been established. Clearly, there are indirect effects of steroids on neurochemical function that are probably of equal importance for developing a comprehensive description of neuroanatomical circuitry. However, the paucity of work on such effects precludes consideration at this time.

Males

Lesion and implant studies in rodents have identified subregions of the amygdala, LS, BNST, anterior hypothalamus, and MPOA as part of a presumptive steroid-sensitive circuit for this behavior. The interconnections among these regions are well documented, as is the presence of target neurons for estrogen and androgen. Conspecific offensive aggression in male rodents is dependent on an olfactory stimulus. There has been progress in defining the chemosensitive circuitry mediating intermale aggression in the hamster and rat. A functional pathway from the vomeronasal organ through the olfactory tracts to the MAMYG, BNST, LS, anterior hypothalamus and other limbic structures has been demonstrated in these model species. The integration of steroid receptor distribution with chemosensitive circuitry provides a basis for studies that can define modulatory effects of the hormonal environment.

Other considerations, however, constrain progress in defining neuroanatomical substrates. The identification of distinct androgen-sensitive and estrogen-sensitive pathways raises the possibility that these systems are at least partially independent. Next, aggression itself is a complex behavior with multiple components, many of which are nonreflexive. This aspect of conspecific aggression led to an approach in which investigators sought to identify regions that were either tied to distinct components of the behavioral sequence or appeared to serve as major integrative sites. Examples of the former include portions of the hypothalamus that constitute an attack area, as well as elegant studies of the neuroanatomy of defensive rage and predatory attack in the cat. However, findings in these studies are not linked to either androgenic or estrogenic effects on intermale aggression.

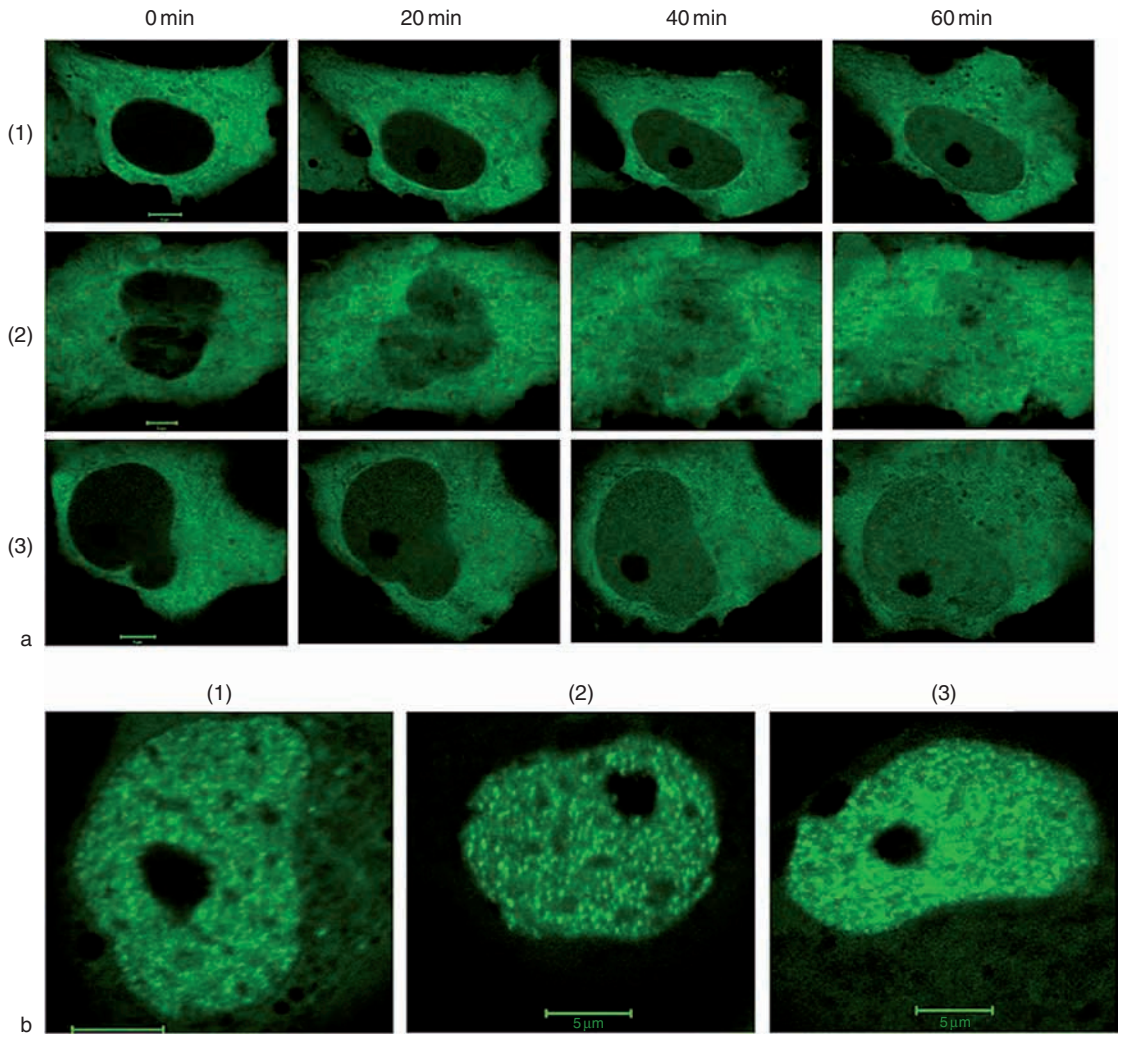

Figure 2 Androgen-dependent AR-GFP intracellular translocation in COS-7 cells: (a) typical intracellular trafficking of AR-GFP in COS-7 cells treated with three steroids, recorded at 0, 20, 40, and 60 min; (b) punctate nuclear distribution of AR-GFP proteins in COS-7 cells (scale = 5 μM). COS-7 cells were transfected with AR-GFP expression plasmid pEGFP-N1-AR. AR-GFP fusion proteins were detected in living cells by excitation with 488-nm line from an argon laser of a confocal microscope. In (a), cells were treated with. (1) 10^{-5} M DHEA, (2) 10^{-6} M Adiol, or (3) 10^{-6} M Adione. A very obvious accumulation of AR-GFP was observed in cells treated with the three compounds in 60 min. In (b), AR-GFP proteins formed fluorescent foci in the nucleus of some cells treated with the same three steroids for approximately 90 min. AR-GFP, androgen receptor-green fluorescent protein; DHEA, dehydroepiandrosterone. Reprinted from Mo Q, Lu SF, and Simon NG (2006) Dehydroepiandrosterone and its metabolites: Differential effects on androgen receptor trafficking and transcriptional activity. *Journal of Steroid Biochemistry and Molecular Biology* 99: 50–58.

Regarding integration, the MAMYG, MPOA, anterior hypothalamus, and LS seem to represent important sites based on findings in rodents. These investigations bridged findings on the chemical neuroanatomy of offensive aggression, in which serotonin and AVP have important regulatory roles, with manipulation of the steroidal environment. Interestingly, putative molecular linkages between androgens, estrogens, and subtypes of AVP and serotonin receptors have been identified. Determining whether these linkages are functional and their relationship to the physiology of conspecific aggression will represent a significant advance in defining neuroendocrine regulation.

An approach that may prove valuable for defining the androgen-sensitive system involves the adult female as a model. The 16–21 day treatment course required for behavioral activation may provide an opportunity to identify androgen-induced changes in neuronal structure and function that may be critical for the display of male-typical aggression.

Studies of neuromodulation by androgens and estrogens in male rodents may have implications for several aversive emotional states in humans. These include, for example, impulsive aggression and violent behavior, the extreme aggression that can accompany steroid or drug abuse, and the depression and anxiety that is sometimes associated with chronic stress.

Females

The neuroanatomical substrates regulating the display of interfemale aggression remain largely undefined. GABA function at the $GABA_A$ receptor, along with modulation by DHEA, appears to represent a critical regulatory component. Although this provides a basis for elucidating at least one major aspect of functional circuitry, the inherent complexity in an analysis of neuroanatomical substrates needs to be acknowledged. For example, the structural heterogeneity of $GABA_A$ across different brain regions, co-localization of $GABA_A$ with steroid receptors in subsets of target cells, and the cellular enzyme phenotype that determines whether neurosteroidal influences on $GABA_A$ are nongenomic, genomic, or both, suggest that multiple mechanisms and processes are at work even within a single neurotransmitter system. When these cell/molecular issues are viewed within the social structure needed to elicit this behavior and the duration of DHEA exposure required to influence interfemale aggression, the lack of progress in this area is perhaps more readily understood.

The analysis of neuroanatomical substrates for aggression in females potentially can advance understanding of aversive emotional states in humans. One potential area is clinically significant anxiety, in which the $GABA_A$ receptor is established as a major target for drug therapies. To the extent that interfemale aggression provides an opportunity to characterize regional differences in the function and modulation of $GABA_A$ receptor, there may be substantive implications for the development of new compounds that treat this disorder. Another area of interest is aging and associated mood changes, in which depression is more commonly seen. DHEA has attracted considerable interest as a replacement therapy. Advancing our understanding of neurosteroidal effects on the $GABA_A$ receptor can add to the understanding of cell/molecular processes that may be influenced by DHEA replacement.

Human Aggression and Aversive Emotional States

Defining the hormonal contribution to human aggressive behavior and aversive emotional states has proven to be a formidable task. Numerous factors that span methodological and conceptual considerations have limited substantive progress, although recent developments, particularly the use of more refined endpoints such as hostility, impulsivity, and dominance suggest that models with greater utility are likely to be forthcoming. Examples of methodological issues include, for example, the use of systemic T assays as a proxy for hormonal effects in the brain and the correlation of hormone levels with endpoints of aggression ranging from paper and pencil tests in college students to histories of violent acts. Conceptual issues include efforts to link hormone levels to a global, catch-all construct of aggression without discriminating between behavioral forms that may be influenced by hormones and those that are not; a view held for many years that female-typical aggression was necessarily also regulated by T; and, with the increasing application of genomic technologies, efforts to identify a single causal gene for aggression and violence. In addition, there is the inherent complexity that accompanies the interaction of physiological systems with social, cultural, cognitive, and experiential factors when attempting to discern neurobiological substrates of human behavior.

The considerations just described suggest that it may be premature to attempt a systematic test of the neuromodulator hypothesis in human males and females. It may be useful, however, to consider principles that have emerged from animal and human studies and use this information to develop more refined hypotheses that may help elucidate neuroendocrine contributions to human aggression, violent behavior, and aversive emotional states.

Characterizing hormonal contributions to sex-typical aggression can inform our understanding of pathological aggression, violence, and aversive states by identifying potential systems in which perturbations may engender atypical behavior. In previous sections, the roles of T, its major metabolites, and their interaction with the 5HT and AVP systems were considered in males, whereas in females the contributions of DHEA, its metabolites, and GABAergic function, particularly at the $GABA_A$ receptor, were presented to illustrate the neuromodulator hypothesis. The question, then, is whether analogous findings have been made in studies of human aggression and, if not, to identify research in those areas as potential priorities.

Males

Results regarding a correlation between T and aggression have been equivocal. This may be due to the breadth of models that have been assessed as correlates of T and a failure to recognize that systemic

blood levels do not necessarily reflect intracellular processes, that is, the importance of E_2 and aromatization (except perhaps in extreme cases, such as large doses of anabolic steroids). It can be argued reasonably that dominance behavior, as well as assertiveness and novelty seeking, is influenced by T in human males. Important mechanistic considerations in determining these hormone-behavior relationships may well include aromatization rates and the dynamics of T secretion. The potential role of aromatization in human male aggression has received scant attention. Competitive interactions induce a consistent change in T levels – a rise in anticipation of competition and sustained elevation as a consequence of winning with lower levels in losers. These changes may produce effects on neurochemical function and subsequent behavior, leading to aversive states in the losing males.

Serotonergic function, with a few reported exceptions, is inversely related to impulsive aggression, hostility, and a life history of aggression. The repeated demonstration of this relationship through studies showing a blunted 5HT response to systemic challenge and lower 5HT metabolites in the CSF in impulsive/aggressive individuals provided the basis for the analysis of genes involved in 5HT function. Several of these studies have identified polymorphisms in the 5HT transporter and $5HT_2$ receptor that are associated with impulsive aggression and an increased risk of violent suicide, respectively. Intronic polymorphisms in the tryptophan hydroxylase 1 gene also have been suggested as a risk factor in suicidal behavior.

Elevated vasopressin function has been linked to impulsive aggression in human males. Higher levels of AVP in the CSF were positively correlated with higher scores aggression in borderline patients. An interesting feature of this study was that these same individuals also had reduced serotonergic function. Several investigations in animal models have demonstrated regulatory interactions among 5HT, AVP, and the expression of aggression. Because there is abundant evidence from studies in other mammals, including nonhuman primates, that establishes hormonal influence on 5HT and AVP function, an assessment of these interactions in humans may help develop an integrative neurobiology of inappropriate aggression and violent behavior.

An excellent example of the potential of the neuromodulator hypothesis and one that can be directly linked to aggression and aversive emotion is androgenic (as well as glucocorticoid) modulation of monoamine oxidase A (MAO-A), a key enzyme in the degradation of 5HT, dopamine, and norepinephrine. Specifically acting androgens regulate the MAO-A promoter via an ARE as well as via Sp1 sites. The association between MAO-A gene expression and impulsive aggression has been documented in humans.

Females

Far less is understood about the hormonal contribution to aggression in human females than in males. A significant factor in this state of affairs has been continuing efforts to identify a systematic relationship between T and various forms of aggression in women. Mixed to negative results generally have been obtained when T levels were correlated with competitive situations, delinquency, assertiveness, and hostility/irritability during the menstrual cycle. Although a focus on dominance-related behaviors might provide some insights, even this more restrictive perspective is uninformative. Women do not exhibit changes in T dynamics comparable to those seen in men in competitive settings, and changes in T levels during the menstrual cycle are not correlated with mood states associated with aggression. Interestingly, a relationship between T and postpartum depression recently was described, which demonstrates the need to carefully define forms of aggression and aversive states that are influenced by the hormonal environment.

Alterations in the serotonin system have been identified in women who exhibit impulsivity. This represents a potential area for the determination of hormonal modulation based on indirect evidence from several studies demonstrating estrogenic influences on serotonin function and mood. The potential role of AVP in human female aggression has not been determined, although results showing an influence of AVP on a broad range of social behaviors suggest that it is an area that merits attention. Overall, however, there is little clear evidence linking hormones, aggression, and aversive states in women.

The equivocal to negative results in human studies suggest that a different approach to the relationship between hormones and offensive aggression in women may be needed. Animal models suggest that sex-typical aggression in females is modulated by DHEA. If this pertains to humans, an emphasis on the GABA system and its modulation by neurosteroids might be useful.

Two examples illustrate the potential utility of this model. The $GABA_A$ receptor is a target for modulation by anabolic steroids. Decreased $GABA_A$ function through elevated androgen exposure may underlie the hostility/rage associated with steroid abuse. How DHEA, the most abundant androgen in women, may influence these responses in normal women remains to be determined. The second example is MAO-A gene regulation, in which DHEA or its metabolites can

directly regulate expression and thus affect serotonergic (and other amine neurotransmitter) function. Together, these and related observations reinforce the potential modulatory role of DHEA in human female aggression and aversive emotions.

Conclusion

Knowledge of hormonal processes in humans linked to the expression of aggression and aversive emotional states during adulthood is limited. The importance of hormone metabolism in males has been demonstrated in animal models but remains an area ripe for study in humans. The intrinsic androgenicity of DHEA and the contribution of its metabolites also may represent important mechanistic steps in females, although this remains an open issue. Although several target neurochemical systems and candidate genes or gene polymorphisms have been identified, substantial work is needed to identify the full range of cellular processes that are affected as well as the genomic and nongenomic mechanisms that mediate these effects. Progress in this area is essential for defining the modulation of neurochemical function by gonadal hormones and neurosteroids. It is an essential step in developing a systems model that eventually should encompass gene regulation, functional circuitry, behavioral expression, and adaptation.

The neuromodulator hypothesis provides an integrative, systems-based framework for determining the hormonal contribution sex-typical aggression, pathological aggression, and aversive emotional states in humans. An interesting feature of the model is that hormonal systems seem to produce different effects in each sex. In males the net effect of T is neutral or facilitative, whereas in females DHEA may be inhibitory.

In males, critical studies defining the molecular interface among T, its metabolites, and components of 5HT, AVP, and no doubt other neurotransmitter systems are needed. In females, nongenomic and genomic modulation of $GABA_A$ receptor function requires characterization as well as the potential interface with MAO-A. This is only a partial list based on the representative systems covered here. Fully characterizing hormonal contributions will require defining multiple effects covering both genomic and membrane-level actions.

Finally, other important aspects of hormone function in aggression and aversive emotion should be noted. We chose to focus on gonadal steroids in males and an adrenal androgen/neurosteroid in females. Other areas of interest, for example, are the effects of corticosteroids and the dopamine system.

The complex nature of hormonal modulation and the need for refined models to assess endocrine contributions to human behavior has been demonstrated in animal models. Progress in defining forms of aggression (e.g., hostility, irritability, impulsivity, and dominance) is a significant improvement over grouping very different behaviors under aggression. An emphasis on a single gene or even a limited subset of physiological markers as the cause of aggression, although sometimes informative, is limiting. A systems perspective that recognizes when hormones may have a role; that physiological effects are modulatory; and that social structure, life events, and subsequent adaptations are reflected in alterations in cellular signaling pathways will be necessary to build an integrative neurobiology of human aggression and aversive emotional states.

See also: Aggression: Neurochemical and Molecular Mechanisms; Emotional Hormones and Memory Modulation; Emotional Control of the Autonomic Nervous System; Gene Expression Regulation: Steroid Hormone Effects; Hormones and Behavior.

Further Reading

Arango V, Huang YY, Underwood MD, and Mann JJ (2003) Genetics of the serotonergic system in suicidal behavior. *Journal of Psychiatric Research* 37: 375–386.

Archer J (2006) Testosterone and human aggression: An evaluation of the challenge hypothesis. *Neuroscience and Biobehavioral Reviews* 30: 319–345.

Baulieu EE (1997) Neurosteroids: Of the nervous system, by the nervous system, for the nervous system. *Recent Progress in Hormone Research* 52: 1–32.

Bethea CL, Lu NZ, Gundlah C, and Streicher JM (2002) Diverse actions of ovarian steroids in the serotonin neural system. *Frontiers in Neuroendocrinology* 23: 41–100.

de Almeida RM, Ferrari PF, Parmigiani S, and Miczek KA (2005) Escalated aggressive behavior: Dopamine, serotonin and GABA. *European Journal of Pharmacology* 526: 51–64.

de Vries G and Simerly RB (2001) Anatomy, development, and function of sexually dimorphic neural circuits in the mammalian brain. In: Pfaff DA, Etgen AM, Fahrbach SE, and Rubin RT (eds.) *Hormones, Brain, and Behavior*, pp. 137–191. San Diego: Academic Press.

Ferris CF (2005) Vasopressin/oxytocin and aggression. *Novartis Foundation Symposium* 268: 190–198; (discussion) 198–200, 242–153.

Gruber CJ, Gruber DM, Gruber IM, Wieser F, and Huber JC (2004) Anatomy of the estrogen response element. *Trends in Endocrinology and Metabolism* 15: 73–78.

Labrie F (2003) Extragonadal synthesis of sex steroids: Intracrinology. *Annals of Endocrinology* 64: 95–107.

Lee R and Coccaro E (2001) The neuropsychopharmacology of criminality and aggression. *Canadian Journal of Psychiatry* 46: 35–44.

Mehta AK and Ticku MK (2001) Unsulfated and sulfated neurosteroids differentially modulate the binding characteristics of various radioligands of GABA(A) receptors following chronic ethanol administration. *Neuropharmacology* 40: 668–675.

Meyer-Lindenberg A, Buckholtz JW, Kolachana B, et al. (2006) Neural mechanisms of genetic risk for impulsivity and violence in humans. *Proceedings of the National Academy of Sciences of the United States of America* 103: 6269–6274.

Mo Q, Lu SF, and Simon NG (2006) Dehydroepiandrosterone and its metabolites: Differential effects on androgen receptor trafficking and transcriptional activity. *Journal of Steroid Biochemistry and Molecular Biology* 99: 50–58.

Ogawa S, Choleris E, and Pfaff D (2004) Genetic influences on aggressive behaviors and arousability in animals. *Annals of the New York Academy of Sciences* 1036: 257–266.

Olivier B (2005) Serotonergic mechanisms in aggression. *Novartis Foundation Symposium* 268: 171–183; (discussion) 183–179, 242–253.

Ou XM, Chen K, and Shih JC (2006) Glucocorticoid and androgen activation of monoamine oxidase A is regulated differently by R1 and Sp1. *Journal of Biological Chemistry* 281: 21512–21525.

Roy-Byrne PP (2005) The GABA-benzodiazepine receptor complex: Structure, function, and role in anxiety. *Journal of Clinical Psychiatry* 66(supplement 2): 14–20.

Simon NG (2002) Hormonal processes in the development and expression of aggressive behavior. In: Pfaff DA, Etgen AM, Fahrbach SE, and Rubin RT (eds.) *Hormones, Brain, and Behavior*, pp. 339–392. San Diego: Academic Press.

Simon NG, Mo Q, Hu S, Garippa C, and Lu S (2006) Hormonal pathways regulating intermale and interfemale aggression. In: Bradley RJ, Harris RA, and Jenner P (eds.) *International Review of Neurobiology*, Vol. 73, pp. 99–124. London: Academic Press.

Trainor BC, Kyomen HH, and Marler CA (2006) Estrogenic encounters: How interactions between aromatase and the environment modulate aggression. *Frontiers in Neuroendocrinology* 27: 170–179.

Aggression: Neurochemical and Molecular Mechanisms

A Siegel, S Bhatt, and S Zalcman, University of Medicine and Dentistry of New Jersey – New Jersey Medical School, Newark, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview – Models of Aggressive Behavior

Aggressive behavior may be defined as behavior that threatens harm or leads to or causes harm, destruction, or damage to another organism. In this context, aggression is not a unitary phenomenon but, instead, reflects a variety of different behavioral processes that are contained under a single heading. A variety of animal research models of aggressive behavior have been used, which include fear-induced, maternal, intermale, irritable, sex-related, territorial, resident-intruder, and predatory aggression. With the exception of predatory aggression, these models of aggressive behavior share the following common features. They reflect a perceived or real threat, are aversive to the organism, are impulsive, display sympathetic signs, and are defensive in nature. Thus, these models can be reduced to the general category of affective (or defensive) aggression. In contrast, predatory aggression is quite different from the others, in particular because it requires planning, shows few autonomic signs, and is positively reinforcing to the organism. Consequently, the remainder of this article considers the neurobiological properties and functions associated with these two forms of aggressive behavior as determined from studies conducted mainly in the cat.

Affective defense (sometimes called defensive rage) occurs in nature in response to the presence of, or perceived presence of, a threatening stimulus such as another species within a cat's territory. It is characterized by arching of the back, retraction of the ears, piloerection, pupillary dilatation, marked hissing, and striking of the target species with the forepaw at the threatening object. This form of aggressive behavior can also be elicited by electrical or chemical stimulation of the midbrain periaqueductal gray (PAG) and medial hypothalamus.

In the animal kingdom, predatory (quite biting) attack behavior is manifested as hunting behavior. The attack is planned and is preceded by stalking of a specific prey object of another species, followed by biting of the back of the neck until the hunter kills the prey. In contrast to defensive rage, the cat displays few autonomic signs aside from mild pupillary dilatation. Predatory attack is induced following electrical stimulation of the lateral hypothalamus or ventrolateral aspect of the PAG and ceases immediately following termination of stimulation.

Regions and Pathways Mediating Defensive Rage and Predatory Attack

A principal goal in the study of the neurobiology of aggression and rage is to identify the underlying mechanisms of these behaviors. This objective requires knowledge of the anatomical substrates of these behaviors. The following discussion identifies the sites within the hypothalamus and midbrain from which each of these forms of aggression is elicited and the pathways mediating these behaviors to other regions of the brain stem and related areas of the central nervous system.

Defensive Rage Behavior

Defensive rage behavior can be elicited by electrical stimulation of wide regions along the rostrocaudal axis of the medial hypothalamus. Defensive rage is also elicited by electrical or chemical (i.e., glutamate analog) stimulation of the dorsolateral quadrant of mainly the rostral half of the PAG. Components or fragments of defensive rage can also be elicited from lower regions of the brain stem. These regions include the caudal PAG and pontine tegmentum and presumably lie along the descending pathways mediating this form of aggression.

The principal descending pathway from the medial hypothalamus subserving defensive rage behavior arises from the anterior medial hypothalamus, and its primary target is the dorsolateral aspect of the rostral half of the PAG. The functions of this pathway are mediated by glutamate acting on *N*-methyl-D-aspartate (NMDA) receptors in the PAG. Of particular interest is that other regions of the medial hypothalamus, such as the ventromedial nucleus, from which defensive rage can also be elicited, project rostrally to the region of the anterior medial hypothalamus from which the descending pathway to the PAG arises. Moreover, the anterior medial hypothalamus also receives significant inputs from components of the limbic system which modulate aggression and rage behavior. The converging inputs into the anterior medial hypothalamus thus enables this region to serve as a major site of integration for the expression of defensive rage behavior.

The second limb of the descending pathway for the expression of defensive rage behavior arises from the region of the dorsolateral PAG, which receives direct inputs from the anterior medial hypothalamus. The

efferent projections of this region of the PAG are directed to structures that mediate autonomic and somatomotor components of defensive rage behavior. There are several routes by which autonomic functions are activated from the PAG. One pathway includes a projection to the locus ceruleus, which in turn projects to the intermediolateral cell column of the thoracic and lumbar spinal cord. Converging inputs to these sympathetic regions of spinal cord are also mediated through projections to the solitary nucleus, whose axons then project to the ventrolateral medulla and from there to the intermediolateral cell column of the thoracic and lumbar cord. There are several regions that mediate the somatomotor components of defensive rage behavior. One set of targets includes the motor nuclei of the trigeminal and facial cranial nerves, which are associated with jaw opening essential for the vocalization aspect of the defensive rage response. A second target includes the nuclei of the reticular formation, which comprise, in part, reticulospinal fibers directed toward alpha and gamma motor neurons. Those neurons directed to the cervical cord presumably affect movements of the upper limbs that make up the striking component of the rage response. It is the collective integration of these two components at the levels of the medial hypothalamus and PAG which make up the defensive rage response. A separate ascending projection of the dorsolateral PAG supplies the rostrocaudal extent of the medial hypothalamus, much of which relates to the expression of defensive rage. This projection probably serves as a substrate for a positive feedback mechanism, thus increasing the likelihood that this response can be prolonged under dangerous conditions, which is of survival value to the animal.

Predatory Attack Behavior

Predatory attack behavior can be elicited by electrical stimulation, most typically of the perifornical lateral hypothalamus, ventrolateral aspect of the PAG, and ventral tegmental area. The principal origin of the descending projections of the hypothalamus is the region of the perifornical lateral hypothalamus, from which predatory attack is elicited. This region supplies the ventrolateral aspect of the PAG, ventral tegmental area, central tegmental fields of the midbrain and pons, locus ceruleus, and motor and main sensory nuclei of the trigeminal complex. The projections to the trigeminal complex are significant in that they provide the anatomical substrate for the jaw-closing reflex critical for the culmination of biting attack. The projections to the brain stem tegmentum presumably provide the initial neuron in a series of descending projections to the lower brain stem and spinal cord which is essential for other motor aspects of the attack response such as stalking and striking at the prey object.

Anatomical and functional relationship between the medial and lateral hypothalamus Although defensive rage behavior and predatory attack clearly reflect distinctly different forms of aggression that use separate and nonoverlapping pathways, they also relate to one another in a unique manner. Within the medial hypothalamus and with respect to defensive rage behavior, there are two classes of neurons. One is a projection neuron (already described), whose target is the dorsolateral aspect of the PAG and that constitutes the descending pathway for this form of aggression. The second is a neuron with a short axon that supplies the lateral hypothalamus. It is γ-aminobutyric acid (GABA)ergic and inhibits neurons in the lateral hypothalamus associated with predatory attack. Likewise, there are at least two classes of neurons in the lateral hypothalamus with respect to predatory attack. The first is a neuron with a long axon that constitutes the descending pathway for the expression of predatory attack, and the second is a GABAergic neuron with a short axon which supplies the medial hypothalamus and inhibits neurons in the medial hypothalamus associated with defensive rage. The likely functional significance of the reciprocal inhibitory pathways linking the medial and lateral hypothalamus is as follows. Because these two responses are mutually exclusive, the effective expression of one requires the suppression of the other. It is intuitive that a successful act of predation can be accomplished only when a predator quietly approaches the prey object, which requires the suppression of hissing and related component responses. Similarly, when defensive rage is required following the presence of a threatening stimulus, elements of predation serve no function and therefore are suppressed in order for the affective components of the response to become manifest. Collectively, the neuroanatomical relationships between the medial and lateral hypothalamus thus provide the essential substrates which are of survival value to the animal.

Limbic Structure-Associated Pathways Modulating Aggression and Rage

The limbic system consists of the following structures: amygdala, hippocampal formation, septal area, prefrontal cortex, and cingulate gyrus. They possess several common anatomical and functional features which distinguishes them from other regions of the brain, including (1) they receive secondary or tertiary

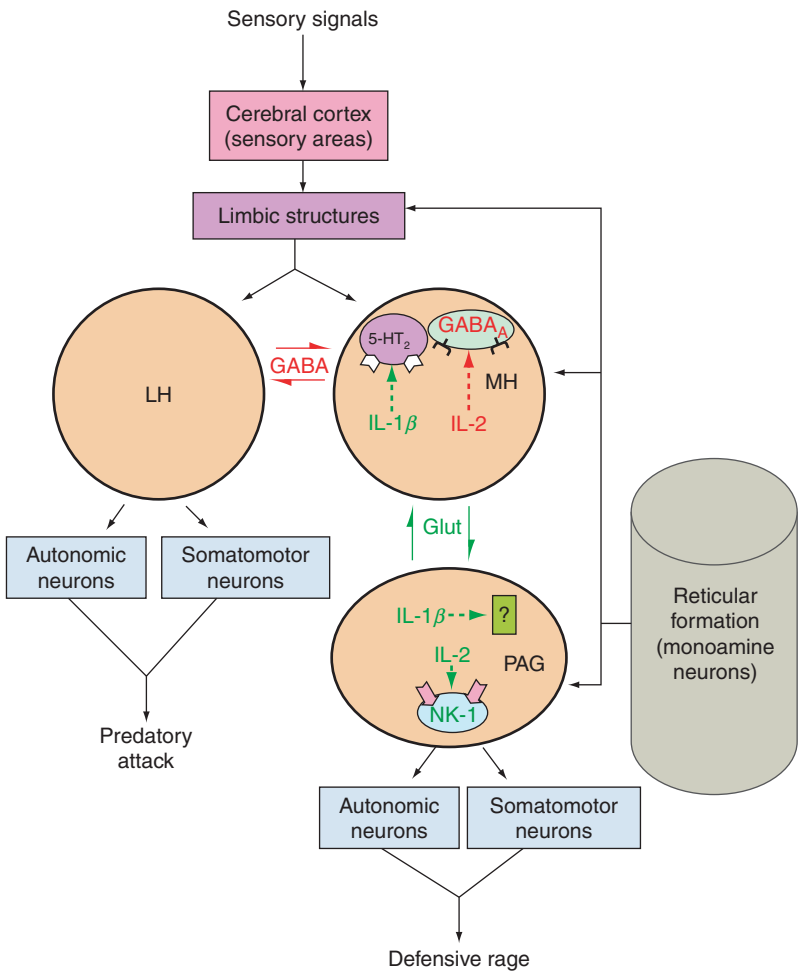

Figure 1 Schematic diagram depicting some of the critical neural mechanisms regulating aggression contained within limbic structures, such as the amygdala, hippocampus, and prefrontal cortex. These are activated by sensory signals that reach them through inputs from sensory regions of cerebral cortex, and these limbic neurons are further modulated by monoaminergic neurons situated within the reticular formation of the brain stem. Subsequent changes in levels of excitability within the limbic system are mediated through efferent pathways of limbic structures such as the fornix and stria terminalis to the hypothalamus, causing changes in excitability levels of hypothalamic neurons and, thus, directly affecting the neural mechanisms that control aggression and rage behavior. The expression of a predatory attack is generated in the lateral hypothalamus and the descending pathways, which engage both autonomic and somatomotor neurons of the lower brain stem. In a similar way, the expression of defensive rage is mediated through neurons in the medial hypothalamus and a glutamatergic descending pathway to the midbrain PAG, which, in turn, provides feedback through an ascending neuron to the medial hypothalamus. The medial and lateral hypothalami mutually inhibit one another through reciprocal GABAergic neurons. Also depicted in this diagram is the role of cytokines IL-1 and IL-2. IL-1 in the medial hypothalamus facilitates defensive rage, acting through 5-HT_2 receptors. IL-2 in the medial hypothalamus suppresses defensive rage, acting through $GABA_A$ receptors, but in the PAG this cytokine facilitates defensive rage, acting through NK_1 receptors. 5-HT, serotonin; GABA, γ-aminobutyric acid; IL, interleukin; LH, lateral hypothalamus; MH, medial hypothalamus; NK-1, neurokinin-1; PAG, periaqueductal gray. Reproduced from Zalcman S and Siegel A (2006) The Neurobiology of aggression and rage: Role of cytokines. *Brain, Behavior* and *Immunity* 20: 507–514, with permission from Elsevier.

sensory inputs which may vary among limbic structures, (2) they receive inputs from brain stem monoaminergic neurons, and (3) they project directly or indirectly to the hypothalamus and related structures of the brain stem. These combined sensory and monoaminergic inputs serve to activate the limbic structures to cause the powerful modulation of aggression and rage by virtue of their efferent projections to the efferent target structures (Figure 1). This section describes the modulating effects of limbic structures on aggression and rage and their associated pathways over which such modulation is mediated.

Amygdala

The amygdala, which consists of a complex of nuclei located in the rostral aspect of the temporal lobe, has received more attention than any other limbic structure

with respect to its relationship to emotional behavior. These studies have revealed that the amygdala is not uniform in its effects on aggression and rage. Instead, the effects are dependent on both the form of aggression and region of amygdala considered.

Excitation of the region of amygdala, including the medial nucleus and medial aspect of the basal complex, in the cat potentiates defensive rage behavior elicited from the medial hypothalamus, whereas excitation of the lateral and central nuclei or lateral aspect of the basal complex suppresses this response. The potentiating effects of the medial amygdala are mediated over the stria terminalis, which projects to the bed nucleus of the stria terminalis and rostral half of the medial hypothalamus, including the dorsomedial region and shell of the ventromedial nucleus. A primary neurotransmitter of this pathway has been identified as substance P (SP), acting on neurokinin (NK)-1 receptors in the medial hypothalamus. In contrast, excitation of the medial amygdala suppresses predatory attack behavior elicited from the lateral hypothalamus. Suppression is manifest via a disynaptic pathway in which the first limb includes the stria terminalis projection to the medial hypothalamus and the second a GABAergic (inhibitory) neuron projecting from the medial to lateral hypothalamus. The inhibitory effects of the amygdala on defensive rage behavior are mediated through a descending projection to the midbrain PAG. The neurotransmitter has been shown to be enkephalin acting through μ-opioid receptors in the PAG. In a parallel manner, excitation of the lateral amygdala potentiates predatory attack. Although the pathway has not been experimentally identified, it is likely to include fibers of the ventral amygdalofugal pathway projecting to the lateral hypothalamus.

Hippocampal Formation and Septal Area

The hippocampal formation in rodents and felines is arranged in a manner that extends from its rostral tip, situated in proximity to the septal area caudally, beneath the corpus callusum, and parallel to the lateral ventricle, entering the temporal lobe where it passes ventrally and rostrally, and ending at a position just caudal to the amygdala. The dorsal region of hippocampus (near the septal pole) suppresses predatory attack, whereas the ventral region (near the temporal pole) facilitates this form of aggression. The modulating properties of the hippocampal formation on aggressive behavior is mediated through the septal area, a structure which may thus be viewed as a relay nucleus of the hippocampal formation.

The pathways over which hippocampal modulation of aggressive responses are mediated probably involve the precommissural fornix – the branch of the fornix that supplies the septal area. The projection from the hippocampal formation is topographically organized in that fibers arising from the dorsal (septal) aspect project to the medial aspect of the lateral septal nucleus, whereas progressively more caudal regions of the hippocampal formation (toward the temporal pole) project to more progressively lateral aspects of the lateral septal nucleus. In turn, the medial aspect of the septal area, which receives inputs from the dorsal hippocampal formation, projects to the medial hypothalamus. In this manner, activation of the dorsal hippocampal formation excites neurons in the medial aspect of the septal area, which in turn excites neurons in the medial hypothalamus. Because the medial hypothalamus communicates with the lateral hypothalamus via a short GABAergic neuron, activation of the medial hypothalamus either directly or indirectly through the medial septal area or dorsal hippocampal formation causes suppression of predatory attack. Activation of the medial hypothalamus via a projection from the medial aspect of the septal area provides the anatomical basis for septal area potentiation of defensive rage behavior elicited from the medial hypothalamus.

Concerning the anatomical basis by which the ventral hippocampal formation and lateral aspect of the lateral septal nucleus potentiate predatory attack behavior, the likely pathways include a direct projection to the lateral aspect of the lateral septal nucleus, which in turn projects to and (presumably) activates neurons in the lateral hypothalamus which mediate the expression of predatory attack behavior.

Prefrontal Cortex and Anterior Cingulate Cortex

The prefrontal cortex and adjoining regions of the anterior cingulate cortex exert the powerful suppression of predatory attack and rage behavior. There are a number of descending fiber systems from the prefrontal cortex that could provide the anatomical substrate for the modulation of aggression. These include a monosynaptic connection consisting of a small number of neurons that project directly to the hypothalamus from the prefrontal cortex and several multisynaptic pathways involving connections with either the amygdala or mediodorsal thalamic nucleus. Of these pathways, there is experimental evidence that the modulating effects from the prefrontal cortex and anterior cingulate gyrus on aggression and rage are mediated primarily through the multisynaptic pathway involving the mediodorsal thalamic nucleus. With respect to this pathway, the mediodorsal thalamic nucleus projects to the hypothalamus through a series of interneurons in the midline thalamus. The

neurotransmitter for this system of neurons is not known but presumably includes a glutamate projection from the prefrontal cortex and anterior cingulate cortex.

Neurochemical and Molecular Mechanisms of Aggression

Neurotransmitters

An increasing body of knowledge has identified several different classes of neurotransmitters that play a role in regulating aggressive behavior. These include the following small molecule neurotransmitters: acetylcholine, GABA, and biogenic amines (dopamine, norepinephrine, and serotonin (5-HT)), and neuropeptides (opioid peptides, SP, cholecystokinin, and vasopressin). Some of these transmitters potentiate aggressive responses, whereas others have inhibitory properties. In addition, several of these neurotransmitters have been identified with specific neuroanatomical pathways, but others have not.

Excitatory neurotransmitters Studies conducted mainly in rodents and felines showed that cholinergic agents generally facilitate aggressive responses. These findings are based on the systemic application of agonists and antagonists that act through muscarinic receptors. Further studies supporting these findings have indicated that cholinergic agents produce their potentiating effects within the region of the medial hypothalamus and can, in fact, induce rage-like responses.

Both dopamine and norepinephrine have similar potentiating effects on both defensive rage and predatory attack. The mechanism presumably involves the activation of catecholaminergic neurons of the brain stem from such regions as the locus ceruleus for norepinephrine and the ventral tegmental area for dopamine, which project to widespread regions of the forebrain, including the hypothalamus and limbic system. Dopamine facilitation is mediated through dopamine D_2 and norepinephrine through α_2 receptors in the medial hypothalamus. Catecholaminergic facilitation of both defensive rage and predatory attack suggests that these neurotransmitters exert generalized potentiating effects on whatever ongoing responses are present during the epoch when these transmitters are activated. Serotonin distributed from brain stem raphe neurons to the PAG, hypothalamus and limbic system differentially affects defensive rage responses. Activation of 5-HT type 2 (5-HT_2) receptors in these regions facilitates defensive rage. The suppressing effects of 5-HT are indicated later in this article.

As already indicated, there are several pathways associated with the expression or modulation of aggressive or rage behavior whose primary neurotransmitters have been identified. These include glutamate neurons projecting from the medial hypothalamus to the PAG that act through NMDA receptors to mediate the expression of defensive rage, and SP neurons in the medial amygdala that project to the medial hypothalamus, which powerfully facilitates defensive rage and suppresses predatory attack (Figure 1). Glutamate neurons that project from the basal amygdala to the PAG, acting through NMDA receptors, also facilitate defensive rage behavior. Several peptides have also been identified that potentiate defensive rage behavior within the PAG. These include SP, acting through NK_1 receptors, and cholecystokinin (CCK), acting through CCK_B receptors. Within the anterior hypothalamus, it is well known that vasopressin (V) neurons play an important role in temperature regulation and drinking behavior. These neurons also potentiate aggressive responses in rodents, acting through V_{1A} receptors in the anterior hypothalamus. The pathway(s) associated with V functions in aggression have not been identified.

Inhibitory neurotransmitters Three neurotransmitters suppress defensive rage behavior. Activation of 5-HT_{1A} receptors in either the PAG or medial hypothalamus by 5-HT released from brain stem raphe neurons suppress this form of aggressive behavior. Powerful suppression of defensive rage behavior is mediated through μ-opioid receptors in the PAG. As already noted, these receptors are activated by enkephalinergic neurons arising in the central nucleus of amygdala. $GABA_A$ receptors in the medial and lateral hypothalamus suppress defensive rage and predatory attack behaviors, respectively. As previously indicated, $GABA_A$ receptors in the medial hypothalamus are activated by GABA neurons projecting from the lateral hypothalamus, and similarly, GABA neurons arising in the medial hypothalamus activate these receptors in the lateral hypothalamus (Figure 1). $GABA_A$ receptors in the PAG also suppress defensive rage behavior when activated by GABA neurons, the origin of which has not yet been identified.

Substances of abuse and psychotropic drugs It is well know that substances of abuse are associated with enhanced levels of aggressive behavior in both humans and animals. Such increases in aggression can be understood in terms of what is known about the link between neurotransmitters and aggression. Opioid peptides, when elevated, normally suppress aggressive behavior. Opioid withdrawal induces aggressive

behavior, and this phenomenon can be understood in terms of the absence of critical levels of enkephalins in the central nervous system. Alcohol, another substance of abuse, has been associated with homicide, child abuse, marital assaults, rape, and verbal abuse. In the animal literature, parallel findings have been observed as dose-dependent increases in aggression have been reported in various species. Altered 5-HT levels have been associated with aggressive alcohol abusers. Other neurotransmitters may be involved, but the data are incomplete at this time. Cocaine, which results in heightened levels of aggression, has been shown to influence dopamine release and retard its removal from the brain. Cocaine's effects on aggression have also been associated with activation of 5-HT_3 receptors. Information concerning the pathways mediating the potentiating effects of substances of abuse on aggressive behavior is lacking in the literature.

Genes and Cytokines

Evidence from genetic studies Two methods have been applied to the study of the role of genetics in aggression and rage behavior. The first involves traditional breeding methods, and the second involves the use of genetic engineering to produce knockout mice in which specific receptors are absent.

Concerning the first approach, increased levels of aggression are present under the following conditions: (1) in animals selectively bred for heightened sensitivity to cholinergic agonists; (2) in animals bred in a manner producing higher levels of brain dopamine levels; and (3) in animals bred for selective loss of 5-HT axons. Thus, these findings have generally supported the findings obtained from pharmacological and neurochemical approaches summarized in the previous section. Concerning the second approach, mutant mice lacking the 5-HT_{1B} receptor or which display decreased 5-HT turnover in the brain have increased levels of aggressive behavior. In contrast, mice lacking the V1b receptor have reduced aggressive behavior.

Cytokines and aggression Cytokines are pleiotropic cell–cell signaling proteins or glycoproteins produced by various cells types in the periphery and in brain. They were initially described for their effects on immunity, including innate and acquired immunity. Cytokines also play fundamental roles in other functions, including development and feeding, among other processes. Cytokines exert their biological effects by binding their receptors, which in turn activates signaling pathways. Cytokine receptors are present on numerous cell types in periphery and the central nervous system, and they may act in autocrine, paracrine, and endocrine manners. Thus, a given cytokine can influence activity in the cell from which it is released, as well as local and distant cells. In the latter regard, some cytokines of peripheral origin may influence brain activity by transporting across the blood–brain barrier. Cytokines may also affect brain activity by entry through the circumventricular organs or by indirect routes, including the activation of sensory afferents and stimulation of cytokine release by brain endothelial cells. Because cytokines and their receptors are present in the brain, they can act as endogenous neuromodulators in regions associated with the expression of aggression and rage, including the hypothalamus and PAG.

A further possible linkage between cytokines and aggression is derived from observations obtained from human studies. Two types of studies have been conducted – the first concerning patient populations and the second involved experiments using healthy adult subjects. Studies involving patient populations, cytokine immunotherapy for the treatment of such disorders as acquired immunodeficiency syndrome (AIDS), cancer, and hepatitis C have shown that such therapy increased aggressive behavior as determined by measures of anger, hostility, and irritability. Similar conclusions were drawn from studies involving healthy adult subjects. In one study, it was demonstrated that experimentally induced enhancement of proinflammatory cytokines was correlated with increases in hostility scores. In another study, a positive correlation was drawn between the presence of hostile marital conflict and the increased production of plasma proinflammatory cytokines.

Animal studies have also indicated a relationship among the enhancement of cytokine production, immune cell activity, and aggression. To cite one example, animals displaying heightened levels of territorial behavior and aggression also revealed higher levels of interferon-γ (IFN-γ) and interleukin-2 (IL-2) production. Several questions can be raised from the results of these studies, including: (1) Which cytokines relate to the control of aggression and rage? (2) Where in the brain are these cytokines located? (3) What are the possible mechanisms underlying cytokine modulation of aggressive behavior? Recent studies have provided partial answers to these questions, indicating that IL-1 and IL-2 do potently affect defensive rage behavior and that the nature of these effects depends on the region of the brain where receptors for these cytokines are activated. In one study, it was demonstrated that IL-1β in the medial hypothalamus facilitates defensive rage behavior and that such facilitation is mediated through 5-HT_2 receptors. In contrast, activation of IL-2 receptors in the same region of the medial hypothalamus suppresses

defensive rage behavior. The suppressive effects of IL-2 appear to be mediated through $GABA_A$ receptors. Thus, these two cytokines use different neurotransmitter mechanisms within the medial hypothalamus to differentially modulate defensive rage behavior. In addition, activation of IL-2 receptors in the PAG potentiates defensive rage behavior, indicating that the locus of activation of this receptor is critical to the nature of its effects on defensive rage behavior. The potentiating effects of IL-2 in the PAG on defensive rage are mediated through SP NK_1 receptors, a receptor that has known excitatory effects on defensive rage behavior (see Figure 1). These findings raise an additional question – namely, how can the same cytokine found in different regions of the brain alter defensive rage behavior in opposite ways? An answer to this and related questions concerning how cytokines modulate aggression and rage can only be answered by future studies directed at understanding the signaling pathways associated with these cytokines and their relationship to the neurochemical milieu of the various regions of the forebrain and brain stem associated with the expression and modulation of aggression and rage behavior.

Conclusion

Two principal forms of aggressive behavior – defensive rage and predatory attack – are induced following excitation of neurons in the hypothalamus and PAG. Defensive rage is elicited from the medial hypothalamus and dorsolateral PAG, whereas predatory attack is elicited from the lateral hypothalamus, ventral tegmental area, and ventrolateral PAG. The primary pathways over which defensive rage occurs include a projection from the medial hypothalamus to the PAG and from the PAG to autonomic and somatomotor neurons of the lower brain stem. Predatory attack is mediated over descending fibers from the perifornical lateral hypothalamus to the midbrain and pontine tegmentum as well as to motor nuclei of cranial nerves V and VII. The expression of these forms of aggression are modulated from limbic structures and their output pathways to the hypothalamus and PAG. These include projections from (1) the amygdala to the hypothalamus via the stria terminalis (containing SP) and ventral amygdalofugal pathways as through a descending pathway to the PAG (containing enkephalin); (2) the hippocampal formation to the hypothalamus via a relay in the septal area; and (3) the prefrontal and anterior cingulate cortex projections via the medial thalamus to the hypothalamus. Neurotransmitters (and neuromodulators) that are known to regulate aggression and rage include monoamines, excitatory and inhibitory amino acids, and various peptides, including cytokines.

See also: Aggression: Hormonal Basis; Emotional Hormones and Memory Modulation; Gene Expression Regulation: Steroid Hormone Effects; Hormones and Behavior.

Further Reading

Adamec RE (1990) Role of the amygdala and medial hypothalamus in spontaneous feline aggression and defense. *Aggressive Behavior* 16: 207–222.

Bandler R and Shipley M (1994) Columnar organization in the midbrain periaqueductal gray: Modules for emotional expression. *Trends in Neuroscience* 17: 379–389.

Hassanain M, Zalcman S, Bhatt S, and Siegel A (2005) Interleukin-1β (IL-1β) and IL-1β type 1 receptors in the medial hypothalamus in defensive rage behavior in the cat. *Brain Research* 1048: 1–11.

Meloy JR (1988) *The Psychopathic Mind: Origins, Dynamics and Treatment*. Northvale, NJ: Jason Aronson.

Meloy JR (1997) Predatory violence during mass murder. *Journal of Forensic Science* 42: 326–329.

Miczek KA and Fish EW (2005) Dopamine, glutamate and aggression. In: Schmidt WJ and Reith MEA (eds.) *Dopamine and Glutamate in Psychiatric Disorders*, pp. 237–266. Totawa, NJ: Human Press.

Miczek KA, Haney M, Tidey J, Vivian J, and Weerts E (1994) Neurochemistry and pharmacotherapeutic management of aggression and violence. In: Reiss A, Miczek KA, and Roth J (eds.) *Understanding and Preventing Violence,* vol. 2, pp. 245–514. Washington, DC: National Academy Press.

Monroe RR (1978) *Brain Dysfunction in Aggressive Criminals*. Lexington, MA: Lexington Books.

Nunn K (1986) The episodic dyscontrol syndrome in childhood. *Journal of Child Psychology and Psychiatry* 27: 439–446.

Petitto JM, Gariepy JL, Gendreau PL, Rodriguiz R, Lewis MH, and Lysle DT (1994) Differences in NK cell function in mice bred for high and low aggression: Genetic linkage between complex behavioral and immunological traits? *Brain Behavior and Immunity* 13: 175–186.

Siegel A (2005) *The Neurobiology of Aggression and Rage*. Boca Raton, FL: CRC Press.

Siegel A and Brutus M (1990) Neural substrates of aggression and rage in the cat. In: Epstein AE and Morrison AR (eds.) *Progress in Psychobiology and Physiological Psychology,* vol. 14, pp. 135–233. San Diego, CA: Academic Press.

Siegel A, Roeling TAP, Gregg TR, and Kruk MR (1999) Neuropharmacology of brain-stimulation-evoked aggression. *Neuroscience and Biobehavioral Reviews* 23: 359–389.

Vitiello B, Behar D, Hunt J, Stoff D, and Ricciuti A (1990) Subtyping aggression in children and adolescents. *Journal of Neuropsychiatry* 2: 189–192.

Zalcman S and Siegel A (2006) The neurobiology of aggression and rage. *Brain Behavior and Immunity* 20(6): 507–514.

Aging and Memory in Animals

P R Rapp, Mount Sinai School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

From a life-span perspective, aging comprises the late component of a genetically determined program of development, maturation, and senescence, interacting with a complex array of environmental factors. Commonly viewed as a process of deterioration, growing old is associated with sharply increased risk for many diseases and disabilities that compromise independent living, placing a heavy burden on families, caregivers, and society. Nonetheless, a majority of people successfully accommodate certain physical signs of aging, and in fortunate cases, old age can represent a rewarding period of new intellectual engagement, novel pursuits, and achievement. Such positive outcomes become increasingly unlikely in the face of failing cognitive function, and partly for this reason, disorders that lead to diminished mental capacity are among the most feared consequences of aging. The leading cause of dementia, Alzheimer's disease, ultimately results in a dense amnesia, gradually robbing patients of the lifetime of memories that define their personal history and identity. Even in the absence of disease, many people experience memory impairment that, although relatively mild, can cause considerable anxiety and compromise the quality of life. As the populations of industrialized countries rapidly age (Figure 1), we face a growing challenge of identifying ways to promote healthy cognitive aging and to maximize optimal functioning.

Research has illuminated many of the key features of age-related cognitive decline in humans; enabled by advances in *in vivo* brain imaging, it has begun to reveal how the neural systems organization of memory is altered. Studies of cognitive aging in humans are complicated by a variety of methodological factors, however, and they are limited by the range of applicable experimental approaches. A vexing issue is that individuals in preclinical stages of Alzheimer's disease and other disorders affecting cognition are difficult to identify with confidence. As a consequence, it is often unclear to what degree observed impairment is attributable to disease rather than normal nonpathological aging. Relating these deficits to underlying biological causes is also problematic, and although noninvasive imaging techniques continue to yield remarkable discoveries, defining the neurobiological mechanisms of cognitive aging requires experimental approaches not suitable for investigation in humans. Research in animal models has played a critical role in efforts to address these issues.

Cognitive Aging – A Neuropsychological Framework

There is a broad consensus on a number of central concepts concerning the neuropsychology of normal memory in young adults. Rather than comprising a unitary, monolithic capacity, multiple forms of memory can be distinguished according to the qualities of the information remembered, endurance, flexibility, and a variety of other characteristics. A major advance in this field is the recognition that the existence of multiple forms of memory follows from the organization of brain systems that mediate these capacities. Converging evidence from rats, monkeys, and amnesic humans, for example, indicates that normal memory for the facts and events of daily life critically requires a collection of anatomically related structures in the medial temporal lobe, including the hippocampus and adjacent parahippocampal cortical areas. Damage to the prefrontal cortex, by comparison, largely spares this declarative or episodic type of memory and, instead, leads to deficits in a variety of memory-related processes, including the strategic use of remembered information when confronted with novel circumstances and the ability to recall the source from which information was acquired.

Informed by this background, research detailing the specific profile of behavioral deficits that emerge during aging can provide a window on the functional status of the neural systems that mediate memory and other cognitive processes. The Morris water maze has been used widely in studies of this sort in rats, partly because memory assessed by this procedure critically requires the functional integrity of the hippocampus. Across trials in this task, animals learn the spatial location of an escape platform hidden in a pool of clouded water, and navigation is guided by the memory of the platform position in relation to cues surrounding the maze. Young adult rats acquire this task rapidly, learning to swim directly to the escape platform from any point around the perimeter of the apparatus. By roughly 24 months of age, acquisition rates are impaired in a substantial proportion of rats, and by comparison with young adults, spatial searching is less tightly focused on the escape location during probe testing, when the platform is removed from the apparatus. There are a variety of reasons to presume that this pattern reflects a genuine impairment in learning and memory rather than the effects of

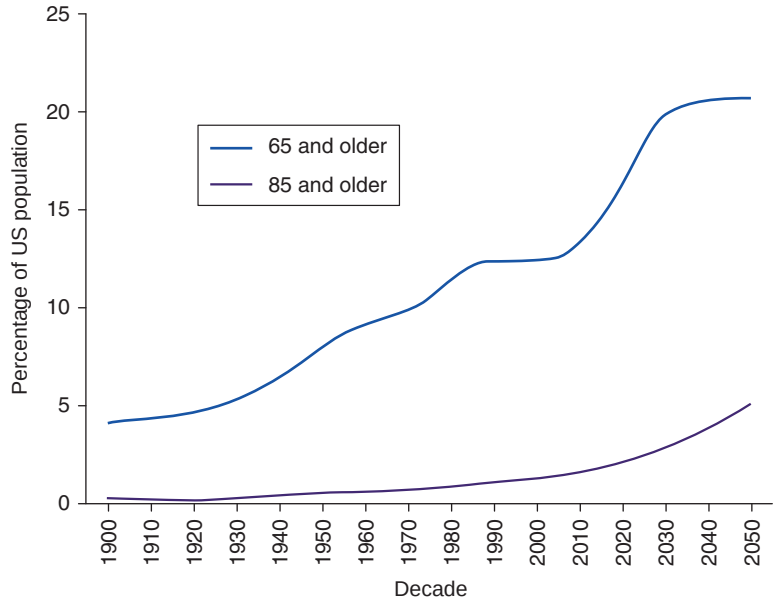

Figure 1 Society is aging rapidly; the historical and projected percentages of the US population that is over 65 and 85 years. Results compiled from US Census Bureau figures.

aging on sensorimotor function, motivation, and other performance factors. In a variant of testing that shares many of the same performance demands as the hidden goal task, rats swim to a visible escape platform that either protrudes slightly above the water's surface or is marked with a prominent visual cue. Performance in this cue-approach version of the water maze is relatively insensitive to aging, suggesting that, in the absence of a requirement to learn and remember information about the spatial location of the escape platform, age-related deficits in swimming ability, the motivation to escape, and visual acuity are not sufficient to prevent performance comparable to that of young adults. Notably, a qualitatively similar pattern of impaired spatial learning, together with preserved cue-approach performance, is seen in rats following the disruption of hippocampal function.

Age-related impairment in learning and memory for spatial information is not peculiar to rats, and parallel deficits have been documented in both humans and monkeys. The hippocampus is not restricted to processing spatial information, however, and in efforts to define how aging influences medial temporal-lobe memory, research in monkeys has relied predominantly on other sorts of assessments. In a particularly widely used procedure, that is delayed nonmatching-to-sample (DNMS), trials consist of two phases; the presentation of a sample item followed by a recognition test. A retention interval is imposed between the two, and monkeys demonstrate recognition by choosing the novel object (or nonmatch) when it is presented together with the previously viewed sample. Training typically proceeds by testing subjects with a short retention interval of 8 or 10 s until they master the nonmatching rule and subsequently introducing more challenging memory delays, up to 10 min or longer.

Visual object recognition in young adult macaque monkeys declines gracefully over increased retention intervals, confirming that the task effectively taxes memory. The average life span in rhesus macaques is less than 25 years, and monkeys at this age and older display modest but reliable deficits in visual recognition. Specifically, whereas older subjects score as well as controls at relatively short delays, recognition accuracy falters at longer intervals. This profile is significant because it demonstrates that sensorimotor function, motivation, and attention in the aged monkey are sufficiently intact to support normal performance and that age-related DNMS deficits emerge specifically in relation to increased demands on memory. In addition, this delay-dependent pattern of impairment is qualitatively similar to the deficits observed in young monkeys with experimental damage involving the hippocampal region. Thus, across both rats and monkeys, behavioral studies converge, concluding that memory mediated by the hippocampal system is vulnerable to normal aging.

Individual Variability in Cognitive Aging

An important insight from recent research is that the cognitive effects of aging vary considerably across individuals such that many experience little or no decline

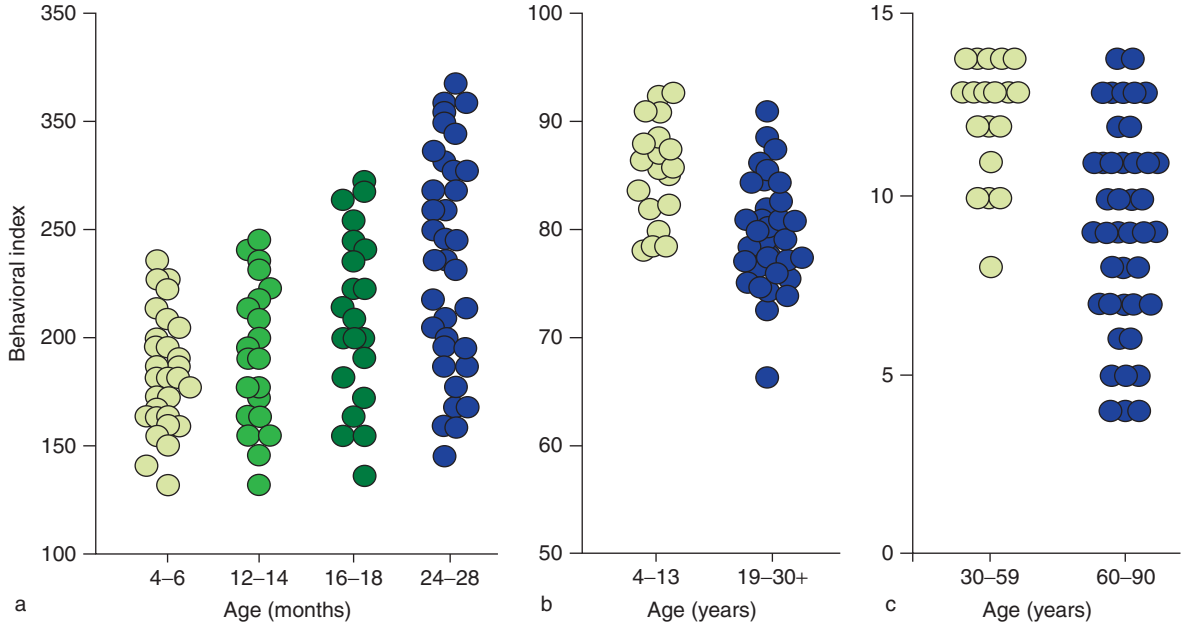

Figure 2 The status of memory varies considerably across aged individuals: (a) spatial memory in rats; (b) object recognition memory in monkeys; (c) delayed recall in humans. For rats (a), low scores represent better memory; in monkeys (b) and humans (c), high values reflect better performance. Symbols signify scores for individual subjects. Note that in all species the status of memory among the aged is distributed across a broad range, from individuals that score on a par with the best young adults to other individuals that exhibit substantial impairment. (a) Data courtesy of M Gallagher, Johns Hopkins University; (b) data courtesy of P Rapp, Mount Sinai School of Medicine; (c) data courtesy of M Albert, Johns Hopkins School of Medicine.

but others develop substantial impairment (Figure 2). Individual variability is widely recognized as a prominent feature of human aging, but it has often been considered a nuisance factor, making it difficult to draw firm conclusions about the likely course of decline. In addition, impaired function in some aged people is likely a prodromal harbinger of disease, obscuring individual variability associated with the process of normal aging itself. Here again, studies in animal models that are spared the spontaneous, dementia diseases of human aging have proved key.

The status of spatial learning and memory varies substantially among aged rats and is continuously distributed from subjects that perform on par with the best young adults to other aged individuals that display marked impairment, outside the range of adult control values (Figure 2(a)). Findings from monkeys are similar, indicating that roughly 40% of aged subjects exhibit marked recognition memory deficits, relative to young animals, as assessed by DNMS (Figure 2(b)). These findings mirror observations in humans, establishing the important concept that, in the absence of frank disease, marked cognitive decline is not an inevitable consequence of growing older (Figure 2(c)). In addition, the recognition of individual differences in the cognitive outcome of aging has enabled a powerful approach for exploring the neurobiological basis of dysfunction. Whereas aging research traditionally has comprised comparisons between groups distinguished on the basis of chronological age, an alternative approach involves the analysis of age-matched groups that differ according to neuropsychological indices of aging. As described in subsequent sections, studies adopting this approach have substantially advanced progress toward defining the neural alterations that specifically distinguish aged individuals with cognitive deficits from others of the same age with preserved function.

Neurophysiology of Cognitive Aging

Neuropsychological studies are an important source of evidence suggesting that memory mediated by the hippocampus is vulnerable to aging. Cognitive data alone, however, are a secondary proxy, and other sorts of approaches are needed to directly document the status of information processing in the aged hippocampus. An illuminating strategy used by a number of investigators involves recording the firing patterns of hippocampal neurons as freely moving rats navigate a maze apparatus. Under these conditions, neuronal activity in a large proportion of hippocampal pyramidal cells reliably and selectively increases when subjects move through a specific spatial location in the testing environment – a phenomenon referred to as place-field firing. Many

features of place-field activity, including the incidence, selectivity, and specificity of location-related firing, appear normal in the aged hippocampus, even among rats with pronounced spatial learning and memory impairment. This sparing counts against the idea that deficits in memory result from generalized, nonspecific deterioration in the functional integrity of the hippocampus. In fact, recent evidence indicates that the influence of aging on hippocampal firing patterns is highly specific and coupled with the status of spatial memory measured behaviorally.

One means of defining the particular constellation of stimuli that controls place cell activity involves the use of probe conditions in which subsets of cues in the testing environment are manipulated independently. After mapping the place fields of neurons in a maze setting that is held constant, for example, the constellation of experimenter-defined tactile and odor cues on the surface of the apparatus can be rotated in the opposite direction of the extra-maze visual and auditory stimuli. Recorded under these conditions, a large percentage of neurons in the young adult hippocampus either stop firing or develop entirely new place fields, unrelated to the rotated configuration of either the intra- or extra-maze cues. Such remapping is thought to enable the flexibility of hippocampal information processing, providing the basis for encoding the relevant predictive relationships between familiar stimuli when they appear in novel circumstances. A similar incidence of flexible place-field remapping is observed in aged rats that perform normally on the hidden platform version of the Morris water maze. In contrast, aged rats with documented deficits in spatial learning and memory exhibit far less flexibility in place-field activity. Neuronal firing in this subpopulation appears abnormally rigid and disproportionately driven by the configuration of extra-maze cues in the testing environment relative to other available stimuli.

As a consequence of encoding a reduced scope of available information, it might be predicted that the aged hippocampus is prone to interference between environments that share a subset of cues. Consistent with this prediction, when aged rats repeatedly explore an environment containing cues that were also present in another familiar setting, place-field remapping occurs substantially more frequently than in young subjects (**Figure 3**). An orthogonal pattern of results is observed when unit activity is recorded in a highly familiar environment, followed by exploration of a novel setting, and subsequent return to the original environment. Place-field firing in memory-impaired aged rats initially persists across the familiar and novel conditions, and repeated exposure is needed to induce the remapping observed in young rats from the outset. Related findings also indicate that at least some features of age-related alteration in place-field firing are regionally selective across the principal cell layers of the hippocampus, appearing to be particularly pronounced in the CA3 field. Considered together, the picture emerges that prominent features of normal cognitive aging are coupled with reliable changes in key computational capacities of the hippocampus and that the specific nature of these alterations contributes to the spatial learning and memory impairment seen in a substantial proportion of aged individuals.

Although the precise cell biological basis of age-related change in hippocampal neurophysiology remains to be defined, recent research has considerably narrowed the range of plausible accounts. A particularly long-standing idea is that the complex constellation of neurobiological alterations that occurs during aging ultimately results in neuron death in vulnerable brain areas such as the hippocampus and that this loss of structural integrity is the proximal cause of cognitive aging. Enabled by the development and validation of efficient methods of morphometric quantification, however, studies taking advantage of these stereological tools have prompted a new consensus. Across all the animal models examined, including rats and monkeys, the findings consistently demonstrate that marked neuron loss in the hippocampus and associated components of the medial temporal-lobe memory system is not a necessary consequence of normal aging. Moreover, among investigations incorporating behavioral assessment prior to sacrifice and stereological analysis, age-related deficits in learning and memory frequently occur in the absence of detectable neuron loss in these structures (**Figure 4**). Alongside this preservation, synaptic connectivity that critically participates in hippocampal functional plasticity is compromised, and in some cases, these changes are correlated with individual variability in the cognitive outcome of aging.

Enduring changes in synaptic strength, mediated by long-term potentiation (LTP)-like mechanisms, are widely presumed to underlie the establishment of place-field firing and other electrophysiological correlates of behavior that are vulnerable to aging. Considerable attention has therefore focused on the possibility that deficits in LTP and other plasticity mechanisms might contribute to the cognitive outcome of aging. Many fundamental electrophysiological properties of neurons throughout the hippocampus are preserved during aging, including the resting membrane potential, the depolarization threshold for action potential firing, and, under maximal nonphysiological stimulation conditions, the induction of LTP. Studies using weaker stimulation, however, reveal that the threshold for LTP induction is increased in both the dentate gyrus

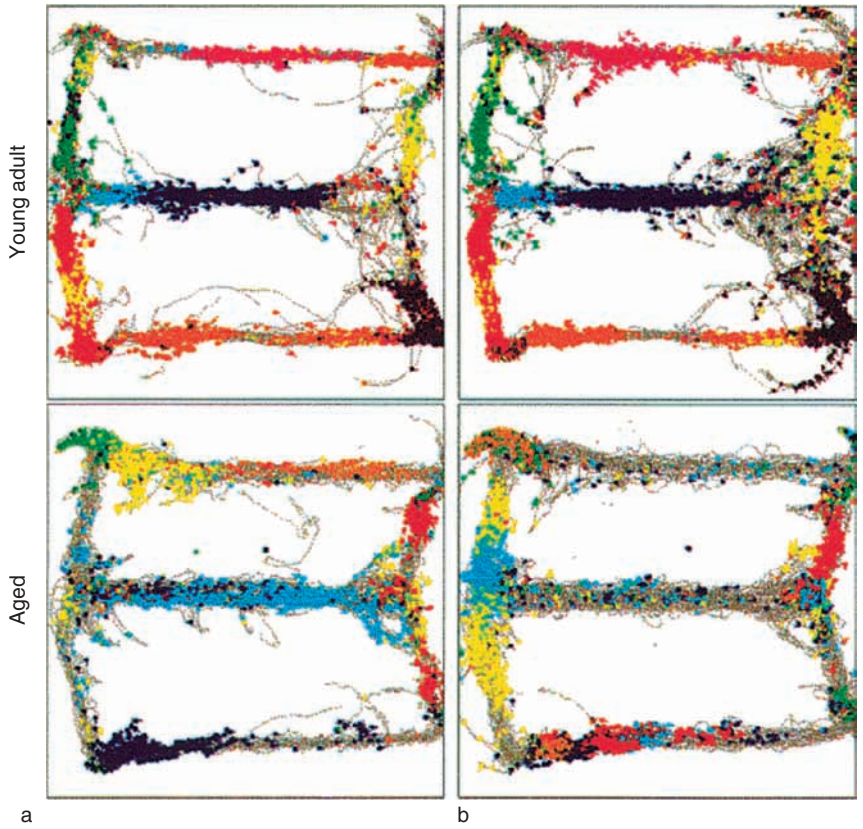

Figure 3 Neuronal firing of multiple individual hippocampal pyramidal neurons as young and older rats explored a figure-8 maze on two occasions: (a) maze exploration A; (b) maze exploration B. Light gray lines illustrate the path subjects took as they navigated the apparatus; the locations at which neurons fired above background levels are represented as color-coded dots. As shown, the coding of spatial location is prone to disruption in the hippocampus of aged rats with memory deficits. Note that in the young rat, the distribution of place-field firing is highly consistent across the two bouts of exploration. In contrast, on roughly 30% of such occasions, the aged hippocampus exhibited completely different place-field mappings, as though the familiar test setting was represented as multiple environments. Reprinted by permission from Macmillan publishers Ltd: *Nature* (Barnes CA, Suster MS, Shen J, and McNaughton BL (1997) Multistability of cognitive maps in the hippocampus of old rats. *Nature* 388: 272–275), copyright (1997).

and CA1 fields of the aged hippocampus. The long-term maintenance of synaptic enhancement, which requires gene expression, protein synthesis, and synaptic remodeling, is also compromised and declines toward the baseline more rapidly in aged subjects than young adults. A significant challenge for future research is to identify the relevant linkages between the effects of aging documented at different levels of analysis (determining whether synaptic alterations in critical circuitry account for changes in the computational capacities of hippocampal neurons) and, ultimately, age-related deficits in learning and memory.

Although it seems reasonable to speculate that age-related changes in synaptic plasticity contribute to observed alterations in hippocampal place-field firing and spatial learning and memory impairment, emerging evidence reveals considerable complexity in the underlying mechanism of these effects. LTP in the CA1 field of the hippocampus reflects the operation of two components: the *N*-methyl-D-aspartate receptor (NMDAR)-dependent pathway and a second, less well-studied form mediated by voltage-dependent Ca^{2+} channel (VDCC) activity. Using pharmacological manipulations to dissect the contribution of these components, current evidence indicates that, even under conditions in which the magnitude and endurance of compound LTP is similar, there is a shift toward a greater role of VDCC-mediated enhancement in the aged hippocampus compared to the young hippocampus. Related findings suggest that aging is also accompanied by a change in the bidirectional balance between synaptic enhancement and long-term depression (LTD). Prolonged low-frequency stimulation that is without effect in the young hippocampus produces an NMDAR-dependent LTD at the aged CA3 to CA1 synapses, and this vulnerability appears at least partly attributable to altered Ca^{2+} signaling. Studies in behaviorally characterized rats extend these observations, suggesting that, whereas age-related change in NMDAR-mediated LTD is unrelated to individual

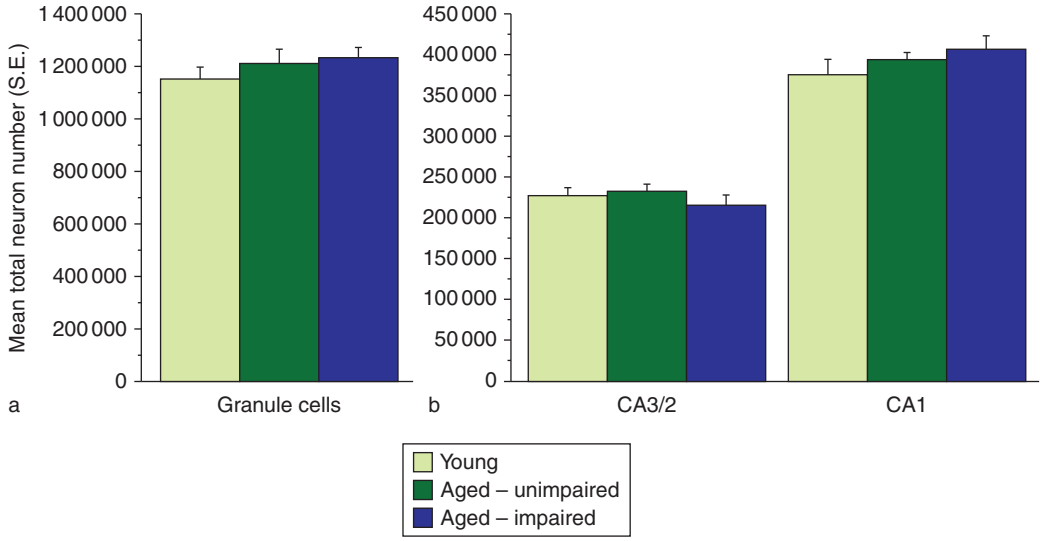

Figure 4 Estimated total number (unilateral) of neurons in the principal cell fields of the hippocampus for young rats and aged (unimpaired and impaired) rats based on their performance in a water maze test of spatial learning and memory: (a) granule cells; (b) CA3/2 and CA1. The results illustrate that neuron number is stable as a function of chronological age and cognitive status. Memory impairment during normal aging does not require neuron death in the hippocampus. Reproduced from Rapp PR and Gallagher M (1996) Preserved neuron number in the hippocampus of aged rats with spatial learning deficits. *Proceedings of the National Academy of Sciences of the United States of America* 93: 9926–9930, copyright 1996, National Academy of Sciences, USA.

differences in the cognitive outcome of aging, NMDAR-independent LTD is increased in magnitude and correlated with preserved spatial memory in aged rats. Together, these findings indicate that the basis of age-related decline in hippocampal memory is likely to involve dynamic interactions between multiple plasticity pathways and that active compensatory mechanisms are available in support of optimal hippocampal function during aging. Defining and harnessing these mechanisms through pharmacological or other means are important goals of current research efforts.

Horizons in Research on Cognitive Aging

Aging beyond the Hippocampus

This article has centered on medial temporal-lobe memory as a particularly well-studied system in animal models of age-related cognitive decline. This focus, however, should not be taken to suggest that this form of memory is affected in isolation. Indeed it is entirely clear that other systems are also vulnerable, prominently including cognitive capacities supported by the prefrontal cortex. Relevant research in rats has begun to emerge only recently, in part because the normative functional organization of the prefrontal cortex has proved challenging to characterize. Nonetheless, comparable to findings in humans, aged rats exhibit deficits in attentional set shifting, or the ability to update and modify behavior in response to altered task contingencies. This capacity is a member of a broad domain, termed executive function, that mediates the strategic use and manipulation of memory in support of adaptive behavior. Parallel findings in monkeys demonstrate age-related impairment on a nonhuman primate adaptation of the Wisconsin Card Sorting Test, a widely used neuropsychological assessment of executive function in humans. Another, particularly well-documented signature of cognitive aging is that older monkeys display robust deficits relative to controls on delayed-response tests that measure the spatiotemporal organization of memory. This procedure challenges animals to remember and select the location of a reward presented prior to a delay interval of several seconds or more. Because each location is rewarded frequently over the course of the many trials that make up daily test sessions, the opportunity for interference is substantial, placing considerable demands on the temporal organization of memory to distinguish information relevant to the current trial from target locations used earlier. Delayed-response performance, like other measures of executive function, critically requires the functional integrity of the prefrontal cortex. Thus, taken together, the available neuropsychological evidence prompts the conclusion that cognitive capacities supported by the prefrontal cortex are vulnerable to normal aging. In addition, because these signatures

of aging are largely unrelated to the status of medial temporal memory, the emerging consensus is that cognitive decline reflects the independent vulnerabilities of multiple memory-related brain structures rather than a process of widespread generalized deterioration.

Intervention

Considerable attention in research on the biology of aging has been directed at efforts to extend the life span. Improving the quality of life is at least as important a goal, however, and in this regard, animal models can be valuable in informing research efforts to promote healthy cognitive aging. This perspective, emphasizing the importance of bringing into convergence evidence from multiple levels of analysis, from basic research to human clinical trials, is illustrated by a consideration of the results from the Women's Health Initiative (WHI). The WHI is among the largest double-blind, placebo-controlled clinical trials ever mounted in the United States and is aimed at defining the health and cognitive effects of hormone replacement therapy in postmenopausal women. A large body of basic research had provided ample reason to suspect that ovarian hormone replacement following menopause might offer significant benefit in terms of neuroprotection and cognitive function, and indeed, results from a number of relatively small observational studies in women supported this prediction. The surprising result of the WHI, in contrast, is that, in addition to significantly increasing the risk for a number negative health outcomes, including cancer and stroke, hormone replacement offered no benefit in terms of cognitive function. In fact, among the few who developed Alzheimer's disease, the incidence was nearly double in treated women compared with placebo controls.

Evidence from animal research suggests a number of factors that might contribute to these unexpected results. In rats, for example, the cognitive benefit of exogenous hormone replacement depends on initiating treatment soon after the onset of estrogen deficiency, and efficacy is compromised with delayed intervention. That there is a temporal window of opportunity for influencing cognitive function with ovarian hormone treatment is relevant to the WHI because, in that study, women averaged 65 years of age at the beginning of the trial, nearly 15 years older than the typical onset of menopause. The specific schedule of dosing may also be important, and adopting the standard clinical practice at the time, the regimen used in the WHI consisted of the chronic, daily administration of conjugated equine estrogen (CEE), with or without a synthetic progestin (medroxyprogesterone acetate (MPA)). Chronic, continuous treatment, however, may lead to the habituation or desensitization of the neurobiological responses that mediate the cognitive influence of estrogen. Supporting this speculation, acute dosing in rats is reportedly at least as, if not more, effective than continuous replacement. Findings in nonhuman primates are similar, demonstrating that a single injection of estradiol every 21 days in aged ovariectomized monkeys substantially reverses a key signature of cognitive aging relative to age-matched vehicle control values. The specific formulations used for replacement may also be critical, based on *in vitro* evidence that estradiol and progesterone produced by the ovaries can have very different, and sometimes opposite, cell biological effects compared to the CEE and MPA compounds typically prescribed in clinical practice. Although the precise influence of these and other variables remains to be defined, current research establishes the value of investigations in animal models for addressing these issues and of broader efforts aimed at the rational development of strategies to promote healthy cognitive aging.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Cognition in Aging and Age-Related Disease; Episodic Memory: Assessment in Animals; Executive Function and Higher-Order Cognition: Assessment in Animals; Functional Neuroimaging Studies of Aging; Gene Expression in Normal Aging Brain; Lipids and Membranes in Brain Aging; Long-Term Depression: Cerebellum; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Rodent Aging; Spatial Memory: Assessment in Animals; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Bachevalier J, Landis LS, Walker LC, et al. (1991) Aged monkeys exhibit behavioral deficits indicative of widespread cerebral dysfunction. *Neurobiology of Aging* 12: 99–111.

Barnes CA, Suster MS, Shen J, and McNaughton BL (1997) Multistability of cognitive maps in the hippocampus of old rats. *Nature* 388: 272–275.

Burke SN and Barnes CA (2006) Neural plasticity in the ageing brain. *Nature Reviews Neuroscience* 7: 30–40.

Gallagher M and Rapp PR (1997) The use of animal models to study the effects of aging on cognition. *Annual Review of Psychology* 48: 339–370.

Hedden T and Gabrieli JD (2004) Insights into the ageing mind: A view from cognitive neuroscience. *Nature Reviews Neuroscience* 5: 87–96.

Morrison JH and Hof PR (1997) Life and death of neurons in the aging brain. *Science* 278: 412–419.

Rapp PR and Gallagher M (1996) Preserved neuron number in the hippocampus of aged rats with spatial learning deficits. *Proceedings of the National Academy of Sciences of the United States of America* 93: 9926–9930.

Sherwin BB (2006) Estrogen and cognitive aging in women. *Neuroscience* 138: 1021–1026.

Tanila H, Shapiro M, Gallagher M, and Eichenbaum H (1997) Brain aging: Changes in the nature of information coding by the hippocampus. *Journal of Neuroscience* 17: 5155–5166.

Aging and Memory in Humans

A M Brickman and Y Stern, Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Memory Systems

Memory is the explicit or implicit recall of information encoded in the recent or distant past. Current conceptualizations of memory, however, do not view the construct as a unitary system but rather divide it into hierarchical taxonomic modules based on duration of retention and the type of information that is being retrieved. Among the more fully elucidated conceptualizations of memory systems is one characterized by Larry Squire and colleagues, in which long-term memory is divided into declarative and nondeclarative subcomponents. Declarative, or explicit, memory refers to the ability to consciously recall facts (semantic memory), events (episodic memory), or perceptual information (perceptual memory). Nondeclarative memory requires the implicit recall of information and is usually divided into procedural, priming, or simple conditioning paradigms. Information that is retained on the order of seconds or minutes is usually referred to as short-term memory and is thought to represent a memory system distinct from long-term memory. Working memory, which comprises short-term memory, refers to the short-term store required to perform certain mental operations during retention. The following sections examine the impact of normal aging on different types of memory, as well as some of the potential moderators and mediators of cognitive aging. The information presented is organized hierarchically, following the memory systems just outlined.

Increased age puts an individual at risk for the development of neurodegenerative disorders, such as Alzheimer's disease (AD). Central to the dementia syndrome that characterizes AD is the gradual and progressive loss of long-term memory functions. Although the vast majority of older adults do not develop dementia, most experience some degree of cognitive change. Following the elucidation of memory systems and their component parts in the cognitive and cognitive neuroscience literature, there has been a recent interest in the impact of age on the different memory systems, independent of the devastating effects of dementia. Among well-screened individuals who do not meet diagnostic criteria for dementia, both cross-sectional and longitudinal studies demonstrate that the different memory component systems do not uniformly age; rather they show differential vulnerability to aging effects.

Long-Term Memory

Declarative Memory

Semantic memory As noted, semantic memory refers the recall of general or factual knowledge. Older adults commonly complain of subjective semantic memory problems when, for example, they report difficulty recalling the names of common objects or other well-learned information. Yet, despite these subjective complaints, semantic memory is among the more stable memory systems across the adult life span. The construct can be operationally defined by requiring subjects to define words or provide the answers to factual questions (e.g., the capital of a certain country), such as on the Vocabulary and Information subtests of the Wechsler Adult Intelligence Scales. Semantic memory is often included as part of the definition of 'crystallized intelligence,' which reflects an accumulation of information acquired over time and that is relatively impermeable to the effects of normal aging or mild brain disease.

It is well established that semantic memory shows very little decline in normal aging. In fact, semantic knowledge accumulation and memory increase into the sixth and seventh decades of life and may show only a gradual decline afterward. Much of our understanding of the impact of age on semantic memory has come from large-scale longitudinal and cross-sectional studies of normal aging. For example, longitudinal data from the Canberra Study , which followed a random sample of adults over the age of 70 years, demonstrated that crystallized abilities remained stable over approximately 8 years. This pattern of stability was apparent when age-associated differences (i.e., cross-sectional analysis) were considered as well. Similarly, Denise Park and her colleagues measured knowledge-based verbal ability, including three semantic memory tasks tapping word knowledge, in a large sample of healthy adults ranging in age from 20 to 92 years, and found a gradual increase in performance across the age groups. This finding is again consistent with the idea that semantic knowledge accumulates across the life span with little or no deleterious effects of normal aging.

Although there is little cross-sectional or longitudinal evidence to suggest that semantic memory changes significantly with normal aging, why are subjective complaints of semantic recall so common among older adults? One phenomenon, termed 'tip-of-the-tongue' (TOT), may explain this occurrence. TOT refers to

the common experience in which individuals have the feeling that they know the correct information (e.g., a person's name, a relatively low-frequency word), yet they are unable to recall it explicitly. The frequency of the TOT phenomenon increases with advancing age and may underlie the perceived difficulty with semantic recall.

Episodic memory The contrast between episodic memory and semantic memory was introduced by Endel Tulving in the early 1970s. Episodic memory refers to the explicit recollection of events, the 'what,' 'where,' and 'when' of information storage, and though it is conceptually distinct from semantic memory, the two memory systems interact. Episodic memory binds together items in semantic memory to form conceptually related time-based events. For example, the explicit recall of a learned story about a cowboy requires episodic memory for the story and semantic memory, or prior knowledge, of the items contained within the story.

Unlike semantic memory, episodic memory declines considerably with age. Older adults, for example, when prompted, have more difficulty recalling what they had for breakfast than do younger adults. The formal observation that episodic memory is affected by aging has existed for decades and has been documented numerous times.

Episodic memory is typically tested by requiring persons to learn information explicitly (e.g., a list or story) and recall it after a delay period. The three aspects of episodic memory include the encoding phase, the storage phase, and the retrieval of the encoded and stored information. These three phases show differential aging effects. Older adults' overall difficulty on tasks of episodic memory may be partially accounted for by a more shallow depth of encoding, as compared to younger adults – that is, older adults recall less information because of more limited processing of the initial study stimuli. This idea is supported by findings of greater age-related decline in acquisition or early retrieval of new information than in the degree of forgetting (i.e., the amount of information lost relative to the amount of information encoded). However, on free-recall tasks of list-learning paradigms, older adults recall fewer absolute words than do their younger adult counterparts; when given the correct stimulus in a recognition paradigm, older adults tend to incorrectly endorse more distractor stimuli, or foils.

Several observations about episodic memory and aging have emerged from the recent literature. First, the pattern of age-related differences in episodic memory appears to be similar across several modalities and domains, such as story recall, paired associate learning, face and word recall, and recognition paradigms of verbal and nonverbal information. Second, cross-sectional data suggest that age-associated episodic memory decline begins as early as age 20 years and decline linearly until about age 60, at which time there is a more precipitous decline. Third, whether the amount of interindividual variability increases systematically as a function of age is still somewhat unclear. Greater variability with aging would suggest that episodic memory decline might be a marker for insipient brain pathology, rather than a primary aspect of normal aging.

There are several competing theories postulating the potential mechanisms for age-associated declines in episodic memory. They can generally be divided into four areas, including an age-related failure of memory monitoring, or metamemory; age-associated decreases in the depth of initial encoding; age-related impairment in processing of contextual information; and age-associated decline in a number of processing resources. While these theories have not been fully substantiated empirically, the latter has received the greatest amount of support in the cognition literature. Proponents of a 'resource reduction hypothesis' argue that central to age-associated changes in episodic memory is a reduction in primary cognitive resources such as attention or working memory, or a reduction in the ability to engage due to an age-associated diminution in attentional inhibitory control. Others argue that age-associated memory decline is not due to a reduction in available attentional resources *per se*, but rather to age-related declines in perceptual processing speed.

Source memory Episodic memory comprises the information that is being recalled as well as the context in which the information was learned. This latter aspect is referred to as source memory, and there is increasing evidence that even when older adults successfully recall information, they may have difficulty recalling the source in which the information was acquired. The phenomenon has been demonstrated in the identification of the temporal context in which an item was learned, as well as the spatial and perceptual context. Studies by Daniel Schacter and colleagues required older adults to listen to different speakers read different blocks of declarative statements, and found that memory for the source of the declarative sentences was disproportionately worse than was memory for the statements. This general finding has been well replicated and has been extended to demonstrate age-associated impairment in both specific-source memory and partial-source memory. For example, older adults have difficulty, relative to younger adults, remembering which of

four people spoke a word (i.e., specific-source memory) as well as remembering partial information about the person who spoke the word, such as his or her gender (i.e., partial-source memory).

Nondeclarative Memory

Nondeclarative, or implicit, memory describes the memory system that allows learning outside of conscious awareness. It is generally divided into procedural memory and priming, each hypothesized to be mediated by a distinct neurobiological system. In general, nondeclarative systems are relatively spared across the adult life span, particularly compared to episodic memory, which shows the greatest aging effects.

Procedural memory Procedural or skill learning is one type of nondeclarative memory that refers to the nonconscious acquisition of motoric sequences. A common example of procedural memory is the process of learning how to drive an automobile. Initially, a novice driver needs to recollect consciously how to control each aspect of the automobile and the sequence in which to do so. With practice, the individual skill required to operate the vehicle enters into procedural memory, and control becomes relatively more automatic.

Little work has been conducted that explicitly examines the impact of chronological age on procedural memory, and results have been somewhat inconclusive; some studies show a decline with age and others show no apparent aging effect. A potential confounding factor in the examination of procedural memory among older adults is how much the experimental paradigm draws upon pure motoric speed and on other nonprocedural cognitive abilities, such as working memory, as well as whether the experimental paradigm addresses procedural learning versus long-term procedural memory. Studies that examine the effects of age on experts, such as pianists or typists, show a general trend of age-associated slowing, but a relative maintenance in measures of performance. Age-associated preservation of procedural memory may also be dependent on the amount of deliberate practice.

When considering age effects on procedural memory, it is important to dissociate performance time, learning, and memory for the task. For example, older adults tend to perform procedural memory tasks more slowly and learn procedural sequences at a slower rate, compared to younger adults. However, after initial acquisition, older adults will relearn a procedural memory task at rates similar to or more rapid than the rates at which younger adults relearn, even after a 2-year interval without practice. While procedural memory and aging remain somewhat understudied, there is some consensus that older adults have lasting preservation of procedural or motor memory. It is important, however, to distinguish between age-associated changes in motoric or perceptual processing speed and age-related changes in procedural memory; when accounting for aspects of the former, the latter appears to be relatively spared.

Priming Repetition priming is a special type of implicit memory that refers to the implicit impact that prior exposure to a stimulus has on later test performance. For example, an individual is more likely to complete the word stem STR__ with ONG (to form the word STRONG) than with EET (to form the word STREET) if he/she had previously studied the word STRONG. There is a long history of the examination of aging effects on priming in multiple modalities, with somewhat mixed results. In general, studies from the 1980s suggested that there are few differences between younger and older adults for priming across a number of tasks, such as word stem completion, picture naming, and word identification. There is evidence of a strong dissociation in the elderly between a preserved ability to perform implicit tasks and a deficit in performance on declarative tasks when compared to younger participants, although there are small, but reliable, age-associated decrements in priming for verbal abilities regardless of whether the dependent measure is accuracy or latency or whether the type of priming is item or associative.

Investigators such as John Rybash have distinguished among five types of priming, including (1) perceptual-item priming, (2) perceptual-associative priming, (3) conceptual-item priming, (4) conceptual-associative priming, and (5) perceptual-motor priming. Perceptual tasks are dependent upon implicit recall of properties of the stimuli presented during encoding, whereas conceptual tasks refer to priming for object meaning. There is some evidence that with normal aging there is relative preservation of performance on tasks that require perceptual priming and a relative decline in tasks that require conceptual priming. However, other investigators, such as Debra Fleischman and John Gabrieli, argue that studies that have shown no aging effect for priming may have had insufficient measurement reliability or power to detect age differences, and positive studies may have included a minority of individuals with insipient dementia. In contrast to Rybash, they argue that perceptual priming is vulnerable to aging effects, whereas conceptual priming remains relatively intact. Differences between these two conclusions may reflect differing positions on inclusion criteria for categories of priming.

For example, Fleischman and Gabrieli exclude all tasks that might include the processes that explicit memory retrieval tasks also engage. Further, they consider tasks such as category-exemplar generation to be conceptual, whereas Rybash considers these to be perceptual. In summary, although the literature is somewhat mixed, the effects of normal aging on repetition priming are of small magnitude or nonexistent, depending on modality.

Short-Term Memory

In contrast to long-term memory, in which information is stored for minutes to years, short-term memory refers to holding information in conscious awareness for the duration of seconds. While short-term memory is qualitatively a type of memory distinct from long-term memory, the two systems interact. For example, in order for information to be stored in long-term memory, the information must first exist within short-term memory. The exact mechanisms by which short-term memory is transferred to long-term storage have yet to be fully elucidated, but the 'modal model' of information transfer, proposed by Atkinson and Shiffrin in the 1960s, provided a theoretical framework that has dominated the cognitive literature for decades. Briefly, information flows from the environment to primary sensory (perceptual) stores and then into a short-term store, which can include rehearsal, coding, or decision. The capacity of the short-term store is limited and information can leave the system (i.e., forgetting), lead to a response output, or enter long-term storage, which is fairly permanent. When information is drawn from long-term memory, it exists in short-term memory while in conscious awareness. In terms of the capacity of short-term memory, a heuristic offered by Miller in the 1950s is that the average short-term memory span for a person is seven, plus or minus two items, or chunks. Although this heuristic has remained popular, more recent studies suggest that short-term memory span is closer to four items, or is contingent upon an individual's processing speed.

The term 'short-term memory' refers to the passive short-term store of information, but speaks little to the process by which information is retained in the store. It is difficult to discuss the effects of age on short-term memory without consideration of these processes. The term 'working memory' is typically applied to the process by which information is held in short-term memory and refers to the cognitive manipulation of information that is contained within short-term memory. An influential model that attempts to explain how information is stored within short-term memory was proposed by Baddeley and Hitch, in which a supervisory 'central executive' controls the information flow among three 'slave systems,' including the phonological or articulatory loop, the visuospatial sketchpad, and the episodic buffer. The first two systems rehearse phonological or visual information, respectively, and the third system integrates information from the other two systems and interfaces with long-term memory. The central executive acts to coordinate the flow of information among the slave systems, to direct attention to appropriate internal and external stimuli, and to suppress irrelevant stimuli.

Working memory paradigms require individuals to perform mental operations on items held in conscious awareness, such as reordering words or numbers. While there is little evidence that short-term memory *per se* significantly declines with normal aging, there is ample evidence that working memory abilities do. Within working memory, efficacy of inhibition and smaller span are particularly vulnerable to the effects of aging.

Working memory in aging studies is often operationally defined by tasks such as letter rotation, reading span, computation span, and line span. Performance on these tasks tends to show a linear age-associated decline that is similar to that seen for tasks of long-term episodic memory and speed of processing. Some research suggests that age-associated decline in working memory abilities mediates age-associated decline in other memory and cognitive domains. Indeed, statistical control of performance on tasks of working memory often attenuates the observed age-associated decline on other cognitive tasks.

Summary and Course

Normal age-associated memory decline is not uniform. Older adults evidence worse performance on long-term memory tasks compared to younger adults, but these differences are relatively greatest on tasks of declarative episodic memory. Semantic memory, on the other hand, remains relatively stable across the adult life span or may even increase as more semantic knowledge is accumulated with age. Similarly, working memory, or the manipulation of information that is held in conscious awareness, shows marked decline in with normal aging, and some theorists propose that working memory deficits mediate age-associated decline in other cognitive domains.

In terms of the course of memory changes across the adult life span, results from cross-sectional and longitudinal studies suggest that subtle memory changes can begin as early as the early or middle

twenties and continue to decline linearly with age. Some authors distinguish between 'lifelong decline' and 'late-life decline.' Performance on tasks of episodic and working memory seems to begin to decline in the twenties and continues to decline linearly across the life span, which is supported by cross-sectional aging and cognition studies. Some longitudinal studies, however, suggest a curvilinear course of memory decline, with a more precipitous decline after about age 60 years, preceded by relatively little decline with age. Short-term memory store appears to remain relatively stable until about age 70, at which point it begins to drop, and, as noted, semantic abilities appear to remain relatively stable, at least until late life.

Moderators and Mediators of Cognitive Aging

A consistent observation in the aging and cognition literature is a greater amount of variability in memory performance with increasing age. This finding is evident both when age is considered as a cross-sectional variable and when individual age trajectories are examined over time; some adults experience great amounts of cognitive decline while others experience relatively little. Increased variability suggests that there are moderators and mediators of cognitive aging. Two factors in particular, including cerebrovascular or cardiovascular risk and cognitive reserve, have received considerable attention as modulators that may account for some of the increased variability, and these are potential targets for intervention or prevention of cognitive decline.

Cerebrovascular or cardiovascular risk factors may mediate the relationship between chronological age and the neurobiological changes that underlie cognitive aging. Increased blood pressure is associated with reduced psychomotor speed, visuoconstruction ability, learning, memory, and executive functioning. Similarly, both insulin resistance and diabetes are associated with diminished cognitive abilities in later life. Hyperlipidemia has also been identified as a potential cerebrovascular risk factor that negatively impacts cognition in older adults. While it is still somewhat unclear by what mechanism cerebrovascular or cardiovascular risk factors impact cognition, it is likely that they interact with age to have a cumulative effect on cognition in later life. Furthermore, intervention studies that target the vascular system, such as cardiovascular fitness regimens, are associated with increased cognitive functioning among older adults.

Cognitive reserve is another factor that may account for the increased variability observed in cognitive aging. The cognitive reserve hypothesis stems from the observation that there is a disconnection between degree of brain pathology and its cognitive manifestation and suggests that the brain actively attempts to cope with brain pathology by using existing processing approaches or by recruiting compensatory networks. Although cognitive reserve is often applied to clinical entities, such as AD or stroke, it may be operative among the normal aged. That is, cognitive reserve may moderate the relationship between normal age-associated neurobiological changes and their cognitive outcome. Indeed, proxies for reserve, such as measures of education or literacy and IQ, are related to the degree of age-associated memory decline. Declarative memory scores of older adults with lower levels of literacy decline at a greater rate than do those with higher levels of literacy. Like vascular risk factors, interventions that target cognitive reserve may improve the course of cognitive aging.

Other factors, such as genes and nutritional exposure, could potentially impact the course of memory decline with normal aging. It is also important to note that the possibility of inclusion of individuals with insipient dementia may have contaminated some studies that attempted to elucidate the pattern of age-associated memory decline, and future studies of normal aging should focus on disentangling normal cognitive decline from pathological aging at the earliest point possible. With the powerful capabilities of neuroimaging techniques, we can begin to understand the complex relationships among normal aging, neurobiology, and cognition while defining potentially modifiable moderators and mediators of cognitive aging.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Aging and Memory in Animals; Basal Forebrain and Memory; Cognition in Aging and Age-Related Disease; Episodic Memory; Functional Neuroimaging Studies of Aging; Humans; Lipids and Membranes in Brain Aging; Metal Accumulation During Aging; Short Term and Working Memory.

Further Reading

Atkinson RC and Shiffrin RM (1968) Human memory: A proposed system and its control processes. In: Spence KW and Spence JT (eds.) *The Psychology of Learning and Motivation: Advances in Research and Theory,* vol. 2, pp. 89–195. London: Academic Press.

Baddeley AD (2000) The episodic buffer: A new component of working memory? *Trends in Cognitive Science* 4: 417–423.

Christensen H (2001) What cognitive changes can be expected with normal ageing? *Australian and New Zealand Journal of Psychiatry* 35: 768–775.

Cowan N (2001) The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24: 87–114; discussion 114–185.

Craik FIM and Salthouse TA (eds.) (2000) *The Handbook of Aging and Cognition,* 2nd edn. Mahwah, NJ: Lawrence Erlbaum.

Fleischman DA and Gabrieli JD (1998) Repetition priming in normal aging and Alzheimer's disease: A review of findings and theories. *Psychology and Aging* 13: 88–119.

Grady CL and Craik FI (2000) Changes in memory processing with age. *Current Opinion in Neurobiology* 10: 224–231.

Hedden T and Gabrieli JD (2004) Insights into the ageing mind: A view from cognitive neuroscience. *Nature Reviews Neuroscience* 5: 87–96.

Krampe RT and Ericsson KA (1996) Maintaining excellence: Deliberate practice and elite performance in young and older pianists. *Journal of Experimental Psychology: General* 125(4): 331–359.

LaVoie D and Light LL (1994) Adult age differences in repetition priming: A meta-analysis. *Psychology & Aging* 9: 539–553.

Light LL (1991) Memory and aging: Four hypotheses in search of data. *Annual Review of Psychology* 42: 333–376.

Park DC, Smith AD, Lautenschlager G, et al. (1996) Mediators of long-term memory performance across the life span. *Psychology & Aging* 11: 621–637.

Rybash JM (1996) Implicit memory and aging: A cognitive neuropsychological perspective. *Developmental Neuropsychology* 12: 127–179.

Salthouse TA (1993) Speed mediation of adult age differences in cognition. *Developmental Psychology* 29: 722–738.

Schacter DL, Kaszniak AW, Kihlstrom JF, et al. (1991) The relation between source memory and aging. *Psychology & Aging* 6: 559–568.

Small SA, Stern Y, Tang M, et al. (1999) Selective decline in memory function among healthy elderly. *Neurology* 52: 1392–1396.

Squire LR (2004) Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory* 82: 171–177.

Stern Y (2002) What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society* 8: 448–460.

Stern Y (ed.) (2006) *Cognitive Reserve: Theory and Applications.* New York: Taylor & Francis.

Tulving E (2002) Episodic memory: From mind to brain. *Annual Review of Psychology* 53: 1–25.

Aging of the Brain

A B Scheibel, Brain Research Institute UCLA, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

Possible Causal Factors

Experience indicates that all living things have finite lives, whether they be the ephemeral summer of the butterfly, a thousand days for the rat, the reported hundred-year-long lives of large tortoises, or the multi-thousand-year histories of giant redwoods and bristle-cone pines. The concept of genetic programming of the aging process seemed to receive support from Hayflick's studies of the late 1950s and early 1960s, in which young, actively dividing connective tissue cells raised *in vitro* appeared limited to a certain maximum number (50) of mitotic divisions, after which degenerative changes invariably set in. In retrospect, these classic experiments now seem less convincing, because of methodologic errors involved, and, although the concept of genetic control remains intuitively appealing, the case is considered far from proven. Nonetheless, the capacity for base excision DNA repair has been shown to be reduced in the brains of aging primates. Furthermore, the activity of polymerase beta in brain cells appears compromised with age. These findings may be bellwethers for future more widely ranging age-related changes in genomic capabilities.

A presently favored alternative theory operating either independently or, more likely, in tandem with the proposed genetic constraints envisages the possibility of progressive damage to the cell from external or internal factors, resulting in a cumulative pattern of dysfunctions. For example, toxic factors within the environment (i.e., chemical carcinogens and background radiation of terrestrial or galactic origin) may slowly affect the cell DNA content and protein-synthesizing machinery, leading to errors in synthesis with resultant progressive changes in cell structure and function. This summation of many small developing errors in the synthetic and enzymatic machinery of the cell (such as the progressive degradation of base excision DNA repair mentioned above) is conceived as mounting to a point beyond which conditions for cell life become impossible (error-catastrophe theory).

Obviously, this putative process might complement and potentiate a set of existing genetic instructions also directed toward the cell's eventual demise. However, normal oxidative metabolic activity of the cell may itself result in cumulative damage. Of particular interest is the development of a family of free radicals of oxygen (e.g., singlet oxygen, hydrogen peroxide, superoxide and hydroxyl radicals) that, through crosslinkage or cleavage reactions, may permanently alter the structure of the cell. In support of this are observations that mitochondrial free radical generation is lower in long-lived species than in short-lived ones. The varied but often overlapping phenomena of repeated stress/glucocorticoid-induced cell damage, especially in the hippocampus, excitotoxic activity of neurotransmitters like glutamate, and programmed cell death (apoptosis) may all contribute to the process. These ideas are still in early stages of development and exemplify the questions that surround the processes of aging in general, and of the nervous system in particular.

Structural Changes

A number of changes, both gross and microscopic, have been reported in the brains of the aged, although variation is the rule rather than the exception. The brain itself is often somewhat reduced in size and weight, especially beyond the eighth decade of life. Decrements of 10–15% are frequently quoted, although interindividual differences are large. Some aged brains show mild to moderate gyral atrophy and sulcal widening, but this is not the rule. The surrounding meninges are frequently more opaque and milky in appearance than those of the young brain, and may be adherent to underlying cortex. Isolated deposits of calcium are found in and around the pacchionian granulations near the vertex of the hemisphere. The ventricular system may show alterations in silhouette, and computed axial tomography and magnetic resonance imaging (MRI) scans often show enlarged ventricular shadows. A large and growing series of studies using T1- and T2-weighted MRIs reveal considerable variation in the aged brain. Although some apparently normal individuals in their seventh and eighth decades show signs of cortical atrophy, ventricular enlargement, and even textural changes in the cerebral white matter near the ventricles (leukoairosis), others show little or no discernible alteration. Recent MRI studies indicate that a slowly developing loss of gray matter, starting as early as the

third decade, is accompanied by increases in white matter volume into the mid- or late 40s, followed then by degeneration. Similar studies at the regional level suggest decrement in size of the hippocampus bilaterally in men but not in women. Localized periventricular white matter changes, in particular, are increasingly thought to result from underlying processes such as hypertension or diabetes, and to have little or no relevance to normal brain aging.

Microscopically, a group of well-known stigmata is seen in routine Nissl or reduced silver-stained preparations. Their incidence varies rather widely among individuals, although there is a general tendency toward increased frequency with age. Neuronal cell bodies gradually accumulate masses of refractile granules with high lipid content, known as lipofuscins. The significance of these so-called aging pigments is not clear, and they have been found as early as the tenth year of life in cells of the inferior olive. The present consensus is that they represent the remains of lysosomal and mitochondrial membranes that have accumulated, due perhaps to gradual failure of mechanisms for turnover and reutilization. As such, they do not necessarily constitute a direct threat to the neuron except to the extent that they preempt increasing amounts of cytoplasmic space previously used for synthesis of glycoproteins, lipoproteins, and neurotransmitters. One interesting and opposing view maintains that the appreciable content of myoglobin and respiratory enzymes allows lipofuscin granules to serve a positive role in providing energy to neurons under conditions of low oxygen tension. Coarser vacuolization of the cytoplasm, usually concentrated in the area of development of the apical dendrite shaft in cortical pyramidal cells, is known as 'granulovacuolar degeneration of Simchowitz' and is of equally unknown etiology.

The neurofibrillary tangle is a structural alteration of neuronal cytoplasm that may involve soma, dendrites, and axon. Initially recognized by Alzheimer and Simchowitz as one of two defining microcriteria of individuals dying with dementia, it is also found in very small numbers in the normal aged. Electron microscope study reveals that the tangle consists of paired helical filaments, about 22 nm in width, which braid loosely around each other with a characteristic series of partial twists, each 80 nm from the next. Tangles are now believed to result from abnormal phosphorylation of tau protein, which is necessary to normal polymerization–depolymerization mechanisms of microtubule-associated proteins. Microtubules provide the infrastructure that facilitates virtually all movement of neurotransmitters, structural proteins, enzymes and endocellular organelles within the neuron and its extensive processes (axonal and dendritic transport). It is therefore easy to imagine how the buildup of neurofibrillary tangles can progressively starve and throttle the neuron. The appearance of tangles (tau aggregates) in the frontal lobe appears to correlate positively with cognitive loss. Experimental neurofibrillary tangle development is being extensively studied in primates, in small laboratory animals, and even in *Drosophila*.

Accompanying these characteristically intracellular alterations is the development of small foci of destruction in the surrounding neuropil, the senile plaque of Alzheimer. These are classically described as containing central cores of Congo red-positive, amyloid-like material, surrounded by radial auras of degenerating dendritic and axonal tissue and a halo of microglia. They constitute the second component of the two major histopathologic criteria of dementia and, like neurofibrillary tangles, are present in more restricted number in the brains of most aged individuals. The density of senile plaques has been said to correlate positively with the degree of cognitive impairment in dementia, although these data have recently been subject to question.

Intensive biochemical and immunocytologic analyses of these structures are presently under way. It is now believed that beta amyloid (a 40- to 42-amino-acid peptide) may represent an unusual by-product of a larger, normally present molecule, called 'amyloid precursor protein', and that production and deposition may result from aberrant enzymatic (secretase) activity. Some data suggest that accumulations of beta amyloid may drive molecular cascades that potentiate neurodegenerative changes, thereby representing a direct threat to the integrity of surrounding neuronal structures. Several experimental interventions are being explored in an effort to prevent its development or mitigate its effects.

All the histopathologic changes so far described appear most intensively in the limbic system, especially the entorhinal cortex and hippocampus. They are also found widely throughout the rest of central nervous system, however, including cerebral neocortex, diencephalon, brain stem, and spinal cord. Since interindividual variation is the rule in the aged brain, these descriptions represent a distillation and summary rather than the expected picture of the brain of any one aged individual.

The problem of neuronal loss has been hotly debated and is still not settled. The few quantitative studies dating from the late 1950s and 1960s suggested neuron loss of up to 30% in the normal aging cerebral neocortex. More detailed recent investigations, based on larger numbers of cases, indicate that such levels of neuronal loss, in normal aging at

least, may have been excessive, based as they were on the study of a few areas in a limited number of brains. More broadly sampled data suggest a pattern of modest loss of neurons, primarily large cortical cells, with compensatory dendritic growth in adjacent neurons. The cholinergic cell masses of the basal forebrain and the noradrenergic cells of the locus coeruleus undergo undoubted change during the aging process, but even here, declining function may be more an expression of waning metabolic vigor and decreased axodendritic dimensions than of massive neuronal loss. This is in sharp contrast to dementing syndromes such as Alzheimer's disease, in which more than half of the complement of locus coeruleus cells may disappear.

Carefully controlled studies indicate that in the rat the only significant cerebral cell losses occur during the first 100 days of life in what might be considered a period of adaptation and fine-tuning to the environment. After this, there is little further discernable neuronal loss, even at 900 days of age, when rats are beginning to die of natural causes. Although the weight of evidence from both animal and human studies now increasingly points to negligible neuronal loss during the normal, uncomplicated aging process, a large number of cells appear to undergo changes in the dendritic (and axonal) extensions of the soma. Many neurons show progressive restriction and atrophy of their more peripheral dendrite branches and, especially in cortical pyramidal cells, among the basilar shafts. Accompanying this is irregular loss of dendrite spines and frequent beaded swellings along the remaining dendritic branches. These changes can be related in general terms to progressive loss of protein-synthesizing capabilities due, perhaps in part, to increasing incursions upon cytoplasmic space by lipofuscin deposits and neurofibrillary tangles. However, it has also been shown that the potential for neuronal growth is not lost during aging. Accompanying the progressive destruction of some dendritic systems, it appears that other neurons grow further dendritic extensions, thereby increasing their available synaptic areas (Figure 1). The concept of two types of neuronal response to aging, one involving dendritic retraction and one reactive dendritic expansion, brings with it a number of exciting implications for providing more effective and fulfilling lives for the elderly.

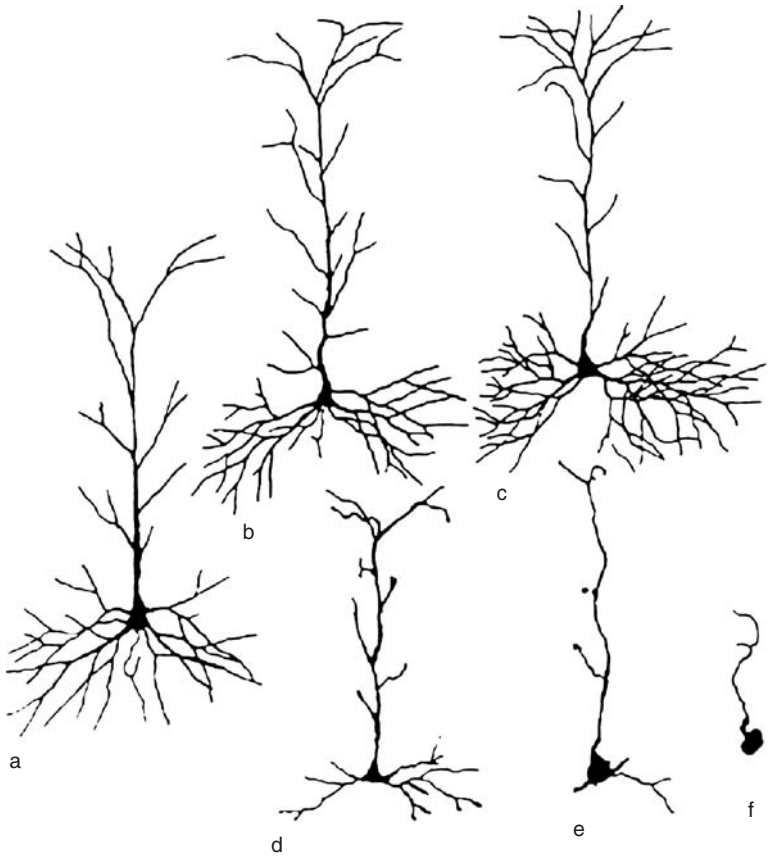

Figure 1 Two possible patterns of age-related alterations in cortical pyramidal cells. The normal mature neuron (a) may show regressive dendritic changes characterized by loss of basilar dendritic branches and eventual loss of the entire dendritic tree (d, e, f). Other neurons (b, c) may show progressive increase in dendritic branching. Drawing based on Golgi impregnations.

During the first half of the twentieth century, brain aging and dementia were usually associated with visible changes in the major arteries of the central nervous system. Cerebral arteriosclerosis was considered a necessary accompaniment to and, in fact, virtually synoymous with aging and senility. Careful studies during the 1950s showed the inaccuracy of such concepts, and today, large-vessel disease is seldom considered to contribute significantly to the picture of general brain aging. Recent investigations with positron emission tomography scanning methods and xenon clearance techniques have again called attention to the relative adequacy of blood flow and oxygen and glucose metabolism in the healthy aging brain. The scanning electron microscope and immunohistochemical studies have, coincidentally, focused interest on the microcirculation (capillary bed) of the brain and on the plexus of neural fibers that normally innervate their walls. Subtle changes in this delicate but all-pervasive system may prove to have significant impact on the maintained vigor or decline of brain structure and function.

Biochemical and Metabolic Changes

It is becoming clear that the aging process entails a broad range of biochemical and metabolic changes. Among these, alterations in neurotransmitter content and activity figure prominently. As many as 90% of neurons are involved, directly or indirectly, in cholinergic mechanisms, synthesizing, transporting, and releasing – or else being synaptically dependent upon – acetylcholine. Cholinergic systems are highly energy dependent, and age-related decreases in several important glycolytic enzymes have been reported. A link might thus conceivably be postulated between altered glycolytic energy mechanisms and cognitive function.

The most important concentration of acetylcholine-rich neurons is found in the basal forebrain area, in and around the nucleus basalis of Meynert and the ventral pallidum. Most studies indicate mild to moderate age-related cell loss in these areas. Associated with (although not entirely dependent upon) this loss are decreases in acetylcholine content in various portions of the brain, lower levels of the synthesizing enzyme, choline acetyltransferase, and decreased numbers of acetylcholine receptor binding sites. The relevance of these alterations to cognitive changes in the elderly, such as impairment of recent memory function, is not entirely clear. Of uncertain import, also, is the apparent loss of fluidity of the cytoplasmic membrane of the cell as the lipoprotein structure changes, a possible result of progressive diminution of choline content. Age-related increases in membrane microviscosity may bring with them a significant train of sequelae including, initially, a higher capacity for receptor binding followed by enhanced rates of receptor loss with eventual overall reduction in receptor binding affinity. Mechanisms responsible for these changes remain uncertain, but it has been postulated that increased microviscosity of the cell membrane leads first to increased exposure of those receptors out of the plane of the membrane, after which the uncovered receptors are progressively sloughed off into the surrounding medium. This suggests that membrane-bound proteins such as receptors and transporters are maintained in dynamic equilibrium by the extent of fluidity of the surrounding membrane lipids. As one example, immunohistochemical studies of a glucose transporter (GLUT 3) revealed significant loss (46%) in the hippocampal-dentate gyrus of the aged rat. Attempts to forestall or modify age-related losses in membrane fluidity through administration of active lipid fractions have been under investigation.

Catecholaminergic systems also show moderate to marked alteration in the brain. Levels of both dopamine and norepinephrine synthesis decrease, as do the numbers of adrenergic and dopaminergic receptors. Marked attenuation of norepinephrine synthesis in the hypothalamus may, in turn, be responsible, at least in part, for decrease in the synthesis of certain hypothalamic hormones.

Significant cell loss has been reported in the locus coeruleus, the single most important source of brain parenchymal norepinephrine. The substantia nigra and adjacent ventral tegmental area (VTA), major sources of dopamine, also undergo histologic alterations that include microvascular changes, iron deposition, and neuronal loss. The serotonin-rich systems of the raphe seem, on the other hand, to be somewhat less vulnerable. Resultant maintenance of minimally disturbed titers of serotonin in a setting of falling catecholaminergic levels may conceivably be related to a high incidence of disturbed sleep patterns and depression in the aged.

A broad range of age-related neuroendocrine changes is known to exist. Although such alterations are documented by innumerable anecdotal observations, regional study of the mechanisms involved is in its early phases. For instance, the importance of the ovary as a pacing factor in the aging female rat is receiving documentation in many laboratories. At an equally obvious level, the high likelihood of occurrence of benign prostatic hypertrophy and prostatic cancer in the aged male is being related in meaningful fashion to age-related alterations in testosterone metabolism that may lead to unbalanced growth

stimulation. In a more global sense, the entire aging process can apparently be slowed by early restriction of nutritional input, which seems to delay pubescence and maturation. Finally, increasing imbalance among the various neurotransmitters, and in particular among those that are aminergic, appears to exert progressive impact upon the hypothalamus, pituitary, and pineal gland, leading to cascades of dysfunction that manifest themselves at every level of organization.

Although the picture presented appears to emphasize regressive elements, a much more positive picture of the normal aging brain can now be painted. Accumulated experience can help make up for age-related loss in the speed of cerebral processing or of memory recall.

Experimental evidence conclusively demonstrates the continued plasticity of even the very old brain. Continual environmental enrichment and challenge appear to enhance cognitive power in the aged. In fact, it has been shown that a subset of those genes whose expression is adversely affected by aging is oppositely affected by exposure to enriched environments. Under conditions of optimally maintained health, increasing numbers of individuals in their 70s and 80s and beyond continue to hold responsible positions, or otherwise distinguish themselves in commerce, the arts, and certain aspects of the sciences. In every case, the touchstone appears to be activity, involvement, and purpose.

See also: Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Apoptosis in Nervous System Injury; Cognition in Aging and Age-Related Disease; Dementia; Gene Expression in Normal Aging Brain; Metal Accumulation During Aging.

Further Reading

Barja G and Herrero A (2000) Oxidative damage to mitochondrial DNA is inversely related to maximum life span in the heart and brain of mammals. *FASEB Journal* 14(2): 312–318.

Buell S and Coleman P (1979) Dendritic growth in the aged human brain and failure of growth in senile dementia. *Science* 206: 854–856.

Coffey CE, Wilkinson WE, Parashos IA, et al. (1992) Quantitative cerebral anatomy of the aging human brain: A cross-sectional study using magnetic resonance imaging. *Neurology* 42(3 pt. 1): 527–536.

Flood D and Coleman C (1988) Neuron numbers and sizes in aging brains: Comparisons of human, monkey and rodent data. *Neurobiology of Aging* 9: 453–463.

Lee J, Duan W, Long JM, Ingram DK, and Mattson MP (2000) Dietary restriction increases the number of newly generated neural cells and induces BDNF expression in the dentate gyrus of rats. *Journal of Molecular Neuroscience* 15(2): 99–108.

Rowe J and Kahn R (1987) Human aging: Usual and successful. *Science* 237: 143–149.

Salmon E, Maquet P, Sadzot B, Degueldre C, Lemaire C, and Franck G (1991) Decrease of frontal metabolism demonstrated by positron emission tomography in a population of healthy elderly volunteers. *Acta Neurologica Belgica* 5: 288–295.

Scheibel A (1996) Structural and functional changes in the aging brain. In: Birren J and Schaie K (eds.) *Handbook of the Psychology of Aging,* 4th edn., pp. 105–128. San Diego: Academic Press.

Ylikowski R, Ylikowski A, Erkinjuntti T, Sulkava R, Raininko R, and Tilvis R (1993) White matter changes in healthy elderly persons correlate with attention and speed of mental processing. *Archives of Neurology* 50: 818–824.

Aging of the Brain and Alzheimer's Disease

D L Price, A V Savonenko, M Albert, J C Troncoso, and P C Wong, The Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Over the past decade, major advances have been made in understanding the causes and mechanisms of age-associated alterations in memory and cognition, including mild cognitive impairment (MCI) and Alzheimer's disease (AD), which is the most common cause of dementia in the elderly. For many reasons (prevalence, lack of mechanism-based treatments, cost of care, and impact on individuals and caregivers), AD is one of the most challenging illnesses in medicine. However, extraordinary progress has been made in deliniating the pathology, biochemistry, and neurobiology of the disease; the clinical and pathological bases of MCI and AD; the utility of new diagnostic approaches and outcome measures; and the value of transgenic models of the genetic forms of familial AD and gene-targeted models which discloses mechanism-based targets for therapy of AD. These discoveries are leading to new treatments for this type of dementia.

Cognitive and Memory Impairments in the Elderly

Although many older individuals remain intellectually intact, a significant number of the elderly show declines in memory and cognitive abilities. These alterations are usually mild, occur relatively late in life, and involve speed of learning, complex problem solving, ability to retain large amounts of new information, and visuospatial skills. Vocabulary, information storing, and comprehension skills often remain relatively stable into old age. Some older individuals have mild cognitive impairments (MCIs) in which memory loss exceeds that expected for age, but they do not meet criteria for AD. However, among individuals with MCIs, there may be considerable heterogeneity (i.e., interindividual variability in the rate and severity of progression). The variability of alterations in memory with age makes it sometimes difficult to determine, after a single examination, whether an individual with mild age-associated memory impairments will remain relatively stable or will progress to severe dementia. The most accurate clinical approach to this problem is to assess cognitive abilities and memory performance repeatedly in the same patient over a period of time. Several neuropsychological tests are valuable in predicting AD: delayed recall tasks (recently learned information), as assessed by the California Verbal Test and Wechsler Memory Scale, and measures of executive functions (ability to organize and plan), as assessed by the Trail-Making Test and Self-Ordering Test. On the basis of clinical assessments correlated with postmortem examinations of the brain, MCI is now regarded as a transitional state between the normal functional state of older persons and early AD. Imaging studies have predictive value regarding the progression to AD.

Alzheimer's Disease: Clinical Features, Diagnostic Studies, Neuropathology and Biochemistry, and Current Treatments

AD is the most common cause of senile dementia, a term that refers to a syndrome occurring in older individuals that results in memory loss and cognitive impairments of sufficient severity to interfere with social, occupational, and personal functions. This type of dementia affects more than 4 million people in the United States. Because of increased life expectancy and the postwar baby boom, this population is the fastest growing segment of society. During the next 25 years, the number of people with AD in the United States will triple, as will the cost of care and treatment.

A majority of individuals with sporadic AD exhibit the first clinical signs during their seventh decade. However, some people develop disease in midlife; in these cases, a family history of the illness is more likely. In both the sporadic and familial forms of AD, affected individuals show difficulties with memory, problem solving, executive functions for language, calculation, visuospatial perceptions, judgment, and behavior; mental functions and activities of daily living are increasingly impaired. Some patients develop psychotic symptoms, such as hallucinations and delusions. In the late stages, individuals are mute, incontinent, bedridden, and usually die of intercurrent medical illnesses.

Other causes of dementia syndromes include cerebrovascular disease, alone or in combination with AD; Lewy body dementia; Parkinson's disease; alcoholism; drug intoxications; infections, such as with human immunodeficiency virus (HIV) (i.e., acquired immunodeficiency syndrome (AIDS)) and syphilis; brain tumors; vitamin deficiencies (e.g., B_{12}); thyroid disease; and a variety of other metabolic disorders. Because some of these disorders are treatable, it is important for physicians to exclude these

entities before making a diagnosis of possible AD. At present, except for brain biopsy, it is not possible to definitively establish the diagnosis of early AD in living humans, and it is extremely important for the physician to exclude other causes of dementia syndromes, because some of these illnesses respond to specific treatments. Imaging studies are of great value in exclusions and are of increasing utility in supporting the diagnosis of AD. At present, the only available treatment for AD is for symptoms.

As mentioned previously, to establish a diagnosis of AD, clinicians rely on histories from patients and informants; physical, neurological, and psychiatric examinations; neuropsychologic testing; laboratory studies, including examinations of cerebrospinal fluid (CSF); and a variety of laboratory tests, including neuroimaging studies. The clinical profile, in concert with a variety of imaging and laboratory assessments (levels of CSF β-amyloid (Aβ) decrease and tau proteins increase), allows the clinician to make a diagnosis of possible or probable AD. Several imaging strategies and studies of biomarkers, particularly of CSF, are useful laboratory approaches. In early AD, magnetic resonance imaging (MRI) often discloses atrophy of specific regions of the brain, while positron emission tomography (PET) with [^{18}F]deoxyglucose, or single-photon emission computerized tomography (SPECT), demonstrates decreased glucose utilization and early reductions in regional blood flow in the parietal and temporal lobes. PET, following administration of brain penetrant ^{11}C-labeled Pittsburgh compound B (PIB), which binds to Aβ with high affinity, discloses patterns which are thought to reflect the Aβ burden in the brain. In concert, these various approaches as applied to patients should increase the accuracy of diagnosis in earlier stages of disease and allow assessments of the efficacies of new antiamyloid therapeutics. The presence of the apolipoprotein E (apoE) ε4 allele confers risk in late-onset disease (see later); apoE genotyping is a useful research tool, but it is not helpful for routine diagnostic purposes in individual patients.

AD is the result of abnormalities associated with dysfunction and death of specific populations of neurons, particularly those cells in neural systems participating in memory and cognitive functions. Abnormalities in the brain selectively involve neurons in the neocortex, entorhinal area, hippocampus, amygdala, nucleus basalis, anterior thalamus, and several brain stem monoaminergic nuclei (in particular the locus coeruleus and raphe complex). Dysfunction of neurons in these brain regions and circuits is reflected by the presence of cytoskeletal abnormalities (neurofibrillary tangles (NFTs)) in these cells, the presence of neuritic plaques in brain regions receiving inputs from these nerve cells, and reductions in transmitter markers of these neurons in their target fields. The cellular pathology is characterized by the presence of intracellular and extracellular protein or peptide aggregates: phosphorylated tau is assembled into the paired helical filaments (PHFs) within NFTs and in swollen neurites (around plaques); and Aβ peptides exist in extracellular β-pleated sheet conformations assembled into oligomers, which are the principal constituent of amyloid plaques. Ultimately, affected neurons die; often 'tombstone' tangles, amyloid deposits, and glial reactions remain to indicate sites of neurodegeneration.

The clinical manifestations of amnestic mild cognitive impairment (aMCI) and early Alzheimer's disease (eAD) result from abnormalities occurring among populations of neurons in neural systems/brain regions essential for memory, learning, and cognitive performance. Damaged circuits include those in the basal forebrain cholinergic system, the amygdala, the hippocampus, the entorhinal and limbic cortex, and the neocortex. In a recent study, the character, abundance and distribution of the lesions (i.e., diffuse plaques, neuritic plaques, and tangles) were correlated with clinical signs in cognitively characterized controls, individuals with aMCI, or cases of eAD. No differences were observed in the number of diffuse plaques among the groups. In persons with aMCI, tangles were significantly increased in regions of the ventral medial temporal lobe as compared to controls; individuals with eAD showed greater numbers of NFTs and neuritic plaques in both frontal and temporal regions. Individuals with aMCI exhibited increased numbers of neuritic plaques in neocortical regions as compared to controls, but not as compared to cases of eAD. Memory deficits correlated most closely with the abundance of NFTs in CA1 of the hippocampus and in the entorhinal cortex, a finding which leads to the interpretation that tangles were more important than amyloid deposition in the progression from normal function to aMCI to eAD, and that tangles in the medial temporal lobe play a key role in memory declines in aMCI. Data from multiple studies are interpreted to indicate that aMCI reflects a transitional state in the evolution of AD. Because the regional distributions of NFTs correlate most closely with the degree of clinical impairments from aged healthy controls to individuals with aMCI to cases of AD, the spread of NFTs beyond the medial temporal lobe is thought to be linked to the development of dementia.

As mentioned previously, clinical signs reflect cellular abnormalities within specific neural circuits responsible for memory and cognition. Affected neurons exhibit conformationally altered isoforms of tau

comprising the PHFs in NFTs, neurites, and neuropil threads. Axonal pathologies include varicosities and terminal clubs, the latter representing disconnected synapses and transected axons (also observed in aged, memory-impaired rhesus monkeys with Aβ deposits, and in some transgenic monkey models of Aβ amyloidosis). The Aβ-containing neuritic plaques are sites of synaptic disconnection in regions receiving inputs from diseased populations of neurons; associated with these lesions are decrements in generic and transmitter-specific synaptic markers in the target fields of degenerating nerve cells and local astroglial and microglial responses (particularly associated with plaques). Ultimately, there is evidence of death of neurons, possibly by apoptosis. Thus, the clinical manifestations of aMCI and AD reflect a progressive disruption of synaptic communications involving subsets of neural circuits, first associated with degeneration of axon terminals and, later, death of neurons.

In one hypothetical model mechanistically linking Aβ and phosphorylated tau, Aβ42 species liberated at terminals oligomerize to form Aβ assemblies or Aβ-derived diffusible ligands (ADDLs), which are linked to synaptic damage. Subsequently, not yet identified retrograde signals, which originate at disconnected terminals, trigger the activation of kinases (or the inhibition of phosphatases) in cell bodies. The phosphorylation of tau at a variety of sites is associated with conformational changes in this protein and the formation of PHFs and, eventually, NFTs. Secondary disturbances in axonal transport can, in turn, compromise the functions and viability of neurons. Eventually, affected nerve cells die and extracellular tangles remain as markers of the nerve cells destroyed by disease.

These cellular abnormalities have profound clinical consequences. Abnormalities that damage the entorhinal cortex, hippocampus, and other circuits in the medial temporal cortex are presumed to be critical for memory impairments. Higher cognitive deficits, such as disturbances in executive functions for language, calculation, problem solving, and judgment, are believed to be related to pathology in the neocortex. Alterations in the basal forebrain cholinergic system may contribute to memory difficulties and attention deficits. The behavioral and emotional disturbances presumably may reflect involvement of the limbic cortex, amygdala, thalamus, and monoaminergic systems.

At present no cure exists for this devastating illness. Available treatments for AD target symptoms, and many present-day therapies focus on treating associated conditions such as depression, agitation, sleep disorders, hallucinations, and delusions. One therapeutic approach has been to try to support the functions of the basal forebrain cholinergic system, which is one of the circuits severely damaged in AD. Several strategies have been developed to influence this neurotransmitter system. Unfortunately, precursor loading (i.e., choline, lecithin) and muscarinic agonist approaches to improve cholinergic functions did not prove effective. Several acetylcholinesterase inhibitors, which prolong the half-life of the transmitter in the synaptic cleft, have been approved for the treatment of AD in the United States. These drugs are sometimes associated with side effects and they have, at best, a very modest effect on cognitive functions and the ability to perform activities of daily living. Clinical trials have tested the efficacies of anti-inflammatory compounds, estrogens, plant extracts, and antioxidants without great benefit. In an antiexcitotoxic approach, memantine, an *N*-methyl-D-aspartate (NMDA) receptor antagonist, has been suggested to reduce the rate of clinical deterioration of patients with moderate to severe AD.

Genetic Causes and Risk Factors for AD

Established genetic risk factors implicated in AD include: mutations in the *amyloid beta (A4) precursor protein* gene (*APP*; chromosome 21), mutations in *presenilin 1* (*PSEN1*; chromosome 14) and *presenilin 2* (*PSEN2*; chromosome 1), and the presence of a risk susceptibility allele (ε4) of *APOE* (chromosome 19), which predisposes to later onset AD and some cases of late-onset familial AD. Autosomal dominant mutations in *APP*, *PSEN1*, or *PSEN2* usually cause disease earlier than occurs in sporadic cases, with the majority of mutations in *APP, PSEN1*, and *PSEN2* influencing β-site APP cleaving enzyme 1 (BACE1) or γ-secretase cleavages of APP to increase the levels of all Aβ species or the relative amounts of toxic Aβ42.

A member of the *APP* gene family (including also *APLP1* and *APLP2*), *APP* encodes a type I transmembrane protein of unknown function which is abundant in the nervous system, rich in neurons, and transported rapidly anterograde in axons to terminals. In the central nervous system (CNS), BACE1 cleaves at the +1 and +11 sites; subsequently, the γ-secretase complex cleaves at a variety of sites (see later), which generate the N- and C-termini of Aβ peptides, respectively. A variety of *APP* mutations alter the processing of APP and influence increase in the production of Aβ peptides or the amounts of the more toxic Aβ42. The APP Swedish double mutation (*swe*) enhances manyfold the BACE1 cleavage at the N-terminus of Aβ (+1 site), resulting in substantial elevations in levels of all Aβ peptides. APP_{717} mutations influence γ-secretase cleavage to increase secretion of Aβ42, which is the most toxic peptide. In contrast, other

mutations may promote local fibril formation and vascular amyloidosis. Moreover, individuals with increased *APP* gene dosage are prone to AD; familial duplications of *APP* or trisomy 21 (Down syndrome) are associated with extra copies of *APP* and develop AD pathology relatively early in life.

PSEN1 and *PSEN2* encode two highly homologous and conserved 43–50 kDa multipass transmembrane proteins that are involved in Notch1 signaling pathways critical for cell fate decisions. Presenilins are endoproteolytically cleaved by a 'presenilinase' to form an N-terminal ~28 kDa fragment and a C-terminal ~18 kDa fragment; both fragments are critical components of the γ-secretase complex. Nearly 50% of early-onset cases of familial AD are linked to >90 different mutations in *PSEN1*. A relatively small number of *PSEN2* mutations also cause autosomal dominant familial AD. The majority of abnormalities in *PSEN* genes are missense mutations that enhance γ-secretase activities and increase the levels of the Aβ42 peptides.

Biochemistry of Amyloidosis: APP and the Secretases

APP is cleaved by β- and γ-secretases, releasing the ectodomain of APP (APPs), liberating a cytosolic fragment termed APP intracellular domain (AICD), and generating several species of Aβ peptides. In the central (but not peripheral) nervous system, Aβ peptides are generated by sequential endoproteolytic cleavages by BACE1 (at the Aβ +1 and +11 sites) to generate APP-β C-terminal fragments (APP-βCTFs) and by the γ-secretase complex (at several sites, varying from Aβ 36 to 38, 40, 42, and 43) to form Aβ species peptides. The intramembranous cleavages of APP-βCTF by γ-secretase releases an AICD, which can form a complex with Fe65, a nuclear adaptor protein; Fe65 and Aβ or Fe65 alone (in a novel conformation) can gain access to the nucleus to influence gene transcription, a signaling mechanism analogous to that occurring in the Notch1 pathway (next to the Notch intracellular domain (NICD)). It has been speculated that the AICD signaling pathway may play a role in learning and memory. In other cells in other organs, APP is cleaved endoproteolytically within the Aβ sequence through alternative, nonamyloidogenic pathways: α-secretase (TNF-α converting enzyme, or TACE) cleaves between residues 16 and 17; BACE2 cleaves between residues 19 and 20 and 20 and 21. These cleavages, which occur in nonneural tissues, preclude the formation of Aβ peptides and serve to protect these cells/organs from Aβ amyloidosis.

BACE1, encoded by a gene on chromosome 11, is a transmembrane aspartyl protease that preferentially cleaves APP at the +1 and +11 sites of Aβ in APP and is critical for the generation of Aβ. BACE1 is present in the brain but is virtually undetectable in nonneural tissues; BACE1-specific immunoreactivities are readily localized in the hippocampus. Importantly, as compared to wild-type APP, BACE1 cleaves APP (swe) cleaved approximately 100-fold more efficiently at the +1 site, resulting in a greater increase in BACE1 cleavage products (elevating cell Aβ species) in the presence of this mutation. Significantly, *BACE1*-deficient mice show no overt developmental phenotype, and in cultures of *BACE1*$^{-/-}$ neurons, the secretion of A$\beta 1_{40/42}$ and A$\beta 11_{40/42}$ is abolished. Thus BACE1 is the principal neuronal β-secretase is responsible for the penultimate Aβ cleavage. In contrast, BACE2, which makes antiamyloidogenic cleavages at sites +19/+20 of Aβ, acts more like α-secretase; it does not appear to play a significant role in APP processing in neurons.

γ-Secretase, essential for the regulated intramembranous proteolysis of a variety of transmembrane proteins, is a multiprotein catalytic complex that includes PSEN1 and PSEN2 (described earlier); nicastrin (Nct), a type I transmembrane glycoprotein; and Aph-1 (anterior pharynx defective) and Pen-2 (presenilin enhancer), two multipass transmembrane proteins. Presenilin contains aspartyl residues that play roles in these intramembranous cleavages; substitutions at D257 in transmembrane (TM) segment 6 and at D385 in TM segment 7 reduce Aβ secretion and cleavage of Notch1 *in vitro*. The functions of the various γ-secretase proteins and their interactions in the complex are not yet fully defined. In one model, Aph-1 and Nct form a precomplex which interacts with presenilin; subsequently, Pen-2 enters the complex, where it is critical for the 'presenilinase' cleavage of presenilin into two fragments. In concert, this complex is responsible for γ-secretase cleavages of APP, Notch, and a variety of other transmembrane proteins. With regard to APP, this complex makes the final cleavage that generates a peptide (especially Aβ) and liberates AICD. Many familial AD-linked *PSEN* mutations promote the cleavages that lead to formation of Aβ42.

Animal Models of Aging and AD

Both spontaneously occurring and genetically engineered diseases in animals have been used to model features of AD.

For example, an age-related spontaneous disorder occurs in elderly rhesus monkeys (*Macaca mulatta*). With an estimated life span of more than 35 years (×2.8, for equivalent human age), these rhesus monkeys show cognitive and memory deficits that

appear at the end of the second and at the beginning of the third decades of life. These animals develop virtually all of the brain abnormalities (amyloid deposits, neuritic plaques, scattered tangles, and modest reductions in transmitter markers) that are observed in older humans and, to a greater degree, in cases of AD. Genetically engineered models relevant to AD are discussed in the following section.

Genetic Models of Aβ Amyloidosis

In mice, expression of *APPswe* or APP_{717} (with or without mutant *PSEN1)* leads to an Aβ amyloidosis in the CNS. Mutant *APP/PSEN1* mice develop accelerated disease secondary to increased levels of Aβ (particularly Aβ42) associated with the presence of diffuse Aβ deposits and neuritic plaques in the hippocampus and cortex. Levels of Aβ peptides, particularly Aβ42, increase in brain with age, and oligomeric species (ADDLs, Aβ*56, etc.) appear in the CNS.

Over time, mice carrying mutant transgenes exhibit Aβ deposits, swollen neurites in proximity to these deposits, and neuritic plaques associated with glial responses. Some lines of mice show evidence of amyloid in vessels (congophilic angiopathy). In forebrain regions, the density of synaptic terminals and levels of several neurotransmitter markers are reduced. In some settings, there are deficiencies in synaptic transmission. Moreover, some lines of mice show evidence of degeneration of subsets of neurons.

Behavioral studies of lines of transgenic mice, including those generated by Dr. David Borchelt, disclose deficits in spatial reference memory (Morris water maze task) and episodic-like memory (repeated reversal and radial water maze tasks). At 6 months of age, *APPswe/PSEN1ΔE9* mice develop plaques, but all genotypes are indistinguishable from nontransgenic animals in all cognitive measures. However, in 18-month-old cohorts, *APPswe/PSEN1ΔE9* mice perform all cognitive tasks less well than do mice of all other genotypes. In these animals, amyloid burdens are high; decreases are detectable in levels of several neurotransmitter markers. The strongest relationships exist between deficits in episodic-like memory tasks and total Aβ loads in the brain. Collectively, these studies suggest that, in *APPswe/PSEN1ΔE9* mice, some form of Aβ (ultimately associated with amyloid deposition) can disrupt circuits critical for memory, particularly episodic-like memory. Some of these impairments have been linked to the presence of Aβ oligomers (see later) and can be reversed by antibody-mediated reductions of levels of brain Aβ. These mice do not recapitulate the complete phenotype of AD, but they are very useful subjects for research designed to examine behavioral consequences of Aβ amyloidosis in the CNS, to delineate disease mechanisms, and to test novel therapies.

Over the past decade, a variety of Aβ species, including monomers, oligomers, Aβ-derived diffusible ligands, structural assemblies, and amyloid deposits in neuritic plaques, have been suggested to play important roles in impairing synaptic communication. For example, in one study, when naturally secreted Aβ peptides were injected into the ventricular system of rats, long-term potentiation (LTP) was inhibited in the hippocampus; the activity of the peptide was completely blocked by the injection of a monoclonal Aβ antibody, a finding consistent with the concept that oligomers are the toxic moiety and that they are both necessary and sufficient to perturb learned behavior. Active immunization was less effective in rescuing function and correlated most closely with the levels of antibodies recognizing oligomers. More recently, studies of Tg2576 mice suggested that extracellular accumulations of a 56 kDa soluble amyloid assembly, termed Aβ*56, purified from the brains of memory-impaired mice, can interfere with memory when administered to young rats.

Models of Tau Abnormalities

The paucity of tau abnormalities in various lines of mutant *APP* mice may be related to differences in tau isoforms expressed in this species, as compared to humans. Early efforts to express mutant *tau* transgenes in mice did not lead to striking clinical phenotypes or pathology. More recently, mice overexpressing *tau* show clinical signs, attributed to degeneration of motor axons. When prion or Thy1 promoters are used to drive tau_{P301L} (a mutation linked to autosomal dominant frontotemporal dementia with parkinsonism), some brain and spinal cord neurons develop tangles. Mice expressing $APPswe/tau_{P301L}$ exhibit enhanced tangle-like pathology in limbic system and olfactory cortex. Moreover, injection of Aβ42 fibrils into specific brain regions of tau_{P301L} mice increases the number of tangles in those neurons projecting to sites of Aβ injection. A triple transgenic mouse (3×Tg-AD), created by microinjecting *APPswe* and tau_{P301L} into single cells derived from monozygous $PSEN1_{M146V}$ knockin mice, develop age-related plaques and tangles as well as deficits in LTP which appear to antedate overt pathology. However, mice bearing both mutant *tau* and *APP* (or *APP/PSEN1*) or mutant *tau* mice injected with Aβ may not be ideal models of familial AD because the presence of the *tau* mutation alone is associated with the development of tangles and disease.

Targeting of Genes in the Amyloidogenic Pathway

To begin to understand the functions of some of the proteins thought to play roles in AD, investigators have targeted a variety of genes encoding BACE1, PSEN1, Nct, and Aph-1.

BACE1$^{-/-}$ Mice

These animals mate successfully and exhibit no overt pathology. *BACE1*$^{-/-}$ neurons do not cleave at the +1 and +11 sites of Aβ, and the production of Aβ peptides is abolished, establishing that BACE1 is the neuronal β-secretase required to generate the N-termini of Aβ. However, *BACE1*$^{-/-}$ mice show altered performance on some tests of cognition and emotion (see later); the former deficits can be rescued by overexpression of *APP* transgenes.

PSEN1$^{-/-}$ Mice

These embryos develop severe abnormalities of the axial skeleton, ribs, and spinal ganglia, a lethal outcome which resembles a partial *Notch1*$^{-/-}$ phenotype. *PSEN1*$^{-/-}$ cells show decreased levels of secretion of Aβ related to the fact that PSEN1 (along with Nct, Aph-1, and Pen-2) is a component of the γ-secretase complex that carries out the final (S3) intramembranous cleavage of Notch1. Without γ-secretase cleavage, the NICD is not released from the plasma membrane and cannot reach the nucleus to provide a signal to initiate transcriptional processes essential for cell fate decisions. Significantly, conditional *PSEN1/2* targeted mice show impairments in memory and synaptic plasticity in the hippocampus, raising the question (posed particularly effectively Dr. Jie Shen and colleagues) as to the roles of loss of presenilin function in neurodegeneration and AD.

Nct$^{-/-}$ Mice

These embryos die early and exhibit several patterning defects, including abnormal segmentation of somites; this phenotype closely resembles that seen in *Notch1*$^{-/-}$ and *PSEN1/2*$^{-/-}$ embryos. Importantly, *Nct*$^{-/-}$ cells do not secrete Aβ peptides, whereas *NctT*$^{+/-}$ cells show reduction of ~50%. The failure of *NctT*$^{-/-}$ cells to generate Aβ peptides is accompanied by accumulation of APP C-terminal fragments. Importantly, *Nct*$^{+/-}$ mice develop tumors of the skin, presumably related to reduced levels of signaling by Notch1, which appears to act as a tumor suppressor in the skin.

Aph-1a$^{-/-}$ Mice

Three murine *Aph-1* alleles (*Aph-1a*, *Aph-1b*, and *Aph-1c*) encode four distinct Aph-1 isoforms: Aph-1aL and Aph-1aS (derived from differential splicing of *Aph-1a*), Aph-1b, and Aph-1c. *Aph-1a*$^{-/-}$ embryos show patterning defects that resemble, but are not identical to, those of *Notch1*, *Nct*, or *PSEN* null embryos. Moreover, in *Aph-1a*$^{-/-}$-derived cells, the levels of Nct, presenilin fragments, and Pen-2 are decreased. There is an associated reduction in levels of the high-molecular-weight γ-secretase complex and a decrease in secretion of Aβ. In *Aph-1a*$^{-/-}$ cells, other mammalian Aph-1 isoforms can restore the levels of Nct, presenilin, and Pen-2.

Experimental Therapeutics

The availability of models of amyloidogenesis provides opportunities to ablate or knock down genes, to modulate cleavages, and to influence clearance; such experiments have set the stage for testing the influences of Aβ production, APP cleavage patterns, peptide neurotoxicity, and promotion of clearance and/or degradation of Aβ. Because it is not possible to discuss all experimental treatment in mouse models of Aβ amyloidosis, we focus on selected studies that illustrate the experimental strategies directed at specific therapeutic targets that we predict will provide mechanism-based therapeutic benefits to patients with AD.

Reductions in BACE1 Activity

Deletion of *BACE1* in *APPswe/PSEN1ΔE9* mice prevents both Aβ deposition and age-associated cognitive abnormalities that occur in this transgenic model. The *BACE1*$^{-/-}$ *APPswe/PSEN1ΔE9* mice do not develop the Aβ deposits or the age-associated abnormalities in working memory that occur in the *APPswe/PSEN1ΔE9* model of Aβ amyloidosis. Similarly, *BACE1*$^{-/-}$ Tg2576 mice show rescue from age-dependent memory deficits and physiological abnormalities. Moreover, Aβ deposits are sensitive to *BACE1* dosage and can be efficiently cleared from regions of the CNS when *BACE1* is silenced at these sites. Inhibitors of β-secretase, conjugated to carrier peptides, can inhibit *in vitro* and *in vivo* (following intraperitoneal injection into Tg2576 mice). Conditional expression systems or RNA interference (RNAi) silencing allow investigators to examine the pathogenesis of diseases and to assess the degrees of reversibility of the disease processes. Clearly, BACE1 is an attractive therapeutic target. However,

BACE1$^{-/-}$ mice manifest alterations in both hippocampal synaptic plasticity and in performance on tests of cognition and emotion; the memory deficits, but not emotional alterations, in *BACE1*$^{-/-}$ mice are prevented by co-expression of *APPswe/PSEN1ΔE9* transgenes, suggesting that APP processing influences cognition/memory and that the other potential substrates of BACE1 may play roles in neural circuits related to emotion. BACE1 and APP processing pathways appear to be critical for cognitive, emotional, and synaptic functions. Moreover, BACE1 cleaves reuregulin (NRC), an axonal signal for myelination; inhibition of β-secretase activity is an exciting therapeutic opportunity, but future studies should be alert to potential mechanism-based side effects that may occur with strong inhibition of the enzyme.

Inhibition of γ-Secretase Activity

Both genetic and pharmaceutical lowering of γ-secretase activity decreases production of Aβ in cell-free and cell-based systems and reduces levels of Aβ mutant mice with Aβ amyloidosis. Thus, γ-secretase activity is a significant target for therapy. Treated mice show reduced levels of Aβ and amyloid plaques. However, γ-secretase activity is also essential for processing of Notch, which is critical for lineage specification and cell growth during embryonic development. Significantly, inhibitors of γ-secretase activity reduce production of Aβ, but can also have profound effects on T and B cell development and on the appearance of intestinal mucosa (proliferation of goblet cells, increased mucin in gut lumen, and crypt necrosis). Significantly, *Nct*$^{+/-}$*APPswe/PSEN1ΔE9* mice develop skin tumors, presumably, in part, because reduced γ-secretase acts, via Notch signaling, as a tumor suppressor in skin. Clinicians carrying out trials of this class of inhibitor will have to be alert to several potential adverse events associated with reduced activity of this enzyme complex.

γ-Secretase Modulation by Nonsteroidal Anti-inflammatory Compounds

Retrospective epidemiological studies suggested that significant exposure to nonsteroidal anti-inflammatory drugs (NSAIDs) reduces risk of AD, an outcome initially interpreted as related to suppression of the well-documented inflammatory process occurring in brains of cases of AD. However, more recent studies indicate that a subset of NSAID compounds in this class can modulate secretase cleavages (to shorter, less toxic Aβ species) without altering Notch or other APP processing outcomes. Short-term treatment of mutant mice appears to have some benefit in terms of lowering Aβ and plaque pathology. This strategy is now being evaluated in drug trials in humans.

Aβ Immunotherapy

Multiple lines of evidence, including studies of lesions of entorhinal cortex or perforant pathway (by lesioning cell bodies or axons/terminals transporting APP to terminals, respectively), indicate that removing the source of Aβ significantly reduces Aβ in target fields. Similarly, increasing local levels of degrading enzymes, including insulin-degrading enzymes (IDEs) and neprolysin (NEP), can cleave Aβ and can reduce levels of the Aβ peptide.

To date, the most exciting findings regarding clearance of Aβ are derived from studies using active and passive Aβ immunotherapy. In treatment trials in mutant mice and in rodents, both Aβ immunization (with Freund's adjuvant) and passive transfer of Aβ antibodies reduce levels of Aβ and the level of plaque burden. Although, the mechanisms of enhanced clearance are not certain, at least two not mutually exclusive hypotheses have been suggested: (1) a small amount of Aβ antibody reaches the brain, binds to Aβ peptides, promotes the disassembly of fibrils, and, via the Fc antibody domain, encourages activated microglia to enter the affected regions and remove Aβ; and (2) serum antibodies serve as a 'sink' to draw the amyloid peptides from the brain into the circulation, thus changing the equilibrium of Aβ in different compartments and promoting removal of Aβ from the brain. Whatever the mechanism, Aβ immunotherapy in mutant mice is successful in partially clearing Aβ, in attenuating learning and behavioral deficits in several cohorts of mutant *APP* mice, and in partially reducing tau abnormalities in the triple transgenic mice. However, several studies have documented that immunotherapy may be associated with brain hemorrhages related to congophilic angiopathy. In these settings, the presence of amyloid presumably can weaken vascular walls; potentially, removal of some intramural vascular amyloid could lead to rupture of damaged vessels and bleeding.

Although mutant mice that received immunotherapy were not reported to develop evidence of meningoencephalitis, a subset of patients in a clinical trial did manifest these problems. Individuals receiving vaccinations with preaggregated Aβ and an adjuvant (followed by a booster) develop antibodies that recognize Aβ in the brain and vessels. Unfortunately, although phase I trials with Aβ peptide and adjuvant vaccination were not associated with any adverse events, phase II trials detected complications (meningoencephalitis) in a subset of patients and were suspended. The pathology in the index case, consistent with T cell meningitis, was interpreted to show some clearance of Aβ deposits, but some regions contained a relatively high density of tangles, neuropil

threads, and vascular amyloid. Aβ immunoreactivity was sometimes associated with microglia, and T cells were conspicuous in the subarachnoid space and around some vessels. In another case, there was significant reduction in amyloid deposits in the absence of clinical evidence of encephalitis. The trial was stopped. Assessment of cognitive functions in a small subset of patients (30) who received vaccination and booster immunizations disclosed that patients who generated Aβ antibodies (as measured by a new Aβ assay) had a slower decline in several functional measures. The events occurring in this subset of patients illustrate the challenges of extrapolating outcomes in mutant mice to human trials. Investigators continue to pursue the passive immunization approaches and are attempting to make new antigen/adjuvant formulations that do not stimulate T-cell-mediated immunologic attack.

Conclusions

Over many years, investigators have more accurately defined mild cognitive impairment and early AD, have developed diagnostic approaches, and have clarified the character and stages of pathology and related the findings to clinical signs. They have greatly enhanced our understanding of the mechanisms underlying the biochemistry of Aβ plaques and tau-related pathology. Following leads from human autopsy studies and from investigations of *in vitro* and *in vivo* models, investigators are now on the threshold of implementing novel treatments based on an understanding of the neurobiology, neuropathology, biochemistry, and genetics of AD. Moreover, a variety of tools, including amyloid imaging and measure of Aβ flux between various compartments, are now available to assess efficacy of treatment. It is anticipated that exciting discoveries over the next few years will lead to the design of new mechanism-based therapies that can be tested in models, and that these approaches will be introduced into the clinic for the benefit of patients with this devastating illness.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Aging of the Brain; Alzheimer's Disease: Neurodegeneration; Alzheimer's Disease: An Overview; Alzheimer's Disease: Transgenic Mouse Models; Animal Models of Alzheimer's Disease; Axonal Transport and Alzheimer's Disease; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Dementia and Language; Dementia; Sleep Architecture.

Further Reading

Bard F, Cannon C, Barbour R, et al. (2000) Peripherally administered antibodies against amyloid β-peptide enter the central nervous system and reduce pathology in a mouse model of Alzheimer disease. *Nature Medicine* 6(8): 916–919.

Barten DM, Guss VL, Corsa JA, et al. (2005) Dynamics of β-amyloid reductions in brain, cerebrospinal fluid, and plasma of β-amyloid precursor protein transgenic mice treated with a γ-secretase inhibitor. *Journal of Pharmacology and Experimental Therapeutics* 312(2): 635–643.

Borchelt DR, Ratovitski T, VanLare J, et al. (1997) Accelerated amyloid deposition in the brains of transgenic mice coexpressing mutant presenilin 1 and amyloid precursor proteins. *Neuron* 19(4): 939–945.

Braak H and Braak E (1994) Pathology of Alzheimer's disease. In: Calne DB (ed.) *Neurodegenerative Diseases*, pp. 585–613. Philadelphia: W.B. Saunders.

Braak H and Braak E (1991) Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica* 82: 239–259.

Buxbaum JD, Thinakaran G, Koliatsos V, et al. (1998) Alzheimer amyloid protein precursor in the rat hippocampus: Transport and processing through the perforant path. *Journal of Neuroscience* 18(23): 9629–9637.

Cai H, Wang Y, McCarthy D, et al. (2001) BACE1 is the major β-secretase for generation of Aβ peptides by neurons. *Nature Neuroscience* 4(3): 233–234.

Cao X and Sudhof TC (2001) A transcriptionally [correction of transcriptively] active complex of APP with Fe65 and histone acetyltransferase Tip60. *Science* 293(5527): 115–120.

Cleary JP, Walsh DM, Hofmeister JJ, et al. (2005) Natural oligomers of the amyloid-β protein specifically disrupt cognitive function. *Nature Neuroscience* 8(1): 79–84.

Cummings JL (2004) Alzheimer's disease. *New England Journal of Medicine* 351(1): 56–67.

Davies P and Maloney AJF (1976) Selective loss of central cholinergic neurons in Alzheimer's disease. *Lancet* 2: 1403.

De Strooper B, Saftig P, and Craessaerts K (1998) Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature* 391(6665): 387–390.

DeMattos RB, Bales KR, Cummins DJ, et al. (2001) Peripheral anti-Aβ antibody alters CNS and plasma Aβ clearance and decreases brain Aβ burden in a mouse model of Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* 98: 8850–8855.

DeMattos RB, Bales KR, Cummins DJ, et al. (2002) Brain to plasma amyloid-β efflux: A measure of brain amyloid burden in a mouse model of Alzheimer's disease. *Science* 295: 2264–2267.

Farzan M, Schnitzler CE, Vasilieva N, et al. (2000) BACE2, a β-secretase homolog, cleaves at the β site and within the amyloid-β region of the amyloid-β precursor protein. *Proceedings of the National Academy of Sciences of the United States of America* 97: 9712–9717.

Ghiso J and Wisniewski T (2004) An animal model of vascular amyloidosis. *Nature Neuroscience* 7(9): 902–904.

Goedert M and Spillantini M (2006) Neurodegenerative alpha-synucleinopathies and tauopathies. In: Siegel GJ, Albers RW, Brady S, and Price DL (eds.) *Basic Neurochemistry: Molecular, Cellular, and Medical Aspects,* 7th edn., pp. 745–759. Boston, MA: Elsevier Academic Press.

Hock C, Konietzko U, Streffer JR, et al. (2003) Antibodies against β-amyloid slow cognitive decline in Alzheimer's disease. *Neuron* 38(4): 547–554.

Jankowsky JL, Fadale DJ, Anderson J, et al. (2004) Mutant presenilins specifically elevate the levels of the 42 residue β-amyloid

peptide *in vivo*: Evidence for augmentation of a 42-specific γ-secretase. *Human Molecular Genetics* 13(2): 159–170.

Jicha GA, Parisi JE, Dickson DW, et al. (2006) Neuropathologic outcome of mild cognitive impairment following progression to clinical dementia. *Archives of Neurology* 63(5): 674–681.

Killiany RJ, Hyman BT, Gomez-Isla T, et al. (2002) MRI measures of entorhinal cortex vs hippocampus in preclinical AD. *Neurology* 58(8): 1188–1196.

Klunk WE, Engler H, Nordberg A, et al. (2004) Imaging brain amyloid in Alzheimer's disease using the novel positron emission tomography tracer, Pittsburgh compound-B. *Annals of Neurology* 55: 1–14.

Klyubin I, Walsh DM, Lemere CA, et al. (2005) Amyloid β protein immunotherapy neutralizes Aβ oligomers that disrupt synaptic plasticity *in vivo*. *Nature Medicine* 11(5): 556–561.

Laird FM, Cai HS, Savonenko AV, et al. (2005) BACE1, a major determinant of selective vulnerability of the brain to Aβ amyloidogenesis is essential for cognitive, emotional and synaptic functions. *Journal of Neuroscience* 25(50): 11693–11709.

Lazarov O, Morfini GA, Lee EB, et al. (2005) Axonal transport, amyloid precursor protein, kinesin-1, and the processing apparatus: Revisited. *Journal of Neuroscience* 25(9): 2386–2395.

Lesne S, Koh MT, Kotilinek L, et al. (2006) A specific amyloid-β protein assembly in the brain impairs memory. *Nature* 440 (7082): 352–357.

Lewis J, Dickson DW, Lin W, et al. (2001) Enhanced neurofibrillary degeneration in transgenic mice expressing mutant tau and APP. *Science* 293: 1487–1491.

Li T, Ma G, Cai H, et al. (2003) Nicastrin is required for assembly of Presenilin/γ-secretase complexes to mediate notch signaling and for processing and trafficking of β-amyloid precursor protein in mammals. *Journal of Neuroscience* 23(8): 3272–3277.

Ma G, Li T, Price DL, et al. (2005) APH-1a is the principal mammalian APH-1 isoform present in γ-secretase complexes during embryonic development. *Journal of Neuroscience* 25(1): 192–198.

Markesbery WR, Schmitt FA, Kryscio RJ, et al. (2006) Neuropathologic substrate of mild cognitive impairment. *Archives of Neurology* 63(1): 38–46.

Martin LJ, Pardo CA, Cork LC, et al. (1994) Synaptic pathology and glial responses to neuronal injury precede the formation of senile plaques and amyloid deposits in the aging cerebral cortex. *American Journal of Pathology* 145(6): 1358–1381.

Morris JC, Storandt M, Miller JP, et al. (2001) Mild cognitive impairment represents early-stage Alzheimer disease. *Archives of Neurology* 58(3): 397–405.

Morris R and Mucke L (2006) Alzheimer's disease: A needle from the haystack. *Nature* 440(7082): 284–285.

Nestor PJ, Scheltens P, and Hodges JR (2004) Advances in the early detection of Alzheimer's disease. *Nature Medicine* 10(supplement): S34–S41.

Nicoll JA, Wilkinson D, Holmes C, et al. (2003) Neuropathology of human Alzheimer disease after immunization with amyloid-β peptide: A case report. *Nature Medicine* 9(4): 448–452.

Oddo S, Caccamo A, Shepherd JD, et al. (2003) Triple-transgenic model of Alzheimer's disease with plaques and tangles: Intracellular Aβ and synaptic dysfunction. *Neuron* 39(3): 409–421.

Petersen RC (2003) Mild cognitive impairment clinical trials. *Nature Reviews Drug Discovery* 2(8): 646–653.

Price DL, Tanzi RE, Borchelt DR, et al. (1998) Alzheimer's disease: Genetic studies and transgenic models. *Annual Review of Genetics* 32: 461–493.

Price JL and Morris JC (1999) Tangles and plaques in nondemented aging and "preclinical" Alzheimer's disease. *Annals of Neurology* 45(3): 358–368.

Rogaev EI, Sherrington R, Rogaeva EA, et al. (1995) Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* 376: 775–778.

Rovelet-Lecrux A, Hannequin D, Raux G, et al. (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genetics* 38(1): 24–26.

Savonenko AV, Laird FM, Troncoso JC, et al. (2005) Role of Alzheimer's disease models in designing and testing experimental therapeutics. *Drug Discovery Today* 2(4): 305–312.

Savonenko A, Xu GM, Melnikova T, et al. (2005) Episodic-like memory deficits in the APPswe/PS1dE9 mouse model of Alzheimer's disease: Relationships to β-amyloid deposition and neurotransmitter abnormalities. *Neurobiology of Disease* 18(3): 602–617.

Schenk D, Hagen M, and Seubert P (2004) Current progress in β-amyloid immunotherapy. *Current Opinion in Immunology* 16(5): 599–606.

Selkoe D and Kopan R (2003) Notch and presenilin: Regulated intramembrane proteolysis links development and degeneration. *Annual Review of Neuroscience* 26: 565–597.

Sheng JG, Price DL, and Koliatsos VE (2002) Disruption of corticocortical connections ameliorates amyloid burden in terminal fields in a transgenic model of Aβ amyloidosis. *Journal of Neuroscience* 22(22): 9794–9799.

Sisodia SS, Koo EH, Beyreuther KT, et al. (1990) Evidence that β-amyloid protein in Alzheimer's disease is not derived by normal processing. *Science* 248: 492–495.

Vassar R, Bennett BD, Babu-Khan S, et al. (1999) β-Secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science* 286: 735–741.

Weggen S, Eriksen JL, Das P, et al. (2001) A subset of NSAIDs lower amyloidogenic Aβ42 independently of cyclooxygenase activity. *Nature* 414: 212–216.

Wolfe MS, Xia W, Ostaszewski BL, et al. (1999) Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and γ-secretase activity. *Nature* 398(6727): 513–517.

Wong PC, Li T, and Price DL (2005) Neurobiology of Alzheimer's disease. In: Siegel GJ, Albers RW, Brady S, and Price DL (eds.) *Basic Neurochemistry: Molecular, Cellular, and Medical Aspects,* 7th edn., pp. 781–790. Boston, MA: Elsevier Academic Press.

Wong GT, Manfra D, Poulet FM, et al. (2004) Chronic treatment with the γ-secretase inhibitor LY-411,575 inhibits β-amyloid peptide production and alters lymphopoiesis and intestinal cell differentiation. *Journal of Biological Chemistry* 279(13): 12876–12882.

Wong PC, Price DL, and Cai H (2001) The brain's susceptibility to amyloid plaques. *Science* 293: 1434–1435.

Aging: Brain Potential Measures and Reaction Time Studies

E Golob, Tulane University, New Orleans, LA, USA
H Pratt, Technion – Israel Institute of Technology, Technion City, Israel
A Starr, University of California at Irvine, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Brain and Cognitive Aging

Normal aging affects brain functions in a variety of ways. Some regions, such as prefrontal cortex, medial temporal lobe, and neuromodulatory systems (cholinergic, noradrenergic, and serotonergic), show marked age-related changes, whereas other regions, such as primary sensory and motor regions, show comparatively little change during normal aging. Age-related changes in memory function are also selective, which is related to regional differences in the impact of brain aging. In general, age-related declines are most evident in certain types of explicit memory tasks that require conscious retrieval of remembered information, but they are small or absent for many implicit memory tasks. For example, episodic memory is one type of explicit memory, and it is defined as memory for specific events and their context, such as remembering what one did last weekend. In contrast, implicit memory can be expressed even if a person does not consciously remember a specific learning episode or cannot articulate what was learned. An example of implicit memory is the gradual development of proficiency in various kinds of motor skills, such as riding a bicycle. Because the adverse effects of aging are most evident for explicit memory, this article focuses on event-related potential (ERP) and electroencephalogram (EEG) findings while subjects perform explicit memory tasks. For most studies of human aging, the comparison group of young subjects consists of young adult college students, typically between the ages of 18 and 25 years. Depending on the study, older subjects can range in age from the 60s, termed 'young-old,' to older than 90 years, termed the 'oldest-old.'

EEG and ERP methods

The EEG is a measure of continuous variations in voltage over time from electrodes placed on the scalp. It was first developed in the 1930s by Hans Berger, and it was one of the first methods for non-invasively measuring brain activity in humans. Event-related potentials can be used to define neural activity measured by the EEG that is time-locked to an event, such as the onset of a stimulus, the onset of a behavioral response, or their correlation with cognitive processing. An example of an ERP following presentation of an auditory stimulus is shown in Figure 1(a). Depending on the component of interest, the EEG waveforms associated with tens to thousands of events (stimuli or behavioral responses) are averaged together to yield the ERP. The main ERP measures employed are peak amplitude and latency of individual components. Amplitude is conventionally defined as the voltage difference of a peak along the waveform relative to a baseline period, such as the mean amplitude during the 100 ms before stimulus presentation (Figure 1(b)). Latency is defined as the time between stimulus onset and the peak of a given component.

Processing Speed and Reaction Time

A basic observation of cognitive aging is that performance of a wide variety of reaction time tasks declines during the course of normal aging, even when subjects are in excellent health. This is often depicted using a Brinley plot, whereby reaction times on multiple tasks in young subjects are arranged in ascending order on the x-axis and plotted against the respective reaction times of older subjects performing the same tasks on the y-axis. Reaction times in young and older subjects show a linear relationship, with a slope >1 indicating larger absolute age effects on reaction time on tasks having longer overall reaction times.

Large-scale cross-sectional and longitudinal studies suggest that a general factor can account for most of the typical cognitive changes associated with normal aging (>90% of variance). Theorists have proposed that the general factor may indicate fluid intelligence or a 'resource' such as processing speed. Processing speed as a general factor is supported by observations that age differences in reaction time can predict age differences on more complicated tests, such as episodic memory or verbal fluency. Thus, slowing of reaction time appears to reflect a fundamental aspect of cognitive aging. As a practical matter, when comparing reaction times among young and older subjects performing two or more tasks, it is often worthwhile not only to express the absolute differences between tasks for each group but also to include a relative measure, such as percentage increase from task A to B in each group, to take into account age-related slowing of overall reaction time.

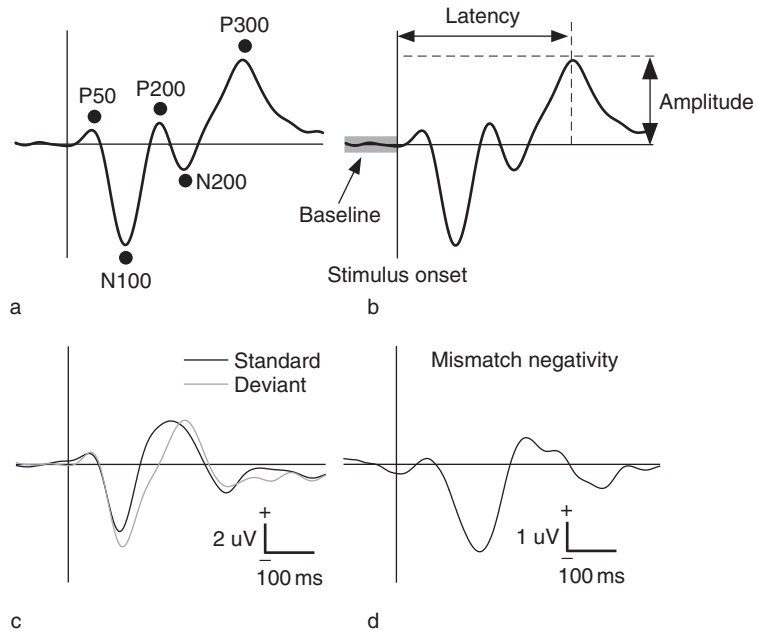

Figure 1 Examples of event-related potential components and measurement of amplitude and latency. (a) Example of auditory event-related potentials elicited by infrequent target stimuli during a target detection task in young subjects. The vertical line indicates stimulus onset. Voltage is plotted as a function of time, from a 100 ms prestimulus baseline to 500 ms after stimulus onset. Components are named by convention using polarity ('P' or 'N') followed by approximate latency. Thus, the 'P50' is a positive component having a latency of ~50 ms. The component labeled 'P300' refers to the P300b (or P3b) component. A different component called the P3a is elicited by distractor stimuli, often a novel sound other than the task-relevant target or nontarget stimuli. (b) Illustration of amplitude and latency measures of the P300 component. (c) Example of the mismatch negativity. Subjects passively listened to standard (90%) and deviant (10%) tone stimuli. Event-related potentials to standard and deviant stimuli are plotted together. Note that potentials to deviants are more negative than standards from ~100 to 240 ms after stimulus presentation, indicating the mismatch negativity. (d) Example of mismatch negativity difference waveform. To better visualize the mismatch negativity, the event-related potential waveform at each time point for standards can be subtracted from the deviant waveform. Note the differences in scale for (d) vs. (a)–(c). To be consistent with event-related potential examples in (a)–(c), negativity is plotted downward. However, in many publications the mismatch negativity is shown with negative polarity plotted upwards.

Sensory Memory

Sensory memory is a relatively automatic form of memory and has a duration of several seconds. In the auditory modality sensory memory is important for the perception of speech and various aspects of auditory scene perception. The mismatch negativity (MMN) is an ERP that has been extensively used to study sensory memory (**Figures 1(c) and 1(d)**). The MMN is elicited by a stimulus that deviates from a previously established pattern of stimuli presented sequentially. The rationale for its use is that sensory memory can be probed by first establishing a standard sensory memory representation and then presenting a stimulus that is predicted to differ from the sensory memory representation in some respect. If the MMN is elicited, it can be inferred that the deviant aspect of the stimulus was not a property of the standard stimulus representation (mismatch). It is a way to infer memory representations by defining the 'limits' of the sensory memory representation, after which a stimulus is considered different.

Studies of the MMN in normal aging consistently report that when the deviant stimulus differs from the standard in terms of acoustic properties such as frequency or intensity, there are little or no age differences. However, if deviant stimuli differ in terms of timing, such as stimulus duration or interstimulus interval, then age differences are often evident, with smaller MMN amplitudes in older compared to young subjects.

Working Memory

Working memory is the process of maintaining and manipulating information that is relevant to performing a task for short periods of time (seconds to minutes). A typical example of maintaining information in working memory is remembering the digits of an unfamiliar phone number prior to placing the call. An example of manipulation is the addition of an area code before the phone number in working memory and reordering the digits to correctly dial the number.

Working memory is frequently studied using variations of a paradigm developed by Saul Sternberg, which provides the ability to document both behavioral and brain activity changes accompanying working memory retrieval. In the Sternberg task, subjects perform a series of working memory trials. In each trial a set of items, typically numbers or letters, are memorized in sequence. After a delay of a few seconds (typically 2–4 s), retrieval is assessed by presenting a probe item. Subjects then determine whether or not the probe was a member of the current memory set, and press one of two buttons to indicate their decision. When a range of memory loads are tested (from one to approximately six to eight items), reaction time shows a positive linear relationship with memory load, with longer reaction times to probes when more items are memorized. Thus, the Sternberg task is a useful tool to define systematically the impact of memory load on reaction time and ERP measures during encoding (memorized items) and memory retrieval (probe), which is important because limited capacity is a defining feature of working memory.

The slope of the memory load versus reaction time function is usually steeper in older subjects than in young subjects, indicating slower retrieval speed. Age effects are sometimes smaller when subjects are rigorously screened to exclude all but the healthiest older subjects or use young-older people (in their 50s), but retrieval speed is still slower in older groups relative to young adults in their 20s. Although not always examined, slower retrieval speed can persist even when there are controls for overall differences in reaction time, if present.

Most ERP studies using the Sternberg paradigm have focused on a prolonged late positive waveform present from approximately 200 to 800 ms that is elicited by probes. The most prominent component is often called the 'late positive wave' or 'P300'; it is similar to the P300 shown in **Figure 1** in response to targets, but it is longer in duration. Detailed analyses indicate that probes elicit a set of components comprising the late positive waveform and likely correlate with somewhat different cognitive processes during memory retrieval. Control tasks show that the late positive wave is not present when the same stimuli are presented as in the working memory task, but probes do not require comparison to memory. In young subjects, increased memory load is associated with increases in P300 latency and decreases in P300 amplitude. Older subjects show a similar pattern of results as a function of memory load, with overall increases in P300 latency and a weaker correlation between P300 latency and reaction time.

In auditory versions of the Sternberg task, a component called the 'sustained frontal negativity' is elicited by memorized items and probes when environmental sounds are used, especially when most probes match memorized items. It is unclear if the sustained frontal negativity is present for other modalities. The amplitude of the sustained frontal negativity is reduced in older subjects relative to young subjects. It has been suggested that a smaller sustained frontal negativity may be associated with age-related changes in prefrontal cortex function during working memory.

Studies have examined potentials that occur before the P300/late positive wave and sustained frontal negativity. The purpose of examining shorter latency components is to determine if modality-specific components associated with sensory cortical activity are also modulated by working memory demands. In the auditory modality, amplitude of the N100 component is affected by memory load during both encoding of items and retrieval, with linear reductions in amplitude as memory load is increased. Amplitude of the N100 to probes is also modulated by the details of encoding, such as the order of item presentation, and is not affected by memory load when memorized items are presented in the visual modality. Both results are consistent with psychological theories emphasizing a linkage between memory retrieval and specific processes during memory encoding, such that memory retrieval may reinstate aspects of perceptual processing.

Compared to young subjects, older subjects exhibit two main differences in auditory potentials in the Sternberg task. During encoding, young subjects have shorter N100 latencies with increasing memory load, whereas N100 latency in older subjects is invariant across memory loads. The second main age difference is that probe N100 amplitudes to memorized items show increases with greater memory load in older subjects, but the opposite is seen in the young as amplitudes decrease with memory load. In addition, N100 amplitudes to probes that were not in the memorized list also exhibit smaller changes in N100 amplitude as a function of memory load relative to young subjects. The significance of age-related changes in auditory potentials is unclear, but it has been proposed that changes in prefrontal cortex function may be relevant because prefrontal cortex can regulate the responsiveness of sensory cortex.

Declarative: Episodic Memory

ERP and EEG studies of episodic memory typically employ a recognition format instead of having subjects freely recall the memorized material. In recognition memory tests, the experimenter presents stimulus items to which the subject has to make a judgment,

such as whether or not the item was previously memorized. A recognition format is preferred for ERP and EEG studies because the ability to precisely control stimulus timing is necessary to define neural responses, such as ERPs, that are time-locked to stimulus presentation. The Sternberg task described previously is also a recognition task because subjects judge whether the probe was a memorized item.

Target detection, or 'oddball,' tasks, first described by Samuel Sutton and colleagues in 1965, can be considered a simple recognition memory test. Subjects are presented a sequence of stimuli and are instructed to detect the infrequent occurrence of a predefined target stimulus. Subjects usually either press a button to each target or count the number of targets presented in the sequence. An example of ERPs recorded during a target detection task is shown in the top of Figure 1. Both target and nontarget stimuli elicit sensory potentials, such as P50, N100, and P200, in response to auditory stimuli. However, targets also elicit two additional components that are small or absent compared to nontargets: the N200 and P300 complex. N200 and P300 are associated with the active identification of targets because they are not present during passive listening tasks, such as those used to elicit the mismatch negativity. When the responses to nontargets are analyzed in detail as a function of the stimulus sequence history, a positive component at 250 ms can be seen when the likelihood for a target to appear is high. Thus, each stimulus in the train appears to be processed slightly differently depending on the expectations of the listener.

Many studies have shown that P300 changes substantially during normal development. P300 latency is shortest in early adulthood (teens) and has a gradual, progressive, increase in latency across age cohorts until at least the 10th decade of life (Figure 2). As a group, P300 latency to auditory stimuli for subjects in their 80s is approximately 100 ms longer than that of young adults in their early 20s (300 vs. 400 ms). The same pattern of age differences is observed for P300 to visual stimuli, which has somewhat longer latencies (by approximately 60 ms for all age groups) relative to auditory stimuli. When the distribution of voltage across the entire scalp at the time of the P300 peak is examined with an array of electrodes, the focus of maximum amplitude is more anterior in older relative to younger subjects. Age differences in P300 latency are probably not attributable to differences in sensory function because stimuli are easily distinguishable by all subjects, and age differences in P300 latency are not affected by manipulations of stimulus discriminability. Response preparation is also unlikely because, unlike most reaction time tasks, reaction times to targets show little change as

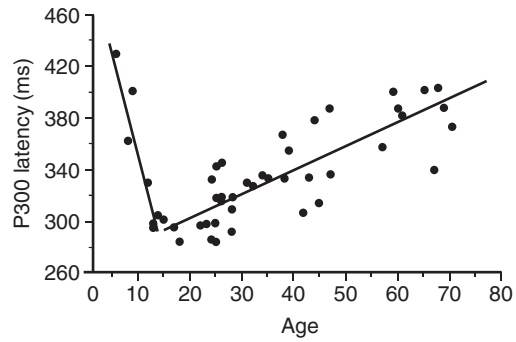

Figure 2 P300 latency as a function of age. Linear decreases in P300 latency are seen during childhood, with shortest latencies during late adolescence. Gradual increases in P300 latency are then observed throughout the adult years. An auditory target detection task was used to assess P300 latency to infrequent targets (20% stimuli) that differed from nontargets according to pitch. Circles indicate measures from individual subjects. Note that 'P300' refers to the P300b component. Data replotted from Goodin DS, Squires KC, Henderson BH, and Starr A (1978) Age-related variations in evoked potentials to auditory stimuli in normal human subjects. *Electroencephalography and Clinical Neurophysiology* 44(4): 447–458.

a function of age in healthy subjects and P300 latency shows little or no change with response demands. Thus, although the precise cognitive correlates of P300 latency are uncertain, age differences are likely associated with memory-dependent stimulus classification processes.

In most episodic memory studies subjects are requested to remember multiple items rather than just one target item as in target detection. Two main paradigms are used extensively in episodic memory studies using ERPs: study/test and continuous recognition. In the study/test paradigm, items are encoded during a study period and memory is later assessed during a test period. One advantage of the study/test paradigm is that it permits separate assessment of potentials during encoding and retrieval. In contrast, in continuous recognition tasks, subjects are presented a sequence of items and each item is typically judged as either old or new. Thus, encoding and retrieval processes are operative during each item presentation.

Presentation of items in both paradigms elicits positive slow wave potentials. The slow waves begin as early as ~250 ms and last for hundreds of milliseconds depending on the component and task. During retrieval in both study/test and continuous recognition previously studied items elicit slow waves that are more positive relative to new items – a phenomenon called the 'ERP old/new effect.' In the study phase of study/test tasks positive slow waves following item presentation are also present. Moreover, items that are later remembered, via either free recall or correct recognition, elicit more positive slow

waves than items that are later not remembered. This finding is called the 'subsequent memory effect.' Here, we focus on episodic memory retrieval because only a few studies have examined subsequent memory effects as a function of aging, and the subsequent memory effect is sometimes not observed in young subjects.

There are two main positive slow waves associated with old/new effects in study/test paradigms: one that is largest over left parietal recording sites and another that is maximal over right frontal areas. The slow waves are also distinguished by onset time and duration, with the left parietal slow wave having an earlier onset and shorter duration compared to the right frontal slow wave. The two slow waves can be dissociated by manipulating task demands. The left parietal slow wave is associated with retrieval of item information, such as old versus new, whereas the right frontal slow wave is associated with item information and contextual information. Item information is specific knowledge presented during the study session, whereas contextual information reflects other aspects of the learning situation that are either not necessarily relevant to the memory test or are common to multiple items. For example, subjects may memorize words presented on two separate lists and be tested later using item recognition for the words (old/new words). In this example, the actual words on the lists would be considered item information, whereas remembering which of the two lists contained a given word would be considered contextual information.

The old/new effect at left parietal sites changes little with aging, suggesting that certain retrieval processes associated with item recognition may not change substantially during aging. In contrast, the right frontal slow wave does exhibit age differences. Although additional study is needed, the right frontal slow wave is usually attenuated in older subjects relative to young subjects when judgments of context are assessed. Age differences in right frontal slow wave amplitude are accompanied by behavioral differences, with less accurate judgments of item context in older subjects. Thus, behavioral and ERPs studies are consistent in showing that normal aging is associated with small differences in recognition memory for items, whereas substantial age differences are evident when memory for context is assessed.

Conclusions

This summary of findings on memory and aging using reaction time and electrophysiological measures underscores the benefit of combining behavioral and physiological tools. Behavioral measures reflect the overt product of cognitive and response selection. Different components of electrophysiological responses reflect specific aspects of covert brain processing. Thus, ERPs and EEG provide a tool to define covert memory processes and compare these memory processes as a function of normal aging.

See also: Aging of the Brain; Aging: Extracellular Space; Brain Volume: Age-Related Changes; Cognition in Aging and Age-Related Disease; Conditioned Reflex; Electroencephalography (EEG); Electrophysiology: EEG and ERP Analysis; Functional Neuroimaging Studies of Aging.

Further Reading

Allison T, Hume AL, Wood CC, and Goff WR (1984) Developmental and aging changes in somatosensory, auditory and visual evoked potentials. *Electroencephalography and Clinical Neurophysiology* 58(1): 14–24.

Friedman D (2000) Event-related brain potential investigations of memory and aging. *Biological Psychology* 54(1–3): 175–206.

Goodin DS, Squires KC, Henderson BH, and Starr A (1978) Age-related variations in evoked potentials to auditory stimuli in normal human subjects. *Electroencephalography and Clinical Neurophysiology* 44(4): 447–458.

Iragui VJ, Kutas M, Mitchiner MR, and Hillyard SA (1993) Effects of aging on event-related brain potentials and reaction times in an auditory oddball task. *Psychophysiology* 30(1): 10–22.

Naatanen R (1992) *Attention and Brain Function.* Hillsdale, NJ: Lawrence Erlbaum.

Sternberg S (1966) High-speed scanning in human memory. *Science* 153(736): 652–654.

Sutton S, Braren M, Zubin J, and John ER (1965) Evoked-potential correlates of stimulus uncertainty. *Science* 150(700): 1187–1188.

Aging: Extracellular Space

E Syková, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, and Charles University, Prague, Czech Republic

© 2009 Elsevier Ltd. All rights reserved.

Gross Anatomic Aging Changes: Extracellular Space

Aging, Alzheimer's disease, and many degenerative diseases are accompanied by serious cognitive deficits, particularly impaired learning and memory loss. This decline in old age is a consequence of changes in brain anatomy, morphology, and volume, as well as functional deficits. Nervous tissue, particularly in the hippocampus and cortex, is subject to various degenerative processes, including a decreased number and efficacy of synapses, a decrease in transmitter release, neuronal loss, a decreased number of dendritic spines, astrogliosis, changes in astrocytic morphology, demyelination, deposits of β-amyloid, changes in extracellular matrix proteins, and, consequently, changes in the brain's diffusion parameters. These age-related changes in the morphology of neural elements have no universal pattern across the entire brain. Despite the fact that aging is associated with cognitive impairment, in most brain areas neuronal loss does not have a significant role in the observed deficits. It has been suggested that rather small, region-specific changes in dendritic branching and spine density are more characteristic changes in neuronal morphology.

Diffusion and Extrasynaptic Transmission

The gross anatomic changes and other subtle morphological changes not only affect the efficacy of signal transmission at synapses, but also the function of glia and extrasynaptic ('volume') transmission mediated by the diffusion of transmitters and other substances through the volume of the extracellular space (ECS). This mode of communication by diffusion (i.e., without synapses) provides a mechanism of long-range information processing in functions such as vigilance, sleep, chronic pain, hunger, depression, long-term potentiation (LTP), long-term depression (LTD), memory formation, and other plastic changes in the central nervous system (CNS). Neurons interact both by synapses and by the diffusion of ions and neurotransmitters in the extracellular space. Since glial cells do not have synapses, their communication with neurons is only mediated by the diffusion of ions and neuroactive substances in the ECS. Neurons and glia release ions, transmitters, and various other neuroactive substances into the ECS. Substances released nonsynaptically diffuse through the ECS and bind to extrasynaptic, usually high-affinity, binding sites located on neurons, axons, and glial cells.

Diffusion in the ECS is critically dependent on the structure and physicochemical properties of the ECS – the nerve cell microenvironment. These properties vary, however, around each cell and in different brain regions. Certain synapses ('private synapses') or even whole neurons are clearly tightly ensheathed by glial processes and by the extracellular matrix, forming so-called perineuronal nets, while others are left more 'naked.' The 'open' synapses are more easily reached by molecules diffusing in the ECS (Figure 1). On the other hand, many mediators, including glutamate and γ-aminobutyric acid (GABA), bind to high-affinity binding sites located on nonsynaptic parts of the membranes of neurons and glia. Mediators that escape from the synaptic clefts at an activated synapse, particularly following repetitive stimulation, diffuse in the ECS and can cross-react with receptors in nearby synapses. This phenomenon, called 'cross-talk' between synapses by the 'spillover' of a transmitter (e.g., glutamate, GABA, glycine), has been proposed to account for LTP, LTD, and memory formation in the rat hippocampus. The cross-talk between synapses, and the efficacy and directionality of volume transmission, are critically dependent on the diffusion properties of the ECS.

There is increasing evidence that changes in neuron–glia interaction (e.g., glial coverage and/or the retraction of glial processes from synapses) occur in many brain regions during physiological and pathological functional changes. The glial environment of neurons is likely to be a key factor in the regulation of intersynaptic communication mediated by glutamate. For example, most synaptically released glutamate is taken up by high-affinity transporters, such as GLT-1 and GLAST, that are located on surrounding astrocytes. Moreover, glial cells represent a diffusion barrier in the ECS, hindering the movement of neuroactive substances within the tissue. Long-term changes in the physical and chemical parameters of the ECS accompany many physiological (e.g., development, aging, lactation) and pathological (e.g., ischemia, epilepsy, hydrocephalus, demyelination, CNS trauma) states.

Measurements of the Diffusion Parameters of the ECS

The diffusion of substances in a free medium, such as water or diluted agar, is described by Fick's laws.

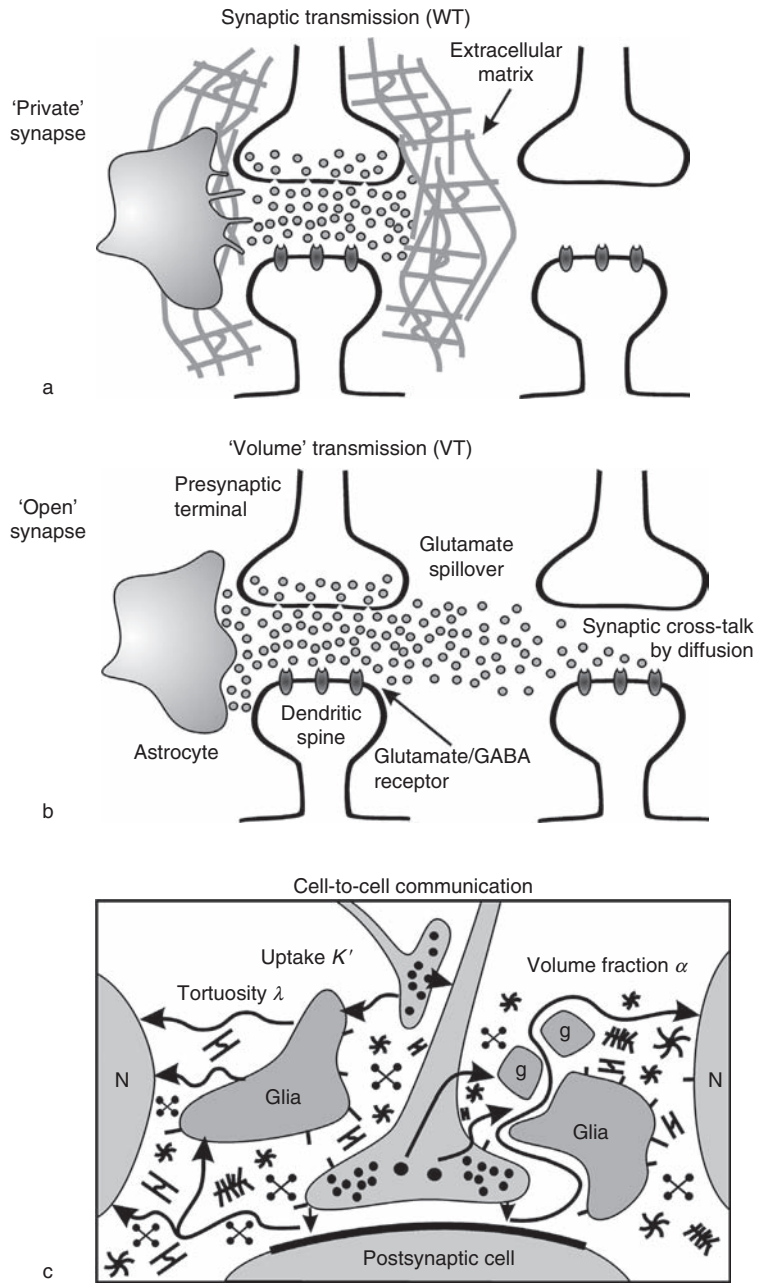

Figure 1 Concept of synaptic 'wired' transmission (WT) and extrasynaptic 'volume' transmission (VT). (a) Closed synapses are typical of synaptic transmission. The synapse is tightly ensheathed by glial processes and the extracellular matrix, forming perineuronal or perisynaptic nets. (b) An open synapse is typical of volume transmission. It allows the escape of transmitter – for example, glutamate and γ-aminobutyric acid (GABA) – from the synaptic cleft (spillover), diffusion in the extracellular space, and binding to receptors on nearby synapses. This phenomenon is known as 'cross-talk' between synapses. Spillover may also lead to plastic changes, inducing the formation of new synapses or eliciting the rearrangement of astrocytic processes around the synapse and the formation of diffusion barriers. (c) Schematic of central nervous system architecture, showing how it is composed of neurons (N), axons, glial cells (g), cellular processes, molecules of the extracellular matrix, and intercellular channels between the cells. The architecture affects the movement (diffusion) of substances in the brain, which is critically dependent on channel size, extracellular space tortuosity, and cellular uptake. Adapted from Syková E (2001) Glial diffusion barriers during aging and pathological states. *Progress in Brain Research* 132: 339–363.

In contrast to a free medium, the diffusion of any substance in the ECS is hindered by the ECS size, the presence of various obstacles, and also by cellular uptake. To take these factors into account, it was necessary to modify Fick's original diffusion equations. First, diffusion in the CNS is constrained by the restricted volume of the tissue available for the diffusing particles – that is, by the extracellular

space volume fraction (α), which is a dimensionless quantity and is defined as the ratio between the volume of the ECS and the total volume of the tissue. It is now evident that the ECS in the adult brain amounts to about 20% of the total brain volume (i.e., $\alpha = 0.2$). Second, the free diffusion coefficient (D) in the brain is reduced by the tortuosity factor (λ). ECS tortuosity is defined as $\lambda = (D/\mathrm{ADC})^{0.5}$, where D is the free diffusion coefficient and ADC is the apparent diffusion coefficient in the brain. As a result of tortuosity, D is reduced to an apparent diffusion coefficient $\mathrm{ADC} = D/\lambda^2$. Thus, any substance diffusing in the ECS is hindered by membrane obstructions, glycoproteins, macromolecules of the extracellular matrix, charged molecules, and fine neuronal and glial cell processes. Third, substances released into the ECS are transported across membranes by nonspecific concentration-dependent uptake (k'). In many cases however, these substances are transported by energy-dependent uptake systems that obey nonlinear kinetics. When these three factors (α, λ, and k') are incorporated into Fick's law, diffusion in the CNS is described fairly satisfactorily.

At the present time, the real-time iontophoretic method is used to determine the absolute values of the ECS diffusion parameters and their dynamic changes in nervous tissue *in vitro* as well as *in vivo*. The second method that is also frequently used to study ECS volume and geometry is diffusion-weighted magnetic resonance imaging (DW-MRI). DW-MRI provides information only about the apparent diffusion coefficient of water. A relationship between an increase in the *ADC* of water and a decrease in ECS volume fraction has recently been found during pathological states as well as during aging.

The real-time iontophoretic method (**Figure 2(a)**) uses ion-selective microelectrodes to measure the diffusion of ions to which the cell membranes are relatively impermeable – for example, tetraethylammonium and tetramethylammonium ions (TEA^+, TMA^+) or choline. These substances are injected into the nervous tissue by pressure or by iontophoresis from an electrode aligned parallel to a double-barreled ion-selective microelectrode (ISM) at a fixed distance. Usually, such an electrode array is made by gluing together an iontophoretic pipette and a TMA^+-sensitive ISM with a tip separation of 130–200 μm. In the case of iontophoretic application, TMA^+ is released into the ECS by applying a current step of +100 nA with a duration of 40–80 s. The released TMA^+ is recorded with the TMA^+-sensitive ISM as a diffusion curve, which is then transferred to a computer to calculate α, λ, and k' from the modified diffusion equation. Values of the ECS volume, ADC, tortuosity, and nonspecific cellular uptake are extracted by a nonlinear curve-fitting simplex algorithm applied on the diffusion curves.

By introducing the tortuosity factor into diffusion measurements in nervous tissue, it soon became evident that diffusion is not uniform in all directions and is affected by the presence of diffusion barriers, including neuronal and glial processes, myelin sheaths, macromolecules, and molecules with fixed negative surface charges. This so-called anisotropic diffusion preferentially channels the movement of substances in the ECS in one direction, (e.g., along axons) and may, therefore, be responsible for a certain degree of specificity in volume transmission. Diffusion anisotropy was found in the CNS in the molecular and granular layers of the cerebellum, in the hippocampus, and in the auditory but not in the somatosensory cortex, and a number of studies have revealed that it is present in the myelinated white matter of the corpus callosum or spinal cord. It was shown that diffusion anisotropy in white matter increases during development. At first, diffusion in unmyelinated tissue is isotropic; it becomes more anisotropic as myelination progresses.

ECS Diffusion Parameters during Aging

Aging is not only accompanied by a decrease in the efficacy of signal transmission at synapses, but also by changes in the ECS diffusion parameters and consequently by a deficit in extrasynaptic transmission. Using the TMA^+ method, the ECS diffusion parameters were measured in the cortex, corpus callosum, and hippocampus (CA1, CA3, and dentate gyrus) of aged rats. The measurements revealed a significant decrease in ECS volume fraction. One of the explanations as to why α in the cortex, corpus callosum, and hippocampus of senescent rats and mice is significantly lower than in young adults could be astrogliosis in the aged brain. Increased glial fibrillary acidic protein (GFAP) staining and an increase in the size and fibrous character of astrocytes have been found in the cortex, corpus callosum, and hippocampus of senescent rats. Other changes could account for the decreases in λ and for the disruption of tissue anisotropy. In the hippocampus in the CA1 and CA3 regions, as well as in the dentate gyrus, changes in the arrangement of fine astrocytic processes were regularly found. These are normally organized in parallel in the x–y plane, but this organization totally disappears during aging. Moreover, decreased staining for chondroitin sulfate proteoglycans and for fibronectin suggests a loss of extracellular matrix macromolecules (**Figure 3**).

If diffusion in a particular brain region is anisotropic, then the correct value of the ECS volume

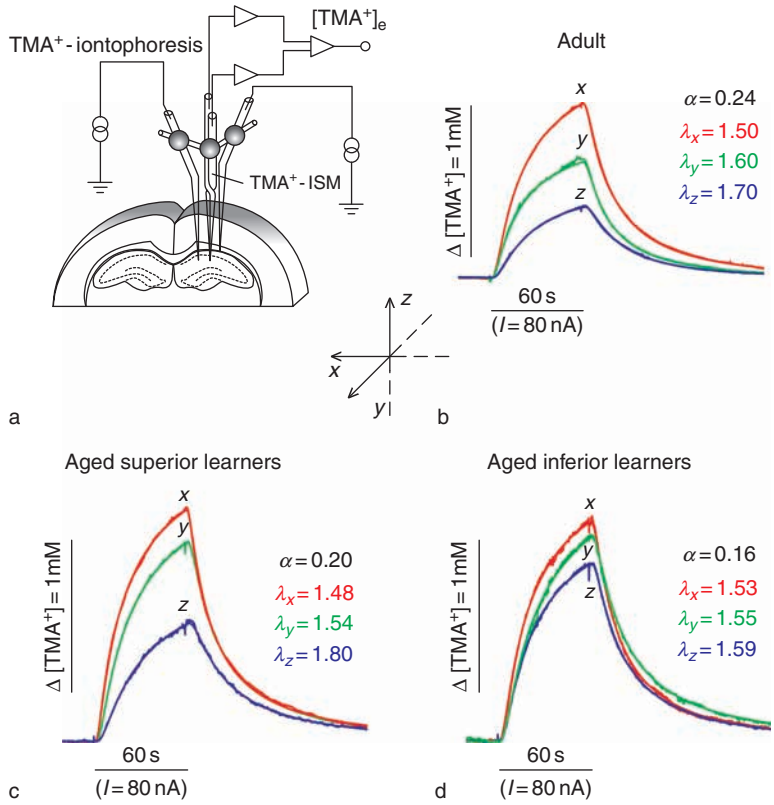

Figure 2 Experimental setup: tetramethylammonium ion (TMA^+) diffusion curves and typical extracellular space diffusion parameters α (volume fraction) and λ (tortuosity) in the hippocampus dentate gyrus of a young adult rat and an aged rat with memory impairment (rats were tested in a Morris water maze). (a) Schema of the experimental arrangement. A TMA^+-selective double-barreled ion-selective microelectrode (ISM) was glued to a bent iontophoresis microelectrode. The separation between electrode tips was 130–200 μm. (b) Anisotropic diffusion in the dentate gyrus of a young adult rat required measurements of TMA^+ diffusion curves (concentration–time profiles) along three orthogonal axes (*x*, mediolateral; *y*, rostrocaudal; *z*, dorsoventral). The slower rises in the *z* axis (compared to the *y* axis) and in the *y* axis (compared to the *x* axis) indicate a higher tortuosity and more restricted diffusion. The amplitudes of the curves show that the TMA^+ concentration, at approximately the same distance from the tip of the iontophoresis electrode, is much higher along the *x* axis than along the *y* axis and is even higher than along the *z* axis (λ_x, λ_y, λ_z). Note that the actual extracellular space volume fraction α is about 0.2 and can be calculated only when measurements are done along the *x*, *y*, and *z* axes. (c, d) Volume fraction and anisotropy decrease during aging. Note the greater decrease in volume fraction and that the anisotropy is almost lost in an aged rat with memory impairment (d). Note that the diffusion curves are higher, showing that α is smaller, and that their rise and decay times are longer when λ (diffusion barriers) increases. Adapted from Syková E, Mazel T, Hasenöhrl RU, et al. (2002) Learning deficits in aged rats related to decrease in extracellular volume and loss of diffusion anisotropy in hippocampus. *Hippocampus* 12: 469–479.

fraction cannot be calculated from measurements done only in one direction. For anisotropic diffusion, the diagonal components of the tortuosity tensor are not equal, and generally its nondiagonal components need not be zero. Nevertheless, if a suitable referential frame is chosen (i.e., if we measure in three privileged orthogonal directions), neglecting the nondiagonal components becomes possible, and the correct value of the ECS size can thus be determined. Therefore, TMA^+ diffusion was measured in the ECS independently along three orthogonal axes (*x*, transversal; *y*, sagittal; *z*, vertical) in the cortex, corpus callosum, and hippocampus of aged rats and mice. These studies revealed that the mean ECS volume fraction α is significantly lower in aged rats (26–32 months old), ranging from 0.17 to 0.19, than in young adults (3–4 months old), in which α ranges from 0.21 to 0.22. Nonspecific uptake k' is also significantly lower in aged rats. From Figure 2 it is evident that the diffusion curves for the hippocampus are larger in the aged rat than in the young one – that is, the space available for TMA^+ diffusion is smaller. Although the mean α values along the *x* axis were not significantly different between young and aged rats, the values were significantly lower in aged rats along the *y* and *z* axes. This means that there is a loss of anisotropy in the aging hippocampus, particularly in the CA3 region and the dentate gyrus.

The three-dimensional pattern of diffusion away from a point source can be illustrated by

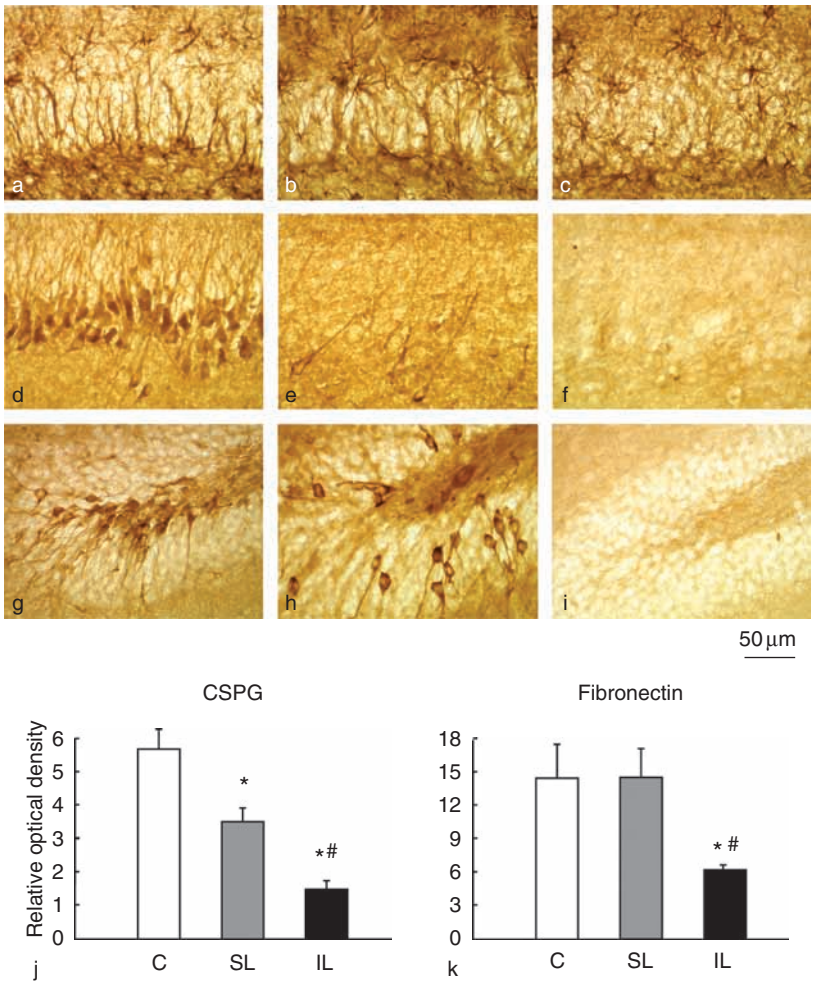

Figure 3 Structural changes in aged superior and inferior learners. (a–c) Astrocytes in the dentate gyrus (inner blade) of an adult rat, an aged superior learner (SL), and an aged inferior learner (IL). Note the loss of radial organization of the astrocytic processes in the aged IL. Staining for chondroitin sulfate proteoglycans (CSPG) in the hippocampal CA3 region (d–f), and for fibronectin in the dentate gyrus (g–i), shows a decrease in perineuronal staining in the aged SL and a loss in the aged IL. The graphs show the relative optical densities of CSPG (j) and fibronectin cell attachment fragment (k) immunoreactivity. Note the significant decrease in CSPG immunoreactivity between each group (C, control) and the decrease in fibronectin immunoreactivity in the aged IL only. *significant difference ($p < 0.05$) compared to control (adults). #significant difference ($p < 0.05$) compared to aged superior learners. Adapted from Syková E, Mazel T, Hasenöhrl RU, et al. (2002) Learning deficits in aged rats related to decrease in extracellular volume and loss of diffusion anisotropy in hippocampus. *Hippocampus* 12: 469–479.

constructing isoconcentration spheres (isotropic diffusion) or ellipsoids (anisotropic diffusion) for extracellular TMA^+ concentration. The surfaces in Figure 4 represent the locations where the TMA^+ concentration first reached 1 mM, 60 s after its application in the center. The ellipsoid in the hippocampus of the young adult rat reflects the different abilities of substances to diffuse along the x, y, and z axes, while the sphere in the hippocampus of the aged rat shows isotropic diffusion. The smaller ECS volume fraction in aged rats is reflected in the sphere's being larger than the ellipsoid. Indeed, we found that there is a significant decrease in the ADC of many neuroactive substances, including the ADC of water (Figure 5), in the aging brain, which accompanies a decrease in ECS volume.

The hippocampus is well known for its role in memory formation, especially declarative memory. Indeed, it has been found that the degree of learning deficit during aging correlates with changes in α, λ, and k'. It is therefore reasonable to assume that diffusion anisotropy leads to a certain degree of specificity in extrasynaptic communication. There is a significant difference between mildly and severely behaviorally impaired rats (rats were tested in a Morris water maze), which is particularly apparent in the hippocampus. The ECS in the dentate gyrus of severely impaired rats (bad learners) is significantly smaller

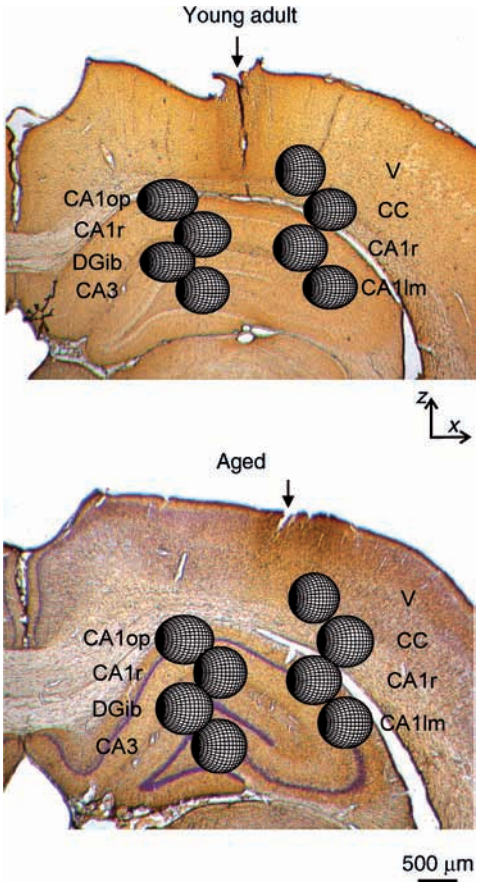

Figure 4 Diffusion parameters in a young adult (3 months old) rat and in an aged (28 months old) rat with a learning deficit in the Morris water maze. Data were recorded in anesthetized animals; microelectrode tracks were verified after the experiments (see arrows); isoconcentration surfaces for a 1 mM tetramethylammonium ion concentration contour 60 s after the onset of an 80 nA iontophoretic pulse. The surfaces were generated using the actual values of volume fraction and tortuosity. The ellipsoids represent anisotropic diffusion in a young adult rat. The larger sphere in an aged rat corresponds to isotropic diffusion and to a small extracellular space volume fraction. It demonstrates that diffusion from any given source will lead to a higher concentration of substances in the surrounding tissue and a larger action radius in aged rats than in young adults. Anisotropy is almost lost in the aged rat. Adapted from Syková E, Mazel T, Vargová L, et al. (2000) Extracellular space diffusion and pathological states. *Progress in Brain Research* 125: 155–178.

than in mildly impaired rats (good learners). Also, anisotropy in the hippocampus of bad learners, particularly in the dentate gyrus, is much reduced, while a substantial degree of anisotropy is still present in aged rats with good learning performance (Figures 2 and 3). Anisotropy might be important for extrasynaptic transmission by channeling the flux of substances in a preferential direction, and its loss may severely disrupt extrasynaptic communication in the CNS, which has been suggested to play an important role in memory formation.

The ensuing decrease in the ECS size could be explained by the disappearance of a significant part of the extracellular matrix, partial neuronal degeneration and loss, a decrease in the number of dendritic processes, demyelination, and changes in glia. Indeed, a decrease in fibronectin and chondroitin sulfate proteoglycan staining was found in the hippocampus of mildly impaired aged rats and almost a complete loss of staining was seen in severely impaired aged rats (Figure 3). Chondroitin sulfate proteoglycans participate in multiple cellular processes, such as axonal outgrowth, axonal branching, and synaptogenesis, which are important for the formation of memory traces. The observed loss of anisotropy in senescent rats could therefore lead to impaired cortical and, particularly, hippocampal function. Moreover, the decrease in ECS volume could be responsible for the greater susceptibility of the aged brain to pathological events, particularly to ischemia, and for the poorer outcome of clinical therapy and the more limited recovery of affected tissue after anesthesia or any severe insult.

The decrease in ECS volume fraction during normal aging might be attributed to the loss of extracellular matrix molecules such as chondroitin sulfate proteoglycans, fibronectin, tenascin-R, or highly sialylated cell adhesion molecules, such as polysialylated neural cell adhesion molecules (PSA-NCAMs). The extracellular matrix molecules act like a sponge, binding a large number of water molecules and tending to occupy a lot of space, due to the mutual repulsion of their numerous negatively charged residues. Thus, their loss during aging might lead to a decrease in ECS volume. ECS shrinkage might also be related to the general decrease in hydration during aging. Life can be viewed as a process during which the highly hydrated state of embryos and neonates is transformed into a gradually more and more dehydrated one. Among other things, this dehydration is associated with an increase in colloid density, resulting in a decrease in catalytic enzyme activity and possibly contributing to the accumulation and deposition of certain substances, both intra- and extracellularly. A reduction in ECS volume would thus contribute, together with neuronal shrinkage, to the decrease in cortical volume observed during aging. These changes might be partially compensated for by concomitant reactive gliosis and glial cell hypertrophy.

Significant differences in ECS volume fraction were also found in aged female mice compared to age-matched males; an even more pronounced difference was found for ADC_w. This was probably the first observation of a gender difference in these two parameters. However, other differences have also been described: female mice have higher numbers of astrocytes and microglia in the hippocampus than

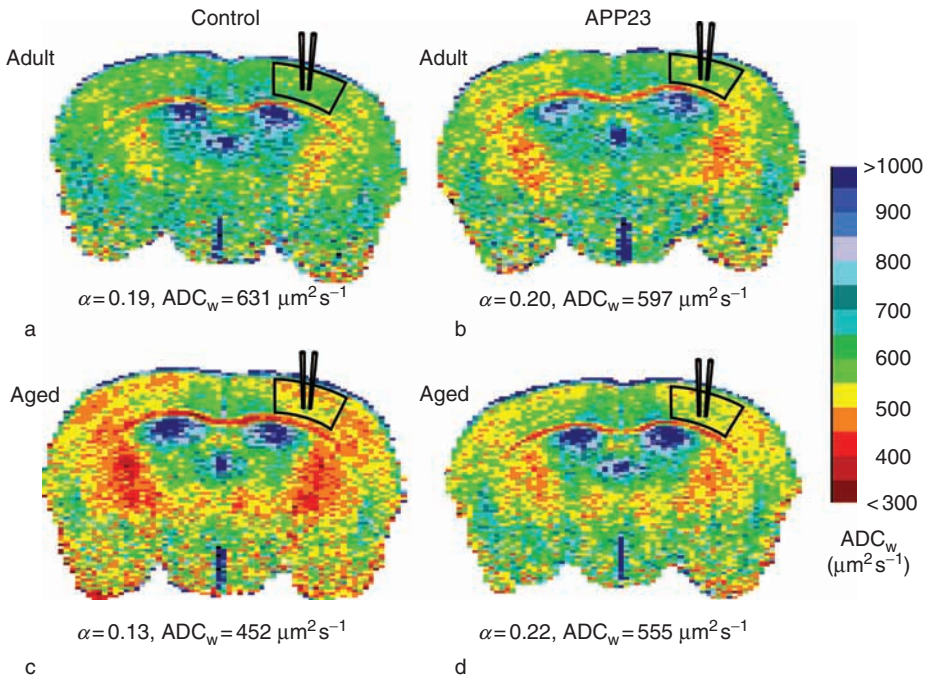

Figure 5 Apparent (water) diffusion coefficient (ADC_w) maps acquired in the brain of control and APP23 female mice. The mean value of ADC_w was calculated in the areas indicated. In the corresponding region, tetramethylammonium ion measurements were performed on the same animals a few days later. The mean values of the ADC_w and the extracellular space volume fraction (α) are given below each map. There were no differences between adult control (a) and adult transgenic (b) mice. (a, c) A decrease in ADC_w and α was found during aging in control mice. In aged APP23 mice (d), there was an increase in both ADC_w and α when compared with age-matched control mice (c). Adapted from Syková E, Voříšek I, Antonova T, et al. (2005) Changes in extracellular space size and geometry in APP23 transgenic mice – A model of Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* 102: 479–484.

males do, and this difference increases with age. These changes might be related to the decrease in 17β-estradiol after estropause, though there is no direct evidence for this. Changes in diffusion parameters may serve as an important indicator of pathological processes, and therefore diffusion changes have been studied under many experimental pathological states using the TMA method as well as diffusion-weighted MRI, which is important for diagnostic purposes.

Changes in ECS Diffusion Parameters in an APP23 Mouse Model of Alzheimer's Disease

Neuronal loss and changes in synaptic transmission are generally considered the main reasons for memory impairment in Alzheimer's disease. However, changes in extrasynaptic transmission might also be important. Changes in ECS diffusion parameters were found in the cortex of transgenic APP23 mice, in which overproduction of mutated human amyloid precursor protein (APP) leads to amyloid plaque formation. In these animals, TMA measurements were compared with data obtained in the same animals using diffusion-weighted magnetic resonance imaging to measure the apparent diffusion coefficient of water (ADC_w) in the tissue (Figure 5). Measurements were performed *in vivo* in 6- to 8-month-old and 17- to 25-month-old hemizygous APP23 mice and in age-matched controls. In 6- to 8-month-old APP23 mice, the ECS volume fraction and ADC_w were not significantly different from these factors in age-matched controls (mean α was 0.20, ADC_w was 618 $\mu m^2 s^{-1}$). Aging in 17- to 25-month-old controls was accompanied by decreases in the ECS volume fraction (to 0.15) and ADC_w (to 549 $\mu m^2 s^{-1}$) decreases significantly greater in females than in males, but there was no change in tortuosity. In aged, 17- to 25-month-old APP23 mice the ECS volume fraction increased to 0.22, and ECS tortuosity and ADC_w increased compared to age-matched controls. These results confirm that in mice as well as in rats, aging leads to a decrease in ECS volume. In contrast, the deposition of β-amyloid is associated with an increase in ECS volume fraction, tortuosity, and ADC_w. The impaired navigation observed in transgenic APP23 mice in the Morris water maze correlated with their amyloid plaque load, which was twice as high in females (20%) as in males (10%). Obviously, the volume of the amyloid plaques, which are located extracellularly, is not included in the ECS volume fraction as measured by the TMA method. If the ECS volume fraction is about 20% and the amyloid plaque load in old female transgenic mice is also about 20%,

the real ECS volume is the sum of both – about 40% of the total tissue volume.

Amyloid deposits, together with altered ECS diffusion properties, may therefore account for the impaired synaptic as well as extrasynaptic volume transmission and spatial cognition observed in old female transgenic mice. The enlargement of the ECS will contribute to a decrease in the extracellular concentrations of many important neuroactive substances, such as acetylcholine, dopamine, and serotonin. The functional radius of protein and peptide molecules, including nerve growth factors, will be compromised. It is therefore quite possible that there is a link between ECS volume, diffusion changes, and behavioral deficits in aged patients with Alzheimer's disease.

Concluding Remarks

The results obtained by diffusion measurements indicate that the degree of learning impairment in aged rats and mice correlates with changes in the ECS diffusion parameters. The changes in diffusion parameters during aging can therefore have an important functional significance. Anisotropy, which, particularly in the hippocampus and corpus callosum, may help to facilitate the diffusion of neurotransmitters and neuromodulators to regions occupied by their high-affinity extrasynaptic receptors, might have crucial importance for the specificity of signal transmission. The importance of anisotropy for the 'spill-over' of glutamate, for 'cross-talk' between synapses, and for LTP and LTD has been proposed. The observed loss of anisotropy in senescent rats could therefore lead to impaired cortical and, particularly, hippocampal function. The decrease in ECS size could be responsible for the greater susceptibility of the aged brain to pathological events, the poorer outcome of clinical therapy, and the more limited recovery of affected tissue after various insults.

See also: Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: MRI Studies; Amyloid: Vascular and Parenchymal; Animal Models of Alzheimer's Disease; Brain Volume: Age-Related Changes; Functional Neuroimaging Studies of Aging; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Long-Term Potentiation (LTP): NMDA Receptor Role; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Fuxe K, and Agnati LF (eds.) (1991) *Volume Transmission in the Brain. Novel Mechanisms for Neural Transmission.* New York: Raven Press.

Nicholson C and Phillips JM (1981) Ion diffusion modified by tortuosity and volume fraction in the extracellular microenvironment of the rat cerebellum. *Journal of Physiology (London)* 321: 225–257.

Nicholson C and Syková E (1998) Extracellular space structure revealed by diffusion analysis. *Trends in Neuroscience* 21: 207–215.

Syková E (1992) Ionic and volume changes in the microenvironment of nerve and receptor cells. In: Ottoson D (ed.) *Progress in Sensory Physiology*, pp. 1–167. Heidelberg: Springer.

Syková E (1997) The extracellular space in the CNS: Its regulation, volume and geometry in normal and pathological neuronal function. *Neuroscientist* 3: 28–41.

Syková E (2001) Glia and extracellular space diffusion parameters in the injured and aging brain. In: de Vellis J (ed.) *Neuroglia in the Aging Brain*, pp. 77–98. Totowa: Humana Press.

Syková E (2001) Glial diffusion barriers during aging and pathological states. *Progress in Brain Research* 132: 339–363.

Syková E (2003) Diffusion parameters of the extracellular space. *Israel Journal of Chemistry* 43: 55–69.

Syková E (2004) Extrasynaptic volume transmission and diffusion parameters of the extracellular space. *Neuroscience* 129: 861–876.

Syková E, Mazel T, Hasenöhrl RU, et al. (2002) Learning deficits in aged rats related to decrease in extracellular volume and loss of diffusion anisotropy in hippocampus. *Hippocampus* 12: 469–479.

Syková E, Mazel T, Vargová L, et al. (2000) Extracellular space diffusion and pathological states. *Progress in Brain Research* 125: 155–178.

Syková E, Voříšek I, Antonova T, et al. (2005) Changes in extracellular space, size, and geometry in APP23 transgenic mice – A model of Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* 102: 479–484.

Zoli M, Jansson A, Syková E, et al. (1999) Intercellular communication in the central nervous system. The emergence of the volume transmission concept and its relevance for neuropsychopharmacology. *Trends in Pharmacological Sciences* 20: 142–150.

Aging: Invertebrate Models of Normal Brain Aging

M Artal-Sanz, K Troulinaki, and N Tavernarakis, Foundation for Research and Technology, Heraklion, Crete, Greece

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The free-living soil nematode *Caenorhabditis elegans* has been extensively used for studying the genetic regulation of aging, in part because of its short life span and genetic homogeneity. In the nematode, neuroendocrine signaling, nutritional sensing, and mitochondrial functions have been shown to play important roles in the determination of life span. However, aging in *C. elegans* is mainly controlled by a neuroendocrine system, the DAF-2/insulin signaling pathway, which also regulates the life span of flies and mammals, indicating that this pathway is a universal longevity regulator.

The insulin signaling pathway in *C. elegans* was first genetically identified for its effects on dauer larva formation (DAF). Dauer is an alternative developmental stage induced by harsh environmental conditions such as starvation, high population density, or high temperature. Under normal conditions *C. elegans* develops to reproductive adulthood through four larval stages (L1–L4), in 3 days. However, when conditions are adverse, larvae arrest development at the second molt, to enter the dauer stage. Dauers do not feed, are resistant to stress, and can survive up to several months. Dauer larvae are considered to be nonaging because postdauer life span is not affected by the duration of the dauer stage. In addition to insulin signaling, another neuronal pathway that regulates the choice between reproductive growth and dauer entry is the DAF-7 transforming growth factor-β (TGF-β) pathway.

The DAF-2 insulin/insulin-like growth factor-1 (IGF-1) receptor pathway is required for reproductive growth and metabolism, as well as for normal life span. *C. elegans* has a single transmembrane insulin receptor kinase, DAF-2. Upon ligand binding to DAF-2, the kinase domain of the receptor phosphorylates and activates AGE-1, which is a phosphatidylinositol 3-kinase (PI3K). Activated AGE-1 PI3K generates 3-phosphoinositides (PtdIns-3,4-P2 and PtdIns-3,4,5-P3), which are second messengers required for activation of downstream kinases. Downstream kinases include pyruvate dehydrogenase kinase and serine/threonine kinases (PDK-1, AKT-1, and AKT-2), which are protein kinase B (PKB) proteins. These protein kinases regulate the forkhead (FOXO) transcription factor DAF-16, which translocates to the nucleus depending on its phosphorylation level. Phosphorylated DAF-16 remains inactive in the cytoplasm, while upon dephosphorylation it enters the nucleus and exerts its effects on transcription. Thus, the insulin signaling pathway functions to block the nuclear localization of DAF-16. An antagonist of the DAF-2/AGE-1 signaling pathway is the DAF-18 phosphatase and tensin homolog (PTEN) lipid phosphatase (**Figure 1**).

Neuronal Insulin-Like Signaling

Mutations in *daf-2*, *age-1*, or in other genes positively regulated by *daf-2* result in constitutive developmental arrest at the dauer stage. Reducing the activity of genes that antagonize insulin signaling, such as *daf-18* PTEN or *daf-16* FOXO, suppresses the dauer-constitutive phenotype of insulin signaling mutants. While severe mutations in the insulin pathway induce dauer arrest, a milder reduction of insulin signaling results in longer life span. For example, when temperature-sensitive *daf-2* and *age-1* mutants are grown at the permissive temperature until past the dauer arrest decision point, and then shifted to higher temperatures, *daf-2* and *age-1* mutants show increased life span that is dependent on DAF-16. DAF-16 is the main downstream target and major effector of DAF-2/insulin-like signaling regulating *C. elegans* life span. Signaling via the DAF-2/insulin-like receptor antagonizes the FOXO transcription factor DAF-16 by promoting its phosphorylation. Similar modulation of insulin-like signaling pathways in the fruit fly and mouse also modify life span.

The insulin/IGF-1 signaling pathway was first linked to aging in *C. elegans* when mutations in *daf-2* were found to double the life span of the worm. Subsequent investigations, aimed at identifying in which specific cells insulin signaling controls animal aging, supported a prominent role for the nervous system. In these studies, the life span of mosaic animals that had lost *daf-2* activity in different cell lineages was examined. After fertilization, the first cell division produces the AB and the P_1 cells. The AB descendants produce most of the neurons, the hypodermis, the pharynx, the excretory glands, and the vulva, while the P_1 descendants give rise to the muscles, the germ line, the somatic gonad, the hypodermal cells, a few pharyngeal cells, and a few neurons. AB mosaics, which had lost *daf-2* activity in the AB lineage but were *daf-2*(+) in the P_1 lineage, lived twice as long

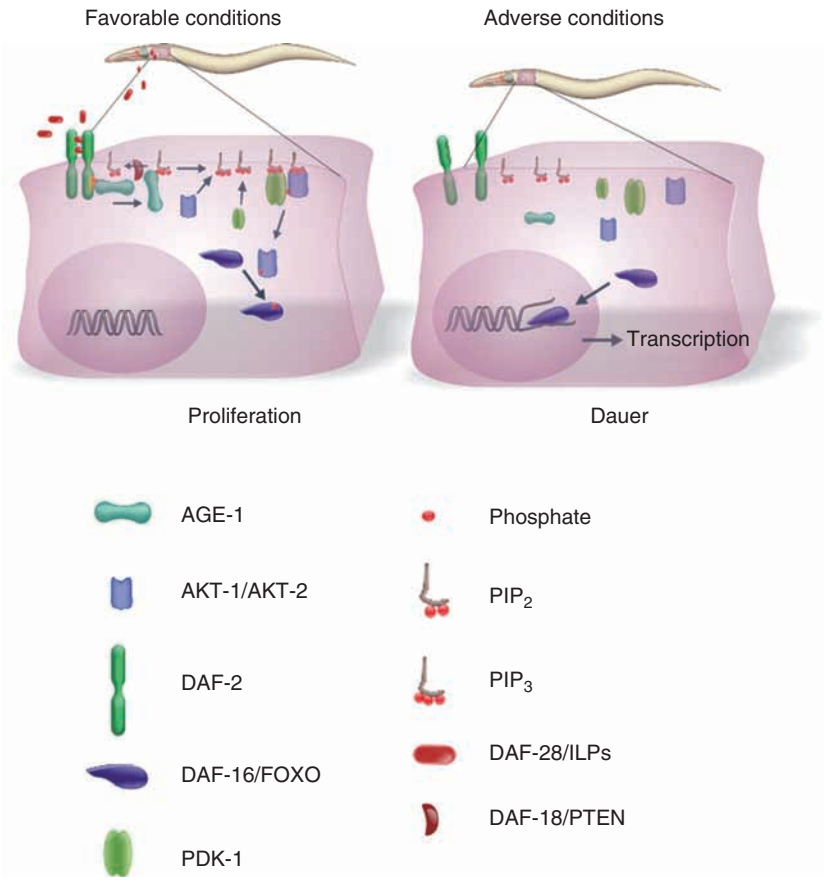

Figure 1 The *C. elegans* insulin-like signaling pathway genes regulate reproductive growth and aging. (Left): Under favorable growing conditions, insulin-like peptides (ILPs) are produced from sensory neurons, to promote the reproductive mode. Binding of ILPs to the DAF-2/insulin receptor results in the phosphorylation of AGE-1/PI3K. Activated AGE-1 generates the phosphoinositides required for the activation of downstream kinases (PDK-1 and AKT-1). This conserved signaling cascade phosphorylates the transcription factor DAF-16/FOXO, preventing its nuclear localization. Retention of DAF-16 in the cytoplasm leads to normal reproductive growth and aging. (Right): Under unfavorable conditions (e.g., crowding or starvation), insulin signaling is inhibited, resulting in the nuclear localization of DAF-16. In the nucleus, DAF-16 regulates the transcription of genes that induce dauer entry and extend life span. Protein nomenclature: AGE, advanced glycosylation end product; AKT, serine/threonine kinase; DAF, abnormal dauer formation; FOXO, forkhead transcription factors, group O; PDK, pyruvate dehydrogenase kinase; PIP, phosphoinositol phosphate; PTEN, phosphatase and tensin homolog. Illustration: Liesbeth de Jong.

as wild-type animals. Although genetic mosaic analyses of *daf-2* support the interpretation that DAF-2 signaling from the nervous system controls longevity, these experiments did not assign longevity control by *daf-2* to specific cell types. In a complementary approach, cell-type-restricted promoters were used to drive the expression of *daf-2* and *age-1* cDNAs in *daf-2* and *age-1* mutants, respectively. Transgenic expression of *daf-2* and *age-1* in neurons suppressed the life span extension phenotype of the corresponding *daf-2* and *age-1* mutants. Life span extension is not rescued when insulin signaling is restored in muscles or in intestinal cells. However, tissue-specific expression, genetic mosaic analysis, and RNA interference (RNAi) experiments indicate that *daf-16* FOXO activity in neurons accounts for not more than 20% of the longevity seen in *daf-16*(−); *daf-2*(−) double mutant animals. Instead, intestinal expression of *daf-16* is sufficient to extend the life span of these animals by 50–60%. These findings indicate that an intricate signaling network regulates aging in *C. elegans* and that neuronal insulin-like signaling controls life span by producing downstream signals that control aging of nonneuronal target tissues.

DAF-2 Insulin Receptor Function in the Nervous System

The *C. elegans* genome encodes more than 30 insulin-like ligands that might mediate input to the *daf-2* pathway through environmental cues, such as nutritional

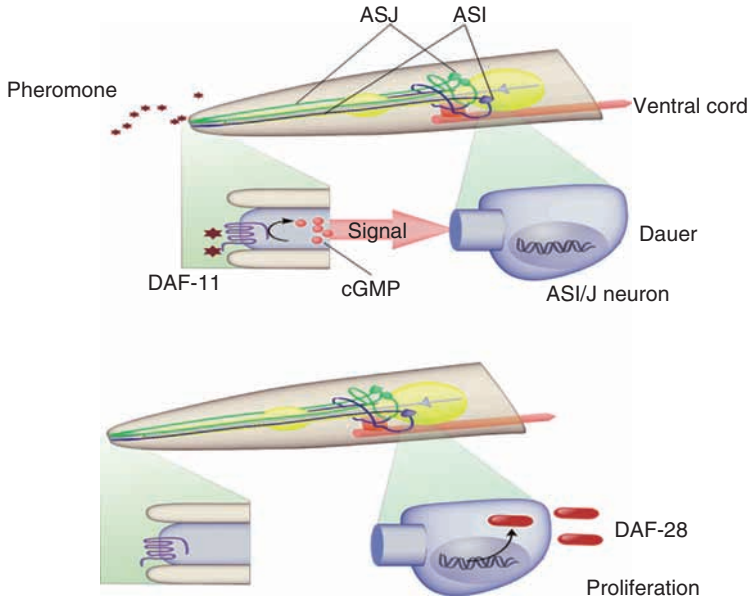

Figure 2 Sensory neurons couple environmental cues to the production of ILPs, such as DAF-28. In the ciliated sensory neurons ASJ and ASI, DAF-11/guanylyl cyclase transduces dauer pheromone signals to cyclic guanosine monophosphate (cGMP) gated ion channels to induce dauer arrest (upper panel). At low pheromone concentration, DAF-28 is produced to activate DAF-2/insulin signaling and reproductive growth (lower panel). Illustration: Liesbeth de Jong.

status or growth conditions. These insulin genes are mainly expressed in neurons, although they are also found in the intestine, epidermis, muscle, and gonad. Some of the insulin-like peptides (ILPs) have been shown to influence longevity. Thus, neuroendocrine control of aging may entail environmental cues that influence neuronal production of insulin-like ligands. For example, *daf-28* encodes an insulin-like protein, which when mutated causes dauer arrest and downregulation of DAF-2 signaling. *daf-28* is expressed in two sensory neurons (ASI and ASJ) that regulate dauer arrest. The presence of dauer pheromone is sensed by DAF-11/guanylyl cyclase, which in turn downregulates *daf-28*. Conversely, in the absence of dauer pheromone, DAF-28 is produced to induce reproductive growth (Figure 2). Although the dauer pheromone does not appear to influence aging, ILPs can act as either agonists or antagonists on DAF-2 to regulate metabolism, reproductive growth, and life span.

In support of this notion, mutations in two genes involved in Ca^{2+}-regulated secretion (*unc-64*, encoding syntaxin, and *unc-31*, encoding calcium-dependent activator protein for secretion, or CAPS), result in an increased life span that is dependent on *daf-16*. *unc-31* is expressed exclusively in neurons, and although *unc-64* is expressed in many secretory tissues, including the nervous system and the intestine, it is the function of *unc-64* in neurons that influences aging. In mammals, insulin secretion by β cells in the pancreas is a Ca^{2+}-regulated process. Therefore, a possible explanation is that the life span extension of *unc-64* and *unc-31* mutants is due to decreased secretion of a DAF-2 insulin-like ligand. Alternatively, *unc-64* and *unc-31* could regulate neurotransmitter input to other insulin-producing cells.

It has also been suggested that oxidative damage to neurons may be a primary determinant of life span. Loss of DAF-2 activity results in the activation of the FOXO transcription factor DAF-16, which controls the expression of antioxidant enzymes, such as superoxide dismutase (SOD) and catalase. Increased expression of these and possibly other free-radical-scavenging enzymes may protect neurons from oxidative damage. Thus, neuronal *daf-2* signaling might regulate animal life span by controlling the integrity of specific neurons that secrete neuroendocrine signals that might regulate the life span of target tissues. In support of this hypothesis, overexpression of Cu/Zn SOD exclusively in *Drosophila* motor neurons extends life span.

DAF-16 Targets

An important question that arises is how the insulin/IGF-1 pathway ultimately regulates aging. As noted earlier, the DAF-16/FOXO transcription factor is a downstream effector of insulin signaling and regulates a wide range of physiological responses by altering the expression of genes involved in metabolism and energy generation, as well as antimicrobial and cellular stress response genes (Table 1).

Table 1 Proteins implicated in neuroendocrine control of *C. elegans* life span, encoded by genes under control of the DAF-16/FOXO, transcription factor

Protein	*Brief description*
SOD-3	Manganese superoxide dismutase
CTL-1	Cytosolic catalase
CTL-2	Peroxisomal catalase
GST-4	Glutathione-*S*-transferase
MTL-1	Metallothionein-related cadmium-binding protein
OLD-1	Putative receptor tyrosine kinase
SCL-1	Secreted protein with sperm-coating protein (SCP) domain
LYS-7	Lysozyme
LYS-8	Lysozyme
DOD-1	Member of the cytochrome P450 family
DOD-13	Member of the cytochrome P450 family
DOD-16	Cytochrome P450, involved in the oxidation of arachidonic acid to eicosanoids
FAT-7	Involved in the biosynthesis of polyunsaturated fatty acid
DOD-11	Putative alcohol dehydrogenase
GPD-2	Glyceraldehydes-3 phosphate dehydrogenase
DAO-3	Putative tetrahydrofolate dehydrogenase/cyclohydrolase
DOD-14	Alcohol dehydrogenase
DOD-15	UDP-glucuronosyl and UDP-glucosyl transferases
GCY-6	Putative guanylyl cyclase
GCY-18	Guanylate cyclase catalytic domain
VIT-2	Vitellogenin
VIT-5	Vitellogenin
PES-2	Unknown function, putative role in ubiquitin-mediated protein degradation
PEP-2	Oligopeptide transporter
HSP16	Heat shock protein family
SIP-1	Heat shock protein
INS-7	Insulin/insulin-like growth factor-1 peptide
PNK-1	Pantothenate kinase
MRP-5	Adenosine triphosphate-binding protein, member of the subfamily C transporters

SOD is one of the most effective intracellular enzymatic antioxidants. This antioxidant enzyme catalyzes the dismutation of superoxide anions (O_2^-) to oxygen (O_2) and to the less reactive species, hydrogen peroxide (H_2O_2). *sod-3* encodes one of two manganese-containing SODs. It is localized in the mitochondrial matrix and is abundant in neural tissues. *sod-3* was one of the first recognized targets of DAF-16. *sod-3* mRNA is increased in *daf-2* mutants and undetectable in the absence of DAF-16; microarray experiments show that expression of this gene is at least tenfold higher in wild-type worms compared to DAF-16-deficient mutants. Moreover, SOD-3 is upregulated in long-lived strains, such as *age-1* and *daf-2* mutants, and in response to exogenously imposed oxidative stress in a DAF-16-dependent manner. Therefore, increased detoxification from damaging free radicals by this enzyme may contribute to life span extension.

DNA microarray analysis shows that *daf-16* affects the expression of stress response genes, the products of which directly influence aging. For example, expression of the catalase genes *ctl-1* and *ctl-2*, the glutathione-*S*-transferase gene *gst-4*, and the small heat shock protein genes increases when the activity of *daf-2* is reduced, whereas the expression of these genes decreases with the reduction of *daf-16* activity. All of these genes function to promote longevity, probably by preventing or repairing oxidative and other forms of macromolecular damage.

The gene *mtl-1*, which encodes the basally expressed form of metallothionein, is another known target of DAF-16. Metallothioneins are small cysteine-rich, metal-binding proteins protecting cells against heavy metal toxicity and reactive oxide species (ROS)-associated damage. They are induced under a variety of stress conditions, such as conditions involving metal ions, inflammation, glucocorticoids, or oxidative stress. The expression of metallothionein genes is increased in *daf-2* mutants, compared to wild-type animals. Elevated levels of these proteins as well as antioxidant enzymes are expected to decrease ROS-associated damage and hence extend life span.

An additional downstream signaling factor that is regulated by insulin/IGF-1 signaling and promotes longevity is OLD-1, a transmembrane tyrosine kinase. OLD-1 expression is increased in long-lived *daf-2* and *age-1* mutants. This increase is dependent

on DAF-16. Moreover, OLD-1 is necessary for the increased longevity of *daf-2* and *age-1* mutants and its overexpression increases life span and stress resistance. *old-1* mutations render animals more sensitive to ultraviolet light, starvation, and heat stress. These results point to a positive regulatory role for OLD-1 in life span and stress resistance.

The gene *scl-1* (which encodes sperm-coating protein (SCP)-like extracellular protein) is another target of DAF-16 that is essential for life span extension. SCL-1 has an SCP domain and is homologous to the mammalian cysteine-rich secretory protein (CRISP) family. These proteins enter the secretory pathway and are either released or anchored extracellularly by a transmembrane domain or a glycophosphatidylinositol (GPI) anchor. Expression of *scl-1* is elevated in long-lived *daf-2* and *age-1* mutants and is required for their extension of life span, since downregulation of *scl-1* reduces both life span and stress resistance of these animals. However, *scl-1* is not expressed in *daf-16* mutants. *scl-1* expression correlates with dauer morphogenesis; *scl-1* expression increases (fivefold) as worms enter dauer morphogenesis and decreases (four- to sevenfold) during dauer exit. This implicates *scl-1* in both aging and dauer formation. CRISP family proteins are involved in host defense systems in various organisms, and several functions have been postulated for SCP domain proteins, including cell adhesion, ligand function, protease inhibition, and other enzymatic activities. SCP proteins function in a variety of biological processes that involve signaling, which can be reconciled with SLP-1 being a secreted protein. However, the biochemical function of SCL-1 remains unknown.

DNA microarray analysis has also revealed a number of other *daf-2*/*daf-16*-regulated genes with substantial effects on life span. These include antimicrobial genes encoding lysozymes (*lys-7* and *lys-8*). Lysozymes are upregulated in *daf-2* mutants, and RNAi with these genes suppresses longevity of *daf-2* mutants. Consistently, *daf-2* mutants are resistant to bacterial pathogens. Other genes induced in *daf-2* and repressed in *daf-16* mutants encode proteins potentially involved in the synthesis of steroid or lipid-soluble hormones (e.g., cytochrome P450s, estradiol dehydrogenases, esterases, alcohol/short-chain dehydrogenases, and UDP-glucuronosyltransferases), and genes involved in fatty acid desaturation.

Genes with expression decreased in *daf-2* mutants and increased in *daf-16* mutants include those (*gcy-6* and *gcy-18*) encoding two receptor guanylate cyclases that are expressed in neurons. Further, RNAi with these genes prolongs life span, indicating a role for insulin signaling in sensing the environment. Other genes in this group include vitellogenin genes (*vit-2* and *vit-5*; yolk protein/apolipoprotein-like) and some genes for proteases and metabolic enzymes, such as amino- and carboxypeptidases; also included are *pes-2* (encoding a protein associated with ubiquitin-mediated protein degradation) and an amino oxidase gene, an aminoacylase gene, and an oligopeptide transporter (*pep-2*) gene. Inhibition of several of these genes results in life span extension, leading to the suggestion that the life span extension of *daf-2* mutants may involve reduced turnover of specific proteins.

Studies utilizing computational tools have also identified putative DAF-16 transcriptional targets. The genomes of both *C. elegans* and *Drosophila* were surveyed for genes with DAF-16 binding sites. Orthologous genes identified were further examined in wild-type and *daf-2* mutants animals; *pnk-1* and *mrp-5* were among several putative targets. *pnk-1* encodes a pantothenate kinase that is involved in the biosynthesis of coenzyme A, which is key to fat metabolism. This gene is upregulated in *daf-2* mutants, and RNAi inactivation results in a dramatic decrease of life span of both wild-type and *daf-2* mutant worms. *daf-2* mutants have increased fat storage, which may reflect *pnk-1* upregulation. *mrp-5* encodes an adenosine triphosphate-binding cassette C transporter. Proteins of this family modulate the secretion of insulin and participate in the transport of glutathione and nucleoside analogs. Inactivation of this gene by RNAi extends life span. It has been proposed that MRP-5 could affect aging by regulating the secretion of insulin or the transport of glutathione, an enzyme required for the antioxidant defense of the organism.

It is known that both heat shock and reduced insulin signaling trigger the nuclear localization of DAF-16, where it promotes the expression of small heat shock protein (*shsp*) genes (e.g., *hsp-16.1*, *hsp-16.49*, *hsp-12.6*, and *sip-1*). The *C. elegans* transcription factor, heat-shock factor-1 (HSF-1), is also required for the life span extension, as observed in *daf-2* mutants. HSF-1 acts in response to heat stress and reduced insulin signaling to activate the expression of *shsp* genes, together with DAF-16. Interestingly, overexpression of the gene encoding HSP70F increases longevity, similar to overexpression of the *shsp* gene *hsp-16*. Heat shock proteins are involved in reparation of misfolded or damaged proteins and are essential for recovery of cells after heat treatment. This indicates that protein misfolding and aggregation are important factors in aging.

In addition to inducing the expression of genes involved in several processes, DAF-16 can also act as a transcriptional repressor. For example, DAF-16 inhibits the expression of *ins-7*. This gene encodes an

insulin/IGF-1 peptide. Its expression is repressed in animals with reduced DAF-2 activity and is elevated in animals with reduced DAF-16 activity. RNAi for *ins-7* increases the life span of wild-type animals and the frequency of dauer formation. Thus, INS-7 behaves as a putative DAF-2 agonist. It is hypothesized that when DAF-2 is active, it inhibits the activity of DAF-16, which allows *ins-7* to be expressed. The production of INS-7 leads to further activation of DAF-2. However, when DAF-2 activity is reduced, DAF-16 is activated and inhibits the expression of *ins-7*.

Sensory Input and Neuroendocrine Signaling

C. elegans senses environmental cues through ciliated sensory neurons. Large numbers of genes required for the development and function of *C. elegans* sensory neurons have been identified. In addition to disrupting sensory neuron function, mutations in some of these genes increase life span.

Amphids are gustatory and olfactory neurons located at the *C. elegans* head. Gustatory neurons sense dauer pheromone, food, and amino acids in the environment. Olfactory neurons are responsible for sensing food-derived substances and volatile chemicals. To define which of these cells are involved in the regulation of life span, individual cells were ablated by a focused laser microbeam. Ablation of gustatory neurons revealed that only a specific subset (ASI and ASG, and not ADF, ASJ, and ASK) may influence life span. This was also the case for the olfactory neurons (AWA and AWC), since only the ablation of AWA extended life span, whereas, ablation of AWC had no effect. This was further confirmed by using mutants with specific defects in these neurons. Combined ablation of gustatory and olfactory neurons results in greater longevity, compared to ablation of either gustatory or olfactory neurons alone, suggesting that these neurons function in distinct pathways to control life span. Many mutations abrogating sensory neurons, including putative chemosensory receptors, extend the life span of *C. elegans*, largely in a *daf-16*-dependent manner. The life span extension caused by gustatory neurons depends on *daf-16*, indicating that these neurons might modulate the *daf-2* pathway. However, in the case of olfactory neurons, effects on life span are only partly dependent on *daf-16*. Many sensory neurons in *C. elegans* produce ILPs. Therefore, it is plausible that perception affects life span by influencing the activity of the insulin signaling pathway. However, these mutations show complicated interactions with the insulin/IGF-1 pathway. Double mutants did not live longer than the *daf-2* single mutant; instead their life span was shorter. This may indicate that sensory control of life span is only partially dependent on the *daf-2* pathway. Likewise, some gustatory and olfactory neurons enhance, whereas others reduce, longevity. Therefore, environmental signals that affect life span may be relayed by a *daf-2*-independent pathway.

A possible additional mechanism for influencing life span by sensory inputs is the regulation of lipid accumulation. *C. elegans* mutant strains with defects in neuroendocrine signaling show increased fat accumulation and extended life span. This is the case for *daf-2* insulin receptor mutants, the tryptophan hydroxylase *tph-1* serotonin defective mutant, and the *tub-1* (tubby ortholog) mutant. In *C. elegans*, *tub-1* is expressed in sensory neurons and *tph-1* is expressed in serotonergic neurons. In addition, a genome-wide RNAi screen identified several genes involved in food sensation and neuroendocrine signaling that, when knocked down, resulted in aberrant fat accumulation. These include genes for glutamate and dopamine receptors, as well as chemoreceptor and olfactory receptor genes. Consistently, *C. elegans* mutants with either structural or functional defects in nine specific ciliated neurons show increased fatty acid accumulation in the intestine. Interestingly, some mutations that increase lipid accumulation also lengthen life span. Thus, ciliated neurons may sense environmental cues and express neuropeptides and insulin ligands to regulate metabolism and life span. However, while the insulin signaling pathway is a major regulator of *C. elegans* fat storage (*daf-2* mutants show increased lipid accumulation), fat accumulation in these sensory mutants is independent of DAF-16/FOXO.

Other Neuroendocrine Mechanisms

In a chemical screen aimed at identifying drugs that delay aging, anticonvulsant medications were found to extend worm life span Anticonvulsants modulate neural activity in mammals and act presynaptically to modulate neuromuscular activity in the worm. Interestingly, anticonvulsants significantly increase the life span of *daf-2* loss-of-function mutants as well as that of *daf-16* mutants. This finding suggests that neural activity regulates aging by an additional mechanism independent of insulin-like signaling Similarly, these compounds further increased the life span of animals with mutations in genes important for the function of sensory neurons (*osm-3* and *tax-4*) and neurotransmission (*unc-31*, *unc-64*, and *aex-3*), highlighting the intricacy of the neuronal pathways that might be involved in the regulation of nematode aging.

The reproductive system also regulates aging in *C. elegans*. Ablation of the two germ line precursor cells, as well as mutations that reduce germ line proliferation, remarkably extend the *C. elegans* life span. This life span extension requires the presence of the somatic gonad, since ablation of both germ line and somatic gonad has no effect on life span. The life span extension of germ-line-ablated animals depends on DAF-16/FOXO and may be mediated hormonally, since it also depends on DAF-12, a nuclear hormone receptor, and DAF-9, a cytochrome P450 involved in the production of steroid hormones (3-keto sterols) that function as DAF-12 ligands. Another gene required for the increased life span associated with germ line loss is *kri-1*. The *kri-1* gene encodes a conserved protein with ankyrin repeats and is expressed in pharynx and intestine. Germ line ablation results in the nuclear localization of DAF-16 in the intestine (the worm's adipose tissue), and intestinal expression is sufficient to account for the observed longevity. KRI-1 and to a lesser extent DAF-12 and DAF-9 are required for the DAF-16 nuclear localization in the intestine of germ-line-defective animals. However, although loss of DAF-2 receptor activity promotes the nuclear localization of DAF-16 in many tissues, including the intestine, DAF-16 localization is not dependent on KRI-1, DAF-12, or DAF-9. Moreover, *kri-1* RNAi completely suppresses the life span extension of germ line mutants, while it has no significant effect on wild-type or *daf-2* mutants, indicating that modulation of life span by KRI-1 is specific to the reproductive-signaling pathway. Thus, the role of lipophilic hormones and KRI-1 on the nuclear localization of DAF-16 is germ line specific and independent of insulin signaling. Therefore, the germ line might possess a specific endocrine system to influence aging.

Nevertheless, neuronal and gonadal endocrine signaling mechanisms appear to interact in a complex manner. Somatic gonad ablation prevents the life span extension of germ-line-ablated wild-type animals, but it does not completely prevent life span extension in animals that lack olfactory neurons. Similarly, germ line ablation further extends the life span of *daf-2* mutants independently of whether or not the somatic gonad is present.

Concluding Remarks

The nervous system performs the task of sensing and integrating environmental cues into coordinated physiological responses that will ensure maximal survival and reproductive fitness. In *C. elegans*, food availability, temperature, and a secreted pheromone are some of the sensory inputs that regulate the decision of entering the metabolically active reproductive mode or shifting to the nonreproducing, nonfeeding dauer larva stage, with large amounts of stored fat.

Despite its apparent simplicity, *C. elegans* has a surprisingly sophisticated neuroendocrine system that regulates development, metabolism, and life span. Both, insulin-like and TGF-β signaling pathways act in parallel to regulate development and metabolism, with the insulin-like signaling pathway playing a major role in life span regulation. Importantly, the regulation of life span by insulin/IGF signaling is conserved across taxa, and reduction of insulin signaling has been shown to extend life span in worms, flies, and mammals. Similarly, the physiological processes involved in the aging process also appear to be conserved. For example, signals from the reproductive system also influence life span in mammals, and dietary restriction has been shown to extend life span in a wide variety of organisms. Likewise, sensory perception could also regulate life span in higher organisms, since blocking the sense of taste reduces insulin secretion in mammals, and the smell of food increases insulin levels in humans.

How physiological processes are coordinated by neuroendocrine signaling to meet the biological demands of an organism is still not completely understood. More than 30 ILPs are encoded in the *C. elegans* genome, some of which are agonists and others of which are antagonists. Some have been shown to influence aging, but many remain to be functionally characterized. It is of fundamental importance to understand which cells or tissues emit or receive signals to coordinate the aging process at the level of the whole organism. *C. elegans* has been instrumental for the discovery of conserved molecular pathways regulating aging. Its relatively short life span and its amenability for genetic and molecular analyses make it an ideal organism to pursue these studies further, aiming to ultimately understand why and how animals age.

See also: Aging of the Brain; Gene Expression in Normal Aging Brain; Insulin-Like Growth Factor Signaling and Actions in Brain; Neuroendocrine Aging: Pituitary Metabolism; Vesicular Neurotransmitter Transporters.

Further Reading

Aamodt E (2006) *The Neurobiology of Caenorhabditis elegans*. Amsterdam; Boston: Elsevier/Academic Press.

Ailion M, Inoue T, Weaver CI, et al. (1999) Neurosecretory control of aging in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7394–7397.

Alcedo J and Kenyon C (2004) Regulation of *Caenorhabditis elegans* longevity by specific gustatory and olfactory neurons. *Neuron* 41: 45–55.

Antebi A (2004) Long life: A matter of taste (and smell). *Neuron* 41: 1–3.

Boulianne GL (2001) Neuronal regulation of lifespan: Clues from flies and worms. *Mechanisms of Ageing and Development* 122: 883–894.

Braeckman BP, Houthoofd K, and Vanfleteren JR (2001) Insulin-like signaling, metabolism, stress resistance and aging in *Caenorhabditis elegans*. *Mechanisms of Ageing and Development* 122: 673–693.

Evason K, Huang C, Yamben I, et al. (2005) Anticonvulsant medications extend worm life-span. *Science* 307: 258–262.

Finch CE and Ruvkun G (2001) The genetics of aging. *Annual Review of Genomics and Human Genetics* 2: 435–462.

Hekimi S and Guarente L (2003) Genetics and the specificity of the aging process. *Science* 299: 1351–1354.

Kenyon C (2005) The plasticity of aging: Insights from long-lived mutants. *Cell* 120: 449–460.

Libina N, Berman JR, and Kenyon C (2003) Tissue-specific activities of *Caenorhabditis elegans* DAF-16 in the regulation of lifespan. *Cell* 115: 489–502.

Mak HY, Nelson LS, Basson M, et al. (2006) Polygenic control of *Caenorhabditis elegans* fat storage. *Nature Genetics* 38: 363–368.

Murphy CT, McCarroll SA, Bargmann CI, et al. (2003) Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* 424: 277–283.

Schreibman MP and Scanes CG (1989) *Development, Maturation, and Senescence of Neuroendocrine Systems: A Comparative Approach*. San Diego: Academic Press.

Tatar M, Bartke A, and Antebi A (2003) The endocrine regulation of aging by insulin-like signals. *Science* 299: 1346–1351.

Thomas JH (1999) Lifespan. The effects of sensory deprivation. *Nature* 402: 740–741.

Walker DW, McColl G, Jenkins NL, et al. (2000) Evolution of lifespan in *Caenorhabditis elegans*. *Nature* 405: 296–297.

Wolkow CA (2002) Life span: Getting the signal from the nervous system. *Trends in Neurosciences* 25: 212–216.

Wood WB (1988) *The Nematode Caenorhabditis elegans*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Yeoman MS and Faragher RG (2001) Aging and the nervous system: Insights from studies on invertebrates. *Biogerontology* 2: 85–97.

Relevant Websites

http://www.arclab.org – Aging Research Centre.

http://www.afar.org – American Federation for Aging Research.

http://www.ncoa.org – National Council on Aging.

http://www.nia.nih.gov – National Institute on Aging (U.S. National Institutes of Health).

http://sageke.sciencemag.org – Science of Aging Knowledge Environment, an interdisciplinary repository of issues related to aging (American Association for the Advancement of Science).

http://www.geron.org – The Gerontological Society of America.

Agnosia

D Tranel and N L Denburg, University of Iowa Hospitals and Clinics, Iowa City, IA, USA

© 2009 Elsevier Ltd. All rights reserved.

'Agnosia,' a neurological term of Greek origin (a + Greek *gnosis*), signifies a lack of knowledge and is virtually synonymous with an impairment of recognition. In the traditional literature, two types of agnosia were commonly described. 'Associative' agnosia referred to as failure of recognition that results from defective activation of information pertinent to a given stimulus. 'Apperceptive' agnosia referred to a disturbance of the 'integration' of otherwise normally perceived components of a stimulus.

Teuber in 1968 gave a narrower definition, in which agnosia was synonymous with having "normal percepts stripped of their meaning." In this sense, agnosia is conceptualized as a disorder of memory, and only associative agnosia qualifies for this stricter definition. In practical terms, however, it has been useful to retain the concept of apperceptive agnosia and to maintain a distinction between apperceptive and associative agnosia. In both conditions, recognition is disturbed. In the apperceptive variety, the problem can be traced, at least in part, to faulty perception, usually in reference to aspects of higher-order perceptual capacities (it is not appropriate to use the term agnosia for conditions in which perceptual problems are severe and obviously preclude the patient's apprehension of meaningful information). In associative agnosia, perception is largely intact, and the recognition defect is strictly or primarily a disorder of memory.

The difficulties of trying to separate apperceptive and associative forms of agnosia underscore the fact that the processes of perception and memory are not discrete. Rather, those processes operate on a physiological and psychological continuum, and demarcation of a clear separation point at which perceptual processes end and memory processes begin is simply not possible. Many patients with recognition defects will have elements of both conditions, that is, high-level perceptual problems and disturbances in memory. Some, however, can be classified unequivocally into one type or the other. For these reasons, the following operational definitions are appropriate. Associative agnosia is a modality-specific impairment of the ability to recognize previously known stimuli (or new stimuli for which learning would normally have occurred) that occurs in the absence of disturbances of perception, intellect, or language, and is the result of acquired cerebral damage. The designation appreciative agnosia applies when the patient meets the preceding definition in all respects except that perception is altered.

The term 'agnosia' should be restricted to situations in which recognition impairments are confined to one sensory modality, for example, vision, or audition, or touch. When recognition defects extend across two or more modalities, the appropriate designation is 'amnesia'. As noted, the term agnosia should not be used for patients in whom recognition defects develop in connection with major disturbances of basic perception. Nor should the term be applied to patients with major impairments of intellect. Finally, the term agnosia should be reserved for conditions that develop suddenly, following the onset of acquired cerebral dysfunction.

One other important distinction is between 'recognition' and 'naming.' The two capacities are often confused. It is true that recognition of an entity, under normal circumstances, is frequently indicated by naming (e.g., that is a 'groundhog' or that is 'Joe Montana'). Studies of brain-injured subjects, however, have shown clearly that recognition and naming are dissociable capacities, and the two terms should not be used interchangeably. Damage in the left inferotemporal region, for example, can render a patient incapable of naming a wide variety of stimuli, while leaving unaffected the patient's ability to recognize those stimuli. For the two preceding examples, for instance, the patient may produce the descriptions of 'that's a roly-poly animal that digs holes under barns and hibernates in the winter,' and 'that's the guy from Notre Dame who was a famous quarterback and won lots of football championships.' Both descriptions indicate unequivocal recognition of the specific entities, even if their names are never produced. In short, it is important to maintain a distinction between recognition, which can be indicated by any number of responses signifying that the patient understands the meaning of a particular stimulus, and naming, which may not, and need not, accompany accurate recognition. The patient with agnosia fails to experience familiarity with the stimulus, and is thus unable to evoke its meaning, use, or relevant relationships in both verbal and nonverbal terms.

In principle, agnosia can occur in any sensory modality, relative to any type of entity or event. In practice, however, some types of agnosia are far more frequent. 'Visual agnosia,' especially agnosia for faces ('prosopagnosia'), is the most commonly encountered form of recognition disturbance affecting a primary sensory modality. Visual agnosia is a disorder of recognition confined to the visual realm, in which a patient cannot arrive at the meaning of some or all categories of previously known nonverbal visual

stimuli, despite normal or near-normal visual perception and intact alertness, attention, intelligence, and language. Most patients manifest a comparable defect in the anterograde compartment; that is, they cannot recognize new nonverbal, visual stimuli that would normally have been learned after adequate exposure.

The condition of 'auditory agnosia' is rarer, followed by the even less frequent 'tactile agnosia.' A frequently encountered condition which also conforms to the designation of agnosia is a disturbance in the 'recognition of illness,' or what has been termed 'anosognosia'.

It is important to distinguish anosognosia from several closely related conditions. One is 'anosodiaphoria', a term that refers to the condition in which a patient acknowledges, but fails to appreciate the significance of, acquired impairments in physical or psychological function. Although anosodiaphoria is not a true form of agnosia, in practice there is a certain degree of overlap between anosodiaphoria and anosognosia. In fact, it is common to observe that blatant forms of anosognosia, for example, denial of hemiplegia, tend to evolve over time, as the patient recovers, into various degrees of anosodiaphoria. Another condition refers to a disorder of body schema. Body schema disturbances are conditions in which patients become unable to localize various parts of their bodies. The most common manifestations are 'autotopagnosia,' 'finger agnosia,' and 'right–left disorientation' (the latter two being essentially partial forms of the first). Autotopagnosia refers to a condition in which the patient loses the ability to identify parts of the body, either to verbal command or by imitation. In its most severe form, the disorder affects virtually all body parts; however, this is quite rare, and it is far more common to observe partial forms of the condition, including deficits in finger localization (finger agnosia) and right–left discrimination.

Accurate detection and diagnosis of agnosia are important on several accounts. Both visual and auditory agnosia are strongly associated with the presence of bilateral cerebral disease, and the presence of one of these conditions can be a useful clue regarding the localization of brain dysfunction. This can be especially helpful in the early stages of acquired cerebral dysfunction, when even modern neuroimaging procedures may fail to detect a lesion. Such conditions furnish additional diagnostic clues because they are typically associated with cerebrovascular disease affecting the territories of the posterior or middle cerebral arteries. Furthermore, unilateral disease involving the dominant parietal lobe has recently been implicated in both tactile agnosia and apraxia. To avoid misdiagnosis, it is important to note that the complaints or behaviors of patients with agnosia can seem so bizarre as to raise questions about their veracity. There was, indeed, a time when it was doubted whether such conditions existed at all. That agnosic conditions do occur is no longer a contentious issue; nonetheless, clinicians may be skeptical of a patient who suddenly claims an inability to recognize familiar faces, despite normal vision, or of a patient who suddenly behaves as though all auditory information had lost its meaning. A particularly unusual case of agnosia was recently reported in which a child with sleep-induced electrophysiological abnormalities involving the occipito-temporal regions and episodic seizure disorder demonstrated stable defects in visual–spatial abilities, as well as visual agnosia.

Despite their relative rarity, agnosias have also proved to be important 'experiments of nature,' and they have assisted with the investigation of the neural bases of human perception, learning, and memory. Careful study of agnosic patients over the past couple of decades, facilitated by the advent of modern structural (computed tomography, magnetic resonance) and functional (positron emission tomography, functional magnetic resonance) neuroimaging techniques, and by the development of sophisticated experimental neuropsychological procedures, has yielded important new insights into the manner in which the human brain acquires, maintains, and retrieves various types of knowledge.

See also: Amnesia: Declarative and Nondeclarative Memory; Prosopagnosia; Recognition Memory; Shape Representation in Inferotemporal Cortex.

Further Reading

Bauer RM and Demery JA (2003) Agnosia. In: Heilman KM and Valenstein E (eds.) *Clinical Neuropsychology,* 4th edn., pp. 236–295. New York: Oxford University Press.

Caselli RJ (1991) Rediscovering tactile agnosia. *Mayo Clinicical Proceedings* 66: 129–142.

Crutch SJ, Warren JD, Harding L, and Warrington EK (2005) Computation of tactile object properties requires the integrity of praxic skills. *Neuorpsychologia* 43: 1792–1800.

Damasio AR and Damasio H (1994) Cortical systems for retrieval of concrete knowledge: The convergence zone framework. In: Koch C and Davis JL (eds.) *Large-Scale Neuronal Theories of the Brain*, pp. 61–74. Cambridge, MA: MIT Press.

Damasio AR, Damasio H, Tranel D, and Brandt JP (1990) Neural regionalization of knowledge access: Preliminary evidence. *Symposia on Quantitative Biology* 55: 1039–1047.

Damasio AR, Tranel D, and Damasio H (1990) Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience* 13: 89–109.

Damasio H, Grabowski TJ, Tranel D, Hichwa RD, and Damasio AR (1996) A neural basis for lexical retrieval. *Nature* 380: 499–505.

Denburg NL and Tranel D (2003) Acalculia and disturbances of the body schema. In: Heilman KM and Valenstein E (eds.) *Clinical Neuropsychology,*, 4th edn., pp. 161–184. New York: Oxford University Press.

Eriksson K, Kylliainen A, Hirvonen K, Nieminen P, and Koivikko M (2003) Visual agnosia in a child with non-lesional occipito-temporal CSWS. *Brain & Development* 25: 262–267.

Teuber H-L (1968) Perception. In: Weiskrantz L (ed.) *Analysis of Behavioral Change*, pp. 274–328. New York: Harper & Row.

Tranel D and Damasio AR (1985) Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics. *Science* 228: 1453–1454.

Tranel D and Damasio AR (1996) The agnosias and apraxias. In: Bradley WG, Daroff RB, Fenichel GM, and Marsden CD (eds.) *Neurology in Clinical Practice,* 2nd edn., pp. 119–129. Stoneham, MA: Butterworth Publishers.

Vignolo LA (1982) Auditory agnosia. *Philosophical Transactions of the Royal Society of London* 298: 49–57.

Warrington EK and McCarthy RA (1987) Categories of knowledge: Further fractionation and an attempted integration. *Brain* 110: 1273–1296.

Agonistic and Affiliative Signals: Resolutions of Conflict

R M Seyfarth and D L Cheney, University of Pennsylvania, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Vocal Displays

Most competitive interactions among animals take the form of vocal, visual, or gestural displays: loud calls, threatening postures, or other behaviors that allow rivals to assess each other without escalating the displays into potentially costly attacks. Regardless of the species involved, displays have one essential property: they involve signals of competitive ability that are difficult, if not impossible, to fake. Male red deer, for example, challenge and assess one another by the amplitude and pitch of their roars. These acoustic features are reliable indicators of size and endurance because only large males in excellent condition can produce loud, low-pitched roars. In much the same way, songbirds challenge and assess one another by the size of their song repertoires. Males with large and complex song repertoires are often older, more experienced, or in better condition than others. Repertoire size is therefore a reliable indicator of a male's age and condition, which in turn is correlated with his competitive ability.

Because natural selection favors the skeptical observer and acts against any individuals who allow themselves to be duped by traits that are unrelated to fighting ability, the only displays that persist over time are 'honest' indicators of a male's condition. Displays will always be more common than actual fighting because, regardless of the competitive ability he brings to the table, it invariably pays a male to display first, before the fight escalates and results in injury. Avoiding injury is of paramount importance because the cost of being injured almost always outweighs the benefits to be gained from any single dispute.

Male baboons' competitive displays take the form of violent chases and loud 'wahoo' calls. Wahoos are low-pitched calls that can be produced only by adult males in good condition. They are costly to produce, not just because of their loudness and low pitch, but also because males give them in long bouts, often as they run or jump through trees. A wahoo display is therefore an exhausting demonstration of a male's stamina and coordination. High-ranking males are more likely than low-ranking males to enter wahoo contests. They also give wahoos at the highest rate and produce wahoos with longer and louder *hoo* syllables than the wahoos of other males. As males age and fall in rank, they are less likely to enter wahoo contests, and the quality of their wahoos deteriorates substantially. The predictable relation between a male's rank and the quality of his wahoos allows competitors to assess each other without actual fighting. Males of very disparate ranks seldom engage in wahoo contests, presumably because subordinate males can assess, through their rival's wahoos and behavior, that they are outmatched. By contrast, wahoo contests involving males of similar rank – whose wahoos sound more alike – occur at high rates. Their contests also involve longer calling bouts, occur for unpredictable reasons, and are more likely to escalate to physical fights and wounding.

Vocalizations as Aggressive Signals

Wahoos allow male baboons to assess each other without escalating the contest to fighting. Other primate vocalizations also function to mediate and influence aggressive interactions. Female baboons, like females in most Old World monkey species, remain in their natal groups throughout their lives and maintain close bonds with their matrilineal kin. They assume dominance ranks similar to their mothers', with the result that the female dominance hierarchy is a stable hierarchy of matrilines. During aggressive disputes against lower-ranking opponents, female monkeys often form alliances both with their close relatives and with higher-ranking individuals. In baboons, most alliances do not involve physical intervention but occur in the form of vocal support: a female utters a series of threat grunts as she observes another female threaten another. In a series of playback experiments, subordinate female baboons were played the threat grunts of their opponent's relative within minutes after they had been threatened. As one control, they heard the same individual's threat grunt in the absence of a recent fight. After receiving aggression, subjects responded strongly to the threat grunts of their opponent's relative, and in the next hour they avoided both their opponent and the relative whose threat grunt they had heard. In contrast, if they had not recently been threatened, subjects ignored the threat grunts, and they did not try to avoid the signaler or her relative. In this case, subjects appeared to regard the threat grunts as directed as someone else.

Vocal 'alliances,' therefore, appear to serve the same function as alliances involving chases and threats. Rather than physically intervening, however, female baboons signal their support by communicating their willingness to do so.

The Resolution of Ambiguity

Communicative signals can also resolve the ambiguity surrounding less aggressive interactions. For example, when a high-ranking female baboon approaches a lower-ranking one, there is always some ambiguity about what will happen next. On the one hand, the high-ranking female might be attempting to groom the lower-ranking one. On the other hand, she might be attempting to supplant or threaten the lower-ranking female from a desirable resource. This uncertainty creates a dilemma, both for high-ranking females who want to behave in a friendly manner and for low-ranking females who are reluctant to give up a resource. In the face of such uncertainty, analyses based on game theory predict the evolution of low-cost, reliable signals that provide accurate information about the signaler's disposition and motivation. In baboons and other monkeys, grunts serve this function. They are individually distinctive and highly predictive of friendly behavior, and listeners respond accordingly. If a high-ranking female baboon grunts while approaching a lower-ranking one, the lower-ranking one is likely to remain seated. High-ranking females almost never threaten a lower-ranking female after grunting to her. Conversely, if the approaching female remains silent, the lower-ranking female usually moves away.

Finally, when aggression does occur, communicative signals can mollify its effects. Nonhuman primates are frequently aggressive toward one another, yet they live in relatively stable, cohesive social groups. A number of studies have shown that opponents mitigate the effects of aggressive competition by grooming or interacting in a friendly manner soon after they fight. In baboons, such interactions are usually preceded by a grunt from the dominant aggressor to her subordinate opponent. Dominant female baboons grunt to their subordinate victims following roughly 13% of all fights. These grunts appear to serve a reconciliatory function because subordinate victims are more likely both to approach their former opponent and to tolerate her approaches if her opponent has grunted to her than if she has remained silent.

Testing the Reconciliatory Hypothesis

The hypothesis that grunts alone, in the absence of other corroborating behavior, can serve a reconciliatory function, has been tested in a playback experiment that mimicked vocal reconciliation. Within minutes of being threatened by a higher-ranking female, subjects were played their opponent's grunt. As controls, subjects heard either no grunt or the grunt of an uninvolved, high-ranking female unrelated to their opponent.

After hearing their opponent's grunt, subjects approached their opponent and tolerated her approaches – by not moving away – at significantly higher rates than they did under baseline conditions, when no aggression had occurred. These results confirmed observational results, which indicated that rates of approaching and grunting increase substantially after conflict. Apparently, a reconciliatory grunt caused subordinate females not just to relax but to seek out their former opponent.

If, however, subjects heard either no grunt or the grunt of a different dominant female, they continued to avoid their opponent and retreated from her approaches. Subjects approached their former opponent after only 2% of the trials involving either no grunt or the control female's grunt. In contrast, they did so in 42% of the trials conducted after playback of their opponent's grunt.

Subjects did not, however, simply change their disposition toward *any* female whose grunt they heard, because hearing the control female's grunt caused no change in their behavior. After playback of the control female's grunt, subjects did not approach that female at higher rates or attempt to interact with her: the change in their behavior was specific to their opponent. They appeared to regard their opponent's grunt as directed at themselves, and they acted as if they regarded the control female's grunt as irrelevant and directed at someone else.

This is not to say, however, that aggression causes baboons to attend only to their opponent's vocalizations and to ignore all other individuals' calls. Although they appear to regard the grunts of unrelated females as irrelevant to the fight, this is not true of the grunts of their opponent's relatives. To test how kinship affects reconciliation, subjects were played the 'reconciliatory' grunt of their opponent's relative soon after being threatened. A grunt from a dominant female from a different matriline served as a control. Once again, subjects responded as if they assumed that the relative's vocalization was directed at themselves and related to the recent fight. Moreover, the calls seemed to serve a reconciliatory function. After hearing the grunt, subjects did not avoid their opponent or her relative; indeed, they often approached both the relative and the opponent. And when they did so, their first interaction with their opponent was less likely to be submissive, and more likely to be friendly, if they had heard a reconciliatory grunt from their opponent's relative than if they had heard a grunt from anyone else.

The Cognitive Mechanisms Underlying Reconciliation

What are the mechanisms that underlie vocal reconciliation in baboons? In its richest interpretation, 'reconciliation' implies that the dominant female recognizes that her victim is anxious and afraid – emotions that she herself is not experiencing – and grunts to alleviate the subordinate's anxiety. The subordinate, in turn, recognizes the dominant's intent to reconcile. In contrast, an alternative hypothesis based on learned contingencies makes no assumptions about animals' ability to recognize other individuals' mental states. It argues only that dominant females grunt to subordinates in order to influence their behavior. Through experience, dominant females learn that subordinate females are less likely to move away from them when they grunt than when they remain silent. Being sensitive to contingencies, subordinate females learn that grunts are associated with friendly behavior, so they do not move away when their former opponent grunts to them.

The playback experiments just described indicated that subordinate females were more likely to approach their former opponent and to tolerate her approaches if they heard her grunt than if they did not. In contrast, hearing the grunt of a different, unrelated dominant female produced no change in their behavior. They appeared to regard that call as directed at someone else. These results are consistent with the hypothesis that baboons have a rudimentary understanding of other individuals' intentions toward themselves, an ability that constitutes a crucial precursor to the attribution of other individuals' mental states. Alternatively, a recent interaction with a particular individual might simply prime baboons to attend preferentially to that individual's vocalizations. This more parsimonious explanation does not require that baboons make inferences about the directedness of calls.

A playback experiment to examine whether baboons' responses to another female's vocalizations are influenced by the nature of prior interactions with that individual again followed a matched-pair design. In the first condition, a subordinate female was played the threat grunts of a dominant female shortly after that female had threatened her. Because females sometimes threaten their victim again soon after their original fight, listeners might interpret these threat grunts as an indicator of renewed aggression directed at them. In the second condition, the same subordinate female was played the same dominant female's threat grunts shortly after the two had groomed. Because females almost never threaten a recent grooming partner, listeners might interpret the call as directed at someone else.

If baboons' responses to threat grunts are simply the result of priming by a recent interaction, subjects' responses after being threatened should have been the same as their responses after being groomed. If, however, listeners take into account the nature of recent interactions when making inferences about the intended recipient of a call, they should interpret the two threat grunts differently – directed at themselves after aggression, but directed at someone else after grooming. Specifically, subjects might respond more strongly to threat grunts after receiving aggression than after a grooming bout. We also predicted that a subject would be less likely to approach her former opponent and more likely to retreat from her approaches after being threatened by her than after having been threatened by a different dominant female. In contrast, when the subject heard the same female's threat grunts after grooming with her, she should have been just as likely to approach and to tolerate her approaches as she was after having been groomed by a different dominant female.

Subjects responded more strongly to a dominant female's threat grunts after being threatened by her than after grooming with her. After aggression, subjects were quicker both to look toward the speaker and to move away from the area. In the ensuing 15 min, they were also less likely to come near their former opponent than they were after they had been threatened by a different female. In contrast, when subjects heard the same female's threat grunt after a grooming bout, they were just as likely to tolerate her approaches as they were after a grooming bout with a different female. Finally, subjects were significantly more likely to approach and to tolerate the approaches of the dominant female if they heard her threat grunts after grooming than after a threat.

As in the experiments testing reconciliatory grunts, subjects' responses were specific to the calls of their former opponent. Hearing their opponent's threat grunt did not affect the likelihood that subordinate subjects would approach another, uninvolved dominant female or the likelihood that they would be supplanted when approached. Taken together, therefore, these experiments suggest that female baboons make inferences about the intended target of a vocalization, even in the absence of visual cues, and that the nature of prior interactions affects subsequent behavior. After a fight, the subordinate assumes that the dominant has aggressive intentions toward her. After grooming, or after hearing a reconciliatory grunt, she makes the opposite attribution.

It seems likely that baboons make inferences about the intended target of a call whenever they hear any vocalization. For example, when subjects heard the threat grunts of their opponent's relative soon after

being threatened, they treated the grunts as a vocal 'alliance' directed at themselves. In contrast, when they heard the same threat grunts in the absence of aggression, they ignored the call and acted as if they assumed that the call was directed at someone else. Similarly, when subjects heard the 'reconciliatory' grunt of their opponent's relative after a fight, they often approached both their opponent and her relative. They did not do so, however, if they had heard the 'reconciliatory' grunt of another, unrelated dominant female. Here again, subjects behaved as if they assumed that a grunt from an opponent's relative must be directed at themselves, as a consequence of the fight. An unrelated female's grunt was irrelevant.

The ability to distinguish signals that are directed at oneself from those directed at someone else appears to be widespread in animals. For example, studies of 'eavesdropping' in birds indicate that listeners readily distinguish between songs that are directed at a third party and ones that are directed at themselves. To date, however, most of the evidence for this ability has come from studies in which individuals are interacting with only one or a few other animals and when factors such as the location of the signaler and the nature and pattern of the song provide information about the intended recipient. The challenge of inferring both the intended target of a signal and the signaler's probable behavior may be considerably more difficult in large social groups.

When deciding, Who, me? on hearing a vocalization, baboons must take into account the identity of the signaler (Who is it?), the type of call given (Was it friendly or aggressive?), the nature of their prior interactions with the signaler (Were they aggressive, friendly, or neutral?), and the correlation between past interactions and future ones (Does a recent grooming interaction lower or increase the likelihood of aggression?). Learned contingencies doubtless play a role in these assessments. Because listeners' responses depend on simultaneous consideration of all these factors, however, this learning is likely to be both complex and subtle.

Moreover, explanations based solely on learned contingencies seem unable to explain some aspects of baboons' behavior. For example, in the reconciliatory grunt experiments, subjects who heard their opponent's grunt following a fight were even more likely to approach their opponent than they were under baseline conditions, in the absence of a fight. If listeners' responses were guided solely by learned contingencies, they should have associated the call only with a low probability of aggression. Hearing the call should have returned their behavior to baseline tolerance levels, but it should not have induced them to approach their former opponent. Instead, females acted as if they interpreted their opponent's grunt as targeted specifically at them, as a directed signal of benign intent. They therefore specifically sought out their opponent.

When attending to vocal signals, female baboons appear to take into account, not only the signaler's identity and her probable subsequent behavior, but also the target of her attention. The ability to integrate these social cues simultaneously may represent a first critical step toward the recognition of other individuals' intentions and motives. In children, inferences about other individuals' intention constitute an early precursor to language learning and full mental state attribution. Do monkeys have a rudimentary understanding about other individuals' intentions toward themselves? If they do, this understanding would represent a crucial first step toward a communication system like language, in which speakers and listeners routinely assess each other's motives, beliefs, and knowledge.

See also: Animal Communication: Honesty and Deception; Communication Networks and Eavesdropping in Animals; Electrical Perception and Communication; Endocrinology of Animal Communication: Behavioral; Pheromones and other Chemical Communication in Animals; Primate Communication: Evolution; Seismic and Vibrational Signals in Animals; Visual Signaling in Animals.

Further Reading

Aureli F and de Waal F (eds.) (2000) *Natural Conflict Resolution*. Berkeley: University of California Press.

Cheney DL and Seyfarth RM (1997) Reconciliatory grunts by dominant female baboons influence victims' behaviour. *Animal Behaviour* 54: 409–418.

Engh AE, Hoffmeier RR, Cheney DL, and Seyfarth RM (2006) Who, me? Can baboons infer the target of vocalisations?. *Animal Behaviour* 71: 381–387.

Kitchen DM, Seyfarth RM, Fischer J, and Cheney DL (2003) Loud calls as an indicator of dominance in male baboons. *Papio cynocephalus ursinus*. *Behavioral Ecology & Sociobiology* 53: 374–384.

MaynardSmith J (1982) *Evolution and the Theory of Games*. Cambridge, UK: Cambridge University Press.

McGregor PK (ed.) (2005) *Animal Communication Networks*. Cambridge, UK: Cambridge University Press.

Searcy WA and Nowicki S (2005) *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton, NJ: Princeton University Press.

Silk JB, Kaldor E, and Boyd R (2000) Cheap talk when interests conflict. *Animal Behaviour* 59: 423–432.

Agraphia

S W Anderson, D Tranel, and N L Denburg,
University of Iowa Health Care, Iowa City, IA, USA

© 2009 Published by Elsevier Ltd.

Agraphia (a + Greek *graphein*) is a term used to denote an impairment of the ability to write caused by cerebral dysfunction. The term generally is reserved for instances of writing impairment due to a cerebral insult, acquired after the person has learned to write, that is, in someone who was previously literate. Writing impairments of developmental origin are usually considered part of dyslexia. Acquired impairments secondary to tremor, hemiparesis, or other basic motor defects are usually excluded from the definition of agraphia. Agraphia may involve impairment of one or multiple components of the psychomotor and linguistic aspects of writing, including grapheme formation, spelling, word selection, grammar, and spatial arrangement. Agraphia rarely occurs in isolation, and more often appears in the context of widespread language impairments, visuospatial defects, apraxia, dementia, or confusional states. The etiology may be virtually any cerebral disease, although stroke, traumatic brain injury, tumor, and Alzheimer's disease are the most frequent causes. Agraphia is often one of the early behavioral manifestations of cerebral dysfunction in acute confusional states of metabolic or toxic origin.

Much of the importance of agraphia in a clinical setting arises from the fact that collection of a writing sample is a simple, quick, and often highly informative procedure to include in a mental status examination, and generally can be obtained from even agitated or marginally cooperative patients. The task typically consists of having the patient copy a sentence, write a sentence to dictation, and spontaneously generate a sentence or two. Highly overlearned writing behaviors, such as writing one's name, may be preserved in agraphic patients and should not be used to infer normal writing ability.

The neural substrates of handwriting, a multifaceted and culturally transmitted skill which has risen to importance very late in the course of brain evolution, are far from 'hard wired.' Lesion studies have demonstrated variability across individuals in the nature and severity of agraphia stemming from damage to relevant brain areas, as well as considerable recovery of handwriting in the months following the onset of focal brain damage. The left dorsolateral frontal lobe, particularly the superior premotor region, is uniquely situated for the convergence of sequential motor activity of the dominant arm and hand together with activity of frontal and posterior language-related cortices, and damage to this region has been associated with severe and relatively circumscribed agraphia. However, agraphia also may arise from focal damage in left temporal and parietal regions, especially the perisylvian language-related cortices, as well as the left basal ganglia. Figure 1 depicts a high left parietal lobe mass which resulted in agraphia.

Patients with severe agraphia may be unable to write even single letters. Milder versions of the condition may be manifest only as more subtle defects, such as an increased frequency of spelling errors. Written output, such as spelling, has been associated with damage to a network of brain regions in the left posterior frontal lobe involving Brodmann's area 44, 45, and 6. The ability to write numbers or other nonletter symbols is often defective in parallel with the impairment in writing letters and words, but in other cases these abilities may be dissociated. For example, number writing may be preserved even if letter writing is impaired. Spatial agraphia involves an inability to maintain appropriate placement of written material on a page; this most often occurs as a progressive upward slant from the left to right side of the page and is associated with superior parietal lesions of the dominant hemisphere. Micrographia involves a progressive diminution of letter size of the course of a sentence, and is associated with parkinsonism or, less commonly, Alzheimer's disease. Figure 2 is an illustration of micrographia: the original stimulus to be copied is in the upper panel, and the patient's rendering of the stimulus is in the bottom panel. A dysexecutive agraphia, or decrease in an individual's ability to express ideas in writing, may also occur following nonfocal brain damage, such as in the case of dementia or traumatic brain injury affecting prefrontal cortex.

We offer the following brief case to illustrate the nature of the writing impairments observed in agraphia. A middle-aged man with a high school degree suffered multiple strokes resulting in damage to each of the left frontal, temporal, and parietal brain regions. Examples of his writing during the acute epoch are presented in Figure 3. Writing during the chronic epoch (i.e., 3 months following the neurological event) indicated improvement, although he continued to perform at a level well below expectations (Figure 4).

Most patients with aphasia also have agraphia which parallels their defect in spoken language, although in individual cases written or spoken production may be relatively preserved. The syndrome of alexia with agraphia involves letter distortion and

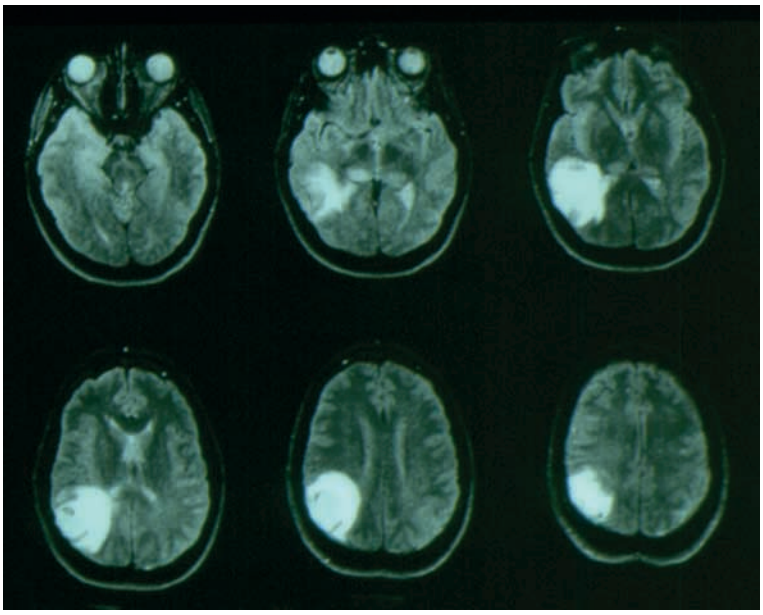

Figure 1

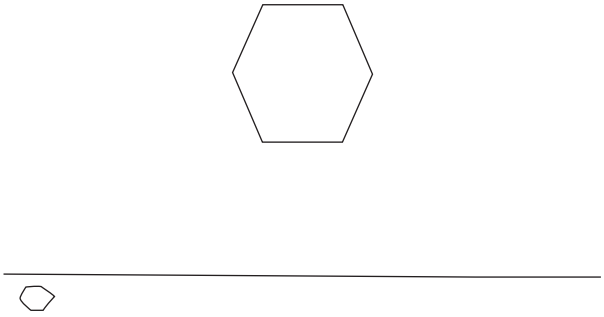

Figure 2

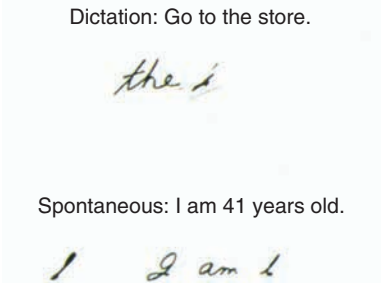

Figure 3

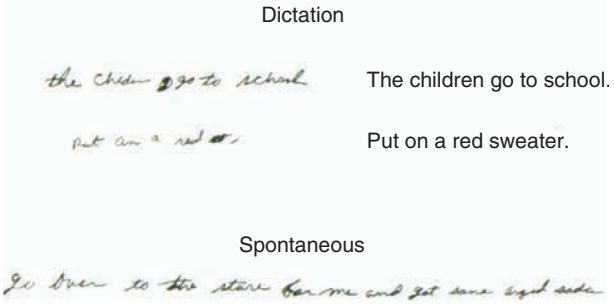

Figure 4

impaired spelling, in addition to reading impairment, and is associated with lesions in the left angular or supramarginal gyrus. In the so-called Gerstmann syndrome, loosely associated with dominant parietal lobe damage, agraphia appears in a constellation with acalculia, finger agnosia, and left–right disorientation. Callosal agraphia affecting the left hand only can follow lesions in the anterior half of the corpus callosum. Phonological agraphia involves a specific defect in phoneme-to-grapheme translation, expressed as an isolated inability to write pronounceable pseudowords, and appears with lesions in various left perisylvian regions.

With regard to treatment, a recent study indicated that rehabilitation is generally beneficial to patients with acquired agraphia, regardless of lesion location.

See also: Aging of the Brain and Alzheimer's Disease; Alexia; Dyslexia: Neurodevelopmental Basis; Language: Cortical Processes; Reading; Stroke; Stroke: Neonate vs. Adult.

Further Reading

Anderson SW, Damasio H, and Damasio AR (1990) Troubled letters but not numbers: Domain specific cognitive impairments following focal damage in frontal cortex. *Brain* 113: 749–766.

Anderson SW, Saver J, Tranel D, and Damasio H (1993) Acquired agraphia caused by focal brain damage. *Acta Psychologica* 82: 193–210.

Ardila A and Surloff C (2006) Dysexecutive agraphia: A major executive dysfunction sign. *Internal Journal of Neuroscience* 116: 653–663.

Bub D and Chertkow H (1988) Agraphia. In: Boller T and Grafman J (eds.) *Handbook of Neuropsychology,* vol. 1. New York: Elsevier.

Chedru F and Geschwind N (1972) Writing disturbances in acute confusional states. *Neuropsychologia* 10: 343–353.

Hillis AE, Chang S, Breese E, and Heidler J (2004) The crucial role of posterior frontal regions in modality specific components of the spelling process. *Neurocase* 10: 175–187.

LaBarge E, Smith DS, Dick L, and Storandt M (1992) Agraphia in dementia of the Alzheimer type. *Archives of Neurology* 49: 1151–1156.

Leischner A (1969) The agraphias. In: Vinken PJ and Bruyn GW (eds.) *Handbook of Clinical Neurology,* vol. 4. Amsterdam: North Holland.

Rapp B (2005) The relationship between treatment outcomes and the underlying cognitive deficit: Evidence from the remediation of acquired dysgraphia. *Aphasiology* 19: 994–1008.

Roeltgen DP (2003) Agraphia. In: Heilman KM and Valenstein E (eds.) *Clinical Neuropsychology,* 4th edn., pp. 126–143. New York: Oxford University Press.

Alcoholism

C L Ehlers, The Scripps Research Institute, La Jolla, CA, USA
J A Chester, Purdue University, West Lafayette, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The majority of people who choose to drink alcohol restrict their intake to amounts that result in no serious health or social consequences. National guidelines for drinking suggest that two drinks a day for men and one drink for women are amounts that do not cause harm. Harmful or binge drinking is defined as drinking the equivalent of five drinks for men and four drinks for women over a 2 h period. Binge drinking typically results in intoxication and impairment as well as blood alcohol concentrations that exceed the legal limit for driving a motor vehicle ($0.08\,g\,dl^{-1}$).

Diagnosis

The Diagnostic and Statistical Manual (DSM) of the American Psychiatric Association defines alcohol abuse as the repeated occurrence of problems associated with drinking, such as legal, social, or interpersonal problems, and the use of alcohol under hazardous conditions. 'Alcohol dependence' is a term that has been shown to be a more valid and reliable diagnosis. A diagnosis of alcohol dependence is made if a person has evidence of three of seven alcohol-related problems that occur over a 12-month or longer time period, including alcohol tolerance, alcohol withdrawal, alcohol use despite medical or psychological problems, preoccupation with obtaining, using, and recovering from drinking, a persistent desire to drink or unsuccessful efforts to cut down or control drinking, drinking larger amounts, or over a longer period than intended, and giving up or reducing important activities because of drinking. Since self-disclosure of symptoms and consequences of heavy drinking may not always be reliable, there has been a quest for valid biological markers of the disorder. Laboratory tests that can reliably index heavy drinking include blood alcohol, serum γ-glutamyltransferase (GGT), carbohydrate-deficient transferrin (CDT), and mean corpuscular volume (MCV) of erythrocytes. Possible new biomarkers include minor alcohol metabolites, such as phosphatidylethanolamine, 5-hydroxytryptophol, and possibly genetic markers, although such markers are as yet to be identified.

Epidemiology

Several national and worldwide surveys have found that the prevalence of alcohol use disorders differs based on gender and ethnicity. Alcohol abuse and dependence are twice as common in men as in women. Alcohol dependence appears to be higher in some European countries and Russia and lower in some countries in Asia. The prevalence of current alcohol abuse and dependence over a 1-year period, as estimated by US surveys, is approximately 5% and 4%, respectively. Alcohol abuse and dependence are also more common among younger people, and, in the US, among Anglo/Caucasians than in African, Asian, and Hispanic Americans. Native Americans/Alaska Natives, First Nation people of Canada, Australian Aborigines, New Zealand Maori, and some Pacific Islanders, as a group, have the highest prevalence of alcohol use disorders. It appears that both genetic and environmental factors contribute to the variance in alcohol dependence rates among men and women and between different ethnic groups.

Clinical Subtypes and Comorbidity

There has long been an effort to subtype alcoholics in order to better understand the etiology of the disorder and perhaps to design more effective treatments. A number of different subtypes have been proposed based on age of onset, severity of alcoholism, presence of severe withdrawal symptoms, comorbidity with other psychiatric disorders, and genetic versus environmental etiology. There is some general consensus that at least two general subtypes exist: (1) a type that is associated with antisocial traits, has a strong genetic component, and has an early age of onset of the disorder; and (2) a type that may occur later in life, is a milder form of the disorder, and may be more environmentally driven. Alcoholism is more common in patients with other psychiatric disorders (comorbidity). Schizophrenia, certain anxiety disorders, bipolar disorder, other drug dependencies (especially tobacco), and antisocial personality disorder have the highest co-occurrence with alcohol dependence. In some individuals alcohol can induce psychiatric illness, most commonly depression, which often resolves following abstinence. Other psychosocial consequences of drinking include increased risk

for suicide, unplanned sex, accidents, interpersonal violence, and child abuse.

Clinical Course and Treatment of Alcoholism

Like many disorders, such as obesity or diabetes, the clinical course of alcoholism progresses in an orderly sequence, with the temporal emergence of several specific alcohol-related problems. These include problems with family/friends and work, physical fights, blackouts, morning drinking, tolerance, craving, psychological impairment, withdrawal, and health problems, all of which may appear over a 5- to 20-year course. Fluctuations between periods of heavy problematic drinking, controlled drinking, and abstinence are common occurrences in alcoholics. However, it has been demonstrated that a high degree of similarity exists in the clinical course of alcoholism between men and women and among different ethnic and treatment groups.

Severe alcoholism is associated with an alcohol withdrawal syndrome. Acute withdrawal from alcohol in a chronic heavy drinker can cause tremulousness, sleep disturbance, sweating, tachycardia, anxiety and depression, headache, nausea/vomiting, and, in severe cases, hallucinations, delirium tremens, and seizures. Signs and symptoms of acute withdrawal usually subside within 2 weeks; however, acute withdrawal may be followed by a protracted abstinence syndrome consisting of hyperarousal, symptoms of anxiety/depression, and insomnia that can persist for several months and may increase risk for relapse to drinking.

In the general population, perhaps 20% or more of alcoholics resolve their drinking problems without recourse to any treatment. Current treatment modalities that have demonstrated some degree of success include alcoholics anonymous, brief interventions, motivational therapy, cognitive behavioral therapy, and medications. Few medications have been shown, as yet, to be more effective than placebo or medical management in the treatment of alcoholism. Naltrexone has been demonstrated to show efficacy in reducing heavy drinking in some studies. Developing effective medications to treat alcoholism using specific neurochemical targets is a particularly strong focus of research.

Alcoholism-Associated Toxicity

Approximately 100 000 people in the United States die each year due to alcoholism. Alcohol poisoning is most likely to occur in young drinkers; rapid consumption of excessive amounts of alcohol can cause mortality through direct depression of the medullary respiratory center. Heavy drinking and alcohol dependence are associated with toxicity in a number of different organ systems. However, the organs with the highest alcohol-associated damage are the liver and brain. Alcoholism is associated with liver cirrhosis and end-stage liver disease, though only about 10% of alcoholics develop cirrhosis. It appears that while alcohol exposure is necessary to produce liver cirrhosis, it is not sufficient, and other genetic and environmental cofactors are necessary to induce the disorder. Alcoholic liver disease occurs in a continuum of severity that begins with fat infiltration that can progress to an inflammatory process involving cytokines (tumor necrosis factor-α) and activation of liver macrophagic-like cells (e.g., Kuppfer cells), and in its most severe form results in collagen synthesis by stellate cells during end-stage disease (cirrhosis). Alcohol can also damage the nervous system both directly and in association with vitamin deficiencies. Magnetic resonance image (MRI) scans of alcoholics in early recovery can demonstrate enlarged ventricles, enlarged sulci, and smaller than normal gyri. These findings can reverse to a certain extent over several months to years with continued abstinence. Neuropsychological studies have shown that alcoholic adults have poorer functioning on tests of memory, executive functioning, and visuospatial functioning. A severe form of alcohol neurotoxicity is Wernicke–Korsakoff syndrome (WK-S). WK-S involves thiamine deficiency and consists of ocular disturbances, ataxia, confusion, and short-term memory loss and is associated with necrotic lesions of the mammillary bodies, thalamus, and other brain stem areas. Other neurological and medical complications of chronic alcohol exposure include peripheral neuropathy associated with multiple vitamin B deficiencies, anterior lobe cerebellar degenerative disease, cardiomyopathy, esophageal varicosities, gastritis, pancreatitis, and myopathy.

Alcoholism and Neural Development

Alcohol may be more toxic during critical neurodevelopmental epochs. It also appears that adolescent brain may be particularly vulnerable to the toxic effects of alcohol caused by heavy bouts of binge drinking, although data in this area are just emerging. Exposure to alcohol *in utero* can result in a spectrum of deleterious effects that appear to exist along a continuum, ranging from subtle behavioral and cognitive dysfunctions to gross neuromorphological anomalies. The most severe result of *in utero* alcohol exposure is fetal alcohol syndrome (FAS). FAS is identified by characteristic facial anomalies, growth retardation, and central nervous system dysmorphology

and behavioral abnormalities. Although heavy drinking is associated with FAS, not all women who drink excessively during pregnancy give birth to offspring with FAS. Both genetic and environmental factors, including nutritional status and the use of other drugs, contribute to the vulnerability to FAS.

Genetic and Environmental Influences on Alcoholism

Twin, family, and adoption studies have all provided evidence that perhaps half of the risk for alcoholism is genetically based. Studies of the heritability of alcohol drinking in twins suggest that the initiation of drinking may be more influenced by environmental factors, whereas development of dependence may have a larger genetic component. There are a number of environmental factors that can influence drinking and alcohol dependence. Efforts to prevent the development of alcohol-related morbidity and mortality often target environmental causes of the disorder. Reducing the availability of alcohol (e.g., by increasing cost or restricting sales) can reduce drinking. Social attitudes and legislation that stigmatize use of a substance can also influence consumption. Increasing the legal minimum age of drinking and lowering the legal blood alcohol concentration for driving also represent successful efforts that have reduced drinking. Restriction of underage drinking is particularly important; it has been demonstrated in several population samples that the earlier the age a person begins to drink, the greater the risk of developing alcohol dependence in their lifetime.

Alcoholism is not a single gene disorder but is rather polygenic, with vulnerability most likely arising from the simultaneous impact of functional variations at several genes. The genes that encode the structures of the enzymes that metabolize alcohol are the best characterized genetic factors that influence alcoholism risk. Alcohol is metabolized primarily in the liver, but also in the stomach and brain by several different enzyme systems. The two main enzymes involved in alcohol metabolism are alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH). ADH breaks down alcohol to the toxic metabolite acetaldehyde, which is then broken down further by ALDH into the less toxic acetate. The genes that code for these enzymes are polymorphic and thus several different forms of the enzymes exist, each differing in their kinetic properties. This means that genetically determined differences in alcohol-metabolizing enzyme structures can cause alcohol to be broken down slower or faster, and acetaldehyde to either be quickly eliminated or to build up and thus increase risk for organ damage. One of the most common variants, a mutation in ALDH that is seen in approximately 40% of Asians, results in facial flushing, nausea, tachycardia, and headache following consumption of alcohol. Individuals with this mutation are at much less risk for developing alcohol dependence. Variants found in ADH, depending on ethnic origin (e.g., ADH1B in individuals of African ancestry), can modify response to alcohol and risk for alcohol dependence as well as alcohol-related organ damage.

A number of other genetic approaches have been taken in order to search for the etiology of alcohol use disorders and other alcohol risk phenotypes (the physical manifestation of a trait that is heritable). Several US research groups have identified alcohol-affected families and conducted genome scans, and evidence for linkage of alcoholism to a number of gene locations has been identified and replicated by several studies. The strongest evidence appears to point toward genes on chromosomes 1, 2, 4, 5, 6, 7, 9, 15, and 16. Another approach to studying the genetics of alcoholism has been to determine if there is an association between alcohol dependence and a polymorphism in a single identified gene that is suspected of contributing to the variance in the etiology of the disorder. These so-called candidate gene approaches have yielded preliminary evidence for a number of neuronal targets, including the γ-aminobutyric acid (GABA) system, dopamine, norepinephrine, cannabinoids, and opiates, as contributing to risk for alcohol dependence, although the studies have not always been robustly replicated. Future efforts in this area will employ the new technology of whole-genome association analyses, which will be a more powerful technique to identify multiple genes simultaneously over the entire genome.

The idea that some phenotypes, called endophenotypes, bear a close relationship to the biological processes that give rise to alcohol dependence has been given much attention in alcoholism research. A number of endophenotypes have been identified as being closely associated with alcoholism, including variants of the electroencephalogram, components of event-related potentials, alcoholism symptom clusters, personality traits, levels of alcohol craving, withdrawal, and sensitivity and tolerance to alcohol. Perhaps one of the most compelling endophenotypes is described as 'low level of response to alcohol.' Individuals with a low level of response to alcohol have a diminution of both objective and subjective responses to alcohol when they are given a standard number of drinks. This allows them to drink larger amounts of alcohol, on any one occasion, before feeling drunk. Prospective studies have demonstrated that individuals with the low level of response are

significantly more likely to develop alcohol-related life problems than are matched individuals with higher levels of response to alcohol.

Alcohol Effects on the Central Nervous System

An initial hypothesis on how alcohol induces its intoxicating and addicting effects is that alcohol acts to fluidize membranes; however, these effects have been shown to be very small at pharmacologically relevant concentrations. Although no specific receptor for alcohol has been discovered, alcohol can bind to proteins in the cell membrane and alter receptor functions and intracellular signaling pathways that are coupled to membrane receptors. In the brain, alcohol decreases neuronal excitability by enhancing the inhibitory effects of the neurotransmitter GABA and by decreasing the excitatory effects of the neurotransmitter glutamate. Alcohol also has significant interactions with other primary brain neurotransmitters, including dopamine and serotonin, as well as with neuropeptide transmitters such as opioids, corticotropin-releasing factor/hormone (CRF/CRH), and neuropeptide Y (NPY) and endocannabinoids. The reinforcing effects of alcohol are thought to be due in large part to its ability to increase the activity of dopamine-containing neurons in the primary reward circuit in the brain, known as the mesolimbic dopamine pathway. Dopamine, in turn, can influence the function of serotonin to produce feelings of well-being and alterations in mood states. Alcohol also increases the release of endogenous opioid peptides, which reduce pain and produce euphoria. Long-term exposure to alcohol can result in relatively permanent changes in neuronal function resulting from alterations in gene expression and intracellular protein levels; this is termed neuroadaptation to alcohol. Neuroadaptation to alcohol is considered one of the primary brain mechanisms of addiction.

Preclinical Models of Alcoholism

A number of valid preclinical models have been developed to study particular characteristics of human alcoholism, including excessive alcohol intake, alcohol craving and relapse behavior, tolerance to alcohol, alcohol toxicity, and withdrawal. Many of these models are studied in combination with genetic strategies to characterize the role of genes, and their interaction with the environment, in shaping in animals those alcohol-related traits that may be relevant to alcoholism in humans.

Genetic Animal Models

Genetic animal models have provided important information about how certain alcohol-related traits, as well as other biological and behavioral traits (e.g., levels of dopamine in the brain, impulsivity, anxiety), may share a common genetic basis. For example, it has been clearly demonstrated that rodents with a genetic propensity toward low alcohol consumption display more pronounced signs of alcohol withdrawal. In addition, genetic propensity toward high alcohol consumption is associated with reduced sensitivity to alcohol's aversive effects. Many rodent lines have been genetically selected to show extreme differences on a particular alcohol-related trait, including voluntary alcohol consumption, the severity of alcohol withdrawal, the hypnotic effects of alcohol, the hypothermic effects of alcohol, and the locomotor-stimulant effects of alcohol. These lines show various neurobiological differences, including brain levels of serotonin, GABA, dopamine, opioids, CRF, and NPY that are associated with high or low alcohol-related behavioral responses for which they were selected. Many genetic animal models have been used to successfully identify single genes or groups of genes that are associated with individual differences in alcohol-related behaviors. One commonly used approach is a gene mapping strategy to determine if a particular chromosomal location containing a gene sequence or cluster of genes is associated with a particular alcohol-related behavior. These gene sequences can then be compared to homologous genes in the human genome. A number of candidate chromosomal locations and genes sequences have been identified, and some of the strongest data indicate that genes that regulate the GABA neurotransmitter system are associated with sensitivity to alcohol withdrawal responses. Converging evidence for a link between alcohol dependence and certain GABA gene variants has also been found in human genetic association and linkage studies.

To date, approximately 100 candidate genes have been studied for their role in regulating alcohol-related traits in preclinical models. For example, mice without functional dopamine D4 receptors (i.e., D4 knockout mice) show enhanced sensitivity to alcohol's locomotor stimulant effects, as well as to cocaine and methamphetamine. Mice with knockouts of the NPY receptor gene show decreased sensitivity to alcohol's effects and increased alcohol consumption. Mice that have a transgene that produces overexpression of CRF are more sensitive to the sedative effects of alcohol and drink less alcohol. Gene array technology has also been used to identify additional candidate genes, as well as to confirm some previously

identified genes, that may be associated with risk for alcohol dependence.

Models of Alcohol Interactions with Stress and Hormones

Stress is one environmental factor that has long been thought to increase the risk for drinking in vulnerable individuals. Findings in animal models suggest that the effects of stress on alcohol drinking are complex and depend on many factors, such as the type and duration of stress exposure, the genetic makeup and sex of the animal, the age of first stress exposure, and whether the animal has previously been exposed to alcohol. Female rodents may be protected against the effects of stress under certain conditions. Male rodents with a genetic propensity toward high alcohol intake appear to be particularly vulnerable to stress-induced alcohol consumption compared to their alcohol-avoiding counterparts. These findings may be partly explained by genetic differences in the level of hypothalamic–pituitary–adrenal (HPA) axis hormones, such as CRH, NPY, and opioid peptides; such differences have been found in animals with innate preference or avoidance of alcohol. Primary HPA axis hormones released in response to stress also appear to increase the rewarding effects of alcohol in certain experimental paradigms. Mice with genetic disruption of the CRH1 receptor show stress-induced alcohol drinking, but only after repeated exposures to stress and prolonged alcohol consumption. This same CRH receptor subtype is associated with stress-induced relapse to alcohol-seeking behavior in rats genetically selected for high alcohol preference. Early exposure to stress may increase subsequent alcohol consumption in adult animals, and these effects appear to depend on interactions with gene variants. For example, the presence of certain genes that regulate serotonin function is associated with increased alcohol preference in female monkeys that were exposed to early-life stress. Animals exposed to alcohol during early periods of development show greater stress-induced increases in alcohol drinking than do animals not exposed to alcohol during development. These results parallel findings in humans that indicate greater stress-reactive drinking and a fivefold increased risk for adult alcoholism in people who began drinking alcohol before the age of 15 years.

Models of Alcohol-Induced Toxicity

Preclinical research using cell culture and animal models, in which variables can be studied under controlled conditions, has led to important knowledge of the mechanisms that contribute to alcohol-induced neuro- and hepatotoxicity. Alcohol damages the liver through a number of biological pathways that produce oxidative stress, disruptions in energy homeostasis, protein breakdown, inflammation, and collagen formation. Investigators have clearly shown the influence of nutritional factors in studies in which animals were fed alcohol directly into the stomach (intragastric feeding models). Animals fed alcohol-containing diets that were low in folate, protein, and carbohydrates, or high in polyunsaturated fats, had more pronounced liver damage as evidenced by greater activity of cytochrome P450 enzymes (CYP2E1 activity), scar tissue formation, and cell death. Conversely, supplementing the diets of alcohol-fed animals with folate or a phospholipid antioxidant has been shown to reduce CYP2E1 activity and associated signs of cell death and fatty cell accumulation. Cultured stellate cells from the livers of alcohol-exposed animals treated with a phospholipid antioxidant showed reduced collagen production, a process that leads to liver scarring. The role of various genes that may increase or protect against liver damage has been an important research area. For example, transgenic mice with overexpression of the CYP2E1 gene show greater alcohol-induced liver injury compared to wild-type controls. Several knockout mouse models in which particular genes have been inactivated (e.g., those encoding tumor necrosis factor and endotoxin receptors) are protected against alcohol-induced liver damage.

Preclinical studies have also provided models of alcohol-induced neurotoxicity. Oral self-administered alcohol over long periods has been shown to reduce synapses in cerebellum, although the effects are not robust. Early studies in cultured cortical neurons demonstrated that exposure to and withdrawal of high doses of alcohol could lead to glutamate-associated neuronal excitotoxicity. *In vivo* studies using a binge alcohol administration model have further shown that high-dose alcohol is selectively toxic and causes neurodegeneration in certain brain regions, such as entorhinal and perirhinal cortical areas in the rat. Binge alcohol treatment also increases neuronal death and reduces neurogenesis in the hippocampal dentate gyrus. The neurodegeneration seen in these studies does not appear to be directly related to glutamate neurotoxicity. Rather, it may involve mechanisms of oxidative stress, inflammatory cytokines, activation of nuclear factor-κB, and reductions in cAMP response element-binding protein (a transcription factor important for inducing key genes that improve resistance to insults, neuronal vitality, and growth). Another model of alcohol-induced brain damage using *in vivo* studies in the rat and organotypic brain slice cultures has

indicated that neuronal degeneration is caused in part by traumalike neuroinflammatory processes involving brain edema and the release of proinflammatory signaling molecules such as tumor necrosis factor-α and arachidonic acid. Studies have also indicated that alcohol can inhibit new cell formation, although how neural degeneration relates to loss of regeneration is not as yet clear. Genetic factors may also influence the degrees of alcohol-induced toxicity, as rats from a selectively bred alcohol-preferring (P) line have a two- to threefold more brain damage in perirhinal and entorhinal cortices than do rats from a nonpreferring (NP) line, after exposure to binge alcohol treatment. Additionally, it has been demonstrated that alcohol-induced neurotoxicity is more extensive following adolescent and *in utero* exposure.

Models of Alcohol Therapeutics

Preclinical research using molecular and cellular techniques and whole-animal models has led to the identification of several biological targets in the brain for drugs to treat alcoholism. In 1995, naltrexone, a drug that blocks opioid receptors, was one of the first drugs approved in the US to treat alcoholism. Many animal studies have shown that naltrexone effectively reduces alcohol drinking and alcohol relapse-type behaviors. Acamprosate, a drug approved in 2004, has been shown to reduce alcohol drinking and signs of craving and withdrawal in animals. Acamprosate appears to work by interacting with the glutamate neurotransmitter system. A promising line of research suggests that targeting neuropeptide transmitter systems such as CRF and NPY may counteract many of the neuroadaptive responses seen after long-term alcohol exposure and may be particularly effective in reducing alcohol craving and relapse-drinking behavior. Many other drugs that target GABA, glutamate, serotonin, dopamine, and endocannabinoid neurotransmitter systems are being intensively studied in animal models of alcoholism to determine their potential for reducing the short- and long-term effects of alcohol and for treating alcohol dependence. Recent data suggest that one reason that only a certain proportion of people respond to medication therapy is in part due to the presence or absence of certain gene variants that regulate responses to the drug. Preclinical research in genetic animal models will advance the current research efforts aimed toward developing individualized drug therapy for alcoholism based on genetic background.

See also: Brain Damage: Functional Reorganization; Dementia; Fetal Alcohol Syndrome; GABA-A Receptor Synaptic Functions; Gamma-Aminobutyric Acid (GABA); Hepatic Encephalopathy.

Further Reading

Adachi M and Brenner DA (2005) Clinical syndromes of alcoholic liver disease. *Digestive Diseases* 23: 255–263.

Chester JA and Cunningham CL (2002) GABA(A) receptor modulation of the rewarding and aversive effects of ethanol. *Alcohol* 26: 131–143.

Crabbe JC, Phillips TJ, Harris RA, et al. (2006) Alcohol-related genes: Contributions from studies with genetically engineered mice. *Addiction Biology* 11: 195–269.

Crews FT, Collins MA, Dlugos C, et al. (2004) Alcohol-induced neurodegeneration: When, where and why? *Alcoholism: Clinical and Experimental Research* 28: 350–364.

Ehlers CL, Gilder DA, Wall TL, et al. (2004) Genomic screen for loci associated with alcohol dependence in Mission Indians. *American Journal of Human Genetics* 129B: 110–115.

Goldman D, Oroszi G, and Ducci F (2005) The genetics of addictions: Uncovering the genes. *Nature Reviews Genetics* 6: 521–532.

Grant BF, Dawson DA, Stinson FS, et al. (2004) The 12-month prevalence and trends in DSM-IV alcohol abuse and dependence: United States, 1991–1992 and 2001–2002. *Drug and Alcohol Dependence* 74: 223–234.

Heilig M and Egli M (2006) Pharmacological treatment of alcohol dependence: Target symptoms and target mechanisms. *Pharmacology and Therapeutics* 111: 855–876.

Hesselbrock VM and Hesselbrock MN (2006) Are there empirically supported and clinically useful subtypes of alcohol dependence? *Addiction* 101(supplement 1): 97–103.

Lieber CS (2005) Metabolism of alcohol. *Clinics in Liver Disease* 9: 1–35.

McBride WJ and Li TK (1998) Animal models of alcoholism: Neurobiology of high alcohol-drinking behavior in rodents. *Critical Reviews in Neurobiology* 12: 339–369.

Niemela O (2007) Biomarkers in alcoholism. *Clinica Chimica Acta* 377: 39–49.

Room R, Babor T, and Rehm J (2005) Alcohol and public health. *Lancet* 365: 519–530.

Schuckit MA (2006) *Drug and Alcohol Abuse: A Clinical Guide to Diagnosis and Treatment*. New York: Springer.

Sullivan EV and Pfefferbaum A (2005) Neurocircuitry in alcoholism: A substrate of disruption and repair. *Psychopharmacology (Berlin)* 180: 583–594.

Relevant Website

http://www.niaaa.nih.gov– National Institute on Alcohol Abuse and Alcoholism (National Institutes of Health).

Alexia

D Tranel and N L Denburg, University of Iowa Hospitals and Clinics, Iowa City, IA, USA

© 2009 Elsevier Ltd. All rights reserved.

Alexia is a neurological term of Greek origin (a + Greek *lex(is)*) that designates a partial or complete inability to read. There are a number of different subtypes of alexia, but all have in common the feature that the affected patient cannot read normally, so that reading is slow or impossible, and comprehension of read material is impaired. In the American literature, alexia is used nearly exclusively to designate acquired defects in reading, that is, reading impairments that occur as the result of a neurological condition in an individual who was previously literate. The British literature tends to use the term dyslexia to refer to this condition. Dyslexia in the American literature is typically applied to persons who have a developmental form of reading disorder – specifically, the not uncommon condition of being unable to learn to read normally, which becomes manifest during childhood and typically affects the acquisition of both reading and writing skills.

Stroke is the most common cause of acquired alexia, especially when the condition appears as an isolated or relatively isolated symptom. Other forms of neurological disease, including cerebral tumors, inflammatory processes, and head injury, can also cause alexia, and there are some cases of degenerative disease (e.g., Alzheimer's) in which a reading disturbance is a prominent early feature of the patient's cognitive dysfunction. Acquired alexia occurs in two main forms. One, termed alexia with agraphia, involves a reading impairment that is accompanied by writing defects; this form is common in patients with various aphasia syndromes. The other, termed alexia without agraphia (or 'pure alexia'), involves an isolated defect in reading that is not accompanied by impaired writing. Pure alexia is associated with lesions that disconnect both visual association cortices from the dominant, language-related temporoparietal cortices. Pure alexia can be caused by a single lesion strategically placed in the region behind, beneath, and under the occipital horn of the left lateral ventricle, by damaging pathways en route from the callosum and pathways en route from the left visual association cortex (Figure 1). Another setting is the combination of a lesion in the corpus callosum, which disconnects right-to-left visual information transfer, and a lesion in the left occipital lobe, which disconnects left visual association cortex from left language cortex (Figure 2). Such lesions are likely to produce a right hemianopia; this sign is a frequent, although not invariable, accompaniment of pure alexia. Another neuropsychological correlate of pure alexia is color anomia, that is, the inability to name colors.

The 'purity' of alexia without agraphia stems from the fact that patients with these lesions do not develop disturbances in writing or in other aspects of speech and linguistic functioning. This separates this type of alexia from the types of reading defects that are common in aphasic patients. (It is rather striking to observe patients who can write a sentence easily and then be unable to read what they have just written.) In this sense, pure alexia can be construed as a disturbance of visual pattern recognition. Patients with pure alexia are unable to read most words and sentences, and in severe cases cannot read even single letters. The problem is not one of visual acuity: the fact that the patients can see the sentences, words, and letters they cannot read can be readily demonstrated by having the patients copy those stimuli, a task that will be executed normally. Thus, most patients with pure alexia have normal visual acuity (although a quadrantanopia or hemianopia may be present), and most have normal recognition of nonverbal visual stimuli such as objects and faces.

In terms of treatment for pure alexia, there are two schools of thought. The first is to attempt to increase the speed and accuracy of letter-by-letter reading. The second, higher-level approach involves an abandoning of letter-by-letter reading for the adoption of whole word recognition, a more efficient approach.

As noted, the nonacquired, developmental varieties of reading disorder are subsumed under the term of developmental dyslexia. These reading impairments, almost invariably accompanied by writing weaknesses (especially spelling defects), constitute a learning disability which in fact is the most common form of learning disability in the general population. Some developmental dyslexias may be related to a disturbance of the cellular organization of language cortices, possibly originating *in utero*. A few older postmortem studies have suggested that abnormalities of cortical organization in the brains of dyslexic patients may have contributed to the reading disability. There is a strong association between developmental dyslexia and left-handedness. The older literature indicated that dyslexia was far more common in boys than girls, but recent and better-controlled studies have not supported this notion; the ratio may even be fairly close to 1:1.

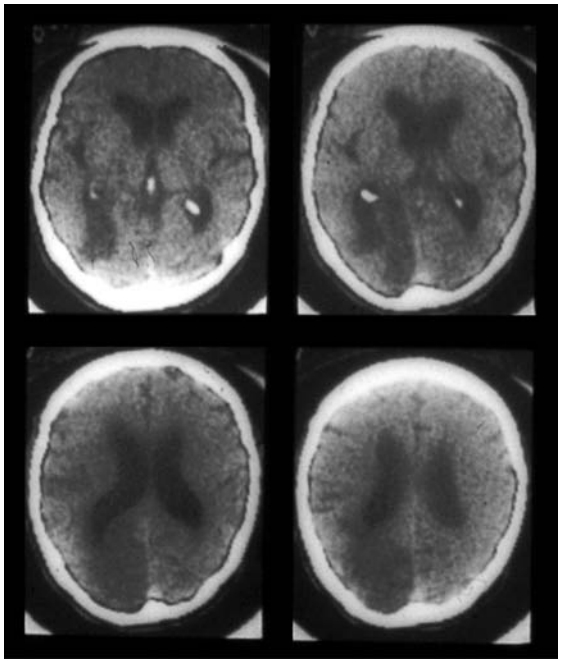

Figure 1 Illustration of the neuroanatomic findings of a patient with pure alexia, arising from a single lesion.

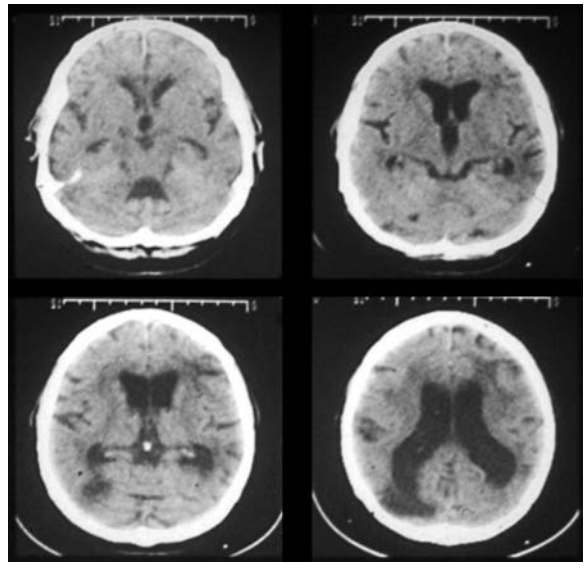

Figure 2 Illustration of the neuroanatomic findings of a patient with pure alexia, arising from a multiple lesion.

See also: Aging of the Brain and Alzheimer's Disease; Agraphia; Cerebrovascular Disease; Dementia; Dyslexia: Neurodevelopmental Basis; Reading; Stroke; Stroke: Neonate vs. Adult.

Further Reading

Benson DF, Brown J, and Tomlinson EB (1971) Varieties of alexia. *Neurology* 21: 951–957.

Coltheart M, Patterson K, and Marshall JC (1980) *Deep Dyslexia*. London: Routledge and Kegan Paul.

Geschwind N (1965) Disconnection syndromes in animals and man. *Brain* 88: 237–294, 585-644.

Greenblatt SH (1973) Alexia without agraphia or hemianopia: Anatomical analysis of an autopsied case. *Brain* 96: 307–316.

Petersen SE, Fox PT, Snyder AZ, and Raichle ME (1990) Activation of extrastriate and frontal cortical areas by visual words and word-like stimuli. *Science* 249: 1041–1044.

Tranel D (1994) Assessment of higher-order visual function. *Current Opinion in Ophthalmology* 5: 29–37.

Tranel D (1996) Disorders of color processing (perception, imagery, recognition, and naming). In: Feinberg T and Farah MJ (eds.) *Behavioral Neurology and Neuropsychology*, pp. 257–266. New York: McGraw-Hill.

Allometric Analysis of Brain Size

H J Jerison, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 H J Jerison. Published by Elsevier Ltd. All rights reserved.

Fossil 'brains' are fossils, but they are not brains. They are casts molded by the cranial cavity of fossils, but such endocasts can be made from any intact skull of a fossil or living species. Avian and mammalian endocasts look enough like brains to be treated as undissected brains. The fossil evidence is from the analysis of the external morphology of the brain as determined from endocasts. A few fossil endocasts are shown in Figure 1.

Analysis of brain evolution begins with measures of brain size. The size of the brain as a whole is a remarkably good estimator for the sizes of many of its parts, as indicated by Figure 2. These graphs show the relationship between the size of the whole brain and the sizes of the cerebellum and of the basal ganglia in living mammals. In a principal components factor analysis, gross brain size accounted for more than 85% of the variance in the size of major structures of the brain. In the sample of mammals for Figure 2, more than 99% of the variance in cerebellum and basal ganglia was accounted for by brain size. This encourages one to use gross brain size in the analysis of brain evolution.

In vertebrates other than birds and mammals, the brain does not fill the cranial cavity, and the evidence is more difficult to analyze. It is conventional to assume that in dinosaurs, as in living reptiles, the brain was approximately half the volume of the endocast. In fish and amphibians, it can be a much smaller fraction. The brain of *Latimeria*, the 'living fossil' fish from the coastal waters of southeast Africa and the Indian Ocean, is suspended in a sack of meninges and occupies only approximately 1% of the volume of the cranial cavity. In the megamouth shark, it is approximately 20% of the volume of the endocast. Precise quantitative evolutionary analysis of the fossil evidence of brain size is necessarily limited to birds and mammals, but it can be performed fairly reliably in reptiles and dinosaurs.

Quantitative Analysis: Convex Polygons

A quantitative analysis of the evolution of the brain is presented in Figure 3. Minimum convex polygons are drawn about much of the available data on living amniotes (mammals, birds, and reptiles). If one imagines the data on individuals in each of these classes of vertebrates as pins on a logarithmically scaled checkerboard, each polygon can be thought of as a string drawn to include all of the pins for a given class of vertebrates. The graph is based mainly on the sample described by Quiring in 1950, but other sources contributed. The polygons for living amniotes can be thought of as maps of brain–body space that indicate the present diversity of adaptations of brain size.

Quantitative analysis with convex polygons is equivalent to regression analysis in which lines are fitted to the clouds of points for each class. The polygons provide a realistic picture of the diversity of brain sizes. Given the logarithmic scaling for Figure 3, most of the within-species variability could be represented by the diameters of data points such as those representing archaic mammal species.

One often encounters a measure of encephalization, the 'encephalization quotient' (EQ), which is a quick way to characterize the relationship among species in relative brain size. To compute this measure, one must calculate the regression of the logarithms of brain size on body size, which is the allometric relationship: $\log Y = 2/3 \log X + \log 0.12$). EQ for a species is the residual from that regression as computed for that species. (Recall that a residual is the difference between a measured Y value and the Y value of the corresponding coordinate (X, Y) on the regression line.) Because the analysis is with logarithmic measures, the difference between the two logarithmic values is their quotient in linear measures. Therefore, EQ is the ratio of measured brain size to expected brain size, where expected brain size for an animal of given body size is determined by the allometric brain–body equation. Median EQ for these archaic mammals was 0.19 and ranged between 0.11 and 0.55.

The smallest mammal that contributed data for the convex polygon for living mammals in Figure 3 was a 3.9 g pygmy shrew (*Sorex minitus*) with a 0.15 g brain. The heaviest mammal was a 59 059 kg finback whale (*Balaenoptera musculus*) with a 6800 g brain. The top vertex of the mammalian polygon is shared by a human (70 kg body and 1300 g brain) and a dolphin (142 kg body and 1735 g brain). Most of the data are from Quiring, whose species names are also used here.

I have seen other published measures that are slightly different from those in Quiring, but they do not significantly affect the picture of the present diversity in brain size. For example, blue whales have been reported as weighing more than 100 000 kg, the largest mammals that ever lived. Logarithmic scaling masks some these large differences in measurements, shifting the polygons only slightly if one uses sizes other than the ones

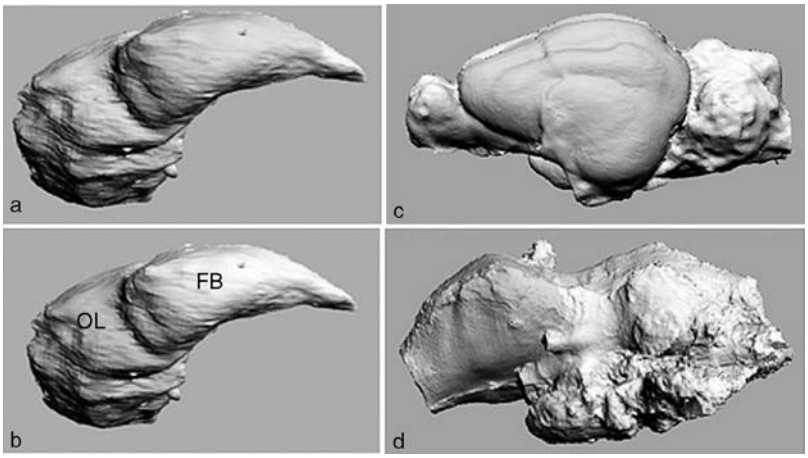

Figure 1 Fossil endocasts. (a) *Archaeopteryx lithographica*, the earliest bird, Late Jurassic, approximately 150 Ma. Facing right; length, 2.7 cm. Natural History Museum of London, Specimen BMNH 37001. (b) *Archaeopteryx*: FB, forebrain; OL, optic lobes (superior colliculi). (c) *Cynodictis*, Late Eocene carnivorous mammal, approximately 40 Ma; a fox-sized species ancestral to wolves and bears. Facing left; length, 5.3 cm. Field Museum of Natural History, Chicago, specimen FMNH PM 59013. (d) *Edmontosaurus*, a Late Cretaceous, approximately 70 Ma, duckbill dinosaur. Facing left; length, 16.4 cm. Museum of the Rockies Specimen, MOR Specimen 639.

in Quiring. If brain and body measurements for the largest blue whale had been available, the mammalian polygon would have been slightly longer than for dinosaurs, but it would have been approximately the same size and oriented in the same way. The relationships between the fossils and the living species are correctly represented in Figure 3.

The smallest bird for the living avian polygon was a 4.8 g hummingbird (*Amazilia tzacatl*) with a 0.2 g brain and the largest was a 123 kg ostrich (*Struthio camelus*) with a 42 g brain. Of living reptiles, the smallest was a 0.68 g lizard, *Phylodactylus gerrophygus*, with a 0.018 g brain. The largest was a 205 kg crocodile with a 14 g brain.

As one can see, when dinosaur data were added to the graph in Figure 3, it was natural to represent them by an extension of the reptilian polygon. The current consensus differentiates dinosaurs from other reptiles and has living birds as surviving small dinosaurs with feathers. The issue is not central to this article. It is appropriate to remember that organisms are mosaics of traits, and traits do not necessarily evolve at the same rate. The earliest bird evidently shared some brain features, such as brain size, with its living descendants; nonavian dinosaurs related to early birds may have been reptilian in brain size. There is one candidate, *Compsognathus*, a small dinosaur known from the same fossil quarry and approximately the same size as *Archaeopteryx*, but it has been impossible to prepare an endocast from this specimen. It would obviously be helpful to have the data. With respect to brain size, dinosaurs appear to have been more like reptiles than like birds.

The largest of the dinosaurs was the herbivorous *Brachiosaurus*, estimated to have weighed 87 000 kg with a brain size of approximately 155 g. The endocast of the famous carnivore, *Tyrannosaurus rex*, is also well-known – approximately 400 ml in volume, suggesting a 200 g brain. These are approximately the same measurements as those for *Edmontosaurus* (Figures 1 and 3; body weights estimated as 7700 kg for *T. rex* and 6000 kg for *Edmontosaurus*, and brain sizes estimated as approximately the same).

Measurements of gross brain size and gross body size in living species, appropriately analyzed, provide the background for understanding the evolution of the brain. Graphs such as those in Figure 3 map present diversity. Adding fossil data then shows features of the evolution of the vertebrate nervous system. When the nervous system is analyzed in more detail for microscopic and neurochemical anatomy, one should keep in mind the evolutionary background provided by fossil endocasts, including the quantitative analysis of Figure 3.

Birds and Dinosaurs

The endocast of the fossil bird graphed at A on the bird polygon in Figure 3 was excavated a few years after Darwin published his great book. It was immediately recognized as evidence of an important transition from reptiles to birds, supporting Darwin's views on natural selection. The present consensus separates dinosaurs from other reptiles and has *Archaeopteryx* as an avian dinosaur adapted for flight or gliding. Living birds are all viewed as surviving feathered

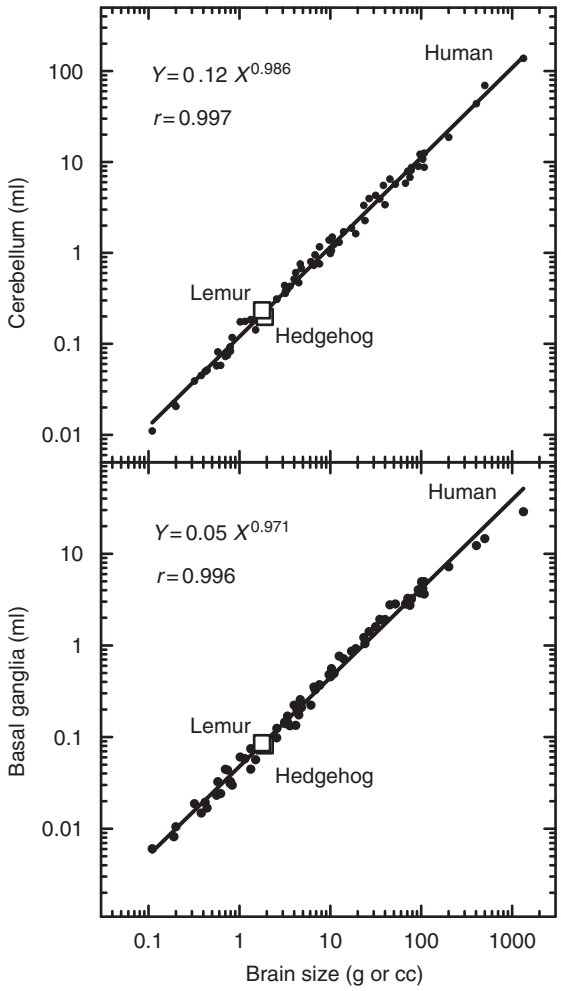

Figure 2 Brain size estimates the size of cerebellum and basal ganglia almost perfectly ($r > 0.99$). The marked points are from a desert hedgehog (*Hemiechinus*) with a 1.9 g brain and 250 g body and from a mouse lemur (*Microcebus*) with a 1.8 g brain and 54 g body, indicating that brain size, rather than body size, is the critical independent variable. Reprinted from Jerison HJ (1991) *Brain Size and the Evolution of Mind: 59th James Arthur Lecture on the Evolution of the Human Brain*. New York: American Museum of Natural History, with permission.

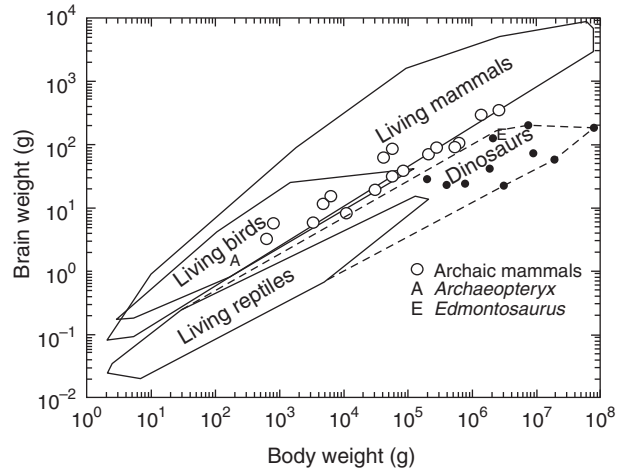

Figure 3 Brain–body relations in living and fossil amniotes. Data points for living species within each polygon are omitted; the convex polygons depict the present diversity of brain size in living amniotes. In drawing the polygons, each species was represented by a single point: 647 species of mammals, 219 species of birds, and 59 species of reptiles. Fossil data for 17 species of archaic mammals as published in Jerison in 1973 are added as open circles. Solid circles represent dinosaur data on 13 species. (For clarity, only 11 species are graphed because the omitted points masked the labels.) Data for the earliest bird, the late Jurassic *Archaeopteryx*, and the late Cretaceous duckbill dinosaur, *Edmontosaurus* (cf. **Figure 1**) are graphed as points A and E.

dinosaurs. The evidence of the brain, nevertheless, separates *Archaeopteryx* from dinosaurs, and the avian from the reptilian polygon of **Figure 3**.

External morphology also separates *Archaeopteryx* from dinosaurs and reptiles with respect to the brain. One superficial feature is the extent to which optic lobes (superior colliculi) are distinct from forebrain. Although on the brain these structures are as distinct on reptiles as on all living vertebrates, they are not distinct on endocasts. Comparing the endocasts of *Archaeopteryx* with that of *Edmontosaurus* illustrates the difference. It is not difficult to distinguish forebrain from optic lobes in the bird, but it is essentially impossible in the dinosaur.

A second superficial feature of endocasts of living birds is a large forebrain hump, the Wulst, which is a visual center in the brains of all living birds. It is probably homologous with mammalian visual cortex. *Archaeopteryx* had no Wulst. Besides *Archaeopteryx*, I know of only three fossil birds in which endocasts are good enough to show Wulst – an Eocene bird probably related to living curlews, a Miocene raptor, and the Pleistocene *Teratornis*. I see no evidence in Wulst in the Eocene curlew ('*Numenius gypsorum*'), but it is clearly evident in the later raptors, pretty much as in living birds. There should probably be a Wulst evident in at least some dinosaur brains. I have seen no convincing evidence of it in the dozen or so dinosaur endocasts that I have examined, but in personal communications, several colleagues have noted seeing signs of Wulst in at least one dinosaur, *Troodon*.

The distinction between forebrain and optic lobes in birds is quantifiable. Current three-dimensional imaging methods permit one to measure surface area, and it should be possible to measure visible surface area of forebrain and that of optic lobes. Qualitative observations are enough to show that forebrain

covered less of the optic lobe in fossils than in living birds, but there are too few specimens to quantify the relationship at this time. A similar quantification on the size of neocortex was easily done for a sample of more than 100 mammalian endocasts, as discussed later.

The final point on birds concerns the relative size of the brain in *Archaeopteryx*. It lies within the convex polygon for living birds, but it is on the lower margin. Its brain was small compared to those of living birds of similar body size. Quiring reported a 282 g pigeon (*Columba livia*) with a 2.7 g brain. *Archaeopteryx* is usually estimated as slightly heavier, 300–400 g, but from its endocast its brain probably weighed approximately 1.5 g. This made it heavier than reptile brains at that body weight. It was just within the lower border of our polygon for living birds. Some encephalization beyond a dinosaurian grade is evident in the data on *Archaeopteryx*.

This summary shows how the fossil evidence contributes to our knowledge of the evolution of the brain even in very small samples – only four species – of fossil birds and their living descendants. Despite the small sample size, the evidence is convincing. It is somewhat inconsistent with the view of birds as surviving dinosaurs, which is based on a mosaic of traits. The brain is an exception. The best guess, widely shared on the basis of overall vertebrate morphology, is that the bird's brain evolved from a pattern shared with living reptiles and is probably fundamentally similar to that of dinosaurs. However, it departed from that pattern with respect to overall size and with respect to the extent to which its shape is reflected by endocasts from the cranial cavity.

Mammals

I reviewed the fossil evidence on mammalian brain evolution in 2006, emphasizing the evolution of neocortex. I summarize it here, and **Figure 3** also provides data on encephalization by comparing relative brain size in archaic and living species of mammals. In this section, as part of a review of the method of measurement (**Figure 4**), I also present the main evidence on neocorticalization (**Figure 5**).

The evolution of encephalization in mammals as shown in **Figure 3** is represented by the clustering of the data on archaic mammals about the lower edge of the convex polygon for living mammals. These brains of mammals that lived more than 40 Ma were between approximately 10% and 50% the size of those of living mammals of comparable body size. There has always been variation in encephalization. Many small-brained species are alive today and have enough gray matter to do quite well in the current world. The Virginia opossum (*Didelphis virginiana*) has approximately one-fifth as large a brain as a living animal of the same size, such as the cat (*Felis catus*).

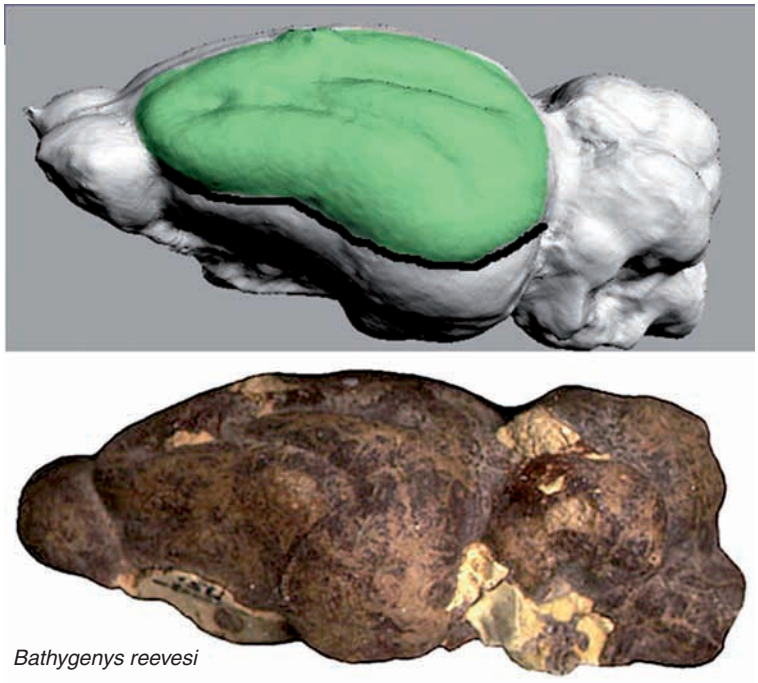

Figure 4 Endocast of a small Late Eocene artiodactyl, *Bathygenys reevesi*, 35 Ma; length, 5 cm. The animal, an oreodont, lived in what is now Big Bend National Park on the Rio Grande in Texas. (Top) Virtual image from a three-dimensional laser scan. Rhinal fissure is marked. (Bottom) Photograph of the same specimen (specimen UT 40209–431, University of Texas, Department of Paleontology).

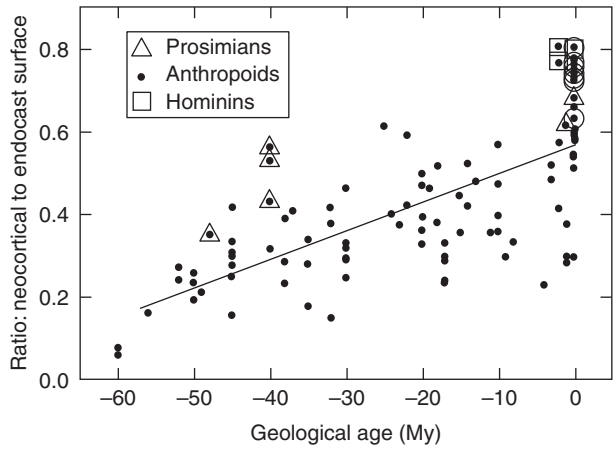

Figure 5 Relative size of neocortex in 106 species of mammals during the past 60 My. Prosimians are marked by triangles, hominins (including australopithecenes) by squares, and simians by circles. Regression of geological age on neocortical ratio: $Y = 0.007\,X + 0.57$; $r = 0.72$. In living mammals, 57% of the endocast surface area is devoted to neocortex; in living monkeys and humans, and in fossil hominins, the ratio averages approximately 75%. Reproduced from Jerison HJ (2006) What fossils tell us about the evolution of the neocortex. In: Kaas JH and Krubizer LA (eds.) *Evolution of Nervous System*, 4 vols. New York: Elsevier, with permission.

The European hedgehog (*Erinaceus europeaus*) has approximately one-third as large a brain. The prevalence of opossums and hedgehogs as road kill on US and European highways is evidence that these animals survive quite well if they stay out of traffic. Although the data in Figure 3 indicate a selective advantage for an enlarged brain, the extent of the advantage differed in different lineages.

The endocast of only one species of mammal, the Eocene 'bear–dog' *Cynodictis*, was shown in the gallery of endocasts in Figure 1. It is an example of how brainlike an endocast can be. *Cynodictis* was not included among the 'archaic' mammals because it was a carnivore, a member of the same order Carnivora that survives today. The reader can verify that it was relatively small brained, however. Its coordinates are for a brain size of 12 g and an estimated 'foxlike' body size of 5000 g (EQ = 0.34). Living bears, wolves, cats, and dogs are 'average' mammals in encephalization (i.e., EQ = 1, approximately). If *Cynodictis* was an average carnivore of its time, there was a tripling in average brain size for carnivores during the 40 My of evolution covered here.

Mammalian endocasts often provide remarkably detailed pictures of the external topography of the brain. The pattern of gyri and sulci in *Cynodictis* is fairly typical for Eocene species, but its rhinal fissure may be difficult to identify. The endocast of a Late Eocene ungulate, *Bathygenys reevesi*, is presented in Figure 4 to show the rhinal fissure less ambiguously.

Neocortex lies dorsal to the rhinal fissure and is distinguishable from paleocortex, which is ventral to the fissure. The top half of Figure 4 is a snapshot of a rendered ditalized image of the endocast, and an area of the surface of this virtual endocast can be marked and measured. Such measurements were almost impossible before the development of digitizing methods with appropriate computer graphics software. The evolution of neocortex can then be determined on a series of mammals from different geological strata.

As I reported in 2006, a diverse sample of 84 fossil and 22 living species, a total of 106 mammalian endocasts, was scanned and measured. The sample included primates, carnivores, ungulates, rodents, insectivores, and marsupials as well as less familiar animals. Many of the species are from archaic orders that are now entirely extinct. To control for body size differences in brain size, neocorticalization was measured by the ratio of neocortex surface area to the surface area of the whole endocast. These ratios, represented by the points in the sample of 106 mammals, are graphed in Figure 5. Recognizing our special interest in human evolution, the primate points in the sample are identified by enclosing them in special symbols.

The most important result of this analysis is the relatively low correlation ($r = 0.72$) between the extent of neocorticalization in mammals and the passage of time. This is evidence of the diversity of specializations in mammals with respect to the brain. Many marsupials are only slightly more neocorticalized than the earliest mammals for which we have evidence. (The koala, despite its winning charms, holds the record in our sample as low beast on the living totem pole, with only 28% of its brain devoted to neocortex.) A variety of niches have been invaded by mammals, and they differ in the extent to which neocorticalization is important.

There are two results documented in Figure 5. First, there was a significant trend toward neocorticalization in mammals as a class. The measure of change in neocorticalizaion is the significant upward slope of the regression line. It averaged an approximately 7% increase in neocorticalization per 10 My. Variation about this trend reflects differences among species in selection for neocorticalization. Second, primates have always been more neocorticalized than other mammals, at least as suggested by this sample. It is worth noting that humans are average anthropoid primates on this dimension. The most neocorticalized of primate species in this sample was a langur, *Cercocebus albigena*, which was approximately the same as the human endocast in the sample, with 80% of its endocast surface area devoted to neocortex.

Marsupials were and are less neocorticalized than placental mammals. Furthermore, among species of carnivores in similar niches (Order Creodonta compared to Order Carnivora), the more neocorticalized species were probably at a selective advantage over less neocorticalized orders. At least they did not go extinct as less neocorticalized species did.

Conclusions

The results discussed here provide an overview of the evidence on brain evolution available from the fossil record. The evidence is primarily by inference from the data on brain size. The most complete evidence is on mammals, and it may be most useful for decisions about which species to study. It should be background for other analyses based on the anatomy, genetics, physiology, and neurochemistry of living brains.

See also: Bird Brain: Evolution; Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brain Fossils: Endocasts; Brain Scaling Laws; Neocortex: Origins.

Further Reading

Butler AB and Hodos W (2005) *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation,* 2nd edn. New York: Wiley–Interscience.

Dominguez PA, Milner A, Ketcham RA, Cookson MJ, and Rowe TB (2004) The avian nature of the brain and inner ear of Archaeopteryx. *Nature (London)* 430: 666–669.

Edinger T (1975) Paleoneurology, 1804–1966: An annotated bibliography. *Advances in Anatomy, Embryology and Cell Biology* 49: 12–258.

Hopson JA (1979) Paleoneurology. In: Gans C, Northcutt RG, and Ulinski P (eds.) *Biology of the Reptilia,* vol. 9, pp. 39–146. Academic Press: New York.

Jerison HJ (1973) *Evolution of the Brain and Intelligence.* New York: Academic Press.

Jerison HJ (1990) Fossil evidence on the evolution of the neocortex. *Cerebral Cortex* 8A: 285–309.

Jerison HJ (1991) *Brain Size and the Evolution of Mind: 59th James Arthur Lecture on the Evolution of the Human Brain.* New York: American Museum of Natural History.

Jerison HJ (2001) Epilogue: The study of primate brain evolution: Where do we go from here? In: Falk D and Gibson K (eds.) *Evolutionary Anatomy of the Primate Cerebral Cortex*, pp. 305–337. Cambridge, UK: Cambridge University Press.

Jerison HJ (2004) Dinosaur brains. In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience,* 3rd edn. Amsterdam: Elsevier [CD edition].

Jerison HJ (2006) What fossils tell us about the evolution of the neocortex. In: Kaas JH and Krubizer LA (eds.) *Evolution of Nervous System,* 4 vols. New York: Elsevier.

Karten HJ (1997) Evolutionary developmental biology meets the brain: The origins of mammalian cortex. *Proceedings of the National Academy of Sciences of the United States of America* 94: 2800–2804.

Nieuwenhuys R, ten Donkelaar HJ, and Nicholson C (1998) *The Central Nervous System of Vertebrates*, 4 vols. New York: Springer.

Quiring DP (1950) *Functional Anatomy of the Vertebrates.* New York: McGraw-Hill.

Stephan H, Frahm H, and Baron G (1981) New and revised data on volumes of brain structures in insectivores and primates. *Folia Primatologica* 35: 1–29.

Ulinski PS (1983) *Dorsal Ventricular Ridge: A Treatise on Forebrain Organization in Reptiles and Birds.* New York: Wiley.

Alternative Splicing in the Nervous System

J Y Wu, Northwestern University, Chicago, IL, USA
J A Potashkin, Rosalind Franklin University of Medicine and Science, North Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

Soon after the discovery of split genes in 1977, evidence began to accumulate to support the hypothesis that multiple mRNAs encoding different protein products could be produced from a single precursor mRNA (pre-mRNA). It was not until the completion of sequencing the human genome, however, that the extent of alternative splicing became apparent. The results from this major undertaking indicated that the number of protein-coding genes is far fewer than expected based on phenotypic diversity. Studies over the past 30 years have revealed the complex multidimensional networks of regulation in mammalian gene expression. Regulation can occur at the level of chromatin structure, transcription, pre-mRNA splicing/alternative splicing, alternative polyadenylation, RNA stability, RNA transport, RNA editing, and post-translational modifications. The following discussions focus on the role of alternative splicing in the nervous system and splicing mutations associated with neurodegenerative diseases.

Pre-mRNA Splicing and Alternative Splicing

Pre-mRNA splicing, the process of removing introns from the nascent transcript and splicing together of exons to produce the functional messenger RNA (mRNA), occurs in a spliceosome, a macromolecular RNA–protein machine. In addition to the pre-mRNA and small nuclear ribonucleoprotein particles (snRNPs), the spliceosome contains over 200 protein factors. Essential to this process is the accurate recognition of the 5′ splice site (5′ss) and 3′ splice site (3′ss) by the spliceosome. In the assembled catalytically active spliceosome, the biochemical reactions of cleavages at splice sites and ligation of exons take place by a two-step transesterification mechanism.

In higher eukaryotes, multiple splicing isoforms can be produced from a single pre-mRNA by alternative usage of different splice sites. This process of alternative splicing can occur in a variety of patterns (Figure 1), including exon skipping/inclusion (cassette alternative exon), intron retention, alternative use of 5′ or 3′ splice sites, mutually exclusive exons, and more complex patterns that are coupled with alternative promoter selection or alternative polyadenylation. Alternative splicing may produce mRNA species with different stability or subcellular localization and/or transcripts that encode distinct protein products. As a result, alternative splicing is an extremely powerful and versatile mechanism for generating functional diversity from a limited number of genes.

In the nervous system, the process of alternative splicing is particularly remarkable because the brain has the highest frequency of alternative splicing of any tissue and it is becoming evident that regulation of alternative splicing in the brain has a significant impact on the development, function, and maintenance/repair of the nervous system. In addition, the largest group of genes that display tissue-specific splicing is expressed in the brain.

Molecular Mechanisms Regulating Alternative Splicing

In higher eukaryotes, especially mammals, pre-mRNA transcripts are usually long, containing multiple introns of variable sizes. Mammalian introns can be as large as 500 kb. In humans, the average size of exons is 150 nucleotides (nt), and that of introns is 3500 nt nucleotides. The basic splicing signals in mammalian pre-mRNAs are degenerate, with only two nucleotides (/GT at the 5′ss and AG/ at 3′ss) that are highly conserved. As a result, the nucleotide sequences surrounding the splice junctions usually contain only a limited amount of information, not sufficient for conferring the specificity required to achieve accurate splice site selection. Therefore, recognition not only of exon–intron junction sequences but also of the regulatory elements in intronic and exonic regions is important for defining splice junctions and maintaining splicing fidelity. This high degree of degeneracy in the splicing signals in mammalian pre-mRNAs provides the flexibility for alternative selection and pairing of different splice sites, a fundamental mechanism for regulating alternative splicing.

The intricate interactions between *cis*-elements in the pre-mRNA substrates and their *trans*-acting splicing factors determine the selective use of different splice sites. This network of interactions between pre-mRNA and *trans*-acting factors involves both snRNPs and non-snRNP splicing regulators. The *cis*-elements include splice sites and splicing regulatory sequences (splicing enhancers or silencers) inside either exons or introns. The sequences and sizes of the exons or introns, as well as secondary structures of the pre-mRNA, also influence splice site selection.

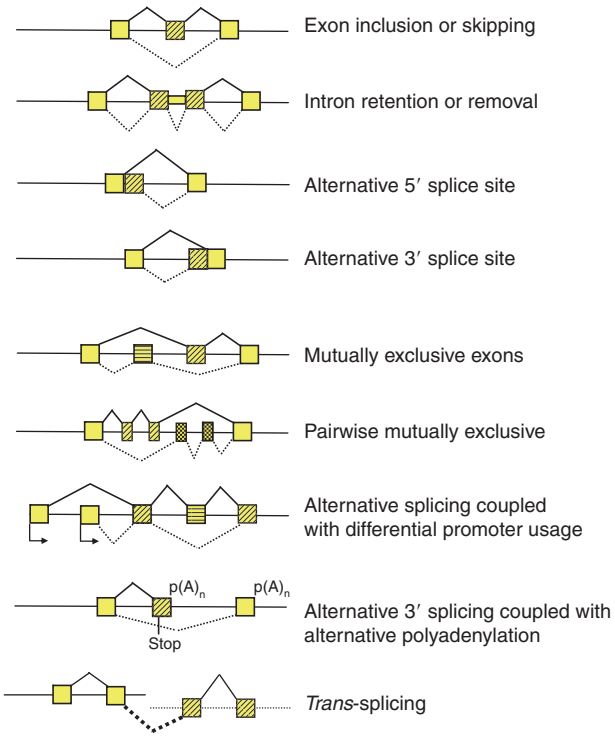

Figure 1 A range of alternative splicing patterns found in the nervous system. Most splicing events are *cis*-splicing, and trans-splicing may occur at a low efficiency in higher eukaryotes. Adapted from Wu JY, Yuan L, and Havlioglu N (2004). Alternatively spliced genes. In: Meyers RA (ed.) *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, 2nd edn., vol. 1, pp. 125–177. Weinheim: Wiley-VCH.

The high level of degeneracy of mammalian splicing signals and the existence of large numbers of trans-acting factors allow versatile RNA–protein and RNA–RNA interactions during different stages of spliceosome assembly. The production and delicate balance of distinct isoforms generated by the alternative splicing of pre-mRNA transcripts are determined by combinatorial effects of multisite interactions among pre-mRNA, essential spliceosomal components, and regulatory factors.

A number of splicing regulators, either positive or negative, modulate alternative splicing. These splicing regulators contain several types of protein sequence motifs, including an RNA recognition motif (RRM) and other nucleic acid binding motifs (K-homology, or zinc finger) as well as a serine/arginine (SR)-rich domain or glycine-rich or RGG-rich sequences. Many SR domain-containing factors interact with exonic splicing enhancer elements to stimulate exon inclusion, whereas proteins of the heterogeneous nuclear ribonucleoprotein (hnRNP) family containing RRMs often act as splicing repressors by interacting with intronic splicing silencer sequences. A number of bifunctional splicing factors have been identified that can act either as activators or repressors for splicing, depending on their interactions with pre-mRNAs and other proteins. The concentration, distribution, composition, and state of modification of these regulatory factors determine whether they enhance or suppress the use of a particular splice site.

The *cis*-acting elements that regulate splicing are enhancer or silencer sequences present in either the intronic or exonic regions. Both types of elements may regulate splicing of a single pre-mRNA. Splicing enhancer and silencer elements may also overlap with each other. The splicing silencer sequences in several well-characterized genes, including those encoding c-src, caspase-2, and the γ-aminobutyric acid $(GABA)_A$ receptor $\gamma 2$ subunit, contain pyrimidine (U/C)-rich elements that interact with the polypyrimidine tract binding (PTB) protein (also referred to as hnRNP I). In these examples, PTB protein binds the repressor sequences in nonneuronal cells so that splice sites are skipped. The PTB gene also undergoes alternative splicing to generate three isoforms. In addition, a neuronal or brain-enriched PTB homolog (nPTB or brPTB) has been identified, and plays a role in relieving the suppression of inclusion of a neuron-specific N1 exon in c-src in neuronal cells. Splicing regulatory elements usually contain multiple binding sites for splicing regulators and function by recruiting other spliceosomal components to form RNP-like complexes. One of the common features of these *cis*-regulatory elements is that they are not simple sequence elements that act independently of each other. Splicing regulatory elements have been identified that can act to stimulate the splicing of one exon but repress another exon.

Neuron-Specific Splicing Regulators

A large number of splicing regulators are expressed in a wide range of tissues or cell types. Several RNA-binding proteins have been reported that are enriched in neurons or have neuron-specific isoforms. These neural splicing regulators include the Nova family (Nova-1 and Nova-2), Hu/ELAV family (Hud, Hel-N), Fox family (Fox-1 and Fox-2), STAR/GSG family, and CELF family (NAPOR/CUGBP2/ETR-3). Some of these factors were originally identified because of their association with neurological disorders.

Only a few of these neuron-specific splicing factors have been well studied. These include the K-homology (KH) domain RNA-binding proteins Nova-1 and Nova-2. Nova proteins were identified by their association with paraneoplastic opsoclonus myoclonus ataxia (POMA). In this syndrome, patients develop antibodies against Nova proteins, and this autoimmune

response leads to a loss of inhibition of motor control in the spinal cord and brain stem. Nova-1 is expressed in the diencephalon, brain stem, and motor neurons of the spinal cord, whereas Nova-2 is present in the cerebral cortex, hippocampus, and dorsal spinal cord. Nova-1 binds to UCAUY elements, and Nova-2 interacts with GAGUCAU elements in their target RNAs (Y represents a pyrimidine). Nova-1 regulates the splicing of $GABA_A$ receptor $\gamma 2$ subunit and the $\alpha 2$ subunit of the glycine receptor (GlyR $\alpha 2$), and also autoregulates its own splicing. Nova has also been reported to co-regulate the splicing of several transcripts important for synaptic function, perhaps accounting for 7% of the brain-specific alternative splicing in the neocortex.

The Hu family is another group of splicing regulators that are autoimmune targets in POMA. Hu proteins bind to mRNAs, including many that are important for neuronal function, thus stabilizing the transcripts and increasing their translation. One member of the family of proteins, HuD, is a homolog of the fruit fly splicing regulator ELAV protein, which is essential for neurogenesis and plays a role in regulating splicing in neurons. Recently, Hu was reported to regulate splicing of calcitonin/calcitonin gene-related peptide (CGRP) pre-mRNA and facilitate exon 4 skipping by inhibiting the binding of TIA-1/TIAR, two ubiquitously expressed factors that promote the nonneuronal pathway.

The neuron-specific splicing of calcitonin/CGRP is also regulated by several other splicing factors, including Fox-1 and Fox-2. These proteins are expressed in muscle, heart, and brain tissues. In the brain, the Fox proteins are neuron specific. Fox-1 and Fox-2 promote exon inclusion in NMHC-B, c-src, FGFR2, and protein 4.1R transcripts. In contrast, within neurons Fox proteins repress exon 4 inclusion in calcitonin/CGRP pre-mRNA by binding to a UGCAUG element and blocking the constitutive splicing factor U2 auxiliary factor 65 ($U2AF^{65}$) from binding to the 3′ splice site upstream of exon 4. The splicing events, which result in the generation of calcitonin/CGRP, also involve several additional regulators, including Tra2β, SRp55, TIAR, SRp20, and PTB.

The STAR (signal transduction and activation of RNA)/GSG (GRP33, Sam68, GlD-1) protein family includes the mammalian Sam68-like protein SLM-1, which regulates splice site selection by binding to a purine-rich enhancer. Rat SLM-1 and SLM-2 are expressed primarily in neurons. The tissue-specific splicing factor SLM-1 is present in the dentate gyrus, whereas SLM-2 is found in the pyramidal cells of the CA1, CA3, and CA4 regions. It is not known if SLM proteins regulate neuron-specific splicing, but SLM-2 regulates the splicing of CD44 and may also regulate tau and tra2β1 pre-mRNA splicing.

The neuroblastoma apoptosis-related RNA-binding protein (NAPOR)/CUG-binding protein 2(CUGBP2)/embryonic lethal abnormal vision type RNA-binding protein 3(ETR-3) has three splice variants; NAPOR-3 is neuron specific while the other two forms are ubiquitously expressed. NAPOR-3 binds to UG-rich regions. The concentration of NAPOR in different brain regions is important for determining its function. Within the forebrain, where it is abundantly expressed, NAPOR promotes exon 5 skipping of E5 and exon 21 inclusion of in *N*-methyl-D-aspartate (NMDA) receptor R1 pre-mRNA.

In addition to these neuron-specific or brain-enriched splicing factors, ubiquitously expressed splicing factors may also play a role in the regulation of splicing in neurons. For example, several splicing factors recognize pyrimidine-rich sequences within the intron similar to the sequence recognized by PTB, including U2 auxiliary factor (U2AF), CUG-binding protein (CUG-BP), and hnRNP F. Partial replacement of general splicing factors by neuronal-enriched splicing factors may lead to neural specific splicing. Another example of a general splicing factor that regulates a neuronal splicing event is KSRP, a factor that binds to an intronic splicing enhancer element downstream of the neuron-specific c-src N1 exon. KSRP is present in both neuronal and nonneuronal cells, but it is more highly expressed in neuronal cells.

A good example of the complexity of neuronal gene alternative splicing regulation is that of exon 10 in the human tau gene. Mutations affecting this splicing event lead to frontotemporal lobe dementia (see later, neurological disorders). In this case, both exonic and intronic regulatory elements play important roles in controlling exon 10. In addition, both positive and negative *cis*-elements are involved and form a multidomain composite regulatory element. A number of *trans*-acting splicing regulators have been identified that interact with the splicing regulatory element.

It is not clear yet how the extremely complex alternative splicing events of different genes in the nervous system are coordinated. In some cases, the alternative splicing events of different genes appear to be co-regulated by the same protein. For example, the splicing regulator PTB can differentially recognize neural and nonneuronal substrates. The expression pattern of PTB in various regions of the brain at different developmental stages supports a role for PTB to act as one alternative splicing coordinator for different splicing target genes. Some regulatory elements have been identified that mediate splicing responses of neurons to extracellular stimuli. In general, very little is known about the molecular mechanisms controlling changes in alternative splicing in response to environmental signals.

Role of Alternative Splicing in the Nervous System

The regulatory role that alternative splicing plays in the nervous system is multifold. Alternative splicing has been reported for genes involved in almost every aspect of the nervous system, from neural development to the function and maintenance of the adult nervous system. The expression of a vast number of genes important for neuronal differentiation, function, and survival undergo alternative splicing. These genes encode trophic factors, neuronal guidance molecules, guidance cue receptors, transmitter receptors, ion channels, intracellular signal transduction molecules, and synaptic components. Similar to the role of splicing in nonneuronal cells, alternative splicing may affect transcript localization/stability or generate proteins of distinct functional activities.

The variety of receptors produced by alternative splicing in neurons is a classic example of how a few genes can produce many protein products. The genes encoding receptors for major neurotransmitters, including serotonin, dopamine, glutamate, and GABA, all undergo alternative splicing. This results in the production of receptors that differ in function and/or subcellular localization. For example, the NMDA R1 receptor has an alternatively spliced C1 exon. When C1 exon is included in the mRNA, the receptor localizes to the cell surface membrane; when C1 exon is skipped, the receptor is cytoplasmic. In addition, the regulation of C1 splicing modulates binding of NMDA R1 to neurofilaments and calmodulin. The splicing regulation of mGluR1 affects both localization and receptor activity. One variant, mGluR1a, has a long C-terminus that targets the protein to dendrites, whereas mGluR1b, containing a short C-terminus, localizes to axons.

Alternative splicing may affect how a receptor responds to drugs. For example, two splice variants of the $GABA_A$ receptor respond to benzodiazepine agonists in different manners. The splice variants of norepinephrine transporter have different rates of norepinephrine uptake and binding affinities to nisoxetine, a norepinephrine reuptake inhibitor. Additionally, the serotonin $5\text{-}HT_4$ receptor has splice variants with different C-terminal domains that respond differently to agonists.

Alternative splicing is also important for the modulation of neuronal signaling. For example, alternative splicing of the gene encoding the α_{1N} N-type calcium channel produces a protein with a two-amino-acid insertion that results in a channel with altered kinetics. Another example is apolipoprotein E receptor 2 (Apoer2, a receptor for Reelin), which forms a functionally active complex with NMDA receptor in the postsynaptic densities of excitatory synapses. An alternatively spliced exon of Apoer2 is required for Reelin to induce phosphorylation of NMDA receptor and thus enhance long-term potentiation (LTP). Interestingly, the splicing of Apoer2 is regulated by synaptic activity.

In addition to modulating synaptic function, alternative splicing plays a role in synapse formation and plasticity by regulating the expression of proteins important for cell adhesion and cell–cell communication. Neural cell adhesion molecule NCAM1 is involved in synaptic plasticity, neurodevelopment, and neurogenesis. Single-nucleotide polymorphisms (SNPs) within NCAM1 splice sites that affect alternative splicing are associated with bipolar disorder and schizophrenia.

The role of alternative splicing in synaptic plasticity is only beginning to be understood. Recently, target genes for the splicing factor Nova were identified, including $GABA_B$ receptors and G-protein-activated inward rectifier potassium channel 2 (GIRK2) channels. These proteins mediate LTP of the slow inhibitory postsynaptic current (siPSC) in dendrites. In mice lacking Nova-2, the LTP of siPSCs, but not the excitatory postsynaptic currents, is absent. The current model is that Nova proteins control the splicing of a number of pre-mRNAs, resulting in the regulation of LTP of the siPSCs. The application of new techniques such as cross-linking immunoprecipitation (CLIP) and microarray assays may reveal additional examples of splicing factors regulating networks that control synaptic plasticity.

Neurological Disorders Associated with Dysregulation and Abnormalities of Pre-mRNA Splicing

Human diseases may result from defective or aberrant splicing caused by mutations in regulatory pre-mRNA sequences and splicing factors. Recent data suggest that approximately 31% of human diseases are caused by mistakes in pre-mRNA splicing, and 15% of disease-causing point mutations affect pre-mRNA splicing. These statistics most likely underestimate the extent of human diseases associated with abnormalities in splicing, because splicing enhancer or silencer sequences, very few of which have been clearly identified, may carry point mutations that disrupt the balance of splice variants and thus result in disease. In addition, environmental stimuli such as stress may change intracellular distribution or posttranslational modifications of splicing factors and thus affect pre-mRNA splicing. To demonstrate the role of splicing dysregulation in human diseases, a few well-studied examples are described in

the following sections. In general, these diseases can be classified in two categories: those caused by *cis*-acting mutations in the affected genes and those associated with secondary splicing defects resulting from *trans*-acting mutations.

Neurological Diseases Caused by *cis*-Acting Splicing Mutations

Dementia Splicing mutations and aberrant pre-mRNA splicing have been associated with both sporadic and familial forms of dementia, including Alzheimer's disease. For example, splicing mutations have been identified in both presenilin-1 and presenilin-2 genes in Alzheimer's disease. Frontotemporal dementia and parkinsonism linked to chromosome 17 (FTDP-17) and related tauopathies are characterized by intraneuronal tau-containing deposits in affected brain regions. Microtubule-binding protein tau, a key player in microtubule stability, is critical for neuronal function. Mutations in human tau genes have been identified in FTDP-17 patients, including not only missense mutations affecting tau protein function but also splicing mutations causing imbalance of different tau splicing isoforms. Three exons in the human tau gene are alternatively spliced. In particular, exon 10 encodes one of four microtubule-binding repeat (R) domains, and alternative inclusion or exclusion of exon 10 leads to the formation of Tau4R or Tau3R, respectively. Exonic or intronic mutations that disrupt the balance of Tau4R versus Tau3R splicing isoforms lead to FTDP-17 and other forms of tauopathy with neurofibrillary tangles. In addition, splicing mutations altering exon 2 and exon 3 inclusion are associated with gliopathy and spinal cord degeneration. These observations demonstrate that an imbalance of different splicing isoforms can result in disease development or progression.

Muscular dystrophy Various mutations that affect alternative splicing of the dystrophin gene lead to Duchenne muscular dystrophy (DMD). Both exonic and intronic splicing mutations have been reported in DMD patients, including those inducing exon skipping and cryptic splice site activation. These mutations act either by disrupting splicing enhancers or by creating splicing silencers. In some cases, nonsense mutation(s) can also affect splicing by generating splicing silencer binding sites. For example, nonsense mutations in the dystrophin gene of some patients produced a binding site for hnRNP A1 (UAGACA) within the resultant pre-mRNA. These recent observations highlight the complexity of splicing errors and suggest that disease-causing splicing mutations may be underestimated.

Ataxia–telangiectasia and neurofibromatosis Ataxia–telangiectasia and neurofibromatosis type 1 are autosomal dominant neurological diseases with malignancy predisposition. Genetic studies have identified splicing mutations as common genetic defects in these patients, with approximately 50% patients carrying splicing mutations. These studies suggest the significant contribution of splicing defects to the pathogenesis of neurological disorders.

Spinal muscular atrophy Spinal muscular atrophy (SMA) is characterized by selective degeneration of motor neurons that leads to progressive paralysis. SMA is an autosomal recessive disease caused by loss of or mutations in the telomeric *survival of motor neuron 1* (*SMN1*) gene. The centromeric *SMN2* gene regulates the severity of the disease. Although *SMN1* and *SMN2* are almost identical, a single translationally silent nucleotide change in *SMN2* disrupts an exonic splicing enhancer and results in exon 7 skipping, with only about 20% of the *SMN2* transcripts producing a functional full-length SMN2 protein. This low level of full-length SMN2 protein is not sufficient to compensate for the loss of *SMN1*. A splice variant that retains intron 3 encodes an axonally localized truncated form of the protein called a-SMN. During development, a-SMN is expressed mainly in the axons of motor neurons and is only produced from the *SMN1* gene. It remains to be elucidated why the loss of *SMN1* primarily affects motor neuron function. Another interesting aspect of SMA is that the SMN1 protein plays a role in snRNP assembly and thus it may also be considered a *trans*-acting splicing disorder (see later).

Neurological Diseases Caused by *trans*-Acting Splicing Defects

A number of neurological syndromes or diseases are caused by *trans*-acting splicing defects. In these disorders, the mutations in genes encoding spliceosomal components or splicing regulatory factors lead to aberrant or defective splicing events.

Retinitis pigmentosa A common cause of blindness is retinitis pigmentosa, which is characterized by a loss of photoreceptor cells and progressive degeneration of the retina. A fraction of patients with autosomal dominant form of retinitis pigmentosa (adRP) have mutations in genes encoding spliceosomal proteins essential for spliceosomal assembly, including PRP3, PRP8, and PRP31. Similar to SMA, in these adRP patients, defects in the ubiquitously expressed RNA processing factors lead to specific neurological diseases. The mechanisms for the specificity of neurodegeneration in these diseases remain to be

elucidated. Current models include haploinsufficiency and dominant-negative effects leading to cell-type-specific functional deficiencies.

Myotonic dystrophy Myotonic dystrophy ('dystrophia myotonica'; DM) is a multisystem syndrome with prominent muscular atrophy. Two types of DM have been revealed by genetic studies: DM1 is caused by the presence of a CUG expansion in the 3′ untranslated region (UTR) of the DM protein kinase pre-mRNA, and DM2 is caused by a CCUG expansion in intron 1 of the *CNBP* (*ZNF9*) gene. CUG/CCUG repeat expansion results in sequestration or upregulation of proteins interacting with CUG/CCUG, including muscle blind-like (MBNL) and CUG-binding protein 1 (CUG-BP1). This in turn causes a dysregulation in the splicing of a subset of pre-mRNA muscle and brain transcripts that require these regulatory factors. DM1 is also associated with changes in splicing of pre-mRNAs that encode amyloid precursor protein, NMDAR1, and tau, suggesting that aberrant splicing of these pre-mRNAs may be the basis of the central nervous system defects such as memory impairments.

Prader–Willi syndrome Patients with Prader–Willi syndrome (PWS) often exhibit behavioral problems such as obsessive–compulsive disorder and autism. A recent report has suggested an explanation for why these patients frequently respond to treatment with serotonin reuptake inhibitors. PWS is associated with deletions of the paternally inherited copy of chromosome 15q11–q13, a maternally imprinted locus encoding multiple copies of a small nucleolar RNA called HBII-52 snoRNA. A stretch of 18 nt in the HBII-52 snoRNA is complementary to the sequence within an alternatively spliced exon (Vb) of the 5-HT_{2c}R serotonin receptor pre-mRNA. HBII-52 snoRNA binds to and antagonizes a splicing silencer, facilitating the inclusion of exon Vb within the mRNA. Changes in splicing of 5-HT_{2c}R were detected in the hippocampal RNA in PWS patients. These data suggest that altered 5-HT_{2c}R pre-mRNA splicing caused by the loss of HBII-52 snoRNA contributes to the clinical manifestations in PWS patients.

Ataxias Fragile X-associated tremor/ataxia syndrome (FXTAS) is a newly recognized late-onset syndrome characterized by cognitive impairment along with tremor and gait problems. FXTAS is associated with moderate expansions of 55–200 CGG repeats in the region of the fragile X mental retardation (*FMR1*) gene encoding the 5′ UTR. Longer expansions (>200) of CGG repeats lead to the fragile X mental retardation syndrome with numerous neurological problems, including autism, learning disabilities, anxiety disorders, and mental retardation. The *FMR1* gene is regulated by alternative splicing and encodes RNA-binding proteins involved in regulation of protein synthesis at postsynaptic sites of dendrites and in the maturation of dendritic spines. The CGG trinucleotide repeat expansions in the 5′ UTR lead to transcription silencing and loss of function of *FMR1*. Intranuclear inclusions in neurons and astrocytes in the postmortem brains of FXTAS patients contain the RNA-binding proteins MBNL1 and hnRNPA2, suggesting that CGG expansions may affect the function of these splicing factors and that FXTAS may involve aberrant pre-mRNA splicing. Further research is needed to understand the role of MBNL1, hnRNPA2, and FMRP in disease development.

In another form of ataxia, spinocerebellar ataxia subtype 2 (SCA2), there is a CAG repeat expansion in the coding region of the gene encoding ataxin-2. Ataxin-2 protein contains RNA-binding domains and interacts with the splicing regulator *A2BP1* (*FOX1*). Disruption of the *A2BP1* locus by chromosomal translocation has been reported in patients with mental retardation and epilepsy. Triplet repeat expansions in the noncoding RNAs have been found in patients with other SCA subtypes, including SCA8, SCA10, and SCA12. These observations suggest the possible involvement of pre-mRNA splicing dysfunction in various forms of ataxia. Further research is necessary to understand if there are common molecular mechanisms among numerous repeat expansion diseases, why these diseases manifest differently in patients, and whether dysregulation of alternative pre-mRNA splicing contributes to the pathogenesis of these diseases.

Paraneoplastic neurologic disorders In paraneoplastic neurologic disorder (PND) patients, malignancies outside of the central nervous system (CNS) induce expression of neuron-specific splicing regulatory factors (SRFs). The autoimmune response generated against these neuron-specific SRFs ultimately results in the dysfunction of neurons that normally express these proteins. Two families of neuron-specific SRFs have been identified as PND target antigens, Hu and Nova. Nova-1 and Nova-2 regulate pre-mRNA splicing of genes encoding synaptic proteins (see earlier). Nova RNA targets correlate with the neurologic defects in PND patients, consistent with observations in Nova-1 knockout mice.

Conclusions and Future Perspective

Overwhelming evidence supports the concept that pre-mRNA splicing and alternative splicing play an

important role in mammalian gene expression and that alternative splicing is a powerful mechanism for genetic diversity. This is particularly evident in the nervous system. Alternative splicing modulates activities of the majority of genes involved in the formation and function of the nervous system. A wide range of neurologic disease is caused by genetic mutations that affect pre-mRNA splicing. Some splicing mutations act *in cis* to cause splicing defects in genes containing the mutations. Other mutations affect spliceosomal components or splicing regulatory factors, which in turn lead to aberrant splicing or dysregulation of splicing of genes regulated by these splicing factors. Trinucleotide repeat expansion mutations can also affect pre-mRNA splicing by sequestering or altering splicing regulators that bind to the trinucleotide repeats and affect the splicing of the genes that are normally controlled by the splicing regulators. In addition, abnormal splicing has been reported in complex diseases such as psychiatric disorders. Although the important role of pre-mRNA splicing and alternative splicing regulation in the nervous system has been established, much work needs to be done to understand the molecular mechanisms underlying the splicing regulation and the molecular pathogenesis of splicing disorders.

Numerous environmental factors influence neuronal survival and function. Very little is known about how the nuclear splicing machinery responds to various extracellular signals, although many studies have documented altered pre-mRNA splicing in response to stresses such as pH change, osmotic or temperature shock, UV exposure, and oxidative stress. For example, brain ischemia induces the translocation of serine-arginine-rich splicing regulators from the nucleus to the cytoplasm and causes changes in alternative splicing. Further research is necessary to elucidate the molecular mechanisms by which environmental factors modulate pre-mRNA splicing and alternative splicing regulation.

The combined application of genetic, molecular, biochemical, and bioinformatic approaches will further our understanding of the role in pre-mRNA splicing and alternative splicing regulation in the nervous system. The ultimate goal is the development of strategies to correct splicing defects that cause human diseases.

See also: Gene Expression Regulation: Activity-Dependent; Gene Expression Regulation: Chromatin Modification in the CNS; Gene Expression Dysregulation in CNS Pathophysiology; Genetic Influence on CNS Gene Expression: Impact on Behavior; Genomic Disorder and Gene Expression in the Developing CNS.

Further Reading

Blencowe B (2006) Alternative splicing: New insights from global analyses. *Cell* 126: 27–46.

Dredge BK, Polydorides AD, and Darnell RB (2001) The splice of life: Alternative splicing and neurological disease. *Nature Reviews Neuroscience* 2: 43–50.

Faustino NA and Cooper TA (2003) Pre-mRNA splicing and human disease. *Genes & Development* 17: 419–437.

Grabowski PJ (1998) Splicing regulation in neurons: Tinkering with cell-specific control. *Cell* 92: 709–712.

Grabowski PJ and Black DL (2001) Alternative RNA splicing in the nervous system. *Progress in Neurobiology* 65: 289–308.

Jeanteur P (ed.) (2006) *Progress in Molecular and Subcellular Biology. Alternative Splicing and Disease,* vol. 44. Berlin, Heidelberg: Springer.

Kar A, Kuo D, He R, et al. (2005) Tau alternative splicing and frontotemporal dementia. *Alzheimer Disease, and Associated Disorders* 19(supplement 1): S29–S36.

Lee CJ and Irizarry K (2003) Alternative splicing in the nervous system: An emerging source of diversity and regulation. *Biological Psychiatry* 54: 771–776.

Lee JA, Xing Y, Nguyen D, et al. (2007) Depolarization and CaM kinase IV modulate NMDA receptor splicing through two essential RNA elements. *PLoS Biology* 5(2): e40 (doi:10.1371/journal.pbio.0050040).

Licatalosi DD and Darnell RB (2006) Splicing regulation in neurologic disease. *Neuron* 52: 93–101.

Potashkin J and Meredith G (2006) The role of oxidative stress in the dysregulation of gene expression and protein metabolism in neurodegenerative disease. *Antioxidant and Redox Signaling* 8: 144–151.

Ranum LP and Cooper TA (2006) RNA-mediated neuromuscular disorders. *Annual Review of Neuroscience* 29: 259–277.

Sharma S and Black DL (2006) Maps, codes, and sequence elements: Can we predict the protein output from an alternately spliced locus? *Neuron* 52(4): 574–576.

Stoilov P, Meshorer E, Gencheva M, et al. (2002) Defects in pre-mRNA processing as causes of and predisposition to diseases. *DNA and Cell Biology* 21: 803–818.

Ule J and Darnell RB (2006) RNA binding proteins and the regulation of neuronal synaptic plasticity. *Current Opinion in Neurobiology* 16: 102–110.

Wu JY, Yuan L, and Havlioglu N (2004) Alternatively spliced genes. In: Meyers RA (ed.) *Encyclopedia of Molecular Cell Biology and Molecular Medicine,* 2nd edn., vol. 1, pp. 125–177. Weinheim: Wiley-VCH.

Aluminum

S C Bondy, University of California at Irvine, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Background and Chemistry

Aluminum (Al) is an abundant metal, composing 8.3% of the Earth's crust by weight. It is the thirteenth element in the periodic table and its electronic configuration is $[Ne]3s^2 3p^1$. It was not until 1825 that the metallic form of the element was first isolated. Aluminum is a simple trivalent cation incapable of undergoing valence changes. It is highly oxophilic, and aluminum within minerals is usually found surrounded by six oxygen atoms. Because of its charge and the fact that it is a Lewis acid, Al^{3+} has a high affinity for electrons and thus can be a good catalyst. However, this element is not known to have a biological function. Since aluminum has a high charge to radius ratio, thermodynamically the metal prefers electrostatic rather than covalent binding. The aqueous chemistry of aluminum is complicated because free Al^{3+} hydrolyzes to form a wide range of complexes with water. There are three main categories of aluminum species relevant to toxicological availability. At neutral pH, Al salts undergo extensive hydrolysis and $Al(OH)_3$ is produced. As the solutions age, $Al(OH)_4^-$ is also present, and this leads to precipitation of polymeric and colloidal aluminum.

Until recently it was generally believed that aluminum released into the environment is harmless. This is because, in solution, Al^{3+} salts form monomeric hydroxy compounds which become chemically inert polymeric and colloidal particles as the solution ages. Because of the formation of these insoluble aluminum species, it was assumed that absorption would be limited and thus the metal would be relatively innocuous. However, the metal has been shown to be toxic to both plants and animals, and there is a rising concern over the metal's potential adverse health effects.

Aluminum Prevalence in the Human Environment

In the Roman Empire, around 2000 years ago, aluminum salts were used for the purification of water, and in the Middle Ages, aluminum was combined with honey for the treatment of ulcers. Aluminum salts are still used today for water purification and medicinal purposes and are present in many antacids. Al is present in a variety of other medicines, including vaccines, buffered aspirin, and phosphate binders used in dialysis. Alum is used medically as a topical astringent and as a styptic to stop bleeding. Parenteral nutrition solutions are often high in aluminum salts.

The main source of inadvertent aluminum intake is food, and the element is found in many food additives, such as processed cheese, baked goods, and grain products. Alum is used as a crisping agent in pickle processing and may be present in baking powder. Aluminum-containing compounds are also used as preservatives, coloring agents, and leavening agents. Antiperspirants containing aluminum chlorohydrate are another source of exposure. Occupational exposure to aluminum can occur and is mainly found where the metal is being processed.

The presence of aluminum in drinking water is due both to natural sources and to widespread water purification procedures involving the use of Al as a coagulant. Water treatment generally increases the content of soluble, low molecular weight, chemically reactive, and possibly more readily absorbed aluminum species. The advent of modern industrial technology and the introduction of a variety of chemicals into the atmosphere have led to the formation of acid rains. Strong mineral acids such as sulfur and nitrogen oxyacids, found in acid rains, can solubilize aluminum from soil and mobilize the metal. Thus the increasing prevalence of acid rain can lead to the release of greater amounts of aluminum salts from insoluble minerals, leading to greater bioavailability.

Aluminum Entry into Nervous Tissues

The most common form of human exposure to Al^{3+} is absorption through the gastrointestinal tract. The rate of absorption is approximately 0.22%, and once in the blood, approximately 90% of the metal is bound to transferrin. Al^{3+} can pass the blood–brain barrier by receptor-mediated endocytosis of the Fe-carrier protein, and in rats, approximately 0.005% of the metal complex enters the brain.

Transdermal absorption of aluminum has been reported in mice after application of Al chlorohydrate, a substance commonly present in commercial underarm deodorants, and this led to an elevated aluminum content in the brain as well as in the serum. Airborne aluminum is capable of entering the brain through the olfactory neurons located in the roof of the nasal cavity. The axons of these sensory neurons form nerve bundles that pass through the cribriform plate of the ethmoid bone and synapse at the olfactory bulb. From here, there are projections to the olfactory cortex, and other cortical areas, including the hippocampus. Absorption of aluminum from the olfactory pathway has been shown in rats

exposed to aluminum acetylacetonate to lead to accumulation of Al in the pons–medulla, olfactory bulb, and hippocampus, implying transneuronal migration.

Aluminum and Neurological Disease

Dialysis Encephalopathy

Acute exposure to aluminum can undoubtedly cause clinical neurotoxicity. Aluminum salts appear to be the cause of encephalopathy in uremic patients receiving chronic hemodialysis. In the past, such patients were routinely treated with aluminum-containing phosphate binders. In brain gray matter from a group of uremic patients on dialysis who died of neurologic syndrome, levels of aluminum were 25 ppm, compared to 6.5 ppm in a group of patients on dialysis who died of other causes, and 2.2 ppm in controls. Recovery from dialysis encephalopathy has been facilitated by application of an Al chelator. Demonstration of improved neurologic function following cessation of aluminum exposure further implicates Al as the cause of the dialysis encephalopathy syndrome. Other clinical procedures, such as bladder irrigation with 1% alum (a double salt of potassium and aluminum sulfate), can also cause symptoms of encephalopathy associated with elevated serum aluminum levels. A case study of a renal failure patient who developed Al-induced encephalopathy after using phosphate-binding Al-hydroxy gels for a prolonged period found increased proliferation of microglia and astrocytes upon postmortem analysis.

Evidence of Aluminum Neurotoxicity following Occupational Exposures

The development of an encephalopathy, characterized by cognitive deficits, incoordination, tremor, and spinocerebellar degeneration, among workers in the aluminum industry is further evidence that exposure to the metal can be profoundly deleterious. 'McIntyre Powder' (finely ground aluminum and aluminum oxide) was used as a prophylactic agent against silicotic lung disease between 1944 and 1979 in miners in northern Ontario. A morbidity prevalence study revealed that miners exposed to aluminum performed less well on cognitive examinations compared to unexposed workers. Furthermore, the likelihood of scores being in the impaired range increased with duration of exposure.

Other Studies Suggesting that Aluminum Salts Can Be Neurotoxic, and the Safety of Alum-Containing Vaccines

Prolonged feeding of infants with aluminum-containing intravenous solutions has led to impaired neurological function as judged by a reduction in the Mental Development Index. Treatment of aluminum-related bone disease with deferoxamine, which mobilizes bone Al and thus elevates serum Al^{3+}, has been reported to precipitate dementia.

Alum is present as an adjuvant in most vaccines, and abnormal neurological symptoms have been observed in several patients receiving intramuscular injections of Al-containing vaccines. The World Health Organization Vaccine Safety Advisory Committee has recognized that there may be a subset of predisposed individuals who are sensitive to Al-containing adjuvants. Macrophagic myofasciculitis secondary to intramuscular injection of aluminum hydroxide-containing vaccines shows both long-term persistence of aluminum hydroxide and an ongoing local immune reaction. These clinical findings are paralleled in several animal systems in which systemically administered aluminum can led to behavioral deficits, including incoordination and cognitive and morphological changes in the central nervous system (CNS) of treated animals. Since alum-containing vaccines have been in widespread use for a long time, the possibility of long-term hazard resulting from this is generally considered low.

Alzheimer's Disease

The presence of excess aluminum in the brains of patients with Alzheimer's disease (AD) has been described, but this issue remains controversial, since several conflicting reports exist. Thus the question as to whether the metal may play a role in AD and in other less common neurological disorders, such as the Guamanian parkinsonism–amyotrophic lateral sclerosis (ALS) constellation and Hallervorden–Spatz disease, is unresolved. However, chelation therapy, in order to reduce the aluminum burden in Alzheimer patients, has been reported as beneficial, and new Al-specific chelators for potential use in AD treatment have recently been developed.

There is a strong epidemiological relation between extended exposure to aluminum in drinking water and the incidence of AD. Increasing numbers of reports have related the aluminum content of drinking water with increased risk of developing AD. Residing in areas where aluminum concentrations in municipal drinking water are $100\,mg\,l^{-1}$ or greater appears to elevate AD incidence, and a dose–response correlation between increasing concentration of Al in drinking water and a higher risk of developing AD has been described in studies emanating from the USA, Canada, and France. A comprehensive literature survey has found 13 reports concerning a significant association between residing in areas where aluminum concentrations in the municipal drinking water are high and an increase in the incidence of AD, and a meta-analysis,

integrating these results, reveals that this association is significant.

Correlative changes are never sufficient to imply causation. Proposals have been made that aluminum entry into the brain is a secondary epiphenomenon, consequent to damage to the blood–brain barrier. However, the finding that dialysis encephalopathy can be treated with deferoxamine with good results-suggests that Al is directly neurotoxic. Thus a causal relationship may exist between circulating Al^{3+} and dementia. Epidemiological findings are strengthened when confirmed by use of an animal model in which the number of potentially confounding variables can be minimized.

Mechanisms of Neurotoxicity

While Al exposure has repeatedly been shown to be correlated to neurodegenerative changes, the mechanisms that underlie this remain unresolved. There is considerable evidence that Al may accelerate the progression of two major and interrelated processes associated with neural senescence. These are oxidative stress and increased levels of basal unprovoked inflammatory activity within the brain.

Oxidative Stress

In both biological and nonbiological isolated systems, aluminum can potentiate production of reactive oxygen species (ROS) by the prooxidant metals iron (Fe) and copper (Cu). In view of the inability of aluminum ions to undergo the valence transitions, or to have a strong affinity for sulfhydryl groups, properties that characterize metals which have the potential to promote excess prooxidant activity within cells, this promotion of ROS generation by Al is puzzling. While the ability of aluminum to induce the generation of free radicals has been reported many times, this seems to depend on the concurrent presence of a transition metal such as copper or iron. These are essential elements and are present in most cell compartments. The mechanism by which this interaction occurs may be by the binding of these transition metals to the surface of colloidal aluminum.

Transition metals can be present at the solid–liquid interfaces of particles, due to their being absorbed onto the surface of insoluble particles, and under these conditions, their prooxidant nature can be enhanced. The ability of Fe and Cu to promote oxidative events based on their redox valence fluctuation is intensified by their incomplete sequestration. Particulates and colloidal compounds (including aggregated amyloid beta peptide) may partially complex prooxidant metals, and thus allow them to participate in Fenton reactions for a prolonged period. This has been reported for several mineral particulates. For example, silica that appears to be free of iron can promote ROS production, but this is prevented by acid washing. This illustrates that very low amounts of trace metals bound to particulate surfaces can initiate ROS generation. High molecular weight aluminum matrices are likely to have extensive solid–liquid interface surface area and could present an appropriate site for iron activation. The neurological hazard of aluminum may then be in part due to its ability to increase the rate of generation of ROS initiated by transition metals.

Other means by which Al may promote cellular ROS generation exist. Exposure of isolated cells of glial, but not neuronal, origin to aluminum sulfate increases formation of prooxidant species. It is possible that the glial-specific aluminum-induced increase in oxidative parameters is due to glial activation by extracellular aluminum complexes. The potentially prooxidant role of Al^{3+} is especially relevant in view of the reported presence of excess oxidative activity, both in the aged brain and, more specifically, in those regions of the Alzheimer's brain showing the most pronounced pathological changes. Iron and aluminum are both present at high concentrations in neurofibrillary tangles and may cooperate in the promotion of physiological aggregation of the amyloid beta peptide.

Inflammation

Elevated levels of intrinsic inflammation are associated with neural aging, and this is exacerbated in several neurodegenerative diseases. In age-related neurodegenerative disorders, such as AD and Parkinson's disease (PD), the further enhancement of age-related inflammatory processes is thought to significantly contribute to pathogenic events. The number of activated astrocytes is increased in AD and these are associated with senile plaques and with cerebral microvessels. In the hippocampus of AD patients, there is an upregulation of proinflammatory genes, and levels of cytokines are elevated in the brain as well as in cerebrospinal fluid and plasma of AD patients. AD is characterized by brain depositions of the toxic amyloid beta peptide (Aβ), which is generated from amyloid precursor protein (APP). In the brain of AD patients, reactive microglia, producing proinflammatory cytokines and acute-phase proteins, are associated with Aβ-containing neuritic plaques. Aluminum salts promote Aβ aggregation *in vitro*, and in the brains of transgenic mice overexpressing amyloid precursor protein, dietary exposure to Al can also exacerbate Aβ deposition and plaque formation.

It has recently been found that low levels of aluminum in drinking water of experimental animals

Figure 1 Microglial activation in the brain of aluminum lactate-exposed animals. (a) Control, (b) Al lactate (0.01 mM), (c) Al (0.1 mM); FC, frontal cortex; CC, corpus callosum; CN/P, caudate nucleus/putamen; GP, globus pallidus. 20× magnification.

elevates basal levels of inflammatory activity within the CNS. In this case, one of the initial events in the cascade leading to inflammatory responses is the activation of transcription factors. Levels of activated nuclear factor-kappa B (NF-κB) were increased in brains of mice treated with Al lactate (0.01, 0.1, and 1 mM) in drinking water for 10 weeks. Several inflammation-associated cytokines, such as tumor necrosis factor-α (TNF-α) and interleukin-1α (IL-1α), were increased in the brains of treated animals. The mRNA for TNF-α was also upregulated. Even the lowest concentration of Al tested increased inflammatory processes in the brain of mice. This level of Al is in the range present in some water supplies. The prevalence of AD is heightened in regions where similar concentrations of the metal are present in residential drinking water.

A pronounced increase in both astrocyte proliferation and microglial activation was observed in mice receiving extended treatment with low levels of Al in water, and these changes were marked in the striatum and globus pallidus as well as the cortex (Figure 1). Since no parallel changes in transcription factors or cytokine levels were observed in the serum or liver of treated animals, the proinflammatory effects of the metal may be relatively selective to nervous tissue. Although oxidative and inflammatory events are independent phenomena, they are related. In addition to the aforementioned indices of inflammation, rates of lipid peroxidation and levels of neuronal nitric oxide synthase (nNOS) were elevated after exposure to 100 μM Al.

In isolated systems, aluminum is able to promote aggregation of Aβ peptides, paired helical filaments, tubulin, and other proteins into insoluble complexes that resist proteolysis, but this has never been demonstrated in animals treated systemically with aluminum. However, direct intracerebral administration of aluminum salts can lead to accumulation of amyloid precursor protein. Aluminum-induced tau protein aggregates (neurofibrillary tangles) in the rabbit can be partially reversed by removal of aluminum with deferoxamine, and aluminum-induced proteinaceous tangles can be reversed by removal of aluminum with silicates.

The possibility that insoluble aluminum complexes may induce glial activation and macrophage activity is supported by the observation that, in rats intracerebroventricularly injected with Al^{3+}, complexes of the metal accumulate largely in the striatum, and this is accompanied by gliosis. Such findings obtained from study of Al-exposed humans or experimental animals can be supported by parallel experiments using isolated systems Treatment of cultured cells derived from glioblastoma, but not neuroblastoma, with aluminum lactate increases cell proliferation, cytokine secretion, and NF-κB activation. This suggests that Al effects on the brain are not systemically mediated but are consequent to activation of the pathway leading from receptors at the cell surface, to phosphorylation of transcription factor-activating kinases, and thence to augmentation of expression of proinflammatory genes (Figure 2).

Conclusions

As the median age of the population rises, protection of the performance of the brain during senescence becomes increasingly relevant. Thus it is important to identify those environmental factors which may accelerate undesirable changes in brain function. Since cortical Al^{3+} levels increase with age, a role for aluminum in the progression of brain aging is possible. In addition, the incidence of age-related

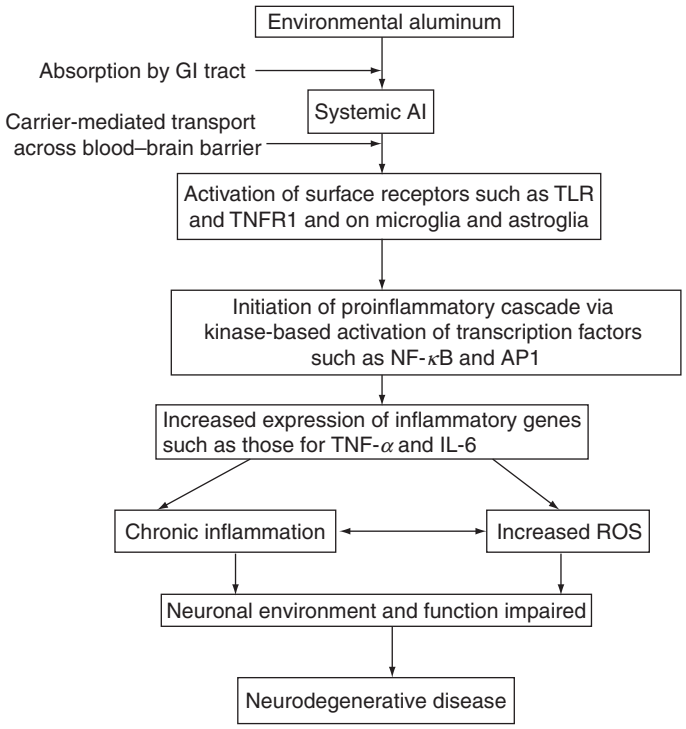

Figure 2 Proposed trajectory of aluminum (Al)-induced promotion of inflammatory processes within the central nervous system. GI, gastrointestinal; TLR, Toll-like receptor; TNFR1, tumor necrosis factor receptor 1; TNF-α, tumor necrosis factor-α; NF-κB, nuclear factor-kappa B; IL-6, interleukin-6; ROS, reactive oxygen species.

neurological diseases can be expected to rise with increasing longevity. The etiology of most neurodegenerative disorders is as yet unknown, but generally may not involve a primarily genetic component.

Aluminum compounds are undoubtedly a factor in the neurological sequelae of dialysis encephalopathy. The role of aluminum in brain aging and in chronic neurological disease is more uncertain. The proposal that excess levels of aluminum within the brain may contribute to the pathogenesis of AD has been extensively studied but remains controversial. This is also true of several other neurodegenerative diseases for which an intracerebral accumulation of aluminum has been reported. While many epidemiological correlations between Al exposure and neurological malfunction have been made, a clear mechanistic pathway that would account for this association is gradually emerging.

In the past, the problems encountered in making a definitive association between Al and CNS disorders have included the unavailability of a convincing aluminum-based animal model relevant to a specific CNS disease and absence of a clear theoretical and mechanistic framework for understanding how Al compounds could be deleterious in the CNS. The discussions herein were intended to illustrate that both of these limitations are likely to be overcome in the near future.

See also: Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Brain Composition: Age-Related Changes; Inflammation in Neurodegenerative Disease and Injury; Lipids and Membranes in Brain Aging; Metal Accumulation During Aging; Neurotoxins and their Neurotoxicology; Oxidative Damage in Neurodegeneration and Injury.

Further Reading

Becaria A, Campbell A, and Bondy SC (2004) Aluminum as a toxicant. *Toxicology and Industrial Health* 18: 309–320.

Campbell A, Becaria A, Lahiri DK, et al. (2004) Chronic exposure to aluminum in drinking water increases inflammatory parameters selectively in the brain. *Journal of Neuroscience Research* 75: 565–672.

Exley C (ed.) (2001) *Aluminum in Alzheimer's Disease*. New York, NY: Elsevier.

Flaten TP (2001) Aluminium as a risk factor in Alzheimer's disease, with emphasis on drinking water. *Brain Research Bulletin* 55: 187–196.

Miu AC and Benga O (eds.) (2006) Metal species and Alzheimer disease, *Journal of Alzheimer's Disease* 10: 133–341.

Perl DP and Moalem S (2006) Aluminum and Alzheimer's disease, a personal perspective after 25 years. *Journal of Alzheimer's Disease* 9(supplement 3): 291–300.

Yokel RA and McNamara PJ (2001) Aluminium toxicokinetics: An updated minireview. *Pharmacology and Toxicology* 88: 159–167.

Yang EY, Guo-Ross SX, and Bondy SC (1999) The stabilization of ferrous iron by a toxic β-amyloid fragment and by an aluminum salt. *Brain Research* 799: 91–96.

Alzheimer's Disease: An Overview

P I Moreira, University of Coimbra, Coimbra, Portugal
X Zhu and M A Smith, Case Western Reserve University, Cleveland, OH, USA
G Perry, University of Texas at San Antonio, San Antonio, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Major Neuropathologic Hallmarks

The distinctive brain lesions, senile plaques (SPs) and neurofibrillary tangles (NFTs), used by Alois Alzheimer together with the clinical deficits to describe Alzheimer's disease (AD), are still used today as the defining features for diagnosis (Figure 1). In addition to these striking changes, there is variable cerebral cortical atrophy, particularly of the temporal and frontal lobes, and associated ventricular dilation, both of which are consequences of the neuron loss and astrocyte proliferation in affected regions.

Both NFTs and SPs are found in normal aged persons, but it is their quantitative increase that defines the pathologic diagnosis of AD. NFTs consist of abnormal 20 nm helical filaments with an 80 nm half-periodicity, termed paired helical filaments, and 12–15 nm straight filaments (Figure 2). NFTs usually occur in large numbers in the brain of a patient with AD, particularly in the entorhinal cortex; hippocampus; amygdala; association cortices of the frontal, temporal, and parietal lobes; and certain subcortical nuclei that project to these regions. The subunit protein of the paired helical filaments is the microtubule-associated protein tau τ. Paired helical filaments are not limited to the tangles found in the neuronal cell bodies but also occur in smaller bundles in many of the dystrophic neurites present around the amyloid plaques. Biochemical studies reveal that the tau protein present in paired helical filaments is a hyperphosphorylated, insoluble form of this normally highly soluble protein. The insoluble tau aggregates in the tangles are often complexed with ubiquitin, a feature they share with numerous other intraneuronal protein inclusions in etiologically diverse disorders, such as Parkinson's disease and diffuse Lewy body dementia. If this complexing with ubiquitin represents an attempt by the neuron to mark the altered tau proteins for degradation by the proteasome, it seems to be largely without benefit for the patient.

The amyloid-β (Aβ) peptides of SPs consist of 7–10 nm helical filaments (Figure 2) that share with paired helical filaments the ability to bind the dye Congo red and appear birefringent green when viewed under cross-polarized light, a property of β-pleated molecular sheets. SPs (or neuritc plaques) contain extracellular deposits of the Aβ42/40 (explained below) surrounded by dystrophic neurites (axons and dendrites), activated microglia, and reactive astrocytes. A large portion of the Aβ in these neuritic plaques is in the form of insoluble amyloid fibrils, but these are intermixed with a poorly defined array of nonfibrillar forms of the protein. Once protein sequencing established that Aβ was the subunit of fibrillar amyloid, Aβ immunohistochemistry revealed several deposits in brains of patients with AD that lacked the dystrophic neurites and altered glia that characterize the neuritic plaques. Such plain Aβ deposits, referred to as 'diffuse' plaques, exist mostly in a nonfibrillar (that is, 'preamyloid') form. The diffuse deposits are composed of Aβ42, which is far more prone to aggregation than the slightly shorter and less hydrophobic Aβ40. In healthy individuals, Aβ40 and Aβ42 make up 90% and about 10%, respectively, of the Aβ peptides that are normally produced by brain cells throughout life. Note that Aβ plaques do not occur simply in these two extreme forms (diffuse and neuritic) but rather as a continuum in which mixtures of nonfibrillar and fibrillar forms of the peptide can be associated with varying degrees of surrounding neuritic and glial alteration.

Neurotransmitter Deficits

Since the demonstration of a 90% drop in cortical choline acetyltransferase activity, there has been considerable interest in understanding whether neurotransmitter replacement therapy would be beneficial in the treatment of AD. The 'cholinergic hypothesis' states that decreased cholinergic transmission plays a major role in the expression of cognitive, functional, and possibly behavioral symptoms in AD. The cholinergic hypothesis rests on pathological, biochemical, and pharmacological observations. Cholinergic neurons in the ventral forebrain are depleted; many of those that remain contain NFTs. As a result of these pathological changes, there are decreases in biochemical indices of cholinergic function in neocortex and hippocampus that correlate with dementia severity. The hypothesis is further supported by an extensive literature from pharmacological studies using cholinergic agonists and antagonists, ablative lesions of the cholinergic pathway, and transgenic animal models that emphasize the close connection between cognition and the cholinergic neurotransmission. It has also been proposed that the cholinergic deficit plays a role not only in the cognitive symptoms

but also in the behavioral changes observed in AD. A limitation of the cholinergic hypothesis is the lack of cholinergic deficit observed in early stages of AD or in patients with mild cognitive impairment.

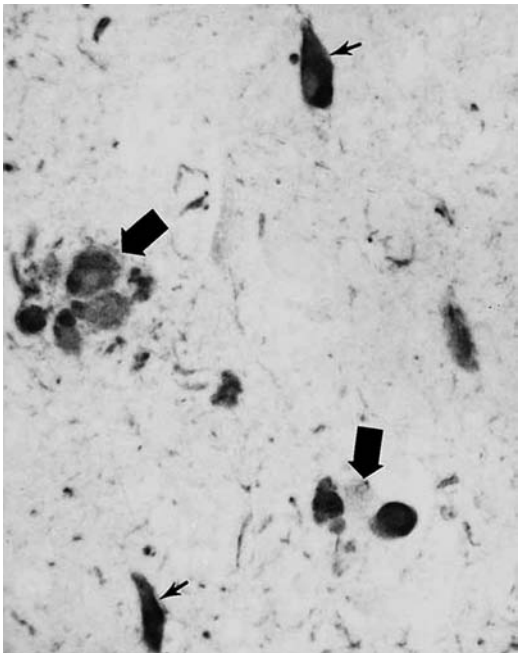

Figure 1 Characteristic pathologic lesions of Alzheimer's disease, neurofibrillary tangles (small arrows), and senile plaques (large arrows) are readily immunostained with antibodies to tau. (×760)

Other neurotransmitters may also be important in AD. For example, levels of three important neurotransmitters, serotonin, somatostatin, and noradrenaline, are reduced in the brains of some Alzheimer's patients. It has been suggested that these abnormalities are related to sensory disturbances and aggressive behavior. However, most neurotransmitter research in dementia continues to focus on acetylcholine because of its close ties to memory and reasoning.

Genetics

Autosomal dominant inheritance of mutations in the amyloid-β protein precursor (AβPP), PSEN1, and PSEN2 genes localized in chromosomes 21, 14, and 1, respectively, are responsible for familial early-onset AD (FAD). PSEN1 and PSEN2 encode for homologous polytopic membrane proteins, termed presenilin 1 and 2 (PS1 and PS2). To date, 15 missense mutations in AβPP, 199 in PS1, and 11 in PS2 have been reported to cause FAD. FAD mutations in PSEN1 cause most aggressive forms of AD, in some cases with onset younger than 30 years. Individuals with trisomy 21 (Down syndrome) have an extra copy of the AβPP gene and develop AD pathology as early as 20 years of age. Even though FAD-linked mutations in AβPP account for less than 5% of total AD cases, these autosomal dominant mutations are highly penetrant, and the clinical and pathologic symptoms of individuals with FAD mutations are nearly identical to those of patients

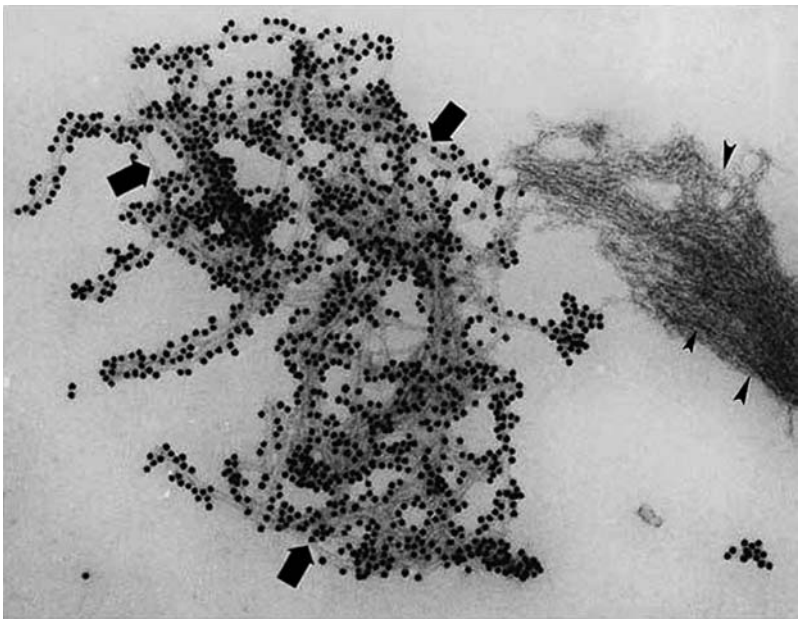

Figure 2 Structural and antigenic differences between paired helical filaments (large arrows) and amyloid-β (Aβ) filaments (arrowheads) are readily apparent in this negatively stained preparation. Heavy subunit of neurofilaments is localized by colloidal gold to paired helical filaments but not to Aβ filaments. (×105 000)

with late-onset sporadic AD. Biochemically, FAD-associated mutants in PSEN1 and PSEN2 lead to selective increase in the levels of Aβ42 species, which readily aggregate *in vitro* and are the initial Aβ species deposited in the brains of individuals with AD and Down syndrome. FAD-linked mutations in AβPP either increase Aβ42 or overall Aβ production or generate highly fibrillogenic Aβ variants. In addition to early-onset FAD mutations, it was also discovered that the presence of ε4 allele of the apolipoprotein E (APOE) gene is a risk factor for familial late-onset AD. APOE plays a vital role in the metabolism and clearance of Aβ along with α2-macroglobulin (α2M) and low-density lipoprotein receptor. The biochemical outcome of harboring the APOE ε4 allele includes increased Aβ aggregation and decreased Aβ clearance.

The number of genes involving distinct proteins suggests that the pathologic and clinical entity we call AD is not unique to abnormalities in a single gene locus or etiology. Significantly, the vast majority of AD is not linked to any established genetic abnormality.

Aβ and Its Protein Precursor

Understanding of the pathogenesis of Aβ deposition was greatly advanced by the sequencing of Aβ by G Glenner and C Wong in 1983. Subsequent cloning showed that Aβ is a 39–42-amino-acid fragment of a larger 695–770-amino-acid, membrane-spanning glycoprotein termed Aβ precursor protein (AβPP). The gene for AβPP resides as a single copy on the long arm of chromosome 21 and forms the locus for some cases of familial AD.

Aβ is proteolytically released from a large type 1 membrane glycoprotein of unknown function, the AβPP, via sequential cleavages by two aspartyl proteases, referred to as the β- and γ-secretases. The Aβ region of AβPP comprises the 28 residues just outside the single transmembrane domain, plus the first 12–14 residues of that buried domain. On this basis, Aβ was originally assumed to arise only under pathological circumstances, in that the second cleavage was thought to require some kind of prior membrane disruption to allow access of γ-secretase and a water molecule to the otherwise intramembranous region. This concept was disproved in 1992, when Aβ was shown to be constitutively released from AβPP and secreted by mammalian cells throughout life and thus occur normally in plasma and cerebrospinal fluid. This discovery enabled the dynamic study of Aβ production in cell culture and animal models, including examination of the effects of AD-causing genetic mutations. Moreover, high-throughput screening could now be conducted on cultured cells to identify Aβ-lowering compounds and determine their mechanism. Most AβPP molecules that undergo secretory processing are cleaved by α-secretase, rather than β-secretase, near the middle of the Aβ region. This releases the large, soluble ectodomain (APPs-α) into the medium and allows the resultant 83-residue, membrane-retained, C-terminal fragment (C83) to be cleaved by γ-secretase, generating the small p3 peptide. α-secretase acts on AβPP molecules at the cell surface although some processing also occurs in intracellular secretory compartments. The precise subcellular loci of the β- and γ-secretase cleavages are unclear but likely include early, recycling endosomes. The functional consequences of the proteolytic processing of AβPP remain ill-defined. The current leading hypothesis is that cleavage by α-secretase followed by γ-secretase enables the release of the AβPP intracellular domain (AICD) into the nucleus, where it may participate in transcriptional signaling. The APPs-α derivative secreted as a result of this processing appears to have distinct extracellular functions. For example, those APPs-α isoforms that contain an alternatively spliced Kunitz protease inhibitor domain function as serine protease inhibitors, including by inhibiting of Factor XIa in the coagulation cascade.

The importance of AβPP in the primary etiology of AD is fairly well established, with FAD associated with a number of mutations in the AβPP gene on chromosome 21. Some of the mutations in AβPP leading to AD have been related to AβPP processing, yielding more Aβ or the longer form of Aβ, which has a greater propensity to form Aβ fibrils. Additional support for the importance of AβPP in the pathogenesis of AD comes from Down's syndrome, in which an extra copy of chromosome 21 leads at midlife to a spectrum of pathologic changes similar to those found in AD.

Cytoskeletal Abnormalities

A fundamental process in the pathogenesis of AD is the breakdown of the cytoskeleton. The main component of NFTs is the paired helical filaments (PHF), which are mainly comprised of the protein tau in an abnormally phosphorylated status. Tau filaments accumulate in dystrophic neurites as fine neuropil threads or as bundles of PHF in neuronal bodies forming the NFTs, which become extracellular ghost tangles after the death of the neuron. The severity of dementia has been correlated with accumulation of NFTs in different brain regions, while with SP, such correlation has not been demonstrated. In AD, tau binds with lower affinity to microtubules, and it self-aggregates into aberrant structures, probably helped by other molecules. As a consequence of hyperphosphorylation, tau shows a loss of microtubule-binding capacity and is accumulated in neuronal bodies.

The finding that neurons are primarily responsible for AβPP production and also contain NFTs highlights a key issue in AD, the relationship of Aβ to NFTs. The application of monoclonal antibodies, immunoelectron microscopy, and antibody affinity techniques has shown that microtubule-associated protein tau and the heavy subunit of neurofilaments are major components of NFTs. Recent studies demonstrated a direct high-affinity interaction between tau and AβPP. More studies are required to understand the pathologic significance of tau-AβPP interaction in Aβ deposition.

Sparing solubility and heterogeneity of enriched fractions have hampered efforts to define NFTs quantitatively, and there is now a considerable effort to understand how the two identified posttranslational modifications of NFTs, phosphorylation and glycation, mediate the transformation of soluble tau into insoluble paired helical filaments. Indeed, although increased phosphorylation would lead to microtubule instability – a key hallmark of AD – phosphorylation does not mediate paired helical filament insolubility. Conversely, oxidation of tau confers the same solubility properties as paired helical filaments. Therefore, it is likely that oxidative cross-links play a role in NFT insolubility. Significantly, oxidation is one of the earliest changes in the disease that leads to neuronal dysfunction. It is not surprising, therefore, that therapeutic efforts to reduce oxidative stress slow disease progression and/or decrease the incidence of the disease.

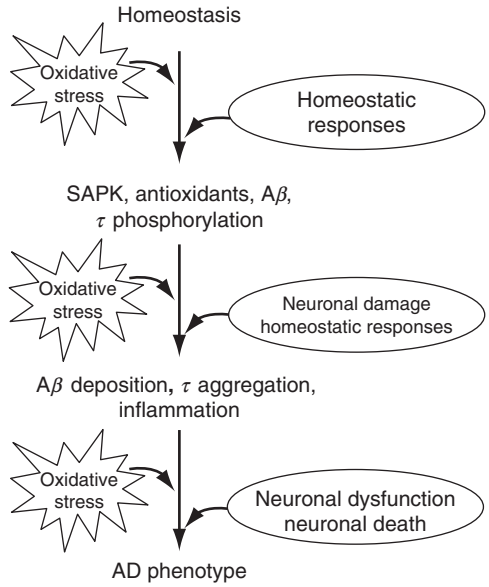

Figure 3 The sequence of events in Alzheimer's disease (AD). Oxidative stress, the earliest event occurring in AD, initially leads neurons to activate neuronal defenses, including stress-activated protein kinases (SAPK), tau (τ) phosphorylation, and amyloid-β (Aβ), to maintain homeostatic balance. However, given the chronic and insidious nature of oxidative stress, progression of the disease through Aβ deposition, tau aggregation, and inflammation overwhelms initial compensatory mechanisms and culminates in neuronal dysfunction and death, that is, the AD phenotype.

Oxidative Stress in AD

It has been shown that increased oxidative damage is a prominent and early feature of vulnerable neurons in AD. Over the past decade, an oxidative stress-related modification of macromolecules has been described in association with the susceptible neurons of AD. Such modifications include advanced glycation end products, nitration, lipid peroxidation adduction products, carbonyl-modified neurofilament protein, and free carbonyls, as well as glycation and glycoxidation products. Levels of these markers are initially elevated following some unknown triggering neuronal event, but these levels soon decrease as the disease progresses to advanced AD. Together these findings suggest that increased oxidative damage is not the terminal sequelae of the disease but instead plays an initial role. They also suggest that damage does not mark further destruction by reactive species and is instead marked by a broad array of increased cellular defenses. It can be argued that in AD, these defenses are responsible for the reduction of damage if we view AD in isolation. However, when seen in the context of other conditions in which reactive oxygen and nitrogen species are involved and damage is either limited or absent, such as Parkinson's disease, this result raises the question of whether oxidative damage noted in AD may be better thought of as homeostatic, that is, that oxidative damage could initiate signal transduction pathways to manipulate cellular responses to stress, which are characterized by increased levels of reactive oxygen and nitrogen species. Furthermore, there is evidence that in the first stage of AD development, Aβ deposition and hyperphosphorylated tau function as compensatory responses and downstream adaptations to ensure that neuronal cells do not succumb to oxidative damage (Figure 3).

Conclusions

The numerous genetic abnormalities as well as the high prevalence of sporadic disease suggest that AD has a defined pathogenesis stemming from numerous etiologies. Consequently, AD is a syndrome now defined by the most striking aspects of that pathogenesis, that is, NFTs and SP. Since oxidative damage is the earliest described cytopathological abnormality that occurs in vulnerable brain regions and selective neuronal populations, it is extremely important to decipher the causes and consequences of neuronal

oxidative stress. Clarification of mechanistic viewpoints concerning AD pathophysiology may provide insights into efficacious therapeutics. In particular, investigation must uncover what the relationships are between increased oxidative stress and other facets of the disease such as regionally selective neuronal degeneration, Aβ deposition as SP, and tau phosphorylation and aggregation as NFTs.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: Neurodegeneration; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: Transgenic Mouse Models; Alzheimer's Disease: MRI Studies; Animal Models of Alzheimer's Disease; Axonal Transport and Alzheimer's Disease; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects.

Further Reading

Glenner GG and Wong CW (1984) Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochemical and Biophysical Research Communications* 120: 885–890.

Hendrie HC (1998) Epidemiology of dementia and Alzheimer's disease. *American Journal of Geriatric Psychiatry* 6: S3–S18.

Hernandez F, Engel T, Gomez-Ramos A, et al. (2005) Characterization of Alzheimer paired helical filaments by electron microscopy. *Microscopy Research and Technique* 67: 121–125.

Katzman R (1993) Clinical and epidemiological aspects of Alzheimer disease. *Clinical Neuroscience* 1: 165–170.

Khachaturian ZS (1985) Diagnosis of Alzheimer's disease. *Archives of Neurology* 42: 1097–1105.

Khachaturian ZS and Mesulam M-M (eds.) (2000) *Special Issue: Alzheimer's disease: A Compendium of Current Theories. Annals of the New York Academy of Sciences* 924.

Levy-Lahad E, Wasco W, Poorkaj P, et al. (1995) Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* 269: 973–977.

Moreira PI, Honda K, Zhu X, et al. (2006) Brain and brawn: Parallels in oxidative strength. *Neurology* 66: S97–S101.

Perry G (ed.) (1993) Alzheimer's Disease. *Clinical Neuroscience* 1: 163–224.

Perry G and Smith MA (1993) Senile plaques and neurofibrillary tangles: What role do they play in Alzheimer disease? *Clinical Neuroscience* 1: 199–203.

Post SG (2000) *The Moral Challenge of Alzheimer Disease: Ethical Issues from Diagnosis to Dying*, 2nd edn. Baltimore: Johns Hopkins University Press.

Roses AD (1994) Apolipoprotein E affects the rate of Alzheimer disease expression: β-Amyloid burden is a secondary consequence dependent on ApoE genotype and duration of disease. *Journal of Neuropathology and Experimental Neurology* 53: 429–437.

Selkoe DJ (1994) Normal and abnormal biology of the β-amyloid precursor protein. *Annual Review of Neuroscience* 17: 489–517.

Selkoe DJ (1994) Alzheimer's disease: A central role for amyloid. *Journal of Neuropathology and Experimental Neurology* 53: 438–447.

Selkoe DJ (2004) American College of Physicians, and American Physiological Society. Alzheimer disease: Mechanistic understanding predicts novel therapies. *Annals of Internal Medicine* 140: 627–638.

Sherrington R, Rogaev EI, Liang Y, et al. (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* 375: 754–760.

Smith MA, Sayre LM, Monnier VM, et al. (1995) Radical ageing in Alzheimer's disease. *Trends in Neurosciences* 18: 172–176.

Smith MA, Siedlak SL, Richey PL, et al. (1995) Tau protein directly interacts with the amyloid β-protein precursor: Implications for Alzheimer's disease. *Nature Medicine* 1: 365–369.

Tanzi RE and Bertram L (2005) Twenty years of the Alzheimer's disease amyloid hypothesis: A genetic perspective. *Cell* 20: 545–555.

Relevant Websites

http://www.alz.org – Alzheimer Association.
http://www.alzforum.org – Alzheimer Research Forum.
http://www.molgen.ua.ac.be – Molecular Genetics Department, VIB, The Flanders Institute for Biotechnology.
http://www.nih.gov/nia – National Institute on Aging.

Alzheimer's Disease: Molecular Genetics

R Sherrington, Unité de Recherche en Neuroscience, Ste-Foy, QC, Canada
P H St. George-Hyslop, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

Introduction

Alzheimer's disease (AD) is the most common form of degenerative dementia of the human central nervous system. It is defined clinically as a progressive loss of cognitive function with the onset of a slowly progressive impairment of memory during mid- to late adult life. The neuropathologic hallmarks of this disease include amyloid deposits, neurofibrillary tangles, astrocytic gliosis, and reductions in the number of neurons and synapses in many areas of the brain, but especially from the cerebral cortex and the hippocampus.

The etiology of AD is complex. Multiple epidemiologic studies have proposed a number of potential risk factors, including environmental (head trauma, smoking, and exposure to heavy metals such as aluminum), sociologic (depression and level of education), biologic (increasing age, hyperthyroidism, late maternal age), and family history (of Down syndrome or AD).

The repeated observation in multiple epidemiologic surveys that a positive family history is a strong risk factor for AD clearly suggest that genetic factors play a role in this disease. A simple autosomal dominant inheritance with age-dependent penetrance has been observed in many pedigrees, especially those with early age of onset. However, in many families multiply affected by AD, the pattern of inheritance is unclear and could fit one of several genetic models, including a high-frequency, low-penetrance single gene disorder or a multifactorial model in which several genes or nongenetic factors interact. Using the powerful tools of genetic linkage analysis, significant progress has been made in unraveling the genetic etiology. Four loci that play a role in the genetic susceptibility of AD have been identified, namely, genes on chromosomes 21 (β-amyloid precursor protein (βAPP)), 19 (apolipoprotein E (APOE)), 14 (presenilin 1 (PS1)), and 1 (presenilin 2 (PS2)).

β-Amyloid Precursor Protein

The βAPP gene was the first gene found to bear mutations capable of causing early-onset familial AD (FAD). The βAPP gene on chromosome 21 encodes an alternatively spliced single spanning transmembrane protein. The longer isoform, APP770 (770 amino acids), contains two exons, which encode a Kunitz protease inhibitor (KPI) domain and a 19-amino acid sequence homologous to the ox-2 antigen. The KPI-containing isoforms are expressed in most tissues. In contrast, the shorter isoform APP695 (695 amino acids), without the KPI domain, is predominantly expressed in the brain. The βAPP protein undergoes a series of endoproteolytic cleavages that give rise to Aβ peptide, a 4 kDa peptide resulting from proteolytic cleavages of the full-length protein from residues 672–711 that spans the last 28 residues of the extracellular domain and the first 14 residues of the transmembrane domain. One of these cleavages, which results from the action of a putative membrane-associated α-secretase, liberates the extracellular N-terminus of βAPP and is thus a nonamyloidogenic pathway, because this cleavage precludes the formation of Aβ peptide. The other cleavage pathway, which occurs in part in the endosomal-lysosomal compartment, involves the recently identified β-secretase (BACE) and the putative γ-secretases, which give rise to a series of peptides that contain the 40–42 amino acid Aβ peptide. Aβ peptides ending at residue 42 or 43 (long-tailed Aβ) are thought to be more fibrillogenic and more neurotoxic than Aβ ending at residue 40, which is the predominant isoform produced during normal metabolism of βAPP. The activity of these secretases, and especially g-secretase, giving rise to the more fibrillogenic and potentially neurotoxic long-tailed Ab1–42, appears to play an important role in the pathogenesis of AD. It has been suggested that processing of βAPP into Aβ and the subsequent accumulation of Aβ in the brain are central events in the pathogenesis of AD in both genetic and nongenetic forms. Although Aβ peptides are secreted by cells under normal physiologic conditions, how they induce neurodegeneration in AD is still unclear.

Five mutations in the βAPP gene (APP V7171, APP V717G, APP V717F, APP KM670/671NL, and APP A692G) can cause early-onset FAD. The mutations occur within or near the Ab peptide. Mutations in this gene account for only a small fraction of early-onset FAD cases (less than 0.1%). The function of the βAPP protein is unknown, although roles in cell adhesion, synaptic growth, and neural repair have been proposed. It has also been suggested that the βAPP may function as a G-coupled receptor linked through an interaction in the C-terminus to G_o protein.

The APOE Gene

The APOE gene is located on chromosome 19q13 and encodes the major apolipoprotein expressed in the central nervous system. APOE plays an important role in triglyceride-rich lipoprotein metabolism and regulation of cholesterol. It has also been suggested that APOE plays a role in repair, growth, and maintenance of myelin and axonal membranes. Studies implicating the APOE gene as an AD susceptibility locus derived from several intersecting lines of investigation (linkage analysis on late-onset AD families, APOE immunoreactivity in senile plaques and NFTs of patient with AD, and APOE binding to Aβ). Association studies revealed that inheritance of the APOE e4 allele (Cys112Arg) is associated with a copy number-dependent increase in risk of both sporadic and late-onset FAD and with a decrease in the age of onset in late-onset AD. Conversely, the e2 allele confers a protective effect against late-onset AD. However, it seems that the inheritance of the e4 allele (APOE4) may be neither necessary nor sufficient by itself to cause AD, and the predictive value of the APOE4 in presymptomatic people is unclear. The mechanism by which the inheritance of the APOE4 increases risk of AD and how this might be addressed therapeutically are still unclear. Moreover, it has been recently reported that APOE represents only 7–9% of the total variation in age at onset of AD, indicating that other loci may have greater importance than does APOE.

The PS Genes

Genetic linkage studies, using large numbers of pedigrees with early-onset autosomal dominant FAD, mapped a common locus (AD3) to chromosome 14q24.3. The AD3 locus is associated with a particularly aggressive form of early-onset (between 30 and 60 years of age) AD and accounts for up to 50% of early-onset FAD cases. Subsequent positional studies led to the isolation of a novel gene termed PS1. Mutational analysis of the PS1 gene have identified at least 70 different predominantly missense mutations in residues that are highly conserved in evolution. The PS1 is predicted to be an integral membrane protein with at least seven hydrophobic membrane-spanning domains (transmembrane domains (TM)) and two acidic hydrophilic domains located at the N-terminus and between TM6 and TM7. The amino acid sequence of the PS1 protein shows a weak homology to the *C. elegans* SPE-4 protein, which is thought to be involved in membrane budding and fusion events in a membrane-bound cytoplasmic organelle derived from the spermatocytes' Golgi network. Subsequently, a strong homology was found between PS1 and SEL-12, another *C. elegans* protein. SEL-12 facilitates signaling mechanisms in the LIN-12/Notch family of intercellular receptors, which are responsible for direct signal transmission from the cell surface to the nucleus during intercellular signal transduction specifying cell fate. These later data suggest that PS1 may have a role in the vertebrate Notch signaling pathway. This is further supported by the fact that PS1 knockout mice show developmental abnormalities similar to those of mice with targeted knockouts of the murine Notch1 gene.

Following the isolation of the PS1 gene causing the AD3 subtype of AD, a second gene, PS2, mapped to chromosome 1, was discovered on the basis of its strong nucleotide and amino acid sequence homology to PS1. The structural organization of the PS2 is very similar to that of PS1, including the transmembrane and acidic hydrophilic domains. These observations suggest that the PS1 and the PS2 genes are members of a gene family. Mutational analysis of the PS2 gene led to the discovery of six different mutations. The phenotype of these families shows a later onset than the phenotype associated with mutations in PS1 or βAPP genes (onset 50–70 years).

All the PS1 or PS2 mutations lead to significantly increased levels of secreted Aβ42 compared with wild-type controls. These observations suggest a dominant gain of function for the mutant PSs because even though the endogenous wild-type alleles are present, the level of Aβ42 is increased in cells expressing mutant cDNAs. How mutant presenilins alter APP processing is unknown. It has been proposed that PS1 (and probably PS2) is part of a complex that includes the recently discovered nicastrin, involved in the proteolytic cleavage of the C-terminal fragment of APP (mediated by γ-secretase). Other functions in intracellular signaling, suppression of apoptosis, and protein/membrane trafficking have been suggested for both PS1 and PS2.

Other AD Genes

Although four different genetic loci associated with inherited susceptibility to AD (βAPP, APOE, PS1, and PS2) have been identified, they account for only about half of the genetic forms of AD. Thus at least one and possibly several other genes remain to be identified. Several polymorphisms in different genes have been associated with AD. Recent findings suggest the presence of new AD genes on chromosomes 9, 10, and 12. However, the lack of replication of these associations in different AD pedigrees makes these results unclear at the moment.

Animal Models

Several transgenic murine model lines have been created using a variety of βAPP constructs and have met with varying degrees of success. The most successful model to date has been a transgenic line in which a mutant human βAPP minigene has been placed under the control of the platelet-derived growth factor receptor beta subunit promoter. This latter model demonstrates abundant Aβ deposition and synaptic pathology. Targeted null mutation (knockout) of the murine βAPP gene was created but is not very illuminating because it leads only to subtle phenotypes including minor weight loss, decreased locomotor activity, abnormal forelimb motor activity, and minor nonspecific degrees of reactive gliosis in the cortex.

Transgenic mice overexpressing wild-type or mutant PSs have been created. All mutants show an increase in Aβ42 production in the brain compared with wild-type. Despite the Aβ42 accumulation in the brain, these PS transgenic mice have not shown any major neuronal death. However, these mice showed reduced spontaneous alternation performance in a Y maze. PS1 knockout mice have also been generated. These knockout mice developed deformations of the caudal axial skeleton, hemorrhages in the central nervous system, and impairment in neurogenesis and usually did not survive beyond the first day after birth. These studies show that, at least, PS1 is required for proper formation of the axial skeleton and is involved in normal neurogenesis and survival of progenitor cells and neurons in specific brain subregions.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: Transgenic Mouse Models; Dementia; Dementia and Language; Gene Expression in Normal Aging Brain; Genomics of Brain Aging: Apolipoprotein E; Genomics of Brain Aging: Twin Studies.

Further Reading

Bertram L and Tanzi RE (2001) Dancing in the dark? The status of late-onset Alzheimer's disease genetics. *Journal of Molecular Neuroscience* 17(2): 127–136.

Farrer LA, Myers RH, Connor L, et al. (1991) Segregation analysis reveals evidence of a major gene for Alzheimer disease. *American Journal of Human Genetics* 48: 1026–1033.

Games D, Adams D, Alessandrini A, et al. (1995) Alzheimer-type neuropathology in transgenic mice overexpressing V717F beta-amyloid precursor protein. *Nature* 373: 523–527.

Goate AM, Chartier-Harlin MC, Mullan M, et al. (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer disease. *Nature* 349: 704–706.

Katzman R (1986) Alzheimer's disease. *New England Journal of Medicine* 314(15): 964–973.

Katzman R and Kawas C (1994) The epidemiology of dementia and Alzheimer disease. In: Terry RD, Katzman R, and Hick KL (eds.) *Alzheimer Disease*, pp. 105–122. New York: Raven Press.

Prince JA, Feuk L, Sawyer SL, et al. (2001) Lack of replication of association findings in complex disease: An analysis of 15 polymorphisms in prior candidate genes for sporadic Alzheimer's disease. *European Journal of Human Genetics* 9(6): 437–444.

Rogaev EI, Sherrington R, Rogaeva EA, et al. (1995) Familial Alzheimer's disease in kindreds with missense mutations in a novel gene on chromosome I related to the Alzheimer's disease type 3 gene. *Nature* 376: 775–778.

Saunders A, Strittmatter WJ, Schmechel S, et al. (1993) Association of apolipoprotein E allele e4 with late-onset familial and sporadic Alzheimer disease. *Neurology* 43: 1467–1472.

Selkoe DJ (1994) Normal and abnormal biology of β-amyloid precursor protein. *Annual Review of Neuroscience* 17: 489–517.

Selkoe DJ (2001) Alzheimer's disease: Genes, proteins, and therapy. *Physiological Reviews* 81(2): 741–766.

Sherrington R, Rogaev EI, Liang Y, et al. (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* 375: 754–760.

Warwick DE, Payami H, Nemens EJ, et al. (2000) The number of trait loci in late-onset Alzheimer disease. *American Journal of Human Genetics* 66(1): 196–204.

Relevant Website

http://www.alzforum.org – Alzheimer Research Forum.

Alzheimer's Disease: MRI Studies

P M Thompson and A W Toga, UCLA School of Medicine, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

This article was reproduced from the previous edition © 2004, Elsevier B.V.

Introduction

Alzheimer's disease (AD) is the leading cause of senile dementia, affecting 10% of those over age 65. The disease causes irreversible memory loss, behavioral and cognitive decline, personality changes, and a decreasing ability to cope with everyday life. Ironically, up to 30 years elapse between the onset of the cellular pathology that causes AD (amyloid plaques and neurofibrillary tangles in the brain) and the clinical changes that lead to diagnosis. To help understand how the disease emerges and progresses, imaging technology can be applied that is safe, repeatable, and widely available. Magnetic resonance imaging (MRI) scans of AD patients reveal profound anatomic changes: severe cortical and hippocampal atrophy, sulcal and ventricular enlargement, and reduced gray matter and white matter volume. These changes occur in a distinct spatial and temporal sequence, and correlate with cognitive and metabolic decline. If patients are scanned repeatedly with MRI as their disease progresses, dynamic maps can be reconstructed that reveal a shifting pattern of cortical changes. This spreading cortical atrophy mirrors the spread of the underlying pathology (as defined by tangle and amyloid plaque deposition). Repeat MRI scanning can monitor disease progression in individual patients and can evaluate how drugs oppose these changes. It can also clarify how anatomical deficits link with cognitive and behavioral deficits as they emerge in individuals and populations.

Impact

AD is a severe and growing public health crisis. The incidence of AD doubles every 5 years after age 60. It afflicts 1% of those aged 60 to 64 and 30% to 40% of those over 85. Without a cure, the number of AD victims will rise from 2.0 to 3.5 million now to an estimated 10 to 14 million by 2030. A number of promising AD treatments are now being developed. These range from acetylcholinesterase inhibitors, which ballast neurotransmitter function, to experimental vaccines, which directly attack the amyloid plaques that are a key element of AD pathology. Most therapeutic trials of new drugs in AD rely primarily on cognitive tests to determine efficacy. Neuroimaging, however, can be extremely beneficial in this research. It supplies a variety of biological markers that measure disease progression. With novel brain mapping techniques, the disease can be tracked as it spreads in the living brain. In Alzheimer's patients, MRI scans show prominent hippocampal atrophy. Diffuse tissue loss is also found in the medial temporal lobes. These deficits are progressive, and their magnitude correlates with cognitive decline. Because it is vital to detect the disease early, there is great interest in developing MRI measures that predict imminent transition to dementia, in healthy elderly subjects. For example, significant hippocampal volume deficits are found in subjects with mild memory impairments, who do not yet have dementia. These deficits shown on MRI can also help to predict how soon an elderly individual will develop AD. Reliable MRI predictors are especially valuable, as cholinergic drugs are most effective in the mildest phases, when widespread neuronal loss has not yet occurred.

Several neuroimaging measures can be used to characterize dementia. For instance, MRI scans can assess the integrity of medial temporal lobe structures involved in memory, such as the entorhinal cortex and hippocampus. The region and rate of atrophic brain changes can be measured as the disease progresses, as can the profile of cortical thinning and gray matter loss. The required three-dimensional MRI scans can be performed in approximately 10 min, on a conventional 1.5-Tesla scanner. Although MRI is not routinely used to diagnose AD, new techniques in brain image analysis can be applied to MRI scans to reveal how the disease emerges, and track how medications affect the disease process. MRI scans are often used in dementia research to (1) screen at-risk populations to find anatomical measures that might help predict each individual's likelihood of developing AD, (2) discriminate AD from normal aging and other dementias (such as frontotemporal and Lewy body dementias, which have distinct anatomic patterns), and (3) monitor disease progress and therapeutic response, gauging the effectiveness of drug treatments.

Brain Tissue Loss and Cognitive Decline

In the 1990s, MRI research in dementia focused on measuring medial temporal lobe structures. This was because AD pathology typically starts in the temporal cortex adjacent to the entorhinal cortex and quickly spreads to the entorhinal cortex before involving the hippocampus. This temporal lobe pathology persists for several years before spreading cortically to engulf

the rest of the temporal, frontal, and parietal lobes. A more recent trend in dementia research has been to move from cross-sectional studies to dynamic measures. Serial MRI scans (acquired from the same patients repeatedly over time) can provide much greater power to detect pathological atrophy, because they provide a baseline reference point to calculate change. Fox et al. found that AD patients lose brain tissue at a faster 'rate' than age-matched controls. Evaluated with MRI for 5–8 years, AD patients lost brain tissue at a median rate of 2.20% per year (range, 0.82–4.19) versus 0.24% per year in controls (range, −0.35–0.64). These rates correlated with the rate of cognitive decline, reflected by worsening performance on the Mini Mental Status Examination (MMSE). In a recent 52 week clinical trial of milameline (a muscarinic receptor agonist), Jack et al. noted that hippocampal volume, measured with MRI, also tracked cognitive decline. Perhaps the most prominent sign of AD seen on an MRI scan is that the lateral ventricles are often greatly enlarged. Bradley et al. measured the ratio of the ventricular volume to the total brain volume (the ventricle-to-brain ratio (VBR) in 39 elderly subjects scanned with serial MRIs over 3 to 6 month intervals. The VBR rate of change was 15.6% ± 2.8% (mean ± SD) per year for AD patients compared with 4.3% ± 1.1% per year in controls ($P < 0.001$). VBR did not separate groups when measured at only a single time point, supporting the value of longitudinal assessments. Power calculations revealed that 135 subjects would be needed in each arm of a placebo-controlled clinical trial if this measure of AD progression were to detect a 20% reduction in the excess rate of atrophy over 6 months, with 90% power.

Gray Matter Deficits

Brain changes in AD can also be visualized using three-dimensional maps. Figure 1 shows the spatial pattern of cortical gray matter loss in mild to moderate AD. This type of image is a composite map; it results from a sequence of image processing steps that compare scans of AD patients with matched healthy elderly subjects. With image analysis techniques, three-dimensional brain MRI scans can be split up

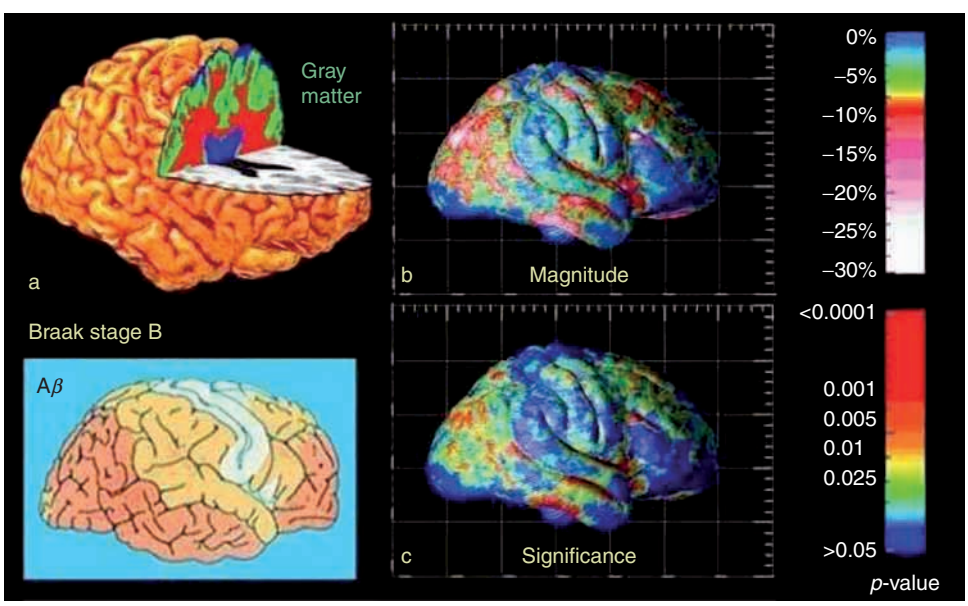

Figure 1 Gray matter deficits in early AD. Here the local amount of cortical gray matter (green colors, (a)) is compared across 26 patients with mild to moderate AD (age, 75.8 ± 1.7 years; MMSE score, 20.0 ± 0.9) and 20 matched elderly controls (age, 72.4 ± 1.3 years). At this stage of AD, 30% of the cortical gray matter has been lost in the temporoparietal regions (b). (c) Statistical significance of these deficits. The pattern of temporal lobe gray matter loss, seen on MRI, spatially matches the pattern of beta-amyloid (Aβ) deposition seen postmortem. The inset panel (Braak stage B) is adapted from data reported by Braak and Braak (1997). It shows regions with minimal (white), moderate (orange), and severe (red) Aβ deposition. Amyloid deposition and gray matter loss may not be synchronized, so these maps may represent different stages of AD; however, there is a clear spatial agreement in the severity of the deficits, between MRI and Aβ maps. Primary sensorimotor regions (white in the amyloid map), and the superior temporal gyri (blue colors in (c)) are spared relative to other temporal lobe gyri. These MRI patterns have been replicated in independent studies by Thompson PM, Mega MS, Woods RP, et al. (2001) Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas. *Cerebral Cortex* 11: 1–16, Baron JC, Chetelat G, Desgranges B, et al. (2001) *In vivo* mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease. *Neuroimage* 14: 298–309, and O'Brien JT, Paling S, Barber R, et al. (2001) Progressive brain atrophy on serial MRI in dementia with Lewy bodies, AD, and vascular dementia. *Neurology* 56: 1386–1388.

into regions representing gray matter, white matter, and cerebrospinal fluid. A measure is then computed that is related to the thickness of the cortical gray matter at each cortical location. Computer analyses then compare the amount of gray matter at each cortical location across subjects, while adjusting for potentially confounding factors, such as age and gender effects, and gyral patterning differences. Differences can be visualized locally in the form of color-coded statistical maps. These show how much gray matter volume is reduced in AD patients relative to healthy controls in each cortical region

Maps of Disease Progression

MRI scanning also reveals a dynamically spreading wave of gray matter loss in the brains of patients with AD. With novel brain mapping methods, the loss pattern can be visualized as it spreads over time from temporal and limbic cortices into frontal and occipital brain regions, sparing sensorimotor cortices. These shifting deficits are correlated with cognitive decline. As shown in Figure 2, cortical atrophy occurs in a well-defined sequence as the disease progresses, mirroring the temporal sequence of beta-amyloid (Aβ) and neurofibrillary tangle accumulation observed at autopsy. The trajectory of deficits also matches the sequence of metabolic decline typically observed with positron emission tomography.

To map AD progression, advancing deficits can be visualized as dynamic video maps that change over time, which distinguish different phases of AD and differentiate AD from normal aging. Frontal brain regions, spared early in the disease, show pervasive deficits later (>15% loss). Local gray matter loss rates ($5.3 \pm 2.3\%$ per year in AD vs. $0.9 \pm 0.9\%$ per year in controls) are faster in the left hemisphere than the right, at least at this stage of AD. Transient

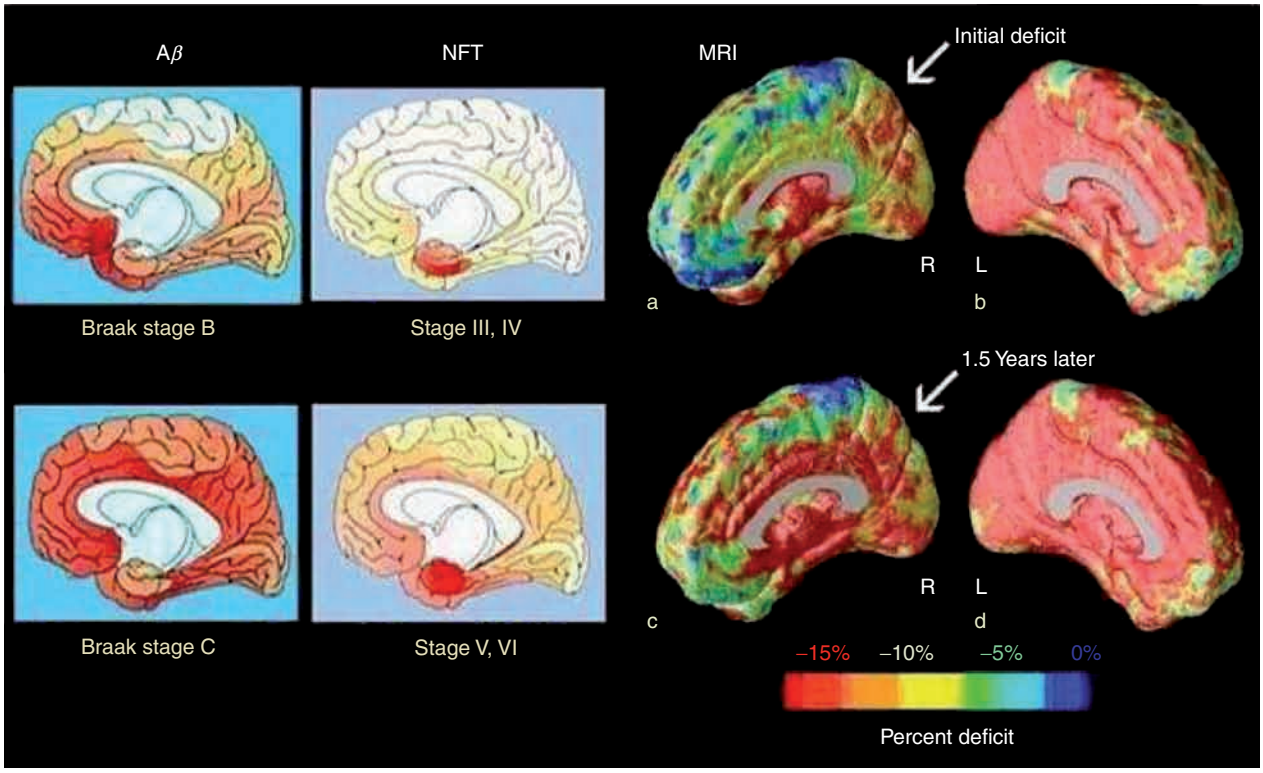

Figure 2 Gray matter deficits spread through the limbic system in moderate AD. Deficits during the progression of AD are detected by comparing average profiles of gray matter volumes between 12 AD patients (age, 68.4 ± 1.9 years) and 14 elderly matched controls (age, 71.4 ± 0.9 years). Colors show the average percent loss of gray matter relative to the control average. Profound loss engulfs the left medial wall (>15% (b), (d)). On the right, however, the deficits in temporoparietal and entorhinal territory (a) spread forward into the cingulate gyrus 1.5 years later (c), after a 5-point drop in average MMSE test scores. Limbic and frontal zones are prominently divided, with different degrees of impairment (c). MRI-based changes, observed in living patients, agree strongly with the spatial progression of Aβ and neurofibrillary tangle (NFT) pathology observed postmortem (Braak stages B and C and III to VI. NFT accumulation is minimal in sensory and motor cortices, but occurs preferentially in entorhinal pyramidal cells, the limbic periallocortex (layers II/IV), the hippocampus/amygdala and subiculum, the basal forebrain cholinergic systems and subsequently in temporoparietal and frontal association cortices (layers III/V). Left four panels adapted from Braak H and Braak E (1997) Staging of Alzheimer-related cortial destruction. *International Psychogeriatrics* 9(supplement 1): 257–271; discussion 269–272.

barriers to disease progression also appear. A frontal band (0–5% loss) is sharply delimited from the limbic and temporoparietal regions that show severest deficits in AD (>15% loss). This pattern is consistent with the hypothesis that AD pathology spreads centrifugally from limbic/paralimbic to higher-order association cortices. This degenerative sequence, observed as it develops in living patients, provides a quantitative, dynamic visualization of cortical atrophic rates in dementia, over a period of cognitive decline lasting 1.5 years.

What Is Gray Matter Atrophy?

Gray matter atrophy observed on MRI is linked with cognitive decline in AD and is attributable to several cellular processes. In healthy aging, age-related neuronal loss does not occur in most neocortical regions and appears specific to the frontal cortex and some hippocampal regions (e.g., CAI and the subiculum). In AD, however, there is substantial neuronal loss, with severe early losses in layer II of the entorhinal cortex.

Aβ and Neurofibrillary Tangle Maps

MRI-based maps of cortical atrophy agree strongly with postmortem maps of Aβ deposition (Aβ, Figure 2). Aβ is an insoluble protein that is a key feature of Alzheimer pathology. The spatial congruence of these two maps supports the hypothesis that Aβ deposition may participate in the cascade of events that leads to regional gray matter atrophy and neuronal cell loss. In both maps, primary sensorimotor cortices are relatively spared until late in the disease, and the superior temporal gyrus is less affected than other temporal lobe gyri. In early AD, intraneuronal filamentous deposits, or neurofibrillary tangles (NFTs), also accumulate within neurons. These deposits are composed of hyperphosphory lated tau protein. This cellular pathology disrupts axonal transport and induces widespread metabolic decline; it eventually leads to neuronal loss, observed as gross atrophy on MRI. Braak and Braak noted on autopsy that NFT distribution was initially restricted to entorhinal cortices, spreading to higher-order temporoparietal association cortices, then frontal, and ultimately primary sensory and visual areas. MRI scans suggest that a similar wave of cortical atrophy can be mapped in patients while they are alive. This provides a biological marker of disease progression that can monitor the effects of therapy.

It remains a mystery why brain changes in AD occur in this sequence. Braak and Braak suggested that the atrophic trajectory in AD is somewhat the reverse of the sequence in which cortical areas are myelinated during development. For example, primary sensory regions myelinate first and degenerate last, and temporal regions mature last but degenerate first in AD. This palindromic sequence is largely supported by the pattern of cortical changes observed on MRI. The selective vulnerability of specific cortical systems in AD may relate to differences in cellular maturational rates and/or plasticity. The most plastic systems may also be most vulnerable to AD.

Conclusion

MRI scans can measure brain change in AD, mapping the disease process in detail. MRI measures of disease progression, along with measures of genetic risk and abnormalities in specific neuropsychological tests, now provide key quantitative predictors to monitor brain degeneration and gauge how well it is decelerated or delayed in clinical trials.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: Molecular Genetics; Axonal Transport and Alzheimer's Disease; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Functional Neuroimaging Studies of Aging; Magnetic Resonance Spectroscopy; Neuroimaging; Numerical Intelligence: Neural Substrates; Perfusion MRI.

Further Reading

Arnold SE, Hyman BT, Flory J, et al. (1991) The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with Alzheimer's disease. *Cerebral Cortex* 1: 103–116.

Baron JC, Chetelat G, Desgranges B, et al. (2001) *In vivo* mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease. *Neuroimage* 14: 298–309.

Braak H and Braak E (1997) Staging of Alzheimer-related cortical destruction. *International Psychogeriatrics* 9(supplement 1): 257–261; discussion 269–272.

Bradley KM, Bydder GM, Budge MM, et al. (2002) Serial brain MRI at 3–6 month intervals as a surrogate marker for Alzheimer's disease. *British Journal of Radiology* 75: 506–513.

Chetelat G and Baron JC (2003) Early diagnosis of Alzheimer's disease: Contribution of structural neuroimaging. *Neuroimage* 18: 525–541.

Convit A, de Asis J, de Leon MJ, et al. (2000) Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. *Neurobiology of Aging* 21: 19–26.

Fox NC, Crum WR, Scahill RI, et al. (2001) Imaging of onset and progression of Alzheimer's disease with voxel-compression mapping of serial magnetic resonance images. *Lancet* 358: 201–205.

Frisoni GB, Laakso MP, Beltramello A, et al. (1999) Hippocampal and entorhinal cortex atrophy in frontotemporal dementia and Alzheimer's disease. *Neurology* 52: 91–100.

Gomez-Isla T, Price JL, McKeel DW Jr., et al. (1996) Profound loss of layer II entorhinal cortex neurons occurs in very mild Alzheimer's disease. *Journal of Neuroscience* 16: 4491–4500.

Jack CR Jr., Slomkowski M, Gracon S, et al. (2003) MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. *Neurology* 60: 253–260.

Jobst KA, Smith AD, Szatmari M, et al. (1994) Rapidly progressing atrophy of medial temporal lobe in Alzheimer's disease. *Lancet* 343: 829–830.

Kaye J, Moore M, Kerr D, et al. (1999) The rate of brain volume loss accelerates as Alzheimer's disease progresses from a presymptomatic phase to frank dementia. *Neurology* 52(supplement): A569–A570.

Laakso MP, Lehtovirta M, Partanen K, et al. (2000) Hippocampus in AD: A 3-year follow-up MRI study. *Biological Psychiatry* 47: 557–561.

Malmgren R (2000) Epidemiology of aging. In: Coffey CE and Cummings JL (eds.) *Textbook of Geriatric Neuropsychiatry*, pp. 17–31. Washington, DC: American Psychiatric Press.

Mesulam MM (2000) A plasticity-based theory of the pathogenesis of Alzheimer's disease. *Annals of the New York Academy of Sciences* 924: 42–52.

Morrison JH and Hof PR (1997) Life and death of neurons in the aging brain. *Science* 278: 412–419.

O'Brien JT, Paling S, Barber R, et al. (2001) Progressive brain atrophy on serial MRI in dementia with Lewy bodies, AD, and vascular dementia. *Neurology* 56: 1386–1388.

Peters A, Morrison JH, Rosene DL, and Hyman BT (1998) Feature article: Are neurons lost from the primate cerebral cortex during normal aging? *Cerebral Cortex* 8: 295–300.

Scahill RI, Schott JM, Stevens JM, et al. (2002) Mapping the evolution of regional atrophy in Alzheimer's disease: Unbiased analysis of fluid-registered serial MRI. *Proceedings of the National Academy of Sciences of the United States of America* 99: 4703–4707.

Thal DR, Rub U, Orantes M, and Braak H (2002) Phases of A-beta deposition in the human brain and its relevance for the development of AD. *Neurology* 58: 1791–1800.

Thompson PM, Hayashi KM, de Zubicaray G, et al. (2003) Dynamics of gray matter loss in Alzheimer's disease. *Journal of Neuroscience* 23: 994–1005.

Thompson PM, Mega MS, Woods RP, et al. (2001) Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas. *Cerebral Cortex* 11: 1–16.

Alzheimer's Disease: Neurodegeneration

N J Cairns, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Neuropathology

Alzheimer's disease (AD) is clinically, neuropathologically, and genetically heterogeneous. Subgroups include familial and sporadic forms; there are also phenotypic differences according to the gene and the defect within a single gene. Not infrequently, atypical cases may be seen, and AD may be found in combination with another neurodegenerative disease, most frequently dementia with Lewy bodies, which may also contribute to the cognitive deficits.

Patients with AD often die of bronchopneumonia and there is often little pathology outside the central nervous system (CNS). Depending on the stage of disease, the brain may appear unremarkable to the naked eye or may be grossly atrophic (Figure 1). Brain weight is typically reduced, often to less than 1000 g, from the average of 1250–1400 g. The atrophy is usually symmetrical and affects the frontal, temporal, and parietal lobes, with relative sparing of the sensorimotor cortices and occipital lobe, although all cortical areas may be affected in the most severe cases. Senile plaques and neurofibrillary tangles are frequently found with other lesions which are not specific to AD, as they are found in the aged brain, including granulovacuolar degeneration, Hirano bodies, gliosis, synaptic and neuronal loss, and white-matter and vascular changes.

The senile or neuritic plaque is one of the signature lesions of the AD brain (Figure 2). These complex extracellular structures range in size from 50 to 200 μm and are readily visualized by silver impregnation methods. In research settings and diagnostic laboratories, immunohistochemistry may be available, and these structures can be detected by antibodies raised against specific epitopes of the pathological protein (Aβ). Immunohistochemistry generally reveals much more extensive pathology than that seen by traditional silver and other staining methods and is now the method of choice for detecting the molecular pathology of most neurodegenerative diseases, not just of AD. The classical senile plaque consists of an amyloid core with a ring or crown, as seen in cross section with the light microscope, of argyrophilic axonal and dendritic processes, amyloid fibrils, astrocytic processes, and microglial cells. The neuritic processes of the senile plaque are often dystrophic and contain abnormal paired helical filaments (PHFs) made up largely of hyperphosphorylated tau protein. The amyloid is composed of 5–10 nm filaments of amyloid β-protein (Aβ), a 39- to 43-amino-acid (4 kDa) protein, a cleavage product of a transmembranous amyloid precursor protein (APP).

Aβ immunohistochemistry reveals a wider spectrum of plaque types than is seen with traditional staining methods. Aβ deposits include subpial, diffuse, ring-with-core, compact, vascular, and dyshoric deposits (Figure 3). Aβ plaques may be found throughout the brain, including the neocortex, amygdala, hippocampus, striatum, pallidum, nucleus basalis of Meynert, thalamus, midbrain, medulla oblongata, cerebellar cortex, and spinal cord. Immunohistochemistry using antibodies that recognize a full-length A$\beta_{1–42(43)}$ or a truncated A$\beta_{1–40}$ detects most Aβ species in the AD brain. The predominant species in sporadic AD is A$\beta_{1–42(43)}$, which is present in most plaques types; A$\beta_{1–42(43)}$ is found only in diffuse plaques, indicating that this species may be more toxic. This widespread deposition of Aβ protein is evidence that AD is an Aβ-amyloidosis of the CNS. *In vitro* and *in vivo* models also provide evidence that soluble Aβ oligomers may also be injurious to neurons and disrupt synaptic function (Figures 4–6).

The second histological signature lesion of AD is the neurofibrillary tangle (NFT; Figures 7 and 8). The NFT is not specific to AD; NFTs occur in aging and other neurodegenerative diseases, including Down syndrome, dementia pugilistica, postencephalitic parkinsonism, amyotrophic lateral sclerosis–parkinsonism–dementia complex of Guam, subacute sclerosing panencephalitis, dementia with tangles with and without calcification, and myotonic dystrophy. NFTs also are characteristic of frontotemporal lobar degeneration (FTLD) with tauopathy, which includes corticobasal degeneration, progressive supranuclear palsy, and argyrophilic grain disease, but their ultrastructure and tau isoform patterns, as demonstrated by immunohistochemistry or biochemistry, differ from those seen in AD and the aging brain.

In AD, as in the aged brain of cognitively normal individuals, NFTs are typically numerous in the medial temporal lobe, but in AD their distribution is much more widespread and may involve neocortex and subcortical nuclei. NFTs are intraneuronal inclusions composed of cytoskeletal components, mainly tau, and their shape is determined by the type of neuron in which they develop – for example, in pyramidal neurons they are flame-shaped, while in subcortical neurons they may be globose in appearance. Mature NFTs are

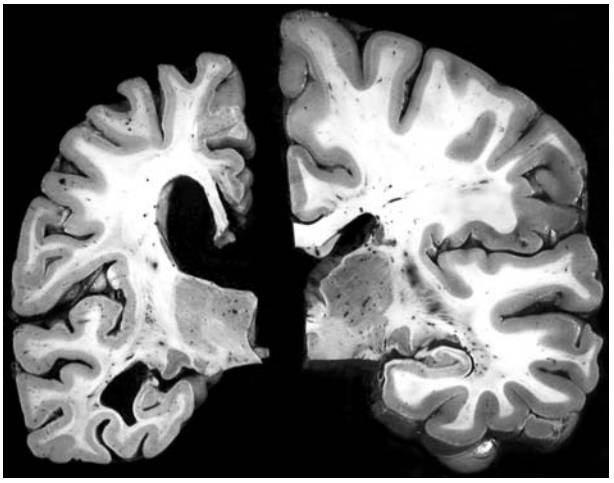

Figure 1 On the left, a coronal slice of the hemi-brain of a 70-year-old patient with severe Alzheimer's disease, showing atrophy: the lateral ventricle is enlarged with rounding of its angle, several gyri are narrowed, the hippocampus is small, and the lateral fissure is widened. On the right, a slice of the right hemi-brain of a normal age-matched individual.

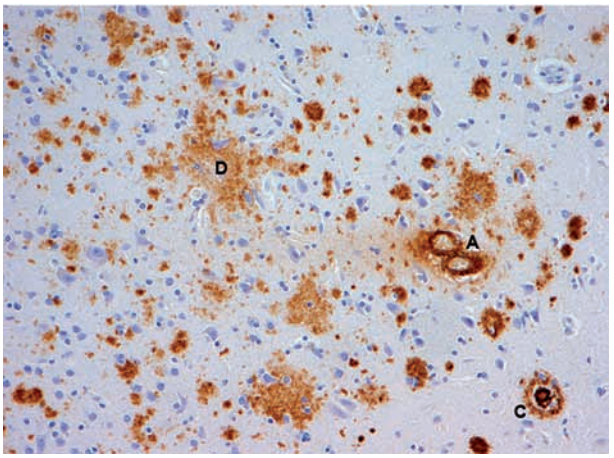

Figure 3 Extracellular amyloid β-protein deposits are present through all layers of the neocortex and have variable morphology: a diffuse plaque (D), a ring-with-core plaque (C), and cerebral amyloid angiopathy (A). Amyloid β-protein (10D5 antibody) immunohistochemistry.

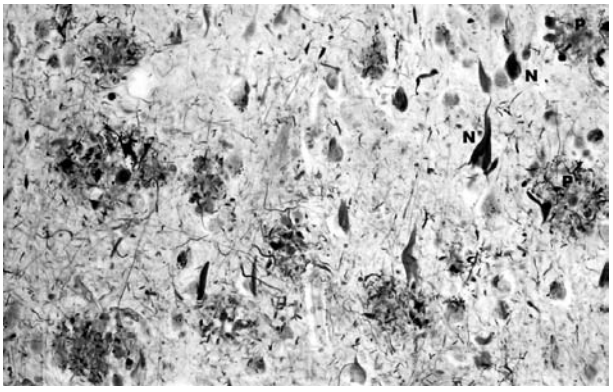

Figure 2 Neurofibrillary tangles (N) and neuritic plaques (P) in the hippocampus. Modified Bielschowsky silver impregnation.

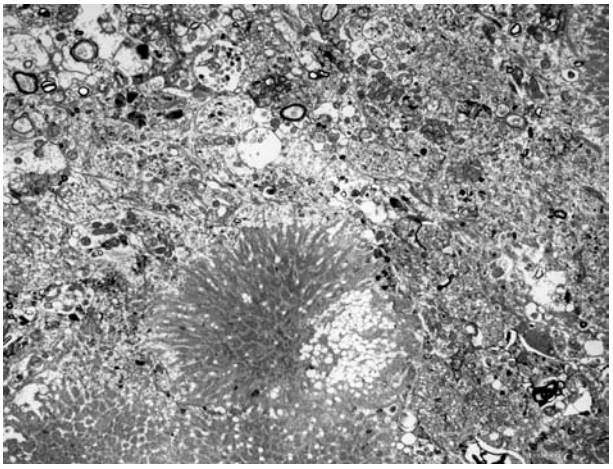

Figure 4 A low-power electron micrograph revealing a compact senile plaque.

argyrophilic, and immunohistochemistry reveals that NFTs are usually ubiquitinated, depending on the stage of evolution of the tangle; pretangles which are labeled by antiphosphorylated tau antibodies are not generally ubiquitinated, nor are they argyrophilic, indicating that the pretangle is an early stage in pathogenesis of this inclusion. Electron microscopy reveals that the NFT is composed largely of PHFs 10 nm in diameter, with cross-over points at every 80 nm, producing the periodicity of a double helix. NFTs also contain straight filaments with a mean diameter of 15 nm. The major component of the NFT is tau protein, a microtubule-associate protein that is expressed predominantly in axons, and at low levels in astrocytes and oligodendrocytes. Human tau proteins are encoded by a single-copy gene on chromosome 17q21. In adult human brain, alternative splicing of exons 2, 3, and 10 generates six tau isoforms, ranging from 352 to 441 amino acids in length, which differ by the presence of either three or four microtubule binding repeats (3R tau or 4R tau, respectively), of 31 or 32 amino acids each. Additionally, alternative splicing of exons 2 and 3 leads to the absence (0N) or presence of inserted sequences of 29 (1N) or 58 (2N) amino acids in the N-terminal third of the molecule. In the adult human brain, the ratio of 3R:4R tau isoforms is approximately 1:1, while in other FTLDs with tauopathy there is a preponderance of 3R tau (Pick's disease), 4R tau (corticobasal degeneration, progressive supranuclear palsy, and agyrophilic grain disease), or 3R, 4R, or 3R and 4R tau, as seen in cases of FTLD with

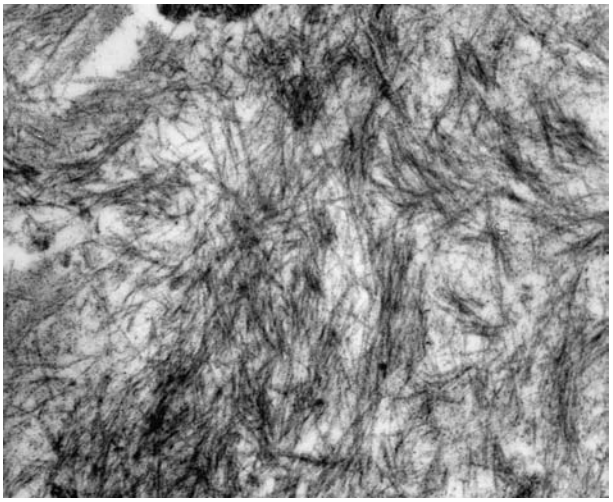

Figure 5 A high-power electron micrograph of loosely aggregated amyloid filaments of the corona of an amyloid β-protein plaque.

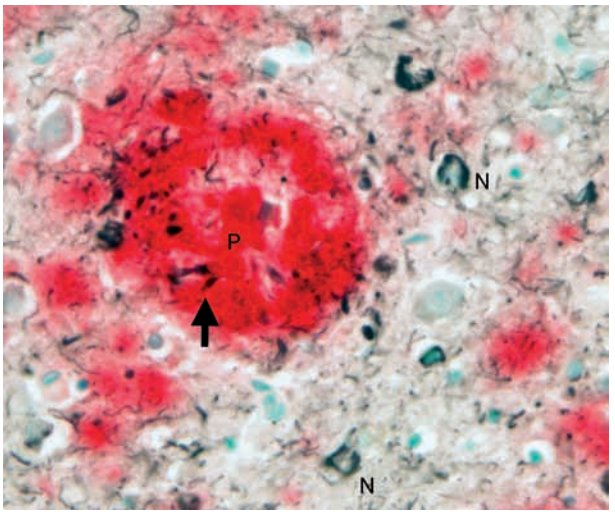

Figure 7 An amyloid β-protein plaque (P) and neurofibrillary tangles (N) in the superior temporal lobe of an Alzheimer's disease brain. The plaque contains tau-immunoreactive dystrophic neurites (arrow). Tau (black, PHF1 antibody) and amyloid β-protein (red, 10D5 antibody) immunohistochemistry.

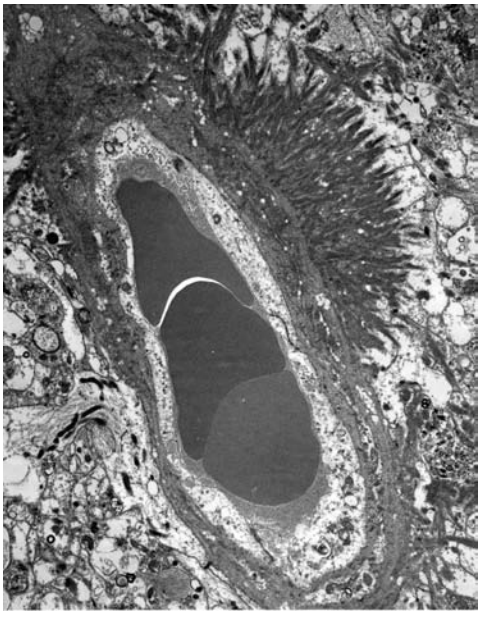

Figure 6 A low-power photomicrograph of dyshoric cerebral amyloid angiopathy with amyloid fibrils extending from the vessel wall and radiating into the surrounding brain parenchyma.

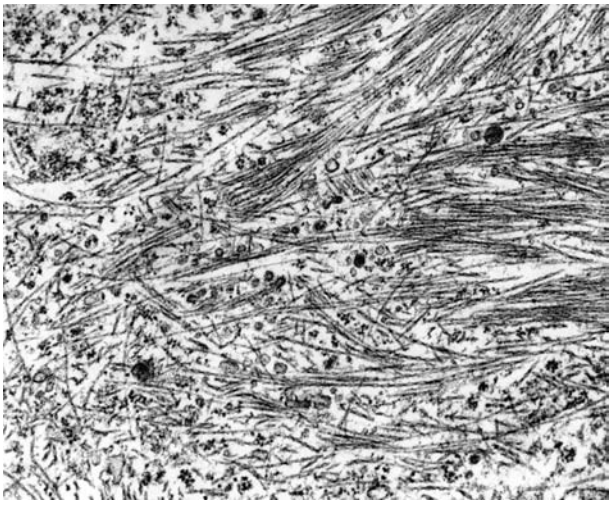

Figure 8 A high-power electron micrograph revealing paired helical filaments of a neurofibrillary tangle. Micrograph kindly provided by the late Professor LW Duchen, University of London, UK.

microtubule-associated protein tau (MAPT) gene mutation.

Tau binds to and stabilizes microtubules (MTs) and promotes MT polymerization. The function of tau as an MT binding protein is regulated by phosphorylation. Phosphorylation at approximately 30 of these sites has been identified in normal tau protein. Several protein kinases and protein phosphatases have been implicated in regulating the phosphorylation state and thus the function of tau. The phosphorylation sites are clustered in regions flanking the MT binding repeats, and increasing tau phosphorylation at multiple sites negatively regulates MT binding. All six isoforms of tau in AD are hyperphosphorylated, and phosphorylation-specific anti-tau antibodies may be used to identify three sites of tau pathology: the NFT, dystrophic neurites of senile plaques, and neuropil threads. In both sporadic and familial AD, tau is hyperphosphorylated, and it is this 'abnormal' tau that is the principal component of the filamentous

aggregates in neurons and one of the pathological hallmarks of this disorder.

The Aβ Cascade Hypothesis

The initial observation that Aβ deposits are present in the blood vessels of young patients with Down syndrome (DS) was followed by the discovery that Aβ plaques and NFTs are present in adult DS and AD. These studies led to the hypothesis that a gene encoding Aβ was located on the long arm of chromosome 21. Subsequently, mutations in the APP gene were identified in a family with early-onset AD, and then additional mutations were reported. These neuropathological, biochemical, and genetic findings led to the hypothesis that increased Aβ production leads to a catastrophic cascade, including synaptic alterations, fibrilization, microglial and astrocytic activation, abnormal phosphorylation of tau proteins to form oligomers and the PHFs of NFTs, progressive synaptic and neuronal loss, loss of multiple neurotransmitters (especially acetylcholine), and ultimately to dementia (Figure 9).

The generation of Aβ peptides is an example of regulated intramembrane proteolysis. APP processing is regulated by either of two membrane-bound proteases: α-secretase (a type of metalloproteinase) and β-secretase, also called β-site APP-cleaving enzyme (BACE). The membrane-associated fragment caused by BACE cleavage may undergo further cleavage by the β-secretase complex, an aspartyl protease composed of presenilin-1 (PS1) or PS2, nicastrin, APH1 (anterior pharynx defective-1), and an anti-presenilin protein enhancer (PEN2). All four proteins are necessary and sufficient to form the γ-secretase complex. Further evidence for the importance of Aβ oligomerization in AD comes from the observation that mutations in PS1 and PS2, components of the γ-secretase complex, lead to enhanced production of Aβ_{1-42}. Neuropathological studies confirm that

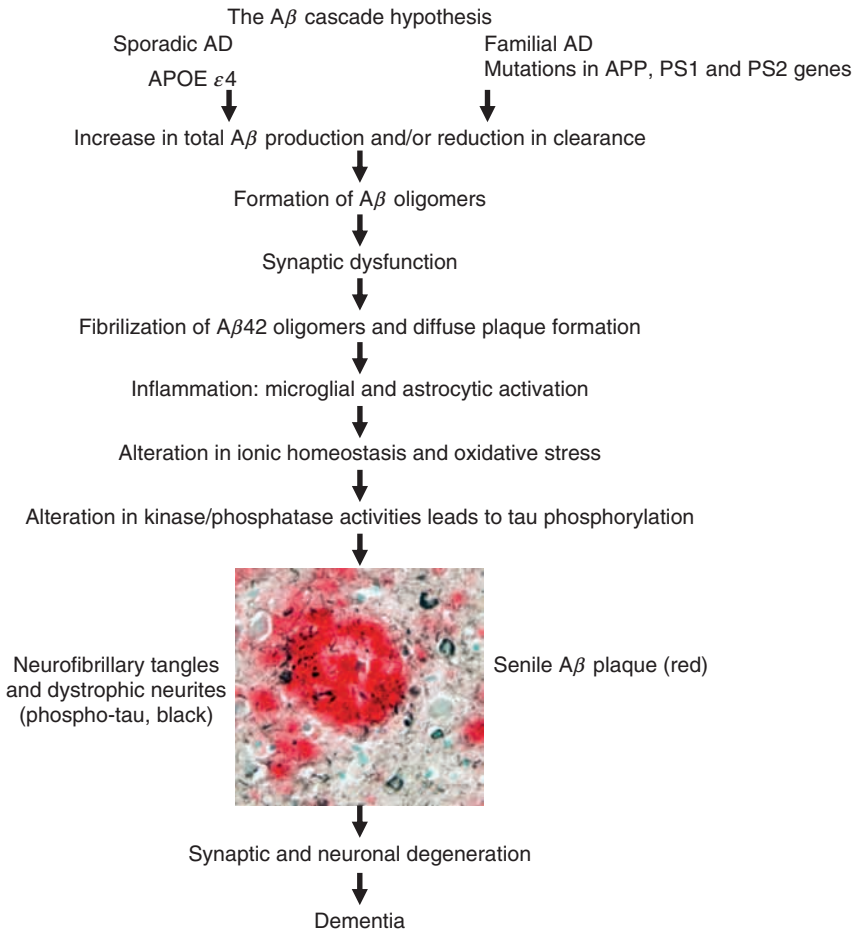

Figure 9 The amyloid β-protein (Aβ) cascade hypothesis of Alzheimer's disease (AD). APOE, apolipoprotein E; APP, amyloid precursor protein; PS1, PS2, presenilins.

A $_{1–42}$ is present in diffuse plaques, an early stage in senile plaque pathogenesis. The molecular dissection of the A cascade has generated several potential targets for therapeutic intervention: inhibitors of - and γ-secretases that generate A , statins, A vaccines, and nonsteroidal anti-inflammatory drugs (NSAIDs) are some of the classes of compounds targeting A -induced neurodegeneration.

The Tau Hypothesis

Several sporadic and familial neurodegenerative disorders, including AD, that are characterized clinically by dementia and/or motor dysfunction are characterized pathologically by abnormal intracellular accumulations of the microtubule-associated protein tau (MAPT); these disorders are collectively called tauopathies. The progressive accumulation of filamentous tau inclusions in the absence of other disease-specific neuropathological abnormalities provides evidence implicating tau dysfunction in disease onset and progression. However, the discovery of pathogenic *MAPT* gene mutations in the heterogeneous group of disorders known as frontotemporal lobar degeneration (FTLD) with *MAPT* mutation, also called frontotemporal dementia with parkinsonism linked to chromosome 17 (FTDP-17), provided confirmation of the central role of tau abnormalities in the etiology of neurodegenerative tauopathies. These findings have opened up novel areas for investigation into the mechanisms of tau dysfunction and the relationship of tau abnormalities to brain degeneration.

Several transgenic models of tau pathology have been generated by overexpressing human tau proteins in mice. However, these mice are either asymptomatic or develop pathology that is localized to the spinal cord and/or lacks many of the key features of tau-based disorders. In contrast, the introduction of the P301L mutation led to the development of transgenic mice that develop age- and gene dose-dependent accumulation of tau tangles in the brain and spinal cord, with associated nerve cell loss and gliosis as well as behavioral abnormalities. Similar to human disease, the tau aggregates are composed of only mutant human tau, further implicating the P301L change in promoting the selective aggregation of mutant tau. Other systems have also been developed to model various aspects of human tauopathies, including a transgenic mouse overexpressing the shortest human tau isoform; this mouse acquired age-dependent tau-pathology similar to that seen in FTLD with *MAPT* mutation. Overexpression of either wild-type or mutant tau in *Drosophila melanogaster* demonstrates features of tauopathy, including adult-onset progressive neurodegeneration with accumulation of abnormal tau. However, the neurodegeneration in this model occurred in the absence of NFT formation. More recent studies have demonstrated NFT-like pathology when tau is co-expressed with *shaggy,* a homolog of glycogen synthase 3-kinase (GSK3), an enzyme implicated in tau phosphorylation. Neurodegeneration and defective neurotransmission have also been demonstrated in *MAPT* transgenic *Caenorhabditis elegans*. In this model, panneuronal expression of normal and mutant tau resulted in altered behavior, accumulation of insoluble phosphorylated tau, age-dependent loss of axons and neurons, and structural damage to axonal tracts. These models recapitulate various features of the tauopathies and highlight targets for disease-modifying therapies, not only for AD but also for related tauopathies. Thus, microtubule-stabilizing drugs might ameliorate the sequestration of tau into NFTs, and drugs such as LiCl, in an animal model of tauopathy, that impede the abnormal phosphorylation of tau by inhibiting GSK3 indicate a fruitful area for drug discovery.

Timeline of Neurodegenerative Changes in AD

Many nondemented older individuals have few or no AD-type changes in the brain. However, about 30% of cognitively normal individuals over the age of 75 years have both diffuse and neuritic plaques throughout the neocortex, identical to those seen in individuals with symptomatic AD (Figure 10). The change in the brain that corresponds most closely to the clinical symptoms of those with very mild dementia (Clinical Dementia Rating = 0.5) is neuronal loss in the medial temporal lobe, including the entorhinal cortex and hippocampus; cognitively normal individuals, in contrast, have minimal or no loss of neurons from these areas. Thus, there is a 'preclinical' stage of AD in which there are sufficient plaques in neocortical areas for a neuropathological diagnosis of AD, but there is not sufficient synaptic and neuronal loss to produce cognitive symptoms. In familial cases, misfolding of A may start at a very young age, and only after two to three decades may result in the neuropathology of AD. Research efforts are now focusing on this preclinical stage of AD, which may extend years or even decades before the onset of clinical symptoms. The development of antecedent biomarkers that detect asymptomatic AD is a priority; efforts include clinical, cognitive, neuropsychological, structural and functional imaging, genetic, and proteomic studies. Tau and A protein levels in cerebrospinal fluid are potentially additional markers of preclinical

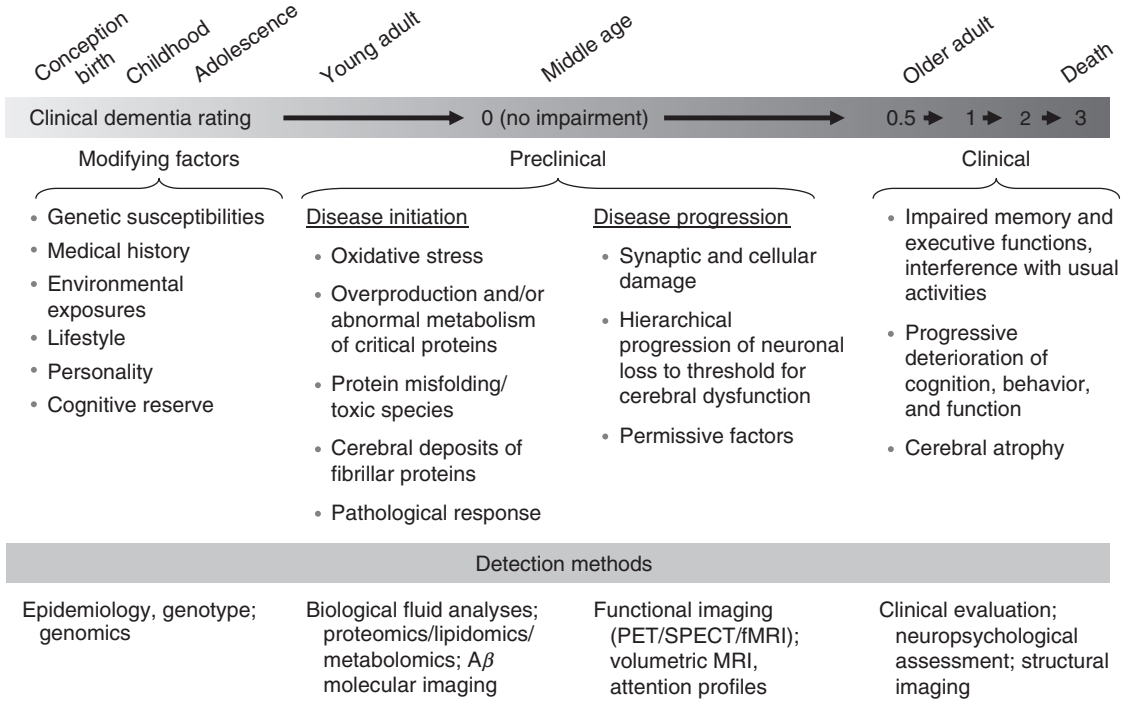

Figure 10 Hypothetical timeline for Alzheimer's disease. PET, positron emission tomography; SPECT, single-photon computed tomography; fMRI, functional magnetic resonance imaging.

AD, as is the Aβ imaging agent, Pittsburgh compound B (PIB). It is likely that a combination of biomarkers will provide greater diagnostic accuracy than is possible using any single analyte or measure.

Conclusions

The accumulations of filamentous Aβ species in the extracellular space, and of abnormally phosphorylated tau proteins within neurons, are the signature lesions of AD. Protein misfolding, fibrilization, and aggregate formation are not unique to AD, but are common features of a wide variety of sporadic and familial neurodegenerative disorders (synucleinopathies, tauopathies, trinucleotide repeat disease, and TDP-43 proteinopathies). These diseases are distinguished by the distinct topographic and cell type-specific distribution of inclusions. The biochemical and ultrastructural characteristics of the inclusions also reveal a significant phenotypic overlap. The discovery that multiple mutations in *APP*, *PS1*, and *PS2* genes lead to abnormal protein aggregation in AD demonstrates that neuronal Aβ mismetabolism is sufficient to produce neurodegenerative disease. Experimental evidence indicates that mutations lead to specific alterations in expression, function, and biochemistry of Aβ and tau proteins. The realization that AD pathology is present many years prior to the onset of clinical symptoms has led to dramatic progress in the search for antecedent biomarkers which herald disease onset. These new markers, once validated, will facilitate the identification of individuals at the preclinical or early stage of AD, and these individuals stand to benefit greatly from emerging therapeutic interventions.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: Transgenic Mouse Models; Alzheimer's Disease: MRI Studies; Axonal Transport and Alzheimer's Disease; Axonal Transport and Neurodegenerative Diseases; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Oxidative Damage in Neurodegeneration and Injury; Transgenic Models of Neurodegenerative Disease.

Further Reading

Braak H, Alafuzoff I, Arzberger T, et al. (2006) Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathologica* 112: 389–404.

Cairns NJ, Lee VM-Y, and Trojanowski JQ (2004) The cytoskeleton in neurodegenerative diseases. *Journal of Pathology* 204: 438–449.

Csernansky JG, Wang L, Swank J, et al. (2005) Preclinical detection of Alzheimer's disease: Hippocampal shape and volume predict dementia onset in the elderly. *NeuroImage* 25: 783–792.

Forman MS, Trojanowski JQ, and Lee VM-Y (2004) Neurodegenerative diseases: A decade of discoveries paves the way for therapeutic breakthroughs. *Nature Medicine* 10: 1055–1063.

Goedert M, Spillantini MG, Cairns NJ, et al. (1992) Tau proteins of Alzheimer paired helical filaments: Abnormal phosphorylation of all six brain isoforms. *Neuron* 8: 159–168.

Haass C and Selkoe DJ (2007) Soluble protein oligomers in neurodegeneration: Lessons from the Alzheimer's amyloid beta-peptide. *Nature Reviews – Molecular Cell Biology* 8: 101–112.

McKhann GM, Albert MS, Grossman M, et al. (2001) Clinical and pathological diagnosis of frontotemporal dementia: Report of the Work Group on Frontotemporal Dementia and Pick's Disease. *Archives of Neurology* 58: 1803–1809.

Mirra SS, Heyman A, McKeel D, et al. (1991) The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* 41: 479–486.

Mirra SS and Hyman BT (2002) Ageing and dementia. In: Graham DI and Lantos PL (eds.) *Greenfield's Neuropathology,* 7th edn., pp. 195–271. London: Arnold.

Price JL, Davis PB, Morris JC, et al. (1991) The distribution of tangles, plaques and related immunohistochemical markers in healthy aging and Alzheimer's disease. *Neurobiology of Aging* 12: 295–312.

Price JL and Morris JC (1999) Tangles and plaques in nondemented aging and "preclinical" Alzheimer's disease. *Annals of Neurology* 45: 358–368.

Relevant Website

http://www.alzforum.org – Alzheimer Research Forum.

Alzheimer's Disease: Transgenic Mouse Models

K H Ashe, University of Minnesota, Minneapolis, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Creation of Transgenic Mouse Models

Biological Foundation

Three scientific breakthroughs made the creation of the earliest transgenic mouse models of Alzheimer's disease (AD) possible. The first was the isolation and sequencing of the amyloid-β (Aβ) peptide in 1984. The second was the cloning of the amyloid precursor protein (APP) gene in 1987 and the elucidation of its role in generating the Aβ peptide. Third was the discovery of the first mutation in autosomal dominant familial AD in APP in 1991 and the subsequent realization that all autosomal dominant mutations causing AD enhance the ability of the Aβ protein to aggregate, either by increasing its overall production or by the generation of amino acid variants that potentiate its aggregation.

This information enabled investigators to create APP transgenic mice modeling AD. The earliest mouse models were developed in the first half of the previous decade.

Altogether about 20 such mice have been published, many of which show age-related amyloid plaque deposition and memory loss. However, APP transgenic mice are incomplete models of AD because they lack neurofibrillary tangles and develop few or no neurodegenerative changes, such as neuronal or synaptic loss. Transgenic mice that develop neurofibrillary tangles accompanied by significant neuronal loss emerged with the creation of tau transgenic mice.

Four important landmarks in tau biology made the generation of tau transgenic mice possible. First was the isolation and characterization in 1975 of tau, which is involved in promoting the aggregation and polymerization of tubulin to form microtubules. Second was the cloning in 1986 of the tau gene. Third was the recognition that tau is the principal protein forming the core of the paired helical filaments of neurofibrillary tangles. Fourth was the discovery in 1998 of mutations in tau linked to familial tauopathies, called frontotemporal dementia with parkinsonism (FDTP). These advances led to the development of tau transgenic mice in the first years of the current decade.

Technical Methodology

The laboratory mouse remains the genetic model organism that is evolutionarily closest to the human. Mice pose advantages over other species, such as rats, dogs, pigs, or primates, in several aspects: (1) it is relatively easy to house large numbers of them; (2) the mouse genome is better characterized than that of any other mammal except humans; (3) many genetically altered transgenic knockout and knockin lines and genetically characterized mouse strains are available for cross-breeding and further genetic analysis; (4) cognitive studies can be done in appropriate strains of mice, such as C57B6/SJL and 129S6, making them especially relevant to the study of AD; and (5) studies in invertebrates, while useful, will lack some relevance in pharmacological studies and pathogenic investigations. To create transgenic mice with salient features of AD, investigators must take care in several aspects of design and development. The most important factors are the selection of the gene or gene variant to be expressed, the promoter driving transgene expression, the background strain of the mice, and the levels and distribution of transgenic protein expression achieved in the brain. The mouse lines with the most robust neuropathological phenotypes have been developed using the method of pronuclear microinjection of transgenes into fertilized oocytes (Figure 1).

Both wild-type and variant human APP and tau genes have been used to generate mice that develop neuropathology related to AD. Transgenes containing mutations linked to autosomal dominant AD or FDTP generate more-robust neuropathology than transgenes encoding wild-type genes. Although mutations in presenilin-1 and presenilin-2 genes lead to early-onset AD in humans, mice expressing presenilin gene variants do not develop neuropathology *per se*. In the presence of human APP transgenes, however, the presenilin gene variants accelerate plaque deposition in transgenic mice.

Characteristics of Transgenic Mouse Models

APP Transgenic Mice

APP transgenic mice often develop age-related cognitive deficits and recapitulate many of the neuropathological features of AD, including amyloid plaques, oxidative stress, astrogliosis, microgliosis, cytokine production, and dystrophic neurites (Figure 2). However, many important neuropathological features of AD are conspicuously absent in APP transgenic mice, including neurofibrillary tangles and gross atrophy.

There are variable degrees of neuronal and synaptic loss among the various transgenic lines, but no line exhibits the severity of neurodegeneration found in

Alzheimer's patients, despite amyloid plaque loads that often exceed those in human brain specimens. For example, Tg2576 mice are virtually devoid of neuronal or synaptic loss, while synaptic loss is present in J20 and PDAPP mice, and there is some neuronal loss in APP23 mice with massive quantities of amyloid plaques. All four of these lines of mice eventually develop amyloid plaque loads that exceed the amount found in typical Alzheimer patient brains at autopsy. Although we do not understand the factors which account for the variations in neurodegeneration between the mouse lines, it is clear that the differences are not related to the quantity of amyloid plaque deposition.

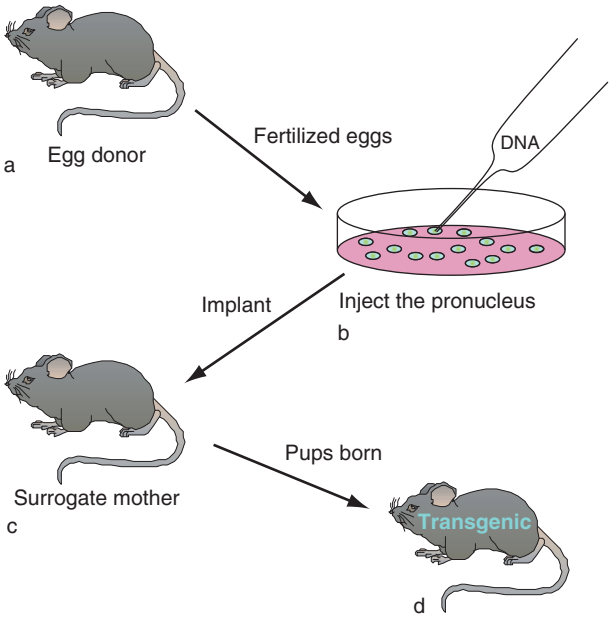

Figure 1 How transgenic mice are made. (a) Fertilized eggs (oocytes) are harvested from a donor mouse. (b) DNA encoding genes linked to Alzheimer's disease is injected into the pronucleus of the fertilized oocytes. (c) The oocytes are implanted into surrogate mothers, where they mature to full-term pups. (d) Of the pups that are born, a certain fraction, usually 10–20%, contain the transgene.

Tau Transgenic Mice

Tau transgenic mice generally develop age-related neurological abnormalities, neurofibrillary tangles, and neurodegeneration but invariably fail to form amyloid plaques. The earliest tau transgenic lines expressing tau ubiquitously in the neurons of the brain developed neurodegeneration mainly in the brain stem and spinal cord, leading to paralysis. Because the paralysis interfered with cognitive testing, tau transgenic mice in which variant tau expression was restricted to the forebrain were created and shown to develop age-related memory impairment.

Although the tau transgenic mice with the most robust neuropathological phenotypes express tau variants linked to FDTP (Figure 3), neurofibrillary tangles and neurodegeneration occur also in mice expressing wild-type human tau on a null tau background, suggesting a propensity for mouse tau to inhibit neurofibrillary tangle formation. Notably, these mice uniquely form the true paired-helical filaments characteristic of neurofibrillary tangles in AD. Straight filaments, rather than paired-helical filaments, are found in tau transgenic mice expressing FDTP tau variants. This distinction, once believed to be critical, has become less important with the discovery that neurofibrillary tangles contribute little or not at all

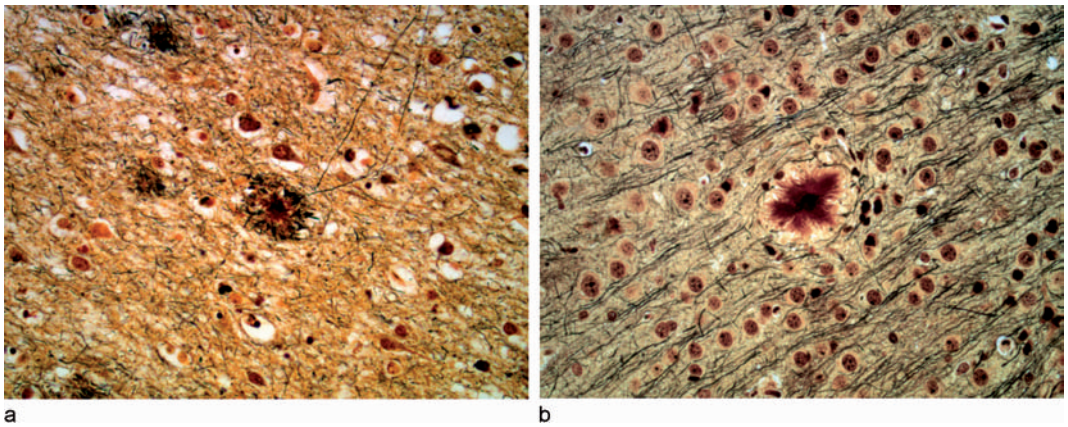

Figure 2 Amyloid plaques in human (a) and mouse (b). Amyloid plaques that are generated in amyloid precursor protein (APP) transgenic mice closely resemble the shape, composition, and size of those found in humans with Alzheimer's disease. The plaque on the left is from a human specimen, while the one on the right is from an aged APP transgenic mouse. Despite the similarities in plaques in the two species, neurodegenerative changes that are clearly present in the human specimen are largely absent in the mouse specimen. Both specimens were stained with the Bielschowsky silver stain. The photomicrographs were taken at the same magnification (40×). Photomicrographs were kindly provided by Dr. Martin Ramsden.

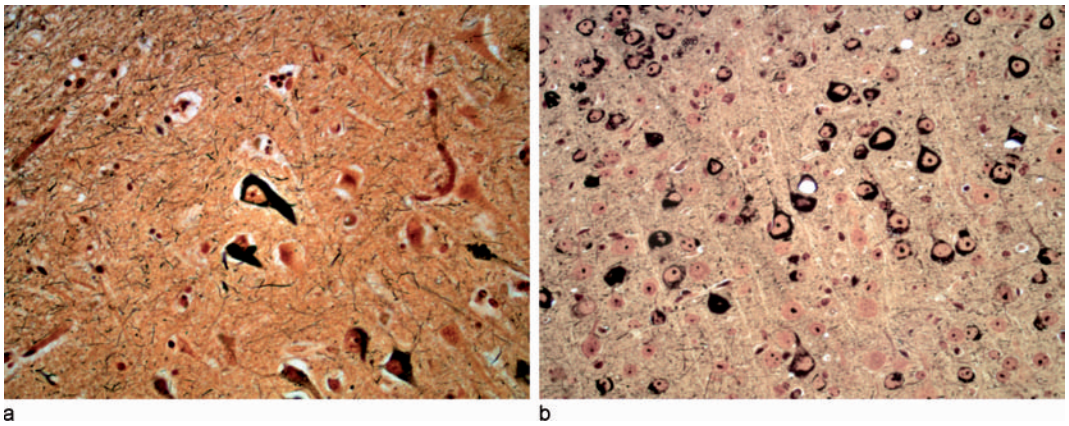

Figure 3 Neurofibrillary tangles in (a) human and (b) mouse. Neurofibrillary tangles that are generated in tau transgenic mice resemble those found in humans with Alzheimer's disease. The tangles on the left are from a human specimen, while those on the right are from an aged tau transgenic mouse. Both specimens were stained with the Bielschowsky silver stain. The photomicrographs were taken at the same magnification (40×). Photomicrographs were kindly provided by Dr. Martin Ramsden.

to neuronal death and memory impairment, as discussed more fully in the section titled 'Mechanism of memory loss in transgenic mouse models.'

Other Types of Transgenic Mice

Efforts to combine the cardinal neurological and neuropathological features of AD, namely amyloid plaques, neurofibrillary tangles, neurodegeneration, and cognitive impairment, within a single mouse resulted in the generation of transgenic mice expressing multiple tau and APP protein variants, sometimes in combination with presenilin 1 variants. Although some such mice develop amyloid plaques, neurofibrillary tangles, neuronal loss, and memory loss, they constitute essentially a hybrid of two separate disorders, AD and FDTP. A number of interesting studies have been carried out using these hybrid mice, but their relevance to AD is not entirely clear. A mouse exhibiting all the hallmarks of AD that does not depend on the use of genes not linked to familial AD remains an elusive creature.

Utility of Transgenic Mouse Models

Validity of Transgenic Mouse Models

The lack of several important features of AD in APP transgenic mice, such as neurofibrillary tangles and prominent neurodegeneration, along with the shortcomings of the tau transgenic mice and the APP-tau hybrid mice, have prompted some scientists, rightly, to challenge the validity of transgenic mouse models of AD. To address this challenge, criteria may be devised for validating Alzheimer's mouse models, three of which are discussed here. First, theoretical validity refers to whether the use of a given transgene is based on sound biological principles. Second, factual validity refers to how accurately various aspects of the human disease are represented. Third, predictive validity refers to whether studies using the model predict outcomes in human trials. Clearly, predictive validity is the main determinant of the value of a given mouse model. Thanks to rapid progress in bench-to-bedside research in AD, we know more about the predictive validity of APP transgenic mice than of transgenic mice modeling any other neurodegenerative illness.

Aβ immunotherapy illustrates this point. Young PDAPP mice vaccinated with Aβ fail to develop amyloid deposits and astrogliosis. This result, reported in 1999, quickly led to an international, multicenter, randomized trial of an experimental Aβ vaccine in Alzheimer patients. The vaccine cleared amyloid plaques away from large regions of brain in Alzheimer patients. However, neurofibrillary tangles were unaffected, and neuronal loss persisted. Furthermore, when the effects of the vaccine on cognitive function were examined, it became clear that memory decline may have been slowed, but it was not restored. The failure to restore memory function in Alzheimer patients is in marked contrast to the effects of either active or passive Aβ immunization on memory function in APP transgenic mice. Memory loss can be prevented, and preexisting memory loss can be reversed and memory function fully restored in several different lines of APP transgenic mice following Aβ immunization treatments.

These studies show that when the effects of the Aβ vaccine on amyloid plaques are studied, then the results in APP transgenic mice predict what occurs in humans. However, when memory function is examined, then there is a striking discrepancy between what

occurs in the mice and what occurs in patients. Thus, the predictive validity of APP transgenic mice as regards amyloid plaques is good. In contrast, the predictive validity of APP transgenic mice in relation to memory function is poor.

Relevance of Transgenic Mouse Models

The value of transgenic models of AD is as good as our understanding of how closely the models mimic various stages of the illness. To appreciate the context in which these mice have helped us study AD, it is useful delineate the natural history of the illness. AD has a very insidious onset; we do not know precisely when neural dysfunction begins. Recent work suggests that the disease process may begin long before symptoms are present or neurodegeneration has occurred. By the time AD is clinically diagnosed, neurodegeneration is already under way.

APP transgenic mice and tau transgenic mice mimic different stages of AD. Each one is useful for exploring the progression of cognitive decline and the pathology of AD. The picture of minimal loss of neurons in many lines of memory-impaired APP transgenic mice suggests a closer resemblance to preclinical stages of AD than to the disease itself. In contrast, the substantial neuronal loss found in tau transgenic mice implies that they may be better models for studying the neurodegenerative aspects of AD.

Testing Experimental Therapies in Transgenic Mouse Models

A major effort in developing AD-modifying therapies has been directed at reducing the amyloid deposits and neurofibrillary tangles in the brain. Transgenic mice displaying plaques and tangles are useful models for testing the effects of therapies on these neuropathological abnormalities and have been used extensively for this purpose. However, as will be discussed in the next section, targeting molecules causing cognitive dysfunction and memory impairment may hold still better promise for treating the symptoms of AD.

Mechanism of Memory Loss in Transgenic Mouse Models

Ever since Alois Alzheimer described the amyloid plaques and neurofibrillary tangles that were evident when he examined the brain of a demented patient in 1906, the scientific research on what has come to be known as AD has focused on how the accumulation of tau and Aβ proteins form amyloid plaques and neurofibrillary tangles, which define the disease neuropathologically, rather than on the functional effects of these molecules on the brain. Why and how tau and Aβ molecules cause memory loss and dementia are only now becoming understood. Our knowledge of the manner in which these molecules disrupt brain function lags a century behind the neuropathological studies of tau and Aβ proteins because suitable tools for examining their effects on the workings of the living brain were lacking until recently. The creation of transgenic mouse models of AD helped fill this void and provided the vital reagents needed to begin to understand the adverse effects that the tau and Aβ proteins exert on memory and cognitive function.

In the Tg2576 APP transgenic mouse model of AD, a major source of heterogeneity in Aβ proteins arises from the aggregation of monomeric proteins to form higher-order structures. Monomeric and trimeric Aβ protein assemblies are located within neurons. In the extracellular space, monomers and trimers coexist with larger Aβ protein assemblies that have similar molecular weights as higher-order species, such as hexamers, nonamers, and dodecamers. Many of the monomeric and oligomeric forms of Aβ protein in the brain are soluble in aqueous buffers. In contrast, amyloid plaques are composed of Aβ fibrils that are insoluble in aqueous buffers and most detergents. In Tg2576 mice, amyloid plaques lag 3–6 months behind the initial loss of memory function, which occurs at 6 months of age, and they do not correlate with memory impairment, indicating that the plaques contribute little to brain dysfunction. Memory impairment is caused, instead, by a specific, soluble Aβ protein assembly, called Aβ*56 (read as A-beta star 56). Aβ*56 correlates strongly with memory dysfunction, and when purified from the brains of impaired Tg2576 mice and administered to young, healthy animals, it transiently disrupts memory function in the healthy subjects. Aβ*56 disrupts memory by a mechanism that does not involve synaptic or neuronal loss. Aβ*56 is present in brain tissue from Alzheimer patients and may contribute to memory loss associated with AD.

Neurofibrillary tangles contribute little, if at all, to neuronal loss and brain dysfunction in a line of transgenic mice modeling FDTP, called rTg4510. By dint of identical hyperphosphorylated tau species in the neurofibrillary tangles of rTg4510 mice and Alzheimer's brain tissue, it may be supposed that the effects of neurofibrillary tangles on dementia and neuronal loss in AD are similarly inconsequential. However, the molecular mechanisms by which tau induces memory impairment and kills neurons in rTg4510 mice or AD are unknown; it is likely that specific, as yet unidentified, tau* (tau star) molecules are responsible for disrupting brain function and interfering with neuronal viability. Finally, the relative contributions of Aβ*56, tau, and neurodegeneration to memory loss, cognitive impairment, and dementia remain unresolved.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: Neurodegeneration; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Animal Models of Alzheimer's Disease; Axonal Transport and Alzheimer's Disease; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Lipids and Membranes in Brain Aging.

Further Reading

Andorfer C, Kress Y, Espinoza M, et al. (2003) Hyperphosphorylation and aggregation of tau in mice expressing normal human tau isoforms. *Journal of Neurochemistry* 86: 582–590.

Games D, Adams D, Alessandrini R, et al. (1995) Alzheimer-type neuropathology in transgenic mice overexpressing V717F beta-amyloid precursor protein. *Nature* 373: 523–527.

Hsiao K, Chapman P, Nilsen S, et al. (1996) Correlative memory deficits, Aβ elevation, and amyloid plaques in transgenic mice. *Science* 274: 99–102.

Ishihara T, Hong M, Zhang B, et al. (1999) Age-dependent emergence and progression of a tauopathy in transgenic mice overexpressing the shortest human tau isoform. *Neuron* 24: 751–762.

Iyadurai SJP and Ashe KH (2005) Creating APP transgenic lines in mice. In: Xia W and Xu H (eds.) *Amyloid Precursor Protein: A Practical Approach*, pp. 185–200. Boca Raton, FL: CRC Press.

Lesné S, Koh MT, Kotilinek L, et al. (2006) A specific amyloid-β assembly in the brain impairs memory. *Nature* 440: 352–357.

Lewis J, McGowan E, Rockwood J, et al. (2000) Neurofibrillary tangles, amyotrophy and progressive motor disturbance in mice expressing mutant (P301L) tau protein. *Nature Genetics* 25: 402–405.

Oddo S, Caccamo A, Shepherd JD, et al. (2003) Triple-transgenic model of Alzheimer's disease with plaques and tangles: Intracellular Abeta and synaptic dysfunction. *Neuron* 39: 409–421.

SantaCruz K, Lewis J, Spires T, et al. (2005) Tau suppression in a neurodegenerative mouse model improves memory function. *Science* 309: 476–481.

Amnesia: Declarative and Nondeclarative Memory

L R Squire, P J Bayley, and C N Smith, University of California at San Diego, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Amnesia refers to difficulty in learning new information or in remembering the past. It is important to distinguish the amnesia that occurs following brain injury or disease (neurological amnesia) from the rarer functional (or psychogenic) amnesia that can occur as the result of an emotional trauma. Neurological amnesia has a variety of origins, including prolonged alcoholism, a temporary loss of blood supply or oxygen to the brain, and diseases such as herpes simplex encephalitis. All these conditions preferentially damage the medial temporal lobe or diencephalon. Neurological amnesia causes severe difficulty in learning new facts and events (anterograde amnesia). Amnesic patients also typically have some difficulty remembering facts and events that were acquired before the onset of amnesia (retrograde amnesia). Functional amnesia shows a different pattern of anterograde and retrograde memory impairment. Functional amnesia is characterized by a profound retrograde amnesia that is transient in some cases, and little or no anterograde amnesia is exhibited.

Functional Amnesia

Functional amnesia, also known as dissociative amnesia, is a dissociative psychiatric disorder that involves alterations in consciousness and identity. Although no particular brain structure or brain system is implicated in functional amnesia, the cause of the disorder must be due to abnormal brain function of some kind. Its presentation varies considerably from individual to individual, but in most cases, functional amnesia is preceded by physical or emotional trauma and occurs in association with some prior psychiatric history. Often, the patient is admitted to the hospital in a confused or frightened state. Memory for the past is lost, especially autobiographical memory and even personal identity. Semantic or factual information about the world is often preserved, though factual information about the patient's life may be unavailable. Despite profound impairment in the ability to recall information about the past, the ability to learn new information is usually intact. The disorder often clears, and the lost memories return. Occasionally, the disorder lasts longer, and sizable pieces of the past remain unavailable.

Etiology of Neurological Amnesia

Neurological amnesia results from a number of conditions, including Alzheimer's disease or other dementing illnesses, temporal lobe surgery, chronic alcohol abuse, encephalitis, head injury, anoxia, ischemia, infarction, and the rupture and repair of an anterior communicating artery aneurism. The common factor in all these conditions is that they disrupt normal function in one of two areas of the brain – the medial aspects of the temporal lobe and the diencephalic midline. Global amnesia results from bilateral damage, whereas material-specific amnesia results from unilateral damage. Typically, left-sided damage affects memory for verbal material, and right-sided damage affects memory for nonverbal material (e.g., the recall of faces and spatial layouts).

Anatomy

Well-studied cases of human amnesia and animal models of amnesia provide information about the neural connections and structures that are damaged in neurological amnesia. Damage limited to the hippocampus itself is sufficient to cause amnesia. For example, in one carefully studied case of amnesia (patient R.B.), the only significant damage was a bilateral lesion confined to the CA1 field of the hippocampus. The severity of memory impairment is exacerbated by additional damage outside the hippocampus. Thus, severe amnesia results when damage extends beyond the hippocampus to include adjacent structures in the medial temporal lobe, including the parahippocampal cortex, entorhinal cortex, and perirhinal cortex. Another well-studied case (H.M.) had surgery in 1953 to treat severe epilepsy. Most of the hippocampus and much of the surrounding medial temporal lobe cortices were removed bilaterally (the entorhinal cortex and most of the perirhinal cortex). Although the surgery was successful in reducing the frequency of H.M.'s seizures, it resulted in a severe and persistent amnesia.

Functional magnetic resonance imaging (fMRI) of healthy individuals who are engaged in learning and remembering reveals neural activity in the same structures that, when damaged, cause amnesia. It is also possible through structural imaging (MRI) to detect and quantify the neuropathology in amnesic patients. Many patients with restricted hippocampal damage have an average reduction in hippocampal volume of about 40%. Two such patients whose brains were available for detailed, postmortem neurohistological analysis (patients L.M. and W.H.) proved to have lost virtually all the neurons in the cornu ammonis (CA)

fields of the hippocampus. These observations suggest that a reduction in hippocampal volume of approximately 40%, as estimated from MRI scans, likely indicates the near complete loss of hippocampal neurons. The amnesic condition is associated with neuronal death and tissue collapse, but the tissue does not disappear altogether because fibers and glial cells remain.

As questions about amnesia and the function of medial temporal lobe structures have become more sophisticated, it has become vital to obtain detailed, quantitative information about the damage in the patients being studied. In addition, single-case studies are not nearly so useful as group studies involving well-characterized patients. In the case of patients with restricted hippocampal damage, one can calculate the volume of the hippocampus itself as a proportion of total intracranial volume. One can also calculate the volumes of the adjacent medial temporal lobe structures (the perirhinal, entorhinal, and parahippocampal cortices), again in proportion to intracranial volume. Last, when there is extensive damage to the medial temporal lobe, it is important to calculate the volumes of lateral temporal cortex and other regions that might be affected. It is important to characterize patients in this way in order to address the kinds of questions now being pursued in memory research.

To understand the anatomy of human amnesia, and ultimately the anatomy of normal memory, animal models of human amnesia have been established in the monkey and in the rodent. In the monkey, following lesions of the bilateral medial temporal lobe or diencephalon, memory impairment is exhibited on the same kinds of tasks of new learning ability that human amnesic patients fail. Cumulative work with animal models suggests that the full medial temporal lobe memory system consists of the hippocampus and adjacent, anatomically related structures, including the entorhinal cortex, parahippocampal cortex, and perirhinal cortex (see Figure 1). When these adjacent structures are damaged, the severity of amnesia is greater than when only the hippocampus itself is damaged.

The important structures in the diencephalon are the mediodorsal thalamic nucleus, the anterior thalamic nucleus, the internal medullary lamina, the mammillary nuclei, and the mammillo-thalamic tract. Because diencephalic amnesia resembles medial temporal lobe amnesia in the pattern of sparing and loss, these two regions likely form an anatomically linked, functional system.

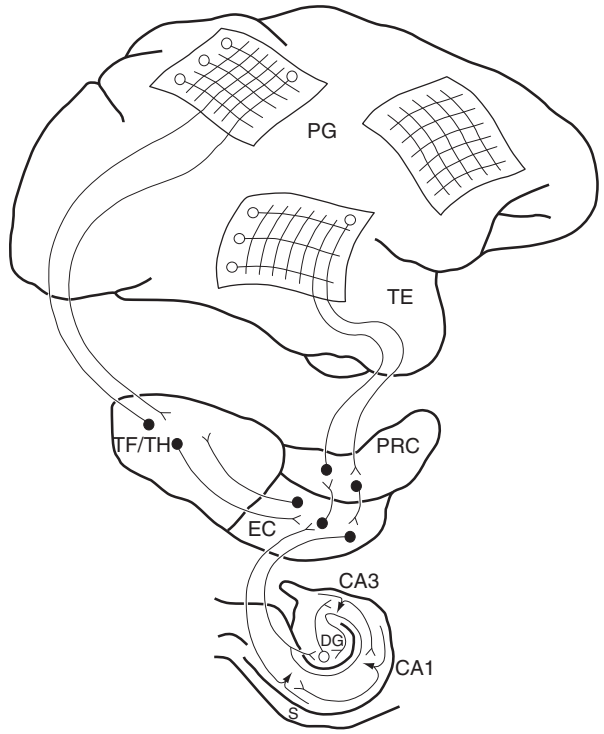

Figure 1 Schematic drawing of primate neocortex together with the structures and connections in the medial temporal region important for establishing long-term memory. The networks in the cortex show putative representations concerning visual object quality (in area TE) and object location (in area PG). If this disparate neural activity is to cohere into a stable long-term memory, convergent activity must occur along projections from these regions to the medial temporal lobe. Projections from neocortex arrive initially at the parahippocampal gyrus (TF/TH) and perirhinal cortex (PRC) and then at entorhinal cortex (EC), the gateway to the hippocampus. Further processing of information occurs in the several stages of the hippocampus, first in the dentate gyrus (DG) and then in the CA3 and CA1 regions. The fully processed input eventually exits this circuit via the subiculum (S) and the EC, where widespread efferent projections return to neocortex. The hippocampus and adjacent structures are thought to support the stabilization of representations in distributed regions of neocortex (e.g., TE and PG) and to support the strengthening of connections between these regions. Subsequently, memory for a whole event (for example, a memory that depends on representations in both TE and PG) can be revivified even when a partial cue is presented. Damage to the medial temporal lobe system causes anterograde and retrograde amnesia. The severity of the deficit increases as damage involves more components of the system. Once sufficient time has passed, the distributed representations in neocortex can operate independent of the medial temporal lobe. (This diagram is a simplification and does not show diencephalic structures involved in memory function.)

The Nature of Amnesia

It is important to appreciate that amnesic patients are not impaired at all kinds of memory. The major distinction is between declarative and nondeclarative memory. Only declarative memory is affected in amnesia. Declarative memory refers to the capacity to remember the facts and events of everyday life. It is the kind of memory that is meant when the term 'memory' is used in ordinary language. A declarative memory can be brought to mind as a conscious recol-

lection. Declarative memory provides a way to model the external world, and in this sense it is either true or false. The stored representations are flexible and can guide successful performance under a wide range of test conditions. Finally, declarative memory is especially suited for rapid learning and for forming and maintaining associations between arbitrarily different kinds of material (e.g., learning to associate two different words).

Anterograde Amnesia

Amnesia is characterized especially by profound difficulty in new learning. This impairment is referred to as anterograde amnesia. Amnesia can occur as part of a more global dementing disorder that includes other cognitive deficits, including impairments in language, attention, visuospatial abilities, and general intellectual capacity. However, amnesia can also occur in the absence of other cognitive deficits and without any change in personality or social skills. In this more circumscribed form of amnesia, patients have intact intellectual functions and intact perceptual functions, even on difficult tests that require the ability to discriminate between similar images containing overlapping features. Patients also have intact immediate memory (as measured, for example, by the ability to repeat a short string of digits). Their intact immediate memory explains why amnesic patients can carry on a conversation and appear quite normal to the casual observer. Indeed, if the amount of material to be remembered is not too large (e.g., a three-digit number), then patients can remember the material for minutes, or as long as they can hold it in mind by rehearsal. One would say in this case that the patients have carried the contents of immediate memory forward by engaging in explicit rehearsal. This rehearsal-based activity is referred to as working memory. The difficulty for amnesic patients arises when an amount of information must be recalled that exceeds immediate memory capacity (typically, when a list of eight or more items must be remembered) or when information must be recalled after a distraction-filled interval or after a long delay. In these situations, patients will remember fewer items than will their healthy counterparts.

Amnesic patients are impaired on tasks of new learning, regardless of whether memory is tested by free recall, recognition (e.g., presenting an item and asking whether it was previously encountered), or cued recall (e.g., asking for recall of an item when a hint is provided). In addition, the memory impairment involves not just difficulty in learning about specific episodes and events that occurred in a certain time and place (episodic memory), but also difficulty in learning factual information (semantic memory). Finally, the memory deficit is present regardless of the sensory modality in which information is presented (visual, auditory, olfactory, and so on).

Retrograde Amnesia

In addition to impaired new learning, amnesia also impairs memories that were acquired before the onset of amnesia. This type of memory loss is referred to as retrograde amnesia. Retrograde amnesia is usually temporally graded. That is, information acquired in the distant past (remote memory) is spared relative to more recent memory. The extent of retrograde amnesia can be relatively short and encompass only 1–2 years, or it can be more extensive and cover a much longer time. For example, an amnesic patient can have retrograde amnesia covering the previous one or two decades. In contrast, memories for the facts and events of childhood and adolescence can be intact. The severity and extent of retrograde amnesia is determined by the locus and extent of damage. Patients with restricted hippocampal damage have a limited retrograde amnesia covering a few years prior to the onset of amnesia. Patients with large medial temporal lobe damage have extensive retrograde amnesia covering decades.

The sparing of remote memory relative to more recent memory illustrates that the brain regions damaged in amnesia are not the permanent repositories of long-term memory. Instead, memories undergo a process of reorganization and consolidation after learning, during which time the neocortex becomes more important. During the process of consolidation, memories are vulnerable if there is damage to the medial temporal lobe or diencephalon. After sufficient time has passed, storage and retrieval of memory no longer require the participation of these brain structures. Memory is at that point supported by neocortex. The areas of neocortex important for long-term memory are thought to be the same regions that were initially involved in the processing and analysis of what was to be learned. Thus, the neocortex is always important, but the structures of the medial temporal lobe and diencephalon are also important during initial learning and during consolidation.

Spatial Memory

Discussions of amnesia have focused especially on the status of spatial memory because of the discovery of 'place cells' in the rodent hippocampus and the possible importance of the hippocampus in forming spatial maps. In human amnesia, spatial memory is impaired along with other forms of declarative memory. Patients have difficulty acquiring new spatial knowledge, and they are impaired in remembering recently acquired spatial knowledge. However, as is the case with other forms of declarative memory, remote spa-

tial knowledge is intact. One well-studied patient with large medial temporal lobe lesions and severe amnesia (E.P.) was able to mentally navigate his childhood neighborhood, use alternate and novel routes to describe how to travel from one place to another, and point correctly to locations in the neighborhood while imagining himself oriented at some other location. These findings show that the medial temporal lobe is not needed for the long-term storage of spatial knowledge and does not maintain a spatial layout of learned environments that is necessary for successful navigation. Accordingly, the available data support the view that the hippocampus and related medial temporal lobe structures are involved in learning new facts and events, both spatial and nonspatial. Further, these structures are not repositories of long-term memory, either spatial or nonspatial.

Nondeclarative Memory

It is a striking feature of amnesia that many kinds of learning and memory are spared. Memory is not a unitary faculty of the mind but is composed of many parts that depend on different brain systems. Amnesia impairs only declarative memory and spares nondeclarative memory. Nondeclarative memory refers to a heterogeneous collection of abilities, all of which afford the capacity to acquire knowledge nonconsciously. Nondeclarative memory includes motor skills, perceptual and cognitive skills, priming, adaptation-level effects, simple classical conditioning, and habits, as well as phylogenetically early forms of experience-dependent behavior such as habituation and sensitization. In these cases, memory is expressed through performance rather than recollection, and performance does not require reflection on the past or even the knowledge that memory is being influenced by past events. For example, in the case of motor skills, one can learn how to ride a bicycle but be unable to describe what has been learned, at least not in the same sense that one might recall riding a bicycle on a particular day with a friend. Perceptual skills include such things as reading mirror-reversed print and searching a display quickly to find a hidden letter. In formal experiments, amnesic patients acquire perceptual skills at the same rate as individuals with intact memory, even though the patients may not remember performing the task.

Priming refers to an improved ability to identify a word or other item as a result of its prior presentation. For example, suppose that a line drawing of a dog, hammer, and airplane are presented in succession, with the instruction to name each item as quickly as possible. Typically, about 800 ms are needed to produce each name aloud. If in a later test these same pictures are presented intermixed with new drawings, the new drawings will still require about 800 ms to name, but now the dog, hammer, and airplane are named about 100 ms more quickly. The improved naming time occurs independent of whether one remembers having seen the items earlier. Furthermore, amnesic patients exhibit this effect at full strength, despite having a poor memory of seeing the items earlier. Priming effects of this kind can persist across intervals as long as several weeks. In formal experiments, severely amnesic patients had intact priming for recently presented words, even when the patients performed at guessing levels (50% correct) on tests that asked them to recognize which words were presented previously and which were not. This result shows that priming is fully independent of declarative memory.

Adaptation-level effects refer to changes in judgments about stimuli (e.g., their heaviness or size) that are caused by recent experience. For example, experience with light-weighted objects subsequently causes other objects to be judged heavier than they would be if the light-weighted objects had not been presented. Amnesic patients show this effect to the same degree as healthy individuals, though they have difficulty remembering what they have done.

Classical conditioning refers to the development of an association between a previously neutral stimulus and an unconditioned stimulus. One of the best-studied examples of classical conditioning in humans is eyeblink conditioning. In a typical conditioning procedure, a tone repeatedly precedes a mild air puff directed to the eye. After a number of pairings, the tone comes to elicit an eyeblink in anticipation of the air puff. Amnesic patients acquire the tone-air puff association at the same rate as healthy individuals do. In both groups, awareness of the temporal contingency between the tone and the air puff is unrelated to successful conditioning. Simple classical conditioning, where the tone overlaps with the air puff and terminates with it, is dependent on the cerebellum.

Habit learning refers to the gradual acquisition of associations between stimuli and responses, such as learning to make one choice rather than another. Habit learning depends on the neostriatum (basal ganglia). Many tasks can be acquired either declaratively, through memorization, or nondeclaratively, as a habit. For example, healthy individuals will solve many trial and error learning tasks quickly by simply engaging declarative memory and memorizing which responses are correct. In this circumstance, amnesic patients are disadvantaged. However, tasks can also be constructed that defeat memorization strategies, for example, by making the outcomes on each trial probabilistic. In such a case, amnesic patients and healthy individuals learn at the same

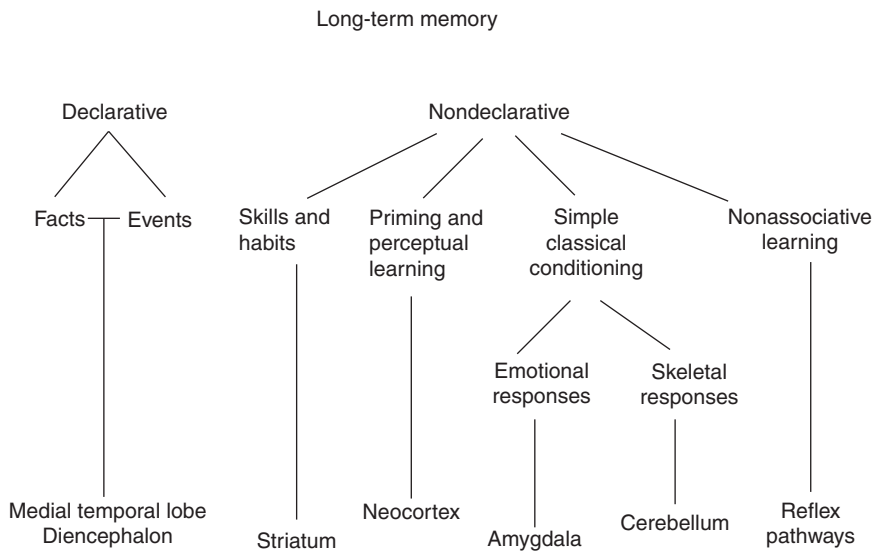

Figure 2 Classification of mammalian long-term memory systems. The taxonomy lists the brain structures thought to be especially important for each form of declarative and nondeclarative memory. In addition to the central role of the amygdala in emotional learning, it is able to modulate the strength of both declarative and nondeclarative learning.

gradual rate. It is also true that severely amnesic patients who have no capacity for declarative memory can gradually acquire trial-and-error tasks, even when the task can be learned declaratively by healthy individuals. In this case they succeed by engaging habit memory.

This situation is nicely illustrated by the eight-pair concurrent discrimination task, which requires individuals to learn the correct object for each of eight object pairs. Healthy individuals learn all eight pairs in a single test session. Severely amnesic patients acquire this same task over many weeks, even though at the start of each session they cannot describe the task, the instructions, or the objects. It is known that this task is acquired at a normal (slow) rate by monkeys with medial temporal lobe lesions and that monkeys with lesions of the neostriatum (basal ganglia) are impaired. Thus, humans appear to have a robust capacity for habit learning that operates outside awareness and independent of the medial temporal lobe structures that are damaged in amnesia.

These examples illustrate that nondeclarative memory is distinct from declarative memory. It is spared in amnesia, and it operates outside awareness. Nondeclarative forms of memory depend variously on the neostriatum, the amygdala, the cerebellum, and on processes intrinsic to neocortex (Figure 2).

Summary

The study of amnesia has illuminated the nature of memory disorders and has also led to a better understanding of the neurological foundations of memory. Experimental studies in patients, neuroimaging studies of healthy volunteers, and related studies in experimental animals continue to reveal insights about what memory is and how it is organized in the brain. As more is learned about the neuroscience of memory, about how memory works, more opportunities will arise for achieving better diagnosis, treatment, and prevention of diseases and disorders that affect memory.

See also: Animal Models of Amnesia; Episodic Memory; Functional Amnesia; Hippocampus: Computational Models; Memory Disorders; Recognition Memory; Semantic Memory.

Further Reading

Bayley PJ, Frascino JC, and Squire LR (2005) Robust habit learning in the absence of awareness and independent of the medial temporal lobe. *Nature* 436: 550–553.

Bayley PJ, Gold JJ, Hopkins RO, et al. (2005) The neuroanatomy of remote memory. *Neuron* 46: 799–810.

Eichenbaum H and Cohen NJ (2001) *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. New York: Oxford University Press.

Gordon B (1995) *Memory: Remembering and Forgetting in Everyday Life*. New York: Mastermedia.

Milner B, Squire LR, and Kandel ER (1998) Cognitive neuroscience and the study of memory. *Neuron* 20(3): 445–468.

Schacter DL (1996) *Searching for Memory: The Brain, the Mind, the Past*. New York: BasicBooks.

Squire LR and Kandel ER (2000) *Memory: From Mind to Molecules*. New York: W. H. Freeman.

Squire LR and Schacter DL (eds.) (2002) *The Neuropsychology of Memory*, 3rd edn. New York: Guilford Press.

Squire LR, Stark CE, and Clark RE (2004) The medial temporal lobe. *Annual Review of Neuroscience* 27: 279–306.

Teng E and Squire LR (1999) Memory for places learned long ago is intact after hippocampal damage. *Nature* 400: 675–677.

Wais PE, Wixted JT, Hopkins RO, and Squire LR (2006) The hippocampus supports both the recollection and the familiarity components of recognition memory. *Neuron* 49: 459–468.

Relevant Websites

http://www.alzforum.org – Alzheimer's Research Forum.
http://whoville.ucsd.edu – Home page of Larry R. Squire.
http://www.nia.nih.gov – National Institute on Aging, Alzheimer's Disease Centers.

AMPA Receptor Cell Biology/Trafficking

P G R Hastie and J M Henley, University of Bristol, Bristol, UK

© 2009 Elsevier Ltd. All rights reserved.

α-Amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) are ligand-gated ion channels that are crucial for normal brain function. They mediate nearly all fast excitatory synaptic transmission in the mammalian central nervous system (CNS). Changes in functional postsynaptic AMPARs mediate the two main forms of synaptic plasticity that are believed to underlie learning and memory. Long-term potentiation (LTP) involves the activity-dependent recruitment of AMPARs to the synapse and a concurrent increase in AMPA-mediated transmission. Conversely, long-term depression (LTD) is a decrease in synaptic AMPAR function. Given their importance, AMPARs are stringently regulated and the mechanisms of this regulation have been, and continue to be, the subject of intense study.

Subunit Composition of AMPARs

AMPARs are tetrameric assemblies of combinations of four individual subunits (glutamate receptor (GluR)1–4). Based on the length of the intracellular C-terminal domains of their predominant splice isoforms, GluR2 and GluR3 are classified as short and GluR1 and GluR4 are classified as long subunits. As discussed here, short and long subunits bind to different sets of intracellular proteins and are trafficked differently.

AMPAR Trafficking

There are three fundamental ways of targeting proteins in polarized cells:

1. Selective delivery in which cargo is sent directly to the target destination.
2. Selective fusion – exocytic machinery packaged with the cargo fuses with a particular membrane subregion.
3. Selective retention – cargo is removed from inappropriate membranes.

In general, dendritic sorting most likely relies on selective delivery because green fluorescent protein (GFP) attached to proteins containing dendritic sorting signals traffics directly to the somatodendritic compartment. Live imaging with GFP-labeled GluR1 and GluR2 suggests that AMPARs can be independently transported intracellularly rather than in clusters. This is consistent with a process of selective fusion acting as the regulated step in synaptic delivery.

Endoplasmic Reticulum to Synapse Trafficking Pathway

AMPARs are delivered to the synapse via the secretory pathway (Figure 1). A YxxLxxR motif targets the translated proteins to the endoplasmic reticulum (ER) for folding, assembly of dimers, and enzyme modification. GluR1 is able to rapidly exit the ER, but RNA editing in GluR2 delays the exit of this subunit. Therefore, a large pool of GluR2 is held available in the ER, ready to combine with other subunits, and thus most hippocampal AMPARs contain GluR2. An extreme N-terminal signal may further promote GluR1 exit from the ER because truncation mutants lacking these nine amino acids co-localize with ER markers despite forming heteromers with GluR2. It is believed that the assembly of dimers into tetramers (a dimer of dimers) leads to masking of the retention signals, allowing only correctly folded, desensitized AMPARs to exit the ER.

AMPARs next enter the *trans*-Golgi network (TGN) and become fully glycosylated. Passage through this pathway is dependent on the C-terminus of the receptor, a site for interaction with a number of proteins, particularly those containing postsynaptic density protein (PSD-)95/Dlg-A/ZO-1 (PDZ) domains. PDZ domains are protein interaction modules that, depending on the type of PDZ motif, can bind selectively to a range of peptide ligand sequences present on target proteins.

Anterograde trafficking of AMPARs occurs along microtubules. One candidate for mediating forward trafficking is mLIN-10/mint1/X11, the *Caenorhabditis elegans* ortholog of which, LIN-10, is required for the correct localization of GLR-1 receptors. The microtubule-associated kinesin motor protein KIF5 binds to the AMPAR binding protein (ABP) glutamate receptor interacting protein 1 (GRIP1) and is involved in selective dendritic delivery. Further evidence that AMPARs travel along microtubules comes from another GRIP1-interacting partner, liprin-α/SYD2, which links to the scaffold protein GIT1 and to microtubules via KIF1A. Interference with either of these interactions specifically reduces GluR2/3 dendritic clustering. In the case of GluR1, linkage to microtubules can occur through a synapse-associated protein (SAP-)97 interaction with KIF1Bα.

AMPAR synaptic delivery to and insertion at dendritic spines are two distinct processes, employing the

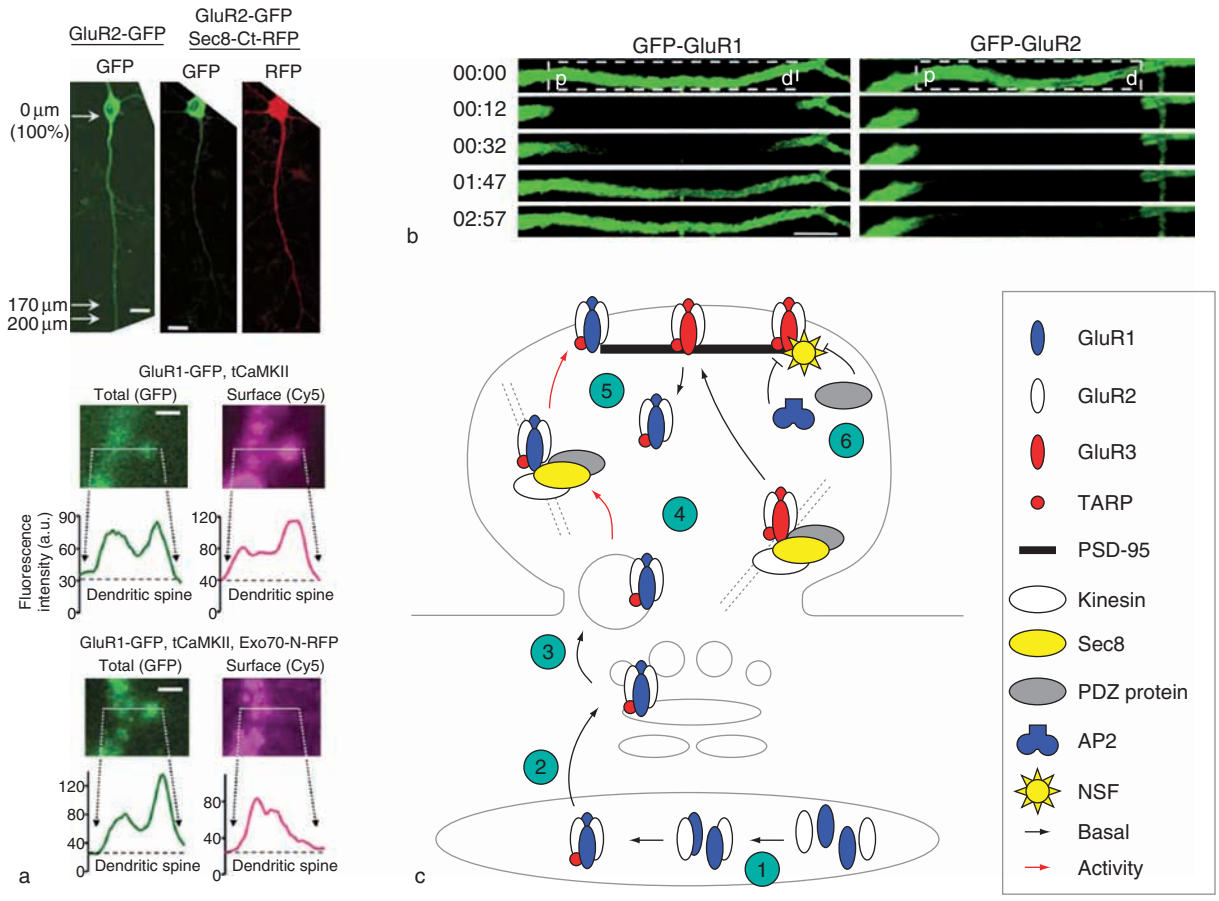

Figure 1 Delivery of AMPARs: (a) AMPAR synaptic delivery and insertion as two distinct processes; (b) differential transport rates of AMPAR subunits; (c) AMPAR synaptic delivery steps. In (a), a dominant negative C-terminal construct of the exocyst component Sec8 prevents the delivery of GluR2-GFP to the distal dendrites (upper panel). A dominant negative construct of another exocyst component, Exo70, prevents the surface insertion of CaMKII phosphorylated GluR1-GFP, causing the accumulation of correctly delivered AMPARs within the spine. In (b), 40 µm segments of cultured hippocampal dendrites expressing either GluR1-GFP or GluR2 were photobleached. Fluorescence recovery from both distal (d) and proximal (p) dendritic ends was greatest for GluR1-GFP, suggesting faster rates of transport and/or greater surface mobility for this subunit. (c) Shows the events in the AMPAR synaptic delivery: (1) individual AMPAR subunits enter the ER and form dimers and then a dimers-of-dimers; Q/R editing in GluR2 prevents GluR2 homomer assembly and thus most AMPARs contain GluR2; (2) ER exit is dependent on the correct assembly of the tetrameric complex, which must be able to enter a desensitized state to exit the ER; TARPs are also required for ER exit; (3) high-mannose sugars on the extracellular surface of the AMPAR are replaced by complex carbohydrates in the Golgi; AMPARs most likely pass through a sorting endosome before being delivered to the surface, with GluR1-containing AMPARs exiting this compartment more slowly than GluR1-lacking AMPARs; (4) anterograde trafficking of AMPARs occurs along microtubules as complexes containing either GluR1-K1F1b-SAP97 or GluR2-KIF5-GRIP; (5) synaptic activity is required for surface delivery of GluR1, which is accompanied by the delivery of slot proteins (e.g., PSD-95) to the synapse; GluR1-containing AMPARs are then replaced by constitutively cycling GluR2/3 heteromers; (6) delivery of AMPARs to the synapse is dependent on an interaction between GluR2 and NSF; this interaction prevents internalization due to AP2 and PICK1 binding to the receptor, thus stabilizing AMPARs at the synapse. AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; a.u., arbitrary units; CaMKII, calmodulin-dependent protein kinase II; ER, endoplasmic reticulum; GFP, green fluorescent protein; GluR, glutamate receptor; NSF, *N*-ethyl maleimide sensitive factor; PDZ, PSD-95/Dlg-A/ZO-1; PICK1, protein interacting with C kinase 1; PSD-95, postsynaptic density protein 95; Q/R, glutamine to arginine; RFP, red fluorescent protein; TARPS, transmembrane AMPAR interacting proteins. (a, middle and bottom) From Gerges NZ, Backos DS, Rupasinghe CN, Spaller MR, and Esteban JA (2006) Dual role of the exocyst in AMPA receptor targeting and insertion into the postsynaptic membrane. *EMBO Journal* 25: 1623–1634. (b) Adapted from Perestenko PV and Henley JM (2003) Characterisation of the intracellular transport of GluR1 and GluR2 α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor subunits in hippocampal neurons. *Journal of Biological Chemistry* 278: 43525–43532.

exocyst, a multisubunit complex required for vectorial targeting of a subset of secretory vesicles. The exocyst complex acts to fuse vesicles to the plasma membrane. Proteins on one end recognize Rab GTPases on secretory vesicles, whereas proteins on the other end localize to the plasma membrane, probably by binding to Rho GTPases that mediate the activity of the cytoskeleton. The proteins in between

may act as a structural core to transmit information between the two membranes. A dominant negative C-terminal construct of the exocyst component Sec8 prevents the delivery of GluR2-GFP to distal dendrites. A dominant negative construct of another exocyst component, Exo70, prevents the surface insertion of GluR2-GFP and calmodulin-dependent protein kinase (CaMK)II-phosphorylated GluR1-GFP, causing the accumulation of otherwise correctly delivered AMPARs within the spine.

Roles of Short and Long AMPAR Subunits

The long and short AMPAR subunits appear to serve different functions in AMPAR surface expression, and a subunit-specific AMPAR trafficking model has been proposed to explain how LTP is expressed and maintained. To summarize, the forward trafficking characteristics of the long-tail GluR1 subunit are dominant and AMPARs containing this subunit are inserted into the synapse during periods of activity. AMPARs lacking GluR1 default to the short-tail GluR2 subunit characteristics of constitutive cycling. GluR2-mediated cycling replaces synaptic AMPARs without changing synaptic efficacy. Therefore, GluR1-containing AMPARs are inserted during the initial phase of LTP, and these are then replaced by GluR2/3 heteromers. This hypothesis is based on the following experimental observations.

Constitutive Cycling of AMPARs Is *N*-Ethyl Maleimide-Sensitive Factor-Dependent

Plasma membrane insertion of $GluR2_{short}$ is much more rapid than for $GluR1_{long}$ under basal conditions, and the delivery of GluR2 continues when neuronal activity is blocked. However, surface expression of GluR2 is short lived; the blockade of exocytic events leads to a rapid run-down of AMPAR responses. GluR2-containing AMPAR insertion and/or stabilization at the synapse requires the hexameric adenosine triphosphatase (ATPase) *N*-ethyl maleimide-sensitive factor (NSF). NSF binds the GluR2 C-terminus upstream of the PDZ binding site between residues Lys844 and Gln853. Blocking this interaction with a peptide (pep2m) corresponding to the interaction site on GluR2 causes a decrease in surface GluR2, indicating the rapid internalization and recycling back to the membrane of GluR2-containing receptors. Pep2m also prevents *de novo* LTD. However, pep2m also blocks an overlapping binding site for the clathrin adaptor protein AP2, which interacts with GluR1–3 subunits and is required for AMPAR endocytosis. A more specific peptide, pep-R845A, which blocks only the GluR2–NSF interaction, leads to the loss of functional AMPARs but has no effect on LTD.

Activity-Dependent Insertion of AMPARs

Insertion of $GluR1_{long}$-containing AMPARs at the hippocampal synapses requires synaptic activity and occurs following *N*-methyl-D-aspartate receptor (NMDAR)-dependent LTP induction. Receptor phosphorylation is critical in the process, and in older neurons CaMKII activation is required, whereas protein kinase A (PKA) activation is more important in younger neurons. Although GluR1 is responsible for the majority of activity-dependent insertion in adults, GluR4 and the long C-terminal variant of GluR2 contribute significantly in younger animals when the expression of these subunits is prevalent.

In GluR1, the PDZ protein binding region comprises a –TGL motif. This is a type I PDZ ligand and therefore binds a different subset of proteins than the type II –SVKI motif present in short AMPAR subunits. SAP-97 directly interacts with the GluR1 PDZ ligand, and the overexpression of SAP-97 increases AMPAR surface delivery; but truncation mutants eliminating GluR1 binding do not affect AMPAR clustering and mice lacking the –TGL motif have normal levels of synaptic GluR1, suggesting that SAP-97 may be important in relieving intracellular retention interactions at this site.

There is close similarity between the signal transduction pathways for LTP and long-tailed AMPAR synaptic insertion. Activation of the Ras GTPase pathway during LTP leads to specific synaptic insertion of long-tailed AMPARs. Low-level activation leads to extracellular signal-regulated kinase kinase (MEK) and p42/44 mitogen-activated protein kinase (MAPK) signaling, leading to delivery of $GluR2_{long}$. Higher levels of activity stimulate the PI3K and protein kinase B (PKB) pathways leading to GluR1 delivery. This action of $GluR2_{long}$ may explain the persistent LTP in mice lacking the GluR1 PDZ ligand.

Maintenance of Increased Synaptic AMPARs Requires Slot Proteins

An important aspect of this model, which depends on an initial increase in surface expression due to long subunits but the maintenance of the increased expression via short subunits, is the need for marker or slot proteins to mark the potentiated synapses. Once trafficked to the PSD, these slot proteins remain there to maintain the increased expression of short subunit AMPARs unless a signal such as LTD induction triggers their removal. They therefore can act as placeholders, defining the number of AMPAR insertion sites at the synapse. If these were not added to the synapse, the internalized GluR1 could not be replaced by GluR2/3 because there would be no scaffolding structure for these new short AMPARs to attach to.

AMPARs Interact with a Complex Network of Scaffold Proteins

Cytoskeletal interactions at the synapse are both dynamic and complex, and there is likely to be redundancy between many of the binding partners. The most direct links with the actin cytoskeleton occur through long AMPAR subunits. Binding of protein 4.1N to the C-terminus of GluR1 and GluR4 upstream of the phosphorylation sites controls the surface expression of these subunits, possibly mediated through SAP-97 because the removal of the 4.1N binding site on this protein reduces AMPAR clustering. SAP-97 also links AMPARs to the actin cytoskeleton through myosin VI, although this minus-end-directed motor protein is involved in removal of AMPARs. A further interaction is provided by reversion-induced Lin11/rat Isl-1/Mec3 (LIM) domain-containing protein (RIL). Unusually, it is the LIM domain that specifically binds GluR1, whereas the PDZ domain links to the actin cytoskeleton via α-actinin, increasing the surface accumulation of AMPARs at the spines.

PSD-95 Regulates Synaptic AMPAR Content

Postsynaptic density protein of 95 kDa (PSD-95), also known as synaptic-associated protein 90 (SAP-90), represents 2.3% of the protein mass of the PSD (equating to ~300 PSD-95 molecules per synapse). This protein is a member of the membrane-associated guanylate kinase (MAGUK) family due to the guanylate kinase (GK) homology domain in the C-terminus. The C-terminus also contains a src homology (SH3) domain. The N-terminus contains a consensus sequence for palmitoylation. Three PDZ domains make PSD-95 an effective scaffold protein.

PSD-95 is required for AMPAR synaptic localization, and there is a correlation between PSD-95 and AMPAR accumulation at the synapse. Expression of PSD-95-GFP increases the synaptic delivery of GluR1, thus occluding LTP, and increases the AMPAR-mediated currents in experience-deprived animals. Furthermore, PSD-95 gene transcription is upregulated by neuronal activity. An attractive model is that activity-dependent delivery of PSD-95 occurs prior to, and enables, the recruitment and insertion of GluR1-containing AMPARs and then acts as the slot protein for subsequent GluR2/3 AMPARs.

Transmembrane AMPAR Interacting Proteins Are Required to Anchor AMPARs to PSD-95

By definition, the five members of the transmembrane AMPAR interacting protein (TARP) family (γ2/stargazin, γ3, γ4, γ7, and γ8) bind AMPARs. This separates them from closely related proteins such as the γ1 calcium channel subunit. Mutations in γ2/stargazin lead to ataxia in waggler mice, which lack surface AMPARs on cerebellar granule cells. It has been suggested that TARPs constitute auxillary AMPAR subunits and that a type I PDZ domain at the C-terminus of TARPs enables interaction with PSD-95, SAP-97, and other MAGUKs. TARPs are involved in both AMPAR surface expression and lateral movement of AMPARs in the membrane to the synapse via an interaction with PSD-95. Whereas stargazin/γ2 is highly expressed in the cerebellum, γ8 is most highly expressed in the hippocampus and is not mutated in the waggler mice; this explains why they show locomotor but not learning deficiencies.

GRIP and ABP Anchor Short AMPAR Subunits

GRIP1 and ABP both contain a maximum of seven PDZ domains and interact with short AMPAR subunits through PDZ domains 4–5 and 3, 5, and 6, respectively. The number of PDZ domains and their ability to form homo- and heterodimers make GRIP1 and ABP ideal candidate scaffolds. Both proteins are enriched in dendritic spines and it is proposed that they anchor AMPARs at the synapse. Ser880Ala mutation in the GluR2 PDZ domain selectively prevents its interaction with GRIP/ABP, leading to a reduction in synaptic accumulation of GluR2 over time. Synaptic targeting was not affected, suggesting GRIP/ABP is specifically required for the synaptic retention of AMPARs. Thus, GRIP and ABP possibly work in tandem with PSD-95 to provide slot proteins for AMPAR insertion and retention at the synapse.

Neuronal Activity Regulated Pentraxin-Induced Clustering of AMPARs

The N-terminus of GluR1 is required for ER exit and synaptic targeting. Neuronal activity regulated pentraxin (NARP), an immediate-early gene released from the stimulated hippocampus, interacts at an extracellular site causing the clustering of GluR1–3 in heterologous cells and co-immunoprecipitates with these subunits, as well as clustering GluR1 in cultured spinal neurons. Endogenous NARP action varies between neurons. Hippocampal axons only secrete NARP at contacts with interneurons, and the exogenous application of NARP fails to cause clustering on pyramidal cells. Spinal neurons cluster AMPARs and NMDARs on cultured hippocampal interneurons where contacts are made on the dendritic shaft but not on pyramidal cells where contacts are made at dendritic spines.

AMPAR Endocytosis

The best understood model for internalizing proteins in neurons is clathrin-mediated endocytosis (CME). Hypertonic sucrose inhibits CME and prevents both agonist-induced GluR1 internalization and insulin-induced GluR2 internalization. Blocking dynamin activity prevents the internalization of GluR1 and GluR2 and blocks LTD. A peptide corresponding to the SH3 domain of amphiphysin also blocks LTD. Disruptions to specific components of the CME machinery (e.g., CPG2 and huntingtin interacting protein 1 (HIP1)) also caused reduced AMPAR internalization. Increased colocalization and interaction of AP2 with AMPARs following agonist treatment suggested AP2 recruits AMPARs to coated pits. This is supported by evidence that the proteins directly interact through the β-adaptin subunit of AP2. This interaction does not occur with the AP1 subunit responsible for cargo recruitment at intracellular sites. It is not clear if non-CME AMPAR endocytosis pathways exist, although lipid rafts, often associated with caveolar structures, are involved in stabilizing surface AMPARs.

Lateral Diffusion to Extrasynaptic Clathrin-Mediated Endocytosis Sites

The development of ecliptic GFP or pHlourins, which fluoresce only at pH > 7.0 allowed the measurement of dynamic changes to the surface expression of proteins. Because intracellular compartments are more acidic than the extracellular environment, super eclipitic phluorin (SEP) fluorescence is suppressed until the protein is surface expressed. This technology was used to show that, in response to NMDAR activation, endocytosis of synaptic GluR2 is preceded by the removal of extrasynaptic receptors (Figure 2). An obvious reason for this is that the endocytosis machinery is located at extrasynaptic sites. Quantum dot-tracking experiments show that the proportion of mobile GluR2 increases following neuronal activity, representative of increased lateral diffusion from synaptic anchors. This suggests that, rather than the regulation of endocytosis itself, the limiting step in generating LTD through internalization is the release of AMPARs from the PSD.

Protein Interacting with C Kinase 1 Releases AMPARs from the Synapse

Protein interacting with C kinase (PICK1) interacts with a number of neuronal receptors, channels, and transporters. A single PDZ domain enables an interaction with short AMPARs leading to LTD. This occurs because PICK1 competes for the GRIP/ABP binding site on GluR2, uncoupling the AMPARs from the synaptic scaffold. This is controlled by protein kinase C (PKC) phosphorylation of GluR2 at Ser880, which prevents GRIP/ABP but not PICK1 binding. Phosphorylation is facilitated by PICK1 itself, which binds PKCα to target the kinase to AMPARs. PKCα binding removes an intramolecular interaction, exposing the BAR domain of PICK1. A second phosphorylation switch with the same function has recently been found at Tyr876, and this residue must be Src phosphorylated in response to drug treatment for internalization to occur. NSF hydrolysis of adenosine triphosphate (ATP) disrupts the PICK1–GluR2 interaction in complexes containing α-SNAP. This stabilizes the AMPARs at the surface. β-SNAP, however, prevents this dissociation and causes receptor internalization. Thus, in addition to actively inserting AMPARs at the synapse, NSF activity prevents their internalization. The presence of a BAR domain in PICK1 suggests this protein can recruit AMPARs to clathrin-coated vesicles (CCVs) in constitutive endocytosis in which AP2 does not seem to be involved.

Ca^{2+} binding to PICK1 increases the affinity of the GluR2 interaction. This has obvious implications for LTD, in which Ca^{2+} concentration is elevated in spines. The increase in affinity may be another switch in the balance between constitutive cycling and long-lasting endocytosis of AMPARs. PICK1 is not the only Ca^{2+}-sensing molecule implicated in the endocytosis of AMPARs during LTD. The neuronal calcium-sensor hippocalcin binds AP2 in a Ca^{2+}-dependent manner, and the Ca^{2+}-sensing region of hippocalcin is required for the generation of LTD. It is thought that Ca^{2+} binding, which exposes a myristyl tail in hippocalcin, recruits AP2 to the membrane, enabling AMPAR sorting into CCVs. Hippocalcin also complexes with transferrin receptors but in a Ca^{2+}-independent manner, suggesting the function of this protein is modified for a specific role in AMPAR endocytosis.

Degradation of Synaptic Scaffolds Accompanies AMPAR Internalization

PSD-95 is ubiquitinated following synaptic NMDA activation. Truncated PSD-95 mutants lacking the ubiquitination motif prevent NMDA-induced GluR2 internalization. One function of ubiquitination is as a signal for proteasomal protein degradation, suggesting that the synaptic scaffold may have to be dismantled for AMPAR release. Although this has not yet been shown for the proteasome pathway, calpain cleavage of PSD-95, SAP-97, and GRIP1 has been shown. AMPARs are themselves cleaved at the C-terminus by calpain, allowing their release from the PSD.

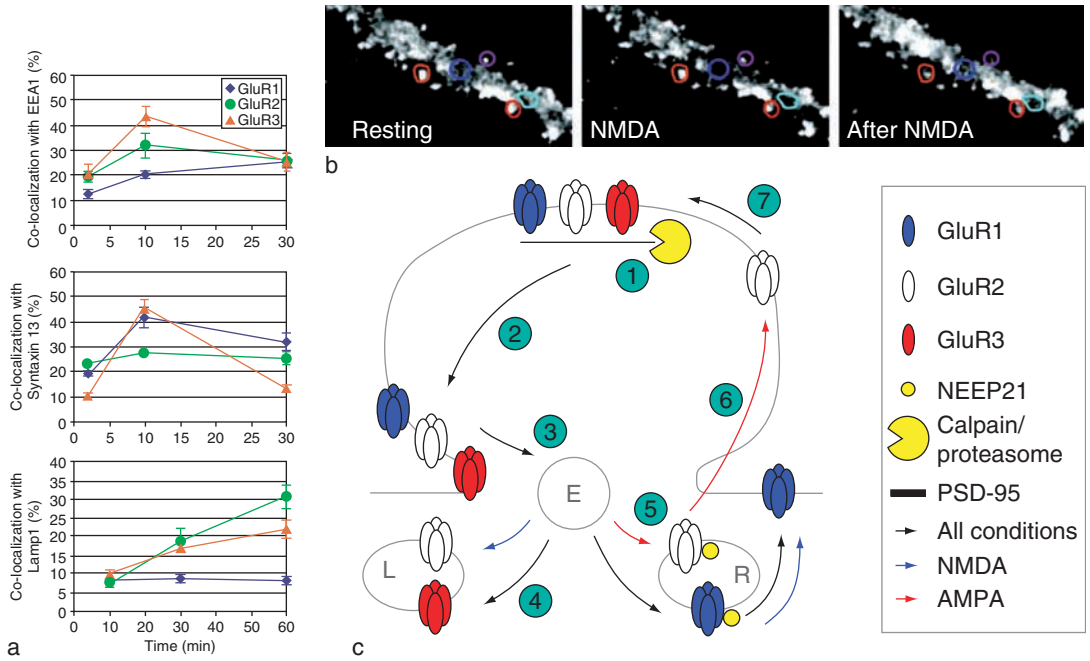

Figure 2 AMPAR internalization and recycling: (a) experiments using recombinant HA-tagged AMPAR subunits; (b) virally expressed pHluorin-GluR2 labeling resting, with NMDA, and after NMDA; (c) main steps in AMPAR recycling. In (a), experiments using recombinant HA-tagged AMPAR subunits demonstrate that they are differentially sorted on internalization. All HA-tagged AMPAR subunits are internalized following the application of 50 μM NMDA for 8 min. All subunits colocalize with the early endosome marker EEA1 after 10 min, but only GluR1 continues to accumulate in this compartment up to 30 min following NMDA treatment. Both GluR2 and GluR3 pass through the syntaxin 13-positive recycling endosome and accumulate in the Lamp1-positive late endosome prior to degradation, whereas GluR1 remains in the recycling endosome. In (b) under resting conditions virally expressed pHluorin-GluR2 labels both punctate spines (red, purple) and diffuse shaft regions (blue). Treatment with 50 μM NMDA causes an immediate loss of the diffuse staining, representative of GluR2 endocytosis. This is followed by the loss of punctate staining after approximately 10 min. (c) Shows a schematic of the main events in AMPAR recycling: (1) AMPARs are released from the PSD by switching binding partners accompanied by degradation of the PSD; (2) receptors exit the spine by lateral diffusion to sites of clathrin endocytic machinery; (3) clathrin-mediated exocytosis transports AMPARs to the early endosome; (4) under all conditions, GluR3 is sent to the late endosome prior to degradation, and GluR2 enters this pathway after NMDA treatment but is sent to the recycling endosome following AMPA treatment; (5) GluR1 constitutively enters the recycling endosome, but its rate of exit from this compartment is increased by NMDA; (6) GluR2 exit from the recycling endosome is promoted by an interaction with NEEP21; (7) GluR2 is rapidly recycled to the synapse following AMPA treatment or in the absence of activity. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; E, early endosome; EEA1, early endosome marker; GluR, glutamate receptor; HA, peptide derived from human influenza haemagglutinin; L, late endosome; Lamp1, lysosome associated membrane protein 1; NEEP21, neuron-enriched endosomal protein of 21 kDa; NMDA, *N*-methyl-D-aspartate; PSD, postsynaptic density protein; R, recyclingendosome. (a) from Lee SH, Simonetta A, and Sheng M (2004) Subunit rules governing the sorting of internalized AMPA receptors in hippocampal neurons. *Neuron* 43: 221–236. (b) Adapted from Ashby MC, De La Rue SA, Ralph GS, Uney J, Collingridge GL, and Henley JM (2004) Removal of AMPA receptors (AMPARs) from synapses is preceded by transient endocytosis of extrasynaptic AMPARs. *Journal of Neuroscience* 24: 5172–5176.

Small GTPase Signaling Triggers AMPAR Endocytosis in LTD

NMDAR activation of the Rap1–p38MAPK pathway via the NR2B subunit causes internalization of GluR2/3 heteromers. Activation of the Rap2–c-Jun N-terminal kinase (JNK) pathway causes the internalization of GluR1/2 heteromers in depotentiation, which is the resetting of synapses that have undergone LTP back to the baseline. Rab5 overexpression specifically depresses AMPAR-mediated excitatory postsynaptic currents (EPSCs) through the removal of GluR1–3 from dendritic spine surfaces. A dominant negative mutant of Rab5 prevents LTD induction. Rab5 activation is likely to occur downstream of p38MAPK activation because this regulates its interaction with Rab-GDI. As would be expected from the localization of Rab5 to endosomes, these effects of overexpression occur downstream of Ser880 phosphorylation and release from the PSD.

Fate of Internalized AMPARs Is Subunit Specific

GluR1 internalizes rapidly and independently of activity, supporting the LTP model in which GluR1-containing subunits are inserted and then replaced by GluR2/3 heteromers. Overexpressed GluR2 and GluR3 are internalized on agonist treatment, leading

to GluR2 being recycled, but NMDAR activation targets GluR2 to lysosomes. GluR1 is recycled and GluR3 is directed to endosomes independent of drug treatment (Figure 2). These subunit rules appear to be determined by the NSF binding site, with GluR2 being dominant over GluR1. NSF may promote the recycling of the receptors, and this further supports a model in which NSF activity or binding to GluR2 is reduced during LTD, allowing targeting to lysosomes. Overall, these data suggest that NMDAR activation during LTD leads to the degradation of GluR2-containing receptors, consistent with an activity-dependent depletion of AMPARs at the synapse. However, exposure to the sodium-channel blocker tetrodotoxin (TTX), which prevents action potentials and therefore evoked presynaptic glutamate release, reverses the effects of NMDA and AMPA. The physiological relevance of this is not clear, but it does highlight the importance of the prior experience of the synapse in determining the effect of plasticity protocols. It has been suggested that this might represent a form of homeostatic plasticity known as synaptic scaling, in which neuronal sensitivity is altered in line with the overall sensitivity of the network.

TTX-treated neuronal cultures have been used to show the role of proteins other than NSF involved in directing AMPARs through the endocytic pathway. Suppression of neuron-enriched endosomal protein of 21 kDa (NEEP21) reduces the rate of GluR1 recycling in response to NMDA. GluR2 internalizes to a NEEP21 positive compartment following NMDA treatment. The presence of Rab4 and syntaxin 13 in this compartment suggests that NEEP21 promotes trafficking to the recycling endosome. In neurons with suppressed NEEP21, AMPAR-mediated EPSCs are decreased, but this effect can be overridden with dynamin mutants to block endocytosis. Because LTP is blocked in the absence of NEEP21, this protein may also have a role in trafficking receptors from intracellular holding pools. The forward trafficking of GluR2 from the early endosome to the recycling endosome is mediated by the formation of a GluR2–GRIP1–NEEP21 complex. Although it is assumed that an as yet unidentified GluR1–NEEP21 complex interacting protein mediates the recycling of GluR1, the lack or low availability of such a protein may explain the retarded recycling of GluR1 to the cell surface.

Summary

The mechanisms of AMPAR trafficking, which in turn influence synaptic efficacy, are complex. However, overall patterns of AMPAR forward traffic, recruitment, and cycling have been established. The insertion of AMPARs at newly formed synapses and at synapses that have received a signal to undergo LTP depends on GluR1. Insertion of these GluR1-containing AMPARs is preceded and/or accompanied by an upregulation of slot proteins at the synapse that hold the AMPARs at the PSD until such time as they are replaced by constitutively cycling GluR2/3-containing AMPARs. During LTD, AMPARs are released from the synapse by alterations to their scaffold binding partners, which are then degraded. AMPARs are then endocytosed at sites separate from the PSD. The fate of the receptors endocytosed by LTD appears to differ. GluR2/3-containing receptors are degraded, whereas GluR1-containing receptors are slowly recycled to extrasynaptic sites. Although this model can explain many of the features of plasticity, clearly a great deal of work remains to be done to obtain a full mechanistic understanding of how AMPARs are regulated. Given the crucial role of these receptors in brain function and dysfunction, we believe that gaining this information is a goal of fundamental importance.

See also: AMPA Receptors: Molecular Biology and Pharmacology; AMPA Receptors: Disease; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; NMDA Receptors and Development; NMDA Receptors and Disease + C464; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation; Receptor Trafficking; Transporter Proteins in Neurons and Glia.

Further Reading

Ashby MC, De La Rue SA, Ralph GS, Uney J, Collingridge GL, and Henley JM (2004) Removal of AMPA receptors (AMPARs) from synapses is preceded by transient endocytosis of extrasynaptic AMPARs. *Journal of Neuroscience* 24: 5172–5176.

Ashby MC, Ibaraki K, and Henley JM (2004) It's green outside: Tracking cell surface proteins with pH-sensitive GFP. *Trends in Neuroscience* 27: 257–261.

Bredt DS and Nicoll RA (2003) AMPA receptor trafficking at excitatory synapses. *Neuron* 40: 361–379.

Choquet D and Triller A (2003) The role of receptor diffusion in the organization of the postsynaptic membrane. *Nature Reviews Neuroscience* 4: 251–265.

Collingridge GL, Isaac JT, and Wang YT (2004) Receptor trafficking and synaptic plasticity. *Nature Reviews Neuroscience* 5: 952–962.

Dingledine R, Borges K, Bowie D, and Traynelis SF (1999) The glutamate receptor ion channels. *Pharmacological Reviews* 51: 7–61.

Gerges NZ, Backos DS, Rupasinghe CN, Spaller MR, and Esteban JA (2006) Dual role of the exocyst in AMPA receptor targeting and insertion into the postsynaptic membrane. *EMBO Journal* 25: 1623–1634.

Horton AC and Ehlers MD (2003) Neuronal polarity and trafficking. *Neuron* 40: 277–295.

Lee SH, Simonetta A, and Sheng M (2004) Subunit rules governing the sorting of internalized AMPA receptors in hippocampal neurons. *Neuron* 43: 221–236.

Malenka RC and Bear MF (2004) LTP and LTD: An embarrassment of riches. *Neuron* 44: 5–21.

Palmer CL, Cotton L, and Henley JM (2005) The molecular pharmacology and cell biology of alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors. *Pharmacological Reviews* 57: 253–277.

Perestenko PV and Henley JM (2003) Characterisation of the intracellular transport of GluR1 and GluR2 α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor subunits in hippocampal neurons. *Journal of Biological Chemistry* 278: 43525–43532.

AMPA Receptors: Disease

R C Carroll and R S Zukin, Albert Einstein College of Medicine, Bronx, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Excessive glutamatergic signaling can result in neuronal death, or excitotoxicity, a mechanism thought to be involved in the pathology of a number of neurological diseases and disorders. Glutamate-induced cell death is often attributed to calcium-mediated toxicity. As *N*-methyl-D-aspartate (NMDA)-type glutamate receptors are highly permeable to calcium, much attention has been focused on their possible role in many forms of neurodegeneration. However, it has become increasingly evident that ion permeation through a subpopulation of non-NMDA-type ionotropic glutamate receptors, the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type receptors (AMPARs), is likely to be the underlying cause of cell injury in several neurological disorders, including seizure, epilepsy, and amyotrophic lateral sclerosis (ALS).

AMPARs mediate fast excitatory synaptic transmission in the central nervous system (CNS). AMPARs are formed as tetramers from a combination of four subunits (GluR1 through 4 or GluRA through D). The messenger RNAs (mRNAs) encoding AMPARs are subject to RNA editing and alternative splicing, thereby increasing the molecular diversity of the receptors. The GluR2 subunit has profound impact on the biophysical characteristics of the heteromeric AMPARs. The presence of an arginine in place of glutamine in the pore-lining M2 region of the GluR2 subunit, resulting from mRNA editing, largely prevents the permeation of Ca^{2+} and Zn^{2+} ions (Figure 1). In contrast, AMPARs lacking the GluR2 subunit are highly permeable to Ca^{2+} and Zn^{2+}, have distinctly fast kinetics, and exhibit a characteristic inwardly rectifying current–voltage (I-V) relation as a result of a voltage-dependent block by intracellular polyamines (Figure 1). The presence of GluR2 also influences the kinetics, conductance, assembly, and trafficking of AMPARs . Therefore, regulation of GluR2 expression can have a substantial impact on synaptic function and neuronal survival.

Whereas the majority of synapses in the CNS express calcium-impermeable, GluR2-containing AMPARs, synapses onto aspiny neurons, as well as glia, throughout the CNS express Ca^{2+}-permeable AMPARs. In these cells, AMPAR-mediated Ca^{2+} signaling is often rapid and can be locally restricted by fast Ca^{2+} extrusion pumps. The function of calcium flux through these AMPARs may be similar to that of NMDA receptors (NMDARs) in other synapses, including contribution to synaptic plasticity. For example, at synapses of dorsal horn neurons in the spinal cord, calcium influx through AMPARs triggers long-term potentiation (LTP) of synaptic efficacy that is associated with nociceptive plasticity and chronic pain.

In most principal neurons of the neocortex, hippocampus, amygdala, and cerebellum, GluR2-lacking, Ca^{2+}-permeable AMPARs are expressed at very low density. However, the subunit composition and Ca^{2+} permeability of AMPARs can be dynamically regulated in both a cell- and a synapse-specific manner during development and in response to neuronal activity. Changes in AMPAR subunit composition can arise as a consequence of the regulated redistribution or trafficking of AMPAR subunits as well as through alterations in protein expression, including activity-dependent local protein synthesis of AMPARs in dendrites. Alterations in AMPAR subunit composition resulting in increased Ca^{2+} permeability are also known to be induced by a number of neuronal insults and diseases including seizures, ischemia, spinal cord injury, drugs of abuse, and ALS. Investigation of the molecular mechanisms underlying activity-dependent remodeling of synaptic AMPAR subunit composition and permeability have begun to highlight the importance of Ca^{2+}-permeable AMPARs in neuronal death associated with neurological disease.

GluR2-Lacking AMPARs in Neuronal Death

Ischemia

Transient global or forebrain ischemia is the complete disruption of blood flow to the brain such as occurs in patients during cardiac arrest, open heart surgery, or profuse bleeding. Transient global ischemia can also be induced experimentally in animal models, providing a clinically relevant experimental model that is well suited to investigation of both the molecular and cellular mechanisms of this disorder and its consequences. A surprising consequence of global ischemic insult is that while all brain regions are equally subject to oxygen deprivation, neuronal death is quite selectively localized. Hippocampal CA1 pyramidal neurons are particularly vulnerable to delayed cell death and are thought to likely contribute to the observed cognitive deficits. For many years it was thought that excessive calcium influx through NMDARs was a critical player in ischemia-induced neuronal death. However, studies involving administration of NMDAR and AMPAR-specific antagonists

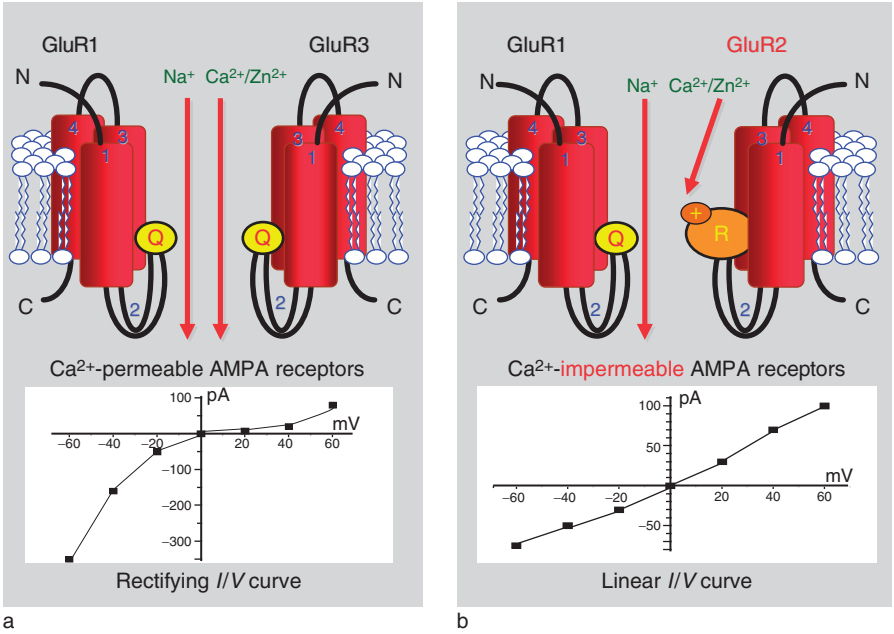

Figure 1 AMPA receptor subunit composition determines Ca^{2+} permeability. (a) GluR2-lacking AMPARs are highly permeable to Ca^{2+} and exhibit doubly rectifying current-voltage relationships. Rectification of GluR2-lacking AMPARs arises as a consequence of voltage-dependent channel block by intracellular polyamines. (b) GluR2-containing AMPARs are Ca^{2+}-impermeable and exhibit electrically linear current–voltage relationships. The presence of the GluR2 subunit in the heteromeric AMPAR channels limits Ca^{2+} and Zn^{2+} influx largely due to the presence of an arginine (R) in place of a glutamine (Q) residue at the Q/R RNA editing site. Reproduced from Liu SJ and Zukin RS (2007) Ca^{2+}-permeable AMPA receptors in synaptic plasticity and neuronal death. *Trends in Neurosciences* 30: 126–134, with permission from Elsevier.

suggest that AMPARs may play a greater role in the cell death associated with global ischemia. Consistent with this, global ischemia induces downregulation of AMPAR GluR2 expression in selectively vulnerable CA1 neurons and a switch in AMPAR phenotype from calcium impermeable to calcium permeable. It is now established that calcium influx through AMPARs is causally linked to the delayed death of hippocampal CA1 neurons.

While principal neurons of the hippocampus express GluR2-containing, Ca^{2+}-impermeable AMPARs under physiological conditions, ischemic insult reduces GluR2 mRNA and protein expression selectively in CA1 neurons. This change results in a long-lasting switch from GluR2-containing to GluR2-lacking AMPARs. By 42 h after ischemia, AMPAR excitatory postsynaptic currents (EPSCs) at CA1 synapses exhibit properties of GluR2-lacking AMPARs, including enhanced rectification, sensitivity to polyamines, and AMPAR-mediated Ca^{2+} influx. In addition to their role in mediating Ca^{2+} entry, GluR2-lacking AMPARs are thought to mediate a late rise in cytoplasmic Zn^{2+}, which reaches toxic levels by 48–72 h following transient global ischemia.

Regulated changes in synaptic AMPAR composition can occur through multiple mechanisms. Synaptic plasticity in cerebellar stellate cells results in rapid trafficking of AMPARs and selective incorporation of GluR2-containing AMPARs at synaptic sites, leading to a switch in the composition of synaptic AMPARs. In contrast, prolonged synaptic inactivity in hippocampal neurons alters AMPAR composition through increased local translation of GluR2-lacking AMPARs. Global ischemia induces a switch in AMPAR phenotype by regulation at the level of transcription. Ischemic insults induce expression of the transcriptional repressor (REST; Figure 2), which in turn silences expression of AMPAR GluR2 in selectively vulnerable CA1 neurons. REST binds to the promoter region of target genes, such as GluR2, and suppresses their expression through chromatin remodeling and epigenetic modifications (or marks). Acute knockdown of REST prevents the suppression of GluR2 and subsequent death of CA1 neurons following ischemia. In addition to changes in GluR2 expression, receptor trafficking and GluR2 RNA editing can be modified in response to neuronal insults. Ischemic insult enhances the clathrin-dependent internalization of GluR2-containing AMPARs and synaptic targeting of GluR2-lacking AMPARs to synapses of insulted hippocampal neurons via exocytosis, leading to a switch in the subunit composition of surface-expressed AMPARs (Figure 3). Global ischemia also inhibits activity of the RNA editing enzyme adenosine deaminase acting on RNA (ADAR)-2 and disrupts glutamine/arginine (Q/R) editing of GluR2,

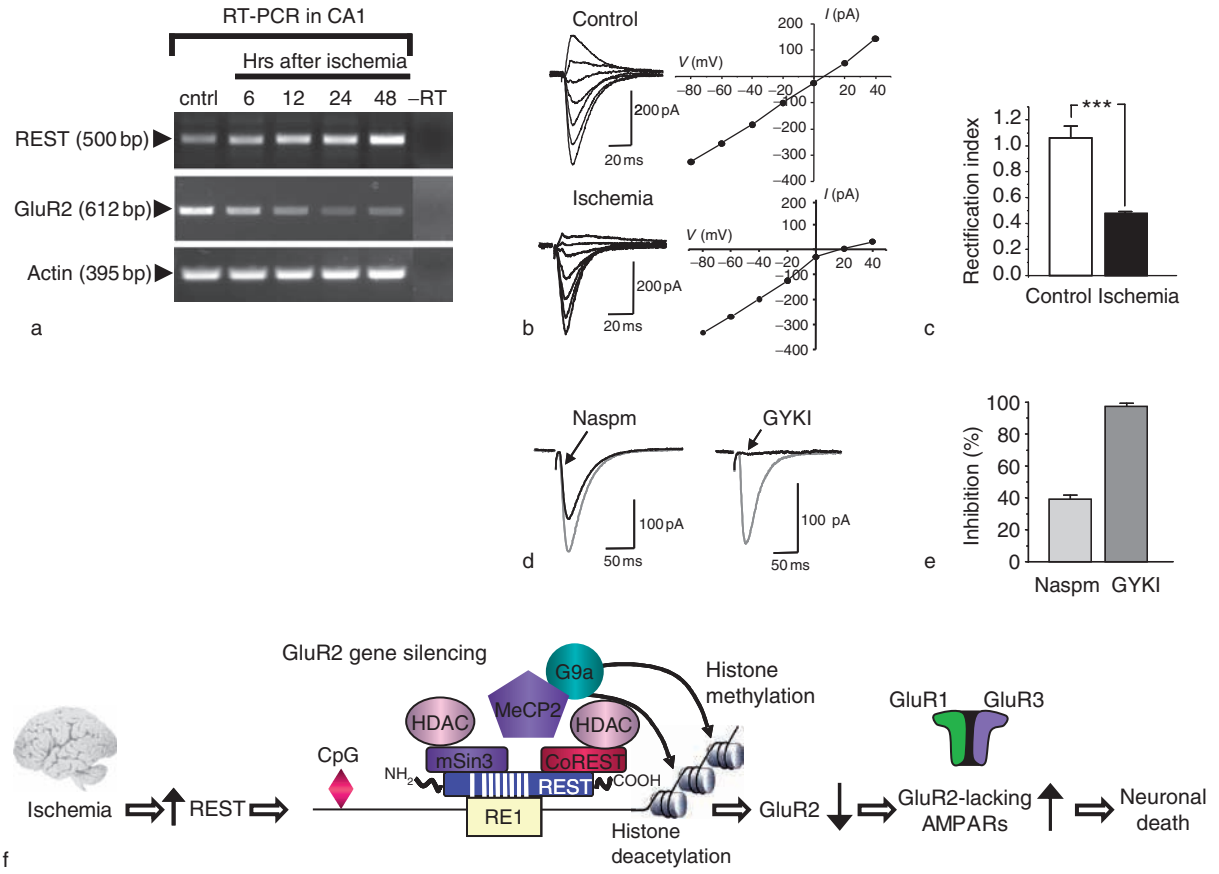

Figure 2 Ischemia silences glutamate receptor 2 (GluR2) expression through the activation of the transcriptional repressor (REST) in CA1. (a) Reverse transcription–polymerase chain reaction (RT-PCR) products amplified from REST, GluR2 and actin RNAs following global ischemia demonstrate the enhancement of REST and the suppression of GluR2 RNA levels. (b, c) Alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) at Schaffer collateral CA1 synapses in the postischemic hippocampus exhibit properties of Ca^{2+}/Zn^{2+}-permeable, GluR2-lacking AMPARs. (b) AMPA-excitatory postsynaptic currents (EPSCs) recorded from control slices exhibit a linear $I-V$ relation whereas AMPA-EPSCs recorded from slices of animals at 42 h after ischemia exhibit inwardly rectifying currents. (c) The rectification index (the ratio of EPSC amplitude at +40 mV to that at 60 mV × 1.5), a measure of the extent of current/voltage relationship rectification, is reduced in ischemic versus control slices (***$p < 0.001$). (d) AMPA-EPSCs in ischemic slices are mediated in part by GluR2-lacking AMPARs as they are inhibited by Naspm, a specific antagonist of these receptors. The currents are completely blocked by the AMPA antagonist GYKI-53655 (right). (e) Summary of data in (d). (f) Scheme representing REST-dependent silencing of GluR2 gene expression in ischemic neurons. Ischemia triggers activation of the repressor REST. REST binds the RE1 element in the GluR2 promoter and silences GluR2 transcription. REST recruits co-repressors mSin3A and CoREST and histone deacetylases (HDACs), which silence gene expression via chromatin remodeling. CoREST recruits methyl-CpG binding protein 2 (MeCP2) and the histone methyltransferase G9a, which promotes DNA and histone methylation. Silencing of GluR2 expression reduces GluR2 subunit number, leading to assembly and insertion of functional, GluR2-lacking AMPARs at CA1 synapses of postischemic hippocampus. Error bars represent SEMs. Reproduced from Liu SJ and Zukin RS (2007) Ca^{2+}-permeable AMPA receptors in synaptic plasticity and neuronal death. *Trends in Neurosciences* 30: 126–134, with permission from Elsevier.

increasing calcium and zinc permeability. Direct delivery of ADAR2 restores Q/R editing and protects vulnerable neurons from cell death. Thus, a variety of processes, all of which enhance the expression of GluR2-lacking AMPARs, contribute to the susceptibility of select neurons to cell death in brain ischemia.

While it is well established that the subunit composition of AMPARs changes with transient global ischemia, understanding of the mechanisms leading to cell death has been more difficult to ascertain. Several lines of evidence link the increase in levels of Ca^{2+}-permeable, GluR2-lacking AMPARs following global ischemia directly to the death of CA1 hippocampal neurons. Reduction in the expression of GluR2 subunits results in enhanced death of pyramidal neurons even under physiological signaling conditions. Overexpression of Ca^{2+}-permeable GluR2(Q) channels or calcium-impermeable AMPARs exacerbates or protects against cell death of CA1 neurons following ischemic insult, respectively. Finally, application of two selective blockers of GluR2-lacking AMPARs, naphthylspermine and philanthotoxin, is

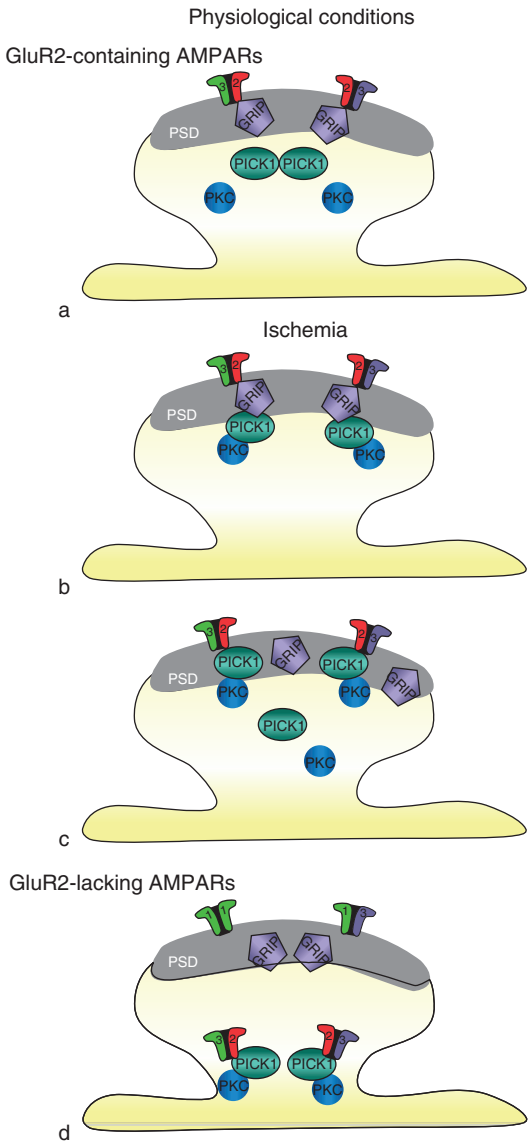

Figure 3 Ischemia-induced switch in AMPAR subtypes at CA1 synapses. In the basal state, GluR2-containing AMPARs are stabilized at the postsynaptic membrane via association with the GluR2/3-interacting AMPA receptor binding protein (ABP). (a) Ischemic insults activate protein kinase C (PKC) and promote binding of PKC to protein interacting with C kinase1 (PICK1), which in turn targets the PICK1-PKC complex to the ABP-GluR2 complex (b). (c) GluR2 binding to ABP is disrupted and its association with is enhanced. Interaction of GluR2-containing AMPARs with PICK1 promotes receptor internalization. (d) Ischemia promotes synaptic incorporation of GluR2-lacking AMPARs, which replace GluR2-containing AMPARs at insulted CA1 synapses. GRIP, GluR2-interacting protein. Reproduced from Liu SJ and Zukin RS (2007) Ca^{2+}-permeable AMPA receptors in synaptic plasticity and neuronal death. *Trends in Neurosciences* 30: 126–134, with permission from Elsevier.

neuroprotective in animal models of global ischemia. While it is likely that the increased calcium permeability into cells expressing GluR2-lacking AMPARs contributes to excitotoxic death, elevation of cytosolic Zn^{2+} via influx through these same receptors now appears to be an important trigger for cell death in these neurons (see the section titled 'Zinc').

Epilepsy

In addition to their role in global ischemia, Ca^{2+}-permeable AMPARs are implicated in the cell death following seizures. Following status epilepticus in adult rats, GluR2 mRNA and subunit expression is downregulated in vulnerable pyramidal neurons found in the CA3 and CA1 region of the hippocampus prior to the onset of their death. Status epilepticus induced in rat pups, however, which does not trigger neurodegeneration, does not cause downregulation of GluR2 expression, providing a further link between these events. Seizures also upregulate REST mRNA and suppress GluR2 promoter activity in hippocampal neurons. As has been observed in global ischemia, REST-dependent silencing of the AMPAR GluR2 gene is thought to contribute to the delayed neurodegeneration associated with seizures. As such, regulated expression of calcium-permeable AMPARs may represent a broad mechanism of insult-induced neuronal death.

ALS

Significant evidence also suggests a role for GluR2-lacking AMPARs in ALS, sometimes called Lou Gehrig's disease. ALS is a rapidly progressive, invariably fatal neurological disease that attacks the neurons responsible for controlling voluntary muscles. In ALS, both the upper motor neurons and the lower motor neurons degenerate or die, and muscles gradually weaken and waste away. Eventually, the ability of the brain to start and control voluntary movement is lost. Individuals with ALS lose their strength and the ability to move their arms, legs, and body and ultimately the ability to breathe. Approximately 5–10% of ALS cases are familial, and of these, one-fifth have been found to carry gain of function mutations in the superoxide dismutase-1 (SOD1) gene.

The selective loss of motor neurons in ALS has been linked to the elevated susceptibility of these neurons to excitotoxic damage. In particular, motor neurons *in vivo* and *in vitro* are subject to injury in response to exposure to glutamatergic agonists. Several factors may contribute to the vulnerability of these neurons. In particular, while these neurons express high levels of GluR2-lacking AMPARs, unlike many other neurons that express the Ca^{2+}-permeable receptors, they have low levels of Ca^{2+}-buffering proteins. The possibility that these receptors play an important role in excitotoxicity in motor neurons is supported by the fact that AMPA/kainate receptor agonists, but not NMDA receptor agonists, were highly effective at inducing cell death.

There are likely many factors that contribute to the pathology of ALS. GluR2-lacking AMPARs appear to play an important role. Some patients with sporadic ALS exhibit a reduced level of editing of the Q/R site of the GluR2 subunit, which controls AMPAR calcium permeability. This mutation is predicted to even further increase the level of calcium permeability through AMPARs in motor neurons of these patients. This deficiency seems to be restricted to the sporadic form of the disease, as GluR2 mRNA is fully edited in SOD1 transgenic rat models of familial ALS and in humans with spinal and bulbar muscular atrophy. However, evidence of a role for Ca^{2+}-permeable AMPARs in familial, SOD1-related forms of ALS has come from transgenic mouse models of the disease. Animals which do not express the GluR2 subunit do not show evidence of motor neuron disease, suggesting the absolute level of GluR2 is not important. However, if these animals are crossed with transgenic animals expressing the G93A gain of function SOD1 mutation, the motor neuron degeneration normally observed in the SOD1 animals is greatly accelerated. Similarly, crossing mice that have a mutation in GluR2 that prevents its editing with mice that have the SOD1 mutation aggravates the course of the motor neuron disease. Overexpression of GluR2, in contrast, delays the onset of motor neuron death in SOD1 mutant mice.

Increased levels of Ca^{2+}-permeable AMPARs may contribute to one apparently SOD1-independent form of juvenile onset ALS, termed ALS2. This autosomal recessive form of ALS is linked to a deficiency in the ALS2 gene that encodes the alsin protein. Alsin binds to a GluR2-interacting protein (GRIP1). In neurons from ALS2-deficient animals, GRIP1 localization at synapses is reduced. GRIP1 is believed to participate in the targeting and membrane stabilization of GluR2 subunits. Consistent with the mislocalization of GRIP1, GluR2-lacking AMPARs at synapses are increased in ALS2 mutant mice, resulting in an increased vulnerability of motor neurons to glutamate receptor-mediated excitotoxicity.

The exact mechanisms by which expression of Ca^{2+}-permeable AMPARs contributes to excitotoxic motor neuron injury are, as yet, unclear. However, a combination of factors likely makes the motor neurons susceptible to death in the course of the disease. Elevated glutamate levels, perhaps related to decreased clearance of glutamate by transporters, elevates intracellular calcium levels via influx through the high levels of calcium-permeable AMPARs in motor neurons. This may be exacerbated by mutations in the editing of GluR2. Because of the poor calcium buffering of motor neurons, calcium uptake into mitochondria can be elevated, resulting in the generation of reactive oxygen species. This again may be aggravated by mutations in SOD1 which can diminish mitochondrial function. The resulting oxidative stress may enhance the degeneration of the motor neurons characteristic of the progression of the disease (Figure 4).

Neuronal Injury

Tumor necrosis factor α (TNFα) is a proinflammatory cytokine reported to have both neuroprotective and neurotoxic functions. It is released following a number of neuronal insults, including trauma, ischemia, and multiple sclerosis. TNFα is released by astrocytes and microglia, and its receptors, TNF-R1 and 2, can be found on both neurons and glia in the CNS. TNFα, at high concentrations, has been found to facilitate glutamate-mediated neurotoxicity. Several mechanisms may account for this effect. TNFα has been found to inhibit glutamate uptake in astrocytes, leading to possible excessive glutamatergic activation. Additionally, TNFα directly enhances synaptic transmission mediated by glutamate receptors. For example, TNFα activation of TNF-R1 elevates the expression of synaptic AMPARs in hippocampal cultures and slices. Notably, this enhancement involves the preferential insertion of GluR2-lacking AMPARs, thereby enhancing Ca^{2+} permeability of synapses in hippocampal neurons (Figure 5). Such a rapid change in the calcium influx in neurons which normally express few calcium-permeable AMPARs would be predicted to contribute to enhanced glutamate pathogenicity and subsequent cell death, particularly when paired with a reduction in glutamate uptake. Future studies will be necessary to establish a direct role of these newly inserted AMPARs in the neurodegeneration associated with insults leading to the release of TNFα.

Mechanisms of Neuronal Death via GluR2-Lacking AMPARs

In all the cases of excitotoxic neurodegeneration discussed here, the exact mechanisms linking the expression or modification of Ca^{2+}-permeable, GluR2-lacking AMPARs to cell death are not completely understood. Elevated Ca^{2+} influx and increased intracellular Zn^{2+} levels are the most likely mediators of cell death. While not all the relevant downstream targets are known, a great deal is understood about the possible mechanisms by which these ions may contribute to toxicity in neurons.

Calcium

Neuronal homeostasis requires that the intracellular concentration of Ca^{2+} be maintained in the

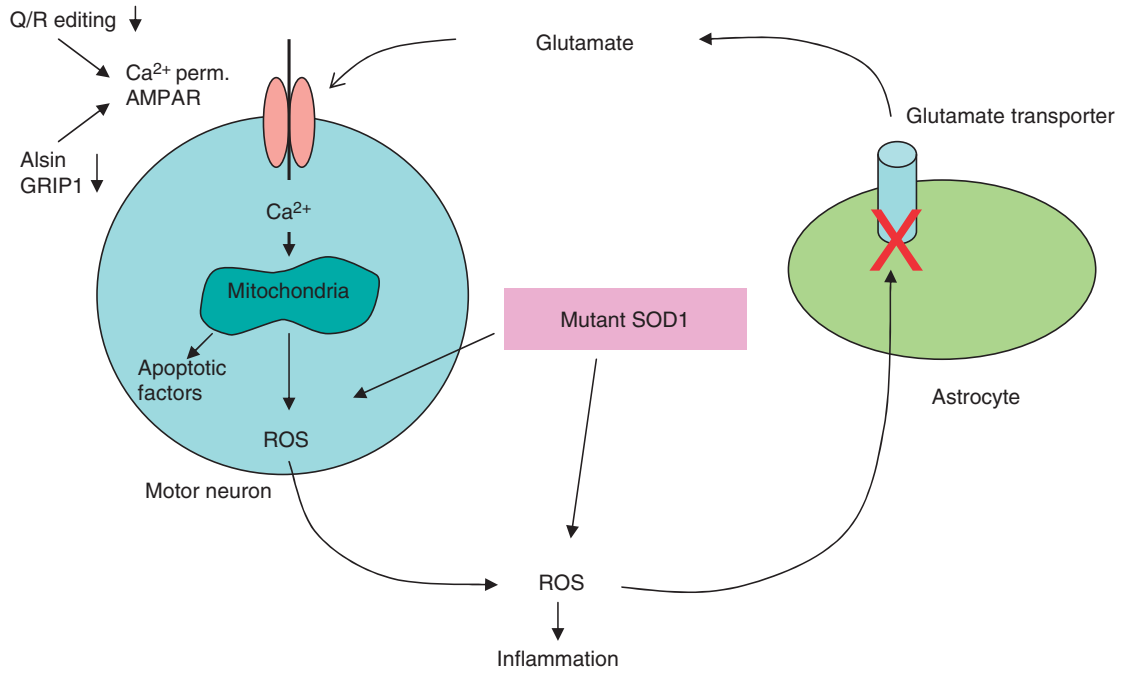

Figure 4 Potential mechanism of selective motor neuron death in amyotrophic lateral sclerosis (ALS). Elevated levels of extracellular glutamate due to reduced glutamate uptake or trauma result in the influx of nonphysiological levels of Ca^{2+} influx through Ca^{2+}-permeable (Ca^{2+} Perm.) α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) channels. The extent of influx may be exacerbated by decreased GluR2 editing in sporadic cases of ALS. Uptake of calcium by mitochondria results in reactive oxygen species (ROS) generation, and possibly apoptosis. The ROS can inhibit astrocytic glutamate transporters, causing additional increases in extracellular glutamate. In many familial cases of ALS, mutations in superoxide dismutase-1 (SOD1) are linked to cell death, perhaps in part through dysregulation of mitochondrial function. AMPAR, AMPA receptor; GRIP, GluR2-interacting protein; Q/R, glutamine/arginine. Reproduced from Kwak S and Weiss JH (2006) Calcium-permeable AMPA channels in neurodegenerative disease and ischemia. *Current Opinion in Neurobiology* 16: 281–287, with permission from Elsevier.

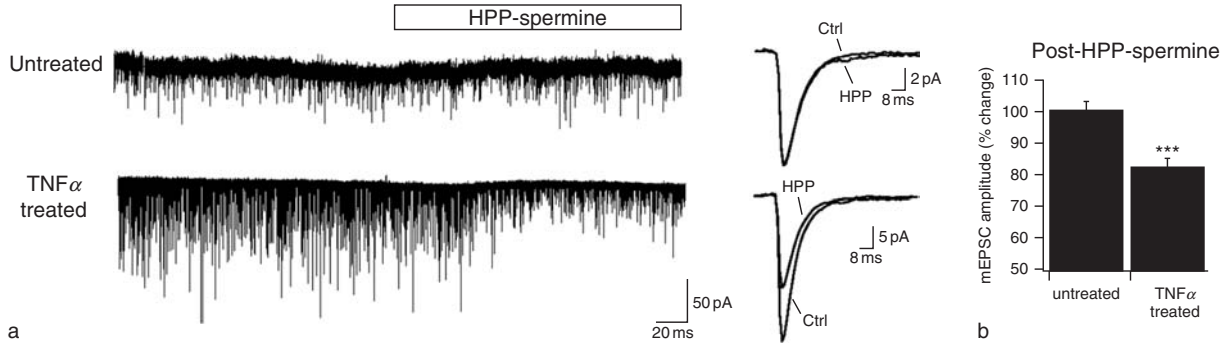

Figure 5 Tumor necrosis factor (TNF) α induces an increase in calcium-permeable α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) in hippocampal neurons. (a) Miniature excitatory postsynaptic currents (mEPSCs) from TNFα-treated cultured hippocampal neurons show enhanced blockade by the calcium-permeable AMPAR blocker *N*-(4-hydroxyphenylpropanoyl) (HPP)–spermine. (b) Averaged normalized mEPSC amplitudes after HPP-spermine for control and TNFα-treated cells demonstrate that the cytokine significantly enhances the expression of GluR2-lacking AMPARs. Error bar represents SEM. Reproduced from Stellwagen D, Beattie EC, Seo JY, and Malenka RC (2005) Differential regulation of AMPA receptor and GABA receptor trafficking by tumor necrosis factor-alpha. *Journal of Neuroscience* 25: 3219–3228, with permission from the Society for Neuroscience.

tens-of-nanomolar range. Unregulated or excessive influx of Ca^{2+} that can result from increases in Ca^{2+}-permeable AMPARs following acute insult or disease could initiate a series of damaging events in the cytoplasm and nucleus. Elevated cytosolic Ca^{2+} activates Ca^{2+}–adenosine triphosphatase, which depletes the energy stores of the cell. High Ca^{2+} levels also uncouple mitochondrial oxidative phosphorylation, causing acute swelling of dendrites and cell bodies and cell death. Ca^{2+} uptake by mitochondria can generate reactive oxygen species and can open the transition pore allowing the release of mediators of apoptosis, including cytochrome C. High cytosolic Ca^{2+} also causes excessive activation of nitric oxide

synthase, which promotes generation of free radicals that destroy cell membranes, damage DNA, and can induce apoptosis. In addition to all these effects, Ca^{2+} activates a host of Ca^{2+}-sensitive transcription factors, phospholipases, endonucleases, and proteases, all of which may trigger specific events associated with necrosis or apoptosis.

Zinc

The transition metal Zn^{2+}, like Ca^{2+}, is an important neuronal signaling molecule under physiological conditions and a critical player in the cell death associated with ischemia and seizures. Zn^{2+} is co-localized with glutamate in presynaptic vesicles at a subset of excitatory synapses and is particularly high in mossy fiber tracts of the hippocampus. Zn^{2+} is co-released with glutamate spontaneously and in an activity-dependent manner. Under physiological conditions, Zn^{2+} release modulates the function of multiple neurotransmitter receptors, including NMDARs, AMPARs and γ-aminobutyric acid $(GABA)_A$ receptors. In the neuronal cytoplasm, Zn^{2+} acts as a functionally important component of metalloenzymes and zinc finger-containing transcription factors. Zn^{2+} can enter neurons via voltage-sensitive Ca^{2+} channels, NMDARs, GluR2-lacking AMPARs, and the Na^+/Zn^{2+} antiporter. Of these, GluR2-lacking AMPARs show the highest permeability to Zn^{2+}. Therefore, upregulation of the expression of these receptors in neurons can greatly enhance the levels of cytosolic Zn^{2+}.

Whereas relatively low concentrations of Zn^{2+} elicit neuronal death with the hallmarks of apoptosis, higher concentrations of Zn^{2+} elicit neuronal death characteristic of necrosis. Zn^{2+} induces neurotoxicity by a number of mechanisms similar to calcium, including production of free radicals, loss of mitochondrial membrane potential, formation of the mitochondrial transition pore, dysregulation of the electron transport chain, and disruption of glycolysis and energy production. It is interesting that Zn^{2+} has been found to be more potent than calcium at disrupting the function of mitochondria.

Studies indicate that Zn^{2+} at high concentrations is a critical mediator of the neuronal death associated with global ischemia, prolonged seizures, traumatic brain injury, and other brain disorders. In the first 1–2 h after global ischemia, Zn^{2+} acts at the level of mitochondria to promote the release of cytochrome c, apoptosis-inducing factor, and second mitochondria-derived activator of caspase/direct inhibitor of apoptosis-binding protein with low isoelectric point from the mitochondria into the cytosol. At delayed late time points following global ischemia or status epilepticus, Zn^{2+} accumulates in selectively vulnerable hippocampal neurons. Administration of Zn^{2+} chelators such as EDTA substantially reduces the rise in intracellular Zn^{2+} in vulnerable neurons and affords robust neuroprotection following both ischemia and epilepsy (Figure 6). These observations implicate Zn^{2+} as a critical mediator of the neuronal death associated with neural insults.

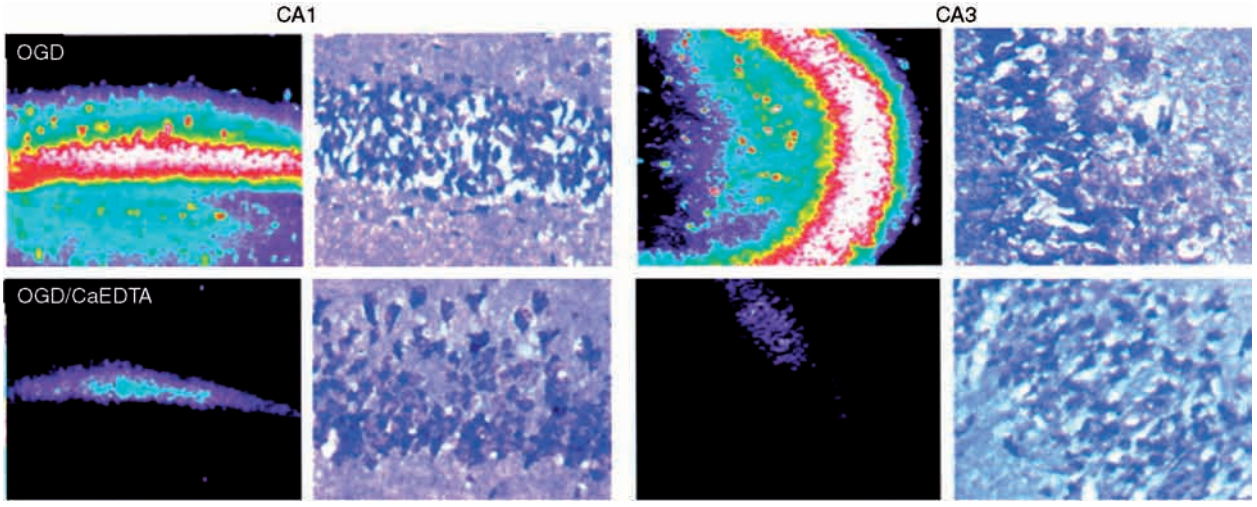

Figure 6 Chelation of zinc is neuroprotective in hippocampal pyramidal neurons following ischemic insult. Slices of adult mouse hippocampus were exposed for 5 min to oxygen glucose deprivation in the absence (OGD), or presence of the extracellular Zn^{2+} chelator Ca^{2+} EDTA (OGD + CaEDTA). Four hours after OGD, propidium iodide (PI) labeling was performed on of hippocampal CA1 (left four panels) and CA3 (right four panels) pyramidal cell layers to measure neuronal degeneration. Images in the left panel for each hippocampal region show pseudocolor images of PI labeling. Right panels show matched slices (25 mm) stained with toluidine blue. Zinc chelation significantly reduced the effects of ischemic injury on neurons. Reproduced from Yin HZ, Sensi SL, Ogoshi F, and Weiss JH (2002) Blockade of Ca^{2+}-permeable AMPA/kainate channels decreases oxygen-glucose deprivation-induced Zn^{2+} accumulation and neuronal loss in hippocampal pyramidal neurons. *Journal of Neuroscience* 22: 1273–1279, with permission from the Society for Neuroscience.

Summary

Rapidly increasing evidence supports the importance of Ca^{2+}-permeable AMPARs in neuronal death. Multiple mechanisms have been identified by which AMPAR subunit composition and consequently their Ca^{2+} permeability is modified in response to neuronal insults and disease. With ischemia and epilepsy, reductions in the protein expression of GluR2 subunits enhance the calcium and zinc permeability in select populations of neurons, leading eventually to cell death. Regulation of the transcription of the GluR2 gene by modulation of REST activity is emerging as an important means by which neural insults can affect the composition of synaptic AMPARs. In motor neuron disease such as ALS, defects in the editing of the GluR2 subunit or the normally elevated levels of Ca^{2+}-permeable AMPARs in motor neurons are linked to cell death. In each of these cases, the exact mechanisms underlying the increased levels of Ca^{2+}-permeable AMPARs are still unknown. However, the influx of calcium and zinc appear to mediate cell death through a variety of mechanisms, including disruption of mitochondrial function and integrity.

Ongoing studies will continue to define both how neuronal insult and disease couple to changes in AMPAR composition and how these changes in AMPARs lead to subsequent cell death. Because calcium-permeable AMPARs exhibit spatially restricted patterns of expression and are normally expressed at very low density by principal neurons of the hippocampus and neocortex, they may be a promising target for possible therapeutic approaches relevant to ischemia, epilepsy, and ALS. Already the ability of specific AMPAR antagonists and Zn^{2+} chelators to reduce neuronal death in models of neural insult provides an exciting glimpse into the possible development of neuroprotective treatments that could lie ahead.

See also: AMPA Receptors: Molecular Biology and Pharmacology; AMPA Receptor Cell Biology/Trafficking; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Apoptosis in Nervous System Injury; Apoptosis in Neurodegenerative Disease; Glutamate Regulation of Dendritic Spine Form and Function; Intracellular Calcium and Neuronal Death; NMDA Receptors and Disease + C464.

Further Reading

Burnashev N, Monyer H, Seeburg PH, and Sakmann B (1992) Divalent ion permeability of AMPA receptor channels is dominated by the edited form of a single subunit. *Neuron* 8: 189–198.

Dingledine R, Borges K, Bowie D, and Traynelis SF (1999) The glutamate receptor ion channels. *Pharmacological Reviews* 51: 7–61.

Gorter JA, Petrozzino JJ, Aronica EM, et al. (1997) Global ischemia induces downregulation of GluR2 mRNA and increases AMPA receptor-mediated Ca^{2+} influx in hippocampal CA1 neurons of gerbil. *Journal of Neuroscience* 17: 6179–6188.

Kawahara Y, Ito K, Sun H, Aizawa H, Kanazawa I, and Kwak S (2004) Glutamate receptors: RNA editing and death of motor neurons. *Nature* 427: 801.

Kuner R, Groom AJ, Bresink I, et al. (2005) Late-onset motoneuron disease caused by a functionally modified AMPA receptor subunit. *Proceedings of the National Academy of Sciences of the United States of America* 102: 5826–5831.

Kwak S and Weiss JH (2006) Calcium-permeable AMPA channels in neurodegenerative disease and ischemia. *Current Opinion in Neurobiology* 16: 281–287.

Liu B, Liao M, Mielke JG, et al. (2006) Ischemic insults direct glutamate receptor subunit 2-lacking AMPA receptors to synaptic sites. *Journal of Neuroscience* 26: 5309–5319.

Liu SJ and Zukin RS (2007) Ca^{2+}-permeable AMPA receptors in synaptic plasticity and neuronal death. *Trends in Neurosciences* 30: 126–134.

Lo EH, Dalkara T, and Moskowitz MA (2003) Mechanisms, challenges and opportunities in stroke. *Nature Reviews Neuroscience* 4: 399–415.

Pellegrini-Giampietro DE, Zukin RS, Bennett MV, Cho S, and Pulsinelli WA (1992) Switch in glutamate receptor subunit gene expression in CA1 subfield of hippocampus following global ischemia in rats. *Proceedings of the National Academy of Sciences of the United States of America* 89: 10499–10503.

Porter BE, Cui XN, and Brooks-Kayal AR (2006) Status epilepticus differentially alters AMPA and kainate receptor subunit expression in mature and immature dentate granule neurons. *European Journal of Neuroscience* 23: 2857–2863.

Seeburg PH and Hartner J (2003) Regulation of ion channel/neurotransmitter receptor function by RNA editing. *Current Opinion in Neurobiology* 13: 279–283.

Stellwagen D, Beattie EC, Seo JY, and Malenka RC (2005) Differential regulation of AMPA receptor and GABA receptor trafficking by tumor necrosis factor-alpha. *Journal of Neuroscience* 25: 3219–3228.

Van Den Bosch L, Van Damme P, Bogaert E, and Robberecht W (2006) The role of excitotoxicity in the pathogenesis of amyotrophic lateral sclerosis. *Biochimica et Biophysica Acta* 1762: 1068–1082.

Yin HZ, Sensi SL, Ogoshi F, and Weiss JH (2002) Blockade of Ca^{2+}-permeable AMPA/kainate channels decreases oxygen-glucose deprivation-induced Zn^{2+} accumulation and neuronal loss in hippocampal pyramidal neurons. *Journal of Neuroscience* 22: 1273–1279.

AMPA Receptors: Molecular Biology and Pharmacology

S M Dravid, H Yuan, and S F Traynelis, Emory University, Atlanta, GA, USA

Published by Elsevier Ltd.

Introduction

Glutamate, one of the fundamental amino acid building blocks of proteins, is also a major excitatory neurotransmitter in the central nervous system (CNS). Neurons synthesize and package glutamate into presynaptic vesicles for release into the postsynaptic cleft. Synaptically released glutamate that diffuses across the 30 nm distance encounters a series of transmembrane postsynaptic proteins that comprise the glutamate receptor family. One class of glutamate receptors, metabotropic glutamate receptors, comprises transmembrane proteins that have extracellular clamshell-like domains that bind glutamate. When activated by glutamate binding, these G-protein-coupled receptors shift intracellular concentrations of signaling molecules to control a diverse set of cell properties. A second class of glutamate receptors, ionotropic glutamate receptors, comprises transmembrane proteins that contain an ion conduction path through the plasma membrane, as well as an array of clamshell-like extracellular ligand-binding domains, some of which bind to glutamate. Mammalian ionotropic glutamate receptors are ligand-gated ion channels encoded by 18 genes, and are subdivided into four major families on the basis of agonist pharmacology and sequence homology. These four receptor classes are known as amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), kainate, *N*-methyl-D-aspartate (NMDA), and δ receptors. This article focuses on the structure and function of AMPA receptors.

The AMPA receptor family is composed of four genes encoding the GluR1–GluR4 subunits (sometimes called GluRA–GluRD). In humans the chromosomal location of GluR1, GluR2, GluR3, and GluR4 encoding genes is 5q33, 4q32–33, Xq25–26, and 11q22–23, respectively. AMPA receptors were the first class of glutamate receptor cloned by screening a rat brain cDNA library for expression of kainate-activated ion channels in *Xenopus laevis* oocytes. After initial identification of GluR1, GluR2–GluR4 were rapidly identified by homology screening. There is about 70% sequence homology among different AMPA receptor subunits. A great deal of information now exists about AMPA receptor structure and function, and it could be argued that more is known about the structure of AMPA receptors than any other class of glutamate receptor.

Expression of AMPA Receptors

AMPA receptors are abundant and widely distributed in the central nervous system. Hippocampus, outer layer of cortex, basal ganglia, olfactory regions, lateral septum, and amygdala of the CNS are all enriched with GluR1, GluR2, and GluR3 subunits. In contrast, GluR4 expression is lower in many regions of the CNS except cerebellum, thalamus, and brain stem, where the expression is high. Immunoprecipitation studies have shown that the pyramidal cells of the hippocampus expressed AMPA receptors composed of GluR2 receptor in complex with either GluR1 or GluR3 subunits. GluR1 homomeric receptors, which have unique ion permeation properties, are thought to be expressed in select neuronal populations.

The expression of AMPA receptors is developmentally regulated. The GluR2 subunit appears as early as embryonic day 16 in rats whereas other receptors are upregulated later during the development. The GluR2 subunit can also be selectively altered during synaptic plasticity as well as during CNS injury, such as global ischemia. These changes in receptor subunit composition are known to change functional receptor properties.

AMPA receptors are present both postsynaptically and presynaptically. AMPA receptors are present on the synaptic membrane; however, 60–70% of AMPA receptors are present intracellularly. Glial cells also express AMPA receptors, which appear to be involved in glutamate-induced cell death. Activation of glial AMPA receptors also leads to release of ATP or nitric oxide.

Topology and Assembly of AMPA Receptor Subunits

All of the ionotropic glutamate receptors share a common topology, which consists of an extracellular N-terminal domain, a ligand-binding domain, three transmembrane domains (M1, M3, and M4), a cytoplasm-facing reentrant membrane loop (M2), and an intracellular C-terminal domain (Figure 1(a)). AMPA receptors are composed of approximately ~900 amino acids and have a molecular mass of ~105 kDa. The location of the N-terminus and C-terminus was first deduced by use of specific antibodies. Because the N- and C-terminal regions were located on the opposite ends of the polypeptide chain, it was proposed that AMPA receptors had an odd number of membrane-spanning domains. Further studies delineated the membrane topology, and showed that the M2 segment is a reentrant loop.

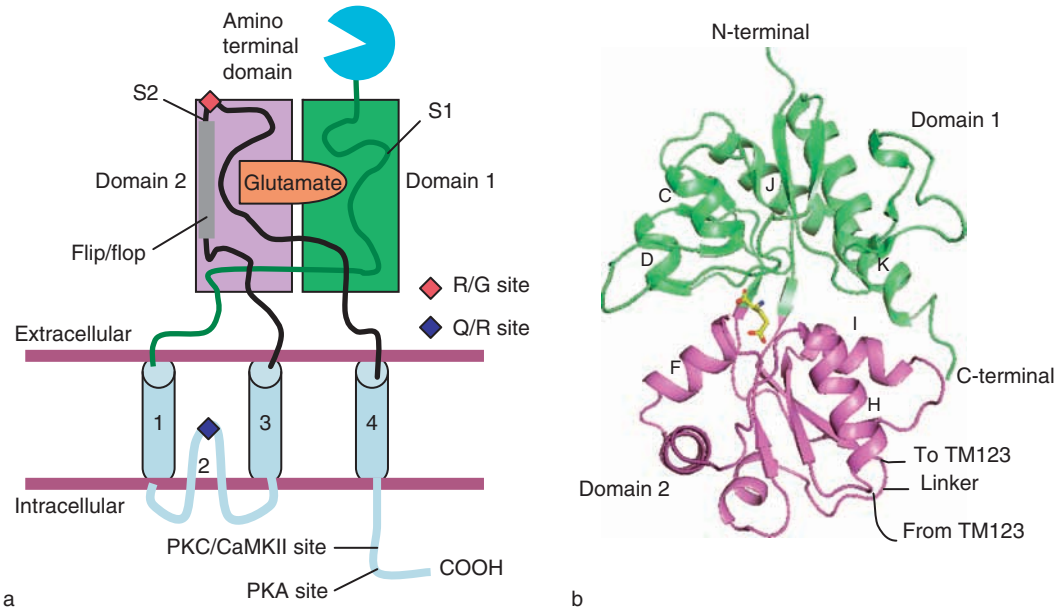

Figure 1 Structure of the AMPA receptor subunit. (a) Transmembrane topology of the AMPA receptor, indicating flip/flop site, Q/R site, R/G site, and phosphorylation sites (PKA, protein kinase A; PKC, protein kinase C; CaMKII, Ca^{2+}/calmodulin-dependent protein kinase II). (b) Crystal structure of the ligand-binding domain of the GluR2 subunit, with glutamate bound in the cleft formed by clamshells (Protein Data Bank code 1FTJ). The lime-colored structure is domain 1 and the violet-colored structure is domain 2 (TM, transmembrane domain).

This transmembrane topology shares a parallel organization to potassium channels, which are now considered a model for AMPA receptor membrane domain structure. The pore diameter of AMPA receptors is about 0.8 nm and, in contrast to potassium channels, permits the entry of Na^+ and, for some subunit combinations, Ca^{2+}.

The N-terminal domain in AMPA represents up to 45% of the mature polypeptide, but its function is poorly understood. Hypothesized functions of this domain include receptor assembly, allosteric modulation of the ion channel (similar to NMDA receptors), and binding of a second ligand. Contradictory to other subunits, the GluR4 subunit can form normally functioning homomeric channels even in the absence of the N-terminal domain.

The semiautonomous ligand-binding domain has been studied in detail using X-ray crystallography. The GluR2 agonist-binding domain is composed of two discontinuous peptide segments of approximately 150 amino acid residues. The first segment (S1) is adjacent and N-terminal to the M1 domain, whereas the second segment (S2) is located between M3 and M4 domains (Figure 1(a)). The agonist-binding domains of the AMPA receptors share similarities in sequence and structural arrangement with the ligand-binding site of several bacterial periplasmic amino acid-binding proteins.

Like potassium channels, AMPA receptors assemble as tetramers. Studies in recombinant and native receptors suggest that AMPA receptors assemble as dimer-of-dimers in a two-step manner. First, the monomers interact through the N-terminal domain to form dimers. Next, the dimers combine via the membrane domains to form the tetramer.

AMPA Receptor Function

A remarkable step forward in understanding the AMPA receptor structure and function occurred in 1995 when a water-soluble mini-receptor that included only the agonist-binding core was described. This was a fusion protein consisting of the S1 and S2 domains of GluR4 joined together via a short hydrophilic linker peptide that replaces the membrane-spanning regions. The engineered agonist-binding domain functionally reproduced the AMPA-binding properties of the GluR4 receptor. This concept paved the way for the subsequent production of soluble agonist-binding core and later generation of crystals for X-ray diffraction, which ultimately could be produced by careful refolding of the denatured ligand-binding core. The first crystal structure of the GluR2 ligand-binding domain complexed with kainate revealed a bilobed clamshell-like shape, with agonist bound deep in the cleft formed by the two lobes. Subsequent descriptions of crystal structures of the nonliganded form (apo state) as well as forms complexed to a variety of ligands were obtained. These studies revealed that the clamshell was geometrically opened widest in the apo state. In the glutamate-bound state, the clamshell was 21° more closed than in the apo

state (Figures 1 and 2). Competitive antagonists of the receptor stabilized the open-cleft state.

Glutamate makes a number of hydrogen-bonded contacts within the binding pocket with both upper (D1) and lower (D2) domains. The upper domain of the clamshell-like structure is formed by segment S1 and the C-terminal portion of segment S2. The N-terminus of S2 forms the lower D2 domain. It has been proposed that when glutamate first encounters the ligand-binding pocket, it initially docks or interacts with the D1 domain, which then promotes the rotation of the D2 domain toward D1 and induces closure of the clamshell. The closed conformation is stabilized by glutamate, itself forming a cross-domain bridge, as well as a number of other hydrogen bonds that form between the domains during glutamate binding. Water appears to persist inside of the agonist binding pocket, particularly for some agonists that interact directly with it.

This closure of the cleft within the isolated ligand-binding domain is considered to be analogous to domain closure in the full-length native receptors. Domain closure has been hypothesized to pull the linker that connects the S1 and S2 domains. In native receptors, this pull or strain is considered to be the force responsible for rearrangement of the pore-forming membrane domains, leading to the opening of the transmembrane conduction path. Thus, each subunit can bind glutamate and undergo conformational changes that contribute to dilatation of the pore. In agreement with this view, single-channel analysis of intact receptor shows at least three conductance states that have been proposed to correspond to two, three, or four liganded subunits within a receptor complex. Pore conductance is therefore conceptually related to the fraction of subunit occupancy and activation within each receptor complex. This suggests that subunits can make incremental contributions to pore opening, gradually shifting the unitary conductance through the channel to higher levels as more subunits become activated. The observed concentration dependence of conductance levels supports this idea.

This view of fourfold rotational symmetry is also supported by the similarities of the pore-forming

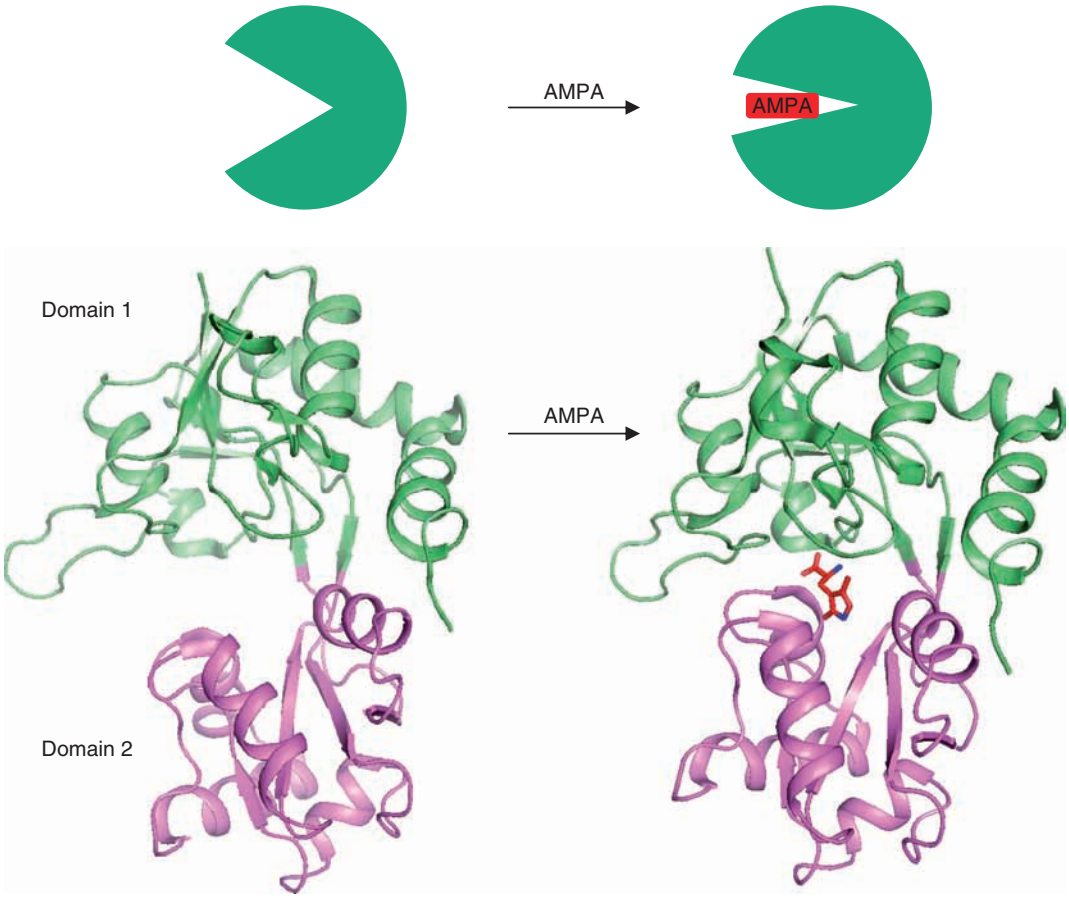

Figure 2 Conformational changes induced by agonist binding to the AMPA receptor. Crystal structure of the ligand-binding domain of the GluR2 subunit shown in nonliganded (apo –state; Protein Data Bank code 1FTO) and liganded (AMPA bound state; Protein Data Bank code 1FTM) states. Agonist binding in the cleft leads to closure of the clamshell (top).

AMPA receptor domain to potassium channels. However, several other features of AMPA receptors point to a twofold rotational symmetry for the tetrameric receptor complex. First, the glutamate-binding domains crystallize as dimers. Second, the D1–D1 dimer interface plays a role in channel activation by forming a structural scaffold that leads to the movement of D2. However, excessive tension can trigger rearrangement of the D1–D1 interface, leading to AMPA receptor desensitization. Mutations that affect desensitization lie on the interface. Additionally the AMPA receptor desensitization inhibitor cyclothiazide (see below) also acts on the dimer interface. Third, the reactivity of cysteine-modifying reagents on cysteine residues inserted into the M3 domain by site-directed mutagenesis fits well with a twofold symmetry rather than a fourfold symmetry. However, additional structural information about the full AMPA receptor complex would be needed to ascertain receptor symmetry.

Partial agonists are typically considered as ligands that induce a response at the maximally effective concentrations, which is lower than that of the endogenous ligand glutamate. The mechanism of action of partial agonists has been studied in AMPA receptors using a series of 5-substituted willardiines, which show lower efficacy compared to glutamate. The 5-substituted willardiines differ in only a single atom at the same position in the molecule. Structural and functional studies suggest that the degree of domain closure of the agonist binding cleft forms the basis for the partial agonist action at the AMPA receptor. Specifically, a combination of crystallographic and functional data show that the degree of domain closure of the clamshell is correlated with the efficacy of the agonist at an individual subunit such that agonists that induce less domain closure appear less effective in opening the channel. However, other conformational rearrangements of the agonist binding domain can also impact agonist efficacy.

Posttranscriptional Modification of AMPA Receptors

Alternative RNA Splicing

All four AMPA subunits undergo alternative RNA splicing in the C-terminal half of the M3–M4 loop, leading to so called flip/flop splice variants. The locations of flip/flop splicing for GluR1, GluR2, GluR3, and GluR4 are 742–793, 736–787, 740–791, and 737–788, respectively. In rats expression of the flip variant predominates up to postnatal day 8, but in adults both forms are expressed to a similar extent in many regions. The flip splice variant endows receptors with a diminished form of desensitization and a faster rate of recovery from desensitization as compared to the flop splice variant. GluR4-flop has the fastest desensitization (<1 ms at room temperature) among all the studied ionotropic glutamate receptors.

GluR2 and GluR4 subunits are also alternatively spliced at their C-terminal ends to generate short or long isoforms. A short isoform for GluR1 has not been reported, whereas GluR3 shows only the short isoform due to absence of splice sites. More than 90% of GluR2 is in the short form while the long form of GluR4 is predominant. The postsynaptic density/disc/zonula occludens-1 (PDZ) motif, which interacts with several proteins necessary for targeting of AMPA receptors to the synapse, is only present in the short form.

RNA Editing

One for the most interesting properties of AMPA receptor subunits is the consequence of differential RNA editing. The GluR2 subunit undergoes RNA editing at the so-called Q/R site (Q607) located at the tip of the reentrant loop of the second membrane-associated region. RNA editing leads to conversion of a glutamine codon to an arginine codon. This residue is a major determinant of the ionic selectivity of the pore. Replacement of the glutamine with an arginine leads to low calcium permeability of GluR2 subunit, low single-channel conductance, and a linear current–voltage relationship that reflects lack of block by intracellular polyamines (see later). Conversely, AMPA receptors lacking the GluR2 subunit have higher unitary conductances and show substantial Ca^{2+} permeability. The latter property may underlie some forms of plasticity at central synapses. Virtually all of the GluR2 subunit RNA is edited with high efficiency, and reduction in editing efficiency can lead to epilepsy and early death in mice. This may reflect the increased Ca^{2+} permeability of AMPA receptors, since it is well established that intracellular calcium overload can lead to cell death during glutamate-mediated neurotoxicity. Reduction in the efficiency with which GluR2 is edited has been proposed to play a role in schizophrenia, Huntington's disease, Alzheimer's disease, epilepsy, and malignant glioma, although more work will be required to evaluate these hypotheses.

Another site of RNA editing is the R/G site, where glycine replaces the arginine in GluR2–GluR4 subunits. The locations of the R/G site for GluR2, GluR3, and GluR4 are R764, R769, and R7665, respectively. This editing leads to reduced desensitization and rate of recovery from desensitization in GluR3 and GluR4 subunits. In adult mammals 50–90% of GluR2–GluR4 subunits are edited at the R/G site. GluR1 only has an arginine at the R/G site (R757).

Posttranslational Modification

Phosphorylation of AMPA Receptors

AMPA receptor function and trafficking is posttranslationally regulated by protein kinases and phosphatases. AMPA receptors are closely localized to cyclic adenosine monophosphate (cAMP)-dependent protein kinase A (PKA) and calcineurin by the binding of the A-kinase anchoring protein (known as AKAP) to scaffolding proteins within the postsynaptic density. In hippocampal neurons PKA phosphorylation potentiates AMPA receptor function by increasing the open probability, increasing the mean open time of the channel, and facilitating receptor trafficking to the membrane surface. PKA phosphorylates Ser845 residue present at the intracellular C-terminal tail of the GluR1 receptor. Mutation of this residue to alanine (S845A) eliminates the ability of PKA to potentiate GluR1-mediated currents. Phosphorylation and dephosphorylation of the Ser845 site are one important component of the molecular events that underlie long-term potentiation (LTP) and long-term depression (LTD) of synaptic strength, respectively, at certain synapses.

AMPA receptors are also phosphorylated by Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII), which is a major constituent of postsynaptic densities at the glutamatergic synapses and forms approximately 1% of the brain protein. CaMKII phosphorylates the Ser831 residue present exclusively in the GluR1 subunit among AMPA receptors. CaMKII enhances the current through native AMPA receptors in hippocampal neurons and increases the single-channel conductance of GluR1 receptors. This change in conductance reflects a shift in the proportion of subconductance to higher levels, rather than creation of new levels. This shift can be accounted for by changes in subunit-dependent gating, in which individual GluR1 subunits that are phosphorylated at Ser831 are more likely to become activated, increasing the probability that the receptor will be in the higher conductance states that require more activated subunits.

Protein kinase C (PKC) is a family of heterogeneous protein kinases activated by Ca^{2+} and/or phospholipids. PKC can be converted to a persistently active form by calpain-mediated proteolysis. GluR1 is phosphorylated by PKC at Ser831, which has similar effects on current response amplitude and conductance to CaMKII phosphorylation at this site.

The GluR2 subunit is phosphorylated by PKC at Ser880 and Ser863. PDZ domains are present in a number of proteins that are capable of meditating interaction with the AMPA receptors. The GluR2 Ser880 residue is present in the C-terminal sequence that is responsible for PDZ domain binding and its phosphorylation leads to decrease in binding to glutamate receptor-interacting protein 1 (GRIP 1), which can synaptically anchor GluR2. GluR2 Ser880 phosphorylation does not affect the binding to other targeting auxiliary proteins, such as PICK 1 (protein interacting with PKC 1). Phosphorylation by PKC at the Ser880 site facilitates the internalization of the GluR2 receptor.

Pharmacology of AMPA Receptors

AMPA Receptor Agonists

A number of natural products have been known for decades to have neurological or psychoactive effects, which were later attributed to actions on glutamate receptors. These include ibotenic acid, quisqualic acid, domoic acid, and willardiine (see **Figure 3** for structures). Ibotenic acid is derived from the mushroom *Amanita muscaria* (the fly mushroom, or toadstool). Quisqualic acid is obtained from the seeds and fruit of *Quisqualis chinensis*, domoic acid is a phycotoxin (algal toxin) associated with certain algal blooms and shellfish poisoning, and willardiine is from *Acacia willardiana* and *Mimosa asperata*. Although L-glutamate is the endogenous agonist of AMPA receptors, these receptors show higher selectivity for the synthetic ibotenic acid analog AMPA, thus deriving their name. Kainic acid can also activate AMPA receptors, although with 10- to 20-fold lower potency for GluR1 and GluR2 over subunits within the kainate receptor family, such as GluR5. AMPA receptors have lower affinity for glutamate compared to NMDA receptors.

Glutamate and AMPA act as full agonists and induce rapid desensitization, which is thought to involve rearrangement of the dimer formed by the ligand-binding domains. Kainic acid, which is the defining agonist for so-called kainate receptors composed of GluR5–GluR7 and KA1 and KA2, functions as a partial agonist and induces little desensitization on AMPA receptors. 5-Fluorowillardiine is not only more selective for AMPA receptors than AMPA itself (5-fluorowillardiine has a 70- to 150-fold higher affinity for GluR1 and GluR2 over GluR5), but also shows some selectivity for different AMPA receptor subunits.

AMPA Receptor Competitive Antagonists

Structurally different classes of competitive AMPA antagonists have been found which bind to the glutamate-binding site of the receptor. The first selective and useful AMPA receptor antagonists were 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) and 6,7-dinitroquinoxaline-2, 3-dione (DNQX). However,

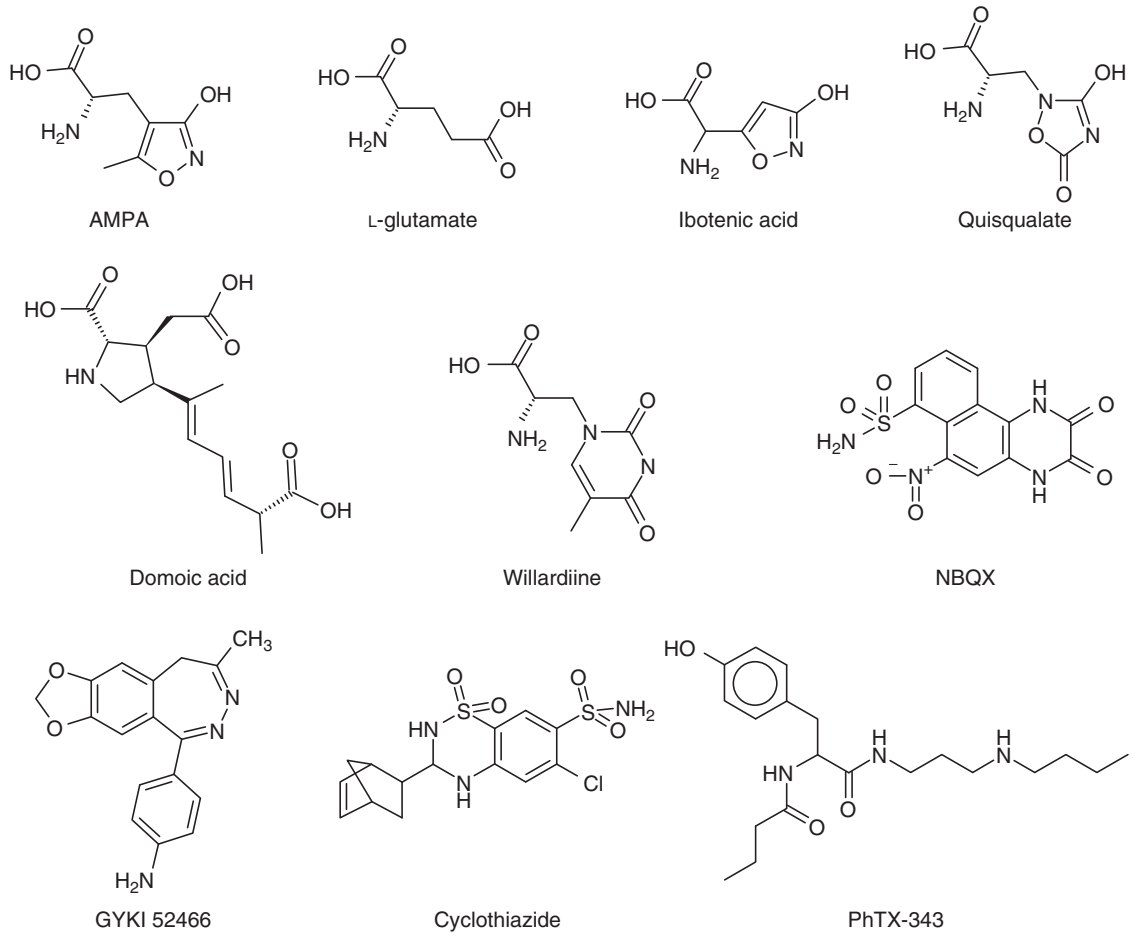

Figure 3 The structures of AMPA receptor agonists α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), L-glutamate, ibotenic acid, quisqualate, domoic acid, and willardiine; AMPA receptor antagonists 2,3-dihydroxy-6-nitro-7-sulfamoyl-benzo[*f*]quinoxaline (NBQX) and GYKI 52466; and AMPA receptor modulator cyclothiazide.

these quinoxaline derivates also have some affinity at the glycine-binding site of NMDA receptors. Newer antagonists have increased potency, higher AMPA receptor specificity, increased water solubility, and a longer duration of action *in vivo*. Examples of some of these competitive antagonists include 2,3-dihydroxy-6-nitro-7-sulfamoyl-benzo[*f*]quinoxaline (NBQX), 1,4,7,8,9,10-hexahydro-9-methyl-6-nitro-pyrido-[3,4-*f*]-quinoxaline-2,3-dione (PNQX), [2,3-dioxo-7-(1*H*-imidazol-1-yl)-6-nitro-1,2,3,4-tetrahydro-1-quinoxaliny]acetic acid (YM872), and [1,2,3,4-tetrahydro-7-morpholinyl-2,3-dioxo-6-(trifluoromethyl)quinoxalin-1-yl]methylphosphonate (ZK200775). Although some of these compounds have entered clinical trials for the treatment of ischemia, side effects and poor water solubility complicated clinical usage.

AMPA Receptor Noncompetitive Antagonists

The first noncompetitive AMPA receptor antagonist described was 1-(4-aminophenyl)-4-methyl-7, 8-methylenedioxy-5*H*-2,3-benzodiazepine hydrochloride (GYKI 52466), which is structurally classified as a benzodiazepine. A related series of selective, noncompetitive antagonists were derived from three classes of compounds: 2, 3-benzodiazepine (GYKI 52466), phthalazine (GYKI 53784/LY303070, GYKI 53773/LY300164), and quinazolinone derivatives (CP-465,022 and CP-526,427). The noncompetitive AMPA antagonists are of interest as potential therapeutic agents because they function effectively even in the presence of high levels of glutamate. Some noncompetitive AMPA receptors antagonists have been evaluated as drug candidates in a number of clinical trials for various neurological disorders.

Positive Allosteric Modulators

When AMPA receptors are activated by glutamate they rapidly desensitize in 1 (or a few) ms (Figure 4(a)). This desensitization has been hypothesized to be triggered by domain closure around the agonist, which exerts strain on the ligand-binding dimer

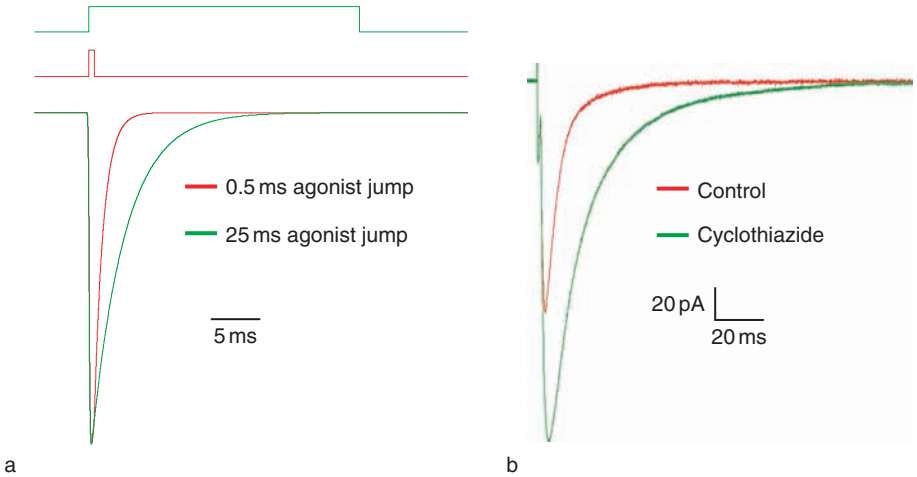

Figure 4 Electrophysiological properties of the AMPA receptor. (a) Simulated traces showing activation of GluR1 homomeric receptors by short (0.5 ms) and long (25 ms) pulses of 10 mM glutamate. The faster decay of current in the red trace activated by short pulse of glutamate represents deactivation. The relatively slower decay of current in green trace represents desensitization. (b) Evoked excitatory postsynaptic currents (eEPSC) from a pyramidal CA1 hippocampal neuron mediated by AMPA receptors in the presence of NMDA and GABA receptor antagonists. The eEPSC in green is in the presence of AMPA receptor desensitization inhibitor cyclothiazide. (a) Model used for simulation from Partin KM, Fleck MW, and Mayer ML (1996) AMPA receptor flip/flop mutants affecting deactivation, desensitization, and modulation by cyclothiazide, aniracetam, and thiocyanate. *Journal of Neuroscience* 16(21): 6634–6647.

interface. The strain can be relieved by opening of the channel or breakdown of the ligand-binding interface, which triggers desensitization. A number of AMPA receptor-positive allosteric modulators have been identified, sometimes referred to as AMPA receptor potentiators. These allosteric potentiators do not activate the AMPA receptors themselves, but enhance the agonist-evoked current by slowing the rate of desensitization (Figure 4(b)).

Drugs that positively regulate AMPA receptors fall into three classes: pyrrolidones (also known as ampakines, including aniracetam and piracetam), the benzothiadiazides (cyclothiazide, diazoxide), and biarylpropylsulfonamides (4-[2-(phenylsulfonylamino)ethylthio]-2,6-difluorophenoxyacetamide (PEPA), *N*-2-[4-(4-cyanophenyl)phenyl]propyl-2-propanesulfonamide (LY404187)). Both aniracetam and cyclothiazide almost completely eliminate desensitization of flip splice variants but only slow the entry into desensitized state for flop splice variant. PEPA potentiates AMPA receptor function by attenuating receptor desensitization without any effect on deactivation; Cyclothiazide and PEPA are approximately 100–500 times more potent compared to aniracetam in potentiating AMPA ion currents. PEPA has varying degrees of selectivity for the flop variants. LY404187 suppresses receptor desensitization with a distinct time dependence in the presence of agonist and shows high potency for flip splice variants. Accumulated evidence suggests that AMPA-positive modulators may act as cognitive enhancers, which could be useful as novel therapeutic agents for treating brain disorders such as Alzheimer's disease and schizophrenia.

The binding site of compounds that relieve desensitization appears to reside within the ligand-binding dimer interface. Crystallographic studies of aniracetam suggest that binding can stabilize the ligand-binding core dimerization. This supports the interpretation that desensitization involves breakdown of the dimer interface.

Pore-Blocking Molecules

The current–voltage relationship for AMPA receptors is controlled by subunit composition, with GluR2 endowing receptors with a linear current–voltage curve that reverses near 0 mV. By contrast, AMPA receptors lacking the GluR2 subunit show inward rectification, a term that means cations can flow through the channels into cells at negative potentials more easily than they can flow out of the channel at positive potentials. Inward rectification occurs due to the presence of intracellular polyamines such as spermine, which can block the flow of current out of the cell by entering and blocking the conduction path. A number of elongated and positively charged molecules such as spermine, spermidine, and polyamine toxins (ArgTX-636, PhTX-343, JSTX-3) are noncompetitive open-channel blockers of cation-conducting channels. Spermine is expressed in many CNS neurons and is implicated in regulation of cell division, protein synthesis, and perhaps specific functions in the nervous system. Polyamine toxins are nonoligomeric, low-molecular-weight compounds isolated from spiders and wasps.

The mechanism by which polyamines block AMPA receptors is thought to involve a deep binding site

within the receptor pore, the binding to which appears voltage dependent. At positive potentials, polyamine binding is enhanced as the membrane potential favors entry of polyamines from the cell into the channel pore. At sufficiently positive potentials, polyamine block can be relieved as polyamines leave their block site to permeate the channel, joining the outward flow of cations. As described previously, GluR2 subunits that have been edited at the Q/R site at the apex of the reentrant loop have a positively charged arginine at the Q/R site in GluR2, as opposed to glutamine in GluR1, GluR3, and GluR4. The presence of an arginine at this site removes inhibition by intracellular polyamines and like molecules, rendering the current–voltage relationship linear.

See also: AMPA Receptor Cell Biology/Trafficking; AMPA Receptors: Disease; Glutamate Receptor Organization: Ultrastructural Insights; Kainate Receptors: Molecular and Cell Biology; Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology; NMDA Receptors, Cell Biology and Trafficking.

Further Reading

Armstrong N and Gouaux E (2000) Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: Crystal structures of the GluR2 ligand binding core. *Neuron* 28: 165–181.

Dingledine R, Borges K, Bowie D, et al. (1999) The glutamate receptor ion channels. *Pharmacological Reviews* 51: 7–61.

Erreger K, Chen PE, Wyllie DJA, et al. (2004) Glutamate receptor gating. *Critical Reviews in Neurobiology* 16: 187–225.

Gouaux E (2004) Structure and function of AMPA receptors. *Journal of Physiology* 554: 249–253.

Kew JN and Kemp JA (2005) Ionotropic and metabotropic glutamate receptor structure and pharmacology. *Psychopharmacology (Berlin)* 179(1): 4–29.

Marenco S and Weinberger DR (2006) Therapeutic potential of positive AMPA receptor modulators in the treatment of neuropsychiatric disorders. *CNS Drugs* 20: 173–185.

Mayer ML and Armstrong N (2004) Structure and function of glutamate receptor ion channels. *Annual Reviews in Physiology* 66: 161–181.

Nicoll RA, Tomita S, and Bredt DS (2006) Auxiliary subunits assist AMPA-type glutamate receptors. *Science* 311: 1253–1256.

Palmer CL, Cotton L, and Henley JM (2005) The molecular pharmacology and cell biology of alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors. *Pharmacological Reviews* 57(2): 253–277.

Partin KM, Fleck MW, and Mayer ML (1996) AMPA receptor flip/flop mutants affecting deactivation, desensitization, and modulation by cyclothiazide, aniracetam, and thiocyanate. *Journal of Neuroscience* 16(21): 6634–6647.

Relevant Website

http://www.iddb.com – Investigational Drugs Database.

Amphetamines

R Kuczenski, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Amphetamine is one of a group of powerful central nervous system psychostimulants which also includes, but is not limited to cocaine, methylphenidate, and methamphetamine. The psychostimulants are highly lipid soluble, and readily enter the brain to promote a wide range of behavioral effects. The prototypic drug, amphetamine (α-methylphenethylamine), was introduced in the 1930s, and during the next decade rapidly became popular both for its therapeutic potential and for its abuse liability. Within a few years of its introduction, behavioral changes ranging from its effects in improving the academic performance of 'problem children' and its antinarcoleptic actions to its ability to induce psychosis were recognized.

All of the amphetamine-like psychostimulants are capable of affecting the biogenic amine transmitters (dopamine, norepinephrine, serotonin), but in general, the most pronounced behavioral effects are mediated primarily by their interactions with dopamine neurons. However, it is important to recognize that relatively small structural modifications can significantly alter the pharmacology of the drug, for example, shifting the primary neurochemical effects to serotonin rather than dopamine (*p*-chloroamphetamine (PCA)) or producing powerful hallucinogens (e.g., 2,5-dimethoxy-4-methylamphetamine (DOM, 'STP')) (Figure 1).

Molecular Mechanisms

In vivo microdialysis studies in awake, behaving animals have revealed that all amphetamine-like stimulants promote a profound increase in extracellular dopamine in all brain regions that receive dopamine projections, and converging evidence has implicated these changes in the variety of behavioral effects induced by these drugs. Most data regarding the mechanisms of action of amphetamine are derived from studies of dopamine projections to striatal areas, including caudate and nucleus accumbens, because these regions receive dense dopamine innervation. In these regions at least, the quantitative features of the temporal profile for the amphetamine-(as well as cocaine-) induced increases in extracellular dopamine closely parallel the drug's pharmacokinetic profile. The dopamine responses in these regions have also provided the basis for structure–activity studies of potency among amphetamine derivatives. For example, amphetamine exists in two stereoisomeric forms, and in terms of its interaction with dopamine systems: the dextrorotatory S(+)– form is about threefold more potent than the levorotatory R(–)– form. The S(+)– form of the *N*-methylated amphetamine derivative, methamphetamine (*N*-methyl-α-methylphenethylamine), is about equipotent to S(+)– amphetamine, whereas the R(–)– form of methamphetamine is considerably weaker.

The specific mechanism underlying the stimulant-induced increase in extracellular dopamine depends on the type of psychostimulant (i.e., 'uptake blocker,' such as cocaine or methylphenidate, or 'releaser' such as amphetamine or methamphetamine). The normal physiological release of dopamine occurs through a calcium-dependent fusion of the synaptic vesicle with the plasma membrane, followed by the release of the transmitter content of the vesicle into the synaptic space. The released transmitter can interact with its target receptors, following which its active life in the synaptic/extracellular space is terminated by transport or uptake through the Na^+/Cl^--dependent dopamine transporter back into the nerve terminal. All amphetamine-like stimulants increase extracellular dopamine as a consequence of their direct binding to the dopamine transporter. First, by occupying the transporter, they compete for the uptake site with dopamine that has been released into the synaptic space, thereby interfering with the normal removal of transmitter from the synaptic space and back into the nerve terminal. As a consequence, the time during which released dopamine can interact with receptors is prolonged. This is the primary mechanism by which the so-called uptake blockers, including methylphenidate and cocaine, enhance dopamine neurotransmission. However, amphetamine (and methamphetamine) is also a substrate for the dopamine transporter. In addition to inhibiting dopamine uptake, the amphetamines can be transported into the nerve terminal. One consequence of the transport and accumulation of amphetamine within the nerve terminal is the subsequent release of cytoplasmic, nonvesicular dopamine from the nerve terminal and into the synaptic/extracellular space through a dopamine transporter-dependent process. The most common model for how amphetamine promotes this release of dopamine involves reversal of transporter function through a facilitated exchange diffusion process in which the transporter moves cytoplasmic dopamine into the synaptic space; that is, transport of amphetamine from the synaptic space and into the cytosol increases the availability of inward-facing

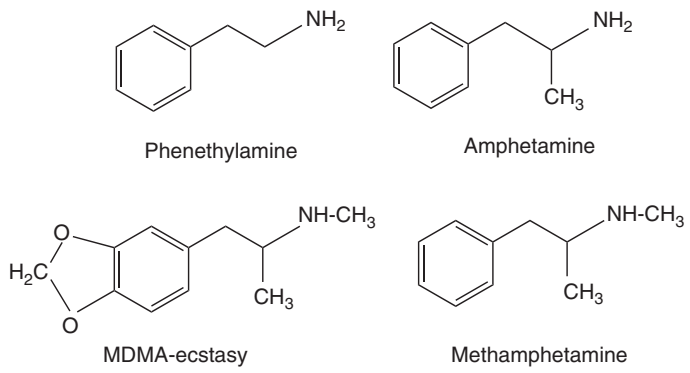

Figure 1 Phenylethylamine is the parent compound for many of the amphetamines.

transporter binding sites within the nerve ending to bind dopamine and then transport dopamine out. Most data support the participation of this process in amphetamine-induced dopamine release, but there are some characteristics of the release that cannot be fully explained only by facilitated exchange diffusion. In this regard, more recent evidence suggests that amphetamine can activate a channel-like mode of the dopamine transporter that supports efflux of dopamine out of the nerve ending. Dopamine efflux through the channel occurs in bursts that approximate a quantum of dopamine equivalent to vesicular release, and these bursts appeared to contribute 10–15% of the total amphetamine-induced release. Unlike amphetamine, dopamine itself does not stimulate this dopamine transporter channel-like activity. Thus, the potential physiological role of the channel-like mode of the transporter in the absence of amphetamine is unclear.

Both the facilitated exchange diffusion and channel mechanisms rely on the availability of dopamine in the cytosol. Although most of the nerve-ending dopamine is confined to vesicles, the enzymes for dopamine biosynthesis are outside the vesicle and in the cytoplasm. Thus, under physiological conditions, some transmitter would be available in the cytosol, in transit to the vesicle, to support the amphetamine-induced release process. However, because amphetamine is highly lipophilic, it can also readily enter the neuron by simple diffusion across the plasma membrane, in addition to its accumulation by the transporter, and once inside the nerve ending at sufficiently high concentrations, amphetamine exerts two actions that can augment the content of dopamine in the cytoplasmic pool. First, amphetamine promotes the redistribution of dopamine from the vesicle into the cytosol by collapsing the vesicular pH gradient. The interior of the vesicle is normally acidified and the electrochemical gradient contributes to the energy required by the vesicle to accumulate and retain transmitter. Amphetamine, a lipophilic weak base, accumulates in the vesicle and is protonated, thereby collapsing the pH gradient, disrupting transmitter storage, and enabling leakage of dopamine into the cytosol. Amphetamine also binds to the vesicular monoamine transporter to inhibit vesicular uptake of dopamine. Because of the relatively high concentration of transmitter within the vesicle, disruption of the storage process contributes substantially to the cytoplasmic pool of transmitter. Furthermore, at sufficiently high concentrations, amphetamine also inhibits monoamine oxidase, the primary catabolic enzyme for dopamine in the nerve terminal. Thus, through the combination of disrupted vesicular storage and inhibited catabolism, the catecholamine releasers enhance the availability of cytoplasmic transmitter to provide sufficient substrate for ongoing amphetamine-mediated, transporter-dependent dopamine release.

In addition to their interactions with the dopamine transporter, the amphetamine-like psychostimulants also exhibit some affinity for the norepinephrine and serotonin transporters, and increase extracellular concentrations of both of these transmitters to varying degrees, depending on the specific stimulant. For example, although methamphetamine and amphetamine have similar potency at the dopamine transporter, methamphetamine exhibits higher potency for affecting serotonin than does amphetamine. In contrast, methylphenidate, with little affinity for the serotonin transporter, has minimal effects on extracellular serotonin, even at relatively high doses, although it readily interacts with the norepinephrine transporter. Similar to their interactions with the dopamine transporter, cocaine and methylphenidate act at the norepinephrine and serotonin transporters strictly as uptake blockers, whereas amphetamine and methamphetamine can release both norepinephrine and serotonin, apparently through transporter-dependent mechanisms comparable to amphetamine-mediated, nonvesicular dopamine release. However, amphetamine-induced release of norepinephrine appears to be restricted to doses of the drug sufficiently high to disrupt vesicular storage of the

transmitter. Because the final step in norepinephrine biosynthesis involves the intravesicular enzyme, dopamine β-hydroxylase, it has been argued that the cytoplasmic pool of norepinephrine is meager under normal conditions, and would not provide sufficient substrate to support amphetamine-induced release. Thus, at lower doses of amphetamine, drug-induced increases in extracellular norepinephrine likely occur through simple amphetamine blockade of uptake. At higher doses which result in sufficient intracellular concentrations of amphetamine to disrupt vesicular storage of norepinephrine, the cytoplasmic pool of this transmitter can be augmented to provide sufficient substrate for transporter-dependent release.

Behavioral Effects

Acute Administration

Amphetamine and the psychomotor stimulants exert a wide-range of dose-dependent behavioral changes ranging from therapeutic effects to toxic psychosis. These effects involve cognitive, affective, and motor processes. One fundamental behavioral effect of the psychomotor stimulants is that they induce arousal, or wakefulness. The term 'arousal' typically refers to the extent to which an organism is responsive to environmental stimuli. Most evidence suggests that both dopamine and norepinephrine each contribute to wakefulness, and in rodents psychostimulant-induced arousal appears to depend on the actions of the drug on both transmitters within a widely distributed network of subcortical regions, including the medial septal area, the medial preoptic area, and the lateral hypothalamus. These alerting effects of the amphetamines likely contribute to their ability to diminish fatigue and their efficacy in the treatment of narcolepsy, and may also play an indirect role in their ability to enhance cognitive function. In addition, however, at the lowest doses, amphetamine also facilitates a variety of cognitive effects, including enhanced information processing, sustained attention, and working memory, effects which may be important in the therapeutic efficacy of stimulants in the treatment of attention deficit/hyperactivity disorder. These influences on cognitive function appear to involve prefrontal cortical processes that depend on separate stimulant-induced contributions from both dopamine, acting at D1 receptors, and norepinephrine, acting at $\alpha 2$ receptors. However, catecholamine facilitation of cortical cognitive function exhibits an inverted U-shaped dose response, such that higher doses of stimulants disrupt these processes.

As the dose increases, amphetamine stimulates behavioral activation which, in rodents, initially presents as increased locomotion, including both horizontal and vertical movements. With further dose increases, the enhanced locomotion first becomes less varied and more perseverative, following repetitive paths and patterns, then ultimately, is replaced by highly repetitive stereotyped movements which become the predominant feature of the response. Depending on the dose, stereotyped movements can persist for prolonged periods (several hours) in the absence of locomotion, and, as the drug is metabolized, the stereotyped behaviors are replaced by a poststereotypy phase of enhanced locomotion. Comparable behavioral characteristics have been described for most mammals, including primates. The specific behaviors that occur during stimulant-induced stereotypy are highly species specific. For example, rats engage in repetitive sniffing, licking, and gnawing, focused on a particular object; pigeons engage in repetitive pecking, and dogs continuously pace. Human behaviors can appear as movement sequences without an apparent purpose, but are frequently more complex and manipulatory, such as repetitive sorting of objects, compulsive cleaning or bathing, or assembling and dissembling a mechanical object. Stimulant abusers have described stereotyped behaviors as highly pleasurable, funny, and relaxing. Individuals appear to perceive stereotypies as have a calming effect, and become irritated and/or anxious when the stereotypy is interrupted. Converging evidence, developed primarily in rodents, but extended to primates as well, implicates dopamine as the primary neurotransmitter in stimulant-induced motor effects. In particular, stimulant-induced increases in dopaminergic neurotransmission in the nucleus accumbens, a major projection area of the dopamine pathway originating in the ventral tegmental area, mediate the locomotor effects, whereas the caudate, receiving its dopamine projection from the substantia nigra, plays a critical role in the stereotypy effects of the drugs.

The amphetamines are particularly notorious for their ability to induce a state of pleasurable affect, elation and euphoria, and are therefore highly abused substances. Although slow-onset routes of drug delivery can promote an enhanced mood, the rapid increase in brain concentrations of the drug associated with intravenous administration, or inhaling the vapors of volatile methamphetamine hydrochloride through smoking, appears to result in a particularly intense pleasurable feeling, or flash, which is more often sought after by compulsive abusers. Subsequent to this stage of euphoria, the pleasurable effects of the drug can persist for prolonged periods because of the slow metabolism and elimination of the drug (amphetamine and methamphetamine have half-lives

in humans of 6–12 h). Neuropharmacological studies have established an important role for dopamine in the reinforcing effects of amphetamine, and, in particular, the mesocorticolimbic dopamine system which projects from the ventral tegmental area to the nucleus accumbens and frontal cortex. For example, most drugs of abuse increase dopamine release in the nucleus accumbens, and laboratory animals will self-administer amphetamine directly into this brain region. In addition, selective destruction of this pathway eliminates amphetamine self-administration. As a consequence, studies directed at understanding psychostimulant addiction, drug craving, and relapse have focused on the adaptational changes in this dopamine pathway associated with chronic exposure to the drug.

Chronic Administration

Sensitization With the repeated, intermittent administration of amphetamine-like stimulants, the motor behaviors induced by the initial treatment with the drug exhibit a sensitized or augmented response to subsequent drug administration. Sensitization (sometimes called reverse tolerance) refers to a progressive enhancement of the locomotor and stereotypy responses with repeated exposure to any of this class of drugs, and the sensitized behavioral response profile persists for prolonged periods of abstinence. Psychomotor sensitization has been reported in all experimental mammalian species that have been examined, including primates, and some evidence suggests that sensitization can be observed in some stimulant response characteristics in humans, as well. In addition, cross-sensitization is evident among the amphetamine-like stimulants; that is, repeated treatment with amphetamine results in an enhanced response to subsequent administration of, for example, cocaine, but not in response to other stimulant-like drugs such as caffeine, whose mechanisms of action are distinct from the psychomotor stimulants. Sensitization does not appear to be a unitary phenomenon and does not simply reflect a shift to the left of the dose–response curve, because the augmented behavioral profile is not reproduced by a higher dose of the drug, and the duration of the response is not extended as would be expected with a higher dose. Rather, behavioral sensitization appears to arise from multiple neuroadaptations in multiple brain nuclei. The mechanisms underlying sensitization appear to include neural restructuring and neurochemical adaptations in the nucleus accumbens and prefrontal cortex. Prominent adaptations involve a long-lasting hyper-responsiveness of mesolimbic dopaminergic and cortical glutamatergic pathways. Because these regions are involved in reward and motivation, sensitization has been speculated to be a factor in the drug craving and compulsive drug-seeking behavior associated with amphetamine abuse, and the neuronal changes underlying the process of sensitization have generated interest as potential adaptations contributing to the process of addiction.

Tolerance Although stimulant-induced motor behaviors can exhibit augmentation with chronic stimulant exposure, depending on the pattern of drug administration, many effects of these drugs exhibit tolerance. For example, the potent anorexic effects of amphetamine, which appear to be mediated by its actions in the lateral hypothalamus, exhibit relatively rapid tolerance, which limits the drug's usefulness as an appetite suppressant. The sympathomimetic effects also exhibit tolerance with chronic exposure, as does the hyperthermia associated with high doses of the drug. Furthermore, the euphoric effects also appear to exhibit tolerance. The tolerance that develops to these effects does not appear to have a pharmacokinetic basis, based on the observations that chronic amphetamine and methamphetamine abusers exhibit drug metabolism profiles similar to those of naïve subjects. Tolerance to the euphoric effects likely contributes to the progressive dose escalation that almost inevitably occurs as those abusers who develop more severe abuse patterns attempt to maintain the intensity of their drug experience. That tolerance develops to the potentially toxic sympathomimetic effects enables compulsive users to self-administer the drug at exceedingly high levels, frequently as much as grams per day, without the lethality that would be experienced by naïve users.

Psychosis Prolonged exposure to high doses of the amphetamines can lead to the development of a paranoid, delusional psychotic state, which can be accompanied by auditory and tactile hallucinations, and which has occasionally been misdiagnosed as schizophrenia. It has been suggested that this drug-induced paranoid state, which frequently evokes considerable anxiety and intense fear, is responsible for the reckless, often aggressive antisocial behavior associated with high-dose stimulant intoxication. Although paranoid psychosis has been elicited with low dose and/or relatively short-term amphetamine exposure, most clinical evidence suggests that this pathological state emerges after chronic stimulant administration. Its full development appears to be related to the amount and duration of drug use, with clear psychosis and severe paranoia more commonly appearing during 'binge' or 'run' patterns of use, when individuals administer the drug throughout the day and night for several days. Some evidence suggests that

paranoid psychosis may not occur during the first binge, but rather requires multiple binges before becoming clearly manifested. However, once elicited, this state appears progressively earlier in subsequent binge exposures. Most available data suggest that the paranoid symptoms dissipate within days after discontinuation of drug administration, and thus may include a component reflecting drug-induced pharmacodynamics.

Neurotoxicity Exposure of experimental animals to acute, high doses of amphetamine or methamphetamine results in relatively persistent alterations in dopamine neurons innervating the dorsal striatum. These changes, which have generally been referred to as neurotoxicity, appear to involve the loss of phenotypic markers of dopamine neurites in the striatum without apparent damage to the dopamine perikarya in the substantia nigra. The decrements include levels of dopamine, the dopamine biosynthetic enzymes, tyrosine hydroxylase and aromatic amino acid decarboxylase, and both the plasma membrane dopamine transporter, and the vesicular monoamine transporter. Loss of the vesicular monoamine transporter is frequently considered as an indicator of damage to the nerve terminals rather than reflecting adaptational changes to the high-dose stimulant treatment. The neurotoxicity is facilitated by the hyperthermia associated with acute high-dose stimulant exposure. Accumulating evidence indicates that drug-induced oxidative stress involving reactive oxygen/nitrogen species, and in particular, high levels of oxidized cytoplasmic dopamine consequential to the amphetamine-induced disruption of vesicular storage, plays an important role in the persistent alterations. In addition, microglial activation accompanies these alterations. Evidence identifying potential behavioral correlates of the drug-induced damage has not been forthcoming. Neuroimaging studies using positron emission tomography (PET) technology and radioligands for the dopamine transporter have also revealed relatively persistent decrements in this marker of striatal dopamine nerve terminals in long-term high-dose methamphetamine abusers. Likewise, limited postmortem data from methamphetamine abusers indicates decreases in the dopamine transporter as well as the dopamine biosynthetic enzymes, although no decrement in the vesicular monoamine transporter has been detected. Nevertheless, these results have been generally interpreted to suggest that methamphetamine abuse can lead to prolonged functional damage and/or injury to striatal dopamine terminals, and that these effects may contribute to the enduring impairments of cognitive and motor functioning that have been reported in abstinent methamphetamine abusers.

See also: Addiction: Neurobiological Mechanism; Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine; Cognition: Neuropharmacology; Dopamine; Drug Addiction: Behavioral Neurophysiology; Drug Addiction: Cellular Mechanisms; 3,4-Methylenedioxymethamphetamine (MDMA, "Ecstasy"); Noradrenaline; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Stimulant and Wake-Promoting Substances.

Further Reading

Angrist B (1994) Amphetamine psychosis: Clinical variations of the syndrome. In: Cho AK and Segal DS (eds.) *Amphetamine and Its Analogs: Psychopharmacology, Toxicology, and Abuse*, pp. 387–414. New York: Academic Press.

Berridge CW (2006) Neural substrates of psychostimulant-induced arousal. *Neuropsychopharmacology.* 31: 2332–2340.

Davidson C, Gow AJ, Lee TH, and Ellinwood EH (2001) Methamphetamine neurotoxicity: Necrotic and apoptotic mechanisms and relevance to human abuse and treatment. *Brain Research. Brain Research Reviews* 36: 1–22.

Kuczenski R and Segal DS (1994) Neurochemistry of amphetamine. In: Cho AK and Segal DS (eds.) *Amphetamine and Its Analogs: Psychopharmacology, Toxicology, and Abuse*, pp. 81–113. New York: Academic Press.

Mehta MA, Sahakian BJ, and Robbins TA (2001) Comparative psychopharmacology of methylphenidate and related drugs in human volunteers, patients with ADHD, and experimental animals. In: Solanto MV, Arnsten AFT, and Castellanos FX (eds.) *Stimulant Drugs and ADHD Basic and Clinical Neuroscience*, pp. 303–331. New York: Oxford University Press.

Nichols DE (1994) Medicinal chemistry and structure–activity relationships. In: Cho AK and Segal DS (eds.) *Amphetamine and Its Analogs: Psychopharmacology, Toxicology, and Abuse*, pp. 3–41. New York: Academic Press.

Robinson TE and Berridge KC (1993) The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research. Brain Research Reviews* 18: 247–291.

Segal DS and Kuczenski R (1994) Behavioral pharmacology of amphetamine. In: Cho AK and Segal DS (eds.) *Amphetamine and Its Analogs: Psychopharmacology, Toxicology, and Abuse*, pp. 115–175. New York: Academic Press.

Sulzer D, Sonders MS, Poulsen NW, and Galli A (2005) Mechanisms of neurotransmitter release by amphetamines: A review. *Progress in Neurobiology* 75: 406–433.

Vanderschuren LJ and Kalivas PW (2000) Alterations in dopaminergic and glutamatergic transmission in the induction and expression of behavioral sensitization: A critical review of preclinical studies. *Psychopharmacology (Berlin)* 151: 99–120.

Amphibian Peptides

P Melchiorri and L Negri, University of Rome, Rome, Italy

© 2009 Elsevier Ltd. All rights reserved.

The secretions of the olocryne granular glands of amphibian skin contain a wide variety of bioactive peptides, nearly all of which are ancestors of mammalian neuropeptides and hormones and are involved in basic cellular processes conserved throughout evolution.

A core of gene clusters within the granular gland cells of frog skin codes for these secreted peptides. A mild transdermal electrical stimulation of the frog releases granular gland contents on the skin surface as a result of glandular syncythia rupture produced by the contraction of myoepithelial cells surrounding the glands. These skin secretions contain the entire peptidome, transcriptome, and genome of the granular gland syncythia (olocryne secretion). The polyadenylated mRNAs constituting the secreted transcriptome and the peptides constituting the secreted proteome are protected from degradation by interactions with co-released amphipathic peptides and mucoproteins endowed with antimicrobial, RNAse-, and protease-inhibitory activities. Thus, amino acid sequencing of secreted peptides, nucleic acid sequencing of peptide-encoding mRNAs, and genomic information retrieving from the secreted DNA can be easily performed in secretion samples collected from few frogs on a regular basis and stored, lyophilized or frozen, for at least 6 years. This is a powerful method of determining evolutionary information on the ancestral sequences of biologically active peptides and proteins and understanding the sequence–function relationship of the human orthologs. Because difficult-to-synthesize bioactive peptides and proteins are currently obtained in large quantities from skin secretions of few amphibian specimens, extensive pharmacological studies have been performed with these amphibian molecules in order to elucidate the functional role of the mammalian orthologs. Whereas some of amphibian skin peptides represent analogs of already known mammalian peptide families (Table 1), others represent prototype peptides not encountered before in nature. In many instances, the discovery of new amphibian skin peptides led to the discovery of novel mammalian neuropeptides; notable examples are caerulein, bombesin, and sauvagin (Table 2). In the amphibian skin, opioid peptides are represented by two prototypes named dermorphin and deltorphin. They differ from mammalian opioid peptides by their amino acid sequences that selectively bind μ-opioid receptors (dermorphin) or δ-opioid receptors (deltorphin) (Table 3). Finally, a new amphibian peptide family, the Bv8-related peptides, has been recently identified (Table 4). Bv8-related peptides represent the ancestors of the novel chemokine-related mammalian peptides, the prokineticins.

Amphibian Skin Opioids

The first peptide family of amphibian opioids was discovered in 1981 and named dermorphins. Until the discovery of mammalian endomorphins by Zadina and colleagues, these peptides represented the most potent and selective μ-opioid receptor agonists identified in living organisms. Nine years later, deltorphins were discovered in the amphibian skin; these peptides are still the most potent and selective δ-opioid agonists available today.

A unique characteristic of amphibian opioid peptides (Table 3) is the presence in the second N-terminal position of a D-amino acid residue that confers on these compounds high resistance to enzyme degradation. Amphibian opioids have been found only in the skin of South American hylid frogs belonging to the subfamily Phyllomedusinae (*Phyllomedusa*, *Agalychnis*, and *Pachymedusa* species). Although pharmacologists discovered these opioids in Amazonian frogs comparatively recently, the Matses of the upper Amazonian basin knew of the pharmacological properties of amphibian skin opioids for a long time. For centuries they had habitually applied the dried skin secretions of *Phyllomedusa bicolor*, called *sapo* (the Spanish word for toad), to cuts in their skin during shamanistic hunting rituals. The abundance of deltorphins and dermorphins acting together with, or probably synergistically with, the other active peptides present in these secretions might have caused analgesia and behavioral excitation in the hunters.

Origin of the Amphibian Skin Opioids

The D-amino-acid-containing opioid peptides issue from precursors showing a common preproregion (a 22-residue signal peptide and a 18- to 25-residue acidic prosequence) with precursors of the peptide antibiotics dermaseptins (24- to 34-residue polycationic and α-helical amphipathic peptides). Of the three types of dermal glands (mucous, lipid, and granular) in the skin of *Phyllomedusa,* only the granular glands are specifically involved in the biosynthesis and secretion of dermaseptins and deltorphins. The granular glands are the largest glands in the *Phyllomedusa* skin; they lie deeper in the epidermis, are lined by epithelium that is more a syncytium, and are

Table 1 Amphibian and mammalian kinins

Amphibian peptides		*Mammalian peptides*
Aromatic tachykinins		
Physalaemus biligonigerus		Substance P (SP) RPKPQQFFGLM-NH_2
Physalaemin	pEADPNKFYGLM-NH_2	Endokinin A (EKA)
Uperoleia rugosa [Lys^5,Thr^6]physalaemin	pEADPKTFYGLM-NH_2	DGGEEQTLSTEAETWVIVALEEGAGPSIQLQLQEVKTGKASQFFGLM-NH_2
Uperoleia marmorata Uperolein	pEPDPNAFYGLM-NH_2	Endokinin B (EKB) DGGEEQTLSTEAETWEGAGPSIQLQLQEVKTGKASQFFGLM-NH_2
Uperoleia inundata Uperin	pEADPNAFYGLM-NH_2	Endokinin-1 (EK-1) GKASQFFGLM-NH_2
Kassina (Hylambates) maculata Hylambatin	NPPDPNRFYGMM-NH_2	Hemokinin-1 (HK-1) SRTRQFYGLM-NH_2
Rana margaratae Ranamargarin	DDASDRAKKFYGLM-NH_2	
Pseudophryne guntheri PG-SPI	pEPNPDEFYGLM-NH_2	
PG-SPII	pEPNPNEFYGLM-NH_2	
Agalychnis callidryas AC-AR1	GPPDPDRFYPGM-NH_2	
AC-AR2	GPPDPDKFYPGM-NH_2	
AC-AR3	pEPDPDKFYPGM-NH_2	
AC-AR4	GPPDPNKFYPVM-NH_2	
Aliphatic tachykinins		
Kasina senegalensis		Neurokinin A (NKA)
Kassinin	DVPKSDQFVGLM-NH_2	HKTDSFVGLM-NH_2
Kassina (Hylambates) maculata		Neurokinin B (NKB) DMHDFFVGLM-NH_2
[G1u2,Pro5]kassinin	DEPKPDQFVGLM-NH_2	Neuropeptide K (NPK)
Phyllomedusa bicolor Phyllomedusin	pENPNRFIGLM-NH_2	DADSSIEKQVALLKALYGHGQISHKRHKTDSFVGLM-NH_2
Pseudophryne guntheri		Neuropeptide γ(NPγ)
PG-KI	pEPHPDGFVGLM-NH_2	DAGHGQISHKRHKTDSFVGLM-NH_2
PG-KII	pEPNPDEFVGLM-NH_2	
PG-KIII	pEPHPNDFVGLM-NH_2	
Agalychnis callidryas AC-AL	GPPDPNKFIGLM-NH_2	
Bradykinins		
Rana temporaria		Bradykinin: RPPGFSPFR Kallidin: KRPPGFSPFR
Ranakinin-1	LLPIVGRPPGFSPFR	
Ranakini-2	RPPGFSPFRIA	
Ranakinin-3	RPPGFSPFRIAPAL	
Ranakinin-4	RPPGFSPFRIAPASI	
Phyllomedusa rhodei phyllokinin	RPPGFSPFRIY(SO_3)	
Bombina orientalis Bombikinin	RPPGFSPFRGKFH	

surrounded by a layer of myoepithelial cells involved in the holocrine rapid discharge of secretory products collected in roundish granules. The granules do not bud off from the membranes on the Golgi apparatus but seem to be generated in the vacuoles of the vacuolated stage during gland development.

All the amphibian skin opioids contain the N-terminal sequence Tyr-D-Xaa-Phe, where the aromatic residues of Tyr^1 and Phe^3 are of L-configuration and the D-Xaa in the second position of the molecule is a D-amino acid (D-Ala or D-Met). The D-enantiomer is encoded, however, by the codon for the L-isomer in the precursor cDNA. Thus L-Xaa^2 must be converted to D-Xaa^2 by an unusual posttranslational reaction that presumably takes place in the precursor itself. Because [L-Xaa^2]-containing peptides have never been found in amphibian skin extracts, the epimerization mechanism probably involves a quantitative inversion of the chirality of the

Table 2 Amphibian peptides that led to the discovery of mammalian neuropeptides

Amphibian peptides	*Amphibians*	*Mammalian neuropeptides*	*References*
Caerulein-related			
pEQDY(SO_3)TGWMDF-NH_2	Litoria caerulea	CCK8:	Erspamer (1994)
pEEY(SO_3)TGWMDF-NH_2	Phyllomedusa sauvagei	DY(SO_3)MGWMDF-NH_2	
pENDY(SO_3)LGWMDF-NH_2	Kassina maculata	Hexagastrin:	Bowie and Tyler (2006)
pEE(OMe)DY(SO_3) TGWMDF-NH_2	Nyctimystes disrupta	Y(SO_3)GWMDF-NH_2	
DY(SO_3)LGWMDF-NH_2	Rana erythraea	Eugenin: pEQDY(SO_3)VFMHPF-NH_2	
Bombesin-related			
pEQRLGNQWAVGHLM-NH_2	Bombina bombina	GRP:	Anastasi, Erspamer and Bucci (1971)
pEQRLGNQWAVGHFM-NH_2	Bombina orientalis	GNHWAVGHLM-NH_2	Spindel (2006)
pEQSLGNQWARGHFM-NH_2	Bombina variegata		
pEGRLGTQWAVGHLM-NH_2	Alytes obstetrica		
pEQWAVGHFM-NH_2	Litoria aurea	Neuromedin B:	
SNTALRRYNQWATGHFM-NH_2	Rana pipiens	GNLWATGHFM-NH_2	
Sauvagine-related		CRF:	
QGPPISIDLSLELLRKMIEIEKQE KEKQQAANNRLLLDTI	Phyllomedusa sauvagei	AEEPPISLDLTFHLLRE VLEMARAEQIAQQA HSNRKLMDIIGK	Erspamer (1994)

Table 3 Amphibian opioid peptides

Amphibian peptides	*Amphibians*	*Mammalian neuropeptides*	*References*
Opioids			
Dermorphins	Phyllomedusa bicolor		Negri and Melchiorri (2006)
μreceptor agonists	Phyllomedusa sauvagei		
YaFGYPS-NH_2, dermorphin	Phyllomedusa rhodei	Analogs not yet identified	
YaFGYPK-NH_2, [Lys^7]dermorphin			
YaFWYPN, [Trp^4,Asn^7]dermorphin			
Deltorphins			
δ-opioid antagonists	Phyllomedusa burmeisteri		
YaFDVVG-NH_2, [D-Ala^2]deltorphin I			
YaFEVVG-NH_2 [D-Ala^2]deltorphin II			
YmFHLMD-NH_2 [D-Met^2]deltorphin			

Lower case indicates D-amino acid residues.

α-carbon of the amino acid residue rather than a racemization, which would yield an equimolar mixture of L- and D-isomers. Enzymes catalyzing the formation of D-amino acids are known so far only in yeast. From *Bombina* skin secretions, Kreil and colleagues recently purified a 52 kDa glycoprotein which catalyzes the reaction Ile-Ile-Gly to Ile-D-allo-Ile-Gly. The partial conversion of Ile to D-allo-Ile in peptide linkage proceeds without the addition of cofactors.

Amphibian Opioids and Opioid Receptors

Despite the common N-terminal tripeptide (Tyr-D-Xaa-Phe), the two families of amphibian opioids differ widely in receptor selectivity, although they bind to their own receptors with similar affinities. The N-terminal domain contains the minimum sequence essential for opioid receptor binding, whereas the C-terminal domain contains the address requisites for receptor selectivity. The presence of negatively charged amino acid residues (Asp 114 and Asp 147) that contribute to ligand binding within the putative transmembrane domains II and III of the μ-opioid receptor protein, may explain why positively charged dermorphins have high μ-opioid receptor selectivity and why amidation of their C-terminal group increases their affinity and potency. Similarly, the negatively charged

Table 4 Bv8-related peptides of frogs, fishes, lizards, and snakes

Peptides	*Primary structure*	*Species*
		Amphibians
	1----------38	
Bv8	AVITGA**C**DKDVQ**C**GSGT**CC**AASAWSRNIRF**C**IPLGNSG	Bombina variegata
Bm8a	AVITGV**C**DRDAQ**C**GSGT**CC**AASAFSRNIRF**C**VPLGNNG	Bombina maxima
Bm8b	AVITGVRDRDAQ**C**GSGT**CC**AASAFSRNIRF**C**VPLGNNG	Bombina maxima
Bm8c	AVITGV**C**DRDAQ**C**GSGT**CC**AASAFSRNVRF**C**VPLGNNG	Bombina maxima
Bm8d	AVITGV**C**DRDAQ**C**GSGT**CC**AASAFSRNIRF**C**VPLGNNG	Bombina maxima
Bm8e	AVITGV**C**DRDAQ**C**GSGT**CC**AASAFSRNIRF**C**VPLGNNG	Bombina maxima
Bm8f	AVITGV**C**DRDAQ**C**GSGT**CC**AASAFSRNIRF**C**VPLGNNG	Bombina maxima
Bo8	AVITGA**C**DRDVQ**C**GSGT**CC**AASAWSRNIRF**C**VPLGNSG	Bombina orientalis
	39----------77	
Bv8	ED**C**HPASHKVPYDGKRLSSL**C**P**C**KSGLT**C**SKSGEKFK**C**S	Bmbina variegata
Bm8a	EE**C**HPASHKVPYNGKRLSSL**C**P**C**NTGLT**C**SKSGEKFQ**C**S	Bombina maxima
Bm8b	EE**C**HPASHKVPYNGKRLSSL**C**P**C**NTGLT**C**SKSGEKYQ**C**S	Bombina maxima
Bm8c	EE**C**HPASHKVPYNGKRLSSL**C**P**C**NTGLT**C**SKSGEKFQ**C**S	Bombina maxima
Bm8d	EE**C**HPASHKVPYNGKRLSSL**C**P**C**NTGLT**C**SKSGEKSQ**C**S	Bombina maxima
Bm8e	EE**C**HPASHKVPYNGKRLSSL**C**P**C**NTGLT**C**PKSGEKFQ**C**S	Bombina maxima
Bm8f	EE**C**HPASHKVPSDGKRLSSL**C**P**C**NTGLT**C**SKSGEKYQ**C**S	Bombina maxima
Bo8	EE**C**HPASHKVPYDGKRLSSL**C**P**C**KSGLT**C**SKSGAKFK**C**S	Bombina orientalis
		Fishes
	1----------38	
Fugu-1	AVITGA**C**ERDVQ**C**GLGL**CC**AVSLWLRGLRM**C**APRGLEG	Takifugu bimaculatus
Fugu-2	AVITGA**C**EKDSQ**C**GGGM**CC**AVSLWIRSLRM**C**TPMGREG	Takifugu chinensis
	39----------77	
Fugu-1	DECHPFSHKVPYPGKRQHHT**C**P**C**LPHLV**C**TRDRDSKYR**C**	Takifugu bimaculatus
Fugu-2	DDCHPMSHTVPFFGKRLHHT**C**P**C**LPNLS**C**IPMDEGRAK**C**	Takifugu chinensis
	78----------94	
Fugu-1	TDDFKNVDLYEVGQTLR	Takifugu bimaculatus
Fugu-2	LSTYKYPDYYL	Takifugu chinensis
		Snakes
	1----------38	
MIT-1	AVITGA**C**ERDLQ**C**GKGT**CC**AVSLWIKSVRV**C**TPVGTSGE	Dendroaspis polylepsis
	39----------77	
MIT-1	D**C**HPASHKIPFSGQRKMHHT**C**P**C**APNLA**C**VQTSPKKFK**C**	Dendroaspis polylepsis
	-80	
MIT-1	LSK	Dendroaspis polylepsis
		Lizards
	1----------38	
VAR-1	AVITGA**C**DKDLQ**C**GEGM**CC**AVSLWIRSIRI**C**TPLGSSGE	Varanus varius
VAR-2	AVITGA**C**DKDLQ**C**GEGM**CC**AVSLWIRSIRICTPLGSSGE	Varanus varius
	39----------77	
VAR-1	DCHPLSHKVPFDGQRKHHTCPCLPNLVCGQTSPGKYKCL	Varanus varius
VAR-2	DCHPLSHKVPFDGQRKHHTCPCLPNLVCGQTSPGKHKCL	Varanus varius
	78---84	
VAR-1	PEFKNVF	Varanus varius
VAR-2	PEFKNVF	Varanus varius

C-terminal tetrapeptide of deltorphins enhances deltorphin selectivity for the δ-opioid receptor by electrostatic attraction to the positively charged binding site of the δ-opioid receptor (Arg 292) and electrostatic repulsion from the negatively charged μ-opioid receptor site. Electrostatic attraction within the peptide molecule may explain why deltorphins have a folded structure, whereas dermorphins have a more distended and flexible structure. Unlike the positively charged C-terminus of dermorphins, the negatively charged C-terminal tail of deltorphins comes into close contact with the positively charged N-terminal tripeptide of the molecule and folds the backbone, thus placing the Tyr^1 and Phe^3 aromatic rings in definite orientations best suited for δ-opioid receptor docking.

Dermorphin affinity for μ-opioid receptors and its opioid potency are 20 and 100 times higher respectively than those of morphine. Among other dermorphins, [Lys^7]dermorphin shows an affinity and selectivity for μ-opioid receptors 6 times higher than dermorphin and

100 times higher than morphine. Like μ-opioid agonists, dermorphins produce not only antinociception but also catalepsy, respiratory depression, constipation, tolerance, and dependence, albeit to a lower degree than morphine. Dermorphin-induced antinociception takes place at the supraspinal and spinal levels. Owing to its low central nervous system (CNS) permeability and bioavailability, the analgesic potency of dermorphin is about 250 times higher than that of morphine after intracerebroventricular (ICV) injection but comparable to that of morphine after subcutaneous (SC) injection. [Lys^7]dermorphin and some synthetic analogs bearing a hydrophilic group (a basic amino acid or a glycosyl residue) at the C-terminal end of the molecule enter the CNS in 7–10 times higher amounts than dermorphin, suggesting facilitated transport across the blood–brain barrier by a carrier or endocytosis.

Although D-Ala-deltorphins have δ-opioid binding affinity similar to D-Met-deltorphin, they consistently have the highest δ-opioid selectivity. The rank order of selectivity ($K_{i\delta}/K_{i\mu}$) is D-Ala-deltorphin-I = D-Ala-deltorphin-II $\gg$ D-Met-deltorphin. The high δ-opioid selectivity of D-Ala-deltorphins can be attributed to their C-terminal tetrapeptide sequence, in which the anionic residue plays an important role. Elimination of the charge at the fourth position normally results in opioids that have similar δ- and μ-opioid affinity and generally lack selectivity. Substitution of Gly in the fourth position permits D-Ala-deltorphins to assume a more extended conformation, dramatically increasing their μ-opioid affinity and potency. The hydrophobic qualities of the residues at the fifth and the sixth positions (Val5–Val6) are crucial in maintaining the affinity and selectivity of D-Ala-deltorphins for δ-opioid receptors, as evidenced in peptide analogs in which the aliphatic quality of the side chain was enhanced.

When D-Ala-deltorphins became available, pharmacological and biochemical studies provided evidence of distinct δ1- and δ2-opioid receptor subtypes in the rodent CNS. Studies comparing the binding properties of [^{3}H]deltorphin-I and D-Pen^2, D-Pen^{5}enkephalin (DPDPE) in rat brain synaptosomes, provided the first biochemical evidence for two δ-opioid receptor subtypes. In the mouse brain, homogenates and in rat brain slices, D-Ala-deltorphin-II binds preferentially to a population of δ-opioid receptors that develop after weaning and correspond to δ2-opioid receptor subtype. The accumulated evidence demonstrates that D-Ala-deltorphin-II exerts antinociceptive action at δ2-opioid receptor subtype, whereas D-Ala-deltorphin-I (and DPDPE) interact with greater specificity at δ1-opioid receptor subtype. Using selective δ-opioid antagonists, tolerance development, and δ-opioid receptor knockdown (antisense oligonucleotide-treated) mice and rats, numerous pharmacological studies also suggested that spinal and supraspinal antinociception produced by DPDPE/D-Ala-deltorphin-I and D-Ala-deltorphin-II in rats and mice are mediated by distinct δ-opioid receptor subtypes.

Despite pharmacological evidence of two distinct δ-opioid receptors, molecular biologists have not yet succeeded in cloning δ-opioid receptor subtypes. The available evidence implies that the relatively small change in the longer side chain on Glu, due to the methylene C–C bond, could influence the global conformation of D-Ala-deltorphin-II, exerting a role in selecting the postreceptor transduction pathway by differentially activating δ-opioid receptors.

D-Ala-deltorphin-II induces δ-opioid receptor-mediated analgesia in frogs and also in the invertebrate land snail (*Cepaea nemoralis*). When administered by intrathecal injection in rats, D-Ala-deltorphin-II produces a dose-related inhibition of nociceptive responses. Its inhibitory effect lasts 10–60 min, depending on the dose, and is naltrindole-reversible. Conversely, when injected ICV in rats, D-Ala-deltorphin-II was a weak partial agonist; only doses higher than 30 nmol produced some degree of antinociception and none of the doses elicited the maximum achievable response. Repeated injection of D-Ala-deltorphin-II induces tolerance to the antinociceptive effect. There is no cross-tolerance between antinociception induced by D-Ala-deltorphin-II and that generated either by μ- or by δ1-opioid receptor agonists. The finding that intrathecal injection of D-Ala-deltorphin-II has a higher analgesic effect than DPDPE is probably related more to the predominance of δ2-opioid receptors in the spinal cord than to a prevalence of supraspinal δ1-opioid receptors. Isobolographic analysis revealed that supraspinal/spinal antinociceptive interactions for both the δ1-opioid agonist DPDPE and the δ2-opioid agonist D-Ala-deltorphin-II were synergistic in many nociceptive tests, suggesting that compounds acting through δ-opioid receptors may have sufficient potency for eventual clinical applications.

Data suggest that the δ-opioid agonists play a predominantly modulatory role in antinociception rather than a primary role. In homozygote mice with a disrupted δ-opioid receptor gene, δ-opioid agonist-induced analgesia is reduced. Many studies showed that in mice and in rats the intensity of δ-opioid analgesia depends on coactivation of μ-opioid receptors by endogenous or exogenous opioids. Stress associated with the ICV injection may activate μ-opioid receptors through the release of endogenous opioids and thus potentiate the antinociceptive responses to δ-opioid agonists. Moreover, δ-opioid agonists can be regarded as potential drugs

for the treatment of chronic pain: in rats; intrathecal administration of D-Ala-deltorphin-II dose-dependently antagonized the cold-water allodynia which developed after sciatic nerve injury, and ICV administration of D-Ala-deltorphin-II significantly reversed the hyperalgesic response associated with peripheral inflammation.

Injections of D-Ala-deltorphin-II into the rat brain ventricles, ventral tegmental area, and nucleus accumbens invariably increase locomotor activity and induce stereotyped behavior . The ambulatory activity is intermittent and usually intercalated by rearing events. The motor activity is antagonized by the δ-opioid-selective antagonist naltrindole, and by high doses of the μ-opioid antagonist naloxone, but is unaffected by the δ1-opioid-selective antagonist naloxonazine. ICV administration of 1.3 nmol/rat of D-Ala-deltorphin-II increases social contacts in rats. Local application of D-Ala-deltorphin-II to the nucleus accumbens, but not to the nucleus caudatus, increases extracellular dopamine concentrations (by up to 120%). It also stimulates locomotor activity and stereotypes. Repeated ICV injection of D-Ala-deltorphin-II in naive rats induces tolerance to the stimulant effects, whereas repeated daily injections or continuous infusion of morphine results in sensitization to the behavioral-activating effects of the δ-opioid agonist. Deltorphin improves memory consolidation in a passive avoidance apparatus in mice; this effect is abolished by naltrindole. D-Ala-deltorphin-II causes hypothermia in cold-adapted animals. In contrast to μ-opioid agonists, D-Ala-deltorphin-I, at low doses, stimulates respiratory activity in fetal lambs, and this effect is blocked by the simultaneous administration of naltrindole.

The discovery of the amphibian opioid peptides, apart from the intriguing problem of the occurrence of their analogs in the mammalian central and peripheral nervous systems, has provided new insights into the functional role of the μ- and δ-opioid systems.

Bv8-Prokineticins

A small protein, named Bv8 to indicate its origin from the skin secretion of *Bombina variegata* and its molecular weight (8 kDa), is the first amphibian member of the Bv8-prokineticin family. Homologs of Bv8 are present in the skin secretions of other amphibians such as *Bombina bombina*, *Bombina orientalis*, *Bombina maxima*, and of lizards and fishes of *Takifugu* species (Table 4). A Bv8 homolog mamba intestinal toxin (MIT-1 or VRPA) is a component of the venom of the black mamba, *Dendroaspis polylepsis*. The striking characteristics of these proteins are their identical N-terminal sequence, AVITG, and the presence of 10 cysteines with identical spacing, which define a five-disulfide-bridged motif called a colipase fold.

The high degree of identity between amphibian Bv8 peptides, fish peptides, and mamba MIT-1 (58%) suggested that similar peptides could also be present in other species, including mammals. In the mouse, rat, cattle, monkey, and human, cDNA cloning identified orthologs of Bv8 (Table 5). The two mammalian proteins similar to Bv8 were named prokineticin 1 (PK1 or EG-VEGF) and prokineticin 2 (PK2 or mBv8). A second form of PK2 has been identified and named PK2b (because of an insert of 21 basic amino acids in its sequence). The name prokineticin refers to the ability of these peptides to contract guinea pig ileum, a property shared with amphibian Bv8.

Bv8-Related Neuropeptides

Expression patterns of rodent and human mRNAs for PK1 and PK2 (PROK1 and PROK2) have been reported in peripheral tissues (dorsal root ganglia (DRG), gastrointestinal tract, endocrine glands, spleen, and human and murine leukocytes) and the CNS.

In the neonatal and adult rat and mouse brain, both PROK1 and PROK2 are clearly expressed in the olfactory bulb. Olfactory bulb neurogenesis may depend on PK2 signaling because PROK2-null mice display a marked reduction in the size of the olfactory bulb, a loss of normal olfactory bulb architecture, and an accumulation of neuronal progenitors in the rostral migratory stream. PROK1 and PROK2 are clearly expressed also in the Calleja islands and in suprachiasmatic nucleus (SCN). The PROK2 expression pattern in the SCN of mice and rats is rhythmic, following the circadian cycle (being lowest in the dark phase), and is severely blunted in mutant mice deficient in clock or cryptochrome genes. PK2, therefore, has been indicated as a candidate SCN clock output signal that regulates circadian locomotor rhythm. The reduction of locomotor rhythms in PROK2-null mice was apparent in both hybrid and inbred genetic backgrounds. PROK2-null mice also displayed significantly reduced rhythmicity for a variety of other physiological and behavioral parameters, including sleep–wake cycle, body temperature, feeding, circulating glucocorticoid and glucose levels, and the expression of peripheral clock genes. In addition, PROK2-null mice showed accelerated acquisition of food anticipatory activity during daytime food restriction. Thus PK2, acting as a SCN output factor, is important for the maintenance of robust circadian rhythms.

Prokineticins are also expressed in the neurons of the medial preoptic area, nucleus of solitary tract, trigeminal and facial nuclei, and DRG.

Table 5 Bv8-related prokineticins of mammals

Peptides	*Primary structure*	*Species*
		Mammals
	1----------45	
mPK1	AVITGACERDIQCGAGTCCAISLWLRGLRLCTPLGREGEECHPGS	Mouse
mPK2	AVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHPLT	Mouse
mPK2b	AVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHPLT	Mouse
rPK1	AVITGACERDVQCGAGTCCAISLWLRGLRLCTPLGREGEECHPGS	Rat
rPK2	AVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHPLT	Rat
rPK2b	AVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHPLT	
bPK1	AVITGACERDVQCRAGTCCAVSLWLRGLRVCTPLGRAGEECHPGS	Cattle
bPK2	AVITGACDRDPQCGGGMCCAVSLWVKSIRICTPMGKVGDSCHPMT	Cattle
bPK2b	AVITGACDRDPQCGGGMCCAVSLWVKSIRICTPMGKVGDSCHPMT	Cattle
mkPK2	AVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGKLGDSCHPLT	Monkey
mkPK2b	AVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGKLGDSCHPLT	Monkey
hPK1	AVITGACERDVQCGAGTCCAISLWLRGLRMCTPLGREGEECHPGS	Man
hPK2	AVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGKLGDSCHPLT	Man
hPK2b	AVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGKLGDSCHPLT	Man
	46----------86	
mPK1	HKIPFLRKRQHHTCPCSPSLLCSRFPDGRYRCFRDLKNANF	Mouse
mPK2	RKVPFWGRRMHHTCPCLPGLACLRTSFNRFICLARK	Mouse
mPK2b	RKSHVANGRQERRRAKRRKRKKEVPFWGRRMHHTCPCLPGLACLRTSFNRFICLARK	Mouse
rPK1	HKIPFFRKRQHHTCPCSPSLLCSRFPDGRYRCSQDLKNVNF	Rat
rPK2	RKVPFWGRRMHHTCPCLPGLACLRTSFNRFICLARK	Rat
rPK2b	RKSHVANGRQERRRAKRRKRKKEVPFWGRRMHHTCPCLPGLACLRTSFNRFICLARK	Rat
bPK1	HKVPFFRKRQHHACPCLPNLLCSRGLDGRYRCSTNLKNINF	Cattle
bPK2	RKVPFLGRRMHHTCPCLPGLACSRTSFNRYTCLAQK	Cattle
bPK2b	RKNHFGNGRQERRKRKRRRKKKVPFLGRRMHHTCPCLPGLACSRTSFNRYTCLAQK	Cattle
mkPK2	RKVPFLGRRMHHTCPCLPGLACLRTSFNRFICLARK	Monkey
mkPK2b	RKNNFGNGRQERRKRKRRKRKKEVPFGRRMHHTCPCLPGLACLRTSFNRFICLARK	Monkey
hPK1	HKIPFFRKRKHHTCPCLPNLLCSRFPDGRYRCSMDLKNINF	Man
hPK2	RKVPFFGRRMHHTCPCLPGLACLRTSFNRFICLAQK	Man
hPK2b	RKNNFGNGRQERRKRKRSKRKKEVPFFGRRMHHTCPCLPGLACLRTSFNRFICLAQK	Man

Bv8-Prokineticin Receptors

The two G-protein-coupled receptors for Bv8-PKs, prokineticin receptor 1 (PKR1) and prokineticin receptor 2 (PKR2) have an overall identity in their amino acid sequences of 85%, with most differences at the N-terminal, and are approximately 80% identical to the previously described mouse orphan receptor gpr73. In specific endothelial cells, neurons, and transfected cells expressing these receptors, Bv8-induced PKR activation stimulates Ca^{2+} mobilization and phosphoinositol turnover, indicating a $G_{q/o}$-protein coupling. Receptor binding studies showed that PKR2 is a MIT-preferring receptor; indeed, the affinity of MIT for PKR2 (in the picomolar range) is approximately 10 times higher than that of PK2 and 50 times higher than that of PK1. PKR1 is a MIT- and PK2-preferring receptor; the affinity of MIT for PKR1 (~5–10 times lower than for PKR2) is comparable to that of PK2 and 60 times higher than that of PK1. The affinity of Bv8 for the receptors is comparable to that of PK2 and is approximately 40 times higher than that of PK1.

In rat embryos from day 12, both receptors are highly expressed in the neuroepithelium lining ventricles, the olfactory bulb, the Gasser-ganglion, and DRG. One day after birth, receptors PKR2 are still expressed at high levels in the olfactory bulb, neuroepithelium lining the ventricles, striatum, hippocampus, thalamic and hypothalamic paraventricular nuclei, SCN, amygdala, and cortex, whereas receptors PKR1 are only found in the cortex. In adult rats, only PKR2 is abundantly or moderately expressed in several discrete brain regions. The presence of PKR2 mRNA and PKR2 protein in the nucleus arcuatus explains the anorexogenic effect of Bv8-prokineticins. The dipsogenic effect of Bv8 depends on its binding to PKR2 in the subfornical organ (SFO). Central and peripheral administration of Bv8 in rats induces growth hormone-releasing hormone (GRH), oxitocin, and vasopressin release and vasopressin-dependent antidiuresis, probably by acting on PKR2 in the paraventricular hypothalamic nucleus. Stimulation of PKR2 in the SCN by Bv8-prokineticins induces sleep in rodents.

Human Kallmann syndrome (KS), which combines anosmia, related to defective olfactory bulb morphogenesis, with hypogonadism due to gonadotropin-releasing hormone deficiency, appears to be related to mutations in the genes coding for PK2 and PKR2. In a

cohort of 192 patients affected by KS, 10 different point mutations were identified in the genes encoding PKR2 and four were identified in the genes encoding PK2. The mutations in PK2 were detected in the heterozygous state, whereas PKR2 mutations were found in the heterozygous, homozygous, or compound heterozygous state. In addition, one of the patients heterozygous for a PKR2 mutation also carried a missense mutation in KAL1, thus indicating a possible digenic inheritance of the disease in this individual. These findings reveal that insufficient prokineticin-signaling through PKR2 leads to the abnormal development of the olfactory system and reproductive axis also in humans.

PKR1 and PKR2 mRNAs are expressed in the DRG of neonatal and adult rats. PKR1 is mainly expressed in small and medium-size neurons and PKR2 in large neurons. PKR proteins are present in the DRG, in the outer layers of the dorsal horns of the spinal cord, and in the peripheral terminals of nociceptor axons. Activation of nociceptor PKRs by Bv8 in rats and mice produces nociceptive sensitization to thermal and mechanical stimuli, without inducing any spontaneous, overt nocifensive behavior or local inflammation. A physiological role of Bv8-prokineticins as peripheral and central pain modulators is supported by the observation that mice lacking the PKRs or PK2 are less sensitive to noxious heat (hot-plate test) than are wild-type mice. PKR1-null mice also exhibited impaired development of hyperalgesia after tissue injury. The inflammatory agents CFA and mustard oil produced comparable paw edema in wild-type and PKR1-null mice, but induced inflammatory nociceptive sensitization to heat and pressure significantly lower in PKR1-knockout than in wild-type mice, demonstrating a role of prokineticins and their receptors in inflammatory pain.

The molecular mechanism of the prokineticin-induced hyperalgesia has been investigated in primary cultures of DRG neurons. Of the neurons responding to Bv8, 90% also responded to capsaicin, showing a very high degree of co-localization of functional PKRs with the heat and capsaicin-activated transient receptor potential cation channel subfamily (TRPV1). Half of the Bv8-responding neurons also express calcitonin gene-related peptide (CGRP) and substance P (SP) and release these neuropeptides on exposure to Bv8-prokineticins. Patch-clamp experiments showed that a brief exposure to Bv8 tremendously potentiated the inward current activated by capsaicin in DRG neurons, an effect that is lacking in neurons from PKR1-knockout mice. Bv8 causes translocation of protein kinase C (PKCε) to the neuronal membrane of nociceptors, and Bv8-activated PKCε sensitizes TRPV1 to activating agents by phosphorylating two serine residues of the channel. Indeed, the Bv8-induced capsaicin potentiation is partially inhibited by the PKC inhibitors staurosporine and RO 31–8220. These data indicate that Bv8-induced hyperalgesia results, at least in part, from increased sensitivity of TRPV1 to heat and acid and from the expression and release of excitatory transmitters (CGRP and SP) at the spinal dorsal horn.

In situ mRNA hybridization experiments demonstrated that PK2 is expressed by inflammatory cells, predominantly neutrophils, in the human inflamed tonsil and appendix and in the rat inflamed paw. PK2 released by inflammatory cells can bind and activate PKRs on the primary sensitive neurons contributing to inflammatory pain. Neutrophil extracts, fractionated using ionic exchange chromatography, gel filtration, and reverse phase (RP) chromatography, displayed Bv8-like activity, displaced ^{125}I-MIT binding from PKR1-transfected Chinese hamster ovary (CHO) cell membranes and produced the Bv8-characteristic hyperalgesia when injected intrathecally in rats.

Prokineticin receptors are potential targets for drugs which block the nociceptive information before it reaches the brain. Identifying of the structural determinants required for receptor binding and hyperalgesic activity of Bv8-prokineticins is thus mandatory for the design of PKR antagonists. The highly conserved N-terminal sequence AVITGA and the triptophan residue in position 24 in all members of the Bv8/PK family are required for biological activity; deletions and substitutions in these conserved residues produce antagonist molecules. Preclinical studies are currently being carried out on these mutated PK peptides as PKR antagonists endowed with analgesic properties.

See also: Invertebrate Neurohormone GPCRs; Mammalian Neuropeptide Families; Neuropeptide Signaling in Invertebrates; Neuropeptides: Pain; Opioid Peptides and Receptors.

Further Reading

Erspamer V (1994) Bioactive secretions of the amphibian integument. In: Heatwole H and Barthalmus T (eds.) *Amphibian Biology, Vol. 1: The Integument*, pp. 178–350. Chipping Norton, NSW: Surrey Beatty & Sons.

Hu WP, Zhang C, Li JD, et al. (2006) Impaired pain sensation in mice lacking prokineticin 2. *Molecular Pain* 15: 35–40 [online].

Li JD, Hu WP, Boehmer L, et al. (2006) Attenuated circadian rhythms in mice lacking the prokineticin 2 gene. *Journal of Neuroscience* 26: 11615–11623.

Matsumoto S, Yamazaki C, Masumoto KH, et al. (2006) Abnormal development of the olfactory bulb and reproductive system in mice lacking prokineticin receptor PKR2. *Proceedings of the National Academy of Sciences of the United States of America* 103: 4140–4145.

Negri L, Lattanzi R, Giannini E, et al. (2002) Nociceptive sensitization by the secretory protein Bv8. *British Journal of Pharmacology* 137: 1147–1154.

Negri L, Lattanzi R, Giannini E, et al. (2006) Impaired nociception and inflamed pain sensation in mice lacking the prokineticin receptor PKR1: Focus on interaction between PKR1 and the capsaicin receptor TRPV1 in pain behavior. *Journal of Neuroscience* 26: 6716–6727.

Negri L, Lattanzi R, Giannini E, and Melchiorri P (2006) Modulators of pain: Bv8 and prokineticins. *Current Neuropharmacology* 4: 207–215.

Negri L and Melchiorri P (2006) Opioid peptides from frog skin and Bv8-related peptides. In: Kastin AJ (ed.) *Handbook of Biologically Active Peptides*, pp. 269–275. Boston: Academic Press.

Vellani V, Colucci M, Lattanzi R, et al. (2006) Sensitization of transient receptor potential vanilloid 1 by the prokineticin receptor agonist Bv8. *Journal of Neuroscience* 26: 5109–5116.

Amygdala: Contributions to Fear

S Maren, University of Michigan, Ann Arbor, MI, USA

© 2009 Elsevier Ltd. All rights reserved.

History

The physiological basis for emotions has intrigued mankind for centuries. Early Greek philosophers held emotion to be a product of visceral activity – a beating heart reflecting passion and love or a clenched gut fueling jealousy and sadness. Over the course of hundreds of years, however, the study of emotion has moved from the trunk to the head. There are many factors that encouraged this shift in thinking about the localization of emotion. Perhaps the most influential one was the discovery of profound emotional changes in a man named Phineas Gage, whose brain was penetrated by an iron spike in a railway construction accident in 1848. Amazingly, Gage survived the injury to his head, but he was not the same person after recovering from the wound. Relative to his gentle demeanor prior to the accident, he angered easily, used profanity, and could not maintain interpersonal relationships after his accident. Emotions, it would appear, had their roots in the brain.

Not long after Gage's condition was described, scientists were embarking on the first neurological studies of brain function in animals. Sanger Brown and E A Shafer at the University College, London, carried one of the first studies of this kind out in 1888. They systematically examined the behavioral consequences of surgical lesions to the monkey cerebral cortex, with a focus on damage to the temporal and occipital lobes of the brain. In one case, they produced a nearly complete and bilateral lesion of temporal cortex that also included underlying structures, including the amygdala (Figure 1). The animal in this case presented with profound behavioral changes. Although its sensory and motor function and general intellectual faculties appeared to be normal, the animal's emotional behavior changed considerably after the lesion. The monkey lost fear of other, dominant monkeys and was tame in the presence of people. It also did not remember an aversive event, readily returning to interact with a monkey that had defeated it in fight, for example.

This observation of emotional changes after temporal cortical damage languished for 50 years until it was rediscovered by Kluver and Bucy in 1937. They similarly observed altered emotional behavior, particularly a loss of fear, in monkeys with temporal cortical lesions. Lawrence Weiskrantz made a key advance in understanding the basis for fear reductions with these lesions in 1956. He showed that the damage to the amygdala, a brain structure underlying the temporal cortex, accounted for the fear loss, and that other changes in behavior were likely the consequence of the cortical damage. Weiskrantz was also the first to suggest that the amygdala has an important role in assigning emotional value (through learning) to stimuli that predict biologically important events (including noxious events). Hence, this seminal work pinpointed a specific brain area that was critical for the genesis of an emotion – in this case, fear. To this day, the amygdala remains a focus for studying the brain mechanisms of fear.

Anatomy of Fear

Fear is a universal emotion that not only has an obvious subjective component ('I feel afraid'), but also a unique behavioral and autonomic signature ('When I am afraid, I freeze in my tracks and my heart pounds in my chest'). The physiological and behavioral aspects of fear are readily studied in both humans and animals and these responses appear to share many similarities, among mammals at least.

Fear is either innate or learned. For instance, primates appear to have an innate predisposition to learn fear of snakes. Monkeys that are born and raised in captivity exhibit fear responses to live or rubber snakes, but not to belts or toy fire trucks (none of which they have seen before). This fear response depends on observing another monkey exhibit fear in the presence of snakes. Laboratory-born deer mice can differentiate predators and nonpredators from their typical habitat, despite never having contacted those animals before. This indicates that fear systems evolved as a defensive behavior system to ward off predators found in an animal's habitat, as well accommodating rapid learning about new threats in the environment. The utility of both types of fear is obvious. Animals that can respond to threats in the environment without requiring past experience with the threat are at an obvious advantage. And rapid learning about novel threats in the environment is similarly adaptive; an animal may only have one opportunity to learn about a threat before it is mortally injured (Figure 2).

Several lines of evidence indicate that the amygdala is involved in both innate and learned fear responses. The amygdala ('almond,' Greek) is an oval-shaped collection of neurons located deep within the temporal lobe of the vertebrate brain. It is prominent in

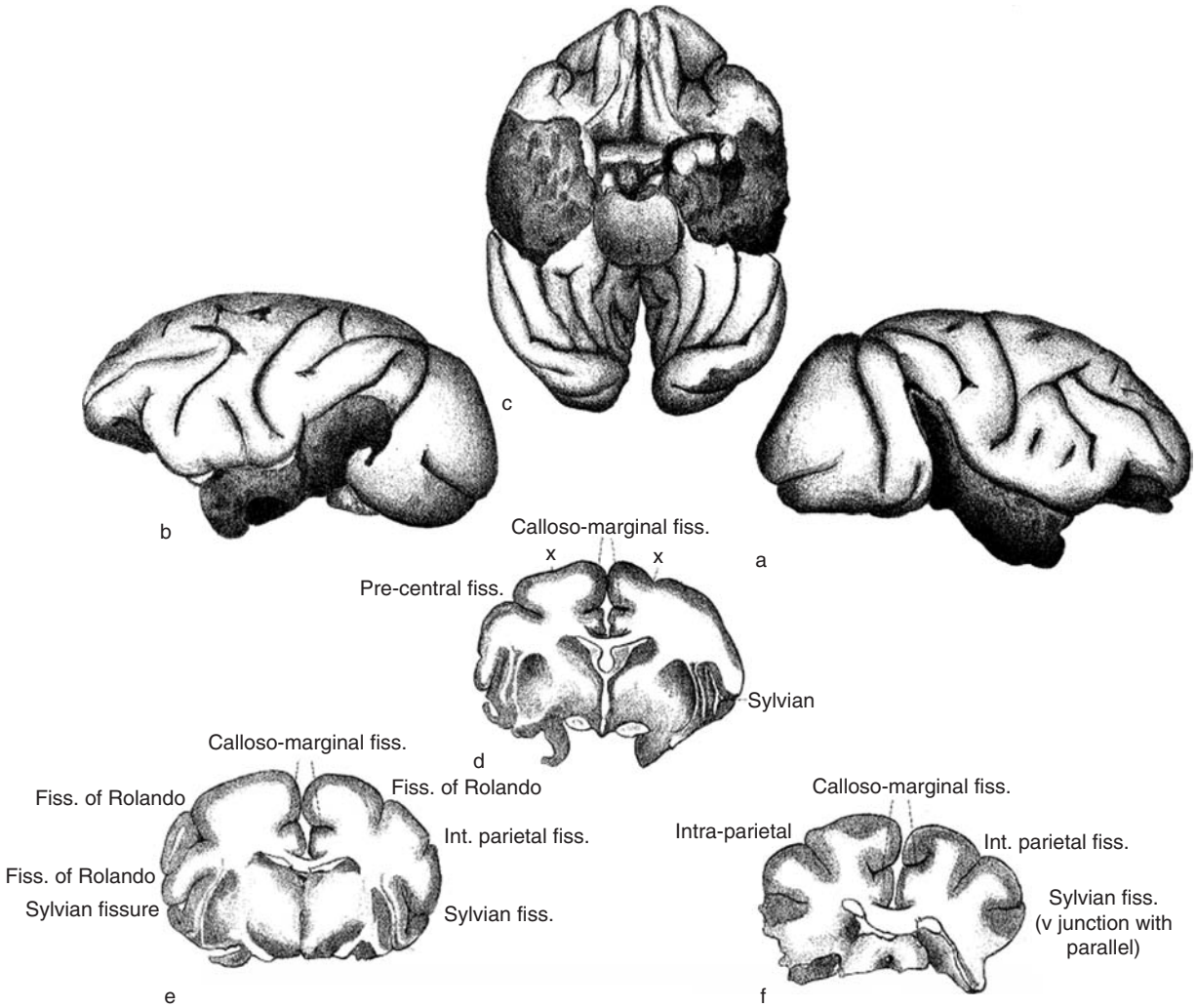

Figure 1 Temporal lobe lesions in a rhesus monkey. Reproduced from Brown S and Schäfer EA (1888) An investigation into the functions of the occipital and temporal lobes of the monkey's brain. *Philosophical Transactions of the Royal Society London, Series B* B179: 303–327, with permission.

Figure 2 A cat stalking and killing a rat. Reproduced from Caldecott R (1878) *The House that Jack Built.* London: Frederick Warne & Co. Ltd. Permission obtained under the Project Gutenberg License.

mammals, but there is also a homologous structure in amphibians and reptiles, among others. In all cases, the amygdala is not a monolithic structure in the brain; it is instead a collection of anatomically and functionally distinct nuclei with different developmental origins. It is therefore commonplace to describe the amygdala as the 'amygdaloid complex,' and to focus on particular nuclei with the amygdala.

In any case, the terms amygdala and amygdaloid complex are synonymous.

Of the ten or so amygdaloid nuclei, there are two major contributors to the fear response: the basolateral complex (including the lateral, basolateral, and basomedial nuclei) and the central nucleus (Figure 3). The basolateral complex receives considerable sensory information from subcortical and cortical areas. It in turn projects heavily to the central nucleus, which sends axons to other brain areas involved in generating specific fear responses, such as freezing (immobility), tachycardia, and stress hormone release. Damage to either the basolateral complex or the central nucleus can impair fear responses under a variety of conditions.

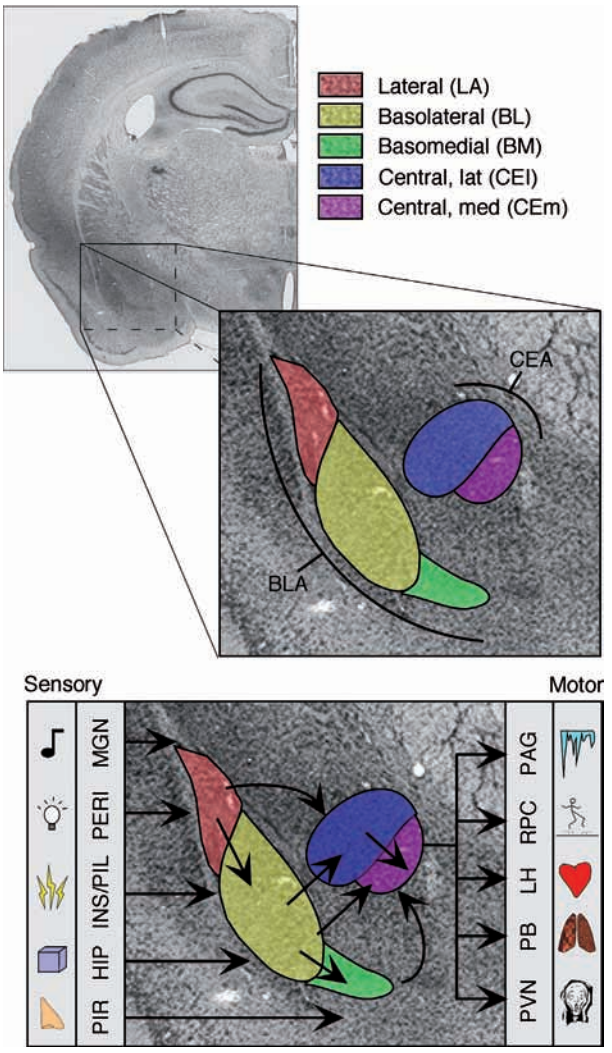

Figure 3 Anatomy of the amygdala. The amygdaloid nuclei can be roughly divided into two subsystems. These include the lateral, basolateral, and basomedial nuclei, which together form the basolateral complex (BLA) and the central nucleus (CEA). The BLA receives and integrates sensory information from a variety of sources. These include the medial and ventral divisions of the thalamic medial geniculate nucleus (MGN, auditory), the perirhinal cortex (PERI, visual), the insular cortex (INS, gustatory and somatosensory), the thalamic posterior intralaminar nucleus (PIL, somatosensory), the hippocampal formation (HIP, spatial and contextual), and the piriform cortex (PIR, olfactory). Intra-amygdaloid circuitry conveys the conditioned stimulus–unconditioned stimulus association to the CEA, where divergent projections to the hypothalamus and brain stem mediate fear responses such as freezing (periaqueductal gray, PAG), potentiated acoustic startle (nucleus reticularis pontis caudalis, RPC), increased heart rate and blood pressure (lateral hypothalamus, LH), increased respiration (parabrachial nucleus, PB), and glucocorticoid release (paraventricular nucleus of the hypothalamus, PVN). For simplicity, all projections are drawn as unidirectional connections, although in many cases these connections are reciprocal.

Innate Fear

Caroline and Robert Blanchard at the University of Hawaii first demonstrated that rats with amygdala lesions lose their innate fear of cats. Normal rats exhibited fear, particularly freezing behavior, in the presence of a cat. This response was markedly reduced by damage to the amygdala. However, the lesions, like those used in the early monkey studies, damaged both neurons in the amygdala, as well as axons passing through the amygdala from one brain area to another. More selective lesions of the amygdala, or reversible inactivation of amygdala neurons, suggest that the amygdala may not be as important in generating innate fear responses as once believed. For example, neurotoxic lesions (that spare axons) or reversible inactivation of neurons in the amygdala do not prevent fear (freezing behavior, in this case) to a predator odor (trimethylthiazoline, a component of fox feces). In primates, neurotoxic lesions of the amygdala do reduce fear responses to real or artificial snakes, but these effects are not nearly as pronounced as those produced by lesions that also damage axons. Similarly, humans with amygdala damage exhibit normal autonomic responses, such as a galvanic skin response, to aversive stimuli. Another brain area that may be involved in innate fear responses is the hypothalamus. Rosen and colleagues have shown that molecular markers for neuronal activity are increased in the rat hypothalamus, but not the amygdala, after exposure to a predator (a cat).

In sum, large lesions that encompass both neurons and axons passing through the amygdala produce profound disturbances in innate fear reactions, although more selective manipulations of amygdaloid neurons within discrete nuclei produce only mild impairments (if any) in innate fear responses. Indeed, rats with amygdala lesions exhibit normal fear reactions in a number of tests of fear and anxiety, such as the open field (which measures locomotor activity in a brightly lit novel environment) or the elevated-plus maze (in which animals avoid well-lit arms that are elevated above the floor in favor of dark and enclosed arms).

Learned Fear

As John Watson showed in the early 1920s, fear is readily learned. He conditioned the infant 'Little Albert' to fear white rats by pairing a noxious, loud noise each time he presented the boy with the rat. Albert, who had no fear of the rat before Watson's intervention, quickly came to be upset and scared by the animal, crying and reeling from it when Watson placed it in his lap (Figure 4). In this case, Albert learned to associate, through classical or Pavlovian conditioning, the white rat and loud noise. Not surprisingly, fear conditioning has become a standard model for understanding how fear is learned in animals and humans.

In the laboratory, fear conditioning proceeds by exposing animals to an innocuous stimulus (conditioned stimulus; CS) that is followed by a noxious event (unconditioned stimulus; US). After as little as a single conditioning trial, animals come to exhibit a learned fear response (conditional response; CR) to the CS. This fear response includes autonomic and behavioral components, and is easily quantified.

An enormous body of work indicates that the amygdala is critical for fear conditioning. In rats, lesions of either the basolateral complex or the central nucleus of the amygdala severely impair both learning and remembering conditional fear responses. Amazingly, lesions of the amygdala disrupt very old fear memories, up to 1-year-old memories in rats. The involvement of these amygdaloid nuclei in fear learning holds for many different stimuli (e.g., tones, lights, places, and odors) and responses (e.g., freezing, heart rate, analgesia, and acoustic startle). And like rats, humans with amygdala lesions (from either surgical interventions for epilepsy or from an amygdala-damaging genetic condition) show deficits in fear conditioning.

The inability of animals and humans with amygdala damage to learn new fears or remember old fears is not simply a problem with expressing fear. As mentioned previously, many fear responses survive amygdala damage. In fact, deficits in learning fear responses can be reproduced with temporary inactivation of the amygdala during the conditioning experience. And interfering with cellular events in the amygdala that are involved in synaptic plasticity shortly after a conditioning experience can impair memory formation even if the amygdala is functioning normally at the time the aversive event is experienced.

The critical region of the amygdala necessary for associating the CS and the US during fear conditioning is under debate. Considerable data suggest that CS and US convergence occurs in the basolateral complex (particularly the lateral nucleus), suggesting that it is where learning occurs in the amygdala (Figure 5). However, sensory information also reaches the central amygdala, and recent work suggests that it may be more important than previously recognized for learning new fears. Indeed, the two structures might work in parallel to mediate different aspects of what is learned during an aversive event.

Given the importance of the amygdala in learning and remembering fearful experiences, it is not surprising that the amygdala becomes active during fear conditioning. In humans, this has been revealed using functional magnetic resonance imaging. Blood flow in the amygdala increases when people view fearful facial expressions or when they are presented with conditioned stimuli that predict aversive unconditioned stimuli. In rats, amygdala activation is observed in the activity of single neurons. In both the basolateral complex and central nucleus, amygdala neurons increase their activity (fire more action potentials) to a newly learned fear stimulus. This activity is closely tied to what the animal has learned about the fear stimulus, and is not simply related to a state of fear itself.

The identification of the amygdala as a key brain area for emotional learning has allowed an unprecedented analysis of the molecular events involved in fear memory formation. Learning is believed to alter synaptic communication in the amygdala, possibly by inducing phenomena such as long-term potentiation, a long-lasting enhancement of synaptic transmission. In line with this possibility, drugs that prevent

Figure 4 John B Watson conditions fear in 'Little Albert.' Reproduced from Watson JB (1920) *Experimental Investigation of Babies*. Chicago: Stoelting, with permission.

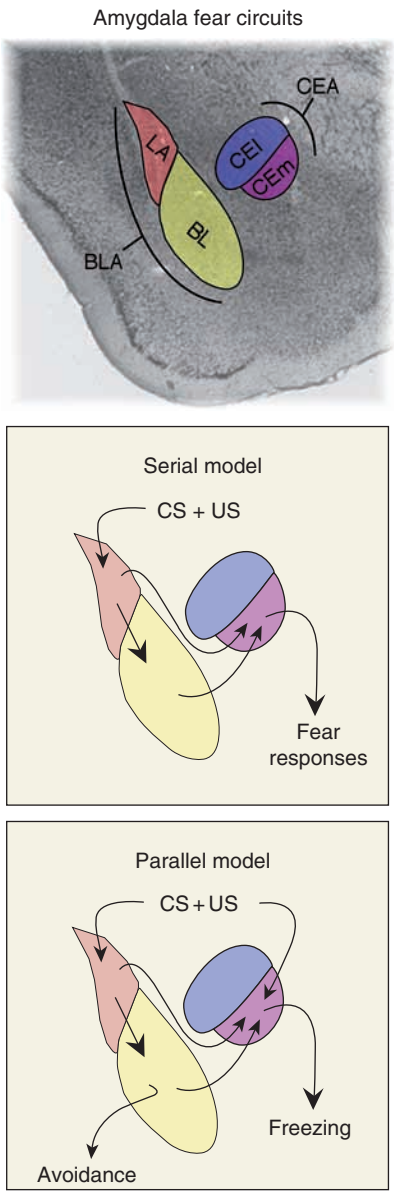

Figure 5 Different models for information processing in the amygdala during fear conditioning. BLA, basolateral complex; LA, lateral nucleus; BL, basolateral nucleus; CEA, central amygdala; CEl, central lateral nucleus; CEm, central medial nucleus; US, unconditioned stimulus; US, unconditioned stimulus. Reprinted from *Neuron*, Vol. 47, Maren S, Synaptic mechanisms of associative memory in the amygdala. 783–786, Copyright 2005, with permission from Elsevier.

long-term potentiation also impede fear learning when they are injected into the amygdala. Hence, fear memories appear to be represented as long-lasting synaptic changes among amygdala neurons.

Pathological Fear

The ability to engage the fear system and associated defensive behaviors after a threat is clearly adaptive. However, fear can also be pathological under some conditions, and excessive fear is at the root of many disorders of fear and anxiety, including panic disorder, posttraumatic stress disorder, and specific phobias, to name a few. In posttraumatic stress disorder, for example, a severe trauma yields a persistent and intrusive state of fear that has debilitating effects on the affected individual.

Pathological fear in posttraumatic stress disorder and other anxiety disorders is often associated with elevated activity in the amygdala. There is some evidence that elevated amygdala activity is the consequence of a loss of cortical inhibition. For example, prefrontal cortical activity is often inversely correlated with amygdala activity in patients with anxiety disorders.

Clinical interventions for anxiety disorders are varied, but many involve behavioral procedures, including exposure therapy to reduce fear in a controlled setting. In its essence, exposure therapy is an extinction procedure in which a fear stimulus is presented without its aversive consequence, in a safe environment. After fear conditioning, CRs can be extinguished by presenting the CS without the US for several trials. Importantly, there is considerable evidence that this procedure does not erase the fear memory, but rather results in new learning (e.g., CS is safe) that inhibits the fear memory. The persistence of the fear memory after extinction is demonstrated by the renewal of the fear when a CS is presented outside of the environment where extinction was performed, and by the spontaneous recovery of fear responses as time passes after the end of extinction training.

The new learning that accompanies extinction also requires the amygdala. Pharmacological manipulations in the amygdala that impair fear conditioning also impair extinction learning. Moreover, drugs that enhance the activity of amygdala neurons can facilitate extinction under some conditions. This may be an important complement to behavioral interventions for treating anxiety disorders.

In addition to the amygdala, there are other brain areas implicated in the suppression of pathological fear. Another medial temporal lobe structure, the hippocampus, is important for learning and remembering the places in which fear and extinction memories are established. The prefrontal cortex exhibits inhibitory control over the amygdala and may be involved in promoting extinction memory when there is conflict between safety and danger signals. It is clear that a wide network of neurons is engaged in both learning new fears and suppression of old ones.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Structure and Circuitry in Rodents and Felines; Aversive Emotions: Molecular Basis of

Unconditioned Fear; Emotion: Neuroimaging; Extinction: Anatomy; Genetics of Human Anxiety and Its Disorders; Panic Disorder as an Emotional Disorder; Pharmacology of Fear Extinction; Phobia and Human Evolution; Prefrontal Cortex: Structure and Anatomy.

Further Reading

Adolphs R, Tranel D, Damasio H, et al. (1995) Fear and the human amygdala. *Journal of Neuroscience* 15: 5879–5891.

Blanchard DC and Blanchard RJ (1972) Innate and conditioned reactions to threat in rats with amygdaloid lesions. *Journal of Comparative and Physiological Psychology* 81: 281–290.

Brown S and Schäfer EA (1888) An investigation into the functions of the occipital and temporal lobes of the monkey's brain. *Philosophical Transactions of the Royal Society London, Series B* B179: 303–327.

Caldecott R (1878) *The House that Jack Built*. London: Frederick Warne & Co.

Cardinal RN, Parkinson JA, Hall J, et al. (2002) Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience & Biobehavioral Reviews* 26: 321–352.

Choi JS and Brown TH (2003) Central amygdala lesions block ultrasonic vocalization and freezing as conditional but not unconditional responses. *Journal of Neuroscience* 23: 8713–8721.

Davis M (1992) The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience* 15: 353–375.

Helmstetter FJ (1992) The amygdala is essential for the expression of conditional hypoalgesia. *Behavioral Neuroscience* 106: 518–528.

Kluver H and Bucy PC (1937) 'Psychic blindness' and other symptoms following bilateral temporal lobectomy in rhesus monkeys. *American Journal of Physiology* 119: 352–353.

LaBar KS and Cabeza R (2006) Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience* 7: 54–64.

LeDoux JE (2000) Emotion circuits in the brain. *Annual Review of Neuroscience* 23: 155–184.

Maren S (2001) Neurobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience* 24: 897–931.

Maren S (2005) Synaptic mechanisms of associative memory in the amygdala. *Neuron* 47: 783–786.

Maren S and Quirk GJ (2004) Neuronal signalling of fear memory. *Nature Reviews Neuroscience* 5: 844–852.

Pare D and Collins DR (2000) Neuronal correlates of fear in the lateral amygdala: Multiple extracellular recordings in conscious cats. *Journal of Neuroscience* 20: 2701–2710.

Rosen JB, Adamec RE, and Thompson BL (2005) Expression of egr-1 (zif268) mRNA in select fear-related brain regions following exposure to a predator. *Behavioural Brain Research* 162: 279–288.

Swanson LW and Petrovich GD (1998) What is the amygdala? *Trends in Neurosciences* 21: 323–331.

Wallace KJ and Rosen JB (2001) Neurotoxic lesions of the lateral nucleus of the amygdala decrease conditioned fear but not unconditioned fear of a predator odor: Comparison with electrolytic lesions. *Journal of Neuroscience* 21: 3619–3627.

Watson JB (1920) *Experimental Investigation of Babies*. Chicago: Stoelting.

Watson JB and Rayner R (1920) Conditioned emotional reactions. *Journal of Experimental Psychology* 3: 1–14.

Weiskrantz L (1956) Behavioral changes associated with ablation of the amygdaloid complex in monkeys. *Journal of Comparative and Physiological Psychology* 49: 381–391.

Amygdala: Structure and Circuitry in Primates

R Kelly and L Stefanacci, Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The primate amygdala comprises a group of cytoarchitecturally, histochemically, and connectionally heterogeneous nuclei in the medial temporal lobes. Clinically, impairments in amygdala function have been implicated in social, emotional, and perceptual disorders such as autism, depression, and schizophrenia. Numerous experimental studies support the hypothesis that the amygdala is critical for forming associations between environmental stimuli and their motivational and emotional significance. The anatomical and physiological substrates of the amygdala's role in learning have been most thoroughly described in fear conditioning paradigms, in which an animal develops a conditioned fear response to a previously neutral stimulus; studies in human and nonhuman primates have also demonstrated that the amygdala is important for forming positive associations. The amygdala's widespread interconnections with cortical and subcortical sites provide the substrate for its role in learning and behavior. Here we summarize these interconnections and suggest their functional significance.

Structure

The individual nuclei of the primate amygdala are distinguished by their histological profiles (Figure 1). The deep nuclei include the lateral, basal, accessory basal, and paralaminar nuclei. The superficial nuclei surround the deep nuclei and, for the most part, are situated on the medial aspect of the amygdala. They include the anterior and posterior cortical nuclei, the periamygdaloid cortex, the nucleus of the lateral olfactory tract, the medial nucleus, and the central nucleus. Other nuclei include the anterior amygdaloid area, which forms a transition between the amygdala and the substantia innominata; the amygdala–hippocampal area, which is situated between the amygdala and the hippocampal formation; and the intercalated nuclei, which are heterogeneous groups of cells that are embedded in white matter within the amygdala. Individual amygdala nuclei have been further partitioned based on immunohistochemical patterns, cell size, and cell packing density (Figure 1).

Intrinsic Connections

The lateral, basal, and accessory basal nuclei are the primary targets of cortical and subcortical projections to the amygdala. From these nuclei, information flows in a general medial and dorsal direction through the amygdala (Figure 2). A substantial portion of output from the lateral, basal, and accessory basal nuclei is directed to the central nucleus. The central nucleus, in turn, projects to regions of basal forebrain, hypothalamus, and brain stem that control autonomic and somatomotor responses. While the central nucleus is the primary source of amygdala output to subcortical sites, the basal and accessory basal nuclei are the main sources of amygdala output to cortical sites.

Cortical Connections

Sensory Cortex

All sensory systems project to the amygdala. The olfactory system provides the only direct primary sensory input to the amygdala. Olfactory inputs are directed to the anterior cortical nucleus, the periamygdaloid cortex, and the deep nuclei. Other sensory input originates from modality-specific association cortices and is directed primarily to the lateral nucleus. The bulk of unimodal sensory input to the amygdala arises from visual processing areas (Von Bonin and Bailey's areas TEO and TE) in the temporal lobe. Auditory information is relayed through auditory association cortex (area TA) in the superior temporal gyrus, and gustatory input and somatosensory information are relayed through anterior (agranular and dysgranular) and posterior (granular) insular cortices. Progressively higher level processing areas in visual and auditory association cortices originate more robust projections that terminate in a more widespread manner in the amygdala. This suggests that highly processed sensory information exerts a relatively greater influence in the amygdala. Sensory afferents terminate in a topographically organized manner in the lateral nucleus. Fibers relaying visual information terminate dorsally and rostrally, those relaying auditory information terminate ventrally and caudally, and fibers relaying somatosensory information terminate in a medial position.

The amygdala projects to all of the sensory cortical areas that innervate it. The projections arise primarily in the basal nucleus. Thus, sensory information accesses the amygdala by way of the lateral nucleus, and the return projections originate in the basal nucleus. This suggests that sensory information undergoes at least one stage of processing within the amygdala before returning to sensory cortex.

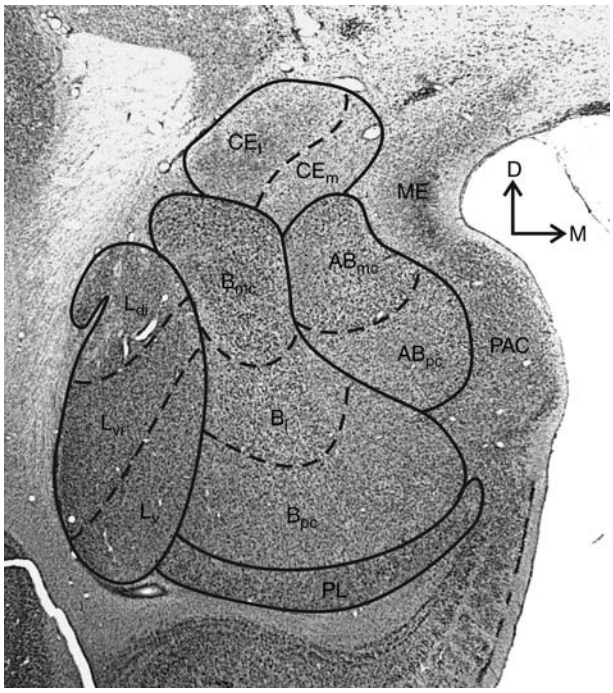

Figure 1 Nissl-stained coronal section through the midrostral–caudal level of the macaque amygdala. AB_{mc}, accessory basal nucleus, magnocellular division; AB_{pc}, accessory basal nucleus, parvicellular division; B_{mc}, basal nucleus, magnocellular division; B_i, basal nucleus, intermediate division; B_{pc}, basal nucleus, parvicellular division; CE_l, central nucleus, lateral division; CE_m, central nucleus, medial division; L_{di}, lateral nucleus, dorsal intermediate division; L_{vi}, lateral nucleus ventral intermediate division; L_v, lateral nucleus, ventral division; ME, medial nucleus; PAC, periamygdaloid cortex; PL, paralaminar nucleus; D, dorsal; M, medial.

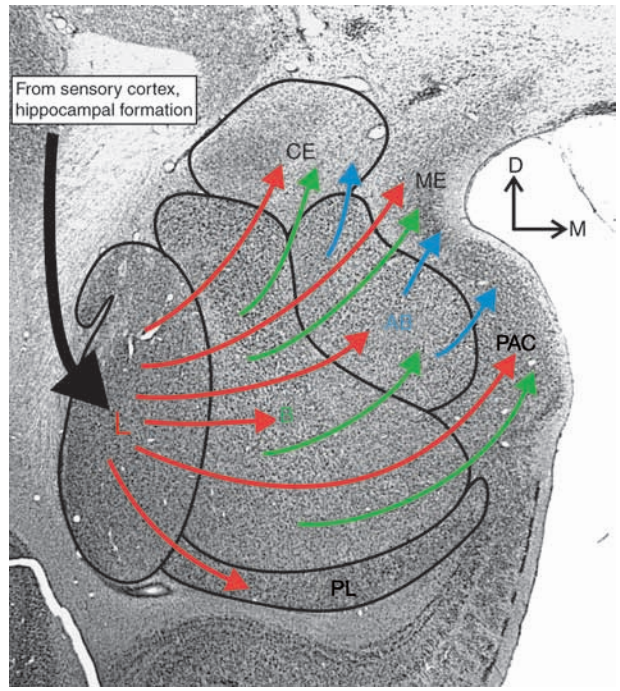

Figure 2 Nissl-stained coronal section through the midrostral–caudal level of the macaque amygdala. Arrows demonstrate the general lateral-to-medial flow of information from the lateral (red), basal (green), and accessory basal (blue) nuclei. The precise topographic organization and strength of the projections are not represented (see text for further details of extrinsic connectivity). AB, accessory basal nucleus; B, basal nucleus; CE, central nucleus; L, lateral nucleus; ME, medial nucleus; PAC, periamygdaloid cortex; PL, paralaminar nucleus; D, dorsal; M, medial.

Consistent with the pattern of sensory input to the lateral nucleus, output from the basal nucleus to sensory cortex is more robust to high-level processing areas in the temporal lobe.The amygdala also projects to areas of sensory cortex that do not project directly to the amygdala. For example, the basal nucleus projects weakly to visual areas V4, MT, and VI, to primary auditory cortex, and to ventral portions of somatosensory cortex, but these areas do not originate projections to the amygdala. Thus, the amygdala has direct access to early stages of sensory processing and may modulate early stages of perception.

Perirhinal and Parahippocampal Cortex

The perirhinal and parahippocampal cortices are major sources of polysensory input to the amygdala. The projection from the perirhinal cortex is directed primarily to the lateral and basal nuclei. The projection from anterior, polar regions of perirhinal cortex is more robust and terminates in a more widespread manner compared to posterior and ventral regions. The projection from the parahippocampal cortex is weaker and is directed primarily to the basal nucleus. The amygdala projection to perirhinal cortex arises from cells throughout the amygdala, particularly the lateral, basal, and accessory basal nuclei. The projection to the parahippocampal cortex arises almost exclusively from the magnocellular division of the basal nucleus.

Perirhinal and parahippocampal cortices are strongly interconnected to the hippocampal formation via the entorhinal cortex, and they receive input from a wide variety of cortical areas. Perirhinal cortex receives projections from visual association areas TE and TEO, dysgranular and granular insular cortex, area 13 of the orbitofrontal cortex, and parahippocampal cortex. The parahippocampal cortex receives projections from visual areas V4, TE, and TEO, and from polymodal association cortices, including the retrosplenial cortex and the dorsal bank of the superior temporal sulcus. The posterior parietal lobe, insular cortex, and prefrontal cortex also project to the parahippocampal cortex. Their connectional organization, together with behavioral studies in monkeys, suggests that the perirhinal cortex supplies the amygdala with information about object recognition, and the parahippocampal cortex relays information about visuospatial memory.

Entorhinal Cortex

The entorhinal cortex projects weakly to the basal nucleus. Efferent fibers from the entorhinal cortex pass through the lateral nucleus, but it is not clear if the fibers form synapses or terminal plexuses within the nucleus. The projection from the amygdala to the entorhinal cortex arises primarily from the lateral nucleus and is most robust passing to anterior portions of the cortex. Unlike its projections to other areas of cortex, the basal nucleus contributes only a weak projection to entorhinal cortex. Two-thirds of all cortical projections to the hippocampus are relayed through the entorhinal cortex. It is not known if the entorhinal cortex provides the same information to the amygdala as it does to the hippocampus.

Prefrontal Cortex

The prefrontal cortex, including anterior cingulate and orbital and medial prefrontal areas, projects primarily to the basal and accessory basal nuclei. The projection from caudal prefrontal cortex (areas 10 and 25) is more robust and terminates more widely in the amygdala relative to rostral prefrontal cortex.

The amygdala projects to the prefrontal cortical areas that innervate it. Cortically directed efferents arise primarily in the basal and accessory basal nuclei. The organization of the pathways connecting the amygdala and prefrontal cortex is partially overlapping. That is, there appear to be some point-to-point interconnections between the medial prefrontal cortex and the amygdala but not between the orbital prefrontal cortex and the amygdala.

The dorsolateral prefrontal cortex does not project to the amygdala. However, the basal nucleus projects weakly to Walker's areas 46, 9, 8, and 6 in the dorsolateral prefrontal cortex. These projections may allow the amygdala to influence working memory and visuomotor and somatomotor behaviors.

The amygdala is not directly connected to parietal cortex. The strength of amygdala–temporal connections, compared to the lack of amygdala–parietal connections, supports the notion that the identification of a stimulus (what?) is more relevant to amygdala processing than is its location (where?). However, visuospatial information from the parietal cortex may reach the amygdala through disynaptic connections with polysensory areas in temporal cortex, such as the parahippocampal cortex.

Subcortical Connections

Basal Forebrain

The amygdala makes a substantial projection to the basal forebrain, directed primarily to large cholinergic neurons in the nucleus basalis of Meynert and the horizontal limb of the diagonal band. The projection arises in the parvicellular portion of the basal nucleus, the accessory basal nucleus, and the central nucleus. The projection from the basal forebrain to the amygdala originates in cholinergic neurons of the nucleus basalis of Meynert and is directed primarily to the magnocellular portion of the basal nucleus. The magnocellular division of the basal nucleus receives the most substantial cholinergic innervation of any forebrain region.

Cholinergic neurons in the basal forebrain project to widespread regions of the cerebral cortex. They are sensitive to the motivational relevance of sensory cues and they have a modulatory effect on cortical activity and attentional processes. Therefore, the basal forebrain provides an additional route for the amygdala to modify cortical activity in response to sensory events.

Striatum

The amygdala makes a substantial projection to the striatum. The projection originates primarily in the basal nucleus and terminates in the ventral striatum, throughout the caudate nucleus and the rostral–caudal extent of ventral putamen. The amygdala–striatal projection is topographically organized, with the parvicellular division of the basal nucleus projecting to the medial portion of the ventral striatum and the magnocellular division of the basal nucleus projecting more extensively to the body and tail of the caudate nucleus and the ventral putamen. The amygdala innervates regions of the striatum that are also targeted by cortical areas that participate in somatosensory, motor, and cognitive operations. This suggests that motivational and emotional information from the amygdala interacts in the striatum to modulate habit and goal-directed behaviors. The striatum does not make a return projection to the amygdala.

Thalamus

The amygdala makes a substantial projection to the thalamus, originating primarily in the basal and accessory basal nuclei and terminating in the magnocellular division of the mediodorsal nucleus. The magnocellular division of the mediodorsal nucleus projects to the same regions of prefrontal cortex that receive direct input from the amygdala Thus there are direct and nondirect routes from the amygdala to the prefrontal cortex. However, axon terminals from the amygdala and thalamus have different laminar distributions in the prefrontal cortex.

The mediodorsal nucleus is part of a diencephalic memory system, and the strong projection from the

amygdala to the mediodorsal nucleus suggests that the amygdala contributes to mnemonic operations in the thalamus. The mediodorsal nucleus does not make a return projection to the amygdala. The central and medial amygdala nuclei project to midline thalamic nuclei, and the central nucleus also makes a weak projection to medial pulvinar. These portions of thalamus project weakly back to the central nucleus.

Hypothalamus

The amygdala projects to the medial and lateral hypothalamic nuclei. The projection to the medial hypothalamus originates in the cortical nuclei and the basal and accessory basal nuclei. The projection to the lateral hypothalamus originates primarily in the central nucleus. Fear conditioning experiments in rats have shown that the projection from the central nucleus to the lateral hypothalamus is necessary for expression of the autonomic conditioned response. The return projection from the hypothalamus to the amygdala is much weaker and is directed primarily to the central, medial, basal, and accessory basal nuclei.

Brain Stem

Projections from the central nucleus extend through the midbrain (where they terminate in the substantia nigra and ventral tegmentum), the pons, and the medulla, and continue into the spinal cord. No other amygdala nuclei participate in this descending pathway. Along their route, fibers innervate a number of structures that control autonomic and somatomotor responses, including the periaqueductal gray, the dorsal vagal nuclei, the parabrachial nuclei, the nucleus of the solitary tract, and the nucleus ambiguous. Many of these nuclei project back to the amygdala.

Stimulation of the central nucleus or its targets results in autonomic and somatomotor responses, such as cessation of spontaneous activity, pupillary dilation, piloerection, posturing for attack, elevation of blood pressure, and changes in respiration and heart rate. This functional connectivity highlights the amygdala's role in driving responses to emotional events.

Chemoarchitecture

The amygdala is unique in the diversity of neuroactive substances that are distributed throughout it, including amino acids, catecholamines, acetylcholine, peptides, and hormones that exert excitatory, inhibitory, and modulatory effects on the amygdala and its efferent targets. The amygdala, and the central nucleus in particular, is a primary site of anxiolytic and anxiogenic action in the brain.

Function

The amygdala, by way of its extensive intrinsic and extrinsic connections, participates in multiple functional systems. Behavioral studies have demonstrated the importance of the amygdala, in conjunction with these systems, for (1) assigning emotional and motivational significance to a stimulus, (2) generating appropriate responses to environmental and social cues, and (3) modifying behavior based on expected outcomes.

The amygdala's circuitry suggests that it integrates incoming information from sensory, memory, and cognitive systems and provides feedback to each of these systems. This organization indicates that in addition to the aforementioned functions, the amygdala has the potential to modulate memory, learning, cognition, and perception.

See also: Amygdala: Structure and Circuitry in Rodents and Felines; Amygdala: Contributions to Fear; Aversive Emotions: Molecular Basis of Unconditioned Fear; Depression and the Brain; Emotional Learning in Humans; Emotional Hormones and Memory Modulation; Emotional Influences on Memory and Attention; Genetics of Human Anxiety and Its Disorders; Pharmacology of Fear Extinction.

Further Reading

Aggleton JP (ed.) (1992) *The Amygdala. Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction.* New York: Wiley-Liss.

Bauman MD and Amaral DG (2005) The distribution of serotonergic fibers in the macaque monkey amygdala: An immunohistochemical study using antisera to 5-hydroxytryptamine. *Neuroscience* 136: 193–203.

Bauman MD, Lavenex P, Mason WA, et al. (2004) The development of social behavior following neonatal amygdala lesions in rhesus monkeys. *Journal of Cognitive Neuroscience* 16: 1388–1411.

Baxter MG and Murray EA (2002) The amygdala and reward. *Nature Reviews Neuroscience* 3: 563–573.

Emery NJ, Capitanio JP, Mason WA, et al. (2001) The effects of bilateral lesions of the amygdala on dyadic social interactions in rhesus monkeys (*Macaca mulatta*). *Behavioral Neuroscience* 115: 515–544.

Gallagher M and Chiba AA (1996) The amygdala and emotion. *Current Opinion in Neurobiology* 6: 221–227.

Ghashghaei HT and Barbas H (2002) Pathways for emotion: Interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Neuroscience* 115: 1261–1279.

LaBar KS and Cabeza R (2006) Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience* 7: 54–64.

Phelps EA (2006) Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology* 57: 27–53.

Pitkanen A and Amaral DG (1993) Distribution of calbindin-d28k immunoreactivity in the monkey temporal lobe: The amygdaloid complex. *Journal of Comparative Neurology* 331: 199–224.

Pitkanen A and Amaral DG (1993) Distribution of parvalbumin-immunoreactive cells and fibers in the monkey temporal lobe: The amygdaloid complex. *Journal of Comparative Neurology* 331: 14–36.

Pitkanen A and Amaral DG (1994) The distribution of GABAergic cells, fibers, and terminals in the monkey amygdaloid complex: An immunohistochemical and *in situ* hybridization study. *Journal of Neuroscience* 14: 2200–2224.

Sadikot AF and Parent A (1990) The monoaminergic innervation of the amygdala in the squirrel monkey: An immunohistochemical study. *Neuroscience* 36: 431–447.

Stefanacci L and Amaral DG (2000) Topographic organization of cortical inputs to the lateral nucleus of the macaque monkey amygdala: A retrograde tracing study. *Journal of Comparative Neurology* 421: 52–79.

Stefanacci L and Amaral DG (2002) Some observations on cortical inputs to the macaque monkey amygdala: An anterograde tracing study. *Journal of Comparative Neurology* 451: 301–323.

Stefanacci L, Buffalo EA, Schmolck H, et al. (2000) Profound amnesia after damage to the medial temporal lobe: A neuroanatomical and neuropsychological profile of patient E.P. *Journal of Neuroscience* 20: 7024–7036.

Stefanacci L, Suzuki WA, and Amaral DG (1996) Organization of connections between the amygdaloid complex and the perirhinal and parahippocampal cortices in macaque monkeys. *Journal of Comparative Neurology* 375: 552–582.

Suzuki WA and Amaral DG (1994) Perirhinal and parahippocampal cortices of the macaque monkey: Cortical afferents. *Journal of Comparative Neurology* 350: 497–533.

Amygdala: Structure and Circuitry in Rodents and Felines

D Paré, Rutgers, The State University of New Jersey, Newark, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The amygdala is a group of nuclei and cortex-like structures located in the rostromedial portion of the temporal lobe of the brain (Figure 1). As a group, amygdala nuclei receive sensory information of all modalities and project to all levels of the central nervous system, from the prefrontal cortex down to brain stem nuclei involved in cardiovascular control. The amygdala has become the focus of intense scrutiny in recent years, to a large extent because it participates in emotional expression and in the formation of emotional memories. Indeed, the amygdala plays a critical role in the acquisition of classically conditioned responses, both aversive and appetitive. Moreover, the amygdala mediates the facilitation of memory by emotions. It is thought that understanding the mechanisms that support the amygdala's involvement in these functions may shed light on the causes of human anxiety disorders and ways to alleviate them.

Main Components of the Amygdala: Structure, Cell Types, and Connectivity

As in primates, the feline and rodent amygdala can be subdivided into several components: superficial cortex-like fields (such as the periamygdaloid cortex, cortical nuclei, amygdalohippocampal area, and medial nucleus) and numerous deep nuclei that have been classified in various ways. Most accounts recognize a basolateral complex (BLA; including the lateral, basolateral, and basomedial or accessory basal nuclei), the central nucleus (itself typically subdivided in two or more subnuclei), and the intercalated cell masses. The nucleus of the lateral olfactory tract and anterior amygdaloid area will not be discussed here.

The heterogeneous composition of the amygdala has led some to suggest that the term 'amygdala' should be abandoned in favor of a functional classification that takes into consideration the distinct developmental and hodological profiles of its constituents. While this view has much merit, the term 'amygdala' will be used here, not only because most neuroscientists know what is meant when the term is used, but also because amygdala nuclei with different developmental origins are interconnected.

Two main structural differences become apparent when comparing the amygdala in primate and subprimate species (Figure 1). First, the relative size of the nuclei changes, with the primate BLA becoming much more voluminous than the central (CE) and medial nuclei. Second, because of the cortical expansion in monkeys, the amygdala rotates medially such that the CE nucleus shifts from a medial position in rodents to a dorsal position in primates.

Cortex-Like Nuclei of the Amygdala

Although the word 'nucleus' is often used when referring to these structures, the term is a misnomer because these nuclei display a laminar organization typical of cortex, including a cell-sparse molecular layer and a prevalent type of neurons with dendrites extending to the pia in a perpendicular manner. Many of these nuclei receive direct olfactory inputs from the main and/or accessory olfactory bulbs, are reciprocally connected with olfactory cortical areas, and contribute substantial projections to the several hypothalalamic nuclei, including the preoptic area, ventromedial nucleus, and lateral hypothalamus.

Basolateral Complex

Despite the random orientation of neurons in the BLA, its cellular composition is similar to that of the cerebral cortex. As in the cortex, two main cell types prevail in the BLA: multipolar spiny (often pyramidal-shaped) projection cells with highly collateralized axons, and a heterogeneous group of aspiny to sparsely spiny local-circuit neurons. Projection cells account for the majority of neurons, they use glutamate as a transmitter, and they contribute most, if not all, projections to other amygdala nuclei and the rest of the brain. There are at least four subtypes of BLA interneurons, all using the neurotransmitter γ-aminobutyric acid (GABA): parvalbumin-expressing neurons, somatostatin positive interneurons, large cholecystokinin interneurons, and small bitufted cells that display various degrees of co-localization of cholecystokinin, calretinin, and vasoactive intestinal polypeptide. Based on the evidence obtained in cortex, it is likely that these different types of interneurons structure the activity of principal cells in distinct ways because they target different postsynaptic domains and receive contrasting sets of inputs.

The cortical connections of the BLA are extensive. On the input side, the BLA receives cortical inputs about most sensory modalities, and these synapses act via α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and *N*-methyl-D-aspartate (NMDA) glutamate receptors. Moreover, the BLA derives auditory and somatosensory information from the posterior thalamic complex. For these reasons, the

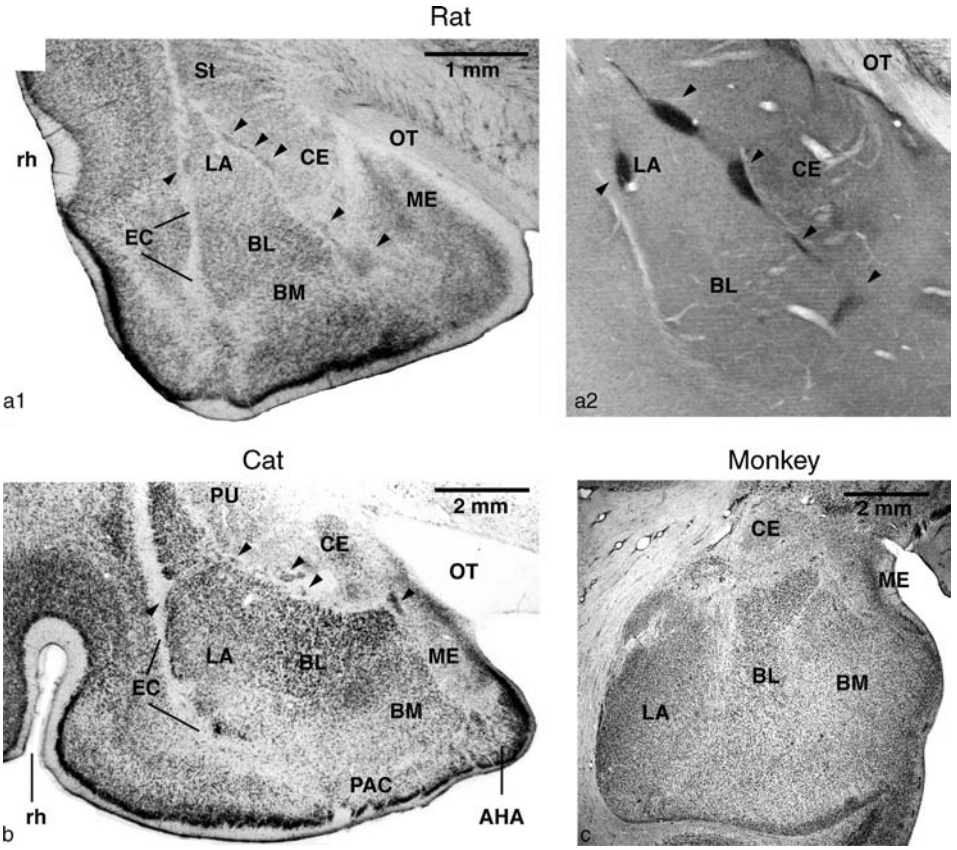

Figure 1 Comparison between the rat (a1 and a2), cat (b), and monkey (c) amygdala. All photomicrographs, with the exception of (a2), show thionin-stained coronal sections at the level of the amygdala. The distribution of μ opioid receptor immunoreactivity in the rat amygdala is shown in (a2). Arrowheads point to intercalated cell masses, where μ opioid immunoreactivity is most intense. AHA, amygdalohippocampal area; BL, basolateral nucleus; BM, basomedial nucleus; CE, central nucleus; EC, external capsule; LA, lateral nucleus; ME, medial nucleus; OT, optic tract; PAC, periamygdaloid cortex; PU, putamen; rh, rhinal sulcus; St, striatum.

BLA (and particularly its lateral nucleus) is typically considered as the sensory gateway of the amygdala. In addition, the BLA has major reciprocal connections with the rhinal cortices, medial prefrontal cortex, and subiculum.

One major species difference seen in the connectivity of the BLA resides in its cortical projections. Whereas the primate BLA projects to extensive neocortical territories, few BLA projections reach cortical fields lateral to the rhinal sulcus in felines and rodents. However, as in primates, the main thalamic projection site of the rodent BLA is the mediodorsal nucleus, but this projection does not exist in felines. Most other BLA connections are highly preserved in primate and subprimate species, including projections to the striatum, lateral hypothalamus, bed nucleus of the stria terminalis, and cholinergic basal forebrain. Last, it should be noted that in contrast to the CE, the BLA contributes few if any projections caudal to the mesencephalon.

Central Nucleus

Traditionally, the CE has been divided in two sections: lateral (CEl) and medial (CEm). However, in rats, CEl has been divided further and there is considerable disagreement regarding the exact borders of these subnuclei. To complicate matters further, some of the components included in the new classifications in fact correspond to cell groups, such as the intercalated cell masses, that were not traditionally included in the CE nucleus. Moreover, these various subnuclei do not appear to form distinct connections with the rest of the brain.

CEm and CEl each contain one prevalent cell type that is believed to use GABA as a neurotransmitter. In the CEm, neurons generally have a fusiform cell body and three to four primary dendrites that branch sparingly and bear a low to intermediate density of spines. In CEl, the main cell type is indistinguishable from the medium spiny neurons of the striatum. These cells generally have smaller somata than do principal CEm neurons and they usually have multiple primary dendrites that branch extensively and bear a high density of spines. Both CEm and CEl contain a low number of aspiny GABAergic neurons, thought to be local-circuit cells.

Even though the lateral nucleus of the amygdala (LA) is generally conceived as the main entry point of

the amygdala for sensory inputs, the CE nucleus also has direct access to sensory information. Indeed, the sensory cortical areas that project to the BLA also typically project to CEl, albeit less intensely. This statement applies to the parietal (somatosensory inputs), piriform (olfactory inputs), and temporal (auditory inputs) and insular (gustatory and visceral inputs) cortices, as well as to several high-order cortical areas (polymodal inputs), such as the prefrontal cortex and rhinal cortices. In contrast, cortical projections to CEm are extremely sparse.

Besides cortical afferents, the CE draws sensory information from subcortical sources. These include afferents from the paraventricular thalamic nucleus (viscerosensory), and the posterior thalamic complex (nociceptive, auditory, polymodal). The posterior thalamic input ends in CEm. Importantly, the CE also receives nociceptive information from the trigeminal complex and spinal cord via the pontine parabrachial nucleus. Last, the BLA relays sensory information sensory from the cortex and thalamus to the CE.

In terms of outputs, CE contributes few if any cortical projections but is the principal amygdaloid source of brain stem projections. These brain stem efferents mainly originate in CEm, CEl projecting only to the parabrachial nucleus. The brain stem targets of CEm include structures controlling motor outputs (reticular formation of the mesencephalon and pons, periaqueductal gray), nuclei that participate in cardiovascular control (dorsal motor nucleus of the vagus, nucleus of the solitary tract), and modulatory cell groups that exert pervasive effects throughout the central nervous system (locus coeruleus, substantia nigra, pedunculopontine nucleus). In addition, CEl is the main amygdaloid source of projections to the cholinergic cell groups of the basal forebrain (substantia innominata, horizontal limb of the diagonal band). Last, the CE exerts indirect control over many of the same structures via the bed nucleus of the stria terminalis.

To summarize, the CE occupies a pivotal situation in the amygdala because it derives sensory inputs from a variety of sources and it targets modulatory systems as well somatomotor and visceromotor control structures. As a result, the CE is typically considered as the output station of the amygdala for fear responses.

Intercalated Cell Masses

In between the BLA (ventrolaterally) and the CE and medial nuclei (dorsomedially) is a fiber bundle that contains small GABAergic neurons in dense clusters known as the intercalated cell masses (Figure 1(a)). Laterally, this fiber bundle pierces the LA and merges with the external capsule, where intercalated cell masses can also be found. In addition, in most species, there is a larger intercalated cluster. This cluster caps the BLA rostrally in felines but is located medial to the BLA in rodents. The available evidence suggests that, irrespective of their position, intercalated clusters receive glutamatergic inputs from various cortical fields and from the BLA. However, their projection sites vary: the main intercalated cluster mainly projects to the substantia innominata and horizontal limb of the diagonal band of Broca, the lateral clusters project to the BLA, whereas the more medial clusters project to CE.

The principal types of intercalated neurons are small cells (8–19 μm in diameter) that use GABA as a neurotransmitter and have spiny dendrites, largely confined to the fiber bundle where their somata are located. Medially located intercalated neurons have two main axon collaterals: one is directed dorsally to the CE and one is directed medially to other intercalated cells. A second, rare type of intercalated cell is characterized by a very large soma (>40 μm in diameter) with extremely long aspiny dendrites that virtually span the entire mediolateral extent of the amygdala.

Electroresponsive Properties of Amygdala Neurons

Most of the data available on the intrinsic properties of amygdala neurons has been obtained in the BLA, CE, and intercalated cell masses. These findings are reviewed in this section.

Basolateral Complex

In terms of electroresponsive properties, three main types of projection cells have been distinguished in the BLA: (1) regular spiking cells that produce trains of action potentials with frequency accommodation, (2) bursting neurons that generate high-frequency spike bursts when depolarized just beyond spike threshold, and (3) late-firing cells that display a striking delay to firing following the onset of depolarizing current pulses.

Regular spiking cells are the prevailing type of projection neuron. Detailed investigations of their properties have revealed that regular spiking cells exhibit a continuum of spike frequency adaptation, from neurons that can fire in a sustained manner throughout depolarizing current pulses with only modest adaptation, to neurons that stop firing after a few spikes, or even that fire only once. These contrasting behaviors are not associated with different morphologies or membrane potentials, but result from variations in the density of calcium-activated potassium currents. In addition, regular spiking projection neurons generate intrinsic subthreshold

membrane potential oscillations in the theta frequency range when depolarized just below firing threshold. These tetrodotoxin-sensitive oscillations allow BLA neurons to reverberate at the theta frequency during arousal.

Two types of aspiny neurons (i.e., presumed inhibitory interneurons) have been described so far. However, on the basis of findings obtained in the cerebral cortex, it is anticipated that future studies will reveal that many more physiological types of interneurons exist in the BLA. The most common type of interneuron is called a 'fast-spiking cell,' by analogy with the corresponding class of parvalbumin-immunoreactive cortical neurons. Like their cortical counterpart, fast-spiking neurons generate brief action potentials and can generate spikes at high rates without frequency accommodation. In addition, a few bursting aspiny neurons have been described. The peptide content of these two types of interneurons remains unknown.

Central Nucleus

Major species differences have been observed between the electroresponsive properties of CE neurons in rats, guinea pigs, and cats. In guinea pigs, the vast majority of neurons in the medial and lateral sectors of the CE exhibit a striking delay to firing when depolarized from hyperpolarized membrane potentials, hence the designation 'late-firing' neurons. This delayed firing onset is due to the activation of a slow voltage-dependent A-like potassium current. Late-firing neurons are believed to be projection neurons.

In contrast, late-firing neurons occur infrequently in rats and cats. In these species, the prevalent type of neuron shows a regular spiking firing pattern and expresses a hyperpolarization-activated mixed cationic current (I_H). Moreover, some rat and cat CE neurons are endowed with a T-current, giving rise to low-threshold spike bursts. Particularly in rats, the shape of spike afterhyperpolarizations (AHPs) has been used to classify CE neurons. Most rat CE neurons show medium-duration AHPs and modest spike frequency adaptation. A minority of CE neurons display long AHPs and pronounced adaptation. Both types appear to be present in CEm and CEl. Last, the CE also contains neurons that have relatively depolarized membrane potentials and a higher input resistance, and display fast-spiking or burst-firing patterns. These rare neurons are thought to be local-circuit cells.

Intercalated Cell Masses

Intercalated neurons display a regular spiking behavior. Moreover, as seen in the CE and BLA, they display a voltage- and time-dependent inward rectification in the hyperpolarizing direction, presumably because they express the hyperpolarization-activated mixed cationic current I_H. However, intercalated cells differ from principal cells of the BLA and CE in two important ways. First, intercalated neurons have an unusually high input resistance (around 800 MΩ), much higher than typically seen in principal BLA and CE cells (usually <200 MΩ). As a result, intercalated cells are more excitable and tend to have higher spontaneous firing rates.

In addition, intercalated cells express an unusual type of potassium current (slowly deinactivating; I_{SD}) that activates at subthreshold membrane potentials, shows inactivation in response to suprathreshold depolarizations, and recovers from inactivation with an extremely slow time course. As a result, the excitability of intercalated neurons is a function of their recent firing history. Indeed, following bouts of intense firing, the inactivation of I_{SD} causes an increase in input resistance and a membrane depolarization – in effect, increasing the likelihood that synaptic inputs will trigger action potentials. Moreover, because each spike renews the inactivation of I_{SD}, this state of increased excitability is self-sustaining.

Neuronal Interactions between the Basolateral Amygdala and Central Nucleus

A dominant theme of research on the amygdala has been its involvement in the acquisition of classically conditioned fear responses. In this model of emotional learning, a neutral sensory stimulus (conditioned stimulus; CS) acquires the ability to elicit fear responses after repeated pairings to a noxious stimulus (unconditioned stimulus; US). According to the prevailing model, convergence of sensory and nociceptive information in the LA would lead to a Hebbian potentiation of synapses conveying information about the CS. As a result, subsequent presentations of the CS would elicit larger responses in CE neurons. In turn, the CE would generate conditioned fear responses via its hypothalamic and brain stem projections. However, synaptic plasticity in the BLA is under tight inhibitory control. In fact, it was shown that inhibitory interneurons effectively gate the induction of long-term potentiation in principal cells. Thus, the following section considers the intrinsic inhibitory pressures that regulate synaptic excitability and associative plasticity within the BLA as well as the propagation of signals from the BLA to the CE.

Intrinsic Control of Neuronal Excitability and Plasticity within the Basolateral Amygdala

Although the cellular composition of the BLA is similar to that of the cerebral cortex, principal BLA

neurons have extremely low firing rates (typically <1 Hz) compared to cortical neurons (average of 10 Hz). This dissimilarity is particularly puzzling because the BLA has an extremely divergent intrinsic connectivity. Indeed, principal BLA neurons have highly collateralized axons, each bearing numerous varicosities that form *en passant* glutamatergic AMPA–NMDA synapses, typically with the dendritic spines of other projection cells. In fact, based on the average length of the intervaricose segments (5–10 μm), it is estimated that for every millimeter of axon, each projection cell forms 100–200 excitatory synapses with other projection cells.

Considering this remarkable potential for divergence within the BLA, the low spontaneous firing rates of principal cells suggest that major inhibitory pressures regulate synaptic excitability within the BLA. In keeping with this, *in vivo* intracellular studies have revealed that the spontaneous activity of BLA neurons is dominated by large-amplitude hyperpolarizing potentials. These inhibitory potentials are generated by the combined influence of synaptically activated intrinsic membrane conductances and inhibitory synaptic conductances. First, principal cells exhibit a calcium-dependent potassium conductance that can be synaptically activated by calcium entry via NMDA receptors. Second, the activity of principal cells is tightly regulated by local-circuit neurons. In response to afferent stimulation, local-circuit cells generate biphasic inhibitory postsynaptic potentials (IPSPs) in principal BLA neurons via GABA-A and GABA-B receptors. In contrast, interneurons are subjected to far less inhibition than principal cells, because they receive fewer inhibitory synapses than principal cells and their GABA-A reversal potential is more depolarized. This results from the fact that the prevalent regulators of intracellular chloride are cation–chloride cotransporters that accumulate chloride in local-circuit cells and extrude chloride in principal neurons.

An intriguing property of the intrinsic inhibitory BLA network is that it is spatially heterogeneous. In the LA, for instance, feedback interneurons tend to inhibit projection cells located in the same coronal plane, their axons extending little rostrocaudally. As a group, the other types of interneurons also show spatial heterogeneity in the control of projection cells. They tend to generate more inhibition in neurons located laterally. Given that different types sensory inputs end in distinct compartments of the BLA, this implies that inscribed in the BLA network is a predisposition to preferentially attend to particular types of sensory information.

In addition to local-circuit neurons, principal LA neurons receive inhibitory inputs from intercalated clusters found in the external capsule. Interestingly, these two populations of GABAergic neurons display opposite responses to dopamine: local-circuit cells are excited whereas intercalated neurons are inhibited by dopamine. Yet, the net effect of dopamine is a disinhibition of principal LA neurons, resulting in a facilitation of long-term potentiation (LTP) induction at thalamic inputs.

Transfer of Sensory Inputs from the BLA to the CE

All BLA nuclei send glutamatergic projections to CE, exciting CE neurons through AMPA and NMDA glutamate receptors. As mentioned previously, it is believed that fear conditioning results in a potentiation of CS-evoked responses in the LA and that, as a result, subsequent presentations of the CS would elicit larger responses in the CE. However, tract-tracing studies indicate that there are no direct connections between the input (LA) and output (CEm) stations of the amygdala for conditioned fear responses. Indeed, the LA nucleus projects only to CEl, not to CEm, the subnucleus at the origin of most amygdala projections to brain stem structures mediating fear responses. Thus, it is likely that the LA transmits CS-related information to the CE indirectly. Consistent with this, the LA sends robust glutamatergic projections to the basal nuclei that, in turn, project to CEl and CEm. Moreover, post- but not pre-conditioning lesions of the basal nuclei block conditioned fear responses, suggesting that in an intact brain, the basal nuclei are critically involved, at least in transferring CS-related signals to the CE. However, the fact that preconditioning lesions of the basal nuclei do not prevent acquisition of conditioned fear responses suggests that CS-related information can reach the CE through other paths. Consistent with this, CEm is the target of auditory and somatosensory inputs from the posterior thalamus, and these inputs can undergo activity-dependent potentiation.

Gating of BLA Projections to the CE by Intercalated Neurons

En route to the CE, BLA axons contact GABAergic intercalated neurons that generate GABA-A IPSPs in CE neurons. Accumulating evidence suggests that despite their small size, intercalated cell clusters play a pivotal role in the synaptic economy of the amygdala. The available data indicate that there is a lateromedial correspondence between the position of intercalated cells, where they project in the CE, and the source of their BLA afferents. Moreover, intercalated clusters are interconnected, but these connections prevalently run in a lateromedial direction. Because of these connections, the amount of feedforward inhibition

intercalated cells generate in the CE, and thus the amplitude of the excitatory postsynaptic potentials (EPSPs) generated by BL inputs, will depend on the spatiotemporal distribution of excitation in the BLA. Thus, inscribed in this network is the possibility of modulating behavior responsiveness as a function of current environmental contingencies. Moreover, because BLA inputs to intercalated cells exhibit activity-dependent plasticity, gating of BLA inputs to the CE by intercalated cells can be modified by experience.

Another factor adding flexibility to this gating function is the presence of glutamatergic inputs from various cortical areas to intercalated neurons. In particular, medial prefrontal inputs to intercalated cells have been implicated in the extinction of classically conditioned fear responses. Extinction refers to the gradual disappearance of conditioned fear responses when the CS is repeatedly presented without pairing to the US. Much behavioral evidence suggests that extinction does not result from unlearning of the original CS–US association, but from a new learning that competes with the first one. This view is based on the fact that extinguished fear memories can undergo spontaneous recovery over time, renewal (recovery when tested in a context different than the extinction context), and reinstatement (recovery when an unsignaled US is presented in the extinction context). Moreover, intra-amygdala injections of NMDA antagonists or of inhibitors of the mitogen-activated protein kinase (MAPK) pathway prevent extinction learning. At present, the processes underlying extinction learning remain obscure. However, mounting evidence suggests that an extrinsic input to the amygdala plays a critical role. Indeed, lesions of the medial prefrontal cortex impair recall of extinction. Moreover, in rats that have not been extinguished, electrical stimulation of the medial prefrontal cortex during the CS reduces conditioned responses. Finally, electrical stimulation of the medial prefrontal cortex produces an inhibition of brain stem-projecting CEm neurons. Because the medial prefrontal cortex sends little projections to CEm, it is believed that the inhibitory effects of medial prefrontal inputs on fear expression are mediated by intercalated neurons. One possibility is that prefrontal inputs, by depolarizing intercalated cells, facilitate NMDA-dependent plasticity of BLA inputs to intercalated neurons. As a result, subsequent CS presentations would evoke more inhibition in CE neurons.

The Amygdala Mediates the Facilitation of Memory by Emotions

Humans generally form more vivid memories of emotionally arousing events than of mundane experiences, and this trend is absent in individuals with amygdala lesions. In part, the impact of emotions on memory involves an enhancement of attention during encoding. Indeed, functional imaging studies have revealed that emotional stimuli activate the amygdala, that attention is enhanced by emotional material, and that this effect is absent in individuals with amygdala lesions. Consistent with this, animal studies have revealed that amygdala stimulation produces an electroencephalogram (EEG) activation associated with an orienting reaction.

However, animal work indicates that besides enhancing attention, the amygdala can also facilitate the consolidation of emotional memories. Overall, the available evidence suggests that stress hormones released in emotionally arousing conditions increase the activity of BLA neurons. In turn, the enhanced activity of BLA cells would facilitate memory consolidation in other brain areas via their widespread cortical and subcortical projections. For instance, posttraining intra-BLA injections of drugs that enhance or reduce BLA activity, respectively, facilitate or impair retention on a variety of emotionally charged learning tasks. Moreover, BLA lesions prevent the facilitation of memory by systemic injections of stress hormones. Importantly, although long-term memory can be affected by altering amygdala activity right after learning, manipulations of BLA excitability at later time points have no effects. These observations imply that the aforementioned memory-modulating effects of the amygdala manipulations do not result from alterations of memory storage in the amygdala, but rather in other structures that presumably constitute the storage site of particular forms of memories.

At present, the mechanisms underlying the facilitation of memory consolidation by amygdala activity remain unknown. One possibility is that the enhancement of BLA activity produced by emotional arousal leads to the recruitment of basal forebrain cholinergic neurons that project to the cerebral cortex. In turn, this amygdala-driven release of acetylcholine would facilitate plasticity in cortical networks. However, it is known that amygdala activity can also facilitate striatal-dependent memories. Since the striatum receives little if any inputs from basal forebrain cholinergic neurons, other mechanisms must be involved. A challenge for future studies will be to determine the distinguishing properties that allow amygdala synapses to facilitate synaptic plasticity in their targets.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Contributions to Fear; Appetitive Systems: Amygdala and Striatum.

Further Reading

Dringenberg HC and Vanderwolf CH (1996) Cholinergic activation of the electrocorticogram: An amygdaloid activating system. *Experimental Brain Research* 108: 285–296.

Everitt BJ, Cardinal RN, Parkinson JA, et al. (2003) Appetitive behavior: Impact of amygdala-dependent mechanisms of emotional learning. *Annals of the New York Academy of Sciences* 985: 233–250.

LeDoux JE (2000) Emotion circuits in the brain. *Annual Review of Neuroscience* 23: 155–184.

Maren S and Quirk GJ (2004) Neuronal signaling of fear memory. *Nature Reviews Neuroscience* 5: 844–852.

McDonald AJ (1992) Cell types and intrinsic connections of the amygdala. In: Aggleton JP (ed.) *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*, pp. 67–96. New York: Wiley-Liss.

McDonald AJ (1998) Cortical pathways to the mammalian amygdala. *Progress in Neurobiology* 55: 257–332.

McGaugh JL (2004) The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience* 27: 1–28.

Pare D (2003) Role of the basolateral amygdala in memory consolidation. *Progress in Neurobiology* 70: 409–420.

Paré D, Royer S, Smith Y, et al. (2003) Contextual inhibitory gating of impulse traffic in the intra-amygdaloid network. *Annals of the New York Academy of Sciences* 985: 78–91.

Paré D, Quirk GJ, and LeDoux JE (2004) New vistas on amygdala networks in conditioned fear. *Journal of Neurophysiology* 92: 1–9.

Phelps EA (2004) Human emotion and memory: Interactions of the amygdala and hippocampal complex. *Current Opinion in Neurobiology* 14: 198–202.

Pitkänen A (2000) Connectivity of the rat amygdaloid complex. In: Aggleton JP (ed.) *The Amygdala: A Functional Analysis*, pp. 31–115. Oxford: Oxford University Press.

Quirk GJ, Garcia R, and Gonzalez-Lima F (2006) Prefrontal mechanisms in extinction of conditioned fear. *Biological Psychiatry* 60: 337–343.

Sah P, Faber ESL, De Armentia ML, et al. (2003) The amygdaloid complex: Anatomy and physiology. *Physiological Reviews* 83: 803–834.

Swanson LW and Petrovich GD (1998) What is the amygdala? *Trends in Neuroscience* 21: 323–331.

Weinberger NM (2004) Specific long-term memory traces in primary auditory cortex. *Nature Reviews Neuroscience* 5: 279–290.

Amyloid: Vascular and Parenchymal

R O Weller, R O Carare, and D Boche, University of Southampton School of Medicine, Southampton, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Age brings many changes to the body and age is also a major risk factor for dementia. Alzheimer's disease is one of the major forms of dementia in the elderly and is characterized by the accumulation of amyloid beta (Aβ) peptides in the brain. Although there are no traditional lymphatics in the brain, interstitial fluid and solutes such as Aβ are eliminated from brain tissue along the basement membranes in the walls of capillaries and cerebral arteries. Here we review how aging of the brain and cerebral arteries may lead to the failure of elimination of Aβ from the brain and how this may contribute to the onset of dementia, particularly in Alzheimer's disease. Immunotherapy for the removal of Aβ from the brain is also briefly discussed.

Aging of the Brain and Cerebral Arteries

Changes occur progressively in the brain and cerebral arteries with advancing age, and age is a major risk factor for Alzheimer's disease and for cerebrovascular disease. Neurons in the brain are very long-lived cells, as they are established before birth and survive for the lifetime of the individual. With age, large neurons in the cerebral cortex and spinal cord accumulate waste products in the cytoplasm in the form of lipofuscin. Damaged mitochondria and membranes are broken down by lysosomal hydrolytic enzymes, and residual material accumulates in the cytoplasm of neurons as secondary lysosomes (lipofuscin). However, apart from a few familial disorders (lipofuscinoses), the lipofuscin in neurons does not seem to be a factor in dementia. Whereas cell organelles are broken down by lysosomes, cytoplasmic proteins are degraded by proteosomes. Damaged proteins become associated with ubiquitin and are delivered to proteosomes for degradation. As will be discussed later, there appears to be a failure of the ubiquitin–proteosomal system in dementias, particularly in Alzheimer's disease. In addition to the changes that occur within aging neurons, astrocytosis is seen in the brain, with advancing age reflecting the damage that occurs throughout life from head injuries, episodes of viral encephalitis, and ischemia.

Cerebral blood vessels and blood vessels elsewhere in the body change with age in a similar way. There is progressive stiffening of arterial walls, with an increase in collagen in the tunica intima and in the tunica media. Major cerebral arteries, such as the vertebral, basilar, and middle cerebral arteries, show progressive dilatation and tortuosity due to fibrosis of the intima and media. Generalized thickening and stiffening of cerebral arteries (arteriosclerosis) is also accompanied by patchy thickening and the accumulation of cholesterol within atherosclerotic plaques of the major cerebral arteries. Occlusion of large or small cerebral arteries may result from emboli arising in the heart or from thrombi associated with atherosclerosis of the carotid arteries. Large infarcts resulting from occlusion of cerebral arteries may result in strokes, with focal neurological signs such as hemiparesis, whereas multiple small infarcts may produce no overt neurological signs but can result in vascular dementia (see Table 1).

Changes in the Brain Associated with Dementia

Since the description in 1907 of the microscopic features in the brain in Alzheimer's disease, many pathological changes associated with different types of dementia have been described. As shown in Table 1, the pathology of dementias is classified broadly based upon presence of intracellular protein inclusions mainly within neurons (tau, synuclein, ubiquitin, and huntingtin), extracellular deposits of amyloid proteins in the brain and in blood vessel walls (Aβ, other amyloids, and prion protein), and small or large infarcts (multi-infarct or vascular dementia). (Lipofuscin has been included in Table 1, although excessive lipofuscin is a rare cause of dementia.)

Intracellular Protein Deposits

Two major proteins, tau and synuclein, accumulate within neurons in dementia patients. Both proteins are associated with ubiquitin, which suggests that disposal of these proteins through the ubiquitin–proteosomal system may fail with age and dementia.

Accumulation of Tau Protein

Tau protein (named after the Greek letter for 't') is a microtubule-associated protein that is concerned with axoplasmic transport in normal neurons. Hyperphosphorylated tau accumulates with ubiquitin in aging neurons as the neurofibrillary tangles that were identified by Alzheimer and others by the use of silver stains (Figure 1). Examination of postmortem

Table 1 Age changes in the brain associated with dementia[a]

Type of dementia	*Intracellular inclusions in neurons*	*Extracellular deposits*	*Cerebral amyloid angiopathy*	*Infarcts*
Aging brain	Lipofuscin +++ Tau +	Aβ +	Aβ +	+
Vascular dementia	Lipofuscin +++ Tau +	Aβ +	Aβ +	**++++**
Alzheimer's disease	Lipofuscin +++ **Tau** +++	**Aβ** ++++	**Aβ** ++++	+/++
Dementia with Lewy bodies	Lipofuscin +++ **Synuclein** +++ Tau ++	Aβ ++	Aβ ++	+/++
Creutzfeldt–Jakob disease	Lipofuscin +++ in elderly	**Prion protein** ++++	+/− (prion protein)	+
Huntington's disease	**Huntingtin** +++	Aβ +/−	+/− (Aβ)	+
Frontotemporal dementias	Lipofuscin +++ **Tau** ++++ **or ubiquitin only** +++	Aβ +/−	+/− (Aβ)	+

[a]Intracellular inclusions, extracellular protein deposits, and vascular damage in the aging brain and in different major types of dementia. The features used as the pathological diagnostic criteria in the various dementias are in boldface type.

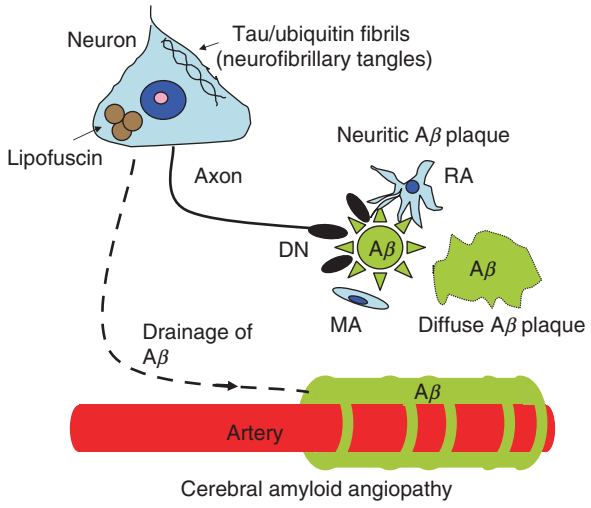

Figure 1 Histopathological features of Alzheimer's disease are characterized by the accumulation of neurofibrillary tangles composed of tau and ubiquitin within neurons, and by the deposition of amyloid beta (Aβ) in brain parenchyma as amyloid plaques and in blood vessel walls as cerebral amyloid angiopathy. Aβ in amyloid plaques is associated with microglial activation (MA) and reactive astrocytosis (RA). Swollen dystrophic neurons (DN) containing tau are associated with many of the amyloid plaques (neuritic plaques).

brains of aged patients with no impairment of cognition reveals some neurons with tau in their cytoplasm, particularly in the hippocampus. In Alzheimer's disease, neurofibrillary tangles composed of hyperphosphorylated tau and ubiquitin complexes are more widespread and more numerous, especially in the temporal cortex, hippocampus, and frontal lobes. Mutations in the tau gene on chromosome 17 are associated with a familial dementia characterized by widespread accumulation, in neurons, of neurofibrillary tangles containing tau protein.

Accumulation of Synuclein

Synuclein is a protein associated with synapses in the brain; it accumulates with ubiquitin as spherical structures (Lewy bodies) within neurons, particularly of the substantia nigra and locus coeruleus in Parkinson's disease. In dementia with Lewy bodies (see Table 1), synuclein and ubiquitin complexes (Lewy bodies) are observed in neurons of the cingulate gyrus, frontal and temporal cortices, and insula; synuclein also accumulates within neuronal processes, particularly in the hippocampus. There appears to be a failure in the ubiquitin–proteosomal system but the reasons for this are unclear in the majority of cases of dementia with Lewy bodies. Some familial cases of Parkinson's disease or dementia with Lewy bodies are associated with mutations in the α-synuclein or parkin genes.

Accumulation of Huntingtin

Inclusions of the protein huntingtin occur within nuclei of neurons in Huntington's disease (see Table 1) in which there is expansion of triplet repeats in the gene encoding huntingtin, on chromosome 4.

Deposition of Amyloid in the Extracellular Spaces of the Brain

A number of different proteins accumulate in the brains of elderly and demented patients as insoluble amyloid deposits in the extracellular spaces of gray matter, and, in many cases, in the walls of arteries as cerebral amyloid angiopathy. The most common of these proteins is amyloid beta (Aβ), which may accumulate in the brains of elderly people who show no evidence of cognitive decline; Aβ is deposited much more abundantly in the brain in Alzheimer's disease.

Other proteins, such as prion protein, accumulate in the gray matter in Creutzfeldt–Jakob disease, and a variety of other amyloid proteins, such as cystatin, transthyretin, and the British and Danish types of amyloid, accumulate in the extracellular spaces of the brain and in blood vessel walls in association with different types of familial dementia. In the following sections we explore the mechanisms by which amyloid proteins accumulate in the brain and blood vessel walls in elderly and demented patients, and in particular in Alzheimer's disease.

Accumulation of Amyloid Proteins in the Brain in Alzheimer's Disease

Alzheimer's disease is the commonest form of dementia in North America and Europe and has recently increased in incidence in many Asian countries. The disease is characterized pathologically by the accumulation of hyperphosphorylated tau within neurons and by the deposition of Aβ in the extracellular spaces of gray matter in the brain and in the walls of cerebral arteries (Figure 1). Overproduction of Aβ and failure of elimination of Aβ from brain tissue appear to be the major reasons why Aβ accumulates in the brain in old age and in Alzheimer's disease.

A number of familial dementias have been identified in which there is either an overproduction of Aβ, particularly of the less soluble form, A$\beta_{1–42}$, or there is a production of a chemically aberrant form of Aβ. Throughout life, Aβ is produced by neurons and virtually all other cells in the body by the cleavage of a transmembrane protein, amyloid precursor protein (APP), encoded on chromosome 21. Three major secretases (α, β, and γ) cleave APP to produce Aβ in a number of forms that differ in the number of amino acids in the Aβ peptide. The two major forms that accumulate in the brain and in the walls of arteries are A$\beta_{1–40}$ (40 amino acids long) and A$\beta_{1–42}$, which contains 42 amino acids and is less soluble than A$\beta_{1–40}$. A$\beta_{1–42}$ is mainly in the insoluble amyloid plaques within brain tissue and A$\beta_{1–40}$ predominates in the blood vessel walls in cerebral amyloid angiopathy. In the familial forms of Alzheimer's disease, genetic mutations in the APP gene affect the cleavage sites of APP or the structure of Aβ. Mutations in the presenilin 1 and 2 genes appear to affect the γ-secretase cleavage of Aβ from APP. In these types of familial Alzheimer's disease, failure of elimination of the excessive amount of Aβ from the aging brain appears to be a major factor in the pathogenesis of the dementia.

Most cases of Alzheimer's disease are sporadic, with familial cases accounting for less than 10% of all cases of Alzheimer's disease. In the more common sporadic form of Alzheimer's disease, there is no firm evidence of overproduction of Aβ, so it is probable that failure of elimination of Aβ from the brain with advancing age is a major pathogenetic factor (Figure 1).

Elimination of Aβ from the Brain

Aβ is produced throughout life in the brain and it is present at an average concentration of 2.75 ng g^{-1} of cerebral cortex in clinically nondemented individuals; Aβ is greatly increased in the brains of patients with Alzheimer's disease. Studies in experimental mice and humans reveal a number of mechanisms by which Aβ is eliminated from the normal brain. Aβ is absorbed into the blood by several known pathways, such as that mediated by low-density lipoprotein receptor-related protein-1 (LRP-1), and it is also degraded by peptidases such as neprilysin, in the parenchyma of the brain. In addition, Aβ is eliminated from the brain with interstitial fluid along perivascular drainage pathways – effectively the lymphatics of the brain.

Perivascular Route for the Elimination of Interstitial Fluid and Aβ from the Brain

There is no traditional lymphatic system in the brain. In other tissues, such as lung and skin, there are well-defined thin-walled lymphatic vessels along which interstitial fluid, solutes, and inflammatory cells drain to regional lymph nodes. Although there are no true lymphatics in the brain, studies in experimental animals have shown that solutes injected into the brain do drain to lymph nodes in the neck. Thus, when a soluble fluorescent tracer of molecular mass equivalent to that of Aβ (4 kDa) is injected into gray matter of the brain, it initially spreads diffusely through the narrow extracellular spaces and rapidly enters basement membranes of capillary and artery walls, to drain out of the brain. It does appear that the perivascular route along basement membranes in the walls of capillaries and arteries is the pathway for lymphatic drainage of the brain.

Aβ is deposited in the walls of cerebral capillaries and arteries in aged humans and in Alzheimer's disease in a pattern that suggests Aβ also drains from the brain along basement membranes of capillary and artery walls. As Aβ progressively accumulates in the artery walls, smooth muscle cells in the tunica media die, so that in severe cerebral amyloid angiopathy the whole thickness of the artery is replaced by amyloid, thus increasing the likelihood of intracerebral hemorrhage.

Figure 2 summarizes the formation and elimination of interstitial fluid and metabolites from gray-matter

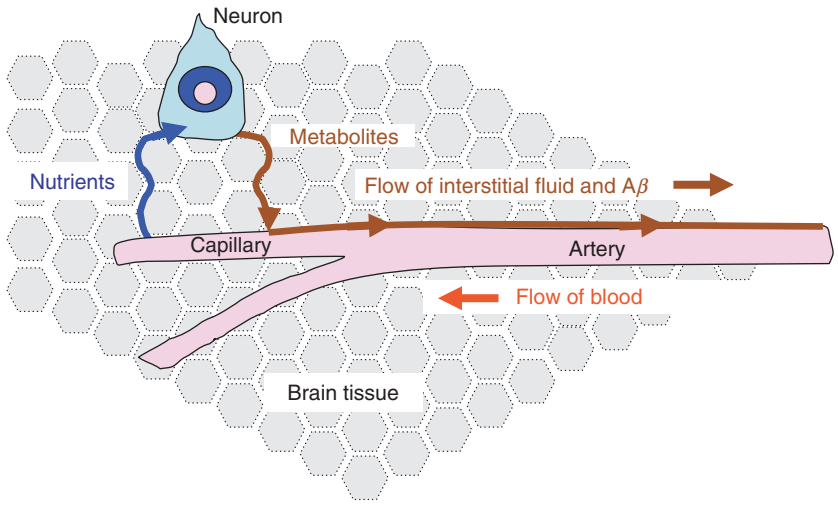

Figure 2 Perivascular drainage of interstitial fluid and metabolites from the brain. Experimental studies suggest that nutrients pass through the blood–brain barrier into capillary endothelial basement membranes and thence diffuse to neurons and other cells in the brain. Metabolites from these cells diffuse back to the capillary basement membranes through the narrow and restricted extracellular spaces of the cerebral cortex and other areas of gray matter in the brain. Once metabolites, including amyloid beta (Aβ), have entered capillary basement membranes, they pass by relatively unrestricted bulk flow into the basement membranes of arteries and out of the brain. The lymphatic flow along artery walls is probably driven by the reflection waves arising from arterial pulsations. Stiffening of artery walls with age may reduce the amplitude of arterial pulsations and reflection waves, thus interfering with the perivascular elimination of metabolites, including Aβ, from the brain.

areas of the brain. Nutrients pass through the blood–brain barrier into capillary basement membranes, which act as conduits or channels for the circulation of nutrients to neurons and other cells within the brain. Metabolites produced by cells in the brain pass through the very narrow extracellular spaces to drain out of the brain, with interstitial fluid, along basement membranes in the walls of capillaries and arteries.

Data from experimental and human studies suggest that interstitial fluid and solutes from the brain drain along artery walls to cervical lymph nodes just under the base of the skull. It is unlikely that lymphocytes and other inflammatory cells can migrate from the brain to lymph nodes along the narrow basement membrane pathways followed by interstitial fluid and solutes. Such failure of lymphocyte migration from the brain may be a factor in the well-known 'immunological privilege' of the brain. The perivascular pathway for drainage of interstitial fluid and solutes from the brain is largely separate from the cerebrospinal fluid (CSF).

Pathways for the Drainage of CSF

Drainage pathways for CSF have been extensively studied in both animals and humans. Tracers injected directly into the CSF of rodents drain largely through channels in the cribriform plate into lymphatics of the nasal mucosae and thence to cervical lymph nodes. In humans, however, CSF drains directly back into the blood through arachnoid granulations and villi in the walls of intracranial and spinal venous sinuses, and the route through the cribriform plate plays a very minor role.

Failure of Elimination of Aβ from the Brain with Advancing Age and Alzheimer's Disease

The various mechanisms and routes for the removal of Aβ from the brain in young animals and young humans appear to fail with advancing age (Figure 3). In mice, degradation of Aβ by neprilysin within the brain parenchyma, and absorption of Aβ into the blood by the LRP-1-mediated mechanism, are reduced with increasing age. Accumulations of insoluble amyloid Aβ in the interstitial fluid drainage pathways in artery walls in elderly individuals and in patients with Alzheimer's disease suggest that the perivascular route for elimination of Aβ also fails with age.

Detailed observations in tracer studies have shown that solutes drain along capillary basement membranes and then along basement membranes between the smooth muscle cells of artery walls. Using these data, theoretical mathematical models indicate that pulsation of artery walls is the motive force for the perivascular drainage of interstitial fluid and solutes from the brain. The model predicts that fluid and solutes would be driven along the vascular basement

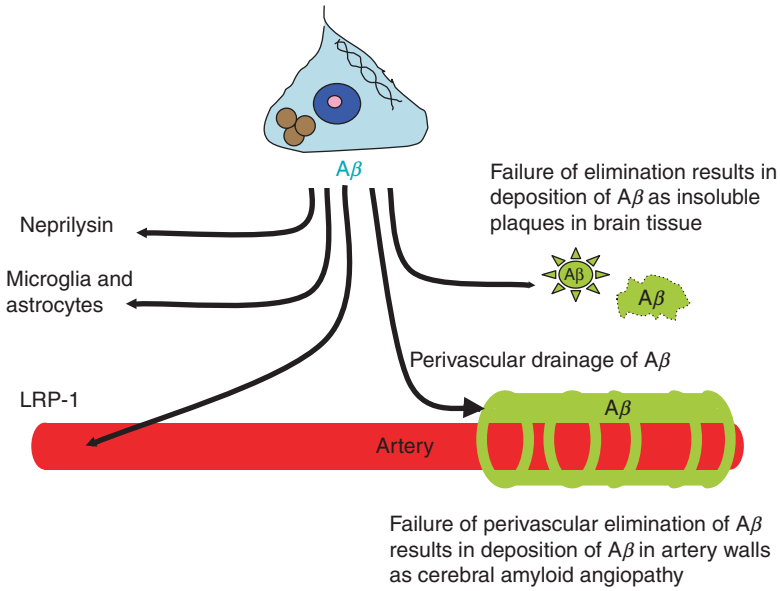

Figure 3 Pathways and mechanisms for the elimination of amyloid beta (Aβ) from the brain fail with advancing age, resulting in the accumulation of Aβ in brain parenchyma and in blood vessel walls. In the young, Aβ is absorbed into the blood by a mechanism mediated by low-density lipoprotein receptor-related protein-1 (LRP-1) and is degraded in brain parenchyma by neprilysin and other enzymes. Aβ also drains from the brain with interstitial fluid along capillary and artery walls. Failure of the LRP-1 mechanism and neprilysin with age and the failure of perivascular drainage of Aβ along stiff, aged arteries are associated with deposition of insoluble Aβ as plaques in brain parenchyma, and in blood vessel walls as cerebral amyloid angiopathy. Ultimately, levels of soluble Aβ, and possibly other metabolites in the brain, rise, and this is associated with decline of cognitive function and dementia in Alzheimer's disease.

membranes in the reverse direction to the blood flow during diastole by the reflection waves generated by the arterial pulse wave. The strength of the reflection waves driving the drainage of interstitial fluid and solutes from the brain would be proportional to the amplitude of the pulsations.

Why then does the perivascular drainage of fluid and solutes such as Aβ fail with age and in Alzheimer's disease? Age is not only a risk factor for Alzheimer's disease, it is also a risk factor for cerebrovascular disease in which there is fibrosis and stiffening of artery walls (arteriosclerosis), atherosclerotic plaque formation, and occlusion of arteries by thrombi or emboli. As cerebral arteries stiffen with age the amplitude of vessel pulsations would be reduced, and this would also reduce the strength of the reflection wave driving the interstitial fluid and solutes from the brain along the walls of arteries. Hypothetically, reduction in the motive force in stiffened arteries would lead to slowing of drainage of Aβ and induce the precipitation of insoluble fibrils of Aβ in the vessel walls. Such deposition would further impede the drainage of Aβ and eventually lead to the accumulation of soluble and insoluble Aβ and other metabolites in the brain.

Direct evidence that cerebrovascular disease and stiffening of artery walls impedes the elimination of Aβ along perivascular pathways is difficult to obtain. Nevertheless, thrombotic occlusion of small cortical arteries with complete abolition of vascular pulsations is associated with deposition of Aβ in the basement membrane of capillaries supplied by the occluded arteries. This suggests that once the pulsations in the vessel cease, Aβ can no longer be propelled along the drainage pathway from capillaries basement membranes into artery walls.

Consequences of Amyloid Deposition in the Brain and Vessel Walls

In cerebral amyloid angiopathy, there are two major complications of the deposition of Aβ in the walls of cerebral arteries. The acute complication is intracerebral hemorrhage and the more long-term complication is related to the onset of dementia, particularly of Alzheimer's disease. Soluble and insoluble forms of Aβ accumulate in the cerebral cortex, and the fluid content in subcortical white matter is increased.

Intracerebral Hemorrhage

There is a close association between intracerebral hemorrhage and cerebral amyloid angiopathy in the elderly. As increasing amounts of Aβ are deposited in the basement membranes of artery walls, smooth muscle cells in the tunica media are destroyed and the vessel wall becomes a brittle tube of insoluble Aβ,

which makes it liable to rupture, causing an intracerebral hemorrhage. This complication occurs mainly in patients over the age of 75 years and can be fatal or may induce neurological deficits. Recent magnetic resonance imaging (MRI) studies have shown that the majority of intracerebral hemorrhages that have been attributed to cerebral amyloid angiopathy occur in the temporal and occipital lobes, supplied mainly by the posterior cerebral arteries.

Dementia

The other major consequence of amyloid deposition in blood vessel walls is the failure of interstitial fluid drainage and the failure of elimination of Aβ from the aging brain. The pathological diagnostic criteria for Alzheimer's disease are mainly based upon the presence of neurofibrillary tangles within neurons and the severity of deposition of insoluble Aβ as plaques in the cerebral cortex and hippocampus. As the elimination of Aβ fails, it is deposited as insoluble plaques in gray-matter areas of the brain. The more compact plaques induce microglial activation and are associated with axonal damage and dystrophic neurites; the term 'neuritic plaque' is used for these structures (Figure 1). Other plaques of Aβ are more diffuse and do not induce the same degree of brain tissue damage.

Deposits of insoluble Aβ in brain tissue impede the elimination of Aβ and other metabolites from the brain. Before entering the drainage pathways in capillary and artery walls, interstitial fluid and metabolites diffuse through the narrow extracellular spaces of gray matter; Aβ plaques block this diffusion. Eventually, drainage of solutes is severely restricted and levels of soluble Aβ in gray matter rise. Recent studies have emphasized that a high level of soluble Aβ in cerebral cortex and the severity of cerebral amyloid angiopathy correlate better with cognitive decline in patients with Alzheimer's disease than does the number of insoluble plaques of Aβ.

Why the failure of elimination of soluble Aβ results in dementia in Alzheimer's disease is still not clear. Many studies suggest that the toxicity of Aβ to neurons is not due to the accumulation of insoluble fibrils but rather is due to the presence of a soluble pool of Aβ, particularly in its oligomeric form. In transgenic mice, dodecameric Aβ assemblies are associated with an impairment of memory independently of Aβ plaques or neuronal loss. Thus increased levels of soluble oligomeric Aβ may be a factor inducing the onset of dementia in Alzheimer's disease. However, it is also possible that other metabolites do not drain adequately from gray matter, resulting in their accumulation in the extracellular spaces, a change in neuronal environment, neuronal malfunction, and dementia.

Creutzfeldt–Jakob disease is another dementia that is characterized by the deposition of insoluble amyloid protein in the extracellular spaces of brain parenchyma, but, in this case, it is prion protein (see Table 1). As in Alzheimer's disease and Aβ, the mechanisms underlying the neurotoxicity of the prion protein are also unclear. Recent studies in experimental models suggest that the insoluble prion protein deposits in brain parenchyma are not toxic but that the conversion of the normal prion protein to disease-associated isoforms within neurons results in the neurotoxic form of the protein. Accumulation of the prion protein within blood vessel walls of the brain is less common than occurs with Aβ. However, in some familial prion diseases and in animal models in which the protein has been truncated, prion protein is deposited within walls of the cerebrovasculature. Whether insoluble deposits of prion protein in brain parenchyma and in artery walls have the same effect of impairing interstitial fluid as do Aβ deposits is not yet certain.

Fluid in Subcortical White Matter

Cerebral amyloid angiopathy has an effect not only on gray matter areas in the brain in Alzheimer's disease, but also on the white matter. Accumulation of fluid in subcortical white matter is a frequent finding on MRI in patients with Alzheimer's disease, and this correlates with the severity of cerebral amyloid angiopathy. Blood vessels supplying the subcortical white matter arise from the leptomeningeal arteries on the surface of the brain, and it appears that deposition of Aβ in the walls of the leptomeningeal arteries impedes the drainage of fluid from subcortical white matter.

Genetic Factors in Alzheimer's Disease and Cerebral Amyloid Angiopathy

Familial Alzheimer's disease is associated with defined mutations in the APP and presenilin genes. Polymorphisms in apolipoprotein E (ApoE) are related to the development of cerebral amyloid angiopathy and sporadic Alzheimer's disease. ApoE is an amyloid-scavenging molecule regulating extracellular concentrations of Aβ through ApoE receptor internalization via the endosomal/lysosomal pathway. ApoE–Aβ complexes are transported from the interstitial fluid in the brain into blood through specific ApoE receptors present at the blood–brain barrier. However, co-localization of ApoE with Aβ in brain parenchyma and in artery walls suggests that ApoE–Aβ complexes are also transported with interstitial fluid along perivascular drainage pathways. In this way, ApoE may act as a chaperone molecule for Aβ. ApoE exists in

three isoforms ($\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$); the $\varepsilon 4$ isoform is the most important genetic risk factor for the development of sporadic Alzheimer's disease and has a strong link with cerebral amyloid angiopathy.

Immunotherapy for Alzheimer's Disease

Alzheimer-like transgenic mice have been generated to reproduce Aβ accumulation in brain parenchyma during their adult life in order to understand the mechanisms underlying Aβ deposition and to test treatments that may remove the amyloid deposits. Immunotherapy studies in Alzheimer-like transgenic mice have shown that active immunization against Aβ_{1-42} in young mice prevents the accumulation of Aβ in brain parenchyma, and active immunization of older mice results in a decrease of Aβ in brain parenchyma. Following the success of Aβ immunotherapy in transgenic mice, a clinical trial was initiated in a cohort of patients affected with mild to moderate Alzheimer's disease. Patients were actively immunized against Aβ_{1-42} protein; however, the clinical trial was halted when 6% of the immunized patients developed a complication of what appeared to be meningoencephalitis. Postmortem examination of the first immunized cases shows that insoluble Aβ plaques were removed from cortical areas by the antibody response generated following the immunization and by activated microglia. However, the tau pathology was not modified by Aβ immunotherapy, with the exception of the dystrophic neurites, which disappeared as Aβ plaques were removed.

One of the consequences of Aβ immunotherapy is an increase of the cerebral amyloid angiopathy. It appears, therefore, that removal of insoluble Aβ plaques from brain parenchyma allows access of Aβ to the perivascular drainage pathways. It is not clear whether this is the Aβ removed from the brain by the immunotherapy or whether it is the Aβ from the increased pool of soluble Aβ in gray matter in Alzheimer's disease. Whichever it is, it becomes entrapped in the perivascular drainage pathways. Such an increase in cerebral amyloid angiopathy also results in an increased frequency of microhemorrhages within the brains of the patients immunized against Aβ; this was also observed in experimental transgenic mouse studies. The first year of clinical follow-up of the immunized patients has shown a beneficial effect, with slowing of the decline of cognitive function, and further follow-up studies will allow clinical data and neuropathological features to be correlated. To date, Aβ_{42} immunotherapy has not induced harmful effects in the large majority of patients, and when there have been complications they have not been lethal. A second approach involving passive immunization with specific antibody or peptide, to avoid the apparent meningoencephalitis, is currently in progress. However, it is still not known whether immunotherapy will be most useful as a means to prevent Aβ plaque formation in familial Alzheimer's disease and Down syndrome, or as a therapy in the established sporadic disease.

Conclusions

Aβ is produced by the brain throughout life by the selective cleavage from amyloid precursor protein. The mechanisms for removal of Aβ appear to fail in elderly humans, and if the failure is severe it is associated with Alzheimer's disease. Failure of elimination of Aβ along perivascular interstitial fluid drainage pathways may result from the stiffening of vessels that occurs with age and cerebrovascular disease in humans. If there is an increase in the production of Aβ and particularly the more insoluble form in familial Alzheimer's disease, the system of drainage may become overloaded and impeded at an early age. Understanding the mechanisms by which Aβ is eliminated from the brain, the possible chaperone molecules that are involved, and the exact reasons why elimination of Aβ fails is essential for the planning of future therapies for Alzheimer's disease. Future research may center upon the factors that induce Aβ to attach to vascular basement membranes. Prevention of such attachment may preserve the patency of the pathways for the drainage of interstitial fluid and solutes from the brain.

See also: Aging of the Brain and Alzheimer's Disease; Aging: Extracellular Space; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: Transgenic Mouse Models; Alzheimer's Disease: MRI Studies; Animal Models of Alzheimer's Disease; Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Axonal Transport and Neurodegenerative Diseases; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Dementia; Variant Creutzfeldt–Jakob Disease.

Further Reading

Abbott NJ (2004) Evidence for bulk flow of brain interstitial fluid: Significance for physiology and pathology. *Neurochemistry International* 45: 545–552.

Esiri MM, Lee VMY, and Trojanowski JQ (eds.) (2004) *The Neuropathology of Dementia,* 2nd edn. Cambridge: Cambridge University Press.

Gallagher PJ and van der Wal AC (2007) Blood vessels. In: Mills SE (ed.) *Histology for Pathologists,* 3rd edn. pp. 218–238. Philadelphia: Lippincott Williams & Wilkins.

Herzig MC, Van Nostrand WE, and Jucker M (2006) Mechanism of cerebral beta-amyloid angiopathy: Murine and cellular models. *Brain Pathology* 16: 40–54.

Lue LF, Kuo YM, Roher AE, et al. (1999) Soluble amyloid beta peptide concentration as a predictor of synaptic change in Alzheimer's disease. *American Journal of Pathology* 155: 853–862.

Nicoll JA, Wilkinson D, Holmes C, et al. (2003) Neuropathology of human Alzheimer disease after immunization with amyloid-beta peptide: A case report. *Nature Medicine* 9: 448–452.

Preston SD, Steart PV, Wilkinson A, et al. (2003) Capillary and arterial amyloid angiopathy in Alzheimer's disease: Defining the perivascular route for the elimination of amyloid beta from the human brain. *Neuropathology and Applied Neurobiology* 29: 106–117.

Roher AE, Kuo Y-M, Esh C, et al. (2003) Cortical and leptomeningeal cerebrovascular amyloid and white matter pathology in Alzheimer's disease. *Molecular Medicine* 9: 112–122.

Schley D, Carare-Nnadi R, Please CP, et al. (2006) Mechanisms to explain the reverse perivascular transport of solutes out of the brain. *Journal of Theoretical Biology* 238: 962–974.

Shibata M, Yamada S, Kumar SR, et al. (2000) Clearance of Alzheimer's amyloid-beta (1–40) peptide from brain by LDL receptor-related protein-1 at the blood–brain barrier. *Journal of Clinical Investigation* 106: 1489–1499.

Weller RO (2005) Drainage pathways of CSF and interstitial fluid. In: Kalimo H (ed.) *Pathology and Genetics. Cerebrovascular Diseases*, pp. 50–55. Basel: ISN Neuropath Press.

Weller RO and Nicoll JAR (2003) Cerebral amyloid angiopathy: Pathogenesis and effects on the ageing and Alzheimer brain. *Neurological Research* 25: 611–616.

Weller RO and Nicoll JA (2005) Cerebral amyloid angiopathy: Both viper and maggot in the brain. *Annals of Neurology* 58: 348–350.

Weller RO, Massey A, Newman TA, et al. (1998) Cerebral amyloid angiopathy: Amyloid beta accumulates in putative interstitial fluid drainage pathways in Alzheimer's disease. *American Journal of Pathology* 153: 725–733.

Wilcock DM, Rojiani A, Rosenthal A, et al. (2004) Passive immunotherapy against Abeta in aged APP-transgenic mice reverses cognitive deficits and depletes parenchymal amyloid deposits in spite of increased vascular amyloid and microhemorrhage. *Journal of Neuroinflammation* 1: 24.

Amyotrophic Lateral Sclerosis (ALS)

J M Bhatt and H Mitsumoto, Neurological Institute of New York, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disorder which affects motor neurons in the central nervous system. Its clinical and pathological entity was established by Jean Martin Charcot in 1874. In the United States, it is also known as Lou Gehrig's disease after the famous New York Yankee baseball player who was forced to retire and succumbed to this disease in 1941. The terms ALS and motor neuron disease (MND) are often interchangeably used, particularly in the United Kingdom and Europe. The classification and the clinical limits of the disease remain to be established. In general, motor neuron diseases represent a number of disorders which affect the motor neurons, and ALS is the prototype of all motor neuron diseases and is the most common motor neuron disease in adults. By definition, the first-order upper motor neurons in the premotor cortex and their corticospinal tracts and second-order lower motor neurons of the anterior horn in the spinal cord are affected, resulting in a fairly consistent clinical syndrome, course, and outcome. There are no established etiologies of the motor neuron degeneration as a cause of familial ALS except for the mutation in the super oxide dismutase 1 (SOD1) gene. Current research supports several, perhaps multifactorial, pathogenic mechanisms which may lead to a common path: a progressive motor neuron degeneration resulting in a clinical phenotype common to ALS. Classic ALS affects both upper and lower motor neurons, whereas progressive muscular atrophy affects only lower motor neurons and primary lateral sclerosis (PLS) affects upper motor neurons and the corticospinal tract. A disease variant which exclusively involves lower brain stem motor neurons is often called progressive bulbar palsy, but most cases are in fact bulbar-onset ALS, in which patients develop progressive weakness in muscles in other areas. Attention has focused on the overlap of ALS with other clinical syndromes, particularly frontotemporal dementia. The prevalence of cognitive and behavioral impairment in ALS patients varies from 30% to 50% depending on the study, but it is clearly more common than previously thought. Based on the current knowledge of disease involvement, it is clear that cortical neurons are widely affected, although ALS is still a distinctively motor neuron disease with regard to its clinical expression.

Epidemiology

Several regional studies throughout the world have shown a uniform incidence of ALS: 1–2.4 per 100 000 people with a prevalence of 5–7 per 100 000. The onset of disease occurs between 20 and 80 or more years of age, with a peak onset between 50 and 60 years. For unknown reasons, it is rare in the eldest segment of the population – those older than 85 years. ALS may occur sporadically in the general population or it may be inherited in families. Familial ALS is clinically indistinguishable from sporadic ALS and occurs in 5–10% of all ALS cases. Males are more affected by sporadic ALS than woman by a ratio of 3:2. Epidemiologic studies in industrialized nations show that mortality rates of ALS have increased during the past few decades, but this may be explained by improved case reporting or increased life expectancy.

There are several environmental and lifestyle factors which may predispose patients to ALS. Military service has been identified as a risk factor for ALS in veterans from the Persian Gulf War and among veterans in general across time periods. Studies have shown that physical activity may increase the risk of ALS; one investigation found increased mortality among a large cohort of professional Italian soccer players. Occupational exposures to electromagnetic fields, heavy metals (lead and mercury), and organic solvents have been associated with increased risk for developing ALS. Diets high in glutamate and fats have been shown to have a positive association with ALS, whereas diets rich in antioxidants (e.g., vitamin E) and fiber have demonstrated a possible protective effect. Investigations also demonstrate cigarette smoking as a risk factor for ALS, with mortality rates higher in smoking populations, which may be due to direct toxic injury of chemicals or the formation of free radicals. These environmental toxins and lifestyle habits likely contribute to ALS, along with other pathological mechanisms and genetic susceptibility.

ALS–Parkinsonism–Dementia Complex

In the 1950s, indigenous populations in Guam, the Kii peninsula of Japan, and other western Pacific islands had an astonishingly high incidence of ALS associated with parkinsonism and dementia – up to 100 times the worldwide rate. Since then, the incidence has declined to nearly the worldwide rate,

suggesting a causative environmental factor. At one point, a high dietary intake of an excitotoxin called β-methylamino-L-alanine (BMAA), produced by cyanobacteria from seeds of a cycad tree, was suspected as a potential cause of Guamanian ALS. More recently, consumption of flying foxes, which feed on cycad seeds and thus increase the concentration of BMAA in a process of 'biomagnification' in this animal, is believed to be associated with this unusual edemic ALS–parkinsonism complex. Although this theory is attractive, solid epidemiological evidence is needed. Other proposed environmental agents of western Pacific ALS include low dietary and water consumption of calcium and magnesium in genetically susceptible individuals. The cause of Guamanian ALS–parkinsonism–dementia complex remains to be determined.

Familial or Hereditary ALS

Several novel mutations with different modes of inheritance have been discovered in families with ALS. The most commonly known mutation is the SOD1 mutation, which accounts for approximately 20% of familial ALS. Approximately 20% of familial cases contain a mutation of the Cu/Zn superoxide dismutase (SOD1/ALS1) gene located on chromosome 21q, which is transmitted in an autosomal dominant manner. More than 130 different mutations of SOD1 have been collectively identified in all five of its exons. Mutations occur at more than 65 of 153 possible amino acid positions and are mostly missense point mutations. Another autosomal dominant mutation in the senataxin (ALS4) gene, which transcribes a DNA/RNA helicase, has been associated with a juvenile-onset ALS on chromosome 9q34; average age of onset is 17 years. Autosomal recessive forms of juvenile-onset ALS have been found in a large Tunisian family with mutations of the alsin gene located on chromosome 2q33, which produces a GTPase exchange factor. Interestingly, other mutations in the alsin gene cause a milder PLS phenotype in families from Kuwait. A rare X-linked dominant form has also been identified on the Xp11–12 chromosome. Linkage studies in families with ALS and frontotemporal dementia (FTD) have identified a new locus at 9q21, and efforts are under way to identify a causative gene and its protein. After gene identification, knockout mouse models of ALS–FTD may be developed to investigate the pathogenic mechanisms of this exciting field.

Histopathology

Gross pathology of brains of patients with ALS exhibits atrophy of cerebral cortex in the precentral gyrus and surrounding cortices. The spinal cord typically shows atrophy of anterior roots. Histopathology shows loss of giant pyramidal cells and surrounding motor neurons of the motor cortex, motor neurons in cranial nerve nuclei, and spinal cord anterior horn cells. In addition, there is a loss of myelinated axons in the lateral and anterior corticospinal tracts. There is also diffuse extensive microglial and macrophage proliferation found by CD68 immunoreactive cells in spinal cord and cerebral cortex, suggesting a chronic microglial inflammatory reaction involved in neurodegeneration.

With regard to anterior horn cell pathology, there are a variety of inclusion bodies and evidence of cytoskeletal abnormalities in sporadic and familial ALS. Inclusion bodies may be found in the cytoplasm, such as eosinophilic hyaline bodies called Bunina bodies and ubiquinated skein-like inclusions, which, if found, seem highly characteristic of ALS pathology. Ubiquitin is a protein thought to mark cells for degradation via nonlysosomal proteolysis, and it is found in other neurodegenerative diseases. Phosphorylated neurofilament aggregates called spheroids are found as markedly swollen axons near the anterior horn cell body. In familial ALS, there are unique intracytoplasmic hyaline inclusions found in surviving motor neurons that contain SOD protein, neurofilaments, and peripherin. In general, the earlier and more aggressive the course of the disease, the greater the number of inclusions and spheroids. It is not known whether the various inclusion bodies are directly related to the pathogenesis of ALS or products of another process, yet these inclusions may potentially suggest a mechanistic importance of protein aggregate diseases. Notably, motor neurons controlling ocular motility and small motor neurons in the sacral spinal cord which innervate muscles controlling pelvic floor and bowel and bladder are largely spared in familial and sporadic cases. These neurons differ with regard to calcium metabolism, electrophysiological firing patterns, and androgen receptors.

Pathogenic Mechanisms

There have been several advances in the hypotheses that have been proposed to explain the cause of ALS. One important concept that has been rigorously scrutinized is the role of excessive neurotoxin, primarily involving glutamate and its receptors. ALS patients have an impaired ability to metabolize orally loaded glutamate; observed levels of serum and cerebrospinal fluid glutamate are increased, whereas glutamate levels in the tissue of brain and spinal cord are reduced in ALS patients. The abnormal levels have been linked to faulty glutamate transport proteins that have been observed in synaptosomal preparations from ALS

patients, leading to a breakdown in glutamate absorption from the extracellular space into glial cells. The defective transport raises extracellular levels of glutamate at the synaptic cleft, allowing it to repeatedly stimulate glutamate receptors, specifically its cation channel receptor. The resulting calcium influx leads to a cascade of events including the activation of catabolic enzymes which produce reactive oxygen and nitrogen species that directly damage structural, enzymatic, and receptor proteins, lipid membranes, and DNA.

Significant discoveries have linked oxidative stress to motor neuron cell death via the production of free radicals. This hypothesis was strongly supported by the discovery of mutations in the SOD1 gene which is linked to 20% of diagnosed familial ALS cases. The initial discovery of SOD1 mutations in 1993 was a watershed moment for ALS research because it led to the swift development of transgenic ALS mice. The mutated SOD1 gene was transferred into a mouse genome, producing a readily available animal model. A number of different transgenic mice carrying mutant SOD1 genes are now available and they have undoubtedly broadened the scope and depth of fruitful ALS research. SOD1 is a dimeric protein that catalyzes superoxide radical O_2^- and H^+ to form H_2O_2. Most mutations in the SOD1 molecule do not involve the active site of the enzyme but affect the structure responsible for proper subunit folding or dimer contact. Hence, enzyme function is variable among SOD1 mutations, ranging from normal to absent. This observation suggests that the SOD1 protein is harmful not because of decreased catabolic activity but because of a toxic gain of function. Mutant SOD1 protein is often monomeric and highly unstable, leading to protein degradation and aggregation, which may interfere with axonal transport. Furthermore, mutant protein may cause copper molecules to detach, which may lead to harmful free radical formation. Free radicals are highly reactive species with unpaired electrons that oxidize cellular structures and make them unstable. Long-term accumulation of free radicals, particularly hydroxyl radicals, results in injury to DNA, cell membranes, and cellular protein. The cumulative DNA and cellular membrane damage may pass a threshold leading to programmed cell death.

Another hypothesis is that neuronal degeneration is caused by abnormal inflammatory responses. There is considerable activation and proliferation of central nervous system immune cells called microglia in regions of motor neuron loss and pyramidal tracts. In ALS, microglial cells markedly express proinflammatory cytokines in the spinal cord, such as tumor necrosis factor-α and cyclooxygenase-2. Activated cytokines and chemokines amplify the inflammatory response and recruit proteins that may produce a harmful cascade of events leading to motor neuron death. There is no known cause of the inflammation, and it is not known whether these inflammatory reactions are primary or secondary in nature.

Pathologic examination of motor neurons in ALS patients and in ALS mice models shows prominent protein aggregation in the form of cytoplasmic inclusions. Sometimes, inclusions are found in surrounding astrocytes as well. Such inclusions are a hallmark of several degenerative diseases, including Alzheimer's disease and Parkinson's disease. It is not clear if protein aggregates are directly toxic to motor neurons and their supporting cells, although several theories have been proposed. Most aggregates are intensely immunoreactive with antibodies to ubiquitin, which suggests dysfunction of the ubiquitin–proteasome system that is normally responsible for protein degradation. Proteasomes, which function to degrade damaged protein, may be overwhelmed with indigestible misfolded protein, or other important proteins may not function properly in the presence of the aggregates. Experiments have shows that aggregates affect the operation of folding chaperones, which help other vital proteins attain their appropriate three-dimensional conformations. There is evidence that inclusions may interfere with mitochondrial or peroxisomal function, leading to cellular demise. It is quite possible that protein aggregates are simply a feature of dying cells; this will require further investigation.

Neurofilament dysfunction has been a proposed mechanism for ALS ever since they were identified as pathologic landmarks in sporadic and familial forms. Neurofilaments are important structural proteins that confer motor neurons with their unique axonal diameters and length. Neurofilament mutations have been identified in a small percentage of sporadic ALS patients and in other neurodegenerative diseases, including infantile spinal muscular atrophy and hereditary sensorimotor neuropathy. However, genetic experiments in mice involving overexpression or deletion of neurofilament subunits have slowed the onset of motor neuron disease. In any case, neurofilament organization may be a significant risk factor, if not a direct cause, of motor neuron diseases.

The investigation of programmed cell death, called apoptosis, has shed light on motor neuron degeneration and potential targets for experimental therapy. Studies have shown activation of a family of enzymes called caspases months before neuron death and clinical manifestation of ALS in animal models. Caspases are proteasomes responsible for the breakdown of several cell components in programmed cell death. Inhibition of caspases has been shown to significantly

slow disease onset and increase survival in mice genetically engineered to acquire motor neuron disease. Furthermore, increasing expression of factors that interfere with apoptosis has yielded similar promising results. The challenge for developing treatments that aim to retard the process of cell death is to selectively target corticospinal and motor neurons. Apoptotic proteins are widely expressed in cells outside the central nervous system, and identification of cell death mediators that specifically affect motor neurons will be important for future therapies.

All the aforementioned hypotheses of pathogenesis in ALS lead to a series of events resulting in neurodegeneration of motor neurons. It is still uncertain whether any one the hypothetical mechanisms described previously is the primary or secondary cause; however, it appears that these mechanisms are closely related to each other. Vigorous research efforts are under way.

Clinical Features and Clinical Course

There is strong evidence to support the hypothesis that the process of motor neuron degeneration precedes the appearance of identifiable weakness in patients with ALS. Electrodiagnostic testing in patients diagnosed early in the course of illness shows peripheral denervation and renervation in muscle. Thus, there is a preclinical stage of varying length before the onset of clinical symptoms. Patients usually present with complaints of focal muscle weakness in the arm or hand (arm onset, ~40%), weakness in the foot and difficulty walking (leg onset, ~35%), or problems with speech and swallowing (bulbar onset, ~25%). The localized weakness usually spreads to adjacent muscle groups before moving to other regions of the body, often on the same side. Most often, muscle weakness and atrophy begin in the small hand muscles in an asymmetric fashion, but weakness can be multifocal or proximal in distribution. Patients with familial ALS are approximately 10 years younger than those with sporadic ALS and present with prominent leg weakness, as opposed to arm weakness with sporadic ALS. Despite skeletal muscle paralysis, the ocular muscles, sensorium and bladder, bowel, and autonomic functions are, as a rule, well preserved. Cognitive dysfunction in ALS patients has been scrutinized in recent years because it may be much more common than previously thought. Indeed, as many as 30–50% of patients with ALS may exhibit cognitive symptoms ranging from mild cognitive impairment to executive and language dysfunction and personality and behavioral changes. Specifically, there is compelling clinical and pathologic overlap between motor neuron disease and frontotemporal dementia, which suggests ALS may be a multisystemic syndrome. It is unknown, however, if ALS with FTD is a spectrum of the same disease or if it represents a different disease.

The neurologic exam shows a combination of upper motor neuron signs, including abnormal hyperreflexia, pathological reflexes, and spasticity, and lower motor neuron signs, including muscle atrophy, muscle weakness, hyporeflexia, hypotonia, and fasciculations. Progressive impairment of motor function rapidly results in physical disability. Paralysis in certain muscles causes progressive impairment with speaking, swallowing, and respiration, resulting in severe chronic debilitation. Alveolar hypoventilation usually leads to terminal demise. Severe weight loss is common, sometimes out of proportion to muscle wasting. Fifty percent of patients die within 3 or 4 years of symptom onset, usually because of respiratory failure or aspiration pneumonia. On the other hand, more than 10% of patients may live more than 10 years. Some forms of ALS are slowly progressive over several decades. Hence, the clinical course is more variable than previously suspected.

In some instances, the diagnosis of ALS may be made by virtue of the history and neurologic examination alone. Often, the diagnosis is not readily apparent and further testing is required to rule out other diseases which may mimic ALS. There is no diagnostic test for ALS. Routine laboratory tests should be obtained at the onset of the workup to rule out other metabolic, inflammatory, hematological, immunological, or hormonal abnormalities. Anti-Gm1 antibodies may be sent to rule out a multifocal motor neuropathy, which is a treatable cause of muscle weakness. Nerve conduction studies and electromyography are also essential in distinguishing ALS from other neuromuscular diseases causing progressive weakness. Finally, neuroimaging is important to rule out any structural or infiltrating or neoplastic disorders of the brain and spinal cord that may resemble ALS, such as cervical spondylosis or a spinal cord tumor. If familial motor neuron disease is suspected, then SOD1 mutation analysis is warranted, but it should be remembered that SOD1 mutation is only positive in approximately 20% of patients with familial ALS. A small percentage of sporadic ALS patients are found to have SOD1 mutations. There are diagnostic guidelines known as the El Escorial criteria, named after the location of an important meeting of the World Federation in Neurology in 1990 and thereafter revised, which have set criteria for the clinical diagnosis of ALS. These criteria have not only facilitated making a uniform diagnosis of ALS in clinical trials and other clinical research, but they have also been used in clinical practice.

Management

Since the outcome of the disease is predictable in most cases, the diagnosis of ALS can be devastating to patients and their families. In making the diagnosis, it is important first to exclude all treatable diseases and, if not, all possible definable diseases before concluding that the disease is ALS. Sensitivity, care, and compassion by the physician are essential when discussing a diagnosis of ALS with patients and family caregivers, who should be available for such a discussion. Giving hope, a positive and compassionate attitude, and commitment to care for the patient are vital in such a desperate situation.

The first line of pharmacotherapy for ALS is an antiglutamate agent, riluzole. Two clinical trials that led to the approval of this medication by the Food and Drug Administration for the treatment of ALS showed statistically significant prolongation of survival by approximately 3 months in patients who took the medication. Patients' symptoms or motor dysfunction did not change. In the United States, less than two-thirds of patients take riluzole. Reasons for this limited use are high drug cost and modest benefits. Several retrospective studies following the drug's approval showed that patients who took riluzole lived longer than those who were in the pre-riluzole era or those who did not take riluzole. Prolongation of survival was as long as 12 months. However, the reliability of such retrospective studies is uncertain. Insurance coverage for riluzole is clearly an important factor for physicians who prescribe this drug.

Maximum symtomatic relief is also important in the treatment of ALS. For example, muscle relaxants can be given for spasticity and muscle cramps, anticholinergic or tricyclic antidepressants can control drooling, antidepressants can treat depression, and analgesics may ease pericapsular joint pains. Other management priorities include constipation, insomnia, and laryngeal spasm.

ALS poses a multitude of medical and psychosocial issues. For managing such a disease, patients are best cared for at a special multidisciplinary care center. At a typical ALS center, physical and occupational therapists evaluate the patient's physical impairment; discuss necessary exercises; and may recommend the use of splints, minor devices for assisting impaired manual dexterity, a neck brace, a wheelchair, or other home equipment to improve function in the activities of daily living. Nutritionists watch the patient's nutritional status by carefully monitoring weight, recommending changes in foods to accommodate dysphagia and maintaining necessary caloric needs. Eventually enteral tube feeding may be recommended. Speech therapists help to improve enunciation and swallowing during the early stages of the disease and later assist patients in using alternative communication methods, including communicators or computers. Social workers not only help to find resources for patients but also often provide psychologic support. Respiratory therapists or nurses assess patients' respiratory status to detect early respiratory difficulty. Physicians and neurologists are now paying serious attention to aggressively managing respiratory dysfunction and nutritional impairment. Nurses specializing in ALS play a key role in arranging all services at the clinic and provide psychologic and personal support to patients and family. A number of regionally and nationally known ALS centers provide such services throughout the United States. The Muscular Dystrophy Association and the ALS Association vigorously support such efforts.

At a certain point – ideally during the early stages of the disease – death and life-support issues (advance directives) should be discussed openly but sensitively. If a patient wishes to be sustained on a ventilator, a ventilator team should be consulted at some point. In patients with impending respiratory failure, caring physicians and pulmonologists recommend the use of a noninvasive positive pressure ventilator, which is a respiratory assist device that uses a mouth, nose, or mouth–nose interface. Positive pressure ventilation provides marked relief of respiratory distress and fatigue, helps with sleeping difficulty, and improves overall quality of life. To relieve terminal respiratory distress, pain, and overwhelming anxiety, palliative care is essential. Hospice services have been used at an increasing frequency. The use of morphine sulfate or other sedatives should be considered for those who do not wish to use life-support measures.

Clinical Trials and Emerging Therapies

Clinical trials in ALS can be an important part of patient management because they can provide hope to patients, their families, and even to the treating physicians. Well-designed clinical trials may shed new light on the pathogenesis of ALS, even if they do not confer a survival benefit. There has been substantial growth in the number of compounds suitable for current trials in ALS patients in the past decade. These medications include glutamate antagonists, antioxidants, anti-inflammatory agents, neurotrophic factors, and miscellaneous 'neuroprotective' compounds that have shown proven benefit in the mouse model. Unfortunately, many promising drugs that showed benefit in animal models of ALS have failed in human trials, including creatine, neurotrophic factors, gabapentin and vitamin E, cefecoxib,

topiramate, and pentoxyfilline. As the number of studies has increased, the use of uniform diagnostic criteria, the availability of reliable and reproducible outcome evaluation techniques, and the study design have become critical issues and there appear to be continuous improvements. More potential pharmacological and nonpharmacological agents have become available in recent years. Further investigations are crucial for developing effective therapeutic agents.

There are quite a few exciting translational research advances on the frontier of ALS research that may deliver therapeutic options in the future. One such approach is called messenger RNA interference, which may potentially treat autosomal dominant forms of ALS. In short, double-stranded RNA molecules are designed to specifically target faulty RNA transcripts to prevent abnormal protein formation. This method allows normal alleles to express wild-type protein for the preservation of the cell and has been used successfully in cells expressing autosomal dominant SOD1 protein. At the same time, oligonucleotides that block the mutated SOD1 gene have been developed and used in SOD1 transgenic mice. This treatment does not address the majority of sporadic ALS cases, but it is a promising start. This treatment appears to delay the disease onset but does not prolong survival in animals. Nevertheless, such treatments may become available in patients with SOD1 mutations. There is also a burgeoning field of vaccine trials for neurodegenerative diseases. Experimental compounds that promote an anti-self cell-mediated response without risk for autoimmune disease have been tested as vaccines in mice. This heightened autoimmune response is protective rather than destructive. A compound known to treat multiple sclerosis called glatiramer acetate is being set up for human trials in ALS because it conferred neuroprotection to mice models of parkinsonism as a vaccine. The process of gene therapy, in which healthy genes are transferred into an individual's cells, has been considered for sporadic ALS. However, there are several practical and theoretical challenges that must be overcome. The correct gene must be identified and then delivered to the proper anatomic site in the motor neuron. Potential transfer candidates are the insulin-like growth factor-I gene and vascular endothelial growth factor gene, perhaps using adeno-associated virus. Furthermore, glial cells which support motor neurons nearby are also affected by mutations and are important in the cascade of events leading to degeneration. Another exciting development in ALS is potential therapy using stem cells. Embryonic or neurogenic stem cells may be able to replace lost motor neurons in animals or patients with ALS. Again, practical aspects of targeting and delivery in a hostile surrounding environment must be investigated in order for effective treatment to be developed.

The future of ALS research and therapy does have its challenges, but it holds great promise. One major difficulty is an overblown expectation created by the media publicizing experimental therapies as if they were readily available for human administration. New mutations are being discovered at a rapid rate and will undoubtedly further our understanding of ALS biology. In addition, the pathogenetic mechanisms of ALS are shared by other neurodegenerative diseases, and other ideas may emerge from those avenues. It will take time and patience before all these exciting developments become a reality.

See also: Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Animal Models of Motor and Sensory Neuron Disease; Apoptosis in Neurodegenerative Disease; Axonal Transport and ALS; Excitotoxicity in Neurodegenerative Disease; Infectious Agents in Neurodegenerative Disease; Inflammation in Neurodegenerative Disease and Injury.

Further Reading

Albert SM, Rabkin JG, Del Bene ML, et al. (2005) Wish to die in end stage ALS. *Neurology* 65: 68–74.

Anderson PM (2006) Amyotrophic lateral sclerosis associated with mutations in the CuZn superoxide dismutase gene. *Current Neurology and Neuroscience Reports* 6(1): 37–46.

Bruijn LI, Miller TM, and Cleveland DW (2004) Unraveling the mechanisms involved in motor neuron degeneration in ALS. *Annual Review of Neuroscience* 27: 723–749.

Mitsumoto H and Munsat T (2001) *Amyotrophic Lateral Sclerosis: A Guide for Patients and Families*. New York: Demos Medical.

Mitsumoto H, Przedborski S and Gordon PH (eds.) (2005) *Amyotrophic Lateral Sclerosis*. New York: Taylor & Francis.

Mitusmoto H, Bromberg M, Johnston W, et al. (2005) Promoting excellence in end-of-life care in ALS. *ALS and Other Motor Neuron Disorders* 6: 145–154.

Noonan CW, White MC, Thurman D, and Wong L (2005) Temporal and geographic variation in United States motor neuron disease mortality, 1969–1998. *Neurology* 64: 1215–1221.

Shaw PJ (2005) Molecular and cellular pathways of neurodegeneration in motor neurone disease. *Journal of Neurology, Neurosurgery & Psychiatry* 76: 1046–1057.

Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms

C Vande Velde and D W Cleveland, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS), known in the United States as Lou Gehrig's disease, was first described by the famous French scientist and physician Jean-Martin Charcot in 1869. ALS, the most common adult-onset motor neuron disease, refers to a heterogeneous group of neurodegenerative disorders characterized by the selective loss of upper and lower motor neurons (specialized cells that control movement) in the brain and spinal cord resulting in inevitable paralysis and death. The typical age of onset for most forms of ALS is between 50 and 60 years with an average survival of less than 3 years. This insidious disease is characterized by progressive muscle weakness, atrophy, and spasticity, and is traditionally viewed as lacking in cognitive impairment. However, recent literature documents a variety of cognitive deficiencies in a subset of ALS patients that occur subsequent to neuromuscular deficits. In general, the final fatal event is the loss of the motor neurons that innervate the respiratory muscles and diaphragm. The human impact of ALS is enormous as it significantly affects a patient's quality of life (the loss of speech and swallowing is inevitable). The life time risk for developing ALS is about 1 in 2000. At present, ALS patients suffer knowing there is no cure, and worse – no truly effective treatment exists to slow disease progression.

While 90–95% of all ALS cases lack an apparent genetic linkage, 5–10% are dominantly inherited disease. Of the familial cases, 15–20% are attributed to mutations in the ubiquitously expressed metalloenzyme copper/zinc superoxide dismutase (SOD1) whose endogenous function is to relieve the oxidative stress generated by normal cellular metabolism. At last count, more than 110 different mutations (scattered throughout the 153 possible positions!) have been identified in familial cases of ALS. Regardless of the nature of the mutation, all provoke the age-dependent and selective loss of motor neurons through acquisition of one or more as yet unidentified toxic properties, not through the loss of enzyme activity. The toxic property itself and the biological basis for the selectivity to motor neurons remain unknown. However, the discovery of disease-causing SOD1 mutations has paved the way for the creation of rodent models on which to model the disease, thus providing a tool to tease out the molecular events that trigger and/or modulate the disease and ultimately develop effective therapies. The extensive use of these models in ALS research has yielded six major themes in the study of motor neuron degeneration (Figure 1).

Mechanisms of Motor Neuron Degeneration

Excitotoxicity

Glutamate is an essential molecule which transmits an excitatory signal across a synapse to an awaiting motor neuron. Deregulated or excessive transmission of glutamate can result in a toxic increase in the intracellular calcium concentration within motor neurons. In general, extracellular glutamate is actively cleared from the synaptic space by closely juxtapositioned astrocytes via the activity of a glial-specific glutamate transporter, excitatory amino acid transporter 2 (EAAT2). In the case of ALS patients and rodent models, a focal loss of this particular transporter correlates well with the loss of motor neurons in the anterior horn of the spinal cord and is further corroborated by the expected increased level of glutamate in the fluid which bathes the brain and spinal cord (cerebrospinal fluid). Furthermore, at-risk motor neurons seem to express a reduced amount of calcium-binding proteins, and thus may be less well equipped to handle a wave of glutamate-mediated calcium influx. To date, excitotoxicity remains one of the few mechanistic links between sporadic and SOD1-mediated familial ALS. In fact, the biological basis for the only food and drug administration (FDA) approved ALS treatment, Riluzole (commercially known as Rilutek), is thought to be through the modulation of the glutamate-mediated excitotoxic response. Despite this, Riluzole offers a very limited extension in lifespan (of ~3 months) and a modest delay in the progression of the disease.

Mitochondrial Dysfunction

Mitochondria are the major energy-producing centers within cells and thus largely responsible for the maintenance of cellular metabolism and survival. The motor neuron is an energy-demanding cell – any interference with its mitochondrial energy production is usually detrimental to its survival. The visual observation of abnormal mitochondrial morphology in the motor neurons of ALS patients in early stages of degeneration was the first implication of

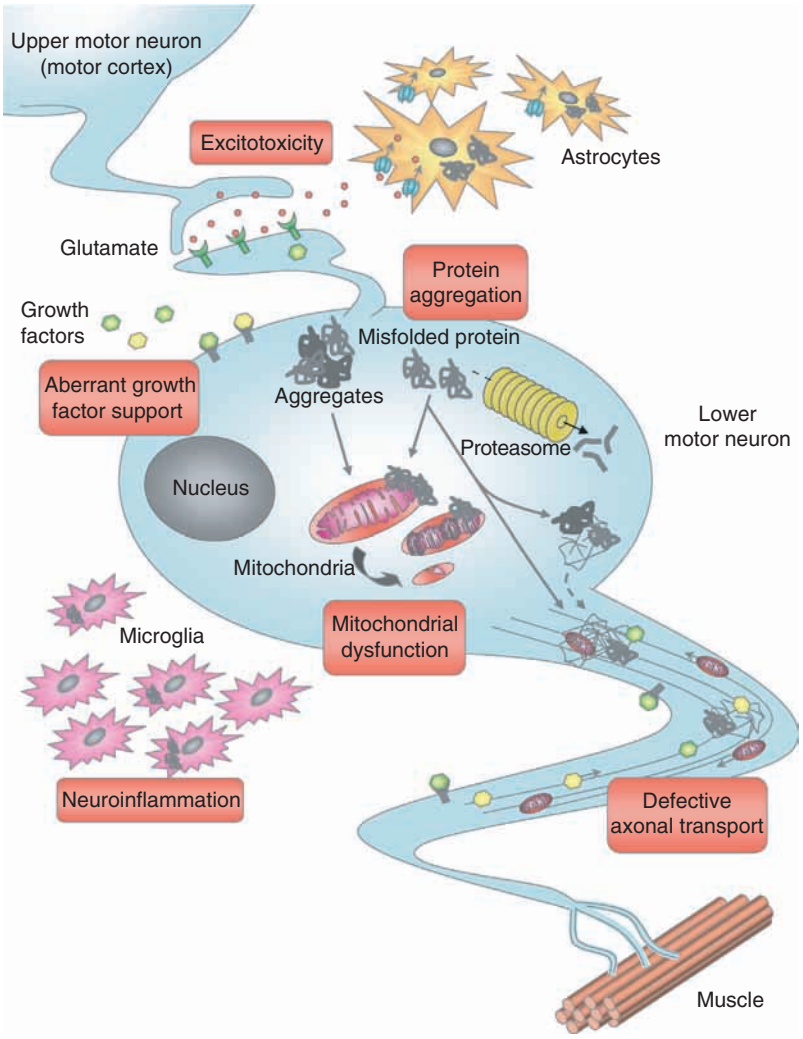

Figure 1 Mechanisms contributing to motor neuron degeneration in ALS.

mitochondria in ALS disease. Similar abnormalities, indicative of mitochondrial damage and dysfunction, were soon identified in rodent models of mutant SOD1-mediated ALS. Using both human (sporadic and familial) cases as well as transgenic models, there has been much focus on components involved in energy production, namely the electron transport chain which drives ATP (the primary energy currency in the cell) generation. While deficits in one or more components of this multicomplex system have been reported, mostly in tissues collected from disease end stage, there has been little agreement between groups on which complex is affected. Moreover, it is unclear whether these defects are causative or merely a consequence of the widespread degeneration of the motor neuron itself that is occurring at this late stage.

The proposal that mitochondrial damage is an initiating event in the failure of motor neurons has gained strong support through the analysis of disease in rodents that express ALS-linked SOD1 mutations. A common feature of all mutant SOD1 proteins (an abundant cytosolic protein) is their preferential association with spinal cord mitochondria purified from affected but not unaffected tissues. This was also observed in samples from familial (SOD1) patients. Moreover, in both cases, the normal endogenous SOD1 was largely excluded from the same spinal cord mitochondria. There is also an intriguing temporal correlation between mitochondrial association and disease progression for the mutant SOD1 proteins. This universal mitochondrial association of various mutant SOD1 proteins may be a critical determinant governing the tissue-specific degeneration of motor neurons (the obligate feature of all ALS cases).

Since optimal mitochondrial function (leading to efficient and sufficient energy production) is central to cellular stability/homeostasis, including the

subsequent proposed mechanisms, understanding how mutant SOD1 proteins impact mitochondrial function is essential and currently of intense interest. Spinal cord mitochondria from presymptomatic mutant SOD1 animals may have a reduced capacity to buffer intracellular calcium fluxes, a concept that fits quite well with the earlier described mechanism of glutamate-mediated excitoxicity. An alternate proposal awaiting confirmation is a proposed direct interaction between mutant SOD1 proteins and BCL-2, an antiapoptotic protein residing on the mitochondrial surface. This proposed interaction may sequester BCL-2, thereby quenching its antiapoptotic activity by preventing BCL-2's interaction with proapoptotic proteins. Since mitochondrial abnormalities are widely reported in sporadic ALS, mitochondrial dysfunction or damage may represent a common pathway in both inherited and sporadic disease.

Protein Aggregation

ALS patient samples typically feature very large intracellular, cytoplasmic inclusions within motor neurons and occasionally in astrocytes. These inclusions are easily identifiable with classic immunohistochemical dyes, but their origin and composition is unresolved. In the mutant SOD1 models, these aggregates are intensely immunoreactive for SOD1 itself, as well as neurofilaments (a structural protein of the motor neuron) and ubiquitin (a component of the protein degradation machinery). Some of these aggregates are detergent resistant and contain misfolded SOD1 and protein folding chaperones (including HSP25 and $\alpha\beta$-crystallin) as well as other proteins including covalent adducts of SOD1 itself. These aggregates occur uniquely in the spinal cords of affected animals, correlate well with disease onset, and may actually be present at the mitochondrial surface.

The prevailing hypothesis is that these large aggregates/inclusions are local enrichments of some toxic species which damage motor neurons. The mechanisms which might contribute to this damage include aberrant chemistry by toxic isoforms of SOD1, co-sequestration of other cellular proteins into these large aggregates (leading to loss of their function), depletion of the protein folding machinery, saturation and eventual disruption of the protein degradation machinery (which has an affinity for misfolded proteins), and interference with mitochondrial function due to aggregation at the mitochondrial surface. The implication of the protein folding pathway is consistent with the co-association of a subset of chaperone proteins, including HSP25, HSP70, HSP40, and $\alpha\beta$-crystallin with SOD1 and SOD1-immunoreactive inclusions. Similar significance can be applied to the proposed defect in protein degradation. Specifically, decreased enzymatic activities of the proteasome, the main organelle involved in protein degradation, has been documented in spinal cord homogenates of mutant SOD1 animals. Finally, it remains controversial as to whether these inclusions are truly damaging. The alternate hypothesis is that these inclusions are in fact protective due to the sequestering of a soluble toxic species. This debate is not unique to ALS, recurring in all of the major examples of human neurodegenerative disease.

Neurofilaments and Axonal Transport

Motor neurons are among the largest and most highly asymmetric cells in nature, with meter long axons generating cell volumes that in humans are 5000 times that of a typical cell. These long processes are specified by an ordered array of neurofilaments, the most abundant structural proteins. Neurofilaments are assembled from neurofilament-L (NF-L), neurofilament-M (NF-M), and neurofilament-H (NF-H). Proper assembly of neurofilaments is essential to the establishment of correct axonal diameters. It is indeed only the largest caliber, neurofilament-rich motor neurons that are most at-risk in all ALS patients (both sporadic and familial) as well as mutant SOD1 rodent models. In the normal motor neuron, neurofilaments form a structural framework that determines axonal caliber (which in turn determines the speed of electrical signal conduction). Furthermore, as discussed earlier, accumulations of neurofilament subunits in motor neuron cell bodies is a frequent feature of ALS cases. In addition, there is some genetic evidence which indicates that mutations in the neurofilament genes may increase the risk of developing ALS. Indeed, mutations in NF-L have already been determined to be causative for one form of motor neuropathy, Charcot–Marie–Tooth disease (type II).

The active movement of cellular proteins and organelles along microtubule tracts in axons, from the cell body all the way to the far-established synapse (and back again), is mediated by ATP-consuming motor proteins. Components traveling anterogradely (toward the synapse) are moved by members of the kinesin family. Defects in the rate of anterograde transport of some components, including SOD1 and neurofilaments, have been documented in mutant SOD1 mice well before the onset of disease. Disturbances in the retrograde movement (from the synapse to the cell body) of cellular constituents, mediated by the single known retrograde motor dynein, have also been implicated in ALS models. In mice, the indirect disruption of dynein function uniquely in postmitotic neurons through disruption of an activation complex (named dynactin) results in impaired

retrograde axonal transport and the eventual development of a late-onset, progressive motor neuron disease that is reminiscent of the ALS mutant SOD1 mouse models. Errors of axonal transport can be a primary cause of late onset motor neuron disease, as demonstrated by point mutations in dynein that provoke motor neuron degeneration in rodents. Interestingly, the introduction of these mutations into mutant SOD1 models leads to a delay in disease onset and extension in lifespan. The molecular basis for this observation is not understood, but perhaps may involve an interplay between the retrograde signaling of toxic and/or trophic factors.

Errors in axonal transport are likely contributors to human ALS. High concentrations of mitochondria and vesicles are found at nerve endings (the neuromuscular junctions where the motor neuron innervates the muscle) in ALS patient samples. Similar distal accumulations have also been reported in mutant SOD1 mouse models. However, what remains unclear is whether these distal accumulations of organelles are due to accelerated anterograde or impaired retrograde transport of these structures. Mitochondria are known to be highly dynamic organelles that are actively transported within cells and through both axonal and dendritic processes of neurons. While the delivery of these organelles is in itself an energy-consuming process, the inefficient delivery of these energy-generating organelles is likely to have significant effects on synaptic activity and neurotransmission, among other previously described cellular processes. The significance of accumulated synaptic vesicles at the nerve terminal remains to be determined but seems to suggest possible problems in synaptic transmission and/or vesicle recycling.

Growth Factor Signaling

Motor neurons require an abundance of both positive and negative signaling molecules to establish and maintain their connections. These signals include those produced by cells in close proximity to their cell bodies (astrocytes and microglia) and those intimately associated with their axons (Schwann cells and oligodendrocytes) and terminals (muscles). The contribution of neighboring cell types to motor neuron degeneration has come to the forefront in ALS research. It is now appreciated that even normal motor neurons (devoid of any disease-causing mutations) are damaged by other nonneuronal cells which do express disease-causing mutant SOD1. This apparent transfer of toxicity to healthy motor neurons may be mediated by a lack of positive trophic signaling from these damage-incurring nonneuronal cells. Candidate molecules that have been implicated include ciliary neurotrophic factor (CNTF), glial cell-line derived neurotrophic factor (GDNF), brain-derived neurotrophic factor (BDNF), and insulin-like growth factor-1 (IGF-1). While these neurotrophins have all been shown to protect motor neurons in culture and/or animal models, all have failed to significantly alter disease in patient trials. However, for IGF-1, CNTF, and GDNF, there are key concerns on whether delivery of appropriate levels of these agents to the nervous system was achieved.

An unexpected involvement of vascular endothelial growth factor (VEGF) has recently been discovered in ALS. VEGF is a growth factor classically thought to participate only in the development of new blood vessels (angiogenesis). Expression of the VEGF gene is regulated by a transcription factor that is itself able to sense low oxygen conditions (hypoxia) and thus upregulate VEGF expression, thereby restoring efficient vascular perfusion in the affected area. Most surprisingly, in a mouse model in which VEGF production was reduced (owing to a mutation in one of its promoter elements which provides normal basal levels of VEGF but does not permit its upregulation in response to hypoxia), a proportion of the mice develop a progressive, late-onset motor neuron disease that is neuropathologically similar to that observed in ALS mice. Furthermore, introducing this mutation of VEGF into mutant SOD1 animals exacerbates the disease. While a firm genetic link of VEGF to ALS pathogenesis remains to be established, an isoform of VEGF has been documented to be neurotrophic for motor neurons grown in culture. Furthermore, the delivery of VEGF whether by viral expression or purified protein into the central nervous system (CNS) of mutant SOD1 animals provides a significant delay in the onset of disease pathology and extends survival. While the exact mechanism involved in VEGF-mediated protection of motor neurons remains to be defined, a clinical trial involving the localized and direct delivery of purified VEGF into the CNS of ALS patients is currently underway.

Neuroinflammation/Glial Activation

It is now appreciated that cell types other than motor neurons contribute to disease initiation and progression. Indeed, the activation of surrounding glial cells, both the trophic-providing astrocytes and the injury-sensing microglia, has been widely reported in both rodent models and patient samples. Microglia share the same ancestry as peripheral tissue macrophages and thus are considered to be the resident immune cells of the CNS. Microglia sense injury in the CNS and rapidly become activated, initiating a neuroinflammatory pathway involving the release of

cytotoxic molecules which affects all neighboring cells, including neurons and astrocytes. Microglial activation correlates well with disease and is observed in human ALS and in all rodent models of ALS and thus is strongly implicated in the disease process. The administration of the FDA-approved antibiotic minocycline to mutant SOD1 mice delays disease progression primarily by preventing microglial activation. Furthermore, the removal of toxic mutant SOD1 from microglial cells limits microglial activation and provides a substantial extension in lifespan in mice which are still expressing the mutant product in all other cell types. This substantial extension in lifespan is actually due to a significant slowing of disease progression after onset. Recognition that microglial activation affects the rate of disease progression offers a significant guidepost to the development of an effective therapy for ALS (as well as other neurodegenerative disease in which neuroinflammation is a component). A clinical trial in which minocycline is included in patient treatment plans is currently underway.

Summary

While many different mechanisms for disease initiation and progression in ALS have been proposed, these should not be seen as mutually exclusive. Many or all are likely to be important contributors that are intricately related to one another. While there may be more than one initiating event occurring within the motor neuron (mitochondrial dysfunction, excitoxicity, protein aggregation, and defective axonal transport), they are likely to converge on one (or at most a few) common final path(s) to motor neuron death. Trophic signaling and neuroinflammation are likely the primary elements mediating disease progression and thus are attractive targets for new therapeutic strategies.

See also: Amyotrophic Lateral Sclerosis (ALS); Axonal Transport and ALS; Axonal Transport and Neurodegenerative Diseases; Excitotoxicity in Neurodegenerative Disease; Inflammation in Neurodegenerative Disease and Injury; Neurofilaments: Organization and Function in Neurons; Oxidative Damage in Neurodegeneration and Injury; Transgenic Models of Neurodegenerative Disease.

Further Reading

Bachman SR, Bradley WG, and Moraes CT (2006) Mitochondrial involvement in amyotrophic lateral sclerosis: Trigger or target? *Molecular Neurobiology* 33: 113–131.

Boillée S, Yamanaka K, Lobsiger CS, et al. (2006) Onset and progression in inherited ALS is determined by motor neurons and microglia. *Science* 312: 1389–1392.

Bruijn LI, Miller TM, and Cleveland DW (2004) Unraveling the mechanisms involved in motor neuron degeneration in ALS. *Annual Review of Neuroscience* 27: 723–749.

Clement AM, Nguyen MD, Roberts EA, et al. (2003) Wild-type nonneuronal cells extend survival of SOD1 mutant motor neurons in ALS mice. *Science* 302: 113–117.

Cleveland DW and Rothstein JD (2001) From Charcot to Lou Gehrig: Deciphering selective motor neuron death in ALS. *Nature Neuroscience Reviews* 2: 806–819.

Strong MJ, Lomen-Hoerth C, Caselli RJ, Bigio EH, and Yang W (2003) Cognitive impairment, frontotemporal dementia, and the motor neuron diseases. *Annals of Neurology* 54: S20–S23.

Relevant Website

http://www.alsa.org – The ALS Association.

Angelman Syndrome

L M Bird, Rady Children's Hospital and University of California, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Most autosomal genes are expressed from both parental copies (biallelic expression), regardless of the parental origin. A small number of genes are expressed from only one parent's allele, a phenomenon known as genomic imprinting. Angelman syndrome (AS) is one of the most well-understood syndromes of genomic imprinting. AS maps to the imprinted gene cluster on chromosome 15q11–q13 and is due to defective expression of the maternally inherited ubiquitin protein ligase 3A gene (*UBE3A/E6-AP*). There are at least four molecular classes of AS, demonstrating the genetic complexity of this disorder. Though AS is a monogenic disorder, it is a prototype for the so-called mixed epigenetic and genetic and mixed *de novo* and inherited (MEGDI) model of disease. Herein follows a summary of the progress made to date in elucidating the mechanisms of imprinting in AS.

Clinical Background

The clinical features of AS are microcephaly, functionally severe mental retardation, ataxic gait, seizures, and electroencephalographic abnormalities. The behavioral profile of AS includes exceptionally happy disposition with episodes of laughter; excessive mouthing of objects, drooling, and tongue thrusting; easy excitability and hypermotoric activity; reduced need for sleep, with sleep disturbance; and an affinity for water.

Seizures begin in infancy or early childhood in the majority of patients, and often show attenuation in adolescence. A wide variety of seizure types have been reported. Movement disturbances (tremors, jerkiness, poor coordination) contribute to the delayed acquisition of motor skills (sitting after 12 months; walking between 2 and 6 years). The incidence of nonambulation is said to be 10%, but this may no longer be true in the modern era of earlier diagnosis and prompt intervention. The early institution of physical therapy may change the natural history of scoliosis (said to occur in 50%) by improving truncal tone.

Virtually all AS patients exhibit a short attention span, and most are hypermotoric. Some aggressive, unwanted behaviors are displayed by the majority of patients, including biting, pinching, hair-pulling, and grabbing. Rarely are these behaviors intended to cause harm; they usually result from easy excitability, poor control over movements, reduced repertoire of need expression, and occasionally frustration over an inability to communicate effectively. Paroxysms of laughter are said to occur in AS, but laughter is rarely unprovoked, though often inappropriate to the triggering stimulus. A happy disposition characterizes most children, but irritability is common in infancy. Gastrointestinal difficulties such as gastroesophageal reflux may contribute to this irritability. AS patients have an apparent reduced need for sleep (5–6 h per night) and many have long or frequent periods of wakefulness during the night. With behavior modification and medication, these difficulties can be overcome in most patients, and sleep patterns improve with age.

Cognitive performance is usually in the range of severe functional impairment. The combination of deficits exhibited by individuals with AS make the commonly used developmental assessment tools difficult to apply, and these tests probably underestimate the abilities of AS children. Receptive language is far superior to expressive language, which is often completely absent. Nonverbal communication using a variety of systems (picture exchange cards; communication devices) is possible in a substantial proportion of AS patients. Poor fine motor coordination usually precludes the use of sign language.

AS adults are not able to live independently, but most are able to using feeding utensils and perform some household tasks. Dressing skills are dependent on the degree of fine motor difficulty. Daytime continence is possible with prompted voiding and habit training. Life expectancy appears to be normal.

Molecular Genetics of AS

Failure of maternally inherited *UBE3A* to be transcribed and translated into a functional protein causes AS. Four molecular mutation classes of AS have been defined: deletion, uniparental disomy (UPD), imprinting defect, and *UBE3A* mutation.

Deletion

Approximately 65% of AS cases exhibit a *de novo* (~4 Mb) microdeletion of chromosome 15q11–q13. Two major classes of deletions exist, one from breakpoint 1 (BP1) to breakpoint 3 (BP3) (class I) and the other from breakpoint 2 (BP2) to BP3 (class II) (see Figure 1). Rarely, patients may have a larger deletion that extends beyond the boundaries of BP1 and BP3 in one or both directions. Deletions are apparently mediated by homologous misalignment and meiotic recombination between low-copy-number repeats

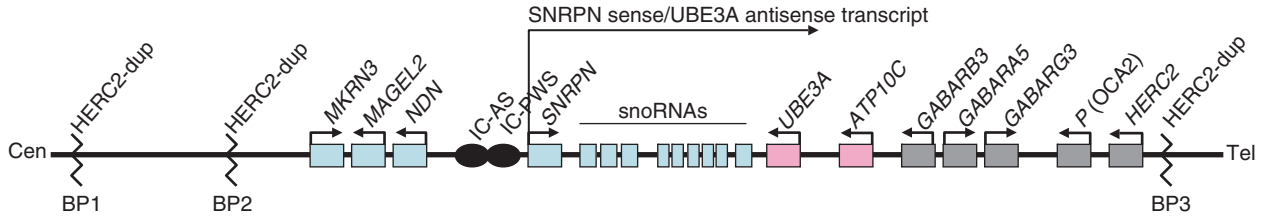

Figure 1 Genomic organization of human chromosome 15q11–q13. Blue loci are paternally transcribed, pink loci are maternally transcribed, and the gray loci are biparentally expressed. Arrows indicate direction of transcription. Black ovals denote bipartite imprinting center (IC-AS, IC-PWS). Cen, centromere; Tel, telomere; BP1, BP2, BP3, common break points. Adapted from Nicholls RD and Knepper JL (2001) Genome organization, function and imprinting in Prader–Willi and Angelman syndromes. *Annual Review of Genomils and Human Genetics* 2: 153–175.

(duplicons) that have been identified in proximal and distal 15q11–q13. These duplicons arose with the amplification of an ancestral gene, *HERC2*. Duplicons having 90–99% identity to the first 79 exons of *HERC2* are found in at least nine copies in 15q11–q13 as well as two copies on chromosome 16. There is one reported case of maternal germ line mosaicism producing familial recurrence of a deletion.

Uniparental Disomy

Paternal UPD accounts for approximately 5% of cases of AS, and in almost all cases, the UPD is isodisomic. The most likely origin of this event is maternal nondisjunction producing a monosomy 15 conception, with postzygotic rescue by duplication of the paternal chromosome 15. As in the deletion cases, this class of AS represents *de novo* mutational events and has a very low risk for recurrence.

Imprinting Defect

In approximately 10% of cases, an imprinting defect underlies AS. In this class of AS, a paternal imprint is erroneously assigned to the maternally inherited allele. Two types of imprinting defects are known: those due to a submicroscopic deletion of the imprinting center (IC) and those with no detectable mutation. Most submicroscopic IC deletions are familial, carrying a 50% risk for recurrence. Thus far all imprinting defects with undetectable mutations have been sporadic events. They are presumed to result from failure to establish or maintain the imprint during oogenesis, due to a stochastic event or perhaps (as yet unknown) environmental factors.

UBE3A Mutation

UBE3A was initially discarded as a potential candidate for the AS gene because it appeared not to be imprinted when studied in lymphocytes and fibroblasts, and its widespread expression ran counter to expectation since the AS phenotype is exclusively neurological. Establishing brain-only imprinting of *Ube3a* in mice resurrected *UBE3A*'s status as a candidate gene for AS. Analysis of *UBE3A* in AS patients with biparental contribution to 15q11–q13 and no imprinting abnormalities showed point mutations in several unrelated patients, identifying it as the gene responsible for the AS phenotype. Intragenic *UBE3A* mutations (insertion, deletion, nonsense, missense, and splice site mutations) cause approximately 10% of cases of AS. Substantial portions of *UBE3A* mutations are inherited from the mother's paternally acquired allele. In this circumstance, a 50% recurrence risk pertains.

Negative Molecular Studies

In addition to the four molecular classes described, approximately 10% of cases of AS have the clinical profile but no discernible molecular defect. Whole-gene or large intragenic deletions of *UBE3A* are not detected by the clinically available diagnostic tests (DNA methylation analysis of chromosome 15q11–q13 and *UBE3A* sequence analysis). Intragenic deletions have been reported in two families, but do not appear to be the explanation for the bulk of cases of AS with negative molecular studies. While some patients with the AS phenotype and negative molecular studies may represent misdiagnoses, it is likely that other molecular mechanisms affecting *UBE3A* expression will be found to explain this class of patients.

Molecular Biology of AS

UBE3A spans 120 kb of genomic DNA. Mutations have been detected throughout all regions of the gene, and there do not appear to be mutational hot spots. Sixteen exons have been identified; additional exons at the 5′ end of the gene are possible, as this region displays alternative splicing. Transcription of *UBE3A* occurs in a telomeric-to-centromeric direction. RNA transcripts of 5–6 kb include approximately 2 kb of 3′-untranslated region (UTR) sequence. There are at least seven mRNA species produced by alternative splicing at exons 5 and 6, at the 5′ end, and possibly at the 3′ end. The various transcripts produce three

isoforms of UBE3A protein. Isoforms II and III have 20 and 23 additional amino acids, respectively, at the N-terminus, compared with isoform I. The function of the different isoforms of UBE3A remains unknown. There are tissue-specific variations in RNA splicing and isoform predominance, but the significance of these differences is unclear.

UBE3A encodes an E3 ubiquitin ligase, E6-AP (E6-associated protein). Though first characterized by its interaction with the E6 protein of papillomaviruses to promote degradation of p53, E6-AP does not maintain a stable association with p53. It participates in the ubiquitin pathway of protein degradation in proteosomes. In this pathway, ubiquitin is covalently attached to one or more lysine residues of proteins destined for destruction. Ubiquitination involves a three-step process (activation of ubiquitin by an E1 enzyme, transfer to an E2 conjugating enzyme, and covalent ligation of ubiquitin to the protein substrate by an E3 ligase), the third step of which is deficient in AS. The C-terminus of UBE3A is a functionally important and highly conserved domain that is shared by a family of proteins (hect domain, so named because it is homologous to the E6-AP carboxy terminus). The last six amino acids of the E6-AP C-terminus are essential for activity *in vitro*. It is unknown how substrate specificity is determined. E6-AP is found in all tissues that have been studied.

Clinical Difference in Classes of AS

Clinical differences between the molecular classes of AS have been recognized. As a group, patients with deletions tend to have a more severe phenotype (later onset of independent walking, earlier onset and increased severity of seizures, and complete absence of speech) compared to patients with UPD, imprinting defect, or *UBE3A* mutation. Though speech is usually absent in deletion AS patients, use of up to 20 words is reported in AS patients with other molecular classes of AS. The contribution of other genes in the deleted region to the increased phenotypic severity is not clear. Three genes for γ-aminobutyric acid (GABA) receptor subunits (*GABRB3*, *GABRA5*, and *GABRG3*) are located telomeric of *UBE3A* and are contained within the common class I and class II deletions. Though these genes show biallelic expression, a role for them in the genesis of epilepsy in AS has not been excluded.

Hypopigmentation occurs in some AS deletion patients. The (unimprinted) *P* gene located telomeric within the commonly deleted region is responsible for autosomal recessive oculocutaneous albinism type 2 (OCA2). Individuals with AS and OCA2 have been reported, the mechanism being a maternal chromosome deletion and paternal *P* gene mutation. Semidominant behavior of the *P* gene product (expression of a hypomorphic allele) has been offered as a potential explanation for the hypopigmentation seen in AS deletion patients. However, hypopigmentation is reported in patients with other classes of AS, including siblings with a maternally inherited intragenic deletion of exons 8–16 of *UBE3A*, so the contribution of the *P* gene to hypopigmentation remains speculative.

Within the deletion class of patients, those with larger class I deletions have lower cognitive scores and are more likely to merit a comorbid diagnosis of autism. Patients with atypical (larger) deletions are usually lower functioning.

Mechanisms of Genetic Imprinting

Genetic imprinting confers functional differences onto the genome such that expression of imprinted genes occurs only from the father or the mother. The mechanisms by which an imprint is established in the germ line, maintained during development and postnatal life, and reversed in the germ line during the subsequent generation are unknown. Additional complexities, such as tissue specificity of imprinting and changes in imprint with age, remain to be elucidated.

Mechanisms to control gene expression include insulators (DNA sequence elements that prevent interaction between nearby chromatin domains or prevent the spread of heterochromatin); transcriptional enhancer competition (promoters of linked imprinted genes competing for access to enhancers); modification of chromatin structure by histone acetylation, phosphorylation, and methylation; and DNA methylation. Each of these mechanisms provides a layer of information beyond what is contained within the nucleotide sequence, and thus contributes to the epigenotype. Imprinted genes cluster together in specific areas of the genome, an observation that suggests coordinate regulation by a regional element, known as the imprinting control region (ICR).

DNA methylation, the modification of DNA by the addition of a methyl group to the cytosine base of a CpG dinucleotide, is the most well studied of these epigenetic mechanisms. DNA methylation occurs in most, but not all, imprinted genes. Loss of DNA methylation occurs during a specific period of germ cell development, and is thought to represent erasure of the imprint from the previous generation. Reacquisition of DNA methylation occurs differentially during spermatogenesis and oogenesis, conferring the parent-of-origin specific imprint or epigenotype. It is not clear whether DNA methylation establishes the imprint or maintains it, or both.

Less is know about control of gene expression through covalent modification of histones and their association with DNA. In general, acetylation of histones H3 and H4 is associated with increased gene accessibility (and therefore increased expression). Methylation of Lys4 of H3 correlates with active gene expression; methylation of Lys9 of H3 usually accompanies gene silencing. These chemical modifications of histones attract effector proteins that apparently mediate the increased or decreased gene activity. It is not clear if DNA methylation directs the histone modifications, or if histone modification drives DNA methylation; it is possible that reciprocal interactions reinforce the effect of each.

Imprinting of the Prader–Willi Syndrome/ AS Region

Both Prader–Willi syndrome (PWS) and AS are caused by defects of the imprinted gene cluster on chromosome 15q11–q13 (see Figure 1). Loss of expression of maternally derived *UBE3A* causes AS, with the clinical phenotype as described in the preceding sections. Loss of expression of paternally derived gene(s) from 15q11–q13 causes PWS. PWS is characterized by hypotonia and failure to thrive as a neonate, small hands and feet, short stature, hypogonadism, childhood-onset hyperphagia and obesity, and cognitive impairment. Whether a single gene or several genes are required to produce the PWS phenotype is still a matter of debate; that no PWS patient with a single gene defect has been reported suggests that PWS is truly a contiguous gene syndrome.

Differential methylation of chromosome 15q11–q13 has been thoroughly studied. Whether methylation directly or indirectly effectuates the imprint of genes within the region continues to be argued and investigated. Differential methylation provides the basis for the diagnostic testing used to confirm a suspected diagnosis of AS or PWS. The highly methylated maternal chromosome 15q11–q13 region can be distinguished from the (largely) unmethylated paternal contribution by Southern blot analysis or polymerase chain reaction (PCR) assay of the promoter region of *SNRPN*. Maternal-only contribution (due to paternal deletion, maternal UPD, or imprinting defect) is diagnostic for PWS. Exclusively paternal contribution is diagnostic for AS. A normal methylation profile does not rule out AS, however, since a substantial fraction of AS is due to maternally inherited *UBE3A* mutations.

The 15q11–q13 cluster is under the bidirectional control of the imprinting center (IC), the designation for the imprinting control region in this area. The IC was defined by the identification of small submicroscopic deletions in a subgroup of PWS and AS multiplex families with biparental inheritance yet uniparental imprint. The IC is thought to regulate in *cis* the establishment and maintenance of the imprint for the entire cluster. The IC has a bipartite structure, the PWS-IC and the AS-IC. All AS patients with IC deletions are missing the AS-IC (880 bp located 35 kb centromeric of PWS-IC). The function of the PWS-IC is to establish and maintain paternal gene expression. During oogenesis, the AS-IC functions to negatively regulate the PWS-IC (through CpG methylation) and prevent the paternal imprint from being established. Deletion of the AS-IC produces a paternal imprint of the entire 15q11–q13 region; when this is transmitted through the maternal germ line, AS results. A unique inversion case has shown that the PWS-IC and AS-IC must be closely and correctly related in order for the maternal imprint to be established.

MKRN3, *NDN*, and *SNURF-SNRPN* are expressed only from the paternal chromosome, where their promoter regions are unmethylated. *MAGEL2* is expressed only from the paternal chromosome, but its methylation status has not been clarified. Between *SNURF-SNRPN* and *UBE3A*, there are more than 70 snoRNA genes. These are noncoding genes that produce small nucleolar RNAs, the main function of which is to modify ribosomal RNA. Though not imprinted directly by methylation, these snoRNAs are imprinted indirectly because they are processed from the paternally expressed *SNURF-SNRPN* sense/ *UBE3A* antisense transcript (discussed in the next section), which is under the control of methylation.

In sporadic imprinting defect cases of AS, the maternal grandfather's and maternal grandmother's chromosomes are equally likely to harbor the unmethylated paternal imprint, suggesting that erasure of the grandmaternal methylation imprint occurs normally, followed by failure to establish (in the case of the grandpaternal chromosome) or reestablish (in the case of the grandmaternal chromosome) the correct maternal imprint.

Much of what we know about methylation status and imprinting of the chromosome 15q11–q13 region comes from investigation of the orthologous region on mouse chromosome 7C. However, important differences exist. The *Frat3* gene is an evolutionarily younger gene which has joined the mouse PWS/ AS region (by retrotransposition) and has acquired the differential methylation and expression of paternally expressed genes in the region. There is no homolog to *Frat3* in the human chromosome 15q11–q13 region. Transgenic studies in which human elements of the PWS/AS region are inserted in the mouse have shown that both the positive and negative regulatory elements of the imprinting machinery have diverged. The mouse equivalent of the human AS-IC has yet to

be identified. Elucidating the mechanism of mouse PWS/AS region imprinting is likely to yield important insights, but this information may not be directly applicable to human *UBE3A* imprinting.

In addition to methylation, histone modifications play a role in establishing and maintaining the imprint. Hyperacetylation of histones H3 and H4, a modification usually signifying gene activation, is found at the PWS-IC on the paternal allele. Methyl-Lys4 histone H3, which is typically found at transcriptionally active loci, associates with PWS-IC and the promoters of *SNRPN* and *NDN* on the paternal allele. Dimethyl-Lys9 histone H3, which is generally found in inactive or heterochromatic regions, associates with the PWS-IC on the maternal chromosome.

Gene Repression of *UBE3A*

The IC is believed to regulate the imprint of 15q11–q13 through establishing or maintaining the methylation profile of the region. However, no site of differential methylation has been reported within the *UBE3A* gene or its vicinity. Though an as-yet-undiscovered site of allele-specific methylation in or around *UBE3A* remains a formal possibility, other mechanisms for imprint regulation are being proposed and evaluated. These include enhancer competition, chromatin state modification, and expression competition by antisense RNA. At present, there is no evidence for chromatin state modification by histone acetylation or methylation. The expression competition model proposes that the IC controls the methylation of an imprintor gene (distinct from the IC). The imprintor gene would be paternally expressed (in a tissue-specific manner) and would 'compete out' expression of *UBE3A*.

There is some evidence to support the expression competition model. A large (>500 kb) transcript oriented in an antisense manner and spanning the 3′ half of *UBE3A,* as well as the region downstream of *UBE3A*, has been identified. This antisense transcript (*UBE3A-AS*) is expressed only from the paternal allele in the brain, is not expressed in tissues with biallelic *UBE3A* expression, and is under the control of the PWS-IC. In this model, *UBE3A-AS* is the direct target of the IC and is imprinted primarily (in brain), and *UBE3A* is secondarily imprinted (in brain) through a repression mechanism. Based on other examples of imprinting regulation by antisense transcripts (e.g., *XIST*, *IGF2R*), *UBE3A-AS* is presumed to silence *UBE3A* by annealing in *cis* to the sense promoter. However, the general concept of *cis* antisense regulation of neighboring imprinted genes remains controversial because of conflicting evidence.

Recent evidence from the mouse indicates that *Ube3a-as* may be regulated by, rather than being a regulator of, *Ube3a*. Using a mouse model of AS generated by targeted disruption of exons 12 and 13 of *Ube3a*, Landers et al. observed that maternal transmission of the defective allele upregulated expression of *Ube3a-as* from the paternal allele. This evidence suggests that *Ube3a* expression, rather than being regulated by *Ube3a-as*, modulates the levels of the *Ube3a-as* and that this interaction occurs in *trans*. The mechanism of this modulation is unknown. Investigating the directionality of the interaction between *UBE3A* and *UBE3A-AS* in the human brain will be difficult given the scarcity of this resource.

Tissue-Specific Imprinting

UBE3A imprinting is tissue specific, in contrast to most other imprinted loci in the 15q11-q13 region. In the brain, there is preferential expression of *UBE3A* from the maternal allele, but a small amount of residual transcription (10%) from the paternal allele occurs. Another gene, *ATP10C*, is maternally expressed in a tissue-specific manner, but so far has not been shown to contribute to the AS phenotype.

The tissue-specific imprinting of *UBE3A* appears to be region specific, with the imprinted expression pattern being found in the hippocampus and cerebellum, and somewhat in the cerebral cortex. This regional specificity has been shown to be due to lineage specificity: *UBE3A* is expressed exclusively from the maternal allele only in neurons, and is biallelically expressed in glial cells and fibroblasts. Because the hippocampus and cerebellum are neuron-rich regions, these regions exhibit the imprint most strikingly.

The molecular basis of this neuron-specific imprinting is unknown. Two hypotheses have been put forward: (1) the epigenetic marks of neurons are the same as those in other tissues, but are interpreted differently by the neuronal transcription machinery; or (2) a specific additional epigenetic mark, a secondary imprint, is acquired during neuronal differentiation.

Summary

AS is a unique neurobehavioral disorder due to defective expression of maternally inherited *UBE3A*. The complex mechanisms of how the *UBE3A* imprint is established, maintained, and reversed remain to be elucidated, as does the mechanism of tissue specificity. *UBE3A* repression apparently does not involve differential methylation, but possibly is mediated by expression competition from an antisense transcript.

See also: Prader–Willi Syndrome; Ubiquitin–Proteasome System and Plasticity.

Further Reading

Boyes L, Wallace AJ, Krajewska-Walasek M, et al. (2006) Detection of a deletion of exons 8–16 of the UBE3A gene in familial Angelman syndrome using a semi-quantitative dosage PCR based assay. *European Journal of Medical Genetics* 49: 472–480.

Clayton-Smith J and Laan L (2003) Angelman syndrome: A review of the clinical and genetic aspects. *Journal of Medical Genetics* 40: 87–95.

Fridman C, Hosomi N, Varela MC, et al. (2003) Angelman syndrome associated with oculocutaneous albinism due to an intragenic deletion of the *P* gene. *American Journal of Medical Genetics* 119A: 180–183.

Guerrini R, Carrozzo R, Rinaldi R, et al. (2003) Angelman syndrome: Etiology, clinical features, diagnosis, and management of symptoms. *Pediatric Drugs* 5: 647–661.

Horsthemke B and Buiting K (2006) Imprinting defects on human chromosome 15. *Cytogenetic and Genome Research* 113: 292–299.

Jiang Y, Bressler J, and Beaudet AL (2004) Epigenetics and human disease. *Annual Review of Genomics and Human Genetics* 5: 479–510.

Jiang Y, Tsai T, Bressler J, et al. (1998) Imprinting in Angelman and Prader–Willi syndromes. *Current Opinion in Genetics & Development* 8: 334–342.

Johnstone KA, DuBose AJ, Futtner CR, et al. (2006) A human imprinting centre demonstrates conserved acquisition but diverged maintenance of imprinting in a mouse model for Angelman syndrome imprinting defects. *Human Molecular Genetics* 15: 393–404.

Kishino T (2006) Imprinting in neurons. *Cytogenetic and Genome Research* 113: 209–214.

Landers M, Calciano MA, Colosi D, et al. (2005) Maternal disruption of *Ube3a* leads to increased expression of *Ube3a-ATS* in *trans*. *Nucleic Acids Research* 33: 3976–3984.

Lossie AC, Whitney MM, Amidon D, et al. (2001) Distinct phenotypes distinguish the molecular classes of Angelman syndrome. *Journal of Medical Genetics* 38: 834–845.

Nicholls RD and Knepper JL (2001) Genome organization, function and imprinting in Prader–Willi and Angelman syndromes. *Annual Review of Genomics and Human Genetics* 2: 153–175.

Nicholls RD, Saitoh S, and Horsthemke B (1998) Imprinting in Prader–Willi and Angelman syndromes. *Trends in Genetics* 14: 194–200.

Rougeulle C and Lalande M (1998) Angelman syndrome: How many genes to remain silent? *Neurogenetics* 1: 229–237.

Sahoo T, Peters SU, Madduri NS, et al. (2006) Microarray based comparative genomic hybridization testing in deletion bearing patients with Angelman syndrome: Genotype–phenotype correlations. *Journal of Medical Genetics* 43: 512–516.

Soejima H and Wagstaff J (2005) Imprinting centers, chromatin structure, and disease. *Journal of Cellular Biochemistry* 95: 226–233.

Angiotensin Actions on and within Brain

M J McKinley and A M Allen, Howard Florey Institute, University of Melbourne, Melbourne, VIC, Australia
B J Oldfield, Monash University, Clayton, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Overview

As well as having important actions on peripheral organs and the circulation, the renin-angiotensin system exerts influences on the central nervous system (CNS) via its effector peptides angiotensin II, III, and IV. Many of these actions, both central and peripheral, relate to maintaining the integrity of the cardiovascular system and body fluid and electrolyte homeostasis. They include peripheral vasoconstriction; stimulation of secretion of aldosterone, vasopressin, and adrenocorticotropic hormone (ACTH); inhibition of renin release; potentiation of sympathetic nerve activity; and stimulation of thirst and salt hunger. Intense study of these actions of angiotensin on both CNS and peripheral organs during the past 50 years has yielded much detailed knowledge of the renin-angiotensin system and its biological actions. In several aspects, however, understanding of this system is still incomplete.

Angiotensin II is generated both peripherally and within the CNS. While the peripherally generated peptide does not readily traverse the blood–brain barrier, it is still able to influence brain function by acting on specific angiotensin receptors in brain sites lacking this barrier: the circumventricular organs. Specific angiotensin receptors are also located in many other parts of the CNS that are inaccessible to circulating angiotensin because of its exclusion by the blood–brain barrier. Centrally generated angiotensin can act at these sites. Thus, in considering actions of angiotensin on the brain, it is necessary to recognize that the peptide products of separate angiotensin generating systems (one peripheral and the other central) can influence CNS function.

The receptors to which angiotensin II binds to exert its biological actions are classified into two subtypes, AT_1 and AT_2. Most of the known actions of angiotensin II on the brain are mediated by the AT_1 receptor, and it is widely distributed in discrete areas of many brain regions. High concentrations of AT_2 receptors occur in the fetal brain, suggesting that these receptors play a developmental role in the CNS. In the adult, the distribution of AT_2 receptors is more restricted; it is predominantly associated with regions involved in sensory processing.

Both AT_1 and AT_2 receptors have been cloned and belong to the 7-transmembrane-spanning G-protein-coupled receptor family. While it can associate with several G-proteins, the AT_1 receptor predominantly couples through $G_{q/11}$, leading to activation of inositol phosphate signaling pathways and elevated intracellular calcium. In addition, the AT_1 receptor is known to signal through G-protein-independent pathways such as production of reactive oxygen species via nicotinamide adenine dinucleotide phosphate oxidase. In neurons, angiotensin II-induced activation of AT_1 receptors leads to excitation, predominantly via inhibition of various outward cation conductances. The AT_2 receptor predominantly couples through G_i, leading to stimulation of phospholipase A_2 and production of arachidonic acid. In neurons this signaling also appears to modulate K^+ channel activity, but both increases and decreases in firing rates have been observed.

Circulating Angiotensin and Its Actions on the Brain

Generation of Circulating Angiotensin

Circulating angiotensin II is the product of the sequential actions of two enzymes, renin and angiotensin converting enzyme, also known as kininase II. Renin, an enzymatic hormone synthesized within the kidney, is released into the circulation from the renal juxtaglomerular apparatus in response to stimuli such as sodium depletion, reduced renal perfusion pressure, and increased renal sympathetic nerve activity. While circulating in the bloodstream, renin cleaves a decapeptide from the terminus of a large circulating plasma protein, angiotensinogen (synthesized by the liver). This decapeptide, known as angiotensin I, is the biologically inactive precursor of angiotensin II. Angiotensin I is then converted into the biologically active octapeptide angiotensin II by the action of angiotensin converting enzyme, located on cell membranes of vascular endothelial cells in several tissues, most abundantly in the lungs (see Figure 1). Once formed within the circulation, angiotensin II is able to reach regions of the brain that are devoid of a blood–brain barrier, the circumventricular organs, and act on its receptors there.

The three circumventricular organs at which angiotensin II from the circulation acts are the subfornical organ and organum vasculosum of the lamina terminalis (OVLT) in the forebrain and the area postrema in the hindbrain (see Figure 2). While both classes of angiotensin receptors (AT_1 and AT_2) are present within these three circumventricular organs, AT_1 receptor density is extremely high there. Studies using specific antagonists show that the responses (thirst, sodium

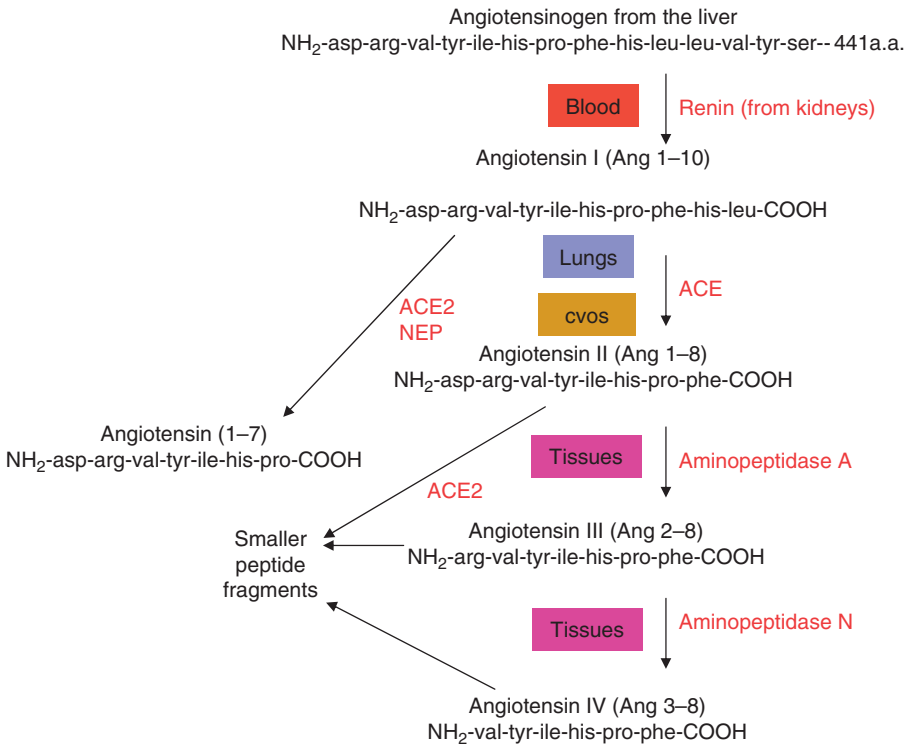

Figure 1 Pathways for the production of circulating angiotensin peptides. Enzymes are indicated in red lettering. ACE, angiotensin converting enzyme; cvos, circumventricular organs; NEP, neutral endopeptidase.

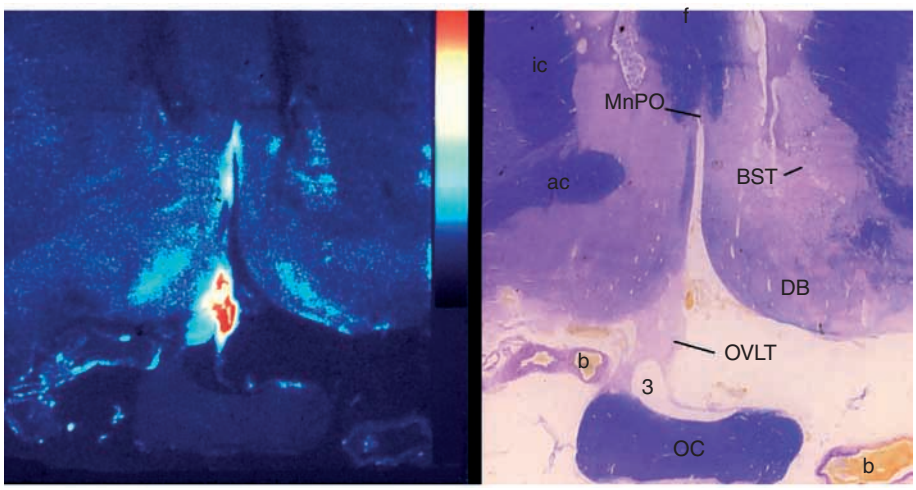

Figure 2 (Left) Angiotensin receptors in the lamina terminalis and other forebrain regions of the human brain. This computer-generated pseudocolor image of an autoradiograph shows high densities of radiolabeled angiotensin II to the vascular organ of the lamina terminalis (OVLT), dorsal part of the median preoptic nucleus (MnPO), bed nucleus of the stria terminalis (BST), and nucleus of the diagonal band. The color bar indicates lowest (blue) to highest (red) receptor-binding densities. (Right) Neuroanatomical features in the corresponding histological transverse section stained with cresyl-violet/luxol fast blue. ac, anterior commissure; b, blood vessel; f, fornix; ic, internal capsule; DB, nucleus of the diagnal band; oc, optic chiasma; 3, third ventricle optic recess.

appetite, vasopressin release, cardiovascular regulation) elicited by the action of blood-borne angiotensin II on these sites are exerted through AT_1 receptors. Extremely high concentrations of angiotensin converting enzyme are found within the circumventricular organs, and angiotensin II can be generated locally within these sites from angiotensin I delivered there from the bloodstream.

Centrally Mediated Cardiovascular Effects of Circulating Angiotensin

The first indication of an action of angiotensin on the brain was the discovery by RK Bickerton and JP Buckley in 1961 of a centrally mediated increase in blood pressure. These investigators showed that infusion of angiotensin II directly into the vertebral artery of dogs elicited a greater increase in arterial pressure than did an infusion of the same amount of angiotensin II into the general circulation. Because the vertebral artery supplied the hindbrain of the dog, they suggested that the pressor response caused by angiotensin infused into this blood vessel was due to an action on the medulla oblongata. Several investigators have shown subsequently that ablation of the area postrema, the only circumventricular organ in the hindbrain, prevents the pressor response to intravertebral arterial infusion of angiotensin II in dogs. Thus, it is likely that blood-borne angiotensin II acts on the area postrema to increase arterial pressure, and there is further evidence to show that it does so by increasing sympathetic vasoconstrictor nerve activity. This increased sympathetic vasoconstrictor activity may be mediated by neural connections between the area postrema, the adjacent nucleus of the solitary tract (NTS), and sympathetic premotor vasoconstrictor neurons in the rostral ventrolateral medulla (RVL). More recently, this idea has been refined to show that circulating angiotensin II acts on the area postrema to inhibit the baroreceptor reflex. Thus chronically administered angiotensin II resets the baroreceptor reflex so that sympathetic vasomotor activity remains elevated even when arterial pressure is high.

Circulating angiotensin may also exert influences on sympathetic vasomotor nerve activity and arterial pressure through AT_1 receptors in the forebrain circumventricular organs, the subfornical organ, and the OVLT. Polysynaptic neural pathways connect neurons in these two sites to both sympathetic premotor and preganglionic neurons in the RVL and the intermediolateral cell column of the spinal cord, respectively. There is evidence to show that the hypothalamic paraventricular nucleus is a relay point in these polysynaptic pathways connecting angiotensin-sensitive neurons in the subfornical organ and the OVLT to sympathetic vasomotor pathways. It should be noted that these forebrain actions of angiotensin II on sympathetic activity and arterial pressure have been implicated in the genesis of some forms of experimentally induced hypertension.

Thirst

Systemic infusion of angiotensin II initiates water drinking in experimental animals, and presumably this behavioral response reflects the stimulation of thirst by angiotensin II action on the brain. When the extracellular fluid is depleted, as for example with hemorrhage, excessive vomiting, or diarrhea, thirst is generated. Such thirst is independent of the osmoreceptors that signal dehydration of the intracellular space. In 1969, JT Fitzsimons showed that the renin-angiotensin system played a crucial role in water drinking resulting from loss of extracellular fluid. He showed that circulating angiotensin II induced water drinking in rats and that it probably acted synergistically with neural signals coming from low pressure volume receptors in the thorax, to initiate thirst. The dipsogenic response to infused angiotensin II was one of the first recognized behavioral effects elicited by a circulating peptide. Virtually all mammals that have been investigated increase water drinking if angiotensin II is administered systemically, although some species (including humans) need quite high blood concentrations of angiotensin II, probably above those encountered under physiological conditions, to stimulate thirst. Baroreceptor stimulation has an inhibitory influence on thirst, and as increased arterial pressure is another consequence of systemic infusion of angiotensin II, the blood levels of angiotensin II that cause drinking behavior may need to be high to overcome the antidipsogenic effect of increased blood pressure. An inhibitory effect of baroreceptor stimulation on thirst does not come into play when endogenous levels of angiotensin II increase under conditions of hypovolemia, however, because arterial pressure does not increase in these circumstances.

The dipsogenic action of angiotensin II may be additive with, potentiated by, or inhibited by factors that also act directly on the subfornical organ. The ovarian hormone relaxin (released during pregnancy) potentiates the dipsogenic action of angiotensin II, while atrial natriuretic peptide (released when extracellular volume is expanded) and estrogen have inhibitory influences on angiotensin-induced drinking. Circulating angiotensin and hypertonicity are additive dipsogenic stimuli. Thus, subjects in a dehydrated condition may experience thirst that results from a combination of both osmotic and angiotensin stimuli. As well as interactions with humoral factors, it is thought that afferent neural signals from the heart and possibly the kidney also interact with angiotensin's dipsogenic action to generate thirst in conditions of hypovolemia.

With the discovery of the dipsogenic action of blood-borne angiotensin II, the question arose as to how a circulating peptide hormone that did not pass across the blood–brain barrier could act directly on the brain to stimulate thirst and water drinking. JB Simpson and A Routtenberg were able to show in 1971 that the subfornical organ, a circumventricular

organ lacking a blood–brain barrier, was the site at which angiotensin II in the circulation stimulated thirst. They demonstrated that the subfornical organ was exquisitely sensitive to locally injected angiotensin II for the induction of water drinking in animals and that ablation of this region abolished water drinking in response to systemic angiotensin II in the laboratory rat. Electrophysiological and *in vitro* autoradiographic binding investigations, as well as more-recent studies of immediate early gene *c-fos* expression, confirm the excitatory action of angiotensin II on neurons of the subfornical organ (see Figure 3).

The neural pathways subserving thirst that are activated subsequent to angiotensin stimulation of neurons within the subfornical organ remain largely unknown. The subfornical organ had efferent neural connections to many other brain sites that included regions such as the lateral hypothalamic area, hypothalamic paraventricular nuclei, OVLT, and median preoptic nucleus (MnPO), all of which have been implicated in thirst responses. The connection to the MnPO has been proposed to relay neural signals coming from angiotensin-sensitive neurons in the subfornical organ to cortical regions for the induction of thirst. Recently, imaging of brains of thirsty humans by positron emission tomography and functional magnetic resonance imaging techniques has drawn attention to the cingulate cortex and insula as cortical sites responsible for the generation of thirst.

Salt Appetite

A hunger for sodium salts develops slowly over a period of several hours when the extracellular fluid becomes depleted. Up to 24 h may elapse following salt and water depletion from the body before salt intake begins to increase. By comparison, thirst is a relatively rapid response to depletion of the extracellular fluid compartment. The most convincing evidence that activation of the systemic renin-angiotensin system is a humoral signal to the brain for salt hunger resulting from sodium depletion is the inhibition of this response by adequate doses of systemically administered angiotensin converting enzyme inhibitors (e.g., captopril) or angiotensin AT_1 receptor antagonists (e.g., losartan). Sodium intake that is blocked by angiotensin converting enzyme inhibitors can be restored by intravenous infusions of small amounts of angiotensin II, evidence that is consistent with a role for circulating angiotensin II in the genesis of sodium appetite. It is thought that the adrenal hormone aldosterone, secreted in response to sodium depletion, acts synergistically with angiotensin to stimulate salt appetite.

Blood-borne angiotensin II may act on AT_1 receptors in both the OVLT and the subfornical organ to influence a hunger for salt because ablation of either of these circumventricular organs reduces salt intake following sodium depletion. Evidence favors the OVLT as the site of action of angiotensin to stimulate salt appetite because microinjection of angiotensin II into the OVLT region (but not the subfornical organ) causes increased salt intake by rats. The subsequent efferent neural pathways distributing from the OVLT that mediate sodium appetite are still largely unknown. Angiotensin-sensitive neurons in the subfornical organ and the OVLT send efferent neural projections to a number of sites, including the bed nucleus of the stria terminalis. This region and the central nucleus of the amygdala have been shown in studies of lesions to have a role in sodium appetite. Thus, it is possible that neurons within the bed nucleus of the stria terminalis and the amygdala may relay signals from the OVLT to higher cortical regions to initiate salt appetite.

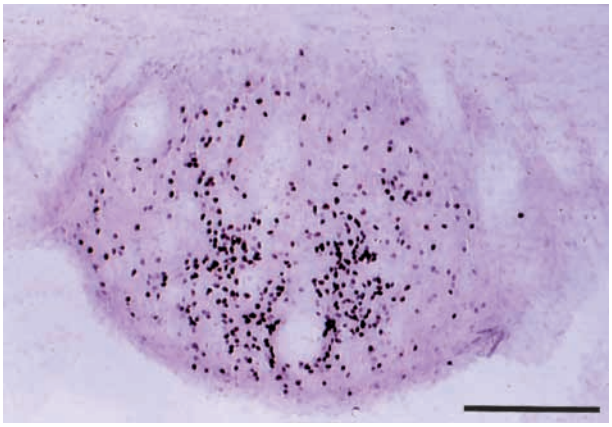

Figure 3 Angiotensin II stimulates neurons in the subfornical organ. Using immunohistochemistry, this transverse section of the rat subfornical organ shows expression of the proto-oncogene *c-fos* in the cell nucleus of neurons within the subfornical organ in response to an intravenous infusion of angiotensin II. Expression of *c-fos* indicates increased electrical activity of neurons. Such expression of *c-fos* is not observed in noninfused rats. Scale bar = 100 μm.

Vasopressin Secretion

Complementary to the dipsogenic action of circulating angiotensin II is its ability to stimulate the release of vasopressin from the posterior pituitary gland. Vasopressin, the antidiuretic hormone, acts on the collecting tubules of the kidney to concentrate urine, thereby reducing urine output and conserving body fluids. The vasopressin-secreting action of circulating angiotensin II is exerted via AT_1 receptors in both the subfornical organ and the OVLT. Some of the angiotensin-sensitive neurons in both these circumventricular organs project directly to the hypothalamic

supraoptic and paraventricular nuclei (sites of vasopressin-synthesizing neurons), suggesting that this is the neural pathway that mediates angiotensin-induced vasopressin release. There is evidence also that neural signals from angiotensin-sensitive neurons are relayed from the subfornical organ and OVLT to the supraoptic and paraventricular nuclei via synaptic connections in the MnPO.

Renal Sympathetic Nerve Activity

Circulating angiotensin II is known to act directly on the sympathetic nerve endings in the kidney to potentiate the release of noradrenaline, but it also has an action on the brain to increase renal sympathetic nerve activity and thereby further promote renal fluid and sodium retention. Angiotensin-sensitive neurons within the subfornical organ and OVLT have polysynaptic neural connections to the renal sympathetic nerves. These neural connections are likely to mediate the effects of circulating angiotensin on renal sympathetic nerve activity, and neural signals from these regions are probably relayed to the kidney via synapses in the hypothalamic paraventricular nucleus and RVL.

Actions of Centrally Generated Angiotensin on the Brain

In addition to physiological responses stimulated by peripherally generated angiotensin II on the circumventricular organs of the brain, angiotensin derived from within the brain may exert a number of different actions on the CNS. These actions are mediated by receptors located at sites with an intact blood–brain barrier and therefore not accessible by circulating angiotensin peptides.

Generation of Angiotensin in the CNS

Several organs and tissues have intrinsic renin-angiotensin systems capable of generating angiotensin peptides, and the brain is no exception. The generation of angiotensin peptides within the CNS appears to be somewhat different from that in the periphery. The precursor protein angiotensinogen is abundant in the extracellular and cerebrospinal fluid of the brain, and astrocytes are the major cellular source of this brain angiotensinogen, as evidenced by the expression of messenger RNA encoding angiotensinogen in astrocytes. Angiotensinogen may also be synthesized within neurons, but the amount is small in comparison with that secreted constitutively by glia. Angiotensinogen-producing glial cells are found throughout the brain, although the hypothalamus and medulla have the highest concentrations. The details of the generation of angiotensin I, angiotensin II, the heptapeptides angiotensin III and angiotensin (1–7), and the hexapeptide angiotensin IV in the brain remain to be elucidated. Renin is present within the brain, but only in exceedingly small amounts. However, angiotensin-converting enzyme is abundant in many brain regions, and its location on external cell membranes facilitates the formation of angiotensin II in the extracellular domain of the brain. Immunohistochemical and electrophysiological evidence suggests that angiotensin II can be released from neurons as a neurotransmitter or neuromodulator. However, whether angiotensin II is generated within neurons or taken up from the brain extracellular space after synthesis in glia is not clear at present. It is also possible that angiotensin peptides diffuse within the brain extracellular fluid to receptor sites distant from where they were formed. Peptidases that generate angiotensin III, angiotensin (1–7), and angiotensin IV are present in the brain.

Angiotensin Receptors within the CNS

A clear indication that centrally generated angiotensin influences brain function is the multitude of different locations within the CNS that have a blood–brain barrier but still express angiotensin receptors, of either the AT_1 or the AT_2 subtype. Neuronal AT_1 receptors have been observed at both pre- and post-synaptic sites. In the hindbrain, AT_1 receptors are highly abundant in the NTS, dorsal motor nucleus of the vagus, rostral and caudal ventrolateral medulla, and medullary raphe nuclei (see Figure 4). In addition, the dorsal horn and intermediolateral cell column of the spinal cord express the AT_1 receptor. In the midbrain, high densities of AT_1 receptors are

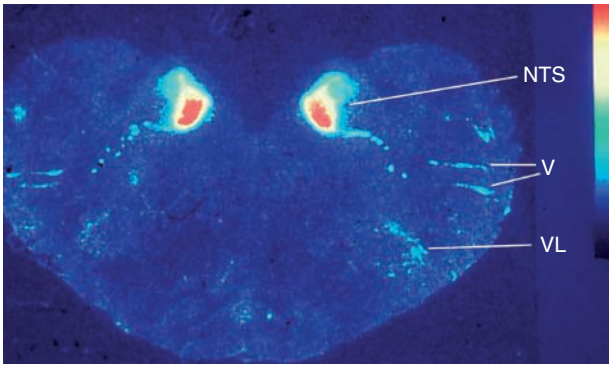

Figure 4 The location of angiotensin AT_1 receptors in the medulla oblongata of a sheep. This pseudocolor image of an autoradiograph of a transverse section shows the binding of radiolabeled angiotensin II to the nucleus of the solitary tract (NTS), ventrolateral medulla (VL) and fibers of the vagus nerve (v) coursing through medulla to the NTS. Highest density of binding is indicated by red and the diminishing gradation by the color bar. Magnification ×10.

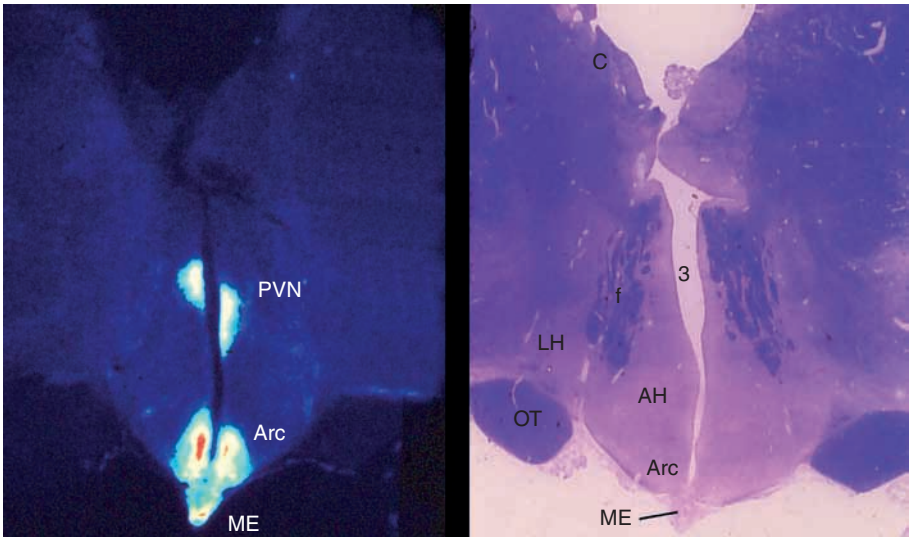

Figure 5 (Left) Location of angiotensin AT_1 receptors in the human hypothalamus. This computer-generated pseudocolor image of an autoradiograph shows high densities of binding of radiolabeled-angiotensin II to the hypothalamic paraventricular nucleus (PVN), arcuate nucleus (Arc), and median eminence (ME). (Right) Neuroanatomical features in the corresponding histological transverse section stained with cresyl violet/luxol fast blue. Magnification ×2.5. AH, anterior hypothalamic area; C, caudate nucleus; f, fornix; LH, lateral hypothalamic area; OT, optic tract.

found in the lateral parabrachial nucleus and substantia nigra pars compacta. In the diencephalon, high densities of angiotensin receptors occur in the hypothalamic paraventricular and supraoptic nuclei, the arcuate nucleus and median eminence, and the MnPO. In the forebrain, the amygdala, septum, caudate nucleus, and olfactory bulb are sites of AT_1 receptor expression (see Figure 5).

Effects of Centrally Administered Angiotensin II

Injection of angiotensin II or III into the lateral or third cerebral ventricles has been shown to elicit many physiological responses, often but not always similar to those obtained with peripherally administered angiotensin II. These responses include thirst and salt appetite, vasopressin secretion, and increased arterial pressure.

Thirst and Vasopressin Secretion

The initial evidence that brain angiotensin could play a role in thirst mechanisms came from AN Epstein and colleagues, who observed that administration of small amounts of angiotensin II directly into the brain could initiate water drinking within seconds. Central injection of angiotensin II also stimulates the secretion of vasopressin from the posterior pituitary gland, resulting in antidiuresis and reduced fluid loss in urine. These responses are similar to those obtained when angiotensin is infused systemically, and they are also blocked by AT_1 receptor antagonists. However, it is clear that the angiotensin-responsive neurons that mediate the responses to centrally injected angiotensin II are not the same as those in the subfornical organ that mediate the dipsogenic and vasopressin responses to circulating angiotensin, because responses to centrally administered angiotensin persist when the subfornical organ is ablated.

Although some neurons in the subfornical organ and OVLT are stimulated when angiotensin II is injected into the lateral cerebral ventricle, the MnPO is the major site of action of intracerebroventricularly infused angiotensin II to stimulate thirst and vasopressin secretion. The MnPO is situated in the anterior wall of the third ventricle (the lamina terminalis) between the subfornical organ and OVLT. Neurons situated in the MnPO have high concentrations of AT_1 receptors, and they receive afferent neural inputs from several other brain regions that are known to mediate fluid and electrolyte homeostasis, such as the subfornical organ, OVLT, lateral parabrachial nucleus, midbrain raphe, and ventrolateral medulla. Since angiotensin-responsive neurons in the MnPO are not directly affected by blood-borne angiotensin because of the blood–brain barrier, it is likely that the endogenous angiotensin that acts on neurons in the MnPO is derived from the brain and acts at the MnPO as a neurotransmitter to engage neural pathways subserving thirst and vasopressin secretion.

In regard to vasopressin release, direct projections from the MnPO to vasopressin-secreting cells of the supraoptic nucleus have been demonstrated,

and ablation of the MnPO disrupts angiotensin-stimulated vasopressin secretion. These observations are consistent with an important role of angiotensin-sensitive neurons in the MnPO to relay signals to the supraoptic and paraventricular nucleus for vasopressin secretion. There are also data that indicate that angiotensin III may be the effector peptide responsible for central angiotensinergic influences on vasopressin release.

Sodium Appetite

Central administration of angiotensin II or renin is a potent stimulus to the ingestion of salt (sodium chloride) by mammals such as rodents, ruminants, and primates. The onset of this angiotensin-stimulated salt appetite is many hours later than the dipsogenic response to centrally administered angiotensin. Experiments in rodents and primates showing inhibition of salt intake by centrally injected angiotensin AT_1 receptor antagonist drugs are further evidence that brain angiotensin has a role in the generation of salt appetite. A powerful synergistic action between central angiotensin and mineralocorticoids in the stimulation of salt intake has been demonstrated. The locus of action of brain angiotensin to induce salt appetite is unclear at present.

Arterial Pressure

Angiotensin increases arterial pressure when microinjected into a number of different brain sites. These include the NTS, the hypothalamic paraventricular nucleus, and the RVL. These two latter regions have direct efferent connections to sympathetic preganglionic neurons in the intermediolateral cell column of the thoracic spinal cord, and stimulation of AT_1 receptors in the paraventricular nucleus or RVL increases arterial pressure by increasing sympathetic vasomotor nerve activity in the periphery.

Pharmacological blockade of brain angiotensin AT_1 receptors in sodium-depleted or hemorrhaged animals results in a failure to maintain arterial pressure. The same treatment drops blood pressure in a number of animal models of hypertension. These observations indicate that brain-derived angiotensin has a role in the regulation of arterial pressure in some physiological and pathophysiological conditions. An angiotensinergic synapse within the RVL may relay neural signals from the hypothalamic paraventricular nucleus to sympathetic preganglionic neurons in the spinal cord to influence arterial pressure.

Another angiotensin peptide, angiotensin (1–7), influences cardiovascular regulation when infused into the brain ventricles or RVL, probably via a non-AT_1 or non-AT_2 receptor. In the NTS, its action is to enhance the baroreceptor reflex, opposite the effect of angiotensin II, whereas in the RVL its action to increase sympathetic activity is similar to that of angiotensin II.

Natriuresis

Injection of angiotensin II into the lateral or third cerebral ventricles causes a rapid and large increase in sodium excretion by the kidneys. This natriuretic response is probably the result of a combination of the increased arterial pressure, reduced renal sympathetic nerve activity, and increased release of vasopressin that also result from the intracerebroventricular administration of angiotensin II. Intracerebroventricular injection of hypertonic saline also causes natriuresis. This effect is blocked by prior central administration of the angiotensin AT_1 receptor antagonist losartan, indicating a possible role of central angiotensinergic pathways in the osmoregulatory excretion of sodium by the kidneys. In combination with increased water intake and fluid retention caused by increased vasopressin levels, this natriuretic effect reduces the sodium concentration and osmolality of the extracellular fluid. Thus, it could be construed that central angiotensinergic action in the short term is attuned to reducing the tonicity of extracellular fluid and, as a consequence of osmotic forces, increasing intracellular volume.

ACTH Secretion

Following the intracerebroventricular injection of angiotensin II, corticotropin-releasing hormone (CRH), synthesized in the parvocellular part of the hypothalamic paraventricular nucleus, is secreted into the pituitary portal blood supply of the median eminence. Subsequently, this CRH stimulates ACTH secretion from the anterior pituitary gland into the systemic circulation. As a result of the action of circulating ACTH on the adrenal cortex, increased secretion of the glucocorticoid hormones cortisol and corticosterone from the adrenal gland occurs. CRH-containing neurons in the paraventricular nucleus that project to the median eminence express high levels of AT_1 receptors and are a likely site at which brain angiotensin influences the hypothalamic–pituitary–adrenal axis.

Body Temperature

Injection of angiotensin II into the cerebral ventricles reduces core body temperature in rats. Pharmacological blockade of brain angiotensin AT_1 receptors results in a higher core temperature of exercising rats or rats exposed to high ambient temperature. These data indicate that angiotensin endogenous to

the brain probably has a role in thermoregulatory function. The thermoregulatory brain regions and mechanisms influenced by brain angiotensin have yet to be identified, although it is known that brain angiotensin may affect the redistribution of blood to the skin and therefore the ability to radiate heat from the body.

Memory

The angiotensin hexapeptide metabolite angiotensin IV has been implicated in memory because central infusions of angiotensin IV agonist drugs reverse experimentally induced memory deficits. The angiotensin IV binding site in the brain is the insulin-regulated amino peptidase (IRAP), and it has been proposed that angiotensin IV enhances memory and learning by inhibiting IRAP enzymatic activity. However, whether sufficient amounts of endogenous angiotensin IV can be generated in the brain to influence memory is undetermined.

Cerebral Blood Flow

In addition to its neurally mediated responses, angiotensin II causes vasoconstriction of cerebral blood vessels and may alter cerebrovascular autoregulation. Administration of AT_1 receptor antagonists improves recovery of brain function following focal cerebral ischemia in animal stroke models by improving vasodilation and restoring cerebrovascular autoregulation.

See also: Angiotensin II; Autonomic Nervous System: Central Cardiovascular Control; Blood Pressure: Baroreceptors; Circumventricular Organs; Circumventricular Organs in Neuroendocrine Control; Neuronal Angiotensin; Osmoregulation; Salt Appetite; Vasopressin/Oxytocin and Receptors.

Further Reading

Allen AM, Oldfield BJ, Giles ME, Paxinos G, McKinley MJ, and Mendelsohn FAO (2000) Localization of angiotensin receptors in the nervous system. In: Quirion R, Bjöorklund A, and Hökfelt T (eds.) *Handbook of Chemical Neuroanatomy, Vol. 16: Peptide Receptors,* Part I, pp. 79–124. Amsterdam: Elsevier.

Andersson B, Eriksson L, Fernandez O, Kolmodin C-G, and Oltner R (1972) Centrally mediated effects of angiotensin II on arterial pressure and fluid balance. *Acta Psychiatrica Scandinavica* 85: 398–407.

Davisson RL (2003) Physiological genomic analysis of the brain renin-angiotensin system. *American Journal of Physiology* 285: R498–R511.

DiBona GF (2001) Peripheral and central interactions between the renin-angiotensin system and the renal sympathetic nerves in control of renal function. *Annals of the New York Academy of Sciences* 940: 395–406.

Fitzsimons JT (1998) Angiotensin, thirst and sodium appetite. *Physiological Reviews* 78: 583–686.

Johnson AK and Thunhorst R (1997) The neuroendocrinology of thirst and sodium appetite: Visceral sensory signals and mechanisms of central integration. *Frontiers in Neuroendocrinology* 18: 292–353.

Llorens-Cortes C and Mendelsohn FA (2002) Organisation and functional role of the brain angiotensin system. *Journal of the Renin-Angiotensin-Aldosterone System* 3(supplement 1): S39–S48.

McKinley MJ, Albiston AL, Allen AM, et al. (2003) The brain renin-angiotensin system: Location and physiological roles. *International Journal of Biochemistry & Cell Biology* 35: 901–918.

McKinley MJ, McAllen RM, Davern P, et al. (2003) The sensory circumventricular organs. *Advances i n Anatomy, Embryology, and Cell Biology* 172: 1–127.

Saavedra JM, Ando H, Armando I, Bregonzio G, Jezova M, and Zhou J (2004) Brain angiotensin II, an important stress hormone: Regulatory sites and therapeutic opportunities. *Annals of the New York Academy of Sciences* 1018: 76–84.

Unger T, and Scholkens BA (eds.) (2004) *Handbook of Experimental Pharmacology 163. Angiotensin,* 2 vols. Berlin: Springer.

Von Bohlen und, Halbach O, and Albrecht D (2006) The CNS renin-angiotensin system. *Cell and Tissue Research* 326: 599–616.

Washburn DL and Ferguson AL (2001) Membrane properties of subfornical organ neurons. *Clinical and Experimental Pharmacology & Physiology* 28: 575–580.

Angiotensin II

M J McKinley, University of Melbourne, Melbourne, VIC, Australia
B J Oldfield, Monash University, Clayton, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The renin–angiotensin system, through the actions of its major effector peptides angiotensin II and angiotensin III (angiotensin peptides), has major influences on the regulation of bodily water and salt balance. Consideration of the role in the brain of angiotensin's action on fluid and electrolyte homeostasis is incomplete unless viewed in the context of its peripheral actions and overall involvement in fluid and electrolyte homeostasis. In this regard, the influences of the renin–angiotensin system range from the hormonal actions of circulating angiotensin II on the kidney, vasculature, adrenal gland, and brain to likely roles of angiotensin peptides as neurotransmitters within the brain. In addition to regulating fluid and electrolyte balance, there are powerful influences of angiotensin on the cardiovascular system and diverse biological actions of locally generated angiotensin peptides in various organs and tissues throughout the body.

Circulating angiotensin II, generated under conditions of sodium depletion, hypovolemia, or hypotension, acts on specific receptors in the adrenal gland, kidney, and brain to regulate fluid and electrolyte homeostasis. Within the adrenal gland, angiotensin peptides act on cells in the zona glomerulosa to stimulate secretion of aldosterone. This steroid hormone acts on the distal tubule of the kidney to cause sodium retention. As well, circulating angiotensin II has direct effects on the kidney, to influence renal sodium and water handling and to provide an inhibitory feedback regulation of renin secretion. Circulating angiotensin II acts on the brain to initiate water drinking (presumably as a response to stimulation of thirst), salt hunger, vasopressin secretion, and increased sympathetic nerve activity. These responses complement the peripheral actions of angiotensin II (vasoconstriction, cardiac stimulation, sodium retention resulting from aldosterone action, and renal Na reabsorption) in restoring the capacity of the cardiovascular system to function optimally. The responses are consistent with the notion that the renin–angiotensin system is the main humoral signal coordinating fluid and electrolyte homeostasis.

In examining the roles of angiotensin peptides in the regulation of fluid and electrolyte homeostasis by the brain, consideration of the actions of angiotensin II derived from within the brain, as well as that circulating in the systemic bloodstream, is also necessary. It is relevant to appreciate the physiological conditions and mechanisms by which angiotensin peptides are generated both in the bloodstream and locally within the brain. Circulating angiotensin peptides do not readily pass across the blood–brain barrier, and therefore the actions of blood-borne angiotensin II on the brain are limited to brain sites lacking this barrier – these are the circumventricular organs (CVOs), such as the subfornical organ and organum vasculosum of the lamina terminalis (OVLT) in the forebrain, and area postrema in the medulla oblongata. In the rest of the brain (separated from blood-borne angiotensin II by the blood–brain barrier), angiotensin peptides may be generated independently of renin secretion from the kidneys. These centrally generated angiotensin peptides may interact with angiotensin receptors that are situated in many parts of the central nervous system (CNS).

Generation of Systemic Angiotensin II

Generation of angiotensin peptides in the bloodstream is dependent on the action of renin, an enzyme that is synthesized in the kidney as preprorenin and processed to its circulating active form, renin. The existence of renin within the kidney had been known since its discovery in 1898. However, it was not until many years later that angiotensin was identified as the effector peptide responsible for the pressor action of circulating renin. Subsequently, angiotensin I was purified and its amino acid sequence determined.

Circulating angiotensin peptides are formed when the aspartyl protease renin is released by the juxtaglomerular apparatus of the kidney into the circulation to act on a circulating protein, angiotensinogen, to enzymatically cleave the decapeptide molecule angiotensin I (Figure 1). (Angiotensinogen has a 452-amino-acid sequence, is synthesized in the liver, and is also known as renin substrate.) Angiotensin I has no biological actions. However, the octapeptide angiotensin II is generated for distribution in the bloodstream when another protease, angiotensin converting enzyme (ACE; also known as dipeptidyl peptidase or kininase II), located on the luminal surface of endothelial cell membranes in many tissues, especially in the lungs, cleaves two more amino acids from the decapeptide. Angiotensin II is the major, but not the exclusive effector peptide of the renin–angiotensin system. Another enzyme, aminopeptidase A, removes a further amino acid from angiotensin II to cause the

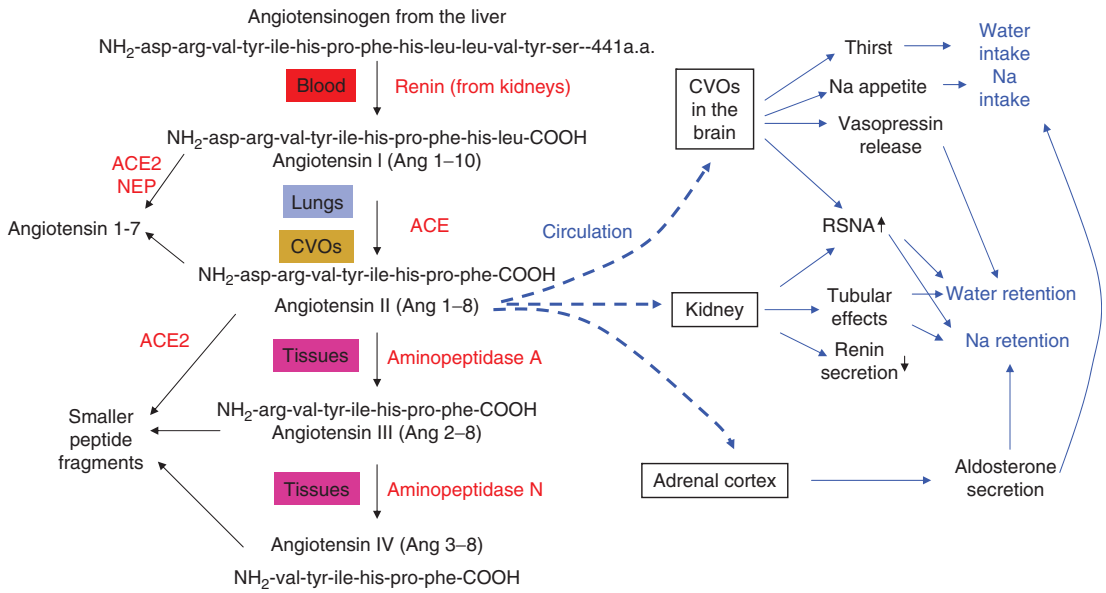

Figure 1 Diagram of the enzymatic reactions that lead to the generation of angiotensin peptides in the peripheral circulation and the circumventricular organs (CVOs) of the brain. Enzymes are indicated in red lettering. The main actions of circulating angiotensin II on salt and water balance are shown on the right. ACE, angiotensin converting enzyme; Ang, angiotensin; NEP, neutral endopeptidase; RSNA, renal sympathetic nerve activity.

formation of the heptapeptide angiotensin III. This molecule is also biologically active and can interact with angiotensin AT_1 receptors in the kidney, adrenal gland, and brain to stimulate sodium reabsorption, aldosterone secretion, and thirst. Further degradation of angiotensin II and III occurs in tissues, with the subsequent formation of the hexapeptide angiotensin IV and smaller peptide fragments that do not act on AT_1 receptors and do not appear to have any actions on salt and water balance.

Stimuli for Renin Secretion and Angiotensin Generation

As previously mentioned, the renin–angiotensin system brings mechanisms into play that both restore depleted fluid and electrolytes (thirst and sodium appetite) and prevent further losses of body water and sodium levels. As sodium chloride is the major ionic component of extracellular fluid, it can be seen that the renin–angiotensin system is a major factor in preventing extracellular fluid depletion. Renin secretion and the subsequent generation of angiotensin II occur when the cardiovascular system is compromised. Reduction of extracellular fluid volume (hypovolemia) is a potent stimulus to renin secretion from the kidney. Hypovolemia can result from sodium depletion caused by loss of fluid and electrolytes from the gastrointestinal tract, excessive sweating, or prolonged exposure to a low-salt diet. Renin secretion also results when there is a decrease in arterial blood pressure. The kidney responds with renin release to three main signals that signify depletion of the extracellular fluid or cardiovascular dysfunction. These signals are a reduction in the renal perfusion pressure, a reduction in the sodium load within the macula densa of the renal distal tubule, and neural signals from the brain in the form of increased activity in the sympathetic nerve supply to the kidney.

Angiotensin Receptors

Subtypes and Structure

The chemical structures of the specific receptors on which angiotensin II and III act in the brain and the periphery to regulate fluid and electrolyte homeostasis were discovered in 1991. Two main classes of angiotensin receptors have been identified and designated as the AT_1 and AT_2 receptors; they have 30% peptide sequence homology. In rodents (but not humans), the AT_1 receptor has two subtypes, AT_{1A} and AT_{1B}, that differ by only a few amino acids. The AT_1 and AT_2 receptors are both membrane-bound G-protein-coupled receptors, with seven membrane-spanning domains, three intracellular and extracellular loops, an intracellular C-terminus, and extracellular N-terminus. Although an AT_4 receptor was proposed as a site of action of the hexapeptide angiotensin IV, in the brain the site of the binding of this peptide has been shown to be insulin-regulated aminopeptidase (IRAP).

Location of AT_1 Receptors in the CNS

The location of angiotensin binding sites within the brain (of the rat) was mapped in 1984, and these sites were later identified as regions containing either AT_1 or AT_2 receptors. The distribution of these receptors in the human brain is similar, although not identical, to those in rat brain. One of the more remarkable features of the distribution of the AT_1 receptor in the brain is the presence of such receptors in virtually all brain regions known to participate in the neural regulation of body fluid and electrolyte homeostasis and cardiovascular control. AT_1 receptors can be located on either pre- or postsynaptic neuronal sites in these regions. In the human brain, high concentrations of AT_1 receptors are found in many regions. These regions include the subfornical organ and OVLT, sensory CVOs that lack a blood–brain barrier and are therefore susceptible to blood-borne angiotensin II. A number of other regions – such as the median preoptic nucleus (MnPO), septum, and bed nucleus of the stria terminalis in the forebrain; supraoptic, paraventricular, and arcuate nuclei in the hypothalamus; lateral parabrachial nucleus, substantia nigra, and locus coeruleus in the midbrain; and nucleus of the solitary tract and ventrolateral medulla in the hindbrain – are rich in AT_1 receptors (Figure 2). Angiotensin AT_1 receptors are also present in the intermediolateral cell column and dorsal horn of the spinal cord and in dorsal root ganglia.

Intracellular Signaling

Angiotensin II acts on neuronal AT_1 receptors to stimulate firing of action potentials. While the intracellular signaling events to which AT_1 receptors are coupled have been studied in only a few brain sites, AT_1 receptor signaling occurs via G-protein (G_q) coupled to the third intracellular loop. Engagement of G_q activates phospholipase Cβ, with subsequent formation of diacylglycerol and inositol triphosphate leading to increased intracellular calcium ions, and activation of protein kinase C and calcium/calmodulin kinase II (CaMKII). These kinases open Ca and close K channels, resulting in neuronal excitation. There is recent evidence that superoxide formation has a role in transducing angiotensin influences on neurons.

The density of AT_1 receptors in the CNS is not static, and they can be up- or downregulated in some regions in response to alterations in sodium and water balance, as well as to other metabolic changes. Specific pharmacological antagonists to both AT_1 (e.g., losartan, candesartan, irbesartan) and AT_2 (e.g., PD123319, CGP42112A) receptors have been developed, and AT_1 receptor antagonists are

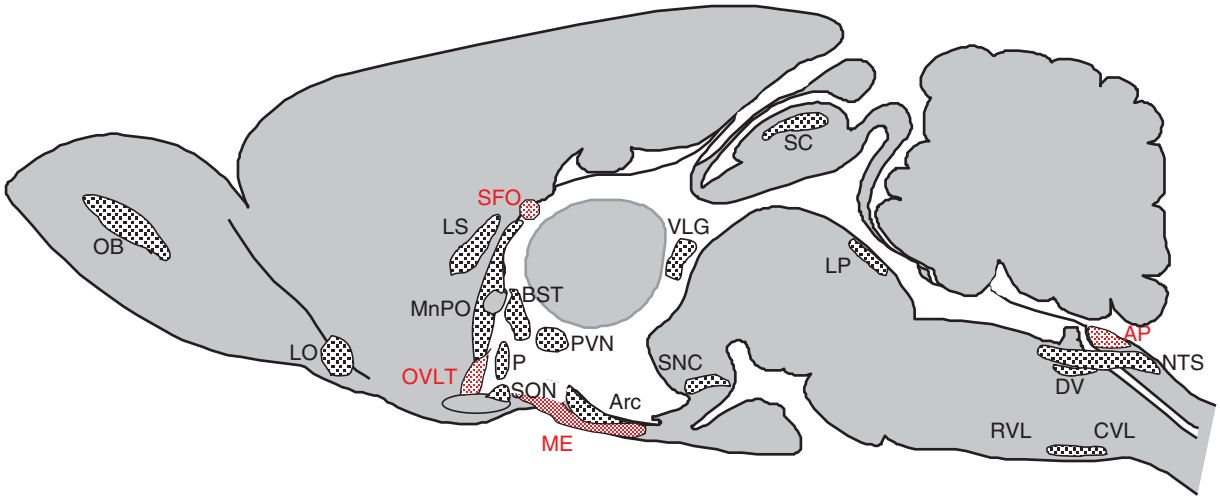

Figure 2 Diagram of the locations of angiotensin AT_1 receptors in the mammalian brain. Circumventricular organs that lack a blood–brain barrier, and are therefore exposed to circulating angiotensin II, are shown in red. Results from several species, including humans, have been projected onto a midsagittal diagram of the rat brain. AP, area postrema; Arc, arcuate nucleus; BST, bed nucleus of the stria terminalis; CVL, caudal ventrolateral medulla; DV, dorsal motor nucleus of vagus; LO, lateral olfactory tract; LP, lateral parabrachial nucleus; LS, lateral septum; ME, median eminence; MnPO, median preoptic nucleus; NTS, nucleus of the solitary tract; OB, olfactory bulb; OVLT, organum vasculosum of the lamina terminalis; P, periventricular nucleus; PVN, hypothalamic paraventricular nucleus; RVL, rostral ventrolateral medulla; SC, superior colliculus; SFO, subfornical organ; SON, supraoptic nucleus; SNC, substantia nigra pars compactus; VLG, ventral lateral geniculate nucleus.

now in wide clinical use. Earlier peptide analog antagonists (e.g., saralasin, sarile, sarthran) blocked both types of angiotensin receptors.

Central Actions of Circulating Angiotensin

In 1961, an increase in arterial pressure in dogs was observed in response to infusion of angiotensin II into the blood supply of their brain. This pressor response was the first recognized action on the brain of angiotensin. Subsequently, other brain-mediated responses of angiotensin have been discovered, including thirst, vasopressin release, sodium appetite, and sympathetic activation.

Thirst

Depletion of the extracellular compartment of the body stimulates thirst. Such thirst is independent of osmoreceptors, the sensors that signal dehydration of the intracellular compartment to initiate thirst and vasopressin secretion. In 1969, JT Fitzsimons showed that circulating angiotensin II was a potent dipsogen, and that it probably acted synergistically with neural signals coming from low-pressure volume receptors in the thorax to initiate thirst.The dipsogenic action of angiotensin II was one of the earliest recognized behavioral responses elicited by a circulating peptide. Virtually all mammals tested increased water drinking when angiotensin II was administered systemically, although some species, including humans, need supraphysiological levels of angiotensin II in blood to stimulate thirst and drinking behavior. Baroreceptor stimulation has an inhibitory influence on thirst, and, as increased arterial pressure is another consequence of systemic infusion of angiotensin II, the blood levels of angiotensin II needed to cause drinking behavior may be high to overcome the antidipsogenic effect of increased blood pressure in normovolemic subjects. An inhibitory effect of baroreceptor stimulation on thirst does not come into play when endogenous levels of angiotensin II increase under conditions of hypovolemia, because arterial pressure does not rise in this condition.

The dipsogenic action of angiotensin II may be additive with, potentiated, or inhibited by factors that also act directly on the subfornical organ. The ovarian hormone relaxin (released during pregnancy) potentiates the dipsogenic action of angiotensin II, while atrial natriuretic peptide (released when extracellular volume is expanded) and estrogen have inhibitory influences on angiotensin-induced drinking. Angiotensin and hypertonicity are additive dipsogenic stimuli. Thus, individuals in a dehydrated condition, occurring as a consequence of water deprivation, may experience thirst that results from a combination of both osmotic and angiotensin stimuli. As well as interactions with humoral factors, it is thought that afferent neural signals from the heart and possibly the kidney also interact with angiotensin's dipsogenic action to generate thirst in conditions of hypovolemia.

With the discovery of the dipsogenic action of blood-borne angiotensin II, the question arose as to how a peptide hormone that did not pass across the blood–brain barrier could stimulate drinking behavior. The answer is that angiotensin II from the circulation stimulates thirst by acting at the subfornical organ, a CVO without a blood–brain barrier. The subfornical organ is exquisitely sensitive to locally injected angiotensin II for the induction of water drinking in animals, and ablation of this region abolished water drinking in response to systemic angiotensin II. Subsequent electrophysiological and *in vitro* autoradiographic binding investigations, as well as studies of immediate-early gene *c-fos* expression, confirm the excitatory action of angiotensin II on subfornical neurons.

The neural pathways that are activated to initiate thirst subsequent to angiotensin stimulation of the subfornical organ remain largely unknown to this point. The subfornical organ has efferent neural connections to many other brain sites including the lateral hypothalamic area, hypothalamic paraventricular nuclei, OVLT, and median preoptic nucleus, all of which have been implicated in thirst responses. The connection to the median preoptic nucleus has been proposed to relay dipsogenic signals coming from angiotensin stimulation of the subfornical organ to cortical regions for the induction of thirst.

Salt Appetite

In addition to thirst, depletion of the extracellular fluid also results in the gradual development of a hunger for sodium salts. Experimentally, depletion of the extracellular fluid is achieved in animals by diuretic treatment, peritoneal dialysis, subcutaneous injection of colloid (e.g., polyethylene glycol), low-salt diet, or uncompensated loss of saliva from a fistula. Up to 24 h may elapse before the onset of increased salt intake following such treatments. In comparison, thirst and increased water intake are relatively rapid responses (minutes to hours) to extracellular fluid depletion. The most convincing evidence that activation of the systemic renin–angiotensin system is the humoral signal to the brain for salt hunger subsequent to sodium depletion is the inhibition of this response by systemically administered ACE inhibitors (see later, for discussion of doses) such as captopril or angiotensin AT_1 receptor antagonists. Salt intake

that is blocked by ACE inhibitors can be restored by intravenous infusions of small amounts of angiotensin II, evidence that is consistent with a role for circulating angiotensin II in the genesis of sodium appetite. It is thought that aldosterone acts synergistically with angiotensin to stimulate salt appetite.

Circulating angiotensin II probably acts on AT_1 receptors in both the OVLT and the subfornical organ to induce a hunger for salt, because ablation of either of these CVOs can reduce salt intake following sodium depletion. Interestingly, microinjection of angiotensin II into the OVLT region, but not the subfornical organ, stimulates salt intake. Since there are reciprocal neural connections between these two CVOs, it is possible that these connections relay signals relevant to salt appetite, and that is why ablation of either of these regions reduces salt hunger. The subsequent efferent neural pathways from these two CVOs that mediate sodium appetite are still largely unknown.

Vasopressin Secretion

Shortly after the discovery of the dipsogenic action of angiotensin, it was shown that systemic infusion of angiotensin II increased the secretion of the antidiuretic hormone vasopressin from the pituitary gland. Again, angiotensin II from the circulation acts on AT_1 receptors mainly in the subfornical organ, but also in the OVLT, to initiate vasopressin release. Some of the angiotensin-sensitive neurons in both of these CVOs project directly to the hypothalamic supraoptic and paraventricular nuclei (sites of vasopressin-synthesizing neurons), implying that this neural pathway mediates angiotensin-induced vasopressin release. Angiotensin-stimulated signals from these CVOs may also be relayed to the supraoptic and paraventricular nuclei via the MnPO.

Renal Sympathetic Nerve Activity

Sodium depletion and hypovolemia result in increased activity of the sympathetic nerves supplying the kidney. Increased renal sympathetic nerve activity (RSNA) leads to reduced excretion of sodium and water in urine. Circulating angiotensin II is known to act directly on the sympathetic nerve endings in the kidney to potentiate the release of noradrenaline. It also has an action on the brain to increase RSNA, and thereby further promote renal fluid and sodium retention. There are angiotensin-sensitive neurons within the subfornical organ and OVLT that have polysynaptic neural connections to the renal sympathetic nerves. These connections are likely to mediate the effects of circulating angiotensin on RSNA, and there is evidence that neural signals from these CVOs are relayed to the kidney via synapses in the hypothalamic paraventricular nucleus and rostral ventrolateral medulla. There is evidence that angiotensin action on the area postrema may also influence renal sympathetic outflow.

Brain-Derived Angiotensin

Angiotensin Generation in the Brain

The generation of angiotensin peptides within the brain contrasts with their formation in the periphery. While details are less clear regarding the central formation of angiotensin II, it is clear that the precursor protein angiotensinogen is abundant within the extracellular and cerebrospinal fluids of the brain. Astrocytes are the major cellular source of this brain angiotensinogen, and the expression of mRNA encoding angiotensinogen in astrocytes confirms the central origin of astrocytic angiotensinogen. There is also evidence that some angiotensinogen may also be synthesized within neurons, but the amount is small in comparison to that secreted constitutively by glia. Angiotensinogen-producing glial cells are found within all divisions of the CNS, but there appears to be a preponderance of these cells in the hypothalamus and medulla, divisions that are of major significance for fluid and electrolyte homeostasis and cardiovascular regulation. A major unresolved question is the manner in which the centrally synthesized angiotensinogen is processed to angiotensins I, II, and III within the brain. The amount of renin that can be detected in the brain is quite small, casting doubt on whether renin is the main processing enzyme for angiotensin I formation in the CNS. On the other hand, ACE is plentiful in several parts of the brain, and its location on external cell membranes facilitates the formation of angiotensin II in the extracellular domain of the brain. Immunohistochemical and electrophysiological evidence suggests that angiotensin II can be released from neurons as a neurotransmitter/neuromodulator. The question of how angiotensin II arises within such angiotensinergic neurons is unresolved, as are the enzymes and cellular domains responsible.

Generation of angiotensin peptides in the subfornical organ and OVLT deserves separate analysis from that described in the previous paragraph, because these sites have extremely high concentrations of ACE and they lack a blood–brain barrier. As a consequence, in rat brain, blood-borne angiotensin I gains access to the ACE in these CVOs, with

subsequent local generation of angiotensin II. It is possible that this locally generated angiotensin II in the subfornical organ and OVLT is in close proximity to the neurons that engage neural pathways that mediate thirst, vasopressin secretion, and salt appetite. Such local generation of angiotensin II in the subfornical organ is the probable explanation of a paradoxical stimulation of water and salt intake that is observed in rats when they are treated with amounts of ACE inhibitors such as captopril or enalopril sufficient to block peripheral formation of angiotensin I, but not the high concentrations of ACE in the CVOs. As a result of peripheral blockade of ACE, circulating angiotensin I levels increase (due to disruption of feedback inhibition of angiotensin II on the kidney), and when this angiotensin I reaches the subfornical organ and OVLT, it is converted to angiotensin II locally, with subsequent stimulation of thirst, vasopressin secretion, and sodium hunger.

Actions of Brain-Derived Angiotensin Peptides on Fluid and Electrolyte Balance

Injections of angiotensin II into the cerebral ventricles or discrete brain sites stimulate copious drinking of water within seconds. Centrally injected angiotensin also initiates a salt appetite, but with a latency of many hours, and can have a powerful synergistic action with mineralocorticoid steroids in the stimulation of salt intake. Central injection of angiotensin II also is a strong stimulus for the secretion of vasopressin from the posterior pituitary gland, resulting in antidiuresis and reduced fluid loss in urine. These actions are similar to those obtained when angiotensin is infused systemically, and they are also blocked by AT_1 receptor antagonists. However, it is clear that the angiotensin-responsive neurons that are mediating the responses to centrally injected angiotensin II are not the same as those in the subfornical organ that mediate the responses to circulating angiotensin, because responses to centrally administered angiotensin persist when the subfornical organ is ablated. As well, not all responses to centrally administered angiotensin II are consistent with maintaining a positive sodium balance. In fact, the initial response of the kidney to centrally administered angiotensin is to increase the excretion of sodium in urine (natriuresis), thus depleting the extracellular fluid of sodium and secondarily stimulating salt appetite. In combination with increased water intake and fluid retention caused by increased vasopressin levels, this natriuretic effect reduces the sodium concentration and osmolality of the extracellular fluid. Thus, it could be construed that central angiotensinergic action in the short term is attuned to reducing the tonicity of extracellular fluid, and as a consequence of osmotic forces, increasing intracellular volume.

Although some neurons in the subfornical organ and OVLT are stimulated when angiotensin II is injected into the lateral cerebral ventricle, the MnPO is the major site of action of intracerebroventricularly infused angiotensin II to stimulate thirst, vasopressin secretion, and natriuresis. The MnPO is situated in the anterior wall of the third ventricle (the lamina terminalis) between the subfornical organ and OVLT. Neurons situated in the MnPO have high concentrations of AT_1 receptors, and they receive afferent neural inputs from several other brain regions that are known to mediate fluid and electrolyte homeostasis, such as the subfornical organ, OVLT, lateral parabrachial nucleus, midbrain raphe, and ventrolateral medulla. Since angiotensin-responsive neurons in the MnPO are not directly affected by blood-borne angiotensin because of the blood–brain barrier, it is likely that the endogenous angiotensin that acts on neurons in the MnPO is derived from the brain, and acts at the MnPO as a neurotransmitter to engage neural pathways subserving thirst, vasopressin secretion, and natriuresis.

Direct projections from the MnPO to vasopressin-secreting cells of the supraoptic nucleus have been demonstrated, and ablation of the MnPO disrupts angiotensin-stimulated vasopressin secretion. These observations are consistent with an important role of angiotensin-sensitive neurons in the MnPO to relay signals to the supraoptic and paraventricular nuclei for vasopressin secretion. There are also data that indicate that angiotensin III may be the effector peptide responsible for central angiotensinergic influences on vasopressin release.

While this discussion has emphasized the importance of AT_1 receptors in fluid and sodium homeostasis, it would be imprudent to totally discount AT_2 receptors in this regard. There is evidence that central administration of AT_2 receptor antagonists can cause a small increase in dipsogenic and pressor responses to centrally administered angiotensin II, suggesting an antidipsogenic and depressor effect of AT_2 receptor stimulation, opposite to the effects of AT_1 receptor stimulation.

See also: Angiotensin Actions on and Within Brain; Atrial Natriuretic Peptide: Fluid/Mineral Balance; Blood Pressure: Baroreceptors; Circumventricular Organs; Fluid and Electrolyte Homeostasis: Clinical Disease; Neurohypophyseal System; Neuronal Angiotensin; Osmoregulation; Salt Appetite; Thirst; Vasopressin/Oxytocin and Receptors.

Further Reading

Allen AM, Oldfield BJ, Giles ME, et al. (2000) Localization of angiotensin receptors in the nervous system. In: Quirion R, Björklund A, and Hökfelt T (eds.) *Handbook of Chemical Neuroanatomy, Volume 16: Peptide Receptors. Part I*, pp. 79–124. Amsterdam: Elsevier Science B.V.

Andersson B, Eriksson L, Fernandez O, et al. (1972) Centrally mediated effects of angiotensin II on arterial pressure and fluid balance. *Acta Physiologica Scandinavica* 85: 398–407.

Davisson RL (2003) Physiological genomic analysis of the brain renin–angiotensin system. *American Journal of Physiology* 285: R498–R511.

Denton DA (1982) *The Hunger for Salt*. Berlin: Springer.

DiBona GF (2001) Peripheral and central interactions between the renin–angiotensin system and the renal sympathetic nerves in control of renal function. *Annals of the New York Academy of Sciences* 940: 395–406.

Fitzsimons JT (1969) The role of a renal thirst factor in drinking induced by extracellular stimuli. *Journal of Physiology* 201: 349–368.

Fitzsimons JT (1998) Angiotensin, thirst and sodium appetite. *Physiological Reviews* 78: 583–686.

Johnson AK and Thunhorst R (1997) The neuroendocrinology of thirst and sodium appetite: Visceral sensory signals and mechanisms of central integration. *Frontiers in Neuroendocrinology* 18: 292–353.

Llorens-Cortes C and Mendelsohn FA (2002) Organisation and functional role of the brain angiotensin system. *Journal of the Renin–Angiotensin–Aldosterone System* 3(supplement 1): S39–S48.

McKinley MJ, Albiston AL, Allen AM, et al. (2003) The brain renin–angiotensin system: Location and physiological roles. *International Journal of Biochemistry & Cell Biology* 35: 901–918.

McKinley MJ, McAllen RM, Davern P, et al. (2003) The sensory circumventricular organs. *Advances in Anatomy, Embryology, and Cell Biology* 172: 1–127.

Sakai RR, Nicolaidis S, and Epstein AN (1986) Salt appetite is suppressed by interference with angiotensin II and aldosterone. *American Journal of Physiology* 251: R762–R768.

Stricker EM and Sved AF (2002) Control of vasopressin secretion and thirst: Similarities and dissimilarities in signals. *Physiology & Behavior* 77: 731–736.

Unger T and Scholkens BA (eds.) (2004) *Handbook of Experimental Pharmacology, Volume 163. Angiotensin*, 2 vols. Berlin: Springer.

Washburn DL and Ferguson AL (2001) Membrane properties of subfornical organ neurons. *Clinical and Experimental Pharmacology and Physiology* 28: 575–580.

Animal Communication: Honesty and Deception

S L Vehrencamp, Cornell University, Ithaca, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Definitions

Communication is defined as the transmission of information from a sender to a receiver via signals. True communication occurs when the information encoded in the signal enables the receiver to make decisions about behavioral responses and actions that subsequently benefit both the receiver and the sender. There are two ways in which this mutually beneficial exchange can break down in such a way that one party exploits the other: deception and eavesdropping.

Deception is the provision of inaccurate information by the sender such that the sender benefits from the interaction but the receiver pays the cost of a wrong decision. Types of deceit include 'exaggeration' or 'bluff' (using a signal whose rank among ordered alternatives is different from that for the corresponding condition values), 'lies' (use of the wrong signal among an unordered set of alternatives), and 'withholding information' (not giving a signal when appropriate). The evolution of such deceptive signaling is the focus of this discussion.

Eavesdropping typically occurs when a third-party receiver detects a signal directed at another receiver and uses the information to make decisions. The effect of eavesdropping on the sender could be positive or negative. For example, a countersinging interaction between two territorial male birds could provide information for eavesdropping females to make mate-choice decisions and for eavesdropping males to learn about the presence, relative dominance, or fighting ability of future rivals.

Both deception and eavesdropping can also take place in interspecific interactions. There are numerous examples of predatory species that mimic the mate attraction signal of their prey, but in such deceptive interactions between two different species, there is no selection pressure on the predator to be honest, and there is little the prey can do to avoid being exploited. The prey, as receiver, can try to improve its discrimination between true mates and impostors, but this process will simultaneously select for better mimicry by the predator. Similarly, predators can eavesdrop on the signals of their prey to locate their next meal. Prey can attempt to reduce the conspicuousness of these signals to predators while maximizing conspicuousness to conspecifics, but predators will counter with increased sensitivity. Such costs of signaling do affect signal design and receiver discrimination but are outside of this discussion of signal honesty. However, as shown below, some signals do appear to transmit honest information to heterospecifics for the benefit of both parties and provide a strong test of honest signaling models.

Honest signals, whether visual, vocal, olfactory, or tactile, are those in which some characteristic of the display (e.g., presence/absence, alternative forms, or a continuously varying parameter) is reliably associated with some attribute of the sender or its environment about which receivers want to know. Identifying this attribute specifies the kind of information transmitted by the signal. To fully demonstrate an honest signaling system, we would need to show that receivers not only attend to the signal and its variants but also benefit from knowing this information to make decisions. No signal is likely to be perfectly accurate. Senders make errors, but if errors lead to both positive and negative payoffs for the sender, we would not call it deception. Receivers will tolerate a certain level of sender error and deception as long as the signal is honest on average (see the section titled 'Do animals cheat?').

Context of Deception

When the sender and receiver in a signaling exchange both rank the payoffs of alternative receiver responses in the same order, selection will favor the accurate exchange of information within the limits of encoding, transmitting, and receiving error. However, sender and receiver often have conflicting interests because they rank the payoffs of alternative receiver responses differently. Under these conditions, animal senders will be tempted to provide misleading information so that the receiver performs the act that most benefits the sender. The strength of the selective pressure to deceive depends on the signaling context and the degree of conflict between the parties. Conflict of interest is greatest when two more or less equal competitors both desire the same nonsharable resource. Each would like the other to back down without a fight, and each would benefit from persuading the other that it is the better fighter by any means possible, including bluff. In the mate attraction context, both male and female benefit from mating with the correct species and therefore agree about the accurate transmission of species information. But females may want to mate only with a high-quality male, putting pressure on low-quality males to hide or exaggerate their quality. Similarly, an

offspring in a brood of siblings may exaggerate its need for food to the parent in order to garner a larger share of the food for itself. Even in cooperative groups, where all members benefit from coordinated flock cohesion, two individuals may disagree about the direction the flock should move and lie to ensure that their directive is heeded.

A Brief History

In the early days of ethology, signals were shown to evolve through the ritualization of behaviors that are or were functionally appropriate to the contexts in which the signals are now given. Signals were believed to be honest indicators of underlying motivations because they were derived from physiologically or anatomically linked sources. With the rise of evolutionary game theory in the 1970s, this notion of signal honesty was questioned. Why should a sender give an honest signal? Similarly, if senders wear their emotions on their shirtsleeves, what prevents a clever receiver from using that information to exploit the sender? Richard Dawkins and John Krebs suggested that senders were best characterized as deceitful manipulators trying to mask their true intentions and trick receivers into actions benefiting senders. Receivers in turn were best viewed as mind readers trying to discount false signals, anticipate the true intent of the sender, and thus identify their own best countermove. This scenario leads to a never-ending arms race, with increasing deceit and concealment of true intentions by senders parried by increased discrimination and exploitation by receivers. Except when sender and receiver have common interests, the resulting signals would be largely deceitful and uninformative.

Amotz Zahavi challenged this pessimistic view of signal honesty. He asserted that receivers have the upper hand and ought not respond to signals unless they carry some guarantee of honesty. One guarantee is to require that the signal impose a cost such that deceitful senders cannot afford to produce an exaggerated signal or that they produce it only in an ineffective way. Signals characterized by such costs are called handicap signals. Although Zahavi's idea was viewed skeptically at first, subsequent game theory models demonstrated the evolutionary feasibility of handicap signaling, and the handicap principle is now widely accepted.

Since the mid-1980s, several dozen game theory models of biological communication have been developed, each depicting different signaling contexts and comprising different game structures and sets of assumptions. The common feature in all the models that found at least some conditions for stable signaling was the assumption of some type of cost imposed on dishonest senders. Without such costs, senders become dishonest, receivers ignore the signals, and no evolutionarily stable state with informative communication signals can be attained. These costs ranged from signal production costs to receiver retaliation costs, reputation costs, and various types of physical and physiological constraints. The realization that the type of cost affects both the form of a signal and the specific kind of information it can encode then led to the useful classification of signals based on the type of cost. Independent attempts to classify signals in this way have largely converged and provide a very powerful framework for understanding the evolution and diversity of animal communication signals.

Classifying Signals Based on the Type of Cost That Guarantees Honesty

Table 1 summarizes the simplest scheme for the relationship between the type of cost that maintains signal honesty, the form or design of the signal, and the kinds of information the signal can encode. Three signal categories have been distinguished here: handicap, index, and conventional signals.

Table 1 Signal classification based on the type of cost that maintains signal honesty

Signal class	*Cost*	*Signal design*	*Information*
Handicap	Signal production, time lost, risk of predation	Graded display, intensity correlated with sender quality	Fighting ability, stamina, condition, territory quality, need, motivation
Index	Physical or physiological constraints	Form linked to sender attributes	Body size, age, strength, pointing at intended receiver
Conventional	Retaliation by receivers	Antithetical discrete or graded display, arbitrary form	Motivation to escalate versus retreat, aggressive intentions, fighting ability, condition

Based on Maynard Smith J and Harper D (2003) *Animal Signals*. Oxford, UK: Oxford University Press; Vehrencamp SL (2000) Handicap, index, and conventional signal elements of bird song. In: Espmark Y, Amundsen T, and Rosenqvist G (eds.) *Animal Signals: Signalling and Signal Design in Animal Communication*, pp. 277–300. Trondheim: Tapir Publishers; Hurd PL and Enquist M (2005) A strategic taxonomy of biological communication. *Animal Behaviour* 70: 1155–1170.

Handicap Signals

The key concept of the handicap principle is that the cost imposed on the sender should be one that 'uses up' the particular sender attribute about which the receiver wants information. Receivers are selected to pay attention only to those types of signals that impose costs linked to the type of information receivers need. Thus the form of the signal is linked to its information content, typically a continuously varying signal parameter correlated with some sender attribute. Directional selection pressure from receivers favoring the most costly signal variants often results in extreme elaboration and exaggeration of the display character. Handicap signals are strategic signals, in the sense that all senders can produce all signal variants in principle, but senders of poor quality or condition tend not to produce the more intense variants because of the high signaling cost. Figure 1 illustrates the classical handicap model, in which poor quality senders pay a higher cost for a given intensity of display. Their best option is to display at a lower intensity than a high-quality individual does, so signal intensity is reliably correlated with sender quality. Such signals are also called quality indicators and condition-dependent signals. Handicapping costs can be subdivided into production costs paid at the time of display and developmental costs paid earlier to grow the structures and organs needed for displaying. Several examples below show how such costs can be linked to useful information for receivers.

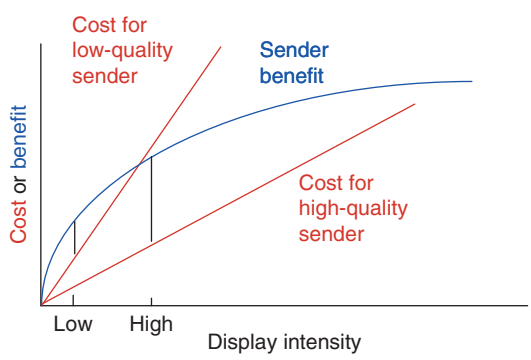

Figure 1 A graphical model for a handicap signal of sender quality. A key assumption is that displaying is costly, but high-quality senders pay a lower cost for a given display intensity than low-quality senders do. High- and low-quality senders receive the same benefit for a given display intensity. The optimal display level occurs at the point where the difference between benefit and cost is greatest. At equilibrium, high-quality senders display at a higher intensity than low-quality senders do, so display intensity is correlated with sender quality or condition. From Johnstone RA (1997) Evolution of animal signals. In: Krebs JR and Davies NB (eds.) *Behavioural Ecology: An Evolutionary Approach*, 4th edn., pp. 155–178. Oxford, UK: Blackwell Science.

Signals with energetically expensive production costs can inform receivers about the health, vigor, or foraging abilities of senders. For example, vocal and visual mate attraction signals must be repeated again and again. Female preference for males that not only produce high-quality displays but also repeat them at a high rate will increase the selection pressure on males to perform at the highest possible energetic level they can sustain. Unhealthy or poor-quality males cannot maintain such an expensive display rate, and this fact will be detectable to females. Choosy females will therefore benefit by selecting vigorous mates who are good genetic fathers or better parental providers or who possess food-rich territories (Figures 2 and 3). Repetitive countercalling contests between rival males can facilitate the assessment of relative condition, fighting ability, or motivation without their having to engage in a physical fight (Figure 4(a)). As a final example, energetic jumping displays are given by gazelles to approaching predators. Only individuals in good condition can afford to perform these actions well, and this provides honest information to the predators, which discourages them from chasing the more able displayers (Figure 4(b)).

Developmental costs are borne before the onset of display and can reveal either nutritional condition during development or more intrinsic aspects of genetic quality. The development of exaggerated display organs can also result in high maintenance costs for the bearer. A well-documented example is the elongated tail feathers of the barn swallow (*Hirundo rustica*) studied by Anders Møller. Females prefer males with longer tails, pairing much more quickly with males having artificially enlarged tails than with males with shortened tails. Long tails are a handicap for males; barn swallows are aerial foragers that capture flying insects on the wing. Artificial tail elongation increases the drag on the tail and reduces agility and foraging efficiency. Males with naturally long tails are stronger, healthier, and more parasite-resistant individuals who can not only grow long tails and cope with the foraging handicap but also pass on their parasite resistance to their genetic offspring. Females thus obtain better-quality offspring by selecting long-tailed males (Figure 5). Another good example of a handicap signal with a development cost is red coloration in birds and fish. Bearers of such color patches may also sustain the maintenance cost of increased conspicuousness to predators. Males in many songbird species possess repertoires of song types that they learn in the first months of life, and females have been shown to prefer larger repertoires and more complex songs. One promising explanation for this preference is the developmental stress

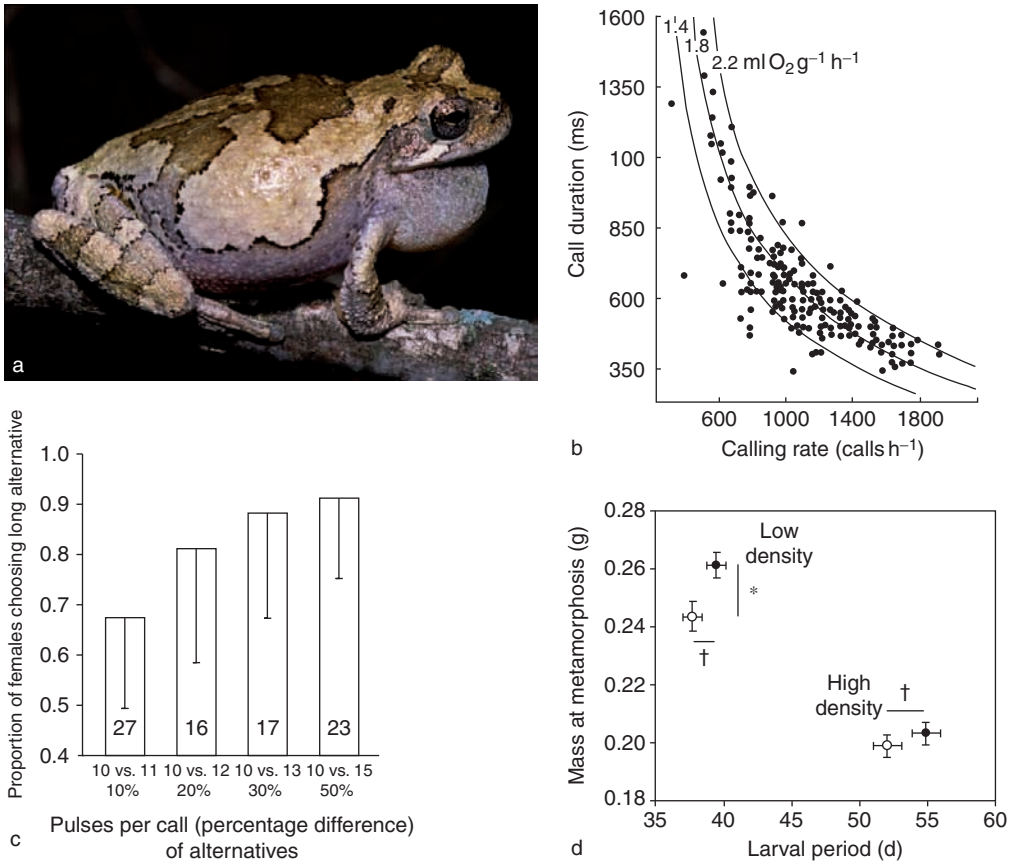

Figure 2 Call duration in the gray tree frog (*Hyla versicolor*) (a) is a costly handicap signal of male quality. (b) Calling is energetically expensive, with call duration and calling rate showing a trade-off in the field. An increase in both parameters is associated with an increase in oxygen consumption. Males increase their call duration when singing in dense choruses and when interacting vocally with other males. (c) Females strongly prefer males and experimental stimuli with longer call duration. (d) The offspring of females mated to long-calling males exhibit greater mass at metamorphosis than offspring of females mated to short-calling males, especially in contexts of low food density. In (c), error bars show the lower 95% confidence limits, and numbers in each bar indicate number of females in the sample. (d) The offspring of females mated to long-calling males (filled symbols) exhibit greater mass at metamorphosis than offspring of females mated to short-calling males (open circles), especially in contexts of low food density; error bars represent ± 1 standard error, asterisk indicates significance at $P < 0.05$, and daggers indicate $P < 0.10$. (a) Photo courtesy of H Carl Gerhardt. (b) From Wells KD and Taigen TL (1986) The effect of social interactions on calling energetics in the gray treefrog (*Hyla* versicolor). *Behavioral Ecology and Sociobiology* 19: 9–18. (c) From Gerhardt C, Tanner SD, Corrigan CM, and Walton HC (2000) Female preference functions based on call duration in the gray tree frog (*Hyla versicolor*). *Behavioral Ecology* 11: 663–669. (d) From Welch AM (2003) Genetic benefits of a female mating preference in gray tree frogs are context-dependent. *Evolution* 57: 883–893.

hypothesis, which argues that young birds require good nutrition to develop the brain nuclei for song learning and production.

Two other types of handicap models involve signaling costs but provide different kinds of information. Models of sender need, developed with the context of begging in mind, require a signal whose intensity is correlated with increasing immediate costs (energetic or predation risk). The key features that stabilize the correlation between signal intensity and need are the greater benefit of food for a needier sender and the fact that the beggar sender and the donor receiver are genetic relatives. The other handicap model variant was developed in the context of aggressive signals, in which the performance of a signal makes the sender vulnerable to attack by the receiver. This has been called a vulnerability handicap or an interaction handicap because it represents a hybrid between the handicap model and the conventional signal model (see the section titled 'Conventional signals'). Signals in this category typically vary in their tactical impact such that a strong signal places the sender at greater risk of injury but is more effective in threatening the opponent.

Index Signals

Handicapping is not the only mechanism for generating honest signals. Recent models suggest that if coevolving senders and receivers can hit on a cost-free signal that reliably correlates with important sender

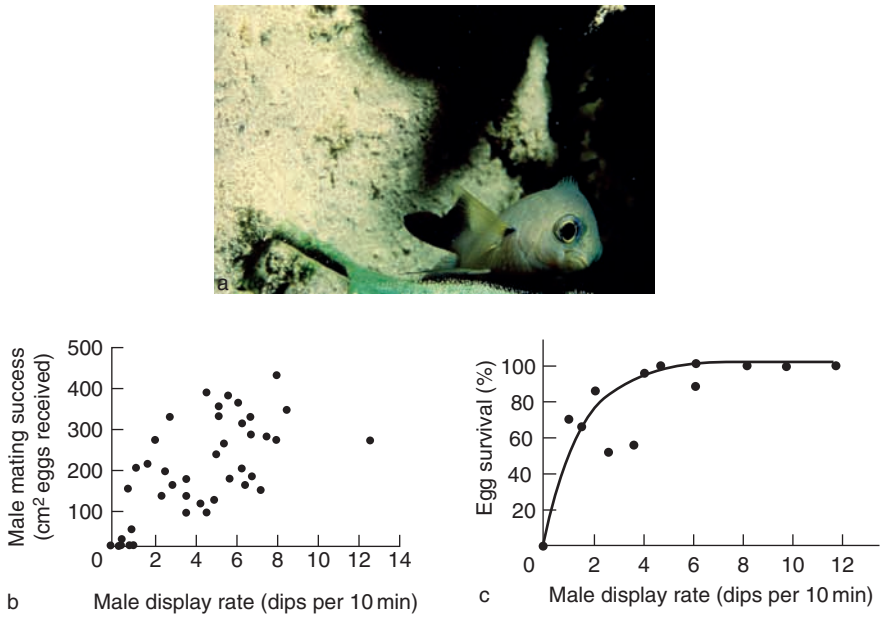

Figure 3 Example of display intensity as an indicator of male parental ability in the bicolor damselfish *Stegastes partitus* (a). This courting male is in the process of fertilizing recently laid eggs visible at the bottom. (b) Females prefer males that display at a high rate. Displaying is energetically costly and uses up calories, but males that are in good condition can afford high display rates and are more successful at guarding eggs (c). (a) Courtesy of Ken Clifton, with permission. (b, c) From Knapp RA and Kovach JT (1991) Courtship as an honest indicator of male parental quality in the bicolor damselfish, *Stegastes partitus*. *Behavioral Ecology* 2: 295–300.

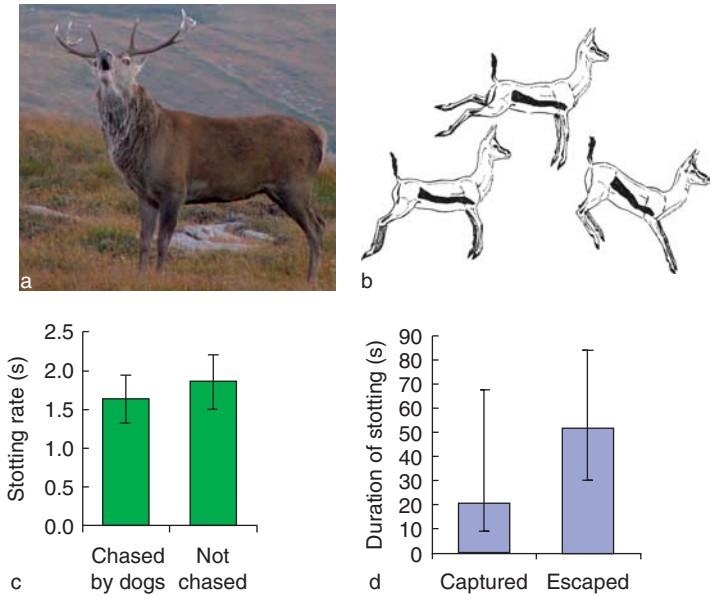

Figure 4 Examples of costly handicap signals in contexts other than mate attraction. (a) Male red deer (*Cervus elaphus*) engage in extended roaring contests with rivals. Roaring is an energetically expensive behavior that uses the actions and muscles employed during fighting. It is thought that only those males in good enough condition to fight can afford to roar at a winning level (Clutton-Brock TH and Albom SD (1979) The roaring of red deer and the evolution of honest advertisement. *Behaviour* 69: 145–170). (b) Stotting in Thomson's gazelles (*Gazella thomsoni*) involves leaping off the ground with all four legs held stiff and straight while running from predators (Walther FR (1969) Flight behaviour and avoidance of predators in Thomson's gazelle (*Gazella thomsoni* Guenther 1884). *Behaviour* 34: 184–221). They are far more likely to stot when chased by coursing wild dogs (78%) than when hunted by stalking cheetahs (9%). Wild dogs concentrated their chases on individuals that stotted at slower rates (c). Captured gazelles also stotted for shorter durations than gazelles that escaped (d). Gazelles were less likely to stot, and stotted at slower rates, during the dry season when food was lower and animals were in poorer condition. Stotting is therefore believed to be an honest signal to coursing predators of the prey's ability to escape capture. The error bars in (c) are ± 1 standard deviation and in (d) are interquartile ranges. (a) Photo courtesy of Alison Donald. (b) Reproduced from Walther FR (1969) Flight behaviour and avoidance of predators in Thomson's gazelle (*Gazella thomsoni* Guenther 1884). *Behaviour* 34: 184–221, with permission from Koninklijke Brill NV. (c, d) From FitzGibbon CD and Fanshawe JH (1988) Stotting in Thomson's gazelles: An honest signal of condition. *Behavioral Ecology and Sociobiology* 23: 69–74.

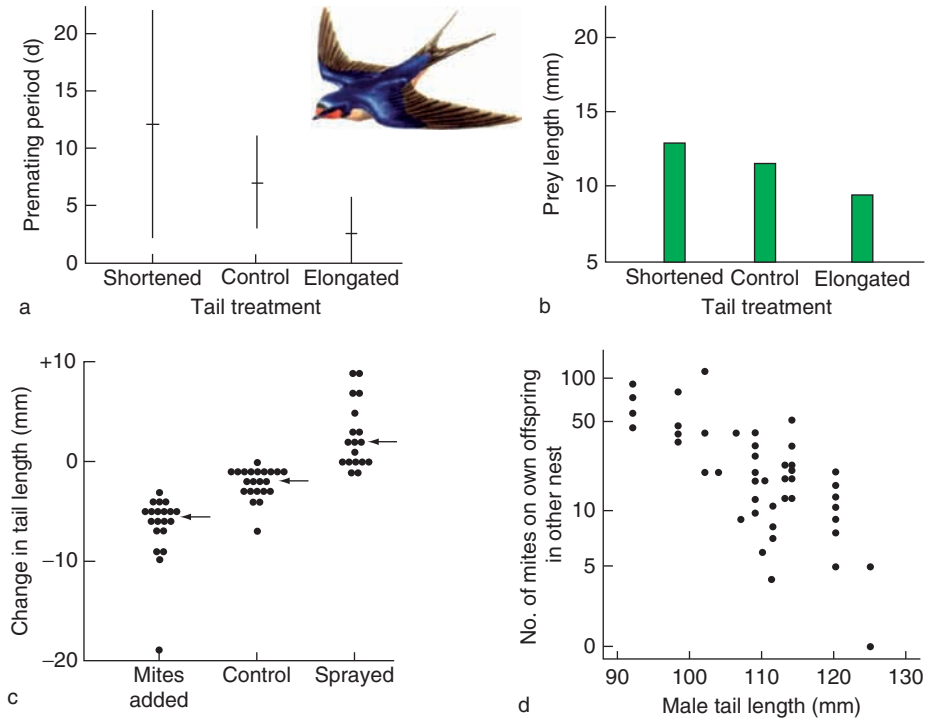

Figure 5 Tail length as a male quality indicator in the barn swallow (*Hirundo rustica*). (a) Females prefer (pair earlier with) males with experimentally lengthened tails. Error bars are ±1 standard deviation. (b) The cost of a lengthened tail in this aerial forager is increased drag, which reduces foraging efficiency and survivorship; males with naturally long tails suffer less of an aerodynamic cost than males with naturally short tails. (c) Long-tailed males possess fewer blood-sucking mites than do short-tailed males; removing the mites results in better tail growth. (d) Nestlings of males with longer tails and fewer parasites pass on their resistance to their offspring, even when they are raised in foster nests. (a) From Møller AP (1988) Female choice selects for male sexual tail ornaments in the monogamous swallow. *Nature* 332: 640–642. (b) from Møller AP, Lope F, and Lopez Caballero JM (1995) Foraging costs of a tail ornament. *Behavioral Ecology and Sociobiology* 37: 289–296. (c, d) from Møller AP (1990) Effects of hematophagous mite on the barn swallow *Hirundo rustica*: A test of the Hamilton and Zuk hypothesis. *Evolution* 44: 771–784, with permission from Blackwell publishing.

attributes, it will be favored over the costly signal. Some signals are constrained by physiology, anatomy, or physical principles to be honest indicators of certain types of sender attributes. Such signals are called unbluffable or index signals. Like handicaps, the form of the signal is strongly linked to the sender attribute of interest to the receiver. Unlike handicaps, they are not strategic signals, because some senders simply cannot produce some signal variants.

Examples of index signals include low-frequency vocal threats that are obligatorily linked with body size in many vertebrates, tailbeating displays in fish, and side-by-side head-high postures in ungulates. Similarly, push-pull and mouth-wrestling forms of ritualized fighting are reliable index signals of weight and strength. Signals that are linked with aging processes or health, such as silver backs in gorillas and song repertoire size in birds, indicate age, experience, and ability to survive (Figure 6). In song bird species with age-restricted (predispersal) song learning, song-type sharing between newly settled males and their neighbors is inversely correlated with dispersal distance and could serve as an index of fighting ability if good fighters are more likely to gain territories close to their tutors. Olfactory signals that are directly derived from reproductive hormones and by-products should also be classified as index signals. They directly reveal reproductive status in both males and females. Similarly, plant volatiles released by herbivores feeding on species-specific host plants are often used to attract conspecific mates. Although still controversial, alarm substances used by many schooling fish to disperse or cluster during predator attacks appear to be an antibacterial agent sequestered in the vacuoles of skin cells that is released on injury and would therefore qualify as an index signal.

Another cost-free constraint that leads to a type of index signal is an informational constraint. In this case, use of the signal is constrained by having access to some information. An example is stalked prey staring at a hidden predator, thereby signaling to the predator both its alerted state and the futility of continuing the hunt. The signal can be performed only by a sender who knows the location of the hidden predator.

Both index and handicap signals may have associated with them traits that make direct assessment of the relevant sender attributes easier or more accurate. Typically these are color patch markings or body

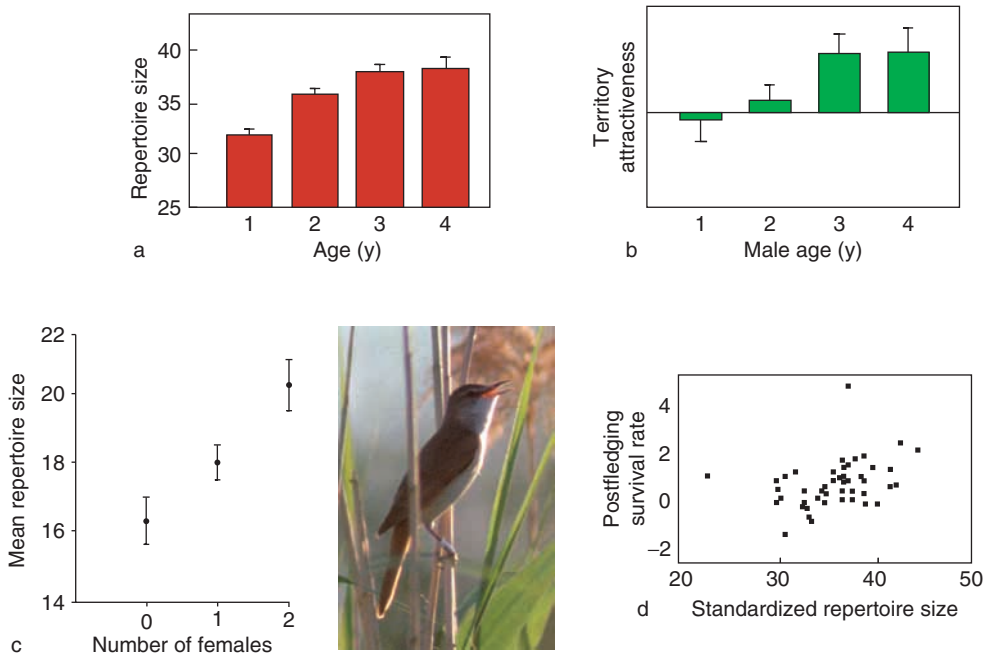

Figure 6 Repertoire size in an open-ended learner, the great reed warbler *Acrocephalus arundinaceus*, as an index of age. (a) Repertoire size increases with age for the first few years of life. (b) Older males possess better territories, as measured by occupancy and vegetation cover. (c) Males with larger repertoires (and better territories) attract more females in this polygynous species. Nine of ten females who sought extrapair fertilizations selected a neighboring male with a larger repertoire than their social mate. (d) The offspring of males with larger repertoires had higher survival, including return rate after migration, suggesting that larger-repertoire males pass on their inherently higher quality to their offspring. Error bars in (a), (b), and (c) are ±1 standard error. (a, b) From Hasselquist D (1994) Male attractiveness, mating tactics and realized fitness in the polygynous great reed warbler. PhD Thesis, Lund University, Sweden. (c) From Catchpole CK, Leisler B, and Dittami J (1986) Sexual differences in the responses of captive great reed warblers (*Acrocephalus arundinaceus*) to variation in song structure and size. *Ethology* 73: 69–77. (d) From Hasselquist D, Bensch S, and von Schantz T (1996) Correlation between male song repertoire, extra-pair paternity and offspring survival in the great reed warbler. *Nature* 381: 229–232. Photo courtesy of Bengt Hansson.

structures, called 'amplifiers,' that enhance visual perception of the physical attribute or movement display. Many fish have contrasting lines demarcating their body margins, which make assessment of body size easier. Similarly, a contrasting triangular marking on the abdomens of spiders makes assessment of nutritional condition, and thus fighting ability, easier. Contrasting stripes on the head in many birds, or outlining of the ears in many mammals, make it easier for a visual receiver to assess where a sender is looking.

Conventional Signals

Some communication signals are neither costly to produce nor obligatorily linked with physical properties of senders. The code by which these signals are associated with contexts is an arbitrary convention, and therefore these are called conventional signals. If there is no conflict of interest between sender and receiver, senders will not be tempted to cheat and conventional signals can be honest and stable without further guarantees. However, conventional signals are seen even during conflicts of interest, and in this case there must be a stabilizing cost that maintains honesty. The cost is receiver retaliation, which must come in the form of receiver skepticism or a retaliation rule in which receivers approach and check the sincerity of senders. In the most compelling model of conventional signaling with conflict of interest, both contenders must signal, and retaliate when both give a similarly strong signal, but retreat when the opponent gives a stronger one. Conventional signals are often discrete, with alternate antithetical signals for the strong versus weak message. This type of signal can convey information about aggressive motivation as well as fighting ability.

Prominent color patches in some birds and lizards are the classic example of this type of signal (Figure 7). The size or hue of the patch is correlated with the dominance rank of the individual, hence their designation as badges of status. Large badge size deters aggressive challenges by small-badged individuals. The cost of possessing a large badge is aggressive retaliation from other large-badged individuals. The evolution of such signals must be accompanied by frequent testing of the honesty of other individuals with a badge size similar to one's own and avoiding or ignoring individuals with larger or smaller badges.

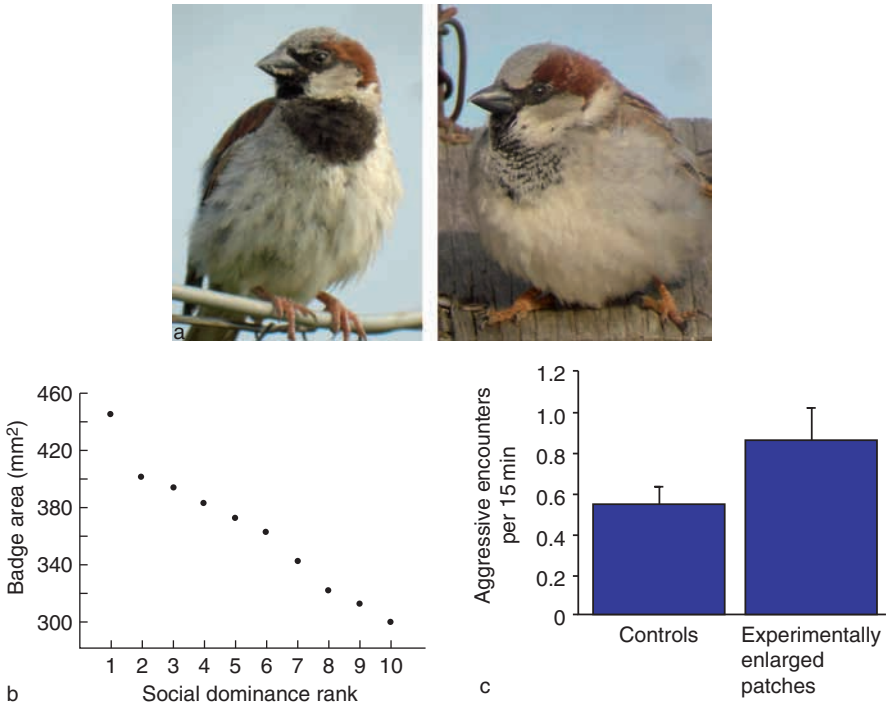

Figure 7 Badge of status in house sparrows, *Passer domesticus*. (a) The black bib on the chest of males varies greatly in size. (b) Size of the patch is strongly correlated with dominance rank. (c) Males with experimentally enlarged patches were involved in more aggressive encounters than were controls, and their attackers had larger than average patches. In Harris sparrows with a similar black chest patch, males with normally large bibs whose patch size was experimentally reduced had to fight very hard, but eventually won. Error bars in (c) are ±1 standard error. (a) Photo courtesy of Kevin McGowan. (b) From Møller AP (1987) Variation in badge size in male house sparrows. *Passer domesticus*: Evidence for status signalling. *Animal Behaviour* 35: 1637–1644. (c) From Møller AP (1987) Social control of deception among status signalling house sparrows *Passer domesticus*. *Behavioral Ecology and Sociobiology* 20: 307–311.

Such a rule makes it very dangerous and costly for a low-status individual to cheat by sporting a large badge. Certain vocal signals, such as song type matching and song type switching rate, are also believed to be conventional signals of aggressive intention.

Receiver skepticism is a type of receiver retaliation cost that can potentially stabilize honesty of conventional signals among group-living animals that recognize each other and interact repeatedly. Thus individuals who are observed to signal deceptively are remembered and tagged with a poor reputation or lack of trust, which may result in their failing to obtain certain benefits in the future. Reputation models with repeated interactions can even maintain honesty of cheap conventional signals when receivers have conflicting interests.

Do Animals Cheat?

Most signals are believed to be honest most of the time because conflicts of interest are minimal or appropriate costs are imposed on cheaters. But dishonest signaling does occur. Receivers will tolerate a certain amount of deception as long as they obtain a net gain by attending to signals (i.e., the probability that the signal is accurate times the payoff of making the right decision, plus the probability that the signal is inaccurate times the payoff of making the wrong decision, is positive). In theory, there are several contexts in which senders may be able to get away with some dishonesty. One context is a high level of perceptual error on the part of receivers, which favors less accurate mapping of sender quality on signal characteristics. A second context arises from a nonequilibrium outcome of dynamic coevolution between sender and receiver. A third model results when there are several classes of sender, each with its own costs and benefits of signaling, and receivers cannot distinguish between sender types; honest sender types must be sufficiently common relative to dishonest types to maintain the signaling system. A fourth situation occurs when receivers pay a very high cost for failing to respond to a signal, which leaves receivers vulnerable to deceit if senders figure out a way to benefit from the response. Documented examples of dishonesty fit several of these models.

Deceitful bluffing and exaggeration of body size or condition have been documented for competitive signals in several species. For example, mantis shrimp (*Neogonodactylus bredini*) defend their burrow from

intruders with a claw-spreading threat display, which is effectively given even when the resident has recently molted and has a soft exoskeleton. In snapping shrimp (*Alpheus heterochaelis*), chela (claw) size is positively correlated with body size and ability to win fights, but individuals with chelae larger than expected for their body size give more open-chela displays than individuals with smaller-than-expected chelae. Males of the mealworm beetle (*Tenebrio molitor*) release a pheromone that attracts female mates, but males whose immune system has been experimentally challenged with a foreign material implant increase their pheromone production and are even more attractive to females than normal males are. These examples may fit the multiple-sender model above, in which receivers cannot accurately distinguish the intrinsic quality of senders and honest senders are more common than dishonest ones.

A good example of an outright lie has been described in birds foraging in flocks; one individual may give a false alarm call to scare competitors away from a rich food find. Initially it was believed that senders would not 'cry wolf' too often, because receivers would learn to ignore the signal and it would cease to be effective in true alarm contexts. However, several studies have shown that the incidence of alarm calls in the absence of a predator is quite high, 55–63%. Although a certain amount of alarm unreliability is caused by errors (e.g., by young individuals who haven't learned which heterospecifics are truly dangerous), senders have been shown to give false alarms in which they clearly benefit, and the likelihood of false alarming varies with the costs and benefits to senders. The reason such a high level of dishonesty can be sustained is clearly a result of the high cost of a miss to receivers (i.e., death) relative to the cost of falsely fleeing and losing a bite of food, supporting the fourth model above. False alarm calls are also given by some male birds and squirrels during the fertile period of female mates to interrupt extra-pair copulations with competitors.

The withholding of information is a much more difficult form of deceit to document than is the provision of false information. Not only must one demonstrate a reliable association between a specific context, a signal, and a specific response, but one must also show that there is some behavioral flexibility on the part of senders to either give or fail to give the signal in different circumstances in a way that benefits the sender. This last point has been termed the 'audience effect' of signaling. Food calls may be given by males in some species (falsely at times) to attract mates, but food calls are always withheld in the presence of rival males. Food calls may also be withheld in group-living species that expect group members to share rich food finds. In this case, there needs to be a significant cost of withholding information to sustain mostly honest notification of food. Such a cost has been described only in primates who fail to advertise a rich food find. If other group members catch an individual feeding on such a find without having called, the individual is aggressively punished. But receivers may have a difficult time detecting a lie in this case, supporting the first model above.

In conclusion, low levels of dishonesty may persist in many signaling systems, but signals must be reliable enough on average to justify receiver response. Otherwise, receivers will ignore signals, senders will no longer benefit from giving dishonest information, and the signals will disappear from the species' repertoire. A variety of costs and constraints maintain signal honesty, and signals must be costly in a way that explains why they provide reliable information.

See also: Communication Networks and Eavesdropping in Animals; Game Theory and the Economics of Animal Communication; Pheromones and other Chemical Communication in Animals; Sexual Selection and the Evolution of Animal Signals; Signal Transmission in Natural Environments; Signal Design Rules in Animal Communication; Visual Signaling in Animals; Vocal Communication in Birds.

Further Reading

Catchpole CK, Leisler B, and Dittami J (1986) Sexual differences in the responses of captive great reed warblers (*Acrocephalus arundinaceus*) to variation in song structure and size. *Ethology* 73: 69–77.

Clutton-Brock TH and Albon SD (1979) The roaring of red deer and the evolution of honest advertisement. *Behaviour* 69: 145–170.

Fitch WT and Hauser MD (2003) Unpacking "honesty": Vertebrate vocal production and the evolution of acoustic signals. In: Simmons AM, Fay RR, and Popper AN (eds.) *Acoustic Communication*, pp. 65–137. New York: Springer.

FitzGibbon CD and Fanshawe JH (1988) Stotting in Thomson's gazelles: An honest signal of condition. *Behavioral Ecology and Sociobiology* 23: 69–74.

Gerhardt C, Tanner SD, Corrigan CM, and Walton HC (2000) Female preference functions based on call duration in the gray tree frog (*Hyla versicolor*). *Behavioral Ecology* 11: 663–669.

Hasselquist D (1994) Male attractiveness, mating tactics and realized fitness in the polygynous great reed warbler. PhD Thesis, Lund University, Sweden.

Hasselquist D, Bensch S, and von Schantz T (1996) Correlation between male song repertoire, extra-pair paternity and offspring survival in the great reed warbler. *Nature* 381: 229–232.

Hasson O (1994) Cheating signals. *Journal of Theoretical Biology* 167: 223–238.

Hurd PL and Enquist M (2005) A strategic taxonomy of biological communication. *Animal Behaviour* 70: 1155–1170.

Johnstone RA (1997) Evolution of animal signals. In: Krebs JR and Davies NB (eds.) *Behavioural Ecology: An Evolutionary Approach*, 4th edn., pp. 155–178. Oxford, UK: Blackwell Science.

Knapp RA and Kovach JT (1991) Courtship as an honest indicator of male parental quality in the bicolor damselfish, *Stegastes partitus*. *Behavioral Ecology* 2: 295–300.

Maynard Smith J and Harper D (2003) *Animal Signals*. Oxford, UK: Oxford University Press.

Møller AP (1987) Social control of deception among status signalling house sparrows. *Passer domesticus*. *Behavioral Ecology and Sociobiology* 20: 307–311.

Møller AP (1987) Variation in badge size in male house sparrows. *Passer domesticus*: Evidence for status signalling. *Animal Behaviour* 35: 1637–1644.

Moller AP (1988) Female choice selects for male sexual tail ornaments in the monogamous swallow. *Nature* 332: 640–642.

Møller AP (1990) Effects of hematophagous mite on the barn swallow. *Hirundo rustica*: A test of the Hamilton and Zuk hypothesis. *Evolution* 44: 771–784.

Møller AP, Lope F, and Lopez Caballero JM (1995) Foraging costs of a tail ornament. *Behavioral Ecology and Sociobiology* 37: 289–296.

Vehrencamp SL (2000) Handicap, index, and conventional signal elements of bird song. In: Espmark Y, Amundsen T, and Rosenqvist G (eds.) *Animal Signals: Signalling and Signal Design in Animal Communication*, pp. 277–300. Trondheim: Tapir Publishers.

Walther FR (1969) Flight behaviour and avoidance of predators in Thomson's gazelle (*Gazella thomsoni* Guenther 1884). *Behaviour* 34: 184–221.

Welch AM (2003) Genetic benefits of a female mating preference in gray tree frogs are context-dependent. *Evolution* 57: 883–893.

Wells KD and Taigen TL (1986) The effect of social interactions on calling energetics in the gray treefrog (*Hyla versicolor*). *Behavioral Ecology and Sociobiology* 19: 9–18.

Animal Intelligence: The Search for Animal Intelligence

A A Wright, University of Texas Health Science Center, Medical School at Houston, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Discoveries of obvious intelligent behavior by animals are appearing at an accelerated rate. Although examples of ape and monkey intelligence have been known for some time, even avians are casting off their 'bird brain' label. Pigeons, for example, can navigate by landmarks, sun compass, magnetic compass, infrasound sources, and possibly by cues from polarized and ultraviolet light. They can compute a navigational fix to determine their location (if unknown) to find 'home' which would be impossible with a compass alone. Galapagos finches and New Caledonian crows (Figure 1) select and fashion tools (e.g., cactus spines, twigs, and leaves) to 'fish' out insects and grubs in tree-limb holes/cracks or under leaf detritus and carry their favorite tools around with them. Clark's nutcracker birds store thousands of pine seeds in hundreds of cache sites (Figure 2), with sites varying yearly and often covered with snow during retrieved (Figure 3). Alex, a parrot, judges pairs of objects in the laboratory that differ in color, material, and/or shape and answers (in English) to the verbal query 'what's different' or 'what's same.'

Despite such astonishing demonstrations of animal intelligence, there is no answer to the question, What is animal intelligence? Theorists have proposed hierarchies of animal intelligence based on learning rates, learning phenomena (e.g., latent inhibition and tool use), discrimination learning (e.g., two-choice 'learning set' and serial discrimination reversals), abstract concept learning, and self-awareness, but none have withstood the test of time. Consequently, most researchers have turned to the study of specific cognitive phenomena (in laboratory settings, natural settings, social settings, etc.). Nevertheless, comparative cognition invariably gravitates to 'who can do what better.' Inevitably, the focus is on species differences and speculation returns to hierarchies, if but limited hierarchies (e.g., a branch of the evolutionary tree such as corvid or paridae birds). Even single-species studies are often (implicitly) compared to humans (e.g., "... the first demonstration in a nonhuman animal"). By contrast, if the behavioral and neurological mechanisms/processes were known, then similarities and differences should be apparent without resorting to a hierarchy ('better than'). In terms of mechanisms, similarities are as important as differences. Similar mechanisms can have somewhat different outcomes or vice versa. For example, a single test of some cognitive ability (e.g., abstract concept learning) might lead to a premature conclusion that some species (e.g., pigeon) does not possess this cognitive ability (or cognitive module). By contrast, if the study had manipulated critical parameters of concept learning, then it might have been revealed that this difference was only a quantitative difference, not the qualitative absence as concluded. These are among the issues that have guided our selection and discussion of animal cognitive phenomena. Some areas of research that have been considered relevant to animal intelligence are mentioned only in passing (encephalization quotient, consciousness, insight, associative learning, self-recognition in mirrors, animal communication, and animal language) because neither theory nor experiments have proven decisive and the field has moved on, and in other cases it is still too early to assess the reliability/validity and relevance to the field (e.g., theory of mind: seeing–begging, knower–guesser, and deception, social theory of intellect: dominance and imitation).

Stimulus-Bound Concepts

The movie '2001: A Space Odyssey' opens with the allegorical scene of primates discovering tool use and the 'touch stone' of intelligence. Tool use has long been considered a behavior whereby humans exploited their (greater) intelligence, passed down tool use over generations, eventually resulting in tools such as the wheel. Some animals in the wild, such as chimpanzees, use twigs to fish for termites and anvil and hammers to break open hard-shelled fruits/nuts. In laboratories, chimpanzees move stools and swing sticks to get hanging fruit or use rakes to pull food within reach, but this behavior does not seem to be learned by imitation. Interpretation of the tube-trap and stick problem (trap hole on one side of the food) is even less clear. Apes (chimpanzees, bonobos, and orangutans) and capuchin monkeys show successful use of the stick solving this problem, but capuchins, at least, apparently learn by trial and error without understanding which direction of stick poke will avoid the trap. With humans and apes atop the hierarchy of tool use, this seemed to be a promising index of intelligence until birds were also shown to use tools. Tool use by Galapagos finches and New Caledonian crows occurs in the laboratory and New Caledonian crows even construct tools, a feat

Figure 1 A New Caledonian crow with hooked tool for coaxing food items out of holes and cracks in tree limbs.

Figure 2 A Clark's nutcracker preparing to cache pinyon nuts. Courtesy of A Kamil.

Figure 3 A Clark's nutcracker in winter. Courtesy of A Kamil.

not shown for apes. But here too, captive finches do not seem to learn this tool use by imitation. If these species do not learn tool use by imitation, then the issue becomes how they learn it (i.e., learning mechanisms/processes). The same can be said for other remarkable animal behaviors that, contrary to early speculation, are apparently not learned by imitation (e.g., pilfering cream from milk bottles by English blue tits, pilfering food from backpacks by New Zealand keas, and washing potatoes by Japanese snow monkeys).

Navigation is typically based on relationships among landmarks and/or vector sources (magnetic, sun, and infrasound compasses). The nutcracker's cache retrieval is thought to be dependent on differential weightings of landmarks by a specialized cognitive module. Similar arguments have been made for rats and their spatial abilities. Some of the most remarkable feats of navigation, however, are by insects – animal species that would not stand very high on anybody's intellectual hierarchy. Honey bees, for example, receiving 'dance' communications about a food source strategically evaluate this information according to their knowledge. If the communicated location is an impossible location (e.g., middle of a lake), then they do not fly. If it is adjacent to an island in the middle of the lake, then they will alter the 'flight plan' and fly to the island. At issue is, do honey bees, or any animals, have cognitive maps (i.e., a mental representation of spatial relationships for flexible use)? Navigation in most instances (ants, wasps, bees, and rats) can often be explained by landmarks joined together by movements and local cues resulting in path integration and vector sums – a none too shabby intellectual feat in itself.

Animals (apes, monkeys, and parrots) can learn about number relationships. Monkeys learn to match similar numbers of objects in pictures independent of object size or cumulative area of the objects. They generalize this relational learning to new objects, new sizes, and new numbers and, in some instances, outside the range of numbers used to train the accurate behavior. A gray parrot vocally responds (in English) to the query "how many" and can accurately identify the number of one object type (e.g., keys) mixed with other objects (e.g., rocks). Apes accurately point to Arabic numerals (4, 5, 6, etc.) corresponding to the number of objects even if the objects (e.g., oranges) are hidden in several places. Monkeys too have been shown to identify correct addition (e.g., lemons) by looking longer at incorrect additions. These latter looking-time procedures have the advantage that they require little training, but it is not always clear what cognitive process accounts for such looking-time differences.

Transitive inference has been considered unique to some of the most intelligent species (e.g., humans, apes, and dolphins). Animals discriminate between successive pairs of stimuli taken from a hierarchical sequence of arbitrary stimuli designated by the following letters: A, B, C, D, E. For each successive pair, the choice of the stimulus highest on the hierarchy is rewarded (+) (e.g., A+ vs. B−, B+ vs. C−, and C+ vs. D−). At issue is whether performance will be accurate on a test of B versus D following training. That is, have they learned the hierarchy and can infer that B should be chosen over D without ever been trained on this pair? Although extensive training and a single test example can make for sometimes tricky interpretations, the evidence does indicate that monkeys, rats, jays, and pigeons in addition to humans and nonhuman primates seem to have this ability. Such information adds to the growing list of animal cognitive skills but it hardly lends itself to ranking animals on anything remotely approaching intelligence, and it is not clear whether this task is particularly well suited to investigating the underlying processes.

Animals can learn stimulus relationships and classify/categorize stimuli bound by a category, prototype, or equivalence class. Even pigeons can learn categories (go/no-go tasks) such as person, water, trees, fish, oak leaves, alphabetical letters, and impressionist paintings and as many as four categories (e.g., person, flower, car, and chair) simultaneously. They can accurately transfer to novel category examples. They learn the go/no-go discriminations more rapidly and more accurately if the pictures are grouped by category than when the same pictures are arbitrarily assigned, but they can learn (i.e., memorize) as many as 830 arbitrarily assigned pictures. Pigeons also learn artificial categories (e.g., artificial 'seeds' and geometrical colored shapes and backgrounds). Artificial categories are more amenable to feature analysis (i.e., controlling mechanisms/processes), but these results have often been at odds with feature theory (from human categorization).

Abstract Concept Learning

Abstract concepts are rules about stimulus relationships (e.g., identity). Animals are tested in the laboratory for abstract concept learning in matching-to-sample (MTS) tasks (e.g., sample plus two comparison/choice stimuli) and same–different (S/D) tasks (pairs of stimuli). Rhesus monkeys and apes tested with either junk objects or slide-picture stimuli have for decades shown complete/full concept learning (equivalent training and transfer performance) when tested with novel stimuli. Chimpanzees allowed to play with six different tin cups and six different metal locks showed immediately excellent performance with other stimuli in an MTS task, interpreted as spontaneous abstract concept learning. Although there is little doubt about primates' (apes, baboons, rhesus monkeys, and capuchin monkeys) abstract concept learning ability, other species have been considerably more problematical. Theories of animal cognitive ability and intelligence have been based on which species can (e.g., apes), cannot (e.g., avians), or partially (e.g., dolphins have this ability with auditory stimuli and monkeys with visual stimuli, but not vice versa) learn abstract concepts. Recently, however, this has changed. Dolphins and monkeys have been shown to learn abstract concepts with both visual and auditory stimuli. Even more damning to the hierarchical theories of abstract concept learning is that parrots, crows, rooks, nutcrackers, jays, and pigeons have all showed good (sometimes full) abstract concept learning. Indeed, the amazing feats of intelligent behavior by avians discussed here and elsewhere in this article are even more significant because of the different neural architecture for avians. For example, hemispheric lateralization of function in avians is at least as strong as it is in humans, yet avians have neither a neocortex nor a corpus callosum (and apparently little interhemispheric communication) and these two structures have been in the past thought to be highly correlated with intelligence. Since intelligence is thought to have evolved just like other bodily traits, if intelligence were related to brain size or certain neural structures, then the size of these brain structures should not be modifiable by experience. However, there is abundant evidence (mostly from rats) that size of the hippocampus, corpus callosum, and overall brain size can be altered by experience.

Regarding abstract concept learning, the pigeon is by far the species of greatest focus as well as the most controversial species. In MTS tasks, pigeons have shown full concept learning with cartoon stimuli; they can learn MTS in any of three different (strategic) ways: configural learning, if–then rule learning, or relational learning. Only relational learning leads to abstract concept learning, whereas the other two strategies produce stimulus-specific learning. Moreover, rats learning MTS with different odors (e.g., cinnamon) apparently learn the task relationally (the strategic process necessary for abstract concept learning), exhibit some characteristics of declarative/episodic memory, and require an intact hippocampus to do so.

Pigeons have shown evidence of S/D concept learning with computer-icon arrays, colored geometrical-shape arrays, and travel-slide pictures. In the first case, the degree of concept learning varied with array entropy, but when the array elements were reduced,

concept learning diminished and all but disappeared with the minimum of two elements. Work in my laboratory has shown that pigeons tested with travel-slide stimuli are capable of learning the S/D abstract concept with pairs of pictures. No concept learning was found with an 8-item training set. However, successive doublings of the training set improved concept learning to the point where concept learning (i.e., novel stimulus transfer) became equivalent to baseline performance with 256 or more training pictures (Figure 4). Similar experiments with capuchin and rhesus monkeys showed more rapid concept learning, and they achieved full concept learning with 128 pictures. These three species showed qualitative similarity in their ability to achieve full S/D abstract concept learning, but they also showed quantitative differences in their rate of concept learning. Pigeons required more exemplars to learn the rule (i.e., larger training set) than monkeys. Further analyses from successive cycles of training set expansion, learning, and transfer testing ruled out generalization (from the training stimulus pairs to the novel transfer pairs) as the mechanism responsible for transfer and concept learning, adding further evidence that these species have learned the higher order concept of S/D.

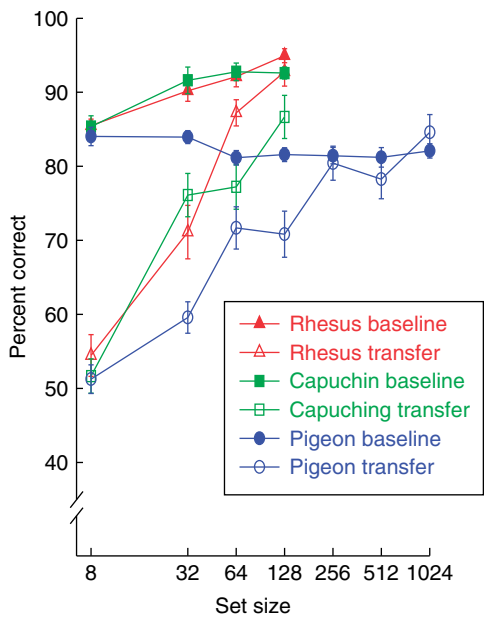

Figure 4 Mean percentage correct performance and standard errors of the mean for baseline performance (solid symbols) and transfer performance (open symbols) at each training set size for rhesus monkeys (triangles), capuchin monkeys (squares), and pigeons (circles). On each trial, a pair of vertically aligned pictures was presented simultaneously along with a white rectangle to the right of the lower picture. If the pair was the same picture, then a response to the lower picture was correct and rewarded. If the pair was two different pictures, then a response to the white rectangle was correct and rewarded. Incorrect choices were unrewarded. All subjects were trained with the same set of eight pictures (100 trials daily) until they were accurate at better than 80% correct and then tested with novel picture pairs (10 novel trials plus 90 training trials) for six sessions. The training set size was then doubled and transfer was tested with novel stimulus pairs after 85% correct performance. This sequence of training-set doubling, retraining, and retesting was repeated several times until transfer was equivalent to baseline (training) performance for each species.

'Higher order' is the watchword for species hierarchies of learning and concept learning in particular. Despite little agreement on most issues, most would probably agree that relations among relations (analogical reasoning) is the highest (most abstract, most difficult) form of abstract concept learning. Supporting the claim of exclusivity would be early evidence that the only animal species capable of this relational feat are chimpanzees because they alone have the ability to recode (tokens and artificial language) abstract relations. However, like previous claims of exclusivity of abstract concept learning, here too failures of other species in analogical reasoning may be the result of training and task failures not cognitive capacity failures. Indeed, baboons have shown similar results using icon arrays, and rhesus monkeys were able to accurately match musical tunes transposed one or two octaves with the first relation being the tunes (a relation among notes) and the second relation being that between the original and transposed tune.

Memory

There is no doubt that many animals have good memory, even by human standards. In addition to the amazing memory ability of nutcrackers to retrieve cached seeds in the natural alpine setting, laboratory studies of monkey memory have shown >70% correct accuracy for sample pictures in a delayed MTS task for 24 h. This good monkey memory occurred with trial-unique pictures so that the animals would not be confused (i.e., proactive interference) from seeing these pictures on previous trials. Although demonstrations of excellent animal memory capture our imagination, research on animal memory has shifted to studies of the memory processes (e.g., rehearsal, decay, consolidation, and interference), different types of memory (episodic, recollection, and familiarity), and brain mechanisms responsible for these different aspects and/or types of memory.

Related to how monkeys in the 24-h memory test might be confused by repeated pictures is what/how animals remember. Do they remember the actual presentation (episode) of the item at the time of test

(i.e., episodic memory), or are they just capable of (a vague sense of) familiarity? Familiarity is considered to be a more primitive memory process (hierarchically simpler), associated with less intelligent species, than is episodic memory. Episodic memory and familiarity may be different kinds of memory possibly mediated by different cognitive modules involving different specialized brain areas and neural circuits. Most research on different memory brain areas has been conducted with monkeys and rats and has focused on the medial temporal lobe structures (hippocampus and perirhinal and parahippocampal cortices) for processing of information, the amygdala for emotional significance, and the prefrontal cortex for awareness and planning based on one's own memory (metacognition) and how the memory information will be used. Despite the vast amount of work, only tentative glimpses of a coordinated picture are beginning to emerge. Perhaps the perirhinal cortex might be most critical for object memory, whereas the parahippocampal cortex might be most critical for spatial memory. The hippocampus and prefrontal cortex may be critical for episodic memory through their role in integrating and abstracting information from many other brain areas, including the perirhinal and parahippocampal cortices, thereby associating context information (e.g., 'when and where' information) with the primary memory events ('what' information) and producing evidence of recollection/episodic memory as opposed to item familiarity.

Although human participants can report "I remember" versus "I just know" to indicate episodic versus familiarity memory, respectively, more objective and clever procedures are required for testing animals for these different types of memory. Although it would seem logical for animal research on episodic/familiarity memory to have been done with (nonhuman) primates due to their relative evolutionary proximity to humans, the best and most clever work is actually being done with avians. An additional advantage of working with avians on this issue is that success will have more profound implications because most hierarchical theorists would have apes and/or monkeys being capable of episodic memory – not the proverbial bird brain. Scrub jays have been shown to cache and recover perishable wax worms and nonperishable peanuts in distinctive halves of sand-filled ice cube trays (Figure 5). After 4-h delays, the jays first recovered the more desirable wax worms, but after 124-h delays they first recovered peanuts because they had learned that wax worms would deteriorate in 124 h. Thus, they remembered what (peanuts vs. wax worms), when (4 vs. 124 h), and where (which tray side) the foods were stored – requirements of episodic memory. Many supporting experiments have replicated these remarkable findings and further explored the processes responsible for this episodic-like memory. Nevertheless, some experts are skeptical that these amazing memory feats by scrub jays may be merely the result of stimulus–response associative learning. However, vast numbers of associative learning studies have shown that associative learning is much more than a collection of stimulus–response reflexes or habits. Rats and other animals in classical conditioning experiments anticipate the future and can even add and subtract anticipated time through complex combinations of forward and backward primary and secondary conditioning. They learn expectations of certain outcomes (what), at certain times (when), and in certain places (where). In these situations, expectations develop over multiple trials (i.e., conditioning) and become reference or generic memory. By contrast, the jays cache a particular item, in a particular place, at a particular time – hence a single episode. Part of the single-episode success may be the act of caching. Caching would be expected to enhance memory for a particular episode, much like the generation effect in human memory. However, the same can be said for taste-aversion memory, which can be a single associative event, a particular stimulus (food taste), and a particular time (as much as several hours or more) before getting sick. Moreover, it has been shown that taste aversion is cognitively flexible (not an automatic stimulus–response association) by manipulations of which of several foodstuffs become associated with sickness. Few would question whether taste aversion is episodic memory because of numerous reports of humans, sometimes decades later, recalling exactly when and under what conditions some previously liked food was inadvertently associated

Figure 5 A scrub jay retrieving previously cached wax worms, a food preferred over peanuts at short cache-retrieval times before they deteriorate. Courtesy of NS Clayton.

with illness. These so-called 'autonoetic' (self-knowing) reports are considered by some to be the hallmark of episodic memory which, by design or theory, would seem to secure humans as the only animal that definitively possesses episodic memory.

Nevertheless, a delayed-MTS experiment showed that rhesus monkeys, at least, can be consciously aware (metacognition) of their own memories. During retention delays monkeys opted for a memory test and a chance for highly desired reward (peanuts) or else a less desired but certain pellet. Surprise tests showed that they were more accurate when they opted for the test, with the interpretation that they knew when their own memory was good. The key to this interpretation was having the monkeys monitor their memory before the test was presented and thereby avoid confounds of recognition familiarity.

The issues of whether animals have episodic memory, metacognition, or autonoetic experiences aside, animals most certainly have working memory. Indeed, how could any animal learn anything without working memory? Working memory is what is remembered on a particular trial; reference memory is the rules of the task. Good memory observed in natural settings has been tested in laboratory analogues. Nutcrackers have been tested in large arenas where they made as many as 25 caches and accurately recovered these cached seeds 6 months later from 180 possible sites (sand-filled cups). Other caching birds (e.g., tits and chickadees) have been tested in arenas filled with artificial tree limbs where the birds cached seeds in several cloth-covered tree limb holes and later remembered these cache sites from as many as 97 available sites. Many caching birds (scrub jays and nutcrackers) will re-cache if they are observed by another bird or alter their retrieval order (e.g., ravens) to beat the would-be pilferer to the cache site. In at least one case (scrub jays), they re-cache only if they themselves have a history of pilfering from others. Thus, they impute to the observing bird (typically a conspecific) their own pilfering intentions – a prototypical example of so-called 'theory of mind.'

Laboratory tests (e.g., radial arm maze analogues and MTS video monitor images) of single-item memory of some of these caching birds have not always found such remarkably good memory compared to other related species that do not cache. As previously noted, perhaps the act of caching is critical to the good memory. Moreover, the size of the hippocampus does not seem to be highly correlated with the degree of caching or performance on laboratory memory tasks. Current work seems to be more focused on how memory works for these bird species (i.e., processes/mechanisms of landmark navigation) rather than which one has the best memory.

Memory is seldom limited to single items. Certainly in the real world there are collections or streams of things to remember. The study of (human) memory began with memory lists. The hallmark of list memory studies is serial position function (SPF) and its primacy and recency effects (good memory for beginning and ending items, respectively), often revealing a U-shaped SPF. Although animals are notoriously difficult to train in list memory studies, U-shaped SPFs with visual stimuli have been shown for apes, rhesus monkeys, squirrel monkeys, capuchin monkeys, rats, and pigeons. Research on some of the mechanisms/processes responsible for the SPFs has shown that with short four-item lists there is a strong recency effect (last items) at very short retention intervals and little or no primacy effect. As the retention interval is lengthened, the primacy effect develops and the recency effect wanes. Research in my laboratory has shown that these same changes occur for rhesus monkeys, capuchin monkeys, pigeons, and humans (**Figure 6**). The qualitative similarity in these SPFs is complemented by a quantitative difference where the time course of changes takes place in approximately 10 s for pigeons, 30 s for monkeys, and 100 s for humans. The developing primacy effect is counterintuitive because it is an (absolute) increase in memory performance with increasing retention delay. SPFs for auditory stimuli have been shown for dolphins and rhesus monkeys. Rhesus monkeys were tested with short natural/environmental sounds at the same retention delays and list lengths as other rhesus monkeys with visual memory. The auditory SPFs were opposite to the visual SPFs, showing initial strong primacy effects and no recency effects. The recency effects grew with retention delay and the primacy effects waned. Other tests revealed that interference among the list items produced retrieval inhibition and that this inhibition dynamically changed with retention interval. These animal SPFs show some of the mechanisms/processes involved in visual and auditory list memory of these species. They also show that memory is not a unitary system, memory does not decay with time, and rehearsal does not produce the primacy effect – all long-standing tenets of the so-called modal model of memory.

The two examples of work from my laboratory have shown qualitative similarity in abstract concept learning and visual memory processing across such diverse species as humans, New and Old World monkeys, and pigeons. Single tests of these cognitive abilities would likely have shown differences that might have been taken as evidence for a species' differences in

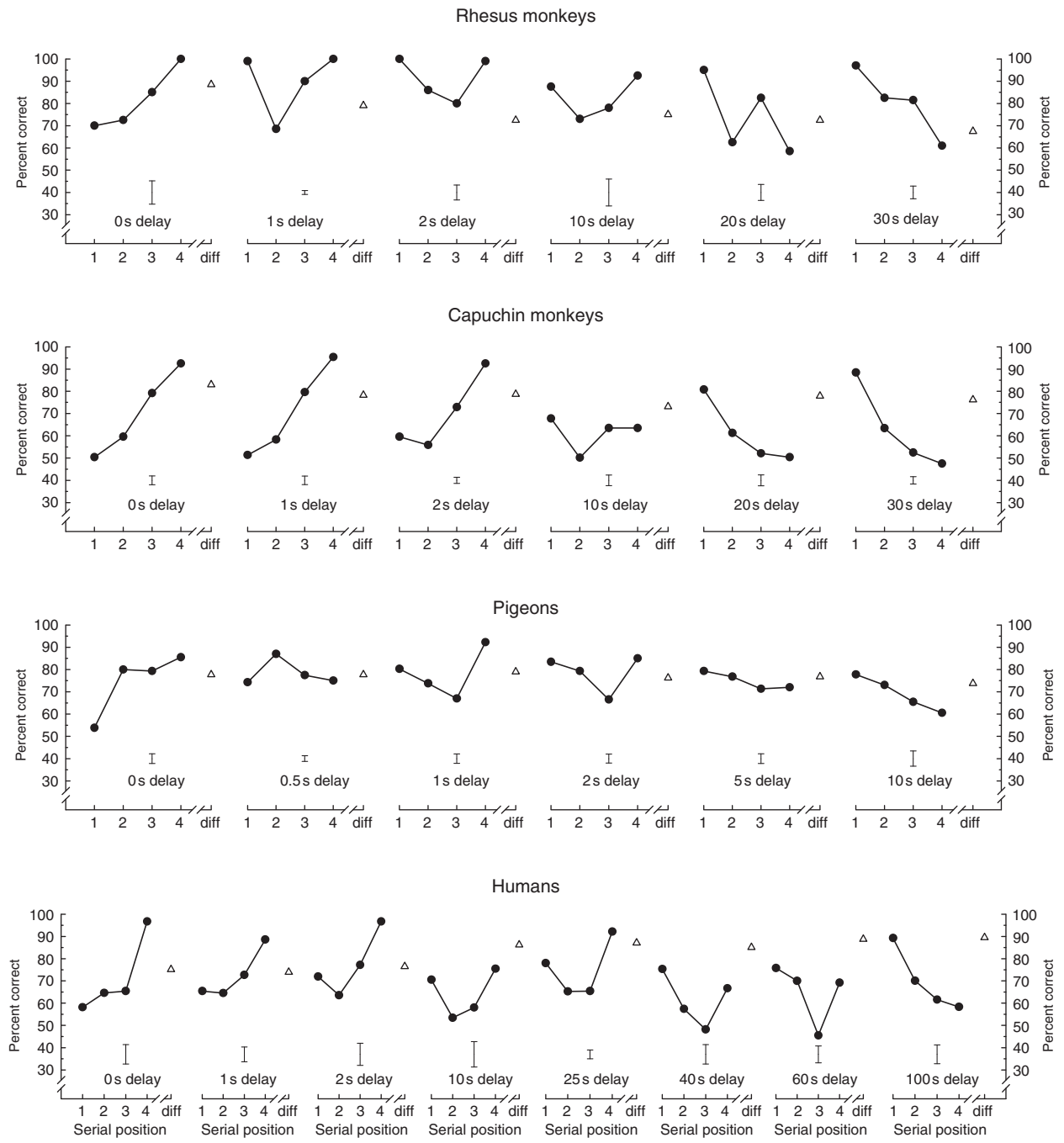

Figure 6 Serial-position memory functions for capuchin monkeys, rhesus monkeys, pigeons, and humans. On each trial, four pictures were briefly presented followed by a retention delay which is indicated in the lower portion of each panel. Subjects responded to indicate their memory of whether a single test picture was or was not in the list. The serial position functions show the results for trials in which the test matched one of the list pictures. (Serial position 4 was the last list item presented.) On half the trials, the test matched no list picture and that performance is shown as 'diff' (different). Test items for monkeys and pigeons were travel slides selected from more than 3000 items; test items for humans were kaleidoscope pictures selected from more than 550 items. The items for all subjects were unique on each daily test session. The error bars are the average standard errors for the four serial positions.

cognitive capability, evolved cognitive modules, and perhaps intelligence. There are differences among species to be sure, but in these cases the differences were quantitative differences. Some of the species (e.g., pigeons) required more exemplars of the rule to learn the abstract concept than monkeys, and in terms of memory, some species had differences in the time course changes of their serial position functions.

See also: Episodic Memory: Assessment in Animals; Executive Function and Higher-Order Cognition: Assessment in Animals; Numerical Intelligence: Neural Substrates; Procedural Learning in Animals; Reasoning and Problem Solving: Models; Referentiality and Concepts in Animal Cognition; Spatial Memory: Assessment in Animals.

Further Reading

Clayton NS, Bussey TJ, and Dickinson A (2003) Can animals recall the past and plan for the future? *Nature Reviews Neuroscience* 4: 685–691.

Emery NJ (2006) Cognitive ornithology: The evolution of avian intelligence. *Philosophical Transactions of the Royal Society B* 361: 23–43.

Olson DJ, Kamil AC, Balda RP, and Nims PJ (1995) Performance of four seed-caching corvid species in operant tests of nonspatial and spatial memory. *Journal of Comparative Psychology* 109: 173–181.

Rogers LJ and Kaplan G (eds.) (2004) *Comparative Vertebrate Cognition*. New York: Kluwer Academic/Plenum.

Shettleworth SJ (1998) *Cognition, Evolution, and Behavior*. New York: Oxford University Press.

Wasserman EA and Zentall TR (eds.) (2006) *Comparative Cognition*. New York: Oxford University Press.

Wright AA (1997) Concept learning and learning strategies. *Psychological Science* 8: 119–123.

Wright AA (1998) Auditory and visual serial position functions obey different laws. *Psychonomic Bulletin and Review* 5: 564–584.

Wright AA and Katz JS (2006) Mechanisms of *same/different* concept learning in primates and avians. *Behavioural Processes* 72: 234–254.

Animal Models of Alzheimer's Disease

J Koenigsknecht-Talboo and D M Holtzman, Washington University, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Alzheimer's disease (AD) is the most common cause of dementia. The pathological hallmarks of this neurodegenerative disease include extracellular structures called diffuse and neuritic plaques as well as amyloid angiopathy. In addition, there are intracellular structures known as neurofibrillary tangles and neuropil threads. The plaques are comprised primarily of aggregated, nonfibrillar, and fibrillar forms of amyloid-β (Aβ) peptide, a 38–43-amino-acid peptide that is a normal cleavage product of the amyloid precursor protein (APP). The neurofibrillary tangles and neuropil threads are composed of hyperphosphorylated, aggregated forms of the microtubule-associated protein tau. In addition to neuropil threads, there is also neuritic dystrophy consisting of large axonal and dendritic varicosities, surrounding fibrillar amyloid deposits in the brain parenchyma. Neuronal cell death and synaptic degeneration also occur in AD. Furthermore, gliosis invariably occurs whereby microglial cells and astrocytes accumulate and cluster around plaques of deposited Aβ. Genetic mutations in the APP and presenilin (PS) genes have been identified and are associated with rare forms of early onset, autosomal-dominant familial AD as well as cerebral amyloid angiopathy (CAA). The normal processing of APP or the sequence of Aβ is altered in the presence of these mutations, thus leading to abnormal production or an increased aggregation propensity of Aβ and its subsequent accumulation. While mutations in tau do not result in AD, they do cause certain forms of fronto-temporal dementia and tangle accumulation. These mutations appear to alter the likelihood that tau will aggregate leading to a tauopathy. Advances in research and further understanding of AD have increased in part due to the generation of animal models that mimic certain aspects of AD, CAA, and fronto-temporal dementia that take place in humans. Introduction of mutations in APP, PS, and tau using transgenic, knockout, and knockin techniques in mouse models has resulted in the production of animals that develop aspects of AD-like pathology and other changes (Figure 1). These transgenic animals directly test whether expression of mutant genes leads to AD-like pathology and behavioral phenotypes, provide insight into molecular mechanisms, and are useful for developing better diagnostic methods as well as testing the efficacy of new treatments.

The presence of a few important criteria is vital to model aspects of AD in animal models. An AD animal model should exhibit at least one if not more than one pathological hallmark of AD. Another criterion is observation of cognitive and behavioral deficits. In models that contain mutations that cause familial AD or that result in expression of genes that lead to AD-like changes, changes in phenotype should be more severe in genetically altered than in age-matched wild-type animals and occur in an age-dependent fashion. The findings seen in a good animal model should be reproducible and be confirmed by several different labs. The generation of a comprehensive animal model of AD has proved to be a challenge for various reasons. AD is a complex disease that thus far involves at least three proven causative genes and one gene that is a known risk factor. In addition, there are also other genes that are required for various aspects of AD pathology. Another obstacle for constructing an AD model is mimicking a human disease that normally occurs in the seventh decade of life or later in mice, which live only 2–3 years. It is also extremely difficult to reproduce all the complexities of human behaviors in rodents. Various model organisms have been used to try and replicate aspects of human AD. Each organism has its strengths and weaknesses in regard to this endeavor. Although the anatomy of a mouse brain is similar to other mammals including humans in many regions, the brain of a rodent is also very different than a primate brain. Mice can be induced to develop plaques and tangles in region-specific areas; however, mice do not normally develop plaques and tangles. Aged rhesus monkeys have been shown to contain similar brain pathology as AD including the presence of aggregated amyloid, reactive glia, and enlarged distal axons. However, monkeys are expensive to maintain and experiments can take a long time to complete. Both worms and flies provide the benefits of easy and fast breeding in addition to the ability to employ powerful genetic tools. Yet, the brain anatomy of worms and flies is vastly different than in humans and AD-like behavior is difficult to address in these organisms. Therefore, much of the advancement in AD research in animal models has come from studies utilizing genetically manipulated mice. These models are reviewed here.

Genetically Modified Animal Models

APP Mouse Models and Aβ-Related Pathology

While multiple different APP transgenic mice have been produced, several examples are described below

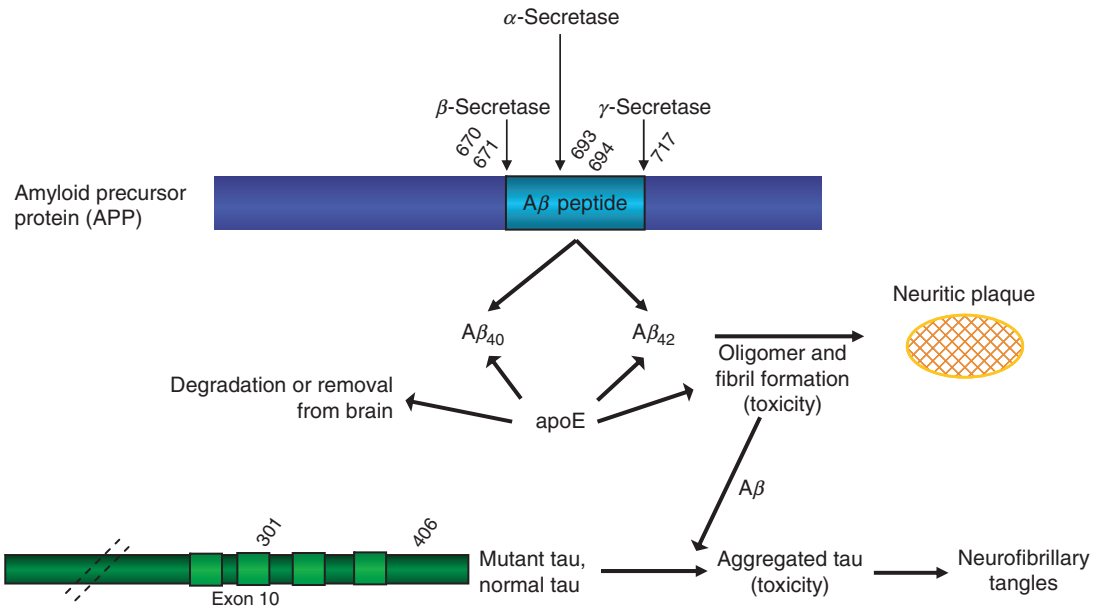

Figure 1 Model for Aβ and tau leading to cell toxicity. Aβ is generated by cleavage of APP by β-secretase and γ-secretase. Oligomers and fibrils of Aβ form and likely lead to neuronal and synaptic damage. Aβ also is hypothesized to induce tau aggregation and toxicity. The numbered sites in APP and tau correspond to amino acids where mutations are present that cause familial autosomal dominant AD or fronto-temporal dementia, respectively, that have been introduced into the mouse models described herein.

to illustrate some key features of the different types of models. One of the first APP transgenic mouse models to develop AD-like pathology described was the PDAPP mouse model of AD. The PDAPP mouse was generated by employing a platelet-derived growth factor (PDGF)-β promoter to induce overexpression of a transgene construct of human APP that contains the V717F mutation predominantly in neurons in the central nervous system (CNS). This point mutation in APP results in a mutation first found in families in Indiana which causes a form of autosomal dominant familial AD. Human APP levels were reported to be 10 times higher than endogenous mouse APP levels in this animal model. PDAPP mice develop age-dependent Aβ accumulation and deposition in diffuse as well as fibrillar neuritic plaques in the hippocampus, cerebral cortex and corpus callosum, as well as to a lesser extent in cerebral arterioles. This pattern of deposition corresponds with the pattern of Aβ plaques revealed in the brains of AD patients. Aβ deposition generally develops between 6 and 9 months and continues to accumulate with age. Different forms of Aβ deposition occur in this animal model. Diffuse plaques with nonfibrillar as well as compacted plaques containing fibrillar Aβ are both observed. In this model, only about 10% of the plaques are fibrillar, as defined with dyes such as Thioflavin-S or Congo red. In addition to Aβ deposition, these animals also exhibit robust gliosis in which plaques are associated with reactive astrocytes and microglial cells. Neuritic dystrophy has been reported in the PDAPP mice; however, studies have not reported a loss of neurons in this model. A decrease in synaptic and dendritic density in the hippocampus has also been observed in this model. It is not clear whether this is a developmental or neurodegenerative phenotype in this model. While behavioral studies have revealed some cognitive deficits in aspects of learning and memory in these mice prior to AD pathology, there are also age-dependent and pathology-dependent behavior deficits in learning and memory, making this likely to be a useful model for certain aspects of cognitive deficits in AD as well as a model of the Aβ-related pathological changes.

The Tg2576 transgenic mouse model was also one of the first APP transgenic mouse models developed. Also known as APPsw, the Tg2576 mouse model was generated by overexpression of the human APP695 construct that contains a double Swedish mutation. This mutation was identified in a large Swedish family that presented with early onset AD. The Swedish mutations (K670N/M671L) are driven by the hamster prion protein in this model. This mouse model expresses 5–6 times the amount of human APP as endogenous mouse APP. Senile plaques with dense cores of fibrillar amyloid in addition to diffuse amyloid deposits have been reported in the Tg2576 animal model. Deposition of amyloid generally occurs between 9 and 12 months of age in Tg2576 mice. Clusters of astrocytes and microglia surround many of the amyloid deposits. These cells induce an inflammatory response surrounding the areas of amyloid

plaques. Despite the presence of amyloid deposits, neurofibrillary tangles are not present in any of the APP transgenic models including the Tg2576 mouse model. Neuritic dystrophy is present in this and other APP transgenic mice but there is no marked neuronal loss. Neuronal loss in the immediate vicinity of fibrillar amyloid has been reported. CAA has been shown to be more prevalent in the Tg2576 model than in PDAPP mice. Notable CAA has also been observed in the APP23 transgenic mouse model along with micro- and sometimes macrohemorrhages. Cognitive decline has been demonstrated to increase with age in this mouse model. For example, spatial learning in a water maze was shown to be impaired in Tg2576 mice. Also, results of a spatial alternation task in this mouse model were consistent with hippocampal damage. These behavior deficits have been reported to occur just prior to the onset of plaque deposition, beginning at ~6 months of age. Learning and memory was normal at 3 months of age in the water maze. Deficits in synaptic plasticity such as long-term potentiation (LTP) have been reported in the hippocampus and the dentate gyrus. Since development of the PDAPP and Tg2576 mouse models, numerous other APP transgenic mouse models that develop Aβ deposition and associated changes have been developed.

The PSAPP mouse model incorporates a PS mutation and was generated by crossing an APP mutated mouse (Tg2576) with a PS1 mutated mouse (M146L). This mouse model exhibits accelerated plaque pathology as well as robust gliosis most likely due to enhanced Aβ42 generation or an altered 42/40 ratio due to the overexpression of a PS1 mutation responsible for familial AD. Abundant diffuse and fibrillar Aβ deposits form in the cortex and hippocampus of these double transgenic mice much earlier than their Tg2576 singly transgenic littermates. By 3 months of age, these mice contain fibrillar plaques within the cortex. Amyloid deposition accumulates and plaque size increases and extends to the hippocampus by 6–8 months of age. This double transgenic mouse model preferentially expresses Aβ42, and brains of these mice have been reported to have a 41% increase in Aβ42 as compared to singly transgenic Tg2576 mice. Neurofibrillary tangles have not been reported in the PSAPP mouse model. Despite the accelerated amyloid pathology in PSAPP mice, the clearest cognitive deficits appear at 15–17 months of age as reported by some. Other models combining expression of a PS mutation together with APP have been developed.

Mouse Models with Mutations in the Aβ Region

A few mouse models have been generated that contain mutations within the Aβ sequence of APP. For example, the APP Dutch mouse model was created by inducing a point mutation, E693Q, in human APP. This mutation, in humans, results in an autosomal dominant form of CAA. Expression of APP Dutch results in a higher level of Aβ40 accumulation than Aβ42. Similar to the human cases, amyloid is deposited predominantly in cerebrovascular plaques in APP Dutch animals in contrast to the parenchymal Aβ deposits that are characteristic of other APP transgenic animal models. The APP Dutch animal has been crossed with animals containing the G384A PS1 mutation, which is known to increase Aβ42 production. The expression of the G348A PS1 mutation in these mice results in a shift of amyloid deposition from the vessels into the brain parenchyma in the cortex and hippocampus. It is hypothesized that a high Aβ42/40 ratio favors parenchymal deposition, whereas a low Aβ42/40 ratio favors vascular Aβ deposition. Similar to the Dutch mutation, the Iowa mutation (D694N) also leads to Aβ accumulation in the brain vasculature. These two mutations have been reported to augment fibrillogenic properties of Aβ *in vitro* and may also result in abnormal clearance of Aβ from the brain leading to its accumulation.

The Arctic mutation, E693G, is another mutation that is within the Aβ sequence and causes an AD phenotype in humans. Transgenic mice that contain the Arctic mutation in addition to the Swedish mutation have been reported to develop intraneuronal Aβ aggregates prior to the deposition of extracellular Aβ. The intraneuronal accumulation of Aβ is not grossly fibrillar. The Arctic lines promote accumulation of shorter species of Aβ. This was demonstrated by a higher Aβ1-38/Aβ1-40 ratio and a lower Aβ1-42/Aβ1-40 ratio. The Arctic mutation is extremely amyloidgenic; one of the Arctic lines generated was reported to contain extracellular neuritic plaque deposition of Aβ as early as 2 months of age. A significant difference between the Arctic and the Dutch mutations is the location of Aβ deposition. The Dutch mutation results in amyloid deposition primarily in vessels; however, the Arctic mutation generally results in parenchymal Aβ plaques within the brain.

All of the APP models discussed result in production of different ratios and amounts of Aβ species, making it difficult to determine the specific effects of Aβ40 versus Aβ42 on pathology and behavior. In addition, in some models the overexpression of APP may be responsible for some of the changes in the animals. An important advancement in AD animal models was achieved by creating mice that specifically produce Aβ40 or Aβ42, known as the BRI-Aβ models. This model employs fusion constructs that contain 243 amino acids of BRI protein followed by a sequence that encodes either Aβ40 or Aβ42. Aβ is produced when the BRI protein is cleaved

by furin in a vesicular compartment in cells. Aβ40 is selectively produced, expressed, and secreted by BRI-Aβ40 mice, whereas Aβ42 is selectively generated, expressed, and secreted by BRI-Aβ42 mice. Only BRI-Aβ42 mice develop plaques demonstrating the critical importance of Aβ42 in the seeding and propagation of Aβ deposition. These models will be useful in sorting out specific effects of Aβ40 or Aβ42 without the confound of APP overexpression.

ApoE Mouse Models

The only proven genetic risk factor for AD is apoE genotype. The ε4 allele of apoE has been demonstrated to be a risk factor for late onset AD and CAA, whereas the ε2 allele of apoE has been shown to be protective against AD. apoE avidly binds Aβ and co-localizes with neuritic plaques and CAA. Evidence suggests that one of the main reasons for the linkage between Aβ and apoE results from the ability of apoE to bind Aβ, acting as a chaperone to influence Aβ metabolism and structure.

In regard to a potential role for apoE in normal brain structure and function, apoE knockout mice have been shown in some studies but not others to contain less synaptophysin and staining with dendritic markers. Interestingly, when either PDAPP or Tg2576 mouse models are bred with apoE knockout mice, the resultant APP transgenic mice lacking apoE exhibit a striking decrease in Aβ deposition compared to animals that express apoE, thereby demonstrating the importance of apoE in amyloid deposition. In addition, the Aβ deposits in mice lacking apoE are not fibrillar until very old ages and contain little to no neuritic dystrophy, microglial inflammation, CAA, or CAA-associated hemorrhages. This suggests that apoE plays an important role in induction of fibrillar Aβ and its consequences.

To examine the effects of individual human apoE isoforms on the brain and on AD mouse models, mice have been generated that contain a specific apoE isoform knocked into the endogenous mouse *apoe* locus. These mice display apoE expression predominantly in glial cells, the normal pattern of endogenous apoE expression. The apoE2, apoE3, and apoE4 mice all express similar levels of apoE. Other mice have been generated that express specific apoE isoforms in astrocytes with a GFAP promoter (GFAP-apoE) and in neurons with a neuron-specific endolase promoter (NSE-apoE). Each of these mice were bred with mouse apoE knockout mice to eliminate the confound of endogenous mouse apoE. In the NSE-apoE4 mice, C-terminal truncated fragments of apoE accumulate in an age-dependent manner. Fragment accumulation occurs to a lesser extent in NSE-apoE3 mice and is not seen in the GFAP-apoE mice. Tau phosphorylation also accumulates in an age-dependent manner in NSE-apoE4 mice. As with apoE fragment accumulation, tau phosphorylation occurs to a lesser degree in NSE-apoE3 mice and is not observed in the GFAP-apoE mice. Co-cultures of hippocampal neurons and astrocytes derived from GFAP-apoE mice displayed greater neurite outgrowth in GFAP-apoE3 mice than in GFAP-apoE4 mice. Importantly, in regard to effects on Aβ, using human apoE isoform-specific transgenic or knockin mice bred to different APP transgenic mice, strong isoform-specific effects of apoE are seen. The isoform-specific effects on Aβ are such that earlier and more Aβ-related pathology occurs in the pattern E4 > E3 > E2. In addition to the isoform-specific pattern, all human apoE isoforms delay the development of Aβ deposition relative to the presence of murine apoE or no apoE. This suggests that human apoE plays a role in Aβ clearance or transport in addition to its role in fibrillogenesis. While the effects of apoE on Aβ are profound, experiments with animals that develop both Aβ deposition and tauopathy will be required to determine if apoE also influences tangle formation.

Tau Mouse Models

Several AD animal models have focused on APP/Aβ; however, other models have been developed that examine other aspects of AD pathology. The first tau transgenic model was generated prior to the identification of pathogenic tau mutations. This model expresses the longest human brain tau isoform under the control of the human Thy-1 promoter. Although this model mimicked some aspects of AD including hyperphosphorylation of tau, neurofibrillary tangles did not develop in these animals. This AD animal model may represent an early AD-like phenotype that occurs prior to tangle formation.

Pathogenic mutations were later identified in the tau protein that cause fronto-temporal dementia and parkinsonism linked to chromosome 17 (FTDP-17). One tau mouse model employs the most common FTDP-17 mutation, P301L, driven by the prion protein promoter. The JNPL3 mouse develops neurofibrillary tangles in addition to astrogliosis, most prominently in the brain stem and spinal cord. Although tangles do not develop in the cortex or hippocampus of JNPL3 mice, there is pretangle pathology in these brain regions. Contrary to the APP mouse models of AD, this mouse model does not develop amyloid plaques. The majority of these mice developed motor and behavioral deficiencies by 10 months of age. The behavioral deficits described in JNPL3 mice include delayed righting reflex, decreased locomotion, and muscular weakness. Another tau mouse model contained a P301L mutation; however, this model

employed the Thy1.2 promoter. These mice expressed P301L tau in neurons in the cortex and hippocampus. Hyperphosphorylation of tau occurs in these mice and neurofibrillary tangles are present. In addition, this mouse model also develops numerous TUNEL-positive neurons in the cortex suggesting neuronal damage.

Another tau mutation, R406W, was utilized to make a tau transgenic mouse model. This transgenic mouse expressed the R406W mutation driven by the α-calcium-calmodulin-dependent kinase-II promoter. Hyperphosphorylated tau inclusions are present in neurons in the forebrain of these animals by18 months of age. These tau inclusions were closely associated with disruptions in microtubules. Associative memory was reported to be impaired in this mouse model through contextual and cued fear conditioning tests. This tau mouse model was not reported to have obvious motor deficits, as has been the case with other tau mutations. A variety of mice with tau mutations causing fronto-temporal dementia have now been generated that have pathological features similar to those described above, although the location of the pathology varies depending on the promoter utilized as well as the mutation.

Htau mice were generated by crossing mice that express a human tau transgene and tau knockout mice. These mice express the normal sequence of human tau. The htau mice therefore express six isoforms of human tau, however, does not express mouse tau. These mice develop pathology despite the fact that they contain nonmutant tau. Hyperphosphorylated tau accumulates within the cell bodies and dendrites of neurons in the hippocampus. Tau aggregated and accumulated inside cells within the cortex and the hippocampus of this mouse model.

Recently, a novel tau model was described that expresses a mutant tau which can be suppressed with doxycycline. These mice overexpress the P301L tau mutation in the forebrain. The expression of mutant tau resulted in the accumulation of hyperphosphorylated tau and a loss of neurons that increased as the animals aged. Spatial memory was also impaired in association with age. Blocking tau expression of tau caused memory function to improve and neuron numbers to stabilize. Importantly, neurofibrillary tangles continued to accumulate, after the mutant tau was suppressed despite improved cognitive performance. Results from this mouse model therefore suggest that a soluble form of tau may be contributing strongly to toxicity.

Other Mouse Models

An advance in AD mouse models was made with the introduction of a novel triple transgenic model (3x Tg-AD) which progressively developed both amyloid plaques and neurofibrillary tangles in brain regions that are affected in AD. The 3x Tg-AD model contained a PS1 mutation (M146V), an APP mutation (APPswe), and a tau mutation (P301L). These mice were generated by microinjecting two transgenes into single cell embryos derived from homozygous M146V knockin mice. One benefit of this method is that the APP and tau transgenes are co-integrated and transmit together to the offspring. Plaque formation was shown to develop in these animals prior to the production of tangles. This pathology is consistent with the amyloid cascade hypothesis. The neocortex of 3x Tg-AD mice has been reported to contain intracellular Aβ immunoreactivity as early as 3–4 months of age which is followed by Aβ aggregation in the hippocampus at 6 months of age. In a similar region-dependent manner, extracellular Aβ deposits formed in the cortex prior to the hippocampus as early as 6 months of age and were easily detectable at the age of 12 months. In contrast to plaque formation, tau pathology of tangle formation appears first in the hippocampus and later is observed in the cortical area of the brains of 3x Tg-AD mice. Synaptic dysfunction was also observed in an age-dependent manner in this mouse model. Interestingly, removal of Aβ by immunotherapy in these mice reverses early changes in tau such as hyperphosphorylation but not later changes such as tangle formation.

Another mouse model that results in both tau and amyloid pathology has been generated by crossing the Tg2576 mouse with the P301L tau mouse model. The resulting mutant tau and APP mice, TAPP mice, develop both amyloid plaques and neurofibrillary tangles. The plaques generated in the TAPP mice were similar in number and distribution as was reported in Tg2576 mice. Aβ plaques were detected as early as 6 months of age and increased as the animals aged. Neurofibrillary tangles were seen in the TAPP mice as early as 3 months of age; however, they were more numerous as the animals aged. There was an acceleration of tau pathology in the TAPP mice as compared to singly transgenic mice only expressing tau with the P301L mutation suggesting that Aβ may somehow exacerbate tangle pathology. Although plaques and tangles were shown to be located in the same brain regions, the neurofibrillary tangles did not form in the immediate environment of the amyloid plaques. Tangle pathology also occurs in brain areas not affected by Aβ deposition in TAPP mice. Dystrophic neurites were observed around the amyloid deposits. Similar to the P301L mutated mice, the TAPP mice demonstrated some disturbances in motor skills. It was also observed in TAPP mice that neurofibrillary tangle pathology developed earlier and more robustly in female mice than in male mice.

Conclusions

Various animal models have been developed which recapitulate one or more aspects of AD. While many of these models are useful, none of the models encompass all of the pathology and cognitive deficits observed in human AD. A key feature of AD that is not readily demonstrated in the mouse models is the presence of plaques and tangles combined with progressive and severe neuronal loss that is seen in certain brain regions in human AD. More models with AD-like pathology can also be bred with mice in which other genes are altered to assess their effects on aspects of AD-like pathology or behavior. Alternatively, expression of genes with viral vectors can be another way to both assess effects of potential gene therapies as well as better model aspects of AD. Important information has been gained from crossing AD models with apoE models and with mice lacking or overexpressing β-secretase, α-secretase, ABCA1, neprilysin, and insulin degrading enzyme. There are many other proteins that potentially play a role in AD that can be assessed in a similar manner. Although there is not yet a perfect AD animal model, valuable information on AD pathogenesis and potential therapies such as secretase inhibitors and immunotherapy have been gleaned from the current available models. It is likely that future improvements in current models as well as a better understanding of AD pathogenesis will continue to be made as the full complement of genes involved in AD is unraveled. Rodent models will likely play a major role in this endeavor.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Alzheimer's Disease: Neurodegeneration; Alzheimer's Disease: An Overview; Alzheimer's Disease: Transgenic Mouse Models; Axonal Transport and Alzheimer's Disease; Dementia.

Further Reading

Allen B, Ingram E, Takao M, et al. (2002) Abundant tau filaments and nonapoptotic neurodegeneration in transgenic mice expressing human P301S tau protein. *The Journal of Neuroscience* 22(21): 9340–9351.

Bales KR, Verina T, Dodel RC, et al. (1997) Lack of apolipoprotein E dramatically reduces amyloid beta-peptide deposition. *Nature Genetics* 17(3): 263–264.

Brendza RP, O'Brien C, Simmons K, et al. (2003) PDAPP; YFP double transgenic mice: A tool to study amyloid-beta associated changes in axonal, dendritic, and synaptic structures. *The Journal of Comparative Neurology* 456(4): 375–383.

Cheng IH, Palop JJ, Esposito LA, et al. (2004) Aggressive amyloidosis in mice expressing human amyloid peptides with the Arctic mutation. *Nature Medicine* 10(11): 1190–1192.

Davis J, Xu F, Deanes R, et al. (2004) Early-onset and robust cerebral microvascular accumulation of amyloid beta-protein in transgenic mice expressing low levels of a vasculotropic Dutch/Iowa mutant form of amyloid beta-protein precursor. *The Journal of Biological Chemistry* 279(19): 20296–20306.

Fagan AM, Murphy BA, Patel SN, et al. (1998) Evidence for normal aging of the septo-hippocampal cholinergic system in apoE (–/–) mice but impaired clearance of axonal degeneration products following injury. *Experimental Neurology* 151(2): 314–325.

Fryer JD, Taylor JW, DeMattos RB, et al. (2003) Apolipoprotein E markedly facilitates age-dependent cerebral amyloid angiopathy and spontaneous hemorrhage in amyloid precursor protein transgenic mice. *The Journal of Neuroscience* 23(2): 7889–7896.

Games D, Adams D, Alessandrini R, et al. (1995) Alzheimer-type neuropathology in transgenic mice overexpressing V717F beta-amyloid precursor protein. *Nature* 373(6514): 523–527.

Gotz J, Chen F, Barmettler R, and Nitsch RM (2001) Tau filament formation in transgenic mice expressing P301L tau. *The Journal of Biological Chemistry* 276(1): 529–534.

Herzig MC, Winkler DT, Burgermeister P, et al. (2004) Abeta is targeted to the vasculature in a mouse model of hereditary cerebral hemorrhage with amyloidosis. *Nature Neuroscience* 7(9): 954–960.

Holcomb L, Gordon MN, Mcgowan E, et al. (1998) Accelerated Alzheimer-type phenotype in transgenic mice carrying both mutant amyloid precursor protein and presenilin 1 transgenes. *Nature Medicine* 4(1): 97–100.

Holtzman DM, Bales KR, Tenkova T, et al. (2000) Apolipoprotein E isoform-dependent amyloid deposition and neuritic degeneration in a mouse model of Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* 97(6): 2892–2897.

Hsiao K, Chapman P, Nilsen S, et al. (1996) Correlative memory deficits, Abeta elevation, and amyloid plaques in transgenic mice. *Science* 274(5284): 99–102.

Lewis J, Dickson DW, Lin W-L, et al. (2001) Enhanced neurofibrillary degeneration in transgenic mice expressing mutant tau and APP. *Science* 293(5534): 1487–1491.

Lewis J, McGowan E, Rockwood J, et al. (2000) Neurofibrillary tangles, amyotrophy and progressive motor disturbance in mice expressing mutant (P301L) tau protein. *Nature Genetics* 25(4): 402–405.

Lord A, Kalimo H, Eckman C, Zhang X-Q, Lannfelt L, and Ng Nilsson L (2006) The Arctic Alzheimer mutation facilitates early intraneuronal Abeta aggregation and senile plaque formation in transgenic mice. *Neurobiology of Aging* 27(1): 67–77.

Oddo S, Caccamo A, Shepherd JD, et al. (2003) Triple-transgenic model of Alzheimer's disease with plaques and tangles: Intracellular Abeta and synaptic dysfunction. *Neuron* 39(3): 409–421.

Santacruz K, Lewis J, Spires T, et al. (2005) Tau suppression in a neurodegenerative mouse model improves memory function. *Science* 309(5733): 476–481.

Sturchler-Pierrat C, Abramowski D, Duke M, et al. (1997) Two amyloid precursor protein transgenic mouse models with Alzheimer disease-like pathology. *Proceedings of the National*

Academy of Sciences of the United States of America 94(24): 13287–13292.

Sun Y, Wu S, Bu G, et al. (1998) Glial fibrillary acidic protein-apolipoprotein E (apoE) transgenic mice: Astrocyte-specific expression and differing biological effects of astrocyte-secreted apoE3 and apoE4 lipoproteins. *The Journal of Neuroscience* 18(9): 3261–3272.

Tatebayashi Y, Miyasaka T, Chui D-H, et al. (2002) Tau filament formation and associative memory deficit in aged mice expressing mutant (R406W) human tau. *Proceedings of the National Academy of Sciences of the United States of America* 99(21): 13896–13901.

Relevant Website

http://www.alzforum.org – Alzheimer Research Forum.

Animal Models of Amnesia

M Moss, Boston University School of Medicine, Boston, MA, USA

© 2009 Published by Elsevier Ltd.

Early Contributions to the Study of Memory: The Pioneers

Perhaps the first major contribution to the formal literature on human memory came from the French psychologist Theodule Ribot in his treatise on the *Diseases of Memory* that produced his 'law' which states that older memories are better remembered than more recent memories. But it is also difficult to introduce this subject without mention of the seminal contribution by William James in his *Principles of Psychology*. James made the case that the processing of memory had a neural substrate and that memory could be fractionated into primary (short-term) and secondary (long-term) memory, concepts we still struggle with today. He also introduced what we now consider the two major systems of memory processing: declarative memory and habit (procedural memory). Nor can any introduction to the subject of memory ignore the contribution of Hermann Ebbinghaus, who made the first major attempt at the quantification and measurement of memory, which culminated in his work *On Memory*. Sergei Korsakoff described a disorder of memory associated with the effects of alcohol and that now bears his name, and Vladimir Bekhterev was the first to point to the possible relationship of memory impairment with damage to the temporal lobe.

While these individuals were laying the foundation for the study of human memory, others began to apply this body of work to animal models. Most notable among these were EL Thorndike, a student of James, and Ivan Pavlov, with his work on classical conditioning. It was also at this time that Shepard Franz began using formal training paradigms in his work on the effects of brain lesions in animals. His student, Karl Lashley, extended this work to a level that many now consider to have constituted the seminal phase of the utility and value of animal models in neuropsychology.

Amnesia: The Paradox and Animal Models

Perhaps the most informative and seminal case study in the field of amnesia and one that initiated major activity using animal models of amnesia was that of patient H.M. In 1953, H.M. underwent surgery for treatment of intractable seizures. The procedure entailed bilateral resection of the medial temporal lobes, including the anterior two-thirds of the hippocampus, the amygdala, and hippocampal gyrus (Figure 1). Although the surgery was successful in ameliorating the frequency and intensity of his seizures, H.M. was left with a devastating impairment in his anterograde short-term memory and to a limited extent his retrograde memory, extending to at least several months preceding the procedure.

Following the surgery, H.M. was able to remember events over very brief intervals. However, it was the fate of this immediate memory that was imperiled. After a period as short as 30 s, or with brief distraction, recall of immediate events was, in virtually all cases, permanently erased. Despite this striking condition, it soon became evident through formal experimentation that H.M. retained some learning capacity. H.M. was able to perform certain motor skill and perceptual learning tasks and could retain much of this learning for as long as a year. It was this case that cemented the theoretical notions that the ability to remember is subserved by at least two memory systems. It was Cohen and Squire who dubbed the impaired and spared memory in H.M. (and other amnesiac patients) as declarative and procedural systems, respectively. Accordingly, declarative memory has been ascribed to medial temporal lobe structures and their anatomic connections to polymodal association areas of the cerebral mantle, whereas procedural memory has been ascribed to nonmedial temporal regions, with the neostriatum and cerebellum as the leading proposed neural substrates.

To date, although studies in animals have explored the neural basis of nondeclarative memory, most have focused primarily on localizing and fractionating the anatomic and functional neuronal substrate of declarative memory. Hence, the thrust of research has focused on the hippocampal formation and the adjacent parahippocampal cortices (including the entorhinal, perirhinal, and parahippocampal cortices; see Figure 2).

Rodent Models of Memory Function

In light of the anatomic and connectional similarities of the medial temporal lobe across species, both rodents and nonhuman primates (as well as canines and felines) have been utilized to assess various aspects of memory function. To a great extent, behavioral paradigms that are equivalent in design have been adapted to each of these species. One such behavioral task is delayed nonmatching to sample.

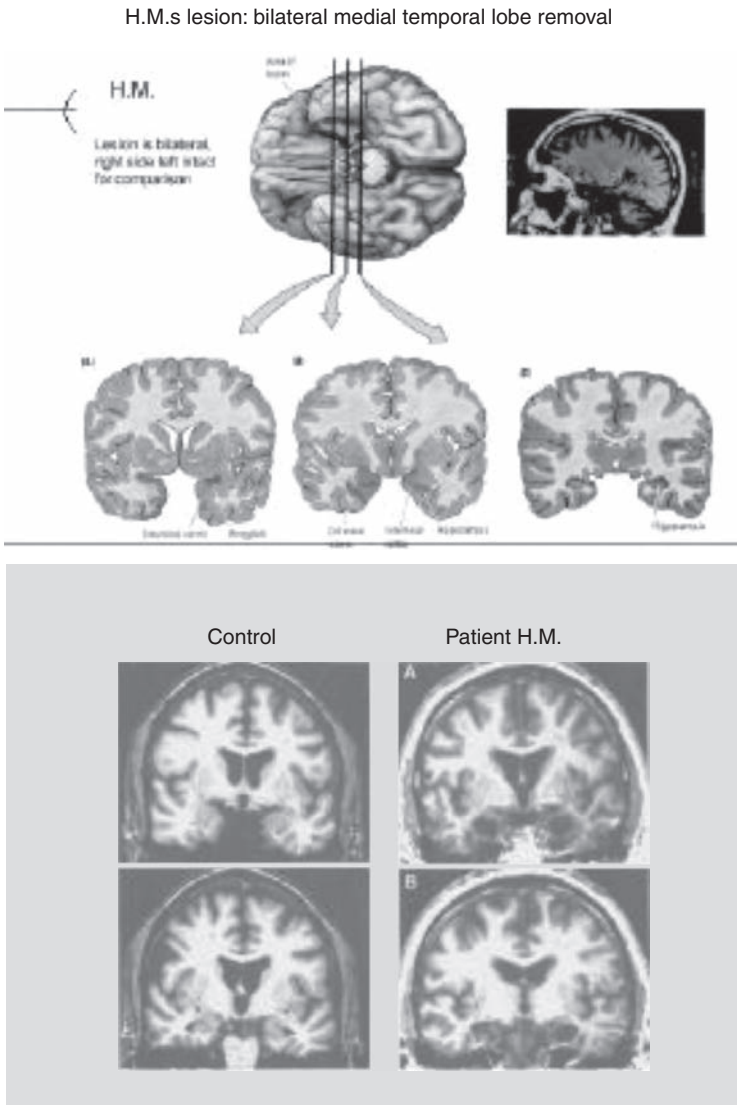

Figure 1 (Above) Representation of area of resection of medial temporal lobe areas in patient H.M. (Below) (A) Magnetic resonance image of coronal section of H.M.'s brain at rostral level showing resection of the amygdala (right) relative to normal control. (B) More-caudal level of H.M.'s brain showing removal of hippocampal formation (right) relative to normal control brain.

In this task, animals must respond to a sample stimulus and then, after a delay, must choose between a novel stimulus and the previously presented familiar one (Figure 3). Rats with damage to the hippocampus can learn the delayed nonmatching to sample task as efficiently as controls with delays of a few seconds but evidence a marked impairment when delays are extended to longer than a minute. These findings support the view that the hippocampus is essential in the ability to form associations between given stimuli and the context in which the stimuli appear.

Perhaps the most widely used method of assessing memory in rodents relies on the use of the Morris Water Maze. In the typical paradigm, the rat (or mouse) is placed into a pool of opaque water that contains an escape platform hidden a few millimeters below the water's surface. Visual cues, such as colored shapes, are placed around the pool in plain sight of the animal. When released, the rat swims around the pool in search of an exit. Several variables can be measured, including the time spent in each quadrant of the pool, the time taken to reach the platform (latency), and the total distance traveled. Rats with damage to the hippocampus are markedly impaired in this task. In contrast, rats with the same lesion can perform successfully on the task when the platform is visible above the water's surface. The data have accumulated to show that removal of the hippocampus in rats results in a marked impairment in tasks requiring spatial navigation and in those that specifically require the animal to remember recent locations it has visited. This effect appears more robust with

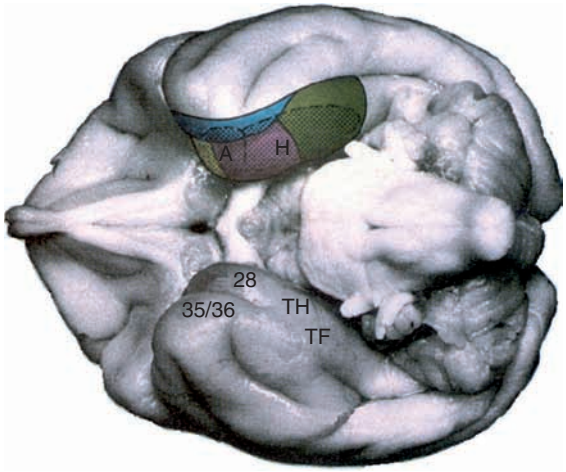

Figure 2 Ventral view of the rhesus monkey brain. Left hemisphere (top of photograph) depicts entorhinal cortex (purple), perirhinal cortex (blue), and parahippocampal cortices (green). Positions of subcortically positioned amygdala (A) and hippocampus (H) are depicted by dotted line. Corresponding Brodmann and Von Bonin and Bailey nomenclature is shown on the right hemisphere (bottom of photograph).

Figure 3 Photograph showing rhesus monkey in the Wisconsin General Testing Apparatus. The monkey is displacing the 'novel' object to obtain a reward (raisin) during the recognition portion of a trial on the delayed nonmatching to sample task.

damage to the hippocampus itself rather than to the surrounding rhinal cortices.

Taking advantage of rodents' exquisitely sensitive olfactory system, Eichenbaum and colleagues have used a novel paired associate odor task to demonstrate that the hippocampus is essential for memory function. In this task, rats first were rewarded for identifying an odor (A) that was paired with another given odor. Several such pairs were readily learned by normal rats as well as those with damage to the hippocampus. This basic stimulus–response learning task could easily be seen as supported by the nondeclarative memory system. On another phase of the task, rats had to infer an association between two odors that were not directly paired but that shared a common associate odor. Whereas control rats learned this task readily, those with hippocampal lesions evidenced marked impairment. Together with the results of related 'transitive' tasks, the data strongly support the view that the hippocampus is key to mediating the connection and expression of memory in a relational context. Eichenbaum has also shown that the hippocampus is essential in mediating sequential events, a working component for episodic memory. Rats with lesions of the hippocampus retain their capacity to recognize individual odors but are severely impaired in their ability to remember the sequential order of odors presented to them.

Nonhuman Primate Models of Memory Function: Medial Temporal Lobe

While evidence from human cases strongly implicated the hippocampal formation in subserving normal memory function, it was unclear whether specific components of this structure had selective roles in memory function. Indeed, the hippocampus represents the endpoint of the transition from a six-layer cortex of the adjacent parahippocampal gyrus to a three-layer cortex in the hippocampus itself. Two of these regions, the perirhinal and parahippocampal cortices, are the recipients of input from secondary sensory association areas, including visual, somatosensory, and auditory regions, that, in turn, project to the entorhinal cortex, which is the major cortical afferent to the hippocampus. The perirhinal, entorhinal, and parahippocampal regions were damaged in patient H.M. (Figure 1), and these regions had also been removed in the surgical approach to the hippocampal formation in the early stages of nonhuman primate studies. As a result, it was difficult to distinguish between deficits traditionally attributed to damage to the hippocampus proper and those that arose from damage to the rhinal and adjacent cortices.

Several studies over the past 10 years have been aimed at fractionating the specific contributions of the hippocampus and rhinal cortices. Findings regarding the effect of damage to the hippocampus in nonhuman primates have been equivocal, particularly regarding the most commonly used task in assessing memory, delayed nonmatching to sample (DNMS). Results from studies of the effects on DNMS performance of damage to the hippocampus of nonhuman primates range from virtually no effect to marked impairment, particularly when delays greater than 30 s are used. However, when certain experimental design variables are taken into account, the contribution of the hippocampus to this task is better revealed.

One striking methodological difference in these studies of memory in nonhuman primates is the point in the experimental paradigm at which the tasks are initially learned. In studies in which monkeys with hippocampal lesions were both trained on the DNMS task and tested on the delay conditions following surgery, a marked impairment obtained. In contrast, in monkeys that were trained on the DNMS task preoperatively and tested postoperatively for retention of performance, little or no impairment was found. This difference in procedure may play an important role in the results because it very likely affects how and where the initial memory is mediated.

Another task used to assess memory, and one that has been widely used with human participants, is the delayed recognition span test (DRST). The DRST is a short-term memory test that requires the participant to identify, trial by trial, the new stimulus within an increasing array of serially presented stimuli. The task is administered using different classes of stimulus material in order to help characterize recognition memory deficits across several stimulus domains. For the spatial condition, identical discs are used as stimuli. A disc is placed over one of 18 wells, which is baited. A screen is raised, and the monkey is allowed to displace the disc to obtain the reward. The screen is then lowered, and a second disc is placed on the board over a baited well while the first disc is returned to its original position over the now unbaited well. After 10 s, the screen is raised, and the monkey is required to displace the new disc in order to obtain the reward. Each successive correct response is followed by the addition of a new disc until the monkey makes an error (i.e., chooses one of the previously chosen discs). With the occurrence of the first error, the trial is terminated, and the number of discs on the test tray minus 1 (i.e., the number of correct responses) constitutes the recognition span score for that trial. The object condition is administered in a similar fashion. Within each trial, the position of the previously correct stimuli is changed in a pseudorandom fashion so that the animal has to identify the new stimulus on the basis of visual cues without the use of spatial cues. Ten trials are administered each day for 10 days (see Figure 4).

Monkeys with bilateral damage restricted to the hippocampal formation made with ibotenic acid injection guided by magnetic resonance imaging are, like patients with Alzheimer's disease, markedly impaired in both the spatial and the object conditions of the DRST. They achieve lower spans than controls, whether the spans are unique or have been presented repeatedly over several sessions.

With regard to the rhinal cortices, Murray and others have performed extensive experiments using neurotoxic lesion methods to produce selective lesions to the perirhinal cortex in the monkey. Damage to this region produces impairment in stimulus recognition and in the maintenance of stimulus–stimulus associations for individual objects that, taken together, suggests that the perirhinal cortices are important for the accurate representation of visual stimuli and have profound effects on both perceptual and memory functions.

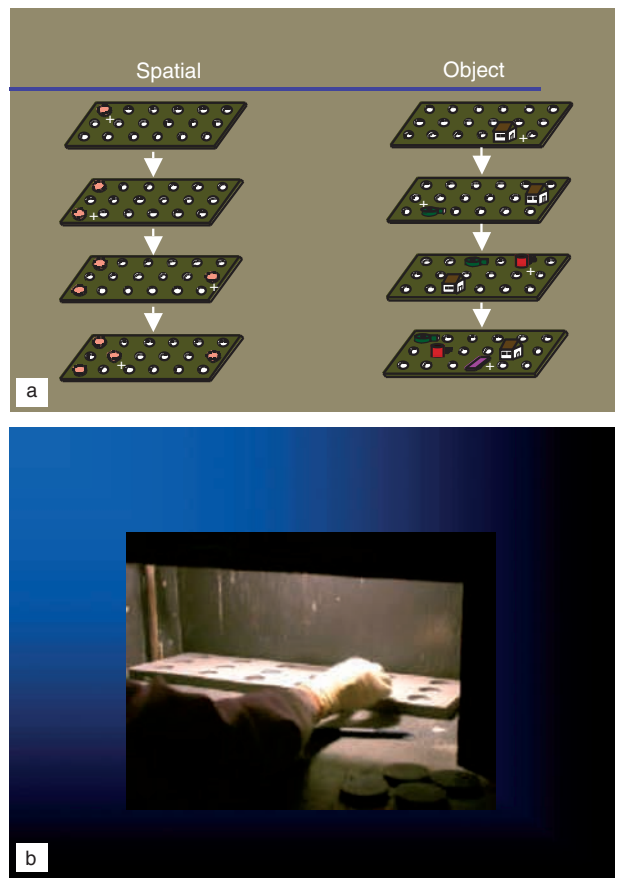

Figure 4 Schematic of the delayed recognition span test (DRST). Left: spatial condition; right: object condition.

Nonhuman Primate Models of Memory Function: Prefrontal Cortices

In the 1930s, researchers began to realize that the frontal lobe was involved in short-term memory processes. Jacobsen, taking the lead from findings of Brown and Schafer, advanced the notion that the frontal lobes may have a role in cognitive function. Jacobsen found that monkeys with damage to the prefrontal cortex were markedly impaired on a delayed response task. The findings were received as the first convincing evidence that the frontal lobes played a role in short-term memory function. Over the past 70 years, evidence for a role of the prefrontal cortices in memory function has been limited. It has become evident form the work of Goldman-Rakic that a small region of the prefrontal cortex plays a key role in working memory and that this area may be modality specific to spatial organization. It has been unclear whether damage to a specific region of the prefrontal cortex (e.g., dorsolateral region (area 9, area 46), ventrolateral region, or ventromedial region) has a specific role in memory or is related more to alteration of response control or some other aspect of so-called executive function.

More-recent evidence has shown that combined damage to areas 9 and 46 of the dorsolateral prefrontal cortex in the monkey results in an impairment on postoperative acquisition and delay conditions of the DNMS task. This deficit is quite severe, as monkeys with dorsolateral prefrontal cortex lesions take more trials to learn the task than do those with lesions of the hippocampus by a factor of almost 2 and perform as badly as monkeys of advanced age.

As is the case with findings from the DNMS task in nonhuman primates following lesions to the hippocampus, the findings following damage to the dorsolateral prefrontal cortex are not consistent. While it may be the case that combined, but not separate, lesions of dorsal prefrontal cortex that include both areas 9 and 46 are necessary to produce a measurable impairment on this task, it is equally plausible that preoperative versus postoperative acquisition of the task may represent the critical difference in findings.

Medical Temporal and Prefrontal Interactions?

While the exact temporal and anatomical sequence of memory processing remains to be elucidated, it is plausible that brain regions associated with visual recognition memory function (such as the hippocampal complex) and other medial temporal lobe structures, along with various parts of the prefrontal cortex, are differentially active in the course of acquisition and retention of qualitatively different cognitive tasks. Further, it is possible that declarative memory function may represent a functionally linked sequential network of mnemonic processing involving both medial temporal and prefrontal areas. According to this notion, one might speculate that during the initial acquisition of the nonmatching principle, both the hippocampus and the dorsal prefrontal cortex are active in processing and storing rules of the nonmatching paradigm. Hence, one would predict that damage to either structure alone would produce impairment in a postoperative acquisition paradigm, and both this report and previous work are consistent with this notion. Accordingly, one might predict that damage sustained by either structure alone following initial acquisition of tasks which employ the nonmatching rule would result in unimpaired or mildly impaired performance on the tasks, and indeed published reports of studies of such a postoperative retention paradigm are consistent with this. Moreover, one would predict that only combined damage to both the medial temporal lobe and the prefrontal cortex would produce significant impairment in such a postoperative retention paradigm, but this has yet to be tested. While these questions have yet to be tested directly, the notion of a complex interaction between prefrontal and medial temporal lobe systems in the acquisition and retention of recognition memory tasks can be tested empirically by making lesions of both the dorsal prefrontal cortex and medial temporal lobe concurrently in a postoperative retention paradigm.

Conclusion

While much progress has been made, greater effort must be made to bring tasks of memory function in animals into parallel with those in humans. It is also here that the great value of animal models can be brought to bear. One can more easily parcellate the anatomic and functional components of memory in animal models with carefully constructed experimental designs and then turn to mapping these onto studies of human memory.

See also: Aging and Memory in Animals; Amnesia: Declarative and Nondeclarative Memory; Episodic Memory: Assessment in Animals; Executive Function and Higher-Order Cognition: Assessment in Animals; Functional Amnesia; Spatial Memory: Assessment in Animals.

Further Reading

Beason-Held LL, Rosene DL, Killiany RJ, and Moss MB (1999) Hippocampal formation lesions produce memory impairment in the rhesus monkey. *Hippocampus* 9(5): 562–574.

Clark RE, West AN, Zola SM, and Squire LR (2001) Rats with lesions of the hippocampus are impaired on the delayed nonmatching-to-sample task. *Hippocampus* 11: 176–186.

Cohen NJ and Squire LR (1980) Preserved learning and retention of pattern analyzing skills in amnesics: Dissociation of knowing how and knowing that. *Science* 210: 207–210.

Eichenbaum H (2004) Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron* 44: 109–120.

Eichenbaum H, Otto T, and Cohen NJ (1994) Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences* 17: 449–472.

Goldman PS and Rosvold HE (1970) Localization of function within the dorsolateral prefrontal cortex of the rhesus monkey. *Experimental Neurology* 27: 291–304.

Mahut H, Zola-Morgan S, and Moss MB (1982) Hippocampal resections impair associative learning and recognition memory in the monkey. *Journal of Neuroscience* 2: 1214–1229.

Mishkin M (1978) Memory in monkeys severely impaired by combined but not separate removal of the amygdala and hippocampus. *Nature* 273: 297–298.

Murray EA (1990) Representational memory in nonhuman primates. In: Kesner RP and Olton DS (eds.) *Neurobiology of Comparative Cognition*, pp. 127–155. Hillsdale, NJ: Erlbaum.

Squire LR (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* 99: 195–231.

Zola-Morgan S and Squire LR (1986) Memory impairment in monkeys following lesions limited to the hippocampus. *Behavioral Neuroscience* 100(2): 155–160.

Animal Models of Huntington's Disease

J Alberch, E Pérez-Navarro, and J M Canals,
University of Barcelona, Barcelona, Spain

© 2009 Elsevier Ltd. All rights reserved.

The generation of animal models is necessary to understand the pathophysiology of neurodegenerative disorders, which will allow the development of new therapeutic strategies. Therefore, it is important to know when and what triggers the signals that induce the neurodegenerative process. The ideal animal model would have reproducible and well-defined behavioral abnormalities and neuropathological features, such as selective loss or dysfunction of specific neuronal populations, all of which must be the closest match to the human pathology. However, neurodegenerative disorders are chronic with different severity and onset, suggesting that there are different factors that modulate to activate the pathogenic mechanisms leading to selective neuronal dysfunction or death. Therefore, different animal models are generated to reproduce different stages of the disease. The generation of animals with a rapid onset and progression will help to determine the earliest molecular and cellular changes associated with the disease, although they can be too aggressive to test new treatments. However, animals with a late onset and milder disease can provide greater specificity and are potentially more useful to test neuroprotective approaches.

Huntington's disease (HD) is a devastating neurodegenerative disorder characterized by irrepressible abnormal movements, cognitive deterioration, and psychiatric disturbances. There is also striking specificity of neuronal loss localized in the striatum (caudate nucleus and putamen), which is the main coordinator of motor activity. The most sensitive cell populations are the striatal medium-sized spiny neurons that project to the globus pallidus and substantia nigra. Moreover, in more advanced stages of the disease, there is also a loss of cortical cells, affecting predominantly the large pyramidal neurons in layers III, V, and VI. Together with the anatomical and neurochemical substrate of HD, the etiology of the neurodegenerative process is well known. The disease is caused by an expansion of a CAG repeat region in the gene encoding huntingtin (IT15) and it is transmitted in an autosomal dominant manner. This mutation results in an increased stretch of glutamines in the N-terminal portion of the protein, which is widely expressed in the brain and peripheral tissues. Asymptomatic individuals have 35 or fewer CAG repeats, with longer expansions causing the illness. There is an inverse relationship between CAG repeat number and the age of onset of symptoms. Longer expansions (>55 repeats) usually give rise to a juvenile form of the disease. However, for medium-length expansions, the age of onset is unpredictable for any individual, which strongly suggests the existence of modifier genes or others factors that influence the age of onset of HD. Among these factors, excitotoxicity, mitochondrial dysfunction, neurotrophic factors, or transcriptional dysregulation have been proposed to modulate the neurodegenerative process.

The mutation has been known for over a decade, but there is still no effective treatment for HD because the pathogenic mechanism is not well understood. Although mutant huntingtin is expressed in all body tissues, the mutation preferentially induces the degeneration of specific neuronal populations in the striatum and cortex. Thus, the development of animal models for HD is required to study the pathogenic mechanisms and to test new treatments.

The first animal models for HD were achieved in the 1980s by either excitotoxic lesioning or metabolic impairment. The intrastriatal injection of glutamate receptor agonists, such as quinolinic acid, induces in rats a selective degeneration of striatal projection neurons, whereas interneurons are relatively spared. Similar results are observed after systemic administration of the mitochondrial toxin 3-nitropropionic acid. These models have been very useful because they produce a similar neuropathological profile as that observed in HD. However, they do not provide any data about the participation of mutant huntingtin in the degenerative process.

Other models have been developed in *Drosophila* and *Caenorhabditis elegans*. It has been shown that flies and worms expressing mutant huntingtin or polyQ peptides alone present a progressive degeneration. These models can be a powerful approach for genetic screens to identify genes that alleviate or modify the disease. However, the mutant protein is not expressed in neuronal types that degenerate in human HD. Hence, cell-specific effects are missing in these models.

However, breakthrough in HD research has been the development of transgenic mice expressing expanded polyglutamine repeats in the huntingtin gene. The first transgenic mice were developed in 1996. These models have been followed by many new HD transgenic and knockin mice that differ with regard to the type of mutation expressed, portion of the protein included

in the transgene, promoter employed, and expression levels of mutant protein and background strain, making each of them unique and related to different degrees of the human pathology. These mouse models can fit in three broad categories: (1) mice that express exon-1 fragments of the human huntingtin gene containing polyglutamine expansions; (2) mice that express the full-length human HD gene; and (3) mice with pathogenic CAG repeats inserted into the endogenous CAG expansion (knockin mice) (Figure 1).

N-Terminal Exon-1 Transgenic Mouse Models

R6/1 and R6/2 transgenic mice were the first mouse models developed to study HD. They both express exon 1 of the human HD gene with 115 and 150 CAG repeats, respectively. The transgene expression in these mice is driven by the human huntingtin promoter. The resulting levels of transgene expression are around 31% and 75% of the endogenous huntingtin

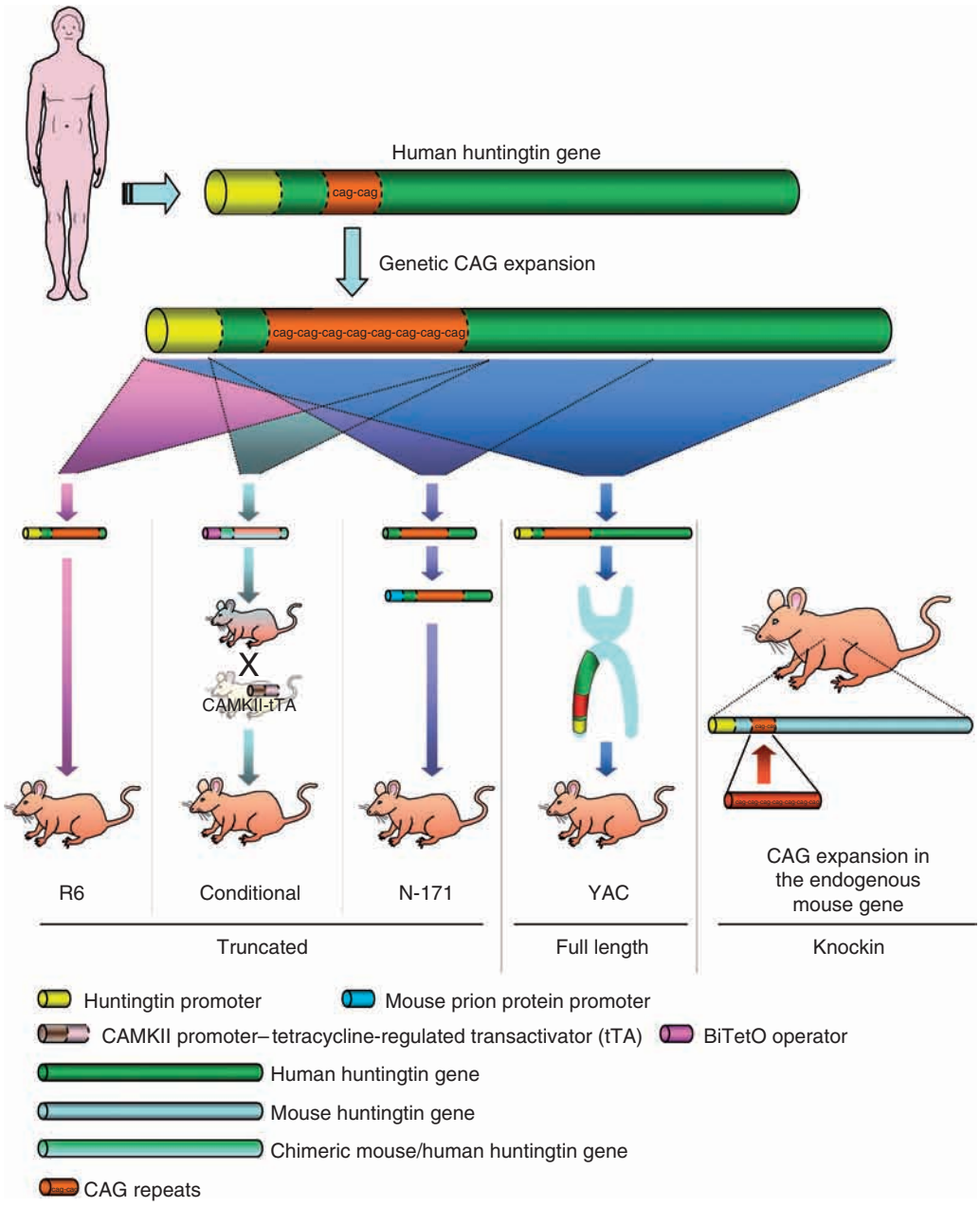

Figure 1 Schematic representation of the human huntingtin gene and constructs used for the generation of mouse models of Huntington's disease. The CAG repeats are localized in the exon 1 of the huntingtin gene, and the normal range is between 19 and 35. In transgenic mouse models the huntingtin gene (A fragment: R6, Conditional, N-171 or full length: yeast artificial chromosome (YAC)) with an expanded CAG repeat is randomly inserted into the genome. Knockin mouse models are obtained by gene targeting of the endogenous mouse huntingtin gene.

in the R6/1 and R6/2, respectively. In accordance with human pathology, R6 mice with the longest CAG repeats (R6/2) have an earlier onset and more severe symptoms like a juvenile form of HD. R6/2 mice are severely impaired by 8–12 weeks of age, whereas R6/1 mice are around 15–20 weeks. A typical sign in these models is the abnormal paw-clasping response. When suspended by the tail, normal mice spread their four limbs, whereas R6 mice clasp their hind- and forelimbs tightly against their thorax and abdomen (Figure 2). Although the pathophysiology of this behavior is not understood, the paw-clasping test is often used in studies examining novel treatments. Other changes in motor function are stereotypical hindlimb grooming, changes in gait patterns, and the emergence of some involuntary movements. As a result, their motor coordination progressively deteriorates, which can be detected as a reduction in the time they can stay on a rotating rod (rotarod).

Neuropathological analysis shows that brain and striatal volume are markedly reduced in R6 mice. However, the significant reduction in brain volume is the result of atrophy of individual neurons with a massive decreased neuropil, because neuronal death is minimal in these mice. Comparison of the time courses of cell death and behavioral anomalies shows that changes in motor activity precede any evidence of neuronal death in these mice, by a matter of several weeks. However, cell death can be observed if the neurotrophin brain-derived neurotrophic factor (BDNF) is reduced in R6/1 mice. The generation and characterization of a double-mutant mouse obtained by crossing R6/1 mice and BDNF heterozygote mice demonstrate that the levels of endogenous BDNF together with mutant huntingtin modulate the onset and severity of motor dysfunction and the degeneration of striatopallidal neurons. Thus, the different levels of endogenous BDNF in various mouse models of HD may also regulate the severity of the phenotype.

The most prominent morphological feature of HD, in addition to the reduction in brain volume, is the development of neuronal intranuclear and intracytoplasmatic inclusions. These are aggregations formed by polyglutamine repeats in the N-terminal huntingtin fragment. They also contain other components such as ubiquitin, and are located within neuronal cell bodies, the nucleus, and their arbors. These inclusions were first observed in animal models with mutant huntingtin and later confirmed in postmortem brains of patients with HD. In contrast to the lack of cell death in early stages in R6 mice, the presence of huntingtin- and ubiquitin-positive aggregates in R6/2 mice is observed as early as postnatal day 1 in the striatum and other brain areas, with the greatest number at 90 days. Thus, there is a close correlation between the degree of motor impairment and the number of striatal neurons exhibiting intranuclear inclusions of mutant huntingtin in this model. However, the function of these aggregates is under discussion, because there is evidence suggesting that they can be either toxic or neuroprotective.

As observed in HD patients, R6 mice also show cognitive deficits in advance of classical motor symptoms. R6/2 mice between 3.5 and 8 weeks of age

Figure 2 Clasping phenotype of R6 mice. Photomicrographs showing 30-week-old wild-type and R6 mice during the tail suspension test. R6 mice display a characteristic full body clasp.

display progressive learning impairment on cognitive tasks sensitive to frontostriatal and hippocampal function, as shown by the Morris water maze, two-choice swim tank, and T-maze tests.

Although the R6 mice are the most extensively studied, other transgenic mouse lines expressing the truncated huntingtin gene with expanded CAG repeats have been generated. N-171 mice express a cDNA encoding an N-terminal fragment (171 amino acids) of human huntingtin with 82, 44, or 18 glutamines. The expression of the transgene is directed by the mouse prion protein promoter, which drives the expression of foreign genes in virtually every neuron of the central nervous system. Mice expressing relatively low steady-state levels of N-171 huntingtin with 82 glutamine repeats (N-171-82Q) develop behavioral abnormalities, including loss of coordination, tremors, hypokinesis, hind limb clasping, and abnormal gait by 3 or 4 months of age, before dying prematurely (4–6 months). In mice exhibiting these abnormalities, diffuse nuclear labeling to the N-terminus of huntingtin, intranuclear inclusions, and neuritic aggregates were found in cortex, striatum, hippocampus, cerebellum, and amygdala as early as 4 weeks of age.

Other lines have also been generated expressing the N-terminal one-third of huntingtin with expanded (HD46, HD100) glutamine repeats under the regulation of the rat neuron-specific enolase promoter. These HD mice exhibit motor deficits at 3 months of age with an exacerbation throughout their lifespan. Neuropathological features, such as accumulation of huntingtin, dysmorphic dendrites, and atrophy in the striatum and cortex, precede the onset of behavioral deficits.

An interesting model would be one in which the mutated gene could be switched off or on. Thus, a conditional model (Tet/HD94) has been developed by the expression of a chimeric mouse/human exon 1 of the HD gene with 94 CAG repeats, under the control of a BiTetO operator that can be switched off after adding tetracycline analogs to the drinking water. The tetracycline-regulated transactivator (tTA), which constitutively activates the BiTetO in the absence of tetracycline analogs, is under the control of the calcium/calmodulin kinase IIα (CAMKII) promoter. Hence, mutant huntingtin expression is restricted to neurons localized in specific brain areas such as the striatum, septum, cortex, and hippocampus. These animals have a slow progression and the average life span is similar to that of wild-type mice (2 years). Until 4 weeks of age, conditional transgenic Tet/HD94 mice are indistinguishable from control littermates. At 4 weeks, some mice begin to clasp, and by 8 weeks all mice are clasping. In addition to this abnormal behavior, at 20 weeks Tet/HD94 mice begin to show a mild tremor, and at 36 weeks they are hypoactive. Neuropathological analysis shows that the brain and striatal volume at 18 weeks of age is smaller than in wild-type mice. This is accompanied by a gross enlargement of the ventricles, although cell death is not observed. The long survival of these animals has allowed the study of aged animals, revealing that at 17 months of age neuronal death is present in the striatum. Interestingly, shutting off expression of the mutant transgene produces significant changes in the neuropathological phenotype of these Tet/HD94 mice. When the gene is turned off, mutant huntingtin is rapidly cleared from the mouse brain and the motor symptoms are reversed. In aged animals, gene silencing can still be beneficial and full motor recovery is possible despite neuronal loss. This model provides evidence that treatments blocking mutant huntingtin expression in the early or even advanced stages of HD might enable the reversal of the clinical symptoms (Table 1).

Full-Length Transgenic Mouse Models

Several mouse models have been generated using a full-length huntingtin gene as the transgene. Mice that express a full-length huntingtin cDNA clone with either 48 (HD48) or 89 (HD89) repeats driven by the cytomegalovirus (CMV) promoter show a progressive motor phenotype. Feet clasping following suspension by their tails is observed in all HD48 and HD89 mice at 8 weeks of age. At 24 weeks of age, decreased motor activity is observed which progresses to locomotor deterioration and akinesia. Importantly, in the hypokinetic stage, neuronal loss, gliosis, and degeneration also occur. Although the number of neurons with nuclear inclusions varies in different brain regions, less than 1% of striatal neurons of HD48 and HD89 mice show intranuclear inclusions. However, the presence of polyglutamine aggregation is seen as early as 12 weeks of age.

Similar features have been observed in other mouse models obtained by the expression of yeast artificial chromosome (YAC) that spans the entire human HD gene with 46 (YAC46), 72 (YAC72), or 128 (YAC128) CAG repeats. YAC46 and YAC72 mice show electrophysiological abnormalities prior to the presence of nuclear inclusions or neurodegeneration indicating a cytoplasmic dysfunction. By 12 months of age, YAC72 mice have a selective degeneration of medium-sized spiny neurons in the striatum. This is associated with the translocation of N-terminal huntingtin fragments to the nucleus. Neurodegeneration can be present in the absence of macro- or micro-aggregates, showing that aggregates are not essential

Table 1 Mouse models of Huntington's disease

	Model design			*Abnormal motor behavior (onset)*	*Neuropathology*		*Life span*
	Promoter	*Gene size*	*CAG repeats*		*Aggregates/inclusions*	*Cell loss*	
Transgenic models							
R6 Mangiarini et al. (1996)	IT15	Exon 1	115 (R6/1)	Clasping (15–21 w) Decline in rotarod performance (13–20 w)	Observed in cortex and striatum from 8 w	Dark cell degeneration in cortex and striatum Reduction of medium-sized spiny neurons in the striatum	32–40 w
			144 (R6/2)	Clasping (8 w) Locomotor hyperactivity (3 w) Hypoactivity (8 w) Decline in rotarod performance (8–12 w)	Present in cortex and striatum from 3–4 w	Dark cell degeneration in cortex and striatum Reduction of medium-sized spiny neurons in the striatum	13–16 w
HD Laforet et al. (2001)	Rat neuron specific enolase	3-kb N-terminal fragment	48	Clasping (3–6 m) Impaired performance in rotarod (3–6 m)	Few intranuclear inclusions in cortex and striatum	Reduced number of neurons in the striatum in some animals	>12 m
			100	Clasping (3–6 m) Impaired performance in rotarod (3–6 m)	Intranuclear inclusions in cortex and striatum	Reduced number of neurons in the striatum in some animals	
N-171 Schilling et al. (1999)	Mouse prion protein	First 171 amino acids from exon 1	44	Normal up to 2 years of age	No	No	Normal
			82	Clasping (not reported) Impaired performance in rotarod (3 m) Abnormal gait (3–4 m) Hypoactivity (5 m)	Cytoplasmatic aggregates at 6 m Nuclear inclusions in the cortex, striatum, hippocampus, cerebellar granular cells, and amygdala at the end stage	Not severe loss of neurons in cortex, striatum, hippocampus, and cerebellum	10–50 w
YAC46 YAC72 Hodgson et al. (1999)	IT15	Full length	46	Normal up to 20 m	No	No	>12 m
			72	Clasping (nondescribed) Hyperactivity (7 m) Circling (9 m)	Cytoplasmic microaggregates in striatal neurons Nuclear inclusions: high number of neurons in the striatum, olfactory tubercle, and nucleus accumbens; small number of neurons in the septum and granule cell layer of the cerebellum	Apoptotic cells in the striatum (12 m)	>12 m
YAC128 Slow et al. (2003)			128	Hyperkinesia (3 m) Motor deficit on the rotarod (6 m) Hypokinesia (12 m)	Inclusions in striatal and cortical neurons	Reduced number of striatal neurons	>12 m
HD94-tet off Yamamoto et al. (2000) Diaz-Hernandez et al. (2005)	CAMKIIα-tTA	Exon 1	94	Clasping (4 w) Mild tremor (20 w) Hypoactivity (36 w) Progressive motor decline (10 m)	Nuclear inclusions in cortex and striatum Extranuclear aggregates in cortex and with higher occurrence in the striatum	Dark cell degenerating neurons in the striatum (14 m) Decreased number of striatal neurons (17 m)	Normal

Continued

Table 1 Continued

	Model design			*Abnormal motor behavior (onset)*	*Neuropathology*		*Life span*
	Promoter	*Gene size*	*CAG repeats*		*Aggregates/inclusions*	*Cell loss*	
Knockin models							
HdhQ92 Wheeler et al. (2000)	Hdh		90	No clasping (up to 17 m)	Nuclear inclusions in striatum, olfactory tubercle, and cortex In older mice, nuclear aggregates are also observed in septum, olfactory bulb, nucleus accumbens, cerebellar granule cell layer, and hippocampus	No	Normal
Hdh94 Menalled et al. (2002a)	Hdh		94	No clasping (up to 24 m) Increased rearing (2 m) Decreased locomotion (4 m)	Nuclear microaggregates in the striatum Nuclear inclusions in the striatum (old mice) Other brain areas nonexamined	No	Normal
HdhQ111 Wheeler et al. (2000)	Hdh		109	Subtle gait deficits (24 m)	Nuclear inclusions in striatum, olfactory tubercle, and cortex In older mice nuclear aggregates are also observed in septum, olfactory bulb, nucleus accumbens, cerebellar granule cell layer, and hippocampus	No	Normal
HdhQ140 Menalled et al. (2003)	Hdh		140	Hyperactivity (1 m) Decreased locomotor activity (4 m) Gait anomalies (12 m)	Nuclear microaggregates in striatum, olfactory tubercle, cortex, and olfactory bulb Nuclear inclusions in striatum, olfactory tubercle, cerebellum, and layer II of piriform cortex Neuropil aggregates in striatum, globus pallidus, substantia nigra pars reticulata, cortex, cerebellum, olfactory tubercle, and olfactory bulb	Not reported	Normal
Hdh150 Lin et al. (2001)	Hdh		150	Gait and rotarod deficits (15–40 w) Clasping (15–40 w) Hypokinesia (40 w)	Nuclear inclusions in the striatum and nucleus accumbens; less frequent in cortex, hippocampus, and cerebellum; sparse in the thalamus, hypothalamus, and brain stem	No loss of striatal neurons	Normal

w, weeks; m, months.

to the initiation of neuronal death. YAC128 mice reveal a hyperkinetic phenotype first manifested at 3 months of age, followed by a progressive motor deficit on the rotarod at 6 months with eventual progression to hypokinesis by 12 months of age. These behavioral changes are followed by striatal atrophy clearly evident by 9 months, cortical atrophy at 12 months, and a progressive loss of striatal neurons accompanied by a decrease in their soma area. In contrast to N-terminal exon-1 transgenic models, the motor deficit in the YAC128 mice is highly correlated with neuronal loss. These mice demonstrate that initial neuronal cytoplasmic toxicity is followed by cleavage of huntingtin, nuclear translocation of huntingtin N-terminal fragments, and selective neurodegeneration.

Mild cognitive deficits have also been described in YAC128 mice that precede motor onset and progressively worsen with age. Rotarod testing reveals a motor learning deficit at 2 months, which progresses until 12 months of age, when YAC128 mice are unable to learn the rotarod task. These mice also show deficits in procedural learning, memory, sensorimotor gating, and strategy shifting.

Knockin Models

Knockin mouse models have been generated by homologous recombination of CAG repeats in the endogenous mouse huntingtin gene. Therefore, knockin mice carry the mutation in its appropriate genomic and protein context, making them more faithful genetic models of the human condition than other transgenic lines, which have the transgene inserted randomly. Several knockin mice have been developed with different CAG expansions and they exhibit similar neuropathological phenotype, although at different ages. All these mice have a late onset and slow progression. However, there are differences in the magnitude of motor abnormalities in knockin and transgenic mice, but both lines show a shift from hyper- to hypoactivity. Thus, for example, knockin mice with 94 repeats display a biphasic motor behavior characterized by increased motor activity at 2 months, followed by hypoactivity at 4 months of age. Gait abnormalities are observed at 24 months. HdhQ111 mice do not differ from wild-type animals in paw clasping and rotarod at 15 months of age, but at 24 months motor function assessed by tunnel walks reveals a subtle gait deficit. In animals with longer CAG repeats, similar patterns of motor abnormalities are observed although at early ages. Gait abnormalities are already observed at 12 months of age. None of the knockin mice develop neuronal loss, although reactive gliosis is observed in the striatum. Nuclear staining and aggregates of the mutant huntingtin are only observed in various knockin models when the mice are old. In mice with 94 repeats, intranuclear inclusions are restricted to the striatum, whereas with longer expansions (140 and 150 CAG repeats) the distribution is more widespread.

Conclusion

In the previous years, several models have been developed showing different phenotypes and timing depending on the genetic construct, length of CAG repeats, and background strain. They all produce aspects of the pathology observed in HD patients, such as movement abnormalities, specific atrophy of the striatum and cortex, and the presence of intranuclear and cytoplasmatic aggregates of mutant huntingtin. Hence, the availability of these mouse models has been an important step in research aimed at dissecting the pathogenesis of HD. Each model can be useful for different types of studies. Knockin models accurately replicate the underlying genetic defect of HD, but they have a slow progression with no robust motor deficits, and brain atrophy is only observed in aged animals. The transgenic mice expressing full-length mutant huntingtin have a less aggressive course of the disease than mice that express a truncated form. The varying degrees of striatal atrophy, brain weight loss, and motor deficit in these animal models allow the study of different stages and intensities (juvenile or adult) of the pathology. Therefore, the different animal models of HD are powerful tools to understand the pathogenic mechanisms and for the assessment of neuroprotective and other therapeutic interventions.

See also: Axonal Transport and Huntington's Disease; Cell Replacement Therapy for Huntington's Disease; Huntington's Disease; Huntington's Disease: Neurodegeneration; Transgenic Models of Neurodegenerative Disease; Triplicate Repeats: Huntington's disease.

Further Reading

Beal MF and Ferrante RJ (2004) Experimental therapeutics in transgenic mouse models of Huntington's disease. *Nature Reviews Neuroscience* 5: 373–384.

Canals JM, Pineda JR, Torres-Peraza JF, et al. (2004) Brain-derived neurotrophic factor regulates the onset and severity of motor dysfunction associated with enkephalinergic neuronal degeneration in Huntington's disease. *Journal of Neuroscience* 24: 7727–7739.

Diaz-Hernandez M, Torres-Peraza J, Salvatori-Abarca A, et al. (2005) Full motor recovery despite striatal neuron loss and formation of irreversible amyloid-like inclusions in a conditional

mouse model of Huntington's disease. *Journal of Neuroscience* 25: 9773–9781.

Hodgson JG, Agopyan N, Gutekunst CA, et al. (1999) A YAC mouse model for Huntington's disease with full-length mutant huntingtin, cytoplasmic toxicity, and selective striatal neurodegeneration. *Neuron* 23: 181–192.

Laforet GA, Sapp E, Chase K, et al. (2001) Changes in cortical and striatal neurons predict behavioral and electrophysiological abnormalities in a transgenic murine model of Huntington's disease. *Journal of Neuroscience* 21: 9112–9123.

Lin CH, Tallaksen-Greene S, Chien WM, et al. (2001) Neurological abnormalities in a knockin mouse model of Huntington's disease. *Human Molecular Genetics* 10: 137–144.

Mangiarini L, Sathasivam K, Seller M, et al. (1996) Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* 87: 493–506.

Menalled LB and Chesselet MF (2002) Mouse models of Huntington's disease. *Trends in Pharmacological Science* 23: 32–39.

Menalled LB, Sison JD, Dragatsis I, Zeitlin S, and Chesselet MF (2003) Time course of early motor and neuropathological anomalies in a knockin mouse model of Huntington's disease with 140 CAG repeats. *Journal of Comparative Neurology* 465: 11–26.

Menalled LB, Sison JD, Wu Y, et al. (2002) Early motor dysfunction and striosomal distribution of huntingtin microaggregates in Huntington's disease knockin mice. *Journal of Neuroscience* 22: 8266–8276.

Rubinsztein DC (2002) Lessons from animal models of Huntington's disease. *Trends in Genetics* 18: 202–209.

Schilling G, Becher MW, Sharp AH, et al. (1999) Intranuclear inclusions and neuritic aggregates in transgenic mice expressing a mutant N-terminal fragment of huntingtin. *Human Molecular Genetics* 8: 397–407.

Slow EJ, van Raamsdonk J, Rogers D, et al. (2003) Selective striatal neuronal loss in a YAC128 mouse model of Huntington disease. *Human Molecular Genetics* 12: 1555–1567.

Wheeler VC, White JK, Gutekunst CA, et al. (2000) Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knockin mice. *Human Molecular Genetics* 9: 503–513.

Yamamoto A, Lucas JJ, and Hen R (2000) Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. *Cell* 101: 57–66.

Relevant Website

http://www.hdfoundation.org – The Hereditary Disease Foundation.

Animal Models of Inherited Retinal Degenerations

J Stone, Australian National University, Canberra, ACT, Australia
J R Heckenlively, University of Michigan, Ann Arbor, MI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The retinal dystrophies are genetic diseases whose mutations target retinal neurons, particularly photoreceptors. The retina has a highly interdependent cellular structure, and diseases that initially affect a particular specialized cell, such as rod or cone photoreceptors, will subsequently damage adjacent cells. The effect of the mutation is readily apparent in higher mammals as they are visually disabled. For mice, however, blindness has not been a selective disadvantage. Wild mice blinded by the *rd1* mutation, for example, have been found all over the world, and the viral insert in rd1 mice is estimated to be more than a million years old.

In many panretinal degenerations, the photoreceptors are the prime target of the mutation, and remodeling of synapses may occur in adjacent cells after the photoreceptors die. Disorders of the retinal pigment epithelium (RPE), the cellular layer adjacent to the outer tips of photoreceptors, which provides essential metabolic support to rods and cones, can also lead to photoreceptor death. Inherited retinal dystrophies (IRDs) have been reported in many mammals, commonly mice, rats, dogs, and cats, and are a clinically important cause of blindness in humans. Genetic mutations have been engineered extensively in mice and, less frequently, in the rat, pig, frog, and fish. The number of mutations identified as causing IRDs is large; the numbers have been documented best in humans (Figure 1), in which thousands of mutations in more than a hundred genes have been identified. The functions of these IRD genes are diverse, and their identification or engineering in animal models has made possible the systematic study of their mutational effects on the physiology of the retina and viability of retinal neurons.

Identification of mutations targeting other retinal cell types is still expanding and has been documented best in humans and mice. Abnormalities of bipolar cells, Müller cells, and retinal astrocytes are all known to cause minor to serious disease. As examples, mutations affecting primarily bipolar cells lead to forms of congenital stationary night blindness, while mutations affecting retinal astrocytes are associated with retinal vasculopathies. Because the large majority of IRDs affect photoreceptors either primarily or at a very early stage, much work in animal models has been dedicated to understanding the function and fragility of photoreceptors, with the aim of identifying key mechanisms that will provide a rationale for therapy.

Organization of the Retina

The mammalian retina has two classes of photoreceptors, rods and cones. The rods function at low-light levels, providing highly sensitive, low-resolution, black-and-white vision. In brighter conditions, such as daylight, the photodetection mechanisms of the rods are saturated, and rods contribute only minimally to visual function. In these brighter conditions, cones are the functional photoreceptor, providing high-resolution color vision. In many primates, including humans, the retina has evolved a specialized macula, or foveal region, where cones are the dominant photoreceptor class and retinal circuitry is specialized to provide very high resolution of detail and color. Together, rods and cones allow vision over a five-log-unit range of light intensities, from near-darkness to bright daylight. Rods and cones have highly organized transmission pathways to the output cells of the retina, the ganglion cells. However, rod and cone pathways interact, beginning with direct electrical synapses between rods and cones.

The photoreceptors of mammalian retina form, counterintuitively, the outer layers of the retina. Further, they are oriented with their photosensitive elements, the outer segments, oriented most externally. This requires light to pass through the inner layers of the retina, which contain neurons processing the electrical signals generated by rods and cones. It has evolved to allow the tips of photoreceptors to be embedded in a bounding epithelium, the RPE. This relationship of rod and cone outer segments to the RPE is critical for the viability and function of photoreceptors. The RPE plays a role in the recycling of the breakdown products of phototransduction and a distinct but equally critical role in phagocytosing membrane segments discarded daily by the outer segments.

The retina forms a delicate, thin (300 μm) neural membrane lining the inner surface of the back of the eyeball. It is protected by the robust outer wall of the eyeball (the sclera) and by specialized ocular media, particularly the vitreous body, which fills much of the cavity of the eyeball. The metabolic needs of the photoreceptors are high and are met by the flow of nutrients from the choroidal circulation, a vascular bed that forms between the RPE and the

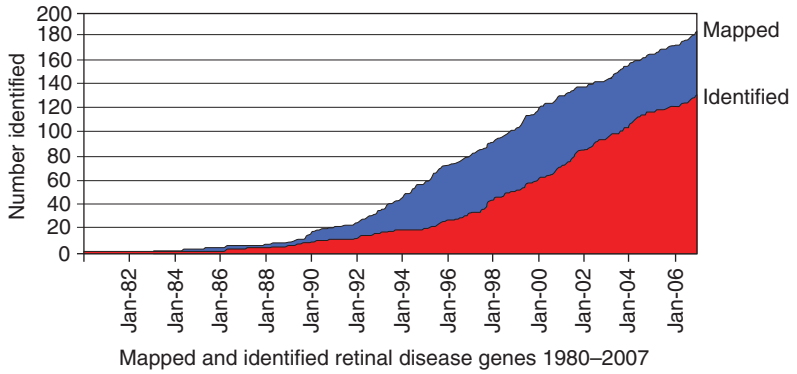

Figure 1 Cumulative graph of the discovery of genes whose mutations cause 'inherited retinal dystrophies'. 'Mapped' means that the chromosomal location of the gene has been defined. 'Identified' means that the specific gene sequence is known. From Daiger SP, Sullivan LS, and Browne SJ (2007) *RetNet: Retinal Information Network*, http://www.sph.uth.tmc.edu/RetNet/ (accessed Jul. 2008).

sclera. The metabolic needs of the inner layers of retina are provided in many species, including the human and the animal species used to model the IRDs, by a retinal circulation spreading through the inner retina from the head of the optic nerve.

Embryologically, the retina and the optic nerve are extensions of the diencephalon of the forebrain and are the only part of the brain to lie outside the cranium. Thus the retina forms a thin layer of central nervous tissue, bounded physiologically by the tight junction-linked endothelium of the retina vasculature and RPE. Within this compartment, the environment of retinal neurons is maintained by radially oriented glial cells, the Müller cells, which play a structure role in forming the limiting membranes of the retina and also participate in the recycling of transmitters and maintaining the ionic environment essential for electrical signaling by retinal neurons. At the inner surface of the retina, an additional layer of glial cells, the astrocytes, is found. These are immigrants to the retina from the optic nerve; they spread over the inner retinal surface, forming a template for the formation of the retinal vasculature.

Mutations causing IRDs have now been described affecting all these major elements of retinal structure: the photoreceptors, the other retinal neurons, the RPE, Müller cells, and astrocytes.

Interaction of Genetically Induced Photoreceptor Death with Normal Photoreceptor Death

Study of rodent models has demonstrated how mutation-induced stresses on photoreceptors interact with normal developmental stress to produce the adult form of disease. In neonatal life, for example, a phase of photoreceptor death has been identified in a developmental 'critical period' soon after the eyes open and photoreceptors commence function. During this period, photoreceptors are subject to hypoxia-induced stress, which culls the population to a sustainable level. In several forms of inherited dystrophy (e.g., the Royal College of Surgeons (RCS) rat, the *rd* mouse, the transgenic P23H rat lines), the high rates of photoreceptor death that characterize the degeneration begin during the normal critical period and persist at varying rates into adulthood. That is, the degeneration induced by a range of mutations (disturbance of recycling in the RCS rat, of a phosphodiesterase (PDE) function in the *rd* mouse, of rhodopsin in the P23H rat) exacerbates a normal process of photoreceptor culling.

In the adult, similarly, there is extensive evidence that the normal (wild-type) retina suffers photoreceptor attrition throughout life, in all mammals examined (rodents and primates). This normal attrition has been documented best in humans, in which the death of rods and cones occurs throughout life, at startlingly high rates – as many as 500 photoreceptors a day, or 150 000 per annum. The loss does not normally become important functionally until the later decades of life, when age-related degeneration can be severe. Again, the inherited dystrophies can be understood as an acceleration of a basal, naturally occurring photoreceptor loss. Photoreceptors operate in a stressful environment of high metabolism, powerful membrane currents, active ion pumps, and rapid turnover of the outer segment membrane. Retinal dystrophies occur when a mutation decreases the evolved stability of adult photoreceptors.

Age-related changes are not unique to the retina; age-related deterioration of many aspects of brain function is well documented. In various ways, the mutations that cause the IRDs destabilize the interactions between retinal neurons and accelerate a naturally occurring loss of photoreceptors.

Insights into the Diversity of Genetically Induced Photoreceptor Death

The diverse forms of the IRDs, listed now on several databases, can be classified in several ways: according to their mode of inheritance (autosomal vs. sex-linked, dominant vs. recessive), by age of onset, by which neuron (rod, cone, or both) is the primary target of the mutation, or by whether the mutation affects more than one sense organ, as in Usher's syndrome in humans, which involves deafness as well as vision loss. As the diversity of mutations that accelerate photoreceptor death has become better known, the function of the gene product has become an alternative way of describing the diversity of the IRDs.

One recent classification, for example, identifies four broad classes of gene products affected by IRD mutations: (1) phototransduction proteins (e.g., rhodopsin, cyclic guanosine monophosphate (cGMP) PDE, arrestin), (2) structural proteins (e.g., peripherin, the adhesion molecule RS1, genes interacting with the photoreceptor cilium), (3) proteins involved in metabolism of photoreceptors and the associated RPE (e.g., phosphorylases such as MertK and RPE65), and (4) transcription factors. This list of the functional classes of IRD genes is likely to grow to include molecular chaperones (proteins which facilitate the correct folding of other proteins); genes that function in every cell but have evolved into retina-specific isoforms, a recent example being inosine monophosphate dehydrogenase 1 (IMPDH1); and ubiquitously expressed, non-retina-specific genes, an example being pre-messenger RNA (mRNA) processing genes, whose products are part of the spliceosome (see Figure 2).

Why Are Photoreceptors Selectively Vulnerable to Mutations in Ubiquitously Expressed Genes?

The selective vulnerability of photoreceptors to mutations in ubiquitously expressed genes is one of the scientific challenges of IRDs, and answers to it are being worked out disease by disease. For example, *IMPDH1* is a highly conserved, ubiquitously expressed gene. The gene product is the enzyme IMPDH1, which regulates the rate-limiting step of guanine synthesis. Although the gene is highly conserved and ubiquitously expressed, the clinical consequences of mutations are

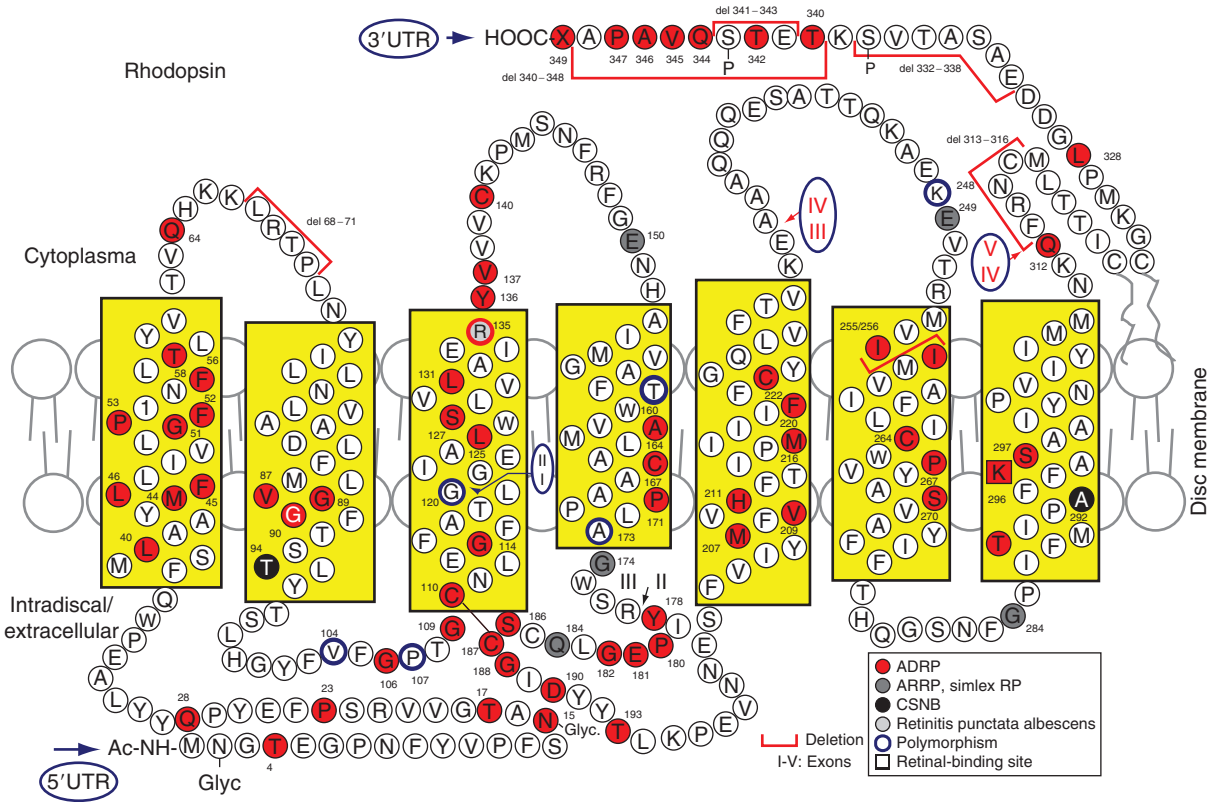

Figure 2 A representation of the rhodopsin protein, embedded in the membrane of the rod outer segment. Sites are shown at which mutations (single-amino-acid substitutions, deletions) have been identified which cause autosomal dominant and recessive forms of retinitis pigmentosa (ADRP, ARRP), congenital stationary night blindness (CSNB), or other conditions, such as retinitis punctata albescens. The diagram emphasizes the large number of mutations that persist into the adult phenotype of a single protein and cause photoreceptor degeneration. del, deletion; UTR, untranslated region. From Preising M (2007) *Retina International Scientific Newsletter*.

retina-specific – the RP10 form of retinitis pigmentosa (RP) and, less commonly, Leber's congenital amaurosis or congenital RP. The selective vulnerability of photoreceptors has been traced to the circumstance that the retina has evolved specific isoforms of the gene, and mutations can be specific to retinal isoforms. The retina also expresses IMPDH1 at high levels, increasing its vulnerability. Diseases related to mutations in nonretinal IMPDH1 have not been reported.

One of the earliest-discovered retinal degeneration mutations in mouse, *rd1*, has a murine viral insert and a second nonsense mutation in exon 7 of the *PDE6B* (β-subunit of cGMP PDE) gene. Again, this enzyme is expressed ubiquitously, but rods and cones have evolved a specific isoform (PDE6), so mutations can be retina specific. The enzyme catalyzes the hydrolysis of cGMP to GMP; because cGMP (but not GMP) gates the ion channels of the outer segment, which conduct the dark current of the photoreceptor, the enzyme's normal action, triggered as part of the phototransduction cascade, is to reduce dark current, creating an electrical signal. The mutation inactivates the enzyme and raises cGMP levels, leaving the channels gated open. This results in elevated intracellular Ca^{2+} levels, which induce apoptotic death, causing the rapid degeneration of rods. *Mertk* is a receptor tyrosine kinase gene expressed in the RPE. Mutations cause a photoreceptor-specific degeneration, first recognized in the naturally occurring RCS rat strain but subsequently identified as a human IRD. Photoreceptors are specifically vulnerable because of their dependence on the RPE for a number of functions, including the phagocytosis of outer segment membrane. This active discarding of membrane and the dependence of photoreceptors on a neighboring epithelium to recycle the ingredients of the photopigment are unique to the photoreceptor.

In another example of retina-specific diseases caused by widely expressed genes, mutations in the pre-mRNA processing factors 31, 8, and 3 cause photoreceptor-specific degenerations known as RP11, RP13, and RP18. Pre-mRNA processing in the spliceosome occurs in every cell, and there is no evidence of a retina-specific isoform of these genes, yet the degenerations are photoreceptor specific.

Thus, the high vulnerability of photoreceptors results in some cases from the evolution of photoreceptor-specific isoforms of the gene and in other cases seems related to photoreceptor-specific parameters, such as the dependence of photoreceptors on the RPE or the high dark current of rods. One common factor is likely the high degree of specialization of photoreceptors, which underlies their high metabolism, the evolution of retina-specific isoforms of some genes, and a multiple dependence on surrounding cells. Contributing to the large number of retinal dystrophies observed is the fact that photoreceptors are not essential for fetal or postnatal development, allowing a large number of retinal dystrophies to be expressed in adults.

Interactions with Retinal Cell Biology

Despite the diverse causes of photoreceptor degeneration, the IRDs show sufficient common symptoms to give some meaning to the clinical name given to many human forms of IRD, RP. The thinning and degeneration of the RP retina result in pigmentary invasion of the RP retina, which becomes apparent on ophthalmic inspection. The vessels of the retina thin, and the retina is intriguingly resistant to hypoxia-driven diseases, such as diabetic retinopathy. Work in rodent models has shown that the retinal vessel thinning, and presumably the resistance to hypoxia, is caused by a chronic rise in choriocapillaris-derived oxygen levels as the photoreceptor population is depleted.

The photoreceptors are the only part of the central nervous system to lack intrinsic vessels, for optical reasons. They are supplied by oxygen and glucose diffusing from the capillaries of the choroidal circulation, the choriocapillaris. The choriocapillaris lies close to the outer retina, separated from the photoreceptors by the RPE and its basement membrane (Bruch's membrane). Critically, however, the choriocapillaris vessels do not lie in the retina, cannot sense the levels of oxygen and other metabolites in the tissue they supply, and therefore cannot adjust their flow to the needs of the photoreceptors. As a consequence, when the photoreceptor population is depleted and its consumption of oxygen falls, oxygen levels in outer retina rise chronically. This hyperoxia causes a slow thinning of retinal vessels and oxidative damage to surviving photoreceptors. It also causes upregulation of genes related to antioxidative damage; resistance to hypoxic disease; upregulation of stress-induced, prosurvival mechanisms; and at the very late stages, abnormal growth of neurons.

In their late stages, then, IRDs are strongly influenced by changes to the cell biology of the retina that create common features to the phenotype. The hyperoxia becomes cumulative, the progressive death of photoreceptors adding to the hyperoxia and reinforcing the oxidative damage.

Rod–Cone Interdependence

The fragility of cones in many IRDs in which the mutation affects a protein (such as rhodopsin) expressed specifically in rods is clinically important and mechanistically puzzling. Good cone vision provides much of the value and pleasure of vision: the ability to read, see faces, and see what others can see

in daylight. Cone vision is typically unstable after rods are depleted, however. Clinically, cone vision can persist for years, even decades, after rod vision fades, but in most cases it is eventually lost. Animal models such as the *rd1* mouse or the rhodopsin-mutant P23H-3 rat confirm that cones are destabilized by rod loss.

The study of IRDs in animals has given insight into the mechanisms of this interdependence, providing evidence that rods and cones rely on diffusible trophic factors for their viability. In mice, evidence has been developed that rods release a 'rod-derived cone viability factory,' identified as a polypeptide consisting of 109 amino acids and essential for the survival of cones. As the rod population is depleted or damaged, cones are starved of this factor and die progressively. A second mechanism, known as the oxygen toxicity hypothesis, came from the study of the effects of photoreceptor depletion in the RCS rat and P23H-3 rhodopsin-mutant transgenic rat and evidence that direct hyperoxia (inhaling oxygen-enriched air) is directly toxic to photoreceptors. Because the supply of oxygen to photoreceptors from the choroidal circulation is unregulated, oxygen accumulates in the photoreceptor layers of retina as a retinal dystrophy progresses. This rise is chronic and results in oxidative damage to the surviving photoreceptors. A third idea is that as the rods disappear, the supporting matrix they provide to the photoreceptor layer is lost, and the cones are distorted or crushed.

Evidence has emerged from mouse IRDs of a converse dependence. In many mouse 'cone' dystrophies, in which cones are affected first, the rods typically show a panretinal degeneration 6–12 months after the cones are lost. There appears to be a co-dependence of rods and cones, neither population being stable without the other.

Pathways to Therapy

The growth of knowledge about the molecular genetics of IRDs has been rapid in the past 15 years, and work on animal IRDs has expanded understanding of the cell biology of the degenerating retina, the effects of oxidative stress on the retina, and the retina's ability to protect its cells against oxidative stress. These bodies of knowledge are being brought together to test numerous avenues to therapy. Some of the major approaches are described in the next sections.

Gene replacement therapy Gene replacement therapy involves delivering a functioning gene to supplement or replace the mutant gene in cells surviving in the degenerating retina. Research in animal models indicates that it is likely to be most useful in IRDs in which mutations such as a deletion or stop codon leave the targeted cell without the needed gene product. Placing small genes in vectors such as lentiviruses or adeno-associated viruses has become routine, and delivery to the retina is usually achieved by subretinal or occasional vitreal injections. A number of successful treatments have been pioneered in animal models of IRDs, including dogs as well as rats and mice. The degeneration of the retina has been slowed or stopped, the function of the retina has been partially restored, and the improvements have been shown to be long-lasting. The use of gene replacement therapy is theoretically more difficult in missense mutations in which the gene product is abnormal, since the abnormal product may continue to be produced and to disrupt the cell even though a correct product is present. In this circumstance, a two-step procedure has been proposed, the first step being to knock out the mutant gene, and the second, to add the normal gene. Because of the diversity of mutations in the IRDs, therapeutic approaches will have to be customized to the mutation and its pathologic effects. Several of the achromatopsia mutations appear to be amenable to possible gene therapy intervention since the cones survive, even though the photopic electroretinogram (ERG) becomes nonrecordable (see Figure 3).

Stem cell or progenitor cell therapies The term 'stem cell therapy' refers to interventions in which embryonic stem cells, or more-differentiated progenitor cells, are introduced to the retina with the hope that they will differentiate into functional neurons, replacing lost cells. Most such interventions reported to date have not been successful. The introduced cells have failed to differentiate into the needed retinal neurons or to integrate into the circuitry of the retina. Often the introduced cells do not survive, and there is no sign of repair of the photoreceptor degeneration. Recently, rod embryonic precursor cells harvested from neonatal mouse retina were shown to differentiate into rods and to integrate morphologically and functionally as rods into the retinas of the peripherin-deficient *rds* mouse and of a rhodopsin knockout strain. Provided the precursor cells were harvested in the first few days of postnatal life, during the period of rod genesis, appropriate synaptic connections were formed within the retina, and the electroretinogram showed restoration of physiologic responses. Other studies using bone marrow-derived progenitor cells have shown reparative properties to inner retinal layers of *rd*10 mice but no significant restoration of photoreceptors. These studies suggest that successful stem or progenitor cell interventions can be attained when the stage of development of the introduced cell is carefully controlled.

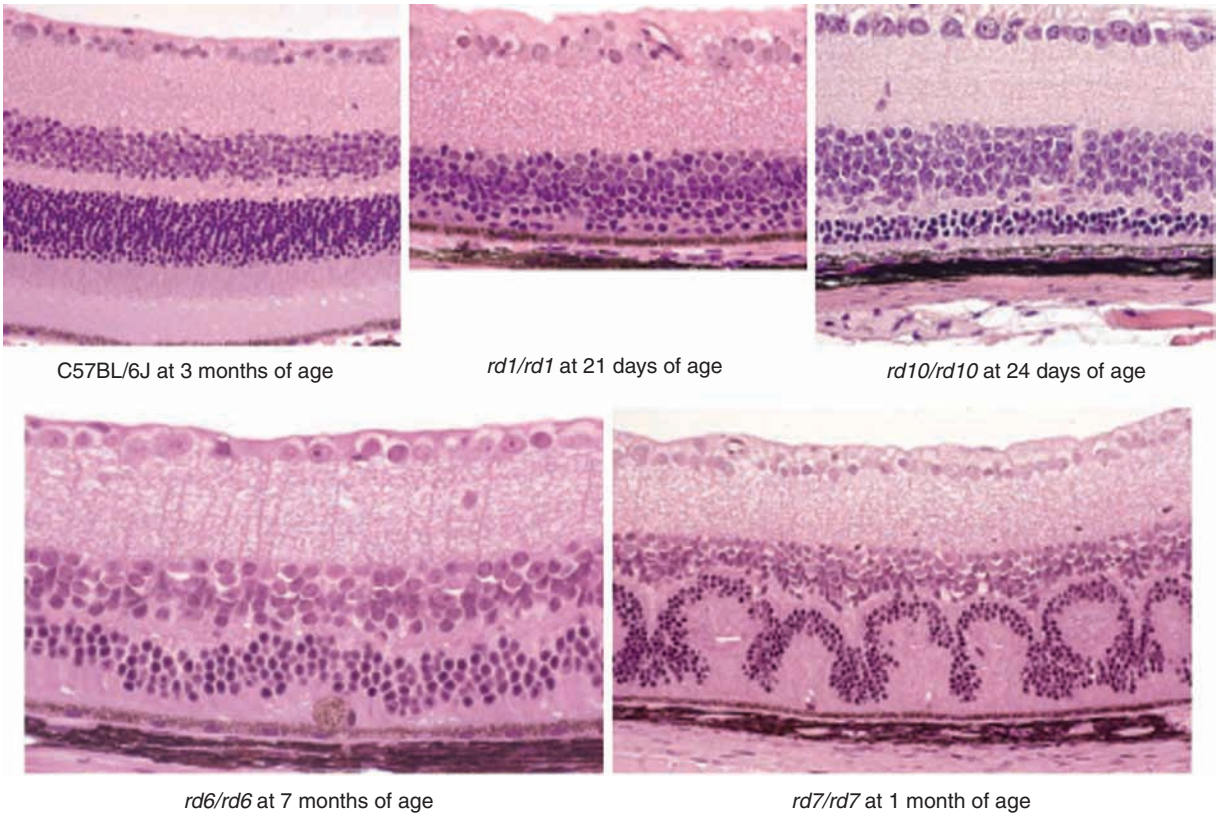

Figure 3 Retinal histologic phenotypes in the mouse, from a set of known mutations. The C57BL/6J is a nondegenerative, wild-type retina, used as a background for many engineered mutations. The retinal layer is normal. The *rd1/rd1* mouse is homozygous for a widely occurring viral insert in exon 7 of the β component for a retina-specific isoform of cGMP phosphodiesterase (PDE6). The outer nuclear layer (the layer of photoreceptor somas) is severely and quickly depleted by 1 month of age. The *rd10/rd10* mouse is homozygous for a missense point mutation in exon 13 of the same gene. The outer nuclear layer depleted, and distorted in the *rd6/rd6* and *rd7/rd7* mouse strains; these IRDs are associated respectively with a 4 bp deletion in the splice donor sequence of intron 4 in the *Mfrp* gene, and a 380-nt deletion from the coding region of a photoreceptor-specific nuclear receptor gene, *Nr2e3*. From Chang B, Dacey MS, Hawes NL, et al. (2006) Cone photoreceptor function loss-3, a novel mouse model of achromatopsia due to mutation in Gnat2. *Investigative Opthalmology & Visual Science* 47: 5017–5021.

Retinal transplantation Many studies using animal models have reported the transplantation of sheets of retinal cells under or on degenerating retina, with the transplanted retinal segments sometimes surviving as the host retina degenerates. No studies have shown that the transplanted cells have integrated into the circuitry of the retina sufficiently to transmit visual information to the brain.

Pharmaceutical interventions Large numbers of antiapoptotic drugs are known, and the various arms of the apoptosis pathway appear to be receptive to pharmacologic interventions. This therapeutic opportunity is attracting attempts to deliver drugs that regulate or inhibit the pathways that trigger apoptotic mechanisms. These drugs have been well tested in other tissues, but few have been tested in the retina.

More progress has been made using endogenous protective factors, proteins, and cytokines produced by the retina itself. Studies of the effectiveness of fibroblast growth factor-2 (FGF-2), ciliary neurotrophic factor (CNTF), and brain-derived neurotrophic factor (BDNF) were pioneered in animal models of retinal dystrophies, both inherited and stress-induced. These factors are endogenous, in the sense that the cells of the retina itself express them. When expressed, they protect the retina, especially photoreceptors, against stress, such as the stress of bright light. Further, their expression is stress inducible. They can be viewed, therefore, as inbuilt self-protective mechanisms, triggered by stress.

Their use as therapy has been inhibited by concerns about the possibility that the factors will have actions beyond photoreceptor protection, exerting trophic effects on blood vessels, on the tunica fibrosa of the eyeball, and on the differentiation of retinal neurons. Further, the factors can reduce the responsiveness of photoreceptors to light. This has been shown most clearly for FGF-2 and for CNTF, which binds to the outer segment, reduces rhodopsin expression, and by

altering the arrestin–rhodopsin ratio, downregulates phototransduction. Whether this downregulation is intrinsic to the protective action of CNTF is not known.

Nevertheless, work to test the value of these factors continues. BDNF is proving effective in protecting photoreceptors in an animal model of the collateral damage caused during photodynamic therapy of human eyes, and trials of CNTF delivered by an encapsulated intraocular implant of cells engineered to produce the factor have passed the initial safety stage. The discovery of these stress-inducible protective mechanisms shows how powerfully the retina can intervene to prevent photoreceptor death caused by both genetic mutations and environmental factors such as bright light. Further understanding of the mechanisms of action may allow their protective actions to be isolated and applied therapeutically.

Light restriction One of the earliest ideas of treatment for IRDs was to reduce the light reaching the retina, for example by wearing sunglasses. Early trials of light restriction in humans gave partly encouraging, partly disappointing outcomes, probably because the effects vary with the mutation. Some forms of rhodopsin mutation, the P23H mutation, for example, make photoreceptors particularly vulnerable to light. A recent survey of rodent IRD models showed that in the majority (15 of 20 strains summarized), dark rearing slows the degeneration and light exposure accelerates it. In other strains, the degeneration appeared to be unaffected by light exposure. In no case did light restriction make the degeneration worse.

In one rodent strain, the rhodopsin-mutant P23H-3 transgenic rat, light restriction has been shown, not just to slow the degeneration, but to allow photoreceptors to recover functional responsiveness to light; that is, the visual loss was partially reversed. The recovery results from the ability of damaged but surviving photoreceptors to rebuild their severely shortened outer segments and resume function. This self-repair ability may be a feature of wild-type photoreceptors; it has been described in the context of photostasis, the ability of photoreceptors to shorten and lengthen their outer segments as ambient light rises and falls, to keep the daily photon catch constant.

For many of the light-responsive strains, the underlying genetic mutation has been shown to occur in humans and cause a form of IRD. The animal data indicate, however, that light restriction is likely to be effective only in some forms of the disease.

Antioxidant therapy There is growing interest in the use of antioxidants, added to the diet or given by injection, in preventing retinal degenerations. Supplements such as β-carotene, vitamins C and D, and zinc have shown promise in reducing complications and may slow the onset and progress of age-related macular degenerations. Prefeeding with dietary β-carotene and saffron has been shown to be effective in protecting rat photoreceptors from light-induced damage, with saffron providing better functional preservation. In mouse IRDs, antioxidants given by injection have been shown to slow the death of cones after rods have died in the rd1 mouse and to slow the death of rods in rd1, rd10, and Q334ter mouse strains. These are distinct mutations, *rd*1 and *rd*10 in cGMP PDE and Q334ter in rhodopsin. The effectiveness of antioxidants in these several cases suggests that oxidative stress is a factor in most degenerations, as predicted by the oxygen toxicity hypothesis.

Near-infrared radiation The irradiation of tissue by light in the far- to near-infrared (NIR) range (670–880 nm) has been shown to improve the recovery of heart muscle from ischemic damage, to promote healing of opportunistic infections in immune deficiency states, and to attenuate degeneration in the injured optic nerve. Mechanistic studies have shown that NIR interacts with cytochrome oxidase, the mitochondrial enzyme which sequesters oxygen from extracellular fluid into oxidative phosphorylation pathways, triggering signaling pathways which result in improved adenosine triphosphate production; upregulation of the expression of protective, antioxidant factors; and improved cell survival. In the retina, NIR has been shown to protect photoreceptors against toxin-induced mitochondrial dysfunction, and preliminary studies have shown a marked slowing of the developmental stage of photoreceptor death in the P23H-3 rat.

The dosages required for these effects are low (fluence of $4\,\mathrm{J\,cm^{-2}}$ for periods of minutes) and can be delivered noninvasively. Further, the mitochondrion is the organelle in which oxygen is sequestered into oxidative metabolism, with the production of damaging oxygen radicals. It is also the source of many cell death signals triggered by stress. This suggests that NIR is acting a critical site, where oxidative stress can lead to cell death.

The Range of Animal Models of Human Diseases

This article has touched on information from only a fraction of the scores of animal models in which mutations cause IRDs. The great majority of these models are murine. Mouse strains with spontaneous or engineered mutations now provide mutation-specific

models of many human IRDs for which the gene responsible has been identified (see Figure 4). Indeed, novel IRDs recognized first in the mouse have provided candidates for gene discovery when a new human family presents with an IRD of unknown cause. More generally, the value of animal models, and particularly the mouse forms, seems likely to grow, for several reasons. The entire genomes of human and mouse are now known,

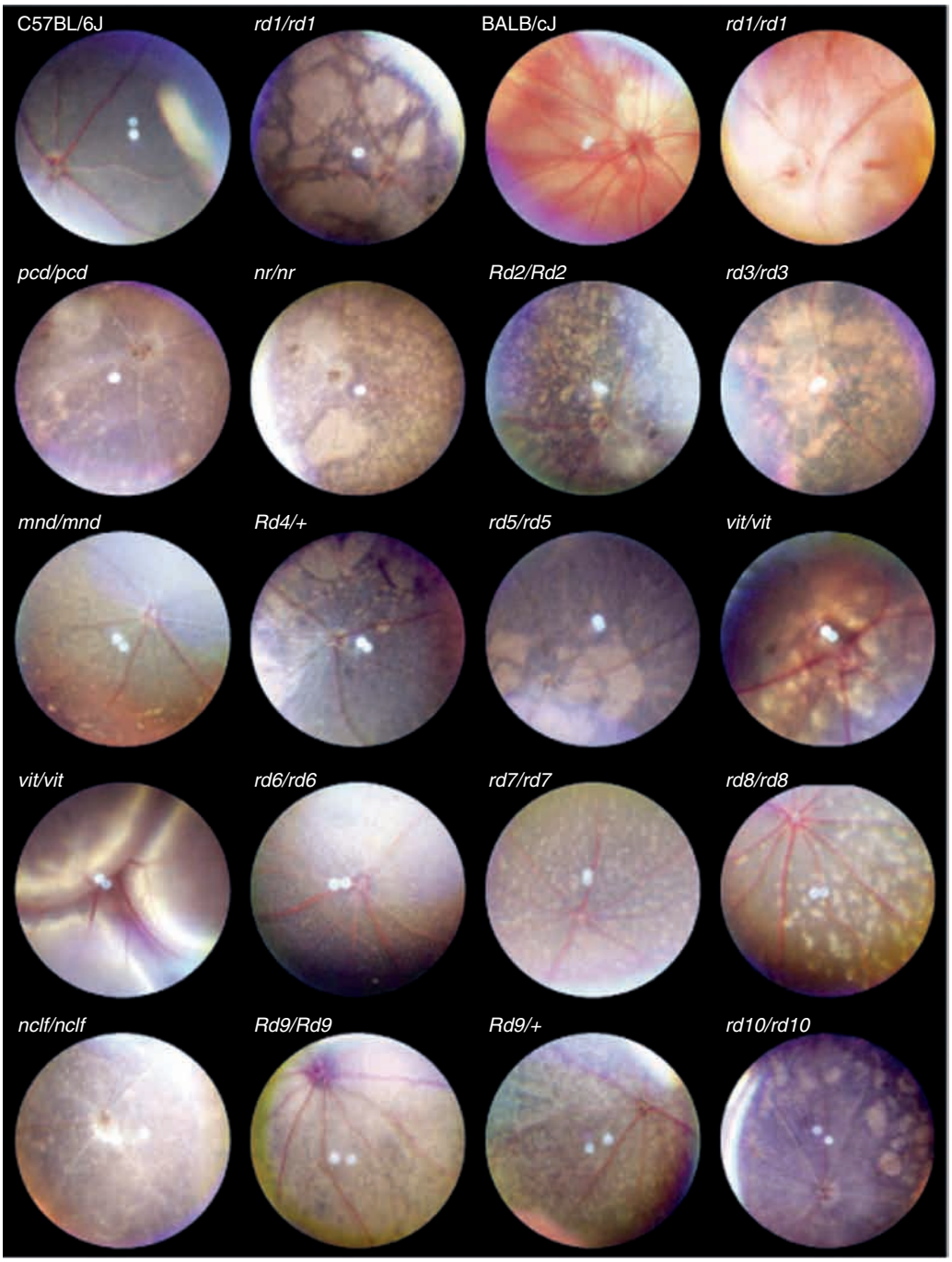

Figure 4 The appearance of the fundus of the mouse eye as seen through an indirect ophthalmoscope. The appearance of the fundus varies characteristically with the form of the inherited retinal dystrophy (IRD). This approach is noninvasive and a valuable screening tool for disease. The C57BL/6J and Balb/CJ retinas are nondegenerative wild types, the former pigmented, the latter albino. The lack of pigment gives the Balb/C fundus its characteristically pink (albino) appearance. The retinal blood vessels radiating from the optic disc are a prominent feature of the fundus. The fundi of IRD retinas occasionally show abnormal spots, thinning of the retinal vessels, and abnormalities of pigmentation. Most have their counterparts in the human IRDs. From Chang B, Hawes NL, Hurd RE, Davisson MT, Nusinowitz S, and Heckenlively JR (2002) Retinal degeneration mutants in the mouse. *Vision Research* 42: 517–525.

and the homology between mouse and human genomes is very high, of the order of 95%. The physiology, structure, and organization of the retina are very similar in the two species, so that the mouse retina is a strong model for the primate retina, except that it lacks a central specialization comparable to the fovea centralis or macula lutea found in primates such as the human. The techniques of engineering transgenic and knockout strains are most highly developed for the mouse. The life cycle of the mouse is naturally short, allowing the relatively rapid observation of developmental and aging processes. Finally, because the mouse genome has been extensively mapped, known markers of gene location are abundant, reducing the time needed to identify mutations (see Figure 5).

The use of other animal models has complemented this reliance on the mouse. The success of gene therapy in two dog strains has been very influential in establishing the potential of this approach, and the larger eyes of larger species provide better models for testing the delivery of drugs and engineered gene therapy constructs to the human eye. The study of

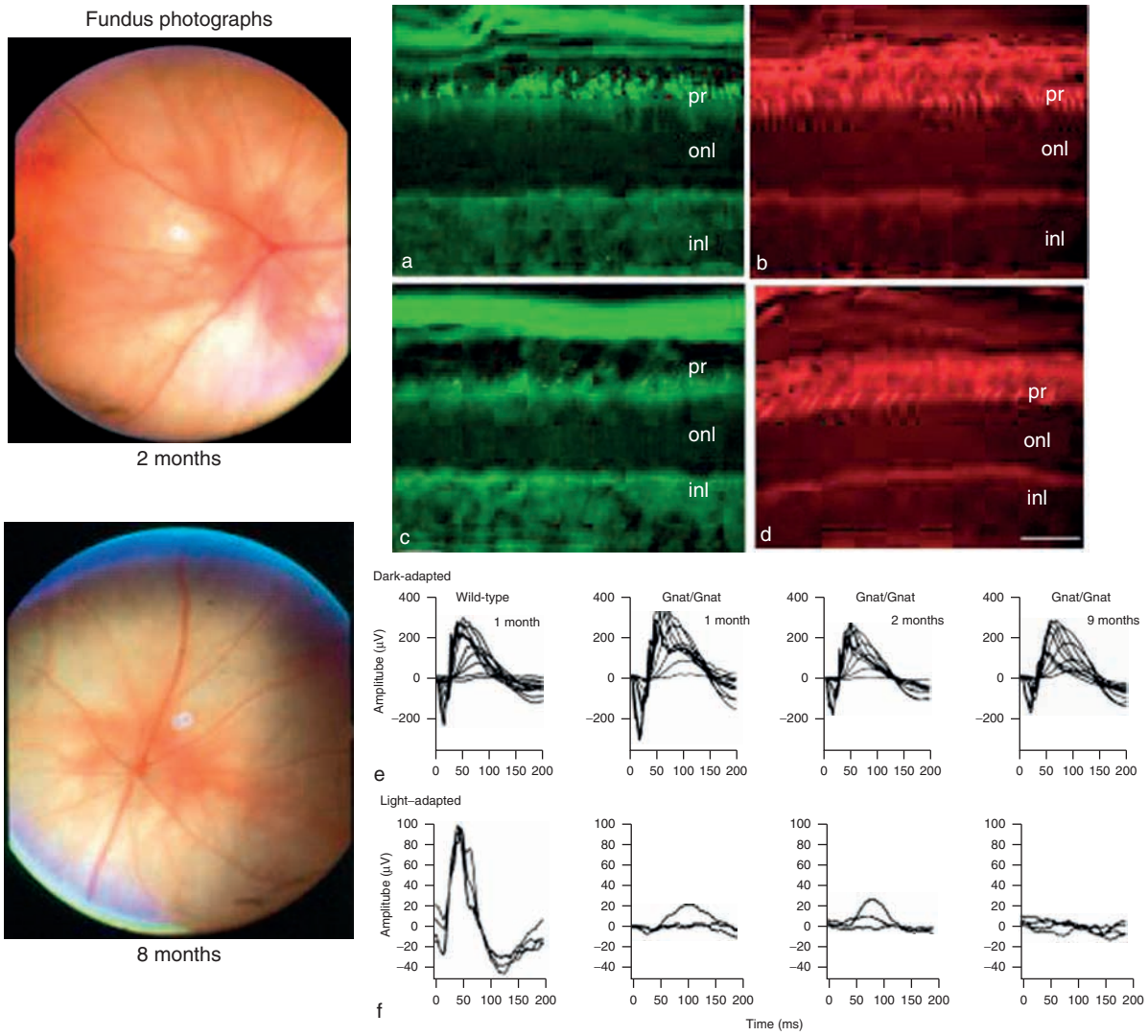

Figure 5 Mutation in mice strain with *GNAT2*, guanine nucleotide binding protein alpha-subunit of transducin, which is necessary for hyperpolarization of the cone photoreceptor. Fundus photograhs of GNAT2 mutant at 2 months and 8 months show progressive vascular attenuation in albino mouse with mutation. Histology Figures 5(a)–5(d), shows staining of α-transducin receptors in wild type (a) and mutant mouse (c) at 4 weeks, and cone staining with rhodamine-conjugated PCNA in wild type (b) and mutant (d). The GNAT-2 mutant shows the presence of large numbers of cones even though the α-transducin expression is greatly reduced and the photopic ERG is very abnormal. The retention of cones in the face of poor cone function suggests patients with this disease may respond to early gene therapy, before degeneration occurs. Scale bar = 50 μm. From Chang B, Dacey MS, Hawes, NL, et al. (2006) Cone photoreceptor function loss-3, a novel mouse model of achromatopsia due to a mutation in Gnat2. *Investigative Ophthalmology & Visual Science* 47: 5017–5021.

fish retina has demonstrated mechanisms of lifelong retinal neurogenesis, which underlie the lifelong growth of the retina in teleosts. In particular, Müller cells act as stem cells, able to generate the full spectrum of retinal cells in order to maintain retinal growth or repair lesions. This knowledge has led to a search for comparable mechanisms in the adult mammalian retina, which is normally unable to regenerate. Evidence is growing that mammalian Müller cells can be induced experimentally to act as stem cells, raising the possibility that regeneration can be induced in human retina to replace photoreceptors lost in a dystrophy.

Because until recently their pathophysiology has been unknown, IRDs have proved intractable to therapy over the century since they were recognized in humans. A range of therapeutic approaches, and many animal models, will be needed as the knowledge that has accumulated about the pathogenesis of IRDs is translated into long-awaited therapies.

See also: Photoreceptor Adaptation; Photoreceptors: Physiology; Phototransduction; Retinal Ganglion Cells: Anatomy; Retinal Ganglion Cells: Receptive Fields; Retinal Horizontal Cells; Retinal Pharmacology: Inner Retinal Layers; Retinal Development: An Overview; Retinitis Pigmentosa.

Further Reading

Bernardos RL, Barthel LK, Meyers JR, and Raymond PA (2007) Late-stage neuronal progenitors in the retina are radial Muller glia that function as retinal stem cells. *Journal of Neuroscience* 27: 7028–7040.

Bowne SJ, Liu Q, Sullivan LS, et al. (2006) Why do mutations in the ubiquitously expressed housekeeping gene IMPDH1 cause retina-specific photoreceptor degeneration? *Investigative Ophthalmology & Visual Science* 47: 3754–3765.

Chang B, Dacey MS, Hawes NL, et al. (2006) Cone photoreceptor function loss-3, a novel mouse model of achromatopsia due to a mutation in Gnat2. *Investigative Ophthalmology & Visual Science* 47: 5017–5021.

Chang B, Hawes N, Davisson M, and Heckenlively J (2007) Mouse models of RP. In: Tombran-Tink J and Barnstable CJ (eds.) *Retinal Degenerations: Biology, Diagnostics, and Therapeutics*, pp.149–160. Totowa, NJ: Humana Press.

Chang B, Hawes NL, Hurd RE, Davisson MT, Nusinowitz S, and Heckenlively JR (2002) Retinal degeneration mutants in the mouse. *Vision Research* 42: 517–525.

Cideciyan AV, Jacobson SG, Aleman TS, et al. (2005) *In vivo* dynamics of retinal injury and repair in the rhodopsin mutant dog model of human retinitis pigmentosa. *Proceedings of the National Academy of Sciences of the United States of America* 102: 5233–5238.

Clarke G, Collins RA, Leavitt BR, et al. (2000) A one-hit model of cell death in inherited neuronal degenerations. *Nature* 406: 195–199.

Clarke G, Heon E, and McInnes RR (2000) Recent advances in the molecular basis of inherited photoreceptor degeneration. *Clinical Genetics* 57: 313–329.

Coleman H and Chew E (2007) Nutritional supplementation in age-related macular degeneration. *Current Opinion in Ophthalmology* 18: 220–223.

Daiger SP, Sullivan LS, and Browne SJ (2007) *RetNet: Retinal Information Network*, http://www.sph.uth.tmc.edu/RetNet/ (accessed Jul. 2008).

Daiger SP, Sullivan LS, Browne SJ, and Rossiter BJF (2008) RetNet: Cloned and/or mapped genes causing retinal diseases. Available at http://www.sph.uth.tmc.edu/RetNet.

Doonan F, Donovan M, and Cotter TG (2005) Activation of multiple pathways during photoreceptor apoptosis in the *rd* mouse. *Investigative Ophthalmology & Visual Science* 46: 3530–3538.

Duncan JL, Paskowitz DM, Nune GC, et al. (2006) Retinal damage caused by photodynamic therapy can be reduced using BDNF. *Advances in Experimental Medicine and Biology* 572: 297–302.

Hitchock P and Raymond PA (2004) The teleost retina as a model for developmental and regeneration biology. *Zebrafish* 1: 257–272.

Jozwick C, Valter K, and Stone J (2006) Reversal of functional loss in the P23H-3 rat retina by management of ambient light. *Experimental Eye Research* 83: 1074–1080.

Komeima K, Rogers BS, and Campochiaro PA (2007) Antioxidants slow photoreceptor cell death in mouse models of retinitis pigmentosa. *Journal of Cellular Physiology* 213(3): 809–815.

Komeima K, Rogers BS, Lu L, and Campochiaro PA (2006) Antioxidants reduce cone cell death in a model of retinitis pigmentosa. *Proceedings of the National Academy of Sciences of the United States of America* 103: 11300–11305.

MacLaren RE, Pearson RA, MacNeil A, et al. (2006) Retinal repair by transplantation of photoreceptor precursors. *Nature* 444: 203–207.

Mervin K and Stone J (2002) Developmental death of photoreceptors in the C57BL/6J mouse: Association with retinal function and self-protection. *Experimental Eye Research* 75: 703–713.

Nusinowitz S and Heckenlively J (2006) Evaluating retinal function in the mouse retina with the electroretinogram. In: Heckenlively J and Arden G (eds.) *Principles and Practice of Clinical Electrophysiology of Vision*, pp. 899–910. Cambridge, MA: MIT Press.

Paskowitz DM, LaVail MM, and Duncan JL (2006) Light and inherited retinal degeneration. *British Journal of Ophthalmology* 90: 1060–1066.

Preising M (2007) *Retina International Scientific Newsletter.*

Rakoczy P, Yu MJ, Nusinowitz S, Chang B, and Heckenlively JR (2006) Mouse models of age-related macular degeneration. *Experimental Eye Research* 82: 741–752.

Stone J, Maslim J, Valter-Kocsi K, et al. (1999) Mechanisms of photoreceptor death and survival in mammalian retina. *Progress in Retinal and Eye Research* 18: 689–735.

Stone J, Mervin K, Walsh N, Valter K, Provis J, and Penfold P (2005) Photoreceptor stability and degeneration in mammalian retina: Lessons from the edge. In: Penfold P and Provis J (eds.) *Macular Degeneration: Science and Medicine in Practice*, pp. 149–165. Berlin: Springer.

Sullivan LS, Heckenlively JR, Bowne SJ, et al. (1999) Mutations in a novel retina-specific gene cause autosomal dominant retinitis pigmentosa. *Nature Genetics* 22: 255–259.

Valter K, Bisti S, and Stone J (2003) Location of CNTFR on outer segments: Evidence of the site of action of CNTF in rat retina. *Brain Research* 985: 169–175.

Wu WW, Wong JP, Kast J, and Molday RS (2005) RS1, a discoidin domain-containing retinal cell adhesion protein associated with X-linked retinoschisis exists as a novel disulfide-linked octamer. *Journal of Biological Chemistry* 280: 10721–10730.

Relevant Websites

http://www.retina-international.org – Retina International Animal Model Database.

http://www.retina-international.org – Retina International Mutation Databases.

http://www.sph.uth.tmc.edu – RetNet: Retinal Information Database.

Animal Models of Motor and Sensory Neuron Disease

P Bomont and D W Cleveland, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Motor and Sensory Tracts

Movements of many parts of the body, from the lips and the eyelids to the hands and toes, have their origin in the brain in a specialized region called the motor cortex (Figure 1). Upper motor neurons receive information there that they transmit to spinal motor neurons, often through an intermediate interneuron. Firing of the lower motor neuron triggers muscle contraction. Upper motor neuron degeneration induces brisk tendon reflexes, spasticity, and hyperreflexia, whereas denervation of the muscle upon lower motor neuron degeneration or loss leads to muscle weakness and loss of tone.

The sensory system works in the opposite direction with information flowing to the brain from the periphery. Receptors, primarily located at the surface of the skin, provide initial sensory input (touch, pressure, vibration perception, pain, and temperature) to the axons of the sensory neurons which carry these signals into the spinal cord. The information is then transferred to the neurons located in the sensory cortex, where the information is decoded by the brain.

Gene Deletion ('Knockout') and Transgenic Mouse Models

Two main principles are used to reproduce inherited human pathologies in the mouse (Figure 2). When the human disease is recessive, a situation usually caused by the absence of one particular protein, a genetic mimic in mice can be constructed by disrupting the corresponding gene in the mouse genome (Figure 2(a)). These models are frequently called 'knockout models,' referring to the disruption or removal of the targeted gene and its encoded protein. A variation of this method, called 'conditional knockout,' permits the removal of the gene of interest only in chosen tissues (Figure 2, box 1). This alternative is preferred in cases in which the deletion of the gene from the entire animal induces lethality in early development.

For dominant neurodegenerative disorders, disease is usually caused by the gain of a toxic property resulting from a modification in the protein (an example of this is amyotrophic lateral sclerosis (ALS) caused by mutation in the gene encoding superoxide dismutase). Two approaches are possible for generating a mouse model. The mouse gene can be replaced by the mutated one (Figure 2(a)), thereby creating a gene 'knock in' model. Alternatively, random insertion of the abnormal gene in the mouse genome can be achieved to produce a transgenic mouse (Figure 2(b)). The latter is largely preferred because it can be constructed more quickly. Importantly, transgene insertion can be readily achieved in mice or rats. An advantage to rat models in neuroscience is that the larger size permits more complex surgical approaches and tissue dissections. Additionally, for some classes of neurons, those derived from rats are more easily grown in culture. Finally, rats are amenable to some behavioral analyses that are more difficult in the mouse.

Neurodegeneration in Human

Motor Disease

Amyotrophic lateral sclerosis ALS, more familiarly known in the United States as Lou Gehrig's disease, is the most prominent adult motor neuron disease, with a typical age of disease onset of 50–60 years. Most (90%) incidences of ALS are referred to as sporadic because there is no family history of disease and hence no evidence for a major genetic component. Disease involves the selective alteration and death of both the upper and the lower motor neurons, resulting in spasticity, hyperreflexia, and progressive weakness of skeletal muscles, atrophy, and death due to respiratory muscle paralysis within 1–5 years after onset.

For the incidences (~10%) of ALS with a genetic origin, determination of the gene(s) responsible has been significantly challenged by the wide heterogeneity of the disease, genetically and clinically. Many genetic loci have been associated with different forms of motor neuron degeneration. Most prominently, mutations in the gene encoding for Cu/Zn superoxide dismutase (SOD1), an enzyme used to detoxify an aberrant oxygen species, have been shown to account for 20% of the familial cases of ALS. Dominantly inherited, analysis of enzymatic activity of mutated SOD1 in ALS patients together with genetic manipulations in mice have clearly revealed that SOD1-mediated toxicity is not due to a reduction in activity of this enzyme but, rather, due to a gain of a new toxic property or properties. Accordingly, transgenic rodent models have been generated by inserting into their genome a human SOD1 gene carrying different ALS-causing mutations. Many models have been created to dissect the disease-associated toxic mechanism(s), the most extensively used being transgenic for the

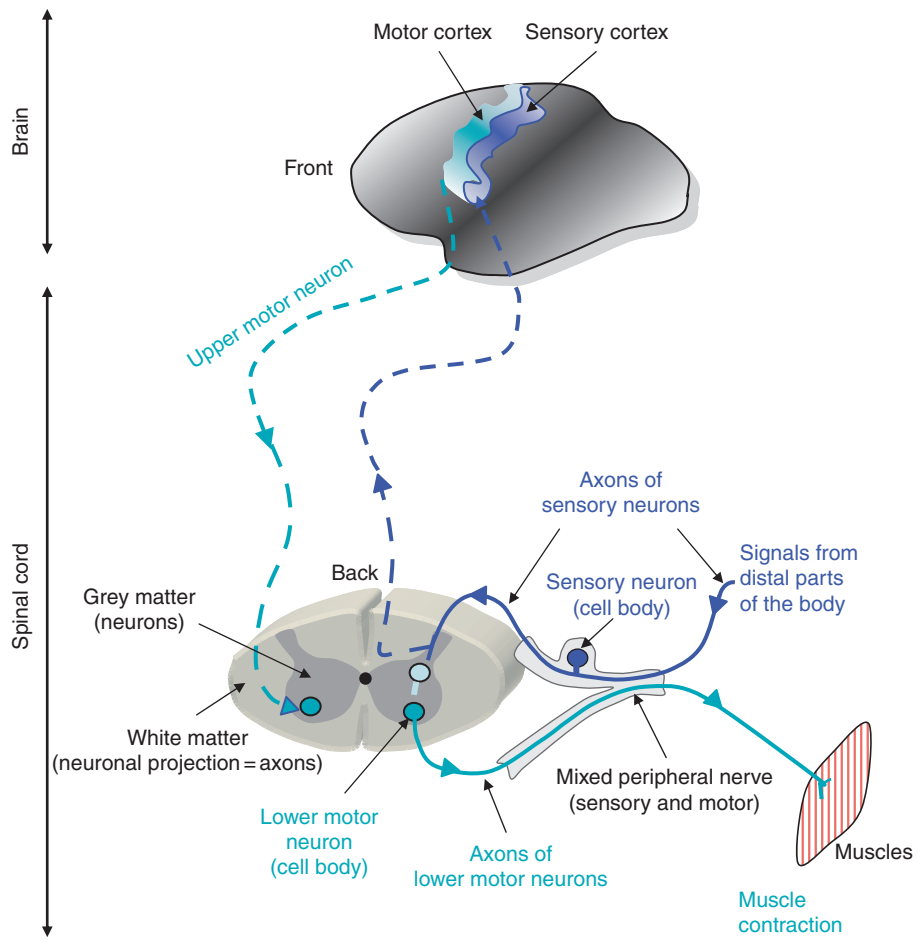

Figure 1 Motor and sensory pathways. Sensory neurons receive information from different parts of the body and transmit it to the spinal cord and then to the brain through the ascending axons of sensory neurons (dark blue). The motor pathway originates in the motor cortex, where upper motor neurons (light blue) connect to transmit information to the lower motor neurons in the spinal cord, which in turn send and deliver signals to the muscles, inducing their contraction.

amino acid substitution mutations G93A (glycine substituted to alanine at position 93), G37R (glycine to arginine at position 37), and G85R (glycine to arginine at position 85). Although none of these completely recapitulate all features of the human disease, they all develop progressive motor neuron degeneration with slightly different symptoms and disease progressions that are characteristic for each mutation, with onset and survival determined by the level of mutant expression. All of these mice develop motor neuron degeneration, limb weakness associated with neurofilament misaccumulation, impaired axonal transport, and axonal swelling.

Mouse and rat models for ALS have been extensively used to test proposals for a wide range of pathological mechanisms. These include impairment of the chaperone/degradation machinery, dysfunction of mitochondria, oxidative stress, alteration of cytoskeletal architecture and axonal transport, glutamate-mediated excitotoxicity, aberrant growth factor signaling, microglial cell involvement, and inflammation. Another aspect particularly important to the identification of therapeutic targets is whether other cell types contribute to motor neuron degeneration. Motor neurons require support provided by several cell types in the central nervous system which also express the mutant SOD1. Use of a transgene that can be eliminated selectively from motor neurons has shown that mutant damage in motor neurons initiates disease onset and an early phase of disease progression. A similar approach has proven that mutant damage within microglial cells, the immune cells of the central nervous system, has no effect on disease onset but accelerates disease progression.

Additional forms of human ALS-like motor neuron disease, all associated with motor neuron degeneration and death, have been designated ALS2–ALS8. Significant divergence in the age of onset and clinical presentation has led to disagreement as to whether these should be classified as forms of ALS. This is

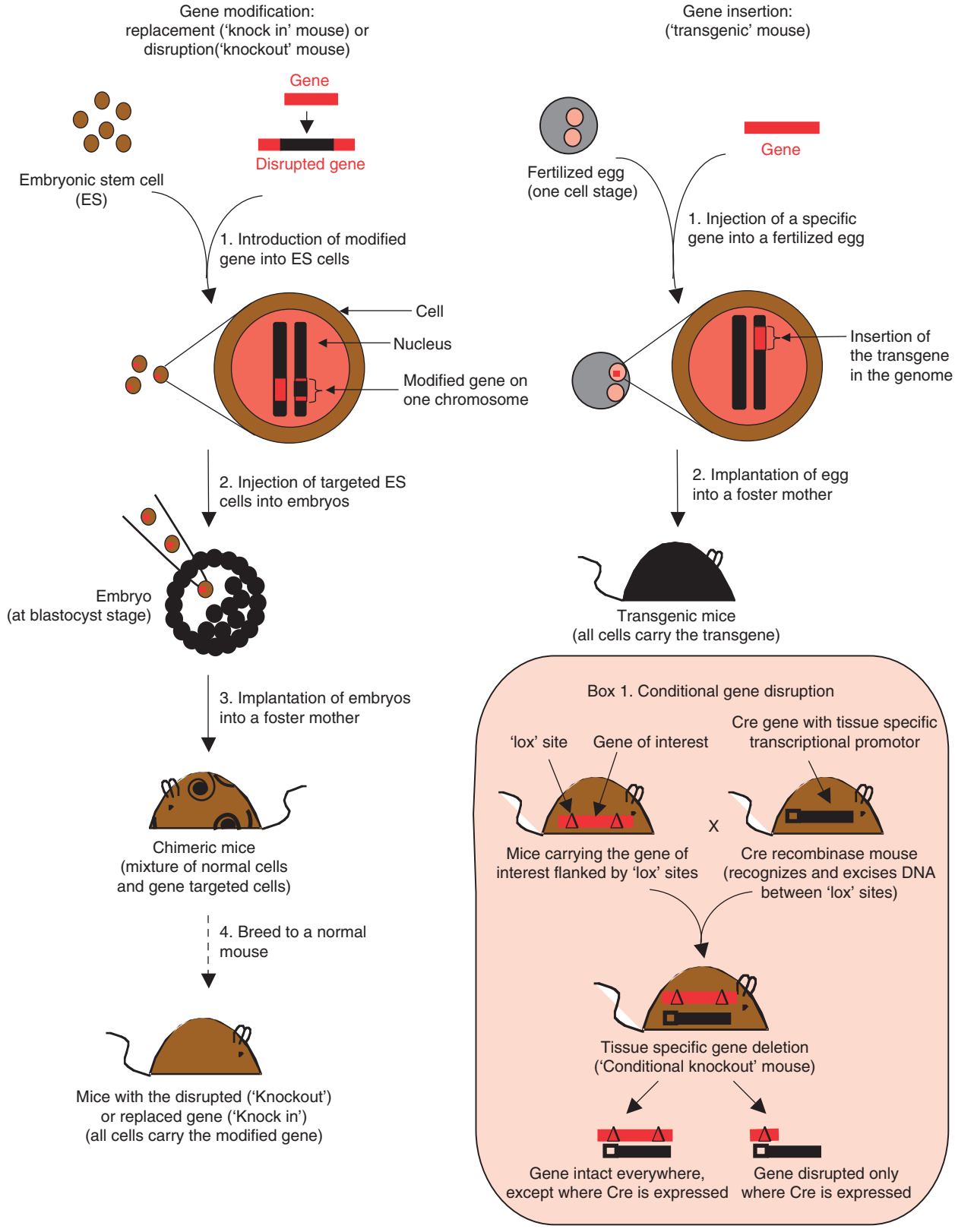

Figure 2 Continued

especially so for ALS2, an infantile and juvenile-onset form of motor neuron disease which is caused by loss of an apparent guanine exchange factor enzyme, termed alsin, that can activate one or more enzymes from the family of small GTPases known as G-proteins. Proposed G-protein partners of alsin include Rab5 and Rac1. Alsin has been shown to modulate processes such as cytoskeleton organization and transport of membrane cargoes.

Recessive AL2 motor neuron disease has been reproduced in mice by deleting the corresponding gene (**Figure 2(a)**). This first revealed that the absence of alsin in the mouse does not produce a juvenile disease that has as severe a phenotype as found in human. The lack of perfect concordance between the human and mouse genetic defect has been proven to be true for many other diseases. Importantly, lower motor neurons (those that directly trigger contraction of skeletal muscles) are almost completely spared in mice deleted of the ALS2 gene, which do not develop muscle weakness. Analysis of voluntary running, however, revealed that these mice move slowly, a well-recognized clinical sign of an upper motor neuron defect. Additionally, the upper motor neurons degenerate. Thus, the animal model reinforces that the pathology and the phenotype of loss of ALS2 provoke an upper motor neuron disease quite distinct from ALS, resembling instead hereditary spastic paralysis.

ALS4 is a dominantly inherited juvenile form of ALS characterized by distal muscle weakness and atrophy before 25 years of age. Rare in humans, this disease is caused by mutation in the senataxin (SETX) gene. The function of the senataxin protein is not known, but its sequence suggests DNA/RNA helicase activity, potentially acting in DNA repair pathways and RNA processing. Senataxin mutations are also causative of a recessive disease termed AOA2 (ataxia–oculomotor apraxia 2); symptoms include balance difficulties and defects in coordination of movements (ataxia) associated with an alteration of eye movement (oculomotor apraxia).

Axonal transport has been shown to play an essential role in neuronal survival. Two families of motor proteins, the kinesins and dynein, which power transport of cargoes along the microtubules away from the nerve cell body (anterograde) or returning components from the nerve ending to the cell body (retrograde), respectively, have been shown to cause or contribute to motor neuron disease in humans. This has been replicated in mice. In humans, mutation in the subunit p150 of the dynactin complex (which interacts with the dynein complex) has been identified to be causative of an unusual form of lower motor neuron disease in early adulthood, characterized by vocal fold paralysis, progressive weakness, and muscle atrophy.

Although a mouse model of the p150 dynactin mutation has not been reported, it is widely established that alteration of the dynein–dynactin complex causes motor deficits. Indeed, overexpression of dynamitin (also called p50 of the dynactin complex) has been shown in mice to disrupt dynein–dynactin interaction leading to late-onset slowly progressive motor neuron degeneration, characterized by muscle weakness, trembling, abnormal gait, and deficits in strength and endurance. Progressive alterations of muscle function and motor coordination have been linked in mice to distinct mutations in the dynein gene. The corresponding mutants, *Loa* (Legs at odd angles) and *Cra1* (cramping 1), develop motor deficits, confirming that alteration in the dynein motor complex and its presumed effects on moving components through axons are of unusual importance for motor axons. A completely unexpected and unexplained finding is that both dynein mutations delay disease progression and increase the life span of mice that develop early onset motor neuron disease caused by expression of an ALS-causing SOD1 mutant. Understanding what is causing the motor deficits in the *Loa* and *Cra1* mice will certainly help us to understand what leads to this amelioration of disease in SOD1 mutant mice and might help to identify new therapeutic approaches for ALS.

Figure 2 Mouse models for human neurodegenerative disease. Strategies for gene modification or replacement. (a) Gene modification. In this approach, the gene of interest is modified *in vitro* and then introduced into (brown) embryonic stem (ES) cells. Following recombination in the ES cells, the modified gene replaces the endogenous gene on one of the two chromosomes. The corresponding targeted ES cells are then injected into mouse embryos at the blastocyst stage and are implanted into a foster mother. Those females produce progeny mice that are mixtures of normal cells and modified cells. Mice carrying the modified gene in all cells (also called 'knockout mice' in the case of a gene disruption and 'knock in' when the gene is replaced) are obtained by standard breeding with normal mice. (Box 1) To selectively disrupt a gene in a specific tissue, the gene region to be deleted is flanked by two lox sites, which are small sequences recognized and cut by an enzyme called Cre recombinase. Breeding mice carrying the modified lox–gene–lox with mice expressing the Cre recombinase under a tissue-specific transcriptional promoter generates the gene disruption only in the tissues that express the Cre recombinase. (b) Gene insertion: 'transgenic' mouse. An exogenous gene is inserted randomly into the mouse genome after direct injection of the gene into the one cell stage fertilized egg. After implantation into a foster mother, the oocyte matures into mice whose cells all carry the transgene.

Spinal muscular atrophy After cystic fibrosis, spinal muscular atrophy (SMA) is the most common autosomal recessive disorder in humans and represents the most common genetic cause of infant mortality. In this disorder, specific neuronal loss of lower motor neurons in the spinal cord results in atrophy of proximal muscles of the trunk and the limbs. SMA cases are classified into three groups: Type I SMA is the most severe phenotype, with an age of onset before 6 months and death occurring before 2 years of age; type II SMA is intermediate, with an age of onset between 6 and 18 months; and type III SMA, which initiates after 18 months, is the mildest form, slowly progressing with a normal life span.

Identifying the genetic cause of SMA was challenging because of the complex and unstable nature of the region of human chromosome 5 where the defective gene resides. For all disease types, symptoms are caused by inactivating mutations in the *SMN1* (survival of motor neuron 1) gene, but the variability in the disease severity depends on the number of copies (ranging from one to five) of an almost identical copy gene, the *SMN2*, located just next to *SMN1*. Each copy of the *SMN2* gene produces only a small amount of functional product so that only a single copy leads to fatal, severe type I disease, and five copies lead to the more benign type III disease. Although the role of SMN protein in lower motor neuron survival is not fully understood, it has been implicated in processing within the nucleus of initial RNA copies of each expressed gene into their mature forms capable of translation into the corresponding proteins. A second likely function is as an RNA-binding protein for RNA transport into axons.

Because the mouse genome contains only the *SMN1* gene and no *SMN2* gene, mouse models for SMA have been attempted by deleting the *SMN1* locus. These animals are not viable, demonstrating an essential role for SMN1 in cell survival. Selective deletion of the *SMN1* gene solely from neurons (**Figure 2**, box 1) produced mice with motor abnormalities and skeletal muscle denervation secondary to motor axon loss, culminating in death at a mean age of 25 days. This is not a very satisfying model, but it does demonstrate an essential role for SMN1 in neurons. SMN1 expression has also been modified only in muscle cells. This has revealed that SMN expression in muscle prevents the SMA phenotype and increases the life span of the animals. This supports targeting muscle as a therapeutic strategy in SMA. Because in human the severity of disease is modulated by the *SMN2* gene, the *SMN2* gene has been introduced (by a transgene) to the mice with the 'neuronal' deletion of *SMN1*. This showed that *SMN2* gene number ameliorates SMN1 absence, closely mimicking what occurs in human disease.

Motor and Sensory Disease

Charcot–Marie–Tooth Charcot–Marie–Tooth (CMT) diseases refer to a heterogeneous class of neuropathies that affect not only motor but also sensory nerves in the peripheral nervous system. In addition to the weakness and atrophy of the distal limb muscles, patients experience impaired sensation and absence of deep tendon reflexes. Representing the most common inherited disorder of the peripheral nervous system, this group of diseases, which are sometimes transmitted through dominant, recessive, or X-linked modes of inheritance, has been classified into two subgroups. The demyelinating form (also called CMT1) results in an impairment of the myelin sheath produced by Schwann cells. CMT2 forms are characterized by degeneration of the axon. The genetic heterogeneity of the disease has led to the identification of at least 17 genes involved in both demyelinating and axonal forms of CMT. The CMT1-causing genes (*PMP22*, *MPZ*, *LITAF/SIMPLE*, and *EGR2*) have revealed an essential role of myelin compaction, the regulation of myelin protein degradation, and the transcriptional control of myelination-specific genes in demyelination.

Axonal survival has been shown to depend on many different pathways. Among these, mutation in the gene encoding the mitochondrial GTPase mitofusin 2 as a cause of CMT has proven an essential role of mitochondrial fusion/transport in axon survival. Maintenance of axonal architecture and transport (for motor neuron survival) is particularly important for sustaining neuron function. Forms of CMT2 are also caused by mutation in the neurofilament subunit NF-L or the motor kinesin KIF1bβ. Impaired intracellular trafficking is also a common theme in axonal degeneration, as revealed by mutation in the regulator of vesicle trafficking Rab7 and the phosphatase MTMR2, which modulate membrane trafficking.

CMT1A is the most frequent form of CMT. Instead of a typical disease-causing mutation that alters or eliminates the encoded protein, CMT1A disease is caused by a duplication of the genomic region surrounding an otherwise normal *PMP22* gene so that affected individuals have an extra copy of the normal gene. Transgenic rat and mouse models containing multiple copies of the *PMP22* gene have proven that overexpression of only this gene is sufficient to cause peripheral demyelination and the symptoms associated with the human disease. This also demonstrated an increased disease severity with increased expression of PMP22. These animal models have been shown to be very useful to test the first rational experimental therapies of CMT1A. Indeed, a synthetic antagonist of the nuclear progesterone receptor was shown to reduce PMP22 expression and to

ameliorate the clinical severity in those animals. Moreover, administration of ascorbic acid, an essential factor of *in vitro* myelination (already approved by the US Food and Drug Administration for other clinical indications), has prolonged survival and restoration of myelination in those models, suggesting its use in human clinical trials for CMT1A.

Giant axonal neuropathy Factors that establish the cytoarchitecture of the axon and its associated axonal transport components are critical for neuronal survival. Disorganized cytoskeletal intermediate filaments constitute the hallmark of another neurodegenerative disease termed giant axonal neuropathy (GAN). This is a rare, recessively inherited condition characterized by a structural deficit in both the central nervous system and the peripheral motor and sensory axons. With disease onset in infancy and marked by gait instability and frequent falls, patients develop diminution of deep tendon reflexes, muscle atrophy, and muscle weakness that progressively evolves to sensory loss and loss of ambulation. Other symptoms, including ataxia, language deficits, and mental retardation, reveal a later impairment of the central nervous system.

GAN is a progressive, fatal disease, with life expectancy of less than 30 years. Identification of the *GAN* gene led to recognition that its encoded protein, gigaxonin, plays a critical role in the ubiquitination machinery whose function, among others, is to tag proteins for degradation. Other errors in other components of this degradation apparatus are implicated in several neurodegenerative disorders, the most prominent of which is Parkinson's disease. Gigaxonin seems to link degradation and impairment of the axonal cytoskeleton. Loss of gigaxonin prevents normal degradation of proteins involved in the assembly and dynamics of microtubules, the protein polymers that serve as the tracks for transport of components up and down the axon. Whether altered microtubules properties are the disease causing damage and, if so, how this provokes the specific aggregation of intermediate filaments seen in patients are not known.

Developing animal models for rare disorders such as GAN is particularly crucial since insights from direct inspection of patient materials are very limited. The study of the mechanisms underlying neurodegeneration through construction and analysis of genetic models mimicking the loss of gigaxonin is at its earliest stages. Mice deleted of the *GAN* gene develop a progressive deterioration of motor function and ataxia, a sign of central nervous system impairment. As in human disease, the absence of gigaxonin generates axonal loss and neurofilament aggregation. Alteration of cytoskeleton architecture has been further confirmed by the decreased density of axonal microtubules and the increased abundance of proteins that affect microtubule assembly.

Conclusion

Since essentially all forms of motor or sensory neuron degeneration are incurable, mouse and rat models represent an indispensable resource to address the origin of neuronal dysfunction, the mechanisms of degeneration, and the determinants of the selectivity of death of subpopulations of neurons. In turn, these key issues will power the design of strategies to ameliorate these life-long, frequently fatal conditions. Thus, for example, for ALS, this fundamental research has led to the development of many experimental therapies in these animals, including antiglutamate drugs, neurotrophic growth factor delivery, SOD1 silencing, and stem cell therapy, which will be tested in human clinical trials. Identifying the pathways of neuron survival will help in understanding the basis of motor/sensory neuron diseases in humans for which the majority of genetic causes remain unknown.

See also: Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Amyotrophic Lateral Sclerosis (ALS); Axonal Transport and ALS; Neurofilaments: Organization and Function in Neurons.

Further Reading

Boillee S, Vande Velde C, and Cleveland DW (2006) ALS: A disease of motor neurons and their nonneuronal neighbors. *Neuron* 52: 39–59.

Boillee S, Yamanaka K, Copeland NG, et al. (2006) Motor neurons and microglia as determinants of disease onset and progression in inherited ALS. *Science* 312: 1389–1392.

Bomont P, Cavalier L, Blondeau F, et al. (2000) The gene encoding gigaxonin, a new member of the cytoskeletal BTB/kelch repeat family, is mutated in giant axonal neuropathy. *Nature Genetics* 26(3): 370–374.

Chen YZ, Bennett CL, Huynh HM, et al. (2004) DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *American Journal of Human Genetics* 74(6): 1128–1135.

Ding J, Allen E, Wang W, et al. (2006) Gene targeting of GAN in mouse causes a toxic accumulation of microtubule-associated protein 8 and impaired retrograde axonal transport. *Human Molecular Genetics* 15(9): 1451–1463.

Monani UR (2005) Spinal muscular atrophy: A deficiency in a ubiquitous protein; A motor neuron-specific disease. *Neuron* 48(6): 885–896.

Rosen DR, Siddique T, Patterson D, et al. (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 362(6415): 59–62.

Shy ME (2004) Charcot–Marie–Tooth disease: An update. *Current Opinion in Neurology* 17(5): 579–585.

Animal Models of Parkinson's Disease

K E Soderstrom, G Baum, and J H Kordower, Rush University Medical Center, Chicago, IL, USA

© 2009 Published by Elsevier Ltd.

Introduction

Parkinson's disease (PD) is the second-most-common neurodegenerative disorder (after Alzheimer's disease), affecting more than 1 000 000 people in the United States. While there is widespread degeneration in the central and peripheral nervous systems in PD, the hallmark pathology remains the dopaminergic striatal insufficiency secondary to degeneration of dopaminergic neurons in the substantia nigra (SN). While a small percentage of PD is of a genetic origin, the vast majority of cases are sporadic and of unknown origin. It is particularly challenging to model a disease whose etiology is, for the most part, unclear. Therefore, PD researchers have attempted to model specific aspects of PD pathology independent of etiology. These attempts have included pharmacological models aiming to replicate striatal dopamine (DA) depletion and neurotoxic models that replicate nigral cell loss. While excellent for evaluating the role of the nigrostriatal system in PD and for testing therapeutic compounds, most of these models are limited in their ability to replicate such important aspects of the human disease as Lewy body formation, nondopaminergic systems affected in PD, and the progressive nature of the disease. Novel models emerging in the field have attempted to model these important facets of the human disease more accurately. Genetic models, the lipopolysaccharide model, and the use of aged animals as a PD model are all emerging alternatives to traditional DA-centric models of PD. The validity, strengths, and weaknesses of these traditional and novel animal models of PD will be addressed in this article, as well as the importance of animal models in general in our understanding and treatment of PD.

PD

PD is a progressive neurodegenerative disorder clinically characterized by the cardinal symptoms of cogwheel rigidity, resting tremor, bradykinesia, stooped posture, and shuffling gait. As stated above, there is a loss of dopaminergic cells in the SN pars compacta (SNpc) that results in insufficient DA innervation of the basal ganglia and subsequent increased inhibition of excitatory thalamo-cortical connections. Lewy bodies, intracellular inclusions principally containing α-synuclein, are also found in the remaining nigral neurons of PD patients.

While motor symptoms are often the most obvious deficits exhibited in PD, it is clear that numerous nonmotor and peripheral nervous system symptoms can significantly affect patient quality of life. Braak and co-workers recently used α-synuclein immunohistochemistry to hypothesize that PD pathology begins in the lower brain stem and olfactory bulb and then progresses in a caudal to rostral fashion. This progressive nature of PD and the involvement of nondopaminergic pathways in its development have traditionally been unrepresented in animal models of PD, despite the growing realization of their role in the human disease.

DA Depletion

The ultimate result of cell loss and cell dysfunction in the SN is the depletion of the neurotransmitter DA in the basal ganglia. This insufficient DA innervation is principally localized to the postcommissural putamen and results in the overdrive of globus pallidus and subthalamic nuclear outputs. The resulting inhibition of thalamocortical function results in the characteristic bradykinesia experienced by PD patients. Researchers first attempted to model PD by replicating this DA depletion in animals. Early studies using pharmacological interventions aimed to selectively deplete monoamine neurotransmitters lost in PD patients. Of these pharmacological models, one of the earliest and most utilized has been the reserpine model, which prevents the storage of DA in presynaptic terminals.

Reserpine

In the late 1950s, Arvid Carlsson demonstrated that the drug reserpine could cause reversible parkinsonism in animals, in a manner similar to that seen in humans. Reserpine blocks the storage of monoamines in catecholaminergic neurons, resulting in DA depletion in the striatum. Treating rabbits with reserpine causes a behavioral syndrome that modeled idiopathic PD and helped elucidate the critical role of DA in the pathogenesis of this disease. This new concept radically changed clinical neurology and lead to the discovery that the DA precursor, levodopa, as well as other prodopaminergic drugs, could profoundly benefit the lives of PD patients. Indeed, decades later, administration of levodopa remains the gold-standard treatment for PD patients.

While clearly groundbreaking, the reserpine model has significant limitations. Most obviously, reserpine

produces only biochemical lesions, and therefore reserpine-treated animals do not model the neuronal dysfunction and degeneration seen in the human disease. In addition, the behavioral effects of reserpine are transient. While this can be advantageous in some circumstances in understanding nigrostriatal function, PD is a progressive chronic disorder, and if possible, progressive functional deficits are most often required for study. Last, reserpine depletes all monoamines, not exclusively DA. While norepinephrine depletion from the locus coeruleus is a facet of PD degeneration, depletion of all monoamines does not accurately model PD.

Nigral Cell Loss

While pharmacological models helped elucidate the role of DA in PD, their inability to model the histopathology of the disease made them poor models for evaluating many novel therapies. More beneficial have been models that aim to replicate the loss of dopaminergic nigral neurons observed in Parkinson's patients. Recognized as a primary neuropathology responsible for PD, SNpc cellular loss has been the prime focus of neurotoxic models that aim to replicate this cellular pathology.

6-Hydroxydopamine

The DA analog 6-hydroxydopamine (6-OHDA), because of its similarity in molecular structure, can be taken up into dopaminergic terminals via the DA transporter. Once inside the cell, it is metabolized, resulting in the production of hydrogen peroxide and free radicals. Ultimately these toxic molecules induce neuronal death via mitochondrial dysfunction.

Like DA itself, 6-OHDA is not able to cross the blood–brain barrier and therefore must be delivered directly to the brain of experimental animals via stereotaxic surgery. In 1968, Ungerstedt and colleagues demonstrated the utility of 6-OHDA lesions as animal models of PD. In their study, 6-OHDA was unilaterally injected into the medial forebrain bundle, extensively depleting the nigrostriatal pathway on one side. Ungerstedt noted that lesioned animals rotated toward the side of their lesions spontaneously as well as after administration of the dopaminergic drug d-amphetamine. Conversely, apomorphine, a drug that acts upon upregulated DA receptors on the side ipsilateral to the lesion, induces rotations contralateral to the lesion.

The number of rotations performed by a lesioned animal can be quantified to serve as an index of the integrity of nigrostriatal function. Experimental therapeutic strategies, such as neural or stem cell transplantation and gene therapy, can use the number of rotations an animal performs as an index of the intervention's efficacy.

Since these initial observations, there have been modifications and improvements on the 6-OHDA model although it is still frequently used in its original form. By varying the position and extent of the lesion, different stages of human PD can be modeled. The late stages of PD have been modeled using 6-OHDA lesions to the medial forebrain bundle and SN (**Figure 1**). Using this model, neuronal loss is detected as soon as 12 h postinjection and peaks at 48 h. In addition, striatal fibers are found to degenerate between 1 and 7 days after 6-OHDA delivery, ultimately resulting in more than 90% striatal DA depletion. This provides an ideal environment to evaluate cellular replacement strategies. Alternatively, 6-OHDA delivery to the striatum can result in levels of DA depletion more representative of early-stage PD. Kirik and colleagues demonstrated the location of striatal injections, either 'terminal' (within the caudate–putamen) or 'preterminal' (at the caudate–putamen boundary), greatly affected the resulting lesion, with preterminal injections creating greater levels of DA depletion. In addition, they found variable reductions in tyrosine hydroxylase-positive (TH+) fiber densities and TH+ SN neurons after either single or multiple 6-OHDA intrastriatal injections. As injection site number

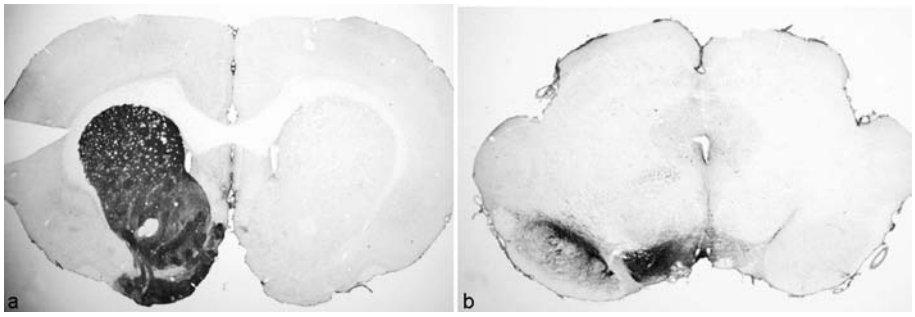

Figure 1 Tyrosine hydroxylase (TH)-stained coronal sections of a 6-OHDA-treated rat striatum (a) and substantia nigra (b) in which the dopaminergic neurotoxin was injected into the right medial forebrain bundle. Note the comprehensive loss of TH-immunoreactive fibers staining in the striatum (a) and the loss of TH-immunoreactive neurons in the nigra ipsilateral to the lesion.

increased, so too did DA depletion. A single preterminal injection site elicited a 60–70% decrease in TH+ fiber density in the rostral striatum and a 25% decrease in TH+ fiber density in the caudal striatum. In contrast, four preterminal site injections resulted in decreases as high as 80% and 90% in TH+ fiber density in the rostral and caudal striatum, respectively. These intrastriatal 6-OHDA models that provide varying degrees of DA depletion have been particularly effective in the evaluation of neuroprotective strategies for human PD.

The 6-OHDA lesion produces a range of functional deficits that have been assessed by means of a number of behavioral tests. Rotational analyses in response to prodopaminergic drugs such as amphetamine or agonists such as apomorphine have been used extensively to assess the extent of an animal's lesion and of its subsequent recovery after therapeutic behavioral analyzes. Nonpharmacological interventions such as open field activity in which a rat is observed for a time within an activity-monitoring chamber are also altered after 6-OHDA. Analysis of the animal's movements can provide a researcher with information about the animal's overall activity, rearing behavior, and any stereotypic behavior. The cylinder test (or limb use asymmetry test) consists of rating the use (and disuse) of each forelimb while an animal navigates itself around a plexiglass cylinder. While an unlesioned rat will typically use each limb equally, a rat with a 6-OHDA lesion will preferentially use the limb ipsilateral to its lesion. The adjusting steps test and its variations involve the restriction of an animal's hindpaws and one forepaw. The unbound forepaw is then moved slowly sideways across a flat surface. In an unlesioned rat, both forepaws will take steps to adjust to the sideways motion. However, in the lesioned animal, the forepaw contralateral to the lesioned striatum will drag along without adjustment. These and many other behavioral tests have been instrumental in the assessment of 6-OHDA lesion-induced deficits and in determining the functional efficacy of therapeutic interventions.

To date, no model of PD has been used as often or, arguably, as effectively as the 6-OHDA model. However, its limitations must also be acknowledged and considered when one is interpreting data obtained from its use. As with any transmitter-specific model, the 6-OHDA lesion does not replicate the constellation of neurotransmitter-based pathologies seen in human PD. The depletion of norepinephrine, due to degeneration of the locus coeruleus seen prior to SN degeneration in human PD patients, is not replicated in the 6-OHDA model. In addition, dysfunction of serotonergic and peptidergic systems seen in PD patients is not mimicked in 6-OHDA-treated animals. Also problematic is the timing of the degenerative process. Whereas Parkinson's is now thought to be a progressive degenerative disorder beginning in the olfactory bulbs and caudal brain stem and progressing over many years rostrally to the midbrain, 6-OHDA lesions are regionally selective and exhibit cell loss in a matter of days to weeks. Finally, while lesioned animals show some of the cellular and behavioral deficits seen in the human disease, 6-OHDA-treated animals fail to develop Lewy bodies or α-synuclein aggregates, features that are pathoneumonic for PD. While 6-OHDA remains a convenient and effective model for the analysis of the nigrostriatal dysfunction and evaluation of many novel therapies, these important shortcomings must be taken into account when translating information obtained from the 6-OHDA-treated animal from the laboratory to the clinic.

1-Methyl-4-Phenyl-1,2,3,6-Tetrahydropyridine

One of the best models of PD results from the administration of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP), which selectively targets dopaminergic neurons in the brain. Its validity as a model is supported by the fact that MPTP is a definitive, albeit rare, environmental cause of PD in humans. Its neurotoxic effects were first reported in the early 1980s by Bill Langston and co-workers after drug users exposed to MPTP began to develop Parkinsonian symptoms. Taken up primarily by astrocytes, MPTP is converted to its metabolite MPP+. MPP+ can then be taken up by DA neurons, where it exerts its toxic effect through interactions with cytosolic proteins and through the interruption of the complex I component of the electron transport chain.

The MPTP mouse model Due to the relative economy and small size of mice, the cost and convenience of using them have made the MPTP mouse model a widely used model of PD. In addition, the ability to genetically modify the mouse genotype has allowed for the analysis of possible genetic factors that may contribute to MPTP sensitivity, a procedure presently not possible in primate models of PD. In this regard, transgenic α-synuclein knockout mice show an increased resistance to MPTP toxicity, while transgenic mice overexpressing α-synuclein show increased susceptibility.

As with 6-OHDA, varying regimens of MPTP delivery can produce varying degrees of nigrostriatal dysfunction. Acute dosing, consisting of a single injection or a short series of high dose injections ($\sim$10–40 mg kg^{-1}), can produce lesions that result in mild to moderate cell loss and adequately represent early stages in human PD. An early study of acute

MPTP toxicity found a 33% reduction of TH+ SN neurons and a 90% depletion in striatal DA in the C57/bl mouse after a single acute injection of MPTP (40 mg kg^{-1}). Alternatively, chronic dosing, consisting of long series of low-dose injections (~5–20 mg kg^{-1}), results in more-robust lesions representative of late-stage human PD. Jackson-Lewis and colleagues have demonstrated as much as 70% nigral cell loss and 90% striatal DA depletion in C57/BL6 mice after a chronic MPTP dosing regimen (80 mg kg^{-1} total). Chronic administration is also associated with lower levels of spontaneous recovery. Research has suggested that neuronal loss and DA depletion incurred after MTPT administration occur through different mechanisms, with acute dosing resulting in necrotic cell loss and chronic dosing leading to apoptotic cell loss. It has also been shown that MPTP toxicity in the mouse can be enhanced through the delivery of MPTP-enhancing agents such as probenecid. In one study, the chronic delivery of MPTP (25 mg kg^{-1}; 10 doses over 5 weeks) with probenecid (250 mg kg^{-1}) to C57/bl mice resulted in the accumulation of α-synuclein immunoreactive aggregations, although these were not phenotypically similar to the Lewy bodies observed in human PD.

A number of behavioral tests have been developed to assess deficits and recovery of function in MPTP-treated mice. As with the 6-OHDA-treated rat, the open-field test has been used extensively in the mouse to assess overall behavioral activity. The rotarod test has also been used to assess the coordination and balance of MPTP-treated and control mice. In this test, mice are taught, prior to MPTP administration, to run on a rod rotating at varying speeds. The time that lesioned mice are able to remain on the rod then decreases with MPTP-induced nigrostriatal dysfunction. MPTP also impairs mouse performance on the pole test, a measure of bradykinesia that involves placing a mouse, head facing upwards, at the top of a textured pole. The amounts of time it takes the mouse to rotate its body toward the ground and to descend the pole have been taken as indices of the mouse's functional impairment.

While the MPTP mouse model has been invaluable in an understanding of the mechanisms of MPTP toxicity, there are some limitations to its uses. Due to the small size of mice, MPTP is typically administered to them systemically. While systemic delivery has the advantage of producing a bilateral lesion more representative of human PD, it is also associated with higher levels of animal morbidity and increases the animal maintenance required of caregivers. In addition, mice may show robust and spontaneous improvement after the initial lesions, making determinations of therapeutic efficacy difficult. These factors, combined with the mouse's insensitivity to MPTP relative to primates, result in the need for high doses to create stable lesions of clinical relevance in the mouse. Also problematic is the variability of MPTP sensitivity that exists between mouse strains. Muthane and colleagues have found that C57/bl mice innately possess fewer TH+ SN cells and furthermore show an increased sensitivity to the toxic effects of MPTP when compared with the CD-1 strain of mouse. Finally, while MPTP-treated mice show some occurrence of motor dysfunction, they have not been particularly useful in the analysis of parkinsonian behavior and recovery after therapeutic intervention.

The MPTP primate model Nonhuman primates are much more sensitive to MPTP toxicity and require smaller doses to create adequate stable nigrostriatal lesions. MPTP has been used experimentally in monkeys to create both bilateral and unilateral lesions. Systemic administration delivered via intraperitoneal, subcutaneous, intravenous, or intramuscular injections creates bilateral lesions that successfully model the cardinal symptoms associated with human PD. However, because these delivery methods are associated with increased animal morbidity and maintenance, researchers have also used primates that receive unilateral MPTP lesions delivered via intracarotid artery (ICA) injection (Figure 2). Despite the necessity of a more invasive surgical procedure, ICA lesions can be more attractive to researchers. Also, despite a small degree of damage to the contralateral hemisphere, it remains relatively intact in unilaterally injected animals and may serve as a useful control and allow for the occurrence of rotational behavior.

In an attempt to combine the benefits of ICA and systemic MPTP treatment and to avoid the weaknesses of each delivery method, recent studies have created an 'overlesioned' model in which an initial ICA unilateral injection is followed by subsequent intravenous injections. This approach elicits an asymmetrical bilateral lesion. In one study looking at the effects of overlesioning in rhesus monkeys, a comprehensive cell loss was seen in the SN ipsilateral to the ICA injection, with a subtotal cell loss in the contralateral SN. Researchers noted a positive correlation between the animals' behavioral deficits and the extent of their lesion contralateral to the ICA injection. The preferential degeneration of one hemisphere over another produced by the overlesioning method is much like the hemispheric biases seen in human PD patients. It is important to note that the deficits in these animals are bilateral and remain stable and that the animals are able to maintain themselves and do not require heroic veterinary intervention.

Again, the method of delivery and dosage of the MPTP toxin may be manipulated to create lesions representative of varying stages of human PD. Acute

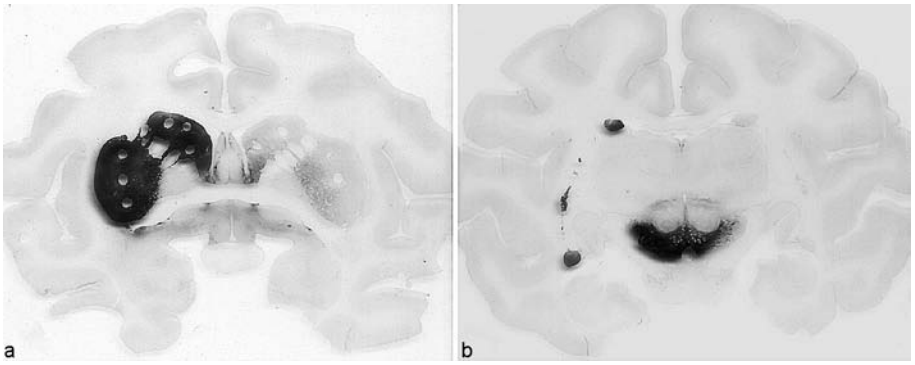

Figure 2 Tyrosine hydroxylase (TH)-stained coronal sections through the striatum (a) and substantia nigra (b) from a rhesus monkey that received an injection of MPTP in the right internal carotid artery. Note the comprehensive loss of TH-immunoreactive fibers staining in the striatum (a) and the loss of TH-immunoreactive neurons in the nigra ipsilateral to the lesion.

MPTP administration results in a significant nigral cell loss, modeling late-stage human PD. In contrast, chronic MPTP administration results in more-moderate cell loss, resembling the middle stages of human PD. In this regard, squirrel monkeys display a 70% nigral cell loss and 95% striatal DA depletion after acute MPTP treatment compared with 40% nigral cell loss and 60–70% striatal DA depletion after chronic systemic administration. Chronic MPTP administration may also be used to model the more progressive nature of cell loss and behavioral deficit observed in the human disease.

As with idiopathic PD, MPTP-induced deficits can be evaluated using clinical rating scales. These assess the severity of symptoms such as tremor, posture, gait, bradykinesia, balance, defense reactions, freezing, and gross motor skills. Fine motor skills testing can also be analyzed through a number of behavioral tests. The forelimb reaching test involves training the monkey to reach toward a recurring target object on which a reward is given, usually food. The time it takes for the monkey to reach toward its reward is an index of its nigrostriatal function. A similar test is the bar-pressing test, in which monkeys are made to press a lever bar to receive a food reward. Tests have also been developed to assess the cognitive functioning of MPTP-treated monkeys. In the barrier–detour retrieval test, monkeys are made to retrieve a food reward from a transparent box with an opening on one side. The monkey must determine which side is the open side of the box before receiving the reward. Both the speed at which the monkey can attain the reward and the number of errors that occur for it to do so can be taken as indices of cognitive functioning.

The effectiveness of MPTP to create lesions of the nigrostriatal pathway can be remarkably variable. One cause for this variability is the age of the animal exposed to the toxin, with older animals showing a much greater sensitivity to MPTP toxicity than their younger counterparts do. This finding may be explained by older individuals' increased levels of MAO-B, the enzyme required for the conversion of MPTP to its toxic metabolite, MPP+. In addition, studies have shown a higher sensitivity to MPTP toxicity in males than in females, suggesting a possible protective effect of estrogen on dopaminergic neurons. Animal species is also a critical predictor of MPTP susceptibility, with some species, such as the nonhuman primate, showing a high sensitivity to MPTP toxicity and others, like the rat, showing little to no effect.

MPTP has done a remarkably better job than 6-OHDA in reproducing many of the nondopaminergic deficits observed in human PD. MPTP-treated monkeys show depletions in norepinephrine and serotonin metabolites and after chronic exposure, cell loss in the locus coeruleus similar to findings in human patients. While true Lewy body formations are not seen in MPTP models of PD, accumulations of α-synuclein protein have been noted in the brains of baboons and squirrel monkeys after MPTP administration. MPTP treatment has been shown to elicit a resting tremor akin to the cardinal symptom manifested in human patients. Resting tremor has been observed in the MPTP-treated vervet monkey although it has not been reported in any other primate models. While the above cases demonstrate that protein aggregation and tremor are possible after MPTP exposure, it should be stressed that these important Parkinson's pathologies are not consistently found to occur within this model.

Environmental Toxin Models

The success of MPTP in replicating the neuropathology of human PD has led researchers to examine other neurotoxins that may similarly affect the central nervous system. Two promising models to emerge

have been the rotenone and combined paraquat and maneb models.

Rotenone Rotenone, an organic pesticide that interrupts complex I of the electron transport chain, also elicits mitochondrial dysfunction and ultimately cell loss in the nigrostriatal pathway. Rotenone may be delivered systemically to create a bilateral lesion or via intranigral infusion to create a unilateral lesion. Like other neurotoxic models, bilateral lesions, though similar to the human disease, frequently result in high animal morbidity, leading researchers to occasionally favor unilateral surgeries. After rotenone exposure, rats can show progressive nigral degeneration, ubiquitin and α-synuclein inclusions in remaining nigral cells, and behavioral deficits similar to those observed in human patients. Because rotenone exerts its effects in a DA transporter (DAT)-independent fashion, it elicits mitochondrial dysfunction in non-dopaminergic systems, although its neurotoxic effects are DA neuron selective.

While rotenone can accurately model human nigrostriatal degeneration in PD, the great variability in individual sensitivity to its effects has hindered its use in research. In one study, only 46% of animals treated with rotenone ($2\,mg\,kg^{-1}$ daily for 21 days) showed decreased TH reactivity, with only one animal showing any nigral cell loss. Recently another study has found that while C57/bl mice displayed some functional dysfunction after chronic rotenone treatment, they failed to develop the nigral degeneration associated with the rotenone rat model. This substantial variability among individuals and species is a major obstacle for the use of the rotenone model for the evaluation of therapies for PD.

Paraquat and maneb Paraquat (PQ; 1,1′-dimethyl-4-4′-bipyridinium) is another pesticide that has been examined as a possible neurotoxic animal model for PD. Paraquat is initially converted to a PQ cation, which is then reoxidized to from a parent compound as well as superoxide radicals. This allows for further redox cycling and further oxidative stress. Paraquat is often administered in combination with the fungicide manganese ethylene-bis-dithiocarbamate (maneb). Exposure to maneb has been linked to the development of parkinsonian symptoms in humans, and its combination with paraquat as an animal model has been useful in the investigation of the role of environmental toxins in the etiology of human PD.

Rodent studies have found significant nigral cell loss after combination paraquat and maneb exposure. Mice receiving biweekly PQ ($10\,mg\,kg^{-1}$) and maneb ($30\,mg\,kg^{-1}$) injections had 25–47% TH+ SN cell loss. PQ-exposed mice have also been shown to develop α-synuclein inclusions in remaining nigral neurons. Paraquat and maneb have also been used extensively in research looking at effects of gestational exposure to toxins. Researchers have found that mice exposed to maneb during gestation show increased sensitivity to PQ toxicity in adulthood, providing support for the 'multihit' hypothesis of PD and suggesting that an initial early insult can sensitize the nigrostriatal system to subsequent 'hits' in adulthood.

Protein Aggregation

The presence of Lewy bodies, α-synuclein-rich intracellular inclusions, in the brains of parkinsonian patients has suggested a role for protein aggregation in the pathogenesis of PD. This has been further supported by the discovery that genetic mutations in the α-synuclein gene can cause genetic forms of PD. In addition, increased levels of α-synuclein in SN neurons have been observed in response to toxin exposure and oxidative stress, suggesting that protein aggregation may have a pivotal importance in the development of idiopathic PD as well. Despite evidence to suggest the central role of protein aggregation in PD development, animal models have traditionally been very poor at exhibiting this important pathology. However, with the development of transgenic animals and the advancement of viral vector delivery systems have come new models for this previously overlooked pathology.

Transgenic Animals

Transgenic models of PD have been emerging that attempt to recreate the known genetic causes of PD. Particularly interesting have been models that knock out, overexpress, and mutate the α-synuclein gene to examine the role of protein aggregation in PD neurodegeneration. In addition, researchers have looked to mutate proteins involved in the ubiquitin–proteosome pathway known to be affected in familial cases to examine the role of proteosome dysfunction in PD.

Alpha-synuclein transgenic models Discovery of the A53T and A30P mutations of the α-synuclein gene in familial PD have led researchers to attempt to manipulate its expression in animals in order to evaluate its role in the human disease. Knockout mice are viable and show decreased striatal DA levels and reduced rearing in the open field, despite DA metabolism's being unaffected. These findings may be explained by the hypothesis that α-synuclein plays a role in synaptic vesicle function. Conversely, mice overexpressing α-synuclein develop intraneuronal inclusions and show decreases in striatal DA levels,

despite showing no nigral cell loss. These effects are generally directly correlated to the amount of α-synuclein overexpressed. Mice expressing mutated forms of α-synuclein have shown varying degrees of pathology, dependent on the forms of construct utilized. Although transgenic animals expressing human α-synuclein with the platelet-derived growth factor β promoter show behavioral deficits and increased SN inclusions, other constructs have shown very little pathology or behavioral deficit.

A major strength of the transgenic α-synuclein models for PD is that expression alterations are not limited to dopaminergic neurons but are expressed throughout the nervous system. In addition, the slow accumulation of protein that occurs in transgenic animals provides a better model of the progression seen in human disease. However, despite these models; ability to replicate the Lewy body inclusions seen in human PD, a major failing is their inability to model SNpc cell loss, making them suboptimal for the evaluation of many novel therapies for PD. In addition, technical restrictions limit the use of transgenic models to the phylogenetically distant mouse, preventing these models' immediate application to the clinic.

Parkin and ubiquitin C-terminal hydrolase L1 transgenic models The discovery of mutations to parkin and ubiquitin C-terminal hydrolase L1 genes, both components of the ubiquitin–proteosome system, in familial cases of PD has led researchers to examine the role of proteosome dysfunction in PD and to develop models that target these genes.

Viral Vector-Delivered α-Synuclein

While transgenic models have been limited to mice, adeno-associated (AAV) and lentiviral delivery systems have enabled the selective overexpression of α-synuclein within the nervous system of higher-ordered species. In rats, delivery of wild-type or mutant AAV-human α-synuclein to the SNpc resulted in cytoplasmic inclusions, behavioral dysfunction, and nigral cell loss. Similar cellular pathology was found in rats receiving lentiviral-delivered wild-type and mutant forms of human α-synuclein. The overexpression of α-synuclein via viral vectors has also been evaluated in the nonhuman primate. Marmosets receiving mutant AAV-human α-synuclein showed a 32–41% reduction of DA neurons in the SN. Behavioral deficits observed included head positional bias and rotation asymmetry, both of which are commonly seen in PD toxicity models.

The ability to overexpress α-synuclein has given researchers the opportunity to evaluate the role of protein aggregation in PD. In addition, it offers a relatively slowly progressing model with which to analyze the gradually progressive nature of the human disease, an advantage not offered by most toxicity models. While it is a seemingly powerful model that can encompass the majority of PD pathologies, the selective overexpression of α-synuclein to one specific brain region limits the observation of nonmotor effects such as olfactory impairments, gastrointestinal dysfunction, depression, sleep disturbances, and cognitive impairments that result from α-synuclein aggregation in other brain regions.

Other PD Models

Lipopolysaccharide

The immune response has been suggested as a major contributor to the progressive nature of nigral cell loss in PD. Researchers have tried to replicate this increased immune activation through exposure to the bacteriotoxin lipopolysaccharide (LPS). Stereotaxic injections of LPS have been used to induce nigral cell loss in the adult animals, and prenatal LPS exposure has been shown to elicit decreased DA neuron number in offspring. In addition, these animals show locus coeruleus cell loss, suggesting nondopaminergic effects, as well as intracellular inclusions. Researchers have also found that prenatal LPS treatment further prolongs the immune response to secondary LPS exposure in adulthood, suggesting a role for neuroinflammation in the 'multihit' hypothesis of PD.

Aged Animals

Aging is the primary risk factor for the development of PD. In normal aging, striatal DA is lost, and there is a loss of TH immunoreactive and DA transporter immunoreactive nigral neurons. These losses are associated with age-related increases in α-synuclein. While controversial, emerging evidence suggests that PD represents the extreme end on a continuum of aging. Some experimenters therefore have utilized aged animals themselves as models for PD. Taken together, these findings suggest that aged animals may represent a useful model for evaluation of therapies aimed at the earliest stages of PD.

Animal Models of Dyskinesia

In addition to mimicking the pathology of PD, models have been used to study treatment-related side effects of PD. In this regard, the laboratory of Angela Cenci has led the field in demonstrating the use of the 6-OHDA rat model of PD for the study

of levodopa-induced dyskinesias. Dyskinesias are disabling, involuntary movements that commonly arise in PD patients after prolonged levodopa therapy. While dyskinesias can take several years to develop in patients, rats may be primed with levodopa to exhibit abnormal involuntary movements in a matter of weeks. The affordability and convenience of the rat model make it ideal for the evaluation of the molecular and cellular causes of this serious side effect.

While demanding much more cost and care, MPTP-treated monkeys have also been one of the best and most widely used animal models in which to study levodopa-induced dyskinesias because of their phylogenetic proximity to humans. Unlike the 6-OHDA-treated rat, the MPTP-treated primate model shows similarities to human anatomy and physiology.

Conclusions

The modeling of PD in animals has been invaluable to our understanding of the human disease. By selectively targeting certain aspects of PD, researchers have been able to evaluate potential therapies aimed at specific pathologies. Pharmacological models, such as reserpine-treated animals, have provided a better understanding of the role of DA in PD. Through its experimental use, the reserpine model has served to evaluate DA replacement strategies such as levodopa treatment, still routinely used effectively to alleviate parkinsonian symptoms. Neurotoxin models, such as the 6-OHDA and MPTP models, replicate the nigral cell loss seen in PD patients and have been used successfully to evaluate neuroprotective strategies and novel therapies such as cell transplantation. A new wave of neurotoxic models, such as the rotenone and paraquat and maneb models, have been invaluable in the investigation of the role of environmental toxins in the etiology of PD.

While the experimental value of the models cannot be understated, they have so far failed to provide a comprehensive model of the human disease. Aspects of human PD that have been particularly poorly modeled by traditional models are Lewy body formation, pathologies of nondopaminergic systems, and the progressive nature of the disease. Novel genetic models mimic the previously unexamined pathology of protein aggregation in PD, allowing for the investigation of its role in PD development. LPS exposure provides an excellent model to observe the role of neuroinflammation in PD and the progressive loss of dopaminergic cells. Finally, aged animals may represent a truly global model for analyzing the very early stages of PD.

See also: Aging of the Brain; Cell Replacement Therapy: Parkinson's Disease; Deep Brain Stimulation and Parkinson's Disease; Dopamine; Dopamine in Perspective; Dopaminergic Agonists and L-DOPA; Parkinsonian Syndromes; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Transgenic Models of Neurodegenerative Disease.

Further Reading

Braak H, Del Tredici K, Rub U, de Vos RA, Jansen Steur EN, and Braak E (2003) Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of Aging* 24(2): 197–211.

Carlsson A (1959) The occurrence, distribution and physiological role of catecholamines in the nervous system. *Pharmacology Review* 11: 490–493.

Chu Y and Kordower JH (2007) Age-associated increases of alpha-synuclein in monkeys and humans are associated with nigrostriatal dopamine depletion: Is this the target for Parkinson's disease? *Neurobiology of Disease* 25(1): 134–149.

Fleming SM, Zhu C, Fernagut PO, et al. (2004) Behavioral and immunohistochemical effects of chronic intravenous and subcutaneous infusions of varying doses of rotenone. *Experimental Neurology* 187(2): 418–429.

Herrera AJ, Castano A, Venero JL, Cano J, and Machado A (2000) A single intranigral injection of LPS as a new model for studying the selective effects of inflammatory reactions on the dopaminergic system. *Neurobiology of Disease* 7(4): 429–447.

Jackson-Lewis V, Jakowec M, Burke RE, and Przedborski S (1995) Time course and morphology of dopaminergic neuronal death caused by the neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydorpyridine. *Neurodegeneration* 4(3): 257–269.

Jellinger KA (1991) Pathology of Parkinson's disease: Changes other than the nigrostriatal pathway. *Molecular Chemical Neuropathology* 14: 153–197.

Jeon BS, Jackson-Lewis V, and Burke RE (1995) 6-Hydroxydopamine lesion of the rat substantia nigra: Time course and morphology of cell death. *Neurodegeneration* 4: 131–137.

Kirik D, Annett L, Burger C, Muzycka N, Mandel R, and Bjorklund A (2003) Nigrostriatal α-synucleinopathy induced by viral vector-mediated overexpression of human α-synuclein: A new primate model of Parkinson's disease. *Proceedings of the National Academy of Sciences of the United States of America* 100(5): 2884–2889.

Kirik D, Rosenblad C, and Bjorklund A (1998) Characterization of behavioral and neurodegenerative changes following partial lesions of the nigrostriatal dopamine system induced by intrastriatal 6-hydroxydopamine in the rat. *Experimental Neurology* 152: 259–277.

Langston JW, Ballard P, Tetrud JW, and Irwin I (1983) Chronic parkinsonism in humans due to a product of meperidine-analog synthesis. *Science* 219: 979–980.

Oiwa Y, Eberling JL, Nagy D, Pivirotto P, Emborg ME, and Bankiewicz KS (2003) Overlesioned hemiparkinsonian non human primate model: Correlation between clinical, neurochemical and histochemical changes. *Frontiers in Bioscience* 8: a155–a166.

Olanow CW and Tatton WG (1999) Etiology and pathogenesis of Parkinson's disease. *Annual Review of Neuroscience* 22: 123–144.

Przedborski S, Jackson-Lewis V, Naini AB, et al. (2001) The parkinsonian toxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP): A technical review of its utility and safety. *Journal of Neurochemistry* 76: 1265–1274.

Przedborski S, Levivier M, Jiang H, et al. (1995) Dose-dependent lesions of the dopaminergic nigrostriatal pathway induced by intrastriatal injection of 6-hydroxydopamine. *Neuroscience* 67: 631–647.

Przedborski S and Vila M (2003) The 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine mouse model: A tool to explore the pathogenesis of Parkinson's disease. *Annals of New York Academy of Science* 991: 189–198.

Schultz W (1988) MPTP-induced parkinsonism in monkeys: Mechanism of action, selectivity and pathophysiology. *General Pharmacology* 19: 153–161.

Thiruchelvam M, Brockel BJ, Richfield EK, Baggs RB, and Cory-Slechta DA (2000) Potentiated and preferential effects of combined paraquat and maneb on nigrostriatal dopamine systems: Environmental risk factors for Parkinson's disease. *Brain Research* 873(2): 225–234.

Ungerstedt U (1968) 6-Hydroxydopamine induced degeneration of central monoamine neurons. *European Journal of Pharmacology* 5: 107–110.

van der Putten H, Wiederhold KH, Probst A, et al. (2000) Neuropathology in mice expressing human alpha-synuclein. *Journal of Neuroscience* 20: 6021–6029.

Animal Models of Stroke

L C Hoyte and A M Buchan, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Among western countries, such as the United States, Canada, and the United Kingdom, stroke is the third leading cause of death and a major source of disability. The risk of having a stroke or a transient ischemic attack increases dramatically with age. These facts, combined with the current shift in age demographics toward a more elderly population in western countries, ensure that stroke will continue to be a health care burden. For example, in the United States, the estimated cost of stroke (from direct and indirect sources) in 2006 was predicted to be $57.9 billion (American Heart Association). It is vital to reduce brain damage and improve quality of life for stroke patients.

Human stroke occurs in two forms: ischemia and hemorrhage. The ischemic stroke results from an occlusive clot, this being from either a thrombus or an embolus that blocks blood flow within an artery of the brain. During a focal ischemic insult, there is a difference in cerebral blood flow (CBF) to the brain depending on the proximity to the clot and the level of collateral flow. The area that is within the main occluded territory is generally termed the core of the infarct, while the surrounding tissue is termed the penumbra. Within the core infarct there is extensive tissue damage, with pannecrosis, gliosis, inflammatory infiltration, breakdown of the blood–brain barrier, cytotoxic and vasogenic edema, and excitotoxic cell death. The penumbra is the area of the brain that has reduced CBF and metabolism, but damage is fundamentally reversible with reperfusion. Hemorrhagic stroke is due to a cerebral vessel bursting and the resultant bleeding into the brain. Hemorrhagic stroke can be further subdivided into two main categories: intracerebral hemorrhage (bleeding from cranial blood vessels into the brain tissue) and subarachnoid hemorrhage (bleeding from cranial blood vessels into the subarachnoid space).

A vast majority of human strokes are ischemic in nature; 85% of all strokes in western countries are ischemic, while 15% are hemorrhagic, however, there is a shift to increased hemorrhage rate among Asian populations. Due to the fact that most strokes are ischemic, experimental stroke research primarily attempts to discover ways to reduce morbidity and mortality among the aging population.

Can Animal Models Mimic Human Stroke?

Most experimental stroke modeling is conducted using rodent models, with surgical procedures used to induce stroke. The rodent brain has a smooth cortex (lissencephalic), and different components of white and gray matter as compared to the human brain. As well, the rodent and human immune systems have marked differences. For example, a prevalence of infiltrating neutrophils is seen in rats following stroke, while in humans generally there is predominantly a macrophage infiltration. However, many rodent species, with the gerbil being a notable exception, possess a complete circle of Willis, similar to the human circulation.

This similarity in circulation allows for experimenters to examine stroke and CBF in the rodent brain and relate it to the human condition. Fundamental brain regions are similar from animal to human, allowing for insight into stroke recovery and plastic adaptation of the brain. However, there has been a plethora of failed human clinical trials; the animal work frequently did not translate to human stroke treatment in the past. Particularly, it has been seen that many experimental results were confounded by alterations in animal physiology; such that the improved outcome was due to an innate physiological response, and was not due to the specific pharmacological action of the drug on its original target. A classic example of such a phenomenon is the use of the *N*-methyl-D-aspartate (NMDA) noncompetitive receptor blocker MK-801. Treatment with this compound was found to be highly efficacious in reducing infarct volumes in the rat cerebral ischemia, but it was later proved that this drug caused a persistent hypothermia in the global model and that an improved collateral flow to the brain in the middle cerebral artery (MCA) occlusion models was responsible for improved outcome. This compound was moved to human clinical trial, without success, and there have been many more, similar examples since. This has led to some despair among researchers that animal modeling will not advance the treatment of human stroke.

However, there has been evidence of successful translation of animal work to the clinic. The only US Food and Drug Administration (FDA)-approved drug for the treatment of acute cerebral ischemia in the human, tissue plasminogen activator (tPA), was first shown to be efficacious in a rabbit model of small clot emboli. Currently there is clinical testing on another compound (NXY-059, a free radical spin trap agent) that shows positive experimental results and has also shown some promise in the clinic.

Cerebral Ischemia Models: Global versus Focal Ischemia

Cerebral ischemia accounts for approximately 85% of the strokes seen in western society. As such, these are extensively studied experimentally. The two forms of cerebral ischemia are global (either complete or incomplete) and focal (either multifocal or focal) (Figure 1).

Global Cerebral Ischemia

Global cerebral ischemia is used to mimic the clinical situation of cardiac arrest, and is also widely used to examine the selective vulnerability of neurons to cerebral ischemia. It is thought that any treatment that prevents the loss of these highly susceptible neurons might be well suited to act as a neuroprotective agent. Complete global ischemia, which can be produced by decapitation, cardiac arrest, neck cuff, or aortic occlusion, results in a loss of blood flow to the entire brain. Incomplete global ischemia; also called transient, severe forebrain ischemia, results in oligemia throughout the forebrain, including the cortex, striatum, and hippocampus. The blood flow to the brain stem is preserved in this model, allowing for the maintenance of heart rate and breathing by the animal, and possibly sparing the systemic effects of such complete global ischemia models as cardiac arrest.

Complete and incomplete global ischemia models are extremely severe, with very exceptional drops in cerebral blood flow. Consequently, the length of the ischemia in these models is generally very short in comparison with focal models. Global ischemia results in a characteristic cell death, without infarct, but with selective neuronal death. Susceptible neurons, such as the CA1 pyramidal cells of the hippocampus and select neurons of the striatum, and cortical layer III, V, and VI, die when exposed to the insult. The complexity arises from this model when one considers that adjacent to the CA1 pyramidal cells in the hippocampus are the CA3 cells – neurons that seem to have similar characteristics in location and cell type, but are resistant to the insult. It has been postulated that the CA3 are more robust than the CA1 due to a difference in the incoming afferents to these areas of the hippocampus, and to differences in the regulation of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor subunit GluR2 following global ischemia.

Complete global ischemia The most commonly studied form of complete global ischemia is the cardiac arrest model. This model is used in both gyroencephalic (convoluted cortex) and lissencephalic (smooth cortex) animals. Frequently, in canine models, it is achieved with exsanguinations, followed by mechanical

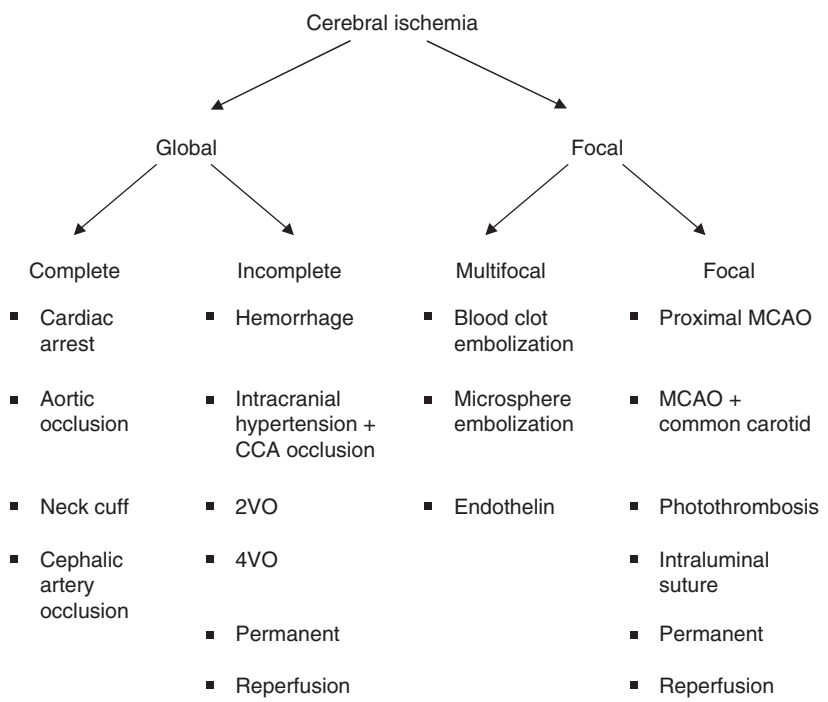

Figure 1 Methods used to generate ischemia in animal models. These include both focal and global models and permanent and reperfusion models. MCAO, middle cerebral artery occlusion; 2VO, two-vessel occlusion; 4VO, four-vessel occlusion. Adapted from Traystman RJ (2003) Animal models of focal and global cerebral ischemia. *ILAR Journal* 44: 85–95 (p. 87). Reprinted with permission from ILAR Journal, Institute for Laboratory Animal Research, National Academy of Sciences, 500 Fifth Street NW, Washington, DC, 2001.

or chemical resuscitation after the specified arrest time. Arrest of the heart can also be achieved by electrical fibrillation, in the cat, with consequent resuscitation. In rodents, the cardiac arrest model is frequently performed using KCl to arrest the heart, with resuscitation. Cardiac arrest models can be used to examine the efficacy of rapid induction of cerebral hypothermia, by aortic or jugular flushing with ice-cold saline. This can rapidly reduce brain temperature and improve outcome in dogs, even following very severe ischemia. The main drawback of the cardiac arrest models is that they generate systemic ischemia, which may, in turn, make interpretation of the cerebral ischemia more challenging due to damage occurring to other body organs.

Incomplete global ischemia Incomplete cerebral ischemia results in transient, but severe, forebrain ischemia, followed by reperfusion. This avoids the systemic effects of cardiac arrest, while still achieving very low cerebral blood flows of less than 10% of baseline. The use of four-vessel occlusion (4VO) is widespread in rat experiments, and can be achieved by occlusion of both vertebral arteries as well as both common carotid arteries (Figure 2). This ischemic insult results in a selective neuronal death that is dependent on ischemia duration. Pyramidal cells of the CA1 region of the hippocampus are especially sensitive to global ischemia, and duration of ischemia has a direct effect on progression of CA1 cell death; shorter durations of ischemia result in slower progression of CA1 damage. The model of global forebrain ischemia creates a lack of blood flow to the hippocampus, striatum, and cortex, while sparing flow to the brain stem. This oligemia can be reversed after 5, 10, 15, or 30 min and a recovery of blood flow and ATP levels is observed immediately. The degree of oligemia is not influenced by infusion of drug during the insult; however, hypothermia, even 6 or 24 h postischemia, is successful in increasing the long-term survival of the sensitive CA1 neurons. One major caveat to the 4VO model is the necessity of a 2-day procedure to generate the ischemia. This model relies on a preparatory day, which constitutes a surgery to isolate and prepare the common carotid arteries (CCAs) for the following day's occlusion along with a cauterization of the vertebral arteries and placement of sutures to limit collateral flow. The animals are fasted overnight, and the following day are reanesthetized to allow for occlusion of the CCAs. Some groups wish to avoid this preparatory day, and to achieve ischemia on the first day. This has led to the development of two-vessel occlusion (2VO) with hypotension.

The process of 2VO with hypotension involves a bilateral occlusion of the common carotid arteries, along with a systemic hypotension by exsanguinations. This model must achieve a level of hypotension sufficient to generate the global ischemia – reduction of blood pressure to 50 mmHg is sufficient in the rat. This generates a similar pattern of selective neuronal death as is seen in the 4VO and the cardiac arrest models, and is highly sensitive to hypothermia. 2VO is also used in the mouse, frequently without systemic hypotension to generate global ischemia, but the duration of ischemia tends to be longer to

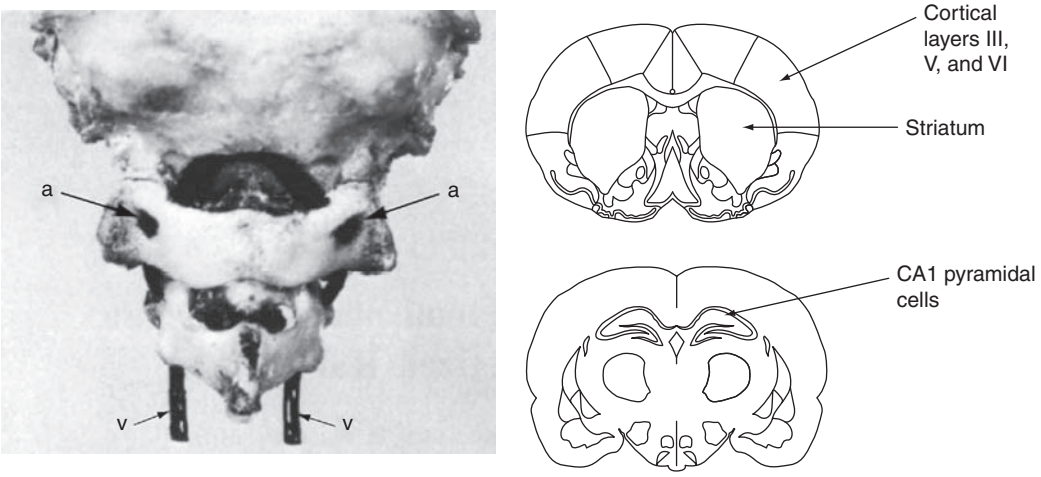

Figure 2 The four-vessel occlusion for incomplete cerebral ischemia in rats. Vertebral arteries (v) are cauterized at the alar foramina (a) in the preparatory surgery, the animals are fasted overnight, and the following day bilateral common carotid artery occlusion generates the transient, but severe, forebrain ischemia. CA1 pyramidal neurons are very susceptible to the ischemia, along with select neurons of the striatum and cortical layers III, V, and VI; this model shows selective neuronal death. Adapted from Pulsinelli WA and Brierley JB (1979) A new model of bilateral hemispheric ischemia in the unanesthetized rat. *Stroke* 10: 267–272.

compensate for the increased flow due to the posterior circulation.

Gerbils can be used to examine the effects of global ischemia, with the 2VO model, but without systemic hypotension. This is because, in general, gerbils lack a posterior communicating artery and do not have a complete circle of Willis. However, new evidence suggests the presence of a posterior communicating artery in some gerbils from specific suppliers, so experimenters must be careful in the assumption that bilateral CCA occlusion will generate global ischemia. The use of the gerbil model has verified the selective vulnerability of CA1 pyramidal cells, and has confirmed the protection of the CA1 by hypothermia. This protection has been shown to markedly improve survival of the CA1 for 18–20 months. One major advantage of this model is the relative ease of preparation and execution of the ischemia.

Currently, a new model for global cerebral ischemia has been developed in the mouse; this model mimics the effects of 4VO in the rat. The model involves the bilateral occlusion of the common carotid arteries as well as an occlusion of the basilar artery. It preserves blood flow to the brain stem while a transient oligemia is induced in the striatum, hippocampus, and cortex. The difficulties with induction of global ischemia in the mouse include an increased risk of respiratory failure due to manipulation of the trachea, which is necessary to access the basilar artery, and the possible breaking of the basilar artery following removal of the microaneurysm clip. The main advantage of generating this global model in mice is to study different transgenic strains following ischemia. This should allow for an in-depth examination of specific genes that might be implicated in selective neuronal death.

Focal Cerebral Ischemia

Focal cerebral ischemia is used to study the tissue damage that results from stroke. Generally, focal cerebral ischemia models examine the blockage of one artery of the brain (e.g., the MCA). Focal cerebral ischemia can also be generated with multiple focal insults, such as occur with embolization. These models result in classic infarcts with core and penumbra areas. The core of the infarct is the area that is closest to the occlusion; this area has the lowest cerebral perfusion and the tissue contained within this area is presumed to be unsalvageable. Seminal work from the late 1970s demonstrated that there was a difference in cerebral perfusion depending on the distance from the occlusion, and the area surrounding the core infarct was termed the ischemic penumbra. The neurons within the penumbral area are structurally-intact, but metabolically paralyzed. There is a direct relationship to the cerebral blood flow and the electrical activity of the brain, such that during cerebral ischemia, the electrical activity of the brain will disappear; however, this loss of electrical activity occurs at a less severe ischemia than is necessary to generate failure of the cell's baseline functions. Thus, the core infarct contains cells that have become structurally and metabolically compromised, while neurons in the penumbra retain structural integrity. The infarct loses the ability for autoregulation, and over time, develops into a glial scar.

Focal ischemia models are considered to have either a single focus (focal) or to have multiple foci (multifocal). The focal insult is generally created by surgical occlusion of an artery of the brain; however, there are cases whereby an artery was occluded by direct placement of clot into the artery. Multifocal insults are produced by injection of emboli (either clot or microspheres) into the brain. This form of insult results in multiple infarcts and is far more variable than focal ischemia.

Middle cerebral artery occlusion By far the most common form of focal ischemia is achieved by occlusion of the MCA. This was first attempted by Harvey and Rasmussen in 1951, when they surgically occluded the MCA with a clip to produce a large infarct and marked hemiparesis in dogs. The same two experimenters showed that the duration of ischemia was important for the development of infarct in the monkey – a 50 min ischemia resulted in an infarct similar to that of permanent ischemia, while ischemia durations of less than 30 min did not produce gross cortical damage. Currently, MCA occlusion is used to induce stroke in many animal species, including rats, mice, pigs, cats, dogs, and nonhuman primates. Rodents are by far the most commonly used experimental animal in MCA occlusion studies due to ethical and financial constraints of using higher order mammals. The occlusion of the MCA can be induced either proximally, such as in the case of intraluminal suture techniques, or more distally by a clipping of the distal MCA.

One of the more commonly used methods for generating MCA ischemia is the intraluminal suture technique. This method achieves ischemia by introduction of an occlusive suture into the internal carotid artery, whereby it is advanced to block the origin of the MCA. The carotid artery is accessed from the ventral cervical surface, and craniotomy is avoided in this model. The occlusion is generated by accessing and isolating the common carotid artery, external carotid artery (ECA), and internal carotid artery (ICA). Originally, the suture was passed directly into the ICA through a small puncture, while the arteries (CCA,

ICA, and ECA) were tied off to prevent bleeding. Further refinements were established to allow for introduction of the suture into the ECA and temporary placement of the suture in the CCA, followed by a movement of the suture into the ICA. This refinement necessitated the permanent ligation and cutting of the ECA, so that the artery could be moved, allowing for advancement of the suture into the ICA. Upon introduction of the suture into the ICA, it is advanced until it blocks the proximal MCA (Figure 3). Various small alterations in this technique exist. They include using different sizes of sutures to block the MCA with a heat-blunted tip, silicone coating of the suture tip, or the use of polylysine-coated sutures. The intraluminal suture model induces severe ischemia in the striatum with milder cortical ischemia. Commonly, the intraluminal suture model results in ischemia within the hypothalamus, and, in the rat, this results in hyperthermia. This hyperthermia can markedly enhance infarct progression and increases the variability of this model. Proximal MCA occlusion in the mouse is conducted in a manner similar to that in the rat, which again results in a severe ischemia of the striatum with milder ischemia in the cortex. The intraluminal mouse model can be variable due to the differences in patency of the posterior communicating artery among strains, and even among animals of the same strain. The possible use of transgenic animals is a great advantage of using the mouse model. However, differences in patency of posterior communicating arteries make the use of transgenic mice more difficult if the background strain used in not the same as the control strain.

Mice are more delicate than rats and frequently succumb to the stress of surgical procedures, and the survival rate of experimental mice is less than that of rats. Physiology of rodents must be carefully controlled, and following focal ischemia by intraluminal suture, mice tend to become hypothermic. This results in smaller infarct volumes and better histological and behavioral scores as compared to temperature-regulated mice.

In rats, direct placement of a clot into the proximal MCA has been shown to generate consistent ischemia, while introducing an occlusive thrombus. In this model, the clot is placed into the MCA by introducing tubing containing the clot into the ECA and advancing it through the ICA, in a manner similar to the placement of the suture in the intraluminal technique. However, the clot is ejected from the tubing, effectively obstructing the MCA. The main advantage of this model is the placement of the clot within the MCA. This should allow for testing of the efficacy of thrombolytic drugs on clot lysis in this model, which is not possible in the intraluminal suture models.

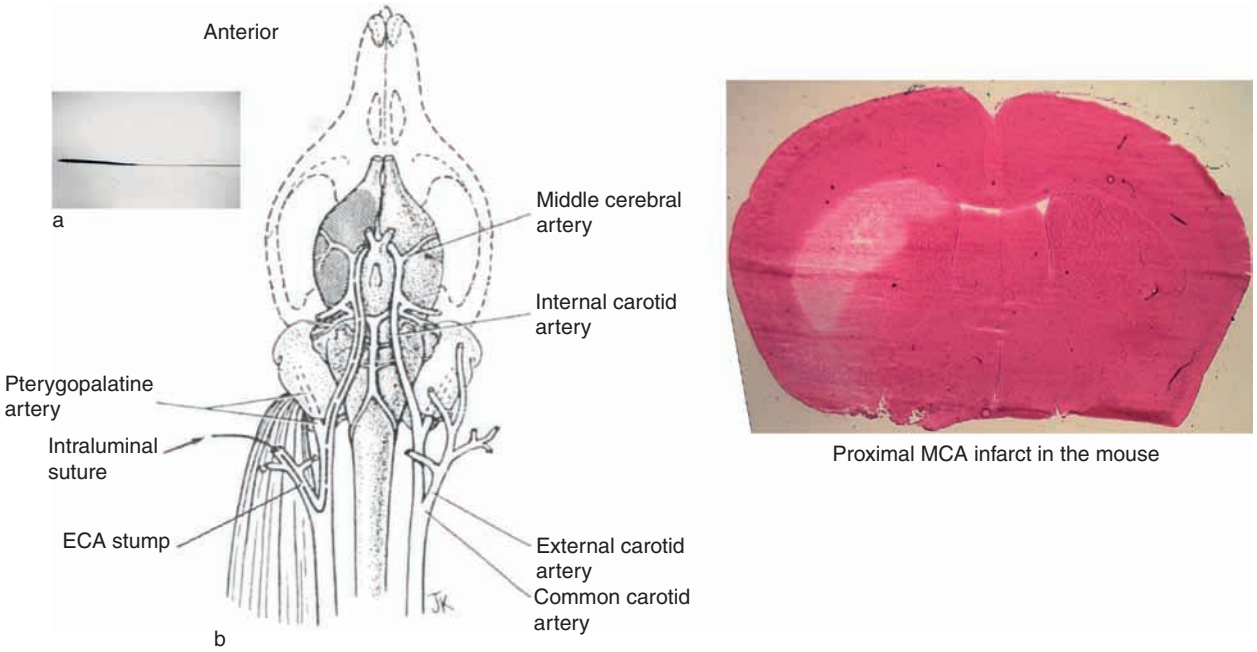

Figure 3 The intraluminal suture middle cerebral artery (MCA) occlusion. The obstructive filament, often coated to reach a specified diameter (a), is introduced in the external carotid artery (ECA) and advanced briefly into the common carotid artery. The ECA is cut to allow for manipulation of the suture. The suture is withdrawn slightly, is maneuvered into the internal carotid artery, and then advanced until it blocks the proximal middle cerebral artery (b). This model generally creates a large striatal infarct, with some cortical involvement. Adapted from Longa EZ, Weinstein PR, Carlson S, et al. (1989) Reversible middle cerebral artery occlusion without craniectomy in rats. *Stroke* 20: 84–91.

Distal MCA occlusion is achieved by accessing the distal MCA, either by craniotomy over the parietal/frontal cortex, in rodents; or by orbital rim approach following enucleation, in gyroencephalic mammals. In the rat model, the skin between the eye and the ear is cut, and the underlying muscles are dissected to expose the skull. Two small burr holes are drilled into the skull, one to access the artery for monitoring of CBF and the other for occlusion of the artery. The distal MCA is then directly visualized through the operating microscope and the more distal burr hole is used for occlusion of the MCA. The dura over the MCA is opened and the MCA is occluded along with bilateral or ipsilateral occlusion of the CCA, depending on the strain. This model was first described in 1986 by Chen and colleagues, who occluded the MCA by ligation with consequent occlusion of one or both CCAs, generating large cortical insults. A newer model of distal MCA occlusion was created that also generated a large cortical insult, but which relied on the placement of a microsurgical clip on the distal MCA, increasing the ease of producing ischemia with reperfusion (Figure 4). One major advantage of this clip model is that it generates the ischemia with only brief periods of anesthesia during the clip placement and removal. Occlusion of one or both CCAs is dependent on rat strain: Wistars require both CCAs to be occluded for successful ischemia, while spontaneously hypertensive rats require only ipsilateral occlusion. The cortical infarction induced by these models correlates with a breakdown of the blood–brain barrier and edema.

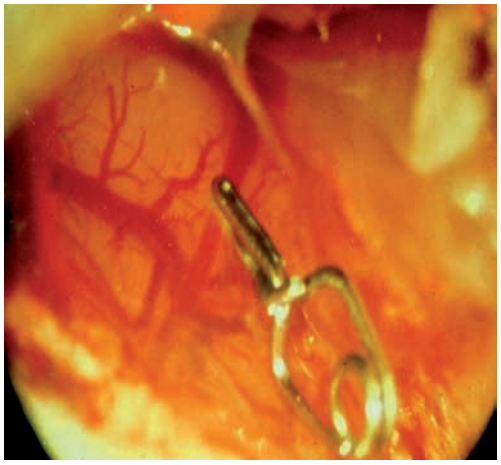

Figure 4 Distal middle cerebral artery occlusion; clip placement on the middle cerebral artery. This model requires a craniotomy and incision through the dura to access the distal middle cerebral artery. The rat is anesthetized during the clip placement and occlusion of one or both common carotid arteries, depending on the strain. The animal is awake during the ischemia, but is reanesthetized for clip removal.

Gyroencephalic models generally attempt to address questions regarding ischemia in an animal with brain structure more similar to that of humans. A new model to assess MCA occlusion in higher mammals is achieved in the miniature pig; the approach to the MCA is over the frontotemporal cortex, accessing the MCA via an orbital rim osteotomy. The cat has also been used to study focal ischemia and the effects of hypothermia on the brain. Generally, the nonhuman primate model, such as in the baboon, generates occlusion of the MCA by transorbital approach. The MCA is located following enucleation and an inflatable silastic balloon cuff is placed around the MCA. This allows for inflation of the balloon and induction of ischemia in an awake animal. Nonhuman primates are important to assess complex behaviors and recovery following stroke and allow for an assessment of injury in a brain that is very similar to a human brain. Nonhuman primate models should only be used to evaluate neuroprotective agents when positive results in rodent models generate enthusiasm before transferal into clinical trial.

Common carotid artery occlusion Occlusion of the common carotid artery can be used to generate focal ischemia in the gerbil, the adult rat with systemic hypotension, or the young rat with hypoxia. By far the simplest method for generating a focal insult is by clipping one of the CCAs in the gerbil, due to their lack of the posterior communicating artery, but investigations must be carefully conducted due to the possibility of gerbils having a PCA, which might confound results. The ventral cervical surface is opened surgically, under general anesthetic, and the CCA is isolated. The artery is then occluded for a specified period of time, with reperfusion. The gerbil model has been used to evaluate motor deficits and examine magnetic resonance image changes following focal ischemia. The main advantage of the gerbil model is the ease of inducing focal ischemia; however, to limit variability in this model, one must be careful to ensure that the posterior communicating arteries are not present in the supplies of gerbils.

Occlusion of one CCA in the rat with systemic hypotension can be used to generate focal insults with minimal surgical invasiveness. The ventral cervical surface of the rat is opened, and the CCA located and isolated. This artery is then occluded and systemic hypotension is achieved by exsanguinations to a blood pressure of 40 mmHg. This model approaches the ease of the gerbil model, but introduces more difficulty due to the removal of blood to generate hypotension. In neonatal rats, it is possible to generate a focal insult by a combination of hypoxia and ischemia. Again, the ventral cervical surface is

opened and the CCA located and dissected. The artery is occluded and hypoxia is created by placing the rat in a low-oxygen environment. This model is very important for modeling neonatal or fetal ischemia in humans; among children, ischemia *in utero* or during parturition is a common mechanism for brain damage.

Pial strip The pial strip method for creating a localized, specific cortical lesion is achieved by removal of the meninges over a particular area of the cortex. This method is generally used in the rat, and is able to generate lesions consistent in size and location. A midline incision in the scalp in created and a small craniotomy is used to access the meninges. The dura and pia mater are removed, thus removing the collateral vessels from that particular portion of cortex, which will go on to die. The main advantage of the pial strip is the reproducibility and the ability to selectively place the lesion on a specific brain region (i.e., motor cortex), creating a specific deficit. This allows for greater power in behavioral studies, and assessment of different treatments in very specific and precise motor behavioral tests. A disadvantage of this model is that it is not a classical focal ischemia/reperfusion model, with a core and penumbral areas of CBF reduction and infarct, so it may be difficult to assess some aspects of ischemic cell death.

Photothrombosis Photothrombosis is a method to induce stroke in rodents; it relies on the injection of light-sensitive dye followed by a laser pulse to create clot within the artery. To generate ischemia in this model, a photosensitizing dye is first injected intravenously into the animal. Following this, the skin and muscle over the skull are retracted, and the area of the brain to undergo ischemia is irradiated through the intact skull, using green light at 560 nm. This model has been shown to create thrombotic plugs and stasis of the red blood cells in vessels, with platelet aggregates. The associated infarction is due to the formation of a thrombus within the vessels and generates a reproducible ischemia, making this model very attractive for investigations requiring a thrombus. As well, this process is much less invasive than the other surgical models and the application of the laser can be directed to ensure ischemia in one particular brain area, such as the motor cortex, making this an attractive model for consequent behavioral testing. A major disadvantage of this model is that it is a permanent ischemia. There is no reperfusion aspect to the injury, limiting the evaluation of reperfusion injury that is possible with other models. This model has also been adapted for use in the mouse, using rose bengal and a krypton laser beam (568 nm). The region of the MCA is irradiated along with a ligation of the CCA to generate a consistent, reproducible ischemia.

Embolic The embolic model is created by introducing a cannula containing occluding substances, such as a clot or microspheres, into the CCA, and injecting these into the artery; the occluding substances generate ischemia in different locations dependent on their size. These embolic models are frequently used in rabbit ischemia. Generally, the rabbit model of focal ischemia induces an occlusion through the injection of a clot into the common carotid artery. This includes both small microclot and larger clot models. The rabbit small clot embolus model (RSCEM) introduces obstructive emboli into random small arteries of the brain, resulting in a heterogeneous stroke, while the rabbit large clot embolus model (RLCEM) causes an obstructive embolus in the MCA. Following embolus injection into the rabbit, behavioral changes can be used to assess damage. The RSCEM can also be used to determine clot load needed to affect 50% of rabbits, which can be used to assess neuroprotection. Because this model involves thrombus formation, it is very useful for determining the effects of proposed thrombolytics. In fact, tissue plasminogen activator was first evaluated as useful therapy for ischemic stroke in the rabbit model.

Rabbit physiology is very stable following the preparatory surgery to induce RSCEM or RLCEM. This is in contrast to the rat and mouse models, which require constant monitoring to maintain physiology within normal levels.

Endothelin The endothelin (ET-1) model of cerebral ischemia is achieved by stereotaxic injection of endothelin, a potent vasoconstrictor that occludes arteries in the area of injection. The application of ET-1 onto the cerebral cortex has been shown to create an ischemic injury that is reproducible, with consistent lesions following injection, and which shows a dose-dependent reduction of flow. ET-1 can be used to induce ischemia in several different brain locations, depending on the delivery. The drug can be applied topically to the cerebral cortex, directly over the MCA, or injected into various brain regions, including the white matter, by advancing a needle into the brain to the correct region using stereotaxic coordinates. Upon application of the ET-1 there is a rapid, but not immediate, reduction in blood flow, which shows gradual reperfusion over time. This mimics the aspect of human stroke without thrombolysis. Increased variability can be introduced into the MCA occlusion model of ET-1 ischemia due to the difficulty in accurate application over the MCA.

Another difficulty with the MCA occlusion model is the necessity for craniotomy, which might cause some trauma to the cortex. The intracerebral injections limit this trauma because craniotomy is not required for the injections. A final consideration must be made on the appropriate application site for experimental study, because certain application sites (i.e., MCA) do not allow for a sufficient motor cortex infarction, necessary for assessment of behavior.

Hemorrhagic Stroke Models

Hemorrhagic stroke models usually attempt to address the effects of cerebral bleeding into either the intracerebral tissue or in the surrounding meninges. These models are usually achieved by either an injection of autologous blood into the brain or injection of collagenase, a chemical that breaks down the vessel walls.

The injection of collagenase into the cerebral tissue will break down type 4 collagen, a major component of the vasculature, and create intracerebral hemorrhage, with a core infarct and a surrounding penumbral area. The core of the hemorrhage has high levels of fluid, red blood cells, and fibrin.

Injection of autologous blood can be performed in various brain areas, including directly into the cerebral tissue or into spaces between the meninges. Following creation of subarachnoid hemorrhage in the monkey, there is a consequent chronic vasospasm. Immediately following hemorrhaging into the brain, there is a consequent reduction of perfusion pressure to the brain and also a reduced CBF. The newest model to generate hemorrhage is by ultrashort laser pulses, which will allow for creation of small, discrete hemorrhage.

Conclusion

Stroke modeling, both ischemic and hemorrhagic, is constantly evolving in an effort to create animal models that most closely mimic human stroke. While these models approximate the human condition, they are not absolute. Thus, it is important to show the efficacy of treatment in more than one model, and for extended durations of animal survival, with extensive behavioral testing. All animal physiology must be carefully monitored and kept within the physiologic levels to ensure that the outcome seen is due to the treatment and not to physiologic perturbations. A clear understanding of the models themselves and what they are capable of testing is also required, and the suitability of the chosen animal model for testing a hypothesis must be carefully considered.

See also: Map Plasticity and Recovery from Stroke; Stroke; Stroke: Injury Mechanisms; Stroke: Neonate vs. Adult.

Further Reading

Astrup J, Siesjo BK, and Symon L (1981) Thresholds in cerebral ischemia – the ischemic penumbra. *Stroke* 12: 723–725.

Barber PA, Demchuk AM, Hirt L, et al. (2003) Biochemistry of ischemic stroke. *Advances in Neurology* 92: 151–164.

Carmichael ST (2005) Rodent models of focal stroke: Size, mechanism, and purpose. *NeuroRx* 2: 396–409.

del Zoppo GJ (1998) Clinical trials in acute stroke: Why have they not been successful? *Neurology* 51: S59–S61.

Dirnagl U, Iadecola C, and Moskowitz MA (1999) Pathobiology of ischaemic stroke: An integrated view. *Trends in Neurosciences* 22: 391–397.

Fisher M (2004) The ischemic penumbra: Identification, evolution and treatment concepts. *Cerebrovascular Disease* 17(supplement 1): 1–6.

Fukuda S and del Zoppo GJ (2003) Models of focal cerebral ischemia in the nonhuman primate. *ILAR Journal* 44: 96–104.

Hoyte L, Kaur J, and Buchan AM (2004) Lost in translation: Taking neuroprotection from animal models to clinical trials. *Experimental Neurology* 188: 200–204.

Longa EZ, Weinstein PR, Carlson S, et al. (1989) Reversible middle cerebral artery occlusion without craniectomy in rats. *Stroke* 20: 84–91.

Pulsinelli WA and Brierley JB (1979) A new model of bilateral hemispheric ischemia in the unanesthetized rat. *Stroke* 10: 267–272.

Richard GA, Odergren T, and Ashwood T (2003) Animal models of stroke: Do they have value for discovering neuroprotective agents? *Trends in Pharmacological Sciences* 24: 402–408.

Traystman RJ (2003) Animal models of focal and global cerebral ischemia. *ILAR Journal* 44: 85–95.

Xi G, Keep RF, and Hoff JT (2006) Mechanisms of brain injury after intracerebral haemorrhage. *Lancet Neurology* 5: 53–63.

Zivin JA, Fisher M, DeGirolami U, et al. (1985) Tissue plasminogen activator reduces neurological damage after cerebral embolism. *Science* 230: 1289–1292.

Animals and the Biology of Music

W T Fitch, University of St. Andrews, St. Andrews, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In recent years, there has been a resurgence of interest in the biology and evolution of human cognitive specializations, including language, speech, and music. One of the dominant issues in all of these fields is the extent to which unusual human abilities depend on general characteristics of vertebrates, mammals, or primates that predated human evolution and to what extent something 'uniquely human' is based on mechanisms which themselves are unique to our species. Logically, an answer to this question requires comparative data: we cannot know what is unique to our own species without first examining others. Whether there are uniquely human mechanisms is a question to be answered by first identifying relevant mechanisms for the domain (e.g., pitch and rhythm perception, vocal learning, and instrumental abilities in the case of music) and then carefully examining multiple animal species to determine whether the trait is present in any of them. Unique mechanisms can only be identified by a failure to discover any corresponding mechanism in a nonhuman species. Currently, few of the mechanisms that underlie human cognitive specializations appear to be unique to humans, and it seems clear that the bulk of human cognitive functioning is based on mechanisms shared in some form with other species.

When mechanisms are found which are shared, we can further address issues of when and why a particular mechanism appeared in evolutionary time. Addressing 'when' requires the examination of homologies among related species (traits shared by descent from a common ancestor who possessed some version of the trait). For example, the hair cells of the inner ear are shared by all vertebrates, and thus we can infer that the common ancestor of all living vertebrates possessed hair cells. A different trait, a complex of three middle ear bones (malleus, incus, and stapes) connecting the tympanum to the cochlea, is found in all living mammals and thus represents a mammalian homology, present in the ancestral mammal. A second type of similarity, often called analogy, arises quite differently by a process of convergent evolution. Such traits as bipedalism in man and birds, wings on birds and bats, or vocal imitation in marine mammals and humans are analogous traits because they were not present in the last common ancestor of these species. Although such traits are not, by themselves, helpful for determining when the trait arose, they can offer valuable clues to help answer the 'why' question: convergent evolution often results from two species independently 'discovering' the same solution to a similar problem (e.g., flight). Thus, both homology and analogy have different complementary roles to play in understanding the evolutionary history of a trait. Understanding cognitive evolution requires that we take a broad comparative approach, and it will be incomplete if focused only on primates.

This article considers the biology and evolution of music from a broad comparative perspective. The evolution of music has been seen as a biological puzzle since Darwin, who wrote in 1871, "As neither the enjoyment nor the capacity of producing musical notes are faculties of the least use to man in reference to his daily habits of life, they must be ranked amongst the most mysterious with which he is endowed." As Darwin was well aware, some important clues to resolving this mystery come from animals related to humans only distantly, such as birds. This is because some core characteristics of human music production are shared with many bird species but not with nonhuman primates. Bird 'song' is thus an example of convergent evolution, and it may provide a means of testing hypotheses about evolutionary function of complex, learned acoustic communication. Music, like spoken language, has two sides that are intertwined: production and perception. This article focuses on the former because without music production there can be no music perception. This discussion is also restricted to vertebrates and acoustic communication due to space limitations. In humans, music is often accompanied by dance, and in some languages there is not even a separate word for these two activities; unfortunately, little is known about the neural mechanisms underlying dance even in our own species, and next to nothing is known about similar conventionalized movement patterns in nonhuman animals. As discussed later, even this somewhat restricted focus on 'musical' vertebrate acoustic communication provides a rich source of comparative data.

Terms and Definitions

The term 'music' is notoriously difficult to define because of the huge range and diversity of musical practices in our own species. For comparative purposes, such suggestions as Blacking's "humanly organized sound" or McDermott and Hauser's

"structured sounds produced directly or indirectly by humans" are clearly inappropriate because they render the concept of animal music impossible, by definition. We can avoid this problem by defining two core subdivisions of human music – song and drumming – which can be unambiguously defined but which have clear parallels in the animal world. Although vocal and instrumental music are often combined, the distinction between them is quite an obvious one among human musical cultures, with virtually all cultures possessing both but with musicians often specializing in producing one or the other form of music. As discussed later, the search for animal parallels of these two subdivisions of music leads in quite different directions.

'Song' in nonhumans can be defined simply as complex, learned vocalizations (for humans, we must add 'nonspeech' to this definition because human speech also constitutes a complex, learned vocalization, but one used for linguistic purposes of expressing propositional meanings). Complexity can be quantified in various ways, such as Kolmogorov or encoding complexity, but more subjective definitions may be appropriate for many purposes than any absolute threshold because what is 'complex' among birds or primates may be different for frogs or fish. Learning can most clearly be demonstrated in laboratory experiments, in which individuals are raised in the absence of auditory stimulation (and do not develop normal song) or given particular songs as input which they then acquire, perhaps with some modification or permutation. By this metric, song exists in humans, birds, and several groups of marine mammals.

'Drumming' can be defined as nonvocal sounds produced by striking an object or some part of the body against another object (the ground, a tree, or, in a few cases, as in human hand clapping or gorilla chest beating, against a different part of the drummer's body). By this definition, drumming has evolved independently in birds, the African great apes, and various rodents and other small mammal species. To my knowledge, there are no other instances of nonvocal communicative sounds in the animal kingdom that could be considered 'instrumental,' although there are many instances of mechanically produced sounds that rely on specialized anatomical mechanisms, such as cricket wing scraping or feather snapping or bill clapping in some bird species. Such uses of specialized body parts for sound-making is in some sense analogous to vocalization rather than to instrumental music; more important, there is no evidence that such mechanical sounds are learned and thus directly relevant to human instrumental music. Therefore, the discussion of nonhuman instrumental music will be restricted to drumming as just defined.

Animal Song

By the definition of song as complex, learned vocalizations, song has evolved convergently in many different vertebrate clades. Birds are the most prominent case, and the ability of birds to learn novel sounds from their environment (e.g., a different species or human speech) has been noted since Aristotle. At least three clades of birds appear to have independently evolved song: parrots, hummingbirds, and the oscine passerines, termed 'songbirds.' It has become clear that several groups of marine mammals can also learn complex novel signals. Among the cetaceans, these include the odontocetes (toothed whales such as dolphins and killer whales) and the mysticetes (baleen whales such as humpback and right whales). Among the pinnipeds, there is good evidence for complex vocal learning among the phocid or 'true' seals and weaker evidence for the otariid or 'eared' sea lions and fur seals, and walruses. Finally, of course, humans have an ability to learn complex vocal signals, which can be divided into linguistic speech signals and musical song signals. Most mammals tested, from guinea pigs to primates, can gate or inhibit vocal signals in an operant context and thus have some voluntary control over their vocal output, and there is evidence for a modicum of control over particular acoustic components of the call (e.g., fundamental frequency or amplitude). Beyond such basic modifications of innate calls, however, no nonhuman primate has been shown to learn novel signals of any complexity. Similarly, some bat species can match aspects of other bats' signals, but these too are rather simple and thus do not constitute song by the definition adopted here. Finally, a number of other species (e.g., elephants) may have vocal learning of complex signals, but data remain sparse.

Birds are the group best known for their vocal learning, and people have been fascinated by 'talking' birds that mimic human speech for centuries. Careful isolation experiments have demonstrated beyond reasonable doubt that many different songbirds must be exposed to normal song of their species if they are to sing normally themselves. In many species, this input must occur within a certain 'sensitive period' early in life, or normal song cannot be attained. For some species, song 'crystallizes' after this period ends, and the bird will sing in the same way for the rest of its life. In other 'open-ended learners,' the bird retains the ability to learn new song as an adult and often produces new songs each breeding season. Prototypically, birdsong is a male activity, concentrated during the breeding season and under strong hormonal control. In these cases, song often has a dual function: it serves both to repel other males (a territorial function) and to attract

females (a courtship/mating function). However, it should not be supposed that these are the only functions of song, nor that only male birds sing. In many species, females also sing (either alone or in duets with their mate). Such well-known species as North American cardinals (*Cardinalis cardinalis*) or European robins (*Erithacus rubecula*) have female song, and duetting may even be typical of sedentary birds living in the tropics. Song can also have various other functions besides mating and territoriality. For example, in many parrot species, both male and female birds sing throughout the year; although the function of parrot song remains poorly understood, it seems to have more to do with building and cementing social relationships than with territoriality or mating. In black-headed Grosbeaks (*Pheucticus melanocephalus*), both parents sing, apparently to maintain contact with their fledgling offspring. The great diversity of bird species with learned song (more than 4000 species) renders sweeping generalizations invalid, and this great diversity allows us to test adaptive and phylogenetic hypotheses about the evolution of birdsong.

Another vertebrate group well-known for its complex songs is the cetaceans, particularly the baleen whales. Although it has been suggested that sailors hearing whale songs through the hulls of their boats gave rise to the Greek myth of sirens' song, whale song was not discovered by science until the advent of underwater hydrophone-based listening systems during the 1940s. It is now clear that many cetaceans are highly vocal. Whereas odontocetes use their vocalizations as the basis for echolocation, and use learned 'whistles' in a variety of communicative situations, the complex songs of baleen whales show the clearest parallel to birdsong and to human music. The songs of the humpback whale (*Megaptera novaeangliae*) are by far the best studied, but some other whales, such as the right whale, also have quite complex, geographically varied vocalizations. The larger baleen whales, such as fin or blue whales, also produce impressive low-frequency vocalizations (20 Hz or below), but these are quite simple and thus do not constitute song by our definition. In all of these Mysticete species, it appears that only males produce vocalizations, and these are concentrated in the time preceding and during the mating season. Because it is difficult to observe whale behavior, much less is known about the function of whale song than about birdsong, but most scientists believe that these vocalizations serve roles similar to that of song in most songbirds and have a dual mate-attraction and competitor-repulsion function. An interesting difference is that, at any given time and place, all the males sing the same song, with the same complex pattern of shared motifs. However, this song changes throughout the mating period, and over a period of decades it changes beyond recognition. This pattern shows that humpback whales, at least, are open-ended learners, capable of adjusting or modifying their song throughout their life span.

The vertebrate group most recently discovered to be capable of learning complex vocalizations is the pinnipeds or seals. Of the main groups of pinnipeds (otariids or eared seals, phocid or earless seals, and the odobenid walruses), the evidence for vocal learning is strong only for phocids. Many phocids produce complex underwater vocalizations that are traditionally termed 'song.' However, the other two pinniped groups have considerable vocal control and are also likely to be vocal learners. The most striking demonstration of vocal learning comes from a male harbor seal (*Phoca vitulina*) named Hoover who lived for many years at the New England Aquarium in Boston. Hoover was orphaned at a very young age and spent the beginning of his life living with some Maine fisherman before being donated to the aquarium, where, upon attaining sexual maturity, he began producing strikingly accurate imitations of human speech, including a convincing New England dialect. Hoover's repertoire was quite limited, and he showed no indication of having any sense of the meaning of the phrases he uttered (which often seemed to be directed at other seals). Beyond this isolated but striking case, the evidence for song learning in phocids is typically more circumstantial, but many phocids show clear geographic differences in song that do not appear to be genetically based and thus are consistent with song learning. Phocid seal 'songs' are almost entirely produced by males during the mating season and thus appear to have a function similar to that of birdsong. Although phocids can and do vocalize in air (e.g., this is an important component of mother–infant communication), the complex vocalizations termed song are typically produced underwater, making behavioral observations and experiments challenging. This, combined with the relatively recent discovery of vocal learning in this group, means that we know frustratingly little about vocal control and vocal learning in this group. However, seals are easily trained and adapt readily to captivity (they are found in zoos and aquariums throughout the world), suggesting that the study of vocal learning in this group has a bright future.

Mechanisms Underlying Complex Learned Vocalizations

Humans possess direct connections between neurons in motor cortex and the primary motor neurons that control the larynx; because such connections

are lacking in other primates, this may represent one key aspect of our increased control over vocalization. Similarly, songbirds have direct monosynaptic connections between neurons in the forebrain and the hypoglossal motor neurons that control their vocal organ, consistent with the hypothesis that monosynaptic connections are an important component of the increased vocal control seen in this group. This hypothesis obviously predicts similar connections in seals or cetaceans, but I know of no anatomical work that tests this prediction. An interesting fact about the vocal production mechanism may provide some clues to the evolution of such motor control mechanisms: In most singing vertebrates, vocalizations are produced by an evolutionarily novel production mechanism rather than the larynx. This is true of all birds, which use the syrinx (a complex organ at the base of the trachea) to produce sounds, and of odontocete cetaceans, which appear to use a novel nasal sac system to produce sounds rather than the larynx. Sound production in the other major cetacean group, the large baleen whales, remains poorly understood. Only two 'singing' groups have evolved adequate control over the primitive tetrapod laryngeal vocal source: humans and seals. In both of these groups, the sound production mechanism is the standard one seen in most other vertebrates: vibrating vocal cords housed within the larynx produce sounds which are then filtered by the vocal tract. This unusual distribution suggests that it may be easier, in an evolutionary sense, to bring a novel production system under complex telencephalic control than to 'emancipate' the primitive laryngeal system from the unlearned and largely innate vocalizations it normally produces. Of course, this puts a special focus on the vocal learning mechanisms of seals, which are more likely to provide insights into the function of the human system than are those of birds or odontocetes. Unfortunately, very little is known about pinniped neuroanatomy and neurophysiology, so this potential remains unexplored.

Animal Drumming

Although animal song has received far more attention, the production of sounds by striking a body part against the substrate or an object has evolved in parallel in an equally diverse group of vertebrates. The most common object struck is the ground: several groups of mammals have convergently evolved a propensity to strike the earth with their limbs to signal alarm or threat (e.g., deer, elephant shrews, agoutis, wallabies, and various desert rodents), and a few use their head or teeth in the same way (e.g., mole rats *Spalax* and mole voles *Microtus*). Although these displays are mostly limited to simple 'thumps,' substrate drumming with elaborate, modifiable rhythms has evolved in some of the kangaroo rats in the genus *Dipodomys* (order Rodentia, family Heteromyidae). Of these, only the banner-tailed kangaroo rat *Dipodomys spectabilis* has been studied in detail by Jan Randall and colleagues. Kangaroo rats of both sexes use footdrumming with their hind legs to drum out complex rhythms that have multiple functions. The most common form appears to serve as a territorial ownership advertisement, although footdrumming patterns can also be used as mating displays and as a snake alarm call. Playback studies demonstrate that kangaroo rats listen to these sounds and respond to them by footdrumming, typically waiting for a pause in the other rat's drumming to begin their own display. Footdrumming patterns consist of a series of individual 'drums' arranged into short bursts termed 'footrolls' which themselves form larger sequences. These patterns are individually distinctive, and rats recognize their neighbors, responding more strongly to novel interlopers. Different kangaroo rat species have different footdrumming patterns. However, when changing territories, rats often change their footdrum signatures to be more different from those of their new neighbors, indicating that these patterns are ontogenetically flexible, suggesting voluntary control. Although the rats have no special adaptations for sensing vibration, they have ears modified for low-frequency hearing. Randall and Lewis have shown convincingly that these signals can be effectively transmitted seismically through ground-borne vibrations which are then converted into airborne sound within the rat's burrows, where environmental acoustic noise (e.g., wind noise) is silenced. This phenomenon may explain why drumming is so common among fossorial or ground-dwelling mammals. In summary, footdrumming in *D. spectabilis* represents a relatively elaborate communication system, with at least an element of learning involved, that provides an interesting parallel to many functional aspects of birdsong.

Another group of species which drum communicatively are the woodpeckers (Order Piciformes, Family Picidae). Although their normal foraging method (rapid, powerful pounding of wood with their chisel-like beaks) generates sound as an unavoidable by-product, many woodpeckers also engage in a communicative drumming in which they seek out particularly resonant hollow trees and drum out species-specific patterns that serve both territorial and mating advertisement functions (these are particularly common during the spring mating season). The different species-specific patterns of these drummed signals are readily audible to human listeners and can be used in species identification; playback studies

with several species have demonstrated that listening birds respond most strongly to the pattern of their own species. Although individual birds must certainly learn which trees to drum, there is no evidence that the patterns are learned. In an interesting example of partial convergence, palm cockatoos *Probosciger atterimus* also drum on hollow logs as a mating and territorial signal; however, they do so with a stick held in the beak. This is the only example of animal drumming known in which a 'drum stick' is used, and which thus it involves a two-part 'instrument.'

Ape Drumming

Another fascinating example of drumming is found in our nearest living cousins and thus provides the only example that is potentially homologous to drumming in our own species: the drumming behavior of the African great apes (chimpanzees, bonobos, and gorillas). I know of no reports of drumming, or even auditory gestures such as repeated clapping, in orangutans, suggesting that this behavior is a homology in the African great ape clade (including humans). Surprisingly, this interesting and conspicuous behavior has received little detailed scientific attention. Perhaps the most prominent and well-known variant of ape drumming is the chest-beating display of adult male silverback gorillas, which is a typical component of aggressive displays that is often preceded by vocalizations and may be followed by branch shaking and charging at the intruder. Drumming in chimpanzees typically seems to have a similar function: it occurs at the culmination of a pant–hoot display. Although chimpanzees do sometimes make percussive sounds by striking their own bodies (e.g., hand clapping), their drumming typically involves repeatedly striking the hands, feet, or both against particularly resonant trees. This sound can carry quite far through the forest and be audible when the chimpanzee's voice is not. Similar drumming behavior has been noted in the bonobo, or pygmy chimpanzee. Thus, all three of the African great apes feature drumming as one component of a multimodal display. However, drumming is not associated exclusively with agonistic or territorial contexts: it is also a component of playful interactions. This type of drumming is particularly prominent in juvenile gorillas, which appear to use drumming as an affiliative signal, often accompanied by the species-typical 'play face.' Drumming frequently initiates a bout of chasing, wrestling, or other rough-and-tumble play in young gorillas, and it is sometimes exchanged in bouts by the two playing individuals. Drumming in these contexts is far more flexible than that seen in silverback displays. Although chest beating is still common, playing gorillas will also drum on hollow objects or the ground, on their own feet, bellies, or heads, or even on the body of their playmate. Young chimpanzees also engage in flexible, relaxed drumming, although this is more often seen in solitary play than in an interactive, dyadic context. This type of playful, flexible drumming, observed in both sexes, seems a more plausible parallel to the affiliative, bonding aspects typical of human music making than the relatively stereotyped aggressive displays seen in adult males. Unfortunately, little is known about the details of ape drumming (e.g., concerning ontogeny and learning, timing and isochronicity, manual and hemispheric lateralization, or other key factors); thus, the degree to which the mechanisms underlying it are true homologs remains unknown. Even the degree to which apes can synchronize or 'entrain' to an externally-generated pulse remains unstudied. However, the drumming behavior of African apes shows enough parallels to human drumming, and thus to human instrumental music, to be clearly worthy of more detailed investigation.

Conclusions

There appear to be two possible parallels to human music in the animal world: song and drumming. Song has evolved in parallel in birds and marine mammals, and it appears to serve similar functions in many of these species. However, the diversity of behavior of singing birds renders any simple generalizations impossible, and this diversity provides an ample source of comparative data for testing phylogenetic and functional theories. In addition, the ease with which birds can be kept in the laboratory means that our understanding of the mechanistic basis of bird song learning, at both neural and genetic levels, is rapidly increasing. By comparison, almost nothing is known about the neural mechanisms underlying vocal learning in whales or seals, and this seems likely to remain the case for the large baleen whale, which produces the most complex songs. However, seals are an example of a poorly studied group that could potentially provide rich dividends for our understanding of the mechanistic basis for vocal learning in mammals. Song learning in seals is particularly relevant to human song and speech because (unlike birds or toothed whales) seals use essentially the same vocal production mechanism (larynx and vocal tract) as humans. Seal vocalization thus provides a neglected but potentially extremely promising field for future research.

Regarding drumming, many vertebrates produce sound by striking some body part against a substrate, but the use of this technique to produce complex, learned signals is less common. In the best studied

example, kangaroo rats of both sexes drum out individually distinctive patterns that can be adapted to the drums of new neighbors and thus appear to be learned. Similar patterns of behavior are likely to be present in several other desert rodents as well. Whereas all of the behaviors just described have evolved by a process of convergent evolution, and thus are (at best) analogs of human musical behaviors, there is one case of animal 'music' that can be plausibly considered a potential homolog of human instrumental music: the drumming behavior of the African great apes. Given the vast amount of information on other aspects of ape behavior, this topic has been oddly neglected, and so the detailed similarities and differences between human drumming and that of gorillas, chimpanzees, and bonobos remain to be empirically determined. However, applications of noninvasive imaging techniques such as positron emission tomography and functional magnetic resonance imaging to nonhuman primates raise the exciting possibility of examining the neural basis of rhythm perception (and perhaps production) in apes, and thus of exploring the function of a neural system homologous to but simpler than that used by humans in musical rhythm processing.

See also: Audiovocal Communication in Bats; Bird Song Systems: Evolution; Birdsong Learning: Evolutionary, Behavioral, and Hormonal Issues; Birdsong Learning; Birdsong: The Neurobiology of Avian Vocal Learning; Game Theory and the Economics of Animal Communication; Referentiality and Concepts in Animal Cognition; Signal Design Rules in Animal Communication; Vocal Communication in Birds.

Further Reading

Arcadi C, Robert D, and Boesch C (1998) Buttress drumming by wild chimpanzees: Temporal patterning, phrase integration into loud calls, and preliminary evidence for individual distinctiveness. *Primates* 39: 505–518.

Dodenhoff DJ, Stark RD, and Johnson EV (2001) Do woodpecker drums encode information for species recognition? *Condor* 103: 143–150.

Fitch WT (2006) The biology and evolution of music: A comparative perspective. *Cognition* 100: 173–215.

Janik VM and Slater PB (1997) Vocal learning in mammals. *Advances in the Study of Behavior* 26: 59–99.

Jurgens U (2002) Neural pathways underlying vocal control. *Neuroscience and Biobehavioral Reviews* 26: 235–258.

Justus TC and Hutsler JJ (2005) Fundamental issues in the evolutionary psychology of music: Assessing innateness and domain specificity. *Music Perception* 23(1): 1–27.

Marler P and Slabbekoorn H (2004) *Nature's Music: The Science of Birdsong*. New York: Academic Press.

McDermott J and Hauser MD (2005) The origins of music: Innateness, uniqueness, and evolution. *Music Perception* 23(1): 29–59.

Ralls K, Fiorelli P, and Gish S (1985) Vocalizations and vocal mimicry in captive harbor seals, *Phoca vitulina*. *Canadian Journal of Zoology* 63: 1050–1056.

Randall JA (2001) Evolution and function of drumming as communication in mammals. *American Zoologist* 41: 1143–1156.

Rehding A (2000) The quest for the origins of music in Germany circa 1900. *Journal of the American Musicological Society* 53(2): 345–385.

Riebel K (2003) The 'mute' sex revisited: Vocal production and perception learning in female songbirds. *Advances in the Study of Behavior* 33: 49–86.

Schaller GB (1963) *The Mountain Gorilla*. Chicago: University of Chicago Press.

Schusterman RJ and Van Parijs SM (2003) Pinniped vocal communication: An introduction. *Aquatic Mammals* 29: 177–180.

Wallin NL, Merker B, and Brown S (2000) *The Origins of Music*. Cambridge, MA: MIT Press.

Anterior–Posterior Spinal Cord Patterning of the Motor Pool

J S Dasen, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Locomotor behaviors in vertebrates depend on the formation of selective connections between motor neurons and their synaptic targets in the spinal cord and periphery. The precision in which spinal circuits are assembled is critical in controlling the precise temporal activation of muscle groups in the limbs. Many of the spinal circuits that control simple locomotor behaviors, such as the alternation of left and right limbs or the coordination extensor and flexor muscles during locomotion, are established during the early stages of embryonic development. Locomotor circuits appear to be shaped initially independent of sensory experience, suggesting a high degree of genetic determinism in their formation. A critical feature of all spinal locomotor circuits is the establishment of precise connections between motor axons and muscle targets in the limb.

Motor neurons share certain basic features that distinguish them from other classes of neurons in the spinal cord, and they also acquire specialized properties that allow them to make selective connections with target cells. For example, all spinal motor neurons possess axons that project outside the spinal cord and release acetylcholine as the primary form of neurotransmission. In many if not all other regards, motor neurons are a highly diverse class of neuron. This diversity is most apparent in the motor neuron subtypes that innervate skeletal muscles in the limb, where each muscle is innervated by dedicated groups of motor neurons called motor pools. The typical vertebrate limb contains more than 50 muscle groups, and each of these targets is innervated by a unique pool of motor neurons.

Anatomical Organization of Motor Neuron Cell Bodies and Their Axonal Projections

During development, motor neurons acquire subtype identities that define their position within the spinal cord and determine their ability to selectively innervate peripheral targets. Motor pool identities emerge over a series of sequential stages, and each step restricts the potential of motor axons to innervate alternate targets. An early step in the differentiation of motor neurons involves the segregation of their cell bodies into longitudinally arrayed columns, each column containing motor neurons that project their axons to a common peripheral target. Five major columnar groups of motor neurons are generated within the spinal cord, four of which localize to specific positions along the rostrocaudal axis (Figure 1). Of particular importance to the specification of motor pool fates is the generation of the lateral motor column (LMC) because the acquisition of an LMC identity directs motor axons toward the limb. Motor neurons within the LMC are generated selectively at brachial (forelimb) and lumbar (hindlimb) levels of the spinal cord and contain the motor pools that innervate specific limb muscles.

Motor neurons within the LMC can be further delineated on the basis of how their axons initially project into the developing limb bud. The cell bodies of motor neurons within the LMC that project dorsally or ventrally into the limb are segregated from one another and define two divisional identities within the LMC. Medially positioned LMC motor neurons (LMC_m) project ventrally within the limb bud mesenchyme while laterally positioned LMC neurons (LMC_l) project dorsally. These divisional subtypes define two coherent subgroups of motor neurons within the LMC and define an initial choice point for motor axons projecting into the limb.

Motor pools are organized within the columnar and divisional identities of motor neurons. A motor pool is defined as the group of motor neurons that project to a single muscle target in the limb. In many cases the cell bodies of motor neurons are clustered in discrete nuclei, although the physiological relevance to motor pool clustering is still uncertain. Anatomical studies of the position of motor pools in the spinal cord have revealed that each motor pool occupies a stereotypical position in the spinal cord. A motor pool within the LMC typically spans two to three segments of the spinal cord, and the number of motor neurons within a given pool is proportional to size of the muscle it innervates.

Classical Studies on Motor Neuron Development in the Chick

Many of the insights into the mechanisms controlling the synaptic specificity of motor neurons emerged from classical manipulations of the neural tube in chick embryos. The idea that motor neurons have intrinsic properties that allow them to selectively innervate specific muscle targets in the limb was supported by studies in which motor neurons were displaced from their normal position within the spinal cord. In one set of experiments, lumbar-level neural tube was rotated

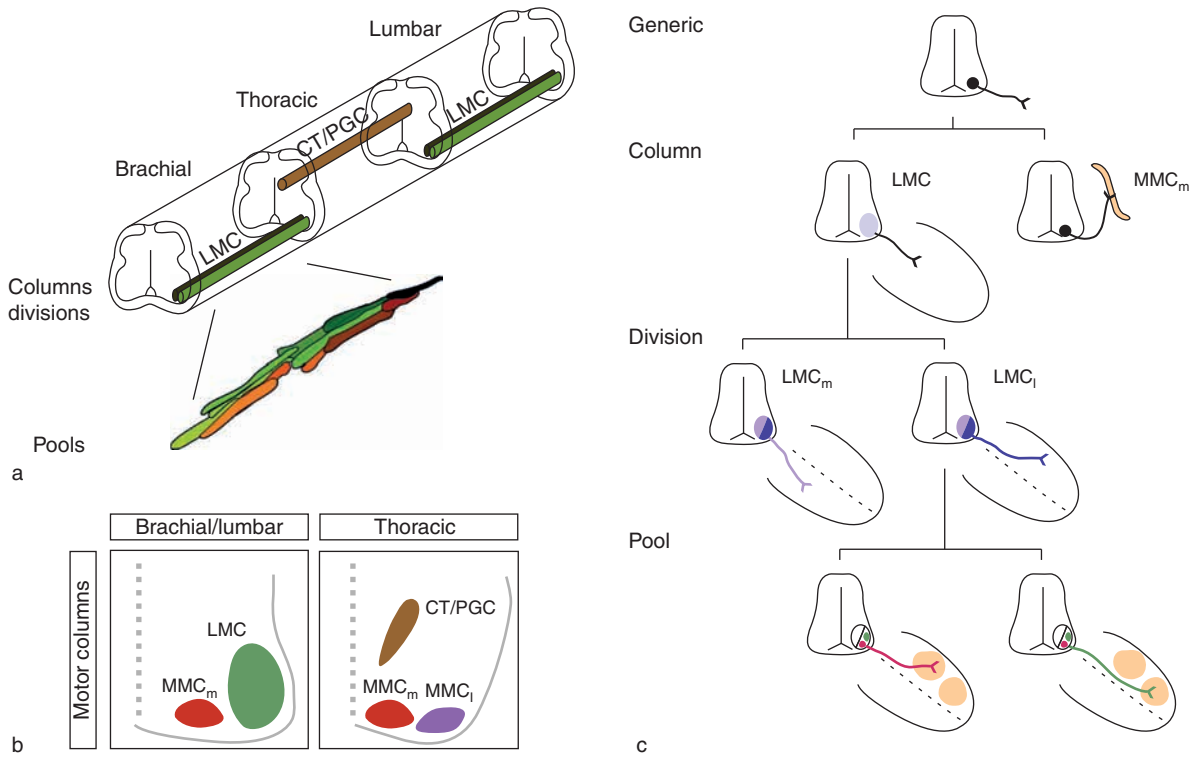

Figure 1 Motor neuron organization within the spinal cord and projection patterns of motor axon. (a) Motor columns and motor pools are generated at specific locations along the rostrocaudal axis. The cell bodies of motor neurons that send axons to the limb are contained within the lateral motor column (LMC) at brachial and lumbar levels of the spinal cord. Preganglionic motor neurons called the column of Terni (CT) in chick or preganglionic columns (PGC) in mouse are found at thoracic levels. The medial and lateral divisions of the LMC are depicted in different colors. Motor pools are generated at specific rostrocaudal positions within the LMC. (b) Schematic of a cross section of brachial/lumbar and thoracic regions of a chick spinal cord, showing the position of motor columns. LMC neurons are found at brachial and lumbar levels of the spinal cord. Motor neurons within the medial division of the medial motor column (MMC$_m$) are generated at all rostrocaudal levels of the spinal cord and extend axons to axial muscles. At thoracic levels, two segmentally restricted columns are generated, a set of lateral MMC (MMC$_l$) neurons that project their axons to body wall muscles and CT/PGC neurons that project axons to sympathetic neuronal targets. (c) Stages of motor neuron specification. Generic motor neurons share features common to all motor neurons, such as the projection of axons away from the spinal cord. Columns are sets of motor neurons that project to distinct regions in the periphery. Divisions are binary subdivisions of columns. Pools are subsets of motor neurons within the LMC that innervate a single muscle target in the limb. (a) Derived from Dasen JS, Tice BC, Brenner-Morton S, and Jessell TM (2005) A Hox regulatory network establishes motor neuron pool identity and target muscle connectivity. *Cell* 123: 477–491.

along the rostrocaudal axis, and the projection patterns of the displaced motor neurons were assessed using retrograde labeling assays. After inversion of the neuron tube along the rostrocaudal axis, motor neurons were still capable of finding their appropriate muscle targets even though they entered the limb from inappropriate positions. Thus, the position in which a motor axon enters the limb is not the primary determinant of its projection pattern or the selection of its synaptic target. In addition, these observations are consistent with the view that aspects of motor pool identity are specified prior to limb innervation.

The intrinsic properties of neurons within a pool presumably allow motor axons to differentially respond to guidance cues in the limb. In experiments in which mirror-image duplications of the limb musculature have been generated, motor axons within a single pool innervate both the normal muscle target and the duplicated muscle. These observations reinforce the idea that motor neurons have intrinsic properties that respond to positional cues and also underscore the importance of limb guidance cues in defining motor axon trajectories. A major challenge over the past 20 years has been to identify the molecules expressed within motor neurons and the limb mesenchyme that control motor axon projection patterns and synaptic specificity.

Positional Information and Cell Type Specification in the Spinal Cord

Motor neurons, like other cell types in the spinal cord, acquire their identities in response to positional cues acting along the dorsoventral and rostrocaudal axis of the neural tube. The pathways controlling

motor neuron subtype identity involve both extrinsic signals, typically in the form of secreted morphogens, and intrinsic signals, in the form of cell type-specific transcription factors. In the mechanisms that control motor pool specification, the intrinsic signals are particularly relevant because transcription factors are differentially expressed by motor neuron subtypes and presumably regulate the downstream genes involved in motor axon guidance decisions.

A major function of extrinsic signals is to establish unique patterns of transcription factor expression in naïve neuronal cell types. Depending on the relative position of neural progenitors from the source of a secreted signal, cells within the neural tube are exposed to different levels of morphogens. In the ventral neural tube, distinct classes of progenitors are specified in response to secreted signals originating from surrounding mesoderm, including sonic hedgehog from the notochord and floor plate and retinoic acid from the paraxial mesoderm. Graded sonic hedgehog and retinoic acid signaling induces the patterned expression of transcription factors in progenitor cells along the dorsoventral axis (Figure 2). These initial patterns of transcription factor expression are further refined through the selective cross-repressive interactions between pairs of transcription factors that act to sharpen the boundaries between progenitor domains and ensure that each progenitor expresses a unique transcription factor profile.

Progenitor cells expressing a specific pattern of transcription factors give rise to distinct classes of postmitotic neurons, including motor neurons. After leaving the cell cycle, all spinal motor neurons express transcription factors that are critical for generic features of their identity. Transcription factors expressed by early postmitotic motor neurons include the homeodomain proteins Hb9, Lhx3, Isl1, and Isl2. Genetic analysis of mice lacking these transcription factors has revealed they are required in each of the subsequent steps in motor neuron differentiation as mice lacking these transcription factors show defects in motor neuron columnar and pool specification. Some of the transcription factors required for early aspects of motor neuron specification are subsequently used in the further diversification of motor neuron subtypes. For example, the LIM homeodomain protein Lhx3 is initially required in all motor neurons and also has later function in the specification of the nonsegmentally restricted motor column that projects to axial muscle.

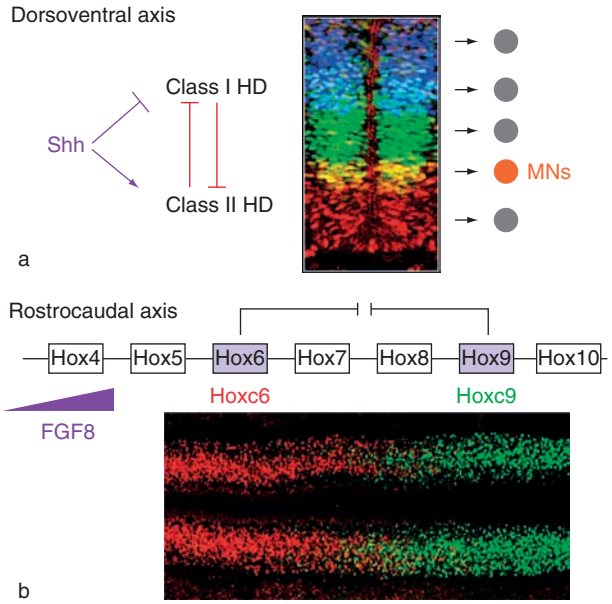

Figure 2 Specification of motor neuron (MN) subtypes along the dorsoventral and rostrocaudal axes. (a) Along the dorsoventral axis of the neural tube, motor neurons are generated in response to the graded activities of sonic hedgehog (Shh). Shh regulates the expression of transcription factors in progenitor cells. Class I transcription factors are induced by Shh, whereas class II transcription factors are repressed. Selective cross-repressive interactions between class I and class II transcription factors sharpens the boundaries between progenitor domains. Each of these progenitor domains gives rise to a distinct class of neuron, including motor neurons. Retinoic acid from the paraxial mesoderm also influences the pattern of transcription factors in neural tube progenitors (not shown). (b) Along the rostrocaudal axis, graded fibroblast growth factor (FGF) signaling induces the expression of chromosomally linked *Hox* genes in the neural tube. Genes located at one end of the cluster are expressed more rostrally, and genes at the opposite end are expressed caudally, in response to higher levels of FGF. The initial domains of Hox expression are further refined through selective cross-repressive interactions. Mutual exclusion of two Hox proteins is shown in top view of a longitudinal section through the spinal cord.

Establishing Patterns of *Hox* Gene Expression along the Rostrocaudal Axis

Although the transcriptional programs mediated by signaling along the dorsoventral axis of the neural tube define how motor neurons as a generic class are specified, additional signaling pathways are necessary for their further diversification into columnar and pool subtypes. Both the columnar and the pool identity of motor neurons requires that certain motor neuron subtypes are generated at specific rostrocaudal positions within the spinal cord. This allows motor neurons to be generated in proximity to their peripheral targets. For example, LMC neurons are generated selectively at limb levels of the spinal cord, while preganglionic column of Terni (CT) motor neurons are generated at thoracic levels in proximity to their synaptic targets in the autonomic nervous system. The columnar and pool identities of motor neurons have been typically defined by their cell body position within the spinal cord, their axonal trajectories, and late profiles of LIM homeodomain transcription

factor expression. However, the early signaling events that control motor neuron columnar and pool specification have only recently been explored.

How are motor neuron subtypes generated at distinct segmental levels of the spinal cord? One class of transcription factors known to be critical for establishing differences in cell identity along the rostrocaudal axis consists of members of the *Hox* gene family. In vertebrates, *Hox* genes encode a large family of homeodomain transcription factors consisting of 39 members organized into four chromosomal clusters. The expression patterns of individual *Hox* genes along the rostrocaudal axis of the spinal cord are closely related to the chromosomal position of a gene within a cluster – a principle called spatial colinearity (Figure 2). *Hox* genes at the 3′-end of a chromosomal cluster are expressed more rostrally in the embryo than genes at the 5′-end, which are expressed more caudally.

Similar to patterning events along the dorsoventral axis, *Hox* gene expression along the rostrocaudal axis is controlled by the actions of secreted signaling molecules which act in graded manner. Genes within a *Hox* cluster are sequentially activated in response to the activities of several signaling molecules, including fibroblast growth factors (FGFs), retinoids, and members of the transforming growth factor-β superfamily. Graded FGF signaling in particular appears to be important for the initial induction of *Hox* gene expression at brachial, thoracic, and lumbar levels of the spinal cord. An organizing region at the posterior end of the embryo, called Hensen's node in chick, is a source of FGF signaling, and as the node regresses caudally, more posterior regions of the spinal cord are exposed to FGF in higher concentration and over longer periods of time. Although FGF signaling appears to be essential in establishing the initial pattern of *Hox* gene expression in the neural tube, the final pattern observed in motor neurons is dictated largely by regulatory interactions between *Hox* genes.

Hox Proteins Function in the Specification of Segmentally Restricted Motor Columns

Several early studies have provided suggestive evidence that *Hox* genes are involved in specifying motor neuron subtype identities along the rostrocaudal axis. Experimental manipulation of mesodermally derived signals is known to affect the specification of segmentally restricted motor columns, such as LMC and CT motor neurons, and these changes in columnar fate are accompanied by alterations in the pattern of Hox protein expression. In addition, *Hox* genes are known to control cell type specification along the rostrocaudal axis of the hindbrain, and certain *Hox* mutants show defects in the projection of motor axons in the limb. However, one of the difficulties in relating the expression patterns of Hox proteins to motor neuron subtype identities was the lack of molecular markers that are specific for columnar subtypes.

The identification of genes expressed by segmentally restricted motor columns has permitted the analysis of the early steps in motor neuron diversification. LMC and CT motor neurons are generated at specific rostrocaudal levels and express unique molecular markers. LMC neurons can be defined by expression of retinaldehyde dehydrogenase-2 (RALDH2), an enzyme involved in retinoid synthesis, while CT motor neurons in chick selectively express bone morphogenetic protein-5, a member of the transforming growth factor-β superfamily. The expression of specific Hox proteins coincides with the position in which these molecularly defined columnar subtypes

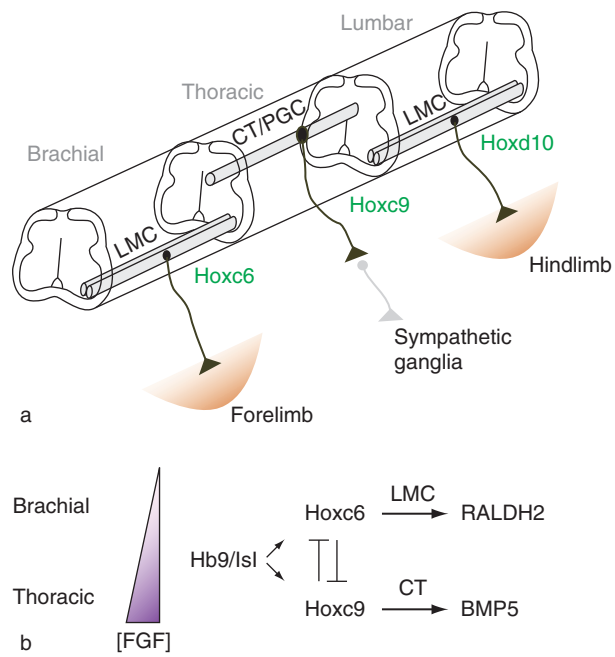

Figure 3 The role of Hox proteins in generating segmentally restricted motor neuron (MN) columnar subtypes. (a) Hoxc6, Hoxc9, and Hoxd10 are expressed in MNs (defined by Hb9 and ISL expression) at distinct rostrocaudal levels of the spinal cord. Each motor column has a distinct peripheral target. (b) Model indicating the roles of fibroblast growth factor (FGF) signaling and Hox expression in the specification of MN columnar identity. FGF establishes an initial pattern of Hox expression in the spinal cord. Cross-repressive interactions between Hoxc6 and Hoxc9 proteins refine the distinct Hox profiles of lateral motor column (LMC) and column of Terni (CT) neurons. Hoxc6 activity in brachial MNs directs expression of retinaldehyde dehydrogenase-2 (RALDH2; a marker for LMC neurons) and late features of LMC identity, whereas Hoxc9 activity in thoracic MNs directs bone morphogenetic protein-5 (BMP5) expression (a marker for CT neurons). Derived from Dasen Js, Liu J-P, and Jessel TM (2003) Motor neuron columnar fate imposed by sequential phases of Hox-c activity. *Nature* 425: 926–933.

are generated. Hoxc6 is expressed by brachial LMC neurons, Hoxc9 by thoracic CT neurons, and Hoxd10 by lumbar LMC neurons (Figure 3). The correlation in the pattern of Hox expression with LMC and CT columnar subtypes suggests that the same signals that control Hox expression may also specify columnar fates and that Hox proteins may function in the specification of columnar identities.

Consistent with a model in which *Hox* gene expression is controlled by graded FGF signaling, elevation of FGF levels at brachial levels of the spinal cord *in vivo* induces a pattern of Hox expression characteristic of thoracic levels. This switch in Hox expression patterns is accompanied by a conversion of brachial LMC neurons to a CT cell fate, as the CT marker bone morphogenetic protein-5 is ectopically expressed by brachial-level motor neurons. These effects of FGF on the columnar identity of motor neurons appear to be directly mediated by changes in *Hox* expression. Misexpression of Hoxc9 at brachial levels is sufficient to convert LMC neurons to CT neurons, while expression of Hoxc6 or Hoxd10 at thoracic levels can convert CT neurons to LMC neurons. In addition to these changes in motor neuron identity based on the expression of molecular markers, switching the pattern of Hox expression leads to alterations in the peripheral pattern of motor axon connectivity. For example, conversion of LMC to CT neurons forces limb-level motor neurons to project to sympathetic chain ganglia.

Hox proteins can transcriptionally cross-repress each other's expression, ensuring that columnar subtypes are generated only at specific segmental levels (Figure 3). Thus some of the mechanistic principles that govern dorsoventral patterning of the neural tube are shared in patterning along the rostrocaudal axis. One important difference is that while transcriptional cross-repression along the dorsoventral axis occurs in neural progenitors, transcriptional cross-repressive interactions between different Hox proteins occur predominantly in postmitotic cells. Nevertheless, these findings reinforce the view that cross-repressive interactions have critical roles in generating cellular diversity in the central nervous system.

Hox Transcription Factors and Motor Pool Identity

Studies on the columnar identity of motor neurons have provided evidence that Hox proteins contribute to neuronal specification along the rostrocaudal axis of the spinal cord. Many additional Hox proteins are expressed by motor neurons in patterns that do not coincide with the rostrocaudal positional boundaries of motor columns, suggesting additional roles in motor neuron diversification. The requirement for more than 50 motor pools to innervate each of the muscles in a vertebrate limb suggests a significant number of Hox proteins would need to be expressed by LMC neurons. Accordingly, of the 39 *Hox* genes, 21 are expressed in discrete subpopulations of motor neurons at brachial, thoracic, and lumbar levels of the spinal cord in a manner consistent with a role in motor pool specification.

In trying to understand the developmental programs that control motor pool specification, two organizational features of motor pools are particularly relevant (Figure 4). First, each motor pool occupies a stereotypic rostrocaudal position within the spinal cord. Thus some aspects of motor pool organization parallel the pattern of motor neuron columnar organization. Second, within a single segmental level, multiple motor pools can be present, and therefore some aspects of motor pool differentiation appear to emerge independent of the early signals that confer rostrocaudal positional information.

Identification of Motor Pools at Early Stages by Transcription Factor Expression

One of the difficulties in trying to define the pathways that control motor pool identity has been relating their anatomical organization with the patterns of Hox protein expression at the time that pools are specified. Motor pools have been defined classically by means of retrograde labeling assays performed after motor axons have reached their muscle targets, yet pool identities appear to be specified several days earlier in development. In addition, the pattern of Hox expression by motor neurons is complex; of the 11 Hox proteins expressed by brachial LMC neurons, a single motor neuron may express up to four different ones. To explore a role for Hox proteins in motor pool identity, what was needed was a set of molecular markers that are expressed by individual pools early in development.

Analysis of transcription factor expression has revealed that brachial and lumbar LMC neurons express an assortment of pool-specific transcription factors. Within the brachial LMC, anatomically defined motor pools can be molecularly defined by expression of the ETS transcription factor Pea3, the runt-related protein Runx1, and the POU-domain factor Scip (Pou3f1). Expression of these transcription factors defines motor pools that occupy stereotypic positions along the rostrocaudal axis and within a given segment. The expression of these pool-specific transcription factors has been used to ascertain the potential contribution of Hox proteins to the anatomical positioning and peripheral connectivity of motor pools.

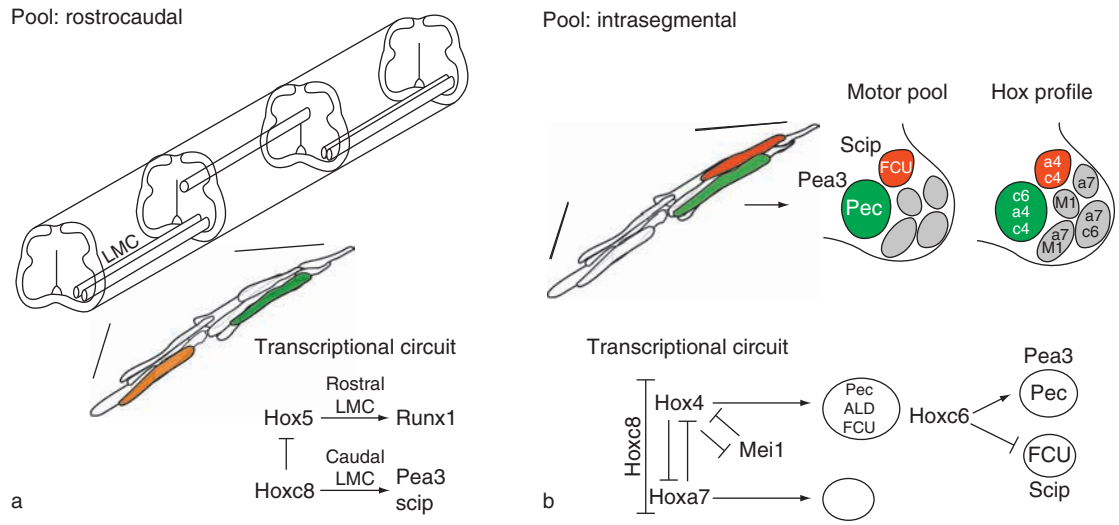

Figure 4 Regulatory networks of Hox proteins control motor pool identity and connectivity. (a) Hox proteins determine the rostrocaudal position of motor pools within the lateral motor column (LMC). At brachial levels of the spinal cord, cross-repressive interactions between Hox5 proteins and Hoxc8 establish the boundary between molecularly defined motor pools. Hox5 proteins are required to generate the motor pool that expresses the transcription factor Runx1 in rostral LMC neurons. Hoxc8 is required in caudal LMC neurons to generate the motor pools that express the transcription factors Pea3 and Scip. (b) Intrasegmental specification of motor pool identity. At a single axial level of the spinal cord, approximately six to ten pools are generated. Motor pools projecting to the pectoralis (Pec) and flexor carpi ulnaris (FCU) can be molecularly defined by expression of the transcription factors Pea3 and Scip, respectively. Both Pec and FCU pools express unique profiles of Hox expression. For simplicity, Hoxa4 and Hoxc4 are shown as Hox4, Hoxa5, and Hoxc5 are shown as Hox5; Mei1 is abbreviated as 'M1', Hoxc6 as 'c6', etc. The patterns of Hox expression in the Pec and FCU pools are established through a transcriptional circuit driven by Hox repressive interactions. Derived from Dasen JS, Tice BC, Brenner-Morton S, and Jessell TM (2005) A Hox regulatory network establishes motor neuron pool identity and target muscle connectivity. *Cell* 123: 477–491.

Hox Proteins and Motor Pool Rostrocaudal Positional Specification

Like motor columns, each motor pool occupies a stereotypic rostrocaudal position within the spinal cord. The position at which motor pools are specified appears to be determined by the pattern of Hox protein expression along the rostrocaudal axis. Brachial LMC neurons, defined by Hoxc6 and RALDH2 expression, can be further subdivided along the rostrocaudal axis by differential expression of Hox3, Hox4, Hox5, Hox7, and Hox8 proteins. For example, motor neurons in the rostral half of the brachial LMC express Hox5 proteins (Hoxa5 and Hoxc5) while motor neurons in the caudal half express Hoxc8. At the boundary between these two regions, expression of Hox5 and Hoxc8 are mutually exclusive.

The exclusive domains of Hox5 and Hoxc8 protein expression within the brachial LMC define a positional boundary between certain motor pools (Figure 4). The motor pool defined by expression of Runx1 is generated within the domain of Hox5 expression, whereas the pools expressing Pea3 and Scip are generated within the domain of Hoxc8 expression. Altering the pattern of Hox expression in these territories leads to changes in motor pool identity, defined by a switch in the molecular profile of pool-specific transcription factors. In addition, changing the pattern of Hox expression alters the peripheral connectivity of motor axons. For example, misexpression of Hoxc8 in the domain of Hox5 expression induces expression of Pea3 in rostral LMC neurons, and these ectopically generated Pea3 motor neurons project to the normal muscle target of this motor pool.

The mechanisms by which the rostrocaudal boundaries of motor pools are established by Hox proteins largely follow the mechanisms of motor neuron columnar identity; motor pool boundaries are established through selective cross-repressive interactions between Hox proteins. Rostral misexpression of Hoxc8 extinguishes expression of Hox5 proteins in a cell-autonomous manner; removing Hoxc8 expression from caudal LMC neurons leads to an expansion in the domain of Hox5 expression. In addition to Hox5 and Hoxc8, several other Hox proteins have rostrocaudal positional boundaries within the brachial LMC, and these boundaries likely correspond to the rostral and caudal limits of other motor pools.

Hox Proteins and the Intrasegmental Diversification of Motor Pool Identities

Multiple motor pools are generated at a given segmental level of the spinal cord, and this intrasegmental diversification appears to be established independent of the early patterning signals acting along the

rostrocaudal axis of the neural tube. Lineage analysis of motor neurons generated from a single rostrocaudal position has revealed that the fate of progenitors is not fixed because progenitors can give rise to motor neurons that occupy several different motor neuron subtypes. In addition, motor neurons that occupy a pool are initially dispersed within the spinal cord, and only relatively late in development do they cluster into discrete nuclei. Together, these observations suggest that there is no prepattern to motor pool specification at intrasegmental levels but that each motor neuron acquires a specific pool identity on a cell-by-cell basis.

Hox proteins appear to contribute to the intrasegmental diversification of motor pool identities. At late stages of development, the pattern of Hox protein expression becomes progressively restricted to distinct motor pools that occupy a given segment. The expression of the transcription factors Pea3 and Scip in brachial LMC neurons has been used to explore the function of Hox proteins in the intrasegmental diversification of motor pools because pools expressing these transcription factors are generated in overlapping segmental positions. The late patterns of Hox protein expression in pools defined by Pea and Scip expression suggest a model in which the intrasegmental diversification is driven by cross-repressive interactions between Hox4, Hox6, and Hox7 proteins (**Figure 4**). Experimental manipulation of the intrasegmental pattern of Hox expression alters the pattern of Pea3 and Scip expression by motor pools and the connectivity of their motor axons to their peripheral muscle targets.

The mechanisms that define the pattern of Hox expression at intrasegmental levels of the spinal cord appear to be distinct from those that control the pattern of Hox expression along the rostrocaudal axis. Along the rostrocaudal axis, cross-repressive interactions between Hox proteins are apparent shortly after motor neurons leave the cell cycle. In contrast, at intrasegmental levels, individual LMC neurons appear to express an initial cohort of Hox proteins based on their position along the rostrocaudal axis. Through continuous cross-repressive interactions that may be biased for the expression of one Hox protein over another, motor pools eventually express specific combinations of Hox proteins.

One corollary of this model is that biases in the efficacies of Hox cross-repression may explain differences in motor pool size. The number of motor neurons allocated to a specific pool is proportional to the size of the muscle it innervates. These differences in pool size emerge independent of trophic signals from the limb and thus appear to be intrinsically determined during an early phase in motor pool specification. Understanding the mechanisms controlling the differences in motor pool size may be relevant in understanding the allocation of cell numbers to specific neuronal fates in other regions of the nervous system.

Conclusions

Together, the studies described above provide evidence that members of the *Hox* gene family have important roles in the specification of motor pool subtypes. Yet many questions remain as to how Hox proteins contribute to the intrinsic programs that determine the specificity of motor neuron connectivity. Hox proteins function in many contexts throughout the embryo, and within the nervous system the same Hox factor can be expressed by multiple classes of neurons. These observations raise the question of how Hox proteins control gene expression in individual neuronal subtypes. Studies in *Drosophila* and other model systems have provided evidence that the specificity of Hox function is determined through interactions with other DNA-binding proteins. It remains to be determined whether Hox target specificity in motor neurons is controlled through interactions with additional motor neuron-restricted transcription factors.

Although these studies of motor pool specification have helped define some of the transcriptional networks that determine motor neuron identity, the pathways downstream of Hox proteins are less clear. The studies described earlier suggest that the combinatorial expression of Hox proteins in motor pools controls the expression of pool-specific transcription factors. Many of these pool-specific transcription factors may in turn control the expression of receptors which guide motor axons to specific muscle targets. Alternatively, Hox proteins themselves may directly regulate the expression of guidance molecules that are involved in intermediate choice points for motor axons projecting along the major axes of the limb.

Hox factors are expressed by several classes of neurons in addition to motor neurons, including sensory neurons and interneurons. One of simplest circuits in the spinal cord is the monosynaptic stretch reflex circuit, which in its most basic form consists of a motor neuron, a sensory neuron, and a muscle target. One possibility is that the precision of connections in this circuit arises from matching profiles of gene expression in motor neurons and sensory neurons. The expression of Hox factors in muscle sensory neurons parallels the segment-specific Hox patterns in motor neurons, raising the intriguing possibility that Hox factors are involved in the formation of sensorimotor circuits. The analysis of Hox function in sensorimotor connectivity provides a starting point for exploring more-complex levels of spinal circuitry, such as the local networks of interneurons and motor

neurons that give rise to rhythmic patterns of activity in the spinal cord and form the central pattern generators required for coordinate locomotor behaviors.

See also: Hox Gene Expression; Motor Neuron Specification in Vertebrates; Segmentation: Segmental Boundaries, Establishment and Morphogenesis (EPH); Segmentation: Spinal Cord Segmentation and A–P Somite Patterning; Transcriptional Networks and the Spinal Cord.

Further Reading

Bel-Vialar S, Itasaki N, and Krumlauf R (2002) Initiating Hox gene expression: In the early chick neural tube differential sensitivity to FGF and RA signaling subdivides the HoxB genes in two distinct groups. *Development* 129: 5103–5115.

Briscoe J, Pierani A, Jessell TM, and Ericson J (2000) A homeodomain protein code specifies progenitor cell identity and neuronal fate in the ventral neural tube. *Cell* 101: 435–445.

Dasen JS, Liu J-P, and Jessell TM (2003) Motor neuron columnar fate imposed by sequential phases of Hox-c activity. *Nature* 425: 926–933.

Dasen JS, Tice BC, Brenner-Morton S, and Jessell TM (2005) A Hox regulatory network establishes motor neuron pool identity and target muscle connectivity. *Cell* 123: 477–491.

Ensini M, Tsuchida TN, Belting HG, and Jessell TM (1998) The control of rostrocaudal pattern in the developing spinal cord: Specification of MN subtype identity is initiated by signals from paraxial mesoderm. *Development* 125: 969–982.

Jessell TM (2000) Neuronal specification in the spinal cord: Inductive signals and transcriptional codes. *Nature Reviews Genetics* 1: 20–29.

Kmita M and Duboule D (2003) Organizing axes in time and space: 25 years of colinear tinkering. *Science* 301: 331–333.

Landmesser LT (2001) The acquisition of motoneuron subtype identity and motor circuit formation. *International Journal of Developmental Neuroscience* 19: 175–182.

Leber SM, Breedlove SM, and Sanes JR (1990) Lineage, arrangement, and death of clonally related motoneurons in chick spinal cord. *Journal of Neuroscience* 10: 2451–2462.

Lin JH, Saito T, Anderson DJ, Lance-Jones C, Jessell TM, and Arber S (1998) Functionally related MN pool and muscle sensory afferent subtypes defined by coordinate ETS gene expression. *Cell* 95: 393–407.

Liu JP, Laufer E, and Jessell TM (2001) Assigning the positional identity of spinal MNs: Rostrocaudal patterning of Hox-c expression by FGFs, Gdf11, and retinoids. *Neuron* 32: 997–1012.

Stirling RV and Summerbell D (1988) Specific guidance of motor axons to duplicated muscles in the developing amniote limb. *Development* 103: 97–110.

Tsuchida T, Ensini M, Morton SB, et al. (1994) Topographic organization of embryonic MNs defined by expression of LIM homeobox genes. *Cell* 79: 957–970.

Antipsychotic Drugs

J P McEvoy, Duke University, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Chlorpromazine, the first of the conventional neuroleptic drugs, was synthesized in the early 1950s as a potential antihistamine. When Henri Laborit gave it to patients being prepared for surgery, their spontaneous movements became rare and slow. Initiative, interest, and affect diminished. They tended to rest, but could readily be aroused, and their intellectual functioning remained intact. Because of these effects, Laborit suggested that chlorpromazine be tried in psychotic patients, targeting agitation, aggression, and excitement; Jean Delay and Pierre Deniker did such trials, and these target features decreased. Unexpectedly, in these patients, the power and intrusiveness of hallucinatory perceptions and delusional beliefs also declined. Contact between the patient and the environment improved. Thoughts and behaviors became more organized. However, chlorpromazine and the other conventional neuroleptic drugs also produced dose-dependent neurological side effects, including dystonic contractions of facial, neck, and trunk muscles, a state of Parkinson-like muscular rigidity and tremors, and distressing motor restlessness. After months or years with these initial, reversible neurological side effects, some patients developed (tardive) dyskinetic or choreiform movements that persisted even if the conventional neuroleptic drug was discontinued.

The fundamental pharmacological action underlying the antipsychotic and neurological effects of the conventional neuroleptic drugs is the antagonism of dopamine neurotransmission. Stereospecific binding sites for conventional neuroleptic drugs in the brain proved to be dopamine (D2) receptors. Antagonism at D2 receptors leads to an initial increase in the firing of presynaptic dopamine neurons, followed by the induction of depolarization block in these neurons. The neurons fire and fail to repolarize; their capacity for phasic ('burst') firing is markedly diminished.

Animal models for psychosis are not available. The initial animal models used to screen for antipsychotic drugs were based on the neurological side effects of chlorpromazine, and all of the conventional neuroleptic drugs share dopamine D2 receptor antagonism and the same neurological side effects. Although the conventional neuroleptic drugs produce clinically meaningful reductions in psychopathology in most treated patients, few patients become symptom free or able to return to full functioning, and approximately 25% have substantial residual psychopathology. Most patients maintained on conventional neuroleptic drugs have at least mild neurological side effects, often with accompanying subjective distress.

Clozapine was the first 'atypical' antipsychotic. Clozapine was atypical because it did not produce neurological side effects in animal models. Initial clinical experience with clozapine suggested that it had a wider scope and greater degree of therapeutic efficacy than did the conventional neuroleptic drugs, and no propensity to cause neurological side effects. However, cases of agranulocytosis led to the withdrawal of clozapine from the market in 1975. In 1988, a definitive study demonstrated clozapine's significantly greater therapeutic effects in patients who had garnered little benefit from conventional neuroleptic drugs. Clozapine was then made available for use in patients unresponsive to other antipsychotic drugs, with the requirement for close monitoring of patients' white blood cell counts throughout the duration of treatment.

Clozapine has a complex pharmacology, including actions at numerous neurotransmitter receptors. Investigations are ongoing to determine if and how each of these actions contributes to clozapine's distinct clinical effects. Clozapine is an antagonist at type 2 serotonin (5-hydroxytryptamine; 5-HT_2) receptors, a property that ameliorates neurological side effects and improves sleep. A series of compounds (the newer atypical antipsychotic drugs) that combine antagonism (or partial agonism in the case of aripiprazole) at dopamine D2 receptors with antagonism at serotonin 5-HT_2 receptors proved to have equivalent or slightly superior therapeutic efficacy and a lower propensity for neurological side effects than did the conventional neuroleptic drugs. However, treatment with clozapine and several of the newer atypical antipsychotic drugs is associated with weight gain, lipid elevations, and an increased incidence of diabetes mellitus, resulting in increased risk for cardiovascular disease.

Conventional Neuroleptics

It is necessary to understand the dose–response relationships of conventional neuroleptic drugs to use them well.

Treatment of Acute Psychotic Exacerbations

Fixed-dose studies suggest that the dose–therapeutic response relationship for neuroleptic treatment is to be found between the daily equivalent of 100 and 700 mg of chlorpromazine. Higher doses are not

only unlikely to be more effective but may also yield inferior outcomes because of increased risk of neurological side effects. A simplistic argument could be made to treat all acutely psychotic patients with the equivalent of 700 mg chlorpromazine (~15 mg haloperidol) daily, since this dose will cover the minimum therapeutic dose range needed for essentially all potentially responsive patients. However, the group dose–response relationship for neurological side effects closely overlaps that for antipsychotic benefit, with a median effective dose in the range of 200–400 mg chlorpromazine equivalents daily. At a chlorpromazine dose of 700 mg daily, most patients experience coarse neurological side effects.

In 1961, Haase proposed that the lowest neuroleptic doses at which individual patients develop a slight increase in rigidity, detectable only by careful examination ('crossing the neuroleptic threshold'), are also the lowest doses at which patients attain maximum therapeutic benefit, and that no greater therapeutic benefit accrues to patients receiving higher doses – only increments in neurological side effects. Simply stated, those few patients who develop slight stiffness at 100 mg chlorpromazine daily are the same few patients who will respond therapeutically to that dose, and so on for the patients initially developing rigidity with each progressive dose increment. There is not a single minimum effective dose for all patients at which sufficient dopamine D2 blockade is achieved to induce depolarization block. The improvements in group mean therapeutic benefit seen across the dose range of 100–700 mg chlorpromazine daily reflect the accrual of more patients responding (i.e., more 'lights being turned on'), rather than all patients showing greater therapeutic benefit (rheostats increasing the brightness of all lights).

Positron emission tomography permits measurement of the occupancy of dopamine D2 receptors by conventional neuroleptic drugs. In keeping with the neuroleptic threshold hypothesis, antipsychotic efficacy becomes available when ~65% of D2 receptors are occupied, whereas coarse neurological side effects appear when >75% are occupied. Simply stated, coarse neurological side effects imply excessive dosing of these drugs.

Preventing Psychotic Relapse

The haloperidol dose that 'just' blocks a rise in plasma growth hormone levels in response to the oral administration of bromocriptine (0.5 mg kg^{-1}) approximates the minimum dose with the maximum therapeutic effect in preventing relapse in chronic schizophrenia. When the haloperidol doses of 16 patients with chronic schizophrenia (originally maintained on doses >20 mg daily) were lowered to 10 mg daily, 3 patients escaped from blockade. Two more escaped from blockade when the haloperidol dose was reduced to 5 mg daily, and 6 more escaped from blockade at 2.5 mg daily. The remaining 5 escaped when haloperidol was discontinued. No patient lost therapeutic antipsychotic benefit until there was an escape from blockade. After escaping from blockade, 9 of the 16 patients exhibited exacerbated psychosis. Lowest effective maintenance doses also appear to be between 100 and 700 mg of chlorpromazine equivalents daily.

Long-Acting Injectable Preparations

Because of the limited acknowledgment of illness and need for treatment that characterizes chronic psychotic disorders, long-acting injectable preparations of conventional neuroleptic drugs were developed. These preparations, which provide real-time information as to whether patients are taking prescribed antipsychotic treatment, have been used in maintenance treatment studies. If active treatment with one of these preparations is replaced by placebo injections, the 1-year relapse rates are >80%.

When 126 patients with chronic schizophrenia were assigned to blinded maintenance treatment with either standard-dose (12.5–50 mg every 2 weeks) or very-low-dose (1.25–5 mg every 2 weeks) fluphenazine decanoate (FD), cumulative 1-year relapse rates were significantly higher for the very-low-dose patients (56%) than for the standard-dose patients (7%). However, the majority of the patients who relapsed in the very-low-dose group responded quickly to additional antipsychotic medication and did not require rehospitalization or experience severe disruption of their social/occupational performance.

There were no differences in cumulative relapse rates at the end of 1 year among 66 outpatients with schizophrenia randomly assigned to low (5 mg every 2 weeks) or standard (25 mg every 2 weeks) dose maintenance treatment with FD, but at 2 years there was significantly better survival in the 25 mg dose group (64%) than in the 5 mg dose group (31%). If clinicians were permitted to make a dose adjustment up to 10 mg every 2 weeks in the low-dose group and up to 50 mg in the standard-dose group, there were no significant advantages for the standard dose in survival. Patients assigned to the standard dose treatment, compared to those given low doses, had more neurological side effects and more subjective distress during the early months of the study.

Among 70 outpatients with schizophrenia randomly assigned to double-blind maintenance treatment, either with the standard dose at which they had been stabilized (on average, 25 mg every 2 weeks) or with 20% of that dose (on average, 4 mg every

2 weeks), patients assigned to reduced doses and who were living in stressful situations (high expressed emotion families) experienced more minor but aborted episodes in the second year, with no increase in rehospitalization rate. Fewer neurological side effects occurred in the reduced-dose group, and, over time, reduced-dose patients were significantly more improved in their social/occupational role performance than were standard-dose patients.

The 1-year relapse rates (>80%) with placebo are progressively reduced as FD doses are increased through the very low (1.25–5 mg) and low (5 mg) dose ranges; relapse rates begin to match those associated with standard doses in the range of 10–12.5 mg every 2 weeks. For the vast majority of patients, no additional relapse prevention accrues at higher 'standard doses,' but more objective neurological side effects, subjective distress, and dampening of social and occupational performance develop.

Tardive Dyskinesia

Tardive dyskinesia (TD) is a syndrome of choreiform or athetoid abnormal involuntary movements that increase with emotional arousal, decrease with relaxation, and disappear during sleep. Limb dyskinesias are more common in younger patients, whereas orofacial dyskinesias are more common in older patients. Although abnormal involuntary choreiform movements and orofacial dyskinesias were described in patients with schizophrenia before the introduction of the conventional neuroleptics, use of the conventional neuroleptics has increased the prevalence by approximately 20–30%, also increasing the severity of the movements in some patients. During continuous treatment with conventional neuroleptic drugs, the incidence rate is 4–6% of unaffected patients per year. Older age, longer duration of treatment, and taking a conventional neuroleptic drug at doses that produce neurological side effects are factors associated with development of TD. Patients whose conventional neuroleptic dose cannot be reduced to a level at which neurological side effects are minimal without losing therapeutic benefit should be switched to a newer atypical antipsychotic, in particular, quetiapine or olanzapine.

In most cases, TD is not progressive, and even with continued treatment with an antipsychotic drug there may be improvement over time, especially if the lowest effective doses can be employed. However, a subgroup of patients will develop severe and disabling forms of TD; they can be identified early by their extreme sensitivity to neurological side effects. When prescribing a conventional neuroleptic drug, clinicians must document: (1) that patients have been informed of the risk for TD and have consented to treatment, (2) that examinations for neurological side effects and early signs of TD, using standard scales and procedures, are completed at regular intervals during treatment, and (3) that reasonable steps are taken if evidence of neurological side effects or TD appear.

Clozapine

Clozapine offers therapeutic efficacy to approximately 50% of patients who get little benefit from conventional neuroleptic drugs, but has a substantial burden of side effects. Among 268 patients with treatment-resistant schizophrenia randomly assigned to 6 weeks of blinded treatment with clozapine up to 900 mg daily or chlorpromazine up to 1800 mg daily, 30% of clozapine-treated patients, but only 4% of chlorpromazine-treated patients, met predetermined 'responder' criteria. Clozapine-treated patients had significantly greater reductions in Brief Psychiatric Rating Scale total scores than did chlorpromazine-treated patients.

Clozapine produced significantly greater reductions of positive psychopathology than did haloperidol among 39 outpatients who had shown partial therapeutic responses to conventional neuroleptics. Eight of 19 clozapine-treated patients, but only 1 of 20 haloperidol-treated patients, met responder criteria. Smaller advantages for clozapine were seen on negative psychopathology, with little benefit to deficit-state patients.

In a study of 227 chronically hospitalized patients randomly assigned to clozapine or continuation of standard care for up to 2 years, those patients assigned to clozapine had significantly greater reductions in neurological side effects, disruptive behaviors, and hospitalization than did patients continued on standard care. There were no differences in psychopathology or quality of life. Both groups were equally likely to be discharged, but, once discharged, patients assigned to clozapine were less likely to be readmitted. Among 423 treatment-resistant patients with schizophrenia randomly assigned to up to 1 year of treatment with either clozapine or haloperidol, patients assigned to clozapine were significantly more likely to complete the year of treatment (57%) than were patients assigned to haloperidol (28%), had lower psychopathology scores and had fewer inpatient days.

In addition to producing superior antipsychotic efficacy, clozapine reduces hostility and violent behaviors, suicidal behaviors, substance use, and the likelihood of incarceration, relative to treatment with other antipsychotics. It is the only antipsychotic that can reduce excessive water drinking and hyponatremia that endanger some patients with chronic schizophrenia. However, clozapine can have severe,

life-threatening side effects. Approximately 1 in 100 patients treated with clozapine will develop agranulocytosis, which, if not detected, can result in overwhelming infection and death. Monitoring of white blood cell counts is required by the Food and Drug Administration as a condition of treatment with clozapine. Approximately 1 in 400 patients treated with clozapine develops myocarditis, usually within the first 2–4 weeks of treatment. Monitoring of inflammatory markers (e.g., eosinophil counts, erythrocyte sedimentation rate), indicators of cardiac damage (e.g., creatine phosphokinase level, troponin level), and the electrocardiogram, weekly for the first 4 weeks, may permit early detection, leading to discontinued treatment. Clozapine can precipitate epileptic seizures in high-risk individuals. Clozapine can produce substantial weight gain, lipid elevations, and diabetes mellitus; these effects are associated with increased risk for cardiovascular disease. Less dangerous side effects include bed-wetting that can be corrected by ephedrine (25 mg) at bedtime, drooling that reflects a decreased swallowing rate rather than excessive saliva production, and marked sedation.

Newer Atypical Antipsychotic Drugs

In 8-week registration trials, risperidone (4–8 mg daily) produced greater reductions in affective and negative psychopathology than did haloperidol (10–20 mg daily), with less neurological side effects. In a maintenance trial in patients with chronic schizophrenia, risperidone-treated patients (mean dose, xx mg daily) had better survival (1-year relapse risks, 34% vs. 60%), greater reductions in psychopathology, and better performance on neurocognitive tests than did haloperidol-treated patients (mean dose yy mg daily). Risperidone showed little advantage over low-dose haloperidol in the treatment of first-episode patients, and was not significantly better than haloperidol for treatment-resistant patients. Risperidone elevates prolactin levels more than other antipsychotics do; sensitive individuals may develop breast engorgement, milk production, loss of menstrual cycle, and sexual dysfunction.

In short-term registration trials, olanzapine (10–15 mg daily) was superior to haloperidol in reducing affective and negative psychopathology. In an extension trial involving patients who had responded to olanzapine or haloperidol acutely, the 1-year relapse rate was estimated to be 20% for olanzapine and 28% for haloperidol. However, in a large collaborative study comparing olanzapine versus haloperidol plus prophylactic benztropine (an anticholinergic anti-Parkinson's drug), little advantage could be demonstrated for olanzapine. First-episode patients treated with olanzapine had less affective and negative psychopathology than did haloperidol-treated patients, and better preservation of brain volume over 2 years of treatment. Olanzapine produced greater reductions in psychopathology compared to haloperidol in treatment-resistant patients. Olanzapine has been associated with more weight gain, more lipid elevation, and more insulin resistance compared to the other antipsychotic drugs, except clozapine.

In two large registration trials, quetiapine was not superior to either haloperidol or chlorpromazine in reducing psychopathology, but had substantially less risk for producing neurological side effects. Ziprasidone was not superior to conventional neuroleptic comparators in registration trials. Ziprasidone produced less neurological side effects. In a 28-week maintenance trial, ziprasidone was equivalent to haloperidol on overall psychopathology, but was associated with less negative psychopathology.

In short-term registration trials, aripiprazole was not superior to haloperidol in therapeutic efficacy. In a 52-week maintenance treatment trial, aripiprazole-treated patients had better survival, and less affective and negative psychopathology, than did haloperidol-treated patients. Aripiprazole was not more efficacious than perphenazine, a conventional neuroleptic drug, for treatment-resistant patients.

The newer atypical antipsychotic drugs are not equivalent in efficacy; only olanzapine and risperidone have evidence of greater therapeutic efficacy than is noted with the conventional neuroleptic drugs. Any advantages for the newer atypical antipsychotic drugs in therapeutic efficacy or survival disappear if the conventional neuroleptic dose is less than 12 mg daily of haloperidol (~600 mg daily of chlorpromazine), although an advantage of less neurological side effects persists.

Large Comparative Trials

To delineate the relative effectiveness and tolerability of the conventional neuroleptic drugs, the newer atypical antipsychotic drugs, and clozapine, the National Institute of Mental Health in the United States and the National Health Service in Great Britain supported large comparative trials. In the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Phase 1 Schizophrenia trial, 1460 patients with schizophrenia were randomly assigned to blinded treatment with olanzapine (up to 30 mg daily), quetiapine (up to 800 mg daily), risperidone (up to 6 mg daily), ziprasidone (up to 160 mg daily), or perphenazine (up to 32 mg daily). Time to treatment discontinuation for any cause, the primary outcome measure for the trial, was longer for olanzapine

(median 9.2 months) than for quetiapine (4.6 months), risperidone (4.8 months), ziprasidone (3.5 months), or perphenazine (5.6 months), and olanzapine produced the greatest initial reductions in psychopathology. Olanzapine appeared to be more effective than the other drugs studied, and there were no significant differences in effectiveness between the conventional neuroleptic, perphenazine, and the other newer atypical antipsychotic drugs. However, olanzapine was associated with the greatest weight gain and increases in cholesterol, triglycerides, and glycosylated hemoglobin levels. Rates of discontinuation for specific side effects differed; olanzapine was associated with more discontinuation for weight gain or metabolic side effects, and perphenazine was associated with more discontinuation for neurological side effects.

In the CATIE Phase 1b trial, 114 patients who had discontinued treatment with perphenazine in the CATIE Phase 1 trial were randomly assigned to continued blinded treatment with either olanzapine, quetiapine, or risperidone. Median time to discontinuation for any cause was significantly longer with quetiapine (9.9 months) and olanzapine (7.1 months) than with risperidone (3.6 months). Individual sensitivity to neurological side effects likely contributed to discontinuation from perphenazine in the Phase 1 trial; survival in Phase 1b was in keeping with the newer atypical antipsychotic drugs' relative propensities to produce neurological side effects (i.e., quetiapine < olanzapine < risperidone).

In the CATIE Phase 2 ziprasidone trial, patients who had discontinued treatment with a newer atypical antipsychotic drug were randomly assigned to treatment with ziprasidone (50%; this was the newest atypical antipsychotic drug at the time, and had the promise to produce less weight gain and metabolic disturbances than produced by the other newer atypical antipsychotic drugs) or a newer atypical antipsychotic (50%; olanzapine, quetiapine, or risperidone), different from the one they had received in Phase 1. Median times to treatment discontinuation were significantly longer for risperidone (7.0 months) and olanzapine (6.3 months) than for quetiapine (4.0 months) and for ziprasidone (2.8 months). Patients receiving ziprasidone did have the greatest weight loss and reductions in lipid levels.

In the Cost Utility of the Latest Antipsychotic Drugs in Schizophrenia Study (CUtLASS 1), 227 patients with chronic schizophrenia were randomly assigned to 1 year of open-label treatment with either a conventional neuroleptic drug or newer atypical antipsychotic drug (clinicians could select the specific drug they preferred from the assigned class). There was no significant difference between the two drug classes on the primary outcome measure, quality of life scale scores. However, only 31 of 58 (53%) of the patients who received sulpiride (the most commonly prescribed conventional neuroleptic drug; not available in the United States), but 37 of 50 (74%) of the patients who received olanzapine (the most commonly prescribed newer atypical antipsychotic drug), completed the full year of treatment on their initially assigned antipsychotic.

In the CATIE Phase 2 Clozapine trial, 99 patients who prospectively failed to improve with an assigned newer atypical antipsychotic drug in Phase 1 were randomly assigned to clozapine (50%) or to another newer atypical antipsychotic (50%; olanzapine, quetiapine, or risperidone), different from what they had received in Phase 1. Clozapine produced longer median time to treatment discontinuation (10.5 months) than was seen with the newer atypical antipsychotic drugs (range 2.7–3.3 months) and greater reductions in psychopathology.

In the CUtLASS 2 trial, 136 patients with treatment-resistant schizophrenia were randomly assigned to clozapine or one of the newer atypical antipsychotic drugs for up to 1 year. Clozapine-treated patients had significantly greater reductions in psychopathology than did patients treated with a newer atypical antipsychotic.

Summary

The fundamental pathophysiology of schizophrenia remains poorly understood. One manifestation of that pathophysiology appears to be abnormal functioning of dopamine tracts that extend from the midbrain to ramify through limbic structures. These dopamine tracts, like those that extend from the substantia nigra to ramify through movement-related structures, appear to be tone setting. If the latter (motor) tracts have decreased functioning, a Parkinson-like decreased readiness to move ensues; if they function excessively, abnormal, involuntary, disorganized movements intrude into normal activity. Psychotic exacerbations are believed to reflect excessive functioning of the limbic dopamine tracts, resulting in the abnormal, involuntary, disorganized application of limbic weighting (emotional valence, veracity, apparent connectiveness) that intrudes into normal thought, resulting in psychotic psychopathology.

All antipsychotic drugs decrease dopamine neurotransmission. The conventional neuroleptics do this solely by blocking D2 dopamine receptors; the brain concentrations that affect limbic dopamine tracts are very close to the brain concentrations that affect the motor dopamine tracts. If doses of these drugs are restricted to levels at which the first evidence of

Parkinson-like rigidity is detected by careful examination, many patients reap therapeutic benefit without coarse neurological side effects. The newer atypical antipsychotics add antagonism at serotonin 5-HT_2 receptors to D2 antagonism, resulting in fewer propensities for neurological side effects, emotional dulling, and subjective distress. Olanzapine, and to a lesser degree quetiapine and risperidone, produce weight gain and increases in lipids and insulin resistance.

Clozapine and, to a lesser degree, olanzapine offer greater control of psychotic psychopathlogy than do the other antipsychotics. It remains unclear what pharmacological mechanisms contribute to their advantage. It is intriguing that the therapeutic response and weight gain produced by these drugs are correlated, and the possibility that some of their metabolic effects contribute to therapeutic benefit should be investigated. When prescribing clozapine or one of the newer atypical antipsychotic drugs that produce weight gain and metabolic abnormalities, clinicians should document that (a) they have informed patients and their family members of these risks, (b) they monitor weight and fasting glucose and lipids, and, if abnormalities occur, (c) they respond to them reasonably.

See also: Cognitive Deficits in Schizophrenia; Cognitive Dysfunction in Psychiatric Disorders; Dopamine; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopaminergic Agonists and L-DOPA; Neuroleptics; Psychotherapeutic Approaches to Psychiatric Disorders; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Schizophrenia: Genetics.

Further Reading

A Task Force, Report of the American Psychiatric Association (1992). *Tardive Dyskinesia*. Washington, DC: American Psychiatric Association.

Csernansky JG, Mahmoud R, and Brenner R (2002) A comparison of risperidone and haloperidol for the prevention of relapse in patients with schizophrenia. *New England Journal of Medicine* 346: 16–22.

Davis JM, Chen N, and Glick ID (2003) A meta-analysis of the efficacy of second-generation antipsychotics. *Archives of General Psychiatry* 60: 553–564.

Kapur S, Zipursky R, Jones C, et al. (2000) Relationship between dopamine D2 occupancy, clinical response, and side effects: A double-blind PET study of first-episode schizophrenia. *American Journal of Psychiatry* 157: 514–520.

Geddes J, Freemantle N, Harrison P, et al. (2000) Atypical antipsychotics in the treatment of schizophrenia: Systematic overview and meta-regression analysis. *British Medical Journal* 321: 1371–1376.

Grace AA, Bunney BS, Moore H, et al. (1997) Dopamine-cell depolarization block as a model for the therapeutic actions of antipsychotic drugs. *Trends in Neuroscience* 20: 31–37.

Haase HJ and Janssen PAJ (1985) *The Action of Neuroleptic Drugs*. New York: Elsevier Science Publishing Company.

Healy D and Savage M (1998) Reserpine exhumed. *British Journal of Psychiatry* 172: 376–378.

Hirsch SR and Weinberger DR (eds.) (2003) *Schizophrenia, Part Three: Physical Treatments*. Oxford: Blackwell Publishing.

Jones PB, Barnes TR, Davies L, et al. (2006) Randomized controlled trial of the effect an quality of life of second-vs first-generation antipsychotic drugs in schizophrenia (CUtLASS 1). *Archives of General Psychiatry* 63: 1079–1087.

Kane J, Honigfeld G, Singer J, et al. (1988) Clozapine for the treatment-resistant schizophrenic; a double-blind comparison with chlorpromazine. *Archives of General Psychiatry* 45: 789–796.

Leucht S, Barnes TR, Kissling W, et al. (2003) Relapse prevention in schizophrenia with new-generation antipsychotics: A systematic review and exploratory meta-analysis of randomized, controlled trials. *American Journal of Psychiatry* 160: 1209–1222.

Leucht S, Wahlbeck K, Hamann J, et al. (2003) New generation antipsychotics versus low-potency conventional antipsychotics: A systematic review and meta-analysis. *Lancet* 361: 1581–1589.

Lieberman JA, Stroup TS, McEvoy JP, et al. (2005) Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE): Effectiveness of antipsychotic drugs in patients with schizophrenia. *New England Journal of Medicine* 353: 1209–1223.

McEvoy JP, Hogarty GE, and Steingard S (1991) Optimal dose of neuroleptic in acute schizophrenia: A controlled study of the neuroleptic threshold and higher haloperidol dose. *Archives of General Psychiatry* 48: 739–745.

Volavka J, Czobor P, Sheitman B, et al. (2002) Clozapine, olanzapine, risperidone, and haloperidol in the treatment of patients with chronic schizophrenia and schizoaffective disorder. *American Journal of Psychiatry* 159: 255–262.

Anxiety Disorders

J J Benson-Martin, University of Cape Town, Cape Town, South Africa
D J Stein, University of Cape Town, Cape Town, South Africa; and Mount Sinai School of Medicine, New York, NY, USA
E Hollander, Mount Sinai School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Almost a century ago, Freud coined the term 'anxiety neurosis,' explained this condition in terms of unconscious drives, and introduced psychoanalytic treatment. Since then there have been important advances in classifying the anxiety disorders, in delineating their underlying psychobiology, and in providing effective intervention. Advances in neuroscience, including rigorous animal models of anxiety, increased understanding of genetic variation, and functional brain imaging have all contributed to knowledge of the anxiety disorders. Advances in pharmacotherapy and psychotherapy, and increased public awareness of the anxiety disorders, have contributed to better outcomes. In this article we briefly review some of these advances.

Phenomenology

Current diagnostic systems attempt to make a distinction between anxiety as a normal emotion and pathological anxiety. Normal anxiety is universally experienced and is characterized by a sense of apprehension (often diffuse and unpleasant), and by accompanying autonomic phenomena. These phenomena vary between individuals but may take the form of abdominal discomfort, restlessness, perspiration, palpitations, etc. Pathological anxiety, on the other hand, consists of excessive fear, worry, physical arousal, and avoidance. The symptoms of pathological anxiety cause clinically significant distress or impairment in important areas of functioning such as social and occupational functioning.

Modern classification systems list several different anxiety disorders, including generalized anxiety disorder, obsessive–compulsive disorder, panic disorder, social anxiety disorder (or social phobia), and specific phobia (see below). The fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) provides specific diagnostic criteria for each of these conditions and is the most widely used nosological system in anxiety disorders research. The *International Classification of Diseases* (ICD-10) similarly has a section on anxiety disorders, although there are some differences (e.g., obsessive–compulsive disorder is not considered as one of the anxiety disorders).

Epidemiology

Nationally representative epidemiological surveys have indicated that anxiety disorders are the most prevalent of the common psychiatric disorders. They are also among the most disabling and costly to society. For example, the Epidemiological Catchment Area (ECA) study in the United States found that specific phobia was the most common psychiatric disorder and obsessive–compulsive disorder (OCD) was the fourth most common. Subsequently, the National Comorbidity Survey (NCS) and the National Comorbidity Survey-Revised (NCS-R) similarly reported that anxiety disorders are the most prevalent of the psychiatric disorders.

Similar findings have been reported in Europe and in the developing world. For example, a review of 27 epidemiological studies carried out in the EU from 1990 to 2004 found that anxiety disorders are the most frequent psychiatric disorders in Europe with a median 12 month prevalence of 12%. The World Mental Health Survey, a set of epidemiological surveys recently conducted around the globe, has again emphasized the prevalence of the anxiety disorders, as well as their early age of onset, high levels of subsequent comorbidity, and considerable underdiagnosis and undertreatment.

Studies in primary care and clinical settings confirm many of the epidemiological findings. Thus, the anxiety disorders have an early onset and chronic duration, they are often associated with subsequent comorbid conditions (particularly depression, substance use disorders, and other anxiety disorders), and they have a significant negative impact on social life, on occupational function, and on family relations. Given that these disorders are characterized by an early age of onset, they can also impact adversely on normal developmental processes. The high costs of the anxiety disorders are due mostly to indirect costs (e.g., loss of occupational productivity, nonpsychiatric medical treatment) rather than to direct treatment costs. There is a clear need for higher awareness and earlier recognition and intervention.

Diagnosis and Evaluation

DSM-IV and ICD-10

The DSM-IV-TR lists 12 anxiety disorders: (1) panic disorder (PD) with agoraphobia, (2) panic disorder

without agoraphobia, (3) agoraphobia without history of panic disorder, (4) specific phobia, (5) social phobia (SP) (social anxiety disorder), (6) OCD, (7) posttraumatic stress disorder (PTSD), (8) acute stress disorder, (9) generalized anxiety disorder (GAD), (10) anxiety disorder due to a general medical condition, (11) substance-induced anxiety disorder, and (12) anxiety disorder not otherwise specified. A thirteenth category, mixed anxiety-depressive disorder (MAD), is currently in the appendix of *Disorders in Need of Further Study*.

There are specific diagnostic criteria for each of the major anxiety disorders (PD, SP, OCD, PTSD, and GAD). In each disorder, there are similar exclusionary diagnostic criteria which emphasize that the diagnosis should not be found to be caused by the physiological effect of a substance or of a general medical condition. The symptoms should also not be better accounted for by another mental disorder. Lastly, the symptoms must cause clinically significant distress or impairment in important areas of functioning such as social and occupational functioning.

In the ICD-10, anxiety disorders are categorized under the class of neurotic, stress-related, and somatoform disorders. This includes the following: (1) phobic anxiety disorders (agoraphobia, social phobias, and specific phobias); (2) other anxiety disorders (including panic disorder, generalized anxiety disorder, and MAD; (3) OCD; (4) dissociative (conversion) disorders; and (5) somatoform disorders. A further category called reactions to severe stress includes (1) acute stress reaction and (2) PTSD. The ICD-10 is less rigorous in its formulation of operational criteria than the DSM.

Categorical and Dimensional Approaches

Although the DSM system focuses on categories, dimensional approaches to measuring anxiety symptoms may also be useful. Arguably, categorical and dimensional approaches to psychiatric diagnosis and evaluation have a complementary role. For many clinical purposes, it may be useful to know whether a diagnosis is present or absent. However, dimensional approaches may be valuable for describing the way in which symptoms fall to a variable extent on different spectrums, the way in which different symptom spectrums intersect or overlap, and the relationship between symptom spectrums and underlying psychobiological mechanisms.

The tripartite model of anxiety symptoms (Clark and Watson) is one influential dimensional approach to the symptoms of anxiety and depression. The relevant dimensions are (1) negative affect – general distress such as worry, tension, and irritability; (2) positive affect – the level of pleasurable agreement with the environment; and (3) autonomic hyperarousal. The Tripartite model proposes that negative affect is shared by both anxiety and mood disorders, while autonomic hyperarousal is peculiar to anxiety. An absence of positive affect (mainly, anhedonia) differentiates mood disorder from anxiety disorders. Spectrum concepts have also been applied to each of the anxiety disorders.

Rating Scales

A number of self-report and interviewer-administered rating scales have been developed to diagnose and evaluate anxiety disorders. Structured diagnostic instruments allow reliable diagnosis of the anxiety disorders and include the Anxiety Disorders Interview Schedule for DSM-IV-TR. Symptom severity measures allow reliable assessment of the characteristic symptoms of a disorder and include the Yale-Brown Obsessive–Compulsive Scale (for OCD), the Liebowitz Social Anxiety Scale (for SAD), the Clinician Administered PTSD scale (for PTSD), the Hamilton Anxiety Rating Scale (for GAD), and the Panic Disorder Severity Scale (PDSS). The development of reliable symptom measures has contributed to the ability of randomized controlled trials to differentiate between effective medications and placebo.

Neurobiology

Significant advances have been made in understanding the psychobiology of anxiety disorders. Lesion studies have long allowed the study of the neurocircuitry of anxiety disorders, and modern structural and functional brain imaging modalities have provided new insights. Anxiety and its disorders are regulated by a range of neurotransmitters and neurochemical systems, and advances in neurogenetics and neuroproteomics should increasingly allow the specific molecular mechanisms in the neurocircuitry underpinning anxiety disorders to be delineated.

Neurocircuitry

Basic studies of fear have demonstrated that the amygdala and related neurocircuitry play a key role. By virtue of afferent and efferent connections to a range of structures, the amygdala is able to play a key role in mediating the fear response. Sensory input reaches the lateral nucleus of the amygdala, while efferents from the central nucleus include the dorsal motor nucleus of the vagus (parasympathetic activation), the locus coeruleus (release of norepinephrine and autonomic arousal), the hypothalamic–pituitary–adrenal (HPA) axis (neuroendocrine activation), the parabrachial nucleus (hyperventilation), the trigeminal and facial nerve nuclei (fearful facial expression), and the para-aqueductal gray matter (escape or freezing behavior).

A range of other structures are also important in mediating fear responses. The hippocampus plays a

crucial role in memory of the context of fear conditioning. The medial prefrontal cortex may be able to inhibit limbically mediated fear responses, and so likely plays a key role in fear extinction. Thus, it can be hypothesized that structures such as the amygdala, hippocampus, and anterior cingulate play a key role in mediating the various anxiety disorders that present in the clinical setting.

Indeed, early lesion studies noted that damage to the amygdala was associated with a loss of fear responses (Kluver–Bucy syndrome), while temporal epilepsy could be associated with increased anxiety. More recently, functional and structural imaging studies have confirmed amygdala hyperactivation in a number of different anxiety disorders. Hippocampal volume appears to be decreased in a number of clinical conditions, including PTSD, and there is some evidence that reduced volume correlates with memory dysfunction. Successful treatment of the anxiety disorders may be accompanied by changes in frontal circuits involved in the suppression of limbic circuitry.

Neurochemistry

A range of neurochemical systems are found in the neurocircuitry thought to mediate anxiety disorders. Noradrenaline, serotonin, glutamate, and γ-aminobutyric acid (GABA) may play a particularly important role in mediating anxiety, and are certainly targets for a number of anxiolytic medications. Nevertheless, a range of other systems, including dopamine, and various neuropeptides are also likely to play an important role in anxiety symptoms. Here we provide only a brief overview of some of the key neurotransmitter systems that are relevant to the anxiety disorders.

The majority of norepinephrine cell bodies are located in the locus coeruleus in the pons. There is evidence for the overactivity of the norepinephrine system in anxiety disorders such as panic disorder and PTSD. For example, there are increased behavioral, physiologic, and neurochemical responses to intravenous challenge with yohimbine, an α 2-adrenoreceptor antagonist, in these disorders. Effective treatment of these conditions with selective serotonin reuptake inhibitors (SSRIs) may decrease yohimbine sensitivity and enable appropriate regulation of α 2-adrenoreceptor function.

Activation of the 5HT system can produce both anxiogenic and anxiolytic effects. Given the complexity of the serotonergic system, it is difficult to make simple generalizations about the relationship between specific receptors and anxiety levels. Nevertheless, stimulation of $5HT_1$ receptors in the hippocampus appears to be anxiolytic, while stimulation of $5HT_{2A}$ receptors in the forebrain can be anxiogenic. Certainly, a range of clinical research with serotonergic agonists, and with serotonin depletors, has confirmed the importance of this system in the mediation of anxiety disorders. SSRIs may work in part by desensitization of auto-receptors, so allowing increased neurotransmission at the level of the synapse.

Glutamate is the main excitatory neurotransmitter and GABA the main inhibitory neurotransmitter in the mammalian brain. The *N*-methyl-D-aspartate (NMDA)/glutamate receptor may play a role in fear conditioning, while the GABA receptor is a key target of benzodiazepines, which are among the most widely prescribed anxiolytics. Molecular imaging technologies will increasingly allow the assessment of a range of neuroreceptors in patients with anxiety disorders, and preliminary work has already indicated that GABA receptors are decreased in anxiety disorders such as PTSD.

Corticotrophin-releasing hormone (CRH) and cortisol receptors are found not only in the paraventricular nucleus of the hypothalamus but also in multiple areas in the cortex. Both play a central role in the mediation of fear and anxiety. In animal studies, early adversity results in long-lasting changes in HPA axis function. In humans with PTSD, there appears to be overproduction of central CRH, with low levels of plasma cortisol, suggesting enhanced negative feedback. This is particularly interesting in view of prior findings that depression is characterized by increased cortisol, and a blunted central response. There is growing interest in the use of CRH antagonists in the potential treatment of anxiety disorders, and trials of these agents are under way.

Neurogenetics

In animals, particular strains may demonstrate increased levels of anxious and avoidant behavior. In humans, family studies have indicated that the anxiety disorders demonstrate significant heritability. Thus, there is growing interest in delineating the neurogenetics of specific behavioral endophenotypes that underpin susceptibility to particular anxiety disorders. For example, behavioral inhibition is a temperament that may predispose to a number of subsequent anxiety disorders. Genetic studies have suggested that variants in the promoter region of the serotonin transporter gene and HPA axis gene variants may play a role in mediating behavioral inhibition.

There is growing interest in assessing the complex relationships between genes, environments, neurocircuitry, and behavior. For example, subjects with the short allele of the serotonin transporter gene promoter region are more likely to have an increased amygdala response to negative facial expressions on functional magnetic resonance imaging. Similarly, while subjects with this particular gene variant are

slightly more likely to have anxiety symptoms, in the presence of significant life stressors, they are much more likely to develop a depressive disorder. Future work would likely focus on assessing a range of different gene polymorphisms and life events in attempting to predict psychopathology.

Treatment

There have been significant advances in both the pharmacotherapy and psychotherapy of the anxiety disorders in recent decades. There is now a substantial database of randomized controlled trials of both psychotropics and specific psychotherapies. Although there remain important gaps in our knowledge, the field now has effective first-line interventions for the major anxiety disorders. There is also increased understanding of how current therapies are able to normalize the underlying neurocircuitry of the anxiety disorders. In the future, additional work is needed on how best to combine different treatment modalities, long-term treatments, and intervention for special populations such as children and adolescents, and the elderly. There is also growing interest in preventive measures, which may ultimately prove effective in forestalling onset of anxiety disorders.

Pharmacotherapy

Benzodiazepines have long been a mainstay of the pharmacotherapy of anxiety disorders. However, these agents have important limitations, including safety problems, and current guidelines emphasize antidepressants, and the SSRIs in particular, as first-line agents for the treatment of major anxiety disorders. There are multiple studies of SSRIs in each of these disorders, and meta-analysis of this work confirms both efficacy and safety. There are subtle differences in the appropriate dose and duration of SSRI treatment across anxiety disorders. Functional imaging studies have provided information on how these agents normalize the underlying neurocircuitry of anxiety disorders.

Although the introduction of the SSRIs represents an important advance, not all patients respond to these agents. There is only a relatively small database of studies on the pharmacotherapy of the treatment-resistant anxiety disorder patient. Options include switching to a different SSRI, to a different class of antidepressant, or augmenting an antidepressant with an additional agent. In the case of OCD, there is substantial evidence that augmentation of a SSRI with low doses of a dopamine receptor blocker can be effective. There are also relatively few studies of SSRIs in children and adolescents with anxiety disorders, although these are encouraging in terms of indicating efficacy and safety.

Psychotherapy

Although early therapists adopted a psychodynamic approach to treatment of the anxiety disorders, there has been relatively little empirical work to support the efficacy of these interventions. In contrast, later cognitive-behavioral perspectives, have been rigorously studied. A large series of controlled studies supports the recommendation of current treatment guidelines to use cognitive-behavioral therapy (CBT) as a first-line psychotherapy in the management of major anxiety disorders. These interventions emphasize the value of behavioral exposure to anxiety-provoking stimuli, and the importance of restructuring cognitive processes and structures. There are subtle differences in techniques used across the anxiety disorders. Functional imaging studies have provided information on how CBT techniques are able to normalize the underlying neurocircuitry of the anxiety disorders.

The introduction of CBT represents an important advance. Although less studied in younger patients, CBT may also be useful in the treatment of child and adolescent anxiety disorders. Nevertheless, not all patients are willing to subject themselves to this kind of treatment, and not all patients respond to this form of therapy. There is only a relatively small database of studies on the psychotherapy of the treatment-resistant anxiety disorder patient. Options include combining psychotherapy with pharmacotherapy. A fascinating series of recent studies, based on the basic neuroscience of fear conditioning, have combined D-cycloserine, a partial agonist at the NMDA receptor, with CBT, in an attempt to increase the efficacy of CBT techniques.

Novel Treatments

Neurosurgical treatments such as anterior cingulotomy and anterior capsulotomy have been used in patients with severe treatment-refractory OCD. The efficacy of these treatments in some cases is consistent with our growing understanding of the neurocircuitry of OCD (where frontostriatal circuitry appears to play a particularly important role). Transcranial magnetic stimulation (TMS) and deep brain stimulation (DBS) are novel procedures under development for treatment-refractory anxiety disorders, including OCD. Further work is needed before these can be recommended for use in routine clinical practice.

Conclusion

In this article, we have briefly discussed the phenomenology, psychobiology, pharmacotherapy, and

psychotherapy of the anxiety disorders. There have been significant advances in each of these areas of investigation in recent decades; we now understand some of the basic mechanisms underlying the pathogenesis of the anxiety disorders, and we have a range of effective treatments. Nevertheless, much additional work is needed to delineate fully the neurocircuitry and molecular basis of the anxiety disorders. Hopefully, such advances will lead to further improvements in our therapeutic armamentarium.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Contributions to Fear; Anxiety: Drug Therapy; Genetics of Human Anxiety and Its Disorders; Obsessive–Compulsive Disorder; Panic Disorder; Panic Disorder as an Emotional Disorder; Posttraumatic Stress Disorder: Neurobiology; Prefrontal Cortex: Structure and Anatomy; Psychotherapeutic Approaches to Psychiatric Disorders; Stress and Cognition.

Further Reading

Baldwin DS, Anderson IM, Nutt DJ, et al. (2005) Evidence-based guidelines for the pharmacological treatment of anxiety disorders: Recommendations from the British Association for Psychopharmacology. *Journal of Psychopharmacology* 19(6): 567–596.

Bartz JA and Hollander E (2006) Is obsessive-compulsive disorder an anxiety disorder? *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 30(3): 338–352.

Carey PD, Warwick J, Neihaus DJ, et al. (2004) Single photon emission computed tomography (SPECT) of anxiety disorders before and after treatment with citalopram. *BMC Psychiatry* 4: 30.

Caspi A, Sugden K, Moffett TE, et al. (2003) Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* 301: 386–389.

Charney DS (2003) Neuroanatomical circuits modulating fear and anxiety disorders. *Acta Psychiatrica Scandinavica* 108(Supplement 417): 38–50.

Charney DS (2004) Psychobiological mechanisms of resilience and vulnerability: Implications for successful adaptation to extreme stress. *American Journal of Psychiatry* 161: 195–216.

Clark LA and Watson D (1991) Tripartite model of anxiety and depression: Psychometric evidence and taxomic implications. *Journal of Abnormal Psychology* 100(3): 313–336.

Furmark T, Tilfors M, Garpenstrand H, et al. (2004) Serotonin transporter polymorphism related to amygdala excitability and symptom severity in patients with social phobia. *Neuroscience Letters* 362: 189–192.

Furmark T, Tillfors M, Marteindottir I, et al. (2002) Common changes in cerebral blood flow in patients with social phobia treated with citalopram or cognitive-behavioral therapy. *Archives of General Psychiatry* 59: 425–433.

Gorman JM, Kent JM, Sullivan GM, and Coplan JD (2000) Neuroanatomical hypothesis of panic disorder, revised. *American Journal of Psychiatry* 157: 493–505.

Greenberg PE, Sisitsky T, Kessler RC, et al. (1999) The economic burden of the anxiety disorders in the 1990s. *Journal of Clinical Psychiatry* 60: 427–435.

Kessler RC (2001) The epidemiology of pure and comorbid generalized anxiety disorder. A review and evaluation of recent research. *Acta Psychiatrica Scandinavica* 406: 7–13.

Krystal JH, D'Souza DC, Sanacora G, Goddard AW, and Charney DS (2001) Advances in the pathophysiology and treatment of psychiatric disorders: Implications for internal medicine. *Medical Clinics of North America* 85(3): 559–577.

Maier W, Gansicke A, Freyberger HJ, et al. (2000) Generalized anxiety disorder (ICD-10) in primary care from a cross-cultural perspective: A valid diagnostic entity? *Acta Psychiatrica Scandinavica* 101: 29–36.

Mathew SJ, Coplan JD, and Gorman JM (2001) Neurobiological mechanisms of social anxiety disorder. *American Journal of Psychiatry* 158: 1558–1567.

Miller LA, Taber K, Gabbard GO, and Hurley RA (2005) Neural underpinnings of fear and its modulation: Implications for anxiety disorders. *The Journal of Neuropsychiatry and Clinical Neurosciences* 17: 1–6.

Mogotsi M, Kaminer D, and Stein DJ (2000) Quality of life in the anxiety disorders. *Harvard Review of Psychiatry* 8: 273–282.

Starcevic V (2006) Anxiety states: A review of conceptual and treatment issues. *Current Opinion in Psychiatry* 19: 79–83.

Stein DJ (2004) *Clinical Manual of Anxiety Disorders*. Washington, DC: American Psychiatric Pub., Inc.

Yehuda R (2002) Post-traumatic stress disorder. *New England Journal of Medicine* 346: 108–114.

Anxiety: Drug Therapy

C K Haas, A Shekhar, and A W Goddard, Indiana University School of Medicine, Indianapolis, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Anxiety disorders are a significant mental health concern. The National Institute of Mental Health reports that anxiety is the most common mental health problem in the United States today, affecting at least 19 million people. Approximately one in four individuals in the United States reports a lifetime history of at least one anxiety disorder, according to the National Comorbidity Survery, and anxiety disorders may be more chronic than either substance abuse or affective disorders. Greenberg and colleagues found that anxiety disorders afflict 15.7 million people in the United States each year and that 75% of these sufferers (11.7 million) also have at least one comorbid psychiatric condition. In addition, international community surveys, such as the Zurich Cohort Study, the World Health Organization, the World Mental Health 2000 Initiative, and the Netherlands Mental Health Survey and Incidence Study, have yielded comparable lifetime prevalence rates of anxiety disorders. There is a disparity in the distribution of anxiety disorders between men and women: women (30.5% lifetime prevalence) are more likely to have an anxiety disorder than are men (19.2% lifetime prevalence). In addition, anxiety disorders are not limited simply to the adult population; the prevalence of anxiety disorders in youth appears to be quite similar to that reported in adults. The economic costs of anxiety disorders are also quite staggering. In 1998 the United States spent an estimated $\$63.1 \times 10^9$ for costs associated with anxiety disorders. Costs include not only the cost of treatment but also the loss of productivity. All anxiety disorders except specific phobia are associated with lost productivity or absenteeism from work. Because anxiety disorders are common and quite costly, the need for effective treatments is apparent.

This article begins with a discussion of the neural fear circuitry and the known and hypothesized sites of action of the major classes of agents used for the treatment of panic disorder, generalized anxiety disorder (GAD), social anxiety disorder (SAD), posttraumatic stress disorder (PTSD), and obsessive–compulsive disorder (OCD). For each anxiety disorder presented, treatment selection, efficacy, and implementation are discussed.

Mechanisms of Drug Action

Because the fear circuit is so complex, it is not surprising that many neurochemical systems modulate the stress response. Included in this list are the serotonergic (5-hydroxytryptamine; 5-HT), noradrenergic (NE), γ-aminobutyricacidergic (GABAergic), glutamatergic, and peptidergic systems. The neurobiology and neuroanatomy of anxiety and fear are complex and beyond the scope of this article. However, it is important to note that, within the fear circuit, the amygdala serves an important role in coordinating the cognitive, affective, neuroendocrine, cardiovascular, respiratory, and musculoskeletal components of fear and anxiety responses. Not only is the amygdala involved in modulating the fear response; in regard to the cognitive components of fear and anxiety, it is also involved in fear conditioning and fear extinction. The amygdala balances incoming stimuli from the environment presented by the thalamus and sensory cortex with memories of past experience presented by the frontal cortex and hippocampus. The amygdala balances inputs from these areas and then orchestrates the anxiety and panic response by stimulating effector brain areas (locus coeruleus, hypothalamus, periaqueductal gray, and parabracheal nucleus) involved in the expression of fear symptoms. The medications used to treat anxiety disorders are thought to modify dysfunction within this circuit, thereby producing anxiolysis.

Serotonergic Mechanisms

There is evidence that 5-HT mechanisms may have an important role in mediating the anxiolytic effects of several classes of therapeutic agents. Important among these options are the selective serotonin reuptake inhibitors (SSRIs), which are thought to exert their long-term therapeutic effects by enhancement of 5-HT neurotransmission. Successful SSRI treatment of panic disorder and PTSD, disorders in which NE system overactivity has been implicated, could be mediated via selective 5-HT-NE interactions. For example, chronic SSRI administration may enhance function of inhibitory 5-HT neurons projecting from the dorsal raphe nuclei to locus coeruleus-NE neurons, thus reducing cell firing of these NE neurons. In addition, effects of SSRIs in the fear circuit may be mediated through inhibitory effects on glutamatergic neurons that regulate locus coeruleus (LC)–NE neuron excitability. It has also been reported that 5-HT, by activating GABA inhibitory interneurons, can modulate glutamatergic inputs to the amygdala, thus inhibiting glutamatergic activation of the amygdala.

Thus, SSRIs may have anxiolytic effects at the level of the amygdala itself. Serotonin release in the prefrontal cortex is implicated in 'coping' with stress and thus could provide additional pathways to reduce anxiety. Other potential SSRI therapeutic effects within the fear circuit may be via 5-HT/NE and 5-HT/cholecystokinin (CCK) interactions in frontal cortical brain regions. Other serotonergic medications, such as the monoamine oxidase inhibitors (MAOIs), also increase net 5-HT neurotransmission with chronic administration and therefore may have anxiolytic mechanisms similar to those of the SSRIs. In other anxiety disorders, such as OCD, the therapeutic mechanism of the SSRIs and the SRI clomimpramine may involve different pathways. For example, pathophysiologic theories of OCD hypothesize that there are dysfunctional interactions between 5-HT and dopaminergic systems that result in neuronal hyperactivity in the orbitofrontal cortex. The anti-OCD action of SSRIs, therefore, may be partly caused by enhancement of 5-HT-related inhibition of orbitofrontal cortex neurons.

At this time, multiple subclasses of 5-HT receptors have been identified. The different subclasses of 5-HT receptors likely result in the various therapeutic effects and side effect profiles of the serotonergic anxiolytics. Anxiolytic and antidepressant effects seems to be associated with 5-HT_1 receptor stimulation. 5-HT_2 receptor stimulation produces sleep disruption, sexual dysfunction, and an increase in anxiety and agitation. Acute stimulation of 5-HT_{2C} receptors by the projections from the raphe to limbic cortex may cause the acute mental agitation, anxiety, or induction of panic attacks that can be observed during initiation of SSRI therapy. 5-HT_3 receptor stimulation produces nausea, diarrhea, and headache but is also implicated in amygdala synaptic plasticity. All three of these 5-HT receptors are found on the postsynaptic neurons. 5-HT_{1A} receptors are found on the cell body of the presynaptic serotonergic neurons and function as somatodendritic autoreceptors that, when stimulated by 5-HT, exert an inhibitory effect that reduces cell firing and thus reduces the release of 5-HT into the synapse. The 5-HT_{1A} receptors are also found postsynaptically. The 5-HT_{1A} partial agonist buspirone is thought to exert therapeutic effect via action at the presynaptic 5-HT_{1A} receptor. As already mentioned, 5-HT_2 receptor stimulation is thought to produce an increase in anxiety and agitation. Trazodone is a robust 5-HT_{2A} receptor antagonist, and its anxiolytic effect may be due to this antagonism. In addition, stimulation of 5-HT_{2A} receptors by 5-HT may reduce the potency of 5-HT at 5-HT_{1A} receptors. Trazodone may be used to boost SSRI efficacy by blocking 5-HT_{2A} receptors so that they are less efficient in inhibiting cell firing through 5-HT_{1A} action. The atypical antipsychotics may produce anxiolytic effects via a similar mechanism in which 5-HT_{2A} receptor antagonism increases 5-HT transmission by desensitizing 5-HT_{1A} autoreceptors, thus increasing 5-HT release. Trazodone belongs to the chemical class of antidepressants known as phenylpiperazines, which also includes nefazodone. Nefazodone's anxiolytic properties, like trazadone's, may be due to antagonism of the postsynaptic 5-HT_{2A} receptor. Nefazodone differs from trazodone in that nefazodone also weakly blocks the reuptake of 5-HT. Another serotonergic agent with anxiolytic properties is mirtazapine. Mirtazapine's anxiolytic properties are thought to be elicited via blocking postsynaptic 5-HT_{2A1} and 5-HT_{2C} receptors. Mirtazapine also blocks postsynaptic 5-HT_3 receptors, which helps prevent the typical serotonergic side effect of nausea.

Noradrenergic Mechanisms

The LC–NE system coordinates fear-induced sympathetic nervous system activation that results in the anxiety symptoms of tachycardia, increased blood pressure, diaphoresis, piloerection, and pupil dilation. NE dysregulation is a key component of Gray's model of pathological anxiety. This model is based on the concept of 'behavioral inhibition,' a constellation of behaviors seen in animals facing an unpredictable stimulus. When exposed to unexpected stimuli, the septohippocampal system assesses the level of threat the stimuli poses. If appropriate, the behavioral inhibition circuit is activated. Within this circuit, when a stimulus is not deemed significant, the septum suppresses hippocampal activation, acting as an 'all-clear' signal. Increased NE output to the septum inhibits signals from the septum to the hippocampus, resulting in prolonged hippocampal activation. Noradrenergic hyperactivity results in chronic activation of this system and may produce the chronic state of fear equivalent to GAD in humans. Another important site of action for norepinephrine is the amygdala. Norepinephrines, acting via β-adrenergic receptors, have been demonstrated to be critical for formation of fear or emotional memory formation. Thus, hyperactive noradrenergic neurotransmission in the amygdala is more likely to enhance fearful memory formation and may contribute significantly to the development of severe anxiety disorders such as PTSD and agoraphobia. Preclinical evidence has strongly suggested that blocking β-receptors at the time of or immediately after a traumatic stimulus interferes with subsequent fear memory formation. This has recently been supported clinically: Treatment with propranalol within a short period after an acute trauma resulted

in significantly reduced likelihood of developing PTSD. Thus, NE system dysregulation may apply to other chronic anxiety syndromes, such as PTSD, panic disorder with severe anticipatory anxiety, and agoraphobia. Antianxiety agents that have the capacity to regulate NE function in these systems, such as the tricyclic antidepressants, MAOIs, and serotonin norepinephrine reuptake inhibitors (SNRIs), have documented clinical efficacy in the disorders mentioned above. In addition, the novel selective NE reuptake inhibitor reboxetine has been found to be anxiolytic in patients with panic disorder. These agents can inhibit NE function by downregulating β-receptors, reducing LC firing rates, and reducing NE turnover. The decrease in firing rates of NE neurons in the LC by selective NE reuptake inhibitors (NRIs) such as the tricyclic imipramine and reboxetine results from the inhibition of NE transporters on the cell body of NE neurons, leading to an accumulation of synaptic NE in the vicinity of α_2-adrenergic autoreceptors, which exert a negative feedback action on NE neuronal firing. With prolonged treatment and in the presence of NE reuptake inhibition, the firing rate of NE neurons does not recover, because the cell body α_2-adrenergic autoreceptors do not desensitize. This is also the case with prolonged administration of MAOI inhibitors when the α_2-adrenergic autoreceptors fail to desensitize and so there is a decrease in NE neuron firing. Chronic administration of dual reuptake inhibitors of 5-HT and NE produces adaptive changes to 5-HT neuronal firing similar to those of SSRIs and attenuation of NE neuron firing, as seen with NRIs.

GABAergic and Glutamatergic Mechanisms

GABA

As discussed earlier, within the fear circuit, the amygdala is situated as a central structure in the coordination of fear and anxiety responses. GABA, a major inhibitory neurotransmitter found in most parts of the brain, reduces fear in animals after its local infusion into the amygdala. In addition, a variety of pharmacologic agents that increase GABA neurotransmission have similar anxiolytic effects when administered directly into the amygdala. These include the full benzodiazepine (BZD)/$GABA_A$ agonists, the benzodiazepines, the GABAergic anticonvulsants such as sodium valproate, gabapentin, and pregabalin, and partial BZD/$GABA_A$ agonists such as bretazenil and abecarnil. With respect to full BZD/$GABA_A$ receptor agonists, the $gamma_2$ subunit of the $GABA_A$ receptor is necessary for benzodiazepine-mediated anxiolysis, and the other subunit of the $GABA_A$ receptor has profound influence on the efficacy of benzodiazepine modulation. Since preclinical and clinical data implicate decreased benzodiazepine receptor binding in panic disorder and GAD, benzodiazepines may restore homeostasis in the GABA system by correcting these deficits. The GABA system may also influence anxiety levels by mediating the release of other neurotransmitters within the fear circuit (e.g. NE, CCK, 5-HT). In addition, there is evidence that GABA/NE interactions may be an important therapeutic mechanism in panic disorder, as benzodiazepine treatment blocks panicogenesis induced by yohimbine. It is conceivable, then, that within the fear circuit, enhancement of GABA function is a common therapeutic pathway for many classes of effective anxiolytics.

Glutamate

It is well known that glutamate is the primary excitatory neurotransmitter within the brain. Regulation of the amygdala neuronal excitability is achieved by a balance between glutamate-induced excitation and GABA-mediated inhibition. Increased amygdala activation has been associated with various anxiety states. In addition, studies have demonstrated that basal levels of anxiety in rats can be decreased by the suppression of excitatory input to the basolateral amygdala by either glutamate antagonists or lesioning of the basolateral amygdala. Thus modulation of glutamatergic function in the amygdala, and other structures within the fear circuit, is likely to promote anxiolysis. Recent animal studies investigating glutamate receptor antagonists showed that both metabotropic glutamate receptor 5 antagonists MPEP and MTEP and the mGlurR1 antagonist LY456236 produced anxiolytic-like effects similar to those of the benzodiazepine chlordiazepoxide. In rodent models, chronic stress leads to repeated activation of the basolateral amygdala, disrupting the balance of GABAergic inhibition and glutamatergic excitation. This changes the sensitivity of the basolateral amygdala neurons and results in the development of a chronic anxiety-like state. The increased sensitivity of the amygdala may cause nonsalient stimuli to elicit anxiety, fear, or neuroendocrine responses leading to syndromes such as panic and PTSD. It is hypothesized that disruption of the inhibitory tone in the basolateral amygdala and a heightened responsivity of the basolateral amygdala could also result in assigning exaggerated salience to novel social stimuli, resulting in a social anxiety-like condition. Modulation of this system by glutamatergic and GABAergic medications may help restore the normal balance between glutamate excitation and GABA inhibition within the amygdala.

Decreasing glutamatergic excitation within the fear circuit, however, may not be the only glutamatergic mechanism that results in anxiolysis. Glutamate is

critically involved in the neuronal plasticity underlying learning and memory. Recent work on the neural basis of fear extinction suggests that extinction is an active learning process in which glutamate plays a significant role. Evidence suggests fear extinction may be mediated by activation of a hippocampal-prefrontal circuit that inhibits amygdala-generated fear responses. Acute enhancement of glutamate function during extinction learning can facilitate this type of learning. For example, in animal and human paradigms, acute administration of the partial agonist at the glycine site of the NMDA receptor, D-cycloserine, facilitates extinction of fears. Moreover, glutamate mediates neurotoxicity and cell death and is thought to be related to injury of hippocampal neurons in patients with PTSD. The drug lamotrigine, which has been shown to decrease presynaptic release of glutamate, may also help modulate the release of glutamate in anxious patients and thus decrease glutamatergic damage to hippocampal neurons, thereby attenuating symptoms of PTSD.

Corticotropin Releasing Hormone and Peptidergic Systems

Corticotropin-releasing hormone (CRH) is significantly involved in effecting the stress response. During stress, CRH levels are increased in the hypothalamus, resulting in increased release of cortisol and dehydroepiansdrosterone. Equally important are the extrahypothalamic effects of CRH throughout the fear circuit. Neurons containing CRH receptors are located throughout the brain, including the prefrontal and cingulate cortices, central nucleus of the amygdala, bed nucleus of stria terminalis, nucleus accumbens, periaqueductal gray, brain stem nuclei such as the major NE-containing nucleus, the LC, and the serotonin nuclei in the dorsal and median raphe. Recent research suggests that corticotrophin-releasing factor (CRF)-mediated glutamate excitation within the amygdala induces long-term synaptic plasticity and increases the excitability of basolateral amygdala neurons, contributing to the etiology of several anxiety disorders such as panic disorder and PTSD. This line of inquiry has led to CRF antagonist drug development programs with the goal of producing a novel class of anxiolytic. This is not an exhaustive list of peptides; neuropeptide Y, CCK, and orexin are currently being investigated to determine their possible role in anxiogenesis.

Pharmacotherapeutic Strategies

GAD

GAD is characterized by excessive anxiety and worry about several events or activities for most days during at least a 6-month period. Patients with GAD have difficulty controlling the worry, and the worry is associated with somatic symptoms such as muscle tension, irritability, difficulty sleeping, and restlessness. When practitioners select agents to treat GAD, it is important that they consider the risk–benefit ratio for the patient. Patients with low levels of anxiety should be encouraged to reserve medications for when they are clearly needed. Comorbidity tends to be the rule rather than the exception in instances of GAD, and therefore comorbid diagnoses may be crucial in determining treating agents. Major depression (>40–60%) or dysthymia (40%) are the most common disorders comorbid with GAD. Medications with anxiolytic and antidepressant actions would therefore be most appropriate first-line therapies.

The newer 5-HT and 5-HT/NE reuptake inhibitors have anxiolytic and antidepressant efficacy and have become first-line agents for the pharmacologic management of GAD (see Table 1). Paroxetine and escitalopram are equally efficacious in treating GAD; however, one study has found that significantly fewer patients withdrew from escitalopram than from paroxetine treatment because of adverse events. SSRIs are safe, nonaddictive, and generally very well tolerated, with only mild adverse side effects, such as sleep disturbance, headache, nausea, and sexual dysfunction. The delay in onset of therapeutic effect with SSRIs as anxiolytics is usually 2–4 weeks, similar to that seen in SSRI treatment of depression; however, once SSRIs begin to work, the therapeutic effect can be dramatic. Venlafaxine, an SNRI antidepressant, has been shown to be efficacious in both short- and long-term treatment of GAD. Side effects associated with venlafaxine are similar to those of SSRIs.

Because many GAD patients have comorbidity with mood and other anxiety disorders, selection of 5-HT/NE agents as first-line treatment is a good initial strategy over other treatment options such as benzodiazepines and buspirone. Benzodiazepines have a rapid onset of efficacy; however, questions have been raised about their long-term utility. One study has shown that the effect of the benzodiazepine diazepam is not significantly different statistically from that of placebo after the initial 4–6 weeks of treatment. Other concerns about benzodiazepine treatment include difficulty in tapering patients; failure to ameliorate worry (the principal cognitive symptom of GAD); possible exacerbation of depressive symptoms in comorbid patients; dependence issues; and side effects, including sedation, hypnotic effects, and cognitive impairment. Also, a higher rate of recurrence has been suspected following benzodiazepine treatment. Buspirone, a partial 5-HT_{1A} receptor agonist, has been shown to be more efficacious than benzodiazepines in

Table 1 Selective serotonin reuptake inhibitors (SSRIs) and serotonin norepinephrine reuptake inhibitors (SNRIs)

Agent	*Normal dosage range (mg day^{-1})*	*$T_{1/2}$ (h)*	*FDA anxiety indications*	*GAD*[a,b]	*PD*[a]	*SAD*[a]	*OCD*[a]	*PTSD*[a]
SSRIs								
Fluoxetine	20–80	24–72	OCD, PD	C	A	A	A	A
Paroxetine	20–60	21	OCD, PD, GAD, SAD, PTSD	A	A	A	A	A
Sertraline	50–200	24	OCD, PD, SAD, PTSD	A	A	A	A	A
Fluvoxamine	50–300	15	OCD	C	A	A	A	C
Citalopram	20–60	35		C	A	C	A	C
Escitalopram	10–20	27–32	OCD, PD, GAD, SAD, PTSD	A	A	A	A	C
SNRIs								
Venlafaxine	75–225	5–11	Extended release – PD, GAD, SAD	A	A	A	B	B
Duloxetine	40–60	12	GAD	A	C	D	D	C

[a]Level of evidence supporting use in anxiety disorders: A, established treatment (replicated, large-scale randomized clinical trials (RCTs)); B, promising treatment approach (e.g. one or more RCTs of moderate size; single site); C, limited evidence (open trials; case series literature); D, very limited evidence (only preclinical evidence, theoretical prediction, or anecdotal reports).
[b]Limited evidence available evaluating fluoxetine, fluvoxamine, and citalopram in GAD; however, efficacy is likely similar to that of other SSRIs.
FDA, US Food and Drug Administration; GAD, generalized anxiety disorder; PD, panic disorder; SAD, social anxiety disorder; OCD, obsessive–compulsive disorder; PTSD, posttraumatic stress disorder; $T_{½}$, half-life.
Data from Allgulander C, Hackett D, and Salinas E (2001) Venlafaxine extended release (ER) in the treatment of generalized anxiety disorder: Twenty-four-week placebo controlled dose-ranging study. *British Journal of Psychiatry* 179: 15–22; Asnis GM and Kohn SR, Henderson M, and Brown NL (2004) SSRIs versus non-SSRIs in post-traumatic stress disorder: An update with recommendations. *Drugs* 64(4): 383–404; Bielski RJ, Bose A, and Chang CC (2005) A doubleblind comparison of escitalopram and paroxetine in the long-term treatment of generalized anxiety disorder. *Annals of Clinical Psychiatry* 17(2): 65–69; Blomhoff S, Haug TT, Hellstrom K, et al. (2001) Randomised controlled general practice trial of sertraline, exposure therapy and combined treatment in generalized social phobia. *British Journal of Psychiatry* 179: 23–30; Bradwejn J, Ahokas A, Stein DJ, Salinas E, Emilien G, and Whitaker T (2005) Venlafaxine extended-release capsules in panic disorder: Flexible-dose, double-blind, placebo-controlled study. *British Journal of Psychiatry* 187: 352–359; Brady K, Pearlstein T, Asnis GM, et al. (2000) Efficacy and safety of sertraline treatment of posttraumatic stress disorder: A randomized controlled trial. *Journal of the American Medical Association* 283(14): 1837–1844; Brady K, Pearlstein T, Asnis GM, et al. (2000) Efficacy and safety of sertraline treatment of posttraumatic stress disorder: A randomized controlled trial. *Journal of the American Medical Association* 283(14): 1837–1844; Davidson J, Baldwin D, Stein DJ, et al. (2006) Treatment of posttraumatic stress disorder with venlafaxine extended release: A 6-month randomized controlled trial. *Archives of General Psychiatry* 63(10): 1158–1165; Davidson JR, Bose A, and Wang Q (2005) Safety and efficacy of escitalopram in the long-term treatment of generalized anxiety disorder. *Journal of Clinical Psychiatry* 66(11): 1441–1446; Davidson JR, Foa EB, Huppert JD, et al. (2004) Fluoxetine, comprehensive cognitive behavioral therapy, and placebo in general social phobia. *Archives of General Psychiatry* 61: 1005–1013; Davidson JR, Rothbaum BO, van der Kolk BA, et al. (2001) Multicenter, double-blind comparison of sertraline and placebo in the treatment of posttraumatic stress disorder. *Archives of General Psychiatry* 58(5): 485–492; Davidson JRT (2006) Pharmacotherapy of social anxiety disorder: What does the evidence tell us? *Journal of Clinical Psychiatry* 67(supplement 12): 20–26; Gelenberg AJ, Lydiard RB, Rudolph RL, et al. (2000) Efficacy of venlafaxine extended-release capsules in nondepressed outpatients with generalized anxiety disorder: A 6-month controlled trial. *Journal of the American Medical Association* 283: 3082–3088; Goddard AW, Brouette T, Almai A, et al. (2001) Earlly coadministration of clonazepam with sertraline for panic disorder. *Archives of General Psychiatry* 58: 681–686; Goodman WK (2004) Selecting pharmacotherapy for generalized anxiety disorder. *Journal of Clinical Psychiatry* 65(supplement 13): 8–13; Gorman JM (2003) Treating generalized anxiety disorder. *Journal of Clinical Psychiatry* 64(supplement 2): 24–29; Hollander E, Allen A, Steiner M, et al., for the paroxetine OCD Study Group (2003) Acute and long-term treatment and revention of relapse of obsessive-compulsive disorder with paroxetine. *Journal of Clinical Psychiatry* 64: 1113–1121; Hollander E, Koran LM, Goodman WK, et al. (2003) A double-blind, placebo-controlled study of the efficacy and safety of controlled-release fluvoxamine in patients with obsessive-compulsive disorder. *Journal of Clinical Psychiatry* 64: 604–647; Liebowitz MR, DeMartinis NA, Weihs K, et al. (2003) Efficacy of sertraline in severe generalized social anxiety disorder: Resultes of a double-blind, placebo-controlled study. *Journal of Clinical Psychiatry* 64: 785–792; Liebowitz MR, Stein MB, Tancer M et al. (2002) A randomized, double-blind, fixed-dose comparison of paroxetine and placebo in the treatment of generalized social anxiety disorder. *Journal of Clinical Psychiatry* 63(1): 66–74; Lydiard RB, and Bobes J (2000) Therapeutic advances: Paroxetine for the treatment of social anxiety disorder. *Depression and Anxiety* 11: 99–104; Marshall RD, Beebe KL, Oldam M, et al. (2001) Efficacy and safety of paroxetine treatment for chronic PTSD: A fixed-dose, placebo-controlled study. *American Journal of Psychiatry* 158(12): 1982–1988; Marshall RD, Beebe KL, Oldham M, and Zaninelli R (2001) Efficacy and safety of paroxetine treatment for chronic PTSD: A fixed dose, placebo-controlled study. *American Journal of Psychiatry* 158: 1982–1988; Martenyi F and Soldatenkova V (2006) Fluoxetine in the acute treatment and relapse prevention of combat-related post-traumatic stress disorder: Analysis of the veteran group of a placebo-controlled, randomized clinical trial. *European Neuropsychopharmacology* 16(5): 340–349; Martenyi F, Brown EB, Zhang H, Koke SC, and Prakash A (2002) Fluoxetine v. placebo in prevention of relapse in post-traumatic stress disorder. *British Journal of Psychiatry* 181: 315–320; Montgomery SA, Sheehan DV, Meoni P, Haudiquet V, and Hackett D (2002) Characterization of the longitudinal course of improvement in generalized anxiety disorder during long-term treatment with venlafaxine XR. *Journal of Psychiatric Research* 36: 209–217; Pallanti S, Hollander E, and Goodman W (2004) A qualitative analysis of nonresponse: Management of treatment-refractory obsessive-compulsive disorder. *Journal of Clinical Psychiatry* 65(supplement 14): 6–10; Pallanti S, Hollander E, Biestock C, et al. (2002) Treatment non-response in OCD: Methodological issues and operational definitions. *International Journal of Neuropsychopharmacology* 5: 181–191;

ameliorating the cognitive aspects of GAD; however, more-recent research has suggested that, like benzodiazepines, buspirone may not be effective in the long-term treatment of anxiety. Other medications that are efficacious in GAD are imipramine and trazodone, which have 5-HT_{2A} agonist properties. A comparative randomized clinical trial (RCT) in 1993 demonstrated that imipramine and trazodone were as effective as diazepam in treatment of GAD.

Panic Disorder

Panic disorder with or without agoraphobia is characterized by the spontaneous, unexpected occurrence of panic attacks and has a lifetime prevalence of 3.5%. Optimal treatment of panic disorder has the aim of reducing or eliminating panic attacks and improving overall function and quality of life for patients. Of all the anxiety disorders, panic disorder is the most responsive to treatment and has a good prognosis.

When practitioners evaluate pharmacotherapy options for the treatment of panic disorder, it is important that they consider the predicted efficacy and tolerability of the agents as well as drug–drug interactions and safety in overdose. A 1995 meta-analysis showed that the effect size of the three major antipanic drug classes, tricyclic antidepressants, SSRIs, and benzodiazepines, are essentially equal and fall in the medium range of effect size. In addition, 5-HT/NE reuptake inhibitors have demonstrated short-term efficacy in panic disorder in many controlled clinical trials. The tricyclic antidepressants, specifically imipramine, were the gold standard in drug therapy of panic disorder for many years until the development of SSRIs and SNRIs. Approximately 60–70% of patients respond to imipramine; however, tolerability of imipramine is less than that of SSRIs. Unpleasant side effects of tricyclic antidepressants include orthostasis, activation, anticholinergic effects, and weight gain. This heavy side effect burden, as well as the potential for toxicity in overdose, has resulted in the relegation of tricyclic antidepressants to second-line agents for panic disorder.

Currently the SSRIs are considered first-line medication treatments for panic disorder because of their demonstrated efficacy and safety. SNRIs, such as venlafaxine, have demonstrated efficacy in the short-term treatment of panic disorder. In addition, SSRIs and SNRIs treat many of the comorbidities associated with panic disorder, including depression, social phobia, GAD, and PTSD. One drawback to SSRIs and SNRIs is the excessive activation, with potential exacerbation of panic symptoms, during the initiation of these medications – an occurrence presumably related to sudden increase in synaptic level of 5-HT and NE. Compared with tricyclic antidepressants, however, SSRIs have less weight gain and fewer anticholinergic side effects, as well as a comparatively benign cardiovascular profile.

Benzodiazepines are considered third-line agents in the treatment of panic. Despite this fact, they are still a widely used treatment option, with one study reporting that the most common type of treatment for panic disorder remains benzodiazepine monotherapy. In general, the high-potency, short half-life benzodiazepines are preferred now in clinical practice

Phelps NJ and Cates ME (2005) The role of venlafaxine in the treatment of obsessive-compulsive disorder. *Annals of Pharmacotherapy* 39(1): 136–140; Physicians' Desk Reference (2007) 61st edn. Montvale, NJ: Thomson PDR; Pollack MH (2005) The pharmacotherapy of panic disorder. *Journal of Clinical Psychiatry* 66(supplement 4): 23–27; Rickels K, Mangano R, and Khan A (2004) A double-blind, placebo-controlled study of a flexible dose of venlafaxine ER in adult outptatients with generalized social anxiety disorder. *Journal of Clinical Psychopharmacology* 24: 488–496; Rocca P, Fonzo V, Scotta M, et al. (1997) Paroxetine efficacy in the treatment of generalized anxiety disorder. *Acta Psychiatrica Scandinavica* 95: 444–450; Rosenbaum JF, Pollock RA, Jordan SK, Pollack MH (1996) The pharmacotherapy of panic disorder. *Bulletin of the Menninger Clinic* 60(2 supplement A): A54–A75; Sheehan DV (1999) Venlafzine extended release (XR) in the treatment of generalized anxiety disorder. *Journal of Clinical Psychiatry* 60(supplement 22): 23–28; Sheehan DV (2001) Attaining remission in generalized anxiety disorder; venlafaxine extended release comparative data. *Journal of Clinical Psychiatry* 62(supplement 9): 26–31; Sheehan DV (2002) The management of panic disorder. *Journal of Clinical Psychiatry* 63(supplement 14): 17–21; Sinajkic A, Weine S, Djuric-Bijedic Z, et al. (2001) Sertraline, paroxetine, and venlafaxine in refugee posttraumatic disorder with depression symptoms. *Journal of Traumatic Stress* 14(3): 445–452; Stein MB, Fryer AJ, Davidson JRT, et al. (1999) Fluvoxamine treatment of social phobia (social anxiety disorder): A double blind, placebocontrolled study. *American Journal of Psychiatry* 156: 756–760; Stein MB, Liebowitz MR, Lydiard RB, et al. (1998) Paroxetine treatment of generalized social phobia (social anxiety disorder): A randomized controlled trial. *Journal of the American Medical Association* 280: 708–713; Stocchi F, Nordera G, Jokinen RH, et al., for the Paroxetine Generalized Anxiety Disorder Study Team (2003) Efficacy and tolerability of paroxetine for the long-term treatment of generalized anxiety disorder. *Journal of Clinical Psychiatry* 64: 240–258; Tucker P, Zaninelli R, Yehuda R, Ruggiero L, Dillingham K, and Pitts CD (2001) Paroxetine in the treatment of chronic posttraumatic stress disorder: Results of a placebo-controlled, flexible-dosage trial. *Journal of Clinical Psychiatry* 62(11): 860–868; Van Ameringen MA, Lane RM, Walker JR, et al. (2001) Sertraline treatment of generalized social phobia: A 20-week, double-blind, placebo-controlled study. *American Journal of Psychiatry* 158: 275–281; Yaryura-Tobias JA, and Neziroglu FA (1996) Venlafaxine in obsessive–compulsive disorder (letter). *Archives of General Psychiatry* 53: 653–654; and Zohar J, Amital D, Miodownik C, et al. (2002) Double-blind placebo-controlled pilot study of sertraline in military veterans with posttraumatic stress disorder. *Journal of Clinical Psychopharmacology* 22(2): 190–195.

due to simpler metabolism and fewer active metabolites. Multiple studies have documented the efficacy of alprazolam for the treatment of panic disorder. Benzodiazepines have the benefits of a favorable side effect profile and a rapid onset of action, producing improvement within 1 week in one study. On the other hand, drawbacks to benzodiazepines include the development of physiologic dependence with sustained use and the risk of abuse in persons predisposed to substance abuse. In contrast, SSRIs and SNRIs have very low abuse or dependence potential. Recent work has shown that coadministration of a benzodiazepine such as clonazepam with an SSRI such as sertraline accelerates the response of patients with moderate to severe panic disorder and may reduce early antidepressant-related stimulation. Thus benzodiazepines may be used effectively for short-term stabilization and SSRIs for longer-term management of symptoms.

In particularly resistant cases of panic disorder, a tertiary option is the use of MAOIs, which have efficacy that is similar to or exceeds that of tricyclic antidepressants. Anticonvulsants such as valproate, gabapentin, and lamotrigine are considered fourth-line agents for treatment of panic disorder. Small studies have shown efficacy of valproic acid, pregabalin, and gabapentin in panic disorder. If a panic disorder patient has comorbid bipolar disorder, one of the anticonvulsants may be preferable because of the risk of inducing mania with an antidepressant (see Table 2).

SAD

SAD is an anxiety syndrome characterized by excessive fears of humiliation or embarrassment in various social situations. It is generally agreed that two subtypes of SAD exist: the generalized subtype, in which the patient fears a number of performance and social interaction situations, and the specific subtype, in which the fear is confined to performance situations

Table 2 Anticonvulsants

Agent	*Usual daily dose range (mg day^{-1})*[b]	*$T_{1/2}$ (h)*	*GAD*[a]	*PD*[a]	*SAD*[a]	*OCD*[a]	*PTSD*[a]
Divalproex	Trough plasma concentration between 50 and 125 μg ml^{-1}	9–16	D	C	C	D	C
Carbamazepine	Further research needed	25–65 initially, then 12–17[c]	D	C	C	D	C
Topiramate	Further research needed	21	D	E	C	E	C
Gabapentin	Further research needed; 900–1800 mg day^{-1} and given in divided doses (3 times a day)	5–7	D	C	B	C	C
Lamotrigine	Further research needed	25–32	D	C	D	D	B

[a]Level of evidence supporting use in anxiety disorders: A, established treatment (replicated, large-scale randomized clinical trials (RCTs)); B, promising treatment approach (e.g., one or more RCTs of moderate size; single site); C, limited evidence (open trials; case series literature); D, very limited evidence (only preclinical evidence, theoretical prediction, or anecdotal reports); E, evidence of negative therapeutic effects or no evidence of efficacy.
[b]No established doses specifically for anxiety treatment.
[c]Because Tegretol induces its own metabolism, half-life is variable; autoinduction is completed after 3–5 weeks of a fixed dosing regimen. Initial half-life values range from 25–65 h, decreasing to 12–17 h on repeated doses.
GAD, generalized anxiety disorder; PD, panic disorder; SAD, social anxiety disorder; OCD, obsessive–compulsive disorder; PTSD, posttraumatic stress disorder; $T_{1/2}$, half-life.
Data from Asnis GM, Kohn SR, Henderson M, and Brown NL (2004) SSRIs versus non-SSRIs in post-traumatic stress disorder: An update with recommendations. *Drugs* 64(4): 383–404; Bartzokis G, Lu PH, Turner J, Mintz J, and Saunders CS (2005) Adjunctive risperidone in the treatment of chronic combat-related posttraumatic stress disorder. *Biological Psychiatry* 57(5): 474–479; Butterfield MI, Becker ME, Connor KM, et al. (2001) Olanzapine in the treatment of post-traumatic stress disorder: A pilot study. *International Clinical Psychopharmacology* 16: 197–203; Hamner MB, Faldowski RA, Ulmer HG, Frueh BC, Huber MG, and Arana GW (2003) Adjunctive risperidone treatment in post-traumatic stress disorder: A preliminary controlled trial of effects on comorbid psychotic symptoms. *International Clinical Psychopharmacology* 18(1): 1–8; McDougle CJ, Goodman WK, Leckman JF, et al. (2000) A double-blind, placebo-controlled trial of risperidone addition in serotonin reuptake inhibitor-refractory obsessive-compulsive disorder. *Archives of General Psychiatry* 57: 794–801; Pallanti S, Hollander E, and Goodman W (2004) A qualitative analysis of nonresponse: management of treatment-refractory obsessive-compulsive disorder. *Journal of Clinical Psychiatry* 65 (supplement 14): 6–10; Petty F, Brannan S, Casada J, et al. (2001) Olanzapine treatment for posttraumatic stress disorder: An open-label study. *International Clinical Psychopharmacology* 16: 331–337; Physicians' Desk Reference (2007) 61st edn. Montvale, NJ: Thomson PDR; Schmidt AW, Lebel LA, Howard HR Jr., and Zorn SH (2001) Ziprasidone: A novel antipsychotic agent with a unique human receptor biding profile. *European Journal of Pharmacology* 425: 197–201; Shapira NA, Ward HE, Mandoki M, et al. (2004) A double-blind, placebo-controlled trial of olanzapine addition in fuoxetine-refractory obsessive-compulsive disorder. *Biological Psychiatry* 550: 553–555; and Stein MB, Kline NA, and Matloff JL (2002) Adjunctive olanzapine for SSRI-resistant combat-related PTSD: A double-blind, placebo-controlled study. *American Journal of Psychiatry* 159: 1777–1779.

(e.g., public speaking anxiety, writing in front of others). Individuals with generalized SAD report greater anxiety and avoidance than do individuals with specific SAD, and individuals with generalized SAD usually require pharmacotherapy.

SSRIs and the SNRI venlafaxine are considered first-line treatments for SAD. Multiple RCTs have shown paroxetine to produce a significant improvement over placebo in patients with SAD and to reduce phobic avoidance, anticipatory fear, and disability. In addition, numerous studies have demonstrated the efficacy of venlafaxine in the treatment of SAD. At this time there is no evidence indicating that either SSRIs or an SNRI is more efficacious in SAD. Irreversible MAOI agents, such as phenelzine, are effective for the treatment of SAD. However, these drugs are no longer considered first-line therapy because of the necessity of patients following a low-tyramine diet to prevent hypertensive crisis. Evidence also exists supporting the efficacy of the benzodiazepines clonazepam and bromazepam. There is also some evidence suggesting the efficacy of the atypical antipsychotic olanzapine (see Table 3) and $\alpha_2\delta$ calcium-channel blockers such as gabapentin (see Table 2) and pregabalin in treating SAD, but further investigation is needed.

OCD

OCD is a chronic psychiatric disorder characterized by recurrent persistent thoughts (obsession) and/or repetitive compulsory behaviors (compulsions).

Table 3 Atypical antipsychotics

Agent	*Quetiapine*	*Risperidone*	*Olanzapine*	*Ziprasidone*	*Aripiprazole*
Receptor binding profile					
D_2	+	++++	++	++++	++++ (partial agonist)
5-HT_{2A}	+	+++++	++++	+++++	++++
5-HT_{2C}	−	++++	++++	+++++	++
5-HT_{1A}	+	+	−	++++	++++
5-HT_{1D}	−	+	+	++++	N/A
α_1-adrenergic	++	++++	++	++	++
M_1-muscarinic	++	−	++++	−	−
H1-histaminergic	++++	++	++++	++	++
NE reuptake	+	−	−	++	N/A
Usual dosage range (mg day^{-1})[b]	100–400	2–6	5–15	20–80	10–30
$T_{1/2}$(h)	6	20	21–54	7	75–94
GAD[a]	D	D	D	D	D
PD[a]	D	D	C	D	D
SAD[a]	C	D	C	D	D
OCD[a]	B	A	B	D	C
PTSD[a]	C	B	C	D	D

[a]Level of evidence supporting use in anxiety disorders: A, established treatment (replicated, large-scale randomized clinical trials (RCTs)); B, promising treatment approach (e.g., one or more RCTs of moderate size; single site); C, limited evidence (open trials; case series literature); D, very limited evidence (only preclinical evidence, theoretical prediction, or anecdotal reports).
[b]No established doses specifically for anxiety treatment.
GAD, generalized anxiety disorder; PD, panic disorder; SAD, social anxiety disorder; OCD, obsessive–compulsive disorder; PTSD, posttraumatic stress disorder; N/A, not applicable; D_2, dopamine D_2 receptor; (+++++), very high affinity; (++++), high affinity; (++), moderate affinity; (+), low affinity; (−), negligible affinity.
Data from Asnis GM and Kohn SR Henderson M, and Brown NL (2004) SSRIs versus non-SSRIs in post-traumatic stress disorder: An update with recommendations. *Drugs* 64(4): 383–404; Berlant J and van Kammen DP (2002) Open-label topiramate as primary or adjunctive therapy in chronic civilian posttraumatic stress disorder: A preliminary report. *Journal of Clinical Psychiatry* 63(1): 15–20; Davidson JRT (2006) Pharmacotherapy of social anxiety disorder: What does the evidence tell us? *Journal of Clinical Psychiatry* 67 (supplement 12): 20–26; Goodman WK (2004) Selecting pharmacotherapy for generalized anxiety disorder. *Journal of Clinical Psychiatry* 65(supplement 13); 8–13; Hertzberg MA, Butterfield MI, Feldman ME, et al. (1999) A preliminary study of lamotrigine for the treatment of posttraumatic stress disorder. *Biological Psychiatry* 45(9): 1226–1229; McGrath MJ, Campbell KM, Parks CR, and Burton FH (2000) Glutamatergic drugs exacerbate symptomatic behavior in a transgenic model of comorbid Tourette's syndrome and obsessive-compulsive disorder. *Brain Research* 877(1): 23–30; Pallanti S, Hollander E, Biestock C, et al. (2002) Treatment non-response in OCD: Methodological issues and operational definitions. *International Journal of Neuropsychopharmacology* 5: 181–191; Pallanti S, Hollander E, and Goodman W (2004) A qualiative analysis of nonresponse: Management of treatment-refractory obsessive-compulsive disorder. *Journal of Clinical Psychiatry* 65: (supplement 14): 6–10; Pande AC, Pollack MH, Crockatt J, et al. (2000) Placebo-controlled study of gabapentin treatment of panic disorder. *Journal of Clinical Psychopharmacology* 20: 467–471; Physicians' Desk Reference (2007) 61st edn. Montvale, NJ: Thomson PDR; Woodman CL and Noyes R Jr. (1994) Panic disorder: Treatment with valproate. *Journal of Clinical Psychiatry* 55: 134–136; Rosenbaum JF, Pollock RA, Jordan SK, and Pollack MH (1996) The pharmacotherapy of panic disorder. *Bulletin of the Menninger Clinic* 60(2 supplement A): A54–A75.

SRIs have been the primary agents of choice in the pharmacotherapy of OCD, suggesting that the pathophysiology of OCD might be related to disturbance in the functioning of brain serotonin systems (see Table 1). This hypothesis was formulated in the late 1960s, when clinicians observed that the serotonergic tricyclic antidepressant clomipramine relieved OCD symptoms but other tricyclics did not. Since then, clomipramine's efficacy compared with placebo has been confirmed by numerous double-blind controlled trials. Side effects such as dry mouth and constipation are a result of the anticholinergic blockade associated with clomipramine. Orthostatic hypotension is the result of adrenoreceptor blockage. Serotonin-related side effects seen with SSRIs, such as nausea, tremor, impotence, and anorgasmia, are also seen with clomipramine. Safety concerns include risk of seizures at higher doses and prolongation of QT interval.

A large percentage of patients with OCD (40–60%) will be nonresponsive to treatment with SRIs. Various strategies have been studied to treat these patients. Pharmacotherapeutic strategies include switching from a serotonergic agent to one with a different mechanism of action or augmenting with another medication. Addition of conventional and atypical antipsychotics to an SRI in treatment-resistant OCD has also been studied (see Table 3). A 1994 RCT demonstrated efficacy of augmenting fluvoxamine with haldol in patients with OCD. In this study, most of the benefit of haloperidol addition to fluvoxamine occurred in the OCD patients with a coexisting chronic tic disorder. The atypical antipsychotic risperidone has also been studied as an augmenting agent to SRIs and found to be efficacious; patients with and without comorbid tics responded at an equal rate (see Table 3).

It is hypothesized that OCD may be partly due to excessive forebrain glutamate output. This hypothesis has led to the study of glutamatergic agents such as riluzole (initially approved as a neuroprotective agent for patients with amyotrophic lateral sclerosis with an inhibitory effect on glutamate release) and an mGluR2/3 agonist, LY354740, that acts on autoreceptors to inhibit release of glutamate. Preliminary data from an open trial of riluzole augmentation in the treatment of patients with refractory OCD suggest promising efficacy of this agent.

PTSD

The goal of pharmacological therapy in treating PTSD is to reduce symptom distress, strengthen resilience, and restore function. The core symptoms of PTSD are intrusive reexperiencing of the original trauma (e.g., through nightmares and flashbacks), avoidance of stimuli associated with the trauma, numbing, estrangement, anhedonia, and hyperarousal. Secondary symptoms associated with PTSD are impaired functioning, poor resilience stress, and comorbid conditions.

Unfortunately, there are no therapies that are highly effective in chronic PTSD. Currently, SSRIs are the most prescribed drugs for PTSD, based on data from RCTs demonstrating modest efficacy (see Table 1). The percentage of patients who respond to SSRIs is between 40% and 60%. The SNRI venlafaxine has been demonstrated to be effective and well tolerated in the treatment of PTSD in a large ($N = 329$) RCT. The antidepressant nefazodone is unique in that its mechanism is presynaptic serotonin and norepinephrine reuptake inhibition as well as 5-HT_{2A} and 5-HT_{2C} postsynaptic antagonism. A small ($N = 15$), randomized, double-blind, placebo-controlled trial has demonstrated the efficacy of nefazodone in treating patients with PTSD. Nefazodone has demonstrated potent anxiolytic properties in patients with depression, and preliminary studies suggest nefazodone is effective in treating insomnia, making this agent attractive to clinicians treating patients with PTSD. Mirtazapine is an antagonist of postsynaptic 5-HT_2 and 5-HT_3 receptors and a potent antagonist of α_2-autoreceptors, and it has antihistaminergic effects. As a result, mirtazapine has antidepressant, anxiolytic, and sedative properties that have led investigators to study its efficacy in PTSD. At this time, small, open-label studies suggest mirtazapine may be efficacious in PTSD, but confirmation via double-blind, placebo-controlled studies is needed.

The anticonvulsants (carbamazepine, valproic acid, topiramate, lamotrigine, and gabapentin) have therapeutic potential for PTSD. The rationale for their use is that patients with PTSD may have affective instability due to a kindling process at the level of the amygdala and medial temporal lobe. The anticonvulsants have been shown to be efficacious in treating affective instability in epilepsy and in nonepileptic disorders, and many anticonvulsants have antikindling properties. There has been only one double-blind, placebo-controlled study of the efficacy of an anticonvulsant in patients with PTSD (see Table 2). Large-scale controlled studies are needed to confirm these promising reports.

Recent work in the neurobiology of anxiety has led to the hypothesis that the central-acting β-adrenergic receptor antagonist propranolol may be beneficial in preventing PTSD if administered early after exposure to trauma. Propranolol blocks recognition and recall of emotionally significant memories and has been shown to block reconsolidation (the conversion

of labile, short-term memory into long-term memory). One promising smaller RCT has been done evaluating its efficacy in the secondary prevention of PTSD.

See also: Anxiety Disorders; Aversive Emotions: Genetic Mechanisms of Serotonin; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Gamma-Aminobutyric Acid (GABA); Mood Stabilizers; Norepinephrine: CNS Pathways and Neurophysiology; Obsessive–Compulsive Disorder; Panic Disorder; Panic Disorder as an Emotional Disorder; Posttraumatic Stress Disorder: Neurobiology; Posttraumatic Stress Disorder as an Emotional Disorder; Stress, Sex and Adolescent Nicotine Response.

Further Reading

Anand A and Shekhar A (2003) Brain imaging studies in mood and anxiety disorders: Special emphasis on the amygdala. *Annals of the New York Academy of Sciences* 985: 370–388.

Asnis GM, Kohn SR, Henderson M, and Brown NL (2004) SSRIs versus non-SSRIs in post-traumatic stress disorder: An update with recommendations. *Drugs* 64(4): 383–404.

Blier P and Abbot FV (2001) Putative mechanisms of action of antidepressant drugs in affective and anxiety disorders and pain. *Journal of Psychiatry & Neuroscience* 26(1): 37–43.

Blier P and Szabo ST (2005) Potential mechanisms of action of atypical antipsychotic medications in treatment-resistant depression and anxiety. *Journal of Clinical Psychiatry* 66(supplement 8): 30–40.

Boone ML, McNeil W, Masia CL, et al. (1999) Multimodal comparisons of social phobia subtypes and avoidant personality disorder. *Journal of Anxiety Disorders* 13: 271–292.

Brawman-Mintzer O (2001) Pharmacologic treatment of generalized anxiety disorder. *Psychiatry Clinics of North America* 24(1): 119–137.

Bruce SE, Vasile RG, Goisman RM, et al. (2003) Are benzodiazepines still the medication of choice for patients with panic disorder with or without agoraphobia? *American Journal of Psychiatry* 160: 1432–1438.

Charney D and Nestler E (2004) *Neurobiology of Mental Illness,* 2nd edn. Oxford, UK: Oxford University Press.

Davidson JRT (2004) Long-term treatment and prevention of post-traumatic stress disorder. *Journal of Clinical Psychiatry* 65(supplement 1): 44–48.

Davidson JRT (2004) Use of benzodiazepines in social anxiety disorder, generalized anxiety disorder, and posttraumatic stress disorder. *Journal of Clinical Psychiatry* 65(supplement 5): 29–33.

Davidson JRT (2006) Pharmacotherapy of social anxiety disorder: What does the evidence tell us? *Journal of Clinical Psychiatry* 67 (supp 12): 20–26.

Davis M and Myers KM (2002) The role of glutamate and gamma-aminobutyric acid in fear extinction: Clinical implication for exposure therapy. *Biological Psychiatry* 52: 998–1007.

Garakani A, Mathew SJ, and Charney DS (2006) Neurobiology of anxiety disorders and implications for treatment. *Mount Sinai Journal of Medicine* 73(7): 941–949.

Gelpin E, Bonne O, Peri T, et al. (1996) Treatment of recent trauma survivors with benzodiazepines: A prospective study. *Journal of Clinical Psychiatry* 57(9): 390–394.

Goddard AW, Brouette T, Almai A, et al. (2001) Early coadministration of clonazepam with sertraline for panic disorder. *Archives of General Psychiatry* 58: 681–686.

Goddard AW and Charney DS (1997) Toward an integrated neurobiology of panic disorder. *Journal of Clinical Psychiatry* 58(supplement 2): 4–11.

Goodman WK (2004) Selecting pharmacotherapy for generalized anxiety disorder. *Journal of Clinical Psychiatry* 65(supplement 13): 8–13.

Gorman JM (2003) Treating generalized anxiety disorder. *Journal of Clinical Psychiatry* 64(supplement 2): 24–29.

Keck PE, Strawn JR, and McElroy SL (2006) Pharmacologic treatment considerations in co-occurring bipolar and anxiety disorders. *Journal of Clinical Psychiatry* 67(supplement 1): 8–15.

McDougle CJ, Goodman WK, Leckman JF, et al. (2000) A double-blind, placebo-controlled trial of risperidone addition in serotonin reuptake inhibitor-refractory obsessive-compulsive disorder. *Archives of General Psychiatry* 57: 794–801.

Mellman TA, Bustamante V, David D, et al. (2002) Hypnotic medication in the aftermath of trauma. *Journal of Clinical Psychiatry* 63(12): 1183–1184.

Morgan CA III, Krystal JH, and Southwick SM (2003) Toward early pharmacologic posttraumatic stress intervention. *Biological Psychiatry* 53: 834–843.

Nutt D, Bell C, Masterson C, and Short C (2001) *Mood and Anxiety Disorders in Children and Adolescents: A Psychopharmacological Approach*. London: Martin Dunitz.

Pallanti S, Hollander E, Biestock C, et al. (2002) Treatment non-response in OCD: Methodological issues and operational definitions. *International Journal of Neuropsychopharmacology* 5: 181–191.

Pallanti S, Hollander E, and Goodman W (2004) A qualitative analysis of nonresponse: Management of treatment-refractory obsessive-compulsive disorder. *Journal of Clinical Psychiatry* 65(supplement 14): 6–10.

Pitman RK, Sanders KM, Zusman RM, et al. (2002) Pilot study of secondary prevention of posttraumatic stress disorder with propranolol. *Biological Psychiatry* 51: 183–188.

Pollack MH (2005) The pharmacotherapy of panic disorder. *Journal of Clinical Psychiatry* 66(supplement 4): 23–27.

Rainnie DG, Bergeron R, Sajdyk TJ, Patil M, Gehlert DR, and Shekhar A (2004) Corticotrophin releasing factor-induced synaptic plasticity in the amygdala translates stress into emotional disorders. *Journal of Neuroscience* 24(14): 3471–3479.

Roy-Byrne PP, Craske MG, and Stein MB (2002) Panic disorder. *Lancet* 369: 1023–1032.

Sheehan DV (2002) The management of panic disorder. *Journal of Clinical Psychiatry* 63(supplement 14): 17–21.

Shekhar A, Truitt W, Rainnie D, and Sajdyk T (2005) Role of stress, corticotrophin releasing factor (CRF) and amygdala plasticity in chronic anxiety. *Stress* 8(4): 209–219.

Stutzmann GE and LeDoux JE (1999) GABAergic antagonists block the inhibitory effects of serotonin in the lateral amygdala: A mechanism for modulation of sensory inputs related to fear conditioning. *Journal of Neuroscience* 19: RC8.

Westenberg HGM, Boer JAD, and Murphy DL (1996) *Advances in the Neurobiology of Anxiety Disorders*. New York: Wiley.

Apelin

A Reaux-Le Goazigo, X Iturrioz, and C Llorens-Cortes, INSERM U691 Collège de France, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Discovery

The apelin story began in 1993 with the cloning of the complementary DNA for the apelin (APJ) receptor (putative receptor protein related to the type 1 angiotensin receptor) from a human genomic library, which was subsequently cloned in rodents. The human receptor is 380 amino acids long and was identified as a member of the family of orphan seven-transmembrane domain G-protein-coupled receptors (GPCRs). Its amino acid sequence was found to be 31% identical to that of the human type 1 angiotensin (AT1) receptor, but it did not bind radiolabeled angiotensin II (AngII), and stimulation of the rat APJ receptor AngII did not modify cyclic adenosine monophosphate (cAMP) production, demonstrating that this receptor was not an angiotensin receptor subtype. It therefore remained an orphan GPCR for which the endogenous ligand had to be isolated. The APJ receptor was deorphanized in 1998 when apelin, its endogenous ligand, was isolated from bovine stomach tissue extracts.

Structure and Processing of the Apelin Precursor

Apelin is a 36-amino-acid peptide (apelin 36) generated from a 77-amino-acid precursor, proapelin (Figure 1), which has been isolated from various species. The human proapelin gene is located on the X chromosome at locus Xq25–q26.1 and contains three exons, with the coding region spanning exons 1 and 2. The 3′-untranslated region also spans two exons (2 and 3). This may account for the presence of transcripts of two different sizes (~3 and ~3.6 kb) in various tissues. The alignment of proapelin amino acid sequences from cattle, humans, rats, and mice has demonstrated strict conservation of the C-terminal 17 amino acids, known as apelin 17 or K17F. *In vivo*, proapelin gives rise to various molecular forms of apelin, probably through the action of prohormone convertases, due to the presence of pairs of basic residues in proapelin. In rat brain and plasma, the predominant forms of apelin are the pyroglutamyl form of apelin 13 (pE13F) and, to a lesser extent, K17F. In rat lung, testis, and uterus and in bovine colostrum, apelin 36 predominates, whereas in the rat mammary gland, both apelin 36 and pE13F have been detected.

Apelin Receptor Internalization and Signaling Cascades

The rat and human apelin receptors are negatively coupled to adenylate cyclase activity. In Chinese hamster ovary (CHO) cells that stably expressed the human or rat apelin receptor, the most potent inhibitors of forskolin-induced cAMP production were found to be apelin 36, K17F, apelin 13 (Q13F), and pE13F, whereas apelin fragments R10F and G5F were inactive. In an Ala scan (individual amino acids of the primary sequence are replaced with an alanine residue) of pE13F, or in N- or C-terminal deletions of K17F, it was shown that the arginine residues in positions 2 and 4 or the leucine in position 5 in pE13F played a critical role in binding affinity or in the inhibition of cAMP production. Interestingly, apelin 36, K17F, and pE13F are also potent inducers of rat and human apelin receptor internalization and suppress the hypotensive effect of K17F, suggesting that apelin receptor endocytosis is required for initiation of a second wave of signal transduction different from adenylate cyclase coupling and is responsible for the biological effect. Deletion of the C-terminal phenylalanine of K17F abolishes internalization without affecting the adenylate cyclase coupling of the apelin receptor. Apelin 36, K17F, and pE13F also increase intracellular calcium mobilization in both NTera-2 human teratocarcinoma (NT2N) cells, which differentiate into postmitotic neurons following retinoic acid stimulation, and RBL-2H3 cells derived from rat basophils stably expressing the human apelin receptor. In CHO cells expressing the mouse apelin receptor, apelin activates extracellular signal-regulated kinases (ERKs) via a pertussis toxin-sensitive G-protein and Ras-independent pathway. In addition, ERK activation by apelin is mediated by an unidentified isoform of protein kinase C (PKC). Apelin was also shown to activate p70S6 kinase in human umbilical vein endothelial cells (HUVECs) or in CHO cells expressing the mouse apelin receptor via two intracellular cascades, one ERK dependent and the other phosphatidylinositol 3-kinase (PI3K) dependent.

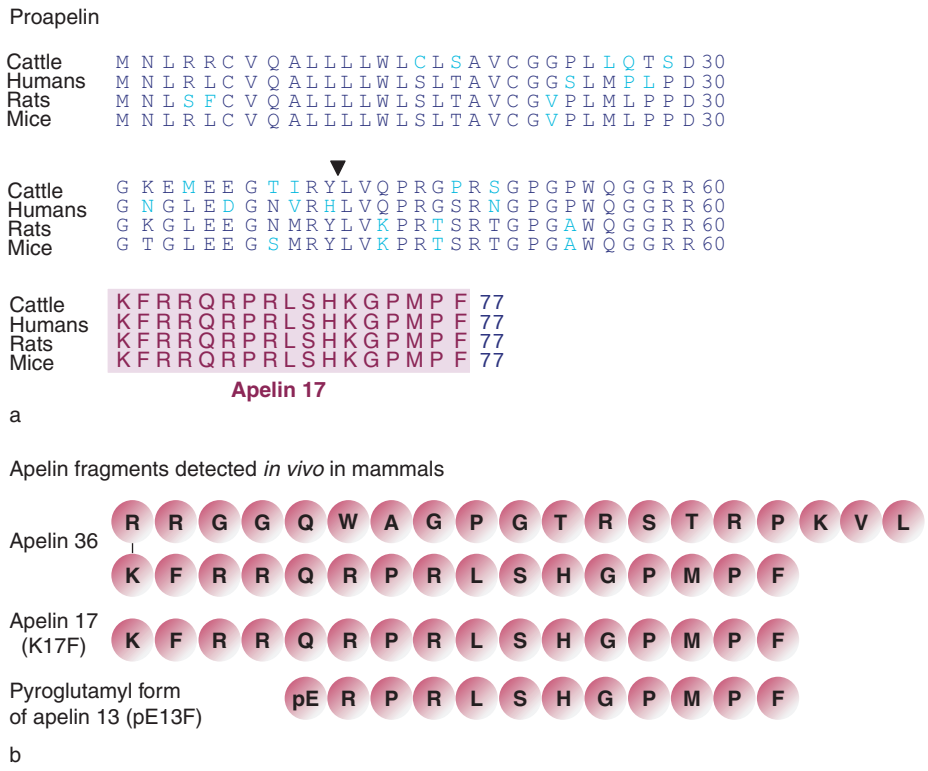

Figure 1 (a) Amino acid sequence of the apelin precursor, proapelin, in cattle, humans, rats, and mice. The first amino acid of apelin 36 is indicated by an arrowhead and the apelin 17 sequence is indicated by pink shading. (b) Apelin fragments detected *in vivo* in mammals: apelin 36, apelin 17 (K17F), and the pyroglutamyl form of apelin 13 (pE13F).

Distribution of Apelin and Its Receptor in Adult Rat Brain

Topographical Distribution of Apelin Immunoreactivity

The precise central topographical distribution of apelin immunoreactivity shows that apelin-immunoreactive (IR) neuronal cell bodies are particularly abundant in the hypothalamus and medulla oblongata, involved in neuroendocrine control, food intake, body fluid homeostasis, and the regulation of cardiovascular functions. These cell bodies are predominantly detected in the hypothalamic supraoptic nucleus (SON) and the magnocellular part of the paraventricular nucleus (PVN), the arcuate nucleus, and the lateral reticular and ambiguus nuclei. Conversely, apelin-IR nerve fibers are much more widely distributed in many brain regions than are neuronal apelin cell bodies. Apelin-IR nerve fibers innervate the mesencephalon, the pons, the medulla oblongata, and several circumventricular organs, such as the vascular organ of the lamina terminalis (OVLT), the subfornical organ (SFO), the subcommissural organ, and the area postrema (Figure 2). The density of IR nerve fibers and apelinergic nerve endings is highest in the inner layer of the median eminence and in the posterior pituitary, suggesting that the apelin neurons of the SON and PVN, like the magnocellular AVP and ocytocin neurons, project into the posterior pituitary via the internal layer of the median eminence. Double immunofluorescence staining confirms this finding, showing that apelin co-localizes with AVP and ocytocin in magnocellular hypothalamic neurons.

Expression of Apelin Receptor mRNA

Apelin receptor mRNA expression, like its ligand, is widely distributed throughout the rat central nervous system. Apelin receptor mRNA is present in the piriform and entorhinal cortices, the nucleus of the lateral olfactory tract, the septum, the hippocampus, and structures containing monoaminergic neuronal cell bodies (pars compacta of the substantia nigra, dorsal raphe nucleus, and locus coeruleus). The apelin receptor is particularly abundant in the apelin-rich hypothalamic nuclei (including the SON, PVN, and arcuate nucleus). The anterior and intermediate lobes of the pituitary are also highly labeled, as well as the pineal gland. Furthermore, like its ligand, apelin receptors (and type 1A and type 1B AVP receptors) are expressed by magnocellular AVP neurons, suggesting an interaction between AVP and apelin (Figure 3).

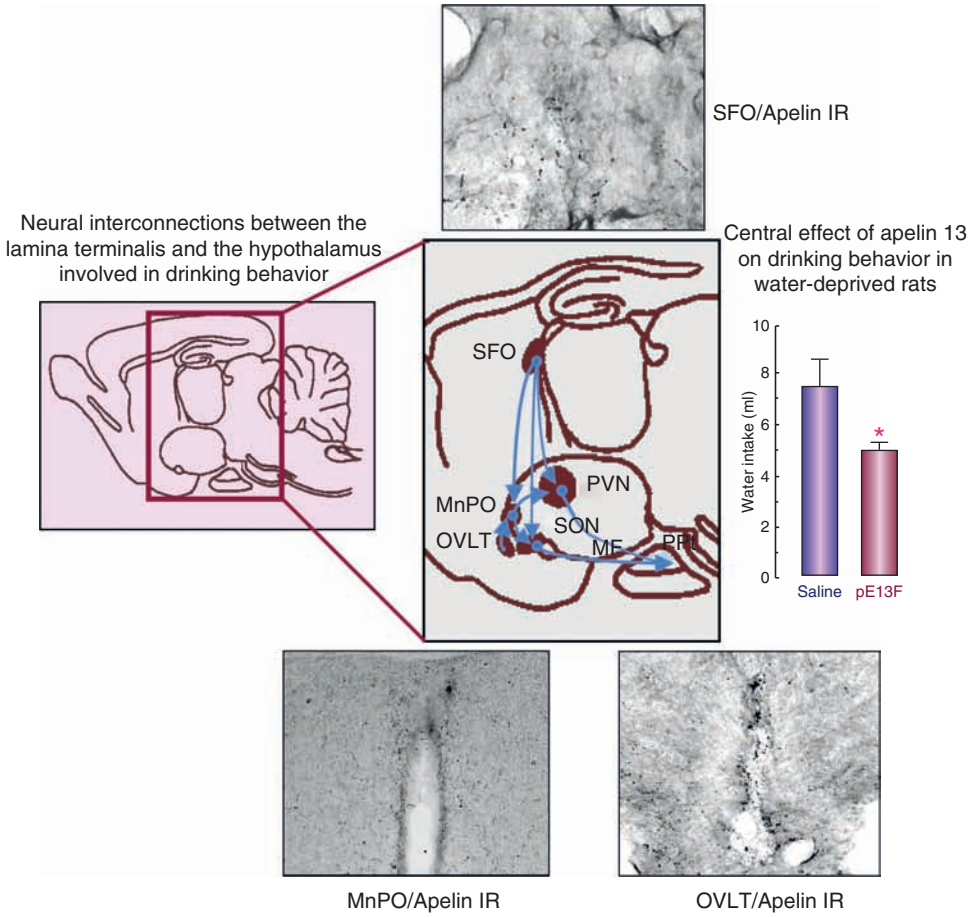

Figure 2 Apelin immunoreactivity in the lamina terminalis: involvement in drinking behavior. Numerous apelin-immunoreactive nerve fibers have been detected in the lamina terminalis, including the subfornical organ (SFO), vascular organ of the lamina terminalis (OVLT), and median preoptic nucleus (MnPO), nuclei involved in the regulation of drinking behavior. Apelin administered into the lateral ventricle in 24 h-water-deprived rats significantly decreased water intake. ME, median eminence; PPIt, posterior pituitary; PVN, hypothalamic paraventricular nucleus; SON, supraoptic nucleus. Adapted from Johnson AK, Cunningham JT, and Thunhorst RL (1996) Integrative role of the lamina terminalis in the regulation of cardiovascular and body fluid homeostasis. *Clinical and Experimental Pharmacological Physiology* 23 (2): 183–191; from Reaux A, De Mota N, Skultetyova I, et al. (2001) Physiological role of a novel neuropeptide, apelin, and its receptor in the rat brain. *Journal of Neurochemistry* 77(4): 1085–1096, Blackwell Publishing and reproduced from Reaux A, Gallatz K, Palkovits M, et al. (2002) Distribution of apelin-synthesizing neurons in the adult rat brain. *Neuroscience* 113(3): 653–662, with permission from Elsevier.

Apelin: Physiological Actions within the Brain and Anterior Pituitary Gland

Involvement of Central Apelin in the Regulation of Food Intake

Apelin-IR cell bodies and nerve fibers have been detected in several hypothalamic nuclei involved in the control of food intake, including the PVN and the suprachiasmatic, arcuate, ventromedial, and dorsomedial nuclei (Figure 4). Whether apelin acts in the brain to modulate feeding behavior is controversial. Central administration of apelin decreases food intake in both fed and starved rats. Similar effects have been reported when apelin is given nocturnally to rats, but the reverse occurs during daytime administration, suggesting a circadian-dependent mechanism of action. In contrast to these findings, another report has shown that intracerebroventricular administration of apelin has either little or no effect on food intake.

Role of Peripheral Apelin in Feeding and Digestion

Apelin precursor mRNA and apelin immunoreactivity have also been detected in the periphery in rats, in the gastrointestinal tract (in neuroendocrine chromogranin A-sensitive cells and in mucosal epithelium). In a murine intestinal enteroendocrine cell line (STC-1 cells) producing and secreting cholecystokinin, apelin stimulates both cholecystokinin secretion and the proliferation of gastric cells. These data suggest that apelin in the gastrointestinal tract may play a role in both exocrine and endocrine functions. Apelin has also been shown to be a novel adipokine released

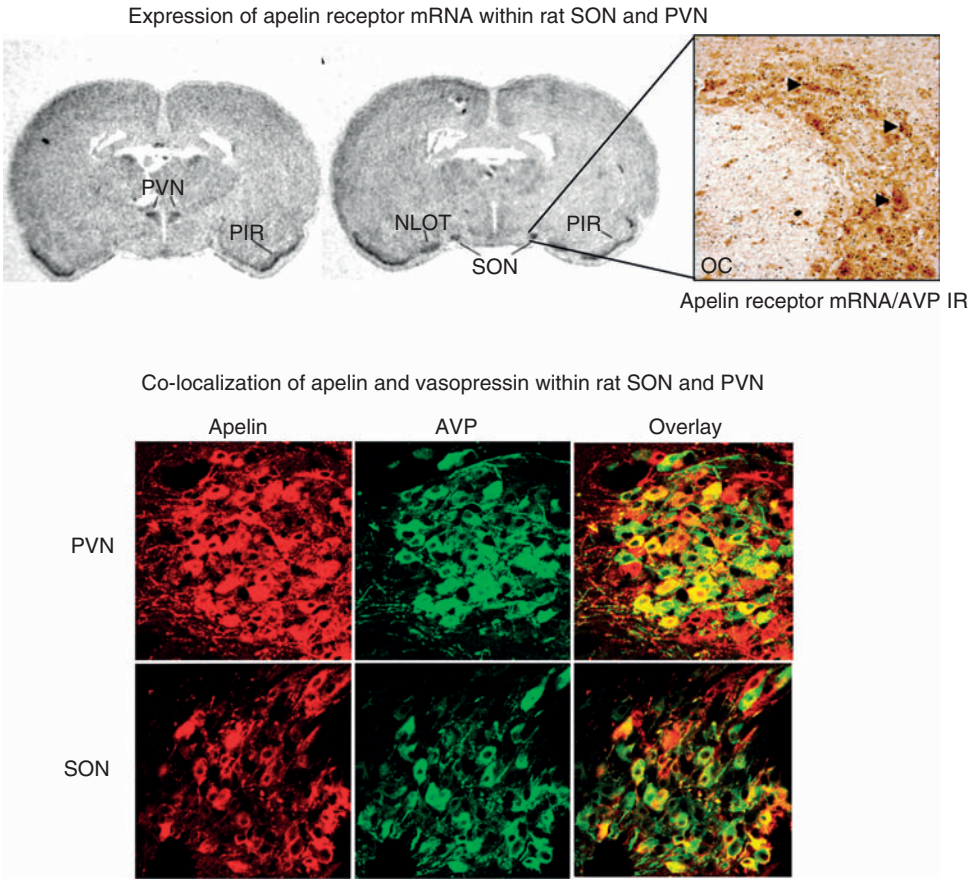

Figure 3 Detection of apelin immunoreactivity and apelin receptor mRNA expression in the rat hypothalamus: co-localization with vasopressin. Distribution of apelin receptor mRNA expression in the rat supraoptic nucleus (SON) and paraventricular nucleus (PVN) by *in situ* hybridization. Apelin receptor mRNA is highly expressed in both nuclei and is synthesized by magnocellular arginine vasopressin-immunoreactive (AVP IR) neurons (arrowhead). Confocal images illustrating the distribution of apelin-immunoreactive cell bodies within the rat SON and PVN. A high co-localization of apelin (red) and AVP (green) has been detected in both SON and PVN magnocellular neurons. PIR, piriform cortex; NLOT, nucleus of the lateral olfactory tract; OC, optic chiasm. Adapted from De Mota N, Lenkei Z, and Llorens-Cortes C (2000) Cloning, pharmacological characterization and brain distribution of the rat apelin receptor. *Neuroendocrinology* 72(6): 400–407 (S. Kargel AG, Basel); from Reaux A, De Mota N, Skultetyova I, et al. (2001) Physiological role of a novel neuropeptide, apelin, and its receptor in the rat brain. *Journal of Neurochemistry* 77(4): 1085–1096, Blackwell Publishing; and from Reaux-Le Goazigo A, Morinville A, Burlet A, et al. (2004) Dehydration-induced cross-regulation of apelin and vasopressin immunoreactivity levels in magnocellular hypothalamic neurons. *Endocrinology* 145(9): 4392–4400, Copyright 2004, The Endocrine Society.

from fat cells and upregulated by insulin and obesity. In experimental animal models of obesity, plasma apelin levels are significantly higher than normal, but only in states associated with hyperinsulinemia.

Apelin appears to play a role in the peripheral and central regulation of food intake. However, given the conflicting nature of the data obtained to date on the central effects, further investigations are required to highlight the mode of action of apelin in the control of this function.

Involvement of Apelin in Regulation of the Hypothalamic–Adrenal–Pituitary Axis

The existence of an apelinergic system within the adult male rat anterior pituitary gland has been recently reported. Apelin is highly co-expressed in corticotrophs and to a much lower extent in somatotrophs, and a high expression of apelin receptor mRNA is also found in corticotrophs. These morphological data suggest a local interaction between apelin and adrenocorticotropic hormone (ACTH). In an *ex vivo* perifusion system of anterior pituitaries, apelin significantly increased basal ACTH release and induced a dose-dependent increase in K^+-evoked ACTH release, outlining the potential role of apelin as an autocrine/paracrine-acting peptide on ACTH release.

Moreover, the detection of apelin-immunoreactive nerve fibers together with apelin receptor mRNA expression in the parvocellular part of the PVN, and the stimulatory action of apelin on corticotropin-releasing

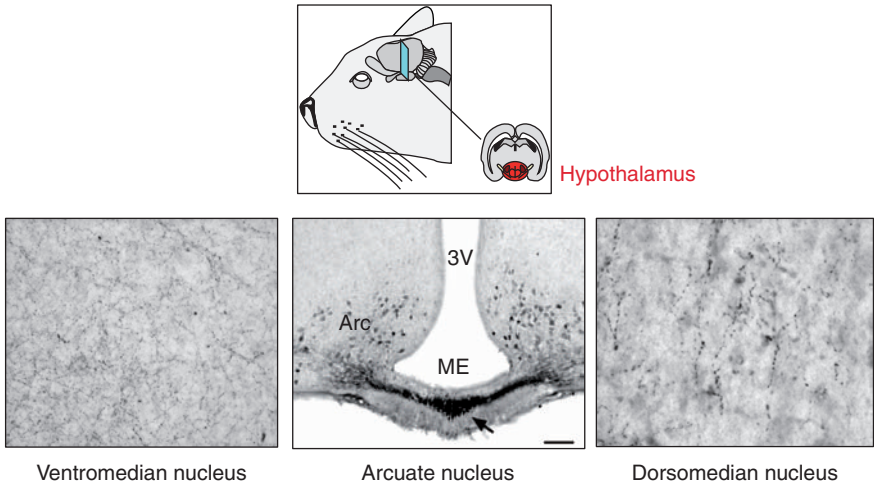

Figure 4 Apelin immunoreactivity in hypothalamic structures involved in the regulation of food intake. The arcuate, ventromedian, and dorsomedian nuclei are brain structures involved in the control of food intake. The presence of numerous apelin-immunoreactive cell bodies in the arcuate nucleus and the presence of numerous apelin nerve fibers in both ventromedian and dorsomedian nuclei suggest that apelin could be involved in the feeding behavior. Adapted from Reaux A, Gallatz K, Palkovits M, et al. (2002) Distribution of apelin-synthesizing neurons in the adult rat brain. *Neuroscience* 113(3): 653–662, with permission from Elsevier.

hormone (CRH) release, indicate that apelin could modulate ACTH release via an indirect action at the level of the hypothalamus. Central injection of apelin is shown to significantly increase plasma ACTH and corticosterone release, at least in part via a stimulatory action on CRH release. Thus, the stimulation of ACTH secretion by apelin occurs by a direct action in the anterior pituitary and by an indirect action via stimulation of CRH release in the hypothalamus. In agreement with the involvement of apelin in the regulation of corticotrophs, it has been reported that in adult rats submitted to acute stress (restraint stress), known to increase the activity of the hypothalamic–adrenal–pituitary (HPA) axis, apelin receptor mRNA expression is increased in the parvocellular division of the PVN. Moreover, dexamethasone, a glucocorticoid agonist, drastically decreases apelin mRNA levels in 3T3-L1 mouse adipocytes, whereas in adrenalectomized rats, apelin receptor mRNA expression is increased, suggesting that glucocorticoids downregulate the expression of apelin and its receptor.

Involvement of Apelin in Maintenance of Body Fluid Homeostasis

The magnocellular neurons release AVP, an antidiuretic vasoconstrictor peptide, into the fenestrated capillaries of the posterior pituitary in response to changes in plasma osmolality and volemia. The recent reports of co-localization of AVP and apelin in the magnocellular neurons of the hypothalamus and the presence of receptors for AVP and apelin on these same neurons suggest a potential apelinergic response to these stimuli.

Independently of the feedback control exerted by AVP on its own release, apelin may also regulate AVP release. This hypothesis has been tested in lactating rats exhibiting a reinforced phasic pattern of AVP neurons during lactation, thereby facilitating systemic AVP release to maintain body water content for optimal milk production. In this model, the intracerebroventricular injection of apelin inhibits the phasic firing activity of AVP neurons, thereby decreasing AVP release into the bloodstream, leading to an increase in aqueous diuresis (Figure 5). Similarly, a marked decrease in systemic AVP release is observed following the intracerebroventricular injection of apelin in mice deprived of water for 24 h, a condition known to increase AVP neuron activity. These data suggest that apelin is probably released from the SON and PVN AVP cell bodies and inhibits AVP neuron activity and release by means of a direct action on the apelin autoreceptors expressed by AVP/apelin-containing neurons. This mechanism probably involves apelin acting as a natural inhibitor of the antidiuretic effect of AVP. The co-localization and opposite biological actions of these two peptides raise questions concerning how these two peptides are regulated to maintain body fluid homeostasis. For the purpose of addressing these questions, the effect of water deprivation on the neuronal content and release of both apelin and AVP were studied.

Water deprivation increases apelin receptor mRNA expression in the magnocellular part of the PVN and SON. Water deprivation largely increases apelin immunoreactivity within both the SON and the PVN and also results in the *de novo* appearance of apelin-IR

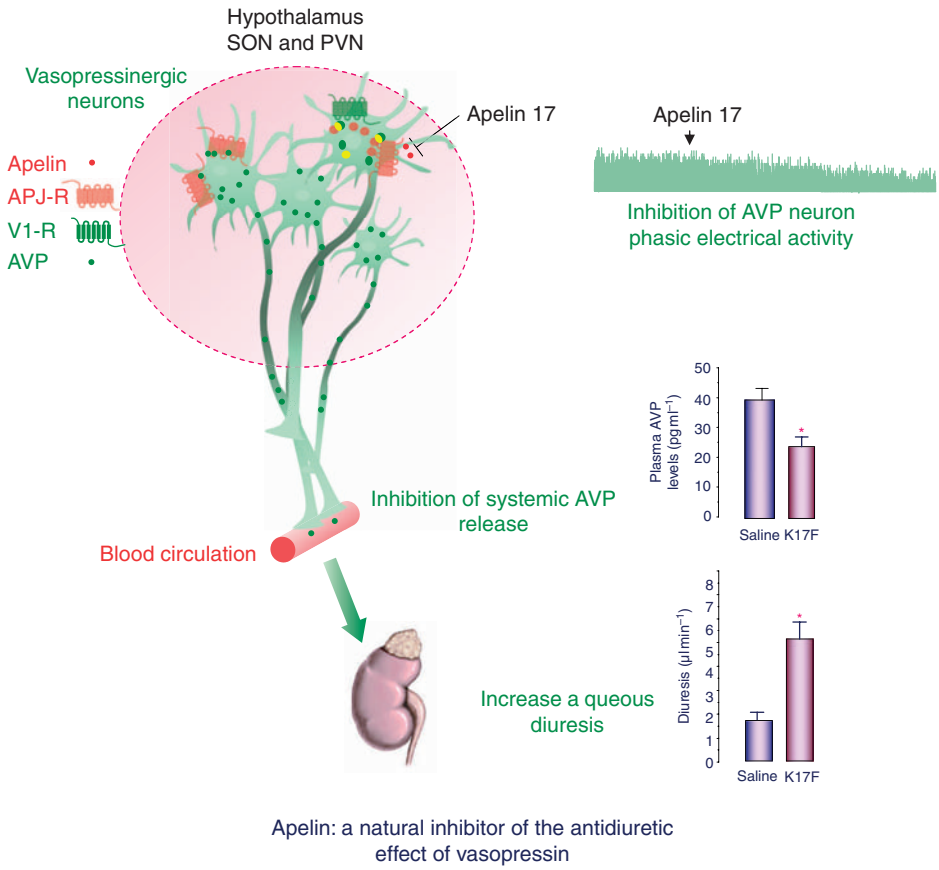

Figure 5 Apelin, a potent diuretic neuropeptide counteracting the effects of arginine vasopressin (AVP) through inhibition of AVP neuron activity and AVP release. In rodents, apelin and its receptor (APJ-R) co-localize with AVP and AVP receptors (V1-R) in the supraoptic nucleus (SON) and paraventricular nucleus (PVN) magnocellular neurons. In lactating animals, the central injection of apelin 17 induces a gradual and sustained inhibition of the phasic electrical activity of AVP neurons, thereby decreasing systemic AVP secretion and increasing aqueous diuresis. Adapted from De Mota N, Reaux-Le Goazigo A, El Messari S, et al. (2004) Apelin, a potent diuretic neuropeptide counteracting vasopressin actions through inhibition of vasopressin neuron activity and vasopressin release. *Proceedings of the National Academy of Sciences of the United States of America* 101: 10464–10469, Copyright (2004), National Academy of Sciences, USA.

neurons in accessory magnocellular nuclei, which contain approximately one-third of the oxytocin and AVP neurons projecting to the posterior pituitary. Among these is the nucleus circularis, which, like the SON and PVN, receives inputs from the SFO, a structure involved in the control of water balance. The increase in apelin immunoreactivity could be due in part to AVP endogenously released in response to dehydration, since chronic treatment with a V1 receptor antagonist significantly reduces the dehydration-induced increase in apelin immunoreactivity in magnocellular neurons.

In rats deprived of water for 24 h, the large increase in hypothalamic apelin content is mirrored by a decrease in plasma apelin levels, suggesting that, under these conditions, apelin accumulates within AVP neurons, rather than being released. The apelin response to dehydration is therefore the opposite of that of AVP, which is released faster than it is synthesized. This interpretation implies that apelin and AVP are released differentially by the magnocellular AVP neurons in which they are produced. Consistent with this hypothesis, high-resolution confocal microscopic images of magnocellular hypothalamic neurons show a marked segregation of apelin and AVP immunoreactivity within SON and PVN neurons. This interpretation is further supported by the presence of apelin-positive/AVP-negative and AVP-positive/apelin-negative varicosities along the same hypothalamo-hypophyseal axons (Figure 6).

These opposite regulatory patterns of apelin and AVP suggest that these molecules act in concert to maintain body fluid homeostasis. During dehydration, increases in the somatodendritic release of AVP optimize the electrical phasic activity of AVP neurons, facilitating the release of AVP into the bloodstream, whereas apelin accumulates in these neurons, rather than being released into the bloodstream and, probably, into the nuclei. Thus, decreases in the local supply of apelin to magnocellular AVP cell bodies may facilitate the expression by AVP neurons of an optimized

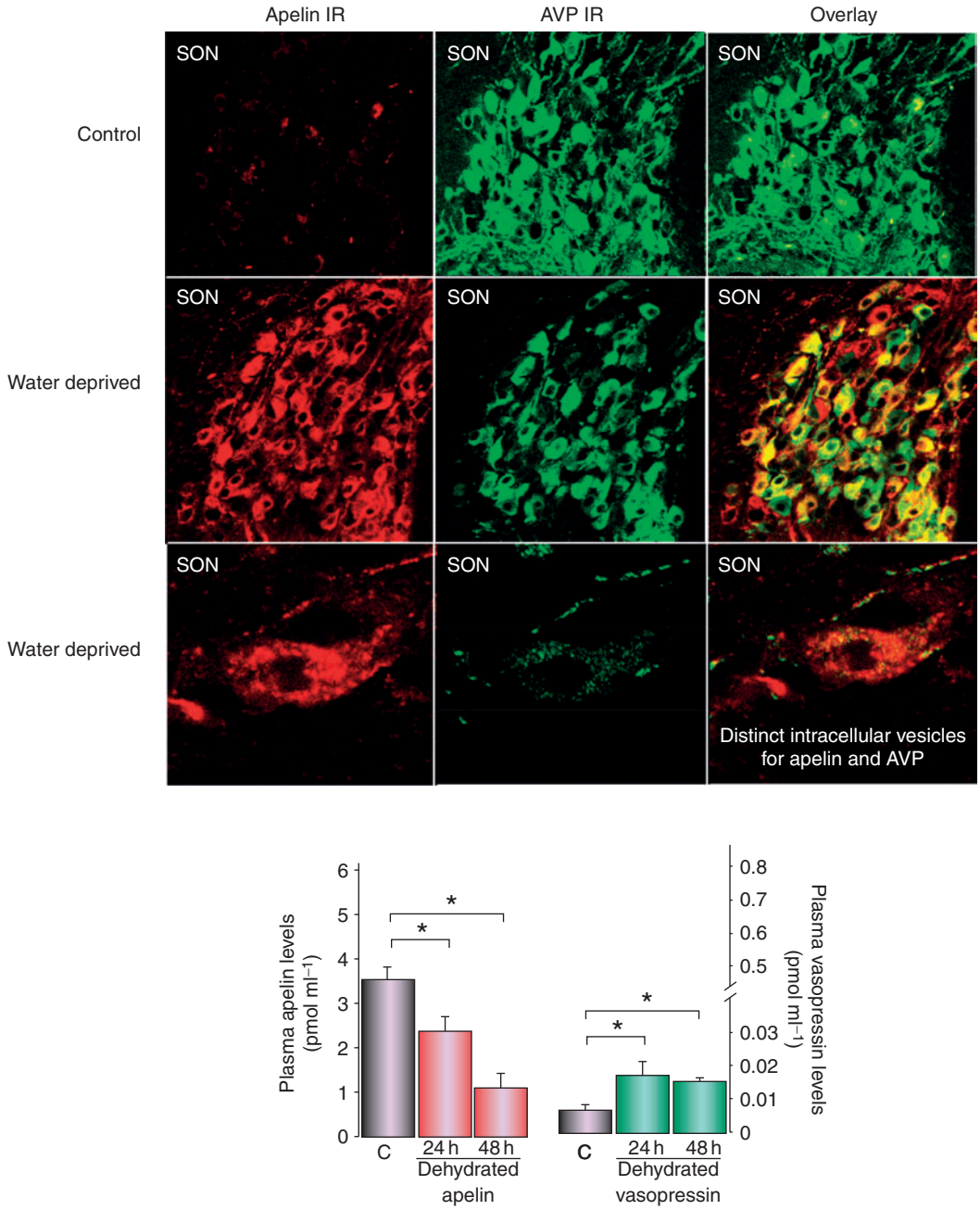

Figure 6 Apelin and vasopressin are conversely regulated during water deprivation. Water deprivation induces a large increase of the number of apelin-immunoreactive cells and in the density of apelin immunofluorescence in the supraoptic nucleus (SON), as compared to control rats with free access to water. In contrast, the number of arginine vasopressin (AVP)-immunoreactive cell bodies is decreased in water-deprived animals. In parallel, water deprivation, which increases systemic AVP release, decreases plasma apelin concentration. The reverse regulation of apelin and AVP during dehydration can occur, since both peptides are detected within a distinct pool of cytoplasmic vesicles. Adapted from De Mota N, Reaux-Le Goazigo A, El Messari S, et al. (2004) Apelin, a potent diuretic neuropeptide counteracting vasopressin actions through inhibition of vasopressin neuron activity and vasopressin release. *Proceedings of the National Academy of Sciences of the United States of America* 101: 10464–10469, Copyright (2004) National Academy of Sciences, USA; and Adapted from Reaux-Le Goazigo A, Morinville A, Burlet A, et al. (2004) Dehydration-induced cross-regulation of apelin and vasopressin immunoreactivity levels in magnocellular hypothalamic neurons. *Endocrinology* 145(9): 4392–4400, Copyright 2004, The Endocrine Society.

phasic activity, by decreasing the inhibitory actions of apelin on these neurons. Antagonistic regulation of apelin and AVP has a biological purpose, making it possible to maintain the water balance of the organism by preventing additional water loss via the kidneys. Consistent with a role for apelin in the control of water balance, which depends not only on AVP secretion regulating renal water fluid loss, but also on the

regulation of water and salt intake, apelin administered centrally clearly and significantly decreases water intake in rats deprived of water for 24 h.

Peripheral Cardiovascular Actions

Apelin also has cardiovascular effects. The mRNA encoding apelin receptors has been detected in the endothelial cells of large conduit arteries, coronary vessels, and endocardium of the right atrium. The injection of apelin into the bloodstream decreases arterial blood pressure (BP), via a mechanism dependent on nitric oxide (NO) production. In normotensive or hypertensive rats, apelin increases the contractile force of the myocardium via a positive inotropic effect. Moreover, apelin receptor knockout mice display an enhanced vasopressor response to systemic AngII, suggesting a counterregulatory action of apelin on AngII. Acute administration of apelin *in vivo* reduces left ventricular preload and afterload while increasing the compliance and contractile reserve of the heart. Chronic treatment with apelin *in vivo* increases cardiac output without causing ventricular hypertrophy. Apelin immunoreactivity has been found to increase in the plasma of patients in the early stages of heart failure and then to decrease during later, more severe stages of heart failure. These data suggest that apelin and its receptor could constitute potential therapeutic targets in the treatment of heart failure. Indeed, the administration of a nonpeptide agonist of the apelin receptor might improve the contractile performance of the myocardium while reducing cardiac loading and increasing aqueous diuresis in patients with heart failure.

Conclusions and Pathophysiological Implications

The identification/discovery of apelin as the endogenous ligand of the orphan APJ receptor constitutes a major advance, both for fundamental research and, potentially, for clinical practice. It demonstrates the validity of the 'deorphanization' approach to orphan receptors for the identification of new bioactive peptides and new potential therapeutic targets. The experimental data obtained to date demonstrate that apelin, by inhibiting the phasic electrical activity of AVP neurons and the systemic secretion of AVP, increases aqueous diuresis. In the periphery, apelin decreases arterial BP and increases the contractile force of the myocardium. Overall, these data show that this new circulating vasoactive neuropeptide may play a key role in the maintenance of water balance and cardiovascular functions. The development of nonpeptide agonists of the apelin receptor, based on the knowledge of the structures of apelin and its receptor, could lead to new therapeutic tools for the treatment of the syndrome of inappropriate secretion of AVP, thirst disorders, and heart and kidney failure.

See also: Angiotensin II; Atrial Natriuretic Peptide: Fluid/Mineral Balance; Blood Pressure: Baroreceptors; Circumventricular Organs; Cyclic AMP (cAMP) Role in Learning and Memory; Invertebrate Neurohormone GPCRs; Osmoregulation; Thirst; Vasopressin/Oxytocin and Receptors.

Further Reading

Ashley EA, Powers J, Chen M, et al. (2005) The endogenous peptide apelin potently improves cardiac contractility and reduces cardiac loading *in vivo*. *Cardiovascular Research* 65 (1): 73–82.

Boucher J, Masri B, Daviaud D, et al. (2005) Apelin, a newly identified adipokine up-regulated by insulin and obesity. *Endocrinology* 146(4): 1764–1771.

De Mota N, Lenkei Z, and Llorens-Cortes C (2000) Cloning, pharmacological characterization and brain distribution of the rat apelin receptor. *Neuroendocrinology* 72: 400–407.

El Messari S, Iturrioz X, Fassot C, et al. (2004) Functional dissociation of apelin receptor signaling and endocytosis: Implications for the effects of apelin on arterial blood pressure. *Journal of Neurochemistry* 90(6): 1290–1301.

Kleinz MJ and Davenport AP (2005) Emerging roles of apelin in biology and medicine. *Pharmacological Therapeutics* 107(2): 198–211.

Lee DK, Cheng R, Nguyen T, et al. (2000) Characterization of apelin, the ligand for the APJ receptor. *Journal of Neurochemistry* 74: 34–41.

Llorens-Cortes C and Beaudet A (2005) Apelin, a neuropeptide that counteracts vasopressin secretion. *Medecine et Sciences (Paris)* 21(8–9): 741–746.

Masri B, Knibiehler B, and Audigier Y (2005) Apelin signalling: A promising pathway from cloning to pharmacology. *Cell Signaling* 17(4): 415–426.

O'Carroll AM and Lolait SJ (2003) Regulation of rat APJ receptor messenger ribonucleic acid expression in magnocellular neurones of the paraventricular and supraoptic nuclei by osmotic stimuli. *Journal of Neuroendocrinology* 15(7): 661–666.

O'Dowd BF, Heiber M, Chan A, et al. (1993) A human gene that shows identity with the gene encoding the angiotensin receptor is located on chromosome 11. *Gene* 136: 355–360.

Szokodi I, Tavi P, Foldes G, et al. (2002) Apelin, the novel endogenous ligand of the orphan receptor APJ, regulates cardiac contractility. *Circulation Research* 91(5): 434–440.

Taheri S, Murphy K, Cohen M, et al. (2002) The effects of centrally administered apelin-13 on food intake, water intake and pituitary hormone release in rats. *Biochemical and Biophysical Research Communications* 291: 1208–1212.

Tatemoto K, Hosoya M, Habata Y, et al. (1998) Isolation and characterization of a novel endogenous peptide ligand for the human APJ receptor. *Biochemical and Biophysical Research Communications* 251: 471–476.

Aphasia: Sudden and Progressive

M-M Mesulam, Northwestern University Medical School, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Aphasia is an acquired disorder of language caused by brain damage. It is diagnosed when deficits are detected in naming, word choice, word comprehension, spelling, or syntax. Dysarthria and mutism do not, by themselves, lead to a diagnosis of aphasia. In approximately 90% of right handers and 60% of left handers, aphasia occurs only after lesions of the left hemisphere. For the majority of the population, the left hemisphere is therefore said to be dominant for language function. In some individuals no hemispheric dominance for language can be discerned, and in others (including a small minority of right handers) there is a right hemisphere dominance for language. A language disturbance occurring after a right hemisphere lesion in a right hander is called crossed aphasia.

Language is controlled by a large-scale distributed network, usually located within the left hemisphere, revolving around two perisylvian nodes. One extends into the temporoparietal junction and is known as Wernicke's area; the other extends into the inferior frontal gyrus and is known as Broca's area. These two areas are connected with each other as well as with multiple regions of the temporal, parietal, and frontal lobes. Wernicke's area can be considered the semantic-lexical pole of the language network, whereas Broca's area can be considered its syntactic-phonological pole. These specializations are relative rather than absolute. The network as a whole links sensory patterns corresponding to the words we hear and read into the distributed associations that encode their meaning. Damage to the language network can therefore have two major consequences. In some patients, the meanings of words cannot be decoded. In others, thoughts and experiences cannot be translated into statements that have the appropriate syntactic structure or semantic content. The location of the damage within the language network determines the type of aphasia experienced by the patient.

There are two major groups of acquired aphasias: those caused by cerebrovascular accidents and those caused by degenerative diseases. Aphasias caused by cerebrovascular accidents start suddenly and display maximal deficits at the onset. The underlying lesion is relatively circumscribed and associated with a total loss of neural function at the lesion site. These are the 'classic' aphasias in which relatively reproducible relationships between lesion site and aphasia pattern can be discerned. Aphasias caused by neurodegenerative diseases have an insidious onset and relentless progression so that the symptomatology changes over time. Since the neuronal loss within the areas encompassed by the neurodegeneration is partial and since it tends to include multiple components of the language network, distinctive clinical patterns and clinico-anatomical correlations are less obvious.

Clinical Examination

The clinical examination of language should include the assessment of naming, spontaneous speech, comprehension, repetition, reading, and writing. A deficit of naming (anomia) is the single most common finding in aphasic patients. When asked to name common objects (e.g., a pencil or wristwatch), the patient may fail to come up with the appropriate word, may provide a circumlocutious description of the object ('the thing for writing'), or may come up with the wrong word (paraphasia). If the patient offers an incorrect but legitimate word ('pen' for 'pencil'), the naming error is known as a semantic paraphasia; if the word approximates the correct answer but is phonetically inaccurate ('plentil' for 'pencil'), the error is known as a phonemic paraphasia. Asking the patient to name body parts, geometric shapes, and component parts of objects (e.g., lapel of coat, cap of pen) can elicit mild forms of anomia in patients who can otherwise name common objects. In most anomias, the patient cannot retrieve the appropriate name when shown an object but can point to the appropriate object when the name is provided by the examiner. This is known as a one-way (or retrieval-based) naming deficit. A two-way naming deficit exists if the patient can neither provide nor recognize the correct name, indicating the presence of a word comprehension impairment.

Spontaneous speech is described as fluent if it maintains appropriate output volume, phrase length, and melody or as nonfluent if it is sparse and halting and if average utterance length is less than four words. The examiner should also note whether the speech is paraphasic or circumlocutious; whether it shows a relative paucity of substantive nouns and action verbs compared to function words (prepositions, conjunctions); and whether word order, tenses, suffixes, prefixes, plurals, and possessives are used appropriately. Comprehension can be tested by assessing the patient's ability to follow the conversation, by asking yes/no

questions ('Can a dog fly?,' 'Does it snow in summer?'), or by asking the patient to point to appropriate objects ('Where is the source of illumination in this room?'). Statements with embedded clauses or passive voice construction ('If a tiger is eaten by a lion, which animal stays alive?') help to assess the patient's ability to comprehend complex syntactic structure. Commands to close or open the eyes, stand up, sit down, or roll over should not be used to assess overall comprehension since appropriate responses aimed at such axial movements can be preserved in patients who otherwise have profound comprehension deficits.

Repetition is assessed by asking the patient to repeat single words, short sentences, or strings of words such as 'No ifs, ands, or buts.' The testing of repetition with tongue-twisters such as 'hippopotamus' or 'Irish constabulary' provides a better assessment of dysarthria than aphasia. Aphasic patients may have little difficulty with tongue-twisters but have a particularly hard time repeating a string of function words. It is important to make sure that the number of words does not exceed the patient's attention span. Otherwise, the failure of repetition becomes a reflection of the narrowed attention span (verbal working memory) rather than an indication of an aphasic deficit. Reading should be assessed for deficits in reading aloud as well as comprehension. Writing is assessed for spelling errors, word order, and grammar. Alexia describes an inability to read or comprehend written words; agraphia (or dysgraphia) is used to describe an acquired deficit in the spelling or grammar of written language.

Aphasias of Cerebrovascular Origin

Aphasias of cerebrovascular origin can be divided into central syndromes, which result from damage to the epicenters of the language network (Broca's and Wernicke's areas), and disconnection syndromes, which arise from lesions that interrupt the functional connectivity of these centers with each other and with the other components of the language network. The syndromes outlined in the following sections are idealizations; pure syndromes occur rarely.

Wernicke's Aphasia

In Wernicke's aphasia, comprehension is impaired for spoken and written language. Language output is fluent but is highly paraphasic and circumlocutious. The tendency for paraphasic errors may be so pronounced that it leads to strings of neologisms, which form the basis of what is known as jargon aphasia. Speech contains large numbers of function words (e.g., prepositions, conjunctions) but few substantive nouns or verbs that refer to specific actions. The output is therefore voluminous but uninformative. The patient does not seem to realize that his or her language is incomprehensible and may appear angry and impatient when the examiner fails to decipher the meaning of a severely paraphasic statement. Patients with Wernicke's aphasia cannot express their thoughts in meaning-appropriate words and cannot decode the meaning of words in any modality of input. This aphasia therefore has expressive as well as receptive components. Repetition, naming, reading, and writing are also impaired. The lesion site most commonly associated with Wernicke's aphasia is in the posterior portion of the language network and tends to involve at least parts of Wernicke's area.

Broca's Aphasia

In Broca's aphasia, speech is nonfluent, labored, interrupted by many word-finding pauses, and usually dysarthric. It is impoverished in function words. Abnormal word order and the inappropriate deployment of bound morphemes (word endings used to denote tenses, possessives, or plurals) lead to a characteristic agrammatism. Speech is telegraphic and pithy but quite informative. Output may be reduced to a grunt or single word ('yes' or 'no'), which is emitted with different intonations in an attempt to express approval or disapproval. In addition to fluency, naming and repetition are also impaired. Comprehension of spoken language is intact, except for syntactically difficult sentences with passive voice structure or embedded clauses.

Global Aphasia

In global aphasia, speech output is nonfluent, and comprehension of spoken language is severely impaired. Naming, repetition, reading, and writing are also impaired. This syndrome represents the combined dysfunction of Broca's and Wernicke's areas and usually results from strokes that involve the entire middle cerebral artery distribution in the left hemisphere. Most patients are initially mute or say a few words, such as 'hi' or 'yes.'

Conduction Aphasia

In conduction aphasia, speech output is fluent but paraphasic, comprehension of spoken language is intact, and repetition is severely impaired. Naming and writing are also impaired. Reading aloud is impaired, but reading comprehension is preserved. The lesion sites spare Broca's and Wernicke's areas but may induce a functional disconnection between the two such that neural word representations formed in Wernicke's area and adjacent regions cannot be transmitted to

Broca's area for assembly into corresponding articulatory patterns. Occasionally, a Wernicke's area lesion gives rise to a transient Wernicke's aphasia that rapidly resolves into a conduction aphasia. The paraphasic output in conduction aphasia interferes with the ability to express meaning, but this deficit is not nearly as severe as the one displayed by patients with Wernicke's aphasia.

Nonfluent Transcortical Aphasia (Transcortical Motor Aphasia)

The features of nonfluent transcortical aphasia are similar to those of Broca's aphasia, but repetition is intact and agrammatism may be less pronounced. The lesion site disconnects the intact language network from prefrontal areas of the brain and usually involves the anterior watershed zone between anterior and middle cerebral artery territories or the supplementary motor cortex in the territory of the anterior cerebral artery.

Fluent Transcortical Aphasia (Transcortical Sensory Aphasia)

The clinical features of fluent transcortical aphasia are similar to those of Wernicke's aphasia, but repetition is intact. The lesion site disconnects the intact core of the language network from other temporoparietal association areas. Infarctions in the posterior watershed zone are common causes.

Isolation Aphasia

Isolation aphasia, a rare syndrome, represents a combination of the two transcortical aphasias. Comprehension is severely impaired, and there is no purposeful speech output. The patient may parrot fragments of heard conversations (echolalia), indicating that the neural mechanisms for repetition are at least partially intact. This condition represents the pathologic function of the language network when it is isolated from other regions of the brain. Broca's and Wernicke's areas tend to be spared, but there is damage in surrounding frontal, parietal, and temporal cortex. Lesions are patchy and can be associated with anoxia, carbon monoxide poisoning, or complete watershed zone infarctions.

Anomic Aphasia

Anomic aphasia may be considered the 'minimal dysfunction' syndrome of the language network. Articulation, comprehension, and repetition are intact, but confrontation naming, word finding, and spelling are impaired. Speech is enriched in function words but impoverished in substantive nouns and verbs denoting specific actions. Language output is fluent but paraphasic, circumlocutious, and uninformative. The lesion sites can be anywhere within the left hemisphere language network, including the middle and inferior temporal gyri.

Pure Word Deafness

In pure word deafness, the most common lesions are either bilateral or left-sided in the superior temporal gyrus. The net effect of the underlying lesion is to interrupt the flow of information from the unimodal auditory association cortex to the language network. Patients have no difficulty understanding written language and can express themselves well in spoken or written language. They have no difficulty interpreting and reacting to environmental sounds since primary auditory cortex and subcortical auditory relays are intact. Since auditory information cannot be conveyed to the language network, however, the patient reacts to speech as if it were in an alien tongue that cannot be deciphered. Patients cannot repeat spoken language but have no difficulty naming objects.

Pure Alexia without Agraphia

Pure alexia without agraphia is the visual equivalent of pure word deafness. The lesions (usually a combination of damage to the left occipital cortex and to a posterior sector of the corpus callosum known as the splenium) interrupt the flow of visual input into the language network. There is usually a right hemianopia, but the core language network remains unaffected. The patient can understand and produce spoken language, name objects in the left visual hemifield, repeat, and write. However, the patient acts as if illiterate when asked to read even the simplest sentence because the visual information from the written words (presented to the intact left visual hemifield) cannot reach the language network. Objects in the left hemifield may be named accurately because they activate nonvisual associations in the right hemisphere, which, in turn, can access the language network through transcallosal pathways anterior to the splenium. Patients with this syndrome may also lose the ability to name colors, although they can match colors. This is known as a color anomia. The most common etiology of pure alexia is a vascular lesion in the territory of the posterior cerebral artery.

Neurodegenerative Aphasia

The language network can also become the target of dementia-causing neurodegenerative diseases. Dementia is a generic term used to designate a neurodegenerative disease that impairs intellect and behavior to the point at which customary activities of daily

living become compromised. Alzheimer's disease is the single most common cause of dementia. The neuropathology of Alzheimer's disease causes the earliest and most profound neuronal loss in memory-related parts of the brain such as the entorhinal cortex and the hippocampus. This is why progressive forgetfulness for recent events and experiences is the cardinal feature of Alzheimer's disease. In time, the neuronal pathology in Alzheimer's disease spreads to the language network, and a progressive aphasia becomes added to the progressive amnesia. There are other patterns of dementia, however, in which neurodegeneration initially targets the language rather than the memory network of the brain, leading to the emergence of a progressive aphasia that becomes the most prominent aspect of the clinical picture during the initial phases of the disease. Primary progressive aphasia (PPA) is the most widely recognized syndrome with this pattern of selective language impairment.

Clinical Presentation and Diagnosis

The patient with PPA comes to medical attention because of word-finding difficulties, abnormal speech patterns, and spelling errors of recent onset. PPA is diagnosed when other mental faculties such as memory for daily events, visuospatial skills (assessed by tests of drawing and face recognition), and comportment (assessed by history obtained from a third party) remain relatively intact, when language is the only major area of dysfunction for the first few years of the disease, and when structural brain imaging does not reveal a specific lesion, other than atrophy, to account for the language deficit. Impairments in other cognitive functions may also emerge, but the language dysfunction remains the most salient feature and deteriorates most rapidly throughout the illness.

Language in PPA

The language impairment in PPA varies from patient to patient. Some patients cannot find the right words to express thoughts; others cannot understand the meaning of heard or seen words; still others cannot name objects in the environment. The language impairment can be fluent (i.e., with normal articulation, flow, and number of words per utterance) or nonfluent. The single most common sign of PPA is anomia, manifested by an inability to come up with the right word during conversation and/or an inability to name objects shown by the examiner. Asking the patient to name geometric shapes, body parts, or components of common objects reveals early stages of anomia. Many patients remain in an anomic phase throughout most of the disease and experience a gradual intensification of word-finding deficits to the point of near mutism. Others, however, proceed to develop distinct forms of agrammatism and/or word comprehension deficits. The agrammatism consists of inappropriate word order and misuse of small grammatical words. One patient, for example, sent the following e-mail to her daughter: 'I will come my house in your car and drive my car into chicago . . . You will back get your car and my car park in my driveway. Love, Mom.' Comprehension deficits, if present, start with an occasional inability to understand single low-frequency words and gradually progress to encompass the comprehension of conversational speech.

The impairments of syntax, comprehension, naming, or writing in PPA are no different than those seen in aphasias of cerebrovascular causes. However, they form slightly different patterns. According to a classification proposed by Gorno-Tempini and colleagues, three variants of PPA can be recognized: an agrammatic variant characterized by poor fluency and impaired syntax, a semantic variant characterized by preserved fluency and syntax but poor single-word comprehension, and a logopenic variant characterized by preserved syntax and comprehension but frequent word-finding pauses during spontaneous speech. The agrammatic variant is also known as progressive nonfluent aphasia, and the semantic variant as semantic dementia. The three variants display overlapping distributions of neuronal loss, but the agrammatic variant is most closely associated with atrophy in the anterior parts of the language network (where Broca's area is located), the semantic variant with atrophy in the temporal components of the language network, and the logopenic variant with atrophy in the temporoparietal component of the language network.

Pathophysiology

Patients with PPA display progressive atrophy (indicative of neuronal loss), electroencephalographic slowing, decreased blood flow (measured by single photon emission computed tomography) and decreased glucose utilization (measured by positron emission tomography) that are most pronounced within the language network of the brain (Figure 1). The abnormalities may remain confined to left hemisphere perisylvian and anterior temporal cortices for many years. The clinical focality of PPA is thus matched by the anatomical selectivity of the underlying pathological process.

Neuropathology

Approximately 30% of patients have shown the microscopic pathology of Alzheimer's disease, presumably with an atypical distribution of lesions. In the majority of cases, the neuropathology falls within

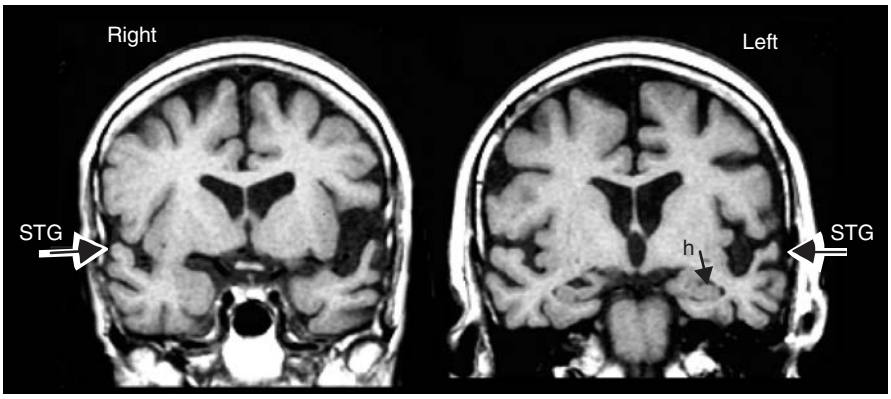

Figure 1 Coronal magnetic resonance image (MRI) of a primary progressive aphasia patient showing neuronal loss in the language network as indicated by atrophy of the left superior temporal gyrus (STG). The hippocampus (h) is intact.

the family of frontotemporal lobar degenerations and displays focal neuronal loss, gliosis, tau-positive inclusions, Pick bodies, and tau-negative ubiquitin inclusions. Apolipoprotein E and prion protein genotyping have shown significant differences between patients with typical clinical patterns of Alzheimer's disease and those with a diagnosis of PPA. The intriguing possibility has been raised that a personal or family history of dyslexia may be a risk factor for PPA, at least in some patients, suggesting that this disease may arise from a background of genetic or developmental vulnerability affecting language-related areas of the brain.

See also: Agraphia; Alexia; Alzheimer's Disease: An Overview; Brain Damage: Functional Reorganization; Brain Injury: Functional Recovery After; Brain Trauma; Cerebrovascular Disease; Dementia and Language; Language: Auditory Processes; Language: Aphasia; Lexical Impairments Following Brain Injury; Sentence Production; Sentence Comprehension; Speech Perception: Adult; Stroke; Word Production; Word Recognition; Word Learning.

Further Reading

Chawluk JB, Mesulam MM, Hurtig H, et al. (1986) Slowly progressive aphasia without generalized dementia: Studies with positron emission tomography. *Annals of Neurology* 19: 68–74.

Gorno-Tempini ML, Dronkers NF, Rankin KP, et al. (2004) Cognition and anatomy in three variants of primary progressive aphasia. *Annals of Neurology* 55: 335–346.

Kertesz A, Hudson L, Mackenzie IRA, and Munoz DG (1994) The pathology and nosology of primary progressive aphasia. *Neurology* 44: 2065–2072.

Knibb JA, Xuereb JH, Patterson K, and Hodges JR (2006) Clinical and pathological characterization of progressive aphasia. *Annals of Neurology* 59: 156–165.

Li X, Rowland LP, Mitsumoto H, et al. (2005) Prion protein codon 129 genotype is altered in primary progressive aphasia. *Annals of Neurology* 58: 858–864.

Mesulam M-M (2001) Primary progressive aphasia. *Annals of Neurology* 49: 425–432.

Mesulam M-M, Johnson N, Grujic Z, and Weintraub S (1997) Apolipoprotein E genotypes in primary progressive aphasia. *Neurology* 49: 51–55.

Mesulam M-M, Johnson N, Krefft TA, et al. (2007) Progranulin mutations in primary progressive aphasia – The PPA1 and PPA3 families. *Archives of Neurology* 64: 43–47.

Rogalski E and Mesulam M-M (2007) An update on primary progressive aphasia. *Current Neurology and Neuroscience Reports* 7: 388–392.

Sonty SP, Mesulam M-M, Thompson CK, et al. (2003) Primary progressive aphasia: PPA and the language network. *Annals of Neurology* 53: 35–49.

Weintraub S, Rubin NP, and Mesulam MM (1990) Primary progressive aphasia. Longitudinal course, neuropsychological profile, and language features. *Archives of Neurology* 47: 1329–1335.

Apoptosis in Nervous System Injury

B A Miller, J C Bresnahan, and M S Beattie, University of California at San Francisco, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

CNS Injury – Overview

The central nervous system (CNS) is generally considered to be refractory to repair and regeneration; although recently it has been realized that neurogenesis does occur in the adult CNS. Nevertheless, the loss of neurons and glial cells after CNS injury results in lifelong deficits. The most recognizable forms of CNS injury include traumatic brain injury (TBI) and spinal cord injury (SCI; Figure 1(a)), but CNS injury also results from ischemic stroke, hemorrhage, infection, and other conditions that cause neural cell death. Since there is little replacement of neurons after damage to the CNS, protection is critical, as evidenced by the evolution of the human skull and vertebral column. It is important to note that in addition to neurons, glial cells (astrocytes, oligodendrocytes, and microglia), along with vascular cells and progenitor cells, are all lost in CNS injury. The final neurological outcome of an individual with CNS injury will depend upon the summation of injury-induced degeneration and repair processes. Therefore, it is important to understand how cells die after CNS injury, which is by a combination of the two principal forms of cell death, necrosis and apoptosis.

Apoptosis – Introduction

Apoptosis, or programmed cell death, occurs naturally during development, and has recently gained attention as an important factor in central nervous system disease and injury. Apoptotic cell death is differentiated from passive, necrotic cell death by both morphological and biochemical features. Apoptotic cells shrink and undergo cellular blebbing, forming small membrane-bound debris which is phagocytosed by immune cells. In contrast, necrotic cells swell and lyse, spilling their contents into the extracellular space. A classic feature of apoptotic cell death is chromatin condensation and fragmentation, as DNA is broken up into smaller fragments by enzymes that are activated as part of the apoptotic program. This DNA fragmentation produces a characteristic pattern that can be visualized by gel electrophoresis, referred to as DNA laddering. A commonly used assay to detect apoptotic DNA fragmentation is terminal deoxynucleotidyl transferase dUTP nick-end labeling, commonly referred to as TUNEL. This assay labels 3′-OH DNA ends that are produced during apoptotic DNA cleavage. Nuclear features of apoptosis can also be seen histologically as fragmented or shrunken nuclei when cells are labeled with DNA-binding dyes such as Hoechst.

Apoptosis occurs normally during embryogenesis. It is interesting to note that this programmed cell death occurs coincidently with cell proliferation and maturation during development, and apoptosis and regeneration also occur simultaneously after CNS injury. Many essential developmental events outside the nervous system take place via apoptosis, such as the elimination of self-recognizing B and T cells and the loss of tissue to create independent digits. Apoptosis also takes place throughout the developing CNS where many different cell types undergo apoptosis, such as Purkinje cells of the cerebellum, retinal ganglion cells, and oligodendrocytes in the optic nerve, cortex, and spinal cord. Apoptosis during CNS development may take place in order to control cell number and regulate the size of CNS tissues.

The biochemical definition of apoptosis is based on the induction of specific intracellular signaling cascades that culminate in enzymatic activation leading to cell death. The enzymes most widely linked to apoptotic cell death are caspase enzymes. Caspase enzymes are cysteine proteases that are present in cells as zymogens until activated. Caspase enzymes are homologs of *ced* (*Caenorhabditis elegans* death) family genes that were originally described in the nematode *C. elegans*, in which they mediate programmed cell death. Once activated, caspases cleave both nuclear and cytoskeletal proteins. Other caspase substrates include antiapoptotic proteins and other regulatory proteins that influence cell survival. Many different events can trigger caspase activation, and different caspase enzyme subtypes can activate each other or undergo autocatalysis. Caspase activation can be detected with antibodies specific to the activated form of caspase enzymes in Western blots and isolated cell or tissue samples. Specific caspase inhibitors can also be used to block apoptotic pathways in experimental preparations. Calpains comprise another set of proteolytic enzymes that play a role in apoptosis. Calpains are calcium-dependent enzymes that can be activated in CNS disease and trauma and are also thought to contribute to apoptotic cell death.

Despite widely accepted morphological and molecular markers of apoptosis, such as DNA fragmentation and caspase activation, the distinction between apoptosis and necrosis is not always clear. It is likely that apoptosis and necrosis exist along a continuum where

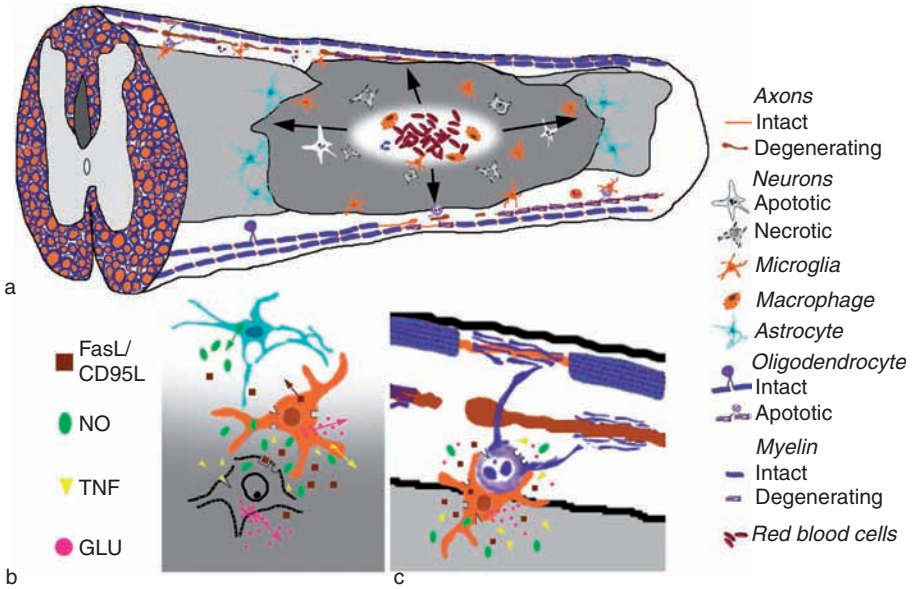

Figure 1 Apoptosis plays a role in the pathogenesis of spinal cord injury, which provides an example of the complex dynamics of central nervous system injuries. (a) Immediately after cord injury, cells are mechanically disrupted and die at the center of the injury site (white area), where petechial hemorrhages form, causing vascular elements to be released into the tissue. Over the next few hours to days, the lesion center expands (dark gray area) and additional tissue is lost via both necrosis and apoptosis. Inflammatory cells, including both microglia and macrophages, are recruited to remove cellular debris. Subsequently, the lesion cavity becomes walled off by reactive astrocytes, which contribute to the chronic scar that forms after spinal cord injury. Distant from the lesion, microglial activation is observed in the axonal tracts undergoing Wallerian degeneration. Both oligodendrocytes and microglia are observed to undergo apoptosis in these regions, as well as in the thin rim of spared white matter around the lesion center. The loss of oligodendrocytes in these areas possibly causes demyelination of surviving axons and contributes to functional loss after injury. (b, c) Examples of secreted molecules – FasL/CD95L, nitric oxide (NO), tumor necrosis factor (TNF), glutamate (GLU) – and their target cells that are thought to play a role in cell death after injury. Activated microglia are observed to be in close apposition to dying neurons (b) and apoptotic oligodendrocytes (c), suggesting an active role for microglia in inducing cell death.

the same initial insult could cause either apoptosis or necrosis, depending on the strength of the stimulus. It is likely that more severe injury results in necrosis, as cells are unable to execute coordinated apoptotic signaling cascades. It is also possible that necrotic cells can initiate apoptotic signaling cascades even as they die by necrosis; thus biochemical events of apoptosis could occur within a cell that ultimately dies via necrosis.

Molecular Mechanisms of Apoptosis in CNS Injury

Developmental apoptosis is likely a result of competition between individual cells for prosurvival signals that are in limited supply, whereas apoptosis after CNS injury is thought to be predominantly due to proapoptotic signals released after injury. Since apoptosis occurs over time, and is meditated by specific signaling cascades, apoptotic cell death could be a viable target for therapies to reduce pathological neuronal and glial loss in CNS injury and disease. Several common mediators of cell death are present in both CNS injury and disease. Among the best studied of these are excitotoxicity, inflammation, and free radicals. Excitotoxicity classically refers to the overactivation of excitatory ionotropic glutamate receptors on neurons, leading to excess intracellular calcium accumulation culminating in cell death. Though oligodendrocytes do not conduct action potentials, oligodendrocytes do express glutamate receptors and are also vulnerable to excitotoxicity. Both neurons and oligodendrocytes have been shown to undergo apoptotic cell death as a result of excitotoxicity, but excitotoxic cell death can be necrotic as well. A key event in excitotoxic cell death, either apoptotic or necrotic, is an increase in intracellular Ca^{2+}. Ca^{2+} that enters cells through ionotropic glutamate receptors can activate nitric oxide synthase enzymes that lead to intracellular free radical production. Additionally, excess intracellular Ca^{2+} can accumulate in mitochondria and cause increased production of the free radical superoxide, which can then rapidly combine with nitric oxide to form the highly reactive free radical peroxynitrite. Intracellular free radicals have been shown to activate apoptotic signaling cascades and can also directly

damage cellular contents. While intracellular free radical formation has been linked to apoptosis, free radical-mediated cell death can be necrotic as well. The net concentration or specific species of intracellular free radicals produced may determine whether a cell dies via apoptosis or necrosis.

One way intracellular free radicals can induce apoptosis is through their effects on mitochondria. Free radicals have been linked to the release of cytochrome *c* from the mitochondrial intermembrane space, which initiates what is referred to as the intrinsic pathway of apoptosis. Cytochrome *c* released into the cytosol can combine with apoptosis-inducing factor-1, which, in the presence of adenosine triphosphate or deoxyadenosine triphosphate, can lead to the activation of caspase-9. Caspase-9 can subsequently activate caspase-3, which is considered to be one of the final effectors of apoptosis.

Inflammation has also been linked to apoptosis in neurological diseases and injury. CNS inflammation can be induced by microglia within the CNS or by cells from the peripheral immune system. Blood–brain barrier (BBB) breakdown occurs after stroke and CNS trauma and could provide a route for peripheral immune cells to access the CNS. Both microglia and peripheral immune cells can produce tumor necrosis factor-α (TNF-α) and other proinflammatory cytokines which can induce apoptosis in both neurons and oligodendrocytes. The pathway by which TNF-α induces apoptosis is known as the extrinsic apoptotic pathway. This pathway is initiated by TNF-α binding to receptors on the cell surface, which causes the receptors to trimerize. Once TNF-α receptors have trimerized they activate caspase-8 via adapter molecules that bind to the intracellular domain of the TNF-α receptors. Caspase-8 can subsequently activate caspase-3, leading to a point of convergence for the extrinsic and intrinsic apoptotic pathways. TNF-α can also induce free radical production, which can lead to apoptotic cell death. Since excitotoxicity and inflammation are present simultaneously in CNS disease and trauma, free radical injury could become a central mediator of cell death in numerous disease states. Although TNF-α has been implicated in apoptosis of both neurons and oligodendrocytes, some studies have shown TNF-α to have a beneficial role in CNS injury and repair. This is just one example of how factors that may induce apoptosis after CNS injury may also have beneficial effects. Further research into the complicated signaling cascades that are activated by TNF-α and other factors present after CNS injury will help to better separate the beneficial and detrimental roles of these molecules.

Another pathway that may induce apoptosis in multiple diseases and trauma is the p75 receptor signaling pathway. p75 is a low-affinity neurotrophin receptor that is part of the TNF receptor superfamily. p75 can bind to several different neurotrophins and increases the response of the neuronal TrkA receptor to nerve growth factor (NGF), which promotes neuronal growth. However, in the absence of the TrkA receptor, p75 activation via NGF or pro-NGF leads to apoptosis. Injuries to the CNS increase p75 levels, and cells undergoing apoptosis after facial nerve axotomy and SCI upregulate expression of p75. Activation of the p75 receptor by neurotrophins that are produced after injury may be another mechanism which induces apoptosis after CNS trauma.

When considering the multitude of factors that may induce apoptosis after CNS injury, it is important to realize that these factors may be activated concurrently or sequentially following one particular injury. There are many points at which these signaling pathways may converge, such as one signaling pathway amplifying the expression of receptors for another, or two or more separate pathways inducing activation of the same caspase pathways. This mechanistic overlap makes it likely that numerous antiapoptotic treatments may be needed to reduce cell death after CNS injury. Furthermore, the mechanisms of apoptosis described here are by no means the only inducers of apoptosis after CNS injury, as molecules such as interferon-γ and Fas/CD95 and its ligand have also been linked to apoptosis after CNS injury. In the future, a more thorough understanding of the extracellular and intracellular signaling events that lead to apoptosis may lead to better treatments for CNS injury.

Neuronal Apoptosis in CNS Injury

Neuronal apoptosis has been described in several CNS disorders and trauma. Apoptotic neuronal death is seen in both human amyotrophic lateral sclerosis (ALS) and animal models of ALS. Apoptotic neurons have also been detected in regions of cell death in Huntington's disease, Alzheimer's disease, Parkinson's disease, and human immunodeficiency virus (HIV)-induced encephalitis. Since these diseases are chronic and may be due to genetic mutations, it would be expected that programmed apoptotic cell death plays a role in neuronal loss in these conditions. However, apoptosis has also been shown to play a role in acute neurological injuries such as stroke, TBI, and SCI. After ischemic stroke, cells are quickly lost in the core of the lesion, due to energy deprivation. It is thought that this cell loss is primarily necrotic. Surrounding the necrotic core of a stroke lesion, there is an area referred to as the penumbra, in which cells are lost over time. Neuronal death in the penumbra is

thought to be apoptotic, and may be more amenable to therapeutic intervention than is acute neuronal loss in the ischemic core.

One of the factors that may lead to neuronal apoptosis in stroke is reperfusion injury. The return of blood flow to an area affected by stroke can paradoxically lead to increased damage. One mechanism behind reperfusion injury is thought to be intracellular free radical generation due to the sudden availability of oxygen to tissue previously deprived of oxygen and glucose. Neurons in a previously hypoxic area may not be able to deal with a sudden increase in oxygen availability and therefore be paradoxically injured when blood flow returns to an ischemic area. The return of blood flow to a previously ischemic area allows for the execution of energy-dependent signaling cascades that mediate apoptosis. Additionally, reperfusion into a damaged area can bring peripheral immune cells into the area, particularly when the BBB has been damaged. This combination of increased free radical production, initiation of apoptotic signaling pathways, and immune infiltration can lead to apoptosis.

Apoptosis is present in animal models of TBI and SCI and is thought to contribute to neuronal loss in human traumatic injury. Apoptotic neurons have been detected in the cortex and hippocampus following experimental TBI. The loss of even a small percentage of neurons in the hippocampus could have a significant effect on quality of life after TBI, when memory impairment is often a result of injury. After SCI, even a small amount of motor neuron loss in critical areas of the cord can lead to functional deficits such as breathing difficulties or loss of hand function. In experimental models of CNS injury, some treatments that reduce apoptosis have been shown to result in better outcomes.

Oligodendrocyte Apoptosis in CNS Injury

In addition to neurons, oligodendrocyte progenitor cells (OPCs) and mature oligodendrocytes undergo apoptosis in CNS trauma and disease. In periventricular leukomalacia (PVL), OPCs are lost as a consequence of hypoxia, ischemia, or intrauterine infection. This loss of OPCs, with the resulting deficiency of mature oligodendrocytes, is thought to be one of the causes of spastic cerebral palsy. Examination of brains from preterm infants has shown apoptosis associated with white-matter lesions and OPC loss. Experimental models of PVL also show OPCs undergoing apoptosis. Mature oligodendrocytes die in multiple sclerosis (MS), and examination of brains of MS patients has revealed apoptotic oligodendrocytes. Experimental models of MS also support oligodendrocyte apoptosis as a feature of this disease. Numerous *in vitro* studies designed to model PVL and MS have shown TNF-α, free radicals, and glutamate receptor activation to induce OPC and oligodendrocyte apoptosis.

Oligodendrocytes also die after CNS trauma such as TBI and SCI. In SCI, it has been shown that oligodendrocytes undergo apoptosis for several weeks after the initial injury. This delayed death of oligodendrocytes is thought to reduce the effectiveness of neural conduction in the spared rim of white matter that often exists in injured spinal cords, and to impair optimal function of neural pathways that may have survived the initial injury. Though many studies have demonstrated delayed oligodendrocyte apoptosis after SCI, the events that cause this remain unknown. It has been shown that axonal transection can lead to oligodendrocyte cell death in the optic nerve, and oligodendrocyte survival *in vitro* can be enhanced by the presence of specific growth factors and hormones. Therefore, it is possible that oligodendrocyte apoptosis in SCI is due to a loss of trophic support from axons as axons degenerate after injury, similar to developmental apoptosis that occurs due to limited availability of growth factors.

An alternate hypothesis is that oligodendrocyte apoptosis after SCI is a result of the presence of substances that are released after injury, rather than a decrease in the amount of survival factors present. After SCI, elevated levels of TNF-α and glutamate are present in the spinal cord. Both of these molecules have been shown to induce oligodendrocyte cell death *in vitro* (Figure 2). *In vivo,* glutamate antagonists have been shown to protect oligodendrocytes after SCI. Though TNF-α is likely to contribute to oligodendrocyte apoptosis after SCI, some studies have shown that knocking out TNF-α or TNF-α receptors results in worse functional outcome and less regeneration after SCI. These findings emphasize that signals that promote apoptosis in some settings may have beneficial roles in other settings or at other times postinjury. One possible source for both glutamate and TNF-α after SCI could be microglia, the resident immune cells of the CNS.

Role of Microglia in Neuronal and Oligodendrocyte Apoptosis

Microglia are known to react quickly in response to CNS injury or disease, proliferating and migrating into an injury site. Microglia secrete a wide array of molecules that have been shown to be toxic to neurons and oligodendrocytes, including glutamate, free radicals, and TNF-α. Microglial activation has been

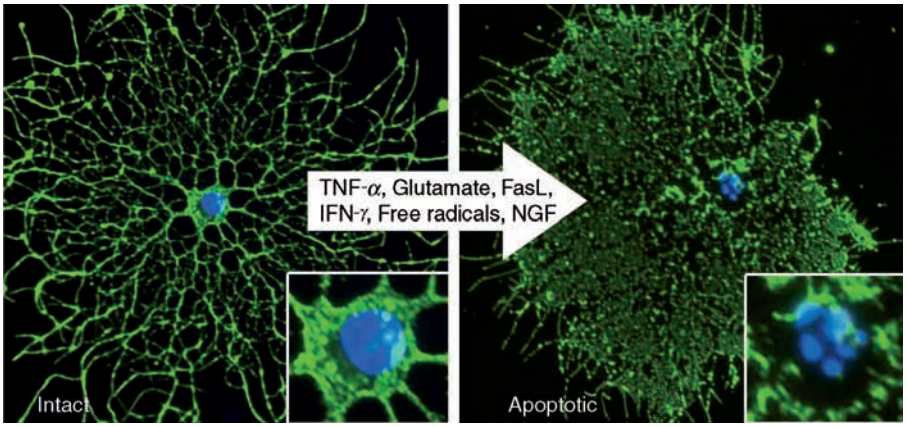

Figure 2 These oligodendrocytes, grown *in vitro*, have been labeled with antibodies to the oligodendrocyte marker galactocerebroside (green) and with Hoechst, which labels DNA and cell nuclei (blue). The nuclei of these cells show the difference between a cell with an intact nucleus (left panel, inset) and a cell with a fragmented apoptotic nucleus (right panel, inset). Some of the factors that have been shown to induce oligodendrocyte apoptosis are identified within the arrow. TNF-α, tumor necrosis factor-α; IFN-γ, interferon-γ; NGF, nerve growth factor.

linked to neuronal apoptosis in several diseases and TBI. Minocycline, a drug that decreases microglial activation, has been shown to reduce neuronal apoptosis in animal models of brain injury. Additionally, treatments antagonizing TNF-α and free radicals, both of which can be produced by microglia, have been shown to improve outcomes in experimental TBI.

It has been shown that activated microglia contribute to oligodendrocyte death in models of MS and PVL. Microglia-induced oligodendrocyte damage may be further exacerbated by the fact that many factors that lead to oligodendrocyte death also lead to microglial activation, such as glutamate and proinflammatory cytokines. Microglia may exert their effects on oligodendrocytes and OPCs both at a distance and via direct cell–cell interaction. *In vitro*, microglia have been shown to be capable of inducing OPC death without being in direct contact with OPCs. However, *in vivo*, microglia have been observed in close proximity to dying oligodendrocytes after SCI (Figure 1(c)). This proximity of microglia to oligodendrocytes after CNS injury may increase the influence of microglia on oligodendrocyte and OPC survival, as it has been shown *in vitro* that physical contact between microglia and oligodendrocytes is a key factor in mediating oligodendrocyte death via microglial TNF-α. While it is possible that microglia induce oligodendrocyte apoptosis by TNF-α, glutamate, or free radical generation, it is also possible that activated microglia near apoptotic oligodendrocytes are only removing the debris left from oligodendrocyte death, similar to apoptosis during development. The relationship between apoptotic oligodendrocytes and activated microglia remains under investigation.

Microglia are also observed to undergo apoptosis in neurological injury. Microglial apoptosis occurs after SCI, peripheral nerve injury, and *in vitro* as a result of activation. This may be caused by excessive free radical production by activated microglia, which could lead to self-induced apoptosis. Microglial apoptosis after CNS injury may be a beneficial self-regulatory process that reduces the number of microglia and the amount of damage microglia do to other CNS cells. On the other hand, microglia have also been suggested to play a beneficial role after CNS injury, and microglial apoptosis may have a detrimental effect on recovery.

Role of Astrocytes in Apoptosis after CNS Injury

Unlike neurons, oligodendrocytes, and microglia, astrocytes are rarely observed to undergo apoptosis after CNS injury. However, astrocytes are thought to play an important role in determining if other cells undergo apoptosis after injury. Astrocytes are connected by gap junctions to form a syncitium that can allow large numbers of astrocytes to react quickly and in unison to signals within the CNS. Under normal conditions, astrocytes protect other cells in the CNS by scavenging excess glutamate, free radicals, and potassium from the extracellular environment. Though not classically considered immune cells, astrocytes can exhibit many of the functions of immune cells after CNS injury and may contribute to the apoptosis of other cell populations. Astrocytes react to CNS injury by becoming activated, much like microglia – altering their cytoskeletal structure and secreting cytokines and

free radicals (Figure 1(b)). Astroglial cytokines and free radicals can activate apoptotic pathways in neurons and oligodendrocytes and also induce microglial activation. Astrocytes also contribute to the formation of the glial scar, which persists chronically after CNS injury and is thought to impair regeneration.

Astrocytes are found in close association with blood vessels in the CNS, and along with endothelial cells constitute the BBB that restricts access of peripheral blood cells and cytokines to the CNS. BBB breakdown occurs in multiple sclerosis, stroke, TBI, and SCI. A compromised BBB allows peripheral immune cells into the injured CNS, where inflammation is thought to induce apoptosis of neuronal and glial cells but may also contribute to repair after injury. Astrocytes are considered more resilient than are neurons and oligodendrocytes in the setting of CNS trauma, and astrocyte apoptosis has not been studied to the same extent as has neuronal or oligodendrocyte apoptosis, although astrocytes do die after CNS injury.

Summary

Apoptosis, or programmed cell death, is morphologically and biochemically distinct from necrosis and is carried out by several intracellular signaling cascades. Neuronal and glial apoptosis processes are thought to play a role in many different neurological diseases and injury. Several molecules may mediate apoptosis in the CNS, including cytokines, neurotrophins, glutamate, and free radicals. The inflammatory response in CNS injury and disease may contribute to apoptosis but may also be important for tissue repair. Since apoptosis occurs over time, and is mediated by distinct intracellular signaling events, apoptotic cell death may be a viable target for therapies designed to improve outcomes after CNS injury. Antiapoptotic therapies have shown promise for the treatment for CNS injury. Treatments such as minocycline, mitochondrial permeability inhibitors, and caspase inhibitors have shown beneficial effects on cell survival and functional outcome in models of stroke, SCI, and excitotoxicity. The only currently approved clinical treatment for spinal cord injury in the United States is methylprednisolone, which has shown to have antiapoptotic effects, although there is controversy surrounding its use. There are currently no clinically approved antiapoptotic treatments for TBI or stroke, and much research continues to focus on developing clinically efficacious treatments for these CNS injuries.

Although much effort has been directed at reducing apoptosis after CNS injury, apoptosis after neurotrauma may have a beneficial role as well, by removing dysfunctional cells and making way for the repair and regeneration that can occur after injury. Apoptosis may also play a role in regulating the immune response to injury and preventing excess inflammation. Although apoptosis has only recently been identified as an important event after CNS injury, our understanding of the mechanisms and mediators of apoptosis is expanding rapidly. The development of therapies that reduce neuronal and glial apoptosis may lead to more effective treatment and recovery after CNS injury.

See also: AMPA Receptors: Disease; Apoptosis in Neurodegenerative Disease; Astrocyte: Response to Injury; Axonal Injury in Demyelinating Disease and CNS Injury; Blood-Brain Barrier and Neurovascular Mechanisms of Neurodegeneration and Injury; Brain Injury: Functional Recovery After; Brain Trauma; Excitotoxicity in Neurodegenerative Disease; Glial Responses to Injury; Inflammation in Neurodegenerative Disease and Injury; Microglial Response to Injury; Mitochondrial Dysfunction in Nervous System Injury; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Sensorimotor Plasticity and Control of Movement Following Spinal Cord Injury +D18; Spinal Cord Regeneration and Functional Recovery: Strategies; Spinal Cord Injuries; Stroke: Injury Mechanisms.

Further Reading

Baptiste DC and Fehlings MG (2006) Pharmacological approaches to repair the injured spinal cord. *Journal of Neurotrauma* 23: 318–334.

Barres BA, Jacobson MD, Schmid R, et al. (1993) Does oligodendrocyte survival depend on axons? *Current Biology* 3: 489–497.

Beattie MS, Hermann GE, Rogers RC, et al. (2002) Cell death in models of spinal cord injury. *Progress in Brain Research* 137: 37–47.

Beattie MS (2004) Inflammation and apoptosis: Linked therapeutic targets in spinal cord injury. *Trends in Molecular Medicine* 10: 580–583.

Casaccia-Bonnefil P (2000) Cell death in the oligodendrocyte lineage: A molecular perspective of life/death decisions in development and disease. *Glia* 29: 124–135.

Casha S, Yu WR, and Fehlings MG (2001) Oligodendroglial apoptosis occurs along degenerating axons and is associated with FAS and p75 expression following spinal cord injury in the rat. *Neuroscience* 103: 203–218.

Crowe MJ, Bresnahan JC, Shuman SL, et al. (1997) Apoptosis and delayed degeneration after spinal cord injury in rats and monkeys. *Nature Medicine* 3: 73–76.

Friedlander RM (2003) Apoptosis and caspases in neurodegenerative diseases. *New England Journal of Medicine* 348: 1365–1375.

Park E, Velumian AA, and Fehlings MG (2004) The role of excitotoxicity in secondary mechanisms of spinal cord injury: A review with an emphasis on the implications for white matter degeneration. *Journal of Neurotrauma* 21: 754–774.

Selmaj KW and Raine CS (1988) Tumor necrosis factor mediates myelin and oligodendrocyte damage *in vitro*. *Annals of Neurology* 23: 339–346.

Shuman SL, Bresnahan JC, and Beattie MS (1997) Apoptosis of microglia and oligodendrocytes after spinal cord contusion in rats. *Journal of Neuroscience Research* 50: 798–808.

Snider BJ, Gottron FJ, and Choi DW (1999) Apoptosis and necrosis in cerebrovascular disease. *Annals of the New York Academy of Sciences* 893: 243–253.

Sugawara T, Fujimura M, Noshita N, et al. (2004) Neuronal death/survival signaling pathways in cerebral ischemia. *NeuroRx* 1: 17–25.

Zajicek JP, Wing M, Scolding NJ, et al. (1992) Interactions between oligodendrocytes and microglia. A major role for complement and tumour necrosis factor in oligodendrocyte adherence and killing. *Brain* 115: 1611–1631.

Relevant Websites

http://www.nih.gov – National Institutes of Health Special Interest Groups (Apoptosis Interest Group).

http://www.edc.pitt.edu/neurotrauma – National Neurotrauma Society (Gainesville, FL, USA).

http://www.sfn.org – Society for Neuroscience, briefing on cell suicide.

Apoptosis in Neurodegenerative Disease

K A Roth and J J Shacka, University of Alabama at Birmingham, Birmingham, AL, USA

© 2009 Elsevier Ltd. All rights reserved.

Apoptosis

In 1972, Kerr and colleagues coined the term apoptosis, which is Greek for 'falling off,' as in petals from a flower or leaves from a tree. This term was used to define a type of cell death characterized morphologically by the discrete budding of self-contained fragments from the cell membrane following a death stimulus and was distinct from necrotic (passive or accidental) cell death. Apoptosis, known also as type I cell death, is now defined by distinct morphological criteria, including pyknosis (condensation or reduction in size) of the cell and/or cell nucleus along with nuclear fragmentation and hyperchromatosis. The DNA of apoptotic nuclei is typically digested into 200–300 bp fragments which can be observed by enzymatic labeling of nicked DNA or by electrophoretic detection of DNA laddering. The plasma membrane of apoptotic cells remains relatively intact, with the dying cell budding off discrete packages of itself into 'apoptotic bodies' that are ultimately phagocytosed. This efficient and discrete phagocytosis in the absence of a ruptured plasma membrane generally limits the inflammatory response around an apoptotic cell and is unlike necrosis, which is characterized by a rupture of the plasma membrane and subsequent leakage of cell contents into the surrounding extracellular space with a resultant inflammatory response. In addition, unlike necrosis, apoptosis is an energy-dependent process and often requires new gene transcription and protein translation for cellular execution.

Molecular Regulation of Apoptosis

Many seminal discoveries have been made in the past quarter century that have enhanced our understanding of cell death. Using the nematode *Caenorhabditis elegans* as a model, Ellis and Horvitz showed for the first time in the 1980s that apoptotic cell death was regulated by distinct genes. Subsequent cloning studies demonstrated that mammals also express homologs of these genes with similar function, although it has been found that the regulation of mammalian cell death is more complex than in the nematode. In 1984, a novel oncogene, B-cell lymphoma-2 (Bcl-2), was isolated and was found to have prosurvival or anti-apoptotic functions. Mammalian cell apoptosis can be classified into distinct pathways, extrinsic (death-receptor mediated) or intrinsic (mitochondrial), and involves the participation of an extensive family of Bcl-2 family proteins (described below) and the activation of caspases (see Figure 1).

Caspases

Caspases are cysteine-containing proteases that cleave substrates after aspartic acid residues and critically regulate both the transduction of apoptotic stimuli and the final execution of apoptotic death. There are 14 mammalian and 12 human caspases that can be divided into two basic groups, the interleukin-1β converting enzyme (ICE)-like and the CED-3-like (named after the *C. elegans* homolog) families. The ICE-like caspases, which include caspase-1, caspase-4, caspase-5, and caspase-12, are involved in the proteolytic processing of cytokines and as such are related indirectly to neuron cell death as a function of modulating inflammatory responses. The CED-3-like caspases are directly involved in the apoptotic signaling cascade and can be further divided into two subgroups, the initiator caspases (caspase-2, caspase-8, caspase-9, and caspase-10) and the effector caspases (caspase-3, caspase-6, and caspase-7). Both initiator and effector caspases exist as inactive zymogens within the cell prior to receipt of a death stimulus. Initiator caspases are activated early in the apoptosis cascade by conformational change resulting from interaction with other apoptosis-associated proteins. Effector caspases are activated by the catalyzed cleavage from 'activated' initiator caspases. Both initiator and effector caspases have been shown to be proteolytically cleaved, but initiator caspases do not typically require cleavage for their activation. It is the selective cleavage of substrates by caspases which induces in large part the appearance of the morphological features of apoptosis.

Bcl-2 Family

There are three major subgroups of proteins in the Bcl-2 family that are classified by composition of their Bcl-2 homology (BH) domains and by function (anti- vs. proapoptotic). It is widely believed that the relative 'ratio' of antiapoptotic versus proapoptotic Bcl-2 family members influences the response of neurons to apoptotic stress. Members of the antiapoptotic subgroup include Bcl-2, Bcl-X_L, Bcl-w, and Mcl-1 and possess four BH domains, except for Mcl-1, which lacks the BH4 domain. The importance of the BH1 and BH4 domains in the maintenance of anti-apoptotic activity of Bcl-2 and Bcl-X_L has been emphasized by molecular studies, where the targeted

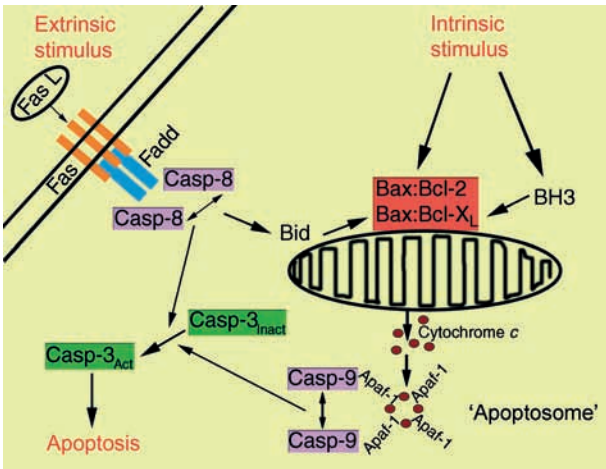

Figure 1 Molecular regulation of cell death in mammals by the extrinsic (death receptor-mediated) versus intrinsic (mitochondrial) apoptotic pathways. Extracellular death receptor ligands (e.g., Fas ligand) bind to and activate their receptors (e.g., Fas), which in turn induces recruitment of adaptor proteins (Fadd) to the complex with subsequent recruitment, dimerization, and activation of caspase-8, an initiator caspase. Activation of caspase-8 in turn can directly cleave and activate caspase-3 resulting in apoptotic death. In addition, caspase-8 can cleave Bid, a proapoptotic, BH3-domain-only member of the Bcl-2 family, to truncated Bid which interacts with Bax at mitochondria and causes the subsequent activation of the intrinsic apoptotic pathway. Intrinsic death stimuli (e.g., DNA damage) can cause changes in expression and/or function of anti- and/or proapoptotic members of the Bcl-2 family. The ratio of anti- to proapoptotic Bcl-2 family members is thought to regulate the release of cytochrome *c* from mitochondria, which in turn binds with Apaf-1 and caspase-9 in a protein complex termed the 'apoptosome.' The binding of caspase-9 in the apoptosome induces its forced dimerization and subsequent activation, which in turn induces activation of caspase-3. The activation of caspase-3 and other effector caspases is responsible for the cleavage of downstream substrates that promotes in part the appearance of apoptotic morphology and ultimate cell death.

deletion of these regions has been shown to abrogate their antiapoptotic activity. Bcl-2 is expressed at high levels during nervous system development and remains high in the adult spinal cord and in sympathetic neurons, but decreases in relative expression in the adult brain. Upregulation of Bcl-2 has been documented in many neurodegenerative diseases, occurring possibly as an antiapoptotic neuron survival response or in reactive glial cell elements. Bcl-X_L and Bcl-w are expressed at high levels in adult central nervous system (CNS) and have also been shown to be altered in neurodegenerative disease; in contrast, Mcl-1 is thought to play a more focused role in CNS development.

The proapoptotic, multidomain subgroup consists of Bax, Bcl-2-homologous antagonist/killer (Bak), Bok, and Bcl-X_S. These proteins share three homologous BH domains (BH1–3) except for Bcl-X_S, which lacks BH1 and BH2 domains but rather possesses a BH4 domain. Both Bax and Bak regulate the release of cytochrome *c* from mitochondria, a critical event in activation of the intrinsic apoptotic pathway. Upregulation of Bax has been documented in many neurodegenerative diseases. In contrast, while present in the CNS and similar in function to Bax, studies have indicated a limited role for Bak in promoting apoptotic neuron death. Bcl-X_S is expressed at low levels in the mammalian nervous system and a regulatory role for Bok in CNS apoptotic death is yet to be described. The proapoptotic BH3 domain-only subgroup includes Bcl-2-associated death protein (Bad), Bid, Bcl-2 interacting mediator of cell death (Bim), Bcl-2 19 kDa interacting protein (BNip), death protein 5/harakiri (DP5/Hrk), Noxa, and p53-upregulated modulator of apoptosis (Puma). Although members of this subgroup are comparably the most diverse, their regulation of cell death depends critically upon the BH3 domain and the presence of Bax and/or Bak. The response of BH3 domain-only proteins to an apoptotic stress is regulated in part by distinct protein modifications, such as phosphorylation (Bim), dephosphorylation (Bad), induction of expression (Bim, DP5, Noxa, Puma), or cleavage (Bid). Altered regulation of BH3 domain-only proteins has been implicated in many neurodegenerative diseases, which are discussed below.

Extrinsic versus Intrinsic Death Pathway

There are two distinct pathways of caspase-dependent cell death, the extrinsic or death receptor-mediated pathway and the intrinsic or mitochondrial-mediated pathway. The extrinsic pathway involves binding of a death receptor ligand (e.g., Fas ligand or TNF-α) to a cell surface death receptor, which in turn stimulates receptor dimerization and the association of adaptor proteins to a newly formed complex that facilitates the association-induced activation of caspase-8. In the intrinsic pathway, a death stimulus regulates in part the levels and/or function of pro- versus antiapoptotic Bcl-2 family members, which induces the mitochondrial release of cytochrome *c* into the cytosol and in turn facilitates the energy-dependent activation of caspase-9 through formation of the 'apoptosome' complex with the protein Apaf-1. Both the intrinsic and extrinsic pathways converge at the level of effector caspase activation. Interestingly, cross-talk between these two pathways occurs via the cleavage of Bid by activated caspase-8 and the translocation of truncated Bid to the mitochondria, where it associates with the proapoptotic protein Bax to induce the release of cytochrome *c* and resultant activation of the intrinsic pathway.

Apoptosis and Neurodegeneration

A variety of evidence for the occurrence of apoptotic death in human neurodegenerative disease has been cited to support the hypothesis that apoptosis-associated molecules such as Bcl-2 and caspase family members play a role in disease pathogenesis. Although the extent and quality of data vary for specific neurodegenerative diseases (see below), much of the data is correlational and fraught with technical and methodological obstacles inherent in the study of human postmortem brain tissue. Poorly validated antibodies to Bcl-2 and caspase family members and nonspecific 'markers' of apoptosis such as terminal deoxytransferase-mediated deoxyuridine triphosphate nick end labeling (TUNEL) have further complicated the analysis of the biological significance of apoptosis in neurodegenerative disease onset and progression. To date, specific mutations in apoptosis-associated genes have not been linked to any human neurodegenerative disease, raising concern about the primacy of altered apoptosis regulation in disease pathogenesis. Despite these caveats, it is clear that alterations in apoptosis-associated molecules occur in a variety of neurodegenerative diseases, that apoptotic death is one of several morphological forms of neuron death detected in human neurodegenerative disease, and that pharmacological inhibition of apoptotic death offers some hope for at least temporary preservation of neuron survival.

Alzheimer's Disease

Alzheimer's disease (AD) is the most common human neurodegenerative disease and is the most prevalent form of dementia among elderly patients. Neuropathological hallmarks of AD include mature amyloid plaques and neurofibrillary tangles, which are considered the classical features of AD neuropathology and were used to identify the first case of AD more than 100 years ago. In addition to plaques and tangles, AD brains show decreased synapse density and profound neuron loss. To date, the majority of research efforts in AD has focused on mechanisms of plaque and tangle formation and their contribution(s) to neuron dysfunction and ultimate neuron death. Amyloid plaques are caused by the aberrant cleavage of the plasma membrane-bound amyloid precursor protein (APP), a soluble form of amyloid beta (Aβ), which ultimately deposits in the extracellular space in its fibrillar form. However, the formation of amyloid plaques correlates weakly at best with the symptoms and progression of AD. Neurofibrillary tangles are caused in part by the hyperphosphorylation of the cytoskeletal protein tau and more closely follow AD progression. Many studies have indicated that a loss in synapse density and neuron number correlate highly with dementia in AD, although the definitive causes of synapse and neuron loss in AD remain speculative. Whether neuron loss precipitates a decrease in synaptic integrity or if a decrease in neurotrophic support at the level of the synapse precipitates neuron death is not entirely clear in AD. Significant decreases in immunoreactivity for the synaptic protein synaptophysin have been detected in early AD prior to pronounced neuron loss, suggesting that neuronal dysfunction precedes neuron death. However, some brain regions, including the nucleus basalis of Meynert, which contains cell bodies of neurons which project to the cortex, show little neuron loss yet still exhibit AD neuropathology, suggesting that synapse function and cell death may not be directly related. Nevertheless, a better understanding of cell death pathways in AD combined with determining the events that trigger cell death in AD will hopefully generate more effective treatment strategies that preserve both neuron numbers and function.

A role for apoptosis in AD brain has been suggested previously by numerous studies. Techniques such as TUNEL staining have been developed that detect *in situ* the fragmentation of DNA, and studies of AD brain show a marked increase in numbers of such 'apoptotic' nuclei relative to control brains. However, subsequent studies have indicated that these detection methods are not specific for apoptotic nuclei, that is, necrotic nuclei have also been shown to exhibit fragmented DNA and are highly sensitive to fixation artifact. In addition, the large number of apoptotic neurons determined by these methods does not correlate well with the chronic neurodegenerative process of AD, which would predict low numbers of apoptotic neurons at any one time that are efficiently phagocytosed and degraded. Thus, despite findings of numerous studies, the morphological assessment of apoptotic neurons in AD remains inconclusive.

Many studies have indicated an altered regulation of Bcl-2 family members in AD brain. Both proapoptotic (Bak, Bad) and antiapoptotic (Bcl-2 and Bcl-X_L) Bcl-2 family members have been shown to be upregulated in AD brain. The increased expression of Bcl-2 and Bcl-X_L may represent a chronic, antiapoptotic stress response of compromised neurons to compensate for increases in Bax. Based on these findings, a proposed therapeutic strategy in AD is the pharmacological delivery of agents that effectively increase the ratio of anti- to proapoptotic Bcl-2 family members in neurons. However, Bcl-2 family members are also expressed in glial cells, and increases in Bcl-2 have been shown to occur in reactive astrocytes in AD brain. Thus, expression data alone may provide an

incomplete picture of the role of Bcl-2 family members and apoptosis in AD pathogenesis. Although they are clearly altered in AD brain, careful studies of Bcl-2 family members in the early stages of AD are required to differentiate possible etiologic from compensatory alterations in Bcl-2 family members specific for neurons.

In vitro models of AD have also implicated altered regulation of Bcl-2 family members. Treatment with Aβ, which causes the death of cultured neurons, induces the expression of proapoptotic Bcl-2 family members including Bax and DP5 while decreasing the expression of Bcl-2 and Bcl-w. In addition, the overexpression of Bcl-2, Bcl-X_L, or Bcl-w and the targeted deletion of Bax prevent Aβ-induced neuron death *in vitro*, which further implicates potential neuroprotective strategies *in vivo*. Future studies are necessary to validate the relative contribution of Bcl-2 family members for Aβ-induced neurotoxicity *in vivo*.

Caspase activation is a critical component of the apoptotic cascade and is largely responsible for the cytological changes that define apoptotic morphology. As such, many *in vivo* studies of human AD brain and *in vitro* models of AD have focused on levels and/or activation of caspases. An increased expression of caspase-1 has been reported in AD brain extracts, which may indicate activation of proinflammatory cytokines such as interleukin-1β that is regulated by active caspase-1 and may lead to the indirect death of neurons in AD. Caspase-12 and caspase-4 are endoplasmic reticulum (ER)-associated proteins that mediate in part ER stress-induced apoptosis. While human isoforms of caspase-4 exist and have been shown along with caspase 12 to induce Aβ-induced neuron apoptosis, caspase-12 was shown recently not to be associated with ER stress-induced apoptosis in humans. In addition, caspase-4 has been shown to cleave APP and may thus be involved in the disease-associated formation of Aβ. Although little, if any, evidence currently exists for its activation in AD, future studies of apoptosis in AD brain warrant further investigation of caspase-4.

A clear role for caspase-2, caspase-8, or caspase-9 in AD has not yet emerged. Of the effector caspases, caspase-7 exists in very low levels and thus is unlikely to play a role in AD pathogenesis. Both caspase-6 and caspase-3 are abundant in human brain and several studies have suggested that caspase-6 may be involved in regulating both APP processing and apoptosis of human neurons. The majority of evidence for caspase activation in apoptosis has focused on caspase-3. Many studies have reported an increased expression and activation of caspase-3 in AD brain, but many conflicting reports have also reported little difference in levels or activity of caspase-3 in AD versus control brains. Caspase-3 has been shown to be activated in cultured neurons treated with Aβ; however, neurons deficient in caspase-3, while lacking apoptotic morphology, are not protected from the death-promoting effects of Aβ, which suggests that caspase-independent death pathways may also exist in AD (see below). While it remains entirely possible that caspase-dependent neuron death is important in AD, a comprehensive analysis of previous studies suggests a degree of caution in defining its relative contribution to AD pathogenesis. A potential disease-promoting role for nonlethal caspase-3 activation in AD neurons has also been proposed. Although cleaved 'activated' caspase-3 is typically associated with apoptotic cell death, cleaved caspase-3 immunoreactivity is consistently observed in AD neurons undergoing granulovacuolar degeneration, a nonapoptotic neurodegenerative process (Figure 2). The significance of this observation is currently unclear since granulovacuolar degeneration occurs in only a small fraction of degenerating neurons in the AD brain. If caspases are indeed involved in both sublethal and terminal events in AD neurons, caspase inhibitors may yet prove to be an effective therapy for preserving neuron function in AD.

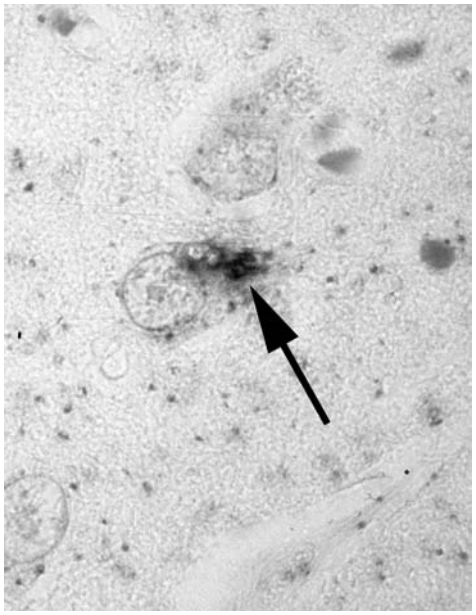

Figure 2 Immunohistochemical detection of activated caspase-3 in AD brain shows strong labeling (dark deposit, indicated by an arrow) in hippocampal pyramidal neurons bearing features of granulovacuolar degeneration. Note that despite the presence of activated caspase-3 immunoreactivity, the nucleus appears intact and the neuron lacks apoptotic histological features.

Parkinson's Disease

Parkinson's disease (PD), the second most common neurodegenerative disease after AD, is characterized

by progressive motor impairment and is defined neuropathologically by the substantial loss of dopamine-containing neurons of the substantia nigra that project to the striatum along with the development of neuronal inclusions called Lewy bodies. Lewy bodies are cytoplasmic aggregates containing many proteins including α-synuclein, ubiquitin, and neurofilaments and occur in all affected brain regions of PD. The majority of PD cases are idiopathic, but approximately 5–10% of PD patients show familial patterns of inheritance, and to date several PD-linked gene mutations have been identified, such as those that encode for proteins including α-synuclein, parkin, and DJ-1.

The contribution of apoptosis in PD has been controversial both due to a limited number of reports suggesting alterations in apoptotic markers and morphology and to reports indicating little, if any, contribution of apoptosis to the neurodegenerative process in PD brain. This may be due to the chronic nature of the disease whereby only a few select neurons undergo apoptosis at any one time. Studies have indicated an upregulation in proapoptotic Bax and also antiapoptotic Bcl-2 and Bcl-X_L in PD brain in comparison to age-matched controls. Increases in Fas ligand and activity of caspase-8 have been reported in PD brain, which may indicate the contribution of the extrinsic apoptotic death pathway. In addition, some studies have indicated an increase in the activity of caspase-3 in the PD substantia nigra along with apoptotic nuclear morphology, but it is not clear whether this is due to activation of the extrinsic and/or intrinsic apoptotic pathways or relates to disease pathogenesis.

Analysis of mutant PD genes, either in genetically modified mice or in cultured dopaminergic neurons, has determined that mutant α-synuclein or the targeted deletion of parkin induce apoptotic neuron death whereas overexpression of the wild-type genes prevent apoptosis. In addition, treatment with the dopaminergic neurotoxins such as 6-hydroxydopamine and 1-methyl, 4-phenyl, 1,2,3,6-tetrahydropyridine (MPTP) induce alterations in Bcl-2 family members, including increases in the expression of proapoptotic Bax and PUMA and decreases in expression of Bcl-2. MPTP-induced neurodegeneration can also be spared by the targeted deletion of Bax. Furthermore, the activation of both initiator (caspase-8 and caspase-9) and effector (caspase-3) caspases has been demonstrated in cultured dopaminergic neurons and in neurons of mice treated with MPTP. Thus, data from both *in vivo* and *in vitro* models of PD suggest a link between PD and apoptotic death and raise the possibility that inhibition of apoptotic death may prove beneficial in PD treatment strategies.

Huntington's Disease

Huntington's disease (HD) induces chronic, progressive motor dysfunction and is caused by a mutation generating expanded CAG repeats in the *huntingtin* gene. In general, HD onset and prevalence is proportional to the number of CAG repeats. In the normal population, the number of CAG repeats varies from 6 to 35, but an increased risk is associated with between 36 and 41 CAG repeats and those with 42 or greater CAG repeats almost always manifest HD. The striatum is the principal area affected in HD with reports from postmortem studies demonstrating as much as 95% loss of striatal GABAergic, medium spiny neurons. In more severe cases of HD, other brain regions including the cortex, thalamus, globus pallidus, and hippocampus may also be affected.

Similar to AD and PD, there are conflicting reports of apoptosis in HD brain. Early reports of apoptosis in HD brain indicated increased TUNEL reactivity, but electrophoretic analysis of DNA fragments in HD brain did not reveal a characteristic apoptotic pattern, which possibly suggests a role for additional death processes, such as necrosis in HD brain. Regardless, there are scattered reports of caspase-8, caspase-9, and caspase-3 activation in HD brain, in addition to an increase in cytosolic cytochrome *c*, suggesting possible activation of both the extrinsic and intrinsic apoptotic pathways in HD. Increased levels of Bax and Bcl-2 have also been reported in HD brain specific to the caudate nucleus, which may be interpreted as an antiapoptotic survival response to an increase of Bax in stressed neurons. More studies are needed to determine the role of extrinsic versus intrinsic apoptotic death pathways in HD brain and the potential altered regulation of Bcl-2 family members in this devastating disease.

Animal models of HD, however, show more convincing evidence of a proapoptotic tone of Bcl-2 - family members in affected neuronal populations. One of the most well-studied mouse models of HD is the R6/2 transgenic mouse, which expresses only exon 1 of the *huntingtin* gene with 145 CAG repeats, shows a dramatic HD phenotype, and dies by 4 months of age. R6/2 mice show increases in mitochondrial Bim and Bax, a decrease in phosphorylated Bad, and activation of caspase-1 and caspase-3. Levels of Bcl-2 or Bcl-X_L are unchanged in these mice, but the overexpression of Bcl-2 has been shown to slow disease progression. Treatment with 3-nitroprussic acid (3-NP), a chemical inhibitor of mitochondrial succinate dehydrogenase used to model HD, induces a selective striatal lesion when administered *in vivo* that increases the ratio of Bax/Bcl-2 and Bax/Bcl-X_L, thus shifting their balance to a proapoptotic state.

Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS) is a fatal disease that affects one to two persons per 100 000 and is characterized by the progressive death of upper and lower motor neurons and eventual loss of motor function. Approximately 10% of ALS cases are familial, and of these 20% are linked to mutations in the gene encoding superoxide dismutase 1 (SOD1). It is well accepted that a percentage of dying motor neurons in ALS undergoes apoptosis, and apoptotic morphology has been identified in motor neurons of the human spinal cord and motor cortex. However, like most chronic neurodegenerative diseases, a major challenge in the study of apoptosis in ALS is the relative rapidity of the apoptotic process in the face of a disease with a slow time course, such that at any one time very few motor neurons are actively undergoing apoptosis. As such, there are reports indicating little, if any, evidence for increased apoptotic morphology in ALS cases when compared to control brains.

Studies of postmortem ALS spinal cord suggest a shift toward proapoptotic members of the Bcl-2 family, such as increased expression of Bax and DP5/Hrk and decreased expression of Bcl-2. Co-immunoprecipitation studies have also indicated an increase in Bax–Bax interactions and a decrease in Bax–Bcl-2 interactions in ALS tissue, which also suggests a shift to a proapoptotic state. A decrease in Bcl-2 may be of greater significance in ALS than in other neurodegenerative diseases since Bcl-2 remains high after neurodevelopment in the spinal cord relative to the brain, which suggests an important regulatory function of Bcl-2 in postmitotic spinal cord motor neurons. Analysis of symptomatic, transgenic mice with mutant SOD1 also shows decreased expression of Bcl-2 and Bcl-X_L and increased expression of Bax and Bad. Truncated Bid is also evidenced in mutant SOD1 mice, which may suggest participation of the extrinsic apoptotic pathway in ALS.

In vitro studies of cultured motor neurons from mutant SOD1 mice indicate a role for Fas-mediated activation of caspase-8 and the extrinsic apoptotic pathway, but these same studies also report release of cytochrome *c* from mitochondria, suggesting activation of the intrinsic apoptotic pathway. Reports of altered regulation of Bax and Bcl-2 in human ALS would clearly support a role for the intrinsic apoptotic pathway in ALS. While there is no current evidence supporting the direct activation or altered expression of caspase-8 in human ALS, recent analysis of sera obtained from human ALS patients indicates an increase in antibodies to Fas that induced death of cultured motor neurons, which further suggests the contribution of the extrinsic apoptotic pathway. In addition to caspase-8, *in vitro* and animal models of ALS have shown activation of caspase-1, caspase-12, caspase-9, and caspase-3, while reports of caspase activation in human ALS tissue are limited to activation of caspase-1 and caspase-3. While good evidence exists supporting a role for apoptosis in ALS, more studies are needed to determine its relative contribution to the overall disease pathogenesis.

Nonapoptotic Neurodegeneration

The reported inconsistencies for the role of apoptosis in neurodegenerative disease have led to an increasing number of studies investigating potential mechanisms of caspase-independent neuron death. Apoptosis-inducing factor (AIF) is a small protein that resides normally within the intermembrane space of mitochondria, and upon certain death stimuli translocates into the cytosol and ultimately the nucleus where it contributes to DNA fragmentation and chromatin condensation. The death-promoting effects of AIF have been shown to be independent of caspase activation, although the mitochondrial release of AIF has been shown to be Bax dependent. A role for AIF in cell death has been implicated in many neurodegenerative diseases including AD, PD, and ALS. Calpains are cytosolic proteases that are activated upon ER stress-induced intracellular calcium influx. While calpains have been shown to act downstream of caspases, many studies indicate caspase-independent, calpain-induced cell death. Calpain activation has been implicated in AD, PD, HD, and ALS. For instance, in an *in vitro* model of AD, the activation of calpains has been implicated in the Aβ-induced, caspase-independent degeneration of axons. α-Synuclein has also been shown to be a substrate of activated calpain, which may lead to its aberrant deposition in PD.

The increased aggregation and deposition of specific cytoplasmic proteins in neurodegenerative diseases has implicated a role for disturbances in lysosomal function and/or type II autophagic cell death. Autophagy is a normal cellular process whereby vesicles shuttle cytoplasmic and organellar macronutrients through a series of fusion events to the lysosome for ultimate proteolytic degradation and recycling. Autophagic cell death is characterized by an aberrant accumulation of vacuoles that result either as a response to metabolic stress or to a disruption in lysosomal function, concomitant with morphological features of both apoptosis and necrosis. An accumulation of autophagic vacuoles has been clearly documented in AD, PD, and HD, which suggests that aberrant autophagic stress plays a significant role in the pathogenesis of these diseases.

Age-related lysosome disturbances may contribute to a compromise in protein degradation that allows for the accumulation of specific proteins in neurodegenerative disease. In addition, a compromise in the integrity of lysosome membranes may also induce leakage of lysosomal proteases, such as cathepsins, into the cytosol. Cathepsins have been shown to induce Bid- and Bax-dependent activation of the intrinsic apoptotic pathway and Bax-dependent release of AIF into the cytosol. In addition, cathepsins can also act as efficient effector proteases similar to that of effector caspases, thus creating another potential mechanism of caspase-independent death in neurodegenerative disease.

Summary

Apoptotic death is a well-characterized and tightly orchestrated cell biological process whose involvement in normal brain development is unquestionable. Apoptotic neurons have been detected to varying degrees in virtually all human neurodegenerative diseases, and alterations in the expression of numerous apoptosis-associated genes and proteins have been widely reported. However, the significance of these findings is unclear, and separating correlation from causation has proven difficult. It remains possible that apoptosis-associated molecules and sublethal intracellular events contribute to neuron dysfunction early in disease pathogenesis, prior to significant loss of neurons, but definitive data on this conjecture are currently lacking for most neurodegenerative diseases. Neuroprotective agents that can inhibit both apoptotic and nonapoptotic neuron death may ultimately be used in conjunction with drugs designed to decrease disease-specific death stimuli, resulting in effective combination therapy to preserve neuron function in AD, PD, HD, ALS, and other neurodegenerative diseases.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Animal Models of Motor and Sensory Neuron Disease; Animal Models of Alzheimer's Disease; Animal Models of Huntington's Disease; Apoptosis in Nervous System Injury; Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autophagy and Neuronal Death; Huntington's Disease: Neurodegeneration; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Boatright KM and Salvesen GS (2004) Mechanisms of caspase activation. *Current Opinion in Cell Biology* 15: 725–731.

Clarke PG (1990) Developmental cell death. *Anatomy and Embryology* 181: 195–213.

Ellis HM and Horvitz HR (1986) Genetic control of programmed cell death in the nematode *C. elegans*. *Cell* 44: 817–829.

Hickey MA and Chesselet MF (2003) Apoptosis in Huntington's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 27: 255–265.

Kerr JF, Wyllie AH, and Currie AR (1972) Apoptosis: A basic biological phenomenon with wide-ranging implications in tissue kinetics. *British Journal of Cancer* 26: 239–257.

Putcha GV, Harris CA, Moulder KL, et al. (2002) Intrinsic and extrinsic pathway signaling during neuronal apoptosis: Lessons from the analysis of mutant mice. *Journal of Cell Biology* 157: 441–453.

Roth KA (2001) Caspases, apoptosis and Alzheimer disease: Causation, correlation, and confusion. *Journal of Neuropathology and Experimental Neurology* 60: 829–838.

Sathasivam S, Ince PG, and Shaw PJ (2001) Apoptosis in amyotrophic lateral sclerosis: A review of the evidence. *Neuropathology and Applied Neurobiology* 27: 257–274.

Shacka JJ and Roth KA (2004) Regulation of neuronal cell death and neurodegeneration by members of the Bcl-2 family: Therapeutic implications. *Current Drug Targets – CNS and Neurological Disorders* 4: 25–39.

Tatton WG, Chalmers-Redman R, Brown D, and Tatton N (2003) Apoptosis in Parkinson's disease: Signals for neuronal degradation. *Annals of Neurology* 53: S61–S70.

Appetitive Systems: Amygdala and Striatum

J A Parkinson, Bangor University, Bangor, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Imagine going to a grocery store for some milk. While strolling the aisles, a carton of Madagascan Vanilla Custard catches your eye; it sounds delicious and would be great to have lashings of this custard with some seasonal fruit for dessert this evening. Into the basket it goes, and before you know it you will be sitting at home indulging your extravagant desires. Even though there was no intention to buy the custard prior to entering the shop, the mere exposure to the salient packaging has captured attention, elicited appetitive arousal, established the Madagascan Vanilla Custard as an important goal for prospective behavior, and even generated an expectancy for the consequences of such a course of action. This simple example captures the essence of what interactions between the amygdala and striatum are believed to underlie.

Major advances in our understanding of emotion have come about through studying model systems in animals, such as fear conditioning in rats. Similarly, our understanding of positive emotions and motivation has also been furthered through the study of appetitive processes in animals and in particular anticipatory, or preparatory, responses to food, sex, and drugs. Thus, whereas fear can be conceived as the anticipation of an unpleasant, or aversive, event, appetitive behavior can be conceived as the anticipation of a pleasant, or desirable, event. The amygdala appears to be a critical mechanism for these anticipatory processes. As far back as the middle of the last century it was suggested that the amygdala complex processed stimulus–reward associations, in that it allowed sensory representations of the world to be linked with affective values. More recently, in the 1980s, researchers suggested that the ventral striatum (VS), a structure that receives signals from the amygdala, acted as a limbic–motor interface allowing motivational states and emotional information to influence action selection. Both of these seminal propositions are broadly accepted; this article describes contemporary neuroscientific research which has refined these ideas and provided a more precise indication of how these two structures interact. Neither the amygdala nor striatum is a unitary structure, and their function can be better understood through an understanding of their anatomical structure and relationships. Indeed, it is likely that subregions of the amygdala and striatum play dissociable yet complementary roles in affective processing.

Anatomical Routes for Amygdalar–Striatal Interactions

Although the amygdala was traditionally viewed as a single structure, detailed cytoarchitectural, neurochemical, and connectional analyses have now supported a multicompartmental view with at least three major subdivisions: (1) a phylogenetically newer cortex-like region (basolateral amygdala regions (BLA)), (2) part of the olfactory system, and (3) an extension of the striatum (central amygdala region (CEA)). The first two regions use glutamate as their primary extrinsic neurotransmitter and can be considered parts of the cortex. The last region shares many features with the neighboring striatum, such as having a dense preponderance of γ-aminobutyric acid (GABA) and a similar pattern of neuropeptide expression. Neuroscientific evidence supports interactions primarily between the BLA and the CEA with the ventromedial regions of the striatum within the context of two organizing models of anatomical connectivity and structure: corticostriatal circuitry and extended amygdala.

Corticostriatal Circuitry

The striatum is primarily viewed as a major input station of the basal ganglia, a system that is critically modulated by dopamine and which supports sensorymotor initiation and control. The basal ganglia consists of a number of independent and parallel circuits (see Figure 1) which are topographically organized. These cortico-striatal-pallidal-thalamic loops (or corticostriatal loops, for the purpose of this article) are capable of selection and amplification, mapping the patterns of sensory input on to response output within a particular organizing context. In effect, these loops can maintain and remap the process of response selection.

The number of functional corticostriatal circuits is not fully resolved and is a matter of considerable debate. Current thinking acknowledges (1) at least one sensorimotor circuit running through lateral striatum (putamen); (2) at least one dorsomedial (caudate) circuit underlying cognitive, or goal-directed, functions; and (3) a ventral striatal circuit that underlies stimulus-induced motivation and is implicated in excessive appetite and addiction. The BLA provides

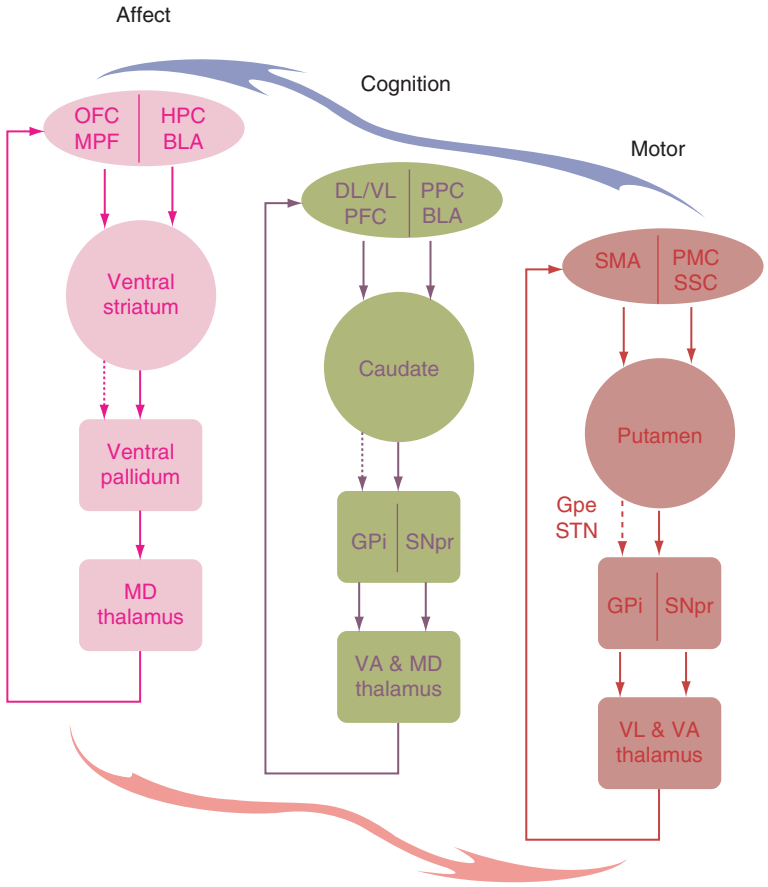

Figure 1 Three segregated, parallel corticostriatal loops are shown with their putative function. Each basal ganglia circuit includes cortical input, striatum, pallidum, and thalamus. The precise manner in which parallel loops interact is not known, although hierarchical control over behavior from medial (affective) to lateral (sensorimotor) circuits is achieved by descending influence of each circuit over the midbrain dopamine (DA) innervation of the next circuit laterally (indicated by red arrow) enabling integration of value and goal information, knowledge-based strategy, and sensorimotor coordination in the pursuit of goal-directed behavior. Further, each circuit may provide feedback (lateral to medial) at the level of cortex (indicated by blue arrow). Finally, local within-loop ensembles may show lateral inhibition across loops (not shown). BLA, basolateral amygdala; DL/VL PFC, dorsolateral and ventrolateral prefrontal cortex; GPe, external segment of the globus pallidus; GPi, internal segment of the globus pallidus; HPC, hippocampal formation; MD, mediodorsal; MPF, medial prefrontal cortex; OFC, orbitofrontal cortex; PMC, premotor cortex; PPC, posterior parietal cortex; SMA, supplementary motor area; SNpr, substantia nigra, pars reticulata; SSC, somatosensory cortex; STN, subthalamic nucleus; VA ventral anterior; VL, ventrolateral.

direct input across the medial striatum, including both the VS and caudate circuits. There are also indirect connections from the CEA to ascending arousal systems and thence to the striatum (see Figure 2). Efferent projections from the VS show a marked divergence across the medial–lateral axis. The lateral core subregion, which merges into overlying caudate, is the more typically striatal, projecting to the ventral pallidum. The medial shell subregion projects not only to the ventral pallidum but also to the lateral hypothalamus and bed nucleus of the stria terminalis and is considered part of the extended amygdala.

Extended Amygdala

The concept of the extended amygdala reflects the developmental similarity between cell groups of the amygdala, specifically the striatal-like centromedial nuclei and surrounding tissue such as the bed nucleus of the stria terminalis. A topographic continuum of cells courses through the forebrain and forms a loop around the internal capsule. In contrast to traditional output systems of the corticostriatal circuits, the extended amygdala is typified by its downstream connections to autonomic and endocrine centers, allowing the coordination of species-typical adaptive responses (labeled brain stem effectors in Figure 2).

Function of Amygdala–Striata Interactions

Overview

Along with the orbital prefrontal cortex (OFC), communication between the amygdala and striatum underlies the anticipation of emotional events and expectations of the consequences of choices and

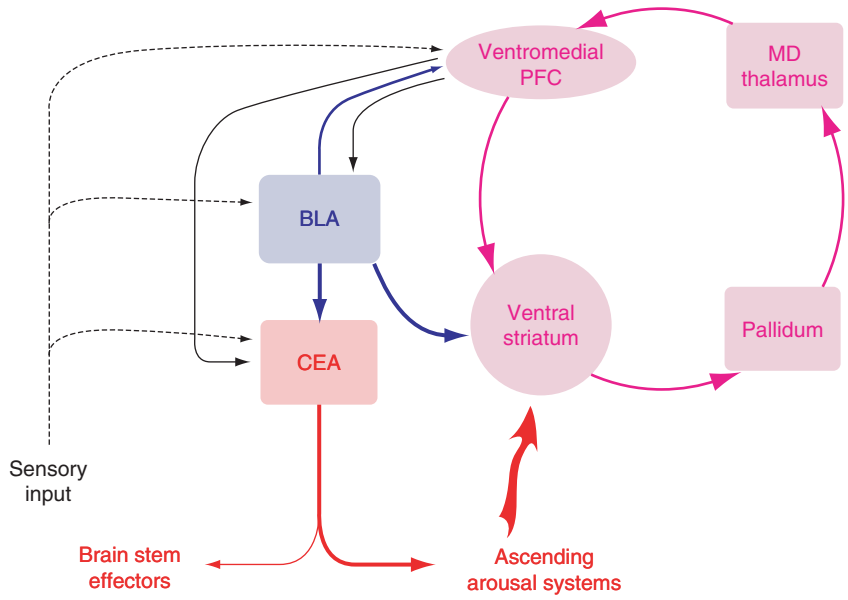

Figure 2 Affective corticostriatal circuit emphasizing two anatomical routes through which the amygdala interacts with the striatum. Orbital and medial prefrontal (ventromedial PFC) afferents converge with amygdala inputs (BLA; blue arrow) at the level of the VS. Modulation of corticostriatal circuitry by dopamine (and other ascending arousal systems) also depends on signals from the amygdala (CEA; red arrow). This interaction between CEA and VS has been implicated in two related processes: (1) providing a reinforcement signal for associative learning and (2) modulating the vigor of a motivated instrumental response (also see the section titled 'Mesolimbic dopamine contributes to both motivational learning and performance' and Figure 3). BLA, basolateral amygdala; CEA, central amygdala; MD, mediodorsal; PFC, prefrontal cortex; VS, ventral striatum. Reproduced from Parkinson JA, Cardinal RN, and Everitt BJ (2000) Limbic cortico-ventral striatal systems underlying appetitive conditioning. In: Uylings HBM, Van Eden CG, De Bruin JPC, et al. (eds.) *Progress in Brain Research*, vol. 126, pp. 263–285. Amsterdam: Elsevier Science, with permission from Elsevier.

behavior. Initial insights into amygdala function came from observations of the effects of gross damage in nonhuman primates. Such animals seemed incapable of learning the association between an object and its emotional, or motivational, value (such as the sight of a banana and its value as a tasty food). This impairment was described as being an inability to form stimulus–reward associations; that is, it is not an impairment in stimulus processing *per se* nor in the ability to experience the hedonic impact of the food outcomes once tasted, but in the relationship between the two: the ability to predict, indeed expect, a particular outcome given its associated stimulus.

Anatomical distinctions within the amygdala have been supported through functional studies. The BLA shares connectivity and structure with the prefrontal and temporal cortex and has been shown to support mnemonic representations of value (e.g., evidence for long-term potentiation underlying associations between a sensory stimulus and its affective value). The BLA may therefore contribute to a system that provides a signal of the affective value of specific objects and events in the world (blue arrows in Figures 2 and 3). The CEA, which sits downstream of the BLA appears to play a greater role in the acquisition and performance of anticipatory conditioned

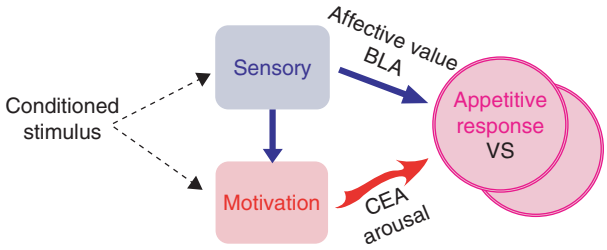

Figure 3 Psychological processes by which affect-laden stimuli can influence cognition and appetitive instrumental behavior. Associations between a conditioned stimulus and an affective outcome can be formed directly between the specific sensory features or with the more general motivational features. The CEA provides the general motivational arousal elicited by conditioned stimuli, whereas the BLA provides the direction for motivational responses based on the specific sensory and motivational properties of a stimulus. BLA, basolateral amygdala; CEA, central amygdala; VS, ventral striatum. Reproduced from Balleine BW (2005) Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology and Behavior* 86: 717–730, with permission from Elsevier.

responses, that is, in the way that reflexive, species-typical responses can be triggered by environmental stimuli in appropriate circumstances (freezing to threat or orienting to a stimulus-predicting reward). The CEA also has access to ascending arousal systems

and may influence affective learning and performance through this route (red arrows in Figures 2 and 3).

The Basolateral Amygdala and Stimulus–Affect Learning

Sophisticated behavioral tests can be employed to dissect distinct psychological processes that an individual may employ to achieve a goal. One such paradigm, outcome devaluation, involves the ability to react to a change in affective value of a stimulus after its initial value has been learned. For example, after having eaten a seafood bouillon and subsequently becoming ill, most people would avoid choosing the same, or similar, dish in future forays to the restaurant, even without tasting the food. This demonstrates an ability to represent the value of food and subsequently modulate or change that value property as a result of new experiences. Witness a more dynamic version at mealtimes when, after experiencing a filling main course, an individual may readily desire a sugary dessert, that is, a transient reduction in value of the main course but a maintenance of value for sugary dessert, sensory-specific satiety. Our choices reflect the changing incentive values of the competing foods. Animal studies of food valuation and revaluation (including outcome devaluation) have been employed to explore the neural mechanisms of goal-directed behavior.

Within this broad paradigm, researchers have employed various neuroscientific techniques including regional lesions, regional inactivations, electrophysiology, and functional magnetic resonance imaging (fMRI) in an attempt to elucidate the neural basis of valuation, expectation, and choice. Disruption in the function of either the BLA or OFC in rats or primates carried out prior to training abolishes the ability of animals to adapt to changes in reinforcer value (e.g., devaluation) on such tasks. In contrast, these lesions do not appear to disrupt motoric ability, general activity, or hedonic responses to rewards themselves. Electrophysiological recording of neurons in awake behaving rodents and primates has further refined our understanding: neurons in the BLA develop selective firing both to the delivery of a particular affective outcome and to the stimulus that predicts it. Further, some neurons that fire to the reinforcer also develop responding during the delay between a behavioral response and the subsequent outcome. As such, the former neurons appear to code the relationship between stimulus and outcome (a form of passive expectancy), whereas others code the expected nature of the outcome following a response (a more active response-based expectation). If the affective nature of the outcomes is reversed, so that a stimulus that used to predict a pleasant outcome now predicts an unpleasant one, the responding of the anticipatory neurons also changes, critically demonstrating that the neurons are tracking the valence of the outcome, not just its sensory properties. Similar activity has also been observed in OFC neurons. However, the selectivity in neuronal firing in the OFC appears to develop later than in the BLA and tracks the emergence of discriminated behavior. Further, OFC neurons tend not to reverse with changing outcome contingencies and, instead, simply lose their selectivity, whereas previously silent neurons acquire selectivity to the new contingencies. This indicates that the OFC is more tightly linked to action selection and the BLA is more tightly linked to stimulus value.

Human neuroimaging studies have arrived at similar conclusions when adopting paradigmatic approaches employed in animal work. Such studies have further shown that activity in the amygdala correlates with the magnitude of arousal elicited by affective stimuli (including pictures, words, tastes and smell of foods, pictures of sexual stimuli, and positively valenced pictures of social interactions). In other words, the level of activity in the amygdala reflects the subjective perception of affective value. OFC activity also responds to the predicted stimulus value, but appears to code for different response options and their expected consequences.

The Central Amygdala and Conditioned Anticipatory Arousal

Much of the foregoing discussion has centered on the BLA and its role in stimulus–affect learning. The CEA appears to play a different role in the learning and performance of appetitive behavior. Damage to the CEA does not impair the ability of animals to perform the sort of tasks that we have described, such as incentive learning and devaluation. As such, many appetitive motivational processes are dissociable between these two amygdala subregions. Lesions of the CEA do abolish learned orienting responses to affective stimuli and also abolish conditioned approach behavior to cues that predict appetitive events – in general terms, the acquisition of attentional and motivational anticipatory responses. The CEA also provides direct input to specific neurotransmitter systems of the ascending arousal systems such as dopamine (DA). Disruption of the CEA–DA–dorsal striatum system impairs attentional learning, whereas manipulations of the CEA–DA–VS system affect the acquisition of motivated approach.

The CEA and BLA can also be dissociated in their influence on the performance of instrumental behavior (see Figure 3). One of the most powerful forms of motivation is the process by which everyday stimuli

can induce motivational arousal, even in the absence of a biological need, and influence the vigor and direction of behavior. Indeed, theories of excessive appetite, including overeating and relapse to drug use, posit a major role for stimulus-induced arousal and desire. An important animal model of such processes examines the way affective stimuli influence ongoing behavior (termed Pavlovian to instrumental transfer (PIT)). PIT can have its effects specifically; that is, the perceived value of a stimulus matches a particular motivational need (known as sensory-specific PIT). Alternatively, it can elicit conditioned motivation, producing a general arousal of behavior irrespective of motivational needs (known as general PIT). Amygdalar–striatal circuitry is critical for both forms of PIT; whereas the CEA is important for the general arousal elicited by a stimulus, the BLA provides specific direction for stimulus-induced motivation (as shown in Figure 3). Further, the VS provides the route for expression of PIT potentially through a combination of direct projections from the BLA and also from downstream projections of the CEA to ascending arousal systems, including DA; intra-VS infusions of DA agonists enhance the PIT response, and systemic blockade of DA abolishes the motivational response.

The Striatum as a Site of Integration and Selection

The VS was famously described as a limbic–motor interface using information about the value of stimuli in the world to guide behavior. Another avenue of research into VS function relates to its involvement, and in particular its dopaminergic innervation, in drug reward. The VS was once thought to underlie the hedonic component of reward; damage to this structure reduced drug-related responses in rodents on tasks modeling drug reward. However, these tasks sometimes confounded dissociable psychological processes such as drug-seeking, anticipatory arousal, and actual drug ingestion, making it difficult to be certain what the effects of VS or dopaminergic manipulations were. Indeed, ingestion of drugs (or food or copulation) is rarely abolished by striatal damage or by manipulation of the striatal DA system. Instead, rises in extracellular VS DA are predominantly seen in anticipation of appetitive events, and manipulations reducing the function of the VS most often reduce anticipatory motivation for reinforcers (again shown with recreational drugs, food, and sex). Other neurochemical systems, notably opioids, may contribute to consummatory/hedonic processes.

Electrophysiological studies in rats and primates have shown that different classes of neuron in the medial striatum code for reward anticipation and for stimulus–response mapping. Both the VS and overlying caudate receive input from the OFC and amygdala, and it is perhaps no surprise that each class of these neurons also appears to exist in one or other of these afferent structures; in this case, the amygdala preferentially shares stimulus–outcome coding, whereas the OFC provides information on responses and their consequences. The VS and overlying caudate therefore serve as integration sites for this information in order to allow appropriate behavior to be initiated and maintained. The distinction between these two corticostriatal circuits is that, whereas the VS circuit underlies stimulus-bound motivational arousal, the caudate circuit contributes to cognitive aspects of goal-directed behavior such as attention and flexibility. Further, the medial and orbital prefrontal cortex (PFC), afferents to the striatum, have been implicated in extinction and habit control (medial) and in inhibition and reversal (orbital). As such, input from the PFC to the striatum appears to provide top-down control of response selection, particularly in allowing flexible behavior as a result of changing environmental contingencies or changing internal motivational and emotional states.

Mesolimbic Dopamine Contributes to Both Motivational Learning and Performance

Seminal electrophysiological work demonstrated that DA neurons in the midbrain, as well as the VS neurons to which they project, show phasic firing patterns to unexpected affective outcomes. If a stimulus predicts the outcome, then neuronal firing begins to track the predictive stimulus rather than the outcome. The omission of an expected stimulus or outcome also elicits phasic firing. This pattern appears to code for a prediction error – a reinforcement signal – which supports learning by predicting the likelihood of an outcome given that a behavioral response has been made or a conditioned stimulus has been presented. Many learning models have incorporated this signal, although at its simplest, DA may act as a reinforcement teaching signal to stamp in the association between a stimulus and the appropriate response, that is, to help create stimulus–response mappings through identifying predictive relationships in the world.

In other work, gross manipulation of DA levels in the VS modulate the vigor, or effort, with which animals respond for reinforcement. Indeed, human subjective experience of psychostimulant drugs (such as cocaine and amphetamine, which enhance DA function in the VS) often includes increased feelings of vigor and behavioral output. Rats with reduced VS DA select less effortful courses of action to gain reinforcement and choose less preferred foods if greater effort is required to get the preferred food. Infusions of DA agonists into the VS specifically potentiate

responding for reward-related stimuli and outcomes, rather than potentiating all behavior, and enhance the desire (or wanting) for valued outcomes.

In summary, phasic DA signals appear to underlie simple affective learning, whereas tonic levels of DA in the extracellular matrix of the VS provide the energy or vigor for responding. It has also been suggested that tonic levels of DA play a permissive role in response selection, and so response vigor may simply be a computational consequence of the response selection process. Interestingly, damage to the CEA abolishes the enhancing effect of intra-VS infusions of DA and impairs the acquisition of appetitive approach behavior (which is also dependent on the VS). As a result, several authors have speculated that the CEA may play a role in stimulus–response learning via influence on the phasic firing of DA neurons and also a role in the selection and arousal of performance by modulating a more tonic level of extracellular DA in the VS (red arrows in Figures 2 and 3). Finally, the extracellular levels of DA appear to be critically modulated by incoming afferents such as the BLA and hippocampal formation and may therefore relate to calculations of value and of context. Again, the VS and its dopaminergic innervation can be seen as a mechanism of integration between sensory input and appropriate behavioral output.

Amygdalar–Striatal Interactions and Drug Addiction

It is likely that drugs exert their effects on emotion and motivation through the same neural circuits that underlie natural rewards, although perhaps in a more potent manner. It is not surprising then that certain aspects of the addictive process relate to amygdalar–striatal function. Drugs and associated drug stimuli (including paraphernalia, incidental stimuli, and contexts) acquire incentive value through association with drug taking. Such stimuli themselves generate affective arousal and lead to increased anticipation and desire for drugs. This process of stimulus-induced arousal and desire can contribute to two important aspects of drug addiction: (1) the abnormal control over motivation and behavior that drugs produce, such as focusing on drug-seeking behavior to the detriment of other natural motivational needs; and (2) the high relapse rates in individuals attempting to abstain from drug use. Stimulus-induced craving is thought to be a major contributor to relapse and the maintenance of abuse. Drug-associated stimuli can therefore be seen as possessing an abnormal level of control over goal-directed processes, can suppress (healthy) inhibition, and can lead to maladaptive motivated behavior.

The motivational impact of drug-associated cues depends on the VS and its dopaminergic innervation. Disruption of VS function, or of incoming glutamatergic signals, reduces drug seeking, probably by reducing motivational arousal normally elicited by affective cues. Indeed, neuroadaptations in amygdalar–striatal circuitry have been observed following chronic drug use, such as sensitization of the DA system; such adaptations are now being documented with food stimuli. Indeed, the release of endogenous opioids in the VS is seen following consumption of energy-dense foods, just as similar changes in opioid levels are observed following heroin self-administration in rats. This growing link between drug and natural rewards highlights the commonality in neural substrates of appetitive motivational control.

Summary

Through experience, individuals learn the affective structure of the world, enabling stimuli endowed with value to provide an important impetus for behavior. The amygdala appears to encode the expected value of a predicted event, and the VS contributes to the formation and expression of motivational responses. These processes are maintained by DA signals that support value-based learning and also modulate the direction and energy underlying behavior. The OFC most likely contributes by providing a top-down control over goal-directed response selection, including functions such as inhibition, flexibility, and choice.

One of the major functions of the brain is to predict and time events. A specific appetitive function of amygdalar–striatal interactions is to predict events of adaptive (emotional and motivational) significance. The position of the amygdala and VS at the anatomical head of corticostriatal circuitry allows them a preeminent role in enabling affect-laden stimuli to influence an individual's choice of goals and course of action.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Structure and Circuitry in Rodents and Felines; Amygdala: Contributions to Fear; Conditioning: Theories; Dopamine Control of Arousal; Dopamine; Drug Addiction: Behavioral Neurophysiology; Drug Addiction: Behavioral Pharmacology of Drug Addiction in Rats; Emotion Systems and the Brain; Goal-Directed Behavior Theories; Prefrontal Contributions to Reward Encoding; Procedural Learning: Striatum; Psychopharmacology of Reward and Appetite in Rats; Reward Systems: Human; Striatum: Internal Physiology.

Further Reading

Alexander GE, Delong MR, and Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience* 9: 357–381.

Balleine BW (2005) Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology and Behavior* 86: 717–730.

Baxter MG and Murray EA (2002) The amygdala and reward. *Nature Reviews Neuroscience* 3: 563–573.

Berridge KC (2004) Motivation concepts in behavioral neuroscience. *Physiology and Behavior* 81: 179–209.

Cardinal RN and Everitt BJ (2004) Neural and psychological mechanisms underlying appetitive learning: Links to drug addiction. *Current Opinion in Neurobiology* 14: 156–162.

Daw ND, Niv Y, and Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8: 1704–1711.

Haber SN, Fudge JL, and McFarland NR (2000) Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience* 20: 2369–2382.

Holland PC and Gallagher M (2004) Amygdala–frontal interactions and reward expectancy. *Current Opinion Neurobiology* 14: 148–155.

Kelley AE, Baldo BA, Pratt WE, and Will MJ (2005) Corticostriatal-hypothalamic circuitry and food motivation: Integration of energy, action and reward. *Physiology and Behavior* 86: 773–795.

Parkinson JA, Cardinal RN, and Everitt BJ (2000) Limbic cortico-ventral striatal systems underlying appetitive conditioning. In: Uylings HBM, Van Eden CG, De Bruin JPC, et al. (eds.) *Progress in Brain Research,* vol. 126, pp. 263–285. Amsterdam: Elsevier Science.

Redgrave P, Prescott TJ, and Gurney K (1999) The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience* 89: 1009–1023.

Salamone JD, Correa M, Mingote SM, and Weber SM (2005) Beyond the reward hypothesis: Alternative functions of nucleus accumbens dopamine. *Current Opinion in Pharmacology* 5: 34–41.

Schultz W and Dickinson A (2000) Neuronal coding of prediction errors. *Annual Review of Neuroscience* 23: 473–500.

Swanson LW and Petrovich GD (1998) What is the amygdala? *Trends in Neurosciences* 21: 323–331.

Yin HH and Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* 7: 464–476.

Apraxia: Disease

G Goldenberg, Krankenhaus Munchen Bogenhausen, Munich, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term apraxia denotes 'higher-order' disturbances of motor control. Apraxia differs from more mundane disorders of motor skill like hemiparesis or ataxia by its bilaterality: unilateral hemisphere lesions cause apraxia of both limbs. Consequently, apraxia can be demonstrated on the limb ipsilateral to the lesioned hemisphere and contrasts with normal motor skill of the same limb in other situations. This definition embraces a wide variety of heterogeneous disturbances, but traditionally diagnosis and research on apraxia concentrate on three domains of human motor actions: imitation of gestures, performance of communicative gestures on command, and skillful use of tools and objects.

A seminal model of apraxia was elaborated by the German psychiatrist Hugo Liepmann about a hundred years ago. He distinguished between two sequential stages of high-level motor control. In the first stage, a concept or mental image of the intended action is elaborated, and in the second stage, this mental image is translated into appropriate motor commands. Liepmann thought that generation of mental images of the intended actions is accomplished by posterior regions of the brain and that their translation into motor commands depends on integrity of fiber paths linking posterior brain regions to the motor cortex. Parietal lesions can sever these fiber paths and are hence a locus of predilection for causing apraxia. Liepmann's most important empirical contribution to the study of apraxia was the observation that it is predominantly a symptom of left brain damage (LBD). Most afflicted patients are aphasic, but the severities of aphasia and apraxia are not tightly correlated, and there are LBD patients who display only one of them. This led Liepmann to speculate that there is left hemisphere dominance for motor control in addition to its well-established dominance for speech and language.

It may appear somewhat untimely to start this article with a theory proposed at the start of the twentieth century, but indeed Liepmann's ideas have remained influential and have shaped the study of apraxia until now. Three basic assumptions have received different interpretation but have remained essentially unchallenged until recently (see Figure 1):

- There is a sequentially ordered 'praxis system,' interruption of which underlies all manifestations of apraxia.
- This praxis system is located in the left hemisphere.
- The parietal lobe has a predominant role within this system.

This article discusses each of the manifestations of apraxia separately, with an emphasis on the role of LBD and particularly of left parietal lesions. Then it questions the existence of a unique praxis system and discusses the usefulness of classifying apraxia as 'ideational' or 'ideomotor.'

Imitation of Gestures

Defective imitation is an easily demonstrable and impressive symptom of unilateral brain damage. Even if patients are severely aphasic, they usually understand the instruction to imitate and obviously strain to copy the gesture made by the examiner, but although they use the otherwise normally skillful hand ipsilateral to their brain lesion, they produce insecure, hesitant, and searching movements which end up in positions quite grossly different from those demonstrated.

The presence and severity of impairment may depend on the kind of gesture that is being tested. A first major divide is between meaningful and meaningless gestures. If the examiner demonstrates a meaningful gesture, patients may understand the meaning and reproduce the gesture out of a repertoire of meaningful gestures stored in long-term memory rather than copying the shape of the gesture. By contrast, meaningless gestures usually have no representation in long-term memory, at least not when tested for the first time. Their imitation must translate the shape of the gesture from visual perception of another person's body to motor commands configuring their own body without support from the repertoire of habitual gestures. The mechanisms employed for meaningless gestures can also support imitation of meaningful gestures, as their shape can be copied without taking account of their meaning. Intermingled presentation of meaningful and meaningless gestures can blur their distinction because participants then tend to make a strategic choice of using the same route for both kinds of gestures and treat meaningful gestures as if they were meaningless. Studies which tested imitation of meaningful and meaningless gestures separately brought forward robust evidence that the two routes

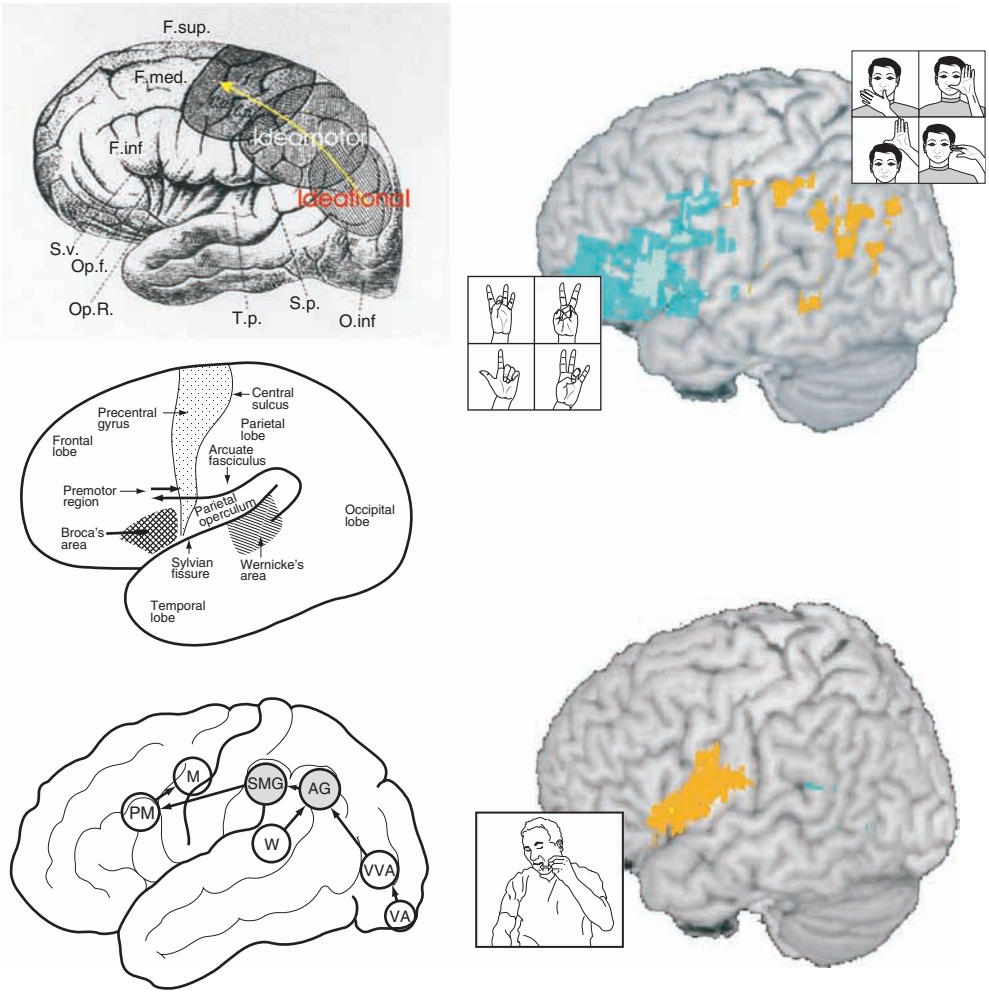

Figure 1 (Left side) Three versions of the 'praxis system.' They concur in the assumption of a posterior to anterior stream of action control with a central role for the parietal lobe. (Right side:) MRI-lesion subtraction identifying left hemisphere regions, damage to which is crucial for different manifestations of apraxia. (Above) Imitation of hand postures is susceptible to lesions in inferior parietal lobe and the parietotemporal junction, while imitation of finger postures is mainly affected by lesions in the inferior frontal lobe. (Below) Disturbed pantomime of object use is predominantly a symptom of left inferior frontal lesions. These lesion locations are hardly compatible with any version of the 'praxis system.' Reproduced from Goldenberg G and Karnath HO (2006) The neural basis of imitation is body-part specific. *Journal of Neuroscience* 26: 6282–6287. Copyright 2006 by the Society for Neuroscience.

of imitation have different neural underpinning. In patients with left parietal lesions, impressive dissociations have been demonstrated between severe impairment of imitating meaningless gestures and perfect imitation of meaningful gestures. Due to the possibility of compensating defective imitation of meaningful gestures by treating them like meaningless ones, the reverse dissociation is less easily demonstrated, but lesion analysis in a group study of imitation has provided evidence that imitation of meaningful gestures depends primarily on integrity of left medial temporal rather than parietal regions.

Meaningless gestures are not a uniform class either. One important distinction is between single gestures and sequences of several gestures. Imitation of single gestures and sequences is strongly correlated in patients with LBD, whereas defective imitation of gesture sequences contrasting with normal imitation of single gestures has been demonstrated in patients with right hemisphere lesions, Parkinson's disease, or supranuclear palsy. It thus appears that LBD affects imitation of single gestures and sequences equally, but that sequences are also vulnerable to lesions elsewhere.

Another feature of the imitated gesture which may be important for the laterality of interfering lesions is the body part involved. In the 1980s, a lively debate arose over whether imitation of whole body and axial movements such as bending the head down or raising the shoulder is generally spared in patients with LBD who commit errors on imitation of limb gestures. The debate has been settled by the conclusion that imitation of axial movements is indeed less error

prone than imitation of limb movements but that their preservation is due to the lower complexity of the axial gestures. There are far fewer degrees of freedom for axial postures than there are for postures of the limbs.

A dissociation which can less easily be referred to different task difficulty has been demonstrated between gestures of the fingers, the whole hand, and the foot. LBD patients have difficulties with all three kinds. By contrast, patients with right brain damage (RBD) have severe difficulties with finger postures and some difficulties with foot postures but imitate hand postures nearly as perfectly as healthy controls. Within the left hemisphere, different locations are important for imitation of hand and finger postures: whereas imitation of hand postures depends on integrity of inferior parietal lobe and parietotemporal junction, imitation of finger postures is predominantly compromised by lesions affecting the inferior frontal lobe and subjacent white matter (Figure 1).

Production of Communicative Gestures on Command

Gestures are a pervasive component of human communication. Their dependence on simultaneous verbal communication varies from gestures which emphasize or modulate the meaning of simultaneous oral speech to sign languages which can completely replace oral language. Diagnosis and research on apraxia traditionally concentrate on a rather small sector of this wide array, namely, gestures which can transmit meaning independently of speech but which are not normally combined into syntactic structures replacing speech, as is the case for sign languages. Such gestures can be 'emblems' which convey a conventionally defined message, like thumb up for approval or the nose thumb for mockery, or pantomimes of object use in which an object and its use are indicated by performing the motor action of use with the empty hand. The majority of research concerns pantomime rather than emblematic gestures. One of the reasons for this preference may be simple pragmatism: comprehension of the verbal instruction to demonstrate a meaningful gesture out of its natural communicative context is notoriously problematic in patients with aphasia. For pantomime of tool use, comprehension can be facilitated by first demonstrating correct pantomimes and then showing pictures of the single objects whose use should be demonstrated. By contrast, there is hardly any way to bypass defects of verbal comprehension when instructing patients which emblematic gesture they should demonstrate. Comprehension of the instruction to pantomime remains nonetheless dubious in severely aphasic patients, who in response to the command to pantomime the use of an object try to name or describe it, to grasp it for actual use, or to outline with the finger a more or less recognizable shape on the table. Independence of apraxic errors from language comprehension becomes plausible when patients make searching movements for the correct grip or movement or when their pantomime displays some but not all distinctive features of the intended pantomime (e.g., pantomiming drinking from a glass with a narrow grip not accommodated to the width of the pretended glass). In severe cases, patients may produce stereotyped circling or swaying movements of the hand, which might be taken to indicate but not specify movement of the object in peripersonal space.

Defective production of communicative gestures to command is invariably bound to left hemisphere lesions and virtually always associated with aphasia. This tight association could lead to the belief that defective production of communicative gestures is a sequela of language impairment or that one common basic disorder underlies both aphasia and defective gesturing. The latter idea was formulated before Liepmann by another German psychiatrist, Finkelnburg, who conceived of such a basic disorder as 'asymbolia,' that is, a general inability to comprehend and produce communicative signs. Such proposals are, however, contradicted by observation of single cases of left-handed patients in whom RBD impaired production of meaningful gestures to command but did not cause aphasia. In these patients, the left hemisphere is dominant for language and the right for communicative gestures. The possibility of such dissociation argues against any hypothesis claiming identity of the neural mechanisms subserving production of speech and communicative gestures.

A recent study subtracted lesions of LBD patients with intact pantomime from those of LBD patients with defective pantomime. Contrary to traditional beliefs, the proportion of parietal lobe lesions was the same in both groups, whereas inferior frontal lobe lesions were distinctly more frequent in patients with defective pantomime. The independence of communicative gestures from integrity of the parietal lobes is supported by single case reports of 'visuoimitative' apraxia. In these patients, defective imitation of meaningless gestures stands in contrast to preserved production (and in some of them, also imitation) of communicative gestures. These patients invariably had left parietal lobe damage. Their preserved production of pantomimes endorses the conclusion that integrity of the left parietal lobe is not obligatory for correct pantomime of tool use.

Use of Single Tools

Defective use of single tools or objects is less frequent than either defective imitation or defective production

of communicative gestures. Affected patients may try to cut paper with closed scissors, write with the wrong end of the pencil, press the knife into the loaf without moving it to and fro, press the hammer on the nail without hitting, or close the paper punch on top of the sheet without inserting the sheet. The responsible lesions are always left-sided and patients are aphasic. Lesions are usually large. They involve the parietal lobes but are not confined to them.

More-detailed insights into the location of responsible lesions come from studies which distinguish between different components of tool use. Conventional tools such as hammers, screwdrivers, knives, or scissors have in common that they are familiar to most persons and that their function is based on transparent mechanical relationships between tools and object. It has been proposed that knowledge about their correct use can be based on either of these properties. As the tools are familiar, knowledge about their prototypical use is stored in semantic memory. Such knowledge presumably specifies the typical purpose of the tool, the object it is associated with, and the motor action of its use. As their function is based on transparent mechanical relationships, possible functions can also be deduced directly from structural properties. Other than retrieval of knowledge from semantic memory, direct inference of function from structure permits detection of nonprototypical uses of familiar tools and discovery of possible uses of novel tools. For example, when there is no hammer for driving in a nail, one may use pliers because they have a flat and rigid surface which can transmit the power of the beat onto the nail.

The observation that only patients with extensive LBD fail use of single conventional tools has led to the conclusion that both sources of knowledge about their correct use are based on left hemisphere function. This conclusion is corroborated by experimental studies which tested integrity of each of the alternative sources. Knowledge about prototypical tool use is a prerequisite for pantomime of tool use (see above). The observation that pantomime is intact in patients with RBD thus indicates that the right hemisphere does not contribute to its retrieval. Likewise, only left brain-damaged patients commit errors when shown the picture of a tool and asked to select in multiple choice the object that is typically associated with the tool or another tool which serves the same purpose. Evidence that the capacity to infer possible function from structure is also bound to left hemisphere integrity comes from studies which required patients to detect the function of novel tools or to select an alternative tool when the tool typically used for a task is absent, such as selecting a coin for turning a screw when there is no screwdriver. Again, patients with left, but not with right, hemisphere damage have difficulties.

Dissociations between both sources of correct use of single familiar objects have been documented between patients with two different degenerative diseases, corticobasal degeneration, and semantic dementia. Whereas the bulk of cortical pathology affects the parietal and superior frontal lobes in corticobasal degeneration, it is concentrated on the temporal and inferior frontal lobes in semantic dementia. Patients with corticobasal degeneration fail completely when asked to find out the use of novel tools but can associate familiar tools with their corresponding object or another tool serving the same purpose. By contrast, patients with semantic dementia who have lost any knowledge about the purpose and prototypical use of common tools can be astonishingly good in inferring possible functions from examination of the structure of the tool. For example, they may find out know that a nail-clipper can be used for cutting but do not know that it is used for cutting nails.

In conclusion, left parietal lobe integrity seems to be necessary for direct inference of function from structure, which is one source of correct tool use. Retrieval of knowledge about prototypical use from semantic memory does not depend on the parietal lobes but may be contingent on integrity of temporal and possibly also inferior frontal regions. Defective use of single tools becomes manifest only when left hemisphere lesions are large enough to destroy both sources.

Multistep Actions with Multiple Tools and Objects

In daily life one is rarely handed a tool and asked to perform its prototypical action on an adequately prepared recipient. Usually the use of the single tool is embedded in a chain of actions involving several tools and objects, frequently including technical devices, and aiming at a superordinate goal transgressing and modulating the purposes of each single action step. Furthermore, daily life may require the parallel completion of two or more multistep actions, such as when preparing the components of a meal.

Such multistep actions tax mental capacities beyond retrieval of knowledge about the use of single tools. They require maintenance of the ultimate goal of the whole action as well as of a record of already completed and outstanding action steps. The completion of steps must be monitored to secure that the next step is not initiated before the previous one is completed. The order in which outstanding steps are initiated may necessitate preplanning for consideration of sequential constraints.

In contrast to the exclusive association of defective use of single familiar tools with LBD, multistep actions are affected also by lesions of the right hemisphere, by diffuse brain damage, and by frontal lobe damage. This lack of localizing specificity seems to fit well with the involvement of rather difficult-to-localize functions such as attention, executive function, or working memory capacity. Alternatively it might indicate that the cognitive complexity of multistep actions renders them vulnerable to multiple cognitive deficiencies and that failure on multistep actions has different causes in different patient groups.

The Praxis System

The left side of Figure 1 shows Liepmann's original version of a left hemisphere praxis system together with two influential modern accounts, by Geschwind and Heilman. They concur in postulating a posterior-to-anterior processing stream leading to motor execution of skillful motor actions. In Liepmann's version this stream had its origin in a visual image of the action generated in occipital regions. Geschwind emphasized the production of gestures to verbal command and put the start of the processing stream into Wernicke's area in the posterior temporal lobe as this area plays a pivotal role in language comprehension. Heilman's model embraced both possibilities but shifted the occipital source nearer to primary visual cortex to account for imitation of visually presented gestures. All these versions postulate a crucial role of parietal lesions. Liepmann's and Geschwind's models are 'hodological.' They ascribe the importance of parietal lesions to the interruption of fiber tracts passing through the parietal lobe on their way from the origin to the target of the praxis system. By contrast, Heilman contended that parietal cortex stores 'temporo-spatial engrams' needed for skillful execution of motor actions.

Of course, all authors were aware that apraxia is not a unitary syndrome but can affect different domains of actions to different degrees. They postulated, however, that such dissociations can be accounted for by different distributions of lesions within the praxis system. The most influential version of such a division had already been proposed by Liepmann. He distinguished between ideational apraxia, afflicting the concept of the action, and ideomotor (in Liepmann's original terminology, 'ideo-kinetic') apraxia, resulting from an interruption of the stream from posterior to motor regions of the brain. He thought that a hallmark of ideomotor apraxia is its restriction to movements made without external objects. "The guidance of the left hand by mental images of the shape of the movement is disturbed. In the majority the disturbance becomes manifest only if patients are required to imitate or to make expressive movements or manipulations without object." Later authors have suggested amendments to this classification and have introduced additional terms, but Liepmann's original division is still the most popular. It classifies defective imitation, together with defective production, of communicative gestures as ideomotor apraxia, and defective use of tools and objects as ideational apraxia. The empirical findings reviewed above do not support such classification: pantomime of tool use has another neural substrate than imitation, and the neural substrate of imitation varies according to which kind of gesture is imitated. The neural substrate of tool use varies according to whether single tools or multistep actions involving several tools and objects are probed.

The central role of left parietal lobe lesions in apraxia does not fare better when the empirical findings are considered. Although some manifestations of apraxia are indeed bound to left parietal lesions, others are predominantly caused by inferior frontal regions. Even the association of apraxia with LBD is limited: for imitation of finger positions and for multistep actions, RBD is as deleterious as LBD. Obviously, there is no such entity as a left hemisphere praxis system, damage to which explains all manifestations of apraxia. It rather seems as if the disorders subsumed under the heading of apraxia are manifestations of different, more or less related, cognitive deficits. Such 'deconstruction' of apraxia does not diminish interest in its exploration. On the contrary, liberating imitation of gestures, production of communicative gestures, and use of tools and objects from their forced unification into a unique praxis system has advantages both for clinical practice and scientific research. For clinical practice, equivocal classifications will be replaced by usable descriptions of what patients can and cannot do. For research, abandoning the traditional classification into ideomotor and ideational apraxia clears the stage for discussing the differences and commonalities between the cognitive and neuronal basis of the affected domains of action.

See also: Apraxia: Sensory System; Basal Ganglia: Motor Functions; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Finger Movements: Control; Map Plasticity and Recovery from Stroke; Posterior Parietal Cortex and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape.

Further Reading

De Renzi E (1990) Apraxia. In: Boller F and Grafman J (eds.) *Handbook of Clinical Neuropsychology,* vol. 2, pp. 245–263. Amsterdam: Elsevier.

Geschwind N (1975) The apraxias: Neural mechanisms of disorders of learned movements. *American Scientist* 63: 188–195.

Goldenberg G (2003) Apraxia and beyond: Life and works of Hugo Karl Liepmann. *Cortex* 39: 509–525.

Goldenberg G (2008) Apraxia. In: Goldenberg G and Miller B (eds.) *Handbook of Clinical Neurology, 3rd Series, Vol 88: Neuropsychology and Behavioral Neurology,* pp. 323–338. Edinburgh: Elsevier.

Goldenberg G (in press) How the mind moves the body: Lessons from apraxia. In: Morsella E, Bargh JA, and Gollwitzer PM (eds.) *The Psychology of Action*, vol. 2. New York: Oxford University Press.

Goldenberg G, Hermsdörfer J, Glindemann R, Rorden C, and Karnath HO (2007) Pantomime of tool use depends on integrity of left inferior frontal cortex. *Cerebral Cortex* 17: 2769–2776.

Goldenberg G and Karnath HO (2006) The neural basis of imitation is body-part specific. *Journal of Neuroscience* 26: 6282–6287.

Hartmann K, Goldenberg G, Daumüller M, and Hermsdörfer J (2005) It takes the whole brain to make a cup of coffee: The neuropsychology of naturalistic actions involving technical devices. *Neuropsychologia* 43: 625–637.

Heilman KM and Rothi LJG (1993) Apraxia. In: Heilman KM and Valenstein E (eds.) *Clinical Neuropsychology*, pp. 141–164. New York: Oxford University Press.

Rothi LJG, Ochipa C, and Heilman KM (1997) A cognitive neuropsychological model of limb praxis and apraxia. In: Rothi LJG and Heilman KM (eds.) *Apraxia: The Neuropsychology of Action*, pp. 29–50. Hove, UK: Psychology Press.

Schwartz MF, Buxbaum LJ, Montgomery MW, et al. (1999) Naturalistic action production following right hemisphere stroke. *Neuropsychologia* 37: 51–66.

Spatt J, Bak T, Bozeat S, Patterson K, and Hodges JR (2002) Apraxia, mechanical problem solving and semantic knowledge: Contributions to object usage in corticobasal degeneration. *Journal of Neurology* 249: 601–608.

Tessari A, Canessa N, Ukmar M, and Rumiati RI (2007) Neuropsychological evidence for a strategic control of multiple routes in imitation. *Brain* 130(4): 1111–1126.

Apraxia: Sensory System

L J Buxbaum, Moss Rehabilitation Research Institute and Thomas Jefferson University, Philadelphia, PA, USA
H B Coslett, University of Pennsylvania School of Medicine and Center for Cognitive Neuroscience, Philadelphia, PA, USA

© 2009 Published by Elsevier Ltd.

Historical Background

Limb praxis is subserved by a complex multicomponent system that provides a processing advantage to previously experienced, purposive movements. The term 'apraxia' was introduced by Steinthal in 1871. While this word, derived from Greek, means 'without action', the term is used to describe a decrease or disorder in the ability to perform skilled movements. Scientific and clinical interest in the disorder dates to the early twentieth century, when Liepmann reported patients with cerebral lesions who were unable to gesture to command or, in some instances, to imitation. Subsequently, Liepmann and Maas (1907) described a patient with a lesion of the corpus callosum who was unable to produce gestures with the left hand to verbal command. On the basis of these findings, Liepmann proposed that the left hemisphere was "dominant" for gesture in that it supported the learned "movement formulae" or "time-space-form picture of the movement" which specified the timing, trajectory, and content of learned movements.

Liepmann's ideas were extended by Geschwind (1965), who proposed a specific left hemisphere-based neural circuitry for movement representations. On his account, failure to produce a movement to command was attributable either to a disruption of Wernicke's area, with resultant failure to understand the command, or to a disconnection of the posterior language areas from motor cortex. A failure to imitate movements was attributed to a lesion involving the arcuate fasciculus, which was assumed to connect the visual association cortex to motor cortices.

Limb Apraxia Subtypes

Hugo Liepmann's description of three major forms of apraxia brought about a 'paradigmatic shift' in our understanding of motor control. These three types were 'limb kinetic apraxia' (LKA), 'ideational apraxia' (IA), and 'ideomotor apraxia' (IMA). To this triad, Gonzalez-Rothi, Heilman, and colleagues added another type, termed 'conceptual apraxia.' These types of apraxia are described briefly below.

LKA

Patients with LKA perform actions with slow, stiff, clumsy movements and exhibit impairment on tasks requiring rapid independent finger movements, such as rotating a coin between the thumb, index, and middle finger. Errors are more apparent in distal (finger) movements than in proximal movements. LKA is associated with lesions that include the primary motor cortex, premotor cortex, or descending corticospinal tract. It frequently occurs in patients with stroke and in degenerative disorders such as progressive supranuclear palsy and corticobasal degeneration.

IA

IA is defined as an impaired ability to carry out a sequence of acts that leads to a goal and that incorporates multiple objects, such as making a sandwich or lighting a candle. For example, a patient with IA might attempt to seal an envelope prior to inserting the letter. Another type of error exhibited in this disorder is illustrated in Figure 1.

IA is most frequently induced by bilateral damage and degenerative dementia. Injury to the frontal lobes is often also associated with temporal order processing deficits as well as impaired working memory, and thus one of the critical foci of dysfunction in IA may be in frontal-subcortical systems. The strongest predictor of errors in multistep, naturalistic action is overall severity of cognitive impairment and not lesion location.

Conceptual Apraxia

Patients with conceptual apraxia make content errors in complex action – that is, they substitute incorrect objects or movements in their actions. For example, patients with conceptual apraxia may eat with a toothbrush. In some cases, underlying deficits in knowledge of specific tools or objects or the association of tools and objects has been demonstrated; these patients may misuse objects because they have lost knowledge regarding the function of the object. Conceptual apraxia also frequently co-occurs with IMA, and it has not been established whether these disorders can be reliably disambiguated.

Deficits in conceptual action knowledge have been associated with the dominant posterior parietal lobe and/or temporal parietal junction. On the other hand, errors apparently attributable to conceptual deficits frequently occur in patients whose lesions entirely spare brain regions typically associated with conceptual action knowledge (e.g., right parietal cortex).

IMA

IMA is a common disorder of complex skilled action not attributable to weakness, incoordination, or other elemental sensory or motor impairments. It is typically observed in individuals who have suffered left hemisphere strokes; IMA is observed in the actions of the 'unimpaired' left hand of approximately 50% of patients with left hemisphere stroke and commonly persists for at least 1 year after stroke. IMA is also common in Alzheimer's disease and in corticobasal degeneration. In stroke, it is usually a consequence of damage to the left inferior parietal lobe (and, on occasion, adjacent intraparietal sulcus and superior temporal gyrus) but has also been observed following left dorsolateral prefrontal, callosal, and subcortical damage (see Figure 2).

IMA is usually diagnosed on the basis of spatiotemporal errors in the production of gesture pantomime both to sight of objects and on imitation of others. That is, IMA is typically seen when a patient is asked to show how an object (e.g., scissors) would be used or when the patient is asked to copy a gesture produced by the examiner. Kinematic analyses have revealed that IMA patients pantomime skilled

Figure 1 Photographs of a patient with conceptual apraxia making a sandwich with meat and mustard. She correctly places meat on a slice of bread, closes the sandwich, and opens a mustard jar. She replaces the mustard jar, reaches into a package of marking pens, retrieves a yellow marker, and proceeds to color the meat yellow.

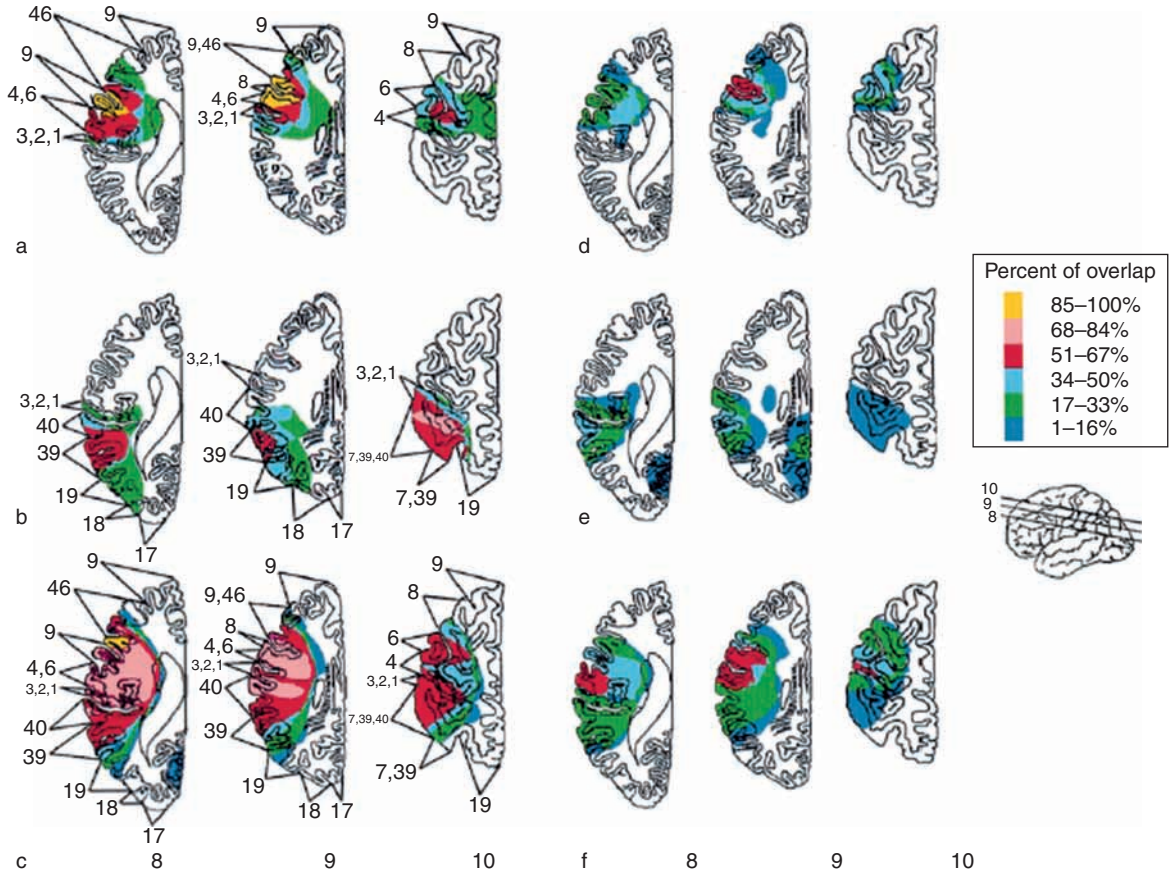

Figure 2 Maximal lesion overlap from 17 apraxic patients is shown in dorsolateral and inferior parietal regions. From Haaland KY, Harrington DL, and Knight RT (2000) Neural representations of skilled movement. *Brain* 123: 2306–2313.

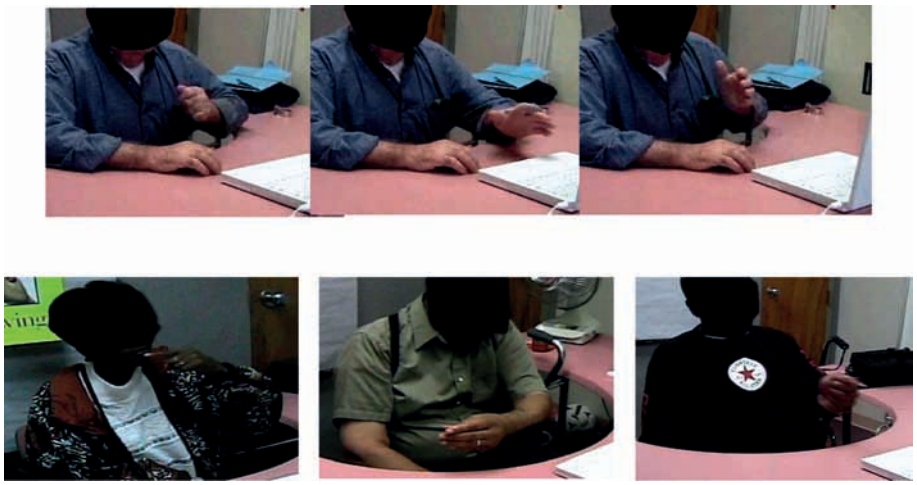

Figure 3 Typical errors in ideomotor apraxia. Top: Three still photographs from a videotape showing a sequence of postures produced by an apraxic patient in imitating a sawing movement. Note the typical hand posture error comprised of repeated hand opening with arm extension, despite the fact that the model maintained a closed grip throughout. Bottom left: Typical 'body-part-as-object' error in a toothbrushing pantomime. Bottom center: Typical arm posture error in imitating a scissoring movement. The model's movement was produced perpendicular to the body wall (from near to far) whereas the patient's movement proceeded left to right. Bottom right: Typical amplitude error in imitation of a hammering gesture. The model to be imitated demonstrated a large swing with peak amplitude at shoulder height, whereas the patient's maximal amplitude was at elbow height.

tool-use movements with abnormal joint angles and limb trajectories and uncoupling of the spatial and temporal aspects of movement (see Figure 3). Spatiotemporal errors persist to a lesser degree with actual tool use. The deficit is not restricted to meaningful movements and has also been observed in meaningless postures and sequences. IMA is also associated with cognitive deficits in declarative knowledge of the manipulation actions appropriate to objects, impairments in mechanical problem solving, deficits in motor planning, and difficulty learning new gestures.

The disorder may be attributed either to damage to stored spatiotemporal gesture representations in the left parietal lobe, sometimes called 'visuokinesthetic engrams,' or to disconnection of intact movement representations from motor output. The integrity of gesture representations is thought by many investigators to bear on the integrity of gesture recognition. In the case of damage to the representations, patients have impaired knowledge of the appropriate motor action to perform, as evidenced by deficits in gesture recognition (representational IMA). In the case of disconnection of intact engrams, patients have unimpaired knowledge of appropriate gestures, as evidenced by intact gesture recognition and ability to discriminate correct from incorrect gestures, but nevertheless perform with spatiotemporal errors (dynamic IMA).

In representational IMA, inability to discriminate correctly from incorrectly performed meaningful object-related hand movements correlates strongly with ability to produce the same movements, suggesting that the same representations may underlie both. In addition, representational IMA patients are significantly more impaired when producing object-related than symbolic, nonobject-related movements. This in turn suggests that the damaged system underlying representational IMA is specialized for movements related to skilled object use.

Disconnection and Dissociation Apraxias

Several apraxia patterns indicate that aspects of input to and output from the skilled action system are dissociable. Verbal–motor dissociation apraxia refers to a pattern of impairment in which patients are unable to gesture in response to command despite adequate comprehension and unimpaired ability to gesture to imitation. Heilman and colleagues posited that the lesion responsible for this apraxia subtype was in the angular gyrus, but they were unable to obtain neuroimaging data. Another reported pattern is seen in the tactile–motor and visuomotor dissociation apraxias, in which patients fail to gesture appropriately when holding tools or viewing tools, respectively, despite unimpaired object recognition and better gesture performance in the unaffected modality. In response to these and other patterns of dissociation, Gonzalez-Rothi, Heilman, and colleagues proposed a detailed diagrammatic model of IMA. Theoretical 'lesions' at various loci in the model appear to explain many of the observed dissociations.

Outstanding Issues in Diagnosis of Apraxia Subtype

Relevance of Recognition and Imitation Deficits for Diagnosis of IMA versus IA or Conceptual Apraxia

Historically, gesture recognition and imitation have both been used to distinguish between IMA and IA/conceptual apraxia. In Liepmann's account, patients with IA fail to reliably activate gesture engrams. Consequently, they perform normally when provided with the 'idea' of the movement; that is, when they are asked to imitate the movement of another person. Liepmann believed that, by contrast, patients with IMA suffered a disconnection of an intact idea (time–space–form picture of the movement) from motor innervatory patterns. Thus, on Liepmann's account, providing IMA patients with the 'idea' in the form of a gesture to imitate would not be of benefit.

On many contemporary accounts, it is representational IMA patients who fail to reliably activate gesture representations and who therefore may be able to imitate gestures. Recent evidence indicates that imitation may be accomplished either via gesture engrams (the so-called indirect or semantic route), or by way of a 'direct route' to action that enables imitation without access to meaning. (The direct route may bear a relationship to the putative 'mirror neuron' system discussed below.) Therefore, the ability to imitate may depend on the integrity of each of these routes. On the other hand, there is evidence that the direct route is not used for meaningful gestures even when it is intact, suggesting that there may be obligatory activation of the semantic route whenever familiar gestures are viewed.

Disagreement persists on whether gesture recognition problems signify IMA or IA. However, recent evidence from monkeys and humans indicates that the same representations are likely used for action recognition and production. In the macaque, cells in the inferior parietal lobule and in a sector of premotor cortex corresponding to Brodmann's areas 44 and 45 in humans respond both when the monkeys produce actions and when they observe the same specific actions performed by others ('mirror neurons'). In humans, there are strong correlations between action production and recognition for the same items. This suggests that gesture recognition problems may reflect degraded or inaccessible sensorimotor representations, a characteristic of IMA.

Relationship of Object Knowledge to Gesture Representations

Continued work is required to clarify the relationship of knowledge of appropriate object-oriented actions to the gesture engram system. In an influential paper on IA, De Renzi and Lucchelli proposed that the problem underlying deficient object use was a loss of knowledge of the manner in which objects are to be used, which they characterized as a semantic deficit. This emphasis on 'manner' of manipulation raises questions about the role gesture engrams might play in object-use knowledge. In contrast, other investigators view deficient recognition of the gestures associated with objects (that is, the manner of use) to be a symptom of the representational type of IM, and not IA.

Recent evidence suggests that different types of object knowledge may bear different relationships to apraxia. For example, knowledge of object function and knowledge of manner of object manipulation are dissociable. One might have knowledge of the function of an object (e.g., a knife is for cutting things) without knowledge that knives are often used with a back-and-forth, sawlike gesture. Patients with IMA tend to have the latter type, but not necessarily the former. Some reports indicate that patients with IMA may also be impaired in mechanical problem solving or the ability to infer function from structure.

Evidence for a relationship between function knowledge and performance on tasks involving multiple objects is equivocal. One potential source of confusion is that the relationship between functional knowledge and object use is sometimes assessed with single objects and sometimes with tasks involving multiple objects. In several studies using single object tests in patients with semantic dementia, a disorder with a predilection for the temporal lobes, a strong relationship has been reported. In other investigations, no relationship between single object use and functional knowledge has been found. There is stronger evidence that function knowledge is not well correlated with performance on tasks involving multiple objects. There are patients who make 'conceptual' errors on these tasks but who perform normally on semantic tests of functional and associative object knowledge, and others who perform nearly normally on tasks involving multiple objects or in real-life action, despite considerable semantic deficits.

Functional Implications of Limb Apraxia

Historically, most clinicians and researchers regarded IMA as a clinical oddity that had little significance in the real world. It appears that this view was derived from the notion that IMA was present when gestures to command and imitation were tested but improved when actions with actual objects were examined. A number of recent studies, however, have suggested that IMA is associated with deficits in activities of

daily living. At least in some studies, participants with IMA are more likely to be impaired in object use, particularly in complex tasks, than nonapraxic participants who have suffered a stroke.

IMA in View of Recent Developments in the Motor Control Literature

Imitation

With the discovery of mirror neurons in the macaque premotor and parietal cortex that respond to both observed and performed actions, action imitation has emerged as an area of considerable interest in the neuroscience community. One important question concerns the degree to which imitation failure in IMA reflects damage to the mirror neuron system. Indeed, the neuroanatomic loci of lesions leading to IMA overlap considerably with the localization of mirror neurons (see Figure 4). In addition, as noted, imitation and recognition impairments show a strong correspondence in IMA. On the other hand, IMA due to left parietal lesions frequently disrupts object-related (transitive) imitation far more than nonobject related, symbolic (intransitive) imitation. In addition, there is evidence of body part specificity in IMA imitation disruption that is not easily accommodated by putative damage to a mirror neuron system. Left hemisphere IMA patients tend to be significantly more impaired in imitation of hand postures than of finger positions, and in general, IMA appears to affect arm more than leg imitation. These dissociations could be accommodated by positing that effector-specific populations of mirror neurons might reside in different cortical regions in each hemisphere, but to this point there is little evidence for this possibility. Future investigations addressing these issues are required.

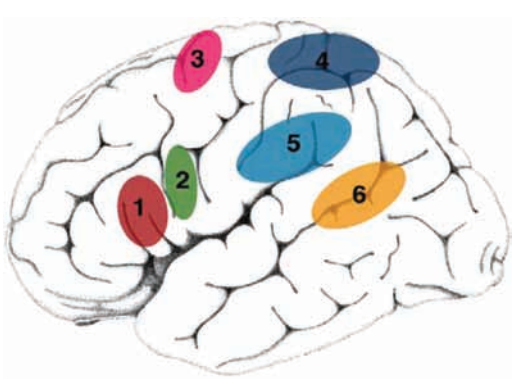

Figure 4 Mirror neuron cortical regions. From Brass M and Heyes C (2005) Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences* 9(10): 489–495.

Object-Related Action

Recent evidence from single cell recordings in monkeys indicates that populations of neurons in the inferior premotor cortex (in an area with probable homolog of areas 44 and 45 in humans) as well as in the anterior intraparietal sulcus are active in response to objects that are graspable by the monkey observer. These have been termed 'canonical' neurons. Complementary studies using functional magnetic resonance imaging and transcranial magnetic stimulation in humans are consistent in suggesting that similar regions in the human brain are responsive to the structural properties (i.e., shape and size) of graspable objects. These populations appear to encode hand movement parameters (e.g., finger thumb aperture) for object grasping. In this context, there is considerable recent evidence that IMA patients, while intact in their ability to position the hand in response to object structure, are disproportionately impaired in hand shaping for functional object manipulation. The relationship of this pattern of performance to the 'canonical' neuron system is an additional area of interest for future investigation.

Spatiomotor Frames of Reference for Action

At least two different frames or reference or coordinate systems have been proposed for action. Many investigators have proposed that action may be planned in workspace-specified extrinsic coordinates. On this account, movements are planned with respect to a target that is coded in external space. Reaching to grasp a target would entail the creation of a spatial vector describing a desired movement's direction and amplitude. An alternative hypothesis proposes that movement control may occur in body-specified intrinsic coordinates; on this account, a movement plan would specify the positions of the shoulder, elbow, and wrist that would be needed to get the hand to the target. Extensive evidence for both types of control has led to a third group of accounts proposing that control is an interactive process that uses both extrinsic and intrinsic coordinate frames, depending in part on the demands of the task.

Recent evidence from IMA patients indicates that movements that may putatively rely strongly on extrinsic control (i.e., object-directed movements) are accurate, whereas movements not having external referents (i.e., body-directed movements) are characterized by spatial errors in hand configuration, wrist angle, hand orientation, and hand location. The possibility that IMA may in part reflect deficient coding of action in a body-centered framework is an area of active investigation in several laboratories.

Feedforward and Feedback-Driven Processes in IMA

The process of motor control is commonly subdivided into planning and online correction components. Planning is the preparation of a movement before movement initiation, whereas online correction refers to the adjustment of the movement plan during movement execution. There is evidence that IMA may be attributable in part to deficits in planning actions with relatively intact online correction. IMA patients are impaired in motor imagery, thought by several investigators to be a proxy for motor planning stages of action. They are also abnormally disrupted when visual feedback of movement is unavailable. This suggests that such patients may rely abnormally on visual feedback in the performance of skilled action.

Treatment of Limb Apraxia

The current literature on apraxia treatment is sparse. Approximately ten treatment efforts have been reported; in many cases, there is but a single study devoted to each treatment approach. The studies uniformly fall into the category of Phase I studies in which feasibility is assessed in small numbers of patients. Thus, it is difficult at this stage to draw conclusions about treatments that may hold particular promise.

In general, the few reported treatment approaches can be grouped into three categories: (1) studies that attempt to directly ameliorate deficient object-related gesture production with a variety of visual and tactile cues and feedback, (2) studies focusing on providing corrective feedback for errors in naturalistic multistep action, and (3) studies that attempt to prevent error from occurring (errorless learning approach). All the studies report at least some treatment benefit, but several difficulties obscure the interpretation of results. For example, apraxia type is frequently poorly characterized. Although gesture recognition is clearly an important index of the integrity of gesture representations (which in turn may have important implications for rehabilitation strategies), recognition testing is usually not performed. Only a few studies report generalization to untreated stimuli (or behaviors), maintenance of treatment effects, or impact on daily activities.

There is preliminary evidence based on these few studies that limb apraxia is amenable to treatment. The purpose of Phase I research, however, is to develop hypotheses, protocols, and methods; establish safety and activity; determine the best-outcome measures; identify responders versus nonresponders; determine optimal intensity and duration; and determine why the treatment is producing an effect. This suggests that further systematic inquiry is required to satisfy the objectives of Phase I research.

Testing for Limb Apraxia

Apraxia cannot be assessed in patients whose comprehension or cognitive deficits prevent them from understanding the task or whose visual deficits preclude identification of an object or gesture; before one tests for apraxia, these disorders must, therefore, be excluded. In order to identify apraxic patients and distinguish between the different types of apraxia described above, a testing battery should include at least the following components:

1. Assessment of manual dexterity (e.g., rotation of coin between fingertips).
2. Testing of gesture to command and to sight of object.
3. Imitation of meaningful and meaningless gestures.
4. Assessment of intrinsic egocentric spatial coding – that is, the ability to imitate meaningless static positions of the body such as holding the dorsum of the left hand against the right cheek.
5. Assessment of extrinsic egocentric spatial coding by reaching to touch or grasp objects.
6. Tests of functional semantic knowledge (e.g., which two of three pictured objects – paper clip, rubber band, and door lock – are used for the same purpose).
7. Tests of manipulation knowledge (e.g., which two of three pictured objects – saw, clothes iron, and watering can – is used with the same or similar gesture).
8. Perform a familiar multistep task such as preparing a cup of instant coffee.
9. Recognize gestures by naming a gesture or selecting which of two gestures is correctly performed.

Conclusions

Apraxia is a complex and heterogeneous disorder that has important clinical and scientific implications. Recent investigations of the disorder that are motivated by emerging accounts of motor control and planning are beginning to explicate the processing impairments underlying the apraxic disorders. The accumulating knowledge offers promise not only for the development of treatments for apraxia but also for greater understanding of the procedures by which actions are generated and the underlying neural basis.

See also: Apraxia: Disease; Basal Ganglia: Motor Functions; Computational Approaches to Motor Control; Finger Movements: Control; Kinematics and Dynamics; Motor Skill Learning; Motor Sequences; Motor Timing; Motor Psychophysics; Neural Coding of Spatial Representations; Sensorimotor Integration: Attention and the Premotor Theory; Sensorimotor Integration: Models; Sensorimotor Control of Manipulation.

Further Reading

Buxbaum LJ, Johnson-Frey SH, and Bartlett-Williams M (2005) Deficient internal models for planning hand–object interactions in apraxia. *Neuropsychologia* 43: 917–929.

Buxbaum LJ and Saffran EM (2002) Knowledge of object manipulation and object function: Dissociations in apraxic and non-apraxic subjects. *Brain and Language* 82: 179–199.

Derenzi E and Lucchelli F (1988) Ideational apraxia. *Brain* 111: 1173–1185.

Goldenberg G (2003) Apraxia and beyond: Life and work of Hugo Liepmann. *Cortex* 39: 509–524.

Goldenberg G and Hagmann S (1997) The meaning of meaningless gestures: A study of visuo-imitative apraxia. *Neuropsychologia* 35: 333–341.

Gonzalez Rothi LJ, Ochipa C, and Heilman KM (1997) *A Cognitive Neuropsychological Model of Limb Praxis and Apraxia*. Hove: Psychology Press.

Haaland KY, Harrington DL, and Knight RT (1999) Spatial deficits in ideomotor limb apraxia: A kinematic analysis of aiming movements. *Brain* 122: 1169–1182.

Haaland KY, Harrington DL, and Knight RT (2000) Neural representations of skilled movement. *Brain* 123: 2306–2313.

Haaland KY, Prestopnik JL, Knight RT, and Lee RR (2004) Hemispheric asymmetries for kinematic and positional aspects of reaching. *Brain* 127: 1145–1158.

Heilman KM, Rothi LJ, and Valenstein E (1982) Two forms of ideomotor apraxia. *Neurology* 32: 342–346.

Jeannerod M (2003) Simulation of action as a unifying concept for motor cognition. In: Johnson-Frey SH (ed.) *Taking Action: Cognitive Neuroscience Perspectives on Intentional Acts*, pp. 139–163. Cambridge, MA: MIT Press.

Johnson-Frey S, Newman-Norland R, and Grafton S (2005) A distributed network in the left cerebral hemisphere for planning everyday tool use skills. *Cerebral Cortex* 15: 681–695.

Klatzky R, Pellegrino J, Mccloskey B, and Lederman S (1993) Cognitive representations of functional interactions with objects. *Memory & Cognition* 21: 294–303.

Leiguarda RC and Marsden DC (2000) Limb apraxias: Higher-order disorders of sensorimotor integration. *Brain* 123(5): 860–879.

Peigneux P, Van Der Linden M, Garraux G, et al. (2004) Imagining a cognitive model of apraxia: The neural substrate of gesture-specific cognitive processes. *Human Brain Mapping* 21: 119–142.

Rizzolatti G and Matelli M (2003) Two different streams form the dorsal visual system: Anatomy and functions. *Experimental Brain Research* 153: 146–157.

Roy EA and Square PA (1985) Common considerations in the study of limb, verbal and oral apraxia. In: Roy EA (ed.) *Neuropsychological Studies of Apraxia and Related Disorders*, pp. 111–164. Amsterdam: North-Holland.

Rumiati RI and Humphreys GW (1998) Recognition by action: Dissociating visual and semantic routes to action in normal observers. *Journal of Experimental Psychology: Human Perception and Performance* 24: 631–647.

Schwartz MF and Buxbaum LJ (1997) Naturalistic action. In: Rothi LJG and Heilman KM (eds.) *Apraxia: The Neuropsychology of Action*. East Sussex: Psychology Press.

Schwoebel J, Buxbaum L, and Coslett HB (2004) Representations of the human body in the production and imitation of complex movements. *Cognitive Neuropsychology* 21: 285–299.

Sirigu A, Duhamel J-R, Cohen L, Pillon B, Dubois B, and Agid Y (1996) The mental representation of hand movements after parietal cortex damage. *Science* 273: 1564–1568.

Artificial Intelligence

J Feldman

© 2009 Elsevier Ltd. All rights reserved.

This article is modified from 'Artificial Intelligence in Cognitive Science' in the *International Encyclopedia of the Social & Behavioral Sciences*, pp. 792–796, © 2001, Elsevier Science Ltd.

Overview

Artificial intelligence (AI) and cognitive science are two distinct disciplines, with overlapping methodologies but with rather different goals. AI is a branch of computer science and is concerned with construction and deployment of intelligent agents as computer programs, and also with understanding the behavior of these artifacts. The core scientific goal of AI is to understand the basic principles of intelligent behavior that apply equally to animal and artificial systems. Almost all of the work is mathematical or computational in character and much of the literature is technique oriented.

Cognitive science is an explicitly interdisciplinary field that has participation not only from AI, but also from linguistics, philosophy, psychology, and subfields of other social and biological sciences. The unifying goal of cognitive science is to understand and model human intelligence, using the full range of findings and methodologies of the complementary disciplines. As one would expect, a wide range of techniques from the mathematical, behavioral, social, and biological sciences are employed. Cognitive science, in contrast with AI, is defined more by phenomena than by methodology. There are research groups that are active in both AI and cognitive science, but they tend to produce different types of reports for journals and conferences in the two areas.

Shared Origins in the Postwar Cognitive Revolution

Both AI and cognitive science evolved after 1950, and in their early development were more tightly integrated than at present. For much of the first half of the twentieth century, the Anglo-American study of cognition was dominated by the behaviorist paradigm, which rejected any investigation of internal mechanisms of mind. The emergence of both AI and cognitive science was part of a general postwar movement beyond behaviorist theories, which also included new approaches in linguistics and the social sciences. The idea of computational models of mind was a central theme of what is sometimes called the 'postwar cognitive revolution.' One of the leading early AI groups, under the leadership of Allen Newell and Herbert Simon at Carnegie Mellon, was explicitly concerned with cognitive modeling using the symbolic processes of AI. The current version of this continuing effort is the symbolic cognitive architecture known as Soar (the term derived originally from the concept 'state, operator, and result,' but those currently involved in Soar development prefer not to represent the term as an acronym). Another traditional symbolic approach to modeling intelligence is known as ACT (an acronym for 'adaptive control of thought'). However, there is very little work in contemporary AI that is explicitly focused on modeling human behavior as opposed to intelligent systems in general. There is some continuing work on human and machine game playing, but it is not integrated into the fields of AI and cognitive science.

The central idea of AI is computational modeling of intelligent behavior – its main contribution to cognitive science. The basic notion of a computational model is now commonplace in all scientific fields and many other aspects of contemporary life. One builds a detailed software model of some phenomenon and studies the behavior of the model, hoping to gain understanding of the original system. Much of the work in AI has the engineering goal of producing practical systems, and there is no sharp boundary between AI and other applied fields of computer science and engineering. AI techniques are now commonplace in the full range of business, scientific, and public applications. While all fields use computational models, researchers, in computer science in general and AI in particular, invent and study computational techniques for constructing models, presenting the results of simulations, and understanding the limitations of the simulation. AI has traditionally studied the modeling of the most complex phenomena – those relating to intelligence. Because of the technical challenges arising in the construction of these complex simulations, many innovations in computing have arisen in AI and then have been more widely applied.

Domain-Focused Research That Cuts Across AI and Cognitive Science

The relationship between AI and cognitive science is further complicated by the fact that there are currently several distinct research fields that cut across both disciplines, but have separate journals, meetings, etc. The most prominent of these research areas are speech, language, vision, and neural networks. Each of these fields has thousands of practitioners, many of

whom are interested in AI, cognitive science, or both. Appropriately, each of these areas is represented by a vast body of literature. As AI and cognitive science have grown, specialized areas such as language and vision modeling have become largely independent, but they do continue to share the development of underlying methodologies. The main areas that have remained as core AI include knowledge representation and reasoning, planning, and problem solving. The study of learning has evolved somewhat differently (see later). There are some common scientific paradigms that cut across all of these fields, and these are discussed in the following sections as they relate to the social and behavioral sciences. The common thread linking AI to cognitive science is reliance on computational models of different kinds.

Role of Formal Logic in AI and Cognitive Science

To a great extent, the early development of AI was based on symbolic, as opposed to numerical, modeling. This led to the introduction of some novel representations such as those of Soar, but the main effect of this was to align AI with formal logic for much of its early history. In fact, much of the driving force for the creation of the field called cognitive science came from people who saw mathematical logic as its unifying theme. This remains a fruitful approach to AI and constitutes one major area of overlap with cognitive science. Mathematical logic is elegant and well developed and can be shown to be, in some sense, general enough to represent anything that can be described formally. There remains a significant community of linguists, philosophers, and computer scientists for whom logic is the only scientific way to study intelligence.

However, the twentieth century was not kind to categorical, deterministic theories in any field, and cognitive science was no exception. A central issue in the study of cognition has been how to describe the meaning of words and concepts. In formal logic, a concept is defined by a set of necessary and sufficient conditions. For example, a bachelor might be defined as a male who never married. The limitations of classical, all-or-none, categories were already recognized by Wittgenstein, who famously showed that concepts such as 'game' could not be characterized by necessary and sufficient conditions, and were better described as family resemblances. Even the definition of 'bachelor' becomes graded when we consider cohabitation, or how old a male needs to be before being considered a bachelor.

Starting in the 1960s, a wide range of cognitive science studies by Rosch and others showed the depth and complexity of human conceptual systems and their relation to language. This helped give rise to the subfield called cognitive linguistics, which overlaps with cognitive science, but has its own paradigms, journals and conferences. The graded and relational nature of human categories undermined the attempt to create a unified science of mind based on formal logic. Another attack on the formalist program arose from the growing understanding of the neural basis of intelligence. A crucial insight was that basic human concepts are grounded in direct experience and that more abstract concepts are mapped metaphorically to more embodied ones. This undercuts the formalist program, and also leads to a separation of AI, which studies intelligence in the abstract, from cognitive science, which is explicitly concerned with human minds. While much excellent work continues to be done using logic, it is now generally recognized that there are a wide range of phenomena that are better handled by biologically based models and/or some form of numerical, often probabilistic, modeling. For a variety of reasons, the movement to quantitative numerical models followed somewhat different paths in AI and cognitive science, but there are some recent signs of reconvergence.

Learning and the Connectionist Approach to Cognitive Science

As it happens, while the formalists were trying to establish a cognitive science based on formal logic, an antithetical neural network movement was also developing, and this approach has become a major force in cognitive science. The two contrasting approaches to cognitive modeling, neural modeling and logic, were mirrored in the two different methods by which early computer scientists sought to achieve AI. From the time of the first electronic computers around 1950, people dreamed of making them 'intelligent' by two quite distinct routes. The first, 'conventional' AI, is to build standard computer programs as models of intelligence. This remains the dominant paradigm in AI and has had considerable success. The other approach is to try to build hardware that is as brainlike as possible and have it learn the required behavior. The history of this 'neural modeling' approach has been well described. After some promising early mathematical results on learning in simple networks, the neural learning approach to modeling intelligence fared much less well for three decades and had little scientific or applied success. Around 1980, a variety of ideas from biology, physics, psychology, and computer science and engineering coalesced to yield a 'new connectionist' approach to modeling intelligence that has become a core field of cognitive science, and also the

basis for a wide range of practical applications. Among the key advances was a mathematical technique (back-propagation) that extended the early results on learning to a much richer set of network structures.

Connectionist computational models are almost always computer programs, but programs of a different kind than those used in, for example, word processing or symbolic AI. Connectionist models are specified as a network of simple computing units, which are abstract models of neurons. Typically, a model unit calculates the weighted sum of its inputs from upstream units and sends to its downstream neighbors an output signal that is a nonlinear function of its inputs. Learning in such systems is modeled by experience-based changes in the weights of the connections between units. The basic connectionist style of modeling is now being used in three quite different ways – in neurobiology, in applications, and in cognitive science. Neurobiologists who study networks of neurons employ a wide range of computational models, from very detailed descriptions of the internal chemistry of the neuron to the abstract units just described. The use of connectionist neural models in practical applications is part of the reconvergence with AI and is discussed in the final section of this article.

In cognitive science, connectionist techniques have been used for modeling all aspects of language, perception, motor control, memory, and reasoning. This universal coverage represents a potential breakthrough; previously, the computational models of, for example, early vision and problem solving used entirely different mathematical and computational techniques. Since the brain is known to use the same neural computation throughout, it is not surprising that neurally inspired models can be applied to all behavior. Unfortunately, the existing models are neither broad nor deep enough to ensure that the current set of mechanisms will suffice to bridge the gap between structure and behavior, but the work remains productive.

Connectionist models in cognitive science fall into two general categories, often called structured and layered networks (also called parallel distributed processor, or PDP, networks). Most modelers are primarily interested in learning, which is modeled as experience-driven change in connection weights. There is a great deal of research studying different models of learning with and without supervision, different rules for changing weights, etc. Because of the focus on what the network can learn, any prewired structure will weaken the results of the experiment. The standard approach is to use networks with unidirectional connections arranged in completely connected layers, sometimes with a very restricted additional set of feedback links. This kind of network contains a minimum of presupposed structure and is also amenable to efficient learning techniques, such as the aforementioned back-propagation method. Most researchers using totally connected layered models do not believe that the brain shares this architecture, but there is an ongoing controversy about the implications of PDP learning models for theories of mind (see the section titled 'Nature and nurture: rules versus connections').

Structured connectionist models are usually less focused on learning than on the representation and processing of information. Essentially all the modeling done by neurobiologists involves specific architectures, which are known from experiment. For structured connectionist models of cognitive phenomena, the underlying brain architecture is rarely known in detail and sometimes not at all at the level of neurons and connections. The methodology employed is to experiment with computational models of the behavior under study that are consistent with the known biological and psychological data and are also plausible in the resources (neurons, computing time, etc.) required. This methodology is very similar to what are called 'spreading activation' models, widely used in psycholinguistics. Some studies combine structured and layered networks, or investigate learning in networks with an initial structure that is tuned to the problem area or the known neural architecture.

Nature and Nurture: Rules versus Connections

Perhaps the most visible contribution to date of connectionist computational models in cognitive science has been to provide a new jousting ground for contesting some age-old issues on the nature of intelligence. Much of the debate has been published in *Science* magazine, which suggests that it is considered to be of major importance by the US scientific establishment. The nature versus nurture question concerns how much of some trait, usually intelligence, can be accounted for by genetic factors, and how much depends on postnatal environment and training. Some PDP connectionists have taken very strong positions, suggesting that learning can account for everything interesting. In the particular case of grammar, an important group of linguists and other cognitive scientists take an equally extreme nativist position, suggesting that humans only need to choose a few parameters to learn grammar. A related issue is whether human grammatical knowledge is represented as general rules or just appears as the rulelike consequences of PDP learning in the neural network of the brain. There is ample evidence against both extreme positions, but the debate continues to motivate a great deal of thought and experiment.

Current and Future Trends

Although the fundamental split between the AI focus on general methods and the cognitive science emphasis on human intelligence remains, there are a growing number of areas of overlapping interest. As already discussed, quantitative neural models are playing a major role in cognitive science. It turns out that the mathematical and computational ideas underlying learning in neural networks have found application in a wide range of practical problems, from speech recognition to financial prediction. The basic idea is that, given current computing power, back-propagation and similar techniques allow large systems of nonlinear units to learn quite complex probabilistic relationships using labeled data. This general methodology overlaps not only with AI but also with mathematical statistics, and is part of a unifying area called computational learning theory. There is also a large community of scientists and engineers who identify themselves as working on neural networks and related statistical techniques for various scientific and applied tasks, along with conferences and journals to support this effort.

While probability was entering cognitive science from the bottom-up through neural models, AI experienced the introduction of probabilistic methods from general theoretical considerations, which only later led to practical application. As discussed earlier, the limitations of formal logic became well recognized in the 1960s. Over the subsequent decades, AI researchers, led by Judea Pearl, at the University of California (Los Angeles), developed methods for specifying and solving large systems of conditional probabilities. These belief networks are now widely used in applications ranging from medical diagnosis to business planning. A growing field that involves both symbolic and statistical techniques is 'data mining,' processing large historical databases to search for relationships of commercial or social importance. Recent efforts to learn or refine belief networks from labeled data are another area of convergence of AI and cognitive science in computational learning theory. Of course, the explosion of Internet activity is affecting AI along with the rest of the computing field. Two AI application areas that seem particularly important to cognitive science are intelligent Web agents and spoken-language interaction.

As the range of users and activities on the Internet continues to expand, there is increasing demand for systems that are both more powerful and easier to use. This is leading to increasing efforts on the human–computer interface, including the modeling of user plans and intentions – clearly overlapping with traditional concerns of cognitive science. One particularly active area is interaction with systems using ordinary language. Whereas machine recognition of individual words is relatively successful, dealing with the full richness of language is one of the most exciting challenges at the interface between AI and cognitive science, and a problem of great commercial and social importance.

Looking ahead, we can be confident that the increasing emphasis on intelligent systems will continue. From the scientific perspective, it is very likely that most of the interdisciplinary research in cognitive science will remain focused on specialized domains such as language, speech, and vision. General issues including representation, inference, and learning will continue to be of interest and will constitute the core of the direct interaction between AI and cognitive science. With the rapid advances in neurobiology, both fields will increasingly articulate with the life sciences, with great mutual benefits.

See also: Animal Intelligence: The Search for Animal Intelligence; Cognitive Neuroscience: An Overview; Cognitive Control and Development; Connectionist Models; Connectionist Models of Language Processing; Executive Function and Higher-Order Cognition: Computational Models; Hippocampus: Computational Models; Memory: Computational Models; Numerical Intelligence: Neural Substrates.

Further Reading

Elman J, Bates E, and Johnson M (1996) *Rethinking Innateness: A Connectionist Perspective on Development (Neural Network Modeling and Connectionism)*. Cambridge, MA: MIT Press.

Lakoff G (1987) *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago, IL: University of Chicago Press.

McClelland JL and Rumelhart DE (1986) *Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Newell A (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Posner MI (ed.) (1985) *Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Regier T (1996) *The Human Semantic Potential*. Cambridge, MA: MIT Press.

Russell SJ and Norvig P (2000) *Artificial Intelligence*. Upper Saddle River, NJ: Prentice-Hall.

Astrocyte: Calcium Signaling

J W Deitmer, FB Biologie, Kaiserslautern, Germany
A Araque, Cajal Institute, Madrid, Spain

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Ca^{2+} signaling has been recognized as one of the major second messenger steps in most cell types, including astrocytes, the major macroglial cell type in vertebrate nervous systems. Astrocytes are endowed with a large number of metabotropic receptors in their cell membrane, most of which are coupled to G-protein-mediated activation of phospholipase C (PLC), and hence the formation of inositol trisphosphate (IP_3) and diacylglycerol. The second messenger IP_3 diffuses within the cytosol, binds to IP_3 receptors in the membrane of the endoplasmic reticulum (ER), and gates Ca^{2+} channels, resulting in a cytosolic Ca^{2+} rise due to release of Ca^{2+} from the ER. This can be a single Ca^{2+} transient with or without a shoulder or plateau phase, repetitive Ca^{2+} transients, so-called Ca^{2+} oscillations, or irregular Ca^{2+} rises, depending on the species of primary messenger (neurotransmitter, hormone, or growth factor) and its concentration. These Ca^{2+} signals may spread along the cell and across cell boundaries to neighboring astrocytes in the form of Ca^{2+} waves, and they can be evoked or modulated by neuronal activity. The signaling pathway leading to a cytosolic Ca^{2+} rise, common to many cell types, may be regarded as a type of 'excitation' in electrically inexcitable cells such as astrocytes. The cytosolic Ca^{2+} signaling may initiate a variety of glial functions and interactions with other cells, for example, by Ca^{2+}-dependent release of transmitters (gliotransmitters) or by affecting vasoconstriction of blood vessels in the brain. This article reviews the types of cytosolic Ca^{2+} signaling in astrocytes, their different modes of initiation, and their functional significance for astrocytes and glia–neuron communication.

Identification of Astrocytes by Ca^{2+} Signals

Identification of astrocytes, in culture as in the tissue, has been established by using staining of glial-specific marker proteins by immunocytochemical techniques. Most astrocytes express an intermediate filament, the glial fibrillar acidic protein (GFAP), particularly in their cellular processes, and a Ca^{2+}-binding protein S100β, which do not occur in neurons. Hence, these two proteins are often detected as glial-specific markers in astrocytes, although they may not be exclusively restricted to this cell type. Brain slices incubated with the Ca^{2+}-sensitive dye (e.g., Fluo-4 AM) can be fixed with EDAC (1-ethyl-3-(3-dimethyl-aminopropyl)carbodiimide) and then stained for GFAP after experiments in which intracellular Ca^{2+} was measured to identify the cells of interest (**Figures 1(a) and 1(b)**).

The use of fluorescent tags to highly expressed genes of glial cells by creating transgenic mice has produced fluorescent astrocytes for identification. In these mice, astrocytes were labeled by the green fluorescent protein (GFP) or the enhanced green fluorescent protein (EGFP) under the control of the human GFAP promoter (**Figures 1(c) and 1(d)**). The level of EGFP expression, however, was variable in different brain areas. High expression was found in the cerebellum, but low expression was found in retina and hypothalamus, where some astrocytes were not labeled at all. Oligodendrocytes and neurons were not stained by these fluorescent tags. The red fluorescent dye sulforhodamine 101 was specifically taken up by a type of astrocytes ('protoplasmic'), which also expressed the EGFP, and was used to quantify morphological characteristics and the close association of astrocytes with cortical microvasculature.

When recording cytosolic Ca^{2+}, astrocytes can be identified by an erratic Ca^{2+} rise at low external K^+ concentrations (<1 mM), observed in rat and mouse astrocytes, but not in neurons. When superfusing a cell culture or brain slices with a low-K^+ solution for 3–5 min, more than 80% of cells that show an immunoreaction to GFAP and S100β antibodies respond with a monophasic or repetitive Ca^{2+} rise (**Figure 1(e)**). This Ca^{2+} response persists after depleting the intracellular Ca^{2+} stores, but it is suppressed in the absence of external Ca^{2+}, indicating that it results from Ca^{2+} influx. Since this low-K^+-induced Ca^{2+} rise can be inhibited by micromolar amounts of Ba^{2+}, it has been suggested that an inward rectifier K^+ channel becomes permeable to divalent cations. This idea is supported by using $K_{ir}4.1$ knockout mice, in which the low-K^+-induced Ca^{2+} response is significantly reduced. The $K_{ir}4.1$ appears to be highly expressed in oligodendrocytes and some astrocytes, in which this channel is believed to be involved in the clearance of extracellular K^+.

In summary, various methods of identifying glial cells may not be completely reliable on their own, but the use of a combination of two or three of these methods can identify astrocytes with a high degree of certainty.

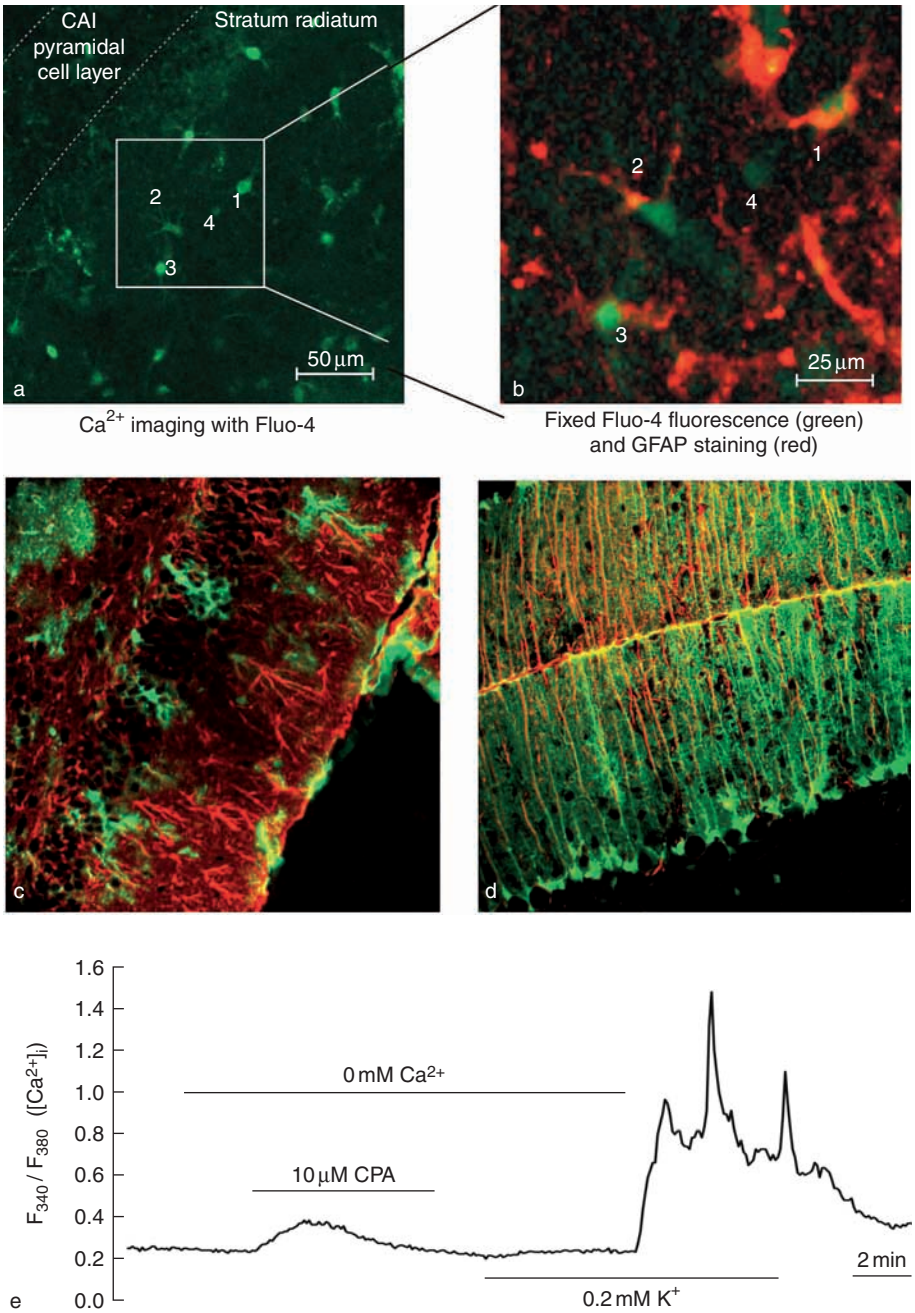

Figure 1 Identification of astrocytes. (a, b) Confocal images of a brain slice after loading cells with Fluo-4 (a) and after fixation of the fluorescence with EDAC (1-ethyl-3-(3-dimethyl-aminopropyl)carbodiimide) and staining with antibodies against GFAP (b). (c, d) Confocal images of GFAP-labeled cryosections from hippocampus with dendate gyrus (c) and from cerebellum showing molecular layer with Bergmann glial cell bodies and radial processes with endfeet (d) of transgenic mice. Scale value for (c) and (d) is same as in (a). (e) Low-K^+-induced Ca^{2+} responses cultured rat astrocytes after depleting intracellular Ca^{2+} stores with CPA (cyclopiazonic acid) are due to Ca^{2+} influx. (a, b) From Beck A, Nieden RZ, Schneider HP, and Deitmer JW (2004) Calcium release from intracellular stores in rodent astrocytes and neurons *in situ*. *Cell Calcium* 35: 47–58. (c, d) From Nolte C, Matyash M, Pivneva T, et al. (2001) GFAP promoter-controlled EGFP-expressing transgenic mice: A tool to visualize astrocytes and astrogliosis in living brain tissue. *Glia* 33: 72–86. (e) From Dallwig R, Vitten H, and Deitmar JW (2000) A novel barium-sensitive calcium influx into rat astrocytes at low external pottassium. *Cell Calcium* 28: 247–259.

Spontaneous Ca^{2+} Transients and Oscillations

Spontaneous Ca^{2+} transients and oscillations have been reported in both neurons and glial cells, in culture, and *in situ* and *in vivo*. Ca^{2+} oscillations are defined as repetitive rises of cytosolic Ca^{2+} and may occur intermittently with single Ca^{2+} transients and with intervals of up to several minutes (Figure 2(a)). These Ca^{2+} signals can propagate as waves along the

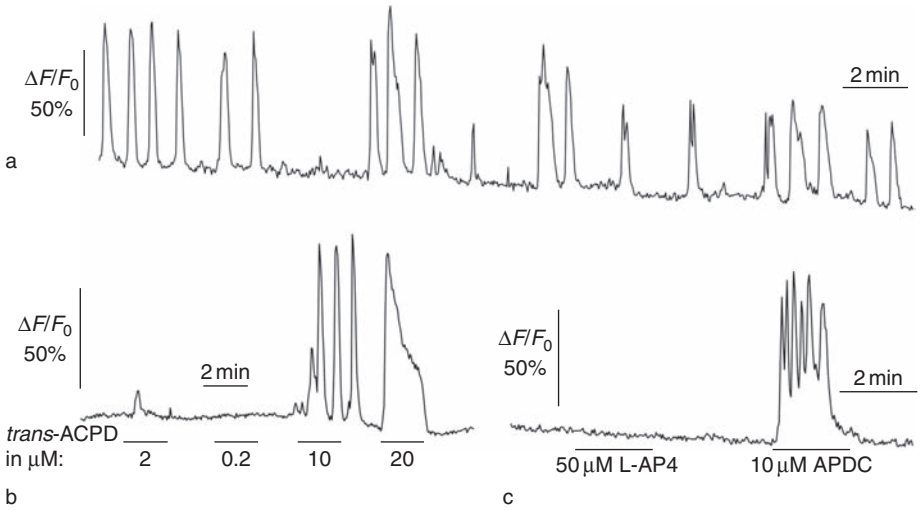

Figure 2 Ca^{2+} oscillations in acute rat hippocampal slices. (a) Continuous recording of spontaneous Ca^{2+} signals in an astrocyte. (b, c) Different concentrations of t-ACPD (b), an agonist of metabotropic glutamate receptors of group I, evoking single Ca^{2+} transients at low concentrations (1 μM), larger Ca^{2+} transients with shoulder at higher concentrations (20 μM), and Ca^{2+} oscillations at intermediate concentrations (5–10 μM). The metabotropic glutamate receptor agonist of group II APDC, but not the agonist of group III L-AP4 (c), can also elicit Ca^{2+} oscillations. From Zur Nieden R and Deitmer JW (2006) The role of metabotropic glutamate receptors for the generation of calcium oscillations in rat hippocampal astrocytes *in situ*. *Cerebral Cortex* 16: 676–687.

cell processes and even beyond cell boundaries to neighboring glial cells. Ca^{2+} transients and Ca^{2+} oscillations are usually evoked by release of Ca^{2+} from intracellular stores (ER) and hence are blocked by inhibiting the Ca^{2+}-ATPase of the ER and/or by IP_3 receptor antagonists. These spontaneous Ca^{2+} transients/oscillations occurred in 35% of astrocytes in the hippocampus of juvenile rats and in 82% of hippocampal astrocytes in juvenile mice with a rate of approximately 1.3 transients/min.

These Ca^{2+} signals require operating mechanisms for refilling the intracellular Ca^{2+} stores, which include store-operated calcium entry (SOCE) through channels in the plasma membrane. Spontaneous Ca^{2+} signals in astrocytes are not immediately affected by neuronal activity, but their synchronous occurrence may be affected after blocking neuronal activity by tetrodotoxin. As discussed later, various neurotransmitters, such as glutamate, noradrenaline, histamine, and ATP, as well as mechanical stimulation can evoke transients, oscillations, and waves of cytosolic Ca^{2+} in astrocytes (Figures 2(b) and 2(c)).

Spontaneous Ca^{2+} transients and oscillations have been associated with nonsynaptic release of glutamate (e.g., via the cystine–glutamate exchanger) and hence with the glutamate background level in the extracellular space, acting on metabotropic glutamate receptors of groups I and II in the astrocyte membrane. Thus, the frequency of these spontaneous Ca^{2+} signals may help to sense the level of extracellular glutamate in nervous tissue. The elaboration of the exact mechanism and of other functions of spontaneous Ca^{2+} transients and oscillations in astrocytes, and their mode of initiation and maintenance, require more experimental analysis.

Molecular Mechanisms of Ca^{2+} Signaling

Calcium signaling in both excitable and nonexcitable cells is based on the existence of a relatively low concentration of cytosolic Ca^{2+} (usually 50–100 nM), and a Ca^{2+} concentration gradient between the cytosol and the extracellular space or the lumen of intracellular organelles (e.g., the ER, where Ca^{2+} concentration may be as high as hundreds of micromolar or even a few millimolar). Cellular Ca^{2+} signaling, manifested as rapid intracellular Ca^{2+} rises, may result from several Ca^{2+} sources. Cytosolic Ca^{2+} levels can increase via Ca^{2+} influx from the extracellular spaces across the plasma membrane and by Ca^{2+} release from intracellular stores. Calcium influx may occur through specific Ca^{2+}-permeable ligand-gated channels, voltage-gated Ca^{2+} channels that are activated by membrane depolarization, and Ca^{2+}-permeable, store-operated channels that are activated by depletion of intracellular Ca^{2+} stores and that help to refill these stores.

Whereas electrically excitable cells use the electrochemical gradient across the plasma membrane to effectively increase the intracellular Ca^{2+} levels due to the high expression of specific ligand- and voltage-gated ion channels that allow the rapid and massive influx of Ca^{2+} ions, nonelectrically excitable cells such as astrocytes mainly use the Ca^{2+} stored

in the ER as the main source for cytoplasmic Ca^{2+} signaling. Ionotropic glutamate/α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acic (AMPA) receptors, known to be expressed by astrocytes from different brain regions, including cortex, hippocampus, cerebellum, and retina, lack the GluR2 subunit, being therefore permeable to Ca^{2+}. However, aside from this special case, Ca^{2+} influx through AMPA receptors does not seem to play a major role in astrocyte Ca^{2+} signaling because in most brain regions astrocytes express AMPA receptors that contain the GluR2 subunit, such as Bergmann glial cells in the cerebellar cortex, and hence display a low Ca^{2+} permeability. Activation of other ligand-gated channels permeable to Ca^{2+}, such as ionotropic purinergic receptors (P2X receptors) or nicotinic acetylcholine receptors containing the α7 subunit, has been shown to lead to influx of Ca^{2+} in cell culture preparations. However, studies *in situ* have failed to observe Ca^{2+} responses in astrocytes under these conditions, which further suggests that Ca^{2+} influx through ligand-gated channels is not a major pathway for astrocyte cytosolic Ca^{2+} signaling.

Calcium influx through voltage-gated Ca^{2+} channels in the plasma membrane constitutes another source for intracellular Ca^{2+} elevation in different cell types. Astrocytes in culture express a variety of voltage-gated channels, including Ca^{2+} channels; however, studies performed *in situ* indicate that functional voltage-gated Ca^{2+} channels are mainly expressed by immature astrocytes, and this expression gradually disappears during development. Hence, the low, if any, level of expression of these channels in mature astrocytes does not serve to increase the intracellular Ca^{2+} levels to any greater extent. Therefore, unlike neurons and other electrically excitable cells that express a high density of ligand- and voltage-gated channels that are permeable to Ca^{2+} and allow an efficient rise of cytosolic Ca^{2+}, astrocytes employ different cellular mechanisms to increase their cytosolic Ca^{2+} concentration.

Astrocytes express a wide variety of functional receptors for many neurotransmitters (including glutamate, norepinephrine, γ-aminobutyric acid (GABA), histamine, ATP, adenosine, and acetylcholine), most of which belong to the metabotropic receptor family. These receptors are associated with G-proteins that upon activation stimulate phospholipase C, which leads to the formation of the intracellular second messenger IP_3. Activation of specific receptors in the ER by IP_3 evokes Ca^{2+} release from the IP_3-sensitive Ca^{2+} stores, thus increasing the cytosolic Ca^{2+} concentration. This cellular mechanism can represent one important mode of neuron-to-astrocyte communication, in which the astrocyte Ca^{2+} signal is controlled by neurotransmitter(s) released during synaptic activity.

In addition to IP_3 receptors, the membrane of the ER of most cells also contains ryanodine receptors, a different type of Ca^{2+}-permeable receptor that is activated by cytosolic Ca^{2+}. An increase in the cytosolic Ca^{2+} leads to the opening of the ryanodine receptors, which can be directly stimulated by caffeine and hence induce the release of Ca^{2+} from the internal stores, a process known as Ca^{2+}-induced Ca^{2+} release. The presence of these receptors in astrocytes, however, is controversial; moreover, the lack of caffeine-sensitive Ca^{2+} release, as found in astrocytes in culture and *in situ*, and in contrast to neurons, suggests a minor, if any, role of these receptors in the astrocyte Ca^{2+} activity.

The maintenance of the high intraluminal Ca^{2+} concentration in the ER depends on Ca^{2+} entry using specific Ca^{2+}-permeable plasma membrane channels, called store-operated channels, that are regulated by the filling state of the ER and that serve to replenish the stores with Ca^{2+}. This SOCE, also called capacitative Ca^{2+} entry, has been observed in many different cell types and recorded electrophysiologically as a persistent current, called I_{CRAC} for 'Ca^{2+} release-activated current.' Although I_{CRAC} has not been identified in astrocytes, SOCE has been clearly demonstrated in these cells by Ca^{2+} imaging studies. SOCE can be detected as a sustained plateau of the intracellular Ca^{2+} levels that depends on the previous Ca^{2+} depletion of intracellular Ca^{2+} stores and that requires the presence of extracellular Ca^{2+}. A diffusible messenger termed Ca^{2+} influx factor (CIF) has been suggested to act through activation of inducible, Ca^{2+}-independent phospholipase A_2 activity. It has been proposed that the EF-hand protein stromal interaction molecule 1 (STIM1), which migrates from the ER to the plasma membrane and activates calcium release-activated channel (CRAC) channels upon Ca^{2+} store depletion, functions as this CIF. An exciting question is whether this mechanism of Ca^{2+} influx serves as a further Ca^{2+} signal beyond the simple role of refilling internal stores.

It is the interplay of the different cellular mechanisms (Ca^{2+} influx and release, Ca^{2+} extrusion and uptake, and Ca^{2+} buffering) that control the intracellular Ca^{2+} levels that is crucial for the generation and shape of astrocyte Ca^{2+} signaling. Mechanisms that are involved in regulating cytosolic Ca^{2+} contribute to shaping the kinetics and amplitude of Ca^{2+} transients. Ca^{2+}-binding proteins, such as parvalbumin, calcineurin, and calmodulin, that are present in the cytoplasm may serve to buffer cytosolic Ca^{2+} concentration. However, this is a low-capacity system that is more effective as a mediator of Ca^{2+}-dependent cellular processes. Energy-consuming

mechanisms are involved in the effective Ca^{2+} clearance. Sarco/endoplasmic reticulum Ca^{2+}-ATPases pump Ca^{2+} ions into the ER using ATP as an energy source. Ca^{2+}-ATPases and Na^{+}/Ca^{2+} exchangers present in the plasma membrane use ATP or the electrochemical gradient of Na^{+} across the plasma membrane, respectively, to extrude Ca^{2+} against a steep concentration gradient across the plasma membrane. Finally, mitochondria are dynamic Ca^{2+} stores that can also sequester Ca^{2+} through the activity of the Ca^{2+} uniporter, with a relatively low affinity but a large capacity to store Ca^{2+} transiently, and protect the cell from cytosolic Ca^{2+} overload, which may lead to cell death.

Intra- and Intercellular Ca^{2+} Signaling

The astrocyte Ca^{2+} signal can be manifested as Ca^{2+} elevations with different temporal patterns (Figure 2). In some cases, astrocytes display a single, fast Ca^{2+} transient that eventually can be followed by a more sustained response. In other cases, astrocytes exhibit irregular Ca^{2+} oscillations that can be modulated in amplitude, duration, and frequency by different stimuli, including the synaptic activity level. The astrocyte Ca^{2+} rise may act as an intra- and intercellular signaling mode that can propagate within and between astrocytes, signaling to different regions of the cell and to different cells (Figure 3).

The astrocyte Ca^{2+} signal that occurs spontaneously or is evoked by synaptic activity can be initiated in spatially restricted areas, called 'microdomains,' of the astrocytic processes. From these discrete regions, the Ca^{2+} signal can propagate along the process to other regions of the astrocyte, generating a wave of intracellular Ca^{2+}. A thorough characterization of the intracellular waves is yet to be accomplished; however, the spatial spread of the intracellular wave seems to be

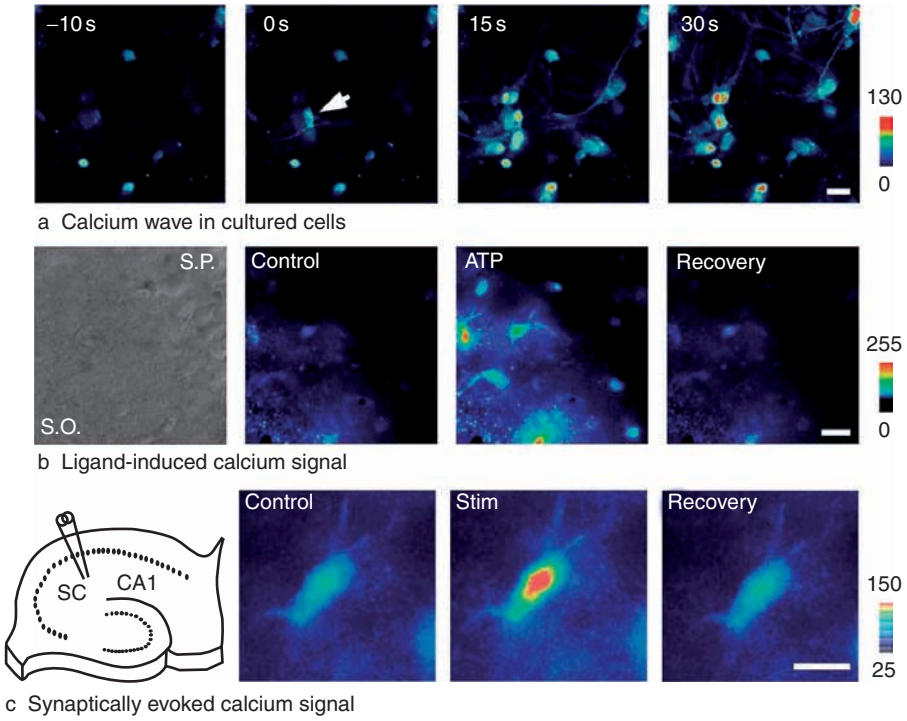

Figure 3 Ca^{2+} signaling in astrocytes evoked by different stimuli. (a) Intercellular Ca^{2+} wave in cultured rat hippocampal astrocytes. Astrocytes were bulk loaded with the fluorescent Ca^{2+} indicator Ca^{2+}-Green 1. Images in pseudocolor mode represent fluorescence intensity of Ca^{2+}-Green 1 emission taken during and after mechanical stimulation of a cell (arrow; time zero) at the times indicated. Mechanical stimulation increased intracellular Ca^{2+} in the directly stimulated cell as well as in neighboring, unstimulated astrocytes by a propagating Ca^{2+} wave. (b) Infrared differential interference contrast image and pseudocolor images representing fluorescence intensities of a Fluo-3-filled rat hippocampal slice in control, after ionophoretical application of ATP for 5 s, and after recovery. S.O., stratum oriens; S.P., stratum pyramidale. (c) Responses evoked by Schaffer collateral (SC) stimulation in rat hippocampal astrocytes. Schematic drawing of the hippocampal slice preparation and pseudocolor images representing fluorescence intensities of a Fluo-3-filled astrocyte in control, 10 s after stimulation of the SC (at 30 Hz for 5 s), and after recovery. Scale bar = 20 μm (a, b), 10 μm (c). (a) Adapted from Perea G and Araque A (2005) Glial calcium signalling and neuron–glia communication. *Cell Calcium* 38: 375–382. (b, c) Adapted from Perea G and Araque A (2005) Properties of synaptically evoked astrocyte calcium signal reveal synaptic information processing by astrocytes. *Journal of Neuroscience* 25: 2192–2203.

under the control of synaptic activity. The intracellular compartmentalization and the control of the propagation of the Ca^{2+} signal are highly relevant for cellular functions and the astrocyte-to-neuron communication. Since 'gliotransmitters' released from astrocytes through Ca^{2+}-dependent mechanisms may act as modulators of synaptic transmission, the mechanisms controlling the intracellular Ca^{2+} wave may be of great significance because they will determine the degree of extension of the signal that triggers the neuromodulatory effects of astrocytes. The fact that the Ca^{2+} signal evoked by synaptic activity is initiated in restricted regions of the astrocyte reveals that the neuron-to-astrocyte communication is a localized event that does not result from uncontrolled spillover of neurotransmitter but more likely from a point-to-point signaling between synaptic terminals and astrocytic processes.

Astrocytes can also communicate with adjacent cells by generating intercellular Ca^{2+} waves, which are propagated along cell processes onto neighboring astrocytes (Figures 3(a) and 3(b)). Different stimuli, such as mechanical stimuli or exogenous, focal, application of neurotransmitters, may induce intracellular Ca^{2+} increases that can propagate as Ca^{2+} waves between astrocytes in cultured cells as well as organotypic and acute brain slices. These waves can extend for long distances ($<500\,\mu m$) at relatively low speed ($\sim 14\,\mu m\,s^{-1}$), and they may constitute a novel form of slow, long-distance cellular communication in the nervous system. Ca^{2+} waves in the astrocytic syncytium were originally thought to spread as a result of gap junction-mediated diffusion of IP_3 between cells. Later studies proposed that, in addition to diffusion of IP_3, other mechanisms that involve extracellular messengers, such as extracellular ATP, may contribute to the propagation of astrocyte Ca^{2+} waves. ATP released from active astrocytes appears to be an important mediator of long-range Ca^{2+} signaling, whereas shorter range signaling may be mediated by gap junctions. More experimental evidence is required to determine whether synaptic processes in different areas that lack immediate neuronal connection are modulated by these astrocytic Ca^{2+} waves.

Ca^{2+} Responses to Transmitters and Other Signaling Molecules

Elevations of the cytosolic Ca^{2+} concentration in astrocytes may be triggered by a large number of signaling molecules, including hormones, prostaglandins, and many neurotransmitters. As described previously, astrocytes express a wide variety of functional receptors for most neurotransmitters, and many of them belong to the metabotropic family; that is, they are coupled to second messenger pathways that lead to the IP_3-mediated Ca^{2+} release from intracellular stores. Activation of these receptors by exogenous application of messengers such as glutamate, norepinephrine, 5-hydroxytryptamine, histamine, acetylcholine, dopamine, nitric oxide (NO), ATP, and GABA can increase the intracellular Ca^{2+} level in astrocytes (Figure 3c). These Ca^{2+} signals can be linked to initiating cellular activity, such as transmitter release, a rise in the K^+ conductance of the cell membrane, or mitochondrial energy production.

Astrocytes may also respond with intracellular Ca^{2+} elevations to other signaling molecules such as chemokines, a class of small proteins that bind to G-protein-coupled receptors that were originally identified as inflammatory mediators of leukocyte chemotaxis but later shown to possess additional roles beyond neuroinflammation. Various types of chemokine receptors are expressed in cells of the nervous system, including astrocytes, in which the chemokine stromal cell-derived factor-1 has been shown to increase astrocytic Ca^{2+} through activation of the receptor CXCR4.

Ca^{2+} Responses to Neuronal Activity

Astrocyte receptors can be activated by neurotransmitters released from synaptic terminals to initiate astrocyte Ca^{2+} signaling (Figure 3c). The control of the astrocyte Ca^{2+} signal by synaptic activity has been reported in retina, cerebellum, hippocampus, and cortex and has been shown to be mediated by glutamate, GABA, acetylcholine, noradrenaline, and NO. Further studies are required to elucidate the involvement of other neurotransmitter systems in the neuron-to-astrocyte communication.

The fact that the astrocyte Ca^{2+} level may be increased by synaptic transmitters does not imply that the Ca^{2+} signal simply reflects synaptic activity; rather, it represents a complex signal that can be tuned by neuronal activity. The amplitude and duration of Ca^{2+} transients as well as the frequency of Ca^{2+} oscillations in astrocytes may vary depending on the level of synaptic activity. Moreover, studies investigating the synaptic control of astrocyte Ca^{2+} signaling have demonstrated that astrocytes display integrative properties for synaptic information processing because astrocytes can discriminate between the activity of synaptic terminals belonging to different axon pathways. Moreover, the synaptically evoked astrocyte Ca^{2+} signal can be bidirectionally modulated by the interaction of different synaptic inputs, being potentiated or depressed depending on the level of synaptic activity. This modulation may also control the intracellular spread of the Ca^{2+} signal in the form of waves, which may have important

consequences on brain function by regulating the spatial extension of the influence of a single astrocyte on different synapses.

Functional Significance of Ca^{2+} Signaling

There are numerous processes in the nervous system that may be initiated and/or modulated by cytosolic Ca^{2+} rises in astrocytes, as has been reported for many other cell types. On the other hand, astrocytic Ca^{2+} signaling is subject to modulation by neuronal activity, particularly at synapses, and therefore is one of the prime mechanisms by which reciprocal neuron–glia signaling is established. One of the prime and most consequential processes initiated by Ca^{2+} rises in astrocytes is the Ca^{2+}-dependent release of transmitters, so-called 'gliotransmitters.' Some of these gliotransmitters and glia-derived messengers, such as glutamate, ATP, tumor necrosis factor-α, or D-serine, have been shown to modulate neuronal excitability, synaptic transmission, and cerebrovascular microcirculation. Some of these transmitters can be released by a Ca^{2+}-dependent mechanism, although Ca^{2+}-independent modes of release have also been reported, such as through large membrane anion channels, purinergic $P2X_7$ channels, or gap junction hemichannels. Therefore, these gliotransmitters may serve as feedback signals that modulate neuronal activity. Astrocyte-induced modulation of synaptic transmission has been observed in cultured cells as well as in tissue slices of several brain areas. Since astrocytes respond to neurotransmitters with calcium elevations that can extend relatively large distances and that induce release of gliotransmitters that modulate neurotrasnmission, astrocyte Ca^{2+} signaling may represent a form of brain information pathway that establishes a functional link between distant synapses.

Neuronal excitability has also been shown to be modulated by glutamate released from astrocytes both *in vitro* and *in situ*. Calcium-dependent glutamate release from astrocytes may have a strong impact on brain pathophysiology because it can lead to the synchronized activity of clusters of neurons, and it may be responsible for the generation of epileptiform activity in neurons.

Other messenger molecules that have been reported to be released from astrocytes in a Ca^{2+}-dependent fashion include NO (e.g., mediated by activation of metabotropic purinergic P2Y receptors). This short-lived volatile messenger, which acts both in an ortho- and a retrograde way, may be a link to other second messenger cascades in neighboring neurons and glial cells, such as the cyclic guanosine monophosphate (cGMP)-mediated pathway. Moreover, NO may enhance SOCE and thereby feeds back on the cellular competence to generate Ca^{2+} signals due to repeated Ca^{2+} release from intracellular stores.

Ca^{2+} signaling in astrocytes has also been associated with the control of cerebral blood flow. Ca^{2+} transients and Ca^{2+} oscillations are propagated to the astrocytic endfeet, which ensheath part of the endothelial layer of blood capillaries, and can elicit vasodilation and vasoconstriction by the release of lipid metabolites. Inhibition of cyclooxygenase-1 activity blocked astrocyte-triggered vasodilation, supporting the notion that Ca^{2+}-dependent release of prostaglandins from astrocytes mediates the control of local blood flow. This modulation may be initiated by neuronal activity in restricted brain areas by evoking astrocyte Ca^{2+} signals that in turn mediate the control of microcirculation and hence the supply of oxygen and glucose.

Furthermore, cytosolic Ca^{2+} rises in astrocytes may gate or modulate ion channels in the cell membrane, particularly K^+ channels, which may be involved in K^+ clearance from the extracellular space following neuronal activity. The list of cellular functions mediated by Ca^{2+} signals in astrocytes will certainly grow in the future. It is the regional control of cellular functions that may be regulated or modulated by Ca^{2+} signals in a unique way. By invading different cellular processes of astrocytes, by being regenerative, and by being a versatile mode of signaling, this form of 'glial excitation' has many assets for allowing a complex dialogue between glial cells and neurons, and it may contribute critically to information processing and metabolism in the brain.

Summary

Cytosolic Ca^{2+} signals in astrocytes can be spontaneous or evoked, and they are often initiated by activation of metabotropic receptors, resulting in Ca^{2+} release from intracellular Ca^{2+} stores. Whereas voltage-dependent Ca^{2+} influx is rare in astrocytes, the refilling of the intracellular Ca^{2+} stores requires store-operated Ca^{2+} entry, a major influx pathway of Ca^{2+} into astrocytes. Astrocyte Ca^{2+} signals are generated in the form of transients or oscillations, which can be evoked by neurotransmitters, hormones, cytokines, and growth factors, often in response to neuronal activity. These Ca^{2+} signals are propagated along cellular processes and can travel from the point of origin across the tissue as Ca^{2+} waves. When propagated into astrocytic endfeet, they may release cyclooxygenase products to control cerebral blood flow. Astrocyte Ca^{2+} signaling may lead to the release of transmitters and may be one of the key mechanisms allowing a complex dialogue between neurons and glial cells.

See also: Astrocyte: Response to Injury; Astrocyte: Neurotransmitter and Hormone Receptors; Astrocyte: Identification Methods; Calcium and Signal Transduction; Neurotransmitter Release from Astrocytes.

Further Reading

Araque A and Perea G (2004) Glial modulation of synaptic transmission in culture. *Glia* 47: 241–248.

Beck A, Nieden RZ, Schneider HP, and Deitmer JW (2004) Calcium release from intracellular stores in rodent astrocytes and neurons *in situ*. *Cell Calcium* 35: 47–58.

Berridge MJ (2005) Unlocking the secrets of cell signalling. *Annual Review of Physiology* 67: 1–21.

Castonguay A, Levesque S, and Robitaille R (2001) Glial cells as active partners in synaptic functions. *Progress in Brain Research* 132: 227–240.

Dallwig R, Vitten H, and Deitmer JW (2000) A novel barium-sensitive calcium influx into rat astrocytes at low external potassium. *Cell Calcium* 28: 247–259.

Deitmer JW, Verkhratsky A, and Lohr C (1998) Calcium signalling in glial cells. *Cell Calcium* 24: 405–416.

Fiacco TA and McCarthy KD (2006) Astrocyte Calcium elevation: Properties, propagation and effects on brain signalling. *Glia* 54: 676–690.

Haydon PG (2001) Glia: Listening and talking to the synapse. *Nature Reviews Neuroscience* 2: 185–193.

Kettenmann H and Ransom BR (2005) *Neuroglia,* 2nd edn. New York: Oxford University Press.

Nedergaard M, Ransom BR, and Goldman SA (2003) New roles for astrocytes: Redefining the functional architecture of the brain. *Trends in Neurosciences* 26: 523–530.

Newman EA (2004) Glial modulation of synaptic transmission in the retina. *Glia* 47: 268–274.

Nolte C, Matyash M, Pivneva T, et al. (2001) GFAP promoter-controlled EGFP-expressing transgenic mice: A tool to visualize astrocytes and astrogliosis in living brain tissue. *Glia* 33: 72–86.

Perea G and Araque A (2005) Glial calcium signalling and neuron–glia communication. *Cell Calcium* 38: 375–382.

Perea G and Araque A (2005) Properties of synaptically evoked astrocyte calcium signal reveal synaptic information processing by astrocytes. *Journal of Neuroscience* 25: 2192–2203.

Porter JT and McCarthy KD (1997) Astrocytic neurotransmitter receptors *in situ* and *in vivo*. *Progress in Neurobiology* 51: 439–455.

Singaravelu K, Lohr C, and Deitmer JW (2006) Regulation of store-operated calcium entry by calcium-independent phospholipase A_2 in rat cerebellar astrocytes. *Journal of Neuroscience* 26: 9579–9592.

Verkhratsky A, Orkand RK, and Kettenmann H (1998) Glial calcium: Homeostasis and signalling function. *Physiological Reviews* 78: 99–141.

Volterra A, Magistretti PJ, and Haydon PG (eds.) (2002) *The Tripartite Synapse: Glia in Synaptic Trasnmission*. New York: Oxford University Press.

Volterra A and Meldolesi J (2005) Astrocytes, from brain glue to communication elements: The revolution continues. *Nature Reviews Neuroscience* 6: 626–640.

Volterra A and Steinhauser C (2004) Glial modulation of synaptic transmission in the hippocampus. *Glia* 47: 249–257.

Zur Nieden R and Deitmer JW (2006) The role of metabotropic glutamate receptors for the generation of calcium oscillations in rat hippocampal astrocytes *in situ*. *Cerebral Cortex* 16: 676–687.

Astrocyte: Identification Methods

A Scheller and F Kirchhoff, Max Planck Institute of Experimental Medicine, Göttingen, Germany

© 2009 Published by Elsevier Ltd.

Introduction

Astrocytes are the major glial cell type of the central nervous system (CNS). Originally, they were described as nerve glue (*Nervenkitt*) by the German neuropathologist Rudolf Virchow in 1858. Work during the past few decades, however, demonstrated that astrocytes play numerous important roles in brain function. Astrocytes are responsible for nourishing neurons, regulation of extracellular ion and transmitter homeostasis, and maintenance of the blood–brain barrier. In addition, they act as guiding structures for neuronal migration during development, contribute to neurotransmitter metabolism, and participate in repair and regeneration processes. Furthermore, it has been shown that astrocytes are in close contact to active synaptic terminals, modulate synaptic transmission, or possess even neurogenic potential. The current view regards astrocytes as equal partners of neurons.

Methods to identify astrocytes are based on three of their basic properties: unique cellular structure, selective gene expression, and characteristic physiological membrane properties (Figure 1).

Astroglial Structure

In the late nineteenth century, the neuroanatomists Karl Weigert, Albert von Kölliker, and Michael von Lenhossek were the pioneers who described astrocytes as distinct cellular elements in the CNS by taking advantage of simultaneous improvements of the light microscope and staining techniques. In the beginning, differential binding properties of dyes such as hematoxylin, carmine, nigrosin, and a series of aniline derivatives were used to distinguish cell types and subcellular structures of the brain. Some of these dyes are still in use today. Later, the silver and gold impregnation technique developed by Camillo Golgi and ingeniously applied by Santiago Ramón y Cajal in the late nineteenth and early twentieth century fully revealed the polarized three-dimensional structure of astrocytes as we know it today. From a small soma less than 10 μm in diameter, major processes extend which give rise to a highly arborized network of finer processes. From these endings, numerous membrane protrusions emanate and enwrap synaptic terminals. In addition, astroglial endfeet of distinct major processes contact brain capillaries by interaction with endothelial cells and pericytes.

In the 1950s and 1960s, the complete fine structure of astrocytes was uncovered by Sanford Louis Palay and colleagues using electron microscopy (EM). This technique highlighted the differences of major proximal processes radiating from the small astroglial soma and distal irregular side branches. Whereas the first are characterized by the presence of intracellular fiber bundles, the distal endings of the other processes are, beside some glycogen granules, almost devoid of electron-dense material and can be as thin as 20 nm. Together with the synapses, these lamellar extensions form functional units throughout the nervous system, which led to the concept of tripartite synapses in which the pre- and postsynaptic terminals are enwrapped by thin astroglial compartments. Reconstruction of hundreds of serial EM sections revealed the extreme complexity of the astroglial process network.

Cell-Type-Specific Gene Expression

Glial Fibrillary Acidic Protein: The Main Constituent of Astroglial Intermediate Filaments

In the 1970s, Lawrence Eng and colleagues purified the intracellular astroglial fibers from gliotic autopsy material. The 50 kDa glial fibrillary acidic protein (GFAP) was identified as its main constituent. This discovery represented an important milestone in astroglia research. Since GFAP turned out to be highly immunogenic, a series of mono- and polyclonal antibodies could be generated. These reagents significantly facilitated subsequent immunohistochemical studies analyzing astrocytes *in situ*. Today, GFAP is the most common marker for astrocytes. Due to the availability and reliability of antibodies against GFAP, it was simple to selectively label astrocytes in normal, as well as in pathological brain tissue. In subsequent years, morphologically distinct astroglial subtypes such as fibrous astrocytes in the white matter or protoplasmic astrocytes in gray matter were identified. Now, at the light microscopic level, primarily using fluorescence microscopy, the direct contacts of astrocytes to blood vessels, the parallel orientation of astroglial fibers along myelinated axons, or the radial interdigitation of cerebellar Bergmann glia fibers through the parallel fiber–Purkinje cell network can readily be visualized. GFAP immunostaining is highly relevant to understanding brain development and pathology. In early work using GFAP antibodies, Amico Bignamy and Doris Dahl demonstrated that in the three mouse mutants *reeler*, *staggerer*, and *weaver*,

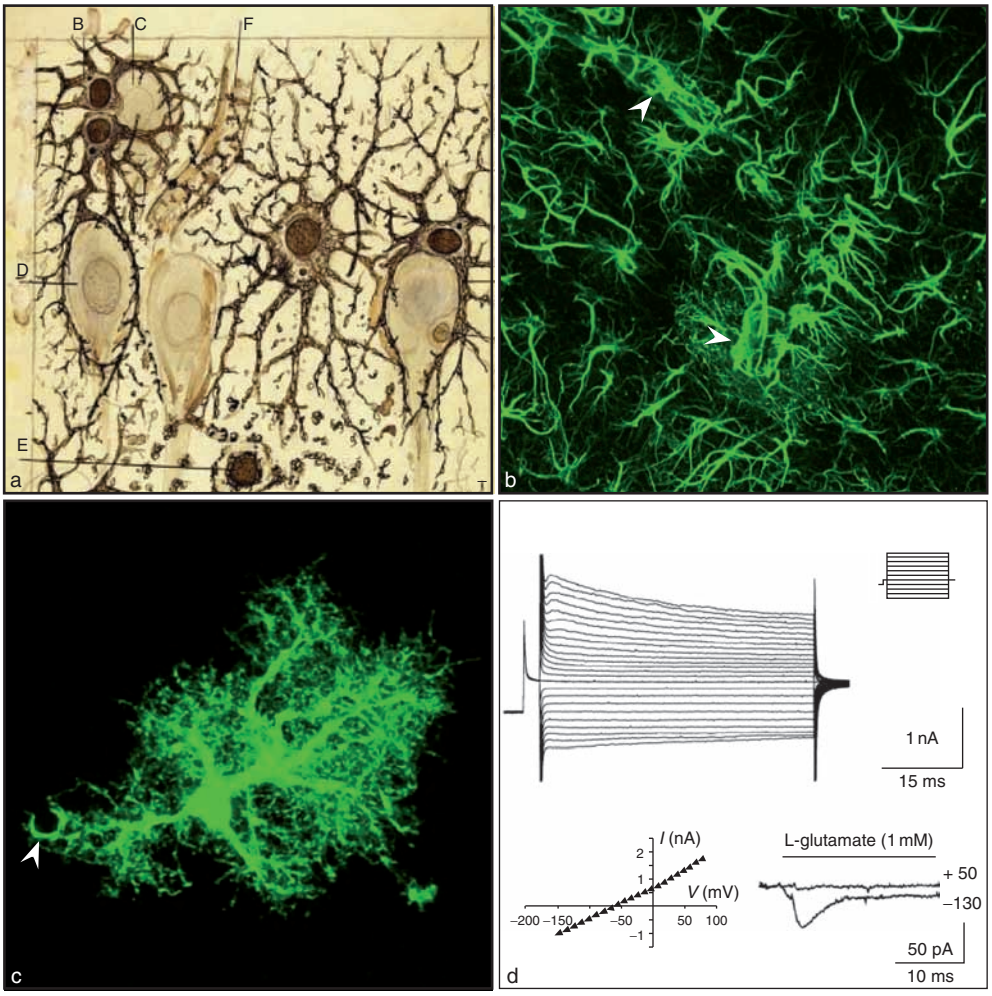

Figure 1 Ways to identify astrocytes. (a) Reproduction of an original drawing of Ramon y Cajal showing gray matter astrocytes visualized by the silver impregnation technique developed by Camillo Golgi (original deposited in the Cajal Institute, CSIC, Madrid). The major astroglial processes intimately enwrap neuronal cell bodies, whereas others contact brain capillaries. (b) Immunostaining of the astroglia-specific intermediate filament protein GFAP recorded by confocal laser scanning microscopy. The arrowheads indicate blood vessels on which GFAP-rich astroglial endfeet impinge. In contrast to the Golgi stain, only GFAP-rich cytoskeletal structures can be revealed. (c) Two-photon laser scanning micrograph of a single cortical astrocyte expressing the enhanced green fluorescent protein in a TgN(hGFAP-EGFP) transgenic mouse. Due to the cytosolic expression of EGFP, very fine details of the complete three-dimensional structure can be recognized. It is evident that astrocytes are highly polarized. Whereas some processes contact the brain capillaries (arrowhead), others are highly branched, forming an extended network of fine membrane protrusions which enwrap neuronal synapses. (d) Electrophysiological characterization of astroglial membrane properties by whole-cell patch clamp recordings. Astrocytes are characterized by a high potassium conductance with a linear current-to-voltage relationship seen when a series of de- and hyperpolarizing voltage steps are applied. In addition, transporter currents can be identified. EGFP, enhanced green fluorescent protein; GFAP, glial fibrillary acidic protein; hGFAP, human glial fibrillary acidic protein. (d) Adapted from Matthias K, Kirchhoff F, Seifert G, et al. (2003) Segregated expression of AMPA-type glutamate receptors and glutamate transporters defines distinct astrocyte populations in the mouse hippocampus. *Journal of Neuroscience* 23: 1750–1758.

Bergmann glia change their morphology with aging, although its formation was not delayed. The authors not only recognized the value of GFAP antibodies but also concluded that mouse strains with neurological mutations are valuable models for neuronal interactions and for the analysis of astroglial responses in neuronal degeneration. Currently, GFAP antibodies are also used for the diagnosis of astroglial tumors and the analysis of astroglia development and differentiation. Since its expression is strongly upregulated upon CNS insults such as traumatic brain injury or neuronal degeneration as in Alzheimer's disease, GFAP is not only the most common marker for gliosis in all types of brain pathology but also a bona fide marker of astrocytes in all vertebrate species from zebra fish to humans. However, *in situ*, particularly in some brain regions such as the cortex, the level of GFAP is significantly downregulated. Numerous

astrocytes appear GFAP negative, although other labeling techniques clearly reveal their abundant presence. Interestingly, some astrocytes, those that contact brain capillaries or touch the pia, display a significant lesser reduction in GFAP expression.

S100β and Glutamine Synthetase as Cytosolic Markers of Astrocytes

Another important marker for astrocytes is S100β. This 11 kDa protein belongs to the superfamily of S100 proteins from which 20 human genes are known. S100 proteins can form calcium-binding, homo- and heterodimer complexes. In addition, they display a remarkable heterogeneous tissue distribution. In mammals, S100β is found predominantly in the brain. It has been implicated in stimulation of neurite outgrowth, neuronal survival during development, and neuroinflammation. S100β is a soluble protein which is expressed throughout the astroglial cytoplasm, but it can also be released into the extracellular space to act as a neurotrophic factor. Due to its cellular localization, antibodies directed against S100β reveal the complete three-dimensional structure of astrocytes, whereas GFAP immunostainings primarily visualize only the major processes. Unfortunately, the expression of S100β is less selective than GFAP. In some areas of the brain, neurons or oligodendrocytes can also be labeled by S100β antibodies.

Similar to S100β, glutamine synthetase (GS) of the brain is located mainly in astroglia. Its primary role is the detoxification of ammonia by generation of glutamine from glutamate. In addition, it serves the metabolism of amino acid neurotransmitters, γ-aminobutyric acid (GABA) and glutamate. The distribution of glial glutamate receptors and glial glutamate transporters parallel to expression of glutamine synthetase suggests a functional coupling between glutamate uptake and the metabolic regulation. Glutamine is released into the extracellular space and is used by neurons as glutamate precursor. GS is considered to be a marker enzyme for astroglia in general. In the adult brain, glutamine synthetase immunoreactivity is observed on astroglial processes surrounding excitatory synapses of most CNS regions. It is also expressed in Müller cells of the retina and embryonic radial glia. However, GS expression has been documented for oligodendroglial subpopulations as well.

Glutamate Transporters GLAST and GLT-1: The Most Abundant Membrane Proteins of Astrocytes

L-Glutamate is the most important excitatory neurotransmitter in the mammalian nervous system. High-affinity glutamate transporters in the plasma membrane terminate the neuronal transmission by removing glutamate from the extracellular space. This mechanism is essential to keep the glutamate concentration in the synaptic cleft below neurotoxic levels. By controlling extracellular glutamate during development, glutamate transporters participate in the regulation of cell differentiation, migration, and synaptogenesis. GLAST (EAAT1) and GLT-1 (EAAT2) are the two astroglial transporters, whereas EAAT3–EAAT5 are primarily expressed on neurons. Although both GLAST and GLT-1 are expressed in almost all brain regions and even in the same astrocyte, GLAST is primarily expressed in Bergmann glia of the cerebellum and GLT-1 predominates in cortex and hippocampus. High-affinity antibodies directed against GLAST and GLT-1 have been instrumental not only to visualize the ultrastructure of astroglial membranes but also to correlate the expression density with uptake currents observed in electrophysiological studies.

Transgenic Labeling of Astrocytes

Advances in molecular biology made provided the possibility to generate transgenic mice in which astrocytes could be labeled not only selectively but also permanently in subsequent generations. Artificial minigenes can be composed of astroglia-selective promoters and reporter genes such as β-galactosidase or the green fluorescent protein. These expression cassettes are then used for transgenesis by injection into fertilized mouse oocytes.

After isolation of the mouse astroglia-specific GFAP gene, in 1991 a 15 kb-long transgene was constructed which included all introns and 2 kb of 5′-flanking and 1.4 kb of 3′-flanking DNA. The *Escherichia coli* β-galacosidase gene (lacZ) was inserted into the first exon as ectopic reporter of the promoter activity. With this construct, transgenic mice were generated by nonhomologous gene recombination. Incubation of brain slices with the β-galacosidase substrate 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) revealed abundant reporter gene expression, primarily in astrocytes. However, variable levels of reporter activity were also observed in neurons. Subsequent *in vitro* studies identified regulatory elements which are essential for astroglial expression. Interestingly, using a 2.2 kb 5′ sequence of the human GFAP promoter for transgenesis revealed a more reliable labeling of astrocytes at high expression levels. Reporter activity of transgenic lines with this promoter mimicked to a large degree the endogenous GFAP levels. The transgenes were expressed beginning on embryonic days 12.5–13.5 and continued to be expressed until adulthood. Also, acute brain lesions resulted in increased reporter expression in parallel to normal gliotic scar

reactions. However, not all astrocytes are labeled by transgenic reporter gene expression. Similar to the normal mouse GFAP levels, transgene expression is most abundant in young mice – in cerebellar Bergmann glia and in astrocytes which contact either pia, the capillaries, or the ventricular walls. The downregulation in cortical brain areas is most dramatic and not understood.

Approximately 10 years ago, transgenic mice were generated in which the green fluorescent proteins S65T-GFP or later EGFP were used to label astrocytes. The unique advantage of fluorescent proteins is buried in their structure, in which the fluorophor is spontaneously generated from three residues of the polypeptide chain (by cyclization and oxidation). No exogenous cofactors are needed as labeling dyes. GFAP–GFP transgenic mice revolutionized the identification of astrocytes in brain tissue. For the first time, it was possible to unequivocally identify astrocytes in all brain regions without prior immunohistochemical labeling, which normally requires tissue fixation. Live astrocytes could be analyzed in acutely isolated brain slices or even *in vivo*. Via fluorescence-activated cell sorting, fluorescent astrocytes could be isolated und subsequently analyzed in cell culture. Using this approach, the neurogenic potential of embryonic radial glia cells was revealed. In addition, acutely isolated tissue slices isolated from brain regions such as the brain stem or hippocampus served for the analysis of the dynamic interaction of astroglial processes with neighboring active synaptic terminals. Time-lapse recordings demonstrated highly motile lamellipodia and filopodia emanating from major process branches. This work strengthened the view of astrocytes as regulators of neuronal transmission. However, astrocytes do not just modulate neuronal function. *In vivo*, the pivotal role of astroglial endfeet associated with neighboring capillaries could be demonstrated.

Transgenic labeling with the human GFAP promoter has been very helpful in identifying cells with astroglial properties in other species such as the zebra fish. Another important approach to label astrocytes uses viral infections. Baculo-, lenti-, adeno-, and adeno-associated viruses carrying GFAP promoter elements have been used to deliver genes of interest into astrocytes. Depending on brain region and tropism of the viral substrain, even preferential infection of astrocytes can be achieved.

Promoter elements of other astroglial genes such as S100β or the gap junction protein connexin 43 have been used successfully to generate transgenic mice. In addition, a series of mice were derived from nonhomologous recombination using bacterial artificial chromosomes carrying large genomic regions of astroglial genes such as the glutamate transporters GLAST and GLT-1.

Interestingly, Fritjof Helmchen and colleagues discovered that the common red fluorescent dye sulforhodamine 101 (SR101; also known as Texas Red) can be used to label astrocytes nontransgenically. Injection of SR101 into the cortical parenchyma *in vivo* or incubation of acutely isolated brain slices in SR101 lead to widespread labeling of all astrocytes in a given area. After uptake, the dye is distributed via gap junctions within the astroglial syncytium. Similarly, green fluorescent calcium indicator dyes such as Oregon Green or Fluo4 have been used to unspecifically label all brain cells simultaneously. Currently, costaining with SR101 allows the analysis of neuronal and astroglial network activities simultaneously. Whereas neurons are stained only by the fluorescent calcium indicator, the astrocytes are highlighted by SR101. An additional advantage of SR101 labeling is that it is species -independent. SR101 has successfully been used to label astrocytes in CNS tissues of rodents, as well as primates.

Inducible genetic labeling of astrocytes has been developed. Astroglial expression of the fusion protein of the Cre DNA recombinase and a modified ligand-binding domain of the human estrogen receptor (CreERT2) was induced either by the GFAP promoter (TgN(GFAP-CreERT2)) or by homologous recombination into the GLAST gene (TgH(GLAST-CreERT2)). Such mice must then be cross-bred to mice which carry a reporter cassette consisting of an ubiquitous promoter, a loxP-flanked stop sequence, and a reporter gene such as lacZ or GFP. Temporal control of astroglial labeling can be achieved by intraperitoneal injection of the estrogen receptor antagonist tamoxifen into these double-transgenic mice. After tamoxifen-induced gene recombination, all CreERT2-positive cells and their progeny are permanently labeled. Such experiments have been performed to investigate the role of astroglial cells as putative neural stem cells during development or disease progression.

Electrophysiological Membrane Properties

Another method to identify astrocytes is based on their electrophysiological properties since they express specific membrane currents composed of ion channels, transporters, and ionotropic receptors. Identification of astrocytes according to their membrane properties was first achieved by Kuffler and colleagues in the 1960s in the amphibian optic nerve and later verified by many other groups in the mammalian

CNS. Mature astrocytes express high levels of symmetrical in- and outwardly rectifying potassium currents with a linear current-to-voltage relationship. The astroglial resting membrane potential is usually very negative (−80 mV or more).

Beside imaging, the most common approach to study the physiological properties of astrocytes is the patch clamp technique. In the early 1990s, patch clamp recordings of astrocytes relied either on a unique localization of astroglial cell populations in defined brain regions (e.g., Bergmann glia in the cerebellum) or on postrecording labeling techniques. Both approaches have major drawbacks. Phase or differential interference contrast does not always allow the distinction of astrocytes from neighboring neurons in acutely isolated brain slices, particularly from interneurons, which very often have similar cell somata. Subsequent cell identification is often hampered by distortion of the cell integrity after patch pipette retraction. The availability of transgenic mice with astroglial GFP expression not only facilitated the physiological analysis but also contributed significantly to define a set of unique membrane properties for astrocytes. In this way, astroglial subpopulations could be distinguished in various brain regions. Patch clamp recordings in GFAP–GFP transgenic mice confirmed the findings of Kuffler and others: a dominance of the membrane conductance by high potassium currents with a linear current-to-voltage relationship and negative membrane potentials of more than −80 mV. In addition, they abundantly express glutamate transporters and are coupled within a syncytial network. Other astroglial cells are characterized by expression of voltage-gated sodium and outwardly rectifying potassium currents. Studies performed on almost all brain regions provide a complex picture of the astrocyte. Although in the past astrocytes were regarded as passive elements of the nervous system acting plainly as sink for potassium or the excitatory transmitter glutamate, AMPA- and NMDA-type glutamate receptors have been found on astroglial cells, as well as GABA receptors and transporters. However, not all receptors, transporters, and ion channels are found on all astrocytes. The current picture reveals astrocytes as a class of cells with some common characteristics and many heterogeneous properties which can be developmentally regulated and brain region dependent.

During patch clamp recordings, astrocytes can also be visualized by simultaneous diffusion of fluorescent dyes from the pipette solution into the cell. Since the majority of astrocytes are coupled via gap junctions consisting primarily of the connexin proteins cx43 and cx30, the complete astroglial syncytium can be revealed.

Conclusion

Since their first description in the nineteenth century, the functional characterization of astrocytes has depended significantly on the development of novel labeling techniques. The current techniques allow the unequivocal identification of astrocytes under almost all experimental conditions: in tissue culture, in postmortem brain sections, or even *in vivo*. We can now identify living astrocytes using transgenic mice with fluorescent protein expression or using astroglia-selective vital dye staining. In addition, we can study the plasticity of astrocytes *in vivo* using advanced imaging techniques such as two-photon laser scanning microscopy. These studies are complemented by the analysis of their signaling molecule repertoire of receptors and transporters at the single-cell level using the whole-cell patch clamp technique. Finally, we can now selectively control gene activity in astrocytes using the inducible Cre/loxP system. This latter approach will help us to understand the impact of a given gene expressed in astrocytes for the living animal.

See also: Astrocyte: Neurotransmitter and Hormone Receptors; Calcium Homeostasis in Glia; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Intermediate Filaments; Neurotransmitter Release from Astrocytes; Potassium Homeostasis in Glia.

Further Reading

Brenner M, Kisseberth WC, Su Y, Besnard F, and Messing A (1994) GFAP promoter directs astrocyte-specific expression in transgenic mice. *Journal of Neuroscience* 14: 1030–1037.

Halassa MM, Fellin T, and Haydon PG (2007) The tripartite synapse: Roles for gliotransmission in health and disease. *Trends in Molecular Medicine* 13: 54–63.

Hirrlinger PG, Scheller A, Braun C, Hirrlinger J, and Kirchhoff F (2006) Temporal control of gene recombination in astrocytes by transgenic expression of the tamoxifen-inducible DNA recombinase variant CreERT2. *Glia* 54: 11–20.

Kuffler SW, Nicholls JG, and Orkand RK (1966) Physiological properties of glial cells in the central nervous system of amphibia. *Journal of Neurophysiology* 29: 768–787.

Lee Y, Su M, Messing A, and Brenner M (2006) Astrocyte heterogeneity revealed by expression of a GFAP-LacZ transgene. *Glia* 53: 677–687.

Matthias K, Kirchhoff F, Seifert G, et al. (2003) Segregated expression of AMPA-type glutamate receptors and glutamate transporters defines distinct astrocyte populations in the mouse hippocampus. *Journal of Neuroscience* 23: 1750–1758.

Misgeld T and Kerschensteiner M (2006) *In vivo* imaging of the diseased nervous system. *Nature Reviews Neuroscience* 7: 449–463.

Nimmerjahn A, Kirchhoff F, Kerr JN, and Helmchen F (2004) Sulforhodamine 101 as a specific marker of astroglia in the neocortex *in vivo*. *Nature Methods* 1: 31–37.

Palay SL, Peters A, and Webster H (1991) *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*. New York: Oxford University Press.

Somjen GG (1988) Nervenkitt: Notes on the history of the concept of neuroglia. *Glia* 1: 2–9.

Astrocyte: Neurotransmitter and Hormone Receptors

C Steinhäuser, University of Bonn, Bonn, Germany
H Kettenmann, Max Delbrück Center for Molecular Medicine, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Background

Until about 20 years ago, the expression of neurotransmitter receptors in the nervous system was considered an exclusive property of neurons. Astrocytes were viewed as electrically silent elements which would not participate in the information processing of the central nervous system (CNS). This view was challenged by a series of studies in cell culture demonstrating that astrocytes had the potential to express transmitter receptors. In the culture dish, astrocytes seemed to express almost all receptors for important transmitter systems. Two exceptions remained: *N*-methyl-D-aspartate (NMDA) and glycine receptors. In the early 1990s, the studies were extended to acute brain slices, and it became evident that astrocytes expressed functional receptors in this more intact preparation. It is interesting to note that the two receptor types lacking in culture, NMDA and glycine receptors, were now found to be functionally expressed by some glial cells from defined regions of the CNS. This article describes the receptor repertoire expressed by astroglial cells. However, a basic limitation has to be mentioned here. It is becoming clear that different types of cells with astroglial properties coexist within a given brain region and that the properties of these cells vary in different areas. So far, we have only rudimentary understanding of this glial diversity, and most of the previous studies describing astroglial properties did not identify the specific cell type. This article refers to different types of cells with astroglial properties as astrocytes.

Ionotropic Glutamate Receptors

Functional α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors have been described in astrocytes *in situ* in various brain regions. In the hippocampus, two types of cells with astroglial properties were shown to coexist, with one type expressing and the other lacking AMPA receptors. While in most brain regions astrocyte AMPA receptors display a low Ca^{2+} permeability, Bergmann glial cells in the cerebellum lack the glutamate receptor 2 (GluR2) subunit and allow passage of divalent cations. The physiological relevance of this subunit combination has been demonstrated by experimentally inducing expression of GluR2 in Bergmann glial cells, which results in neuronal rearrangement. The pharmacological properties and single channel conductance of astroglial AMPA receptors mimic those of their neuronal counterparts. In the hippocampus, astroglial receptors possess an intermediate Ca^{2+} permeability and are primarily assembled from the subunits GluR1, GluR2, and GluR4. Receptor activation leads to reversible blockage of K^+ channels, through intracellular accumulation of Na^+. Early in postnatal development, astrocytes co-express AMPA receptors with variable Ca^{2+} permeability while later on, only receptors with low Ca^{2+} permeability, carrying GluR2, are formed. In addition, changes in receptor splicing occur within the first 3 postnatal weeks: in contrast to hippocampal neurons, the portion of flip splice variants increases in astrocytes, resulting in prolonged receptor opening. Joint stimulation of AMPA and metabotropic glutamate receptors (mGluRs) in hippocampal and neocortical astrocytes triggers Ca^{2+} oscillations and mediates glutamate release. Non-NMDA receptor activation in astrocytes was reported to trigger ATP release from these cells, which in turn caused homo- and heterosynaptic suppression in CA1 neurons of the hippocampus. Recently, cells with astroglial properties in the hippocampus were shown to receive direct synaptic input from glutamatergic neurons. In these glial cells, nerve fiber stimulation led to the activation of AMPA receptor-mediated postsynaptic currents. Fiber stimulation in the cerebellum leads to ectopic presynaptic release of glutamate, which rapidly activates AMPA receptors in Bergmann glial cells. While the physiological impact of this rapid form of neuron-to-glia signaling is poorly understood, increasing evidence suggests that dysfunctional astroglial AMPA receptors are involved in the pathogenesis of neurological disorders.

Clear evidence for a functional expression of kainate receptors in astrocytes is lacking. Transcripts of all five kainate receptor subunits, GluR5-7, KA1, and KA2, have been found in adult bovine white matter astrocytes *in situ*. Antibodies against GluR5-7, GluR6,7, and KA2 confirmed the expression of kainate receptor protein, but functional receptors have yet to be identified.

Some astrocytes may express functional NMDA receptors. Bergmann glial cells express messenger RNAs (mRNAs) for NR2B, and a physiological study in acute brain slices reported NMDA-induced membrane currents in Bergmann glial cells. However, these responses displayed unusual properties (e.g., no Mg^{2+} and glycine sensitivity) and indirect effects cannot be excluded. Evidence for the presence of functional

NMDA receptors *in situ* comes from cortical astrocytes. In these cells, astroglial NMDA responses were observed that resembled NMDA response in neurons, including sensitivity to extracellular Mg^{2+} and blockade by MK-801. In contrast, another study described Mg^{2+}-insensitive NMDA receptors in cortical astrocytes. Moreover, activation of NMDA receptor currents were observed in hippocampal astrocytes after ischemia *in vivo*, due to upregulation of the NR2B subunit.

mGluRs

The mGluR family consists of eight members which couple to G-proteins and can be classified into three groups: group I comprises mGluR1 and mGluR5, and their activation leads to stimulation of phospholipase C (PLC), increase in inositol triphosphate (IP_3), and release of Ca^{2+} from internal stores, and group II (mGluR2, mGluR3) and group III (mGluR4, mGluR6–8) receptors couple to adenylate cyclase. mGluR3 and mGluR5 are the predominant subtypes expressed by astrocytes. mGluR1 has been reported in hippocampal astrocytes and in the spinal cord. Accordingly, activation of these receptors led to an increase in intracellular Ca^{2+} and inhibition of cyclic adenosine monophosphate (cAMP) accumulation, although Gs-coupled cAMP stimulation has also been reported. Stimulation of astroglial mGluRs leads to intracellular Ca^{2+} oscillations and Ca^{2+} wave propagation within the astrocyte network, activates Ca^{2+}-dependent K^+ channels, and induces prostaglandin-mediated glutamate release from astrocytes that activates neuronal receptors. These responses are likely to occur under physiological conditions because astroglial mGluR activation and subsequent Ca^{2+} responses could be evoked by electrical stimulation of fiber tracts in acute brain slices and through sensory stimulation *in vivo*. Activation of mGluRs induced other astrocyte responses, including swelling, activation of phospholipase D and glutamine synthetase, release of arachidonic acid, cAMP-dependent block of K^+ currents, modulation of proliferation, and regulation of the expression of the glutamate aspartate transporter (GLAST) (**Figure 1**). Moreover, mGluR-induced intracellular Ca^{2+} elevations in astrocytes are involved in the regulation of cerebral blood flow. Together, these findings suggest that mGluRs play a pervasive role in modulating astrocyte function and intercellular communication in the CNS. Dramatic changes in astrocyte mGluR expression occur after spinal cord injury and in neurological disorders.

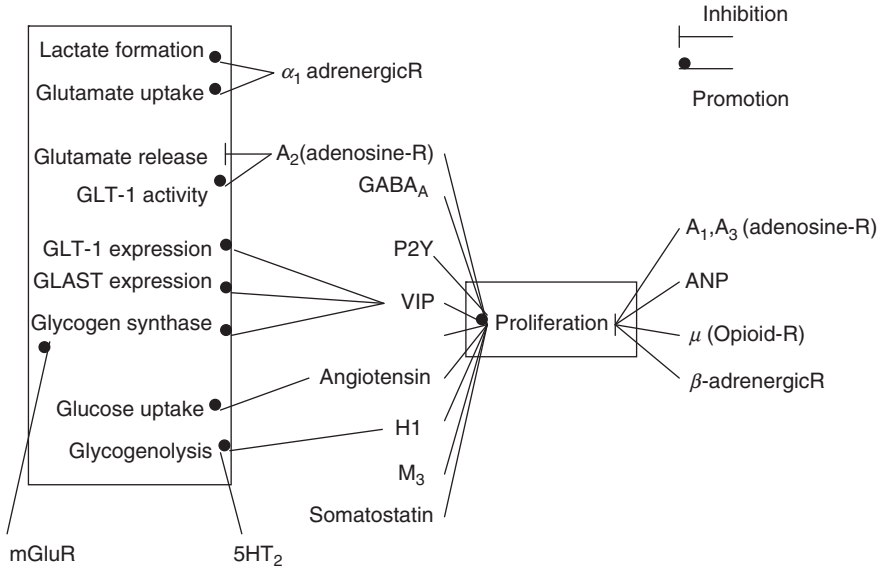

Figure 1 Astrocyte receptors linked to proliferation and glutamate/glucose metabolism. This figure summarizes the influence of different astrocytic receptors on glutamate and glucose metabolism and transport and combines this information with the impact of receptors on proliferation. It becomes evident that only proliferation-promoting receptors also control glutamate or glucose metabolism and transport, in contrast to receptors which inhibit astrocyte proliferation. It is evident that only proliferation-promoting receptors also control glutamate or glucose metabolism and transport, in contrast to receptors that inhibit astrocyte proliferation. A_1, A_2, and A_3, adenosine receptors; α_1 adrenergicR, α_1 adrenergic receptor; ANP, atrial natriuretic peptide; β-adrenergicR, β adrenergic receptor; GLAST, glutamate aspartate transporter; $GABA_A$, γ-aminobutyric $acid_A$; GLT-1, glutamate transporter 1; H_1, histamine receptor; $5\text{-}HT_2$, 5-hydroxytryptamine-2; M_3, muscarinic acetylcholine receptor; mGluR, metabotropic glutamate receptor; P2Y, a type of purinergic receptor; R, receptor; VIP, vasoactive intestinal polypeptide.

γ-Aminobutyric Acid Receptors

Ionotropic γ-aminobutyric acid ($GABA_A$) receptors are abundantly expressed by astrocytes. *In situ*, functional $GABA_A$ receptors have been demonstrated in Bergmann glial cells, astrocytes of the spinal cord, optic nerve, retina, hippocampus, and pituitary gland. Astrocytic $GABA_A$ receptors form Cl^- channels with a conductance of about 30 pS. In contrast to mature neurons, $GABA_A$ receptor activation leads to depolarization in astrocytes. Indeed, local stimulation of interneurons in the hippocampus, which form functional synapses with astroglial cells, led to $GABA_A$ receptor-mediated glial depolarization. In astrocytes of the pituitary gland, postsynaptic potentials were activated by neuronally released GABA. Besides opening Cl^- channels, $GABA_A$ receptor activation in astrocytes triggers long-lasting blockade of K^+ conductances, thereby augmenting depolarization and activating voltage-gated Ca^{2+} channels with subsequent cytosolic Ca^{2+} increase.

Bergmann glial cells in cerebellar slices showed GABA responses which were benzodiazepine-insensitive. Immunocytochemistry revealed that these cells lack the γ subunit; they express instead the δ subunit. In contrast, hippocampal astrocytes are benzodiazepine-sensitive, indicating functional heterogeneity among astrocyte subtypes. In Bergmann glial cells, $GABA_A$ receptors are downregulated during postnatal maturation. In these cells, immunolabeling preferentially identified receptors along the processes. There is evidence that $GABA_A$ receptor activation promotes differentiation of astrocytes. In culture, GABA triggers formation of processes and leads to a more complex morphological shape of the astrocytes. A similar effect has been observed in the hypothalamus *in situ*: GABA released from neurons activated astroglial $GABA_A$ receptors and induced differentiation.

$GABA_B$ Receptors

The most convincing demonstration of functional expression of astroglial $GABA_B$ receptors was obtained in acute hippocampal slices. Here, these receptors mediated intracellular Ca^{2+} elevations that were accomplished by GABA release from inhibitory interneurons. This indicates that astrocytes *in situ* can sense inhibitory neuronal activity via $GABA_B$ receptors. It has also been demonstrated that $GABA_B$ receptor activation in astrocytes has an impact on neuronal function by potentiating inhibitory postsynaptic currents in pyramidal neurons. Other studies reported that astroglial $GABA_B$ receptor activation mediates heterosynaptic neuronal depression. Thus, in the hippocampus, astrocytic $GABA_B$ receptors are part of a functional circuit involving inhibitory neurons, astrocytes, and pyramidal neurons.

Purinergic Receptors

Astrocytes express a large number of purinergic receptor subtypes linked to different effector mechanisms. The presence of adenosine receptors *in situ* has been established in the hippocampus and in acutely isolated cortical astrocytes. In both preparations, activation of adenosine receptors led to the accumulation of intracellular Ca^{2+} via release from cytoplasmic stores. All three types of adenosine receptors (A_1, A_2, A_3) can be expressed by astrocytes. A_2 receptor activation increased internal cAMP concentration and stimulated astroglial proliferation, whereas stimulation of A_1 and A_3 receptors had opposite effects. A_1 receptor activation stimulated the production of neurotrophic factors (nerve growth factor (NGF), S100β, transforming growth factor-β (TGF-β)) and enhanced mGluR5-induced intracellular Ca^{2+} responses. Receptor subtypes can have opposing actions: while A_1 receptor stimulation inhibited purine release, A_2 receptor activation stimulated release *in situ*. A_3 receptor stimulation not only inhibits proliferation but also leads to apoptosis and induces reorganization of the astroglial cytoskeleton. Thus, adenosine triggers a complex response in astrocytes which can lead either to proliferation or to apoptosis, depending on the expression levels of A_1 and A_3 receptors, respectively. A_{2a} receptor signaling is linked to the control of glutamate levels in the extracellular space: Receptor activation inhibited glutamate transporter 1 (GLT-1)-mediated glutamate uptake and facilitated glutamate release, leading to increased extracellular levels of the transmitter.

Both types of ATP receptors, P2X and P2Y, have been identified in astrocytes. In cell culture, activation of P2X receptors leads to the opening of a cationic conductance. So far, however, P2X receptor-mediated cationic conductances have not been observed in astrocytes *in situ* despite positive immunolabeling for various P2X subunits. Only in Müller cells of the human retina have $P2X_7$-induced currents been recorded, a property that seems to be species specific. Here, $P2X_7$ receptor activation triggered Ca^{2+}-activated K^+ (BK) channels which might be involved in $P2X_7$-dependent regulation of proliferation.

P2Y-mediated signaling has been identified in Bergmann glial cells in acute slices of cerebellum, in freshly isolated astrocytes from cortex and hippocampus, and in Müller cells and astrocytes from the retina. P2Y receptor activation leads to a G_q-PLC mediated increase in IP_3 and Ca^{2+} release from intracellular stores. Moreover, these receptors are coupled

to pertussis toxin (PTX)-sensitive G_i/G_o proteins linked to the synthesis of phospholipase A2, arachidonic acid, and protein kinase C (PKC), extracellular signal-regulated kinase, cyclooxygenase 2, release of arachidonic acid and prostaglandin E2, stimulation of mitogen-activated protein kinases (MAPKs), and induction of immediate early genes. Astrocytes of the postnatal hippocampus express the $P2Y_1$ and $P2Y_2$ receptor subtypes, with $P2Y_2$ increasing during development. P2Y receptor activation is also involved in the induction of astrogliosis *in vivo* since lesion-induced astrogliosis could be blocked by P2Y receptor antagonists.

Glycine Receptors

Functional glycine receptors have been detected in astrocytes from spinal cord slices. A combined patch clamp reverse transcription polymerase chain reaction (RT-PCR) approach revealed the expression of $\alpha 1$ and β subunits. As in neurons, glycine activated a Cl^- conductance. It is interesting to note that glycine receptor expression has not been reported for cultured astrocytes and is also not a common property of astrocytes *in situ*: Bergmann glial cells, for instance, do not express functional glycine receptors.

Acetylcholine Receptors

α-Bungarotoxin-sensitive nicotinic acetylcholine (ACh) receptors containing the $\alpha 7$ subunit were described for cultured astrocytes. Receptor activation led to Ca^{2+} influx through the receptor channels and Ca^{2+}-induced Ca^{2+}-release from caffeine-sensitive stores. This subunit was also localized to astrocytes of human cerebellum *in situ*. This receptor was upregulated in astrocytes from the brains of Alzheimer's patients, suggesting a role in the pathogenesis of this disease.

The expression of muscarinic ACh (mACh) receptors by astrocytes is well established. There is evidence for the expression of transcripts for the mACh receptor subtypes M_1–M_5 in cultured astrocytes. Acetylcholine induces IP_3 formation, release of Ca^{2+} from cytoplasmic stores and inhibits adenylate cyclase. As tested in culture, acetylcholine induced an increase in astroglial proliferation due to activation of M_3 receptors and phosphorylation of MAPK. Neuronal activity can stimulate astrocytic mACh receptors: stimulating cholinergic fibers projecting from the septum to the hippocampus triggered Ca^{2+} responses in hippocampal astrocytes. Excessive neuronal activity leads to upregulation of mACh receptors *in vivo*, as has been shown in amygdala kindling, indicating that neuronal activity can regulate astrocyte receptor expression. During development, mACh receptor responses and receptor protein expression is upregulated.

Adrenergic Receptors

Both α- and β-adrenergic receptors have been described in astrocytes in culture and *in situ*. The α_1 receptors trigger the release of Ca^{2+} from internal stores. They affect astrocyte function by inhibiting gap junction coupling via PLC. This is in line with the finding that mechanically induced Ca^{2+} waves in hippocampal astroglial cells are inhibited after stimulation of α_1 receptors. Moreover, α_{1b} receptor activation stimulated glutamate uptake in cultured astrocytes. There is also a hint for the control of metabolic activity by α_1 receptors because receptor activation induced lactate formation. The α_1 receptors regulate expression of α_2 receptors and thus form a complex feedback circuit. Both α_1 and α_2 receptors have been identified *in situ*; α_1 receptors are expressed postnatally in the hippocampus (after postnatal day 8 in rat). In the hippocampus, α_2 receptors are located on astroglial processes, near terminals forming asymmetric excitatory synapses. Neuronal stimulation in acute cerebellar slices led to the release of noradrenaline from afferent fibers and activation of α_1 receptors in Bergmann glial cells.

Both subtypes of β receptors, β_1 and β_2, are expressed by astrocytes *in situ*, with β_2 being more prominently expressed under pathological conditions. In astrocytes of rat visual cortex, β receptors appear after the second postnatal week. In the injured brain, adrenergic receptors are differentially regulated. While α_1 receptors decrease in areas of neuronal degeneration and gliosis, β receptors are upregulated, as evidenced by binding assays. The change in β receptor density and astrogliosis seems to be functionally linked. Blockade of β receptors suppressed glial scar formation, indicating that adrenergic receptors are part of the cascade leading to astrocyte activation. This suggestion is substantiated by the finding that β_2 receptors are activated or upregulated in the transected optic nerve *in vivo*, confirming that β-adrenergic signaling is an important feature of the astrocyte response to injury. Further support comes from the observation that stimulation of β receptors leads to astrogliosis and cell proliferation in the optic nerve *in vivo*. This effect might be mediated via the control of growth factor expression (e.g., fibroblast growth factors 1and 2, BDNF, and ciliary neurotrophic factor (CNTF)). In contrast, an inhibition of proliferation on β-adrenergic stimulation was observed in cultured astrocytes, which, however, does not contradict gliosis. The involvement of β receptors in pathologic processes is further substantiated by the finding that their stimulation leads to production of amyloid precursor protein, changes in morphology, and increase in glial fibrillary acidic protein (GFAP).

Stimulation of β_1 receptors is also accompanied by enhanced glycogen levels which could be part of the gliotic response. The observation that β adrenergic receptor stimulation leads to a cAMP-mediated inhibition of astroglial inwardly rectifying K^+ channels points to a role in the regulation of the extracellular K^+ homeostasis.

Dopamine Receptors

D1 receptor activation leads to a PTX-sensitive cAMP formation via protein kinase A stimulation. Prolonged receptor activation led to a reduction of astroglial dopamine sensitivity. D2 receptors in astrocytes are linked to the actin cytoskeleton via interaction with filamin A. A recent study provided first evidence for the presence of dopamine receptors on astrocytes *in vivo*. Ligand binding and ultrastructural analysis found strong expression of D2 receptors in GFAP-positive cortical astrocyte processes surrounding interneurons. These receptors were indeed functional since their activation induced intracellular Ca^{2+} elevations in the astrocytes.

Histamine Receptors

In cultured astrocytes, binding sites for H_1 and H_2 subtypes of histamine receptors have been identified which couple to G_q and G_s types of G-proteins, respectively. Accordingly, H_1 receptors couple to PLC, leading to IP_3 production and Ca^{2+} release while H_2 receptor activation in astrocytes is linked to adenylate cyclase and intracellular cAMP accumulation. H_1 receptors were mainly found on the processes of astrocytes, and stimulation led to locally restricted Ca^{2+} responses. Activation of histamine H_1-receptor enhances neurotrophic factor secretion from cultured astrocytes. Histamine also acts as a mitogen in cortical astrocytes. Activation of H_1 receptors has been shown to stimulate glycogen breakdown. Although *in situ* hybridization has not yet unequivocally associated histamine receptors with astrocytes, pharmacological studies indicate their expression in cerebellum and hippocampus. H_1 receptor activation triggered an increase in $Ca^{2+}{}_i$ due to release from thapsigargin-sensitive intracellular pools. Based on ultrastructural studies suggesting glial apposition to histaminergic neurons, astrocytes were supposed to respond to histamine released from hippocampal neurons.

Serotonin Receptors

RT-PCR in human and rat brain cultures revealed that astrocytes can express various serotonin receptors. In cell culture, stimulation of 5-hydroxytryptamine $(5HT)_2$ led to phosphatidylinositol (PI) hydrolysis, cAMP accumulation, and upregulation of glycogenolysis. A distinct expression of 5HT subtypes also occurs *in situ* since the subtypes $5HT_{1A}$, $5HT_{2A}$, and $5HT_{5A}$ have been identified in different brain areas. 5HT receptors have been speculated to be involved in pathogenesis. The $5HT_{5A}$ subtype is upregulated in gliosis, and the $5HT_{2A}$ subtype is enhanced in schizophrenia. Stimulation of 5HT receptors leads to the release of S100β from astrocytes, and this release has been hypothesized to have an impact on the ongoing pathologic event. $5HT_{5A}$ transcripts are developmentally regulated: Receptor mRNA is detected prior to birth, and expression peaks at postnatal day 20 in the rat.

Angiotensin Receptors

Expression of angiotensin (AT) receptors in astrocytes *in situ* seems to be restricted to white matter. Antibody staining identified AT_1 and AT_2 receptors in astrocytes of white matter tracts in cerebellum and subcortical regions, in the optic nerve, and in the corpus callosum, while no immunoreactivity was found in grey matter. The diversity in expression is also reflected when astrocytes are harvested from different brain areas: AT_1 receptors were found in astrocytes from medulla oblongata and cerebellum, whereas astrocyte cultures from hypothalamus and cortex did not express functional receptors. Activation of astroglial AT receptors stimulates PI hydrolysis and PKC and triggers Ca^{2+} release from internal stores. Receptor activation as studied in culture stimulated proliferation, increased glucose uptake, and induced prostaglandin release.

Atrial Natriuretic Peptide Receptors

All three types of atrial natriuretic peptide (ANP) receptors have been identified in cultured astrocytes: ANP_A and ANP_B (biological receptor), as well as ANP_C (clearance receptor). So far, only the ANP_A receptor has been clearly identified *in situ* (in the brain stem). ANP receptors, without subtype specification, have been found in several other brain structures, including olfactory bulb, hippocampus, and amygdala. In astrocytes, the clearance receptor has a possible biological function: ANP inhibits MAPK via ANP_C. MAPK is stimulated by growth factors such as endothelin-3 (ET-3) or platelet derived growth factor, and thus ANP_C counteracts their activation. This antimitogenic action of ANP in astrocytes is mediated via inhibition of ET-3 induced G-protein activation. In consequence, since these growth factors stimulate

proliferation, ANP acts as an antiproliferative substance on astrocytes. Accordingly, if astrocytes are stimulated with a proliferative agent such as AT_2 or other growth factors, ANP acts as an antagonist to restrict cell number. The specific role of ANP_A and ANP_B receptors is less evident. ANP has been shown to stimulate cyclic guanosine monophosphate production and activates guanylate cyclase in astrocytes.

Bradykinin Receptors

So far, functional bradykinin receptors (B_1 and B_2 receptors) have been identified only in cultured astrocytes. Receptor activation leads to an increase in intracellular Ca^{2+}, induced by release from intracellular stores, and to a blockade of a constitutive K^+ conductance. Pharmacological data indicate that this block is mediated via B_2 receptors. B_2 receptor-induced increase in intracellular Ca^{2+} leads to the release of glutamate and aspartate from astrocytes, which stimulates neuronal glutamate receptors, substantiating that astrocytes can directly modulate neuronal activity.

Endothelin Receptors

Astrocytes in culture and in tissue express endothelin B receptors. Stimulation of these receptors is linked to control of proliferation, in particular in the context of brain injury. There is even evidence that pathologic conditions such as a hypoxic insult increase endothelin B receptor expression. Endothelin stimulates astrocytic neurotrophin-3 and BDNF production. Since astrocytes are also capable of releasing endothelins, an autocrine pathway is possible which might be relevant in pathology.

Opioid Receptors

Early work indicated the presence of opioid receptors on tanycytes and pituicytes. There is evidence for the expression of all three subtypes of the opioid receptors, μ, δ, and κ, both in cell culture and *in situ*. Activation of μ and κ receptors activation DNA synthesis, proliferation, and astroglial growth. It is interesting to note that the inhibitory action of opioids on astrocyte proliferation was restricted to GFAP-positive astrocytes while S100β-positive cells were not affected. Receptors κ and δ are linked to Ca^{2+} signaling, and their stimulation inhibited forskolin-induced cAMP stimulation. Circumstantial evidence indicates that the effect of opioids on astroglial proliferation is Ca^{2+}-dependent. Indeed, application of opioid receptor agonists induced an increase in intracellular Ca^{2+}, but the underlying mechanisms are not yet fully resolved. Receptors μ and δ were shown to be preferentially expressed on astrocytic processes *in situ*. While μ receptors are more frequently present on immature astrocytes, the expression of δ and κ receptors is increased in adult tissue.

Oxytocin and Vasopressin Receptors

Both vasopressin and oxytocin can trigger Ca^{2+} responses in astrocytes, indicating the presence of their receptors. Neuronal production and release of TGF-β led to an increase in astroglial oxytocin receptor mRNA in astrocytes, indicating that neurons were able to upregulate the astrocyte receptor. V_{1A} vasopressin receptors were identified by binding and biochemical studies. There is evidence that vasopressin is important for Cl^- homeostasis and volume regulation of astrocytes. An increased Cl^- uptake was observed on vasopressin application, presumably mediated by V_{1A} receptors.

Somatostatin Receptors

Prominent expression of somatostatin (sst) receptors in astrocytes of hippocampus, amygdala, and hypothalamus *in situ* has been identified in binding studies on brain slices. The subtypes have been more stringently identified in culture, and it is evident that astrocytes express the sst_{2A} splice variant. Receptor activation is linked to inhibition of cAMP accumulation and leads to enhanced proliferation rates. Moreover, it leads to a reduction in forskolin-induced interleukin-6 (IL-6) release.

Tachykinin Receptors

Astrocytes express all subtypes of tachykinin receptors, NK_1 to NK_3. In cell culture, activation of NK_1 receptors leads to membrane depolarization. This depolarization is due to blockade of a constitutive K^+ conductance and opening of Cl^- channels. Patch clamp recordings demonstrate that channel activity is linked to the tachykinin receptor via intracellular pathways. Receptor localization on the light and electron microscopic level in different species, including humans, has identified NK_2 and NK_3 receptors in astrocytes of various areas of the CNS, such as spinal cord, cortex, and hippocampus. NK_2 receptors have been found to cluster close to axon terminals in spinal cord. Substance P, as a prominent ligand, enhances secretion of IL-6 and prostaglandin E_2 on IL-1β stimulation in cultured spinal cord astrocytes. Moreover, lipopolysaccharide-induced secretion of tumor necrosis factor-α and IL-1 from cultured astrocytes was augmented by the ligand known as

substance P. A proinflammatory function of tachykinin receptors in reactive astrocytes has been suggested since a lesion-induced upregulation of the receptors was observed in astrocytes of transected optic nerve *in vivo*.

Thyrotropin-Releasing Hormone Receptors

Thyrotropin-releasing hormone (TRH) receptors have been identified in culture and *in situ* by immunocytochemistry and Western Blot. Astrocytes express the TRH_1 subtype. In the adult rat spinal cord, expression *in situ* is stronger in white than in grey matter. The receptors are present in TRH-synthesizing astrocytes, indicating that this hormone acts in an autocrine and paracrine fashion.

Vasoactive Intestinal Polypeptide Receptors

Vasoactive intestinal polypeptide (VIP) receptors are expressed by astrocytes, and the activation leads to Ca^{2+} signaling. All three subtypes of VIP receptor have been identified in cultured astrocytes. VIP receptor activation induces release of IL-6 and neurotrophic factors and stimulates proliferation. There is a link between receptor activation and glutamate uptake: Stimulation of VIP receptors leads to the promotion of GLT-1 and GLAST expression. This finding is in line with another study showing that VIP enhances glutamate uptake in astrocytes. Together, these data indicate that VIP can increase the strength of glutamate-mediated neurotransmission. In addition, VIP receptors may play an important role in regulating energy metabolism. VIP depletes astrocyte glycogen initially, followed by delayed reaccumulation to a level beyond baseline. These effects are mediated by regulating a number of related genes, such as glycogen synthase, via the transcription factor family CCAAT enhancer binding protein. However, there is not yet convincing evidence of the presence of VIP receptors in astrocytes *in situ*.

See also: AMPA Receptors: Molecular Biology and Pharmacology; Astrocyte: Identification Methods; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-B Receptors: Molecular Biology and Pharmacology; Glycine Receptors: Molecular and Cell Biology; Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology; Nicotinic Acetylcholine Receptors; Purinergic Receptors; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors; Somatostatin and Receptors; Substance P/Tachykinins and its/their Receptors; Vasoactive Intestinal Peptide and Pituitary Adenylate Cyclase Activating Peptide Receptors; Vasopressin/Oxytocin and Receptors.

Further Reading

Hatton GI and Parpura V (2004) *Glial Neuronal Signaling*. Boston: Kluwer.

Kettenmann H and Ransom BR (2005) *Neuroglia*. New York: Oxford University Press.

Matthias K, Kirchhoff F, Seifert G, et al. (2003) Segregated expression of AMPA-type glutamate receptors and glutamate transporters defines distinct astrocyte populations in the mouse hippocampus. *Journal of Neuroscience* 23: 1750–1758.

Schipke CG and Kettenmann H (2004) Astrocyte responses to neuronal activity. *Glia* 47: 226–232.

Seifert G, Schilling K, and Steinhäuser C (2006) Astrocyte dysfunction in neurological disorders: A molecular perspective. *Nature Reviews Neuroscience* 7: 194–206.

Seifert G and Steinhäuser C (2001) Ionotropic glutamate receptors in astrocytes. *Progress in Brain Research* 132: 287–299.

Verkhratsky A, Orkand RK, and Kettenmann H (1998) Glial calcium: Homeostasis and signaling function. *Physiological Reviews* 78: 99–141.

Volterra A, Magistretti PJ, and Haydon PG (2002) *The Tripartite Synapse*. Oxford, UK: Oxford University Press.

Volterra A and Meldolesi J (2005) Astrocytes, from brain glue to communication elements: The revolution continues. *Nature Review. Neuroscience* 6: 626–640.

Wang X, Lou N, Xu Q, et al. (2006) Astrocytic Ca(2+) signaling evoked by sensory stimulation in vivo. *Nature Neuroscience* 9: 816–823.

Astrocyte: Response to Injury

C J Feeney and P K Stys, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

© 2009 Published by Elsevier Ltd.

Introduction

Astrocytes are the most numerous cell type in the central nervous system (CNS) and make up about 50% of human brain volume. A central role of glial cells (in particular astrocytes) in regulation of the extracellular ionic environment has been well established, with key functions including K^+ buffering, homeostasis of H^+ and Ca^{2+}, detoxification of ammonia, and free radical scavenging. In addition, astrocytes are important in the production and release of growth factors, in supportive roles such as supply of energetic metabolites (lactate, pyruvate, and tricarboxylic acid cycle metabolites), neurotransmitter uptake, and precursor provision. Very recent data have even identified a role for astrocytes in promoting myelination of central axons in response to electrical activity. Thus it is clear that healthy astrocyte functioning is of paramount importance to proper CNS operation, and indeed survival, after insults such as ischemia and traumatic injuries.

In general, a perception predominates that astrocytes are relatively resistant to insults such as those caused by ischemia (oxygen–glucose deprivation; OGD) and oxidative stress. Recent work, however, points to significant astrocyte dysfunction and also cell death in response to these injuries. This article focuses on the various pathophysiological changes that can occur in astrocytes in response to hypoxia and ischemic-like insults.

How Do Astrocytes Fare in the Face of Ischemic-Like Insults?

Hypoxia

Cultured astrocytes have a well-documented prolonged resistance to exposure to lowered O_2 tension. In general, these studies utilize cultured cortical astrocytes derived from fetal or early neonatal-stage rodents. The elimination of oxygen alone (often termed *in vitro* hypoxia) has been shown to be relatively innocuous to these cultures, illustrating the astrocytes' ability to maintain viability for many hours of O_2 deprivation. These cells are able to survive such extended periods of hypoxia through the utilization of glycolytic adenosine triphosphate (ATP) production. Prolonged oxygen deprivation (12–24 h) is not, however, without effect on the cultured astrocyte, as subsequent reoxygenation of cultures leads to pronounced reactive oxygen species (ROS) production, mitochondrial damage, and cell death, at least in part, via apoptosis.

OGD (*In Vitro* Ischemia)

As with hypoxia, cultured astrocytes can also survive many hours of OGD, often termed *in vitro* ischemic conditions. Astrocyte cultures show an age-dependent resistance to OGD, with no significant cell death occurring in astrocyte cultures exposed to OGD for 6–12 h. Younger cultures (1 week *in vitro*) were found to be much more resistant to such OGD than older astrocyte cultures (3 weeks old). As well, the elimination of exogenous glucose also leads to relatively slow astrocyte demise, as cultured astrocytes can survive tens of hours in the absence of glucose when oxygen is present. Again, immature (younger) astrocyte cultures were found to be more robust than mature cultures. These findings suggest that maturational increases in metabolic rates in the developing brain *in vivo* and increases in oxidative metabolism in astrocytes *in vitro* with increasing age of culture confer a greater susceptibility of these cells to hypoxic/ischemic damage.

In Vivo Studies ('True' Ischemia)

Numerous published reports point to the susceptibility of neuronal cells to ischemia-induced injury *in vivo*, while far fewer have focused on astrocytic susceptibility. Those that have reported on the nature of astrocytic vulnerability to ischemia most generally document a greater resistance of astrocytes to ischemia compared with their neuronal counterparts. For example, a number of studies have shown that a brief transient cerebral ischemia sufficient to produce substantial neuronal cell death in CA1 of the hippocampus produces little to no astrocytic cell death. A more substantial insult, such as prolonged (2 h) or permanent middle cerebral artery occlusion (MCAO), has been shown to lead to rapid and significant astrocyte injury within 30–60 min, in some cases preceding neuronal demise. It should be noted that it is not clear whether these findings reflect a differential vulnerability of certain astrocyte populations or a different (i.e., faster) rate of cellular demise (necrosis vs. apoptosis). Finally, some have shown that particular subsets of astrocytes may be more susceptible to ischemic insults than others. For example, protoplasmic astrocytes appear to be more rapidly compromised than fibrous astrocytes in response to ischemia.

Astrocytic Responses to Ischemic-Like Insults

Mitochondrial Dysfunction

Ischemia is well known to induce rapid and sustained loss of mitochondrial function in numerous cell types. Of particular relevance in this article is the rapid and seemingly complete loss of neuronal mitochondrial potential ($\Delta\Psi_M$) in response to OGD. Thus, in cultured neuronal cell models, hippocampal slice cultures, and acute brain slices, neurons lose their mitochondrial potential within minutes of an ischemic-like insult. While prolonged exposure to oxygen and substrate deprivation must inevitably lead to mitochondrial dysfunction, conflicting evidence exists as to the early responses (if any) of astrocytic mitochondria to OGD. For example, some have shown that cortical astrocyte cultures exposed to 1 h of OGD do not lose their $\Delta\Psi_M$ or experience ATP depletion, nor does the exposure induce free radical production or decrease nicotinamide adenine dinucleotide phosphate (critical in the maintenance of reduced glutathione levels in the cell), whereas the same insult does so in neuronal cultures. Contrary to this finding, others have clearly shown that both cultured neurons and astrocytes lose their mitochondrial potential within 45–60 min when exposed to OGD. Of importance is that astrocytes that showed 'rapid' loss of $\Delta\Psi_M$ were cocultured with neurons, and the astrocytes cultured alone showed massive $\Delta\Psi_M$ loss only at 90 min of OGD (see Figure 1). The rapid mitochondrial dysfunction induced in these astrocytes was found to be, at least in part, related to the induction of permeability transition, as cyclosporin A partially protected astrocytes against $\Delta\Psi_M$ loss. It is interesting that astrocytes that find themselves amid neurons (e.g., cocultures or slice cultures) respond to OGD with a more profound mitochondrial dysfunction than those in isolation. In addition, the reestablishment of astrocyte $\Delta\Psi_M$ after ischemia-induced collapse appears to be gradual in nature. On 'reperfusion' with O_2 and glucose, astrocytic $\Delta\Psi_M$ is restored to 'normal' after several hours, possibly owing to some long-lasting structural changes in the astrocytic mitochondria themselves.

Whether these differences in astrocytic response are a result of differential culturing environments or a synergistic interaction between astrocytes and neurons leading to a heightened astrocytic sensitivity remains to be firmly established. What is clear, though, is that the loss of astrocytic $\Delta\Psi_M$ would likely have detrimental effects on CNS recovery after an ischemic-like insult. As astrocyte mitochondrial dysfunction proceeds, the homeostatic functions

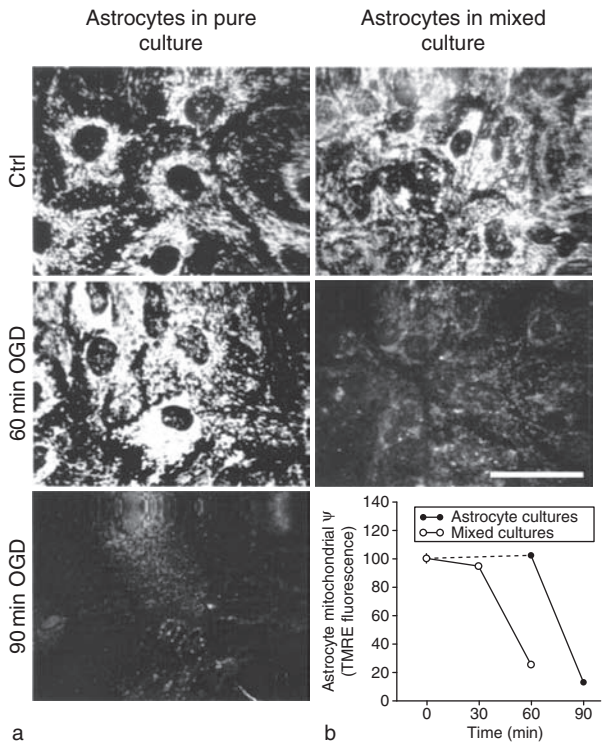

Figure 1 Astrocytes in pure and cocultures with neurons (mixed culture) lose mitochondrial potential in response to oxygen–glucose deprivation (OGD). Decreases in the mitochondrial potential dye tetramethylrhodamine ethylester (TMRE) reflect loss of mitochondrial potential ($\Delta\Psi_M$), which occurs faster in the cocultured astrocytes. Scale bar = 20 μm. From Reichert SA, Kim-Han JS, and Dugan LL (2001) The mitochondrial permeability transition pore and nitric oxide synthase mediate early mitochondrial depolarization in astrocytes during oxygen–glucose deprivation. *Journal of Neuroscience* 21(17): 6608–6616, copyright 2001 by the Society for Neuroscience.

astrocytes perform would be compromised, thereby leading to dysregulation of K^+ and H^+ ion concentrations, neurotransmitter uptake, and metabolite transfer to neighboring neurons. While short-term loss of astrocyte mitochondrial function may be well tolerated by the astrocytes themselves (by utilization of glycogen stores, alternate energetic metabolites), the relatively modest capacity of energy stores would soon lead to cellular compromise in both neuronal and nonneuronal compartments.

Reactive Oxygen Species

The production of injurious free radical compounds (e.g., ROS) has been linked to several neuropathological states, such as ischemic insults, Parkinson's disease, and Alzheimer's disease. Regardless of the exact mechanism and species of ROS produced, the oxidative stress these compounds initiate leads to the compromise of metabolic pathways, eventually leading to structural breakdown of cellular components.

ROS-initiated oxidative stress has been shown to be an essential process in excitotoxicity-induced degeneration. In ischemia, the neuronal ROS species produced in response to glutamate-dependent increases in cytosolic Ca^{2+} have been shown to cause elevated mitochondrial free Ca^{2+} concentrations, and this is perhaps an initiator of primary mitochondrial dysfunction in neurons after reperfusion. In astrocytes, this type of sustained elevation of intracellular Ca^{2+} concentration has been documented. Thus, activation of voltage-gated Ca^{2+} channels and receptor-mediated release of endoplasmic reticulum (ER) stores has been reported after ischemic-like insults (see Figure 2), sufficient to load astrocytic mitochondria. In addition, it has been demonstrated that astrocytes isolated from neuronal elements experience a less significant elevation in intracellular Ca^{2+}, suggesting the possibility that neuronally derived factor(s) may promote astrocytic dysfunction (or stress) after ischemic-like insults.

Most of the available data suggest that astrocytes are much more resistant to oxidative stress than are neuronal cell types. This may stem from the fact that they have been reported to contain higher concentrations of antioxidants such as α-tocopherol and ascorbate and higher levels of glutathione and the enzymes involved in glutathione metabolism. This, coupled with a relatively low level of membrane polyunsaturated lipids, suggests that these cells would be quite resistant to ROS-induced lipid peroxidation and oxidative stress. In light of these findings, many have suggested that astrocytes, being more competent in scavenging ROS, play a key role in the detoxification of oxidative processes in the CNS and the promotion of neuronal survival. In fact, astrocytes have been shown to be resistant to and 'neuro-protective' in the face of oxidative stress induced by the free radical generator H_2O_2.

Astrocytes are, however, not impervious to the effects of ROS or sustained oxidative stress. These cells will respond to an oxidative insult with an increase in their intracellular Ca^{2+} concentrations and loss of $\Delta\Psi_M$ and thereby oxidative phosphorylation. While some have shown that elevated intracellular Ca^{2+} is not a requisite for oxidative stress-induced astrocytic mitochondrial dysfunction, others have reported that increases in mitochondrial ROS production can lead to increased Ca^{2+} release from internal stores (ER), which can in turn lead to mitochondrial Ca^{2+} loading and subsequent dysfunction (via mitochondrial permeability transition induction) under oxidative stress (see Figure 3).

Thus, it may be that very local increases in Ca^{2+} concentration, at sites of ER/mitochondrial contact shown in other cell types, could lead to this destructive cycle of Ca^{2+} release, mitochondrial Ca^{2+} loading, and further ROS production, leading to progressive mitochondrial potential collapse and necrotic cell death.

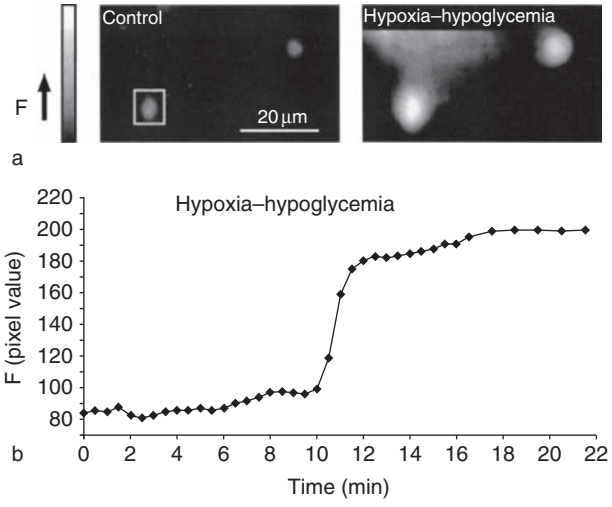

Figure 2 (a) Astrocytes from hippocampal slices were loaded iontophoretically with fluorescent Ca indicators. (b) Brief oxygen–glucose deprivation (termed Hypoxia–Hypoglycemia in the figure) leads to an abrupt rise of $[Ca^{2+}]_i$ in these cells. F, calcium orange fluorescence (arbitrary units). From Duffy S and MacVicar BA (1996) *In vitro* ischemia promotes calcium influx and intracellular calcium release in hippocampal astrocytes. *Journal of Neuroscience* 16(1): 71–81.

Glutamate Uptake

Astrocytes are well known for their ability to sequester extracellular neurotransmitters such as glutamate and are thus vital in maintaining excitatory–inhibitory network balances in brain activity. Glutamate is the primary excitatory neurotransmitter in the mammalian CNS, and excess release and unmanaged uptake have been linked to a number of neuropathological states, such as epilepsy, trauma, and ischemia. The existence of astrocytic high-affinity glutamate transporters (EAAT1 and EAAT2) is well established and is required for the rapid termination of transmitter action at presynaptic receptors. The uptake of glutamate into astrocytes is not chemically favored, given a ~1000-fold higher intracellular glutamate concentration in astrocytes. Thus, the uptake of glutamate is coupled to the electrogenic inward transport of Na^+ and H^+ in exchange for outward K^+ flow, which is dependent on a large Na^+ gradient generated by the action of the membrane Na^+/K^+-ATPase. Thus, under conditions of low ATP concentrations (e.g., mitochondrial dysfunction, ischemia), ionic gradients collapse, and these glutamate transporters become conduits for glutamate release (see Figure 4).

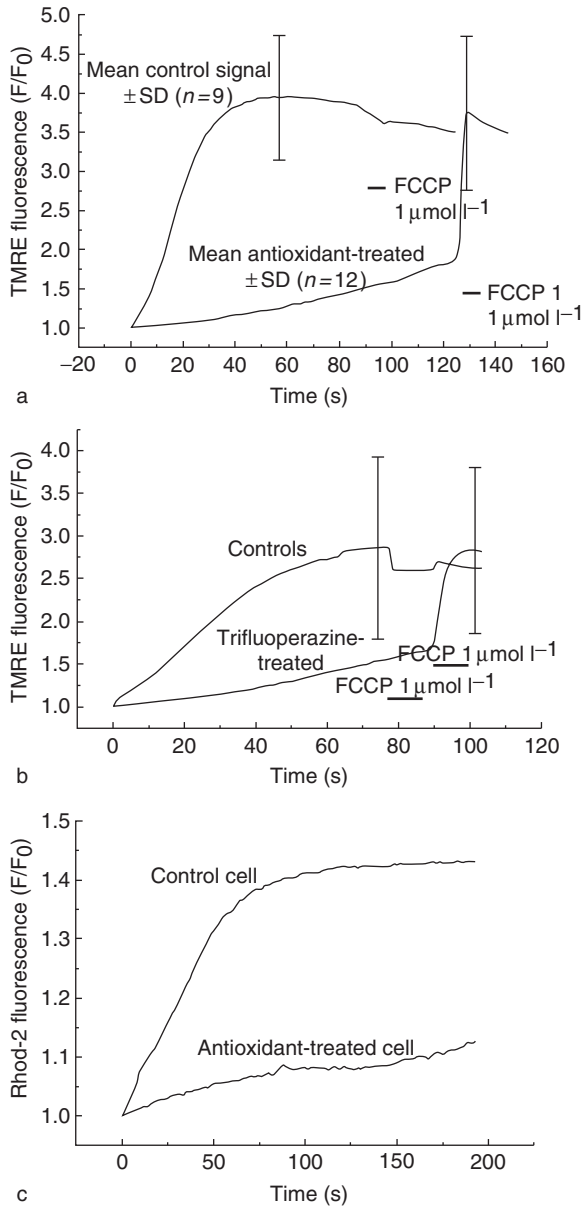

Figure 3 Reactive oxygen species (ROS)-induced astrocyte mitochondrial dysfunction. The antioxidants ascorbic acid, catalase, and Trolox, the spin trap TEMPO (a), and the mitochondrial permeability transition inhibitor trifluoperazine (b) were effective in reducing the rate of mitochondrial depolarization as measured by tetramethylrhodamine ethylester (TMRE; increasing fluorescent signal indicates loss of mitochondrial membrane potential). Large fluorescence increases with the addition of the protonophore FCCP indicate prior preservation of mitochondrial polarization, as with the antioxidants, or mitochondrial permeability transition inhibition. (c) ROS-induced astrocytic mitochondrial Ca^{2+} loading was significantly reduced with antioxidant treatment. In all experiments, ROS production was initiated ($t = 0$) by brief exposure to excitation light. F/F_0, fluorescence ratio compared to time zero. Modified from Jacobson J and Duchen MR (2002) Mitochondrial oxidative stress and cell death in astrocytes: Requirement for stored Ca^{2+} and sustained opening of the permeability transition pore. *Journal of Cell Science* 115(Pt. 6): 1175–1188.

The functional abilities of astrocyte glutamate transporters under less than complete ischemic conditions is unclear. While some have shown that astrocyte glutamate transporters are inhibited by incomplete ischemia, others have reported an initial stimulation or a maintained activity for substantial periods (hours) of ischemia. These differences may be attributable to degrees of acidosis experienced by the astrocytes, with greater decreases in pH leading to a more profound reduction in glutamate uptake. Thus, the extent of the ischemia experienced by the astrocytes, and the precise environment the cells are exposed to, will likely influence the activity of their glutamate transporter activity.

The effects of increased glutamate concentrations in astrocytes under pathologic conditions may not be without consequences to the astrocytes themselves. Both glutamate-induced swelling and cell death have been illustrated in cultures of astrocytes and are likely as a result of glutamate-stimulated K^+ uptake, and glutathione depletion and oxidative stress, respectively.

Cell Swelling

Ischemic-type insults are also accompanied by significant astrocytic swelling, occurring initially at astrocytic end-feet processes, then involving cell body regions. Astrocytes are the primary regulators of brain K^+ homeostasis. The removal of 'excess' extracellular K^+ is accomplished by two types of processes: K^+ spatial buffering and K^+ uptake, which work to rapidly redistribute high K^+ loads from active regions. Under ischemic-like environments, metabolically challenged cells respond to increases in extracellular K^+ with a rapid water accumulation, preceded by energy-independent influx of K^+, Cl^-, and HCO_3^- in response to membrane depolarization (see **Figure 5**). The load of extracellular K^+ can be substantial, climbing from about $3\,mmol\,l^{-1}$ up to $80\,mmol\,l^{-1}$ with ischemia. Recent evidence has suggested a significant role for the Na–K–Cl cotransporter (i.e., NKCC1) in ischemic-induced astrocytic swelling. Also, the combined action of the inwardly rectifying K channels (i.e., $K_{ir}4.1$) and the water channel protein aquaporin-4 (AQP4) has been shown to be particularly important in astrocyte K^+ siphoning at perivascular astrocytic end-feet. Thus, under the conditions of escalating extracellular $[K^+]$ observed in ischemia, homeostatic water influx is increased along with K^+ uptake.

Astrocytic swelling *per se* may not be lethal for the cells involved, as the reestablishment of normal extracellular $[K^+]$ results in repolarization of cell membranes. One of the consequences of continued K^+-induced cell swelling is the release of compounds such as glutamate and aspartate, via volume-regulated

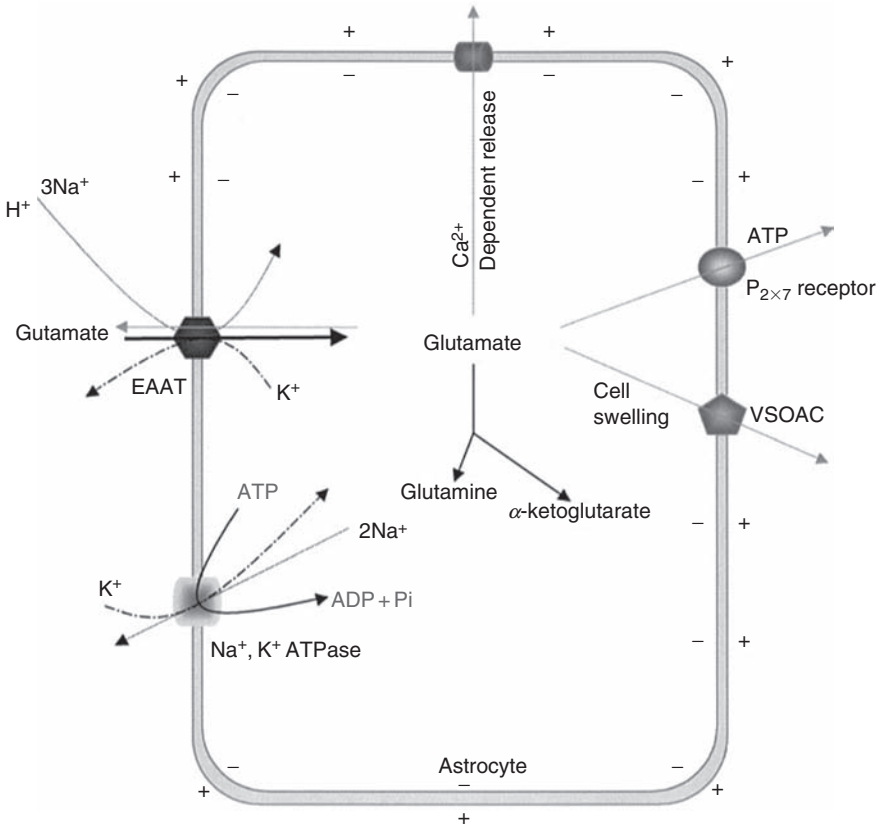

Figure 4 Glutamate handling in the ischemic astrocyte. In response to an ischemic-like insult, astrocyte mitochondria depolarize (with perhaps various time courses, depending on cell type, microenvironment, etc.), and glycogen stores are depleted. As energy metabolism is compromised, adenosine triphosphate (ATP) levels decline, and ATP-supported ionic membrane gradients collapse, leading to a halt in uptake and the release of glutamate through uptake reversal. EAAT, excitatory amino acid transporter; VSOCA, volume-sensitive organic anion channels. From Chen Y and Swanson RA (2003) Astrocytes and brain injury. *Journal of Cerebral Blood Flow and Metabolism* 23(2): 137–149.

anion channels, in an effort to reduce intracellular water loads. In fact, this adaptive homeostatic pathway may be a dominant mechanism of excitatory amino acid release in ischemic insults and may, ironically, lead to astrocyte-mediated neuronal excitotoxicity (see Table 1).

Apoptosis

Cell death, both *in vitro* and *in vivo*, may progress by two mutually nonexclusive ways. Thus, necrotic cell death is characterized by cell and organelle (mitochondria) swelling, leading to a rapid depletion of energy levels, breakdown of homeostatic regulation, and cell membrane rupture and release of the intracellular contents. In contrast, apoptotic ('programmed') cell death can exhibit varying degrees of chromatin condensation and DNA fragmentation, general membrane preservation (with possible blebbing), cysteine aspartate specific protease (caspase) activation, and cell shrinkage.

Astrocyte apoptosis has been demonstrated more clearly in focal ischemia, rather than global ischemic models. Thus, a MCAO for a few hours has been shown to cause extensive DNA laddering and activation of procaspase-12 in glial fibrillary acidic protein-positive cells (presumably astrocytes). In addition, a permanent occlusion of the same artery has been shown to upregulate numerous proapoptotic caspases in both neurons and astrocytes. It should be noted that others have shown that the morphological/ultrastructural changes that occur in astrocytes susceptible to ischemia resemble necrosis rather than apoptosis. From younger rat models, very strong evidence exists of an apoptotic astrocyte cell death in response to focal ischemic insults. Focal ischemia has been shown to upregulate Bax, to release cytochrome *c*, and to activate caspase-3 and -9 in astrocytes. However, it has also been shown that these types of insults do not lead to apoptotic astrocyte cell death. For instance, MCAO in adult mice leads to the induction of active caspase-3 and DNA laddering in neurons but not in astrocytes.

Cultured astrocytes can be induced to undergo apoptosis. The list of apoptotic inducers in astrocytes is long. In general, these factors and treatments lead

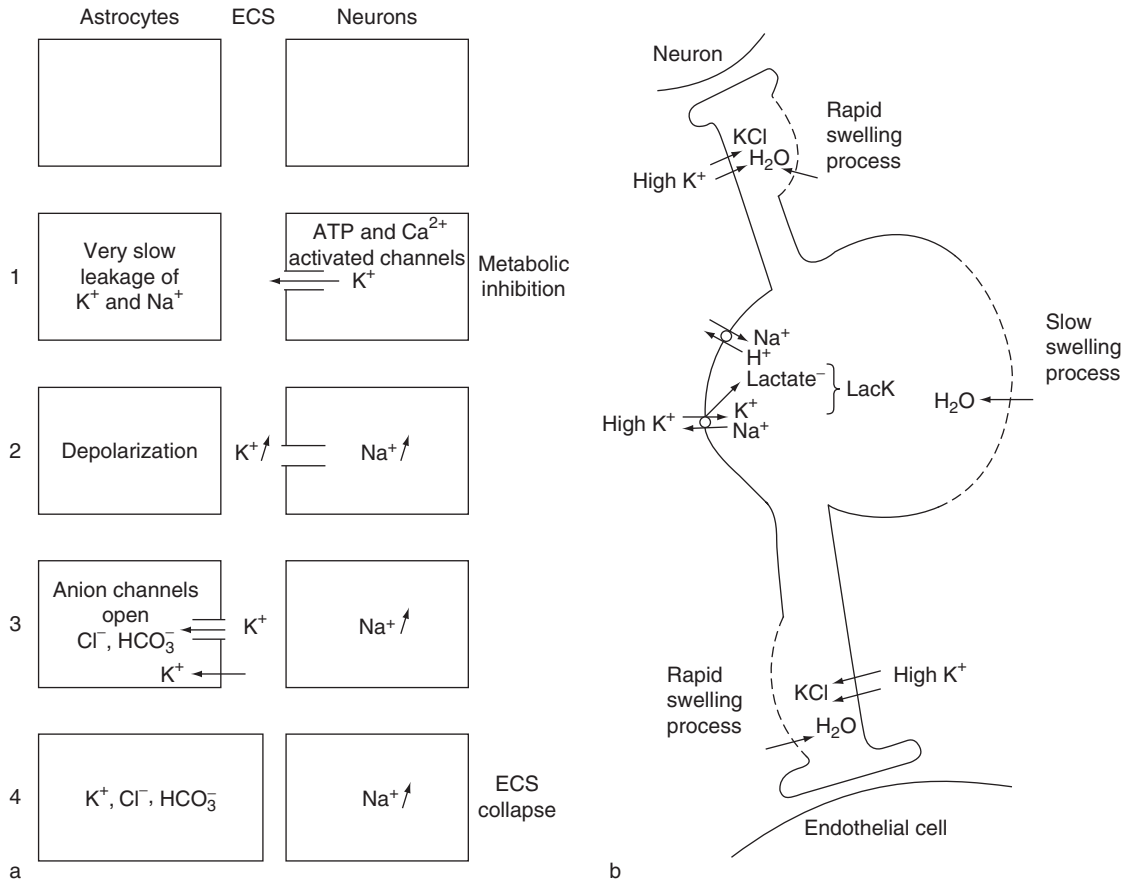

Figure 5 Ischemia-induced inhibition of Na^+/K^+ adenosine triphosphatase leads to progressive increases in extracellular $[K^+]$ (1) and membrane depolarization (2); continued elevation in $[K^+]_O$ leads to opening of anion channels and passive fluxes of K^+, Cl^-, and HCO_3^- (3), and water accumulates in astrocytes, causing them to swell (4). As ischemia continues, the short-term benefits of K^+ clearance become outweighed by the detrimental effects of reversal of glutamate uptake and/or cell swelling-induced excitatory amino acid release and possible effects on neuronal excitability (see **Table 1**). (b) A rise in extracellular K^+ depolarizes astrocytes, which in turn activate astrocytic anion channels. Passive, Donnan-mediated astrocytic K^+, Cl^-, and HCO_3^- accumulation causes significant cellular swelling. ATP, adenosine triphosphate; ECS, extracellular space; LacK, lactate–pottasium. (b) From Walz W, Klimas-Zewski A, and Paterson IA (1993) Glial swelling to ischemia: A hypothesis. *Developmental Neuroscience* 15(3–5): 216–225.

Table 1 Astrocytic pottasium homeostasis

Coping mechanism	*Primary beneficial effects*	*Secondary detrimental effects*
Increased spatial buffering (gap junction coupling)	↑ K^+ clearance	↑ Spreading depression and secondary injury
Alkalinization (Na^+/HCO) and sustained Na^+/K^+ ATPase[a] (glycogen)	↑ K^+ clearance	↑ Na^+ entry and glutamate release (reversal of uptake)
Increased passive K^+ uptake and cell swelling	↑ K^+ clearance	↑ Release of excitatory amino acids (regulatory volume decrease; RVD)

[a]ATPase, adenosine triphosphatase.
From Leis JA, Bekar LK, and Walz W (2005) Potassium homeostasis in the ischemic brain. *Glia* 50(4): 407–416.

to increases in intracellular calcium concentrations, oxidative stress, mitochondrial dysfunction, and protease activation. Of particular interest for this article, a number of studies have pointed to the involvement of oxidative stress in astrocytic apoptosis. The ROS producer H_2O_2 causes a loss of mitochondrial potential, cytochrome *c* release, and caspase-3 activation, a pathway likely to involve activation of transcription factor nuclear factor κB as a death-promoting factor in astrocyte apoptosis. The importance of elevated cytosolic calcium levels to this process has been established. Strategies to limit these increases of intracellular calcium, such as inhibition of astrocyte Na^+–Ca^{2+} exchanger and intracellular Ca^{2+} chelation, lead to reduction in astrocytic apoptosis *in vitro*.

Conclusions

Research over the past two decades has illustrated the importance of astrocyte functioning in management of the CNS microenvironment. Alterations in astrocytic physiology during pathological states such as ischemia and traumatic brain injury could have profound implications for the progression of these insults. Understanding the precise mechanism(s) of astroctyic demise or dysfunction should help in directing future clinical interventions aimed at reducing the extent of neuropathological damage.

See also: Apoptosis in Nervous System Injury; Astrocyte: Calcium Signaling; Astrocyte: Neurotransmitter and Hormone Receptors; Astrocyte: Identification Methods; Glial Responses to Injury; Neurotransmitter Release from Astrocytes.

Further Reading

Alberdi E, Sanchez-Gomez MV, and Matute C (2005) Calcium and glial cell death. *Cell Calcium* 8(3–4): 417–425.

Almeida A, Delgado-Esteban M, Bolanos JP, and Medina JM (2002) Oxygen and glucose deprivation induces mitochondrial dysfunction and oxidative stress in neurones but not in astrocytes in primary culture. *Journal of Neurochemistry* 81(2): 207–217.

Bender AS, Schousboe A, Reichelt W, and Norenberg MD (1998) Ionic mechanisms in glutamate-induced astrocyte swelling: Role of K^+ influx. *Journal of Neuroscience Research* 52(3): 307–321.

Benjelloun N, Joly LM, Palmier B, Plotkine M, and Charriaut-Marlangue C (2003) Apoptotic mitochondrial pathway in neurones and astrocytes after neonatal hypoxia-ischaemia in the rat brain. *Neuropathology and Applied Neurobiology* 29(4): 350–360.

Chen CJ, Liao SL, and Kuo JS (2000) Gliotoxic action of glutamate on cultured astrocytes. *Journal of Neurochemistry* 75(4): 1557–1565.

Chen Y and Swanson RA (2003) Astrocytes and brain injury. *Journal of Cerebral Blood Flow and Metabolism* 23(2): 137–149.

Duffy S and MacVicar BA (1996) *In vitro* ischemia promotes calcium influx and intracellular calcium release in hippocampal astrocytes. *Journal of Neuroscience* 16(1): 71–81.

Erecinska M and Silver IA (1990) Metabolism and role of glutamate in mammalian brain. *Progress in Neurobiology* 35(4): 245–296.

Fern R (2001) Ischemia: Astrocytes show their sensitive side. *Progress in Brain Research* 132: 405–411.

Ishibashi T, Dakin KA, and Stevens B, et al. (2006) Astrocytes promote myelination in response to electrical impulses. *Neuron* 49(6): 823–832.

Jacobson J and Duchen MR (2002) Mitochondrial oxidative stress and cell death in astrocytes: Requirement for stored Ca^{2+} and sustained opening of the permeability transition pore. *Journal of Cell Science* 115(Pt. 6): 1175–1188.

Juurlink BH, Hertz L, and Yager JY (1992) Astrocyte maturation and susceptibility to ischaemia or substrate deprivation. *Neuroreport* 3(12): 1135–1137.

Kimelberg HK (2005) Astrocytic swelling in cerebral ischemia as a possible cause of injury and target for therapy. *Glia* 50(4): 389–397.

Leis JA, Bekar LK, and Walz W (2005) Potassium homeostasis in the ischemic brain. *Glia* 50(4): 407–416.

Somjen GG (1979) Extracellular potassium in the mammalian central nervous system. *Annual Review of Physiology* 41: 159–177.

Swanson RA, Farrell K, and Stein BA (1997) Astrocyte energetics, function, and death under conditions of incomplete ischemia: A mechanism of glial death in the penumbra. *Glia* 21(1): 142–153.

Takuma K, Baba A, and Matsuda T (2004) Astrocyte apoptosis: Implications for neuroprotection. *Progress in Neurobiology* 72(2): 111–127.

Walz W (1989) Role of glial cells in the regulation of the brain ion microenvironment. *Progress in Neurobiology* 33(4): 309–333.

Atomic Force Microscopy Methodologies

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

Atomic force microscopy (AFM) is a technique used to produce quantitative topographic images of surfaces at a magnification of up to 10^7 (subnanometer scale, some 10^3 times better than possible using an optical microscope). The underlying principle is surface scanning using a sharp tip or stylus and the measurement of atomic forces between the stylus and the specimen. The stylus on an atomic force microscope is sensitive to both attractive and repulsive surface forces in the 0.1 μN range, but in practice the repulsive forces are the most stable, and these are the most widely used to obtain images. Scanning can be carried out in a vacuum, but a major advantage of AFM over scanning electron microscopy (SEM) is that AFM can be performed in air or with the sample immersed in liquid. Also, in contrast to SEM, no pretreatment of the sample, such as gold coating, is required. The resulting data also provide a true three-dimensional profile, not a two-dimensional image of a surface, as obtained by SEM. The data produced by AFM are a collection of points characterized by a constant force between the tip and the specimen, analogous to isobars on a weather map.

AFM is based on a related technique called scanning tunneling microscopy (STM), and some instruments can be fitted with swappable detectors to facilitate both types of analysis. The main difference between AFM and STM is that the latter requires a conducting surface that can sustain a stable tunneling current, making it unsuitable for most applications in neuroscience. Since AFM works on nonconducting surfaces, it is suitable for the analysis of cells, membranes, surface structures, protein complexes, individual proteins, and nucleic acids. Applications in neurobiology include the measurement of forces corresponding to receptor–ligand interactions, structural analysis of neuronal membranes, cell adhesion, and protein folding.

The information collected by surface scanning is quantitative in three dimensions, with vertical resolution on the order of 1 nm and horizontal resolution on the order of 0.1 nm. Samples that provide the best data have a smooth, flat surface, whereas rough or tilted samples generate poor images. Data are presented as topographical maps, often shaded or rendered in false color to indicate contours. Software can be used to rotate the image, generate virtual sections through the sample, and measure distances between features.

Atomic Forces and Their Measurement

The force between the stylus and the surface of the specimen is almost zero when the two are separated by a distance equivalent to more than a few hundred atoms. As the stylus gets closer to the surface, an attractive force builds up based primarily on van der Waals interactions. This attractive force is maximal when the tip of the stylus and the surface of the specimen are a few atoms apart, and then it is rapidly dispelled by a repulsive force as the two are pushed closer together. Contact equilibrium is established at the point where the net force between the stylus and the specimen surface is zero. At this point, any further movement toward the surface must involve an external force which causes elastic deformation and possible damage to the tip and the specimen.

If the stylus is maintained at a constant height relative to the surface at its starting position, then any peaks and troughs in the specimen's surface landscape encountered during scanning could result in loss of contact or collision, neither of which is desirable. Therefore, in most applications, and certainly in those where a variable surface landscape is anticipated, the stylus is mounted on a cantilever which allows its absolute height to be varied during scanning. A feedback mechanism is used to maintain a constant distance between the tip and the sample, and it is these changes in height that are used to map the specimen surface (**Figure 1**). The most common approach to measuring the small forces produced by surface scanning is to fabricate by microlithography a silicon nitride cantilever beam with an integral silicon nitride pyramidal tip and measure displacement by laser reflection, capacitance, or the detection of tunneling currents. Mechanical scanning is accomplished by using a piezoelectric transducer; if it is fabricated in the form of a hollow cylinder, with one electrode on the interior wall and four separate electrodes on each quadrant of the exterior wall, motion in three dimensions can be achieved. One end of the cylinder is fixed to the microscope base, and the specimen is mounted on the opposite end. A voltage applied between the interior electrode and all the outer electrodes will cause a dimensional change along the axis of the cylinder, or z-direction. Potentials applied between opposite quadrants of the outer electrodes will cause bending of the tube and subsequent deflection of the free end of the tube in the x- or y-direction. Typical ranges are 10 μm along the z-axis and 100 μm

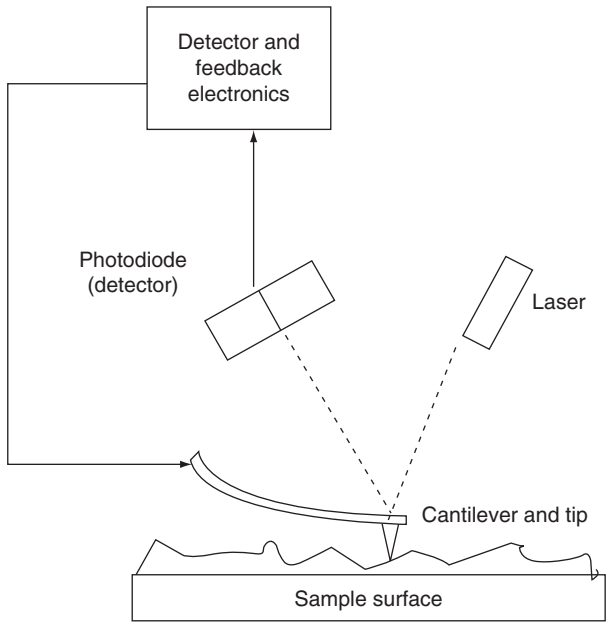

Figure 1 The principle of atomic force microscopy. Reproduced from Morris WG (2001) Atomic force microscopy In: *Encyclopedia of Materials: Science and Technology*, pp. 365–371. London: Elsevier, with permission from Elsevier.

along the x- and y-axis. For atomic resolution, actuators with smaller ranges in x- and y-direction are preferred because the scan may cover only a few nanometers. Alternatively, a tripod configuration can be used with three separate piezoelectric crystals, each responsible for scanning along one of the three axes. This can address image distortion issues that are caused by cross-talk when a traditional tube scanner is used.

Scanning Modes

The description above refers to atomic force microscopes operating in what is known as contact mode, in which the stylus remains in constant contact with the surface at the contact equilibrium point and differences in repulsive forces can be measured as the stylus tracks across the specimen. One disadvantage of this mode is that the measurements are somewhat prone to noise, although cantilevers that are more flexible than usual are employed to optimize the deflection signal. Another problem with contact mode is that under ambient conditions, a meniscus layer tends to form on most surfaces, and this strongly influences the interaction between tip and specimen. The tip can be held in contact with the specimen by the surface tension, and at the point of contact where the meniscus covers the specimen and stylus, attractive forces of up to 1 μN can be experienced. The repulsive force between the tip and specimen balances the surface tension, so the tip is almost always in the repulsive force mode when operating in air. If the liquid molecules remain between the tip and specimen, they can alter the force strength, which causes image distortion. This can be addressed through the use of liquid cells such that both the tip and the specimen are totally immersed in a liquid to eliminate forces generated by liquid–air surface tension.

An alternative approach which addresses the problems of static measurements is carrying out AFM in dynamic mode. Instead of remaining in static contact with the sample beneath, the stylus in dynamic mode oscillates at or near its resonance frequency. Feedback is provided, not by the z-displacement of the cantilever beam, but by differences in oscillation amplitude, phase, and resonance frequency relative to the external reference oscillation caused by tip–sample interaction forces. There are several varieties of dynamic mode, depending on which characteristics of the oscillating probe are measured. In frequency modulation, the changing oscillation frequency provides sensitive data about the forces between the stylus and sample, and very stiff cantilevers can be used that provide stability close to the sample and allow high-resolution scanning. Amplitude modulation is the most widely used mode; it uses changes in the oscillation amplitude or phase to provide the feedback for image creation. Amplitude modulation is particularly useful for addressing the problems caused by surface tension when scanning is done in air: The intermittent contact allows the stylus to be kept close enough to the sample for short-range repulsive forces to become detectable while the restoring force provided by the cantilever spring is strong enough to break contact with the surface meniscus. This mode is also known as tapping AFM because the stylus taps against the surface rather than dragging along it.

Image Collection

Feedback from sample scanning is fed into a computer which converts the forces into image data. However, before the sample is scanned, the same feedback system must be used to determine the correct operational position of the tip. The operating distance is established in two phases, a coarse stage, controlled by stepping motors, and a fine stage, in which the sample is moved toward the tip with a piezoelectric actuator. Data collection at this point is used to measure the atomic forces between tip and sample solely to establish the contact equilibrium point prior to the initiation of the scan.

Once scanning begins, the most straightforward way to present the data is height versus distance along a given line such that repeated scans along

parallel courses with a given vertical or horizontal offset can generate a crude topographical map. This results in a top–down topographical view showing the x–y area with different contours along the z-axis shown in gray scale (or rendered in false color) to indicate peaks and troughs. The maximum value of the gray scale can be changed, which in turn will change the contrast of the image, but it does not actually change the quantitative height information. The gray scale can be set to optimize the contrast in each individual image, making most of the details visible while minimizing regions of saturated black or white. Alternatively, it is easier to compare images if they are displayed with the same maximum gray-scale values. On a given specimen, wide-area scans will frequently require a larger maximum gray-scale value than small-area scans.

More-sophisticated computer programs can convert the simple topographical map into a three-dimensional perspective that can be viewed from any angle. The relative scale of the z-axis can also be increased or decreased with respect to the surface area so that surface features can be stretched and magnified. However, while enabling surface features to be studied in greater detail, this technique would also lead to the distortion and overemphasis of inconsequential features and artifacts.

Besides being an alternative to gray scale, the addition of color to an AFM image can help enhance certain image features, such as edges and clefts, which may be difficult to identify in monotone. At higher resolutions, coloring can be used to identify functionally important features of molecules, such as particular amino acid residues in proteins or points of contact between a receptor and its ligand. False color can also be used for trivial reasons, for instance, to add shadows to three-dimensional landscapes as if the surface were illuminated by a directional light source. While color can be used for clarification in many circumstances, the overuse of color has the counterproductive effect of making images more difficult to interpret.

Quantitative Data

One disadvantage of AFM compared with SEM is the relatively long time it takes to collect image data, in part reflecting the necessity to carry out several scans along the horizontal and vertical axes to establish topography. This can also lead to some loss of data integrity through a phenomenon known as thermal drift, which makes it difficult to collect accurate quantitative data, such as to measure distances between features on the surface landscape. Another potential source of inaccuracy is hysteresis of the piezoelectric material. Several AFM designs with faster scan times have been developed to overcome the data problems caused by thermal drift, including video-AFM and the use of feature-oriented scanning and closed-loop scanners which collect and process images in near real time. Software has also been developed to correct distortions caused by thermal drift and hysteresis.

The accuracy of calibration can also limit the accuracy of any quantitative information extracted from the image, so the instrument must be calibrated regularly using a specially constructed mesh with 1-μm periodicity along both the x- and y-axis as well as a peak-to-valley distance of 0.2 μm. Corrections for other more complex types of distortion, described by terms like pinch, punch, pincushion, shear, and swirl that are familiar to Adobe Photoshop users, may also be necessary.

Further software fixes can be used to process images and enhance the information content. For example, the Fourier transform can be applied to a height profile to determine the frequency of regularly spaced surface features, and it can help clean up and smooth out the features of rough, noisy images, thus preserving the relevant features while removing artifacts.

Limitations and Drawbacks

Although AFM has many advantages over SEM in terms of the convenience of sample preparation and scanning, there are several limitations which restrict the applications of this method and which are the subject of current research. One significant issue is the limited image area and depth. The SEM can image an area more than 100 times the size of that possible using an AFM because it has a displacement of several millimeters along the x- and y-axis, whereas the maximum useful horizontal and vertical displacement possible with an AFM is about 150 μm. Newer piezodrivers extend this to approximately 250 μm, but such wide scans often result in greater image distortion, which must be corrected with software. Mechanical stages permit the cantilever–detector assembly to be positioned within a 5-mm^2 area so that larger areas can be covered with adjoining multiple images when needed.

The depth of field for an SEM is also in the millimeter range, whereas for AFM it is a few micrometers. This means that although the AFM has great sensitivity for small changes in elevation, it has a very limited vertical range, restricting the AFM to the analysis of surfaces which are relatively smooth and flat when observed using other methods and making the analysis of convex surfaces challenging.

The resolution and characteristics of an AFM image are also dependent on the integrity of the stylus. The

radius of curvature at the tip must be on the order of 30 nm to achieve the best spatial resolutions, but extended use can cause wearing and damage which will generate artifacts in image data. For example, a worn tip with a flat region on it may generate artifactual square features on the image, and a new tip or a damaged tip with transferred material on it may contain rough protrusions (asperities) that generate shadows.

See also: Cognition: An Overview of Neuroimaging Techniques; Neuroimaging.

Further Reading

Eibl RH and Moy VT (2005) Atomic force microscopy measurements of protein–ligand interactions on living cells. *Methods in Molecular Biology* 305: 439–450.

Gadegaard N (2006) Atomic force microscopy in biology: Technology and techniques. *Biotechnic & Histochemistry* 81: 87–97.

Hansma HG (2001) Surface biology of DNA by atomic force microscopy. *Annual Reviews of Physical Chemistry* 52: 71–92.

Li G, Xi N, and Wang DH (2006) Probing membrane proteins using atomic force microscopy. *Journal of Cellular Biochemistry* 97: 1191–1197.

Meyer E, Hug H-J, and Bennewitz R (2004) *Scanning Probe Microscopy: The Lab on a Tip*. Berlin: Springer.

Morita S, Wiesendanger R, and Meyer E (2002) *Noncontact Atomic Force Microscopy, Nanoscience and Technology*. Berlin: Springer.

Morris VJ, Kirby AR, and Gunning AP (1999) *Atomic Force Microscopy for Biologists*. London: Imperial College Press.

Shao Z, Mou J, Czajkowsky DM, Yang J, and Yuan J-Y (1996) Biological atomic force microscopy: What is achieved and what is needed. *Advances in Physics* 45: 1–86.

Zlatanova J and van Holde K (2006) Single-molecule biology: What is it and how does it work? *Molecular Cell* 24: 317–329.

Atrial Natriuretic Peptide: Fluid/Mineral Balance

W K Samson and M M White, Saint Louis University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Originally isolated from extracts of the cardiac atrium, atrial natriuretic factor (ANF) was the first peptide identified in what is now known to be a family of homologous peptides, products of unique genes, which formed the basis of our initial understanding of the heart as an endocrine organ and the science of cardiovascular endocrinology. Released primarily in response to stretch (increased venous return or increased afterload), ANF and its homolog B-type natriuretic peptide (BNP) exert potent adrenal, vascular, and renal actions all seemingly coordinated to unload the vascular tree (e.g., inhibition of aldosterone release, increased glomerular filtration, natriuresis, diuresis, and inhibition of sympathetic tone). A third member of the family of natriuretic peptides (NPs), C-type natriuretic peptide (CNP), is predominantly produced in the vascular endothelium and shares some of the actions of ANF and BNP. All three peptides are present in the central nervous system (CNS) and exert potent pharmacologic actions there. Evidence for a coordination of the CNS and peripheral actions of these peptides is quite convincing. Thus, complementary actions in several organ systems indicate that these peptides are important contributors to the physiologic regulation of fluid and electrolyte homeostasis.

Discovery of the Natriuretic Peptides

It was recognized more than 40 years ago that cardiac myocytes were more than simply contractile elements, based on the histological identification of secretory granules in the cytoplasm. In 1981, pioneering studies by AJ deBold and colleagues in Kingston, Ontario, provided evidence that those granules contained a vasoactive substance that not only caused hypotension when injected into bioassay animals but also stimulated increased urine volume (diuresis) and urinary sodium excretion (natriuresis). Two years later, TG Flynn and AJ deBold provided the presumptive amino acid sequence for what is now known as ANF, the factor extracted from those myocytes which is the renotropic agent identified in their bioassay. Soon thereafter, K Kangawa and colleagues in Osaka identified the second member of the family, BNP, and in 1990 that same group reported the sequence of CNP. All three peptides are products of unique genes and, largely because of their structural homology, they share common receptors with varying degrees of affinity (Figure 1). The type A NP (NPR-A) and NPR-B receptors are membrane-bound guanylyl cyclases (GCs) that signal through the formation of cGMP. They are also referred to as GC-A and GC-B, respectively. A third receptor, NPR-C, shares some structural homology with the NPR-A and NPR-B in the extracellular domains, but it is devoid of the intracellular guanylyl cyclase domain. This receptor was originally thought to function as primarily a biologically inactive, clearance receptor; however, it is now recognized that ligand binding to this receptor results in inhibition of cAMP formation, probably via a G_i protein, which in isolated hypothalamic neurons results in inhibition of L-type calcium channels. The three receptors display selective ligand binding (Table 1).

Control of Natriuretic Peptide Production and Release

The major stimulus for the release of ANF and BNP from the heart is volume expansion (i.e., stretch). However, numerous neurohumoral factors stimulate release, including endothelin, angiotensin, adrenal steroids, inflammatory cytokines, growth factors, prostaglandins, thyroid hormones, and alpha-adrenergic agents. Although constitutive release of the peptides occurs under normal conditions, production and secretion are upregulated in volume overload states such as congestive heart failure or after ischemic events (e.g., acute myocardial infarction). In fact, cardioprotective effects of the peptides imply that the upregulation of ANF and BNP production and secretion represents a physiologically relevant compensatory mechanism that protects cardiac function (anti-mitogenic actions and increased coronary artery perfusion) and facilitates restoration of normal vascular volume (renotropic actions and effects on capillary permeability). It is not surprising, then, that both ANF and BNP have been examined as potential therapeutic agents for use immediately following acute myocardial infarction and in congestive heart failure. In addition, the increased secretion of the peptides into the circulation in these pathologies has led to the use of rapid assays for the peptides as diagnostic tools.

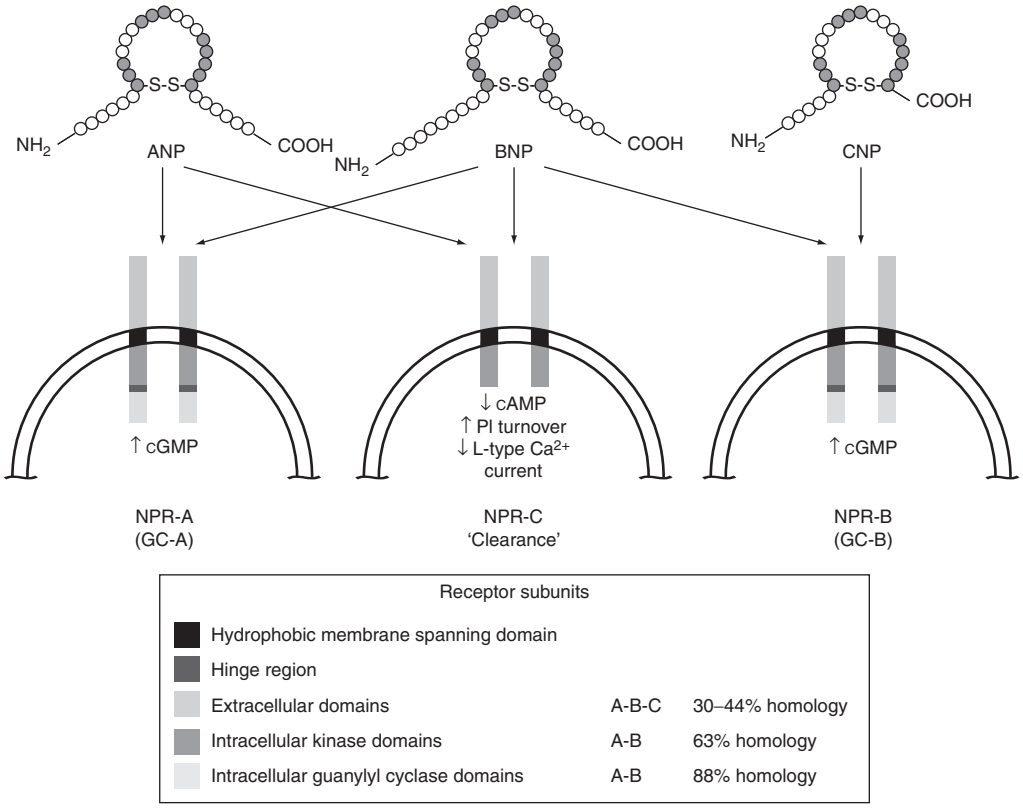

Figure 1 Structure of the natriuretic peptides and their receptors. Signaling mechanisms recruited by the three receptors are indicated.

Table 1 Natriuretic peptide receptor: order of binding potency

NPR-A (GC-A)	ANF $\geq$ BNP $>$ CNP
NPR-B (GC-B)	CNP $\gg$ ANF $\geq$ BNP
NPR-C ('clearance')	ANF $\geq$ CNP $\geq$ BNP

CNS Localization of NPs and NP Receptors

Soon after the initial description of ANF in heart, it became clear that neurons in the CNS also produce the peptide and that it exerts actions within brain that complemented its peripheral actions. Indeed, the first example of a CNS action of ANF was the demonstration that the peptide inhibits arginine vasopressin (AVP) secretion. ANF also acts in brain to inhibit thirst and sodium appetite. The anatomic framework for these pharmacologic actions was established with the demonstration by autoradiography of ANF binding in CNS sites known to be important in the control of fluid and electrolyte homeostasis, cardiovascular/renal function, and the hypothalamic–pituitary–adrenal axis.

Immunoreactive perikarya were identified in lateral septal nucleus, bed nucleus of the stria terminalis, and medial amygdaloid nucleus in the telencephalon. In diencephalon, ANF-positive cell bodies were localized in numerous sites, including organum vasculosum lamina terminalis (OVLT), periventricular and medial preoptic nuclei, preoptic suprachiasmatic nucleus, parvocellular paraventricular nucleus, arcuate nucleus, ventro- and dorsomedial hypothalamic nuclei, habenula, zona incerta, mammillary nuclei, and the dorsal aspect of the lateral hypothalamic area. Dense innervation of these same diencephalic sites as well as the subfornical organ (SFO), external lamina of the median eminence, periventricular thalamus, and the medial lemniscus has been described. Peptide localization to amygdala and hypothalamus implied a role for ANF in the hypothalamic control of thirst, salt appetite, and AVP secretion and even hinted at possible actions of the peptide in the anterior pituitary gland.

In midbrain, ANF-positive cell bodies were identified in interpeduncular nucleus, periaqueductal gray, ventral tegmental area, dorsal raphe, and substantia nigra pars lateralis. ANF-immunoreactive perikarya in the interpeduncular nucleus were surrounded by a dense network of ANF-positive fibers that extended into the neighboring medial longitudinal fasciculus. In hindbrain, ANF-positive cell bodies were detected in dorsal parabrachial nucleus, nucleus of the spinal tract of V, the dorsal tegmental nucleus, pedunculopontine tegmental nucleus, lateral tegmental nucleus, Barrington's nucleus, and locus coeruleus. Caudally, cells were stained positively in nucleus

prepositus hypoglossi and nucleus tractus solitarius (NTS). Immunoreactive fibers were abundantly present in NTS and dorsal motor nucleus of the vagus (DMV). A more diffuse pattern of fiber staining was observed in spinal tract of V, nucleus ambiguus, central gray, and locus coeruleus. The presence of ANF-positive parikarya and nerve fibers in these brain stem sites suggested a potential role for the peptide in autonomic function, both the afferent (e.g., baroreceptor input) and the efferent (both sympathetic and parasympathetic) components.

As with the initial studies identifying CNS sites of ANF immunoreactivity, early work from several groups identified receptors for the peptide based on binding studies. It is impossible with this approach to identify the specific subtype of receptor since ANF binds to all three cloned NP receptors. However, these localization studies further suggested a role for endogenous ANF (or BNP or CNP) in the CNS control of fluid and electrolyte homeostasis and autonomic function. Dense autoradiographic binding was reported in olfactory structures and at least two circumventricular organs, the SFO and area postrema (AP). In hypothalamus, abundant binding was described particularly in the paraventricular and arcuate nuclei. These findings, as well as those of peptide localization studies, predicted subsequent findings that revealed potent actions of ANF on AVP and prolactin secretion, respectively. The medial preoptic area also displayed autoradiographic binding sites, perhaps providing an anatomic framework for the subsequent demonstration of the effects of the NPs on the hypothalamo–pituitary–gonadal axis. In hindbrain, binding was detected in NTS, AP, DMV, locus coeruleus, and raphe nuclei and broadly dispersed throughout the reticular formation.

The development of receptor subtype selective probes led to the hypothesis that the major NP receptor within the brain was the NPR-B, in agreement with data implying that the most abundant NP produced in CNS was CNP. In fact, the majority of the NP binding sites behind the blood–brain barrier are NPR-B and NPR-C receptors, whereas those present in areas without the barrier (e.g., the SFO) are predominantly the NPR-A receptor. This finding implies that circulating ANF or BNP of cardiac origin affects neural function at those barrier-free sites, whereas the majority of the pharmacologic actions described to date for the various NPs when administered behind the blood–brain barrier (i.e., into the cerebroventricles or directly into the brain interstitium) reflect an action on the NPR-B receptor. Indeed, the opposing actions of ANF and CNP on water drinking support this hypothesis.

Pharmacologic Effects of the NPs in Brain

The major focus of the initial studies on the actions of the NPs in brain was anatomic localization. Thus, early reports detailed the effects of intracerebroventricular (i.c.v.) administration of ANF on thirst, salt appetite, AVP secretion, and blood pressure regulation. With the eventual realization that most of the NP produced in brain was in fact CNP, those initial studies were repeated using CNP as the pharmacologic agent. In general, these studies can be summarized as follows: NPs act, with one notable exception, either at circumventricular organs or in CNS sites behind the blood–brain barrier to complement the peripheral actions of the peptides on intravascular volume, plasma osmolality, and blood pressure. As in the periphery, many of these CNS actions rely on the ability of the peptides to abrogate the actions of angiotensin II (Ang II).

NPs and Vasopressin Secretion

In 1985, it was reported that the diuretic action of ANF in kidney mirrored an action of the peptide in brain to inhibit AVP secretion. Dehydration- and hemorrhage-induced AVP release, as well as pharmacologically driven (i.e., Ang II-stimulated) AVP secretion, was inhibited by ANF. The site of action was the hypothalamus since ANF inhibited AVP release from isolated hypothalamo-neurohypophyeal explants but not from isolated neural lobes. In addition, ANF inhibited neuronal activity in the paraventricular nucleus. Whether these data reflect a physiologic action of endogenously produced ANF or CNP is not known; however, dehydration, a clear stimulus for AVP production and secretion, also increased ANF-like immunoreactivity in the rat hypothalamus.

There is electrophysiologic evidence for direct effects of ANF on synaptic transmission in hypothalamic explants. Osmotic stimulation of the OVLT increased frequency of spontaneous, glutamatergic excitatory postsynaptic potentials (EPSPs) in magnocellular neurons, and the amplitude of the glutamatergic EPSPs evoked by electrical stimulation of the OVLT was reduced by ANF. The effect was mimicked by 3′,5′-dibutyryl cGMP, implying mediation by either the NPR-A or NPR-B receptors. These effects were presynaptic since ANF did not alter depolarization of the magnocellular cells in the presence of glutamate receptor agonists. Direct, postsynaptic effects of the NPs on magnocellular neurons of the paraventricular and supraoptic nuclei have been demonstrated. CNP and the NPR-C receptor selective agonist cANF-4–23 produced significant inhibitions of L-type calcium currents that were mimicked by application of a peptide that activates G_i proteins

and prevented by nicarpidine pretreatment. In current clamp mode, CNP decreased the number of action potentials resulting from depolarizing stimuli and decreased repolarization time. Although the morphologic phenotype of the cells recorded was magnocellular elements dispersed from or located within these nuclei, the chemical identity was not established and thus these results may have been generated in both vasopressin- and oxytocin-expressing cells. Importantly, these experiments demonstrate the contribution of the G_i signaling pathway in the inhibitory effect of NPs acting on the 'clearance' receptor (NPR-C) in hypothalamus.

ANF and Thirst

The first descriptions of a CNS action of ANF were reports that the peptide inhibits water drinking in response to dehydration or Ang II administration. These pharmacologic actions of ANF have a physiologic correlate since pretreatment of rats i.c.v. with an ANF antiserum exaggerated water drinking in response to Ang II administration or water deprivation. These studies imply the SFO as the site of the antidipsogenic action, which is consistent with the presence of the NPR-A receptor there. The location of the SFO at the confluence of the lateral and third cerebroventricles predicts exposure of the receptors to both circulating and i.c.v.-administered ANF. Earlier localization studies described ANF-positive nerve fibers innervating the SFO; thus, ANF of central origin may act there as well, explaining the results of the passive immunoneutralization study.

However, not all NP effects on thirst are inhibitory. CNP, when administered i.c.v., increased water drinking in response to Ang II and dehydration. This dipsogenic action of CNP was in all likelihood expressed on the NPR-B receptor since both ANP and the NPR-C receptor selective ligand, cANF 4–23, inhibited water intake in these models. It should be remembered that CNP is the main NP produced in brain and that the NPR-B and NPR-C receptors are abundant in blood–brain barrier-protected sites. Thus, although circulating ANF, or BNP, may exert antidipsogenic actions in SFO (intravenous infusion of ANF did inhibit vasopressin secretion), the hypothalamic effect of the NPs on thirst, particularly those of endogenous CNP, may be exerted via the NPR-B and therefore be stimulatory in nature. Thus, the CNS actions of brain-derived NPs on thirst may not complement the diuretic action of ANF in kidney.

ANF and Salt Appetite

The natriuretic action of ANF in kidney is complemented by an action in brain to inhibit salt appetite. The physiologic relevance of this CNS action has been established. Acute and chronic i.c.v. administration of ANF inhibits salt appetite induced by NaCl deprivation in normotensive animals and in spontaneously hypertensive rats (SHRs). The results in the SHR model were striking since these were *ad libitum*, two-bottle preference protocols and the effect on salt appetite was not present in the Wistar Kyoto controls. In addition, the effect to inhibit salt appetite was not related to any change in blood pressure in these hypertensive animals. Both the SHR and Wistar Kyoto animals consumed less water in response to ANF infusion, demonstrating selectivity of action of ANF on thirst versus salt appetite during *ad libitum* fluid ingestion.

ANF is not the only neuropeptide reported to act within brain to inhibit salt appetite. A physiologically relevant role for oxytocin and for adrenomedullin in the control of salt intake has been established. In fact, oxytocin neurons may mediate the inhibitory effect of adrenomedullin, implying the importance of both peptides in the neural network responsible for the control of salt appetite.

A series of experiments examined the potential interactions of ANF and oxytocin on salt appetite. Rats were pretreated with cytotoxin (ricin A chain) conjugates designed to compromise cells expressing either oxytocin or ANF receptors or they were treated with unconjugated ricin A chain (nontoxic control). Salt appetite in these animals was studied following a combination of hypovolemia and hyperosmolar challenges using either 2.0 M mannitol (which raises osmolality but lowers plasma sodium levels) or 1.0 M NaCl (which raises both osmolality and plasma sodium level). The induction of hypovolemia stimulates both water and NaCl ingestion in control animals and the superimposition of the hyperosmolar load inhibits salt appetite, even in the face of the hypovolemic challenge. In contrast, exaggerated saline drinking in response to hypertonic mannitol administration was observed in animals in which either the oxytocin or the ANF receptor expressing cells were compromised by cytotoxin targeting. Thus, in the absence of either oxytocin or ANF signaling in brain, these animals could no longer appropriately terminate their appetite for salt in the hypovolemic but hyperosmotic state induced by mannitol.

On the other hand, in animals stimulated to drink water and saline by the colloid pretreatment, hyperosmotic challenge with sodium chloride resulted in appropriate inhibition of saline drinking in controls and in rats pretreated with the oxytocin–ricin A chain cytotoxin conjugate. However, exaggerated saline intake was observed in animals unable to respond to the signal of endogenous ANF (ANF–ricin A chain

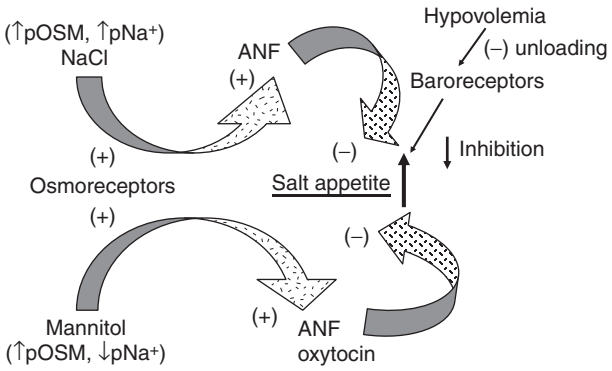

Figure 2 In rats with thirst and salt appetite induced by hypovolemia, hyperosmotic stimulation of the osmoreceptors with mannitol inhibits salt appetite by recruitment of central oxytocin and/or ANF. Hyperosmotic stimulation of the osmoreceptors with NaCl, in contrast, selectively activates central ANF pathways that inhibit salt appetite. pNa^+, plasma sodium concentration; pOSM, plasma osmolality.

pretreated rats). Thus, a divergence appeared between the inhibitory actions of oxytocin and ANF on saline drinking depending entirely on the nature of the hyperosmotic stimulus (Figure 2).

In a final experiment, pharmacologically driven salt appetite was similarly affected in both oxytocin and ANF receptor compromised rats when administered a dose of Ang II. This finding implies that Ang II acts not only to stimulate water and saline drinking but also to stimulate the central release of both oxytocin and ANF, which then buffer the effect of Ang II on salt appetite. Interestingly, the dipsogenic effect of Ang II was not enhanced in the ANF–ricin A chain conjugate pretreated animals, which might have been expected since ANF inhibits water intake stimulated by Ang II in control animals. This finding implies that the sensitivity of the inhibitory effect of ANF on salt appetite is much greater than that for thirst.

ANF and Central Autonomic Control

The initial localization of ANF binding in CNS sites known to be important in central cardiovascular control predicted the actions of ANF on blood pressure regulation in conscious rats. Indeed, the inhibitory effect of ANF on Ang II-stimulated mean arterial blood pressure in conscious rats has been demonstrated and extended to include an inhibitory effect of ANF on the blood pressure elevation observed following central hypertonic saline injection. Potential sites of action of ANF to lower blood pressure and reduce heart rate were identified to be the NTS and rostral ventrolateral medulla (RVLM). These two areas receive afferent ANF-positive input from neurons in the PVN and lateral hypothalamus and the lateral parabrachial nucleus. In addition to inhibitory effects of ANF on putative sympathoexcitatory neurons of the NTS and RVLM, the peptide has been demonstrated to stimulate preganglionic parasympathetic neurons in the dorsal motor nucleus of the vagus.

In addition to inhibiting the central action of Ang II on blood pressure, thirst, and vasopressin secretion, ANF may act in brain stem sites to reduce the action of Ang II on renal sympathetic nerve activity. In this way, the peripheral natriuretic actions of ANF, exerted directly in the tubule or via blockade of the peripheral renin–angiotensin system, may be complemented by CNS actions as well. Indeed, when ANF is administered centrally, diuresis and natriuresis occur.

Summary

The natriuretic peptides exert CNS actions that complement those expressed in the periphery, particularly in heart, kidney, and adrenal gland. These actions seem well coordinated to protect the cardiovascular system against volume or pressure overload. The physiologic relevance of the CNS actions has been established in animal models; however, it is not clear that these effects are expressed under physiologic conditions in humans. Although there is evidence for coordinated production and release of ANF in brain and heart, the conditions under which CNS production of ANF or CNP may vary have not been exhaustively catalogued. On the other hand, circulating levels of ANF and BNP are elevated in several disease states, and thus physiologically or pathologically relevant actions of the peptides may be exerted in the circumventricular organs under those conditions. Therefore, therapeutic use of NP analogs or enzyme inhibitors that prolong the plasma half-life of the endogenous peptides may result in CNS actions on thirst, vasopressin secretion, salt appetite, or autonomic function.

See also: Angiotensin II; Blood Pressure: Baroreceptors; Circumventricular Organs; Fluid and Electrolyte Homeostasis: Clinical Disease; Natriuretic Peptides; Neurohypophyseal System; Osmoregulation; Salt Appetite; Thirst.

Further Reading

Ackermann U and Azizi N (2000) Increased central AT1 receptor activation, not systemic vasopressin, sustains hypertension in ANP knockout mice. *American Journal of Physiology* 278: R1441–R1445.

Blackburn RE, Samson WK, Fulton RJ, Stricker EM, and Verbalis JG (1995) Central oxytocin and ANP receptors mediate osmotic inhibition of salt appetite in rats. *American Journal of Physiology* 269: R245–R251.

John SW, Veress AT, Honrath U, et al. (1996) Blood pressure and fluid–electrolyte balance in mice with reduced or absent ANP. *American Journal of Physiology* 271: R109–R114.

Levin ER, Gardner DG, and Samson WK (1998) Natiuretic peptides. *New England Journal of Medicine* 339: 321–328.

Lopez MJ, Wong SK, Kishimoto I, et al. (1995) Salt-resistant hypertension in mice lacking the guanylyl-cyclase-A receptor for atrial natriuretic peptide. *Nature* 378: 65–68.

McGrath MF, Kuroski de Bold M, and de Bold AJ (2005) The endocrine function of the heart. *Trends in Endocrinology and Metabolism* 16: 469–477.

Melo LG, Steinhelper ME, Pang SC, Tse Y, and Ackermann U (2000) ANP in regulation of arterial pressure and fluid–electrolyte balance: Lessons from genetic mouse models. *Physiological Genomics* 3: 45–58.

Munagala VK, Burnett JC, and Redfield MM (2004) The natriuretic peptides in cardiovascular medicine. *Current Problems in Cardiology* 29: 707–769.

Okada J, Takayama K, Xiong Y, and Miura M (1994) Influence of humoral control peptides on medullary vasomotor control neurons: Microstimulation and double-labeling studies using SHR and WKY rats. *Journal of the Autonomic Nervous System* 49: 171–182.

Potter LR and Hunter T (2001) Guanylyl cyclase-linked natriuretic peptide receptors: Structure and regulation. *Journal of Biological Chemistry* 276: 6057–6060.

Rademaker MT and Richards AM (2005) Cardiac natriuretic peptides for cardiac health. *Clinical Science* 108: 23–36.

Richard D and Bourque CW (1996) Atrial natriuretic peptide modulates synaptic transmission from osmoreceptor afferents to the supraoptic nucleus. *Journal of Neuroscience* 16: 7526–7532.

Rose RA, Anand-Srivastava AB, Giles WR, and Bains JS (2005) C-type natriuretic peptide inhibits L-type Ca^{2+} current in rat magnocellular neurosecretory cells by activating the NPR-C receptor. *Journal of Neurophysiology* 94: 612–621.

Samson WK and Levin ER (eds.) (1997) *Natriuretic Peptides in Health and Disease*. Totowa, NJ: Humana Press.

Trachte G (2005) Neuronal regulation and function of natriuretic peptide receptor C. *Peptides* 26: 1060–1067.

Attention and Eye Movements

E Kowler, Rutgers University, Piscataway, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Eye movements are inextricably linked to visual attention because both are the principal tools available for selecting interesting portions of visual scenes for enhanced perceptual and cognitive processing. Selection is crucial for animals that rely on foveal images for making decisions about the visual environment or for guiding action. As a result, most of what is known about eye movements and attention derives from studies of human and nonhuman primates.

Eye movements are often taken to mark the path of attention through a scene, where attention refers to the internal (or covert) distribution of processing resources. Eye movements and attention are assumed to serve useful purposes connected to the visual task, an assumption that has fueled decades of efforts to use eye movements to study how people search, read, study pictures of scenes, or carry out all manner of visually guided actions involving reaching, pointing, manipulating objects, walking, or driving. All these tasks require serial processing of different parts of a scene, and all routinely call upon eye movements.

For eye movements to be useful, they must be able to bring the line of sight to relevant locations quickly and accurately, and to keep it there, with no deliberate effort or laborious decision stages beyond those already incorporated into the task itself. In order to meet these requirements – speed, accuracy, and low cognitive load – eye movements must be sensitive to both the physical structure of the visual array and to the momentary needs of the task. To see why this is the case, consider the two extreme possibilities. On the one hand, if eye movements were completely under the control of the stimulus, with gaze attracted to the most vivid or salient portions in a scene, cognitive load would be minimal, but the eye would continually be dragged off to useless and irrelevant locations. It would be impossible to choose targets according to the needs of the task. (Nevertheless, there are many models based on just this assumption.) On the other hand, too much emphasis on volition, choice, or effort to control gaze would create undue cognitive demands and interfere with the ongoing flow of the task. The effort devoted to eye movement control would be a continual distraction from the job of recognizing objects and making task-related decisions and plans. Over the past decades, oculomotor scientists have struggled to find ways of integrating these two extreme modes of eye movement control. The focus of these attempts, in behavioral, psychophysical and neurophysiological studies, has been the attempt to find and understand the relationship between the control of eye movements and the mechanisms of visual attention.

What Is Attention?

Before considering the relationship between eye movements and attention, it is useful to briefly summarize some aspects of attention as they apply more generally to perception.

In his classical treatment of attention, William James acknowledged that, although everyone knows what attention is ("experience is what I agree to attend to", p. 402) the question of what attention does is difficult to answer. Attention is credited with being able to enhance certain experiences at the expense of others, but at the same time it must do so without distorting the nature of these experiences. Attention, James said, might make it possible to notice a faint sound or light, but paying attention to faint sounds or lights does not make them suddenly appear unusually loud or bright.

Modern researchers grappling with James's paradox have devised a number of clever experimental paradigms to explore the effects of attention on perception in many different contexts. Attention can be summoned to locations or to objects by various sorts of cues or signals, or it can be drawn away by imposition of a competing task. Studies using these techniques have led to broad agreement about several characteristics of attention:

1. Attention can lower detection thresholds and increased perceived contrast, producing improvements resembling those produced by modest increases in the contrast of the stimulus itself. These effects may correlate with attention-induced changes in cell firing patterns (e.g., increased contrast gain) at neural levels as early as V1 or the lateral geniculate nucleus (LGN).
2. Improvements in perceptual thresholds can be produced by paying attention either to a selected location (the 'attentional spotlight'), to a selected feature, or to a selected object. In the case of feature-based attention, the attentional benefits can spread across broad regions of space; in the case of object-based attention, the benefits of attention can extend to the different features (color, size, orientation) making up the attended object.
3. Attention to an object serves to overcome potential perceptual interference from visual objects

nearby. The importance of attention in the presence of competing stimuli has contributed to the view that attention should be viewed as an internal processing resource that can be distributed among features, objects or regions, as needed, to accomplish task goals. The importance of competition for limited resources has also been reflected at the neural level (areas V4 or the middle temporal area (MT), for example) with observations showing that, in the presence of multiple stimuli falling in a cell's receptive field, attention determines the relative contribution of each to the response of the cell.

4. Attention may also control access to limited capacity visual short-term or visual working memory, including the visual memory that preserves information from one fixation to the next.

All four of these characteristics – demonstrated in experiments that took pains to eliminate a role for eye movements, often by restricting stimulus presentations to very brief intervals of time – have influenced thinking and research about the role of attention in oculomotor control and the role played by eye movements and attention in visual processing.

Attention and Gaze Stabilization

One of the most important functions of eye movements is to maintain stable gaze. Even modest rotations of the head, if not compensated by counterrotations of the eye in the orbit, can create sufficient motion of the retinal image to impair visual resolution, carry important visual details away from the fovea, and create the illusion that the environment itself is moving. The oculomotor system is equipped with sophisticated systems that use both visual and vestibular signals to compensate for head movements and maintain stable gaze. These traditional full-field gaze stabilization reflexes, however, are not ideally suited to environments containing multiple objects at various distances and locations, viewed by observers who continually pick up and handle some of these objects, and who continually shift their own position in myriad ways. These situations require the maintenance of a stable gaze on only one thing at a time, and this means overriding the global visual or vestibular gaze stabilization reflexes, or at least restricting the portion of the scene on which they are allowed to operate.

In the early twentieth century, physicist and philosopher Ernst Mach described the situation well. The problem, said Mach, is that the retinal motion generated by head movements and body movements as we move through the environment may "exert a peculiar motor stimulus upon the eye, and draw our attention and our gaze after them" (p. 146). But this attraction can be overridden:

> No special apparatus is necessary for observing the foregoing phenomena. They are to be met with on all hands. I walk forward by a simple act of the will. My legs swing to and fro without my having to attend to them particularly. My eyes fixed steadfastly upon my goal without suffering themselves to be drawn aside by the motion of the retinal images consequent upon progression. All this is brought about by a single act of the will. . . . The same process must also be set up if the eyes are to resist for any length of time the stimulus of a mass of moving objects. (Mach, 1906/1959: 146)

Mach's impression that we can override flows of motion on the retina and maintain stable gaze on chosen targets has been confirmed and extended many times in the modern oculomotor laboratory. It is possible, for example, to maintain gaze on even very small stationary targets superimposed on large and vivid moving backgrounds. It is also possible to smoothly track a target moving across a stationary background (although this has proven to be the more difficult task) and to track a target moving in the presence of a nearby moving nontarget. In the case of the target and nontarget, it may require about 100 ms or so for complete selectivity to be achieved. Observations such as these have been made in many studies since 1930, with different kinds of backgrounds and with fixation targets of different sizes and retinal locations, each time showing that it is possible to maintain fixation on selected targets and override the effects of the more plentiful motion signals originating from the background (**Figure 1**).

The ability to select the target for smooth eye movements has also been demonstrated with superimposed transparent sheets of dots, each moving at a different velocity (**Figure 1(b)**). With such patterns, attending to a discrete location would not be sufficient to select one sheet as the target because dots are superimposed everywhere. Thus, effective selection with transparent motions requires attending to a selected object or surface rather than to a discrete location within the pattern. Interestingly, the selection of one of the transparent surfaces does not disrupt the perceptual impression of transparency, nor does it disrupt perceptual interactions between the superimposed fields. For example, if one of the fields is moving and the other is stationary, the stationary field seems to be moving opposite in direction to the other, ignored moving field. The percept of induced motion shows that the selection of the target for the eye movements leaves intact basic operations of sensory and perceptual motion processing. This raises the possibility that the selective mechanism that determines the target for

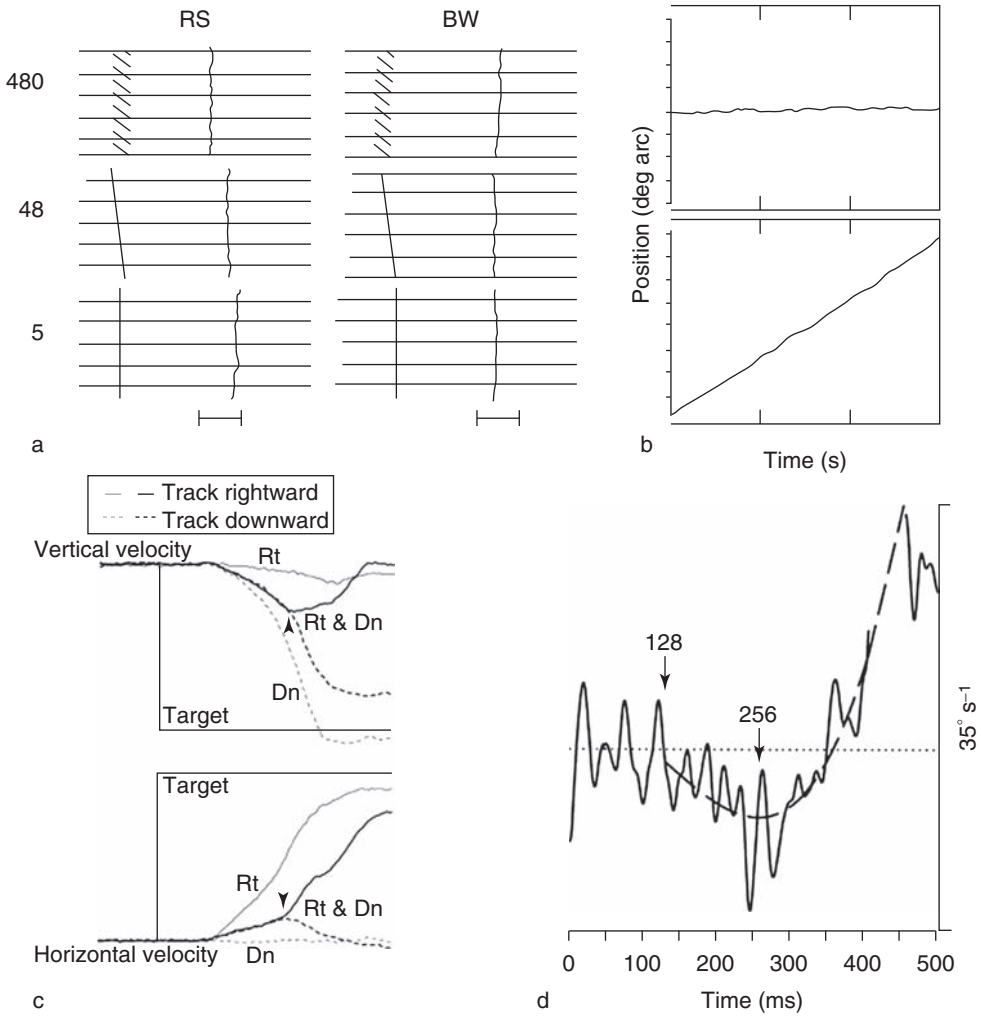

Figure 1 Smooth eye movements and attention: (a) examples of horizontal smooth eye movements maintaining a stable line of sight on a stationary fixation target superimposed on a background moving grating; (b) example of the independence of smooth eye movements from background motion with stationary and moving dots; (c) average eye velocity in response to a single target moving either downward (top) or rightward (bottom) or to a pair of targets, one moving rightward and the other downward; (d) horizontal eye velocity in the presence of two visual arrays of noise elements moving in opposite directions, one located above and the other below the line of sight. In (a), the data are from two observers, RS and BW. The left trace in each panel shows the stimulus motion (5, 48, or 480 min arc s^{-1}); the right trace shows horizontal eye movements. Horizontal lines are 1 s time markers, with time beginning at the bottom. The horizontal bar below the panels for each observer represents a 1° eye rotation. In (b), each trace shows horizontal smooth eye movements with two full-field superimposed arrays of random dots, one array stationary and the other moving at 1.2° s^{-1}. In the upper panel, the observer was told to fixate the stationary array and, in the lower panel, to pursue the moving array. There was virtually no effect of the unattended field on performance. In (c) average eye velocity is in response to a single target moving either downward (top) or rightward (bottom) (thin eye traces, labeled Rt or Dn) or to a pair of targets, one moving rightward and the other downward (bold traces labeled Rt and Dn). Traces labeled Target shows the stimulus which traveled over 20° for 400 ms. After 150 ms, one of the target motions of the pair disappeared, leaving only a single stimulus moving downward (top) or rightward (bottom). Eye velocity when both target motions were present was the average of the two target velocities until about 80 ms after one of the motions was removed (marked by the arrows) when the traces diverged. In (d), the eye trace shows the change in pursuit direction when first one, and then the other, moving field is selected as the pursuit target. (a) Reproduced from Murphy BJ, Kowler E, and Steinman RM (1975) Slow oculomotor control in the presence of moving backgrounds. *Vision Research* 15: 1263–1268, with permission from Elsevier. (b) Reproduced from Kowler E, Van der Steen J, Tamminga EP, and Collewijn H (1984) Voluntary selection of the target for smooth eye movements in the presence of superimposed, full-field stationary and moving stimuli. *Vision Research* 24: 1789–1798, with permission from Elsevier. (c) From Lisberger SG and Ferrera VP (1997) Vector averaging for smooth pursuit eye movements initiated by two moving targets in monkeys. *Journal of Neuroscience* 17: 7490–7502. (d) From Liston D and Krauzlis RJ (2003) Shared response preparation for pursuit and saccadic eye movements. *Journal of Neuroscience* 23: 11304–11314.

smooth eye movements may use its own dedicated attentional filter whose operation has minimal consequences for perception.

The Role of Attention in Perception versus the Role of Attention in Smooth Eye Movements

One way to find out whether smooth eye movements and perception share a common attentional mechanism is to expand the eye movement experiment to include a concurrent perceptual task. In one representative situation, an observer is presented with two sets of moving targets and asked to smoothly track one set and ignore the other (Figure 2). In this situation, perceptual judgments made about the tracked target are more accurate than those made about the untracked background (even after taking into account any differences in retinal position and retinal velocity), implying that a single filter and a single attentional decision determine the strength of signals reaching both perceptual and oculomotor systems. But there are important differences between eye movements and perception. Whereas the smooth tracking eye movements are nearly perfect, with virtually no effect of the unselected background motion, perceptual judgments about the unselected background are above chance, showing that some perceptual registration of the background remains. Thus, there is an asymmetry between the effects of attention on perception and eye movements. Analogous results have been obtained from neurons in the extrastriate cortical areas MT and medial superior temporal region (MST), the presumed sources of the motion signals that guide smooth eye movements and serve perception. Neurons in these areas respond more vigorously to attended and tracked motion than to unattended untracked motion, but the differences in firing patterns are small and do not explain the high degree of selectivity that can be demonstrated in smooth pursuit eye movements (although they are in general agreement with the magnitude of selectivity observed

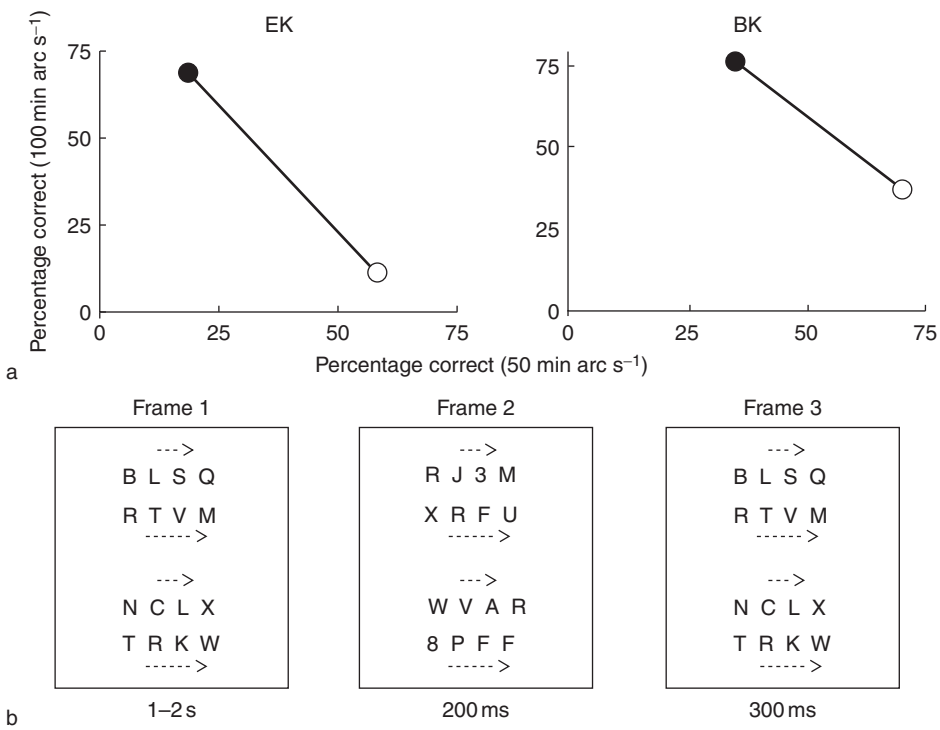

Figure 2 Smooth eye movements with a concurrent perceptual task: (a) perceptual performance for two observers; (b) representative sequence of display frames for a trial. The three display frames in (b) show four rows of letters moving rightward with the top and third rows moving at half the velocity of the second and bottom rows. Letters remained within the confines of the frame so that as portions reached an invisible boundary on the right they re-appeared immediately on the left side of the display. Vertical eye position remained in the gap between rows 2 and 3 while subjects attended and tracked either the faster or slower pair of rows. Frames 1 and 3 are masks. Frame 2 is the critical display, containing two numerals that had to be identified at the end of the trial. Tracking gains were >0.8. The perceptual performance shown in (a) for two observers depended on instructions. When observers attended and tracked the slower pair of rows (open symbols), identification accuracy was better for the slower rows than the faster rows. When they attended and tracked the faster pair of rows (closed symbols), identification improved for the faster pair at the expense of the slower pairs. Separate analyses showed that these results were not due to differences in retinal speed of tracked and untracked rows and were due solely to attention. Reproduced from Khurana B and Kowler E (1987) Shared attentional control of smooth eye movements and perception. *Vision Research* 27: 1603–1618, with permission from Elsevier.

at the perceptual level). The observed differences between the effects of attention on eye movements and perception, and the small effects observed in the motion areas MT and MST during the selective tracking tasks, imply that a given distribution of attention across target and background has different consequences for perception and for eye movements. This rules out early selection models, in which attention operates primarily at an early sensory level of visual processing shared by perception and eye movements.

The ability to eliminate almost all influence of the unselected motion signals from the eye movements suggests the involvement of a winner-take-all (WTA) network operating on the set of attentionally weighted motion signals emerging from areas MT and MST. WTA networks are often used in models of attention or perception and are effective ways to remove weak, and potentially interfering, signals. Although such a network operating downstream from sensory motion areas could account for the perfect or near-perfect selectivity observed for eye-tracking performance, an unconstrained WTA network places a strict limit on the options. It would not, for example, allow people to divide attention between a target and background and to track at a velocity intermediate between the two. An alternative to the WTA network is an executive controller that determines the relative weights assigned to different sets of sensory motion signals. A single executive controller could be used by both perceptual and oculomotor systems with different effects on each. For example, assigning maximum weight to a target could completely attenuate the representation of nontargets in neural areas devoted to smooth tracking and only partially attenuate the nontarget representation in perceptual areas. Different assignments of weight would produce different performance outcomes. Given that the neural circuitry that controls smooth eye movements includes several high-level cortical areas (frontal eye field (FEF), supplementary eye field (SEF), and lateral intraparietal area (LIP)) that also play roles in perceptual attention, there is considerable opportunity for executive intervention in the attentional weighting of motion signals from target and background.

Summary: Attention and Gaze Stabilization

Selective attention plays a crucial role in the control of smooth eye movements by ensuring that stable gaze can be maintained on either stationary or moving targets, regardless of the signals present in the background. Attention ensures that eye movements are controlled only by signals originating from the selected object. Performance of concurrent oculomotor and perceptual tasks have shown that a single attentional decision affects both – it is not possible to maintain gaze on one object while fully attending another. Unselected objects nevertheless remain perceptible and continue to generate neural signals in areas sensitive to motion. This outcome rules out early selection models of attentional control and raises still unanswered questions of how a common attentional decision can have different consequences for smooth eye movements and for perception.

Attention and Saccadic Shifts of Gaze

Saccadic eye movements are used to inspect the visual environment, with saccadic shifts of gaze occurring anywhere from once every several seconds to several times a second. Visual processing occurs during the pauses between saccades. Unlike smooth eye movements, which cannot be generated without some sort of representation of smooth motion, saccadic eye movements can be made at will to look at arbitrary locations.

Saccades are typically directed to useful or informative locations and are often taken to disclose the path of attention during performance of natural tasks. Attention, however, need not be locked to the fovea even as the eye jumps from place to place. Attention to locations remote from the line of sight is valuable for broad surveys of the scene and is indispensable for ensuring that saccades land on the selected object, regardless of the visual structure of the immediate surround.

Eye Position as an Overt Marker of the Locus of Attention

Figure 3 shows examples of saccadic patterns made during various visual tasks: reading, counting, geometry problem-solving, and visual scene recall. Figure 4 shows eye movements during visual search. Figure 5 replicates and extends the well-known experiment described by Yarbus in his 1967 book *Eye Movements and Vision*, in which people were asked to inspect the same painting in order to answer different questions. The scan patterns differed depending on the question, showing that the physical features of the pictures do not by themselves determine the landing locations of the eye but that there is considerable influence of purpose and intent. Figure 6 looks at this issue from another angle, showing scan patterns obtained when looking at two versions of the same photograph, where in one version the image has been filtered and the fine detail removed. Although the task is the same in each, the scatter of the fixated locations is different.

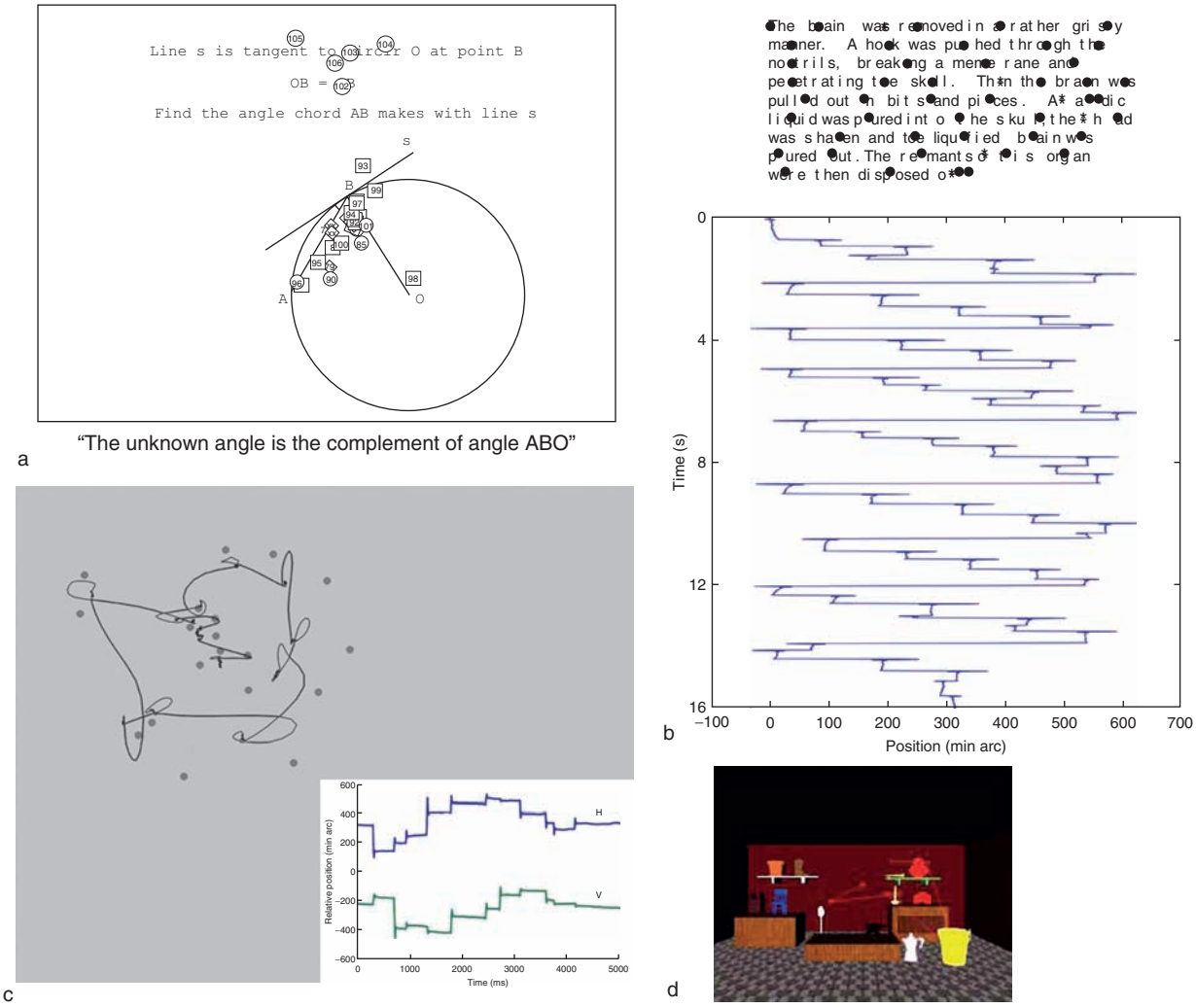

Figure 3 Examples of eye movements during various visual and cognitive tasks: (a) fixation positions of one observer during one stage of solving a problem in geometry; (b) eye movements during reading; (c) eye movements while counting an array of dots; (d) eye movements during a visual recall task. In (a), the problem is presented above the diagram and the excerpt from the protocol is below it. Important features of the diagram were refixated at regular intervals. Analyses of the timing and locations of fixations showed frequent revisits, consistent with a working memory capacity of five visual features of the problem. In (b), the same pattern of eye movements is shown as individual fixation locations superimposed on the text (top) and as a trace of eye positions over time (bottom). Eye position is on the abscissa, and time moves from top to bottom. These eye movements are typical of reading in that saccades were about seven characters long and intersaccadic pauses were about 275 ms. More than 40 readings of this same text showed little change in the pattern of eye movements and a high level of consistency in the distribution of landing locations. In (c), eye movements are shown both superimposed on the dot array and also as horizontal and vertical traces over time. In (d), subjects had to report the contents of the picture after periods of scanning ranging up to 4 s. Analyses showed that memory for the content of the pictures was preserved across successive viewings separated by several minutes and several intervening viewed scenes. (a) Reproduced from Epelboim J and Suppes P (2001) Eye movements during problem solving in geometry. *Vision Research* 41: 1561–1574, with permission from Elsevier. (b) Reproduced from Schnitzer BS and Kowler E (2006) Eye movements during multiple readings of the same text. *Vision Research* 46: 1611–1632, with permission from Elsevier. (d) Reproduced from Melcher D and Kowler E (2001) Visual scene memory and the guidance of saccadic eye movements. Vision Research, 41: 3597–3611, with permission from Elsevier.

Similar efforts to document sequences of fixations in different tasks have been made in more complex and dynamic situations, in which observers move about the environment, handling various objects for some specific purpose. The sequences of fixation positions of the two eyes are required when observers perform three different tasks: tapping a set of rods in prescribed order, grooming a fellow primate, and assembling a doll.

Data such as these have always been intriguing. They prove that ordinary visual experience, which seems smooth, continuous, and flowing, is really made up of sequences of brief, still frames and snapshots, whose detailed contents escape awareness. We do not notice

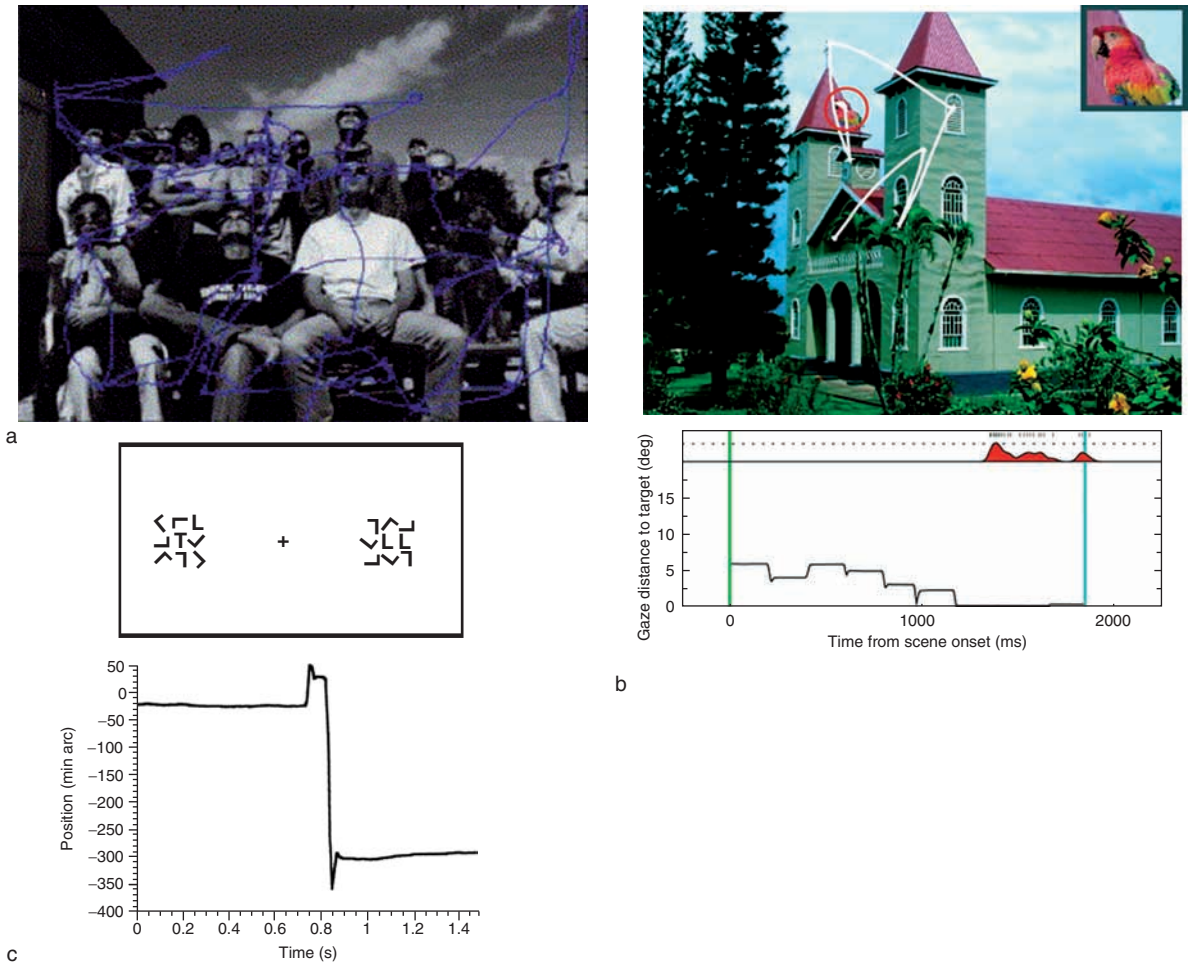

Figure 4 Eye movements during visual searche: (a) search for small gray crosses superimposed on the scene; (b) search in monkey for a small embedded target (reproduced in inset); (c) search for a target letter T embedded in one of two clusters of L's. In (a), analyses showed little or no bias to avoid returning to previously fixated location. In (b), eye traces show the distance to the target over time. Spike density functions show little response until just before the animal fixates the hidden target. Selection of the target as a saccadic goal was shown to be a necessary condition for the firing bursts. In (c), the probability that one of the clusters contained the T (0.8/0.2) was disclosed by a brightness cue. The eye traces show that often subjects looked first at the cluster with the lower probability, particularly when the cluster was closer to the line of sight, and then quickly with an unusually short latency interval (<100 ms) made a second saccade to the high probability cluster containing the target letter. (a) Reproduced from Hooge ITHC, Over EAB, van Wezel RJA, and Frens MA (2005) Inhibition of return is not a foraging facilitator in saccadic search and free viewing. *Vision Research* 45: 1901–1908, with permission from Elsevier. (b) Reproduced from Sheinberg DL and Logothetis NK (2001) Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *Journal of Neuroscience* 21: 1340–1350, with permission from the Society for Neuroscience (c) Reproduced from Araujo C, Kowler E, and Pavel M (2001) Eye movements during visual search: The costs of choosing the optimal path. *Vision Research* 41: 3613–3625, with permission from Elsevier.

each individual glance because we remember only what is most essential – the cumulative number of dots counted or the meaning of a word or phrase – or we use what we see to guide an immediate action. Observing sequences of eye movements reveals the sequential nature of vision and provides a view of underlying goals and strategies that would be hard to recover through other means.

Decisions, Attention, and Saliency Maps

Understanding the rules that govern saccadic planning in everyday life tasks is an enormous undertaking that must be based on models of the task, combined with knowledge of the relevant visual, cognitive, and motor capacities that limit performance and affect the strategies adopted. Although we still lack comprehensive and general models, progress has been made toward the goal. One of the influential constructs developed to account for the eye movements made during scanning of a scene is the salience map. As originally proposed by Koch and Ullman, the salience map is derived from the output of banks of visual filters, modeled after those present in early levels of the

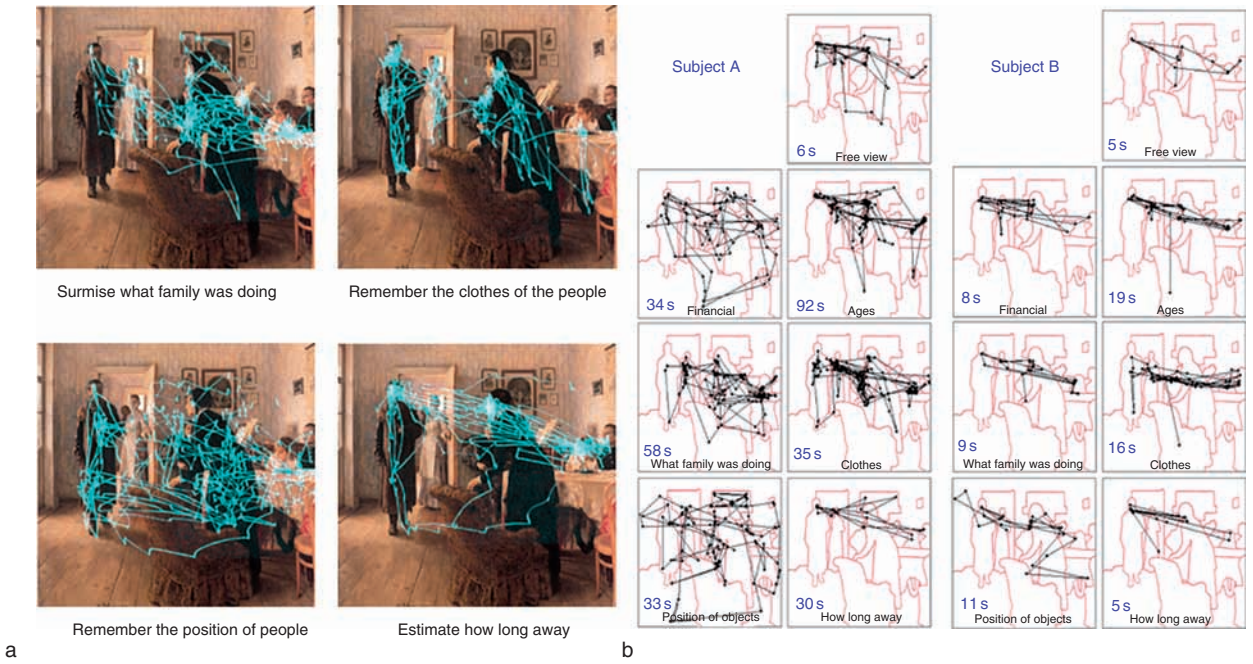

Figure 5 Effect of task motivation on eye movements: (a) people looking at Repin's painting *The Unexpected Visitor* altered the choice of fixation positions depending on the question they had to answer; (b) a replication of Yarbus's experiment, in which subject A shows the same sensitivity to task as Yarbus demonstrated but subject B tends to look in the same places regardless of the task. (a) Adapted from Yarbus A (1967) *Eye Movements and Vision.* New York: Plenum. (b) Figure provided by J Pelz and reproduced with permission. For discussion see Lipps M and Pelz JB (2004) Yarbus revisited: Task-dependent oculomotor behavior [abstract]. *Journal of Vision* 4: 115a, with permission from ARVO.

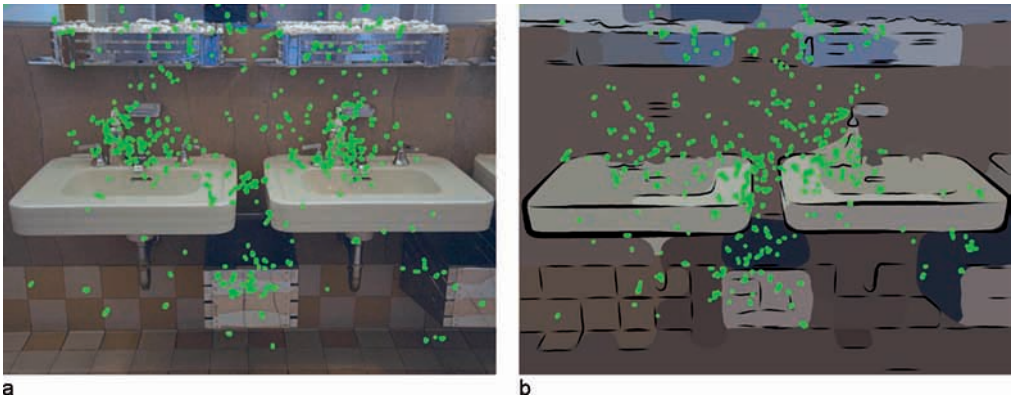

Figure 6 Fixations made by 10 viewers while looking at two variants of the same photograph: (a) original photo; (b) version in which computer graphics techniques were used to create a drawing-like version of the scene and remove detail throughout the image. Subjects were instructed to remember the picture for a subsequent recognition task. In (a), fixations are tightly clustered in a few key areas. In (b), fixations are more uniformly distributed. Figures provided by A Santella; original photo courtesy of Philip Green spun. For discussion see DeCarlo D and Santella A (2002) Stylization and abstraction of photographs. *ACM Transactions on Graphics* 21: 769–776; Santella A and DeCarlo D (2004) Visual interest and NPR: An evaluation and manifesto. In: *Proceedings of the Third International Symposium on Nonphotorealistic Animation and Rendering*, pp. 71–78. Annecy, France, 7–9 June. New York: ACM Press.

visual system (V1). These filters respond on the basis of local feature contrast (luminance, color, or orientation contrast) present in the scene. Models for the generation of salience maps integrate all three types of feature contrasts, taking the effects of retinal eccentricity into account. Saliency maps have been used to predict where people direct saccades and how they distribute visual attention (i.e., processing capacity) across a scene even when eye movements are not made.

In order to use a global salience map to predict a sequence of fixations, two additional assumptions have to be made. The first is that a WTA network (similar to that already discussed in the context of smooth eye movements) establishes the ordering of the fixations, with the region with the highest salience

level designated as the next one to be fixated. The second assumption is that the salience map is subjected to inhibition of return; that is, once a location is examined its strength in the salience map will abruptly diminish, thus allowing the line of sight to proceed to new places and avoiding the need to store a list of already-scanned locations. Armed with a salience map and these two supporting assumptions, it has been possible to generate precise predictions about where people look during unconstrained free-viewing. The predictions have proven to be reasonably accurate, at least when computed salience is compared to the global aggregate of fixated locations over some period of inspection. Nevertheless, the approach has not been without controversy.

The main criticism of the salience map approach is the obvious one, recognized and discussed even by proponents of the approach. Strict control of eye movements or attention based solely on physical salience becomes counterproductive as soon as people are faced with real-world tasks (Figures 4–6). In real-world tasks, people look at the objects and details that are important, regardless of the physical salience levels. Gaze and attention may be drawn to novel or unexpected objects or to highly familiar objects, depending on the situation. None of these characteristics is represented by physical salience. Physical salience also does not capture the long-term changes in gaze preferences that develop as the contents of the scene are learned, nor does it capture effects of motivation. Do you look at something because it attracts your interest, because it contains some new information you need to perform a task, or because you need to confirm a detail that you already know is present? All of these so-called top-down variables that influence gaze and attention can be viewed as displacing the concept of a saliency map or, alternatively, as imposing an additional set of weights on the physical salience levels.

Neurophysiological studies using a variety of active saccadic tasks involving visual search or visual discriminations have addressed a portion of the problem of incorporating top-down variables by showing that neurons in areas related to saccadic planning (LIP, FEF, and superior colliculus (SC)) may be able to encode the top-down salience level at a given location. The evidence for this assertion is that neurons in these areas show enhanced activity when objects in their receptive fields contain information relevant to the guidance of an upcoming saccade, even when the location of enhanced activity does not correspond to the location of the saccadic target. Visual areas (the V4 or inferotemporal visual area (IT), for example), also show sensitivity to top-down salience. In these areas, neurons tuned to specific features show effects of attention, as well as patterns of presaccadic enhancement. Thus, there has been considerable progress in identifying brain areas that modulate responses to objects depending on their immediate behavioral significance, and these areas may constitute the core of a network that directs attention and saccades to objects or locations relevant to a task.

Incorporating top-down factors and physical salience levels into a single map may be a physiologically plausible means of deriving a single message to guide saccades and attention. Alternatively, there may be multiple maps that compete for control. Studies of eye movements during visual search have provided some support for the notion of competing maps. In these studies, saccades are drawn to physically salient locations that are not likely to contain the search target, followed quickly by corrective saccades to more appropriate, but less salient, places (Figure 4(c)). These rapid saccadic sequences could represent the results of the competition between different maps for momentary control of the neural centers that produce saccades. Although such competition seems inefficient, the amount of time lost by an occasional errant saccade may be small enough that an investment of time and resources in more prudent saccadic planning may not be warranted.

Saccades and Attention

Saliency – either stimulus-driven or top-down – is said to predict the distribution of both attention (internal processing resources) and saccades over time and space. Is there any important or necessary distinction between the distribution of attention and saccades? According to premotor theories of attention, saccadic eye movements and shifts of attention are essentially the same process, originating in the same neural areas, with an attention shift occurring only when the saccade is suppressed or inhibited. The premotor theory is supported by finding several neural areas (SC and FEF) whose activity can evoke or is correlated with both attention shifts and saccades. Nevertheless, it is troubling for premotor theories that attention can be distributed in parallel across the visual field or aligned symmetrically about the line of sight; spatial patterns do not readily map on to trajectories of saccades.

If we view attention not as a subthreshold saccade but rather as an internal processing resource that influences both perception and motor control, we can then ask how attention and saccades interact. Attention plays an important role in saccadic guidance in that, although it is possible to shift attention independently of the saccade, it does not appear possible to plan and carry out an accurate saccade without shifting some attention to the saccadic goal.

Dual-task studies, similar to those described earlier for smooth eye movements, have examined the distribution of visual attention prior to saccades, using various psychophysical tasks as indicators of local attentional strength. These dual-task studies require observers to do two things at once (in as brief a time as possible), namely, prepare to look at a specified object and identify a perceptual target presented while the preparation of the saccade is in progress. The perceptual target appears either at the saccadic goal or elsewhere. These experiments can succeed only if care is taken that the two tasks are done in the same brief interval of time because otherwise observers can decide to delay saccades in order to improve perceptual performance.

Dual-task experiments have shown that perceptual performance is better at the saccadic goal than elsewhere (see illustrative study in **Figure 7**), showing that saccades and attention do not functional independently. Yet, as was the case with smooth eye movements, perceptual performance at nongoal locations remained well above chance even when efforts were made to produce the saccades as quickly and accurately as possible (i.e., with the same latency and accuracy as observed with no concurrent perceptual task). Moreover, it only required a modest sacrifice in saccadic performance (10–20% latency increases) to achieve pronounced improvement in perceptual performance at nongoal locations.

Similar connections between attention and saccades are found during the execution of sequences of saccades (**Figure 8(a)**). Best perceptual performance is found at the saccadic goal, with elevations in perceptual thresholds at other locations, the amount of elevation depending on when and where the perceptual target appears. The pattern of activity observed during saccadic sequences suggests that the reduction in the attentional levels at nongoal locations is important for providing sufficient attentional contrast to guide an accurate saccade and prevent the line of sight from landing at salient, but incorrect, locations during scanning tasks. Thus, the main role of top-down attention in saccadic guidance may be to suppress the attraction of physically salient, but irrelevant, locations.

When the scan path to be followed is marked by a feature difference, rather than being followed from memory, the distribution of attention changes significantly. Attention spreads to locations along the feature-cued path, including targets of subsequent, as well as previous saccades (feature-based attention) (**Figure 8 (b)**). An interesting question raised by these results is how the saccades remain accurate when so many locations are attended at once. As was the case with smooth eye movements, there are two main alternatives. The first is a WTA network that directs the eye to the location where attentional strength is greatest. The second alternative requires the intervention of a central executive controller that imposes its own attentional weights. The controller could, for example, focus all active effortful attention on the upcoming target. The perceptual enhancement observed elsewhere along the featured-cued path could result from a passive process that facilitates the spread of attention across space to the selected features.

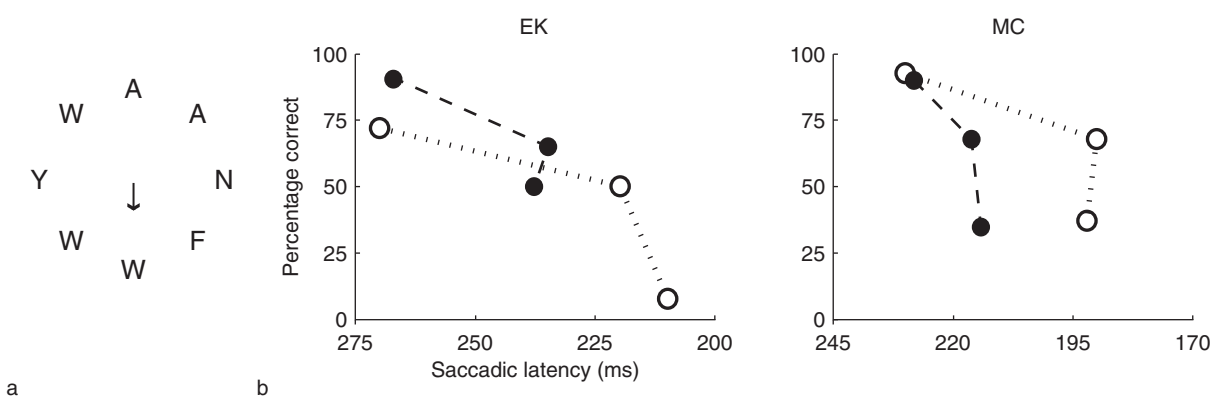

Figure 7 Saccades and attention: (a) the task display; (b) attentional-operating characteristics for two subjects showing the trade-off of saccadic and perceptual performance under two different conditions: saccade to a randomly chosen location (filled circles) and saccade to the same location on each trial (open circles). Shown in (a), the display contained eight randomly chosen letters. The task was to make a saccade (downward, in this case) while simultaneously identifying the letter in the right-hand position. Display timing was adjusted so that the critical letters appeared only during the saccadic preparation interval. In (b), The three data points in each function show performance under three instructions: (1) minimize saccadic latency (lowermost symbols), (2) sacrifice latency for improved perceptual accuracy (uppermost symbols), and (3) perform at a level intermediate between these two extremes. Achieving best perceptual identification accuracy required a sacrifice in saccadic latency (analysis of saccadic precision showed analogous results). Perceptual identification reached near-perfect levels with latency increases of only 10–20%. Reproduced from Kowler E, Anderson E, Dosher B, and Blaser E (1995) The role of attention in the programming of saccades. *Vision Research* 35: 1897–1916, with permission from Elsevier.

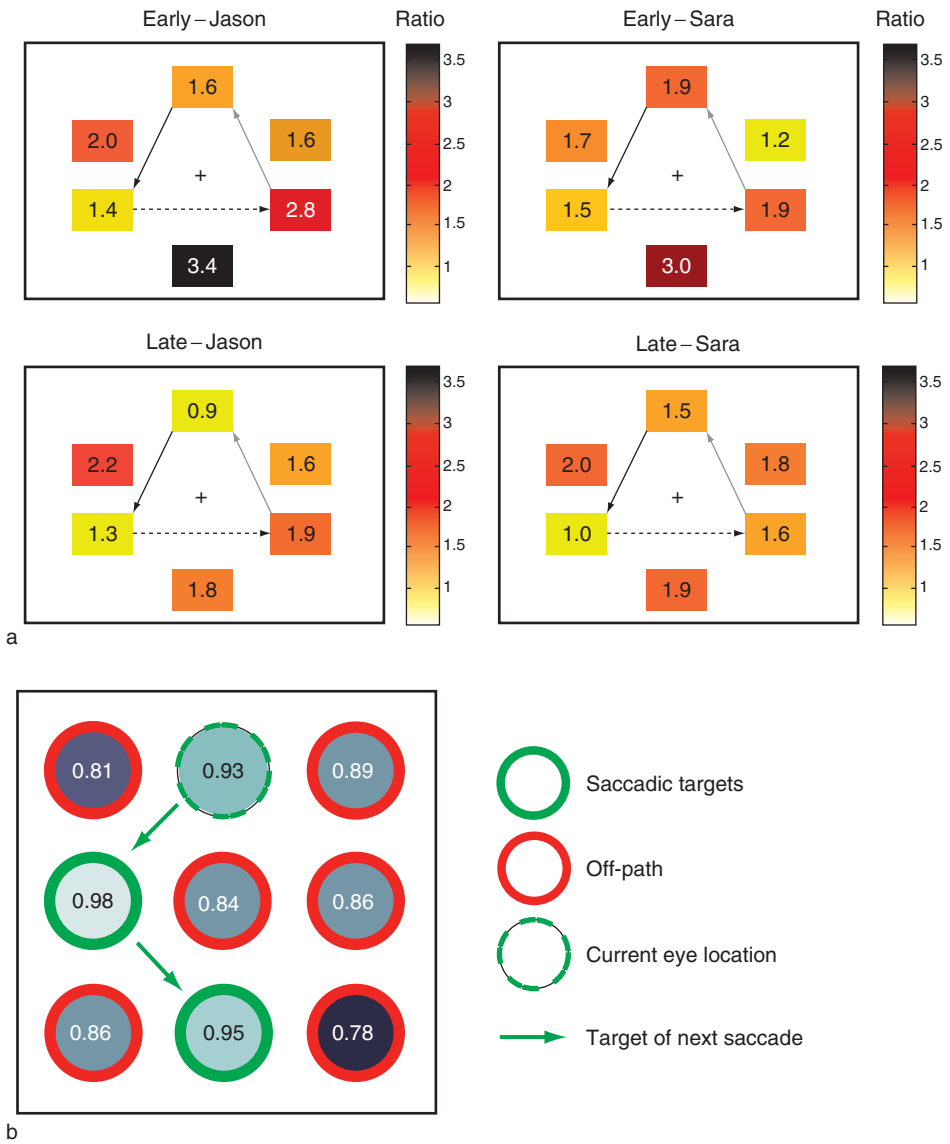

Figure 8 Attention during performance of sequences of saccades: (a) repetitive sequences of saccades; (b) nonrepetitive sequences of saccades. In (a), two subjects made saccades to rectangular boxes following a triangular pattern (shown by arrows) in which saccades were made to every other box. During a randomly selected fixation pause, an oriented Gabor patch was shown in one of the boxes either during the early (top row) or later (bottom row) portion of the pause. Current fixation position is in the top box. The colors of the boxes and the numbers in the boxes show the ratios of contrast threshold for correct orientation identification relative to thresholds during steady maintained fixation. Thresholds were significantly elevated at all locations except the currently fixated position (top of each graph) and target of the next saccade (lower left box in each graph). In (b), arrows represent a segment of a saccadic path. Current eye position is the top circle (dotted outline). Shading and numbers in the circles show percentage of correct orientation identification (two alternative forced-choice) for a Gabor patch shown briefly in one of the circles during a randomly selected intersaccadic pause. Perceptual performance is best for locations on the saccadic path, including the targets of the next two saccades. Performance for on-path locations is better than for off-path locations, including off-path locations of equal or smaller retinal eccentricity. (a) Reproduced from Gersch TM, Kowler E, and Dosher B (2004) Dynamic allocation of visual attention during the execution of sequences of saccades. *Vision Research* 44: 1469–1483, with permission from Elsevier (b) From Gersch TM (2007) *Attentional Filter Used during Scanning Influent Both Perception and Memory*. Doctoral thesis, Rutgers University.

Summary: Attention and Saccades

The picture of saccades and attention that emerges from these psychophysical and behavioral studies, which is in agreement with neurophysiological work, is that saccades and attention normally operate in a connected fashion, with saccades landing at the object that is the main focus of attention. Nevertheless, attending to the goal of saccades still allows significant perceptual processing at different locations, perhaps with the aid of other mechanisms that passively

distribute attention to relevant locations in parallel across the visual array.

Overall Summary and Conclusion

The connection between eye movements and attention described in this article gives some support for the prevalent assumption that eye movements can provide reliable indicators of the main focus of attention in active visual tasks, such as reading, search, or object manipulation. Attention plays a crucial role in oculomotor control by enhancing signals from selected targets relative to those in the background, thus ensuring that eye movements are planned and programmed on the basis of selected visual signals, regardless of the visual configuration of objects nearby. Following the classical treatment of William James, it is useful to distinguish between the distribution of attention and the effects of attention. Attending to a given target can completely eliminate the influence of visual backgrounds on eye movements (smooth or saccadic) while continuing to allow perceptual processing, albeit at reduced levels. This asymmetry between perception and eye movements has the desirable consequence of allowing eye movements to be accurate without disrupting global perceptual analysis across the visual field.

See also: Attentional Functions in Learning and Memory; Contextual Interactions in Visual Perception; Perception and Eye Movements; Saccades and Visual Search; Saccadic Eye Movements; Sensorimotor Integration: Attention and the Premotor Theory; Vision for Action and Perception; Visual Attention.

Further Reading

Araujo C, Kowler E, and Pavel M (2001) Eye movements during visual search: The costs of choosing the optimal path. *Vision Research* 41: 3613–3625.

Corbetta M, Akbudak E, Conturo TE, et al. (1998) A common network of functioning areas for attention and eye movements. *Neuron* 21: 761–773.

DeCarlo D and Santella A (2002) Stylization and abstraction of photographs. *ACM Transactions on Graphics* 21: 769–776.

Deubel H and Schneider WX (1996) Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research* 36: 1827–1837.

Epelboim J and Suppes P (2001) Eye movements during problem solving in geometry. *Vision Research* 41: 1561–1574.

Gardner JL and Lisberger SG (2001) Linked target selection for saccadic and smooth pursuit eye movements. *Journal of Neuroscience* 21: 2075–2084.

Gersch TM (2007) *The Attentional Filter Used during Scanning Influences Both Perception and Memory.* Doctoral Thesis, Rutgers University.

Gersch T, Kowler E, and Dosher B (2004) Dynamic allocation of attention during sequences of saccades. *Vision Research* 44: 1469–1483.

Hayhoe M and Ballard D (2005) Eye movements in natural behavior. *Trends in Cognitive Science* 9: 188–194.

Hooge ITHC, Over EAB, van Wezel RJA, and Frens MA (2005) Inhibition of return is not a foraging facilitator in saccadic search and free viewing. *Vision Research* 45: 1901–1908.

James W (1890/1950) *Principles of Psychology*, Vol. 1, p. 402. New York: Dover.

Khurana B and Kowler E (1987) Shared attentional control of smooth eye movements and perception. *Vision Research* 27: 1603–1618.

Koch C and Ullman S (1985) Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology* 4: 219–227.

Kowler E (1990) The role of visual and cognitive processes in the control of eye movement. In: Kowler E (ed.) *Eye Movements and Their Role in Visual and Cognitive Processes*, pp. 1–70. Amsterdam: Elsevier.

Kowler E, Anderson E, Dosher B, and Blaser E (1995) The role of attention in the programming of saccades. *Vision Research* 35: 1897–1916.

Kowler E, Van der Steen J, Tamminga EP, and Collewijn H (1984) Voluntary selection of the target for smooth eye movements in the presence of superimposed, full-field stationary and moving stimuli. *Vision Research* 24: 1789–1798.

Lipps M and Pelz JB (2004) Yarbus revisited: Task-dependent oculomotor behavior [abstract]. *Journal of Vision* 4: 115a.

Lisberger SG and Ferrera VP (1997) Vector averaging for smooth pursuit eye movements initiated by two moving targets in monkeys. *Journal of Neuroscience* 17: 7490–7502.

Liston D and Krauzlis RJ (2003) Shared response preparation for pursuit and saccadic eye movements. *Journal of Neuroscience* 23: 11305–11314.

Mach E (1906/1950) *Principles of Psychology, Vol. 1, The Analysis of Sensations*. New York: Dover.

Melcher D and Kowler E (2001) Visual scene memory and the guidance of saccadic eye movements. *Vision Reearch* 41: 3597–3611.

Murphy BJ, Kowler E, and Steinman RM (1975) Slow oculomotor control in the presence of moving backgrounds. *Vision Research* 15: 1263–1268.

Santella A and DeCarlo D (2004) Visual interest and NPR: An evaluation and manifesto. In: *Proceedings of the Third International Symposium on Nonphotorealistic Animation and Rendering*, pp. 71–78. Annecy, France, 7–9. June New York: ACM Press.

Schnitzer BS and Kowler E (2006) Eye movements during multiple readings of the same text. *Vision Research* 46: 1611–1632.

Sheinberg DL and Logothetis NK (2001) Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *Journal of Neuroscience* 21: 1340–1350.

Sperling G and Dosher BA (1986) Strategy and optimization in human information processing. In: Boff KR, Kaufman L, and Thomas JP (eds.) *Handbook of Perception and Human Performance Vol.1: Sensory Processes and Perception.* New York: Wiley.

Steinman RM, Menezes W, and Herst AN (2005) Handling real forms in real life. In: Jenkin MRM and Harris LR (eds.) *Seeing Spatial Form*, pp. 187–212. New York: Oxford University Press.

Yarbus A (1967) *Eye Movements and Vision*. New York: Plenum Press.

Relevant Website

http://www.philip.greenspun.com – Philip Greenspun.

Attention Deficit Hyperactivity Disorder

L Tamm, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Attention-deficit/hyperactivity disorder (ADHD) is the most prevalent disorder in childhood. Children with ADHD demonstrate age-inappropriate difficulties maintaining attention and concentration, as well as hyperactivity and impulsivity. These deficits occur in multiple settings and result in impairments in academic performance and social functioning. In approximately 60% of children with ADHD, symptoms persist into adolescence and may even continue into adulthood. Negative sequelae of ADHD include academic underachievement, substance abuse, mood disorders, antisocial behavior, and higher rates of accidents and hospitalizations. This article provides an overview of ADHD, covering diagnosis, prevalence, neurobiology and etiology, and current empirically supported treatments.

Diagnosis

All children are sometimes restless, sometimes act without thinking, and sometimes daydream. Because everyone shows these behaviors at one time or another, the guidelines for determining whether a person has ADHD are very specific. When hyperactivity, distractibility, poor concentration, or impulsivity begin to affect performance in school, social relationships with other children, or behavior at home (i.e., cause functional impairment), ADHD may be suspected.

While there are no biological, physiological, or genetic markers or independently valid tests that can reliably identify the disorder, the scientific consensus is that a diagnosis of ADHD can be made reliably using well-tested diagnostic interview methods. Currently, the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (DSM-IV), is utilized as the basis for diagnostic process in ADHD in the United States. The DSM-IV includes categorizations for Predominantly Inattentive Type, Predominantly Hyperactive – Impulsive Type, and Combined Type within the ADHD diagnosis. Table 1 lists the DSM-IV symptoms for ADHD.

Determining if a child has ADHD is a multifaceted process. Many biological and psychological problems can contribute to symptoms similar to those exhibited by children with ADHD. For example, anxiety, depression, and certain types of learning disabilities may cause similar symptoms. Thus, a comprehensive evaluation is necessary to establish a diagnosis, rule out other causes, and determine the presence or absence of coexisting conditions. Such an evaluation should include a clinical assessment of the individual's academic, social, and emotional functioning and developmental level. A careful history should be taken from the parents, the teachers, and, when appropriate, the child. Checklists for rating ADHD symptoms and ruling out other disabilities are often used by clinicians.

Information about each of the DSM-IV ADHD symptoms is collected to determine presence or absence of the disorder. In addition, the behaviors must create significant difficulty in at least two areas of life, such as home, social settings, school, or work. Symptoms must be present for at least 6 months. Different symptoms may appear in different settings, depending on the demands the situation may pose for the child's self-control. Symptoms of ADHD will appear over the course of many months, often with the symptoms of impulsiveness and hyperactivity preceding those of inattention, which may not emerge for a year or more, or until the child begins school, where demands on attention are increased.

Neuropsychological or psychological assessment of executive functioning is also useful, as children with ADHD typically perform poorly on tests of executive function, including inhibition and working memory tasks. A medical exam by a physician is important and should include a thorough physical examination, including hearing and vision tests, to rule out other medical problems that may be causing symptoms similar to ADHD (e.g., thyroid problems).

In children and teenagers, the symptoms must be more frequent or severe than in other children the same age. In very young children, more stringent criteria may be applied (e.g., increasing the duration of symptoms from 6 to 9 months, considering moderate to severe symptomatology by multiple raters, and using higher ratings of impairment). Diagnosing ADHD in an adult requires an examination of childhood academic and behavioral history as well as reviewing current symptoms. In adults, the symptoms must affect the ability to function in daily life and persist from childhood.

Prevalence

The prevalence rates for ADHD vary depending on diagnostic practices (i.e., structured interview, inclusion

Table 1 DSM-IV criteria for ADHD

I. Either A or B:

A. Six or more of the following symptoms of inattention have been present for at least 6 months, to a point that is disruptive and inappropriate for developmental level:

- Inattention

1. Often does not give close attention to details or makes careless mistakes in schoolwork, work, or other activities.
2. Often has trouble keeping attention on tasks or play activities.
3. Often does not seem to listen when spoken to directly.
4. Often does not follow instructions and fails to finish schoolwork, chores, or duties in the workplace (not due to oppositional behavior or failure to understand instructions).
5. Often has trouble organizing activities.
6. Often avoids, dislikes, or doesn't want to do things that take a lot of mental effort for a long period of time (such as schoolwork or homework).
7. Often loses things needed for tasks and activities (e.g., toys, school assignments, pencils, books, or tools).
8. Is often easily distracted.
9. Is often forgetful in daily activities.

B. Six or more of the following symptoms of hyperactivity-impulsivity have been present for at least 6 months to an extent that is disruptive and inappropriate for developmental level:

- Hyperactivity

1. Often fidgets with hands or feet or squirms in seat.
2. Often gets up from seat when remaining in seat is expected.
3. Often runs about or climbs when and where it is not appropriate (adolescents or adults may feel very restless).
4. Often has trouble playing or enjoying leisure activities quietly.
5. Is often 'on the go' or often acts as if 'driven by a motor.'
6. Often talks excessively.

- Impulsivity

1. Often blurts out answers before questions have been finished.
2. Often has trouble waiting one's turn.
3. Often interrupts or intrudes on others (e.g., butts into conversations or games).
4. Some symptoms that cause impairment were present before age 7 years.
5. Some impairment from the symptoms is present in two or more settings (e.g., at school/work and at home).
6. There must be clear evidence of significant impairment in social, school, or work functioning.
7. The symptoms do not happen only during the course of a pervasive developmental disorder, schizophrenia, or other psychotic disorder. The symptoms are not better accounted for by another mental disorder (e.g., mood disorder, anxiety disorder, dissociative disorder, or a personality disorder).

II. Based on these criteria, three types of ADHD are identified:

ADHD, Combined Type: if both criteria 1A and 1B are met for the past 6 months.

ADHD, Predominantly Inattentive Type: if criterion 1A is met but criterion 1B is not met for the past 6 months.

ADHD, Predominantly Hyperactive–Impulsive Type: if criterion 1B is met but criterion 1A is not met for the past 6 months.

of data from both parents and teachers), sampling techniques, and sampling populations (school, clinic, community). A behavioral definition of ADHD, based on symptoms shown at one point in time (which does not indicate actual psychiatric disorder), is found in 10–20% of the general population in several countries. A psychiatric definition in the DSM tradition, with specific inclusion criteria for symptom onset, duration, pervasiveness, and impairment, is found in 5–10% of the general population, whereas this frequency is 1–2% with the International Classification of Diseases (ICD) tradition, which restricts diagnosis to the full syndrome with limited comorbidity. The prevalence rate for adults is estimated at 4%, for preschoolers, 2–5%; and for the school-aged population, 3–5% using the DSM.

ADHD has been extensively studied in the US, and the predominance of American research in this field and apparent differences in the prevalence of ADHD or hyperkinesis, as defined by ICD, has also led to the impression that ADHD is largely an American disorder. However, a review of the prevalence literature indicates the rate of ADHD is at least as high in many non-US children as in US children, with the highest prevalence rates being seen when using DSM-IV diagnoses, compared with previous versions (DSM-III and DSM-III-R).

With regard to gender, boys are nearly three times as likely as girls to receive a diagnosis of ADHD. With regard to race and ethnicity, prevalence rates seem to be somewhat lower in Hispanic and African-American families compared to White non-Hispanic families, who were almost twice as likely to receive a diagnosis of ADHD. These prevalence differences do not seem to be accounted for by birth weight, income, and insurance coverage.

Neurobiology

The neurobiology of ADHD is not completely understood, however, imbalances in dopaminergic and noradrenergic systems have been implicated in the core symptoms that characterize this disorder. Although there are inconsistencies among studies, there is converging evidence to suggest that deficits in frontal lobe function and the connections between the frontal lobe and key subcortical regions underlie this disorder. Because of the complexity of prefrontal circuitry, it is not yet clear whether the prefrontal abnormalities in ADHD are secondary to abnormalities of the prefrontal cortex or to brain areas with prefrontal projections.

Structural imaging studies, with computerized tomography or magnetic resonance imaging (MRI), suggest structural brain abnormalities among ADHD patients, with the most common findings being smaller whole brain volumes, and smaller gray-matter volumes in frontal cortex, cerebellum, and subcortical structures (e.g., caudate, globus pallidus). These differences have been shown to remain stable with age for ADHD patients independent of medication status.

Functional magnetic resonance imaging (fMRI) studies are consistent with structural studies in implicating the frontosubcortical systems. The caudate, putamen, and globus pallidus subcortical structures are part of the neural circuitry underlying motor control, executive functions, inhibition of behavior, and the modulation of reward pathways. These frontal–striatal–pallidal–thalamic circuits provide feedback to the cortex for the regulation of behavior.

Functional imaging studies also implicate the cerebellum and corpus callosum in the pathophysiology of ADHD. The cerebellum contributes significantly to cognitive functioning, presumably through cerebellar-cortical pathways involving the pons and thalamus. The corpus callosum connects regions of the two cerebral hemispheres. Size variations in the callosum and volume differences in the number of cortical neurons might degrade communication between these two hemispheres, which might account for some of the cognitive and behavioral symptoms of ADHD.

Etiology

No single etiology for ADHD has been determined to date. A vast majority of studies indicate a strong genetic contribution to ADHD, with heritability rates estimated at 0.6 to 0.8 and an approximately fivefold elevated risk for ADHD in first-degree relatives. However, it is clear that other factors contribute. In a diathesis–stress model, a genetic vulnerability or predisposition (diathesis) interacts with the environment and life events (stressors) to trigger behaviors or psychological disorders. A diathesis–stress model probably best explains the etiology of ADHD.

Environmental Agents

Alcohol and cigarette use during pregnancy has been shown to increase risk for ADHD in the offspring of that pregnancy. High levels of lead in the body also seem to be associated with risk for ADHD, particularly in preschoolers. However, lead does not account for the majority of ADHD cases, and many children with high lead exposure do not develop ADHD. Furthermore, recent regulations prohibiting lead in paint have resulted in significantly fewer cases of lead exposure (except for those who live in older buildings).

Food and Food Additives

Popular lore suggests that there may be a relationship between various foods, food additives, and sugar and ADHD symptoms. However, no controlled trial has demonstrated a causal relationship between ADHD and food intake. For example, 12 double-blind, placebo-controlled studies of sugar challenges failed to provide any evidence that sugar ingestion leads to untoward behavior in children with ADHF or in normal children. Likewise, no studies have found any negative effect of candy or chocolate on behavior. Furthermore, diet restrictions were only shown to help approximately 5% of children with ADHD (most of these with food allergies), and sugar restrictions were not shown to have any impact on core ADHD symptoms.

Genetics

Data from family, twin, and adoption studies show that genes play a significant role in the etiology of ADHD. The familial risk of ADHD is two- to eightfold greater in parents and siblings of ADHD probands. Adoption studies have demonstrated that biological relatives of nonadopted children have higher rates of symptoms than do adopted relatives. Furthermore, adopted relatives show a risk similar to that of controls.

Twin studies show perhaps the most compelling data for understanding heritability. Monozygotic twins share 100% of their genes, whereas fraternal twins and other siblings share 50% of their genes. Therefore, heritability can be computed by determining the extent to which monozygotic twins are more concordant for ADHD compared with fraternal twins. A pooled analysis of 20 international twin studies has revealed the heritability of ADHD to be 0.76, thus demonstrating that ADHD is one of the most heritable psychiatric disorders. Although the mode of inheritance of ADHD is unknown, it is likely

to be polygenic based on its modest relative risks and high population prevalence. The individual risk contribution per gene, therefore, may be quite small.

Two approaches are used to evaluate the genetic etiology of ADHD: (1) the genome scan, which examines all chromosomal locations without *a priori* guessing as to which genes underlie ADHD, and (2) the candidate gene approach, which examines one or more genes based on theoretical considerations and empirical evidence from neurobiological studies. Few genome-wide linkage scan studies have been conducted, in part due to the large number of study participants required for sufficient power. The convergence of data from the studies which have been conducted implicate chromosomal regions 5p13, 16p13, and 17p11, although other linkage studies have implicated 7p, 10q26, 12q23, 15q, 15q15, 5p12, and 11q22–25. There are inherent challenges to drawing inferences from few studies with relatively small samples, in which it is likely that the genomic regions suggestive of linkage will differ appreciably across studies for statistical reasons alone. That is, although there may be other reasons for the discrepant findings, such as differences in the populations sampled or in the assessment or diagnostic methods used, the stochastic fluctuations associated with few studies of small sample size are sufficient to cause such discrepancies. Additional research is required to substantiate and replicate these linkages, as well as their contribution to ADHD.

Candidate gene studies have particularly focused on neurotransmitters such as dopamine and norepinephrine. These neurotransmitters are found in abundance in the prefrontal cortex and are thought to mediate executive function and hyperactive/impulsive behaviors. Furthermore, these neurotransmitters are implicated in the mechanism of action in pharmacological agents used to treat ADHD. Thus, genes that produce these neurotransmitters and arbitrate their function are logical candidates for investigation in ADHD probands.

Candidate genes for neurotransmitter systems may include (1) precursor genes that affect the rate at which neurotransmitters are produced from precursor amino acids (e.g., tyrosine hydroxylase for dopamine), (2) receptor genes that are involved in receiving neurotransmitter signals (e.g., genes corresponding to the five dopamine receptors, DRD1, DRD2, DRD3, DRD4, and DRD5), (3) transporter genes that are involved in the reuptake of neurotransmitters back into the presynaptic terminal (e.g., DAT1), (4) metabolite genes that are involved in the metabolism or degradation of these neurotransmitters (e.g., the gene for catechol-*O*-methyl-transferase, COMT), and (5) genes that are responsible for the conversion of one neurotransmitter into another (e.g., dopamine β-hydroxylase, which converts dopamine into norepinephrine). The majority of candidate gene studies in ADHD have implicated dopamine transport (DAT), dopamine β-hydroxylase (DBH), dopamine receptor D4 (DRD4), and dopamine receptor D5 (DRD5) in the pathogenesis of ADHD. By far, the gene most strongly implicated in ADHD is the seven-repeat allele of the DRD4 gene, confirming a strong dopamine component in the pathogenesis of ADHD. In addition, emerging evidence for other candidate genes has implicated serotonin transporter 5-hydroxytryptamine (5-HT), serotonin receptor 1B (HTR1B), and synaptosomal-associated protein 25 (SNAP-25).

Although genetic studies have contributed significantly to the understanding of ADHD, none of the studies accounted for more than 5% of ADHD. It is likely that several gene polymorphisms with moderate to small effect sizes contribute to the phenotype of ADHD; different combinations of such predisposing variants presumably underlie ADHD in different individuals. It is also possible that some of the discrepancies in the genetic studies could be due to differences in geographic, clinical characteristics and sample size used in the experiments. Therefore, large samples for molecular genetic studies are mandatory to detect these polymorphisms. Accordingly, several of today's findings have to be regarded as preliminary. The understanding of ADHD's neurobiology may be advanced by new technologies, such as single nucleotide polymorphism (SNP)-based genome scans performed with gene chips comprising 10 000–1 000 000 SNPs, as well as using more sophisticated animal model designs.

Assessments of comorbid psychiatric disorders in family studies further support a polygenic theory for ADHD. Results of analyses from independent samples of children with ADHD suggest that (1) ADHD and major depression share common familial vulnerabilities (e.g., children with ADHD often have mothers with depression) and (2) ADHD children with comorbid conduct disorders and bipolar disorders might be a distinct familial subtype of ADHD. Thus, dividing ADHD samples into those with conduct and bipolar disorders and those without might help define more familially homogeneous subgroups.

Treatment

Stimulant Medication

The most common treatment for ADHD is medication, with the vast majority of children receiving some form of stimulant medication (e.g., methylphenidate or Ritalin). Stimulant medication is effective in 70–80% of children receiving medication, and safety

and efficacy have been well documented. Of those children who do not respond to one stimulant medication, approximately 20–25% respond to an alternative stimulant medication. Recent trends in stimulant medication treatments include long-acting formulations (to reduce stigma and need to administer medication in the school setting) and patches (to reduce the need for oral administration and decrease the dosing schedule).

The frontosubcortical systems pathways associated with ADHD are rich in catecholamines, which are involved in the mechanism of action of stimulant medications used to treat this disorder. Stimulants, such as methylphenidate, seem to reduce core ADHD symptoms (i.e., inattention, hyperactivity, impulsivity) by inhibiting the dopamine transporter and blocking dopamine and norepinephrine reuptake into the presynaptic neuron, thereby increasing the release of these monoamines into the extraneuronal space. Changes in dopaminergic and noradrenergic function seem to be necessary for the clinical efficacy of pharmacologic treatments of ADHD. A plausible model for the effects of medications in ADHD suggests that, through dopaminergic and/or noradrenergic pathways, these agents increase the inhibitory influences of frontal cortical activity on subcortical structures. In contrast, effects on serotonin metabolism seem to be marginally related to the clinical efficacy of ADHD treatments, and serotonergic drugs have few effects in mitigating core symptoms of the disorder.

Side effects Most side effects of the stimulant medications are minor and are usually related to the dosage of the medication being taken. Higher doses produce more side effects. The most common side effects are decreased appetite, insomnia, increased anxiety, and/or irritability. Some children report mild stomach aches or headaches. A few children develop tics during treatment. Recent studies show consistent use of stimulant medication is associated with mild growth suppression (particularly height). Rare but potentially severe adverse effects, including sudden cardiac death and cancer, following long-term treatment have been reported; however, these effects have not been adequately demonstrated to be of significant concern at this time.

Nonstimulant Medication

In some individuals it may be necessary to try nonstimulant medication treatments. Reasons include (1) stimulants do not relieve symptoms, (2) stimulants cause intolerable side effects, (3) comorbid medical problems contraindicate treatment with stimulants, or (4) comorbid psychiatric conditions suggest treatment with nonstimulants.

A common nonstimulant alternative is atomoxetine (Strattera), which is a selective norepinephrine reuptake inhibitor. Its mechanism of action in ADHD seems to be the highly specific presynaptic inhibition of norepinephrines. Atomoxetine does not appear to affect the dopamine systems as directly as do the stimulants. Given its pharmacokinetic half-life of 5 h, it is generally dosed twice a day. Atomoxetine may also reduce tics and may be effective in children with ADHD who have comorbid anxiety. The Food and Drug Administration has issued warnings regarding rare side effects of hepatotoxicity and suicidal ideation.

Antihypertensives (e.g., Tenex), designed to treat high blood pressure, can also help control aggressive and impulsive behaviors in some people and are a second line of treatment for ADHD. Alternatively, antidepressants (e.g., Wellbutrin) may be needed if psychostimulants do not improve symptoms.

Nonmedication Interventions

Stimulant medications are particularly effective in remediating ADHD symptomatology. However, medications do not cure ADHD, nor do they increase knowledge or improve academic skills. The medications help the child to use those skills he or she already possesses, and allow the child sufficient self-regulation to learn new skills. Moreover, there is not compelling evidence to suggest that the stimulants improve the rather guarded long-term prognosis of the disorder. Thus, nonmedication treatments are also warranted in the treatment of ADHD. The most common and successful treatments for ADHD include behavior therapy (reinforcement and environmental controls), parent training, and social skills training. Due to the large evidence base consisting of rigorous experimental investigations, behavioral parent training and behavioral classroom interventions have been designated 'empirically supported' psychosocial treatments for ADHD.

Treatment sequencing Nonmedication interventions should be started early in a child's life and, ideally, before treatment with stimulant medication. Parent training should be considered a first-line treatment for ADHD, particularly with younger children, as there is evidence that such intervention may prevent later school problems, in addition to improving overall family functioning and parental ability to manage behavioral problems. At school entry, behavioral classroom interventions should be implemented as soon as possible. If treatment with stimulant medication is initiated, it should be administered in conjunction with these effective nonmedication treatments for ADHD. A large multisite randomized control trial showed that maintenance of gains was better in

children who received combination behavioral and medication treatment compared to those receiving medication only. Furthermore, combined treatments are effective in targeting other dysfunctions associated with ADHD, including oppositional behavior and anxiety.

Behavioral parent training A large evidence base exists for the use of parent behavioral interventions to reduce ADHD symptoms, to improve parenting skills and parent sense of competence, and to diminish family distress. Behavioral parent training interventions are based on a foundation of social learning principles that teach the child socially acceptable behavior. Parents and primary caregivers are trained in contingency management strategies, emphasizing behavior modification, cues, and consequences, reward systems, and discipline. Parents learn how to identify and manipulate the antecedents and consequences of a child's behavior (functional behavioral analysis), target and monitor problematic behaviors, reward positive and prosocial behavior through praise, positive attention, and tangible rewards, and decrease unwanted behavior through planned ignoring, time out, response cost, and other nonphysical discipline techniques. These approaches focus on reducing any positive reinforcement (e.g., parental attention) provided to the child for engaging in disruptive/defiant behavior, while simultaneously increasing the reinforcement parents provide for appropriate and compliant behavior. Parents are trained to provide punishment contingent on the display of disruptive or unacceptable behavior in a predictable, contingent, and immediate (to the behavior which precedes it) manner. Courses typically include training in specific evidence-based techniques for giving commands, reinforcing adaptive and positive social behaviors of the child while ignoring minor inappropriate behaviors to reduce or eliminate them, training in techniques for establishing and enforcing rules and establishing time-out procedures, training on initiating a point system with reward and response cost, and training in how to enforce contingencies across settings, problem-solving techniques, and strategies for maintenance and relapse prevention.

Behavioral classroom interventions As with parent training, behavioral classroom interventions generally involve regular consultation with the child's teacher regarding the use of behavior modification strategies. Consultation usually begins with psychoeducation about ADHD and identification of specific target behaviors, based upon a functional assessment of behavior (i.e., examination of antecedents, behaviors, and consequences). Teachers are then instructed regarding the use of specific behavioral techniques, including praise, planned ignoring, effective commands, and time out, as well as the daily report card (DRC) and/or more extensive individualized or classroom-wide contingency management programs. The consultant and teacher collaboratively develop specific behavioral interventions. Research suggests that teachers widely implement behavioral classroom interventions that target ADHD symptoms and associated functional difficulties, such as complying with classroom rules, engaging in appropriate interactions with classmates, displaying disruptive behavior, and complying with teacher commands. Furthermore, direct contingency management strategies employed in the classroom setting have been shown to be more effective than is traditional outpatient treatment for ADHD-related behaviors. A meta-analysis found that behavioral classroom interventions showed a very large effect size (ES = 1.44) on measures of treatment outcome, with a larger effect on child behavior than on academic or clinic performance.

Social skills training Peer rejection of children with ADHD is quite common and may lead to serious long-term consequences. Further, children with ADHD who overcome their social problems have better outcomes in the long term than do those children who continue to experience problems with peers. Thus, investigators have focused on designing psychosocial interventions that specifically target peer relationships. Social skills training interventions include instruction in social skills, social problem solving, and behavioral competencies. In addition, the interventions attempt to enhance social competence by encouraging close friendships, and decreasing undesirable and antisocial behaviors. Interestingly, although children with ADHD are less socially effective than are their peers, they typically do not perceive themselves as such. Thus, the primary goal of social skills training is to promote prosocial behaviors that include cooperation, communication, participation, and validation. Social skills training represents the most common approach to treating social problems in children, with groups typically being conducted at a clinic, a summer treatment program, or in school-based settings, and often including parent and teacher participation. The literature shows that combining social skills training with other interventions (e.g., behavior management, parent training) is more effective than is providing it as a stand-alone treatment.

Alternative therapies The following promising interventions must be tested in controlled, randomized trials in order to become established as empirically supported or evidence-based treatments.

• *Academic interventions*. Academic interventions for ADHD focus primarily on manipulating antecedent conditions such as academic instruction or materials in order to improve both behavioral and academic outcomes. Task and instructional modifications, peer tutoring, computer-assisted instruction, and strategy training have all been shown to improve outcomes in ADHD in preliminary studies. A recent meta-analysis of school-based interventions for children with ADHD found that both behavior management and academic interventions had similar positive effects on ADHD-related behaviors, although the impact of academic interventions on academic performance was more difficult to discern due to the relatively few studies employing academic outcome measures. Additional research is warranted to document the efficacy of academic interventions.

• *Attention training* (*ATT*). ATT is an innovative intervention based on the concept that adaptive activities can be directed to enhance skills related to sustained and executive attention functions. ATT originated from the field of cognitive rehabilitation. ATT activities emphasize many different aspects of attention, including sustained, selective, divided, alternating, working memory, self-regulation, overall body awareness, and metacognitive strategies. Most ATT approaches are adaptive, continually increasing in difficulty to challenge a child and provide practice in the domain being trained. Improvements in executive function, untrained academic skills, sustained attention, cognitive function, inhibition, and intelligence have been reported in various ATT studies with children and adolescents. Furthermore, transfer of improvements (or generalization) to untrained areas in individuals with ADHD has also been reported. These studies provide support for the notion that ATT may positively impact the developing attention skills of children with ADHD, and such increases may generalize to ecologically valid assays of real-world effortful task performance and expression of ADHD symptoms. However, more controlled clinical trials are needed before ATT can be endorsed as an effective, reliable treatment.

• *Biofeedback/neurofeedback*. Among the alternative treatments suggested for ADHD are electroencephalographic (EEG) biofeedback (neurofeedback or neurotherapy) and metacognitive therapy. These approaches are a form of behavioral training aimed at developing skills for self-regulation of brain activity. EEG biofeedback attempts to treat ADHD by decreasing g their slow wave activity and/or increase their fast-wave EEG activity, often using behavioral principles such as operant condition (i.e., positive reinforcement). Essentially, electrodes are positioned on the patient's head; a computer detects EEG activity and provides auditory or visual feedback in the targeted frequency bands. When the person is producing the desired EEG pattern, the computer provides a positive response or reward. After 20–50 training sessions, the person is hypothesized to be able to produce the desired brain activity independently via his or her own awareness of physiological processes. Despite some promising results, treatment effects may be due to nonspecific or placebo effects. Despite anecdotal and preliminary evidence for these alternative treatments, reviews of the applicability of neurofeedback for ADHD generally conclude that more controlled clinical trials are needed before it can be endorsed as an effective, reliable treatment. Studies are necessary to address flaws such as small sample size, diagnostic status of participants, lack of control or placebo procedures, failure to use blinding techniques as to subject status, practice effects, and failure to use random assignment. Additionally, few studies have been published in journals that are subject to rigorous peer review. Additional research is required to document the efficacy of this approach, as well as the effective agents of change in ADHD. Thus a National Institutes of Health consensus report on ADHD treatment noted that the empirical evidence for biofeedback and neurofeedback treatments was uneven and recommended more controlled studies before this treatment could be endorsed.

See also: Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine; Attention: Models; Attentional Mechanisms in Ventral Pathway; Attentional Networks; Attentional Networks in the Parietal Cortex; Attentional Functions in Learning and Memory; Sensorimotor Integration: Attention and the Premotor Theory; Visual Attention.

Further Reading

Bush G, Valera EM, and Seidman LJ (2005) Functional neuroimaging of attention-deficit/hyperactivity disorder: A review and suggested future directions. *Biological Psychiatry* 57: 1273–1284.

Chronis AM, Jones HA, and Raggi VL (2006) Evidence based psychosocial treatments for children and adolescents with attention deficit/hyperactivity disorder. *Clinical Psychology Review* 26: 486–502.

Conner DF (2002) Preschool attention deficit hyperactivity disorder: A review of prevalence, diagnosis, neurobiology, and stimulant treatment. *Developmental and Behavioral Pediatrics* 23: S1–S9.

Himelstein J, Schulz KP, Newcorn JH, et al. (2000) The neurobiology of attention-deficit hyperactivity disorder. *Frontiers in Bioscience* 5: 461–478.

Hinshaw SP (2007) Moderators and mediators of treatment outcome for youth with ADHD: Understanding for whom and how interventions work. *Ambulatory Pediatrics* 7: 91–100.

Khan SA and Faraone SV (2006) The genetics of ADHD: A literature review of 2005. *Current Psychiatry Reports* 8: 393–397.

Krain AL and Castellanos FX (2006) Brain development and ADHD. *Clinical Psychology Review* 26: 433–444.

Spencer TJ, Biederman J, and Mick E (2007) Attention-deficit/hyperactivity disorder: Diagnosis, lifespan, comorbidities, and neurobiology. *Ambulatory Pediatrics* 7: 73–81.

Waldman ID and Gizer IR (2006) The genetics of attention deficit hyperactivity disorder. *Clinical Psychology Review* 26: 396–432.

Relevant Websites

http://www.chadd.org – Children and Adults with Attention Deficit/Hyperactivity Disorder.

http://www.nimh.nih.gov – National Institute of Mental Health, National Institutes of Health, US Department of Health and Human Services: Attention Deficit/Hyperactivity Disorder booklet.

Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine

C J Vaidya, Georgetown University and Children's National Medical Center, Washington, DC, USA
P S Lee, Georgetown University, Washington, DC, USA

© 2009 Elsevier Ltd. All rights reserved.

Attention-deficit hyperactivity disorder (ADHD) is observed in 3–9% of school-aged children and 4% of adults worldwide at higher rates in males than females (2.5:1) and in children older than 9 years relative to younger children. Although family, twin, and adoption studies indicate high heritability (0.76), the mode of transmission is unknown but suspected to be polygenic. Molecular genetic studies suggest that susceptibility to ADHD involves multiple small-effect genes coding for proteins involved in catecholaminergic transmission. Catecholaminergic etiology is consistent with the treatment of choice for ADHD – psychostimulants such as amphetamines and methylphenidate hydrochloride (MPH) that increase synaptic levels of dopamine and norepinephrine by somewhat different mechanisms. Psychostimulants are highly effective for temporary alleviation of symptoms, starting at 30 min and peaking 60–90 min following oral administration of immediate-release formulations. Attempts to elucidate the therapeutic efficacy of MPH for ADHD have shaped current hypotheses about the central role of catecholamines, particularly dopamine (DA), in the neuropathophysiology of ADHD.

Behavioral Profile of ADHD

Clinical Characteristics

Based on the *Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition* (*DSM-IV*), diagnostic criteria for ADHD include parent and teacher reports of symptoms of inattention (e.g., making careless mistakes and difficulty concentrating), hyperactivity (e.g., fidgeting and excessive running or climbing), and/or impulsivity (e.g., difficulty waiting one's turn and talking excessively). These symptoms must appear prior to age 7 years, persist for at least 6 months in two settings (e.g., school and home), and cannot be explained by other psychiatric or neurological conditions. Symptom presentation varies widely among children and although the definition of subtypes in *DSM-IV* (e.g., predominantly inattentive, predominantly hyperactive/impulsive, and combined type with both inattention and hyperactivity/impulsivity) recognizes the heterogeneity, it does not fully account for it.

Phenotypic heterogeneity in ADHD stems from a variety of sources. First, assessment of symptoms varies across clinical settings such that primary care physicians rely on parent/teacher reports and direct observation of behaviors but specialty care settings (e.g., neuropsychology services) supplement this information with performance-based measures of cognitive functioning, academic achievement, and symptoms (e.g., Test of Variables of Attention (TOVA)) (Table 1). Performance-based measures alone, however, have limited discriminant validity for diagnostic use. Second, the severity or nature of symptom presentation may depend on gender (e.g., greater motor restlessness in boys but excessive talking in girls), age (e.g., reduced motor restlessness in adolescence relative to childhood), and context (e.g., reduced symptoms in structured environments). Third, ADHD symptoms co-occur with other psychiatric conditions such as mood disorders (e.g., depression and anxiety), conduct disorder, oppositional defiant disorder, obsessive–compulsive disorder, Tourette's, developmental dyslexia, autism spectrum disorders, as well as nonpsychiatric conditions such as sleep disorders. This complex pattern of comorbidity is difficult to resolve because ADHD symptoms could be primary or secondary or define a distinct phenotype that includes the comorbid disorder.

Cognitive Characteristics

It is commonly believed that the primary cognitive domain impaired in ADHD is executive function, an umbrella term that subsumes multiple dissociable 'top-down' operations that control attention and actions in the service of goal-directed behavior. Operations that have been the focus of ADHD research include inhibitory control (e.g., suppressing prepotent or irrelevant responses), working memory (e.g., temporary maintenance and/or manipulation of information), sustained attention or vigilance, and switching of task set (e.g., adapting to changing task demands). Studies using neuropsychological tasks of executive function that involve multiple operations (e.g., Tower of Hanoi, a problem-solving task requiring set maintenance and switching, working memory, and inhibition of inappropriate responses) indicate reduced performance in ADHD relative to control subjects. A meta-analysis of 83 studies involving 6700 subjects reported reductions in all executive operations sampled with effect sizes ranging from 0.4 to 0.7. Furthermore, effect sizes were larger for studies of spatial working memory

Table 1 Neuropsychological and laboratory measures of clinical symptoms exhibited by children with ADHD

Symptom	*Measure*	
	Neuropsychological	*Laboratory*
Inattention	Test of Variables of Attention (TOVA)	N-back working memory
	Continuous Performance Test (CPT)	Sternberg memory search
	Test of Everyday Attention for Children (TEA-Ch)–Score!	
	TEA-Ch–Sky Search/Map Mission	
	Finger-Windows test–spatial span	
	Letter/number cancellation	
	Symbol Search subtest of WISC-IV	
Impulsivity	TEA-Ch–Walk, Don't Walk	Stop Signal task
	TEA-Ch–Opposite Worlds	Go/No-go task
	Stroop task	
	Eriksen flanker task	
	Antisaccade task	
Hyperactivity	None	None

WISC-IV, Wechsler Intelligence Scale for Children–Fourth Edition.

($d = 0.85-1.14$). Drawing conclusions from these studies about which executive operations may be core deficits in ADHD is challenging because most neuropsychological tasks do not measure executive operations in a process-pure manner.

An influential theory of ADHD has proposed that a core deficit of response inhibition underlies executive dysfunction in ADHD. Response inhibition can be measured in a relatively process-pure manner using tasks that require subjects to withhold a prepotent response to a cue that occurs unpredictably following initiation of a response on the Stop Signal task or infrequently on the Go/No-go task. A meta-analysis of 17 studies involving 1200 children reported longer Stop Signal reaction times, an index of poor response inhibition, in ADHD relative to control children (effect size, $d = 0.58$). Use of response inhibition tasks in frontal lobe-damaged patients and in functional neuroimaging studies of healthy adults and children indicates involvement of dopamine-rich frontal–striatal circuitry. In ADHD children, those frontal–striatal regions are recruited to a lesser extent relative to controls. Thus, converging behavioral and neuroanatomical findings provide evidence for involvement of dopamine-rich regions underlying response inhibition deficits in ADHD.

Three lines of evidence have called into question the view that executive dysfunction is the cardinal cognitive deficit in ADHD. First, effect sizes of group differences in executive task performance are modest ($d = 0.50$–1.0) and suggest considerable overlap between score distributions of ADHD and control groups. In a study of five executive function measures, only 10% of ADHD children were impaired on all operations, whereas 21% of ADHD and 53% of control children were unimpaired on all operations. Such results have led researchers to conclude that executive dysfunction is neither necessary nor sufficient to account for all cases of ADHD.

Second, a variety of nonexecutive processes show impairments in ADHD subjects. On Go/No-go tasks, responses on Go trials that do not require inhibitory control are slower in ADHD than control children ($d = 0.58$). Reaction time performance of ADHD children deteriorates over the course of the task, is often slow on initial trials suggesting underarousal, and is more variable from trial to trial. Atypical response characteristics may relate to problems with temporal processing because ADHD children perform poorly on tasks of time estimation, time duration, and motor timing.

Third, motivational deficits have gained attention in ADHD with the observation that ADHD subjects are delay averse – that is, they tend to prefer small immediate rewards than to wait for larger delayed rewards. Delay aversion is posited to reflect atypical functioning of dopaminergic reward systems with motor hyperactivity reflecting compensatory behavior in response to unavoidable delays. Indeed, results of functional imaging study of reward function in ADHD showed that relative to controls, adolescents with ADHD showed reduced ventral striatal activation during anticipation of monetary rewards; furthermore, it was negatively associated with severity of impulsive/hyperactive symptoms. Impaired executive function and delay aversion appear to relate independently to ADHD because when tested in the same group of subjects, the two deficits were uncorrelated but together accounted for 90% of the cases. Thus, executive and motivational dysfunction may represent distinct phenotypes of ADHD.

Summary

The behavioral profile of ADHD suggests considerable phenotypic heterogeneity that is not fully captured by current diagnostic taxonomy and cognitive models of ADHD. At the clinical end, reliance on subjective reports of the child's behavior is unlikely to provide fine-grained symptom profiles unless supplemented by performance-based measures derived from cognitive models of ADHD. At the cognitive end, executive and motivational function appears to be independently affected in ADHD. A unifying factor, however, is that dopaminergic pathways figure centrally in the functional anatomy of both executive and motivational functions.

Evidence for Dopaminergic Dysfunction in ADHD

Whereas convergent lines of evidence suggest abnormal dopaminergic transmission in ADHD, its specific nature is debated. Furthermore, it is believed that the pathophysiology of ADHD is complex, with some interaction between dopaminergic and noradrenergic systems. This view comes from findings showing that pure DA agonists such as levadopa do not attenuate ADHD symptoms. Although the nature of its interaction with the noradrenergic system remains to be specified, strides made in understanding dopaminergic dysfunction in ADHD have been useful for elucidating the neurobiology of ADHD. Sources of evidence suggesting dopaminergic dysfunction in ADHD are discussed next.

Therapeutic Action of MPH

The primary source of evidence for hypothesizing dopaminergic dysfunction in ADHD comes from the mechanism of action of MPH. MPH blocks the dopamine transporter (DAT) that reuptakes DA following release, thereby increasing synaptic levels of DA. Indeed, ligand-based imaging studies in primates and humans showed that MPH bound to DAT in concentrations that were highest in the striatum and lower in thalamus, cortex, and cerebellum, leading to increases in magnitude of extracellular DA. At clinically relevant doses of 0.25–0.5 $mg\,kg^{-1}$, MPH blocked 50–60% of DAT in the striatum. MPH-induced (dose, 0.3 $mg\,kg^{-1}$) increases in extracellular DA in the striatum were positively correlated with severity of impulsivity and inattention on a performance-based measure (e.g., TOVA) in ADHD adolescents. Thus, *in vivo* imaging in humans provides evidence for increased synaptic DA following inhibition of DAT by MPH and its association with ADHD symptom expression.

Hyper- and hypodopaminergic status in ADHD In light of the pharmacodynamics of MPH, two lines of evidence have been important in shaping current views about dopaminergic dysfunction in ADHD. First, measurements of homovanillic acid (HVA), a metabolite of DA in cerebrospinal fluid (CSF), correlated positively with hyperactivity ratings from parents and teachers, and CSF levels of HVA decreased following MPH (and other stimulants) treatment. This finding suggests that ADHD reflects increased dopaminergic activity. Second, ligand-based imaging studies of DAT, the primary target of MPH, showed increased binding in the striatum in ADHD relative to control subjects in several studies. This finding suggests that ADHD reflects reduced dopaminergic activity due to increased DA reuptake resulting from higher DAT availability.

Both trait- and state-related factors may underlie increased DAT availability in ADHD. Trait-related factors resulting in higher DAT availability include 'hypertrophy' of dopaminergic neurons due to insufficient pruning during development and/or genetic differences. Alternately, increased DAT availability in ADHD may be state related, resulting from adaptive processes compensating for greater DA release, reduced vesicular storage, or other differences in DA transmission. In some studies, DAT availability in the striatum did not differ between ADHD and control subjects but, rather, was reduced in the midbrain of ADHD subjects. Differences in sample characteristics (e.g., nicotine use) and properties of ligands (e.g., detection of internalized or external DAT) may have contributed to the inconsistent findings across studies. It is also likely that heterogeneity among individuals may emerge from some combination of trait- and state-related factors.

Working hypothesis of dopaminergic dysfunction in ADHD Increased DA release and increased DAT availability in ADHD have been reconciled in a model by Seeman and Madras that considers both tonic and phasic dopaminergic activity. Phasic activity reflects acute DA release stimulated by a nerve impulse, whereas tonic activity reflects chronic DA accumulation in synaptic space. The two types of DA activity are reciprocal such that tonic DA activity stimulates autoreceptors on the presynaptic neuron that serve to attenuate phasic DA release. Seeman and Madras posit that tonic DA activity is reduced in ADHD. Findings of elevated DAT in ADHD provide some support for this view because greater reuptake of synaptic DA is likely to result in reduced tonic DA activity. Furthermore, it is posited that low tonic DA activity is unlikely to stimulate presynaptic

autoreceptors, resulting in higher phasic release of DA. Greater phasic DA release is consistent with findings of greater HVA levels in CSF of hyperactive children. The therapeutic action of MPH in Seeman and Madras's model is mediated through increased tonic DA and reduced phasic DA. Specifically, they posit that the blockade of DAT reuptake by MPH increases tonic DA activity that leads to increased autoreceptor stimulation, which in turn attenuates phasic DA release. Seeman and Madras's model, therefore, posits that ADHD is characterized by both hypo- and hyperdopaminergic status that is normalized by MPH.

Variability in MPH response Two properties of MPH effects, context dependency and individual variability, should be considered by any model of ADHD pathophysiology. First, anecdotal and clinical reports have emphasized that efficacy of MPH varies across situations. Indeed, in a placebo-controlled study, symptom improvements in ADHD children were greater in a classroom than playground setting. Ligand-based imaging of D2 receptors in healthy adults suggests that context dependency in MPH response is a property of dopaminergic function. Specifically, oral administration of MPH at clinically relevant doses (20 mg) increased extracellular DA in the striatum selectively for a mathematical task with monetary incentives. DA levels were associated with subjects' reports of greater interest and engagement in the mathematical task following MPH administration. These effects were not observed following MPH and placebo administration for a neutral task (viewing nature scenes without monetary incentives) and were not observed on placebo for the mathematical task. Thus, these results suggest stimulus selectivity in MPH-induced enhancement of attentional performance and interest and associated DA modulation.

Second, a majority (60–70%) but not all ADHD individuals respond favorably to MPH. Whereas responders and nonresponders did not differ in symptom characteristics, nonresponders had lower than normal DAT availability. Furthermore, at similar levels of DAT blockade, the magnitude of extracellular DA increase induced by MPH varied across subjects due to differences in rates of DA cell firing. Thus, it is important to consider etiological factors underlying individual differences in DAT availability and DA cell firing.

Candidate Dopaminergic Genes

In light of the molecular targets of stimulant medications and results of transporter imaging studies in ADHD, molecular genetic studies have searched for susceptibility genes relating to catecholamine function. Studies of the dopaminergic system have focused on polymorphisms of D2, D5 receptors – DAT, dopamine beta-hydroxylase (DBH), tyrosine hydroxylase, catechol-O-methyltransferase (COMT), and monoamine oxidase A (MAO-A).

Although findings are not conclusive, the most consistent evidence for association with ADHD comes from polymorphisms of the dopamine transporter gene (DAT1) and D4 receptor gene (DRD4). DAT1, located on chromosome 5p15.3, contains a 40-base pair variable number of tandem repeats ranging from 3 to 13, with 9 and 10 being most common, in the 3′-untranslated region. Greater prevalence of the 10-repeat allele in subjects with ADHD has been reported by many studies but not by all studies. DRD4 contains a 48-base pair variable number of tandem repeats in exon 3, and greater prevalence of the 7-repeat allele in ADHD subjects has been reported by some studies but not by all. Furthermore, some studies have found association of ADHD with other DRD4 alleles (e.g., 2–5 repeat). Mixed findings across studies may result from differences in allelic variability associated with ethnicity of samples, study designs for ascertaining linkage, methods of ADHD diagnosis, and from limited statistical power due to small sample sizes. Examination of odds ratios (a statistic assessing magnitude of association) pooled across studies, a method that circumvents statistical limitations of small samples, showed that polymorphisms of several dopaminergic genes (DRD4, DRD5, DAT, and DBH) confer a statistically significant, albeit small, risk for ADHD.

Genetic polymorphisms have the potential to elucidate dopaminergic pathophysiology of ADHD because allelic variation in DAT1 and DRD4 influences DA transmission. DAT expression was greater for the 10-repeat allele in *in vitro* studies but findings of *in vivo* studies are mixed, reporting higher, lower, and similar striatal DAT expression in homozygous carriers of the 10-repeat allele relative to carriers of the 9-repeat allele. *In vitro* studies of DRD4 alleles indicate that response to DA was reduced for the 7-repeat allele. Selective ligands for D4 receptors are currently unavailable and, therefore, it is unknown whether D4 receptor density differs between carriers of the 7-repeat and other DRD4 alleles.

Variability in ADHD symptoms and MPH response Allelic variation in DAT1 and DRD4 relates to ADHD symptoms selectively. Homozygosity of the 10-repeat DAT1 allele was associated with higher hyperactivity symptoms, whereas that of the 7-repeat DRD4 allele was associated with higher inattention symptoms, in ADHD children and their parents. Furthermore, children homozygous for the 10-repeat

DAT1 allele performed worse on performance measures of impulsivity, the Test of Everyday Attention for Children (TEA-Ch) Opposite Worlds task and errors of commission on the Continuous Performance Task, relative to heterozygous ADHD children. Also, homozygous 10-repeat DAT1 children exhibited greater intraindividual response time variability than heterozygous children. It is noteworthy that DAT is primarily expressed in the striatum, whereas D4 is abundant primarily in prefrontal cortex. Thus, selective effects of DAT1 and DRD4 polymorphisms on inattention and impulsivity in ADHD may reflect functional characteristics of the striatum and prefrontal cortex, respectively.

Pharmacogenetic studies suggest that the efficacy of MPH depends on DAT1 and DRD4. Based on *in vitro* studies indicating that DAT expression was greater for the 10-repeat allele, carriers of that allele should show a better response to MPH. Indeed, homozygosity of the 10-repeat DAT1 allele was associated with a better response to MPH in some studies. Other studies, however, did not replicate this finding. Discrepant findings among studies probably relate to differences in sample selection and characteristics of treatment regimens, such as dosage, duration, and assessment of MPH response. Based on the *in vitro* observation that DA response is blunted for the 7-repeat allele, carriers of that allele should show a reduced MPH response. Indeed, a study examining variability in dose–response to MPH found that carriers of the 7-repeat allele required higher doses to attain normalization of symptoms relative to noncarriers. Thus, examination of allelic variability in MPH effects is a worthwhile approach to elucidate etiological heterogeneity in DA function in ADHD, but larger well-controlled studies are needed to draw definitive conclusions.

Functional Brain Imaging of MPH Effects

Functional magnetic resonance imaging (fMRI) performed with methylphenidate challenge is a noninvasive way to visualize dopaminergic function *in vivo*. fMRI capitalizes on changes in blood oxygenation secondary to neural activity evoked by a cognitive challenge. Results of several studies indicate that frontal and striatal activation during performance of attentional and inhibitory tasks differs between ADHD and control subjects. Comparison of fMRI measures in the same subjects during performance of a cognitive challenge, once with and once without administration of MPH, provides an index of changes in neural activity induced by alterations in catecholamine levels (generally referred to as pharmacological fMRI). Evidence from animal studies has validated the ability of fMRI to index metabolic changes induced by dopaminergic stimulation. Specifically, fMRI signal in response to amphetamine or its analog, CFT (2B-carbomethoxy-3B-[4-flurophenyl] tropane, WIN 36 528), was correlated with changes in extracellular DA concentrations in the striatum (confirmed by microdialysis) and was lost following denervation of the striatum by 6-OHDA lesioning. Furthermore, the fMRI signal changes were restricted to regions with high expression of DA receptors, such as striatum and cingulate and frontal cortex (confirmed by receptor-labeled positron emission tomography). Thus, fMRI with MPH challenge provides a noninvasive assay for dopaminergic function relevant to cognitive deficits observed in ADHD.

Findings of pharmacological fMRI studies indicate that MPH has selective effects on brain regions involved in executive function in subjects with ADHD. Results of these studies suggest that MPH normalizes activation in striatal structures but not in other brain regions. Striatal activation was reduced relative to control children and increased following administration of MPH during performance of response inhibition and divided attention tasks. However, other regions with reduced activation in ADHD subjects, such as middle temporal and anterior cingulate gyri, did not change following administration of MPH. Furthermore, effects of MPH on prefrontal activation in ADHD subjects differed across studies such that activation was increased during a difficult response inhibition condition, did not change during an easy response inhibition condition, and was reduced during working memory performance. Thus, effects of MPH on metabolism in circuitry involved in executive function are not uniform across regions.

The influence of MPH on brain metabolism appears to vary by symptom expression. In contrast to increased striatal activation following MPH administration, fMRI imaging during response inhibition in control children showed that MPH reduced striatal activation. These opposite effects of MPH in ADHD and control children were not observed elsewhere in the brain. Furthermore, these opposite effects did not relate to performance differences because MPH improved response inhibition to the same extent in both ADHD and control groups. Results of other studies indicate that MPH increased blood volume in children who were more hyperactive, whereas it decreased blood volume in those who were less hyperactive, in the cerebellar vermis and putamen. In light of evidence relating symptom expression to functional properties of DA transmission, MPH effects on brain metabolism probably reflect differences in baseline dopaminergic function.

Summary

Convergent evidence from studies of MPH effects and *in vivo* imaging suggests that DA transmission is dysfunctional in ADHD and may relate to polymorphisms of dopaminergic genes. The most direct evidence for presynaptic DA dysfunction in ADHD comes from ligand-based imaging studies showing elevated DAT density in subjects with ADHD. Whether this is a trait related to abnormal neuronal pruning or a state reflecting adaptive response to abnormalities in other aspects of DA transmission remains to be resolved. Resolution may come from molecular genetic studies because inheritance of the 10-repeat allele of the DAT gene has been linked to individual differences in DAT expression and susceptibility to ADHD. Polymorphism of D4 receptor genotype also confers some risk for ADHD, suggesting that postsynaptic DA function is also part of the ADHD pathophysiology. The leading model of DA dysfunction in ADHD posits that the disorder involves both lower and higher dopaminergic activity. Specifically, Seeman and Madras hypothesize that ADHD reflects reduced tonic but increased phasic DA activity. Efficacy of MPH in this model is mediated by reduced phasic DA activity via autoreceptor stimulation due to increased tonic DA activity that results from increased synaptic DA following DAT blockade by MPH. However, MPH response varies depending on contextual and subject-related factors, a finding that is not accounted for by current models of ADHD. Use of pharmacological fMRI provides a noninvasive assay of DA function in ADHD. Its use has revealed that changes in neural activity induced by MPH during executive functions serve to 'normalize' striatal function in ADHD subjects but depend on symptom severity.

Future Direction: Endophenotypes for ADHD

Researchers agree that phenotypic heterogeneity of ADHD is the primary limiting factor in elucidating the pathophysiology of ADHD. Currently, most researchers endorse an approach that characterizes individual differences in terms of endophenotypes rather than symptom profiles. Endophenotypes refer to cognitive functions that can be specified on performance-based measures that are heritable and draw upon neurophysiology relevant to the disorder. Current research has identified response inhibition measured by Stop Signal or Go/No-go tasks as a possible endophenotype for ADHD because functional variability relates to dopaminergic modulation in prefrontal–striatal regions and to heritability via dopaminergic genes. Other candidate endophenotypes for ADHD are intraindividual response variability on reaction time tasks and a shortened delay gradient on tasks involving rewards. Pharmacological fMRI is a promising tool in the search for endophenotypes for ADHD because it can characterize individual variation in the functional anatomy of a cognitive function with respect to MPH response. Furthermore, candidate genes can be incorporated into this approach by examining whether MPH effects on prefrontal–striatal function vary for polymorphisms of dopaminergic genes. A characterization of ADHD phenotypes with an approach that unifies genetic, neurophysiological, and cognitive factors is also likely to be promising for the development of objective diagnostic criteria and for treatment planning.

See also: Amphetamines; Attention Deficit Hyperactivity Disorder; Cognition: An Overview of Neuroimaging Techniques; Dopamine – CNS Pathways and Neurophysiology; Dopamine in Perspective; Neuroimaging.

Further Reading

Castellanos FX, Sonuga-Barke EJ, Milham MP, and Tannock R (2006) Characterizing cognition in ADHD: Beyond executive dysfunction. *Trends in Cognitive Science* 10(3): 117–123.

Faraone SV, Perlis RH, Doyle AE, et al. (2005) Molecular genetics of attention-deficit/hyperactivity disorder. *Biological Psychiatry* 57(11): 1313–1323.

Faraone SV, Sergeant J, Gillberg C, and Biederman J (2003) The worldwide prevalence of ADHD: Is it an American condition? *World Psychiatry* 2(2): 104–113.

Grace AA (2001) Psychostimulant actions on dopamine and limbic system function: Relevance-related behavior and impulsivity. In: Solanto MV, Arnsten AFT, and Castellanos FX (eds.) *Stimulant Drugs and ADHD: Basic and Clinical Neuroscience*, pp. 134–157. New York: Oxford University Press.

Krause KH, Dresel SH, Krause J, la Fougere C, and Ackenheil M (2003) The dopamine transporter and neuroimaging in attention deficit hyperactivity disorder. *Neuroscience & Biobehavioral Reviews* 27(7): 605–613.

Madras BK, Miller GM, and Fischman AJ (2005) The dopamine transporter and attention-deficit/hyperactivity disorder. *Biological Psychiatry* 57(11): 1397–1409.

Nigg JT, Willcutt EG, Doyle AE, and Sonuga-Barke EJ (2005) Causal heterogeneity in attention-deficit/hyperactivity disorder: Do we need neuropsychologically impaired subtypes? *Biological Psychiatry* 57(11): 1224–1230.

Sonuga-Barke EJ (2002) Psychological heterogeneity in AD/HD—A dual pathway model of behaviour and cognition. *Behavioural Brain Research* 130(1–2): 29–36.

Spencer TJ, Biederman J, Madras BK, et al. (2005) *In vivo* neuroreceptor imaging in attention-deficit/hyperactivity disorder: A focus on the dopamine transporter. *Biological Psychiatry* 57(11): 1293–1300.

Swanson JM and Volkow ND (2001) Pharmacokinetic and pharmacodynamic properties of methylphenidate in humans. In: Solanto MV, Arnsten AFT, and Castellanos FX (eds.) *Stimulant Drugs and ADHD*. New York: Oxford University Press.

Vaidya CJ (2002) Application of pharamacological fMRI to developmental psychiatric disorders. *Developmental Science* 5(3): 310–317.

Vaidya CJ, Austin G, Kirkorian G, et al. (1998) Selective effects of methylphenidate in attention deficit hyperactivity disorder: A functional magnetic resonance study. *Proceedings of the National Academy of Sciences of the United States of America* 95(24): 14494–14499.

Volkow ND, Wang GJ, Fowler JS, et al. (2004) Evidence that methylphenidate enhances the saliency of a mathematical task by increasing dopamine in the human brain. *American Journal of Psychiatry* 161(7): 1173–1180.

Willcutt EG, Doyle AE, Nigg JT, Faraone SV, and Pennington BF (2005) Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry* 57(11): 1336–1346.

Attention: Models

P H E Tiesinga, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
T J Sejnowski, Salk Institute and University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Our sensory environment is represented in the cortex as the electrical activity of billions of neurons. The neural activity in the visual pathway reflects not only the current visual environment but also the goals and expectations of the organism as well. Suppose that one is at the airport picking up a friend who is wearing a red coat. In general, red objects will activate specific neurons in the visual cortex, but, because they are now relevant to the current goal, their responses will be enhanced and preferentially processed. Selective attention is a general strategy for selecting the currently relevant information out of the overwhelming amount of information that enters the brain via the senses. This article reviews models for selective attention in the visual system at the level of single neurons and the local circuit they are embedded in.

Parametric Models for Stimulus Response Properties of Early Visual Neurons

The response of a visual neuron depends both on the stimulus that is driving it (bottom-up) and attention (top-down). The purely sensory responses of cortical neurons were originally characterized in anesthetized cats and primates and with simple stimuli such as oriented bars and gratings. In the primary visual cortex (V1), many neurons respond best to a stimulus of a specific orientation placed at a specific location in the visual field. The receptive fields (RF) of middle temporal (MT), V2, and V4 cortical areas downstream to V1 are much larger than those of V1 neurons and involve more complex combinations of features.

When an oriented bar is placed at the center of the RF of a V1 simple cell, the mean firing rate varies approximately as a bell-shaped curve with the stimulus orientation (Figure 1(a)) that can be approximated mathematically as

$$R(\theta) = R_0 + A\,\exp\left(\frac{(\theta - \theta_{\text{pref}})^2}{2\sigma^2}\right) \qquad [1]$$

where R is the firing rate in response to a stimulus of orientation θ, θ_{pref} is the neuron's preferred orientation, σ is the tuning width (related to the half width at half height), R_0 is the neuron's firing rate in the absence of the stimulus, and A represents the strength of firing rate modulation. Studies using anesthetized cats and macaques as well as alert macaques to investigate the orientation selectivity of V1 neurons have revealed spontaneous activity R_0 ranging from zero to about 20 Hz and A varying from a few hertz to 100 Hz. The tuning width σ could be as low as 10° with a median value of 30° across all layers in V1 in alert macaque monkeys.

Luminance contrast is the absolute difference in luminance values across the stimulus expressed relative to the background. It is a measure of stimulus strength, and it takes values between zero and 100%. The firing rate increases monotonically with stimulus contrast (Figure 1(b)). Mathematically, this relationship is summarized in terms of the contrast response function (CRF), a commonly used expression for which is

$$R(c) = R_0 + R_m \frac{c^n}{c_0^n + c^n} \qquad [2]$$

where c is the contrast, R_m is the maximum firing rate, c_0 is the contrast for which the firing rate above baseline is half R_m, and n is the power of the nonlinearity. Typical values for cat and primate V1 are $c_0 \approx 10\%$, $n \approx 2$. The CRF has three regimes. For low contrast ($c \ll c_0$), the firing rate increases supralinearly as a power of the contrast c. The largest rate of change in firing rate is at medium contrast, around c_0, and the firing rate saturates at high contrasts ($c \gg c_0$).

Orientation tuning is approximately contrast invariant. For a range of contrast values, the orientation-tuning curve has approximately the same shape but a different maximum firing rate (the parameter A). Mathematically, this implies that the rate above baseline can be written as a product of the orientation-tuning curve and the contrast response curve (Figure 1(c)):

$$R(\theta, c) = R_0 + R_m \frac{c^n}{c_0^n + c^n} \exp\left(\frac{(\theta - \theta_{\text{pref}})^2}{2\sigma^2}\right) \qquad [3]$$

This functional form reflects a normalization model. A given stimulus activates not only the neuron that is recorded but also a pool of other neurons. The recorded neuron receives input proportional to the orientation-tuned output of a filter L and the

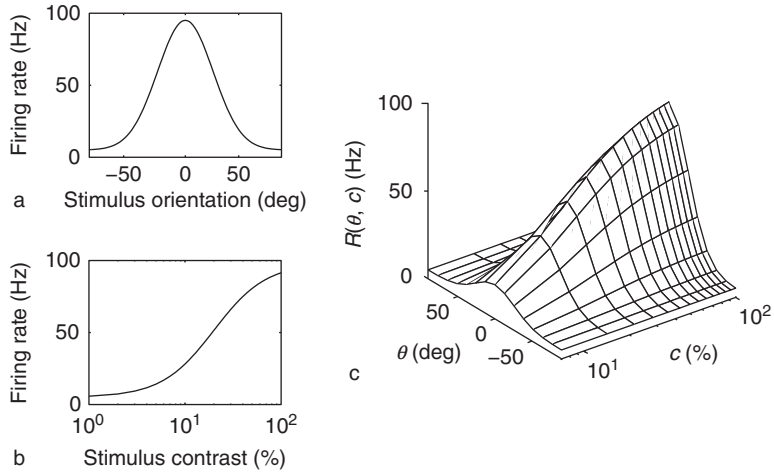

Figure 1 (a) Orientation-tuning curve. (b) Contrast response function. (c) The firing rate, according to eqn [3], as a function of orientation θ and contrast c.

contrast c. The combined activity of neurons in the pool is independent of stimulus orientation, but it increases with contrast as c^n. This activity divisively normalizes the response of the recorded neuron. The normalization model can account for a large number of experimentally measured responses. Model studies and *in vitro* experiments show that a number of biophysical mechanisms yield divisive normalization. An alternative means of achieving normalization involves the short-term plasticity of the synapses onto the neuron.

Parametric Models for Attentional Modulation of Early Visual Responses

The modulation by selective attention of neural responses has been studied in awake nonhuman primates. The strongest modulation of individual neurons was recorded in the MT, V4, and inferotemporal cortical areas, which all are downstream of V1. In one experimental paradigm, two equivalent arrays of stimuli are presented, typically on opposite sides of the vertical meridian. The monkey is rewarded for detecting a subtle change in the stimulus at a cued location. It is not rewarded for reporting a change in a stimulus at the noncued location. Since this is a difficult task, the monkey needs to pay attention to the cued location in order to observe and respond to the stimulus change. The neuron that is recorded has its receptive field at one of the two locations. Hence, the response of that neuron to the same stimulus with attention can be compared with when attention is directed at the location outside the neuron's RF.

Attention modulates the firing rate. For spatial attention, the firing rate typically increases when the neuron's receptive field overlaps with the area where attention is directed. The neuron receives feed-forward synaptic inputs and transforms them into a firing rate. Attention could act in several ways: it could make the feed-forward input stronger, or it could change the gain of the input-to-output transformation. The former is referred to as contrast gain and the latter as response gain. In Figure 2(a), illustrating contrast gain, the CRF with attention (left curve) is plotted together with the curve for attention directed away from the RF (right curve). Attention shifted the curve to the left, which implies that a neuron responds to a stimulus as if it had a contrast higher than its actual (veridical) contrast. Contrast gain has three distinguishing properties. First, with attention the neuron will respond to low-contrast values that it did not respond to without attention. Second, the largest change in firing rate will occur for moderate firing rates corresponding to contrasts around c_0. Third, the saturation firing rate remains the same, but it is reached for a lower value of the contrast than without attention.

Response gain is illustrated in Figure 2(b). The CRF with attention is obtained by multiplying the CRF without attention by a gain factor larger than unity. Response gain can be distinguished from contrast gain. First, if the neuron does not respond to a low-contrast stimulus without attention, neither will it do so with attention. Second, the largest change in firing rate will occur at the highest value of the contrast where the neuron fires at its highest rate. Third, the saturation rate will be higher for attention compared with no attention.

Experiments in a spatial attention paradigm in cortical areas MT and V4 of macaque show that attentional modulation of the CRF is best described as contrast gain. This matches the results of psychophysical experiments: Human observers report that the perceived contrast of a stimulus is higher if it is in the focus of attention. Other experiments on the

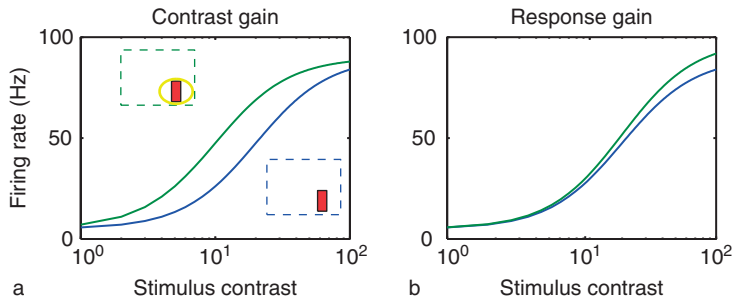

Figure 2 Attentional modulation of the contrast response function when attention is directed away from (blue) or directed toward (green) the receptive field of the recorded neurons. Illustration of (a) contrast gain and (b) response gain.

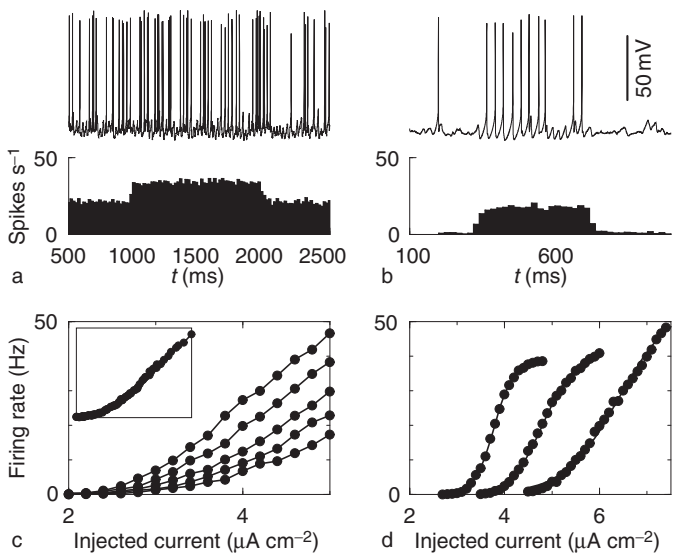

Figure 3 Response gain and contrast gain can be mediated by modulation of inhibitory synaptic inputs. (a, b): (top) Example voltage trace and (bottom) spike time histogram based on 500 trials. (a) The model neuron was driven by synchronous inhibitory and asynchronous excitatory inputs. During the interval between 1000 and 2000 ms, the jitter of the inhibitory volleys was decreased from 4 to 2 ms. During the period of increased inhibitory synchrony (low jitter), the firing rate was increased from 22 to 35 Hz. (b) The model neuron was driven by synchronous inhibition and a constant depolarizing current. During the interval between 300 and 700 ms, the jitter of the inhibitory volleys was decreased from 8 to 2 ms. The inhibitory inputs act as a gate. During periods of low inhibitory synchrony (high jitter), the neuron does not spike, whereas during periods of high inhibitory synchrony, it does spike. (c) and (d) The firing rate vs. current curves. Response gain: in (c) the synchronous inhibitory volleys consisted on average of ten presynaptic spikes, the jitter was, from top to bottom: 1, 2, 3, 4, and 5 ms. Inset: the five curves were made to overlap by scaling the firing rate and by small shifts in the current. Contrast gain: in (d) the synchronous inhibitory volleys consisted on average of 50 presynaptic spikes, and jitter was, from left to right, 1, 3, and 5 ms. Reproduced from Tiesinga PH, Fellous JM, Salinas E, Jose JV, and Sejnowski TJ (2004) Inhibitory synchrony as a mechanism for attentional gain modulation. *Journal of Physiology-Paris* 98: 296–314, with permission from Elsevier.

modulation of orientation tuning curves were best described as a response gain with attention. Overall, there is evidence in support of response gain, as well as contrast gain, depending on the specific details of the behavioral task by which attentional modulation was studied. In experiments designed to study feature-based attention in area MT, effects consistent with response gain were found. The neuron's response was multiplied by a gain factor that monotonically increases with the difference between the attended feature (direction, color, or shape) and the feature value preferred by the neuron. A similar effect of feature-based attention was obtained in fMRI experiments with human participants. The response of MT to a moving random dot pattern was increased when it was task-relevant. In macaque visual cortex, attention led in some, but not all, cases to an increase in baseline rate (R_0).

Thus, the response to a stimulus of contrast c and orientation can be described by eqn [3]. In this expression, attention 'a' enters via the baseline rate $R_0(a)$ and the contrast threshold $c_0(a)$. The effects of attention are consistent with either response gain or contrast gain depending on which stimulus parameter

is varied. This form makes a prediction that has recently been confirmed experimentally: the response to a stimulus of nonoptimal orientation should saturate at the same contrast as a stimulus with optimal orientation.

Biophysical Models for Attentional Response Modulation

A biophysical model must account for the inputs a neuron receives from three sources: feed-forward input that represents stimulus aspects (bottom-up), modulatory feedback input (top down), and recurrent connections within the same cortical area (but potentially different layers). The recurrent activity depends both on bottom-up and top-down inputs. Furthermore, the bottom-up input includes the influence of top-down projections on the earlier cortical area. Response mechanisms can be studied theoretically and *in vitro* at the single-neuron level under the assumption that the three sources of synaptic inputs can be split up in two sets of independent inputs, one of which depends only on the stimulus identity and contrast, the other only on the locus of attention. The rate of stimulus-related synaptic inputs depends, as a bell-shaped tuning curve, on the stimulus orientation, with excitatory inputs dominating inhibitory inputs. The attention-related inputs consist of approximately equally strong excitatory and inhibitory inputs. Thus the main contribution of attention consists of modulating fluctuating inputs. Two mechanisms have been proposed that are based on the idea that attention modulates the amount of fluctuation. First, in gain modulation by balanced synaptic inputs, attention causes a proportional change in rates of excitatory and inhibitory inputs such that the variance of the membrane potential changes but its mean remains the same. Second, in gain modulation by correlation, there is a change in correlation between inputs such as the synchrony of inhibitory inputs or the delay between excitatory and inhibitory inputs. Both of these proposed mechanisms are supported by *in vitro* experiments. The basic paradigm is as follows: in the experiments, feed-forward inputs are presented by a constant depolarizing input current I injected at the soma, whereas the modulatory inputs are injected at the soma using dynamic clamp. The firing rate versus current curve (f–I) is then constructed by calculating the firing rate for a set of different values of I in the presence of the modulatory inputs. Chance and co-workers have shown that an increase in the level of balanced synaptic inputs yields a decrease in the gain of the f–I curve. Tiesinga et al. have shown that an increase in inhibitory synchrony also leads to an increase in gain of the f–I or a shift in the f–I along the current ordinate (Figure 3).

The single-neuron studies show how the response could be modulated, but they do not provide a mechanism at the local circuit level by which the necessary changes in the modulatory inputs come about. Furthermore, in the model as well as *in vitro*, the inputs were somatic, but the modulatory and feed-forward inputs may impinge at different locations on the neuron's extensive dendritic tree, which could introduce more complex dynamical interactions.

Discussion

This article has focused on the effects of attention on the response of a single neuron to a simple stimulus. Current experimental investigations are extending these results in two ways. First, in higher cortical areas, the receptive fields are so large that they accommodate multiple stimuli. Attentional modulation is much stronger when there are multiple stimuli in a receptive field. Phenomenological models are presently being developed to quantitatively account for these data. Second, even when attention does not change the firing rate, it may change the correlation between neurons. Experiments in V4 show that attention increases the coherence in the γ-frequency range between neurons and the local field potential. This suggests an important role for the synchrony in networks mediating the effects of attention. Future theoretical research is directed at exploring the synchrony hypothesis at the level of the cortical circuit.

See also: Attention and Eye Movements; Attentional Networks; Attentional Functions in Learning and Memory; Neglect Syndrome and the Spatial Attention Network; Sensorimotor Integration: Attention and the Premotor Theory; Visual Attention; Visual Motion Models.

Further Reading

Albrecht DG, Geisler WS, Frazor RA, and Crane AM (2002) Visual cortex neurons of monkeys and cats: Temporal dynamics of the contrast response function. *Journal of Neurophysiology* 88: 888–913.

Bichot NP, Rossi AF, and Desimone R (2005) Parallel and serial neural mechanisms for visual search in macaque area V4. *Science* 308: 529–534.

Chance FS, Abbott LF, and Reyes AD (2002) Gain modulation from background synaptic input. *Neuron* 35: 773–782.

Ferster D and Miller KD (2000) Neural mechanisms of orientation selectivity in the visual cortex. *Annual Review of Neuroscience* 23: 441–471.

Fries P, Reynolds JH, Rorie AE, and Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291: 1560–1563.

Maunsell JH and Treue S (2006) Feature-based attention in visual cortex. *Trends in Neurosciences* 29: 317–322.

Reynolds JH and Chelazzi L (2004) Attentional modulation of visual processing. *Annual Review of Neuroscience* 27: 611–647.

Salinas E and Sejnowski TJ (2001) Correlated neuronal activity and the flow of neural information. *Nature Reviews Neuroscience* 2: 539–550.

Tiesinga PH, Fellous JM, Salinas E, Jose JV, and Sejnowski TJ (2004) Inhibitory synchrony as a mechanism for attentional gain modulation. *Journal of Physiology* 98: 296–314.

Williford T and Maunsell JH (2006) Effects of spatial attention on contrast response functions in macaque area V4. *Journal of Neurophysiology* 96: 40–54.

Attentional Functions in Learning and Memory

M Sarter and C Lustig, University of Michigan, Ann Arbor, MI, USA

© 2009 Elsevier Ltd. All rights reserved.

Attention and Memory: Conceptual Issues

Attention describes a set of cognitive processes which act to optimize stimulus detection, discrimination, and processing. Attention operates in part by top-down tuning of sensory systems to facilitate the detection of selected stimulus characteristics, such as location and modality, by switching the cortical processing from associational to input modes, and by allocating attentional resources to these operations. Different forms of attention have been categorized, such as sustained, selective, and divided attention. Sustained attention describes the subject's state of readiness to detect rarely and unpredictably occurring changes in the stimulus situation over extended periods of time. Selective attention brings targeted information into the focus of consciousness, while suppressing the detection and processing of nontarget signals. Divided attention emphasizes the allocation and the management of limited attentional resources in situations that require attention to multiple stimuli or tasks.

The assumption that attended stimuli are encoded more effectively into memory than less-attended ones is straightforward and supported by substantial evidence. However, theoretical frameworks that more fully describe the relationships between attention and learning and memory (L&M) have remained unexpectedly rare. Furthermore, empirical analyses of interactions between attention and L&M focused on selected aspects of attention and a rather small number of experimental paradigms. For example, increased demands on the division of attention impair stimulus encoding, but generally do not impede the retrieval of previously learned information. Conversely, high demands on memory operations impair attentional capacities, particularly the ability to filter nontarget stimuli from being processed. A second major paradigm used in research on the interactions between attention and L&M concerns the impaired learning of extradimensional shifts, as such shifts require the processing of a previously unattended stimulus dimension (e.g., shape, when one had been attending to color).

Attentional processes and capacities represent a cluster of variables contributing to the efficacy of L&M. At the extreme, it is difficult to envision meaningful acquisition of declarative information in the absence of attention. Furthermore, levels of attentional performance vary considerably, and this variation affects the rate of learning and thus the efficacy of memory. Therefore, it is not unexpected that brain mechanisms mediating attentional functions and capacities are also part of the neuronal circuitry mediating L&M.

Functional neuroimaging studies provide strong evidence for the interplay between attention, learning, and memory. Meta-analytic and experimental studies reveal large overlaps between prefrontal and parietal regions activated in attention and memory tasks that are not shared with other (e.g., language processing) domains. Dividing attention at learning leads to reduced activity in prefrontal brain regions associated with subsequent memory, and has long been known to reduce memory performance. Conversely, recent evidence suggests that higher levels of tonic or baseline activity in regions involved in stimulus perception (e.g., parahippocampal regions for perception of scenes) are associated with higher levels of attention to those stimuli, and with better subsequent memory than those preceded by lower tonic activity. Memory also serves to guide attention: memory-guided visual search activates similar brain regions as does cue-guided visual search, and may even provide greater performance benefits than a visual cue.

In summary, attention, learning, and memory are highly integrated, dynamically interactive processes. This interactivity is reflected both in heavy overlap between the brain systems involved in attention and those involved in L&M, and in the mutual influences of attention on memory processing. The following sections describe the neurobiological mechanisms of attention–memory interactions in more detail.

Neuronal Macrosystems Mediating Attentional Functions and Capacities: Major Research Themes

Both lesion studies and functional neuroimaging provide evidence for a distributed cortical network of prefrontal and posterior parietal regions involved in attention. The specific regions involved vary according to the nature of the attentional task. Demands on sustained attention, irrespective of the modality of targets, activate frontal and parietal regions primarily in the right hemisphere. Selective attention and demands for nontarget filtering recruit cingulate and other prefrontal regions, again primarily in the right hemisphere. Bilateral frontoparietal regions typically are activated by tasks requiring the division of attention between multiple targets.

Ongoing research, particularly experiments employing functional magnetic resonance imaging (fMRI), continues to refine the attribution of aspects of attentional performance to specific cortical networks. Results from this research confirm the conceptualization of anterior and posterior attention systems. Posner and Peterson proposed that the anterior attention system, consisting mainly of cingulate and other prefrontal regions, mediates target detection or the processes involved in the subject's consciousness of a signal's presence through producing a response documenting its detection. The posterior attention system involves the posterior parietal cortex and collicular and thalamic regions and controls management of the visual-attentional space.

Essential insights into the neuronal mechanism mediating attentional processes have been gained from research focusing on bottom-up versus top-down selection of stimuli. Bottom-up selection is a function of the intrinsic properties of the stimulus; stimuli compete and cooperate for detection and processing depending on their salience. Bottom-up, stimulus-driven attention is associated with increases in activity in sensory and sensory-associational regions representing the actual stimuli and their location. In addition, increased activity in frontoparietal networks indicates that stimulus-driven attentional control may also influence the executive management of attentional priorities and resources. In contrast, top-down selection is a function of experience, expectations, or instructions. It involves the tuning of receptive field properties, (sustained) anticipatory activity, and suppression of activity in irrelevant regions or modalities, all in order to optimize the detection and processing of expected targets. Top-down attentional control is generally thought to be executed by prefrontal modulation of parietal networks.

The actual mechanisms allowing prefrontal regions to initiate top-down effects, and the cognitive mechanisms and neuronal circuits mediating such top-down effects have remained largely unclear. Likewise, the neuronal circuitries mediating the bottom-up enhancement of stimulus processing, usually in a highly topographic fashion, are not known. As discussed further below, evidence on the role of neurotransmitter-specific cortical input systems provides the basis for hypotheses about the neuronal mechanisms modulating the detection of stimuli as a function of top-down, voluntary attention versus the bottom-up influences based on stimulus properties (see also **Figure 1**).

Evidence from recent fMRI experiments questioned the widely held view that the suppression of irrelevant information or modalities represents an important aspect of prefrontally controlled top-down regulation of attentional mechanisms. These studies demonstrated that prefrontal mechanisms act to amplify detection of task-relevant information and did not find consistent evidence for top-down suppression of irrelevant stimuli. However, other investigators have found evidence that instructions to ignore one stimulus set in order to more successfully memorize another caused brain activity associated with the ignored stimulus set to be suppressed below a perceptual baseline. As will be further discussed below, the view that selective detection is primarily mediated via enhancement of target stimuli, with little contribution of suppression of nontargets, corresponds with the attentional functions of the cortical cholinergic input system.

Intuitively, motivation is an important factor for engaging top-down mechanisms in order to combat fatigue- or distractor-related decline in attentional performance, to stabilize impaired performance, or to recover from performance impairments. The effects of motivational processes on top-down attention and the neuronal mechanisms via which motivation accesses the anterior attention system to facilitate top-down effects in challenging situations are emerging as an important theme. Increasing the incentives for attentional performance is associated with enhanced activity in frontoparietal attentional networks, and with activation in additional limbic regions involved in processing errors and response outcomes. Hypotheses concerning the interactions between motivational and attentional processes focus on the regulation of frontal–parietal attention systems via interactions with ventral striatal circuitry known to process reward, reward expectation, and prediction errors.

While motivation is a critical factor for top-down attention, stimuli associated with affective information exhibit superior potency for the bottom-up capture of attentional resources. For example, people in a crowded bar manage to ignore the TV until an emotionally charged symbol or action shifts their attention effectively from ongoing activities to the screen. The neuronal mechanisms allowing such stimuli to capture attentional resources and then drive the subsequent engagement of top-down mechanisms are not well known. However, it is safe to hypothesize, as illustrated in **Figure 1**, that such stimuli access the brain's main attention systems via reciprocal connections with prefrontal regions (not shown) and with ascending neuronal systems crucially involved in the mediation of attentional functions.

Neurotransmitter-Specific Projection Systems in Attention and Learning

Substantial evidence supports the hypothesis that the cortical cholinergic input system contributes

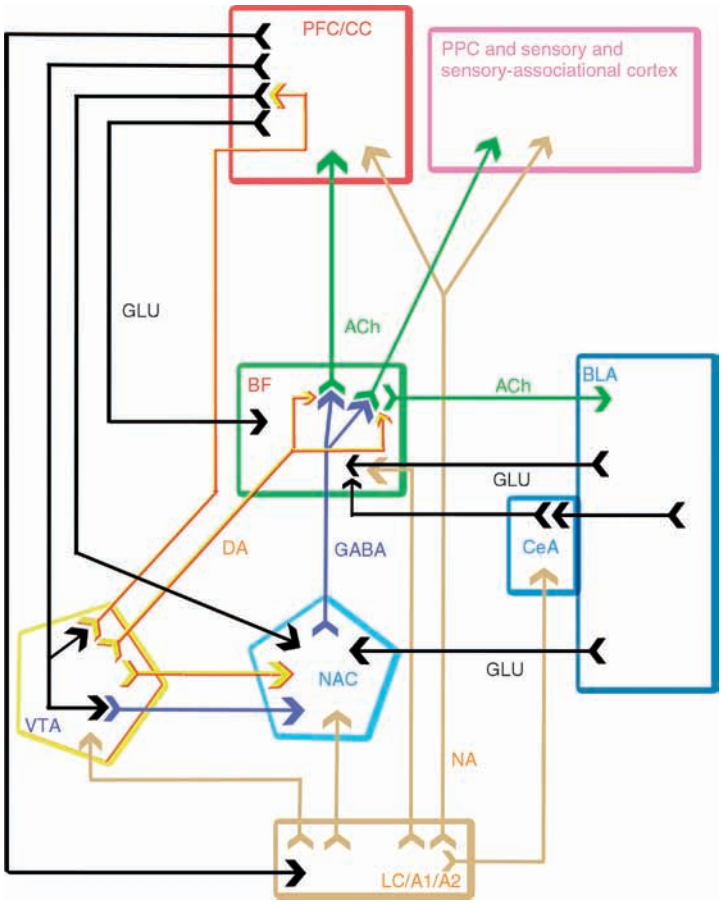

Figure 1 Schematic illustration of the main components of a neuronal network mediating attentional performance and the motivated activation of the cortical cholinergic input system in order to counteract the effects of challenges on attentional performance. The illustration depicts and emphasizes certain direct neuronal projections on the basis of anatomical evidence and importance in the present context. As discussed in the text, neurons using acetylcholine (ACh) as transmitter originate from basal forebrain (BF) regions and innervate all cortical regions, including the prefrontal cortex (PFC), cingulate cortex (CC), and somatosensory and posterior associational regions, including the posterior parietal cortex (PPC). Increases in cortical cholinergic activity contribute to the recruitment of anterior and posterior attention systems and the cholinergic amplification of input processing in sensory and sensory-associational regions. Cholinergic modulation of the PFC and CC is also involved in the implementation of top-down mechanisms, and the cholinergic inputs to posterior cortical regions are a component of the prefrontal efferent circuitry mediating such top-down effects. Increases in top-down effects that are a result of challenges on attentional performance, as well as the subjects' motivation to stabilize residual performance or regain normal levels, depend in part on the regulation of BF neurons by direct projections from the dopaminergic (DA) ventral tegmental region (VTA) and indirectly, via the nucleus accumbens (NAC) and its GABAergic (GABA; γ-aminobutyric acid) projections to BF cholinergic neurons. These midbrain and ventral-striatal circuits mediate incentive information and, via feedback from the CC and PFC, information about errors and unpredicted reward contingencies. Therefore, cholinergic modulation by outputs from these regions forms the basis for the integration between motivational and attentional processes. Corresponding with this hypothesis is evidence indicating that control over cortical cholinergic activity by ventral-striatal regions is confined to prefrontal regions. In addition to signals predicting reward or reward loss, stimuli associated with salient affective qualities exhibit potent attention-capturing qualities. The cholinergic and glutamatergic (GLU) reciprocal connections between the basolateral amygdala (BLA), as well as additional projections from the amygdala to the BF via the central amygdaloid nucleus (CeA), are thought to be involved in the mediation of such processes. Finally, ascending noradrenergic (NA) projections originating from the locus coeruleus (LC) and the catecholaminergic cell groups in the medulla (A1/A2) innervate major components of this circuits and also receive afferent feedback from the PFC. Thus, ascending noradrenergic projections are hypothesized to contribute to specific aspects of attentional processing in parallel with those attributed to the BF cholinergic system.

essentially to the mediation of attentional performance. Cholinergic inputs contribute to both top-down and bottom-up modulation of detection processes, and represent a link between motivational systems and cortical attention systems. Originating from basal forebrain regions, cholinergic neurons receive main inputs from midbrain and telencephalic regions, including the prefrontal cortex, and innervate all cortical areas and layers.

Removal of cholinergic inputs results in persistent impairments in sustained, selective, and divided attentional abilities. Furthermore, attentional tasks selectively increase the release of acetylcholine (ACh) from cortical terminals. Evidence suggests

that the right hemispheric cortical cholinergic input system is dominant for sustained attention performance; this finding corresponds with evidence from human fMRI studies and from patients with lateralized cortical damage (as described above).

Recent data indicate that in the prefrontal cortex, a transient cholinergic signal is elicited specifically by successfully detected targets, but not by targets that were missed. The temporal characteristics of these cholinergic signals confirm that essential aspects of the detection process are mediated via increases in cholinergic neurotransmission in the prefrontal cortex. These aspects include disengagement from noncontingent ongoing behaviors or from internal, associational processing, as well as the initiation of target-associated behavioral responses. These findings correspond with Posner's original hypothesis that the anterior attention system mediates the detection of attention-demanding targets (as described above).

The cortical cholinergic input system is involved both in the activation of the top-down attention system itself and in the implementation of its downstream effects. Cholinergic inputs to prefrontal regions, including cingulate cortex, specifically contribute to the activation of the anterior attention system, thereby enhancing target detection and activating top-down mechanisms. From there, cholinergic inputs to more posterior cortical regions, including parietal cortex, serve as a branch of the prefrontal efferent circuitry that implements top-down influences on input processing in sensory and sensory-associational regions.

The idea that cholinergic function can be separated into top-down and bottom-up components is supported by the dissociable interactions of these components with other brain systems. For example, ventral-striatal modulation of cortical cholinergic activity is limited to the prefrontal cortex. As described above, ventral-striatal modulation, particularly by the nucleus accumbens, is considered critical in situations characterized by the motivated, top-down increase in attentional effort. The modulation of cholinergic inputs to prefrontal regions corresponds with the role of those prefrontal cholinergic inputs in activating top-down effects.

In contrast, bottom-up, stimulus-driven recruitment of the cortical cholinergic input system is believed to involve recruitment of the entire cholinergic projection system, thereby directly influencing sensory processes while also allowing the stimulus to influence the executive management of attentional priorities. Such broad recruitment of the basal forebrain cholinergic system by salient stimuli is mediated in part by ascending noradrenergic projections targeting the cholinergic neurons in the basal forebrain. Noradrenergic neurons in the brain stem are wired to receive information about visceral correlates of salient stimuli and thus 'import' information about salience to forebrain regions. Support for the hypothesis that noradrenergic–cholinergic interactions are involved in bottom-up capturing of attentional resources by salient stimuli was generated by experiments demonstrating, for example, that the cortical processing of salient stimuli is attenuated by loss of cholinergic inputs to the cortex or blockade of receptors for noradrenaline in the basal forebrain.

However, the ascending noradrenergic system may exert additional and more specific contributions to attentional performance. There are direct prefrontal projections to noradrenergic neurons in the brain stem, and noradrenergic neurons of the locus coeruleus are activated by attended stimuli. Both of these findings indicate that this ascending system is involved in attentional functions that parallel, at least in part, the attentional functions attributed to the cortical cholinergic input system. The overlaps, interactions, and dissociations between the attentional functions of basal forebrain cholinergic and brain stem noradrenergic systems represent an important research topic.

Likewise, the interactions between ascending dopaminergic, cholinergic, and noradrenergic systems likely are essential mechanisms in the mediation of attentional processing. The processing of stimuli capable of predicting reward, including the processing of prediction errors, involves mesolimbic dopaminergic systems. Reward and error signals are capable of capturing significant attentional resources, particularly in response to prediction errors. Thus, interactions between dopaminergic systems and the two other ascending modulatory systems may be necessary to optimize arousal and attentional processing. The dopaminergic recruitment of basal forebrain cholinergic neurons is based on tegmental projections to the basal forebrain, and indirectly on dopaminergic projections to the prefrontal cortex and nucleus accumbens and the innervation of the basal forebrain by these regions (see **Figure 1**). The dopaminergic regulation of the cortical cholinergic input system supports the integration of motivation- and attention-related processing. The prefrontal regulation of brain stem noradrenergic neurons and noradrenergic–cholinergic links represents a second major branch of the neuronal systems which, in concert, regulate attentional processes and resources (see **Figure 1**).

Ascending Modulatory Systems Mediating Attention: Involvement in Learning

Failure to attend and thus to detect a stimulus involves a failure to produce a representation of this stimulus

for encoding. Thus, it is not unexpected that the available evidence on brain regions involved in attention matches, at least partly, the prefrontal–parietal neuronal systems critical for L&M. Indeed, given the critical role of attention for effective encoding, efforts designed to dissociate brain regions involved in attention versus learning pose conceptual challenges. It appears that a system of cortical regions processes stimuli at levels giving rise to awareness about these stimuli; however, the source of such processing of stimuli, whether it is a result of attentional or mnemonic operations, is not indexed in these regions.

The overlapping roles of ascending modulatory systems in attention and L&M are less well documented. In fact, several animal studies on the effects of loss of cortical cholinergic inputs have suggested that L&M do not depend on the integrity of these neurons, and that therefore the neuronal circuitries mediating attention and L&M could be dissociated. However, the lack of effects of, for example, lesions of the basal forebrain cholinergic system on L&M in animals may reflect, at least in part, the limited degree to which conventional animal tasks for L&M assess attention-requiring encoding and retrieval of declarative information. By contrast, more recent studies using tasks that require involving attention to stimuli or attention shifts find that learning is readily impaired by manipulations of levels of cortical cholinergic neurotransmission. Recent studies in humans have begun to simultaneously measure the cognitive and hemodynamic effects of drugs modulating cholinergic neurotransmission ('pharmaco-fMRI'). These studies indicated that drug-induced increases in cholinergic transmission enhance attentional mechanisms and the selectivity of stimulus processing during encoding. These results confirm the overlapping role of the cholinergic system in attention and learning, at least in situations requiring the encoding of attended stimuli. Similar conclusions have been drawn based on data from experiments in animals and humans on the role of the noradrenergic system in attention and learning.

Attention and Learning: Relevance for Aging and Neurodegenerative and Neuropsychiatric Disorders

Normal and pathological aging, including mild cognitive impairments and the dementias, are characterized by coinciding impairments in attentional functions and learning.

In nondemented aging, reduced attentional function is typically thought to be the major source of age differences in forgetting. There is some debate as to the nature of age deficits in attention function, with some theories emphasizing age differences in the amount of attentional resources, whereas others emphasize age differences in the allocation of attention, particularly deficits in inhibition. Across these perspectives, there is broad agreement that the controlled, top-down aspects of attention are those that show the greatest declines, whereas more automatic, bottom-up influences are relatively spared. Decreases in top-down attention are often ascribed to atrophy and reduced function in prefrontal cortex and basal ganglia structures, the structures that typically show the largest volume reductions in both longitudinal and cross-sectional studies of normal aging. These changes are thought to be largely responsible for age differences in most areas of cognition, although recent longitudinal evidence suggests that changes in select medial temporal lobe structures may be greater than suggested by earlier cross-sectional studies. Moreover, age-related attentional impairments may be a result of dysregulation and eventual degeneration of basal forebrain cholinergic projections to cortical regions.

The idea that the reduced availability of attentional resources is a major contributor to age differences in L&M receives substantial support from studies showing that asking young adults to divide their attention between multiple tasks (thus reducing the attentional resources available to any single task) often leads to results similar to those of older adults under single-task conditions. This is especially the case when the performance measure is subsequent memory for material studied under divided attention. Like older adults under single-task conditions, young adults studying under divided attention conditions show less activity in left frontal brain areas associated with later successful memory than do young adults studying under single-task conditions. On behavioral tests, the memory costs of divided attention are often similar for young and older adults, but older adults show larger impairments on the secondary task.

On the positive side, reducing the demand for top-down control often improves both memory and brain function for older adults. For example, intentional memory instructions ('memorize the words') require participants to engage top-down control to choose and implement a strategy for processing words that will support their later memory. Older adults typically perform much worse than do young adults under these conditions, and show less activation in left prefrontal cortex areas involved in subsequent memory. By contrast, when given instructions that guide attention to the memory-supporting semantic aspects of a stimulus ("Does the word mean something abstract, or concrete?"), memory is

improved, and older adults often show prefrontal activations that are at least as great as those shown by young adults.

While initial findings focused on age-related underactivations of frontal brain regions involved in controlled attention and memory, more recent data show that in many cases, older adults show more activation, or activate additional brain regions that young adults do not. In many cases, this additional activation is linked to better performance, especially on memory tests. Patterns of under- and overactivation of frontal brain regions by older adults may be related to failures of top-down, 'proactive' control, and later 'reactive' attempts to compensate.

A recent meta-analytic review suggested that reduced or disrupted activations may be more common in right anterior frontal regions (often thought to be involved in high-level controlled strategic processing), perhaps reflecting a failure to adequately engage top-down processing to organize efficient task execution. By contrast, left frontal regions were more often associated with greater activation by older adults, perhaps reflecting greater engagement of lower-level processes in an attempt to compensate. Further supporting the idea of an age-related shift to reactive control, memory-related frontal brain activations have an extended time course for older adults under conditions with high control demands (remembering words studied once), but equivalent time courses to young adults under conditions with low control demands (remembering words studied 20 times). This shift is specific to frontal regions; parietal brain regions involved in successful recognition do not show a similar sensitivity to control demand.

Whereas attention and dopamine function are the focus of most research on normal aging, deficits in memory and cholinergic function are the characteristic features of Alzheimer's disease (AD). However, AD is also associated with marked deficits in attention. For example, the ability to divide attention between two ongoing tasks is substantially impaired in AD, even when performance on the individual tasks has been matched with that of healthy controls.

Attention deficits are not easily detected in the earliest stages of the disease, which is primarily marked by memory deficits. However, difficulties with attention precede the breakdowns in language and visuospatial function that occur at the moderate and severe stages of the disease. It is not clear which aspect of attention (selective, divided, or sustained) is the first to deteriorate in AD. Several reports indicate that very demanding selective or divided attention tasks show declines early in the disease process, whereas simpler sustained attention tasks only show deficits at later stages or under conditions (e.g., stimulus degradation) that increase difficulty. However, there have also been reports of preserved performance on selective attention tasks in patients with reduced sustained attention performance.

Within a domain, the detection of AD-related deficits in attention may depend not only on disease severity but also on the specific processes being tested. In selective attention, patients are often able to shift and engage attention appropriately in response to a valid cue. However, they appear to be impaired at disengaging attention from an inappropriate location following an invalid cue, especially if the task involves discrimination of a target stimulus from nontargets rather than simple detection.

For both normal aging and AD, a heuristic description is that attention performance deficits become more obvious with increased complexity or control demand. Functions that require little or no top-down control generally show little or no deficit. Examples include the engagement of spatial attention by the sudden onset of a peripheral cue, or the inhibition of attention's return to a previously attended location.

Deficits can be especially large on so-called 'executive function' tests, which require top-down control in order to override automatic responses driven by bottom-up stimulus characteristics (e.g., moving one's attention away from a peripheral cue in the antisaccade task, responding to ink color rather than word identity in the Stroop task) or to maintain and rapidly switch between multiple response rules (e.g., dual-task procedures). The memory problems of AD may contribute to poor performance on tasks designed to measure attention, as patients may have difficulty both with remembering the response rule and with its execution.

The specific abnormalities in the regulation of integrity of neuronal systems responsible for the age- or dementia-related impairments in attention and encoding are still not well understood. However, structural and regulatory age differences in basal forebrain cholinergic projection have been extensively documented in brains of humans and animals. Furthermore, the development and maintenance of forebrain cholinergic neurons depend on nerve growth factor (NGF) signaling via several neurotrophic receptors expressed selectively by cholinergic neurons. The availability of such receptors is dramatically reduced in the brains of subjects with mild cognitive impairments. Given the central role of the cholinergic system in attention (as described above), these reductions support the hypothesis that the attentional and related cognitive impairments observed in these patients are due, at least to a significant extent, to dysregulation of forebrain

cholinergic neurons. Subsequent loss of forebrain cholinergic neurons during the onset and progression of dementia contributes essentially to the severity of impairments in attention and memory.

Impairments in both the activation of relevant stimuli and the inhibition or filtering of irrelevant stimuli are among the fundamental cognitive dysfunctions of schizophrenia. These disruptions of attentional abilities are reflected in impaired L&M. For example, the exhaustion of attentional resources for encoding of relevant stimuli contributes to impaired maintenance and updating of patients' memory, thereby perhaps contributing to the development of aspects of positive symptoms. Abnormal metabolic responses to cognitive challenges have been observed in frontoparietal networks, although the specific neuronal mechanisms underlying the attentional dysfunctions of schizophrenia are not fully understood. However, available evidence strongly suggests an abnormally reactive dopamine system in schizophrenia. Given the role of dopaminergic–cholinergic interactions in the motivational regulation of attentional performance (as described above), both cortical cholinergic recruitment of the anterior attention system and the cholinergic mediation of input processing in sensory and sensory-associational regions are expected to be highly abnormal in schizophrenia. Evidence from animal models of this disease corresponds with this hypothesis, as does the demonstration of down-regulated muscarinic receptors in the cortex of schizophrenic patients. It is widely assumed that abnormalities in the development of cortical, particularly prefrontal, circuits represent a primary neuropathological foundation for schizophrenia. These abnormalities may in turn lead to the dysregulation of the ascending, bottom-up modulatory systems (as described above). Moreover, the dysregulated release of these neuromodulators interacts with defective cortical target circuits, collectively mediating the attentional and encoding dysfunctions characteristic for schizophrenia.

See also: Aging and Memory in Humans; Attentional Networks; Attentional Networks in the Parietal Cortex; Basal Forebrain and Memory; Cognition: An Overview of Neuroimaging Techniques; Cognitive Deficits in Schizophrenia; Executive Function and Higher-Order Cognition: Neuroimaging; Humans; Memory Representation; Prefrontal Cortex: Structure and Anatomy; Prefrontal Contributions to Reward Encoding; Psychophysics of Attention; Reward Systems: Human.

Further Reading

Aston-Jones G and Cohen JD (2005) An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neurosciences* 28: 403–450.

Buckner RL (2004) Memory and executive function in aging and AD: Multiple factors that cause decline and reserve factors that compensate. *Neuron* 44: 195–208.

Cabeza R and Nyberg L (2000) Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience* 12: 1–47.

Corbetta M and Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* 3: 201–215.

Craik FIM and Byrd M (1982) Aging and cognitive deficits: The role of attentional resources. In: Craik FIM and Trehub S (eds.) *Aging and Cognitive Processes*, pp. 191–211. New York: Plenum.

Duclukovic NM and Wagner AD (2006) Attending to remember and remembering to attend. *Neuron* 49: 784–787.

Fernandes MA and Moscovitch M (2000) Divided attention and memory: Evidence of substantial interference effects at retrieval and encoding. *Journal of Experimental Psychology: General* 129: 155–176.

Kruschke JK (2003) Attention in learning. *Current Directions in Psychological Sciences* 12: 171–175.

Mesulam M (2004) The cholinergic lesion of Alzheimer's disease: Pivotal factor or side show? *Learning and Memory* 11: 43–49.

Mufson EJ, Kroin JS, Sendera TJ, and Sobreviela T (1999) Distribution and retrograde transport of trophic factors in the central nervous system: Functional implications for the treatment of neurodegenerative diseases. *Progress in Neurobiology* 57: 451–484.

Perry RJ, Watson P, and Hodges JR (2000) The nature and staging of attention dysfunction in early (minimal and mild) Alzheimer's disease: Relationship to episodic and semantic memory impairment. *Neuropsychologia* 38: 252–271.

Reuter-Lorenz PA and Lustig C (2005) Brain aging: Reorganizing discoveries about the aging mind. *Current Opinion in Neurobiology* 15: 245–251.

Sarter M, Bruno JP, and Givens B (2003) Attentional functions of cortical cholinergic inputs: What does it mean for memory? *Neurobiology of Learning and Memory* 80: 245–256.

Sarter M, Gehring WJ, and Kozak R (2006) More attention must be paid: The neurobiology of attentional effort. *Brain Research Reviews*. 51: 145–160.

Sarter M, Hasselmo ME, Bruno JP, and Givens B (2005) Unraveling the attentional functions of cortical cholinergic inputs: Interactions between signal-driven and top-down cholinergic modulation of signal detection. *Brain Research Reviews* 48: 98–111.

Sarter M, Nelson CL, and Bruno JP (2005) Cortical cholinergic transmission and cortical information processing following psychostimulant-sensitization: Implications for models of schizophrenia. *Schizophrenia Bulletin* 31: 117–138.

Wager TD, Jonides J, and Reading S (2004) Neuroimaging studies of shifting attention: A meta-analysis. *Neuroimage* 22: 1679–1693.

Attentional Mechanisms in Ventral Pathway

L Chelazzi, C Della Libera, and E Santandrea, University of Verona, Verona, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Perhaps the most obvious form of visual selective attention is when individuals turn their gaze toward a salient or otherwise interesting object in their surroundings to align it with the high-resolution fovea of the retina. This allows more detailed processing of the fixated object at the expense of competing objects falling on peripheral regions of the retina. However, it is well established that selective attention can be aimed at extrafoveal locations and objects, thus effectively decoupling the high-resolution power of the fovea from enhanced central processing due to selective attention.

There appear to be several computational reasons why the brain implements selective attention mechanisms. First and foremost, selective attention can be viewed as the mechanism that mediates selection of the next target for preferential, foveal analysis. In this vein, selective attention primarily assists the oculomotor system to optimize sensory sampling of the visual environment, given the current goal. Second, motor systems in general, including those for reaching and grasping movements, are physically constrained and can act only on one (or a few) objects at any given moment. Therefore, selective attention is needed to focus processing onto a single object in order to plan coherent behavioral responses targeted at the selected object. Third, it is probably impossible, or perhaps simply disadvantageous, for the memory systems of the brain to store each and every single object and event occurring within a crowded environment. Therefore, selective attention is necessary to gate access of perceptual representations to memory systems. Finally, perceptual awareness is inherently limited in nature, and it unfolds serially, with a single perceptual representation gaining dominance at any instant in time. Therefore, selective attention is needed to allow entrance of the selected representation into working memory and conscious perception. On top of all the reasons above, students of vision and selective attention raise two further reasons why selective attention may be indispensable, and they are probably related. One reason is that processing of incoming retinal input must be focused on a single object at a time simply because, otherwise, processing and recognition of all objects simultaneously would overcome the limited processing capacity of the system. In addition, and more specifically, directing attention toward a single object at a time might serve the important function of aiding the correct conjoining of all its elemental features, therefore preventing the erroneous binding of features belonging to separate objects in a cluttered scene. In brief, selective attention appears to be a key mechanism aiding efficient object recognition, perceptual awareness, goal-directed behavior, and selective memory storage.

Given the key role of selective attention in visual processing, in particular its role in building and gating object representations, it comes as no surprise that much of the relevant experimental work over the past two decades has been devoted to the investigation of the neuronal correlates of selective attention along the ventral pathway of cortical visual processing. The ventral pathway originates at the level of primary visual cortex, or V1, and nearby secondary visual cortex, or V2, and further extends through extrastriate area V4 and posterior inferotemporal cortex, or area TEO, to culminate in a relatively vast cortical territory occupying the middle and anterior segments of the inferotemporal (IT) cortex. The ventral pathway represents a network of interconnected areas, largely organized according to a hierarchical scheme, whereby object representations are created with an ascending level of complexity and representational invariance. Ultimately, patterns of activity within IT cortex are now known to represent with remarkable speed and efficiency the various objects we are able to recognize. Key nodes along this pathway are represented by area V4 and the various sectors of IT cortex, and therefore this article will focus on these nodes of the pathway.

Studying the Manifestations versus the Control

In speaking of selective attention, one should distinguish between the manifestations and the causal control mechanisms. Specifically, one may use the term 'selective attention' to refer to the modulation of sensory processing along the ventral pathway in relation to concurrent changes in behavioral performance. In this case, the term would index the manifestations of visual selective attention at the neuronal, as well as at the behavioral, level. In contrast, one may use the term selective attention to refer to the signals that, within a given behavioral context, bring about the manifestations of attention considered above. The latter signals may or may not originate within the visual system, and the available evidence suggests that in most cases they do not. This article

mainly concentrates on the manifestations of selective attention within the ventral pathway but also briefly discusses available evidence concerning the signals impinging on the ventral pathway to exert attentional control. Several forms of selective attention are distinguished, including enhanced processing of individual attended items, selective processing among competing items (or biased competition), feature-based guidance of target selection in visual search, and finally, feature-selective attention.

Enhanced Processing of Attended Objects: The Beneficial Effects of Spatially Directed Attention

Psychophysical studies of human observers have documented robust effects of attention on visual sensitivity at selected regions of space. Sensitivity has been shown to increase at attended versus unattended locations in the visual field, with relatively shorter reaction times to detect an item at the attended location, as well as greater accuracy. In particular, attentional facilitation entails better detection of faint, low-contrast stimuli and improved discrimination of their features, as if attention to the stimulus led to enhancement of signal strength. In turn, these effects are reminiscent of those produced by an increase in stimulus contrast, and it has been reported recently that indeed attention increases perceived stimulus contrast. Consistent with these behavioral results, single-unit recording studies in the behaving macaque have found enhanced neuronal responses (e.g., in area V4) to a single stimulus presented inside the receptive field (RF) of the recorded neuron when the animal's attention is aligned with the stimulus location, relative to when attention is directed elsewhere in the visual field. As a result, stimuli at an attended location engender stronger central representations than unattended stimuli do. These neural effects likely represent part of the mechanism underlying enhanced behavioral performance, as previously described. Overall, however, enhancement of neuronal responses to individual stimuli presented inside the RF is not very strong, typically on the order of 20%. Furthermore, the effect has not been found in all reported studies, and a possible account of this variability is offered below.

Recent studies of neuronal responses in area V4 have shed further light on the above modulation of responses to single RF stimuli as a function of attention. If the effects of attention are akin to those brought about by increased stimulus contrast, then one might predict that directed attention changes the contrast response function of neurons. Neurons at many stages of the visual system produce increasing responses as a function of stimulus contrast, up to a plateau, and the function takes the form of a sigmoid. If attention acts by increasing the effective contrast of the RF stimulus, then one predicts a leftward shift in the contrast response function of the neurons. In line with this prediction, it was found that attention directed to an RF stimulus causes a leftward shift of the sigmoid relative to when the stimulus is unattended. As a consequence, responses to an attended stimulus will not differ reliably from those to an unattended stimulus at or beyond the point of saturation in the contrast response function. Instead, effects of attention will be greatest within – or just below – the dynamic range of the contrast response function of the neuron. These findings may explain, at least in part, why not all single-cell recording studies have found enhanced responses to attended compared with unattended single RF stimuli along the ventral pathway, including area V4, since attentional effects may be minimal, if any, when stimuli of high contrast are employed. In summary, prevailing evidence indicates that stimuli presented at attended locations will elicit greater responses compared with stimuli at ignored locations. However, the effect is relatively large with stimuli of low contrast, whereas it tends to decrease with contrast of the stimulus.

Notice that there is an important difference between the effects of directed attention and contrast on responses of neurons in visual cortex, including area V4; namely, while attention mimics the effect of contrast in terms of response magnitude, it does not do so in terms of response onset latencies. Response latency has been shown to increase considerably for low-contrast stimuli, whereas no detectable change in response latency is associated with manipulations of attention. This imposes some caution in likening effects of attention to changes in effective stimulus contrast.

An important question is whether directed attention, in addition to changing the strength of neuronal responses, will modify neuronal tuning for the stimulus features, for example, stimulus orientation. This has been addressed in a number of studies, and the prevailing view is that tuning properties of neurons are relatively immune to the influence of directed spatial attention, although they may be modified as a result of extensive discrimination training with perceptual learning protocols. Instead, spatially selective attention has been shown to cause a multiplicative scaling of tuning curves. Responses throughout the tuning curve will be multiplied by a constant factor, with no appreciable changes in the filter properties of neurons. Again, this is similar to the known effect on tuning curves of varying stimulus contrast. Nonetheless, it is conceivable that at the population

level, a gain modulation of tuning curves allows finer encoding of features at an attended location than at an unattended one, for instance by reducing the signal-to-noise ratio at the attended location.

Top-Down Control: Biases and Baseline Shifts

Given the distinction between manifestations of attention in sensory processing areas and control signals, researchers have sought evidence for control signals that may cause the manifestations of attention summarized previously. A key feature of these signals is that they ought to precede onset of task-relevant stimuli, that is, they should be present while the animal is attending to a given visual field location in preparation for performing a task on some relevant item. In practice, people have compared baseline activity of the neurons during the waiting period of the task between conditions in which the animal's attention was directed toward the RF of the studied neuron versus when attention was directed toward some location outside the RF. Single-unit recording studies have shown that neurons in area V4 (and V2) display elevated baseline firing during periods when the animal is attending to a location inside the RF of the neuron in anticipation of RF stimulus onset. It is interesting that analogous changes in baseline activity depending on the direction of spatial attention can be observed even when one compares attention to different locations inside the single RF, provided that the locations to be compared are not equally sensitive. Specifically, baseline activity has been found to covary with the strength of the visually evoked responses at any given location within the RF. Increases in baseline activity are typically small in absolute terms, on the order of a few spikes per second. However, in fractional terms, they can amount to a 50% increase in firing rate in the absence of visual stimulation. Therefore they represent a substantial percentage increment in neural activity over a relatively large population of neurons, those neurons with RFs encompassing the attended location.

The accepted account of elevated baseline activity due to spatial attention is that it reflects the influence of an incoming signal, originating in areas of the brain responsible for exerting control over spatially directed attention. In the case of area V4, these likely include areas in the posterior parietal and prefrontal cortices, although subcortical sources (e.g., the superior colliculus) have also been implicated. Recent evidence obtained with low-current electrical microstimulation has directly demonstrated that signals of this sort may originate at the level of the frontal eye fields, and they are capable of enhancing visual responses within area V4 at selected visual field representations. The same type of microstimulation was also shown to improve the animal's performance in a demanding stimulus detection task. Analogous effects on behavior have been obtained with electrical microstimulation of the superior colliculus, although it remains to be established whether microstimulation of the superior colliculus would also enhance neuronal responses in area V4 or other areas along the ventral pathway. In summary, the available data suggest that a number of cortical and subcortical regions are involved in delivering control signals for spatially directed attention, and they largely overlap with critical nodes of the circuit controlling saccadic eye movements. These signals are likely responsible for elevated baseline firing in areas along the ventral pathway when attention is directed to a location inside the RF of the recorded neuron. The elevated baseline firing, in turn, may be part of the mechanism that confers to the neurons increased sensitivity to visual stimulation.

Coherent Firing at the Population Level

Recent work has shown that attention to an RF location may entail, not only elevated baseline firing and enhanced responses to an RF stimulus, but also increased synchronization of firing among the relevant neurons. Increased synchronization may or may not take the form of oscillatory activity, but typically it does. Therefore evidence is rapidly accruing to indicate that when attention is directed to a given location in the visual field, neurons with RFs encompassing that location will entertain an enhanced coherent firing, usually in the gamma-band frequency range, around 50 Hz. In turn, increased synchronization of firing at the attended location may enhance synaptic transmission downstream of the considered neural population, effectively amplifying transmission of information in a spatially selective manner. It is interesting that this effect has also been observed under task conditions in which there was no consistent change in the magnitude of visual responses as a function of spatial attention. Therefore, enhanced processing and spike transmission at the attended location may take the form of increased firing, increased synchronization, or both. It remains to be established whether there are specific task parameters that lead preferentially to one or the other manifestation of attentional modulation.

Competitive Interactions among Multiple Visual Stimuli

A special problem for perceptual and attentional mechanisms to solve is one in which multiple stimuli

are presented together (crowding) and an individual must select the relevant stimulus while at the same time discarding any potential distracter, in particular nearby distracters. In neurophysiological terms, this translates to conditions in which multiple stimuli impinge simultaneously onto the RF of an individual neuron and they compete for controlling the neuron's firing pattern. It has been demonstrated that neurons along the ventral pathway, including area V4 and the IT cortex, produce responses to two or more stimuli falling inside their RF that approximate the average of the responses elicited by the component stimuli presented in isolation. In other words, neurons in areas of the ventral pathway seem to be incapable of clutter invariance, a property that, if present, would allow neurons to encode the single most preferred stimulus inside the RF while automatically discarding other nearby stimuli, effectively implementing a MAX operation. In contrast, it appears that multiple stimuli falling inside a single RF, or in its immediate surroundings, compete for the encoding capacity of the neuron, and the neuron's firing is ambiguous as to which stimulus is encoded. It seems that competitive interactions among multiple RF stimuli are only weakly affected, if at all, by the specific nature of the stimuli involved, including the degree of their similarity (but see the discussion of luminance contrast in the section titled 'Top-down versus bottom-up in selective attention'), except that the greatest competitive interactions occur with stimuli far apart in their ability to drive a neuron's visual response, as when a highly preferred or a null stimulus is involved. Under these circumstances, the presence of the null or ineffective stimulus can drive the response to the preferred stimulus well below the level that it would have elicited if presented alone. An important notion that has emerged from these studies is that a stimulus, which causes only modest changes in firing rate when presented in isolation, can nevertheless exert a profound (suppressive) influence on the neuronal firing when presented in combination with an effective stimulus, thus demonstrating clear-cut decoupling between effectiveness of stimuli in driving a response from the given neuron and their effectiveness in determining the firing rate of the same neuron. In terms of the latter property, an ineffective stimulus can be no less effective than a highly preferred, or optimal, stimulus in exerting control over the neuron's activity. In area V4, competitive interactions of this sort have been shown to span a limited extent of visual space, covering the RF size and extending only a small distance beyond the boundary of a neuron's RF. In contrast, competitive interactions sometimes span a much larger extent of the visual field in IT cortex, including portions of the visual hemifield ipsilateral to the recorded hemisphere. However, it has been reported that competitive interactions in IT cortex are much weaker, or nearly absent, when competing stimuli are placed across the vertical meridian, as if competitive interactions could not come about at full strength when they involve the midline commissures (e.g., the corpus callosum). Selective attention mechanisms are needed to resolve these competitive interactions – the core notion of the biased competition model of attention.

Resolving the Competition: Selection and Filtering

The biased competition model of attention has been highly influential over the past decade, as it can account for a great deal of experimental observations obtained with a variety of approaches and techniques, in both human and animal studies of perception and attention. Mathematical and neural network implementations of the model have been developed. The model rests on two tenets. First, as discussed in the previous section, multiple stimuli falling within the RF of a given neuron (or in its immediate surround) compete for controlling the neuron's firing rate. The most compelling evidence of this takes the form of suppressed responses to an effective stimulus falling inside the RF of a cell when it is paired with a second, ineffective stimulus for the cell, with responses to the pair approaching an average of the responses elicited by each of the two stimuli in isolation. Second, competition among stimuli can be resolved when a signal biases the competitive interaction in favor of either stimulus in the pair, thus causing the cell's firing to be primarily determined by the favored stimulus. When this occurs, selective attention is enacted: one of the competing stimuli is selected; the other is filtered out of the RF, or ignored. In cell physiological terms, a neuron's firing to multiple stimuli impinging on its RF as if only one of them were present – the favored one – corresponds to selecting a salient or otherwise relevant stimulus while discarding distracters.

As already considered, the biasing signal for spatially directed attention may take the form of elevated baseline activity of the relevant neural population, but the proposal has been made that increased synchronization of firing across the population of neurons with RFs encompassing the attended location may as well bias competition in favor of the relevant stimulus location. It remains to be established to what extent increased baseline firing and enhanced synchronization are related phenomena in functional terms. Regardless of this, we have already mentioned that likely sources of signals biasing competition in favor of the attended location include cortical areas

such as the frontal eye field and lateral intraparietal area, in the frontal and posterior parietal cortex, respectively, as well as subcortical structures, such as the superior colliculus. Future work might well reveal that other parts of the brain, at the cortical and subcortical level, play as important a role in controlling spatial attention.

Selection of a relevant object (or target) among competing stimuli can be achieved not only on the basis of its location in space, but also on the basis of its feature composition. For example, in visual search tasks, an observer is asked to find a target object among irrelevant distracters. Under some conditions, the target may be found easily, at no increasing cost as a function of the number of distracters, such as when it is characterized by some unique property (known as 'pop-out'). In contrast, under less efficient conditions, locating the target may take some effort and increasing time as a function of the number of distracters. By means of search tasks of the latter kind, it has been shown that neurons in areas V4 and IT may contribute significantly to the search process. In particular, as the search process unfolds, neurons in both areas come to encode the target but much less, or not at all, the distracters. Specifically, while neural activity shortly after search array onset to some extent represents all items in the array, later on, in anticipation of the behavioral response, only the target item activates the neural population which is selective for its constituent features, while neural populations activated by the features of the distracters are strongly suppressed. This form of selective attention has been shown to engage underlying mechanisms similar to those engaged by spatially directed attention, except that here, selection is guided by feature information. It has been further suggested that control signals for feature-based selection of a target object likely originate in at least partly different brain regions from those involved in delivering control signals for spatially selective attention. The proposal has been made that feature information specifying the target item and guiding its ultimate selection is represented within brain networks responsible for holding object feature information online during the execution of the task (working memory).

Behavioral evidence obtained following lesion or deactivation of area V4 (and/or TEO) in the monkey is in full agreement with the biased competition model of attention. This work has elegantly shown that, when selective attention mechanisms are knocked out, the animal is at the mercy of stimulus salience. In other words, when multiple stimuli are presented and the animal must select a high-salience target among low-salience distracters, behavior is largely unimpaired. Conversely, when the animal is required to select a low-salience target among high-salience distracters, performance shows a dramatic drop. Consistent with a key role of area V4 in the implementation of attention mechanisms, this deficit has been observed following lesion of area V4 in the macaque, as well as following damage to the homolog of area V4 in the human brain. These findings suggest that, when the mechanisms for cognitively mediated selection are compromised, such as can be obtained through damage to area V4 (and TEO), competitive mechanisms and selection are primarily controlled by the intrinsic salience of objects, the topic of the next section.

Top-Down versus Bottom-Up in Selective Attention

There is now evidence at the single-cell level that competitive interactions among multiple stimuli falling inside the RF of an individual V4 neuron are directly modulated by stimulus salience, such as can be obtained by varying the luminance contrast of the stimuli. As already noted, with attention directed well outside the RF of the recorded neuron (e.g., to the opposite visual hemifield), adding an ineffective stimulus reduces responses of V4 (and IT) neurons to a concurrently presented effective stimulus for the neuron. It has also been recently shown that the suppressive effect is progressively stronger as the luminance contrast of the ineffective stimulus is increased, with the contrast of the effective stimulus held constant at an intermediate level (40%). Although the suppressive effect increases with contrast of the ineffective stimulus, notice that at the same time, the ineffective stimulus presented alone elicits a progressively larger, albeit weak, visual response when its contrast is increased. This again indicates a remarkable dissociation between the efficacy of a stimulus to drive a visual response from a neuron and efficacy of the same stimulus to control the neuron's firing. A stimulus that, for its feature composition, may be largely ineffective in driving a visual response from a given neuron, can nonetheless be highly effective in determining the neuron's response, due to its salience, or strength, such as its high luminance contrast. Moreover, within the same experimental context, attention directed to the ineffective stimulus in the pair has been shown to further enhance the suppressive effect exerted by this stimulus to the point that attention to a high-contrast ineffective stimulus almost completely dominates the cell's firing, namely, it almost completely silences the cell. These findings indicate that competitive interactions are entertained automatically within visual cortex and that competition can be resolved in favor of a high-salience

(e.g., high-contrast) stimulus in bottom-up, in the absence of top-down signals reflecting the current volitional control on selective attention.

Feature-Based Attention

As noted previously, selective attention can be directed toward a specific spatial location, or it can be guided by feature information specifying the target-defining properties. Furthermore, behavioral evidence in humans indicates that feature-based attention can affect processing throughout the entire visual field, in a parallel fashion. Consistent with this, single-unit recordings from area V4 of the macaque have revealed the correlates of this form of nonspatial selection. It has been discovered that neuronal responses to any potential target in the visual field – that is, any element that shares one or more of the target-defining features, including the target itself – are enhanced as the search process progresses, long before the animal actually locates the designated target. In other words, this form of feature-based attention is able to 'highlight' all the objects in the visual array that are potentially relevant for the task at hand. Essentially, the mechanism allows privileged processing of these objects, while other objects are effectively filtered out in parallel across the visual array. Although findings of this kind have come in slightly different flavors in the literature, perhaps related to specific characteristics of the experimental protocols, all converge to indicate that among the entire population of neurons in area V4 activated by the array elements, the neurons firing at the highest rate will be those directly stimulated by a feature in the RF that matches the feature preference of the neurons (e.g., red) while the animal is searching for a target item defined by the same feature (e.g., red). Evidently, depending on the currently relevant features, a specific control signal can target the neuronal populations with RFs anywhere in the visual field that are selective for the corresponding features.

Gating-Feature Information, or Feature-Selective Attention

Unlike the form of feature-based attention discussed in the previous section, feature-selective attention is engaged under task conditions in which an individual is asked to identify, or otherwise respond to, a specific object feature while at the same time ignoring other features of the same object. This form of feature-selective attention, therefore, entails that the unity of perceptual objects be broken down in order to cope with the current task. Feature-selective attention plays an important role in many real-life situations, for instance when an individual wishes to sort, or classify, objects on the basis of one elemental feature (e.g., color) and other features (e.g., shape and texture) must be ignored. In addition, this type of feature-selective processing is tapped by a number of classical neuropsychological tests, such as the Stroop test and the Wisconsin card sort test. In both cases, performance must be guided by selective feature information, and interference from the irrelevant feature or features must be blocked. The neuronal underpinnings of the latter form of feature-selective attention have been systematically explored in a recent single-unit recording study in which the activity of V4 neurons was recorded while an animal was attending to either one or the other feature of differently colored, oriented bars. It was found that, under these task conditions, responses of V4 neurons to otherwise identical stimuli are modulated depending on the component feature of the stimulus being currently attended. Most important, it turns out that a large fraction of the recorded neurons are able to cluster the attended features of the stimuli into one or the other of two behaviorally relevant response categories, indicating that area V4 may be important in the process of converting selected feature information into a categorical code available to guide the animal's behavioral responses.

Conclusions

Research over the past 25 years has allowed impressive progress in the understanding of the brain mechanisms underlying the ability to concentrate mental resources on a single location or object at any given time – an essential component of the ability to implement goal-directed behavior. Fundamental pieces of evidence have come from neurophysiological investigations in the awake, behaving macaque monkey. Science is very close to a full understanding of what it means at the single-neuron level to pay selective attention to a specific location or object, or object feature, including the fine details of the circuitry that brings about attentional modulation of firing in visual cortical areas, as well as the source and nature of the signals that control the same circuitry, thus initiating attention-related phenomena at the neuronal and behavioral level. The investigation of the neuronal correlates of visual selective attention along the ventral pathway of cortical visual processing has been particularly successful at identifying specific ways in which mechanisms for selective attention are intertwined with perceptual mechanisms for feature analysis and object recognition.

See also: Attention and Eye Movements; Attention: Models; Attentional Networks; Attentional Networks in the Parietal Cortex; Attentional Functions in Learning and Memory; Decision-Making and Vision; Neglect Syndrome and the Spatial Attention Network; Psychophysics of Attention; Vision for Action and Perception; Visual Attention.

Further Reading

Bichot NP, Rossi AF, and Desimone R (2005) Parallel and serial neural mechanisms for visual search in macaque area V4. *Science* 308: 529–534.

Chelazzi L, Duncan J, Miller EK, and Desimone R (1998) Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology* 80: 2918–2940.

Desimone R and Duncan J (1995) Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18: 193–222.

De Weerd P, Peralta MR III, Desimone R, and Ungerleider LG (1999) Loss of attentional stimulus selection after extrastriate cortical lesions in macaques. *Nature Neuroscience* 2: 753–758.

Fries P, Reynolds JH, Rorie AE, and Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291: 1560–1563.

Luck SJ, Chelazzi L, Hylliard SA, and Desimone R (1997) Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology* 77: 24–42.

Maunsell JH and Cook EP (2002) The role of attention in visual processing. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1063–1072.

Maunsell JH and Treue S (2006) Feature-based attention in visual cortex. *Trends in Neuroscience* 29: 317–322.

McAdams CJ and Maunsell JH (1999) Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience* 19: 431–441.

Mirabella G, Bertini G, Samengo I, et al. (2007) Neurons in area V4 of the macaque translate attended visual features into behaviorally relevant categories. *Neuron* 54: 303–318.

Moore T and Armstrong KM (2003) Selective gating of visual signals by microstimulation of frontal cortex. *Nature* 421: 370–373.

Motter BC (1994) Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience* 14: 2178–2189.

Reynolds JH and Chelazzi L (2004) Attentional modulation of visual processing. *Annual Review of Neuroscience* 27: 611–647.

Reynolds JH, Chelazzi L, and Desimone R (1999) Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience* 19: 1736–1753.

Zoccolan D, Cox DD, and DiCarlo JJ (2005) Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience* 25: 8150–8164.

Attentional Networks

N U F Dosenbach and S E Petersen, Washington University in St. Louis School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Attention is the brain's ability to selectively allocate cognitive resources to those stimuli, responses, memories, and trains of thought that are behaviorally most relevant, at the expense of less relevant ones. As William James pointed out in 1890, "Each of us literally chooses, by his ways of attending to things, what sort of universe he shall appear to himself to inhabit."

Attention can have widespread effects on behavior. It has been known since the time of William James that attention can improve our ability to perceive, conceive, distinguish, remember, and respond. For example, attending to the spatial location at which a visual target will occur decreases the time it takes to respond to it, even when subjects only covertly shift their attention to the cued target location in the absence of overt eye or head movements.

Attention has been shown to alter neural activity. Single-unit recording studies in macaque monkeys placed stimuli in the receptive fields of single cells and compared two conditions: (1) when the monkey was attending to the stimulus and (2) when it was attending somewhere else in the visual field. The classical finding has been that selectively attending to a stimulus increases the neuronal response to it. Similar effects have been seen in neuroimaging studies of selective attention in humans. Selectively attending to the color, shape, and motion of an object, for example, increases activity in extrastriate visual regions specialized for processing these features.

The attention-driven modulation of neural activity is not believed to be inherent to early sensorimotor regions of the brain. Instead, it is thought that so-called biasing signals from higher order source regions influence moment-to-moment processing in sensorimotor regions such as extrastriate visual cortex.

The notion that anatomically separate brain regions may control the selective allocation of attention was initially driven by studies of neglect patients. Unilateral lesions, particularly in temporoparietal and frontal cortex, often cause patients to neglect the contralateral half of extrapersonal space, a condition known as spatial neglect. Such patients may only dress the ipsilesional half of their body, only eat the food on the ipsilesional half of their plate, and only attend to stimuli in the ipsilesional half of their visual field. Studies showed neglect to be a primary deficit of attention, not of sensorimotor processing.

Principles of the Human Attention System

The combination of connectional anatomy, electrophysiology, lesion research, cognitive psychology, and positron emission tomography (PET) imaging allowed Posner and Petersen to build a theoretical cognitive neuroscience model of human attention. This model proposed attention to be the emergent property of a network of functional areas. Posner and Petersen formulated three principles of attentional networks that still seem relevant: (1) The brain's attention system is anatomically separate from those downstream systems that process specific inputs independent of whether these inputs are being attended to or not; (2) attention is the emergent property of networks of distinct anatomical areas, not a single area; and (3) these areas and networks carry out separable attentional functions.

Posner and Petersen focused on three putative classes of attentional processes: alerting, orienting, and detecting targets for conscious processing. Much research since then has expanded and refined this initial set of proposed classes.

Alerting

Alerting is thought to constitute the most basic attentional function. It appears to consist of distinct subfunctions or processes. Intrinsic alertness describes the ability to maintain certain levels of arousal in the absence of cues. It is thought to reflect general increases in excitability, mediated by top-down control signals. Changes in intrinsic alertness can be assessed over longer periods of time (minutes to hours) by measuring simple reaction times to perceptual stimuli that occur without warning.

Alerting subjects to an upcoming target decreases reaction time and error rate. Phasic alerting effects can be documented by comparing reaction time and neural activity on uncued trials and trials for which a nonspecific warning cue provided temporal information about the upcoming trial (**Figure 1(a)**). Phasic alerting may ready task-specific processing pathways for the next stimulus.

Alerting is thought to be supported by the widespread cortical distribution of the brain's norepinephrine system arising in the locus coeruleus (LC-NE) of the midbrain. The reticular thalamus may relay the effect of LC-NE activity to the cerebral cortex. Imaging studies of phasic alerting indicate that the thalamus strongly responds to alerting cues. LC-NE neurons are most active during wakefulness and become silent during rapid eye movement sleep.

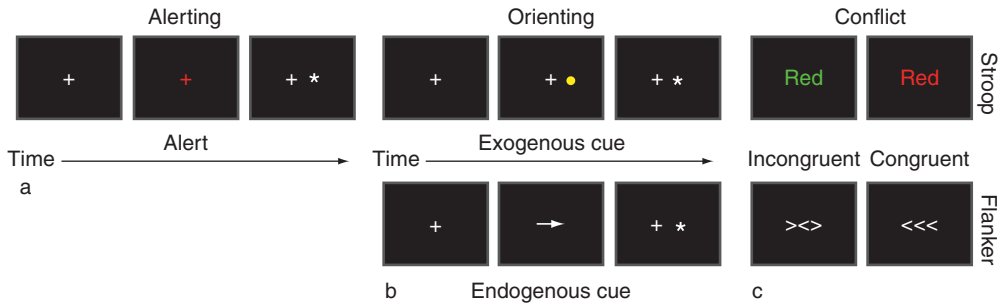

Figure 1 Examples of tasks commonly used to study attentional phenomena. (a) When asked to respond to a target (asterisk) by pressing a button, simple warning cues (e.g., color change of the fixation cross) that phasically alert subjects to an upcoming target can decrease reaction times and error rates. Such warning cues do not carry any spatial information about the target, only temporal information. (b) An exogenous orienting cue, such as a peripheral flash of light, will automatically capture selective attention and facilitate responses at the cued location, starting 50 ms after the cue. An endogenous orienting cue, such as a centrally placed arrow pointing to the right, will also facilitate responses at the cued location, but only after approximately 150 ms. Endogenous cueing effects are mediated by top-down mechanisms, whereas the effects of exogenous cues are thought to be entirely bottom-up or stimulus driven. (c) When the stimulus properties are incongruent, such that the task demands come in conflict with well-trained stimulus-driven responses, performance worsens. In the classic Stroop task, subjects are asked to report the ink color of words. When the word meaning (red) and ink color (green) are incongruent, the overtrained response of reading the word interferes with naming the ink color. In the Eriksen–Flanker task, subjects are asked to report the direction of a central arrow. Responses are faster when flanking arrows point in the same direction as the central arrow (congruent) than when they point in the opposite direction (incongruent). Since executive control processes are needed to overcome stimulus-driven responses, the Stroop, Eriksen–Flanker, and other conflict tasks have been used to study the brain's executive control networks. However, overcoming conflict from incongruent stimuli is only one of the many aspects of executive control.

Salient stimuli cause phasic LC-NE activations and norepinephrine release. LC-NE neurons project to parietal cortex, primary motor cortex, the pulvinar nucleus of the thalamus and the superior colliculus. Anterior cingulate cortex (ACC) and orbitofrontal cortex send strong projections to the LC. Therefore, it has been argued that LC activity might, in part, be controlled by the ACC.

Shifting Selective Attention: Cue Interpretation and Orienting

Attention can be selectively directed toward locations in space; intervals of time; features such as frequency, volume, color, and motion; semantic categories; abstract concepts; memories; and different output modalities. Although spatial attention has been the focus of many experimental studies, it may have its own specific mechanisms and represent a special case. It is important to keep in mind selective attention in its broader sense, which also includes selecting specific internal representations and response configurations.

In studies of spatial attention, the most commonly used analogy for focused selective attention is that of a spotlight. Although simple, this analogy captures the dynamics of selecting specific types of information initially proposed by Posner: disengaging attention from its current focus, moving attention, and engaging it at a new focus. Early PET and functional magnetic resonance imaging (fMRI) studies implicated regions in dorsal parietal and dorsal frontal cortex (DFC) under conditions of selective attention. Since then, lateral and medial parts of posterior parietal cortex (PPC), especially the intraparietal sulcus (IPS) and, more inferiorly, the temporoparietal junction (TPJ), have consistently been associated with shifts of selective attention. Regions in DFC, potentially constituting the human homologue of the frontal eye field, as well as parts of more ventral frontal cortex (VFC) are also widely believed to be important for selective attention.

A distinction can be made depending on whether the focus of attention is shifted involuntarily by a salient stimulus or voluntarily through top-down mechanisms. Stimulus-driven selective attention effects can be measured by flashing a cue on a screen (exogenous cue) and comparing how quickly subjects respond to targets at the cued location compared to a second location (Figure 1(b)). As early as 50 ms after the cue, reaction times are facilitated at the cued location. This is in contrast to situations in which cognitive or endogenous cues are used, such as a centrally placed arrow pointing toward one side of the visual field (Figure 1(b)) or the word 'right.' If the endogenous cue is valid and correctly predicts the location of the upcoming target, responses are also facilitated. However, these effects develop more slowly and are not measured until 150 ms after the cue.

Based largely on single-unit recording and event-related fMRI studies that can dissociate cue from target-related activity, Corbetta and Shulman proposed that stimulus-driven and goal-oriented shifts

of selective attention might be mediated by partially distinct brain networks.

In this model, the right TPJ and parts of the right VFC appear to mediate stimulus-driven shifts of selective attention. The TPJ is relatively unresponsive to endogenous cues that provide information about an upcoming target, such as its likely location or direction of motion. Yet, the right TPJ and VFC are strongly activated by unexpected targets that are thought to automatically capture selective attention. This response to unexpected targets appears to be independent of the task's input and output modalities. Corbetta and Shulman have conceptualized the stimulus-driven attention control system as a circuit breaker that can disrupt the current attentional focus and redirect selective attention in the absence of voluntary control. This apparent right lateralization of stimulus-driven attention control is consistent with the clinical finding that neglect is most commonly due to right hemisphere lesions.

In contrast, bilateral regions in the IPS and DFC appear to be important for the voluntary engagement of selective attention. The IPS and DFC consistently show greater activity when subjects have been endogenously cued toward a specific target property or location than when they have been cued to passively view a display. Furthermore, the bilateral IPS and DFC show anticipatory pretarget activity that is time-locked to the presentation of the endogenous cue. This cue-related activity is more extended in duration than cue-related activity in occipital cortex, which likely reflects purely visual processing of the cue. It has been demonstrated that the IPS and DFC are also active when visuospatial attention is shifted based on long-term memory in the absence of cues.

Goal-directed selective attention signals in dorsal parietal and frontal cortex are not limited to visual tasks. Studies have demonstrated dorsal frontoparietal activity during endogenous shifts of selective attention between different locations in auditory space and different auditory features. Regions in PPC are also activated by voluntary shifts of selective attention between audition and vision. Furthermore, dorsal parietal and frontal regions are thought to covertly orient goal-directed selective attention toward specific movements. Such selective motor attention leads to reaction time facilitation analogous to selective visual attention.

In addition, DFC is thought to select specific movements in anticipation of a response. Single-unit recordings in adjacent parts of macaque PPC, as well as frontal cortex, have documented effector specific anticipatory activity triggered by endogenous cues. Dorsal frontoparietal regions may even be important for selectively attending to specific sensorimotor transformations or stimulus-response mappings, especially when they are simple and well practiced. This is consistent with task-switching experiments that show PPC and dorsal frontal regions to be active when subjects switch between different tasks while the attentional focus remains constant for both input and output.

Executive Control

Executive control of selective attention is believed to be the third major function of attentional networks. Executive control is composed of distinct processing classes related to the instantiation, maintenance, monitoring, and adjustment of attentional sets. The clearest evidence that humans have some voluntary control over the selective allocation of cognitive resources comes from the fact that we can choose to perform many different operations on a given stimulus. Top-down control is critically important for the flexibility of behavior because it allows us to implement arbitrary criteria for input and output selection, as well as stimulus categorization. Without executive control, humans would be automatons limited to a finite number of preset stimulus-response mappings.

The term executive implies that this system is informed about the current task goals. The brain's executive control system likely transmits top-down biasing signals to downstream moment-to-moment information processors, such as visual and sensorimotor cortex. The executive system should have some knowledge of the organizational states of downstream processors. Thus, executive control systems also include a bottom-up component. Executive control regions receive ongoing performance feedback signals that can be used to adjust top-down signals for improved task performance. Behavioral observations have provided insight into top-down and bottom-up signals related to executive control and their interplay. Task-switching paradigms, for example, have shown that under certain conditions reaction times increase on postswitch trials. It is thought that this response slowing reflects the time it takes executive control to clear the previous attentional set and implement a different one.

Despite the importance of executive control systems for behavioral flexibility, their control over the selection of processing pathways is incomplete. When voluntary task goals come in conflict with strong prepotent stimulus-response mappings, behavioral performance worsens. Several well-studied paradigms, such as the Stroop and Eriksen–Flanker tasks, demonstrate the behavioral decrements caused by such conflict (Figure 1(c)). For these tasks, reaction times increase when the stimulus properties are such that the task demands are incongruent with a well-learned

prepotent response. In the Stroop task, for example, subjects are instructed to name the color of ink in which a word is written. Reaction times are slowest when the ink color and the meaning of the word are incongruent – for example, the word 'green' written in red ink. This response slowing is likely caused by competition between the voluntarily adopted task goals and the prepotent trained behavior of automatically reading the word.

Because top-down control over information processing is incomplete, executive systems are thought to receive feedback information about action outcomes so that top-down biasing signals can be adjusted as needed. The executive control system is thought to adjust the attentional set when performance, as indexed by slow reaction times or errors, is poor. Reaction time changes that are dependent on the nature of the previous trial are thought to reflect such trial-by-trial adjustments of control parameters. Often, reaction times will be systematically slower on the trial following an error, suggesting top-down adjustments of the attentional set. In the Stroop task, the average response time to an incongruent stimulus is faster if the preceding stimulus was incongruent than if it was congruent (**Figure 1(c)**). It has been proposed that adjustments of top-down biasing signals triggered by incongruent stimuli facilitate performance on subsequent trials.

Regions in dorsal anterior cingulate cortex/medial superior frontal cortex (dACC/msFC), dorsolateral prefrontal cortex (dlPFC), and anterior prefrontal cortex (aPFC) have been consistently associated with different executive control processes by a wide range of experimental approaches. The anterior insula/frontal operculum (aI/fO) has also been shown to play an important role in executive control. Although the brain's executive control network comprises all of these regions, each one of them likely carries out a slightly different executive control function.

The dACC/msFC is thought to play a central role in the exertion of executive control. It is believed to be essential for relating goal-oriented behavior to its outcomes. Evidence for the notion that the dACC/msFC carries out several important executive control processes comes from a variety of sources. Lesions of the dACC/msFC can lead to difficulties with the voluntary initiation and maintenance of complex behaviors, whereas more automatic stimulus-driven behavior is spared. In addition, dACC/msFC lesions can affect one's ability to correct task performance following an error.

Since executive control maintains the task goals, sustained neural activity is thought to be its hallmark feature. Maintenance activity has been measured in the dACC/msFC between a cue and the subsequent trial, as well as sustained across a whole block of trials. The dACC/msFC is activated by instructional cues that provide information about the task demands indicating that it is important for the instantiation of attentional sets. It also shows activity related to switching between tasks. Meta-analyses of neuroimaging studies have shown the dACC/msFC to be active for a wide variety of cognitive operations, which is consist with the idea that the dACC/msFC exerts executive control independent of the specific processing domain.

Besides signals thought to be related to the exertion of top-down control, a range of apparent feedback signals have also been measured in the dACC/msFC. Human neuroimaging studies have shown activation differences in the dACC/msFC related to errors, error likelihood, conflict, monetary loss and gain, pain, dread, social rejection, and expectancy violations. In addition, the dACC/msFC also carries domain-independent target detection signals.

Several experimental approaches besides human fMRI have shown that the dACC/msFC may help monitor behavior for errors. Event-related potentials (ERPs), for example, show a negative difference wave localized to the dACC/msFC when correct and error trials are compared. This effect, termed error-related negativity (ERN), is time-locked to the occurrence of the error. It has also been suggested that the dACC/msFC monitors conflict on a trial-by-trial basis. ERP studies have shown a difference wave when comparing congruent and incongruent stimuli localized to the dACC/msFC, labeled conflict-related-negativity (CRN; N450). Event-related fMRI studies have shown greater activity for high-conflict (incongruent) than low-conflict (congruent) trials in the dACC/msFC independent of input and processing domain.

It has been suggested that closely adjacent frontal midline structures may carry out different executive control functions. More posterior and dorsal msFC may implement attentional sets, whereas a slightly more anterior and ventral region in the dACC helps guide behavior by integrating actions and their outcomes. Imaging studies that compared self-guided actions to experimenter-guided ones showed greater activity for self-guided actions in the msFC. In contrast, the dACC did not show a preference for self-initiated movements. Instead, activity related to contingency-based learning and decision making has been documented in the dACC.

A region on the border of the anterior insula (aI) and frontal operculum (fO) has shown executive control properties very similar to the dACC/msFC. Imaging studies have shown the aI/fO to be coactivated with the dACC/msFC across a wide range of tasks.

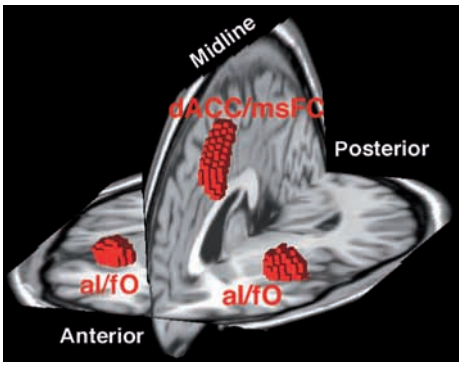

Figure 2 The dorsal anterior cingulate cortex/medial superior frontal cortex (dACC/msFC) has consistently shown activity related to attentional control. A cross-studies analysis of fMRI experiments documented attention-related signals in the dACC/msFC and bilateral anterior insula/frontal operculum (aI/fO). Across tasks, the dACC/msFC and aI/fO showed cue-related alerting/orienting as well as sustained task set-maintenance activity and error-related feedback signals important for executive control. Reproduced from Dosenbach NU, Visscher KM, Palmer ED, et al. (2006) A core system for the implementation of task sets. *Neuron* 50: 799–812, with permission from Elsevier.

Activation maps reveal that the bilateral aI/fO carries goal-maintenance and instantiation activity, as well as many of the same performance monitoring signals as the dACC/msFC (Figure 2). Similar to dACC/msFC lesions, strokes of the aI can lead to a reduction of self-initiated goal-directed activity.

It stands to reason that some brain regions may integrate executive control processes related to the instantiation, maintenance, monitoring, and adjustment of top-down biasing signals. Consistent with this notion, a cross-studies analysis showed that the dACC/msFC and aI/fO carried activity related to cue processing and the maintenance and monitoring/adjustment of attentional sets across a wide range of tasks. Therefore, it has been proposed that the bilateral aI/fO and dACC/msFC could form a domain-independent core of the human executive control network (Figure 2).

Regions in aPFC and dlPFC appear to support somewhat different executive control functions from the dACC/msFC and aI/fO. It is thought that the aPFC is especially important for the generation and maintenance of strategic plans, especially when they involve subgoaling and the integration of different types of information. It has also been proposed that aPFC neurons may adaptively code for different stimulus categorizations depending on the task goals. Humans with aPFC lesions often have difficulties solving complex tasks that require strategic planning. The maintenance of task contingencies, rules, and plans seems to be especially impaired by aPFC lesions. A series of fMRI experiments have documented sustained signals in the aPFC likely related to the maintenance of attentional sets and complex rules. However, this attentional set-maintenance activity was not common across all tasks. aPFC maintenance signals seem to be selectively recruited by more complex categorizations. Consistent with these findings, a meta-analysis of neuroimaging studies showed that the contrasts highlighting lateral aPFC had compared activity on trials with slow reaction times to activity on faster trials. In addition, neuroimaging studies have also shown activity related to switching between different task sets and stimulus categorizations in aPFC.

dlPFC may be specifically important for trial-by-trial adjustments of attentional control parameters. Consistent with this idea, an ERP study has shown that right dlPFC lesions increase errors and alter the ERN. Some dlPFC lesion patients fail to correct their errors and show no improvements during extended performance of complex cognitive tasks. Human fMRI studies have also shown greater dlPFC activity on error than correct trials. In contrast to the dACC/msFC, the dlPFC does not appear to carry conflict-related signals as measured by fMRI. According to the conflict-monitoring hypothesis, dlPFC functions as 'active memory in the service of control.' This hypothesis postulates that the dACC/msFC sends signals about needed adjustments in control parameters to the dlPFC, which helps to adjust attentional control parameters accordingly. In contrast, based on ERP data it has been argued that the dlPFC may be sending signals to the dACC/msFC, which in turn adjusts top-down biasing signals. Perhaps dlPFC maintains information about planned control adjustments from one trial to the next, whereas the dACC/msFC and aI/fO maintain the basic task parameters for the entire time period during which subjects are performing a specific task. The dlPFC may send adjustment-related signals directly to downstream processors, or it may transmit them to the dACC/msFC for downstream implementation.

Integration of Attentional Processes

The conceptual separation of attentional processes related to alerting, selective attention shifting (orienting), and executive control was initially based on differences in the behavioral effects of alerting, visuospatial cueing, and conflict. Although brain regions preferentially carry signals related to alerting, selective attention shifting, and executive control, these functions likely do not occur completely independently from each other. The brain's putative alerting, attention shifting, and executive control networks must communicate.

The executive control network needs to receive information about the meaning of sensory cues from both the alerting and the attention shifting networks so that it can implement the appropriate attentional settings. Conversely, the executive control network is thought to maintain the task goals according to which the attention-shifting network directs endogenous selective attention. It has even been shown that some stimulus-driven shifts in selective attention are contingent on the underlying attentional set. A stimulus that would not normally capture selective attention may do so if the subject is already searching for a feature shared by the stimulus. Voluntary changes in intrinsic alertness indicate that the executive control network can, at least indirectly, affect LC-NE activity.

Some data suggest that dACC/msFC and aI/fO may integrate several different executive control functions. It seems possible that dACC/msFC and aI/fO may also receive warning signals from the brain's alerting network, as well as information about the meaning of cues and attentional shifts from regions in PPC and frontal cortex. Perhaps specific nodes in each of the attentional networks form bridges between them that allow for an integration of attentional functions. Although distinct networks of brain regions support alerting, attention-shifting, and executive control functions, attention is likely a property that emerges from the interactions between these networks. One speculative idea is that the integration of attentional functions may be helped by the implementation of a basic goal-oriented task mode. This focused attention mode may stand in a mutually exclusive push–pull relationship with the brain's default mode.

See also: Attention and Eye Movements; Attention: Models; Attentional Functions in Learning and Memory; Attentional Mechanisms in Ventral Pathway; Attentional Networks in the Parietal Cortex; Decision-Making and Vision; Executive Function and Higher-Order Cognition: Definition and Neural Substrates; Frontal Lobe Syndrome; Neglect Syndrome and the Spatial Attention Network; Prefrontal Cortex: Structure and Anatomy; Psychophysics of Attention; Spatial Cognition and Executive Function; Visual Attention.

Further Reading

Aston-Jones G, Rajkowski J, and Cohen J (1999) Role of locus coeruleus in attention and behavioral flexibility. *Biological Psychiatry* 46: 1309–1320.

Braver TS, Reynolds JR, and Donaldson DI (2003) Neural mechanisms of transient and sustained cognitive control during task switching. *Neuron* 39: 713–726.

Bunge SA, Wallis JD, Parker A, et al. (2005) Neural circuitry underlying rule use in humans and nonhuman primates. *Journal of Neuroscience* 25: 10347–10350.

Corbetta M, Kincade MJ, Lewis C, et al. (2005) Neural basis and recovery of spatial attention deficits in spatial neglect. *Nature Neuroscience* 8: 1603–1610.

Corbetta M and Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* 3: 201–215.

Corbetta M, Shulman GL, Miezin FM, and Petersen SE (1995) Superior parietal cortex activation during spatial attention shifts and visual feature conjunction. *Science* 270: 802–805.

Dosenbach NU, Visscher KM, Palmer ED, et al. (2006) A core system for the implementation of task sets. *Neuron* 50: 799–812.

Duncan J and Owen AM (2000) Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences* 23: 475–483.

Eriksen CW and Hoffman JE (1973) The extent of processing of noise elements during selective encoding from visual displays. *Perception and Psychophysics* 14: 155–160.

Gehring WJ and Knight RT (2000) Prefrontal–cingulate interactions in action monitoring. *Nature Neuroscience* 3: 516–520.

James W (1890) *The Principles of Psychology.* New York: Holt.

Lau HC, Rogers RD, Haggard P, and Passingham RE (2004) Attention to intention. *Science* 303: 1208–1210.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24: 167–202.

Posner MI and Petersen SE (1990) The attention system of the human brain. *Annual Review of Neuroscience* 13: 25–42.

Posner MI, Snyder CRR, and Davidson BJ (1980) Attention and the detection of signals. *Journal of Experimental Psychology: General* 109: 160–174.

Raz A and Buhle J (2006) Typologies of attentional networks. *Nature Reviews Neuroscience* 7: 367–379.

Rushworth MF, Walton ME, Kennerley SW, and Bannerman DM (2004) Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Science* 8: 410–417.

Sakai K and Passingham RE (2003) Prefrontal interactions reflect future task operations. *Nature Neuroscience* 6: 75–81.

Stroop JR (1935) Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18: 643–662.

Weissman DH, Gopalakrishnan A, Hazlett CJ, and Woldorff MG (2005) Dorsal anterior cingulate cortex resolves conflict from distracting stimuli by boosting attention toward relevant events. *Cerebral Cortex* 15: 229–237.

Attentional Networks in the Parietal Cortex

G H Patel, B J He, and M Corbetta, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The brain is continuously flooded with information from different senses: vision, audition, touch, smell, and taste. As the inflow of sensory information is much greater than what the brain can process at any given time, one of the fundamental interests of neuroscience is to understand which brain mechanisms are responsible for the selection of those few bits of information that are relevant to the ongoing goals of an individual, and how irrelevant information is filtered out. Attended objects tend to be perceived and remembered much better than unattended objects are, and only objects that are attended become the target of motor plans, such as when we look at and reach for an apple. Accordingly, attention is defined as the ensemble of psychological and neural operations that mediate selection of sensory stimuli and link them to response and memory systems.

The parietal lobe, from the Latin *paries*, or 'walls of a house,' is the part of the brain that sits between the occipital, temporal, and frontal lobes, and that includes the lateral superior part of each hemisphere. The parietal lobe is situated between sensory (visual, tactile, and auditory) areas and contains cells that respond predominantly to behaviorally relevant stimuli. The parietal lobe is also active when individuals prepare to look or move toward a stimulus, and is heavily connected with the frontal lobe, in which the actions are planned. The parietal lobe of the brain is therefore well suited to perform operations that are neither strictly sensory nor motor, but rather operations that integrate sensory and motor information. Because of the integrative nature of the parietal lobe, the functions of the cortical areas contained within it are complex and diverse, and damage to the parietal lobe often results in multifaceted deficits.

In this article we first describe the anatomical organization of the parietal lobe; next, we turn to what we know about the functional organization of parietal cortex, and finally, we consider some of the behavioral deficits that arise after it is damaged.

Organization and Connectivity

The parietal lobe is bordered on the posterior and ventral sides by visual and auditory cortex, respectively, and the anterior portion of the parietal lobe is occupied by somatosensory cortex. Anterior to the parietal lobe is the frontal lobe, much of which is devoted to the planning and execution of movements. The parietal lobe is divided into smaller parts based on gross anatomy. The most prominent anatomical feature of the parietal lobe is the intraparietal sulcus (IPS), which runs anterior–posteriorly along the lateral aspect of the parietal lobe. The IPS divides the parietal lobe into two lobules: the superior parietal lobule (SPL), which encompasses the lateral aspect of the parietal lobe dorsal to the IPS as well as the medial wall, and the inferior parietal lobule (IPL), which encompasses all of the parietal lobe ventral to the IPS. The part of the parietal lobe on the medial wall is also often called the precuneus.

Using histological techniques, investigators have subdivided the parietal lobe in several ways. Perhaps one of the best known schemes for subdividing parietal cortex was proposed by Brodmann in 1909. In this scheme, the human parietal cortex is subdivided into seven areas: areas 1–3, which comprise the anterior edge of the parietal lobe and cover somatosensory cortex; area 5, which is immediately posterior to area 2 and covers the anterior part of the SPL; area 7, which is posterior to area 5 and covers much of the lateral and medial SPL and some of the IPL; and areas 39 and 40, which cover much of the IPL (see Figure 1(a)). Beyond these coarse divisions, however, not much was known for many years about the anatomical organization of parietal cortex, in part because invasive techniques necessary to trace connections and study the function of this part of the brain in humans were not available. As a result, most of the detailed information that we have on the parietal lobe's organization and on other parts of the human brain comes from anatomical and physiological studies of nonhuman primates, especially macaques, the brains of which share many of the same sensory and motor functions of the human brain.

The gross and histological organization of macaque parietal lobe is similar to that of the human parietal lobe in many ways: an IPS separates the parietal lobe into an SPL and IPL, somatosensory cortex makes up the anterior edge of the parietal lobe, and visual areas lie along the posterior edge. Areas 1, 2, 3, 5, and 7 are also present in macaque; however, the relative position of areas 5 and 7 is more ventral, given that areas 39 and 40, which occupy the ventral part of the human IPL, are not present in the macaque (see Figure 1(b)). Despite these potential differences, the macaque continues to serve as a useful model of how human parietal areas may be involved in attention.

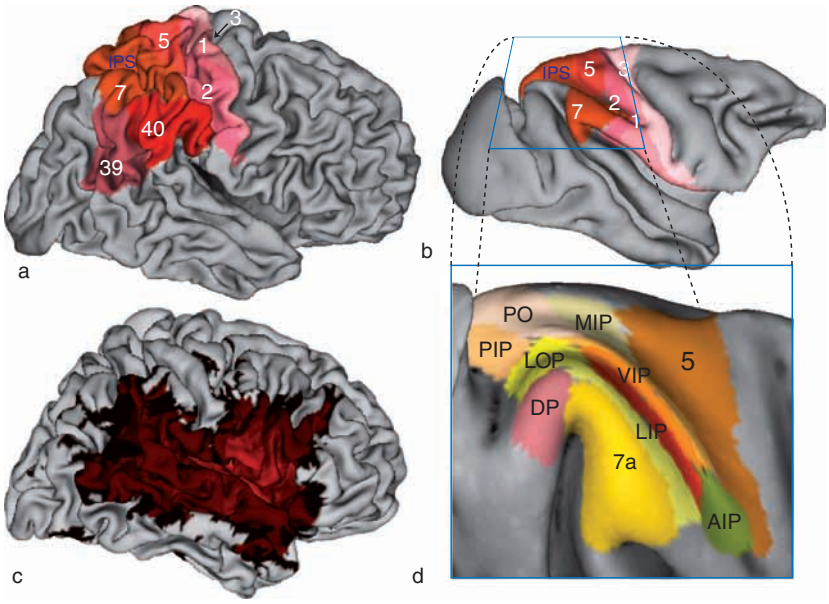

Figure 1 (a) Human Brodmann areas 1–3, 5, 7, 39, and 40, and the intraparietal sulcus (IPS). (b) Macaque Brodmann areas. (c) Cortical damage underlying neglect. (d) Areas in and around the macaque IPS (PO, parietal-occipital; MIP, medial intraparietal; PIP, posterior intraparietal; LOP, lateral occipitoparietal zone; VIP, ventral intraparietal; DP, dorsal prelunate; AIP, anterior intraparietal; 5, area 5; 7a, area 7a). Adapted from Lewis JW and Van Essen DC (2000) Mapping of architectonic subdivisions in the macaque monkey, with emphasis on parieto-occipital cortex. *Journal of Comparative Neurology* 428: 79–111.

In the 1970s, the advent of anatomical methods to trace connections between areas allowed for the partition of posterior parietal cortex into multiple areas based on their profile of feed-forward and feedback connections. Feed-forward connections refer to connections from lower to higher levels of a sensory hierarchy of cortical areas, whereas feedback connections refer to connections from higher to lower levels. There are several proposed schemes to divide the parietal cortex. Along the IPS, one common scheme divides the cortex into ten areas: dorsal prelunate (DP), posterior intraparietal (PIP), parietal-occipital (PO), the lateral occipitoparietal zone (LOP), medial intraparietal (MIP), ventral intraparietal (VIP; divided into medial and lateral), lateral intraparietal (LIP; divided into dorsal and ventral), anterior intraparietal (AIP), area 5, and area 7a (see Figure 1(d)). Other similar schemes have also been proposed. Tracer studies have shown that these areas receive input from other brain structures involved in the processing of vision, audition, and sensation, and may send output to the same sensory areas and/or to various motor and motor-planning areas. The IPS areas residing more posteriorly and laterally receive most of their input from visual areas and send output to oculomotor structures, whereas the more medial and anterior areas also receive input from somatosensory areas and send output to hand and arm motor areas.

For an area to be involved in the selective operations of visual spatial attention, it is likely to be connected to the visual areas responsible for the processing of incoming stimuli. It is also likely to be connected to areas involved in the planning and execution of saccades, since after the selection of a stimulus of interest, a saccade is often made to bring the stimulus into the fovea for further scrutiny. Area LIP in the macaque fits this description. It occupies the caudal half of the lateral bank of the IPS, and is demarcated histologically from surrounding parietal areas by increased myelination of layers 3–5. This area receives feed-forward connections from and sends feedback connections to many extrastriate visual areas, including areas V4, V3a, and middle temporal complex (MT), an area known to be involved in processing visual motion. LIP is also heavily interconnected with oculomotor structures, such as the superior colliculus in the midbrain, the pulvinar in the thalamus, and the frontal eye fields (FEFs) in prefrontal cortex.

Area 7a has a profile of connections with cortical areas similar to that of LIP, except rather than FEF, it is connected with area 46 in prefrontal cortex. Area 46 is an area known to be involved in spatial working memory, implying that 7a may also play a role in working memory and attention. Other posterior parietal areas, such as DP and PO, connect mainly with extrastriate areas and more anterior IPS areas, such as LIP, indicating that they play a role in relaying information from visual cortex to the higher level planning areas in IPS.

In general, the many anatomical studies of parietal cortex over the past several decades have shown that it appears to be divided into many areas, each of which appears to play a role in translating sensory information into motor plans. But what exactly this role entailed remained a mystery until the technology to study these areas *in vivo* was developed, first in macaques and then later in humans.

Function

In 1975 Mountcastle and colleagues first reported on electrophysiological recordings of macaque parietal cortex neurons while the monkey performed various tasks. They found that neurons in the IPS "appear[ed] to direct visual attention to objects of interest and motivational power, and to issue commands for maintaining directed fixation of the object when it [was] stationary, and to track it if it mov[ed]." In the decades since these early experiments, electrophysiological recordings of parietal cortex neural activity have determined that the SPL and IPL are subdivided into many smaller functional areas that appear to roughly correspond with the subdivisions determined by histological techniques.

Each of these areas appears to extract the spatial information from one or more sensory inputs, and then play some role in transforming that information into a general movement plan in the same spatial coordinates. Because of its involvement in spatial processing, the parietal cortex is said to be part of the 'where' pathway (as opposed to the ventral 'what' pathway, which appears to be involved in processing the identity of a stimulus regardless of spatial location). For instance, the area LIP is involved in the planning of upcoming saccadic eye movements. The neurons in this area respond transiently to visual stimuli presented within a specific part of the visual field, which is known as a receptive field. Moreover, if a saccade is to be performed to that stimulus at some point in the future, the neurons continue to fire until the saccade has been executed. In this way, they are said to represent the planned target of the saccade until the action is completed, even if the visual stimulus is no longer in the receptive field. This characteristic is often used to define the boundaries of LIP in electrophysiological recording experiments. LIP neurons have large visual receptive fields that cover on average a quarter of the visual field, and these receptive fields move with the position of gaze (gaze centered). They are also more likely to respond to stimuli in the contralateral visual hemifield than in the ipsilateral, and may contain a coarse but continuous map of the contralateral visual field (otherwise known as being retinotopically organized). LIP neurons, then, appear to represent the plan of an upcoming saccadic eye movement in a rough spatial map of the visual environment.

LIP neurons will also respond to a visual stimulus that is not the target of a planned saccade, but is otherwise task relevant. For instance, if the monkey is instructed to maintain fixation on a central point, and to indicate with his hand when a visual stimulus in the periphery dims just slightly, the LIP neurons representing the location of this stimulus continue to fire until the task is completed, and this increased level of firing is directly correlated to increased performance of whatever perceptual decision needs to be made at that location. Moreover, chemically deactivating LIP will result in the reduction of the monkey's ability to discriminate an oddball stimulus from other stimuli (such as a red circle among a field of green circles). It appears, then, that LIP neurons also represent the current locus of visuospatial attention.

Are these two purported functions of LIP neurons in conflict, or do they represent two ways of saying the same thing? The answer to this has been the source of a contentious debate. Thus far we have seen that LIP is involved both in covert shifts of attention (marking a location of interest so that processing of objects in that location can be enhanced) and in overt shifts of attention (making a saccade to a location so that it can be studied in more detail). One side of the debate holds that all covert shifts of attention represent potential saccades, and that the enhanced processing at that location is merely a side effect of planning a movement to that location. As evidence for this view, they point to other parietal areas, which appear to play the same planning role for other effectors, such as arm and hand movements. The 'intentional map,' then, might be a general principle of what the parietal cortex contributes to a sensorimotor transformation. The other side, however, contends that the LIP is a 'salience map' of the visual world, marking locations of interest for any number of systems to use, including the visual attention and oculomotor system. Part of the evidence for this view is that LIP neurons track the location of salient objects independently of stored oculomotor plans. While the differences between these two sides may seem minute, the resolution of this debate will give us insight into both what specific calculations parietal cortex neurons are performing on incoming information and in what terms the brain represents the external world – as a distorted version of reality, skewed toward the most interesting and relevant stimuli, or always in terms of a motor plan for interacting with the world?

Other parietal areas in the macaque may also play roles in selective attention, though these areas have

not been as thoroughly investigated as LIP. Areas caudal to LIP, such as DP, appear to play a role in visuospatial processing, though probably at an earlier stage than LIP (again fitting with their profiles of connections). Area 7a, which covers much of the IPL lateral to LIP, also appears to be involved in visuospatial processing. Like LIP, its neurons have large receptive fields, though it appears to evenly represent both the contralateral and ipsilateral hemifields. The neurons in this area appear to respond to the novel appearance of behaviorally relevant visual stimuli, but not as many of the 7a neurons have presaccadic activity as compared to LIP cells. This may indicate that rather than representing the current focus of attention, 7a neurons may be involved in detecting novel but potentially relevant stimuli, a counterpart of sorts to LIP. How other macaque parietal areas are involved in spatial attention is less clear, as they have generally been studied under the rubric of motor planning. Another important organizing principle for thinking about parietal cortex is to consider how space is coded in this region of the brain. There is evidence that parietal cortex subregions may be specialized for coding spatial location away from the body (or extrapersonal space) or near the body (or peripersonal space).

This wealth of information about monkey's parietal cortex until recently did not have a counterpart in the human work, because studies were limited to clinical observations (see later). This state of affairs changed in the late 1980s with the advent of positron emission tomography (PET), and then again in the 1990s with the advent of functional magnetic resonance imaging (fMRI). These neuroimaging technologies allowed, for the first time, *in vivo* studies of neural activity with sufficient spatial and temporal resolution to discern areas of the brain that were active in different tasks. Early PET studies confirmed that in humans the parietal lobe is also part of the dorsal 'where' pathway and is involved in the control of both spatial attention and eye movements. Several subsequent fMRI studies have also shown that, like in the macaque, the parietal region activated during covert shifts of attention largely overlaps with the region activated during saccades. This region includes much of the cortex in and around the IPS. Parts of this region are also activated during nonspatial shifts of attention, such as changing the focus of attention from the direction of moving dots to the color of the moving dots. Because of this profile of activity, the region along the human IPS has been thought of as generally homologous to the macaque IPS.

However, due to the combination of relatively poor spatial resolution of fMRI and the high degree of variability in the location of sulci and gyri among human individuals, it has been difficult to subdivide this region into functional areas as has been done in the macaque. Two foci of neural activity are consistently activated in different tasks requiring shifting or maintenance of attention. The first is in the SPL, on the dorsal–medial bank of the IPS. Because it is thought to be homologous to the macaque LIP, it is sometimes called human LIP (hLIP). Like macaque LIP, hLIP appears to be gaze centered in its spatial reference frame, responds more strongly to stimuli in the contralateral visual hemifield than in the ipsilateral, and appears to be loosely retinotopically organized. This area is activated if there is an attention shift to a peripheral location, and activity is sustained in this area if either attention remains focused on, or a saccade is being planned to, the peripheral location. Activity in this area seems to track both the locus of visuospatial attention and the target of an upcoming eye movement, extending the debate about parietal function to the human parietal cortex. While much of the functional profile of hLIP appears to be broadly similar to that of macaque LIP, the lack of direct comparisons of histological data or neural activity during tasks requiring shifts of attention prevents a more thorough assessment of the homology of these areas.

Areas posterior to hLIP are also often activated in tasks requiring shifts of spatial attention. One of the most prominent is in the fundus of the caudal section of the IPS, and is variously termed ventral IPS (vIPS) or V7. This area is preferentially activated by attending to the contralateral hemifield, but appears to be more visual in nature than hLIP. V7 most likely represents an earlier stage of the 'where' pathway that probably relays visual information to hLIP, and is a potential homolog to one of the posterior parietal areas in the macaque.

While the aforementioned areas in the human SPL are activated during voluntary and involuntary shifts of attention, the IPL appears to play a different role in attention. Regions in the IPL are only activated in response to the appearance of a salient, behaviorally relevant object, and even more so if attention has to be reoriented from another location (known as stimulus-driven reorienting of attention). Moreover, while activity in the SPL areas is increased during sustained covert shifts of attention, activity in the IPL is suppressed. It has been proposed that these areas play a role in stimulus-driven attention, especially when the focus of attention is captured by a salient and relevant novel stimulus. In macaques, area 7a (on the macaque IPL) is one possible candidate for a homolog to the human IPL.

Another important feature of human parietal cortex is that it is anatomically and functionally asymmetrical. Several studies have now shown that

functional regions of the supramarginal and angular gyrus at the temporoparietal junction (TPJ) in humans show hemispheric asymmetries of attention (right-dominant) or verbal memory (left-dominant) tasks. This asymmetry is even more dramatic when one considers the behavioral effects of stroke in the middle cerebral artery (MCA) distribution that feeds the parietal cortex. Whereas left MCA strokes commonly result in aphasia (language deficits), strokes of the right MCA commonly result in a syndrome termed 'neglect,' which is a collection of spatial attention–perception–premotor deficits (see Figure 1(c)).

An important goal for the future will be to reconcile monkey and human data in a common evolutionary framework that takes into account both commonalities and differences in functional organization. An important technological advance that will help this line of research is the development of fMRI studies in awake-behaving monkeys.

Lesions

Damage to the right IPL often results in spatial neglect, a common syndrome following stroke in which patients are biased toward attending to the right side of visual space more than to the left side. Neglect can leave patients unable to perform simple tasks requiring spatial attention, and occurs in about 30% of all stroke-affected individuals.

Patients with neglect often act as if part of the left side of their world does not exist. This inattention to the left side can occur in extrapersonal space or about the patient's own body, and can occur in different reference frames: gaze-centered frame (left with respect to the center of gaze), body-centered frame (left with respect to the midline of the body), or world-centered frame (left with respect to the environment). Less commonly, patients can manifest so-called object-centered neglect, where the left side of an object is ignored, no matter if it is presented in the left or right side of the visual space. For instance, if a page of text with three columns is presented, a patient with body-centered or world-centered neglect will miss the words on the left side of the page, whereas a patient with object-centered neglect will miss the left side of each column or the left side of each word.

Functionally, neglect patients may 'forget' to shave, groom, or dress the left side of body. They also have major problems with driving a car. Severely affected patients might also display a tonic rotation of the body or eyes toward the right, due to a coexisting motor imbalance. Some of these patients may still be able to detect a stimulus presented alone in the left hemifield, but when presented with stimuli simultaneously to both visual fields, patients tend to see only those in the right hemispace. This deficit is referred to as 'extinction.' Extinction is often tested by snapping fingers in one or both of the patient's visual fields and asking the patient to point to which hand they saw or heard snapping. When fingers are snapped simultaneously in both visual fields, neglect patients will inexorably point to the hand in their right visual field. This test is one of the many routine exams in clinical settings to assess the severity of neglect. Other commonly used neuropsychological tests for diagnosis of neglect include line bisection, in which patients with neglect bisect a horizontal line to the right of the true center; a series of cancellation tasks, in which neglect patients commonly miss marking targets on the left side of a paper; clock drawing, in which neglect patients may draw only the right half of a clock; and a baking-tray task, in which neglect patients usually cluster cookies on the right side of a baking tray.

In addition to the lack of awareness of the left side of space, neglect patients can also manifest deficits in planning hand or eye movements toward the left side of space. In this type of neglect, patients may not demonstrate a bias in awareness to either visual field, but will be less inclined to reach or look for objects located in the left visual field. This form of neglect is called motor or premotor neglect, or directional hypokinesia, to emphasize its relationship to action. In addition to lateralized (i.e., left worse than right space) deficits, a number of so-called nonlateralized deficits (i.e., similar across left and right visual fields) have also been described, including impairment of spatial working memory, sustained attention, and an overall lower level of alertness.

An important focus of current research is to assess the functional effect of different deficits, their importance for final outcome, their recovery over time, and their localization in the brain. Shall we think of neglect as a homogeneous or heterogeneous syndrome? What are the regions more responsible for one or another deficit? We know that spatial neglect can occur as a clinical syndrome for lesions in many parts of the brain, including parietal, temporal, frontal cortex, basal ganglia, thalamus, and in various white-matter tracts. It is still unknown how these different lesions may or may not cause different behavioral profiles of neglect, and how the effect of lesions in the brain relates to the functional organization observed in healthy individuals.

An alternative view considers neglect not to be the effect of damage to specific cortical areas, but rather the result of a distributed and combined anatomical–functional dysfunction of large parts of parietal and frontal cortices and their connections. In general, the areas that are damaged in neglect are located ventrally in the brain, including the IPL regions

specialized for stimulus-driven attention. Conversely, damage to more dorsal IPS and SPL regions does not produce strong neglect; rather, these lesions produce primarily eye- or arm-movement planning problems, which is also consistent with the results of functional imaging studies in healthy volunteers. However, only a few patients with damage restricted to the SPL have been studied.

One model of neglect postulates that damage by stroke or trauma to the IPL in the region of the TPJ could give rise to both spatial and nonspatial deficits by interrupting input to the SPL attentional areas from the TPJ. This disruption would then reduce the brain's ability to detect important sensory events in both visual fields. In addition, this lack of input into ipsilesional dorsal areas may induce a relative interhemispheric imbalance in the dorsal attentional areas that code for spatial locations and eye and arm movements. The imbalance due to the competitive nature of interhemispheric processing will lead to a relative hyperorienting toward right space, with consequent problems of attention and detection in left space. This model of spatial bias based on interhemispheric competition is also supported by a number of animal studies in which neglect has been cured by inactivating homologous parietal regions of the intact hemisphere.

Conclusion

Parietal cortex is subdivided into different areas, based on both anatomical and functional criteria; these areas have connections with both sensory processing and motor planning regions of the brain. Some of these areas appear to be involved in the selection and representation of locations and objects of interest in the immediate environment. The spatial information in parietal cortex can then be 'read out' by other brain areas, both for the planning of movements and for the enhancement of sensory processing. Attention can be directed to a location or object by a premeditated plan, or by the appearance of a novel stimulus that may require additional processing. Area LIP in the macaque and its putative homolog in the SPL of humans appear to be involved in representing the locations of interest, whereas IPL areas appear to play the role of reorienting attention whenever a novel stimulus enters awareness. Damage to these areas results in the syndrome of neglect, which fundamentally involves a lack of awareness for spatial information, as well as problems with vigilance and motor planning.

See also: Attention and Eye Movements; Attention: Models; Attentional Functions in Learning and Memory; Attentional Mechanisms in Ventral Pathway; Attentional Networks; Decision-Making and Vision; Neglect Syndrome and the Spatial Attention Network; Parietal Cortex and Spatial Attention; Psychophysics of Attention; Visual Attention.

Further Reading

Andersen RA and Buneo CA (2002) Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience* 25: 189–220.

Colby CL and Goldberg ME (1999) Space and attention in parietal cortex. *Annual Review of Neuroscience* 22: 319–349.

Corbetta M and Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* 3: 201–215.

Desimone R and Duncan J (1995) Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18: 193–222.

Egeth HE and Yantis S (1997) Visual attention: Control, representation, and time course. *Annual Review of Psychology* 48: 269–297.

Goldman-Rakic PS (1988) Topography of cognition: Parallel distributed networks in primate association cortex. *Annual Review of Neuroscience* 11: 137–156.

Hillis AE (2006) Neurobiology of unilateral spatial neglect. *Neuroscientist* 12: 153–163.

Husain M and Rorden C (2003) Non-spatially lateralized mechanisms in hemispatial neglect. *Nature Reviews Neuroscience* 4: 26–36.

Kastner S and Ungerleider LG (2000) Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience* 23: 315–341.

Lewis JW and Van Essen DC (2000) Mapping of architectonic subdivisions in the macaque monkey, with emphasis on parieto-occipetal complex. *Journal of Comparative Neurology* 428: 79–111. Avaliable at URL: http://sumsdb.wustl.edu:8081/sums/directory.do?id=679531.

Mesulam MM (1999) Spatial attention and neglect: Parietal, frontal and cingulate contributions to the mental representation and attentional targeting of salient extrapersonal events. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 354: 1325–1346.

Mountcastle VB, Lynch JC, Georgopoulos A, et al. (1975) Posterior parietal association cortex of the monkey: Command function for operations within extrapersonal space. *Journal of Neurophysiology* 38: 871–908.

Orban GA, VanEssen D, and Vanduffel W (2004) Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Science* 8: 315–324.

Pashler HE (1998) *The Psychology of Attention*. Cambridge, MA: MIT Press.

Posner MI and Petersen SE (1990) The attention system of the human brain. *Annual Review of Neuroscience* 13: 25–42.

Attractor Network Models

X-J Wang, Yale University School of Medicine, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term attractor is being increasingly used by neurophysiologists to characterize stable, stereotyped spatiotemporal neural circuit dynamics. Examples include readily identifiable rhythmic activity in a central pattern generator; well-organized propagation patterns of neuronal spike firing in cortical circuits *in vivo* or *in vitro*; self-sustained persistent activity during working memory; and neuronal ensemble representation of associative long-term memory. In these examples, the rich and complex neural activity patterns are generated largely through regenerative mechanism(s), and emerge as collective phenomena in recurrent networks. In part, interest in attractor networks arises from our growing appreciation that neural circuits are typically endowed with an abundance of feedback loops and that the attractor theory may provide a conceptual framework and technical tools for understanding such strongly recurrent networks.

The concept of attractors originates from the mathematics of dynamical systems. Given a fixed input, a system consisting of interacting units (e.g., neurons) typically evolves over time toward a stable state. Such a state is called an attractor because a small transient perturbation alters the system only momentarily; afterward, the system converges back to the same state. An example is illustrated in Figure 1, in which a neural network is described by a computational energy function in the space of neural activity patterns and the time evolution of the system corresponds to a movement down hill, in the direction of decreasing the computational energy. Each of the minima of the energy function is thus a stable (attractor) state of the system; a maximum at the top of a valley is an unstable state. Such a depiction is not merely schematic, but can be rendered quantitative for certain neural models.

The characterization stable and stereotyped is sometimes taken to imply that an attractor network is not sensitive to external stimuli and is difficult to reconfigure. On the contrary, as was shown by recent studies, attractor networks are not only responsive to inputs, but may in fact to instrumental to the slow time integration of sensory information in the brain. Moreover, attractors can be created or destroyed by (sustained) inputs; hence, the same network can serve different functions (such as working memory and decision making), depending on the inputs and cognitive control signals. The attractor landscape of a neural circuit is readily modifiable by changes in cellular and synaptic properties, which form the basis of the attractor model for associative learning.

In this article, we first introduce the basic concepts of dynamical systems, attractors, and bistability using simple single-neuron models. Then, we discuss attractor network models for associative long-term memory, working memory, and decision making. Modeling work and experimental evidence are reviewed and open questions are outlined. We show that attractor networks are capable of time integration and memory storage over timescales much longer than the biophysical time constants of fast electrical signals in neurons and synapses. Therefore, strongly recurrent attractor networks are especially relevant to memory and higher cognitive functions.

The Neuron Is a Dynamical System

A passive nerve membrane is a dynamical system described by a simple resistance–capacitance (RC) circuit equation:

$$C_m(\mathrm{d}V_m/\mathrm{d}t) = -g_L(V_m - E_L) + I_{app} \qquad [1]$$

where V_m is the transmembrane voltage, C_m the capacitance, g_L the leak conductance (the inverse of the input resistance), E_L the leak reversal potential, and I_{app} the injected current. In the absence of an input, the membrane is at the resting state, $V_{ss} = E_L$, say -70 mV. This steady state is stable; if V_m is transiently depolarized or hyperpolarized by a current pulse, after the input offset it will evolve back to V_{ss} exponentially with a time constant $\tau_m = C_m/g_L$ (typically 10–20 ms). Thus, V_{ss} is the simplest example of an attractor. More generally, for any sustained input drive I_{app}, the membrane always has a steady-state $V_{ss} = V_L + I_{app}/g_L$, given by $\mathrm{d}V_m/\mathrm{d}t = 0$ (i.e., V_{ss} does not change over time). The behavior of this passive membrane changes quantitatively with input current, but remains qualitatively the same; the dynamics is always an exponential time course (determined by τ_m) toward the steady-state V_{ss}, regardless of how high is the current intensity I_{app} or how large is the capacitance C_m or the leak conductance g_L. Moreover, the response to a combination of two stimuli I_1 and I_2 is predicted by a linear sum of the individual responses to I_1 or I_2 presented alone. These characteristics are generally true for a linear dynamical system, such as a differential equation with only linear dependence on V_m.

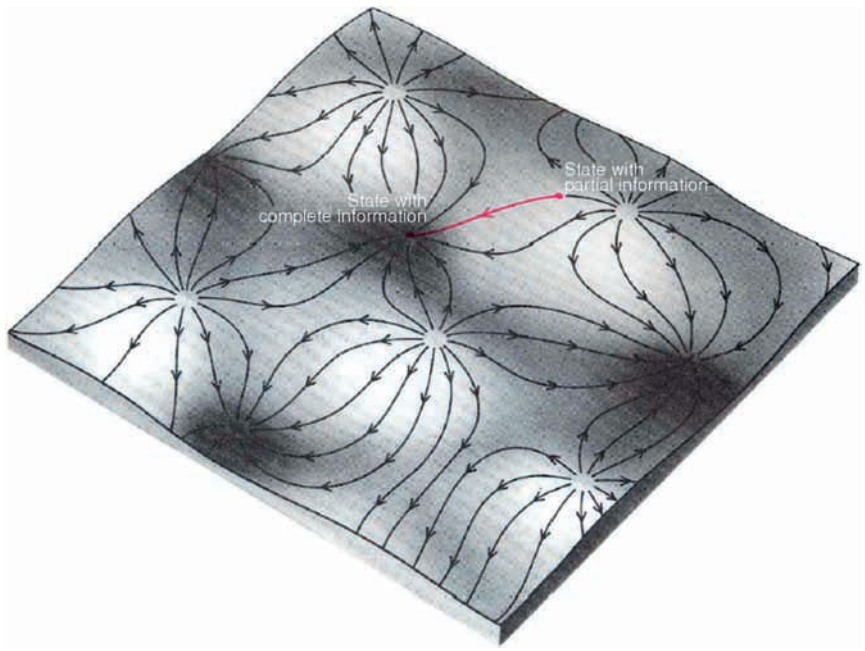

Figure 1 Schematic of an attractor model of neural networks. Computational energy function is depicted as a landscape of hills and valleys plotted against the neural activity states (on the *XY* plane). The synaptic connections and other properties of the circuit, as well as external inputs, determine its contours. The circuit computes by following a path that decreases the computational energy until the path reaches the bottom of a valley, which represents a stable state of the system (an attractor). In an associative memory circuit, the valleys correspond to memories that are stored as associated sets of information (the neural activities). If the circuit is cued to start out with approximate or incomplete information, it follows a path downhill to the nearest valley (red), which contains the complete information. From Tank DW and Hopfield JJ (1987) Collective computation in neuronlike circuits. *Scientific American* 257: 104–114.

More interesting behaviors become possible when nonlinearity is introduced by the inclusion of voltage-gated ionic currents. For instance, if we add in the RC circuit a noninactivating sodium current $I_{\mathrm{NaP}} = g_{\mathrm{NaP}} m_{\mathrm{NaP}}(V_{\mathrm{m}})(V_{\mathrm{m}} - E_{\mathrm{Na}})$, where the conductance exhibits a nonlinear (sigmoid) dependence on V_{m}, the membrane dynamics becomes

$$\begin{aligned} C_{\mathrm{m}}(\mathrm{d}V_{\mathrm{m}}/\mathrm{d}t) = &- g_{\mathrm{L}}(V_{\mathrm{m}} - E_{\mathrm{L}}) \\ &- g_{\mathrm{NaP}} m_{\mathrm{NaP}}(V_{\mathrm{m}})(V_{\mathrm{m}} - E_{\mathrm{Na}}) \\ &+ I_{\mathrm{app}} \end{aligned} \qquad [2]$$

This system is endowed with a self-excitatory mechanism: a higher V_{m} leads to more I_{NaP}, which in turn produces a larger depolarization. If g_{NaP} is small, the weak positive feedback affects the membrane dynamics only slightly (Figure 2(b), red lines). With a sufficiently large g_{NaP}, the steady state at $V_{\mathrm{Down}} \simeq -70\,\mathrm{mV}$ is still stable because at this voltage I_{NaP} is not activated. However, the strong positive feedback gives rise to a second, depolarized plateau potential (at $V_{\mathrm{Up}} \simeq -20\,\mathrm{mV}$) (Figure 2(b), blue lines). Therefore, the membrane is bistable; a brief input can switch the system from one attractor state to another (Figure 2(a)). As a result, a transient stimulus can now be remembered for a long time, in spite of the fact that the system has only a short biophysical time constant (20 ms). Unlike linear systems, in a nonlinear dynamical system gradual changes in a parameter (g_{NaP}) can give rise to an entirely new behavior (bistability).

Attractor states are stable under small perturbations, and switching between the two can be induced only with sufficiently strong inputs. How strong is strong enough? The answer can be found by plotting the total ion current $I_{\mathrm{tot}} = I_{\mathrm{L}} + I_{\mathrm{NaP}} - I_{\mathrm{app}}$ against V_{m}, called the *I–V* curve (Figure 2(b), top, with $I_{\mathrm{app}} = 0$). Obviously, a V_{m} is a steady state if $I_{\mathrm{tot}}(V_{\mathrm{m}}) = 0$ (thus $\mathrm{d}V_{\mathrm{m}}/\mathrm{d}t = 0$). As seen in Figure 2(b) (blue lines), the two attractors (filled circles) are separated by a third steady state (open circle). The third state is unstable – if V_{m} deviates slightly from it, the system does not return but converges to one of the two attractors. Indeed, if V_{m} is slightly smaller, I_{tot} is positive (hyperpolarizing), so V_{m} decreases toward V_{Down}. Conversely, if V_{m} is slightly larger, I_{tot} is negative (depolarizing), so V_{m} increases toward V_{Up}. Therefore, an external input must be strong enough to bring the membrane potential beyond the unstable steady state to switch the system from one attractor state to the other.

It is worth noting that the attractor landscape depends not only on the strength of the feedback mechanism (the value of g_{NaP}) but also sustained inputs. As shown in Figure 2(c), with a fixed g_{NaP} a constant applied current I_{app} shifts the *I–V* curve up

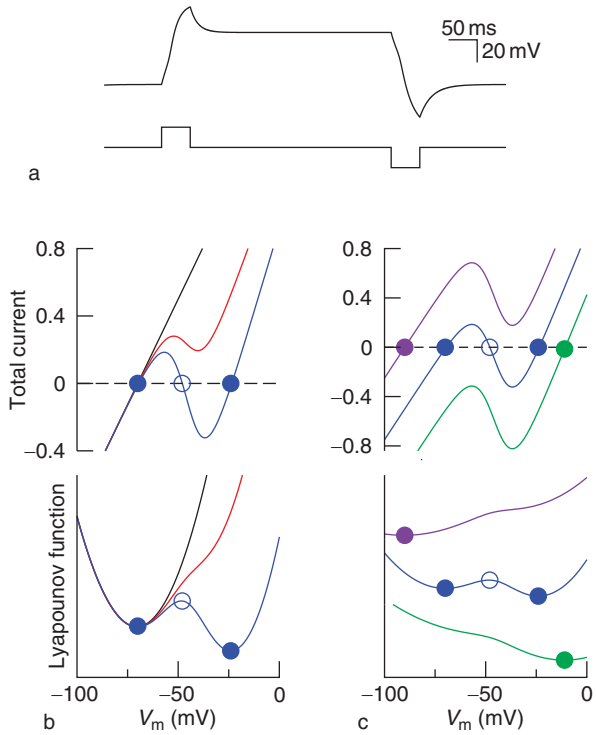

Figure 2 Positive feedback and attractor dynamics in a simple neural membrane model: (a) bistability with $g_{NaP} = 0.015\,\mu S$; (b) I–V curve (top) and computational energy function (Lyapounov function) (bottom); (c) bistability modulated by inputs, with $g_{NaP} = 0.015$ fixed, I–V curve (top) and computational energy function (Lyapounov function) (bottom). The membrane voltage is described by an RC circuit with the addition of a fast noninactivating sodium current $I_{NaP} = g_{NaP} m_{NaP}(V_m)(V_m - E_{Na})$, where $m_{Nap}(V_m) = 1/(1 + \exp[-(V_m + 45)/5])$ is a sigmoid function of V_m. The interplay between I_{NaP} and membrane depolarization produces an excitatory regenerative process. In (a), the system is initially at rest ($V_{Down} \simeq -70$ mV). A depolarizing current pulse switches the membrane to a plateau potential ($V_{Up} \simeq -20$ mV), which persists after the input offset. A second, hyperpolarizing current pulse switches the membrane back to the resting state. In (b), I–V curve (upper) is the total current $I_{tot} = I_L + I_{NaP} - I_{app}$ as a function of V_m, with $I_{app} = 0$); a steady state is given by an intersection with $I_{tot} = 0$. In the computational energy function (lower) $U(V_m)$, a steady state corresponds to a maximum (unstable) or a minimum (stable). For $g_{NaP} = 0$ (black, passive membrane) or $g_{NaP} = 0.08$ (red), there is only one steady state ($\simeq -70$ mV). For $g_{NaP} = 0.015$ (blue), there are three steady states; two are stable (filled circles) and the third is unstable (open circle). In (c), the injected current intensity is varied (blue, violet, and green for $I_{app} = 0$, -0.5, and 0.5 nA, respectively). Other parameter values are $C_m = 0.5$ nF, $g_L = 0.025\,\mu S$, $V_L = -70$ mV, and $V_{Na} = 55$ mV. The energy function $U(V_m)$ is defined by rewriting the circuit Eqn. (2) as $dV_m/dt = F(V_m) = -dU/dV_m$; hence, the energy function $U(V_m)$ is the integral of $-F(V_m)$. For instance, with $g_{NaP} = 0$, Eqn. (2) is reduced to Eqn. (1) and can be rewritten as $dV_m/dt = (V_{ss} - V_m)/\tau_m = F(V_m)$. By integrating $-F(V_m)$, we have $U(V_m) = (V_{ss} - V_m)^2/(2\tau_m) + U_0$, with an arbitrary constant U_0. Therefore, $U(V_m)$ is a parabola with V_{ss} at the bottom of the valley of the energy function (**Figure 1(b)**, black). For $g_{NaP} = 0.015$ (**Figure 1(b)**, blue), the energy function $U(V_m)$ displays two valleys, at V_{Down} and V_{Up}, separated by a peak at the unstable steady state. RC, resistance-capacitance; V_m, membrane voltage.

or down. Either a hyperpolarization or depolarization can destroy the bistability phenomenon. This simple example demonstrates that neuronal bistable dynamics can be readily reconfigured by external inputs. This is generally true for neural networks as well and has important computational implications. Attractors need not be steady states. In neurons, a plateau potential is typically not stable as a steady state. Instead, on depolarization the Hodgkin–Huxley-type sodium and potassium currents produce repetitive action potentials, which represent another (oscillatory) type of attractor behavior. The attractor nature of periodic spiking is shown in **Figure 3** using the classic Hodgkin–Huxley model; regardless of the initial states, the membrane system always converges to the same periodic attractor state. If the system is perturbed by a transient stimulus, it resumes the same firing pattern after the stimulus offset, except for a shift of spiking time or the phase of the periodic attractor state. Thus, a periodic attractor is robust (the amplitude and periodicity are not sensitive to transient perturbations). At the same time it is sensitive to phase shift; hence, the clock can be readily reset. This general property of nonlinear oscillators is a key to understanding synchronization among neurons and coherent brain rhythms. It is also a cornerstone of the neurobiology of biological clocks, such as the circadian rhythm, sleep cycle, and central pattern generators for locomotion such as walking, swimming, and breathing. In each of these systems, the rhythm is amenable to being reset by a transient input which leads to a phase shift in time but otherwise does not alter the stereotypical activity pattern in the network.

The oscillatory attractor illustrates that the behavior of a dynamical system (the Hodgkin–Huxley model) is determined by direct observables (the membrane voltage) as well as internal dynamical variables (ion channel-gating variables). The space of all dynamical variables form a phase space of the system, which is typically multidimensional. A neural circuit in the mammalian brain consists of many thousands of cells; its phase space is enormous. The dynamics of such systems can be very complex, exhibiting a wide gamut of spatiotemporal activity patterns. It is generally not possible to define an energy function for such systems. Nevertheless, the concept of computational energy landscape is still helpful for developing intuitions about attractor networks.

Synaptic Plasticity and Associative Memory

Bistable switches or, more generally, multiple attractors can be realized on all scales, from the molecular

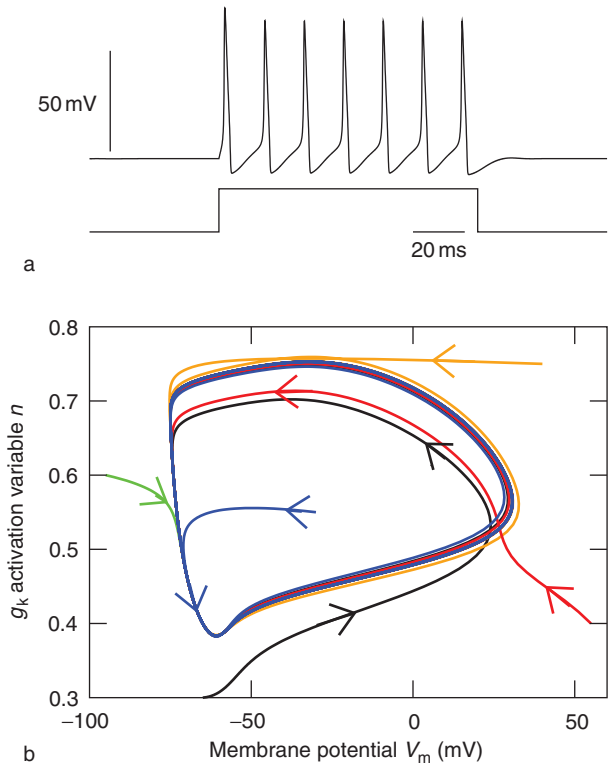

Figure 3 Oscillatory attractor in the original Hodgkin–Huxley model of action potentials: (a) repetitive firing of action potentials with a current pulse $I_{app} = 9\,\mu A/cm^2$; (b) potassium activation variable n plotted against V_m, showing that the oscillation forms a closed trajectory in this projected phase space. The model has a leak current I_L, a fast sodium current I_{Na}, and a noninactivating potassium current I_K. Its dynamics is described by four coupled differential equations (the membrane voltage V_m, the activation and inactivation gating variables m and h for I_{Na}, and the activation variable n for I_K). In (b), the trajectory roughly consists of three portions: the upstroke of an action potential (when both V_m and n increase), the downstroke (when V_m decreases while n keeps increasing and then starts to decrease), and the refractory period (when n continues to decrease while V_m starts to increase). Different colors correspond to five different initial conditions (with different V_m and n values at $t = 0$), in all cases the system dynamically evolves into the oscillatory attractor state. The Hodgkin–Huxley model exhibits bistability between a steady resting state and an oscillatory state, which is not shown for the sake of clarity.

machinery of individual synapses and the electrical activity of single neurons to large neural circuits. It is well known that synapses that form connections between neurons are highly plastic, and experience-dependent synaptic modifications are believed to be a physiological substrate of learning and memory. A single synapse comprises hundreds of proteins that interact with one another in a highly connected signal transduction network; therefore, a synapse is a dynamical system. The biochemical time constants in such a network, and the typical protein lifetimes, range from seconds to hours. So how can synapses store memories that may be retained for many years? One possibility is that the expression of memory maintenance involves changes in the molecular composition of the synapse that are mechanically stable over a long time. Alternatively, memories could be stored in molecular switches, with two or more states that are stable over durations beyond the intrinsic molecular time constants. Recent physiological experiments have yielded evidence for switchlike behavior during synaptic modifications. Molecular studies and modeling have revealed several candidate protein kinases that may exhibit switchlike behavior at single synapses, such as calcium/calmodulin-dependent protein kinase II (CaMKII), protein kinase C (PKC), and mitogen-activated protein kinase (MAPK). Memory switches do not necessarily have an infinite lifetime because an active state may be turned off by subsequent synaptic changes during ongoing neural activity by protein turnover or by molecular fluctuations due to the small synaptic volume (approximately 0.1 fl). Nevertheless, the key point here is that, if individual synapses exhibit stable self-sustained active states (attractors), the lifetime of a memory trace is not directly limited by the biochemical time constants of synaptic signaling pathways.

Whereas synaptic plasticity provides a structural basis for memory formation, the stored information is encoded in a distributed manner by the synaptic connection patterns in a neural circuit. Theoretically, it has been proposed that associative memories are learned through the creation of stable neural activity patterns (attractors). In this view, a memory network has many attractor states, each representing a particular memory item, and has its own basin of attraction within which other states evolve dynamically into the attractor state (Figure 1). The stability of attractor states ensures that memory storage is robust against small perturbations. These memory states are imprinted in the network by long-lasting changes of synaptic connections through Hebbian learning, and much theoretical work has been devoted to the analysis of storage capacity (the number of memory items that can be stored and retrieved reliably as a function of the network size). Memories thus established are associative because a partial cue brings the network into the basin of attraction of an attractor with the information content close to that of the sensory cue. Thus, memory retrieval can be done by association between a cue and the corresponding memory item, and the recall process is error-correcting (incomplete information still leads to the correct memory retrieval). This capability for pattern completion is a hallmark of associative memory. At the same time, an attractor model is also capable of pattern separation, in the sense that two slightly different input patterns near the boundary of two basins of attraction may

drive the network to two distinct attractor patterns, leading to the retrieval of two separate memories. To experimentally test the attractor model, many neurons must be simultaneously monitored so that distributed activity patterns in memory circuits can be assessed. One of such circuits is the hippocampus, which is known to be critical to the formation of episodic memory and spatial memory. In rodents, during exploration in a familiar environment, pyramidal place cells in the hippocampus are activated when the animal passes through a specific location, called a place field. This spatial selectivity is characterized by a bell-shaped tuning curve (neuronal firing rate as a function of the animal's location). Place fields of hippocampal cells cover the entire surface of the environment with about an equal density, so that the neural ensemble firing can be decoded to read out the animal's location in space. The hippocampus thus implements a neuronal representation of a spatial map. Moreover, when the animal is exposed to another environment, the place fields of cells undergo great changes (remapping) until a representation of the new environment is established in the hippocampus. Therefore, the hippocampus stores distinct maps, with each map being activated as the rat enters a different environment. It has been proposed that these spatial representations reflect distinct attractor states, based on the observation that activity of place cells is preserved in the dark (without external visual cues). The attractor model has recently been directly tested by combining a clever navigation task design with simultaneous recording from many single cells in the hippocampus. In one experiment, rats were trained to forage in two distinct environments, a square and a circle that differed in color, texture, and shape. This led to the remapping of place fields in the majority of cells recorded in the hippocampal area CA1. Then, in probe trials, the rat commuted between a series of environments of the same color and texture, but the shape morphed gradually between the square and the circle. These environments were chosen randomly from trial to trial. Remarkably, it was found that the place fields of the recorded cells abruptly and coherently changed from squarelike to circlelike (Figure 4). This observation provides strong support for the attractor model, which predicts that the hippocampal coding of space should exhibit both pattern completion (minor changes from a familiar environment – a square or a circle – do not alter place fields) and pattern separation (similar inputs intermediate between squarelike and circlelike result in drastically different place fields, due to a switch between the two learned maps).

Persistent neural firing observed in a brain area may not necessarily be generated locally but is a mere reflection of mnemonic activity elsewhere. The area CA1 is commonly viewed as an readout circuit, whereas memories are believed to be stored upstream. It has long been hypothesized that autoassociative memories are formed in the area CA3, which projects to CA1 and is endowed with strong recurrent excitatory connections among pyramidal cells, a prerequisite for the generation of attractor states. Furthermore, recent discoveries have drawn attention to the entorhinal cortex, the major input area for the hippocampus. Cells in the dorsolateral entorhinal cortex fire whenever the rat is on any vertex of a triangular lattice spanning the whole surface of the environment. This grid-cell activity is preserved in the dark, when the visual cues are absent. These findings have led to the proposal that the entorhinal cortex embeds an attractor network with a gridlike representation of the environment and that the hippocampus transforms this periodic firing pattern in the input into a nonperiodic firing pattern that encodes the animal's current spatial position. If so, it remains to be seen whether the hippocampus still needs to operate as an attractor network and whether spatial navigation is subserved by interconnected attractor circuits.

In order for an animal to navigate without reference to external cues, the neural instantiation of a spatial map must be constantly updated by the integration of linear and angular self-motion. Whether an attractor model is capable of carrying out such computations robustly is not well understood. Finally, little is known about precisely how the attractor landscape in the space of neuronal firing patterns is shaped by the details of the animal's training process. Another study, using a somewhat different learning and probe procedure but also with morphed environmental shapes, found that place fields of CA3 and CA1 cells switched more gradually and less coherently. This finding does not necessarily mean that the attractor paradigm is incorrect; we expect from computational work that a different learning history gives rise to different sets of attractor states. Elucidation of the general principles and cellular mechanisms of this learning process in future experiments and modeling will help determine whether the attractor model represents a sound theoretical framework for the neurobiology of learning and memory.

Persistent Activity and Working Memory

Attractor models have also been applied to working memory, our brain's ability to actively hold information online for a brief period of time (seconds). Neurons that maintain working memory must be manifestly active in a sustained manner. How can such persistent activity be generated in the absence

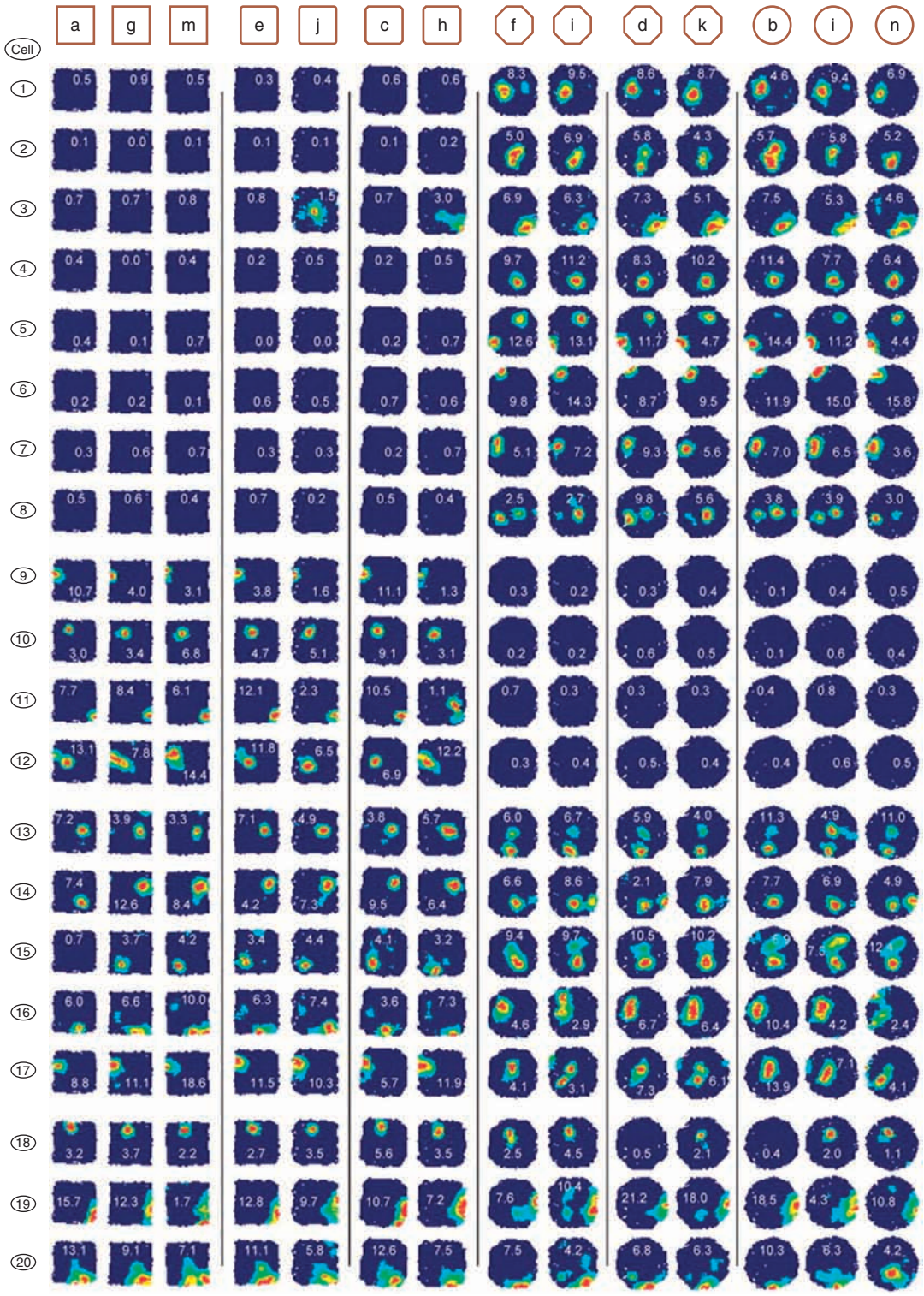

Figure 4 Abrupt and coherent expression of spatial representation of hippocampal neurons. A rat explores a series of environments with morphed shapes (top icons) between a square (the three left-most columns) and a circle (the three right-most columns). Twenty single cells were simultaneously recorded, shown in rows from top to bottom. Warm colors indicate locations (place field) of the rat when the cell's spiking activity is high. Each field is scaled to peak firing rate shown in red. The 17 of 20 simultaneously recorded place cells with different (remapped) firing patterns in the square and the circle, almost all switch from the squarelike to the circlelike pattern between the h and f octagons. Eight cells had fields in the circle but not the square (cells 1–8); four had fields in the square but not the circle (9–12); five fired in both but in different places (13–17); and three did not reach the criterion for remapping (18–20). From Wills TJ, Lever C, Cacucci F, Burgess N, and O'Keefe J (2005) Attractor dynamics in the hippocampal representation of the local environment. *Science* 308: 873–876.

of direct external input? R Lorente de Nó in the 1930s and DO Hebb in the late 1940s proposed that the answer lies in the feedback loop connections. Thus, in a working memory network, every cell receives excitatory drive from both afferent inputs and intrinsic synaptic connections. Inputs activate neurons in a selective cell assembly; the triggered spike activity reverberates through excitatory synaptic circuit, which is enough to sustain an elevated firing when the inputs are withdrawn. This general idea has been made rigorous in attractor models, according to which a working memory circuit exhibits multiple attractor states (each coding a particular memory item) that coexist with a background (resting) state. All the attractor states are self-maintained and relatively stable in the face of small perturbations or noise. Yet memory states can be turned on or switched off by brief external stimuli.

Stimulus-selective neural persistent activity has been observed in awake animals performing delayed-response tasks that depend on working memory. For example, in a delayed match-to-sample task, two visual objects are shown consecutively separated by a time gap of a few seconds, and the subject is asked to judge whether the two stimuli are the same. Or after the presentation of the first object and the delay, an array of visual objects are shown and the subject must indicate by a motor response which of them is identical to the first visual cue. In both cases, the subject's performance relies on the working memory about the first object across the delay. The stored information involves a collection of discrete items. Other delayed-response tasks engage working memory of an analog quantity, such as spatial location or stimulus amplitude. While a monkey was performing such a task, the neurons in the prefrontal, posterior parietal, inferotemporal, and premotor cortices were found to exhibit elevated persistent activity that was selective to stimuli. Three types of mnemonic coding has been observed: (1) object working memory cells are tuned to one or a few of discrete memory items, (2) spatial working memory cells typically exhibit a bell-shaped (Gaussian) tuning function of the spatial location or directional angle, and (3) cells that store parametric working memory of magnitudes (e.g., vibration stimulus frequency) are characterized by a monotonic tuning function of the encoded feature.

These experimental observations lend support to the attractor model inasmuch as stimulus-selective persistent firing patterns are sustained internally in the absence of direct sensory input, are dynamically stable, and are approximately tonic in time (e.g., across a delay). However, experiments show that delay neural activity is often not tonic but exhibits time variations such as ramping up or ramping down. How to account for the heterogeneity and time courses of mnemonic persistent activity represents a challenge to the attractor network model. Moreover, it remains an open question as to what cellular or circuit mechanisms are responsible for the generation of persistent activity. This question is now being addressed using biologically constrained models of persistent activity. **Figure 5** shows such a recurrent network model for spatial working memory in which spiking neurons and synapses are calibrated by the known cortical electrophysiology. The key feature is an abundance of recurrent connections (loops) between neurons, according to a Mexican-hat-type architecture – localized recurrent excitation between pyramidal cells with similar preference to spatial cues and broader inhibition mediated by interneurons (**Figure 5(a)**). In a simulation of a delayed oculomotor task (**Figure 5(b)**), the network is initially in a resting state in which all cells fire spontaneously at low rates. A transient input drives a subpopulation of cells to fire at high rates. As a result, they send recruited excitation to one another via horizontal connections. This internal excitation is large enough to sustain elevated activity, so the firing pattern persists after the stimulus is withdrawn. Synaptic inhibition ensures that the activity does not spread to the rest of the network, and persistent activity has a bell shape (bump attractor). At the end of a mnemonic delay period, the cue information can be retrieved by reading out the peak location of the persistent activity pattern; and the network is reset to the resting state. In different trials, a cue can be presented at different locations. Each cue triggers a persistent firing pattern of the same bell shape but with the peak at a different location (**Figure 5(c)**). A spatial working memory network thus displays a continuous family of bump attractors.

Biophysically realistic models have specific predictions about the circuit properties required for the generation of stimulus-selective persistent activity. In particular, it was found that a network with strong recurrent loops is prone to instability if excitation (positive feedback) is fast compared to negative feedback, as is expected for a nonlinear dynamical system in general. This is the case when excitation is mediated by the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs), which are approximately two to three times faster than inhibition mediated by γ-aminobutyric acid A ($GABA_A$) receptors (time constant, 5–10 ms). The interplay between AMPARs and $GABA_A$ receptors in a excitatory-inhibitory loop naturally produces fast network oscillations. In a working memory model, the large amount of recurrent connections, needed for the generation of persistent activity, often leads to excessive oscillations that are detrimental to network stability.

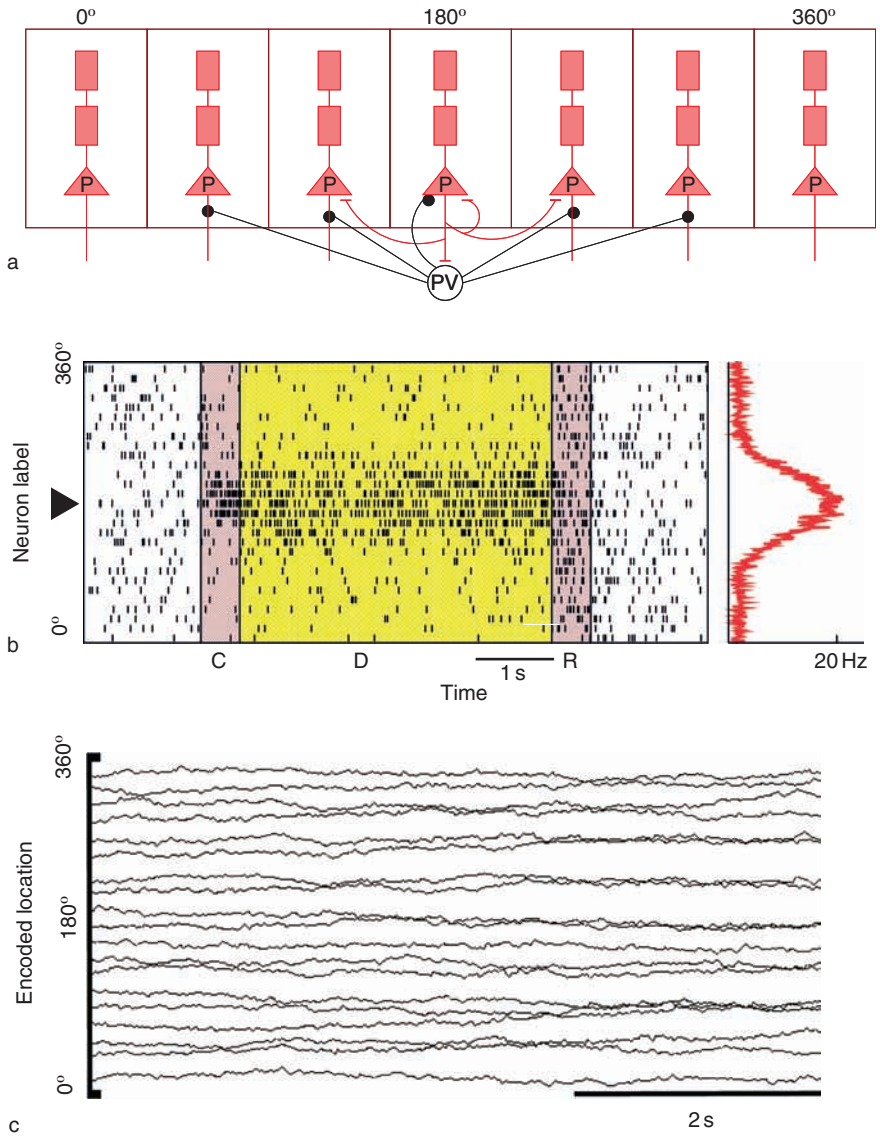

Figure 5 Bump attractor model for spatial working memory: (a) Mexican-hat-type connectivity of the model circuit; (b) spatiotemporal activity pattern of the pyramidal cell population in a simulation of delayed oculomotor response task; (c) temporal evolution of the peak location of mnemonic persistent activity pattern in 20 trials with transient stimuli at different locations. In (a), pyramidal (P) neurons are arranged according to their preferred cues (0–360°). Recurrent excitatory connections are strong between cells with similar cue preference and decrease with the difference in their preferred cues. Local excitation is counteracted by broad synaptic inhibition mediated by paravalbumin (PV) expressing inhibitory interneurons. In (b), each dot is an action potential; the elevated and localized neural activity is triggered by the cue stimulus at 180° and persists during the delay period. On the right is shown the spatial pattern, in which the average firing rate during the delay period is plotted vs. the neuron's preferred cue. The memory of the spatial cue is stored by the peak location of this bell-shaped persistent activity pattern (bump attractor). In (c), the remembered cue (defined by the peak location of the network activity pattern) as a function of time during a 6 s delay period, for a number of trials, each with a different initial spatial cue. R, response period. Panel (b) From Compte A, Brunel N, Goldman-Rakic PS, and Wang X-J (2000) Synaptic mechanisms and network dynamics underlying visuospatial working memory in a cortical network model. *Cerebral Cortex* 10: 910–923; panel (c) from Renart A, Song P, and Wang X-J (2003) Robust spatial working memory in a heterogeneous network model with homeostatic synaptic scaling. *Neuron* 38: 473–485.

Working memory function can be rendered stable if excitatory reverberation is slow, that is, contributed by the *N*-methyl-D-aspartate (NMDA) receptors (time constant 50–100 ms) at recurrent synapses. Thus, the model predicts a critical contribution of NMDA receptors to working memory. Other processes with time constants of hundreds of milliseconds, such as short-term synaptic facilitation or intrinsic ion channels in single cells, may also contribute to reverberatory dynamics underlying working memory.

On the other hand, the feedback mechanism cannot be too slow. An alternative to the attractor model is the scenario in which persistent activity actually is not stable but represents slowly decaying afterdischarges mediated by some intracellular mechanisms such as second-messenger pathways or kinetics of protein kinases. However, this scenario predicts that triggering inputs must be long lasting (lasting for seconds), which is incompatible with physiological experiments in which working memory states have been shown to be switchable quickly by brief external inputs (a few hundreds of a millisecond). The recurrent (attractor) network mechanism achieves the stability and long persistence time of memory storage, as well as rapid flexible memory encoding and erasure, that are behaviorally desirable.

Winner-Take-All and Vector Averaging

Attractor dynamics naturally instantiate winner-take-all (WTA); when several neural pools, each selective for a different sensory stimulus, are simultaneously activated by the presence of several stimuli, competition through recurrent attractor dynamics may ultimately lead to an output pattern with a high firing activity in one neural pool while all the other neural pools are suppressed. WTA could subserve a number of neural processes, such as categorical decision making. In a continuous neural network, such as the ring model shown in Figure 6, WTA operation is not always categorical but depends on the similarity between simultaneously presented stimuli. For instance, if two inputs are very close to one another, instead of WTA the network's response is expected to be a vector average of the two individual responses to the stimuli presented alone. Moreover, in a working memory network, WTA often involves different stimuli that occur at different times; this interaction across temporal gaps is possible because an earlier input can trigger persistent activity that interacts with responses to later inputs.

These points are illustrated in Figure 6, which shows the interaction of a remembered stimulus and a distractor in the continuous (bump) attractor network model of spatial working memory (same as in Figure 5). Significantly, the to-be-remembered stimulus (located at θ_S) and the distractor (θ_D) have the same amplitude but presented at different times, one at the beginning of a trial and the other during the delay. As we intuitively expect, the impact of a distractor depends on its strength (saliency). If the stimulation amplitude is sufficiently large, the distractor is powerful enough to overcome the intrinsic dynamics of the recurrent circuit, and the network is always perturbed to a location close to the intervening stimulus (Figure 6(a), top; red curve in Figure 6(b)). In this case, the network can be reset by every new transient stimulus and keeps a memory of the last stimulus in the form of a refreshed selective persistent activity state. On the other hand, if external input is moderately strong relative to recurrent synaptic drive, a distractor triggers only a transient response and the memory of the initial cue is preserved (Figure 6(b), bottom). The resistance to distractors can be understood by the fact that, in a memory delay period, active neurons recruit inhibitions which project to the rest of the network. Consequently, those cells not encoding the initial cue are less excitable than when they are in the resting state and hence are less responsive to distracting stimuli presented during the delay. Moreover, WTA takes place only when the spatial locations of the initial cue and the later distractor are sufficiently distant from one another (more than 90°) (blue curve in Figure 6(b)). When they are closer, the activated neural pools (selective for θ_S and θ_D) overlap one another and the peak location of the resulting population activity is roughly the vector average of θ_S and θ_D (blue curve in Figure 6(b)). Hence, the same network can perform both WTA and vector-averaging computations, depending on the similarity of stimuli.

Time Integration and Categorical Decision Making

Cortical areas that are engaged in working memory – such as the prefrontal cortex – are also involved in other cognitive functions such as decision making, selective attention, and behavioral control. This suggests that microcircuit organization in these areas is equipped with the necessary properties to subserve both the internal representation of information and dynamical computations of cognitive types. As it turns out, models originally developed for working memory can account for decision-making processes as well. An example is shown in Figure 7 from model simulations of a visual motion-discrimination experiment. In this two-alternative forced-choice task, monkeys are trained to make a judgment about the direction of motion (say, left or right) in a stochastic random dot display and to report the perceived direction with a saccadic eye movement. A percentage of dots (called motion strength) move coherently in the same direction, so the task can be made easy or difficult by varying the motion strength (from close to 100 to 0%) from trial to trial. While a monkey is performing the task, single-unit recordings revealed that neurons in the posterior parietal cortex and prefrontal cortex exhibit firing activity correlated with the animal's perceptual choice. For example, in a trial

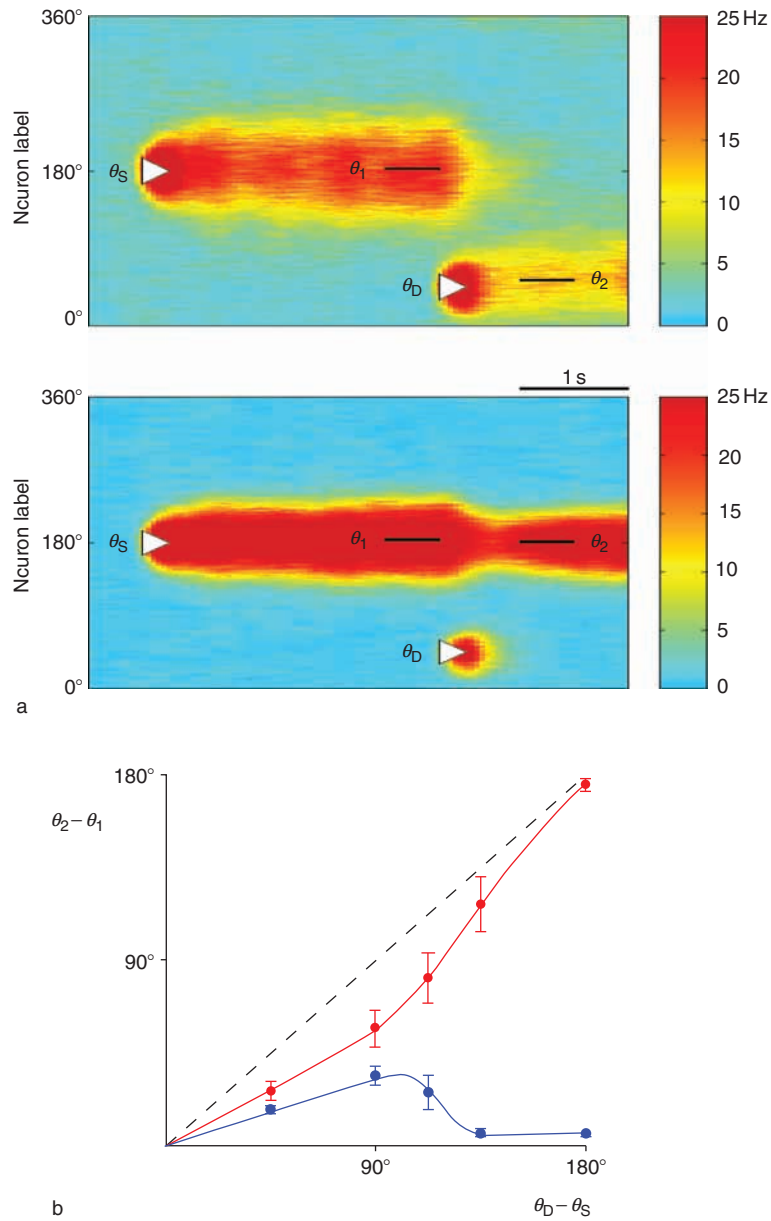

Figure 6 Winner-take-all and vector averaging in the continuous attractor network model of spatial working memory (**Figure 5**): (a) two sample simulations; (b) dependence of network distraction on the distance between the cue and distractor and on the stimulation intensity. In (a), the network spatiotemporal firing pattern is plotted in a color-coded map. A cue is presented initially at θ_S, triggering a tuned persistent activity. After a 2.5 s delay, a distractor stimulus is presented at θ_D, with the same intensity and duration as the cue stimulus. The peak location of the population activity pattern is computed just before the distractor (θ_1) and 500 ms after the distractor (θ_2). The network is completely distracted by a strong stimulus (top), but is resistant to a distractor of moderate intensity (bottom). In (b), the distracted angle $\theta_2 - \theta_1$ is plotted vs. the distraction angle θ_D–θ_S for several distractor cues. The dashed line indicates perfect distraction (as in (a), top), whereas points on the *x*-axis imply the complete absence of distraction (as in (a), bottom). The red curve indicates large input intensity; the blue curve indicates moderate input intensity. In the latter case, the network exhibits a switch from winner-take-all (when $\theta_D - \theta_S$ is large) to vector averaging (when θ_D–θ_S is small). Adapted from Compte A, Brunel N, Goldman-Rakic PS, and Wang X-J (2000) Synaptic mechanisms and network dynamics underlying visuospatial working memory in a cortical network model. *Cerebral Cortex* 10: 910–923.

in which the motion strength is low (say, 6.4%), if the stimulus direction is left but the monkey's choice is right, the response is incorrect. In that case, cells selective for right display a higher activity than those selective for left. Hence, the neural activity signals the animal's perceptual decision rather than the actual sensory stimulus. This experiment can be simulated using the same model designed for working

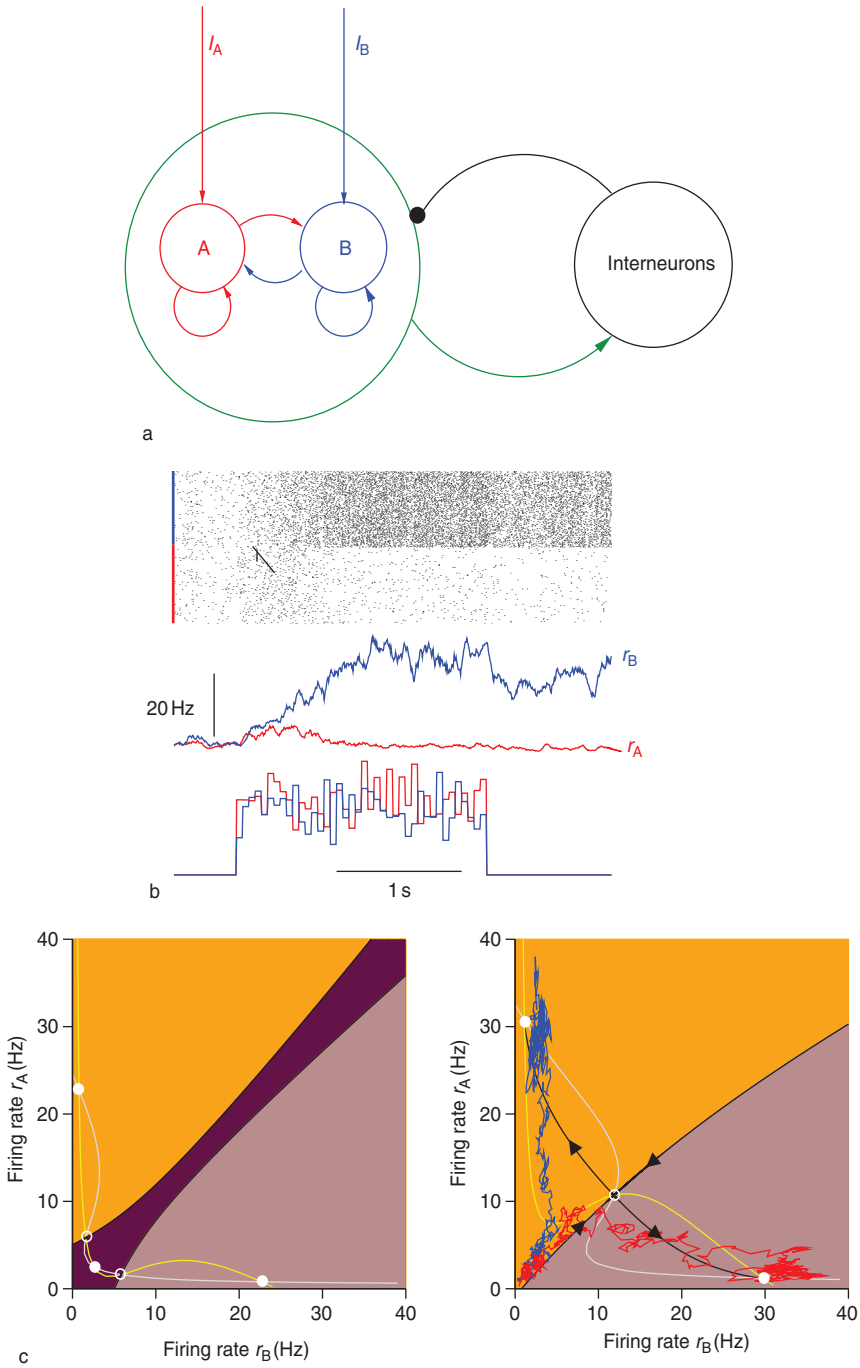

Figure 7 A spiking neuron circuit model for two-alternative forced-choice tasks: (a) model scheme; (b) network simulation with $c' = (I_A - I_B)/(I_A + I_B) = 6.4\%$; (c) decision dynamics shown in the two-dimensional plane where firing rates r_A and r_B are plotted against one another. In (a), there are two groups of spiking pyramidal cells, each of which is selective to one of the two directions (A = left, B = right) of random moving dots in a visual motion-discrimination experiment. Within each pyramidal neural group, there are strong recurrent excitatory connections which can sustain persistent activity triggered by a transient preferred stimulus. The two neural groups compete through feedback inhibition from interneurons. In (b), the population firing rates r_A and r_B exhibit an initial slow ramping (time integration) followed by eventual divergence (categorical choice). In this sample trial, although the input A is larger, the decision is B (an erroneous choice). As shown in (c), in the absence of stimulation (left), there are three attractor states (filled circles) and two unstable steady states (open circles); colored regions are the basins of attractions (maroon for the resting state; orange and brown for the persistent activity states). With a stimulus of $c' = 6.4\%$ in favor of choice A (right), the decision trajectory is shown for two trials (correct trial in blue; error trial in red). Panels (a,b) from Wang X-J (2002) Probabilistic decision making by slow reverberation in neocortical circuits. *Neuron* 36: 955–968; panel (c) from a simulation by KF Wong using the model published in Wong KF and Wang X-J (2006) A recurrent network mechanism for time integration in perceptual decisions. *Journal of Neuroscience* 26: 1314–1328.

memory. The only difference between a working memory simulation and a decision simulation is that in a delayed-response task only one stimulus is presented but for a perceptual discrimination task conflicting sensory inputs are fed into competing neural subpopulations in the circuit. This is schematically depicted in Figure 7(a), in which the relative difference in the inputs $c' = (I_A - I_B)/(I_A + I_B)$ mimics the motion strength in the visual motion discrimination experiment. Figure 7(b) shows a simulation with $c' = 6.4\%$. At the stimulus onset, the firing rates of the two competing neural populations, r_A and r_B, initially ramp up together for hundreds of milliseconds before diverging from one another when one increases while the other declines. The perceptual choice is decided based on which of the two neural populations wins the competition. Therefore, consistent with the physiological observations from the monkey experiment, the decision process proceeds in two steps. Sensory data are first integrated over time in a graded fashion, which in the model is instantiated by the NMDA receptor-dependent slow reverberation. This is followed by WTA competition produced by synaptic inhibition, leading to a categorical (binary) choice.

Figure 7(c) shows attractor dynamics in the decision space, where r_A is plotted against r_B. In the absence of stimulation (Figure 7(c), left), three attractors coexist (filled circles): a spontaneous state (when both r_A and r_B are low), and two persistent activity states (with high r_A and low r_B , or vice versa). On the presentation of a stimulus (Figure 7(c), right), the attractor landscape is altered and the spontaneous steady state disappears, so the system is forced to evolve toward one of the two active states which represent perceptual decisions (A or B). In this graph, the sensory evidence is in favor of the choice A (with $c' = 6.4\,\%$), so the attractor A has a larger basin of attraction (orange) than that of the attractor B (brown). The system is initially in the spontaneous state which now falls in the basin of attraction A and evolves toward decision state A in a correct trial (blue). However, at low c', the bias is not strong, and noise can induce the system's trajectory to travel across the boundary of the two attraction basins, in which case the system eventually evolves to decision state B in an error trial (red). The crossing of a boundary between attraction basins is slow, which explains why the reaction times are longer in error trials than in correct trials, as was observed in the monkey experiment. After the offset of the stimulus, the system's configuration reverts to that in Figure 7(c) (left). Because a persistently active state is self-sustained, the perceptual choice (A or B) can be stored in working memory for later use, to guide behavior. In this way, the attractor dynamics model offers an unified account for working memory and decision-making computations.

Concluding Remarks

In summary, the language of attractors is natural for describing the electrical activity of neurons and neural circuits. It provides a plausible theoretical framework for both short-term working memory and long-term associative memory. Significantly, nonlinearity due to feedback loops makes it possible that graded changes of a cellular or synaptic parameter lead to the emergence of qualitatively different behaviors (e.g., with or without persistent activity). The functional implications of this insight are potentially far-reaching because it suggests that cortical areas dedicated to distinct functions (e.g., sensory processing vs. working memory) may share a same canonical cortical circuit layout but with subtle differences in the cellular/molecular makeup, connectivity properties, and neuromodulatory influence.

Qualitatively speaking, working memory requires neurons to convert a transient input pulse into a sustained persistent activity, such as a time integral of the stimulus. Similarly, in perceptual decisions, approximate linear ramping activity, at a rate proportional to input strength, can also be conceptualized as time integration. It is worth noting, however, that a genuine integrator implies that, after a transient input is turned off, the activity is persistent at a firing rate proportional to the input strength, spanning a continuous range. This is not the case in Figure 7, in which after the stimulus offset the neural activity is binary (representing one of the two categorical choices), independent of the input motion strength. This is what has been observed in posterior parietal neurons, and it is the kind of neural signals needed to accomplish categorical choices. Integration by neural circuits over long timescales represents an important topic of active research.

An open question concerning working memory is whether persistent activity is primarily generated by local circuit dynamics, a single-cell property, or a large-scale network composed of several brain regions. For both working memory and long-term memory, a major challenge is to elucidate the biological substrates that underlie the robustness of continuous attractors and integrators. Furthermore, there is a dichotomy between rapidly switchable attractors, on the one hand, and an intracellular signaling network with

multiple time constants, on the other hand. The interplay between cellular processes and collective network dynamics will turn out to be an exciting topic in future research.

Acknowledgments

I thank K-F Wong for making Figure 7(c). This work is supported by the NIH grant MH62349.

See also: Hodgkin–Huxley Models; Neural Integrator Models; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Short Term and Working Memory; Visual Associative Memory; Working Memory: Capacity Limitations.

Further Reading

Amari S (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics* 27: 77–87.

Amit DJ (1992) *Modelling Brain Function: The World of Attractor Neural Networks*. Cambridge, UK: Cambridge University. Press.

Ben-Yishai R, Bar-Or RL, and Sompolinsky H (1995) Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 92: 3844–3848.

Bhalla US and Iyengar R (1999) Emergent properties of networks of biological signaling pathways. *Science* 283: 381–387.

Brunel N (2005) Network models of memory. In: Chow CC, Gutkin B, Hansel D, Meunier C, and Dalibard J (eds.) *Methods and Models in Neurophysics*, pp. 407–476. Amsterdam: Elsevier.

Camperi M and Wang XJ (1998) A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *Journal of Computational Neuroscience* 5: 383–405.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America* 79: 2554–2558.

Machens CK, Romo R, and Brody CD (2005) Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science* 307: 1121–1124.

McNaughton BL, Battaglia FP, Jensen O, Moser EI, and Moser MB (2006) Path integration and the neural basis of the "cognitive map." *Nature Reviews Neuroscience* 7: 663–678.

Miller P, Zhabotinsky A, Lisman J, and Wang X-J (2005) The stability of a stochastic CaMKII switch: Dependence on the number of molecules and protein turn over. *PLoS* 3: 705–717.

O'Connor DH, Wittenberg GM, and Wang SS (2005) Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proceedings of the National Academy of Sciences of the United States of America* 102: 9679–9684.

Seung HS (1996) How the brain keeps the eyes still? *Proceedings of the National Academy of Sciences of the United States of America* 93: 13339–13344.

Strogatz SH (1994) *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Reading, MA: Addision-Wesley.

Tsodyks M (1999) Attractor neural network models of spatial maps in hippocampus. *Hippocampus* 9: 481–489.

Wang X-J (2001) Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neuroscience* 24: 455–463.

Wang X-J (2002) Probabilistic decision making by slow reverberation in neocortical circuits. *Neuron* 36: 955–968.

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience* 16: 2112–2126.

Audiovocal Communication in Bats

J S Kanwal, Georgetown University, Washington, DC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The advanced status of audiovocal communication in bats, in general, rivals the sophisticated communication strategies reported in mammalian species with dramatically larger brains, such as primates and cetaceans. Audiovocal communication fulfills an important survival function for insectivorous bats. These bats forage under low light conditions, and in some species social interactions and breeding may occur in complete darkness, where coordination via communication sounds is critical. With technical advances enabling a rapid analysis of complex sounds, there has been a proliferation of information on audiovocal communication in several species of bats. The social behavior of bats, however, is relatively difficult to study in their natural environment and much of our knowledge of audiovocal social communication in bats has been obtained from observations of interactions between conspecifics in a colony of captive, free-flying bats. In some species, such as *Carollia*, field studies and observations made on captive bats indicate that under seminatural conditions, their social organization and interactions remain relatively unchanged. The vocal apparatus and vocal control circuits in the brain of insectivorous bats are especially well developed and underlie their ability to emit complex vocalizations. Their auditory apparatus and the accompanying neural machinery are equally advanced at all levels, from the elaborate structure of the ear pinnae to the specializations of the cochlea and hair cells in the periphery and of neurons in the central nervous system.

Audiovocal communication can be studied at three distinct levels. First is the acoustic level, which requires a clear understanding of the acoustic structure of communication sounds. Second is the behavioral level, which is intricately connected to the social interactions between conspecifics, and the third is the sensorimotor (audiovocal) neural level. Audiovocal communication in bats, as in other animals, evolves through communication behavior as an adaptive interface of the organism with its biological and physical environment. This is especially well developed in many insectivorous, microchiropteran bat species since sound continues to be a viable means of signaling when vision and vision-mediated tactual interactions are eliminated or minimized in the darkness of caves and the night sky.

The ability to produce a rich variety of sounds for communication, especially via vocal learning, appears to be independent of brain size among mammals and birds. For example, songbirds with relatively small brains are more advanced in their vocal learning skills than other species with larger brains, such as ostriches and penguins. Whereas echolocation has been studied extensively in many species of bats, auditory communication is only now being studied in an increasing number of species. This article summarizes data on a few species of bats that have been extensively studied and/or exhibit characteristics that provide new representative examples of advancements in audiovocal communication among the taxonomic group of Chiroptera.

Echolocation versus Communication

Frequently, in the published literature, both echolocation and communication sounds emitted by bats have been referred to as 'calls,' blurring the important and clear differences in the function of these two types of signals. Whereas echolocation sounds are produced nearly continuously during flight and occasionally during roosting, communication sounds are largely produced in a roost and during flight-mediated social interactions. Spectrally, echolocation sounds may not be very different from some types of communication sounds, but as the name implies, echolocation sounds are meaningless to the emitter until paired with their respective echoes. They are also relatively stereotypic, with systematic differences observed in their structure during different phases of food search and insect capture in the case of insectivorous bats. Communication sounds, as a group, cover a wide range of acoustic forms, varying from pure tone-like to various levels of frequency modulations and noise bursts as well as mixtures of these three basic acoustic elements. Using 'discriminant' or 'principal components' analysis, most calls can be grouped into discrete clusters, with each cluster representing a wide range of variations in multiple parameters (Figure 1). Clearly, the factors shaping the acoustic structure of communication sounds in bats are different from those that shape their echolocation sounds. For example, the latter are greatly influenced by the foraging strategy of a species and its ecological niche. Variation in communication sounds may follow motivational–structural rules as first proposed by Morton in 1977.

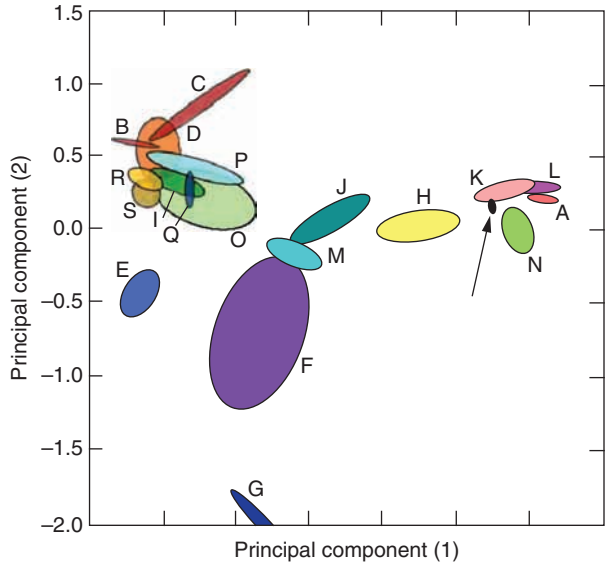

Figure 1 Scatter plot obtained from a principal-components analysis for mustached bat social calls. Bivariate sample ellipsoids drawn at the 75% confidence interval based on scatter plots of the first two principal components of a three-dimensional solution for the multivariate parameters of syllables (plotted with letters). Ellipsoids that overlap in the first two components are separated by the third component. Echolocation pulses (arrow) are the most tightly grouped of all vocalizations. Letters correspond to the following syllable abbreviations (see **Table 1**): A, TCFs; B, QCFs; C, QCFl; D, tQCF; E, NNBs; F, NNBl; G, rBNB; H, bUFM; I, bDFM; J, cDFM; K, sAFM; L, dAFM; M, sHFM; N, dRFM; O, fRFM; P, fSFM; Q, qSFM; R, sWFM; S, WFMl. Adapted from Kanwal JS, Matsumura S, Ohlemiller K, and Suga N (1994) Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *Journal of the Acoustical Society of America* 96: 1229–1254, with permission.

To emphasize the functional distinction between the two types of sounds, echolocation sounds are best referred to as 'echolocation pulses' rather than calls, and communication sounds are best referred to as social calls or simply calls. This distinction is consistent with the neurophysiological literature on processing of echolocation signals. Per this terminology, neurons in the auditory system are tuned to combinations of various parameters of a pulse–echo pair, with the pulse being the relatively brief, stereotypic sound produced by a bat for echolocation. Suga and co-workers have shown that this tuning extracts important information relevant for echolocation (e.g., target range during insect pursuit).

Eavesdropping

Eavesdropping or 'secretly listening in' may be considered as a passive form of communication or gathering information from the vocal signals of conspecifics. An echolocating animal gives away a lot of information about itself, its position, and flight path by the timing and structure of its echolocation pulses. High-intensity echolocation pulses may also reveal the identity of the emitter and possibly the presence of a food source to an attentive conspecific. Therefore, echolocation signals can provide useful cues to eavesdroppers that may use this information for various purposes. For example, it has been shown that instead of focusing on one's own echolocation pulse and echo information, bats may use the echolocation pulses of conspecifics to obtain positional information about preferred roost mates. Möhres made this observation in the horseshoe bat, *Rhinolophus ferrumquinuum*, in 1967. Similarly, *Myotis lucifugus* has been experimentally shown to use the presence of a high density of echolocation pulses emitted by conspecifics in the field as an indicator of various resources such as a food source at a particular location. This is intriguing because several neural mechanisms have been enunciated that show how CF–FM bats (species with constant frequency (CF) and frequency modulated (FM) elements in their echolocation pulse) may avoid hearing and therefore jamming of their echolocation signal by conspecifics when thousands of them located in a roost or in close proximity emit high-intensity (~120 dB SPL) echolocation pulses at the same time. Accordingly, cortical neurons of at least some bat species, such as mustached bats, are sharply tuned and highly specialized in that they respond well only to specific parameters in the echo following the pulse. Interestingly, the same neurons are not as sharply tuned to the frequency of the fundamental in the echolocation pulse. Therefore, it is plausible that individuals can eavesdrop on the fundamental frequency in echolocation pulses emitted by conspecifics to locate a food source or be alerted to other foraging and social activities, such as mating. Young bats have been shown to be more responsive than adults to playback presentations of echolocation pulses. This could result from a general increased level of curiosity and playfulness observed in the young of many mammalian species. In some larger bat species, adults use echolocation pulses produced by foraging conspecifics to achieve spacing of at least 50 m from conspecifics on feeding grounds. This type of transfer or gathering of information from echolocation sounds is best labeled as an indirect one-way form of communication.

Mother–Infant Interactions

In bats, echolocation pulses, sometimes considered as a form of autocommunication or communication with the self, may also inform conspecifics and modify their behavior in some contexts. In these contexts, they function as calls. A *bona fide* use of echolocation pulselike sounds for the purpose of communication is

during mother–infant interactions in which the mother may use them to identify its pup and vice versa. Among all communication behaviors that involve the production of sounds, mother–infant communication is one of the best studied. Calls produced during these interactions have been studied in at least 24 different species. Since orientation or echolocation signals are among the first to be produced by infants in many species, these play a major role in communication via antiphonal calling between infants and their respective mothers in a bat colony. Some of the studies documenting this behavior were conducted in the little brown bat, *M. lucifugus*, and in Japanese horseshoe bats, *Rhinolophus ferrumequinum Nippon*. In addition, Habersetzer has shown that *Rhinopoma hardwicki*, found in India, can decrease the slope of their normal echolocation pulses when emitting them in the roost compared to the pulse emitted in flight. In noctilionids (*Noctilo leporinus*) and a few other species, two bats on a collision course drop the frequency sweep of the echolocation pulse by one octave to warn an approaching bat and avoid a collision. This drop in frequency may activate neurons that are tuned broadly to low frequencies and are not engaged in processing echolocation information. For purposes of echolocation, the second harmonic of the echolocation pulse in mustached bats has the most energy so it can provide information from the echo returning from a target. Sometimes, however, mustached bats produce bursts of echolocation pulselike sounds in which the predominant energy is shifted to the fundamental instead of the second harmonic. These variants of the echolocation pulse are produced only in the presence of conspecifics.

Echolocation pulses likely coevolved by modification of a communication call, perhaps associated with foraging, since the latter are also present in nonecholocating bats and other mammalian species. A scaled multidimensional representation of several acoustic parameters present in calls and echolocation pulses of mustached bats shows that echolocation pulses share parameters with several different calls (Figure 2). The observation that echolocation pulses and their variants can serve a communicative role in various contexts even in modern-day bats, as shown by Fenton and Gould, is consistent with this idea. The evolutionary relationship between echolocation and communication sounds, however, needs to be investigated in a number of different species.

Acoustic Structure of Social Calls

From an acoustics standpoint, all natural sounds, including communication calls, consist of three basic types of elementary sounds, namely CF sounds, FM sounds, and noise the burst (NB)-type sounds. Pure tones are rare in nature; accordingly, very few calls are strictly tonal, although many have a clear harmonic structure (Figure 2). Most complex calls consist of combinations of tones and/or noise bursts; the latter can be either broad- or narrowband, whereas the former contain multiple (harmonic) bands of predominant frequencies. Additional complexities may be present in both the spectral and the temporal domains. Both animal and human sounds may contain varying levels of noisiness, making it difficult to identify distinct classes of calls. Also, sounds with two different fundamental frequencies may be emitted at the same time and independently modulated. The simple syllables in mustached bats are defined on the basis of the statistically significant differences in the overall spectrographic pattern of calls so that any one pattern is present in only one call type. These basic patterns of acoustic structure are also present in the calls of the other species, such as monkeys and cats, as well in the phonemes in speech sounds, although the vocal frequency range and audiogram in these species is shifted down by nearly an order of magnitude and the meaning of each sound is obviously species specific.

Social calls in bats are grouped into categories or call types using different criteria depending on an investigator's method of analysis. Traditionally, social calls have been grouped according to the social behavior with which they are associated – for example, aggressive or warning calls, mating and song flight calls, distress calls, or isolation and direction calls. In a few cases, they may be grouped according to the subjective description of the sound (e.g., a twitter, scream, or whistle), but this anthropocentric categorization is not very meaningful because, unlike bird calls, many of the acoustic elements in bat calls may lie in the ultrasonic range and are therefore not heard. A more accurate, although behaviorally less-meaningful way is to categorize calls on the basis of the acoustic elements present within a call and the geometric pattern these may take. For example, calls may be considered as being largely CF, FM, or NB call types (Table 1). This relatively objective classification scheme has been used to define the calls of at least two bat species in detail and allows one to visualize roughly the spectrographic structure of calls simply by knowing its name. This also allows a more objective way to compare call structures in related species of bats from an acoustics standpoint, which can be later used to compare their behavioral correlates as well. Prefixes such as long and short give additional information about the call. Dissociating call names and behaviors also makes it possible to assign a call to two or more different behaviors with which it may be associated. Alternatively, variants of the same

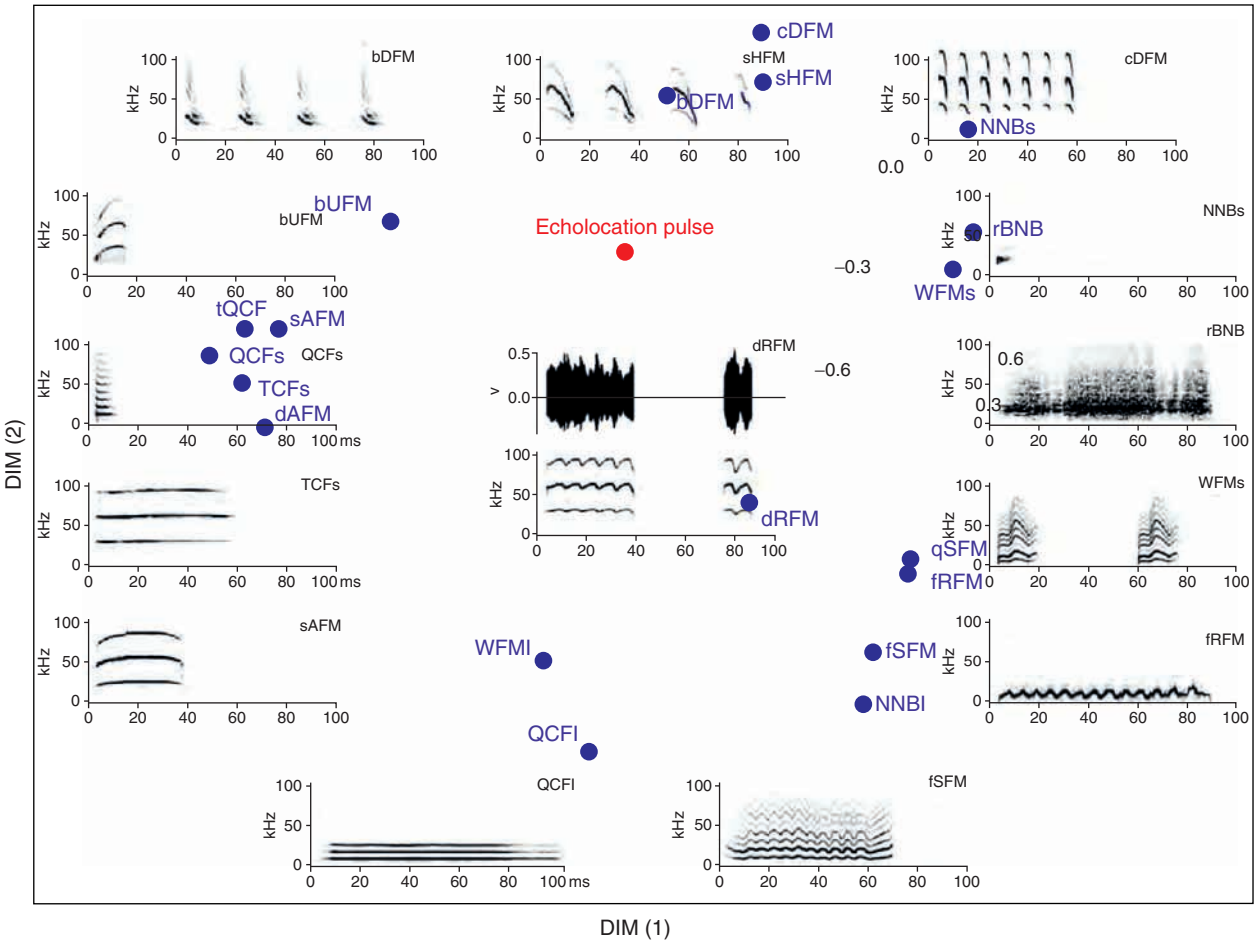

Figure 2 Multidimensionally scaled (MDS) configuration of syllable types (blue dots) and echolocation pulses (red dot) obtained from mean values of nine parameters of mustached bat calls. The structural similarities captured by the plot are illustrated by superimposed spectrograms of several syllable call types roughly corresponding to its position in the two-dimensional MDS plot. More than 90% of the variation was captured for the first three dimensions. All spectrograms are drawn to the same scale. Spectrogram for the descending, rippled FM (dRFM) is shown at the center of the contour plot close to the location of the echolocation pulse. Note the separation of CF and CF-like syllables from FM syllables. Adapted from Kanwal JS, Matsumura S, Ohlemiller K, and Suga N (1994) Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *Journal of the Acoustical Society of America* 96: 1229–1254, with permission.

call type may be used in two different contexts. This usage of one call type in more than one behavior has been shown in primates and a few other species. Relatively complete acoustical and statistical analysis of vocalization repertoires has been performed for the mustached bat, *Pteronotus parnellii*, and the horseshoe bat, *R. ferrumequinum*. Earlier studies also provided information on different call types produced by *Carollia perspicallata*, *Pteropus poliocephalus*, *Myotis lucifugus*, and the Mexican free-tailed bat, *Tadarida brasilensis*. Also, the social calls of 16 species of European bats have been recorded to compare their structure and usage.

Syllables and Syntax

For social communication, two or more different categories of sounds may be combined in the time domain to generate 'composites' that do not belong to any one single class of elemental sounds. This ability has been demonstrated in at least two different species of bats. Furthermore, simple syllables are combined within composites according to some rules, which for primate vocalizations has been referred to as a 'phonetic-like' syntax and in birdsong as syllabic syntax, although in the latter case short silent intervals are included as part of a syllable. From recordings of mustached bat calls, only 11 of the 19 simple syllables reported to date in a single subspecies are used to construct composites. For example, the bent upward FM (bUFM) component in a composite is never followed by an attached syllable whose fundamental frequency is lower than the mean terminal frequency of the bUFM. Similarly, a fixed, sinusoidal FM (fSFM) component in a composite typically

Table 1 Abbreviations of syllable types

Abbreviation	*Syllable name*
TCFs	short, True CF
QCFs	short, Quasi CF
QCFl	long, Quasi CF
tQCF	trapezoidal, Quasi CF
NNBs	short, Narrowband NB
NNBl	long, Narrowband NB
rBNB	rectangular, Broadband NB
bUFM	bent, Upward FM
bDFM	bent, Downward FM
cDFM	checked, Downward FM
sAFM	single, Arched FM
dAFM	double, Arched FM
sHFM	single, Humped FM
dRFM	downward, Rippled FM
fRFM	fixed, Rippled FM
fSFM	fixed, Sinusoidal FM
qSFM	quasi, Sinusoidal FM
WFMs	short, Wrinkled FM
WFMl	long, Wrinkled FM

precedes a short, quasi CF (QCFs) or a QCFs-like component and is not attached to syllables with very high (>20 kHz) fundamental frequency. In mustached bat calls, only the duration parameter of a simple syllable appears to be significantly altered when that syllable is combined within a composite (e.g., bUFM). This total duration of a syllable is probably constrained by the lung volume and respiratory rate of the animal; therefore, the duration of one of the components in a composite may be significantly modified.

Simple syllabic calls emitted by mustached bats may also be combined in different forms and contexts to produce a train or 'stream' of sounds. Repetition of similar syllables during such vocalization bouts frequently indicates the urgency of a desired response. In mouse pups, wriggling (isolation) calls emitted at high repetition rates and sometimes with increasing amplitudes trigger a retrieval response on the part of maternal females. This type of sequencing of sounds is less common in animals compared to speech in humans, except in the context of song as in songbirds, bats, dolphins, and whales or chorus calls in some species of frogs and primates. It has been postulated both for bats and for some species of primates that these types of syllable combinations in the time domain follow elementary syntactical rules not unlike those in the production of a stream of speech sounds.

Call Variability

Categorization of call types strictly on the basis of the acoustic structure of calls is accompanied by quantification of the variability in the acoustic structure of each call type. Data indicate that for most calls there is considerable variability among different examples of the same call type, as uttered by different individuals, in terms of both tempo and spectral position (Figure 3). The source of this variability may be the sex, mood, and/or identity of the emitter as well as the social and environmental context in which a call type is emitted. The receivers may use variations in calls to gauge motivational levels of others and to modulate their own vocal behavior so as to interact with conspecifics. Alarm calls may be used to signal an environmental threat, allowing others to behave in the most adaptive manner. Variations in the tonality, pitch, and frequency modulation may also be used to identify individuals and their mood in the dark, especially when olfactory cues are unavailable from a short distance away. Males, females, and infants differ in their fundamental frequencies or pitch. Yet, what is remarkable about auditory communication is that pitch-shifted variants of the same call may reveal the identity/mood of the emitter without altering its contextual identity or meaning. The receiving individual must therefore be capable of ignoring information that is irrelevant identifying a call and derive a caller-invariant representation of each call type. At the same time, complementary information within a call type can be used to define caller identity and other factors that produce the variation. Neuroscientific studies are needed to understand how the brains of bats and other animals extract the different information-bearing elements from the acoustic structure of a call.

Geographic Dialects and Vocal Learning

Geographic variation has been studied most extensively in bird species; for example, trills of the blue tit, *Parus caeruleus*, vary geographically depending on environmental factors, competition, and food availability. Among mammals, much of the research on geographic variation of communication calls has focused on primate species; for example, differences in the acoustical repertoires of free-ranging langurs, *Presbytis entellus*, have been documented and attributed to phylogenetic adaptations to their respective habitats or to the different social organizations of the langur populations. Although geographic variation has been shown to exist in the echolocation calls of certain bat species such as the hoary bat *Lasiurus cinereus* and mustached bat, no such comparative study has been done on bat communication calls even though they represent the vast majority of a bats' vocal repertoire. There is only preliminary evidence of the existence of dialects for calls in bats. These data derive from an analysis of calls

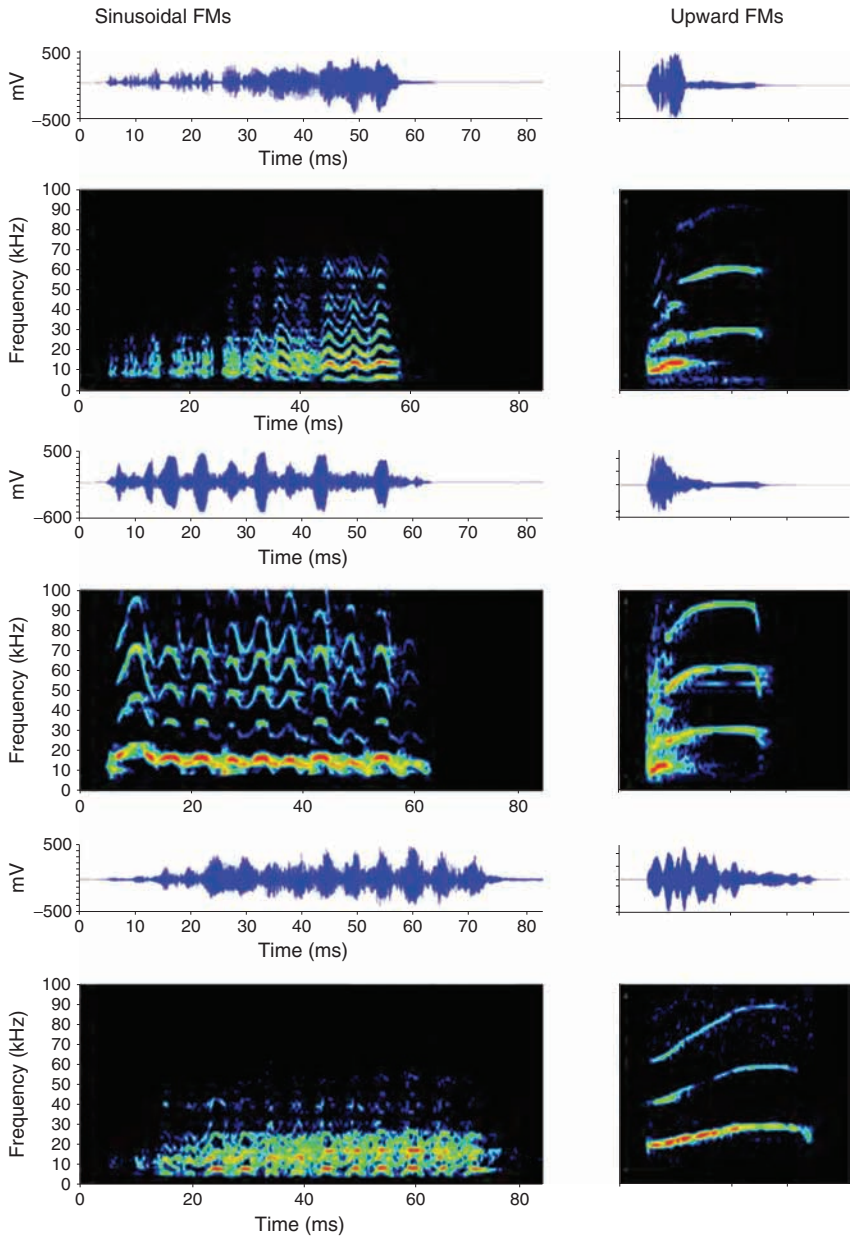

Figure 3 Amplitude envelops and spectrograms of variants of two different call types (same timescale), fixed, sinusoidal FM (fSFM) and bent, upward FM (bUFM), emitted by mustached bats. Multiple parameters are modified between different examples of a call type.

in mustached bats collected from different islands in the Caribbean (Jamaica, Puerto Rico, and Trinidad). The vocal repertoire of Jamaican mustached bats includes four types of CF calls: short, true CF calls (TCFs), long and short quasi-CF calls (QCFl and QCFs, respectively), and a trapezoidal quasi-CF (tQCF). Of these, Trinidad bats have been observed to vocalize only the QCFl and QCFs types of CF calls. Mustached bats obtained from Puerto Rico, *Pteronotus pamellii puertoricensis*, demonstrate a surprising paucity of CF calls and instead emit a CF subsyllable as part of a frequently occurring composite syllable. This call, an upward FM sweep preceding the CF segment, resembles a reversed echolocation pulse, with its harmonics at 30, 60, and 90 kHz and maximal intensity in the second harmonic. Additional CF calls may also be emitted by this species, but they were not readily recorded under seminatural laboratory conditions. Males of *Saccopteryx bilineata* exhibit variation at both the geographic and the individual level in the acoustic profile of their vocalizations.

Dialects usually develop within an ontogenetic time frame and their presence within subspecies of a taxonomic group may be taken as indirect evidence of vocal learning. Several species in four separate vertebrate groups have been shown to exhibit vocal learning.

These include songbirds, whales and dolphins, elephants, and bats. Of these, songbirds and cetaceans have been shown to have dialects and exhibit vocal learning. Vocal learning in bats has been demonstrated both in infants and in adults. Interestingly, in the primate species studied so far, vocal learning of species-specific calls has not been demonstrated.

In bats, studies of mother–pup interactions in *Carollia* show that the pup learns to match the sinusoidal FM call of the mother. Other forms of vocal learning in pups include adaptation or tuning of the resting frequency in echolocation pulses. Studies in horseshoe bats, *R. ferrumequinum*, showed that vocal development hinges on the perfection of the laryngonasal junction and consists of a shift from oral to nasal calls, low to higher frequencies, and noisy to pure tone type of calling. One of the major hurdles to parental care in bats that are highly social is that the pups grow up in crèches that consist of dozens to millions of young ones, many of which are born at approximately the same time. These pups mingle with each other while the parents are out foraging. To find and feed its young is therefore a difficult task for the mother. In evening bats, *Nycticeius humeralis*, it has been shown that the infants produce an isolation call whose distinctive pattern of frequency modulation can be recognized across the life span of the individual even though several other parameters may change with age. Moreover, these patterns are variable between individuals and repeatable within individuals. Thus, heritable signatures provide potential information about genetic relatedness in these species. This method of identification of young ones by mothers on the basis of vocal signatures has been demonstrated in birds, such as penguins, as well as dolphins.

Spear-nosed bats, *Phyllostomus hastatus*, produce different types of screech calls during foraging and for coordination of hunting movements, and these are highly group distinctive, clustering around a group average. When bats from different groups are housed together, they modify their screech calls to converge on a new unique screech call that is indicative of the new group identity. Thus, the structure of screech calls shifts following changes in the group composition. Since these signature calls result from socially mediated learning rather than from shared genes, they may serve to facilitate recognition of unrelated social group members. These group-distinctive calls can also vary significantly within different roosts of the same species even within a relatively confined geographic space. For example, distinct variations in the 'screech calls' of *P. hastatus* were obtained from three separate caves in Trinidad. Because the emergence of regional song dialects in birds is considered to be a result of vocal song learning, these data from *P. hastatus* suggest that differences on the group level may represent a microcosm of larger-scale geographic dialect patterns. Thus, signature calls are used not only to indicate identity of infants for promoting mother–infant interactions but also as group-distinctive calls to signify the identity of the social group.

Social Communication Behavior

From a behavioral perspective, acoustic social signals in bats are grouped into aggressive or warning calls, mating and song flight calls, distress or isolation calls, and directive or attraction calls. Individuals of the leaf-nosed bat, *Carollia perspicillata*, rely heavily on acoustic signals for social communication. Many acoustically distinct call types have been recorded from observations in a free-flying, captive colony of *Carollia*. FM signals that are produced by both females and infants have a communicative function during mother–infant interactions. The FM signals do not change much in this species as the infant matures. Whines, warbles, and trills are emitted in the context of male-to-female interactions within a harem. Whines, which are similar to the fixed rippled FM (fRFM) in mustached bats, are emitted as a male approaches a female during mating.

False vampire bats, *Megaderma lyra*, interact socially in various contexts and emit different types of communication sounds. Unlike some of the other species studied, many of the auditory communication behaviors observed in *Megaderma* occur in flight. One of these behaviors is labeled 'grumbling flight' and involves several bats hovering for a few seconds in a head-to-head formation emitting a series of short downward FM sounds that may terminate in a shallow downward FM. The spectral structure of this sound sequence is remarkably similar to the checked downward FM (cDFM) sounds observed in mustached bats. This may be preceded by an initial side-to-side flight by two or more bats in an excited state. Hovering behavior accompanied by calling is also exhibited during mating by males of *S. bilineata* and in *Carollia*. Reports on social interactions during flight are usually based on chance observations apart from honk calls, which are emitted when flying bats are about to collide.

Many of the Megachiropteran species that do not echolocate are nevertheless highly vocal and emit loud calls when disturbed as well as in other social contexts. Detailed studies of their social calls have been made in only a few species. For example, the Australian flying foxes (Pteropodidae) emit at least

22 different types of calls, rivaling the variety of calls emitted by primates. The distance calls emitted by *Pteropus* species are threat, alarm, or warning calls.

Singing and Babbling

The occurrence of singing and babbling in some species also attests to the specialized audiovocal abilities of bats. Like songbirds and some species of whales and dolphins, a few species of bats have been shown to sing and engage in courtship displays. Singing in bats was first reported by Vaughan in 1976 in *Cardioderma cor*, an African megaderma bat species. Singing is strongly seasonal in nature and probably establishes territoriality of the foraging bat. False vampire bats, *M. lyra*, engage in song flight behavior. This behavior is displayed only by the dominant males and can occur at any time of the year. It consists of a stereotyped flight pattern that is continuously accompanied by vocalization. The behavior is clearly one aspect of courtship and is directed only at nonlactating females within the group. In the only study describing this behavior, it was also elicited by the introduction of new females to the colony. The song flight in *Megaderma* consists of three stages that have been named introductory, advancing, and final flight. Spectrally distinct segments or strophes accompany each of these stages.

Singing is used to establish roosting territories in the sac-winged bat, *S. bilineata*, which employs an unusually large vocal repertoire. In this species, males emit tonal calls while interacting with females and other types of calls consisting of composites when actively defending their territories. Their songs consist of short repeated tones that do not appear to have any obvious context other than advertising the quality of singing by a male. The phenomenon of babbling has also been demonstrated in male and female pups of this species. As with infant babbling in humans, these pups produce renditions of all known adult vocalization types during bouts of vocalizations. These appear to be independent of a distinct social context and provide evidence of training to communicate effectively (Figure 4).

Motivational Basis of Call Structure

A study of social calls produced by 16 different European species suggests that calls commonly emitted by each of these species can be grouped into four major sonographic structures or types. Type A (long, narrowband NB) calls, also labeled as a 'squawk,' are generally agonistic or aggressive calls and may also function as a threat display. Type B calls (train of steep downward FM), such as a 'repeated trill' and a 'curved cheep,' express different levels of irritation. Type C (downward gliding FM) calls are characteristically used for interactions between females and pups. Pups continuously emit these isolation calls in the absence of their mothers, and the mothers use these calls to recognize their pups upon their return from foraging. Type D (humped, wrinkled, and sinusoidal FM) calls are involved in mate attraction behaviors and may also function during agonistic interactions between conspecifics. Both sexes of

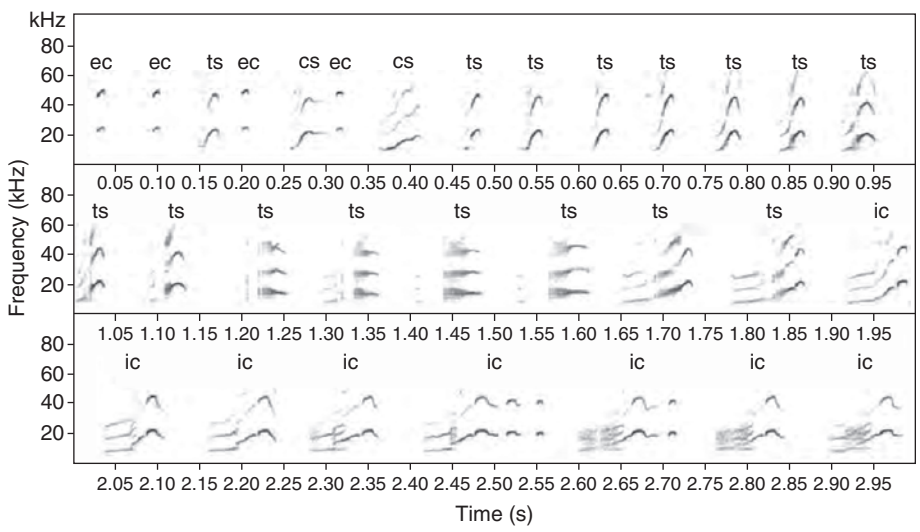

Figure 4 Sequence of spectrograms of babbling bouts from an individual of the sac-winged bat, *Saccopteryx bilineata*, depicting spectral frequency as a function of time: echolocation pulses (ec) and different elements of territorial song (ts), courtship song (cs), and isolation calls (ic) in a single babbling bout. Adapted from Knörnschild M, Behr O, and von Helversen O (2006) Babbling behavior in the sac-winged bat (*Saccopteryx bilineata*). *Naturwissenschaften* 93: 451–454, with permission.

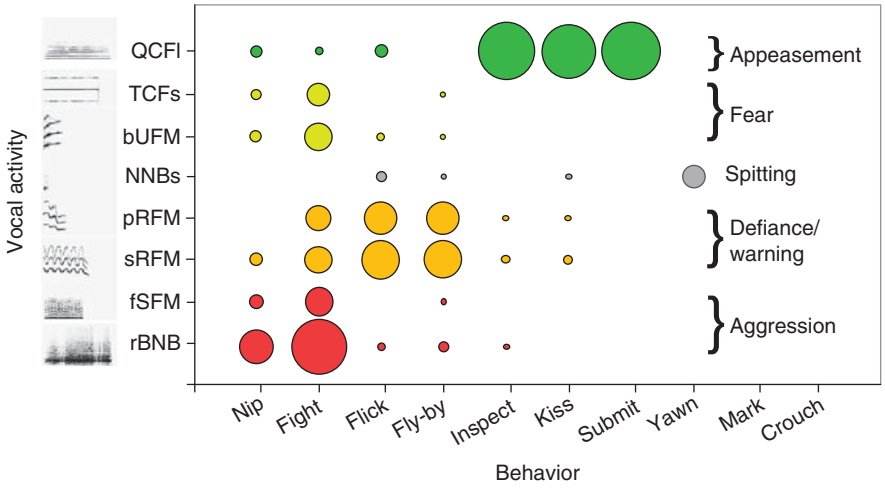

Figure 5 A bubble plot indicating the association of call types with behavioral states or emotions that influence group cohesion. Color coding is from red for aggression to green for appeasement. An act or state of appeasement impacts positively on the affiliation between two individuals and may be triggered either spontaneously or because of an impending uncertainty. A state of fear implies increased awareness to external stimuli in preparation for fight or flight and preempts a sensation of pain or stress. A state of defiance or to warn implies readiness to combat when confronted by another individual and may be triggered by either fear or insecurity; a state of aggression is one possible outcome of a state of defiance and warning. The size of the bubble is proportional to the normalized (percentage of total events) frequency of occurrence of a call type. The light gray circle indicates an 'organic' (spitting) sound that is sometimes associated with yawning but is not a vocalization. No calls were emitted during yawning, marking, and crouching. Fighting, which can be a complex and long-lasting behavior, produced the largest variety of call types. Adapted from Clement MJ, Dietz N, Gupta P, and Kanwal JS. (2006) Audiovocal communication and social behavior in mustached bats. In: Kanwal JS and Ehret G (eds.) *Behavior and Neurodynamics for Auditory Communication*, pp. 57–84. Cambridge, UK: Cambridge University Press, with permission.

M. lyra, an Old World microchiropteran species, use a low, multiharmonic 'grumble' as an aggressive call, and males use a mix of tonal CF and FM calls in a display for females. Male *S. bilineata* use harsh, broadband calls to threaten other males and direct tonal calls toward females.

Studies of the European species as well as mustached bats indicate that the acoustic structure of calls emitted by bats appears to follow the rules of the motivation–structure hypothesis for signal design as proposed by Morton in 1977 and August and Anderson in 1987 for various avian and mammalian species. Accordingly, harsh broadband calls in bats are widely used during aggression, although buzzes and trills also fill this role (Figure 5). Tonal calls in bats are typically used between mothers and pups, whereas more complex calls are used during mate attraction behavior.

Conclusion

Audiovocal communication in bat species is as advanced as that of any nonhuman mammal. In fact, speech in humans may owe its origin to selection pressures that shaped behavioral and neural adaptations for audiovocal communication in ancestors common to modern-day bats and humans.

See also: Communication Networks and Eavesdropping in Animals; Echolocation I: Behavior; Echolocation II: Neurophysiology; Signal Design Rules in Animal Communication.

Further Reading

August PV and Anderson JGT (1987) Mammal sounds and motivation–structural rules: A test to the hypothesis. *Journal of Mammalogy* 68: 1–9.

Balcombe JP (1990) Vocal recognition of pups by mother Mexican free-tailed bats: Do pups recognize their mothers? *Animal Behaviour* 39: 980–986.

Barclay RM, Fenton MB, and Thomas DW (1979) Social behavior of the little brown bat, *Myotis lucifugus*: II. Vocal communication. *Behavioral Ecology and Sociobiology* 6: 137–146.

Boughman JW (1998) Vocal learning by greater spear-nosed bats. *Proceedings of the Royal Society of London, Series B* 265: 227–233.

Clement MJ, Dietz N, Gupta P, and Kanwal JS (2006) Audiovocal communication and social behavior in mustached bats. In: Kanwal JS, and Ehret G (eds.) *Behavior and Neurodynamics for Auditory Communication*, pp. 57–84. Cambridge, UK: Cambridge University Press.

Davidson SM and Wilkinson GS (2002) Geographic and individual variation in vocalizations by male *Saccopteryx bilineata* (Chiroptera: Emballonuridae). *Journal of Mammalogy* 83: 526–535.

Eser KH and Schmidt U (1989) Mother–infant communication in the lesser spear-nosed bat, *Phyllostomus discolor* – Evidence for acoustic learning. *Ethology* 82: 156–168.

Fenton MB (1984) Echolocation: Implications for ecology and evolution of bats. *Quarterly Review of Biology* 59: 33–53.

Fenton MB (1985) *Communication in Chiroptera*. Bloomington: Indiana University Press.

French B and Lollar A (1998) Observations on the reproductive behavior of captive *Tadarida brasiliensis mexicana* (Chiroptera: Molossidae). *Southwest Nature* 43: 484–490.

Gould E (1971) Studies of maternal–infant communication and development of vocalizations in the bats *Myotis* and *Eptesicus*. *Communications in Behavioral Biology* 5: 263–313.

Habersetzer J (1981) Adaptive echolocation sounds in the bat *Rhinopoma-hardwickei* – a field-study. *Journal of Comparative Physiology* 144: 559–566.

Kanwal JS, Matsumura S, Ohlemiller K, and Suga N (1994) Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *Journal of the Acoustical Society of America* 96: 1229–1254.

Knörnschild M, Behr O, and von Helversen O (2006) Babbling behavior in the sac-winged bat (*Saccopteryx bilineata*). *Naturwissenschaften* 93: 451–454.

Ma J, Kobayasi K, and Metzner M (2007) Common themes in communication call design in five species of echolocating bats. Eighth International Congress of Neuroethology, Vancouver, Canada, Abstract no. PO328, p. 208.

Matsumura S (1981) Mother–infant communication in a horseshoe bat (*Rhinolophus ferrumequinum nippon*). Vocal communication in three-week-old infants. *Journal of Mammalogy* 62: 20–28.

Möhres FP (1967) Communicative characters of sonar signals in bats. In: Busnel RG (ed.) *Animal Sonar Systems Biology and Bionies,* vol. 2, pp. 939–945. NATO Advanced Study Institute, Joy-en-Josas: INRA-CNRS.

Morton ES (1977) On the occurrence and significance of motivation–structural rules in some bird and mammal sounds. *American Naturalist* 111: 855–869.

Nelson JE (1964) Vocal communication in Australian flying foxes (Pteropodidae; Megachiroptera). *Zeitschrift fur Tierpsychologie* 121: 857–870.

Pfalzer G and Kusch J (2003) Structure and variability of bat social calls: Implications for specificity and individual recognition. *Journal of Zoology* 261: 21–33.

Porter FL (1979) Social behavior in the leaf-nosed bat, *Carollia perspicillata*: II. Social communication. *Zeitschrift fur Tierpsychologie* 50: 1–8.

Thomas DW, Fenton MB, and Barclay RMR (1979) Social behavior of the little brown bat, *Myotis lucifugus*: I. Mating behavior. *Behavioral Ecology and Sociobiology* 6: 129–136.

Wilkinson GS and Boughman JW (1998) Social calls coordinate foraging in greater spear-nosed bats. *Animal Behaviour* 55: 337–350.

Auditory/Somatosensory Interactions

S E Shore, University of Michigan, Ann Arbor, MI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When stimulated with an auditory (sound) or somatosensory (touch) stimulus, distinct sensations are perceived that are modality specific (e.g., pitch and itch). However, sensory modalities are also able to influence each other through anatomical convergence of more than one modality onto single neurons. Such 'multisensory' neurons are evident throughout the nervous system, as first-order neurons in the brain stem, second-order neurons in the midbrain and higher-order neurons in the cerebral cortex (Figure 1). Multisensory neurons are usually excited by stimuli from more than one modality and can often 'integrate' this bimodal information by responding with a greater (or lesser) number of action potentials to combined stimulation than to either modality alone. The corresponding behavioral consequence of these cross-modal interactions is an increase in the probability of a correct response to an event accompanied by a decrease in reaction time for bimodal stimulation. In some cases, a neuron may respond only to one modality but its response may be suppressed or enhanced by the simultaneous addition of a stimulus from another modality. This form of integration has been demonstrated in the auditory cortex, inferior colliculus (IC), and cochlear nucleus (CN).

Somatosensory Influences on Auditory Cortex

Figure 1 is a schematic of the distribution of multisensory neurons in the rat sensory neocortex. It is evident from this figure that the majority of multisensory interactions occur at the borders between primary sensory regions, a feature that is attaining increasing support across several species. Many of the multisensory neurons in these border regions are capable of integrating information from each modality. The auditory cortex of primates and cats contains a core region of three primary areas surrounded by surrounding 'belt' and 'para-belt' regions of secondary areas (Figure 2). Neurons in the belt regions respond to both auditory and somatosensory stimuli.

Connections between Auditory and Somatosensory Cortex

Auditory cortex receives input from somatosensory cortex The anatomical basis for these responses is provided by tract tracing studies demonstrating that the belt regions receive inputs from somatosensory cortex as well as from auditory cortex core regions. The caudomedial belt region (CM) especially is densely connected with the retroinsular (RI) area of the somatosensory cortex, the superior temporal sulcus, and the posterior parietal and entorhinal cortex. Bimodal auditory and somatosensory responses have been reported in both CM and RI in response to electrical stimulation of the median nerve or mechanical stimulation of the upper body. Magnetic resonance imaging reveals an analogous belt region in humans, the planum temporale, which also shows evidence of auditory somatosensory convergence. Interestingly, there is sparse or no input to these belt regions from the medial geniculate complex, supporting the concept that the major functions of these belt areas are integrative.

Responses of bimodal neurons in auditory cortex Bimodal (or 'multisensory') neurons in auditory cortex typically are excited by stimulation of either modality and often integrate this information by producing responses that are substantially greater than the sum of the responses to each modality alone (Figure 3). However, in some cases, responses of auditory cortical neurons to sound can be suppressed by stimulation of somatosensory inputs to this area, even when the somatosensory stimulation by itself does not elicit a response (Figure 3). These subthreshold cross-modal interactions may arise through activation of inhibitory interneurons in the auditory cortex that receive somatosensory input, which, in itself, designates a neuron as 'multisensory.'

Neurons in the belt region of marmoset monkeys show a particular tendency to respond preferentially to the vocalizations of other monkeys, and do not respond to self-vocalizations. This preferential responsiveness to external vocalizations coupled with the suppression of self-vocalizations may be initiated at earlier stages in the auditory system, since neurons in the IC and even CN show similar tendencies (see below). The anatomical substrates for preservation of these patterns across levels is indeed present in the projection patterns that originate in the dorsal cochlear nucleus (DCN), terminate in the ICX, which in turn projects to the dorsal/posterior geniculate nucleus (MG). Whereas the primary (lemniscal) auditory pathway projects mainly to the core regions of auditory cortex (A1) via the ventral MG, projections to the belt and parabelt regions arise largely from the dorsal/posterior MG, and also other thalamic regions that demonstrate bimodal responses to auditory and

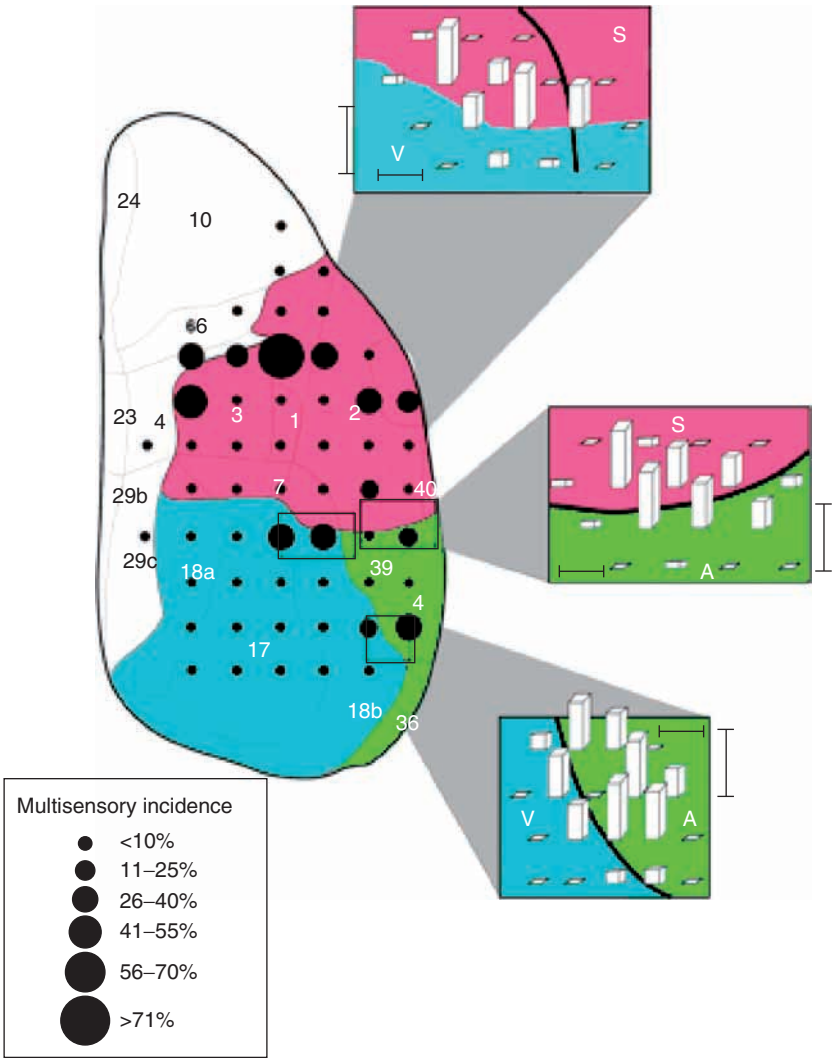

Figure 1 The distribution of multisensory neurons in rat sensory neocortex. The line drawing depicts the dorsal surface of cortex. Numbers and solid lines designate major subdivisions (17) (parietal, red shading; temporal, green shading; and occipital, blue shading). Filled circles show locations of electrode penetrations in a coarse-grain analysis that was conducted in 22 animals, and circle size indicates the relative incidence of multisensory neurons at each site. Insets show the results of higher-resolution sampling through each of the transitional regions that was conducted in a total of nine animals. Bar height indicates the relative incidence of multisensory neurons. Horizontal scale bar = 250 μm, and vertical scale bar = 50% multisensory incidence. V, visual cortex; A, auditory cortex; S, somatosensory cortex. Reproduced from Wallace MT, Ramachandran R, and Stein BE (2004) A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Sciences of the United States of America* 101: 2167–2172. Copyright (2007) National Academy of Sciences, USA.

somatosensory stimuli, such as the Po, Sg, Lim, and PM. It is paradoxical that these latter areas also project to A1, in which there is no evidence for bimodal integration. This suggests that the bimodal responses observed in CM are either solely the result of inputs from the RI somatosensory area, or co-incident inputs from several of these regions. CM sends projections to the IC, terminating primarily in ventrolateral regions including the external cortex, which, itself, demonstrates bimodal properties (see below).

Activity-dependent plasticity in cortical auditory–somatosensory interactions Determination of the size of a cortical sensory field may be genetically determined depending on the environmental demands. For example, the duck billed platypus has a large portion of its cortex assigned to somatic inputs from the bill, while squirrels devote much of the cortex to visual processing. However, recent studies suggest that activity from sensory receptors can also play a role in the determination of cortical fields and their connections. In congenitally deaf mice, the auditory

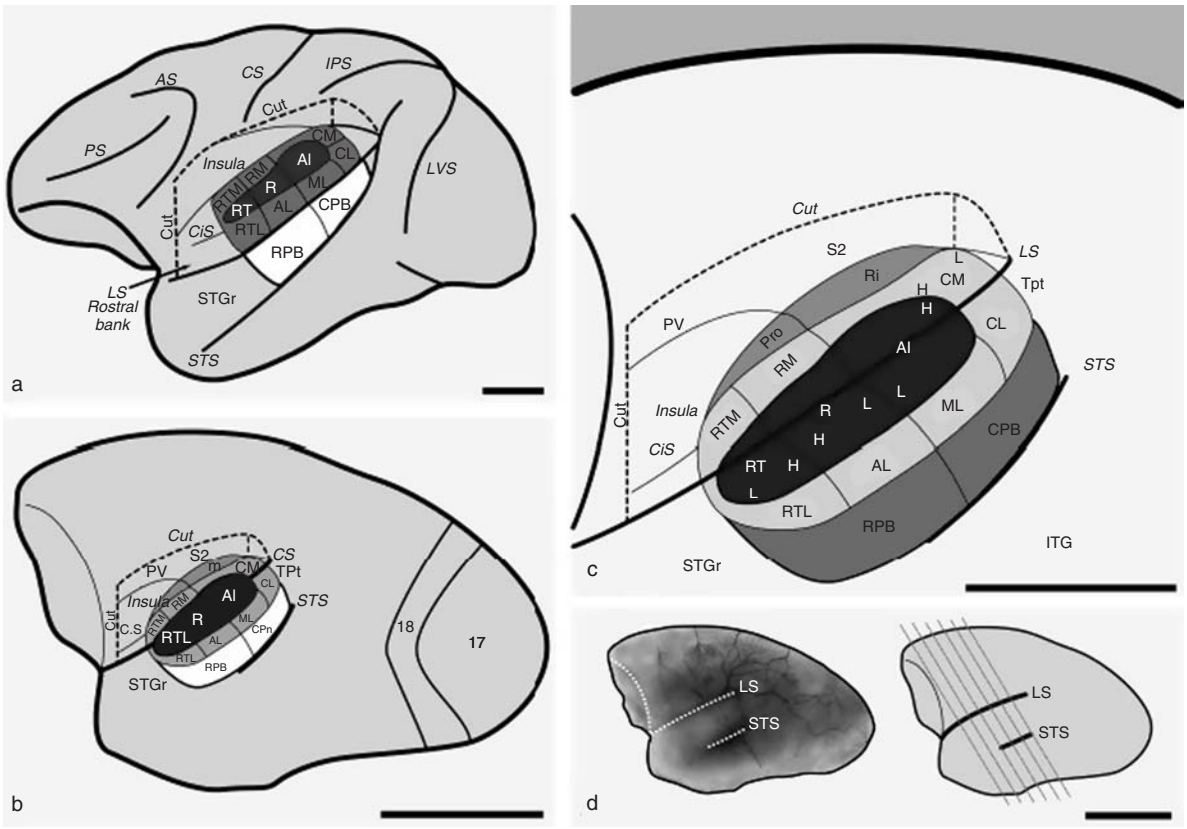

Figure 2 Schematic models of macaque (a) and marmoset (b, c) monkey auditory cortex. The lateral sulcus (LS) of the left hemisphere was graphically opened (cut) to reveal the locations of auditory cortical areas on its lower bank. The circular sulcus (CiS) was flattened to show the position of the rostromedial (RM) and rostrotemporal medial (RTM) areas that occupy its lateral wall. The upper bank of the LS was partly opened (cut) to show the locations of the retroinsular area (Ri) in the fundus, second (S2) and parietoventral (PV) somatosensory areas on the upper bank, and insula. The three areas that comprise the core region of the auditory cortex (dark shading) are located on the lower bank (A1, auditory area 1; R, rostral; RT, rostrotemporal). The core is surrounded by seven or eight areas that belong to the belt region (light shading) (CM, caudomedial; CL, caudolateral; ML, middle lateral; RM, rostromedial; AL, anterolateral; RTM, rostrotemporal medial; RTL, rostrotemporal lateral). The proisocortex area (Pro) is a putative addition to the medial belt. The core and lateral belt regions are mostly contained within the lateral sulcus in macaques but extend onto the superior temporal gyrus (STG) in the marmoset. On the surface of the STG are two areas that make up the parabelt region (medium shading; RPB and CPB, rostral and caudal parabelt). The rostral part of the STG (STGr) extends to the temporal pole. The temporal parietotemporal area (Tpt) occupies the caudal end of the STG and extends onto the supratemporal plane within the LS. Tonotopic gradients within areas are indicated by H (high frequency) and L (low frequency). Other sulci shown include the arcuate sulcus (AS), central sulcus (CS), intraparietal sulcus (IPS) inferior temporal gyrus (ITG), principal sulcus and superior temporal sulcus (STS). (d) Photographic image of the marmoset left hemisphere and schematic showing the plane of section (diagonal lines) used in the present study for histological processing. Scale bar = 10 mm (a, b, d); 5 mm (c). From de la Mothe LA, Blumell S, Kajikawa Y, and Hackett TA (2006) Cortical connections of the auditory cortex in marmoset monkeys: Core and medial belt regions. *Journal of Comparative Neurology* 496: 27–71.

cortical area that normally contains unimodal auditory neurons is replaced largely by neurons responsive to somatosensory stimuli and to a lesser extent by neurons responsive to visual or combined visual–somatosensory stimulation. Furthermore, most of the receptive fields of these neurons represent the face as opposed to the limbs, trunk, or tail. This extensive multisensory plasticity, which affects the deprived cortical area, as well as the other sensory cortices, is evident also in humans through the use of functional magnetic resonance imaging (FMRI) studies. In congenitally deaf mice, cortical areas that would normally be responsive to auditory stimulation are responsive to both somatosensory and visual stimulation. Mechanisms underlying these cortical alterations may be partly determined by alterations in subcortical connections observed in deaf animals.

Somatosensory Influences on IC

The IC is a major synaptic station of the auditory midbrain. Almost all ascending and descending

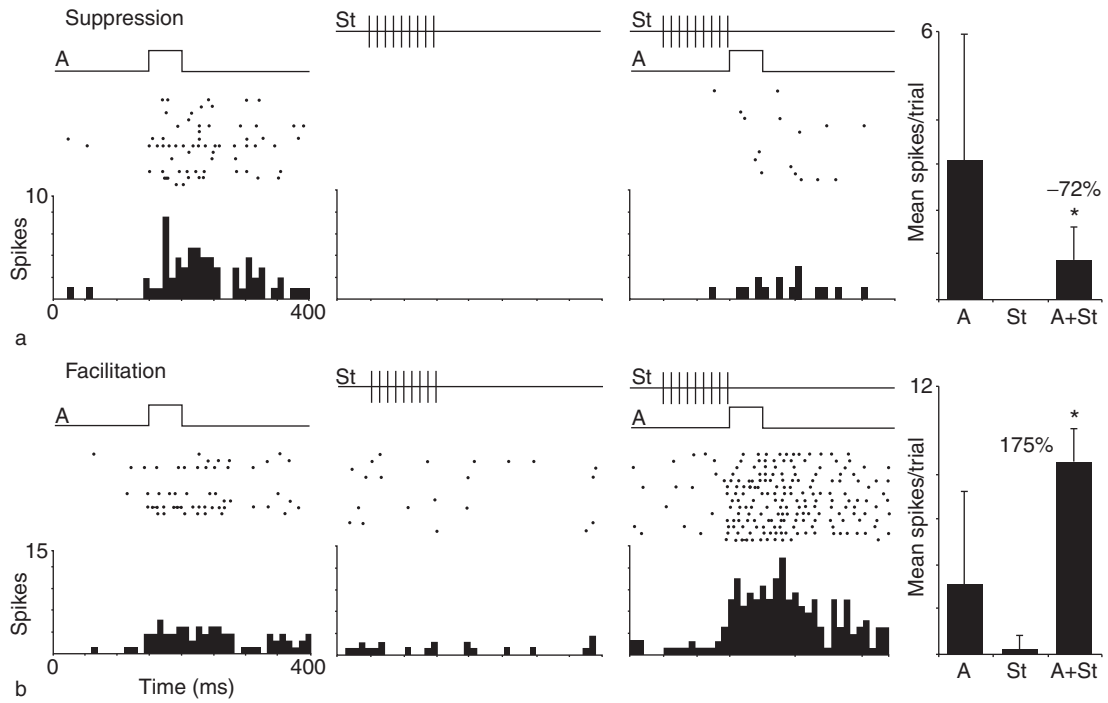

Figure 3 Subthreshold effects of SIV activation on auditory responses of FAES neurons. (a) (left-most panel), recordings from an FAES neuron (each raster dot = 1 spike; each raster row = 1 presentation; histogram = activity summary, 10 ms bin) show that an auditory stimulus (white noise denoted by the square-wave, 55 dB SPL, 50 ms duration) evoked a modest response on most presentations. For the same neuron, when an indwelling SIV electrode was activated (600 μA, 0.1 ms duration, 100 Hz for 100 ms, pulses labeled 'St'), no neuronal activity was recorded. However, when the two stimulation elements were combined ('St + A'), the response to the auditory stimulus was substantially diminished. The bar histogram to the far right summarizes these results, showing that combined SIV and auditory stimulation significantly (*$P < 0.05$, paired *t*-test) reduced the response to the auditory stimulus alone by 72%. These same conventions are used in (b), where the auditory stimulus alone (far left) elicited a modest response, but the SIV stimulus did not affect the neuron. However, when the SIV and auditory stimuli were combined, there was a vigorous and reliable response. The 'bar histogram' summarizes these results and indicates that combined SIV-auditory stimulation produced a large (175%) and significant facilitation of neuronal activity. From Meredith MA, Keniston LR, Dehner LR, and Clemo HR (2006) Crossmodal projections from somatosensory area SIV to the auditory field of the anterior ectosylvian sulcus (FAES) in cat: Further evidence for subthreshold forms of multisensory processing. *Experimental Brain Research* 172(4): 472–484.

auditory pathways synapse here. The IC is comprised of three subdivisions: the central nucleus (CIC), surrounded by a dorsal cortex (DCIC) and a laterally located external cortex (ICX). The CIC forms part of the 'primary' or 'lemniscal' auditory pathway that originates in the auditory nerve and terminates in primary auditory cortex. It receives input from the majority of auditory ascending projections and displays sharply tuned, low-threshold neuronal properties. The DCIC and ICX form part of the 'nonclassical' auditory pathway that receives not only auditory inputs, but also nonauditory projections, including somatosensory, and descending inputs from the auditory cortex and CIC. In contrast to the sharply tuned neuronal responses in CIC, neurons in the ICX are broadly tuned and respond preferentially to spectrally complex stimuli such as vocalizations. ICX neurons are further distinguished by their responsiveness to multimodal stimulation.

Connections between IC and Somatosensory Nuclei

The anatomical bases for responses to multimodal stimulation are projections that originate in the spinal trigeminal nuclei (Sp5), the cuneate nuclei, and the medullary reticular formation, and terminate primarily in the ICX.

Inputs from the trigeminal nuclei The dorsomedial and marginal areas of Sp5I (interpolaris), and the deep and marginal layers of Sp5C (caudalis), project directly to the ventral ICX (ICXV). In addition to their involvement in the modulation of pain, these subdivisions of the Sp5 receive low-threshold nonnociceptive afferents from head and face, such as those sensitive to gentle pressure and vibrissa deflection. These regions also receive inputs from vocal tract/intra oral structures including the temporomandibular joint and tongue muscles. Furthermore,

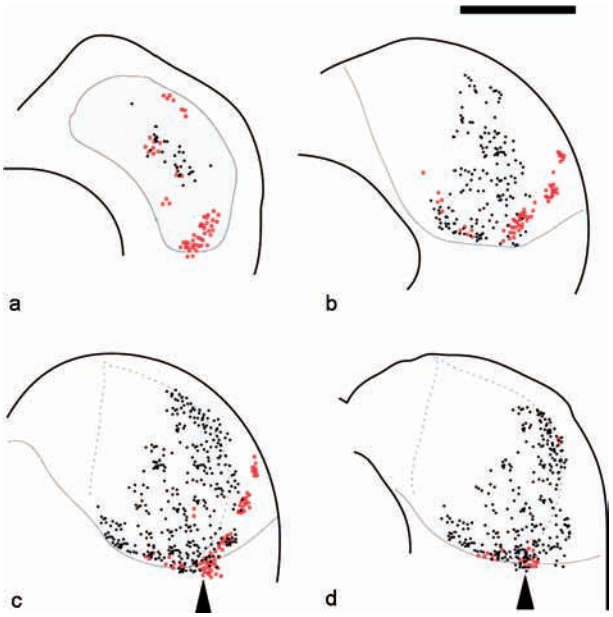

Figure 4 Sp5 terminations in ICX overlap with those from the CN. (a–d) Rostral to caudal sequence. Red squares represent the Sp5 terminations and black dots represent the CN terminations. Most of the convergence was observed in the ICXV (triangles in c, d) and some convergence in the rostral ICX. Each dot/square represents 1–3 terminal endings. Scale bar = 2 mm (b) (applies to (a–d)). From Zhou J and Shore S (2006) Convergence of spinal trigeminal and cochlear nucleus projections in the inferior colliculus of the guinea pig. *Journal of Comparative Neurology* 495: 100–112.

the terminals of these projections overlap with direct inputs to the ICX from the DCN (Figure 4). As a region receiving auditory inputs from ascending auditory and somatosensory pathways, it is ideally placed as the primary IC region for processing multisensory information.

Bimodal Responses of Neurons in ICX

Neurons in the ICX of guinea pigs respond to both acoustic and tactile stimuli. In some ICX neurons, electrical stimulation of the Sp5 can modulate (usually suppress) sound-evoked responses, even when Sp5 stimulation alone does not evoke responses. This subthreshold suppression of sound-evoked activity by somatosensory stimulation is similar to that described above in the ectosylvian cortex and the DCN (see below), the latter projecting to the ICX. Furthermore, recordings from the ICX in awake monkeys demonstrate that these neurons distinguish between the vocalizations of other monkeys and self-vocalizations: external vocalizations produce vigorous neuronal responses, while self-produced vocalizations suppress neuronal responses to spontaneous rate or below (Figure 5). These studies suggest that bimodal neurons in the ICX may play a role in the suppression of self-generated sounds. Anatomical support for this hypothesis can be obtained from findings that the regions of Sp5I and Sp5C that project to ICX receive somatosensory inputs from vocal structures. Thus, sounds associated with chewing, vocalization, and respiration may be suppressed by the concurrent activity from the somatosensory system, which likely originates in the proprioceptive afferents or broad dynamic range mechanoceptors in deep tissues during these activities.

Somatosensory Influences on the CN

The CN is the obligatory nucleus in which all ascending information from the cochlea forms synaptic connections with the auditory brain. Once considered a simple 'relay nucleus,' the CN performs complex computations to transform the incoming signals from the auditory nerve. The well-established somatosensory influences on higher auditory structures have only recently been extended to include the CN where it is now evident that somatosensory and auditory inputs converge. Neurons both in the ventral and dorsal CN can be modified by activating somatosensory neurons.

Connections between CN and Somatosensory Nuclei

Inputs to the CN from somatosensory structures originate in the trigeminal nerve, trigeminal nucleus, dorsal column, dorsal column nuclei, and indirectly from the medullary reticular formation that receives input from the trigeminal nerve (Figure 6).

Inputs from the trigeminal system The trigeminal nerve conveys sensory information from the head and face to the central trigeminal sensory nuclear complex via the ophthalmic, maxillary, and mandibular branches of the trigeminal nerve. The ophthalmic nerve innervates the forehead, upper eyelid, and extra ocular muscles; the maxillary nerve supplies the upper lip, the lower eyelid, and the upper jaw and roof of the mouth; and the mandibular nerve innervates the lower lip, the mucous membranes of the lower jaw, the floor of the mouth, and the anterior two-thirds of the tongue. Trigeminal neurons that project directly to the CN are located in the medial portion of the trigeminal ganglion (TG) at the origin of the ophthalmic nerve, as well as in the mandibular division of the ganglion. The axons of TG projection neurons are thin (~1 μm) and typically form en passant boutons that are concentrated in the ipsilateral marginal cell areas that contain small cells and granule cells. Some terminals are located around cell bodies in the magnocellular regions of ventral

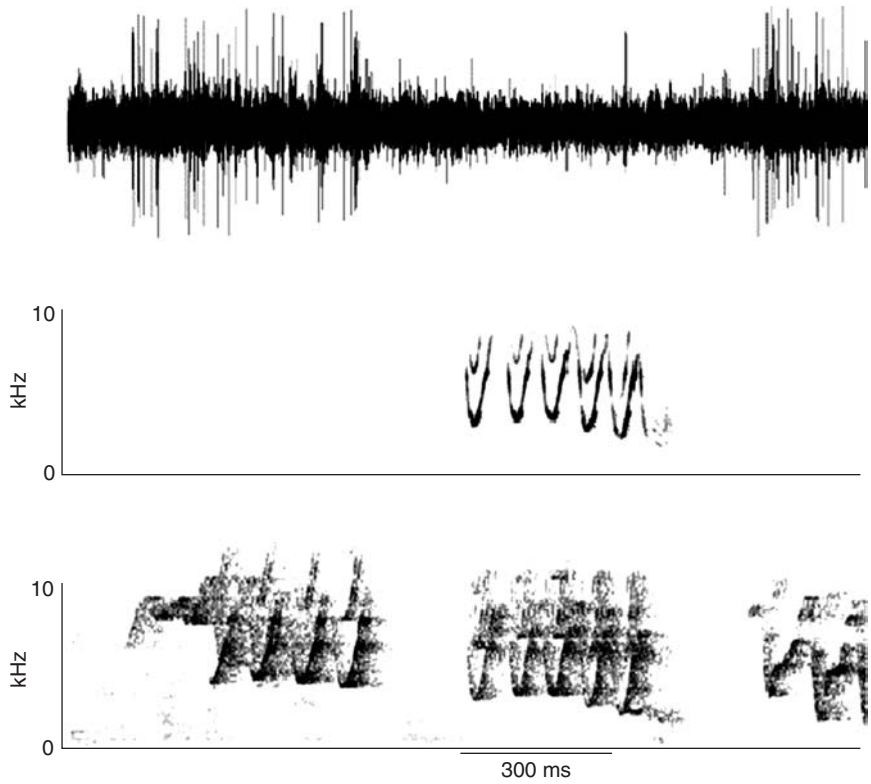

Figure 5 Neuronal activity in the external nucleus of the IC during self-produced vocalization and perception of vocalizations produced by other animals. Upper trace: extracellular recording. Middle trace: vocalization of the experimental animal in the form of a sonagram. The signal was picked up by the piezo-electric skull vibration sensor and transmitted telemetrically. Lower trace: sonagrams of the vocalizations recorded with the room microphone. Vocalization 2 stems from the experimental animal, vocalizations 1 and 3 from group mates. Reproduced from Tammer R, Ehrenreich L, and Jurgens U (2004) Telemetrically recorded neuronal activity in the inferior colliculus and bordering tegmentum during vocal communication in squirrel monkeys (*Saimiri sciureus*). *Behavioural Brain Research* 151: 331–336, with permission from Elsevier.

cochlear nucleus (VCN). Electron microscopic evaluation of these endings reveals round synaptic vesicles and asymmetric synaptic specializations indicative of excitatory synapses.

The spinal trigeminal nucleus (Sp5) is comprised of three nuclei: pars oralis (Sp5O), pars interpolaris (Sp5I), and pars caudalis (Sp5C). Each subdivision receives nociceptive and nonnociceptive afferents from the head and face that are sensitive to pressure and vibrissa deflection. The Sp5 also receives proprioceptive inputs from vocal tract/intra oral structures including the temporo-mandibular joint and tongue muscles. Projections from Sp5 to the CN originate primarily in Sp5I and Sp5C (see Figure 6). The paucity of projection cells in the subnucleus gelatinosus, which receives nociceptive afferents, suggests that the neurons in Sp5 that project to the auditory system convey primarily vocal structure mechanosensory information, and not pain information. Fibers and terminal endings of the Sp5 projection are small or medium in size with *en passant* or large, irregular terminal swellings. Postsynaptic targets include the dendrites of granule cells in the marginal regions of VCN and principal cells in the molecular and deep layers of DCN and magnocellular regions of VCN. As observed for TG projections to the CN, electron microscopic evaluation reveals round synaptic vesicles and asymmetric synaptic specializations indicative of excitatory synapses.

Projections from other somatosensory structures to the CN Somatosensory innervation to the CN also originates in the dorsal root ganglion of the spinal cord and dorsal column nuclei (cuneate and gracile) of the brain stem, which receive proprioceptive afferents from head, trunk, and limbs. These projection cells project also to the granular cell domain of the CN, terminating on the dendrites of granule and other small cells. Electron microscope examination indicates that these terminals are excitatory, containing round

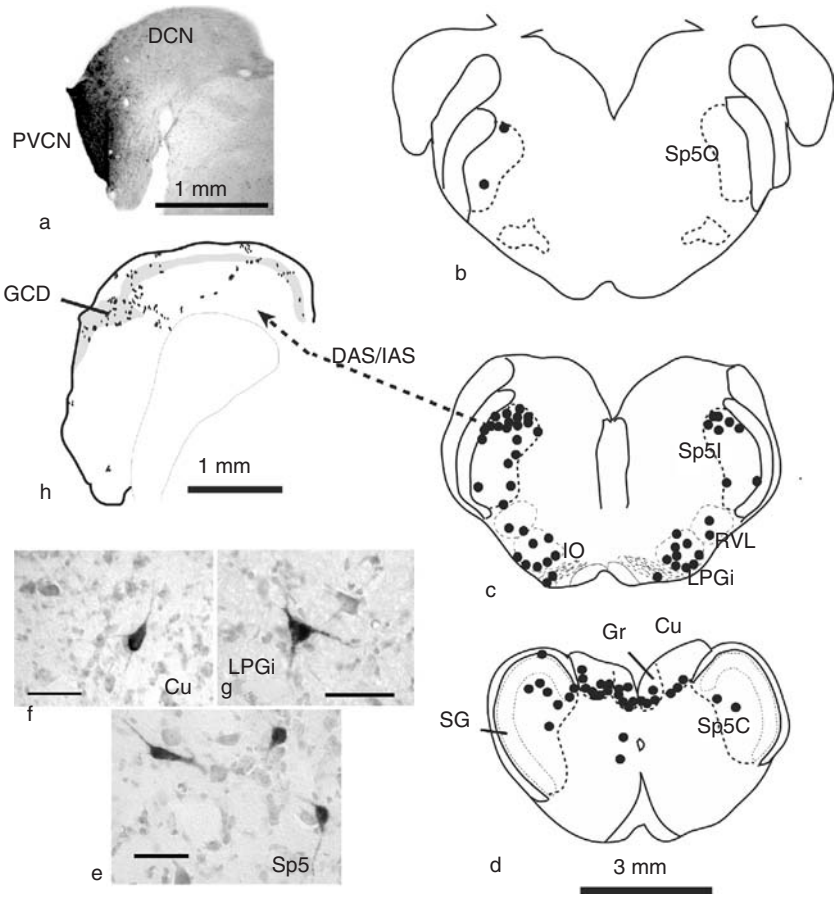

Figure 6 Projections from brain stem somatosensory regions to the CN in the guinea pig. (a)–(g) Retrograde labeling in the brain stem after an injection of biotinylated dextran amine (BDA) into the CN. (a) Photomicrograph of the injection site. The injection site is restricted to the granule cell domain of the PVCN. (b)–(d) Drawings of 1 mm transverse sections across the medulla. Each dot represents one labeled cell. The labeled neurons are located primarily in the Sp5I and Sp5C. Few labeled cells are located in the SG (d). Labeled neurons are also seen in the medullary reticular formation (RVL and LPGi, (c)), inferior olive (IO, (c)), and dorsal column nuclei (Gr and Cu, (d)). Projection neurons in Sp5 have polygonal or elongated somata (e). Projection neurons in dorsal column nuclei and reticular formation are multipolar (f and g). (h) Terminal labeling in the CN after placement of an anterograde tracer into Sp5I. Most Sp5 fibers enter the CN via DAS/IAS and terminate primarily in the granule cell domain (gray area), but also in deep DCN. Each dot represents one to three labeled terminal endings. Scale bars = 25 μm (e–g). CN, cochlear nucleus; Cu, cuneate nucleus; DAS, dorsal acoustic striae; DCN, dorsal cochlear nucleus; GCD, granule cell domain; Gr, gracile nucleus; IAS, intermediate acoustic striae; IO, inferior olive; LPGi, lateral paragigantocellular reticular nucleus; PVCN, posteroventral cochlear nucleus; RVL, rostral ventrolateral reticular formation; SG, subnucleus gelatinosus; Sp5, spinal trigeminal nucleus; Sp5C, pars caudalis of Sp5; Sp5I, pars interpolaris of Sp5; Sp5O, pars oralis of Sp5. Reprinted from Shore SE and Zhou J (2006) Somatosensory influence on the cochlear nucleus and beyond. *Hearing Research* 216–217: 90–99, with permission from Elsevier.

synaptic vesicles and forming asymmetric synapses. Some of the proprioceptive information conveyed by these systems, such as pinna and head position, might provide important spectral cues for sound localization. In conjunction, neurons that innervate the muscle spindles of the extra ocular muscles are located in the ophthalmic division of TG, which also projects to the CN. It is thus likely that trigeminal inputs to the CN provide proprioceptive information related to eye-to-head orientation while dorsal column inputs provide head-to-space and eye-to-head orientation, for the purpose of locating sound sources important for prey tracking or recognizing danger.

Responses of CN Neurons to Stimulation of Somatosensory Pathways

Some of the terminations from the TG that have characteristics of excitatory synapses end on somata in the magnocellular regions of VCN. Thus, it is not unexpected that TG stimulation primarily excites cells in the VCN. However, granule cells in the marginal layers of the CN are the major recipients of somatosensory terminations. The glutamatergic parallel fiber axons of granule cells throughout the CN terminate on the apical dendrites of the principal output neurons, fusiform cells as well as inhibitory

interneurons, the cartwheel and superficial stellate cells. Thus, activation of granule cells by somatosensory stimulation could excite principal cells directly through their apical dendrites or inhibit them via cartwheel or stellate cells. In addition, direct activation of cells in the deep DCN may occur. Likely candidate cells for this activation are vertical cells or giant cells.

Responses of CN neurons to somatosensory stimulation in the absence of sound Stimulation of the dorsal column nuclei and manual manipulation of the pinna produce a complex pattern of inhibition and excitation of DCN neurons. In the cat, stimulation of pinna regions evokes stronger responses in DCN than stimulation of other areas of the face, leading to the suggestion that dorsal column nuclear input to the DCN may be involved in sound localization, which in the cat is aided by the mobility of the pinna. DCN neurons can be affected in a similar manner by stimulation of somatosensory nuclei innervating the vibrissae and peripheral nerves innervating the neck, perhaps providing feedback regarding head movements or position as an aid to sound localization.

Stimulation of the TG, Sp5 or reticular formation (RF) produces similar complex patterns of inhibition and excitation in DCN neurons. Units responding to trigeminal stimulation show primarily pauser-buildup, buildup or chopper responses to best frequency (BF) toneburst stimulation, responses that are associated with fusiform cells and giant cells in the deeper layers of DCN. The latencies of inhibitory responses are consistently longer (by >2 ms) than excitatory responses, consistent with the inhibition being mediated through an additional interneuron such as the cartwheel or superficial stellate cell.

Multisensory integration in the CN: responses to combined somatosensory and acoustic stimulation When exposed to bimodal stimulation of the auditory and the trigeminal nerves, DCN neurons are capable of multisensory integration as described above for cortical and IC neurons. More than half of DCN output neurons demonstrate multisensory integration, of which about two-thirds are suppressive and about one-third enhancing. Figure 7 shows PST histograms from one DCN unit to broadband noise (BBN) stimulation (1) and combined trigeminal-BBN stimulation (2). A strong suppression of the firing rate to the noise burst occurs when it is preceded by trigeminal stimulation (2).

Significance of bimodal integration in the CN and beyond The suppression of responses to sound by activating the trigeminal system provides evidence that the DCN may be involved, not only in sound localization, but also may act as an adaptive filter to reduce neuronal responses to body-generated sounds. The stimulus paradigm illustrated in Figure 7 simulates a 'natural' condition in which the CN would receive overlapping auditory and somatosensory information, as would occur in chewing, respiration, and self-vocalization. The findings that trigeminal stimulation can strongly suppress sound-evoked

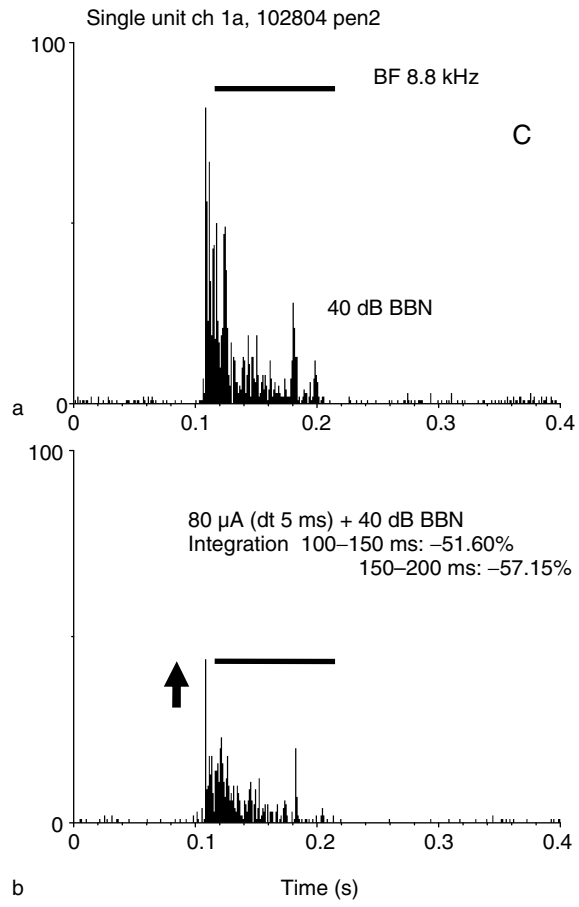

Figure 7 TG stimulation can suppress responses to sound in single units. (a) Unimodal stimulation. Poststimulus time histogram of responses of an isolated single unit to a broadband noise (BBN) stimulus (40 dB SPL, 100 ms). (b) Bimodal stimulation. Poststimulus time histogram of responses of the same single unit to the BBN noise stimulus preceded by electrical stimulation of the TG (onset 5 ms preceding BBN, 80 μA, 100 μs/phase). Arrow indicates onset of electrical stimulation at 95 ms; solid bar indicates 100 ms duration of BBN, 200 presentations. Bin width 0.5 ms. Multisensory integration to the bimodal stimulus is calculated for times 100–150 ms or 150–200 ms. Suppression of more than 50% occurs for both measures in this unit indicating maximal trigeminal suppression of activity to the BBN at the beginning of the response and continuing throughout its duration. Bin width 1 ms, 100 repetitions. From Shore S (2005) Multisensory integration in the dorsal cochlear nucleus: Unit responses to acoustic and trigeminal ganglion stimulation. *European Journal of Neuroscience* 21: 3334–3348.

activity, and even modify the temporal firing pattern evoked by a noise stimulus after the addition of a trigeminal stimulus, are consistent with the behavior of other cerebellar-like systems with granule cell – parallel fiber circuits that subtract predictable stimulation produced by the animal's own movements. For example, in the electrosensory lateral line of weakly electric fish, electric fields generated by the animal's own respiration are 'subtracted out' in principal cells by motor and proprioceptive information carried via the parallel pathways analogous to those seen in the mammalian DCN. Thus, an additional function of the parallel fiber inputs to the DCN, arising in the trigeminal system, may be to suppress internally generated sounds produced by chewing, respiration, and also self-vocalization. These combined acoustic-somatosensory activations would occur on an ongoing basis and therefore represent expected signals, as contrasted to environmentally generated or unexpected signals. The bimodal integration, in which trigeminal input enhances the responses of both VCN and DCN units to noise stimuli may be important in improving signal-to-noise ratios when attention is directed to a particular location. Thus, the CN could act as an 'adaptive filter' to suppress self-generated sounds, and also enhance perception of behaviorally relevant sounds, such as the vocalizations of other animals, generated externally, and therefore not combined with internal somatosensory stimulation.

These findings add to the emerging view that the 'merging of the senses' begins at early stages of sensory processing and not, as previously thought, at later stages of cortical hierarchical schemes. Early studies of multisensory integration in the superior colliculus add to the view that multisensory integration serves the purpose of enhancing the detection of behaviorally relevant stimuli and speeding up reaction time, orientation, and accuracy of responses to these stimuli.

See also: Auditory Cortex Structure and Circuitry; Auditory System: Central Pathways; Canal–Otolith Interactions; Cochlear Development; Somatosensory Perception; Somatosensory Cortex; Visual–Vestibular Interactions.

Further Reading

de la Mothe LA, Blumell S, Kajikawa Y, and Hackett TA (2006) Cortical connections of the auditory cortex in marmoset monkeys: Core and medial belt regions. *Journal of Comparative Neurology* 496: 27–71.

Jain R and Shore S (2006) External inferior colliculus integrates trigeminal and acoustic information: Unit responses to trigeminal nucleus and acoustic stimulation in the guinea pig. *Neuroscience Letters* 395: 71–75.

Meredith MA, Keniston LR, Dehner LR, and Clemo HR (2006) Crossmodal projections from somatosensory area SIV to the auditory field of the anterior ectosylvian sulcus (FAES) in cat: Further evidence for subthreshold forms of multisensory processing. *Experimental Brain Research*. 172(4): 472–484.

Shore S (2005) Multisensory integration in the dorsal cochlear nucleus: Unit responses to acoustic and trigeminal ganglion stimulation. *European Journal of Neuroscience* 21: 3334–3348.

Shore SE and Zhou J (2006) Somatosensory influence on the cochlear nucleus and beyond. *Hearing Research* 216–217: 90–99.

Stein B and Meredith M (1993) *The Merging of the Senses*. Cambridge, MA: MIT Press.

Tammer R, Ehrenreich L, and Jurgens U (2004) Telemetrically recorded neuronal activity in the inferior colliculus and bordering tegmentum during vocal communication in squirrel monkeys (*Saimiri sciureus*). *Behavioural Brain Research* 151: 331–336.

Wallace MT, Ramachandran R, and Stein BE (2004) A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Sciences of the United States of America* 101: 2167–2172.

Zhou J and Shore S (2004) Projections from the trigeminal nuclear complex to the cochlear nuclei: A retrograde and anterograde tracing study in the guinea pig. *Journal of Neuroscience Research* 78: 901–907.

Zhou J and Shore S (2006) Convergence of spinal trigeminal and cochlear nucleus projections in the inferior colliculus of the guinea pig. *Journal of Comparative Neurology* 495: 100–112.

Auditory Cortex Structure and Circuitry

D Barbour, Washington University in St. Louis, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The auditory system comprises neuronal circuits responsible for evaluating distant objects through mechanical (acoustic) vibrations transmitted to the observer via air or water. Acoustic vibrations are converted into neuronal signals by the ear, and these signals undergo multiple processing stages so that the organism can infer relevant information about the vibrations' source. Auditory cortex (ACx) of mammals represents the highest level of the conventionally described auditory system and one of the final stages of auditory processing before acoustic information has been sufficiently decoded to achieve cognitive awareness or to be acted upon in complex fashion. Many subcortical circuits extract various features from acoustic signals prior to ACx processing. These circuits have received the bulk of auditory research attention and consequently are better understood than are cortical circuits.

ACx represents the region of mammalian neocortex dedicated predominantly to the processing of auditory information and is located on the lateral surface of the brain, typically lateral or anterolateral to primary somatosensory cortex and anterior to primary visual cortex. Generally, ACx occupies a prominent position on the lateral surface of the telencephalon, and in primates (including humans) lies on the temporal lobe just inferior to and within the lateral sulcus (see Figure 1). In heavily convoluted primate brains, such as that of humans, much of ACx lies within the operculum of the lateral sulcus and is thus not visible from the external surface of the brain. Neuroanatomical studies indicate that ACx comprises several distinct areas that vary by species and that have individual functions that are still rather poorly understood. Organization of ACx demonstrates considerably more variation among species than does that of subcortical auditory nuclei.

Neuronal projections to the auditory cortex arise largely from the thalamus and other cortical areas, although inputs from other structures have been demonstrated. The various divisions of the medial geniculate body (MGB) of the thalamus represent ACx input arising directly from the ascending auditory pathway. These forward projections and (generally) reciprocal connections among distinct cortical areas define the region collectively referred to as ACx. Many of these distinct areas have unclear functions and await further physiological investigation to determine their contributions to auditory processing. Projections from ACx target other cortical areas, such as the frontal lobe, as well as each of the subcortical auditory nuclei.

Organization of Auditory Cortex

Thalamic Input

Distinct cortical areas are generally identified by architectonic differences and confirmed by neuroanatomical tracing studies and neurophysiological properties. Nonspecialized mammals have 10–15 distinct areas identifiable as ACx by their connections. These areas are typically subdivided into groups based upon their proximity and the types of subcortical projections they receive. One major source of thalamic input to ACx arises from the ventral division of the medial geniculate body (MGv), often referred to as the lemniscal pathway. This pathway appears to transmit strictly auditory information at short latencies using projection neurons having small receptive field sizes and arranged topographically in the thalamus. These neurons receive the bulk of their ascending input from topographically arranged neurons in the central nucleus of the inferior colliculus (ICc). Of the distinct thalamocortical pathways, the ventral pathway projects to fewer cortical areas and connects to these areas with relatively patchy distributions predominantly in cortical laminae 3b and 4.

Another major source of thalamic input to ACx arises from the dorsal division of the medial geniculate body (MGd), often referred to as one component of the paralemniscal or extralemniscal pathway. In some species, MGd has been shown to comprise multiple nuclei, each with potentially distinct cortical targets. The dorsal pathway contrasts with the ventral pathway in most if not all anatomical and physiological features. Compared with the ventral pathway, the dorsal pathway receives different, nontonotopic inputs from the inferior colliculus, projects largely to distinct areas of cortex, and sends less dense axonal collaterals more widely throughout most cortical laminae, although the largest projections are also to layers 3b and 4. The neurons composing this pathway exhibit less topographic organization, broader tuning, longer and more variable response latencies, and tendencies to habituate to ongoing stimuli. The projection patterns of the individual nuclei of the dorsal division appear to be more diverse than those of the ventral division.

The medial division of the medial geniculate body (MGm) comprises diverse cell types that represent

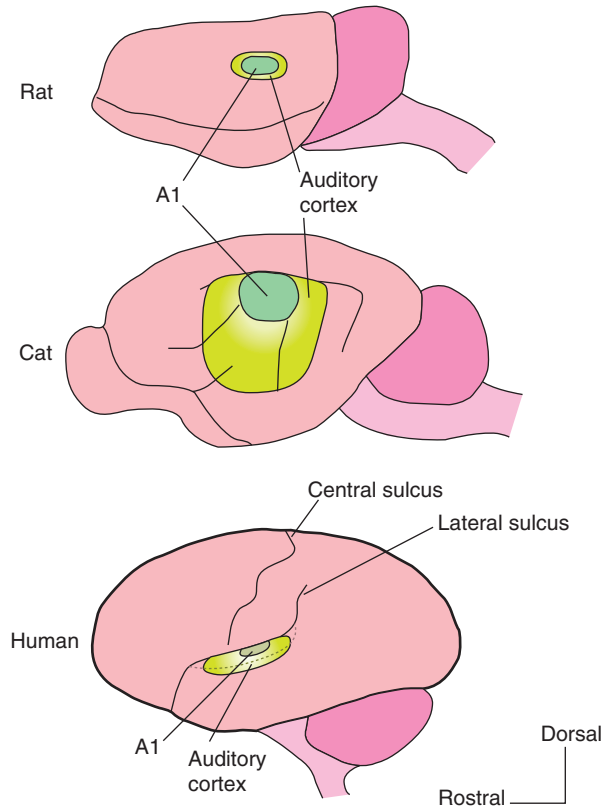

Figure 1 Approximate locations of auditory cortex on the surface of rat, cat, and human cortex. Some areas in the human brain are visible only by cutting away overlying brain tissue. A1 is the primary auditory cortex. Drawings not to scale.

another paralemniscal pathway. The MGm receives largely nontonotopic ascending auditory input and projects rather diffusely to multiple auditory cortical areas as well as nonauditory cortical areas and the amygdala. Its neurons typically respond at relatively short latencies (although with potential cell-to-cell variation) and some project with large axons, possibly indicating a fast connection from thalamus to cortex. Medial pathway projections to ACx are densest in cortical lamina 1a and possibly 6 and otherwise tend to be very sparse and fairly evenly distributed over the remaining laminae.

Other subcortical structures project to ACx in at least some species, although relatively little is known about the nature and effects of most of these projections. These projections likely underlie associative, learning, and sensory integration functions.

Area Parcellation

Distinct areas of auditory cortex are typically discerned by cytoarchitectural differences with adjacent areas and the pattern of thalamic input. Additional criteria for distinctiveness typically include variation in functional responses, although most of the anatomically determined areas of auditory cortex have yet to be explored physiologically in any detail. Primary auditory cortex (A1) receives its largest subcortical input from MGv, as typically do several other areas, the designations of which vary by species. Determination of A1 follows historically from its local architectonic distinctiveness, MGv input, and similarity in appearance to other koniocellular primary sensory areas. In addition, certain developmental stages in some species yield differential staining properties of A1, with stains differentiating a variety of molecules, such as cytochrome oxidase, acetylcholinesterase, parvalbumin, and neurofilament protein. Homology of areas labeled A1 among multiple species may be open to reinterpretation because of the plurality of cortical areas receiving substantial organized MGv input, particularly in primates.

Nomenclature of both area names and groupings varies by species. In primates, three groups of areas have been delineated based largely upon architecture, thalamocortical connections, interconnections with one another, and physiology. The core group stretches from anteriolateral to posteromedial along the lateral sulcus, contains tonotopically organized areas (including A1), and receives most of its thalamocortical input from MGv (see Figure 2). Two additional area groups situated just medial and lateral to the core are collectively referred to in modern parcellation schemes as 'the belt.' Belt areas receive little MGv input and substantially greater MGd and MGm input than do the core areas. These areas appear to be tonotopically organized but are difficult to study with classical, simple sound stimuli. Lateral to the lateral belt areas lie the parabelt areas, which receive no MGv projections and scattered MGd and MGm input. The internal organization of parabelt areas remains largely unknown.

Connections of Auditory Cortex

Local Structure of Primary Auditory Cortex

A1 represents both the most studied auditory cortical area and is the focus of the most research. For many years even such a basic organizational structure as tonotopy – the topographic representation of the cochlea – of A1 was debated because of the great variety of functional responses encountered during physiological experiments. Many modern lines of evidence, however, demonstrate tonotopy in A1, but other potential topographic structures remain unclear. In tonotopic areas, each sound frequency is represented in a narrow strip along the surface of the cortex

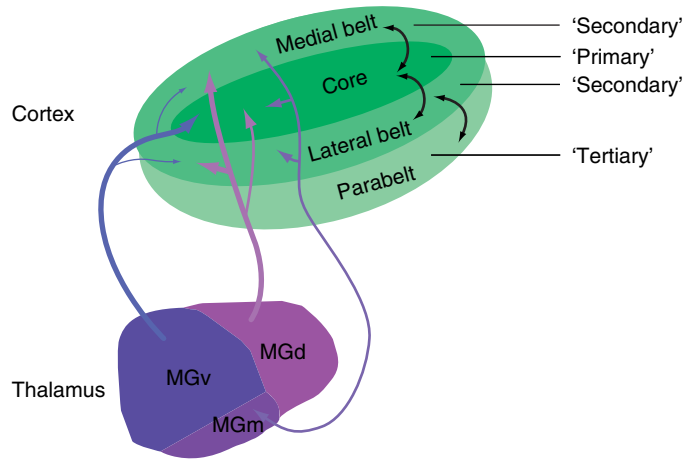

Figure 2 Thalamic input schematic to primate auditory cortex, including corticocortical connections. Arrow size approximates projection size; MGv, ventral division of medial geniculate; MGd, dorsal division of medial geniculate; MGm, medial division of medial geniculate.

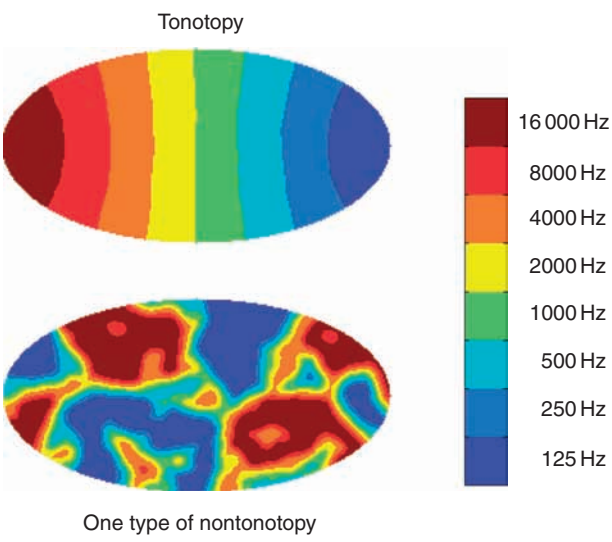

Figure 3 Model tonotopic and nontonotopic representations of sound frequency in a cortical area.

(see Figure 3), and many attempts have been made to identify other sound properties that may be mapped along each of these strips. Many sound properties are known to be mapped nonrandomly onto the surface of A1, and computational studies imply that this relatively large number (>5) makes it unlikely that any one of these parameters is mapped orthogonally to the tonotopic axis.

Local horizontal connections in A1 tend to be patchy along the lengths of the isofrequency strips; when coupled with the patchiness of the MGv inputs, this finding implies that nonfrequency processing subunits may be spatially segregated in A1. Spatial segregation of processing appears to be true for frequency, as well, and evidence exists for substantial local frequency interactions in A1, currently of unknown function but potentially for the sensitive detection of frequency modulation sweeps and extraction of harmonicity underlying the perception of pitch – both important components of species-specific vocalizations. The patchiness of other stimulus parameters, such as neuronal receptive field bandwidth, is predicted by computational topographic models.

Local connections within cortical columns of A1 remain to be further elucidated, but, as with other primary sensory areas, it is likely that specificity of interlaminar connections in A1 will become more apparent with further experimentation. Layer 4 cells tend to receive most of their excitatory input from other layer 4 cells. Layer 2/3 pyramidal cells appear to receive most of their excitatory input from either other layer 2/3 cells or layer 4 cells, reminiscent of the lemniscal/paralemniscal projection dichotomy apparent in rodent barrel cortex. How these local intracolumnar connections interface with thalamocortical and corticocortical connections remains unknown, but potentially they represent multiple modes of information transfer through A1 to other cortical areas.

Interconnections between Areas of Auditory Cortex

Auditory cortical interconnections follow the parcellation created by thalamocortical projections. The core areas in primates, which receive their major thalamocortical input from MGv, tend to interconnect densely. Core areas also interconnect with nearby belt areas, but not with parabelt areas. Belt areas appear to interconnect with most or all adjacent auditory areas (Figure 4). In terms of hierarchical auditory

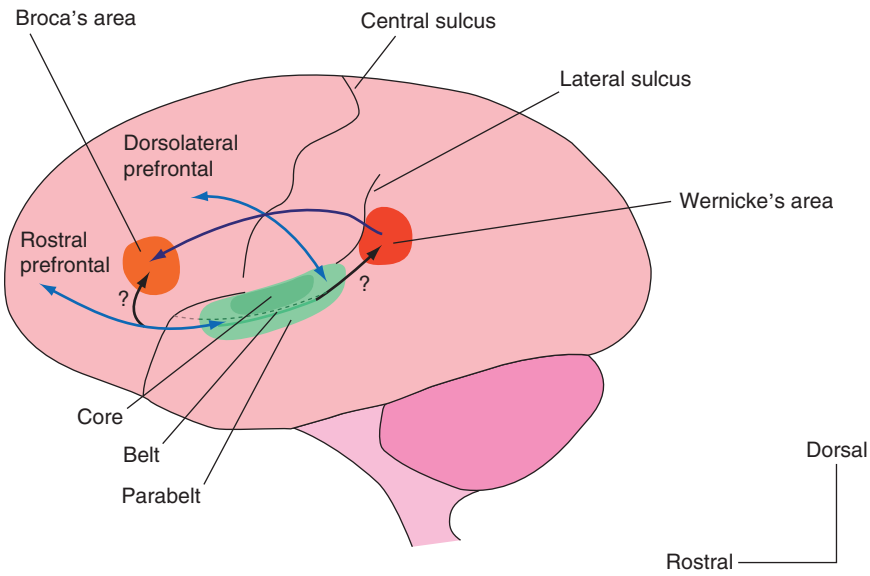

Figure 4 Projections from lateral belt to frontal and parietal lobes in primates. Frontal lobe projections are from experiments in monkeys. Wernicke's and Broca's areas have been demonstrated only in humans, but auditory cortex projections to Wernicke's and Broca's areas have not been conclusively demonstrated in humans.

cortical processing, therefore, the belt areas appear to represent an obligatory intermediate step. This arrangement of cortical area interconnection may be unique to the auditory system.

Projections among auditory cortical areas tend to be topographic in nature, connecting subregions of each tonotopic area that correspond to similar frequencies, as assessed by physiological recordings. A similar topography appears to exist even for areas that show little or no clear frequency tuning, implying that the topographic nature of the projections reflects cortical circuits implementing more than simple frequency extraction. The persistence of topographical projections in absence of clear frequency tuning may ultimately provide some insight into the sensory transformations occurring between interconnected areas of auditory cortex.

Topography of projections across the midline to the contralateral auditory cortical areas follows the same general guidelines for ipsilateral connections, including area specificity and topography. Layer 2/3 pyramidal neurons project to other auditory cortical areas and to the contralateral hemisphere, as do some layer 5a pyramidal cells and possibly some cell types in layer 4. The proportion of auditory interhemispherical projections far outweighs the proportions of similar connections in other sensory modalities, indicating that ACx of both hemispheres represents fairly integrated processing circuits. It should be noted that unlike the visual system, many subcortical auditory nuclei receive input from both sensory organs (ears), as many of these circuits are used to compute the spatial layout of sound sources in the environment relative to the organism's head.

Cortical Projections Beyond Auditory Cortex

Auditory cortical areas in primates are known to project to various nonauditory frontal and parietal lobe areas shown to be involved in cognitive functions rather than audition. While core auditory cortex areas appear to interconnect substantially only with one another and nearby belt fields, the lateral belt and parabelt areas have been shown in monkeys to project to multiple targets in the frontal lobe, in addition to interconnections with one another. At least two projection pathways from lateral belt and parabelt areas to the frontal lobe have been elucidated in macaque monkeys. The anterior lateral belt and parabelt areas appear to send projections predominantly to the rostral prefrontal areas, while the posterior belt and parabelt areas appear to target mainly the dorsolateral prefrontal cortex and adjacent areas. The parabelt areas appear to send more projections to the frontal lobe than do the belt areas. All of the projections appear to be topographic in nature, with nearby auditory cortex neurons projecting to nearby frontal lobe neurons.

It has been proposed, based upon primate functional imaging studies, visual cortex projections, and limited electrophysiology, that the posterior belt/parabelt projections to the dorsolateral prefrontal

cortex constitute a 'dorsal stream' representing predominantly spatial information. Conversely, the anterior belt/parabelt projections to the rostral prefrontal areas have been proposed to constitute a 'ventral stream' representing predominantly identifying information about the physical objections producing the sounds. Alternate interpretations for the information transmitted by these projections also exist, including representations of speaker identity. It is likely that future experimentation will reveal a greater complexity of information processing and transfer through auditory cortex to higher cognitive cortical areas than is implied by the 'two-stream' model.

Projections to Subcortical Nuclei

The ascending auditory pathway appears to have multiple projection streams that all converge initially in the inferior colliculus (IC) and again in the MGB. In other words, the IC and MGB are obligatory synapses for ascending auditory projections. In contrast, descending corticofugal projections, which represent a numerically large proportion of total auditory projections, project monosynaptically all the way to the earliest auditory brain stem nuclei. It appears, however, that few individual cortical neurons project to multiple subcortical auditory stations.

Most of the descending corticothalamic projections tend to mirror the ascending projections. In other words, thalamic nuclei that tend to send large projections to a particular auditory cortical area also tend to receive large projections from that cortical area. Like the ascending projections, these descending projections tend to be topographic, even for areas currently without clear functional organization. Not surprisingly, core areas such as A1 tend to project largely to MGv, although substantial divergence in corticothalamic projections exist such that A1 sends some projections to multiple nuclei of the MGB.

Descending corticocollicular projections appear to be organized topographically, much as are the corticothalamic pathways, but are smaller in total number and tend to target paralemniscal IC nuclei. Corticofugal projections to the brain stem represent a numerically small subset of the overall descending projections, terminating largely in the superior olivary complex (SOC) and the cochlear nuclei (CN). Most of these descending projections to all subcortical targets appear to be ipsilateral.

The cortical laminar locations of the corticofugal cells are infragranular (laminae 5 and 6) and tend to be located in specific sublaminar distributions. Corticothalamic cells tend to reside in layer 5b and throughout layer 6. Corticocollicular cells tend to reside in laminae 5 in regions that contain relatively few corticothalamic projections, although some IC-projecting cells are found in deep layer 6b. Few cortical cells appear to project to both the MGB and the IC. Cortical projections to SOC and CN are typically located in deep layer 5b.

The lack of extensive populations of cortical neurons sending projections to multiple subcortical stations argues that these corticofugal projections represent parallel descending pathways, each performing a distinct function. Details of the nature of these functions await further experimentation, but apparently dynamic and plastic modification of spectral tuning and neuronal dynamic range represents a major role of at least some of these pathways.

Other Circuitry

Other interconnections with ACx not described in detail here exist to various degrees throughout development and in the adult. Most of these connections are presumed to have modulatory roles on the encoding of acoustic stimuli, but most remain to be studied. Cholinergic projections from the basal forebrain play an active role in cortical plasticity, particularly in the developing animal but to some extent also in the adult. Dopaminergic, serotonergic, and norepinephrinergic projections to auditory cortex also exist and have yet to be studied extensively. The thalamic reticular nucleus (TRN) receives both ascending and descending auditory collateral input and appears to negatively modulate MGB responses. Interconnections with the limbic system, especially prevalent in nonprimary auditory cortex, presumably play a critical role in adding proper behavioral context to acoustic perception and memories, although little is known about how this phenomenon actually occurs. The auditory system at several levels, including cortex, interfaces with other sensory systems, most likely to integrate perception of space and object identity.

Organization of Specialized Auditory Cortex: Echolocating Bats

Extensive neuroanatomical and neurophysiological research on the acoustic processing and behavior of echolocating bats over the past half-century has revealed a tightly integrated neuronal system exquisitely adapted for one critical behavior. When seeking prey, the echolocating bat emits a series of vocalizations with stereotypical energy signatures. These energy signatures are modified by the size, shape, and velocity of the target and are reflected back to the bat's ears. Bats process this information extremely rapidly and use it to alter their own flight velocity to intercept desirable targets.

The subcortical auditory nuclei and auditory cortical areas of several species of echolocating bats have been studied successfully, thanks largely to the presence of their prey-seeking ethology and their stereotyped vocalizations. Use of species-specific vocalizations to probe physiological responses has resulted in marked advances in understanding the hierarchical processing of echolocating calls throughout the auditory system. In fact, more is known about the encoding of this specialized behavior in ACx than about any other auditory cortical phenomenon.

Knowledge of bat auditory processing of more general acoustic processing tasks, such as passive listening or species-specific vocalizations, generally does not surpass the equivalent knowledge in less specialized species. In fact, the auditory areas that seem to respond in such specific ways in echolocation tasks often respond in rather complex ways to more general tasks – much as is the case for other mammals. Nevertheless, it is hoped by many that adopting similar experimental techniques of studying behaving animals with relevant tasks and acoustic stimuli, such as species-specific vocalizations, will eventually lead to an improved knowledge of acoustic coding in these less specialized species.

Physiology of Unspecialized Auditory Cortex

Early studies of A1 using cortical surface electrodes revealed an apparent cochleotopic (tonotopic) topography. Later studies using penetrating microelectrodes recoding single neurons revealed much greater complexity than the surface electrodes hinted at and fueled claims that no such tonotopy existed. Tonotopy of A1 was eventually demonstrated unequivocally, and today many lines of evidence demonstrate this feature. The difficulty of determining such a basic organizational structure as tonotopy foreshadowed the difficulties in studying physiological responses from a cortical region where complexity, variability, and lack of clear organizing principles appear to dominate. Many attempts to assign unique processing features to A1 and other auditory cortical areas have been made, although a viable comprehensive theory regarding even the major role of A1 remains elusive.

The extraction of sound pitch or the perception of a single fundamental frequency for complex sounds represents a well-known psychophysical phenomenon. Imaging studies that show a functional signal in or near A1 in response to acoustic stimuli eliciting a sensation of pitch fail to reveal a similar source of subcortical activity. It is possible that one role of auditory cortex is to reassemble acoustic information decomposed in previous processing stages into relevant percepts, including the dimension of pitch. It has been speculated that pitch may represent only one dimension of an inherent musical processing ability implemented by circuits in ACx designed to process naturally occurring stimuli.

Experimentally, individuals exposed to many examples of repeated sounds with an occasional low-probability sound typically reveal a particular electroencephalogram (EEG) signal localized to auditory cortex in a phenomenon called mismatched negativity (MMN). Research into this phenomenon has led to theories that auditory cortex may carry recent acoustic history in its firing pattern in order to reassemble multiple streams of information over time, such as melodies, vocalization patterns, or sentences. Auditory stream analysis is a popular theory for the role of auditory cortex, although many of the properties of phenomena such as MMN can be found in subcortical structures or even simpler neural networks, such as the retina.

Different auditory cortex areas exhibit different responses to classes of species-specific vocalizations, leading to the theory that cortex is responsible for extracting information relevant to vocalizations. No models have been proposed for how this type of hierarchical processing may take place, and data upon which to build such models are currently sparse.

Nonauditory cortical areas that are known to be involved with the processing of human speech and language include Wernicke's area at the temporoparietal junction and Broca's area in the lateral frontal cortex (see **Figure 4**). These areas are located in the same cerebral hemisphere as one another and define the dominant hemisphere for humans, who have high degrees of brain lateralization. (Most individuals are left-hemisphere dominant.) Wernicke's area is involved with language comprehension, and impairment of function in this area induces a language deficit or aphasia that results in fluent speech lacking syntactical meaning. Broca's area is involved with speech production, and impairment of function in this area induces an aphasia that results in nonfluent, halting speech. These areas are connected by a projection pathway called the arcuate fasciculus, and impairment of function of this pathway preserves language comprehension and speech production but induces a so-called conduction aphasia whereby affected individuals have difficulty repeating complex utterances. (Conduction aphasias are likely to be caused by damage to other structures, as well.) Nonhuman primate analogs of Wernicke's and Broca's areas have yet to be convincingly demonstrated. It is assumed that auditory cortex connects to Wernicke's and Broca's areas in humans, but these projections have not been fully characterized.

Outstanding Questions

The auditory system differs from other sensory systems in the number and complexity of its subcortical nuclei and circuits. Bilateral information is represented at a very early stage in the auditory pathway, largely to calculate sound direction by circuits accurately measuring timing differences to microsecond resolution. Many other stimulus features appear to be extracted subcortically as well, including those that would be useful in acoustic recognition tasks. An obvious question arising from this observation is the appropriate analogy between auditory cortical areas and the processing regions of other sensory systems, notably the somatosensory and visual systems. All three of these systems project through the thalamus, but A1 is situated at least four or five spiking synapses from the periphery, unlike other primary sensory areas. The cytoarchitectural appearance of koniocellular primary auditory cortex, its cellular arrangement, and its subcortical interconnections appear to match well with those of the other two sensory systems, so the anatomical analogy seems appropriate.

On the other hand, statistical evaluations of the physiological filter elements most efficient in encoding natural stimuli indicate that the auditory nerve may actually generate an output analogous to that of primary visual cortex. If the assumption is correct – that the role of the early visual system is to neuronally construct efficient filters to encode visual scenes – then this comparison has some merit. Much of the subcortical auditory circuitry has no relevance for visual processing, especially given that binocular interactions are first seen in primary visual cortex.

Describing the responses of A1 neurons has been challenging and at times frustrating for decades – even more so for nonprimary auditory cortical areas. Is it appropriate to think of A1 in nonspecialized mammals as equivalent to a visual association area that requires appropriate stimuli and behavioral contexts to successfully study? Or perhaps a reductionist view is more appropriate: A1 may be responsible for creating particular 'filters' used to deconstruct stimuli and pass on to other areas for further processing, as V1 and bat A1 appear to do.

Questions and proposals such as these reflect at least part of the future research on high-level sensory coding in the auditory system, and begin to be resolved with continued efforts to discern what unique processing is performed by A1.

See also: Auditory Scene Analysis; Auditory System: Efferent Systems to the Auditory Periphery; Auditory System: Central Pathways; Auditory/Somatosensory Interactions; Cochlear Development; Language: Auditory Processes; Musical Illusions; Sensory Aging: Hearing; Sound Localization: Neural Mechanisms.

Further Reading

Doucet JR, Molavi DL, and Ryugo DK (2003) The source of corticocollicular and corticobulbar projections in area Te1 of the rat. *Experimental Brain Research* 153(4): 461–466.

Doucet JR, Rose L, and Ryugo DK (2002) The cellular origin of corticofugal projections to the superior olivary complex in the rat. *Brain Research* 925(1): 28–41.

Griffiths TD (2003) Functional imaging of pitch analysis. *Annals of the New York Academy of Sciences* 999: 40–49.

Hu B (2003) Functional organization of lemniscal and nonlemniscal auditory thalamus. *Experimental Brain Research* 153(4): 543–549.

Huang CL and Winer JA (2000) Auditory thalamocortical projections in the cat: Laminar and areal patterns of input. *Journal of Comparative Neurology* 427(2): 302–331.

Kaas JH and Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America* 97(22): 11793–11799.

Kaur S, Lazar R, and Metherate R (2004) Intracortical pathways determine breadth of subthreshold frequency receptive fields in primary auditory cortex. *Journal of Neurophysiology* 91(6): 2551–2567.

Kaur S, Rose HJ, Lazar R, et al. (2005) Spectral integration in primary auditory cortex: Laminar processing of afferent input. *in vivo and in vitro. Neuroscience* 134(3): 1033–1045.

Lee CC and Winer JA (2005) Principles governing auditory cortex connections. *Cerebral Cortex* 15(11): 1804–1814.

Lewicki MS (2002) Efficient coding of natural sounds. *Nature Neuroscience* 5(4): 356–363.

Penagos H, Melcher JR, and Oxenham AJ (2004) A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *Journal of Neuroscience* 24(30): 6810–6815.

Peretz I and Zatorre RJ (2005) Brain organization for music processing. *Annual Review of Psychology* 56: 89–114.

Romanski LM, Bates JF, and Goldman-Rakic PS (1999) Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. *Journal of Comparative Neurology* 403(2): 141–157.

Schreiner CE, Read HL, and Sutter ML (2000) Modular organization of frequency integration in primary auditory cortex. *Annual Review of Neuroscience* 23: 501–529.

Schwartz O and Simoncelli EP (2001) Natural signal statistics and sensory gain control. *Nature Neuroscience* 4(8): 819–825.

Smith PH and Populin LC (2001) Fundamental differences between the thalamocortical recipient layers of the cat auditory and visual cortices. *Journal of Comparative Neurology* 436(4): 508–519.

Suga N (1990) Biosonar and neural computation in bats. *Scientific American* 262(6): 60–68.

Suga N (1994) Processing of auditory information carried by species-specific complex sounds. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 295–313. Cambridge, MA: MIT Press.

Suga N, Xiao Z, Ma X, et al. (2002) Plasticity and corticofugal modulation for hearing in adult animals. *Neuron* 36(1): 9–18.

Winer JA (2005) Decoding the auditory corticofugal systems. *Hearing Research* 207(1–2): 1–9.

Winer JA, Miller LM, Lee CC, et al. (2005) Auditory thalamocortical transformation: Structure and function. *Trends in Neuroscience* 28(5): 255–263.

Auditory Cortex: Models

S A Shamma and J B Fritz, University of Maryland, College Park, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The auditory cortex plays a critical role in the perception and localization of complex sounds. Incoming acoustic stimuli arrive in the ears, are transformed into a neural code in the cochlea, and ascend in an interwoven and bidirectional network of processing centers, passing through the cochlear nuclei and the superior olivary complex, the lateral lemniscus, the inferior colliculus (IC), and the medial geniculate body (MGB) en route to auditory cortex. Despite the rapidly expanding knowledge of the neuroanatomy and connectivity of the auditory cortex, relatively little is known about its functional organization, especially compared with the visual and the motor systems. One exception has been the auditory cortex of the mustache bat, in which functional maps have been reported, serving the stereotypical echolocating behavior of this species. In other mammals, it is more difficult to isolate a natural auditory behavior and its associated relevant stimulus features with comparable specificity. Nevertheless, a few tasks have been broadly accepted as being vital for all species, such as sound localization, timbre recognition, pitch perception, and of course survival and reproduction! For each, evidence of various functional and stimulus feature maps has been found or postulated, a significant number of them in the past few years. This article elaborates on a few examples of such organization. In each example, the goal is to determine how and whether models of the underlying neural networks can deepen our understanding of the auditory cortex.

Auditory Cortical Fields

The layout and neural structure of the auditory cortex is in many respects similar to that of other sensory cortices. For instance, based on physiological and cytoarchitectonic criteria, and patterns of connectivity, it is subdivided in the cat into a primary auditory field (A1) and several other surrounding fields, such as the anterior auditory field and the secondary auditory cortex. The number and specific arrangement of surrounding fields varies among different species, reflecting presumably the complexity of the animal's acoustic behavior (and the number of anatomical and physiological mapping studies performed). For ballpark figures, both the macaque and the cat have at least 14 auditory cortical fields, the ferret has at least seven, the mustache bat has at least 12, the Mongolian gerbil has at least eight, and the rodent has at least five. The most obvious physiological organizational feature of the core and other tonotopic fields is the presence of a systematic frequency sensitivity map inherited originally from the cochlea. In A1, responses are spatially ordered based on the tone frequency to which they respond best (or their best frequency, BF). The A1 area devoted to biologically significant frequencies may be enlarged, as exemplified by the 'acoustic fovea' in the Doppler-shifted constant frequency (DSCF) area in the mustache bat. But there may also be an increase in frequency representation in A1 to encode the learned importance of sound. Within the isofrequency bands, there is an interleaved modular organization that includes tuning curve bandwidth, binaural properties, and intensity threshold and tuning, which suggests significant differences in local network architecture. In the awake animal, A1 neurons respond vigorously to the onset of a tone and show sustained responses to preferred stimuli. In some cortical belt areas, although there is a poor response to pure tones, cochleotopic frequency organization can be demonstrated by the use of bandpassed noise stimuli. In other fields, responses are even less frequency-selective, more adaptive, or totally absent to single tones, preferring instead more spectrally or temporally complex stimuli. A sudden change in these response patterns or in the gradual spatial order of the tonotopic map is usually taken to signify a border between different fields. Finally, A1 and possibly other fields are further subdivided into smaller regions, serving perhaps different functional roles. One promising approach to the study of cortical functionality is illustrated by a set of recent reversible lesion studies.

General Cortical Mechanisms

The auditory cortex shares with other cortical structures basic neuronal mechanisms and architectures that are thought to be actively involved in a wide range of perceptual processes. For example, the input layers in A1 contain spectrotemporal receptive fields that share strong analogies to those of the visual cortex. This suggests that common organizational and functional principles underlie the excitatory and inhibitory thalamocortical projections and the representation of sensory cues in the primary areas. Anatomical and

physiological support of this hypothesis was provided by experiments in which optic nerve projections to the auditory thalamus (MGB) of newborn ferrets resulted in the eventual development of classical visual sensitivity and receptive fields in the target A1 that supported functional visual behavior. However, one curious finding from intracellular recordings in the auditory cortex is the apparent complete overlap of the inhibitory and excitatory inputs (which may be as important for increasing temporal precision as for spectral receptive field shape). Although some of the rich spectral structure of the receptive field is already established in the thalamus, it appears that the thalamocortical input from the lemniscal pathway preferentially confers responses to BF of cortical neurons. However, it is also clear that there are highly influential horizontal fiber corticocortical inputs that also shape spectral receptive field properties, especially for non-BF responses, thus creating a much more complex spectral receptive field structure for A1 neurons.

Adaptation and synaptic depression are two important physiological properties of cortical responses in general that have found strong resonance in psychoacoustics. In particular, they have been hypothesized to be the neural correlates of perceptual phenomena such as 'forward masking,' 'buildup of perceptual streams' in auditory scene analysis (see below), and the 'multiple looks hypothesis' for the integration of cues leading to detection. Interesting possible physiological correlates of these perceptual phenomena have been recently described. One intriguing discovery is that stimulus-specific adaptation takes place on multiple timescales ranging from hundreds of milliseconds to tens of seconds, which may play a role in encoding auditory memory.

Another important property of auditory cortex is redundancy reduction in representation of complex spectrotemporal stimuli compared with stimulus-induced redundancy observed in IC and MGB. Neural correlation in spiking patterns between adjacent neurons is more likely when there is overlap of spectrotemporal receptive fields or when the difference in characteristic frequency is small.

Finally, rhythmic gamma oscillations (20–120 Hz) abound in the auditory cortex, much like those found in other areas. The functional role of oscillatory synchronization of neural activity remains mysterious but may enhance coincidence detection or noise tolerance or play a role in plasticity and attention.

Encoding of Timbre

Timbre is the perceptual attribute of sound responsible for recognizing and classifying complex sounds, such as the distinctions among speech phonemes and different instruments playing at the same pitch. While a multitude of acoustic cues underlie this ability, the most important is the shape of the spectral profile of sound and its evolution in time. The spectral profile emerges early in the auditory system as the sound is analyzed into different frequency bands, in effect distributing its energy along the one-dimensional tonotopic axis. However, in the central auditory system, the spectral profile is expanded into a two-dimensional response pattern, with each frequency represented by an entire sheet of neurons – the so-called isofrequency axis. Many hypotheses have been proposed as to the nature of this added dimension and how the response patterns in the primary auditory cortex might be representing the spectral profile and its dynamics. One early hypothesis is the best-intensity model, motivated primarily by the strongly nonmonotonic responses as a function of stimulus intensity observed in many cortical cells. One can view such a cell's response as being selective to (or encoding) a particular intensity. Consequently, a population of such cells, tuned to different frequencies and intensities, can provide an explicit representation of the spectral profile by their spatial pattern of activity. The most compelling example of such a representation is in the DSCF area of A1 in the mustache bat. However, an extension of this hypothesis to multicomponent stimuli has not been demonstrated in any species.

The second model is inspired by extensive data and ideas gained from physiological and psychoacoustical experiments over the past decade. Specifically, insight has been gained from measurements of the so-called spectrotemporal response fields (STRFs) of A1 cells. An STRF summarizes the dynamics and sensitivity of a cell or, more precisely, the impulse response of the cell at each frequency. Thus, an STRF displays the excitatory and inhibitory interactions that give the cell its selectivity to spectrotemporal patterns. Some STRFs are responsive (excited or suppressed) over a broad range of frequencies, exceeding an octave, while others are quite narrowly tuned. Dynamically, some STRFs' responses decay rapidly after an impulse, while others last twice as long. Finally, this combined time–frequency sensitivity can take more complex forms that are 'inseparable,' as in oriented STRFs that are sensitive to frequency modulations (FM).

STRFs have been measured in many ways, including reverse correlation with random tone chords or spectrotemporally modulated noise. Another method is the 'ripple analysis method' that employs broadband noise with sinusoidally modulated spectrotemporal envelopes with different parameters. Ripples serve the same function as regular sinusoids in measuring the transfer function of linear filters,

except that they are two dimensional (spectral and temporal). A1 cells respond well to ripples and are usually selective to a narrow range of ripple parameters that reflect details of their receptive fields. By compiling a complete description of the responses of a cell to all ripple densities and velocities, it is possible by an inverse Fourier transform to compute the corresponding STRF.

From a functional perspective, the rich variety of STRFs found in A1 implies that each STRF acts as a modulation-selective filter of its input spectrogram, specifically tuned to a particular range of spectral resolutions (also called scales) and a limited range of temporal modulations (or rates). The collection of all such STRFs then would constitute a filter bank spanning the broad range of psychoacoustically observed scale and rate sensitivity in humans and animals. Evidence of the importance of spectrotemporal modulations in the perception of complex sounds has come from experiments in which systematic degradations of the speech signal were correlated with the gradual loss of intelligibility. In fact, the relationship between the temporal modulations and speech intelligibility has long been codified in the formulation of the widely used speech transmission index and the spectrotemporal modulation index, which assesses the integrity of both the spectral and temporal modulations in a signal as a measure of intelligibility.

Pitch Representation in the Cortex

A sound complex consisting of several harmonics is heard with a strong pitch at the fundamental frequency of the harmonic series, even if there is no energy at all at that frequency. This percept has been variously called the missing fundamental, virtual pitch, or residue pitch. A large number of psychoacoustical experiments have been carried out to elucidate the nature of this percept and its relationship to the physical parameters of the stimulus. Basically, all models fall into one of two camps. The first believes that the pitch is extracted explicitly from the harmonic spectral pattern. This can be accomplished in a variety of ways, for instance by finding the best match between the input pattern and various harmonic templates assumed to be stored in the brain. The second group claims that the pitch is extracted from the periodicities in the time waveform of responses in the auditory pathway that can be estimated, for example, by computing their autocorrelation functions. In these latter models, some form of organized delay lines is assumed to exist in order to do the computations, much like those that seem to exist in the FM–FM area of the mustache bat.

In all pitch models, however, it is assumed that the extracted pitch is finally represented as a spatial map in higher auditory centers. This is because many studies have confirmed that neural synchrony to the repetitive features of a stimulus, whether it is the waveform of a tone or its AM modulations, becomes progressively worse toward the cortex. It is a remarkable aspect of pitch that, despite its fundamental and ubiquitous role in auditory perception, only a few reports exist of physiological evidence of spatial pitch maps, and none have been independently confirmed. One source is functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) scans of the human auditory cortex. Another is from single-unit and multiunit responses in various precortical auditory structures and recently in the primate auditory cortex. One key difficulty in all experiments seeking to demonstrate physiological correlates of pitch is the cochlear nonlinearity that produces distortion components at the fundamental of the upper harmonics and that unintentionally excites low BF cells. This 'artifact' has cast a shadow of doubt over all discoveries of physiological pitch maps because of the experimental difficulties in avoiding or masking it.

Of course, the difficulty in finding a spatial pitch map in the auditory cortex may be due to the fact that it does not exist! This possibility is counterintuitive, given the results of ablation studies that show that bilateral cortical lesions in the auditory cortex severely impair the perception of pitch of complex sounds but do not affect the fine discrimination of frequency and intensity of simple tones. Another possibility is that the maps sought are not at all as straightforward as imagined. For example, harmonic complexes may evoke stereotypical patterns that are distributed over large areas in the auditory cortex and not localized, as the simple notion of a pitch map implies. Finally, it is also possible that A1 simply functions as one stage that projects sufficient temporal or spectral cues for later cortical stages to extract the pitch explicitly.

Models of Sound Localization

It has been recognized for many years that the auditory cortex (and especially the A1) is involved in sound localization. Detailed physiological studies further confirmed that A1 cells are sensitive to manipulations of the binaural stimulus, such as the interaural level difference (ILD) and interaural time difference, a sensitivity that has its origins early in the auditory pathway, where the first convergence of binaural inputs occurs. In fact, a combined auditory–visual map of auditory space has been

found in the superior colliculus of several mammals and in the IC of the barn owl. However, the exact role of the auditory cortex in the perception of auditory space remains unclear since no obvious topographic organization of auditory space has been found. Instead, responses from the two ears have been found to alternate between excitatory–excitatory (EE) and excitatory–inhibitory (EI) interactions in bands along the isofrequency planes of A1. One possible functional model that utilizes such maps assumes that EI cells are tuned to particular ILDs and hence encode the location of a sound source based on this cue. EE cells would in contrast encode the absolute level of the sound. However, there is little evidence to support this hypothesis in the sense that neither EE nor EI cells are particularly stable encoders of specific ILD or absolute sound levels. An alternate hypothesis is that these cells encode the absolute levels of the stimulus at each ear rather than the difference and average binaural levels, as previously postulated. Location coding has recently been modeled by an opponent process theory in which sound source locations are represented by population differences in activity of broadly tuned contralateral and ipsilateral space neurons. Another hypothesis uses a likelihood approach to spatial coding.

Finally, it has also been proposed that A1 units encode the spatial location of a stimulus by unique patterns of temporal firing, ones that can be discerned by more elaborate pattern recognition neural networks. Recent findings reveal a role for auditory cortex in long-term plasticity of auditory localization. Attentional mechanisms may operate in a much shorter time frame; for example, a recent study in the barn owl suggests that focused spatial attention may tighten auditory spatial receptive fields in the IC.

Neural Correlates Auditory Scene Analysis

Humans and many animals can attend to one sound source and stream (or segregate) it from a background of many sources rapidly and with no learning or prior exposure to the specific sounds. For humans, this is the essence of the well-known 'cocktail party' phenomenon, in which a person can effortlessly conduct a conversation with a new acquaintance in a crowded and noisy room. For frogs, songbirds, or penguins, this ability is vital for locating a mate or an offspring in the midst of a loud chorus.

The question of how the acoustic scene is parsed by the auditory system into auditory objects and streams is one of the most fundamental in perceptual science. Despite its importance, the study of the underlying neural mechanisms and their computational models remains in its infancy. Several studies have looked for neural correlates of auditory streaming in humans, using electroencephalography (EEG), MEG, and fMRI. The earlier studies focused on the mismatch negativity (MMN) as a potential electro- or magnetoencephalographic index of streaming and found indications that certain features of the MMN (its presence, amplitude, or latency) are related to how the evoking sequence was perceptually organized. However, because the experimental manipulations in these studies always involved changes in the physical stimulus characteristics, differences in the MMN across experimental conditions could reflect factors other than streaming *per se*. Other EEG investigations of streaming have overcome some of these criticisms by exploring the properties of streaming and the systematic differences that emerge when listeners were in different auditory perceptual states. A recent study examines the way in which attending to one sound stream modifies the extent to which the unattended sounds are processed. Some investigators have also recently begun to use fMRI in order to identify what regions of the brain are activated during the perception of one or two auditory streams. Findings thus far point to overall activation in nonprimary auditory cortex and parieto-occipital regions.

While fMRI (and to some extent MEG and EEG) studies may reveal what regions of the brain are involved during streaming, they provide only limited insight into the actual neuronal mechanisms underlying auditory streaming and are hence only indirectly helpful in the formulation of computational and functional models of auditory scene analysis. A more fine-grained view is achieved through single- and multiunit recording studies in primates, bats, and birds. The results have generally pointed to a model in which enhanced segregation of neural activity plays a critical role in mediating the streaming percept and to mechanisms that are fairly basic to cortical physiology, including 'forward suppression' or 'adaptation,' synaptic depression, and spiking threshold. Despite these encouraging results, two criticisms can be made of all neural studies of streaming so far. First, none has provided behavioral tests of the percepts that the animals were experiencing with the stimuli used in the experiments. Second, all the above studies involved awake but nonbehaving animals. This may be particularly important for streaming-related issues as it is possible that attention is critically involved in at least some aspects of auditory stream formation. Combined physiological and behavioral experiments, although generally difficult to perform, are not impossible. Such experiments have been successfully performed and have led to important insights into the neural mechanisms of perception.

There are numerous computational models of auditory scene analysis that can partially disentangle mixtures of sources. These models vary substantially in flavor, complexity, and details, depending on the specific phenomena they address. For instance, computational models of streaming with relatively simple periodic tonal sequences, noise, and other abstract stimuli have emphasized data-driven low-level processing and the notion that peripheral cochlear channeling is crucial in stream formation. Auditory scene segmentation and edge detection have been modeled based on a study of amplitude transients, and sparse overcomplete linear representations in A1 have been proposed to be the basis for solving the cocktail party problem. By contrast, models for segregating mixtures of speech, music, and other complex nonperiodic sources have sought to include more-complex perceptual approaches, including a role for context and expectations, integration of grouping cues over relatively long times, and even explicit training with information beyond that normally available to a listener, such as prior exposure to clean exemplars of the voices in the mixture.

An alternate computational view of streaming is a purely data-driven process that, at the same time, incorporates top-down perceptual processes such as expectations. In this view, the auditory representation of sound undergoes an iterative clustering process that integrates cortical dynamics and Kalman-filtering prediction to form the final streams. Integrating bottom-up feedforward transformations with top-down expectations provides a computational account of (1) how low-level grouping cues such as harmonicity, onsets, and binaural cues can influence the formation of perceptual streams; (2) how cortical dynamics can give rise to the observed dynamics of streaming in simple stimuli as well as musical and speech rhythms; and (3) why the 'old-plus-new' heuristic has proven exceptionally influential in explaining the perceptual organization of complex sounds.

Emerging Functional Views of the Cortex

The classical conceptual view of the auditory system (and other sensory systems) as a sequence of feedforward transformations and feature extractions leading to the ultimate 'cognitive' recognizers is rapidly being replaced by a more integrated view that takes into account (1) the adaptive nature of its receptive fields and their sensitivity to contextual cues and stimulus 'meaning'; (2) the vast networks of feedback from higher centers down to the earliest processing stages; (3) the substantial multimodal interactions among different sensorimotor pathways; (4) interactions with reward, limbic, and cognitive pathways; and (5) the global rhythmic brain activity that may reflect and (in turn) enhance the attentional state of the animal. Consideration of all these important factors is essential for a complete understanding of active auditory processing. One hypothesis is that A1 is not a feature analyzer or detector but rather that the role of auditory cortex is to organize these disparate feature properties into discrete auditory objects in the context of a coherent auditory scene.

The Chameleon Brain

Long-term auditory experience or learning has been shown to cause profound global effects, such as reshaping of tonotopic maps, and significant local effects by transforming receptive field properties of neurons in A1. A computational model has been developed to describe the experience-dependent reorganization of tonotopic maps. Convergent studies of plasticity in the auditory, visual, and motor systems have recently also demonstrated the capacity for dynamic modulation of representational maps and shown that cortical cells in these systems can undergo rapid, task-dependent, and context-specific changes of their receptive field properties during attentive behavior. The key elements of this form of adaptive plasticity appear to be (1) directed attention to salient task-related cues, which leads to (2) selective functional reconfiguration of the underlying cortical circuitry that occurs simultaneously with task performance and causes (3) changes in receptive field properties of individual neurons and the cortical ensemble, which may enhance behavioral performance in the current task. These findings suggest that cortical receptive fields are not fixed but may be constantly adapting and reorganizing dynamically to meet the challenges of an ever changing environment and new behavioral demands and may play an important role in information processing and storage. In this functional model, each primary sensory cortical neuron participates in multiple behavioral contexts, and it is likely that its receptive field properties are differentially modified by top-down influences in each case; its network connectivity may also be reconfigured in an immediate and reversible manner as the animal switches between behavioral states. In this way, the same neuronal ensemble can mediate entirely different perceptual functions. The basic adaptive mechanisms that underlie this plasticity may be similar in perceptual and motor learning and also during optimal performance of a previously learned task. It is interesting that it appears that flexibility in task-dependent processing of similar acoustic stimuli is a fundamental principle, not only at the level of single cells and local networks, but also at the level of hemispheric activation.

Multimodal Influences

Recent neuroanatomical and neurophysiological studies have shown a convergence of multisensory (visual and somatosensory) inputs to auditory cortex. Task-related motor and reward-related responses have also been reported in auditory cortex. It is interesting that nonauditory inputs appear to be gated by task relevance in the auditory cortex and IC. These results suggest that rather than a purely unisensory processing stream, responses in the auditory cortex must be understood as an interwoven tapestry of relevant multimodal contextual inputs.

Feedback and Top-Down and Cognitive Influences

It is well established that cognition is simultaneously bottom-up and top-down, involving interactions between bottom-up, sensory-driven information and top-down, attentional memory and executive processes that modulate the bottom-up processing. This bottom-up versus top-down distinction is roughly consistent with the anatomical and physiological evidence of a cortical architecture that abounds with forward and backward axonal projections in the neocortex and associated structures (such as the thalamus, the amygdala, the striatum, and the hippocampus). In fact, there is increasing evidence that most higher brain functions, including the brain's ability to learn from experience, depend on the integration of such forward and feedback signals. Consequently, a complete understanding of how auditory cortical responses encode the acoustic environment must take into account the behavior of the animal within it.

For instance, one simply cannot obtain a true understanding of auditory cortical responses to a threatening sound by merely playing it to an anesthetized animal. Instead, when an animal recognizes and escapes threatening sounds, it enters a highly aroused and attentive state in which it categorizes its predators' calls as salient foreground targets to distinguish them from harmless background sounds. Simultaneously, it likely integrates other sensory cues (visual, olfactory) as well as its stored acoustic memories into its auditory judgment of the nature of the calls. In addition, neuronal correlates of category formations in the prefrontal cortex would likely feed back and adapt the receptive fields of the auditory cortex so as to enhance the perception of the target sounds (against the background) and subsequently to generate an appropriate multimodal representation of the scene and plan motor actions to respond to the threat. All these interactions significantly alter auditory responses in the cortex, and a massive descending corticofugal feedback system dynamically reshapes cortical inputs, perhaps influencing precortical structures all the way down to the cochlea, and hence must be taken into account when dissecting the nature of auditory cognition.

See also: Auditory Cortex Structure and Circuitry; Auditory Scene Analysis; Auditory System: Efferent Systems to the Auditory Periphery; Auditory System: Central Pathway Plasticity; Auditory Evoked Potentials; Auditory System: Central Pathways; Auditory Localization; Sound Localization: Neural Mechanisms; Speech Perception: Neural Encoding; Temporal Processing in the Auditory Pathway.

Further Reading

Bakin JS and Weinberger NM (1996) Induction of a physiological memory in the cerebral cortex by stimulation of the nucleus basalis. *Proceedings of the National Academy of Sciences of the United States of America* 93: 11219–11224.

Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organisation of Sound*. Cambridge, MA: MIT Press.

Carlyon RP (2004) How the brain separates sounds. *Trends in Cognitive Neurosciences* 8: 465–471.

Clarey J (1992) Physiology of thalamus and cortex. In: Popper AN and Fay RR (eds.) *The Mammalian Auditory Pathway: Neurophysiology*, pp. 232–334. New York: Springer.

Edeline JM (2003) The thalamo-cortical auditory receptive fields: Regulation by the states of vigilance, learning and the neuromodulatory system. *Experimental Brain Research* 153: 554–572.

Elhilali M, Chi T, and Shamma S (2003) Intelligibility and the spectrotemporal representation of speech in the auditory cortex. *Speech Communication* 41: 331–348.

Fritz JS, Shamma S, Elhilali M, and Klein D (2003) Rapid task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience* 6: 1216–1223.

Ghazanfar A and Schroeder C (2006) Is neocortex essentially multisensory? *Trends in Cognitive Neurosciences* 10: 278–285.

Goldstein J (1973) An optimum processor theory for the central formation of pitch of complex tones. *Journal of the Acoustical Society of America* 54: 1496–1516.

Kaas JH and Hackett TA (2005) Subdivisions and connections of the auditory cortex in primates: A working model. In: König R, Heil P, Budinger E, and Scheich H (eds.) *The Auditory Cortex: A Synthesis of Human and Animal Research*, pp. 7–26. Mahwah, NJ: Lawrence Erlbaum.

Knudsen EI and Konishi M (1978) A neural map of auditory space in the owl. *Science* 200(4343): 795–797.

Li C-SR, Padoa-Schioppa C, and Bizzi E (2001) Neuronal correlates of motor performance and motor learning in the primary motor cortex of monkeys adapting to an external force field. *Neuron* 30: 592–607.

Naatanen R, Tervaniemi M, Sussman E, Paavilainen P, and Winkler I (2001) "Primitive intelligence" in the auditory cortex. *Trends in Neurosciences* 24: 283–288.

Sur M, Garraghty PE, and Roe AW (1988) Experimentally induced visual projections into auditory thalamus and cortex. *Science* 242: 1437–1441.

Ulanovsky N, Las L, Farkas D, and Nelken I (2004) Multiple time scales of adaptation in auditory cortex neurons. *Journal of Neuroscience* 24: 10440–10453.

Auditory Evoked Potentials

J R Melcher, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When sounds reach our ears, they are transduced by peripheral hearing structures into electrical signals in the auditory nerve. The nerve transmits these signals to the central auditory system, where they are processed by neurons in the brain stem, thalamus, and cerebral cortex. The result of this processing is the perception of sound. Many thousands of neurons are engaged in the transformation from sound to perception. Neuronal activity produces in the surrounding brain tissue a voltage that can be recorded from electrodes on the surface of the head. A measurement of the voltage produced collectively by the active neurons is an auditory evoked potential (AEP).

AEPs Produced by Brief Stimuli

Many sounds produce a measurable AEP, but AEPs produced by brief sounds, such as acoustic pulses (called clicks), brief tones, and short bursts of broadband noise, are especially robust. Figure 1 shows a stylized AEP in response to a click. The response consists of a series of voltage fluctuations continuing for hundreds of milliseconds following the sound stimulus. In response to a single stimulus presentation, most, if not all, of these fluctuations are swamped by the voltage produced by ongoing activity in the brain (i.e., the electroencephalogram). An AEP is extracted from this ongoing signal by averaging the voltage following many presentations of the same stimulus.

The AEP illustrated in Figure 1 is based on recordings in people. However, AEPs have also been recorded in many other species, including mice, guinea pigs, rats, cats, and dolphins.The AEP waveform differs across species, reflecting, for instance, interspecies differences in head size, axonal path lengths, physical location and orientation of cells in the head, and the specific auditory cell types present in the auditory pathway. In all cases, however, there is a major commonality: the AEP comprises a time-varying voltage following the stimulus.

The AEP produced by a brief stimulus is generally divided into three parts:

1. The auditory brain stem response (ABR) consists of the shortest latency fluctuations occurring within approximately 10 ms of the stimulus (green in Figure 1). The voltage fluctuations of the ABR are each ~1 ms in width. As the name of this response indicates, these fluctuations are generated by auditory neurons in the brain stem.
2. The middle latency response (MLR) consists of a series of voltage fluctuations immediately following the ABR and preceding the long latency response (red in Figure 1). The voltage fluctuations comprising the MLR are much broader than are the fast fluctuations of the ABR. The MLR partly reflects cellular activity in the thalamus.
3. The late latency response (LLR) begins 50–100 ms after the stimulus (blue in Figure 1). It includes one of the most widely studied components of the AEP, the N1 or N100. The complete response consists of a complex of positive and negative fluctuations generated by the cortex.

Dependencies on Stimulus Parameters

The ABR, MLR, and LLR all depend on physical aspects of the sound stimulus. For instance, increases in the rate of stimulus presentation generally decrease response amplitude. Increases in stimulus intensity generally increase the amplitude of responses and decrease the latency (i.e., shorten the time between stimulus and response). These rate and intensity dependencies are partly attributable to the behavior of the auditory nerve, which generates the first wave of the ABR and provides input, directly or indirectly, to neurons contributing to the remainder of the AEP. For example, individual auditory nerve fibers are more likely to discharge in response to high-intensity stimuli and will do so with a shorter latency. Wave I of the ABR, which reflects the summed response from all auditory nerve fibers, is correspondingly greater in amplitude and shorter in latency, the typical intensity dependence for AEPs.

Dependencies on Arousal and Attention

The MLR and LLR differ from the ABR in their sensitivity to whether individuals exposed to the stimulus are asleep or awake, whether they are paying attention to the stimuli, and whether they are anesthetized. For instance, the LLR can change substantially across different stages of sleep. It also shows distinct changes in amplitude when individuals perform a task that requires concentrated listening (i.e., attention) to the stimuli used to evoke the LLR. The ABR is highly resistant to changes in state, persisting, for instance,

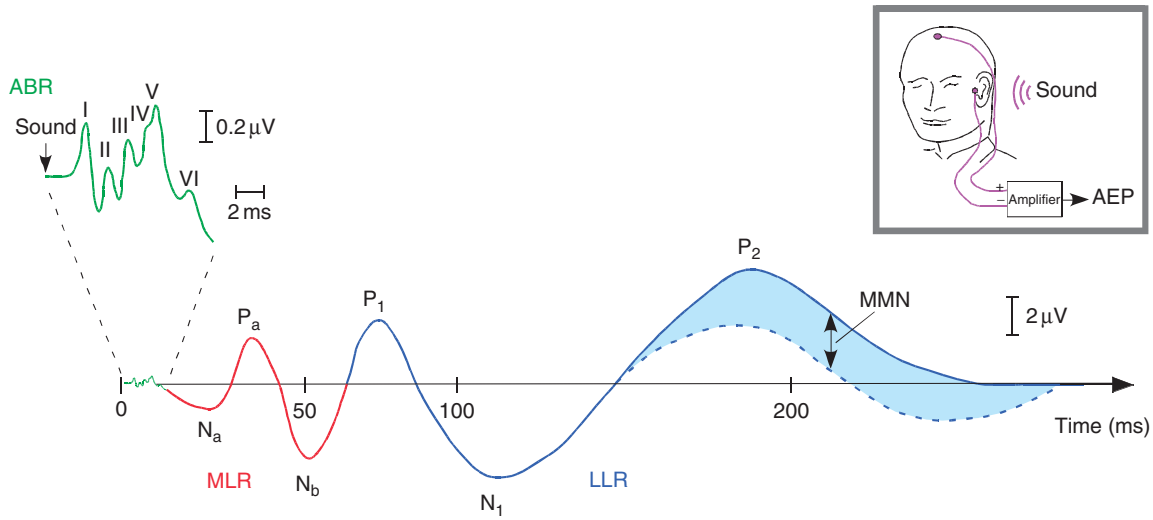

Figure 1 Stylized auditory evoked potential (AEP) in humans. The illustration is modeled after responses to brief stimuli recorded between two electrodes: at the vertex or top of the head and near the stimulated ear (inset at top right). Stimulus presentation is at 0 ms (at arrow in top waveform). The entire AEP includes the auditory brain stem response (ABR; green), the middle latency response (MLR; red), and the late latency response (LLR; blue). The ABR has been expanded (top left) so that its individual wave components (labeled I–V) can be resolved. The dashed trace indicates how the response to the same stimulus would differ if presented as the deviant stimulus in an oddball paradigm. The dashed waveform, minus the solid, is the mismatch negativity (MMN).

even when individuals are anesthetized deeply enough for surgery. The resilience of the ABR has made it an effective tool for monitoring the health of the auditory system during neurosurgery (e.g., to remove a tumor from the vestibulocochlear nerve) and for infant hearing tests (ideally performed during sleep).

Neural Generators of AEPs

While the ABR, MLR, and LLR differ in many ways (amplitude, latency, duration of individual voltage fluctuations) and are generated by different neurons, their mechanism of generation is fundamentally the same, since all represent the net potential produced by large populations of neurons. Several factors determine whether any given neuronal population will contribute substantially to the overall AEP:

1. *The total number of neurons in the population.*
2. *The amplitude of the potential produced by individual neurons at the AEP recording electrodes.* This amplitude can be dependent on neuronal morphology. For example, when current flows across the membrane of a neuronal cell body and returns through the dendrites, a so-called closed field results if the dendrites uniformly surround the cell body (i.e., zero potential at the AEP electrodes), whereas an open field results if the dendrites are distributed nonuniformly (i.e., nonzero potential).
3. *Similarity of morphology and orientation across neurons.* For example, two oppositely oriented neurons next to one another will produce equal but opposite (i.e., canceling) potentials.
4. *The degree to which activity is temporally synchronized across neurons.* The potentials produced by the individual neurons of a population will only add effectively if the neuronal responses to the AEP stimulus are similar in latency (i.e., temporally aligned with one another).

A variety of approaches have been used to identify neural generators of the AEP. One widely used approach involves making focal brain lesions in animals and examining the effect on the AEP. Lesioning techniques used for this purpose include making surgical cuts, passing electric current, or injecting a neurotoxin. The effects of neurotoxic lesions, combined with modeling, have led to the conclusion that most of the click-evoked ABR in cats is generated by neurons in two of the many pathways of the auditory brain stem. These pathways originate with the bushy cells of the cochlear nucleus, the most peripheral nucleus of the central auditory system and recipient of direct inputs from the auditory nerve. They extend centrally to the superior olive, nuclei of the lateral lemniscus, and inferior colliculus. In cats, there are two types of bushy cells, spherical and globular, which mark the beginning of the two pathways involved in ABR generation. The pathways emanating from bushy cells

have specialized features for maintaining the timing of neuronal activity (e.g., large-diameter axons and unusually large terminal endings called endbulbs), and hence maintain the high degree of temporal synchronization across neurons present in the auditory nerve. This temporal synchronization is one of several features that make bushy cells (spherical and globular), as well as the direct and indirect targets of these cells, robust contributors to the cat ABR. In contrast to cats, the globular bushy cell population in humans is proportionately quite small, and according to some is nonexistent. Therefore, the human ABR (after wave II) may largely reflect activity in just one pathway consisting of spherical bushy cells and their primary and secondary targets (e.g., medial superior olive, inferior colliculus).

The potential contributed to the AEP by brain stem neurons likely reflects discharge activity on the cell body and axon. (This is consistent with the brevity of the fluctuations comprising the ABR, ~1 ms.) In contrast, the traditional view of cortically generated potentials (e.g., P1, N1, P2, of the LLR) is that they arise from synaptic currents flowing between the cell body and dendrites of pyramidal cells (a synaptic origin is consistent with the longer duration of the fluctuations comprising the LLR). The cell bodies of pyramidal cells, lying in the deeper layers of cortical gray matter, give rise to an apical dendrite extending into the superficial layers and oriented perpendicular to the cortical surface. The morphology of pyramidal cells (resulting in an open, or dipole, field) and the consistent dendritic orientation across pyramidal cells mean that large groups of synchronously activated pyramidal cells are capable of generating a robust evoked potential.

The model for pyramidal cells just described is generally assumed by one of the most extensively used approaches for identifying cortical generators of AEPs. The approach involves (1) measuring the AEP at many finely spaced locations over the head (e.g., 100 or more), (2) modeling the underlying cortical activity using the dipole model suggested by the morphology and dendritic orientation of pyramidal cells, (3) assuming an electric model for the head, and (4) determining the strength, location, and orientation of dipole(s) that yield a best fit between the measured AEP and AEP calculated based on the head and dipole models. The location of AEP generators estimated by this approach can be further refined using information from other imaging modalities, including structural and functional magnetic resonance imaging (MRI). The results are vivid spatial mappings of cortical activity unfolding over time.

Magnetic Analog of the AEP

When neurons responding to sound produce an AEP, they also produce a magnetic field recordable outside the head. While small (about 100 000 times less than the magnetic fields in our everyday surroundings), these auditory evoked magnetic fields (AEMFs) can be measured routinely when specialized, low-noise magnetic detectors are used. Both AEPs and AEMFs are produced directly by neuronally generated currents in the brain. This distinguishes AEP and AEMF recordings from other modalities for measuring brain function, such as functional MRI (fMRI) and positron emission tomography (PET), which detect hemodynamic changes accompanying changes in neuronal activity, rather than neuronal activity itself.

While both AEPs and AEMFs are a direct reflection of neuronal activity, the two types of recordings differ and are, in many respects, complementary. For instance, AEMFs, but not AEPs, are preferentially sensitive to cortical activity in pyramidal cells having a particular dendritic orientation relative to the surface of the head (dendrites parallel to the head surface, creating what is called a tangential dipole source). Much of the gray matter on the human superior temporal lobe contains pyramidal cells with this orientation, including the crown of Heschl's gyrus, where primary auditory cortex is typically located, and planum temporale, which contains nonprimary areas. Because the orientation of auditory cortex is well suited for magnetic recordings, AEMFs have been used extensively to investigate human auditory cortical function.

Other, Specialized AEPs

In order to obtain a specific indicator of activity in limited neuronal populations, responses are sometimes derived from multiple AEP measurements. The binaural difference potential (BDP) illustrates the selectivity provided by a derived response; it specifically reflects activity involved in comparing sounds between the two ears. A BDP is derived by summing the ABRs produced by left and right monaural stimulation and subtracting the ABR produced by stimulating both ears simultaneously. For click stimuli, the summed monaural and binaural ABRs show no difference during waves I–III, but diverge during waves IV/V. This difference is the BDP. A BDP would not be detected if sounds at the two ears were processed independently; the BDP reflects an interaction between the two ears. The coincidence of the BDP with the ABR indicates that it is generated by brain stem neurons.

Animal lesion data and the prominence of a particular binaural nucleus in the brain stem of humans, the medial superior olive (MSO), have led to suggestions that MSO neurons contribute substantially to the human BDP.

While AEPs can be elicited by discrete stimuli presented in silence, they can also be elicited by brief changes in ongoing sound. One example is the lateralization shift response (LSR), a cortically generated potential produced by changing the interaural time difference of binaurally presented continuous noise. A paradigm for eliciting this response involves initially presenting identical noise to the two ears (the noise is heard diffusely, but centered in the head), then, over a brief interval, temporally shifting the noise at one ear relative to other (by ~1 ms), and then quickly shifting it back (the diffuse, centered noise is still heard without interruption, but a brief sound lateralized to the ear with the leading noise is also heard). The result is an evoked response 100–300 ms after the shift, the LSR. Since the noise presented to either ear alone continues without disruption, it does not produce an AEP. The LSR only occurs when the noise is presented to both ears. Thus, like the BDP, the LSR specifically reflects binaural processing within the auditory central nervous system, although at a cortical rather than brain stem stage.

In addition to depending on the physical features of the stimuli used to evoke them, evoked potentials can also depend on the context in which the stimuli are presented. An example of a context-dependent AEP is the mismatch negativity (MMN). An MMN is elicited in a paradigm that involves presenting a standard stimulus (e.g., a tone burst) at regular intervals but, every so often, replacing the standard with a deviant sound (e.g., a tone burst of slightly different frequency; also called a rare stimulus). Responses to the deviant sound presented in the context of this paradigm (called an oddball paradigm) are different from the responses to the same stimulus presented over and over again on its own. The difference is the MMN (difference between the solid trace, which is the stimulus presented on its own, and the dashed trace, which is the stimulus presented as deviant, in Figure 1). The MMN, generated largely by the auditory cortex, can be thought of as a form of sensory memory, 'sensory' because it originates within cortex that specializes in encoding one of our senses, and 'memory' because a history of the stimuli appears to be retained by cortex; each incoming sound is compared to its predecessors and an MMN is produced when they differ. An MMN can be recorded even when individuals are distracted from the stimuli (e.g., by watching a movie), indicating that the MMN reflects, at least in part, neuronal processing prior to conscious perception.

AEP Correlates of Auditory Perception

Since AEPs can be recorded in awake people while they listen to sounds, they offer a way to understand the neural processing that determines how sounds are perceived. Two examples of this important role for AEPs follow.

The first example concerns the binaural difference potential and its correlation with certain perceived aspects of binaurally presented clicks. When clicks of equal intensity are presented simultaneously to the two ears, they are perceived as a single, fused object that is centered with respect to the head. As the clicks presented to one ear are made greater in intensity or shorter in latency, compared to the clicks presented to the other ear, the perceived object moves to the side of the more intense or leading clicks. As the interaural click delay is increased beyond about 1 ms, the object stops moving, and eventually splits in two such that two separate clicks are perceived. Studies have compared these perceptual properties with measurements of the BDP in human listeners. They showed that (1) the BDP was only detectable when the clicks at the two ears were fused into a single object and (2) the latency of the BDP covaried with the perceived location of the object. These correlations between percept and BDP indicate that the neuronal activity underlying the BDP contains a representation of both the fusion and the location of binaurally presented sounds. This implies that binaural fusion and perceived sound location may be at least partially encoded in the brain stem, where the BDP is generated, and, interestingly, well before cortical processing stages, where the activity dictating our perceptions is usually assumed to reside.

The second example concerns the N1 of the LLR and its correlation with the perceived rate of successive sounds and the segregation of sounds into distinct perceptual streams. A nice illustration of stream segregation, and the changes in perceived rate that sometimes accompany it, is provided by a sequence of tone bursts alternating in frequency (ABAB…, where A and B denote different frequencies). When the frequency difference between tones is small, a single, high-rate sequence is heard. However, when the frequency difference is large, the tones perceptually segregate into two lower rate sequences, comprising the high- and the low-frequency tones, respectively. For this and similar tone sequences, N1 evoked by the tones changes with increasing frequency separation in a manner predictable from the perceived rate and typical AEP rate dependencies described earlier (see the

section titled 'Dependencies on stimulus parameters'). In particular, N1 is greater for large frequency separations (perceived rate is low), and less for smaller ones (perceived rate is high). Especially interesting are the changes in N1 that occur when there is a change in percept without a corresponding change in physical stimulus. Such changes can be seen using tone sequences with an intermediate frequency separation, chosen such that the percept spontaneously fluctuates between 'one high-rate stream' and 'two low-rate streams.' In experiments selectively averaging N1 during each of these two perceptual conditions, N1 (or more specifically, its AEMF analog) was found to be greater during the perception of the low-rate streams. In other words, N1 was again greater during the perception of low-, as compared to high-rate sounds, but this time the difference in N1 occurred in the absence of physical alterations to the stimulus. The implication is that some of the neural activity underlying N1 specifically encodes the percept elicited by sound, rather than physical attributes of sound.

These examples are two of many in which auditory evoked responses have been used to identify, and study, the neural activity leading to human sound perception.

See also: Auditory Cortex Structure and Circuitry; Auditory System: Central Pathways; Evoked Potentials: Recording Methods; Sound Localization: Neural Mechanisms.

Further Reading

Burkard RF, Don M, and Eggermont JJ (2006) *Auditory Evoked Potentials: Basic Principles and Clinical Application.* Philadelphia: Lippincott Williams & Wilkins.

Butler RA (1968) Effect of changes in stimulus frequency and intensity on habituation of the human vertex potential. *Journal of the Acoustical Society of America* 44: 945–950.

Cohen D and Cuffin BN (1983) Demonstration of useful differences between magnetoencephalogram and electroencephalogram. *Electroencephalography and Clinical Neurophysiology* 56: 38–51.

Ridgway SH, Bullock TH, Carder DA, et al. (1981) Auditory brainstem response in dolphins. *Proceedings of the National Academy of Sciences of the United States of America* 78: 1943–1947.

Dale A and Halgren E (2001) Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. *Current Opinion in Neurobiology* 11: 202–208.

Furst M, Levine RA, and McGaffigan PM (1985) Click lateralization is related to the B component of the dichotic brainstem auditory evoked potentials of human subjects. *Journal of the Acoustical Society of America* 78: 1644–1651.

Gutschalk A, Micheyl C, Melcher JR, et al. (2005) Neuromagnetic correlates of streaming in human auditory cortex. *Journal of Neuroscience* 25: 5382–5388.

Hall JW III (1992) *Handbook of Auditory Evoked Responses.* Boston: Allyn and Bacon.

Hari R (1990) The neuromagnetic method in the study of the human auditory cortex. In: Grandori P, Hoke M, and Romani GL (eds.) *Auditory Evoked Magnetic Fields and Electric Potentials*, pp. 222–282. Basel: Karger.

Llinás R and Nicholson C (1974) Section IV. Analysis of field potentials in the central nervous system. In: Rémond A (ed.) *Handbook of Electroencephalograpy and Clinical Neurophysiology,* vol. 2B, pp. 61–83. Amsterdam: Elsevier.

Melcher JR and Kiang NYS (1996) Generators of the brainstem auditory evoked potential in cat III: Identified cell populations. *Healing Research* 93: 52–71.

Melcher JR (1996) Cellular generators of the binaural difference potential in cat. *Hearing Research* 95: 144–160.

Näätänen R (1995) The mismatch negativity: A powerful tool for cognitive neuroscience. *Ear & Hearing* 16: 6–18.

Näätänen R and Picton T (1987) The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology* 24: 375–425.

Nunez P and Srinivasan R (2005) *Electric Fields of the Brain,* 2nd edn. New York: Oxford University Press.

Picton TW, McEvoy LK, and Champagne SC (1991) Human evoked potentials and the lateralization of a sound. *Acta Otolaryngology (Stockholm)* 491(supplement): 139–144.

Auditory Localization

M Konishi, California Institute of Technology, Pasadena, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Basic Concepts and Terms

Complex sounds consist of simple 'spectral' components, which are represented by sine or cosine functions. Different spectral components are distinguished by frequency, amplitude, and phase (Figure 1). Amplitude is also referred to as intensity or level in psychoacoustic literature. Phase is a circular variable ranging from 0° to 360°. The time to go from 0° to 360° is called the period (T). The number of periods in a second is called frequency (F), so that $F = T^{-1}$. The mammalian and avian inner ear breaks down complex sounds into their frequency components, and primary auditory neurons carry information about the amplitude and phase of each frequency band. The narrow range of frequency to which a neuron responds is not coded by the number of nerve impulses *per se* but by its site of innervation on the inner ear basilar membrane. The sound intensity of each frequency band is coded by the rate of nerve impulses emitted by each neuron tuned to that frequency range. The phase of each frequency is coded by 'phase locking,' in which each primary auditory neuron fires impulses at or near a specific phase angle, such as 90° (Figure 2).

Many animals can determine the direction from which sound signals arrive. The two main cues for sound localization are the interaural time difference (ITD) and the interaural level difference (ILD). An ITD results when a particular point or phase angle of a sound wave reaches one (near) ear earlier than the other (far) ear (Figure 3). When a source is located directly in front, sound reaches both ears simultaneously. The ITD is zero in this case. As the source moves to one side, sound arrives earlier at the ear on that side and later at the other ear. Such a difference can occur in different aspects of sound waves, including sound onset times, phase angles, and points on the pattern of amplitude modulation (cf. Figure 2). Of these potential cues, the use of phase differences by humans and animals has been most extensively studied. Note that the same ITD corresponds to different binaural phase disparities in different frequency channels. An ILD results from a greater loss of sound energy by distance, absorption, and reflection en route to the far ear than to the near ear.

Behavioral Studies of Auditory Localization

The simplest way to test one's ability to localize sounds is to use earphones. When the ITD is zero in signals delivered by earphones, humans perceive not two but a single phantom source at the midpoint between the ears or sometimes straight ahead outside the head. The brain fuses the two separate signals into one internal representation. As the ITD is varied, the source appears to shift toward the ear at which the signal or waveform arrives earlier. Humans are also very sensitive to differences in the quality of sound signals between the two ears. The clarity and size of a perceived phantom source depend on the degree of similarity between the sounds. Similarity is measured by the degree of correlation between the sounds in the two ears. A sound signal on a television screen looks like a complex curve across the screen. If the same curve is drawn twice, two curves obviously match point by point, that is, they are perfectly 'correlated.' Binaural correlation is 1 when the signal waveform is identical in the two ears. Addition of random noise to correlated signals reduces the degree of binaural correlation. The amount of difference between correlated signals and random noise determines the degree of correlation. Partially decorrelated signals produce a blurred source, and the blur increases as the signals are further decorrelated. In one series of experiments, human participants used an electronic delay device to bring a sound source to the midpoint between the ears from an arbitrary initial locus. As the degree of correlation became smaller than about 0.3, the participants' choice of delays became more variable; that is, the standard deviations of the mean delays chosen by the participants increased.

Finding an animal model for human sound localization has not been easy, because most laboratory animals are difficult to train to perform perceptual tests such as those mentioned above. The barn owl is an ideal animal for the study of sound localization, however, because both its brain and its behavior are organized to perform this task. The owl uses its ears to pinpoint prey in total darkness. Owls listening through specially made earphones rapidly turn their heads in the direction that is predictable from the ITD contained in the signals. Thus, owls also perceive a single phantom source from separate sounds delivered to the two ears. Tests with partially decorrelated noise signals in owls also showed that the standard deviations of mean localization scores remained almost

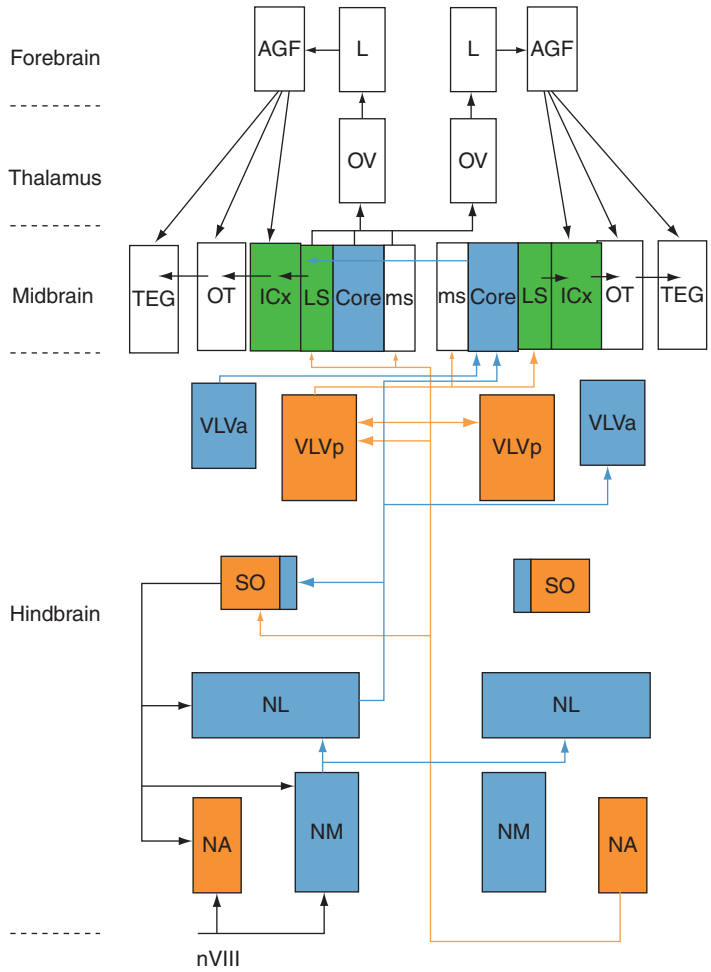

Figure 1 Simple acoustic terms. Complex sounds such as human voice consist of simple components, which can be represented by sine and cosine waves. The amplitude of a sine wave such as shown in this figure varies periodically as a function of time, which is expressed in terms of phase, because the amplitude changes cyclically. The period is the time to go from 0° to 360°. The number of periods or cycles per second is called frequency.

constant until the degree of correlation decreased to about 0.3. The curves for the standard deviations closely resembled those obtained for humans. The necessity for binaural correlation suggests that the brain needs it to compute the ITD from waveforms (phase) instead of signal onset times (cf. Figure 2). The theory of binaural cross-correlation assumes that the human brain performs a process similar to mathematical cross-correlation to compute the ITD. The results obtained in owls are also consistent with this theory.

A simple relationship between ITD and location holds when signal bandwidth is broad. With narrow-band signals such as tones, both humans and owls may perceive sounds coming from directions other than those of real sources. This phenomenon is called 'phase ambiguity.' If the sound incidence angles encoded by ITDs are known, phase-ambiguous ITDs are predictable from the following relationship: the true ITD and phase-ambiguous ones ITD $\pm nT$, where T is the period of the stimulus tone and n is an integer. Thus, if the tonal frequency is 8 kHz and $n = 1$, T is $1/8000 = 125\,\mu s$. Because the range of ITDs that barn owls experience is from 0 to about 150 µs, they have to distinguish only 0 from 125 µs because its multiples ($n = 2$, etc.) on both sides of the midline are outside the 0–150 µs range. Neither humans nor owls perceive illusory sound sources when signal bandwidth is broader than a certain value, which is 300 Hz in humans and 3 kHz in owls. This difference is correlated with the frequency range in which each species can use the ITD for localization. Owls can use the ITD for sound localization at frequencies as high as 8.5 kHz, whereas the highest frequency for humans is about 1200 Hz.

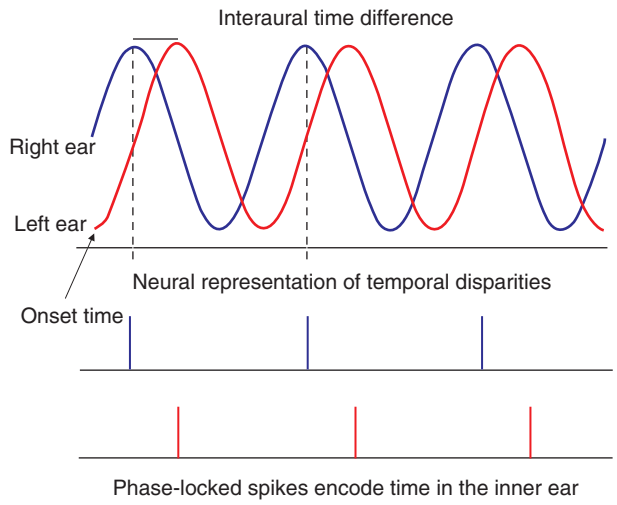

Figure 2 Interaural time difference and phase locking. Phase or time differences between the two ears are important cues for sound localization. Primary auditory neurons of many animals produce impulses at or near a particular phase angle. This phenomenon is called phase locking. This figure shows a highly simplified example in which each neuron fires exactly at 90°. The horizontal lines indicate time, and the vertical bars on them represent nerve impulses. Since a single neuron cannot fire more than about 800 impulses per second, the phase of a high-frequency tone is thought to be encoded by multiple neurons. The height of waves is called amplitude. Differences in amplitude between the two ears, as in this figure, are also available for sound localization.

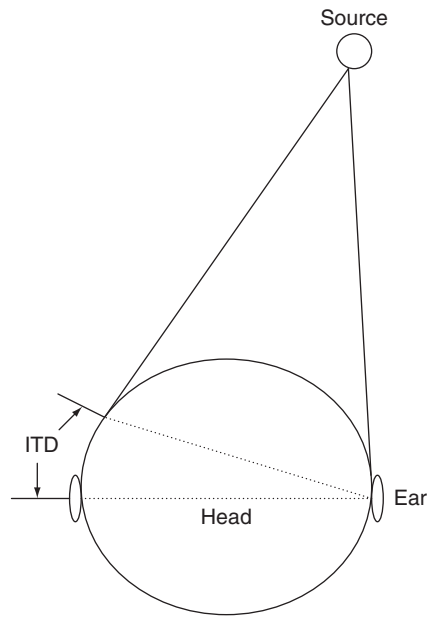

Figure 3 Interaural time differences. An interaural time difference results when a particular point of a sound wave reaches one (near) ear earlier than the other (far) ear. The reference point can be a sound onset time, a phase angle, or a point on the pattern of amplitude modulation.

Humans can use both the ITD and the ILD for localization in the horizontal plane, but barn owls use the ILD for the vertical plane. Owls do not have any external ear, but the feather ruff that surrounds their face serves as a sound-collecting device. The facial ruff consists of left and right halves that are separated by a tall ridge of feathers along the midline of the face. In barn owls, the left ear opening is located higher in the ruff than the right one. This asymmetry and other differences between the left and right halves of the ruff make the left and right ears more sensitive to sound coming from below and above eye level, respectively. This difference in the directional sensitivity of the ear causes the ILD to vary as a function of source elevation. In owl species with symmetrical ears, the ILD does not change along the vertical plane. Given a constant size of the head and sound collectors, shorter wavelengths and thus higher frequencies produce greater intensity differences than do lower frequencies. The barn owl needs relatively high frequencies (5–8 kHz) for accurate localization in elevation. Thus, to localize sound in two dimensions, the owl must be able to derive both interaural cues from the same high-frequency signal. Mammals including humans, which use both cues for azimuth, detect phase-based ITD only in low-frequency ranges and ILD in high-frequency ranges. Humans can also use monaural spectral cues for localization in the vertical plane.

Parallel Brain Pathways for Processing ITD and ILD

The owl's auditory system processes the two cues in separate pathways, and a similar design has been suggested for humans (Figure 4). Birds have two anatomically separate hindbrain areas on each side called nucleus angularis and nucleus magnocellularis. Each primary auditory fiber divides into two branches, one projecting to nucleus angularis and the other to nucleus magnocellularis. Neurons of nucleus magnocellularis 'phase lock' to each spectral component of complex sounds, whereas those of nucleus angularis generally do not. Thus, each cochlear nucleus is more specialized to process one cue than the other cue. The two pathways that start from the cochlear nuclei are anatomically separate until they converge in the lateral shell of the inferior colliculus (cf. Figure 4).

Detection and Coding of the ITD

Temporal disparities between left and right trains of phase-locked impulses provide the data for extracting

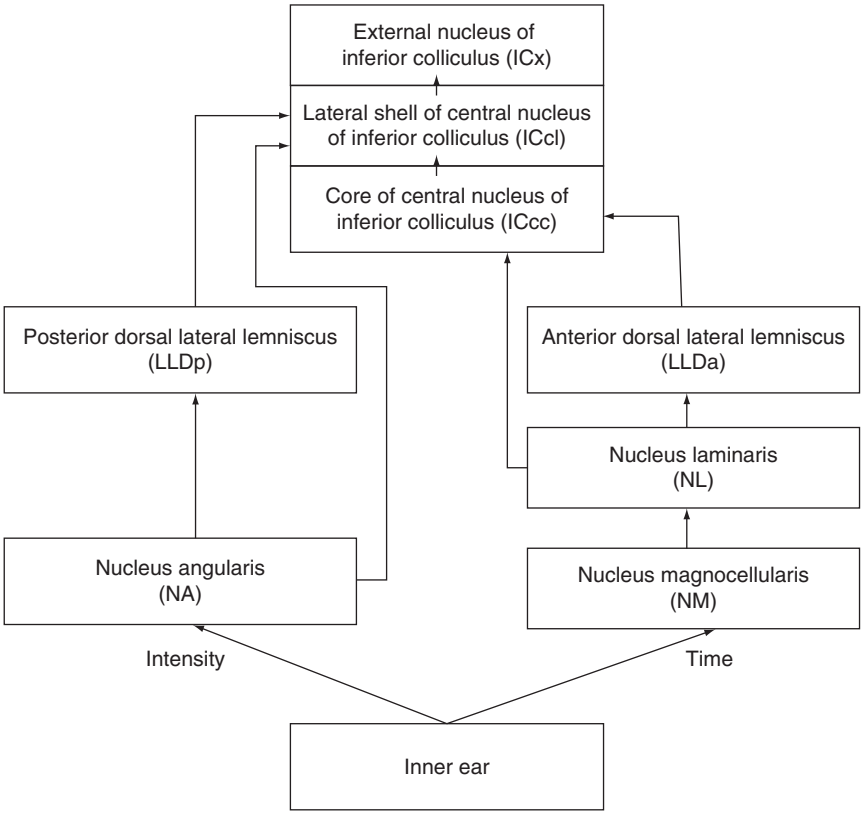

Figure 4 Parallel pathways for time and intensity. This figure shows only the brain areas whose role in sound localization is known and the connections that indicate ipsilateral or contralateral projections. The owl's auditory system processes interaural time differences (ITDs) and interaural level differences (ILDs) in separate parallel pathways. Each primary auditory fiber (nVIII) divides into two branches; one innervates cochlear nucleus angularis (NA) and the other, cochlear nucleus magnocellularis (NM). These two nuclei are the starting stations for the two pathways. Nucleus laminaris (NL), which receives inputs from the left and right magnocellular nuclei, is the first site for processing ITD. NL projects contralaterally to one of the lemniscal nuclei (LLDa) and the core of the central nucleus of the inferior colliculus. The core projects to the lateral shell (LS) on the contralateral side, where the two pathways meet. Different frequency bands converge on single neurons in the external nucleus of the inferior colliculus (ICx). This area contains a map of auditory space which is composed of neurons selective for combinations of ITD and ILD.

ITDs (cf. Figure 2). A model for the detection of ITDs uses delay lines and an array of coincidence detectors to measure the degree of match between the two trains (Figure 5). Loyd Jeffress was the first to propose this idea, in 1949, when little was known about the neurophysiology of the auditory system. He reasoned that the place of coincidence detectors in the array encodes the direction of sound sources. This model involves a transformation of codes from time to place, meaning that neither the rate nor the pattern of firing by single neurons encodes ITDs, although somewhere there must be mechanisms to compare the rates of discharge along the array. In the barn owl's auditory system, the axons from nucleus magnocellularis serve as delay lines, and the somata of nucleus laminaris serve as coincidence detectors (cf. Figure 5). The magnocellular axons convey phase-locked impulses to an array of laminaris cells with delays. These delays increase in the opposite direction for the ipsilateral and contrateral inputs because the length of axonal paths for the two sides increases toward the opposite ends of the array. When impulses travel to a laminaris cell by a longer path on one side than on the other side, they reach the cell at different times. In general, a left–right difference in impulse arrival times (Ta) consists of an ITD in the stimulus and a difference in impulse conduction delays (Tc). Hence, $Ta = ITD + Tc$. Coincidence means $Ta = 0$, which is obtained when $-ITD = Tc$; that is, ITD equals Tc in magnitude but is opposite in sign. Because the length of axonal paths varies systematically in both left and right inputs, Tc should also vary regularly. This argument is consistent with the finding that neurons sensitive to different ITDs are systematically arranged in the owl's nucleus laminaris. These neurons respond mostly to sounds coming from the contralateral side, as indicated in Figure 5.

Laminaris neurons are, however, not all-or-nothing coincidence detectors, for their responses decrease gradually from maximum to minimum as stimulus

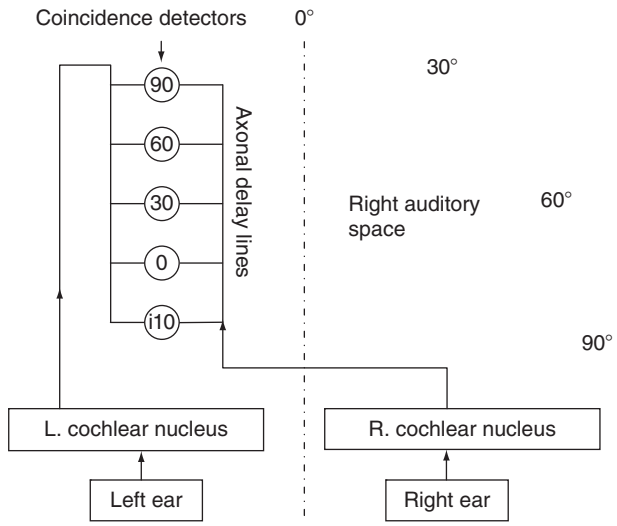

Figure 5 Neural encoding of interaural time differences in owls. The neural network that detects and encodes interaural time differences consists of cochlear nucleus (magnocelluaris) and nucleus laminaris on each side of the brain. The network consists of delay lines, which are axons of magnocellular cells, and coincidence detectors, which are laminaris neurons, as shown by 0, 30, and so forth, in circles. The coincidence detectors fire when impulses from the two sides arrive simultaneously. The length of axons to the coincidence detectors varies systematically from one end of the neuron array to the other. The longer the axon, the longer the time for nerve impulses to travel. The right and left sets of axons provide every possible pair of conduction delays, which range from 0 μs to about 150 μs in barn owls. An ITD of 90 μs, right side leading, must be delayed by 90 μs so that the neural impulses reach coincidence detectors simultaneously. In the owl, the left set of coincidence detectors encodes mostly sound locations on the right side. Thus, neuron 60 in the figure responds to sounds coming from 60° (corresponding to ~120 μs) on the right side. Jeffress used the term 'place coding' to emphasize the point that neither the number of nerve impulses nor the temporal pattern of their discharge encodes ITDs but the place of the coincidence detectors in the array. Therefore, if the neuron at 0 fires, the individual perceives a sound source straight ahead.

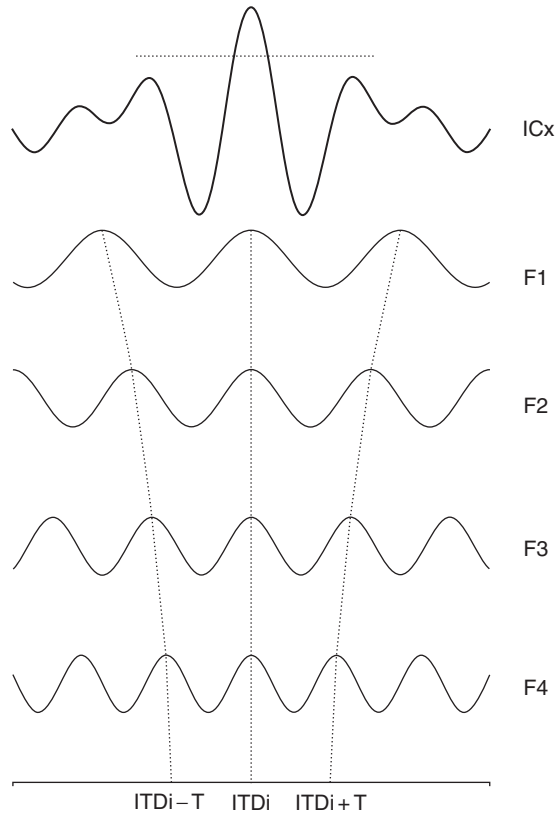

Figure 6 Detection of frequency-independent interaural time difference (ITDi). Single neurons in all stations below the external nucleus of the inferior colliculus (ICx) show phase ambiguity (responding to ITDi $\pm$ nT, where T is the period of the stimulus tone and n is an integer). Although ICx neurons also show phase-ambiguous responses to ITD carried by tones or narrow-band signals, they respond to a single ITD carried by broadband signals such as noises. An ICx neuron receives inputs from contralateral lateral shell of the central nucleus of the inferior colliculus (ICcl) neurons tuned to the same ITD from different frequency channels. The peak representing this ITD is the highest when ITD curves for different frequencies (F1, F2, etc.) are aligned and summed. This process can be seen by recording postsynaptic potential from an ITD neuron. The firing threshold (horizontal dotted line) of an ICx neuron appears to be set such that it is crossed only by the highest peak representing the frequency-independent ITD.

timing changes from perfect coincidence to out of phase by 180°. Laminaris neurons respond best to a single interaural time difference (ITDi) which is independent of frequency and its phase equivalents (ITDp). The relationship between these two variables is ITDp = ITDi $\pm$ nT, where T is the period of the stimulus tone and n is an integer. This phenomenon is the neural equivalent of phase ambiguity. Laminaris neurons fire maximally when they receive binaurally synchronous trains of phase-locked impulses. However, the coincidence of the two trains recurs every time ITD is changed by an integer multiple of the period of the stimulus tone, causing the laminaris neurons to discharge maximally. Nucleus laminaris projects both to the core of the central nucleus of the inferior colliculus (ICc) and to one of the lemniscal nuclei (the anterior dorsal lateral lemniscus (LLDa); cf. Figure 4). The core projects to the contralateral lateral shell of the central nucleus of the inferior colliculus (ICcl).

Resolution of Phase Ambiguity

Human psychoacousticians proposed a scheme that would solve the problem of phase ambiguity. ITD is independent of frequency, whereas ITD $\pm$ T varies with frequency because T changes with frequency. Therefore, comparison of ITD responses of neurons across different frequency bands should discriminate between frequency-dependent and -independent ITDs. How this comparison is carried out in the human brain has

been a subject of debate. In the owl, one can observe the process at the level of single neurons in the inferior colliculus (Figure 6). In the external nucleus (ICx), different frequency bands carrying ITD information converge on single neurons. We can explain how the convergence of different frequency bands results in the identification of the frequency-independent ITD. When ICx neurons receive signals from lateral shell neurons tuned to different frequencies, they gather input from those tuned to the same ITD. However, because the afferent neurons respond to ITD $\pm T$, they confer this property on the ICx neurons. When the ICx neuron adds the inputs (postsynaptic potentials) across all frequency bands, the peaks at the frequency-independent ITD add, whereas the peaks and troughs of F1–Fn partially cancel each other out. The ICx neuron fires impulses only when the sum of the inputs crosses the threshold, as shown by the horizontal dotted line in Figure 6. This linear summation across frequencies is not the only method that the owl's auditory system uses. Evidence indicates that nonlinear integration such as inhibition is also involved.

Processing of ILDs

The nucleus angularis, the first station in the intensity-processing pathway, projects to a lemniscal nucleus called the posterior dorsal lateral lemniscus (LLDp), in which all neurons are excited by sound in the contralateral ear and inhibited by sound in the ipsilateral ear (Figure 4). The excitatory input comes from the contralateral nucleus angularis, and the inhibitory input from the other LLDp. The louder the sound in the contralateral ear is, the greater the response, whereas the louder the sound in the ipsilateral ear is, the greater is the degree of inhibition. Thus, the response of an LLDp neuron varies as a function of the difference in sound level between the two sides, even though the neuron still responds to monaural stimulation of the contralateral ear alone. Furthermore, the threshold and degree of inhibition change systematically from the dorsal to ventral direction of LLDp. This gradient suggests a map of ILDs in this nucleus.

Convergence of Time and Intensity Pathways

The lateral shell of the inferior colliculus (LS) is the site of convergence between the time and intensity pathways (Figure 4). The LS receives direct inputs from nucleus angularis and LLDp and indirect inputs from nucleus laminaris and LLDa via the core of the inferior colliculus. LS neurons respond to combinations of ITD and ILD within a narrow range of frequencies, indicating that the convergence of the two pathways occurs initially in each frequency band. The LS projects to the adjoining external nucleus of the inferior colliculus (ICx), where neurons respond to broad frequency ranges (3–9 kHz). ICx neurons do not respond to ITD or ILD alone but require specific combinations of the two cues, and unfavorable combinations may induce inhibition.

A Map of Auditory Space

The tuning of ICx neurons to specific pairs of ITD and ILD makes these cells selective for the direction of sound propagation because such combinations are obtained only when sound comes from particular areas in an owl's auditory space. These areas are called receptive fields, and cells with such areas are called space-specific neurons. ICx neurons are systematically arranged according to the location of their receptive fields, thus forming a neural map of auditory space (Figure 7). The topographical representation of auditory space is due, not to anatomical projections of the inner ear, as with the retinotopic map of the visual field, but to central synthetic processes. Nevertheless, the auditory space map projects onto the retinotopic map of the optic tectum. The tectal neurons making up the bimodal map respond to both auditory and visual stimuli coming from the same area in space; that is, their auditory and visual receptive fields are mutually aligned. Similar maps are found in the superior colliculus of the ferret, although the mechanisms underlying these maps appear to differ from those in owls.

Owl species with symmetrical ears also have space-specific neurons in their ICx, but these cells map only azimuth, because their receptive fields do not have elevational boundaries, although the cells are tuned to combinations of ITD and ILD. In these owls, the ILD does not change as a function of sound elevation, because the highest frequency audible to them is too low for this change to occur.

A Motor Map for Sound Localization

The overt behavior that characterizes sound localization in owls is rapid head rotation. The bimodal map of space in the optic tectum projects to a motor map for gaze control, which is found in the optic lobe. Neurons of the motor map innervate the medullary nucleus involved in the control of head rotation. The tectal motor map encodes the direction, speed, and amplitude of head rotations. Experiments reveal that the map contains separate circuits for head rotations in azimuth and elevation.

Another new development in the study of sound localization in owls concerns the role of the forebrain

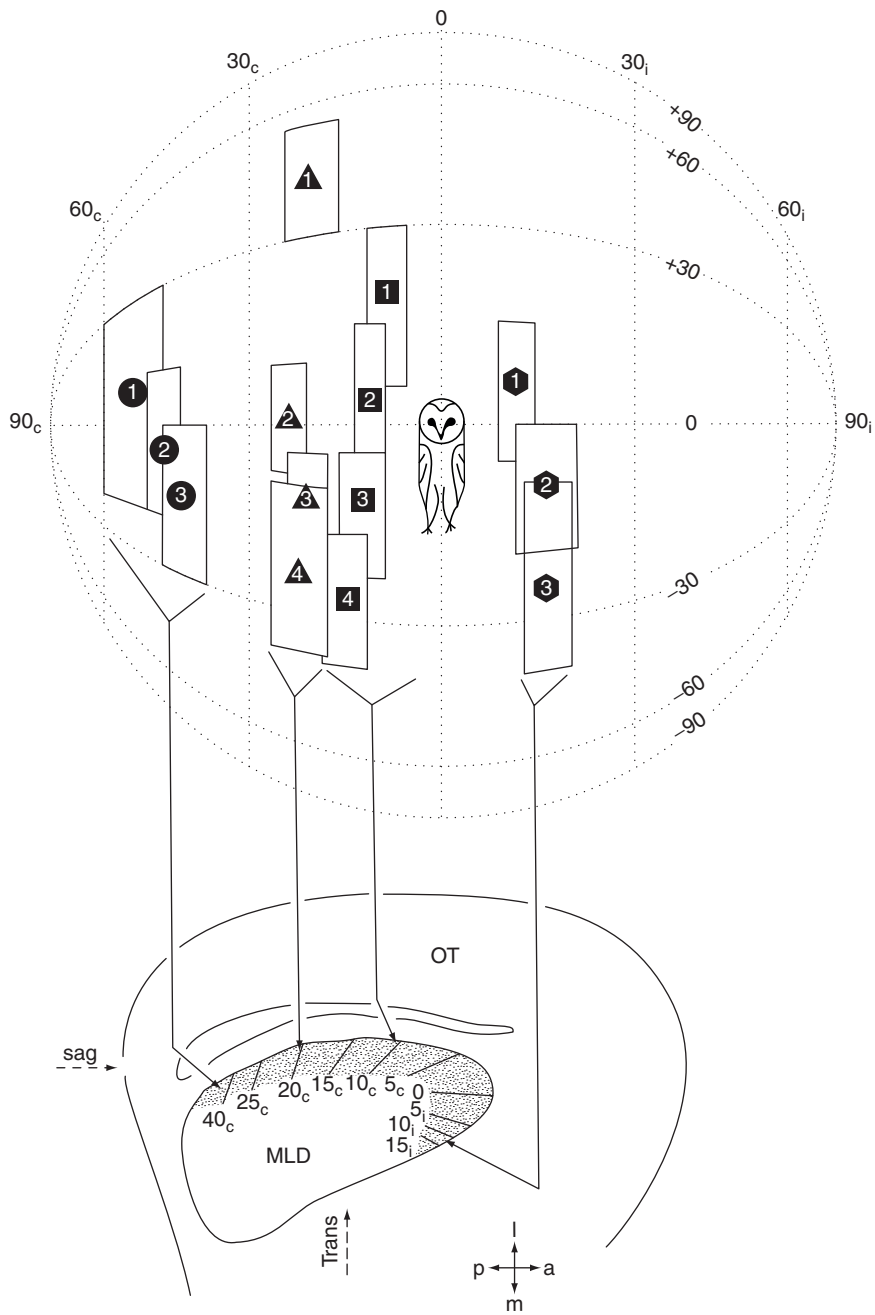

Figure 7 A neural map of auditory space. Space-specific neurons have spatial receptive fields the centers of which are shown by rectangular boxes with numbers indicating the sequences in which the neurons were recorded during a single electrode penetration. The shaded area in the bottom panel is the external nucleus of the inferior colliculus (ICx). The map represents mostly the contralateral (c) side of the auditory space, although a small extension to the ipsilateral (i) side is shown. The auditory space map is not due to topographic projections of sensory surfaces such as the body surface, the retina, and the cochlea but is created by means of neural computations. Nevertheless, the owl's auditory map projects to the optic tectum (OT) to join the topographically made map of the retina to form a bimodal map of space in which each neuron responds to both light and sound coming from the same direction.

in this behavior. Lesions of the auditory space map affect sound localization for a while, but the owl's ability to localize sound persists. Ablations of both the map and the nucleus ovoidalis, the avian homolog of the medial geniculate body, completely abolish the ability of owls to localize sound. The nucleus ovoidalis receives inputs from all parts of the inferior colliculus except ICx. Many ovoidalis neurons are tuned to combinations of ITD and ILD. Nucleus ovoidalis projects to the forebrain auditory area field L, which in turn projects to the archistriatum (new term: acropallium). This structure contains many neurons tuned to combinations

of ITD and ILD. But neither field L nor archistriatum neurons form an auditory space map. Focal electrical stimulation of the archistriatum induces head rotations even when the tectal motor map is disabled. It is interesting that lesions of the archistriatum cause owls to turn their head toward sound sources like an automaton. Barn owls determine and memorize sound locations before orienting toward them. When the archistriatum of barn owls is inactivated with a drug, the owls can no longer remember the computed sound location, even though they can still localize the sound source if it continues to emit the signal. Thus, one of the functions of the owl's archistriatum may be storage of working memory.

See also: Auditory Scene Analysis; Auditory System: Central Pathway Plasticity; Auditory System: Giant Synaptic Terminals, Endbulbs and Calyces; Auditory Cortex: Models; Localizing Signal Sources; Sound Localization: Neural Mechanisms.

Further Reading

Blauert J and Lindemann W (1986) Spatial mapping of intercranial auditory events for various degrees of interaural coherence. *Journal of the Acoustical Society of America* 79: 806–813.

Carr CE and Konishi M (1990) A circuit for detection of interarual time differences in the brain stem of the barn owl. *Journal of Neuroscience* 10: 3227–3246.

Du Lac S and Knudsen IE (1990) Neural maps of head movement vector and speed in the optic tectum of the barn owl. *Journal of Neurophysiology* 63: 131–149.

Jeffress LA (1949) A place theory of sound localization. *Journal of Comparative and Physiological Psychology* 41: 35–39.

Jeffress LA, Blodgett HC, and Deatherage BH (1962) Effect of interaural correlation on the precision of centering a noise. *Journal of the Acoustical Society of America* 32: 1122–1123.

Knudsen IE (2002) Instructional learning in the auditory localization pathway of the barn owl. *Nature* 417: 322–328.

Konishi M (2000) Study of sound localization by owls and its relevance to humans. *Comparative Biochemistry and Physiology. A, Comparative Physiology* 126: 459–469.

Konishi M (2003) Coding of auditory space. *Annual Review of Neuroscience* 26: 31–55.

Peña JL and Konishi M (2001) Auditory spatial receptive fields created by multiplication. *Science* 292: 249–252.

Saberi K, Takahashi Y, Farahbod H, and Konishi M (1999) Neural bases of an auditory illusion and its elimination in owls. *Nature Neuroscience* 2: 656–659.

Saberi K, Takahashi Y, Konishi M, Albeck Y, Arthur BJ, and Farahbod H (1998) Effects of interaural decorrelation on neural and behavioral detection of spatial cues. *Neuron* 21: 789–798.

Stern RM and Trahiotis T (1997) Models of binaural perception. In: Gilkey RH and Anderson TR (eds.) *Binaural and Spatial Hearing*, pp. 499–531. Mahwah, NJ: Erlbaum.

Takahashi T and Konishi M (1986) Selectivity for interaural time difference in the owl's midbrain. *Journal of Neuroscience* 6: 3413–3422.

Takahashi TT and Konishi M (1988) Projections of the cochlear nuclei and nucleus laminaris to the inferior colliculus of the barn owl. *Journal of Comparative Neurology* 274: 190–211.

Takahashi TT and Konishi M (1988) Projections of nucleus angularis and nucleus laminaris to the lateral lemniscal nuclear complex of the barn owl. *Journal of Comparative Neurology* 274: 212–238.

Volman SF and Konishi M (1990) Comparative physiology of sound localization in four species of owls. *Brain, Behavior and Evolution* 36: 196–215.

Auditory Scene Analysis

A S Bregman, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

The Scene Analysis Problem

In the natural world that surrounds us, and in which we evolved, it is rarely the case that only one sound is present at a time. What reaches our ears is a mixture of all the sounds present at a given moment: a voice speaking, the recorded music in the background, a car passing by, a bird singing, etc. If you ask people how they can hear an individual sound in this mixture, they say that they just focus their attention on whatever they want to listen to. This answer, however, ignores a fundamental difficulty. The mixture that reaches each of our ears is a single pressure waveform that is the moment-by-moment arithmetic sum of the pressure patterns that arise from individual events. This summed wave does not have written on it how many sounds contributed to it or how each sound is buried in it. Yet our auditory systems, and those of other animals as well, have the capacity to find the individual sounds in the mixture – a capacity called auditory scene analysis (ASA).

Every sound of finite duration can be thought of as the sum of a set of frequency components of different amplitudes and phases (a spectrum). An example of a mixture is shown in Figure 1 in the form of a spectrogram (showing time on the x-axis, frequency on the y-axis, and the intensity at any time-by-frequency position as darkness). This representation is relevant to ASA because there is evidence that the first stage in the neural processing of an incoming acoustic signal involves analyzing its frequency composition. The ASA problem is equivalent to finding a set of spectrograms, which when superimposed and summed, gives us the observed spectrogram. This decomposition would be made easier if natural sounds were compact in frequency or in time, but generally, they are not. Think of two voices heard at the same time. Far from being compact in frequency or time, the frequency components of the two are intermingled on both dimensions.

It should be evident that there are a virtually infinite number of ways in which this spectrogram can be decomposed into two or more component spectrograms, if the decomposition is not guided by principles. But what sort of principles? Engineers have used mathematical procedures, such as principal components analysis, that solve the ASA problem decisively, but only in very restricted circumstances. Since no single computation will always solve the problem in a broad range of environments, animals are forced to work with sets of principles that are called 'heuristics' because they are very helpful but are not guaranteed to always work. These heuristics exploit regularities in the world such as the following: since it is unlikely that two unrelated sounds in the world will start at exactly the same time, if the auditory system registers a set of frequency components whose onsets are approximately simultaneous, it is highly probable that they are parts of the same sound and therefore should be grouped together.

One cannot be sure that every type of animal solves the problem in the same way. Some undoubtedly have specialized mechanisms to extract critical sounds (for detecting prey, predators, or mates in the mixture). The specialized mechanisms may well be neural circuits that act as detectors for the required information, responsive to certain time and frequency relations, and unresponsive to sounds that lack these relations. However, larger-brained animals deal with sound in much more complicated ways, and can learn about the dangers and affordances of new sounds, but only if they can extract them from mixtures. For these animals, including ourselves, it is of great value to have general methods for decomposing a mixture into its component sounds, regardless of whether the latter are familiar or not.

Primitive and Schema-Based Processes

There seem to be at least two types of brain mechanisms involved in the grouping of information. The first is a set of primitive (unlearned) bottom-up processes that is probably shared with nonhuman animals. As described by Albert Bregman, in his book *Auditory Scene Analysis*, these processes group the information by using similarities and discontinuities in the signal, similar to those described by Gestalt psychology in their study of vision.

The second type of mechanism consists of a set of brain processes (called schemas) for dealing with frequent and important patterns in the environment. They may be completely, or in part, innate, but large-brained animals such as humans can modify them and develop new ones through learning. These are the mechanisms that permit recognition of conspecific animal sounds, familiar words, melodies, etc., and probably assist in segregating these patterns from their backgrounds. Schemas are extremely numerous in humans, numbering at least in the hundreds of thousands for a particular adult (consider just the schemas for the tens of thousands of

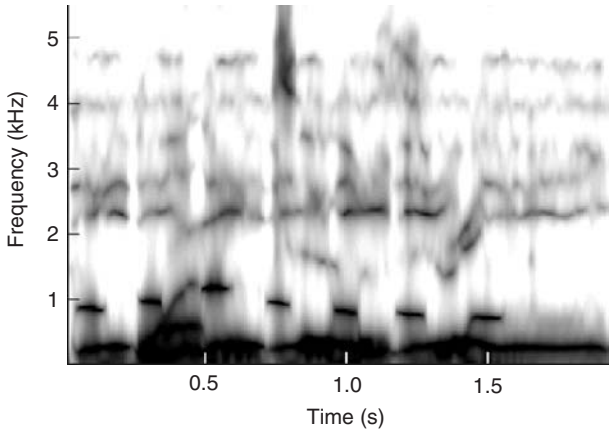

Figure 1 Spectrogram of a mixture of four sound sources: the words, "one, two, three"; a voice singing "da-da-da"; a person whistling; and a computer fan. Reproduced from Bregman AS and Woszczyk W (2003) Controlling the perceptual organization of sound. In: Greenebaum K and Barzel R (eds.) *Audio Anecdotes: Tools, Tips, and Techniques for Digital Audio*, pp. 35–66, with permission from A K Peters, Ltd.

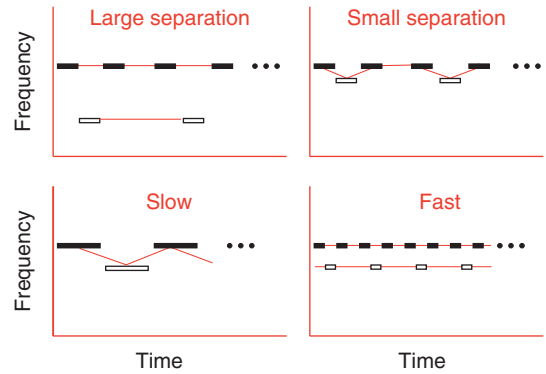

Figure 2 Diagram of four conditions in the galloping pattern of a higher tone (H) and a lower one (L) played repetitively in a galloping pattern: HLH–HLH–HLH–···. The connecting lines indicate the perceived streams. The top two panels show the effects of frequency separation; the bottom panels show the effects of speed. Reproduced from Bregman AS and Woszczyk W (2003) Controlling the perceptual organization of sound. In: Greenebaum K and Barzel R (eds.) *Audio Anecdotes: Tools, Tips, and Techniques for Digital Audio*, pp. 35–66. with permission from A K Peters, Ltd.

words that an adult can recognize). They can operate in conjunction with attention, as when we are trying to listen to a familiar voice in a crowded room, or prior to attention, as when, in that same room, our own name pops out of the background sound, attracting our attention. Despite the probable importance of schema-based perceptual processes in isolating familiar patterns from their contexts, the present discussion focuses on the primitive processes of ASA.

ASA as Grouping

The decomposition of the information in the mixed spectrogram of Figure 1 can be viewed as a problem of grouping. This grouping has two aspects, simultaneous and sequential. Simultaneous grouping determines which parts of the complex information, presented simultaneously to the senses, should be allocated to the same description of an environmental event. One can think of this as sharing out the energy of the spectrum that is present at a given moment. In the spectrogram, this represents the grouping of energy on the *y*-dimension, and even on the *z*-dimension (intensity), since all the energy at a single frequency-by-time location may not have come from a single environmental sound.

Sequential grouping is responsible for answering the question as to which spectral components should be connected up over time (the horizontal dimension of the spectrogram) to yield a representation of a distinct train of events in the environment. The two types of groupings interact, but it is convenient, for exposition, to describe them one at a time.

Sequential Integration and Segregation

A laboratory phenomenon called stream segregation, or streaming, illustrates sequential integration. The reader can view an illustration of a pattern of sounds in Figure 2. The stimulus is a rapid cycle composed of two tones, a higher-frequency one (H) and a lower-frequency one (L), formed into triplets, separated by short silences: HLH–HLH–HLH–(repeated). The H and L tones are far apart in frequency. This cycle gradually speeds up. At slow rates, the HLH triplets are heard as repeating units with a galloping rhythm; but as the sequence gets faster, the high tones seem to segregate from the low ones and we hear two parallel sequences, one consisting of a sequence of high tones and the other, a slower sequence of low tones. These two streams of sound seem to be going on at the same time, yet are perceived independently. Attention seems to be enable to focus on one stream or the other, but not on both at the same time. When the high and low tones are close to one another in frequency, the perception of the gallop persists even at higher speeds, and the sequence remains as a single coherent stream. This shows the importance of frequency separation, as well as speed, in causing the H and L tones to form separate streams. These effects are illustrated in Figure 2. This stream formation is a form of ASA. When the streams are segregated, the system is betting that there are two sources of sound in the environment, not one. While sequential grouping of the sensory data in a complex environment may not operate in as simple a way as in the streaming phenomenon of the laboratory, it is thought that this

phenomenon is a glimpse of the sequential process of ASA in a pure form. Consequently, this artificial version of ASA has been used extensively to study sequential grouping.

The streams need not be as simple as those of Figure 2. One can interleave the notes of a melody with distractor tones, destroying our ability to recognize it. However, if the melody and the distractors are separated in pitch range, one can hear each sequence in a separate auditory stream. Incidentally, the number of concurrent streams is not restricted to two: the upper limit is unknown. If one uses only musical notes, varying only their pitch, the limit seems to be three or four. However, if one were to add quite different sounds to the mixture, such as the clicking of a clock, the ringing of a telephone, spoken digits, and so on, the number would undoubtedly be higher.

Factors Contributing to Sequential Segregation

We can think of the tones in each panel of Figure 2 as laid out on a two-dimensional surface (time by frequency) and can imagine that grouping occurs as a result of relative proximity on this surface. When the frequency separations are small and the time separations large (at low speeds), the time dimension – having the greater range of values – will dominate the grouping, and tones will group with their nearest temporal neighbors, regardless of the small frequency differences. However, when the temporal separations are small (high speeds) and the frequency separations large, then frequency differences will be dominant and tones will group with tones that are close in frequency. It is as if the auditory system forms clusters in the frequency-by-time space that minimize the distances within clusters and maximize the distances between them.

However, two dimensions, for example, the frequency and time separations of pure tones, are not enough to represent all possible differences between sounds. Other differences also contribute. The segregation of complex (as opposed to pure) tones can be based on (1) differences in fundamental frequency F_0 (even when the harmonics are restricted to the same frequency range); (2) differences in spectral shape, with F_0 held constant, perceived as differences in timbre; (3) differences in spatial location; (4) differences in intensity; and (5) differences in amplitude envelope (e.g., rise and fall times).

The effects of these differences combine. For example, if, in a rapid sequence of alternating A and B tones, A and B are different in two ways, say in timbre and in spatial location, they will segregate more readily than if different in only one of these ways. It is as if the best clusters are formed in a multidimensional similarity space, including time as one of its dimensions.

A separate factor that contributes to grouping is continuity. If changes from A to B in the repeating sequence ABAB . . . are abrupt, the A's and B's are more likely to form separate streams than if the properties of A smoothly transform into those of B. For example, if A and B differ in frequency, then frequency glides joining A and B tend to hold the tones together as a single stream of sounds.

Cumulative Effects

There is a gradual increase in the tendency of A and B to form their own streams with increased numbers of repetition of the A–B alternation. However, the alternation of two tones in a regular pattern is not a requirement for the buildup of such streams. Two sets of tones all different, but separated into two distinct frequency bands, can also segregate into separate streams. The segregation tendency dies away gradually during a silent gap and starts building up again after the silence.

It has been proposed by Stuart Anstis and Shinya Saida that the effects of repetition can be explained by the existence of frequency-transition detectors, whose function is to integrate successive tones into a single stream. Repetition of ABAB . . . transitions lead to the habituation of these detectors, so that they can no longer perform this integrative function. A problem with this theory is that it appears that any perceptible difference at all between two tones may promote their segregation into two streams; so the habituation theory would require the auditory system to have a very large number of types of transition detectors.

An alternative theory, functional, rather than physiological, argues that the default condition of grouping is to assign all incoming sounds to a single stream, but the repeated occurrence of tones in different frequency regions builds up evidence that the sounds are coming from two different sources and should be assigned to separate perceptual streams.

Van Noorden's Two Boundaries

Stream segregation seems to take two forms, shown by the two curves in Figure 3, based on the data of Leon van Noorden who studied segregation of the HLH–HLH– . . . pattern. The x-axis represents time between adjacent tones, while the y-axis represents the frequency separation of the H and L tones. Each of the two curves represents the time-by-frequency threshold between hearing one versus two streams. At frequency–time values above the curve one hears two streams, and below the curve one stream. However, there are two different curves. The upper

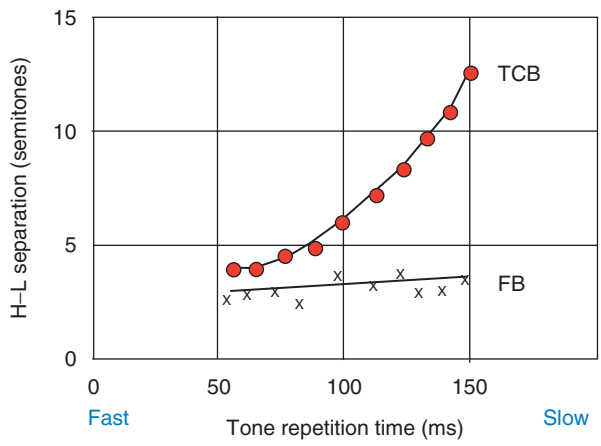

Figure 3 Curves by Van Noorden showing the thresholds for stream segregation. For each curve, the area above it shows the region (in frequency-by-time) in which two streams are heard; in the area below it, only one stream is heard. The temporal coherence boundary (TCB) is obtained when the listener is trying to hear all the tones as a single stream. The fission boundary (FB) is found when the listener is trying to hear the tones in two separate streams. The tonal pattern used was a galloping sequence of the form LHL–LHL–···, where L and H are lower- and higher-frequency tones, respectively. Adapted with the permission of LPAS Van Noorden.

one, the temporal coherence boundary (TCB), was obtained when the listeners were trying to hold onto a single-stream percept, and the lower one, labeled fission boundary (FB), was obtained when they were trying to hear two streams. Two facts are evident: the first is that when trying to hold onto a single stream they could do so at higher frequency separations and speeds than when they were not. This is not surprising. More remarkably, when they were trying to segregate the streams, the rate of presentation had a very small effect.

The different shapes of the two curves reveal the activity of two different mechanisms of segregation. When trying to hold onto all the tones as a single stream, the process that interferes with this goal is probably the primitive, bottom-up process of grouping. When trying to hear separate high and low streams, the process employed is one of selective attention, trying to hear either the high or the low tones – a top-down process, which is limited only by the listener's ability to discriminate the higher and lower tones at high speeds.

Effects of Sequential Grouping

When streams are strongly segregated from one another, the effects are numerous: (1) melodies and rhythms seem to include only within-stream sounds and (2) the temporal relations between concurrent streams become uncertain. Indeed, it appears that for any pattern of sound to be clearly perceptible, it must involve only the elements of a single stream.

Simultaneous Integration

Cues Favoring the Grouping of Simultaneous Components

One of the most important effects on the allocation of spectral energy to simultaneous sounds is 'harmonicity.' Many natural sounds including the vowel-like segments of human speech, and the pitch-possessing portions of the calls of other animals, have repetitive waveforms. So also do certain manufactured sounds, such as the sound of a violin. Repetitive waves have a pitch, and are composed of harmonics – frequencies that are integer multiples of the lowest frequency (the fundamental). The auditory system has mechanisms for detecting and grouping a subset of frequency components that are multiples of the same fundamental (a harmonic series). Furthermore, it can find more than one harmonic series at a time, allowing the listener to hear concurrent sounds that have different pitches. Both humans and computers depend strongly on harmonicity to segregate simultaneous components. For example, it is easier for us to segregate a man's voice from a woman's than from another man's, and computer models of the segregation of speech from other interfering sounds typically use the harmonic structure of the vocalic sounds as the main grounds for finding a stream of speech in the mixture.

Harmonicity is not the only basis for segregating concurrent sounds. Another important one is synchrony or asynchrony of onset. It is very likely that all the frequency components from a single environmental sound will start together, and it is very unlikely that the components of unrelated sounds will start at the same moment. So synchrony and asynchrony are used by the ASA system to determine whether or not frequency components should be allocated to the same sound. A third factor is the frequency separation of concurrent components. Components that are further apart are less likely to be treated as part of the same sound.

Another difference between components that is used by ASA is their difference in spatial location. This difference, by itself, does not powerfully group and segregate sets of concurrent components (although it is very powerful in sequential stream segregation). However, when other principles, such as asynchrony of onset, act to segregate concurrent components, spatial differences seem to multiply the strength of this segregation. It is surprising that spatial differences have such a limited effect in human ASA. Engineers use a technique called blind separation, in which the separation of voices is primarily or even

exclusively based on spatial separation. However, spatial information is not always reliable in natural environments (e.g., when the sounds are coming around a corner). Perhaps this is why spatial cues are not used by humans as a primary basis for grouping simultaneous components.

Achievement of Stability in the Face of Unreliable Acoustic Evidence

It is not only spatial cues that are unreliable. While harmonicity is a good cue, many sounds are not harmonic. There is no pitch involved in footsteps, accelerating cars, scratches, or bumps; yet these sounds are informative and we need to know how to group their components across the spectrum and over time. Because even the best cues are not always reliable, the ASA system assesses many cues and allows them to reinforce, or compete with, one another in controlling the decisions about grouping (as if the various cues could vote for their own preferred organizations).

Another method that the ASA system uses because cues are less than perfectly reliable is the conservative strategy of maintaining an existing interpretation until evidence piles up that it is wrong. The system seems to start off with the hypothesis that all acoustic input is part of a single sound. As evidence builds up, an organization in terms of a number of distinct sounds emerges. Maintaining a stable percept, despite the fact that cues can suffer interference, means not altering the grouping of the sounds upon encountering a brief drop in intensity, or a momentary change in interaural spatial cues, as the listener passes behind an obstruction. If each brief glitch caused a reorganization, our perceptions would be highly unstable. Hanging on to an existing organization is a valuable strategy in a world that is more stable than the cues about it are.

What, then, are the perceptual effects of simultaneous grouping – sometimes called fusion? The most obvious one is on the number of perceived sounds and the distinctness of their qualities. When segregated, we hear more distinct sounds, each with its own qualities of pitch, timbre, loudness, location, etc., whereas when fused, the full set of frequency components create the qualities of a single global unit, which, at any given moment, has only one pitch, one loudness, one spatial location, and one timbre.

Competition between Sequential and Simultaneous Grouping

Although we have been looking separately at the two major dimensions of grouping – sequential and simultaneous (the horizontal and vertical dimensions, respectively, of a spectrogram) – it is important to recognize that each of these affects the other. Often they compete. This competition is illustrated in Figure 4, a simplified spectrogram whose horizontal and vertical dimensions show time and frequency, respectively. A, B, and C are three pure-tone components. First A occurs alone, followed by B and C together, and this pattern is repeated cyclically. The BC spectrum can be interpreted as a complex tone with two frequency components or else as two simple tones, B and C, that happen to occur at about the same time. One can hear the total cycle as a two-tone stream formed of A and B, repeating over and over, accompanied by a one-tone stream, consisting of repetitions of C. Alternatively, one can hear a pure tone, A, alternating with a rich tone, BC.

This is a good stimulus for demonstrating the competition of sequential and simultaneous organizations. First, we can increase the sequential grouping of A with B by moving A closer to B in frequency. This grouping not only increases the AB sequential grouping, but also weakens the BC simultaneous grouping, so that the BC spectrum is heard as less rich. Conversely, if the simultaneous grouping of B with C is manipulated by altering the BC asynchrony, not only is the fusion of B and C affected, but also the sequential grouping of A with B. As the BC fusion becomes weaker due to greater BC asynchrony, B becomes more available to form a

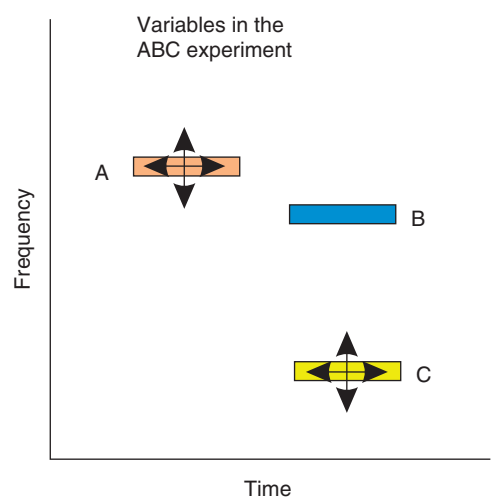

Figure 4 Diagram of a stimulus pattern in which a pure tone, A, alternates repeatedly with a pair of pure tones, B and C. The arrows show that the temporal positions and frequencies of both A and C may be varied. As A comes closer to B in frequency or time, it captures the latter more strongly into a sequential stream. As C comes closer to B in frequency or becomes more synchronous with it, it captures B more strongly into a fused perceptual unit.

sequential pure-tone stream with A. It is as if A and C were competing to make a connection with B.

The Old-Plus-New Heuristic

Another example of an interaction between simultaneous and successive organization is the old-plus-new heuristic. In natural environments, sounds do not strictly follow one another in time, nor are they exactly superimposed on one another. Instead they are typically overlapped. The old-plus-new heuristic uses the moment when a new sound enters a mixture to derive a very good description of the newly entering sound. It works as follows: when a spectrum suddenly becomes more intense or more complex, the auditory system carries out an analysis of this spectrum to determine whether the spectrum just before the change – or one very similar to it – is still present after the change. If so, it treats the changed spectrum as the sum of a continuing old spectrum plus a new sound (hence the name old-plus-new). Since it can compute the difference between the earlier and later spectra, it can derive, with some accuracy, the properties of the new sound. Note that it is not necessary that the earlier spectrum contains only a single sound, as long as no sound within it drops out at exactly the instant that the new sound begins. So the moment of onset of a new sound plays a critical role in identifying it and it would be surprising if the nervous system did not have mechanisms that were specialized for the analysis of sudden increases in spectral intensity or complexity.

Explanations

Explanations of stream segregation and other manifestations of ASA have fallen into two categories: functional and neurophysiological. The discussion up to this point has been about the job of the ASA system and which of its properties allows it to carry out its job. What cues does it use? How does it put them together? How does evidence accumulate? How do top-down and bottom-up processes interact? The concept of an ASA system is not actually a theory but a set of concepts and observations that can act as constraints on the form of such a theory.

An example of a physiological theory is one offered by Leo van Noorden and a closely related one proposed by William Hartmann and Douglas Johnson. The basic argument is that for stream segregation to occur, the members of the two streams must activate nonoverlapping populations of hair cells in the cochlea of the listener (Van Noorden), or, different frequency channels in the cochlea (Hartmann and Johnson). It is argued that the reason that streams can form on the basis of frequency differences is because each frequency has its strongest effect on a different population of hair cells. The reason that streams can form on the basis of differences in ear-of-arrival is that there are different populations of hair cells (or frequency channels) in the cochleas of the two ears.

This theory is attractive because it is simple and makes the peripheral sensing apparatus responsible for stream segregation. If true, it would open streaming to investigation by well-understood physiological techniques. Unfortunately, a number of findings fail to support this theory. The most clear-cut evidence comes from an experiment in which complex tones containing exactly the same spectral components were made to sound different in timbre and pitch by manipulating the phases of their components. When two tones that had different phase relations among their components were rapidly alternated in the galloping pattern described earlier (ABA–ABA– ···), each tone formed its own stream, despite the fact that it activated the same frequency channels within the cochlea as the other tone did.

In some trivial sense, the external sense organ is indeed responsible for stream segregation. After all, unless sounds differ in some way at the early receptive levels, the rest of the brain will never hear about it. However, this is far from saying that the neural computation that establishes the separate streams is close to the sensory periphery. One reason to believe that stream segregation, and in general, ASA, is computed higher up in the brain is that the instructions given to listeners, which activate top-down processes, can affect grouping in ambiguous cases. Another is that various auditory differences among the set of incoming sounds (e.g., pitch, spatial location, onset asynchronies) have a combined effect on the segregation of streams. Yet these cues may be provided by different sorts of feature analyzers. Bringing together the 'votes' of the different feature analyses has to be done at a level beyond the one at which the features are first detected, and at a level at which the effects of prior learning can play a role (probably in auditory cortex).

While adequate physiological explanations of ASA have not yet been forthcoming, physiological methods have been able to attack questions that are not easily answered using behavioral methods. One such question is whether the creation of auditory streams (the brain's representation of distinct environmental sounds) out of a sequence of sounds can precede the involvement of attention or requires the participation of attention. It is very hard to solve this problem using behavioral methods, because whenever human listeners have to carry out a task, such as reporting whether a particular set of sounds forms a separate

stream, this task focuses their attention on the sounds. Elyse Sussman and her colleagues have studied the involvement of attention in stream formation by recording event-related potentials (ERPs) from scalp-mounted electrodes when people are exposed to sequences of sounds while their attention is distracted by a visual task. She has used a component of the ERP called mismatch negativity (MMN) to detect whether the sounds are forming separate streams. The results suggest that at least three separate streams can be formed while the listener is not paying attention to the sounds, but as soon as attention is focused on one stream, the MMN evidence for the existence of the other streams disappears. So at least in some circumstances, stream segregation may be pre-attentive.

The ERP method is particularly well suited for studying auditory organization in persons who are not capable of reporting on their experiences, such as young infants. Do newborns organize sound the way adults do? If they do not, it suggests that the ASA methods may have to be learned. István Winkler and his colleagues played sequences of sounds to sleeping newborn infants 2–5 days of age. There was clear evidence for auditory stream segregation. This supports the idea that the mechanisms of ASA are primitive, in the sense of being inborn, and can give an initial boost to the infants' early learning about the important sounds in their environments, preventing them from memorizing the accidental combinations of properties exhibited by fortuitous combinations of sounds.

ASA in Other Animals

Nonhuman just like human animals live in a world where sounds come mixed with others. Yet they must respond only to certain sounds in the mixture. An important motivation for studying ASA in nonhuman animals is that its physiological basis could be investigated. Accordingly, ASA has been studied in a number of species, including birds, bats, frogs, fish, pinnapeds, and insects. However, a word of caution is in order. ASA is an accomplishment, not a mechanism. Even if different animals succeed in partitioning the sound mixture to the extent needed to detect their particular predators, prey, or potential mates, this does not mean that each of them does it by the same mechanism that the others do or that humans do. It is probable that many species have mechanisms to extract specific features and patterns that are important to the particular type of animal involved, and probably prepare certain actions that are appropriate responses to the detected patterns. Even humans may have them: some theorists have claimed that we possess specific mechanisms for extracting speech patterns from mixtures. Another example is that bats use specific bands of acoustic energy in the echolocation of their prey and have neural mechanisms specialized in the extraction of specific patterns from the echo data.

Macaque monkeys are large-brained animals and might be expected to share the general ASA mechanisms with humans. In one study with macaques, a rapid alternation of tones of two different frequencies, A and B, was delivered to them while recordings were made from their auditory cortexes. In humans, as the rate of alternation of A and B is increased, stream segregation increases. In the macaque auditory cortex, cells that respond best to A also respond a little to B. But as the rate of A–B alternation increases, these A-sensitive cells stop responding to B. It is as if the rate increase had caused stream segregation to occur, so that the cells that responded to A no longer 'saw' B. The same finding has also been reported in bats and starlings. Of course, even if this effect really is part of the stream-segregation mechanism, it does not automatically imply that the A-sensitive cell is, by itself, responsible for stream segregation. It may merely be a point at which one of the effects of stream segregation can be detected by an outside observer.

An interesting physiologically motivated neural network model, the ARTSTREAM model of Stephen Grossberg and his colleagues, has attempted to give a foundation for ASA in terms of neural computation. The process is much more complex than the activity of feature-sensitive cells.

It is possible to conceptualize ASA as the binding together of acoustic information of various types, in order to create acoustic objects, be they sequences or single sounds. Remember that more than one acoustic object (or stream) is being formed at the same time. So the brain has to register, for example, that it is the loud sound that has the rich timbre, and the soft sound that has the purer timbre, and not the reverse. It is possible that this binding of the right features to individual sounds could be carried out by temporarily recruiting some cells, activated by all the to-be-bound features, to represent the object as a whole (sort of a Hebbian phase sequence).

However, the existence of binding need not imply the convergence of all the information about the acoustic object to a common pool of neurons. This may not be the way 'objectness' is encoded in the brain. It has been proposed by DeLiang Wang, following the approach of Christoph von der Malsberg, that individual auditory features activate oscillatory processes in the brain and that their binding occurs when the oscillations representing particular features are driven into synchrony.

Computational Auditory Scene Analysis

Wang's theory, implemented in the form of a computational model, is one of many such models, inspired by the findings about human ASA. A new field of research, known as computational auditory scene analysis (CASA), attempts to create computational models of sound segregation. Most such models focus on the strongest cues for the grouping of sensory input: spatial location and harmonicity. However, a beginning has been made by Darryl Godsmark and Guy J Brown, on a more open system that allows a multiplicity of features to influence the final organization, using a blackboard architecture.

There are important practical benefits in creating effective computational systems for ASA: computer systems have great difficulty in recognizing speech mixed with other sounds, so a system that segregated sounds in the course of recognizing them would have a greater chance of successful recognition.

Conclusions

ASA affects all perceptible features of sound. The perceived loudness, position in space, pitch, rhythm, and timbre of sounds all depend on how the sensory input is organized. As we have become increasingly aware of this fact, the topic of ASA has stimulated research in psychophysics, cognitive science, biology, neuroscience, mathematical and computational modeling, speech, hearing science, audio engineering, and the psychology of music.

See also: Auditory Cortex Structure and Circuitry; Auditory System: Efferent Systems to the Auditory Periphery; Auditory System: Central Pathways; Auditory/Somatosensory Interactions; Auditory Cortex: Models; Cochlear Mechanics; Dichotic Listening Studies of Brain Asymmetry; Language: Auditory Processes; Otoacoustic Emissions; Sensory Aging: Hearing.

Further Reading

Alain C, Reinke K, He Y, Wang C, and Lobaugh N (2005) Hearing two things at once: Neurophysiological indices of speech segregation and identification. *Journal of Cognitive Neuroscience* 17(5): 811–818.

Bee MA and Klump GM (2004) Primitive auditory stream segregation: A neurophysiological study in the songbird forebrain. *Journal of Neurophysiology* 92: 1088–1104.

Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press (paperback 1994).

Bregman AS and Ahad P (1996) *Demonstrations of Auditory Scene Analysis: The Perceptual Organization of Sound*. Audio compact disk. Montreal: Authors (distributed by MIT Press).

Cooke MP and Brown GJ (1993) Computational auditory scene analysis: Exploiting principles of perceived continuity. *Speech Communication* 13(3–4): 391–399.

Darwin CJ and Carlyon RP (1995) Auditory grouping. In: Moore BCJ (ed.) *Hearing*, pp. 387–424. San Diego, CA: Academic Press.

Ellis DPW (1999) Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication* 27(3–4): 281–298.

Fishman YI, Reser DH, Arezzo JC, and Steinschneider M (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hearing Research* 151: 167–187.

Grossberg S, Govindarajan KK, Wyse LL, and Cohen MA (2004) ARTSTREAM: A neural network model of auditory scene analysis and source segregation. *Neural Networks* 17(4): 511–536.

Hartmann WM and Johnson D (1991) Stream segregation and peripheral channeling. *Music Perception* 9(2): 155–184.

Hulse S (2002) Auditory scene analysis in animal communication. In: Slater PJB, Rosenblatt JS, Snowdon CT, and Roper TJ (eds.) *Advances in the Study of Behavior,* vol. 31, pp. 163–200. San Diego, CA: Academic Press.

Klump G (2005) How does the hearing system perform auditory scene analysis? In: Van Hemmen JL and Sejnowski T Jr. (eds.) *23 Problems in Systems Neuroscience,* ch. 15. Oxford: Oxford University Press.

Moss CF and Surlykke A (2001) Auditory scene analysis by echolocation in bats. *Journal of the Acoustical Society of America* 110(4): 2207–2226.

Rosenthal DF and Okuno HG (eds.) (1998) *Computational Auditory Scene Analysis*. Mahwah, NJ: Erlbaum.

Sussman E (2005) Integration and segregation in auditory scene analysis. *Journal of the Acoustical Society of America* 117(3, Pt 1): 1285–1298.

Van Noorden LPAS (1975) Temporal coherence in the perception of tone sequences. Doctoral disseration, Eindhoven University of Technology, Eindhoven, The Netherlands (available online from http://aboxandia.tue.nl/extra1/PRF2A/7707058.pdf).

Wang D (1996) Primitive auditory segregation based on oscillatory correlation. *Cognitive Science* 20: 409–456.

Wang D and Brown GJ (2006) *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Piscataway, NJ: Wiley–IEEE Press.

Winkler I, Kushnerenko E, Horváth J, et al. (2003) Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences of the United States of America* 100(20): 11812–11815.

Yost WA (2004) Determining an auditory scene. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences,* 3rd edn., pp. 385–396. Cambridge, MA: MIT Press.

Auditory System: Central Pathway Plasticity

D R F Irvine, Monash University and Bionic Ear Institute, Melbourne, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

One of the most striking discoveries in sensory neuroscience in the past 40 years has been the fact that the functional response characteristics of single neurons in sensory central nervous system (CNS) structures, and the functional organization of those structures, can be modified by changes in the organism's sensory experience. This capacity for functional change is commonly referred to as 'plasticity.' Plasticity is demonstrated experimentally by changing either the organism's sensory environment or the significance for the organism of particular stimuli in that environment, or by modifying either the ancillary structures that deliver energy to the receptors, the receptor function itself, or the transmission of information from the receptors to the CNS. The first demonstrations of such plasticity were of changes that occurred only during restricted 'critical' periods during development, and it was initially assumed that the capacity for plastic change was restricted to these periods. More recently, however, it has been established that considerable plasticity is retained in adult sensory systems. The study of this plasticity has been complemented by major changes in the way in which the receptive fields (RFs) of sensory neurons are conceptualized. The evidence for both developmental and adult plasticity in central auditory system (CAS) structures, and on the nature of the mechanisms involved, is reviewed in the following sections.

It should be emphasized that the definition of 'plasticity' as changes in the structural and functional characteristics of neurons that occur in response to altered patterns of input is intended to distinguish it from changes that occur as passive consequences of the altered input or as a direct consequence of changes in the organism's state (e.g., emotional state, age). In the auditory system, for example, changes in the frequency tuning of auditory nerve fibers and central neurons occur as a direct (or passive) consequence of destruction of the outer hair cells (and thus of the cochlear amplifier), but these changes are not manifestations of plasticity. Plasticity involves some form of dynamic change in neural properties as a consequence of the altered input. Although the distinction between passive and plastic changes as a consequence of altered input is usually easily made, there are borderline cases where the distinction is difficult.

Developmental Plasticity

Effects of Neonatal Cochlear Ablation

Neonatal ablation of one cochlea in mammals, which eliminates afferent input from that ear to the CAS, results in structural and functional changes in the projections from the intact ear to binaural neurons in the CAS. Axons from the ventral division of the cochlear nucleus (CN) normally terminate in restricted regions of the major nuclei of the superior olivary complex (SOC). Following neonatal cochlear ablation, CN neurons on the side of the intact ear project to, and make synapses on, SOC regions normally innervated by the CN on the side of the ablated cochlea. Similarly, the terminal fields in the central nucleus of the ipsilateral inferior colliculus (IC) of axons from the CN on the side of the intact ear are much larger than those in normal animals. Both of these results indicate that, following unilateral cochlear ablation, axons from the CN on the intact side sprout to innervate additional territory. These structural changes are associated with increased excitatory responses to stimulation of the intact ear in the IC and primary auditory cortex (AI) ipsilateral to that ear. Although the effects of cochlear ablation are apparently restricted to a critical period preceding the onset of hearing, there is recent evidence indicating that CN lesions after the onset of hearing result in similar structural changes in the SOC.

Changes in Maps of Auditory Space

In barn owls, and in many mammals, the deep layers of the superior colliculus (SC) (or optic tectum) contain maps of auditory and visual space. These maps are in register: bimodal neurons, or auditory and visual neurons at the same locus, have corresponding auditory and visual spatial RFs (Figure 1(a)). The auditory space map is derived via orderly projections from the external nucleus of the IC (ICX). It is based on neural sensitivity to interaural time and level differences (ITDs and ILDs, respectively) and, in mammals, on sensitivity to spectral cues produced by the effects of the head and pinnae on the sound field. If a sound-attenuating plug is placed in one external meatus, the values of the interaural disparities associated with any given spatial location are altered. Such ear plugging in adult animals results in the auditory and visual space maps being misaligned, and the owls fail to recover normal localization, for as long as the plug is in position. When the ear plug is removed, the maps are aligned and the owls localize with normal accuracy. However, when barn owls or ferrets are

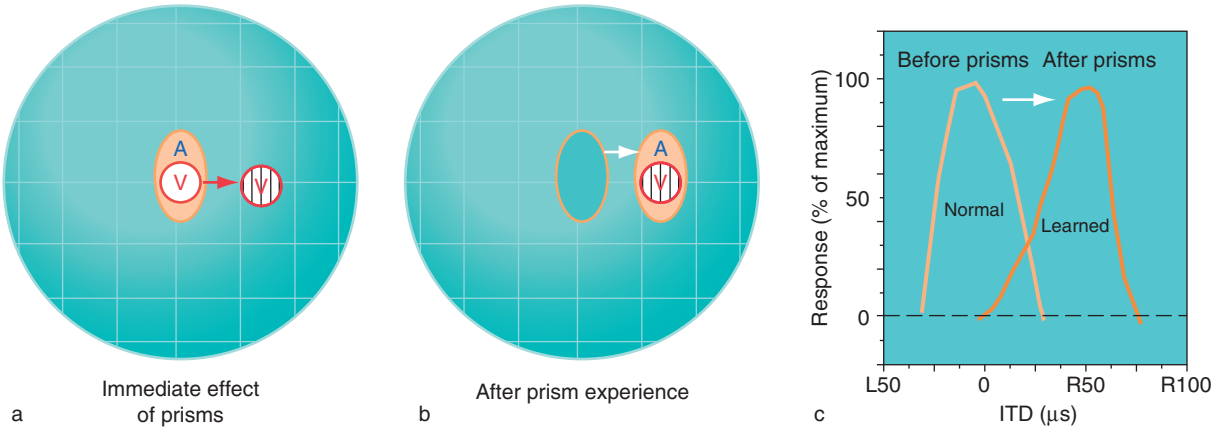

Figure 1 Plasticity of auditory tuning in the optic tectum of a juvenile barn owl, resulting from prism experience. Most neurons in the optic tectum respond to both auditory and visual stimuli and have auditory and visual receptive fields that are mutually aligned in space. (a) An example of the effect of 23° prisms on the location of a neuron's visual receptive field (encircled V; prismatically displaced field is striped). The globe represents space relative to the owl's line of sight. The auditory receptive field (A) is orange. (b) After the owl has experienced prisms for 8 weeks, the auditory receptive field has shifted to align with the prismatically displaced visual receptive field. (c) Plasticity of interaural time difference (ITD) tuning. ITD tuning of 2 units, both with visual receptive fields located at 0° azimuth, with the prisms removed (as shown in a), measured before (normal) and after (learned) 8 weeks of prism experience. On the ITD scale, L values represent left ear leading; R values represent right ear leading. Adapted from Knudsen EI (2002) Instructed learning in the auditory localization pathway of the barn owl. *Nature* 417: 322–328, with permission.

reared from infancy with one ear plugged, the visual and auditory maps in the adult are in register. This registration between the maps has been shown in barn owls to reflect the fact that the tuning to ITDs and ILDs of neurons in ICX and SC has changed to maintain alignment of the auditory and visual RFs. This change in neural tuning is associated with the appearance of novel projections from ICC to ICX: axons sprout into, and form synapses in, regions outside their normal projection zone. Immediately after ear plugging, young owls mislocalize sounds, but in parallel with the map changes they recover accurate localization ability. When the plug is removed, they exhibit systematic localization errors, and normal localization is recovered only if plug removal occurs within a critical period. Analogous developmental plasticity occurs when owls are raised wearing prismatic spectacles that displace the visual field in the horizontal plane: the tuning of SC neurons to ITDs (which serve as the cue for azimuthal location in barn owls) shifts to maintain the correspondence between auditory and prismatically displaced visual RFs (Figure 1). The plasticity demonstrated in these studies using experimental manipulations of sensory input is analogous to plastic changes that must occur in all animals during development, as the size of the head and pinnae increases and the nature of the ITDs, ILDs, and spectral transformations associated with particular spatial locations consequently changes.

Adult Plasticity

Plasticity Induced by Restricted Cochlear Damage

Lesions of a restricted region of the cochlea (which result in a partial hearing loss) in adult animals result in a reorganization of the frequency map in the AI. The nature of this reorganization is that the region deprived of its normal input by the cochlear lesion is occupied by an enlarged area in which neurons have characteristic frequency (CF) at the frequency represented at the edge(s) of the cochlear lesion (Figure 2). The thresholds, latencies, and sharpness of frequency tuning of neurons in this expanded representation at their new CF indicate that the changed frequency organization is not a passive consequence of the peripheral lesion but instead reflects a dynamic process of reorganization. Such reorganization has been described in a number of species, including nonhuman primates, and as a consequence of cochlear lesions produced by different procedures (mechanical, noise trauma, ototoxic drugs). After mechanical cochlear lesions in adults, similar reorganization is seen in the major auditory thalamic nucleus (the ventral division of the medial geniculate nucleus), but reorganization either does not occur or is weak in the ICC, suggesting that this form of CAS plasticity is a characteristic of thalamocorticothalamic circuitry. However, reorganization is seen in both the AI and the ICC of animals partially deafened neonatally (by ototoxic injections), suggesting that the

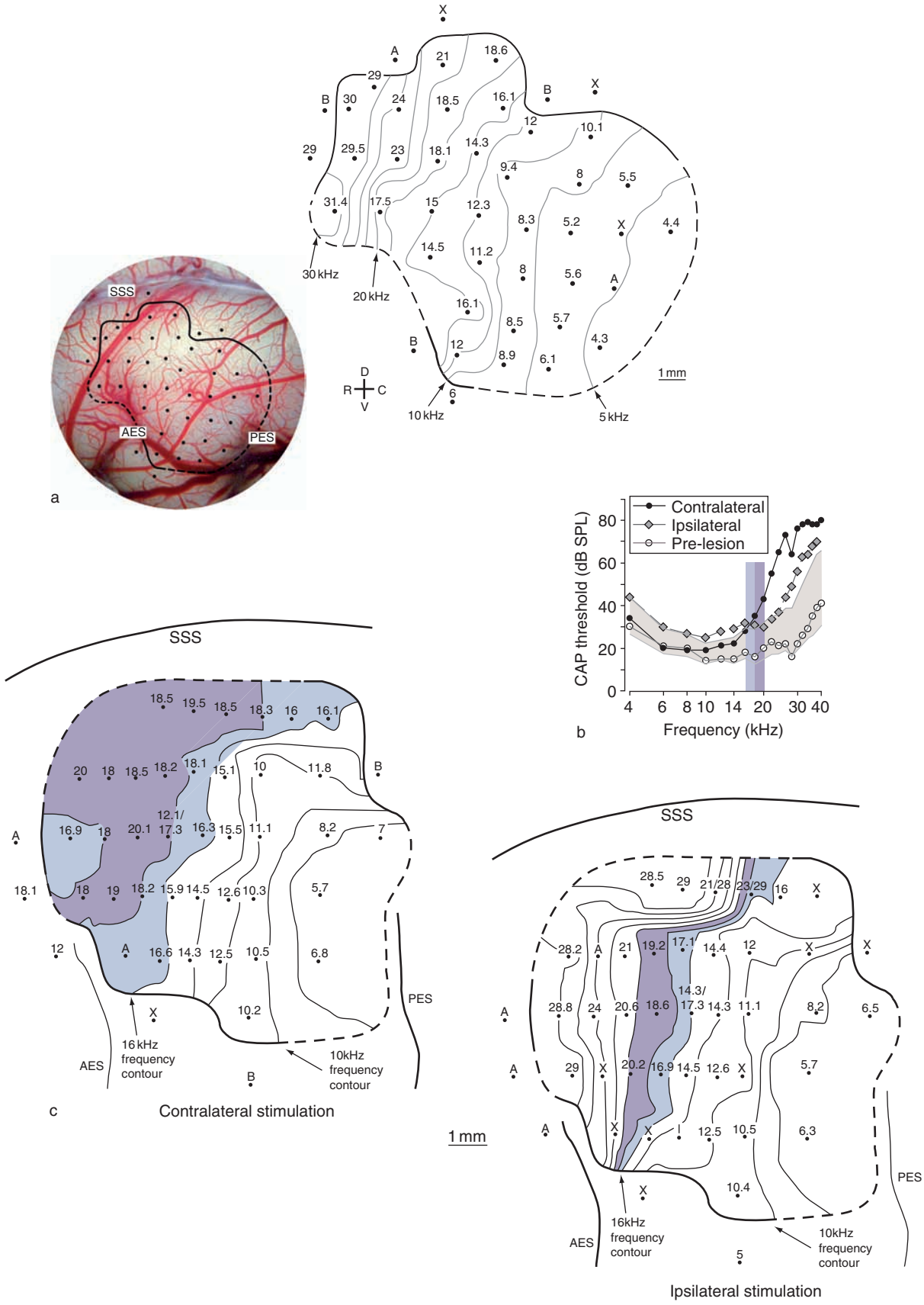

Figure 2 Continued

capacity for such reorganization is greater during development than in adults.

In all mammals, auditory cortex comprises a number of fields, commonly classified as 'core' (AI and one or more other tonotopically organized fields), and as surrounding 'belt' and 'parabelt' fields. Studies of injury-induced cortical reorganization have almost exclusively focused on the AI, with incidental observations on a second core field in some cases. The reason for this is mainly technical, in that the precise tonotopy in the core fields makes it easier to both identify and quantify reorganization. It is likely, however, that the capacity for plastic change is at least as great, and probably greater, in the less precisely organized belt and parabelt fields. Finally, reorganization analogous to that seen in the AI after restricted cochlear lesions is seen in visual cortex after restricted retinal lesions and in somatosensory cortex after digit amputation or after a variety of other procedures for eliminating input from a restricted portion of the skin. It is clear that this capacity for map reorganization in adults is a general characteristic of sensory systems.

Plastic Effects of Cochlear Electrical Stimulation in Profoundly Deaf Animals

The occurrence of plastic changes in the CAS of profoundly deaf animals (or humans) can obviously not be investigated using acoustic stimuli. However, electrical stimulation via intracochlear electrodes, which directly activates surviving spiral ganglion neurons, can be used to assess the responsiveness of the CAS in such cases. This stimulation is directly comparable to that provided by cochlear implants, which have been used to restore functional hearing to many thousands of profoundly deaf humans. Plasticity in the CAS of congenitally deaf or neonatally deafened animals has been demonstrated by comparing responses to cochlear electrical stimulation in adult animals that have received chronic electrical stimulation, for some specified period, with those in unstimulated controls. In some cases the stimulation is presented over weeks or months prior to the final recording, in which case differences undoubtedly reflect adult plasticity. In other cases, the stimulation is initiated at an early stage of development (as soon as it is possible to carry out implant surgery), but continues during adulthood, so it is unclear to what extent the plastic changes occur during development or adulthood. Although there appear to be no studies directly comparing the effects of cochlear electrical stimulation restricted to an early developmental period with that restricted to adulthood, it is likely that the capacity for plastic change is in fact greater in the former.

In congenitally deaf or neonatally deafened adult animals, the cochleotopic organization of the IC is normal, but that in the AI is rudimentary at best. Chronic cochlear electrical stimulation via a multichannel electrode (i.e., at a number of cochlear places) can restore (or maintain) basic cochleotopy. Stimulation of a restricted region of the cochlea results in an expansion of the representation of that cochlear region in the IC and in the AI. This effect is analogous to the expansion of the representation of lesion-edge frequencies after restricted cochlear lesions, but the fact that only limited effects are seen in the IC after mechanical cochlear lesions suggests that subcortical changes contribute in different ways to the two forms of plasticity.

The temporal pattern of AI neuronal responses to cochlear electrical stimulation is also modified by chronic stimulation. Responses in congenitally deaf adult animals are characterized by small

Figure 2 (a) Frequency map of the primary auditory cortex and digital photograph of the exposed cortical surface of a cat with normal hearing (AES, anterior ectosylvian sulcus; PES, posterior ectosylvian sulcus; SSS, suprasylvian sulcus). The dots on the photograph indicate the sites at which microelectrode penetrations were made, and the solid black line indicates the physiological boundary of the auditory cortex as defined from the data shown in the map. In the frequency map, the characteristic frequency of the neuron cluster recorded in each penetration is indicated above the dot other penetrations are labeled X' (no response to acoustic stimulation), A (acoustically responsive, but characteristic frequency could not be determined), B (broadly tuned), or I (inhibitory response). The line defining the physiological boundary of the auditory cortex is broken where this boundary was not determined unequivocally. Thin lines indicate iso-characteristic frequency contours (characteristic frequency identified by figures at lower boundary of the auditory cortex) fitted to the data at 2.5 kHz intervals using an inverse-distance smoothing function; R, C, D, and V indicate rostral, caudal, dorsal, and ventral directions, respectively. (b) Compound action potential (CAP) audiograms for a chronically lesioned cat. CAP thresholds (in decibels sound pressure level; dB SPL) prior to cochlear lesioning (prelesion) in the lesioned ear (open circles), and for both ears, lesioned (contralateral; filled circles) and normal (ipsilateral; diamonds), at the time of cortical mapping. The shaded area gives mean normal ± 1 SD CAP thresholds from a large pool of cats ($n > 50$). The lesion produced a steeply sloping hearing loss with edge frequency in the region of 18 kHz. The normal ear has a mild high-frequency loss at the time of cortical recording, but all thresholds are close to 1 SD above the mean. The vertical colored bars indicate the frequency bands at the edge of the lesion for which the representations (in c) are enlarged. (c) Frequency maps of the auditory cortex in the hemisphere contralateral to the (unilateral) cochlear lesion for stimulation of the contralateral (lesioned) ear and the ipsilateral (normal) ear in the cat for which the CAP audiogram is shown in b; conventions as in a. Blue and purple bands indicate the area of cortex containing neurons with characteristic frequencies in the range 16–18 and 18–20 kHz, respectively. Modified from Kamke MR, Brown M, and Irvine DRF (2004) Basal forebrain cholinergic input is not essential for lesion-induced plasticity in mature auditory cortex. *Neuron* 48: 675–686, with permission.

middle-latency responses and the complete absence of long-latency responses. Chronic cochlear electrical stimulation (initiated early in development) results in increased amplitude and a broader spatial distribution of middle-latency responses, and the appearance of long-latency responses. The temporal resolution of IC responses, as manifested in response latency and in the ability to follow the frequency of pulse trains, is also degraded by prolonged deafness, but chronic intracochlear electrical stimulation can reverse this degradation, and for some stimulation parameters, produce temporal resolution better than that seen in control animals.

Plasticity Associated with Various Forms of Learning

Changes in the stimulus selectivity of neurons in the higher levels of the auditory pathway as a consequence of behavioral conditioning procedures which alter the significance of particular stimuli for the organism were the first demonstrations of CAS plasticity in adult animals, and this remains the most active area of research on this topic. The most frequently reported finding in studies in which a tonal stimulus serves as the conditioned or discriminative (henceforth 'training') stimulus is that the response of cortical neurons at the training frequency increases, while the response at other frequencies decreases, such that the training frequency becomes the neurons' 'best' frequency (i.e., the frequency evoking the largest response) (Figure 3). Such effects have been described in the AI and in belt (secondary) auditory cortical fields, and in the auditory thalamus. These studies have commonly used fear conditioning procedures, but appropriate controls, and the specificity of the effects to the training frequency, establish that the observed effects are manifestations of plasticity rather than consequences of changes in state variables. Similar changes in the spectrotemporal receptive fields of auditory cortical neurons have been described in animals trained to detect a target tone of a particular frequency embedded in a sequence of broadband noiselike stimuli.

Another form of auditory learning that might involve CAS plasticity is perceptual learning, the improvement in discriminative capacity with training that is a common observation in all sensory modalities, both in psychophysical experiments and in everyday life. The specificity of many forms of visual perceptual learning to particular features of the training stimuli, or to the region of the receptor to which the stimuli are presented, led to the proposal that the learning might involve changes in neural tuning in primary sensory cortex. In the auditory system, the evidence from animal studies for changes in the AI associated with perceptual learning on frequency discrimination tasks is equivocal, and it is possible that the learning involves changes in higher-order decision-making processes. However, evidence for auditory cortical changes is provided by reports that auditory evoked responses to musical tones (measured by either electro- or magnetoencephalography) are of greater amplitude in trained musicians than in nonmusicians. Although this correlation could simply reflect the fact that people with this characteristic are more likely to undertake musical training, the fact that in some studies the effect is specific to the trained instrument (i.e., it is timbre specific), and that similar enhancement is seen when nonmusicians undergo training, suggests that the correlational evidence reflects plasticity.

The cortical responses exhibiting plasticity in these human studies are mainly localized to non-core (viz., belt and parabelt) areas of auditory cortex, in contrast to most of the animal evidence which has been derived from the AI. As noted previously, it is likely that the capacity for plastic change is generally greater in non-core than in core sensory cortical fields.

Two forms of auditory perceptual learning that are of great practical significance relate to speech perception. The first is the effect of language experience on the perception of speech sounds, and thus on language acquisition, during a critical period of development. The second is the improvement in speech discrimination shown by people with cochlear implants over the months and years following implantation. It is not clear, however, whether these forms of learning reflect plasticity in the CAS or in regions involved in higher-order cognitive processing of auditory information.

Plasticity Associated with Environmental Enrichment

One of the most intensively studied aspects of structural brain plasticity over more than 50 years has been the effects of so-called environmental 'enrichment': the brains of rats raised in enriched environments are structurally different in many respects from those of rats raised in (admittedly often impoverished) standard conditions. There has been only a limited investigation of the effects of environmental enrichment on the response characteristics of auditory cortical neurons, but the available evidence indicates that a combination of generalized and specifically acoustic enrichment results in changes in the response strength, RF characteristics, and temporal response properties of AI neurons.

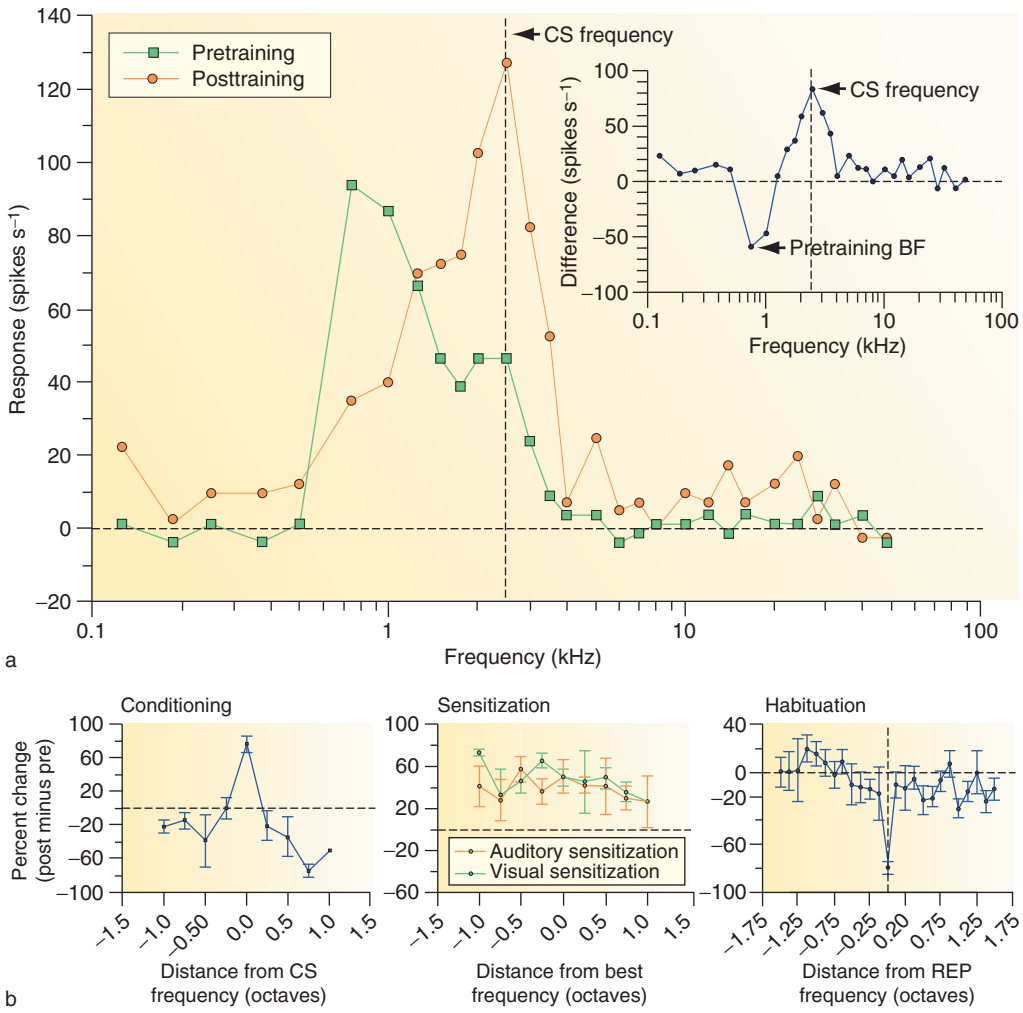

Figure 3 The effects of classical conditioning on the frequency selectivity of neurons in the primary auditory cortex of the guinea pig. (a) Receptive field plasticity of a single auditory cortex neuron is manifested in isolevel functions obtained (at 70 dB sound pressure level) before and after tone–shock conditioning (pretraining and posttraining, respectively). The best frequency (that evoking the maximum response) shifts from below 1 kHz to the frequency of the conditioned stimulus (CS). The inset shows the difference in frequency selectivity (posttraining minus pretraining), with the maximum increase in response at the frequency of the CS. (b) Normalized group difference functions (posttraining minus pretraining) showing change in response as a function of the distance of the probe frequency from the frequency used in behavioral training in conditioning, sensitization, and habituation conditions. Conditioning (left) produces a selective increase in response at the CS frequency, with reduced responses at most other frequencies. Sensitization training produces a nonselective increase in response at all frequencies, both for auditory sensitization (tone–shock unpaired) and visual sensitization (light–shock unpaired), showing that this nonassociative effect is transmodal. Habituation (repeated presentation of the same tone alone) produces a selective decrease in response at that frequency. Adapted from Weinberger NM (2004) Specific long-term memory traces in primary auditory cortex. *Nature Reviews Neuroscience* 5: 279–290, with permission.

Plasticity Induced by Microstimulation

Studies in a range of species indicate that focal electrical stimulation of a region of the AI can change the frequency selectivity of neurons around the stimulation site. In some experiments, the frequency selectivity of neurons in the tonotopically corresponding area of the central nucleus of the IC has also been shown to change as a consequence of cortical microstimulation. Centrifugal fibers from the AI to IC have been shown to play a role in the latter form of plasticity.

Plasticity Induced by Direct Activation of Neuromodulatory Systems

The neocortex receives diffuse extrathalamic projections from five different subcortical cell groups which use different neurotransmitters. These systems act to modulate the sensitivity of cortical neurons and have been implicated in cortical plasticity. The most extensively studied is the system of cholinergic fibers originating in the basal forebrain (BF). Pairing electrical stimulation of the BF with tonal stimulation at a

particular frequency has been reported to shift the frequency selectivity of cortical neurons toward the paired frequency and to result in an enlarged representation of that frequency in AI. These effects of experimental activation of the BF cholinergic system suggest that it is almost certainly involved in the use-related plasticity in the auditory cortex, as described in previous sections. This hypothesis is supported by a substantial body of evidence that the cholinergic basal forebrain plays an important modulatory role in various forms of auditory learning in both animals and humans. Lesions of the cholinergic BF do not, however, prevent injury-induced reorganization of AI tonotopy.

Cellular Mechanisms of CAS Plasticity

The evidence reviewed so far for the plasticity of the RFs of neurons in the CAS has been complemented by changing conceptions of RFs in all sensory modalities over the past two decades. In the early period of sensory neuroscience, RFs were described simply in terms of the range of the given stimulus parameters that evoked an increase in action potential frequency above the background resting activity. In the absence of spontaneous activity, these studies could not identify inhibitory components of RFs. Subsequent studies using forward-masking paradigms or the iontophoretic application of inhibitory transmitter antagonists revealed that many neurons at all levels of the CAS receive suprathreshold excitatory input, which is masked by inhibition, outside the 'classically' defined RF. More recently, intracellular recordings *in vivo* have revealed that many cortical neurons receive subthreshold excitatory and inhibitory inputs over an even broader range of stimulus values beyond their 'classical' RFs, much of which is derived from intrinsic corticocortical connections.

The cellular mechanisms underlying plasticity in adult sensory systems undoubtedly comprise a hierarchy of changes, differing in time course and complexity, and different forms of plasticity might involve different subsets of these mechanisms. The supra- and subthreshold inputs outside the aforementioned classic RF provide a broad matrix of inputs from which short-term RF changes can be sculpted. The first stage of injury-induced plasticity in adult sensory cortices appears to be an immediate expansion of RFs, the rapidity of which indicates that it reflects an unmasking of normally inhibited excitatory inputs. A similar unmasking is involved in space map plasticity in the barn owl. The subsequent establishment of smaller RFs in injury-induced plasticity presumably involves changes in the efficacy of both excitatory and inhibitory synapses (i.e., in synaptic weights). These changes are likely to involve an increase in the efficacy of previously subthreshold inputs, although changes of this sort have not yet been experimentally demonstrated. In both developmental plasticity in the barn owl and adult plasticity in the mammalian visual and somatosensory systems, there is evidence for the strengthening of excitatory synapses via *N*-methyl-D-aspartate (NMDA) receptor-mediated long-term potentiation (LTP). LTP itself, and this stage of plasticity, are also thought to involve the activation of previously 'silent' synapses. As discussed previously, there is evidence that neuromodulatory systems contribute to plastic changes in the cortex, and that centrifugal fibers are involved in some forms of subcortical plasticity.

Finally, on a longer timescale, structural changes in axon collaterals and in dendritic orientation have been shown to contribute to some forms of cortical plasticity. Long-term changes in the barn owl involve axonal sprouting and synaptogenesis, and axonal sprouting is also involved in the changes observed after unilateral cochlear ablation in neonatal animals. Axonal sprouting of horizontal fibers in the superficial layers of the cortex has been shown to be associated with visual cortical plasticity after retinal lesions in adults. In the barrel cortex of adult rats, vibrissal deafferentation, which results in an expansion of the influence of intact vibrissae into the area in which the deafferented vibrissae were represented, has been shown to result in changes in the dendritic orientation of neurons in this area toward the sources of this new input. Structural changes of this sort have not yet been shown in injury-induced auditory plasticity in adults, but the similarity of many forms of adult plasticity in different sensory systems suggests that such changes are likely to be involved.

Functional Significance of Central Auditory System Plasticity

Auditory developmental plasticity and learning-induced plasticity are undoubtedly adaptive, in that they enhance the organism's ability to adjust to altered patterns of input and, in the case of human language acquisition and adaptation to a cochlear implant, to make important auditory discriminations. It is less clear that adult injury-induced plasticity is adaptive, as the organism remains deaf in the frequency range affected by the cochlear lesion. Although there is evidence for enhanced frequency discrimination ability at lesion-edge frequencies in humans with cochlear damage (and associated hearing losses) of the sort that results in cortical reorganization in animal studies, it is unlikely that this slight change constitutes any form

of compensation for the deafness. It is likely that this form of plasticity is an extreme manifestation of the processes that underlie other forms of plasticity in response to altered input, but it is also possible that similar processes are involved in recovery of function after central damage such as that produced by stroke.

The possibility that some forms of learning impairment might involve deficiencies in aspects of auditory processing that exhibit plasticity, such that these deficiencies can be modified by training, has resulted in the recent development of a number of auditory training software packages. As discussed previously, the success of cochlear implants undoubtedly rests in part on the plasticity of CAS structures, and plasticity therefore contributes to the therapeutic effects of these devices.

It is also possible that some forms of CAS plasticity might have pathological consequences. In the somatosensory system, there is evidence that cortical reorganization might be responsible for some features of phantom limb experiences in amputees. It has similarly been suggested that tinnitus might be a consequence of cortical reorganization consequent on a cochlear lesion.

See also: Adult Cortical Plasticity; Auditory Cortex Structure and Circuitry; Auditory System: Central Pathways; Auditory Cortex: Models; Auditory Localization; Cortical Plasticity and Learning: Mechanisms and Models; Perceptual Learning: Neural Mechanisms; Perceptual Learning and Sensory Plasticity; Somatosensory Plasticity.

Further Reading

Atwood HL and Wojtowicz JM (1999) Silent synapses in neural plasticity: Current evidence. *Learning and Memory* 6: 542–571.

Buonomano DV and Merzenich MM (1998) Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* 21: 149–186.

Calford MB (2002) Dynamic representational plasticity in sensory cortex. *Neuroscience* 111: 709–738.

Engineer ND, Percaccio CR, Pandya PK, et al. (2004) Environmental enrichment improves response strength, threshold, selectivity, and latency of auditory cortex neurons. *Journal of Neurophysiology* 92: 73–82.

Gilbert CD (1998) Adult cortical dynamics. *Physiology Review* 78: 467–485.

Gu Q (2002) Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. *Neuroscience* 111: 815–835.

Hayes EA, Warrier CM, Nicol TG, et al. (2003) Neural plasticity following auditory training in children with learning problems. *Clinical Neurophysiology* 114: 673–684.

Irvine D, Brown M, Martin R, et al. (2005) Auditory perceptual learning and cortical plasticity. In: Koenig R, Heil P, Budinger E, et al. (eds.) *The Auditory Cortex: A Synthesis of Human and Animal Research*, pp. 409–428. Mahwah, NJ: Lawrence Erlbaum.

Irvine DRF and Wright BA (2005) Plasticity in spectral processing. In: Malmierca M and Irvine DRF (eds.) *Auditory Spectral Processing*, pp. 435–472. San Diego: Elsevier.

Kaas JH and Florence SL (2001) Reorganization of sensory and motor systems in adult mammals after injury. In: Kaas JH (ed.) *The Mutable Brain*, pp. 165–242. Amsterdam: Harwood.

Kamke MR, Brown M, and Irvine DRF (2004) Basal forebrain cholinergic input is not essential for lesion-induced plasticity in mature auditory cortex. *Neuron* 48: 675–686.

Knudsen EI (2002) Instructed learning in the auditory localization pathway of the barn owl. *Nature* 417: 322–328.

Kuhl PK (2000) A new view of language acquisition. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11850–11857.

Parks TN, Rubel EW, Popper AN, et al. (eds.) (2004) *Plasticity of the Auditory System*. New York: Springer.

Suga N and Ma X (2003) Multiparametric corticofugal modulation and plasticity in the auditory system. *Nature Reviews Neuroscience* 4: 783–794.

Syka J (2002) Plastic changes in the central auditory system after hearing loss, restoration of function, and during learning. *Physiological Reviews* 82: 601–636.

Weinberger NM (2004) Specific long-term memory traces in primary auditory cortex. *Nature Reviews Neuroscience* 5: 279–290.

Auditory System: Central Pathways

J C Middlebrooks, University of Michigan, Ann Arbor, MI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The central auditory system consists of interconnected processing stations at brain stem, midbrain, thalamic, and cortical levels (Figure 1). Input to the central auditory system originates in the cochlea of the inner ear. Active mechanical processes in the cochlea perform a frequency analysis of sound such that an inner hair cell at any point along the cochlear spiral is tuned to a particular frequency, with hair cells at the base of the cochlear spiral tuned to high frequencies and hair cells at the apex tuned to low frequencies. The systematic map of high-to-low frequency onto basal-to-apical cochlear place is known as 'tonotopic organization' ('tono' referring to frequency and 'topic' referring to cochlear place). Tonotopic organization is preserved in the projection to the brain and forms a functional organizing principle throughout all levels of the central auditory pathway.

The frequency-tuned input from each cochlea is carried to the brain by the left and right auditory nerves, each nerve containing ~30 000 fibers (in humans). Approximately 95% of auditory nerve fibers are 'type I' afferents that contact only one inner hair cell, thereby inheriting the frequency tuning of that hair cell. Action potentials in auditory nerve fibers tend to fire in phase with low-frequency sounds, up to about 4 kHz, and fire in a non-phase-locked manner to higher-frequency sounds. Based on this frequency-tuned, partially phase-locked auditory-nerve input, the central auditory system performs a multitude of tasks, including the following.

Spectral analysis. The auditory system preserves and enhances the frequency analysis that begins in the cochlea. It distinguishes sound spectra that characterize communication calls (including speech sounds) and environmental sounds. Whereas single auditory nerve fibers and some central neurons are tuned to single frequencies, many neurons in the central pathway are selective for complex spectra that contain multiple frequency components.

Intensity analysis. The auditory system discriminates sound pressure levels over nearly six orders of magnitude of pressure (i.e., 120 dB). Single auditory nerve fibers have dynamic ranges of only ~20–50 dB between the threshold sound level that barely elicits a reliable response and the saturated level at which further increases in level produce no further increase in response. Because of the limited dynamic ranges of single fibers, the full dynamic range of hearing must involve multiple nerve fibers that differ in threshold and in frequency sensitivity.

Temporal analysis. The auditory system analyzes temporal features of sounds on multiple timescales, from tens of microseconds for sound localization, to tens of milliseconds for discrimination of certain speech sounds (e.g., /da/ vs. /ta/), to hundreds of milliseconds for segmentation of speech syllables. It detects periodicity on the scale of tens to hundreds of cycles per second to distinguish pitch or to recognize voices.

Efferent control of the cochlea. The central auditory system enhances processing of signals in noise and protects the ear against intense sounds by modulating cochlear mechanics and the sensitivity of afferent fibers. This task involves pathways that originate in the auditory cortex and brain stem, ending with olivocochlear efferent fibers that terminate on outer hair cells and afferent fibers in the cochlea.

Sound localization. The auditory system computes the location of a sound source by integrating acoustical cues provided by the interaction of the incident sound wave with the head and external ears. The principal cues for localization are the relative timing and level of sounds at the two ears (for left/right localization) and direction-dependent spectral shapes imposed by the directional filter properties of the external ears (for up/down and front/back localization). Location is a fundamental attribute of a sound source, and sound localization is a topic that is amenable to both physiological and psychophysical research. For those reasons, the mechanisms for sound localization are among the most thoroughly studied of central auditory mechanisms.

Central auditory processing is accomplished by parallel processes involving anatomically discrete, physiologically specialized pathways. In some cases, one can point to a particular pathway that performs a specific task. In other cases, functions are carried out in multiple parallel pathways, and the result of a specific auditory function is distributed widely across multiple areas of the auditory and multimodal cortex.

Auditory Brain Stem

Cochlear Nucleus

The auditory nerve terminates in the cochlear nucleus complex at the junction of the medulla and pons (Figure 2). The cochlear nucleus comprises two

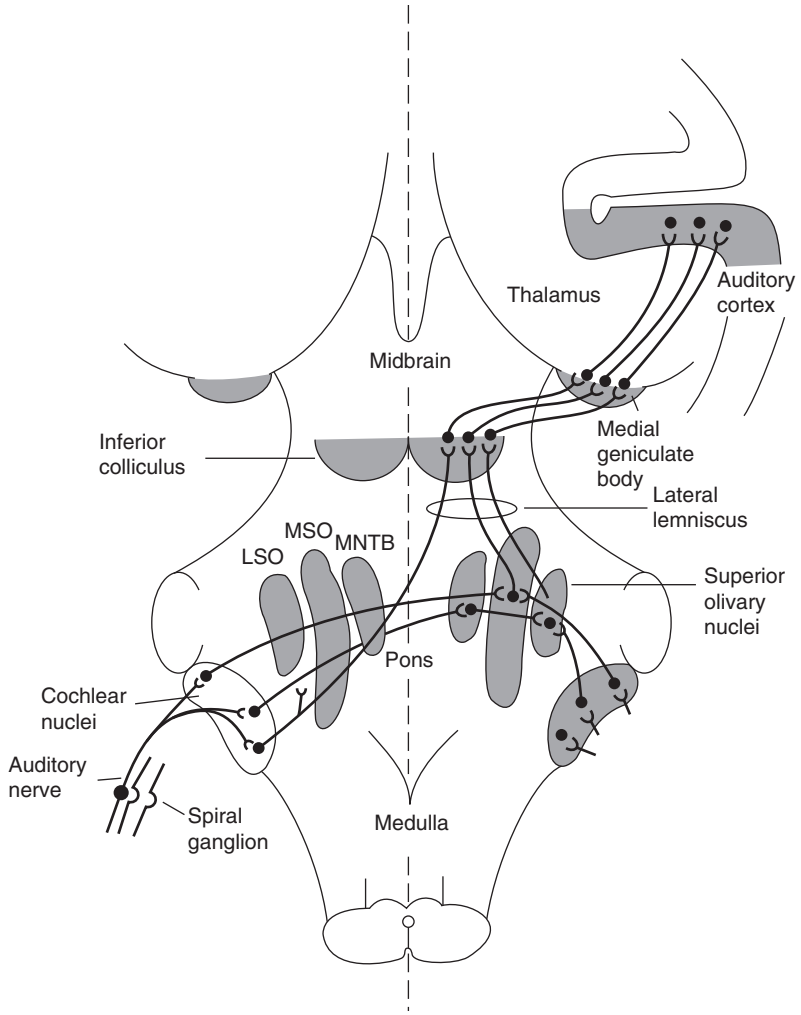

Figure 1 Central auditory pathway. All structures are bilaterally symmetrical, although the illustrated connections show only the pathways to the right auditory cortex. LSO, lateral superior olive; MSO, medial superior olive; MNTB, medial nucleus of the trapezoid body. Reproduced from Imig TJ, Irons WA, and Samson FR (1990) single-unit selectivity to azimuthal direction and sound pressure level of noise bursts in cat high-frequency primary auditory cortex. *Journal of Neurophysiology* 63: 1448–1466, with permission from Elsevier.

physically distinct divisions: the dorsal cochlear nucleus (DCN) and the ventral cochlear nucleus. The ventral cochlear nucleus is further divided, on the basis of the pattern of termination of auditory nerve fibers, into the posteroventral cochlear nucleus (PVCN) and anteroventral cochlear nucleus (AVCN). Each auditory nerve fiber divides into two major branches as it enters the cochlear nucleus. The ascending branch terminates in the AVCN. The descending branch gives off terminals in the PVCN, then continues to the DCN where it terminates. In that way, each auditory nerve fiber forms three terminal regions. The fibers terminate in a tonotopic pattern within each region, resulting in three maps of sound frequency within the cochlear nucleus on each side of the brain.

The cochlear nucleus complex traditionally has been regarded as a strictly monaural structure, responding only to stimulation of the ipsilateral ear. Recent studies, however, demonstrate a projection from the contralateral ear by way of the contralateral cochlear nucleus. The contralateral input is predominantly inhibitory, although a period of conductive hearing loss can result in an increase in excitatory inputs. These observations indicate that the cochlear nucleus plays at least a minor part in binaural hearing.

The cochlear nucleus contains many distinct classes of neurons that are distinguished by the morphology of cell body and dendrites, by patterns of auditory nerve synapses, and by intrinsic physiological properties. In the AVCN and PVCN, some of the prominent projection neurons (i.e., cell types that send projections out of the cochlear nucleus) are the spherical bushy cells, the globular bushy cells, the octopus cells, and the multipolar (or stellate) cells. Spherical bushy cells, found in the AVCN, have large spherical cell

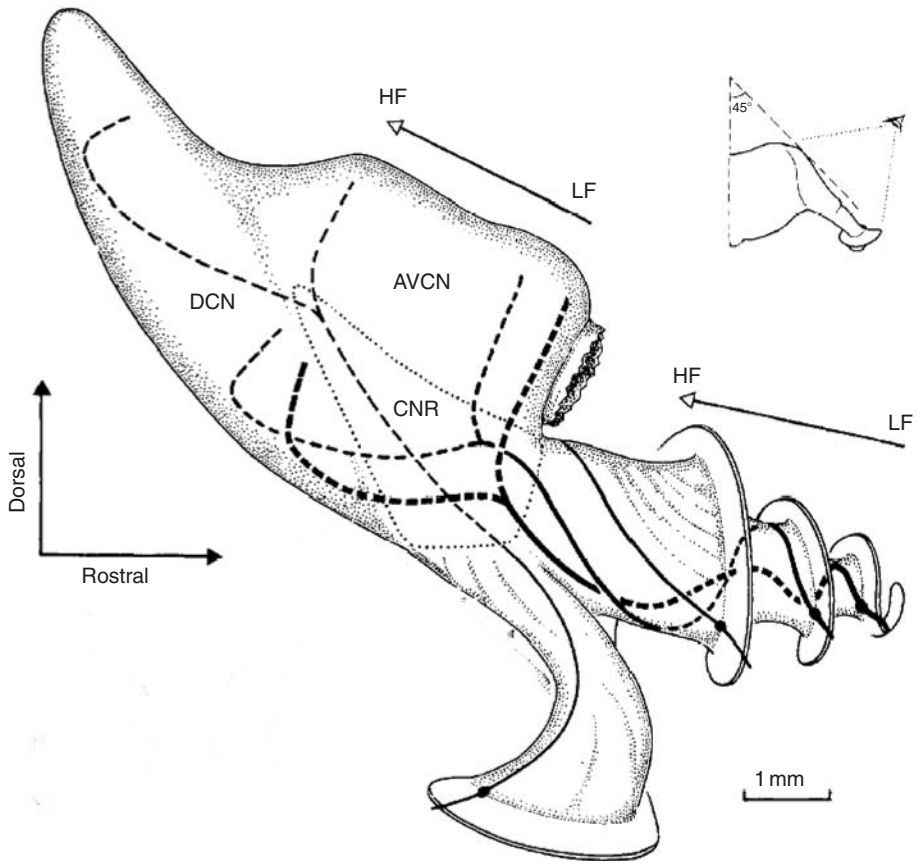

Figure 2 Cochlear nucleus complex and auditory nerve. This diagram from the right side of the cat shows auditory nerve fibers originating at four levels of the spiral ganglion. Each fiber forms major ascending (rostral) and descending (caudal) branches in the cochlear nucleus complex, resulting in maps of characteristic frequency from low frequency (LF) to high frequency (HF). The posterior ventral cochlear nucleus (PVCN), not shown here, lies deep to the cochlear nerve root (CNR). DCN, dorsal cochlear nucleus; AVCN, anterior ventral cochlear nucleus. Arnesen AR and Osen KK (1978) The cochlear nerve in the cat: Topography, cochleotopy, and fiber spectrum. *Journal of Comparative Neurology* 178: 661–678.

bodies that receive input from auditory nerve fibers by way of one or two large axosomatic endings, the endbulbs of Held. These synapses transmit temporal information with great fidelity, so the frequency selectivity and temporal firing patterns of spherical bushy cells closely resemble the responses of auditory nerve fibers – these are 'primary-like' firing patterns. Globular bushy cells receive axosomatic endings from a few auditory nerve fibers and respond with a slightly modified primary-like temporal pattern ('primary with notch'). Octopus cells, found in a restricted portion of the PVCN, receive input onto their dendrites from a large number of auditory nerve fibers spanning a wide frequency range, resulting in wide frequency response areas. The responses of octopus cells are tightly locked to the onsets of sounds. Multipolar cells are widely distributed throughout the AVCN and PVCN. They have elaborately branched dendritic trees that receive numerous synaptic boutons from auditory nerve fibers. They respond to sound with a periodic 'chopping' spike pattern; the chopping rate is determined by the electrical properties of the cell membrane rather than by the period of the sound.

The major projection cells of the DCN are the fusiform cells (also known as pyramidal cells) and some of the giant cells. They receive input directly from auditory nerve fibers as well as projections from granule cells and multipolar cells in the ipsilateral PVCN and AVCN. Fusiform cells and giant cells have complex frequency response areas consisting of multiple excitatory and inhibitory frequency regions. One likely role of neurons with complex frequency response areas is in detecting spectral shape cues for the up/down or front/back location of a sound source. That is, the external ear filters sounds differentially depending on the angle by which the sound enters the ear. Recognition of particular direction-dependent sound spectra, perhaps performed by the DCN, provides cues to vertical sound location. Indeed, experimental lesions of the DCN result in deficits in sound localization behavior in the vertical dimension.

The ascending projections from the cochlear nucleus form three major fiber bundles. The 'trapezoid body' is formed by fibers from the AVCN, destined primarily for the superior olivary complex or the inferior colliculus (IC). The 'intermediate acoustic stria' is formed primarily by fibers from the PVCN that will project to the contralateral IC and nuclei of the lateral lemniscus. The 'dorsal acoustic stria' contains the projection fibers of the DCN, which terminate primarily in the contralateral IC and nuclei of the lateral lemniscus.

Superior Olivary Complex

The superior olivary complex in the pons is the first major site of binaural convergence. The complex comprises three prominent nuclei: the medial nucleus of the trapezoid body (MNTB), the medial superior olive (MSO), and the lateral superior olive (LSO), as well as several periolivary nuclei. Neurons within the periolivary nuclei primarily make connections within the superior olivary complex or send projections to the cochlea and cochlear nucleus.

The MSO and LSO each have distinct roles in binaural hearing, which is important for sound localization in the horizontal dimension. Horizontal sound localization has long been understood in the context of a 'duplex theory' in which low-frequency sounds are localized on the basis of differences in the time of arrival of sounds at the two ears ('interaural time differences'; ITDs) and high-frequency sounds are localized on the basis of differences in sound level at the two ears ('interaural level differences'; ILDs). Based on the distinctive physiology of constituent neurons, one can associate ITD and ILD sensitivity with the MSO and the LSO, respectively. ITDs result from differences in the length of the path from a lateral sound source to the near and far ear. Humans utilize ITDs cues in the range of $\pm 700\,\mu s$ with resolution of $\sim 10\,\mu s$; ITD cues are most salient at frequencies below ~ 1 kHz. ILDs result from the tendency of the head to shadow the ear on the side away from the sound source. Humans process ILD cues for localization in the range of ± 30 dB with resolution of ~ 1 dB; ILDs are most informative at frequencies > 3 kHz.

The MSO has a tonotopic organization that primarily represents low frequencies. Individual neurons receive direct inputs from spherical bushy cells on both sides of the brain. The MSO cells are sensitive to the relative timing of inputs originating at the two ears, responding strongly to action potentials that arrive simultaneously at the MSO. A recent study in gerbils indicates that an inhibitory input from the MNTB acts to shift the ITD tuning of MSO neurons so that maximum spike rates are produced by binaural sounds leading at the contralateral ear and that maximum 'changes' in spike rates are associated with near-zero ITDs. That arrangement would result in maximum spatial acuity for sounds near the midline of auditory space, which is what is observed psychophysically.

The tonotopic organization of the LSO primarily represents high frequencies. LSO neurons receive excitatory inputs directly from spherical bushy cells of the ipsilateral AVCN. The contralateral pathway to the LSO is indirect, involving an excitatory projection from globular bushy cells in the contralateral AVCN to the ipsilateral MNTB, then an inhibitory projection from the ipsilateral MNTB to the LSO. The net effect is that LSO neurons are inhibited by sounds at the contralateral ear and excited by sounds at the ipsilateral ear. The binaural sensitivity of LSO neurons results in sensitivity to differences in sound levels at the two ears. Analogous to the situation in the MSO, the greatest change in LSO spike rates occurs for ILDs passing through zero, corresponding to sound-source locations near the midline of auditory space.

The 'olivocochlear' system consists of diffusely organized neurons in the superior olivary complex and their efferent fibers that terminate in the cochlea. There are lateral and medial components. Generally, the neurons of the lateral olivocochlear system are located in or around the LSO. The majority of these neurons have fine unmyelinated fibers that travel in the uncrossed olivocochlear bundle toward the ipsilateral cochlea. They eventually enter the cochlea by way of the vestibular nerve. Fibers of the lateral system terminate on the dendrites of primary afferent fibers near their synapses with inner hair cells. The function of the lateral efferent system is not well understood, but is thought to serve a protective role. The neurons of the medial olivocochlear system are located in the medial and rostral periolivary nuclei. Their projection fibers are myelinated, and the majority travel in the crossed olivocochlear bundle toward the contralateral cochlea. The fibers terminate with cholinergic synapses on outer hair cells. Activity in the medial olivocochlear system modulates the active mechanical feedback that outer hair cells contribute to the process of frequency analysis in the cochlea.

Nuclei of the Lateral Leminiscus

The ascending projections of most brain stem auditory nuclei coalesce on each side of the brain in a major fiber bundle known as the lateral lemniscus. Amid those fibers are nuclear groups known as the 'nuclei of the lateral lemniscus.' The nuclear groups vary widely among species, but generally one can distinguish ventral (VNLL), intermediate (INLL), and dorsal

(DNLL) nuclei of the lateral lemniscus. The VNLL receives its major input from the contralateral cochlear nucleus. Neurons in the VNLL typically are excited by stimulation of the contralateral ear and show no response to ipsilateral stimulation. Tonotopic organization in the VNLL is rather indistinct. The INLL also receives predominately contralateral inputs and is excited by stimulation of the contralateral ear. The INLL is particularly well developed in echolocating bats, in which it is thought to contribute to processing of temporal characteristics of sounds. The DNLL is a strongly binaural nucleus. It receives bilateral input from the AVCN and LSO, ipsilateral input from the MSO, and contralateral input from the opposite DNLL. High-frequency neurons in the DNLL display sensitivity to ILDs, similar to the responses of LSO neurons. Many DNLL neurons that respond to lower frequencies also show ITD sensitivity similar to that of neurons in the MSO.

The fibers of the lateral lemniscus, originating in bilateral cochlear nuclear complexes, superior olivary complexes, and nuclei of the lateral lemniscus, terminate in a key structure of the midbrain, the IC.

Midbrain

The major midbrain structures involved in hearing are the IC and superior colliculus (SC). The major divisions of the IC are the central nucleus (ICC), the lateral nucleus (roughly equivalent to the 'external nucleus' in other nomenclatures), the dorsal cortex, and the dorsomedial nucleus (Figure 3). Nearly all ascending auditory pathways synapse in the ICC en route from the brain stem to the forebrain. The ICC shows a prominent laminar organization formed by sheets of disk-shaped neurons and by the distribution of dendrites and axons. These sheets form functional 'isofrequency laminae' in that neurons within a particular sheet have similar characteristic frequencies. Perpendicular to the plane of the isofrequency lamina, there is a steady tonotopic progression in characteristic frequency, with characteristic frequencies increasing from low to high frequencies along a roughly dorsolateral to ventromedial axis.

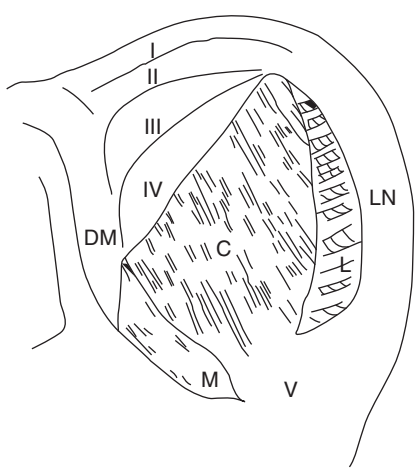

Figure 3 Subdivisions of the inferior colliculus, shown in transverse section. DM, dorsomedial nucleus; C, M, and L, central, medial, and lateral parts of the central nucleus; LN, lateral nucleus; I, II, III, IV layers of the dorsal cortex. Oliver DL and Morest DR (1984) The central nucleus of the inferior colliculus in the cat. *Journal of Comparative Neurology* 222: 237–264.

Each isofrequency lamina in the ICC receives projections from multiple brain stem nuclei. Some areas of isofrequency laminae receive projections that converge from multiple sources, whereas other areas are dominated by inputs from a particular brain stem source. The partial segregation of inputs within particular regions of isofrequency laminae permits parallel processing of specific attributes of sounds. The segregation of inputs is manifest as a diversity of response properties. The frequency response areas of neurons, for instance, show three major classes. Type V neurons generally have low characteristic frequencies and have V-shaped tuning curves similar to those of primary afferents. Type I neurons have narrow excitatory frequency bands that are flanked by inhibition. The type O neurons show sharply tuned excitatory responses to low-level tones and inhibition to high-level tones – this suggests a direct input from the DCN. Binaural response properties also vary, showing strictly contralateral excitatory responses (suggestive of a direct input from the contralateral cochlear nucleus), contralateral excitation and ipsilateral inhibition (suggestive of input from the contralateral LSO), sensitivity to interaural phase differences in low-frequency tones (suggestive of input from the MSO), or other more complex characteristics. The major output from the ICC to the forebrain travels through a fiber bundle known as the brachium of the IC.

Subdivisions of the IC outside of the central nucleus integrate input from the central nucleus of the IC, from descending cortical projections, and/or from the somatosensory system. Convergence of information across frequencies in those subdivisions partially obscures tonotopic organization. Neurons from the lateral nucleus of the IC and/or the nucleus of the brachium of the IC provide auditory input to a sensorimotor integrative structure of the midbrain, the SC.

The SC has the task of integrating auditory localization information with information from other sensory modalities to initiate orienting movements of the eyes and head. It receives visual, somatosensory, and auditory input. The ascending auditory input from the IC produces relatively short-latency responses of

SC neurons to the onsets of sounds. There is also descending input from auditory cortical areas that is more likely to have a modulatory role. The SC is organized functionally to form a dynamic map of motor error. That is, the receptive fields of neurons reflect the disparity between the angle of gaze (determined by head and eye position) and an auditory, visual, or somatic target. Studied under anesthetized conditions, this dynamic motor map can be seen as a complex of mutually aligned auditory, visual, and somatic sensory maps. In response to acoustic noise bursts presented from a moveable loudspeaker, single neurons show spatial receptive fields that are restricted in the horizontal and vertical dimensions. The centers of auditory spatial receptive fields vary according to the locations of neurons in the SC such that the horizontal dimension of auditory space maps onto the rostrocaudal dimension of the SC and the vertical dimension maps onto the mediolateral dimension. It is important to appreciate that the auditory spatial topography in the SC is not a simple representation of the auditory sensory epithelium – the cochlea maps frequency, not space. Instead, the auditory spatial map in the SC must reflect the integrated activity of multiple brain stem nuclei that are sensitive to specific auditory spatial cues, such as interaural phase difference, ILD, and direction-dependent spectral shape cues. Auditory spatial topography is particularly evident in the barn owl in a homologue of the SC, the optic tectum.

Auditory Thalamus and Cortex

The auditory forebrain consists of the medial geniculate body (MGB) of the thalamus and multiple auditory cortical areas in the temporal lobe. The MGB and cortical areas are linked by reciprocal thalamocortical and corticothalamic projections. The MGB has three major divisions: ventral, dorsal, and medial. The ventral division receives direct projections from the central nucleus of the IC. As in the central nucleus of the IC, neurons of the ventral division are sharply tuned for frequency and there is a prominent laminar tonotonic organization. Indeed, the ascending pathway from the ICC to the ventral division of the MGB to the primary auditory cortex constitutes the highest-fidelity representation of sound frequency in the form of tonotopically organized neurons with narrow frequency tuning curves. The inputs to the dorsal division of the MGB are primarily from IC subdivisions outside of the central nucleus. Dorsal division neurons tend to be broadly tuned for frequency. Dorsal division neurons project to the 'belt' areas of the auditory cortex (defined below), outside of the tonotopically organized 'core' regions. The medial division of the MGB receives convergent inputs from many auditory sources as well as some somatosensory input. The medial division projects rather diffusely to multiple auditory cortical areas. A rather unique population of very-large-diameter myelinated axons originates in the medial division and terminates in the superficial lamina of the auditory cortex.

The auditory cortex has a hierarchical organization consisting of core, belt, and parabelt levels (**Figure 4**). The core area receives the principal tonotopic projection from the ventral division of the MGB. The core area contains the primary auditory cortex (area A1) and, in some species, one or more additional tonotopically organized fields. The core area in the macaque monkey, for instance, contains two or three maps of frequency representations: in area A1, in a rostral area, and possibly in a rostrotemporal area. The core areas exhibit a pronounced columnar organization in which radial cell columns extending through the cortical layers share common characteristic frequencies and binaural response properties.

The core area of the auditory cortex is surrounded by multiple belt areas. The belt areas receive feedforward projections from the core as well as direct projections from the MGB, primarily from the dorsal division. Neurons in belt areas tend to respond more strongly to noise bands than to pure tones and often show complex frequency response areas with multiple excitatory and inhibitory regions. At least in primates, the parabelt areas form a third level of the hierarchy, with parabelt areas receiving inputs from belt areas but not from core areas. The parabelt areas project widely to temporal, parietal, and frontal cortical areas for association of auditory processes with other sensory and motor processes.

Auditory responses in the cortex reflect a chain of processing that is a minimum of five synaptic levels beyond the hair cells of the cochlea. One might expect that these many levels of processing would result in cortical neurons that are highly selective for specific sounds or specific features of sounds. Contrary to that expectation, most cortical neurons tend to show graded responses to multiple stimulus parameters. In regard to sound localization, for instance, there is no evidence in the cortex of a point-to-point map like that found in the SC. Instead, the responses of many cortical neurons vary systematically according to sound-source location through as much as 360° of location so that individual neurons represent auditory space panoramically. As a result, the cortical representation of any point in auditory space is distributed widely through the cortex. Our current understanding points to the representation of any sound by the coordinated activity of widespread populations of cortical neurons. The way in which

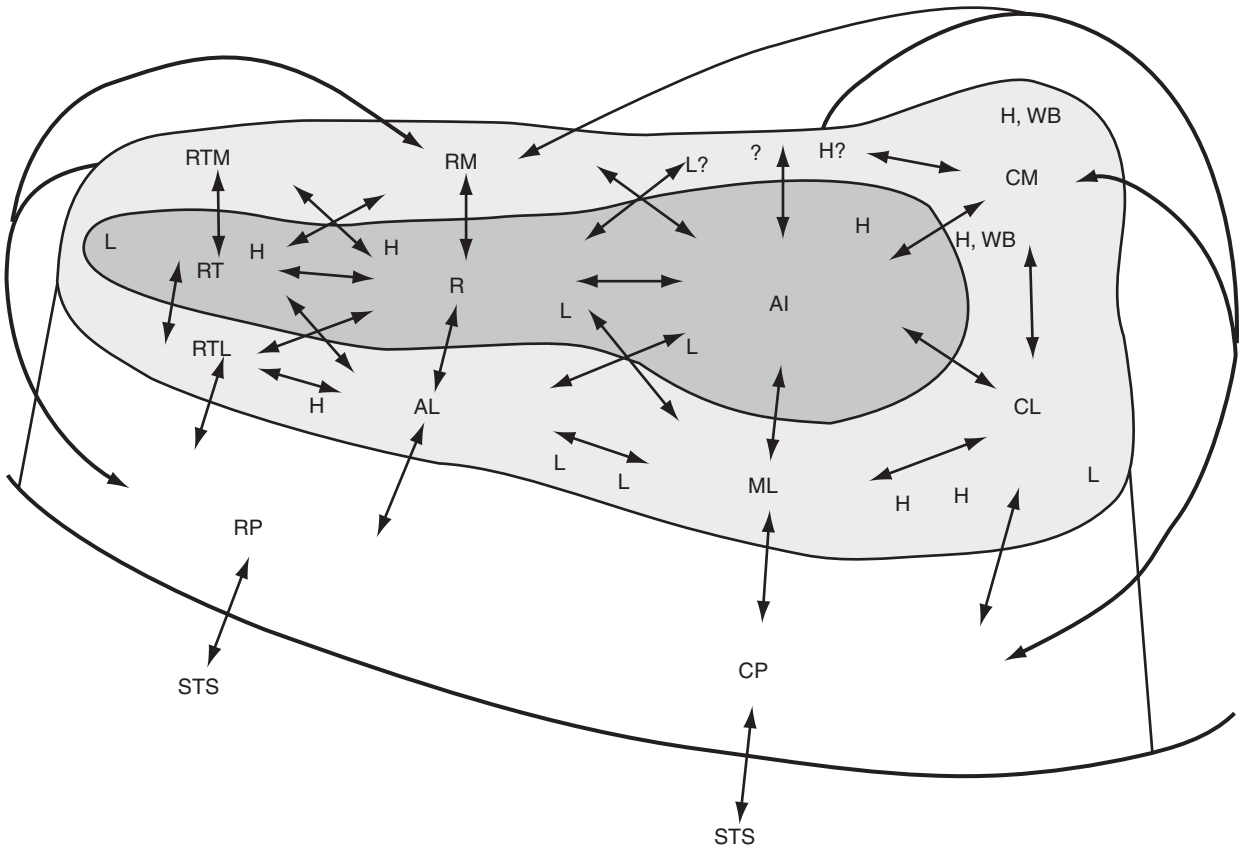

Figure 4 Auditory cortical areas in the macaque monkey. The auditory area lies on the dorsal surface of the temporal lobe; this is a view from the left hemisphere. Dark, medium, and light shading indicate core, belt, and parabelt areas, respectively. Broad and thin arrows indicate heavy and light connections, respectively. H, L, and WB indicate areas responsive to high, low, and wideband stimuli, respectively. Core areas: AI, primary auditory cortex; R, rostral; RT, rostrotemporal. Belt areas: RTL, lateral rostrotemporal belt; RTM, medial rostrotemporal belt; AL, anterior lateral belt; RM, rostromedial region; ML, middle lateral belt; CL, caudal lateral belt; CM, caudal medial belt. Parabelt areas: RP, rostral parabelt; CP, caudal parabelt. STS, superior temporal sulcus. Reproduced from Kass JH, Hackett TA, and Tramo MJ (1999) Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology* 9: 164–170, with permission from Elsevier.

patterns of activity in neural populations are translated to auditory perception and behavior remains a key topic in ongoing research.

See also: Auditory Cortex Structure and Circuitry; Auditory System: Giant Synaptic Terminals, Endbulbs and Calyces; Auditory/Somatosensory Interactions; Cerebral Cortex; Cochlear Development; Echolocation I: Behavior; Reticular Formation and the Brain Stem; Sound Localization: Neural Mechanisms; Temporal Processing in the Auditory Pathway.

Further Reading

Arnesen AR and Osen KK (1978) The cochlear nerve in the cat: Topography, cochleotopy, and fiber spectrum. *Journal of Comparative Neurology* 178: 661–678.

Brand A, Behrend O, Marquardt T, McAlpine D, and Grothe B (2002) Precise inhibition is essential for microsecond interaural time difference coding. *Nature* 417: 543–547.

Davis K (2005) Contralateral effects and binaural interactions in dorsal cochlear nucleus. *Journal of the Association for Research in Otolaryngology* 6: 280–296.

Ehret G and Romand R (eds.) (1997) *The Central Auditory System*. New York: Oxford University Press.

Hackett TA, Stepniewska I, and Kaas JH (1998) Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *Journal of Comparative Neurology* 394: 475–495.

Huang CL and Winer JA (2000) Auditory thalamocortical projections in the cat: Laminar and areal patterns of input. *Journal of Comparative Neurology* 427: 302–331.

Kaas JH, Hackett TA, and Tramo MJ (1999) Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology* 9: 164–170.

Loftus WC and Sutter ML (2001) Spectrotemporal organization of excitatory and inhibitory receptive fields of cat posterior auditory field neurons. *Journal of Neurophysiology* 86: 475–491.

May BJ (2000) Role of the dorsal cochlear nucleus in the sound localization behavior of cats. *Hearing Research* 148: 74–87.

Middlebrooks JC and Green DM (1991) Sound localization by human listeners. *Annual Review of Psychology* 42: 135–159.

Middlebrooks JC and Knudsen EIA (1984) Neural code for auditory space in the cat's superior colliculus. *Journal of Neuroscience* 4: 2621–2634.

Middlebrooks JC, Xu L, Eddins AC, and Green DM (1998) Codes for sound-source location in non-tonotopic auditory cortex. *Journal of Neurophysiology* 80: 863–881.

Oertel D, Bal R, Gardner SM, Smith PH, and Joris PX (2000) Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11773–11779.

Oertel D, Fay RR and Popper AN (eds.) (2002) *Integrative Functions in the Mammalian Auditory Pathway.* New York: Springer.

Oliver DL and Morest DK (1984) The central nucleus of the inferior colliculus in the cat. *Journal of Comparative Neurology* 222: 237–264.

Pickles JO (1988) *An Introduction to the Physiology of Hearing*, 2nd edn. London: Academic Press.

Ramachandran R and May BJ (2002) Functional segregation of ITD sensitivity in the inferior colliculus of decerebrate cats. *Journal of Neurophysiology* 88: 2251–2261.

Sparks DL (1989) The neural encoding of the location of targets for saccadic eye movements. *Journal of Experimental Biology* 146: 195–207.

Strutt JD (Lord Rayleigh) (1907) On our perception of sound direction. *Philosophical Magazine* 13: 214–232.

Warr WB (1992) Organization of olivocochlear efferent systems in mammals. In: Webster DB, Popper AN and Ray RR (eds.) *The Mammalian Auditory Pathway: Neuroanatomy*, pp. 410–448. New York: Springer.

Auditory System: Efferent Systems to the Auditory Periphery

M C Brown, Massachusetts Eye & Ear Infirmary, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

The auditory central pathway contains a number of descending systems that allow higher centers to modulate the processing performed at lower centers. The actions and functions of these descending systems are generally not known. Progress toward an understanding of descending systems has progressed most rapidly on the descending systems to the auditory periphery, since these are most accessible to investigation. This article reviews the two main efferent systems to the auditory periphery: (1) the olivocochlear neurons that innervate the organ of Corti in the inner ear and (2) the middle ear muscles that are linked to the ossicles of the middle ear.

Olivocochlear Neurons

Hair cell receptors within the auditory, vestibular, and lateral line systems receive an abundant efferent innervation from the central nervous system. This characteristic sets them apart from receptors in other sensory modalities. The auditory receptors receive innervation from the olivocochlear (OC) system, named because of its origin in the brain stem's superior olivary complex and termination in the cochlea (inner ear) (Figure 1). Two separate systems of OC neurons exist. Lateral olivocochlear (LOC) neurons originate in and near the lateral superior olive and project to auditory-nerve dendrites beneath the inner hair cells. Medial olivocochlear (MOC) neurons originate in the more medial parts of the complex and project to outer hair cells (Figure 1). LOC neurons are located predominantly on the same side of the brain as the cochlea that they innervate, whereas MOC neurons are distributed bilaterally.

Little is known about the functional role of the LOC neurons, although they may excite auditory nerve fibers. LOC neurons may be divided into subgroups that release acetylcholine, gamma-amino butyric acid (GABA), calcitonin gene-realated peptide, and/or dopamine. Much more is known about MOC neurons. They release acetylcholine, which binds to an alpha 9/alpha 10 nicotinic cholinergic receptor located in the outer hair cell membrane. Calcium inflow activates a calcium-activated potassium channel to open, which hyperpolarizes the hair cell. This effect decreases the outer hair cell electromotility, so they provide less amplification of the movement of the organ of Corti. The sensory inner hair cells thus receive less stimulation and respond

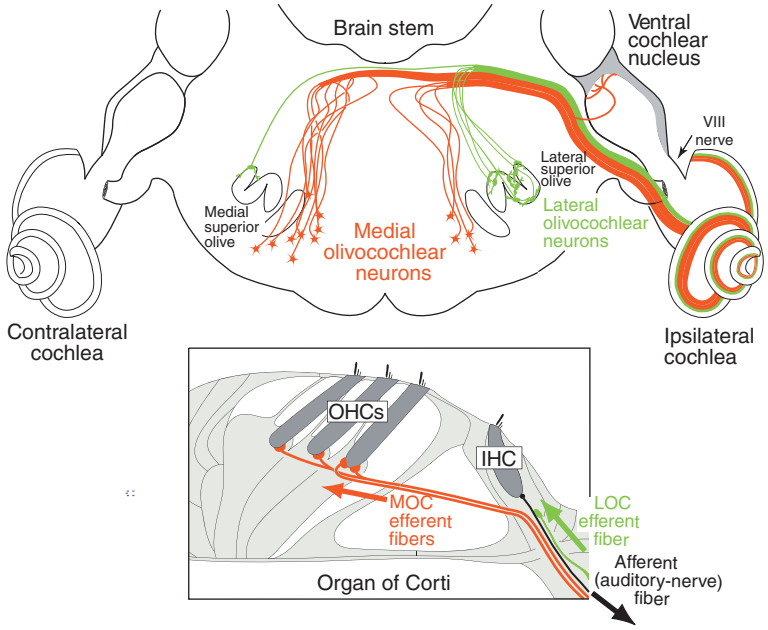

Figure 1 Neural pathways of the medial and lateral olivocochlear neurons in the brain stem and cochlea (inset). Only the efferent pathways to the right (ipsilateral) cochlea are illustrated. Large arrows indicate the direction of spike transmission. An afferent auditory-nerve fiber (type I) from an inner hair cell is illustrated in black; auditory nerve fibers (type II) from outer hair cells are not illustrated. Adapted from Liberman MC (1990) Effects of chronic cochlear de-efferentation on auditory-nerve response. *Hearing Research* 49: 209–224, with permission from Elsevier.

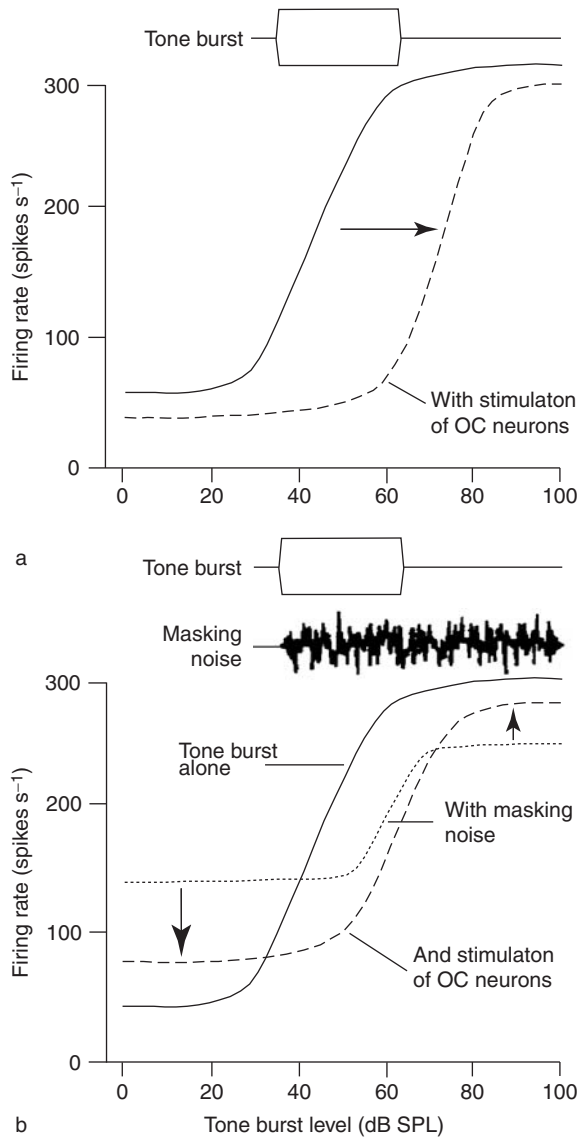

Figure 2 Rate vs. level functions for an auditory-nerve fiber. Insets above the graphs show the stimuli. (a) Response to tone bursts alone (solid curve) increases up to saturated rate, where further increases in level cannot be signaled by the fiber. Response to tone bursts plus stimulation of the MOC neurons (dashed curve). The effect of olivocochlear stimulation is to shift the function to the right (arrow), shifting the dynamic range of the fiber to higher sound levels. (b) Response to tone bursts in a continuous noisy background is 'compressed' (dotted function, solid curve reproduced from panel 'a'). Here, the fiber responds to the ongoing noise, boosting its rate even at low tone burst levels, and causing adaptation and less response at high tone burst levels. In this condition, stimulation of the MOC neurons (dashed curve) decreases the response to the ongoing noise (left arrow), decreasing adaptation and allowing a higher response to the transient tone burst (right arrow). The effect is to 'uncompress' the function, allowing the fiber to better signal changes in level of the tone.

less. Less excitatory neurotransmitter is released and there is less response of the auditory nerve fibers. The effect is a shift in the rising portion of the rate/level function (Figure 2(a)), with little effect on spontaneous rate and usually not much change on saturated rate. Overall, the most sensitive range of hearing is shifted to higher levels. Thus, controlling the dynamic range of hearing is an important function of the MOC neurons.

MOC neurons are likely to perform an additional function: to reduce the masking effects of noise on transient signals. To understand the role of MOC neurons, the effects of noise on rate versus level functions of auditory-nerve fibers will be briefly reviewed. The rate-level function to a transient tone burst alone is shown in Figure 2 (see solid lines). The effect of noise is a 'compression' of this function (compare dotted vs. solid functions in Figure 2(b)). The compression is caused at low tone burst levels by the fiber's response to the ongoing noise. Also, at high tone burst levels, the fiber is in an adapted state because of its response to the noise and thus responds less to the tone burst. MOC neurons 'unmask' this compressed level function by decreasing the response to the ongoing noise. This decreases the noise response at low tone burst levels (left arrow in Figure 2(b)). Because there is less response to the noise, the fiber is less adapted and can better signal the tone burst when it comes along at moderate and high levels (right arrow on Figure 2(b)). Overall, the function is less compressed and better able to signal changes in the tone even when masked by noise.

One additional function of both MOC and LOC neurons is to protect the inner ear from damage due to high-level sounds. However, these systems have a relatively slow time course of action. Thus, their protective effects are probably mostly seen for sounds of long duration rather than for brief, impulsive sounds. Finally, OC neurons may mediate selective attention, for example, by decreasing cochlear sensitivity during a visual task.

Middle Ear Muscles

Two muscles attach to the ossicles of the middle ear: the tensor tympani that attaches to the malleus and the stapedius that attaches to the stapes (Figure 3). The tensor tympani is innervated by a branch of the fifth cranial nerve, whereas the stapedius is innervated by a branch of the seventh cranial nerve. The muscles receive a dense innervation from motor neurons. The cat stapedius receives almost 1200 motor neurons in total, which is about one motor neuron per muscle fiber. Such high innervation ratios are only found in muscles requiring a high degree of control such as the muscles controlling movements of the eye. This innervation ratio implies a high degree of central control on the action of these muscles.

The effect of contraction of the middle ear muscles is generally to reduce transmission of sound through

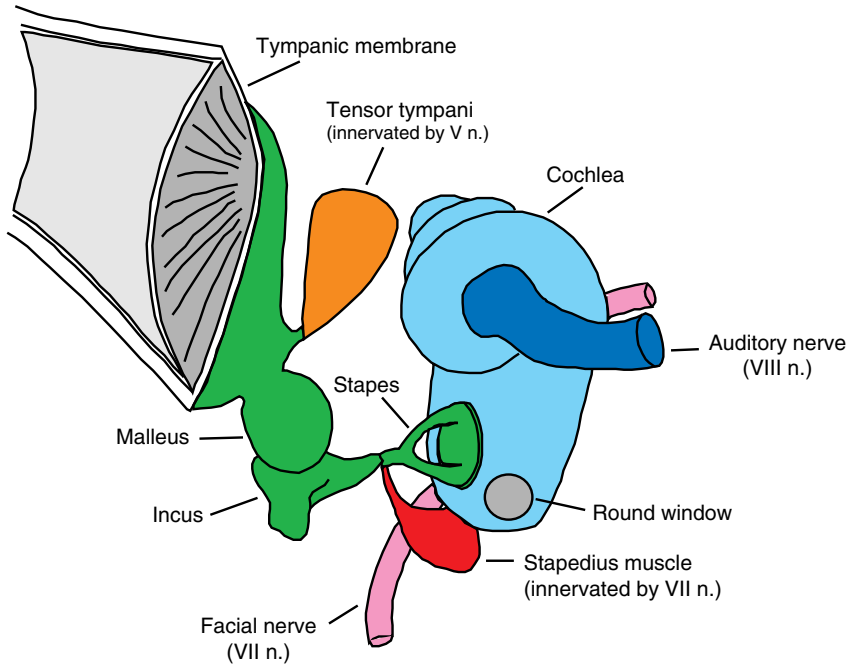

Figure 3 Diagram of the two middle ear muscles and their attachments to the ossicles. The stapedius muscle is attached to the stapes; it is innervated by a branch of the VII nerve. The tensor tympani is attached to the malleus; it is innervated by a branch of the Vth cranial nerve. The action of both muscles is to decrease sound transmission through the middle ear.

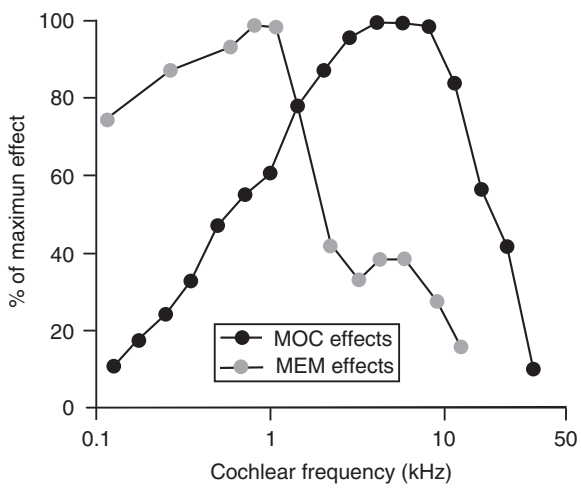

Figure 4 Relative strengths of the MOC and MEM effects. MOC data are the effects on auditory-nerve fibers during electrical activation of the OC bundle; MEM data are the changes in sound transmission during electrical stimulation of the stapedius muscle. All data are from anesthetized cats. Reproduced from Liberman MC and Guinan JJ, Jr. (1998) Feedback control of the auditory periphery: Anti-masking effects of middle ear muscles vs. olivocochlear efferents. *Journal of Communication Disorders* 31: 471–483, with permission from Elsevier.

the middle ear. The maximal effect is about 20 dB. While this is generally the same effect as stimulation of the OC neurons, the frequency ranges of the two systems are different and complimentary. Whereas the middle ear muscle contractions cause loss of sensitivity mainly for low frequencies, the action of the MOC neurons is mainly at middle and high frequencies (Figure 4). The middle ear muscles probably have generally similar functions to the OC systems. The middle ear muscle (MEM) reflex is particularly effective at decreasing the masking of auditory-nerve fibers tuned to high frequencies in the presence of low-frequency noise. This is an important situation because much environmental noise contains low frequencies. In addition, activity in the middle ear muscles of some animals precedes their own vocalizations, evidence that this system prevents self-stimulation.

The action of both MOC neurons and MEM motor neurons causes changes in otoacoustic emissions (OAEs). These emissions are sounds that emanate from the ear either spontaneously or evoked by sound stimulation. OAEs are thought to be generated by outer hair cells, which are electromotile. OAEs can be recorded with a microphone placed in the ear canal in animal studies as well as in human subjects. They can hence be used as a 'window' to indicate changes taking place in the outer hair cells. In one type of experiment, an OAE is recorded in one ear in response to a low-level sound stimulus and a reflex elicitor is presented to the opposite ear to evoke a reflex via a crossed central pathway (see below). As both MOC neurons (via their direct endings) and MEM muscles (via changing the sound transmission through the middle ear) can influence the outer hair

cells, both systems must be considered when changes in OAEs are observed.

Olivocochlear and Middle Ear Muscle Reflexes

Both MOC neurons and stapedius motor neurons respond to sounds such as pure tones and noise. The response characteristics of these neurons have been determined in recordings from anesthetized animals. The majority (about two-thirds) of MOC neurons respond to sound presented in the ear that they innervate. These units are called Ipsi units. About one-third of the units respond to sound in the opposite ear and are called Contra units. A small number, called Either-Ear units, respond to sound in either ear. Regional separation has been established for the Ipsi and Contra groups in labeling experiments. Ipsi units are located in the side of the brain stem opposite the cochlea they innervate, whereas Contra units are located on the same side as the cochlea they innervate (Figure 5). MOC neurons are sharply tuned to sound frequency. Their projections onto the cochlea are tonotopic and generally follow the cochlear frequency mapping set by the cochlear tuning. Thus, an MOC neuron that responds to high sound frequencies feeds back onto a high-frequency location in the cochlea.

The 'wiring diagram' of the MOC reflex is fairly well established. The afferent limb of this reflex is the type I auditory-nerve fibers sending messages centrally from inner hair cells. These fibers synapse in the cochlear nucleus. Cochlear nucleus neurons within the posteroventral subdivision, probably the stellate/multipolar cells, are the interneurons of the reflex. There are direct projections from the cochlear nucleus to the MOC neurons, which then constitute the efferent limb of the reflex. Thus, the reflex consists of three neurons each of which has been fairly well identified: type I auditory-nerve fiber, stellate/multipolar neuron of the cochlear nucleus, and MOC neuron.

The middle ear muscle with the biggest response to sound in many species (including humans) is the stapedius. Studies in animals demonstrate that like MOC neurons, stapedius motor neurons are divided into Ipsi, Contra, and Either-Ear groups. There is an additional group that responds only to sound in both ears. These different subgroups show regional clustering around and rostral to the facial motor nucleus. All motor neurons are located on the same side as the innervated muscle. For the middle ear muscle reflex, less is known about the interneurons. The reflex pathway probably consists of three or four neurons. The reflex interneurons within the cochlear nucleus have not been specified. The stapedius motor neurons

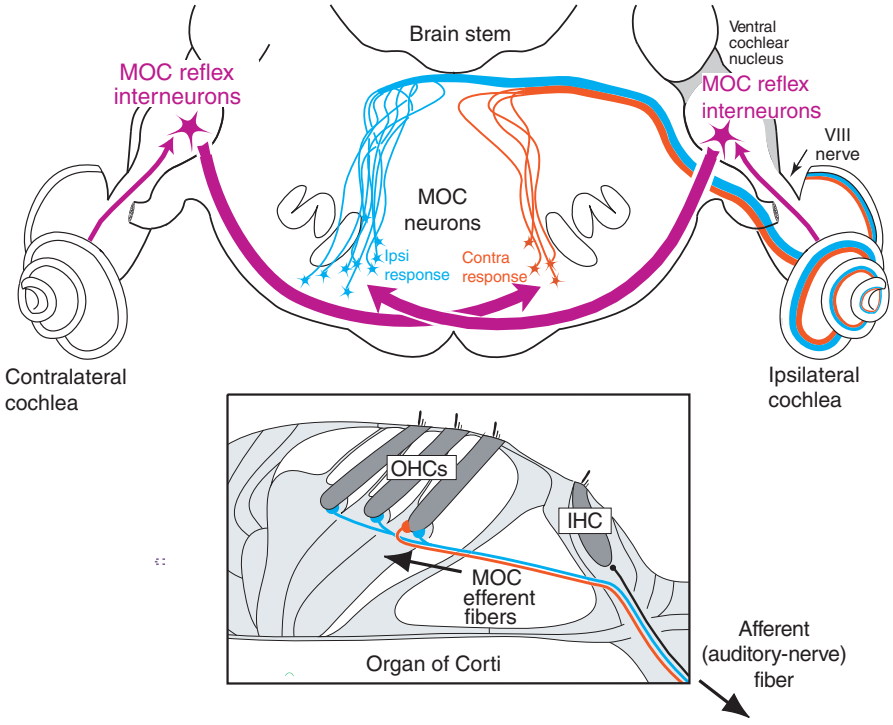

Figure 5 Neural pathways for the MOC reflex. The reflexes begin with activation of the auditory-nerve fibers. This information is transmitted to the cochlear nucleus, where interneurons are activated. MOC Ipsi neurons (blue color) receiving crossing information from the cochlear nucleus of the opposite side and in turn project back to the cochlea on that side. This ipsilateral reflex loop is thus double-crossed. MOC Contra neurons (red color) receive crossing information from the opposite cochlear nucleus but their axons in turn project back to the cochlea on the same side. This contralateral reflex loop is thus single crossed.

differ from MOC neurons in that they are fairly broadly tuned to sound frequency. They also have fairly high thresholds to sound (75–90 dB SPL), whereas MOC neurons can respond at much lower sound levels.

The simple reflex pathways shown in Figure 5 are under the control of higher centers. Evidence in support of this idea is that OC neurons receive inputs from higher centers such as the inferior colliculus and auditory cortex. In fact, work by Suga in the bat has shown that electrical stimulation of the bat cortex results in changes in cochlear physiology, a demonstration of the effectiveness of the higher centers on this brain stem pathway. The descending projections may enable the brain stem reflexes to be actively controlled – for instance, turned off when listening to low-level sounds near threshold and left on in noisy situations or perhaps when vocalizing. The details of this higher control remain to be worked out.

See also: Auditory Cortex Structure and Circuitry; Auditory System: Central Pathways; Auditory Cortex: Models; Hair Cells: Sensory Transduction; Hair Cell Differentiation.

Further Reading

Delano P, Elgueda D, Hamame C, and Roblest L (2007) Selective attention to visual stimuli reduces cochlear sensitivity in chinchillas. *The Journal of Neuroscience* 27(15): 4146–4153.

de Venecia RK, Liberman MC, Guinan JJ, Jr., and Brown MC (2005) Medial olivocochlear reflex interneurons are located in the posteroventral cochlear nucleus. *Journal of Comparative Neurology* 487: 345–360.

Guinan JJ, Jr. (1996) The physiology of olivocochlear efferents. In: Dallos P, Popper AN, and Fay RR (eds.) *The Cochlea*, pp. 435–502. New York: Springer.

Kobler JB, Guinan JJ, Jr., Vacher SR, and Norris BE (1992) Acoustic reflex frequency selectivity in single stapedius motoneurons of the cat. *Journal of Neurophysiology* 68: 807–817.

Liberman MC (1990) Effects of chronic cochlear de-efferentation on auditory-nerve response. *Hearing Research* 49: 209–224.

Liberman MC and Guinan JJ, Jr. (1998) Feedback control of the auditory periphery: Anti-masking effects of middle ear muscles vs. olivocochlear efferents. *Journal of Communication Disorders* 31: 471–483.

Suga N, Xiao Z, Ma X, and Ji W (2002) Plasticity and corticofugal modulation for hearing in adult animals. *Neuron* 36: 9–18.

Warr WB (1992) Organization of olivocochlear efferent systems in mammals. In: Webster DB, Popper AN, and Fay RR (eds.) *The Mammalian Auditory Pathway: Neuroanatomy*, pp. 410–448. New York: Springer.

Auditory System: Giant Synaptic Terminals, Endbulbs, and Calyces

D K Ryugo, Johns Hopkins University School of Medicine, Baltimore, MD, USA
G A Spirou, West Virginia University, Morgantown, WV, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The inescapable relationship between form and function in biological systems has focused attention on the giant synaptic endings in the central auditory system. Two of these giant endings are the subject of this article: endbulbs of Held that arise from myelinated auditory nerve fibers and calyces of Held that arise from globular bushy cells of the cochlear nucleus. These endings form hundreds of synapses with their targets and are implicated in fail-safe synaptic transmission that faithfully couples neural activity to environmental auditory events. The importance of this linkage is the preservation of neural timing that codes for all aspects of sound. Sounds only have significance to us when they occur over time. This preservation of timing within auditory signals enables the translation of prosodic utterances into perceptible speech and the processing of two cues used to localize sounds in space, interaural time differences (ITDs) and interaural level (intensity) differences (ILDs).

The ITD and the ILD pathways contain some of the largest, fastest, and most powerful synaptic endings in the central nervous system. The ITD pathway is initiated by the myelinated auditory nerve fibers that give rise to large and highly branched axosomatic endings called the endbulbs of Held. One or two of these endbulbs converge onto the cell body of a spherical bushy cell in the anteroventral cochlear nucleus (AVCN) that, in turn, transmits high-fidelity temporal information bilaterally to structures in the superior olivary complex. The ILD pathway is initiated by myelinated auditory nerve fibers that give rise to many smaller endbulbs that converge onto globular bushy cells in the AVCN. Globular bushy cells project to the contralateral medial nucleus of the trapezoid body (MNTB) and form a giant axosomatic synaptic ending called the calyx of Held. These two giant endings represent the key components for preserving timing in the auditory system.

Endbulbs of Held: Morphology

Large endings of auditory nerve fibers were originally described in Golgi-stained material by Hans Held in 1893. They were characterized in kittens as axosomatic, spoon-shaped endings with many filopodia that marked the end of the anterior branch of the fiber (Figure 1, left). When endings of the anterior branch were labeled in the same location of the adult cat, however, they appeared vastly different (Figure 1, right). The question was whether the differences were due to staining techniques or age of the animal studied. These endings were analyzed in an age-graded series of cats using the Golgi stain and horseradish peroxidase (HRP), demonstrating that the structural differences were due to the age of the animal examined (Figure 2). Auditory nerve fibers from every vertebrate examined have exhibited large axosomatic endbulbs of Held. These include the red-eared turtle, alligator lizard, chickens, owls, guinea pigs, rats, cats, monkeys, and humans. Endbulbs are found in the anteroventral cochlear nucleus and make synaptic contact with neurons known as spherical bushy cells.

The phylogenetically conserved ending implies functional significance for the processing of acoustic information. The size of the ending suggests powerful synaptic drive on the postsynaptic cell. The faithful transmission of presynaptic activity to the postsynaptic cell ensures that neural activity in the brain is yoked in time to acoustic events. The precision of the timing forms the substrate for sound localization as well as for translating sequences of sounds into perceptible speech.

Synapses: Presynaptic Endings

The reliability of transmission at this synapse was inferred from the observation that a small positive potential, called the prepotential, occurred 0.5 ms prior to the action potential. The prepotential was interpreted to represent the depolarization of the endbulb, and the coupling between prepotential and spike resulted in the notion of a fail-safe synapse that may nonetheless be modulated by inhibition. This ending has between 400 and 1500 release sites as calculated through serial section electron microscopy. The release sites of endbulbs have the morphology typical of normal synapses in the central nervous system (Figure 3). They are characterized by the presynaptic assemblage of clear, round synaptic vesicles, approximately 50–55 nm in diameter, that accumulate across from convex, dome-shaped postsynaptic densities (PSDs) of the postsynaptic membrane.

The postsynaptic density is visible as a small (on the order of 0.5 μm in diameter or less) aggregation of electron-dense material on the cytoplasmic side of the

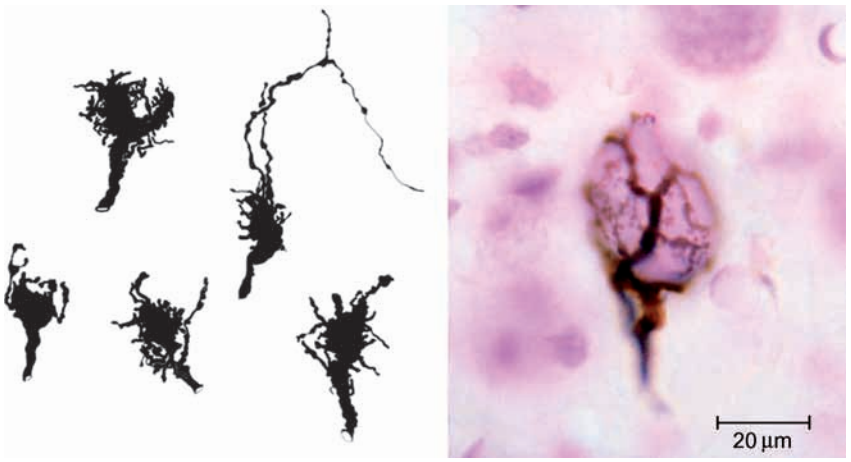

Figure 1 Left panel: Golgi-stained endbulbs of Held from neonatal kittens. They appear as spoon-shaped swellings with filopodia, identical to the original description by Hans Held in 1893. Right panel: HRP-labeled endbulb from an adult cat. The contrast in morphology of neonatal and adult cat endbulbs raised the question as to whether the differences were due to staining methods or age of the animals.

target membrane. The density is composed of neurotransmitter receptors, ion channels, signal transduction proteins, cytoskeletal proteins that anchor receptors to the postsynaptic site, and adhesion molecules for proper alignment of pre- and postsynaptic membranes. All of this machinery is essential for chemical synaptic transmission. The main demand on this transmission is speed. The auditory system processes sound for spatial localization, identification, and communication. Inherent to these functions is the ability to faithfully transmit rapid changes in the acoustic signals. The amino acid glutamate is the candidate for facilitating rapid transmission.

The unmistakable identification of glutamate as the neurotransmitter of the auditory nerve has been impeded by the widespread distribution of glutamate in nervous tissue and the inability to measure its release during synaptic activity. As a result, research strategies have been to analyze the effects of pharmacologic agonists and antagonists of glutamate, to study the properties of glutamate receptor subunits, and/or to characterize glutamate transporters. Quantitative immunohistochemistry demonstrated greater labeling over primary auditory nerve endings compared to nonprimary endings (containing flat or pleomorphic synaptic vesicles) or glia. Moreover, potassium-induced depolarization that depleted glutamate significantly lowered such staining in endbulb terminals. The presence of glutamate and one of its metabolic precursors in the endbulb and the complementary distribution of glutamate receptors (GluRs) in cochlear nucleus neurons represent strong inferential evidence that glutamate is the neurotransmitter. The involvement of glutamate in auditory nerve synapses is greatly strengthened by observations that large postsynaptic currents in cochlear nucleus slice recordings are blocked by glutamate receptor antagonists.

Synapses: Postsynaptic Targets

Receptors that reside in the PSD of the target cell seem to be primarily of the ionotropic type that are formed by several subunits surrounding a central ion pore. Immunogold and immunoperoxidase methods demonstrate a strong reaction for GluR2/3 and GluR4. Since GluR1 shows little or no immunolabeling at endbulb synapses, it is inferred that most synaptic α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors are composed of GluR3 and GluR4 subunits. These subunits facilitate fast transmission because of rapid desensitization. Moreover, calcium permeability due to the lack of GluR2 could account for the rapidly decaying responses of the postsynaptic spherical bushy cells. Immunolabeling of metabotropic glutamate receptors (mGluR1α) is modest at the endbulb synapse.

N-Methyl-D-aspartate (NMDA) receptors feature a voltage-dependent calcium channel. Opening of the NMDA channel occurs as a result of depolarization via the AMPA receptors. There are a number of different NMDA receptor subunits, including NR1, NR2A–D, and NR3. Receptors are composed of NR1 plus one or more variants of NR2, the combination of which determines physiological properties. The role of these receptors at the endbulb synapse is not known; they are present in the developing auditory system but they diminish by weaning.

The spiking pattern of auditory neurons is determined in part by the type of voltage-sensitive potassium channels expressed. The activation and deactivation

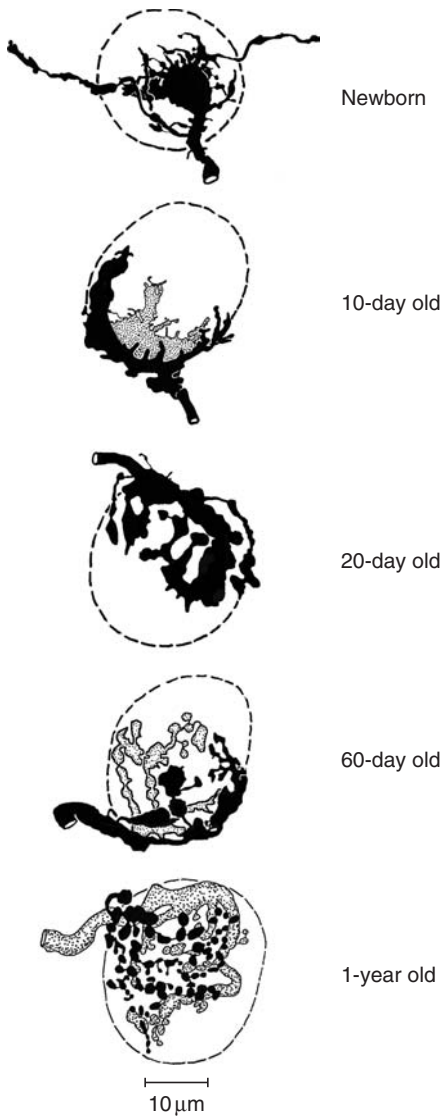

Figure 2 Developmental sequence of endbulbs of Held. Endbulbs were stained using either the Golgi or HRP method. Note the progression of endbulb structure, from a spoon-shaped ending to a highly branched arborization. This result emphasizes the importance of age on structure and presumably function.

kinetics of potassium channels provide for rapid repolarization of action potentials. These potassium channels are abundant in spherical bushy cells. The fast response is necessary because auditory neurons maintain timing while responding to high rates of synaptic input. The likely candidates for the low voltage-activated and high voltage-activated potassium currents in auditory neurons are Kv1.1 and Kv3.1, respectively. The combination of AMPA receptors and these potassium channels accounts for the brevity and rapidity of auditory neuronal signaling.

Physiology

The time-varying nature of natural sounds requires that accurate identification depends on the ability of auditory neurons to encode timing. Vertebrates have the capacity to discriminate between pitches that differ by as little as 2 μs in the period of the sound waves, and to resolve differences on the order of 10 μs in the time of arrival of a sound at the two ears. The duration of single action potentials is significantly longer than these timing differences, so it is remarkable that there are biological mechanisms that reliably preserve and transmit this information. There are, however, a number of specializations in auditory neurons that facilitate the processing demands.

One to two endbulbs of Held converge upon single spherical bushy cells in the anteroventral cochlear nucleus. It is estimated that a single auditory nerve fiber will activate between 2000 and 10 000 AMPA receptors. Excitatory postsynaptic potentials (EPSPs) in bushy cells are brief and have very rapid rise times. The consequences are (1) threshold is reached quickly, (2) action potentials occur reliably and with little temporal jitter, (3) endbulb depolarization produces only one postsynaptic spike, and (4) refractory period is short. As a result, spherical bushy cells exhibit the ability to 'follow' repeated stimulation up to 300 Hz, which is near the maximum firing rate of auditory nerve fibers, but there is also 'enhancement' – these cells perform better than the fibers. These factors contribute to the coupling in time of neural discharges to acoustic events by the auditory system.

Plasticity

The efficiency of synaptic transmission at the endbulb synapse is well known and is determined by the amplitude and time course of evoked excitatory postsynaptic currents (EPSCs). Both spontaneous and evoked EPSCs, however, exhibit large fluctuations in amplitude and time course. Because receptor desensitization did not play a role in this variability, it was proposed that variability in response characteristics was due to intrinsic fluctuations in release probability and numbers of available receptors. Modulation of release probability is achieved by varying calcium and protein kinase C levels, which in turn can elicit short-term plasticity of responses.

Evidence for long-term plasticity arises in different forms. Auditory nerve fibers exhibit separate groupings on the basis of spontaneous discharge rates and thresholds to activation. Low-spontaneous-rate (SR) fibers have high thresholds, represent approximately 40% of the population of myelinated fibers, and give

Figure 3 Electron micrograph through a section of endbulb containing a synapse (flanked by arrowheads). The ending contains mitochondria and many clear, round synaptic vesicles. At the synapse, there is the signature arch of the postsynaptic membrane accompanied by a fuzzy membrane thickening known as the postsynaptic density. The postsynaptic density contains neurotransmitter receptors, ion channels, and structural proteins. Reproduced from Ryugo DK, Huchhton DM, Pongstapron T, and Niparko JK (1997) Ultrastructural analysis of primary endings in deaf white cats: Morphologic alterations in endbulbs of Held. *Journal of Comparative Neurology* 385: 230–244, with permission. Copyright (1997) by Wiley-Liss. Inc.

rise to endbulbs having very complicated arborizations. In contrast, high-SR fibers have low thresholds, represent roughly 60% of the population of myelinated fibers, and give rise to endbulbs have less complicated shapes (Figure 4). Under normal conditions, these two populations of endbulbs exhibit different morphology.

In the event of congenital deafness, however, the auditory nerve fibers have little spike activity, and the resulting endbulbs appear withered (Figure 5). Moreover, the endbulb synapses of congenitally deaf cats have a pathologic appearance when compared to those of normal hearing cats (Figure 6(a), asterisks). Those of congenitally deaf cats lack synaptic vesicles and show hypertrophied, flattened postsynaptic densities (Figure 6(b), arrowheads). These synapses could be rescued if miniaturized human cochlear implants were surgically placed into congenitally deaf cats (Figure 6(c), asterisks). The implanted cats were stimulated 7 h per day, 5 days per week for 10–12 weeks using the same processing strategy and programming software used in children. Environmental sounds had biological significance because the implanted cats could routinely be 'called' for a food reward. Since the cochlear nucleus gives rise to all ascending auditory pathways, the preservation of the endbulb–bushy cell circuit would support the faithful transmission of temporal cues contained within auditory signals. It is proposed that the changes in endbulb synapses by cochlear implants represent one key to the development of integrative and cognitive brain functions reflected in aural and oral communication in deaf children.

Endbulbs of Held serve as a near one-to-one pipeline from a single inner hair cell to a single spherical bushy cell. They are found in a wide variety of vertebrates, emphasizing evolutionary pressure for their preservation. They are known for their reliability but also exhibit structural and functional plasticity under different conditions. Much remains to be learned about these important synaptic endings.

Calyces of Held

The calyx of Held is perhaps the largest nerve terminal in the mammalian central nervous system (CNS) and covers 25–50% of the postsynaptic somatic surface (Figure 7). It was recognized as the terminal portion of an axon by Hans Held (1893), whose name became associated with this structure via reference to the 'calyces of Held' by other neuroanatomists of his time. This structure figured in the great debate over the neuron doctrine that blazed at the end of the nineteenth century. Held and Camillo Golgi, two leaders in the nascent field of neuroanatomy, adhered to the theory that the nervous system was a continuous reticulum, resembling a nerve net. Held interpreted rodlike structures seen in the postsynaptic neuron to be the injection of calyx fibrils into the cytoplasm of its target. Ramón y Cajal claimed that the rodlike structures belonged strictly to the postsynaptic neuron. His analysis of the calyx as supporting the integrity of individual nerve cells was eventually proved correct.

The advent of the electron microscope and single-neuron recording techniques in the mid-twentieth century opened a second phase of study of this large terminal and its functional role in hearing. Round synaptic vesicles, structural evidence for an excitatory connection, fill the terminal and associate with

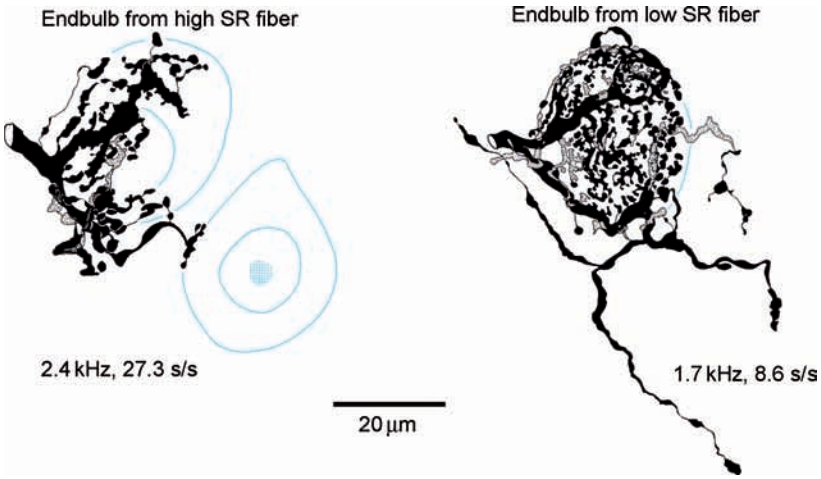

Figure 4 Camera lucida drawings of HRP-labeled endbulbs stained by intracellular injections. Two groups of auditory nerve fibers may be distinguished on the basis of spontaneous discharge rate (SR) and sensitivity. Because they are found at all frequency ranges, they are hypothesized to underlie separate aspects of stimulus processing. Note that the endbulb from the low-SR fiber exhibits more complex branching compared to that of the high-SR fiber. Reproduced from Sento S and Ryugo DK (1989) Endbulbs of Held and spherical bushy cells in cats: Morphological correlates with physiological properties. *Journal of Comparative Neurology* 280: 553–562, with permission. Copyright 1989 by Alan R. Liss.

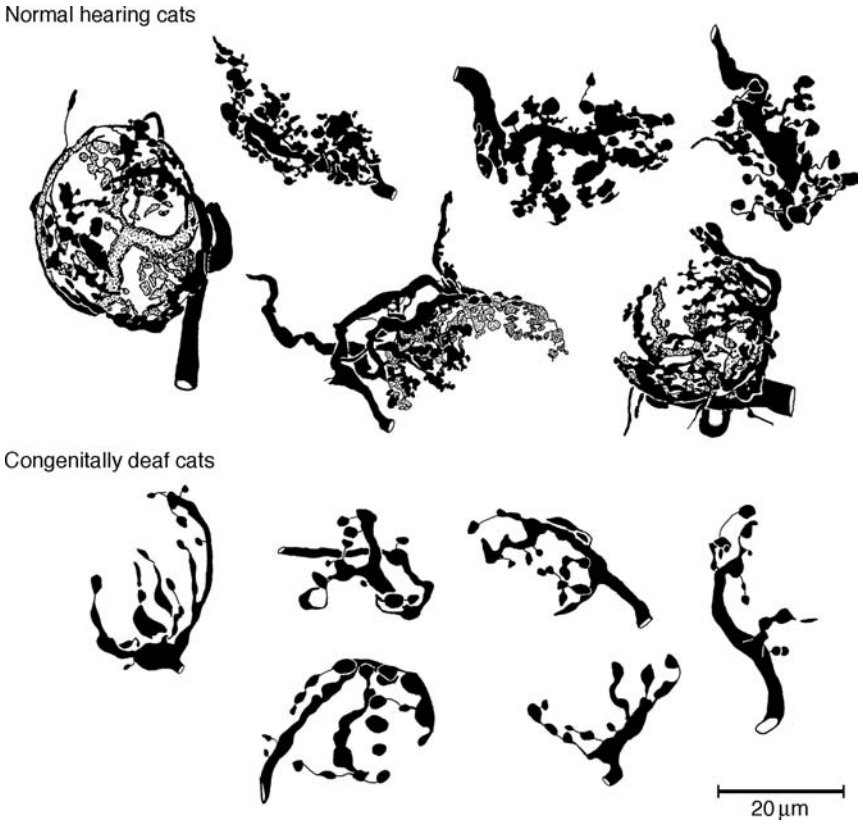

Figure 5 Camera lucida drawings of HRP-labeled endbulbs from normal hearing cats and congenitally deaf cats. There is no spontaneous or driven spike activity in auditory nerve fibers of congenitally deaf cats. Note that the endbulbs of deaf cats appear withered and atrophic. Spike activity has an influence on endbulb shape. Reproduced from Ryugo DK, Rosenbaum BT, Kim PJ, Niparko JK, and Saada AA (1998) Single unit recordings in the auditory nerve of congenitally deaf white cats: Morphological correlates in the cochlea and cochlear nucleus. *Journal of Comparative Neurology* 397: 532–548, with permission. Copyright 1998 by Wiley-Liss, Inc.

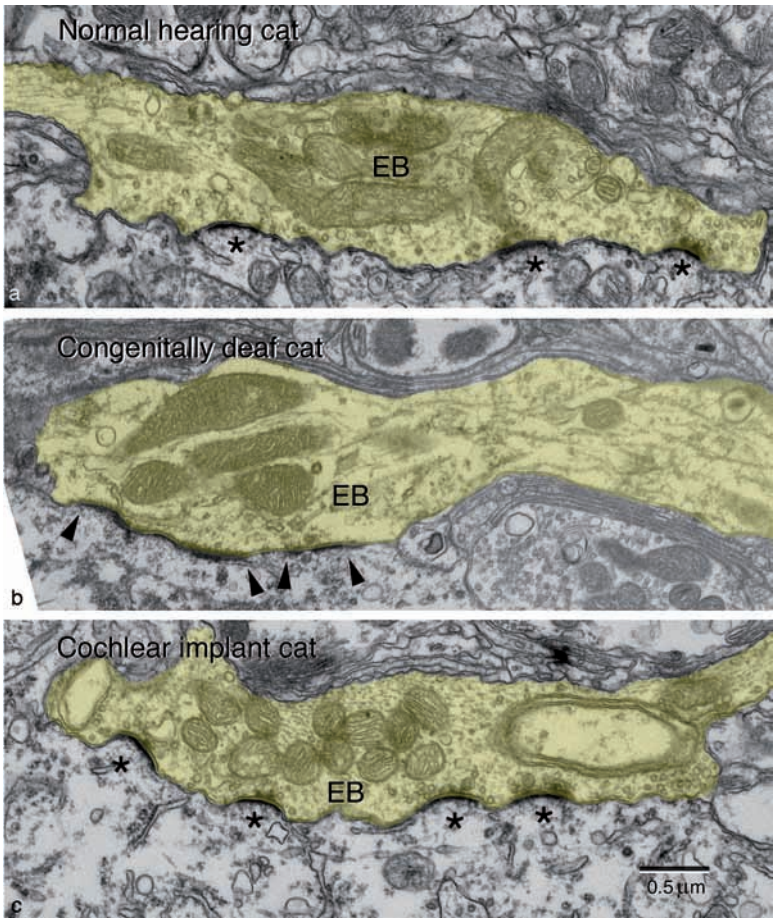

Figure 6 Electron micrographs of endbulbs (EB; yellow) illustrate the restorative effect of activity on synapses. (a) The endbulb of a normal hearing cat exhibits the typical arched postsynaptic density with accumulations of synaptic vesicles (asterisks). (b) The endbulb of congenitally deaf cats, by contrast, contains few synaptic vesicles and the postsynaptic densities are flat and hypertrophied (arrowheads). (c) The endbulb of a congenitally deaf cat that received stimulation from a cochlear implant exhibits synapses with normal morphology (asterisks). Reproduced from Ryugo DK, Kretzmer EA, and Niparko JK (2005) Restoration of auditory nerve synapses in cats by cochlear implants. *Science* 310: 1490–1492, with permission. Copyright 2005 by American Association for the Advancement of Science.

multiple synaptic sites, later shown to range in number from 400 in mice to 600 in rat to 2500 in cat. Resembling the endbulb, the calyx was shown to exhibit a physiological signature in extracellular recordings, whereby a short prepotential preceded an action potential in the postsynaptic neuron, located in the medial nucleus of the trapezoid body within the superior olivary complex, by about 0.5 ms. Each postsynaptic action potential is typically associated with a prepotential. Intracellular recording and labeling of single neurons unambiguously associated the calyx nerve terminal with its origin from the globular bushy cell located in the ventral cochlear nucleus, and revealed that globular bushy cells innervate other cell groups in the superior olivary complex via small bouton endings (Figure 8). As described in the previous section of this article, globular bushy cells are themselves contacted by multiple large endings of auditory nerve fibers. The ability of bushy cells to encode temporal fine structure in the incident acoustic wave and their involvement in brain stem circuits that mediate sound localization implicated the calyx in functional roles for localizing sound in space.

A third phase of investigation began in 1994 with the first intracellular recording directly from the calyx in brain slices using patch electrode techniques (Figures 9(a) and 9(b)). Over the ensuing years the calyx has been studied as a model for glutamatergic neurotransmission throughout the brain. The ability to record pre- and postsynaptic signals simultaneously (Figure 9(c)) has made this a powerful preparation. To date, bulk labeling techniques using tract tracing molecules, viral transfection, and immunocytochemistry have not revealed a single instance of more than one calyx terminating onto a single postsynaptic neuron. Contrary to the pervasiveness of endbulbs of Held, calyces are typically mammalian structures. Their presence in humans, however, can be

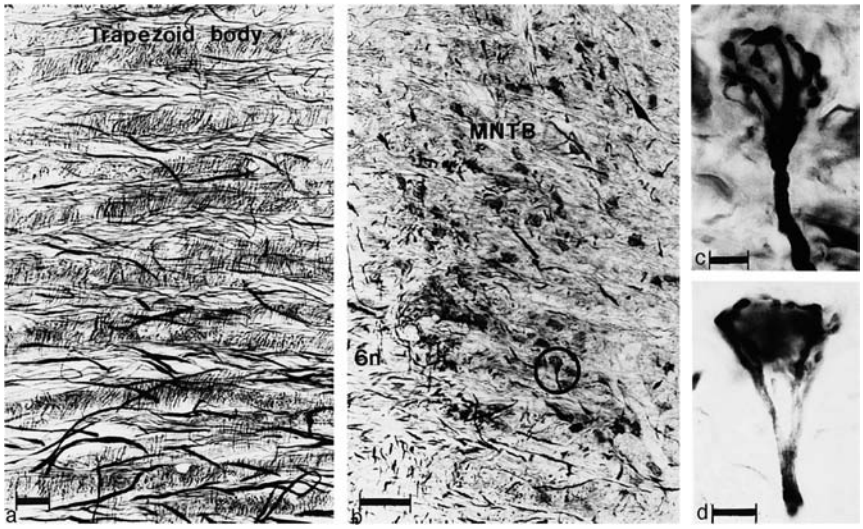

Figure 7 Globular bushy cells in cats are immunolabeled by antisera to the putative calcium-binding protein PEP-19. (a) Thick, immunolabeled fibers in the ventral trapezoid body belong to globular bushy cells. (b) Calyces of Held are immunolabeled throughout the medial nucleus of the trapezoid body (MNTB), and are easily seen because the MNTB neurons are immunonegative. (c) The large-diameter calyceal axon branches near its postsynaptic target. (d) In some cases, large branches occur further from the soma. Scale bar = 50 μm (a), 100 μm (b), 10 μm (c, d). Reproduced from Berrebi AS and Spirou GA (1997) PEP-19 immunoreactivity in the cochlear nucleus and superior olive of the cat. *Neuroscience* 83(2): 535–554, with permission from Elsevier Science Ltd.

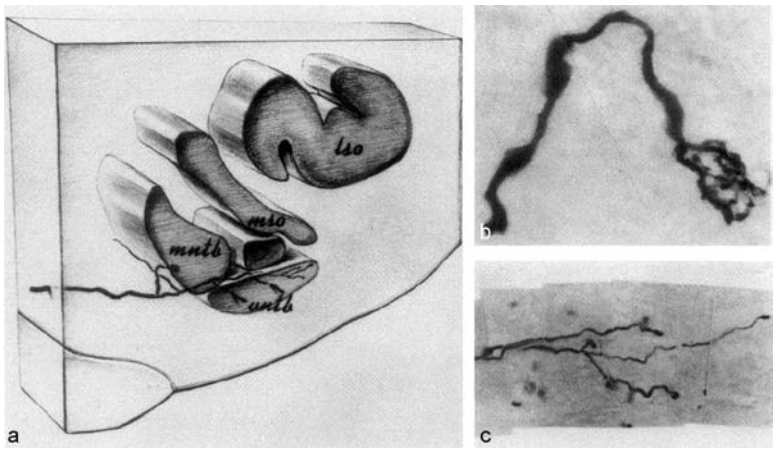

Figure 8 Branching pattern and terminal profiles of a single globular bushy cell input to the superior olivary complex of the cat. (a) Artist's rendering of the branching and terminal profile for the labeled terminals shown in (b) and (c). A thick section through the superior olive in the transverse plane is pictured, viewed from a caudal perspective (mso, lso: medial and lateral superior olive). The large-diameter axon branches and delivers a calyx into the medial nucleus of the trapezoid body (mntb). The calyx is pictured in panel (b). The parent axon turns rostrally, but also branches further within the superior olive. These branches terminate as boutons; those within the ventral nucleus of the trapezoid body are pictured in panel (c). Reproduced from Spirou GA, Brownell WA, and Zidanic M (1990) Recordings from cat trapezoid body and HRP labeling of globular bushy cell axons. *Journal of Neurophsyiology* 63(5): 1169–1190, with permission. Copyright 1990 by The American Physiological Society.

questioned, since their postsynaptic cell group, the MNTB, has not been unambiguously identified in the human brain stem.

Internal Structure

Mature calyces resemble clawlike structures of complex geometry. Each calyx is formed from a large-diameter axon that branches repeatedly, sometimes via thin necks, into single swellings or series of swellings. Early in development the calyx is less branched (see later), and hence derived its name in reference to a floral calyx, the collection of green sepals that surround the petals of the flower. All parts of the calyx except the necks contain active zones and are filled with round synaptic vesicles. The large size of the calyx and number of active zones pose challenges for neurotransmission at high rates, where ionic

Figure 9 Intracellular recordings from the calyx and its postsynaptic target can be performed simultaneously. (a) Differential interference contrast (DIC) image of a medial nucleus of the trapezoid body (MNTB) cell (asterisk) and calyx (arrow points to calyx covering the right side of the cell body). A patch pipette is positioned to record from the calyx. (b) The calyx pictured in panel (a) is filled through the patch pipette with fura dye, which fluoresces in the presence of calcium. This technique confirms the selective recording from the calyx. (c) DIC image showing simultaneous recording from the MNTB cell (left pipette electrode) and the calyx (right pipette electrode). Scale bar = 10 μm (a, b), 15 μm (c). Reproduced from Billups G and Forsythe ID (2002) Presynaptic mitochondrial calcium sequestration influences transmission at mammalian synapses. *Journal of Neuroscience* 22(14): 5840–5847, with permission. Copyright 2002 by the Society for Neuroscience.

gradients must be maintained and neurotransmitter rapidly cleared from the synaptic cleft. Early electron microscopic studies revealed regions of the calyx that are apposed to the postsynaptic membrane and that are studded with synaptic sites and puncta adherentia. These contacts form directly onto the somatic membrane or onto somatic appendages, which are hairlike protrusions of the postsynaptic somatic membrane. Interdigitated with these contact regions are separations of the pre- and postsynaptic membranes, called extended extracellular spaces (EESs; Figure 10). These spaces can be infiltrated with glial processes and may clamp extracellular ionic concentration in the cleft and facilitate diffusion of neurotransmitter away from the cleft. These factors minimize desensitization of postsynaptic AMPA receptors because glutamate transporters, which recover synaptically released molecules, are found on glial membranes in the EES and in greater number on glial membranes surrounding the calyx. The fenestrated geometry of mature calyces also supports rapid diffusion from the synaptic cleft. A specialized organelle complex, called the mitochondrion-associated adherens complex (MAC), tethers mitochondria to a punctum adherens within 200 nm of the presynaptic membrane. This complex can be ringed by synaptic sites, and also is hypothesized to support high-fidelity synaptic transmission.

Development

The 1:1 innervation ratio between calyces and their postsynaptic targets provides a developmental model for studying the precise formation of topographic connections between cell groups. The calyx proceeds through four stages of growth: (1) appearance of the migratory growth cone in the vicinity of the MNTB, (2) expansion of growth cone processes into a single large ending called a protocalyx, (3) further expansion of the protocalyx into a spoon-shaped structure called the young calyx, which envelopes about 50% of the postsynaptic somatic surface, and (4) the refinement of the spoon-shaped structure into a more clawlike calyx (Figure 11). The timing for these stages of calyx formation has been most studied using rodent animal models. Postsynaptic neurons of the MNTB are born between embryonic days 12 and 14 and within a day or two, axons of ventral cochlear nucleus cells have crossed the midline into the region where the MNTB will arise. Since the MNTB and other cell groups of the superior olivary complex do not coalesce into recognizable entities until P17, innervating axons may appear in this vicinity prior to the postsynaptic targets.

Young calyces appear in relative abundance by P5, so physiological studies of the calyx typically commence at this age. Synapses onto MNTB cells, evident in the electron microscope as early as the day of birth, generate postsynaptic currents which grow in amplitude from 50 pA at P0 to as large as 10 nA by P4. The growth in postsynaptic current accelerates between P2 and P4. Serial section electron microscopy reveals that protocalyces appear at P2, and can expand into young calyces within 48 h. This rapid growth period prompted the question of whether competition among inputs was occurring at the earliest age of young calyx formation at P4. Minimal stimulation techniques suggest multiple inputs onto a minority of MNTB neurons (15%). Three-dimensional reconstructions of MNTB cells and their inputs from electron micrographs reveal multiple inputs onto a

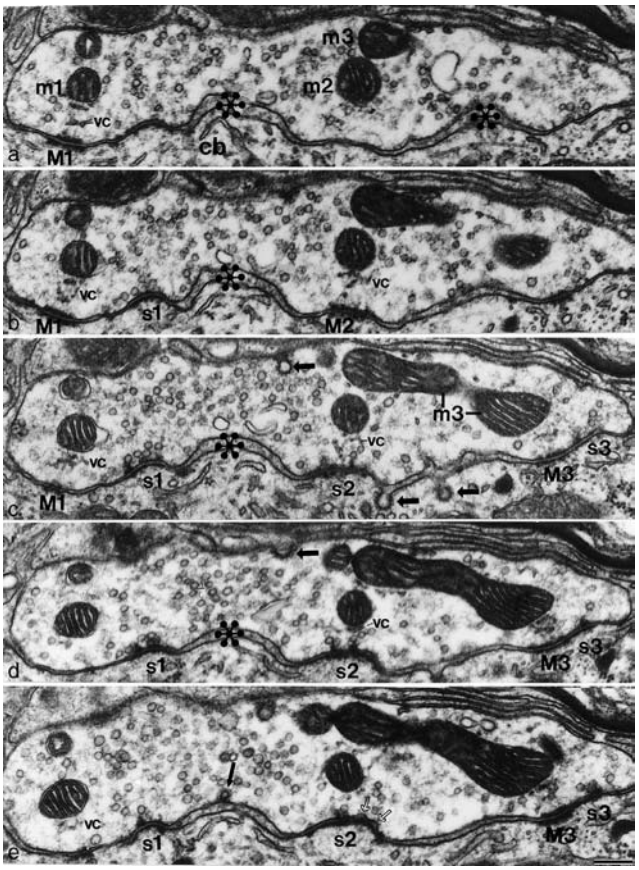

Figure 10 Serial electron micrographs from a calyceal swelling contacting a cell body (cb) in a cat. Synapses (s) and mitochondrion-associated adherens complexes (MACs; labeled M) lie adjacent to one another. These functional contacts can be separated by extended extracellular spaces (asterisk), where pre- and postsynaptic membranes are separated. MACs are comprised of a mitochondrion (m), vesicular chain of membrane between the mitochondrion and cell membrane (vc), and punctum adherens (adjacent to M label). Individual synapses and MACs are numbered. Arrows indicate coated, and presumably endocytosing, vesicles. Scale bar (shown in panel E for all panels) = 0.25 μm. Reproduced from Rowland KC, Irby NK, and Spirou GA (2000) Specialized synapse-associated structures within the calyx of Held. *Journal of Neuroscience* 20: 9135–9144.

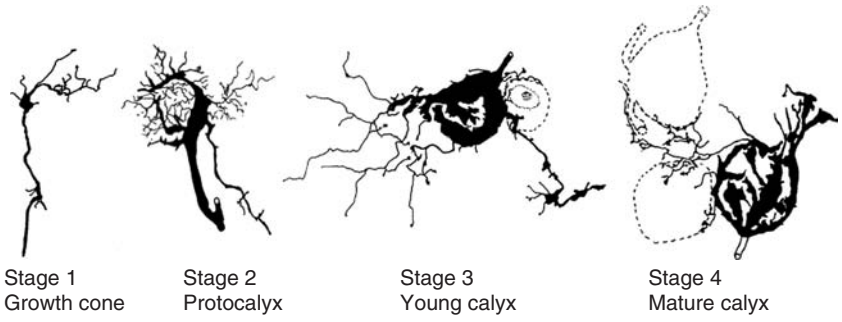

Figure 11 Stages of calyx growth, defined using Golgi stain of pouch-young opossum. Stage 1: Migratory growth cones enter the region of the medial nucleus of the trapezoid body. Stage 2: Growth cones expand into a protocalyx. Stage 3: The protocalyx further expands into a cup-shaped structure called a young calyx, which envelops much of the postsynaptic cell body. Stage 4: The young calyx is pruned into a highly branched, mature calyx. Reproduced from Morest DK (1968) The growth of synaptic endings in the mammalian brain: A study of the calyces of the trapezoid body. *Zetschrift fur Anatomie und Entwicklungsgeschichte* 127: 201–220, with kind permission of Springer Science and Business Media.

similarly small percentage of cells (12%; Figure 12). Therefore, specificity in matching of axonal inputs to their targets occurs at early stages of nerve terminal formation. In other neural systems, axons typically compete for innervation territories, and these are refined via experience-driven processes. In the calyx system, the cochlea is only beginning to transduce sound when young calyces are formed, so calyx

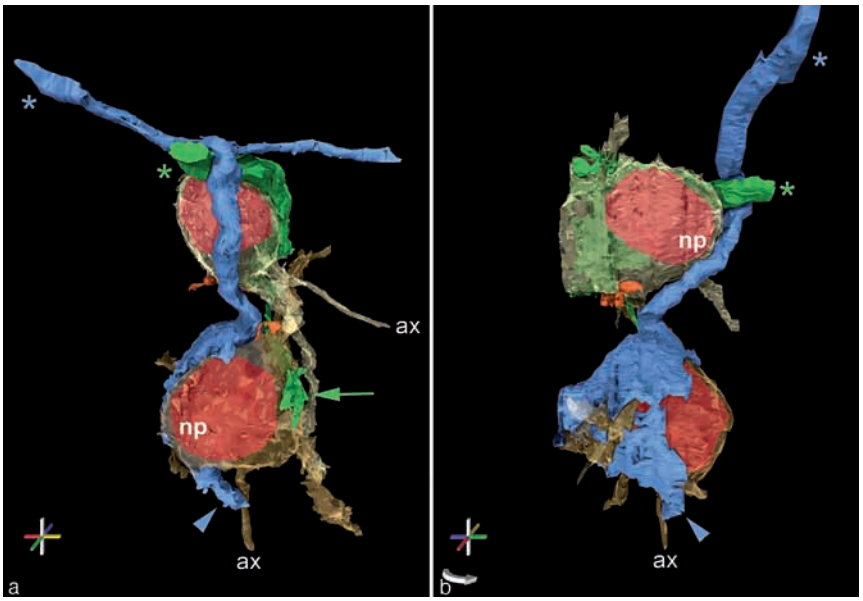

Figure 12 A 3-D reconstruction of two neighboring MNTB cells at P4. (a, b) Two views of a pair of representative MNTB cells that were contacted by single large terminals. Panel (b) is a 90° clockwise rotation of panel (a) about the vertical axis. Rotated scale axes and arrows are provided for reference. MNTB cells are transparent and colored tan (top cell) or brown (bottom cell) for distinction of overlapping dendrites and axons. Cell bodies are ovoid in shape and have two to three dendrites (only proximal dendrites, an axon (ax), and an eccentric nucleus (red) are included in the reconstruction). The blue calyx terminal consists of two large pieces linked by a narrow bridge (panel (b)) and has a solitary process extending away from the cell body (blue arrowhead). The axon (ax) of this terminal was traced to its branch point from a trapezoid body fiber. The green terminal on the top cell extends a collateral that contacts the bottom cell (green arrow in panel (a)). Innervation territories avoid the nuclear pole (np) of the MNTB cells (top cell, panel (b); bottom cell, panel (a)). The nuclear pole is defined by the close apposition of the nuclear membrane and cell membrane. Orange terminals could not be linked to other inputs onto the cell or to an axon. Asterisks indicate the axons for each terminal of the corresponding color. Scale axes are 2 μm from origin to end, 4 μm in total length. Note that scale axes in the 3-D perspective views apply most accurately to the middle depth of field. Reproduced from Hoffpauir BK, Grimes JL, Mathers PH, and Spirou GA (2006) Synaptogenesis of the calyx of Held: Rapid onset of function and one-to-one morphological innervation. *Journal of Neuroscience* 26(14): 5511–5523, with permission. Copyright 2006 by the Society for Neuroscience.

growth is an experience-independent process. Competition among calyx-forming inputs may occur earlier among the growth cones or protocalyces.

Short-Term Synaptic Plasticity

In mammals, the mechanisms that underlie synaptic plasticity have proved difficult to study, especially in the CNS. However, the calyx of Held has offered a solution to this quandary. The power of simultaneous pre- and postsynaptic recording is increased when these techniques are combined with other approaches, such as pharmacologic manipulations and fluorescence imaging. Important information has been gleaned from the calyceal connection with the MNTB cell, tempered by the reminder that studies are conducted during the maturation process of the terminal and its postsynaptic partner, and often not at physiological temperatures. Most studies employ the rat as the animal model for physiological studies, and recordings are made from animals between P5 and P15. This age range brackets the onset of hearing in rodents between P8 and P12 and includes the early stages of myelination.

The calyx-to-MNTB neuron connection has been considered an efficient converter of excitation into inhibition, although recent *in vivo* demonstrations of prepotentials that do not elicit postsynaptic spikes challenge that view. Cellular studies *in vitro* reveal modulatory events, such as direct glycinergic and GABAergic effects on the calyx, that also provide a different picture of synaptic processing and integration by MNTB cells. Within a week after its formation, the calyx exhibits a nearly adultlike ability to follow high rates of stimulation with good temporal fidelity. Seemingly contrary to this observation, postsynaptic EPSCs depress rapidly during repetitive stimulation, although depression is lessened by 2 weeks of age. The calyx connection offers an experimental paradigm to relate presynaptic Ca^{2+} levels, which are strongly

linked to short-term plasticity, to postsynaptic currents in a mammalian nerve terminal. Experimental manipulations of Ca^{2+} concentrations using fluorophores, Ca^{2+} chelators, and caged Ca^{2+} have led to quantitative relationships between Ca^{2+} concentration within the calyx and vesicle release. Resting Ca^{2+} levels are estimated to increase at the active zone to approximately 10–20 μM following a single action potential. In adult animals, Ca^{2+} enters the terminal primarily via P/Q-type Ca^{2+} channels. Facilitation of postsynaptic currents can be elicited by prepulse depolarization of the calyx terminal or by posttetanic potentiation. During potentiation, postsynaptic current amplitudes follow presynaptic Ca^{2+} levels, which increase by an order of magnitude during the tetanic stimulus. Ca^{2+} levels are thought to couple with highly nonlinear Ca^{2+} sensors to facilitate vesicle release. These studies provide strong evidence for a causal relationship between Ca^{2+} levels and potentiation. Although the vesicle release probability is likely to vary among active zones, the average release probability across the multiple active zones is estimated to be on the order of 0.25, with a readily releasable pool of vesicles numbering between 1500 and 4000. Given the sensitivity of the vesicle release mechanism to small changes in Ca^{2+}, action potential amplitude and duration must be tightly controlled, especially during high rates of activity. Specializations that mediate these properties include the exclusion of sodium channels from the terminal and the positioning of these channels along a lengthy heminode of the axon leading into the calyx. In sum, many mechanisms operate in concert to affect Ca^{2+} levels at microdomains surrounding individual active zones and regulate vesicle release across hundreds of synaptic sites.

Conclusion

The large axosomatic endings, the endbulbs and calyces of Held, are more than independent, synaptic transmission machines. Recent evidence suggests that they are subject to modulatory influences. Glycine acts presynaptically to depolarize the terminal and facilitate neurotransmitter release. Retrograde signaling via endocannabinoids occurs at these endings, perhaps enhancing the sensory experience. Certain second messenger systems can influence vesicle mobilization, and activation of metabotropic glutamate receptors can affect vesicle release probability and the size of the readily releasable pool of vesicles during trains of activity. The next few years promise advancement of our knowledge of the mechanistic elements of synaptic transmission at these large synaptic terminals. Studies at the cellular level will lead to progress at the molecular level to complement *in vivo* exploration of the role of these remarkable structures in the neural encoding of natural sounds.

See also: Auditory System: Central Pathway Plasticity; Auditory System: Central Pathways; Auditory Localization; Cochlear Mechanics; Cochlear Development; Deafness; Sound Localization: Neural Mechanisms; Temporal Processing in the Auditory Pathway.

Further Reading

Berrebi AS and Spirou GA (1997) PEP-19 immunoreactivity in the cochlear nucleus and superior olive of the cat. *Neuroscience* 83(2): 535–554.

Billups G and Forsythe ID (2002) Presynaptic mitochondrial calcium sequestration influences transmission at mammalian synapses. *Journal of Neuroscience* 22(14): 5840–5847.

Choe S (2002) Potassium channel structures. *Nature Reviews Neuroscience* 3: 115–121.

Ehret G and Romand R (eds.) (1997) *The Central Auditory System*. Oxford: Oxford University Press.

Forsythe ID and Barnes-Davies M (1997) Synaptic transmission: Well-placed modulators. *Current Biology* 7: R362–R365.

Geisler CD (1998) *From Sound to Synapse*. New York: Oxford University Press.

Hoffpauir BK, Grimes JL, Mathers PH, and Spirou GA (2006) Synaptogenesis of the calyx of Held: Rapid onset of function and one-to-one morphological innervation. *Journal of Neuroscience* 26(20): 5511–5523.

Kopp-Schneinpflug C, Dehmel S, Dörrscheidt GJ, et al. (2002) Interaction of excitation and inhibition in anteroventral cochlear nucleus neurons that receive large endbulb synaptic endings. *Journal of Neuroscience* 22: 11004–11018.

Lu Y, Monsivais P, Tempel BL, et al. (2004) Activity-dependent regulation of the potassium channel subunits Kv1.1 and Kv3.1. *Journal of Comparative Neurology* 470: 93–106.

Morest DK (1968) The growth of synaptic endings in the mammalian brain: A study of the calyces of the trapezoid body. *Zetschrift fur Anatomie und Entwicklungsgeschichte* 127: 201–220.

Niparko JK, Kirk KI, Mellon NK, et al. (eds.) (2000) *Cochlear Implants: Principles & Practices*. Philadelphia: Lippincott Williams & Wilkins.

Oleskevich S, Youssoufian M, and Walmsley B (2004) Presynaptic plasticity at two giant auditory synapses in normal and deaf mice. *Journal of Physiology* 560: 709–719.

Petralia RS, Wang Y-X, Zhao H-M, et al. (1996) Ionotropic and metabotropic glutamate receptors show unique postsynaptic, presynaptic, and glial localizations in the dorsal cochlear nucleus. *Journal of Comparative Neurology* 372: 356–383.

Rowland KC, Irby NK, and Spirou GA (2000) Specialized synapse-associated structures within the calyx of Held. *Journal of Neuroscience* 20: 9135–9144.

Ryugo DK and Parks TN (2003) Primary innervation of the avian and mammalian cochlear nucleus. *Brain Research Bulletin* 60: 435–456.

Ryugo DK, Rosenbaum BT, Kim PJ, Niparko JK, and Saada AA (1998) Single unit recordings in the auditory nerve of congenitally deaf white cats: Morphological correlates in the cochlea and cochlear nucleus. *Journal of Comparative Neurology* 397: 532–548.

Ryugo DK, Huchhton DM, Pongstapron T, and Niparko JK (1997) Ultrastructural analysis of primary endings in deaf white cats: Morphologic alterations in endbulls of Held. *Journal of Comparative Neurology* 385: 230–244.

Ryugo DK, Kretzmer EA, and Niparko JK (2005) Restoration of auditory nerve synapses in cats by cochlear implants. *Science* 310: 1490–1492.

Sento S and Ryugo DK (1989) Endbulbs of Held and spherical bushy cells in cats: Morphological correlates with physiological properties. *Journal of Comparative Neurology* 280: 553–562.

Spirou GA, Brownell WA, and Zidanic M (1990) Recordings from cat trapezoid body and HRP labeling of globular bushy cell axons. *Journal of Neurophysiology* 63(5): 1169–1190.

Trussell LO (1997) Cellular mechanisms for preservation of timing in central auditory pathways. *Current Opinions in Neurobiology* 7: 487–492.

VonGersdorff H and Borst JG (2002) Short-term plasticity at the calyx of Held. *Nature Reviews Neuroscience* 3: 53–64.

Auditory Systems in Insects

H C Hughes and S S Wang, Dartmouth College, Hanover, NH, USA

© 2009 Elsevier Ltd. All rights reserved.

Many people might not give much thought about whether insects can hear. After all, none of them has anything that look like ears. On the other hand, anyone who has heard the cascade of cricket songs that begin promptly at dusk might suppose that crickets can hear – it is unlikely all that singing is for our benefit alone. Careful research has demonstrated all of the insects that produce various forms of calls or songs hear those songs very well. These include crickets, grasshoppers, and bushcrickets. However, even more surprising is the fact that many insects that appear silent to us also hear. Among these silent listeners are various species of moths and flies.

Insects are the most diverse and most numerous animals on Earth. Entomologists have identified approximately 900 000 different species of insects, but they estimate that 80% of the species remain undiscovered. The actual number of insect species probably lies somewhere between 3 million and 30 million! There are probably 10 quintillion (10 000 000 000 000 000 000) individual insects alive at this moment – that comes to an astonishing 200 million insects for every living human on earth. Many of those insects can hear.

This article describes the variety of insects that can hear, the function of insect hearing, the anatomical organization of insect auditory systems, and the ways in which the insect nervous system processes acoustic information.

The Physical Properties of Sound Waves

A sound is the propagation of a mechanical disturbance through an elastic medium. For terrestrial animals, that medium is air. A vibrating object displaces air molecules, and those displacements are transmitted through the air as variations in air pressure. Very close to the vibrating object, the vibrations produce high-velocity movements of air molecules that are called near-field sounds. These near-field sounds are detected in some insects by very fine mechanoreceptors that are actually displaced in a manner analogous to the way shafts of wheat bend with the wind. These molecular movements become negligible once the sound has moved more than several wavelengths of distance from the source however. As a result, far-field sounds exist only as variations in air pressure. Far-field sounds are detected using a different kind of mechanoreceptor that is sensitive to variations in air pressure.

The velocity of sounds varies with atmospheric conditions, but under typical conditions sound waves travel at approximately 340 m s^{-1} (1225 km h^{-1} or 761 mph). Sounds have three basic properties that any hearing animal needs to be able to identify. The first property is the frequency, which is rate with which the pressure changes as the sound moves past the listener. The unit of frequency, called hertz (Hz), is the number of pressure cycles that occur in 1 s. The second important property of sound waves relates to the amplitude of the pressure variations. This is the intensity of the sound, and is related to the perceptual quality we call loudness. The third important feature of a sound is the location relative to the listener – whether the sound is to the left or the right, in front of or behind the listener. The small size of insects complicates the process of sound localization, which has led to the evolution of some very clever adaptations.

Diversity of Auditory Mechanisms Within the Class Insecta

There are 29 different orders of insects and a sense of hearing has been identified in seven: Neuroptera (green lacewings), Lepidoptera (moths and butterflies), Coleoptera (tiger beetles), Dictyoptera (praying mantis and American cockroach), Orthoptera (migratory locusts, field crickets, and katydids), Hemiptera (bees, wasps, ants, sawflies), and Diptera (flies).

Diversity of ear location A remarkable aspect of insect hearing is the fact that the ears can be found on virtually any part of the body. Like vertebrates, the insect body is based on a segmental plan, and each of these segments has stretch receptors and other kinds of mechanoreceptors that sense movements of various organs within each segment. It appears that insect ears evolved from a mutation in genes that control the development of these so-called chordotonal mechanoreceptors, and those mutations could occur in any of the body segments. Thus, insect ears evolved independently in many different species, and these independent mutations produced ears at many different locations in different species. Insect ears can be found on the wing base, abdomen, metathorax, base of forehead, hindwing, forewing base, cervical membranes, metathoracic leg, first abdominal segment, prothoracic leg, mesothorax, and ventral prosternum (see Figure 1). Crickets and katydids have eardrums

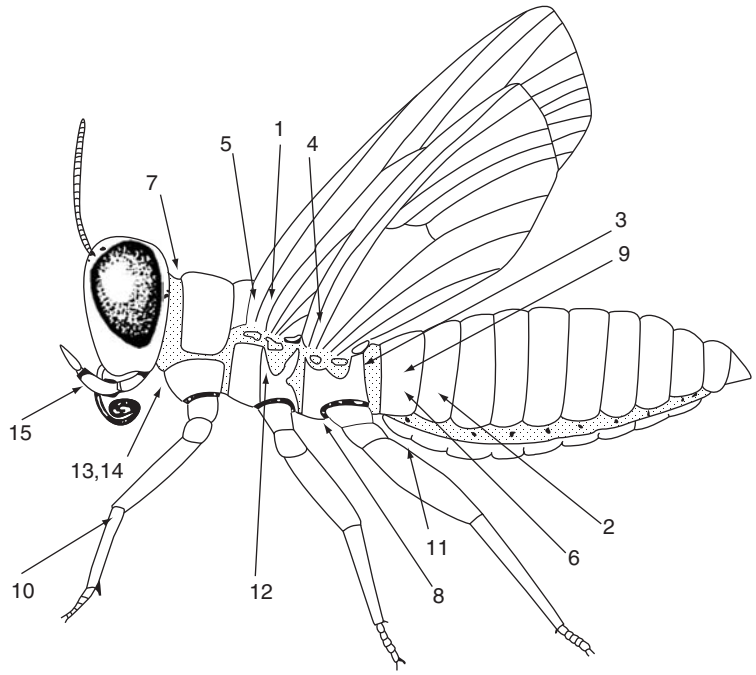

Figure 1 The variety of locations of tympanal organs as illustrated in a 'generic insect'. The locations indicated are: 1, wing base; 2, abdomen; 3, metathorax; 4, base of hind-wing; 5, fore-wing base; 6, abdomen; 7, cervical membranes; 8, ventral metathorax/metathoracic leg; 9, first abdominal segment; 10, prothoracic leg; 11, abdomen; 12, mesothorax; 13 and 14, ventral pro-sternum; 15, pilifer organ. Reprinted from Hoy RR and Robert D (1996) Tympanal hearing in insects. *Annual Review of Entomology* 41: 443–450 (figure 1, p. 438).

on their forelegs, grasshoppers and locusts have eardrums on their first abdominal segments, and cicadas have exposed eardrums on the abdomen.

Diversity of ears Although all insect auditory receptors are a variety of chordotonal receptor, there are two basic types of insect ears: tympanal and flagellar. Tympanal auditory organs are pressure detectors and are designed to detect far-field sounds. The basic principle of their operation is similar to the tympanic membrane (ear drum) of vertebrate ears. The variations in air pressure provide a force that is applied to the surface of the tympanum, which makes it move. Those movements are detected by the auditory receptors. Three morphological features characterize most tympanal organs. First, the tympanal membrane appears as a localized thinning of the external cuticle at the location of the hearing organ (see Figures 2 and 3). Second, an air-filled sac of tracheal origin is situated close to the internal surface of the tympanum. Third, the presence of a chordotonal organ composed of a sensory neuron with a dendrite embedded within a specialized support structure called the scolopidia.

Flagellar auditory organs function as particle velocity detectors for near-field sounds. They are known as Johnston's organs, and consist of a fine cilia that protrudes from the insect's body and bends in response to the motion of air molecules. In mosquitoes, the antennas are covered with hairs that serve as near-field sound detectors. Johnston's organs are generally used to detect the wing-beats of conspecifics or nearby predators.

The function of auditory receptor cells is to convert the acoustic mechanical energy into an electrochemical response in the nervous system. In tympanal organs, the sensory cell dendrite and its associated scolopidia are attached to the tympanal membrane, and undergo a mechanical disturbance when sound waves displace the tympanal membrane. The detailed biophysical processes that underlie transduction tympanal ears is not yet understood.

Purposes of Audition in Insects

Insects use audition to communicate with members of their own species, detect predators, and, in some cases, detect prey. Many insects, including grasshoppers, crickets, bushcrickets, katydids, fruit flies, bees, and cicadas, produce acoustic signals that serve to attract mates, proclaim territories, and exchange information such as the location of a food source. Some moths, lacewings, beetles, mantids, and flies can hear the ultrasonic sonar signals emitted by predatory bats. Detection of ultrasonic signals usually initiates evasive maneuvers in these species.

It is usually found that the auditory systems are specialized for processing species-specific communications. Thus, the auditory receptors often show their

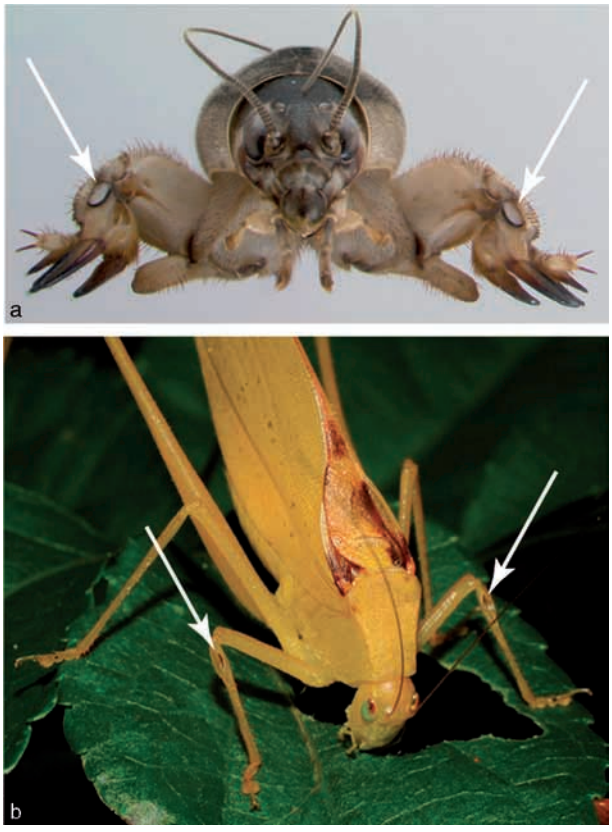

Figure 2 Tympanal organs on the forelegs of two species of insects. (a) Southern mole cricket *Scaptericsus borellii*. (b) Oblong-winged katydid *Amblycorypha oblongifolia*. (a) Photo reproduced with permission of TJ Walker, University of Florida. http://buzz.ifas.ufl.edu/341ppl.htm. (b) Photo reproduced with permission of TG Forest, University of North Carolina – Asheville. http://buzz.ifas.ufl.edu/007pmy2.htm.

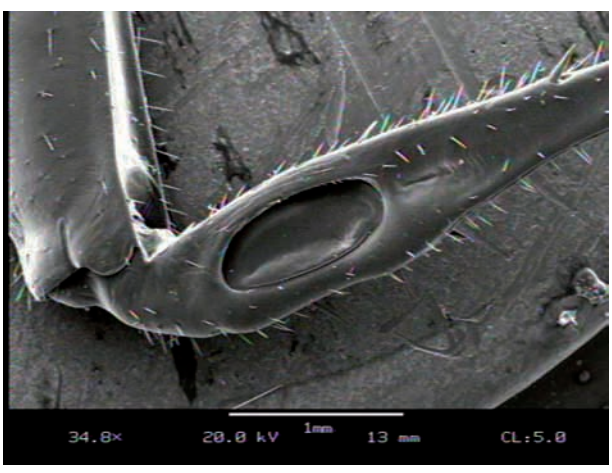

Figure 3 A scanning electron micrograph of tympanal organ of a katydid. Photo reproduced with permission of WE Conner, Lake Forest University. http://www.wfu.edu/biology/batsandbugs/bugs_ears.htm.

greatest sensitivity for the frequencies used in intraspecies auditory signaling. Other important aspects of insect auditory systems appear to maximize the range over which auditory communications will be effective. Cicadas and crickets produce mating calls that appear to be detected by females at distances between 10 and 20 m. Amazingly, one form of African grasshopper produces a mating call that can be heard by prospective mates at distances up to 1.9 km. An equivalent range in human auditory communication would allow us to converse with people that were over 40 km away.

In most species, the males produce the mating calls, usually by rubbing their wings or legs against each other – a technique called stridulation. Potentially receptive mates detect these calls with their tympanal organs. Most insect calls are comprised of rapid variations in the amplitude of what is termed a carrier frequency. Both the carrier frequency and the temporal patterning or rhythm of the bursts varies with different species. Thus, the receiving insect must encode both the carrier frequency and the rhythm, so that communication is confined to members of the same species. Examples of the mating calls of several cricket species are illustrated in Figure 4. The females respond to these calls by either approaching (positive phonotaxis) or turning away from them (negative phonotaxis). The calls of other species are either ignored or result in negative phonotaxis. Scientists use these phonotaxic responses to determine which aspects of a mating call the females find most attractive. The ability to orient toward the source of a mating call demonstrates that the insects can both recognize conspecific calls and localize their source. This requires specialized neural processes that are described below.

Insects that broadcast their presence by emitting these calls and songs take a certain risk: potential predators might hear them. In some instances, conspecific communication attracts not only mates but also deadly parasites. The female tachinid fly (*Ormia ochracea*) has sternal tympanal ears for acute directional hearing, which allows them to locate singing male crickets. Once a cricket is located, the female fly climbs onto the host and deposits small and undetectable larvae that enter the cricket's body and develop into large maggots that devour the cricket from the inside out (see Figure 5).

Audition in insects is also important for predator detection. Moths use their metathoracic tympanal organs to detect the sonar signals of bats that prey upon them. These organs are remarkably sensitive to ultrasonic sound waves, as they can detect a

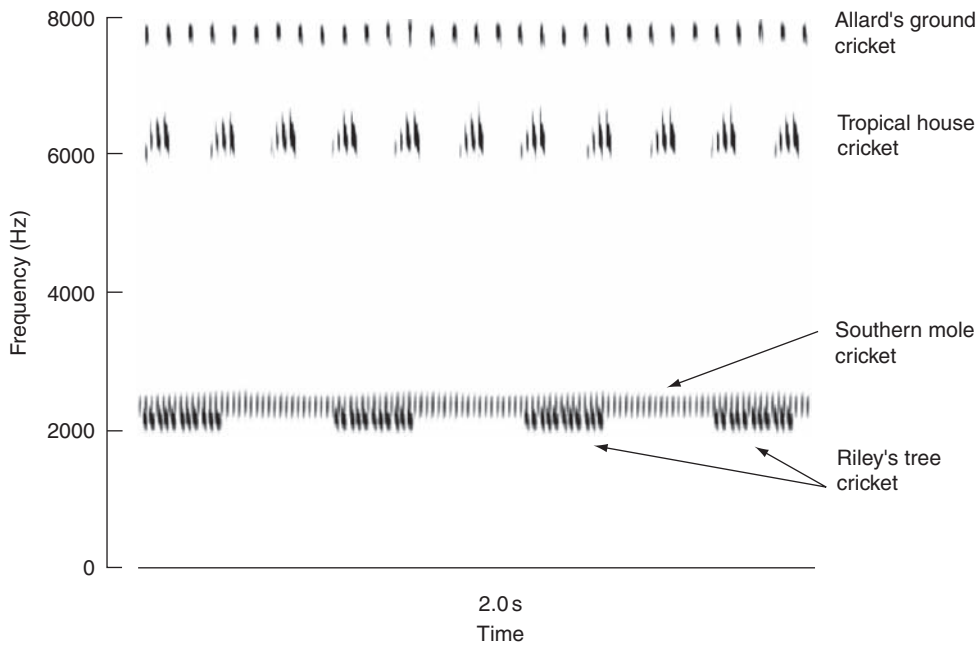

Figure 4 Sound spectrograms of four different species of crickets. Each song consists of brief pulses centered on a carrier frequency. The carrier frequencies and the temporal patterns differ for each of the four species although the carrier frequencies of Riley's tree cricket and Southern mole cricket are very similar. The ranges of Riley's tree cricket and the southern mole cricket do not overlap. The range of the tropical house cricket overlaps with both Riley's and Southern mole cricket. The spectrograms were obtained from http://buzz.ifas.ufl.edu and are reprinted here with permission of TJ Walker.

Figure 5 *Ormia ochracea* (on top) laying a larva on a host crcket. See text for further details. Obtained from http://hoylab.cornell.edu/.

movement of the tympanal membrane of as little as 0.000 000 000 001 m, which is substantially smaller than the diameter of a hydrogen atom. Detection of bat sonar signals initiates evasive countermaneuvers. Other nocturnal insects, including green lacewings, field crickets, praying mantises, locusts, katydids, and beetles display startle responses or fold up their wings and fall from the sky when exposed to simulated bat sonar signals.

Drosophilas (fruit flies) use flagellar organs to discriminate species-specific courtship songs within a close range. Similarly, bees use their Johnston's organs in the flagellar antennas to decode the acoustic signals of the waggle dance and song within the hive.

Central Processing of Auditory Information in Insects

Insects have a much more primitive central nervous system (CNS) than most vertebrate species. However, like the vertebrate nervous system, the insect CNS is an elaboration upon a basic segmental design. Each body segment has a corresponding ganglia (collection of nerve cells), and the ganglia are interconnected via nerves. Sometimes the ganglia fuse into larger groups. The head ganglia is the insect's brain.

As in all nervous systems, we can identify three main classes of neurons: sensory neurons (afferent neurons), interneurons, and motor neurons (efferent neurons). The sensory neurons convey information from the various sensory receptors to the CNS. The motor neurons have cell bodies in the ganglia, with axons that project to and innervate muscles. In some cases, sensory neurons make direct synaptic contact

with motor neurons. These direct sensory–motor connections mediate the simplest forms of reflexes. Frequently, however, the relationship between sensory inputs and motor responses are more complicated and requires more elaborate neural circuitry. In those cases, circuits of interneurons lie between the sensory afferents and motor efferents. In general, the auditory system of insects is composed of a relatively small number of sensory neurons and a larger set of interneurons. The neurons that actually innervate the auditory receptor are called primary auditory afferents: they are the neurons that actually generate neural impulses in response to a mechanical disturbance in the auditory receptor. In some species, each ear may be innervated by a single sensory neuron. Typical insect ears have 50–100 primary auditory afferents. In general, more primary afferents allow for more elaborate processing of the auditory signal. By way of comparison, the human auditory system contains about 30 000 primary afferents from each ear. The primary afferents make synaptic contact with interneurons. Interneurons that receive an input from primary afferents are often called second-order afferents. The patterns of synaptic connections between neurons are what enable the computations that are the hallmark of all nervous systems. In general, each higher-order set of interneurons continues a trend of increasing response specificity.

The human auditory system can be considered a general purpose processor of auditory information. In contrast, the insect auditory system appears specialized for extracting very specific information about the location of possible mates, predators, or prey. Yet the insect nervous system still has to extract information about the frequencies, loudness, temporal patterns, and locations of biologically relevant sounds.

Frequency Coding

Each species of animal hears a limited range of frequencies, although the range differs in different species. The range of human hearing goes from 20 to 20 000 Hz. In general, insect auditory systems respond to a higher and broader band of frequencies, but generally with less resolution than humans. Many insects can hear sounds from around 1000 Hz to frequencies as high as 40 000 Hz. In general, those that can hear very high frequencies (30 000 Hz or higher) evolved this ability to detect when they are in the sonar beam of predatory bats.

Studies of the phonotaxic responses of crickets and other insects have shown that they will orient toward a simulated mating call with a carrier frequency of say 5000 Hz, but turn away from an identical call if the carrier frequency is changed to 15 000 Hz. Thus, the cricket can discriminate the calls on the basis of a difference in frequency. The way auditory systems achieve frequency discrimination is by having different auditory fibers respond somewhat selectively to different frequencies. This is known as frequency tuning. Because different neurons are tuned to different frequencies, a change in frequency produces a corresponding change in the ensemble of active neurons – we can say that identity of responding neurons is the neural code for frequency.

Because insects have relatively few primary auditory afferents, their ability to discriminate frequencies is much more limited than humans. Insects that detect the ultrasonic emissions of bats usually possess only two primary afferents that are tuned to those high frequencies – one for each tympanal organ. These insects cannot perform any detailed analysis of the bat sonar signal; they simply function to detect those signals and make a rough determination of the bat's location. This is an example of a special purpose auditory system in which a minimal amount of information processing needs to be performed in order to produce a very specific motor response (avoidance).

Other insects have more than one primary afferent, so in principle they could discriminate different frequencies. As early as 1966, Michelsen discovered that locusts' auditory receptors are tuned to different sound frequencies. This gives these insects their ability to discriminate between sound frequencies. Since Michelsen's original discovery, frequency-sensitive auditory receptors have been found in other insects such as bushcrickets and crickets. It is not known how frequency selectivity is achieved in insect primary auditory afferents, but it appears that the mechanism is somewhat different that the one employed in vertebrate auditory systems. The mechanism in insects might be related to anatomical specializations of the support structures surrounding the receptor dendrites, or it may related to the biophysical properties of the receptor cell membrane or of the tympanal membrane itself. Most insects can discriminate among different frequency bands within their own hearing range, and the hearing range is generally determined by the evolutionary needs of the particular species.

For many insects, the neurons that are tuned to different frequencies are organized in a specific spatial arrangement – adjacent cells prefer nearby frequencies. This is called tonotopic organization. Tonotopic organization implies that the nervous system contains something analogous to a frequency map, so that the location of the activity within the map is part of a neural code for frequency. Take, for example, the locust ear: it has four receptor groups, each group is activated by distinctive sound frequencies and the groups are arranged in an orderly tonotopic map. Auditory

receptors in the bushcricket ears and cricket ears are also tonotopic. In crickets, a large clustering of neurons tuned to the carrier frequency of the calling-song frequency results in an overrepresentation of the carrier frequency within the array of primary afferents. This overrepresentation of particularly important frequencies is commonly found in many different sensory systems in many different species. Tonotopic organization aids in the formation of appropriate connections between neurons because neural pathways are often topographically organized (adjacent afferents make synaptic contact with adjacent neurons at the next synaptic station).

The principle of increasing complexity in higher-order sensory neurons is well illustrated in the insect auditory system. In crickets, two identified interneurons, designated AN1 and AN2, receive inputs from primary afferents and project to the brain. AN1 neurons are most sensitive to low frequencies, whereas AN2 neurons are sensitive to high frequencies and the species-typical calling song frequency. In field crickets, AN1 neurons are tuned to the frequency of the calling and aggression song, while AN2 neurons are tuned to the higher frequencies of the courtship song. In female katydids, interneurons TN-1 respond to both the frequencies of male songs and the frequencies of bat calls. However, TN-1 neurons also function as temporal pattern analyzers: they respond to the bat call pulse rate but not to male katydid pulse rate. This suggests that the TN-1 neurons function as bat detectors. The patterns of connectivity between primary afferents and interneurons appear to perform a primitive and automatic form of auditory stream segregation. This allows the insects to quickly identify and differentiate behaviorally important sound signals, even when frequency content alone is insufficient to make the distinction.

Loudness Coding

Studies of positive phonotaxis have shown that female crickets prefer louder mating calls to softer ones. Thus, they can discriminate the calls based on loudness. Loudness is usually represented by the frequency with which the responding neurons generate action potentials. As the intensity of the sound increases, the frequency of action potentials increases as well. This is known as a rate code for intensity.

Encoding Temporal Patterns

Most insect songs do not have patterned changes in frequency. Unlike vertebrate communication signals, insect songs are monotonous and lack melodies. The main information that distinguishes the individual song type from the others is in the rhythm and temporal structure of amplitude modulations of the carrier frequency. Thus, although an analysis of carrier frequency is performed, the insect nervous system must also be capable of recognizing species-specific temporal patterns of the song. The structure is defined by (1) duration and shapes of sound pulses, (2) the spacing between the sound pulses, and (3) the organization of sound pulses into higher-order groupings. A typical song consists of repetitions of a serious of brief sounds called pulses or syllables (see **Figure 4**). By subtly varying the temporal structures of synthetic calling songs and noting the phonotaxic responses they produced, previous studies were able to establish the aspects of temporal patterns most important for song recognition.

For many crickets, song models with the species-typical pulse period induce the strongest phonotaxic responses. In some cases, different parameters of structural features trade off against one another, suggesting a parallel process mechanism where the output response is the summation of all the feature information in the input. Similarly, grasshoppers use temporal structures to decode the meaning of songs.

Using synthetic song models, behavioral experiments established the properties of the temporal-pattern filters used to identify conspecific acoustic signals. This filtering is for the most part performed in the CNS. In crickets, neither the primary afferents nor the second-order interneurons demonstrate selectivity for temporal pattern. In the cricket, neurons that are selective for the temporal pattern of conspecific songs have so far only been found in the brain. The general trend is that higher-order interneurons tend to respond selectively to temporal patterns that produce positive phonotaxis in behavioral experiments. In contrast, the responses of lower-order neurons faithfully reproduce the temporal pattern of both effective and ineffective temporal patterns. It is not yet known how this specificity for temporal pattern in achieved. One interesting possibility that these higher-order interneurons receive two copies of the temporal pattern, but one is delayed relative to the other. If the delays are carefully calibrated, the two copies of temporal pattern will arrive at the interneuron in the same temporal phase. In this case, they can interact synergistically, but only for the correct species-specific calling pattern.

Sound Localization

Detecting the presence of a potential mate or a potential predator is of little value if the location of

the mate or the predator cannot be determined. Thus, sound localization is an essential function for any auditory system. Consider the process of mate finding in short-horned grasshoppers. The male sings, and, if he is lucky, a receptive female sings a response song. Depending on the direction of the response song, the male jumps either straight ahead, or to the left or right. The male continues walking for a short distance and sings again, listens for a reply, and jumps again. By repeating this sequence, the male eventually finds the female. Female field crickets use a similar strategy when they walk toward the source of a mating call while repeatedly determining its direction. Various measurements of the direction of orienting provide scientists with detailed information concerning the accuracy of sound localization in different insects. Many of them can distinguish sound sources that are about 20° apart. By contrast, the minimum sound separation that humans can discriminate is about 1°. However, the localization accuracy of certain insects is equivalent to the limits of human sensitivity.

Cues for Sound Localization

It has been known since the latter part of the nineteenth century that there are two cues available for detecting the direction of sound waves. The first is diffraction and the second is the time of arrival. Both cues require comparisons between two detectors (ears) in different locations. No matter where an animal's ears are located, they are almost always on opposite sides of the body. Diffraction refers to the bending of waves around an occluding object. Diffraction is heavily dependent on the size of the occluding object relative to the wavelength of sounds, and the small size of insect bodies complicates the problem of sound localization. Significant diffraction occurs when the distance between the ears is greater than one-tenth of the wavelength of the sound. In this case, the sound bends around the body, which produces changes in both the amplitude and the phase of the sound wave arriving at each ear. When the wavelengths of the sound are very small relative to the size of the head, less diffraction occurs, which means that the sounds do not bend around the body as readily and this creates a sound shadow: the ear that is farther from the source receives a less intense signal that the ear that is nearer to the source. The time of arrival cue simply results from the fact that the sound must travel farther to arrive at the more distant of the two ears, so it arrives at the distant ear later in time. Because their bodies are so small, the time of arrival cue is not useful to the insects – the small distances between their two ears means that the difference in arrival times would be less than a few microseconds, which is too small a difference for the nervous system to detect. The insect's small bodies mean that, unless the frequencies are exceedingly high (and have correspondingly short wavelengths), their bodies do not produce much of a sound shadow. In such cases, the sound waves diffract around the insect's body and arrive with very little attenuation of sound pressure on the other side. The extra distance does introduce a small change in phase however.

The small size of insects makes sound localization especially difficult for them to achieve. Insects (and some small vertebrates, including frogs and small birds) have evolved an ingenious solution to the problem. Instead of using a simple pressure-sensitive receiver like the human ear, the insects have evolved a pressure difference receiver. The key element of a pressure difference receiver is that the sound is applied to both sides of the tympanal membrane, rather than just the outside surface, as occurs in larger vertebrates. Of course, this means that there must be an air-filled cavity that extends from the outside of the animal's body all the way through to the inside of the tympanum. This is easier to achieve when the ear is on a foreleg, but when the ear is located on one of the body segments, the canal must extend completely through the segment (e.g., from the left side of the body to the inside of the right tympanum). This allows the sound to apply a force on both the outside and the inside of the membrane. When the sound waves reach both the outer- and inner-surface tympanal membrane, the difference in phase that is produced by the sound arrival time difference between the two surfaces provides information about directionality using the pressure difference receiver. Careful measurements of sound pressure at the external (front) and internal (back) surfaces of a grasshopper's eardrum demonstrated that the movement of the tympanum was well approximated by the sum of two vectors. One vector represents the pressure on the outside of the membrane and the other represents the pressure on the inside, after it passed through the air-filled cavity that runs from the opposite side of the body to inside of the tympanal membrane. Using this two-input model, and combining the pressure information obtained by the ipsilateral and the contralateral tympanum, the direction of the sound source can be determined.

The pressure difference receiver thus provides a difference in the response of the two tympanal membranes that varies with the location of the sound source. It is therefore a cue to location of the source. To extract the information inherent in this cue, a comparison between the two ears is required. That comparison is performed by interneurons that receive inputs from the primary auditory afferents arising from each ear.

Intracellular recordings confirm that some interneurons in the auditory pathway integrate excitatory input from the ipsilateral ear with the inhibitory input from the contralateral side. In some cases, there are mechanical linkages between the two ears that serve to amplify the interaural signals even before any nervous processing takes place. These specializations endow small insects with enough sound localization ability to do what needs to be done: find mates and avoid predators.

The auditory system of insects appears in many ways rudimentary when compared to many vertebrate species. However, it demonstrates a number of ingenious adaptations that have permitted the insects to condense a great deal of auditory processing capacity into a very small and efficient package. It is a marvel of biological engineering, and has helped insects become the most numerous animals on the planet.

See also: Auditory Cortex Structure and Circuitry; Auditory System: Central Pathways; Auditory Cortex: Models; Auditory Localization; Sound Localization: Neural Mechanisms.

Further Reading

Eberl DF (1999) Feeling the vibes: Chordotonal mechanisms in insect hearing. *Current Opinion in Neurobiology* 9(4): 389–393.

Green DM (1976) *Introduction to Hearing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Helversen D and Helversen O (1983) Species recognition and acoustic localization in acridid grasshoppers: A behavioral approach. In: Huber F and Markl H (eds.) *Neuroethology and Behavioral Physiology*, pp. 95–107. Berlin: Springer.

Hoy RR and Robert D (1996) Tympanal hearing in insects. *Annual Review of Entomology* 41: 443–450 (figure 1, p. 438).

Hoy RR (1998) Acute as a bug's ear: An informal discussion of hearing in insects. In: Hoy RR, Popper AN, and Fay RR (eds.) *Comparative Hearing: Insects*, pp. 1–17. New York: Springer.

Hughes HC (2001) *Sensory Exotica: A World Beyond Human Experience*. Cambridge, MA: MIT Press.

Jack JE (2004) The structure and function of chordotonal organs in insects. *Microscopy Research and Technique* 63(6): 315–337.

Michelsen A (1998) Biophysics of sound localization in insects. In: Hoy RR, Popper AN, and Fay RR (eds.) *Comparative Hearing: Insects*, pp. 18–62. New York: Springer.

Pollack GS (1998) Neural processing of acoustic signals. In: Hoy RR, Popper AN, and Fay RR (eds.) *Comparative Hearing: Insects*, pp. 139–196. New York: Springer.

Pollack GS (2000) Who, what, where? Recognition and localization of acoustic signals by insects. *Current Opinion in Neurobiology* 10(6): 763–767.

Rayleigh JWS (1877) *The Theory of Sound*, 2nd edn. London: Macmillan.

Robert D, Amoroso J, and Hoy RR (1992) The evolutionary convergence of hearing in a parasitoid fly and its cricket host. *Science* 258(5085): 1135–1137.

Romer H (1998) The sensory ecology of acoustic communication in insects. In: Hoy RR, Popper AN, and Fay RR (eds.) *Comparative Hearing: Insects*, pp. 63–97. New York: Springer.

Schul J and Sheridan RA (2006) Auditory stream segregation in an insect. *Neuroscience* 138: 1–4.

Young D and Hill KG (1997) Structure and function of the auditory system of the cicada, *Cystosoma Saundersil*. *Journal of Comparative Physiology A* 117: 23–45.

Autism

G Dawson and M Murias, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Autism Spectrum Disorder

It is now recognized that autism is part of a broad and heterogeneous group of disorders, referred to as autism spectrum disorders (ASD), that involve impairments in reciprocal social interaction and communication, and the presence of restricted, stereotyped, and repetitive interests and behaviors. Impairments in the domain of social relationships are considered a core feature of ASD and can be manifest as a lack of social and emotional reciprocity; atypical eye–eye-gaze, facial expression, and gestures; lack of interest in and/or difficulty relating to others, particularly peers; and a failure to share enjoyment and interests with others. Asperger syndrome, a subtype of ASD, is characterized by intact use of formal language skills, such as vocabulary and syntax, along with impairments in social use of language (e.g., conversational skills), social relationships, and idiosyncratic and consuming interests. Diagnosis of most forms of ASD is based on behavioral observation and developmental history. Although the diagnosis of ASD is quite stable from early childhood to adulthood, outcome is extremely varied. Whereas approximately 25% of individuals do not develop communicative speech and the majority function in the mentally retarded range, a substantial proportion of individuals have average to above average intelligence and are able to live independent and productive lives. IQ and language ability in childhood are the strongest predictors of long-term outcome.

A strong focus on current research is early recognition of autism. Screening questionnaires for toddlers exist that can be used by health professionals. Studies suggest that infants at risk for autism show very few behavioral symptoms at 6 months but that by 12 months, many core autism symptoms are apparent for the majority of infants. Evidence for this comes from both prospective and retrospective home videotape studies, which have shown that by 12 months, infants later diagnosed with autism exhibit poor eye contact and reduced joint attention, communicative babbling, and social reciprocity. Behavioral screening methods for detection of infants at risk for autism appropriate for use with 12-month-olds are currently being developed. Scientists are searching for biomarkers that can identify infants at risk for autism, but no reliable biomarkers have yet to be discovered. A subgroup, estimated to be approximately 25%, shows autistic regression, characterized by apparently normal development until the second or third year of life, followed by a loss of language and social skills and onset of autism symptoms, such as repetitive behaviors. Recent genetic linkage studies suggest that the susceptibility genes associated with autistic regression differ from those associated with early-onset autism.

Prevalence and Etiology

The prevalence of ASD has increased significantly in the past few decades. A large epidemiological study conducted by the Centers for Disease Control (CDC) found a prevalence rate of 34 per 10 000 among 3- to 10-year-old children in metropolitan Atlanta, Georgia (USA). However, this rate of 34 per 10 000 is likely to be an underestimate. Higher functioning individuals may have been missed and younger children may not have been identified. The rate reported for 5- to 8-year-olds in the CDC study – that of 41–45 per 10 000 – may be more accurate, and is similar to other surveys that report a prevalence rate of 60 per 10 000. Autism affects males at rates three to four times higher than females.

Genetic Influences

Evidence for genetic influences on autism is strong: 5–10% of autism cases are due to an identifiable medical disorder with a known inheritance pattern, such as fragile X syndrome, untreated phenylketonuria (PKU), tuberous sclerosis, and neurofibromatosis. Fragile X syndrome in particular accounts for about 8% of cases of autism.

Heritability estimates for idiopathic autism range from 91 to 93%, with concordance rates for autism of 60–95% for monozygotic (MZ) twins versus 3–7% concordance in siblings and dizygotic (DZ) twins. In contrast, rates in the general population are ~0.1%. When a broader phenotype is used, concordance increases for both MZ (88–91%) and DZ (9–30%) twins. The rapid decrease in concordance rates from identical twins to siblings, together with the differential risk rates for male versus female siblings, suggests interactions (epistasis) among several genes, estimated to be between 5 and 10 or more. Over 100 genes have been tested as candidates for autism susceptibility loci. Promising candidate genes include *Engrailed 2* (*EN2*) where a positive association was reported in two independent samples. *SLC6A4* has been tested

as a candidate gene in multiple studies that have yielded positive but not entirely consistent results. Several mutations have been described in the neuroligin 3 and neuroligin 4 genes, though these appear to be an extremely rare cause of autism. A recent study implicates reduced *MET* gene expression in autism susceptibility. A number of genome scans for linkage have been reported using autism families. While no single region has been unambiguously associated with autism, coincident signals for specific chromosomal regions have been identified in multiple studies. The 2q, 7q, and 17q regions are associated with the strongest signals. The challenges posed by mapping the risk alleles in autism have led some scientists to consider epigenetic mechanisms, including genomic imprinting and epimutations. Involvement of such mechanisms in ASD is supported by the central role of epigenetic regulatory mechanisms in the pathogenesis of Rett syndrome and fragile X syndrome, both single-gene disorders associated with ASD.

The complex behavioral expression of ASD also poses challenges to genetic research. ASD presentation is heterogeneous, involving at least three symptom domains (communication, social, and restrictive behaviors/flexibility), with evidence that the autism phenotype extends beyond classic autism to 'lesser variant' phenotypes. Parents and siblings of individuals with autism exhibit higher than normal rates of autism-related impairments, as do more extended relatives. Older unaffected siblings exhibit impairments in receptive and expressive language skills, as well as specific learning disabilities. A family history study of 195 parents and 137 siblings of autistic probands found that 12–20% of siblings demonstrated broader phenotype impairments, including language and communication deficits, social impairments, and/or learning disabilities, as compared to only 2–3% of siblings of individuals with Down syndrome. Relatives of autistic probands from multiplex families (with $\geq$2 affected children with autism) are at greater risk of having broader phenotype impairments than are relatives of probands from families with only one child with autism. A study of 149 families of autistic probands reported excess rates of broader phenotype deficits in the 3095 relatives studied, including second- and third-degree relatives. Like autism, broader phenotype characteristics are more common in males than in females.

The use of quantitative measures of autism in genetic studies may be more informative than discrete diagnoses. Most linkage analyses of autism have used qualitative diagnoses. There have been few genetic studies that have attempted to measure autism-related traits along a continuum. In one study of 201 autism multiplex families (694 individuals), two traits, social motivation and restricted activities/flexibility, showed the highest heritability as well as strong genetic correlation, suggesting that these traits are promising for gene mapping and share a genetic basis. Both clinical progress and basic science likely will profit from the identification of endophenotypes. Diagnostic categories of autism were created for clinical classification and are based on behavioral symptoms that show considerable heterogeneity across individuals. Such symptoms might not be the best phenotype for genetic studies, because of the complexity of the cascade of events leading from genes to these behavioral symptoms. Intermediary measures that assess neural pathways and function might be more useful. The use of electrophysiological endophenotypes, in addition to clinical diagnoses, has led to successful identification of genes in other complex psychiatric disorders, such as alcoholism. Furthermore, similar endophenotypic approaches are in widespread use in other areas, such as in genetic studies of cardiovascular disease and migraine. Previous studies have identified elevated rates of specific neurocognitive impairments and atypical electrophysiological responses to social stimuli (e.g., faces) in family members of autistic probands. These include impairments on tests of executive function, reading ability, pragmatic language ability, and face processing, and atypical event-related brain potentials to facial stimuli. Future linkage studies will increasingly focus on these more refined, quantitative endophenotypes.

Brain Development

Cerebral Enlargement

In the first two years of life, head circumference (HC) measurements accurately reflect brain volumes. At birth, HC measurements in children with autism are not atypically large; however, by adulthood, approximately 20% of individuals with autism exhibit larger than average HC. A period of exceptionally rapid head growth occurs during the first year of life in children who are later diagnosed with autism, as well as in many younger siblings of children with autism who are at risk for the disorder. Patterns of HC trajectory appear to be an early risk marker for the development of ASD symptoms. Increased rate of growth may index aberrant processes during early development, and may precede the onset of autism-specific symptoms. HC is easily measured in the course of routine check-ups and is potentially useful as an early biological risk marker for ASD.

Imaging studies have consistently found increases in overall brain volumes in children and adults with ASD. Results from a longitudinal imaging study of brain

growth demonstrate that the cerebral enlargement found in ASD children at 3–4 years old were no longer significant at 6–7 years of age, compared to age-matched developmentally delayed and typically developing children. However, within individuals, this growth curve can be quite variable. Alterations in early cellular developmental processes, such as failure of apoptosis or synaptic pruning, have been hypothesized to account for early cerebral enlargement, but recent longitudinal magnetic resonance spectroscopic imaging (MRI) findings of brain chemical changes in ASD children have failed to support such mechanisms.

Structural Findings

Abnormalities of the medial temporal lobe (MTL) structures, in particular the amygdala, have been strongly implicated in ASD and autism-related symptom expression. Recent reviews examining structural and functional imaging findings have highlighted the potential role of early amygdalar developmental deficits impacting social perception and social behavior, which are hallmark symptoms of ASD. Imaging studies have identified volumetric amygdalar abnormalities, which, similar to cerebral volume, vary as a function of age. An imaging study of 3- to 4-year-old males with ASD exhibited disproportionate right amygdalar enlargement, which was more accentuated in children with more severe symptoms. A subsequent cross-sectional study that assessed amygdalar volume across a broad age range also observed amygdalar enlargement among young, but not older, individuals with ASD.

A fairly consistent finding in quantitative and diffusion tensor MRI studies is reduced callosal volumes in ASD. In particular, several studies have found the genu, the anterior portion of the callosum containing axons that interconnect portions of the frontal lobes. This suggests changes in cortical connectivity may be associated with ASD.

Volumetric MRI evidence for hippocampal differences in ASD has been inconsistent; however, postmortem studies have identified abnormalities of neuron size and density within the hippocampus. Although there is not compelling evidence of direct hippocampal involvement in core symptoms of ASD, the hippocampus has been implicated in the degree of mental retardation associated with ASD, and strong links between mental retardation and increased seizure risk are present in ASD. A 3-D hippocampus surface topography study in 3- to 4-year-olds with ASD found that a characteristic pattern of hippocampal shape alterations was strongly associated with lower intellectual ability and poorer performance on neuropsychological tests specifically tapping the MTL, but not to tests of prefrontal function. The shape alteration found in these ASD children, who did not have seizures at the time they were imaged, was similar to a pattern found to characterize adults with MTL epilepsy.

The basal ganglia system is associated with a variety of functions that are altered in ASD, including motor control, cognition, and emotions. In particular, the presence of repetitive and stereotyped behavior in ASD suggests abnormalities in basal ganglia circuitry. Several MRI studies have demonstrated increased mean basal ganglia volume in ASD.

Cerebral water content can be characterized using quantitative MRI T2 relaxation times, and the technique has recently been used to investigate possible pathologic processes in ASD that are associated with cerebral white matter and cortical gray matter. One study compared T2 relaxation in 3- to 4-year-old children with ASD with age-matched children with typical development and developmental delay. Whole-brain cortical gray matter T2 was prolonged in the children with ASD, whereas whole-brain white matter T2 did not differ. These findings may reflect brain mechanisms involving neuroinflammation, which is typically accompanied by edema and would result in increased T2 relaxation. The presence of autoimmune markers, such as microglial activity, elevation in cytokines, and increases in glial fibrillary acidic protein, could provide a plausible mechanism for the prolongation of gray matter T2 observed in the ASD sample. These results support altered gray matter cytoarchitecture or pathology in ASD.

Both gray and white matter deviations are present in ASD, and volumetric MRI data suggest that the patterns of alteration are strongly age related. In 2- to 3-year-old children with ASD, MRI measurements indicate that cerebral gray and white matter volumes appear abnormally large. Enlargements of gray and white matter volumes in frontal and temporal lobes is apparent by 2 to 4 years of age, with white matter enlargements maximal in the frontal and parietal lobes. Gray and white matter differences are apparently less pronounced in older children. In ASD children aged 6 to 11, total white matter volumes have appeared enlarged, with fewer differences found in gray matter. Recent studies have observed white matter increases among 6- to 12-year-olds, with short-range white matter underlying frontal cortex showing the greatest enlargements, while deeper long-range fibers (callosal and anterior–posterior projections) did not differ between ASD and controls. Cerebellar developmental abnormalities appear to be more complex. In a study of 2- to 4-year-old ASD children, cerebellar gray matter volume did not appear enlarged. In the same children, however, cerebellar white matter volume was dramatically

greater than in controls. Among older ASD patients, reduction in cerebellar vermis has been a fairly consistent finding.

Neurochemical Imaging

Magnetic resonance spectroscopy (MRS) allows non-invasive, *in vivo* study of cellular neurochemistry. *N*-Acetyl aspartate (NAA), found only in the nervous system, is a sensitive marker for neuronal integrity as well as for homeostatic interactions between neurons and glia. Increases in NAA are thought to parallel myelination and to reflect neuronal maturation. In ASD, decreases in NAA have been reported, but the specific brain regions contributing to these reductions have been inconsistent. One MRS study of 3- to 4-year-olds with ASD revealed regional and global decreases in NAA, and lower levels of other chemicals, and these difference appeared largely specific to gray matter. Recent longitudinal findings suggest that this pattern of chemical alterations in ASD generally persists between 3–4 and 6–7 years of age.

Autopsy Findings

Sparse tissue availability has made postmortem studies difficult to perform; however, several studies have described anatomic abnormalities in ASD brains. Consistent findings have been observed in the limbic system, cerebellum, and related inferior olive. Decreased cell sizes and increased cell packing density have been shown in hippocampal, amygdala, and entorhinal cortex. Cerebellum findings include significantly reduced numbers of Purkinje cells, primarily in the posterior inferior regions of the hemispheres. In ASD, frontal white matter may be compromised by neuroinflammatory responses, with evidence for inflammation noted especially at the junction of the cortex and white matter. Pathologies among cortical minicolumns have been reported in ASD, in which minicolumns appear narrower and more numerous. Minicolumns appear to be packed closer together in ASD brains, but the total number of cells does not differ from that in controls.

Postmortem studies in ASD have also documented histopathological features reflecting reduced numbers of neurons in the amygdala and especially the lateral nucleus of the amygdala, particularly in adults. Taken together with volumetric MRI studies, the apparent lower number of neurons in the amygdala suggests that the amygdala appears to undergo an abnormal pattern of development that includes early enlargement and a decreased number of neurons in adulthood. This suggests that either fewer amygdalar neurons are generated during early development, or a normal or even excessive number of neurons form in early development, with subsequent degeneration.

Functional Brain Imaging Studies

Functional brain imaging studies in autism have focused on understanding the neurofunctional bases of the core social impairments and the well-documented impairments in complex, higher order, abstract thinking found in this disorder.

Social Brain Circuitry

Specific areas of the brain, including the occipital and temporal cortices (e.g., superior temporal sulcus, fusiform face area), the amygdala, the orbito-frontal cortex, and the anterior cingulated cortex, are critically involved in social processing, forming the 'social brain circuitry.' Each of these brain regions subserves different aspects of social processing, and each has been implicated in autism. Many of the early social impairments in autism, such as eye contact, joint attention, responses to emotional displays, and face recognition, involve the ability to attend to and process information from faces. Face processing may play a fundamental role in the dysfunction of the brain systems underlying the other impairments in social cognition in autism. The neural systems that mediate face processing come online very early in life. Behavioral studies have demonstrated face processing impairments in children, adolescents, and adults with autism, including deficits in face discrimination and face recognition. Functional magnetic resonance imaging (fMRI) studies have demonstrated abnormal activation of the fusiform face area during face processing in individuals with autism. Electrophysiological studies also document atypical responses to faces in children and adults with autism. Specifically, autism appears to be associated with an atypical N170 response to faces and to emotional facial expressions. Thus, face processing impairments may be one of the earliest indicators of abnormal brain development in autism. Several fMRI studies also have documented abnormalities in amygdala activation during facial expression processing tasks. Current models of abnormalities in social brain circuitry in autism posit impairments in the amygdala-fusiform system.

A recent area of investigation in autism involves brain regions related to motor imitation. Imitative deficits in individuals with autism are consistently observed and imitation deficits are one of the core impairments of autism. In typically developing children and adults, passive observation of actions performed by others activates some of the same cortical regions that are activated during execution of action, likely contributing to action recognition. Electroencephalogram (EEG) studies with typical adults and children show that the mu (8–13 Hz) rhythm is attenuated both when individuals execute and simply observe actions.

Recent studies with children and adults with autism found that attenuation of mu was present during the execution, but not during the observation, of movement. Mu wave attenuation has been interpreted as reflecting a neural system for action observation–execution matching, also referred to as a 'mirror neuron' system. fMRI studies have shown that observation, execution, and imitation of actions in humans generate activity in the premotor cortex and Broca's area (homolog for area F5 in monkeys). It has been hypothesized that the action observation–execution matching system serves as the neural substrate for imitation, understanding of others' intentions, development of empathy, theory of mind, and even the evolution of language, and thus may be important for understanding the neurofunctional basis of ASD.

Functional Connectivity

Functional MRI and EEG studies indicate that autism is associated with abnormal cortical connectivity, particularly underconnectivity of the frontal regions with other higher cortical regions (e.g., the parietal region during an executive function task, the left lateroposterior temporal area during a sentence comprehension task). Underconnectivity, as measured by fMRI, has been found during a wide range of tasks, including executive function tasks, sentence comprehension, visual motor performance, and resting states. Reduced corpus callosum volume has been found to be correlated with lower connectivity in individuals with ASD. A recent EEG study utilized measures of EEG coherence to assess functional connectivity in adults with ASD during a resting state. Functional connections can be estimated in EEG signals by assessing the consistency of phase relationships between them via coherence measurements. High coherence between two EEG signals reflects synchronized neuronal oscillations and suggests functional integration, while low coherence suggests independently active neural populations. This study documented reduced long-range EEG coherence in the alpha frequency range, and increased short-range EEG coherence in the theta range in adults with ASD. These alpha range findings suggest that the frontal lobe has selectively weak functional connections with the rest of the cortex, and may be consistent with metabolic studies showing reduced correlated blood oxygenation between frontal and other regions.

Summary and Future Directions

In autism, the overall cerebral volume appears to increase at an unusually rapid rate during early childhood, suggesting that the disorder must be caused by prenatal or early postnatal events. Clear evidence of functional abnormalities in several regions involved in the social brain has been documented. Such evidence includes abnormal activation of the fusiform gyrus, superior temporal sulcus, amygdala, and orbit-frontal regions. Current studies are utilizing event-related potential (ERP) paradigms to study the development of social and language brain circuitry in infants at risk for autism. Of particular interest is whether the abnormalities previously documented in older children with autism are related to a lack of critical input to early developing neural systems during the first few years of life. Ongoing studies of young children with ASD are examining the effects of early behavioral interventions, designed to increase social engagement and communication, on functional cortical organization and neural responses (e.g., ERPs to social and linguistic stimuli). Furthermore, recent pathology studies have provided new insights into underlying mechanisms that might help explain anatomical abnormalities, such as enlarged cerebral volume, and functional impairments, such as reduced functional connectivity. Evidence for postnatal processes in abnormal brain development in autism, such as postnatal abnormalities in head growth, suggest the possibility that autism could be amenable to treatment, both medical and behavioral, during early life, and that such very early treatment might prevent or significantly ameliorate the onset of autism symptoms. As autism susceptibility genes are identified, the hope is that these genes, in combination with other behavioral, electrophysiological, and MRI indices, might allow very early identification of autism, thus offering the potential of prevention or reduction of the full-blown syndrome. Meanwhile, behavioral interventions that are appropriate for very young children with autism are becoming increasingly sophisticated and effective, at least for a substantial subgroup of children with this complex and challenging disorder.

See also: Cognition: An Overview of Neuroimaging Techniques; Development of Behavior; Executive Function and Higher-Order Cognition: Neuroimaging; Fragile X Syndrome.

Further Reading

Baron-Cohen S and Belmonte MK (2005) Autism: A window onto the development of the social and the analytic brain. *Annual Review of Neuroscience* 28(1): 109.

Belmonte MK, Cook EH Jr., Anderson GM, et al. (2004) Autism as a disorder of neural information processing: Directions for research and targets for therapy. *Molecular Psychiatry* 9(7): 646–663.

Casanova MF (2006) Neuropathological and genetic findings in autism: The significance of a putative minicolumnopathy. *The Neuroscientist* 12(5): 435.

Courchesne E, Redcay E, and Kennedy DP (2004) The autistic brain: Birth through adulthood. *Current Opinion in Neurology* 17(4): 489–496.

Dawson G, Webb S, and McPartland J (2005) Understanding the nature of face processing impairment in autism: Insights from behavioral and electrophysiological studies. *Developmental Neuropsychology* 27(3): 403–424.

Dawson G, Webb S, Schellenberg G, et al. (2002) Defining the phenotype of autism: Genetic, brain, and behavioral perspectives. *Development and Psychopathology* 14: 581–611. (Special Issue: Multiple Levels of Analysis.)

Dawson G, Webb SJ, Wijsman E, et al. (2005) Neurocognitive and electrophysiological evidence of altered face processing in parents of children with autism: Implications for a model of abnormal development of social brain circuitry in autism. *Development and Psychopathology* 17: 679–697.

DiCicco-Bloom E, Lord C, Zwaigenbaum L, et al. (2006) The developmental neurobiology of autism spectrum disorder. *Journal of Neuroscience* 26(26): 6897–6906.

Friedman SD, Shaw DW, Artru AA, et al. (2003) Regional brain chemical alterations in young children with autism spectrum disorder. *Neurology* 60(1): 100.

Herbert MR (2005) Large brains in autism: The challenge of pervasive abnormality. *The Neuroscientist* 11: 417–440.

Just MA, Cherkassky VL, Keller TA, et al. (2004) Cortical activation and synchronization during sentence comprehension in high-functioning autism: Evidence of underconnectivity. *Brain* 127(8): 1811.

Pardo CA, Vargas DL, and Zimmerman AW (2005) Immunity, neuroglia, and neuroinflammation in autism. *International Review of Psychiatry* 17(6): 485–495.

Schultz RT (2005) Developmental deficits in social perception in autism: The role of the amygdala and fusiform face area. *International Journal of Developmental Neuroscience* 23(2–3): 125–141.

Schumann CM and Amaral DG (2006) Stereological analysis of the amygdala neuron number in autism. *The Journal of Neuroscience* 26: 7674–7679.

Sparks BF, Friedman SD, Shaw DW, et al. (2002) Brain structural abnormalities in young children with autism spectrum disorder. *Neurology* 59: 184–192.

Wassick TH, Brzustowics LM, Barlett CW, et al. (2004) The search for autism disease genes. *Mental Retardation Developmental Disabilities Review* 10(4): 272–283.

Autoimmune Autonomic Neuropathy

S Vernino, University of Texas Southwestern Medical Center, Dallas, TX, USA
P A Low, Mayo Clinic College of Medicine, Rochester, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Acquired autonomic neuropathy can involve dysfunction of peripheral autonomic nerves or ganglia and can take many forms. Chronic progressive autonomic neuropathy may occur in the context of a more diffuse peripheral neuropathy associated with medical conditions such as diabetes or amyloidosis. A chronic, slowly progressive, idiopathic presentation of peripheral autonomic failure without sensorimotor neuropathy has customarily been called pure autonomic failure (a degenerative neurological disorder first described by Bradbury and Eggleston), but this arbitrary classification probably encompasses disorders with several different pathophysiologies, including chronic autoimmune autonomic neuropathy (AAN). Many cases of subacute autonomic neuropathy, where the maximal autonomic deficit occurs within 3 months, appear to be caused by neurological autoimmunity. Subacute autonomic neuropathies can be further divided into dysautonomia associated with sensory and motor neuropathy (acute inflammatory neuropathies such as Guillain–Barré syndrome; GBS), autonomic neuropathy associated with malignancy (paraneoplastic AAN), and AAN.

Autoimmune Autonomic Neuropathy

Clinical Description

The presentation of AAN (formerly known as acute pandysautonomia, idiopathic subacute autonomic neuropathy, or pure autonomic variant of GBS) is often highly characteristic. The typical syndrome occurs in a previously healthy individual in whom autonomic failure develops over the course of a few days or weeks. The most common pattern is severe generalized autonomic failure. Sympathetic failure is manifested as severe orthostatic hypotension and anhidrosis, and parasympathetic failure as dry mouth, dry eyes, sexual dysfunction, constipation, urinary retention, impaired pupillary light response, and fixed heart rate. Gastrointestinal dysmotility is common (70% of patients) and manifests as anorexia, early satiety, postprandial abdominal pain and vomiting, constipation, or diarrhea. Orthostatic hypotension, anhidrosis, and gastrointestinal dysfunction each occur in more than 50% of patients.

We now recognize that some cases of AAN have a more insidious onset. In those cases, the autonomic syndrome may be difficult to distinguish from an autonomic-predominant form of multiple-system atrophy or from pure autonomic failure. Clinical features which can help differentiate these disorders are presented in Table 1. Even if the time course is unclear, the presence of prominent gastrointestinal dysmotility as well as impaired pupillary light reflexes suggests AAN.

Due to loss of postural autonomic reflexes in AAN, a common chief complaint is light-headedness and fainting when upright. Some patients are unable to stand for more than a few minutes, and the most severely affected may even faint when sitting up. In addition to gravitational pooling of blood in the dependent limbs, impaired autonomic innervation of the splanchnic-mesenteric vascular bed contributes significantly to failure to maintain blood pressure when standing. When these patients stand up (or are brought to a head-up posture on a tilt table), there is a progressive decline in blood pressure without a compensatory rise in heart rate. This pattern is indicative of diffuse autonomic failure. Due to failure of the cardiovagal control of heart rate, there may be a slight resting tachycardia and minimal change of heart rate during deep breathing, Valsalva maneuver, or standing.

Gastrointestinal symptoms are another common and disabling symptom in AAN. Patients may have severe gastroparesis, which leads to early satiety and nausea after eating. Some patients regurgitate undigested food many hours after eating. A feeding jejunostomy tube may be required to provide nutrition distal to the stomach. Lower bowel hypomotility is also frequent, typically causing severe constipation. Abdominal X-ray may suggest intestinal obstruction with dilated loops of small bowel or distended colon. In some cases, laparotomy and bowel resection have been performed. The gastrointestinal symptoms lead to abdominal pain and weight loss due to inability to maintain nutrition.

Heat intolerance due to loss of sweating can be potentially life-threatening for patients in warm, dry climates. Patients are at risk for hyperthermia, and so must avoid extreme heat and cool their skin with water. Urinary retention (neurogenic bladder) is common in patients with AAN. Many other autonomic symptoms occur but are rarely life-threatening. These should be recognized as features of autonomic failure and not ignored. Dry eyes and dry mouth due

Table 1 Autoimmune autonomic neuropathy (AAN), pure autonomic failure (PAF), and multiple system atrophy (MSA)

Parameter	*AAN*	*PAF*	*MSA*
Onset	Subacute or insidious	Insidious	Insidious
First symptom	Multiple	Orthostatism	Neurogenic bladder
Gastrointestinal symptoms	Common	Absent	Uncommon
Pupillary involvement	Common	Absent	Uncommon
Central nervous system involvement	Absent	Absent	Present
Somatic neuropathy	Mild/minimal[a]	Absent	Present in 15–20%
Pain	Often present	Absent	Absent
Autonomic findings	Widespread	Limited	Relatively widespread
Progression	Often monophasic	Slow	Inexorably progressive
Prognosis	Relatively good	Relatively good	Poor
Lesion	Postganglionic	Postganglionic	Preganglionic; central
Supine plasma			
Norepinephrine	Reduced	Markedly reduced	Normal
Electromyogram	Usually normal	Normal	Usually normal
Ganglionic AChR antibody	Positive (50%)	Negative[b]	Negative

[a]More common in paraneoplastic cases.
[b]Chronic AAN may be indistinguishable from PAF.

to loss of parasympathetic secretomotor function are occasionally misdiagnosed as sicca or Sjögren's syndrome. Patients with impaired pupillary light reflexes may complain of blurry vision and light sensitivity. Impotence is a common early autonomic symptom in men.

In the laboratory, objective autonomic abnormalities are usually obvious and include orthostatic hypotension, tonic pupils, and lack of heart rate variability. The spectrum and severity of autonomic failure, however, varies from patient to patient. Less common presentations are a selective cholinergic (parasympathetic) failure, selective adrenergic neuropathy, or isolated gastrointestinal dysmotility.

Some case reports have demonstrated a dramatic and complete spontaneous autonomic recovery. In larger case series, the typical course is monophasic worsening followed by stabilization or partial remission without recurrences. Many patients report spontaneous improvement, but recovery is typically incomplete. Only a third of patients experience major functional improvement of autonomic deficits.

It is apparent that a subset of patients present with a more insidious and progressive form of AAN rather than with the typical subacute monophasic presentation. Some of these patients are readily identified as AAN by their pattern of autonomic symptoms or by the presence of specific serum autoantibodies (especially ganglionic acetylcholine receptor (AChR) antibodies; see later). Seronegative patients with an insidious presentation may be difficult to distinguish from degenerative forms of dysautonomia. Features which help differentiate AAN from pure autonomic failure and multiple-system atrophy are shown in Table 1.

Associated Features

When subacute autonomic neuropathy occurs in isolation (AAN), the syndrome follows a viral prodrome in about 60% of cases, with a flulike illness or upper respiratory infection being the most frequent association. Specific preceding viral infections have been reported, with Epstein–Barr virus being the most common association. Other recognized antecedent events include minor surgical procedures or routine immunizations. The spinal fluid protein is often elevated. Motor and sensory nerve abnormalities are minimal or absent. Neuropathic symptoms, such as tingling in distal extremities, are reported by about 25% of patients, but these symptoms are not accompanied by objective signs or electrophysiologic evidence of sensory neuropathy.

Ganglionic AChR Antibodies

About 50% of patients with the typical clinical features of AAN have high titers of autoantibodies directed against the ganglionic AChR. This receptor is a pentameric transmembrane complex consisting of two AChR $\alpha 3$ subunits in combination with AChR β subunits. The $\alpha 3$-type ganglionic AChR mediates fast synaptic transmission in all peripheral autonomic ganglia and is homologous but genetically and immunologically distinct from the AChR at the neuromuscular junction (Figure 1). Serum ganglionic AChR antibody levels in AAN cases correlate with the severity of autonomic neuropathy clinically and with the severity on laboratory testing. Although the finding of high levels of ganglionic AChR antibody is specific for the diagnosis of AAN, a negative antibody test does not rule out the diagnosis.

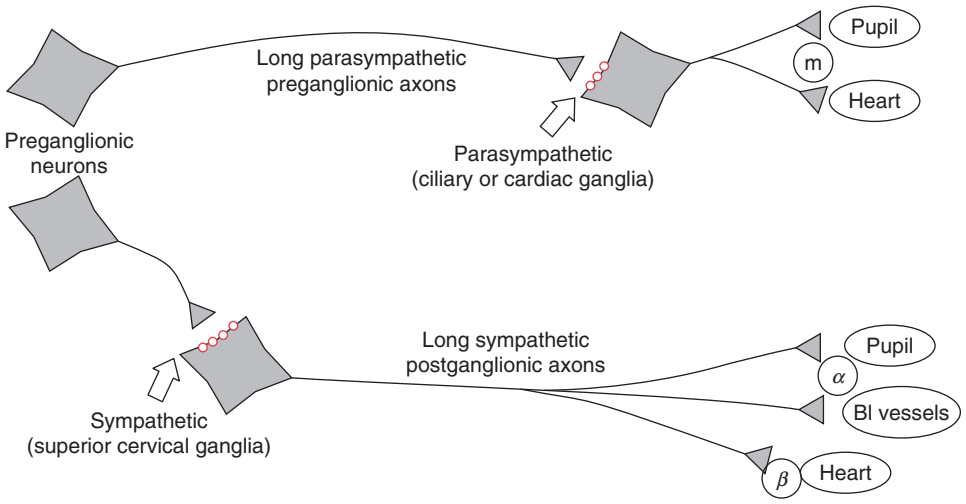

Figure 1 Autonomic system. The two major limbs of the peripheral autonomic nervous system both have a synapse in peripheral autonomic ganglia. Ganglionic synaptic transmission is mediated by acetylcholine, which binds to neuronal receptors on the postganglionic neuron (ganglionic AChR). Postganglionic autonomic neurons innervate multiple targets. Antibodies against ganglionic AChR could interfere with synaptic transmission in autonomic ganglia and cause diffuse autonomic failure (α, α3 subunits; β, β subunit). m, muscarinic; α, alpha-adrenergic receptors; β, beta-adrenergic receptors.

Clinically, patients with high levels of ganglionic AChR antibodies more often have a subacute onset and generally show more prominent cholinergic dysautonomia (sicca complex, pupillary abnormalities, gastrointestinal dysmotility, and bladder symptoms) compared to seronegative autonomic neuropathy patients. Recognition of ganglionic AChR antibodies has allowed for the serological detection of AAN and led to a better appreciation of the spectrum of this disorder, including the observation that some cases are characterized by insidious symptom onset and gradual progression without antecedent event. Such chronic cases may be initially indistinguishable from degenerative forms of autonomic failure. Some features that help distinguish AAN from degenerative forms of autonomic failure are shown in Table 1. A high serum level of ganglionic AChR antibody is most characteristic of idiopathic AAN. Low levels of antibodies are found in a minority of patients with limited forms of acquired dysautonomia. Ganglionic AChR antibodies are found in 5–10% of patients with postural tachycardia syndrome, idiopathic gastrointestinal dysmotility, or Lambert–Eaton myasthenic syndrome.

Treatment

The mainstay of treatment for AAN is symptomatic, including volume expansion for blood pressure support, bowel management, and supplemental moisture for dry eyes and mouth. Acetylcholinesterase inhibitors may be used to alleviate neurogenic orthostatic hypotension, and this class of drugs might be particularly appropriate to treat AAN where the pathophysiology is presumed to be impaired ganglionic cholinergic synaptic transmission.

Based on evidence of an autoimmune pathogenesis and the presence of a specific autoantibody, it is reasonable to consider plasma exchange or intravenous immune globulin. Single case reports of a successful intravenous immune globulin therapy have all involved early therapeutic intervention. It remains to be determined whether this treatment is effective for patients with a chronic progressive or chronic stable course. Plasma exchange has more recently been used to successfully treat AAN and was effective even in one patient with severe autonomic failure for over 10 years.

Animal Models of Experimental Autoimmune Autonomic Neuropathy

Several animal models of autoimmune or antibody-mediated autonomic neuropathy have been developed, and these have allowed insights into autonomic pathophysiology in humans. The most convincing way to prove an antibody-mediated disease is to transfer the disease to a healthy animal by administering antibody (passive transfer). AAN can be induced in mice by passive transfer of immunoglobulin G (IgG) from affected rabbits or humans. An animal model of experimental autoimmune autonomic neuropathy (EAAN) can also be induced in rabbits by immunization with the ganglionic AChR. Rabbits with EAAN manifest symptoms of chronic autonomic failure similar to those seen in AAN patients, including impaired blood pressure and heart rate variability, impaired pupil constriction, decreased lacrimation, and gastrointestinal hypomotility. As in patients with AAN, more

severe experimental autonomic dysfunction occurs with higher antibody levels.

Closer evaluation of the EAAN animal models provides some insight into the pathophysiology of AAN. Histologic studies of EAAN indicate that this autoimmune form of autonomic neuropathy is associated with a loss of ganglionic AChR from peripheral (ganglionic) autonomic neurons. Most likely, the binding of specific ganglionic AChR antibodies causes internalization and accelerated degradation of the synaptic receptors. This leads to failure of ganglionic synaptic transmission without significant loss of autonomic neurons. Since there is no permanent neuronal loss, EAAN and AAN may be reversible. Indeed, mice treated with ganglionic AChR antibodies demonstrate signs of transient autonomic failure and then recover fully.

Paraneoplastic AAN

Paraneoplastic neurological disorders are remote immunological effects of systemic malignancy that can affect any part of the nervous system. Sensory peripheral neuropathy, cerebellar ataxia, and limbic encephalitis are some well-recognized syndromes. A paraneoplastic autonomic disorder can take several different forms. As with other paraneoplastic disorders, the symptoms usually precede the diagnosis of cancer, and the tumors, when found, are limited in stage or only locally metastatic (regional lymph nodes). The autonomic symptoms, therefore, cannot be attributed to direct effects of the tumor, to nonspecific consequences of chronic illness, or to chemotherapy-induced neuropathy. In many cases, paraneoplastic dysautonomia is accompanied by symptoms of limbic encephalitis or elements of other paraneoplastic syndromes.

Paraneoplastic autonomic neuropathy can be a subacute diffuse autonomic failure indistinguishable from idiopathic AAN. More commonly, the autonomic syndrome specifically targets the enteric autonomic nerves and ganglia causing severe gastrointestinal dysmotility (gastroparesis or intestinal pseudoobstruction). Esophageal dysmotility (including achalasia) has also been reported. Patients present with nausea, early satiety, bloating, abdominal pain, constipation, and weight loss. Patients have difficulty getting adequate calorie intake and may even have difficulty maintaining fluid intake. This is complicated by the loss of appetite and taste that often accompanies cancer and chemotherapy. Radiographs show dilated loops of bowel, and motility studies reveal delayed gastric emptying, diffuse intestinal hypomotility, and absent or incoordinated motor complexes. Endoscopy and exploratory laparotomy fail to identify an obstruction. Pathologically, paraneoplastic dysmotility has been associated with inflammation and destruction in the myenteric plexus throughout the gut, with reduction in neurons and axons, and lymphocytoplasmic infiltration.

Several different autoantibody specificities may be encountered in patients with paraneoplastic autonomic neuropathy. The best recognized serological finding is the antineuronal nuclear antibody type 1 (ANNA-1, also known as anti-Hu). Small-cell lung carcinoma is found in at least 80% of patients who are seropositive for ANNA-1. Autoantibodies against collapsing-response mediator protein (CRMP-5) can be found with many neurological syndromes, including autonomic neuropathies. CRMP-5 antibodies are associated with small-cell lung carcinoma as well as thymoma. Antibodies specific for neuronal cation channels (ganglionic or muscle AChR, voltage-gated N-type or P/Q-type calcium channels, or neuronal potassium channels) are also commonly found. A minority of patients will lack any currently recognized autoantibody marker. Small-cell lung carcinoma is the tumor most commonly associated with paraneoplastic autonomic neuropathy. Dysautonomia, especially gastrointestinal dysmotility, may also be encountered with thymoma with or without associated symptoms of myasthenia gravis.

Treatment of paraneoplastic autonomic neuropathy generally consists of supportive symptomatic treatments to alleviate the most problematic symptoms, orthostatic hypotension and gastrointestinal dysmotility. Paraneoplastic dysmotility is usually refractory to treatment with motility-enhancing agents or surgical decompression. Many patients require jejunostomy feeding tubes or intravenous nutrition. Gastrostomy feeding is usually not tolerated because of gastroparesis and vomiting. Every effort should be made to locate and treat the underlying malignancy. In some cases, autonomic function improves once the malignancy is effectively treated. Plasma exchange or intravenous immunoglobulin has also been reported to be effective in individual cases.

Other Autoimmune Disorders Affecting Autonomic Function

Autonomic symptoms are commonly encountered in other recognized autoimmune disorders. Dry mouth, impotence, and constipation are common in patients with Lambert–Eaton syndrome, a presynaptic disorder of neuromuscular junction transmission caused by antibodies against the P/Q-type voltage-gated calcium channel. Autonomic instability (manifested as fluctuations in heart rate and blood pressure) and

gastrointestinal dysmotility are often encountered in patients with acute inflammatory demyelinating polyradiculoneuropathy (GBS syndrome). Autonomic overactivity is a typical feature of autoimmune neuromyotonia and Morvan syndrome. Both disorders are associated with antibodies against voltage-gated potassium channels and present with muscle stiffness and spontaneous muscle twitching. Future studies are likely to uncover other ion channels and other proteins in the autonomic nervous system that are targets for autoimmunity.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Central Control of the Gastrointestinal Tract; Autonomic Nervous System; Blood Pressure: Baroreceptors; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Buckley C and Vincent A (2005) Autoimmune channelopathies. *Nature Clinical Practice Neurology* 1: 22–23.

Etienne M and Weimer LH (2006) Immune-mediated autonomic neuropathies. *Current Neurology & Neuroscience Reports* 6: 57–64.

Klein CM, Vernino S, Lennon VA, et al. (2003) The spectrum of autoimmune autonomic neuropathies. *Annals of Neurology* 53: 752–758.

Lee HR, Lennon VA, Camilleri M, et al. (2001) Paraneoplastic gastrointestinal motor dysfunction: Clinical and laboratory characteristics. *American Journal of Gastroenterology* 96: 373–379.

Lennon VA, Ermilov LG, Szurszewski JH, et al. (2003) Immunization with neuronal nicotinic acetylcholine receptor induces neurological autoimmune disease. *Journal of Clinical Investigation* 111: 907–913.

Mathias CJ and Bannister R (eds.) (2006) *Autonomic Failure*. Oxford: Oxford University Press.

Robertson D and Low PA (eds.) (2004) *Primer of the Autonomic Nervous System*. Amsterdam: Elsevier Academic Press.

Schroeder C, Vernino S, Birkenfeld AL, et al. (2005) Plasma exchange for primary autoimmune autonomic failure. *New England Journal of Medicine* 353: 1585–1590.

Vernino S, Ermilov LG, Sha L, et al. (2004) Passive transfer of autoimmune autonomic neuropathy to mice. *Journal of Neuroscience* 24: 7037–7042.

Vernino S, Low PA, Fealey RD, et al. (2000) Autoantibodies to ganglionic acetylcholine receptors in autoimmune autonomic neuropathies. *New England Journal of Medicine* 343: 847–855.

Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development

R O Heuckeroth, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Published by Elsevier Ltd.

Introduction

Programmed cell death (PCD) is an evolutionarily conserved process in multicellular organisms that is important for morphogenesis during development and for the maintenance of tissue homeostasis in organs with ongoing cell proliferation. PCD also occurs in many tissues in response to injury. Apoptotic cell death is one form of PCD that provides a mechanism to remove individual cells from tissues without damage to adjacent cells. Defects in molecular mechanisms that regulate cell death cause both cancer and abnormal development. Because many aspects of apoptosis have been reviewed over the past decade, this article focuses on the developmental role of apoptosis in the sympathetic, parasympathetic, and enteric nervous system and provides only a brief overview of apoptotic mechanisms.

Apoptotic Cells and Molecular Mechanisms

Apoptosis is a form of cell death characterized by cell shrinkage, plasma membrane blebs, nuclear chromatin condensation, and DNA fragmentation. Phosphatidylserine, usually restricted to the inner face of the plasma membrane in live cells, also partially redistributes to the extracellular face of the plasma membrane. These changes largely result from activation of a set of cysteine-aspartyl-specific proteases called caspases. Caspases can be activated via either an extrinsic pathway or an intrinsic pathway. The extrinsic pathway is activated by binding of extracellular ligands to cell surface death receptors like Fas (Apo-1 or CD95), tumor necrosis factor receptor-1 (TNFR-1/p55/CD120a), interferon receptor, and TRAIL (TNF-related apoptosis-inducing ligand or Apo2-L) receptor. This binding leads to the recruitment of a variety of proteins that activate procaspase-8 or procaspase-10. In contrast, the intrinsic pathway is characterized by cytochrome c release from mitochondria. Cytochrome c in the cytoplasm induces heptamerization of Apaf-1, that then binds to and activates procaspase-9. These initiator caspases (-8, -9, and -10) activate the effector caspases (caspase-3 and caspase-7). The effector caspases have many substrates (estimated to include 0.5–5% of cellular proteins), induce DNA cleavage and mitochondrial permeabilization, and activate additional proteolytic cascades that eventually result in cell death.

Apoptosis in the Nervous System and the Discovery of Neurotrophic Factors

PCD is common in most parts of the developing nervous system, where 20–80% of all neurons produced during embryogenesis die before adulthood. Typically, these cells are eliminated by PCD at the time that they would normally innervate their targets. In this setting, PCD is thought to occur to allow matching of the neuronal population to the size of the innervation target via target-derived, trophic factor-dependent cell survival. This mechanism also efficiently eliminates cells with abnormal migration or abnormal axon targeting because cells that fail to innervate appropriate targets do not receive adequate trophic factor. The observation that developing sympathetic and sensory neurons were dependent on exogenous trophic factor for survival led to the discovery of nerve growth factor (NGF) by Rita Levi-Montalcini and Stanley Cohen. This discovery prompted work that in turn resulted in the discovery of the related proteins brain-derived neurotrophic factor (BDNF), neurotrophin-3 (NT-3), and NT-4/5. An *in vitro* sympathetic neuron survival assay was also critical for discovery of neurturin (NRTN) and for the identification and characterization of the related proteins glial cell line-derived neurotrophic factor (GDNF), artemin (ARTN), and persephin (PSPN). Furthermore, molecular mechanisms of neonatal sympathetic neuron survival *in vitro* have been studied extensively and form the basis of much of what we know about neuronal apoptosis. In the rat, for example, superior cervical ganglion (SCG) sympathetic neurons are NGF dependent from embryonic day (E)16 to postnatal day (P) 7. NGF deprivation during this time results in PCD within 48 h. With this assay, investigators demonstrated that apoptosis is triggered after NGF withdrawal by increased c-Jun protein levels and c-Jun phosphorylation, increased translocation of the proapoptotic protein Bax from the cytosol to mitochondria, and Bax-induced cytochrome c release that causes apoptosis via the intrinsic pathway. Release of Smac/DIABLO and HtrA2/Omi from mitochondria is also important to relieve caspase inhibition by inhibitor of apoptosis proteins so that caspases can become activated.

Lessons from Mutant Mice

The neurotrophic factors NGF, BDNF, NT-3, GDNF, NRTN, and ARTN are critical for the normal development and maintenance of the peripheral nervous system. Depending on the specific cell type and developmental time point, these factors determine nervous system structure and function via a number of distinct effects including preventing or in some cases promoting cell apoptosis. These proteins have other important roles as well:

- Promotion of cell survival
- Promotion of cell proliferation
- Support for neuronal differentiation
- Promotion of directed cell migration
- Promotion of neurite extension
- Axon guidance

Each of these actions is mediated by binding of trophic factors to specific cell surface receptors. There is also some ability for specific trophic factors to activate multiple receptors at least *in vitro*. Preferred receptor–ligand interactions, alternate receptor–ligand interactions, and a summary of interactions that are important *in vivo* are presented in Figure 1. Briefly, NGF, BDNF, and NT-3 bind to and activate the transmembrane tropomyosin-related kinase (Trk) receptors TrkA, TrkB, and TrkC, respectively. In addition, NT-4 activates TrkB, whereas NT-3 can directly activate TrkA and TrkB, but with lower efficiency than it activates TrkC. All these NTs also bind to the low-affinity receptor p75NTR. The interaction with p75NTR inhibits the activation of Trk receptors by their nonpreferred ligands and in some cases improves intracellular signaling. However, in the absence of specific Trk receptor activation, binding of proneurotrophins to p75NTR promotes cell death. For example, pro-BDNF binding to p75NTR on sympathetic neurons, which express TrkA but not TrkB, promotes cell death, but NGF promotes cell survival by activating both TrkA and p75NTR. This dual role for p75NTR in sympathetic neurons is thought to improve the specificity of axon targeting by eliminating cells whose axons innervate the wrong target and arrive at a source of BDNF but not of NGF. The proteins GDNF, NRTN, ARTN, and PSPN form a separate family of trophic factors called the GDNF-related ligands (GFLs). GFLs all activate the transmembrane tyrosine kinase Ret. Instead of binding directly to Ret, however, GFLs activate Ret by binding to a glycosylphosphatidylinositol-linked cell surface receptor (GFRα1–4). Each GFL preferentially binds to a specific GFRα protein, with GDNF, NRTN, ARTN, and PSPN interacting best with GFRα1, GFRα2, GFRα3, and GFRα4, respectively. There is some *in vitro* cross-talk between GFLs and nonpreferred GFRα proteins, as indicated in Figure 1, but these interactions generally require higher trophic factor concentrations and do not appear to be physiologically important *in vivo*. However, receptor

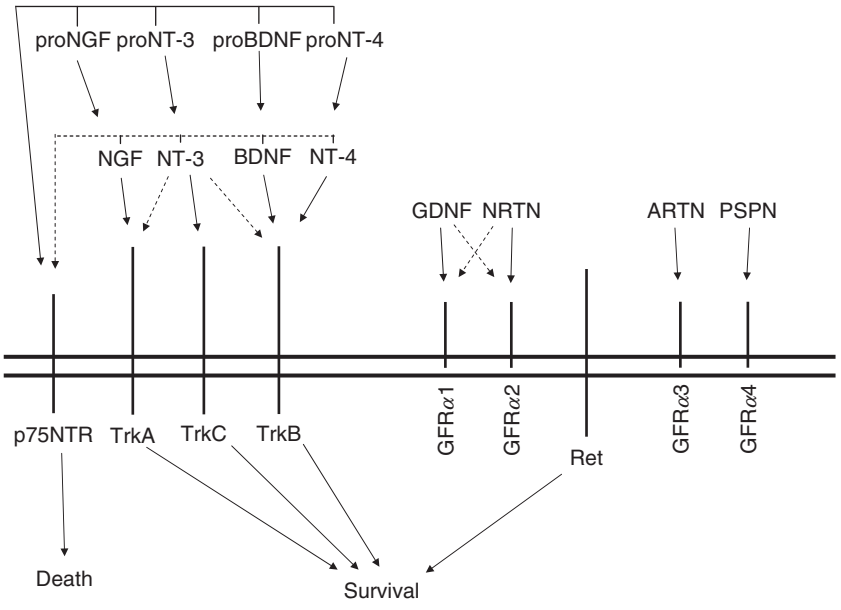

Figure 1 Summary of neurotrophic factors and their receptors. Solid arrows show preferred receptor–ligand interactions. Dashed arrows show nonpreferred interactions between ligands and receptors. For neurturin (NRTN) and glial cell line-derived neurotrophic factor (GDNF), the nonpreferred interactions do not appear to be important *in vivo*. For neurturin (NT)-3, nonpreferred interactions are important *in vivo*. Interactions of neurotrophins with p75NTR are also likely to be physiologically relevant. ARTN, artemin; BDNF, brain-derived neurotrophic factor; GFR, growth factor receptor; NGF, nerve growth factor; PSPN, persephin.

cross-talk appears important in the pharmacological response to exogenously administered factors, presumably acting at supraphysiological levels.

Given the complexity of these systems, determining the relevance of specific trophic factors and their receptors for neuronal development and function has required a combination of *in vitro* studies and the analysis of mutant animals. While many conclusions can be drawn from this work, several important themes have emerged:

1. Patterns of trophic factor receptor and ligand expression correlate well with trophic factor dependence *in vivo* but incompletely predict function.
2. For many developing neurons, trophic factor dependence changes during development such that cells initially dependent on one factor become dependent on a different trophic factor at a later developmental stage.
3. Trophic factors may promote both survival and proliferation early in development or may direct precursor migration. At later stages, these factors may support survival (e.g., at the time of target innervation) or provide trophic support without being essential for survival.
4. Trophic factor dependence *in vitro* does not imply developmental stage-specific programmed cell death *in vivo*. Cell death *in vivo* requires both trophic factor dependence and trophic factor deficiency.
5. Trophic factor receptor expression also does not imply trophic factor-dependent survival since specific cell populations may respond to multiple trophic factors simultaneously.

A few examples will clarify the strategies employed to determine cell number in the autonomic and enteric nervous system (ENS). Effects of specific receptor and ligand mutations on sympathetic, parasympathetic, and enteric neurons are summarized in Table 1.

Trk Receptors and Ligands Regulate Cell Death *In Vivo* in the Sympathetic Nervous System

The sympathetic chain in the mouse embryo arises at E11.5, and at that time, TrkC is expressed, but TrkA is not detectable. By E13.5, when the SCG forms, both TrkC and TrkA are detected in the SCG. TrkA expression becomes more robust by E15.5 and continues to be expressed at high levels throughout postnatal development. TrkC levels in the SCG fall significantly by P0, when only a few cells have detectable TrkC expression. In contrast, TrkB is not detected in sympathetic neurons at any developmental stage.

The pattern of TrkA receptor gene expression correlates well with the normal role of TrkA and NGF in development of the mouse sympathetic nervous system. In TrkA−/− mice, for example, neuron number is normal in the sympathetic chain at E11.5 and in the SCG and sympathetic chain at E13.5. By E15.5, there are 15% fewer SCG neurons in TrkA-deficient mice than wild-type (WT). By E17.5, TrkA−/− mice had 35% fewer SCG neurons than WT mice did, and almost all SCG neurons are lost in TrkA−/− mice by P9. At each of the periods investigated, the reduction in SCG neuron number in TrkA−/− mice was attributable to increased cell death. Similarly, in NGF-deficient mice, all SCG neurons are absent by P14, and increased cell death was observed at P3. These changes in neuron number demonstrate the critical role for NGF/TrkA signaling in sympathetic neuron survival and are consistent with results obtained in the 1960s with exposure to neutralizing antibodies to NGF. They also define a period of trophic factor-dependent cell death that correlates well with the timing of trophic factor dependence of SCG neurons in primary culture. Moreover, the striking similarity of the SCG phenotype for NGF−/− and TrkA−/− mice highlights the importance of this preferred ligand–receptor interaction.

In contrast to these results with TrkA−/− and NGF−/−mice, however, the effect of TrkC or NT-3 deficiency is less easily predicted based on gene expression patterns. For example, although TrkC is expressed in the sympathetic nervous system as early as E11.5, the SCG is normal at all developmental stages in TrkC−/− mice. This contrasts with the 50% reduction in SCG neurons found in NT-3-deficient mice. The time during development when NT-3 is required for sympathetic neuron survival, however, has been controversial, with an early report suggesting increased SCG precursor cell death between E11 and E17 and a later study using more animals failing to reproduce this observation and demonstrating increased programmed cell death after birth. Furthermore, injection of anti-NT-3 antibodies into neonatal rats caused a 60–80% loss of SCG neurons, suggesting a role for NT-3 in postnatal SCG survival. Together these observations demonstrate the complexity of determining how specific trophic factors regulate neuronal apoptosis, especially in cells with multiple trophic factor receptors. They further suggest that although NT-3 critically regulates SCG apoptosis, it may perform this function primarily via TrkA receptor activation.

The role TrkB and BDNF in SCG neuron survival is even more remarkable. Since TrkB is not expressed in the sympathetic nervous system, it might initially have been predicted that the SCG would be normal in

Table 1 Neuron number in autonomic and enteric ganglia of mutant mice

	Sympathetic		*Parasympathetic*				
	SCG	*Chain ganglia*	*Ciliary*	*Submandibutar*	*Sphenpalatine*	*Otic*	*Enteric*
TrkA KO	100% loss by P9						
NGF KO	100% loss by P7						
TrkB KO	Normal						
BDNF KO	36% increase						
TrkC KO	Normal						Selective neuron loss
NT-3 KO	50% loss						Selective neuron loss
p75NTR KO	Increased						
Ret KO	Abnormal, but variable	Abnormal	48% loss	30–47% loss	99–100% loss	99–100% loss	99% loss
GFRα1 KO	Normal			33% loss	99–100% loss	99–100% loss	99% loss
GDNF KO	30% loss		40% loss	36% loss	99% loss	86% loss	99% loss
GFRα2 KO	Normal			42% loss	Normal	40% loss	
NRTN KO	Normal		50% loss	45% loss	Normal	Normal	Normal
GFRα3 KO	Abnormal, but variable	Abnormal					
ARTN KO	Abnormal, but variable	Abnormal					Normal

BDNF−/− mice. Remarkably, SCG neuron number is increased by 36% in BDNF−/−compared to WT animals at P15. This was hypothesized to occur because pro-BDNF promotes sympathetic neuron cell death by binding to p75NTR. Indeed, WT SCG neuron number decreases by 42% because of PCD between P0 and P23, but SCG neuron number increases between P0 and P23 in P75NTR−/− mice. These results suggest that both BDNF and p75NTR are required for naturally occurring cell death in the SCG of WT mice. Further support for this hypothesis is the observation that mice deficient in both p75NTR and TrkA have markedly reduced sympathetic neuron apoptosis compared to TrkA−/− animals. This mechanism allows for efficient elimination of cells with misguided axons that encounter pro-BDNF and reinforces the importance of TrkA activation for SCG survival.

Ret Signaling Promotes Sympathetic Neuron Precursor Migration and Supports Axon Extension Required for Sympathetic Neurons to Innervate Targets and Encounter Trk Ligands

Ret activation via GFRα3 and ARTN is essential for normal sympathetic nervous system development. In this case, however, Ret signaling appears to be important for neuronal precursor migration and axon extension, but not directly for cell survival. This conclusion is based on the following observations. Ret−/− mice have smaller-than-normal SCG ganglia that appear in a more caudal position than in WT animals. Furthermore, sympathetic fiber innervation of targets in the nasal mucosa, eye, and skin of Ret−/− mice is almost completely absent, while sympathetic innervation density of the submandibular salivary gland is significantly reduced. In addition, there are defects all along the sympathetic chain in Ret−/− mice, including smaller-than-normal ganglia, abnormally located ganglia, and reduced target innervation. Detailed evaluation of the mechanism of these defects demonstrated that Ret−/− sympathetic neuron precursor lineage commitment appears normal, but differentiation of these cells is delayed, and neuronal cell death increases between E16.5 and P0 compared with WT mice. This increased cell death was particularly interesting since Ret is expressed in all sympathetic precursors at E11.5, but most of these cells lose Ret expression by E15.5. Furthermore, although increased cell death occurred in Ret−/− SCG, it did not occur preferentially in Ret-expressing cells.

Additional analysis demonstrated that the Ret ligand ARTN is expressed in blood vessels along the normal route of sympathetic axons where ARTN potently stimulates SCG axon growth and is a chemoattractant that directs axon pathfinding. It is interesting that ARTN- and GFRα3-deficient mice have variable defects in the SCG. Specifically, SCG size was near normal in ARTN−/− and GFRα3−/− mice if the ganglia were normal and in position and peripheral targets were innervated, but SCG size was markedly diminished and apoptosis was increased in mice with abnormally located SCG and reduced peripheral target innervation. Defects in ARTN−/− and GFRα3−/− mice were detected as early as E10.5, when sympathetic neuron precursors emerge from the neural crest, demonstrating an important role for Ret/ARTN/GFRα3 signaling in precursor migration and neurite extension. Overall, these data suggest that increased cell death in the sympathetic nervous system of Ret−/−, ARTN−/−, and GFRα3−/− mice, compared with WT mice, occurs because of failure to properly innervate targets and subsequent deficiency in target-derived trophic factor (i.e., NGF deficiency). The Ret ligand GDNF also appears to be important for SCG development; since GDNF−/− mice have been reported to have a 30% loss of SCG neurons. Although GDNF supports the survival of cultured SCG neurons, the precise role of GDNF for SCG neuron survival *in vivo* has not been defined. Furthermore, mice missing the preferred GDNF receptor GFRα1 have a normal SCG at P0, which suggests that GDNF effects on the SCG are mediated via an alternate signaling pathway. These analyses demonstrate the complexity of interpreting mouse phenotypes and primary culture data and the importance of careful mechanistic time-course studies to define the role of specific trophic factors and their receptors in neuron survival. They also demonstrate robustly that programmed cell death in response to trophic factor deficiency at the time of target innervation provides a powerful mechanism to ensure survival of only the sympathetic neurons that correctly innervate their targets.

Programmed Cell Death in the Parasympathetic Nervous System

Cell death in parasympathetic neurons is much less well studied than in sympathetic neurons, but these cells do not appear to rely on neurotrophins (NGF, BDNF, or NT-3) for survival. Furthermore, although ciliary ganglion cell survival is supported by ciliary neurotrophic factor (CNTF) *in vitro*, CNTF does not appear to perform this function *in vivo* during normal development. Instead, Ret activation by GDNF and NRTN appears to be the most important determinant of parasympathetic neuron number. Unfortunately, a detailed analysis of cell apoptosis and proliferation

within the parasympathetic nervous system of mutant mice has not been performed for most cell populations. Nonetheless, several important conclusions can be drawn about the role of GDNF and NRTN in parasympathetic neurons:

1. Different parasympathetic neuron populations respond differently to the loss of Ret signaling. For example, otic and sphenopalatine ganglia are essentially completely absent from Ret−/− and GDNF−/− mice, but the reduced cell number in the ciliary and submandibular ganglia of these mice is much less dramatic (Table 1).
2. Reduced number of ganglion cells does not imply increased apoptosis. This is demonstrated by analysis of the sphenopalatine ganglion in GDNF and GFRα1−/− mice. In these animals, the sphenopalatine ganglia are absent as early as E12.5. Furthermore, bromodeoxyuridine labeling in GDNF−/− mice demonstrated reduced proliferation of sphenopalatine ganglion cell precursors, but terminal transferase deoxyuridine triphosphate nick end labeling (TUNEL) analysis failed to demonstrate any increase in cell death. Thus, at least in the sphenopalatine ganglion, GDNF/Ret signaling is required for precursor proliferation and not for survival of mature neurons.
3. Reduced parasympathetic ganglion cell survival may result from abnormal ganglion cell precursor migration. This is illustrated by the otic ganglion in GDNF−/− mice. Like the sphenopalatine ganglion, most otic ganglion cells are absent in GDNF−/− animals. In this case, however, otic ganglion cell precursors migrate abnormally, and TUNEL staining demonstrated increased cell death in the abnormally migrating precursors.

Thus, while programmed cell death may be important for parasympathetic nervous system development, it is much less well understood than in the sympathetic nervous system. In particular, mechanisms for eliminating cells with abnormal axon targeting have not been well documented in the parasympathetic nervous system. Indeed, target innervation by parasympathetic sphenopalatine neurons is dramatically reduced in NRTN−/− and GFRα2−/− mice, but sphenopalatine ganglion numbers are normal, which suggests that these cells are not dependent on target-derived trophic factors for survival.

Mechanisms Governing Neuron Number in the ENS

The ENS is a complex network of neurons and glia within the bowel wall that controls intestinal motility, responds to sensory stimuli, and regulates intestinal secretion and blood flow. To perform these functions, there are roughly as many neurons in the ENS as in the spinal cord, and the ENS comprises many distinct neuron subtypes that differ in function, transmitter phenotype, and pathways of axon pathfinding. Mechanisms of axon targeting in the ENS, however, and mechanisms to ensure that enteric neurons have correctly innervated their targets, are not yet understood. Furthermore, the ENS presents challenges that do not occur in other regions of the nervous system. This is especially true within the myenteric plexus since specific subtypes of myenteric neurons must extend their axons either orally (toward the mouth) or aborally (toward the end of the bowel) for the gut to function normally. Although there must be axon guidance cues present during development to direct these axons, adjacent regions of the bowel wall appear remarkably similar in the mature organism. Thus, unlike the sympathetic nervous system, where targets of innervation are far from the neuronal cell bodies and target-derived trophic factor dependence is an excellent mechanism for ensuring that only correctly targeted neurons survive, it is difficult to imagine how apoptosis could be used in the ENS to eliminate neurons whose axons project in the wrong direction. It is easier to imagine that apoptotic pathways could be important for enteric neurons projecting outside the muscular gut wall (e.g., to villi), but this has not yet been investigated.

With these ideas in mind, it is perhaps not surprising that apoptosis does not appear to occur within the developing or adult ENS of WT mice. This is not to suggest that enteric neurons are trophic factor-independent. Indeed, TrkA, TrkB, TrkC, p75NTR, Ret, GFRα1, and GFRα2 are all expressed within the ENS. Furthermore, enteric neurons undergo apoptotic cell death in primary culture when they are deprived of trophic factors and clearly respond to a variety of neuronal survival factors, including GDNF, NRTN, and NT-3, *in vitro*. In addition, Ret−/−, GFRα1−/−, and GDNF−/− mice miss essentially all enteric neurons from small bowel and colon. In fact, defective Ret signaling is the most commonly identified etiology of distal colon aganglionosis in humans (i.e., Hirschsprung disease). NT-3- and TrkC-deficient animals also have abnormal ENS development, with striking reduction in some subpopulations of enteric neurons. In contrast to these results, NRTN−/− and GFRα2−/− mice have a normal density of enteric neurons, but as in the parasympathetic nervous system, enteric neurons of NRTN−/− and GFRα2−/− mice are smaller than normal, with reduced neuronal projections, at least in some subtypes of neurons. Finally, apoptosis commonly occurs in ENS precursors of mice with a variety of mutations that cause intestinal aganglionosis, including in Ret−/−,

$Sox10^{Dom}/Sox10^{Dom}$, and Phox2b−/− animals. Thus, since ENS precursors depend on trophic factors for survival, but apoptosis does not appear to occur during normal ENS development, these factors must be produced in an adequate supply in WT animals to prevent programmed cell death. This further implies that ENS precursor proliferation must be carefully regulated to avoid producing more neurons than can be supported by available trophic factor. One way this occurs is that the availability of GDNF directly determines the rate of ENS precursor proliferation. Both increases and decreases in GDNF availability alter enteric neuron number via changes in precursor mitotic rates.

Apoptosis in Neuronal Injury

In addition to the physiologic role of apoptosis during normal development, cellular injury may also cause cell death via apoptotic pathways. In the ENS, for example, apoptosis has been reported in age-related myenteric neuron loss, anti-HuD-associated paraneoplastic syndrome, colitis-induced neuronal injury, and diabetes-associated ENS injury. Furthermore, at least in mouse models of diabetes, apoptosis can be reduced by providing additional trophic factor (GDNF) *in vivo*. It is interesting that in contrast to the effect of diabetes on the ENS, sympathetic neuron cell death does not appear to occur in diabetic rats. Thus, once again, the importance of apoptotic pathways in the autonomic nervous system and the ENS is specific to the age of the animal, the apoptotic trigger, and the neuronal subtypes evaluated. Presumably, these differences in the extent of apoptosis in different regions of the nervous system and at different times during life reflect both the availability of trophic factors to support survival and the abundance of intracellular pro- and anti-apoptotic proteins.

Summary

Programmed cell death is important for tissue morphogenesis during development and for the maintenance of tissue homeostasis during adult life. In part, apoptosis is valuable because it provides a way for the organism to specifically eliminate single cells that are no longer needed. This is important in the nervous system, where correct axon targeting and matching of the target size to the number of innervating nerve fibers is critical for function. Because both too much and too little cell death could be detrimental, carefully regulated intracellular and extracellular control mechanisms have been established to control PCD. Many of these mechanisms have been studied in detail in the sympathetic nervous system, where target-derived trophic factors are required for neuron survival during a defined developmental period when neurons are innervating their targets. Because trophic factors are active at the axon tip and are produced in the axon target this strategy effectively eliminates neurons that fail to innervate an appropriate target. Similar strategies are employed in most regions of the nervous system.

Mechanisms of PCD in the autonomic nervous system and the ENS highlight several important themes. First, for PCD to be useful for tissue morphogenesis, the axon tip and neuron cell body must be in different environments. For this reason, PCD tends to occur as neurons are innervating their final targets instead of when they first begin to extend axons. Because many parasympathetic neuron cell bodies are embedded in their targets and many enteric neurons project axons to targets in an environment similar to that around the cell body, it is more difficult to see how PCD could be used in the parasympathetic nervous systen and the ENS to ensure proper axon targeting. Indeed, it is difficult to find evidence that PCD occurs as a part of normal development in either the ENS or the parasympathetic nervous system. Second, for PCD to be effective, different targets need to produce different trophic factors, and subsets of neurons must respond to only a limited array of trophic factors. Even more effective would be a strategy to actively eliminate cells exposed to the 'wrong' trophic factor. This explains the variety of trophic factors and receptors present in the nervous system and the role of p75NTR to induce cell death in the absence of Trk receptor activation. Indeed, given the wide array of neuron subtypes and targets, it is remarkable that the nervous system can be established with so few trophic factors and receptors. It is probably for this reason that developing neurons may respond to a combination of trophic factors or change trophic factor dependence during development. Finally, most postmitotic neurons are needed throughout the life of the organism. For this reason, resistance to apoptosis in neurons that have correctly innervated their targets is important for longevity. In the nervous system, this is accomplished by limiting trophic factor-dependent cell survival to a small developmental period. It is important to note that this requires changes in the intracellular machinery for apoptosis once targets are innervated.

Tremendous advances have been made over the past decade in understanding the role of apoptosis in development and disease, but many challenges remain. In particular, it would be valuable to develop additional strategies to prevent neuronal cell death after injury. Ideally, these strategies should target specific cell populations since global inhibition of apoptosis is likely to be carcinogenic. Continued detailed analysis of the molecular mechanisms of apoptosis

and survival in different defined cell populations is therefore critical to allow targeted therapy and reduce disease-related morbidity and mortality.

See also: Apoptosis in Nervous System Injury; Apoptosis in Neurodegenerative Disease; Enteric Nervous System: Neurotrophic Factors; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Parasympathetic Nervous System; Programmed Cell Death.

Further Reading

Airaksinen MS and Saarma M (2002) The GDNF family: Signalling, biological functions, and therapeutic value. *Nature Reviews Neuroscience* 3(5): 383–394.

Baloh RH, Enomoto H, Johnson EMJ, and Milbrandt J (2000) The GDNF family ligands and receptors: Implications for neural development. *Current Opinion in Neurobiology* 10: 103–110.

Enomoto H, Crawford PA, Gorodinsky A, Heuckeroth RO, Johnson EM Jr, and Milbrandt J (2001) RET signaling is essential for migration, axonal growth, and axon guidance of developing sympathetic neurons. *Development* 128(20): 3963–3974.

Enomoto H, Heuckeroth RO, Golden JP, Johnson EM Jr, and Milbrandt J (2000) Development of cranial parasympathetic ganglia requires sequential actions of GDNF and neurturin. *Development* 127: 4877–4889.

Fagan AM, Zhang H, Landis S, Smeyne RJ, Silos-Santiago I, and Barbacid M (1996) TrkA, but not TrkC, receptors are essential for survival of sympathetic neurons *in vivo*. *Journal of Neuroscience* 16(19): 6208–18.

Francis N, Farinas I, Brennan C, et al. (1999) NT-3, like NGF, is required for survival of sympathetic neurons, but not their precursors. *Developmental Biology* 210(2): 411–427.

Gewies A (2003) ApoReview: Introduction to apoptosis. http://www.celldeath.de/encyclo/aporev/aporev.htm (accessed 3 August 2007).

Gianino S, Grider JR, Cresswell J, Enomoto H, and Heuckeroth RO (2003) GDNF availability determines enteric neuron number by controlling precursor proliferation. *Development* 130(10): 2187–2198.

Honma Y, Araki T, Gianino S, et al. (2002) Artemin is a vascular-derived neurotropic factor for developing sympathetic neurons. *Neuron* 35(2): 267–282.

Majdan M, Walsh GS, Aloyz R, and Miller FD (2001) TrkA mediates developmental sympathetic neuron survival in vivo by silencing an ongoing p75NTR-mediated death signal. *Journal of Cell Biology* 155(7): 1275–1285.

Newgreen D and Young HM (2002) Enteric nervous system: Development and developmental disturbances: Part 2. *Pediatric and Developmental Pathology* 5(4): 329–349.

Oppenheim RW (1991) Cell death during development of the nervous system. *Annual Review of Neuroscience* 14: 453–501.

Putcha GV, Harris CA, Moulder KL, Easton RM, Thompson CB, and Johnson EM Jr. (2002) Intrinsic and extrinsic pathway signaling during neuronal apoptosis: Lessons from the analysis of mutant mice. *Journal of Cell Biology* 157(3): 441–453.

Putcha GV and Johnson EM Jr. (2004) Men are but worms: Neuronal cell death in *C. elegans* and vertebrates. *Cell Death and Differentiation* 11(1): 38–48.

Reichardt LF (2006) Neurotrophin-regulated signalling pathways. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 361(1473): 1545–1564.

Sastry PS and Rao KS (2000) Apoptosis and the nervous system. *Journal of Neurochemistry* 74(1): 1–20.

Wikipedia (2007) Apoptosis. http://en.wikipedia.org/wiki/Apoptosis (accessed 3 August 2007).

Yuan J and Yankner BA (2000) Apoptosis in the nervous system. *Nature* 407(6805): 802–809.

Autonomic Disorders

E L Phillips and P D Donofrio, Vanderbilt University Medical Center, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The autonomic nervous system regulates cardiovascular, gastrointestinal, genitourinary, thermoregulatory, exocrine, and pupillary function to maintain internal homeostasis. The autonomic system has two major components, the sympathetic and parasympathetic divisions. The sympathetic system sends efferents from T1–L2 to end organs, producing a 'fight-or-flight' response. This causes an increase in heart rate, cardiac contractility, and blood pressure; dilated coronary arteries and pupils; bronchodilation; piloerection; sweating; constriction of intestinal sphincters; and ejaculation. The parasympathetic system's efferents emanate from cranial nuclei and S2–S4 nerve roots to end organs and regulate 'rest and digest' functions. Increased parasympathetic activity results in heightened gastric secretions and peristalsis, decreased heart rate and cardiac contractility, bronchoconstriction, constricted pupils, bladder contraction, and maintenance of erection.

Both of these systems are controlled by higher central nervous system (CNS) centers. The nucleus tractus solitarius (NTS) is considered one of the main central processors. It receives input from many central structures, including the forebrain, hypothalamus, hippocampus, amygdala, periaqueductal gray, raphe nucleus, and pontomedullary reticular formation. The NTS also receives sensory afferent information from the periphery, including signals from internal nuclei (i.e., chemoreceptors, osmoreceptors, thermoreceptors, and baroreceptors). The efferents from the NTS project to the brain stem and spinal cord.

The sympathetic system has short preganglionic and long postganglionic neurons (Figure 1). Preganglionic sympathetic fibers originate in the interomediolateral cell column in lamina VII of the spinal cord from levels T1 to L2. From there, they project to paravertebral and prevertebral ganglia, also called the sympathetic chain or trunk. This chain runs from cervical to sacral levels on each side of the cord. The prevertebral ganglia are located in the celiac plexus and include the celiac ganglia, superior mesenteric ganglion, and inferior mesenteric ganglion.

The sympathetic preganglionic nerves secrete acetylcholine, and most of the sympathetic postganglionic fibers secrete norepinephrine. The exception is the sympathetic postganglionic fibers that innervate sweat glands and adrenal medulla, which secrete acetylcholine.

The parasympathetic preganglionic fibers originate from cranial nuclei 3, 7, 9, and 10 and sacral parasympathetic nuclei in the lateral gray matter from S2–S4 (Figure 2). The preganglionic fibers project to terminal ganglion in or near the effector organs. In contradistinction to the sympathetic system, the parasympathetic system has long preganglionic fibers and short postganglionic fibers. The parasympathetic preganglionic and postganglionic fibers secrete acetylcholine.

Symptoms and Signs of Autonomic Dysfunction

Autonomic dysfunction may result in many different symptoms. One common symptom is orthostatic hypotension, which patients may describe as a feeling of dizziness, lightheadedness, weakness, confusion, blurred vision, or syncope after a postural change. To test for orthostasis, the blood pressure must be measured while lying, sitting, and standing. If there is a decrease in systolic pressure of at least 20 mmHg or a decrease in diastolic pressure of at least 10 mmHg within 3 min of standing accompanied by symptoms of cerebral hypoperfusion, then the patient meets criteria for orthostatic hypotension proposed by the American Autonomic Society and the American Academy of Neurology. Orthostasis associated with an appropriate rise in heart rate is consistent with hypovolemia, which is seen with dehydration, medication side effects, or diuretics. In sympathetic adrenergic failure, the characteristic feature is a drop in blood pressure without an appropriate rise in heart rate.

Intolerance to position changes can be seen without a drop in blood pressure. In positional tachycardia syndrome, there is an increase in heart rate of more than 30 beats per minute when changing from a reclined to upright position. This is most commonly observed in women younger than age 50 years. Patients complain of dizziness, usually without syncope, during postural change or modest exertion.

Gastrointestinal dysfunction is a common complication of autonomic neuropathy. Patients may have symptoms arising from the proximal gastrointestinal tract, including dysphagia and gastroparesis, or from the distal gastrointestinal tract, including chronic intestinal pseudoobstruction, constipation, diarrhea, and fecal incontinence. Patients may complain of

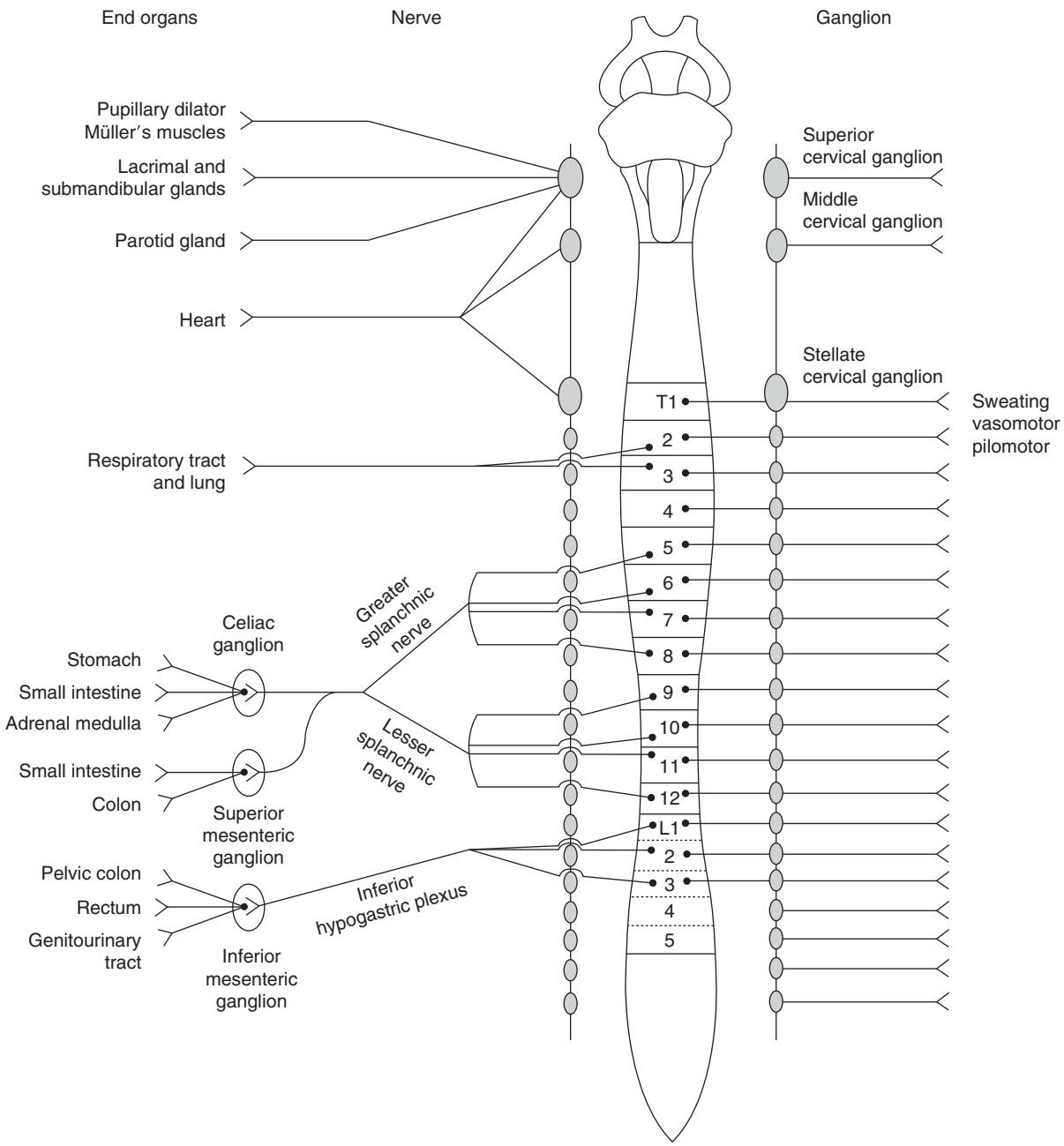

Figure 1 Anatomic framework of the sympathetic system. Reprinted from Donofrio PD (2001) Autonomic disorders. *The Neurologist* 7: 220–233, with permission.

heartburn, nausea, early satiety, loss of appetite, postprandial emesis, bloating, belching, constipation, or diarrhea.

Autonomic dysfunction can be associated with a variety of urologic complications, including bladder atony, nocturia, frequency, urgency, incontinence, and retention. The differences in symptoms depend on the location of the lesion in the sympathetic or parasympathetic system. Lesions above the sacral cord can cause detrusor overactivity. In this situation, the parasympathetic system dominates and causes the detrusor muscle to contract, resulting in increased intravesicular pressure, which the patient senses as imminent micturition and can lead to urge incontinence. When there is a lesion between the pons and the sacral cord, detrusor–sphincter dyssynergia occurs. This is characterized by a nonrelaxing sphincter during detrusor contraction. Dyssynergia can lead

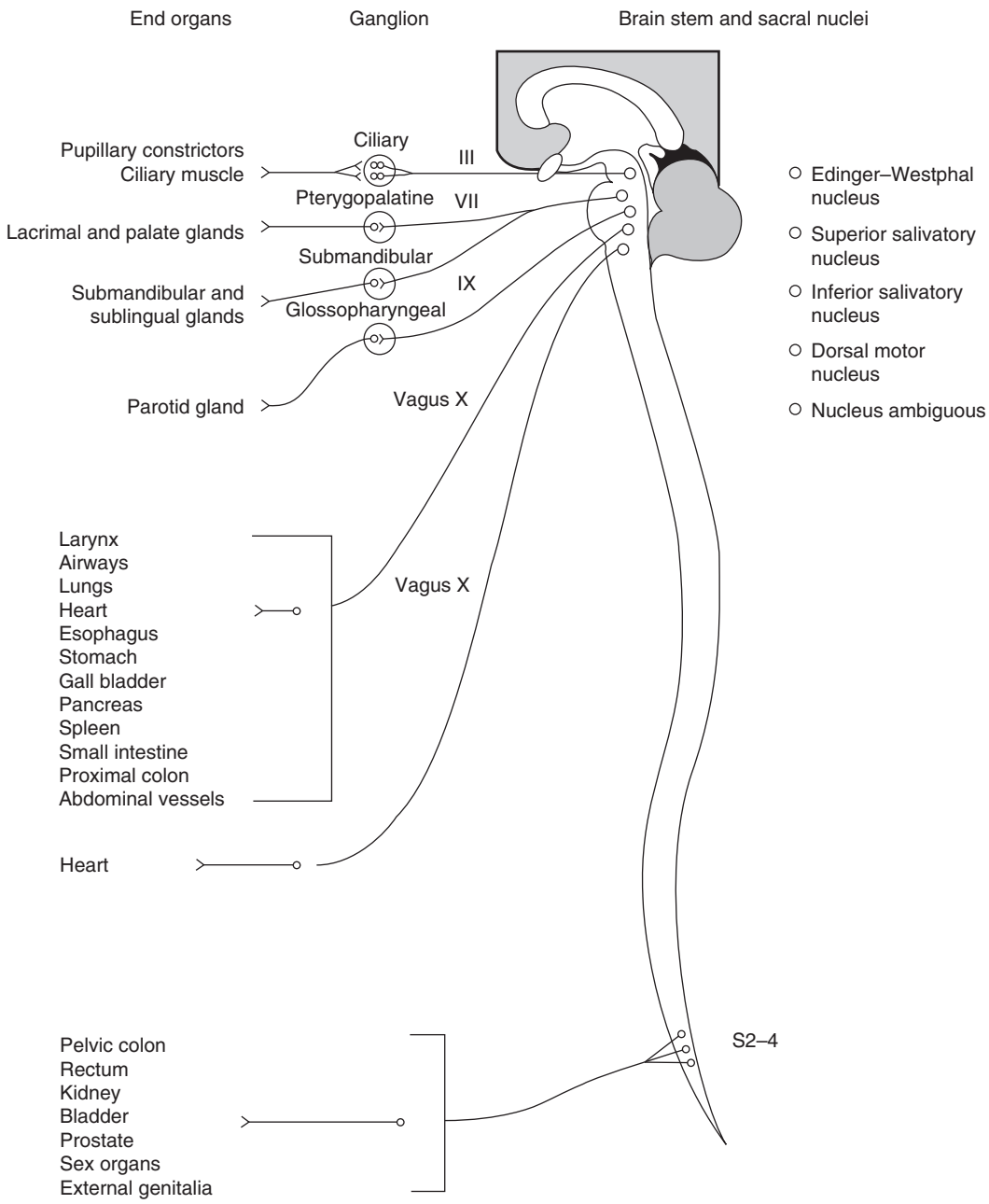

Figure 2 Anatomic framework of the parasympathetic system. Reprinted from Donofrio PD (2001) Autonomic disorders. *The Neurologist* 7: 220–233, with permission.

to incomplete bladder emptying, frequent urinary tract infections, and chronic renal failure from hydronephrosis. Bladder atony and overflow incontinence occur with more peripheral lesions of autonomic system. In this situation, the parasympathetic pelvic nerves are affected and are unable to cause contraction of the detrusor muscle.

Often, erectile dysfunction is one of the earliest symptoms of autonomic disease in men and is usually followed by ejaculatory failure. Women usually describe decreased libido, decreased ability for orgasm, and vaginal dryness.

When pupillomotor function is affected, patients complain of blurry vision, poor night vision, or a glare after bright lights. Patients with parasympathetic dysfunction have dry eyes and dry mouth, secondary to secretomotor paralysis of the salivary and lacrimal glands.

Another complication seen with autonomic dysfunction is an abnormality of sweating, either hypohidrosis

(extreme dryness) or hyperhidrosis (excessive perspiration). Patients may not notice a change in sweating; therefore, probing questioning regarding sweating on hot days, after a hot bath, or during a fever may be necessary.

When vasomotor instability is present, patients may complain of extremely cold hands and feet. On examination of the distal extremities, patients may manifest changes in the appearance of the skin, including mottling, redness, or pallor, or trophic changes in the skin, nails, and hair, including thinning of the skin, brittle nails, and thickening or thinning of the hair.

Diseases Affecting the Autonomic System

Many diseases affect the autonomic nervous system; therefore, a systematic approach to classifying them is needed. The simplest approach is to divide them into two groups – those that affect the CNS and those that affect the peripheral nervous system. This section briefly discusses a few of these diseases.

Progressive autonomic failure, Parkinson's disease, and Shy–Drager syndrome are all neurodegenerative disorders that cause autonomic dysfunction. Progressive autonomic failure usually presents in men between the ages of 40 and 60 years. Onset is usually insidious, with slow progression leading to death within 10–15 years. Clinical findings are typically limited to the autonomic system. Patients with Parkinson's disease usually present with resting tremor, rigidity, bradykinesia, and postural instability and later develop autonomic symptoms. Shy–Drager syndrome typically presents with significant autonomic symptoms along with parkinsonian features; however, tremor is usually minimal. Other diseases of the CNS associated with autonomic dysfunction include multiple sclerosis, strokes, brain stem tumors, spinal cord lesions, tabes dorsalis, and Adie syndrome (Table 1).

Diseases of the peripheral autonomic system can further be categorized depending on the coexistence of a somatic peripheral neuropathy. Some of the peripheral neuropathies commonly associated with autonomic symptoms include diabetes, Guillain–Barré syndrome, Riley–Day syndrome, infectious neuropathies, amyloidosis, and acute intermittent porphyria (Table 2).

The most common cause of autonomic neuropathy in developed countries is diabetes mellitus. The neuropathy in diabetes typically presents as a distal sensorimotor polyneuropathy, and the autonomic neuropathy presents late in the course of the illness.

Guillain–Barré syndrome (acute inflammatory demyelinating polyradiculoneuropathy) is an acquired monophasic illness that begins as an acute ascending demyelinating sensorimotor polyneuropathy. Autonomic dysfunction occurs in approximately 60% of patients with Guillain–Barré syndrome as cardiac arrhythmias, postural hypotension, transient hypertension, urinary retention, and ileus.

Familial dysautonomia (Riley–Day syndrome or HMSN III) is an autosomal recessive neuropathy with prominent autonomic symptoms, most commonly observed in Ashkenazi Jewish children. It presents in infancy with loss of pain and temperature sensation, hypotonia, and hyporeflexia. Patients have absence of tear production and lingual fungiform papillae. Autonomic symptoms can present at any time during the disease course.

Table 1 Central nervous system disorders affecting autonomic function

Progressive autonomic failure (idiopathic orthostatic hypotension)
Multiple-system atrophy (Shy–Drager syndrome)
Parkinson's disease
Spinal cord lesions
Wernicke's encephalopathy
Cerebrovascular disease
Brain stem tumors
Multiple sclerosis
Adie's syndrome
Tabes dorsalis

Adapted from McLeod JG and Tuck RR (1987) Disorders of the autonomic nervous system: Part I. Pathophysiology and clinical features. *Annals of Neurology* 21: 419–430.

Table 2 Peripheral autonomic neuropathy with somatic neuropathy

Autonomic dysfunction clinically important
Diabetes mellitus
Amyloidosis
Guillain–Barré syndrome
Acute intermittent porphyria
Familial dysautonomia (Riley–Day syndrome, HMSN III)
Chronic sensory and autonomic neuropathy
Infections: HIV, leprosy, Chagas disease, diphtheria, and Lyme disease
Autonomic dysfunction usually clinically unimportant
Hereditary sensory and autonomic neuropathy type I, II, IV, and V
Fabry's disease
Toxic neuropathies (vincristine sulfate, acrylamide, heavy metals, perhexiline maleate, organic solvents, marine toxins, vacor, cisplatin, paclitaxel, amiodarone, pentamidine)
Metabolic (B_{12} deficiency, alcohol-induced neuropathy, uremia)
Connective tissue diseases (systemic lupus erythematosus, rheumatoid arthritis, Sjogren's disease)
Chronic inflammatory demyelinating polyneuropathy

HIV, human immunodeficiency virus; HMSN, hereditary motor and sensory neuropathy.
Adapted from McLeod JG and Tuck RR (1987) Disorders of the autonomic nervous system: Part I. Pathophysiology and clinical features. *Annals of Neurology* 21: 419–430.

Infectious neuropathies have been described in patients with human immunodeficiency virus (HIV), Chagas disease, leprosy, and diphtheria. The autonomic dysfunction associated with HIV infection can arise from direct viral effects, malnutrition, or drug side effects. Chagas disease is caused by the protozoan *Trypanosoma cruzi*. Severe gastrointestinal and cardiovascular autonomic dysfunction is typically observed during the chronic phase of the illness. Leprosy is contracted from the infectious agent *Mycobacterium leprae*. In the cooler parts of the body, there is loss of pain and temperature sensation along with focal anhidrosis. Several weeks after a cutaneous or pharyngeal diphtheria infection, a toxin-mediated neuropathy develops. This can present with multiple cranial mononeuropathies, hyporeflexia, sensory ataxia, and accommodation paralysis in the setting of a preserved light response.

In amyloidosis there is a deposition of insoluble protein fibers in β-pleated sheets within many tissues and organs. The neuropathy found in primary and hereditary amyloidosis is frequently associated with an autonomic neuropathy. However, secondary amyloidosis is not usually associated with autonomic dysfunction. Amyloidosis can be diagnosed by gingival biopsy, aspiration of subcutaneous fat pad, or a biopsy of rectal mucosa.

Acute intermittent porphyria is an autosomal dominant disease due to a defect in the enzyme porphobilinogen deaminase, which leads to a buildup of porphobilinogen and aminolevulinic acid. Most patients present between 18 and 40 years of age with intermittent attacks of abdominal pain, neuropathy, and constipation lasting several days. The neuropathy is predominantly motor and autonomic. Diagnosis is confirmed by an elevated urine porphobilinogen. Treatment of an acute attack consists of hematin 4 mg kg^{-1} day^{-1} for 4 days.

The peripheral autonomic neuropathies not typically associated with a peripheral neuropathy include botulism, pure adrenergic neuropathy, acute pandysautonomia, cholinergic dysautonomia, acute paraneoplastic autonomic neuropathy, and chronic idiopathic anhidrosis (Table 3).

Botulism is a neuromuscular junction disorder caused by the binding of *Clostridium botulinum* toxin to the presynaptic terminal. This toxin prevents the release of acetylcholine, leading to autonomic symptoms and a severe quadriplegia from failure of the neuromuscular junction. Dilated pupils with poor response to direct light and accommodation are classic features, in contrast to diphtheria, in which the light reflex is not affected.

Acute pandysautonomia is a disorder of both sympathetic and parasympathetic systems. The etiology is likely an autoimmune process or postviral reaction since the onset is usually after a viral illness. Approximately 50% of patients will have an autonomic ganglionic acetylcholine receptor antibody.

Table 3 Peripheral autonomic neuropathy without significant sensorimotor neuropathy

Acute and subacute autonomic neuropathy
Acute pandysautonomia
Cholinergic dysautonomia
Acute paraneoplastic autonomic neuropathy
Chronic idiopathic anhidrosis
Pure adrenergic neuropathy
Botulism

Adapted from McLeod JG and Tuck RR (1987) Disorders of the autonomic nervous system: Part I. Pathophysiology and clinical features. *Annals of Neurology* 21: 419–430.

Acute cholinergic neuropathy is similar to acute pandysautonomia, but it only affects the postganglionic cholinergic neurons. Pure adrenergic neuropathy is very rare, affecting postganglionic sympathetic neurons.

Acute paraneoplastic autonomic neuropathy is frequently associated with anti-Hu antibodies. Small cell lung cancer is the most common malignancy associated with these antibodies. Other antibodies detected with paraneoplastic autonomic neuropathies include Purkinje cell cytoplasmic antibodies type 2 and antibodies to collapsing response-mediator protein 5.

Chronic idiopathic anhidrosis is an acquired neuropathy producing diffuse anhidrosis without other signs of autonomic failure. When exposed to heat or after exercise, patients become hot, lightheaded, flushed, and weak because of their inability to release heat through sweating.

Autonomic Testing

Patients with autonomic signs or symptoms are candidates for autonomic testing if verification is necessary. Depending on the clinical presentation of the patient, evaluation should begin with hematologic, chemistry, autoantibody, and pathologic studies. Table 4 lists some of the more commonly used tests for evaluating autonomic dysfunction.

Further testing can be pursued to define the physiology of the sympathetic and parasympathetic systems. All patients who are scheduled to undergo autonomic testing need to take certain measures in preparation for the testing. They should not consume any caffeine within the previous 24 h. They should avoid nicotine and food for 3 h prior to testing. Medications that affect autonomic function should be avoided for at least 48 h prior to testing, and preferably 5 half-lives before testing. Medications to be avoided

include antihypertensives, volume contractors (e.g., diuretics), volume expanders (e.g., fludrocortisones), anticholinergics (e.g., antihistamines, antidepressants, and decongestants), and analgesics. Also discontinued should be anything that mechanically changes blood flow (e.g., compressive stockings and corsets) 24 h prior to testing.

One means to assess the function of the sympathetic and parasympathetic systems is by evaluating cardiovascular physiology. As mentioned previously, activation of the sympathetic system leads to tachycardia, elevation of blood pressure, increased cardiac contractility, and dilated coronary arteries. Conversely, stimulation of the parasympathetic system reduces heart rate and cardiac contractility. The baroreflex arc modulates these two systems. The baroreflex arc consists of afferent input from arterial baroreceptors in the aortic arch, carotid sinus, thoracic arteries, cardiac mechanoreceptors, and pulmonary stretch receptors. Activation of the baroreceptors leads to increased afferent input to central processors. This results in a decrease in sympathetic and increase in parasympathetic activity. Conversely, when the afferent activity decreases, there is an increase in sympathetic and decrease in parasympathetic activity.

Table 4 Laboratory assessment of autonomic neuropathy

- Chemistry, hematology, and pathology
  - Complete blood count and differential
  - Fasting blood glucose, glucose tolerance test if diabetes suspected
  - HIV testing
  - Immunoelectrophoresis of blood and urine
  - Plasma norepinephrine (supine and standing)
  - Porphyria investigations
    - Urinary porphyrin concentration (24 h urine collection for aminolevulinic acid, porphobilinogen, and porphyrins)
    - Erythrocyte porphobilinogen deaminase activity (patients with type I and type III disease have approximately 50% enzyme activity)
  - Genetic testing for inherited neuropathies
  - Fat aspirate, rectal biopsy, or gingival biopsy for amyloid
- Autoantibody assessment
  - Anti-nuclear antibody
  - Rheumatoid factor
  - Anti-Ro/SS-A
  - Anti-La/SS-B
  - Antibodies to neuronal nicotinic acetylcholine receptor
  - Antibodies to P/Q-type calcium channel
  - Antibodies to acetylcholine receptor
  - Paraneoplastic antibodies
    - Anti-Hu (type 1 anti-neuronal nuclear antibody, ANNA-1)
    - Purkinje cell cytoplasmic antibodies type 2 (PCA-2)
    - Collapsin response-mediator protein 5 (CRMP-5)

Adapted from Freeman R (2005) Autonomic peripheral neuropathy. *Lancet* 365: 1259–1270.

Valsalva Maneuver

The Valsalva maneuver evaluates the sympathetic and parasympathetic responses of the baroreflex arc. The patient exhales into a manometer and maintains a pressure of 40 mmHg for 15–20 s. Heart rate and blood pressure are monitored 30 s before and 2 min after testing. There are four phases of the Valsalva maneuver (Table 5). During phase 1, there is an increase in intrathoracic pressure that mechanically causes a brief increase in blood pressure and decrease in heart rate. In early phase 2, there is a reduction of venous return and a subsequent decrease in stroke volume, causing a decrease in blood pressure. In late phase 2, the decreased blood pressure activates the baroreflex that causes a sympathetically mediated increase in heart rate and blood pressure back toward baseline levels. When the patient terminates the Valsalva maneuver, blood refills the pulmonary vasculature. This causes the change seen in phase 3 – a temporary further decline in blood pressure. During phase 4, there is an increase in venous return, which leads to a compensatory decrease in heart rate and increase in blood pressure that may overshoot baseline blood pressure. By monitoring the heart rate during the Valsalva maneuver, the Valsalva ratio can be measured. This is calculated by dividing the highest heart rate during Valsalva by the lowest heart rate usually occurring during phase 4. In autonomic failure, the reflex bradycardia and blood pressure overshoot is typically absent. Valsalva ratio varies with age, but a ratio less than 1.1 is abnormal at any age.

Sinus Arrhythmia (Deep Metronomic Breathing; Heart Rate Response to Respiration)

Respiratory sinus arrhythmia is a normal phenomenon; the term relates to the increase in heart rate with

Table 5 Phases of Valsalva

Phase	*Timing*	*Physiology*	*Heart rate*	*Blood pressure*
1	Early valsalva	Increase in intrathoracic pressure	Decrease	Increase
2 – early		Decreased venous return	Decrease	Decrease
2 – late		Baroreflex	Increase	Increase
3	End of valsalva	Refills pulmonary vasculature	Increase	Decrease
4		Increase in venous return	Decrease	Increase

inspiration and decrease during expiration. Measuring this variability in heart rate is a reliable way to assess the parasympathetic innervation of the heart mediated by the vagus nerve. The heart rate is monitored while the patient is asked to breathe six respirations per minute (each breath lasting 10 s) for seven or eight cycles. After resting for 5 min, this sequence is repeated again. Age, pulmonary disease, hypocapnia, hyperventilation, cardiac failure, tachycardia, and CNS depression can all lead to a decrease in sinus arrhythmia.

There are several ways to assess sinus arrhythmia. One way is to measure heart rate variation, which is determined by taking the difference between the average heart rate during inspiration from the average heart rate during expiration. Heart rate variation greater than 15 beats per minute (bpm) is considered normal. Less than 10 bpm is abnormal in younger patients (<40 years), and less than 5 bpm is abnormal at any age. Another measure includes calculating the expiratory/inspiratory ratio (E/I ratio). This is the longest R-R interval during expiration divided by the shortest R-R interval during inspiration. The E/I ratio varies depending on age. However, an E/I ratio less than 1.1 is abnormal at any age.

Sympathetic Skin Test

The sympathetic skin response (SSR) can be used to evaluate the integrity of the entire sympathetic sudomotor response, including the central and peripheral components. The preoptic anterior hypothalamus, spinal cord, viscera, and skin have thermoreceptors that input to the posterior hypothalamus. The posterior hypothalamus then determines a set point. Efferent output is directed through the brain stem down to the intermediolateral cell column. The preganglionic sympathetic cholinergic fibers synapse on the paravertebral sympathetic ganglia. Postganglionic sympathetic cholinergic neurons terminate on eccrine sweat glands. The resulting sweat response produces the SSR.

SSRs occur spontaneously or can be provoked by a stimulus. The stimulus can be direct, such as electrical activation of the peripheral C-fibers, or indirect (auditory, visual, startle, cough, or inspiratory gasp), which activates the centrally mediated sympathetic sudomotor response leading to a change in skin potential. The recording electrodes are usually placed on the palms and soles, with references on the dorsal surfaces. The skin potential is the difference in voltage between two skin surfaces. This potential is largely determined by the sweat glands and the epidermis. The initial change in potential is likely due to sweating and the following changes are likely secondary to changes in skin potential. Normally, the SSR amplitude is 0.5–3 mV and the latency is 1.5 s in the hand and 2 s in the foot. The SSR is easy to perform using conventional electromyography equipment; however, interpreting the results is challenging due to a large variability in responses. Most electromyographers agree that an absent SSR is abnormal. However, there are no well-accepted criteria for interpreting responses that have a reduction in amplitude or latency. Soliven et al. published reference values of 0.791 ± 0.346 mV for the palm and 0.388 ± 0.227 mV for the sole of the foot. The SSR decreases with age, likely due to the reduction in the number of sweat glands during aging. Other factors that interfere with the amplitude include skin creams and cold, dry, or oily skin.

Quantitative Sudomotor Axon Reflex Test

The quantitative sudomotor axon reflex test (QSART) assesses the postganglionic cholinergic sympathetic innervation of eccrine sweat glands. This test is performed by applying acetylcholine to the skin through a multicompartmental sweat cell. Through iontophoresis, acetylcholine binds to the muscarinic receptors on the sweat gland, causing an immediate sweat response. Acetylcholine also binds to the postganglionic nicotinic receptors, which causes an antidromic activation. When the antidromic activation reaches a branch point, it travels orthodromically to release acetylcholine in other sweat glands near the initial iontophoretic stimulation. This is the indirect axon reflex sweat gland response. One of the compartments of the sweat cell measures the sweat production from this indirect axon reflex using a sudorometer. This recording is usually done prior to, during, and after iontophoresis. The indirect axon–reflex sweat response usually occurs 1 or 2 min after iontophoresis. The usual recording sites for the QSART include the proximal dorsal foot (sural nerve), the lateral proximal calf (saphenous nerve), the distal medial leg (peroneal nerve), and the middle forearm (ulnar nerve). A normal response in men is 2 or 3 $\mu l^{-2} cm^{-2}$ and that in women is 0.25–1.2 $\mu l^{-2} cm^{-2}$, with some variation depending on the site tested.

QSART evaluates postganglionic sympathetic sudomotor function in different regions of the body. The most distal sites may have a small or absent response in the early stages of a small fiber neuropathy, whereas proximal sites have a normal QSART. However, as the disease progresses, the more proximal sites will begin to show abnormal responses. When the lesion is preganglionic or central, such as in spinal cord lesions or in multiple system atrophy, the QSART response should be normal at all sites. However, if the preganglionic lesion has been present for an extended period of time, the QSART response

may be abnormal. Theories for this phenomenon include a transsynaptic defect or sweat gland atrophy.

Other abnormal findings in central lesions include resting sweat activity (RSA) and persistent sweat activity (PSA). RSA is defined as sweat response at a skin temperature that is below the normal temperature for a sweat response. PSA is the continuation of the sweat response even after the stimulus is removed. Both RSA and PSA are thought to arise from damaged sympathetic sudomotor axons that spontaneously fire. This is commonly seen in early or painful neuropathies, reflex sympathetic dystrophy, hyperhidrosis, or anxious patients.

Treatment of Autonomic Dysfunction

If possible, treatment of autonomic dysfunction should be directed at the underlying disease. For example, Guillain–Barré syndrome can be treated with plasma exchange or intravenous immunoglobulin (IVIG). In Lambert–Eaton myasthenic syndrome (LEMS), the underlying malignancy should be treated. Immunosuppressants, plasma exchange, IVIG, and 3,4-diaminopyridine can improve symptoms in LEMS. A trivalent antitoxin exists for the treatment of botulinum; however, it is only beneficial if given within the first 30 min of exposure. In porphyria, hematin, phenothiazines, and meperidine may provide some relief. Tight glycemic control can prevent and stabilize diabetic autonomic neuropathy. Unfortunately, many autonomic diseases are not treatable, and therefore management must be symptomatic.

Orthostatic Hypotension

Treating orthostatic hypotension can be as simple as increasing the intake of salt (10 g of sodium per day) and fluids (2–2.5 l per day in adults), wearing elastic stockings, rising slowly, and avoiding hot environments, large meals, and alcohol. Sleeping with the head of the bed elevated 10–20 degrees attenuates nocturnal diuresis, which decreases the morning symptoms of orthostatic hypotension.

If simple lifestyle changes are not effective, then multiple medications can be prescribed (Table 6). Fludrocortisone (0.1–1.0 mg qday) is a mineralocorticoid that increases blood volume. Side effects include peripheral edema, hypokalemia, and headache. It should not be used in patients with chronic heart failure.

Midodrine, a direct peripherally acting α_1-adrenoreceptor agonist, can be added to fludrocortisone. It causes arterial and venous constriction and has a duration of action of 4 h. Midodrine does not cross the blood–brain barrier, so central sympathomimetic side effects are absent. Initial dosing is 2.5 mg tid titrating up to 10 mg tid. Ephedrine and pseudoephedrine are direct and indirect α_1-adrenoreceptor agonists. They not only act directly on the α_1-adrenoreceptor but also cause norepinephrine release from postganglionic sympathetic neurons. Because of the indirect action, these medications are best used in patients with partial or incomplete lesions. They may produce central sympathomimetic side effects, including anxiety, tachycardia, and tremulousness.

Table 6 Pharmacotherapy of orthostatic hypotension

Mineralocorticoids	
Fludrocortisone	0.1–1.0 mg po daily
Sympathomimetic agents	
Midodrine	2.5–10 mg po tid-qid
Clonidine	0.1–0.3 mg po bid
Ephedrine	15–45 mg po tid
Yohimbine	2.5–5.4 mg po tid
V2 receptor agonists	
Desmopressin (DDAVP)	100–400 mg po at bedtime or 5–40 μg intranasally at bedtime
Erythropoietin	25–75 U kg^{-1} sc three times a week
Dopamine blocking agents	
Metoclopramide	10–15 mg po up to qid
Prostaglandin synthetase inhibitors	
Indomethacin	25–50 mg po tid
Ibuprofen	300–400 mg po tid
Somatostatin analog	
Octreotide	25–50 μg sc tid

Erythropoietin has been shown to improve symptoms in patients with orthostatic hypotension. The exact mechanism for this is unknown, but it is likely explained by changes in blood viscosity, an increase in hematocrit and central blood volume, and neurohumoral effects on blood vessels. Desmopressin acts on V2 receptors in the kidney to inhibit water excretion and to increase volume expansion. Clonidine is an α_2 agonist which reduces sympathetic tone centrally; thus, in healthy patients it decreases blood pressure. However, in severe central autonomic dysfunction, central sympathetic receptors are depleted and postsynaptic α_2 receptors in the periphery may be activated by clonidine, leading to increased blood pressure. Caution must be used when treating orthostatic hypotension with clonidine because of its propensity to cause hypotension.

When treating orthostatic hypotension, one must be cognizant of coexisting supine hypertension. Treatment of supine hypertension in patients who have orthostatic hypotension includes elevating the head of the bed or using short-acting antihypertensives at night. In this situation, captopril or nitrates could be good choices. If supine hypertension becomes a problem, ambulatory monitoring may be needed to design a treatment plan.

Gastrointestinal Atonomic Dysfunction

If gastroparesis is related to diabetes, controlling hyperglycemia can improve gastric motility. Patients should be advised to eat frequent small meals with decreased fat content. Agents that promote motility can be used, including metoclopramide, domperidone, and erythromycin. Rarely, a jejunostomy tube is needed.

Bowel hypomotility initially should be treated with increased fluid intake, increased fiber intake (5–25 g per day), and bulk agents (bran and psyllium). Pharmacological treatment usually begins with stool softeners (docusate) and laxatives (polyethylene glycol or lactulose) (Table 7). Lactulose is a disaccharide and should be avoided in diabetics since a small amount may be absorbed and cause elevations in blood glucose. Osmotic laxatives (sodium and magnesium salts) should not be used for extended periods due to a concern for producing fluid and electrolyte imbalance. Stimulant laxatives (senna, cascara, bisacodyl, and phenolphthalein) should also be prescribed only on a short-term basis since these agents can cause smooth muscle atrophy and neuronal damage.

Neurogenic diarrhea may be treated with clonidine. Psyllium hydrophilic mucilloid may be used as a bulk-forming agent. Loperamide may also be helpful to treat diarrhea. In diabetic diarrhea, bowel stasis may lead to bacterial overgrowth, in which case treatment with tetracycline is useful. For sphincter incontinence, phenylephrine ointment 1% may be applied topically to improve sphincter tone for approximately 4 h.

Sexual Dysfunction

Erectile dysfunction requires a thorough medical history to exclude other causes, especially medications. One treatment option is a phosphodiesterase type 5 inhibitor, which inhibits the breakdown of cyclic GMP. This results in relaxation of smooth muscle and increased blood flow in the corpus cavernosum. Sildenafil 50 mg, tadalafil 20 mg, and vardenafil 20 mg are the three drugs currently marketed in the United States for erectile dysfunction. These medications are contraindicated in patients with unstable heart disease, orthostatic hypotension, significant hypotension, or in patients on nitrates. Other treatments for erectile dysfunction include intracorporeal injections of phentolamine, papaverine, or prostaglandin E. Vasoactive drugs can be delivered transurethrally. Mechanical devices such as constricting rings or a vacuum erection device are other options. If these therapies do not provide benefit, then penile prosthetic implants may be considered. Women with autonomic dysfunction may benefit from vaginal lubricants and estrogen creams. Table 8 lists treatments for erectile dysfunction.

Urinary Dysfunction

Treatment of incontinence should begin with scheduled voiding every 3 or 4 h. Combining this plan with Valsalva or suprapubic pressure can increase bladder contractions and bladder emptying. In many cases, pharmacologic treatment is needed (Table 9). Anticholinergics and tricyclic antidepressants may be used to help decrease detrusor muscle activity. If oral medications do not provide relief, Botox injections into the detrusor muscle can be used to decrease detrusor overactivity by inhibiting acetylcholine release. Alpha-blockers may be used for sphincter dysfunction. Occasionally, desmopressin can provide relief by decreasing urine production for 6–8 h. Often, clean intermittent catheterization or indwelling catheterization is required. Many surgical procedures may be performed to treat incontinence, including bladder neck suspension, augmentation cystoplasty, and artificial sphincter and urinary diversion with stoma collection bag.

Table 7 Pharmacotherapy of bowel hypomotility

Bulk agents	
Bran	1–6 tablespoons po daily
Psyllium	2.2–45 g po daily
Methylcellulose	1 tablespoon po daily to tid
Stool softeners and lubricants	
Mineral oil	15–45 ml po prn
Docusates	50–200 mg po daily
Osmotic laxatives and cathartics	
Lactulose	10–40 g po daily
Sorbitol	30–150 ml po prn
Magnesium citrate	150–300 ml po daily
Contact cathartics	
Bisacodyl	10–15 mg po prn
Senna	2–8 tabs po daily
Prokinetic agents	
Metoclopramide	10–15 mg po up to qid
Domperidone	10–20 mg po qid
Erythromycin	250 mg tid

Table 8 Pharmacotherapy of erectile dysfunction

Phosphodiesterase type 5 inhibitors	
Sildenafil	50 mg po qday prn
Tadalafil	5–20 mg po qday prn
Vardenafil	5–20 mg po qday prn
Intracorporeal injections	
Phentolamine	
Papaverine	
Prostaglandin E	
Transurethral suppository	
Alprostadil	
Vacuum erection device	
Penile prosthesis	

Table 9 Pharmacotherapy of urinary dysfunction

Adrenergic blocking agents (sphincter dysfunction)	
Prazosin	1–5 mg po bid to tid
Terazosin	1–5 mg po daily
Anticholinergic agonists (detrusor overactivity)	
Oxybutynin chloride	5 mg po tid
Dicyclomine hydrochloride	20 mg po tid
Tricyclic antidepressants (detrusor overactivity)	
Imipramine	25 mg po qday to tid
Detrusor Botox injections	

Anhidrosis

Anhidrosis usually leads to dry skin, which can be treated with emollients. It may produce hyperthermia, which can be prevented by avoiding hot environments. If hyperpyrexia occurs, fans, cool drinks, and cool baths can be helpful.

Hyperhidrosis

Treatments for hyperhidrosis include anticholinergics, such as trihexyphenidyl and propantheline. However, their use is limited by anticholinergic side effects. Glycopyrrolate does not cross the blood–brain barrier and might have less side effects. Beta-blockers and benzodiazepines can be of some benefit. Localized hyperhidrosis may be treated with topical anticholinergic creams (hyoscine hydrobromide or glycopyrrolate), tap water iontophoresis, or Botox injections. Surgical treatment includes percutaneous endoscopic transthoracic sympathectomy, with ablation of prevertebral sympathetic ganglia from T2 to T4.

See also: Autoimmune Autonomic Neuropathy; Autonomic Failure; Autonomic Nervous System: Neuroanatomy; Autonomic Nervous System: Central Control of the Gastrointestinal Tract; Autonomic Nervous System; Erectile Dysfunction; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Apostolidis AN and Fowler CJ (2003) Evaluation and treatment of autonomic disorders of the urogenital system. *Seminars in Neurology* 23: 443–452.

Chelimsky G and Chelimsky TC (2003) Evaluation and treatment of autonomic disorders of the gastrointestinal tract. *Seminars in Neurology* 23: 453–458.

Donofrio PD and Caress JB (2001) Autonomic disorders. *The Neurologist* 7: 220–233.

Freeman R (2003) Treatment of orthostatic hypotension. *Seminars in Neurology* 23: 435–442.

Freeman R (2005) Autonomic peripheral neuropathy. *Lancet* 365: 1259–1270.

Gibbons C and Freeman R (2004) The evaluation of small fiber function – Autonomic and quantitative sensory testing. *Neurology Clinics* 22: 683–702.

Goldstein DS, Robertson D, Esler M, Straus SE, and Eisenhofer G (2002) Dysautonomias: Clinical disorders of the autonomic nervous system. *Annals of Internal Medicine* 137: 753–763.

Hilz MJ and Dutsch M (2006) Quantitative studies of autonomic function. *Muscle and Nerve* 33: 6–20.

Low AL (1993) *Clinical Autonomic Disorders: Evaluation and Management*. Boston: Little, Brown.

Low PA (2003) Testing the autonomic system. *Seminars in Neurology* 23: 407–421.

Mansoor GA (2006) Orthostatic hypotension due to autonomic disorders in the hypertension clinic. *American Journal of Hypertension* 19: 319–326.

Mathias CJ (2003) Autonomic diseases: Clinical features and laboratory evaluation. *Journal of Neurology, Neurosurgery and Psychiatry* 74: 31–41.

Mathias CJ (2003) Autonomic diseases: Management. *Journal of Neurology, Neurosurgery and Psychiatry* 74: 42–47.

Mathias CJ (2004) Role of autonomic evaluation in the diagnosis and management of syncope. *Clinical Autonomic Research* 14: 45–54.

McLeod JG and Tuck RR (1987) Disorders of the autonomic nervous system: Part I. Pathophysiology and clinical features. *Annals of Neurology* 21: 419–430.

Autonomic Dysfunction: Drug-Induced

A L Tonkin, University of Adelaide, Adelaide, SA, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Many drugs and other compounds are capable of causing autonomic dysfunction by a variety of mechanisms. Acute exposure may cause autonomic overactivity in some cases, but most drug-induced autonomic dysfunction results in reduced autonomic activity. This article focuses primarily on drugs and other compounds that reduce autonomic activity, either directly by transiently inhibiting autonomic nerve function or indirectly by causing autonomic neuropathy. It concludes with a discussion of compounds that can cause an abnormal transient increase in autonomic activity.

Direct Inhibition of Autonomic Activity

Drugs that reduce autonomic activity are very widely used in clinical medicine, particularly in management of cardiovascular disorders but also in areas such as gastroenterology and urology. Many different sites of action have been targeted, including central autonomic neurons, autonomic ganglia, postganglionic neurons (neurotransmitter synthesis, storage, and release), and receptors on effector cells. Both the sympathetic and parasympathetic systems have provided fertile ground for the development of inhibitory drugs for a variety of therapeutic purposes. These drugs are described in detail in pharmacological textbooks and only a brief summary is provided here.

Drugs That Reduce Both Sympathetic and Parasympathetic Activity

Drugs acting on preganglionic nerve fibers and autonomic ganglia Since both sympathetic and parasympathetic preganglionic cells utilize acetylcholine as their neurotransmitter, drugs that act on preganglionic cholinergic neurons or nicotinic receptors in autonomic ganglia have very wide-ranging autonomic effects and are usually poorly tolerated. Some of these, such as hexamethonium, have been used in the past as antihypertensives, but with very poor tolerability records. Because of their lack of specificity, none are used as first-line therapeutic agents currently.

Drugs That Reduce Sympathetic Activity

Antagonists of the sympathetic nervous system may act at central autonomic pathways, and/or peripheral α- or β-adrenoceptors. Many of these have been clinically important at various times in the management of hypertension. In general, their effects on autonomic function are well known and underpin their therapeutic efficacy.

Centrally acting drugs Inhibitors of central sympathetic outflow have been used in the management of hypertension for several decades. These drugs include α-methyldopa, clonidine, and, more recently, moxonidine. They act as α_2-receptor agonists, reducing sympathetic outflow from the central nervous system (CNS) by stimulating presynaptic α_2-receptors that control the release of noradrenaline (NA). Clonidine and moxonidine also bind to imidazoline receptors, which may mediate some of the effects of these drugs, although this is a controversial area.

Many other centrally acting drugs have autonomic effects, the pathogenesis of which remains poorly understood. Levodopa and dopamine D_2 receptor agonists, such as bromocriptine and pergolide, can cause severe orthostatic hypotension, particularly in the early phase of therapy of Parkinson's disease, which itself is associated with abnormal cardiovascular autonomic control. Combination anti-parkinsonian therapy, such as L-dopa with selegiline or bromocriptine, can cause more severe autonomic dysfunction than single drug treatment. The effects of L-dopa and the D_2 agonists are presumed to be related directly to an agonist effect at presynaptic D_2 receptors, resulting in reduced sympathetic outflow from the vasomotor centers of the brain stem. Other centrally acting drugs with similar clinical effects include the centrally acting anorectic drugs such as dexfenfluramine and the nonspecific monoamine oxidase inhibitors. Selegiline, an inhibitor of monoamine oxidase B (MAO-B), is metabolized to amphetamine compounds which deplete NA from sympathetic terminals, and selegiline can sometimes cause orthostatic hypotension acutely.

Peripherally acting drugs The vast majority of clinically useful sympathetic antagonists act on postsynaptic receptors. Three different major types of receptors are involved in mediating sympathetic activity: α- and β-adrenoceptors that respond to NA and adrenaline, and muscarinic receptors present in sweat glands which, uniquely, are innervated by cholinergic postganglionic neurons that are part of the sympathetic nervous system. Muscarinic receptors will be considered further in the section on the parasympathetic division of the autonomic nervous system.

Drugs can reduce sympathetic activity relatively selectively by antagonizing the effects of NA at

Table 1 Drugs antagonizing peripheral adrenoceptors

Receptor type	*Example antagonist*	*Common indications*
α_1-Receptors and α_2-receptors	Phenoxybenzamine (nonselective, irreversible)	Pheochromocytoma
	Phentolamine (nonselective, reversible)	
α_1-Receptors	Prazosin (selective, reversible)	Hypertension, bladder outflow obstruction
	Chlorpromazine	Psychosis
	Prochlorperazine	Nausea and vomiting
	Amitriptyline	Depression
α_2-Receptors	Yohimbine	Not used
β_1-Receptors and β_2-receptors	Propranolol (nonselective)	Migraine prophylaxis, essential tremor
β_1-Receptors	Metoprolol (selective)	Angina, hypertension
β_2-Receptors	No selective agents available	

peripheral receptors. Table 1 shows examples of such drugs, with their common indications for use. Many of these inhibitors of sympathetic transmission are used with the specific intention of reducing sympathetic activity to achieve therapeutic outcomes, such as reduction in blood pressure or heart rate, or relief of bladder obstruction. However, in practice, the drugs that most commonly lead to clinical presentations suggesting autonomic dysfunction are those that used for other effects, and they inhibit sympathetic activity as an adverse effect. Symptoms are particularly likely to occur when such drugs are used in the elderly. Some important examples are discussed here.

The clinical effects of sympathetic inhibition depend upon the balance of α- and β-adrenoceptors affected by the particular drug. Drugs with peripheral α-antagonist properties (such as the phenothiazines and tricyclic antidepressants) commonly cause orthostatic hypotension, particularly in elderly individuals. These drugs can also cause exacerbation of urinary incontinence, particularly in elderly women. This effect may have significant consequences, including institutionalization, if the causative agent is not identified and removed.

β-Blocker drugs are well known to have negative inotropic effects, and may sometimes lead to significant exacerbations of cardiac failure. In addition, they have negative chronotropic effects, a particular risk when administered in conjunction with other drugs that slow atrioventricular nodal conduction, such as digoxin or verapamil. Such a combination may cause clinically significant atrioventricular conduction delay, even in individuals with normal baseline electrocardiograms (ECGs). On the other hand, metoprolol and other β-blockers improve long-term outcome in chronic cardiac failure, and many patients with heart failure can tolerate β-blockade if the drug is introduced at a low dose and up-titrated slowly. Exacerbation of symptoms of obstructive airways disease is also well recognized as a complication of the use of β-blockers in susceptible people, although cardioselective β-blockers can be used safely in some patients with mild to moderate reactive airways disease.

Local administration of drugs with autonomic effects may also cause systemic symptoms. A well-recognized example of this phenomenon is the local instillation of timolol to the eye for the treatment of glaucoma, with subsequent exacerbation of bronchospasm or cardiac failure due to its systemic β-adrenoceptor antagonist effect.

Drugs That Reduce Parasympathetic Activity

Reduced parasympathetic activity occurs frequently in response to drug therapy. In some cases the drug is specifically intended to inhibit the parasympathetic system, but many drugs commonly used for other purposes also have peripheral antimuscarinic effects. Common examples are listed in Table 2.

Anticholinergic drugs are sometimes given with the aim of producing muscarinic antagonism. A common example is the use of benztropine in the management of parkinsonian symptoms, either due to Parkinson's disease or the movement disorders associated with antipsychotic use. Recently there has been an increase in the use of muscarinic receptor antagonists, such as oxybutynin and tolterodine, in the management of urinary incontinence due to detrusor instability. This condition is found most commonly in elderly people, who are particularly sensitive to antimuscarinic adverse effects such as dry mouth, sleep disruption, confusion, memory impairment, constipation, nausea, dizziness, and headache. The advent of more selective antagonists with less capacity to cross the blood–brain barrier may improve tolerability.

Protriptyline and amitriptyline are the most potent antimuscarinic receptor antagonists among the tricyclic antidepressants, and should be considered as possible contributors to presentations such as urinary retention or constipation. The newer antidepressant drugs, such as the selective reuptake inhibitors

Table 2 Drugs antagonizing peripheral muscarinic receptors

Example antagonist drug/drug group	*Usual indication*
Antimuscarinics	Prevention of extrapyramidal motor symptoms
	Bladder instability; nocturnal enuresis or urge incontinence
Phenothiazines	Psychosis; sedation in acute delirium
Tricyclic antidepressants	Depression
Antihistamines	Atopy, allergic reactions
Quinidine, disopyramide	Ventricular arrhythmias
Antispasmodics	Gastrointestinal motility disorders

fluoxetine and venlafaxine, are much less potent antagonists at muscarinic receptors. Class I antiarrhythmic agents such as quinidine and disopyramide also have significant antimuscarinic effects.

In all of these cases the drug is administered systemically and can be expected to have widespread autonomic effects. These include blurred vision, which is more symptomatic in young people with preserved lens accommodation, as well as dry mouth and eyes, severe constipation, and urinary retention, particularly in elderly men with prostatic hypertrophy. At very high doses, gastric emptying and gastric secretion may also be inhibited, leading to epigastric discomfort. The concurrent use of more than one anticholinergic drug is most common in the elderly and any change in CNS functioning in particular, which may range from a subtle impairment of memory to an acute anticholinergic-induced delirium, should be a stimulus to a careful review of drug therapy. At higher doses, anticholinergic drugs can precipitate frankly psychotic reactions with hallucinations.

Topical or local use of drugs with inhibitory effects on the parasympathetic nervous system can also lead to significant symptoms. The local use of botulinum toxin for conditions such as cervical dystonia or hyperhidrosis has been associated with a significant risk of dose-dependent anticholinergic adverse effects such as dryness of the mouth, accommodation difficulties, reduced sweating, swallowing difficulties, constipation, and urinary retention.

Other Compounds That Directly Reduce Autonomic Activity

A wide variety of naturally occurring and synthetic compounds also influence autonomic function.

Nicotine Nicotine has a variety of pharmacological actions and may be a source of toxicity in tobacco smokers. Nicotine binds to nicotinic receptors in autonomic ganglia, as well as at neuromuscular junctions and in the CNS. At usual doses associated with smoking there is autonomic stimulation, with effects including increased heart rate, blood pressure, cutaneous vasoconstriction, and reduced baroreflex sensitivity. Exposure to toxic doses (e.g., in tobacco workers or those exposed to nicotine-containing pesticides) can result in ganglionic paralysis, leading to reduction in sympathetic, parasympathetic, and neuromuscular transmission, with bradycardia, hypotension, coma, and eventual respiratory muscle paralysis. The role of nicotine in the pathogenesis of the known toxic effects of long-term tobacco smoking is unclear, but it is possible that autonomic dysregulation may contribute to the well-documented susceptibility of smokers to a variety of cardiovascular disorders.

Anticholinesterase compounds Compounds with anticholinesterase activity, such as organophosphorus esters and carbamate compounds, are commonly used in agriculture as insecticides. Acute ingestion by humans causes inhibition of the enzyme that removes acetylcholine from synaptic clefts, thus increasing neurotransmission at autonomic ganglia and parasympathetic terminals (discussed later). In some circumstances, however, these compounds may reduce autonomic function. Such a situation is observed after acute exposure to sarin gas, a highly toxic organophosphorus compound originally developed during World War II and more recently used in terrorist attacks. The autonomic effects of sarin are long-lasting (up to several weeks) in some individuals, and observations on workers involved in handling similar compounds during World War II have indicated that autonomic dysfunction may persist for years after acute exposure.

Herbal remedies Many freely available proprietary herbal remedies contain alkaloids with anticholinergic activity, such as atropine and scopolamine, and there have been case reports of anticholinergic poisoning resulting in confusion, dry mouth, tachycardia, and dilated pupils.

Illicit drugs Acute suppression of vasomotor outflow may occur following ingestion of amphetamines, with effects that include pronounced drowsiness and severe orthostatic hypotension, possibly due to NA depletion, with preservation of normal vagal function. The chronic use of amphetamine derivatives has been shown to cause significant impairment of parasympathetic cardiovascular control, still detectable about a week after last ingestion of the drug.

Contamination of illicit drugs may also lead to acute autonomic dysfunction. A number of cases of acute anticholinergic poisoning have been reported in individuals who had self-administered material sold as street 'heroin,' but which had been mixed with scopolamine. The clinical picture of hallucinations, tachycardia, pupillary dilatation, dry skin and mucous membranes, and urinary retention responds rapidly to treatment with the anticholinesterase compound, physostigmine.

Venoms Much progress has been made over recent years in understanding the molecular structure and mechanisms of action of the toxic components of venoms of various types. Many have acute inhibitory effects on autonomic activity. For example, some snake toxins bind to muscarinic receptors, leading to parasympathetic inhibition; others bind to ganglionic nicotinic receptors, leading to reduced sympathetic and parasympathetic transmission, and others bind to neuromuscular nicotinic receptors, causing flaccid paralysis.

Components of venoms from snakes and other venomous creatures, such as scorpions, are also known to interact with monoaminergic neurotransmission at multiple sites, including synthesis, release, or reuptake of monoamines, or agonist or antagonist activity at postsynaptic receptors. Venoms of cone snails and green mamba snakes, for example, contain small peptides that are antagonists at α-adrenoceptors. Many other neurotoxins have stimulatory effects and will be discussed later.

Autonomic Neuropathy Leading to Impaired Autonomic Function

Drug- or toxin-induced autonomic neuropathy is an important differential diagnosis in the assessment of patients with evidence of subacute or chronic autonomic dysfunction. Many neurotoxic compounds have selective effects on particular components of the nervous system, and not all compounds that cause peripheral neuropathy affect the autonomic system. Examples are arsenic, which causes a primarily sensory neuropathy without autonomic dysfunction, and lead, which commonly causes motor neuropathy and CNS effects (encephalopathy, behavioral changes) with little, if any, effect on autonomic function.

In specific clinical situations, damage to the autonomic nervous system is induced for therapeutic purposes by introduction of neurolytic agents into specific regions. Agents such as phenol are used for this purpose, and the resulting autonomic dysfunction includes the intended effect (e.g., in sympathectomy for peripheral vascular disease or hyperhidrosis) in addition to any unintended sequelae. More commonly, autonomic neuropathy is an unintended complication of exposure to drugs or chemicals in a therapeutic or occupational context.

Cytotoxic Agents Used in Cancer Chemotherapy

Vinca alkaloids Vincristine and vinblastine are used in the chemotherapy of hematological malignancies such as lymphomas and leukemias. Vincristine, and to a lesser extent, vinblastine, frequently cause a dose-dependent peripheral sensory neuropathy due to axonal degeneration, which is common at cumulative doses over 40 mg. Motor symptoms are rare but autonomic involvement occurs in more than a third of patients, usually manifest as orthostatic hypotension, constipation, paralytic ileus, erectile dysfunction, and urinary retention. Most of the neurological symptoms are reversible after months or years but in some cases are irreversible.

Platinum compounds The platinum compounds, including cisplatin, carboplatin, and oxaliplatin, react with DNA and cause apoptotic cell death in rapidly dividing cells, such as malignant cells. Platinum compounds frequently cause a predominantly sensory peripheral neuropathy associated with axonal degeneration and secondary atrophy of the dorsal root, but autonomic involvement is uncommon, probably affecting less than 10% of recipients of cisplatin.

Taxanes Paclitaxel and docetaxel, chemotherapeutic agents with activity against breast and ovarian cancer, commonly cause a peripheral sensorimotor neuropathy, probably by preventing the normal dissociation of axonal microtubles involved in fast axonal transport, although the precise mechanism of neurotoxicity remains unknown. Autonomic involvement is relatively uncommon, but paclitaxel has been reported to cause acute severe orthostatic hypotension associated with both parasympathetic and sympathetic dysfunction. Paralytic ileus and orthostatic hypotension are the most common manifestations of paclitaxel-induced autonomic neuropathy, and patients with diabetes mellitus may be more susceptible to this complication.

Other Therapeutic Drugs

Perhexiline maleate Perhexiline, used as a last-line agent in the management of refractory angina, is known to cause a peripheral sensorimotor neuropathy. In a few of these cases the neurotoxicity extends to the autonomic nervous system, causing postural hypotension and abnormal heart rate control.

Amiodarone Chronic treatment of cardiac arrhythmias with amiodarone causes peripheral neuropathy (predominantly sensory) in a small proportion of patients. In some of these there may also be autonomic dysfunction, manifest as orthostatic hypotension.

Pentamidine Pentamidine, commonly used in the management of *Pneumocystis carinii* pneumonia secondary to AIDS, has been associated with acute autonomic insufficiency manifested as severe orthostatic hypotension with a fixed heart rate and loss of baroreflex function.

Occupational Exposure

Organic solvents Parasympathetic function appears to be disturbed in some workers exposed to a variety of organic solvents, including hydrocarbons, alcohols, ketones, esters, and ethers, although long-term exposure to toluene has not been associated with peripheral or autonomic neuropathy. Long-term exposure (over years to decades) to low levels of carbon disulfide, for example in the viscose/rayon industry, may cause mild parasympathetic dysfunction, but not sufficient to cause clinical effects.

Acrylamide Acrylamide, which is widely used in paper manufacture, water treatment, building construction, and laboratory research, is known to cause a neuropathy affecting somatic and autonomic nerves. Clinically detectable autonomic disturbances are uncommon in humans but have been studied extensively in experimental animals, in which damage to sympathetic vasomotor nerves as well as parasympathetic fibers has been demonstrated.

Heavy metals Chronic exposure to lead is known to cause a peripheral neuropathy with axonal degeneration, particularly affecting motor fibers. Very long-term exposure may also result in mild sensory and autonomic polyneuropathy, and similar effects, mainly involving the parasympathetic system, have also been reported following mixed exposure to lead, zinc, and copper.

Other Forms of Exposure

Alcohol Alcohol ingestion has both acute and chronic effects on autonomic function in humans. Acute exposure to alcohol reduces parasympathetic function, possibly due to alterations in central cardiovascular control centers rather than a direct effect on vagal function. Chronic alcohol abuse causes an axonal neuropathy, affecting both somatic (sensory and motor) and autonomic function. About 30% of chronic alcohol abusers have detectable abnormalities in the parasympathetic control of heart rate and, in others, sympathetic function may also become affected, resulting in orthostatic hypotension and anhidrosis. Long-term follow-up of chronic alcohol abusers suggests that the presence of autonomic abnormalities on clinical testing is associated with an increased mortality, possibly related to cardiac dysregulation.

Pesticides Most insecticides in common use in agricultural, industrial, or domestic settings are neurotoxic to the target organisms. In general they are not species selective, and can also affect mammalian nervous systems, the outcome depending on the level of exposure in relation to body size. Organochlorine insecticides (such as DDT) affect sensory and motor nerves with few, if any, autonomic effects after acute or chronic exposure.

Organophosphorus compounds (discussed earlier in relation to their acute anticholinesterase activity) may also cause various subacute or chronic neuropathies following a single high-intensity exposure, as in the case of the use of organophosphorus-based chemical weapons in warfare, or during chronic exposure – for example, in manufacturing processes. Long-term chronic low-level exposure to organophosphorus ester insecticides may produce persistent autonomic dysfunction. Organophosphorus compounds bind to acetylcholinesterase for a much longer period than do the more recently developed carbamate compounds, and there is little evidence of prolonged toxicity with the latter group.

Botulinum toxin Although rare, botulism due to ingestion of contaminated food is an important differential diagnosis for the presentation of acute autonomic neuropathy, particularly when there is predominant parasympathetic dysfunction. The toxin acts on cholinergic synapses to prevent the calcium-mediated release of acetylcholine from the presynaptic terminal. Its effects are seen primarily at neuromuscular junctions, causing paralysis, and, to a lesser extent, at autonomic ganglia, where the effect is to block autonomic transmission acutely.

Compounds That Increase Autonomic Activity

Several therapeutic drugs are used to produce increased autonomic activity in situations such as asthma, heart conduction disorders, and hypotension. Excessive autonomic activity is frequently seen in individuals with denervation hypersensitivity (e.g., due to preexisting autonomic neuropathy) during treatment with autonomic agonists. Examples include the use of sympathetic α-adrenoceptor agonists in orthostatic

hypotension, and the parasympathomimetic agent, bethanechol, in the management of bladder atony. In other situations, a toxic compound may cause abnormally increased autonomic activity in an individual with previously normal autonomic function.

Sympathetic Stimulation

Drugs used in psychiatry Acute hyperactivity of the sympathetic division of the autonomic nervous system is seen occasionally as a consequence of exposure to psychoactive drugs, both therapeutic and illicit. Neuroleptic malignant syndrome is thought to be mediated by changes in central dopaminergic transmission, causing increased sympathetic outflow manifested as hyperthermia, muscle rigidity, and instability of blood pressure and heart rate. It occurs in fewer than 1% of patients receiving antipsychotic drugs, but can be fatal if untreated. A similar clinical picture, with occasional fatalities, occurs in some individuals who ingest the so-called designer drug Ecstasy (3,4-methylenedioxymethamphetamine (MDMA)) and other amphetamine derivatives that stimulate the sympathetic system.

An interaction between antidepressant drugs, which inhibit NA and/or serotonin reuptake from the synapse, and tyramine, which releases catecholamines from storage vesicles, can result in a marked increase in neurotransmitter concentrations within sympathetic synapses. Clinically there is marked sympathetic hyperactivity with sweating, tachycardia, and severe hypertension, which may lead to intracranial hemorrhage. A similar picture can be seen following the use of the combination of a monoamine oxidase inhibitor and a serotonin uptake inhibitor (such as fluoxetine). Known as the 'CNS serotonergic syndrome,' it comprises general CNS overactivity, muscle spasms, hyperthermia, and autonomic instability, resulting in hyper- or hypotension, tachycardia, and profuse sweating. Fatalities have been reported. The mechanism is believed to be an increase in serotonergic activity, particularly involving serotonin ($5\text{-}HT_{1A}$) receptors in the brain stem and spinal cord.

Ketamine The neurolept anesthetic/analgesic agent, ketamine, primarily an *N*-methyl-D-aspartate (NMDA) receptor antagonist, has an interesting combination of stimulatory and inhibitory cardiovascular and respiratory effects. In isolated organ experiments, in the absence of an intact autonomic nervous system, it has a direct negative inotropic effect on the myocardium. However, in the intact animal or human, this is counterbalanced by a central stimulatory effect, probably mediated by increased sympathetic outflow, which results in an elevation in blood pressure, heart rate and cardiac output, and bronchodilatation.

Venoms and toxins Although snake venom may have a variety of autonomic effects, it is very likely that much of the autonomic activation following snake bite is fear related, resulting in nausea, vomiting, diarrhea, syncope, tachycardia, and cold, clammy skin, even in the absence of actual envenomation. Some venoms, such as those of the stonefish, South American viper snake, and various spiders and scorpions, contain agonists for α- and β-adrenoceptors, which are likely to contribute to venom-induced hypertension and tachycardia. Some venomous creatures, including snakes and scorpions, cause prominent sympathetic hyperactivity, resulting in the potentially fatal clinical syndrome sometimes known as 'autonomic storm.' Components of venoms of scorpions, Australian funnel-web spiders, and the box jellyfish contain neurotoxins that induce massive release of NA, resulting in sympathetic hyperstimulation, which may cause severe hypertension, hypertensive encephalopathy, tachycardia, and cardiac arrhythmias.

Parasympathetic Stimulation

Therapeutic drugs Denervation hypersensitivity has been observed in association with autonomic neuropathy in patients given bethanechol, which produced a markedly exaggerated effect. Anticholinesterase treatment for Alzheimer's disease frequently leads to adverse effects attributable to increased cholinergic neurotransmission in the periphery, such as nausea, vomiting, and diarrhea. More rarely, more serious cholinergic effects, such as heart block, hallucinations, and confusion, may occur. The available drugs appear to be similar in their spectrum of adverse effects.

Other compounds Similar muscarinic stimulatory effects are seen after poisoning with wild mushroom species that contain neurotoxins with muscarinic agonist activity. The effects occur rapidly and, predictably, include lacrimation, salivation, nausea, vomiting, abdominal pain, bronchospasm, headache, miosis, blurred vision, bradycardia, and hypotension. A similar clinical picture is seen acutely following anticholinesterase poisoning with agricultural chemicals such as organophosphorus compounds, which inhibit acetylcholinesterase. The inhibition results in the typical picture of increased secretions, bronchoconstriction, miosis, abdominal cramps, and bradycardia (all related to muscarinic receptor stimulation), and hypertension, muscle fasciculations, tremor, and eventual muscle paralysis (due to nicotinic receptor stimulation at autonomic ganglia and at neuromuscular junctions).

Less common causes of this constellation of features include acute exposure to snake venom toxins known as fasciculins, which also have anticholinesterase activity, both at neuromuscular junctions, where

they produce a prolonged muscular contraction, and at autonomic ganglia, where they increase autonomic transmission. Some scorpion toxins have a similar effect, most commonly by binding to neuronal voltage-sensitive sodium and potassium channels and increasing the spontaneous release of acetylcholine, although some venom components have direct muscarinic agonist effects.

Conclusion

While the mechanisms of action of drugs which interact reversibly with autonomic nerve fibers or receptors as part of their pharmacodynamic activity are generally well known, more remains to be established about the effects of drugs and other compounds that cause autonomic neuropathy. Drugs, chemicals, and toxins should always be considered as a possible cause of autonomic dysfunction or autonomic neuropathy in patients presenting with the clinical manifestations of sympathetic and/or parasympathetic dysfunction.

See also: Autoimmune Autonomic Neuropathy; Autonomic Disorders; Autonomic Failure; Autonomic Nervous System; Parasympathetic Nervous System.

Further Reading

Bouhaddi M, Vuillier F, Fortrat JO, et al. (2004) Impaired cardiovascular autonomic control in newly and long-term-treated patients with Parkinson's disease: Involvement of L-dopa therapy. *Autonomic Neuroscience: Basic and Clinical* 116: 30–38.

Dressler D and Benecke R (2003) Autonomic side effects of botulinum toxin type B treatment of cervical dystonia and hyperhidrosis. *European Neurology* 49: 34–38.

Gerhardt U, Hans U, and Hohage H (1999) Influence of smoking on baroreceptor function: 24 hour measurements. *Journal of Hypertension* 17: 941–946.

Gold BS, Dart RC, and Barish RA (2002) Bites of venomous snakes. *New England Journal of Medicine* 347: 347–356.

Gouzoulis-Mayfrank E, Thelen B, Habermeyer E, et al. (1999) Psychopathological, neuroendocrine and autonomic effects of 3,4-methylene dioxyethylamphetamine (MDA), psilocybin and *d*-methamphetamine in healthy volunteers. *Psychopharmacology* 142: 41–50.

Gwee MCE, Nirthanan S, Khoo H-E, et al. (2002) Autonomic effects of some scorpion venoms and toxins. *Clinical and Experimental Pharmacology and Physiology* 29: 795–801.

Jamal GA, Hansen S, and Julu POO (2002) Low level exposures to organophosphorus esters may cause neurotoxicity. *Toxicology* 181–182: 23–33.

Jerian SM, Sarosy GA, Link CJ, et al. (1993) Incapacitating autonomic neuropathy precipitated by taxol. *Gynecologic Oncology* 51: 277–280.

Kay GG and Granville LJ (2005) Antimuscarinic agents: Implications and concerns in the management of overactive bladder in the elderly. *Clinical Therapeutics* 27: 127–138.

Ketch T, Biaggioni I, Robertson RM, et al. (2002) Four faces of baroreflex failure: Hypertensive crisis, volatile hypertension, orthostatic tachycardia, and malignant vagotonia. *Circulation* 105: 2518–2523.

Nicolosi C, DiLeo R, Girlanda P, et al. (2005) Is there a relationship between somatic and autonomic neuropathies in chronic alcoholics? *Journal of Neurological Sciences* 228: 15–19.

Quasthoff S and Hartung HP (2002) Chemotherapy-induced peripheral neuropathy. *Journal of Neurology* 249: 9–17.

Smit AAJ, Wieling W, Voogel AJ, et al. (1996) Orthostatic hypotension due to suppression of vasomotor outflow after amphetamine intoxication. *Mayo Clinic Proceedings* 71: 1067–1070.

Russell JW and Reading PH *Acute Autonomic Neuropathy* and *Chronic Autonomic Neuropathies* Available at: http://www.medlink.com/medlinkcontent.asp.

Tjeerdsma G, Szabo BM, van Wijk LM, et al. (2001) Autonomic dysfunction in patients with mild heart failure and coronary artery disease and the effects of add-on β-blockade. *European Journal of Heart Failure* 3: 33–39.

Tonkin AL and Frewin DB (1999) Drugs, chemicals and toxins that alter autonomic function. In: Mathias CJ and Bannister R (eds.) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System,* 4th edn. Oxford: Oxford University Press.

Autonomic Dysregulation During REM Sleep

S M Caples and V K Somers, Mayo Clinic, Rochester, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Sleep Stage Transitions and the Autonomic Nervous System in Normal Individuals

In humans, wakefulness normally transitions to sleep with the onset of non-rapid eye movement (NREM) sleep, which comprises the majority of total sleep time. Compared with wakefulness, NREM sleep is characterized by cardiovascular stability marked by a reduction in blood pressure (BP) and heart rate (HR). Slowing of the HR is primarily under the influence of vagal activation that is characteristic of NREM sleep. Because of the limited parasympathetic influences on the peripheral vasculature, the decrease in BP is probably a result of concurrent sympathetic withdrawal, as supported by animal models of sympatholysis. During NREM sleep, there is a prominent sinus arrhythmia as cardiac output couples with the respiratory rhythm. With inspiration, as venous return to the heart increases, there is an acceleration of HR followed by a deceleration with exhalation. BP is tightly controlled on account of high baroreceptor gain, which is particularly responsive to BP increments during NREM sleep, serving to ensure the maintenance of stable low BP, referred to as the 'dipping' phenomenon. The absence of this decrement in BP ('nondipping') has, in some studies, been associated with cardiovascular morbidity and mortality – outcomes perhaps mediated by autonomic dysregulation.

Progression of NREM sleep from stage 1 through 4 is characterized by increasingly synchronous cortical and brain stem electrical discharges as exhibited by low-frequency (LF), high-amplitude output on electroencephalography (EEG). This natural sequence is accompanied by further reduction in sympathetic neural traffic such that by stages 3 and 4 (slow wave), the output may be half that encountered in wakefulness. Stage 2 sleep is punctuated by characteristic 'K-complexes' – high-amplitude EEG discharges associated with transient increases in peripheral sympathetic neural activity and an attendant rise in BP (Figure 1). Bursts of vagal activity sometimes accompany the transition out of NREM sleep, resulting in HR pauses.

Sharply distinct from NREM sleep, rapid eye movement (REM) sleep occurs cyclically (four to six times) throughout sleep, is more concentrated during the second half of a sleep cycle, and accounts for approximately 25% of total sleep time. With the exception of the respiratory diaphragm and extraocular muscles, REM sleep is heralded by skeletal muscle atonia along with high-frequency (HF), low-amplitude waves on EEG. Because the EEG may otherwise resemble that of wakefulness, REM sleep has been referred to as 'paradoxical sleep.' REM sleep may be subdivided into tonic and phasic stages, the latter of which features characteristic darting eye movements in association with pontogeniculo-occipital EEG spikes, and muscle twitches, predominantly seen in the extremities and face, which break through skeletal muscle atonia.

As discussed later, REM sleep, particularly the phasic component, is considered a state of autonomic instability, marked by surges in both sympathetic and vagal activity with attendant variations in HR, BP, and peripheral vascular resistance. Animal and human studies have demonstrated more concentrated autonomic changes associated with phasic REM sleep, during which intense electrical discharges have been recorded from the brain stem. As a testament to the complexity of the origins of REM sleep and its associated autonomic findings, measurable physiologic parameters representative of autonomic fluctuations have been shown to precede, by several seconds, characteristic eye movements and an active, desynchronized EEG.

Dreaming

Dreaming is thought to occur primarily during REM sleep. Human experiments show that the majority of subjects awoken during REM sleep report a dream experience, compared with a minority reporting the same when awoken from NREM sleep. It has been postulated that the paradoxically active cerebral cortex associated with REM sleep relates to dreams, during which powerful emotions such as anger and fear are frequently generated. Since psychological stress has been associated with cardiac ischemia and rhythm disturbances during wakefulness, it may follow that similar vulnerability could occur during dreaming, as suggested by Mac William in 1923 and also by case reports of patients awakening from dreaming with angina pectoris. There is evidence that brain electrical discharges may directly impact cardiac activity and are perhaps mediated by autonomic tone. Early animal experiments, confirmed by Verrier, have shown that electrical stimulation of certain areas of the brain could induce ventricular arrhythmias and that β-adrenergic blockade, but not vagotomy, reduced vulnerability.

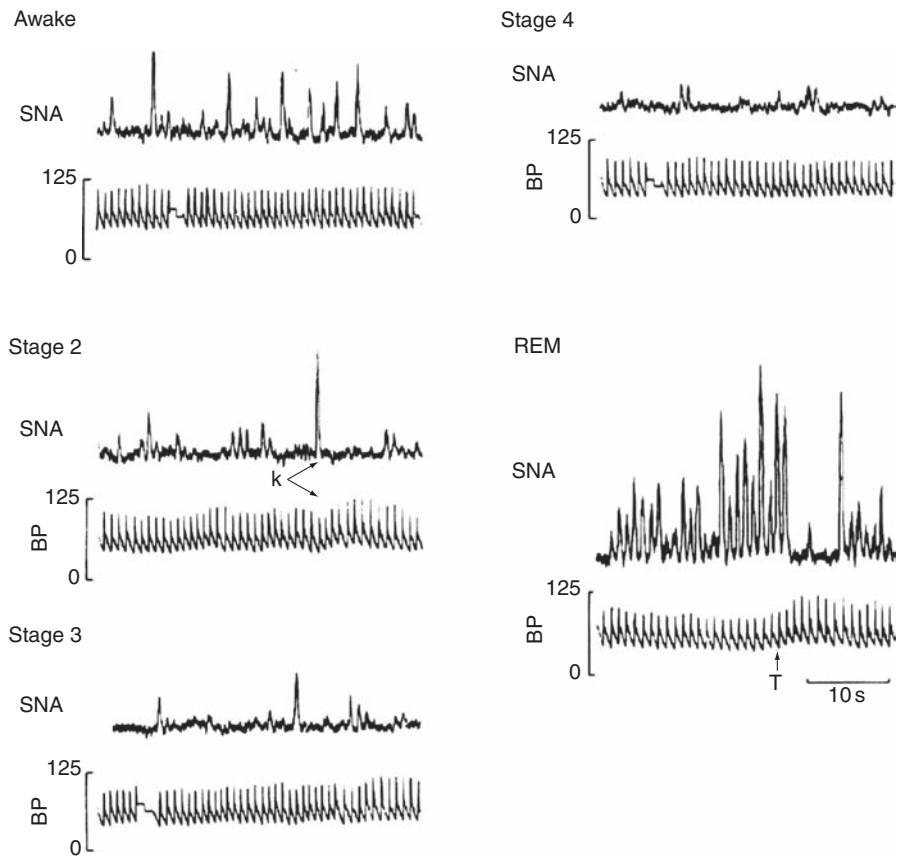

Figure 1 Recordings of sympathetic nerve activity (SNA) and mean blood pressure (BP) in a single subject while awake and while in stages 2, 3, 4, and REM sleep. Note the progressive reduction in SNA and BP as sleep progresses through the stages of NREM sleep. The 'K' during stage 2 corresponds to simultaneous K-complexes noted on the electroencephalogram. In contrast, REM sleep is marked by higher frequency and amplitude of SNA as well as variability in BP. The 'T' during REM sleep refers to a muscle twitch characteristic of phasic REM sleep, associated with abrupt inhibition of SNA and an increase in BP. Reproduced from Somers VK (1993) Sympathetic-nerve activity during sleep in normal subjects. *New England Journal of Medicine* 328(5): 303–307.

Other Physiologic Responses to REM Sleep

In addition to directly observed cardiovascular irregularities, REM sleep is associated with other physiologic perturbations which may indirectly contribute to autonomic dysfunction and, therefore, cardiovascular risk. The respiratory diaphragm usually escapes skeletal muscle atonia, but loss of tone in accessory muscles of respiration and those of the upper airway may pose an increased workload. The ventilatory pattern fluctuates during REM sleep. As a consequence, even though respiratory rate may increase during REM sleep, tidal volumes and minute ventilation usually decrease compared with NREM sleep. There are blunted ventilatory responses to hypoxia and hypercapnia and a heightened arousal threshold, all contributing to prolonged reductions in blood oxygen content, particular in those with impaired respiratory pump function (neuromuscular weakness) or gas exchange capabilities (chronic obstructive lung disease). Hypoxemia stimulates the chemoreflex, resulting in further sympathetic neural outflow. Increased upper airway resistance related to muscle atonia is operative in the pathogenesis of obstructive sleep apnea (OSA), as detailed later. Partial loss of thermoregulatory control during REM sleep further disrupts homeostasis. Finally, cortical influences related to dream activity during REM sleep may contribute to the irregular breathing pattern, along with increased cortical and brain stem electrical discharges seen in association with dreaming.

Cardiovascular Reflexes Mediating Autonomic Activity

A number of reflexes feed back to the autonomic nervous system to influence cardiac and vascular function during periods of stress or physical exertion.

From the standpoint of sleep, these mechanisms may be particularly important during changes in stage of sleep (NREM to REM), position, or with the occurrence of arousals, all of which may take place repeatedly throughout a sleep cycle and are increasingly common with advancing age, a population vulnerable to cardiovascular disease. Dysregulation of these reflexes may be operative in the pathophysiology of certain disease states, such as sleep disordered breathing.

The Chemoreflexes

The peripheral arterial chemoreceptors, the most important of which are located in the carotid bodies, respond primarily to changes in the blood partial pressure of oxygen. Hypoxemic stimulation elicits a brain stem-mediated increase in respiratory muscle output. By virtue of connections between the carotid bodies and sympathetic ganglia, there is an increase in sympathetic outflow to peripheral blood vessels. Homeostasis is preserved by within-breath activation of pulmonary stretch receptors, which are mediated through the vagus nerve, restraining the overall adrenergic response. This sequence is often part of a normal physiologic response under conditions of hypoxia, such as at high altitude. However, an exaggerated ventilatory response, with attendant further increases in sympathetic tone, has been found in certain disease states, such as heart failure and OSA.

The Arterial Baroreflex

The arterial baroreflex system, composed of sensory receptors in the aortic arch and carotid sinuses that relay signals to the brain stem, provides powerful beat-by-beat negative feedback regulation of arterial BP. Increases in BP stretch the receptors and dampen efferent sympathetic output to cardiac and vascular smooth muscle, thereby relatively increasing cardiac parasympathetic tone. The acute increases in BP associated with peripheral vasoconstriction are partially tempered by the baroreflex. This may occur simultaneously with activation of cardiac vagal drive resulting in bradycardia, collectively referred to as the 'diving reflex,' so-called because of its detailed description in diving mammals and also described in humans. Baroreceptor sensitivity appears to be heightened during NREM sleep, but findings in very small studies that measure REM sleep are inconsistent, with some showing a blunting of baroreceptor sensitivity, no change, or a time-dependent increase in sensitivity. Influences of comorbid illnesses, such as hypertension or heart failure, on baroreceptor function are well documented.

Measures of Autonomic Cardiovascular Regulation

Although there are a number of methods by which to measure autonomic neural output, there have been, with some exceptions, concordant findings of parasympathetic predominance during NREM sleep and sympathetic surges with variable states of parasympathetic tone during REM sleep. A commonly utilized method, heart rate variability (HRV), is defined as the oscillation in the interval between consecutive heartbeats (the R-R interval) as well as the oscillations between consecutive instantaneous heart rates. During NREM sleep, a near sinusoidal modulation of HR variation occurs due to the normally occurring respiratory sinus arrhythmia. Spectral analysis allows quantification of the short-term oscillatory components of cardiovascular variability, which are organized in two frequency bands – the LF (approximately 0.1 Hz) and the HF (>0.15 Hz) respiratory bands. The HF components of R-R variability primarily reflect vagal modulation of cardiac rhythm, whereas the LF rhythm has been attributed, albeit not without controversy, to cardiac sympathetic modulation.

Utilizing HRV analysis, a number of studies have demonstrated increased parasympathetic tone and sympathetic withdrawal with the transition from wakefulness to NREM sleep. These studies also demonstrate evidence for increased sympathetic output with the onset of REM sleep, although not all studies have shown reductions in parasympathetic neural output during REM sleep.

Invasive and technically demanding, microneurographic recording of sympathetic nerve activity (MSNA) involves the direct recording of postganglionic sympathetic nerve traffic which may originate from muscle and skin. MSNA has yielded fairly concordant sleep stage-dependent findings. Peripheral sympathetic neural activity, HR, and BP all decrease in concert from sleep onset and progression of NREM sleep to deep, synchronized slow-wave sleep, such that by stage 4, sympathetic measured output was half that encountered in wakefulness. REM sleep, on the other hand, has been associated with SNA exceeding that encountered during wakefulness, particularly during clusters of eye movements (Figure 1). Muscle twitches that break through REM atonia have been associated with abrupt attenuation of sympathetic output, probably due to baroreceptor-mediated inhibition during transient rises in BP. That REM sleep is associated with minimal changes in HR and BP may relate to the dissociation of peripheral muscle and cardiac sympathetic activity and underscores the complexities of cardiac sympathovagal balance

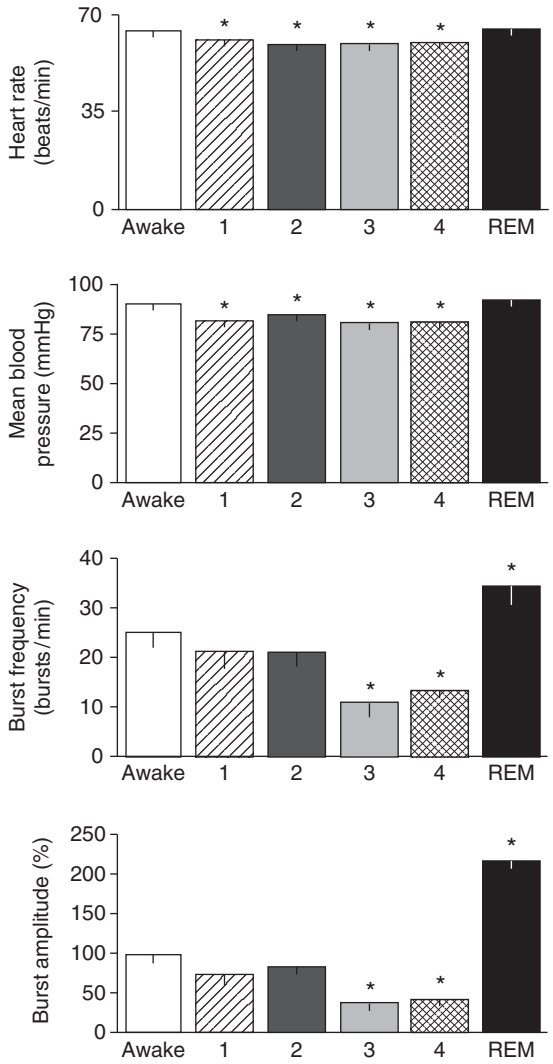

Figure 2 Physiologic indices in a group of normal subjects during wakefulness and NREM and REM sleep. Burst frequency and amplitude refer to sympathetic nerve activity (SNA) measured by intraneural microneurography. Heart rate (HR) and blood pressure (BP) were significantly lower during all stages of NREM sleep than during wakefulness, and SNA was significantly lower during slow-wave sleep (asterisk denotes $p < 0.001$). During REM sleep, sympathetic activity increased significantly, but BP and HR were similar to those recorded during wakefulness. Reproduced from Somers VK (1993) Sympathetic-nerve activity during sleep in normal subjects. *New England Journal of Medicine* 328(5): 303–307.

(Figure 2). Thus, this technique provides insight into regional sleep stage-dependent autonomic mechanisms, although it may not necessarily reflect global sympathetic tone.

Other methods for measuring autonomic neural activity, all with their own limitations, include arterial baroreflex function, plasma and urine catecholamine levels, peripheral arterial tonometry, and pulse transit time.

REM Sleep and Heart Rate/Rhythm

Animal experiments have demonstrated marked HR instability during REM sleep, where increases of more than 30% are seen in association with phasic REM sleep, accompanied by an elevation in BP. The frequency of HR surges is higher during phasic REM than tonic REM sleep, with particular concentration during sleep epochs in which an associated muscle twitch was recorded. In animals with coronary artery stenosis, REM sleep causes further decrements in coronary artery blood flow. That these findings were prevented with sympathectomy in cats and dogs suggests an effect attributable to sympathetic overdrive rather than parasympathetic withdrawal. In rats, REM, particularly during phasic episodes, is associated with a higher incidence of bradycardia.

In humans, the average HR is higher during REM than NREM sleep, as is sympathetic neural output, although there is evidence that sympathetic activation is not a global phenomenon during REM sleep. Case reports and series suggest that healthy young adults, sometimes athletes, have periods of exaggerated vagal tone during REM sleep, manifesting as first- or second-degree atrioventricular (AV) block or sinus pauses that may exceed 9 s in duration. Further insight comes from feline experiments demonstrating the highest occurrence of these pauses during the transition from slow-wave sleep to REM sleep as well as during phasic REM sleep. Reversal of the pauses has been reported with atropine or vagotomy. REM sleep therefore appears to be marked by fluctuations in autonomic activity, both to the cardiac conduction system and to peripheral blood vessels.

Patients with established cardiac conduction disease may have distinct pathophysiologic mechanisms during REM sleep, suggested by a small group of subjects with preexisting AV block demonstrating shortening of AV delay during periods of REM sleep. Ambulatory cardiac monitoring (non-polysomnographic) data in healthy elderly populations demonstrate less common occurrences of sinus pauses and AV blocks during sleep episodes, with relatively common supraventricular and ventricular ectopic beats. The circadian variation in paroxysmal atrial fibrillation noted by some suggests a potential influence of the autonomic nervous system. Analysis of HRV in patients with atrial fibrillation demonstrated abrupt shifts in sympathovagal balance preceding the onset of the arrhythmia. Finally, gender-selective effects of REM sleep on cardiac electrical activity were demonstrated by Lanfranchi et al., who measured prolongation of the QT interval exclusive to women.

Arousals/Morning Awakening

Arousal from sleep, either spontaneous or induced by exogenous stimuli, is associated with sympathetic neural surges, leading to increases in HR and BP. Sympathetic tone, however, remains elevated long after return of HR and BP to prearousal levels. Nonspecific arousals and sleep fragmentation increase with advancing age, where the presence of comorbid cardiovascular disease is prevalent and provides a milieu potentially more vulnerable to the effects of sympathoexcitation. Morning awakening induces a stepwise activation of the sympathoadrenal system, with increased HR, BP, and blood catecholamines, and further increases occur with position change and physical activity. Population-based studies have shown that the risk of sudden death from cardiac causes in the general population is significantly greater during the morning hours after waking (i.e., from 6 a.m. to noon) than during other times of the day or night. Because REM is concentrated in the second half of a night's sleep, it is conceivable that autonomic dysregulation associated with REM sleep, among other mechanisms, may be an important contributor, particularly in those with a vulnerable predisposition, such as in the setting of ischemic heart disease. OSA, perhaps due to autonomic dysregulation, has been associated with disruption of this day–night pattern of sudden death.

Obstructive Sleep Apnea

The hallmark of OSA is repetitive upper airway collapse resulting in oxyhemoglobin desaturation. Central nervous system arousals intervene with termination of each breathing event, followed by compensatory hyperventilation. As a result of upper airway/pharyngeal muscle atony as well as reduced lung functional residual capacity, REM sleep usually manifests more upper airway instability and profound deoxygenation than that which occurs during NREM sleep. Further contributing to prolonged hypoxia is the high arousal threshold characteristic of REM sleep. These multiple physiologic stressors all contribute to acute and repetitive surges in sympathetic neural output, the detection of which by measurement of peripheral arterial tone has been used as a diagnostic tool. Some studies report augmented sympathetic responses when central nervous system arousals are accompanied by hypoxia, hypercapnia, and upper airway occlusion. Further evidence for autonomic imbalance in patients with OSA is the finding that the usual decrements in sympathetic activity seen during NREM sleep are lost.

It is possible that there are cumulative effects of long-standing repetitive upper airway events that lead to disruption of autonomic homeostatic mechanisms. In fact, there is abundant evidence for autonomic dysregulation in OSA, in which high levels of MSNA during daytime normoxic wakefulness have been measured (Figure 3). HRV is altered in OSA, independent of comorbid illnesses including obesity. Sleep apneics have abnormal baroreceptor function and exaggerated peripheral and central chemoreflex responses, resulting in additive sympathetic stimulation. Deficits in compensatory neural responses to upper airway collapse have been detected in humans with OSA.

Treatment of OSA, typically with the application of continuous positive airway pressure (CPAP) which prevents upper airway collapse, has been shown to acutely attenuate MSNA. Longer term treatment with CPAP reduces daytime sympathetic neural traffic and may also restore, to some degree, autonomic regulatory mechanisms, including baroreceptor function and HRV. Such effects may mediate the salutary effects of CPAP on cardiovascular diseases commonly associated with OSA, such as hypertension and heart failure. Sympathetic neural output may be particularly important in the pathophysiology of heart failure, with OSA having an additive effect and CPAP treatment resulting in acute reductions in sympathetic activity that remain measurable after 3 months of therapy. The long-term implications of OSA treatment on HF outcomes are unknown.

Sleep Deprivation/Short Sleep Duration

National surveys have reported increasingly shorter sleep time per night; as such, there has been intense interest in the systemic effects of short sleep duration. In tightly controlled human experiments, acute restriction of sleep time has been shown to result in metabolic dysregulation, a finding which, in part, could have a physiologic basis in autonomic control. Large population-based studies suggest a higher risk of hypertension in those who self-report shorter sleep duration. Evidence for mediation through autonomic tone has come from experiments demonstrating REM rebound following acute sleep deprivation in young adults and associated increased sympathetic tone by HRV analysis. The blunting of the reduced ventilatory response to hypoxia and hypercapnia associated with sleep deprivation could further contribute to autonomic imbalance.

REM Sleep and Autonomics in the Setting of Cardiovascular Disease

Whereas the acute surges in sympathetic activity associated with REM sleep may be well tolerated in

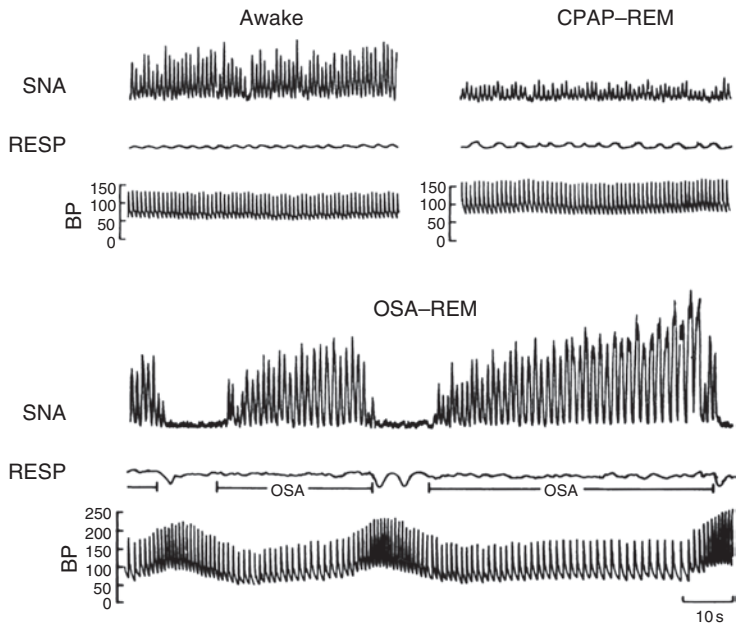

Figure 3 Recording of sympathetic nerve activity (SNA) by microneurography, respiratory excursion (RESP), and intra-arterial blood pressure (BP) in a subject with obstructive sleep apnea (OSA) when awake, during obstructive apneas while in REM sleep, and with elimination of obstructive apnea by continuous positive airway pressure (CPAP) therapy during REM sleep. SNA is high even during normoxic wakefulness, with further increases during REM sleep with obstructive apnea. Note increases in BP as apnea progresses and instantaneous inhibition of SNA with ventilation. Elimination of apneas by CPAP is associated with decreased SNA. Reproduced from Somers VK (1995) Sympathetic neural mechanisms in obstructive sleep apnea. *Journal of Clinical Investigation* 96(4): 1897–1904.

healthy individuals, those with preexisting cardiovascular disease may be vulnerable to adrenergic-mediated vasoconstriction and increments in HR and BP. Polysomnographic recordings of patients with nocturnal angina show ST segment changes on electrocardiography occurring with HR surges associated with REM sleep, as well as with changes in sleeping position that may accompany the transition between NREM and REM sleep. Some patients with diabetes and a recent history of myocardial infarction (MI) have abnormal HRV profiles consistent with higher nocturnal sympathetic tone or reduced vagal tone, which may help explain a higher rate of unfavorable cardiovascular outcomes in these patients. It should also be noted that NREM sleep, which may be associated with hypotension and reduced coronary perfusion pressure, may not be risk-free in susceptible individuals with coronary artery stenosis or after MI. The normal dipping of BP associated with sleep is thought to result from the dominance of parasympathetic tone during NREM sleep over the course of the night. The blunting or absence of dipping suggests a possible role of autonomic dysregulation and has been associated with a higher risk of stroke and congestive heart failure.

Conclusion

There are undoubtedly complex interactions between the autonomic nervous system and REM sleep. The study of cardiovascular function in health and of disease states such as OSA has furthered our understanding of such mechanisms. Additional research will be needed to better characterize these complexities to determine if targeted sleep interventions have any impact on clinical outcomes.

See also: Autonomic Nervous System; Endocrine Function During Sleep and Sleep Deprivation; Immune Function During Sleep and Sleep Deprivation; Metabolic Syndrome and Sleep; Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep Apnea; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function; The AIM Model of Dreaming, Sleeping, and Waking Consciousness; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Adlakha A (1998) Cardiac arrhythmias during normal sleep and in obstructive sleep apnea syndrome. *Sleep Medicine Reviews* 2(1): 45–60.

Bonnet MH (1997) Heart rate variability: Sleep stage, time of night, and arousal influences. *Electroencephalography and Clinical Neurophysiology* 102(5): 390–396.

Elsenbruch S (1999) Heart rate variability during waking and sleep in healthy males and females. *Sleep* 22(8): 1067–1071.

Guilleminault C (1984) Sinus arrest during REM sleep in young adults. *New England Journal of Medicine* 311(16): 1006–1010.

Narkiewicz K (1999) Enhanced sympathetic and ventilatory responses to central chemoreflex activation in heart failure. *Circulation* 100(3): 262–267.

Narkiewicz K (1999) Selective potentiation of peripheral chemoreflex sensitivity in obstructive sleep apnea. *Circulation* 99(9): 1183–1189.

O'Driscoll DM (2004) Cardiovascular response to arousal from sleep under controlled conditions of central and peripheral chemoreceptor stimulation in humans. *Journal of Applied Physiology* 96(3): 865–870.

Somers VK (1989) Influence of ventilation and hypocapnia on sympathetic nerve responses to hypoxia in normal humans. *Journal of Applied Physiology* 67(5): 2095–2100.

Somers VK (1993) Sympathetic-nerve activity during sleep in normal subjects. *New England Journal of Medicine* 328(5): 303–307.

Somers VK (1995) Sympathetic neural mechanisms in obstructive sleep apnea. *Journal of Clinical Investigation* 96(4): 1897–1904.

Tank J (2003) Relationship between blood pressure, sleep K-complexes, and muscle sympathetic nerve activity in humans. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 285(1): R208–R214.

Van de Borne P (1994) Effects of wake and sleep stages on the 24-h autonomic control of blood pressure and heart rate in recumbent men. *American Journal of Physiology – Heart and Circulatory Physiology* 266(2): H548–H554.

Verrier RL (1996) Sleep, dreams, and sudden death: The case for sleep as an autonomic stress test for the heart. *Cardiovascular Research* 31(2): 181–211.

Autonomic Failure

E M Garland and D Robertson, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Autonomic Dysfunction

The clinical features of autonomic dysfunction are pervasive, involving the cardiovascular, gastrointestinal, urogenital, thermoregulatory, sudomotor, and pupillomotor systems. Much of the testing evaluates the integrity of the cardiovascular reflexes. Under normal conditions, cardiovascular reflexes act through the autonomic nervous system to maintain blood pressure and cerebral perfusion at appropriate levels. For example, when stretch or baroreceptors detect a decrease in blood pressure, they relay this information to the central nervous system, which decreases parasympathetic tone and increases sympathetic outflow. Increases in norepinephrine (NE) secretion and peripheral resistance restore blood pressure and provide adequate cerebral perfusion. Disturbances in one or several parts of this reflex arc may result in syncope, a sudden, transient loss of consciousness with spontaneous recovery that may be associated with hypotension, bradycardia (reduced heart rate), and loss of postural tone. Conservative estimates suggest that 30% of the general population has experienced at least one syncopal spell and that syncope is responsible for over 1% of hospital admissions. Infrequently, syncope may be a sign of impaired cardiovascular reflexes resulting from autonomic failure.

Disorders of autonomic failure include central neurodegenerative diseases such as multiple-system atrophy (MSA), primary peripheral autonomic nervous system degeneration (as in autonomic neuropathy or pure autonomic failure (PAF)), and congenital diseases such as familial dysautonomia (FD) and dopamine β-hydroxylase (DBH) deficiency. Patients with severe autonomic failure experience profound hypotension either after assuming the upright position or after food consumption (orthostatic and postprandial hypotension, respectively). In baroreflex failure, exaggerated blood pressure and heart rate fluctuations are exacerbated by emotional or physical stress, whereas patients with postural tachycardia syndrome (POTS) endure orthostatic symptoms and enhanced increases in heart rate with standing. Although Parkinson's disease (PD) is primarily a disorder of motor control, cardiac sympathetic dysregulation has been demonstrated and symptoms of autonomic failure occur in 20–50% of patients. Other systemic illnesses, such as diabetes and amyloidosis, may disturb autonomic function.

Pure Autonomic Failure

Clinical Manifestations

PAF, or Bradbury–Eggleston syndrome, is a persistent, degenerative disorder of the peripheral autonomic nervous system presenting in middle to late life, and affecting men more often than women. The disorder appears to be confined to the sympathetic and parasympathetic nervous systems; the adrenal medulla is relatively spared. Failed attempts to identify the cause of a patient's autonomic failure trigger a diagnosis of PAF.

Although the initial feature of PAF in men is often erectile dysfunction, patients predominantly present with orthostatic hypotension. The orthostatic hypotension, which may be described as unsteadiness, dizziness, or faintness upon standing, is worse in the morning, after meals or exercise, or in hot weather. Patients also complain of pain in the neck or back of the head, relieved by lying down. The orthostatic hypotension is often accompanied by supine hypertension. Nocturia may cause the patient to get up as many as 5 times per night. Urinary hesitancy, urgency, dribbling, and occasional incontinence might also occur. Some patients develop signs of neurogenic urinary retention associated with repeated urinary tract infections. A sudden decline in daily functions in a patient with PAF suggests an infection, usually of the urinary tract. Hypohidrosis, or at least an asymmetrical distribution of sweating, is typical in PAF. Basal metabolic rate may be reduced.

The pathology of PAF has not been completely elucidated, but lesions in the sympathetic ganglia and deficient catecholamine uptake and catecholamine fluorescence in sympathetic postganglionic neurons have been observed. Cell loss has also been reported in the intermediolateral column of the spinal cord, and Lewy bodies, which stain for ubiquitin and α-synuclein, have been found in ganglia, as well as in peripheral autonomic nerves and the central nervous system. Neuroimaging studies confirm the absence of functional sympathetic nerve terminals in the myocardium of patients with PAF.

PAF patients have greatly reduced levels of catecholamines (Figure 1) while lying down and have little increase upon standing. Plasma and urinary NE levels are usually considerably below normal, as is plasma dihydroxyphenylglycol (DHPG; an intraneuronal metabolite of NE). Plasma levels of epinephrine are

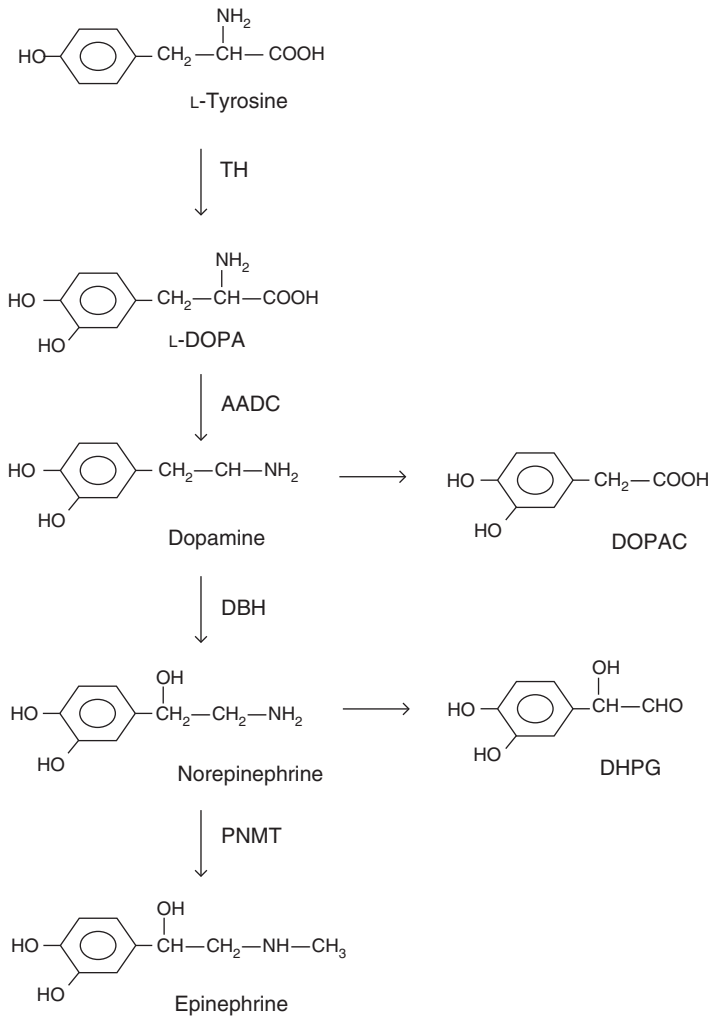

Figure 1 Catecholamine biosynthetic pathway. TH, tyrosine hydroxylase; L-DOPA, levodopa; AADC, aromatic amino acid decarboxylase; DBH, dopamine β-hydroxylase; PNMT, phenylethanolamine-*N*-methyltransferase; DOPAC, 3,4-dihydroxyphenylacetic acid; DHPG, dihydroxyphenylglycol.

also reduced but usually to a lesser extent than NE. Reductions in plasma 3,4-dihydroxyphenylalanine (DOPA) and dopamine (DA) levels are consistent with diminished catecholamine synthesis. In PAF, drugs that release NE from sympathetic terminals have little effect on plasma NE or blood pressure, whereas drugs that directly act on adrenoreceptors produce an exaggerated response.

Diagnosis

The definitive diagnosis of orthostatic hypotension is usually based on a drop in systolic blood pressure of $\geq$20 mmHg and diastolic blood pressure of $\geq$10 mmHg after at least 1 min of standing. A decrease in systolic blood pressure of 50 mmHg or greater is not uncommon in patients with PAF. However, a diagnosis of PAF cannot be excluded based on a single normal upright blood pressure measurement. Several measurements of orthostatic blood pressure should be made.

PAF should be distinguished from two other disorders of autonomic failure: MSA and idiopathic PD. PAF is less progressive and generally less disabling than these other syndromes. For a diagnosis of PAF, there should be no indication from the history or physical examination of cerebellar, striatal, pyramidal, or extrapyramidal dysfunction. Symptoms of hoarseness and sleep apnea are also highly suggestive of MSA. PAF has not been definitively associated with any genetic or environmental causes, nor is it infectious.

Management

Patients with PAF have a generally good outlook; many live for 20 years or more after the onset of their disease. The most common cause of death in

these patients is pulmonary embolus or recurrent infection. The incidence of both myocardial infarction and stroke appears to be significantly reduced.

Some patients with PAF find that leg crossing helps forestall presyncopal symptoms during upright posture. Other nonpharmacological treatment options include high-salt diet, consumption of small meals, elevating the head of the bed, squatting, abdominal compression, bending forward, and compression stockings. Pharmacological treatment is directed mainly at the orthostatic hypotension. Medication options include fludrocortisone, somatostatin analogs, yohimbine, midodrine, erythropoietin, and other vasopressor agents. Patients should be encouraged to avoid factors that can exacerbate orthostatic hypotension, such as large meals, alcohol, drugs, straining during micturition and defecation, and exposure to a warm environment. Finally, a recently discovered treatment option for hypotension is water. Sixteen ounces of water can raise blood pressure by as much as 40 mmHg, with a peak at 30 min after ingestion. Consistent with residual sympathetic tone contributing to supine hypertension, clonidine can be effective at decreasing nighttime blood pressure and natriuresis.

Multiple-System Atrophy

Clinical Manifestations

MSA is a sporadic, progressive, neurodegenerative disease characterized by extrapyramidal, pyramidal, cerebellar, and autonomic dysfunction in any combination. MSA is classified according to predominant features: Shy–Drager syndrome, when autonomic failure predominates; striatonigral degeneration (MSA-P), when extrapyramidal features prevail; and olivopontocerebellar atrophy (MSA-C), for predominantly cerebellar features. The autonomic dysfunction in MSA can be viewed as a primarily central defect with a generally intact peripheral autonomic nervous system.

The average age of onset of MSA is in the sixth decade. Men are affected more frequently than are women, with a male:female ratio of 3–9:1. Up to 75% of MSA patients present with complaints related to autonomic dysfunction, with motor symptoms generally developing within 2 years. Patients who initially have principally movement symptoms develop autonomic dysfunction within 5 years of diagnosis. The most common major autonomic problem of MSA patients is orthostatic hypotension. Other symptoms and signs of autonomic failure include syncope, erectile dysfunction, urinary incontinence or retention, frequent urinary tract infections, constipation, and decreased thermoregulatory sweating. Orthostatic hypotension can be worsened by nighttime pressure natriuresis induced by associated supine hypertension, which occurs in approximately 60% of patients with MSA.

Parkinsonian signs of MSA-P include progressive akinesia and rigidity, with facial stare, and tremor. MSA-C is characterized by gait ataxia with a wide-based gait, cerebellar oculomotor disturbances, intention tremor, and poor performance of rapid, alternating movements. Other symptoms and signs associated with neurodegeneration in MSA include hoarse voice, headache, neck pain, change in writing style, slurred speech, sleep apnea, and recurrent aspiration.

Diagnosis

MSA is classified as possible, probable, or definite on the basis of features and criteria in three clinical domains: autonomic and/or urinary dysfunction, parkinsonism, and cerebellar dysfunction (Table 1). Possible MSA is diagnosed when one criterion and two features from separate clinical domains are found. The diagnosis of probable MSA requires autonomic and/or urinary dysfunction and poorly

Table 1 Clinical diagnosis of MSA

Clinical domain	*Incidence (%)*	*Criterion*	*Features*
Autonomic or urinary dysfunction		Orthostatic hypotension (decrease by ≥30/15 mmHg) or persistent urinary incontinence with erectile dysfunction or both	Orthostatic hypotension ≥20/10 mmHg, urinary incontinency, incomplete bladder emptying
Parkinsonism	87	Bradykinesia and rigidity, postural instability, or tremor	Bradykinesia, rigidity, postural instability, tremor; minimal or reduced response to L-DOPA
Cerebellar dysfunction	54	Gait ataxia and limb ataxia, ataxic dysarthria, or sustained gaze-evoked nystagmus	Gait ataxia, ataxia dysarthria, limb ataxia, sustained gaze-evoked nystagmus
Corticospinal tract dysfunction	49	Not a criterion in diagnosing this disease	Extensor plantar responses with hyperreflexia

levodopa (L-DOPA)-responsive parkinsonism or cerebellar dysfunction. MSA can be definitively diagnosed only at autopsy, by the loss of neuronal and oligodendroglial cells in numerous regions of the brain and spinal cord and by the presence of ubiquitin-containing glial cytoplasmic inclusions (GCIs) in both glial cells and neurons. GCIs stain for α-synuclein, and MSA is classified as a 'synucleinopathy' along with PD and dementia with Lewy bodies.

In patients whose initial symptoms resemble those of PD, MSA should be suspected if there is a poor response to L-DOPA, if there is loud snoring, if disease progresses rapidly to wheelchair confinement, and if prominent urinary tract symptoms are present. Neuroimaging studies have enabled distinction between MSA and PD. All patients with PD and orthostatic hypotension have markedly decreased uptake and retention of sympathoneuronal imaging agents in the heart, whereas all patients with MSA have intact cardiac sympathetic innervation, as measured by sympathetic neuroimaging. Patients with MSA who present with only autonomic and/or urinary dysfunction may be incorrectly diagnosed with PAF. The distinctive features of MSA compared to PAF are presented in Table 2.

The etiology of MSA is unknown, but a genetic component has been considered unlikely because no familial clustering has been reported for the disease. The results of genetic association studies, however, indicate that genetic polymorphisms associated with interleukin-1A, interleukin-8, intercellular adhesion molecule-1, tumor necrosis factor, and α1-antichymotrypsin increase the risk for MSA. An autoimmune process has been suggested to cause the neurodegeneration, although neurotoxicity related to oxidative stress, environmental neurotoxins, and infection has also been proposed.

Management

Patients rarely survive longer than 10 years after a diagnosis of MSA. The most common causes of death are pulmonary embolus, apnea, and intercurrent infection. No current therapy can reverse or halt progression of MSA. Management involves treatment of the depression, tremor and gait disturbances, supine hypertension, orthostatic hypotension, and urinary tract problems. Orthostatic and postprandial hypotension and supine hypertension respond to the interventions mentioned for PAF. Treatment of the movement-disorder component of MSA is less effective than when the same drugs used in PD are administered.

DBH Deficiency

Clinical Manifestations

DBH deficiency, also known as NE deficiency, is a rare disorder due to the congenital absence of the enzyme that converts DA to NE. It is characterized by normal parasympathetic and sympathetic cholinergic function in the absence of sympathetic noradrenergic function. The prevalence of DBH deficiency is unknown. Fewer than 20 cases, all of Western European descent, have appeared in the literature.

Patients with DBH deficiency have experienced vomiting, dehydration, hypotension, hypothermia, and hypoglycemia perinatally. Orthostatic symptoms tend to worsen during adolescence, leading to evaluation and eventual diagnosis. Physical examination of individuals with DBH deficiency reveals poor cardiovascular regulation with a low normal supine blood pressure, a low supine heart rate, and profound orthostatic hypotension with an inadequate compensatory orthostatic tachycardia. The orthostatic hypotension may be accompanied by presyncopal

Table 2 Comparison of MSA with PAF

Characteristic	*Multiple-system atrophy*	*Pure autonomic failure*
Site of autonomic impairment	Mainly preganglionic, central	Mainly postganglionic; peripheral
Prognosis	Poor; median survival 6.5–9.5 years	Good; slow disease progression with survival >10–15 years
Parkinsonian and cerebellar symptoms	Common	Not present
Neuropathological marker	Glial cytoplasmic inclusions in glia and neurons	Lewy bodies in neurons
Orthostatic hypotension	Common	Common
Respiratory disturbances	Common	Uncommon
Urinary tract disturbances	Early in disease	Late in disease
Supine plasma norepinephrine	Normal	Very low
Orthostatic rise in norepinephrine	Subnormal	Subnormal
Response to norepinephrine-releasing drugs	May be increased	Low
Postsynaptic adrenoreceptor sensitivity	Mildly increased	Increased

Table 3 Common clinical features of DBH deficiency

Severe orthostatic hypotension
Postprandial hypotension
High palate
Anemia
Impaired ejaculation
Ptosis of eyelids
Hyperflexible or hypermobile joints
Nasal stuffiness

symptoms, including dizziness, blurred vision, dyspnea, nuchal discomfort, and occasionally chest pain. Although patients are profoundly affected by the defects in autonomic regulation of the cardiovascular system, there is little evidence of central nervous system dysfunction. Other clinical features that have been reported in DBH deficiency are included in Table 3.

Patients with DBH deficiency are unique in that they have minimal or undetectable plasma NE and epinephrine and a five- to tenfold elevation of plasma DA. In addition, metabolites of NE are low or absent in plasma, urine, and cerebrospinal fluid, whereas the DA precursor, DOPA, and metabolites of DA are elevated. The plasma DA concentration responds to various physiological and pharmacological stimuli, similarly to NE in normal individuals. For example, a change from supine to upright posture normally doubles plasma NE levels, but in patients with DBH deficiency, plasma NE remains undetectable while the plasma DA concentration increases two- to threefold. These observations are consistent with release of DA into the synapse in place of NE. Responses to α_1-adrenoceptor agonists and β-adrenoceptor agonists are exaggerated in this disorder.

DBH deficiency is inherited in an autosomal recessive manner. A number of variants in both the coding and noncoding regions of the *DBH* gene have been identified in patients and in unaffected family members but not in unrelated control individuals or in patients with other autonomic disorders.

Diagnosis

Although DBH deficiency appears to be present from birth, the diagnosis is not generally recognized until adolescence or later. Measurement of plasma catecholamines (NE, DA, and their metabolites) is required to verify the diagnosis. Results of autonomic function testing are consistent with sympathetic noradrenergic failure and adrenomedullary failure but intact vagal and sympathetic cholinergic function. DBH enzyme activity varies over a wide range in healthy individuals, and most individuals with low plasma DBH enzyme activity do not have DBH deficiency. Not only is DBH enzyme activity undetectable in the blood of individuals with DBH deficiency, but immunoassay and immunocytochemistry also indicate an absence of DBH protein.

Management

The life expectancy of patients with DBH deficiency is unknown. At least two patients are now in their fifties, and no reported patient has died as yet. For the most part, treatment of DBH deficiency is supportive and directed at relieving orthostatic symptoms. When it was determined that these patients lacked NE, an effective treatment was developed in the form of droxidopa (dihydroxyphenylserine; DOPS). Droxidopa is metabolized directly to NE by L-aromatic amino acid decarboxylase, bypassing DBH. In patients with DBH deficiency, it increases blood pressure and restores plasma NE to the normal range. Plasma DOPA is also lowered to normal levels by droxidopa, but DA and its metabolites remain somewhat elevated and plasma epinephrine continues to be at or below the limits of detection.

Individuals with DBH deficiency do not respond as well to standard medications used for autonomic failure. Fludrocortisone and indomethacin have been mildly effective at raising blood pressure. Phenylpropanolamine elicits some pressor response, presumably due to the denervation hypersensitivity of vascular α-adrenoreceptors.

Familial Dysautonomia

Clinical Manifestations

FD, or Riley–Day syndrome, is a disease of inadequate development and progressive neuronal degeneration that affects central and peripheral components of the autonomic nervous system as well as sensory neurons. It is an autosomal recessive disorder seen almost exclusively in individuals of Ashkenazi heritage. FD is one of a group of rare neurodevelopmental disorders termed hereditary sensory and autonomic neuropathies. More than 500 patients had been diagnosed with FD by 2002.

Patients with FD have gastrointestinal dysfunction, vomiting crises, recurrent pneumonia, decreased pain and temperature perception with intact sensitivity to visceral pain, and cardiovascular instability (Table 4). Patients can exhibit both extreme hypertension and profound and rapid orthostatic hypotension without appropriate compensatory changes in heart rate. Cardiac autonomic dysfunction may be related to hypoplastic cervical and thoracic sympathetic ganglia. Dysautonomic crises with vomiting, hypertension,

Table 4 Clinical features of FD

System	*Common features*
Sensory	Decreased pain and temperature sensation
	Impaired vibration sense
	Intact visceral sensation
Autonomic nervous	Dysphagia
	Esophageal and gastric dysmotility
	Gastroesophageal reflux
	Vomiting crises
	Aspiration
	Insensitivity to apoxia
	Orthostatic hypotension
	Hypertensive crises
	Blotching
	Reduced or absent tears
	Corneal analgesia
Central nervous	Emotional lability
	Normal intellect
Orthopedic	Short stature
	Spinal curvature
Other	Ashkenazi Jewish extraction
	Lack of lingual fungiform papillae
	Decreased patellar reflexes
	Ataxia
	Pneumonia

tachycardia, sweating, blotching of skin, puffy hands, and behavior abnormalities may be triggered by emotional or physical stress.

FD adversely affects growth and development, and the incidence of scoliosis is as high as 85%. Intelligence is normal. An age-related decline in renal function is frequently observed. In individuals with FD, while plasma NE might be normal when patients are supine, the orthostatic increase in NE is reduced. Plasma levels of the NE metabolite DHPG are also low. Plasma DOPA is elevated by about 25%, consistent with stimulation of tyrosine hydroxylase activity, and DA may exhibit dramatic surges during 'crises.' The high DOPA:DHPG ratio in FD is not seen in other autonomic disorders. An exaggerated hypertensive response to infused NE is consistent with supersensitivity resulting from a decrease in the number of postganglionic terminals on peripheral blood vessels.

Diagnosis

The clinical diagnosis of FD is supported by the presence of five 'cardinal' criteria: lack of overflow tears, absent lingual fungiform papillae, decreased patellar reflexes, no axon flare following intradermal histamine, and documentation of Ashkenazi Jewish extraction. However, a definitive diagnosis can now be made using genetic testing of the *IKBKAP* gene.

Management

With greater understanding of the disorder and more effective treatment, quality of life and survival are improving for patients with FD. However, as these patients age, they frequently complain of poor balance, unsteady gait, and difficulty with concentration. Psychiatric disorders may develop, and renal function may decline. Cardiovascular autonomic function may worsen. Due to their increased risk of aspiration, FD patients can develop chronic lung disease.

Treatment of FD is presently supportive and directed toward the various symptoms. Diazepam is used during crises to treat vomiting, hypertension, and agitation. Clonidine can also be administered to help control blood pressure. The orthostatic hypotension that occurs between crises may respond to standard interventions, such as increased salt and fluid intake as well as use of fludrocortisone and midodrine. Decreased sensation requires vigilance, and orthopedic therapy may be needed for scoliosis.

Baroreflex Failure

Clinical Manifestations

Baroreflex failure, a rare disorder, is a result of damage along the baroreflex arc at the baroreceptors, along the glossopharyngeal or vagal nerves, or in the brain stem nuclei. It is generally a consequence of neck surgery, head or neck irradiation for carcinoma, brain stem lesions, or bilateral carotid body tumors.

Acute interruption of the baroreflex arc, with damage isolated to the afferent limb, may be associated with severe, sustained hypertension, tachycardia, and headache. Diaphoresis and apnea may occur, especially in the first 48 h postoperatively. Labile, episodic hypertension (volatile hypertension), on the other hand, is often seen in patients who develop baroreflex failure gradually or during a more chronic phase of the disorder. Surges in blood pressure are elicited by mental or physical stress, during which sympathetic outflow is increased. These patients experience sensations of warmth or flushing, dizziness or light-headedness, palpitations, headache, and diaphoresis. Tremulousness, anxiety, and irritability may also be present. In patients with baroreflex failure, periods of volatile hypertension can be interrupted by hypotensive episodes, especially during periods of quiet, sedation, or sleep, when sympathetic outflow is diminished.

In some cases of baroreflex failure, with interruption of efferent vagal function, tachycardia is provoked by mild sympathetic activation. In other patients, with selective baroreflex failure (Jordan syndrome), the lesion is in the afferent input from the carotid sinus to the nucleus tractus solitarii, and

efferent sympathetic and parasympathetic output remains intact. Malignant vagotonia with hypotension, bradycardia, and asystole characterize selective baroreflex failure. Accompanying symptoms include fatigue and dizziness, with possible syncope. Plasma NE levels in baroreflex failure tend to parallel the blood pressure changes. During periods of relatively low sympathetic activity, NE may be normal (111–360 pg ml^{-1} (0.66–2.13 nM)). During hypertensive-tachycardic episodes, plasma NE may rise to levels as high as 2260 pg ml^{-1} (13.36 nM).

Diagnosis

The key diagnostic feature in baroreflex failure is a parallel increase in blood pressure and heart rate with stress, and a parallel decrease with sedation or rest. In addition, pressor agents do not cause a reflex decrease in heart rate nor does reflex tachycardia occur after vasodilators. A history of neck surgery, irradiation, or some other traumatic event would support this diagnosis.

Management

The primary goal of therapy of patients with baroreflex failure is to decrease the frequency and magnitude of surges in blood pressure and heart rate. Clonidine is the treatment of choice for blood pressure surges. Once symptoms have been well controlled, diazepam can be instituted to control stress and decrease central sympathetic input. Medication, such as fludrocortisone, may also be required to prevent hypotensive episodes. In patients with selective baroreflex failure, a pacemaker may be needed. Finally, because of the excessive levels of plasma NE in baroreflex failure, agents that prevent release of NE may also be helpful, and agents that can increase synaptic NE should be avoided.

Postural Tachycardia Syndrome

Manifestations

POTS differs from other forms of autonomic failure in that upright posture is associated with little or no decrease in blood pressure but marked orthostatic tachycardia. Also known as orthostatic intolerance (OI), POTS is relatively common (approximately 500 000 Americans are affected) and tends to occur in individuals between 15 and 50 years old, with a female:male ratio of 5:1. POTS is the most frequently encountered dysautonomia among patients referred to centers specializing in autonomic disorders.

Posture-related symptoms in POTS are consistent with impaired cerebral perfusion and include lightheadedness, clouding of thought, blurred vision, anxiety, substernal chest pain, fatigue, palpitations, and occasionally syncope. Warm dry feet, excessive orthostatic blood pooling and dusky skin in the lower leg, reduced galvanic skin response, abnormal sweating on the extremities, and hyperresponsiveness of leg veins to NE are other clinical signs of POTS. The supine heart rate is usually normal or slightly elevated, whereas patients typically have a greater than 30 beats per minute increase in heart rate on standing. Blood pressure in patients is similar to that in healthy individuals in both supine and upright postures. Supine plasma catecholamines are high, and their responses to standing are exaggerated.

Pathophysiology and Diagnosis

POTS is a heterogeneous disorder, and a number of different pathophysiologies underlie symptoms, perhaps with elevated sympathetic drive to the cardiovascular system as a final common pathophysiological mechanism in the majority of patients. Some cases of POTS may result from a patchy impairment of autonomic neuronal function that allows the heart to remain normally innervated in the face of impaired innervation of the distal extremities. High plasma NE levels support the view that increased sympathetic nervous system activation can underlie POTS in some patients. These patients may have reduced plasma renin activity and aldosterone, as well as supine and dynamic orthostatic hypovolemia. A low absolute or orthostatic blood volume would be expected to elicit tachycardia on standing. Other potential pathophysiological mechanisms include excessive venous pooling, ganglionic-receptor-binding antibodies, exaggerated gravity-dependent fluid shift, cardiac β-adrenoceptor hypersensitivity, and diminished cardiovagal baroreflex sensitivity with impaired central autonomic regulation. The onset of POTS is often preceded by a recent viral infection. Patients can undergo extensive clinical evaluation by a variety of specialists, but the basis of POTS remains obscure in most afflicted individuals and it remains undiagnosed in many patients.

A small percentage of patients diagnosed with POTS have norepinephrine transporter (NET) deficiency, a condition that was first described in 2000. The substantial orthostatic tachycardia in NET deficiency is accompanied by an increase in the plasma NE concentration to almost 4 times its value in the supine position. The relatively low concentrations of plasma DHPG, relative to NE, as well as decreased NE clearance, increased NE spillover and tyramine resistance, are consistent with impaired NET function. A missense mutation has been identified in the NET gene that renders the transporter nonfunctional. In case-control studies, genes encoding the β1- and β2-adrenoreceptors, nitric oxide synthase, and endothelin have been associated with the development of POTS and the

magnitude of orthostatic responses. Contributions of these genes have been relatively minor, however, and a number of other genes are likely to contribute to POTS.

Management

A majority of patients with POTS have a relatively mild disorder, which improves over time. Those patients with potentially precipitating events, such as a viral infection, appear to do better overall than others. Volume loading and increases in sodium intake blunt the orthostatic tachycardia, normalize the plasma NE response to standing, and improve orthostatic symptoms in POTS. Effective therapeutic regimens include orthostatic 'exercise,' water, fludrocortisone, low-dose propranolol, clonidine, midodrine, and pyridostigmine.

See also: Autoimmune Autonomic Neuropathy; Autonomic Disorders; Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Metabolic Function; Autonomic Nervous System: Neuroanatomy; Autonomic Nervous System: Central Cardiovascular Control; Autonomic Nervous System; Blood Pressure: Baroreceptors; Dopamine; Dopamine in Perspective; Dysautonomia: Familial; Sympathetic Nervous System.

Further Reading

Axelrod FB (2004) Familial dysautonomia. *Muscle and Nerve* 29: 352–363.

Garland EM, Hahn MK, Ketch TP, et al. (2002) Genetic basis of clinical catecholamine disorders. *Annals of the New York Academy of Sciences* 971: 506–514.

Gilman S, Low PA, Quinn N, et al. (1999) Consensus statement on the diagnosis of multiple system atrophy. *Journal of the Neurological Sciences* 163: 94–98.

Goldstein DS (2001) *The Autonomic Nervous System in Health and Disease*. New York: Marcel Dekker, Inc.

Goldstein DS, Robertson D, Esler M, et al. (2002) Dysautonomias: Clinical disorders of the autonomic nervous system. *Annals of Internal Medicine* 137: 753–763.

Grubb BP (2005) Neurocardiogenic syncope and related disorders of orthostatic intolerance. *Circulation* 111: 2997–3006.

Hilz MJ and Dutsch M (2006) Quantitative studies of autonomic function. *Muscle and Nerve* 33: 6–20.

Ketch T, Biaggioni I, Robertson R, et al. (2002) Four faces of baroreflex failure: Hypertensive crisis, volatile hypertension, orthostatic tachycardia, and malignant vagotonia. *Circulation* 105: 2518–2523.

Man in't Veld AJ, Boomsma F, Lenders J, et al. (1988) Patients with congenital dopamine beta-hydroxylase deficiency. A lesson in catecholamine physiology. *American Journal of Hypertension* 1: 231–238.

Mathias CJ (2003) Autonomic diseases: Clinical features and laboratory evaluation. *Journal of Neurology, Neurosurgery & Psychiatry* 74: iii31–iii41.

Mitsky V and Robertson D (1995) Failure of the autonomic nervous system. *Comprehensive Therapy* 21: 529–534.

Parikh SM, Diedrich A, Biaggioni I, et al. (2002) The nature of the autonomic dysfunction in multiple system atrophy. *Journal of the Neurological Sciences* 200: 1–10.

Robertson D and Biaggioni I (1995) *Disorders of the Autonomic Nervous System*. Luxembourg: Harwood Academic Publishers.

Robertson D, Flattem N, Tellioglu T, et al. (2001) Familial orthostatic tachycardia due to norepinephrine transporter deficiency. *Annals of the New York Academy of Sciences* 940: 527–543.

Robertson D, Haile V, Perry SE, et al. (1991) Dopamine beta-hydroxylase deficiency. A genetic disorder of cardiovascular regulation. *Hypertension* 18: 1–8.

Autonomic Nervous System

J B Furness, University of Melbourne, Parkville, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

The autonomic nervous system (ANS) is the system of neurons that control peripheral organs, other than striated muscle that is under voluntary control (Figure 1). Thus the ANS controls the visceral organs of the thoracic, abdominal, and pelvic cavities, which includes the lungs, heart, digestive organs, kidneys, urinary bladder, and internal generative organs. It also controls endocrine and exocrine glands, the blood vessels that supply all organs, and, within the eye, the diameter of the iris (Table 1).

The purpose of the ANS is to adjust the activities of these organs so that they function at levels that are most favorable to the state of the body and to its environment. The ANS is thus one of two systems, the other being the endocrine system, that control the functions of the internal and surface organs, including the skin. The two control systems would be better thought of as one, because they act in synergy to control organs.

Autonomic control of organs is through reflexes and through cortical control centers. To elicit a reflex, the relevant states of organs must be detected. This detection is through visceral afferent neurons, that are properly regarded as part of the ANS. Many visceral afferent neurons also communicate other information, for example, pain from the viscera, satiety from the digestive tract, or temperature. Thus, autonomic visceral afferent neurons, while part of the ANS, may carry signals to the central nervous system that serve other functions. Some autonomic visceral afferent neurons (e.g., baroreceptor neurons and intrinsic primary afferent neurons of the intestine) seem to have roles exclusively in autonomic reflexes.

Since the late nineteenth century and early twentieth century, it has been common to divide the ANS into three divisions – the sympathetic, parasympathetic, and enteric divisions. There were pragmatic reasons for this separation of parts of what is essentially one control system. The efferent (motor) autonomic outflows from the central nervous system have gaps, that is, there are nerves emerging from the brain and spinal cord that do not carry autonomic motor pathways (Figure 1). Autonomic fibers are absent from the first two cranial nerves, olfactory and optic; they are then present in cranial nerves III, VII, IX, and X. Then a gap in outflow occurs, with no autonomic contribution to cranial nerves XI and XII or the cervical nerves (there can rarely be a contribution to C7). The next group of nerve roots, T1 to L2 or -3 (the thoracolumbar outflows), all have autonomic components, and then there is a small gap where there are few autonomic fibers, which become prominent again in sacral roots 2–4. The outputs are thus considered as cranial, thoracolumbar, and sacral. The number of segments covered by the divisions varies slightly between mammalian species; the levels quoted are for human.

Some of the pathways that emerge from the cranial autonomics slow the heart and dilate blood vessels, and there are also dilator nerves in the sacral outflows. Thus, these were grouped for classification purposes as the craniosacral autonomic. Conversely, there are cardio-accelerator and vasoconstrictor pathways in the thoracolumbar outflows. The nerves of the thoracolumbar outflows were termed sympathetic because they were proposed to maintain a sympathy between the organs. The autonomic nerves of the craniosacral division of the ANS were later called parasympathetic, because of their opposite effects on the cardiovascular system, being associated with slowing or lowering cardiovascular function.

The division of the ANS into sympathetic and parasympathetic has led to enormous misconceptions, the most serious being the concept that the two divisions are somehow in opposition to each other. This is quite a wrong idea. Autonomic nerves, whatever their anatomical origin, act in concert to control visceral organs and the vasculature.

The third division of the ANS is the enteric division, which is the system of autonomic ganglia and nerve fibers that is contained within the walls of the digestive organs. This is given a status as a separate division because it contains complete reflex circuits which can operate in the absence of connections with the central nervous system. In terms of numbers of neurons, the enteric is the largest autonomic division. In humans, it contains 200–600 million neurons.

The motor pathways of the ANS that arise in the central nervous system pass through autonomic ganglia and make synapses on the way to the organs that they innervate, with the exception of the pathways to the adrenal glands. Synaptic transmission in autonomic ganglia is mediated primarily by acetylcholine (ACh) that acts on nicotinic receptors. The use of nicotine and other drugs that block these receptors has been important in analyzing the nerve pathways. Neurons with cell bodies in the central nervous system that make synapses in peripheral ganglia are known as autonomic (sympathetic or parasympathetic)

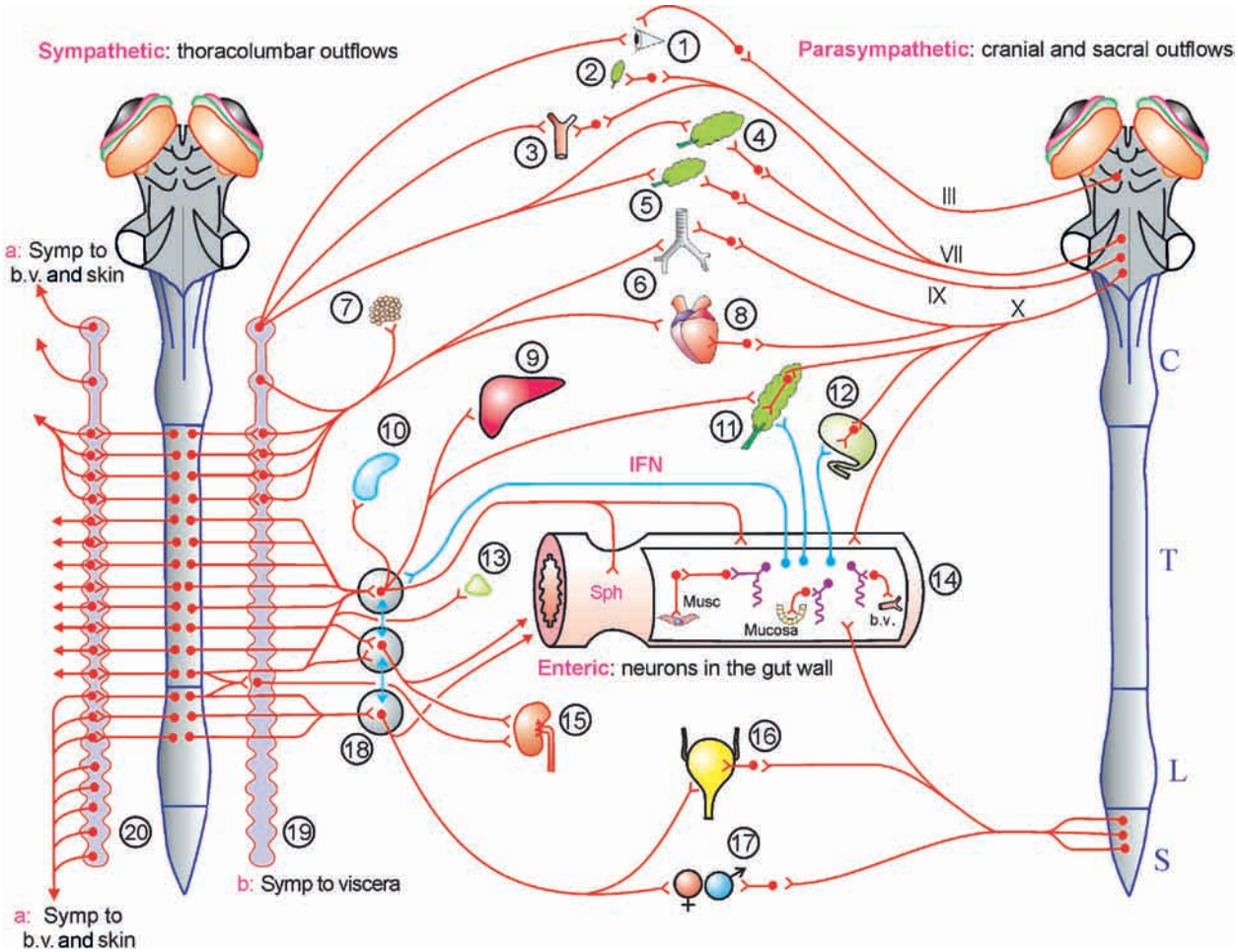

Figure 1 Schematic representation of the efferent (motor) pathways of the ANS, including illustration of some peripherally confined primary afferent neurons (autonomic visceral afferent neurons relaying signals to the central nervous system (CNS) have been omitted to simplify the diagram). The brain stem and spinal cord are represented twice, on the left to show sympathetic connections and on the right to show parasympathetic connections. The enteric division is within the gut wall (center of diagram (14)). *Sympathetic outflows.* (a) To the left side of the spinal cord, sympathetic chain (paravertebral) ganglia are represented; some neurons of these ganglia supply blood vessels (b.v.) throughout the body and effectors in the skin (sweat glands pilomotor muscles). These pathways have synapses in the paravertebral ganglia. For simplicity of illustration, pathways that run rostrally and caudally within the sympathetic chains are not illustrated. (b) On the right side of the cord are the connections that pass first through the sympathetic chains and then through prevertebral ganglia and plexuses (18) to supply visceral organs, as well as pathways that supply structures in the head and neck and intracranial arteries. Synapses occur in either prevertebral and paravertebral ganglia. *Parasympathetic outflows.* These emerge from cranial and sacral levels and innervate structures in the head, neck, abdomen and pelvis, but not in the limbs. *Enteric neurons.* These are represented within the outline of the intestine (14). The enteric reflex circuits contain intrinsic primary afferent neurons (IPANs, purple), interneurons and motor neurons (red). As illustrated, these control muscle (musc), the secretory epithelium of the mucosa and blood vessels (b.v.). Sph: this indicates the sphincter regions of the intestine; these are controlled by enteric and extrinsic neurons. *Peripherally confined neural connections between organs.* These are marked in blue. One of the neurons that contribute to these circuits is the intestinofugal neuron (IFN) that projects from the intestine to prevertebral ganglia. *Target tissues and organs.* (1) Eye; (2) lacrimal glands; (3) intracranial arteries; (4) and (5) salivary glands; (6) airways; (7) brown fat; (8) heart; (9) liver; (10) spleen; (11) pancreas; (12) gallbladder; (13) adrenal gland; (14) tubular gastrointestinal tract; (15) kidney; (16) urinary bladder; (17) genital organs; (18) prevertebral ganglia and plexuses; and (19, 20) sympathetic chains (paravertebral ganglia and their interconnections). *Spinal cord levels.* C, cervical; T, thoracic; L, lumbar; S, sacral.

preganglionic neurons. The neurons with which they connect are called postganglionic neurons. Enteric neurons are innervated by parasympathetic preganglionic neurons and sympathetic postganglionic neurons.

The ANS is associated with specific collections of neurons within the central nervous system. A series of nuclei in the brain stem (including the Edinger–Westphal nucleus, the salivatory nuclei, and the dorsal motor nucleus of the vagus) contain the cell bodies of preganglionic neurons. Within the spinal cord, autonomic cell groups are in the intermediolateral column nuclei at levels corresponding to autonomic outflows,

Table 1 A summary of functions controlled by the ANS

Heart rate, force, and conduction
Arterial diameter (all vascular beds)
Mesenteric venous capacity
Pupillary diameter, accommodation of lens
Exocrine gland secretion: lacrimal gland, salivary glands, gastric glands, exocrine pancreas, sweat glands, glands of genital organs
Endocrine secretion: adrenal medulla, endocrine pancreas
Secretion into organs: intestinal water and electrolyte secretion, pulmonary secretion, nasal secretion
Gastrointestinal wall movement
Gall bladder contraction and biliary tract motility
Regulation of the urinary bladder and control of micturition
Tracheal and bronchial diameter
Contraction of vas deferens, vagina, other internal genitalia, dartos muscle
Penile erection, clitoral and labial engorgement
Fat mobilization
Immune system[a]
Piloerection

[a]See especially Jänig (2006).

from the upper thoracic to mid-lumbar levels and in the sacral spinal cord. The central autonomic motor nuclei receive inputs from autonomic integrative cell groups, including the nucleus of the tractus solitarius, autonomic cell groups of the ventrolateral medulla, the paraventricular nucleus, and other cell groups within the hypothalamus. At higher levels, important autonomic control is exerted through the parabrachial nuclei, amygdala, insular cortex, and cingulate cortex. At these higher levels, it is no longer possible to distinguish the centers as simply autonomic. They are also involved in endocrine control and affective behavior, both of which are closely related to autonomic function.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Neuroanatomy; Autonomic Neuroplasticity: Development; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Blessing WW (1997) *The Lower Brainstem and Bodily Homeostasis*. Oxford: Oxford University Press.

Furness JB (2006) *The Enteric Nervous System*. Oxford: Blackwell Publishing.

Furness JB (2006) The organisation of the autonomic nervous system: Peripheral connections. *Autonomic Neuroscience* 130: 1–5.

Jänig W (2006) *The Integrative Action of the Autonomic Nervous System*. Cambridge, UK: Cambridge University Press.

Loewy AD and Spyer KM (1990) *Central Regulation of Autonomic Functions*. 390pp. Oxford: Oxford University Press.

Morrison SF (2001) Differential control of sympathetic outflow. *American Journal of Physiolology* 281: R683–R698.

Saper CB (2002) The central autonomic nervous system: Conscious visceral perception and autonomic pattern generation. *Annual Review of Neuroscience* 25: 433–469.

Autonomic Nervous System Development

D F Newgreen, Royal Children's Hospital, Parkville, VIC, Australia
M J Howard, Medical University of Ohio, Toledo, OH, USA
R Nishi, University of Vermont College of Medicine, Burlington, VT, USA

© 2009 Elsevier Ltd. All rights reserved.

Origin of the Autonomic Nervous System from the Neural Crest

All neurons and glia of the autonomic nervous system – the sympathetic, parasympathetic, and enteric nervous systems – are derived from the neural crest (NC). The NC, a migratory cell population unique to vertebrates, arises during early embryogenesis, becoming microscopically distinguishable in a rostral-to-caudal wave in human embryos approximately from embryonic (E) days E18 to E40, in mice from E8.5 to E11, and in chicks from E1.3 to E3.5. The induction of the NC has been elucidated by, among others, Marianne Bronner-Fraser's group (bird embryos) and Robert Mayor's group (frog embryos). This cell population is induced by earlier signals from neighboring tissues, and it consists of epithelial cells at the border between the medial neural ectoderm, which gives rise to the central nervous system (CNS), and the more lateral epidermal ectoderm, which contributes to the skin. This involves signaling via secreted growth factors: fibroblast growth factors (FGFs), Wnts, and transforming growth factor (TGF)-β family members, especially the bone morphogenetic proteins (BMPs). This sets up a pattern of gene activity that specifies the NC cells as different from the ectodermal cells on each side. This involves expression in the nascent NC cells of genes such as *Pax7* and *Zic3* (in mice) for transcription factors (which control the expression of other genes). In a positive feedback, the gene for the inducer BMP-4 is also activated in the NC. This is rapidly followed by the transient expression of transcription factor genes such as *Snail1* (in mice; *SNAI2/SLUG* in birds and reptiles), the forkhead gene *Foxd3* (mice), and the Sox group E genes (in mice, *Sox8, Sox9*, and *Sox10*).

The transcription factors encoded by these genes promote the onset of cell migration, a classic example of the epithelial–mesenchymal transition. NC cells lose the homophilic cell–cell adhesion molecule N-cadherin first at translational then at transcriptional levels. This is accompanied by reorganization of the cytoskeleton into a motile form and by the ability to adhere to and also digest extracellular matrix involving, among others, expression of genes for RhoA (actin modulator), integrin (matrix adhesion molecule), and matrix metalloproteases (MMPs). This allows the formerly coherent epithelial cells to assume a quasi-individualistic migratory behavior.

The fate of the NC cells after migration has been demonstrated in most detail by Nicole Le Douarin and colleagues using chick–quail orthotopic transplantation as a device for labeling and tracking cells. These form not only the peripheral autonomic nervous system, but also many other neural and related endocrine cells, and even connective tissues. These fates can be mapped to position of origin along the neuraxis (Figure 1). The differentiation competence of NC cells has been tested by heterotopic transplantation. These experiments indicate that competence is broader than the fate, and the particular lines of differentiation are controlled by the tissues with which the NC cells interact. However, superimposed on this are also some early-imposed restrictions on differentiation. More rostrally originating cells have a wider range of options; this is epitomized by the restriction of connective tissue competence to cranial NC cells. This early spatial restriction also extends to some autonomic competences such as enteric neurogenesis capacity. These cells can be elicited when NC cells are combined with intestinal tissues, but only cranial NC cells are competent to produce large numbers of neurons. In contrast, sympathetic neurons can differentiate from all of the NC, including cranial levels that are not fated to do so.

The nascent NC cells, like the forming CNS, express multiple homeobox genes in a nested rostro-to-caudal pattern. There are four clusters of these genes (A, B, C, and D), each cluster being variously numbered (from 1 to 13). The equivalently numbered genes of each cluster are similar and tend to be co-expressed. The most rostral and earliest formed neural tissues including NC, at forebrain and midbrain levels, express no homeobox genes. Progressively higher number homeobox genes are expressed in the NC at progressively more caudal levels. Thus, *HOXB1* is expressed maximally at hindbrain levels rostral to the ear, *HOXB4* in the hindbrain caudal to the ear, and *HOXB9* at the level of the neck. A combinatorial code of homeobox gene expression is important in specifying an early positional memory that influences later differentiation processes. It is possible that the rostral-to-caudal patterned restrictions on later differentiation in early NC cells, noted earlier, are regulated by the homeobox code.

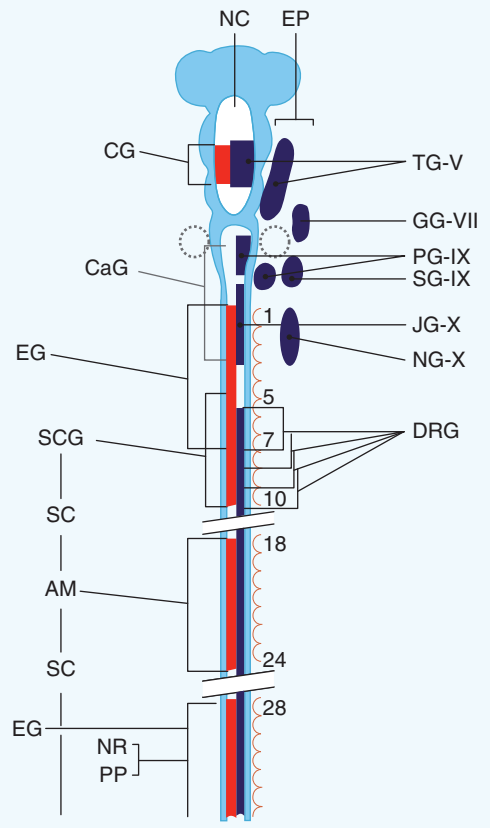

Figure 1 Scheme of neural crest (NC) derivatives mapped onto the neuraxis (light blue). Autonomic nervous system derivatives (red) include ciliary ganglion (CG), cardiac ganglion (CaG), enteric ganglia (EG), superior cervical ganglia (SCG), sympathetic ganglia (SC), nerve of Remak (NR; avian only), pelvic ganglia (PP), and adrenomedullary (AM) gland ganglia. Related sensory neural structures (dark blue), dorsal root sensory ganglia (DRG), and epibranchial placodes (EP) give rise to cranial sensory ganglia. The somites are numbered, and the otic (ear) vesicle position is circled. Adapted from Newgreen DF and Erickson CA (1986) The migration of neural crest cells. *International Review of Cytology* 103: 89–145.

Migratory Morphogenesis of Neural Crest Cells

The outstanding characteristic of NC cells is their migratory morphogenesis. The sympathetic and parasympathetic cells require a migration of only several 100 μms from their NC origin, and in mouse and chick embryos this takes about half a day. Trunk NC cells that form the sympathetic system migrate in lateroventral chains through the mesenchymal cell mass of the rostral half of the adjacent sclerotome (a subregion of the somitic body segments) to reach the dorsal aorta (Figure 2). Recent live cell imaging *in vivo* has shown that these cells then turn rostrally and caudally parallel to the aorta, so that cells from one segmental level become distributed over four or more segments.

The enteric nervous system is by far largest division of the autonomic nervous system (ANS) in cell numbers and is much more widely distributed. Yet it arises from more restricted NC regions, mostly stemming from the vagal neuraxis (in the caudal hindbrain and slightly overlapping the trunk) with a numerically minor component from lumbosacral levels (Figure 1). The vagal cells migrate over and through the vagal-level somites to the nearby foregut (esophageal, gastric, and duodenal primordia), then migrate through the dense gut mesenchyme of the midgut (future small intestine and cecum) and hindgut (future colon). This colonization takes about 4 weeks in humans and 4–5 days in mouse and chick embryos. Live cell imaging from Heather Young and Miles Epstein and their colleagues has shown that these cells also migrate as dynamic chains (Figure 3). Biomathematical modeling and experimental studies by Kerry Landman and co-workers have indicated that proliferation in the NC wavefront is a major driver of colonization of the gastrointestinal tract. However, completion of colonization to the distal colon is made more difficult, since the gut is simultaneously elongating. Failure to complete colonization leads to Hirschsprung's disease, in which the distal gut lacks neural ganglia, cannot perform peristalsis, and, after birth, becomes massively distended with fecal contents proximal to the aganglionic region. Mutations in many genes predispose to neural dysplasias of the intestine, and many of these defects are Hirschsprung-like (Table 1). Many of these genes control NC cell numbers by affecting cell proliferation or survival; this may lead to an apparent migration defect.

Molecular Control of Migration

Migration routes of NC cells are controlled by receptor-mediated interactions with their surroundings, including extracellular matrix molecules, growth factors, and molecules on the surface of non-NC cells. Initial migration is into extracellular matrix that forms a migratory substrate. Important matrix molecules include fibronectin, laminins, and collagens, for which NC cells possess multiple integrin class adhesion and signaling receptors. Treatment with peptides and antibodies that block matrix interactions can cause NC cell migration to stall *in vitro* and *in vivo*. Chemoattraction of NC cells may also occur: the glial cell line-derived neurotrophic factor (GDNF) appears to attract enteric NC cells in *in vitro* tests, and is produced by intestinal mesenchyme cells (Figure 4). However, analysis is complicated, because growth factors often have several functions; in this

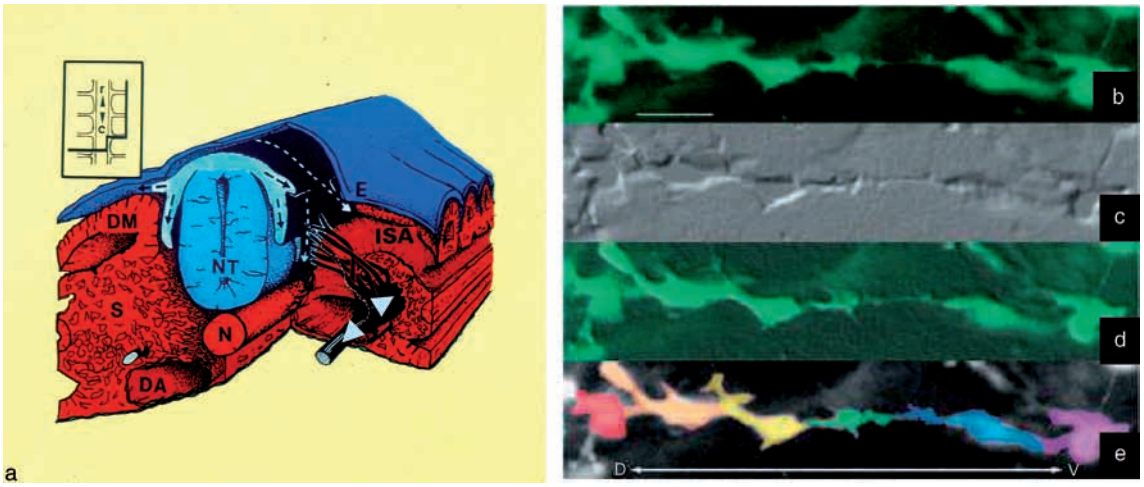

Figure 2 (a) Cutaway scheme of neural crest cells migrating as chains through the rostral sclerotome (S) to the dorsal aorta (DA), where they turn rostrally and caudally (arrows; see inset: r, rostral; c, caudal). Other areas depicted are the demomyotome (DM), epidermal ectoderm (E), intersegmental artery (ISA), notochord (N), and neural tube (NT). (b–e) Single neural crest cell chain, extending from neural tube to dorsal aorta. (b) Green fluorescent protein-labeled neural crest cells. Scale bar = 20 μm. (c) Embossed image of chain. (d) Fluorescence and embossed overlay. (e) Individual cells within chain colored separately; D↔V, dorsal to ventral. (b–e) From Kasemeier-Kulesa JC, Kulesa PM, and Lefcort F (2005) Imaging neural crest cell dynamics during formation of dorsal root ganglia and sympathetic ganglia. *Development* 132: 235–245.

example, as well as a chemotactic role, GDNF is a required survival factor, a potent mitogen, and a differentiation driver for NC cells possessing its receptor, Ret (the product of *ret proto-oncogene*).

Negative regulators of migration are of equal importance to positive regulators. These molecules repel NC cells from regions such as the perinotochordal zone, the caudal half of the somites, parts of the gut mesenchyme, and elsewhere. These repulsive molecules include ephrinB1 (receptor on NC cells: EphB3), semaphorins 3A and 3F (receptors: neuropilins 1/2), and slit 2 (receptor: Robo), the large chondroitin sulfate proteoglycan aggrecan, the glycoprotein tenascin-C, and the growth factor sonic hedgehog (Shh).

Migrating NC cells contact each other, and these contacts are important for migration and directional choice. Isolated NC cells show little translocation. NC cells at the forefront of migration, and at certain 'decision points,' extend filopodia simultaneously in many directions, and move erratically, but NC cells in contact in chains move more consistently. Major cell–cell adhesion molecules N-cadherin and N-CAM are reduced on NC cells in the migratory phase, but lower affinity cadherins may maintain the transient adhesions observed. Interference with the cell adhesion molecule L1CAM on mouse NC cells disrupts the chainlike connections of these cells in the intestine and delays migration. Interestingly, *L1CAM* is a modifier gene for Hirschsprung's disease genes (see Table 1), and since it is located on the X chromosome, it may contribute to the 4:1 male:female ratio of Hirschsprung's disease.

Gangliogenesis

Less is known about gangliogenesis in the ANS than about the process of migration, but upregulation of the homophilic cell adhesion molecules N-cadherin and N-CAM often occurs, and genetic disruption of these systems leads to less compact aggregates. In the forming of sympathetic ganglia, initially hemisegmental NC cells migrate longitudinally across segmental boundaries to form a relatively uniform chain (Figure 2(a)). This chain then, via N-cadherin upregulation, forms a string of aggregates. These aggregates assemble at the anterior half of each hemisegment, and this is based on avoidance of repulsive interactions with ephrinB1 that is increasingly expressed in the alternate posterior hemisegment.

In the enteric nervous system the ganglia are uniformly small and regularly spaced in two dimensions, in two layers in the gut mesenchyme. These ganglionated plexuses form at a distance from the gut endoderm. When Shh expression is reduced, neurons are found closer to the endoderm, suggesting that Shh is involved in establishing a domain that excludes NC-derived cells. The cell behavior in ganglion formation, visualized by time-lapse microscopy, suggests that the uniformity of size and spacing could result from a balance between contact-mediated NC cell cohesion and repulsion-at-a-distance between

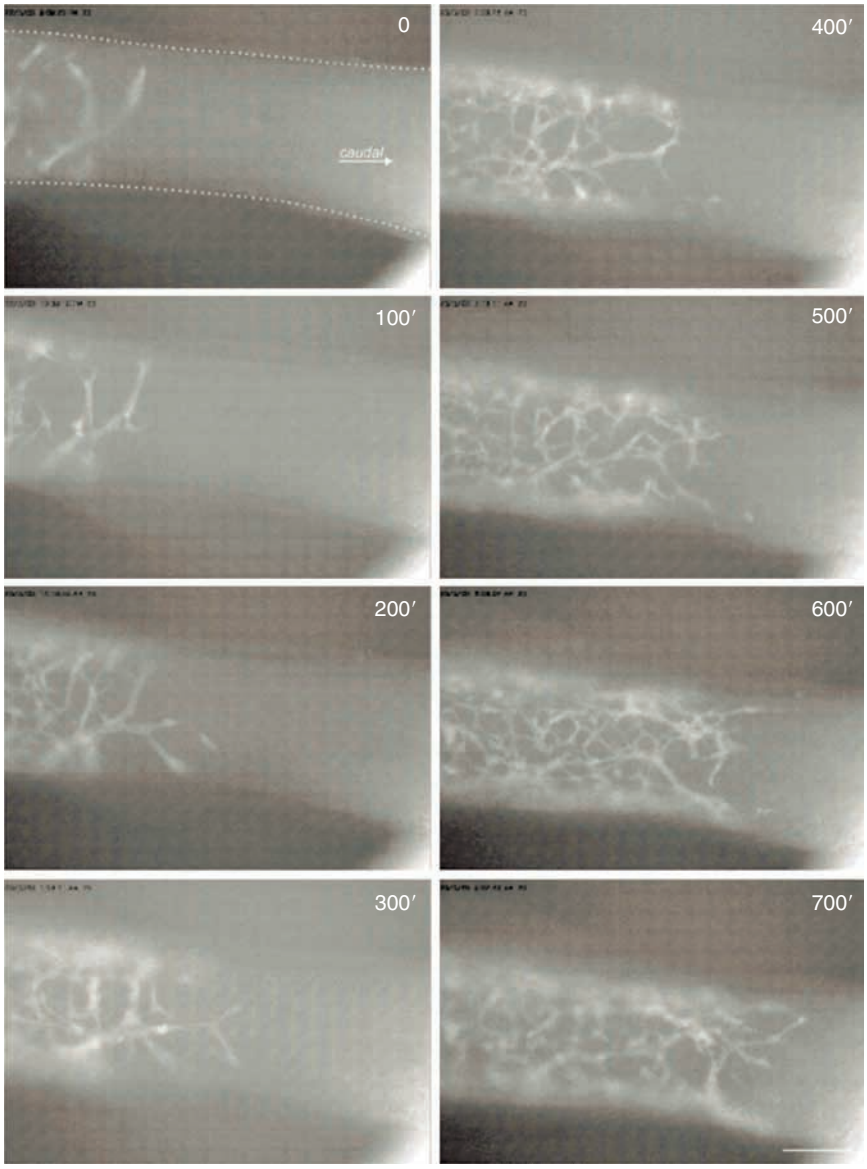

Figure 3 Selected time-lapse frames of the hindgut of an E12.5 mouse showing the caudal progression of green fluorescent protein-expressing neural crest cells. Time is noted in minutes. Most of the cells are present in intersecting chains that follow a variety of trajectories. Scale bar = 100 μm. From Young HM, Bergner AJ, Anderson RB, et al. (2004) Dynamics of neural crest-derived cell migration in the embryonic mouse gut. *Developmental Biology* 270: 455–473.

like cells. However, the molecules responsible for this reorganization are not yet known.

Sympathetic Ganglia: Neuronal Differentiation and Connections

The sympathetic chain ganglia, derived from trunk NC (Figure 1), innervate all of the organs, smooth muscle, skeletal muscle, and glands, and serve to modulate their function. Innervation proceeds in a rostral-to-caudal pattern, with more rostral neurons innervating rostral structures and caudal neurons innervating caudal organs.

The preganglionic motor neurons that innervate sympathetic chain ganglia arise at thoracic levels of the spinal cord. Their axons leave the spinal cord via the ventral roots and project both rostrally and caudally. The rostral-to-caudal projection pattern is determined by the location of the cell bodies along the neural axis. Several factors contribute to motor column-specific motor neuron identity as well as generation of a segment-specific projection pattern. The generation of neurons in the Column of Terni (designation for preganglionic motor neurons in the chick) depends upon loss of expression of the homeodomain

Table 1 Genes implicated in enteric nervous system formation and dysplasias

Gene	*Human chromosomal location*	*Phenotype of ENS in mice in which the gene is homozygously inactivated*
RET	10q11.2	Absence of neurons from small and large intestines
GDNF	5p12-p13.1	Absence of neurons from small and large intestines
GFRAI	10q25	Absence of neurons from small and large intestines
EDNRB	13q22	Absence of neurons from distalmost large intestines
EDN3	20q13.2-q13.3	Absence of neurons from distalmost large intestines
ECE1	1p36.1	Absence of neurons from distalmost large intestines
PHOX2B	4p12	Absence of neurons from entire gastrointestinal tract
SOX10	22q13	Absence of neurons from entire gastrointestinal tract
PAX3	2q37	Absence of neurons from small and large intestines
ASCL1	12q22-q23	Absence of neurons from esophagus
IHH	2q33-q35	Absence of neurons from parts of the small intestine and colon
SHH	7q36	Ectopic neurons within mucosa
ZEB2 (*SIP1*)	2q22	Absence of neurons from distalmost large intestine
TLX2	2p13.1	ENS hyperplasia in colon and hypoplasia in small intestine

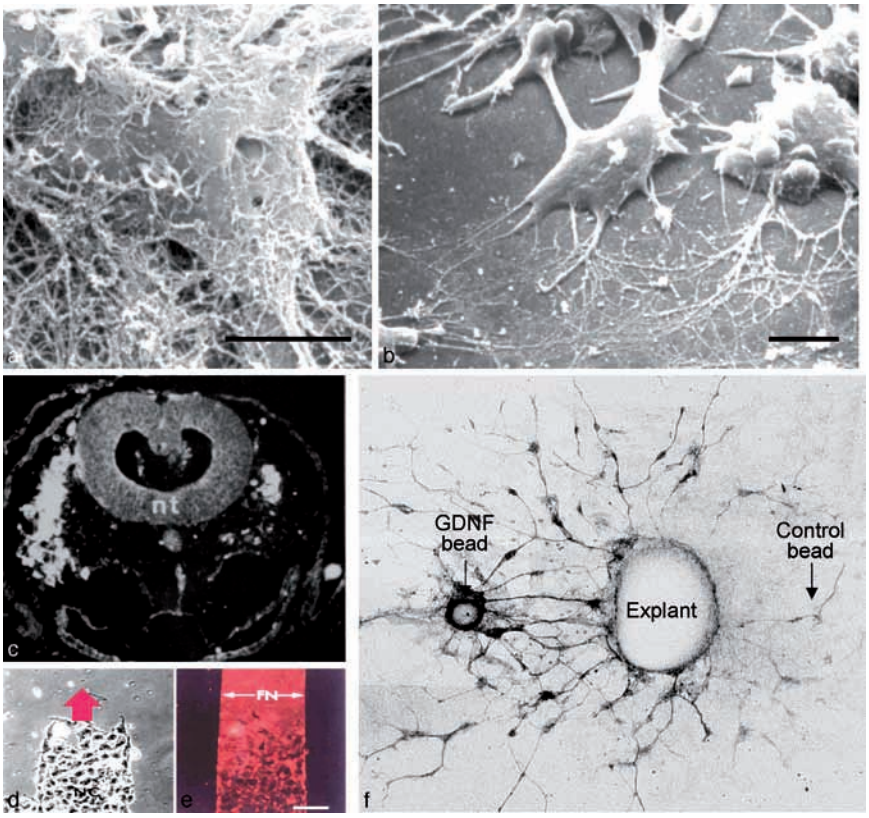

Figure 4 The microenvironment influences neural crest cell migration. (a) Scanning electron micrograph of neural crest cells *in vivo*, migrating in a fibrillar fibronectin extracellular matrix. (b) Fibrillar fibronectin strongly promotes neural crest cell migration in *in vitro* assays. (c) Transverse section of the head of a chick embryo injected on the right side (*) with a function-blocking antibody to the fibronectin receptor. The neural tube (nt) is labeled. Neural crest outgrowth (white area) labeled with the HNK-1 antibody to avian neural crest cells is reduced on the injected side. (d, e) Neural crest (NC) cells migrating *in vitro* (d, arrow) accurately follow stripes of fibronectin (FN); (e) shows same field shown in (d), but is immunolabeled (red) for fibronectin. (f) Enteric neural crest cells labeled with PGP9.5 antibody migrate *in vitro* from a gut segment explanted onto a collagen gel. This migration is strongly biased toward a bead loaded with glial-derived neuronal factor (GDNF), compared to migration toward a control bead. Scale bars = 20 μm (a), 20 μm (b), 50 μm (c). (a) From Newgreen DF (1985) Control of the timing of commencement of migration of embryonic neural crest cells. *Experimental Biology and Medicine* 10: 209–221; (c) from Bronner-Fraser M (1985) Alterations in neural crest migration by a monoclonal antibody that affects cell adhesion. *Journal of Cell Biology* 101: 610–617; (f) adapted from Young HM, Hearn CJ, Farlie PG, et al. (2001) GDNF is a chemoattractant for enteric neural cells. *Developmental Biology* 229: 503–516.

DNA-binding proteins MNR2, Lim3, and HB9. Few molecular markers that distinguish visceral from somatic motor neurons have been identified, but expression of BMP-5 appears to be specific for neurons in the Column of Terni. Although mechanisms underlying the segment-specific projection pattern observed in sympathetic preganglionic fibers are not completely understood, it appears that soluble signals present in the somitic mesoderm contribute to segment-specific identity. Correct patterning is necessary in order to ensure that preganglionic input is received by the appropriate postganglionic neurons. Following early phases of development, preganglionic inputs are necessary for regulation of neurotransmitter biosynthesis in postganglionic neurons.

Postganglionic sympathetic neurons are located in three anatomically distinct sets of ganglia. The majority of the principal sympathetic neurons are located in the paravertebral ganglia that are bilaterally distributed along each side of the spinal cord. The prevertebral ganglia are located at the midline; these cells lie anterior (ventral) to the dorsal aorta. The previsceral (terminal) ganglia, in a pattern more similar to parasympathetic ganglia, are situated in close proximity to several organs in the pelvis, mainly the bladder and rectum. Interestingly, although noradrenaline (norepinephrine) is the major neurotransmitter utilized by principal sympathetic ganglion neurons, similar to that found in the enteric nervous system, there is chemical coding of the neurotransmitters and neuropeptides expressed by sympathetic ganglion neurons. This chemical template is dependent upon the ganglia in which the cell bodies reside and the identity of their target. Neurons that innervate sweat glands (sudomotor) and the periosteum of bone are cholinergic; these represent 10–15% of sympathetic ganglion neurons. There is now general agreement that for some rather specialized neurons that neurotransmitter identity is plastic and changes with development once target innervation is complete. In addition to the now classic case of the rodent footpad, where the neurons are initially noradrenergic and become cholinergic in response to target-derived factors, it appears that the expression of neuropeptides in the prevertebral and previsceral ganglia also depends upon target for their expression. Neuropeptides, including neuropeptide Y (NPY) and vasoactive intestinal polypeptide (VIP), have been co-localized primarily with noradrenaline (NPY) and acetylcholine (VIP), and additional molecules such as substance P, dopamine, and serotonin have been found co-localized with these neurotransmitters in neurons of the previsceral and prevertebral chains.

Choice of neurotransmitter is an attribute that is acquired early in development and is dependent upon instructive cues encountered by the NC-derived precursor cells during their migration from the neural tube, as well as at sites along the dorsal aorta where these cells segregate into ganglia (Figure 5). There are two sources of instructive soluble signals required for appropriate migration and differentiation of NC-derived precursor cells as sympathetic ganglion neurons. The neural tube synthesizes and secretes an as yet unidentified member of the TGF-β family that is necessary for migration and differentiation as sympathetic neurons. BMP is a second required instructive factor synthesized and secreted by the dorsal aorta. BMP is an essential determinant of the noradrenergic phenotype. BMP is a proximal signal that induces expression of a network of DNA-binding proteins required for both neurogenesis and cell type-specific expression of noradrenergic marker genes. The core network of DNA-binding proteins that support differentiation of NC-derived precursor cells as sympathetic ganglion neurons includes the homeodomain (HD) proteins Phox2b and Phox2a, the basic helix–loop-helix (bHLH) DNA-binding proteins achaete-scute homolog 1 (MASH1 in mouse, CASH1 in chick), and HAND2, and the zinc finger protein GATA3 (GATA2 in chick). These proteins function together for both cell determination and differentiation; it is common to find that bHLH and HD proteins function together in networks of cross-regulated DNA-binding proteins. An essential function for both Phox2b and HAND2 has been demonstrated by gene knockout in mice. Deletion of Phox2b results in loss of autonomic neurons in each branch of the ANS; this DNA-binding protein is a master regulator of specification for the ANS. Deletion of HAND2 results in loss of sympathetic ganglion neurons with no apparent effect on parasympathetic ganglion neurons. Interestingly, loss of HAND2 affects migration and differentiation of NC-derived precursors that will contribute to the enteric nervous system. The differential effects on HAND2 in the generation of sympathetic, compared to enteric, neurons suggests that additional instructive cues necessary for specification of neurotransmitter phenotypic characteristics remain to be elucidated. The roles of various molecules in sympathetic, parasympathetic, enteric, and sensory neuron differentiation are illustrated in Figure 6.

Although neurotransmitter specification and expression are early developmental events, this aspect of phenotypic choice and expression is remarkably plastic. The adult neurotransmitter phenotype for some neurons is not achieved prenatally. Sympathetic ganglion neurons innervating the eccrine sweat glands and periosteum are noradrenergic prior to birth. Expression of noradrenergic characteristics does not depend upon target innervation; this sets

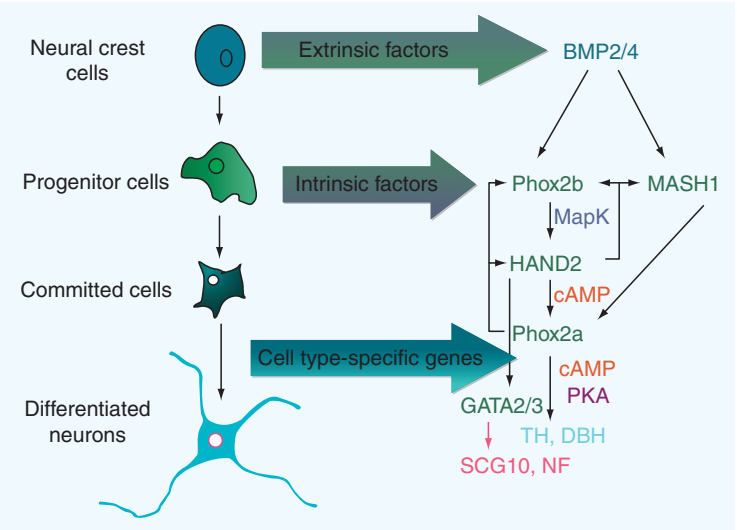

Figure 5 Specification and differentiation of peripheral autonomic neurons are dependent upon the interplay between cell extrinsic and cell intrinsic factors. Initial instructive cues from the neural tube influence neural crest cells that then respond to bone morphogenetic proteins (BMP2/4) derived from the dorsa aorta or retro-orbital mesenchyme. Induction of paired-like homeobox 2b (Phox2b) and mammalian achaete-scute homolog 1 (MASH1) proteins is followed by the induction of heart and neural crest derivatives-expressed protein 2 (HAND2) and Phox2a, resulting in expression of genes encoding pan-neuronal proteins (SCG10, superior cervical ganglion-10 protein (also called stathmin-like 2 protein); NF, neurofilament protein) and cell type-specific proteins (TH, tyrosine hydroxylase; DBH, dopamine β-hydroxylase; choline acetyltransferase and vesicular acetylcholine transporter, not shown). The recognized transcription factors required for the differentiation of noradrenergic sympathetic and cholinergic parasympathetic neurons are the same, with the notable exception of HAND2. Consequently, it is possible that this diversity could be in part the result of exclusive expression of HAND2 in precursors of noradrenergic sympathetic ganglion neurons. Cross-regulation of transcription factor expression suggests patterns of regulation based on generation of local gradients. MapK, mitogen-activated protein kinase; cAMP, cyclic adenosine monophosphate; PKA, protein kinase A; GATA2/3, GATA transcription factors 2 and 3. From Howard MJ (2005) Mechanisms and perspectives on differentiation of autonomic neurons. *Developmental Biology* 277: 271–286.

these neurons apart from other cholinergic neurons whose neurotransmitter phenotype is determined early in development and does not depend upon instructive target-derived signals. Interestingly, in the adult, the neurotransmitter phenotype of these neurons is altered following target innervation. This switch in neurotransmitter expression from noradrenergic to cholinergic depends upon retrograde transport of a target-derived factor that affects expression of a number of genes; the cholinergic differentiation factor has not been definitively identified, but is likely cardiotrophin 1. In addition to expression of choline acetyltransferase (ChAT), these neurons also express the vesicular acetylcholine transporter (VAChT), the choline transporter (ChT), as well as VIP. The expression of each of these molecules can be increased (induced) in sympathetic ganglion neurons by a variety of cytokines signaling through gp130. Expression of ChAT and VAChT is coordinately regulated and these two gene products comprise the cholinergic gene locus. Expression of both the cholinergic gene locus and VIP is regulated, in part, by the zinc finger DNA-binding protein REST. Although expression of ChAT, VAChT, ChT, and VIP appear to increase in parallel, relatively little is known about the underlying transcriptional mediators or how expression of these genes is coordinately regulated.

As development proceeds, sympathetic ganglion neurons become dependent upon nerve growth factor (NGF) for their maturation and survival. Precursor cells and young neurons express the neurotrophin receptor TrkC and have some dependence upon neurotrophin-3 (NT-3) for their survival. Mature postmitotic sympathetic ganglion neurons express TrkA and later acquire dependence upon NGF. These trophic factors are target derived and are retrogradely transported to the neuron somata, where they affect many cellular processes involved in cell death, cell survival, and maturation. Adult sympathetic ganglion neurons depend upon both NGF and NT-3. Together these trophic factors affect neuron survival. In addition, NGF influences axon targeting, the size and extent of dendritic arbors, and expression of the biosynthetic enzymes required for synthesis of noradrenaline. A little appreciated aspect of aging, dependent in part upon NGF and NT-3, is a decrease in levels of noradrenaline as well as a decrease in the innervation of many sympathetic ganglion target tissues, including heart, spleen, and cerebral blood vessels. Although not well understood, it appears that there are

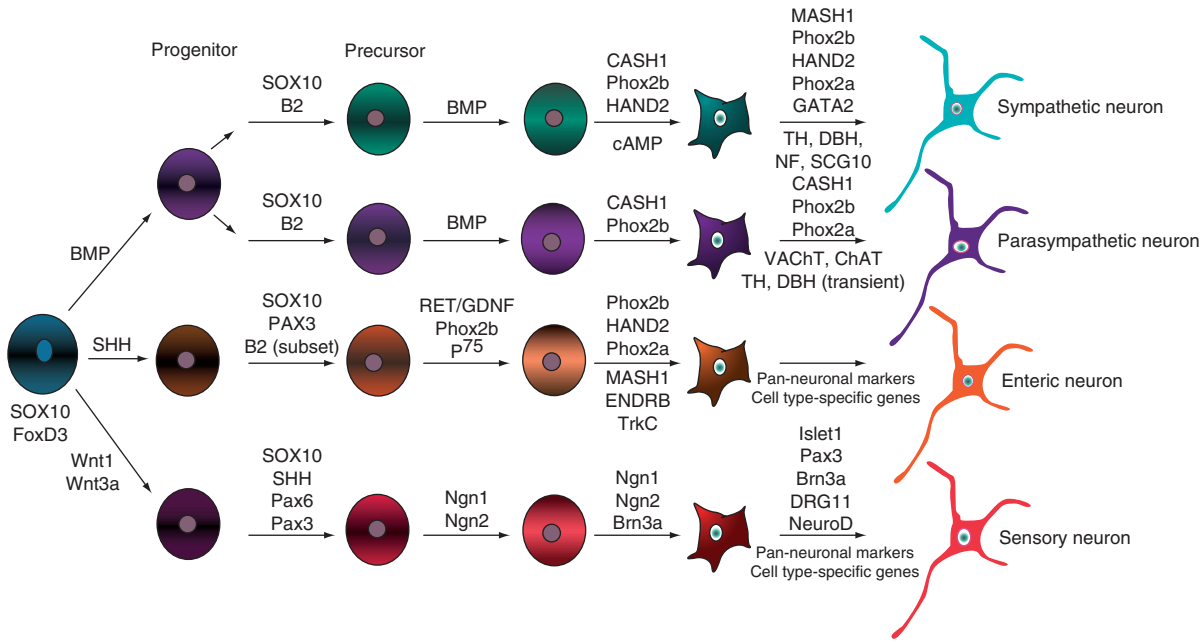

Figure 6 Analyses *in vitro* and *in vivo* have identified a series of growth factors and transcriptional regulators (markers) affecting different stages of neurogenesis of neural crest-derived progenitor cells. This schematic diagram summarizes work from many laboratories and includes some information not explicitly described in the body of the text. The compiled data provide a roadmap for following important hallmark events in the development of autonomic, enteric, and sensory neurons. Neural crest cells segregated from the neuroepithelium can be identified by the expression of FoxD3 and Sox10 (among other markers); cells expressing FoxD3 give rise to neurons and not melanocytes. Sox10 maintains multipotency in neural crest-derived cells as well as neurogenic potential. In the enteric nervous system, Sox10 and Pax3 together regulate Ret, which is required for normal development of these neurons. Progenitor cells differentiate into sympathetic, parasympathetic, enteric, or sensory neurons, in part dependent upon instructive signals encountered early at or near the time of egress from the neural tube. Additionally, extrinsic cues encountered during migration or at sites where neural crest-derived cells differentiate influence patterns of gene expression. In autonomic ganglia, expression of HAND2 appears to select cells as noradrenergic sympathetic ganglion neurons, as well as functioning in cell type-specific gene expression. In the sensory neuron lineage, the POU domain transcription factor Brn3a, expressed downstream of neurogenins 1 and 2 (Ngn1, Ngn2), regulates a large array of genes influencing cell death, neurotransmitter expression, and axon guidance. The signaling molecule sonic hedgehog is necessary for the expression of neurogenin. In the enteric nervous system, HAND2 is expressed downstream of Phox2b in all segments of the developing gut; the function of HAND2 in development of enteric neurons is unknown. The neurotrophin receptor TrkC is expressed early by neural crest-derived cells, the potential of which is restricted to neuronal or glial lineages as well as a subset of enteric neurons. From Howard MJ (2005) Mechanisms and perspectives on differentiation of autonomic neurons. *Developmental Biology* 277: 271–286.

alterations in the transcriptional and translational regulation of expression of both NGF and NT-3 coincident with the changes in function and survival of sympathetic ganglion neurons with increased age.

Parasympathetic Ganglia: Neuronal Differentiation and Connections

Of the components of the ANS, the least is known about parasympathetic ganglia because they are small, diffusely structured ganglia embedded within the tissues that they innervate (Figure 7). Parasympathetic ganglia are derived from the NC; however, the detailed axial origins of many of the ganglia are not known. The cranial parasympathetic ganglia, including the ciliary, lacrimal, otic, and sphenopalatine, arise from the cranial NC rostral to the otic (ear) placode. Caudal to this, the cardiac ganglion originates from the NC near the otic placode to the third somite, overlapping the region giving rise to the enteric nervous system (see later). Most of the trunk NC contributes the sympathetic ganglia (as previously mentioned), but the lumbosacral NC also gives rise to the parasympathetic pelvic ganglia and, in birds, the ganglion of Remak near the hindgut that is thought to be an extension of the pelvic ganglia (these origins are diagrammed in Figure 1).

The specification and differentiation of parasympathetic ganglia have many pathways in common with sympathetic ganglia (see Figure 6): both rely on signaling by BMP and require the expression of the transcription factors MASH1, Phox2a, and Phox2b. Specification of the cholinergic phenotype typical of parasympathetic neurons may occur prior to migration and is reinforced by suppression of the

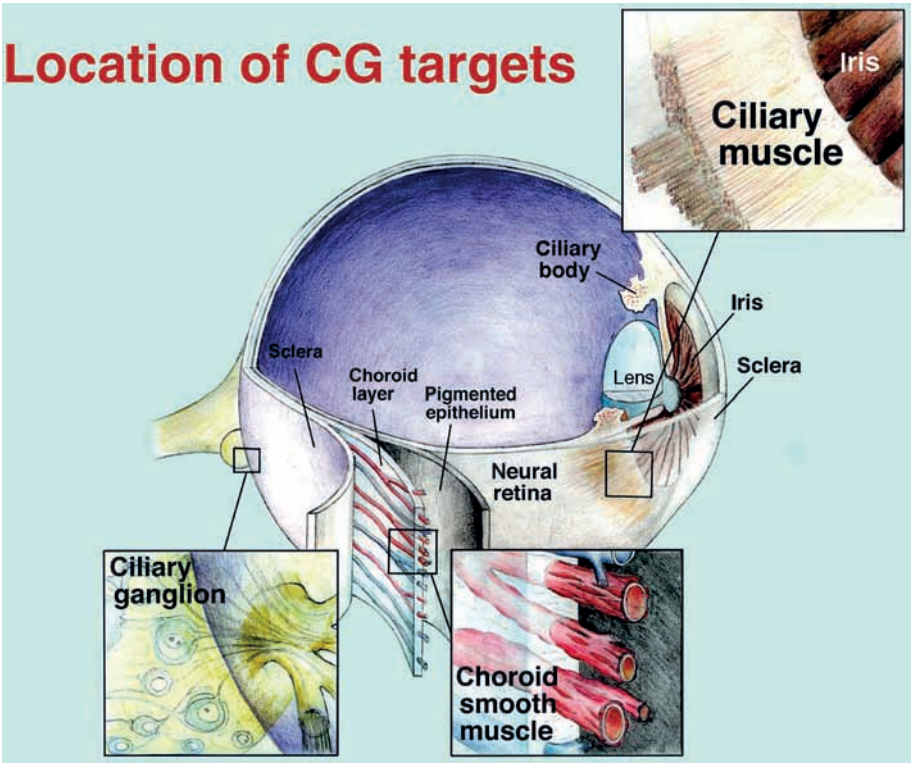

Figure 7 The ciliary ganglion (CG) and its targets of innervation, illustrating the relatively close relationship between parasympathetic ganglia and their innervation targets. The schematic shows the location of the ciliary ganglion and its target tissues, the vascular smooth muscle of the choroid layer, and the striated muscle of the iris and ciliary body. The enlargements show the details of the histological structures of the tissues. The ciliary ganglion contains small, unmyelinated neurons that innervate the vascular smooth muscle (choroid neurons) and large, myelinated neurons that innervate the iris and ciliary muscle (ciliary neurons).

transcription factor HAND2 (dHAND), which is necessary for the expression of catecholamines. In the examples that have been best studied (ciliary ganglion, cardiac ganglion, and submandibular ganglion), the precursors of parasympathetic ganglia migrate to their targets of innervation prior to the onset of organogenesis of the particular organ. As morphogenesis commences, neuronal differentiation occurs. For example, as the salivary gland epithelium branches and forms ducts, many of the neurons associated at the base of the organ rudiment are already postmitotic and extend axons as epithelial buds form (Figure 8). Likewise, cardiac ganglion neurons undergo neurogenesis within the developing heart, and ciliary ganglion neurons elaborate axons as the optic vesicle forms the eye and its associated structures.

This intimate differentiation of parasympathetic neurons along with their targets of innervation suggests that trophic interactions reciprocally guide development. In fact, the expression of GDNF and the closely related molecule neurturin in target tissues is essential for the differentiation and development of a number of parasympathetic ganglia. These include ciliary, otic, sphenopalatine, submandibular, lacrimal, penile, and pancreatic islet neurons. In many instances, differentiation of parasympathetic ganglia appears to require sequential action of GDNF, followed by neurturin.

By far the best-studied parasympathetic ganglion is the avian ciliary ganglion (see Figure 7). This ganglion contains two populations of principal neurons: ciliary neurons, which innervate the iris and ciliary muscle, and choroid neurons, which innervate the arterial smooth muscle in the choroid layer. Both populations differentiate, innervate their respective target tissues, and undergo programmed cell death in midgestation (between E6 and E14 of chick development). Target tissues control the degree of programmed cell death that occurs as well as other aspects of neuronal differentiation, such as the expression of the neuromodulatory neuropeptide somatostatin, and the expression of ion channels, such as the calcium-activated potassium channel and the nicotinic acetylcholine receptor (Table 2). Afferents from the accessory oculomotor nucleus of the midbrain also regulate differentiation

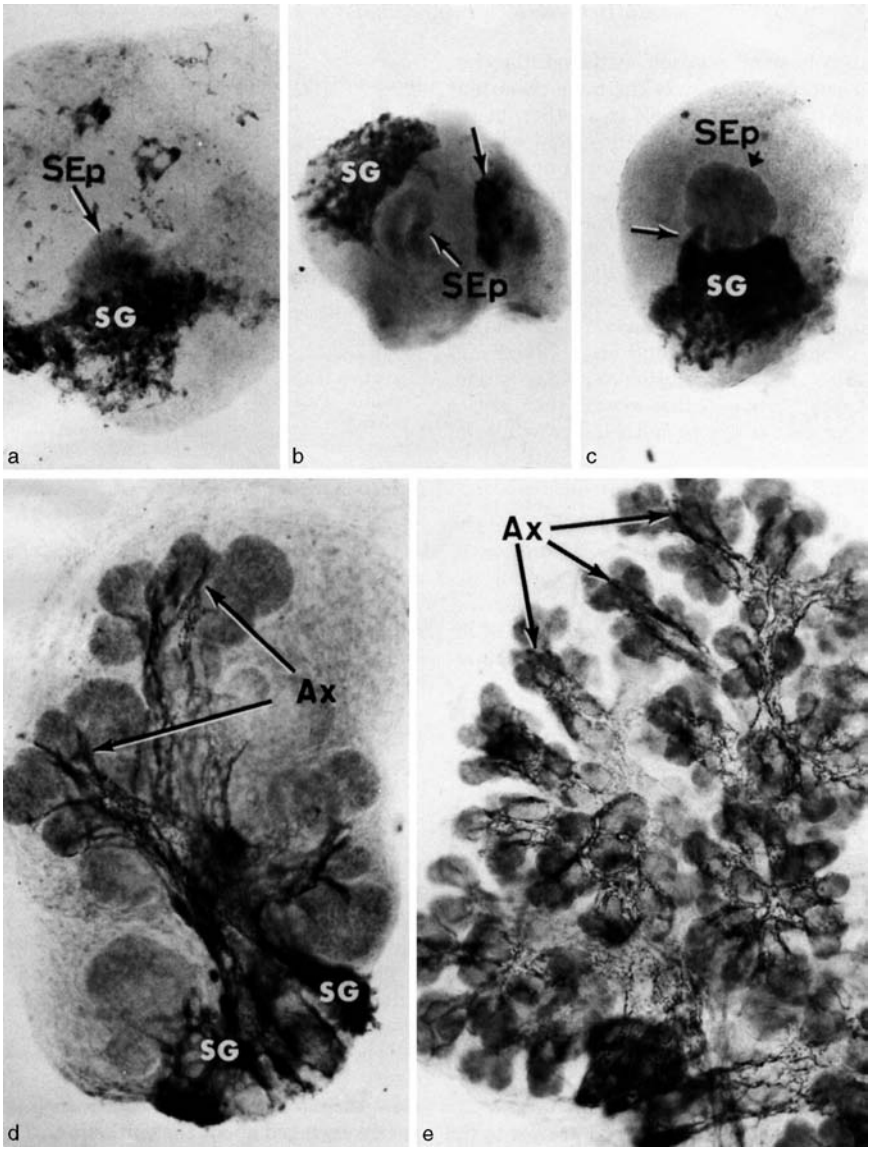

Figure 8 Early development of parasympathetic nerves matches the morphogenesis of the target, shown in whole mounts of the submandibular salivary gland stained for acetylcholinesterase activity. (a) At E12, the salivary gland epithelial bud (SEp) is surrounded by darkly staining salivary ganglion cells (SG); (b) E12, looking down on the preparation shown in a; (c) early E13, when the lobule of the SEp has just begun to form clefts (arrowhead) and axons begin to extend into the epithelium (arrow); (d) E14 gland; the epithelium has grown and has begun to branch extensively, and axons (Ax) course over the epithelium and travel in the clefts between the buds; (e) a portion of an E15 gland; axon outgrowth continues to parallel the growth of the epithelium. From Coughlin MD (1975) Early development of parasympathetic nerves in the mouse submandibular gland. *Developmental Biology* 43: 123–139.

of the ciliary ganglion neurons. These inputs influence cell death and differentiation through nicotinic cholinergic activation of ciliary neurons. Recently, nicotinic activation has been tied to the expression of a chloride transporter, which changes the magnitude of the chloride gradient across the plasma membrane, thereby influencing γ-aminobutyric acid (GABA)ergic signaling in neurons as well as axonal morphology. Whether the principles of development uncovered by studies of the ciliary ganglion apply to other parasympathetic ganglia is a matter of ongoing investigation.

Enteric Ganglia: Neuronal Differentiation and Connections

The enteric nervous system (ENS) is by far the largest division of the ANS, and the most complex. The enteric nervous system consists of numerous small ganglia placed as nodes in a lattice of interconnections. Two

Table 2 Anterograde and retrograde influences of development in the avian ciliary ganglion

Effector molecule	*Mode*	*Neuronal property*	*Effect*
ACh (all nAChRs)	Anterograde	Chloride transporter	Increase
Neuregulin	Anterograde	Ca^{2+}-activated K^+ channels	Increase
Neuregulin	Anterograde	nAChRs	Increase
ACh (α7 nAChRs)	Anterograde	Neuronal survival	Decrease
ACh (all nAChRs)	Anterograde	Neuronal survival	Increase
CNTF	Retrograde	Neuronal survival	Increase
GDNF	Retrograde	Neuronal survival	Increase
TGF-β4	Retrograde	Ca^{2+}-activated K^+ channels	Increase
TGF-β3	Retrograde	Ca^{2+}-activated K^+ channels	Decrease
Activin	Retrograde	Somatostatin	Increase
Unknown	Retrograde	nAChRs	Increase

such layers occur, the myenteric plexus and the submucosal plexus, with extensive radial connections. The ENS has a full reflex circuitry with sensory, motor, and interneurons, and to a degree can function without CNS input, which derives from visceral motor neurons in vagal and sacral levels of the CNS (the same levels from which the enteric NC precursors arise). At least 15 neurotransmitters (but not noradrenaline) and neuromodulatory peptides occur in mature enteric ganglia, and each small ganglion has several neuron types differentiated by neurochemical code, morphology, and projection pattern. How so many different neuronal types are specified in each ganglion is not well understood. As explored by Michael Gershon and co-workers, differences in timing of differentiation of different neuron classes could be involved, and neuron type-related differences in median birth days does occur, as determined by ascertaining neuronal birth days using pulse-delivered tritiated thymidine or bromodeoxyuridine (BrdU). However, neurons of each class are born over a time range, and the ranges for different classes overlap strongly. Moreover, some neuron types, such as 5-hydroxytryptamine neurons, are specified prior to or early in NC migration. As with other parts of the ANS, Phox2b is a decisive requirement, and deletion of HAND2 has revealed its essential role in the specification and differentiation of enteric nervous system neurons that express VIP.

Differentiation of enteric neurons has similarities to that of the neurons of the rest of the ANS (Figure 6). For example, it is marked by the loss of the transcription factor Sox10 and the appearance of the neuron-specific transcription factor family member HuC/D. Sox10 remains present in enteric precursors and enteric glia, and the latter also express glial fibrillary acidic protein (GFAP) like CNS glia, but unlike glia of the other ANS divisions. One of the first neuron classes to appear is nitric oxide synthase-expressing neurons, and in mice these cells are initially transiently catecholaminergic, showing dopamine (DOPA) metabolism such as DOPA decarboxylase and dopamine-β-hydroxylase. These early differentiating neurons cease to migrate but extend axons distally, in parallel with the migration of vagal enteric NC cells.

Unlike the rest of the ANS and CNS, the enteric neuron population size does not normally overshoot and then undergo apoptotic pruning. Instead, the population proliferates up to a density set by the gut tissue. Instrumental in this so-called logistical growth is GDNF. This factor, via its receptor Ret on NC cells and its co-receptor GDNF family receptor-α (GFR-α) has a complex role, since it is not only a mitogen for enteric NC cells, but also induces differentiation into postmitotic neurons. *In vitro* experiments suggest that the small peptide endothelin-3, via its receptor EdnrB on NC cells, retards the differentiation effect of GDNF while preserving or even increasing its mitogenic function. Mutation of all the genes for these can cause Hirshsprung's disease (Table 1) either by directly decreasing GDNF mitogenic signaling (for mutations in *RET*, *GFRA1*, and *GDNF*) or by removal of the brake on GDNF-stimulated mitotic withdrawal (for mutations in *EDN3*, *EDNRB*, and *ECE1*).

See also: Autonomic Nervous System: Neuroanatomy; Autonomic Nervous System; Enteric Nervous System Development; Enteric Nervous System: Neurotrophic Factors; Nerve Growth Factor; Neural Crest.

Further Reading

Anderson RB, Newgreen DF, and Young HM (2006) Neural crest and the development of the enteric nervous system. *Advances in Experimental Medicine and Biology* 589: 181–196.

Bertrand N, Castro DS, and Guillemot F (2002) Proneural genes and the specification of neural cell fates. *Nature Reviews Neuroscience* 3: 517–530.

Bronner-Fraser M (1985) Alteration in neural crest migration by monoclonal antibody that affects cell adhesion. *Journal of all Biology* 101: 610–617.

Caughlin MD (1975) Early development of parasympathetic nerves in the mouse submandibalar gland. *Developmental Biology* 43: 123–139.

Enomoto H, Heuckeroth RO, Golden JP, et al. (2000) Development of cranial parasympathetic ganglia requires sequential actions of GDNF and neurturin. *Development* 127: 4877–4889.

Guillemot F, Lo L-C, Johnson JE, et al. (1993) Mammalian achaete-scute homolog-1 is required for the early development of olfactory and autonomic neurons. *Cell* 75: 463–476.

Howard MJ (2005) Mechanisms and perspectives on differentiation of autonomic neurons. *Developmental Biology* 277: 271–286.

Kasemeier-Kulesa JC, Kulesa PM, and Lefcort F (2005) Imaging neural crest cell dynamics during formation of dorsal root ganglia and sympathetic ganglia. *Development* 132: 235–245.

LeDouarin NM, Creuzet S, Couly G, et al. (2004) Neural crest cell plasticity and its limits. *Development* 131: 4637–4650.

Muller F and Rohrer H (2002) Molecular control of ciliary neuron development: BMPs and downstream transcriptional control in the parasympathetic lineage. *Development* 129: 5707–5717.

Newgreen DF (1985) Control of the timing of Commencement of migration of embryonic neural crest cells. *Experimental Biology and Medicine* 10: 209–221.

Newgreen DF and Erickson CA (1986) The migration of neural crest cells. *International Review of Cytology* 103: 89–145.

Nishi R (2003) Target-mediated control of neural differentiation. *Progress in Neurobiology* 69: 213–227.

Simpson MJ, Landman KA, Hughes BD, et al. (2006) Looking inside an invasion wave of cells using continuum models: Proliferation is the key. *Journal of Theoretical Biology* 243: 343–360.

Stanke M, Duong CV, Pape M, et al. (2006) Target-dependent specification of the neurotransmitter phenotype: Cholinergic differentiation of sympathetic neurons is mediated *in vivo* by gp130 signaling. *Development* 133: 141–150.

Young HM, Bergner AJ, Anderson RB, et al. (2004) Dynamics of neural crest-derived migration in the embryonic mouse gut. *Development Biology* 270: 455–473.

Young HM, Hearn CJ, Farlie PG, et al. (2001) GDNF is a Chemo attractant for enteric neural cells. *Developmental Biology* 229: 503–516.

Autonomic Nervous System: Cardiovascular Control

J A Armour, University of Montreal, Montreal, QC, Canada

© 2009 Published by Elsevier Ltd.

Introduction

The function of the neuronal hierarchy involved in cardiovascular control is ultimately to match cardiac output to whole body blood flow demands. The peripheral and central components that make up this hierarchy interact on an ongoing basis to coordinate regional cardiac indices such that regional demands for blood flow are met in an efficient manner. Dysfunction of any population within this hierarchy can result in cardiac or vascular malfunction, such as the induction of atrial or ventricular arrhythmias or the genesis of essential hypertension.

The cardiovascular neuronal hierarchy is a massively parallel and, for the most part, stochastic control system such that stable control usually occurs in the absence of obvious cause and effect. Apparently that is because the individual components that make up this hierarchy display emergent properties, while interacting in a highly optimized fashion to tolerate normal cardiovascular perturbations. On the other hand, the functional interconnectivity of the hierarchy is so organized that the whole can be catastrophically disabled by cascading failures initiated by sometimes relatively minor abnormal inputs. In this article, an overview of the anatomy and the function of its afferent and efferent components are presented first, followed by how these neurons putatively interact to coordinate autonomic motor neuron outputs to the heart and peripheral vasculature. It is proposed that defining the relevance of each of its populations may be essential to understanding the role of the whole during the evolution of cardiovascular disease.

Peripheral Autonomic Neurons Regulating the Heart and Major Blood Vessels

Cardiovascular Afferent Neurons

Pain associated with myocardial ischemia is frequently referred to a patient's left upper limb and/or anterior thoracic wall. Thus, it has been assumed that most cardiac afferent neurons are located in left-sided, cranial thoracic dorsal root ganglia. Anatomical evidence indicates that cardiac afferent neurons are distributed relatively evenly throughout both nodose ganglia and the right and left C7 to T4 dorsal root ganglia. Cardiac afferent neurons are also located in intrathoracic ganglia, including those intrinsic to the heart. The sensory neurites of such neurons are distributed in the walls of the major veins adjacent to the heart, all four chambers of the heart, and along the arch of the aorta. Select ones also lie concentrated in the region of the carotid bulbs.

Cardiac Motor Neurons

Cardiac parasympathetic efferent neurons The somata of parasympathetic efferent preganglionic neurons that synapse with cholinergic efferent postganglionic neurons on the heart are located primarily ventral lateral to the nucleus ambiguous of the medulla; lesser numbers are located in the dorsal motor nucleus and the intermediate zone in between these two medullary nuclei. Cardiac efferent preganglionic neurons in individual medullary loci project axons to parasympathetic efferent postganglionic neurons distributed throughout all major atrial and ventricular ganglionated plexuses.

Cardiac sympathetic efferent neurons Sympathetic efferent preganglionic neurons in the spinal cord that are involved in cardiac regulation project axons via the T1 to T5 rami to synapse with cardiac sympathetic efferent postganglionic neurons located in the cranial poles of the stellate ganglia, throughout the right and left middle and superior cervical ganglia, as well as in mediastinal ganglia adjacent to the heart. They also project to sympathetic efferent postganglionic neurons in intrinsic cardiac ganglia.

Intrathoracic Local Circuit Neurons

Interneurons are interposed between cardiac afferent and efferent neurons in the various intrathoracic ganglia. As these project axons to neurons not only within their own ganglia, but also those in other intrathoracic ganglia, they have been called local circuit neurons. Some of these larger-diameter neurons (i.e., up to ~30 μm) form rosettes within intrathoracic ganglia. The somata of many of them are frequently found in the periphery of an intrinsic cardiac ganglion, their centrally projecting dendrites lying adjacent to one another to form synaptic connections with other neurons of that rosette. Such anatomy may represent the substrate for local information processing within one ganglion. Many intrathoracic neurons, including those on the heart, display immunoreactivity to multiple peptides. These data indicate that intrathoracic cardiac neurons can be involved in a multiplicity of interactions that collectively coordinate regional cardiac indices (cf. section on cardiac neuroaxis).

Peripheral vascular sympathetic efferent neurons Sympathetic efferent preganglionic neurons in the spinal cord synapse with sympathetic efferent postganglionic neurons located not only in the superior cervical, middle cervical, and stellate ganglia, but abdominal sympathetic chain and midline ganglia. Many of their sympathetic efferent postganglionic neurons project axons to the major vessels (arteries and veins) in the limbs, for instance, to control regional blood flow throughout the body.

Physiology of Cardiovascular Neurons

Cardiovascular Afferent Neurons

Much of the activity generated by neurons throughout the cardiovascular neuroaxis is dependent on tonic inputs arising from the various cardiovascular sensory neurons depicted in the earlier sections. Thus, the activity generated by many central and peripheral cardiac neurons is ultimately reflective of the transduction capabilities of cardiac and major vascular mechano- and/or chemosensory neurons. For instance, most (~75%) nodose ganglion cardiac afferent neurons transduce chemical stimuli. About 35% respond to mechanical stimuli while fewer (~10%) are multimodal in nature, responding to mechanical and chemical stimuli. In contrast, the vast majority of dorsal root ganglion cardiac afferent neurons (~95%) are multimodal in nature. That capacity permits ventricular sensory neurites associated with individual dorsal root ganglion afferent neurons to transduce not only regional ventricular dynamics (Figure 1), but altered release of local chemicals such as occurs during ventricular ischemia as well. The varied transduction capabilities displayed by individual dorsal root ganglion afferent neurons presumably permit such a relatively limited population of afferent neurons to simultaneously transduce multiple cardiac signals to spinal cord neurons.

Cardiac sensory neurites are also associated with afferent neuronal somata in intrathoracic extracardiac (Figure 2) and intrinsic cardiac ganglia. Most of these are multimodal in nature, responding thereby to regional ventricular ischemia. Limited populations of intrathoracic cardiac afferent neurons also transduce the postischemia (reperfusion) state, becoming activated early on during reestablishment of regional coronary arterial blood flow.

With respect to cardiac chemotransduction, adenosine is known to be released in increasing quantities from the ischemic myocardium. This chemical activates the sensory neurites associated with many ischemia-sensitive cardiac afferent neurons. In fact, it has been proposed that purinergic dorsal root ganglion cardiac afferent neurons are involved in the genesis of symptoms in patients subjected to myocardial ischemia.

Numerous mechanosensory neurites associated with other intrathoracic afferent neuronal somata are located on the major intrathoracic vessels or coronary arteries. These transduce local vascular wall deformation subsequent to underlying blood pressure changes

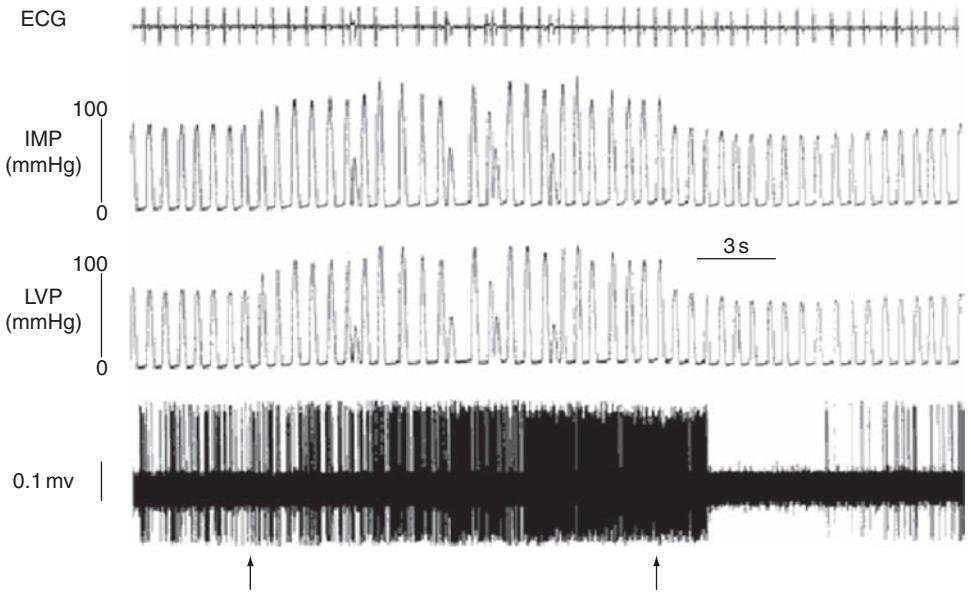

Figure 1 Behavior of a T3 left-sided dorsal root ganglion afferent neuron associated with left ventricular mechanosensory neurites in response to altered left ventricular dynamics. In this anaesthetized canine preparation, afferent neuronal activity was increased by partial occlusion of descending aorta (between arrows, below), as left ventricular intramyocardial (IMP) and chamber (LVP) systolic pressures rose. The lowest trace represents the afferent neuronal activity. ECG, electrocardiogram; LVP, left ventricular cavity pressure.

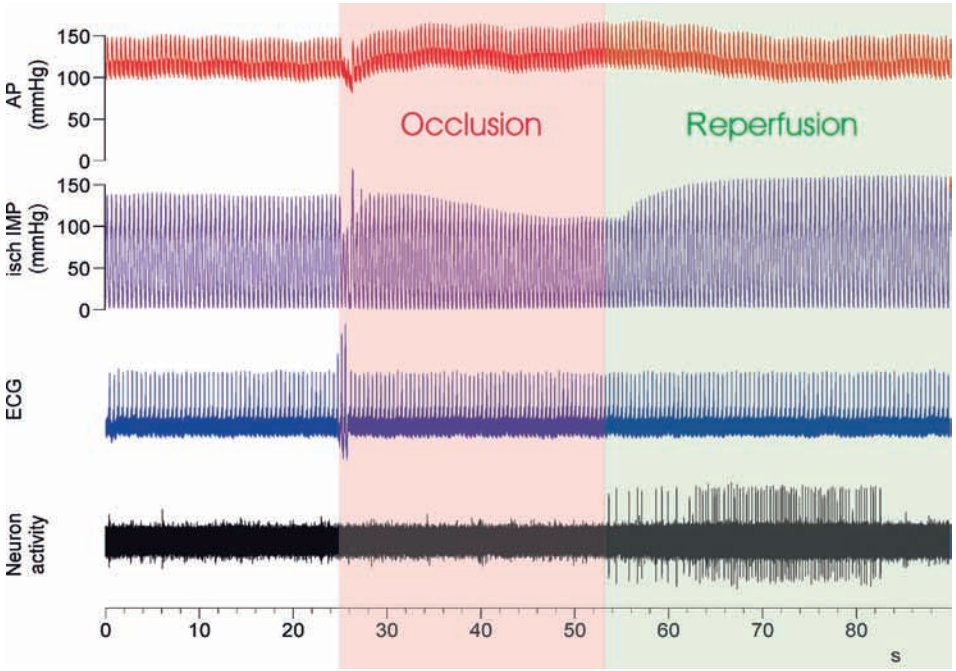

Figure 2 Left-sided middle cervical ganglion afferent neurons associated with left ventricular sensory neurites were activated in an anesthetized canine immediately following (early reperfusion; green section), but not during occlusion of its left anterior descending coronary artery. Occlusion (pink section lasting for 28 s) reduced left ventricular wall regional dynamics (ischemic LV intramyocardial systolic pressure), while aortic pressure (AP) remained stable. The lowest trace represents raw middle cervical ganglion neuronal activity. (These data are presented in Acrobat Distiller format for video viewing with sound.)

with the degree of fidelity that many carotid artery baroreceptor neurons projecting to the nucleus solitarius exhibit. In sum, cardiovascular afferent neuronal somata located in various ganglia continuously transduce not only arterial dynamics (i.e., carotid artery and aortic baroreceptors), but also the heart's regional mechanical and chemical environment.

Intrathoracic Local Circuit Neurons

The cardiac and major vascular milieu transduced by intrathoracic afferent neurons influence some cardiac motor neurons on a beat-to-beat basis, primarily via intrathoracic local circuit neurons. Intrathoracic, extracardiac, and intrinsic cardiac local circuit neurons process information arising not only from cardiac and major intrathoracic vascular mechanosensory and/or chemosensory afferent neurons, but also descending information from sympathetic (spinal cord) and parasympathetic (medullary) efferent preganglionic neurons. Due to the fact that some receive direct inputs from cardiac and aortic mechanosensory neurons, they may display activity that is phase-related to the cardiac cycle. Their capacity to processing cardiac sensory inputs persists even after all of their central neural connections have been chronically eliminated. That is one reason why it has been proposed that much of the daily transduction of cardiovascular milieu to cardiac motor neurons can occur independent of the central nervous system, relying to a considerable degree on intrathoracic local circuit neurons.

Cardiac Motor Neurons

Parasympathetic efferent neurons Medullary efferent preganglionic neurons innervate parasympathetic postganglionic neurons distributed throughout the atrial and ventricular ganglionated plexuses. When activated, these cholinergic motor neurons suppress not only atrial rate and force and atrioventricular nodal conduction, but also ventricular contractile force. Cardiac cholinergic efferent neurons are under the particular, but not exclusive, reflex control of arterial mechanosensory inputs (particularly those arising from the carotid artery mechanosensory neurites). Thus, some of these neurons generate activity that reflects changing arterial pressure. Indeed, their effectiveness in modulating cardiac rate frequently relates to arterial dynamics as a result of the short-latency reflexes that they initiate.

Sympathetic efferent neurons Sympathetic efferent postganglionic neurons in each intrathoracic ganglion receive inputs from sympathetic efferent preganglionic neurons in the caudal-cervical and cranial-thoracic spinal cord. These postganglionic neurons, in turn, innervate widely divergent regions of the heart. Cardiac

adrenergic motor neurons also receive inputs from intrathoracic sensory neurons via intrathoracic local circuit neurons. Adrenergic motor control of cardiac chronotropism, dromotropism, and regional inotropism ultimately depends upon the integration of cardiovascular sensory and central neuronal inputs via complex information processing within the intrathoracic neuroaxis.

Vascular Motor Neurons

Control of blood flowing through peripheral arteries and veins resides in intrinsic (local metabolic effects) and extrinsic (neurohumoral) mechanisms. Classically, the control of the peripheral circulation resides primarily with local factors (cf. locally released metabolites). Sympathetic efferent neurons that innervate arteries and veins act to constrict most major vessels. An exception to this rule is the control that neurons exert over arteries supplying oxygenated blood to skeletal muscles whereby, when activated, vasodilatation occurs that increases local perfusion.

In the coronary circulation, intrinsic mechanisms predominate. In the skin extrinsic control predominates, while in skeletal muscles both are important. Arterial baroreceptors concentrated on the carotid arteries and along the arch of the aorta, acting via medullary, spinal cord, and intrathroacic reflexes, play a key role in short-term adjustments of arterial pressure. Central chemosensory neurons, when activated, for instance, by local hypoxemia, exert profound influences on arterial resistance and venous capacitance, in addition to cardiac work. In order to maintain normal cardiovascular status, a balance of these cardiac and vascular reflex forces must be achieved via ongoing neurohumoral balance.

The Cardiovascular Neuroaxis

The target organ nervous system constantly interacts with neurons in intrathoracic and cervical ganglia, as well as central neurons, to process cardiovascular sensory information via multi-tiered reflexes. Excitatory and inhibitory reflexes within the cardiac neuroaxis coordinate cardiac motor neuronal outputs that regulate regional cardiac indices on a beat-to-beat basis. Neurons in the intrathoracic neuronal hierarchy interact via a host of chemicals that include acetylcholine, butyrylcholine, alpha- and beta-adrenoceptor agonists, histamine, nitric oxide donors, peptides such as angiotensin II (Figure 3), purinergic agents (adenosine and adenosine 5′-triphosphate (ATP)), excitatory and inhibitory amino acids, and serotonin.

Some intrinsic cardiac local circuit neurons receive inputs from both sympathetic and parasympathetic efferent preganglionic neurons. In other words, populations of neurons on the heart process inputs from both 'limbs' of the autonomic efferent nervous system and not necessarily in a reciprocal fashion. Intrathoracic local circuit neurons receive indirect inputs via spinal cord neurons derived from sensory neurites in extrathoracic tissues (i.e., in tissues of the neck or each of the four limbs) such that alterations in the extrathoracic milieu can influence the intrinsic cardiac nervous system.

In sum, it has been proposed that the mammalian intrinsic cardiac nervous system functions collectively as a 'little brain,' processes these multiple inputs on an ongoing basis in the coordination of regional cardiac indices. Short-latency reflexes (40–100 ms) involving neurons within and near the target organ influence cardiac motor neurons throughout each

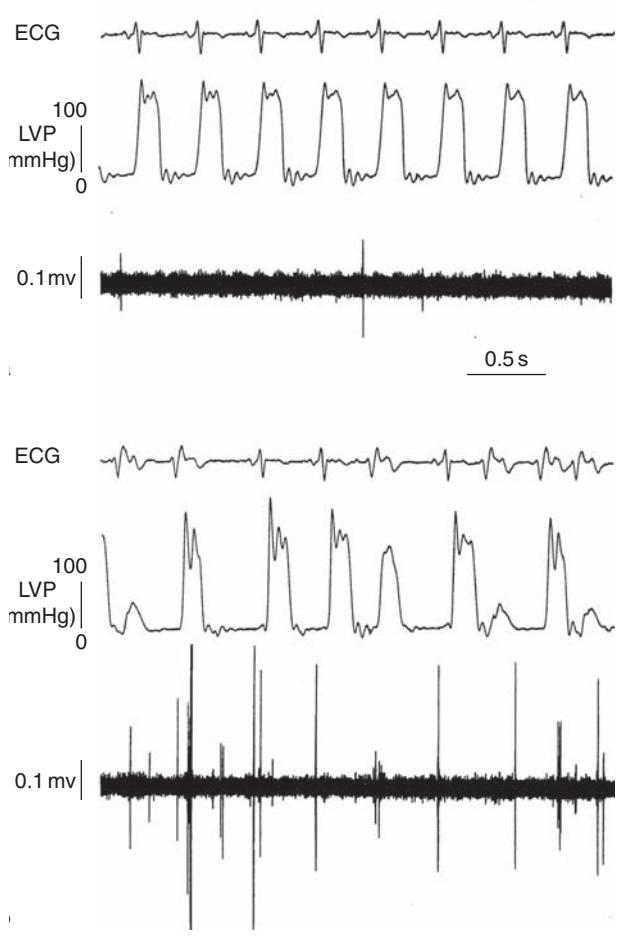

Figure 3 Angiotensin II-induced changes in intrinsic cardiac neuronal activity in an anesthetized, canine preparation. Spontaneous activity generated by canine right atrial neurons *in situ* was enhanced following local application of angiotensin II (10 μl of a 10 μm solution over a 5 s period between panels). Note that previously inactive neurons became active and that ventricular ectopy was induced. ECG, Lead II electrocardiogram; LVP, left ventricular cavity pressure. The lowest trace represents neuronal activity.

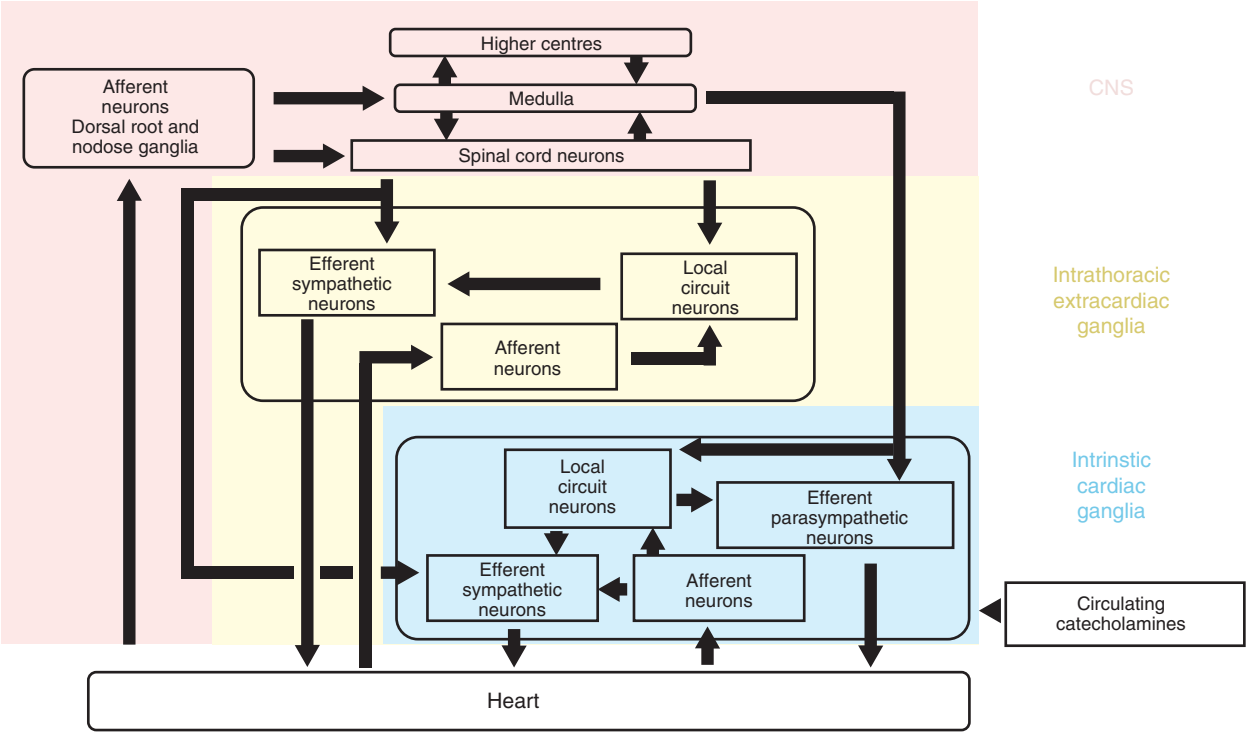

Figure 4 Schematic representation of the locations of afferent, local circuit, and efferent (parasympathetic and sympathetic) neurons in peripheral autonomic ganglia involved in regulating cardiac function, along with some of their putative interconnections. This peripheral neuronal hierarchy is under the tonic influence of medullary and spinal cord neurons, as well as circulating catecholamines.

phase of the cardiac cycle. On the other hand, longer latency cardiocardiac reflexes (300 ms to 2 s) involving central neurons influence cardiodynamics over longer timescales, that is, during subsequent cardiac cycles.

Clinical Implications

All of the above indicate that the cardiac neuroaxis can be represented as a massively parallel and, for the most part, stochastic control system that is involved in transducing the cardiac mechanical and chemical milieu to cardiac motor neurons on a beat-to-beat basis (Figure 4). The varied intrathoracic neuronal interactions so engendered are under the tonic influence of central cardiovascular reflexes that rely to a considerable extent on arterial baroreceptor initiated reflexes. The unforeseen relevance of such intrathoracic reflexes lies in their stability such that relatively little demand may be placed on central neurons in the daily maintenance of adequate cardiac output. In fact, the interactions that occur among their varied neural elements appear to be optimized to tolerate normal cardiac perturbations. As individual units of the whole collectively display emergent properties, the whole cannot be represented as a simple positive and negative reflex control system.

On the other hand, such an organizational arrangement means that it can be catastrophically disabled by cascading failures initiated by relatively minor abnormal inputs as that is not a design feature of the system. Reductions in regional coronary arterial blood supply affect the function of various intrinsic cardiac neuronal populations in either a direct (ischemic damage) or indirect (altered sensory inputs) manner. The latter occurs when regional cardiac sensory inputs arising from an ischemic myocardium are transduced throughout this hierarchy. Excessive activation of some populations may have devastating effects on the capacity of cardiac motor neurons to coordinate regional cardiac indices. For instance, it has been proposed that excessive activation of inputs to the intrinsic cardiac nervous system may be a predisposing factor for cardiac arrhythmia formation (cf. Figure 3). Because neurons within this hierarchy communicate in part via angiotensin II or β-adrenergic receptors, therapy targeting these receptors may indeed modify cardiomyocyte function not only directly but also indirectly by influencing the cardiac neuroaxis.

See also: Autonomic Nervous System: Central Cardiovascular Control; Cardiovascular Function: Central Nervous System Control; Cerebrovascular Disease; Cotransmission; Noradrenaline; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Adams DL and Cuevas J (1994) Electrophysiological properties of intrinsic cardiac neurons. In: Armour JA and Ardell JL (eds.) *Neurocardiology*, pp. 1–60. New York: Oxford University Press.

Andresen MC, Kuntz DJ, and Mendelowitz D (2004) Central nervous system regulation of the heart. In: Armour JA and Ardell JL (eds.) *Basic and Clinical Neurocardiology*, pp. 187–219. New York: Oxford University Press.

Ardell JL (1994) Anatomy and function of mammalian intrinsic cardiac neurons. In: Armour JA and Ardell JL (eds.) *Neurocardiology*, pp. 95–114. New York: Oxford University Press.

Armour JA (1991) Anatomy and function of the intrathoracic neurons regulating the mammalian heart. In: Zucker IH and Gilmore JP (eds.) *Reflex Control of the Circulation*, pp. 1–37. Boca Raton, FL: CRC Press.

Gray AL, Johnson CI, Ardell JL, and Massari VJ (2004) Parasympathetic control of the heart. A novel interganglionic intrinsic cardiac circuit mediates neural control of the heart. *Journal of Applied Physiology* 96: 2273–2278.

Levy MN and Martin PJ (1979) Neural control of the heart. In: Berne RM (ed.) *Handbook of Physiology: The Cardiovascular System*, pp. 581–620. Bethesda, MD: American Physiological Society.

McAllen RM and Spyer KM (1976) The location of cardiac vagal preganglionic motorneurons in the medulla of the cat. *Journal of Physiology* 258: 187–204.

Randall WC, Wurster RD, Randall DC, and Xi-Moy S (1996) From cardioaccelerator and inhibitory nerves to a heart brain: An evolution of concepts. In: Shepherd JT and Vatner SF (eds.) *Nervous Control of the Heart*, pp. 173–200. Amsterdam: Harwood Academic Publishers.

Steele PA, Gibbins IL, Morris JL, and Mayer B (1994) Multiple populations of neuropeptide containing intrinsic neurons in the guinea-pig heart. *Neuroscience* 62: 241–250.

Sylvén C (2004) The genesis of pain during myocardial ischemia and infarction. In: Armour JA and Ardell JL (eds.) *Neurocardiology*, pp. 298–314. New York: Oxford University Press.

Autonomic Nervous System: Carotid Body and Chemoception

C A Nurse, McMaster University, Hamilton, ON, Canada
V A Campanucci, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The mammalian carotid bodies are small, oval-shaped sensory organs that sense chemical components in arterial blood and initiate corrective reflex adjustments so as to maintain homeostasis. For example, they can detect changes in partial pressure of the blood gases O_2 and CO_2, that is, PO_2 and PCO_2, acidity (pH), osmolarity, and glucose concentrations. They can also sense temperature, further emphasizing their role as polymodal sensors of blood-borne stimuli. The bilateral location of these organs at the bifurcation of the common carotid artery is strategic, since they are ideally positioned to sample blood chemicals just before they reach the brain, a critical organ that is very sensitive to both O_2 and glucose deprivation. Consistent with its principal role as an arterial chemosensor, the carotid body is richly vascularized and receives an enormous blood flow, reputed to be the highest per unit weight of all body tissues and organs. It receives sensory (afferent) innervation from neurons located in the petrosal ganglion (PG) via the carotid sinus nerve (CSN), a branch of the glossopharyngeal nerve (GPN). Additionally, it receives 'efferent' innervation from autonomic neurons embedded within the GPN and CSN. The functional and structural properties of the carotid body, and the roles of these nervous elements in carotid body physiology, are discussed in more detail in subsequent sections.

Carotid Body as an Arterial Chemosensor

Since the pioneering studies of Heymans and collaborators in the 1930s, it has been well established that the carotid body is activated by a fall in oxygen tension or PO_2 (i.e., hypoxia) in arterial blood. Additionally, increases in blood PCO_2 (i.e., hypercapnia) and decreased pH (i.e., acidosis) have long been recognized as potent carotid body stimuli. These three stimuli can separately excite the carotid body, and initiate corrective changes in ventilation such that blood levels of PO_2, PCO_2, and pH are restored to normal. Interactions between PO_2 and PCO_2 are known to occur at the level of the carotid body and are synergistic, resulting in a more than additive chemosensory response when the two stimuli are combined. Though central chemoreceptors located in the brain stem are the main CO_2/pH sensors and regulators of minute-to-minute control of ventilation in the whole animal, carotid bodies appear to be the main PO_2 sensors. Moreover, they contribute to respiratory adjustments due to increased blood PCO_2, and are at least responsible for the initial changes in ventilation during metabolic acidosis, caused by changes in blood acidity at constant PCO_2 levels. These reflex responses originate at carotid body receptor cells, which are directly innervated by sensory fibers of the CSN (Figure 1). These receptor cells, commonly known as glomus or type I cells, release neurotransmitters that excite the afferent endings of the CSN, causing an increase in sensory discharge. This increase in discharge is relayed centrally, where it influences the 'central pattern generator,' a local neuronal network in a region of the brain stem (i.e., nucleus tractus solatarius) responsible for controlling breathing rhythm and ultimately contraction of the diaphragm muscles. Thus, conditions that result in low blood oxygen (hypoxemia), for example, during cardiovascular disease or ascent to high altitude, may cause an increase in CSN sensory discharge, especially when arterial PO_2 falls from its normal level of ~95 mmHg to values below 60 mmHg. The hypoxia-induced increase in CSN sensory discharge can be readily demonstrated *in vitro*, using an isolated but intact rat carotid body-attached nerve (CSN) preparation (Figure 2). *In vivo*, increased CSN activity ultimately leads to an increase in frequency and depth of breathing, such that blood PO_2 is restored toward normal levels, thereby correcting the initial disturbance.

More recent studies indicate the carotid body is also a glucosensor, capable of detecting low blood glucose (hypoglycemia) and activating a counter-regulatory neuroendocrine response. Together with the pancreas, liver, and portal vein, it contributes to a class of physiologically important glucosensors located outside the brain. Though central hypothalamic neurons are also known to sense hypoglycemia, their physiological significance has been questioned. An important role of the carotid body as a systemic glucosensor was initially demonstrated by the observation that in order to obtain an appropriate counter-regulatory neuroendocrine response to insulin-induced mild hypoglycemia, the innervation of the carotid body by the CSN needed to be intact. Additionally, insulin-induced mild hypoglycemia in adult rats results in a significant increase in spontaneous ventilation that is abolished if the CSNs are transected. Though recent studies have demonstrated that

low glucose can stimulate carotid body receptors *in vitro*, the underlying mechanisms are not completely understood.

Anatomical Organization of the Carotid Body

The principal cell components of the carotid body are clusters of endocrine-like glomus or type I cells that are intimately associated with glial-like, sustentacular type II cells. Type I cells are polymodal receptor cells that derive from the sympathoadrenal sublineage of the embryonic neural crest. Like sympathetic neurons and adrenal chromaffin cells, they synthesize catecholamines, of which dopamine is the one that predominates in type I cells. Immunolocalization of tyrosine hydroxylase (TH), the rate-limiting enzyme in catecholamine biosynthesis, has frequently been used to identify and characterize type I cells in carotid bodies of several species *in situ* (Figure 1). At the ultrastructural level, the occurrence of large cytoplasmic dense-cored vesicles is the most recognizable feature of these cells. Another morphological characteristic of clustered type I cells is that they form chemical and electrical synapses (i.e., gap junctions) with each other, and chemical synapses with sensory terminals of the petrosal afferent fibers. Additionally, electrical synapses have also been reported between type I and type II cells. The function of the type II cells is presently unclear, though it is possible they help coordinate activity within the cluster. Blood vessels, with fenestrated capillaries, are intimately associated with type I clusters, and this presumably facilitates delivery of the sensory stimuli to the receptor sites. Few autonomic ganglion-like cells have also been found inside the carotid body structure but their physiology is not well understood.

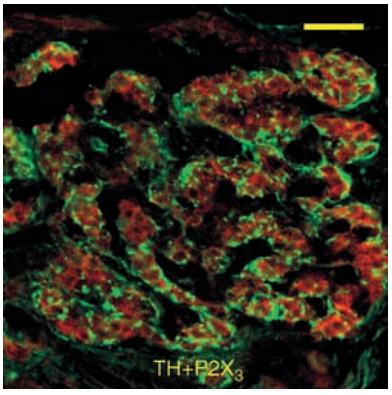

Figure 1 Innervation of chemoreceptor type I cells by petrosal sensory nerve terminals. In this tissue section from a 2-week-old rat carotid body, type I cell clusters are immunopositive for tyrosine hydroxylase (TH; red fluorescence), the rate-limiting enzyme in catecholamine biosynthesis. Nerve terminals surrounding type I cells are immunopositive for the purinergic receptor $P2X_3$ subunit (green fluorescence). Calibration bar represents 50 μm.

Neural Inputs to the Carotid Body

The main sensory innervation to the carotid body originates from neurons located in the PG (Figure 3). Their peripheral branches travel along the glossopharyngeal and CSNs to the carotid body (Figure 3), whereas their central projections terminate in the nucleus tractus solatarius of the brain stem. In rats, the chemoafferent neurons that directly innervate type I cells are located in the more distal region of the ganglion (Figures 3 and 4(a)). The majority of these chemoafferent neurons express TH (Figure 4(a)), and synthesize dopamine similar to carotid body type I cells. Neurons in the more proximal region of the ganglion send fibers that terminate near blood vessels and express other peptidergic markers such as substance P and calcitonin

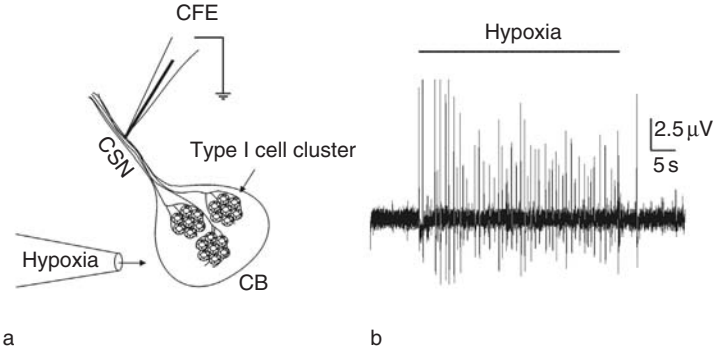

Figure 2 Hypoxia-induced increase in CSN activity in the isolated rat carotid body-attached nerve preparation *in vitro*. A schematic diagram showing an extracellular carbon fiber electrode (CFE) applied to CSN while a hypoxic solution is superfused over the preparation (a). A recording trace of the hypoxia-induced increase in CSN sensory discharge is shown in (b).

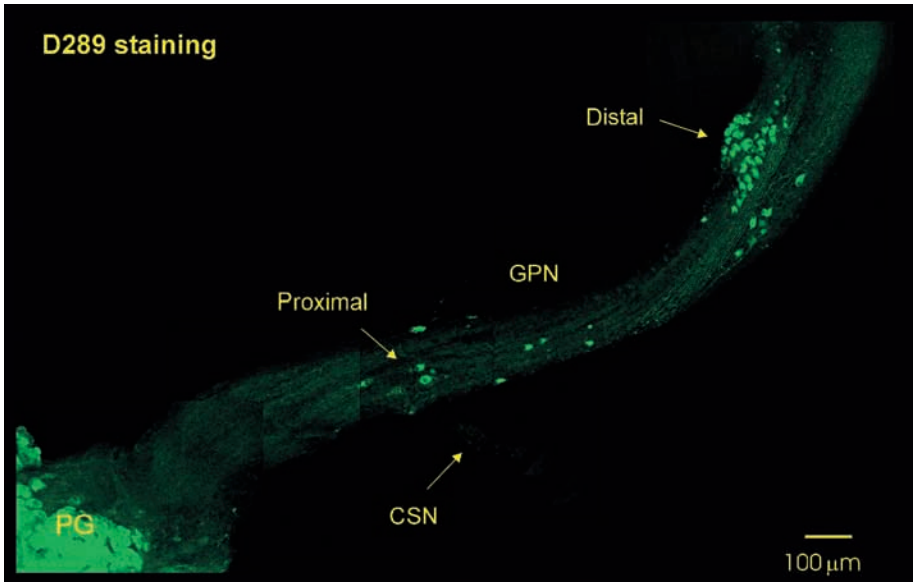

Figure 3 Vital staining of neuronal elements in rat PG and GPN. This whole-mount preparation was stained by the vital styrylpyridinium dye D289, following intraperitoneal injection. Sensory neurons in distal region of the PG, and two groups of autonomic neurons in the GPN are stained. The proximal GPN neuronal population is located near the branch point with the CSN; the distal population is located further along the GPN. The carotid body, normally attached to the CSN, was removed in this preparation. Reproduced from Campanucci VA and Nurse CA (2005) Biophysical characterization of whole-cell currents in O2-sensitive neurons from the rat glossopharyngeal nerve. *Neuroscience* 132(2): 437–451, with permission from Elsevier.

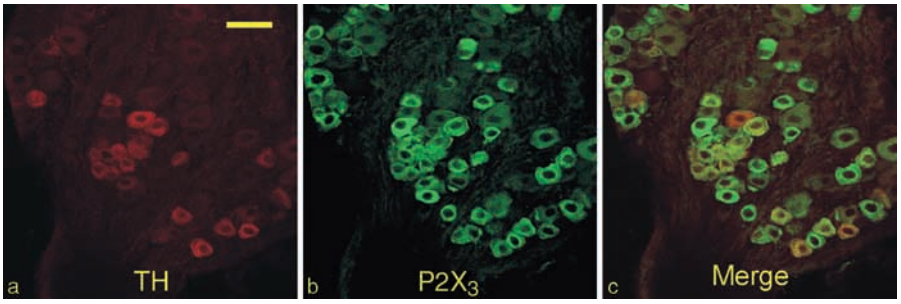

Figure 4 Immunostaining of carotid body chemoafferent neurons in the distal region of the PG. These chemoafferent neurons comprise a subpopulation that are mostly immunopositive for TH (red fluorescence) as illustrated in (a). Many petrosal neurons in the same section are immunopositive for the $P2X_3$ purinergic receptor subunit (green fluorescence) (b), including the TH-positive neurons. Co-localization of the two markers in chemoafferent neurons are shown in the merged images (c). Calibration bar represents 50 μm.

gene-related peptide. Additionally, autonomic neurons that express neuronal nitric oxide synthase (nNOS) are found in discrete populations within the glossopharyngeal and CSNs, and supply efferent innervation to the carotid body. These neuronal groups can be conveniently labeled and visualized in whole-mount preparations following injection of the vital styryl pyridinium dye D289 (Figure 3). Their role in chemoreceptor inhibition is discussed in a later section.

Transduction of Chemical Signals in Arterial Blood

While significant progress has been made toward our understanding of the transduction mechanisms in type I receptor cells during hypoxia, gaps in knowledge still remain. The prevailing view is that the hypoxia-induced receptor potential in type I cells is due to closing of K^+ channels that are open at rest. The particular type of PO_2-regulated K^+ channel may

differ among species, and it is likely that more than one K^+ channel subtype is involved in the hypoxic response even in the same cell. However, the identity of the O_2-sensor and the signaling pathways leading to K^+ channel closure are not completely resolved, and it is plausible that a single cell may use more than one pathway depending on the PO_2 level. Candidates for the O_2 sensor include heme proteins of the mitochondrial electron transport chain, heme oxygenase II, and AMP kinase. These are thought to generate second messenger signals (e.g., changes in ADP/ATP ratio, CO) during hypoxia, leading to K^+ channel closure and membrane depolarization (Figure 5). The latter in turn causes opening of voltage-gated Ca^{2+} channels, Ca^{2+} entry, and neurotransmitter release.

Transduction of CO_2 stimuli (hypercapnia) and acidosis also appear to involve K^+ channel inhibition, and the second messenger signal is thought to be an acidic intracellular pH (pH_i). In the case of hypercapnia, acidification of pH_i is catalyzed by intracellular carbonic anhydrase present in type I cells. The signaling pathways involved in low glucose sensing by type I cells are not completely understood, though inhibition of K^+ channels may also contribute.

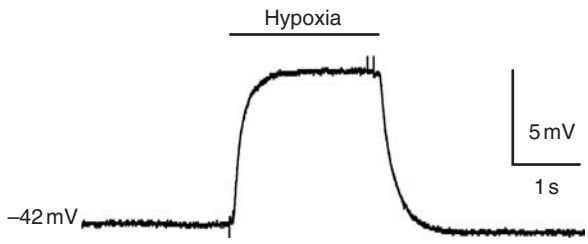

Figure 5 Hypoxia-evoked receptor potential in a type I cell. Current clamp recording showing depolarizing receptor potential evoked in a type I cell following exposure to acute hypoxia ($PO_2 \sim 5$ mmHg). The cell was a member of a cluster and the hypoxic stimulus was applied by rapid perfusion during the period indicated by the horizontal bar.

Role of Neurotransmitters in Sensory Processing

The neurotransmitter mechanisms involved in processing of chemosensory stimuli in the carotid body have attracted much attention, and in recent years substantive progress has been made. Traditional approaches to this problem relied on attempts to pharmacologically block the chemosensory discharge in the isolated, perfused, and/or superfused sinus nerve-carotid body preparation *in vitro* (see Figure 2). One disadvantage of this preparation is that synaptic events at the nerve terminal are not directly monitored, and block of spike activity could occur while subthreshold transmission still remains intact. A cell-culture model based on reconstruction of the chemosensory synapse *in vitro* has proved especially powerful in elucidating these neurotransmitter events. In co-cultures containing dissociated rat petrosal neurons and type I cell clusters, neurofilament-positive neuronal processes ramify over the culture substrate and contact TH-positive type I cells (Figure 6(a)). Importantly, in some cases, the petrosal soma is fortuitously situated adjacent to a type I cluster (Figure 6(b)). This allows the monitoring of electrical events, including subthreshold potentials, close to the synaptic sites. In such coculture models, it is possible to show that the same functional chemosensory neuron can be robustly activated by different stimuli including hypoxia, hypercapnia, and low glucose (Figure 7(a)). These models led to the first suggestion that corelease of the neurotransmitters ATP and

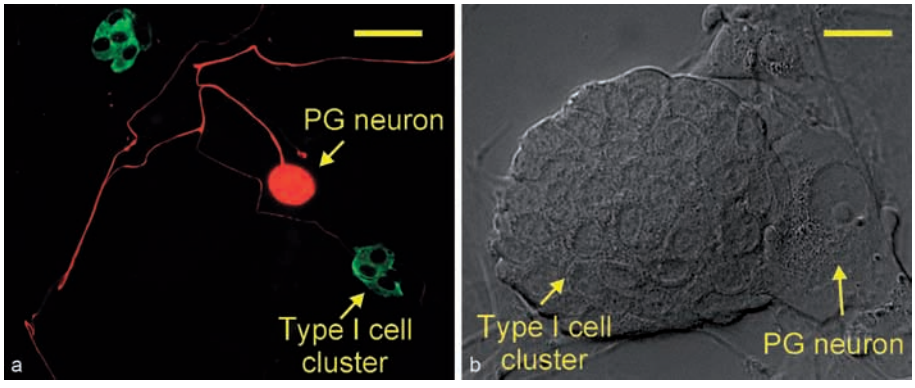

Figure 6 Coculture model of rat type I cell clusters and dissociated petrosal neurons. In (a), immunostaining of the coculture revealed positive neurofilament (NF) immunoreactivity in the petrosal neuron (PN) soma and its processes (red fluorescence), and positive TH immunoreactivity in two type I cell clusters (green fluorescence). Note that neuronal process appears to contact lower type I cell cluster. In (b), the differential interference contrast image shows a living type I cell cluster juxtaposed to a petrosal neuron, as typically used in electrophysiological recordings of synaptic interactions (see Figure 7). Calibration bar represents 20 μm in (a) and 10 μm in (b).

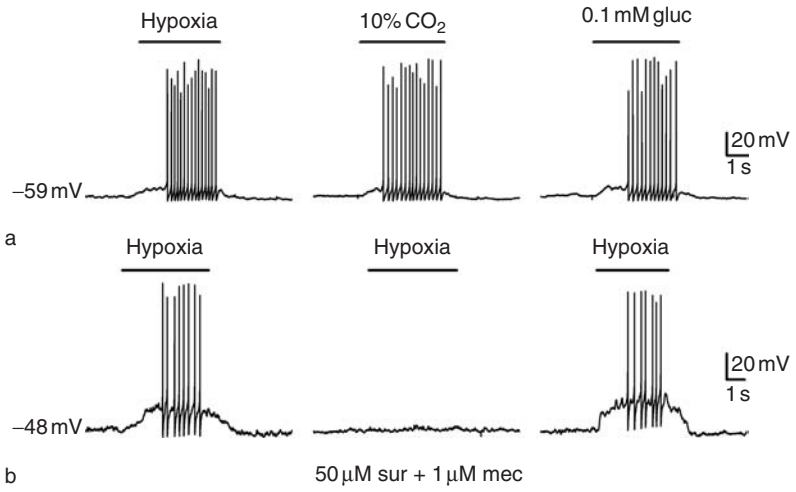

Figure 7 Chemosensory signaling in coculture and role of ATP and ACh as cotransmitters. In all traces, current clamp recordings of membrane potential were obtained from petrosal neurons juxtaposed to type I clusters as in Figure 6(b). In (a), the same chemosensory unit was excited by three stimuli, that is, hypoxia (left), isohydric hypercapnia (10% CO_2; pH = 7.4) (middle), and low glucose (0.1 mM) (right). Stimuli were applied during the period indicated by upper horizontal bar. In (b), hypoxic chemotransmission was reversibly abolished by a combination of purinergic (50 μM suramin) and nicotinic (1 μM mecamylamine) receptor blockers.

acetylcholine (ACh) was the main mechanism mediating hypoxic chemotransmission in the rat carotid body. Thus, in many cases, the hypoxic response recorded in cocultured neurons could be partially inhibited by purinergic (e.g., suramin) and nicotinic (e.g., mecamylamine) blockers applied separately, and completely inhibited when both blockers were present together (Figure 7(b)). Similarly, the hypoxia-induced sensory discharge in the isolated rat sinus nerve-carotid body preparation can be blocked by a combination of purinergic and nicotinic blockers, suggesting that the corelease mechanism is not an artifact of culture conditions. The neurotransmitter ATP, released from type I cells during hypoxia, acts on ligand-gated, purinergic P_2X receptors located on petrosal afferent nerve endings. Consistent with this idea, confocal immunofluorescence studies indicate that purinergic P_2X_2 and P_2X_3 subunits are localized on sensory nerve terminals apposed to type I cell clusters *in situ* (e.g., Figure 1). Moreover, co-localization of P_2X_2 and P_2X_3 receptor immunofluorescence in TH-positive chemoafferent neurons can be demonstrated in the distal region of the PG (Figures 4(a)–4(c)). A pivotal role of P_2X_2 subunits in hypoxia chemotransmission has been elegantly demonstrated with the use of P_2X_2 receptor knockout ($P_2X_2^{-/-}$) mice. These animals showed a markedly attenuated ventilatory response to hypoxia, and there was a dramatic reduction in hypoxia-induced nerve activity in the *in vitro* sinus nerve-carotid body preparation from $P_2X_2^{-/-}$ mice. Surprisingly, P_2X_3 receptor-knockout ($P_2X_3^{-/-}$) mice behaved similar to wild-type mice in their ventilatory response to hypoxia, though the resting chemoafferent activity in the sinus nerve was significantly reduced in these animals. Further, the hypoxia-induced sinus nerve discharge was further attenuated in $P_2X_{2/3}^{Dbl-/-}$ mice relative to $P_2X_2^{-/-}$ mice. The combined data from these knockout mice indicate an essential (and sufficient) role for P_2X_2 subunit in the ventilatory response to hypoxia, with smaller contributions from the P_2X_3 subunit. Comparable knockout experiments have not been reported to validate the more controversial role of ACh in chemotransmission, though such experiments are complicated by the presence of several and similar nicotinic ACh receptor subunits in both petrosal neurons and type I cells. The possibility that other neurotransmitters (e.g., 5-hydroxytryptamine (5-HT)) may contribute to the fast chemoexcitatory response in some cases has not been ruled out since ligand-gated 5-HT_3 receptors are expressed by many petrosal neurons.

In addition to the fast-acting neurotransmitters discussed, there is evidence for a complex regulation of carotid body sensory discharge by slower-acting neuromodulators that signal through G-protein-coupled pathways. These include positive regulators which facilitate chemoreceptor output such as 5-HT acting via presynaptic 5-HT_{2A} receptors, and adenosine acting via presynaptic A_{2A} receptors on type I cells. On the other hand, negative regulators appear to inhibit chemoreceptor output including dopamine acting on presynaptic D2 receptors and γ-aminobutyric acid (GABA) acting via presynaptic $GABA_B$ receptors. Histamine is also released during carotid body excitation and though various histamine receptors (H1, H2, and H3) have been identified in the organ, its

physiological role is presently unclear. To add to this complexity, there is evidence that released ATP can also act presynaptically on P2Y receptors located on both type I and type II cells, and ACh can produce opposing excitatory and inhibitory actions via nicotinic and muscarinic type I cell autoreceptors respectively. Neuropeptide modulators including substance P and enkephalins appear to have stimulatory and inhibitor actions respectively on carotid body discharge. The reasons for such a plethora of neuroactive chemicals in such a tiny organ remain obscure, but may be the basis for chemoreceptor plasticity that results in short- and long-term changes in organ sensitivity, following exposure to chronic or intermittent stimuli, for example, hypoxia.

Modulation of Sensory Discharge by Nitric Oxide Produced via Autonomic Neurons

In addition to the chemical neuromodulators of carotid body sensory discharge discussed in the previous section, gaseous modulators including nitric oxide (NO) are also involved. Sources of NO include endothelial cells, which express endothelial nitric oxide synthase, and efferent autonomic neurons within the GPN/CSN nerves that express nNOS (Figures 3 and 8(a)). The nNOS-containing fibers from autonomic GPN neurons or 'paraganglia' travel along the CSN and are a key component in the efferent inhibition of carotid body chemoreceptors via NO release. This idea had been suggested for many years, and was supported by the observations that electrical stimulation of the CSN, or carotid body exposure to hypoxia, caused an elevation in NO production that was prevented by specific NOS antagonists. Also, nNOS mutant mice displayed augmented ventilatory responses to hypoxia, and NO donors were reported to inhibit the carotid body chemosensory responses to hypoxia, as well as L-type Ca^{2+} currents in type I cells. Despite these compelling data for an inhibitory role of NO in chemoreception, the mechanisms that underlie activation of GPN neurons during hypoxia were not completely understood. Recent data have shed new light on this issue with the demonstration that isolated GPN neurons are O_2 sensitive and depolarize during hypoxia due to inhibition of background K^+ channels. This provides a potential pathway for Ca^{2+} entry, nNOS activation, and NO synthesis and release during hypoxia. Additionally, GPN neurons express at least four different types of P_2X receptor subunits (P_2X_2, P_2X_3, P_2X_4, and P_2X_7), allowing them to respond to a wide range of ATP concentrations (Figure 8(b)). This is particularly important since, as discussed earlier, ATP is a key excitatory neurotransmitter involved in carotid body chemoreceptor function. Therefore, ATP released from type I cells during hypoxia could excite GPN nerve terminals, leading to nNOS activation. The latter could occur following the rise in intracellular Ca^{2+} due to Ca^{2+} entry through Ca^{2+}-permeable P_2X receptors and/or voltage-dependent Ca^{2+} channels activated by ATP-induced membrane depolarization (Figure 8(b)). Indeed, the direct involvement of ATP on efferent inhibition of carotid chemoreceptors was recently demonstrated in a novel coculture model of type I cells and GPN neurons. In these conditions, application of ATP (or hypoxia) to GPN neurons caused a robust type I cell hyperpolarization that was prevented by preincubation with the NO scavenger carboxy 2-(4-carboxyphenyl)-4,4,5,5-tetramethyl-imidazoline-1-oxyl-3-oxide potassium (PTIO) or the NOS inhibitor L-nitro arginine methyl ester. Significantly, the NO donor sodium nitroprusside alone, but not ATP, could strongly hyperpolarize type I cells grown without GPN neurons in otherwise similar cultures. Therefore ATP, released from type I cells or erythrocytes during hypoxia, could lead to activation of GPN neurons and/or their

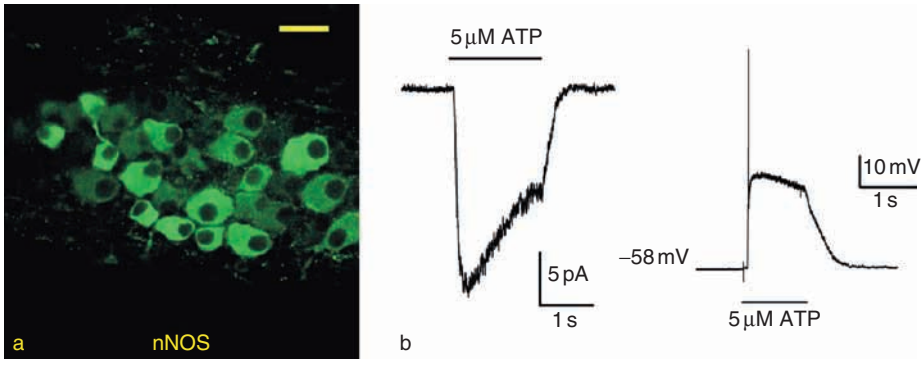

Figure 8 ATP sensitivity in nNOS-positive autonomic neurons of the distal population of GPN neurons. In (a), GPN neurons in a whole-mount of the GPN are immunopositive for nNOS (green fluorescence). In isolated GPN neurons in culture (b), ATP (5 μM) induced an inward current during voltage clamp at –60 mV (left), and membrane depolarization with superimposed spike activity during current clamp recording (right). Calibration bar represents 30 μm.

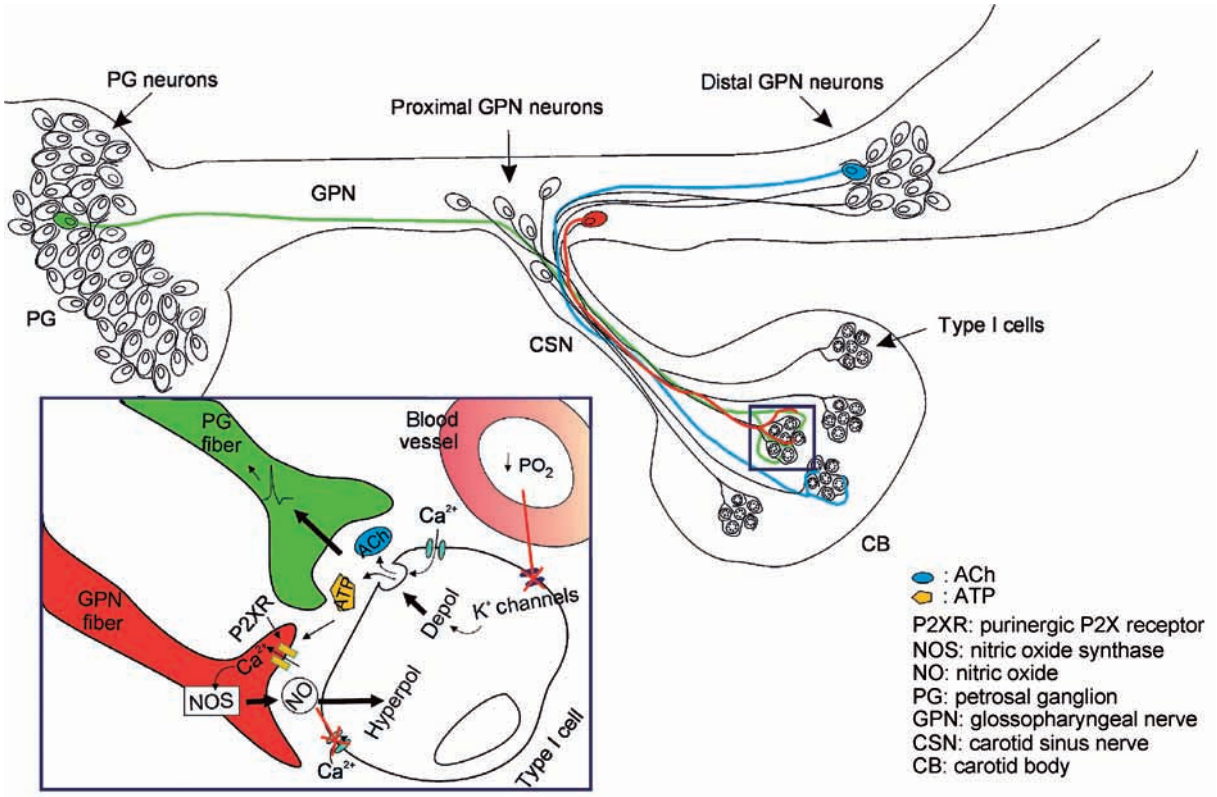

Figure 9 Summary diagram of afferent and efferent innervation of rat carotid body type I cells. Note that PG neurons (green) send afferent fibers that synapse with type I cell clusters. Also, proximal (red) and distal (blue) GPN autonomic neurons supply efferent fibers that terminate near type I clusters. During a decrease in PO_2 (hypoxia), ATP and ACh are coreleased from type I cells and interact with $P2X_2$–$P2X_3$ and nicotinic ACh (nAChR) receptors respectively (not shown) on petrosal terminals (green), causing CSN excitation (lower inset). This ATP (or ATP released from red blood cells during hypoxia) can also act on P2X receptors on GPN 'efferent' terminals (red), causing a rise in intracellular calcium, activation of nNOS, followed by synthesis and release of NO. Finally, NO causes type I cell hyperpolarization, suppression of neurotransmitter release, and inhibition of CSN sensory discharge.

terminals via multiple P_2X receptors, leading to nNOS activation, NO production, and chemoreceptor inhibition. The diagram in Figure 9 summarizes a proposed mechanism for carotid body efferent inhibition.

Carotid Body Plasticity

A remarkable feature of the carotid body is its ability to adapt to conditions of chronic or intermittent hypoxia. Chronic hypoxia can lead to time-dependent changes in the respiratory system including alterations in carotid body sensitivity to hypoxia. Ventilatory acclimatization to hypoxia involves both morphological and functional changes in the carotid body. For example, there is organ enlargement accompanied by both hypertrophy and hyperplasia of type I cells, as well as neovascularization. Functionally, there is an alteration in ion channel expression and neurotransmitter or neuromodulator functions in type I cells. These result in an increased sensitivity of the organ to hypoxia. Similarly, intermittent hypoxia, as occurs during recurrent apnea in humans, results in a selective enhancement of carotid body sensitivity to hypoxia, without obvious morphological changes. These adaptations also appear to involve changes in neurotransmitter functions, though these are not fully understood. The importance of understanding these mechanisms is underscored by the fact that intermittent hypoxia during recurrent apneas can lead to serious cardiovascular abnormalities in humans.

See also: Adenosine Triphosphate (ATP); Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Nervous System: Metabolic Function; Autonomic Nervous System; Gamma-Aminobutyric Acid (GABA); Nitric Oxide; Role of NO in Neurodegeneration.

Further Reading

Campanucci VA, Zhang M, Vollmer C, and Nurse CA (2006) Expression of multiple P2X receptors by glossopharyngeal neurons projecting to rat carotid body O_2 chemoreceptors: Role in nitric oxide-mediated efferent inhibition. *Journal of Neuroscience* 26: 9482–9493.

Iturriaga R and Alcayaga J (2004) Neurotransmission in the carotid body: Transmitters and modulators between glomus cells and petrosal ganglion nerve terminals. *Brain Research Brain Research Reviews* 47: 46–53.

Lopez-Barneo J (2003) Oxygen and glucose sensing by carotid body glomus cells. *Current Opinion in Neurobiology* 13: 493–499.

Nurse CA (2005) Neurotransmission and neuromodulation in the chemosensory carotid body. *Autonomic Neuroscience: Basic & Clinical* 120: 1–9.

Prabhakar NR (2006) O_2 sensing at the mammalian carotid body: Why multiple O_2 sensors and multiple transmitters. *Experimental Physiology* 91: 17–23.

Prabhakar NR and Peng YJ (2004) Peripheral chemoreceptors in health and disease. *Journal of Applied Physiology* 96: 359–366.

Rong W, Gourine AV, Cockayne DA, et al. (2003) Pivotal role of nucleotide $P2X_2$ receptor subunit of the ATP-gated ion channel mediating ventilatory response to hypoxia. *Journal of Neuroscience* 23: 11315–11321.

Autonomic Nervous System: Central Cardiovascular Control

M C Andresen, Oregon Health & Science University, Portland, OR, USA
D Mendelowitz, George Washington University, Washington, DC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Vital nourishment of all tissues is assured by a properly regulated cardiovascular system. Consistent and coordinated contractions of the heart provide pressure for the circulation system and vessel conduits distribute the blood to flow throughout the body. Thus, the goal of cardiovascular control is to maintain and allocate blood flow that meets tissue demands over a diverse range of circumstances. Cardiac contractions are normally initiated by electrical pacemaker activity within the sinus node of the right atrium, and without external stimuli, intrinsic pacemaking automatically maintains rates that are fairly constant. However, the autonomic nervous system is responsible for adjustments that meet changing circulatory demands such as exercise or postural challenges. Thus, the nervous system orchestrates optimal adjustments in the pump function of the heart. Both rate and the timing of events between atria and ventricles are altered by released neurotransmitters to facilitate efficient pumping. Neurotransmitter activation of smooth muscle in blood vessels adjusts vessel caliber and thus flow patterns. These cardiovascular adjustments are accomplished by neurotransmitters released by autonomic nerves. Overall, the level of arterial blood pressure at all times is a dominant factor in driving the circulation of blood. Multiple neural reflexes rapidly control blood pressure to accommodate a wide range of systemic and regional demands and priorities. Acute postural change, for example, can require changes in pump rate at the very next heart beat, and such rapid adjustments use the autonomic nervous system to quickly fine-tune the heart and the vascular distribution of blood.

The organization of the central nervous system (CNS), directed toward control of the cardiovascular system, follows a general, prototypic pattern for these reflex pathways. Afferent information arises from a network of afferent neurons distributed within key portions of the cardiovascular system. The peripheral endings of cardiovascular afferent neurons are unequally distributed and are particularly concentrated in the centralmost portions of the vascular tree (aorta, common carotid artery, vena cava, and jugular vein) and within the heart. Mechanosensitive afferent neurons (e.g., arterial baroreceptors) transduce local vessel and chamber stretch as an index of filling or local pressure. This 'filling' information plays a key role in rapid cardiovascular adjustments on a beat-to-beat basis. Afferent neurons arising from the heart as with many other visceral organs have cell bodies in cervical ganglia associated with the IX and X cranial nerves (nodose, jugular, and petrosal ganglia). Additional cardiovascular afferents arise from neurons of the dorsal root ganglia and often these spinal afferent neurons send axons that travel via the sympathetic nerve trunks and have been termed sympathetic afferents in the literature. Visceral afferents dominate reflexes that adjust the autonomic outflow back to the organ of origin, but often additionally contribute to the control of entirely different organs (e.g., respiratory afferents that alter cardiac control). Both cardiac and arterial mechanosensory neurons initiate rapidly responding, relatively short pathways for parasympathetic efferent adjustments as well as longer latency reflexes via the spinal cord that modulate sympathetic efferent neurons. Cardiovascular reflex pathways predominantly involve networks of neurons in the brain stem and spinal cord. Control of heart rate and arterial pressure is the primary focus of this overview.

The arterial baroreceptor reflex (baroreflex) control of heart rate is perhaps one of the most basic autonomic regulatory reflexes and serves as a central pathway prototype for reflexes arising from many other cranial visceral afferents. The parasympathetic reflex control of the heart is fundamentally better understood than control of arterial pressure or venous capacitance. The consensus view of the core circuit for the cardiac parasympathetic control suggests that this reflex arc relies on structures confined to the medulla. Baroreceptor afferents project to brain stem neurons in the nucleus of the solitary tract (NTS). NTS neurons, in turn, contact output neurons within the nucleus ambiguus (NA) that exit the brain stem to contact final parasympathetic motor neurons within the heart itself. While this constitutes the simplest pathway of cardiovascular control, additional, more convoluted, brain stem pathways exist as well as indirect contributions arising from ascending spinal viscerosensory pathways.

Medulla Oblongata

The arterial baroreflex is organized as a classic negative feedback system in which increases in arterial blood pressure activate mechanosensitive arterial baroreceptors (Figure 1) that trigger reflex responses

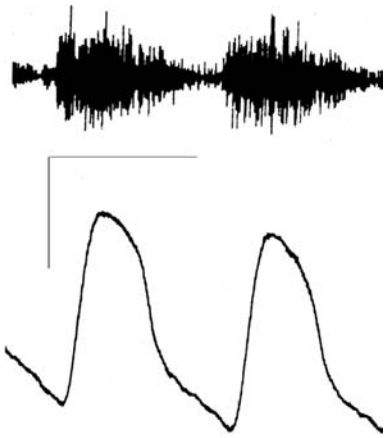

Figure 1 Pulse synchronous arterial baroreceptor discharge. Aortic depressor nerve activity recorded from whole nerve in a conscious rabbit (upper trace) rises abruptly with the systolic phase of femoral arterial blood pressure (lower trace). Scales are 20 mmHg (vertical) and 500 ms (horizontal).

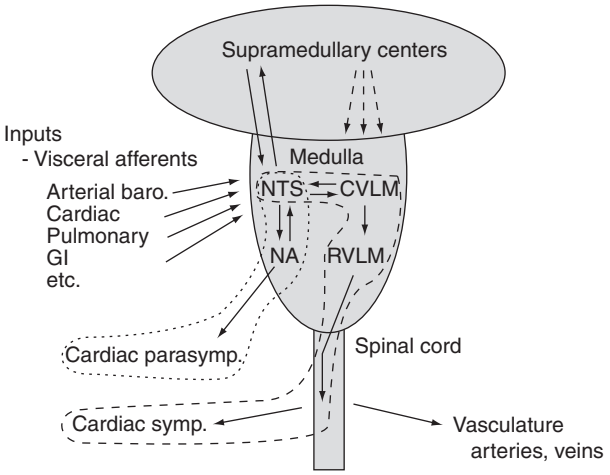

Figure 2 Schematic of core baroreflex pathways for cardiovascular control. Reflex responses are initiated from visceral primary afferent neurons such as cardiac afferent and arterial baroreceptor neurons (left) that converge on initial central neurons (second-order neurons) within the nucleus of the solitary tract (NTS). From there, information diverges to two key autonomic centers along the parasympathetic pathway to nucleus ambiguus (NA) and along the sympathetic pathway to the caudal ventrolateral medulla (CVLM). The parasympathetic pathway (dotted outline) leaves NA to target the intrinsic cardiac network including postganglionic neurons at the heart. The sympathetic regulatory pathway (dashed outline) goes on to neurons in the rostral ventrolateral medulla (RVLM) before descending to the spinal cord at the intermediolateral cell column and out to the sympathetic postganglionic neurons. NTS is reciprocally connected to a wide range of CNS regions (bidirectional arrows) including many supramedullary regions that modulate reflex function but are not part of the essential core circuit. Forebrain regions, for example, often act to modify activity within the core brain stem circuits. A wide range of pathways (excitatory and inhibitory) contacts neurons of the core baroreflex pathway. Note that the negative feedback and slowing of the heart rate occurs peripherally in the parasympathetic pathway by actions of the postganglionic neurotransmitter acetylcholine, but centrally in the sympathetic pathway as CVLM-released GABA inhibits RVLM neurons (and sympathetic activity).

that ultimately reduce arterial blood pressure. The discharge of arterial baroreceptors is conducted along cranial nerve axons (IX and X) projecting directly into the brain stem to neurons in the NTS where they make the first synapse in the baroreflex pathway (Figure 2). Processed afferent signals exit NTS to activate vagal preganglionic neurons located primarily in NA that project to cardioinhibitory parasympathetic postganglionic neurons in the heart. NTS neurons also activate sympathoinhibitory neurons located in the caudal ventrolateral medulla (CVLM). Thus, activity along dual, cardiac directed autonomic pathways is oppositely regulated to produce negative feedback control of cardiac pumping by (1) increased release of acetylcholine by parasympathetic postganglionic neurons and (2) decreased release of norepinephrine by sympathetic postganglionic neurons.

In the resting, conscious state, midcollicular decerebration does not alter the gain (sensitivity) of this heart rate baroreflex, suggesting that brain stem mechanisms are sufficient for reflex function. Although brain structures above the medulla are not required, information from the forebrain and other higher CNS sites modulate the baroreflex during various behaviors including stressful emotional responses. Thus, the core baroreflex circuit appears to reside below the pons and contains both short and longer latency pathways. The shortest anatomical pathway activates cardiac parasympathetic efferent neurons by engaging, at a minimum, (1) primary afferent neurons associated with cardiovascular mechanoreceptors, (2) second-order NTS neurons, (3) cardiac parasympathetic preganglionic motor neurons located primarily in the NA, and (4) parasympathetic postganglionic neurons intrinsic to the heart. Such a parasympathetic pathway is responsible for relatively short-latency (<100 ms) cardiovascular-cardiac reflexes that adjust heart rate within one cardiac cycle (Figure 2).

Primary Afferent Neurons of the Arterial Baroreflex

Cardiovascular regulation depends on the constant monitoring of arterial pressure, central venous pressure, and cardiac rate and force. Populations of cardiovascular-selective sensory neurons supply a constant stream of information and these viscerosensory neurons belong to cellular/molecular phenotypic classes based on axon myelination that are broadly similar to sensory neurons of the skin (somatic afferents). Baroreceptors are concentrated near the bifurcation of the common carotid artery (carotid sinus) and just outside the left ventricle in the arch of the thoracic aorta. Neurites from arterial baroreceptors lie within the adventitia, the outermost layer of the artery wall.

This position within the artery wall means that they are stretched with each expansion of the artery wall that accompanies each expulsion of blood from the heart. A-type baroreceptors rapidly conduct responses via myelinated fibers (2.5–25 m s^{-1}) in brief bursts of activity during each pressure pulse (Figure 1). A-type baroreceptors discharge within the normal range of blood pressures even at rest and their discharge activity is quite regular and encodes mean arterial pressure, as well as the rate of rise and amplitude of the pressure pulse. A much larger number of arterial baroreceptors are C-type and conduct more irregular discharge patterns along slowly conducting (<1 m s^{-1}), unmyelinated axons. C-type baroreceptors are often silent at resting blood pressures and become active with sparse, intermittent discharge only at higher pressures than the A-type baroreceptors. A minority of C-type baroreceptors is activated at normal arterial pressures. The relationship between the amplitude of sensory terminal stretch and the frequency of evoked discharge defines the 'baroreceptor sensitivity.' This is not a fixed relationship, but displays 'plastic' changes both acutely and chronically in the face of sustained blood pressure changes. Under physiological and pathological conditions, the roles played by specific afferent neuronal populations can become 'reset' and change along with changes in cardiovascular regulation.

Solitary Tract Nucleus – Arterial Baroreflex Second-Order Neurons

Neurons within dorsal medial portions of the caudal NTS mark the central first step of the medullary baroreceptor reflex. Sensory inputs from baroreceptor afferent neurons release glutamate as their primary neurotransmitter to activate non-*N*-methyl-D-aspartate (NMDA) receptors on second-order NTS neurons. Inhibitory transmission via γ-aminobutyric acid (GABA) is prominent at second-order neurons in the medial NTS with inhibitory postsynaptic potentials arriving shortly after visceral afferent glutamatergic excitatory synaptic potentials. Since a majority of these GABAergic neurons are themselves directly coupled to viscerosensory afferent inputs, this organization provides effective short-latency feedback inhibition within this cardiovascular-cardiac reflex. These two amino acid transmitters activate ionotropic (ion channel coupled) receptors to convey fast transmission but both glutamate and GABA, in addition, activate metabotropic receptors that are coupled to intracellular second messengers. Thus, heterosynaptic cross talk between neurotransmitters is common and provides afferent-linked mechanisms that are very sensitive to the frequency content of afferent patterns of discharge. A wide range of accessory transmitters such as peptides like ATP, substance P, and calcitonin gene-related peptide are also involved in such transmission, although their precise roles are less certain.

Cardiac Parasympathetic Neurons

Cardiac parasympathetic motor neurons exert powerful control on cardiodynamics. Heart rate is dominated by the activity of the cardioinhibitory parasympathetic nervous system (Figure 2). Cardiac parasympathetic neurons are tonically active at rest but little, if any, cardiac sympathetic activity is commonly present under normal conditions. Preganglionic cardiac vagal neurons have cell bodies located mostly in the NA and, to a lesser extent, in the dorsal motor nucleus of the vagus. These parasympathetic preganglionic neurons send axons that exit the brain stem via the vagus nerves to contact postganglionic neurons located in cardiac ganglia. These cardiac ganglia are found within the connective and fatty tissue that surround the right atrium and vena cava. Parasympathetic postganglionic fibers emerge from these ganglia to make extensive synaptic contacts with the nearby sinoatrial and atrioventricular nodes of the heart where they release the neurotransmitter acetylcholine.

Activity within the arterial baroreflex is a major determinant of the tonic activity of the parasympathetic efferent preganglionic neurons that target the heart. During normal states, parasympathetic efferent preganglionic neuronal activity occurs during specific phases of the cardiac cycle initiated by this fast-responding reflex. Heart rate is dominated by the activity of the cardioinhibitory parasympathetic nervous system. Measures of these ongoing, tonic, and rapid adjustments in parasympathetic activity (e.g., heart rate variability) are indicators of cardiac health and, when absent, can predict an important risk of cardiac sudden death. During increases in arterial pressure, the initial reflex-induced slowing of the heart (bradycardia) is caused primarily, if not exclusively, by increases in cardiac vagal nerve activity. During decreases in arterial pressure, the baroreflex-induced cardiac acceleration (tachycardia) is caused mostly by decreases in parasympathetic along with somewhat later-arriving increases in sympathetic nerve activity. When both parasympathetic and sympathetic activities are present, parasympathetic activity generally dominates the control of heart rate. Increases in parasympathetic activity evoke a cardiac slowing that is more pronounced when there is a high level of sympathetic firing. When there is a moderate or high level of parasympathetic activity, changes in sympathetic firing elicit negligible changes in heart rate.

The CNS system normally controls the parasympathetic activity reaching the heart. Preganglionic cardiac

vagal fibers tonically fire with a cardiac pulse synchronous pattern that is reduced during inspiration. Interruption of links to the CNS by cutting the cardiac vagal preganglionic fibers leads to an increase in heart rate demonstrating that the remaining, intact cardiac parasympathetic postganglionic neurons by themselves cannot sustain normal cardiac rates. Similarly, synaptic blockade silences parasympathetic neurons within the brain stem indicating an absence of pacemaker-like activity such as repetitive or phasic depolarizations or action potential. However, even small depolarizing currents (100 pA) evoke sustained, repetitive firing in cardiac vagal neurons with minimal spike frequency adaptation. The voltage-gated currents and firing characteristics of cardiac vagal neurons enable them to closely follow fast synaptic drive as well as to integrate long-lasting modulatory influences. Together, the characteristics of both these preganglionic and postganglionic cardiac parasympathetic neurons suggest that heart rate at rest is determined by activity within the parasympathetic pathways that arises from the integration of afferent and other drives to produce normal levels of cardiac vagal tone.

The activity of cardiac vagal preganglionic neurons is controlled by three major classes of synaptic inputs: glutamatergic, GABAergic, and glycinergic. NTS neurons provide a major glutamatergic input pathway that activates both NMDA and non-NMDA postsynaptic currents in cardiac vagal neurons. Owing to baroreceptor activation of NTS neurons, this glutamatergic pathway to cardiac vagal neurons likely constitutes the essential link between increases in blood pressure and the parasympathetic component of the arterial baroreceptor reflex. Additionally, NTS activation triggers GABAergic inputs to cardiac vagal neurons. Electrical stimulation of afferents in the vagus nerve evokes both GABAergic and glutamatergic responses in cardiac vagal neurons. In addition, the pungent active ingredient in hot peppers, capsaicin, can chemically activate and then, with prolonged exposure, inactivate visceral sensory C-type afferents including baroreceptors. Blockade of C-type afferents by capsaicin increases the latency of the GABAergic response without changing the latency of the glutamatergic responses. Such findings suggest that an inhibitory GABAergic pathway is involved in patterning of cardiac vagal activity to produce bursting synchronous with the cardiac cycle.

Cardiac vagal nerve activity also has pronounced respiratory modulation that mediates respiratory sinus arrhythmia, a pattern of slowing and accelerating heart rate in phase with lung inflation and deflation. Cardiac vagal fibers fire most rapidly in the postinspiration period, and often fall silent during inspiration. A form of this firing activity of cardiac vagal neurons during the phases of respiration persists in slices of the medulla that contain cardiac vagal neurons, pre-Bötzinger complex, and local circuits for motor output generation, and inspiratory motor discharge patterns can be recorded in hypoglossal cranial nerves (Figure 3). The firing of cardiac vagal neurons is inhibited during inspiratory activity.

Nicotinic cholinergic mechanisms have important actions on cardiac vagal neurons. During inspiration,

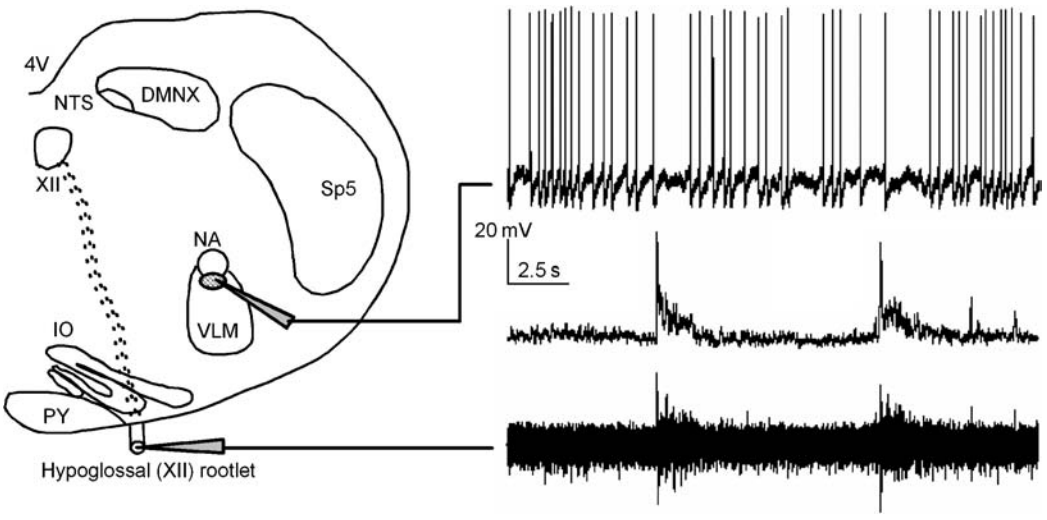

Figure 3 The medullary slice preparation that contains the cardiac vagal neurons, pre-Bötzinger complex, and respiratory hypoglossal motor neurons with intact hypoglossal cranial nerves is illustrated. Inspiratory activity recorded from the hypoglossal rootlet is shown bottom trace, right, and integrated, middle trace, right. The firing activity of a cardiac vagal neuron, induced to fire repetitively with a constant injection of inward current (~50 pA), top right, is inhibited during each inspiratory burst.

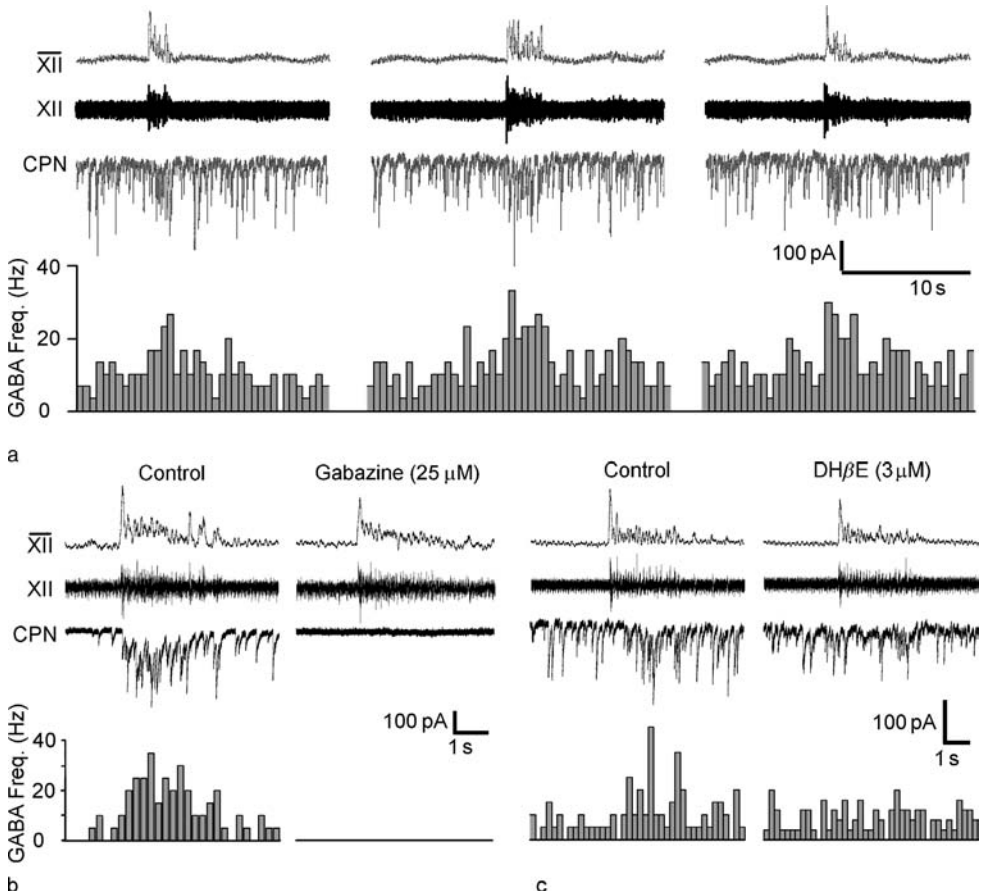

Figure 4 Inspiratory-related bursting activity was recorded from the hypoglossal rootlet (XII) and rectified and adjacent averaged ($\overline{\text{XII}}$). GABAergic inhibitory postsynaptic currents (IPSCs) were isolated by focal application of the NMDA, non-NMDA, and glycine receptor antagonists AP-5 (50 μM), CNQX (50 μM), and strychnine (1 mM), respectively. (a) During inspiratory activity, the frequency of GABAergic IPSCs in cardiac vagal neurons was significantly increased ($p < 0.05$). (b) The $GABA_A$ antagonist gabazine blocked all IPSCs. (c) This inspiratory-related increase in GABA synaptic frequency was abolished by focal application of the nicotinic receptor antagonist dihydro-β-erithroidine (DhβE) at a concentration selective for the $\alpha 4\beta 2$ receptor subtype DHβE did not significantly change GABA synaptic frequency between inspiratory bursts.

the frequency of both GABAergic and glycinergic synaptic events increases substantially (Figure 4). A nicotinic antagonist, selective for a specific subset of nicotinic receptors ($\alpha 4\beta 2$), abolishes this respiratory-evoked increase in GABAergic events supporting the critical role of cholinergic inputs in shaping cardiac parasympathetic control. This is a highly specific interaction since this intervention has no effect on the increase in glycinergic synaptic events during inspiration. Together, such findings provide a basic outline of interactions between glutamate, glycine, and acetylcholine in the control of heart rate in this core NTS–NA pathway (Figure 5).

Spinal Cord

Sympathetic preganglionic neurons related to the cardiovascular system are located within the intermediolateral cell column of the spinal cord and constitute a second major autonomic effector pathway. These sympathetic preganglionic neurons send processes from the spinal cord to activate postganglionic neurons within peripheral ganglia. The postganglionic sympathetic neurons themselves send axons to contact visceral organs and blood vessels where they release the neurotransmitter norepinephrine. Activity in these sympathetic neurons influences all aspects of cardiac function including heart rate, cardiac conduction, the strength of cardiac contraction, and ventricular stroke volume. Many sympathetic motor neurons are directed to blood vessels resulting in activation of smooth muscle and changes in vessel diameter and local arterial resistance or venous capacitance.

Although reflexes within the medulla exert substantial influence on cardiovascular sympathetic efferent neurons, at present these reflexes are incompletely defined. Sympathetic premotor neurons within the

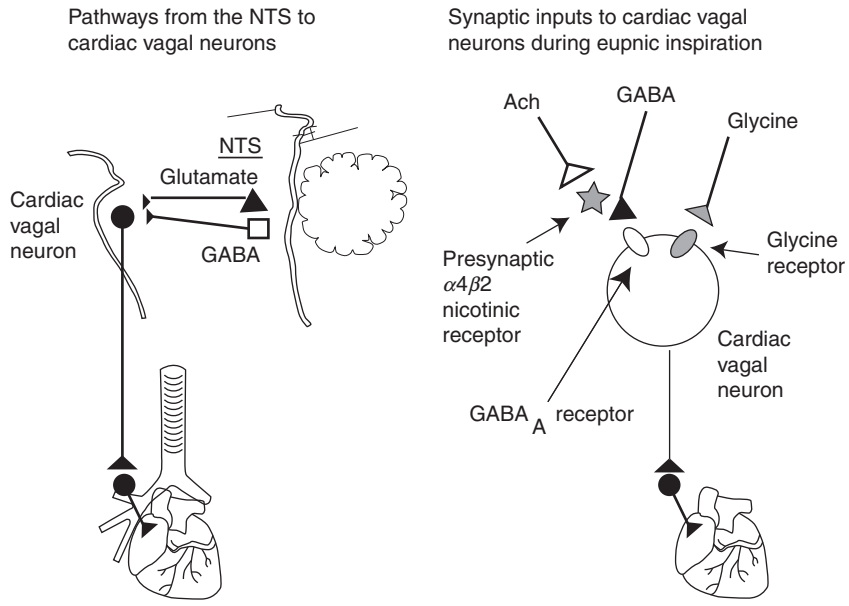

Figure 5 Neurotransmitter actions in the core reflex pathway in the control of cardiac parasympathetic neurons. NTS sends glutamatergic and GABAergic but not glycinergic pathways to cardiac vagal neurons (left). Inspiratory activity increases both GABAergic and glycinergic inputs to cardiac vagal neurons (right). Nicotinic $\alpha 4\beta 2$ receptors presynaptically facilitate GABAergic, but not glycinergic, neurotransmission.

rostral ventrolateral medulla (Figure 2) send axons to modulate spinal cord cardiovascular sympathetic preganglionic neurons. An important determinant of cardiovascular sympathetic regulation arises from sympathoinhibitory neurons lying within the CVLM that secrete GABA onto sympathetic premotor neurons within the rostral ventrolateral medulla. The full range of mechanisms of the control of sympathetic premotor neurons remains to be established. However, baroreflex pathways provide substantial sympathoinhibitory input and thus remain part of a functionally negative feedback system that, for example, reduces heart rate by inhibiting (i.e., withdrawing) ongoing norepinephrine release. Thus, NTS neurons that receive tonic inputs from cardiovascular afferent neurons activate sympathoinhibitory neurons that, in turn, phasically depress sympathetic premotor neuronal activity. These premotor neurons, in turn, project to sympathetic efferent preganglionic neurons in the intermediolateral cell column of the spinal cord. The cranial thoracic spinal cord preganglionic neurons project to cardiac sympathetic efferent postganglionic neurons distributed throughout the intrathoracic and superior cervical ganglia. Interestingly, such spinal cord sympathetic motor neurons also receive inputs from cardiac afferent neurons in dorsal root ganglia. Unlike the medullary cardiovascular reflexes depicted above, activation of spinal cardiac afferent neurons results in the activation of spinal cord sympathetic premotor neurons involved in enhancing cardiac rate and contractile force – a positive feedback mechanism. Spinal cord sympathetic preganglionic motor neurons also innervate peripheral blood vessels and influence arterial vascular resistance and venous capacitance. The interrelationship between positive and negative feedback pathways controlling cardiovascular sympathetic motor neurons is poorly understood, although often suspected to be involved during pathophysiological circumstances.

Higher Centers

Under many normal and pathophysiological conditions, the gain or sensitivity of the heart rate baroreflex can be altered by influences arising from supramedullary brain structures such as the hypothalamus or amygdala. Under stress and even during physiological states, these supramedullary inputs transform the operation of the baroreflex to accommodate integrative responses for exercise or strong emotional behaviors. Thus, cardiovascular regulation including the major baroreflex pathways is often an integral component of major homeostatic and highly complex patterned responses. These descending inputs from higher center neurons modulate both the fast-responding parasympathetic reflexes that provide beat-to-beat adjustments in cardiac performance as well as the longer latency cardio-cardiac reflexes involving spinal cord (sympathetic) neurons.

Clinical Implications

As a powerful, classic negative feedback system, the baroreflex provides rapid cardiovascular adjustments to the changing demands on the heart and great vessels. Cardiovascular feedback is constantly responding and remodeling to altered physiological status such as occurs during prolonged bed rest or zero gravity conditions. Substantial remodeling and reshaping of these reflex pathways also occurs in pathological states such as essential hypertension or heart failure. Prominent alterations in neuronal structures and neurotransmitter density within the NTS are associated with cases of sudden infant death syndrome, for example. The quantitative performance characteristics of the cardiac baroreflex have been linked to postmyocardial infarction survival as well as risk of sudden death from lethal cardiac arrhythmias. The sensitivity of the multiple reflexes involved in blood pressure regulation is reshaped at all stages of the cardiovascular neuroaxis during development of hypertension. The mechanisms for integration of all of the central neuronal reflexes involved in the maintenance of normal arterial pressure are not fully understood but presumably neural system remodels during the evolution of diseases hold keys to understanding essential hypertension as well as the development of novel and optimally effectively targeted therapies to treat such syndromes.

See also: Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System; Blood Pressure: Baroreceptors; Cardiovascular Function: Central Nervous System Control; Cerebrovascular Disease; Parasympathetic Nervous System.

Further Reading

Andresen MC and Kunze DL (1994) Nucleus tractus solitarius: Gateway to neural circulatory control. *Annual Review of Physiology* 56: 93–116.

Andresen MC, Kunze DL, and Mendelowitz D (2004) Central nervous system regulation of the heart. In: Armour JA and Ardell JL (eds.) *Basic and Clinical Neurocardiology,* 2nd edn., pp. 187–219. New York: Oxford University Press.

Armour JA (1976) Instant to instant reflex cardiac regulation. *Cardiology* 61: 309–328.

Cole CR, Blackstone EH, Pashkow FJ, Snader CE, and Lauer MS (1999) Heart-rate recovery immediately after exercise as a predictor of mortality. *New England Journal of Medicine* 341: 1351–1357.

Kinney HC, Filiano JJ, and Harper RM (1992) The neuropathology of the sudden infant death syndrome. A review. *Journal of Neuropathology and Experimental Neurology* 51: 115–126.

Kunze DL and Andresen MC (1991) Arterial baroreceptors. In: Zucker IH and Gilmore JP (eds.) *Excitation and Modulation, Reflex Control of the Circulation*, pp. 141–166. Boca Raton: CRC Press.

LaRovere MT, Pinna GD, Hohnloser SH, et al. (2001) ATRAMI investigators: Baroreflex sensitivity and heart rate variability in the identification of patients at risk for life-threatening arrhythmias: Implications for clinical trials. *Circulation* 103: 2072–2077.

Loewy AD (1990) Central autonomic pathways. In: Loewy AD and Spyer KM (eds.) *Central Regulation of Autonomic Functions*, pp. 88–103. New York: Oxford University Press.

Mayr U and Keele S (2000) Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General* 129: 4–26.

Mendelowitz D (1999) Advances in parasympathetic control of heart rate and cardiac function. *News in Physiological Sciences* 14: 155–161.

Spooner PM, Albert C, Benjamin EJ, et al. (2001) Sudden cardiac death, genes, and arrhythmogenesis: Consideration of new population and mechanistic approaches from a national heart, lung, and blood institute workshop. Part I & II. *Circulation* 103: 2447–2452.

Autonomic Nervous System: Central Control of the Gastrointestinal Tract

W Jänig, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The functions of the gastrointestinal tract (GIT) are ingestion, transport, and enzymatic breakdown of its content, reabsorption, and secretion of fluid (regulation of fluid balance), defense of the body against poisonous substances, and evacuation of waste. These functions are exerted by groups of effector cells of the GIT, such as smooth nonvascular muscle cells, secretory epithelia, interstitial cells of Cajal, endocrine cells, and cells related to the immune system (gut-associated lymphoid tissue). The enteric nervous system (ENS) is a specialized autonomic nervous system, consisting of motor neurons, intrinsic primary afferent neurons, and interneurons; the ENS regulates the activity of these effector cells, leading to spatial and temporal coordination of the functions of the GIT independent of the brain.

The brain adapts the intestinal functions at any moment to the behavior of the organism. This control is particularly powerful at the oral site (esophagus, stomach) and at the anal site (hindgut). It is exerted by various parasympathetic pathways, most of them originating in the dorsal motor nucleus of the vagus (DMNX) and the sacral spinal cord, and various sympathetic nonvasoconstrictor pathways originating in the thoracolumbar spinal cord. The vasculature of the GIT is, in addition to being locally controlled by enteric neurons, under direct control of the brain by visceral vasoconstrictor neurons. This control is particularly important in the regulation of arterial blood pressure (Figure 1).

Information from the GIT to the brain is transmitted via visceral afferent neurons and hormones secreted by gastrointestinal (GI) endocrine cells. Afferents projecting through the vagus nerves convey detailed information on various functions of the GIT. Visceral afferent neurons projecting from the GIT to the spinal cord are involved in evacuation and continence of the hindgut, in nociception and pain (including protective reflexes), and in some other functions. The endocrine signals derive from endocrine cells in the mucosa of the GIT (most of them in the anterior part) and pancreas and act on the brain mainly via the area postrema (AP) in the medulla oblongata and/or on the arcuate nucleus (ARN) in the hypothalamus. These GI hormonal signals are particularly important in the long- and short-term regulation of metabolism as well as of food intake, and therefore of hunger, satiety, and body weight (Figure 1).

The biological global functions of the control of the GIT by the brain are (1) adaptation of the functions of the GIT to the behavior of the organism and (2) regulation of energy balance and food intake (this includes regulation of body weight, hunger, and satiety). Pathobiologically, these global neural regulations by the brain may fail and result in (1) dysregulation of transport and fluid balance (secretion, reabsorption), being obvious at the oral and anal section of the GIT, where the control by the brain is particularly powerful, (2) dysregulation of energy balance, food intake, body weight, hunger, and satiety, and (3) psychosomatic diseases related to the GIT.

In the following sections, the neuronal pathways by which the central nervous system (CNS) controls the GIT are described, concentrating mainly on the medulla oblongata.

Afferent and Autonomic Pathways Connecting GIT and Brain

The parasympathetic systems have their dominant effects and overall access to target cells at the oral and anal sites of the GIT. Otherwise they do not interfere with single groups of effector cells in the GIT but influence enteric sensorimotor programs. The sympathetic nonvasoconstrictor systems act throughout the GIT, do not interfere directly with single groups of effector cells, and also modulate the enteric sensorimotor programs. The numbers of preganglionic parasympathetic and sympathetic neurons involved in GI functions is low (in the range of $\leq 1\%$!) in comparison to the numbers of enteric neurons.

Visceral Afferent Neural Pathways

The GIT gives detailed information about mechanical and intraluminal chemical events by the activity in visceral afferents to the brain. These visceral afferent neurons consist of three sets with different although overlapping functions:

1. About 80% of all axons in the vagal branches innervating the GIT are afferent. They are involved in regulation of motility, digestion, secretion, food intake, and possibly protection of the GIT. Their activation also triggers general sensations associated with the GIT, such as satiety, hunger, and fullness, and they are involved in the generation of emotional feelings. They are not directly involved in the generation

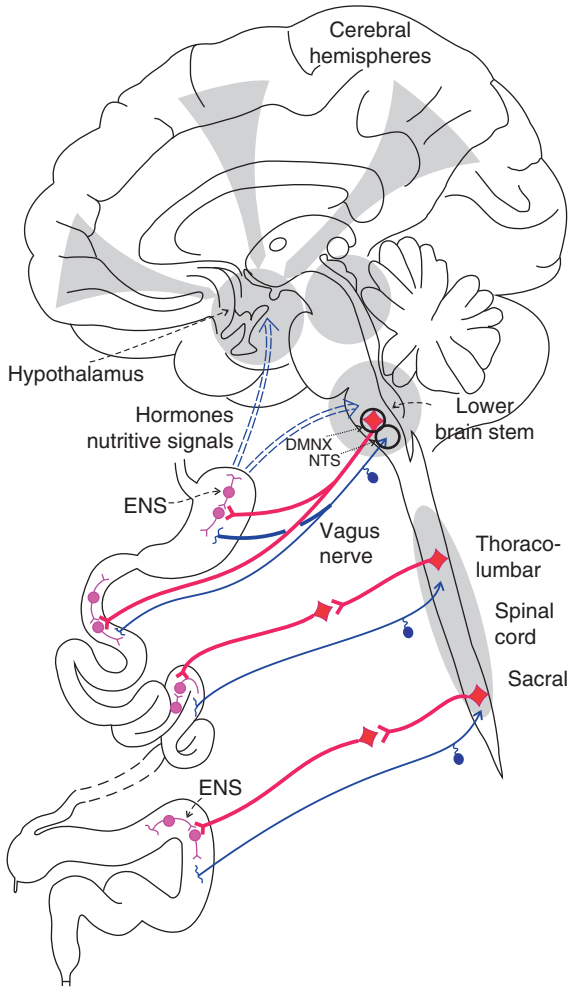

Figure 1 Afferent and efferent pathways connecting GIT (ENS, enteric nervous system) and brain. Afferent connections: vagal afferents to the lower brain stem (NTS, nucleus tractus solitarii); spinal afferents to the thoracolumbar or sacral spinal cord; and GI hormones acting on the neurons in the dorsal vagal complex via the AP in the lower brain stem and/or on the hypothalamus (mainly on the neurons of the ARN); nutritive signals (glucose and lipids in the blood) acting on neurons in the ARN and possibly the dorsal vagal complex. Efferent connections: preganglionic vagal (parasympathetic) neurons in the dorsal motor nucleus of the vagus (DMNX); sympathetic pathways originating in the intermediate zone of the thoracolumbar spinal cord; parasympathetic (spinal) pathways originating in the intermediate zone of the sacral spinal cord.

of visceral pain, although they participate in the control of transmission of nociceptive impulses in the spinal cord and possibly elsewhere. Most of them exhibit relative specificity in their responses to physiological stimulation. They respond to mechanical (distension, contraction, shearing stimuli acting on the mucosa) or/and intraluminal chemical stimuli. The chemical stimuli are mediated by enterochromaffin cells releasing serotonin, or by enteroendocrine cells releasing cholecystokinin (CCK), for example. They project viscerotopically to the nucleus tractus solitarii (NTS), the oral part of the anterior GIT being represented rostrocaudally and the caudal part (e.g., the cecum in the rat), caudomedially.

2. Thoracolumbar visceral afferent neurons innervating the GIT are involved in local and centrally organized protective reflexes, in visceral pain and discomfort, and in shaping emotions.

3. Sacral visceral afferent neurons innervating the GIT are involved in the same functions as are thoracolumbar visceral afferent neurons, but in addition they are important in the regulation of pelvic organs and in nonpainful sensation associated with the hindgut.

Afferent Hormonal and Nutritive Signals Related to the GIT

The brain receives hormonal and nutritional signals from the GIT or from cell groups (e.g., adipose tissue) related functionally to the GIT. These signals are important in the regulation of metabolism and food intake of the organism and therefore also of hunger, satiety, and body weight. Hormonal signals either act on neurons of the dorsal vagus complex (DVC) via the AP as satiety signals – CCK, glucagon-like peptide-1 (GLP-1), pancreatic polypeptide (PP), and peptide YY (PYY) (see Table 1) – or on neurons in the ARN of the hypothalamus as adipositas-related signals (leptin from the fat tissue, insulin from the endocrine pancreas, PYY from the ileum–hindgut) to inhibit food intake and metabolism or to initiate food intake (ghrelin from the stomach). Nutritional afferent signals are glucose and lipids (free fatty acids) in the blood. They act on neurons of the ARN of the hypothalamus as adipositas-related signals and possibly also on the lower brain stem via the AP.

Parasympathetic Pathways

Rather little is known about parasympathetic preganglionic DMNX neurons that project to the GIT. The DMNX is practically a motor nucleus of the GIT, and the neurons are specialized with respect to various motility, secretomotor, and enteric endocrine functions of the GIT. They may also be involved in protection of the GIT, innervating enteric circuits that regulate GI defense. They innervate motor neurons and interneurons of the ENS. Figure 2 summarizes the way in which excitation of parasympathetic preganglionic neurons in the DMNX may generate or modulate different motility patterns of the GIT. This somewhat hypothetical scheme shows that the brain does interfere, via the preganglionic parasympathetic neurons, with many different functions in which the ENS is involved, and aids in understanding the complex peripheral mechanisms that are activated by the different functionally defined classes of preganglionic

Table 1 Integration of endocrine and neural signals in the DVC in the control of GI functions

Peptide	*Main origin*	*Released by/ during*	*Effect*	*Activation/inhibition of neurons*	*Global function(s)*
CCK*+	I cells in duodenum and small intestine	Protein, fat digests in duodenum	Gastric emptying↓, exocrine pancreas↑, food intake↓, satiety↓	Activation of neurons in the cNTS	Enhancement of satiety and digestion
CRH	Circulating, CRH neurons of the paraventricular nucleus of the hypothalamus and Barrington's nucleus	Stress	Gastric motility↓, acid secretion↓, transit time small intestine↓, transit time large bowel↓	Activation of inhibitory (NANC) pathways	Response of the GIT during stress
GLP-1*#	L cells of distal small intestine	Food ingestion (proportional to meal caloric intake)	Insulin secretion↑ (glucose dependent), glucagon↓, gastric emptying↓, caloric intake↓, satiety↑	Activation of neurons of DMNX (probably NANC pathway)	Enhancement of satiety, facilitation of disposal of ingested nutrients
PP*	Endocrine cells in pancreas	Fasting in anticipation of feeding (in cephalic phase)	Gastric motility↑, gastric secretion↑, gastric transit↑	Activation of secretomotor and motility-regulating neurons in DMNX (excitatory pathways)	Anticipatory preparation of feeding
PYY*#	Endocrine cells in the ileum, colon–rectum	Fatty acids in the ileum	Acid secretion↓, gastric motility↓, pancreatic secretion↓, food intake↓	Inhibition of secretomotor and motility-regulating neurons in DMNX (excitatory pathways)	Antagonizes effect of TRH/5-HT neurons, inhibits activity of stomach
TNF-α^+	Macrophages, T lymphocytes, glia cells	Inflammation of the GIT	Gastric functions↓, nausea, emesis, anorexia	Activation of neurons in NTS and DMNX (NANC pathway)	Defense of the GIT
TRH	Circulating (thyroid gland), TRH/5-HT neurons in VMM	Cold stress	Gastric secretion↑, feeding behavior↑, metabolism↑	Activation of neurons in DMNX (excitatory pathways)	Coordination of thermoregulation and regulation of metabolism

References: general (Stanley et al., 2005); CCK (Moran and Kinzig 2004); CRH (Lewis et al., 2002; Taché et al., 2001); GLP-1 (Yamamoto et al., 2003); PP (Browning et al., (2005); Rogers et al., 1995; PYY (Yang et al., 2000); TNF-α (Emch et al., 2000; Emch et al., 2002; Hermann and Rogers 1995; Hermann et al., 2002; Travagli and Rogers 2001); TRH (Rogers et al., 1980; Rogers et al., 1995; Travagli and Rogers 2001; Yang et al., 2000).
Key to notation and abbreviations: *, also present in central neurons and/or vagal afferents (CCK); #, also acting on ARN; +, effects also, or mainly mediated, by vagal afferents; CCK, cholecystokinin; cNTS, central nucleus tractus solitarii; CRH, corticotropin-releasing hormone; DMNX, dorsal motor nucleus of the vagus; GIT, gastrointestinal tract; GLP-1, glucagon-like peptide-1; 5-HT, 5-hydroxytryptamine; NANC, nonadrenergic noncholinergic; PP, pancreatic polypeptide; TNF-α, tumor necrosis factor-α; TRH, thyrotropin-releasing hormone; VMM, ventromedial medulla; PYY, peptide YY.

parasympathetic neurons. Parasympathetic preganglionic neurons in the DMNX may influence the motility of the GIT (above all, stomach, duodenum, and ileum) via enteric neurons that excite the smooth muscle directly, enteric neurons that excite the smooth muscle cells via interstitial cells of Cajal, enteric neurons that inhibit smooth muscle cells, and enteric neurons that activate endocrine cells to release a hormone (e.g., gastrin), which in turn influences smooth muscle cells and enteric neurons involved in regulation of motility.

In addition to vagal preganglionic neurons integrated in regulation of motility of the GIT, there exist other types of parasympathetic preganglionic neurons that innervate enteric secretomotor neurons in the submucosal plexus or enteric neurons supplying endocrine cells of the GIT (e.g., G cells secreting gastrin and S cells secreting secretin; endocrine cells in the pancreas secreting insulin, glucagon, or PP). The cell bodies of these preganglionic neurons are possibly also situated in the DMNX. The organization of these peripheral pathways is poorly understood. Finally, preganglionic parasympathetic neurons that innervate the liver are involved in regulation of hepatic glucose production. Activation of these neurons inhibits hepatic glucose production.

The sacral (parasympathetic) spinal pathways to the colon–rectum are important to regulate continence

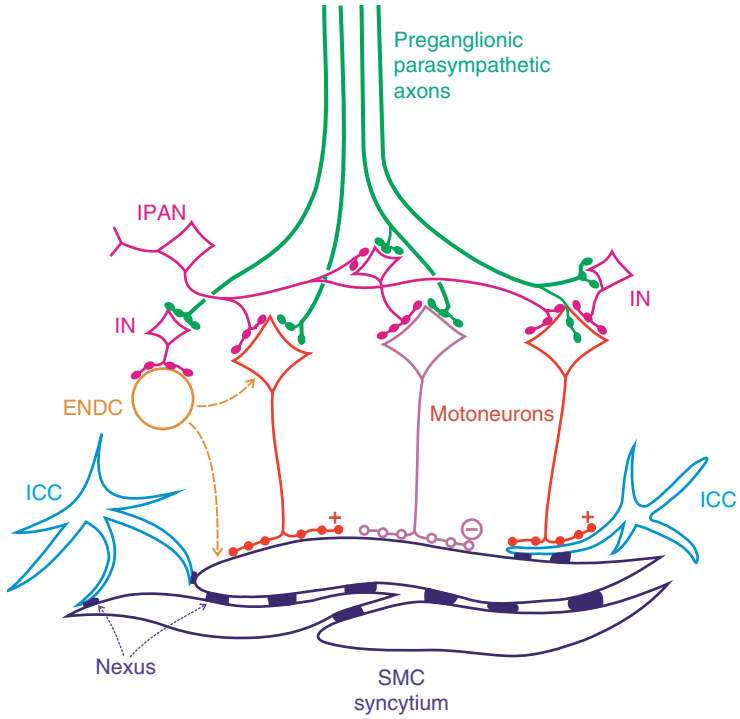

Figure 2 Diagram depicting the relation between vagal parasympathetic preganglionic neurons (cell bodies in the DMNX) and neurons of the ENS in regulation of motility. Various parasympathetic (vagal) pathways are involved in regulation of the motility of the smooth muscle syncytia of the GIT. ENDC, endocrine cell; ICC, interstitial cell of Cajal; IN, interneuron; IPAN, intrinsic primary afferent neuron; SMC, smooth muscle cell. +, excitation; −, inhibition. Adapted from Jänig W (2006) *The Integrative Action of the Autonomic Nervous System. Neurobiology of Homeostasis.* Cambridge, NY: Cambridge University Press.

and evacuation of the hindgut. Most postganglionic parasympathetic neurons do not innervate the target cells directly, but instead innervate enteric neurons in the myenteric plexus (but not in the submucosal plexus) or in the serosal plexus.

Sympathetic Pathways

The postganglionic sympathetic neurons are located in the prevertebral ganglia, and some are also in the paravertebral ganglia. Most postganglionic visceral vasoconstrictor neurons innervating the vasculature (arteries, arterioles, veins) of the GIT seem to be located in the paravertebral ganglia. They receive synapses from preganglionic neurons in the intermediate zone of the spinal cord but not from peripheral intestinofugal neurons. Furthermore, the postganglionic visceral vasoconstrictor neurons do not receive synaptic input from collaterals of spinal peptidergic primary afferent fibers (Figure 3).

The population of sympathetic nonvasoconstrictor neurons consists of motility-regulating neurons and secretomotor neurons (Figure 3). These two groups of neurons are most likely further subdifferentiated with respect to various functions of the GIT. Most of the motility-regulating neurons and secretomotor neurons innervate neurons of the myenteric or submucosal plexus, respectively. Activation of these sympathetic neurons inhibits enteric neurons; this inhibition occurs mainly presynaptically by decrease of release of excitatory transmitter, but also postsynaptically. A few postganglionic sympathetic secretomotor neurons innervate the mucosa directly and a few postganglionic sympathetic motility-regulating neurons innervate the nonsphincteric smooth muscles, leading in both cases to inhibition (decrease of contraction or secretion) when activated. Sphincter muscles are directly innervated by sympathetic postganglionic fibers and contract when these fibers are excited.

Postganglionic nonvasoconstrictor neurons in the prevertebral ganglia receive synaptic input from enteric intestinofugal neurons and form extraspinal reflex circuits. These intestinofugal neurons are cholinergic and may have vasoactive intestinal peptide co-localized. In the guinea pig, collaterals of spinal peptidergic afferents containing substance P (SP) appear to influence only postganglionic neurons in prevertebral ganglia that express somatostatin (in addition to noradrenaline). Only these postganglionic neurons, which have secretomotor function, express the tachykinin neurokinin-receptor 1 for SP (Figure 3).

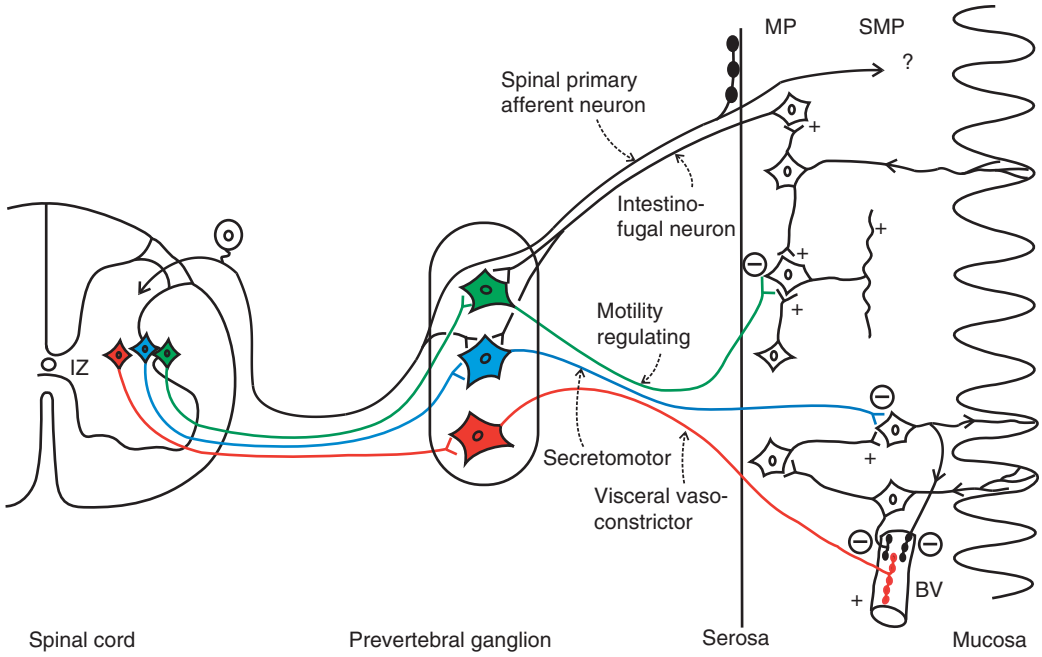

Figure 3 Sympathetic outflow to the small intestine. Most sympathetic postganglionic neurons projecting to the intestine are situated in the prevertebral ganglia. These neurons consist of three classes, each receiving separate synaptic inputs from functionally distinct classes of preganglionic neurons located in the intermediate zone (IZ) of the thoracolumbar spinal cord: vasoconstrictor neurons (red) innervating blood vessels, secretomotor neurons (blue) innervating neurons of the submucosal plexus (SMP), and motility-regulating neurons (green) innervating the myenteric plexus (MP). The vasoconstrictor neurons do not get synaptic input from collaterals of spinal primary afferent neurons (peptidergic) or from enteric intestinofugal neurons (cholinergic). The cell bodies of many visceral vasoconstrictor neurons are also located in paravertebral ganglia. The two latter classes of neurons, secretomotor and motility regulating, exert (mainly pre- but also postsynaptically) inhibitory effects on enteric neurons. They are probably divided into functional subtypes. The postganglionic nonvasoconstrictor neurons integrate preganglionic and peripheral synaptic inputs. They receive synaptic input from intestinofugal neurons of the ENS (cholinergic nicotinic). Postganglionic secretomotor neurons additionally receive peptidergic synaptic input from collaterals of spinal primary visceral afferent neurons. BV, blood vessels; +, excitation; −, inhibition. Adapted from Furness JB (2005) *The Enteric Nervous System*. Oxford: Blackwell Science Ltd.

Activity of sympathetic secretomotor neurons is related to the balance of fluid volume and electrolytes in the body. Decrease of both leads to activation of these neurons and consequently to inhibition of chloride and bicarbonate secretion by the intestinal mucosa. This is followed by a decreased transport of sodium and water into the intestinal lumen and conservation of body water. The central circuits linked to sympathetic secretomotor and motility-regulating neurons are unknown. They are unlikely to be identical to those linked to the visceral vasoconstrictor pathway.

Organization of the Dorsal Vagus Complex

The topographical organization of vagal afferents projecting to the NTS and of the efferent neurons projecting to the anterior part of the GIT is the anatomical basis of the neural control of the GIT (stomach, duodenum, small intestine) by the CNS. The NTS subnuclei related to the GIT and the ventrally located DMNX, including the AP, are also called the dorsal vagal complex (DVC) (see Figure 4(a)).

Functional Anatomy

Nucleus tractus solitarii Afferents from the subdiaphragmatic GIT project preferentially to the gelatinous, medial and commissural nuclei, but not to the central, interstitial, intermedial and ventrolateral nuclei of the NTS. Afferents from the esophagus project to the central nucleus of the NTS. The projection of afferent neurons to the NTS in the rat from the soft palate, pharynx, esophagus, stomach, and cecum is organized topographically, with limited overlap. This organization has three distinct characteristics: (1) afferents from different sections of the alimentary canal project to different subnuclei of the NTS (Figure 4(a)); (2) the afferent projections have a mediolateral and caudorostral segregation (Figure 4(b)); and (3) the organization of these afferent projections

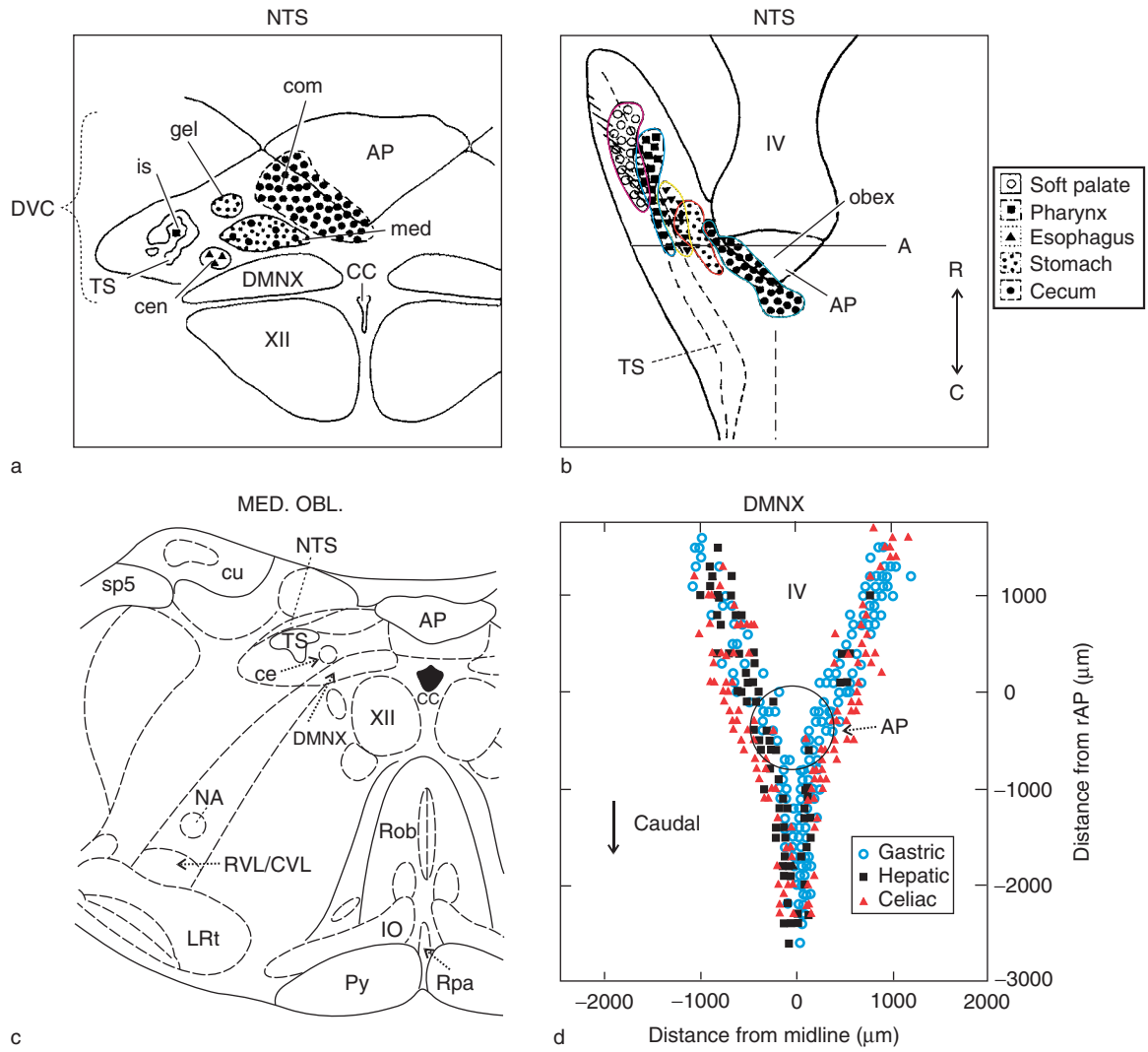

Figure 4 Topographic projection of vagal afferents from different sections of the GIT to the nucleus tractus solitarii (NTS) and topographic organization of preganglionic neurons supplying the GIT in the dorsal motor nucleus of the vagus (DMNX) of the rat. (a) Schematic transverse section through the dorsomedial medulla oblongata at the level of the area postrema (AP) containing the NTS and DMNX (location: see line A in (b)); nuclei of the NTS are central (cen), commissural (com), gelatinous (gel), interstitial (is), and medial (med). (b) Horizontal (longitudinal) section through the NTS in the rat, showing the rostrocaudal and mediolateral topography of the projections of afferents from the rostral part of the GIT up to the cecum. (c) Transverse section through the medulla oblongata at the level of the AP, showing the location of the DMNX in relation to other nuclei, in particular the NTS, and the AP. (d) The distribution of preganglionic neurons projecting through the gastric (circles), celiac (triangles), and hepatic (squares) branches of the subdiaphragmatic vagus nerves, as viewed in the horizontal plane. The neurons were labeled with Fast Blue injected intraperitoneally in rats 4–6 days before perfusing the animals. Various combinations of abdominal vagal branches (two gastric, two celiac, one hepatic) were transected, leaving one or two vagal branches intact. Fast Blue is taken up by the terminals of the intact vagal preganglionic axons and transported to the cell bodies. Preganglionic cell bodies were visualized in horizontal 100-µm-thick sections of the DMNX under ultraviolet epifluorescence, and were intracellularly filled via a glass micropipette with the dye Lucifer yellow. In this way, cell bodies, dendrites, and axons of the neurons were visualized and reconstructed. CC, central canal; cu, cuneate nucleus; DVC, dorsal vagal complex; IO, inferior olive; LRt, lateral reticular nucleus; NA, nucleus ambiguus; Py, pyramid; Rob, raphe obscurus; Rpa, raphe pallidus; RVL/CVL, rostral ventrolateral medulla/caudal ventrolateral medulla; sp5, spinal trigeminal tract; TS, tractus solitarius; IV, fourth ventricle; XII, hypoglossus. (a, b) Adapted from Altschuler SM, Ferenci DA, Lynn RB, et al. (1991) Representation of the cecum in the lateral dorsal motor nucleus of the vagus nerve and commissural subnucleus of the nucleus tractus solitarii in rat. *Journal of Comparative Neurology* 304: 261–274. (c) Adapted from Paxinos G and Watson C (1998) *The Rat Brain*. San Diego: Academic Press. (d) Modified from Fox EA and Powley TL (1992) Morphology of identified preganglionic neurons in the dorsal motor nucleus of the vagus. *Journal of Comparative Neurology* 322: 79–98.

is related to the mediolateral organization of the more ventrally located DMNX preganglionic neurons that project to the GIT.

The projection pattern of the GI vagal afferents to the subnuclei of the NTS is distinct from the projection pattern of cardiovascular afferent neurons and afferent neurons innervating the respiratory tract, although there is no simple relation between the subnuclei in the NTS and the projecting afferents from the three systems (cardiovascular, GI, respiratory). Furthermore, it is unknown whether functionally distinct types of vagal afferents from the GIT (mechanosensitive, chemosensitive, etc.) exhibit distinct projection patterns.

Synaptic transmission from vagal afferents to the NTS neurons is excitatory and the transmitter is glutamate. In the rostrocaudal area of the NTS, to which vagal afferents project, some 40–50% of the NTS neurons are synaptically activated by electrical stimulation of the subdiaphragmatic vagal nerve. Inhibitory responses are practically absent. Neurons activated from the GIT do not seem to receive additionally convergent synaptic input from arterial baro- or chemoreceptor afferents.

Dorsal motor nucleus of the vagus Preganglionic parasympathetic neurons innervating the proximal part of the GIT (stomach, duodenum, and small intestine, including pancreas and liver) are located in the DMNX. This nucleus is the motor nucleus of the foregut, in which the stomach has the highest representation. In the rat, it contains about 10 000 neurons (5000 neurons on each side), 7500 neurons being preganglionic and projecting to the GIT; the remaining neurons are interneurons or other types of preganglionic neurons projecting to visceral organs in the thoracic cavity.

The preganglionic DMNX neurons projecting through the subdiaphragmatic vagus nerves to the GIT are topographically organized into rostrocaudal cell columns. The neurons projecting in the gastric branches are located medially and the neurons projecting in the celiac branches are located laterally. Neurons projecting in the hepatic branch are located in a cell column between the medial (gastric) and the lateral (cecal) cell column of the left DMNX. The dendrites of these preganglionic neurons are almost confined to the cell columns projecting rostrocaudally, but also extend dorsally to the subnuclei of the NTS (**Figures 4(c) and 4(d)**). Most preganglionic parasympathetic neurons innervating the lower esophageal sphincter probably are also located in a medial rostrocaudal cell column of the DMNX.

Intestino-Intestinal Reflex Circuits in the DVC

The intimate anatomical relationship between the NTS and the DMNX, both forming, together with the AP (see later), the DVC, implies that afferents from the GIT form disynaptic contacts with the preganglionic neurons projecting to the GIT. Systematic neurophysiological analysis of preganglionic DMNX neurons that project to the GIT have shown that several functional types of preganglionic neuron can be identified by their reflex responses to physiological stimuli applied to stomach, duodenum, or small intestine – for example, distension of organ sections (esophagus, stomach, duodenum) and intraluminal chemical stimulation. The patterns of reflex responses are correlated with other properties of the neurons, such as the rate of spontaneous activity, morphology of dendrites, and peptide content. At least four functional types of preganglionic neurons have been identified in the DMNX, indicating the functional differentiation of the parasympathetic pathways that are represented in the DMNX and project to the GIT.

The predominant synaptic transmission between the NTS neurons and the preganglionic DMNX neurons projecting to the GIT is inhibitory, the transmitter being mainly γ-aminobutyric acid. Other synaptic connections between NTS neurons and DMNX neurons are excitatory, the transmitter probably being glutamate. Gastric relaxation induced by esophageal distension is mediated by a reflex pathway involving the central nucleus of the NTS and the DMNX. NTS neurons involved in this reflex are noradrenergic, and this receptive relaxation reflex is significantly reduced after blockade of α-adrenoceptors (α_1, α_2) in the DMNX. Of the DMNX neurons projecting to the stomach, 75% are either excited or inhibited by noradrenaline. Thus noradrenaline is believed to be the transmitter mediating this reflex in the DMNX.

For most of these parasympathetic pathways, it is at present uncertain or a matter of speculation as to which type of target cell or functionally distinct enteric circuit they innervate. However, it is reasonable to assume that circuits of the ENS, which are related to nonvascular smooth muscles, exocrine glands, endocrine cells, or even the vasculature, receive differential signals from parasympathetic preganglionic neurons in the DMNX (see **Figure 2**). Thus it is also not far-fetched to assume that there are many more than four functional types of preganglionic parasympathetic neurons in the DMNX. Based on cyto- and chemoarchitectonic criteria, the DMNX in the human contains nine types of neuron.

Integration of Endocrine Signals by the AP in Neural Control of GI Functions

The reflex neural circuits in the DVC are also modulated by bloodborne signals that reach the neurons of the DVC by way of the AP and neighboring parts of the DVC. These signals are, for example, hormones from the GIT, such as GLP-1, PP, or PYY, or other hormones, such as corticotropin-releasing hormone (CRH) or thyrotropin-releasing hormone (TRH), or cytokines (e.g., tumor necrosis factor-α(TNF-α)). The AP and neighboring medial parts of the DVC do not have a vascular diffusion barrier, thus the blood–brain barrier is absent (Figure 5), allowing the hormonal signals (peptides) to have free access to the neural circuits of the DVC. The dendrites of both NTS neurons and DMNX neurons project into the area that lacks a blood–brain barrier. The circuits in the DVC are furthermore influenced by neuronal circuits that use neuroendocrine hormones as transmitters or neuromodulators (such as TRH or CRH).

Table 1 shows examples that demonstrate the close integration, in the lower brain stem, of autonomic (parasympathetic) systems and neuroendocrine systems in the regulation of GI functions. Each of the peptides listed has multiple functions, related to the GIT, that are mediated by peripheral neural and hormonal pathways and central pathways involving the DVC. With the exception of TNF-α, these peptides are also localized in neurons or act additionally or mainly via vagal afferents on the DVC (e.g., CCK). They may also act on the hypothalamus via circumventricular organs. Some circulating peptides acting mainly on the hypothalamus in the context of regulation of metabolism and satiety, such as leptin, ghrelin or insulin, are not listed in Table 1. However, these peptides may also act on the neural circuits of the DVC via the AP. It is important to note that each peptide listed in Table 1 (see last column in the table) represents an organizing principle for global functions in which the brain–gut axis is involved.

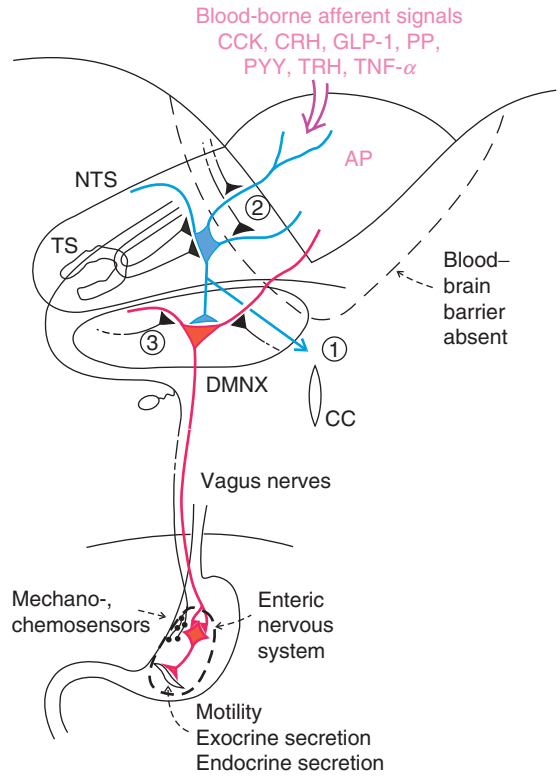

Figure 5 Dorsal vagal complex basic vago-vagal reflex circuits involved in regulation of functions of the GIT, and their modulation by hormones via the area postrema (AP). Neurons in the nucleus tractus solitarii (NTS) form inhibitory or excitatory synapses with preganglionic neurons in the dorsal motor nucleus of the vagus (DMNX) and are synaptically activated from the periphery by mechano- and/or chemosensitive vagal afferent neurons. The preganglionic neurons project through the vagus nerves to the ENS and have various functions. NTS neurons project additionally to various regions of the brain stem, hypothalamus, and telencephalon (1). Both neurons in the NTS and neurons in the DMNX are under multiple synaptic controls from brain stem, hypothalamus, and telencephalon (2 and 3). Signals in the blood (e.g., hormones) have access to the vago-vagal circuits via the AP and neighboring areas in which the blood–brain barrier is absent. CC, central canal; CCK, cholecystokinin; CRH, corticotropin-releasing hormone; GLP-1, glucagon-like peptide-1; PP, pancreatic polypeptide; PYY, peptide YY; TNF-α, tumor necrosis factor-α; TS, tractus solitarius; TRH, thyrotropin-releasing hormone. Adapted from Rogers RC, McTigue DM, and Hermann GE (1995) Vagovagal reflex control of digestion: Afferent modulation by neural and endoneurocrine factors. *American Journal of Physiology* 268: G1–G10, and Travagli RA and Rogers RC (2001) Receptors and transmission in the brain–gut axis: Potential for novel therapies. V: Fast and slow extrinsic modulation of dorsal vagal complex circuits. *American Journal of Physiology* 281: G595–G601.

Control of the DVC by Upper Brain Stem and Forebrain: A Concept

In order to account for the many types of functional reflexes mediated by the preganglionic neurons in the DMNX – as defined by the excitatory or inhibitory effect on the different target cells (different types of smooth muscle cells, exocrine glands, endocrine glands etc.) – and generated by the different types of vagal afferent neurons from the GIT, Powley and co-workers have created an interesting hypothesis on the spatial organization of the DVC. This hypothesis is based on the mediolateral representation of the GIT in the DMNX and on the rostrocaudal organization of the projections of vagal afferents to the NTS. It is proposed that these two layers of the DVC form a sensorimotor lattice that would allow for many specific reflexes elicited in the parasympathetic preganglionic neurons, by stimulation of distinct types of

afferents from the GIT, and for various combinations (patterns) of reflexes.

The functionally distinct reflex circuits formed in the DVC between the afferents from and the preganglionic neurons to the GIT are the basic building blocks for the brain to control GI functions (Figure 6). The idea of the sensorimotor lattice of the DVC as proposed by Powley is a heuristic model and an approximation and does not apply to all situations. For example, esophageal distension elicits proximal gastric relaxation, the basis being the 'receptive relaxation reflex' arc consisting of vagal afferents from the esophagus, the second-order neurons in the pars centralis of the NTS (NTS_{cen}) that receive synaptic input from these afferents, and the vagal preganglionic neurons projecting to the stomach (probably two pathways). It has been shown that the neurons in the NTS_{cen} project extensively throughout the rostrocaudal DMNX. Thus the interesting hypothesis of the sensorimotor lattice in the DVC has to be tested rigorously before it can be accepted universally.

Anatomical studies have shown that several nuclei in the brain stem, hypothalamus, and telencephalon have reciprocal connections with the circuits of the DVC. The functions of most of these neural connections are poorly understood. This situation is conceptually quite similar to the role of spinal autonomic circuits in the control of spinal autonomic final pathways by supraspinal centers. Thus, these reflex pathways of the DVC are under modulatory control of neurons in the medulla oblongata and supramedullary brain centers (so-called 'executive' neurons; for example, neurons in the ventrolateral and ventromedial medulla, in the parabrachial and Kölliker–Fuse nuclei and A5 area, in the paraventricular nucleus of the hypothalamus and lateral hypothalamus, in the central nucleus of the amygdala and bed nucleus of the stria terminalis, and in the insula and medial prefrontal cortex), which also receive detailed afferent information from the GIT (via the NTS), from other visceral organs, and from somatic body domains. Executive neurons and basic autonomic circuits in the DVC associated with the GIT represent the internal state of the organism as far as the GIT is concerned (Figure 6).

An essential component of this internal state of the organism is the feedback by hormonal and nutritive signals from the GIT to the DVC via the AP and to the hypothalamus (mainly via the ARN). This hormonal feedback consists of several components (see Table 1 for the DVC) and is integrated by the brain into the regulation of the metabolic (nutritional) state, the thermoregulatory state, the fluid matrix, the reproductive state, and the protective state of the organism.

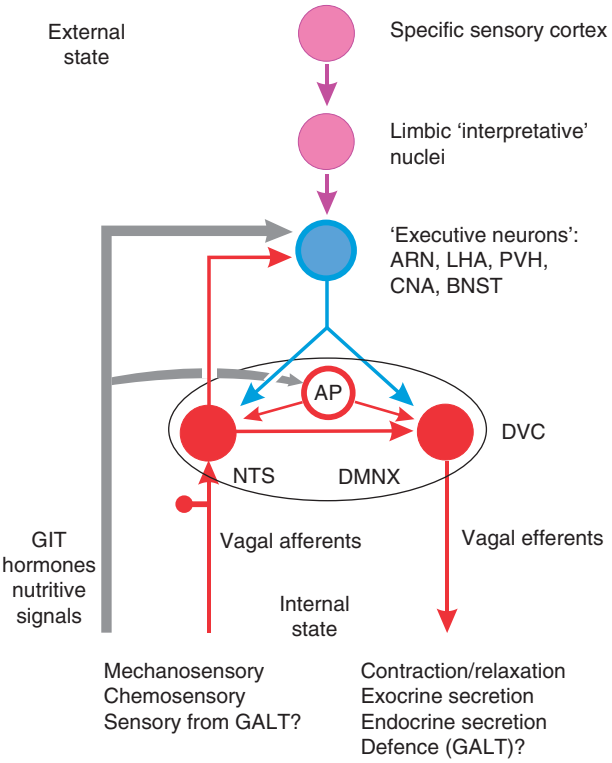

Figure 6 Proposed relationship between gastrointestinal vago-vagal reflex pathways, autonomic 'executive' neuronal circuits, limbic 'interpretative' neuronal circuits, and exterosensory systems (special cortical sensory systems). Several functionally specific vago-vagal reflex pathways organized in the dorsal vagal complex (DVC) of the medulla oblongata are the basic neuronal building blocks of the control of the gastrointestinal tract (GIT) by the brain. Vagal afferents measure mechanosensory, chemosensory, and other sensory events – for example, in the gut-associated lymphoid tissue (GALT) – and project to the nucleus tractus solitarii (NTS); vagal preganglionic neurons are located in the dorsal motor nucleus of the vagus (DMNX) and are involved in regulation of motility, exocrine secretion, endocrine secretion, and probably other events related to protection of the GIT (e.g., those associated with the GALT). Multiple hormonal afferent inputs from the GIT occur via GI hormones to the DVC, via the area postrema (AP), and to the hypothalamus, mostly via the arcuate nucleus (ARN) of the hypothalamus. Nutritive signals in the blood (glucose, lipids) act via the AP and ARN on the neural circuits in the DVC and the hypothalamus, respectively. Neurons of the 'executive' centers – for example, ARN, lateral hypothalamic and parafornical area (LHA), paraventricular nucleus of the hypothalamus (PVH), central nucleus of the amygdala (CNA), and bed nucleus of the stria terminalis (BNST) – evaluate the state of the internal milieu (by way of inputs from visceral afferents and hormonal inputs) as well as the current or anticipated behavioral state (via input from limbic nuclei that evaluate the significance of exteroceptive signals). These executive centers adapt the internal state (e.g., GI functions) to the behavioral state of the organism. Adapted from Rogers RC and Hermann GE (1992) Central regulation of brainstem gastric vago-vagal control circuits. In: Ritter S, Ritter RC, and Barnes CD (eds.) *Neuroanatomy and Physiology of Abdominal Vagal Afferents*, pp. 99–134. Boca Raton: CRC Press.

These regulations are homeostatic, involving the lower brain stem and the hypothalamus. However, they are adapted to the external state of the organism by the telencephalon, as occurring, for example, during heavy exercise, extreme environmental temperature changes, food and fluid deprivation, and exposure to toxic compounds or invasion by bacteria (sepsis, toxemia). The temporary adaptation of the homeostatic regulation to potentially life-threatening situations is called allostasis. The extra costs to meet these allostatic adaptations are called the allostatic load. Disease may result if this allostatic adaptation of the homeostatic regulations either does not occur or cannot be switched off and is too strong once it has been initiated and is not any longer needed or is switched on by trivial events that are not threatening for the organism.

The internal state is adapted to the behavior of the organism by cortical and limbic system structures that monitor and represent the external state of the organism (Figure 6). However, the internal state also modulates the central representation of the external state, leading to changes of sensory perception, body feelings, and emotional responses. This concept shows that there is a close integration between the homeostatic regulation of GIT functions and the higher nervous system functions related to body perception, emotions, and adaptation of behavior. The highly specific autonomic reflex pathways in the brain stem, hypothalamus, and cerebral hemispheres are the basis of this integration. This general concept of the control of GI functions by the brain has been propagated and worked out by Rogers and co-workers, and is the biological basis for the changes of GI function during various behaviors, including stress.

See also: Autonomic Nervous System: Gastrointestinal Control; Autonomic Nervous System: Neuroanatomy; Autonomic Nervous System; Gastrointestinal Tract role in Neural Control of Metabolism, Food Intake and Body Weight: A Summary; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Altschuler SM, Ferenci DA, Lynn RB, et al. (1991) Representation of the cecum in the lateral dorsal motor nucleus of the vagus nerve and commissural subnucleus of the nucleus tractus solitarii in rat. *Journal of Comparative Neurology* 304: 261–274.

Berthoud HR and Neuhuber WL (2000) Functional and chemical anatomy of the afferent vagal system. *Autonomic Neuroscience* 85: 1–17.

Browning KN, Coleman FH, and Travagli RA (2005) Effects of pancreatic polypeptide on pancreas-projecting rat dorsal motor nucleus of the vagus neurons. *American Journal of Physiology* 289(2): G209–G219.

Emch GS, Hermann GE, and Rogers RC (2002) Tumor necrosis factor-alpha inhibits physiologically identified dorsal motor nucleus neurons *in vivo*. *Brain Research* 951: 311–315.

Emch GS, Hermann GE, and Rogers RC (2005) TNF-alpha activates solitary nucleus neurons responsive to gastric distension. *American Journal of Physiology* 279: G582–G586.

Fox EA and Powley TL (1992) Morphology of identified preganglionic neurons in the dorsal motor nucleus of the vagus. *Journal of Comparative Neurology* 322: 79–98.

Furness JB (2005) *The Enteric Nervous System*. Oxford: Blackwell Science Ltd.

Hermann G and Rogers RC (1995) Tumor necrosis factor-alpha in the dorsal vagal complex suppresses gastric motility. *Neuroimmunomodulation* 2: 74.

Hermann RC, et al. (2002) *American Journal of Physiology* 283: G634.

Jänig W (2005) Vagal afferents and visceral pain. In: Undem B and Weinreich D (eds.) *Advances in Vagal Afferent Neurobiology*, pp. 461–489. Boca Raton: CRC Press.

Jänig W (2006) *The Integrative Action of the Autonomic Nervous System. Neurobiology of Homeostasis*. Cambridge, NY: Cambridge University Press.

Lewis MW, Hermann GE, Rogers RC, and Travagli RA (2002) *In vitro* and *in vivo* analysis of the effects of corticotropin releasing factor on rat dorsal vagal complex. *Journal of Physiology* 543: 13.

McEwen BS and Wingfield JC (2003) The concept of allostasis in biology and biomedicine. *Hormones and Behavior* 43: 2–15.

Moran TH and Kinzig KP (2004) Gastrointestinal satiety signals. II: Cholecystokinin. *American Journal of Physiology* 286: G183.

Paton JF and Kasparov S (2000) Sensory channel specific modulation in the nucleus of the solitary tract. *Journal of the Autonomic Nervous System* 80: 117–129.

Paxinos G and Watson C (1998) *The Rat Brain*. San Diego: Academic Press.

Powley TL, Berthoud HR, Fox AP, et al. (1992) The dorsal vagal complex forms a sensory-motor lattice: The circuitry of gastrointestinal reflexes. In: Ritter S, Ritter RC, and Barnes CD (eds.) *Neuroanatomy and Physiology of Abdominal Vagal Afferents*, pp. 55–79. Boca Raton: CRC Press.

Rogers RC and Hermann GE (1992) Central regulation of brainstem gastric vago-vagal control circuits. In: Ritter S, Ritter RC, and Barnes CD (eds.) *Neuroanatomy and Physiology of Abdominal Vagal Afferents*, pp. 99–134. Boca Raton: CRC Press.

Rogers RC, Hitoshi K, Larry B, and Donald N (1980) Afferent projections to the dorsal motor nucleus of the vagus. *Brain Research Bulletin* 5: 365–373.

Rogers RC, McTigue DM, and Hermann GE (1995) Vagovagal reflex control of digestion: Afferent modulation by neural and endoneurocrine factors. *American Journal of Physiology* 268: G1–G10.

Stanley S, Wynne K, McGowan B, et al. (2005) Hormonal regulation of food intake. *Physiological Reviews* 85: 1131–1158.

Taché Y, Martinez V, Mulugeta M, and Wang L (2001) Stress and the gastrointestinal tract. III: Stress-related alterations of gut motor function: Role of brain corticotropin-releasing factor receptors. *American Journal of Physiology* 280: G173–G177.

Travagli RA, Hermann GE, Browning KN, et al. (2006) Brainstem circuits regulating gastric function. *Annual Review of Physiology* 68: 279–305.

Travagli RA and Rogers RC (2001) Receptors and transmission in the brain-gut axis: Potential for novel therapies. V: Fast and slow extrinsic modulation of dorsal vagal complex circuits. *American Journal of physiology* 281: G595–G601.

Udem B and Weinreich D (2005) *Advances in Vagal Afferent Neurobiology.* Boca Raton: CRC Press.

Yamamoto H, Kishi T, Lee CE, et al. (2003) Glucagon-like peptide-1-responsive catecholamine neurons in the area postrema link peripheral glucagon-like peptide-1 with central autonomic control sites. *Journal of Neuroscience* 23: 2939.

Yang H, Kawakubo K, Wong H, Ohning G, and Tache Y (2000) Peripheral PYY inhibits intracisternal TRH-induced gastric acid secretion by acting in the brain. *American Journal of Physiology* 279: G575–G581.

Yang H, Yuan PQ, Wang L, and Taché Y (2000) Activation of the parapyramidal region in the ventral medulla stimulates gastric acid secretion through vagal pathways in rats. *Neuroscience* 95: 773–779.

Autonomic Nervous System: Central Respiratory Control

A V Gourine and K M Spyer, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Breathing is the vital automatic rhythmic activity that ensures appropriate rates of oxygen uptake and carbon dioxide elimination. Neurophysiological data indicate that the respiratory rhythm is a coordinated alternation of inspiratory, postinspiratory, and expiratory activities. Their coordination controls lung ventilation and is also important for vocalization, swallowing, and motor control. Respiratory activity in mammals is generated by a network of neurons located within the lower brain stem – in the medulla and pons. The respiratory neuronal network has to be fully operational before birth and must ensure adaptation and survival of the organism in variable environmental and physiological conditions. In this article we discuss the central nervous mechanisms controlling breathing (Figure 1). We describe the anatomy of the respiratory network, discuss mechanisms of respiratory rhythm generation and cardiorespiratory interactions, and focus on the control of breathing exerted by chemo- and mechanoafferent inputs. We also give a brief overview on how central nervous control of respiratory activity is modified in some common physiological states.

Central Nervous Control of Respiratory Activity

Anatomical Organization

The core neuronal network responsible for the generation and shaping of the respiratory rhythm, as well as transmitting this rhythm to the spinal motor neurons controlling respiratory muscles, is located in the lower brain stem – within a bilaterally organized dorsal respiratory group and ventral respiratory column (VRC) of neurons in the pons and the medulla oblongata (Figure 2).

The VRC is essential and sufficient to generate the respiratory rhythm. It extends rostrocaudally in the ventrolateral medulla oblongata in close proximity to the nucleus ambiguus. The VRC consists of 'distributed' neuronal circuits, involving assemblies of adjacent classes of respiratory neurons. The VRC can be divided into several functional compartments, including (from rostral to caudal) the overlapping retrotrapezoid nucleus/parafacial respiratory group (RTN/pFRG), the Bötzinger complex, the pre-Bötzinger complex (PBC), and rostral and caudal ventral respiratory groups (Figure 2). These functional divisions of the VRC have been identified using specific neurochemical markers and physiologically by monitoring changes in respiratory activity following activation and/or blockade of neurons in these regions of the brain stem. Lesioning of the PBC has identified it as the principal kernel for generating respiratory activity. It contains neurons that are essential and sufficient for generation of the basic inspiratory rhythm. The PBC can be visualized immunohistochemically using antibodies directed against neurokinin-1, μ-opioid receptors, somatostatin, or connexins. Recent evidence suggests that RTN/pFRG may contribute to the generation of the respiratory rhythm – it appears to produce an expiratory rhythm, in addition to inspiratory activity that is generated by the PBC. It has been suggested that the activity of the latter is dominant in mammals at rest, when there is little or no active expiration (discussed in more detail later). The other functional compartments of the VRC are involved in shaping the basic rhythm (produced by PBC), generation of the respiratory pattern and relaying this pattern to cranial and spinal resistance and respiratory motor neurons. Neuronal interactions within the dorsolateral pontine parabrachial and Kölliker–Fuse nuclei (pontine respiratory group; referred to as 'pneumotaxic center' in the older literature) (Figure 2) provide input to the VRC and play an important role in reflex regulation and shaping of respiratory activity patterns.

Activity of the brain stem respiratory network is powerfully influenced by peripheral afferent inputs originating from the arterial chemoreceptors, which are sensitive to changes in the arterial pH and levels of oxygen and carbon dioxide ($[H^+]$, PaO_2, $PaCO_2$), and from the mechanoreceptors that provide information about the mechanical status of the respiratory tract. Both arterial chemoreceptor afferents and afferents originating from the pulmonary mechanoreceptors terminate in the dorsal medullary nucleus tractus solitarii (NTS). NTS neuronal circuits which involve neurons within the dorsal respiratory group (located in the ventrolateral NTS) integrate these afferent inputs, modulate respiratory activity patterns in relation to these inputs, and contribute in particular to promoting the transition from inspiration to expiration.

The primary central CO_2 chemosensitive areas are located on the ventral surface of the medulla just beneath the VRC and in the close proximity to the brain stem respiratory rhythm and pattern generator (Figure 2). There is evidence suggesting that the PBC, pontile regions, and NTS may also contain functional respiratory CO_2 chemoreceptors.

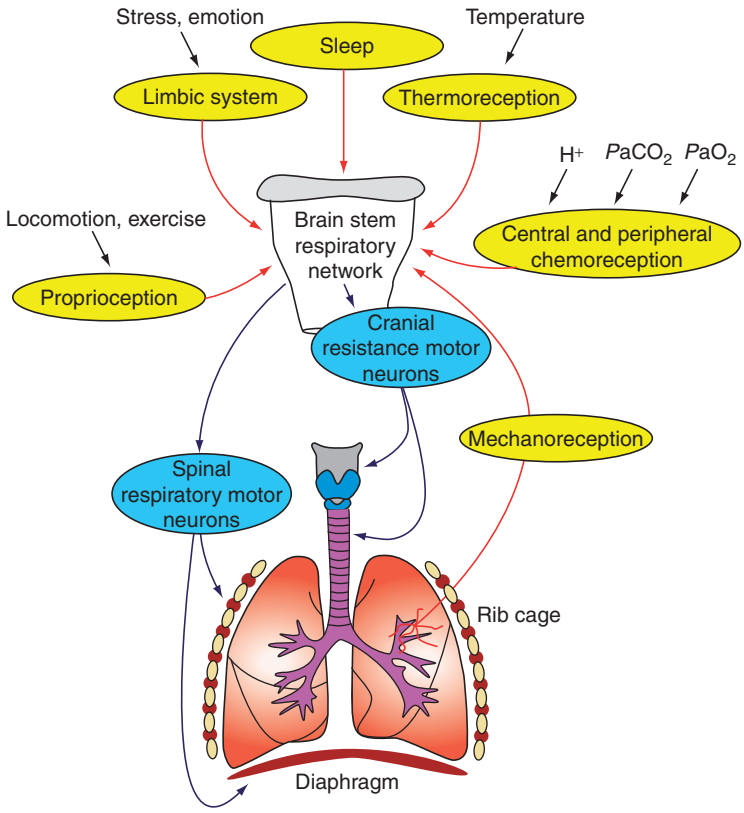

Figure 1 Regulation of breathing by the central nervous system: main afferent inputs and motor output. Refer to the text for details.

Generation of the Respiratory Rhythm

Neurophysiological data indicate that the medullary respiratory neuronal network oscillates in a three-phased respiratory pattern (Figure 3). These three phases are defined as inspiration, postinspiration (passive expiration), and expiration (active expiration). From a motor control perspective these phases represent successive activation of different respiratory muscles. During inspiration, the diaphragm and other inspiratory muscles contract. During postinspiration, contraction of the inspiratory muscles ceases and laryngeal muscles contract to control the outflow of the air from the lungs. Expiratory intercostal and abdominal muscles contract if there is an active expiration. Breathing in mammals is dominated by inspiration, and at least two antagonistic phasic activities are necessary – for example, inspiratory and postinspiratory activities – during rapid shallow breathing. These phases are usually supplemented by a third, expiratory activity, which is either weak or absent during quiet breathing. The respiratory network is dominated by synaptic inhibition, which determines the phases of the respiratory cycle. Most of the medullary respiratory neurons show evidence of synaptic inhibition at some time of the respiratory cycle. When synaptic inhibition is blocked experimentally, phase relationships between inspiratory, postinspiratory, and expiratory activities are markedly disturbed; however, as discussed later, generation of the respiratory rhythm *per se* is not completely abolished.

This brain stem respiratory neuronal network has the ability to generate different respiratory patterns (Figure 3). Normal breathing, or eupnea, is characterized by a slow augmenting sequential activation of the respiratory muscles innervated by cranial and spinal nerves during inspiration and expiration. Gasping is an inspiratory pattern characterized by a rapid decrementing synchronous activation of cranial and spinal respiratory muscles with minimal expiratory activity. Gasping is expressed in pathological conditions – for example, during severe hypoxia – and serves as a powerful mechanism of autoresuscitation. Emotions, partial lung collapse, and some other conditions increase the incidence of large-amplitude inspirations or sighs (augmented breaths), which occur on top of the 'normal' eupneic inspirations. There is evidence that the PBC neuronal network isolated within a brain stem slice is capable of generating all these qualitatively different breathing patterns, suggesting that normal breathing, gasps, and sighs may result from reconfiguration of activity of

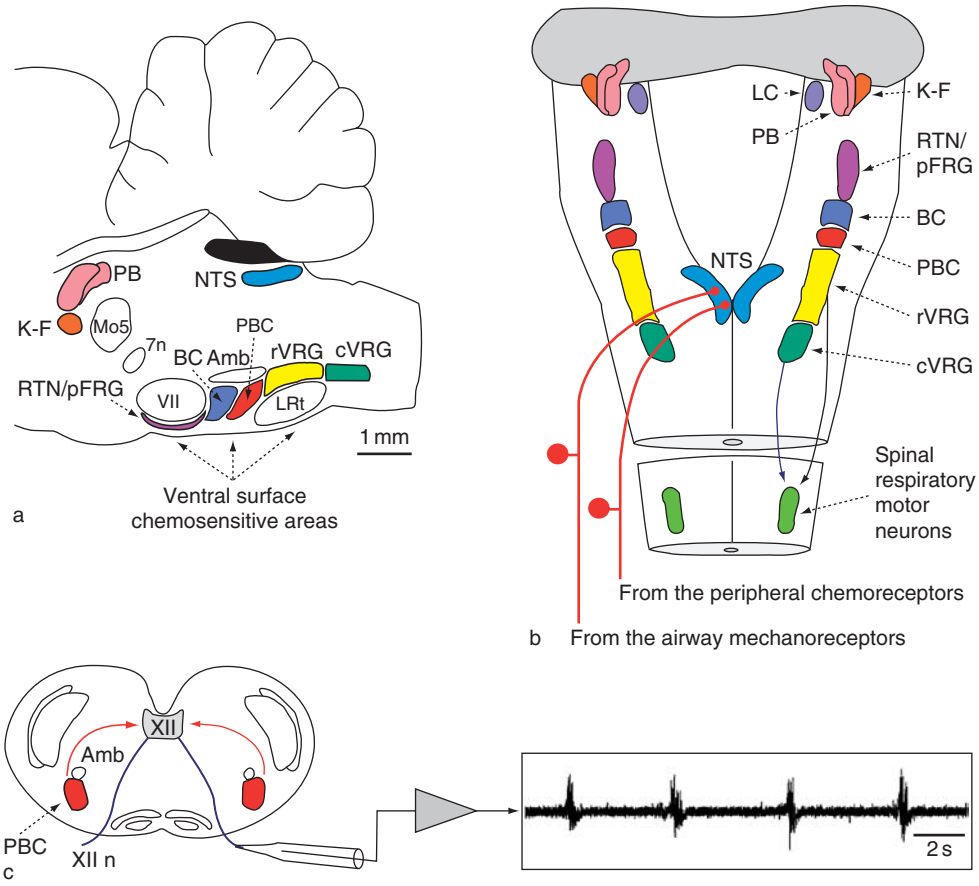

Figure 2 Neuroanatomy of the brain stem respiratory network. Sagittal (a) and horizontal (b) views of the rat brain stem, showing locations of the main groups of respiratory neurons in the mammalian central nervous system. (c) Diagram of a 'rhythmic' 300- to 700-μm-thick transverse slice prepared from the neonatal rat medulla oblongata and containing PBC, which retains rhythmic activity after *in vitro* isolation. Rhythm generated by the PBC is transmitted to the XIIth nucleus allowing recordings of inspiratory motor discharges with a suction electrode from the rootlet of the hypoglossal nerve (XII n). VII, facial nucleus; XII, hypoglossal nucleus; Amb, nucleus ambiguus; BC, Bötzinger complex; cVRG, caudal ventral respiratory group; K-F, Kölliker–Fuse nucleus; LC, locus coeruleus; LRt, lateral reticular nucleus; Mo5, motor trigeminal nucleus; NTS, nucleus tractus solitarii; PB, parabrachial nucleus; PBC, pre-Bötzinger complex; RTN/pFRG, retrotrapezoid nucleus/parafacial respiratory group; rVRG, rostral ventral respiratory group.

the same rhythmogenic neuronal circuits. Underlying synaptic and cellular (membrane) mechanisms responsible for generation of different respiratory patterns appear to be quite distinct. During severe hypoxia, gasping is induced when synaptic inhibitory mechanisms (essential for eupnea) are suppressed. It has also been shown that gasping relies on persistent sodium current (which is augmented by hypoxia), whereas the generation of eupnea may not require this cellular mechanism.

As mentioned previously, the PBC is believed to contain the principal kernel, or 'noeud vital,' for producing basic inspiratory rhythm (Figure 2). The central role of the PBC in neurogenesis of breathing was demonstrated in experimental animals from evidence showing that the rhythmic respiratory output from the brain stem disappears when the PBC is destroyed. All PBC neurons which discharge with preinspiratory firing pattern (i.e., those neurons which fire during inspiration and for a short period before that) express neurokinin-1 receptors. In the adult rat there are approximately 300 such neurons per side, and selective experimental bilateral elimination of >80% of them results in an abnormal (ataxic) breathing pattern. Less extensive loss of neurokinin-1-expressing PBC neurons leads to a significant disruption of breathing during sleep.

The PBC neuronal circuitry continues to generate rhythmic inspiratory activity when isolated *in vitro* within coronal or sagittal brain stem slices prepared from the brains of neonatal rodents (Figure 2). Although the respiratory network is known to be dominated by synaptic inhibition, *in vitro* respiratory rhythm continues after blockade of postsynaptic inhibition, suggesting that pacemaker neurons may generate this rhythm. Indeed, respiratory pacemaker-like neurons have been identified within the PBC, supporting the pacemaker theory for central respiratory

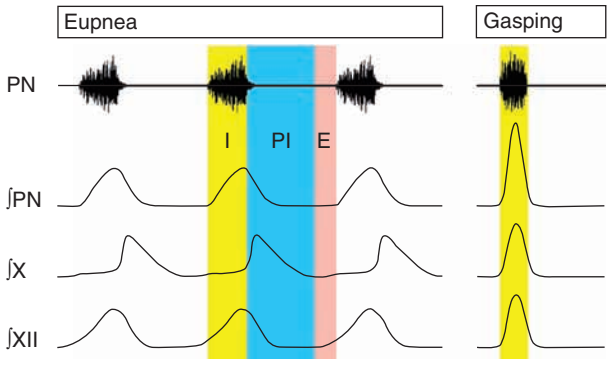

Figure 3 Three-phased respiratory activity. Schematic showing the activity patterns of spinal (phrenic) and selected cranial (vagus and hypoglossal) nerves during normal breathing (eupnea) and hypoxic gasping. Medullary respiratory neurons which discharge during either inspiration, postinspiration, or expiration can be recorded in various functional divisions of the VRS. Note how activity patterns of cranial nerves change during hypoxia-induced gasping (refer to the text for details). PN, activity of the phrenic nerve; ∫PN, integrated activity of the phrenic nerve; ∫X, integrated activity of the vagus nerve; ∫XII, integrated activity of the hypoglossal nerve; I, inspiratory phase; PI, postinspiratory phase; E, expiratory phase.

rhythm generation. This theory postulates that neurons with voltage-dependent pacemaker properties (they show spontaneous rhythmic oscillations in membrane potential) are responsible for generation of the respiratory rhythm. Voltage-independent pacemaker neurons have also been identified, suggesting that the PBC pacemaker populations are heterogeneous, with various burst-generating conductance mechanisms. However, the proportions of both populations of pacemaker-like neurons within the region are rather low. Furthermore, a stable rhythm still remains in slices in which the individual key ion channels for pacemaker activity are blocked, suggesting that respiratory rhythmic activity could be generated by neuronal networks exhibiting oscillatory behavior through chemical and electrotonic synaptic interactions.

At present many investigators favor a unifying 'hybrid pacemaker-network model' for respiratory rhythm generation; this model suggests that under normal conditions regular respiratory activity results from excitatory and inhibitory synaptic interactions between respiratory neurons within a network which shape the drive of the pacemaker units embedded within this network. Finally, when developmental changes in the operational conditions of the respiratory network are also taken into account, a 'maturational network-bursting theory' has been developed, which proposes that the principles for respiratory rhythm generation change significantly as the network matures. This concept emerged when data obtained using various neonatal and adult preparations were analyzed systematically. Understanding the connectivity within the PBC respiratory network has also been aided significantly by mathematical modeling. However, the detailed neuronal organization of the PBC remains to be very difficult to dissect experimentally.

There is evidence indicating that there might be more than one central respiratory rhythm generator in the mammalian brain stem. A second respiratory oscillator has been identified within the VRC, in the RTN/pFRG (Figure 2). The functional role of this more rostral oscillator has been a subject of debate. However, the available data suggest that it may be responsible for generation of the expiratory rhythm, in addition to the inspiratory rhythm generated by the PBC. Brain stem transection caudal to the RTN/pFRG, but rostral to the PBC (i.e., removal of the RTN/pFRG input; see Figure 2), does not affect inspiratory rhythm, but completely and irreversibly eliminates active expiration. Conversely, partial blockade of the PBC inspiratory oscillator activity by opiates disturbs inspiratory rhythm but does not affect expiratory activity, revealing the input and importance of the expiratory oscillator located in the RTN/pFRG. It was proposed that in mammals under normal conditions the activities of these different respiratory oscillators are tightly coupled but are dominated by the inspiratory rhythm generated by the PBC. It was also suggested that in other vertebrate species in which breathing is dominated by expiration (amphibians and reptiles have a breathing pattern of active expiration and passive inspiration), the RTN/pFRG oscillator may play a dominant role in neurogenesis of the respiratory activity. However, this hypothesis has not been tested experimentally.

The role of the pons in generation of eupnea has been a subject of debate for many years. A role of the pontile mechanisms in rhythmogenesis of normal respiratory activity is supported by the evidence obtained in anesthetized experimental animals showing that interruption of the vagal afferent input (vagotomy) and complete removal of the pons acutely eliminate eupnea and evoke gasping. This was observed in many mammalian species examined (but not in rodents) and suggested that pontile mechanisms may play an important role in the neurogenesis of eupnea. However, in unanesthetized animals, eupneic respiratory activity continues following transections at midpontine level or caudally through the facial (VIIth) nucleus. Similarly, studies in vagotomized and anesthetized juvenile rats have demonstrated that complete transection of the brain stem just rostral to the VIIth nucleus (i.e., complete removal of the pons; see Figure 2) does not significantly affect respiratory activity. Furthermore, as

discussed previously, in the absence of pontile input, medullary neuronal circuitry appears to be capable of generating both gasping and eupneic inspiratory activity.

The basic respiratory rhythm generated by the respiratory oscillators is subsequently shaped, modified, and transmitted to bulbospinal inspiratory (located within the rostral ventral respiratory group) and expiratory (within the caudal ventral respiratory group) premotor neurons, which in turn relay the resultant respiratory pattern to spinal motor neurons innervating inspiratory and expiratory muscles, respectively.

Motor Output

The diaphragm (major inspiratory muscle in mammals) and intercostal muscles are innervated by the motor neurons located in the midcervical and thoracolumbar segments of the spinal cord. Inspiration occurs when phrenic motor neurons innervating the diaphragm and thoracic motor neurons innervating the external intercostal muscles are activated by descending command from the VRC. Expiratory motor neurons, which innervate internal intercostal muscles, are responsible for active expiration. There are mono- and polysynaptic pathways, which transmit respiratory pattern from the brain stem respiratory network to the spinal respiratory motor neurons.

Adequate ventilation also depends on airway patency, which is regulated by the 'resistance muscles' controlled by several cranial nerves. Striated muscles of the pharynx and larynx as well as smooth muscles of the trachea and bronchi all modulate airway resistance. The glottis works as a valve and plays an important role in regulating upper airway patency in eupnea. Glottal abductor and adductor muscles, which contract during inspiration and postinspiration, respectively, are innervated by the recurrent and superior laryngeal nerves. Different laryngeal motor neurones which discharge during inspiratory and postinspiratory phases of the respiratory cycle have been identified in the medulla and have been found to receive synaptic input from the respiratory network. Contraction of the glottal adductor muscles in postinspiration slows outflow of the air from the lungs, thus increasing the time of gas exchange and preventing lung collapse by maintaining its functional residual capacity. Laryngeal activity is also important for other related activities such as vocalization and swallowing.

The smooth muscles of the airways are controlled by parasympathetic (vagal and glossopharyngeal) and nonadrenergic, noncholinergic innervations. Vagal preganglionic neurons are located in the dorsal vagal nucleus and nucleus ambiguus of the medulla oblongata. The latter neurons discharge predominantly during the inspiratory phase of the respiratory cycle, contributing to an increase in airway smooth muscle tension during inspiration. Most reflex changes in the airway resistance, as well as maintenance of resting airway tone, are mediated via the excitatory cholinergic parasympathetic smooth muscle innervation.

Respiratory Afferent Inputs

The medullary respiratory network receives modulatory inputs continuously from the other parts of the brain as well as afferent inputs from the periphery (Figure 1) and ensures appropriate ventilation of the lungs in variable environmental and physiological conditions. Peripheral arterial chemoreceptors (sensitive to changes in PaO_2 and $PaCO_2$/pH), brain stem CO_2/pH chemoreceptors, and mechanoreceptors located in the airways control the activity of the brain stem respiratory network. Respiratory activity is also influenced by the activity of cardiovascular and other afferents; however, the inputs arising from chemo- and mechanosensors are of primary importance.

Chemosensory inputs Experiments conducted in animals in which the carotid and aortic bodies have been denervated have shown different functional roles for the peripheral and central chemoreceptors. Hypoxia fails to stimulate ventilation centrally, suggesting that carotid and aortic bodies are the primary oxygen-sensitive sites. In contrast, although peripheral chemoreceptors are sensitive to changes in PCO_2 and pH, the ventilatory response to CO_2 is largely preserved in animals after denervation of the peripheral chemoreceptors.

In adult mammals the specialized chemosensitive neurosecretory glomus cells of the carotid body (located strategically at the place where the common carotid artery bifurcates) are the primary O_2 sensors. Glomus cells monitor changes in PO_2 and, when stimulated by hypoxia, release excitatory mediator(s) to activate afferent fibers of the carotid sinus nerve. Information about arterial PO_2 relays to the brain via the glossopharyngeal nerve and is processed by second-order neurons in the NTS. Under basal conditions (PaO_2 ~75–100 mmHg) the drive from the peripheral chemoreceptors is low. When PaO_2 decreases below 60 mmHg, carotid chemoreceptor activity and ventilation increase exponentially. This ensures adequate supply of O_2 to the brain, which is highly sensitive to hypoxia.

CO_2 provides the major tonic drive for ventilation. Even a small increase in $PaCO_2$ results in a profound increase in respiratory activity to restore the normal

level of CO_2 in the arterial blood (~40 mmHg). Up to 80% of the CO_2-evoked respiratory response is mediated by the action of CO_2 at the chemoreceptors located within the lower brain stem. The primary central CO_2 chemosensors are believed to be located on the ventral surface of the medulla oblongata (Figure 2), although other regions of the brain stem, such as the PBC, the medullary raphe nuclei, the NTS, and the locus coeruleus, have been reported to contain functional respiratory chemoreceptors. Ventral surface CO_2 chemosensors are located just beneath and in a close proximity to the VRC. There is evidence suggesting that upon stimulation these chemosensors release the purine nucleotide adenosine triphosphate (ATP), which may act on the VRC respiratory neurons to evoke appropriate changes in breathing. There is also evidence indicating that respiratory neurons within the VRC could be intrinsically sensitive to CO_2.

Mechanosensory inputs Mechanoreceptors are found throughout the entire respiratory tract. They supply the brain stem respiratory network with information about the mechanical status of the lungs and chest, contribute to the control of respiratory activity, and initiate protective reflexes such as cough. Mechanoreceptors in the lungs and airways have both myelinated and unmyelinated nerve fibers. These fibers travel to the brain mainly within glossopharyngeal and vagus nerves and terminate in the NTS. Slowly adapting pulmonary stretch receptors are activated when the lungs inflate and play a critical role in termination of inspiration and prolongation of expiration (Breuer–Hering reflex). These receptors are located in the airway smooth muscles and have large myelinated fibers. The information is transmitted to the pontine nuclei (pneumotaxic center) and VRC via the NTS second-order relay neurons located within the dorsal respiratory group. Other mechanoreceptors – rapidly adapting pulmonary stretch receptors and bronchopulmonary C-fiber receptors – initiate defensive respiratory reflexes in response to inhaled irritants or rapid lung inflations or deflations. Lung congestion and edema are considered to be the main stimuli for pulmonary C-fiber afferents.

Other inputs Inputs arising from the limbic system influence the activity of the brain stem respiratory network and evoke modifications of breathing during stress, affective behavior, etc. Activation of proprioceptors in muscles, tendons, and joints stimulates breathing and contributes to the increases in ventilation during locomotion and exercise. There are also pathways originating from the neocortex (some of them going directly to the spinal motor neurons innervating the respiratory muscles) which allow voluntary control of breathing.

Plasticity

Breathing is continually adjusted to ensure appropriate ventilation of the lungs in changeable environmental and physiological conditions. Respiratory motor output exhibits significant plasticity, both short-term, when changes in the activity outlast the duration of the stimulus (e.g., hypoxic episode) by seconds and minutes, and long-term, when these changes persist for several hours. Respiratory plasticity can be elicited by hypoxia, hypercapnia, exercise, deafferentation, neural injury, and some other stimuli. Several neurotransmitters, such as serotonin, dopamine, and nitric oxide, have been proposed to be responsible for the respiratory neuroplasticity. For example, serotonin acting within the brain stem as well as at the spinal level (phrenic nucleus) appears to play a central role in the long-term facilitation of breathing triggered by recurrent hypoxic episodes.

Interactions between the Respiratory and Cardiovascular Systems

Activities of the cardiovascular and respiratory systems are intimately linked (Figure 4). Respiratory and cardiovascular rhythms are controlled in relation to one another, ensuring adequate ventilation–perfusion matching within the lungs, thus maintaining optimal respiratory gas exchange. Respiratory sinus arrhythmia – increases in heart rate during inspiration due to a rhythmic partial withdrawal of the inhibitory vagal tone – represents a classical example of cardiorespiratory integration. In variable physiological conditions appropriate changes always occur in both systems. Stimulation of the 'respiratory afferents' has a profound effect on the activity of the cardiovascular system. Similarly, stimulation of the 'cardiovascular afferents' modifies central respiratory drive (for example, increases in blood pressure reduce respiratory activity by stimulation of the arterial baroreceptors). Rhythmic respiratory modulation of vagal parasympathetic and sympathetic activities can be recorded from the appropriate efferent nerves. It is probably not a coincidence that anatomically both cardiac vagal preganglionic neurons and presympathetic neurons are located in the ventrolateral medulla at sites in a close proximity to the brain stem respiratory network. The evidence of central cardiorespiratory integration can be seen at the level of individual neurons. For example, cardiac vagal preganglionic neurons of the nucleus

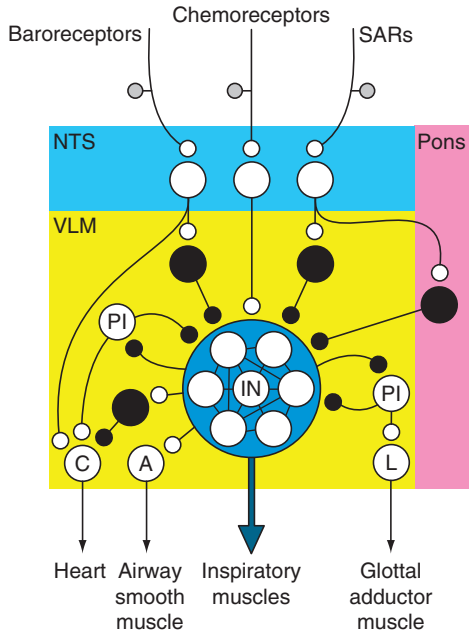

Figure 4 Diagramatic representation of central nervous mechanisms of cardiorespiratory interactions, showing some hypothesized central pathways of the pontomedullary cardiorespiratory network controlling inspiratory activity and motor output in selected cranial nerves. Connections shown may involve mono- or polysynaptic pathways. Open circles (o) depict excitatory connections. Filled circles (•) show inhibitory connections. A, airway vagal preganglionic neurons; C, cardiac vagal preganglionic neurons; IN, neuronal circuitry generating inspiratory activity; L, laryngeal motoneurones ('postinspiratory'); NTS, nucleus tractus solitarii; PI, postinspiratory neurons; SARs, slowly adapting pulmonary stretch receptors; VLM, ventrolateral medulla.

ambiguus receive powerful inhibitory inputs during inspiration as well as excitatory inputs during postinspiration, and these rhythmic changes in their activity underlie respiratory sinus arrhythmia.

Central Nervous Control of Breathing in Some Common Physiological States

Sleep

Sleep–wake cycles have profound influence on the activity of the brain stem respiratory network. Respiratory sensitivities to arterial levels of PO_2 and PCO_2 are reduced during sleep, and the apneic threshold is raised. There is a mild decrease in ventilation, resulting in an increase in arterial level of PCO_2 by around 2–5 mmHg. Withdrawal of a wakefulness-associated excitatory drive to the VRC is believed to be responsible for this decrease in respiratory activity during sleep. Therefore, breathing during sleep is often disordered and frequently interrupted by periods of apnea even in healthy individuals. The incidence of central sleep apneas increases significantly with age. Sleep is also associated with increases in upper airway resistance, which in severe cases can lead to complete upper airway obstruction and profound hypoxia (see later). Because sleep is an oscillatory state, the resultant changes in respiratory activity may be recurrent and/or amplified with repeated episodes.

Exercise and Locomotion

An adequate increase in lung ventilation is crucial to support the metabolic demands of the exercise. Respiratory activity increases even before the onset of the exercise, independently of inputs from the respiratory chemoreceptors or muscle proprioceptors. Hypothalamic structures are believed to be the source of this 'central anticipatory command' to the brain stem respiratory center. The major inputs into the hypothalamus originate from the cortical motor areas associated with locomotion. During exercise (and locomotion in general), respiratory and locomotor patterns are coupled. Respiratory activity is entrained by the somatic afferent input which ascends to the VRC via the lateral parabrachial nucleus of the pons and causes resetting of the respiratory rhythm.

Temperature Regulation

Low and high ambient temperatures are known to have a profound effect on ventilation. During cold exposure, as well as during fever, heat production increases and this increase in metabolism is accompanied by appropriate increases in ventilation. During a febrile response, mediators of inflammation (e.g., cytokines) may also increase respiratory activity by their action on the peripheral chemoreceptor sites and/or within the brain stem respiratory network. At high ambient temperatures many animal species increase the frequency of the respiratory movements to facilitate evaporative heat loss from the respiratory tract (polypnea and panting). These adaptive reflex modifications in breathing can be evoked from the hypothalamic thermoregulatory center following activation of the peripheral (skin) thermoreceptors, even in the absence of changes in core body temperature.

Emotions and Stress

Emotions and stress result in a significant alterations in breathing patterns. Inputs from the limbic system to the brain stem regions concerned with central respiratory control evoke changes in breathing activity during stress, affective behavior, etc.

Hypoxia

The initial increase in respiratory activity in response to hypoxia is evoked by activation of the peripheral

chemoreceptors. If hypoxia is sustained, this increase in breathing is followed by a decline in central respiratory output. Structures located in the hypothalamus, mesencephalon, and the brain stem have been shown to be involved in this secondary respiratory decline. In neonatal experimental animals, brain stem transection or lesions in the lateral upper pons help to maintain respiratory activity that is normally abolished during sustained hypoxia. Several brain sites within the thalamus, hypothalamus, pons, and the medulla have been proposed to function as central O_2 sensors. Their physiological role, however, remains largely unclear.

Summary

Breathing is controlled by a network of neurons located within the lower brain stem. In adult mammals the basic rhythm of respiratory activity is generated by a network of neurons within the so-called PBC complex of the ventral medulla oblongata and is then subsequently shaped, modified, and transmitted to bulbospinal premotor neurons which relay the resultant respiratory pattern to spinal motor neurons innervating respiratory muscles. The activity of the respiratory network is influenced by continuous modulatory inputs from the other parts of the brain and afferent inputs from the periphery. Peripheral and central chemoreceptors (which supply information about arterial pH and levels of O_2 and CO_2), mechanoreceptors located in the respiratory tract (which provide information about the mechanical status of the lungs and chest), proprioceptors in muscles, tendons, and joints (which are responsible for coupling of respiratory and locomotor patterns), as well as from the higher centers of the central nervous system, such as the limbic system, are all essential for adaptive changes (both short- and long-term) in the respiratory motor output to ensure appropriate ventilation of the lungs in variable environmental and physiological conditions.

See also: Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Respiratory Control; Autonomic Nervous System: Neuroanatomy; Autonomic Nervous System: Central Cardiovascular Control; Brainstem Respiratory Circuits; Cardiovascular Function: Central Nervous System Control; Respiration.

Further Reading

Daly M and de Burgh B (1997) Peripheral arterial chemoreception and respiratory-cardiovascular integration. *Monograph for the Physiological Society.* Oxford: Oxford University Press.

Duffin J (2004) Functional organization of respiratory neurons: A brief review of current questions and speculations. *Experimental Physiology* 89: 517–529.

Feldman JL and Del Negro CA (2006) Looking for inspiration: New perspectives on respiratory rhythm. *Nature Reviews Neuroscience* 7: 232–242.

Feldman JL and McCrimmon DR (1999) Neural control of breathing. In: Zigmond MJ, Bloom FE, Landis SC, et al. (eds.) *Fundamental Neuroscience*, pp. 1063–1090. San Diego: Academic Press.

Feldman JL, Mitchell GS, and Nattie EE (2003) Breathing: Rhythmicity, plasticity, chemosensitivity. *Annual Review of Neuroscience* 26: 239–266.

Jordan D (2001) Central nervous pathways and control of the airways. *Respiration Physiology* 125: 67–81.

Lieske SP, Thoby-Brisson M, Telgkamp P, et al. (2000) Reconfiguration of the neural network controlling multiple breathing patterns: Eupnea, sighs and gasps. *Nature Neuroscience* 3: 600–607.

Putnam RW, Filosa JA, and Ritucci NA (2004) Cellular mechanisms involved in CO_2 and acid signaling in chemosensitive neurons. *American Journal of Physiology – Cell Physiology* 287: C1493–C1526.

Ramirez JM and Viemari JC (2005) Determinants of inspiratory activity. *Respiratory Physiology & Neurobiology* 147: 145–157.

Richter DW and Spyer KM (2001) Studying rhythmogenesis of breathing: Comparison of *in vivo* and *in vitro* models. *Trends in Neurosciences* 24: 464–472.

Rybak IA, Shevtsova NA, Paton JF, et al. (2004) Modeling the ponto-medullary respiratory network. *Respiratory Physiology & Neurobiology* 143: 307–319.

Smith JC, Butera RJ, Koshiya N, et al. (2000) Respiratory rhythm generation in neonatal and adult mammals: The hybrid pacemaker-network model. *Respiration Physiology* 122: 131–147.

Speck DF, Dekin MS, Revellette WR, et al. (eds.) (1993) *Respiratory Control – Central and Peripheral Mechanisms*. Lexington: The University Press of Kentucky.

St-John WM and Paton JF (2004) Role of pontile mechanisms in the neurogenesis of eupnea. *Respiratory Physiology & Neurobiology* 143: 321–332.

Taylor EW, Jordan D, and Coote JH (1999) Central control of the cardiovascular and respiratory systems and their interactions in vertebrates. *Physiological Reviews* 79: 855–916.

Autonomic Nervous System: Central Thermoregulatory Control

C J Gordon, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

The Thermoregulatory System

Temperature regulation in mammals and birds has evolved with autonomic and behavioral motor responses, termed thermoeffectors, to defend the core body temperature against changes – heat gain from and heat loss to the environment, as well as heat production from exercise and fever. The body's core temperature (i.e., in brain, heart, lungs, viscera, etc.) is usually maintained within narrow limits and is distinct from the thermal shell, which represents the skin and mucosal surfaces of the body that engage in heat exchange with the environment. The thermal shell also includes those tissues under the surfaces whose temperature may deviate from the core due to heat exchange with the environment.

Behavioral versus Autonomic Mechanisms

Thermoregulation is indeed a unique homeostatic system because it relies on higher level central nervous system (CNS) processes for the conscious sensation and elicitation of corrective motor responses. There is little conscious awareness of most other homeostatic processes, such as those involved in the regulation of blood pressure, respiration, blood pH, and other systems, whereas we are almost always aware and keenly sensitive to changes in our external thermal environment. We continuously strive to maintain a comfortable thermal environment, primarily by using behavioral thermoregulation. Behavioral thermoregulatory mechanisms in humans, such as adding or removing clothing, adjusting a thermostat, or setting a forced convective environment with a fan, all involve high-level, cerebral processing. Other species of mammals and birds will also preferentially use behavioral rather than autonomic processes to thermoregulate. The regulation of body temperature via behavioral mechanisms involves little metabolic energy and is thus preferred over that of autonomic mechanisms. Most species of the reptile, amphibian, and fish phyla depend entirely on behavioral mechanisms to thermoregulate.

Behavior alone is insufficient to maintain a regulated core temperature under the wide range of thermal environments that may be inhabited by birds and mammals. Behavioral thermoregulation can be suppressed during sleep, sickness, injury, and other conditions where behavior cannot be used to provide an optimum thermal environment. With the evolution of a CNS-mediated control of autonomic thermoeffectors, birds and mammals have developed a tightly controlled core temperature that is maintained day and night from soon after birth until the impending point of death. The thermal milieu is largely responsible for the expansion of mammal and bird fauna into environments that are uninhabitable for most other species.

The Autonomic Thermoregulatory System

The autonomic aspect of the temperature regulatory system is generally defined as the involuntary responses to heat and cold stress which modify the rates of heat production and heat loss. The motor responses, termed thermoeffectors, include sweating, panting (thermal tachypnea), or saliva spreading to increase evaporative heat loss; shivering, and nonshivering thermogenesis to increase heat production; adjustments in circulation of blood to the skin and other sites for heat exchange; and involuntary bristling of the fur (piloerection) to curtail heat loss. From this definition, one should note that autonomic thermoeffectors are not limited to responses controlled solely by sympathetic and parasympathetic efferents. Autonomic systems to defend the body core against heat and cold stress are recruited when behavioral options are unavailable or ineffective for maintaining thermal homeostasis. The autonomic systems for thermoregulation use organ systems that have evolved for other purposes: the cardiovascular system for controlling skin blood flow and heat loss, the respiratory system for evaporative heat loss in panting species (dog, cattle, sheep), and salivary glands of the digestive system in rodents that spread saliva on the fur for evaporative water loss. Brown adipose tissue (BAT) appears to be the only organ that has evolved with the main purpose of thermoregulation.

Basic Neural Pathways

Our understanding of the neural networks involved in the autonomic control of thermoeffectors has seen remarkable progress in the past decade. The cutaneous vasculature of the rat tail and rabbit ear and the sympathetic activation of BAT are key experimental models to elucidate the final motor pathways for the control of heat loss and heat production. Decades of research using microelectrodes to record the activity of thermoregulatory neurons *in vivo* and *in vitro* have

led to a thorough understanding of the sensory and integrative aspects of the thermoregulatory system in the peripheral and CNSs. Microinjection of neurotransmitters, modulators, and hormones into specific locations of the CNS has provided data that tie together the neurophysiological and neurochemical mechanisms of thermoregulatory control.

A general model for thermal sensation, integration, and elicitation of thermoeffector responses can be applied to most mammals (Figure 1). Sensory inputs from the temperature of the skin, spinal cord, and other locations have been shown to impinge on thermoregulatory neurons in the preoptic area and anterior hypothalamus (POAH) that serve as sensors of core temperature and show exaggerated changes in firing rate with either an increase or decrease in POAH temperature. A relatively large percentage of neurons in the POAH are warm-sensitive. It is thought that cold-sensitive neurons increase their firing rate with POAH cooling because they are inhibited by synaptic connections from warm-sensitive neurons. Many neurons in the POAH and other select areas of the CNS (spinal cord, and/or other locations in the core) are sensitive to changes in skin temperature. There are also populations of neurons that are insensitive to changes in local temperature but respond vigorously to elevations or reductions in skin temperature. All together, the integration of the afferent thermal sensory processes with core temperature in a network of warm-sensitive, cold-sensitive, and thermal-insensitive neurons leads to the generation of effector signals that control the autonomic effectors (see later). These same integrated signals within the POAH are also thought to drive behavioral thermoregulatory processes through activation of higher level cortical pathways. Thermoregulatory neurons in the POAH are also influenced by nonthermal factors that are associated with or induce changes in the core temperature. For example, circulating reproductive hormones that control menstrual- and estrous-related

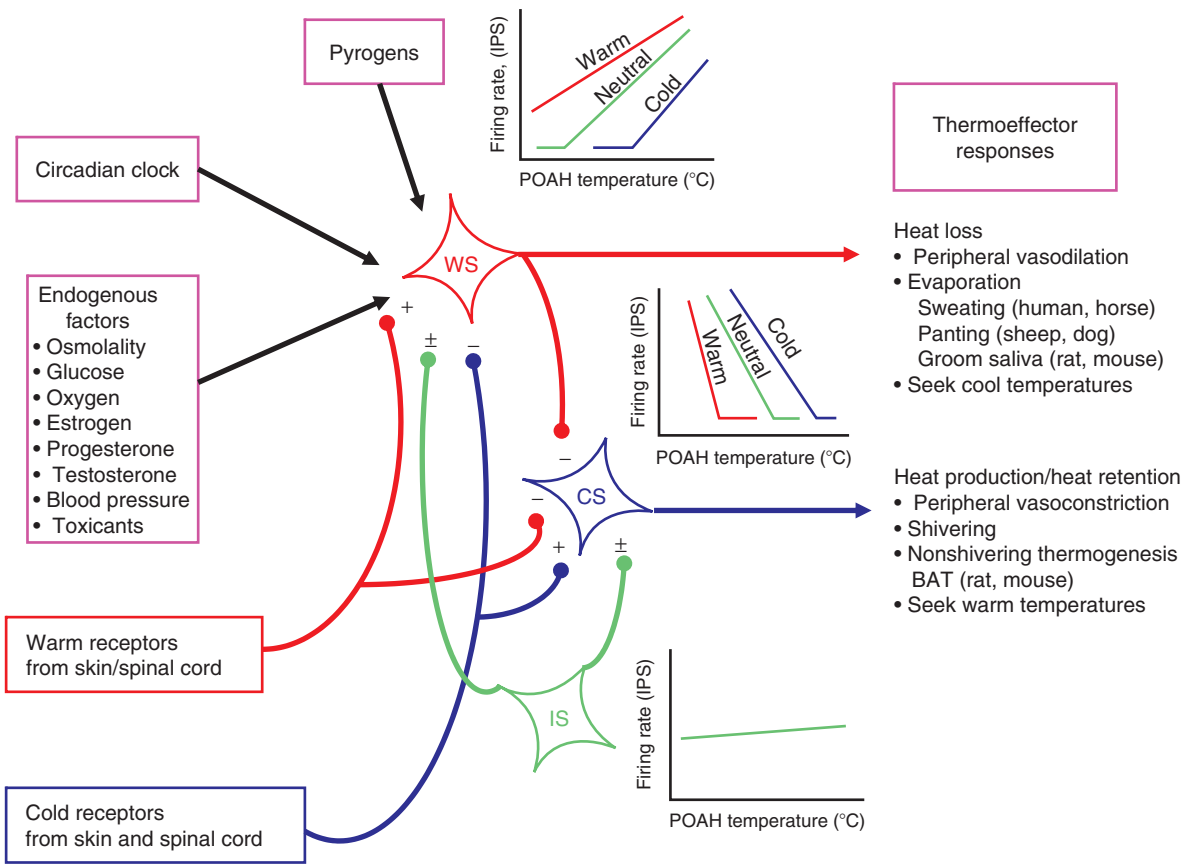

Figure 1 A neural model showing synaptic connections of warm-sensitive (WS), cold-sensitive (CS), and thermally insensitive neurons in the preoptic area and anterior hypothalamus (POAH). Activity of WS and CS neurons is modulated by ambient temperature, resulting in effector signals to modulate the activity of autonomic and behavioral motor outputs of the thermoregulatory system. For example, a WS neuron has higher activity when exposed to warm ambient temperature combined with elevated POAH temperature. IPS, Impulses/sec; + indicates stimulation; − indicates inhibition; ± indicates stimulation or inhibition.

rhythms of the core temperature have been shown to selectively alter activity of warm- and cold-sensitive neurons in the POAH. Glucose, oxygen, blood pressure, and tissue osmolarity have also been found to alter activity of some thermoregulatory neurons. The circadian pacemaker located in the suprachiasmatic nucleus also modulates activity of warm- and cold-sensitive neurons, leading to a 24 h regulated rhythm in core temperature. Environmental toxicants and other insults can affect the activity of the thermoregulatory centers.

The integration of the peripheral and central thermal sensory capacity of the thermoregulatory system was elucidated years ago with the application of stereotaxically implanted thermodes to heat or cool the POAH or other areas of the CNS. For example, heating of the POAH stimulates a heat loss response characterized by reduction in shivering (if the animal is exposed to cold), peripheral vasodilation, and sweating or panting. The animal behaves as if it was hot and seeks cooler temperatures. The thermoregulatory response to POAH heating can reach a magnitude such that heat loss exceeds heat production and the animal becomes hypothermic. Cooling the POAH area elicits a thermogenic response, cutaneous vasoconstriction to restrict heat loss, and the animal seeks warmer ambient temperatures. Studies employing POAH thermodes and manipulations of ambient temperature contributed to the concept of reciprocal inhibition of warm and cold neural pathways in the POAH.

Integrative neurons in the CNS are generally 5–10 times more sensitive to a change in local temperature than to changes in skin temperature. That is, a 1 °C reduction in brain temperature has about the same effect on a thermoeffector response as would a 5–10 °C reduction in skin temperature. When maintained within ambient temperatures where core temperature is controlled (i.e., zone of normothermia), autonomic and behavioral thermoeffectors are controlled primarily by changes in the shell or skin temperature. However, thermal sensors in the CNS and other core sites are called upon in the response to the hyperthermia from exercise and fever as well as when thermoregulation begins to fail during excessive cold or heat stress. The thermal sensory properties of neurons in the CNS of mammals and birds may be considered a vestige of evolution whereby the 'lower vertebrates' rely primarily on CNS thermal sensors to elicit thermoeffector responses.

Integration in the CNS

Neurons in the POAH integrate thermal information as well as nonthermal stimuli that can have a bearing on thermoregulation. Blood pH, blood pressure, oxygen level, osmotic tonicity, and glucose levels can all influence the activity of POAH neurons in a manner that would be expected based on their thermoregulatory effects. For example, dehydration has been shown to increase the threshold core temperature for evaporative water loss in mammals exposed to heat stress. This is an adaptive mechanism that conserves precious body water under conditions of prolonged heat exposure but also results in a higher regulated core temperature. Warm-sensitive neurons in the POAH respond to osmotic stress *in vitro* with a reduction in thermal sensitivity, a response that essentially raises the set point or threshold temperature to activate sweating. Overall, one can visualize a network of warm-sensitive, cold-sensitive, and thermally insensitive neurons in the POAH as well as in other locations in the CNS; all interact with nonthermal stimuli to generate effector signals that drive a series of thermoeffectors to control autonomic thermoregulation.

Thermal-sensitive neurons in the POAH are integrated with the control of blood pressure. Baroreceptor neurons that monitor blood pressure and blood volume have projections to thermal-sensitive neurons. The capacity of thermoregulatory neurons to integrate pressure/volume of the circulatory system is adaptive in times of cardiovascular crisis. For example, a reduction in blood pressure has been shown to preferentially stimulate warm-sensitive neurons and suppress cold-sensitive neurons in the POAH. This should result in a heat dissipatory response and subsequent reduction in core temperature. Rats subjected to hemorrhagic shock (an acute loss in blood volume) become hypothermic and select cooler ambient temperatures, an indication of a reduction in the set point. The hypothermic response improves the ability to survive the hemorrhagic shock, suggesting that the integration of thermal and baroreceptor inputs in the POAH is an adaptive response.

Set Point

The concept of a thermostat with a set-point temperature as depicted in Figure 2 is an extremely useful analogy for explaining how autonomic and behavioral thermoeffectors are controlled by the CNS. Thermal physiologists define set point as follows: "The value of the regulated variable which a healthy organism tends to stabilize by the processes of regulation. When external or internal interferences tend to alter the regulated variable (i.e., body temperature), the resulting thermoeffector activities counter the alterations." In other words, if an organism uses its thermoeffectors to maintain core temperature at

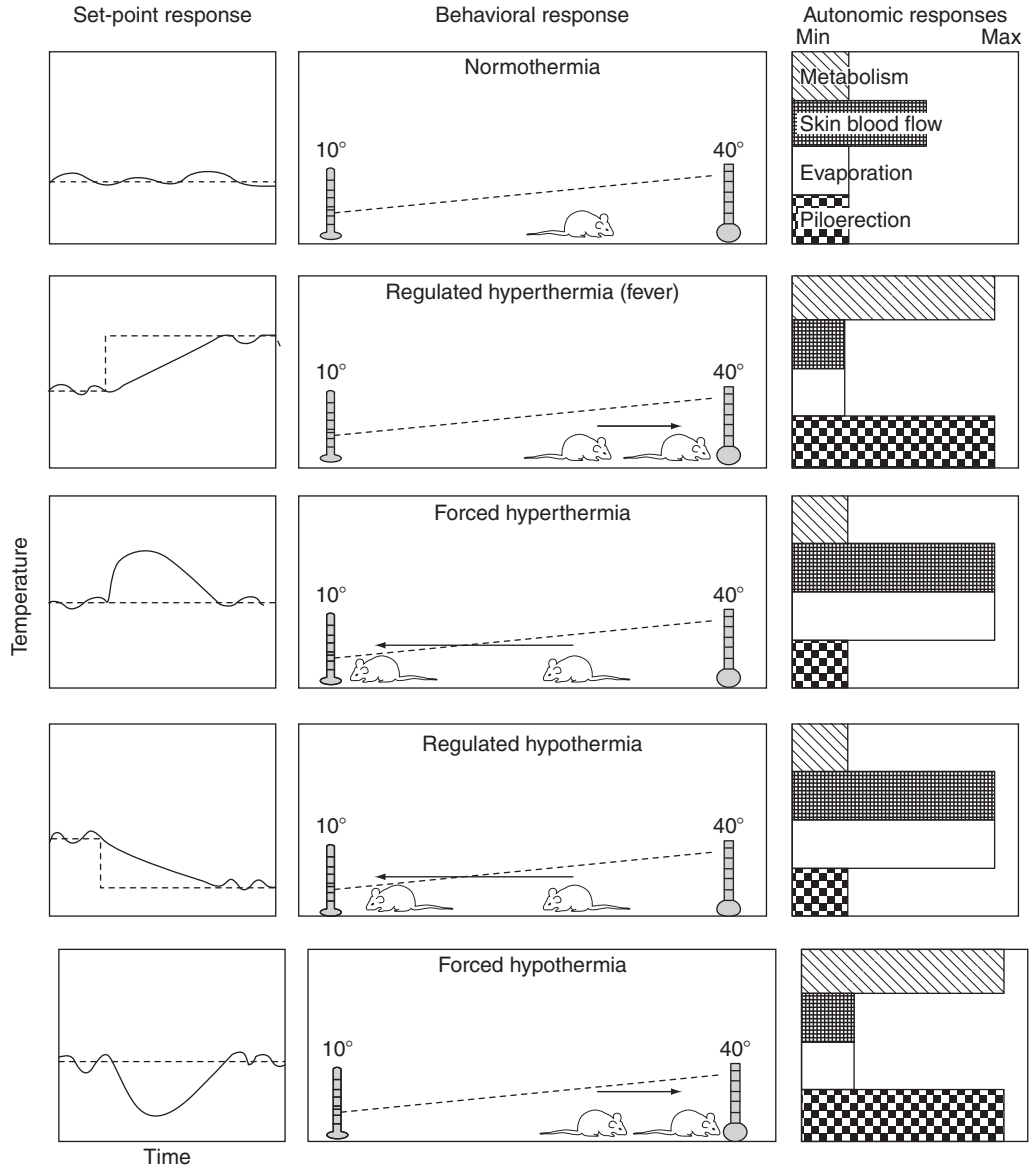

Figure 2 Summary of behavioral and autonomic responses of a homeotherm when subjected to manipulation of body and set-point temperature. Graphs on the left represent the relationship between the set point (dashed line) and core temperature (solid line).

37.5 °C, then it is assumed that its set point or reference temperature is set at 37.5 °C. The set point for temperature regulation may change with certain endogenous and environmental stimuli, such as fever, starvation, and dehydration. The circadian variation in core temperature is considered to be a result of a 24-hr oscillation in the set-point temperature.

The set-point theory allows one to distinguish between stimuli that elicit regulated changes in thermoregulatory control from ones that simply impart deficits in thermoeffector function (Figure 2). The thermoregulatory system of homeotherms and poikilotherms attempts to maintain a core temperature equal to the set-point temperature. This process is continuous in homeotherms but is intermittent in poikilotherms when options are available to behaviorally thermoregulate (e.g., shuttling between sun and shade). In a thermoneutral environment for a healthy homeothermic species, set-point temperature is equal to core temperature and thermoeffectors for heat gain and heat loss are balanced and maintained at a minimal level of activity. The animal has a normothermic body temperature and selects an ambient temperature that is comfortable and associated with minimal energy expenditure. Normothermy means core temperature is controlled within ±1 standard deviation of the range associated with normal, resting, thermoneutral conditions.

Autonomic Thermoeffectors

Brown Adipose Tissue and Nonshivering Thermogenesis

Nonshivering thermogenesis, defined as the heat produced from metabolic processes not involving contracting muscles, has been one of the most intensively studied topics in thermal physiology. A key strategy for mammals subjected to chronic cold exposure is to either reduce heat loss by adding insulation or accelerate oxygen consumption and produce heat, but without muscular contraction. The former mechanism is not a viable option in relatively small mammals that are limited in the amount of insulative fur that can be added from summer to winter. Smaller species of mammals must either develop better thermogenic capacity in the winter, which also requires more foraging for food, or hibernate, to avoid the metabolic costs of thermoregulation in the winter. The heat produced from shivering is effective at counteracting the heat loss from cold exposure, but shivering is metabolically inefficient and uncomfortable. Ideally, nonshivering thermogenesis is the most effective way to adapt to a cold environment. Many tissues, including heart and liver, from cold-acclimated mammals, compared to warm-climate species, have increased aerobic capacity and higher concentrations of enzymes involved in cellular respiration.

BAT is a key thermoeffector for the autonomic control of nonshivering thermogenesis in rodents. There is a renewed interest in understanding the control of BAT thermogenesis because of possible links between the function of this metabolically active tissue and human obesity. BAT function in rodents is intimately related to total energy balance. The neural and hormonal contributions to obesity (e.g., leptin) are now approached in rodent models with consideration to the neural control of BAT thermogenesis. The activity of thermoregulatory neurons in the POAH modulates the activity of BAT thermogenesis through descending excitatory and inhibitor pathways utilizing the neurotransmitters glutamine and γ-aminobutyric acid (GABA), respectively (**Figure 3**). For example, microinjection of bicuculine, a $GABA_A$ receptor antagonist, into the rostral raphe pallidus evokes a marked elevation in sympathetic nerve activity and a prolonged rise in BAT temperature, reflecting an explosive rise in its metabolism and heat production. Currently, the focus is on warm-sensitive POAH neurons that evoke tonic inhibition in the dorsomedial hypothalamus. Releasing this inhibition by exposure to a cold environment, selective cooling of the POAH, or induction of a fever blocks warm-sensitive neuron activity in the POAH, leading to BAT thermogenesis.

BAT is found in other small mammals and also occurs in appreciable amounts in the newborn of large mammals, including cows, goats, and humans. Because the ability to shiver is not developed until about 10 days of age, BAT is the main source for heat production in newborn rats and mice. BAT is also critical to stress-induced hyperthermia and energy balance. In addition to the possible role in obesity discussed previously, the neural control of BAT is becoming a crucial model to study how dysregulation of thermogenesis contributes to hyperthermia and heat stroke. Rodents and other species respond to stressors such as noxious stimuli, placement in a novel environment, and restraint with a marked elevation in body temperature. Such an autonomic response prepares the animal adequately to respond to potential danger and to escape.

Peripheral Vasomotor Tone

The rate of dry heat loss (i.e., nonevaporative) from the skin to the surrounding environment is proportional to the difference between skin and air temperature. The neural control of the distribution of warm blood in the core to the thermal shell represents an effective means of regulating heat loss under a wide range of ambient temperature. The regulation of heat loss through the variation in skin blood flow is most effective in locations that are concentrated with arterial–venous anastomoses (AVAs). AVAs are under sympathetic control and are capable of shunting large amounts of blood to the skin for heat dissipation. For example, the hands and feet of humans are replete with AVAs. Indeed, blood flow to the extremities of humans can vary by 100-fold and mean skin blood flow varies by about tenfold, depending on ambient conditions and body temperature.

The rat and mouse tail and rabbit ear are also key sites for the regulation of dry heat loss. The rat tail and rabbit ear abound in AVAs, have little or no fur, and a relatively high surface area:volume ratio, making them ideal sites for regulating heat loss. Studies of the sympathetic control of peripheral vasomotor tone in rabbit and rat have led to the development of neural models involving neurons of the raphe magnus/pallidus and parapyramidal area (**Figure 4**). Neurons from these regions innervate preganglionic, sympathetic cutaneous vasomotor neurons. Their activation is critical to evoke vasoconstriction when exposed to cold temperatures. These pathways are also essential for eliciting peripheral vasoconstriction when subjected to nonthermal, stressful stimuli such as a startle, noxious fumes, or pain. Interestingly, the thermoregulatory aspects of vasomotor control involve subtle differences in activation from the CNS as compared to other

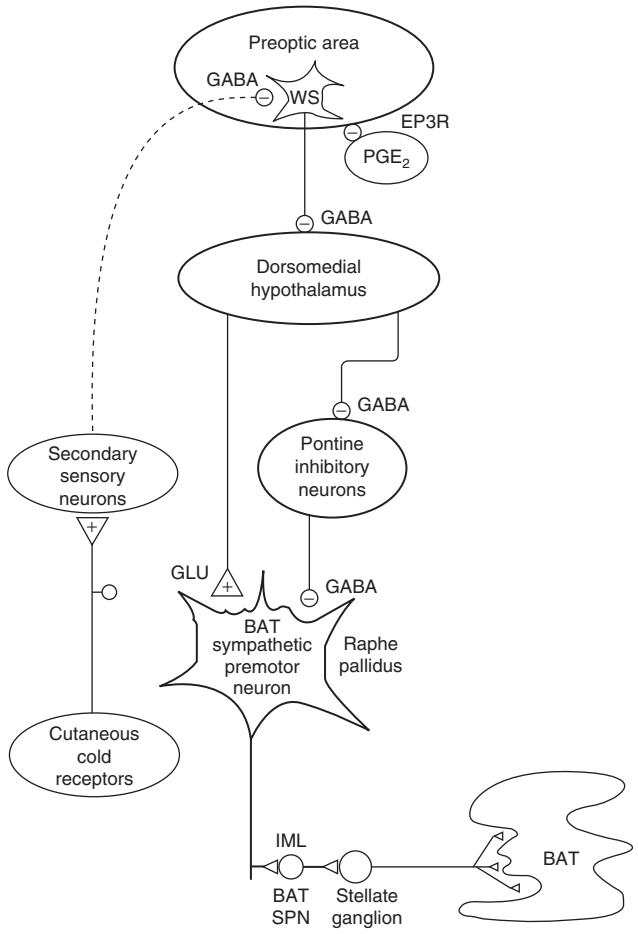

Figure 3 Neural model for the organization of the central pathways that control brown adipose tissue (BAT) thermogenesis. Warm-sensitive (WS) neurons in the POAH are proposed to impart tonic inhibition of the dorsomedial hypothalamus that normally activates BAT thermogenesis through glutamate (GLU) excitatory pathways, which drive BAT sympathetic premotor neurons in the raphe pallidus. γ-Aminobutyric acid (GABA) pathways inhibit BAT thermogenesis via pontine inhibitory pathways (IML, intermediolateral column; SPN, sacral parasympathetic nucleus). Prostaglandin E_2 (PGE_2), a final modulator of fever, inhibits activity of WS neurons, resulting in a marked rise in BAT thermogenesis to raise core temperature (EP3R, prostaglandin E receptor subtype).

nonthermal stimuli. Overall, thermal stimuli, from the core and periphery, are thought to be integrated in the POAH area. The resulting effector signals are further processed in the aforementioned areas of the brain stem and spinal cord to elicit the proper thermoeffector signals to regulate heat loss by modulating skin blood flow.

Evaporative Heat Loss

The effectiveness of the transfer of heat from the body to the surrounding environment by radiation, convection, and conduction decreases with increasing ambient temperature. Thus, evaporative heat loss becomes the key thermoeffector for heat loss if thermal homeostasis is to be maintained in warm environments. Insensible or passive evaporation of water accounts for approximately 20% of the total heat loss in mammals maintained in a comfortable thermal environment. However, when exposed to heat stress, passive water loss is insufficient to maintain thermal homeostasis. Active mechanisms of evaporative water loss include sweating, panting, or saliva spreading, depending on the species. Each thermoeffector achieves a rise in evaporative heat loss, to thermoregulate in the heat, but with unique neural and anatomical mechanisms.

Rodents exposed to heat stress will actively groom saliva on their tail and fur to increase evaporative heat loss, allowing for effective regulation of core temperature. Grooming saliva is in fact a complex thermoeffector involving the autonomic activation of salivary glands combined with higher level participation by cerebral processes to apply the saliva to the skin. The submaxillary salivary gland of the rat is selectively stimulated by heat stress via activation of parasympathetic cholinergic neurons. Mammals that sweat, including humans, secrete copious amounts of sweat on the skin from sweat glands that are controlled by the sympathetic nervous system. Sympathetic nerve

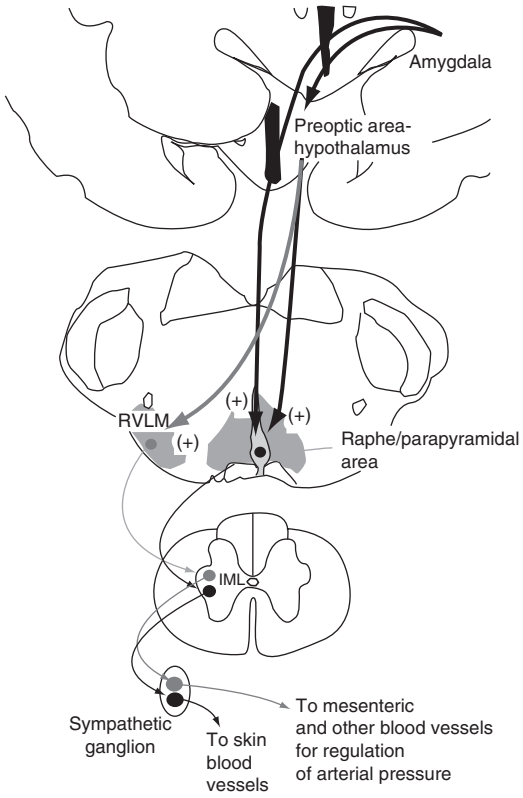

Figure 4 Proposed CNS pathways that mediate control over peripheral vasomotor tone. Thermally sensitive neurons in the POAH modulate activity of the raphe/parapyramidal area and rostral ventrolateral medulla (RVLM) neurons. IML, Intermediolateral column.

fibers innervating human sweat glands are uniquely cholinergic and are under control of thermoregulatory pathways in the POAH (see Figure 1). Panting species such as dog and sheep achieve an increase in evaporative water loss by moving air rapidly across mucosal surfaces to maintain a high rate of evaporative heat loss. The neural control of respiration in the medulla in panting species can be modulated by the thermoregulatory centers in the POAH and possibly other sites of the CNS.

Neuroendocrine Pathways

The hypothalamic–pituitary axis is critical in the autonomic control of thermoregulation. The combined stimulation of cold-sensitive neurons and suppression of warm-sensitive neurons in the POAH elicits effector signals that trigger a cascade of neuroendocrine events to raise heat production. Skin cooling or direct cooling of the POAH triggers the release of thyrotropin-releasing hormone from the hypothalamus, leading to the release of thyroid-stimulating hormone (TSH) from the anterior pituitary gland. TSH stimulates the release of thyroxine (T_4) from the thyroid gland, which is deiodinated to triiodothyronine (T_3) that is bound to cellular and nuclear receptors, stimulating heat production. Corticotropin-releasing factor is also released from the hypothalamus during POAH heating and cooling and stimulates the release of adrenocorticotropic hormone from the anterior pituitary, leading to release of corticosteroid hormones from the adrenal cortex. Corticosteroids are critical for the maintenance of thermal homeostasis during cold stress, heat stress, and fever.

Pathology

Menopausal Hot Flashes

Millions of women worldwide undergoing menopause experience the thermal discomfort of hot flashes. The hot flash is a feeling of thermal discomfort that is accompanied by a widespread thermal dissipation, including sweating and cutaneous vasodilation. Recent application of functional magnetic resonance imaging (fMRI) in symptomatic women suggests that the insular cortex and anterior cingulate play a critical role in the response to hot flashes. These areas of the CNS show a remarkable rise in activity in women as they exhibit hot flashes. Interestingly, the hypothalamus does not emerge in the fMRI as a key region of activation during a hot flash. Hence, the hot flash is an autonomic motor response with obvious conscious sensation elicited through the cortex.

Aging

Aging is associated with generalized dysfunction of the autonomic nervous system characterized by impaired baroreflex function, reduced parasympathetic reflexes, increase in blood pressure, and reduced parasympathetic reflexes. With these deficits, it is not surprising to find impaired autonomic control of thermoregulatory processes with aging. Aging seems to have no effect on baseline body temperature in humans and is associated with a minor attenuation in the amplitude of the circadian temperature rhythm. Aged humans do have reduced capacity to generate heat during cold exposure and diminished ability to reduce skin blood flow, which together imply greater susceptibility to hypothermia during cold exposure. Older adults subjected to heat stress are unable to produce the increase in sweating and rise in skin blood flow as seen in younger adults, which could mean a greater susceptibility to hyperthermia during heat stress. There is increased mortality in the aged from hyperthermia and hypothermia, but this incidence appears to be attributed to nutritional, medication, and disease effects and not to a direct malfunction of thermoregulation.

Fever

The development of and recovery from a fever are mediated through circulating inflammatory cytokines that modulate the CNS control of autonomic and behavioral thermoeffectors, and through immune systems. The fever pathway is initiated by exposure to an exogenous pyrogen, which is defined as any substance that causes a fever when it enters the body. The most potent pyrogens are infectious agents or their components (e.g., bacterial endotoxins such as lipopolysaccharide). These pyrogens activate circulating lymphocytes and/or fixed macrophages (e.g., Kupffer cells) to produce anti-inflammatory and pro-inflammatory cytokines that enter the circulation. Interleukins (IL-1 and IL-6) are key pro-inflammatory cytokines that play a major role in the development of an elevated body temperature. Anti-inflammatory cytokines, such as IL-10 and tumor necrosis factor-α, act as antipyretics (or cryogens) and serve, among many functions, to limit the magnitude of the fever. Circulating cytokines are thought to enter the hypothalamus and preoptic area via fenestrated (i.e., leaky) areas of the organovasculosum lamina terminalis or by active transport mechanisms. Constitutive and inducible cyclooxygenases are activated by IL-1 and IL-6, leading to the conversion of arachidonic acid to prostaglandins and other products. Prostaglandin E_2 (PGE_2) is a final mediator of fever. When released into the POAH, PGE_2 selectively stimulates the activity of cold-sensitive neurons and suppresses the activity of warm-sensitive neurons, leading to an explosive elevation in BAT thermogenesis and marked reduction in skin blood flow. The fever, in conjunction with an increase in white blood cell motility and lymphocyte proliferation, enhances the host's defense to the infectious agent. The high body temperature of the host also suppresses growth of some pathogens. The fever pathway can also be activated by other traumatic insults that cause cellular injury, such as a severe sunburn, physical injury to the CNS, exposure to chemical toxicants, and stroke.

Toxicology

Exposure to pesticides, heavy metals, air pollutants, and many other xenobiotic chemicals can alter thermoregulatory function. Organophosphate-based insecticides, which are similar in structure to nerve gases, inhibit acetylcholinesterase (AChE) activity, leading to a marked reduction in body temperature. Inhibition in AChE is thought to stimulate the warm-sensitive neurons and/or suppress cold-sensitive neurons in the POAH (see Figure 1), leading to a regulated hypothermic response in rodents. The relatively large body mass of adult humans impedes the development of the hypothermic response characteristic of small rodents. Humans and other large species may nonetheless exhibit a brief period of mild hypothermia following intoxication. Both humans and rodents exposed to organophosphates can sustain a prolonged fever following exposure that persists for several days. Air pollutants such as ozone and carbon monoxide also induce marked hypothermia in rats and mice. The hypothermic response has been shown to alleviate the toxic symptoms and prolong survival, suggesting that the autonomic and behavioral thermoregulatory response is an adaptive response.

See also: Autonomic Nervous System: Neuroanatomy; Energy Homeostasis: Adiposity Signals; Energy Homeostasis: Thermoregulation; Sympathetic Nervous System; Thermoregulation During Sleep and Sleep Deprivation; Thermoregulation: Autonomic, Age-Related Changes.

Further Reading

Blessing WW (2003) Lower brainstem pathways regulating sympathetically mediated changes in cutaneous blood flow. *Cellular and Molecular Neurobiology* 23: 527–538.

Brown JW, Whitehurst ME, and Carroll RG (2005) Thermoregulatory set point decreases following hemorrhage in rats. *Shock* 23: 239–242.

Freedman RR, Benton MD, Genik RJ II, et al. (2006) Cortical activation during menopausal hot flashes. *Fertility and Sterility* 85: 674–678.

Gordon CJ (1993) *Temperature Regulation in Laboratory Rodents*. New York: Cambridge University Press.

Gordon CJ (2005) *Temperature and Toxicology: An Integrative, Comparative, and Environmental Approach*. Boca Raton, FL: CRC Press.

Gordon CJ, Mohler FS, Watkinson WP, et al. (1988) Temperature regulation in laboratory mammals following acute toxic insult. *Toxicology* 53: 161–178.

Hori T, Nakashima T, Koga H, et al. (1988) Convergence of thermal, osmotic and cardiovascular signals on preoptic and anterior hypothalamic neurons in the rat. *Brain Research Bulletin* 20: 879–885.

Kenney WL and Munce TA (2003) Aging and human temperature regulation. *Journal of Applied Physiology* 95: 2598–2603.

Morrison SF (2004) Central pathways controlling brown adipose tissue thermogenesis. *News in Physiological Science* 19: 67–74.

Nalivaiko E and Blessing WW (2001) Raphe region mediates changes in cutaneous vascular tone elicited by stimulation of amygdala and hypothalamus in rabbits. *Brain Research* 891: 130–137.

The Commission for Thermal Physiology of the International Union Of Physiological Sciences (2001) Glossary of terms for thermal physiology. *Japanese Journal of Physiology* 51: 245–280.

Autonomic Nervous System: Central Urogenital Control

W C de Groat, University of Pittsburgh Medical School, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Many functions of the urogenital system are controlled by complex neural pathways in the brain and spinal cord. These central pathways in turn regulate the activity of peripheral autonomic (sympathetic and parasympathetic) and somatic nerves that innervate the smooth muscle, striated muscle, epithelial cells, and exocrine glands in the urogenital organs. Some urogenital functions (penile erection) are purely involuntary and are mediated by reflex pathways in the spinal cord or brain stem; whereas others (micturition) are more complex, involving voluntary control by the cerebral cortex. This article provides a brief review of the central neural circuitry and neurotransmitters involved in (1) the control of urine storage and release by the lower urinary tract and (2) the control of male sexual organs.

Innervation of the Urogenital Organs

The lower urinary tract and sex organs are innervated by three sets of peripheral nerves that arise at the level of the lumbosacral spinal cord (Figure 1). Visceral structures (urinary bladder, urethra, cavernous tissue of the penis and clitoris, vas deferens, and exocrine glands) receive an innervation from both divisions of the autonomic nervous system. Parasympathetic efferent axons that originate from preganglionic neurons in the sacral spinal cord pass through the pelvic nerves to synapse on peripheral ganglion cells that, in turn, innervate the organs. Activation of parasympathetic pathways induces a bladder contraction, urethral relaxation, and penile and clitoral erection. Sympathetic efferent axons that originate from preganglionic neurons in the rostral lumbar segments of the spinal cord project to ganglion cells in the prevertebral and paravertebral ganglia, which then send axons through various nerves to the target organs. Activation of sympathetic pathways relaxes the bladder, contracts the bladder neck and urethra, produces seminal emission, and can elicit penile erection or detumescence. Somatic efferent axons originate in motor neurons in the sacral spinal cord and pass through the pudendal nerves to striated muscles of the urethral sphincter and periurethral striated muscles (bulbocavernosus and ischiocavernosus) involved in ejaculation.

Afferent axons innervating the urogenital organs arise in the lumbosacral dorsal root ganglia and are contained in the three sets of peripheral nerves. The most important afferents for initiating micturition are those which travel in the pelvic nerve to the sacral spinal cord. These afferents consist of small myelinated (A-δ) and unmyelinated (C) axons that convey impulses from tension receptors and nociceptors in the bladder wall. Afferent activity arising from mechanoreceptors in the penis is carried in the pudendal nerves and triggers penile erection and emission–ejaculation.

Anatomy of Central Nervous Pathways Controlling the Urogenital System

The reflex circuitry controlling the urogenital organs consists of four basic components: primary afferent neurons, spinal efferent neurons, spinal interneurons, and neurons in the brain that modulate spinal reflex pathways.

Pathways in the Spinal Cord

Afferent projections in the spinal cord Afferent pathways from the bladder project through the pelvic nerves into Lissauer's tract at the apex of the dorsal horn in the caudal lumbosacral spinal cord and then send collaterals laterally and medially around the dorsal horn into laminae V–VII and X at the base of the dorsal horn (Figure 2). The lateral pathway terminates in the region of the sacral parasympathetic nucleus and also sends some axons to the dorsal commissure (Figure 2). Pudendal afferent pathways from the penis and clitoris project into the deeper layers of the medial lumbosacral dorsal horn.

Efferent neurons Parasympathetic preganglionic neurons innervating the urogenital organs are located in the intermediolateral gray matter (laminae V–VII) in the caudal lumbosacral segments of the spinal cord (Figures 3 and 4), whereas sympathetic preganglionic neurons are located in medial (lamina X) and lateral sites (laminae V–VII) in the rostral lumbar spinal cord. Motor neurons innervating the periurethral striated muscles are located in lamina IX in the ventral horn (Figure 4).

Spinal interneurons Interneurons involved in spinal reflex circuits controlling urogenital functions have been identified by retrograde transneuronal transport following injection of pseudorabies virus into the urinary lower urinary tract or genital organs of the rat. Large populations of interneurons are located just dorsal and medial to the preganglionic neurons as well as in the dorsal commissure and lamina I (Figure 2(c)). The spinal neurons involved in processing afferent

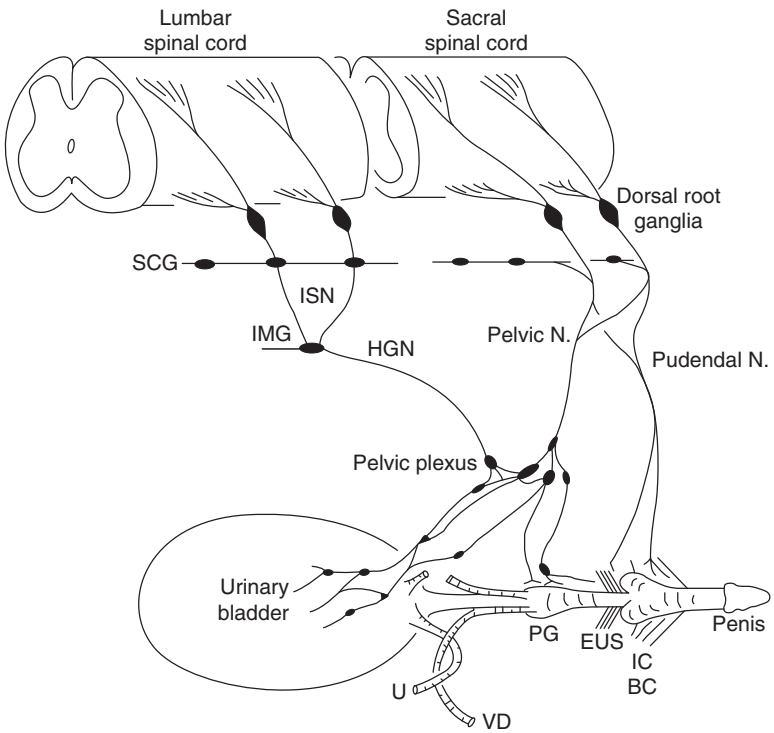

Figure 1 Diagram showing the sympathetic, parasympathetic, and somatic innervation of the urogenital tract of the male cat. Sympathetic preganglionic pathways emerge from the lumbar spinal cord and pass to the sympathetic chain ganglia (SCG) and then via the inferior splanchnic nerves (ISN) to the inferior mesenteric ganglia (IMG). Preganglionic and postganglionic sympathetic axons then travel in the hypogastric nerve (HGN) to the pelvic plexus and the urogenital organs. Parasympathetic preganglionic axons, which originate in the sacral spinal cord, pass in the pelvic nerve to ganglion cells in the pelvic plexus and to distal ganglia in the organs. Sacral somatic pathways are contained in the pudendal nerve, which provides an innervation to the penis, the ischiocavernosus (IC), bulbocavernosus (BC), and external urethral sphincter (EUS) muscles. The pudendal and pelvic nerves also receive postganglionic axons from the caudal sympathetic chain ganglia. These three sets of nerves contain afferent axons from the lumbosacral dorsal root ganglia. U, ureter; PG, prostate gland; VD, vas deferens.

input from the lower urinary tract have been identified by the expression of the immediate-early gene, c-*fos* (Figure 2(b)). In the rat, chemical or mechanical stimulation of the bladder and urethra increases the levels of Fos protein primarily in the dorsal commissure, in the superficial dorsal horn, and in the area of the sacral parasympathetic nucleus (Figure 2(b)). Some of these interneurons make local connections in the spinal cord and participate in segmental spinal reflexes, whereas others send ascending projections to regions in the brain that are involved in the supraspinal control of urogenital function.

Pathways in the Brain

In the rat, transneuronal virus tracing methods have identified many populations of neurons in the brain that are involved in the control of bladder, urethra, and urethral sphincter. Labeled areas include Barrington's nucleus (the pontine micturition center (PMC)); medullary raphe nuclei, which contain serotonergic neurons; the locus coeruleus, which contains noradrenergic neurons; periaqueductal gray (PAG); and the A5 noradrenergic cell group (Figure 3). Several regions in the hypothalamus (including the paraventricular nucleus and medial preoptic nucleus) and the cerebral cortex also exhibit virus-infected cells. Neurons in the cortex are located primarily in the medial frontal cortex. Similar areas in the brain have been identified after injection of pseudorabies virus into the penis or prostate gland.

Other anatomical studies in which anterograde tracer substances were injected into brain areas and then identified in terminals in the spinal cord are consistent with the virus tracing data (Figure 4). Tracer injected into the paraventricular nucleus of the hypothalamus labeled terminals in the sacral parasympathetic nucleus as well as the sphincter motor nucleus. On the other hand, neurons in the anterior hypothalamus project to the PMC. Neurons in the PMC in turn project primarily to the sacral parasympathetic nucleus, the lateral edge of the dorsal horn, and the dorsal commissure. Conversely, projections from neurons in the lateral pons, an area implicated in the control of the urethral sphincter, terminate rather selectively in

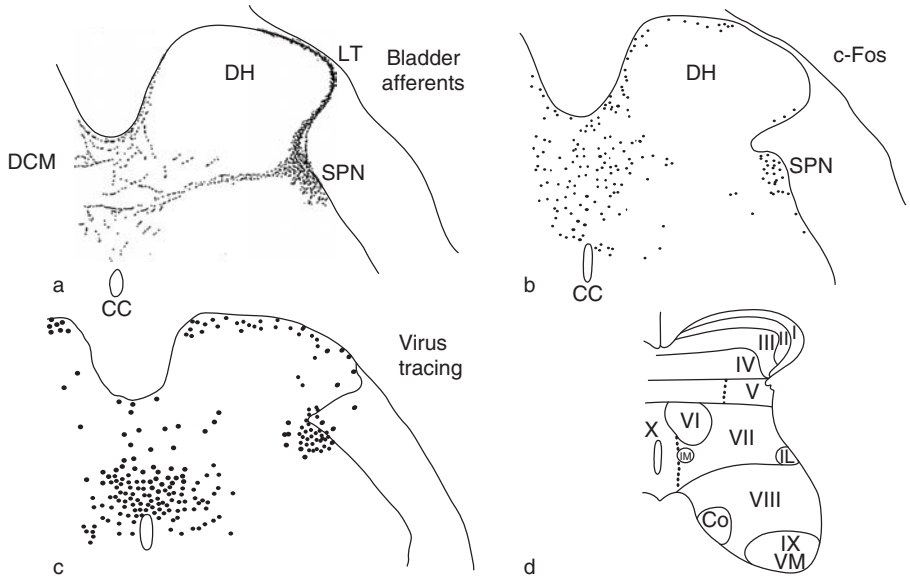

Figure 2 Comparison of the distribution of rat bladder afferent projections to the L6 spinal cord (a) with the distribution of rat c-Fos-positive cells in the L6 spinal segment following chemical irritation of the lower urinary tract (b) and the distribution of interneurons in the L6 spinal cord labeled by transneuronal transport of pseudorabies virus injected into the urinary bladder (c). Afferents are labeled by wheat germ agglutinin-conjugated horseradish peroxidase injected into the urinary bladder. c-Fos immunoreactivity is present in the nuclei of cells. (d) The laminar organization of the cat spinal cord. DCM, dorsal commissure; DH, dorsal horn; LT, Lissauer's tract; SPN, sacral parasympathetic nucleus; CC, central canal.

the sphincter motor nucleus, also known as Onuf's nucleus (Figure 4).

Central Neural Control of the Lower Urinary Tract

The central pathways controlling lower urinary tract function are organized as simple on–off switching circuits that maintain a reciprocal relationship between the urinary bladder and urethral outlet (Figure 5). The principal reflex components of these switching circuits are listed in Table 1 and illustrated in Figure 6. Intravesical pressure measurements during bladder filling in both humans and animals reveal low and relatively constant bladder pressures when bladder volume is below the threshold for inducing voiding (Figure 5). The accommodation of the bladder to increasing volumes of urine is primarily a passive phenomenon dependent upon the intrinsic properties of the vesical smooth muscle and quiescence of the parasympathetic efferent pathway. In addition, in some species urine storage is also facilitated by sympathetic reflexes that mediate an inhibition of bladder activity as well as closure of the bladder neck and proximal urethra (Table 1, Figure 6). During bladder filling the activity of the sphincter electromyogram also increases (Figure 5), reflecting an increase in efferent firing in the pudendal nerve and an increase in outlet resistance which contributes to the maintenance of urinary continence. Both sympathetic and pudendal nerve reflex activity are activated by low-level bladder afferent activity induced by bladder distension.

The storage phase of the urinary bladder can be switched to the voiding phase either involuntarily (Figure 5(a)) or voluntarily (Figure 5(b)). The former is readily demonstrated in the human infant (Figure 5 (a)), when, at a certain level of bladder filling, increased afferent firing from bladder tension receptors produces firing in the sacral parasympathetic pathways and inhibition of sympathetic and somatic pathways. The expulsion phase consists of an initial relaxation of the urethral sphincter (Figure 5(a)) followed in a few seconds by a contraction of the bladder, an increase in bladder pressure, and flow of urine. Relaxation of the urethral outlet is mediated by activation of a parasympathetic reflex pathway to the urethra (Table 1) that triggers the release of nitric oxide, an inhibitory transmitter, as well as by removal of adrenergic and somatic excitatory inputs to the urethra.

Storage and voiding reflexes require the integrative action of neuronal populations at various levels of the neuraxis (Figure 6). Reflexes mediating the excitatory outflow to the sphincters and the sympathetic inhibitory outflow to the bladder are organized at the spinal level (Figure 6(a)), whereas the parasympathetic outflow to the bladder has a more complicated central organization involving spinal and spinobulbospinal pathways (Figures 6(b) and 7).

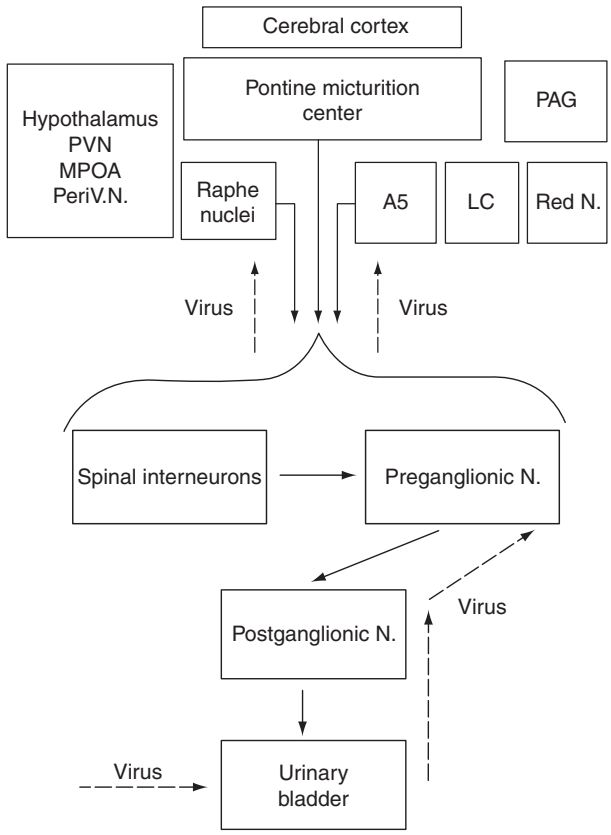

Figure 3 Structures in the brain and spinal cord of the adult and neonatal rat, labeled after injection of pseudorabies virus into the urinary bladder or the urethra. Virus is transported transneuronally in a retrograde direction (dashed arrows). Normal synaptic connections are indicated by solid arrows. At long survival times the virus can be detected in neurons at specific sites in the spinal cord and brain, extending to the PMC in the pons (i.e., Barrington's nucleus) and to the cerebral cortex. Other sites in the brain labeled by the virus are the paraventricular nucleus (PVN), medial preoptic area (MPOA), and periventricular nucleus (PeriV.N.) of the hypothalamus; the periaqueductal gray (PAG); the locus coeruleus (LC) and subcoeruleus; the red nucleus (N); the medullary raphe nuclei; and (A5) the noradrenergic cell group.

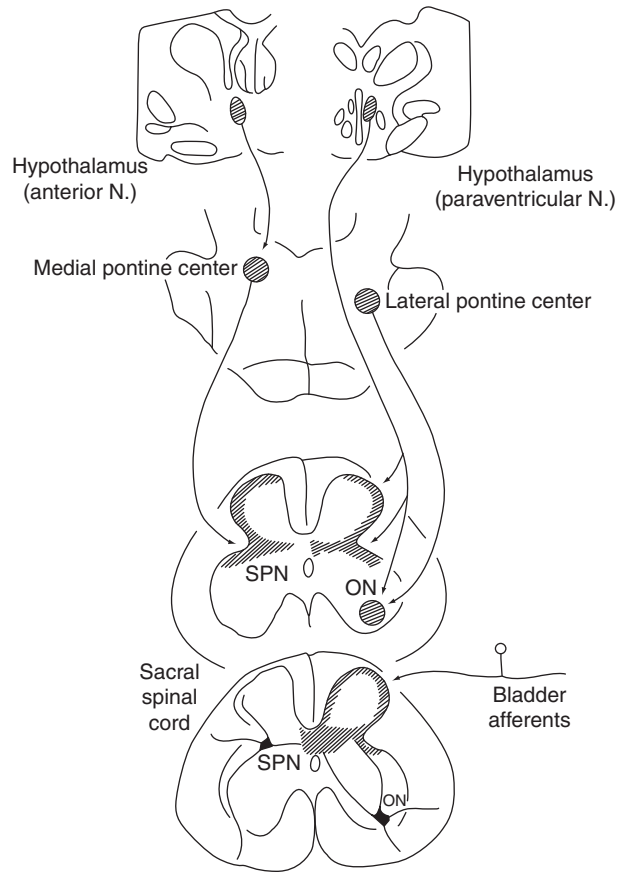

Figure 4 Neural connections between the brain and the sacral spinal cord that may be involved in the regulation of the lower urinary tract in the cat. The lower section of spinal cord shows the location and morphology of a preganglionic neuron in the sacral parasympathetic nucleus (SPN), a sphincter motor neuron in Onuf's nucleus (ON), and the sites of central termination of afferent projections from the urinary bladder. The upper spinal cord section shows the sites of termination of descending pathways arising in the medial PMC, the lateral pontine storage center, and the paraventricular nuclei of the hypothalamus. Section through the pons shows the projection from the anterior hypothalamic nuclei to the PMC.

Organization of Urine Storage Reflexes

Sympathetic storage reflex Although the integrity of the sympathetic input to the lower urinary tract is not essential for the performance of micturition, it does contribute to the storage function of the bladder. Surgical interruption or pharmacological blockade of the sympathetic innervation can reduce urethral outflow resistance and reduce bladder capacity. Sympathetic reflex activity is elicited by a sacrolumbar intersegmental spinal reflex pathway that is triggered by vesical afferent activity in the pelvic nerves (Figure 6(a)). This reflex pathway is inhibited when bladder pressure is raised to the threshold for producing micturition. The inhibitory response is abolished by transection of the spinal cord at the lower thoracic level, indicating that it originates at a supraspinal site. Thus, the vesicosympathetic reflex represents a negative feedback mechanism that allows the bladder to accommodate larger volumes during bladder filling, but is turned off during voiding to allow the bladder to empty completely (Figure 6(b)).

Urethral sphincter storage reflex Motor neurons innervating the striated muscles of the urethral sphincter exhibit a tonic discharge that increases during bladder filling (Figure 5). This activity is mediated in part by low-level afferent input from the bladder. During micturition the firing of sphincter motor neurons is inhibited. This inhibition is dependent in part

on supraspinal mechanisms, since it is less prominent after spinal cord injury.

Organization of Voiding Reflexes

Spinobulbospinal micturition reflex pathway Micturition is mediated by activation of the sacral parasympathetic efferent pathway to the bladder and the urethra, as well as by reciprocal inhibition of the somatic pathway to the urethral sphincter (Table 1, Figure 6(b)). Studies in animals using brain-lesioning and electrophysiological techniques revealed that the micturition reflex is mediated by a spinobulbospinal pathway that passes through the PMC in the rostral brain stem (Figures 6(b) and 7). Electrical stimulation in the PMC evokes bladder contractions and voiding, whereas application of inhibitory agents directly to the PMC suppresses micturition. Single-unit recording in the PMC has revealed increased neuronal activity during bladder distension. It has been proposed that the micturition reflex pathway functions as an on–off switch that is activated by a critical level of afferent activity arising from tension receptors in the bladder and is, in turn, modulated by inhibitory

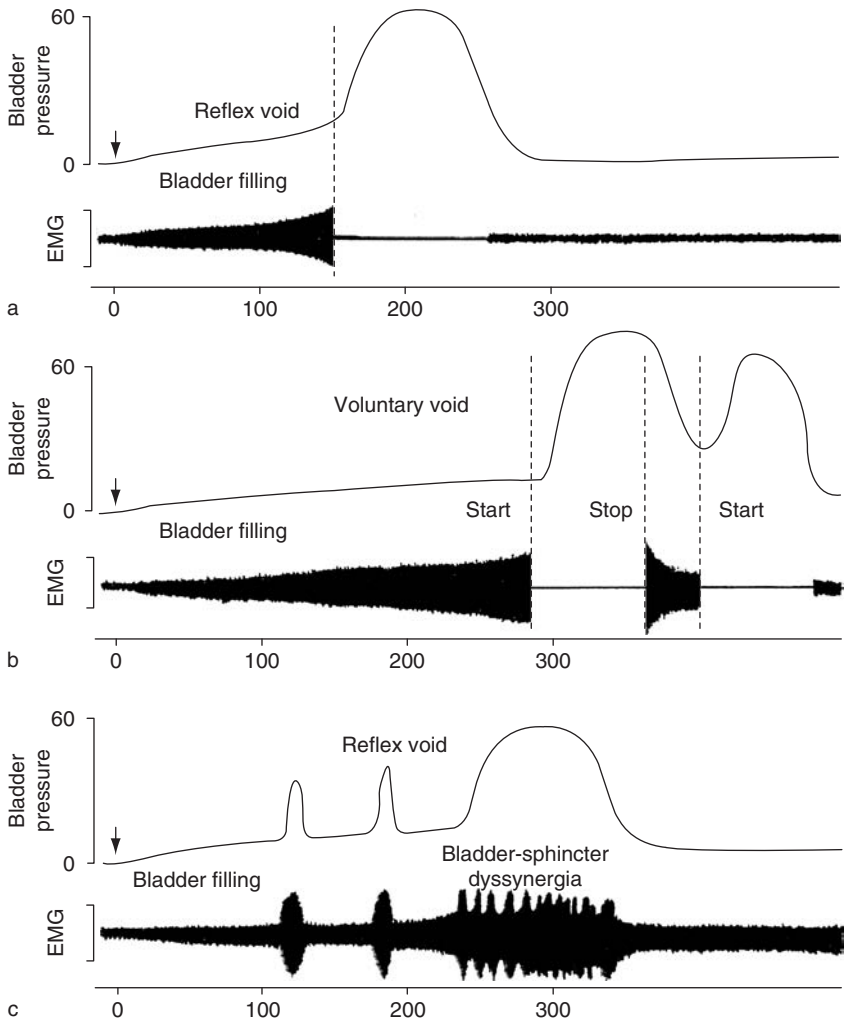

Figure 5 Combined cystometrograms and sphincter electromyograms (EMGs) comparing reflex voiding responses in an infant (a) and in a paraplegic patient (c) with a voluntary voiding response in an adult (b). The abscissa in all records represents bladder volume in milliliters and the ordinates represent bladder pressure in cmH_2O and electrical activity of the EMG recording. On the left side of each trace the arrows indicate the start of a slow infusion of fluid into the bladder (bladder filling). Vertical dashed lines indicate the start of sphincter relaxation, which precedes by a few seconds the bladder contraction in parts (a, b). In part (b), note that a voluntary cessation of voiding (stop) is associated with an initial increase in sphincter EMG followed by a reciprocal relaxation of the bladder. A resumption of voiding is again associated with sphincter relaxation and a delayed increase in bladder pressure. On the other hand, in the paraplegic patient (c) the reciprocal relationship between bladder and sphincter is abolished. During bladder filling, transient uninhibited bladder contractions occur in association with sphincter activity. Further filling leads to more prolonged and simultaneous contractions of the bladder and sphincter (bladder–sphincter dyssynergia). Loss of the reciprocal relationship between bladder and sphincter in paraplegic patients interferes with bladder emptying.

Table 1 Reflexes to the lower urinary tract

Afferent pathway	*Efferent pathway*	*Central pathway*
Urine Storage		
Low-level vesical afferent activity (pelvic nerve)	1. External sphincter contraction (somatic nerves) 2. Internal sphincter contraction (sympathetic nerves) 3. Detrusor inhibition (sympathetic nerves) 4. Ganglionic inhibition (sympathetic nerves) 5. Sacral parasympathetic outflow inactive	Spinal reflexes
Afferent activity from the external urethral sphincter	6. Inhibition of parasympathetic outflow	Spinal reflex
Micturition		
High-level vesical afferent activity (pelvic nerve)	1. Inhibition of external sphincter activity 2. Inhibition of sympathetic outflow 3. Activation of parasympathetic outflow to the bladder	Spinobulbospinal reflexes
	4. Activation of parasympathetic outflow to the urethra	Spinal reflex

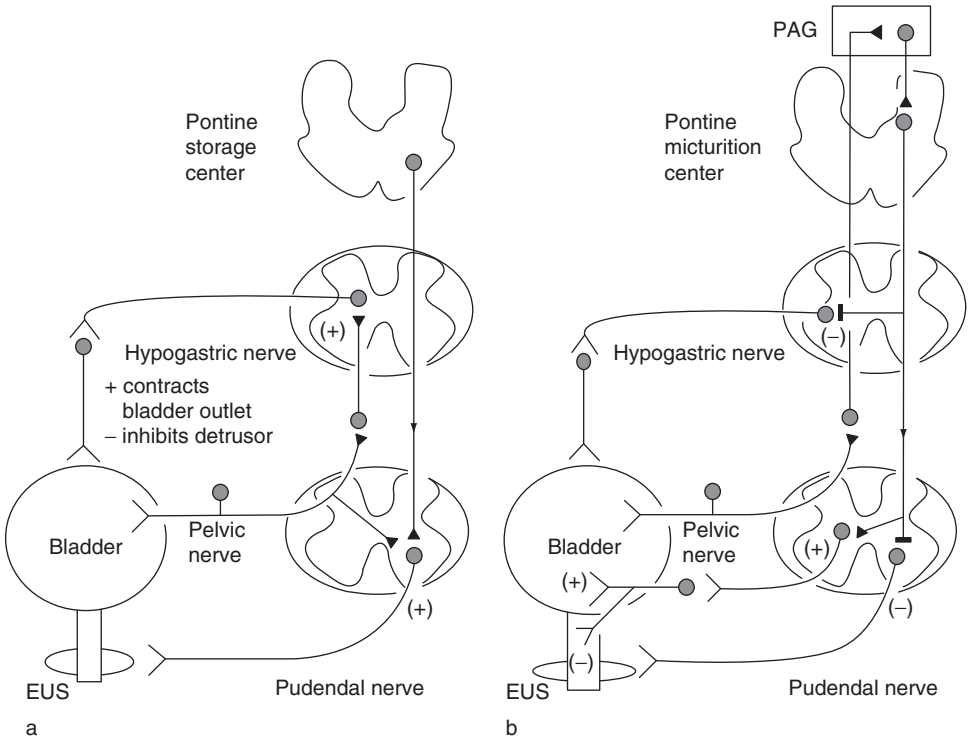

Figure 6 Diagram showing neural circuits controlling continence and micturition. (a) Urine storage reflexes. During the storage of urine, distension of the bladder produces low-level vesical afferent firing, which in turn stimulates the sympathetic outflow to the bladder outlet (base and urethra) and the pudendal outflow to the external urethral sphincter (EUS). These responses occur by spinal reflex pathways and represent guarding reflexes, which promote continence. Sympathetic firing also inhibits detrusor muscle and modulates transmission in bladder ganglia. A region in the rostral pons (the pontine storage center) increases external urethral sphincter activity. (b) Voiding reflexes. During elimination of urine, intense bladder afferent firing activates spinobulbospinal reflex pathways passing through the PMC, which in turn stimulates the parasympathetic outflow to the bladder and urethral sphincter smooth muscle and inhibits the sympathetic and pudendal outflow to the urethral outlet. Ascending afferent input from the spinal cord may pass through relay neurons in the periaqueductal gray (PAG) before reaching the PMC.

and excitatory influences from areas of the brain rostral to the pons (e.g., diencephalon and cerebral cortex) (Figure 7). Afferent input from the bladder is thought to reach the PMC after passing through a relay station in the PAG (Figure 6(b)).

Suprapontine control of micturition Studies in humans and animals indicate that voluntary control of micturition depends on connections between the frontal cortex, hypothalamus, and other forebrain structures, such as anterior cingulate gyrus, amygdala,

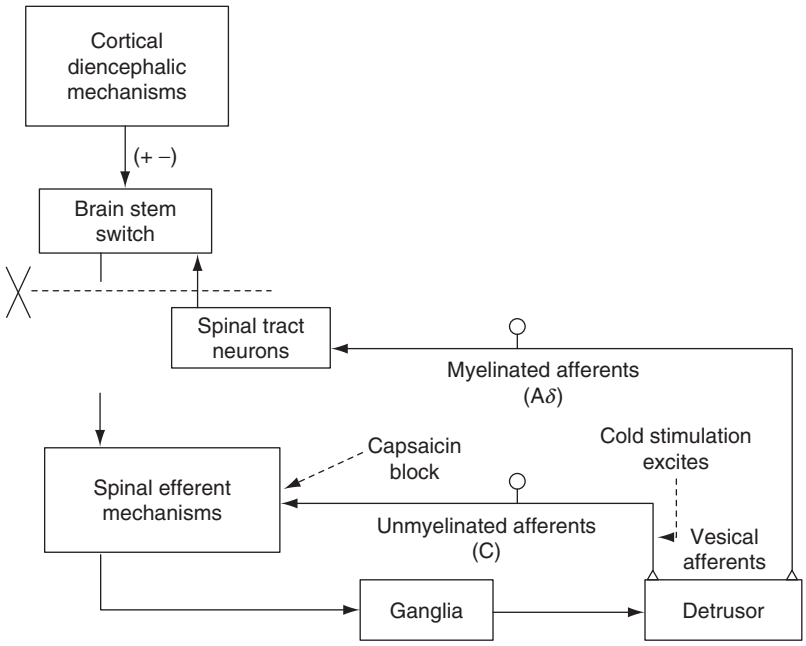

Figure 7 Diagram showing the organization of the parasympathetic excitatory reflex pathway to the detrusor muscle. Scheme is based on electrophysiologic studies in cats. In animals with an intact spinal cord, micturition is initiated by a supraspinal reflex pathway passing through a center in the brain stem. The pathway is triggered by myelinated afferents (A-δ fibers), which are connected to the tension receptors in the bladder wall. Injury to the spinal cord above the sacral segments (X) interrupts the connections between the brain and spinal autonomic centers and initially blocks micturition. However, over a period of several weeks following cord injury, a spinal reflex mechanism emerges, which is triggered by unmyelinated vesical afferents (C, C-fibers); the A-fiber afferent inputs are ineffective. The C-fiber reflex pathway is usually weak or undetectable in animals with an intact nervous system. Stimulation of the C-fiber bladder afferents by instillation of ice water into the bladder (cold stimulation) activates voiding responses in patients with spinal cord injury. Capsaicin (20–30 mg, subcutaneously) blocks the C-fiber reflex in chronic spinal cats, but does not block micturition reflexes in intact cats. Intravesical capsaicin also suppresses detrusor hyperreflexia and cold-evoked reflexes in patients with neurogenic bladder dysfunction.

bed nucleus of the stria terminalis, and septal nuclei, where electrical stimulation elicits excitatory bladder effects. Damage to the cerebral cortex due to tumors, aneurysms, or cerebrovascular disease appears to remove inhibitory control of the PMC, resulting in bladder overactivity.

Human brain imaging studies using single-photon emission computed tomography (SPECT), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) have examined the areas of the brain involved in the control of micturition. Some studies evaluated the brain areas responsible for the perception of bladder fullness and the sensation of the desire to void during bladder filling, whereas others examined brain activity during micturition or voluntary contractions of the pelvic floor during urine withholding. PET scan studies revealed that two cortical areas (the right dorsolateral prefrontal cortex and the anterior cingulate gyrus) were active (i.e., exhibited increased blood flow) during voiding (Figure 8). The hypothalamus, including the preoptic area as well as the pons and the PAG, also showed activity in concert with voluntary micturition. Other PET studies that examined the changes in brain activity during filling of the bladder revealed that increased activity occurred in the PAG, midline pons, mid-cingulate gyrus, and bilaterally in the frontal lobes. The results were consistent with the notion that the

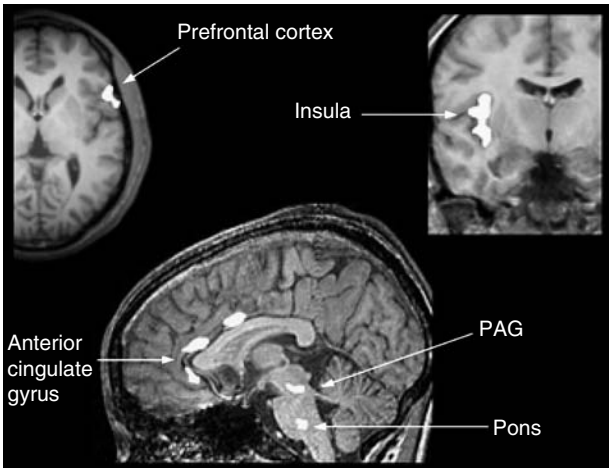

Figure 8 Regions in the human brain identified in PET imaging studies that exhibit differences in activity based on whether the bladder was full or empty (PAG, periaqueductal gray).

PAG receives information about bladder fullness and then relays this information to other brain areas involved in the control of bladder storage.

Spinal micturition reflex pathway Spinal cord injury rostral to the lumbosacral level eliminates voluntary and supraspinal control of voiding, leading initially to an areflexic bladder and complete urinary retention followed by a slow development of automatic micturition and bladder hyperactivity (Figure 5(c)) mediated by spinal reflex pathways. However, voiding is commonly inefficient due to simultaneous contractions of the bladder and urethral sphincter (bladder–sphincter dyssynergia) (Figure 5(c)). Electrophysiologic studies in animals have shown that the micturition reflex pathway changes after spinal cord injury. For example, the afferent limb of the micturition reflex in cats with chronic spinal transection above the lumbar level consists of unmyelinated (C-fiber) axons (Figure 7), whereas in cats with an intact spinal cord, myelinated (A-δ) afferents activate the micturition reflex (Figure 7). In normal cats, capsaicin, a neurotoxin known to disrupt the function of C-fiber afferents, did not block reflex contractions of the bladder or the A-δ-afferent fiber-evoked bladder reflex. However, in cats with chronic spinal injury, capsaicin completely blocked C-fiber-evoked bladder reflexes. The emergence of C-fiber-evoked bladder reflexes seems to be mediated by several mechanisms, including changes in central synaptic connections and alterations in the properties of the peripheral afferent receptors that lead to sensitization of the mechanoinsensitive C-fiber afferents and the unmasking of mechanosensitivity. Evidence for a role of C-fiber bladder afferents in neurogenic bladder overactivity and involuntary voiding in humans has been obtained in studies in which capsaicin or resiniferatoxin, another C-fiber afferent neurotoxin, was administered intravesically to patients with detrusor overactivity due to multiple sclerosis or spinal cord injuries. In these patients the toxins increased bladder capacity and reduced the frequency of incontinence.

Neurotransmitters in Central Micturition Reflex Pathways

Excitatory Neurotransmitters

Excitatory transmission in the central pathways controlling the lower urinary tract involves several types of neurotransmitters, including glutamic acid, neuropeptides, nitric oxide, and adenosine triphosphate (ATP). Pharmacological experiments in rats revealed that glutamic acid is an essential transmitter in the ascending, pontine, and descending limbs of the spinobulbospinal micturition reflex pathway and in spinal reflex pathways controlling the bladder and external urethral sphincter. *N*-Methyl-D-aspartate (NMDA) and non-NMDA glutamatergic synaptic mechanisms appear to interact synergistically to mediate transmission in these pathways.

Inhibitory Neurotransmitters

Several types of inhibitory transmitters, including inhibitory amino acids (γ-aminobutyric acid (GABA) and glycine) and opioid peptides (enkephalins), can suppress the micturition reflex when applied to the central nervous system. Experimental evidence in anesthetized animals indicates that GABA, glycine, and enkephalins exert a tonic inhibitory control in the PMC and regulate bladder capacity. These substances also have inhibitory actions in the spinal cord.

Neurotransmitters with Mixed Excitatory and Inhibitory Actions

Some transmitters (dopamine, 5-hydroxytryptamine, norepinephrine, acetylcholine, and nonopioid peptides, including vasoactive intestinal polypeptide and corticotropin-releasing factor) have both inhibitory and excitatory effects on reflex bladder activity. In some instances the different effects are mediated by different types of receptors. For example, the inhibitory effects of dopamine are mediated by D_1-like (D_1 and D_5) receptor subtypes, and the facilitatory effects are mediated by D_2-like (D_2, D_3, and D_4) receptor subtypes. Loss of forebrain dopaminergic inhibitory mechanisms in patients with idiopathic Parkinson's disease is associated with bladder hyperactivity.

5-Hydroxytryptamine has complex effects on the lower urinary tract; these vary in different species. For example, in the cat, activation of central 5-HT_{1A} receptors inhibits reflex bladder activity; whereas activation of 5-HT_2 receptors enhances urethral sphincter activity. On the other hand, activation of central 5-HT_{1A} receptors in the rat enhances bladder and sphincter reflexes.

Central Neural Control of the Urogenital Organs

The physiologic changes initiated by erotic stimuli can be divided into four distinct phases (excitement, plateau, orgasm, and resolution) that have been designated collectively as the sexual response cycle. Although anatomic differences obviously preclude identical responses in males and females during each phase of the cycle, it is clear that similar vascular responses (skin flush, penile, and clitoral erection),

secretory responses (stimulation of the prostate, bulbourethral gland, glands of Littre in the male; Bartholin's and paraurethral glands in the female), and responses of smooth and striated muscles occur in both sexes. The following sections review the central nervous system control of the principal autonomic and somatic responses in the male sexual response cycle (erection, secretion, emission, and ejaculation).

Penile Erection

Penile erection, which is one of the first responses to occur during sexual arousal, is a vascular phenomenon resulting from neurally mediated increase in blood flow to the penile erectile tissue (corpora cavernosa and corpus spongiosum). The erectile tissue consists of large venous sinuses that contain very little blood when the penis is flaccid, but distend considerably when blood flow is increased. Dilation in the arterial supply to the cavernous tissue, coupled with a relaxation of the sinusoidal smooth muscle in the trabecular tissue, is responsible for erection.

Penile erection is primarily an involuntary or reflex phenomenon that can be elicited by a variety of stimuli and by at least two distinct central mechanisms: psychogenic and reflexogenic (Table 2). Psychogenic erections are initiated by supraspinal centers in response to auditory, visual, olfactory, tactile, and imaginative stimuli and can be elicited by peripheral neural pathways arising in either the sacral parasympathetic or the thoracolumbar sympathetic nuclei of the spinal cord (Table 2). Reflexogenic erections, which are initiated by exteroceptive stimulation of the genital regions, are mediated by a sacral spinal reflex mechanism having an afferent limb in the pudendal nerves and an efferent limb in the sacral parasympathetic nerves. Electrophysiological studies in animals have revealed that the sacral parasympathetic reflex mechanism involved in penile erection is not altered by transection of the spinal cord above the lumbar level, indicating that the reflex circuitry is contained in the spinal cord.

In patients with lower motor neuron lesions involving the sacral spinal cord, reflexogenic erections are abolished but psychogenic erections may still occur via the sympathetic innervation to the penis. In patients with spinal cord lesions above the level of T12, psychogenic erections are abolished but reflexogenic reactions persist. Under normal conditions it is likely that psychogenic and reflexogenic stimuli act synergistically in producing erections. It is also known that psychologic factors, such as guilt and hostility, or endocrine disturbances that influence libido or supraspinal centers, can interfere with erectile reflexes.

Brain imaging studies using fMRI or PET during sexual arousal induced by erotic visual stimuli have identified several brain regions in both males and females that are activated during sexual stimulation. PET studies in males detected activation in three general regions, including limbic–paralimbic areas (anterior cingulate gyrus, orbitofrontal cortex), the striatum (head of the caudate nucleus), and the posterior hypothalamus. Strong signals specifically associated with penile turgidity were observed in the right subinsular region (claustrum, caudate, cingulate gyrus) using fMRI. A PET study also showed increased blood flow in the right prefrontal cortex during orgasm in males, whereas all other cortical areas showed decreases in blood flow.

It has been suggested that the activation of paralimbic areas is correlated with the emotional and motivational states associated with sexual arousal, whereas the activation of the anterior cingulate and hypothalamic regions is related to the affective,

Table 2 Male sexual reflexes

Response	*Afferent nerves*	*Efferent nerves*	*Central pathway*	*Effector organ*
Penile erection Reflexogenic	Pudendal nerve	Sacral parasympathetic	Sacral spinal reflex	Dilation of arterial supply to corpus cavernosum and corpus spongiosum
Psychogenic	Auditory, imaginative, visual, olfactory, tactile	Sacral parasympathetic, lumbar sympathetic	Supraspinal origin	
Glandular secretion	Pudendal nerve	Sacral parasympathetic, lumbar sympathetic	Sacral spinal reflex	Seminal vesicle and prostate
Seminal emission	Pudendal nerve	Lumbar sympathetic	Intersegmental spinal reflex (sacrolumbar)	Contraction of vas deferens, ampulla, seminal vesicles, prostate, and closure of bladder neck
Ejaculation	Pudendal nerve	Somatic efferents in pudendal nerve	Sacral spinal reflex	Rhythmic contractions of bulbocavernosus and ischiocavernosus muscles

autonomic, and endocrine responses. Recent fMRI studies in males have also detected activation in parietal areas known to be involved in attentional processes. The human data correlate well with results from animal experiments which indicate that connections between the limbic system and the hypothalamus (medial preoptic and paraventricular nuclei) are essential for the stimulation of the descending autonomic projections to the spinal cord that trigger psychogenic penile erections.

Glandular Secretion

During the second phase of the sexual response cycle (plateau), activity in parasympathetic pathways stimulates mucus secretion from bulbourethral and Littre's glands and secretion from the seminal vesicles and the prostate gland. Mucus secretion contributes to lubrication of the penis, whereas secretions from the seminal vesicles and prostate provide the bulk of the fluid and chemical factors that contribute to the viability and motility of the spermatozoa. Glandular secretion is thought to be mediated by the parasympathetic system; however, the central reflex mechanisms and the neurotransmitters have not been identified.

Emission–Ejaculation

The third phase of the sexual act (orgasm), which is accompanied by emission and ejaculation of semen, involves the coordination of autonomic and somatic reflex mechanisms at different levels of the lumbosacral spinal cord. During the first step in this process (emission), reflex activity in the thoracolumbar sympathetic outflow elicits rhythmic contractions of the smooth muscle of the seminal vesicles, prostate, ductus deferens, and ampulla, resulting in the ejection of sperm and glandular secretions into the urethra and at the same time a closure of the bladder neck, to prevent the backflow of semen into the bladder. Pharmacologic studies have shown that these responses are mediated by sympathetic neurotransmitters, norepinephrine and ATP, interacting with α-adrenergic receptors and purinergic receptors, respectively, in the peripheral organs. After emission of semen into the proximal urethra, rhythmic contractions of the bulbocavernosus, ischiocavernosus, and paraurethral striated muscles result in ejaculation. The afferent and efferent limbs of the ejaculation reflex are contained in the pudendal nerve (Table 2). The sensations accompanying ejaculation constitute the orgasm. Orgasm is not necessarily affected by sympathectomy, provided that the pudendal nerves remain intact. Thus, neither afferent fibers in the sympathetic nerves nor contractions of smooth muscles of the seminal vesicles and ductus deferens are essential for the occurrence of orgasm.

In the spinal cord, ascending fibers carrying sensory information from the sex organs seem to lie primarily within the anterolateral tracts. Thus, bilateral cordotomy for the treatment of chronic pain usually diminishes or completely abolishes orgasm. Cordotomy also severely compromises ejaculatory and erectile mechanisms, although the latter may recover over a period of time. With more extensive damage of the spinal cord in paraplegics, when the site of injury is located rostral to T12, ejaculation occurs in a relatively small percentage of patients in comparison to reflexogenic erections, which are readily elicited. This observation is consistent with the greater complexity of spinal reflex pathways underlying ejaculation and indicates a considerable dependence of these pathways on supraspinal coordinating mechanisms.

Recent studies in rats have identified a population of lumbar spinal neurons located around the central canal in lamina X and medial lamina VII that participate in the ejaculatory reflex. These neurons express neurokinin-1 receptors, receive afferent input from the penis, and project to the thalamus. It was initially hypothesized that these neurons might function primarily in transmitting sensory information from the penis to pleasure centers in the brain. However, it was discovered that these neurons were only activated during ejaculation and not during other components of male sexual behavior. When the neurons were destroyed by intrathecal administration of a toxin (saporin conjugated to a substance P analog) that selectively destroys neurons expressing neurokinin-1 receptors, ejaculatory responses were completely abolished without altering other components of copulatory behavior (penile erection, mounting, intromission). It was concluded that the neurons may be part of a spinal ejaculation generator.

Central Neurotransmitters

Pharmacological studies in animals have implicated many neurotransmitters in the central control of sexual function. The dopaminergic system in the medial preoptic area and the oxytocin and nitric oxide systems in the paraventricular nucleus exert major excitatory effects on erection, while the serotonergic pathways arising in the raphe nuclei seem to be inhibitory. Elsewhere in the brain, noradrenergic, GABAergic, and serotonergic pathways are generally inhibitory, whereas dopaminergic, nitric oxide, and oxytocin pathways facilitate sexual function. Nitric oxide appears to tonically modulate the hypothalamic excitatory circuitry, because intracerebroventricular injections of nitric oxide synthase inhibitors prevent the penile erectile responses in rats induced by dopamine agonists and oxytocin. It is uncertain whether all of these findings in animals are entirely applicable

to man. However, recently, apomorphine, a nonselective dopamine receptor agonist that stimulates sexual behavior and penile erection in animals, has been marketed to treat penile erectile dysfunction in patients.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Urogenital Control; Autonomic Nervous System: Neuroanatomy; Autonomic Neuroimmunology; Autonomic Neuroplasticity: Development; Autonomic Nervous System; Sympathetic Nervous System.

Further Reading

Argiolas A and Melis MR (2005) Central control of penile erection: Role of the paraventricular nucleus of the hypothalamus. *Progress in Neurobiology* 76: 1–21.

Coolen LM, Allard J, Truitt WA, et al. (2004) Central regulation of ejaculation. *Physiology & Behavior* 83: 203–215.

de Groat WC (2006) Integrative control of the lower urinary tract: Preclinical perspective. *British Journal of Pharmacology* 147 (supplement 2): S25–S40.

de Groat WC and Yoshimura N (2001) Pharmacology of the lower urinary tract. *Annual Review of Pharmacology and Toxicology* 41: 691–721.

de Groat WC and Yoshimura N (2006) Mechanisms underlying the recovery of lower urinary tract function following spinal cord injury. *Progress in Brain Research* 152: 59–84.

Ferretti A, Caulo M, Del GC, et al. (2005) Dynamics of male sexual arousal: Distinct components of brain activation revealed by fMRI. *NeuroImage* 26: 1086–1096.

Giuliano F and Rampin O (2004) Neural control of erection. *Physiology & Behavior* 83: 189–201.

Kavia RB, Dasgupta R, and Fowler CJ (2005) Functional imaging and the central control of the bladder. *Journal of Comparative Neurology* 493: 27–32.

Marson L and Foley KA (2004) Identification of neural pathways involved in genital reflexes in the female: A combined anterograde and retrograde tracing study. *Neuroscience* 127: 723–736.

Marson L, Platt KB, and McKenna KE (1993) Central nervous system innervation of the penis as revealed by the transneuronal transport of pseudorabies virus. *Neuroscience* 55: 263–280.

Morrison J, Birder L, Craggs M, et al. (2005) Neural control. In: Abrams P, Cardozo L, Khoury S, et al. (eds.) *Incontinence*, pp. 363–422. Jersey: Health Publications Ltd.

Nadelhaft I and Vera PL (1995) Central nervous system neurons infected by pseudorabies virus injected into the rat urinary bladder following unilateral transection of the pelvic nerve. *Journal of Comparative Neurology* 359: 443–456.

Saenz de Tejada I, Cavidad NG, Heaton J, et al. (2001) Anatomy, physiology, and pathophysiology of erectile function. In: Jardin A, Wagner G, Khoury S, et al. (eds.) *Erectile Dysfunction*, pp. 65–102. Jersey: Health Publications Ltd.

Sasaki M (2005) Properties of Barrington's neurones in cats: Units that fire inversely with micturition contraction. *Brain Research* 1033: 41–50.

Truitt WA and Coolen LM (2002) Identification of a potential ejaculation generator in the spinal cord. *Science* 297: 1566–1569.

Autonomic Nervous System: Clinical Testing

C J Mathias, Imperial College London at St. Mary's Hospital, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Basic Principles

The autonomic nervous system, through its major efferent components, the parasympathetic and sympathetic nervous systems, supplies every organ in the body (Figure 1(a)). It influences local organ function and integrated systems such as those that control blood pressure and body temperature. The parasympathetic nervous system has a cranial division that innervates the lacrimal glands, iris musculature, salivary glands, heart, and upper gastrointestinal tract; the sacral division supplies the urinary bladder, sexual organs, and lower gastrointestinal tract. The sympathetic nervous system emerges from the thoracic and upper lumbar segments of the spinal cord and innervates every organ in the body. Both systems have pre- and postganglionic segments. Some preganglionic fibers continue through the paravertebral ganglia, forming the greater and lesser splanchnic nerves, with synapses in the celiac, superior mesenteric, inferior mesenteric, and renal ganglia, from where postganglionic unmyelinated fibers emerge; these comprise the enteric nervous system, considered the third component of the autonomic nervous system.

Autonomic activity is controlled and modulated at a number of levels within the neural axis. There are major cerebral centers, especially in the hypothalamus and brain stem, which are influenced by a variety of peripheral sensory inputs, in addition to signals from cortical and other areas of the brain. These central centers and pathways increasingly are being delineated in humans using modern neuroimaging and allied technological advances. Autonomic efferent activity is a composite response to a variety of information derived externally and internally, with integration at different levels of the neuraxis that maintain functioning of the organs and systems in a wide range of environments and stimuli.

Acetylcholine (ACh) is the major transmitter at ganglia acting through the nicotinic receptor and affecting both parasympathetic and sympathetic activity (Figure 1(b)). ACh also is the neurotransmitter within the adrenal medulla, which functions as a ganglion, but with an endocrine role. It secretes adrenaline and noradrenaline (NA), which enter the circulation and act on α-adrenoreceptors (in blood vessels) and β-adrenoreceptors (in the heart and lungs). Neurons often contain more than one neurotransmitter and a range of substances, including neuropeptides, that play important roles in influencing autonomic activity and target organ function. The major neurotransmitter at sympathetic postganglionic sites is noradrenaline, formed through a series of metabolic steps from the amino acid tyrosine (Figure 1(c)). ACh acts at parasympathetic postganglionic nerve endings and at sympathetic terminals on sweat glands through the muscarinic receptor.

Clinical Features

Localized lesions (Table 1), such as Horner's syndrome, cause few symptoms but characteristic signs such as ptosis, a small pupil, and anhidrosis, indicating damage to sympathetic pathways to the face. The underlying cause, however, may be a tumor at the apex of the lung, a lesion within the spinal cord, or dissection of the internal carotid artery. Determining the underlying disorders that contribute to morbidity and mortality therefore is of importance. Generalized disorders (Table 2) often are caused by irreversible damage to the autonomic nervous system. There are many manifestations resulting from impairment of different regions and systems (Table 3). The symptoms (Table 4) and confirmation of postural (orthostatic) hypotension (Figure 2(a)) often lead to consideration of autonomic involvement.

In secondary autonomic dysfunction, the clinical manifestations of the primary disorder, such as muscle paralysis in tetraplegia, may predominate. Autonomic features include severe paroxysmal hypertension with headache and facial flushing due to autonomic dysreflexia. In diabetes mellitus, gustatory sweating (hyperhidrosis over the head and neck during food ingestion), impotence, and nocturnal diarrhea occur. Drugs are a frequent cause of autonomic dysfunction (Table 5).

Some disorders result in intermittent autonomic impairment. They include neurally mediated syncope, where hypotension and bradycardia contribute to fainting. There often are no discernible abnormalities, autonomic or otherwise, between attacks. Of importance are the history and the association of attacks with provocative factors. In the recently described disorder, the postural tachycardia syndrome (PoTS), palpitations occur during standing and exertion, with an abnormal elevation of heart rate to over 30 beats min^{-1} on postural challenge.

Clinical Investigation of Autonomic Function

The clinical investigative aims are to ascertain the presence and degree of autonomic dysfunction, define the etiological factors, and delineate the site of the lesion. Each is important for diagnosis, prognosis, and management. A variety of tests relating to different systems may need to be performed (Table 6). This depends on the presenting problem, the possible underlying disorder, and the strategies planned for treatment and its evaluation. Advances in technology now enable measurements noninvasively and continuously, as exemplified in relation to blood pressure (BP), heart rate (HR), and other cardiovascular variables.

Cardiovascular System

Physiological

Gravitational challenge Postural hypotension is a cardinal manifestation of sympathetic failure. A fall of 20 mmHg systolic or 10 mmHg diastolic blood pressure, or less in the presence of symptoms, warrants further investigation. Measurements are first made with the individual lying flat, with adequate readings to determine the stability of blood pressure before postural change. This is induced either by using a tilt table (to 60°) or by making the individual stand upright. Prolonged tilt-table testing, for up to 45 min, also is used in the investigation of neurally mediated syncope, sometimes in combination with

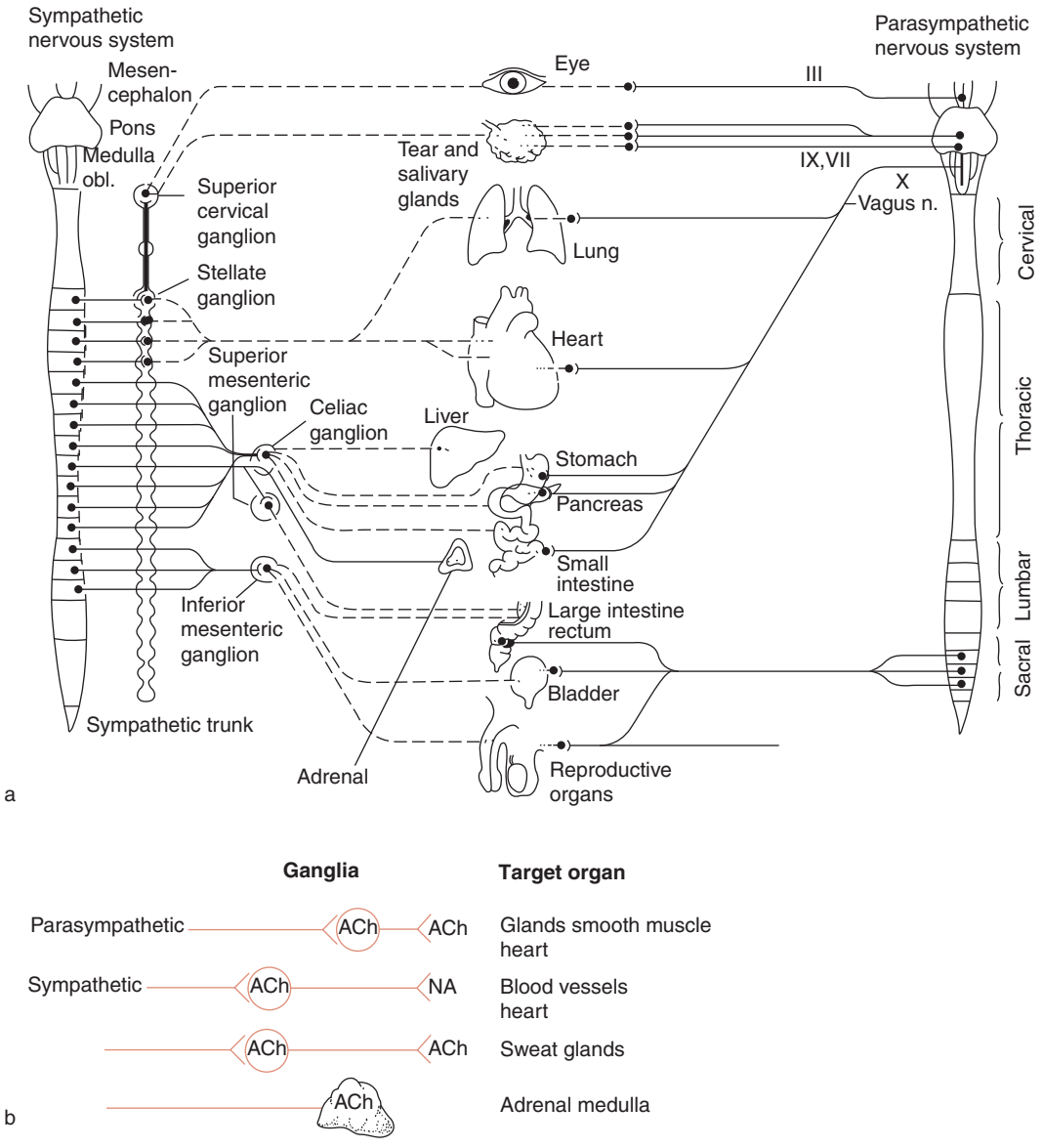

Figure 1 Continued

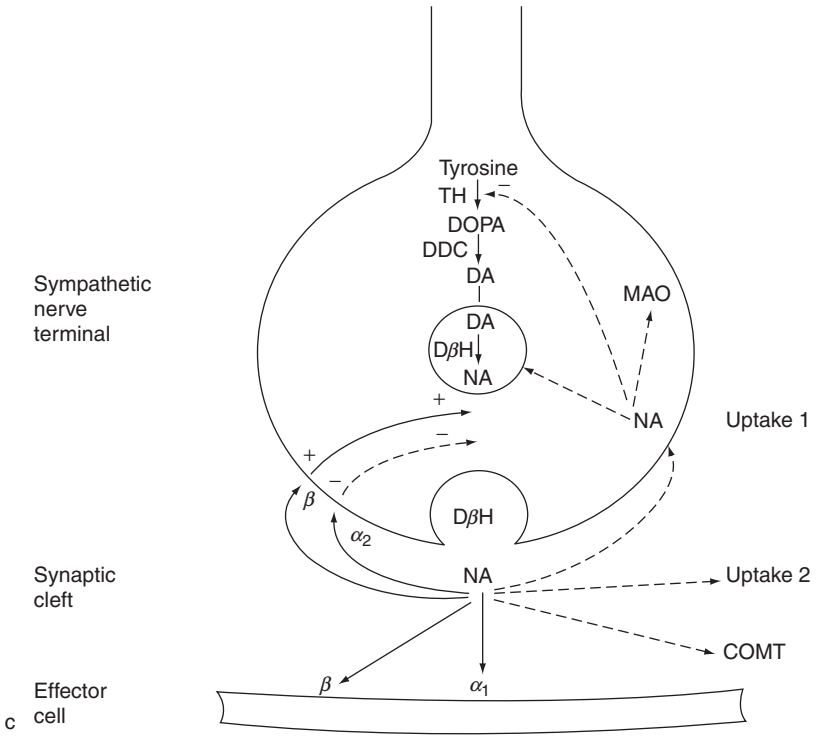

Figure 1 (a) Schema outlining details of the craniosacral parasympathetic and thoracolumbar sympathetic outflows to various target organs. (b) Neurotransmitters subserving the parasympathetic and sympathetic pathways at ganglia and target organs. (c) Steps involved in the formation of noradrenaline (NA) from tyrosine within a sympathetic nerve terminal. NA in granules is released by a process of exocytosis into the synaptic cleft, following which it acts on various α- or β-adrenoreceptors, either pre- or post-synaptically. NA is subject to various processes, which involve uptake 1 into the nerve terminal, following which it is either incorporated into granules, exerts negative feedback on TH, or is metabolized by monoamine oxidase (MAO). Some is taken up into nonneuronal tissues (uptake 2) and some is metabolized by catechol-*O*-methyltransferase (COMT), while the rest spills over into the circulation. ACh, acetylcholine; DA, dopamine; DβH, dopamine β-hydroxylase, DDC, dopa decarboxylase; DOPA, dihydroxyphenylalanine; TH, tyrosine hydroxylase. (a) From Janig W and McLachlan EM (2001) Neurobiology of the autonomic nervous system. In: Mathias CJ and Bannister R (eds.) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System*, 4th edn., pp. 3–15. Oxford: Oxford University Press. (b) From Mathias CJ (1987) Autonomic dysfunction. *British Journal of Hospital Medicine* 38: 238–243. (c) From Mathias CJ (2000) Disorders of the autonomic nervous system. In: Bradley WG, Daroff RB, Fenichel GM, et al. (eds.) *Neurology in Clinical Practice*, 3rd edn., pp. 2131–3265. Boston: Butterworth-Heinemann.

Table 1 Examples of localized autonomic disorders

Horner's syndrome
Holmes–Adie pupil
Crocodile tears (Bogorad's syndrome)
Gustatory sweating (Frey's syndrome)
Reflex sympathetic dystrophy
Idiopathic palmar, plantar, or axillary hyperhidrosis
Chagas' disease (*Trypanosoma cruzi*)[a]
Surgical procedures[b]
Sympathectomy (regional) causing compensatory hyperhidrosis
Vagotomy and gastric drainage procedures in 'dumping' syndrome
Organ transplantation (heart, lungs)

[a]Listed here because it specifically targets intrinsic cholinergic plexuses in the heart and gut.
[b]Surgery also may cause other localised disorders, such as Frey's syndrome, after parotid surgery.
From Mathias CJ and Bannister R (2001) Investigation of autonomic disorders. In: Mathias CJ and Bannister R (eds.) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System,* 4th edn., pp. 169–195. Oxford: Oxford University Press.

supraphysiological (negative body pressure) or pharmacological (glyceryl trinitrate) stimuli.

In normal individuals there are only small changes in blood pressure during head-up tilt or standing. In autonomic failure, the fall is often rapid and a return to horizontal restores blood pressure levels (Figure 2(a)). There may be considerable variability in the extent of postural hypotension, as many factors in daily life adversely or favorably influence blood pressure in these individuals. Postural hypotension is greater in the morning, after food ingestion, with exercise, in hot weather, and if vasodilator drugs are used (Table 7). Water ingestion (500 ml) and various physical maneuvers (such as squatting, leg crossing, and tensing of leg muscles) have the opposite effect and reduce postural hypotension. It is important to exclude nonneurogenic causes that can exacerbate neurogenic postural hypotension (Table 8).

There is normally a modest rise in heart rate during postural change. A lack of change in heart rate despite

Table 2 Outline classification of various causes of autonomic failure

Primary
Acute/subacute dysautonomias
Pure cholinergic dysautonomia
Pure pandysautonomia
Pandysautonomia with neurological features
Chronic autonomic failure syndromes
Pure autonomic failure
Multiple system atrophy (Shy–Drager syndrome)
Autonomic failure with Parkinson's disease
Diffuse Lewy body disease
Secondary
Congenital
Nerve growth factor deficiency
Hereditary
Autosomal dominant trait in familial amyloid neuropathy
Autosomal recessive trait
Familial dysautonomia – Riley–Day syndrome
Dopamine β-hydroxylase deficiency
Metabolic diseases
Diabetes mellitus
Chronic renal failure
Chronic liver disease
Alcohol induced
Inflammatory
Guillain–Barré syndrome
Transverse myelitis
Infections
Bacterial – tetanus
Viral – human immunodeficiency virus infection
Prion – fatal familial insomnia
Neoplasia
Brain tumors – especially of third ventricle or posterior fossa
Paraneoplastic, to include adenocarcinomas of lung and pancreas
Trauma
Cervical and high thoracic spinal cord transection
Drugs
By their direct effects
By causing a neuropathy
Neurally mediated syncope
Vasovagal syncope
Carotid sinus hypersensitivity
Situational syncope
Postural tachycardia syndrome

From Mathias CJ (2003) Disorders of the autonomic nervous system. In: Bradley WG, Daroff RB, Fenichel GM, et al. (eds.) *Neurology in Clinical Practice*, 3rd edn., pp. 2403–2420. Boston, MA: Butterworth-Heinemann.

Table 3 Clinical manifestations in chronic primary autonomic failure[a]

Cardiovascular	Postural hypotension
Sweating	Anhidrosis, heat intolerance
Gastrointestinal	Oropharyngeal dysphasia, constipation, occasionally diarrhea
Renal and urinary bladder	Nocturia, frequency, urgency, incontinence, retention
Genital	Erectile and ejaculatory failure in the male
Eye	Anisocoria, Horner's syndrome
Respiratory	Stridor, involuntary inspiratory gasps, apneic episodes
Other neurological deficits	Parkinsonian, cerebellar, and pyramidal features

[a]The respiratory and additional neurological deficits do not occur in pure autonomic failure and usually are part of multiple system atrophy/Shy–Drager syndrome.
From Mathias CJ (1997) Autonomic disorders and their recognition. *New England Journal of Medicine* 10: 721–724.

Table 4 Some of the symptoms resulting from orthostatic hypotension and impaired perfusion of various organs

Cerebral hypoperfusion
Dizziness
Visual disturbances
Blurred – tunnel
Scotoma
Graying out – blacking out
Color defects
Syncope
Impaired cognition
Muscle hypoperfusion
Paracervical and suboccipital ('coathanger') ache
Lower back/buttock ache
Calf claudication
Chest wall and cardiac hypoperfusion
Chest pain
Angina pectoris
Spinal cord hypoperfusion
Lower limb weakness and parasthesiae
Renal hypoperfusion
Oliguria
Nonspecific
Weakness, lethargy, fatigue
Falls

From Mathias CJ, Mallipeddi R, and Bleasdale-Barr K (1999) Symptoms associated with orthostatic hypotension in pure autonomic failure and multiple system atrophy. *Journal of Neurology* 246: 893–898.

a substantial fall in blood pressure usually is indicative of cardiac parasympathetic failure or afferent baroreceptor block. In tetraplegia the heart rate rises in response to the fall in blood pressure, as the sympathetic efferent components of the baroreceptor reflex are interrupted within the cervical spinal cord, while the vagi from the brain stem are intact (Figure 2(b)). The heart rate, however, does not usually rise above 110 beats min^{-1}, which is similar to cardiac parasympathetic withdrawal as observed after vagal blockade with atropine. Further elevation probably is dependent on adrenomedullary stimulation and elevation of plasma adrenaline levels, which does not occur in tetraplegics. This adrenal component probably accounts for the greater tachycardia observed in subjects with an intact sympathetic nervous system when blood pressure is lowered substantially, as in hemorrhagic shock.

In neurally mediated syncope (such as vasovagal syncope), autonomic testing indicates normal autonomic function between attacks. Blood pressure often is maintained initially during head-up tilt, but falls during provocation; this may be preceded by a fall in heart rate (Figure 2(c)). The former results from withdrawal of sympathetic activity and the latter from vagal overactivity. This differs markedly from the responses to

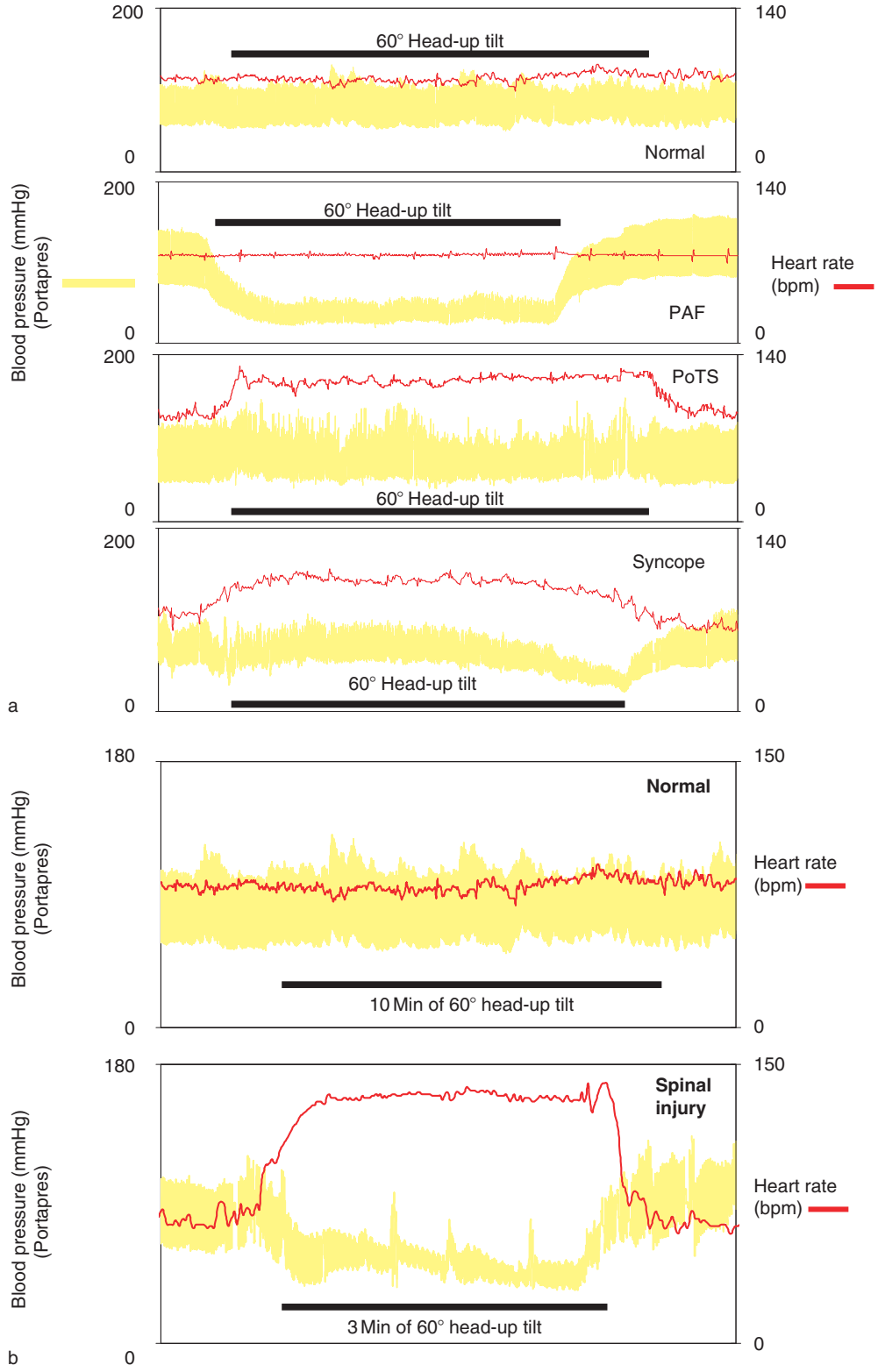

Figure 2 Continued

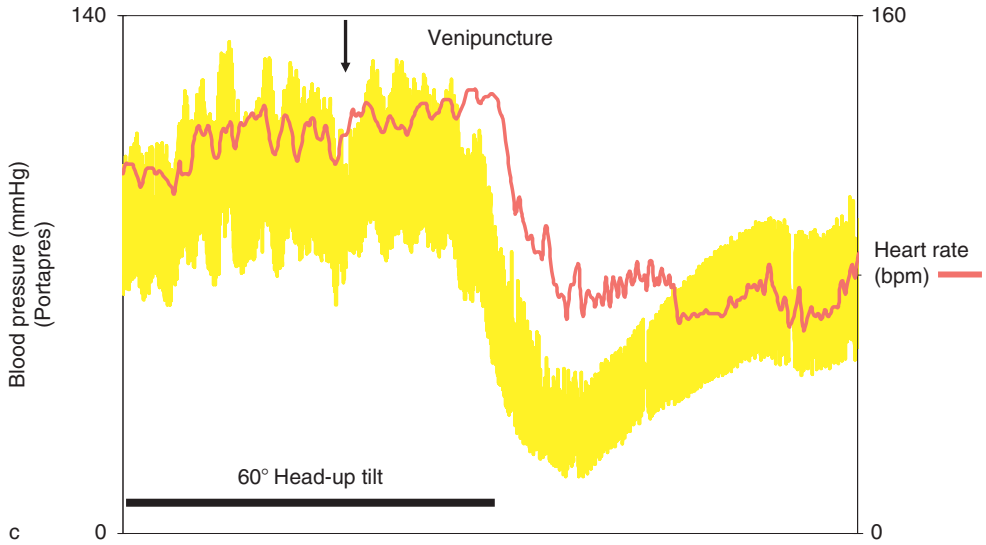

Figure 2 (a) Blood pressure and heart rate (beats per minute, bpm) measured continuously before, during, and after 60° head-up tilt by the Portapres II in a normal person and in individuals with three different autonomic disorders: with pure autonomic failure (PAF), postural tachycardia syndrome (PoTS), and vasovagal syncope. (b) Blood pressure and heart rate measured continuously with the Portapres II in a normal person and in a patient with a high cervical spinal cord lesion. In the spinal injury patient there is a fall in blood pressure because of impairment of the sympathetic outflow disrupted in the cervical spine. Heart rate rises because of withdrawal of vagal activity in response to the rise in pressure. (c) Blood pressure and heart rate with continuous recordings from the Portapres II in a patient with the mixed (cardioinhibitory and vasodepressor) form of vasovagal syncope. (a) From Mathias CJ (2002) To stand on ones' own legs. *Clinical Medicine* 2: 237–245. (b) From Mathias CJ (2006) Orthostatic hypotension and orthostatic intolerance. In: De Groot LJ, Jameson JL, de Kretser D, et al. (eds.) *Endocrinology*, 5th edn., pp. 2613–2632. Philadelphia: Elsevier. (c) From Mathias CJ (2006) Orthostatic hypotension and orthostatic intolerance. In: De Groot LJ, Jameson JL, de Kretser D, et al. (eds.) *Endocrinology*, 5th edn., pp. 2613–2632. Philadelphia: Elsevier.

head-up postural change observed in orthostatic hypotension due to autonomic failure.

In the syndrome PoTS, also called the neuropathic PoTS, the heart rate rises by over 30 beats min^{-1} during standing or head-up tilt, without postural hypotension. There is no evidence of widespread autonomic failure. However, in some cases there is partial denervation, predominantly affecting the lower limbs. A genetic abnormality, linked to a noradrenaline transporter defect has been described in one family. Psychogenic components, including hyperventilation, may contribute. There is an association with mitral valve prolapse, the chronic fatigue syndromes, physical decompensation, and the joint hypermobility syndrome (Ehlers–Danlos III).

Pressor stimuli Pressor stimuli raise blood pressure by increasing sympathetic efferent activity. They act via the periphery (isometric exercise and cutaneous cold (Figure 3(a)), or directly by cerebral stimulation (sudden noise and mental arithmetic). Isometric exercise is performed by sustained hand-grip using a dynamometer. In the cold pressor test the hand is immersed in ice for 90 s. In both situations the autonomic responses depend on peripheral stimulation but with a central component. Cerebral stimulation is induced by a sudden noise, such as a starting pistol, or mental arithmetic, utilizing subtraction or addition of serial 7 s or 17 s (or more complex) tasks. In normal individuals both blood pressure and heart rate increase.

Pressor stimuli do not elevate blood pressure in central or efferent sympathetic lesions. In complete cervical and high thoracic spinal cord lesions, stimuli above the level of the lesion (cutaneous cold and mental arithmetic) do not raise blood pressure, in contrast to stimuli below the lesion, which can elevate blood pressure markedly by inducing spinal cord sympathetic reflexes (Figure 3(b)). The heart rate falls in response to stimulation activation of baroreflex afferents and an increase in vagal activity. Stimuli that cause these changes include cutaneous stimuli (such as cold), activation of abdominal or pelvic viscera (by urinary bladder or large bowel distension), and skeletal muscle spasms. The cardiovascular changes to such stimulation are an important component of the syndrome of autonomic dysreflexia. This may be mistaken for a hypertensive crisis as occurs with pheochromocytoma tumors, which secrete excessive amounts of catecholamines. Measurement of plasma

Table 5 Drugs, chemicals, poisons, and toxins causing autonomic dysfunction

Decreasing sympathetic activity
Centrally acting
Clonidine
Methyldopa
Moxonidine
Reserpine
Barbiturates
Anesthetics
Peripherally acting
Sympathetic nerve ending (guanethidine, bethanidine)
α-Adrenoceptor blockade (phenoxybenzamine)
β-Adrenoceptor blockade (propranolol)
Increasing sympathetic activity
Amphetamines
Releasing noradrenaline (tyramine)
Uptake blockers (imipramine)
Monoamine oxidase inhibitors (tranylcypromine)
β-Adrenoceptor stimulants (isoprenaline)
Decreasing parasympathetic activity
Antidepressants (imipramine)
Tranquilizers (phenothiazines)
Antidysrhythmics (disopyramide)
Anticholinergics (atropine, probanthine, benztropine)
Toxins (botulinum)
Increasing parasympathetic activity
Cholinomimetics (carbachol, bethanechol, pilocarpine, mushroom poisoning)
Anticholinesterases
Reversible carbamate inhibitors (pyridostigmine, neostigmine)
Organophosphorous inhibitors (parathion, sarin)
Miscellaneous
Alcohol, thiamine (vitamin B_1) deficiency
Vincristine, perhexiline maleate
Thallium, arsenic, mercury
Mercury poisoning (pink disease)
Ciguatera toxicity
Jellyfish and marine animal venoms
First dose of certain drugs (prazosin, captopril)
Withdrawal of chronically used drugs (clonidine, opiates, alcohol)

From Mathias CJ (2003) Disorders of the autonomic nervous system. In: Bradley WG, Daroff RB, Fenichel GM, et al. (eds.) *Neurology in Clinical Practice*, 3rd edn., pp. 2403–2420. Boston, MA: Butterworth-Heinemann.

noradrenaline and adrenaline levels during these episodes excludes the possibility of catecholamine hypersecretion. In autonomic dysreflexia there is a modest rise in plasma noradrenaline but not adrenaline levels, the former reflecting sympathoneural activation.

Heart rate responses to respiratory change The cardiac vagi rapidly respond to respiratory changes, with a rise in heart rate during inspiration and a fall during expiration (sinus arrhythmia) (Figure 3(c)). In normal individuals, deep breathing (6 breaths min^{-1}) and hyperventilation raise heart rate because of vagal withdrawal. This does not occur in autonomic failure affecting cardiac parasympathetic pathways. In diabetes mellitus and alcoholism, where cardiac vagal lesions may occur prior to sympathetic impairment, the heart rate responses to such stimuli often are abnormal.

Valsalva maneuver In the Valsalva maneuver, the blood pressure and heart rate responses to raised intrathoracic pressure provide a measure of the integrity of the baroreflex pathway (Figure 3(d)). The individual blows with an open glottis into a sphygmomanometer and maintains a forced expiratory pressure of up to 40 mmHg for 10 s. The rise in intrathoracic pressure impairs venous return, reduces cardiac output, and lowers blood pressure, with a consequent rise in heart rate and blood pressure recovery by increasing sympathetic neural activity. At the end of the maneuver, when intrathoracic pressure returns to normal, there is a blood pressure overshoot because sympathetic stimulation persists and baroreflex activation reduces the heart rate. In sympathetic vasoconstrictor failure the Valsalva maneuver causes blood pressure to fall without stabilization. Following reduction in intrathoracic pressure, there is no blood pressure overshoot and no compensatory bradycardia.

Food challenge Food releases vasodilatatory gut peptides, causing dilatation in the splanchnic circulation, which normally results in an increase in cardiac output, and constriction of skeletal muscle resistance vessels that prevents blood pressure from falling. These compensatory responses to splanchnic vasodilatation do not occur in autonomic failure and may cause postprandial hypotension even in the supine position (Figure 4(a)). Measuring the blood pressure during head-up tilt before and 45 min after a balanced 300 ml liquid meal of 530 kcal determines if food unmasks or enhances postural hypotension.

Exercise testing Exercise raises blood pressure in normal individuals. In autonomic failure even modest exercise on a bicycle while supine lowers blood pressure and can, on standing postexercise, aggravate postural hypotension (Figure 4(b)). This is due to vasodilatation in exercising muscles, not accompanied by compensatory changes in other vascular beds.

Carotid sinus massage Carotid sinus hypersensitivity is an important cause of falls, especially in the elderly. Carotid sinus massage normally causes minor changes in heart rate with no decrease in blood pressure. In carotid sinus hypersensitivity, bradycardia and hypotension occur especially when massage is performed during head-up tilt (Figure 4(c)).

Table 6 Outline of investigations in autonomic failure

Cardiovascular	
Physiological	Head-up tilt (60°)*
	Standing*, Valsalva maneuver*
	Pressor stimuli – isometric exercise,* cutaneous cold,* mental arithmetic*
	Heart rate responses – deep breathing,* hyperventilation,* standing,* head-up tilt*
	Carotid sinus massage
	Liquid meal challenge, modified exercise testing
Biochemical	Plasma noradrenaline – supine and standing, urinary catecholamines and their metabolites
Pharmacological	Noradrenaline – α-adrenoreceptors (vascular)
	Isoprenaline – β-adrenoreceptors (vascular and cardiac)
	Tyramine – pressor and noradrenaline response
	Atropine – parasympathetic cardiac blockade
	Clonidine – noradrenaline suppression and growth hormone stimulation
Sweating	Thermoregulatory sweat test – increase core temperature by 1 °C and assess sweating
	Modified sweat testing for localized abnormalities
	Gustatory sweating
	Sweat gland response – intradermal acetylcholine, quantitative sudomotor axon reflex test (Q-SART)
	Sympathetic skin response
Gastrointestinal	Barium studies, video-cinefluoroscopy, endoscopy, gastric emptying studies
Respiratory	Laryngoscopy, sleep studies to assess apnea and oxygen desaturation
Renal function and urinary	Day and night urine volumes and sodium/potassium excretion
	Urodynamic studies, intravenous urography, ultrasound examination, urethral or anal sphincter electromyography
Sexual	Penile plethysmography
	Intracavernosal papaverine response
Eye	Pupillary function – physiological and pharmacological
	Schirmer's test for lacrimal secretion

Asterisk denotes inclusion of the screening tests used in our London laboratories.
From Mathias CJ and Bannister R (2001) Investigation of autonomic disorders. In: Mathias CJ and Bannister R (eds.) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System*, 4th edn., pp. 169–195. Oxford: Oxford University Press.

Table 7 Factors influencing orthostatic hypotension

Speed of positional change
Time of day (worse in the morning)
Prolonged recumbency
Warm environment (hot weather, central heating, hot bath)
Raising intrathoracic pressure – micturition, defecation, or coughing
Food and alcohol ingestion
Water ingestion[a]
Physical exertion
Physical maneuvers and positions (bending forward, abdominal compression, leg crossing, squatting, activating calf muscle pump)[b]
Drugs with vasoactive properties (including dopaminergic agents)

[a]This raises blood pressure in autonomic failure.
[b]These maneuvres usually reduce the postural fall in blood pressure, unlike the others.
From Mathias CJ (2003) Autonomic diseases – clinical features and laboratory evaluation. *Journal of Neurology, Neurosurgery & Psychiatry* 74: 31–41.

Table 8 Nonneurogenic causes of postural hypotension

Low intravascular volume	
Blood/plasma loss	Hemorrhage, burns, hemodialysis
Fluid/electrolyte loss	Inadequate intake – anorexia nervosa, vomiting
	Diarrhea – including losses from ileostomy
	Renal/endocrine – salt-losing nephropathy, adrenal insufficiency (Addison's disease), diabetes insipidus, diuretics
Vasodilatation	Drugs – glyceryl trinitrate
	Alcohol
	Heat, pyrexia
	Hyperbradykininism
	Extensive varicose veins
Cardiac impairment	
Myocardial	Myocarditis
Impaired ventricular filling	Atrial myxoma, constrictive pericarditis
Impaired output	Aortic stenosis

From Bannister R and Mathias CJ (2002) Management of postural hypotension. In: Mathias CJ and Bannister R (eds.) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System,* 4th edn., pp. 342–356. Oxford: Oxford University Press.

Other methods of assessing cardiovascular autonomic and baroreceptor reflex function In addition to autonomic screening and allied tests that are used in clinical autonomic laboratories, there are newer

techniques mainly used in clinical research laboratories; these techniques enable continuous noninvasive measurement of cardiac function, peripheral vascular resistance, and blood flow in various regions. Computer and spectral analytical techniques assess spontaneous or induced fluctuations in blood pressure and heart rate, providing a noninvasive measure of the different autonomic systems influencing cardiovascular control.

A variety of techniques determine baroreflex function. Cardiopulmonary baroreceptor afferents can be stimulated by negative pressure to the lower body, induced by placing the individual in a box with an airtight seal, extending up to the midthoracic region, which allows suction to 'unload' the receptors. This normally increases sympathetic neural activity, with a rise in heart rate constriction of resistance vessel, and maintenance of blood pressure. In sympathetic failure, blood pressure rapidly falls during induction of negative pressure.

Carotid sinus afferents may be tested with a cervical collar, whereby elevation or lowering of pressure enables inhibition or stimulation of receptors. Pressor agents (phenylephrine) or vasodilators (glyceryl trinitrate), given either as bolus injections or sublingually, cause transient changes in blood pressure and heart rate, enabling pharmacological assessment of baroreflex sensitivity.

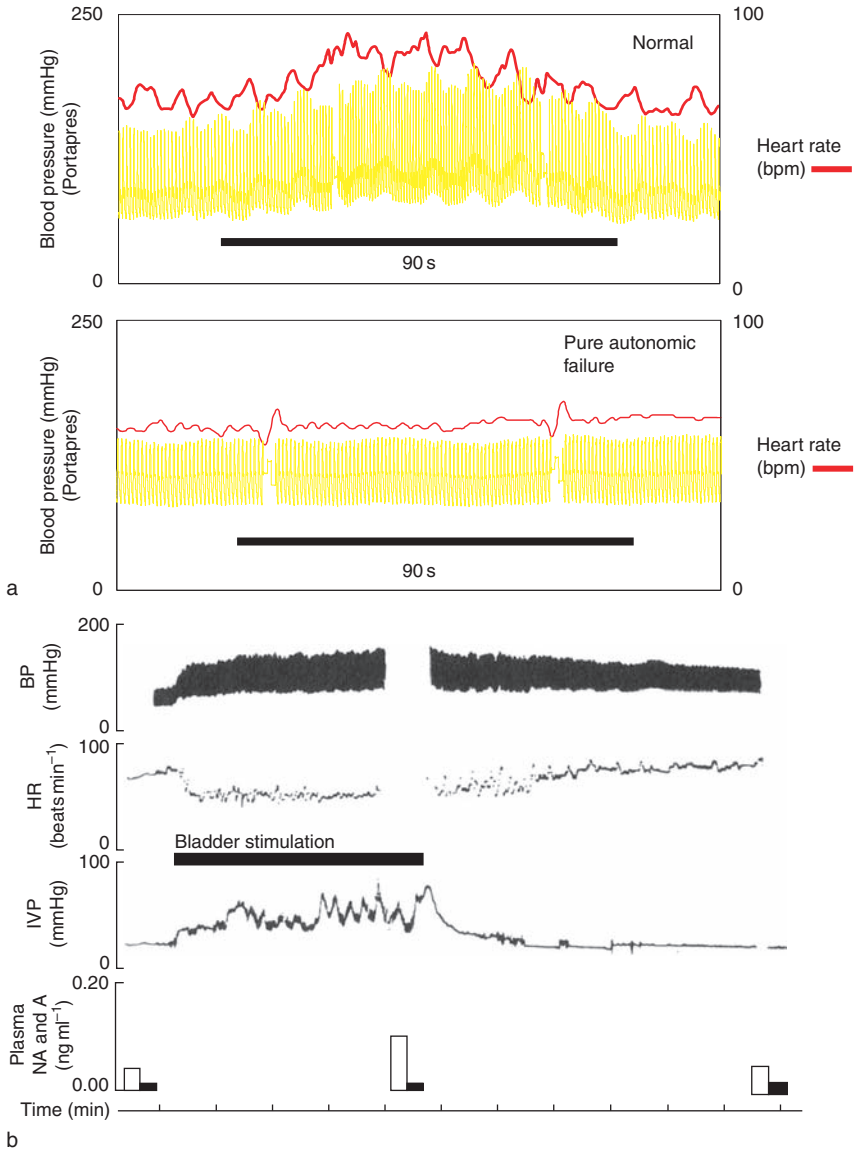

Figure 3 Continued

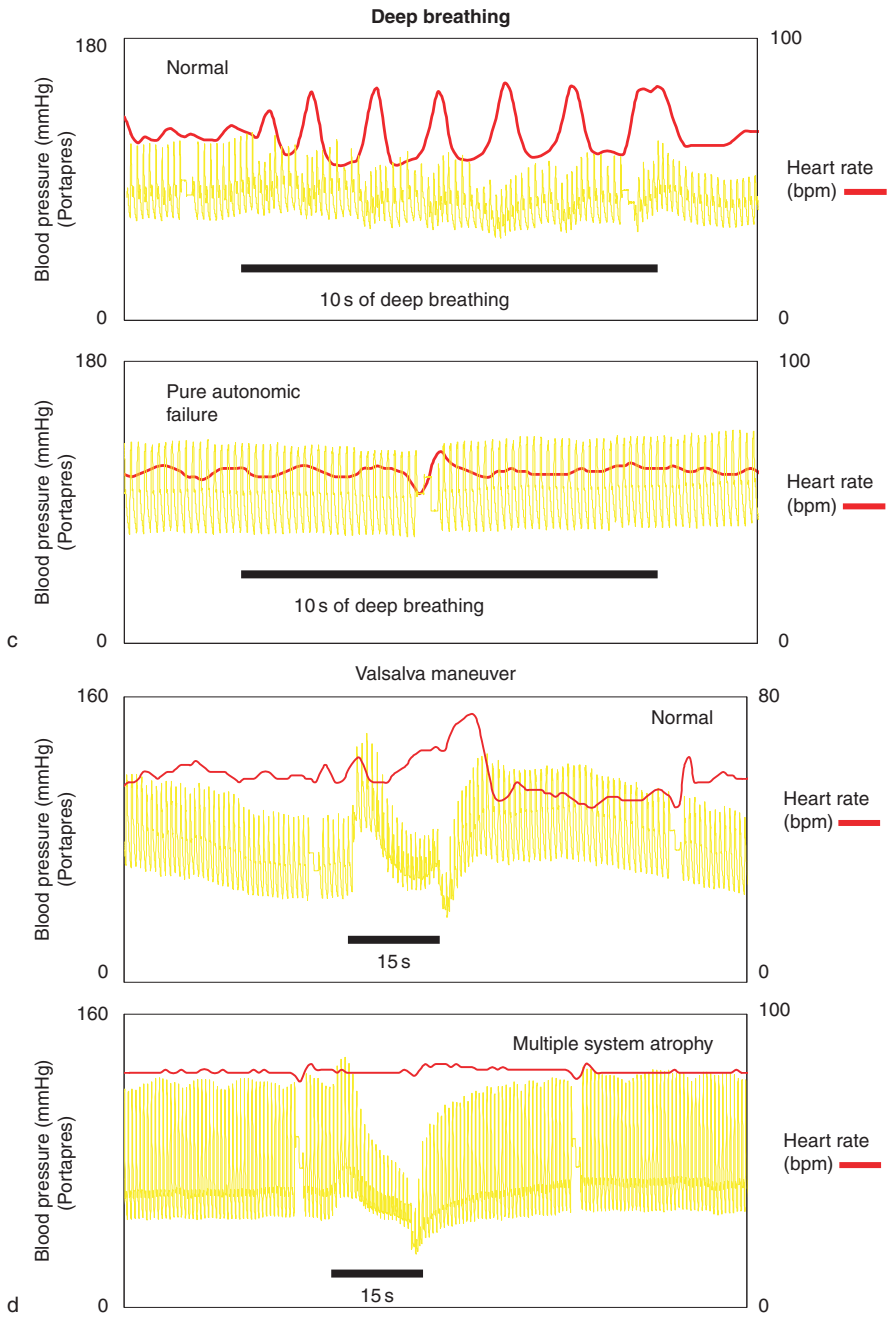

Figure 3 (a) Rise in blood pressure and heart rate (beats per minute, bpm) during cutaneous cold in a normal individual but not in pure autonomic failure. (b) Blood pressure (BP) and heart rate (HR) in a tetraplegic patient before and during bladder stimulation induced by suprapubic percussion of the anterior abdominal wall. This is expected to cause neurogenic bladder contraction. There is a marked rise in BP with a corresponding fall in HR. The breaks in the record indicate where blood was taken for measurement of plasma noradrenaline (NA) and adrenaline (A). There is no change in A (filled rectangles) levels, while NA (open rectangles) levels rise, indicating that the rise in pressure is the result of reflex sympathetic nervous activity occurring via the isolated spinal cord. (c) Rise and fall in heart rate (sinus arrhythmia) in a normal individual and in pure autonomic failure. Normal responses are diminished in autonomic failure affecting the cardiac parasympathetic pathway. (d) Intraarterial recording of blood pressure and heart rate during the Valsalva maneuver in a normal person. There is a fall in blood pressure (with a rise in heart rate), while intrathoracic pressure is elevated. There is a partial recovery in blood pressure with time. Following release of intrathoracic pressure there is a blood pressure overshoot and a corresponding fall in heart rate below basal levels; compare to multiple system atrophy. (a, c, and d) From Mathias CJ and Bannister R (2001) Investigation of autonomic disorders. In: Mathias CJ and Bannister R (eds.) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System*, 4th edn., pp. 169–195. Oxford: Oxford University Press. (b) From Mathias CJ and Frankel HL (1983) Clinical manifestations of malfunctioning mechanisms in tetraplegic man. *Journal of the Autonomic Nervous System* 7: 303–312.

Pharmacological

The responses to agonist or antagonist drugs provide information on the integrity of autonomic pathways and sensitivity of receptors on target organs. They often predict the response to drug therapy.

Noradrenaline predominantly stimulates α-adrenoreceptors and in sympathetic failure results in an enhanced pressor response. This is of greater magnitude in peripheral lesions, as compared to central lesions. Isoprenaline increases heart rate by stimulation of β_1-adrenorceptors and lowers blood pressure through vasodilatatory β_2-adrenoreceptors, if the latter response is unopposed by sympathetic reflexes, blood pressure falls to a greater extent. The drugs are given intravenously, by bolus injection or continuous infusion. Ideally a dose–response curve to three or more doses should be constructed.

The ability of sympathetic nerve endings to release noradrenaline can be tested by using tyramine, which releases noradrenaline from both granules and cytosol in the nerve terminal. The lack of a pressor response indicates postganglionic denervation and absence of noradrenaline (as in dopamine β-hydroxylase (DBH) deficiency). In incomplete lesions, release of even small amounts of noradrenaline may elicit a response in the presence of pressor supersensitivity.

Cardiac parasympathetic efferent pathways are tested by using the muscarinic blocker atropine. An intravenous bolus of 5 $\mu g\,kg^{-1}$ every 2 min (to a maximum of 4 mg) normally raises heart rate to 110–120 beats min^{-1} because of vagal withdrawal. Vagal efferent lesions result in minimal or no change in heart rate. Atropine testing also helps determine whether vasovagal syncope has a predominantly cardioinhibitory component.

The α_2-adrenoceptor agonist clonidine reduces sympathetic activity predominantly through its central effects. Clonidine lowers levels of plasma noradrenaline if elevated due to increased neural activity such as stress. This suppression does not occur with autonomous secretion, as in pheochromocytoma (Figure 5). Another central action of clonidine, mainly through the hypothalamus, is stimulation of growth hormone-releasing hormone that acts on the anterior pituitary to release growth hormone. A rise in serum growth hormone levels occurs in normal individuals and in distal autonomic lesions (such as pure autonomic failure (PAF)), but not in central lesions (such as multiple system atrophy (MSA)) (Figures 5(a) and 5(b)). The clonidine–growth hormone test is used to differentiate these disorders and also in separating MSA from parkinsonian disorders at an early stage when autonomic failure may not be present.

Biochemical

Activation of the sympathetic nervous system, as during head-up tilt, increases synthesis of the neurotransmitter noradrenaline and results in elevation of its plasma levels in normal individuals but not in sympathetic failure. Basal plasma noradrenaline levels are low when there is damage to postganglionic nerve terminals such as PAF, unlike central lesions such as MSA, when they often are within the normal range (Figure 6). Basal levels of plasma adrenaline in normal individuals are much lower (around 50 $pg\,ml^{-1}$), and near the detection limit of many assays. Exercise and hypoglycemia cause a greater degree of adrenal stimulation and predominantly raise plasma adrenaline levels. Levels of plasma noradrenaline and/or adrenaline both at rest and during stimulation often are greatly elevated in pheochromocytoma. In DBH deficiency, a condition characterized by postural hypotension due to sympathetic adrenergic failure, present from birth or early childhood, plasma noradrenaline and adrenaline levels are undetectable while plasma dopamine levels are elevated. The enzyme DBH is undetectable in plasma and tissue.

Urinary measurements of catecholamines and their metabolites provide a useful measure of secretion over a period, and are of value in certain situations, as in pheochromocytoma, in which random plasma measurements may miss intermittent secretion. In DBH deficiency, metabolites of noradrenaline and adrenaline (vanillylmandelic acid, normetadrenaline, and metadrenaline) are extremely low, whereas dopamine metabolites (homovanillic acid and 3-methoxytyramine) are normal or elevated.

Noradrenaline spillover techniques, although invasive, provide information on sympathetic activity in the whole body and also regionally in heart, splanchnic, renal, and brain circulations. Radionuclide ([^{123}I] *meta*-iodobenzylguanidine) and positron emission tomography (PET) scanning (using 6-[^{18}F]fluorodopamine) allow imaging of sympathetic innervation of cardiac tissue. Some of these research techniques are applied in the clinical setting.

Sudomotor System and Thermoregulation

Sweat glands influence thermoregulation and are innervated by sympathetic cholinergic nerves. In the thermoregulatory sweat test, central temperature is elevated by 1 °C and results in widespread activation of eccrine glands. Sweating is detected by dyes such as quinazarine or Ponso red (which turns from pale pink to bright red on exposure to moisture) (Figure 7(a)), or starch iodine (which turns dark brown). In

autonomic failure, sweating often is impaired or absent. The sweat glands may be directly tested by applying ACh intradermally or by iontophoresis. This results in sweating via the axon reflex. Sudorometers measure sweat production in localized areas and with appropriate studies help differentiate preganglionic from postganglionic sympathetic lesions.

Local abnormalities of sudomotor function due to denervation and anhidrosis, as in leprosy, can be detected by using quinazarine powder covered over

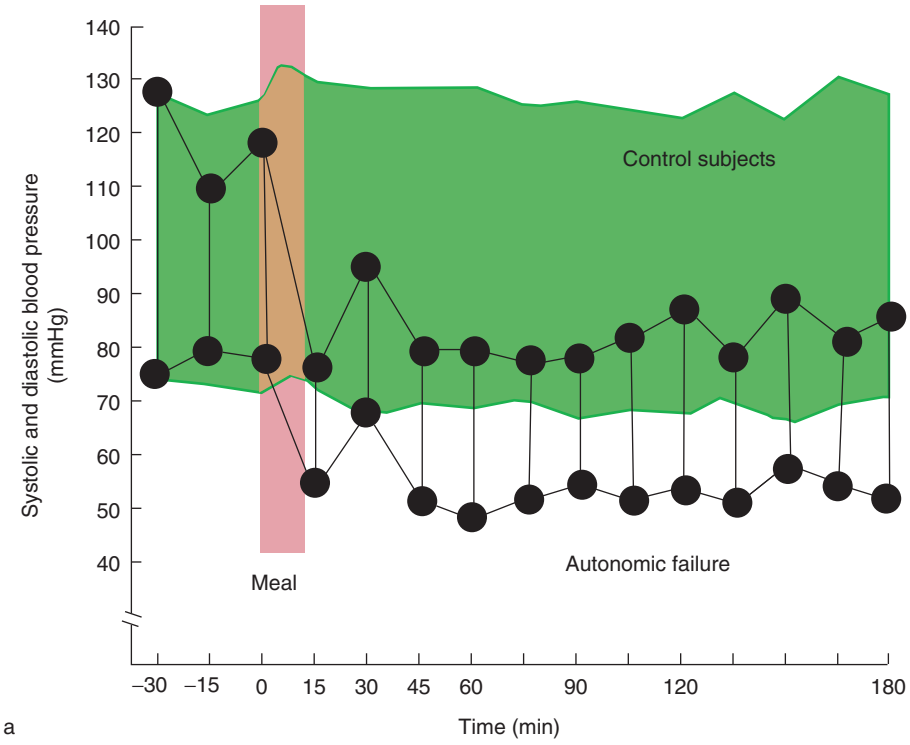

a

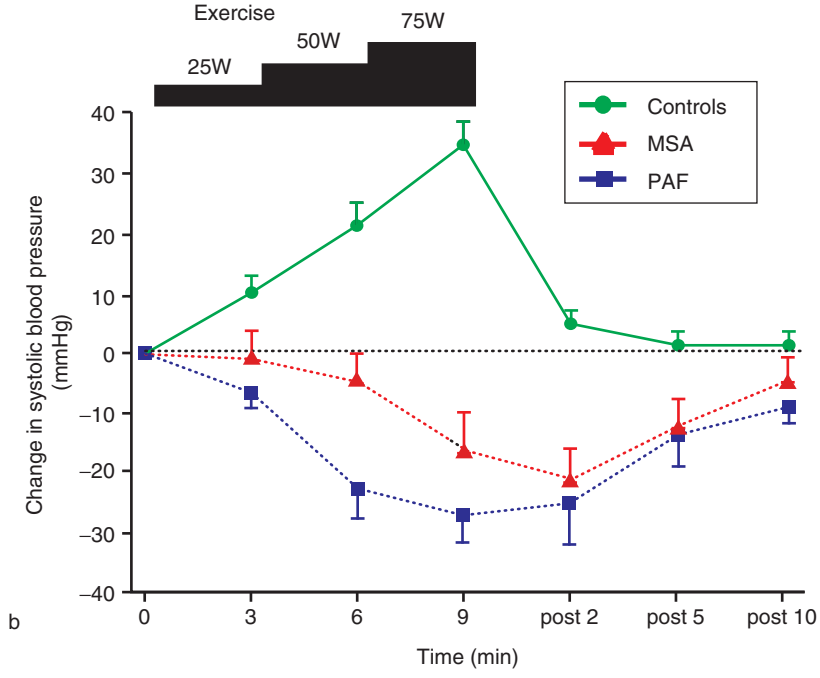

b

Figure 4 Continued

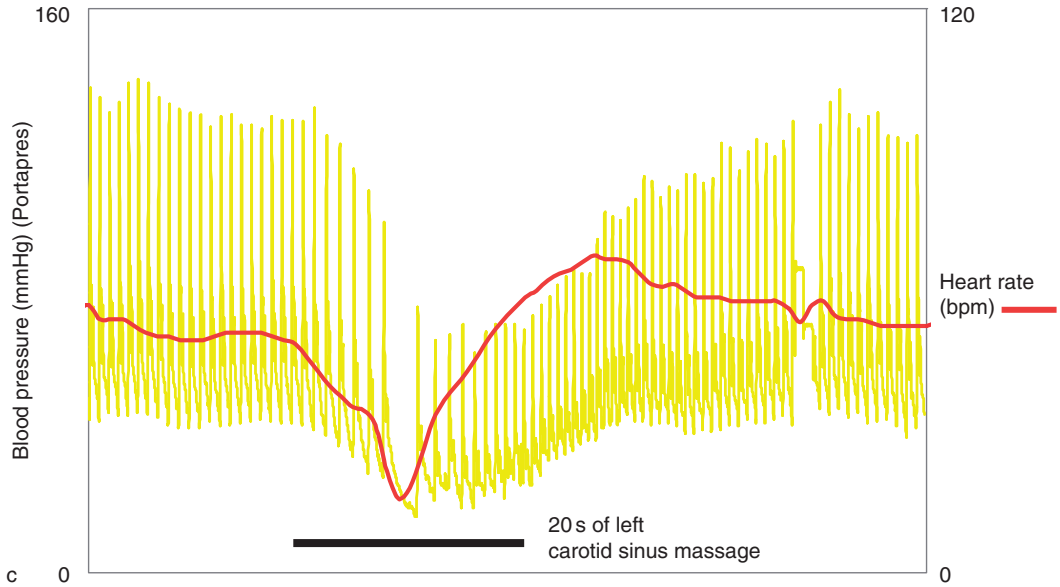

Figure 4 (a) Changes in blood pressure before and after a standard meal in a group of normal persons (controls; green area) and in a patient with autonomic failure (black circles), while in the supine and horizontal position. Bars indicate ± standard error of mean. In the normal persons there is no change in blood pressure after the meal. In the patient with autonomic failure there is a marked fall in blood pressure soon after food ingestion, with levels falling to around 80/50 mmHg and remaining low for 3 h, even in the supine position. (b) Changes in systolic blood pressure during horizontal bicycle exercise at three incremental levels (and on standing, postexercise) in normal persons (controls) and in patients with multiple system atrophy (MSA) and pure autonomic failure (PAF). In both MSA and PAF, unlike controls, there is a fall in blood pressure. (c) Continuous blood pressure and heart rate measured noninvasively (by Portapres) in a patient with falls of unknown etiology. Left carotid sinus massage caused a drop in both heart rate and blood pressure. The findings indicate the mixed (cardioinhibitory and vasodepressor) form of carotid sinus hypersensitivity. (a) From Mathias CJ and Bannister R (2001) Postprandial hypotension in autonomic disorders. In: Mathias CJ and Bannister R (eds.) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System*, 4th edn., pp. 283–295. Oxford: Oxford University Press. (b) From Smith GDP, Watson LP, Pavitt DV, et al. (1995) Abnormal cardiovascular and catecholamine responses to supine exercise in human subjects with sympathetic dysfunction. *Journal of Physiology (London)* 485, 255–265. (c) From Mathias CJ and Bannister R (2001) Investigation of autonomic disorders. In: Mathias CJ and Bannister R (eds.) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System*, 4th edn., pp. 169–195. Oxford: Oxford University Press.

with transparent adhesive tape. Gustatory sweating (or Frey's syndrome), which mainly affects the head and neck, may be severe and socially embarrassing and occurs after ingestion of spicy or acidic foods. It occurs in diabetes mellitus or after surgery to the parotid gland. This is because nerve damage is followed by neural regeneration, with aberrant reinnervation of parasympathetic fibers (to the salivary glands with sympathetic cholinergic fibers that supply the face).

Denervation of sweat glands causes patchy anhidrosis and may cause troublesome compensatory hyperhidrosis in innervated areas. This occurs in the Holmes–Adie syndrome associated with progressive anhidrosis (Ross' syndrome) and after percutaneous transthoracic endoscopic sympathectomy performed for palmar, axillary, and facial hyperhidrosis.

Measurements of electrical potentials from electrodes on the foot and hand record sympathetic cholinergic activity to sweat glands (sympathetic skin response (SSR)). Physiological (inspiratory gasps, loud noise, or touch) or electrical (median nerve) stimulations induce the SSR. In distal sympathetic lesions caused by PAF and pure cholinergic dysautonomia (Figure 7(b)), the SSR is absent. The SSR above and below the level of the lesion may provide a sensitive marker of the integrity of descending sudomotor pathways in spinal cord injuries.

Hypothalamic lesions and sympathetic failure result in impaired thermoregulation. Hyperpyrexia may occur because of the inability to appropriately sweat and vasodilate in the periphery. When exposed to radiant heat, body temperature rises without change in peripheral blood flow or sweating. The reverse, hypothermia, occurs as sympathetic vasoconstrictor failure reduces the capacity to retain heat, especially when muscle paralysis, as in tetraplegics, impairs heat production because of reduced shivering thermogenesis. Hypothermia may be missed unless a low-reading rectal thermometer, or tympanic membrane measurement, is used.

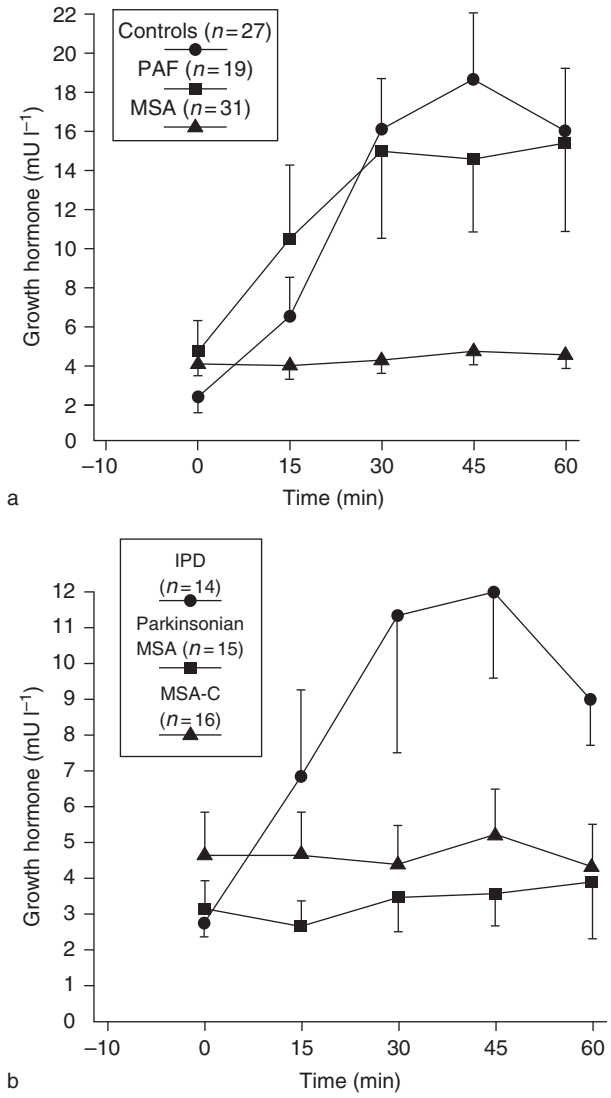

Figure 5 (a) Serum growth hormone concentrations before (0) and at 15 min intervals for 60 min after clonidine (2 μg/kg/min) in normal persons (controls) and in patients with pure autonomic failure (PAF) and multiple system atrophy (MSA). Growth hormone concentrations rise in controls and in patients with PAF with a peripheral lesion; there is no rise in patients with MSA with a central lesion. (b) Lack of serum growth hormone response to clonidine in the two forms of MSA, the cerebellar form (MSA-C) and the parkinsonian form, in contrast to patients with idiopathic Parkinson's disease (IPD) with no autonomic deficit, in whom there is a significant rise in growth hormone levels. (b) From Kimber JR, Watson L, and Mathias CJ (1997) Distinction of idiopathic Parkinson's disease from multiple system atrophy by stimulation of growth hormone release with clonidine. *Lancet* 349: 1877–1881.

Intraneural Recordings of Sympathetic Activity

Skin and muscle sympathetic nerve activity can be recorded using a tungsten microelectrode inserted into a peripheral nerve. Skin sympathetic nerve activity is affected by thermal stimuli, with cold increasing and heat decreasing activity. It is not time-locked to blood pressure changes during the cardiac cycle, which is characteristic of muscle sympathetic nerve activity. In tetraplegia, basal sympathetic discharge is diminished, consistent with low basal plasma noradrenaline levels, while during autonomic dysreflexia there is only a modest and transient increase in sympathetic discharge, despite the disproportionately greater and prolonged elevation in blood pressure. The latter probably is due to decentralization pressor supersensitivity and impaired baroreflex activity. In the Guillain–Barré syndrome, episodes of hypertension and tachycardia are closely correlated with increased sympathetic discharge. In neurally mediated syncope, sympathetic discharge diminishes, causing vasodilatation and hypotension.

The disadvantage of sympathetic microneurography is that it is an invasive technique involving measurements in a restricted region of the body. It is ideal when sympathetic activation is expected, but not when sympathoneural activity is low or absent, as in autonomic failure.

The Eye

Pupillary responses can be tested physiologically or pharmacologically. In Horner's syndrome, infrared television pupillometry indicates reduced pupillary dilatation in the dark. There is a constrictor response to light as occurs normally, but with a redilatation lag in the latter part of the recovery phase. With parasympathetic impairment, as in diabetes mellitus, the speed of the light reflex is reduced and its latency prolonged.

Drugs that stimulate or block autonomic papillary responses can be instilled onto the conjunctiva. Cocaine (4%) blocks the neuronal uptake of noradrenaline and normally causes pupillary dilatation; in sympathetic failure there is a diminished response. Hydroxyamphetamine (0.5%) releases noradrenaline and causes pupillary dilatation in preganglionic but not in postganglionic lesions, thus helping localize the site of lesion. In sympathetic denervation there is dilator supersensitivity of the pupil to 1:1000 adrenaline solution. In parasympathetic denervation (Holmes–Adie pupil) there is constrictor supersensitivity, unlike the normal pupil (Figure 8).

Lacrimal secretion measured with a special blotting paper (Schirmer's test) is diminished in parasympathetic failure. The reverse, excessive lacrimation, is found in 'crocodile tears' syndrome (gustolacrimal reflex, or Bogorod's syndrome) during eating.

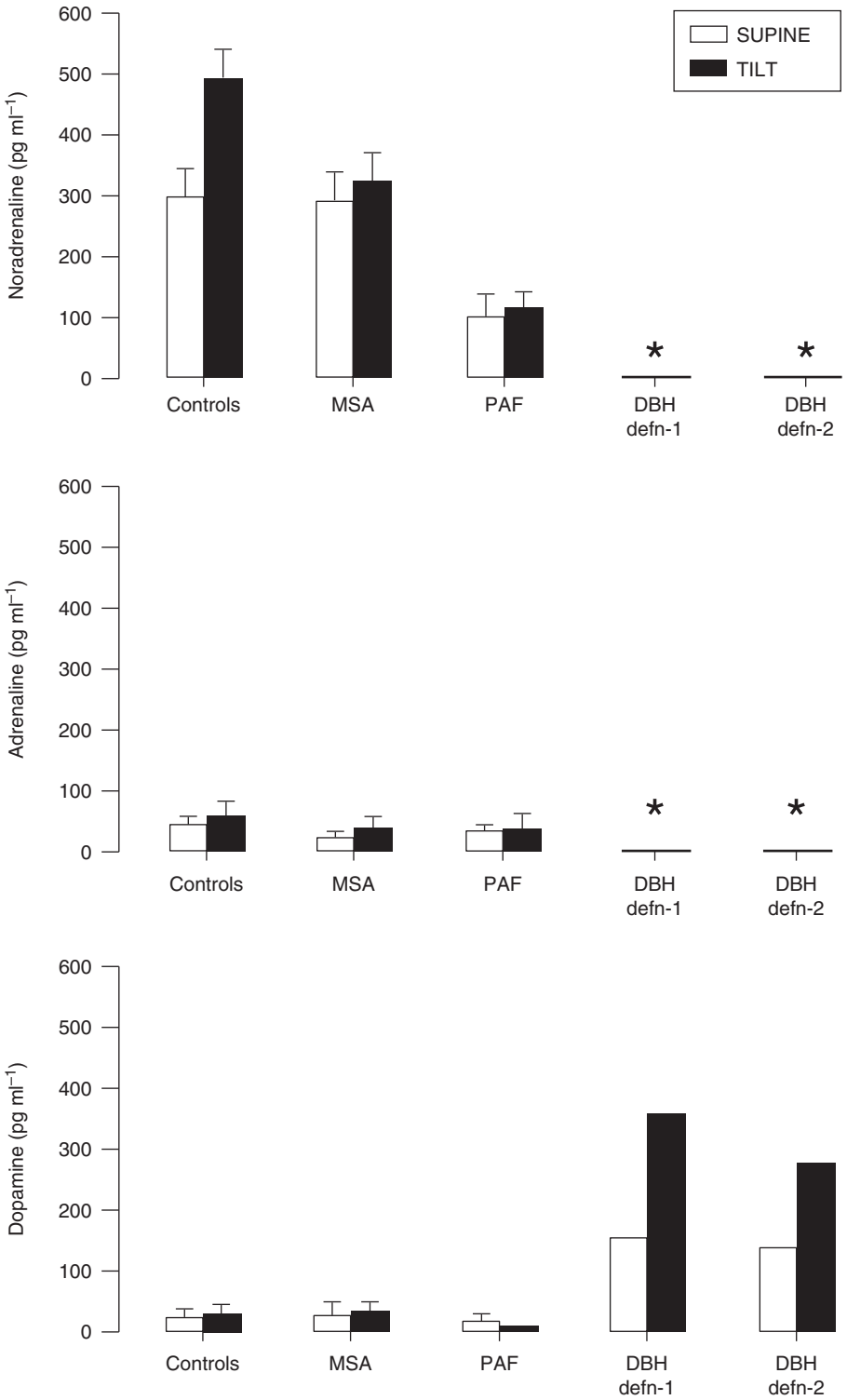

Figure 6 Basal levels of plasma noradrenaline, adrenaline, and dopamine measured by high-pressure liquid chromatography with electrochemical detection, in normal persons (controls), in patients with multiple system atrophy (MSA) and pure autonomic failure (PAF), and in two siblings with a deficiency of dopamine β-hydroxylase (DBH). Levels of noradrenaline are low in PAF and undetectable in DBH deficiency. In the DBH pair, plasma dopamine levels are clearly elevated, indicating a block at this particular stage in the formation of noradrenaline. The asterisks indicate levels of sensitivity, which are $<5\,\mathrm{pg\,ml^{-1}}$ for noradrenaline and adrenaline and $<20\,\mathrm{pg\,ml^{-1}}$ for dopamine. From Mathias CJ, Bannister R, Cortelli P, et al. (1990) Clinical autonomic and therapeutic observations in two siblings with postural hypotension and sympathetic failure due to an inability to synthesize noradrenaline from dopamine because of a deficiency of dopamine beta hydroxylase. *Quarterly Journal of Medicine* 278: 617–633.

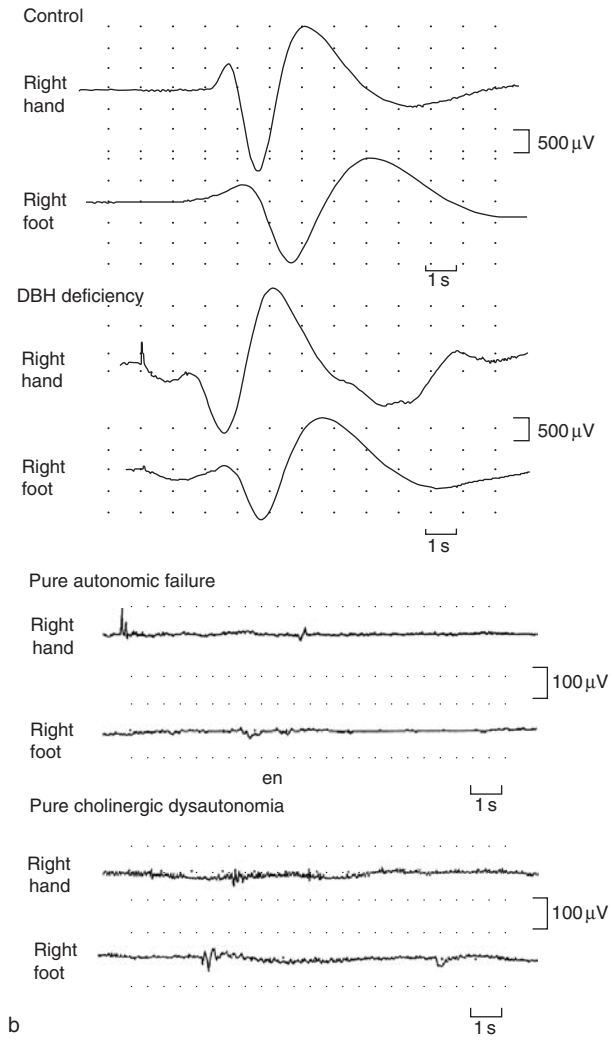

Figure 7 (a) Segmental hyperhidrosis as a presenting feature in a patient in whom there were large areas without sweating. He had the Holmes–Adie syndrome with Ross's variant. (b) The sympathetic skin response (in microvolts) from the right hand and right foot of a normal individual (control) and of a patient with dopamine β-hydroxylase (DBH) deficiency (upper panel). The sympathetic skin response could not be recorded in two patients, one with pure autonomic failure and the other with pure cholinergic dysautonomia (lower panel). (a) From Mathias CJ (1998) Autonomic disorders. In: Bogousslavsky J and Fisher M (eds.) *Textbook of Neurology*, pp. 519–545. Newton, MA: Butterworth-Heinemann. (b) From Magnifico F, Misra VP, Murray NMF, et al. (1998) The sympathetic skin response in peripheral autonomic failure – evaluation in pure autonomic failure, pure cholinergic dysautonomia and dopamine-β-hydroxylase deficiency. *Clinical Autonomic Research* 8:133–138.

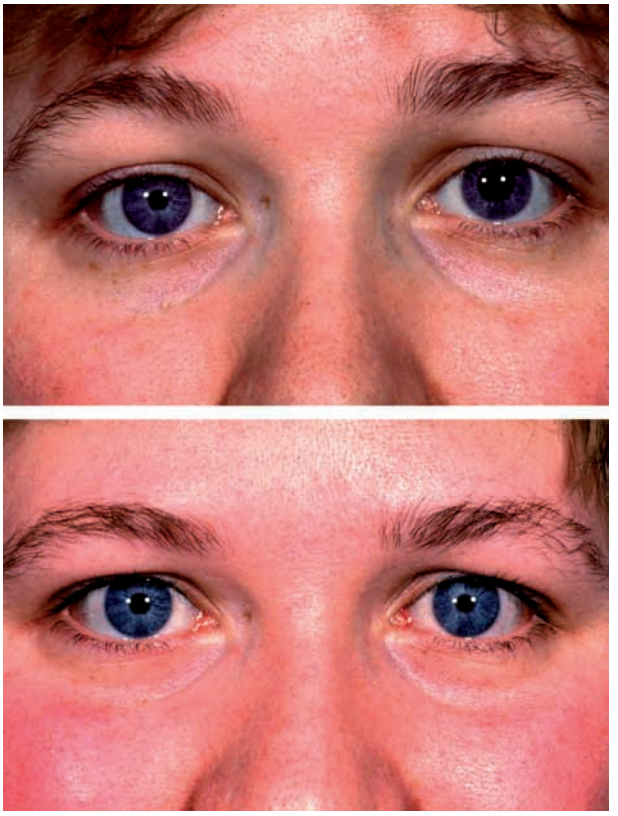

Figure 8 Pupillary dilatation in Holmes–Adie syndrome and its response to pilocarpine. Upper panel: There is a diminished light response, causing a larger left pupil. Lower panel: The response of the left pupil to dilute pilocarpine is greater than that of the normal right pupil. From Mathias CJ (1998) Autonomic disorders. In: Bogousslavsky J and Fisher M (eds.) *Textbook of Neurology*, pp. 519–545. Newton, MA: Butterworth-Heinemann.

Gastrointestinal Tract

Swallowing often is affected in MSA and can be detected by video-cinefluoroscopy. A barium swallow determines esophageal involvement, as in Chagas' disease, a trypanasomal infection that causes dilatation of the lower esophagus. Measurement of esophageal pressure is valuable in achalasia. Upper gastrointestinal endoscopy enables biopsy of tissue, which is diagnostic in systemic amyloidosis. Gastric motility is measured using a technetium-labeled meal and progress of the gamma emitter is tracked by a scintiscan camera. Delayed gastric emptying may complicate amyloidosis and diabetes mellitus.

In diabetes mellitus, diarrhea may be a problem. The reasons include motility abnormalities and bacterial overgrowth in dilated, poorly functioning segments of gut. The response to broad-spectrum antibiotics provides a therapeutic diagnosis. Jejunal aspiration, bacterial culture, and the [^{14}C]glycocholate test may be needed. Gastrointestinal motility and activity can be continuously monitored using indwelling tubes or capsules with telemetric devices.

Large bowel dysfunction and constipation are common in autonomic failure, and a barium enema or colonoscopy may be needed. A rectal biopsy, with Congo red staining, provides definitive evidence in amyloidosis.

Renal Function and Urinary Tract

Nocturnal polyuria is common in autonomic failure and may cause an overnight weight loss of over a kilogram, with a reduction in extracellular fluid volume, which aggravates morning postural hypotension. Day and night measurements of urine volume and sodium excretion provide measurement of diurnal changes. With urinary bladder and sphincter involvement, frequency of micturition and incontinence may occur. An intravenous urogram and/or ultrasound examination of the urinary tract may be needed. Urodynamic measurements provide information as to the nature of bladder dysfunction such as detrusor areflexia or hyperreflexia. Urethral or anal sphincter electromyography is abnormal in MSA but is unaffected in Parkinson's disease.

Sexual Function

In autonomic failure, erectile and ejaculatory sexual dysfunctions are common in the male. Differentiation from psychogenic impotence can be difficult in the

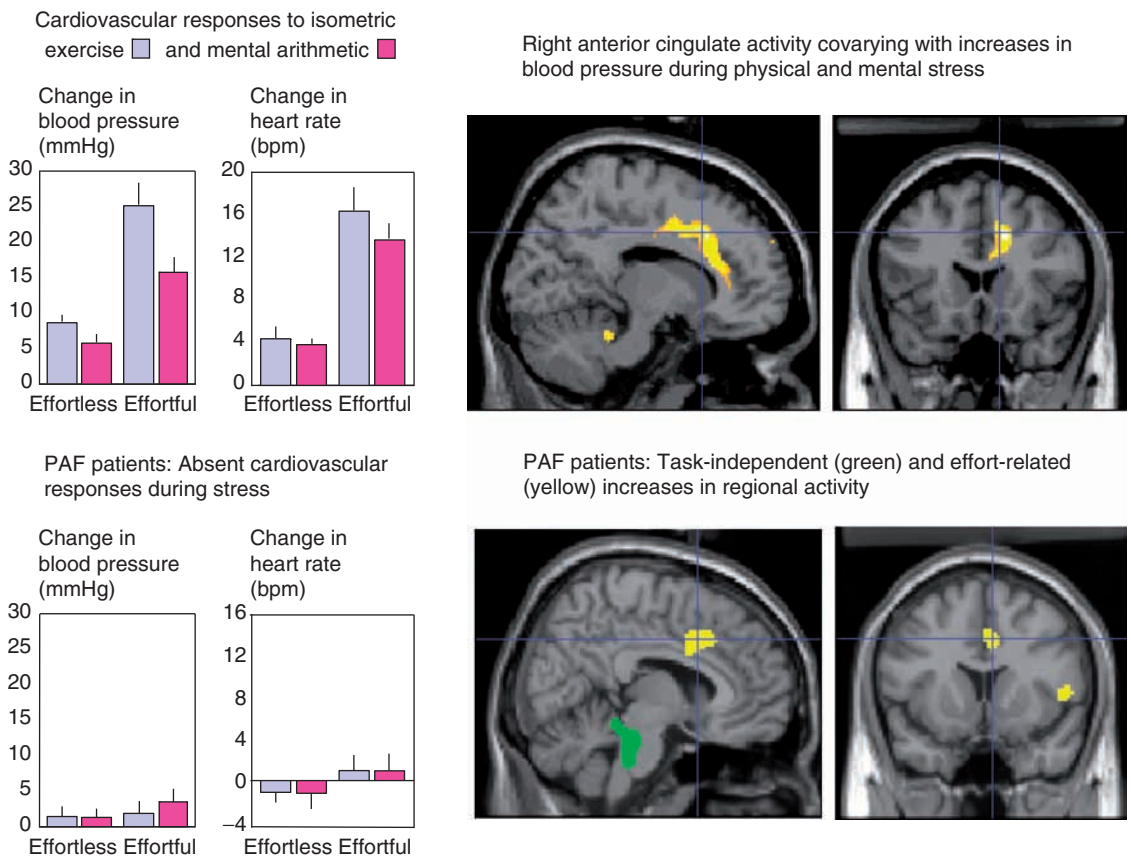

Figure 9 Simultaneous neuroimaging and physiological measurement during cardiovascular autonomic activation in normal persons and in pure autonomic failure (PAF) patients. Changes in blood pressure and heart rate during control (effortless) tasks and effortful isometric exercise and mental arithmetic are shown (top left, normal persons; lower left, PAF patients). Alongside are images showing changes in regional cerebral (mainly right anterior cingulate) activity. Normal persons show increased activity in the right anterior cingulate area and positive covariance with mean arterial blood pressure during isometric exercise and mental arithmetic. In PAF there is an additional increase in activity in the right insula and brain stem (pontine) areas despite no change in blood pressure. The cerebral autonomic centers in PAF are preserved, and the additional areas activated indicate attempts to increase peripheral autonomic activity, which is not possible because of their postganglionic sympathetic lesion. From Mathias CJ (2002) To stand on ones' own legs. *Clinical Medicine* 2: 237–245.

absence of other autonomic features, as both respond well to intracorporeal injections of papaverine. In diabetes mellitus, vascular factors may cause or contribute to impotence.

Respiratory Tract

Nocturnal stridor due to denervation of abductors of the vocal cords may occur in the later stages of MSA. Hypoxia results from laryngeal abductor paresis and/or brain stem dysfunction. The former is detected by laryngoscopy. Blood gas monitoring during sleep is necessary to determine if significant oxygen desaturation warrants procedures such as continuous positive-pressure ventilation or a tracheostomy.

Miscellaneous Investigations

Additional nonautonomic investigations may be needed to determine the etiology of the autonomic disorder and to distinguish between the different groups with primary autonomic failure. Computed tomography (CT) or magnetic resonance imaging (MRI) of the brain may indicate cortical, cerebellar, and/or brain stem atrophy in MSA. Advances in neuroimaging (especially PET and event-related functional MRI), as used in clinical research (**Figure 9**) increasingly are being applied to determine pathophysiological mechanisms, and also aid diagnosis. The electroencephalogram (EEG) is of value when postural hypotension is associated with, or precipitates, seizures and in separating epilepsy from neurally mediated syncope, which results in hypoxia-induced convulsions or myoclonic jerks. If peripheral nerve involvement is suspected, electrophysiological studies are indicated, together with a sural nerve biopsy. A rectal or renal biopsy (especially if there is proteinuria) may confirm the diagnosis of amyloidosis. In amyloidosis, gene mapping to determine mutations is important, especially in familial amyloid polyneuropathy and also in sporadic amyloidosis. Cardiac causes of syncope are common, especially in the elderly, and may necessitate a 24 h ambulatory electrocardiogram (ECG) recording.

Concluding Remarks

Autonomic disorders involve a number of systems and complicate many diseases. Clinical autonomic testing, utilizing a combination of physiological, biochemical, and molecular biological methods along with modern neuroimaging techniques, have advanced our ability to investigate autonomic disorders. Autonomic investigation often provides important information for the clinician and patient, and aids diagnosis, prognosis, and appropriate treatment.

See also: Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Central Cardiovascular Control; Autonomic Nervous System: Central Urogenital Control; Autonomic Nervous System: Central Thermoregulatory Control; Autonomic Nervous System: Central Control of the Gastrointestinal Tract; Autonomic Nervous System: Central Respiratory Control; Autonomic Dysfunction: Drug-Induced; Autonomic Nervous System; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Appenzeller O and Oribe E (eds.) (1997) *The Autonomic Nervous System. An Introduction to Basic and Clinical Concepts,* 5th edn. Amsterdam: Elsevier Medical Press.

Cariga P, Catley M, Savic G, et al. (2002) Organisation of the sympathetic skin response in spinal cord injury. *Journal of Neurology, Neurosurgery & Psychiatry* 72: 356–360.

Critchley HD, Corfield DR, Chandler MP, et al. (2000) Cerebral correlates of autonomic cardiovascular arousal: A functional neuroimaging investigation in humans. *Journal of Physiology* 523: 259–270.

Critchley HD, Mathias CJ, and Dolan RJ (2001) Neuroanatomical basis for first- and second-order representations of bodily states. *Nature Neuroscience* 4: 207–212.

Johnson RH, Lambie DG, and Spalding JMK (1984) *Neurocardiology. The Interrelationship between Dysfunction in the Nervous and Cardiovascular Systems.* London: W.B. Saunders.

Low PA (ed.) (1997) *Clinical Autonomic Disorders,* 2nd edn., Philadelphia: Lippincott, Raven.

Mathias CJ (1991) Disorders of the autonomic nervous system. In: Bradley WG, Daroff RB, Fenichel GM, et al. (eds.) *Neurology in Clinical Practice,* 2nd edn., pp. 1661–1685. Boston, MA: Butterworth Publishers.

Mathias CJ (2000) Autonomic dysfunction. In: Grimley-Evans J, Franklin Williams T, Lynn Beattie B, et al. (eds.) *Oxford Textbook of Geriatric Medicine,* 2nd edn., pp. 833–852. Oxford: Oxford University Press.

Mathias CJ (2002) To stand on ones' own legs. *Clinical Medicine* 2: 237–245.

Mathias CJ (2003) Autonomic diseases – clinical features and laboratory evaluation. *Journal of Neurology, Neurosurgery & Psychiatry* 74: 31–41.

Mathias CJ (2003) Autonomic diseases – management. *Journal of Neurology, Neurosurgery & Psychiatry* 74: 42–47.

Mathias CJ and Bannister R (eds.) (2002) *Autonomic Failure: A Textbook of Clinical Disorders of the Autonomic Nervous System,* 4th edn., Oxford: Oxford University Press.

Mathias CJ, Deguchi K, and Schatz I (2001) Observations on recurrent syncope and presyncope in 641 patients. *Lancet* 357: 348–353.

Mathias CJ and Williams AC (1994) The Shy–Drager syndrome (and multiple system atrophy). In: Calne DB (ed.) *Neurodegenerative Diseases*, pp. 742–767. London: Saunders Scientific Publications.

Mathias CJ and Young TM (2004) Water drinking in the management of orthostatic intolerance due to orthostatic hypotension, vasovagal syncope and the postural tachycardia syndrome. *European Journal of Neurology* 11: 613–619.

Autonomic Nervous System: Gastrointestinal Control

J C Bornstein, University of Melbourne, Parkville, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The gastrointestinal system is the most complex organ in the body, outside the brain. Its function is digestion and absorption of nutrients. This requires a wide variety of enabling behaviors, including movement of food via muscular contractions (motility), secretion of water and salt to provide an appropriate medium, and secretion of digestive enzymes. Freed nutrients are separated from waste by absorption through the mucosa. Water and salt are then reabsorbed to maintain whole-body water and electrolyte balance. Blood flow through the intestinal wall is tightly regulated both for these purposes and to ensure effective transport of nutrients. Finally, waste is eliminated via defecation. All these processes are regulated by the autonomic nervous system, although the details vary among regions. This article focuses on the small intestine, with reference to the colon where appropriate.

Neural control of intestinal functions operates at multiple levels (Figure 1). The major control is exerted locally by a network of intrinsic neurons, the enteric nervous system (ENS). This is the largest and most complex part of the autonomic nervous system, containing 30 000 000–100 000 000 neurons in humans. The ENS produces complex behaviors in response to physiological stimuli independently of the central nervous system (CNS) and so includes sensory neurons, motor neurons, and integrative neurons (interneurons). Other levels of neural control include coordination of the behavior of different gut regions by reflex pathways running through the CNS and an intermediate-level pathway that passes through prevertebral sympathetic ganglia (intestino-intestinal reflexes).

Intestinal behavior depends critically on whether the animal is fed or fasted. Even the sight or smell of food converts intestinal behavior from a fasted to a fed state. This clearly involves the brain, but both fed and fasted states depend on the ENS for their specific behaviors. In the fed state, intestinal movements fall into three distinct patterns. The most common, segmentation, involves localized constrictions repeated rhythmically at specific sites, where they remain stationary or propagate only short distances. Segmentation mixes chyme with digestive enzymes and with water secreted into the lumen from the blood. Occasionally, segmentation is replaced by a constriction that propagates anally, forcing the content ahead of it (peristalsis). Orally propagating constrictions occur; but are rarer than peristalsis. When food is present, water and salt are secreted into the lumen of the duodenum and jejunum. The water and salt are then reabsorbed in the ileum and colon. Both secretion and absorption require increased blood flow to the intestine.

Once digestion and absorption have been completed, the fasted state appears. This manifests as the interdigestive migrating motor complex (MMC) in the small intestine. MMCs have three distinct phases that appear in the antrum or upper duodenum and propagate slowly along the small intestine to the ileocolonic junction, taking over an hour to traverse the human gut. Phase I is a period of quiescence with no muscle contractions; this slowly converts to a period of irregular activity (phase II), which then converts to a strong wave of contraction lasting up to 5 min (phase III). In herbivores, phase III contractions interrupt the fed-state behavior and may be important for intestinal propulsion. In the fasted state, phase III contractions clear mucosal debris and accumulated intestinal bacteria. All three phases of interdigestive MMCs depend on enteric neural activity.

The role of the colon is distinctly different: reabsorption of water and storage and excretion of waste. It also regulates small intestinal transit of digesta: colonic distension or bile acids in the colonic lumen cause significant slowing of propulsion along the jejunum via a reflex pathway running through the prevertebral sympathetic ganglia.

Structure of the ENS

Locations of Neurons and Effectors

The basic structure of the gastrointestinal tract (Figure 2) is maintained along its length. The intestinal tube is lined by the mucosal epithelium and associated connective tissues and blood vessels. Adjacent to the mucosa is a layer of smooth muscle, the muscularis mucosae. This is virtually nonexistent in small animals, but is substantial in larger species, including humans. The next layer is the submucosa, a layer of connective tissue within which are embedded arterioles, venules, and the cell bodies and processes of neurons grouped into a ganglionated submucous plexus. The major muscle layer, the circular muscle, lies adjacent to the submucosa. This is the thickest muscle layer and its smooth muscle cells are oriented

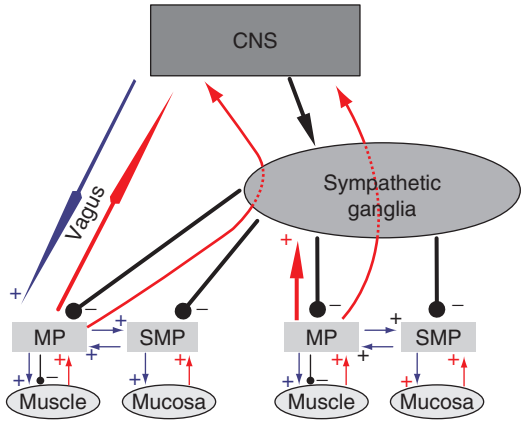

Figure 1 The hierarchy of control of the gastrointestinal tract and its interconnections. Excitatory connections are illustrated by the blue arrows, sensory pathways are indicated by the red arrows, and inhibitory pathways are indicated by the black 'clubs.' The major control of gastrointestinal movements and secretion is exerted within the enteric nervous system. The muscle is directly regulated by the myenteric plexus (MP) and the mucosa is regulated from the submucous plexus (SMP). There are strong excitatory inputs to the SMP from the MP and weaker excitatory connections in the opposite direction. Activity of enteric neurons is modulated by input from the central nervous system (CNS) via the vagus (excitatory) and sympathetic nerves (inhibitory). Enteric ganglia receive sensory input from both muscle and mucosa. Sensory information reaches the CNS via vagal afferents and visceral afferents running with the sympathetic innervation. Intestinofugal neurons, most prominent in the large bowel, send axons to the prevertebral sympathetic ganglia via the sympathetic nerve trunks, activating inhibitory reflexes that regulate activity in the colon and also in the upper small intestine.

parallel to the gut's circumference, so their contractions constrict the lumen. The next layer outward, the myenteric plexus, contains many neurons and their processes in an ordered network of nerve trunks and ganglia. The final muscle layer is the longitudinal muscle, whose smooth muscle cells run parallel to the long axis of the intestine. This is bounded by the serosa, the outer wall of the intestine. The blood supply, with the extrinsic nerves that run with it, enters the intestine via the mesentery and ramifies within the intestinal wall.

The myenteric plexus extends from the upper esophageal sphincter to the anus. The submucous plexus only becomes ganglionated at the pyloric sphincter of the stomach. The submucous plexus can sometimes be subdivided into an inner and outer plexus, each containing different types of neurons.

Effector Targets of Enteric Neurons

The myenteric plexus provides most of the innervation of the muscle coat. In smaller mammals, this comes exclusively from myenteric neurons, but submucosal neurons contribute in larger animals. Smooth muscle cells are not the major cellular target of neurons innervating the circular muscle layer; rather, they contact a different cell type, the interstitial cells of Cajal (ICCs) (black cells in the CM in Figure 2). These are organized into a functional network via gap junctions with each other and also form gap junctions with circular muscle cells. Transmitters released from motor neurons apparently alter muscle behavior indirectly via actions on ICCs. This contrasts with the innervation of the longitudinal muscle, which involves direct connections between nerve terminals and muscle cells. A second population of ICCs, adjacent to the myenteric plexus (Figure 2), consists of pacemaker cells generating rhythmic depolarizations – slow waves – in the smooth muscle. These depolarizations are small in guinea pigs (1–3 mV), but are much larger in most species (up to 40 mV) and impose an underlying rhythm on contractions evoked by neural activity.

The efferent innervation of the mucosa is largely from submucous neurons (Figures 1 and 2), although some myenteric neurons, notably the intrinsic sensory neurons, send axons to the mucosa. Submucous neurons also innervate submucosal arterioles and act as vasodilators. Extrinsic autonomic neurons, whether parasympathetic or sympathetic, typically terminate within myenteric or submucous ganglia, indicating that they modulate the activity of enteric neurons, rather than having a direct effect on the smooth muscle. This raises an interesting dichotomy because sympathetic neurons that modulate intestinal movement or secretion act as interneurons, while neurons in the same ganglia that supply intestinal blood vessels are final motor neurons.

Functional Types of Enteric Neuron

At least 12 functional types of enteric neurons are present in every mammal (Table 1). Intrinsic sensory neurons detect the chemistry, texture, and volume of the intestinal contents. Excitatory and inhibitory motor neurons separately innervate the circular and longitudinal muscle layers. There are also orally directed (ascending) interneurons, anally directed (descending) interneurons, two types of secretomotor neurons, and vasodilator neurons. In addition, intestinofugal neurons project to prevertebral sympathetic ganglia as the afferent arm of reflexes coordinating behavior between different intestinal regions. In guinea pigs, the most comprehensively studied species, many of these neurons can be further subdivided on the basis of their neurochemistry and connections (Table 2). For example, three classes of anally projecting (descending) interneurons can be discriminated neurochemically in guinea pig ileum: some are

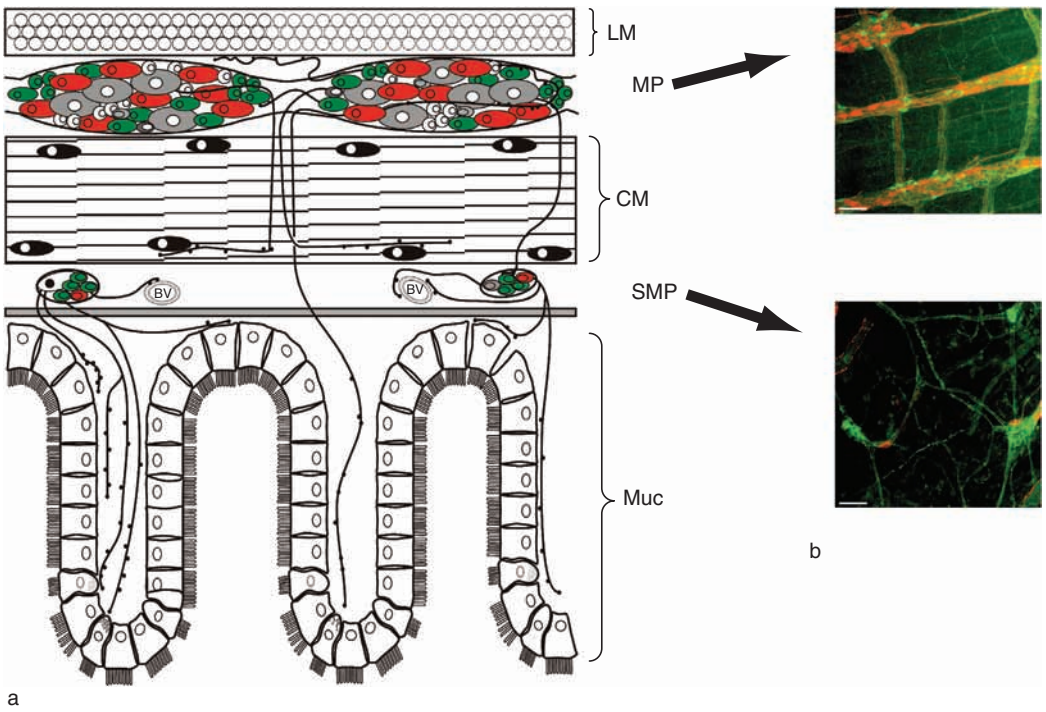

Figure 2 The layers of the intestine and their innervation. (a) Schematic drawing showing the layers of the intestine and where their innervation arises in the intrinsic ganglia. In this case, the projection pattern corresponds to guinea pig small intestine; villi are absent from other parts of the gastrointestinal tract. Neurons are color coded to match the neurons stained immunohistochemically in the micrographs (b), and neuron types not immunoreactive in the micrographs are shown as gray or white. Interstitial cells of Cajal are shown in the CM as interconnected black cells next to the MP and in the deep muscular plexus. Sources of innervation of longitudinal muscle, other ganglia, circular muscle, blood vessels, and mucosa are shown via representative axons. (b) Micrographs showing the MP (top) stained to reveal immunoreactivity for nitric oxide synthase (red) and calretinin (green), and the SMP (bottom) stained to reveal immunoreactivity for neuropeptide Y (red) and vasoactive intestinal peptide (green). Neurons in the MP are grouped into large ganglia (about 100 neurons in guinea pig), while SMP ganglia are much smaller (about 10 neurons). Note the extensive interconnections via interganglionic connectives and the neuropeptide Y reactive innervation of submucosal blood vessels, which arises from sympathetic vasoconstrictor neurons. LM, longitudinal muscle; MP, myenteric plexus; CM, circular muscle; SMP, submucous plexus; Muc, mucosa.

Table 1 Types of neurons that can be deduced from physiological and pharmacological studies to be present in the ENS of all species

Intrinsic sensory neurons
Excitatory motor neurons innervating the circular muscle
Excitatory motor neurons innervating the longitudinal muscle
Inhibitory motor neurons innervating the circular muscle
Inhibitory motor neurons innervating the longitudinal muscle
Motor neurons innervating the muscularis mucosae
Orally directed (ascending) interneurons
Anally directed (descending) interneurons
Cholinergic secretomotor neurons
Noncholinergic secretomotor neurons
Vasodilator neurons
Intestinofugal neurons

immunoreactive for nitric oxide synthase (NOS), vasoactive intestinal peptide (VIP), and other neuropeptides; others are immunoreactive for somatostatin and choline acetyltransferase (ChAT); another group express serotonin (5-hydroxytryptamine; 5-HT) and ChAT. However, neurochemical codes from one location cannot be confidently generalized to other preparations because regional differences occur within the same species (e.g., somatostatin interneurons project orally in guinea pig colon).

Electrophysiological studies in the small intestine and colon of guinea pigs and mice have identified two broad classes of neurons, AH and S neurons (Figure 3). Both classes are seen in human and rat tissue, but AH neurons have not been identified in pig tissue. AH neurons have prominent, prolonged, after-hyperpolarizations (AHPs) following their action potentials. They have a common shape (Figure 3), whether they are in the myenteric or submucous plexus, with large cell bodies and multiple axons that project circumferentially and ramify in myenteric ganglia, the submucous plexus, and the mucosa. There is strong evidence that they have a sensory function (Figure 4).

Table 2 Subtypes of neurons in guinea pig ileum and their distinguishing features

General function	*Subtype*	*Location*	*Neurochemistry*	*Electrophysiology*
Intrinsic sensory neuron	Chemoreceptor	Myenteric plexus	ChAT/tachykinins/calbindin	AH neuron
	Stretch/tension receptor	Myenteric plexus	ChAT/tachykinins/calbindin	AH neuron
	Mucosal mechanoreceptor	Submucous plexus	ChAT/tachykinins	AH neuron
Excitatory circular muscle motor neuron	Short oral projection	Myenteric plexus	ChAT/tachykinins/ENK	S neuron
	Long oral projection	Myenteric plexus	ChAT/tachykinins	S neuron
Excitatory longitudinal muscle motor neuron	Local projections	Myenteric plexus	ChAT/calretinin/tachykinins	S neuron
Inhibitory circular muscle motor neuron	Short anal projection	Myenteric plexus	NOS/VIP/PACAP/NPY	S neuron
	Long anal projection	Myenteric plexus	NOS/VIP/GRP	S neuron
Inhibitory longitudinal muscle motor neuron	Local projection	Myenteric plexus	NOS/VIP/GABA	S neuron
Ascending interneuron	Oral projection	Myenteric plexus	ChAT/calretinin/tachykinins	S neuron
Descending interneuron	Supplies myenteric plexus	Myenteric plexus	NOS/VIP/GRP	S neuron
	Supplies submucous plexus	Myenteric plexus	NOS/VIP/GRP	S neuron
	Supplies myenteric plexus	Myenteric plexus	ChAT/SOM	S neuron
	Inhibitory interneuron supplying submucous plexus	Myenteric plexus	SOM/ChAT	S neuron
	Supplies myenteric plexus	Myenteric plexus	ChAT/5-HT	S neuron
	Supplies submucous plexus	Myenteric plexus	ChAT/5-HT	S neuron
Secretomotor neuron	Cholinergic	Submucous plexus	ChAT/NPY/SOM/CGRP/GAL	S neuron
	Noncholinergic	Submucous plexus	VIP/Dyn/GAL	S neuron
Vasodilator neuron	Cholinergic	Submucous plexus	ChAT/calretinin	S neuron
	Noncholinergic	Submucous plexus	VIP/Dyn/GAL	S neuron
Interplexus interneuron	Noncholinergic	Submucous plexus	VIP/Dyn/GAL	S neuron

ChAT (choline acetyltransferase) is the enzyme that synthesizes acetylcholine; 'tachykinins' refers to substance P and neurokinin-A, which are products of the same gene and are difficult to distinguish immunohistochemically; ENK, enkephalin; PACAP, pituitary adenylyl cyclase-activating peptide; NPY, neuropeptide Y; GRP, gastrin-releasing peptide; GABA, γ-aminobutyric acid; SOM, somatostatin; CGRP, calcitonin gene-related peptide; GAL, galanin; Dyn, dynorphin.

Unlike AH neurons, S neurons have prominent fast excitatory postsynaptic potentials (EPSPs; Figure 3). Myenteric S neurons are morphologically diverse, with single axons that project orally or anally, and ramify in the muscle (motor neurons), supply other myenteric ganglia (interneurons) or submucous ganglia (interneurons), or run to the mucosa (secretomotor neurons). Most submucosal S neurons project to the mucosa. Others are vasodilator neurons.

Putting the Neurons in Their Circuits

The functions of neurons can only be fully defined when neurons are placed in physiologically meaningful circuits. Relevant evidence largely comes from analysis of neuronal behavior during simple motility reflexes. Mechanical or chemical stimulation of the intestinal wall elicits polarized reflexes: circular muscle contractions oral to and relaxations anal to the stimulus (ascending excitation and descending inhibition, respectively). Direct recordings from identified neurons during these reflexes, together with pharmacological analyses of transmission in reflex pathways and neurochemical studies of different neuronal subpopulations, have been combined to analyze the enteric neural circuitry. Collectively, the studies reveal the polarized circuit shown in Figure 5 along with the nature of transmission at most classes of synapses within the circuit. Key features include sensory neurons organized into recurrent excitatory networks around the intestinal circumference and with outputs to interneurons in two polarized feed-forward pathways running orally to excite and anally to inhibit the circular muscle. The sensory neurons have monosynaptic outputs to local excitatory and inhibitory motor neurons, but these also lead to polarized activity because the former run orally and the latter run anally.

The intrinsic sensory neurons are AH neurons and respond to several different stimuli, including muscle tension, increases in muscle length (distension), chemicals in the lumen, and mucosal deformation. The last stimulus is detected by submucosal sensory neurons, but the others are largely the province of myenteric sensory neurons. The myenteric sensory neurons that respond to distension have axons that run anally in addition to their normal circumferential projections. However, most myenteric sensory neurons have only circumferential projections.

The intrinsic sensory neurons transmit to ascending interneurons (S neurons), via acetylcholine (ACh)

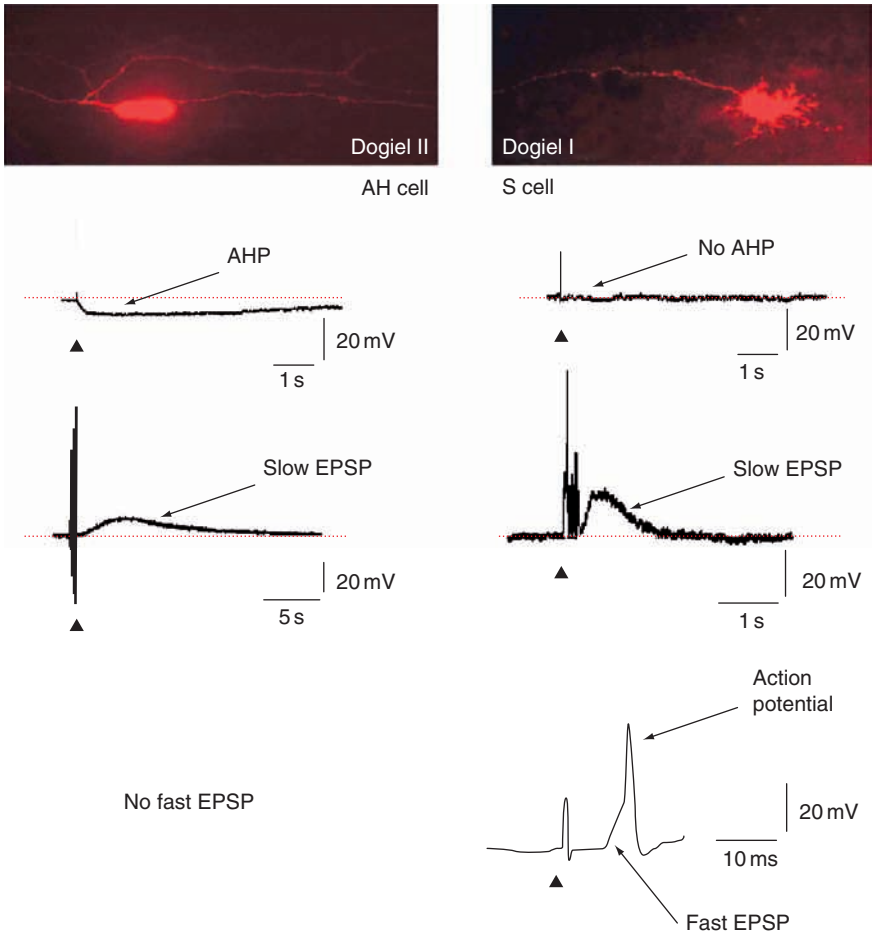

Figure 3 Types of neurons in the enteric nervous system: morphology and electrophysiology. The left side of this figure shows the shape and electrophysiological properties of AH/Dogiel type II neurons. The micrograph shows a neuron of this type injected with biocytin from an intracellular recording electrode and later fixed and stained with streptavidin–Texas red. The neuron is a typical Dogiel type II cell with a large oval cell body, lacking dendrites, but with at least two distinct axons that ramify within the neuron's own ganglion. The traces below show the action potential of one of these neurons, with its characteristic prolonged afterhyperpolarization (AHP) lasting well over 10 s (upper trace) and a slow excitatory postsynaptic potential (slow EPSP) evoked in the same neuron by a train of stimuli delivered to an interganglionic connective (lower trace). Fast EPSPs are not seen in these cells. The right side shows a micrograph of an S neuron, a typical Dogiel type I cell with one axon and short lamellar dendrites. The top trace shows the action potential of such a neuron, which lacks the prolonged AHPs of AH/Dogiel type II neurons. These neurons also exhibit slow EPSPs (middle trace), while they also have prominent fast EPSPs (bottom trace).

acting at nicotinic ACh receptors (nAChRs), ACh acting at muscarinic ACh receptors (mAChR), and tachykinins acting at tachykinin (neurokinin; NK-3) receptors. They transmit to other sensory neurons via NK-3 and/or NK-1 tachykinin receptors. Transmission between ascending interneurons is exclusively via nAChRs. Transmission from interneurons to excitatory motor neurons is via nAChRs, supplemented by tachykinins acting at NK-3 receptors. Transmission in the descending reflex pathways depends on the species and the location. In guinea pig ileum, transmission from intrinsic sensory neurons to descending interneurons and local inhibitory motor neurons is via an, as yet, unidentified transmitter/receptor combination. However, transmission from interneurons to nitric oxide synthase (NOS) interneurons may be mediated by ATP acting at purinergic ($P2Y_1$) receptors, while transmission from these neurons to inhibitory motor neurons is mediated by ATP acting at P2X receptors. No role for the nAChR has been identified in this pathway. By contrast, transmission within the descending inhibitory pathway of both guinea pig and rat colon depends on nAChRs at every functionally defined synapse.

Motor neurons supplying the muscle are S neurons. Their axons run longitudinally within the myenteric plexus before entering the circular muscle, where they branch and run circumferentially. Their projections are polarized according to function. Excitatory motor neurons run orally. Inhibitory motor neurons run

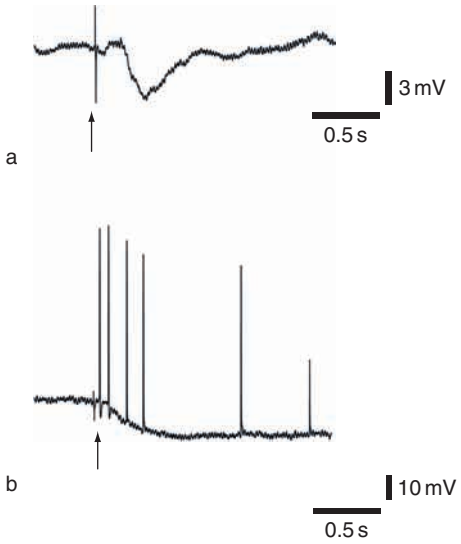

Figure 4 Sensory stimuli can excite both local reflexes and direct firing of action potentials in AH/Dogiel type I neurons. (a) Intracellular recording of the circular muscle response to local application of 10^{-3} mol l^{-1} L-alanine (arrow) to mucosal villi near the recording electrode. A robust and highly repeatable inhibitory junction potential is evoked by the stimulus. (b) The same stimulus evoked a burst of action potentials in an AH/Dogiel type II neuron, the cell body of which lies in the myenteric plexus close to the smooth muscle, in which reflexes of the type shown in (a) can be recorded. This burst of action potentials is characteristic of an intrinsic sensory neuron sensitive to chemical stimuli applied to the mucosa.

anally. Neurons innervating the longitudinal muscle divide close to their cell bodies and ramify locally.

Excitatory transmission to the smooth muscle is largely via mAChRs, but tachykinins play a significant, if lesser, role. Many myenteric neurons contain both ChAT and tachykinins. Some project to the muscle and are almost certainly excitatory motor neurons. Others are ascending interneurons. In guinea pig, all excitatory motor neurons contain and release ACh and most also contain tachykinins, but fewer cholinergic motor neurons may use tachykinins in other species. Inhibitory neuromuscular transmission also involves cotransmission, with at least two transmitters being released from inhibitory motor neurons. Both nitric oxide and ATP have been identified as transmitters from these neurons, although their relative contributions vary between species and locations within species. All inhibitory motor neurons contain NOS, but some interneurons also contain this enzyme. There is no reliable marker for neurons releasing ATP as a transmitter, as ATP is co-stored in vesicles with other neurotransmitters. The NOS inhibitory motor neurons also contain several neuropeptides (e.g., vasoactive intestinal peptide, pituitary adenylyl cyclase-activating peptide) that relax or inhibit the muscle and these may also be cotransmitters from these neurons.

Motility

Segmentation

Segmentation is the most enigmatic of the major behaviors of the intestine. What progress has been made arises from high-resolution mapping of intestinal diameter in video recordings as a function of both time and location. Such maps allow the relationships between the activity of adjacent regions to be determined, and hence whether constrictions propagate and, if so, their direction and speed of propagation. Luminal nutrients, either fatty acids like decanoic and dodecanoic acid or amino acids like L-phenylalanine and L-tryptophan, induce a fed-state-like behavior *in vitro*. This is blocked by tetrodotoxin, which blocks neural activity, and hyoscine, which largely blocks excitatory neuromuscular transmission, and so is dependent on neural activity. In guinea pig intestine, the stationary contractions result from excitatory junction potentials that generate action potentials, while the region outside the contracted area is actively inhibited. Thus, there is a local oscillator circuit that produces periodic local excitation of the circular muscle with an inhibitory surround. The identity of the oscillator is unknown, but it is probably the recurrent network of intrinsic sensory neurons. By contrast, in mice, neural activity enhances or depresses slow-wave-mediated contractions so that two pattern generators, one neural and the other myogenic, interact to produce the final pattern of constriction and dilation. There are no studies of segmentation in human intestine *in vitro*, and *in vivo* studies lack the spatial resolution to determine the relationship between neural and slow-wave activity. Both probably contribute to segmentation.

Peristalsis

Segmentation induced by luminal nutrient is occasionally interrupted by powerful constrictions that propagate rapidly ($20\,mm\,s^{-1}$) along the intestinal segment, peristalsis and antiperistalsis. Peristalsis is a strong constriction that appears close to the oral end of the segment and then extends anally. Antiperistalsis appears at the anal end and extends orally. Stretching the intestinal wall evokes propagating constrictions similar to those evoked by nutrients, but these always run anally.

Recent data challenge the simple model of peristalsis as arising from ascending excitation and descending inhibition. Importantly, localized stimuli (distension, mucosal distortion) activate anally propagating excitation of circular muscle although the stimuli do not move. This implies that the activity is carried by descending neurons, rather than being due to continuous reactivation of reflexes. This anally

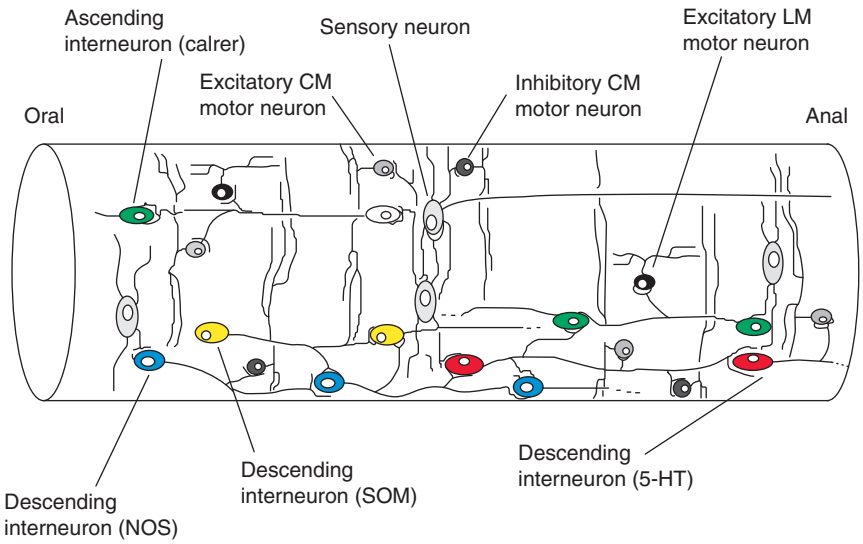

Figure 5 The neural circuitry that can be deduced from physiological, pharmacological, and neuroanatomical studies to be responsible for propulsive motility reflexes in the guinea pig ileum. There are three basic components to the circuit. The intrinsic sensory neurons (pale gray, not subdivided by sensory modality for the sake of some simplicity) project predominantly circumferentially and make synaptic connections with each other as well as having outputs to virtually all other neuron types. Some project anally, but their targets have not been identified. These neurons provide excitatory outputs to ascending interneurons (green), which in turn excite other ascending interneurons, excitatory motor neurons that project orally to supply the circular muscle (CM) (dark gray), and locally projecting longitudinal muscle (LM) motor neurons (excitatory, black). The intrinsic sensory neurons also excite descending interneurons (blue) that contain nitric oxide synthase (NOS) and form the major pathway for descending inhibitory reflexes. The NOS neurons excite inhibitory motor neurons (black) that project anally, ascending interneurons, and two other populations of descending interneurons, those that contain choline acetyltransferase and somatostatin (SOM; yellow) and those that contain choline acetyltransferase and 5-hydroxytryptamine (5-HT; red). The latter contact both excitatory motor neurons and ascending interneurons, but the role and connections of the former are not yet clear. At each point along the pathways, all neurons except the choline acetyltransferase/SOM neurons receive inputs from local intrinsic sensory neurons.

directed pathway probably includes interneurons that contain 5-HT and transmit to excitatory motor neurons via 5-HT_3 receptors, and other interneurons that are excited by ATP acting at P2X receptors. A key finding is that longitudinal muscle is excited simultaneously with the circular muscle, contrary to predictions of earlier models.

Peristalsis in the fed state depends on both the composition and the volume of the digesta at the point of initiation. How these stimuli interact is unknown. However, a neural switch is needed to shift the ratio of segmentation to propulsion to account for changes in intestinal transit resulting from varying luminal content. Despite the myogenic slow waves that are prevalent in many species, peristalsis is predominantly regulated by neural activity. Slow waves impart an underlying rhythm rather than determining the direction or speed of propagation of the constriction.

Interdigestive Motor Complex

Interdigestive migrating motor complexes are difficult to identify *in vitro*, although contraction complexes similar in form and timing to phase III type are seen in isolated mouse colon and small intestine (Figure 6). Contraction complexes seen in the jejunum *in vitro* often propagate from anal to oral, so are not identical to the MMCs seen *in vivo*. Nevertheless, the neural circuits regulating these two behaviors must have many neurons in common and the behaviors share several key properties, including very low cycling rates (1 cycle every 2–5 min) and propagation speeds (about 1 mm s^{-1}). The latter are too slow to be accounted for by activity passing along chains of interneurons, because all identified interneurons run longitudinally. By contrast, AH neurons send axons around the intestinal circumference with a slight bias in the anal direction and make excitatory synapses with each other. These transmit via slow EPSPs that can initiate action potentials and can be evoked by physiological stimuli. This arrangement automatically slows longitudinal conduction of activity within the network, so that activity in a recurrent AH neuron network can account for the slow propagation of the contraction complexes. Modeling networks of these neurons indicates that interactions between their AHPs and the slow EPSPs are essential for encoding ongoing stimuli applied to the intestinal wall. Complete suppression of the AHPs causes the network to produce propagating waves of excitation that

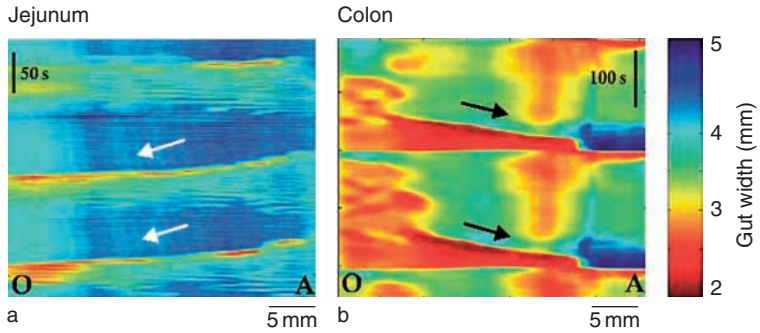

Figure 6 Spatiotemporal maps showing propagating contractile complexes in the mouse jejunum (a) and colon (b). These maps are constructed from video recordings of segments of jejunum and colon *in vitro* as they contract spontaneously. In each frame of the recording, computer software converts the intestine's image into a silhouette, rotates the silhouette until it is horizontal, and then for each horizontal pixel measures the cross section of the silhouette. This value is converted to a color scale, with the minimum diameter represented by dark red and the maximum diameter represented by dark blue. The diameter of the intestine at each point along the segment is then plotted as a multicolored line. The result of this calculation for the next frame is then added below, and so on. In these maps, time increases from top to bottom and the oral end of the segment is on the left. Arrows indicate the direction of apparent propagation of constrictions. Contraction complexes in the jejunum repeat at approximately 2 min intervals and appear to propagate orally. Similar complexes appear in the colon, but propagate anally.

migrate along the intestine at a speed similar to that of MMCs. The activity can propagate in either direction from the point of initiation, but oral to anal propagation is imposed on the network if the point of initiation is at the oral end. Hence, both interdigestive MMCs and the propagating contraction complexes seen *in vitro* may be driven by activity in recurrent AH neuron networks. In this state, AH neurons cannot encode the magnitude of a stimulus and instead act as coordinating interneurons and not sensory neurons. This suggests that there is a physiological switch that converts neurons that detect the composition, consistency, and volume of the digesta into interneurons with no sensory function, and back again.

Secretion of Water and Electrolytes

Like intestinal motility, secretion of water and electrolytes into the intestinal lumen is essential for digestion and is directly controlled by the ENS. Blockade of neural activity leads to net absorption of water and salt. Several neurally mediated secretory responses to physiological stimuli, including distension of the intestinal wall and mechanical deformation of the mucosa, have been identified. Furthermore, the hypersecretion produced by diarrhea-producing toxins, such as cholera toxin, heat-labile *Escherichia coli* enterotoxin, and heat-stable *E. coli* enterotoxin type A, depends critically on enteric neural activity, despite the same toxins directly activating mucosal enterocytes to promote secretion.

The secretomotor neurons are predominantly located within the submucous plexus. There are two classes of submucosal secretomotor neurons: cholinergic secretomotor neurons and noncholinergic secretomotor neurons. The latter contain vasoactive intestinal peptide (VIP), a potent secretagogue, and this is their major neurotransmitter. Their inputs and their potential sources are illustrated in Figure 7. While cholinergic secretomotor neurons are under the direct control of circuits arising in the myenteric plexus, the noncholinergic secretomotor neurons have a much more complex role. VIP neurons receive at least seven types of synaptic inputs from several sources (Table 3). These neurons receive excitatory input producing fast EPSPs via nAChR from submucosal neurons and from myenteric neurons. Many VIP neurons also exhibit fast EPSPs mediated by ATP acting at P2X receptors, although their source is unknown. In addition, these neurons receive two or more pharmacologically distinct slow EPSPs from myenteric and submucosal neurons. Much of this slow input is mediated by ATP acting at $P2Y_1$ receptors, with the rest being probably mediated by a neuropeptide, possibly VIP. The complexity of their input is underlined by the observation that VIP secretomotor neurons also have strong inhibitory inputs from two distinct sources. Sympathetic axons produce inhibitory synaptic potentials mediated by norepinephrine acting at α_2-adrenoceptors, and this is the only direct input from outside the gut into the secretomotor pathways. A second inhibitory input comes from myenteric neurons. The roles of the diverse pathways converging on these neurons are yet to be clarified.

Intestino-intestinal Reflexes

There are several long pathways by which the behaviors of distant regions of the gut are coordinated.

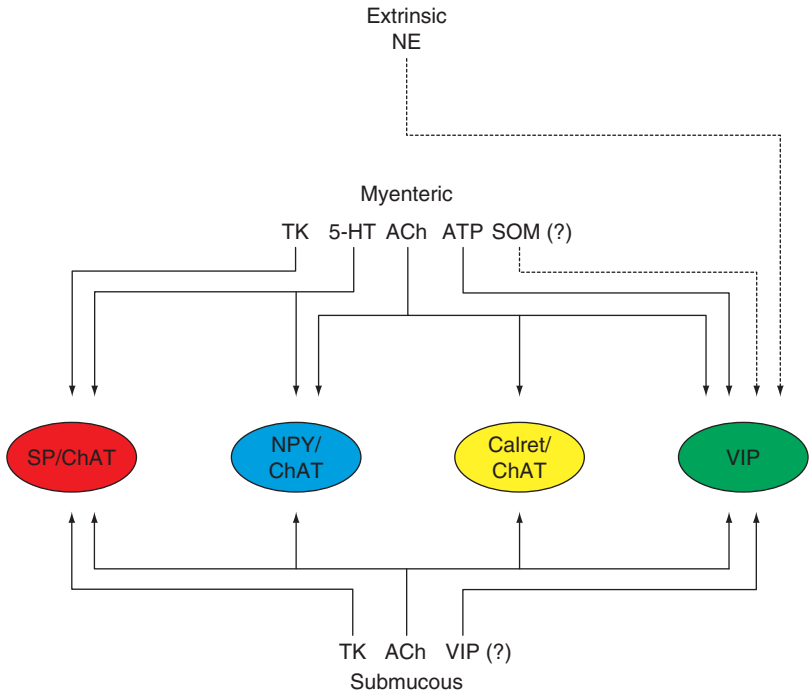

Figure 7 Sources and transmitters of inputs to neurochemically identified neurons in the submucous plexus. Schematic diagram showing the different sources and transmitters of the synaptic inputs to neurochemically identified submucous neurons of the guinea pig ileum. Each discrete class of submucous neurons has a distinct pattern of input, with all receiving excitatory inputs from myenteric and submucous neurons (solid arrows), but only VIP neurons receiving inhibitory inputs (dashed arrows) from both myenteric neurons (probably mediated by somatostatin (SOM)) and the prevertebral ganglia (mediated by norepinephrine (NE)). All classes of neurons appear to receive excitatory input from submucous and myenteric neurons via acetylcholine (ACh) acting at nicotinic receptors. However, the transmitters employed by other excitatory inputs are diverse. Excitatory postsynaptic potentials mediated by tachykinins (TK; notably SP) appear to be confined to the SP/ChAT neurons. Excitatory inputs mediated via serotonin (5-HT) are largely confined to NPY/ChAT neurons, although there may also be an input to SP/ChAT neurons. Excitatory inputs mediated by ATP are largely confined to the VIP neurons. SP/ChAT, substance P/choline acetyltransferase; NPY/ChAT, neuropeptide Y/choline acetyltransferase; Calret/ChAT, calretinin/choline acetyltransferase; VIP, vasoactive intestinal peptide.

Table 3 Sources of inputs to VIP secretomotor neurons

Source	*Synaptic potential*	*Transmitter*	*Receptor*
Myenteric neuron	Fast EPSP	ACh	Nicotinic
Submucous neuron	Fast EPSP	ACh	Nicotinic
Myenteric neuron	Fast EPSP	ATP	P2X
Myenteric neuron	Slow EPSP	ATP	$P2Y_1$
Myenteric neuron	Intermediate EPSP	ATP	$P2Y_1$
Submucous neuron	Slow EPSP	?VIP?	Unknown
Myenteric neuron	IPSP	Somatostatin	Unknown
Sympathetic ganglion cells	IPSP	Norepinephrine	α_2-Adrenoceptor

ACh, acetylcholine; ATP, adenosine 5′-triphosphate; EPSP, excitatory postsynaptic potential; IPSP, inhibitory postsynaptic potential; P2X, $P2Y_1$, purinergic; VIP, vasoactive intestinal peptide.

In general, this coordination takes the form of inhibition of proximal regions by stimuli applied distally. A classic example is the ileal brake, in which fats in the ileum cause marked slowing of transit along the jejunum. This allows more time for the digestion and absorption of other nutrients, thereby increasing the efficiency of the system. Similar reflexes are triggered by bile salt metabolites in the colon or by colonic distension and also slow the arrival of undigested material into regions not involved in digestion or nutrient absorption. Similar reflexes regulate gastric emptying.

A major pathway for these long-distance reflexes is via myenteric neurons whose axons run to, and

make synapses in, prevertebral sympathetic ganglia (Figure 1). These intestinofugal neurons excite sympathetic neurons within the celiac, superior mesenteric, and inferior mesenteric ganglia, which in turn project to the myenteric plexus of more proximal intestinal regions to inhibit enteric neural circuits regulating motility. Most intestinofugal neurons are interneurons receiving input from local intrinsic sensory neurons and exciting their targets via nAChRs. They also contain various neuropeptides that may act as cotransmitters, producing slow EPSPs in some physiological circumstances. Some intestinofugal neurons may also have sensory functions of their own.

There also appears to be a pathway by which stimuli in the colon can inhibit secretion in the small intestine. Evidence for this is less direct than for the regulation of motility via the intestino-intestinal pathway, but is nevertheless compelling. Celiac neurons that contain somatostatin and norepinephrine have only one identified target in the ileum, the submucous ganglia, and these neurons receive a major portion of the input from intestinofugal neurons. Thus, activation of intestinofugal pathways should inhibit VIP secretomotor neurons in the submucous plexus, as these are the only submucous neurons with inputs from sympathetic axons. This implies that the content of the colon can regulate secretion of water and salt in the small intestine.

Vasomotor Control

Intestinal blood flow is critical for digestion, as well as being a key element of overall blood pressure control. During exercise and other environmental stresses, blood flow to the gut is restricted to allow maintenance of appropriate flows to the brain, heart, and skeletal muscles, for example. This process is regulated by the CNS via vasoconstrictor neurons with cell bodies in the para- and prevertebral sympathetic ganglia, which innervate both mesenteric blood vessels and submucosal arterioles. Vasoconstrictor neurons in prevertebral ganglia do not receive input from intestinofugal neurons, so vasoconstriction is not modulated by direct input from the gut, although pathways that run via the CNS are certainly involved.

On the other hand, there are also enteric vasodilator pathways. Submucosal vasodilator neurons have been identified by stimulation of individual neurons in ganglia and measuring dilation of nearby arterioles. These neurons are excited by cholinergic neurons in distension-activated reflex pathways running from the myenteric plexus, and act upon the blood vessels via mAChRs. How vasodilation, motility, and mucosal secretion are coordinated by the intrinsic pathways will be a fertile ground for research in the near future.

Neural Regulation of Digestive Enzyme Secretion

Secretion of digestive enzymes was thought to be primarily regulated by circulating hormones, notably cholecystokinin (CCK), released by nutrients in the duodenum. However, pancreatic enzyme secretion is also mediated via the autonomic nervous system. Locally released CCK activates vagal afferent nerve terminals within the duodenal wall, which leads to activation of vagal efferents supplying pancreatic acinar cells. This reflex is supplemented by enteric neurons that project to the pancreas to modulate enzyme release.

Summary

All major processes involved in digestion depend on the autonomic nervous system. The intestinal movements and secretion of water and salt depend on the ENS. Coordination of different intestinal regions depends on reflex pathways running through sympathetic ganglia. Coordination of whole-body functions and secretion of digestive enzymes depend on central sympathetic and parasympathetic pathways. Thus, the innervation of the gut reflects its role as 'the way to a man's heart.'

See also: Adenosine Triphosphate (ATP); Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Nervous System: Neuroanatomy; Autonomic Nervous System: Central Control of the Gastrointestinal Tract; Autonomic Neuroeffector Junction; Cotransmission; Enteric Nervous System: Sensory Pathways; Enteric Nervous System: Neural Circuits and Chemical Coding; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Gastrointestinal Signals: Stimulation; Gastrointestinal Signals: Satiety; Gastrointestinal Tract role in Neural Control of Metabolism, Food Intake and Body Weight: A Summary; Muscarinic Receptors: Autonomic Neurons; Nitric Oxide; Peptidergic Receptors; Purinergic Receptors.

Further Reading

Bornstein JC, Furness JB, Kunze WAA, et al. (2002) Enteric reflexes that influence motility. In: Brookes SJH and Costa M (eds.) *Innervation of the Gastrointestinal Tract*, pp. 1–55. London: Taylor & Francis.

Brookes SJH and Costa M (2002) Cellular organisation of the mammalian enteric nervous system. In: Brookes SJH and Costa M (eds.) *Innervation of the Gastrointestinal Tract*, pp. 393–467. London: Taylor & Francis.

Costa M, Brookes SJH, Steele PA, et al. (1996) Neurochemical classification of myenteric neurons in the guinea-pig ileum. *Neuroscience* 75: 949–967.

Furness JB (2006) *The Enteric Nervous System*. Oxford: Blackwell Publishing.

Furness JB, Jones C, Nurgali K, et al. (2004) Intrinsic primary afferent neurons and nerve circuits within the intestine. *Progress in Neurobiology* 72: 143–164.

Galligan JJ (2002) Pharmacology of synaptic transmission in the enteric nervous system. *Current Opinion in Pharmacology* 2: 623–629.

Gwynne RM, Thomas EA, Goh SM, et al. (2004) Segmentation induced by intraluminal fatty acid in isolated guinea-pig duodenum and jejunum. *Journal of Physiology (London)* 556: 557–569.

Owyang C and Logsdon CD (2004) New insights into neurohormonal regulation of pancreatic secretion. *Gastroenterology* 127: 957–969.

Sanders KM, Koh SD, and Ward SM (2006) Interstitial cells of Cajal as pacemakers in the gastrointestinal tract. *Annual Reviews of Physiology* 68: 307–343.

Thomas EA, Sjövall H, and Bornstein JC (2004) Computational model of the migrating motor complex of the small intestine. *American Journal of Physiology – Gastrointestinal and Liver Physiology* 286: G564–G572.

Wood JD (2004) Enteric neuroimmunology and pathophysiology. *Gastroenterology* 127: 635–657.

Autonomic Nervous System: General Overview

N R Keller and D Robertson, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The autonomic nervous system (ANS) is composed of the sympathetic, parasympathetic, and enteric neurons, and the adrenal medulla. It is one of two major divisions of peripheral nerves, the other being the sensorisomatic nervous system. Reciprocally integrated with central autonomic foci, the ANS maintains homeostasis and adaptive responses to physical and mental stress virtually beyond the awareness of the host. The sympathetic system controls fight-or-flight reactions by increasing sympathetic outflow to the heart and other viscera, while the parasympathetic system maintains basal regulation of bodily functions. The enteric nervous system, on the other hand, manages peristalsis and influences glandular secretions. Among the many different aspects of the ANS, this article focuses primarily on neurotransmission within the sympathetic branch (Figure 1).

Neurotransmitters were first proposed as chemical stimulants released by nerve terminals early in the twentieth century. Research in the past 10 to 20 years has generated substantial progress in identifying intra- and extracellular signaling molecules, developmental regulatory transcription factors, metabolic pathways, targeted pharmacologic agents, and the influence of genetic mutations in the pathogenesis of autonomic dysfunction. Along with these discoveries, our understanding of the mechanisms underlying various neurotransmitter disorders has also grown. A sampling of genetic, metabolic, and transport errors is highlighted herein.

Synthesis

Dopamine β-Hydroxylase

Dopamine β-hydroxylase deficiency Dopamine β-hydroxylase (DBH) is an oxidoreductase of the copper type II, ascorbate-dependent monooxygenase family, and is responsible for conversion of dopamine (DA) to norepinephrine (NE). DBH deficiency is an autosomal recessive disorder of heterogeneous nature. The human DBH gene contains 12 exons and is situated at the 9q34 locus, immediately adjacent to the blood group locus ABO. All DBH-deficient patients characterized thus far carry one copy of an IVS1 + 2T → C mutation in intron 1 that prevents appropriate splicing. *In vitro* analysis, however, has shown that this mutation does not entirely explain undetectable levels of DBH protein in plasma, because some properly spliced DBH message can be detected. Instead, concomitant genetic changes such as –1021C → T, which is linked with very low plasma DBH activity, are suspected as inhibitory transcriptional modulators. Thus, NE deficiency may only occur in the context of a specific haplotype, not merely as the result of an individual mutation. Missense mutations of exon 2 (D100E) and of exons 1 and 6 residing *in cis* (V87M and D331N) have also been suggested as pathogenic accomplices.

Both norepinephrine and epinephrine (EPI) are completely absent in plasma, urine, and cerebrospinal fluid (CSF) of humans with DBH deficiency (Figure 2), since the enzyme is vital for conversion of DA to NE. Untreated individuals experience a lifetime of severe orthostatic hypotension and mild eyelid ptosis, with some skeletal muscle hypotonia and nasal stuffiness. Perinatal progress is understandably precarious; near-fatal events in the first few days of life include hypotension, hypoglycemia, hypothermia, and seizures. In survivors, delay in eye opening may occur. Because neonates with undetected enzyme insufficiency may succumb to such complications, it is possible that the incidence of DBH deficiency is greater than predicted based on the number of affected adults, although it is still quite rare.

The lack of sympathetic tone severely limits the hemodynamic response to upright posture, with systolic blood pressure (BP) routinely falling well below 100 mmHg with a relatively small increase in heart rate (HR). Presyncopal dizziness is often accompanied by dyspnea, nuchal discomfort, and occasional chest pain. Dizziness is exacerbated by warm environmental temperature. For these reasons, individuals with DBH deficiency are not able to tolerate physical exercise. Hypothermia is seen in infants. Male reproduction is complicated by retrograde ejaculation.

Autonomic testing in DBH deficiency is abnormal for sympathetic function. No increase in BP is elicited in the cold pressor test, isometric handgrip, or phase IV of the Valsalva maneuver. Patients are normal, however, with respect to sweating, lacrimation, salivation, corneal and deep tendon reflexes, sensory function, and the senses of taste and smell. Gastrointestinal function is within normal limits, and bladder function is preserved. Nocturnal polyuria with increased dopamine-mediated natriuresis is a shared feature of the disease.

Sympathetic nerve endings in DBH deficiency appear normal and stain positive for cytosolic

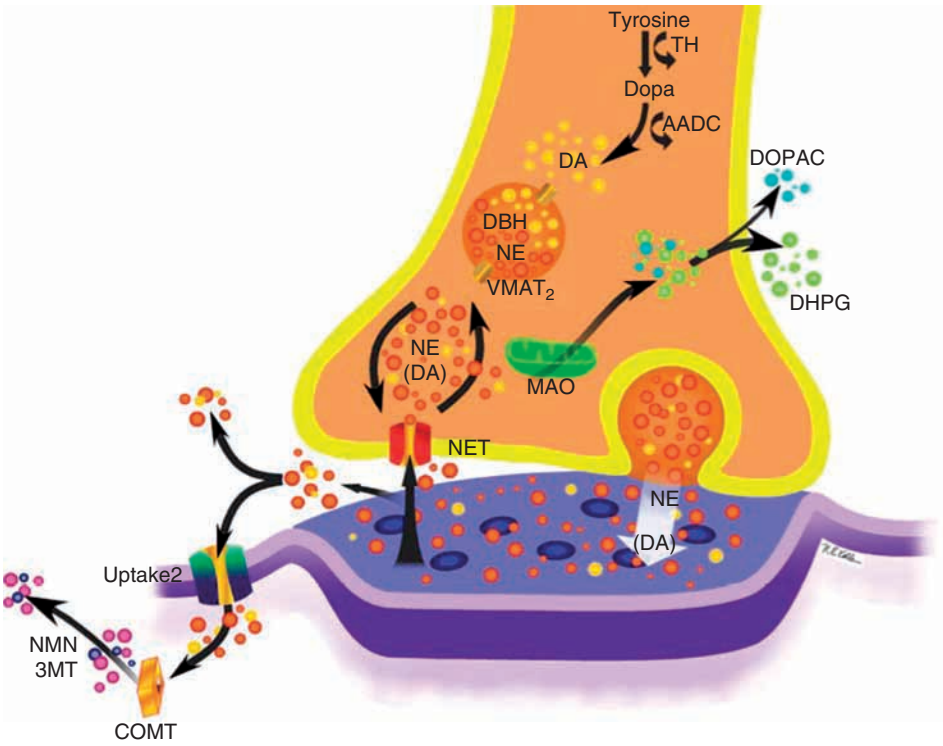

Figure 1 Normal sympathetic neurotransmission. Norepinephrine (NE) is the major neurotransmitter in postganglionic sympathetic and central noradrenergic neurons. Enzymatic conversion of tyrosine to dopamine (DA) takes place in the cytoplasm, whereas transformation of dopamine to NE occurs within the storage vesicles. After release into the synapse, residual catecholamines are pumped back into the presynaptic neuron by the norepinephrine transporter (NET), and returned to storage vesicles by the vesicular monoamine transporter 2 ($VMAT_2$). A smaller portion of the surplus transmitters diffuse into the extracellular space; these transmitters are cleared by the low-affinity, high-capacity 'Uptake 2' transporter, which has recently been renamed the 'extraneuronal monoamine transporter' in humans. Breakdown of cytosolic catecholamines is accomplished in part by monoamine oxidases (MAOs), while extracellular metabolism is conducted by catechol-*O*-methyltransferase (COMT). NMN, normetanephrine; 3MT, 3-methoxytyramine; DBH, dopamine β-hydroxylase; TH, tyrosine hydroxylase; Dopa, dihydroxyphenylalanine; DOPAC, 3,4-dihydroxyphenylacetic acid; AADC, L-aromatic amino acid decarboxylase; DHPG, 3,4-dihydroxyphenylglycol.

tyrosine hydroxylase, the rate-limiting enzyme in DA synthesis. Normally, DA is taken into the vesicles for conversion to NE by DBH. Without dopamine transformation, however, it is released from stimulated noradrenergic nerve terminals in place of NE. DBH is also present in the normal adrenal medulla, thus the effects of enzyme failure are undetectable tissue EPI and NE.

Without NE-mediated negative feedback on tyrosine hydroxylase, cytosolic DA production is amplified. Plasma levels of DA, therefore, are notably elevated; approximately five- to tenfold. Urine DA is more than doubled and CSF levels are also increased. The DA metabolites homovanillic acid (HVA) and 3-methoxytyramine (3MT) are roughly double normal values in urine. Plasma and CSF dihydroxyphenylalanine (DOPA) levels are double to triple normal values. As might be expected, prolactin is low. The NE metabolites 3-methoxy-4-hydroxyphenylglycol (MHPG) and normetanephrine (NM) are below detectable limits in urine.

Patients can be successfully treated with droxidopa (dihydroxyphenylserine, or L-DOPS), which is endogenously converted to NE by L-aromatic amino acid decarboxylase (AADC). Both intraneuronal and extraneuronal conversions take place, since AADC is present in most tissues. Adrenomedullary conversion of DA to NE does not occur, however, and EPI levels remain undetectable. With drug administration, BP increases dramatically and NE concentrations approach the normal range in plasma and urine. Consequently, standing time is greatly enhanced, as is quality of life – so improved, in fact, that in 2005 a 26-year-old patient treated with droxidopa for less than 2 years of her life was able to compete in and finish a 26-mile marathon.

Menkes Disease

Menkes disease is an X-linked fatal neurodegenerative disease caused by impaired copper transport in the gastrointestinal tract, placenta, and the blood–brain barrier, resulting in systemic deficiency of the

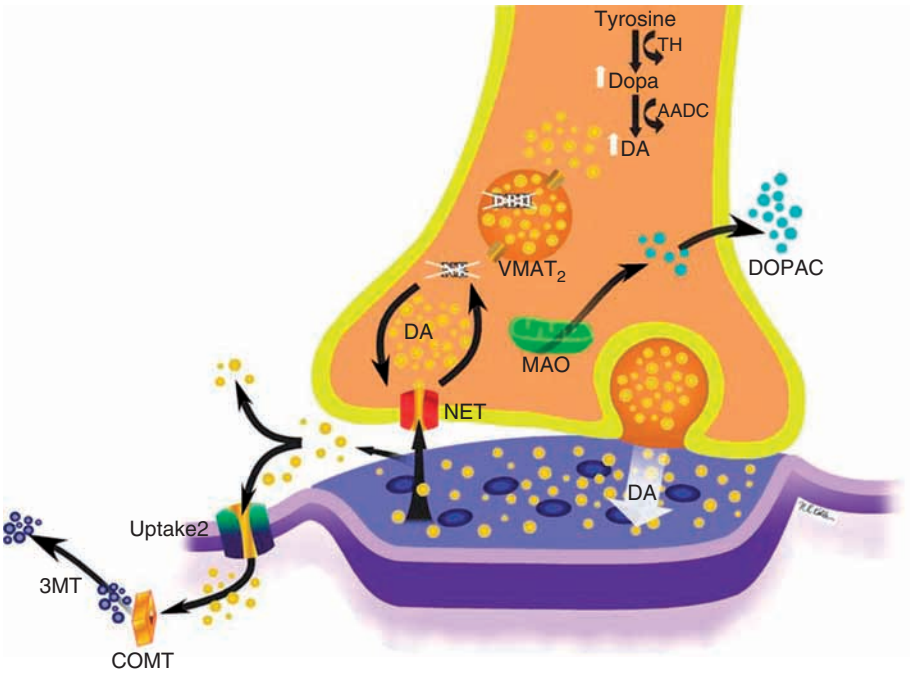

Figure 2 Dopamine β-hydroxylase (DBH) deficiency. Lack of DBH causes the total absence of norepinephrine (NE) in noradrenergic nerve terminals. Sympathetic neurotransmission, therefore, results in the release of dopamine (DA). Because there is no NE-mediated negative feedback on tyrosine hydroxylase (TH), DA production is increased. Reuptake of synaptic DA occurs through the norepinephrine transporter (NET). Enzymatic breakdown yields increased DA metabolites. (Other abbreviations as in **Figure 1**.)

element. Mutations in the *ATP7A* gene are responsible for the defective protein, MNK, a P-type ATPase that belongs to the CPx-type family of transmembrane enzymes necessary for ATP-dependent translocation of metal cations across cell membranes. MNK typically resides in the *trans*-Golgi network, where it delivers copper to local enzymes, including DBH. The ATPase can also be found in cytosolic vesicular compartments and the plasma membrane, where it mediates efflux of excess copper.

Copper is an essential cofactor for DBH. Consequently, plasma catecholamines are abnormally low, although NE levels in the CSF may be normal or elevated. Typically, however, because DBH activity is impaired, plasma and CSF DOPA, DA, and dihydroxyphenylacetic acid (DOPAC) are elevated. Ratios of plasma DOPA to dihydroxyphenylglycol (DHPG), the chief neuronal metabolite of NE, are increased approximately sevenfold, and DOPAC to DHPG ratios are about fivefold greater. CSF ratios follow a similar pattern. In addition, the proportion of HVA to vanillylmandelic acid (HVA:VMA) increases in the urine.

Early detection and treatment of the disease is crucial, since copper deficiency begins during gestation. Affected infants often appear normal until 2 to 3 months of age. Clinical symptoms vary somewhat depending on the exact mutation, but most neonates experience seizures, severe hypotonia, hypothermia, and extreme developmental delay. Connective tissue abnormalities due to impaired lysyl oxidase may appear as hyperelastic skin, hernias, and aortic aneurysms. Copper administration, however, does not correct these developmental defects. The most striking outward symptom may be the appearance of the hair, which is unusually short, sparse, coarse, and twisted, and is often lightly pigmented. It has been compared to steel wool, and the disease was formerly referred to as the 'kinky hair syndrome.' The first indicators of autonomic dysfunction are likely temperature instability and hypoglycemia. Without treatment using parenteral copper-histidine, death typically occurs by 3–4 years of age. Copper-histidine therapy for 3 months ordinarily normalizes serum copper, ceruloplasmin, DA, and NE levels.

Tetrahydrobiopterin

Hyperphenylalaninemia (HPA) and phenylketonuria (PKU) are the result of impaired enzymatic action of phenylalanine-4-hydroxylase (PAH). Underlying the PAH malfunction, however, is an array of direct and indirect mechanisms ultimately linked to tetrahydrobiopterin (BH4) deficiency. BH4 is a cofactor for PAH, tyrosine-3-hydroxylase (TH), and tryptophan-5-hydroxylase (TPH). Impaired hydroxylation of tyrosine and tryptophan results in catecholamine

and serotonin deficiencies. Determination of pterins and neurotransmitter metabolites in CSF is paramount in establishing an early treatment plan.

Clinical features, along with specific biochemical criteria, distinguish disease variants. The terms 'severe' (typical, general) and 'mild' (atypical, peripheral, partial) are used to describe symptoms, and indicate the need for treatment with neurotransmitter precursors. Symptoms may not be obvious within the first few weeks of life, but are usually apparent by the fourth month. Severe forms of BH4 deficiency manifest as mental retardation, tone and posture abnormalities, convulsions, somnolence, irritability, dyskinesias, hypersalivation and difficulty swallowing, and intermittent hyperthermia without underlying infection. Neurologic symptoms may follow a diurnal pattern, being worst in the evenings and less bothersome in the mornings after rest and neurotransmitter precursor replenishment.

Cofactor deficiency is attributable to genetic mutations in the enzymes responsible for BH4 biosynthesis or regeneration. Here, focus is limited to faulty metabolic pathways causing catecholamine deficiency. Catechol replacement usually consists of droxidopa, with carbidopa added to inhibit peripheral DA production. Levels of plasma phenylalanine warrant close attention when administering neurotransmitter therapy, since high concentrations of phenylalanine can hinder precursor membrane transport or compete with tyrosine.

The form of HPA encountered most is 6-pyruvoyl-tetrahydropterin synthase (PTPS) deficiency, which also exhibits the largest amount of genetic variability. Over 50% of BH4 deficiencies are PTPS and about 80% of affected individuals have the severe form of the disease. Phenotypic expression ordinarily reflects homozygous, compound heterozygous, and obligate heterozygous enzyme activities. Mutations in the PTS gene have been located in all six exons and generally result in a deficient or inactive enzyme product. An exception is the K129E mutant, which demonstrates heightened enzyme activity in a cell-specific manner *in vitro*, suggesting that posttranslational modification plays a role in inducing the upper limits of enzyme performance.

Analysis of urinary pterins discloses that severe PTPS produces the highest amount of neopterin of the BH4 deficiencies, while levels in the mild form are comparable to the pterin-4α-carbinolamine dehydratase (PCD) deficiency profile. Neurotransmitter metabolites are often normal in the CSF of persons with the mild version, as they are in the rarer intermediate form of PTPS deficiency, whereas urinary pterin concentrations are similar to the severe variety. Milder symptoms can worsen with age, and CSF neurotransmitter levels, although normal soon after birth, decrease in accordance with neurological decline.

Most patients with dihydropteridine reductase (DHPR) deficiency lack enzyme activity entirely, or nearly so. Thus, the majority of patients have the severe phenotype. About half are found in southern Europe. Again, mutations are distributed fairly evenly throughout the gene, and most are of the missense type. Urine biopterin levels are higher than in any other BH4 deficiency, and pterin synthesis in general is increased due to lack of feedback inhibition of GTP cyclohydrolase I (GTPCH). Neopterin is normal or slightly elevated. The CSF neurotransmitter metabolic profile is similar to severe PTPS deficiency, with HVA levels more than twice those of 5-hydroxyindoleacetic acid (5HIAA). An unrelated consequence of DHPR deficiency is central folate depletion, presumed to be due to accumulation of 7,8-dihydrobiopterin that competitively interferes with reduction of 7,8-dihydrofolic acid to the active tetrahydrofolic acid.

Both GTP cyclohydrolase I (GTPCH) deficiency and PCD are relatively rare compared to other BH4 disease varieties, with the former being somewhat variable clinically but in general, more severe. Pterins are virtually undetectable in urine from GTPCH patients since the enzyme participates in early pterin synthesis, while neopterin is elevated in PCD. In CSF, pterins follow a similar trend, although both neopterin and biopterin are increased in PCD. Neurotransmitter metabolites are below normal in both disorders as well.

Metabolism

Monoamine Oxidase

Monoamine oxidase (MAO) is located in the outer mitochondrial membrane where oxidative deamination of neuroactive, vasoactive, and xenobiotic amines occurs (**Figure 3**), producing hydrogen peroxide and aldehydes. MAO-A and MAO-B are encoded by two adjacent gene products on the X-chromosome, with different molecular weights and immunological properties. They share about 70% homology and neurons routinely contain both forms. Norepinephrine and serotonin (5-hydroxytryptamine; 5-HT) are preferential substrates for MAO-A, while dopamine, phenylethylamine, and benzylamine are metabolized by MAO-B. Dopamine, tyramine, and tryptamine are substrates of both isoenzymes.

MAO has been implicated in the pathogenesis of neurodegeneration, including that underlying Parkinson's disease. Active transport of catecholamines into neuronal vesicles is dynamically opposed

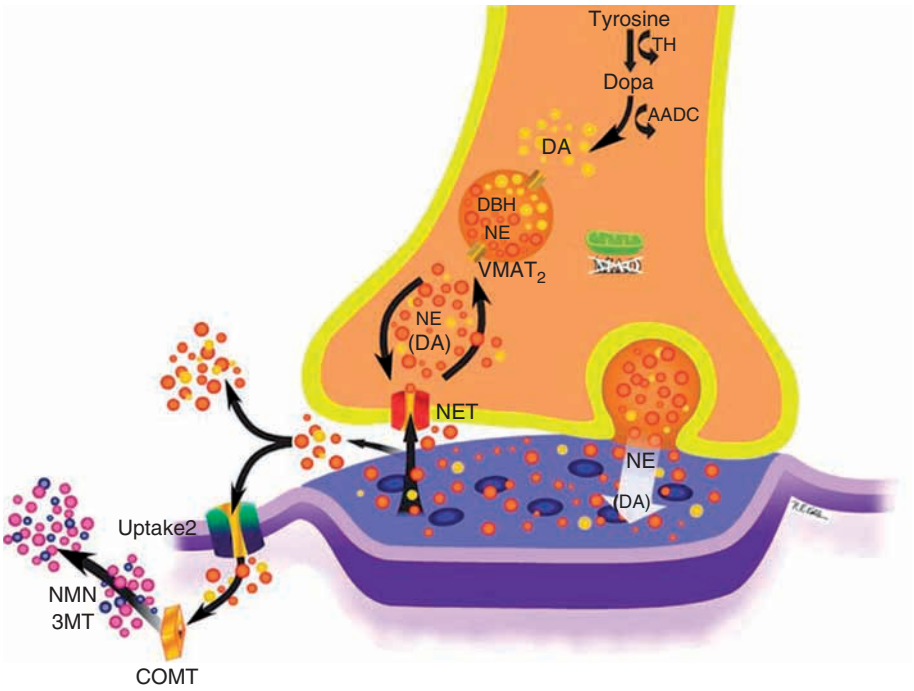

Figure 3 Monoamine oxidase (MAO) deficiency. Norepinephrine (NE) and 5-hydroxytryptamine are metabolized by MAO-A, while dopamine (DA), phenylethylamine, and benzylamine are broken down by MAO-B. Dopamine, tyramine, and tryptamine are substrates of both isoenzymes. Deficiency of only one gene product is reflected by the absence of its corresponding metabolites. Without oxidative deamination of NE by MAO-A in sympathetic nerves, for example, 3,4-dihydroxyphenylglycol will not be formed. Lack of both enzyme subtypes would result in the loss of both 3,4-dihydroxyphenylglycol and 3,4-dihydroxyphenylacetic acid. Metabolites of catechol-*O*-methyltransferase (COMT) will increase. (Other abbreviations as in **Figure 1**.)

by passive leakage into the cytoplasm. Historically, neurotoxicity has been explained by nonenzymatic autooxidation of cytoplasmic catecholamines and the formation of reactive quinones, free radical production, and formation of alkaloids via condensation reactions. The oxidative by-products of MAO, on the other hand, are deaminated catecholaldehyde metabolites. In fact, 3,4-dihydroxyphenylacetaldehyde (DOPAL) and 3,4-dihydroxyphenylglycolaldehyde (DOPEGAL) are formed in such a manner by deamination of DA and NE/EPI, respectively, by MAO and are highly reactive. Prompt aldehyde metabolism to deaminated acids by aldehyde dehydrogenase or to deaminated alcohols by aldose or aldehyde reductase normally neutralizes the deleterious effects. However, any cytoplasmic imbalance that interferes with vesicular uptake, increases vesicular leakage, or inhibits aldehyde metabolism theoretically favors accumulation of neurotoxins and impending cell damage. Centrally, this could be associated with neurodegenerative changes in dopaminergic neurons and parkinsonian-like movement disorders, whereas buildup of toxic NE metabolites in cardiac sympathetic nerves has been postulated to play a role in orthostatic hypotension due to sympathetic denervation of the heart. Unfortunately, we do not have experimental evidence which implicates all these mechanisms in human disease. If such mechanisms are found to be operative, then MAO inhibitors might reduce neurodegeneration caused by overproduction and/or reduced metabolism of cytoplasmic catecholaldehydes.

MAO inhibitors have long been used in the treatment of depression, and the significance of MAO polymorphisms in psychiatric illness has been widely investigated. Deficiency of either isoenzyme alone or the combination of both will yield characteristic neurochemical profiles. One MAO-A mutation found in a Dutch pedigree yields no active enzyme and is associated with mental retardation and impulsive–aggressive behavior in males. Follow-up animal models have also revealed an aggressive phenotype in male mice lacking MAO-A.

In other human studies, significant variations in upstream coding sequences of the MAO-A gene were found across ethnic groups, consisting of 30 bp repeats present in 2, 3, 3.5, 4, or 5 copies. The polymorphism proved to be functional in regulating the MAO-A promoter, and alleles with a higher number of repeat copies (3.5, 4, and 5) demonstrate increased transcriptional activity. High-activity alleles have

been associated with panic and generalized anxiety disorders in females, neuroticism in males, and impulsivity and aggression in males. Low-activity alleles have been associated with male alcoholics with antisocial personality disorder. Finally, recent evidence suggests that increased MAO-A gene activity due to incomplete X chromosome inactivation may underlie the greater prevalence of depression in women compared to men. While these data strongly suggest an association of specific genetic mutations with psychiatric disorders, the reader should keep in mind that different groups often report contrasting results and resolution has proved difficult in some instances. Repeat studies and larger sample sizes are clearly necessary before a definitive association can be claimed for any given diagnosis.

MAO-A and-B deficiency (and Norrie disease gene) The MAO genes are situated next to the Norrie disease (ND) gene on the X chromosome. ND is a neurodevelopmental disorder characterized by congenital blindness caused by incomplete retinal vascularization. Approximately one-third of patients also suffer sensorineural deafness. Mental retardation is not uncommon, but the degree of severity is variable. Patients with atypical ND, by contrast, exhibit profound mental retardation, autistic-like behavior, hypogonadism, atonic seizures, altered peripheral autonomic function, and aberrant biogenic amine metabolism. Such severity is ascribed to large deletions encompassing the MAO and ND gene loci together. Combined MAO-AB deficiency results in >93% reductions in plasma DHPG, MHPG, and VMA, along with elevated plasma normetanephrine and metanephrine resulting from increased catechol-*O*-methyltransferase (COMT) metabolism. Individuals lacking the ND and MAO-B genes, on the other hand, show no signs of mental retardation or disruptive behavior. MAO-B deficiency is characterized by undetectable levels of platelet MAO-B activity and increased urinary phenylethylamine.

Catechol-*O*-Methyltransferase

Most extraneuronal catecholamine metabolism is accomplished by COMT in the liver and kidney. In addition to catecholamines, COMT also metabolizes L-DOPA, DHPG, DOPAC, and catechol estrogens. Conversion of catechols to their O-methylated metabolites requires *S*-adenosyl-L-methionine as a methyl donor. Two isoenzyme forms of COMT exist, one soluble and the other membrane bound. Soluble COMT is present in most tissues and has higher activity levels compared to the membrane-bound form. The main exception is the adrenal medulla, in which membrane-bound COMT is most prevalent. Pheochromocytomas, catecholamine-producing tumors at adrenomedullary and extraadrenal sites, contain vast amounts of membrane-bound COMT and generate high concentrations of O-methylated metabolites, primarily metanephrine and MHPG. Diagnosis is best accomplished by assaying plasma free metanephrines.

A functional polymorphism of the COMT gene (Val158Met) has been identified that causes amino acid substitution. The wild-type Val allele is associated with high enzyme activity, while the Met substitution produces three- to fourfold lower activity. Globally, there is wide allelic variation, with persons of European descent having the greatest heterozygosity compared to those of African or Asian heritage. Specifically, Europeans have nearly equal frequencies of the two alleles, whereas the majority of populations in other parts of the world express the high-activity allele. Furthermore, high-activity COMT is primarily found in nonhuman primates, implying it is the ancestral allele in humans.

A large genetically homogeneous population-based study conducted in Norway recently revealed an association of the Val/Val genotype with systolic BP $\geq$140 mm Hg compared to either Val/Met or Met/Met genotypes. Although the underlying mechanism was not accounted for, the authors postulate that altered dopamine catabolism may be involved. Dopamine-mediated natriuresis is believed to be important in long-term control of BP. Therefore, elevated COMT activity in the kidney would likely increase dopamine breakdown, thus increasing sodium levels and, ultimately, BP.

The COMT polymorphism has been extensively investigated in anxiety disorders. Despite early promising findings, later studies revealed conflicting results and failed to replicate a significant association of this polymorphism with various subtypes of anxiety, including obsessive–compulsive, general anxiety, and panic disorders, or phobias.

A secondary role for COMT is methylation of 2-OH and 4-OH catechol estrogens. Transformation to methoxy estrogens lowers the risk of breast cancer associated with estrogen-mediated DNA adduct formation or oxidative damage arising from catechol estrogen quinone–semiquinone redox cycling. Cancer risk examined by *in vitro* analysis of the COMT Val → Met polymorphism demonstrated that the low-activity Met variant produced two- to threefold lower levels of methoxy estrogens in breast cancer cells compared to the wild-type variant. Therefore, estrogenized women with lower COMT activity may be at increased risk for breast cancer due to elevated tissue concentrations of prooxidative 4-OH catechol

estrogens, and lower levels of 2-methoxyestradiol, an antiproliferative metabolite.

Reuptake

Norepinephrine Transporter Dysfunction

Norepinephrine is rapidly recycled by the presynaptic neuron for rerelease. Most NE is removed from the synaptic cleft by active transport via the norepinephrine transporter (NET). The NET is a member of Na^+/Cl^--dependent transporters sharing similar amino acid sequences and pumping mechanisms. Transporters were historically considered to be static proteins intrinsic to the cell surface. Recent evidence, however, has revealed that NET is usually tethered to neuronal vesicles until transmitter release, and that phosphorylation is crucial to cell surface expression. NET is dynamically regulated by both intra- and extracellular signaling molecules, as well as targeted by a host of long-term therapeutic agents and drugs of abuse. Indeed, NET expression is recognized as a means of neuronal adaptation to changes in transmission. Endogenous regulators of NET expression include norepinephrine, acetylcholine, γ-aminobutyric acid (GABA), angiotensin, insulin, natriuretic peptide, nerve growth factor, estrogens, and nitric oxide.

A substitution mutation in the NET gene was discovered in a large family in which affected individuals had orthostatic tachycardia and postural tachycardia syndrome (POTS). The A457P mutation resulted in decreased plasma membrane concentrations of NET in heterozygous individuals. In fact, the altered protein was shown to bind native transporter and reduce its trafficking to the cell surface, making this a dominant-negative mutation. (General NET deficiency is shown in Figure 4.) Analysis of 254 additional POTS patients has not yielded similar functional NET mutations to account for diminished NE reuptake. Therefore, POTS remains a dysautonomia of heterogeneous etiology, with NET deficiency a very rare etiology.

POTS is an often debilitating autonomic disorder characterized by excessive tachycardia on standing, elevated plasma NE spillover, and presyncopal/syncopal episodes. Affected individuals are young and primarily women. In addition to syncopal symptoms, common complaints include dizziness, visual changes, head and neck discomfort, poor concentration while

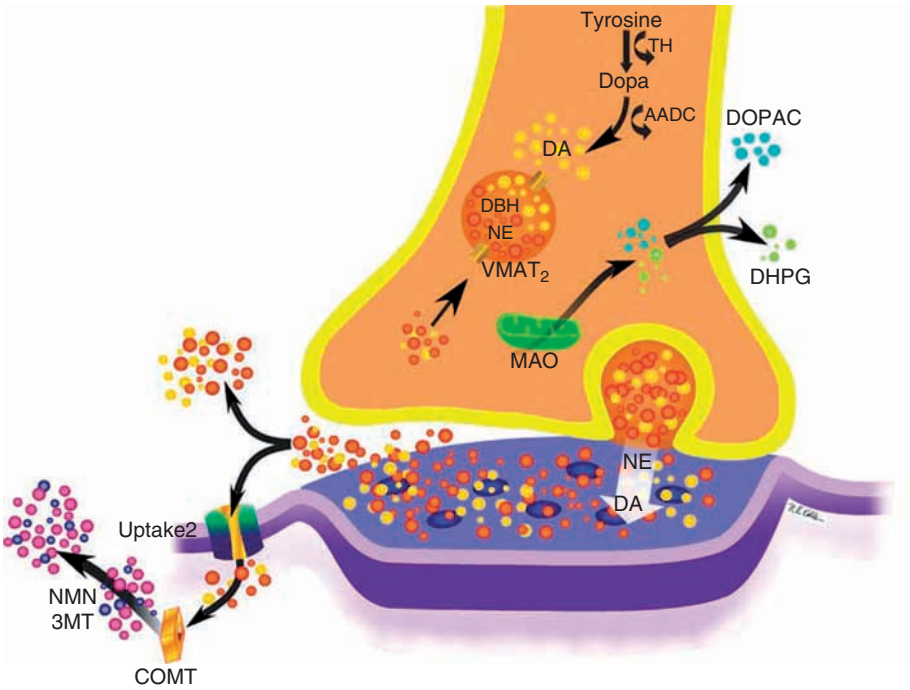

Figure 4 Norepinephrine (NE) transporter deficiency. Residual NE and dopamine (DA) are removed from the synaptic cleft by active transport through the NE transporter. The absence of NE transporter results in an increased concentration of catecholamines in the synapse, prolonging noradrenergic signaling and elevating levels of plasma NE and DA. Because monoamine recycling does not occur, conversion of DA to NE is increased. Continuous neurotransmission can deplete intraneuronal stores of NE, resulting in greater release of precursor and decreased production of 3,4-dihydroxyphenylglycol (DHPG). Metabolites of catechol-*O*-methyltransferase (COMT) are increased. (Other abbreviations as in Figure 1.)

standing, fatigue, palpitations, tremor, and anxiety. In familial POTS, levels of plasma NE increase approximately twice as high as in normal individuals on standing. DHPG concentration is decreased supine and upright, thereby lowering the DHPG:NE ratio. This is in contrast to idiopathic POTS, in which NE spillover is similar to that in normal individuals, both supine and standing.

Numerous naturally occurring single-nucleotide polymorphisms (SNPs) have been identified in the NET gene; these SNPs alter protein synthesis or maturation, transport mechanics, antagonist interactions, and kinase-mediated regulation. Many were discovered in cardiovascular phenotypes, and the role of similar functional mutations has been examined in psychiatric profiles. One such case is an A → T polymorphism at –3081 upstream of the transcription initiation site. The T allele was shown to decrease NET promoter function by allowing the binding of transcriptional repressors. A significant association of the mutation was demonstrated with attention-deficit/hyperactivity disorder (ADHD) in a relatively small sample size, but more extensive investigation is needed to confirm these findings.

Although detection of functional NET mutations has thus far been limited, pharmacological NET inhibition is very common. Many antidepressants and medications used in treating attention-deficit disorders block or alter NET transport. This includes first-generation tricyclics as well as newer specific NET inhibitors (e.g., atomoxetine or reboxetine). Pharmacological NET blockade can mimic an orthostatic tachycardia phenotype in susceptible normal subjects. Yohimbine, a central α-2 antagonist that increases synaptic norepinephrine, can also cause orthostatic tachycardia.

Conclusion

In summary, neurotransmitter disorders occur when there is insufficient or exaggerated neurochemical transmission, and can be due to either direct or indirect causes. For example, genetic mutations may affect essential enzymes directly in the synthesis pathway, or indirectly by limiting cofactor function. Multiple mechanisms exist for the clearance and recycling of catecholamines, and gain-of-function or loss-of-function mutations that alter synaptic, cytosolic, or vesicular concentrations often underlie aberrant neurotransmission. Continued investigation of genotype–phenotype associations will provide a better understanding of the pathophysiology of neurotransmitter disorders, allowing new methods for diagnosis and adjustment of treatment options to meet individual needs.

See also: Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Clinical Testing; Autonomic Nervous System; Cell Culture: Autonomic and Enteric Neurons; Dopamine; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopamine Neurons: Reward and Uncertainty; Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Sympathetic Nervous System; Trace Monoamines and Receptors in Mammalian CNS.

Further Reading

Blakely RD, DeFelice LJ, and Galli A (2005) Biogenic amine neurotransmitter transporters: Just when you thought you knew them. *Physiology* 20: 225–231.

Blau N, Thony B, Cotton RRH, et al. (2001) Disorders of tetrahydrobiopterin and related biogenic amines. In: Scriver CR, Beaudet AL, Sly WS, et al. (eds.) *The Metabolic and Molecular Bases of Inherited Disease*, pp. 1725–1776. New York: McGraw-Hill.

Carrel L and Willard HF (2005) X-Inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434: 400–404.

Cox DW and Moore SDP (2002) Copper transporting P-type ATPases and human disease. *Journal of Bioenergetics and Biomembranes* 34: 333–338.

Daniel KG, Harbach RH, Guida WC, et al. (2004) Copper storage diseases: Menkes, Wilson's, and cancer. *Frontiers in Bioscience* 9: 2652–2662.

Eisenhofer G, Kopin IJ, and Goldstein DS (2004) Catecholamine metabolism: A contemporary view with implications for physiology and medicine. *Pharmacological Reviews* 56: 331–349.

Garland EM, Hahn MK, Ketch TP, et al. (2002) Genetic basis of clinical catecholamine disorders. *Annals of the New York Academy of Sciences* 971: 506–514.

Goldstein DS, Eisenhofer G, and Kopin IJ (2003) Sources and significance of plasma levels of catechols and their metabolites in humans. *Journal of Pharmacology and Experimental Therapeutics* 305: 800–811.

Hagen K, Pettersen E, Stovner LJ, et al. (2007) High systolic blood pressure is associated with Val/Val genotype in the catechol-O-methyltransferase gene. *American Journal of Hypertension* 20: 21–26.

Hahn MK, Mazei-Robison MS, and Blakely RD (2005) Single nucleotide polymorphisms in the human norepinephrine transporter gene affect expression, trafficking, antidepressant interaction, and protein kinase C regulation. *Molecular Pharmacology* 68: 457–466.

Keller NR and Robertson D (2006) Familial orthostatic tachycardia. *Current Opinion in Cardiology* 21: 173–179.

Mandela P and Ordway GA (2006) The norepinephrine transporter and its regulation. *Journal of Neurochemistry* 97: 310–333.

Oeltmann T, Carson R, Shannon JR, et al. (2004) Assessment of O-methylated catecholamine levels in plasma and urine for diagnosis of autonomic disorders. *Autonomic Neuroscience* 116: 1–10.

Palmatier MA, Kang AM, and Kidd KK (1999) Global variation in the frequencies of functionally different catechol-O-methyltransferase alleles. *Biological Psychiatry* 46: 557–567.

Robertson D, Biaggioni I, Burnstock G, et al. (eds.) (2004) *Primer on the Autonomic Nervous System*, 2nd edn., San Diego: Elsevier Academic Press.

Robertson D, Goldberg MR, Onrot J, et al. (1986) Isolated failure of autonomic noradrenergic neurotransmission: Evidence for impaired beta-hydroxylation of dopamine. *New England Journal of Medicine* 314: 1494–1497.

Sabol SZ, Hu S, and Hamer D (1998) A functional polymorphism in the monoamine oxidase A gene promoter. *Human Genetics* 103: 273–279.

Shannon JR, Flattem NL, Jordan J, et al. (2000) Orthostatic intolerance and tachycardia associated with norepinephrine-transporter deficiency. *New England Journal of Medicine* 342: 541–549.

Shih JC, Chen K, and Ridd MJ (1999) Monoamine oxidase: From genes to behavior. *Annual Review of Neuroscience* 22: 197–217.

Vincent S, Bieck PR, Garland EM, et al. (2004) Clinical assessment of norepinephrine transporter blockade through biochemical and pharmacological profiles. *Circulation* 109: 3202–3207.

Zahniser NR and Doolen S (2001) Chronic and acute regulation of Na^+/Cl^--dependent neurotransmitter transporters: Drugs, substrates, presynaptic receptors, and signaling systems. *Pharmacology and Therapeutics* 92: 21–55.

Zhu BT (2002) Catechol-O-methyltransferase(COMT)-mediated methylation metabolism of endogenous bioactive catechols and modulation by endobiotics and xenobiotics: Importance in pathophysiologoy and pathogenesis. *Current Drug Metabolism* 3: 321–349.

Relevant Websites

http://www.genetests.org – Medical genetics information resource (funded by the National Institutes of Health).

http://www.bh4.org – Tetrahydrobiopterin deficiencies research (Zürich Research Group).

Autonomic Nervous System: Metabolic Function

N E Straznicky, P J Nestel, and M D Esler, Baker Heart Research Institute, Melbourne, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The autonomic nervous system is a key neuromodulator of cardiovascular, metabolic, and other physiological functions in humans. As such, it provides an important vehicle by which the central nervous system (CNS) maintains homeostasis during periods of acute and chronic changes in the physiological state. Dysfunctions in autonomic regulatory systems may lead to disease. Abdominal obesity, characterized by the accumulation of visceral adipose tissue, is commonly associated with a constellation of risk factors for cardiovascular disease and type 2 diabetes, termed the 'metabolic syndrome.' Evidence shows that sympathetic nervous system (SNS) overactivity plays at least an important, if not pivotal, role in the pathogenesis and complications of the metabolic syndrome. The hyperadrenergic state of the metabolic syndrome may be mechanistically complex because of the involvement of metabolic, central, and reflex factors. A better understanding of the causes and consequences of autonomic dysfunction in the metabolic syndrome may facilitate efforts to prevent and treat this condition. This article provides a brief overview of the anatomical and physiological aspects of the autonomic nervous system as a prelude to considering the evidence showing a role for autonomic dysfunction in the metabolic syndrome.

Organization of the Autonomic Nervous System

There are two main effector or efferent limbs of the autonomic nervous system. (1) Sympathetic innervation is derived from preganglionic fibers whose cell bodies lie in the intermediolateral column of the spinal cord at the level of the thoracic and upper lumbar roots. These nerves synapse with postganglionic sympathetic neurons in the bilateral chain of sympathetic ganglia located close to the spinal cord. Postganglionic sympathetic nerves innervate almost all of the vital organs and tissues of the body by release of neurotransmitters from their varicosities. (2) The parasympathetic nervous system consists of preganglionic fibers that originate in the CNS, namely the midbrain, the medulla oblongata and sacral part of the spinal cord, and their postganglionic connections. The parasympathetic system has terminal ganglia very near or within the organs innervated. The neurotransmitter of all preganglionic autonomic fibers, all postganglionic parasympathetic fibers, and a few postganglionic sympathetic fibers is acetylcholine. Adrenergic fibers comprise the majority of postganglionic sympathetic fibers and noradrenaline is the primary neurotransmitter. The adrenal medulla can be considered as a modified sympathetic ganglion. It is cholinergically innervated and releases catecholamines (primarily epinephrine) directly into the circulation instead of having a postganglionic nerve. The effects of norepinephrine and epinephrine are the result of activation of adrenergic receptors and subtypes thereof (α_1, α_2, β_1, β_2, and β_3). There are two distinct cholinergic receptors, termed nicotinic and muscarinic.

Peripheral autonomic activation usually occurs as part of reflex mechanisms, which involve afferent signals and processing in the CNS. Information on the status of visceral organs is transmitted to the CNS through two main sensory systems: the cranial nerve (parasympathetic) visceral sensory system and the spinal (sympathetic) visceral afferent system. In addition, diverse hormonal signals supply information about the peripheral metabolic state to the brain, for example, insulin from the pancreas and leptin from adipose tissue. The hypothalamus is regarded as the principal CNS site for the integration of autonomic function important in the control of metabolism. These integrated patterns of response involve autonomic, endocrine, and behavioral components (Table 1). The lateral hypothalamic area and ventromedial hypothalamic nucleus (VMH) are recognized as the parasympathetic and sympathetic centers respectively, and are functionally important in appetite regulation, glucose, and lipid metabolism. The paraventricular nucleus regulates the outflow of sympathetic and parasympathetic innervation of visceral organs, including the liver, pancreas, and the adrenal glands.

Organs of Metabolism, Their Innervation, and Key Metabolic Processes

The autonomic nervous system regulates metabolic processes through both direct neural effects (Table 2) and hormonal effects. The effects of the SNS on glucose and lipid metabolism are mediated by circulating glucagon, epinephrine, and the direct sympathetic innervation of the liver, adipose tissues, and skeletal muscle. The effects of the parasympathetic nervous system on glucose and lipid metabolism are mediated by insulin and the direct parasympathetic innervation of the liver. In general, sympathetic stimulation

Table 1 Hypothalamic areas and function

Area		*Nucleus*	*Function*
Lateral			Sending parasympathetic signals to the body via autonomic nerves; anabolic response: glycogen synthesis, lipogenesis; decreased appetite.
Medial	VMH	Ventromedial hypothalamic nucleus	Sending sympathetic signals to the body via autonomic nerves; catabolic response: gluconeogenesis, lipolysis; increased appetite.
Paraventricular	PVN	Paraventricular hypothalamic nucleus	Receiving information from SCN and ARN; sending information to autonomic nerves and pituitary gland.
	SCN	Suprachiasmatic nucleus	Circadian clock of the brain; receiving afferent input from retina; sending some signals to the PVN.
	ARN	Arcuate nucleus	Sending signals to the PVN and other nuclei in the hypothalamus; sensitive to insulin, leptin and glucose in the blood; playing a key role in energy metabolism.

Adapted from Uyama N, Geerts A, and Reynaert H (2004) Neural connections between the hypothalamus and the liver. *The anatomical Record A* 280: 808–820.

Table 2 Responses of metabolic organs to autonomic nerve impulses

Organ system	*Sympathetic effects*	*Adrenergic receptor type*	*Parasympathetic effects*	*Cholinergic receptor type*
Liver	Glycogenolysis and gluconeogenesis (+++)	α_1, β_2	↑ Glucogenesis ↑ HISS release	M
WAT	↑ Glucose uptake	α_1	↑ Glucose and FFA uptake	
	Lipolysis (+++)	β_1, β_2, β_3		
	Inhibition of lipolysis	α_2	↓ HSL activity ↑ Release of leptin and resistin	
Skeletal muscle	Vasoconstriction	α_1		
	Vasodilation (++)[a] and ↑ glucose uptake	β_2		
	↑ Contractility	β_2		
	Glycogenolysis	β_2		
	↑ HSL activity			
Pancreas				
α cells	↑ Glucagon secretion	β_2		
β cells	↓ Insulin secretion (+++)	α_2		
	↑ Insulin secretion (+)	β_2		
Adrenal medulla	Secretion of epinephrine and norepinephrine			N; M (secondarily)

[a]β_2 vasodilation in skeletal muscle is mediated by epinephrine. Cholinergic receptors, nicotinic (N) and muscarinic (M); HISS, hepatic insulin sensitizing substance; HSL, hormone sensitive lipase; WAT, white adipose tissue.
Responses are designated + to +++ to provide an approximate indication of the importance of sympathetic and parasympathetic nerve activity.
From Flechtner-Mors M, Jonkinson CP, Alt A, Biesalski HK, Adler G, and Ditschuneit HH (2004) Sympathetic regulation of glucose uptake by the alpha1-adrenoreceptor in human obesity. *Obesity Research* 12: 612–620; and Romjin JA and Fliers E (2005) Sympathetic and parasympathetic innervation of adipose tissue: Metabolic implications. *Current Opinion in Clinical Nutrition and Metabolic Care* 8: 440–444.

produces catabolic effects while parasympathetic stimulation produces anabolic effects.

Liver

More than any other organ, the liver contributes to the maintenance of metabolic equilibrium, including glucose and lipid homeostasis. The liver is innervated by the splanchnic sympathetic nerves and vagal parasympathetic nerves. Sympathetic stimulation promotes hepatic glucose production by activating glycogenolysis in fed states and gluconeogenesis in fasted states. Glucose output is increased through rapid activation of the key glycogenolytic enzyme, glycogen phosphorylase, with a concomitant decrease in liver glycogen content. In addition, sympathetic stimulation increases the activity of the gluconeogenic enzyme phosphoenol pyruvate carboxykinase. Parasympathetic stimulation leads to hepatic glycogen synthesis through activation of the enzyme glycogen synthase and an inhibition of glucagon-induced

glucose output. Recent research suggests that hepatic parasympathetic nerves may also play a permissive regulatory role in the release of hepatic insulin sensitizing substance, a liver-derived factor that increases glucose deposition in skeletal muscle.

In addition to direct control, autonomic nerves can indirectly influence liver function via the hypothalamus-pancreas and hypothalamus-adrenal axes. Sympathetic outflow from the hypothalamus stimulates the release of catecholamines from the adrenal medulla and the release of glucagon from the pancreatic alpha cells. Parasympathetic outflow stimulates the release of insulin from pancreatic beta cells.

Adipose Tissue

White adipose tissue is the principal tissue for energy storage in humans. When energy needs cannot be met by circulating fuels or stored carbohydrate, lipid is mobilized through the process of lipolysis, a breakdown of triglycerides into glycerol and free fatty acids, to provide fuel to other organs and to deliver substrates to the liver for gluconeogenesis (glycerol) and lipoprotein synthesis (free fatty acids). Adipose tissue lipid metabolism is highly regulated by hormones (insulin and catecholamines) and other factors such as nutritional status (feeding, fasting) and exercise. Sympathetic nerve fibers innervate white adipose tissue and stimulate lipolysis. The catecholamines noradrenaline and epinephrine activate lipolysis via β_1-, β_2-, and β_3-adrenoceptors, initiating a chain of events that activates hormone sensitive lipase, the principal intracellular enzyme responsible for the hydrolysis of triglycerides. A counter-regulatory system causing inhibition of lipolysis is mediated via α_2-adrenoceptors. The lipolytic activity of catecholamines is highest in the visceral fat depot, followed by abdominal subcutaneous fat, and lowest activity occurs in peripheral subcutaneous fat tissue. The SNS innervation of white adipose tissue also regulates fat cell number, as noradrenaline inhibits fat cell proliferation *in vitro*. Insulin is the most potent antilipolytic hormone in adipose tissue. It regulates the anabolic actions of fat cells, by stimulating glucose uptake, free fatty acid uptake (via action of lipoprotein lipase (LPL) on circulating triglycerides), inhibiting lipolysis, reesterifying fatty acids to triglyceride, and possibly by stimulating *de novo* fatty acid synthesis (lipogenesis). The SNS also has stimulatory effects on glucose uptake in adipose tissue, mediated by β- and α_1-adrenoceptors.

LPL is a rate-limiting enzyme, located on the endothelial surface of adipose tissue and skeletal muscle that hydrolyzes the triglyceride component of circulating lipoproteins. Studies in humans indicate that the response of LPL to catecholamines is tissue specific, with no effect or an inhibition in adipose tissue, but stimulation in skeletal muscle.

Recent neuroanatomical studies in rats have demonstrated parasympathetic innervation of white adipose tissue. Parasympathetic input was found to increase adipose tissue insulin sensitivity and modulate endocrine function, namely the secretion of the adipokines leptin and resistin. Interestingly, there is differential innervation of visceral and subcutaneous fat depots by both sympathetic and parasympathetic neurons.

Skeletal Muscle

The complex action of catecholamines is evident in skeletal muscle. Here the hormones have catabolic effects, which lead to mobilization of energy through the breakdown of glycogen (glycogenolysis) and lipids (lipolysis) via β_2-adrenoceptors. Catecholamines also increase muscle contraction and blood flow. β-2 Vasodilation in skeletal muscle is mediated by epinephrine. Skeletal muscle is the principal site of insulinstimulated glucose uptake. There is also evidence of noninsulin-mediated pathways of glucose utilization, activated by increased sympathetic activity. Electrical stimulation of the VMH is accompanied by increased glucose uptake independent of insulin.

Adrenal Medulla

Epinephrine, released from the adrenal medulla, can also directly affect lipid, protein, and carbohydrate metabolism. During physiological (e.g., heavy exercise) and stressful conditions (e.g., hypoglycemia, cold), epinephrine can affect hepatic glucose production by increasing glycogen degradation. It can also stimulate lipolysis of adipose tissue and proteolysis of muscle. Epinephrine affects gluconeogenesis indirectly by increasing the availability of substrates such as glycerol.

Autonomic Regulation of Energy Balance

Energy balance is maintained via a homeostatic system involving both the brain and the periphery. A key component of this system is the hypothalamus. Hormonal signals reflecting the availability and demand for metabolic fuel are relayed via neurons in the arcuate nucleus of the hypothalamus. One neuron group expresses neuropeptide Y and another proopiomelanocortin. Activation of the former leads to increased food intake and decreased energy expenditure, while the opposite effect is elicited by activation of the latter. A large number of circulating hormones and metabolites modulate neuronal function. Under stable conditions, equilibrium exists between orexigenic peptides that stimulate feeding behavior (e.g., ghrelin) and anorectic peptides (e.g., leptin, insulin)

that function to restrict excessive body fat accrual by limiting food intake and upregulating SNS activity. The second main input for information relating to energy balance is the brain stem. Vagal afferents that are sensitive to gastrointestinal signals synapse onto and excite neurons in the nucleus tractus solitarus, causing satiation.

Linked to the effects on appetite are thermogenic effects that dissipate heat. Components of human energy expenditure include (1) the obligatory energy expended on basic cellular and physiological functions that require adenosine triphosphate (ATP) (resting metabolic rate); (2) the energy expended in physical activity (exercise and spontaneous physical activity); and (3) adaptive thermogenesis, or that component that is stimulated in response to external stresses such as cold or changes in the amount of food ingested (thermic effect of food). Since the demonstration that sympathetic activity is increased during overfeeding and decreased during starvation, the SNS has been recognized as a pivotal efferent system in the link between diet and thermogenesis. Adaptive thermogenesis requires an uncoupling between substrate oxidation and biological work (phosphorylation). It is mediated in brown adipocytes by a mitochondrial uncoupling protein (UCP-1), which allows protons to leak back across the inner mitochondrial membrane, with a resultant increase in heat production. In small rodents, noradrenaline acts via β-adrenoceptors to activate and increase expression of UCP-1. Although a number of human studies have linked polymorphisms of UCP-1 homologs (UCP-2 and UCP-3) with obesity or low rates of energy expenditure and/or fat oxidation, there is considerable uncertainty about their physiological roles and function in adaptive thermogenesis.

The Metabolic Syndrome

The metabolic syndrome consists of a constellation of factors that raise the risk of cardiovascular disease and type 2 diabetes, namely, visceral obesity, insulin resistance, glucose intolerance, dyslipidemia, hypertension, and a pro-inflammatory state. The global obesity epidemic has been associated with a striking increase in the prevalence of the metabolic syndrome. The age-adjusted prevalence in the US population is estimated at 24% and increases to 44% in adults aged $\geq$60 years. Three competing hypotheses have been advanced to explain the primary etiology of this complex syndrome. In all of these, sympathetic overactivity plays a fundamental role:

1. Insulin resistance, at a cellular or receptor level leads to hyperinsulinemia, sympathetic activation, and hypertension. Obesity, dyslipidemia, and impaired glucose tolerance are components of the syndrome.
2. Obesity, particularly visceral obesity, is the primary problem leading to secondary insulin resistance, sympathetic activation, and attendant features.
3. Increased sympathetic activity of central origin is the primary defect leading to insulin resistance and other metabolic and cardiovascular sequelae.

Evidence for Sympathetic Overactivity in Obesity and the Metabolic Syndrome

Direct intraneural recording of sympathetic nerve traffic by microneurography and the measurement of systemic and regional noradrenaline spillover using tracer methods are currently preferred techniques for the assessment of SNS activity in humans. Microneurography records postganglionic sympathetic nerve action potentials targeted at skeletal muscle vasculature (muscle sympathetic nervous system activity, MSNA). Multifiber recordings of 'bursts' of nerve activity synchronous with the heart beat, and more recently single fiber traces are generated. The use of arterial blood samples in the spillover technique allows an assessment of whole-body SNS activity, while regional venous sampling allows SNS activity to be determined in a variety of organs.

Application of these sensitive newer methods has yielded substantial evidence that human obesity is characterized by increased SNS activity. The magnitude of this increase is influenced by coexistent risk factors and body fat distribution. Furthermore, there is evidence to show that regional SNS activity is heterogeneous in the obese state. In normotensive obese subjects, sympathetic outflow to the kidneys and skeletal muscle vasculature are increased by up to twofold, outflow to skin and the hepatomesenteric circulation are unchanged, while the sympathetic outflow to the heart is reduced, with cardiac noradrenaline spillover being only about 50% of that found in age-matched lean controls. Additional augmentation in MSNA has been demonstrated in obese patients with hypertension or obstructive sleep apnea (OSA). Among the adiposity indexes, abdominal visceral fat is most strongly associated with sympathetic overactivity. MSNA is greater in individuals with central obesity than in those with peripheral obesity and significant positive associations between waist circumference, MSNA, and whole-body noradrenaline spillover have been reported.

Findings in metabolic syndrome subjects mirror those in obesity: higher 24 h urinary noradrenaline excretion, elevated whole-body noradrenaline spillover, and increased MSNA (including single unit

discharges), suggesting the occurrence of sympathetic overactivity.

Power spectral analysis of variations in heart rate is based on sophisticated mathematical models to determine the balance between sympathetic and vagal nerve activities in the heart. Low heart rate variability is a risk factor for coronary heart disease. These indirect measures of cardiac autonomic tone show that obesity and the metabolic syndrome are associated with reduced heart rate variability and a relative sympathetic predominance. One criticism of this technique, however, is that it seems to have poor correlation with measurements of noradrenaline spillover from the heart.

Potential Mechanisms Contributing to Sympathetic Activation

Several factors have been implicated in the sympathetic activation that occurs in obesity. Since obesity is a heterogeneous condition, the relative contribution of these factors may vary based upon other modifying biological factors, both genetic and environmental.

Hyperinsulinemia Insulin resistance, defined as a decreased ability of target tissues to increase glucose uptake in response to insulin, is accompanied by compensatory hyperinsulinemia. Animal, human, and epidemiological (e.g., The Normative Ageing Study) studies indicate that sympathetic overactivity and hyperinsulinemia are closely interrelated. According to the Landsberg hypothesis, the insulin resistance of obesity is a mechanism evolved to limit further fuel storage as muscle glycogen and adipose tissue triglyceride, and to increase metabolic rate, thereby restoring energy balance. Thus, hyperinsulinemia-mediated sympathetic nervous activation was viewed as an adaptive response to limit further weight gain. In support of this hypothesis, insulin infusion under euglycemic conditions has been demonstrated to enhance MSNA in healthy subjects and borderline hypertensives. However, there is some doubt as to whether the obese retain their sensitivity to the stimulatory effect of insulin on the SNS. Vollenweider observed a specific impairment of sympathetic neural responsiveness (MSNA) to physiological hyperinsulinemia in obese subjects. Moreover, euglycemic insulin infusion does not activate renal sympathetic nerves in obese insulin-resistant humans and fasting insulin and renal noradrenaline spillover are not correlated. An alternate explanation for the association of hyperinsulinemia and sympathetic activity is that sympathetic activation is the primary etiopathogenetic factor, with insulin resistance being a secondary phenomenon. This notion is discussed below.

Downregulation of β-Adrenergic receptors Several studies have indicated that the response of the SNS to various physiological (under- and over-eating, cold exposure) and pharmacological stimuli is blunted in the obese. Reduced *in vivo* lipolytic response to intraneural electrical stimulation and reduced β_2-adrenoceptor numbers in isolated subcutaneous adipocytes have also been reported in obese compared with lean controls. It is currently unclear whether adrenergic receptor defects lead to SNS overactivity or whether enhanced sympathetic tone downregulates β-adrenergic receptors, which would promote weight gain and obesity. A polymorphism at Arg16Gly of the β_2-adrenoceptor gene has been linked to SNS overactivity. In other studies, the Gly16 allele of Arg16Gly has been associated with blunted vasodilatory responses to β_2-agonist. Thus, it could be speculated that β_2-adrenoceptor polymorphisms might contribute to blunted β_2-adrenoceptor function and increased sympathetic activity.

Baroreflex impairment Impaired baroreflex modulation of heart rate and MSNA at rest, and following vasoactive stimuli, have been consistently demonstrated in obese and metabolic syndrome populations. Potential mechanisms that may influence baroreflex function in this clinical setting include reduced arterial distensibility, altered central integration of afferent vagal nerve traffic, or impaired cardiac muscarinic receptor function. However, since individuals with peripheral and central obesity have a similar degree of baroreflex impairment, the greater sympathetic activation of individuals with a visceral fat depot cannot be explained by this reflex mechanism.

Leptin Leptin is a 167-amino-acid protein hormone produced primarily by adipose tissue that acts on specific receptors in the hypothalamus to decrease appetite and increase energy expenditure. Based on findings in rodents, which showed that leptin administration increases catecholamine turnover in brown adipose tissue, kidneys, and hindlimb vasculature, it has been speculated that the sympathetic activation of obesity may be mediated in part by increased plasma leptin levels. Obesity is known to be associated with circulating hyperleptinemia, reflecting a high fat mass and partial resistance to leptin. It has been postulated that leptin resistance in obesity can be selective, with preservation of sympathetic responsiveness despite resistance to satiety and the metabolic actions of leptin. To date there are no definitive descriptions of the sympathetic actions of leptin in humans. We have previously reported a strong positive correlation between renal noradrenaline spillover and plasma leptin

in men of widely differing adiposity. In a subsequent study in overweight and obese metabolic syndrome subjects, plasma leptin was significantly associated with whole-body noradrenaline spillover, but not MSNA (Figure 1(a)). These results may reflect differences in regional sympathetic responses to leptin, in concordance with animal studies. Furthermore, animal studies indicate that regional sympathoactivation may be mediated by different neuropeptide pathways: with melanocortins mediating renal sympathoactivation and corticotrophin-releasing hormone mediating brown adipose tissue SNS activation.

Obstructive sleep apnea OSA is common in the obese and has been increasingly implicated in the progression of cardiovascular disease and the increase in blood pressure that results from heightened sympathetic drive. OSA is associated with a marked increase in MSNA not only in the obese but also in lean individuals. It is believed that this increase originates from the stimulation of peripheral chemoreceptors brought about by repetitive nocturnal hypoxemia. Nasal continuous positive airway pressure is the treatment of choice and has been shown to effectively decrease apneic episodes, improve insulin sensitivity, and increase noradrenaline clearance from the circulation, but was without effect on noradrenaline release.

Nonesterified fatty acid Abdominal obesity is linked to increased nonesterified fatty acid (NEFA) concentration and turnover, which are resistant to suppression by insulin. NEFA contributes to the metabolic aspects of the metabolic syndrome by impairing insulin-mediated glucose disposal. Based on epidemiological data that showed an association between NEFA levels and blood pressure (e.g., the Paris Prospective Study), it has been hypothesized that increased delivery of NEFA to the portal circulation could trigger a reflex increase in sympathetic tone. Studies in lean and obese human subjects have demonstrated that acute elevations in NEFA result in decreased baroreceptor sensitivity and an increase in α_1-adrenoceptor-mediated pressor sensitivity. However, in a recent study in healthy adults, whole-body and renal noradrenaline spillover did not change during fatty acid infusion, despite a threefold rise in NEFA.

Angiotensinogen A bidirectional interaction between the SNS and renin–angiotensin system, both of which are activated in the obese state, has been long appreciated. The expression of the angiotensin II precursor angiotensinogen is increased in visceral adipose tissue. Previous work in animals suggests that angiotensin II can trigger peripheral and central sympathoexcitatory effects. Evidence that angiotensin II activates the SNS in humans is less strong. A 3-month treatment with the angiotensin II receptor blocker candesartan in obese hypertensives was accompanied by a 21% reduction in MSNA. This was attributed to the concomitant improvement in insulin sensitivity rather than a direct sympathomodulatory influence, because in lean insulin sensitive

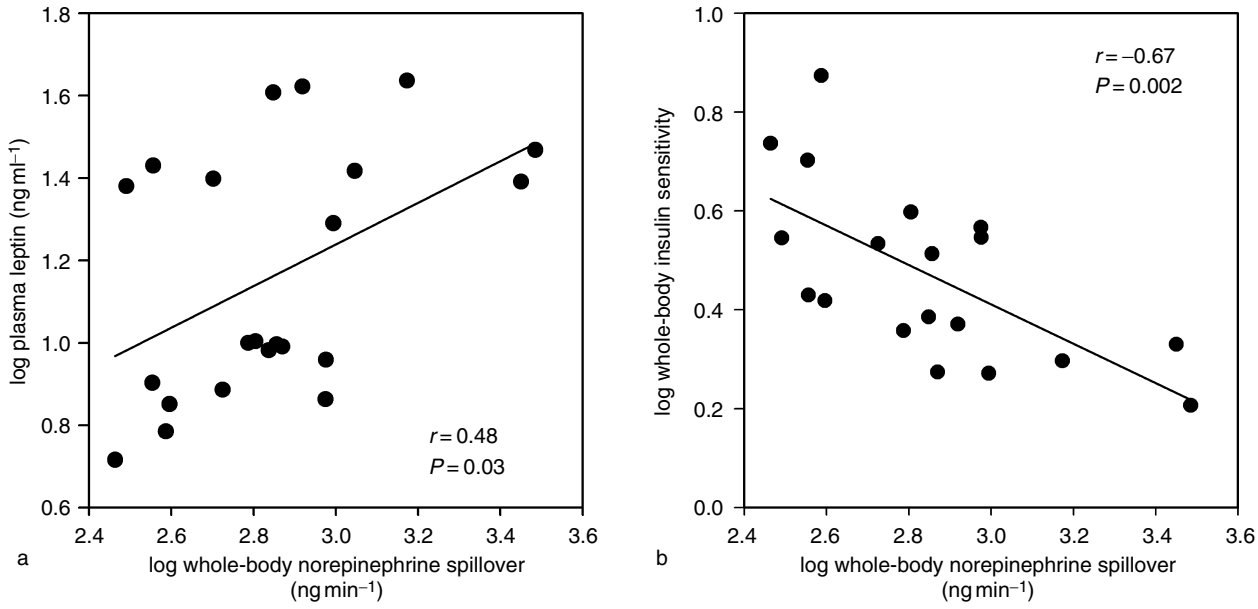

Figure 1 Univariate correlates of whole-body plasma noradrenaline spillover rate with fasting plasma leptin (a) and whole-body insulin sensitivity (b) in obese metabolic syndrome subjects. From Straznicky NE, Lambert EA, Lambert GW, et al. (2005) Effects of dietary weight loss on sympathetic activity and cardiovascular risk factors associated with the metabolic syndrome. *Journal of Clinical Endocrinology and Metabolism* 90: 5998–6005.

hypertensive subjects, angiotensin II receptor blockade did not result in sympathoinhibition.

Metabolic and Cardiovascular Consequences of Sympathetic Overactivity

Insulin resistance/glucose intolerance The notion that sympathetic overactivity precedes the development of insulin resistance is supported by longitudinal data in Japanese men. Masuo et al. reported that elevated plasma noradrenaline levels at baseline predicted future higher blood pressure readings, gain of weight, and higher insulin values in men followed for 10 years. The possible primacy of autonomic dysfunction is also supported by epidemiological data. Healthy participants in The Atherosclerosis Risk in Communities Study with reduced low-frequency power heart rate variability (a marker of decreased parasympathetic input) were at increased risk for developing type 2 diabetes.

Alterations in skeletal muscle blood flow, and hence the rate of delivery of both glucose and insulin, may be one mechanism by which sympathetic activation may modulate insulin sensitivity. Insulin sensitivity is also related to the number of slow-twitch muscle fibers (type 1) and capillary density. Experimental maneuvers that acutely stimulate endogenous noradrenaline release (e.g., lower body negative pressure and bilateral thigh cuff inflation) are accompanied by a concomitant reduction in insulin-mediated glucose uptake in skeletal muscle. These effects have been attributed to a reduction in skeletal muscle blood flow (via α-receptors) and/or a direct metabolic effect of noradrenaline. Conversely, mental stress which increases skeletal muscle blood flow (via β-receptors) improves glucose uptake during euglycemic hyperinsulinemia. Increased sympathetic drive also induces loss of capillary density and facilitates the shunting of glucose and insulin to less metabolically active skeletal muscle beds.

A further mechanism by which chronic sympathetic stimulation may cause insulin resistance is via the promotion of lipolysis. An increase in plasma-free fatty acid concentration can produce insulin resistance of glucose utilization by skeletal muscle and inhibition of the normal suppression by insulin of hepatic glucose production. Together, this could lead to glucose intolerance and insulin resistance. We have previously demonstrated a strong inverse relationship between whole-body plasma noradrenaline spillover rate and insulin sensitivity in obese metabolic syndrome subjects (**Figure 1(b)**).

Dyslipidemia The atherogenic dyslipidemia of metabolic syndrome obesity is characterized by increased liver secretion of very low-density lipoprotein (VLDL), high fasting and postprandial triglyceride levels, low high-density lipoprotein (HDL), and small dense low-density lipoprotein particles. Multiple factors contribute to the observed alterations in lipoprotein metabolism, including excess adiposity, insulin resistance, adipose tissue cytokines, and sympathetic overactivity. A major determinant of increased VLDL secretion in the metabolic syndrome is higher hepatic triglyceride content, derived in part from increased free fatty acid delivery from adipose tissue, and return of triglyceride rich lipoprotein remnants to the liver. Kinetic studies have shown that low HDL concentration is the consequence of increased catabolism, possibly secondary to triglyceride enrichment of the particles, and increased hepatic lipase activity.

Hypertension Elevated arterial blood pressure is common in metabolic syndrome subjects. Contributing mechanisms include (1) increased SNS activity, (2) impaired endothelial function with impaired vasodilation, and (3) increased renal sodium reabsorption, which would increase plasma volume and thus cardiac output.

Altered cytokine production There is some evidence to suggest that the SNS may regulate adipocytokine production. β-Adrenergic agonist induced secretion of interleukin-6 has been demonstrated *in vitro* in murine cultured adipocytes and *in vivo* in humans following isoproterenol infusion. Similarly, epinephrine, noradrenaline, and isoproterenol all induced a concentration-dependent increase in the production of tumor necrosis factor alpha in human mononuclear cells. Conversely, sympathetic antagonism with β-blocker treatment decreases the level of both of these cytokines. Adiponectin is another adipocytokine that is associated negatively with components of the metabolic syndrome, including obesity, insulin resistance, and dyslipidemia. In contrast to the aforementioned cytokines, β-adrenergic agonists reduce adiponectin mRNA in adipose tissue and adiponectin concentration in plasma.

Clinical Management of the Metabolic Syndrome

The underlying risk factors that promote development of the metabolic syndrome are overweight and obesity, physical inactivity, and an atherogenic diet. All current guidelines on the management of the metabolic syndrome emphasize that lifestyle modification (weight loss and physical activity) are first-line therapy. Several large clinical trials such as the Diabetes Prevention Program have found that lifestyle modification is the most efficacious strategy to prevent

progression to type 2 diabetes. In this study intense lifestyle changes were accompanied by a 58% risk reduction. Both weight loss and exercise have been documented to exert marked sympathoinhibitory effects and improve metabolic syndrome components. Drug treatment for associated risk factors may be indicated for some patients.

Weight Loss

We have recently reported that moderate weight loss (7% of body weight) is accompanied by a 43% reduction in whole-body noradrenaline spillover and an improvement in all metabolic syndrome components. It is likely that both the decrease in body fat and caloric restriction (negative energy balance) contributed to these changes. It is of note that only subjects who were insulin resistant at baseline observed a reduction in noradrenaline spillover after weight loss, despite the fact that both insulin sensitive and insulin resistant groups lost similar amounts of weight (Figure 2). This suggests that insulin resistance is a key driver of sympathetic activity in the metabolic syndrome.

Exercise

Regular aerobic exercise enhances caloric expenditure and is thought to have a synergistic effect on weight loss during a hypocaloric diet in obese individuals. In addition, exercise training provokes beneficial cardiovascular and autonomic adaptations, namely a reduction in blood pressure and sympathetic activity (preferentially involving renal sympathetic outflow) and improvements in insulin sensitivity and lipoprotein profile.

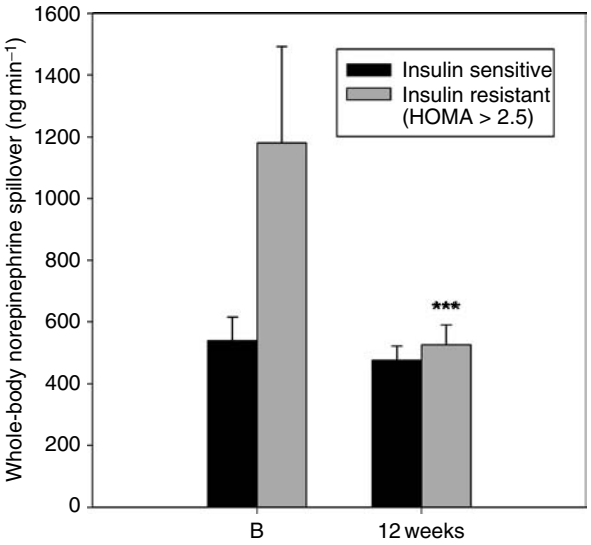

Figure 2 Whole-body noradrenaline spillover rate at baseline (B) and after a 12-week hypocaloric diet, in insulin-resistant and insulin-sensitive obese metabolic syndrome subjects. Insulin-resistant ($n=10$) and insulin-sensitive ($n=9$) subjects were defined as HOMA index ≥ 2.5 and <2.5, respectively. Weight loss was similar in the two groups ($6.4+1.1$ vs. $6.6+1.0$ kg). Data are shown as the mean + SEM. ***$P<0.001$ versus baseline values. From Straznicky NE, Lambert EA, Lambert GW, et al. (2005) Effects of dietary weight loss on sympathetic activity and cardiovascular risk factors associated with the metabolic syndrome. *Journal of Clinical Endocrinology and Metabolism* 90: 5998–6005.

Drug Treatment

Modulating sympathetic overactivity is a worthwhile therapeutic goal not only in the management of hypertension itself but also in the cluster of metabolic abnormalities with which it is associated. In choosing antihypertensive drug treatment, potential metabolic side effects need to be considered. Vasoconstricting β-blockers and diuretics can worsen insulin resistance and atherogenic dyslipidemia. β-Blockers also inhibit gluconeogenesis and lipolysis and are typically associated with some weight gain. Antihypertensive drugs for which concomitant beneficial effects on glucose metabolism and SNS activity have been demonstrated include the angiotensin II antagonist candesartan, and imidazoline-I_1-receptors agonists such as moxonidine and rilmenidine, which reduce central sympathetic drive. The latter class of drugs improve insulin-mediated glucose disposal, decrease plasma triglycerides, and do not increase body weight in obese patients with hypertension.

Insulin-sensitizing drugs represent a second-line therapeutic option for metabolic syndrome subjects in whom lifestyle interventions have been unsuccessful. Thiazolidendiones activate nuclear peroxisome proliferator-activated receptor-gamma and improve insulin sensitivity at key metabolic sites (skeletal muscle, adipose tissue, and liver). They also reduce glucose production and have a beneficial effect on the dyslipidemia associated with the metabolic syndrome. Furthermore, antihypertensive and sympathoinhibitory effects have been demonstrated after long-term treatment. Improvement in insulin sensitivity and reduced leptin expression are two potentially contributing mechanisms.

See also: Autonomic Disorders; Autonomic Failure; Autonomic Dysfunction: Drug-Induced; Autonomic Nervous System; Neuropathy: Metabolically-Induced; Sympathetic Nervous System.

Further Reading

Broberger C (2005) Brain regulation of food intake and appetite: Molecules and networks. *Journal of Internal Medicine* 258: 301–327.

Brunton LL, Lazo JS, and Parker K (eds.) (2006) *Goodman and Gilman's Pharmacological Basis of Therapeutics*, 11th edn. New York: McGraw-Hill.

Dulloo AG, Seydoux J, and Jacquet J (2004) Adaptive thermogenesis and uncoupling proteins: A reappraisal of their roles in fat metabolism and energy balance. *Physiology and Behaviour* 83: 587–602.

Eikelis N, Schlaich M, Aggarwal A, Kaye D, and Esler M (2003) Interactions between leptin and the human sympathetic nervous system. *Hypertension* 41: 1072–1079.

Esler M, Straznicky N, Eikelis N, et al. (2006) Mechanisms of sympathetic activation in obesity-related hypertension. *Hypertension* 48: 787–796.

Flechtner-Mors M, Jenkinson CP, Alt A, Biesalski HK, Adler G, and Ditschuneit HH (2004) Sympathetic regulation of glucose uptake by the alpha1-adrenoceptor in human obesity. *Obesity Research* 12: 612–620.

Grundy SM, Hansen B, Smith SC, Cleeman JI, and Kahn RA (2004) Clinical management of the metabolic syndrome. *Circulation* 109: 551–556.

Landsberg L (1986) Diet, obesity and hypertension: An hypothesis involving insulin, the sympathetic nervous system and adaptive thermogenesis. *Quebec Journal of Medicine* 236: 1081–1090.

Masuo K, Katsuya T, Fu Y, et al. (2005) β2- and β3-Adrenergic receptor polymorphisms are related to the onset of weight gain and blood pressure elevation over 5 years. *Circulation* 111: 3429–3434.

Masuo K, Mikami H, Ogihara T, and Tuck ML (1997) Sympathetic nerve hyperactivity precedes hyperinsulinaemia in young, nonobese Japanese population. *American Journal of Hypertension* 10: 77–83.

Nestel P (2003) Metabolic syndrome: Multiple candidate genes, multiple environmental factors – multiple syndromes? *International Journal of Clinical Practice* 134(Supplement): 3–9.

Nonogaki K (2000) New insights into sympathetic regulation of glucose and fat metabolism. *Diabetologia* 43: 533–549.

Puschel GP (2004) Control of hepatocyte Metabolism by sympathetic and parasympathetic hepatic nerves. *The Anatomical Record A* 280: 854–867.

Romjin JA and Fliers E (2005) Sympathetic and parasympathetic innervation of adipose tissue: Metabolic implications. *Current Opinion in Clinical Nutrition and Metabolic Care* 8: 440–444.

Straznicky NE, Lambert EA, Lambert GW, et al. (2005) Effects of dietary weight loss on sympathetic activity and cardiovascular risk factors associated with the metabolic syndrome. *Journal of Clinical Endocrinology and Metabolism* 90: 5998–6005.

Uyama N, Geerts A, and Reynaert H (2004) Neural connections between the hypothalamus and the liver. *The Anatomical Record A* 280: 808–820.

Vollenweider P, Randin D, Tappy L, et al. (1994) Impaired insulin-induced sympathetic neural activation and vasodilation in skeletal muscle in obese humans. *Journal of Clinical Investigation* 93: 2365–2371.

Autonomic Nervous System: Neuroanatomy

G Gabella, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

While the body moves around pursuing its behavior and interacting with the environment by motor and sensory mechanisms, its internal organs, such as the heart, stomach, glands, and bronchi, carry out their activities under the control and monitoring of the autonomic nerves. These constitute the autonomic nervous system (ANS), which is connected to, but relatively independent from, the rest of the nervous system. It has central components, part of the central nervous system (CNS), and peripheral components, spread throughout the body; all its parts, with the exception of solar plexus and enteric nerves, are bilaterally symmetrical. Somatic and visceral functions are distinct from each other but are connected and are part of a whole organism. The activities of viscera are spontaneous (or generated without voluntary intervention, based on reflexes, or 'automatic') and any sensitivity is poorly localized. Yet, these activities and the sensations generated from deep organs, including pain, are integrated with other bodily functions and reach to some extent our awareness. The internal state of the body is monitored and signaled to the brain.

Thus, the ANS controls the motility of the heart, gastrointestinal tract, urinary tract, and reproductive organs; the tone of blood vessels, lymphatic vessels, and airways; movements of the pupil and lens; the activity of the minute muscles of hairs; and the secretion of glands (from the large salivary glands and lachrymal glands to innumerable small glands in the gut, airways, genital organs, and skin) and the secretion of the adrenal medulla. Several of these activities are coordinated with somatic behavior. A major difference with somatic behavior is that visceral functions are not fully initiated and controlled (or dominated) by nerves, as is the case with somatic functions. The heart would continue to beat without nerves (although with a different pattern and a reduced range of adaptations); the intestine would still contract if all nerves were destroyed. Some visceral functions are in fact little influenced by nerves (e.g., the tone of large vessels or the motility of the uterus); other functions, however, are under tight control from autonomic nerves (e.g., ejaculation, motility of urethra and bladder, or movement of the intrinsic muscles of the eye).

General Plan

The central component of the ANS (within the CNS) consists of nuclei in the spinal cord, brain stem, and hypothalamus and nerve pathways connecting them to each other and to somatic nuclei. From the autonomic nuclei of the spinal cord and brain stem emerge nerve fibers that exit the CNS directed to peripheral organs.

The peripheral component consists of ganglia and nerves (ganglionated chains and plexuses) forming an extensive web that spreads from the spinal cord and brain stem to deep inside peripheral organs. Only the sympathetic chain is clearly recognizable by the naked eye; most of the other structures are visible only under a microscope.

The sensory component includes sensory neurons located in dorsal root ganglia and in some cranial nerve ganglia, mixed with somatic sensory neurons; these neurons project simultaneously to the spinal cord (dorsal horn) and to peripheral organs. There are many autonomic sensory neurons in ganglia associated with the vagus and glossopharyngeal nerves.

The peripheral component, however intricate and widespread, consists of nerve pathways leading from motor neurons in the spinal cord and brain stem to contractile and secretory elements (know as effectors) in peripheral organs. Whereas in somatic nerves there is a direct connection between motor neuron and muscle fiber, in autonomic nerves the pathway is interrupted by a ganglion. The autonomic motor neurons issue a fiber that terminates in a ganglion (hence the terms preganglionic neuron and preganglionic fiber). A second set of neurons within a ganglion, the ganglion neurons, which are driven by the preganglionic fibers that reach them, issue nerve fibers, known as postganglionic fibers, that reach the peripheral organ and terminate onto effector cells (muscle cells and secretory cells). Both preganglionic fibers and postganglionic fibers are very thin (0.2–1.0 μm in diameter) and are predominantly nonmyelinated: The conduction velocity is much slower (especially in postganglionic fibers) than in somatic nerve fibers.

Despite their diminutive diameter (and given the physical properties of cell membrane and the size of organelles such as vesicles and microtubules, these fibers are close to the physical limit of miniaturization) and the absence of a myelin sheath, postganglionic fibers can extend from a ganglion near the spine to an arteriole at the end of the foot or in the middle of the scalp.

The two-neuron pathway connecting the brain stem and spinal cord to muscular and secretory effectors is

thus relatively slow compared to the somatic motor neurons; by this arrangement, each preganglionic neuron (motor neuron) affects more than one ganglion neuron and each of these affects many effector cells by means of vast spreading of peripheral terminals. The interruption at the level of the ganglion, where the preganglionic fibers form synapses abutting ganglion neurons, adds further delay to transmission, and this matches the relatively slow contractile response of smooth muscle (which also carries large savings in energy cost).

Preganglionic Neurons

In the brain stem, the main bilateral groups of preganglionic neurons are the accessory oculomotor nucleus of Edinger–Westphal which projects to the ciliary ganglion; the superior salivatory nucleus projecting to the pterygopalatine and submandibular ganglia; the inferior salivatory nucleus projecting to the otic ganglion; the dorsal vagal nucleus projecting to ganglia in the trachea, esophagus, and gastrointestinal tract; and part of the nucleus ambiguus projecting to cardiac ganglia.

In the spinal cord, preganglionic neurons are grouped into symmetrical columns at the base of the ventral horn. The largest columns are part of the sympathetic pathways and extend from the last cervical level (C8) through the thoracic levels to the second lumbar level (L2). Other columns, which are part of the parasympathetic pathways, are located in the sacral segments (between S2 and S4).

Preganglionic neurons are all cholinergic and receive two main inputs: one from higher centers, including the micturition center in the pons, the cardiovascular centers in the brain stem, and several autonomic nuclei in the hypothalamus, and the other from the periphery via afferent (sensory) autonomic neurons in dorsal root ganglia. The latter input is either direct (monosynaptic reflex) or mediated by an interneuron.

Preganglionic fibers from preganglionic neurons gather into bundles and exit the cord (or brain stem) within ventral roots. Preganglionic neurons, ganglion neurons, and pre- and postganglionic fibers together form the sympathetic and parasympathetic pathways.

Sympathetic Pathways

The sympathetic pathways extend from the autonomic columns of the thoracolumbar spinal cord to the peripheral organs, including the adrenal gland, heart, and all the blood vessels of the body. They represent the autonomic thoracolumbar outflow and include the sympathetic ganglia.

The axons of preganglionic neurons of the sympathetic pathways (preganglionic fibers) reach the corresponding ventral roots and emerge as separate nerves (called white rami communicants), which in turn reach the paravertebral sympathetic chain (or trunk). The chain is made of ganglia and connecting nerve strands, and it extends bilaterally from the base of the skull to the sacral region in front and to the side of the vertebral bodies. Within the chain, preganglionic fibers spread out either cranially or caudally; with this arrangement, preganglionic fibers, which issue only from the middle part (the thoracic and some of the lumbar part) of the spinal cord, can reach all the ganglia of the chains, which extend for the full length of the spine. The largest ganglia are the superior cervical ganglion, which provides sympathetic innervation to organs of the head including cerebral blood vessels and skin, and the stellate ganglion, the main target of which is the heart.

Some preganglionic fibers reach the adrenal gland. Others travel further away from the spinal cord, form the splanchnic nerves, and reach sympathetic ganglia located in front of the abdominal aorta – the prevertebral ganglia. These ganglia, also known collectively as the celiac or solar plexus, are large and close to each other, and they are densely interconnected. They issue nerves directly to all the organs in the upper part of the abdominal cavity.

All the ganglia previously mentioned, which are an integral part of the sympathetic pathways, contain ganglion neurons that have fibers projecting to the peripheral organs. These are called postganglionic fibers, and they complete the sympathetic pathways.

Parasympathetic Pathways

Some parasympathetic pathways involve the brain stem and cranial ganglia, and some are centered on the sacral region of the spinal cord and the pelvic ganglia. In the brain stem, the main nuclei of the parasympathetic pathways (the cranial parasympathetic outflow) are the Edinger–Westphal nucleus, the superior and inferior salivatory nuclei, the dorsal vagal nucleus, and the nucleus ambiguus. The neurons of these nuclei issue axons and then enter the oculomotor nerve, the facial nerve, the glossopharyngeal nerve, and the vagus nerve, respectively. The oculomotor branch projects to the ciliary ganglion (its neurons innervate the iris and ciliary muscle, thus providing accommodation and pupil constriction). The facial nerve branch projects to the pterygopalatine and submandibular ganglia (their neurons innervate the lachrymal, submandibular, and sublingual glands), and the glossopharyngeal branch projects to the otic ganglion (its neurons innervate the

parotid gland). The fibers from the nucleus ambiguus and dorsal vagal nucleus enter the vagus nerve, of which they represent the motor component, and extend a long distance in the neck, thorax, and upper part of the abdomen. The fibers terminate synapsing on neurons in small ganglia close to the tracheal and bronchial muscles and the esophagus and in some enteric ganglia of stomach and intestine. From these minute intramural ganglia, postganglionic fibers emerge that innervate glands and smooth musculature.

In the spinal cord, preganglionic parasympathetic neurons are assembled into columns in the second, third, and fourth sacral segments. Their preganglionic fibers, predominantly cholinergic, project onto pelvic ganglia, whose postganglionic fibers innervate mainly the urogenital organs.

In some organs, typically the heart or the pupil, which receive both sympathetic and parasympathetic fibers, these exert an antagonistic effect. Many organs, however, are controlled predominantly by one or the other of the two pathways.

Autonomic Ganglia

These ganglia range in size from the largest, which in man are a few millimeters wide, to microscopic aggregates of only a few neurons. The largest ones – those of the sympathetic chain, the solar plexus and the pelvic plexus, and the named ganglia of the cranial parasympathetic pathways – have a capsule of connective tissue, an important protection against mechanical stimuli for structures situated close to large pulsating arteries and close to motile organs. The capsule is made of layers of connective tissue and is perforated by nerves entering and exiting the ganglion (also carrying the blood supply). The ganglion is in essence an aggregate of neurons (ganglion neurons). There are large and small neurons (partly related to the length which their axon has to travel to each the target organ and the extent of peripheral spreading). Each ganglion neuron has an axon (which becomes a nerve fiber after it is bundled up and associated with a glial support) that reaches the periphery, the postganglionic fiber. Ganglion neurons have dendrites spreading like trees around their cell body – long and highly branched arborization from large neurons (especially in the sympathetic ganglia of large animal species), short, barely branching arborization from small neurons (especially in small animal species), or even the absence of dendrites in certain ganglion neurons (some sympathetic neurons of the mouse). All ganglion neurons are fully cloaked by glial cells (satellite cells making the neuronal capsule), interrupted only where cell processes leave the cell and where terminal incoming fibers reach the neuron to synapse on it. The satellite cells, connected to each other by gap junction (hence, metabolically coupled), provide a tight barrier around the neuron and isolate it from the extracellular space of the ganglion and from other neurons.

In addition to ganglion neurons, satellite glial cells, and axons in transit and their glial sheath, autonomic ganglia contain fibroblast and other elements of connective tissue and some chromaffin cells. The ganglia are highly vascularized, in contrast to the nerve trunks.

The space between neurons is a busy place; at any one site there are hundreds of nerve fibers carrying impulses in one direction and in the opposite direction preganglionic fibers carrying impulses to the ganglion neurons, postganglionic fibers carrying impulses to the periphery, and many more fibers passing to and from other ganglia. Locally, the input from a preganglionic fiber is distributed to several ganglion neurons (divergence), and each ganglion neuron receives an input from several preganglionic neurons (convergence).

Postganglionic Fibers and Innervation of Viscera

Ganglion neurons of autonomic ganglia all issue an axon (which turns into a nerve fiber), and virtually all these fibers exit the ganglion directed toward a peripheral target organ, along different paths depending on the ganglion of origin. Some form discrete nerves (splanchnic, pelvic, and urinary), whereas others penetrate into somatic nerves; the mixed nerves of autonomic and somatic fibers thus formed reach the periphery, particularly the limbs and all the skin. Upon reaching the target organ, they branch extensively and make contact with muscle elements (mainly smooth muscle cells) and with secretory elements (glands).

Postganglionic fibers exert their effect on muscle and glands by releasing neurotransmitters that stimulate (and sometimes inhibit) contraction and secretion. The neurotransmitters are released from 'terminals' scattered along a substantial part of the terminal portion of the axons. Each ganglion neuron issues one fiber traveling to the periphery, which has many branches within the terminal organ, which in turn have many thousands of terminals. In the bladder muscle, for example, a few thousand ganglion neurons directly innervate many millions of muscle cells.

The relationship between varicosities (nerve endings) and muscle cells (or gland cells) varies from a close contact (a close and specialized junction) in some tissues (e.g., the bladder) to a loose relationship with a wide gap (some blood vessels). Chemical transmission from these endings – the release of the content of a few synaptic vesicles each time an ending

is depolarized – completes the activity of the efferent autonomic pathway.

The problem posed by the discrepancy between a small number of postganglionic fibers and a very large number of effector cells is solved by means of various structural arrangements. First, postganglionic fibers branch extensively inside the target organ (but not before). Second, nerve endings (which contain the neurotransmitters packed into axonal vesicles) are found not just as expansions at the anatomical end of each axonal branch: The entire terminal part of each branch (the last 100 μm of its length or more) presents bulbous expansion (or varicosities) that gives a beaded appearance to the axon and functions as nerve endings. Varicosities are expansions 1 or 2 μm wide, regularly spaced in a chain and separated by intervaricose segments that can measure as little as 0.3 μm in diameter. Each expansion is packed with axonal vesicles, and even if it is not the end of an axon, it functions like a nerve ending. The number of functional nerve endings of these so-called varicose axons is therefore 10–100 times greater than the number of anatomical end points of the postganglionic fiber.

Two other mechanisms assist in the uniform activation (or inhibition) of a large number of effector cells. These cells are often coupled to each other by gap junctions that allow the two cells to directly exchange metabolites and ions (which influence the electrical status of a cell). In addition, many effectors display some spontaneous activity (e.g., myogenic contractions of some muscles), and the action of nerves is a modulation and enhancement of some background activity. Lastly, some effectors are able to contract in the absence of nerves (e.g., the muscle of large arteries), and many effectors are affected by chemical agents that do not derive from nerves but may derive from endocrine glands or from secretion at adjacent sites.

Sensory Fibers

The ANS comprises an important sensory component. Its neurons lie in dorsal root ganglia, mixed with the vastly more numerous somatic sensory neurons. Each ganglion neuron issues an axon that splits, sending a branch centrally to the spinal cord and a branch peripherally to a peripheral organ. The peripheral branches travel together with sympathetic and parasympathetic fibers in mixed nerves and are distributed to the same tissues. Some of the central branches terminate onto the autonomic preganglionic neurons, in which case a reflex pathway is formed composed of an afferent and an efferent arm. Others synapse on interneurons, which in turn project to preganglionic neurons or form long pathways to higher centers of the CNS. The latter pathways convey sensations from the viscera, including pain.

Within the spinal cord there is a certain degree of convergence of visceral and somatic afferent nerve fibers, and some incoming fibers may synapse on the same interneuron, despite their different origin. This anatomical feature may explain the phenomenon of referred pain – that is, pain stimuli originating from a peripheral organ that are perceived by the subject as if they were coming from the surface of the body.

Enteric Nerves

Enteric nerves (or the enteric nervous system) are the set of autonomic nerves dedicated to the control of the activity of the stomach and intestine (including the lower part of the esophagus, anal canal, and gallbladder). They are located entirely within the wall of these organs (intramural nerves); they are linked to the sympathetic and parasympathetic outflows but are relatively independent of them. The enteric nerves consist of myriad small aggregates of neurons (enteric ganglia, which are quite different in size and structure from the other autonomic ganglia) connected by strands and forming a mesh or plexus, a relatively thin and flat structure, like a net with knots and meshes of a microscopic texture.

The most extensive and best known of these plexuses is the myenteric, or Auerbach's, plexus, located within the muscle coat, usually between circular and longitudinal muscle layers and extending continuously from the lower part of the esophagus to the stomach, small intestine, colon, rectum, and over the entire circumference of these organs.

The submucosal, or Meissner's, plexus is situated in the submucosa (the layer of dense connective tissue and blood vessels between muscle and mucosa), close to the inner border of the musculature and in proximity to the mucosa (in some animal species there is an additional component of this plexus nearer the mucosa).

There are differences in the pattern of the meshes in the plexuses, size of ganglia, and types of neurons in different parts of the gut, and these possibly reflect the differences in the activities of the different organs. The continuity of the plexuses over the entire length of the gut allows the activity in adjacent segments to be coordinated (even when they are separated by a sphincter); long-distance coordination, however, is achieved through reflexes involving extrinsic neurons (i.e., neurons located outside the gut wall) in the prevertebral ganglia. Still, the continuity of the plexus along the entire gut is compatible with the generation of different patterns of activity, especially motor activity, in different regions.

Invisible to the naked eye, and remarkably autonomous from other parts of the nervous system, the enteric neurons constitute an immense population of cells. Many thousands of neurons are found in every square centimeter of intestinal surface, and the total number in the gut amounts to many millions. Their spatial density is highest in the smallest mammals and decreases in animals of larger body size, in which connections between ganglia become longer. The total number of neurons, however, is highest in the larger animal species.

The ganglia are situated amid other tissues of the gut wall. They contain no connective tissue but consist exclusively of neurons with their processes, glial cells, sometimes other undifferentiated cells, and many nerve fibers in transit. The glia is highly developed, not as individual wrapping around axons and neurons (as in sympathetic ganglia) but as flattened cell bodies and expansions that permeate and fill all the spaces in the ganglion. Axons are usually not individually wrapped by a glial process (but are bundled together in direct contact with each other), and there are no myelinated fibers.

Many enteric neurons have axons and dendrites that may be very long but never leave the plexus, and they connect synaptically with other neurons situated upstream or downstream or elsewhere around the circumference or with neurons situated in deeper or more superficial ganglia. All these should be regarded as interneurons, and they are very abundant. The long axons of other neurons (motor or efferent neurons) exit the plexus and reach the musculature and the glands of the organ.

A small number of fibers exit the ganglia and the gut wall and project to the prevertebral ganglia. Also, some fibers from sympathetic and parasympathetic ganglia reach the intramural plexuses. The sensory fibers originate in the dorsal root ganglia, but there are also afferent fibers entirely intramural, supporting local reflexes.

Enteric neurons are a large population of nerve cells comprising several different categories of neurons. Due to the intricate packing of neurons and supporting glial cells, much research has been performed to identify nerve cell 'types' on the basis of cell appearance, incoming and outgoing connections, location, the presence of chemical features (neuronal markers), the types of transmitter released, its electrical properties, and its metabolic properties. There are descriptions that reveal the remarkable complexity of this nervous tissue.

An intriguing aspect of enteric neurons is that they seem to have highly specific distribution of chemical transmitters and markers (as in the endocrine system) but also have precise synaptic connections (as generally applies in the nervous system).

Central Components of the Autonomic Nervous System

The columns of preganglionic neurons in the spinal cord are the lowest level of a hierarchy of structures in the CNS that preside over the autonomic functions and their coordination. Higher up, in the brain stem, several nuclei and some less distinct groups of neurons constitute the interconnected reticular centers that regulate activities that are entirely or predominantly autonomic, such as cardiac, respiratory, and gastrooesophageal functions. These neurons receive an input from higher centers in the CNS, notably the hypothalamus, and an input from neurons (e.g., in the nucleus of the solitary tract) that receive sensory input from heart, intestine, and airway. The main outputs of the brain stem nuclei are (1) parasympathetic fibers to the heart, bronchi, stomach, etc. in the glossopharyngeal and vagus nerves and (2) sympathetic fibers to the sympathetic preganglionic neurons of the spinal cord. In certain systems, somatic and autonomic functions are tightly linked and integrated to each other (e.g., in respiration or vomiting).

The largest structure in the CNS controlling autonomic functions is the hypothalamus, located between the mesencephalon and the telencephalon, below the thalamus, bordered laterally by the internal capsule and the optic tract. Complex functions such as body temperature (homeostasis), circadian rhythms (the oscillations of certain bodily functions, mainly autonomic, with a 24 h period), feeding, drinking, and reproductive behaviors are regulated essentially from the hypothalamus. The organ exerts the highest and direct control both on the ANS and on the endocrine system; the hypothalamus is linked to the pituitary gland via the infundibular stalk and some of its neurons secrete the hormones oxytocin and vasopressin. Unlike in the rest of the brain, there is no blood–brain barrier in parts of the hypothalamus. In addition to projections to the thalamus, the forebrain, and the pituitary, many hypothalamic neurons project to cardiovascular centers in the brain stem (parabrachial nucleus, nucleus ambiguus, nucleus of the solitary tract, and dorsal motor nucleus of the vagus) and to the preganglionic neurons in the spinal cord.

The hypothalamus is divided, conventionally, into topographic zones that contain distinct nuclei as well as some diffuse groups of neurons. The periventricular zone includes the supraoptic nucleus, the preoptic nucleus (active in temperature control), and the suprachiasmatic nucleus (active in circadian rhythms). Located in the infundibulotuberal zone is the arcuate nucleus, and located in the posterior (or mammillary) zone are the medial mammillary nucleus (that has a major projection to the thalamus)

and the tuberomammillary nucleus. The lateral zone contains the lateral hypothalamic nucleus and the lateral tuberal nucleus. In the intermediate (or medial) zone are theparaventricular nucleus, the supraoptic nucleus, the mammillary body, the tuberomammillary nucleus, and the ventromedial nucleus. The ventromedial nucleus contains neurons sensitive to the plasma level of glucose and other metabolites; by integrating sensory inputs from the viscera, this part of the hypothalamus participates in the control of metabolism and food intake and of feeding behaviors in general. The preoptic nucleus contains neurons that are sensitive to small changes in the temperature of the blood circulating through the hypothalamus.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Disorders; Autonomic Failure; Autonomic Neuroimmunology; Autonomic Nervous System; Energy Homeostasis: Visceral Control; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Brading A (1999) *The Autonomic Nervous System and Its Effectors*. Oxford: Blackwell.

Furness J (2005) *The Enteric Nervous System*. Oxford: Blackwell.

Mathias CJ and Bannister R (2001) *Autonomic Failure. A Textbook of Clinical Disorders of the Autonomic Nervous System*, 4th edn. Oxford: Oxford University Press.

Robertson D (2004) *Primer on the Autonomic Nervous System*. San Diego: Elsevier.

Saper CB (1990) Hypothalamus. In: Paxinos G (ed.) *The Human Nervous System*, pp. 389–411. London: Academic Press.

Wilson-Pauwels L, Stewart PA, and Akesson EJ (1997) *Autonomic Nerves, Basic Science, Clinical Aspects, Case Studies*. London: Dekker.

Autonomic Nervous System: Ophthalmic Control

J Pintor, Universidad Complutense de Madrid, Madrid, Spain

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The eye is the sense organ that permits the detection of light owing to the existence of a sophisticated neuronal array, called the retina, which is responsive to photons. This ocular structure sends the acquired information through the optic nerve to central areas such as the lateral geniculate nucleus and the superior colliculus, on the way to the visual cortex.

The correct functioning of this complex system requires the coordination of several intraocular structures that ultimately permit the perfect focusing of images on the neural retina. Light has to pass through four different media before it reaches the retina: the cornea, aqueous humor, lens, and vitreous humor. Moreover, the composition and structure of some of these media can change due to several physiological mechanisms that are mostly controlled by the nervous system.

Although it is not yet clear whether some pathologies occur as a consequence of ocular nerve malfunction, it is apparent that disorders such as glaucoma may be a result of an imbalance between the formation and drainage of aqueous humor, due to a problem in the ocular innervation.

In this article, the importance of the nervous system controlling the iris and the ciliary body is reviewed.

The Iris and the Ciliary Body Are Located in the Anterior Uvea

The uvea is the part of the eye that is located between the scleral coat and the neuroepithelial cell layer. This region contains two main structures called the iris and the ciliary body (Figure 1).

The iris is the colored structure that works like a diaphragm regulating the amount of light passing toward the retina. It is important to control the quantity of light that arrives at the retina otherwise it will be overexposed and accurate vision of the object will be impossible. In brief, the iris works in a similar way to the diaphragm of a photographic camera. The iris has two main muscles, the iris dilator (dilator pupillae) and the iris sphincter (sphincter pupillae). These structures are innervated by the sympathetic and parasympathetic nervous systems, respectively, and work in coordination to control the size of the pupil in order to permit the best optic accuracy under different light conditions.

The ciliary body is located behind the iris. This structure is connected to the ciliary muscle which controls the lens shape in order to properly focus all the images in the retina in a process termed accommodation. Apart from the importance of the ciliary body in accommodation, the main task of this ocular structure is the synthesis of the aqueous humor. This transparent fluid serves as the source of nutrients for the avascular structures within the eye, such as the lens and the cornea. It is also important to remember that an increase in the synthesis of aqueous humor or a reduction in aqueous humor outflow can produce an increase in the intraocular pressure (IOP) leading to the pathology known as glaucoma. The ciliary processes are formed by two main epithelial cells types, pigmented and the nonpigmented ciliary epithelial cells, which are in contact with the aqueous humor. Under these cells vascular and neural plexuses provide nutrients and control the physiology of this ocular area.

Uvea Innervation: General Aspects

Sympathetic, parasympathetic, and sensory divisions of the peripheral nervous system are responsible for the uveal innervation, and they provide the three components regulating the physiology of this ocular area.

The Sympathetic Nerves and Its Neurotransmitters

Uveal sympathetic nerves are derived from the superior cervical ganglion and they innervate both the iris and the ciliary body. The sympathetic innervation originates in the spinal cord. From the segments C_8–T_3, the preganglionic neurons exit the spinal cord, contributing to thoracic nerves I–III, synapsing in the superior cervical ganglia. Postganglionic neurons in these ganglia project to the trigeminal ganglion and cavernous sinus. Some of the axons pass through the trigeminal ganglion and enter the ocular globe via the long posterior ciliary nerves. When in the eye, these axons spread and project to the iris muscle and to blood vessels. Other axons travel along the carotid artery, pass through the ciliary ganglion, and innervate the ciliary body as well as the uveal blood vessels.

The iris contains a dense population of adrenergic fibers, which innervate the iris dilator muscle. In human ciliary processes, the adrenergic nerves are located just beneath the ciliary epithelial cells, while in some animal models it seems that these nerve endings penetrate more and establish contacts with nonpigmented epithelial cells.

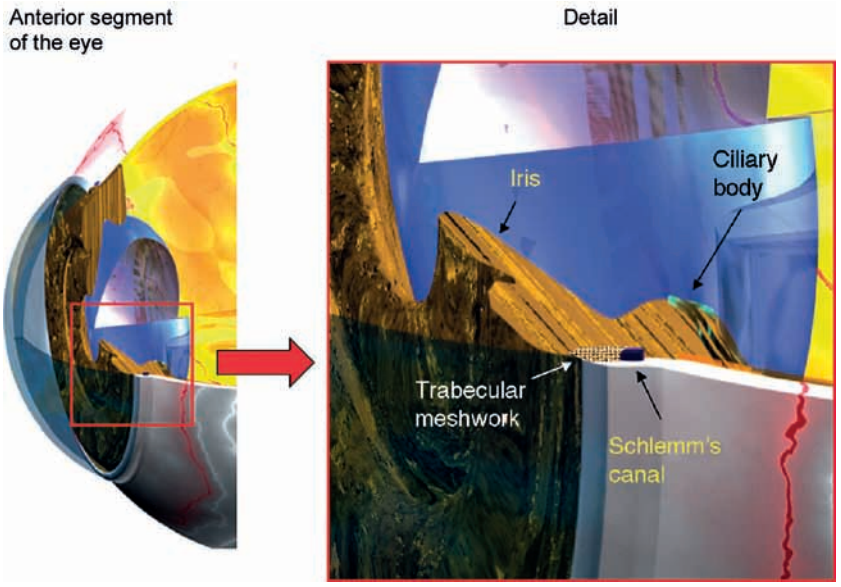

Figure 1 Three-dimensional models of the anterior segment of the human eye. In the selected area the main structures described in this article are depicted.

Noradrenaline (NA) is probably the best representative of sympathetic neurotransmitters in the eye. NA is synthesized in the classical pathway starting from the amino acid tyrosine. With the intermediate steps which produce the substances dihydroxyphenylalanine (DOPA) and dopamine, it is transformed into NA by means of the enzyme dopamine β-hydroxylase.

NA is an important transmitter in the iris since its concentration can be above 20 nmol g^{-1} tissue. This is in clear contrast with the concentration found in the aqueous humor, which is about 0.5 ng ml^{-1}. The origin of NA in the aqueous humor seems to be mainly from the iris since iridectomy reduces the concentrations of this transmitter in aqueous humor to below 80%. This does not mean that NA is inconsequential in the ciliary processes; on the contrary when analyzed in detail, NA is important since this transmitter is released after local electrical stimulation. The control of NA release in experimental models has been investigated. These studies demonstrate that α_2-adrenergic receptors, muscarinic, dopamine, and neuropeptide Y (NPY) receptors among others, can all inhibit NA release. In contrast, angiotensin can enhance NA release.

Dopamine is present at low concentrations in the aqueous humor (0.4 ng ml^{-1}) and its origin, in clear contrast to NA, does not come from the iris. It has been suggested that dopamine originates from the ciliary body.

5-Hydroxytryptamine (5-HT), also termed serotonin, is derived from the amino acid tryptophan. Its presence in the aqueous humor is very low although some authors have been able to identify some serotonergic nerve fibers in the ciliary body. The existence of receptors for 5-HT has been demonstrated in areas such as the ciliary body and the iris.

NPY is formed from 36 amino acids, and is amidated at its C-terminal. It is a neurotransmitter which runs in parallel with NA-containing fibers, and may be contained in the same nerve fibers as NA. The iris dilator muscle and the ciliary processes seem to present nerve fibers containing this transmitter.

Co-stored with NPY, opioids have been identified in postganglionic sympathetic nerves that project to the iris, and both sphincter and dilator muscles. To date, little is know about the role of these peptides in the iris, although receptors have been found in this tissue.

Parasympathetic Nerves and Its Neurotransmitters

The parasympathetic innervation that controls the iris and ciliary body comes from the preganglionic nerves located in the Edinger–Westfal nucleus which is located in the midbrain. Axons of these nerves leave the brain through the inferior oculomotor and oculomotor branches and synapse in the ciliary ganglion. The neurons which project to the iris and ciliary body are located in the periphery of the ciliary ganglion. More than 95% of all the fibers leaving this ganglion project to the ciliary body, while the remaining 5% reach the iris sphincter and the blood vessels. Parasympathetic terminals do not have a myelin coat and run in parallel with either the blood vessels or the muscle fibers.

It is clear that the main parasympathetic neurotransmitter is acetylcholine (ACh). Synthesized by the enzyme choline acetyltransferase, this enzyme has been extensively used to localize ocular nerves containing ACh as neurotransmitters. In particular, the presence of this enzyme has been described in the iris sphincter and ciliary muscle. The ACh concentration in the iris of some animals is between 100 and 200 ng mg^{-1} tissue. Nevertheless, one needs to take into account that ACh can be rapidly degraded by acetylcholinesterase (AChE). This enzyme rapidly reduces the local concentrations of Ach, thus explaining the acute effects of AChE inhibitors on miosis and accommodation.

ACh release is strongly controlled by other transmitters by means of prejunctional α_2-adrenoceptors, adenosine A_1, and opioid and nicotinic receptors (amongst others). Vasoactive intestinal polypeptide (VIP), a C-terminal amidated polypeptide containing 28 amino acids, is present in parasympathetic terminals coming from the pterygopalatine ganglion. Since these fibers mainly innervate blood vessels, it is possible to detect them in the iris stroma and ciliary processes. In the ciliary body it has been possible to detect receptors for this peptide, which are positively coupled to adenylate cyclase.

Sensory Innervation and Its Neurotransmitters

Starting in the trigeminal ganglia, the trigeminal neurons extend their axons centrally into the pons and peripherally, via the ophthalmic, maxillary, and mandibular branches to the target organs. The ophthalmic branch divides into frontal, lachrymal, and nasociliary nerves, the latter providing a major branch that travels to the eye with the ciliary nerves. The location of these nerves within the eye demonstrates that in the iris their location is mainly in the sphincter, as well as other ocular areas.

In the ciliary body, sensory nerves are widely distributed, being present in the ciliary processes, blood vessels, ciliary muscle, and scleral spur. The nature of these nerve terminals permits them to be considered as mainly mechanoreceptors and polymodal receptors.

A representative of the neurotransmitters present in sensory terminals is substance P (SP). Nerve fibers that are positive to this transmitter are located in the iris sphincter muscle as well as in the blood vessels of the iris stroma. It has also been possible to find nerves containing SP in the ciliary processes and ciliary muscle. SP is released after the application of certain compounds including capsaicin, bradykinin, or nicotine, which excite the nerve-ending triggering the release of the neurotransmitter. SP release in ocular structures can be modulated by presynaptic transmitters via the activation of α_2-adrenergic and κ-opioid receptors, and its release can be stimulated by means of nicotinic and adenosine receptors.

Calcitonin gene-related peptide is a polypeptide formed by 37 amino acids, which is often co-expressed with SP and therefore both are co-distributed in the iris and ciliary body.

Cholecystokinin (CCK) terminals are not very abundant but they can be found at the iris stroma and are also associated with blood vessels in the ciliary body. The 29-amino-acid polypeptide galanin is present as well, in terminals in the sphincter, dilator, and iris stroma.

Vasopressin (the antidiuretic hormone), a small, cyclized, C-terminal amidated peptide of nine residues, is the transmitter of some nerve fibers which arrive at the iris sphincter muscle. The application of exogenous vasopressin produces contraction in this iris structure.

The Neural Control of the Iris

Although the structure of the iris could be simplified to the iris dilator muscle, iris sphincter muscle, and the stroma, the main part of the control of the physiology of the iris is driven by the harmonization of sympathetic and parasympathetic terminals.

Sympathetic innervation of the iris sphincter muscle is mainly by NA. Neuropharmacological studies reveal that NA, acting through β-adrenoceptors, is able to induce relaxation in the sphincter by means of a mechanism which involves the activation of adenylate cyclase and the corresponding formation of cyclic AMP (cAMP). In some animal species the participation of α-adrenoceptors has also been quoted, but in this case mediating contraction.

The sympathetic regulation of the iris dilator muscle is mainly driven by α_{1B}-adrenoceptors, which contribute to the contraction of this muscle. This role has been demonstrated by means of selective antagonists that clearly abolish transmural contraction of the iris. The mechanism underlying the activation of this receptor seems to be phospholipase C (PLC) activation with the generation of inositol triphosphate (IP_3) and diacylglycerol (DAG). IP_3 releases Ca^{2+} from intracellular stores that together with DAG stimulates protein kinase C (PKC), which in turn phosphorylates myosin light chain (Figure 2).

The existence of β-adrenoceptors has been demonstrated in the iris dilator muscle. The effect that is mediated through this receptor is the relaxation of the muscle by means of the cAMP production. This apparent contradictory effect is supposed to exist in order to control the contractility, which is triggered by the α-adrenoceptor.

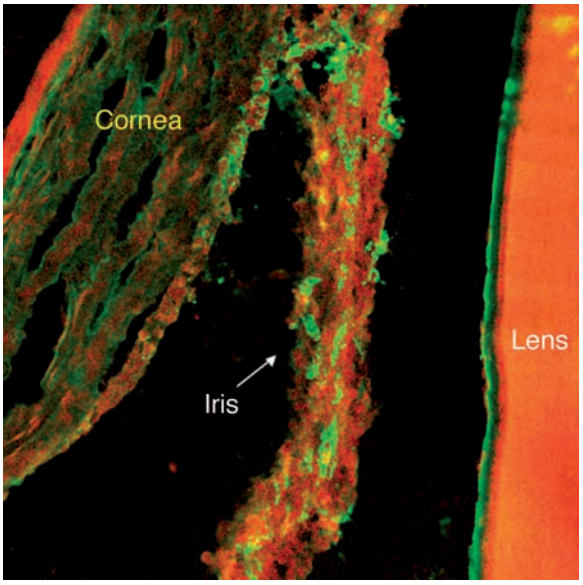

Figure 2 Immunohistochemical section of the iris labeled with an antibody which shows the nerve endings in green .

The neurotransmitter NPY seems to work closely with the α_1-adrenoceptor. On its own, NPY does not modify the physiology of the dilator muscle; nevertheless, it can potentiate the responses of NA acting via the α-adrenoceptor. Considering the mechanism involved in this potentiation it is possible that NPY, in some way, inhibits adenylate cyclase.

Parasympathetic control of the iris sphincter muscle is well known since muscarinic antagonists were able to block the miosis induced by light exposure. Atropine is isolated from the plant *Atropa belladonna*. The word belladonna comes from Italian, meaning beautiful lady; sixteenth-century Italian courtesans used to dilate their pupils by instilling belladonna into their eyes, making them more beautiful according to the perceptions of that time.

Indeed, the sphincter is contracted after treatment with ACh or its analogs. The receptors responsible for these actions are mainly of the muscarinic M3 subtype with fewer M2 receptors. M3 receptors are coupled to PLC and the corresponding generation IP_3 and DAG, while M2 receptors are negatively coupled to adenylate cyclase. Activation of both receptors would induce a more efficient contraction of the iris sphincter muscle.

The muscarinic contraction is a process which develops in two steps. There is one which is dependent on the release of Ca^{2+} from extracellular stores (triggered by IP_3), and a second in which calcium channels are involved. DAG activates PKC, and phosphorylation of calcium channels permits a higher cytosolic Ca^{2+} concentration that, together with PKC, permit a more efficient myosin light-chain phosphorylation.

VIP, which behaves as a co-transmitter in parasympathetic terminals, reverses the contraction triggered by cholinergic agents due to the production of cAMP, which probably inhibits the release of ACh from the parasympathetic nerve terminals.

The parasympathetic regulation of the iris dilator muscle is predominantly relaxant, in this case the effect being mediated by M3 receptors involving the activation of calcium-dependent adenylate cyclase. Also, the presence of M2 receptors has been defined on noradrenergic terminals; therefore, ACh controls the release of NA.

Sensory regulation of the iris sphincter has been identified by the use of selective antagonists such as SP receptor blockers. Capsaicin-induced SP release produces contraction in the iris sphincter muscle, but it has been demonstrated that the effect of SP can vary according to species, and in humans SP does not contract this muscle. These two behaviors are also corroborated by the putative activation of different SP receptors; while in contracting tissues SP receptor seems to be coupled to PLC activation, in humans the SP receptor is coupled to adenylate cyclase.

CCK seems to be more important than SP in humans since this transmitter is able to produce contractions in the human sphincter by a mechanism which involves CCK-A receptors.

The Neural Control of Aqueous Humor Formation

Aqueous humor is a transparent nutritional fluid that supplies nutrients to those intraocular structures that do not have a blood supply, especially the lens and the cornea. This fluid also provides the eye with an internal pressure, the IOP, that is necessary to preserve its shape.

Regulating IOP is a physiological process which depends on the right balance between the production and the drainage of the aqueous humor. Aqueous humor is produced at the ciliary processes while the drainage occurs in the iridocorneal angle via the trabecular meshwork, canal of Schlemm, and epiescleral veins. Equilibrium between production and elimination of the humor is a guarantor of ocular health. In contrast, imbalance brings complications that can lead to pathologies such as glaucoma, defined as an increase in IOP. When this happens it is probable that, to some extent, the elevation in pressure reduces retinal blood supply. If the pressure is strong it can even strangle the optic nerve. In both cases the result is a gradual blindness, starting from the peripheral retina towards the fovea (Figure 3).

As can be inferred, the control of the production and drainage of the aqueous humor is a tandem

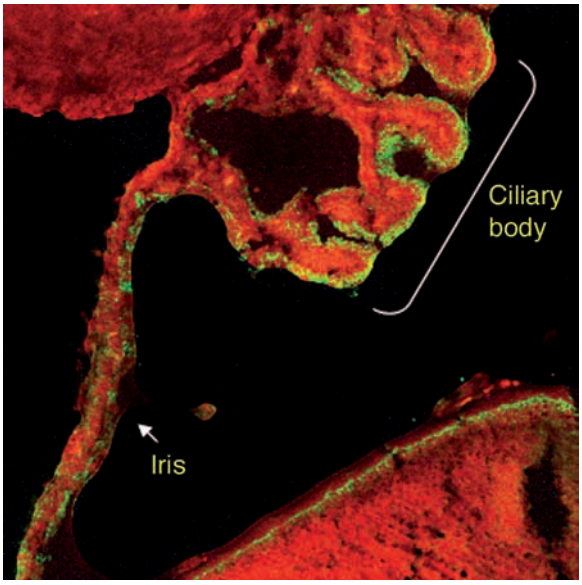

Figure 3 Immunohistochemical section of the iris and ciliary processes. The image shows the ciliary processes with the non-pigmented epithelium in green and the pigmented epithelium and stroma in red.

process that needs to be highly controlled to avoid the unpleasant consequences. The nervous system controls aqueous humor dynamics mainly by means of the sympathetic and parasympathetic pathways.

The Sympathetic Component

Stimulation or denervation are classical methods to study the involvement of the nervous system in the control of IOP. It is necessary to take into consideration that stimulation of the superior cervical ganglion will activate all the sympathetic pathways that arrive at ocular structures, making it difficult sometimes to understand the results obtained. It is, for example, typical that sympathetic denervation produces the release of transmitters from degenerating or degenerated nerve terminals, altering all the intraocular physiology. This may explain some contradictory results described in the literature.

Sympathetic stimulation by, for example, exciting the superior cervical ganglion reduces IOP mainly because it causes a reduction in uveal flow. This effect seems to be controlled by α-adrenoceptors producing vasoconstriction. Paradoxically, high frequencies of stimulation produce the opposite effect: an increase in IOP. It looks as if there is a selective release of neurotransmitters; at low frequencies NA is released while at high frequencies NPY is preferentially released.

Sectioning the superior cervical ganglion produces changes in the circadian variations of aqueous humor production. Also, this ganglionectomy produces an increase in IOP due to catecholamine release from degenerating sympathetic terminals.

Most of the drugs developed to reduce IOP in patients presenting with glaucoma is based on the use of more-or-less selective antagonists for adrenergic receptors, mostly β-adrenergic receptors. The presence of α- and β-receptors in the ciliary processes makes nonselective adrenergic agonists produce an increase in the secretion of the aqueous humor, an effect which is mainly blocked by β-adrenergic antagonists, such as timolol. Due to the circadian involvement of the sympathetic system, the ability of β-antagonists to reduce the production of aqueous humor is effective during the day but not during sleep.

Among the possible candidates responsible for these effects, β_2-adrenergic receptors appear as the best candidates. This receptor is present in the ciliary processes and is positively coupled to adenylate cyclase. Thus, it is cAMP that is the second messenger triggering a series of intracellular changes that , in the end, produce an increase of aqueous humor production and an increase in IOP.

α_2-Adrenergic agonists are also interesting agents to take into account when studying the involvement of the sympathetic nervous system and the control of IOP. It has been shown that clonidine among others can reduce the rate of aqueous humor production reducing concomitantly IOP. Biochemically speaking, α_2-adrenergic receptors are negatively coupled to adenylate cyclase diminishing the levels of cAMP stimulated by the corresponding activation of the β_2-adrenoceptors. There is also a prejunctional effect of α_2-adrenoceptors, diminishing the release of NA from the nerve endings and this may affect all the previously described physiology mediated by α- and β-adrenergic receptors.

Apart from NA, NPY seems to present an important role in the control of IOP. This peptide reduces the blood flow at the ciliary processes. Also, in this ocular structure the presence of receptors for NPY has been described. These receptors are negatively coupled to adenylate cyclase functioning in a similar way to α_2-adrenoceptors.

The Parasympathetic Component

The cholinergic influences on IOP are well known. Indeed, a vast pharmacology for the treatment of glaucoma has been developed, in which cholinomimetics clearly reduce IOP. Cholinergic agonists such as pilocarpine or inhibitors of AChE lower IOP by facilitating the outflow via the trabecular meshwork and Schlemm's canal. Apparently, this facilitation is a consequence of the ciliary muscle contraction and the subsequent pull of the tendon in the scleral spur which opens the corneal–scleral angle. This effect produces a wider spacing between the trabecular sheets, reducing the outflow resistance. The existence

of muscarinic receptors in the ciliary processes has been demonstrated. The best candidates are M2 and M4 receptors that mediate adenylate cyclase inhibition. Nevertheless, the results obtained in different experimental animals provide contradictory results on the direct effect of ACh in the ciliary epithelium.

Denervation of some cranial nerves such as facial produces a reduction in the uveal blood flow and consequently a reduction in IOP. The stimulation of this nerve causes just the opposite. These two effects are related with changes in the vasculature (vasodilatation) rather than local effects in the ciliary processes/trabecular meshwork. A transmitter which appears to participate in the vasodilatation is VIP, although intraocular application stimulates the production of the aqueous humor, since this transmitter stimulates adenylate cyclase.

The Purinergic Component of the Eye Autonomic Nervous System

Although purines have been described as co-transmitters of sympathetic and parasympathetic nervous system, little is known about the presence of purines in autonomic nerves of the eye. Evidences nevertheless indicate that they may be stored together with catecholamines in the sympathetic nervous system, and with ACh in the parasympathetic nervous system.

In any case, and before exploring the features of nucleotides in the eye, it is important to note that nucleotides and dinucleotide actions are due to the existence of receptors in cell membranes termed P2 and dinucleotide receptors, respectively. These receptors are divided in ionotropic (P2X and dinucleotide receptors) and metabotropic (P2Y) receptors. All these purinoceptors contribute to the regulation of important processes such as the vascular tone or neurotransmission in the periphery and central nervous system.

In the eye, and in particular in the rabbit iris dilator muscle, the phasic contractions of this muscle present two components, one adrenergic and other nonadrenergic The second component of the iris contraction is mimicked by ATP and the analog α,β-methylene ATP (α,β-meATP). ATP may play a prejunctional action in the iris, modulating the release of sympathetic neurotransmitters. It is not known which receptor is involved in the action of ATP in the iris, but some evidence suggests that it may be a P2Y receptor.

Nucleotides and dinucleotides have interesting actions on aqueous humor dynamics. Indeed, the presence of mononucleotides and dinucleotides in aqueous humor has been described in the relevant literature. This fluid contains measurable amounts of both mono- and dinucleotides. The most relevant adenine mononucleotides found in the whole aqueous humor were AMP, ADP, and ATP, presenting concentration values in the micromolar range. A detailed study has described the concentration of ATP in the immediate area where it is released (the ciliary epithelium), between 4 and 8 μM. Diadenosine polyphosphates have also been discovered, their concentrations being in the nanomolar range.

The effect of mono- and dinucleotides presents a clear pattern of modulation on IOP. Thus, the experiments performed by applying nucleotides to the eye allowed to classify the effect of these molecules under two main groups: on the one hand those elevating IOP and on the other those reducing it. Among the first, the synthetic analogs 2-methylthio ATP and ATP-γ-S presented a clear increase of IOP. This profile fits well with P2Y receptor pharmacology, although the slow destruction of these nucleotides by means of ectonucleotidases cannot be ignored.

On the contrary, β,γ-methylene ATP (β,γ-meATP) and α,β-meATP produced a clear and marked reduction in rabbit IOP. The hypotensive effect produced by these two mononucleotides was blocked by the P2 antagonist PPADS but was unaffected by the adenosine antagonist DPCPX.

β,γ-MeATP and α,β-meATP activate P2 receptors, presumably $P2X_2$ receptors, present in cholinergic terminals that are innervating the ciliary muscle. The activation of this $P2X_2$ receptor would generate an increase in ACh, which contracts the ciliary muscle, pulls the scleral spur, and opens the iridocorneal angle facilitating the drainage of the aqueous humor.

Diadenosine polyphosphates (ApnA) present a similar behavior as that observed for mononucleotides. Of all the tested dinucleotides, only Ap_4A presented a clear reduction in IOP, with the others presenting a hypertensive effect.

Recent results indicate that $P2Y_1$ receptors present in trabecular meshwork cells also modulate IOP. Nucleotides and dinucleotides such as Ap_4A can activate this receptor. The role of this purinergic receptor is to facilitate the drainage of the aqueous humor by changing the shape of these cells and reducing the hydrodynamic resistance of the aqueous humor evacuation.

The Retina

The retina is a very sophisticated neural structure which does not belong to the autonomic nervous system. It is considered as an extension of the brain that is specialized in responding to environmental light by means of a specialized group of neurons termed photoreceptors. The retina presents a clear

cellular architecture in which four other neuronal types help to transmit visual information to central visual areas. The organization of the retina presents two cell planes, one vertical consisting of photoreceptors, bipolar, and ganglion cells and another which is horizontal formed by horizontal and amacrine cells. All these cells define two synaptic regions, one formed with the photoreceptors, bipolar, and horizontal cells, and another with bipolar, amacrine, and ganglion cells.

When light stimulates the photoreceptor (cones and rods), a biochemical process occurs so the release of the neurotransmitter glutamate that is released during the darkness stops. Bipolar cells, which contain different types of glutamate receptors, respond to the lack of glutamate by either triggering an action potential and releasing more glutamate or inhibiting it. The presence or absence of glutamate can modify the activity of ganglion cells that projects their axons to central visual areas.

Horizontal cells as well as amacrine cells connect peripheral photoreceptors and peripheral bipolar cells with ganglion cells. These two types of neurons can present different transmitters since in some cases they are excitatory but in others they are inhibitory. Amacrine cells, for instance, can contain ACh or γ-aminobutyric acid (GABA) depending on its function.

Altogether, the retina is an extremely amazing neural array that permits the acquisition of light and its transmission, under the form of an electrical impulse, to arrive to central areas where it will be processed to construct sophisticated images.

The retinal pigmented epithelium (RPE) is part of what is call nonneural retina. This monolayer of cubic epithelial cells is in close relation with the photoreceptors since it is in charge of cone and rod turnover. The RPE is a noninnervated tissue, although some chemical signals can come from photoreceptors or from the choroids. Molecules such as SP and NPY released from the ciliary nerve in the choroids, and with a vasoactive action, may also change the physiology of RPE cells. On the other hand, it has been demonstrated that ATP can also modulate the activity of RPE cells by activating P2Y receptors. The ATP necessary to stimulate these receptors comes from the disc phagocytosis carried out by the RPE cells. So, although from neural origin since photoreceptors are neurons, ATP may not be considered in this case as a neurotransmitter in a classical way. One of the most interesting roles of ATP in the eye occurs in the RPE. During the pathology termed retinal detachment, a variable amount of fluid accumulates between photoreceptors and the RPE. This is potentially dangerous since the detachment of the neural retina drives to blindness. It has been demonstrated that the activation of P2Y receptors by ATP, UTP, and other dinucleotides (as well as NA) can reverse this process permitting the reabsorption of the fluid and therefore permitting the restoration of the right retinal conditions.

Conclusions

It is not easy to join all the experimental data coming from different animal models when describing the role of the nervous system in the iris and ciliary body. Very often, there are differences among species, which need to be taken into account, and that makes it difficult to unify concepts. This article aims to provide a general view of how the nervous system can control relevant processes within the eye, but one should be aware that there are cases in which the results presented here may not fit with the role of the nervous system in certain animal models.

See also: Autonomic Disorders; Autonomic Neuroplasticity and Aging; Diabetic Neuropathy; Retinal Pharmacology: Inner Retinal Layers; Retinal Development: An Overview; Sensory Ganglia; Vision: Light and Dark Adaptation.

Further Reading

Abdel-Latif AA (1989) Calcium-mobilizing receptors, polyphosphoinositides, generation of second messengers and contraction in the mammalian iris smooth muscle: Historical perspectives and current status. *Life Sciences* 45: 757–786.

Bergmanson JPG (1982) Neural control of intraocular pressure. *American Journal of Optometry and Ophysiological Optics* 59: 94–98.

Bognar IT, Wesner MT, and Fuder H (1990) Muscarine receptor types mediating autoinhibition of acetylcholine release and sphincter contraction in the guinea-pig iris. *Naunyn-Schmiedeberg's Archives of Pharmacology* 341: 22–29.

Burnstock G and Sillito AM (2000) *Nervous Control of the Eye*. Amsterdam: Harwood Academic Publishers.

Davson H (1993) The aqueous humour and the intraocular pressure. In: *Physiology of the Eye*, 5th edn., pp. 34–95. New York: Pergamon Press.

Fuder H and Muth U (1993) ATP and endogenous agonists inhibit evoked [^{3}H]-noradrenalin from field stimulated rat iris via A_1 and P2Y like receptors. *Naunyn-Schmiedeberg's Archives of Pharmacology* 348: 352–357.

Gallego R and Belmonte C (1974) Nervous efferent activity in the ciliary nerves related to intraocular pressure changes. *Experimental Eye Research* 19: 331–334.

Jumblatt JE (2000) Innervation and pharmacology of the iris and ciliary body. In: Burnstock G and Sillito AM (eds.) *Nervous Control of the Eye*, pp. 1–40. Amsterdam: Harwood Academic Publishers.

Kaufman PL, Wiedman T, and Robinson JR (1984) Cholinergics. In: Sears ML (ed.) *Handbook of Experimental Pharmacology*, pp. 149–191. Berlin: Springer.

Lütjen-Drecoll J and Rohen E (1989) Morphology of the aqueous outflow pathways in normal and glaucomatous eyes. In: Klein EA (ed.) *The Glaucoma*, pp. 89–123. St. Louis: CV Mosby.

Muramatsu I, Kigoshi S, and Oda Y (1994) Evidence for sympathetic, purinergic transmission in the iris dilator muscle of the rabbit. *Japanese Journal of Pharmacology* 66: 191–193.

Peterson WM, Meggyesy C, Yu K, and Miller SS (1997) Extracellular ATP activates calcium signaling, ion, and fluid transport in retinal pigment epithelium. *Journal of Neuroscience* 17: 2324–2337.

Rohen J (1964) Das auge uns seine hilforgane. In: von Möllendorff W and Bargmann W (eds.) *Handbuch der Mikroskopischen Anatomie des Menshen*, Bd III/2, pp. 189–328. Berlin: Springer.

Autonomic Nervous System: Respiratory Control

B J Undem, The Johns Hopkins University School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The autonomic nervous system assists in maintaining homeostasis of the internal milieu. In the respiratory system this aids in optimizing the function of gas exchange. Perhaps more important than this, however, is the role the respiratory nervous system plays in host defense. Quietly breathing while reading this article, you are inhaling about 600 ml of air with each breath, which amounts to nearly 10 l of air every minute. The air we breathe contains myriad potential harmful substances, the nature of which depends on the environment in which we are breathing. It is little wonder that a majority of nerves in the respiratory tract are afferent polymodal 'nociceptive' nerves designed to detect potentially noxious chemical perturbations. The afferent nerves communicate with the central nervous system, thereby evoking obvious defensive reflexes such as apnea, sneezing, and coughing. More subtly, these afferent nerves can assist in defending the airspace by initiating autonomic reflexes, resulting in bronchoconstriction, mucus secretion, and alterations in blood flow. The following sections provide a basic overview of the innervation of the respiratory tract with these functions in mind.

Upper Respiratory Tract (Nasal Innervation)

A schematic of the innervation of the nasal respiratory mucosa is depicted in Figure 1.

Afferent Innervation of Upper Airways

Sensory nerves innervating the nonolfactory respiratory nasal epithelium are derived from the ophthalmic and maxillial branches of the trigeminal ganglia (the sensory nerves in the olfactory epithelium are not reviewed here). Neurons situated in the trigeminal ganglia project both nociceptive capsaicin-sensitive C-fibers and mechanosensitive A-fibers to the nasal epithelium. Little is known about the specific consequences of stimulation of nasal A-fibers. Stimulation of nasal C-fibers with capsaicin, however, causes sensations of intense burning, and increases the parasympathetic autonomic outflow to the nasal passages and tear ducts.

Autonomic Innervation of Upper Airways

Parasympathetic Preganglionic parasympathetic innervation of the nasal airways originates from the facial nucleus of the brain stem and the superior salivatory nucleus. Preganglionic fibers follow the greater superficial petrosal nerve and the vidian nerve until they synapse on neurons clustered in the sphenopalatine ganglion (also known as the pterygopalatine ganglion). The preganglionic nerves are cholinergic, and upon stimulation cause nicotinic fast excitatory potentials on the postganglionic neurons.

Postganglionic fibers are distributed to the nasal mucosa via the branches of the posterior nasal nerve, where they ultimately innervate serous and mucous glands, arteries, veins, and arteriovenous anastomoses. Postganglionic parasympathetic nerves in the nasal mucosa release acetylcholine, and in some cases nitric oxide and vasoactive intestinal peptide (and related peptides), onto effector cells. The most extensively distributed of all of the muscarinic receptor subtypes in the nose appears to be the M3 receptor, which is responsible for cholinergic glandular secretions.

Nasal secretions are derived from serous/mucous glands and plasma extravasation from blood vessels in the mucosa. Most of the autonomically mediated secretory effects are due to glandular secretion and are inhibited by drugs that block cholinergic muscarinic receptors, indicating a major role of parasympathetic cholinergic nerves. Consistent with this hypothesis, experimental stimulation of the parasympathetic nerves causes glandular secretions rich in mucous glycoproteins, lactoferrin, lysozyme, secretory leukoprotease inhibitor, neutral endopeptidase, and secretory immunoglobulin A (IgA). In experimental animal models, direct stimulation of parasympathetic nasal nerves can also elicit vasodilatation of vessels in the nasal mucosa through both cholinergic and noncholinergic mechanisms. The noncholinergic component (atropine resistant) of the vascular response to parasympathetic nerve stimulation is likely due to either vasoactive intestinal peptide and/or nitric oxide release from the nerves, as both of these transmitters have been localized to parasympathetic nerves in the nose, and both are effective vasodilators. Although immunohistochemical evidence indicates that there are nitric oxide synthase-containing nerves innervating the human nasal mucosa, a nonadrenergic, noncholinergic (NANC) neural control of the nasal blood flow has not yet been explored.

Sympathetic Most of the sympathetic neural output to the human nose originates from cholinergic

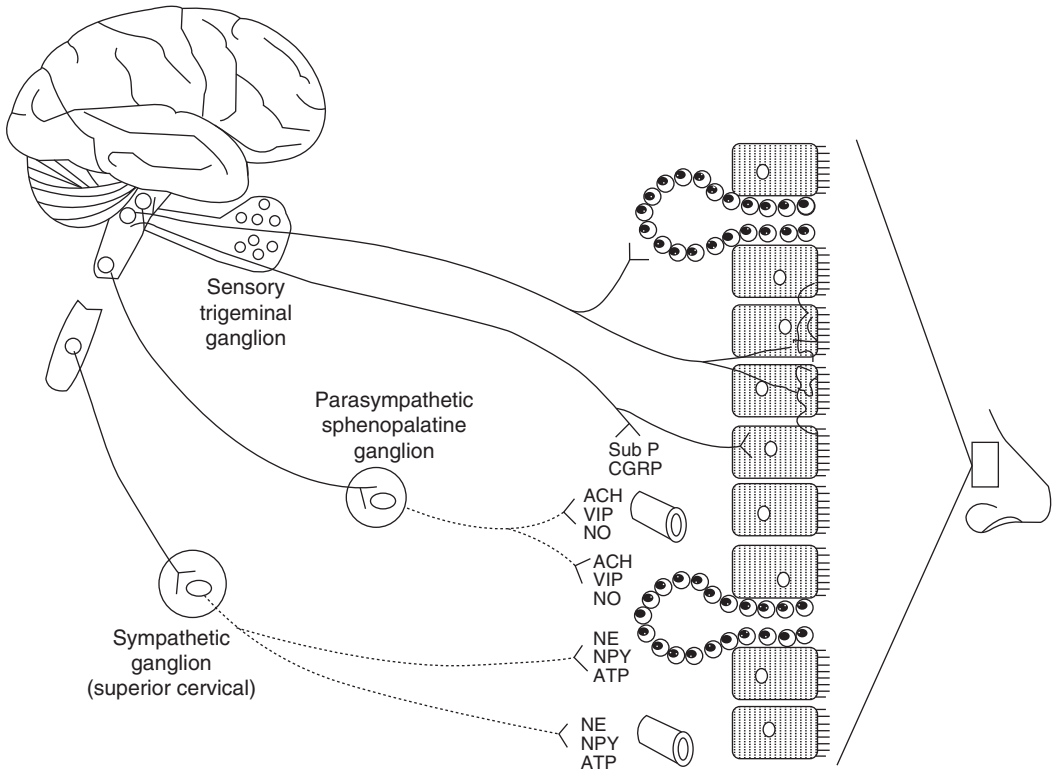

Figure 1 Schematic diagram of the innervation of the nasal respiratory mucosa. The transmitters noted are acetylcholine (ACH), adenosine 5′-triphosphate (ATP), vasoactive intestinal peptide (VIP), nitric oxide (NO), substance P (Sub P), calcitonin gene-related peptide (CGRP), norepinephrine (NE), and neuropeptide Y (NPY). These transmitters and neuropeptides are commonly found in the nasal mucosa, but should not be considered an exclusive list. The cholinergic and the nonadrenergic, noncholinergic (VIP and NO) branches of the parasympathetic system are depicted as arising from the same neurons, but it is also possible that the cholinergic nerves are distinct from the nonadrenergic, noncholinergic nerves, as has been observed in the lower airways.

preganglionic fibers in the thoracolumbar region of the spinal cord; these fibers synapse on neurons in the superior cervical ganglion. The postganglionic fibers form the petrosal nerve, which joins the greater superficial nerve to form the vidian nerve. The vidian nerve therefore contains both the parasympathetic and the sympathetic innervation to the nasal mucosa. The postganglionic sympathetic nerves in the mammalian nose typically contain catecholamines, but may also contain adenosine 5′-triphosphate (ATP) and neuropeptides, most notably neuropeptide Y (NPY).

The major consequence of sympathetic nerve stimulation is vasoconstriction and increases in nasal airway patency. There is also some evidence that sympathetic activity can induce airway glandular secretion through stimulation of serous cells. The effects of sympathetic nerves on nasal blood flow are mediated mainly through adrenoceptors. Stimulation of α1- and α2-adrenoceptors on the smooth muscle of resistance vessels (which control blood flow) and of venous sinusoids (which are responsible for blood pooling, leading to mucosal engorgement and modulation of nasal airway resistance) leads to vasoconstriction and, consequently, to reduced blood flow and reduced blood pooling. Some individuals that take α-adrenoceptor antagonists for treatment of their hypertension experience nasal congestion as a side effect. Conversely, α-adrenoceptor agonists are commonly used in nasal sprays to treat 'stuffy' noses. Activation of β1- and β2-adrenoceptors, also present on the vasculature of the nasal mucosa, can lead to vasodilation of resistance vessels and increased blood flow. The effects of β-receptor stimulation, however, are less pronounced than are those induced by α-receptor stimulation. Both NPY and ATP, transmitters often co-localized with norepinephrine in the nasal mucosa, also cause vasoconstriction and decongestion. These transmitters likely account for any nonadrenergic component of sympathetic nasal vasoconstriction.

Autonomic Reflexes in the Nose

Activation of afferent nerves in the respiratory mucosa can lead to sensations of temperature changes, touch, pain, sneezing, changes in breathing pattern, and changes in autonomic outflow to the nose. Experimental stimulation of nociceptive C-fibers with capsaicin application to one nostril invariably

leads to secretions from both the ipsilateral and the contralateral nostrils. The contralateral secretion can be prevented by treatment with muscarinic receptor antagonists, indicating a C-fiber-initiated cholinergic reflex. Activation of nasal afferent nerves can also alter autonomic outflow to the tear ducts and to the lower respiratory tract.

In a majority of human beings, there is a cycling of resistance to nasal airflow between the two nostrils. This so-called nasal cycle is thought to be due to unilateral cyclical changes in parasympathetic and sympathetic drive to the nasal vasculature. The congestion occurs concurrently with a decrease in sympathetic drive, but some evidence also implicates a role for a cyclical increase in parasympathetic drive.

Axon reflexes Generally speaking, stimulation of nasal sensory nerves regulates end-organ activity by increasing or decreasing autonomic reflex tone. However, there is substantial evidence for the existence of local reflex regulation of end-organ activity that occurs independently of centrally mediated autonomic response via sensory axon reflex mechanisms. Subsets of C-fibers innervating the nasal mucosa contain neuropeptides such as substance P, neurokinin A, and calcitonin gene-related peptide in their peripheral (and central) terminals, which can release neuropeptides upon activation. The axon-reflex concept envisages an action potential discharge initiated at a terminal site of the axon reaching a branch in the nerve, where it then antidromically (away from the central nervous system) travels back down a collateral branch of the same nerve, resulting in the local release of bioactive peptides. Release of neuropeptides from sensory nerve endings can lead not only to vasodilatation, but also to increased leukocyte infiltration and increased vascular permeability. These events are collectively referred to as 'neurogenic inflammation.' Stimulation of neuropeptide-containing C-fibers with capsaicin has been reported to evoke neurogenic inflammatory responses in the human nose, and these effects may be amplified when studied in the context of ongoing nasal inflammation (rhinitis).

Lower Respiratory Tract

A schematic of the innervation of the bronchial wall is illustrated in Figure 2. The trachea and bronchi are richly innervated by autonomic and sensory nerve fibers. In guinea pigs and rodents a majority of airway nerves are found in the subepthelial nerve plexus, and in a more superficial plexus near the smooth muscle layer. This gets more complex in larger mammals. In the dog trachea, for example, five separate nerve plexi have been identified. Beginning from the serosal surface there is paratracheal plexus, an intramuscular plexus, an outer submucosal plexus, an inner submucosal plexus, and, finally, a mucosal plexus.

Sensory Nerves in the Lower Respiratory Tract

The majority of afferent fibers reach the lower respiratory tract via the vagus nerves. Among the 8000 or so vagal afferent fibers innervating cat lungs, morphological studies reveal that about 5000 of these are C-fibers, with the remaining 3000 being myelinated A-fibers.

The vagal C-fibers innervating the lower respiratory tract have been subdivided into those innervating large airways (bronchial C-fibers), and those innervating more peripheral structures (pulmonary C-fibers). Based on studies in guinea pigs and rodents it appears that these C-fiber subtypes may originate in different vagal ganglia, with the C-fibers innervating the large airways being derived mainly from the jugular or supranodose ganglia, and those found deeper in the tissue arising from nodose ganglia. The nodose and jugular ganglia have different embryological origins, and this may explain why the nodose and jugular C-fibers in the respiratory tract have distinctly different phenotypes.

Most A-fibers in the lungs are relatively low-threshold mechanosensors that can be activated by lung distension. These 'stretch receptors' have been subdivided into rapidly adapting receptors (RARs) and slowly adapting receptor (SARs) based on the action potential accommodation to prolonged supra-threshold lung distension. Another vagal fiber phenotype that conducts action potentials at a velocity faster than C-fibers do, but slower than RAR or SAR fibers do, has been described in the extrapulmonary airways of the guinea pig. These fibers are not stretch receptors, and evoke the cough reflex when stimulated. Myelinated afferent fibers that do not fit into either the RAR or SAR class are also found, along with a subset of C-fibers, innervating neuroepithelial bodies (NEBs) in the airway epithelium. The function of NEB-associated afferent nerves has not yet been worked out.

Although the vast majority of afferent fibers innervating the lower respiratory tract are vagal, sensory neurons situated in the thoracic dorsal root ganglia also project fibers to the lungs. These are thought to be primarily C-fibers, but to date they have not been studied in detail.

Autonomic Innervation of Lower Respiratory Tract

Parasympathetic The parasympathetic branch of the autonomic nervous system is the dominant neural regulator of intra- and extrapulmonary airways. The preganglionic parasympathetic neurons are situated

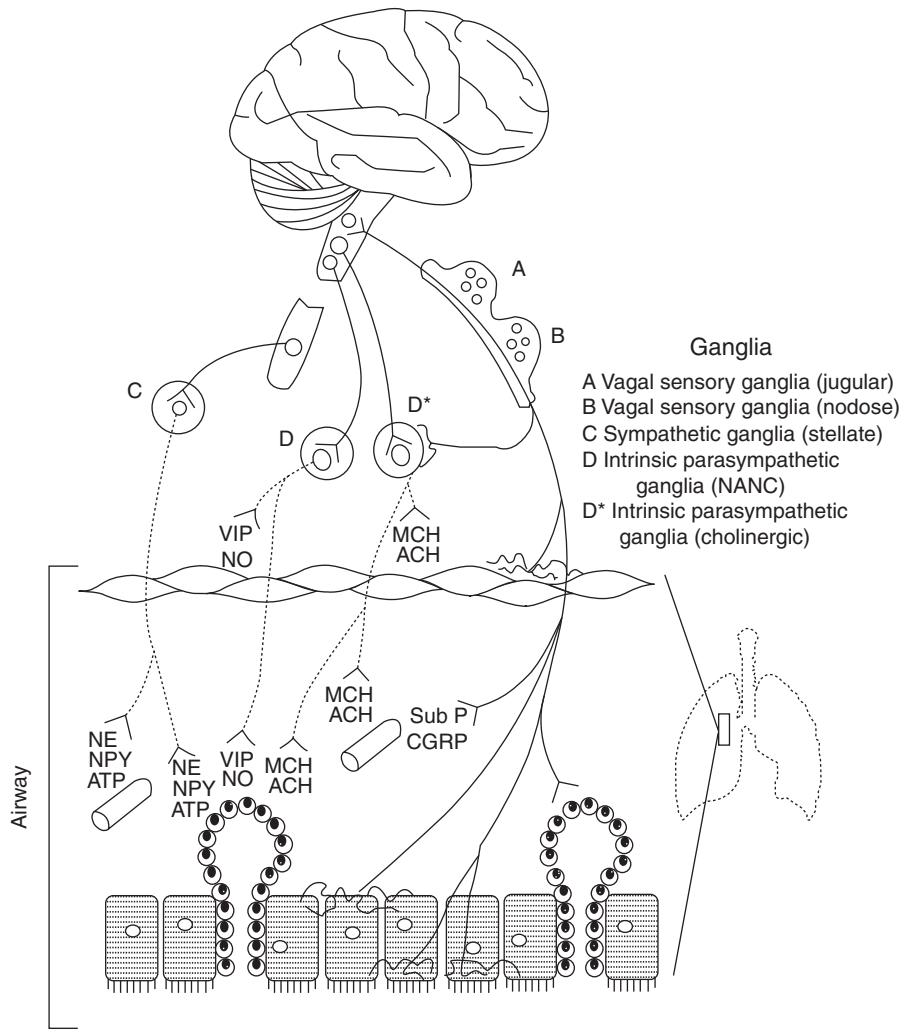

Figure 2 Schematic diagram of the innervation of the bronchi. The transmitters noted are acetylcholine (ACH), adenosine 5′-triphosphate (ATP), vasoactive intestinal peptide (VIP), nitric oxide (NO), substance P (Sub P), calcitonin gene-related peptide (CGRP), norepinephrine (NE), neuropeptide Y (NPY), and methacholine (MCH); the intrinsic parasympathetic ganglia are nonadrenergic, noncholinergic (NANC). The transmitters and neuropeptides depicted are the most commonly studied ones, but should not be considered an exclusive list. Not depicted is the afferent innervation derived from spinal (dorsal root) sensory ganglia. The spinal ganglia are found mainly in the lungs, but where and if the terminate in the airway wall have not been clearly established.

in the brain stem in and around the compact formation of the nucleus ambiguous and the dorsal motor nucleus of the vagus, and in the reticular formation located between these nuclei.

The preganglionic parasympathetic fibers arise at the airway via the vagus nerves, where they synapse with principal ganglia neurons (postganglionic neurons) that are located in small ganglia within or near the airway wall. Much of the preganglionic input is filtered at the parasympathetic ganglia because the excitatory postsynaptic potentials fail to reach action potential threshold. Parasympathetic ganglia are located mainly in the larger airways, but the postganglionic fibers innervate the conducting airways to the bronchioles. As a consequence, the parasympathetic ganglion neurons function to filter, integrate, and distribute parasympathetic activity throughout the branches of the airway tree. The number of neurons within an intrinsic airway parasympathetic ganglion ranges from a few to nearly hundred. Investigators have estimated the total number of intrinsic parasympathetic neurons in the airways of several species, including mouse (about 200), guinea pig (about 250), rat (about 750), ferret (about 4000), and dog (about 15 000).

Neurons in the bronchial parasympathetic ganglia are cholinergic and NANC or both. As in the upper airways, the major parasympathetic NANC neurotransmitters identified thus far in human airways are nitric oxide and vasoactive intestinal peptide, but in

some species tachykinins such as substance P have also been localized in intrinsic airway parasympathetic neurons. Several studies have provided evidence that the cholinergic innervation and NANC parasympathetic innervation reflect discrete parasympathetic pathways. In ferrets, the NANC neurons are situated in ganglia that are more superficial than are the ganglia containing mainly cholinergic neurons. In the guinea pig trachea and main bronchi, the parasympathetic NANC nerve fibers are actually derived from neurons situated in ganglia associated with the myenteric plexus of the esophagus (this may explain why the trachea has relatively few intrinsic neurons compared to the rat; see earlier). The preganglionic fibers that innervate the NANC airway-targeted neurons in the guinea pig esophagus are also distinct from those that innervate the cholinergic neurons within the airway. Therefore, in this species at least, separate reflex pathways may exist for parasympathetic cholinergic responses and parasympathetic noncholinergic responses. Circumstantial evidence also supports the hypothesis of separate cholinergic and NANC parasympathetic pathways in cat airways.

The parasympathetic nervous system is the dominant regulator of airway smooth muscle tone in all mammals, including humans. Stimulation of vagal parasympathetic nerves leads to bronchial smooth muscle contraction in all mammals. Invariably, the parasympathetic contractions are the result of cholinergic muscarinic receptor activation. Although bronchial smooth muscle expresses both muscarinic M2 and M3 receptors, the latter are responsible for the smooth muscle contractions. This explains the utility of inhaled antimuscarinic drugs such as ipratroprium, and the pseudo-irreversible tiotropium, in the treatment of obstructive lung diseases. There are muscarinic M2 receptors on the cholinergic terminals that serve to inhibit acetylcholine release. Thus, selectively blocking muscarinic M2 receptors in human bronchi can actually enhance nerve-evoked cholinergic contractions.

In many species, including humans, the parasympathetic nervous system also provides relaxant innervation to bronchial smooth muscle. Drugs that block the production or action of nitric oxide largely inhibit the parasympathetically mediated relaxations. As mentioned earlier, at least in guinea pigs, the parasympathetic relaxant innervation is distinct from the cholinergic innervation. In theory, acetylcholine and the noncholinergic transmitters may be co-released from the same nerve terminals. Such co-release, however, has not yet been observed at the level of airway smooth muscle.

Tracheobronchial submucosal glands and goblet cells receive both parasympathetic and sympathetic efferent innervations, though they are dominated by parasympathetic tone. The neurogenic increase in mucus secretion in human airways is largely cholinergic via muscarinic M3 receptor stimulation; however, noncholinergic transmitters, such as vasoactive intestinal peptide, associated with parasympathetic nerves also stimulate mucus secretion. The parasympathetic system may play a role in mucociliary responses. Administration of atropine, a muscarinic receptor antagonist, decreases ciliary beat frequency, suggesting that the role is stimulatory.

Increases in parasympathetic outflow to the airways results in vasodilatation and increases in blood flow. This occurs in both the pulmonary and bronchial (systemic) circulation. In experimental animals, parasympathetic nerve-mediated relaxation of the bronchial vasculature is mediated through both cholinergic and noncholinergic (likely nitric oxide and vasoactive intestinal peptide) mechanisms.

Sympathetic The mammalian respiratory tract receives sympathetic innervation mainly from cell bodies situated in the stellate ganglia. Postganglionic sympathetic nerves typically use norepinephrine as the major neurotransmitter. In some cases NPY and ATP are released along with norepinephrine as co-transmitters. Much of the early work on airway innervation was carried out on dogs. This work led to the conclusion, still reprinted in many textbooks, that the parasympathetic system provides cholinergic contractile innervation to the bronchial smooth muscle, and this is apposed by the sympathetic nervous system, which provides adrenergic relaxant innervation. Indeed, stimulation of sympathetic nerves in the dog leads to β-adrenoceptor-mediated bronchodilation. This is not the case, however, in most species. In guinea pigs, cats, monkeys, and humans, the major, if not only, relaxant innervation to the airway smooth muscle is provided by the parasympathetic nervous system, as already described. Sympathetic, β-adrenoceptor-mediated, bronchial smooth muscle relaxation can occur, however, in all species thus far studied via circulating epinephrine derived from the adrenal medulla.

As in other organs, the sympathetic nervous system innervates the blood vessels in the airways and lungs. The pulmonary and bronchial circulations are under adrenergic control, and perhaps also under sympathetic nonadrenergic (e.g., NPY) control. Studies on intrapulmonary pulmonary arteries have indicated that ATP may also contribute to sympathetic nonadrenergic vasoconstriction. Sympathetic transmitters can constrict and relax vascular smooth muscle, but in the vast majority of studies in experimental animals, the net effect of sympathetic nerve stimulation is vasoconstriction and decreased blood flow.

Autonomic Reflexes in the Lower Airways

The parasympathetic preganglionic neurons are rhythmically active during eupnic breathing and send volleys of action potentials down the preganglionic axons. Intracellular recordings of tracheal ganglion neurons in the cat reveal rhythmic preganglionic volleys in excess of 25 Hz. This baseline activity appears to be due to the effect of pattern generators within the central nervous system, and to reflex activity evoked through the stimulation of certain vagal mechanosensors as a consequence of lung expansion following inspiration. Artificially increasing the respiratory rate increases the baseline parasympathetic drive. This is thought to be due to activation of vagal RAR stretch receptors. The baseline parasympathetic drive to the airways can be essentially terminated by a prolonged deep inspiration, presumably due to activation of vagal SAR stretch receptors. In other words, RARs and SARs have apposing effects on reflex parasympathetic tone.

Both cholinergic and NANC parasympathetic fibers are active during normal respiration in airways, but the net effect is one of cholinergic constriction. Studies using computed tomography indicate that the luminal diameters of small and midsized human bronchi are reduced about 15% by the tonic descending parasympathetic tone. When the vagi are cut (in experimental animals), or if large doses of antimuscarinic drugs are administered, the baseline smooth muscle tone is essentially abolished and the airways maximally dilate.

Inhalation of an irritant or potentially noxious stimulus that is capable of activating nociceptive C-fibers in the airway mucosa will lead to increases in parasympathetic drive and the consequential airway narrowing, mucus secretion, and vasodilation. Several inflammatory mediators are adept at evoking action potential discharge in airway vagal nociceptors, and it is thought that the resulting increase in parasympathetic reflex nerve activity may contribute to the pathophysiology of inflammatory airway diseases (e.g., chronic obstructive pulmonary disease, asthma).

Hypoxic pulmonary vasoconstriction is a major mechanism in the prevention of arterial hypoxemia. There is some evidence that reflex sympathetic outflow to the pulmonary circulation may contribute to this important phenomenon.

Peripheral reflexes A peripheral autonomic reflex in which sensory nerves directly communicate with intrinsic autonomic nerves, independent of the central nervous system, may occur in the lower airways, although this is certainly less prevalent than in the enteric nervous system. Neurokinin-containing C-fiber terminals are found near many bronchial parasympathetic ganglion neurons in several species, including humans. Direct electrophysiological recordings have demonstrated that stimulation of these C-fibers can evoke membrane depolarization and occasionally action potential discharge in the principal ganglion neurons. In guinea pig and human airways, the C-fiber-mediated excitatory postsynaptic potentials are due to stimulation of neurokinin (NK-3) receptors. Consistent with the electrophysiological studies, blockers of NK-3 receptors can inhibit cholinergic bronchoconstriction *in vivo* caused by vagus nerve stimulation (presumably by decreasing the synaptic efficacy within the intrinsic ganglia). There is anatomical evidence of intraganglionic communication within the airways. Specifically, postganglionic parasympathetic NANC fibers have been observed within cholinergic parasympathetic ganglia. The functional significance of this type of intraganglionic integration is not known.

Axon reflexes C-fiber axon reflex-induced neurogenic inflammation (as discussed previously) has been thoroughly investigated in the airways of guinea pigs and rodents. In these species, antidromic stimulation of vagal C-fibers leads to pronounced plasma extravasation and infiltration of leukocytes. The plasma leakage occurs at the postcapillary venules of the bronchial circulation. In guinea pigs, axon reflex responses may also include bronchial smooth muscle contraction via activation of neurokinin NK-1 and NK-2 receptors on the airway smooth muscle.

The effect of axon reflexes in human airways is less obvious. The density of peptidergic C-fiber innervation in the airways of guinea pigs and rats is much greater than that seen in humans. It is therefore likely that the neurogenic inflammatory response that occurs in the lower airways of these species overstates the case seen in humans.

See also: Autonomic Nervous System: Central Respiratory Control; Autonomic Nervous System; Brainstem Respiratory Circuits; Evolution of Vertebrate Respiratory Control; Parasympathetic Nervous System; Respiration; Sympathetic Nervous System.

Further Reading

Brouns I, De Proost I, Pintelon I, et al. (2006) Sensory receptors in the airways: Neurochemical coding of smooth muscle-associated airway receptors and pulmonary neuroepithelial body innervation. *Autonomic Neuroscience* 126–127: 307–319.

Canning BJ (2006) Reflex regulation of airway smooth muscle tone. *Journal of Applied Physiology* 101: 971–985.

Canning BJ and Mazzone SB (2005) Reflexes initiated by activation of the vagal afferent nerves innervating the airways and lungs. In: Undem BJ and Weinreich D (eds.) *Advances in Vagal Afferent Neurobiology*, pp. 403–430. Boca Raton, FL: CRC Press and Taylor & Francis Group.

Carr MJ and Undem BJ (2003) Bronchopulmonary afferent nerves. *Respirology* 8: 291–301.

Coleridge HM and Coleridge JC (1994) Neural regulation of bronchial blood flow. *Respiratory Physiology* 98: 1–13.

Coleridge HM and Coleridge JC (1994) Pulmonary reflexes: Neural mechanisms of pulmonary defense. *Annual Review of Physiology* 56: 69–91.

Ellis JL and Undem BJ (1994) Pharmacology of non-adrenergic, non-cholinergic nerves in airway smooth muscle. *Pulmonary Pharmacology* 7: 205–223.

Kubin L, Alheid GF, Zuperku EJ, et al. (2006) Central pathways of pulmonary and lower airway vagal afferents. *Journal of Applied Physiology* 101: 618–627.

Laitinen LA and Laitinen A (1997) Innervation of the airways. In: Barnes PJ (ed.) *Autonomic Control of the Respiratory System*, pp. 1–21. Amsterdam: Harwood Academic Publishers.

Rogers DF (2002) Pharmacological regulation of the neuronal control of airway mucus secretion. *Current Opinions in Pharmacology* 2: 249–255.

Sarin S, Undem B, Sanico A, et al. (2006) The role of the nervous system in rhinitis. *Journal of Allergy and Clinical Immunology* 118: 999–1016.

Undem BJ and Myers AC (1997) Autonomic ganglia. In: Barnes PJ (ed.) *Autonomic Control of the Respiratory System*, pp. 1–21. Amsterdam: Harwood Academic Publishers.

Widdicombe JG (1991) Neural control of airway vasculature and edema. *American Review of Respiratory Disease* 143: S18–S21.

Wine JJ (2007) Parasympathetic control of airway submucosal glands: Central reflexes and the airway intrinsic nervous system. *Autonomic Neuroscience* 133: 35–54.

Autonomic Nervous System: Urogenital Control

W C de Groat, University of Pittsburgh Medical School, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The urogenital system has two major subdivisions, the urinary tract and the genital tract. The urinary tract, consisting of the kidneys, ureters, urinary bladder, and urethra, is responsible for the production, storage, and elimination of urine. The genital tract, consisting of the penis, clitoris, vagina, testes, ovaries, and uterus, is involved in reproductive activity. Many functions of the urogenital system are controlled by autonomic (parasympathetic and sympathetic) and somatic efferent pathways originating in the lumbosacral spinal cord (Figure 1). These pathways regulate the activity of smooth muscle, striated muscle, epithelial cells, and exocrine glands in the urogenital organs via the release of multiple neurotransmitters, including acetylcholine (ACh), norepinephrine, adenosine triphosphate (ATP), nitric oxide (NO), and neuropeptides. This article reviews the innervation of the urogenital organs, the functions controlled by the different neural pathways, and the physiology and pharmacology of neurotransmission.

Lower Urinary Tract

Anatomy

The storage and periodic elimination of urine is controlled by the activity of two functional units in the lower urinary tract, (1) a reservoir (the urinary bladder) and (2) an outlet (consisting of bladder neck, urethra, and striated sphincter muscles). Under normal conditions, the urinary bladder and urethral outlet exhibit a reciprocal activity. During storage, the bladder neck and proximal urethra are closed and the detrusor muscle is quiescent, allowing intravesical pressure to remain low (5–15 cmH_2O) over a wide range of bladder volumes. During voluntary micturition the initial event is a reduction of intraurethral pressure, which reflects a relaxation of the pelvic floor and the paraurethral striated muscles. These changes are followed in a few seconds by a detrusor contraction and a rise in intravesical pressure that is maintained until the bladder empties.

Innervation

The innervation of the lower urinary tract is derived from three sets of peripheral nerves: sacral parasympathetic (pelvic nerves), thoracolumbar sympathetic (hypogastric nerves and sympathetic chain), and sacral somatic nerves (primarily the pudendal nerves) (Figure 1).

Parasympathetic pathways The sacral parasympathetic outflow, which in humans originates from the S2 to S4 segments of the spinal cord, provides the major excitatory input to the bladder. Cholinergic preganglionic neurons located in the intermediolateral region of the sacral spinal cord send axons to cholinergic ganglion cells in the pelvic plexus and in the bladder wall (Figure 2). Transmission in bladder ganglia is mediated by a nicotinic cholinergic mechanism, which is sensitive to modulation by various transmitter systems, including muscarinic, adrenergic, purinergic, and peptidergic (Table 1; Figure 2). The ganglion cells in turn excite the bladder smooth muscle. A large proportion of the ganglia and nerves supplying the human lower urinary tract contain acetylcholinesterase (AChE) as well as the vesicular ACh transporter (VAChT), and therefore must be cholinergic. AChE- and VAChT-positive nerves are abundant in all parts of the bladder but are less extensive in the urethra. Neuropeptide Y (NPY) and nitric oxide synthase (NOS) have also been identified in a large percentage (40–95%) of intramural ganglia of the human bladder. Several populations of axonal varicosities have been detected in close proximity to intramural ganglion cells, including (1) substance P-, vasoactive intestinal polypeptide (VIP)-, and calcitonin-gene related peptide-positive axons which are presumably collaterals of extrinsic sensory nerves; (2) tyrosine hydroxylase and NPY axons which are likely to be sympathetic axons; and (3) galanin- and NPY-containing axons; which are thought to be preganglionic nerve terminals.

Parasympathetic neuroeffector transmission in the bladder is mediated by ACh acting on postjunctional muscarinic (M) receptors. Both M_2 and M_3 muscarinic receptor subtypes are expressed in bladder smooth muscle (Figure 3); however, studies with subtype-selective muscarinic receptor antagonists and muscarinic receptor knockout mice have revealed that the M_3 subtype is the principal receptor involved in excitatory transmission.

In bladders of various animals, stimulation of parasympathetic nerves also produces a noncholinergic contraction that is resistant to atropine and other muscarinic receptor-blocking agents. ATP (Table 1) has been identified as the excitatory transmitter mediating the noncholinergic contraction. ATP excites the bladder smooth muscle by acting on P2X receptors

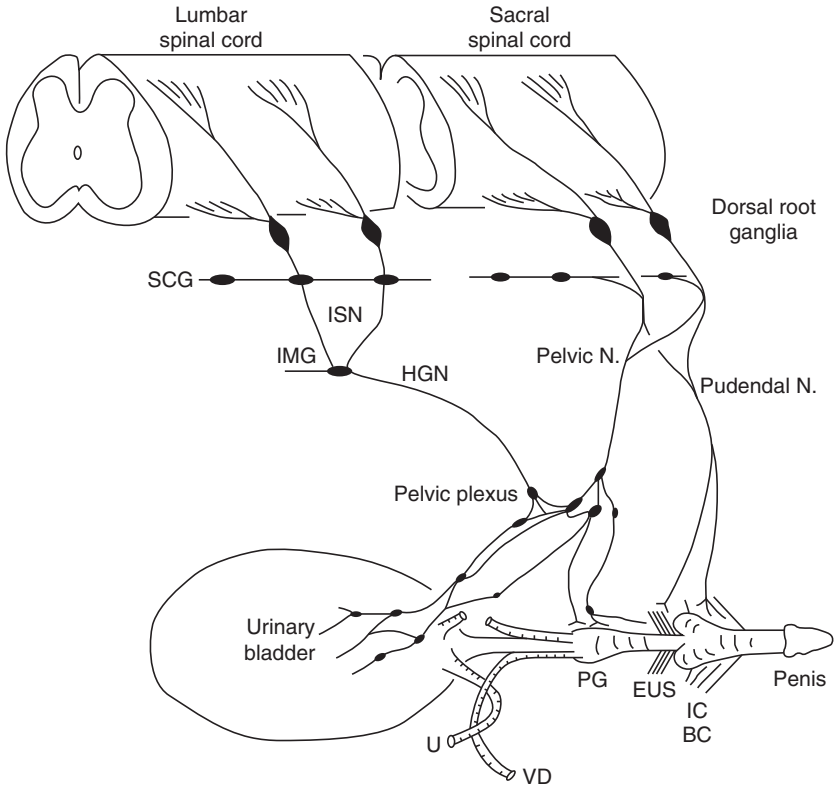

Figure 1 Diagram showing the sympathetic, parasympathetic, and somatic innervation of the urogenital tract of the male cat. Sympathetic preganglionic pathways emerge from the lumbar spinal cord and pass to the sympathetic chain ganglia (SCG) and then via the inferior splanchnic nerves (ISN) to the inferior mesenteric ganglia (IMG). Preganglionic and postganglionic sympathetic axons then travel in the hypogastric nerve (HGN) to the pelvic plexus and the urogenital organs. Parasympathetic preganglionic axons, which originate in the sacral spinal cord, pass in the pelvic nerve to ganglion cells in the pelvic plexus and to distal ganglia in the organs. Sacral somatic pathways are contained in the pudendal nerve, which provides an innervation to the penis, the ischiocavernosus (IC), bulbocavernosus (BC), and external urethral sphincter (EUS) muscles. The pudendal and pelvic nerves also receive postganglionic axons from the caudal sympathetic chain ganglia. These three sets of nerves contain afferent axons from the lumbosacral dorsal root ganglia. Abbreviations: U, ureter; PG, prostate gland; VD, vas deferens.

that are ligand-gated ion channels. Among the seven types of P2X receptors that have been identified in the bladder, the $P2X_1$ subtype is the major subtype expressed in the rat and human bladder smooth muscle. Although purinergic excitatory transmission is not important in the normal human bladder, it has been identified in bladders from patients with pathological conditions such as chronic urethral outlet obstruction or interstitial cystitis.

Smooth muscle contractions are initiated by an increase in intracellular Ca^{2+} concentration that can occur by intracellular release of Ca^{2+} from the sarcoplasmic reticulum or by influx of Ca^{2+} from the extracellular fluid. The former mechanism is an essential step in the cholinergic activation of the detrusor muscle. It has been shown that stimulation of M_3 receptors triggers the formation of inositol trisphosphate (IP_3) and this in turn activates IP_3 receptors on the sarcoplasmic reticulum, which then causes the release of Ca^{2+} (Figure 3). On the other hand, activation of P2X purinergic receptors causes the influx of extracellular Ca^{2+} as well as depolarization of the cells, leading to an opening of voltage-gated Ca^{2+} channels. This triggers intracellular Ca^{2+}-induced Ca^{2+} release by activation of ryanodine-sensitive receptors in the sarcoplasmic reticulum. Intracellular Ca^{2+} combines with calmodulin to activate the contractile proteins. Activation of M_2 muscarinic receptors also appears to enhance contractions by suppressing β-adrenergic inhibitory mechanisms by blocking adenylyl cyclase (Figure 3) or K^+ channels.

Parasympathetic pathways to the urethra induce relaxation during voiding. In various species, the relaxation is not affected by muscarinic antagonists and therefore is not mediated by ACh. However, inhibitors of NOS block the relaxation *in vivo* during reflex voiding or block the relaxation of urethral smooth muscle strips induced *in vitro* by electrical stimulation of intramural nerves, indicating that NO is the inhibitory transmitter involved in relaxation. NO increases the levels of cyclic guanosine monophosphate (GMP) by stimulating guanylyl cyclase. Cyclic

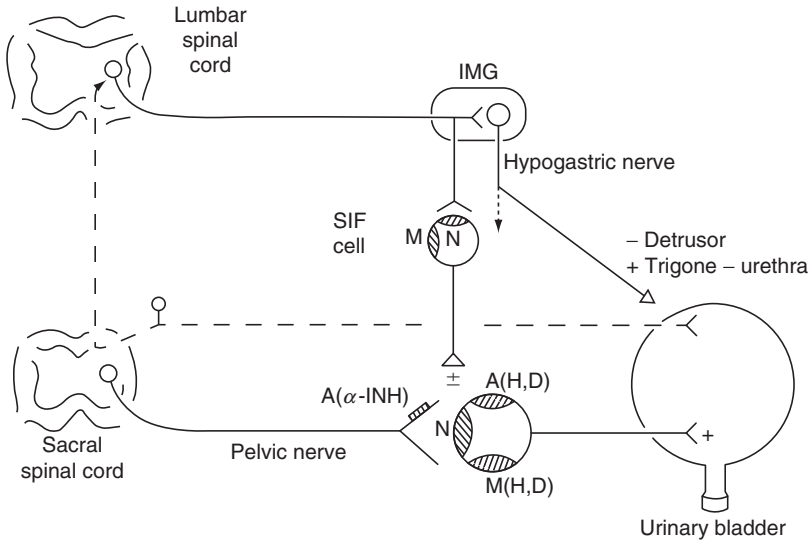

Figure 2 Diagram showing the autonomic innervation of the urinary bladder of the cat and the synaptic mechanisms within bladder ganglia. Nicotinic (N), muscarinic (M), and adrenergic (A) receptors are depicted on a principal ganglion cell and a small intensely fluorescent (SIF) cell. Receptors mediating hyperpolarization (H) and depolarization (D) are also indicated. An α-adrenergic receptor (A, α-INH) mediating presynaptic inhibition is indicated on the preganglionic nerve terminal. Inhibitory and excitatory synaptic mechanisms are designated by $-$ and $+$, respectively. Postsynaptic adrenergic and muscarinic receptors mediate both hyperpolarizing and depolarizing responses. Sympathetic preganglionic axons make synaptic contact with cells in the inferior mesenteric ganglion (IMG) and send axons through the IMG to make synaptic contact with SIF cells in bladder ganglia. SIF cells have both nicotinic and muscarinic excitatory receptors. Sympathetic efferent pathways can be activated by afferent projections from the urinary bladder.

Table 1 Receptors for putative transmitters in the lower urinary tract

Tissue	*Cholinergic*	*Adrenergic*	*Other*
Bladder body	$+(M_2)$	$-(\beta_2)$	$+$Purinergic ($P2X_1$)
	$+(M_3)$	$-(\beta_3)$	$-$VIP
			$+$Substance P (NK-2)
Bladder base	$+(M_2)$	$+(\alpha_1)$	$-$VIP
	$+(M_3)$		$+$Substance P (NK-2)
			$+$Purinergic (P2X)
Urothelium	$+(M_2)$	$+\alpha$	$+$TRPV1
	$+(M_3)$	$+\beta$	$+$TRPM8
			$+$P2X
			$+$P2Y
			$+$Substance P
			$+$Bradykinin (B2)
Urethra	$+$(M)	$+(\alpha_1)$	$+$Purinergic (P2X)
		$+(\alpha_2)$	$-$VIP
		$-(\beta)$	$-$Nitric oxide
Sphincter striated muscle	$+$(N)		
Adrenergic nerve terminals	$-(M_{2/4})$	$-(\alpha_2)$	$-$NPY
	$+(M_1)$		
Cholinergic nerve terminals	$-(M_{2/4})$	$+(\alpha_1)$	$-$NPY
	$+(M_1)$		
Afferent nerve terminals			$+$Purinergic ($P2X_{2/3}$)
			$+$TRPV1
Ganglia	$+$(N)	$+(\alpha_1)$	$-$Enkephalinergic (δ)
	$+(M_1)$	$-(\alpha_2)$	$-$Purinergic (P_1)
		$+(\beta)$	$+$Purinergic (P2X)
			$+$Substance P

VIP, vasoactive intestinal polypeptide; NPY, neuropeptide Y; TRP, transient receptor potential. Letters in parentheses indicate receptor type (M, muscarinic; N, nicotinic; NK-2, neurokinin-A). Plus and minus signs indicate excitatory and inhibitory effects.

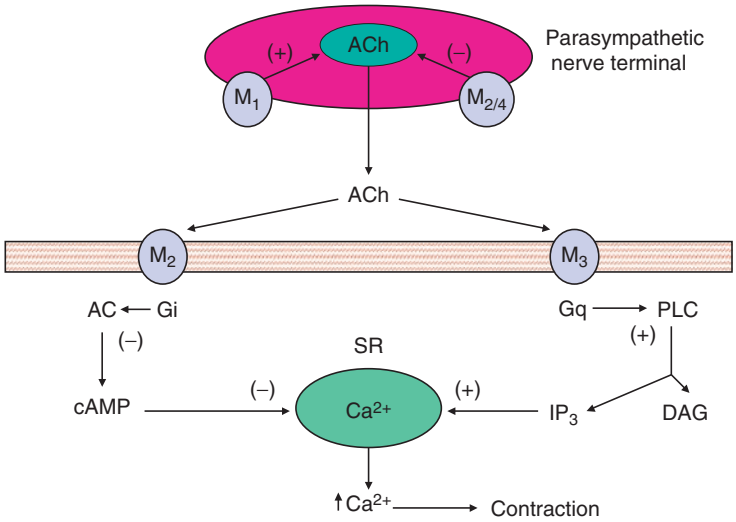

Figure 3 Diagram showing prejunctional modulatory muscarinic receptors on a parasympathetic cholinergic nerve terminal in the urinary bladder. Acetylcholine (ACh) released from the nerve terminals acts on postjunctional M_2 and M_3 muscarinic receptors on the bladder smooth muscle. Activation of M_3 receptors acts via a G-protein (Gq) to stimulate phospholipase C (PLC) and induce formation of diacylglycerol (DAG) and inositol trisphosphate (IP_3); IP_3 in turn triggers the release of Ca^{2+} from the sarcoplasmic reticulum (SR). Increase in Ca^{2+} initiates a muscle contraction. Stimulation of M_2 receptors acts through another G-protein (Gi) to inhibit adenylyl cyclase (AC) and reduce cyclic adenosine monophosphate (cAMP) inhibition of contractile mechanisms.

GMP in turn activates protein kinase G, which produces smooth muscle relaxation by several mechanisms, including activation of K^+ channels and desensitization of the contractile machinery to Ca^{2+}.

Sympathetic pathways Sympathetic preganglionic pathways that arise from the T11 to L2 spinal segments pass to the sympathetic chain ganglia and then to prevertebral ganglia in the superior hypogastric and pelvic plexuses, and also to short adrenergic neurons in the bladder and urethra. Sympathetic postganglionic nerves that release norepinephrine provide an excitatory input to smooth muscle of the urethra and bladder base, an inhibitory input to smooth muscle in the body of the bladder, as well as inhibitory and facilitatory inputs to vesical parasympathetic ganglia (Figure 2). Histofluorescence microscopy in animals and humans has shown that the smooth muscle of the bladder base is richly innervated by adrenergic terminals, but the bladder body has a considerably weaker adrenergic innervation. Vesical parasympathetic ganglion cells also receive an extensive adrenergic innervation. Radioligand receptor binding studies show that α-adrenergic receptors are concentrated in the bladder base and proximal urethra, whereas β-adrenergic receptors are most prominent in the bladder body (Table 1). These observations are consistent with pharmacological studies showing that sympathetic nerve stimulation or exogenous catecholamines produce β-adrenergic receptor-mediated inhibition of the body and strong α-adrenergic receptor-mediated contractions of the base and urethra and weak contractions of the bladder body. Molecular and contractility studies have shown that β_3-adrenergic receptors elicit inhibition and α_1-adrenergic receptors elicit contractions. The α_{1A}-adrenergic receptor subtype is most prominent in normal bladders, but the α_{1D} subtype is upregulated in bladders from patients with outlet obstruction due to benign prostatic hyperplasia. This finding raises the possibility that enhanced α_1-adrenergic receptor excitatory mechanisms in the bladder body might contribute to irritative lower urinary tract symptoms in these patients with prostate disease.

Activation of β-adrenergic receptors in bladder smooth muscle stimulates adenylyl cyclase and increases cyclic adenosine monophosphate (cAMP), which in turn activates protein kinase A. Protein kinase A is thought to act in part by inducing a hyperpolarization of the cells, either by opening of K^+ channels or by stimulating an electrogenic ion pump. Excitatory responses in the urethra and bladder neck mediated by α_1-adrenergic receptors are attributed to an increased release of Ca^{2+} from intracellular stores.

Neural Modulatory Mechanisms

Presynaptic modulatory mechanisms and synaptic communication between parasympathetic and sympathetic pathways to the bladder have been demonstrated in bladder ganglia and at postganglionic nerve terminals. In the cat, stimulation of sympathetic nerves elicits an initial inhibitory and a delayed facilitatory modulation of cholinergic transmission in parasympathetic bladder ganglia mediated by α_2- and α_1-adrenergic receptors, respectively (Table 1;

Figure 2). Transmission in cat bladder ganglia is also modulated by enkephalins released as cotransmitters with ACh from preganglionic nerve terminals. The inhibitory effect of enkephalins can be blocked by the opioid receptor antagonist naloxone and occurs by a presynaptic inhibitory mechanism. Adenosine, presumably derived from ATP released in the ganglia, also exerts an inhibitory action on ganglionic transmission.

Adrenergic and enkephalinergic inhibitory mechanisms in cat bladder ganglia are frequency dependent, being prominent at low frequencies (0.25–0.5 Hz) and markedly reduced at higher frequencies (1–10 Hz) of stimulation. This phenomenon is presumably related to the marked temporal facilitation, which occurs in cat bladder ganglia at frequencies of preganglionic nerve stimulation above 0.5 Hz (Figure 4). It has been speculated that cat bladder ganglia function as high-pass filters eliminating parasympathetic input to the bladder when preganglionic activity is low during urine storage, but significantly amplifying the input to the bladder when firing is increased during voiding. Thus the ganglion acts like a gating circuit. Frequency-dependent adrenergic and enkephalinergic inhibitory mechanisms complement this gating function by inhibiting transmission during urine storage, but turning off during voiding to allow complete bladder emptying.

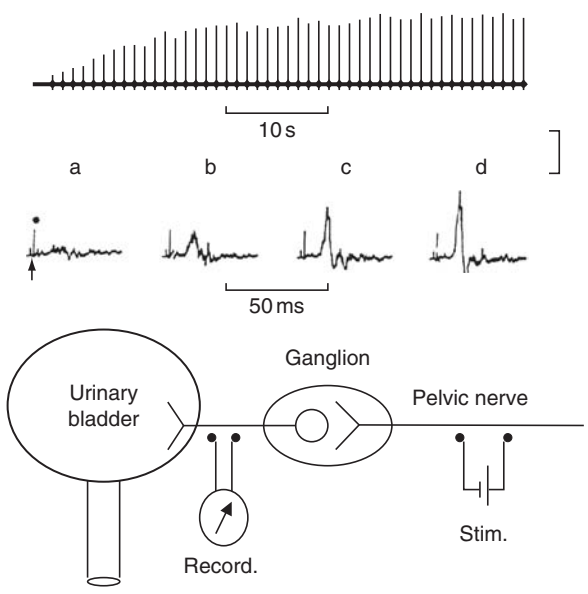

Figure 4 Extracellular recordings on a bladder postganglionic nerve, demonstrating facilitation of bladder ganglionic transmission during repetitive stimulation of preganglionic axons in the pelvic nerve. Top: Action potentials evoked with submaximal (5 V, 1 Hz) stimulation, recorded with a slow time base, showing gradual increase in the amplitude during repetitive preganglionic nerve stimulation. Middle: Sample responses from the preceding action potentials (a, 1st; b, 5th; c, 10th; and d, 20th responses in the series), recorded with a fast time base. Bottom: Diagram showing sites for nerve stimulation and recording.

In the rat, the autonomic pathways to the lower urinary tract are susceptible to modulation at postganglionic sites. For example, NPY, a cotransmitter in cholinergic and adrenergic nerves, acts on postganglionic nerve terminals to suppress the release of both ACh and norepinephrine in the bladder and urethra. The NPY inhibition is also frequency dependent, being most prominent at low frequencies of nerve stimulation. In addition, in the rat urinary bladder, ACh released from parasympathetic nerves can induce a heterosynaptic facilitation and inhibition of norepinephrine release from sympathetic nerves by acting, respectively, on prejunctional M_1 and $M_{2/4}$ muscarinic receptors on the adrenergic terminals. Prejunctional M_1 facilitatory and $M_{2/4}$ inhibitory muscarinic modulatory mechanisms have also been identified on parasympathetic nerve terminals in the rat bladder (Figure 3; Table 1). These receptors mediate positive and negative cholinergic feedback mechanisms, respectively, that regulate the release of ACh. Inhibitory $M_{2/4}$ mechanisms are dominant at low frequencies of nerve activity and therefore could contribute to urine storage; whereas M_1 facilitatory mechanisms are dominant at high frequencies of nerve stimulation and could contribute to an enhancement of neurally evoked bladder contractions during micturition.

Somatic pathways The external urethral sphincter (EUS), which is composed of striated muscle, receives a somatic innervation via the pudendal nerve from anterior horn cells in the third and fourth sacral segments. Branches of the pudendal nerve and other sacral somatic nerves also carry efferent impulses to muscles of the pelvic floor. Reflex activation of the EUS occurs during bladder filling or during sneezing and coughing, thereby contributing to the maintenance of urinary continence.

Urothelial–Afferent Interactions

The luminal surface of the bladder is covered by a multilayered urothelium, which functions as a highly efficient barrier to movement of water and ionized substances across the bladder wall. Recent studies have also revealed that the urothelium has specialized sensory and signaling properties that allow urothelial cells to respond to their chemical and physical environment and to engage in reciprocal chemical communication with neighboring nerves in the bladder wall or with smooth muscle cells (Figure 5). These properties include (1) expression of nicotinic, muscarinic, tachykinin, adrenergic, and capsaicin (TRPV1) receptors; (2) sensitivity to transmitters released from

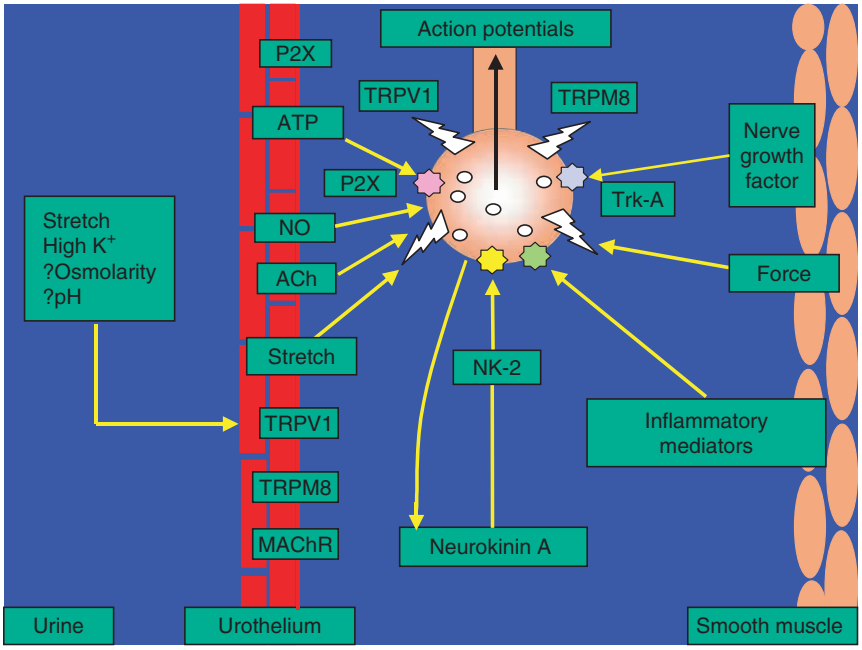

Figure 5 Diagram showing receptors present in the urothelium and in sensory nerve endings in the bladder mucosa, and putative chemical mediators that are released by the urothelium, nerves, or smooth muscle that can modulate the excitability of sensory nerves. Urothelial cells and sensory nerves express common receptors (purinergic, P2X; transient receptor potential vanilloid receptor 1 (sensitive to capsaicin), TRPV1; and transient receptor potential menthol/cold receptor, TRPM8). Distension of the bladder activates stretch receptors and triggers the release of urothelial transmitters, such as adenosine triphosphate (ATP), acetylcholine (ACh), and nitric oxide (NO), that may interact with adjacent nerves. Receptors in afferent nerves or the urothelium (e.g., TRPV1 or P2X) can respond to changes in pH, osmolality, high K^+ concentration, chemicals in the urine, or inflammatory mediators released in the bladder wall. Neuropeptides (neurokinin-A) released from sensory nerves in response to distension or chemical stimulation can act on neurokinin-A (NK-2) autoreceptors to sensitize the mechanosensitive nerve endings. The smooth muscle can generate force, which may influence some mucosal endings. Nerve growth factor released from muscle or urothelium can exert an acute and chronic influence on the excitability of sensory nerves via an action on tyrosine kinase A (Trk-A) receptors.

sensory nerves; (3) close physical association with afferent nerves; (4) ability to release chemical mediators such as ATP, ACh, and NO that can regulate the activity of adjacent nerves and thereby trigger local vascular changes and/or reflex bladder contractions; and (5) release of inhibitory factors that relax smooth muscle.

The role of ATP in urothelial–afferent communication has attracted considerable attention because bladder distension releases ATP from the urothelium and intravesical administration of ATP induces bladder hyperactivity, an effect blocked by administration of P2X purinergic receptor antagonists that suppress the excitatory action of ATP on bladder afferent neurons. Mice in which the $P2X_3$ receptor was knocked out exhibited hypoactive bladder activity and inefficient voiding, suggesting that activation of $P2X_3$ receptors on bladder afferent nerves by ATP released from the urothelium is essential for normal bladder function. It has also been reported that urothelial cells obtained from patients or cats with a chronic painful bladder condition (interstitial cystitis) released significantly larger amounts of ATP in response to mechanical stretching than did urothelial cells from normal patients. This raises the possibility that ATP-mediated signaling between the urothelium and afferent nerves (Figure 5) is involved in the triggering of painful bladder sensations.

Transmitter release from the urothelium appears to be mediated by exocytosis associated with vesicle trafficking, similar to the release mechanisms occurring in nerve terminals. Botulinum toxin A blocks release of transmitters from nerves by destroying soluble *N*-ethylmalemide-sensitive factor attachment protein receptors (SNARE proteins; involved in synaptic vesicle trafficking) and also blocks transmitter release from urothelial cells. Based on these observations it has been proposed that the beneficial effects of intravesical injections of botulinum toxin A in the treatment of overactive bladder symptoms may be related in part to the suppression of chemical communication between the urothelium and afferent nerves.

Sex Organs

The physiologic changes initiated by erotic stimuli can be divided into four distinct phases (excitement, plateau, orgasm, and resolution) that have been

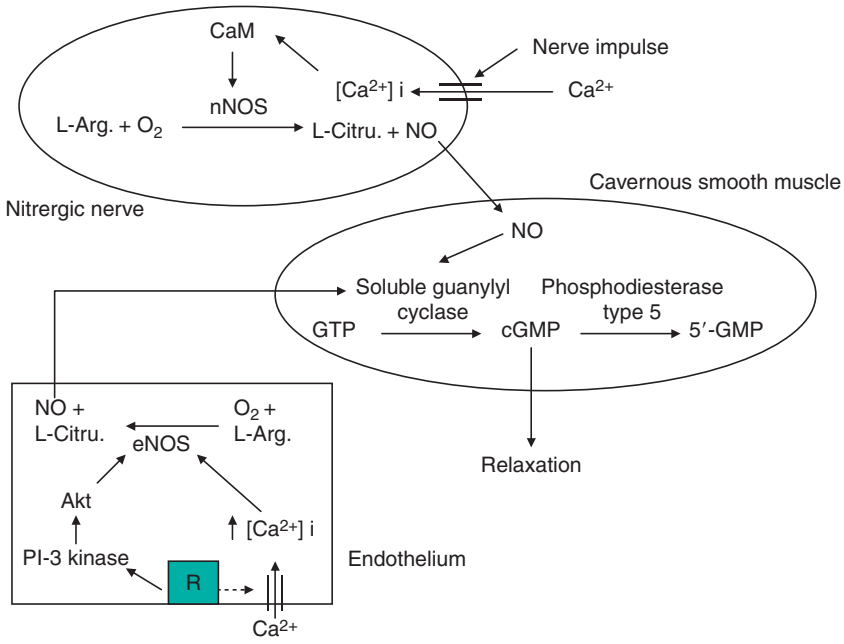

Figure 6 Diagram showing the regulation of cavernosal smooth muscle function by NO liberated from nitrergic nerves and the sinusoidal endothelium. A nerve impulse stimulates Ca^{2+} influx into the nerve terminal followed by synthesis of nitric oxide (NO) by neuronal NO synthase (nNOS). NO released from the nerve terminal activates cyclic guanosine monophosphate (cGMP) formation and relaxation of the cavernous smooth muscle. Changes in blood flow activate shear stress receptors in the endothelium, enhancing the synthesis of NO that elicits further relaxation of the smooth muscle. Nerve stimulation has also been reported to stimulate receptors (R), which activates a signaling mechanism involving phosphatidylinositol 3-kinase (PI-3 kinase) and a serine/threonine protein kinase (Akt), leading to the phosphorylation and activation of endothelial NOS (eNOS) and NO production. Abbreviations: CaM, calmodulin; Arg, arginine; Citru, citrulline.

designated collectively as 'the sexual response cycle.' Although anatomic differences obviously preclude identical responses in males and females during each phase of the cycle, it is clear that similar vascular responses (skin flush, penile and clitoral erection), secretory responses (stimulation of the prostate, bulbourethral gland, glands of Littre in the male; Bartholin's and paraurethral glands in the female), and responses of smooth and striated muscles occur in both sexes. The following sections review the neural control of the peripheral autonomic and somatic pathways involved in male sexual responses.

Innervation

Parasympathetic pathways As described earlier with regard to micturition, three sets of nerves provide an innervation to the urogenital system (Figure 1). Parasympathetic preganglionic axons from the sacral spinal cord provide the efferent outflow to erectile tissue in the penis and to the seminal vesicles, prostate, and urethral glands. The sacral pathways have cholinergic ganglionic relay stations in the pelvic plexus and possibly in the effector organs. The postganglionic parasympathetic neurons synthesize and release several transmitters, including NO, ACh, VIP, and ATP. NO is thought to be the major transmitter mediating neurally induced erections (Figure 6), whereas ACh appears to be involved in stimulating glandular secretion. The functions of VIP and ATP are uncertain.

Sympathetic pathways The sympathetic innervation of the genital organs, consisting of preganglionic neurons in the thoracolumbar segments of the spinal cord and postganglionic neurons in the paravertebral and prevertebral ganglia (inferior mesenteric and pelvic ganglia), provides an input to the penile erectile tissue as well as to the smooth muscle of ductus deferens, seminal vesicles, urethra, and prostate. Sympathetic postganglionic axons are also carried in the pudendal nerves and in nerves arising in the pelvic plexus. Most sympathetic postganglionic neurons are noradrenergic and release norepinephrine, ATP, and neuropeptides such as NPY. These nerves produce constriction of blood vessels and cause a contraction of the vas deferens and urethra-bladder neck. Some sympathetic nerves arising in the pelvic plexus release NO and presumably induce vasodilation and penile erection.

Somatic pathways The pudendal nerve arising from the S2 to S4 segments of the spinal cord provides an efferent excitatory input to the striated muscles (bulbocavernosus and ischiocavernosus) involved in

ejaculation (Figure 1). The pudendal nerve also contains the principal afferent pathway from the penis and from the clitoris and vagina in the female. Afferent innervation to the uterine cervix and uterus travel in the pelvic nerves and sympathetic nerves, respectively.

Erection

Penile erection, which is one of the first responses to occur during sexual arousal, is a vascular phenomenon resulting from neurally mediated increase in blood flow to the penile erectile tissue (corpora cavernosa and corpus spongiosum). The erectile tissue consists of large venous sinuses that contain very little blood when the penis is flaccid, but distend considerably when blood flow is increased. Dilation in the arterial supply to the cavernous tissue, coupled with a relaxation of the sinusoidal smooth muscle in the trabecular tissue, is responsible for erection. During the initiation of erection, corpus blood flow increases 7- to 30-fold.

The neural pathways producing erection arise from both the sacral (parasympathetic) and the thoracolumbar (sympathetic) segments of the spinal cord. The postganglionic neurons in these pathways express neuronal NOS (nNOS; Figure 6). All smooth muscle in the penis is richly innervated by nerves containing nNOS. The endothelial cells in the penis also express endothelial NOS (eNOS) and can release NO in response to mechanical stimuli (shear stress) associated with changes in blood flow (Figure 6). NO directly activates soluble guanylyl cyclase in the penile smooth muscle to increase the formation of cGMP, which in turn activates cGMP-dependent protein kinases (cGK I). Inactivation of cGK I in mice severely reduces reproductive function and markedly reduces the ability of corpus cavernosal tissue to relax in response to neurally, endothelially, or exogenously administered NO. cGK I is thought to act by multiple mechanisms, including suppression of membrane and sarcoplasmic reticulum Ca^{2+} channels and suppression of IP_3-induced intracellular Ca^{2+} release. The effects of NO are terminated by the enzymatic breakdown of cGMP by phosphodiesterase (Figure 6). Pharmacologic studies in animals have shown that erections elicited by stimulation of autonomic nerves are reduced by NOS inhibitors and are enhanced by phosphodiesterase (PDE) inhibitors. One type of PDE (PDE5) is highly expressed in penile tissues. Several PDE5 inhibitors are currently used to treat erectile dysfunction.

Endothelial NOS has also been implicated in neurally mediated erections (Figure 6). While erections are initiated by NO synthesized in nerves by nNOS, it has been proposed that this mechanism triggers a very transient increase in blood flow and expansion of the penile vasculature and sinusoidal spaces. However, the resulting shear force on the endothelium activates a phosphatidylinositol 3-kinase (PI3K) pathway that in turn stimulates a serine/threonine protein kinase (Akt, protein kinase B), causing direct phosphorylation of eNOS. After phosphorylation, the Ca^{2+} requirement for eNOS activity is reduced, causing increased production of NO, which induces a prolonged penile erection. Intercellular communication via gap junctions is thought to promote the spread of signals throughout the smooth muscle of the penis and amplify the relaxation.

Exogenous ACh acts on postjunctional muscarinic receptors to elicit a contraction or relaxation of *in vitro* preparations of the corpus cavernosum. The latter effect is mediated by M_3 receptors and release of NO from the endothelium. However, exogenous ACh does not produce erection *in vivo*, and atropine, a muscarinic receptor antagonist, does not block erections, indicating that cholinergic mechanisms are not essential for neurally mediated increases in penile blood flow. On the other hand, ACh may act on prejunctional inhibitory receptors on adrenergic nerve terminals to suppress the release of norepinephrine, leading to a decrease in noradrenergic vasoconstriction in the penis and a facilitation of NO-mediated erections.

VIP activates G-protein-coupled receptors to stimulate adenylyl cyclase and increase cAMP, which in turn stimulates cAMP-dependent kinases to suppress contractile mechanisms in vascular smooth muscle and smooth muscle of the trabecular tissue of the penis (Figure 7). While it is clear that exogenous VIP can relax human cavernosal tissue strips *in vitro*, it has never been shown that VIP released from nerves is responsible for relaxation of penile smooth muscle *in vitro* or *in vivo*.

The sympathetic adrenergic innervation of the penis that provides an excitatory input to penile blood vessels is thought to be involved primarily in detumescence. Electrical stimulation of sympathetic axons in either the hypogastric or pudendal nerves in various animals produces a substantial reduction in penile blood flow. The effect is blocked by α-adrenergic receptor-blocking agents. In some animals, a partial erection can be elicited by the administration of α-adrenergic blocking agents, suggesting that tonic sympathetic vasoconstrictor mechanisms have an inhibitory influence on erection. Several mechanisms have been implicated in the noradrenergic vasoconstrictor effect, including (1) prejunctional inhibition of parasympathetic nitrergic nerve terminals mediated by α_2-adrenergic receptors (Figure 7); (2) activation of postjunctional α_2-adrenergic receptors, which then

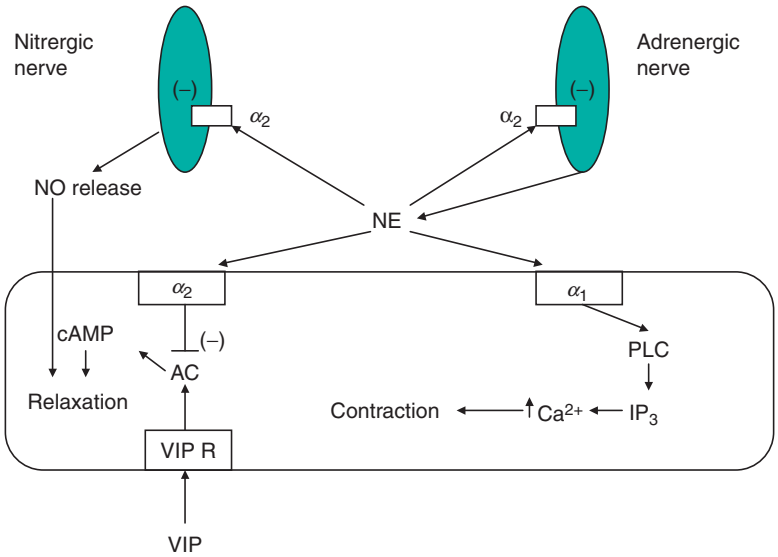

Figure 7 Diagram showing the reciprocal regulation of the contractile tone of penile smooth muscle by norepinephrine (NE) and vasoactive intestinal polypeptide (VIP). NE released from an adrenergic nerve acts on postjunctional α_1-adrenergic receptors in smooth muscle to stimulate the formation of inositol trisphosphate (IP_3) via a phospholipase C (PLC) mechanism. IP_3 triggers the intracellular release of Ca^{2+} that induces a muscle contraction and penile detumescence. NE also can act on α_2-adrenergic receptors located on nerve terminals to suppress transmitter release and on postjunctional receptors to inhibit adenylyl cyclase (AC), and in turn suppress the production of cyclic adenosine monophosphate (cAMP) that induces smooth muscle relaxation and penile erection. VIP acts on receptors (VIP R) to induce relaxation and erection by stimulating AC and increasing cAMP.

inhibits adenylate cyclase and formation of cAMP; (3) activation of postjunctional α_1-adrenergic receptors, which stimulates phospholipase C activity followed by formation of IP_3 and diacylglycerol, leading to release of intracellular Ca^{2+} as well as sensitization of contractile mechanisms to Ca^{2+} (Figure 7). Ca^{2+} sensitization has been linked to activation of an intracellular signaling pathway involving Rho kinase, which can be stimulated by G-protein-coupled α-adrenergic receptors. Activation of Rho kinase leads to a change in the phosphorylation state of myosin light-chain kinase, resulting in phosphorylation of myosin and subsequent smooth muscle contraction. Administration of a Rho kinase inhibitor can trigger an increase in intracavernous pressure and an NO-independent erection. Similarly, the expression of a dominant negative construct to downregulate Rho kinase enhances erectile function in rats. Other studies have revealed that NO can relax penile smooth muscle by suppressing Rho kinase activity. These studies have raised the possibility that antagonism of Rho kinase activity may yield new treatments for erectile dysfunction.

Glandular Secretion

During the second phase of the sexual response cycle (plateau), activity in parasympathetic pathways stimulates mucus secretion from bulbourethral and Littre's glands and secretion from the seminal vesicles and the prostate gland. Mucus secretion contributes to lubrication of the penis, whereas secretions from the seminal vesicles and prostate provide the bulk of the fluid and chemical factors that contribute to the viability and motility of the spermatozoa. Glandular secretion is regulated by parasympathetic nerves; however, the transmitters have not been identified. ACh may be involved since cholinomimetic agents stimulate secretion from some glands and AChE-containing nerves have been identified in the prostate and seminal vesicles. However, VIP-containing nerves are also present in these organs.

Emission–Ejaculation

The third phase of the sexual act (orgasm), which is accompanied by emission and ejaculation of semen, involves the coordination of autonomic and somatic reflex mechanisms at different levels of the lumbosacral spinal cord. During the first step in this process (emission), reflex activity in the thoracolumbar sympathetic outflow elicits rhythmic contractions of the smooth muscle of the seminal vesicles, prostate, ductus deferens, and ampulla, resulting in the ejection of sperm and glandular secretions into the urethra and at the same time a closure of the vesical neck to prevent the backflow of semen into the bladder. Pharmacologic studies have shown that these responses are mediated by the sympathetic nerve transmitters, norepinephrine and ATP, interacting with α-adrenergic

receptors and purinergic receptors, respectively. Surgical interruption of sympathetic nerves, or the administration of drugs that block α-adrenergic receptors, deplete norepinephrine stores, or block norepinephrine release, blocks emission.

After emission of semen into the proximal urethra, rhythmic contractions of the bulbocavernosus, ischiocavernosus, and the paraurethral striated muscles result in ejaculation. The afferent and efferent limbs of the ejaculation reflex are contained in the pudendal nerve. The sensations accompanying ejaculation constitute the orgasm.

See also: Autonomic Nervous System: Central Urogenital Control; Autonomic Nervous System; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Andersson KE (2003) Erectile physiological and pathophysiological pathways involved in erectile dysfunction. *Journal of Urology* 170: S6–S13.

Andersson KE and Arner A (2004) Urinary bladder contraction and relaxation: Physiology and pathophysiology. *Physiological Reviews* 84: 935–986.

Beckel JM, Kanai A, Lee SJ, et al. (2006) Expression of functional nicotinic acetylcholine receptors in rat urinary bladder epithelial cells. *American Journal of Physiology – Renal Physiology* 290: F103–F110.

Birder LA, Nakamura Y, Kiss S, et al. (2002) Altered urinary bladder function in mice lacking the vanilloid receptor TRPV1. *Nature Neuroscience* 5: 856–860.

Burnett AL (2001) Novel pharmacological approaches in the treatment of erectile dysfunction. *World Journal of Urology* 19: 57–66.

Burnett AL, Chang AG, Crone JK, et al. (2002) Noncholinergic penile erection in mice lacking the gene for endothelial nitric oxide synthase. *Journal of Andrology* 23: 92–97.

Chancellor MB and Yoshimura N (2002) Physiology and pharmacology of the bladder and urethra. In: Walsh PC, Retik AB, Vaughn ED, et al. (eds.) *Campbell's Urology*, pp. 831–886. Philadelphia: WB Saunders.

Cockayne DA, Hamilton SG, Zhu QM, et al. (2000) Urinary bladder hyporeflexia and reduced pain-related behaviour in P2X3-deficient mice. *Nature* 407: 1011–1015.

de Groat WC (2004) The urothelium in overactive bladder: Passive bystander or active participant? *Urology* 64: 7–11.

de Groat WC and Booth AM (1993) Synaptic transmission in pelvic ganglia. In: Maggi CA (ed.) *The Autonomic Nervous System, Volume 3, Nervous Control of the Urogenital System*, pp. 291–347. London: Harwood Academic Publishers.

Fry CH, Brading AF, Hussain M, et al. (2005) Cell biology. In: Abrams P, Cardozo L, Khoury S, et al. (eds.) *Incontinence*, pp. 363–422. Jersey: Health Publications.

Matsui M, Motomura D, Karasawa H, et al. (2000) Multiple functional defects in peripheral autonomic organs in mice lacking muscarinic acetylcholine receptor gene for the M3 subtype. *Proceedings of the National Academy of Science USA* 97: 9579–9584.

Morrison J, Birder L, Craggs M, et al. (2005) Neural control. In: Abrams P, Cardozo L, Khoury S, et al. (eds.) *Incontinence*, pp. 363–422. Jersey: Health Publications.

Saenz de Tejada I, Cavidad NG, Heaton J, et al. (2001) Anatomy, physiology, and pathophysiology of erectile function. In: Jardin A, Wagner G, Khoury S, et al. (eds.) *Erectile Dysfunction*, pp. 65–102. Jersey: Health Publications.

Toda N, Ayajiki K, and Okamura T (2005) Nitric oxide and penile erectile function. *Pharmacology & Therapeutics* 106: 233–266.

Autonomic Neuroeffector Junction

G Burnstock, Royal Free and University College Medical School, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Within the past 30 years, new discoveries have changed our understanding of the organization of the autonomic nervous system (ANS), including the structure of the autonomic neuroeffector junction and the multiplicity of neurotransmitters which take part in the process of autonomic neuroeffector transmission, as well as cotransmission, neuromodulation, receptor expression, and long-term (trophic) signaling. An outstanding feature of autonomic neurotransmission is the inherent plasticity afforded by its structural and neurochemical organization and the interaction between neural mediators and environmental factors. In this way, autonomic neurotransmission is matched to ongoing changes in demands and can sometimes be compensatory in pathophysiological situations.

Structure of the Autonomic Neuromuscular Junction

The autonomic neuromuscular junction differs in several important respects from the better known skeletal neuromuscular junction; it is not a synapse with the well-defined prejunctional and postjunctional specializations established for the skeletal neuromuscular synapse or ganglionic synapses. A model of the autonomic neuroeffector junction has been proposed on the basis of combined electrophysiologic, histochemical, and electron-microscopical studies. The essential features of this model are that the terminal portions of autonomic nerve fibers are varicose, transmitters being released *en passage* from varicosities during conduction of an impulse, although excitatory and inhibitory junction potentials are probably elicited only at close junctions. Furthermore, the effectors are muscle bundles rather than single smooth muscle cells and are connected by low-resistance pathways (gap junctions) that allow electrotonic spread of activity within the effector bundle. In blood vessels, the nerves are confined to the adventitial side of the media muscle coat, and this geometry appears to facilitate dual control of vascular smooth muscle by perivascular nerves and by endothelial relaxing and contracting factors. Neuroeffector junctions do not have a permanent geometry with postjunctional specializations, but rather the varicosities are continuously moving, and their special relation with muscle cell membranes changes with time, including dispersal and reformation of receptor clusters. For example, varicosity movement is likely to occur in cerebral blood arteries, where there is a continuously increasing density of sympathetic innervation during development and aging and in hypertensive vessels or those that have been stimulated chronically *in vivo*, where there can be an increase in innervation density of up to threefold.

Varicose Terminal Axons

In the vicinity of the effector tissue, axons become varicose, varicosities occurring at 5–10 μm intervals (Figure 1(a)), and branches intermingle with other axons to form the autonomic ground plexus, first described by Hillarp in 1946. The extent of the branching and the area of effector tissue affected by individual neurons vary with the tissue. Autonomic axons combined in bundles are enveloped by Schwann cells. Within the effector tissue, they partially lose their Schwann cell envelope, usually leaving the last few varicosities naked.

The density of innervation, in terms of the number of axon profiles per 100 muscle cells in cross-section, also varies considerably in different organs. For example, it is very high in the vas deferens (Figure 2(a)), iris, nictitating membrane, and sphincteric parts of the gastrointestinal tract but low in the ureter, uterus, and longitudinal muscle coat of the gastrointestinal tract. In most blood vessels, the varicose nerve plexus is placed at the adventitial border, and fibers rarely penetrate into the medial muscle coat (Figure 2(b)).

Junctional Cleft

The width of the junctional cleft varies considerably in different organs. In the vas deferens, nictitating membrane, sphincter pupillae, rat parotid gland, and atrioventricular and sinoatrial nodes in the heart, the smallest neuromuscular distances range from 10–30 nm. The minimum neuromuscular distance varies considerably in different blood vessels. Generally, the greater the vessel diameter, the greater the separation of nerve and muscle. Thus, minimal neuromuscular distances in arterioles and in small arteries and veins are about 50–100 nm, in medium to large arteries the separation is 200–500 nm, whereas in large elastic arteries where the innervation is sparser, the minimum neuromuscular distances are as wide as 1000–2000 nm. Serial sectioning has shown that at close junctions in both visceral and vascular organs, there is fusion of prejunctional and postjunctional

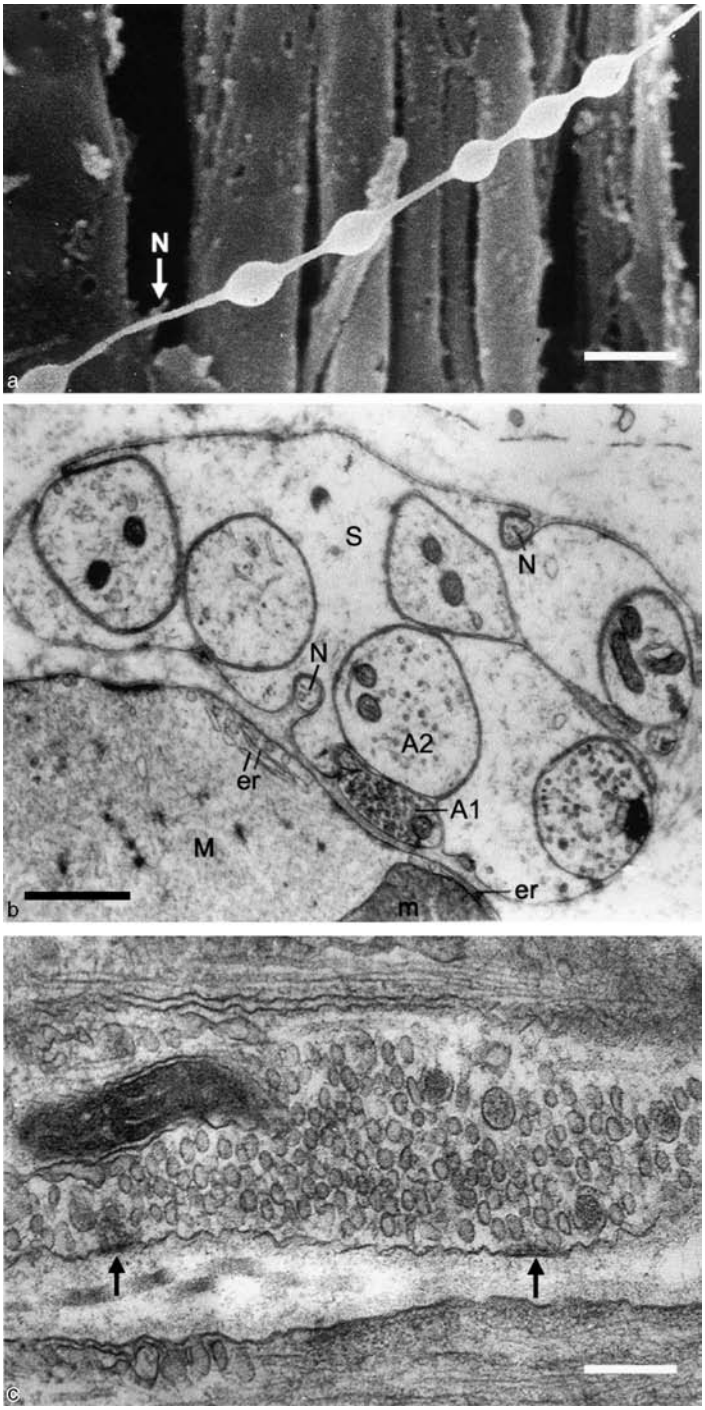

Figure 1 (a) Scanning electron micrograph of a single terminal varicose nerve fiber(N) lying over smooth muscle of small intestine of rat. Intestine was pretreated to remove connective tissue components by digestion with trypsin and hydrolysis with HCl. (b) A medium-size intramuscular bundle of axons within a single Schwann cell (S). There is no perineurial sheath. Some axons, free of Schwann cell processes, contain synaptic vesicles (e.g., A1 and A2). For nerve profile A1, there is close proximity (about 80 nm) to smooth muscle (M) with fusion of nerve and muscle basement membranes. Most of the axons in bundles of this size have few vesicles in the plane of section, but they resemble the vesicle-containing axons of the larger trunks in that they have few large neurofilaments. The small profiles (N), less than 0.25 μm in diameter, are probably intervaricosity regions of terminal axons. m, mitochondria; er, edoplasmic reticulum. (c) Autonomic varicosities with dense prejunctional thickenings and bunching of vesicles, probably representing transmitter release sites (arrows), but there is no postjunctional specialization. Scale bar = 3 μm (a), 1 μm (b), and 0.25 μm (c). (a) Reproduced from *The Airways*: *Neural Control in Health and Disease*, 1988, 1–22, Autonomic neural control mechanisms: With special reference to the airways, Burnstock G, figure 1, copyright Marcel and Dekker. With kind permission of Springer Science and Buisness Media. (b) Reproduced from *The Journal of Cell Biology*, 1963, vol. 19, pp. 529–550, by copyright permission of The Rockfeller University Press. (c) Courtesy of Phillip R Gordon-Weeks.

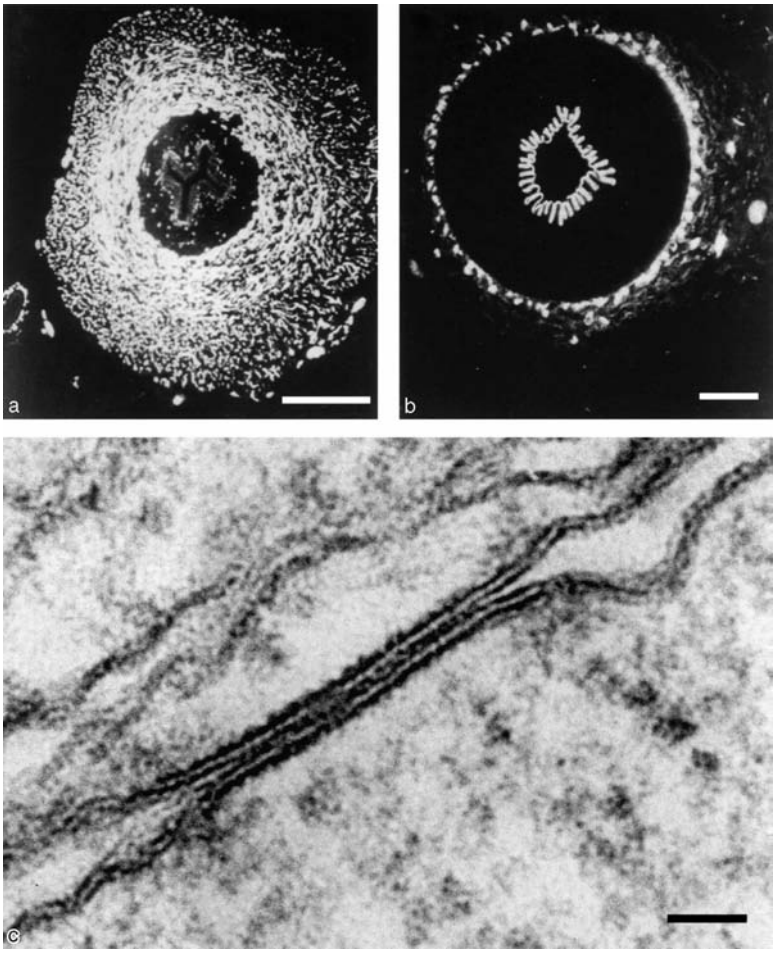

Figure 2 Comparison between the adrenergic innervation of the densely innervated vas deferens of the guinea pig (a) and the rabbit ear artery (b), in which the adrenergic fibers are confined to the adventitial-medial border. The inner elastic membrane shows a nonspecific fluorescence (autofluorescence). (c) A gap junction between two smooth muscle cells grown in tissue culture. Scale bar = 500 μm (a), 50 μm (b), and 50 nm (c). (a, b) Reproduced from *Adrenergic Neurones: Their Organisation, Function and Development in the Peripheral Nervous System*, 1975, Burnstock G and Costa M, plate 9, copyright Chapman and Hall. With kind permission of Springer Science and Buisness Media. (c) Reproduced from *The Journal of Cell Biology*, 1971, Vol. 49, pp. 21–34, by copyright permission of The Rockfeller University Press.

basal lamina (see Figure 1(b)). In the longitudinal muscle coat of the gastrointestinal tract, autonomic nerves and smooth muscle are rarely separated by less than 100 nm. However, in the circular muscle coat, close (20 nm) junctions are common, sometimes several axon profiles being closely apposed with single muscle cells.

Prejunctional and Postjunctional Specialization

Although there are many examples of prejunctional thickenings of nerve membranes in varicosities associated with accumulations of small synaptic vesicles, representing sites of transmitter release (see Figure 1(c)), there are no convincing demonstrations of postjunctional specializations, such as membrane thickening or folding or indeed absence of micropinocytic vesicles; this is in keeping with the view that even close junctions might be temporary liaisons.

Muscle Effector Bundles and Gap Junctions

The smooth muscle effector is a muscle bundle rather than a single muscle cell – that is, individual muscle cells are connected by low-resistance pathways that allow electrotonic spread of activity within the effector bundle. Sites of electrotonic coupling are represented morphologically by areas of close apposition between the plasma membranes of adjacent muscle cells. High-resolution electron micrographs have shown that the membranes at these sites consist of gap junctions (see Figure 2(c)). Gap junctions (or nexuses) vary in size between punctate junctions,

which are not easily recognized except in freeze-fracture preparations, and junctional areas more than 1 μm in diameter. The number and arrangement of gap junctions in muscle effector bundles of different sizes in different organs and their relation to density of autonomic innervation have not been fully analyzed. It is interesting that partial denervation has been shown to result in an increase in gap junctions.

Receptor Localization on Smooth Muscle Cells

The distribution of P2X purinoceptors on smooth muscle cells in relation to autonomic nerve varicosities in urinary bladder, vas deferens, and blood vessels has been examined recently by using immunofluorescence and confocal microscopy. Antibodies against the $P2X_1$ receptor, the dominant receptor subtypes found in smooth muscle, and an antibody against the synaptic vesicle proteoglycan SV2 showed clusters of receptors (about 0.9×0.2 μm in size) located beneath varicosities. Many more small clusters (about 0.4×0.04 μm) were present on the whole surface of smooth muscle cells unrelated to varicosities; they may represent pools of receptors that can migrate toward varicosities to form large clusters. In blood vessels, small clusters of P2X receptors are present on cells throughout the medial muscle coat, whereas large clusters are restricted to the muscle cells at the adventitial surface. α_2-Adrenoceptors appear to be located only in extrajunctional regions, so the possibility that noradrenaline (NA) is released from more distant varicosities has been raised. There are hints from studies of receptor-coupled green fluorescent protein chimeras that the receptor clusters are labile, dispersing when a varicosity moves to a new site where clusters reform, perhaps within a 20–30 min timescale.

Model of Autonomic Neuroeffector Junction

A model of the autonomic neuromuscular junction has been proposed on the basis of combined electrophysiologic, histochemical, and electron-microscopical studies described earlier (**Figures 3(a) and 3(b)**). The essential features of this model are that the terminal portions of autonomic nerve fibers are varicose, transmitter being released *en passage* from varicosities during conduction of an impulse, although excitatory junction potentials (EJPs) and inhibitory junction potentials are probably elicited only at close junctions. Furthermore, the effectors are muscle bundles rather than single smooth muscle cells, which are connected by low-resistance pathways (gap junctions) that allow electrotonic spread of activity within the effector bundle. In blood vessels, the nerves are confined to the adventitial side of the media muscle coat, and this geometry appears to facilitate dual control of vascular smooth muscle by endothelial relaxing and contracting factors and perivascular nerves.

Neuroeffector junctions do not have a permanent geometry with postjunctional specializations, but rather the varicosities are continuously moving, and their special relation with muscle cell membranes changes with time. For example, varicosity movement is likely to occur in cerebral blood arteries, where there is a continuously increasing density of sympathetic innervation during development until old age, and in vessels that have been stimulated chronically *in vivo*, where there can be an increase in innervation density of up to threefold, including an increase in the number of varicosities per unit length of nerve from 10–20 per 100 μm to 30 per 100 mm.

Autonomic effector junctions appear to be suitable not only for neurotransmission but also for neuromodulation. A neuromodulator is defined as any substance that modifies the process of neurotransmission. It may achieve this either by prejunctional action that increases or decreases transmitter release or by postjunctional action that alters the time course or extent of action of the transmitter or by both (**Figure 3(c)**).

Finally, it should be emphasized that if this model of the autonomic effector junction is true, then the earlier emphasis on looking for images of specialized nerve-cell close apposition may not be appropriate. If a varicosity has a passing close relation with a cell and releases transmitter to act on receptors expressed on that cell (e.g., most cells, epithelial cells, or even immune cells), then, in effect, that cell is innervated.

Autonomic Neurotransmission

The Multiplicity of Neurotransmitters in the Autonomic Nervous System

A neurotransmitter is a chemical substance released from nerves on electrical stimulation and which acts on specific receptors on adjacent effector cells to bring about a response, thus acting as a chemical messenger of neural activation. In early studies, acceptance of a substance as a neurotransmitter required satisfaction of the following criteria: (1) the presynaptic neuron synthesizes and stores the transmitter; (2) the transmitter is released in a calcium-dependent manner; (3) there should be a mechanism for terminating the activity of the transmitter, either by enzymatic degradation or by cellular uptake; (4) local exogenous application of the substance should mimic its effects following release due to electrical nerve stimulation; and (5) agents that block or

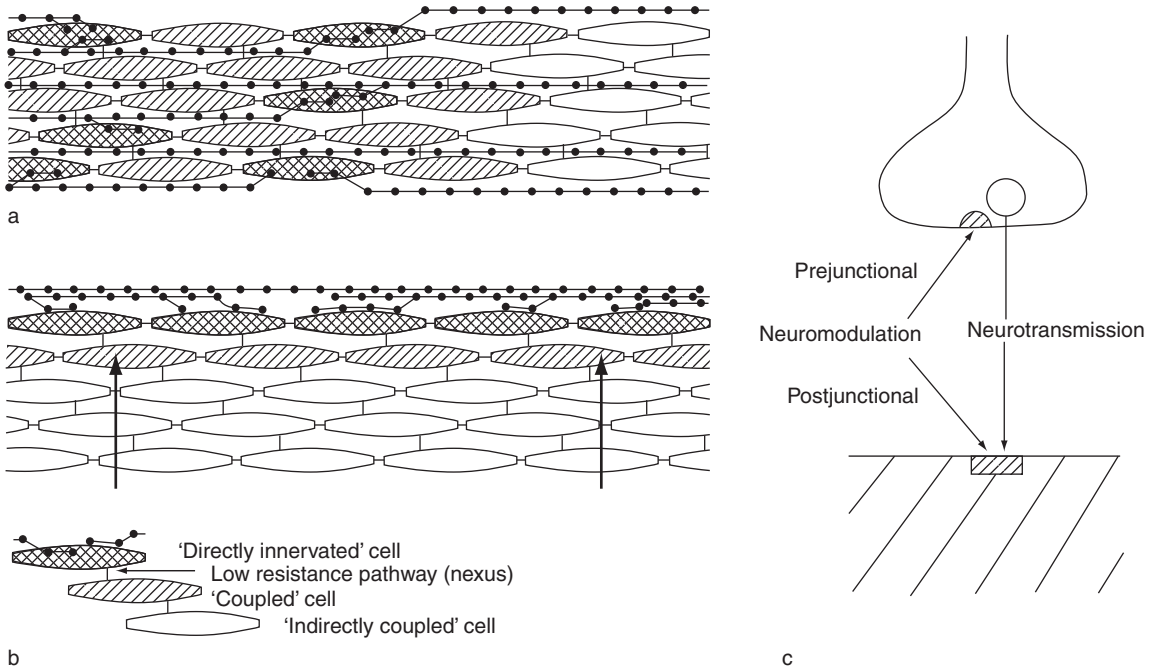

Figure 3 (a) Schematic representation of control of visceral smooth muscle. Directly innervated cells (cross-hatched) are those that are directly activated by neurotransmitter; coupled cells (hatched) are those where junction potentials spread from directly innervated cells, when a sufficient area of the muscle effector bundle is depolarized, a propagated action potential will activate the indirectly coupled cells (white). (b) Schematic representation of control of vascular smooth muscle by nerves (-•-•-) and endothelial factors (arrows). (c) Schematic representation of prejunctional and postjunctional neuromodulation. (a, b) Reproduced from *Adrenergic Neurones: Their Organisation, Function and Development in the peripheral Nervous System*, 1975, Burnstock G and Costa M, figure 18 (a) and (b), copyright Chapman and Hall. With kind permission of Springer Science and Buisness Media. (c) Reproduced from: Royal College of Physicians. *Advanced Medicine 18*. Sanner M (ed). London: Pitman Medical, 1982. Copyright © 1982 Royal College of Physicians. Reproduced by permission.

potentiate the endogenous activity of the transmitter should also affect local exogenous application in the same way.

The classical view of autonomic nervous control as antagonistic actions of NA and acetylcholine (ACh) causing either constriction or relaxation, depending on the tissue, was changed in the early 1960s when clear evidence of a nonadrenergic, noncholinergic (NANC) system was presented. About a decade later, studies of autonomic neurotransmission revealed a multiplicity of neurotransmitters in the ANS. Neurally released substances, including monoamines, amino acids, neuropeptides, adenosine 5′-triphosphate (ATP), and nitric oxide (NO) were identified (see Table 1). Since NO does not conform to the constraints of the criteria outlined earlier, although it certainly acts as a rapid chemical messenger in the ANS, a reappraisal of the criteria for defining a neurotransmitter was proposed by Hoyle and Burnstock in 1996, taking into account evidence for nonvesicular, Ca^{2+}-independent release of some classical neurotransmitters and the intracellular site of action of NO. The rapid expansion of the number of proposed autonomic neurotransmitters in recent years, including endothelin, secretoneurin, pituitary adenylate cyclase-activating peptide (PACAP), which is similar in structure to vasoactive intestinal polypeptide (VIP), glutamate, and carbon monoxide, makes it likely that the list is still incomplete.

Cotransmission

The concept of cotransmission was first formulated by Burnstock in 1976, incorporating hints in the earlier literature from both vertebrate and invertebrate systems. It is now well established. Immunohistochemical evidence of coexistence of more than one neurotransmitter should not necessarily be interpreted as evidence of cotransmission, since in order for substances to be termed cotransmitters, it is essential that postjunctional actions to each substance be shown to occur via their own specific receptors. For example, many neuropeptides have slow trophic actions on surrounding tissues, and this may be their primary role, or they may act as neuromodulators. The relative contribution of each transmitter to neurogenic responses is dependent on the parameters of

Table 1 Established and putative neurotransmitters/neuromodulators in the autonomic nervous system

Noradrenaline (NA)
Acetylcholine (ACh)
Adenosine 5′-triphosphate (ATP) and other nucleotides
Nitric oxide (NO)
Carbon monoxide (CO)
5-Hydroxytryptamine (5-HT)
Dopamine (DA)
γ-Aminobutyric acid (GABA)
Glutamate (GLU)
Neuropeptides
Neuropeptide Y (NPY)/pancreatic polypeptide (PP)
Enkephalin (ENK)/endorphin (END)/dynorphin (DYN)
Vasoactive intestinal polypeptide (VIP) and related peptides PHI and PHM
Pituitary adenylate cyclase-activating peptide (PACAP)
Substance P (SP)/neurokinin A (NKA)/neurokinin B (NKB)
Calcitonin gene-related peptide (CGRP)
Somatostatin (SOM)
Galanin (GAL)
Gastrin releasing peptide (GRP)/bombesin (BOM)
Neurotensin (NT)
Cholecystokinin (CCK)/gastrin (GAS)
Angiotensin II (AII)
Adrenocorticotrophic hormone (ACTH)
Secretoneurin
Endothelin (ET)

Reproduced from Burnstock G (2007) Structural and chemical organization of the autonomic nervous system with special reference to nonadrenergic, noncholinergic transmission. In Mathias CJ and Bannister R (eds.) *Autonomic failure. A textbook of clinical disorders of the autonomic nervous system.* Oxford: Oxford University Press, by permission of Oxford University Press.

stimulation. For example, short bursts (1 s) of electrical stimulation of sympathetic nerves at low frequency (2–5 Hz) favor ATP release whereas longer periods of nerve stimulation (30 s or more) favor NA release.

Peptides, purine nucleotides, and NO (identified by localization of nitric oxide synthase (NOS)) are often found together with the classic neurotransmitters, NA and ACh. In fact, the majority of nerve fibers in the ANS, if not all, contain a mixture of different neurotransmitter substances that vary in proportion in different tissues and species and during development and disease. The widespread use of double and triple immunohistochemical labeling techniques has been critical to the demonstration of co-localization of potential cotransmitters within the same nerve fiber and has been invaluable when combined with electron microscopy. Different neurotransmitters within the same varicosity may be localized in the same or separate vesicular populations using post-embedding colloidal gold techniques. In the gastrointestinal tract, many neurons contain multiple transmitters. ATP is a cotransmitter with calcitonin gene-related peptide (CGRP) and substance P (SP) in many sensory–motor nerves and with NO and VIP in enteric NANC inhibitory nerves. Transmitters with seemingly diverse and opposing effector action are sometimes co-localized in the same neuron, but generally they act in the same way and usually synergistically.

The precise combinations of neurotransmitters (and neuromodulators) contained in individual neurons and their projections and central connections, termed their 'chemical coding' by Furness and Costa in 1997, has been defined in studies of the enteric nervous system and peripheral autonomic and sensory ganglia.

Neurotransmission at the sympathetic neuroeffector junctions: Evidence for co-release and roles of NA, ATP, and neuropeptide Y It is now recognized that the main neurotransmitters/neuromodulators in postganglionic sympathetic nerves are NA, ATP, and neuropeptide Y (NPY). These substances are co-released in varying proportions, depending on the tissue and species, and also on the parameters of stimulation. Short bursts at low frequency particularly favor the purinergic component whereas longer periods of nerve stimulation favor the adrenergic component, and NPY release is optimal with high-frequency intermittent bursts of stimulation. A considerable variability in the contribution of a purinergic component to sympathetic neurotransmission has been demonstrated in different blood vessels; for example, rabbit saphenous and mesenteric arteries have a substantial purinergic component, whereas in the rabbit ear artery, the purinergic component is relatively small. In intestinal submucosal arteries, the responses to sympathetic nerve stimulation are mediated solely by ATP, with NA acting as a prejunctional modulator via α_2-adrenoceptors. The initial electrophysiological postjunctional response to sympathetic nerve stimulation is a rapid, transient EJP, which is mediated by ATP. In some vessels, the EJP is followed by a slow depolarization, which is mediated by NA. Postjunctionally, the effects of ATP and NA released as sympathetic cotransmitters are generally synergistic. NA and ATP can depress sympathetic neurotransmission by prejunctional modulation, via α_2-adrenoceptors or predominantly via P1 receptors following extracellular breakdown to adenosine, but also via P2 receptors in some vessels. Prejunctional P2 receptor-mediated increase in NA release has also been reported.

In most tissues, including the vas deferens and many blood vessels, NPY does not act as a genuine neurotransmitter, having little direct postjunctional action, but rather acts as a neuromodulator, often by prejunctional attenuation of NA and ATP release

and/or postjunctional potentiation of responses to adrenergic and purinergic components of sympathetic nerve responses. In tissues in which NPY does have a direct vasoconstrictor effect, such as in blood vessels of the spleen and kidney and in coronary and cerebral arteries, the response is characteristically slow in onset and long lasting.

Other substances localized within sympathetic nerves include 5-hydroxytryptamine (5-HT), which is largely taken up by sympathetic nerves and released as a false transmitter. Opioid peptides are also widely distributed in sympathetic neurons where their functional role appears to be related to their prejunctional inhibitory effects on sympathetic neurotransmission.

Neurotransmission at the parasympathetic neuroeffector junctions: The atropine-resistant components of parasympathetic neurotransmission ACh, VIP, ATP, and NO are cotransmitters commonly synthesized in and released from parasympathetic nerves. As with sympathetic cotransmission, the relative functional importance of the cotransmitters in parasympathetic neurotransmission is variable in different tissues and species. For example, NO may be the main mediator of neurogenic vasodilation in cerebral vessels, whereas VIP may be of more importance during neurogenic vasodilation in the pancreas. The coordinated roles of VIP and ACh in parasympathetic neurotransmission were demonstrated in an elegant study of the cat exocrine salivary gland innervation. It showed that VIP and ACh were stored in separate vesicles in the same nerve terminal and were both released on transmural nerve stimulation, but with different stimulation parameters. ACh was released during low-frequency stimulation to increase salivary secretion from acinar cells and to elicit some minor dilatation of blood vessels in the gland. At high stimulation frequencies, VIP was released to produce marked dilatation of the blood vessels in the gland and to act as a neuromodulator postjunctionally on the acinar gland to enhance the actions of ACh and prejunctionally on the nerve varicosities to enhance the release of ACh. ACh was also found to have an inhibitory action on the release of VIP. VIP has since been shown to have a direct vasodilatory action in the submandibular gland in man. PACAP also seems to be present in VIP-containing parasympathetic nerves. NOS is often co-localized with ACh and VIP in parasympathetic nerves innervating blood vessels. Postganglionic nerves from pelvic ganglia containing VIP, ACh, and NOS project to the urethra, colon, and penis. The human bladder body receives a dense parasympathetic innervation comprised predominantly of ACh-containing nerves. In the rodent bladder, ATP is a major cotransmitter in these nerves. However, only a small purinergic component is present in human bladder, except in pathological conditions (discussed later).

Neurotransmission at sensory–motor neuroeffector junctions: The roles of SP, CGRP, and ATP The motor function of sensory nerves, whereby antidromic impulses down collateral fibers result in local release of sensory neurotransmitters, is widespread in autonomic effector systems and forms an important physiological component of autonomic control. To distinguish these nerves from the other subpopulation of afferent fibers that have an entirely sensory role and have terminals containing few vesicles and a predominance of mitochondria, they have been termed 'sensory–motor' nerves.

SP and CGRP are cotransmitters in many unmyelinated, primary afferent nerves. They often coexist in the same large granular vesicles in capsaicin-sensitive nerve terminals. The proportions of coexistence of SP and CGRP vary with species; for example, in the guinea pig, most sensory neurons containing CGRP also contain SP, but in the rat, about 50% of CGRP-containing neurons do not contain SP. In the vasculature, unlike CGRP, SP does not appear to act directly on receptors of the vascular smooth muscle but rather acts via occupation of receptors on endothelial cells lining the lumen to bring about NO release and consequent vasodilation. This action of neurally released SP may be particularly important in the microvasculature, but access of neurally released SP to the endothelium in large vessels is questionable; it is largely released from endothelial cells to act on receptors on endothelial cells to release NO, resulting in vasodilation. ATP is now also established as a cotransmitter with glutamate in small primary sensory nerves mediating mechanical and/or nociceptive signals.

Other neuropeptides and transmitters have been localized in sensory–motor nerves. For example, in the human urinary bladder, VIP, cholecystokinin (CCK), and dynorphin (DYN) are present, together with SP and CGRP, in the afferent projections to the lumbosacral spinal cord. In the guinea pig, dorsal root ganglion neurons containing SP, CGRP, CCK, and DYN project to the epidermis and small dermal blood vessels. NOS has been localized in populations of primary sensory neurons of trigeminal and dorsal root ganglia. Endothelin, a potent vasoconstrictor peptide with mitogenic actions, is also localized in neurons of these sensory ganglia, often co-localized with SP.

There are increasing examples in the literature of cross-talk between sensory–motor, sympathetic, and parasympathetic nerves. In the heart, SP has

excitatory effects on cardiac parasympathetic innervation, in contrast to CGRP, which is inhibitory.

Neurotransmission involving intrinsic neurons: Special reference to neurotransmitters localized in nerve cell bodies in the heart, bladder, intestine, and lung
Many intrinsic neurons localized within autonomic neuroeffector tissues are part of the postganglionic parasympathetic system, but there are also intrinsic neurons derived from neural crest tissue that is different from that which forms sympathetic and parasympathetic neurons, such as intrinsic neurons abundant in the gut and possibly subpopulations in the heart and airways.

The most extensive system of intrinsic neurons is in the myenteric and submucous plexuses of the gastrointestinal tract. These enteric neurons contain numerous neuroactive substances, of which the majority are involved in neurotransmission or neuromodulation at the ganglion level and/or have a trophic role; only a small percentage are involved in neuromuscular transmission. The chemical coding of enteric neurons has been examined in detail, particularly in the guinea pig. ATP, NO, and VIP mediate NANC inhibitory neurotransmission in the gut in varying proportions depending on the region. ACh and SP are cotransmitters in enteric excitatory neurons.

There are many intrinsic neurons in the heart, particularly in the right atrium. The neurochemical makeup of the intrinsic cardiac ganglia is heterogeneous and includes a variety of neurochemical markers. For example, subpopulations of atrial intrinsic neurons from newborn guinea pigs immunostain for NPY, 5-HT, heme oxygenase-2, and NOS, and these neurons probably also utilize ACh and ATP.

Most airway intrinsic neurons contain choline acetyltransferase, but NOS and VIP are also found in these neurons in humans. Intrinsic ganglia in the human urinary bladder wall contain a number of neuroactive substances (VIP, NOS, NPY, ATP, galanin, and occasionally tyrosine hydroxylase); in the bladder neck, a few intrinsic neurons contain enkephalin and SP. Intramural ganglia containing NPY and VIP have been identified in human urethra.

Autonomic Neuromodulation

Some substances stored and released from nerves do not have direct actions on effector muscle cells but alter the release and/or the actions of other transmitters; these substances are termed neuromodulators. Many other substances (e.g., circulating neurohormones; locally released agents such as prostanoids, bradykinin, histamine, and endothelin; and neurotransmitters from nearby nerves) are also neuromodulators in that they modify the process of neurotransmission. Many substances that are cotransmitters are also neuromodulators. The wide and variable cleft characteristic of autonomic neuroeffector junctions makes them particularly amenable to the mechanisms of neural control mentioned earlier.

Plasticity of the Autonomic Nervous System

There are some examples of altered expression of neurotransmitters/neuromodulators in autonomic nerves during development and aging; following trauma, surgery, and chronic exposure to drugs; and in disease. Neurons possess the genetic potential to produce many neurotransmitters. The particular combination and quantity that result are partly preprogrammed and partly determined by 'trophic' factors and hormones that trigger the expression or suppression of the appropriate genetic machinery. The plasticity of expression of neural substances cocoordinated to environmental cues allows rapid and precise matching of neurotransmission to altered demands. Several neurotransmitters/neuromodulators are themselves trophic molecules, with mitogenic or growth-promoting/-inhibiting properties.

Conclusions

A combination of the variety of neurotransmitters involved in autonomic neurotransmission and the interactions between sympathetic, parasympathetic, and sensory–motor nerves and those arising from intrinsic ganglia, via mechanisms of cotransmission and pre- and postjunctional neuromodulation, indicate the complexity of peripheral autonomic control and the variety of ways by which autonomic dysfunction can occur. Recent advances in the unraveling of these mechanisms, together with molecular identification of specific receptor subtypes and localization and characterization of their expression, and of the long-term effects of dysfunction, will bring advances toward the design of treatment regimes to combat autonomic failure.

See also: Autonomic Neuroplasticity and Aging; Autonomic Neuroplasticity: Development; Autonomic Neuroplasticity and Regeneration; Autonomic Nervous System; Cotransmission; Gap Junction Abnormalities and Disorders of the Nervous System.

Further Reading

Burnstock G (1976) Do some nerve cells release more than one transmitter? *Neuroscience* 1: 239–248.

Burnstock G (1982) Neuromuscular transmitter and trophic factors. In: Samer M (ed.) *Advanced Medicine 18*, pp. 143–148. London: Pitman Medical, Royal College of Physicians.

Burnstock G (1986) Autonomic neuromuscular junctions: Current developments and future directions. *Journal of Anatomy* 146: 1–30.

Burnstock G (1986) The changing face of autonomic neurotransmission. *Acta Physiologica Scandinavica* 126: 67–91.

Burnstock G (1988) Autonomic neural control mechanisms: With special reference to the airways. In: Kaliner MA and Bames PJ (eds.) *The Airways: Neural Control in Health and Disease*, pp. 1–22. New York: Dekker.

Burnstock G (1975) *Adrenergic Neurones: Their Organisation, Function and Development in the Peripheral Nervous System*. London: Chapman and Hall, Elsevier.

Burnstock G (1990) Changes in expression of autonomic nerves in aging and disease. *Journal of the Autonomic Nervous System* 30: S25–S34.

Burnstock G (1992) *The Autonomic Nervous System, Vol. 1: Autonomic Neuroeffector Mechanisms*. Chur, Switzerland: Harwood.

Burnstock G (2004) Cotransmission. *Current Opinions in Pharmacology* 4: 47–52.

Burnstock G and Iwayama T (1971) Fine structural identification of autonomic nerves and their relation to smooth muscle. In: Eränkö O (ed.) *Progress in Brain Research, 34, Histochemistry of Nervous Transmission*, pp. 389–404. Amsterdam: Elsevier.

Burnstock G and Ralevic V (1994) New insights into the local regulation of blood flow by perivascular nerves and endothelium. *British Journal of Plastic Surgery* 47: 527–543.

Campbell GR, Uehara Y, Mark G, et al. (1971) Fine structure of smooth muscle cells grown in tissue culture. *Journal of Cell Biology* 49: 21–34.

Furness JB and Costa M (1987) *The Enteric Nervous System*. Edinburgh: Churchill Livingstone.

Hansen MA, Balcar VJ, Barden JA, et al. (1998) The distribution of single P2X1-receptor clusters on smooth muscle cells in relation to nerve varicosities in the rat urinary bladder. *Journal of Neurocytology* 27: 529–539.

Hillarp NA (1946) Functional organization of the peripheral autonomic innervation. *Acta Anatomica* 2: 1–153.

Hoyle CHV and Burnstock G (1996) Criteria for defining enteric neurotransmitters. In: Gaginella TS (ed.) *Handbook of Methods in Pharmacology*, pp. 123–140. London: CRC Press.

Luff SE (1996) Ultrastructure of sympathetic axons and their structural relationship with vascular smooth muscle. *Anatomy and Embryology (Berlin)* 193: 515–531.

Lundberg JM (1996) Pharmacology of cotransmission in the autonomic nervous system: Integrative aspects on amines, neuropeptides, adenosine triphosphate, amino acids and nitric oxide. *Pharmacological Reviews* 48: 113–178.

Merrillees NCR, Burnstock G, and Holman ME (1963) Correlation of fine structure and physiology of the innervation of smooth muscle in the guinea pig vas deferens. *Journal of Cell Biology* 19: 529–550.

Sandow SL, Whitehouse D, and Hill CE (1998) Specialised sympathetic neuroeffector associations in rat iris arterioles. *Journal of Anatomy* 192: 45–57.

Autonomic Neuroimmunology

H P M van der Kleij, P Forsythe, and J Bienenstock, McMaster University, Hamilton, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Structural Associations

Immune cells occur in close proximity to neural processes at sites where these structures are abundant. Anatomical nerve–immune associations have been reported involving mast cells as well as other bone marrow-derived cells, such as eosinophils and plasma cells in the lamina propria. Further work has demonstrated similar morphologic mast cell–nerve associations in the liver, mesentery, urinary bladder, and skin. In the skin, observations of Egan and co-workers showed that especially unmyelinated axons were associated with mast cells, as well as Langerhans cells (LCs) in primate and murine skin. Many if not most epidermal LCs appeared to be closely associated anatomically with calcitonin gene-related peptide (CGRP)-containing nerves, as determined by confocal laser scanning microscopy. In addition, there is morphological evidence for neural modulation of immune activity *in vivo*. It has been demonstrated clearly that the regions in which lymphocytes reside receive direct sympathetic neural input, and substance P may interact with thymocytes, mast cells, and other cells in the thymus, affecting their development and function.

Nerve–Immune Systems: Interaction

The nervous and immune systems can communicate in a bidirectional fashion as a result of a common set of signal molecules and their receptors. It has been established that the central nervous system (CNS) can regulate immune functioning. For instance, immune responses can be classically conditioned involving a CNS learning paradigm. Accordingly, information can flow not only from the CNS to the immune system but also in the opposite direction. Immune cells can produce neuropeptides such as substance P and β-endorphin. In addition, cells of the immune system possess receptors for a host of hormones and neurotransmitters, including receptors for catecholamines, acetylcholine (ACh), adrenocorticotropic hormone (ACTH), opioid peptides, substance P, CGRP, somatostatin (SOM), and vasoactive intestinal peptide (VIP). These receptors have been shown to respond *in vivo* and/or *in vitro* to the neurotransmitter substances, and their manipulation can alter immune responses. Furthermore, the CNS shares receptors with the immune system for neuropeptides and neurotransmitters as well as for cytokines.

Immune regulation is mediated via a cascade of events requiring cell–cell interactions and subsequent release of their contents, such as cytokines and antibodies, to regulate these responses. Cells of the immune system and their products have been shown to influence peripheral and central neurotransmission, leading to the conceptualization of a bidirectional neuroimmune communication system.

Innate immunity and adaptive (acquired) immunity are the two major immune defense responses of the body. Microbe-associated molecular patterns (MAMPs) derived from bacteria are recognized by pattern recognition receptors such as Toll-like receptors (TLRs). Detection of microorganisms through MAMPs by TLRs is crucial in triggering protective immunity. Studies have shown an essential role for TLRs in the activation of innate and adaptive immunity in animals. Ligand recognition induces a host defense response, which includes the production or regulation of inflammatory cytokines, upregulation of co-stimulatory molecules, and induction of antimicrobial defenses. Dendritic cells (DCs) are key players in the orchestration of immune events. Their function is crucial in both the early innate responses and the subsequent adaptive response. Activation of DCs by TLR ligands is necessary for their maturation and consequent ability to initiate immune responses.

Peripheral neuropeptides can directly attract immature DCs and at the same time they may capture mature DCs at sites of neurogenic inflammation. Low neuropeptide concentrations can improve the motility of DCs to the source of the neuropeptides, primarily sensory nerve fibers, where they may undergo functional and phenotypical maturation, probably due to changes in signal transduction pathways of neuropeptide receptors in immature and mature DCs. These findings provide evidence for a connection between adaptive immunity and the nervous system. The adaptive immune response can be altered by the nervous system releasing neuropeptides at specific local sites of infection or challenge, which may then influence the direction of the T helper cell (Th1/Th2) response. For instance, CGRP can inhibit interferon-γ (IFN-γ) production markedly in a dose-dependent manner; substance P and VIP can suppress interleukin-4 (IL-4) production. Catecholamines may play an important role in regulating adaptive immune responses as Th1 cells express adrenergic receptors. Furthermore, adrenergic receptors on B cells suggest that the sympathetic nervous system may also

regulate antibody production in addition to natural killer (NK) cell activity.

Recent reports have identified neural pathways that regulate the peripheral immune response. For instance, two immunomodulatory peptides, VIP and the pituitary adenylate cyclase-activating polypeptide (PACAP), are present in and released from both nerves and immune cells. VIP/PACAP have a general anti-inflammatory effect, both in innate and adaptive immunity. In innate immunity, VIP/PACAP inhibit the production of proinflammatory cytokines and chemokines from macrophages, microglia, and DCs. In adaptive immunity, VIP/PACAP promote Th2-type responses, and reduce the proinflammatory Th1-type responses.

Cytokine-induced illness (behavioral changes) can be seen as a neuroimmune response to activation of innate immunity. The nervous system can regulate immune function and inflammation. The CNS alerts the immune system to environmental changes using the shared neuropeptide, neurotransmitter, and cytokine receptors on immune cells. Neuropeptides, neurohormones, and neurotransmitters of the CNS interact with the immune system, which, in turn, feeds back to the brain, inducing behavioral changes and changes in the immune system. Cytokines produced in the periphery can affect the brain, producing illness and activation of neuroendocrine and autonomic stress-related circuits. Recognition of stimuli such as viruses and bacteria by the immune systems results in transmission of information to the CNS, causing physiological responses, such as fever associated with illness behavior. In the abdomen these responses are a direct result of immune cell-derived proinflammatory cytokines signaling the CNS via the subdiaphragmatic vagus nerve.

Neuropeptides

Neuropeptides are produced primarily in the brain and nervous system, although every tissue in the body can produce and exchange neuropeptides. The immunological activity of neuropeptides is mostly mediated through specific receptors. The presence of specific receptors for substance P, SOM, CGRP, corticotropin-releasing hormone (CRH), melanocortin peptides, VIP, and related peptides has been reported on immune cells. For example, macrophages, DCs, T lymphocytes, and B lymphocytes all have the ability to express neurokinin type 1 (NK-1) receptors. The existence of these specific neuropeptide receptors on immune cells represents the framework for neuropeptides functioning as mediators of neuroimmune interactions.

Substance P, the most important neuropeptide of the tachykinin family, is widely distributed in the central, peripheral, and enteric nervous systems of many species. Substance P functions in the CNS as a neurotransmitter and its NK-1 receptors are localized in distinct areas of the brain important in affecting behavior and the neurochemical response to both psychological and somatic stress. It may also coordinate the response to stress by interacting with the hypothalamic–pituitary–adrenal (HPA) axis and the sympathetic nervous system. Neuronal substance P is stored in vesicles and released from sensory nerves in response to various stimuli, such as leukotrienes, prostaglandins, and histamine. Release of substance P is often accompanied by release of CGRP, since they are often co-localized in autonomic nerves.

The tachykinin family has recently been extended by the discovery of a third tachykinin gene preprotachykinin-C mRNA, encoding a tachykinin called hemokinin-1. An important observation that has been made is the fact that constitutive expression of hemokinin mRNA occurs preferentially in leukocytes and lymphoid tissue, not in the nervous system. Microglial cells, described as resident macrophages in the brain, are a potential source of hemokinin-1. There is the possibility that hemokinin-1 could play a significant role in communication between immune cells in the CNS, and between immune cells and neuronal cells expressing the NK-1 receptor.

Neuropeptides can influence the innate immune response, especially since NK-1 receptors have been found on macrophages and DCs. *In vivo* studies have demonstrated a role for the NK-1 receptor in the initiation of the host response against bacterial and viral pathogens. Upregulation of NK-1 receptor expression by *Salmonella* may significantly increase the substance P-mediated macrophage response, directly killing the bacteria. In addition, substance P can enhance IL-1, IL-6, and tumor necrosis factor (TNF) secretion by macrophages.

CGRP can also be found by itself in some motor and enteric neurons. CGRP inhibits T cell proliferation by inhibiting IL-2 production. In macrophages, it inhibits antigen presentation and other macrophage functions, including phagocytosis. It also stimulates production of a number of cytokines, such as IL-6, IL-10, and TNF. *In vivo* injection of CGRP was shown to inhibit the onset of delayed-type hypersensitivity reactions, supporting the role of CGRP in systemic regulation of immune activation. Endogenous CGRP may generally modulate immune function, presumably, in part, through an inhibitory effect on LC function in the skin. CGRP is capable of inhibiting both the induction and expression of

fundamental cellular immune functions. This suggests an interaction between the nervous system and immunological function. These observations imply the likelihood that the CNS may be capable of suppressing the initiation of immune responses, and is also effective in their modulation. Overall, there is substantial experimental evidence demonstrating that CGRP can influence the function and development of inflammatory and immune cells in local microenvironments by specific receptor-mediated mechanisms.

Although classically neuropeptides are released from autonomic and sensory nerves, there is increasing evidence that they may also be synthesized and released from immune cells such as macrophages and eosinophils, particularly in disease. The regulation of synthesis and secretion of these mediators is largely undetermined. Therefore any description of homeostasis, especially under conditions in which immune stimulation or inflammation is present, must by necessity remain incomplete until this knowledge is available.

Neurotrophins

Based on their expression profile, neurotrophins are good candidates for mediating immune–nerve cell interactions. Nerve growth factor (NGF) is the best characterized neurotrophic factor, but the term 'neurotrophin' includes brain-derived neurotrophic factor (BDNF), glial-derived nerve growth factor (GDNF), and neurotrophins 3/5. Neurotrophins influence the proliferation, differentiation, survival, and death of neuronal and nonneuronal cells. However, several nonneuronal functions have also been characterized. For instance, neurotrophic factors are essential for enteric nervous development. Especially GDNF has been identified to control the migration, morphogenesis, and differentiation from precursors to the enteric nervous system. Furthermore, neurotrophins can regulate neuropeptide expression, interact with immunoregulatory cells and epithelial cells, and control motility during inflammation. Neurotrophins also act as an important intermediate in inflammatory pain, contributing to both peripheral and central sensitization.

Neurotrophin receptors are widely expressed in the peripheral and the CNSs, as well as on cells of the immune system. Many immune cells express the high-affinity NGF tyrosine receptor kinase A (TrkA) receptor. This allows NGF to promote the release of inflammatory mediators. Several of these mediators, such as IL-1, IL-4, IL-5, TNF, and IFN, can, in turn, induce the release of NGF. Therefore, NGF seems to be a mediator with functions on both immune cells and nerve cells and is likely an important factor in integrating communication between the nervous and immune systems.

There is increasing evidence that NGF acts on cells of the immune system. NGF promotes differentiation, activation, and cytokine production of mast cells and macrophages, and activates eosinophils. It promotes survival of different cell types, such as mast cells, by their prevention of apoptosis. NGF is a chemoattractant, thereby causing an increase in the number of mast cells as well their degranulation. Furthermore, NGF influences activity of basophils, eosinophils, neutrophils, macrophages, and T cells and is also produced by a wide range of immune cells. Mast cells, T cells, B cells, eosinophils, lymphocytes, and epithelial cells can synthesize NGF.

NGF can have proinflammatory as well as anti-inflammatory effects, dependent on its location and concentration. In mice, nasal treatment with NGF induces airway hyper-responsiveness, measured by electrical field stimulation. NGF-transgenic mice, overexpressing NGF in Clara cells, develop hyper-reactivity in the absence of airway inflammation, suggesting that NGF by itself determines the induction of airway hyper-responsiveness. On the other hand, there is evidence indicating that the increased production of NGF in the CNS during brain diseases, such as multiple sclerosis, can suppress inflammation by switching the immune response to an anti-inflammatory, suppressive mechanism. The fact that NGF may under different circumstances be either proinflammatory or anti-inflammatory is an important example of the complexity of interactions between neuroactive molecules and the immune and inflammatory systems.

Mast Cells–Nerve Interaction

Histological studies reveal an intimate association between mast cells and neurons in both the peripheral and CNSs. Besides the anatomical link, mast cells also form a functional link between the immune and nervous systems. Neuronal mechanisms are involved in mast cell activation and mast cells act as principal transducers of information between peripheral nerves and local inflammatory events.

Direct bidirectional communication between mast cells and superior cervical neurites has been supported, in an *in vitro* co-culture model of murine sympathetic neurons cultured with rat basophilic leukemia (RBL-2H3) cells. Neurites show neurotrophic and neurotropic activity for RBL cells. Intimate synapse-like contact occurs followed by cessation of division, membrane electrophysiologic alterations, maturation

of granules in RBL cells, and accumulation of vesicles in neurons close to the point of contact. Activation of nerves with scorpion venom elicited degranulation of RBL cells via substance P, and substance P activation is initiated only at their point of contact. Co-cultures of bone marrow-derived mast cells (BMMCs) and neurites have demonstrated that expression of NK-1 receptors by mast cells lowers the threshold of activation induced by nerve stimulation.

Mast cells and nerves in many tissues are in constant contact with each other and share a number of activating signals, for some of which both cells express receptors (such as vanilloids). Additionally, both mast cells and nerves respond to stimulation by secretion of preformed mediators, many of which are produced by both cells (NGF, neuropeptides, and endothelin-1). Moreover, a large proportion of primary spinal afferent neurons, which contain CGRP and substance P, express the proteinase-activated receptor 2 (PAR-2). Proteases, such as tryptase from degranulated mast cells, have recently been shown to cleave PAR-2 on primary spinal afferent neurons, causing release of substance P, activation of the NK-1 receptor, and amplification of inflammation and thermal and mechanical hyperalgesia. This mechanism of protease-induced neurogenic inflammation may contribute to the proinflammatory effects of mast cells in human disease.

Many studies have shown that mast cell-derived mediators, such as histamine, serotonin, and cytokines, modulate nonadrenergic noncholinergic (NANC) neurotransmission. NANC nerve endings express receptors for histamine (H1 and H3) and serotonin (5-HT_{2A}), and histamine H1 receptor expression at least is upregulated on primary NANC nerves in inflammation. Mast cell mediators such as TNF can sensitize afferent C fibers by lowering their threshold, but can also cause direct release of substance P, neurokinin A, and CGRP from unmyelinated fibers.

In addition to IgE and antigen, cytokines, hormones, and neuropeptides can trigger mast cell secretion. The latter include SOM, neurotensin, PACAP, CGRP, and substance P. CRH, stem cell factor (SCF), and NGF can promote mast cell growth and can also trigger mast cell degranulation. SCF has also been reported to induce mast cells to become responsive to PACAP. Keratinocytes, LCs, fibroblasts, mast cells, and endothelial cells express functional neurokinin receptors, while G-proteins on mast cells can, in addition, be directly activated by substance P. Functional NK-1 receptors are expressed on murine mast cells in the presence of IL-4 and SCF, but are not constitutively expressed by human intestinal mast cells; this changes upon stimulation by IgE receptor cross-linking. This suggests that, *in vivo*, specific tissue conditions such as allergic inflammation may lead to mast cell expression of NK-1 receptors.

Nerve–mast cell interactions are engaged in both homeostatic and pathologic regulation. Synaptic cell adhesion molecule (SynCAM) is involved in most CNS synaptic associations. Furthermore, SynCAM was reported to mediate the attachment of mast cells to neurites by expression in both cell types. This homophilic binding not only results in this immune–nerve association, but also facilitates successful neurotransmission from neurite to mast cell.

Neurogenic Inflammation

Neurogenic inflammation involves a change in function of sensory neurons due to inflammatory mediators, inducing an enhanced release of neuropeptides from the sensory nerve endings. The effects produced by tachykinins and CGRP released from peripheral endings of capsaicin-sensitive primary sensory neurons are collectively referred to as neurogenic inflammation. There is also histological evidence for the presence of cytokines, especially IL-6, within both sensory and autonomic nerves. The release of these cytokines would cause local inflammation and indicates a direct link between neurogenic stimulation and the release of proinflammatory cytokines. Experimental research has shown a role for neurogenic inflammation in disease processes such as arthritis, colitis, bladder inflammation, and asthma.

The postganglionic sympathetic neuron may also play a role in neurogenic inflammation. In the presence of subdiaphragmatic vagotomy, the potency of nociceptor afferents to inhibit the sympathetically dependent bradykinin-induced plasma extravasation was significantly increased. These findings suggest that there is a vagal mechanism affecting neurogenic inflammation. Afferent signals to the brain can generate a reflex, causing an anti-inflammatory response, which is partly mediated by the efferent branch of the vagus nerve. This cholinergic anti-inflammatory pathway is mediated primarily by nicotinic ACh receptors expressed on tissue macrophages and blood monocytes. ACh can interact specifically with macrophage $\alpha 7$ subunits of nicotinic ACh receptors, leading to cellular deactivation and inhibition of cytokine release.

CNS and Stress

Several pathways have been shown to link the brain and the immune system: the autonomic nervous system, via direct neural influences, and the neuroendocrine humoral outflow, via the pituitary. The CNS can and does activate or even inhibit immune and

inflammatory events. Stress activation of the HPA axis has such a bimodal effect. CRH activates peripheral mast cells. Corticosterone may inhibit inflammatory events and catecholamines have complex effects. Hypnosis has been shown to inhibit the expression of skin reactivity to allergen in classical tests of allergic hypersensitivity, and it has been shown that classical Pavlovian conditioning could cause mast cell secretion.

In the CNS, mast cells may participate in the regulation of inflammatory responses through interactions with the HPA axis. In the dog, degranulation of CNS mast cells evokes HPA activation in response to histamine release and CRH. Control or peripheral administration of specific antigen results in corticosterone release from the adrenal glands. CRH is also thought to be involved peripherally in tissue responses to stress in the skin, respiratory tract, intestine, and bladder. Many, if not all, of the recorded changes have involved mast cells and neuronal activation, the latter being often mediated by neurotensin and/or substance P.

Central neuropeptides initiate a systemic stress response by activation of neuroendocrinological pathways such as the sympathetic nervous system, the hypothalamic–pituitary axis, and the renin–angiotensin system, with the release of the stress hormones (i.e., catecholamines, corticosteroids, growth hormone, glucagons, and renin). Immobilization stress increases colonic motility and many other physiologic parameters through release of CRH. Both mast cells and nerves are involved, and the effects have been shown to occur through neurotensin and substance P, mediated by cholinergic, adrenergic, and ganglionic pathways. These results provide support for an important link between the nervous and the immune systems in stress-mediated intestinal responses.

Active immune responses represent a source of systemic stress, which impacts the brain and modifies various neuroendocrine and behavioral functions. Therefore, the immune system has been conceived of as a potential contributor to stress-related behavioral abnormalities, such as clinical depression. It is suggested that immunologically induced changes in the brain activate neuropeptides, thereby sustaining an adaptive state of arousal that promotes appropriate behavioral adjustments during infectious illness.

Immune cells such as macrophages are involved in certain types of stress. The cold-water swim test results in the increase of peritoneal substance P and its receptor in peritoneal macrophages. Immune–nerve interaction has been shown to be involved in the skin during stress. Stress can inhibit hair growth *in vivo* and significantly increases the number of hair follicles containing apoptotic cells, activated perifollicular macrophage clusters, and degranulated mast cells, whereas it downregulates the number of intraepithelial lymphocytes. Substance P seems to be a key mediator of stress-induced hair growth inhibition *in vivo*. Increased numbers of substance P-immunoreactive sensory fibers, seen in the dermis of stressed mice, are a result of transient high levels of NGF, suggesting that NGF is a central element in the perifollicular neurogenic inflammation that develops during the murine skin response to stress.

Several inflammatory skin conditions, including atopic dermatitis and psoriasis, are exacerbated by stress. Recent evidence suggests that cross-talk between mast cells, neurons, and keratinocytes might be involved in such exacerbation. CRH and its receptor are present in the skin and their levels are increased following stress. Human mast cells synthesize and secrete CRH in response to IgE receptor cross-linking. Mast cells also express CRH receptors, activation of which leads to the selective release of cytokines and other proinflammatory mediators. CRH receptor antagonists could be used to inhibit stress-induced mast cell activation and provide new therapeutic options for chronic inflammatory conditions exacerbated by stress.

Summary

The brain and nervous system have a bidirectional communication with the immune system. There is a consistent finding of associations between nonmyelinated nerves and lymphocytes, mast cells, eosinophils, and LCs, at which locations bidirectional interaction clearly occurs. Furthermore, it has become apparent that many cells in these different systems possess the same receptors and can make the same molecules. Behavioral changes can be initiated as a result of peripheral inflammation (illness behavior) and fever can be caused by peripheral infections and inflammation as a result of communication with and through the nervous system. The brain routinely receives communication signals from peripheral immune events and can coordinate responses through the HPA axis and the sympathetic, parasympathetic, and NANC nervous systems. Hypoactivity or hyperactivity of these mechanisms could contribute to diseases characterized by either a decreased or an exaggerated inflammatory response. Much remains to be done, but scientific progress promises much greater complexity as well as understanding of these processes. What is clear, however, is that immune mechanisms cannot be ignored in dealing with a functional understanding of the nervous system or even behavior, and the reverse is also true.

See also: Immune System–Neuroendocrine Interactions; Inflammation in Neurodegenerative Disease and Injury;

Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neuroimmune System: Aging; Neuropeptides in Autonomic Neurons; Neurotrophins: Physiology and Pharmacology; Stress and Parasympathetic Control; Stress, the HPA Axis and Depressive Illness.

Further Reading

Bienenstoc J, Goetzl EJ, and Blennerhassett MG (2003) Autonomic neuroimmunology, In: Burnstock G (ed.) *The Autonomic Nervous System*, vol. 15, pp. 1–392. London: Taylor & Francis.

Black PH (1994) Central nervous system–immune system interactions: Psychoneuroendocrinology of stress and its immune consequences. *Antimicrobial Agents and Chemotherapy* 38: 1–6.

Blalock JE (1984) The immune system as a sensory organ. *Journal of Immunology* 132: 1067–1070.

Chrousos GP and Gold PW (1992) The concepts of stress and stress system disorders. Overview of physical and behavioral homeostasis. *JAMA: The Journal of the American Medical Association* 267: 1244–1252.

Dantzer R (2001) Cytokine-induced sickness behavior: Mechanisms and implications. *Annals of the New York Academy of Sciences* 933: 222–234.

Felten DL, Felten SY, Bellinger DL, et al. (1987) Noradrenergic sympathetic neural interactions with the immune system: Structure and function. *Immunological Reviews* 100: 225–260.

Maier SF and Watkins LR (1998) Cytokines for psychologists: Implications of bidirectional immune-to-brain communication for understanding behavior, mood, and cognition. *Psychological Review* 105: 83–107.

Marshall JS and Waserman S (1995) Mast cells and the nerves – Potential interactions in the context of chronic disease. *Clinical & Experimental Allergy* 25: 102–110.

McDonald DM, Bowden JJ, Baluk P, et al. (1996) Neurogenic inflammation. A model for studying efferent actions of sensory nerves. *Advances in Experimental Medicine and Biology* 410: 453–462.

Purcell WM and Atterwill CK (1995) Mast cells in neuroimmune function: Neurotoxicological and neuropharmacological perspectives. *Neurochemical Research* 20: 521–532.

Steinhoff M, Stander S, Seeliger S, et al. (2003) Modern aspects of cutaneous neurogenic inflammation. *Archives of Dermatology* 139: 1479–1488.

Tracey KJ (2002) The inflammatory reflex. *Nature* 420: 853–859.

van der Kleij HP, Blennerhassett MG, and Bienenstock J (2003) Nerve–mast cell interactions – Partnership in health and disease. In: Bienenstock J, Goetzl EJ, and Blennerhassett MG (eds.) *Autonomic Neuroimmunology*, pp. 139–170. London: Taylor & Francis.

Watkins LR and Maier SF (1999) Implications of immune-to-brain communication for sickness and pain. *Proceedings of the National Academy of Science of the United States of America* 96: 7710–7713.

Williams RM, Bienenstock J, and Stead RH (1995) Mast cells: The neuroimmune connection. *Chemical Immunology* 61: 208–235.

Autonomic Neuroplasticity and Aging

T Cowen, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Autonomic neurons, in common with neurons from all parts of the nervous system, undergo extensive and continual adaptation in response to intrinsic and extrinsic stimuli. The nature of neural plasticity and the factors that influence it vary throughout life. Broadly, these changes fall into three stages: development, maturation and adulthood, and old age. During development, neurons undergo differentiation toward their adult phenotype. This process includes acquiring appropriate morphology, expression of neurotransmitters and receptors, and patterns of activity and signaling pathways adapted to particular functions and connectivity. Toward the end of this period, usually during the perinatal period in mammals, neurons that have not made appropriate connections with their target tissues (which may be other neurons or effector organs such as muscles and glands) die by programmed cell death. During maturation and adulthood, plasticity takes different forms. The broad outline of the phenotype acquired during development – morphology, neurotransmitter, and other functional characteristics – remains stable. However, neurons retain responsiveness to altered demand. During adulthood and aging, many autonomic neurons retain a remarkable capacity to adapt to changing circumstance. However, a minority of cells become subject to selective vulnerability, undergo atrophy, and may die, probably contributing to the functional losses of old age which include deficits in autonomic control of homeostasis. The molecular mechanisms that underlie autonomic neural plasticity are beginning to be understood. A recurring theme throughout the life span is the interaction between neurons and their target tissues, which has profound effects on many aspects of neural plasticity.

Plasticity in the Developing Autonomic Nervous System

Neuronal Growth

During embryonic development, differentiating neurons of the autonomic nervous system (ANS) migrate to appropriate peripheral locations, such as the pre- and paravertebral autonomic ganglia and the enteric nerve plexuses of the gut. From these locations, the neurons extend dendrites to form synapses with spinal autonomic neurons of the intermediolateral column. They also extend axons toward their peripheral targets which may be smooth muscle effectors in organs such as blood vessels, gut, or iris, or glands such as salivary, pineal, or sweat glands, or glands of the gastrointestinal tract. Guidance cues for migratory neurons and growing neurites are provided by bound molecules of the extracellular matrix (ECM) as well as by diffusible factors. In recent years, it has become increasingly clear that guidance, like other epigenetic processes in the nervous system, is regulated by a combination of positive and negative influences. In the case of neural guidance, bound and diffusible factors may act as chemoattractants or chemorepellents. Chemoattractants include the ECM glycoprotein laminin and the diffusible neurotrophic and other growth factors, while chemorepellents include members of the semaphorin family, netrins, and ECM glycoproteins. Some but not all of these molecules have been investigated in relation to autonomic neurons. The combination of attractive and repulsive signals ensures that growing neurites follow appropriate pathways, including established nerve fascicles, grow toward appropriate targets, and avoid inappropriate ones.

Around the time that neurons approach their target areas, the peripheral tissues start to synthesize a locally specific cocktail of factors, which attract the growing axons and stimulate synaptogenesis. Bound factors such as laminin stimulate growth of autonomic neurites, for example, through contact between receptors on axon terminals and laminin located in glial or smooth muscle basal laminae. Laminin binds to integrin receptors of the tyrosine kinase family, which initiate neuronal growth and survival signals through ERK/MAP-kinases and PI3-kinase. Diffusible growth-promoting factors produced by targets include the neurotrophin family of growth factors, nerve growth factor (NGF), and neurotrophin-3 (NT3) having the most clearly defined functions in the ANS. These molecules generate growth responses in sympathetic and other peripheral neurons by activating the Trk family of tyrosine kinase receptors, of which TrkA and TrkC are the receptors principally responsible for mediating the effects of NGF and NT3, respectively. The contribution of the nonspecific neurotrophin receptor, p75, remains controversial. Some suggest a synergistic role with Trk receptors in promoting growth and survival; others argue for an inhibitory role for p75 in neuronal growth and survival, probably acting in situations where Trk's are absent. A characteristic of these signaling systems is that they often act in combination, with bound and diffusible factors either working synergistically or providing a balance of positive and negative influences.

Locally Specific Phenotype

Target specificity of innervation is achieved by secretion by target tissues of locally specific combinations of factors, which stimulate particular patterns of neuronal growth as well as neuronal synthesis of particular neuropeptides. Growth of large, complex sympathetic neurons (supplying, e.g., the iris where fast physiological reflexes are required) correlates with high levels of synthesis of NGF. In contrast, smaller sympathetic neurons supplying sparse innervation to blood vessels (where much slower, less specific neuromuscular responses are required) are associated with low levels of NGF synthesis in the vessel wall (Figure 1). Target-derived factors also ensure the adoption by the neurons of a suitable neurotransmitter phenotype The mechanism underlying establishment of neurotransmitter phenotype has been investigated widely, particularly in the minority of sympathetic nerves of the ANS which utilize acetylcholine as a neurotransmitter. Target-derived molecules are clearly involved in this process; however, their identity remains elusive and may include the cytokines ciliary neurotrophic factor, leukemia inhibitory factor, or cardiotrophin-1 and/or the neurotrophic factor, glial cell-derived neurotrophic factor (GDNF).

Neuronal Survival

The nervous system at large generates more neurons during embryogenesis than are required in maturity.

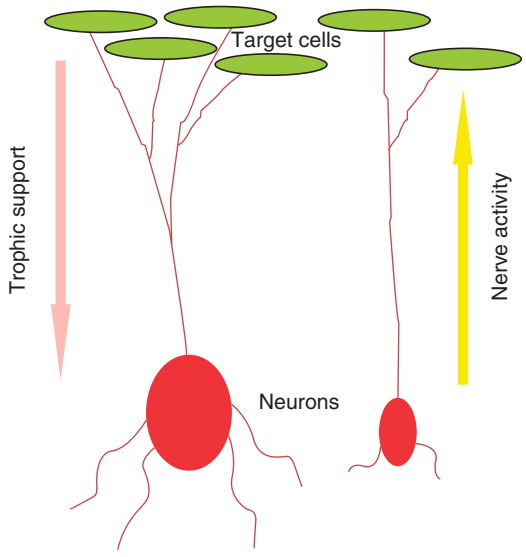

Figure 1 The neurotrophic hypothesis. Target tissues support neuronal growth, survival, and phenotype through their production of neurotrophic and other factors which are taken up by neurons and retrogradely transported to the cell body where they initiate gene transcription. Neurons regulate the supply of these factors through their activity. Targets producing high levels of neurotrophic factors become innervated by large, complex neurons; conversely, targets producing lower levels of these factors become innervated by smaller ones. These different phenotypes are likely to be functionally adaptive. However, they appear to have pleiotropic effects on the survival trajectories of these neurons in old age (see text).

Fine-tuning of the numbers of surviving neurons to the physical size and physiological demands of peripheral organs occurs by ensuring that the cocktail of diffusible growth factors produced by the target tissue provides sufficient but not superfluous support for the innervating neurons. Neurons that reach their target compete for this scarce resource by growing axons bearing receptors which uptake the appropriate factors, generating further growth and survival signals. Those that fail to take up a sufficient supply undergo programmed cell death or apoptosis. This process was extensively investigated in sympathetic neurons of the ANS and only later understood to be important in many areas of the nervous system. This model of developmental regulation of neuron survival has become encapsulated in the 'neurotrophic theory' (Figure 1). An interesting recent addition to this concept concerns the emerging evidence that physiological activity of the target, which is at least partly a result of the activity of the innervating neurons, can itself stimulate synthesis or secretion of growth factors. This important observation suggests that neurons have the capacity to regulate their target-derived trophic support.

Similar groups of molecules contribute to the survival as well as to the growth of ANS neurons. Thus, target-derived NGF and NT3 signaling through Trk receptors regulate survival of numbers of sympathetic neurons appropriate to the particular target tissues during development while the survival of gut neurons is regulated partly by laminin, and subsequently by GDNF and NT3. Pruning of neuron numbers takes place at different periods of perinatal life depending on the target area. In general, it starts in the few days before birth and continues for a variable number of weeks after birth. The signaling pathways mediating developmental survival are similar to those driving neurite growth, that is, tyrosine kinase signaling through ERK/MAP-kinase and PI3-kinase. It seems most likely that downregulation of these pathways (i.e., by subthreshold levels of the ligand neurotrophins), particularly of PI3-kinase, is the key signal for neuronal death. The death signal involves a pro-apoptotic shift in the downstream effectors of tyrosine kinase signaling, which include expression of the BCL-2 and caspase families of proteins. Knockout studies in mice indicate that the p75 receptor, signaling through NFκB, is also involved in the developmental regulation of cell death in NGF-dependent autonomic and other neurons.

Plasticity in the Adult and Aging ANS

Neuronal Growth

Adult neurons retain the capacity to respond sensitively to extrinsic growth signals. Body organs continue to

grow beyond the period of naturally occurring postnatal neuronal cell death. Existing and relatively stable populations of neurons are therefore required to expand their axonal and dendritic connections by continued growth and synaptogenesis. *In vitro* studies have shown that adult neurons are at least as sensitive as early postnatal neurons in terms of their growth responsiveness to neurotrophins. Indeed, even those neurons that survive into old age retain growth responsiveness to neurotrophins; strong nerve sprouting has been observed *in vivo* in response to treatment with neurotrophic factors. However, there is an age-related reduction in the dose-response curve *in vivo* and *in vitro*.

That plasticity is a physiological feature of adult autonomic neurons is supported by several lines of evidence. Pregnancy and the estrous cycle are accompanied by substantial cyclical changes in the autonomic neurons innervating the uterus. Sympathetic nerve fibers undergo atrophy in late estrus and pregnancy, followed by regeneration. Pregnancy-induced neuronal atrophy may protect the fetus against vasoconstriction caused by homeostatic responses such as fight and flight. Multiple pregnancies may result in the death of some of the innervating sympathetic neurons. Transplantation studies and experimental enlargement of organs demonstrate further that neurons, particularly autonomic neurons, retain an extraordinary ability to expand their field of innervation. Partial obstruction of the bladder can result in a roughly tenfold increase in muscle mass to which the innervating autonomic neurons respond with an increase in size of around 80%. Removal of the obstruction results in reversion of neuron size to control values over a few weeks. Comparable manipulations in the gut and iris – both autonomically innervated tissues – have yielded similar results.

The same neurotrophic factors, which regulate growth of neurons during development, are retained by the adult target tissues, generally at unchanged levels. In the case of adult sympathetic neurons, these factors include NGF and NT3. Of these, NGF produces the substantially stronger growth response, although both factors are present in many of the target tissues of sympathetic neurons. TrkA and TrkC receptors are present on adult sympathetic neurons to mediate the growth responses. The increased target area resulting from the physiological or experimental expansion of organs described above is therefore likely to induce nerve growth by increasing the availability of neurotrophins in the vicinity of existing nerve terminals. The role of the nonspecific p75 receptor remains as controversial in the adult nervous system as it is at earlier stages of development. Some argue that p75 inhibits neuronal growth in adult rodent neurons while others consider that p75 supports the trophic role of NGF by heterodimerization with Trk. These apparently contradictory roles for this receptor may result from different expression ratios of p75 and Trk in different neurons and under different circumstances. Local availability of neurotrophins can affect the expression of both p75 and Trk receptors in the adult nervous system, thus providing a further level at which nerve growth can be regulated.

The signaling pathways involved in adult neuron growth are less well defined than those regulating survival. While inhibition of tyrosine kinase signaling largely inhibits extension of neurites from adult sympathetic neurons *in vitro*, there is less conclusive definition of the downstream pathways involved. This may be because neurite extension involves integrin-as well as neurotrophin-mediated activation of tyrosine kinase signaling, with resulting convergent activation of ERK/MAP-kinase and PI3-kinase.

During aging, some autonomic neurons undergo atrophy (shrinkage of their dendritic and axonal arbors and cell soma), while others continue to grow into old age. This phenomenon has been called 'selective vulnerability' and affects sympathetic neurons projecting to vascular targets and sweat and pineal glands. In contrast, neurons projecting to iris and submandibular gland show no evidence of atrophy. To the extent that the subpopulations of neurons innervating each target tissue exhibit specific patterns of growth or atrophy during aging, selective vulnerability involves interactions between particular target tissues and neurons. However, it has not proved possible to demonstrate that reduced levels of NGF availability from targets causes age-related atrophy of fibers.

Locally Specific Phenotype

Adult neurons retain locally specific morphology and neurochemical phenotype. Thus, subpopulations of adult sympathetic neurons, retrogradely traced from particular target tissues, exhibit different sizes which appear to relate to physiological function in the same way that developing neurons adapt to the physiological function of the end organ. In addition, adult neurons continue to express locally specific patterns of neuropeptides and transmitters. For example, neuropeptide Y expression is highest in superior cervical ganglion (SCG) neurons projecting to iris, lower in those projecting to the middle cerebral artery, and lowest in neurons projecting to the submandibular gland. There is evidence that these patterns of neurotransmitter expression remain plastic in adulthood. Transplantation in adult rats of sweat glands (normally innervated by cholinergic, vasoactive intestinal peptide (VIP)-expressing sympathetic neurons) into contact with catecholaminergic sympathetic neurons induces a shift to a cholinergic, VIP-expressing phenotype in those neurons, which extend processes to innervate the

transplanted tissue. There is no obvious loss of this form of plasticity in old age.

Neuronal Survival in Adulthood

While neurons retain growth responsiveness to extrinsic factors in adulthood and even in old age, the factors that determine survival, at least of sympathetic and other (sensory) peripheral neurons, appear to change dramatically during maturation and in aging. Between birth and 12 weeks of age, rodent sympathetic neurons, from being dependent for their survival on an extrinsic source of neurotrophin, become almost completely independent and able to survive in dissociated cell culture for many days and even weeks under serum- and NGF-free conditions (Figure 2). The loss of survival dependence on target-derived signals may be seen as an adaptation allowing adult neurons a degree of resistance against an axonal injury that separates the neuron from its target, thus allowing the neuron a period of respite in which to initiate regeneration.

The principal signaling pathway that mediates neuronal survival independence, as this phenomenon is called, is PI3-kinase. Pharmacological and/or genetic inhibition of this pathway at different points leads to the rapid death of cultured adult sympathetic neurons; inhibition of ERK/MAP-kinase is without significant effect. This suggests strongly that while neurons lose their dependence on signaling through Trk receptors to regulate their survival, the downstream PI3-kinase pathway remains a key determinant of neuronal resistance to cell death. It is feasible that autonomic and perhaps other neurons protect their survival in adulthood by cell-autonomous activation of PI3-kinase in a way that resembles a mechanism by which cancer cells avoid apoptosis.

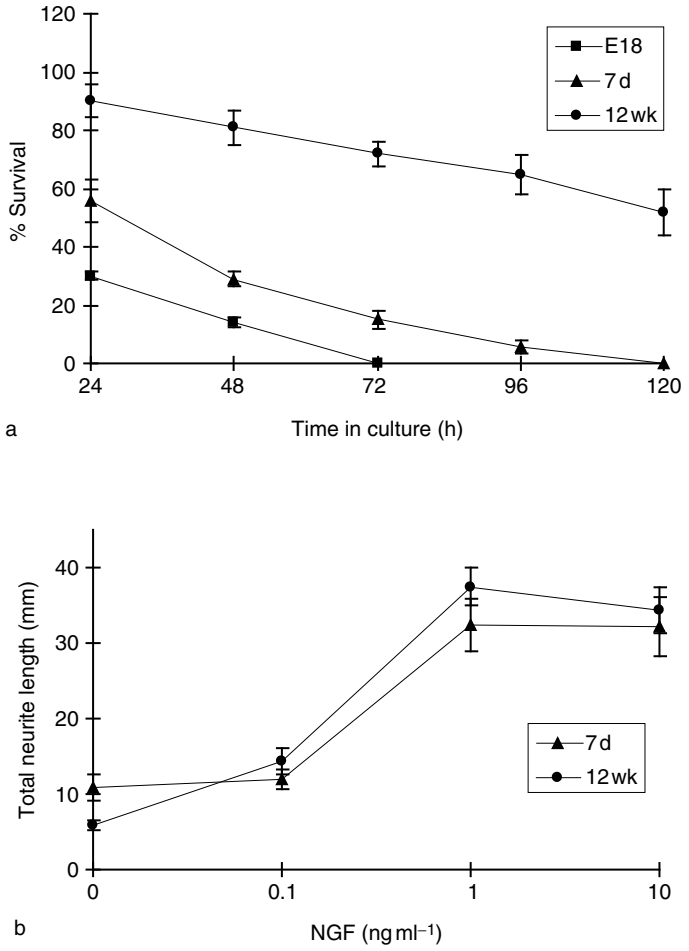

Figure 2 Differential regulation of neuronal growth and survival during maturation. During postnatal maturation, sympathetic neurons lose their survival dependence on NGF. Growth, however, remains NGF dependent. (a) Sympathetic neurons become independent of NGF for their survival between embryonic day 18 (E18) and 12 weeks of age, as demonstrated by a dramatic increase in survival with age in dissociated, pure neuron cell cultures maintained over several days in the absence of serum and NGF. Note that by 12 weeks, approximately 60% of neurons survive for longer than 5 days *in vitro*, whereas no E18 neurons survive longer than 3 days. (b) Postnatal sympathetic neurons require NGF to initiate growth and to extend neurites *in vitro*. Note that neurons fail to grow in the absence of NGF; dose responsiveness is similar in 12-week-old and 7-day-old neurons.

It has recently become clear that production of neuronal progenitors, at least in the central nervous system (CNS), continues throughout life under the influence of a range of epigenetic factors, including physical exercise, brain activity, and injury. It is not clear, however, whether the ANS can continue to generate neuronal precursors in adulthood or whether plasticity mechanisms similar to those seen in development and maturation regulate the differentiation, survival, and integration of progenitor cells into the functioning adult nervous system.

Neuronal Survival in Old Age

Selective vulnerability to neuron cell death is a well-established characteristic of the aging nervous system as a whole, as well as of the ANS. The groups of neurons affected are similar to those which exhibit age-related atrophy (see the earlier section on neuronal growth). Among sympathetic neurons, only certain populations undergo cell death, while many survive into extreme old age. Roughly 50% of vascular-projecting neurons in the SCG are lost in aging rats, while neurons projecting to the iris are less significantly reduced in numbers. Autonomic neurons from the myenteric plexus of the gut are also vulnerable to age-related cell death, with roughly 50% of neurons lost from for small and large intestine of aging rats and guinea pigs (Figure 3). There is limited but less secure evidence that similar patterns of myenteric neuron loss occur in humans, at least from the large intestine.

In line with evidence that adult neurons become relatively independent of target-derived trophic support, it is probable that locally selective patterns of neuron death during aging are not the result of

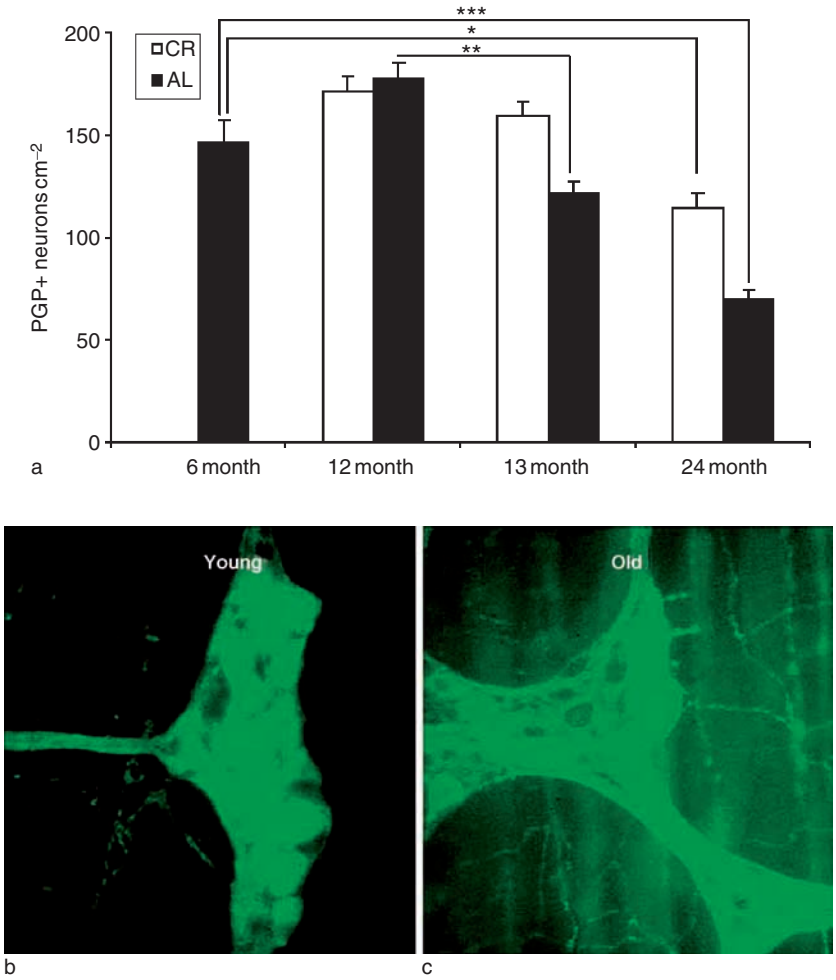

Figure 3 Age- and diet-related loss of myenteric neurons from the rat small intestine. (a) Counts of myenteric neurons using the pan-neuronal marker, PGP 9.5. The histogram shows significant losses of total numbers of neurons, commencing at 12–13 months of age in *ad libitum*-fed (AL) animals, which reaches a peak of 51% at 24 months. There is a much smaller but statistically significant ($p < 0.05$) reduction seen in animals of the same age (24 months) but fed a calorie-restricted (CR) diet, demonstrating substantial protection by CR against age- and diet-related loss of myenteric neurons. Statistically significant differences are indicated by $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. (b, c) Photomicrographs of young (6 months; b) and old (24 months AL; c) PGP-immunostained whole-mount preparations of rat myenteric plexus. Note the large spaces in the old ganglia indicating where neuron cell loss has occurred.

reduced availability of neurotrophins. Studies of different target areas show no evidence of altered levels of NGF or NT3 protein or mRNA in targets of aging sympathetic neurons, mirroring similar observations in targets of vulnerable NGF-dependent neurons in the CNS. The search for causes of age-related neurodegeneration is a high priority in aging research. A number of avenues are being explored including whether expression of the p75 or Trk receptors are selectively altered in subgroups of aging sympathetic neurons vulnerable to age-related cell loss. In the absence of changes in levels of extrinsic factors, these are likely to be cell-autonomously regulated. Subpopulations of neurons that exhibit age-related cell loss project to target areas that throughout life express relatively low levels of neurotrophic factors; conversely neurons that survive successfully into old age tend to project to targets expressing high levels of neurotrophic factors (see the above discussion on locally specific phenotype). These differential levels of neurotrophin synthesis are retained throughout life and are reflected in altered patterns of p75 and Trk receptor expression, which correlate directly with neurotrophin availability. It therefore seems likely that growth factor signaling is set during maturation at levels that are adapted to particular physiological requirements of subgroups of neurons. This maturational setting affects the capacity of neurons to survive into old age.

Oxidative and other forms of free radical damage are heavily implicated in neurodegenerative disease as well as in normal aging. It is now known that neurotrophic factors can influence antioxidant defense in neurons and that free radical levels increase in neurons, such as those of the myenteric plexus, prior to the onset of age-related cell loss. A working hypothesis, then, is that differential, lifelong patterns of neurotrophin signaling are established during early postnatal life, and become more or less fixed in maturity. These differential and adaptive levels of signaling then determine, among other things, neuronal capacity to resist prolonged exposure to free radical damage. The neurons with low levels of signaling succumb first (Figure 4).

An alternative way in which neurotrophin signaling might be altered in adulthood and old age concerns the recent discovery that the precursor pro-form of NGF, previously thought to be expressed within target cells prior to its cleavage by enzymes of the proconvertase family, is widely distributed in the adult nervous system, including in nonneural targets of autonomic nerves. ProNGF levels are elevated in the pregnant uterus, in association with the atrophy and degeneration of the innervating sympathetic neurons (see the earlier section on neuronal growth). ProNGF is also the predominant form of NGF present in the Alzheimer's diseased brain and in the spinal cord after injury – two situations known to

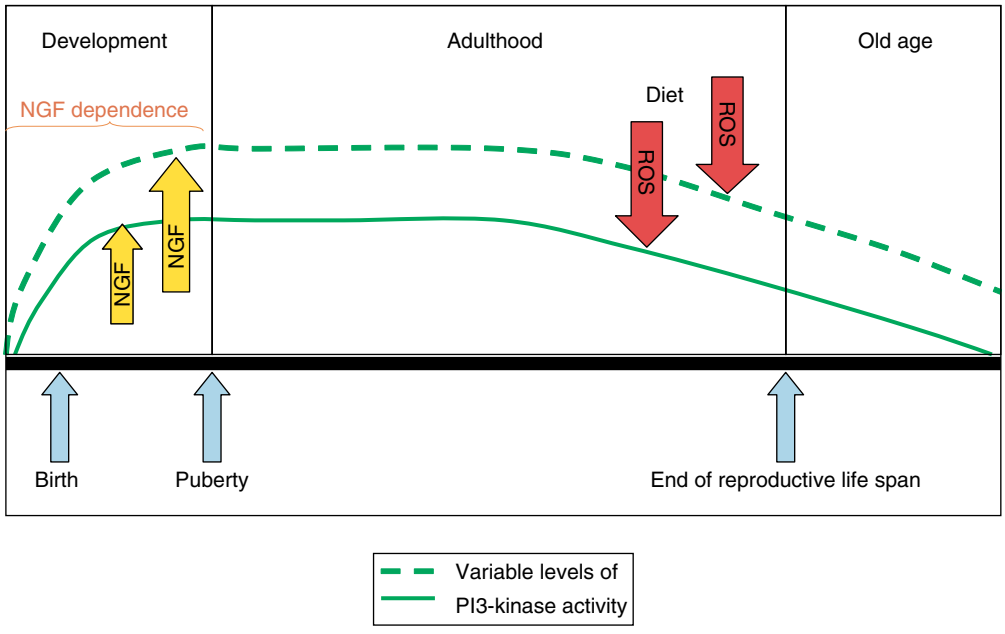

Figure 4 Model illustrating mechanisms regulating lifelong survival of autonomic neurons. During perinatal life, autonomic neurons become increasingly independent of target-derived NGF for their survival as a result of upregulation of intraneuronal PI3-kinase activity. The degree of upregulation is variable (see dashed and solid green lines), locally specific and adaptive, being set initially by the levels of the target-derived NGF (large and small arrows). Later in life (around puberty and after), PI3-kinase activity is maintained independently of NGF availability. Toward the end of the reproductive life span, exposure to reactive oxygen species (ROS) downregulates PI3-kinase activity, increasing vulnerability to age-related neuron cell death. AL feeding (see Figure 3) increases exposure to ROS and thereby accentuates this process, while CR provides protection.

be associated with widespread neuron cell death. ProNGF has been shown to induce apoptotic cell death in PC12 cells as well as SCG neurons from young animals. ProNGF signals via the receptor sortilin (first discovered as a member of the Vsp10 family of sorting receptors and later identified as the receptor for neurotensin) acting in combination with p75. At present, there is little evidence from the aging nervous system that proNGF contributes to age-related cell death. However, the sortilin receptor is expressed in the adult (including human) brain; therefore, this possibility cannot be ruled out.

Extrinsic Factors Affecting Neuron Survival

Factors that influence life span have been shown, unsurprisingly, to also affect neuron survival. The classical extender of life span in rodents, caloric restriction (CR), has remarkable effects in rescuing myenteric neurons of rats from age-related cell death. CR also reduces free radical levels in neurons, and appears to achieve this by influencing neurotrophic factor-mediated antioxidant defense. Exercise and the environmental stimuli to which laboratory rodents are exposed (including group vs single animal housing) may also affect health, life span, and, perhaps, cell survival. Further studies are required to quantify these phenomena and to explore the underlying mechanisms.

Conclusions

Autonomic neuronal plasticity is regulated by core mechanisms which are set up during development, modified and consolidated during maturation, and maintained into adulthood and old age. These mechanisms, which are adaptive, may have deleterious pleiotropic consequences for some groups of aging neurons, making them vulnerable to neurodegeneration. Neurodegeneration probably occurs initially as loss of terminal nerve fibers (axons and dendrites) and shrinkage of the cell soma and may progress ultimately to cell death. Interactions between neurons and their target tissues, including the provision of nerve growth and survival promoting factors by targets and the related expression of appropriate receptors in neurons, comprise an intimate and essential interaction at the core of plasticity responses throughout life. Intraneuronal signaling pathways have specific roles in regulating plasticity, including growth and survival responses, and, probably, maintenance of neurotransmitter phenotype and antioxidant defense. Extrinsic factors, including diet and other behaviorally regulated phenomena, have major effects on neural plasticity, probably by acting on intraneuronal signaling pathways.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Neuroanatomy; Autonomic Neuroplasticity: Development; Autonomic Neuroplasticity and Regeneration; Autonomic Nervous System.

Further Reading

Anderson DJ (2000) Genes, lineages and the neural crest: A speculative review. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 355: 953–964.

Brauer MM, Shockley KP, Chavez R, Richeri A, Cowen T, and Crutcher KA (2000) The role of NGF in pregnancy-induced degeneration and regeneration of sympathetic nerves in the guinea pig uterus. *Journal of the Autonomic Nervous System* 79: 19–27.

Bruno MA and Cuello AC (2006) Activity-dependent release of precursor nerve growth factor, conversion to mature nerve growth factor, and its degradation by a protease cascade. *Proceedings of the National Academy of Sciences of the United States of America* 103: 6735–6740.

Cowen T (2002) Selective vulnerability in adult and ageing mammalian neurons. *Autonomic Neuroscience* 96: 20–24.

Cowen T and Gavazzi I (1998) Plasticity in adult and ageing sympathetic neurons. *Progress in Neurobiology* 54: 249–288.

Dontchev VD and Letourneau PC (2003) Growth cones integrate signaling from multiple guidance cues. *Journal of Histochemistry and Cytochemistry* 51: 435–444.

Fahnestock M, Michalski B, Xu B, and Coughlin MD (2001) The precursor pro-nerve growth factor is the predominant form of nerve growth factor in brain and is increased in Alzheimer's disease. *Molecular and Cellular Neuroscience* 18: 210–220.

Gabella G, Berggren T, and Uvelius B (1992) Hypertrophy and reversal of hypertrophy in rat pelvic ganglion neurons. *Journal of Neurocytology* 21: 649–662.

Gershon MD (1998) V. Genes, lineages, and tissue interactions in the development of the enteric nervous system. *American Journal of Physiology* 275: G869–G873.

Glebova NO and Ginty DD (2005) Growth and survival signals controlling sympathetic nervous system development. *Annual Review of Neuroscience* 28: 191–222.

Miller FD and Kaplan DR (2001) Neurotrophin signalling pathways regulating neuronal apoptosis. *Cellular and Molecular Life Sciences* 58: 1045–1053.

Nykjaer A, Lee R, Teng KK, et al. (2004) Sortilin is essential for proNGF-induced neuronal cell death. *Nature* 427: 843–848.

Orike N, Middleton G, Buchman VL, Cowen T, and Davies AM (2001) Role of PI 3-kinase, Akt and Bcl-2-related proteins in sustaining the survival of neurotrophic factor-independent adult sympathetic neurons. *Journal of Cell Biology* 154: 995–1005.

Thrasivoulou C, Soubeyre V, Ridha H, et al. (2006) Reactive oxygen species, dietary restriction and neurotrophic factors in age-related loss of myenteric neurons. *Aging Cell* 5: 247–257.

Xu XM, Fisher DA, Zhou L, et al. (2000) The transmembrane protein semaphorin 6A repels embryonic sympathetic axons. *Journal of Neuroscience* 20: 2638–2648.

Autonomic Neuroplasticity and Regeneration

P J Kingham and G Terenghi, University of Manchester, Manchester, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuronal plasticity can be defined as the responsiveness of a neuron to its surroundings and how this changes following injury. The normal phenotype and function of a neuron depends on both the integrity of the neuron and also its connection with other neurons and end targets. During development of the nervous system, the terminals of neurons project to the same target and compete for limited quantities of neurotrophic growth factors, which then bind to surface receptors and are internalized and transported to the cell body to provide key signals for survival and differentiation. Disruption of these connections, through injury, initiates a sequence of molecular and cellular responses which act to promote regeneration and if successful reestablish contact with end targets. Therefore, the plastic changes which remodel the autonomic nervous system following injury involve a balance between cell death, due to trophic starvation, and regeneration of surviving neurons which switch from a differentiated to growing phenotype.

Cell Survival and Growth Factors in the Autonomic Nervous System

Nerve growth factor (NGF) is the classic target-derived neurotrophin synthesized by end targets. Sympathetic neurons were the first neurons to be identified as NGF responsive and their targets, vascular and smooth muscle, skin cells such as keratinocytes and melanocytes, and various endocrine tissues produce large quantities of the growth factor. NGF has two known receptors, TrkA and $p75^{NTR}$, which are both expressed by adult sympathetic neurons. TrkA is a transmembrane protein which serves as a receptor tyrosine kinase, and the neurotrophic properties of NGF are largely attributable to its effects on this receptor. $p75^{NTR}$ interaction with TrkA further modulates the responsiveness to NGF. The use of primary cell cultures of sympathetic neurons or the PC12 cell line has led to the identification of numerous intracellular signaling pathways activated by NGF. In brief, binding of NGF to TrkA initiates receptor dimerization leading to phosphorylation of integral tyrosine residues, which results in activation of kinase activity and subsequent phosphorylation of adaptor proteins which link to the mitogen-activated protein (MAP) kinase signaling cascade. Activation of these proteins contributes to altered expression of NGF-responsive genes via the mobilization of transcription factors from the cytoplasm to the nucleus. In addition to the local effect of NGF on the axon terminals, it is retrogradely transported from the end target to neuronal cell bodies within the sympathetic ganglia via postganglionic axons. This occurs through endocytosis of the ligand receptor complex and the formation of signaling endosomes. These vesicles then accumulate in the neuronal soma where high levels of phosphorylated TrkA activate the signaling events necessary for survival.

The importance of NGF signaling for cell survival is illustrated by studies of NGF–/– animals, which show almost complete loss of sympathetic nerves. Lesion to a sympathetic nerve isolates it from the end target source of NGF and leads to a reduction in neuron number. Nevertheless, adult sympathetic neurons do not die acutely after such a challenge but instead undergo a gradual cell death. This resistance to growth factor deprivation suggests that protective mechanisms can be activated following nerve injury. This involves changes both at the lesion site and within the cell bodies. Ultimately, this leads to activation of signaling cascades which act to mimic the effects of NGF and other growth factors which are important for the regulation of autonomic nerve survival.

Early Responses to Nerve Injury

Nerve injury causes damage to axonal membranes and interrupts the retrograde flow of signals from the normal innervation target. Before the ends can be resealed, Ca^{2+} enters the axoplasm and activates calcium-dependent proteases such as calpains, leading to extensive cytoskeletal remodeling. This may underlie new growth cone formation and may direct new intra-axonal protein synthesis, preventing the axons from simply withering away. Both myelinated and unmyelinated (sympathetic postganglionic nerves) axons are surrounded by Schwann cells. In the healthy nerve, Schwann cells appear quiescent and their function is tightly regulated by the axon they ensheath. When this interaction is lost, Schwann cells become highly active. Within hours of nerve injury, Schwann cells produce elevated levels of inflammatory cytokines such as tumor necrosis factor α and interleukin-1 which act to attract infiltrating macrophages, T-cells, and neutrophils. Both resident and recruited macrophages are critical for removal of nerve debris, which might impede the regeneration of axons.

Schwann cells proliferate in large numbers at the distal stump, forming a strand of cells called the band of Büngner, which provides an indispensable pathway for guiding regenerating axons to their end target. Perhaps more importantly, the Schwann cells release a number of neurotrophic factors to compensate for the decline occurring as a result of lost contact with the end target. Following nerve injury, Schwann cells produce increased levels of NGF which remain 10- to 15-fold elevated for at least 2 weeks. High levels of NGF are maintained by interactions between Schwann cells and invading macrophages. NGF may subsequently be taken up by the newly forming growth cones at the proximal stump. It is likely that the growth factor is dispersed diffusely in a gradient fashion around regenerating axons and therefore boosts axonal elongation. Brain-derived neurotrophic factor (BDNF) is another growth factor important for the survival of sympathetic neurons. Quiescent Schwann cells contain low levels of BDNF, which is upregulated following axotomy. In contrast to NGF, the response to denervation is slower and peak levels of BDNF are reached weeks following injury. Ciliary neurotrophic factor (CNTF) is a growth factor important for the survival of parasympathetic ciliary ganglion neurons. Schwann cells in intact nerves contain high levels of CNTF. Following injury, the levels decline and are reestablished only upon reinnervation, suggesting that contact with axons is necessary for the synthesis of CNTF in Schwann cells. Studies exploring the downstream targets of CNTF such as signal transducer and activator of transcription-3 (STAT-3) indicate that this could be an important mediator of early responses to nerve injury. Local release of CNTF by denervated Schwann cells activates the CNTF receptor on nearby axons, leading to phosphorylation of STAT-3 and its retrograde transport to the cell bodies of injured neurons. Therefore, the release of neurotrophic factors by Schwann cells at the injury site can help maintain a viable population of neurons which can switch to a growth phenotype, enabling their axons to regenerate across the lesion and back to the periphery.

Signaling Pathways and Changes in Gene Expression in the Injured Neuron

Chromatolysis, a histological change involving the rough endoplasmic reticulum and nucleus, was first noted in axotomized cell bodies many years ago. It was suggested that this process reflected a change in the types of proteins the neurons were synthesizing. It is now well accepted that nerve injury induces changes in intracellular signaling pathways which result in the induction of transcription factors and gene activation leading to upregulation of growth-associated molecules and a switch from neurotransmission. The recent development of microarray technology has led to the identification of large numbers of neuronal genes regulated by nerve injury. With regard to the autonomic nervous system, these changes have been best studied in the sympathetic, superior cervical ganglion (SCG). This is partly due to the difficulties of injuring selectively parasympathetic axons, given that most of the peripheral neuron cell bodies lie in widely dispersed ganglia closely associated with, or in the end target. Additionally, *in vitro* SCG explants provide an intact ganglion maintaining the association between neurons and nonneuronal cells and allow the subsequent investigation of the role of individual regeneration-induced genes. Indeed, it has been shown that upregulation of leukemia inhibitory factor (LIF) in the satellite cells of the ganglion within 6 h of axotomy mediates the expression of many of the induced genes. Studies are now underway to elucidate the role of the gene products produced during regeneration, a selection of which are discussed below.

Transcription Factors

These proteins play a major role in coupling extracellular signals to altered gene expression in the neuron and contribute to the underlying neuroplasticity of the autonomic nervous system. STATs are transcription factors which are phosphorylated by janus (JAK) kinases in response to cytokine activation of cell surface receptor tyrosine kinases. Activated STATs dimerize and translocate to the nucleus where they activate transcription of cytokine-responsive genes. To date, at least three JAK kinases and six STAT proteins have been shown to participate in this complex signaling pathway. Axotomy of SCG results in the rapid and sustained activation of STAT-1α and STAT-3 proteins in neuronal cell bodies. This activation is dependent on LIF release from satellite cells. The signaling cascade may also in part be mediated by reduced trophic support since NGF and other growth factors have been shown to reduce STAT expression. The STAT pathway may be particularly important for the regulation of neuropeptide expression in the SCG (see below).

Nuclear factor kappa B (NF-κB) is another transcription factor that plays a key role in the regulation of genes involved in nervous system function. NF-κB is a heterodimeric protein, principally composed of a 50 kDa subunit and a 65 kDa subunit. Typically, NF-κB is sequestered in the cytoplasm by the specific inhibitory protein IkB. The phosphorylation and subsequent proteasomal degradation of IkB releases

NF-κB and permits its translocation into the nucleus in which it can bind DNA and regulates transcription. NF-κB is constitutively activated in sympathetic neuron cultures maintained in the presence of NGF and withdrawal of the growth factor downregulates NF-κB activity. This correlates with increased apoptosis and therefore NF-κB might play a role in the survival of sympathetic neurons.

Like NF-κB, activating protein-1 (AP-1) transcription factors are dimeric molecules and comprise two major protein families: the Jun family and the Fos/activating transcription factor (ATF) family. Activation of the transcription factor c-Jun through phosphorylation by the MAP kinase, c-Jun N-terminal kinase (JNK), has been implicated in neuronal stress responses. Neurons contain low basal levels of the c-jun protein and axonal damage results in dramatic increases in expression levels, most likely as a consequence of decreased trophic support. The relative role of c-jun in neuronal cell death and regeneration is still unknown. In sympathetic neuron culture, c-jun is activated following NGF withdrawal and correlates positively with the onset of apoptosis. In contrast, *in vivo* studies of peripheral nerves have shown that deletion of c-jun hinders the expression of genes and proteins associated with axonal regeneration and reduces the speed of target reinnervation. The activity of other transcription factors such as NF-κB and ATF-3 can suppress the c-Jun-dependent transcription of cell death genes. Microarray data from axotomized SCG indicate that all three transcription factors are upregulated, suggesting a complex interaction for the balance of neuronal cell death and regeneration.

Cell Survival Proteins

The expression of a number of genes encoding anti-apoptotic proteins are increased in axotomized SCG. These proteins may contribute to the survival of injured neurons in response to growth factor deprivation. Heat shock proteins (HSPs) are a group of molecules expressed after cellular stress such as nerve injury. It has been suggested that these proteins serve a protective role by stabilizing other protein structures that are sensitive to denaturation and to help them to retain or to restore their native, functional conformations. The production of HSP70 mRNA is increased rapidly following SCG injury, and although protein studies have not been performed, it is likely that the ganglion cells expressing high HSP70 levels show a greater resistance to apoptosis. In mixed peripheral nerves, HSP70 protein remains elevated for days and positively correlates with surviving neurons. One particular stress a neuron encounters following axotomy is an increase in oxidative free radicals. Elevated expression levels of superoxide dismutase following sympathetic axotomy may protect against cell death. Of the pathways leading to apoptosis, mitochondria play a central role releasing a number of proapoptotic molecules resulting in the activation of caspases and fragmentation of DNA. Studies of cultured sympathetic neurons indicate that the translocation of Bax protein, from cytoplasm to mitochondria, is a critical and irreversible step in the cell death pathway following trophic factor deprivation. However, the extent to which the same pathway is used *in vivo* is still not clear. Nevertheless, the observation that Bax inhibitor-1 expression is increased in the SCG following axotomy suggests a protective response.

Neural Transmission Molecules

Noradrenaline (NA) is the principal neurotransmitter of peripheral sympathetic neurons and is stored in vesicles, which are formed in the cell body and then transported into the axon, toward the nerve terminal, by fast axonal transport. Here the neurotransmitter is released upon the arrival of a nerve impulse. Both large (80–100 nm diameter) and small (40–50 nm diameter) vesicles exist, which in addition to NA contain varying types of peptides (see below). ATP is also an important co-transmitter with NA. The phenotype of sympathetic neurons is highly plastic both *in vitro* and *in vivo*. For example, dissociated neurons exposed to medium conditioned by nonneuronal cells can upregulate choline acetyltransferase activity and develop cholinergic characteristics. Likewise, the sympathetic axons innervating rat sweat glands are initially noradrenergic but as the sweat gland matures, NA levels fall and cholinergic properties are acquired. This indicates that sympathetic neurons are highly responsive to extracellular factors. This plasticity is especially apparent in neuropeptide expression levels following nerve injury.

Neuropeptide Y (NPY) is a 36-amino-acid peptide, which is highly concentrated in the large vesicles of sympathetic postganglionic neurons and is co-released with NA. Following axotomy, NPY levels are increased both *in vivo* and in SCG explants. This occurs in response to as yet unidentified factors released from injured nonneuronal cells in the ganglion. In contrast to the high number of adult sympathetic neurons positive for NPY, other neuropeptides such as galanin and vasoactive intestinal polypeptide (VIP) are rarely expressed in healthy neurons. Both are significantly increased in response to axotomy as the result of LIF release in the ganglion. In the case of VIP, it has been shown that LIF promotes the binding of transcription factors to a cyclic AMP responsive element located within the 5′-flanking region of the

VIP gene. Other cytokines such as interleukin-1 can also potentiate the number of VIP-positive sympathetic neurons. In contrast, administration of NGF partially suppresses the increases in galanin and VIP positive neurons following axotomy. This suggests that under physiological conditions, NGF may play a role in controlling the neuropeptide phenotype of sympathetic neurons. Recent microarray data have identified two additional upregulated neuropeptides, cholecystokinin (CCK) and PTHrP, which had not previously been observed in sympathetic neurons. It is tempting to speculate that in addition to their classic transmitter role, neuropeptides are involved in the growth response during sympathetic nerve regeneration. The observation that galanin knockout mice display retarded regeneration of the sciatic nerve indicates this may be the case. Furthermore, VIP increases the neurite extension of sympathetic neuroblasts and the survival of cultured sympathetic neurons deprived of NGF. In addition to affecting neurons, VIP can also act on Schwann cells to increase the expression of extracellular matrix (ECM) molecules necessary for promoting regeneration.

Regulators of Neurite Outgrowth

Although sprouting at the lesion site occurs within hours of injury, it takes several days before axonal outgrowth emerges from the proximal stump. During this time, there is upregulation of many proteins necessary for remodeling the growth cone of regenerating axons. Organized cytoskeletal structures in association with activated intracellular signal transduction pathways determine the shape and motility of growth cones. Many years ago, it was noted that while molecules involved in the synthesis of neurotransmitters such as tyrosine hydroxylase were decreased, important structural proteins including actin and tubulin were increased, indicating an active remodeling process. GAP-43 is a small acidic protein, which is highly associated with outgrowth of neurites during development. Following injury to the adult nerve, GAP-43 is synthesized in the cell body and transported to the regenerating tip where it accumulates at high levels in the growth cones. The N-terminal of GAP-43 binds to the plasma membrane of the growth cone and the C-terminal with the cytoskeletal components of the axon. Calmodulin, an important mediator of signal transduction, is also attached to GAP-43. Phosphorylation of GAP-43 by protein kinase C releases calmodulin and leads to growth cone extension. In contrast, activation of G-proteins at the N-terminal end can lead to inhibition of growth cone extension. Thus, GAP-43 plays an important role in signaling and controls the spreading, branching, and adhesion of the growth cone.

The extension and path finding of growth cones are mediated by substrates such as ECM molecules with which they interact. The basal lamina is the ECM to which regenerating axons attach preferentially. The most potent adhesion molecule for promoting axonal outgrowth in this matrix is laminin, which binds with integrin receptors on the axon surface. An integrin is a heterodimer protein composed of α and β chains, and peripheral nerve injury results in the upregulation of the $\alpha6$, $\alpha7$, and $\beta1$ integrin subunits. Neuronal jun-deficient mice do not show an upregulation of $\alpha7$ integrin after injury, which contributes to the poor regenerative response seen in these animals. In addition, the expression of the fibronectin gene is significantly increased following axotomy of sympathetic neurons, and this may contribute to Schwann cell migration to the regenerating front. Therefore, the increased expression of cell adhesion molecules is likely to lead to activation of intracellular signal transduction pathways and remodeling of the cytoskeleton, facilitating the growth of regenerating axons. This is in part mediated by the nonreceptor tyrosine kinase, pp60c-src, which is also upregulated following nerve injury, and concentrates within the growth cone. Pp60c-src directly links a complex of integrins, vinculin and actin filaments, which can control cytoskeletal transformations. Transient activation of this complex, by phosphorylation events, enables growth cone motility. In addition, phosphorylation of tubulins by other members of the c-src family of tyrosine kinases suppresses the formation of microtubules creating a highly plastic environment in the regenerating axon.

Since regrowth is necessary for functional recovery following axotomy, it is interesting to note that microarray data often produce 'hits' for molecules that are known to regulate neurite outgrowth. For instance, axotomy of the SCG leads to increased expression of the arginase 1 gene, an enzyme involved in polyamine synthesis. Arginine is hydrolyzed to ornithine, which is then converted to the diamine putrescine and then the polyamines spermidine and spermine. Levels of these molecules increase in axotomized ganglia. Pharmacological blockade of polyamine synthesis inhibits the regenerative response, while administration of exogenous polyamines promotes neurite outgrowth and functional recovery following injury to *in vivo* SCG neurons. In contrast, the levels of inhibitors of regeneration such as the proteinase activated receptor or thrombin receptor fall in response to axotomy. Together, these changes lead to the promotion of outgrowth and an enhanced regenerative response.

Conclusions

The aim of this article is to highlight how autonomic nerves respond to nerve injury and produce a regenerative response. The recent use of microarray technology combined with classic protein biochemistry studies have helped to elucidate many of the signaling pathways involved and demonstrate the highly plastic nature of the peripheral nervous system. Activation of mechanisms that compensate for neurotrophic factor deprivation aim to maintain sufficient neuronal numbers which can then switch from a differentiated to growing phenotype and subsequently reinnervate the end target. Further investigation of these endogenous responses to injury should lead to a new generation of therapies to enhance nerve repair.

See also: Apoptosis in Nervous System Injury; Autonomic Neuroplasticity and Aging; Autonomic Neuroplasticity: Development; Axonal Regeneration: Role of Growth and Guidance Cues; Nerve Growth Factor; Peripheral Nerve Regeneration: An Overview; Schwann Cells and Axon Relationship; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Bennet MR, Gibson WG, and Lemon G (2002) Neuronal cell death, nerve growth factor and neurotrophic models: 50 years on. *Autonomic Neuroscience* 95: 1–23.

Boeshore KL, Schreiber RC, Vaccariello SA, et al. (2004) Novel changes in gene expression following axotomy of a sympathetic ganglion: A microarray analysis. *Journal of Neurobiology* 59: 216–235.

Glebova NO and Ginty DD (2005) Growth and survival signals controlling sympathetic nervous system development. *Annual Review of Neuroscience* 28: 191–222.

Hendry IA (1992) Responses of autonomic neurones to target deprivation. In: Hendry IA and Hill CE (eds.) *Development, Regeneration and Plasticity of the Autonomic Nervous System*, pp. 415–462. Chur, Switzerland: Harwood Academic Publishers.

Hou XE and Dahlstrom A (2000) Synaptic vesicle proteins and neuronal plasticity in adrenergic neurons. *Neurochemical Research* 25: 1275–1300.

Ide C (1996) Peripheral nerve regeneration. *Neuroscience Research* 25: 101–121.

Klimaschewski L, Tran TD, Nobiling R, and Heym C (1994) Plasticity of postganglionic sympathetic neurons in the rat superior cervical ganglion after axotomy. *Microscopy Research and Technique* 29: 120–130.

Kury P, Stoll G, and Muller HW (2001) Molecular mechanisms of cellular interactions in peripheral nerve regeneration. *Current Opinion in Neurology* 14: 635–639.

Makwana M and Raivich G (2005) Molecular mechanisms in successful peripheral regeneration. *FEBS Journal* 272: 2628–2638.

Raivich G, Bohatschek M, Da Costa C, et al. (2004) The AP-1 transcription factor c-Jun is required for efficient axonal regeneration. *Neuron* 43: 57–67.

Terenghi G (1999) Peripheral nerve regeneration and neurotrophic factors. *Journal of Anatomy* 194: 1–14.

Zigmond RE and Sun Y (1997) Regulation of neuropeptide expression in sympathetic neurons. Paracrine and retrograde influences. *Annals of the New York Academy of Sciences* 814: 181–197.

Autonomic Neuroplasticity: Development

P G Smith, University of Kansas Medical Center, Kansas City, KS, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The concept that the nervous system can undergo substantial reorganization has been appreciated for over a century. In 1909, in *Modern Problems in Psychiatry,* Ernesto Lugaro used the term 'plasticity' in the context of the nervous system to convey the idea that chemical interactions responsible for the developmental patterning of the nervous system also provide adaptive mechanisms for behavioral modification and nervous system repair.

While plasticity can occur throughout the nervous system, the autonomic nervous system arguably has been most instructive in revealing the underlying principles. One reason is that the peripheral autonomic nervous system provides a particularly tractable system for studying neuroplasticity. Because its principal neurons are located outside of the central nervous system, it is relatively easy to isolate and remove relatively pure neuronal populations, either for cell culture under regimented conditions or to examine the effects of extirpation or transposition. A second reason is that sympathetic neurons are responsive to the prototypic neurotrophic factor, nerve growth factor (NGF); thus the role of target-derived proteins and autonomic neuronal survival and growth has been recognized and manipulated in this system for over half a century. Moreover, this system has proved not only to be highly complex, with multiple factors controlling development, but also very amenable to investigations using genetic manipulation to elucidate the role of various genes and proteins. Accordingly, the autonomic nervous system has served as a primary model system for advancing our knowledge of cellular and molecular processes regulating nervous system development. Perturbations to these processes have allowed us to more fully appreciate the repertoire of responses that comprise developmental neuroplasticity.

The process of nervous system development is itself inherently plastic. During development, undifferentiated cells undergo variable degrees of migration and cell division, giving rise to pools of precursor cells which provide the anatomical substrate of the nervous system. Precursor cells differentiate, taking on specific properties and becoming responsive to chemical cues in their environment, and eventually elaborate axons which find their way to appropriate targets. Neuron numbers are further refined through apoptosis in response to limitations in available target-derived survival factors, thus matching neuronal numbers to target mass and cell type. Target innervation density is then established through the actions of both propulsive and repulsive signals, and retrogradely transported proteins contribute to final determination of transmitter phenotype. In the perinatal period, hormonal factors play increasingly important roles in refining nerve–target relationships. Perturbations to any of these dynamic processes can lead to significant alterations in nervous system properties, which are appropriately defined as 'developmental neuroplasticity.'

Developmental autonomic neuroplasticity is defined here broadly to include pre- and postnatal developmental processes associated with sympathetic and parasympathetic nervous system formation and those perturbations that affect structure, phenotype, and function. These include (1) molecular signaling events regulating autonomic neuronal ontogeny and the impact of their disruption through genetic manipulation and transpositioning or removing individual structures, (2) the establishment and maintenance of autonomic projections to peripheral targets, and (3) establishment, maintenance, and alterations of neurochemical phenotype during development or in response to injury.

Autonomic Gangliogenesis

A defining event in the ontogeny of the autonomic nervous system is the formation of motor ganglia. These consist of aggregates of sympathetic and parasympathetic postganglionic neurons, located in the periphery, which receive excitatory cholinergic synaptic input from preganglionic neurons located in the central nervous system. Autonomic postganglionic neurons derive from the neural crest, a transient structure that originates from cells migrating rostrocaudally from the neural tube shortly after closure. Neural crest precursor cells give rise to a variety of cell types, including melanocytes, peripheral glia, cartilage, smooth muscle, adrenomedullary chromaffin cells, and sensory and autonomic neurons. Those destined to become autonomic ganglion neurons of the peripheral nervous system migrate ventrally through the rostral somitic mesoderm to populate sympathetic and parasympathetic ganglia. Any perturbation that disrupts normal ganglion formation will severely impact autonomic nervous system organization and function.

The generation of autonomic neuronal precursors is complex, involving a variety of transcription factors and signaling proteins (Figure 1, Table 1). A pool of

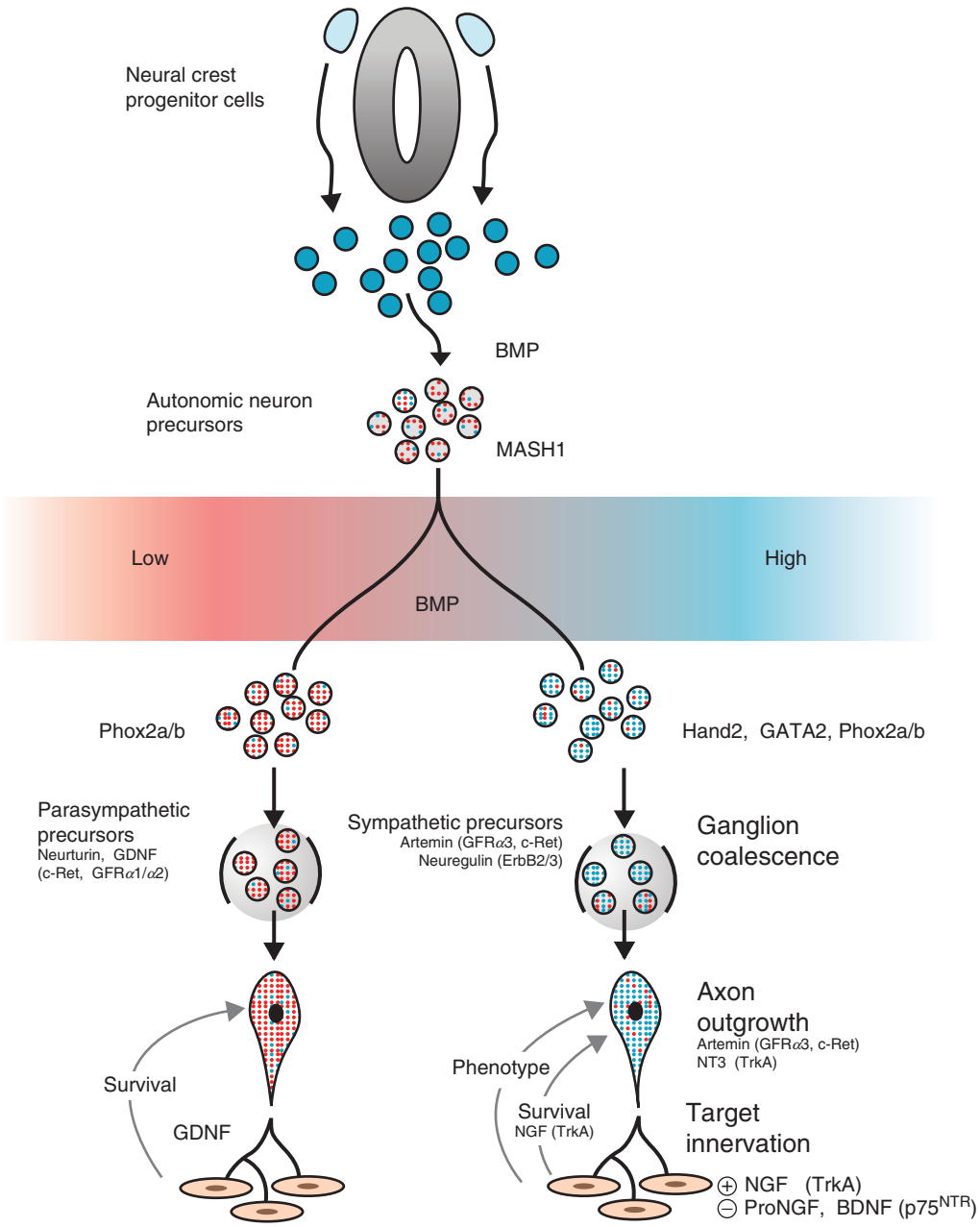

Figure 1 Schematic diagram indicating processes involved in the ontogeny of the autonomic nervous system, and the roles of some regulatory proteins. Neural crest gives rise to progenitor cells which differentiate into mammalian achaete–scute homolog-1 (MASH1)-expressing autonomic neuronal precursors under the influence of members of the bone morphogenetic protein (BMP) family. Neuronal identity appears to be determined within a BMP gradient, with sympathetic neurons differentiating within high concentrations and parasympathetic neurons differentiating within low concentrations. Parasympathetic neurons require expression of members of the Phox2 family of transcription factors, which are also required for sympathetic neurons, in addition to Hand2 and Gata2. Coalescence of precursors into ganglia requires artemin acting on the GFRα3/c-Ret receptor complex and neuregulin acting on ErbB receptors for sympathetic ganglia, and glial cell-derived neurotrophic factor (GDNF) and neurturin acting on GFRα1 or GFRα2 complexed with c-Ret. Sympathetic axon outgrowth requires both artemin and neurotrophin-3 (NT3), which acts on the TrkA receptor, whereas parasympathetic axon outgrowth may be dependent on GDNF or related proteins. Target-derived factors regulate target innervation: in sympathetic neurons, nerve growth factor (NGF) acting on TrkA can promote neuronal survival and increase innervation density, whereas proNGF and brain-derived neurotrophic factor (BDNF) acting selectively on the $p75^{NTR}$ diminish target innervation density. GDNF regulates innervation density of some parasympathetic targets. Target-derived factors can also influence neuronal neurochemical phenotype.

Table 1 Factors regulating autonomic nervous system development, and effects of mutations of regulatory genes

Factor	*Role in development*	*Effect of deletion*
MASH1, CASH1	Transcription factors specifying autonomic neuron identity	Autonomic neurons fail to form
Bone morphogenetic protein	Regulates formation of autonomic precursors, and differentiation of sympathetic and parasympathetic neurons	Sympathetic and parasympathetic neurons fail to form
Phox2a/b	Transcription factors required for catecholaminergic phenotype	Sympathetic and parasympathetic neurons fail to develop
Hand2, Gata2	Transcription factors specifying sympathetic fate	Absence of sympathetic neurons
Neuregulin, ErbB2/3	Promote sympathetic precursor migration	Inappropriate sympathetic trunk ganglion formation
Hepatocyte growth factor	Neuroblast survival and differentiation	Reduction in sympathetic ganglion neurons
Semaphorin 3A	Neural crest cell migration	Abnormal sympathetic ganglion formation
Neurotrophin-3	Sympathetic axon outgrowth	Impaired axon extension, fewer surviving sympathetic neurons due to inaccessibility to target-derived trophic factors; impaired cholinergic sympathetic phenotype
Glial cell-derived neurotrophic factor, c-Ret, GFRα1/2	Parasympathetic neuron survival, neuroblast migration; axon guidance	Reduced numbers of parasympathetic postganglionic neurons, impaired axon outgrowth
Nerve growth factor	Target-derived protein regulating sympathetic neuronal survival in early postnatal development, and target innervation density	Augmented cell death of sympathetic neurons; absent or diminished sympathetic innervation of some targets
Artermin, c-Ret, GFRα3	Sympathetic neuroblast migration; rostral sympathetic ganglion formation; sympathetic axon outgrowth	Misplaced superior cervical ganglion; impaired sympathetic axon outgrowth
Neurturin	Maintenance of parasympathetic target innervation	Reduced parasympathetic innervation

autonomic neuron progenitor cells differentiates from neural crest precursors under the influence of local expression of bone morphogenetic proteins (BMPs). Differentiation of both sympathetic and parasympathetic precursor cells is dependent upon their expression of the mouse or chick achaete–scute homolog transcription factors, MASH1 or CASH1, respectively, and animals lacking these proteins fail to form autonomic neurons. The transcription factors Phox2a and Phox2b are also involved in expression of catecholaminergic properties, which are characteristic of both sympathetic and parasympathetic neuronal precursors. Phox transcription factors are required for early survival of both sympathetic and parasympathetic autonomic neuron precursors. Sympathetic neuronal precursors are further dependent upon the Hand2 and Gata2 transcription factors.

Autonomic neuronal precursors migrate ventrally toward their ultimate destination, guided by local signaling proteins which, for truncal sympathetic neurons, include neuregulin and Sema3A, acting on the ErbB and neuropilin 1 receptors, respectively. Artermin, acting on the c-Ret and GFRα3 receptors, appears to play an important role in sympathetic ganglion coalescence and in providing chemotactic signals responsible for rostral migration of superior cervical ganglion precursor neurons. The factors regulating parasympathetic neuron migration are less understood, but evidence indicates that in some cranial parasympathetic ganglia, precursors are dependent initially upon glial cell-derived neurotrophic factor (GDNF) via the c-Ret and GFRα1 receptors for ganglion cell proliferation and migration, and subsequently on neurturin and the GFRα2 receptor for neuronal survival and ganglion development.

During normal development, different regions of the neural crest give rise to specific structures. In chick–quail chimeras, for example, rostral parasympathetic ganglion neurons derive from mesencephalic neural crest, whereas sympathetic ganglion neurons derive from the truncal neural crest extending below somite 4. However, premigratory precursor cells apparently are not committed to specific fates. Thus, neural crest cells from the vagal region that would normally give rise to enteric ganglion cells, when transplanted caudally, give rise to apparently normal sympathetic ganglion neurons and adrenomedullary cells. This suggests that neural crest derivatives at this stage remain multipotent and substantially plastic, and differentiation into mature neural crest derivatives is determined largely by peri- and postmigratory environments.

While most migrating neural crest cells become irrevocably committed during the process of migration and gangliogenesis, it is now clear that a small population of precursor cells remains undifferentiated

and may give rise to nascent neurons and glia throughout development and maturity. These presumptive stem cells may serve as a source of cells that can contribute to continued plasticity in the developing and adult organism.

Plasticity of Autonomic Neuroeffector Pathways

Axon Outgrowth

Following migration and coalescence of ganglion cells, connectivity between the central nervous system and peripheral targets must be established. Autonomic neurons elaborate axons in the prenatal and early postnatal periods. Preganglionic axons traverse the anterior region of the somitic mesoderm. Somitic mesoderm appears to play a critical role in preganglionic axon guidance, and removal or transpositioning of somites leads to major alterations in axon trajectory. Preganglionic axons project to the region of their intended ganglion, even if the ganglion has been removed. Once having entered the ganglion, preganglionic axons synapse with the postganglionic neurons; while some synapses are present in the prenatal period, the first postnatal week is characterized by dramatic synaptogenesis and establishment of a functional pathway, at least in the sympathetic inferior cervical ganglion of the rat. Ganglion neurons are initially polyinnervated by convergent inputs from multiple preganglionic neurons, and normal reductions through axon pruning taking place postnatally.

Autonomic postganglionic neurons elaborate axons shortly after exiting the cell cycle, and postganglionic axons first reach their intended targets in the late postnatal period in rodents; substantial increases in numbers of target projecting neurons occur through the first 2–3 postnatal weeks. In the case of sympathetic axonal outgrowth, this appears to be regulated predominantly by two locally produced proteins. Artemin, a member of the GDNF family of trophic factors, is produced by vascular smooth muscle cells in early development, and sympathetic axons at this developmental stage express the artemin receptors c-Ret and GFRα3. Neurotrophin 3 is also produced locally and activates sympathetic TrkA receptors, and is required for normal postganglionic axon outgrowth. The association of these growth factors with vascular smooth muscle is believed to account for the fact that sympathetic pathways associate closely with blood vessels as they course to their target destinations. Parasympathetic axon outgrowth is dependent upon GDNF and its receptors c-Ret and GFRα1. Disruptions of these growth factors or their receptors have serious consequences; without sufficient axon outgrowth, not only are targets hypoinnervated, but neurons are also in turn deprived of required target-derived survival factors and hence undergo abnormally extensive cell death, leading to fewer ganglion neurons. In some cases, diminished postganglionic axon outgrowth affects the mature neurochemical phenotype because of deprivation of target-derived proteins that influence neurotransmitter properties.

One feature of the postganglionic sympathetic system is a dramatic ability to reorganize in response to injury in the postnatal period. For example, innervation of the orbit normally derives exclusively from the ipsilateral superior cervical ganglion, and damage to this pathway in the adult results in sustained denervation and impairment, known clinically as Horner's syndrome. In this syndrome, the pupil is abnormally constricted (loss of excitatory sympathetic innervation to the pupil dilator muscle) and the upper eyelid fails to retract normally (ptosis, due to loss of excitatory sympathetic innervation to the superior tarsal smooth muscle which elevates the eyelid). However, in the neonatal rat, removal of one superior cervical ganglion results in sprouting of spared intracranial vascular sympathetic axons from the contralateral intact ganglion, which take atypical pathways to reach the orbital targets. Despite having significant phenotypic differences from normal resident innervation, such as expression of neuropeptide Y, as is typical of vascular sympathetic neurons but not normal resident tarsal muscle innervation, this aberrant pathway largely restores functional control of the target (**Figure 2**). The ability to establish contralateral orbital innervation is lost completely during the first postnatal month, largely through differentiation of smooth muscle-like myofibroblasts that define their trajectory to the orbit; this coincides with the time course reported for the loss of artemin expression during development in similar cell types, implicating a developmental reduction in artemin expression as a cause of the postnatal diminution in pathway plasticity.

Neuronal Survival: The Neurotrophic Hypothesis

Once the terminal axon reaches the target, a critical process takes place in which the relationship between target and neuron is established and refined, thus ensuring that the neuroeffector pathway is effective and appropriate. This process relies on target-derived proteins that can modify both neuronal survival and terminal axon growth.

The prototypical neurotrophin NGF has long been associated with sympathetic postganglionic axon sprouting *in vitro*, so it is not surprising that this protein would be considered to be the primary candidate for controlling sympathetic innervation. An abundance of

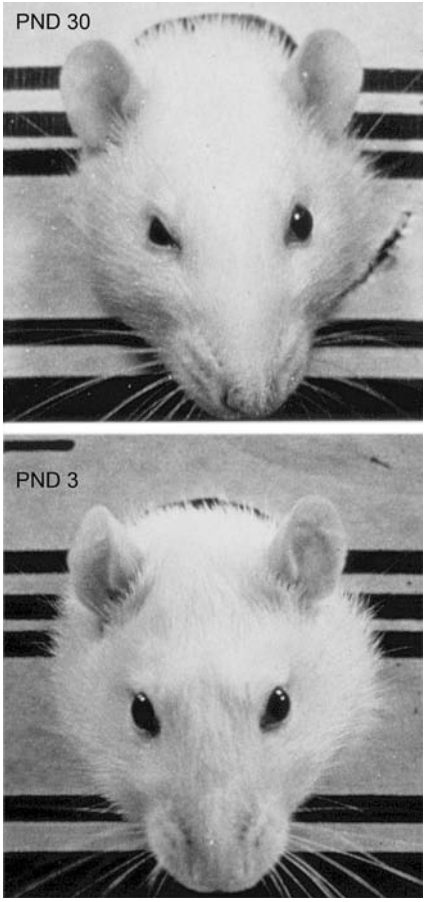

Figure 2 Plasticity in postganglionic pathway formation in the rat. The right superior cervical ganglion was removed from the rat in the top panel on postnatal day (PND) 30, whereas the same ganglion was removed from the rat in the bottom panel on PND 3. The photographs are of the awake rats at 3 months of age. Following PND 30 ganglionectomy, contraction of the superior tarsal muscle in the right eyelid is impaired, leading to ptosis and reduced palpebral fissure width. In the rat receiving ganglionectomy on PND 3, there is minimal evidence of impaired contraction; recovery is due to the formation of an atypical pathway deriving from intact contralateral neurons that normally project to cranial blood vessels to the denervated orbital targets.

information supports the idea that NGF (and many other trophic proteins in different contexts) is required for neuronal survival. Sympathetic neurons are produced in overabundance during development, and a relatively large proportion (in some cases >50%) undergoes developmental cell death in the perinatal period, leading to the ganglion cell numbers ultimately observed in the adult. Those neurons that do survive are believed to be the ones that have obtained access to adequate amounts of a target-derived survival protein (trophic factor, from the Greek term *trophos*, meaning 'to nourish'). However, trophic proteins are believed to be produced in limited quantities. In those sympathetic axons that reach the target first and have access to the greatest amounts of NGF, which binds to the NGF-selective receptor TrkA and is transported retrogradely to the cell body as receptor–ligand complex, programmed cell death (apoptosis) is prevented. Axons that cannot compete for adequate neurotrophic factor because they are misguided or arrive after target innervation has already achieved optimal density ultimately are eliminated through programmed cell death. The principle that an oversupply of developing neurons competes for target-limited trophic factor, thus matching neuronal numbers to target needs, represents the essence of the 'neurotrophic hypothesis.' While less is known about the trophic requirements of parasympathetic neurons, members of the GDNF family of ligands appear to serve similar roles. Modulating the amount of trophic factor available can induce considerable plasticity in the numbers of postganglionic neurons; addition of exogenous NGF or upregulation of its synthesis in peripheral tissues, such as epidermis, results in increased numbers of surviving sympathetic ganglionic neurons. Similarly, depletion of endogenous NGF during development, inactivation of the NGF gene, or null mutations of the TrkA gene encoding the NGF receptor all result in marked reductions in numbers of sympathetic ganglion neurons. A comparable situation has been shown to be the case for cranial parasympathetic neurons, the survival of which is reliant first upon GDNF and subsequently upon neurturin, acting, respectively, on GFRα1 and GFRα2 complexed with the coreceptor c-Ret.

Regulation of Target Innervation

The establishment of effective neuroeffector transmission is dependent upon the presence of appropriate numbers of terminal autonomic axons within the target tissue. Visceral targets show a wide range in the numbers of sympathetic and parasympathetic fibers present per unit tissue mass (innervation density). In light of NGF's well-documented role, not only in inducing sympathetic neuronal survival but also in eliciting robust outgrowth *in vitro*, it is not surprising that this neurotrophin would be a logical candidate for regulating sympathetic innervation density in peripheral targets. Indeed, early studies of several different targets revealed a positive correlation between levels of NGF mRNA or protein and the relative densities of sympathetic innervation. In keeping with this idea, mice with null mutations for NGF or TrkA fail to develop normal target innervation. Conversely, conditions favoring increased target levels of NGF can lead to increased target innervation. For example, spontaneously hypertensive rats display abnormally elevated levels of NGF in the mesenteric vasculature, and these vessels are also hyperinnervated by

sympathetic axons. Further, in keeping with the central tenet of the neurotrophic hypothesis, that neurons compete for limited amounts of trophic factors, neonatal destruction by capsaicin of sensory nociceptor nerves (which express TrkA and compete for target NGF) results in sympathetic axon hyperinnervation; this presumably occurs because reduced competition makes more NGF available to sympathetic axons. Accordingly, alterations in target NGF levels can result in significant plasticity in target innervation density.

While there is little doubt that NGF plays an important role in sympathetic target innervation, recent lines of investigation show that the relationship is far more complex. As indicated earlier, NGF is required for neuronal survival; therefore, loss of innervation in the absence of NGF or TrkA could be due solely to cell death rather than to an absence of local growth-promoting properties. Indeed, using mouse mutants lacking genes for both NGF and the apoptotic factor Bax (thus circumventing cell death that would normally occur in the absence of NGF), Ginty and colleagues found that the dependency on NGF for development of target innervation was highly variable, with innervation of some targets being drastically reduced (e.g., salivary glands, heart) while others were unaffected (trachea). This supports the idea that growth-promoting factors other than NGF contribute to target sympathetic innervation.

A second complicating factor is that, in addition to propulsive effects of proteins such as NGF and GDNF, autonomic target innervation appears to be controlled by proteins that are repulsive in nature. For example, sympathetic neurons are responsive to semaphorins and can be repulsed from targets expressing high levels of these proteins. Moreover, neurotrophins show surprising complexity. In addition to the TrkA receptor mediating sympathetic neuron survival and axon outgrowth, the pan-neurotrophin receptor $p75^{NTR}$ can facilitate ligand binding to TrkA, but can also mediate repulsive cues. Brain-derived neurotrophic factor (BDNF), a neurotrophin closely related to NGF, shows little affinity for TrkA but binds avidly to $p75^{NTR}$. When $p75^{NTR}$ is selectively activated, this inhibits sympathetic axonal outgrowth and contributes to sympathetic axon pruning (and in some cases can cause cell death). Importantly, in addition to BDNF, the precursor form of NGF, proNGF, shows relatively little affinity for TrkA but strongly binds $p75^{NTR}$, and, like BDNF, can be repulsive and pro-apoptotic to sympathetic neurons. Thus the extent to which NGF regulates target innervation density will depend upon which form predominates.

The balance between TrkA and $p75^{NTR}$ activation has important implications for developmental plasticity of sympathetic innervation. Some targets, such as those of the female reproductive tract, undergo considerable innervation plasticity in the postnatal period. For example, uterine innervation changes through puberty and beyond, under the influence of the gonadal steroid hormone estrogen; in fact, changes in estrogen lead to reductions in myometrial sympathetic innervation at the estrus stage of each estrous cycle. Available evidence shows that estrogen increases uterine expression of BDNF, and that fluctuations in this hormone underlie physiological axon pruning that takes place during puberty and throughout the reproductive cycle.

Neurochemical Plasticity

Neural crest progenitor neurons give rise to a richly diversified range of peripheral autonomic neurons. Fully differentiated parasympathetic neurons typically express a cholinergic phenotype, which includes choline acetyltransferase and vesicular acetylcholine transporter (VAChT), a nitrergic phenotype (neuronal nitric oxide synthase), and the neuropeptide vasoactive intestinal polypeptide (VIP). Most mature sympathetic neurons are noradrenergic, expressing the proteins tyrosine hydroxylase (TH), dopamine β-hydroxylase (DBH), vesicular monoamine transporter-2 (VMAT2), and the norepinephrine transporter; a significant proportion of noradrenergic sympathetic nerves, mainly those innervating blood vessels, also contain neuropeptide Y. About 5% of mature sympathetic neurons, which project to eccrine sweat glands and periosteum, are cholinergic. Neural crest progenitor cells also give rise to autonomic ganglionic small intensely fluorescent (SIF) cells, which are dopaminergic (express TH but not DBH), and adrenomedullary chromaffin cells (synthesize epinephrine and contain phenylethanolamine *N*-methyltransferase (PNMT), in addition to TH and DBH).

The differentiation of parasympathetic and sympathetic neurons appears to be regulated by way of a BMP gradient; progenitors most proximate to BMP-producing tissues (e.g., dorsal aorta) acquire a noradrenergic phenotype, while those exposed to lower concentrations become cholinergic parasympathetic neurons. Differentiation of both types of neurons requires MASH1 and Phox2a/b transcription factors, and noradrenergic neurons additionally require the Hand2 and Gata2 transcription factors. Differentiation of neural crest progenitors into adrenomedullary chromaffin cells is, in part, under control of the local environment; corticosteroid derived from the surrounding adrenal cortex is required for PNMT expression. Local factors have also been implicated in phenotype acquisition and maintenance in SIF cells,

which are located proximate to blood vessels and may require endothelial factors for differentiation.

The common origins of both sympathetic and parasympathetic neurons underlie sustained similarities throughout development and presage responses observed with certain perturbations. Experimental observations in neural crest-derived precursor cells reveal that the biochemical machinery necessary for acetylcholine synthesis is present well before commitment to final cell fate. Similarly, catecholaminergic properties have also been described as a feature of autonomic neuron progenitors prior to expression of other pan-neuronal properties.

Transmitter Phenotype 'Switching'

While genetic programming and instructive signals during precursor migration and gangliogenesis appear to play a major role in determining autonomic neuronal phenotype, target-derived proteins are also important in determining or refining mature neuronal phenotype. The best studied example of this is the transition from an adrenergic to a cholinergic phenotype in some sympathetic neurons. Early studies showed that neonatal sympathetic neurons cultured under certain conditions (e.g., in culture medium conditioned by cardiac myocytes) lose their noradrenergic phenotype, and this is replaced over time by cholinergic properties. In an extensive series of experiments, Story Landis and colleagues extended these findings *in vivo* by showing that sweat glands, which are innervated by cholinergic sympathetic axons in the adult, are actually innervated by axons with a catecholaminergic phenotype at postnatal day 4; however, these catecholaminergic properties (dense core vesicles, catecholamine histofluorescence, and TH immunoreactivity) are downregulated over the next 10 days and are replaced by cholinergic properties (VIP immunoreactivity and cholinesterase activity). An impressive body of evidence has been accrued to support the idea that these neurons undergo a 'phenotype shift' from catecholaminergic to cholinergic. Evidence in favor of this hypothesis includes several findings: (1) destruction of catecholaminergic fibers by neonatal administration of selective neurotoxins 6-hydroxydopamine, guanethidine, or antisera to NGF prior to the shift prevents sweat gland cholinergic innervation, even though cholinergic nerves should be unaffected by these treatments, (2) transplantation of the sweat glands to ectopic sites where sympathetic nerves normally do not show a cholinergic phenotype (trunk hairy skin, ocular anterior chamber co-cultures with the superior cervical ganglion) induces cholinergic properties, (3) replacement of sweat glands with a target that normally is not innervated by cholinergic sympathetic axons (salivary gland) prevents cholinergic properties from appearing in the resident innervation, and (4) mouse mutants that fail to develop sweat glands also fail to develop cholinergic fibers to the footpad. Further studies have shown that the shift is facilitated by an interactive loop with the target, in which noradrenergic transmission accelerates acquisition of cholinergic properties, and that a similar phenotypic shift occurs to sympathetic innervation to the periosteum; periosteal cells induce a VIP and VAChT phenotype in resident innervation and also when transplanted to ectopic sites. The factors responsible for this shift have not been definitively identified, but members of the gp130 cytokine family, which includes leukemia inhibitory factor and cardiotropin, promote cholinergic properties in cultured neurons and are strong candidates.

Recent studies imply that the establishment of cholinergic traits in sympathetic neurons is more complex. Using antibodies with high sensitivity to VAChT, it appears that cholinergic properties are demonstrable in subpopulations of sympathetic neurons prior to target innervation, and that sweat gland-innervating fibers can express VAChT at an age supposedly preceding the phenotype shift. Thus, the extent to which cholinergic properties are conferred *de novo* by the target or are instead an intrinsic property of sympathetic neurons undergoing normal development *in vivo* remains uncertain. Nonetheless, it is clear that some targets produce factors that are capable of inducing neurochemical phenotype plasticity in developing sympathetic neurons.

Another factor that can influence neurochemical phenotype is injury. Normally, VIP-immunoreactive neurons are only rarely encountered in the superior cervical ganglion. Following axotomy, however, VIP is dramatically upregulated *in situ* or under culture conditions. This implies that the factors normally responsible for providing a stable neurochemical phenotype can be affected by injury.

See also: Autonomic Nervous System: Neuroanatomy; Autonomic Neuroplasticity and Aging; Autonomic Neuroplasticity and Regeneration; Autonomic Nervous System Development; Nerve Growth Factor; Neural Crest.

Further Reading

Ernsberger U (2001) The development of postganglionic sympathetic neurons: Coordinating neuronal differentiation and diversification. *Autonomic Neuroscience* 94: 1–13.

Francis NJ and Landis SC (1999) Cellular and molecular determinants of sympathetic neuron development. *Annual Review of Neuroscience* 22: 541–566.

Glebova NO and Ginty DD (2005) Growth and survival signals controlling sympathetic nervous system development. *Annual Review of Neuroscience* 28: 191–222.

Honma Y, Araki T, Gianino S, et al. (2002) Artemin is a vascular-derived neurotropic factor for developing sympathetic neurons. *Neuron* 35: 267–282.

Howard MJ (2005) Mechanisms and perspectives on differentiation of autonomic neurons. *Developmental Biology* 277: 271–286.

Huber K (2006) The sympathoadrenal cell lineage: Specification, diversification, and new perspectives. *Developmental Biology* 298: 335–343.

Hyatt-Sachs H, Schreiber RC, Bennett TA, et al. (1993) Phenotypic plasticity in adult sympathetic ganglia *in vivo*: Effects of deafferentation and axotomy on the expression of vasoactive intestinal peptide. *Journal of Neuroscience* 13: 1642–1653.

Keast JR (2006) Plasticity of pelvic autonomic ganglia and urogenital innervation. *International Review of Cytology* 248: 141–208.

LeDourin NM, Creuzet S, Couly G, et al. (2004) Neural crest cell plasticity and its limits. *Development* 131: 4637–4650.

Mione MC, Cavanagh JF, Lincoln J, et al. (1990) Long-term chemical sympathectomy leads to an increase of neuropeptide Y immunoreactivity in cerebrovascular nerves and iris of the developing rat. *Neuroscience* 34: 369–378.

Morrison SJ, White PM, Zock C, et al. (1999) Prospective identification, isolation by flow cytometry, and in vivo self-renewal of multipotent mammalian neural crest stem cells. *Cell* 96: 737–749.

Smith PG, Fan Q, and Zhang R (1999) Divergence of smooth muscle target and sympathetic pathway cell phenotypes in the orbit of the developing rat. *Journal of Comparative Neurology* 408: 352–364.

Young HM, Anderson RB, and Anderson CR (2004) Guidance cues involved in the development of the peripheral autonomic nervous system. *Autonomic Neuroscience* 112: 1–14.

Autophagy and Neuronal Death

P G H Clarke, University of Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Autophagy and Its Role in the Mediation of Autophagic Cell Death

Autophagy

Although this article is primarily concerned with neurons, and the implication of autophagy in their death, it will be necessary to draw general principles from studies on other cell types, because autophagy is a general phenomenon occurring in virtually all types of cell, and the most convincing molecular analyses of its role in cell death have been done in nonneuronal cell lines.

Autophagy is the mechanism by which cells degrade parts of their own cytoplasm using the lysosomal machinery. There are several types of autophagy, including microautophagy, the direct capture of tiny portions of cytosol by invagination of lysosomal membranes; chaperone-mediated autophagy, a specific mechanism for degrading cytosolic proteins containing a particular pentapeptide consensus motif; pexophagy, the specific autophagocytosis of peroxisomes; and macroautophagy, which involves the engulfment of sizeable regions of cytoplasm, including organelles, in double-membrane vesicles called autophagosomes. Macroautophagy is the best-studied type of autophagy, and the only type that has been studied in detail in relation to cell death. This article will therefore deal primarily with macroautophagy.

Macroautophagy (Figure 1) is initiated by the formation of autophagosomes (about 400–800 nm in diameter) from cup-shaped double-membranous structures called isolation membranes or phagophores, which engulf cytosolic components, including organelles. The isolation membrane then closes to form the autophagosome. The origin of the isolation membrane is still a matter of debate. There is evidence that it may arise from various sources including smooth endoplasmic reticulum and the *trans*-Golgi network, but recent research on yeast indicates that a major source of its membrane is an independent punctate structure called the 'pre-autophagosomal structure.' The autophagosome fuses with a lysosome to form an autolysosome, where the enclosed material is broken down. The term 'autophagic vacuole' includes both autophagosomes and autolysosomes.

Autophagy (including macroautophagy) is involved in the normal turnover of cell contents and is enhanced by cellular stresses, against which it provides protection, for example, by replenishing the pool of free amino acids in the case of amino acid depletion, or by eliminating damaged proteins. Also, by reducing the size of stressed cells, autophagy reduces their metabolic burden. Thus, in many situations, autophagy promotes the health and survival of cells.

Autophagic Cell Death: Origins of the Concept

Despite the life-promoting roles of autophagy, macroautophagy has also been associated with cell death, and the term 'autophagic (or type 2) cell death' is used as a morphological classification for dying cells containing numerous autolysosomes. This occurs frequently during embryonic development and is probably the most common type of cell death in insect metamorphosis, but (macro)autophagic features are also associated with many cases of pathological cell death including heart failure, excitotoxicity, and neurodegenerative diseases.

Historically, the development of electron microscopy permitted the discovery of (macro)autophagy in the early 1960s, and this was soon followed by numerous ultrastructural studies from the mid-1960s onward, showing an abundance of autolysosomes in dying cells in many situations, including most cases of metamorphosis. Nevertheless, even as late as the 1990s, only a few authors considered that the autophagy was instrumental in the cell death.

The reasons for this reluctance were multiple. One was that autophagy had from the moment of its discovery been understood to play physiological roles in healthy cells, for example, the provision of breakdown products for reuse and the elimination of abnormal proteins, and several authors interpreted its presence in the dying neurons to reflect an unsuccessful survival-promoting mechanism for eliminating damaged regions of cytoplasm. Many other authors were influenced by the (then) widely accepted 'suicide bag hypothesis' of De Duve, discoverer of the lysosome, according to which cell death is achieved by the release of hydrolases from the lysosomes; the status of this hypothesis is still controversial. Then, as the suicide bag hypothesis gradually fell out of favor in the 1970s and 1980s, the simultaneous rise in popularity of a somewhat rigid dichotomy according to which all cell death had to be apoptosis or necrosis did not encourage openness to alternative mechanisms of cell death. Indeed, proponents of the apoptosis–necrosis dichotomy maintained that

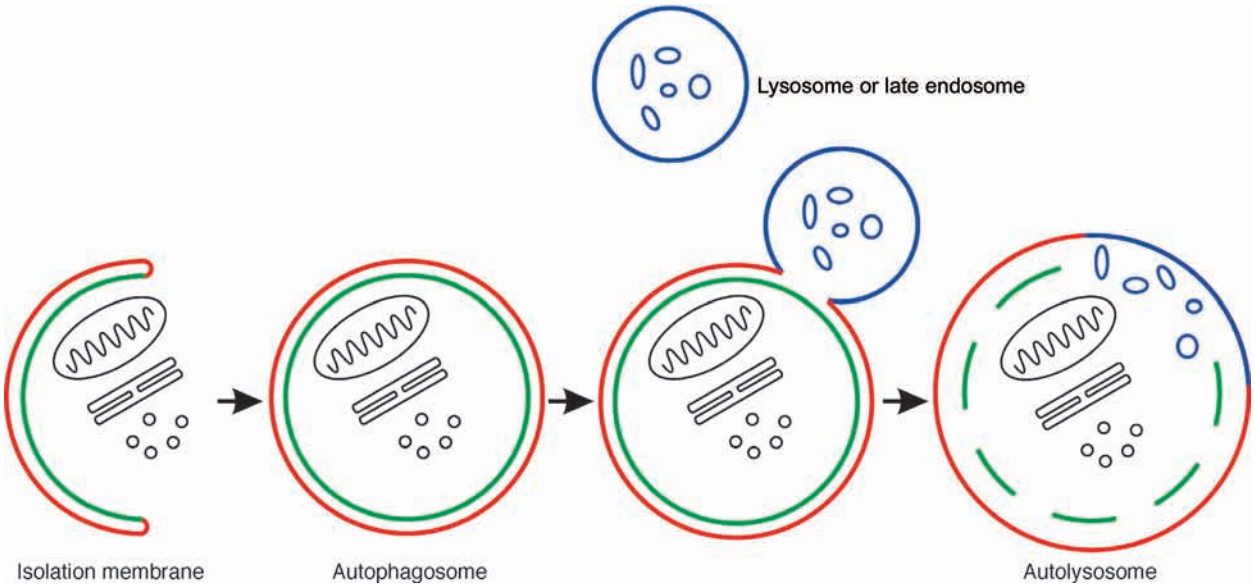

Figure 1 Macroautophagy. An isolation membrane (formed from smooth endoplasmic reticulum, the *trans*-Golgi network, or the 'pre-autophagosomal structure') enwraps cytosolic components and closes to form the two-membraned autophagosome, which then fuses with a lysosome or late endosome to give an autolysosome, in which the engulfed contents and the inner autophagosomal membrane (shown in green) are degraded.

autophagic dying cells were in fact undergoing apoptosis and that the autolysosomes were either a protective reaction or an irrelevant epiphenomenon. And, finally, it has to be admitted that a death-mediating role for the autophagy had not been proved, and in several cases very strong autophagy can occur without neuronal death.

The idea of autophagy-mediated cell death was, however, supported in the 1980s by experiments on neuronal death in the target-deprived isthmo-optic nucleus in chick embryos. (The isthmo-optic nucleus is the source of efferents from the brain to the retina in birds.) This neuronal death was characterized by abundant autolysosomes that ultimately filled most of the cytoplasm, and also by the loss of DNA from the nucleus to neighboring lysosomes. The fact that a cell's own DNA was being degraded by autophagy went against the view that the autophagy was a survival-promoting reaction to cellular stress.

Prevention of Autophagic Cell Death by Pharmacological Inhibitors of Autophagy

Nevertheless, a death-promoting role for autophagy gained only limited acceptance until it could be proved that inhibiting it prevented cell death. Initial evidence for this was provided in the 1990s by the death-preventing effects of 3-methyladenine (3-MA), an inhibitor of the formation of autophagic vacuoles that has been described as 'specific' but only in the limited sense that it does not alter the overall level of protein synthesis. Sandvig and van Beurs first showed, in 1992, that cell death, in this case toxin induced, could be prevented by 10 mM 3-MA. Subsequently, similar doses of 3-MA were shown to prevent (partially or completely) or delay cell death with autophagic characteristics in many situations including sympathetic neurons deprived of nerve growth factor, telencephalic neurons exposed to chloroquine, and cerebellar granule neurons deprived of serum and potassium. In all cases, the dying cells were shown to contain numerous autophagic vacuoles, and their rescue by 3-MA was accompanied by a reduction in their content of autophagic vacuoles. The suppression by 3-MA of autophagy is probably due to its inhibition of class III phosphatidylinositol 3-kinase (PI3-K), but it was uncertain whether this is also the basis of its protection against autophagic cell death, because its pharmacological profile is poorly characterized and it probably affects other enzymes. It was therefore important to test whether better-characterized inhibitors of PI3-K (LY294002 and wortmannin) could have similar protective effects. In many situations, these inhibitors are proapoptotic, because they inhibit the powerfully protective class I PI3-K pathway, so a protective effect due to inhibition of class III PI3-K (and hence autophagy) can easily be masked; but in serum-deprived PC12 cells, LY294002, wortmannin, and 3-MA have all been shown to be protective, apparently through the blockade of autophagy.

Prevention of Autophagic Cell Death by Interference with Autophagy Genes

However, even the better-characterized PI3-K inhibitors affect other cellular processes as well as autophagy, and definitive evidence for the death-mediating role of autophagy was provided only recently, by studies involving RNA interference of specific autophagy genes.

Our understanding of the control mechanisms of autophagy (particularly macroautophagy) depends to a great extent on intensive studies on autophagy in yeast, where about 30 genes controlling the initiation and execution of autophagy have been identified during the last decade. Until recently, these were grouped into three main gene families (apg, aut, and cvt), according to the genetic screens in which they were detected, but the functional distinctions between these families do not appear to be very clear-cut, and in the current terminology all the genes are grouped into the single 'atg' (autophagy gene) family (atg1–atg29 at the time of writing, but new members continue to be discovered). A detailed description of how these genes control autophagy would be beyond the scope of this chapter, but it is highly relevant to our present concerns that many of the yeast genes have vertebrate (including mammalian) homologs, and that certain of them, including atg5, atg6 (beclin 1), and atg7, are essential for the formation of autophagosomes.

This fact was used in two key papers in 2004 in which macroautophagy was blocked by RNA interference of atg5, atg6 (beclin 1), and atg7 (as well as by pharmacological inhibitors of autophagy) in cell lines whose apoptotic machinery had been deactivated genetically or pharmacologically. In both papers, pure autophagic cell death occurred and both the autophagy and the accompanying cell death were prevented by the RNA interference (and by the pharmacological inhibitors). Although a role for the autophagy genes in processes other than autophagy cannot be entirely ruled out, the fact that silencing each of the three genes prevented the autophagic cell death is strong evidence that the (macro)autophagy is not merely an epiphenomenon, or a defensive reaction, but is actually involved in mediating the cell death.

The Importance of Autophagy-Mediated Cell Death in Relation to Apoptosis

Autophagic cell death, as judged morphologically, seems to be the commonest type of cell death in physiological situations of massive cell death leading to the destruction of a tissue, as in many cases of metamorphosis and in some radical cases of mammalian embryonic tissue remodeling, whereas apoptosis appears to be the usual mechanism where sporadic dying cells occur in a tissue destined to survive. Thus, if autophagy could be assumed to mediate cell death in all cases of morphologically identified autophagic cell death, one could conclude that the autophagic death mechanism was of almost equal importance to the apoptotic mechanism.

Unfortunately, this is currently uncertain. While the reliability of 3-MA in protecting against many different cases of autophagic cell death does suggest that the autophagic death mechanism is of widespread importance, the more specific studies with RNA interference (or antisense) are still few in number, and situations have been reported in which massive autophagy can occur in cells without them dying. Moreover, there is evidence that a lysosomal, presumably autophagic, mechanism can initiate caspase activation and apoptosis. This is clearly different from autophagic cell death, which in many cases has been shown to be caspase independent, but does mean that morphological evidence for autophagy cannot be taken as proof of autophagy-mediated cell death. Thus, although the existence of an autophagic death mechanism is now difficult to deny, its generality and importance are still matters of debate.

Indeed, it has recently been argued that autophagy may mediate cell death only in very artificial situations where apoptosis has been deactivated (as in the two RNA interference studies mentioned above). Even if this were true, it would not detract from the importance of autophagic cell death in many pathological situations, where apoptosis may indeed have been deactivated either genetically (as in cancers that have become resistant to apoptosis-promoting chemotherapy) or pharmacologically (as in anticaspase neuroprotective protocols). But it has recently been shown that downregulation of atg5 by antisense technology protected against interferon-γ-induced autophagic cell death in HeLa cells whose apoptotic machinery had not been inhibited. Moreover, pharmacological blockade of autophagy by inhibition of PI3-kinase actually enhances the apoptotic machinery by increasing caspase-3 activation, but it can still prevent or delay cell death.

Thus, the autophagic death mechanism can be effective without the artificial deactivation of apoptosis, but its generality and importance are still not entirely clear.

Autophagy in Neuronal Death and Neurodegeneration

Autophagy in Neuronal Death

Although our mechanistic understanding of autophagic cell death has come largely from studies of

Table 1 Reports of neuroprotection by 3-MA

Situation	*Autophagic morphology?*	*Protection by 3-MA?*	*Caspase-dependent?*	*Reference*
Cerebellar granule neurons deprived of serum and potassium	Yes	Partial protection	Partially	Kaasik et al. (2005)
Cerebellar granule neurons in low potassium	Yes	Yes	Yes	Canu et al. (2005)
Telencephalic neuronal cultures exposed to chloroquine	Yes	Yes	No	Zaidi et al. (2001)
Sympathetic neurons treated with cytosine arabinoside	Yes	Cell death delayed	No	Xue et al. (1999)
Sympathetic neurons deprived of NGF	Yes	Cell death delayed	No	Xue et al. (1999)

nonneuronal cells, there is considerable morphological evidence for autophagic 'neuronal' death in all the main situations where neurons die: in natural development, in various pathological situations, and in experimental models, as is discussed below. In addition, there are a few studies showing the prevention of autophagic neuronal death by 3-MA (see Table 1).

Autophagic Neuronal Death during Development

Reports of autophagic neuronal death occurring naturally during development are relatively few, and most concerned anuran metamorphosis, including the death of the Rohon–Beard neurons, a transient population of sensory neurons that undergoes 100% cell death. In mammals, one is able to find only one relevant report; it concerned autophagic neuronal death in the developing cerebral cortex. This paucity of reports suggests that autophagic cell death plays only a relatively minor role in naturally occurring neuronal death in mammals (and other higher vertebrates). This fits with the generalization made above, that autophagic cell death occurs most commonly in physiological situations of massive cell death leading to the destruction of a tissue. However, caution is required, because in many studies isolated autophagic dying cells may have been mistaken for phagocytes, which they resemble morphologically and in their expression of autophagic markers.

Failure in competition for retrograde neurotrophic support is believed to be a major cause of naturally occurring neuronal death, and numerous studies of neuronal death in development have involved axotomy or other means of depriving neurons of retrograde support. In some cases, the resulting neuronal death was autophagic, but in many others it was clearly not. The reasons for the differences are unclear, but one factor may be the developmental stage. This was first indicated by an elegant study by Decker in 1978 on motor neuronal death in larval frogs. He found that very early axotomy caused a pyknotic (apoptosis-like) morphology, whereas very late axotomy caused classic chromatolysis. But axotomy at an intermediate stage caused the 'genesis of numerous secondary lysosomes in degenerating cells' – in other words, cell death with an autophagic morphology.

Studies on the isthmo-optic nucleus of chick embryos showed an age dependence that was similar to the above but not quite so clear cut. Early deprivation of retrograde support by blocking axonal transport in the isthmo-optic axons led to isthmo-optic neuronal death with a mixed morphology that was both pyknotic and autophagic, whereas later transport blockade caused a purer form of autophagic cell death with only minimal pyknosis (Figure 2). This neuronal death was also characterized by strong endocytic activity, a phenomenon that has since been observed in several subsequent studies of stressed, but not necessarily dying, neurons. Isthmo-optic neuron death could also be provoked by de-afferentation, but this caused no signs of autophagy, and when combined with blockade of retrograde support it decreased the autophagic characteristics of the dying neurons.

Neuronal Autophagy in Acute Neurological Conditions

The neuronal cell death in virtually all acute neurological conditions (e.g., stroke, traumatic brain injury, and neonatal asphyxia) shares a common mechanism: excitotoxicity, excessive depolarization that is usually due to the excessive activation of glutamate receptors, especially the *N*-methyl-D-aspartate (NMDA) subtype. Excitotoxic neuronal death is generally considered to be necrosis (most commonly) or apoptosis or a combination of the two, and, until recently, the presence of enhanced autophagy in these conditions was largely ignored. However, over the last few years, morphological evidence for intense autophagy and an increase in the autophagosomal marker LC3-II have been reported in several experimental models of cerebral hypoxia–ischemia, and an increase in the autophagy gene beclin 1 has been reported in a model of traumatic brain injury. NMDA receptor activation has likewise

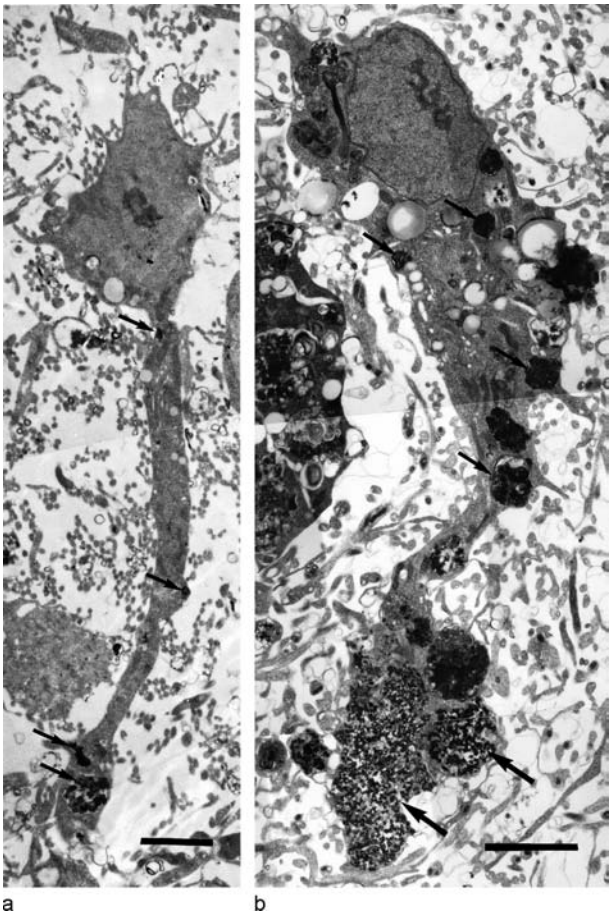

Figure 2 Electron micrographs of two autophagic dying neurons in the isthmo-optic nucleus of 14-day-old chick embryos. The death of both neurons was provoked by an injection of colchicine in their axonal target territory, which blocked axoplasmic transport, depriving them of retrograde support. (a) Early stage of neuronal death. Several vacuoles are found in the cytoplasm of the soma and main dendrite. The nucleoplasm and cytoplasm are somewhat denser than in a healthy neuron. (b) Advanced stage of neuronal death. The vacuoles are larger and more numerous, many containing membranous whorls. The arrows indicate vacuoles labeled with intravascularly injected horseradish peroxidase (in this and several other cases autophagic dying neurons have been shown to be strongly endocytic). Scale bar = 2 μm. From Hornung JP, Koppel H, and Clarke PGH (1989) Endocytosis and autophagy in dying neurons: An ultrastructural study in chick embryos. *Journal of Comparative Neurology* 283: 425–437.

been shown to induce autophagic neuronal death, in organotypic hippocampal cultures. This neuronal death was also characterized by strong endocytosis of exogenous horseradish peroxidase (Figure 3). However, it is currently unknown whether the autophagy in acute neurological conditions and excitotoxicity mediates cell death.

Autophagy in Neurodegenerative Diseases

In contrast to acute neurological conditions, neurodegenerative diseases involve progressive neuronal degeneration over periods of many months or years. Changes in the endosomal–lysosomal system, including increased macroautophagy, have been reported in virtually all neurodegenerative diseases including Alzheimer's, Huntington's, and Parkinson's diseases, prion diseases, and amyotrophic lateral sclerosis. The causes and roles of the increased macroautophagy are difficult to establish in human diseases, but additional information from experimental models provides some preliminary hypotheses. From models of Alzheimer's, Huntington's, and Parkinson's diseases, there is evidence that the macroautophagy may in many cases be involved in clearing protein aggregates from affected neurons, and hence be protective, but may also lead to autophagic neuronal death.

In Huntington's disease, the autophagy seems to be primarily protective. This disease involves massive neuronal death in the striatum as a result of the presence of an expanded polyglutamine repeat in the Huntington gene product. The dying neurons have a strongly autophagic morphology, and the autophagy appears to be a defense mechanism because the experimental enhancement of autophagy in fly and mouse models of Huntington's disease reduces the accumulation of polyglutamines as well as the neuronal death, whereas inhibition of autophagy has the opposite effect on both.

In Parkinson's disease, the situation is more ambiguous. The best-known neuropathological characteristics of this disease are the degeneration of dopaminergic neurons of the substantia nigra, and the presence of cytoplasmic inclusions called Lewy bodies in these neurons before they die. Lewy bodies contain ubiquitinated aggregates of α-synuclein and other proteins. There are reports that this neuronal death can have an autophagic morphology. Some cases of early-onset Parkinson's disease involve a mutation in the α-synuclein gene. In cultured PC12 cells, overexpression of mutant but not wild-type α-synuclein causes an impairment in the ubiquitin–proteasome system and the presence of ubiquitinated protein aggregates, an accumulation of autophagic vacuoles, and increased nonapoptotic autophagic cell death. Thus, although the increased autophagy may be an attempt to protect the cells by clearing the protein aggregates, it may also be involved in mediating the neuronal death.

Alzheimer's disease is characterized by the presence of (extracellular) β-amyloid plaques and filamentous tangles, primarily in the hippocampus and cerebral cortex. Both are currently believed to be involved in the degenerative changes in these brain regions. Pronounced macroautophagy has been demonstrated in the affected neurons, and β-amyloid has been shown to be generated by the proteolytic cleavage of β-amyloid precursor protein. In a mouse model of the

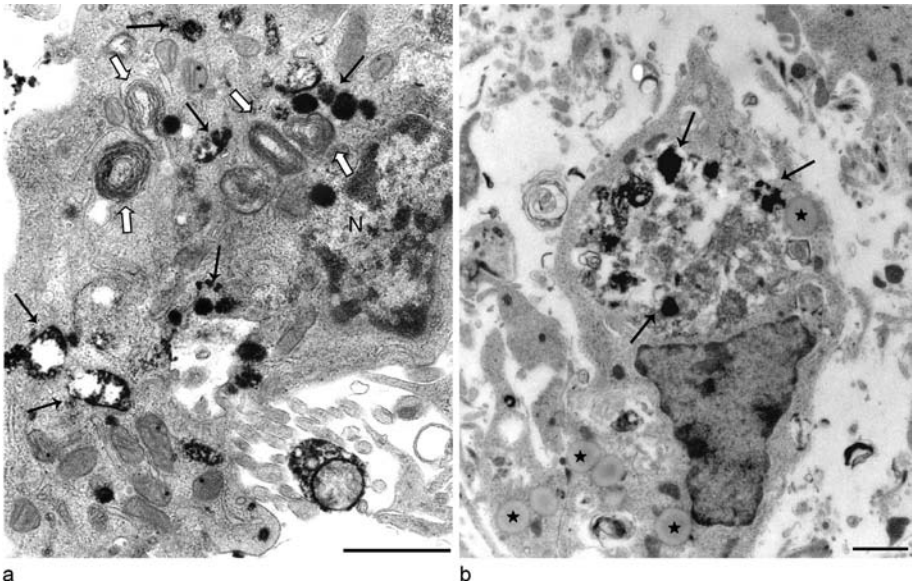

Figure 3 Autophagic degenerating neurons in the CA1 region of organotypic hippocampal cultures after 2 h (a) or 8 h (b) of exposure to 100 μM NMDA. (a) Thick white arrows, membranous whorls (autolysosomes); black arrows, endosomes labeled with horseradish peroxidase; N, nucleus. (b) Clumps of peroxide labeling within a large vacuole; stars, putative unlabeled endosomes. Scale bar = 1 μM. Reproduced from Borsello T, Croquelois K, Hornung JP, and Clarke PGH (2003) *N*-methyl-D-aspartate-triggered neuronal death in organotypic hippocampal cultures is endocytic, autophagic and mediated by the c-Jun N-terminal kinase pathway. *European Journal of Neuroscience* 18: 473–485, with permission from Blackwell Publishing.

disease, a similar neuronal macroautophagy occurs, and this happens rather early, before the extracellular β-amyloid deposits, but the maturation of autophagosomes to autolysosomes appears to be impaired. At later stages, there is a further accumulation of autophagosomes, and these are rich in β-amyloid. Inducing or inhibiting macroautophagy elicits parallel changes in macroautophagy and β-amyloid production, suggesting that in this case the macrophagy may contribute to the disease process, but not necessarily through autophagic cell death.

Neuronal Autophagy in Lysosomal Storage Diseases

Lysosomal storage diseases are caused by mutations in the genes encoding various lysosomal hydrolases, leading to the accumulation (or 'storage') of partially digested substances in lysosomes. Different lysosomal storage diseases cause degenerative and other changes in different organs of the body, including in some cases the brain (e.g., in Tay–Sachs disease and Niemann–Pick C disease). Whereas most neurodegenerative diseases involve increased lysosomal digestion, lysosomal storage diseases are caused by a 'decrease' in one particular component of lysosomal digestion, but this can lead to complex changes in many different cellular signaling pathways. Since the genetic mutation directly affects the lysosomal system, autophagic digestion must presumably be affected. There have been few studies of autophagy in neuronal death in these diseases, but in a mouse model of Niemann–Pick C disease there was massive degeneration of cerebellar Purkinje cells, which had features consistent with autophagic cell death.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Epilepsy: Neuronal Death; Huntington's Disease: Neurodegeneration; Intracellular Calcium and Neuronal Death; Neuroprotection: Pharmacological Approaches; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Programmed Cell Death.

Further Reading

Borsello T, Croquelois K, Hornung JP, and Clarke PGH (2003) *N*-methyl-D-aspartate-triggered neuronal death in organotypic hippocampal cultures is endocytic, autophagic and mediated by the c-Jun N-terminal kinase pathway. *European Journal of Neuroscience* 18: 473–485.

Bursch W and Ellinger A (2005) Autophagy – a basic mechanism and a potential role for neurodegeneration. *Folia Neuropathologica* 43: 297–310.

Canu N, Tufi R, Serafino AL, Amadoro G, Ciotti MT, and Calissano P (2005) Role of the autophagic-lysosomal system on low potassium-induced apoptosis in cultured cerebellar granule cells. *Journal of Neurochemistry* 92(5): 1228–1242.

Clarke PGH (1990) Developmental cell death: Morphological diversity and multiple mechanisms. *Anatomy and Embryology* 181: 195–213.

Hornung JP, Koppel H, and Clarke PGH (1989) Endocytosis and autophagy in dying neurons: An ultrastructural study in chick embryos. *Journal of Comparative Neurology* 283: 425–437.

Kaasik A, Rikk T, Piirsoo A, Zharkovsky T, and Zharkovsky A (2005) Up-regulation of lysosomal cathepsin L and autophagy during neuronal death induced by reduced serum and potassium. *European Journal of Neuroscience* 22(5): 1023–1031.

Larsen KE and Sulzer D (2002) Autophagy in neurons: A review. *Histology and Histopathology* 17: 897–908.

Nixon RA, Wegiel J, Kumar A, et al. (2005) Extensive involvement of autophagy in Alzheimer disease: An immuno-electron microscopy study. *Journal of Neuropathology and Experimental Neurology* 64: 113–122.

Shimizu S, Kanaseki T, Mizushima N, et al. (2004) Role of Bcl-2 family proteins in a non-apoptotic programmed cell death dependent on autophagy genes. *Nature Cell Biology* 6: 1221–1228.

Stefanis L, Larsen KE, Rideout HJ, Sulzer D, and Greene LA (2001) Expression of A53T mutant but not wild-type alpha-synuclein in PC12 cells induces alterations of the ubiquitin-dependent degradation system, loss of dopamine release, and autophagic cell death. *Journal of Neuroscience* 21: 9549–9560.

Xue L, Fletcher GC, and Tolkovsky AM (1999) Autophagy is activated by apoptotic signalling in sympathetic neurons: An alternative mechanism of death execution. *Molecular and Cellular Neuroscience* 14(3): 180–198.

Yu L, Alva A, Su H, et al. (2004) Regulation of an ATG7-beclin 1 program of autophagic cell death by caspase-8. *Science* 304: 1500–1502.

Yu WH, Cuervo AM, Kumar A, et al. (2005) Macroautophagy – a novel beta-amyloid peptide-generating pathway activated in Alzheimer's disease. *Journal of Cell Biology* 171: 87–98.

Zaidi AU, McDonough JS, Klocke BJ, et al. (2001) Chloroquine-induced neuronal cell death is p53 and Bcl-2 family-dependent but caspase-independent. *Journal of Neuropathology and Experimental Neurology* 60(10): 937–945.

Aversive Emotions: Genetic Mechanisms of Serotonin

S E Ahmari, M D Alter, and R Hen, Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Anxiety and fear responses are necessary components of adaptive behavior. Fear of predators and normal vigilant behavior are essential for survival in the wild. It is when these responses are uncontrollable that they stop being protective, instead becoming maladaptive and pathological, ultimately impairing an individual's ability to function in the world.

Though a complete understanding of the neurobiological mechanisms underlying normal and pathological fear and anxiety remains a future goal and motivator for research, a role for serotonin, also known as 5-hydroxytryptamine (5-HT), in these emotions has been firmly established. Both clinical evidence, ranging from empirical data (serotonin reuptake inhibitors have been found effective in treating a wide range of anxiety disorders; see Table 1) to human genetic studies, and animal experiments have supported a role for this neurotransmitter in both normal anxiety and pathological disease states. For example, there is evidence from human linkage analyses implicating monoamine oxidase A (MAO-A; a molecule which plays a key role in the degradation of serotonin) in the pathophysiology of panic disorder, and also suggesting a role for the serotonin transporter and two serotonin receptors (5-HT1B and 5-HT2A) in obsessive–compulsive disorder (OCD). It is likewise clear that neurotransmitters other than serotonin play a role in the development of anxiety and fear responses. However, though the story is ultimately likely to be very complex, with particular neurotransmitters and receptors being important at different times in development and with gene–environment interactions and epigenetics playing important roles in the ultimate determination of behavioral phenotypes, focusing on the role of serotonin will serve as a starting point for discussion. One way of shedding light on the role of this neurotransmitter in fear and anxiety is through examination of the combination of the effects of serotonin-related genetic factors and early life stress on development and the impact of these factors on the development of anxiety disorders in later life.

Relationship between Fear and Anxiety

Though there is a close relationship between fear and anxiety, and animal models of fear conditioning have allowed us to gain important insights into the neural circuits underlying anxiety, it is apparent that there is a distinction between these responses despite substantial overlap in involved brain regions. Past models have differentiated fear as being a response to real threats and anxiety as being an abnormal response to a perceived threat in the absence of immediate danger to the organism. Alternatively, one can understand fear as being stimulus specific (akin to 'phobia' in humans), whereas anxiety is more generalized. Despite this distinction, it seems apparent that there is substantial overlap in the neural circuitry of fear and anxiety. For example, rodent studies of fear conditioning have clearly demonstrated the involvement of cortical and thalamic projections to the lateral nucleus of the amygdala, where substantial intra-amygdala processing occurs. Projections then arise from the central nucleus of the amygdala to areas of the brain stem, hypothalamus, and hippocampus, leading to fear responses. However, there is evidence that different amygdala output tracts may differentially subserve fear versus anxiety responses. For example, by using bright light exposure and corticotropin-releasing hormone-enhanced startle paradigms to model anxiety in comparison to stimulus-specific fear conditioning, Davis and colleagues demonstrated differential involvement of the bed nucleus of the stria terminalis in anxiety versus the central nucleus of the amygdala in fear, despite very similar output projections to the hypothalamus and brain stem. Human neuroimaging studies of perception of fearful faces (trait anxiety) and fear conditioning have clearly demonstrated involvement of the amygdala, hippocampus, and thalamus; similarly, neuroimaging studies of anxiety disorders have also implicated the amygdala (post-traumatic stress disorder (PTSD), phobia, OCD, panic disorder), prefrontal cortex (PTSD, generalized anxiety disorder (GAD), OCD, phobia), hippocampus (PTSD, phobia, panic disorder), and striatum (OCD, phobia). The remainder of this article will focus on the relationship between serotonin and the development of anxiety responses in both animals and humans, with the understanding that there is likely a close relationship between anxiety and fear responses, particularly in disorders such as PTSD.

Early Life Stress and Mental Health

A multitude of research studies associate early life stress with the development of mental health problems. Morbidity and mortality from a wide array of medical and mental illnesses are increased in abused children. Even less severe forms of early-life

Table 1 Pharmacological dissection of anxiety disorders

Disorder	*BDZ*	*Buspirone*	*IMI*	*CMI*	*MAOI*	*SSRI*
GAD	+	+	+	+	0	+
Social phobia	+	(+)	0	(+)	+	+
OCD	0	(+)	(+)	+++	(+)	+++
Panic disorder	+	0	+	+++	+	+

GAD, generalized anxiety disorder; OCD, obsessive–compulsive disorder; BDZ, benzodiazepines, compounds with activity at the γ-aminobutyric acid (GABA) receptor; buspirone, partial 5-hydroxytryptamine (5-HT) 1A agonist (partial agonist at a particular serotonin receptor); IMI, imipramine, a tricyclic antidepressant without activity at serotonin receptors; CMI, clomipramine, a tricyclic antidepressant with significant activity at serotonin receptors; MAOI, monoamine oxidase (MAO) inhibitors, compounds which inhibit MAO and therefore inhibit the breakdown of serotonin; SSRI, selective serotonin reuptake inhibitors, which prevent reuptake of serotonin into neurons, leading to increased concentrations in the synaptic cleft and enhanced postsynaptic activation of serotonin receptors; +, effective anxiolysis; (+), small or discrepant effects; +++, extensive anxiolysis; 0, no effect. Adapted from McNaughton N and Corr PJ (2004) A two-dimensional neuropsychology of defense: Fear/anxiety and defensive distance. *Neuroscience and Biobehavioral Reviews* 28(3): 285–305, with permission from Elsevier.

stress have been associated with an increase in adverse outcomes. For example, children brought up in families characterized by a lack of parental warmth or either over- or underregulation of children's behavior are more likely to have a host of medical or psychiatric problems. One could speculate that such children might start to have aberrations in the regulation of normal anxiety and fear responses early in development, which ultimately puts them at greater risk for development of disordered mood in later life.

Still, not all children exposed to these stressful environments have adverse outcomes. A particularly interesting line of research has demonstrated that at least a portion of these effects is mediated by genetic variability of the serotonin system. In a landmark study, Caspi and colleagues demonstrated that boys with a hypofunctioning polymorphism in the MAO-A gene are at increased risk for adolescent and adult antisocial behavior if they were abused during childhood. Since MAO-A is one of the primary enzymes responsible for the breakdown of serotonin in the central nervous system, this finding demonstrates a potential interaction between serotonin genetics and adolescent and adult psychopathology. Similarly, in a separate study in the same prospective cohort but utilizing both boys and girls, individuals homozygous (s/s) or heterozygous (l/s) for the short allele of a hypofunctioning polymorphism in the serotonin transporter (SERT) gene were found to have a markedly increased risk for depression if subjected to child abuse, but individuals homozygous for the normal functioning long allele (l/l) had no increase in depression when exposed to abuse. Many studies have since examined the relationship between SERT and MAO polymorphisms and various forms of psychopathology, demonstrating associations in OCD, panic disorder, specific and social phobia, GAD, and anxious personality traits.

When examining the multiple studies demonstrating the detrimental effects of early life stress and genetic variance on mental health, the logical next question is, how are these detrimental effects being mediated? An attractive hypothesis is that early life represents a particularly sensitive period for the development of neural circuits and that disruption of the normal processes of establishing these circuits makes a person more vulnerable to the development of psychopathology.

Sensitive Periods of Development: A Role for Serotonin?

A sensitive period is a general term that applies whenever experience has a particularly strong and lasting effect over a limited time (Figure 1). A classic example is that of filial imprinting, which occurs shortly after birth as a young animal learns to recognize and bond with its parent. An important characteristic of sensitive periods is that the effects of experience are not readily amenable to change after the sensitive period has occurred. For example, once an animal has undergone imprinting, the attachment to the parent will be strong and persistent; moreover, attachment to an alternative parental figure will be much more difficult to achieve. This is thought to occur because, though sensitive periods are marked by behaviors such as attachment to a parent or speaking a language, they are actually a property of underlying neural circuits. The properties of the neural circuits are shaped by experience during the sensitive period, and once formed, they may be modulated but do not easily change their basic structure. Furthermore, because behaviors, complex behaviors like language or anxiety responses in particular, represent the coordination of many neural circuits, the true effect of the persistence of circuit-level changes occurring during a sensitive period will tend to be underestimated.

Sensitive periods are of particular interest to researchers and clinicians because they represent periods in development during which certain capacities are readily shaped or altered by experience and therefore may serve as key points during which interventions may have a lasting impact on the development of normal versus abnormal brain circuits.

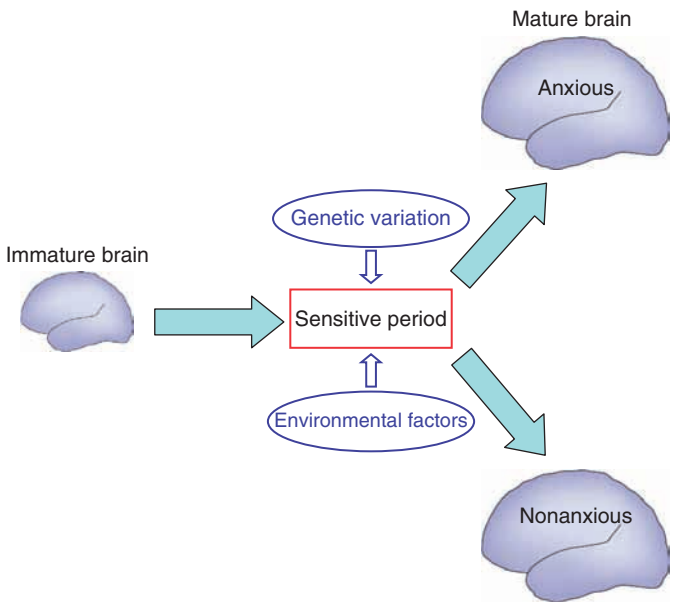

Figure 1 Sensitive period. The immature brain goes through a sensitive period of development during which it is particularly sensitive to environmental and genetic factors. Changes that occur during this period are persistent and are reflected as differences in adult behavior.

Research on the serotonin system underscores an important role of serotonin in the development of anxiety. Many regions innervated by the serotonin system, including the amygdala, hippocampus, anterior and posterior cingulate, and prefrontal cortex, have been suggested to play a role in anxiety and anxiety-like behavior. Studying the genetics of the serotonin system has helped to outline circuits important in anxiety and to further suggest that serotonin affects these circuits during sensitive periods of their development. This work has repeatedly pointed to early life as a sensitive period of development and the amygdala and hippocampus as regions of importance. Together, these lines of research are lending insight into how a molecule traditionally believed to be involved only in synaptic transmission can have multiple roles in shaping brain development and complex behavioral responses.

Serotonin Has Multiple Roles

Initially isolated from serum in 1949 and detected in the mammalian brain in 1953, serotonin is traditionally viewed as a modulatory neurotransmitter. Serotonergic neurons cluster in the brain stem, primarily in the dorsal and median raphe nuclei, and send projections throughout the brain and spinal cord. Dorsal raphe neurons preferentially send projections to the basal ganglia (striatum, substantia nigra); in addition, they typically have small fusiform boutons, diffusely branching fibers, and likely more release of extrasynaptic serotonin. In contrast, neurons in the median raphe nuclei mainly project to components of the limbic system (hippocampus, amygdala) and have large, spherical varicosities with extensive repeated synapses. Notably, there is significant overlap in projection areas from these two sets of nuclei, and single serotonergic neurons have been found to innervate multiple regions in a synaptically connected circuit.

Through its diverse projections, the serotonin system plays a role in modulating sleep, appetite, memory/cognitive function, impulsivity, sexual behavior, motor function, and limbic/affective responsiveness. In addition to this traditional understanding of serotonin as a neuromodulator, increasing evidence demonstrates that neurotransmitters, including serotonin, have important roles in development. Serotonin in particular has been closely studied in this regard. Pharmacologic studies have demonstrated that serotonin can modulate a number of developmental events, including cell division, neuronal migration, cell differentiation, and synaptogenesis.

At first glance, it appears unlikely that one neurotransmitter could have such diverse modulatory functions and also play a role in the regulation of complex neural circuit formation. However, the serotonin system seems ideally suited for this position as it is itself quite complex and diverse; it therefore appears perfectly capable of carrying out multiple tasks. This complexity is clearly demonstrated by the multitude of receptors through which serotonin exerts its effects. Fifteen distinct genes encode for serotonin receptors; these are further classified into five families on the basis of their second messenger systems. Splice variants and posttranslational processing expand the

repertoire of viable receptor products to 30. In addition, receptors are located throughout the body and brain, and their expression is both temporally and spatially dynamic. Depending on the receptor type, its location, and the developmental stage of the organism, stimulation of serotonin receptors can be excitatory or inhibitory. Thus, combinations of variables can lead to vastly different downstream effects from this single neurotransmitter, which may be particularly important in the development of anxiety disorders, including abnormal fear responses.

Disruption of Serotonin Homeostasis during Development Leads to Anxiety

Adding further to the complexity of the serotonin receptor system, serotonin synthesis, reuptake, and degradation are further levels at which serotonin function is regulated. Disruption at each of these dynamically regulated levels of serotonin synthesis has been demonstrated to have long-standing effects on anxiety behavior. These effects appear to be mediated at least in part through changes that occur during development.

Excessive extracellular serotonin during development is associated with increased anxiety. Deletion of the MAO-A or SERT gene in mice leads to an accumulation of extracellular serotonin in the brain. MAO-A is the principal enzyme responsible for degradation of monoamines in early postnatal life. MAO-B compensates for the deletion later in life, but mice lacking MAO-A have a ninefold increase in brain serotonin levels during the first postnatal week. SERT is another molecule responsible for the removal of serotonin from the synaptic cleft. In the absence of SERT, serotonin continues to be released but is unable to be removed and accumulates in the synaptic cleft, leading to significant alterations in receptor binding and downstream effects. Knockout mice with deletions of either MAO-A or SERT genes have marked alterations in behavior characterized by increased anxiety and aggressiveness.

Despite this clear anxiety behavioral phenotype, putative anxiety circuits have not been carefully examined in these mice. Notably, SERT and MAO-A knockout mice have been found to have defects in the formation of brain circuits involved in sensory processing. It is interesting that both these mutations lead to abnormal development of thalamocortical axons important in the formation of barrel receptive fields and retinal axons important in visual processing. This abnormal development has been shown to be mediated by the serotonin 5-HT1B receptor. Double knockouts of MAO-A and 5-HT1B or triple knockouts of MAO-A, SERT, and 5-HT1B resulted in normal axonal development. It is tempting to speculate that excessive serotonin might mediate abnormal development of anxiety circuits in a manner similar to the serotonin-dependent disruption of thalamocortical and retinal circuits in these mice. Disruption of anxiety circuits during a developmental sensitive period would help to explain the persistence of anxiety-like behavior in these mice throughout their life span.

Serotonin levels which are abnormally low can also lead to the development of abnormal behaviors. For example, Pet-1 is a transcription factor that regulates transcription of several genes important in serotonergic function. In mice, knockout of Pet-1 leads to an 80% reduction in brain serotonin levels. It is interesting that although the levels of serotonin in these mice move in directions opposite those in the MAO-A and SERT knockout mice, at a behavioral level Pet-1 knockout mice resemble MAO-A and SERT knockout mice, with increased anxiety and aggressive behavior.

The recent identification of a neuronal-specific form of tryptophan hydroxylase provides another example of the effects of reduced serotonin in the developing brain. Functionally significant polymorphisms in the gene that encodes for this neuron-specific form, TPH2, have been found in humans and mice. In mice, the hypofunctioning allele is found in the BALB strain, which is known for its significantly increased baseline levels of anxiety. However, this allele is not found in the less anxious strain, C57. There is good evidence that these behavioral effects of decreased serotonin may be mediated through effects on developmental processes. For example, in rodents, pharmacologic depletion of serotonin in the mother or embryo results in alterations in neurogenesis, neuronal migration, and dendritic maturation.

Taken together, the evidence from mouse models shows that disruption of serotonergic homeostasis in either direction can have clear effects on anxiety behavior. There is also good evidence that serotonergic dysregulation can have developmental effects on circuit formation, neurogenesis, neuronal migration, and dendritic maturation. Putting these observations together lends credence to the idea that serotonin's effects on development, at least in part, underlie serotonin's effects on behavior.

Anxiety Phenotype Develops in Early Life

In the above models, it is not clear when particular gene deletions are mediating their effects. For example, germline deletions of MAO-A and SERT result in changes in serotonin levels throughout the life of an

animal, thus making it difficult to determine when these changes might be most influential. To address this issue of timing, researchers have used fluoxetine, a selective inhibitor of serotonin transporter activity, to mimic the effects of disrupting the SERT gene during a restricted time. In one particular study, newborn mice were given fluoxetine for only the first 3 weeks of life. In this way, serotonin levels were altered only during a limited time of postnatal development. Nonetheless, throughout the life span, these mice were behaviorally indistinguishable from the SERT knockout mice with respect to exploration in novel environments, one measure of anxiety behavior.

Mice deficient in the 5-HT1A receptor have been demonstrated to have increased anxiety-related behavior. In an elegant study to address the timing of this effect, Gross and colleagues showed that genetic replacement of the serotonin 5-HT1A receptor during only the first 3 weeks of life was able to completely rescue normal anxiety behavior in mice deficient in 5-HT1A. On the other hand, replacement of the 5-HT1A receptor throughout life with the exception of the first 3 weeks did not restore normal behavior. Mice deficient in 5-HT1A receptor during only the first 3 weeks of life continued to display the characteristic 5-HT1A knockout anxiety phenotype that is similar to the anxiety phenotype of MAO-A and SERT knockout mice. There is also substantial evidence that experience has potent effects during this same period. Rats exposed to high levels of maternal care during the first 3 weeks of life have decreased levels of anxiety relative to rats that receive low levels of maternal care. Similarly, lengthy maternal separation during the same period results in increased anxiety-related behavior.

In humans and other primates, this developmentally sensitive period appears to be somewhat longer, with research suggesting a time course beginning during the third trimester of pregnancy and extending to periadolescence. Nonetheless, as discussed previously, early life appears to be a particularly potent period for the effects of experience and genetic variation.

Early Life: What Is Happening at the Circuit Level?

It is important to consider what is occurring in the brain during the developmentally sensitive period of early life. Numerous studies have demonstrated that different brain regions develop and mature at different rates and different times. In a general sense, phylogenetically primitive regions such as the hippocampus or amygdala develop earlier than the phylogenetically more advanced regions such as the human frontal cortex. It is therefore important to consider not only timing but location in order to understand the effects of experience on development.

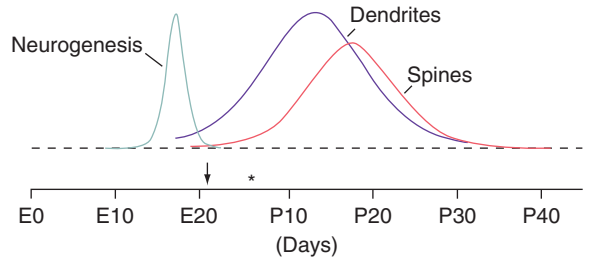

Figure 2 Development of rat hippocampus. Development of the rat hippocampus proceeds in a stepwise fashion in which late prenatal neurogenesis is followed by early postnatal dendritic outgrowth and spine development. These postnatal changes are coincident with the sensitive period for the development of an anxiety phenotype. E, embryonic; P, postnatal; *, birth. Adapted from Ben-Ari Y (2001) Developing networks play a similar melody. *Trends in Neurosciences* 24(6): 353–360, with permission from Elsevier.

The rat hippocampus has been particularly well studied in regard to its developmental time course, and as discussed previously, the hippocampus has also been well demonstrated to be involved in anxiety and anxiety-related behavior. Within the first 3 weeks, there is a dramatic rise and fall of dendritic growth and dendritic spine formation in the rat hippocampus that is fundamental to the formation of synapses and circuits (Figure 2). During this same time, other events occur that are likewise important for the development of brain circuits. In the rat hippocampus, there is a dramatic increase in the spontaneous synchronous firing of neurons that is a hallmark of developing neuronal networks. The timing of this spontaneous network activity directly precedes the emergence of hippocampal-dependent functions such as context-dependent learning.

In concordance with the broadened period of vulnerability, in humans these processes are proposed to occur over an extended period beginning earlier, in the third trimester of fetal development, and continuing later, throughout much of childhood. In summary, the period during which animals are most vulnerable to experience and genetic variability is the same period during which brain circuits important in anxiety are developing.

Anxiety Circuits Disrupted?

We have seen that early life represents a period of increased vulnerability to the effects of both experience and genetic variation. Furthermore, it is clear that this period is an important time in brain development and circuit formation. But what about the final piece of the hypothesis? Do changes in serotonin

function during early postnatal life affect circuit formation in the anxiety system? The developmentally sensitive period for the establishment of an anxiety phenotype that persists into adulthood is highly suggestive of an effect at the circuit level.

Examination of 5-HT1A knockout mice, which display more anxiety-related behaviors than do wild-type mice, lends support to this hypothesis. Lesion studies in these mice have specifically implicated the hippocampal circuit in anxiety-related behaviors. In humans, a functional polymorphism in the 5-HT1A gene promoter is associated with increased anxiety-related personality traits, as well as an increased incidence of a subtype of panic disorder characterized by the additional presence of agoraphobia.

Also in support of the hypothesis that changes in serotonin system gene expression can modulate anxiety and fear circuits, Hariri and colleagues have shown that in humans, a polymorphism in the SERT gene is associated with particularly robust changes in the sensitivity of the amygdala to fearful or threatening stimuli. Individuals with the hypo-functioning SERT allele (s/s) had a fivefold increase in amygdala activation in response to fearful stimuli. Another group of researchers also demonstrated this increased amygdala activation and further suggested changes in neural circuitry by demonstrating that individuals with at least one copy of the short allele had increased coupling of amygdala and prefrontal cortex activity. Because of the likely relation between fear and anxiety, this change in fear responsiveness may underlie an increased risk for anxiety disorders in these individuals. It is important to note that these differences in regional brain activation were much more readily apparent than were overt behavioral differences in these individuals, underscoring the point that a change in neural circuit function does not necessarily result in a behavioral phenotype in the case of an animal model, or in an anxiety disorder in the case of humans.

Sensitive Period Revisited: Gene–Environment Interactions

Current evidence suggests that anxiety circuits appear to be subject to a sensitive period of development. Gene knockout studies and phenotypic characterization of individuals with functional polymorphisms have highlighted several genes of the serotonin system that affect the developmental course of circuits involved in anxiety. Evidence for this is both direct, as with functional neuroimaging of individuals with SERT polymorphisms, and indirect, as evidenced by the lifelong persistence of anxiety phenotypes based on genetic variation and early experience.

The existence of a sensitive period in the development of neural circuits involved in anxiety becomes particularly interesting when the myriad influences are examined for their combined effects. These effects can be antagonistic, as seen when the 5-HT1B receptor is removed in the MAO-A and SERT knockout mice. Removal of this receptor blocks the detrimental effects of increased serotonin on the formation of thalamocortical and retinal circuits. The combined effects can also be additive, exemplified by an exacerbation of the anxiety phenotype in SERT knockout mice when BDNF, another molecule important in circuit formation, is also removed.

It is important to note that these combined effects are not limited to gene–gene interactions. As discussed at length above, the normal functioning (l/l) allele of the SERT polymorphism appears to be protective against many of the behavioral consequences of early life stress in humans and in primates, whereas those that are homozygous (s/s) or heterozygous (l/s) for the hypo-functioning short allele have less resilience in the face of early life adverse events.

Conclusions, Future Directions, and Resilience

Understanding that psychopathology, particularly anxiety responses, can represent developmentally mediated changes in neural circuits is not an end but rather a beginning. It affords an opportunity to more carefully dissect not only the factors involved in psychopathology, but also the resilience factors that protect individuals with circuit-level dysfunction from developing illness.

Understanding brain function and pathology as an intermingling of individual circuits, each subject to its own developmental trajectory, also serves as a framework to probe for deeper understanding of the process of brain development. Defining circuits for a particular research focus opens the door for a careful dissection of genetic and environmental influences on neural function that is not available when using the gross outcome measures of behavior or disease.

Furthermore, it is possible that dysfunction in neural circuit formation may represent a more fundamental dysfunction of molecular circuits. This certainly may be the case in the 5-HT1A knockout mouse, in which the 5-HT1A receptor represents a critical regulator of negative feedback on serotonin function. Understanding dysfunction of molecular circuits would not only enhance understanding of the fundamentals of brain function but may also provide a yet more sensitive phenotypic marker of risk of psychopathology and a potential marker for treatment responsiveness.

Ultimately, however, attempts must be made to relate neural changes to behaviors or cognitive functions. Having defined changes in molecular or neural circuits, one then has a carefully defined subset of individuals to probe for subtle differences in higher-order functions. It is through this iterative process that progress in understanding brain function and behavioral pathology can be made.

See also: Aggression: Hormonal Basis; Aggression: Neurochemical and Molecular Mechanisms; Amygdala: Contributions to Fear; Anxiety Disorders; Aversive Emotions: Molecular Basis of Unconditioned Fear; Obsessive–Compulsive Disorder; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors.

Further Reading

Ansorge MS, Zhou M, Lira A, Hen R, and Gingrich JA (2004) Early-life blockade of the 5-HT transporter alters emotional behavior in adult mice. *Science* 306(5697): 879–881.

Ben-Ari Y (2001) Developing networks play a similar melody. *Trends in Neurosciences* 24(6): 353–360.

Caspi A, McClay J, Moffitt TT, et al. (2002) Role of genotype in the cycle of violence in maltreated children. *Science* 297(5582): 851–854.

Caspi A, Sugden K, Moffitt TT, et al. (2003) Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* 301(5631): 386–389.

Charney DS (2003) Neuroanatomical circuits modulating fear and anxiety behaviors. *Acta Psychiatrica Scandinavica. Supplementum* (417): 38–50.

Gaspar P, Cases O, and Maroteaux L (2003) The developmental role of serotonin: News from mouse molecular genetics. *Nature Reviews Neuroscience* 4(12): 1002–1012.

Gingrich JA, Ansorge MS, Merker R, Weisstaub N, and Zhou M (2003) New lessons from knockout mice: The role of serotonin during development and its possible contribution to the origins of neuropsychiatric disorders. *CNS Spectrums* 8(8): 572–577.

Gingrich JA and Hen R (2001) Dissecting the role of the serotonin system in neuropsychiatric disorders using knockout mice. *Psychopharmacology* 155(1): 1–10.

Gordon JA and Hen R (2004) The serotonergic system and anxiety. *Neuromolecular Medicine* 5(1): 27–40.

Gross C and Hen R (2004) The developmental origins of anxiety. *Nature Reviews Neuroscience* 5(7): 545–552.

Gross C, Zhuang X, Stark K, et al. (2002) Serotonin1A receptor acts during development to establish normal anxiety-like behaviour in the adult. *Nature* 416(6879): 396–400.

Horn G (2004) Pathways of the past: The imprint of memory. *Nature Reviews Neuroscience* 5(2): 108–120.

Knudsen EI (2004) Sensitive periods in the development of the brain and behavior. *Journal of Cognitive Neuroscience* 16(8): 1412–1425.

Levitt P, Harvey JA, Friedman E, Simansky K, and Murphy EH (1997) New evidence for neurotransmitter influences on brain development. *Trends in Neuroscience* 20(6): 269–274.

McNaughton N and Corr PJ (2004) A two-dimensional neuropsychology of defense: Fear/anxiety and defensive distance. *Neuroscience and Biobehavioral Reviews* 28(3): 285–305.

Nemeroff CB (2004) Neurobiological consequences of childhood trauma. *Journal of Clinical Psychiatry* 65(supplement 1): 18–28.

Noda M, Higashida H, Aoki S, and Wada K (2004) Multiple signal transduction pathways mediated by 5-HT receptors. *Molecular Neurobiology* 29(1): 31–39.

Phelps EA and LeDoux JE (2005) Contributions of the amygdala to emotion processing: From animal models to human behavior: Review. *Neuron* 48: 175–187.

Ressler KJ and Nemeroff CB (2000) Role of serotonergic and noradrenergic systems in the pathophysiology of depression and anxiety disorders. *Depression and Anxiety* 12(supplement 1): 2–19.

Whitaker-Azmitia PM (2001) Serotonin and brain development: Role in human developmental diseases. *Brain Research Bulletin* 56(5): 479–485.

Zhuang X, Gross C, Santarelli L, Compan V, Trillat AC, and Hen R (1999) Altered emotional states in knockout mice lacking 5-HT1A or 5-HT1B receptors. *Neuropsychopharmacology* 21(supplement 2): 52S–60S.

Aversive Emotions: Molecular Basis of Unconditioned Fear

J B Rosen, University of Delaware, Newark, DE, USA

© 2009 Elsevier Ltd. All rights reserved.

Interest in the scientific study of emotions in psychology, biology, and neuroscience has waxed and waned for many years. In the latter part of the nineteenth century, Darwin brought emotion to the forefront of biology with a functionally and biologically grounded analysis of emotional expression. Darwin considered that emotion and the expressions of emotional behavior developed because of their importance in intra- and interspecies communication that promotes survival. Many emotions were considered innate and developed through evolutionary and functional mechanisms, just as body morphology evolved through environmental pressures. However, the scientific study of emotions has always been hampered by the subjectivity of emotional experience and therefore was not as prominent a field of study in psychology and neuroscience as other fields of behaviorism and cognition were throughout much of the twentieth century. In the last few decades, the study of the neurobiology and molecular biology of emotions, particular aversive emotions such as fear, has recently become tractable in both animals and humans because of the ability to measure behavioral and neural emotional responses, and it consequently has become a major field of neuroscientific study in the twenty-first century.

Great strides in our knowledge of the neuroscience of fear stems from the use of simple conditioning paradigms that measure fear with species-specific defensive responses to elucidate a fundamental and detailed understanding of the neuroanatomy of fear that can then be followed by the study of the molecular biology within these neuroanatomical circuits of fear. Combining a number of learning procedures, particularly Pavlovian fear conditioning paradigms in rodents that produce robust learning and memory of fear, with neuroscience techniques has been very successful in developing an understanding of neurobehavioral systems in the brain that are crucial for learning, memory, and the expression of the emotion fear. On the heels of the study of learned fear, in recent years the neuroscientific study of unconditioned fear has begun and is now accumulating data that suggest the neural and molecular mechanisms of unconditioned fear may have similarities to, but also important differences from, those of conditioned fear.

Conditioned and Unconditioned Fear

Fear conditioning paradigms in animals have a long history in psychological research and theory. Pavlov not only paired a bell with the arrival of food to induce conditioned or learned appetitive responses to the bell, but also paired bells and other neutral stimuli to unconditioned aversive events, such as shocks, to induce aversive responses to these formerly neutral, but now conditioned, stimuli. Thus, just as stimuli contingently paired with food become predictive signals for the arrival of food, stimuli contingently paired with shock develop into predictive signals for the arrival of the shock. Pavlovian fear conditioning therefore is an adaptively important process that helps organisms survive and is one of the most rapid types of learning in many animal species, with significant learning in a single trial and typically reaching asymptote within a handful of trials. Although Pavlovian conditioning by pairing lights, tones, contexts, and shocks is a fairly simple paradigm to perform, the associative process is not merely a transfer of the ability to elicit behavioral responses from one stimulus to another but is, instead, cognitive in nature and involves animals making predictions, representations, and relations between events, situations, and circumstances. Pavlovian learning and memory processes have immense biological significance by increasing the survivability of animals in a world where the predictive value of stimuli changes.

Because Pavlovian fear conditioning is learning about predictive relationships between conditioned and unconditioned stimuli, and not about learning new behavioral responses, innate, functionally adaptive species-specific defensive responses are typically displayed during fear and experimentally are used as measures of conditioned fear. These include autonomic changes in heart rate and blood pressure, and increases in behavioral responses such as freezing and reflexes such as startle. All these measures are considered preparatory responses to the anticipated aversive stimulus that are induced by presentation of the conditioned stimulus and are, thus, thought to be indicative of a state of fear.

Whereas learning to be afraid of stimuli that we encounter is adaptive and has great significance for survival, the ability to be innately fearful of some stimuli has similar significance and may be important for encounters and circumstances that organisms typically do not survive (and therefore cannot learn

from). It is suggested that in humans fear of heights, the dark, and small animals may be examples of unconditioned fears. Unconditioned fear may also be a basis for some anxiety disorders, particularly some specific phobias. However, it is very difficult, if not impossible, to determine whether these fears are innate or are due to learning. However, this may be resolved in laboratory animals that have been bred for generations in a laboratory environment with no contact with putative unconditioned fear stimuli. Although there are a number of different types of unconditioned fear stimuli or situations that have been studied, such as neophobia and brightly lit environments, this article focuses on fear of predators and predator odors. It has been know since 1913 that laboratory rats display unconditioned fear responses to cats, and unconditioned fear to predators and predator odors have been studied most extensively.

The use of ecologically and biologically based stimuli, such as exposure to predators and odors from predators, allows for the controlled, scientific study of unconditioned fear. Use of these innate fear-inducing stimuli exploits the evolutionarily shaped unconditioned fear of predators and predator odors. The physiological and behavioral responses to these predators and predator odors are similar to those evoked by conditioned fear stimuli (e.g., cardiovascular changes, freezing, startle, and vigilance), indicating that the same species-specific defensive responses are used during unconditioned fear states. What distinguishes unconditioned from conditioned fear is that it is unnecessary to initially induce unconditioned fear with a painful stimulus and, significantly, that unconditioned fear appears without learning or prior exposure to the unconditioned stimulus. Also, fear responses in rats to repeated exposure to live predators (e.g., cat) and some predator odors (e.g., ferret fur odor or trimethylthiazoline, a synthetic fox odor) do not habituate. This does not mean that learning cannot occur with exposure to predators and predator odors. Robust conditioned flavor avoidance is supported by synthetic fox odor. Learning of cue and contextual fear can be supported by exposure to a live cat and some predator odors (e.g., cat fur), but this is difficult to demonstrate with other predator odors (e.g., synthetic fox odor, ferret fur odor, and cat feces), except in specific types of environments. However, sensitization can be produced by exposure to these predators and predator odors that may last for at least 3 weeks. These learning effects of predators and predator odors (both nonassociative sensitization and associative conditioning) may recruit different neural circuits from those necessary for unconditioned fear responses.

Neuroanatomical Circuits of Unconditioned Fear

Before discussing the molecular biology of unconditioned fear, I delineate the neural circuits where molecular changes are to be found. If we were to envision the neurocircuitry of unconditioned fear, it would probably be similar to that subserving conditioned fear. Pavlovian fear conditioning in rodents has been very important for delineating specific neuroanatomical circuits of fear conditioning (**Figure 1**). Numerous studies have demonstrated that the lateral, basal, and central nuclei of the amygdala are all necessary for fear conditioning. The lateral nucleus is particularly important as a site of stimulus integration because it receives information from association cortices and from sensory and pain-relay nuclei of the thalamus. The basal nucleus also integrates afferents from the lateral nucleus and other brain regions. The central nucleus receives input from the lateral and basal nuclei and is the major output pathway to many subcortical regions that mediate numerous fear-related responses and behaviors.

Whereas the neuroanatomy of conditioned fear is fairly well established, the neuroanatomy of unconditioned fear is just beginning to be dissected. Early studies of exposure to a cat or cat fur supported the view that conditioned and unconditioned neural circuits are the same by showing that fear behavior diminished in rats with large lesions or inactivation of the amygdala. However, more recent studies with smaller lesions of discrete nuclei of the amygdala found that the lateral, basal, and central nuclei of the amygdala are not necessary for the display of

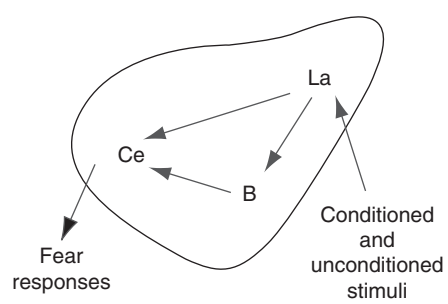

Figure 1 An amygdala fear circuit for fear conditioning. Input from thalamus and cortex with information about conditioned and unconditioned stimuli enters the lateral nucleus of the amygdala (La). The lateral nucleus projects to the basal (B) and central (Ce) nuclei of the amygdala. Output from the Ce nucleus projects to midbrain and brain stem areas to generate different kinds of conditioned fear behavior. This circuit is hypothesized to be important for fear conditioning, but not for unconditioned predator fear. However, it is recruited for associative and nonassociative fear induced by unconditioned predator fear.

unconditioned fear to various predator odors. A different neural fear circuit running through the medial hypothalamus may underlie innate defensive behavior to predators and predator odors (Figure 2). The core of this medial hypothalamic defense circuit includes the anterior hypothalamic nucleus (AHN), dorsomedial part of the ventromedial hypothalamic nucleus (VMHdm), and the dorsal premammillary nucleus (PMd). These nuclei are highly interactive, where the AHN has strong reciprocal connections with the VMHdm and PMd, and the VMHdm sends efferents to the PMd. This system also lies as an interface between sensory inputs and motor outputs, similar to the way that the amygdala does. The circuit is thought to play a major role in the somatomotor integration and initiation of aversively motivated behavior. For example, rats with PMd lesions display significantly reduced unconditioned freezing and escape during exposure to a live cat or cat fur odor. Also, lesion or inactivation of major inputs to this system (the medial nucleus of the amygdala or bed nucleus of the stria terminalis) also reduces freezing to predator odors.

The AHN and VMHdm have substantial olfactory afferents from main and accessory olfactory systems via the medial and accessory basal nuclei of the amygdala and the bed nucleus of the stria terminalis. VMHdm afferents from posterior intralaminal thalamus supply auditory inputs, paralleling thalamic auditory afferents of the lateral nucleus of the amygdala. Indirect input from the lateral nucleus of the amygdala by way of hippocampus and ventral subiculum provides processed visual information to the AHN. The VMHdm also receives substantial visceral and nociceptive information from the parabrachial nucleus in the brain stem. Efferents from the hypothalamic nuclei to the amygdala, septum, bed nucleus of the stria terminalis, and medial prefrontal cortex are able to influence processing in numerous regions of the telencephalon. Finally, the three core hypothalamic nuclei of the circuit each have substantial projections to the periaqueductal gray and other brain stem areas to generate and control a variety of fearlike behaviors.

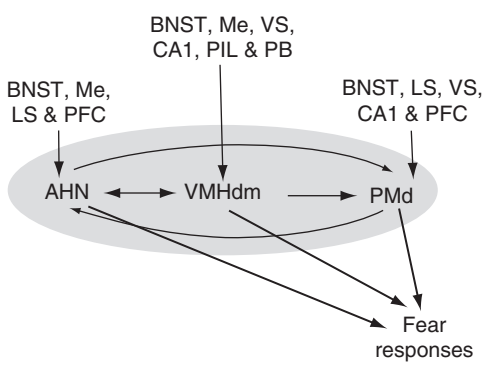

Figure 2 A medial hypothalamic fear circuit for unconditioned predator and predator odor fear. The anterior hypothalamic nucleus (AHN), the dorsomedial ventromedial hypothalamic nucleus (VMHdm), and the dorsal premammillary nucleus (PMd) make up the core of this circuit. Input into this system comes for numerous areas. Inputs from the bed nucleus of the stria terminalis (BNST), medial nucleus of the amygdala (Me), lateral septum (LS), ventral subiculum (VS), CA1 region of hippocampus (CA1), and several regions of the medial prefrontal cortex (PFC), posterior intralaminar nucleus of thalamus (PIL), and parabrachial nucleus (PB) are shown, but other inputs also exist. All core nuclei of the medial hypothalamic circuit project to brain stem regions (particularly strong to the periaqueductal gray) that control and generate fear responses.

Although the fear circuits for conditioned and unconditioned fear may be different, continued research will surely find a functional overlap between these circuits. Neuroanatomical studies demonstrate that the lateral nucleus of the amygdala receives projections from the VMHdm. The prefrontal cortex has been shown to project to both amygdala nuclei and the AHN and PMd. It has been known for many years that electrical and chemical stimulation of the ventromedial hypothalamus elicits fear behavior and that this is modulated by stimulation of the amygdala. Furthermore, lesions of the bed nucleus of the stria terminalis disrupt contextually fear-conditioned freezing and lesions of the medial nucleus of the amygdala interfere with fear-potentiated startle. Probably there is much interaction between fear conditioning and unconditioned fear circuits during both associative and nonassociative learning to unconditioned predator stimuli that will be elucidated in the coming years.

Molecular Biology of Unconditioned Fear

The study of the molecular and cellular mechanisms of unconditioned fear to predators and predator odors is in its infancy and sketchy at best, but it will eventually include the expression of molecules at all levels of neural transmission, signaling transduction, and gene expression. To data, the data demonstrate altered expression and participation of molecules at the levels of gene expression, neurotransmitters, and receptors following the exposure of rodents to these unconditioned fear stimuli.

Inducible Transcription Factors (Immediate-Early Genes)

The neural circuitry of unconditioned fear to predators and predator odors suggested by brain-lesion techniques is corroborated by measures of immediate-early

genes that are highly responsive to environmental stimulation. Changes in the expression of immediate-early gene mRNA or protein levels, such as c-fos and egr-1, not only can be used as markers for transient neural activity, but, because these genes are inducible transcription factors, they provide a pattern of gene expression in these brain regions that may regulate some of the long-lasting changes in behavior (e.g., nonassociative sensitization, and associative conditioning) induced by exposure to predator stimuli.

Exposing rats to cats induces c-fos mRNA in regions of the medial hypothalamic defensive circuit, including the ventromedial hypothalamus, dorsal premammillary nucleus, bed nucleus of the stria terminalis, and lateral septum. Significantly, c-fos in these regions does not increase with restraint or hypoxia, other common stress-inducing manipulations. Furthermore, c-fos in the medial nucleus of the amygdala is increased with both cat exposure and restraint stress. Other studies of fear conditioning also demonstrate that c-fos increased in the medial nucleus of the amygdala, indicating that the activation of the medial nucleus of the amygdala is not specific for unconditioned fear to predators or predator odors but is activated during many types of fear or stress.

Studies investigating exposure to predator odors (cat and ferret fur and synthetic fox odor) induce similar patterns of c-fos mRNA and Fos protein expression as seen with exposure to a live cat, but more brain regions have been examined. Thus, these predator odors are found to induce c-fos mRNA or Fos protein in the main olfactory bulbs, paraventricular nucleus of the hypothalamus, locus coeruleus, cuniform nucleus, and all divisions of the periaqueductal gray (the brain region important for the fear behaviors freezing and escape). Interestingly, the subnuclei of the amygdala (lateral and basal nuclei), regions critical for all types of fear conditioning (i.e., auditory, visual, olfactory, and contextual fear conditioning induced by shock), do not display increased c-fos expression following live predator or predator odors. In addition, the expression of another immediate-early gene, egr-1, also is not induced in the lateral and basal nuclei of the amygdala of rats exposed to a live cat. These data begin to delineate and corroborate lesion studies showing differences in brain circuits for conditioned and unconditioned fear.

The expression of these inducible transcription factor genes not only provides patterns of neural activation during behavior, but also suggests that these changes may alter long-term responses to predators and predator odors and, in the wild, may facilitate the survival of the animal. To date, few studies have examined the role of inducible transcription factors in unconditioned fear. However, these are some data demonstrating that the reduction of translation of egr-1 in the lateral nucleus of the amygdala interferes with fear conditioning but not with unconditioned fear to synthetic fox odor. Further research will elucidate the importance of inducible transcription factors in unconditioned and conditioned fear.

Phosphorylated Cyclic AMP Responsive Element-Binding Protein

Expression of other inducible transcription factors has not been investigated. However, there are a couple of studies examined the phosphorylation of cyclic AMP responsive element-binding protein (CREB) induced by cat exposure. CREB is a constitutive factor that on phosphorylation (pCREB) becomes an active transcription factor that has been shown to be involved in many forms of learning and memory, including fear conditioning within the amygdala. pCREB had been found to increase in the lateral, basal, central, and medial nuclei of the amygdala, as well as in the periaqueductal gray of rats following exposure to a live cat. This is interesting because live-cat exposure induces a long-lasting sensitization that can be demonstrated in a number of anxiety tests. Of further interest is that pCREB does not increase in the ventromedial hypothalamus, a nucleus of the medial hypothalamic defense circuit. Taken together with the findings that the inducible transcription factors, c-fos and egr-1, are not induced in the lateral and basal nuclei of the amygdala by a live cat, cat odor, ferret odor, or synthetic fox odor but are induced in the ventromedial hypothalamus, the data suggest that unconditioned predator fear and predator-induced sensitization engage different neural circuits. Circuits for predator-induced sensitization appear to be similar to those necessary for shock-induced fear conditioning.

Extracellular Neurotransmitter Signaling

Dopamine

The effect of unconditioned predator fear on the dopamine system has been studied more than other neurotransmitter systems. Dopamine release and turnover in the prefrontal cortex and amygdala (subnuclei of the amygdala were not specified), which are known to increase in response to stress, also increase with exposure to synthetic fox odor. Dopamine neurons in the ventral tegmental area that project to the prefrontal cortex and amygdala are activated by synthetic fox odor, whereas dopamine neurons in the substantia nigra do not. Unconditioned fear induced by synthetic fox odor also interferes in a non-match-to-sample short-term working memory task that is dependent on prefrontal cortex processing. Dopamine in other regions, such as hypothalamus, striatum, and nucleus

accumbens, has also been shown to increase with exposure to predator odors but not as consistently as the increase in the prefrontal cortex. Pharmacologically, dopamine D1 receptor agonist infused in the prefrontal cortex of mice increases fear and anxious behavior to predator and predator odor exposure. These data suggest that unconditioned predator and predator odor exposure, similar to other fear- and stress-inducing stimuli, enhance dopamine neurotransmission in the prefrontal cortex and interfere with working memory tasks that are dependent on the prefrontal cortex.

Serotonin

In general, pharmacological serotonin 5-$HT1_A$ receptor agonism and blockade of serotonin reuptake reduce fear and anxious behavioral responses in rats and mice to predators and predator odors. In mice, there appears to be differential effects of 5-$HT1_A$ receptor agonists to reduce predator risk-assessment behavior, and 5-$HT2_A$ receptor antagonism to reduce flight or escape behavior. Mutant mice with a knockout of the serotonin transporter were shown to have increased vulnerability to predator-induced sensitization, as shown by the increase anxiety-like behavior in the elevated-plus maze and a light–dark box test after exposure to a cat. The increased vulnerability to unconditioned fear in these mice may be similar to the enhanced vulnerability to stress in humans with the short allele of the serotonin transporter.

Norepinephrine

Norepinephrine plays a role in vigilance and conditioned fear, and therefore it should also modulate unconditioned fear. Norepinephrine was shown to increase in the hippocampus in mice exposed to a cat and in the bed nucleus of the stria terminalis of rats exposed to synthetic fox odor. Complementary to increases in norepinephrine, the blockade of norepinephrine neurotransmission in the bed nucleus of the stria terminalis with a local infusion of clonidine decreased unconditioned fear-related freezing responses to synthetic fox odor. Furthermore, in a drug schedule of intermittent methylphenidate treatment for 3 weeks that sensitized dopamine and norepinephrine systems, rats given yohimbine, which facilitates norepineprhine neurotransmission, displayed increased avoidance of a predator odor. Together, these studies confirm that norepinephrine neurotransmission is important for unconditioned fear to predators and predator odors.

γ-Aminobutyric Acid

γ-Aminobutyric acid (GABA) levels have been shown to increase in the amygdala and nucleus accumbens with exposure to a predator or predator odor. These increases are probably compensatory because the facilitation of $GABA_A$ neurotransmission with the administration of systemic benzodiazepine agonist reduces defensive threat/attack and risk-assessment behaviors to cat and cat odors in mice. Unconditioned freezing in rats to synthetic fox odor is blocked by local infusion of muscimol, a $GABA_A$ agonist, into the bed nucleus of the stria terminalis or medial nucleus of the amygdala. Interestingly, muscimol infusion into the basal nucleus of the amygdala of rats only delays freezing in response to synthetic fox odor for a few minutes, but freezing reaches levels-seen in vehicle-treated rats. Furthermore, in a study investigating fear behavior of sheep in response to a predator (a dog), an infusion of bicuculline, a $GABA_A$ antagonist into the sheep's central nucleus of the amygdala had only a marginal effect on predator-elicited fear, but sensitized predator fear when measured 2 days after the initial exposure to the dog. These data suggest that the basal and central nuclei of the amygdala play only a minor modulatory role in unconditioned fear to predators, but may have a more substantial role in long-lasting fear sensitization induced by predator fear.

Glutamate

Both competitive and noncompetitive antagonism of *N*-methyl-D-aspartate (NMDA) receptors has effects on fear induced by exposure to a predator and to predator odor. In rat pups, CGP 43487 given at postnatal day 5 reduces freezing elicited by exposure to an adult male rat at postnatal days 13 and 20 but not at postnatal day 30. In adult rats, CPP and AP7 do not affect rat defensive behavior to cat exposure, but they significantly interfere with long-lasting enhancement of fear on a number of anxiety tests. This is reminiscent of the role of NMDA receptors in long-term potentiation, in which NMDA receptors are involved in long-term synaptic plasticity by not short-term potentiation. The role of NMDA glutamate receptors in the amygdala appears to be involved in the long-lasting fear sensitization induced by predator exposure, possibly by strengthened synaptic transmission in amygdala–periaqueductal gray circuits. Glutamate has also been shown in increase in the rat nucleus accumbens in response to synthetic fox odor. Other brain regions have not been examined yet.

Corticotropin-Releasing Hormone and Corticosterone

Corticotropin-releasing hormone (CRH) and corticosterone are part of the hypothalamic–pituitary–adrenal axis. This system is important for the regulation of glucose metabolism to provide fuel for diverse behavioral and physiological demands,

including fear. Blood plasma corticosterone levels are increased by exposure to predator and predator odors. Exogenous CRH induces fear and anxiety-like behavior. Both CRH and corticosterone are involved in the consolidation of fear conditioning, and the amygdala is a major locus of action. Chronic high levels of corticosterone enhance memory of learned and unconditioned fear, and they also induce CRH expression in the central nucleus of the amygdala and the bed nucleus of the stria terminalis. Corticosterone enhancement of unconditioned startle appears to be through CRH in the bed nucleus of the stria terminalis but not the amygdala. However, exposure to predator and predator odor increases CRH release in the amygdala. In sheep exposed to a dog (a fear stimulus in sheep), CRH release in the amygdala occurs in two waves. The first wave lasts for only a few minutes and is independent of corticosterone, whereas the second wave of CRH release is corticosterone dependent. Interestingly, repeated exposure to the dog potentiates fear to the dog and enhances only the corticosterone-dependent CRH release in the amygdala. Furthermore, infusion of a CRH antagonist into the amygdala had little effect on unconditioned fear behavior to the dog, but interfered with the potentiation of fear to subsequent dog exposure. These data suggest that CRH in the amygdala is not critical for unconditioned predator fear but that corticosterone-regulated CRH in the amygdala is important for nonassociative sensitization to predator stimuli. Whether corticosterone-induced enhancement of CRH in the bed nucleus of the stria terminalis is involved in sensitization to predators, as it is in sensitization to cocaine, awaits further research.

Cholesystokinin

Cholesystokinin (CCK) is known to induce anxiety-like feelings in humans. In rats, it appears to play a role in unconditioned predator fear and sensitization induced by predator exposure. CCK is increased in the basal nucleus of the amygdala in mice exposed to synthetic fox odor compared to mice not exposed to this odor, but only after a light–dark box anxiety test. Blockade of CCK-2 receptors, but not CCK-1 receptors, decreases fearfulness in rats during exposure to a cat and decreases cat-exposure-induced sensitization demonstrated in the elevated plus maze. An interesting difference in response to cat exposure has been found in PVG hooded and Sprague-Dawley rats, in which PVG hooded rats freeze more during exposure to a cat than do Sprague-Dawley rats. PVG hooded rats also have a greater abundance of CCK-2 receptors than and several genetic variations in the code for CCK-2 receptors different from those of Sprague-Dawley rats. The importance of these differences for unconditioned fear is not known at present, but it is interesting to note that Sprague-Dawley and Long-Evans rats display robust freezing to synthetic fox odor, whereas Wistar rats do not. There are also mouse strain differences in response to predator odors. Whether these are related to CCK or other neurotransmitter systems is not known.

Conclusion

Investigation of the molecular basis of unconditioned fear, particularly fear of predators and predator odors, has seen substantial progress in the last few years. As shown by lesion and expression of immediate early genes, a neural circuit for unconditioned predator fear through the medial hypothalamus is emerging that may be distinct from an amygdala circuit for fear conditioning. However, there appears to be an overlap of these circuits and recruitment of the amygdala fear-conditioning circuit for associative and nonassociative conditioning to predator stimuli. Although inducible transcription factors seem to be expressed in the unconditioned fear circuit during exposure to predator stimuli, nonasssociative sensitization of fear induced by predator exposure appears to induce pCREB, a phosphorylated constitutive transcription factor known to be involved in fear conditioning and other types of learning, in amygdala nuclei of the fear-conditioning circuit. A number of neurotransmitter systems also seem to be involved in unconditioned fear and sensitization induced by predator exposure. The list of neurotransmitter systems involved will certainly grow. Investigation of signal transduction pathways has not started, but will probably contribute to our understanding of unconditioned fear in the coming years. Continued research with ethologically based fear stimuli should provide more neuroanatomical and molecular detail of unconditioned fear and should have a substantial influence on our understanding of normal fear and related anxiety disorders.

See also: Amygdala: Contributions to Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Emotion: Neuroimaging; Emotional Control of the Autonomic Nervous System; Fear Conditioning and Synaptic Plasticity ; Genetics of Human Anxiety and Its Disorders.

Further Reading

Adamec RE, Blundell J, and Burton P (2006) Relationship of the predatory attack experience to neural plasticity, pCREB expression and neuroendocrine response. *Neuroscience and Biobehavioral Reviews* 30: 356–375.

Adamec RE, Burton P, Blundell J, Murphy DL, and Holmes A (2006) Vulnerability to mild predator stress in serotonin transporter knockout mice. *Behavioural Brain Research* 170: 126–140.

Blanchard DC, Canteras NS, Markham CM, Pentkowski NS, and Blanchard RJ (2005) Lesions of structures showing FOS expression to cat presentation: Effects on responsivity to a cat, cat odor, and nonpredator threat. *Neuroscience and Biobehavioral Reviews* 29: 1243–1253.

Blanchard RJ, Blanchard DC, and Hori K (1989) An ethoexperimental approach to the study of defense. In: Blanchard RJ, Brain PF, Blanchard DC, and Parmigiani S (eds.) *Ethoexperimental Approaches to the Study of Behavior,* vol. 48, pp. 114–136. Dordecht: Kluwer Academic.

Blundell J, Adamec R, and Burton P (2005) Role of NMDA receptors in the syndrome of behavioral changes produced by predator stress. *Physiology & Behavior* 86: 233–243.

Canteras NS (2002) The medial hypothalamic defensive system: Hodological organization and functional implications. *Pharmacology, Biochemistry & Behavior* 71: 481–491.

Day HE, Masini CV, and Campeau S (2004) The pattern of brain c-fos mRNA induced by a component of fox odor, 2,5-dihydro-2,4,5-trimethylthiazoline (TMT), in rats, suggests both systemic and processive stress characteristics. *Brain Research* 1025: 139–151.

Dielenberg RA, Hunt GE, and McGregor IS (2001) 'When a rat smells a cat': Distribution of Fos immunoreactivty in rat brain following exposure to predatory odor. *Neuroscience* 104: 1085–1097.

Dielenberg RA and McGregor IS (2001) Defensive behavior in rats towards predatory odors: A review. *Neuroscience and Biobehavioral Reviews* 25: 597–609.

Fendt M, Endres T, Lowry CA, Apfelbach R, and McGregor IS (2005) TMT-induced autonomic and behavioral changes and the neural basis of its processing. *Neuroscience and Biobehavioral Reviews* 29: 1145–1156.

Masini CV, Sauer S, and Campeau S (2005) Ferret odor as a processive stress model in rats: Neurochemical, behavioral, and endocrine evidence. *Behavioral Neuroscience* 119: 280–292.

Morrow BA, Elsworth JD, and Roth RH (2002) Fear-like biochemical and behavioral responses in rats to the predator odor, TMT, are dependent on the exposure environment. *Synapse* 46: 11–18.

Morrow BA, Roth RH, and Elsworth JD (2000) TMT, a predator odor, elevates mesoprefrontal dopamine metabolic activity and disrupts short-term working memory in the rat. *Brain Research Bulletin* 52: 519–523.

Rosen JB (2004) The neurobiology of conditioned and unconditioned fear: A neurobehavioral system analysis of the amygdala. *Behavioral and Cognitive Neuroscience Reviews* 3: 23–41.

Takahashi LK, Nakashima BR, Hong H, and Watanabe K (2005) The smell of danger: A behavioral and neural analysis of predator odor-induced fear. *Neuroscience and Biobehavioral Reviews* 29: 1157–1167.

Walker DL, Toufexis DJ, and Davis M (2003) Role of the bed nucleus of the stria terminalis versus the amygdala in fear, stress, and anxiety. *European Journal of Pharmacology* 463: 199–216.

Awareness: Functional Imaging

G Rees, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

This article describes attempts to identify experimentally the neural correlates of the contents of conscious experience in humans. In general usage, the term 'conscious' can be used in both a transitive and intransitive sense. For example, it is possible to be conscious (as opposed to being unconscious, asleep, or in some other state). But it is also possible, when awake, to be conscious of a particular thing (as opposed to not being conscious of that thing, or conscious of some other thing at that particular moment). This distinction means that any experimental account of the neural correlates of consciousness must at a minimum specify the neural structures and processes associated with each sense of the word conscious.

The Waking State

The contents of consciousness are accessible and can be reported by individuals who are in the waking state. Changes in the level of consciousness, such as those induced artificially by anesthesia or pathologically by brain damage leading to coma, lead to individuals' being rendered unconscious and unable to report the contents of their consciousness. Sensory processing is altered, leading to an absence of conscious representation. Neural structures mediating level of consciousness are therefore surmised to facilitate the representation of conscious contents. Understanding how sensory processing is altered during changes in the level of consciousness can therefore help understand how the contents of consciousness are represented in the human brain.

Sensory processing is profoundly altered by changes in level of consciousness resulting from different causes. Typically, responses can still be obtained from the primary sensory cortices, but responses of association and the higher cortices are reduced or absent. For example, primary auditory cortex activation can still be elicited during sleep, during anesthesia, or after brain damage leading to coma. Similarly, less dramatic modifications of level of consciousness associated with meditation modify activity only in areas of the posterior sensory cortex, apart from primary visual cortex.

Anesthesia represents perhaps the most common deliberate manipulation of the level of consciousness, and it is crucial for surgical practice. Although there is increasing knowledge about the molecular and cellular targets of anesthetic agents, their effects on sensory processing in the human brain at a systems level remain poorly understood. Inhalational agents modulate brain stem nuclei and (by inference) their ascending modulatory efferent pathways. For example, isoflurane and halothane cause relative reductions in regional cerebral glucose metabolism in midbrain reticular formation, and propofol-induced depression of thalamic and reticular blood flow are closely linked. Similarly, halothane and enflurane or isoflurane depress spontaneous single-unit firing in the reticular formation of cat. However, brain stem responses to sensory stimulation (e.g., auditory evoked potentials) appear intact.

In the auditory cortex, most investigations of the effects of anesthesia on sensory processing have focused on auditory stimulation and are consistent with the general position outlined previously. In contrast, very little is known about processing in the other sensory cortices under anesthesia. For the visual system, a dose-dependent reduction in activation of the primary visual cortex is observed with thiopental, but that study did not measure the depth of anesthesia so we cannot correlate such findings with the level of consciousness. Subanesthetic isoflurane affects task-induced activation in the parieto-frontal (but not visual) cortices during performance of a visual search task. Visual evoked potentials can still be obtained during anesthesia, although somewhat unreliably in the operative environment, indicating some preservation of cortical processing. Similarly, little is known about brain areas involved in tactile processing in the unconscious human, despite the obvious clinical requirement to eliminate somatosensory sensation during surgical procedures. The clinical utility of monitoring somatosensory evoked potentials indicates that cortical processing can persist, although presumably without awareness. Two neuroimaging studies have examined localized cortical responses to somatosensory stimuli. Both showed diminished responses in the primary somatosensory cortex to electrical or vibrotactile stimulation but showed preserved responses in either the thalamus or secondary somatosensory cortex. This suggests impaired thalamocortical information transfer during anesthesia, but it is somewhat inconsistent with observations from other techniques. For example, intraoperative intrinsic signals from human somatosensory cortices to vibrotactile stimuli delivered during anesthesia are topographically specific to the stimulated digit. The reasons for this inconsistency are unclear, but it may suggest that some

somatosensory cortical responsiveness is retained during anesthesia.

Taken together, these studies suggest that the waking state (compared to the anesthetized state) is associated with significant modifications in the responsiveness of sensory cortices to stimulation, particularly affecting secondary and association cortices. One unifying model for the actions of anesthetic agents is that the interruption of the dynamic connectivity between the thalamus and cortex represents a common mechanism underlying the observed changes in sensory processing and unconsciousness. Because changes in thalamocortical connectivity can also be observed in persistent vegetative state and resolve on recovery, it is conceivable that a single underlying mechanism underlies all changes in sensory processing associated with modification of the level of consciousness.

Rarely, awareness during anesthesia is reported subsequent to the administration of an anesthetic. The neural substrates of such experience remain poorly understood due to its rarity. A single case has recently been reported of an individual in a minimally conscious state who could not report her experiences overtly; however, when the patient was asked to imagine one of two different experiences, brain imaging revealed different patterns of brain activation. One possible explanation of these findings is that the individual was able to formulate an intention to follow the instructions given by the experimenter, implying concomitant awareness. How this can arise in an apparently comatose patient unable to report by any overt means and its implications for understanding the mechanisms of coma and changes in level of consciousness more generally remain to be explored.

Conscious Access: Neural Correlates of Perceptual Awareness

When people are awake, a characteristic feature of awareness is that its contents can be reported. We have subjective conscious access to the contents of our consciousness. But it is equally apparent that much of our mental life remains unconscious and inaccessible to report. Experimental enquiry must therefore distinguish what is special concerning the neural activity specifically associated with the contents of conscious experience, in contrast to other activity that merely reflects the unconscious correlates of perception or behavior. The principal methodological approach of the work reviewed next is thus to contrast the neural activity evoked by conscious versus unconscious processing, using a number of experimental paradigms.

Spontaneous Variation in the Contents of Consciousness

Bistable Stimulation

When a physical stimulus allows two different perceptual interpretations, awareness typically fluctuates spontaneously between the two possible perceptual interpretation. For example, binocular rivalry occurs when two different images are presented, one to each eye separately. Conscious experience fluctuates spontaneously, each monocular view being in turn visible for a few seconds while the other is suppressed. Because these fluctuations in awareness occur without any change in the physical stimulus, binocular rivalry (and other forms of bistable perception) form a useful model system with which to distinguish the neural correlates of these changes in the contents of consciousness from brain activity related solely to physical stimulus characteristics.

Investigations of the neural correlates of bistable stimulus perception identified two distinct patterns of findings. First, brain activity time-locked to perceptual transitions can be identified in the prefrontal and parietal cortices for both binocular rivalry and a wide variety of bistable stimuli. Moreover, the involvement of these cortical areas is independent of any requirement to make overt reports of the perceptual transitions. Similar areas are also active during spontaneous changes in the contents of consciousness in very different paradigms, during the emergence of a figure when the subject views either fragmented figures or stereo pop-out stimuli. Activation in these areas of the parietal and prefrontal cortices is associated with selective attention. More recently, activation of different parietal and prefrontal loci, thought to be involved in working memory, have been associated with the maintenance of a stable bistable percept. Although these studies do not identify whether these different psychological processes are specifically associated with such prefrontal and parietal activation, they offer a rich line of future investigation.

A second major pattern of findings associated with studies of bistable perception is that fluctuations in brain activity specifically associated with the particular contents of consciousness can be identified in the ventral visual cortex. Strikingly, such fluctuations can be identified at all levels of the human geniculostriate pathway, from the lateral geniculate nucleus (LGN) to V1 to category-specific areas such as the fusiform face area. Indeed, the local pattern of activity in cortically evoked responses carries sufficient information that automated algorithms can be used to blindly decode and accurately predict the fluctuations in perceptual experience that the observer reports. Significantly, response fluctuations that are

further along the ventral visual pathway are generally larger than those observed at earlier stages, and at later stages in processing they are equal in magnitude to responses evoked to physical (nonrivalrous) stimulus alteration. This pattern of findings has been interpreted as suggesting that, at anatomically early stages of the ventral visual pathway, both seen and perceptually suppressed images are physically represented, but that competition between the two is increasingly resolved at higher stages of the visual system. Such findings are broadly consistent with event-related potential (ERP) measurements in humans and single-unit recordings in monkeys.

The involvement of both ventral visual cortical structures (representing the current contents of consciousness) and the frontoparietal structures (whose activity is either time-locked to perceptual transitions or associated with perceptual maintenance) suggests the involvement of a distributed network in perceptual rivalry. Consistent with such a hypothesis, large-scale changes in intra- and interhemispheric neuronal synchronization can be observed during rivalrous fluctuations. One model that might account for these data is that distributed object representations in the ventral visual pathway compete for perceptual dominance, biased by top-down signals reflecting working memory and attention from the frontal and parietal cortices. In the context of bistability, a conscious percept is thus represented by both an activated representation of the perceptual content and activity in the frontal and parietal cortices. Such a hypothesis is consistent with emerging theoretical models of consciousness that posit the involvement of both sensory representations and a global workspace, the latter tentatively identified with the frontal and parietal cortices.

Near-Threshold Stimuli

A detection threshold can be defined for any type of sensory stimulus, for example, by presenting it in temporal proximity to a mask or by physically degrading it in some fashion. At detection threshold, only a proportion of stimuli presented to observers will be detected, whereas the others fail to reach awareness. This spontaneous trial-to-trial variability in awareness of stimulation can be used to determine the neural correlates of stimulus awareness without confounding by changes in physical stimulus characteristics. Alternatively, brain responses to different types of fully masked and invisible stimuli can be used to determine neural correlates of unconscious processing.

Fully invisible stimuli can nevertheless evoke modest but reliable activation throughout the visual cortex. For example, activity specific to the orientation of a simple visual grating can be decoded from the primary visual cortex activity even when the stimulus is completely invisible to the observer. Similarly, when a simple target such as an achromatic disc is briefly flashed, robust activity is elicited in the corresponding retinotopic location of V1 (plus V2 and V3), even when the target is rendered completely invisible by a surrounding mask. Invisible images of tools evoke both activity in the primary visual cortex and category-specific activity in visually responsive dorsal stream structures. Invisible motion can activate the V5/middle temporal (MT) visual motion area. At higher levels of the visual system, activation selective for the presence (although undetected and invisible to the observer) of a word or for invisible faces and objects can also be observed. When an object or word is not noticed during the attentional blink, unconscious activation of the ventral visual pathway corresponding to the object identity or semantic processing can be observed. Unconscious sensory activation by complex visual stimuli is not confined to the cortical visual pathways but extends to subcortical structures. For example, the amygdala can be activated by fearful-face stimuli that have been rendered invisible through masking, in response to the emotional content of invisible words or during the suppression in binocular rivalry.

Complementing these studies in normal observers, unconscious activation of the visual cortex can also be identified in patients with visual extinction resulting from damage to the right parietal cortex. Areas of both the primary and extrastriate visual cortex that are activated by a seen left-visual-field stimulus are also activated by an unseen and extinguished left-visual-field stimulus. The unconscious processing of an extinguished face stimulus extends even to the face-selective cortex in the fusiform face area. Taken together, these findings indicate that, even without awareness, extensive processing of visual stimuli can take place in the normal visual cortex, including the generation of category-specific responses.

Invisible stimuli evoke activity associated with their mere presence and/or identity; it is equally clear that, when stimuli become visible, this is associated with an increase in activity in the corresponding category-specific areas of the visual cortex. For example, simple detection of a low-contrast grating stimulus is associated with retinotopically specific enhancement of activity in the primary visual cortex. Visually presented objects can be made difficult to identify by degrading them, and in such circumstances occipitotemporal activity shows a close correlation with recognition performance. At higher levels of the ventral visual cortex, successful detection of a face stimulus during the attentional blink is associated with significantly higher

activity in the fusiform face area. Such successful detection of visually presented stimuli in the attentional blink is associated with event-related potentials that occur relatively late in stimulus processing (after several hundred milliseconds), suggesting that such activation may be associated with a late wave of activation that spreads through a distributed network of the cortical association areas. A similar pattern of enhanced activation in category-specific areas of the visual system is seen in other paradigms comparing conscious and unconscious perceptions. For example, the introduction of a flicker between successively presented visual images makes it particularly difficult to detect large-scale changes between the images. When such a change is successfully perceived, however, enhanced activity in regions of the visual cortex with category specificity for the detected change can be observed that may precede conscious detection of the change.

Successful conscious detection of stimuli is associated not just with changes in activation of the ventral visual cortex but also with changes in coupling between the early retinotopic cortex and higher visual areas. When the interval between a target and a nonoverlapping metacontrast mask is varied over a few hundred milliseconds, observers report that visibility of the target first declines before rising again, in a U-shaped function. Strikingly, the degree of coupling (or effective connectivity) between the early retinotopic cortex and higher visual areas shows a pattern of changes with target–mask stimulus onset asynchrony (SOA) that closely parallel this behavioral function, suggesting that effective connectivity may play a role in determining visibility.

As for bistable stimuli, conscious detection of stimuli close to the threshold involves not just the sensory cortices but areas of the frontal and parietal cortices. For example, successful detection of a change is associated with parietofrontal activation that may reflect the successful deployment of attention. At the detection threshold, differences in parietal and prefrontal activity emerge following electrical activity associated with conscious perception at approximately 100 ms. Parietal cortical activation is also associated with conscious recognition of visual verbal stimuli. Successful identification of an object evokes event-related negativity and is associated with occipital responses and modulation of the parieto-occipital alpha rhythm. Finally, conscious detection of flicker is associated with the activation of the frontal and parietal cortices, indicating that these areas are important not just for awareness of spatially extended objects but also for conscious detection of temporally distinct visual events.

Hallucinations

A hallucination is a false sensory perception in the absence of an external stimulus, and it is often associated with drug use, sleep deprivation, and neurological or psychiatric disease. During visual hallucinations, brain activation can be observed in object-selective regions of the ventral visual pathway whose selectivity corresponds to the content of the hallucination. During auditory hallucinations in schizophrenia, the subcortical nuclei and limbic structures and the primary auditory cortex are activated. Finally, hallucination of a supernumerary limb after stroke is associated with activation of the medial prefrontal cortex in the supplementary motor area. Hallucinations are therefore associated with the activation of either the sensory or motor cortex whose modality and neuronal specificities corresponds closely to that of the hallucinatory perceptual content.

Modifying the Contents of Consciousness

Changes in the contents of consciousness can be driven by spontaneous fluctuations. They can also be systematically altered by either top-down signals (e.g., directing attention toward or away from some aspect of a stimulus) or by altering the spatial or temporal context in which a stimulus is presented.

Attention

When subjects are engaged in a demanding task, irrelevant but highly salient stimuli outside the immediate focus of attention can go entirely unnoticed. This is known as inattentional blindness, and it suggests that consciousness may depend on attention. Consistent with this hypothesis, brain activity evoked by irrelevant sensory stimulation in the ventral occipital and temporal cortices is reduced when attention is withdrawn. Moreover, when inattentional blindness results for unattended words, then brain activity no longer differentiates between such meaningful words and random letters. This suggests that attention is required both for brain activity associated with the higher processing of sensory stimuli and for their subsequent representation as the contents of consciousness. However, although attention may be necessary for conscious representation, it is not sufficient. Varying the availability of attentional resources changes the activity evoked in the primary visual cortex by even completely invisible stimuli. Thus, attention can act on invisible stimuli and need not change their visibility as a consequence.

Illusions

An illusion occurs when real external stimuli are misperceived and represented in consciousness in an

incorrect fashion. The content of the illusory perception typically depends on the spatial and temporal context in which it occurs. For example, sensory aftereffects are illusory sensory perceptions in the absence of sensory stimulation that occur following an extended period of adaptation to a sensory stimulus. Aftereffects contingent on prior adaptation to color or motion activate V4 or V5/MT, respectively, and the time course of such activation reflects phenomenal experience. Perception of illusory or implied motion in a static visual stimulus results in the activation of V5/MT, and the perception of motion at unstimulated regions of the visual field on the path of apparent motion is associated with retinotopically specific activation of V1 mediated by feedback from V5/MT. Perception of illusory contours activates areas of the early retinotopic extrastriate cortex. Similarly, when a moving grating is divided by a large gap, observers report seeing a moving phantom in the gap and there is enhanced activity in the locations in the early retinotopic visual cortex corresponding to the illusory percept. The primary visual cortex represents the perceived angular size of an object rather than its actual size on the retina. Finally, when a uniformly illuminated surface is placed eccentrically on a dynamic textured background, after a few seconds, it is perceived to disappear and be replaced by the background texture. This texture filling-in is associated with the reduction of activity in the primary visual cortex associated with the representation of the now-invisible surface.

Common to all these experimental paradigms are changes in phenomenal experience without corresponding changes in physical stimulus. Perceptual illusions can also be used to manipulate feelings of ownership of a rubber hand. When participants feel that the hand is theirs, then activity in the premotor cortex reflects this conscious perception of ownership. Thus, both perceptual and motor illusions lead to a specific modification of brain activity. Altered brain activity is observed in areas of the brain known to contain neurons whose stimulus specificities encompass the attribute represented in consciousness.

Imagery

Conscious perception can be created by the act of imagination. In patients with implanted electrodes for presurgical epilepsy mapping, single neurons in the human medial temporal lobe that fire selectively when particular visual stimuli are presented are also activated when the individual imagines the same stimuli. Similarly, neuronal populations elsewhere in the ventral visual pathway with stimulus specificity for faces or places are activated during imagery of these categories of object. These findings are consistent with the notion that activity in functionally specialized neuronal populations in sensory cortices is associated with the representation of an object or image in consciousness.

Summary and Outstanding Problems

Findings from a wide variety of neuroimaging experiments are remarkably consistent in reaching the following conclusions. First, sensory stimuli that remain unconscious can nevertheless undergo extensive processing in the sensory cortices. Second, compared to stimuli that remain unconscious, consciously perceived stimuli elicit greater levels of activity in functionally specialized areas of the sensory cortex corresponding to the attributes represented in consciousness. Finally, consciously perceived stimuli also elicit activity in specific areas of the prefrontal and parietal cortices that may be associated with higher-level cognitive processes such as attention and working memory.

This article has focused exclusively on studies of awareness employing a single method, functional neuroimaging. In practice, findings from a single technique are never considered in isolation and identifying experimentally the neural correlates of the contents of conscious experience in humans depends on considering converging evidence from multiple methodologies. It is therefore appropriate to consider some of the important limitations of neuroimaging methodologies in order to understand the potential contribution of these data in the context of findings from other experimental techniques.

Causality

All the work discussed in this article has attempted to correlate changes in brain activity with changes in the contents of consciousness. Such correlational studies cannot determine whether such neural activity plays a causal role in determining the contents of consciousness. In particular, they cannot show whether the areas that have been identified in association with particular contents of consciousness are either necessary or sufficient for such conscious experiences to occur. The identification of necessary and sufficient brain activity requires explicit experimental manipulation of that activity. If manipulation of brain activity changes consciousness, then a causal role for that brain activity can be inferred. Manipulations of brain activity can be performed with direct electrical stimulation, using transcranial magnetic stimulation and as a consequence of brain lesions (although in humans this depends on the presence of pathological conditions and is not under direct experimental control!).

Report

The studies discussed in this article seek to relate neural activity to visual awareness, and in doing so they must distinguish the neural correlates of such activity from the merely unconscious concomitants, such as unconscious perception or behavior. Typically, keeping unconscious aspects of perception or behavior constant can allow this distinction to be made. For example, many of the studies reviewed here employed physically identical stimuli that elicited different conscious percepts. However, with very few exceptions, such studies also required that participants report whether they were aware of the stimulus. Because different behaviors are required to make different reports of awareness, it is possible that this might confound the interpretation of such studies. For example, when masked words are consciously identified and then named (as opposed to when they are not consciously identified and not named), then activity is observed in the frontoparietal cortex. However, such activity may reflect the (unconscious) processes associated with generating verbal output rather than the neural correlates of awareness. Although the differences in behavior elicited by the various reports of awareness are typically much more subtle, and, for example, frontal and parietal activsergenity associated with awareness can be observed in the absence of report or when performance confounds are controlled, further study of the unconscious correlates of behavior associated with report is indicated.

Conclusion

The most parsimonious account of currently available data is that the current contents of visual consciousness consist of a representation in the primary visual cortex and ventral visual pathway corresponding to the attributes represented in consciousness, together with activity in specific parietal (and perhaps) prefrontal structures. Future work must seek to more closely delineate the qualitative and quantitative differences between sensory representations of stimuli that either reach consciousness or remain unconscious, determine whether the frontal and parietal cortices play a causal role in representing the contents of consciousness, and more closely define the relationship between particular higher cognitive processes involved in awareness and corresponding patterns of brain activation.

See also: Binocular Rivalry; Cognition: An Overview of Neuroimaging Techniques; Consciousness: Neural Basis of Conscious Experience; Consciousness: Neurophysiology and Visual Awareness in; Consciousness: Philosophy; Consciousness: Theories and Models; Neuroimaging; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Alkire MT, Haier RJ, and Fallon JH (2000) Toward a unified theory of narcosis: Brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness. *Consciousness and Cognition* 9(3): 370–386.

Antognini JF, Buonocore MH, Disbrow EA, and Carstens E (1997) Isoflurane anesthesia blunts cerebral responses to noxious and innocuous stimuli: A fMRI study. *Life Sciences* 61(24): PL 349–354.

Baars BJ (2005) Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research* 150: 45–53.

Bahrami B, Lavie N, and Rees G (2007) Attentional load modulates responses of human primary visual cortex to invisible stimuli. *Current Biology* 17(6): 509–513.

Barnes J, Howard RJ, Senior C, et al. (1999) The functional anatomy of the mccollough contingent colour after-effect. *NeuroReport* 10(1): 195–199.

Beck DM, Rees G, Frith CD, and Lavie N (2001) Neural correlates of change detection and change blindness. *Nature Neuroscience* 4(6): 645–650.

Blake R and Logothetis NK (2002) Visual competition. *Nature Reviews Neuroscience* 3(1): 13–21.

Bonhomme V, Fiset P, Meuret P, et al. (2001) Propofol anesthesia and cerebral blood flow changes elicited by vibrotactile stimulation: A positron emission tomography study. *Journal of Neurophysiology* 85(3): 1299–1308.

Cannestra AF, Black KL, Martin NA, et al. (1998) Topographical and temporal specificity of human intraoperative optical intrinsic signals. *NeuroReport* 9(11): 2557–2563.

Carmel D, Lavie N, and Rees G (2006) Conscious awareness of flicker in humans involves frontal and parietal cortex. *Current Biology* 16(9): 907–911.

Dehaene S, Kerszberg M, and Changeux JP (1998) A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the United States of America* 95(24): 14529–14534.

Dehaene S, Naccache L, Cohen L, et al. (2001) Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience* 4(7): 752–758.

Dierks T, Linden DE, Jandl M, et al. (1999) Activation of Heschl's gyrus during auditory hallucinations. *Neuron* 22(3): 615–621.

Dueck MH, Petzke F, Gerbershagen HJ, et al. (2005) Propofol attenuates responses of the auditory cortex to acoustic stimulation in a dose-dependent manner: A fMRI study. *Acta Anaesthesiologia Scandinavica* 49(6): 784–791.

Ehrsson HH, Spence C, and Passingham RE (2004) That's my hand!: Activity in premotor cortex reflects feeling of ownership of a limb. *Science* 305(5685): 875–877.

Fang F and He S (2005) Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience* 8(10): 1380–1385.

Ffytche DH, Howard RJ, Brammer MJ, David A, Woodruff P, and Williams S (1998) The anatomy of conscious vision: An fMRI study of visual hallucinations. *Nature Neuroscience* 1(8): 738–742.

Fiset P, Paus T, Daloze T, et al. (1999) Brain mechanisms of propofol-induced loss of consciousness in humans: A positron

emission tomographic study. *Journal of Neuroscience* 19(13): 5506–5513.

Frith C, Perry R, and Lumer E (1999) The neural correlates of conscious experience: An experimental framework. *Trends in Cognitive Sciences* 3(3): 105–114.

Grill-Spector K, Kushnir T, Hendler T, and Malach R (2000) The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience* 3(8): 837–843.

Hadjikhani N, Liu AK, Dale AM, Cavanagh P, and Tootell RB (1998) Retinotopy and color sensitivity in human visual cortical area v8. *Nature Neuroscience* 1(3): 235–241.

Haynes JD, Deichmann R, and Rees G (2005) Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature* 438(7067): 496–499.

Haynes JD, Driver J, and Rees G (2005) Visibility reflects dynamic changes of effective connectivity between v1 and fusiform cortex. *Neuron* 46(5): 811–821.

Haynes JD and Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience* 8(5): 686–691.

Haynes JD and Rees G (2005) Predicting the stream of consciousness from activity in human visual cortex. *Current Biology* 15(14): 1301–1307.

He S, Cohen ER, and Hu X (1998) Close correlation between activity in brain area mt/v5 and the perception of a visual motion aftereffect. *Current Biology* 8(22): 1215–1218.

Heinke W, Fiebach CJ, Schwarzbauer C, Meyer M, Olthoff D, and Alter K (2004) Sequential effects of propofol on functional brain activation induced by auditory language processing: An event-related functional magnetic resonance imaging study. *British Journal of Anaesthesia* 92(5): 641–650.

Heinke W and Schwarzbauer C (2001) Subanesthetic isoflurane affects task-induced brain activation in a highly specific manner: A functional magnetic resonance imaging study. *Anesthesiology* 94(6): 973–981.

Hirsch J, De LaPaz RL, Relkin NR, et al. (1995) Illusory contours activate specific regions in human visual cortex: Evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America* 92(14): 6469–6473.

Huettel SA, Guzeldere G, and McCarthy G (2001) Dissociating the neural mechanisms of visual attention in change detection using functional MRI. *Journal of Cognitive Neuroscience* 13(7): 1006–1018.

Kerssens C, Hamann S, Peltier S, Hu XP, Byas-Smith MG, and Sebel PS (2005) Attenuated brain response to auditory word stimulation with sevoflurane: A functional magnetic resonance imaging study in humans. *Anesthesiology* 103(1): 11–19.

Kleinschmidt A, Buchel C, Zeki S, and Frackowiak RS (1998) Human brain activity during spontaneously reversing perception of ambiguous figures. *Proceedings: Biological Sciences* 265 (1413): 2427–2433.

Kourtzi Z and Kanwisher N (2000) Activation in human mt/mst by static images with implied motion. *Journal of Cognitive Neuroscience* 12(1): 48–55.

Kreiman G, Koch C, and Fried I (2000) Imagery neurons in the human brain. *Nature* 408(6810): 357–361.

Kumar A, Bhattacharya A, and Makhija N (2000) Evoked potential monitoring in anaesthesia and analgesia. *Anaesthesia* 55(3): 225–241.

Lau HC and Passingham RE (2006) Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the United States of America* 103(49): 18763–18768.

Laureys S, Faymonville ME, Degueldre C, et al. (2000) Auditory processing in the vegetative state. *Brain* 123(part 8): 1589–1601.

Laureys S, Faymonville ME, Luxen A, Lamy M, Franck G, and Maquet P (2000) Restoration of thalamocortical connectivity after recovery from persistent vegetative state. *Lancet* 355 (9217): 1790–1791.

Lee SH and Blake R (2002) V1 activity is reduced during binocular rivalry. *Journal of Vision* 2(9): 618–626.

Lee SH, Blake R, and Heeger DJ (2005) Traveling waves of activity in primary visual cortex during binocular rivalry. *Nature Neuroscience* 8(1): 22–23.

Leopold DA and Logothetis NK (1996) Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* 379(6565): 549–553.

Lou HC, Kjaer TW, Friberg L, Wildschiodtz G, Holm S, and Nowak M (1999) A 15o-h2o pet study of meditation and the resting state of normal consciousness. *Human Brain Mapping* 7(2): 98–105.

Luck SJ, Vogel EK, and Shapiro KL (1996) Word meanings can be accessed but not reported during the attentional blink. *Nature* 383(6601): 616–618.

Lumer ED, Friston KJ, and Rees G (1998) Neural correlates of perceptual rivalry in the human brain. *Science* 280(5371): 1930–1934.

Lumer ED and Rees G (1999) Covariation of activity in visual and prefrontal cortex associated with subjective visual perception. *Proceedings of the National Academy of Sciences of the United States of America* 96(4): 1669–1673.

Marois R, Yi DJ, and Chun MM (2004) The neural fate of consciously perceived and missed events in the attentional blink. *Neuron* 41(3): 465–472.

Martin E, Thiel T, Joeri P, et al. (2000) Effect of pentobarbital on visual processing in man. *Human Brain Mapping* 10(3): 132–139.

McGonigle DJ, Hanninen R, Salenius S, Hari R, Frackowiak RS, and Frith CD (2002) Whose arm is it anyway? An fMRI case study of supernumerary phantom limb. *Brain* 125(Pt. 6): 1265–1274.

Mendola JD, Conner IP, Sharma S, Bahekar A, and Lemieux S (2006) fMRI measures of perceptual filling-in in the human visual cortex. *Journal of Cognitive Neuroscience* 18(3): 363–375.

Mendola JD, Dale AM, Fischl B, Liu AK, and Tootell RB (1999) The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *Journal of Neuroscience* 19(19): 8560–8572.

Meng M, Remus DA, and Tong F (2005) Filling-in of visual phantoms in the human brain. *Nature Neuroscience* 8(9): 1248–1254.

Morris JS, Ohman A, and Dolan RJ (1999) A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences of the United States of America* 96(4): 1680–1685.

Moutoussis K and Zeki S (2002) The relationship between cortical activation and perception investigated with invisible stimuli. *Proceedings of the National Academy of Sciences of the United States of America* 99(14): 9527–9532.

Moutoussis K and Zeki S (2006) Seeing invisible motion: A human fMRI study. *Current Biology* 16(6): 574–579.

Muckli L, Kohler A, Kriegeskorte N, and Singer W (2005) Primary visual cortex activity along the apparent-motion trace reflects illusory perception. *PLoS Biology* 3(8): e265 [online].

Murray SO, Boyaci H, and Kersten D (2006) The representation of perceived angular size in human primary visual cortex. *Nature Neuroscience* 9(3): 429–434.

Myles PS, Leslie K, McNeil J, Forbes A, and Chan MT (2004) Bispectral index monitoring to prevent awareness during anaesthesia: The b-aware randomised controlled trial. *Lancet* 363 (9423): 1757–1763.

Naccache L, Gaillard R, Adam C, et al. (2005) A direct intracranial record of emotions evoked by subliminal words. *Proceedings of the National Academy of Sciences of the United States of America* 102(21): 7713–7717.

Nachev P and Husain M (2007) Comment on "detecting awareness in the vegetative state." *Science* 315(5816): 1221 [author reply 1221].

Niedeggen M, Wichmann P, and Stoerig P (2001) Change blindness and time to consciousness. *European Journal of Neuroscience* 14(10): 1719–1726.

O'Craven KM and Kanwisher N (2000) Mental imagery of faces and places activates corresponding stiimulus-specific brain regions. *Journal of Cognitive Neuroscience* 12(6): 1013–1023.

Ogawa T, Shingu K, Shibata M, Osawa M, and Mori K (1992) The divergent actions of volatile anaesthetics on background neuronal activity and reactive capability in the central nervous system in cats. *Canadian Journal of Anaesthetics* 39(8): 862–872.

Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, and Pickard JD (2006) Detecting awareness in the vegetative state. *Science* 313(5792): 1402.

Pasley BN, Mayes LC, and Schultz RT (2004) Subcortical discrimination of unperceived objects during binocular rivalry. *Neuron* 42(1): 163–172.

Pessoa L and Ungerleider LG (2004) Neural correlates of change detection and change blindness in a working memory task. *Cerebral Cortex* 14(5): 511–520.

Pins D and Ffytche D (2003) The neural correlates of conscious vision. *Cerebral Cortex* 13(5): 461–474.

Plourde G, Belin P, Chartrand D, et al. (2006) Cortical processing of complex auditory stimuli during alterations of consciousness with the general anesthetic propofol. *Anesthesiology* 104(3): 448–457.

Polonsky A, Blake R, Braun J, and Heeger DJ (2000) Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neuroscience* 3(11): 1153–1159.

Portas CM, Krakow K, Allen P, Josephs O, Armony JL, and Frith CD (2000) Auditory processing across the sleep-wake cycle: Simultaneous eeg and fMRI monitoring in humans. *Neuron* 28(3): 991–999.

Portas CM, Strange BA, Friston KJ, Dolan RJ, and Frith CD (2000) How does the brain sustain a visual percept? *Proceedings: Biological Sciences* 267(1446): 845–850.

Ress D and Heeger DJ (2003) Neuronal correlates of perception in early visual cortex. *Nature Neuroscience* 6(4): 414–420.

Rees G (2001) Seeing is not perceiving. *Nature Neuroscience* 4(7): 678–680.

Rees G, Frith CD, and Lavie N (1997) Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science* 278(5343): 1616–1619.

Rees G and Lavie N (2001) What can functional imaging reveal about the role of attention in visual awareness? *Neuropsychologia* 39(12): 1343–1353.

Rees G, Russell C, Frith CD, and Driver J (1999) Inattentional blindness versus inattentional amnesia for fixated but ignored words. *Science* 286(5449): 2504–2507.

Rees G, Wojciulik E, Clarke K, Husain M, Frith C, and Driver J (2000) Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain* 123 (Pt. 8): 1624–1633.

Rees G, Wojciulik E, Clarke K, Husain M, Frith C, and Driver J (2002) Neural correlates of conscious and unconscious vision in parietal extinction. *Neurocase* 8(5): 387–393.

Ritzl A, Marshall JC, Weiss PH, et al. (2003) Functional anatomy and differential time courses of neural processing for explicit inferred, and illusory contours: An event-related fMRI study. *Neuroimage* 19(4): 1567–1577.

Rodriguez E, George N, Lachaux JP, Martinerie J, Renault B, and Varela FJ (1999) Perception's shadow: Long-distance synchronization of human brain activity. *Nature* 397(6718): 430–433.

Rudolph U and Antkowiak B (2004) Molecular and neuronal substrates for general anaesthetics. *Nature Reviews Neuroscience* 5(9): 709–720.

Sakai K, Watanabe E, Onodera Y, et al. (1995) Functional mapping of the human colour centre with echo-planar magnetic resonance imaging. *Proceedings: Biological Sciences* 261(1360): 89–98.

Senior C, Barnes J, Giampietro V, et al. (2000) The functional neuroanatomy of implicit-motion perception or representational momentum. *Current Biology* 10(1): 16–22.

Silbersweig DA, Stern E, Frith C, et al. (1995) A functional neuroanatomy of hallucinations in schizophrenia. *Nature* 378(6553): 176–179.

Sterzer P, Haynes JD, and Rees G (2006) Primary visual cortex activation on the path of apparent motion is mediated by feedback from hmt+/v5. *Neuroimage* 32(3): 1308–1316.

Sterzer P, Russ MO, Preibisch C, and Kleinschmidt A (2002) Neural correlates of spontaneous direction reversals in ambiguous apparent visual motion. *Neuroimage* 15(4): 908–916.

Tong F and Engel SA (2001) Interocular rivalry revealed in the human cortical blind-spot representation. *Nature* 411(6834): 195–199.

Tong F, Nakayama K, Vaughan JT, and Kanwisher N (1998) Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* 21(4): 753–759.

Tononi G, Srinivasan R, Russell DP, and Edelman GM (1998) Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. *Proceedings of the National Academy of Sciences of the United States of America* 95(6): 3198–3203.

Tootell RB, Reppas JB, Dale AM, et al. (1995) Visual motion aftereffect in human cortical area mt revealed by functional magnetic resonance imaging. *Nature* 375(6527): 139–141.

Vuilleumier P, Sagiv N, Hazeltine E, et al. (2001) Neural fate of seen and unseen faces in visuospatial neglect: A combined event-related functional MRI and event-related potential study. *Proceedings of the National Academy of Sciences of the United States of America* 98(6): 3495–3500.

Weil RS, Kilner JM, Haynes JD, and Rees G (2007) Neural correlates of perceptual filling-in of an artificial scotoma in humans. *Proceedings of the National Academy of Sciences of the United States of America* 104(12): 5211–5216.

Wiedemayer H, Fauser B, Armbruster W, Gasser T, and Stolke D (2003) Visual evoked potentials for intraoperative neurophysiologic monitoring using total intravenous anesthesia. *Journal of Neurosurgical Anesthesiology* 15(1): 19–24.

Wunderlich K, Schneider KA, and Kastner S (2005) Neural correlates of binocular rivalry in the human lateral geniculate nucleus. *Nature Neuroscience* 8(11): 1595–1602.

Yi DJ, Woodman GF, Widders D, Marois R, and Chun MM (2004) Neural fate of ignored stimuli: Dissociable effects of perceptual and working memory load. *Nature Neuroscience* 7(9): 992–996.

Zeki S, Watson JD, and Frackowiak RS (1993) Going beyond the information given: The relation of illusory visual motion to brain activity. *Proceedings: Biological Sciences* 252(1335): 215–222.

Axon Guidance by Glia

A Faissner, Ruhr-University, Bochum, Germany

© 2009 Elsevier Ltd. All rights reserved.

Evidence for Conducive Functions of Astroglia

Astroglial Cells Constitute Axonal Growth Pathways *In Vivo*

Astrocyte-derived basal laminae Astrocyte surfaces were originally recognized as a favorable substrate for axon outgrowth. Müller cells, for example, produce a specialized structure toward the inner surface of the retina, the glial endfeet. Retinal ganglion cell axons elongate along the endfeet-derived basal lamina that contains characteristic extracellular matrix (ECM) components such as laminin-1, collagen IV, heparan sulfate proteoglycan (HSPG), and nidogen. The response of axonal growth cones to laminin-1 is regulated by integrins, heterodimeric receptors of the ECM.

Astrocyte monolayers Astrocyte monolayers constitute an excellent growth substrate for axons *in vitro*. For example, type I protoplasmic astrocytes with extended surfaces prepared from embryonic or perinatal central nervous system (CNS) tissues are more efficient than those obtained from later postnatal stages. Primary astrocytes are more efficient soon after plating and tend to lose beneficial properties with time. Subtypes may exist that are less supportive for axon extension. The neurite growth-promoting properties are also regulated by the spatial arrangement of astrocytes.

Astrocytes in three dimensions It has been observed that astrocytes that still support axon growth when used as monolayers may have inhibitory properties when assembled in tissuelike structures. For example, dorsal root ganglion axons confronting astrocytes lodged in three dimensions in a cellulose acetate tube will grow readily into tubes filled with embryonic but not with aged or matured astrocytes. The model in this regard mimics the dorsal root entry zone, a reputed stop area for centripetal axons in the adult. In this context, the astrocyte–meningeal cell interface may construct a growth barrier.

Multiple Control Mechanisms Regulate the Astrocyte-Dependent Establishment of Cortical Connections via the Corpus Callosum

The Corpus Callosum and the Glial Sling

Anatomy of the corpus callosum and the blueprint hypothesis The corpus callosum connects the left with the right hemisphere in the developing nervous system. In the mouse, it consists of approximately 3 million myelinated fibers that link corresponding regions of the cortices. More that 50 human genetic aberrations have been described that lead to some form of dysgenesis of the corpus callosum. Two populations of midline glia, the indusium griseum and the glial wedge, are necessary for the adequate guidance of axons in this process. The emerging glial structures construct a transient bridge of astrocytes that connects the left with the right hemisphere of the developing telencephalon. This glial bridge, also called the glial sling, supports the reciprocal growth of cortical axons, and the experimental interruption of the sling leads to the formation of acallosal mice. In this situation, the cortical connecting axons role up on either side of the cerebral midline and form the bundles of Probst, longitudinal fascicles of misdirected axons. Growth promotion of cortical axons can be restored by the implantation of nitrocellulose filters that are covered with embryonic astrocytes-derived membranes. The blueprint hypothesis of axon growth states that channels walled by astrocyte surfaces might provide a mechanical growth and guidance substrate for growth cones. Molecular specializations of both growth cone and astrocyte surfaces are presumably implicated in the regulation of these interactions.

Molecular bases of neuron–glia interactions The mechanistic concept of axon guidance has been modified and geared toward an interpretation which emphasizes molecular signals in the growth environment, the readout by specific growth cone-based receptors, and integration of these influences by signal transduction cascades which eventually modulate growth cone movements (Table 1). Thus, a conduit function of astroglia based on the chemorepellent slit-2 has been proposed as an additional guidance principle in the corpus callosum. The molecular analysis of the signaling system involved in corpus callosum formation has progressed and additional gene families have been identified. Thus, the wnt5 protein is expressed by cells of the glial wedge and plays a role in guidance of connecting axons on the contralateral side of the developing corpus callosum formation. Wnt5a-dependent guidance in this context is mediated by the protein tyrosine kinase Ryk that is required in the axon.

Cell Adhesion Molecules Play a Pivotal Role in Neural Cell Interactions

General Definition of Cell Adhesion Molecules

Cell adhesion molecules (CAMs) were first identified during the study of sponge aggregation and

Table 1 Families of adhesion molecules involved in axon guidance

Adhesion molecules	*Features*	*Functions*
Ig superfamily (IgSf)	At least one immunoglobulin (Ig-like) domain, often in combination with fibronectin type III domains	Ca-independent adhesion, either in the homophilic or in the heterophilic mode
L1CAM, NCAM	First IgSf members found in the nervous system	Axon growth and guidance, synapse formation, and plasticity; human L1CAM mutated in the human MASA syndrome that is associated with mental retardation
TAG-1	GPI-linked member of the IgSf	Axon fasciculation
Cadherin superfamily	Characterized by at least one cadherin domain	Primarily homophilic, Ca^{2+}-dependent adhesion; roles in sorting out of cells, synapse formation
N-cadherin	Classical cadherin	Neuron–neuron and neuron–astrocyte adhesion; elimination leads to premature death and cardiovascular malformations
Laminin-1	Founding member of the laminin gene family that comprises more than ten members; laminins are heteromultimers formed by α, β, and γ chains; several genes for each subunit have been distinguished	Excellent axon growth promoting substrate and, for this reason, obligatory cell substrate constituent in many cell culture protocols; as a component of basal lamina, laminin-1 promotes axon growth in the peripheral nervous system; the outgrowth promotion is read out via integrins
Integrins	Large family of heterodimeric receptors formed by α and β subunits; many genes for either type of subunit have been described	Integrins are receptors of extracellular matrix constituents; a subgroup interacts with the sequence RGD in target molecules; more than 20 Itg receptors have been identified
Netrin-1, -2	Homologies to the γ_1 subunit of laminin-1	Netrins are released by midline glia and build up a chemotactic gradient that attracts axons from commissural neurons

subsequently discovered in mammals. One distinguishes Ca^{2+}-dependent and Ca^{2+}-independent adhesion mechanisms that are served by separate gene families, the immunoglobulin (Ig) and the cadherin gene superfamilies. CAMs interact either in the homophilic mode with identical members of a gene family or in the heterophilic mode with distinct partners that may or may not belong to the same gene family.

The Cadherins and Ca^{2+}-Dependent Adhesion

N-cadherin was the first classical cadherin discovered in the CNS and it mediates neuron–neuron and neuron–astrocyte interactions. Classical cadherins contain five cadherin repeat motifs, calcium binding sites, and a transmembrane domain that result in an overall molecular mass of approximately 100 kDa. Subgroups of cadherin-related neuronal receptors and of protocadherins as characterized by the cadherin domain have been discovered, many of which are expressed in the CNS. There, expression patterns correlate with neuroanatomical networks, and current concepts propose a functional role in wiring and synaptogenesis within these systems. This interpretation is consistent with the fact that classical cadherins function in a homophilic, calcium-dependent manner. The constitutive cadherin domain motif has also been found in so-called unconventional cadherin-like molecules that play a role in epithelial cell junction formation or in *Drosophila* tumors (e.g., the gene *fat*). The cytoplasmic domain of classical cadherins interacts with specialized linker proteins called catenins. These are required for the functional activity of cadherins, and the component β-catenin is involved in signal transduction to the cell nucleus and is an important mediator of the wnt signaling pathway, which links cadherins to differentiation processes and to cancer pathology.

Immunoglobulin Superfamily

The Ig superfamily constitutes the other dominating family of CAMs in the nervous system. Its characteristic feature is the Ig domain that consists of 90–100 amino acids arranged in seven antiparallel β-pleated sheets that fold into a globular structure. In many molecules, the Ig loop is combined with one or several fibronectin type III (FNIII) domains, followed by a transmembrane domain. Some IgCAMs are connected to the membrane by a glycosylphosphatidylinositol (GPI) anchor that confers augmented mobility within the membrane plane. GPI-linked proteins are thought to be associated with transmembrane glycoproteins that convey signal transduction processes – for example, the neurexin family member Caspr/paranodin in the case of the IgCAM contactin. On the mechanistic level, Ig superfamily members mediate calcium-independent adhesion processes that may be either homophilic or heterophilic. Several neuronal adhesion molecules with pronounced expression

on axonal surfaces, such as L1CAM or TAG-1/axonin-1 CAM, have been grouped as AxCAMs, highlighting their prominent role in axon fasciculation.

IgCAM Interactions Launch Signal Transduction

Functional participation of IgCAMs in neurite outgrowth requires the activation of downstream signaling mechanisms, including the variation of intracellular calcium in the growth cone. The transduction mechanisms elicited by IgCAMs converge with those launched via N-cadherin and necessitate the basic fibroblast growth factor (bFGF) receptor that *cis* interacts with these CAMs within the membrane. Selected isoforms of NCAM are expressed by astrocytes *in vitro* and mediate neuron–astrocyte adhesion. With regard to heterophilic interactions, the small isoform of the receptor protein tyrosine phosphatase (RPTP)-β/ζ is expressed by astrocytes, a transmembrane protein that interacts with the neuronal adhesion molecule contactin, and other CAMs of the Ig superfamily. These examples illustrate the functional involvement of IgCAMs in conducive neuron–astrocyte interactions. Mutations of the human L1CAM lead to mental retardation, hypotrophy of the corticospinal tract, and hydrocephalus, emphasizing the importance of the Ig superfamily for development of CNS structures.

Integrins as Extracellular Matrix Receptors

The integrins constitute the third prominent gene family of membrane-based CAMs. These consist of α and β subunits that assemble to heterodimers and mediate the interactions of neural cells with ECM components. In some cases, interactions between integrins and Ig superfamily members have been reported, and the activation of integrins involves signal transduction pathways with small GTP-binding proteins such as *ras*.

Evidence for Segregating Functions of Glia

Gene Families That Mediate Axon Growth Inhibition Counterbalance Adhesion Molecule Gene Families

Growth cone collapse The investigation of neural cell interactions resulted in the identification of the Ig and cadherin superfamilies and of growth- and motility-promoting ECM constituents. However, in the 1980s it became clear that in addition to growth promotion, growth inhibitory molecules also contribute to the regulation of cell migration and growth cone movement (Table 2). The phenomenon of growth cone collapse had been described in the context of the interaction of sympathetic with retinal neurites, or when growth cones were confronted with myelin fractions or membrane preparations obtained from inhibitory territories. These interactions invariably resulted in collapse and retraction of the growth cone that remained paralyzed for 30–60 min *in vitro* prior to resuming growth and exploratory behavior. It was quickly realized that this inhibitory effect might equally affect guidance and inhibition of regeneration.

Eph kinases and ephrins The systematic investigation of the visual projection led to the identification of RAG (retinal axon guidance molecule), a GPI-linked protein with a gradient-like distribution in the tectum that proved homologous to a gene family subsequently renamed ephrins. One distinguishes the GPI-linked ephrins A, which interact with complementary EphA-type tyrosine kinases, from the transmembrane ephrins B, which recognize the EphB-type tyrosine kinases. Both groups contain a large number of both ephrin- and Eph-type kinase genes, and a certain degree of freedom in the mutual combinations has been noted. A remarkable trait of the Eph kinases and the corresponding ligands is that the molecules are expressed in reciprocal gradients, which is similar to the concept that Sperry developed in his chemoaffinity hypothesis that posited a chemical cell surface expressed code of positional information. Manipulation of expression levels of Eph genes resulted in a graded repositioning of axonal projections in the visual system. These studies provided impressive support both for the Sperry hypothesis and for the notion that the Eph–ephrin signaling system is an important contributor to the code. Whether the members of the pairs are expressed in neuronal and glial lineages, respectively, remains to be established.

Semaphorins and neuropilins Another gene family mediating growth cone collapse is the semaphorins, which comprise an increasing number of members and have been described in *Drosophila*, chicken, mouse, and human. A constitutive structural feature of these proteins is the sema domain of 500 amino acids that is shared by all family members. Semaphorin-dependent growth cone collapse is mediated by the plexin receptors, which can be categorized into subtypes A–D. In some cases, the plexins interact in the membrane in *cis* with the neuropilins NP1 or NP2 that act as coreceptors in these cases. The semaphorin IIIa dimer, for example, interacts with plexin A1 and neuropilin 1 to launch growth cone collapse using a signal transduction pathway that

Table 2 Gene families involved in repulsion and inhibition

Inhibitory molecules	*Features*	*Functions*
Semaphorins	Characterized by the semaphorin domain; seven structural classes reported in vertebrates and invertebrates	Sema3 molecules inhibit axon growth from several classes of neurons.
Plexins and neuropilins	Transmembrane semaphorin receptors of the growth cone	Detect and mediate Sema-dependent axon growth and guidance.
Ephrins	Family of genes that is divided into the GPI-linked ephrin A and the transmembrane located ephrin B molecules	Ephrins are the long sought-after ligands of the Eph kinases, originally described as orphan receptors.
Eph kinases	Tyrosine kinases located in the membranes of growth cones; one distinguishes Eph-A and Eph-B kinases; Eph kinases activated by interaction with complementary ephrin-A or -B ligands	In the CNS, activation of Eph kinases by ephrins in certain (but not all) combinations leads to growth cone collapse. Eph kinases and ephrins construct complementary gradients in the nervous system that encode positional information, for example, in the visual system.
Nogo proteins	Nogo proteins A–C have been described; they belong to the inhibitory components of myelin that are involved in the prevention of regeneration (in addition to MAG and OMgp)	Nogo proteins activate the Nogo receptor NgR that interacts with the low-affinity NGF receptor p75. Activation of these receptors results in growth cone collapse.
Chondroitinsulfate proteoglycans (CSPGs)	Composed of a core glycoprotein and at least one chondroitin sulfate carbohydrate chain; CSPGs include the lectican family, NG2, and phosphacan	CSPGs are found enriched in glial boundaries during development and in glial scars of the lesioned CNS. They are associated with axon growth inhibition. Axon growth stimulatory chondroitin sulfates have also been described.
Tenascin-C and tenascin-R	Composed of egf-type repeats and fibronectin-type 3 modules and represent the best-characterized glycoproteins of the neural extracellular matrix; tenascins assemble to multimers and are distributed in discrete patterns	Tenascins are inhibitory in certain situations and are capable of forming boundaries *in vivo* and *in vitro*. On the other hand, stimulation of axonal growth has also been reported. These molecules are thus best suited to mediate the ambivalent influences of astrocytes on axon growth and guidance.

involves small GTP-binding proteins, such as Rho and Rac1. The contributions of these inhibitory molecules in the context of neuron interactions with astrocytes remain to be worked out in detail.

Nogo and myelin-based inhibitors By comparison, more is known about the myelin sheath-derived inducers of growth cone collapse that are believed to underlie myelin-dependent inhibition of regeneration. These include the Nogo glycoproteins, with Nogo-A as the principal myelin-based inhibitor of growth, a member of the larger Reticulin gene family. Nogo-A contains a region called Nogo-66, a looplike structure which induces growth cone collapse by interacting with the complementary Nogo receptor NgR. NgR is GPI anchored to the growth cone membrane and part of a receptor complex that also contains the low-affinity NGF receptor p75. Interestingly, two other myelin components inhibitory to axon growth have been detected – the Ig superfamily member MAG (myelin-associated glycoprotein) and OMGP (oligodendrocyte–myelin glycoprotein). Both are also able to activate the Nogo receptor complex, which suggests a common downstream pathway of myelin-dependent inhibition. This pathway involves the downstream activation of RhoA family-type proteins that play a major role in growth cone collapse (Table 3).

Glial Boundaries between Developmental Compartments of the Central Nervous System

Rhombomeres of the hindbrain The brain stem is subdivided into rhombomeres, and the paraxial mesoderm is substructured into somites. The rhombomeres in the hindbrain emerge beginning with the six somite (6s) stage and can be identified as a series of eight swellings along the neuraxis which are separated by grooves at stage 16s. The axons of motor neurons located in pairs of rhombomeres are destined to innervate specific branchial arches and leave the even but not the odd-numbered rhombomeres at defined exit points. The motor nuclei of the cranial nerves and their motor tracts are confined to subsets of rhombomeres along the rostrocaudal axis. Cells of neighboring compartments do not mingle, and intercellular communication by gap junctions is limited to cells within a given compartment. The apparent segmentation of the hindbrain correlates with

Table 3 Gene families involved in choice decisions at the midline

Floor plate-based signals	*Growth cone-based receptors*	*Functions and comments*
Netrin-1, -2 (laminin gene SF; unc-6 in the nematode)	Dcc ('deleted in colorectal cancer', IgSF)	Netrin attracts commissural neurons toward the floor plate. The gene products unc-5 and unc-6 were found to control circumferential axon growth in the nematode.
Sonic hedgehog (Shh)	Patched and smoothened	Shh is a morphogen with fate-instructing functions in early development (e.g., induces the floor plate). It serves as chemoattractant at this stage.
Slit-1, -2, -3 (extracellular matrix components)	Robo-1, -2, -3 (IgSF members)	Slit builds a chemorepellent gradient to drive axons away from the floor plate after crossing. When the robo receptor is mutated, the axons circle incessantly back and forth across the midline ('roundabout' mutation in *Drosophila*). There are several *slit* and *robo* genes in mammals.
Sema3a	Neuropilin and plexin-A, -B, and -C receptors	Acts as chemorepellent that prevents return to the floor plate after crossing.
Ephrin B2	EphB kinases	The Eph–ephrin connection mediates membrane-based reciprocal avoidance behaviors in certain combinations.
NrCAM (IgSF member)	Tag-1/axonin-1 (IgSF member)	The NrCAM–Tag-1 interactions are required to enable the growth cone to cross the midline.

distinct expression patterns of transcription factors of the homeobox gene family that code positional identities in *Drosophila* and vertebrates. The limits of expression of these genes in the mouse spinal cord progress in a caudorostral direction and are strictly coherent with boundaries of rhombomeric pairs in ascending order. Also, other genes are expressed in register with rhombomere boundaries, such as various members of the Eph tyrosine kinase and complementary ephrin ligand gene families and zinc finger transcriptional activators such as *Krox-20*. The Hox code can be altered by treatment with retinoic acid, which caudalizes the spinal cord, including cranial nerve nuclei. These findings concur to qualify the rhombomeres as true compartments with homeotic identity.

The prosomeric model of the prosencephalon Analogous patterns of transcription factors at more rostral positions of the neuraxis have prompted the view that compartments may also exist in diencephalic or telencephalic structures. A detailed analysis of the distribution of numerous transcription factors, such as Otx2, Emx1, Emx2, and Gbx2, in the developing rostral CNS has founded the view that one compartment yields the mesencephalon and six prosomeres preconfigure the diencephalon and the telencephalon with its cortical hemispheres. A more detailed assessment of compartments of the CNS is beyond the scope of this article and is available elsewhere. Interestingly, early axonal pathways often follow the boundaries between transcription factor expression territories. This is also true for interrhombomeric boundaries, where the extension of axonal fiber pathways has been documented.

Specialized cells in the compartment interface Closer inspection of the cellular populations in boundary regions revealed a reduced interkinetic nuclear migration of neuroepithelial cells and an increased intercellular space. IgCAMs such as NCAM that are expressed in rhombomeres may mediate intercellular adhesion in this context. The boundary cells exhibit an unusual fan-shaped array and abut with their endfeet on both the pial and the ventricular surfaces. In zebra fish, a so-called glial curtain has been proposed to separate individual rhombomeres. It has been envisaged that the specialized boundary cells construct a privileged pathway for outgrowing axons. An instructive influence of glial cordones on axon pathways is supported by transplantation experiments in which fiber tracts follow ectopic boundaries. The mechanism and molecular bases of hypothesized boundary functions for axon guidance are unknown, and differential adhesion or local inhibition concepts are being discussed.

Midline Glia as a Signaling Center for Growth Cone Guidance Decisions

Genetic control of the midline in *Drosophila* In *Drosophila*, the midline of the ladderlike nervous system involves the so-called midline glia, which is controlled by a battery of genes. Upstream, *single-minded* regulates the fate determination of midline cells in general. Glial cells emerge under the regulatory influence of *spitz*, a homolog of transforming growth factor-α, and the transcription factor *pointed*, which intervenes in the expression of *gcm* (glial cell missing), an additional gene required for the generation of glial cells in *Drosophila*. Homologs of *gcm*

have been described, but they seem to play a different role in vertebrates. Elimination of the midline glia prevents the separation of commissural fibers and eliminates the longitudinal connections of the characteristic ladderlike nervous system. The commissural fibers of the system begin to extend between three pairs of midline glia cells and the pair of MP1 neurons. Separation of the commissural fiber systems involves migration of the midline glia in the correct direction. A genetic screen has led to the identification of a number of additional genes which control this migratory behavior, one of which, called *klötzchen*, seems to implicate the spectrin cytoskeleton of midline glia. In all cases, deficits in midline glia migration lead to errors in the separation of the connecting commissural fiber systems.

The midline glia in the optic nerve chiasm A comparable class of boundaries associated with glial cell types has been described in the midline of the developing nervous systems of vertebrates. Thus, glial cells separate the left and right axonal projection systems at the optic chiasm. Advancing axons are attracted by netrin, interact with this glial population, and are directed either to the ipsilateral or to the contralateral cortex, as required in the context of binocular vision. Several candidate molecules have been considered to be involved in the directional growth cone choice. Among these are the glycoprotein L1, the hyaluronate receptor CD44, and chondroitin sulfate proteoglycan(s), as visualized by the expression of chondroitin sulfate epitopes in boundary glia. Chondroitin sulfate proteoglycans (CSPGs) are thought to indicate inhibitory territories in the developing nervous system that are also secured by ephrins.

The floor plate of the vertebrate neural tube Analogous situations are also observed in the floor and roof plates of the developing spinal cord (Figure 1). In the latter case, the dorsal midline cells express keratan sulfate and chondroitin sulfate epitopes linked to proteoglycan core proteins. Thereby, the CSPGs outline a boundary region that is not traversed by the commissural axons. When chondroitin/keratan sulfate-expressing proteoglycans are exposed as patterned substrates alternating with the glycoprotein laminin-1, various cultured neurons and their processes avoid the proteoglycan-rich regions and grow out on the laminin-1 substrates.

In the ventral half of the spinal cord, commissural axons of the dorsal horn are attracted toward the floor plate that releases the chemoattractant netrin-1. Netrin-1 binds to the growth cone-based receptor deleted in colorectal cancer (*dcc*) of the Ig superfamily. This mechanism is highly conserved because it has already evolved in the nematode *Caenorhabditis elegans*, in which unc-5 guides circumferential axons via the unc-6 receptor – homologs of netrin-1 and dcc, respectively. Sonic hedgehog (*Shh*) represents a second chemoattractant that acts in a similar direction. When the axons reach the midline, the commissural fibers interact with the floor plate via an adhesive mechanism that involves the Ig superfamily members NrCAM as expressed in the midline and TAG-1/axonin-1 as heterophilic ligand in the growth cone membrane. The GPI-linked axonin-1/TAG-1 protein is downregulated after the crossing has occurred and L1CAM appears on the longitudinally oriented fibers. Concerted interactions of these Ig superfamily members thus seem to be required to regulate the crossing step. Subsequently, the axons lose responsiveness to the attractant netrin-1 and develop

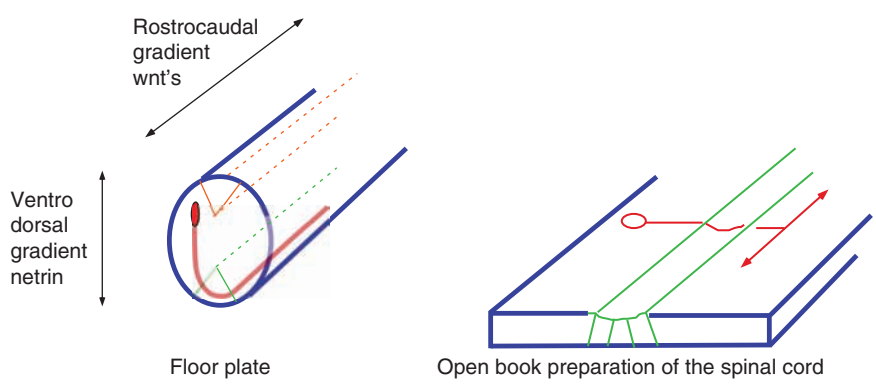

Figure 1 Axon guidance at the midline of the developing spinal cord. The figure shows the developing spinal cord of the rat at approximately embryonic day 13. At this stage, axons emanating from the commissural neurons migrate toward the ventral midline. There, axons contact the floor plate, midline glia (outlined in green) that defines an axon growth choice point. The growth cone migration and its behavior at the midline can be monitored *in vitro* with the so-called 'open book' preparation (right). Many genes are involved in growth cone guidance at the midline, some of which are listed in Table 3.

a sensitivity to the chemorepulsive signaling protein slit. This glycoprotein is released by midline glia and induces growth cone repulsion mediated by the Ig superfamily receptor roundabout (*robo*). *Robo* as well as *slit* were originally discovered in *Drosophila*, in which axons fail to stay on the ipsilateral side after crossing the midline, leading to the characteristic picture of incessantly circling axons. Several *slit* and *robo* genes have since been identified in mammals. A further inhibitory system is provided by semaphorins of the sema3 class and the complementary plexin and neuropilin receptor complexes NP1 and NP2, which mediate growth cone collapse. In addition, evidence has been presented that Eph kinase–ephrin interactions are also involved in the midline choice decision of the growing corticospinal projection. Thus, ephrin B3 is a constituent of the midline and prevents recrossing of these axons when they enter the gray matter after having migrated to the contralateral side of the spinal cord. Subsequently, the wnt proteins are involved in stimulating neurite outgrowth and regulating anterior–posterior guidance of commissural axons. These proteins utilize at least three signal transduction pathways and are implicated in many aspects of axon growth and interneuronal wiring (Table 3).

The Ambivalent Properties of Astrocytes Are Reflected by Extracellular Matrix Components That Embody Both Axon Inhibitory and Stimulatory Properties

Glycoproteins of the Extracellular Matrix

The pericellular space is organized by a superstructure of interacting glycoproteins and proteoglycans of the ECM. Astrocytes *in vitro* produce many of the ECM glycoproteins originally described in other tissues, namely fibronectin, laminin-1, vitronectin, thrombospondin, and tenascin-C. Laminin-1 is a functional component of astroglial endfeet in limiting membranes, for example, in the developing retina, and forms an excellent growth substrate for axon extension of many neuronal cell types. The chemodiffusible chemoattractants netrin-1 and netrin-2, which guide outgrowing commissural axons toward the floor plate of the midline in the spinal cord, are structurally related genes. Fibronectin has been found in association with blood vessels, a structure in which astrocytes contribute to the formation of the blood–brain barrier which isolates the CNS from the bloodstream.

The Tenascin Gene Family

Tenascin-C is transiently expressed by immature astrocytes in the developing CNS in which its distribution follows functional neuroanatomical subdivisions. For example, in the barrel field it delineates the emerging barrel field structure in layer IV. The glycoproteins of the tenascin gene family are characterized by structural motifs common to tenascin-C (Tnc), tenascin-R (Tnr), tenascin-X (Tnx), tenascin-Y (Tny), and tenascin-W (Tnw). A cysteine-rich N-terminus is followed by a series of egf-type repeats, fibronectin type III modules, and, finally, homologies to fibrinogen-β and -γ at the C-terminus. Different from this organization, a tenascin-like pair-rule gene in *Drosophila* contains the characteristic egf-type repeats but is devoid of other structural elements. The egf-type repeats of tenascins display a characteristic arrangement of cysteines that has also been found in the ECM protein reelin and is distinct from the one in the egf-type repeat modules of Notch or the laminin genes. The N-terminus links monomers in Tnr to trimers and in Tnc to hexamers under nonreducing conditions. As viewed with the electron microscope, the hexamer appears as a typical six-armed structure that has been designated hexabrachion and seems to be conserved during evolution. Two isoforms which are distinguished by one FNIII motif have been described in Tnr, a gene which is expressed in oligodendrocytes at later stages of development.

Multiple Isoforms of Tnc

Tnc possesses an alternative splice site between the fifth and the sixth FNIII module, where as many as six and nine additional FNIII repeats can be inserted in mouse and human Tnc, respectively. Up to 30 alternatively spliced variants in this region of Tnc have been revealed, approximately 50% of the theoretically possible 64 isoforms assuming a binary combinatorial code. In the human, theoretically 512 or 2^9 splice variants are conceivable on the ground of 9 potential alternatively spliced modules. Therefore, Tnc seems suited to specify pericellular microenvironments or to distinguish glial sublineages. Tnc is associated with numerous pathological conditions, including glial tumors.

Tnc Is a Multimodular and Multifunctional ECM Component

The characteristic hallmark of Tnc is its antiadhesive property for many cell types. In situations that expose a Tnc-rich environment alternating with laminin-1, the glycoprotein deflects growth cones and neuronal cell bodies, consistent with its boundary-like distribution in several CNS territories. On the other hand, homogeneous substrates containing Tnc promote neurite outgrowth of most neuronal cell types studied to date. Neurite outgrowth promotion could be mapped to the distal splice site that surrounds

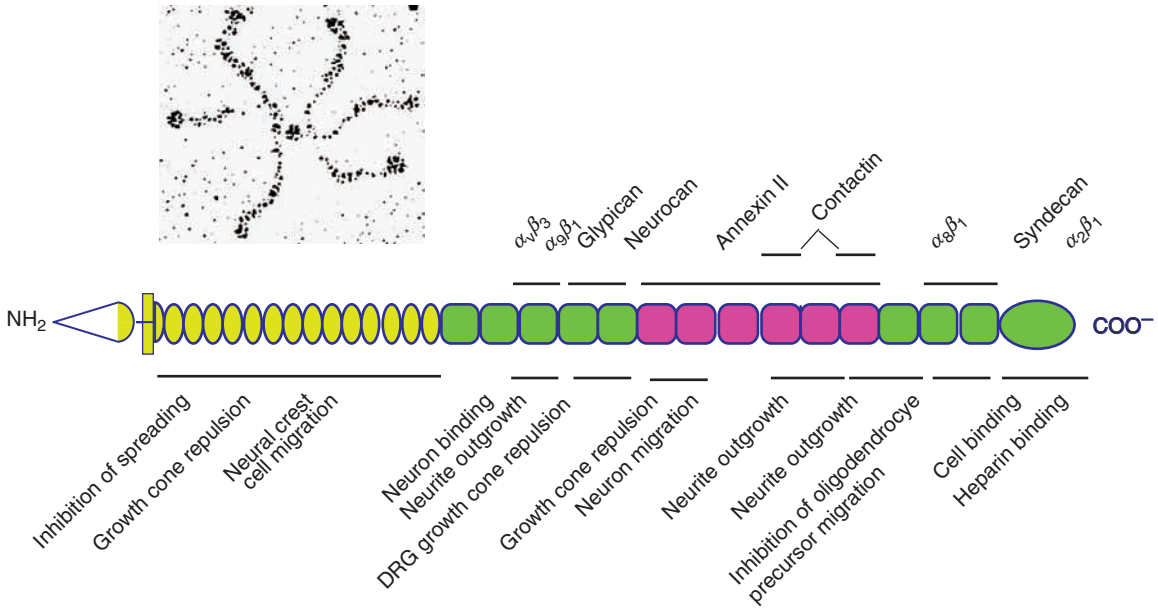

Figure 2 Structure–function relationship of tenascin-C. The glycoprotein of the extracellular matrix (ECM) tenascin-C is characterized by a modular structure with fibronectin type III domains (boxes), some of which are alternatively spliced (pink boxes). On the N-terminal end, the sequence comprises egf-type repeats (yellow elipses). Many functions have been ascribed to tenascin-C that are mediated by distinct receptors and/or ligands. The corresponding binding sites are underlined and complementary receptors are indicated. $\alpha_n\beta_m$ pairs of letters indicate specific integrin heterodimers.

cassette D. A site interfering with the motility of oligodendrocyte precursors was located to the FNIII module pair TNfn78, whereas the antiadhesive qualities could be mapped to other domains (Figure 2). Several Tnc receptors have been described, including the Ig superfamily member contactin, which mediates Tnc-dependent stimulation of neurite outgrowth, and the integrins $\alpha_v\beta_3$, $\alpha_7\beta_1$, $\alpha_8\beta_1$, and $\alpha_9\beta_1$. In the ECM, Tnc interacts with the proteoglycans phosphacan and neurocan. A number of studies on the Tnc −/− mutants suggest altered responses to stress, behavioral modifications, and deficits in the stem cell compartments of several organs, including the brain, during development, in the adult, and in response to lesion. These observations and the strong association of Tnc with human pathology in CNS cancer, hippocampal sclerosis, and various types of lesion warrant further studies of this versatile multifunctional glycoprotein. The antiadhesive glycoproteins of the CNS also comprise the thrombospondins, which have been found to mediate an astrocyte-derived signal for synaptic maturation.

Proteoglycans of the Extracellular Matrix

Key features of proteoglycans and heparan sulfate proteoglycans Proteoglycans represent the second class of ECM components expressed in the CNS. These components are defined as glycoproteins that carry at least one covalently linked glycosaminoglycan chain. One can distinguish HSPGs, which are mainly membrane bound, from CSPGs and keratan sulfate proteoglycans (KSPGs) of the nervous system, which are preferentially recovered from saline detergent-free extracts.

Two important HSPG subfamilies are the membrane-based syndecans and the GPI-linked glypicans. Central roles of the HSPGs may reside in the support of signaling processes in the context of tissue development. Thus, bFGF binds to a specific motif in heparan sulfate carbohydrate chains that expose the factor to its (the FGF receptor is a membrane bound tyrosine kinase) receptor. Similar mechanisms have been proposed for wnt signaling that plays important roles in axon growth and synapse formation.

CSPGs and KSPGs of the central nervous system With regard to CSPGs, the members of the lectican family – brevican, neurocan, versican, and aggrecan – have been identified in the CNS. These CSPGs possess a binding site for hyaluronic acid, a lectin-type sequence, and further distinct structural motifs. Versican is expressed by mature oligodendrocytes, whereas neurocan and aggrecan have been detected in neurons. The core glycoproteins may carry the HNK-1 epitope, a carbohydrate structure also expressed by neural recognition molecules, or other N-linked carbohydrates, for example, of the Lewis X-type, which are

recognized by specific monoclonal antibodies. Further CSPGs that have been described in the CNS with the help of specific monoclonal antibodies include CAT 301 and NG2, a marker of oligodendrocytes. CAT 301 upregulation during early embryonic development of the spinal cord depends on activation of the NMDA receptor, suggesting a role in plasticity. The possible roles of CSPGs in synaptic reorganization have been highlighted by the finding that injection of chondroitinase ABC, which degrades the particular glycosaminoglycans of CSPGs, restores plasticity of the adult visual cortex.

Inhibition of axon outgrowth by CSPGs On the functional level, many studies support the view that CSPGs inhibit axon outgrowth in an otherwise supportive environment. Thus, DRG axons extend profusely on a laminin-1-coated substrate but strictly avoid territories that have been replenished with CSPG preparations. Comparable findings have been obtained with defined CSPGs such as neurocan and versican and a variety of cell types, including neural crest, which led to the conclusion that CSPGs play a central role in axon growth inhibition and guidance. These observations motivated the analysis of CSPGs in the context of regeneration inhibition. Numerous studies concur that CSPGs are strongly upregulated by reactive astrocytes in a broad spectrum of lesion paradigms. Therefore, these components are considered an important inhibitory compartment that plays a pivotal role in the lack of regeneration of the CNS. This interpretation has gained support by the observation that injection of chondroitinase ABC into the lesioned spinal cord enhances plasticity and reactive sprouting and hence entails some improvement of the afferent sensory function. Several possibilities are conceivable to explain the mechanistic aspects of the inhibitory properties of CSPGs, and it is plausible that these components bind growth cone collapse, inducing molecules such as semaphorins to particular structural sequences in the glycosaminoglycan chains. On the other hand, CSPGs have been found associated with axon growth and regeneration in the peripheral nerve, which emphasizes that overall matrix composition and the lineage and age of the neurons involved need to be considered.

Phosphacan and Receptor Tyrosine Phosphatases of the Central Nervous System

The CSPG phosphacan/DSD-1-PG DSD-1-PG/phosphacan is one of the more abundant soluble CSPGs in postnatal mouse brain and is homologous to the secreted proteoglycan phosphacan from rat tissues. The GAGs of phosphacan are composed of chondroitin sulfate (CS)-A and CS-C motifs, a keratan sulfate chain, and the DSD-1-epitope. This unique structure was discovered with the MAb 473HD, requires the sulfation of the carbohydrate backbone, and contains CS-D dimers and dermatansulfate. The DSD-1-epitope displays neurite outgrowth promoting properties, which possibly involves its capacity to bind pleitrophin. Phosphacan is a splice variant and corresponds to the complete extracellular region of the largest isoform of the transmembrane receptor protein tyrosine phosphatase-β (RPTP-ζ/β). RPTP-ζ/β proteins occur as large or short receptors which possess a transmembrane domain and two cytoplasmic tyrosine phosphatase modules. The additional phosphacan short isoform (PSI) that corresponds to the N-terminal sequence has been described in the mouse, and several isoform variants have been found in *Xenopus*.

Receptor protein tyrosine phosphatases in neuron–glia interactions The different isoforms of RPTP-ζ/β are developmentally regulated, and astrocytes from various parts of the CNS express the short RPTP-ζ/β receptor. RPTP-ζ/β_{long} is expressed in the ventricular zone of the developing and the subventricular zone of the adult CNS and also by oligodendrocyte precursors. Neuronal expression has also been observed which may partially be due to PSI that is strongly expressed by cortical neurons. The spatiotemporal expression patterns of RPTP-ζ/β isoforms during development, maintenance, and pathology of the CNS have been correlated with cell–cell signaling, cellular proliferation, migration, differentiation, axon outgrowth, synaptogenesis, synaptic function, and tissue regeneration. Based on the prominent glial expression of phosphacan and RPTP-ζ/β receptors, the proteins have been considered as possible mediators of neuron–glia interactions. In the adult CNS, phosphacan occurs in the perineuronal nets of parvalbumin-expressing neurons, surrounding axon terminals and glial endfeet but not the synaptic clefts. It has been hypothesized that CSPGs associate with hyaluronic acid in these structures to build a neural ECM comparable to that in connective tissue. CSPGs in the perineuronal ECM may constitute an important element in limiting synaptic plasticity. On the functional plane, phosphacan interacts with the Ig superfamily members contactin/F3/F11, axonin-1/TAG-1, NrCAM, and NgCAM and hence might intervene in both homophilic and heterophilic interactions. Therefore, it is plausible to assume that IgSF constituents represent neuronal ligands of RPTP-ζ/β receptors expressed in glial membranes and serve as molecular mediators at the interface between these two cellular lineages. Interestingly, both RPTP-ζ/β

and the IgCAMs are linked to signal transduction pathways and hence will act back on the expressing cells in the context of reciprocal signaling mechanisms. In particular, the phosphotyrosine–phosphatase modules of RPTP-ζ/β may antagonize tyrosine kinase-based activities. ECM ligands of phosphacan include tenascin-C and tenascin-R. The integration into ECM superstructures might explain to some extent why the elimination of the phosphacan gene does not yield serious developmental deficits in mice.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Axonal Pathfinding: Extracellular Matrix Role; Cadherins and Synapse Organization; Cell Adhesion Molecules at Synapses; Glia and Stroke; Glial Responses to Injury; Glial Responses to Virus Infection.

Further Reading

Bandtlow CE and Zimmermann DR (2000) Proteoglycans in the developing brain: New conceptual insights for old proteins. *Physiological Reviews* 80: 1267–1290.

Brembeck FH, Rosario M, and Birchmeier W (2006) Balancing cell adhesion and Wnt signaling, the key role of beta-catenin. *Current Opinion in Genetics & Development* 16: 51–59.

Carulli D, Laabs T, Geller HM, and Fawcett JW (2005) Chondroitin sulfate proteoglycans in neural development and regeneration. *Current Opinion in Neurobiology* 15: 116–120.

Ciani L and Salinas PC (2005) WNTs in the vertebrate nervous system: From patterning to neuronal connectivity. *Nature Reviews Neuroscience* 6: 351–362.

Doetsch F (2003) The glial identity of neural stem cells. *Nature Neuroscience* 6: 1127–1134.

Faissner A, Heck N, Dobbertin A, and Garwood J (2006) DSD-1-Proteoglycan/Phosphacan and receptor protein tyrosine phosphatase-beta isoforms during development and regeneration of neural tisssues. *Advances in Experimental Medicine and Biology* 557: 25–53.

Gotz M and Huttner WB (2005) The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology* 6: 777–788.

Granderath S and Klambt C (1999) Glia development in the embryonic CNS of *Drosophila*. *Current Opinion in Neurobiology* 9: 531–536.

Joester A and Faissner A (2001) The structure and function of tenascins in the nervous system. *Matrix Biology* 20: 13–22.

Johnson KG, Ghose A, Epstein E, Lincecum J, O'Connor MB, and Van Vactor D (2004) Axonal heparan sulfate proteoglycans regulate the distribution and efficiency of the repellent slit during midline axon guidance. *Current Biology* 14: 499–504.

Kiecker C and Lumsden A (2005) Compartments and their boundaries in vertebrate brain development. *Nature Reviews Neuroscience* 6: 553–564.

Kleene R and Schachner M (2004) Glycans and neural cell interactions. *Nature Reviews Neuroscience* 5: 195–208.

Kullander K and Klein R (2002) Mechanisms and functions of Eph and ephrin signalling. *Nature Reviews Molecular Cell Biology* 3: 475–486.

Lemke G (2001) Glial control of neuronal development. *Annual Review of Neuroscience* 24: 87–105.

Lemke G and Reber M (2005) Retinotectal mapping: New insights from molecular genetics. *Annual Review of Cell and Developmental Biology* 21: 551–580.

Levine JM, Reynolds R, and Fawcett JW (2001) The oligodendrocyte precursor cell in health and disease. *Trends in Neurosciences* 24: 39–47.

Liu BP and Strittmatter SM (2001) Semaphorin-mediated axonal guidance via Rho-related G proteins. *Current Opinion in Cell Biology* 13: 619–626.

Orend G (2005) Potential oncogenic action of tenascin-C in tumorigenesis. *International Journal of Biochemistry & Cell Biology* 37: 1066–1083.

Rhodes KE and Fawcett JW (2004) Chondroitin sulphate proteoglycans: Preventing plasticity or protecting the CNS? *Journal of Anatomy* 204: 33–48.

Sandvig A, Berry M, Barrett LB, Butt A, and Logan A (2004) Myelin-, reactive glia-, and scar-derived CNS axon growth inhibitors: Expression, receptor signaling, and correlation with axon regeneration. *Glia* 46: 225–251.

Yamaguchi Y (2001) Heparan sulfate proteoglycans in the nervous system: Their diverse roles in neurogenesis, axon guidance, and synaptogenesis. *Seminars in Cell and Developmental Biology* 12: 99–106.

Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems

F Wang, Duke University Medical Center, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The functions and signaling mechanisms of many important axon guidance molecules may be described individually, but here we attempt to take a more systematic view to describe how various chemotropic guidance cues and adhesion molecules work together to help build precise connections in sensory systems, from the periphery to the brain. We focus on the cell surface/secreted molecules that regulate important pathfinding events in the vertebrates, especially in the mammalian sensory systems. Among the sensory systems, the pathfinding mechanisms are best understood in the visual and the olfactory systems, and to a lesser extent in the somatosensory system. The auditory and the gustatory systems, however, have not been very well characterized.

Building the Visual System

In vertebrate animals, visual stimuli are transmitted by retinal ganglion cells (RGCs) from the retina to subcortical relay stations and eventually to the primary visual cortex. The underlying pathway is formed during development by axons navigating through complex environments, involving decisions at multiple steps, in a remarkably stereotypical and precise manner.

Pathfinding and Topographic Mapping by RGCs

RGC axons project from the retina to the superior colliculus (SC) in mammals, or to the optic tectum (OT) in amphibians and avians, before reaching the lateral geniculate nucleus (LGN) of the thalamus. Significant progress has been made in revealing how RGC axons are guided to make several important choices: where to exit the retina, where to form the optic nerve, to cross or not to cross at the optic chiasm, and how to map the visual field topographically onto the SC and the LGN.

Finding and exiting the optic disk The first step in building the visual pathway is for the RGC axons to exit the retina and project into the optic nerve. RGC axons travel in radial routes inwardly toward the fovea of the retina, where the optic disk is located. Studies in zebra fish have revealed that RGCs express the chemokine receptor CXCR4, while the optic disk expresses the ligand stromal cell-derived factor-1 (SDF-1), and that SDF-1 attracts RGCs toward the optic disk (Figure 1(a)). In addition, cell adhesion molecules and bone morphogenetic protein (BMP) receptor type I, chondroitin sulfate proteoglycan (CSPG), and Sonic hedgehog (SHH) molecules have also been implicated in directing this process. Once the RGC axons reach the optic disk, they require netrin-1 to project into the optic nerve head (Figure 1(b)). In netrin-1 and DCC (deleted in colorectal cancer) mutant mouse embryos, many RGC axons fail to exit the eye and instead they project aberrantly into other regions of the retina, resulting in the hypoplasia of the optic nerve.

Forming the optic nerve and growing toward the optic chiasm After the RGC axons exit the eye, their growth along the optic nerve is guided in part by Slit proteins (Slit1 and Slit2). Slit1 and Slit2 are repellent molecules for RGC axons and are expressed in overlapping and complementary domains surrounding the optic nerve and adjacent to the optic chiasm (Figures 1(b) and 1(c)). In this way, Slit1 and Slit2 help establish repulsive barriers for a narrow corridor that channels retinal axons toward the optic chiasm. In the absence of both Slit1 and Slit2, retinal axons are defasciculated and project ectopically into the preoptic area. Slits also help to prevent RGC axons from growing back toward the other eye after they reach the chiasm. In zebra fish, mutations in the Slit receptor gene *Robo2* lead to random projections of RGC axons and complete disruption of optic nerve formation. The responses and sensitivities of RGC axons to Slit proteins appear to be regulated by heparan sulfotransferases.

Crossing or not crossing at the optic chiasm At the optic chiasm, RGC axons have to decide whether to stay on the ipsilateral side or to cross the midline to the contralateral side. The percentage of RGC axons crossing at the optic chiasm differs among different species. Fish and birds have no binocular vision and all of their axons cross the midline. In higher mammals, RGCs located in the nasal retina project their axons to the contralateral side, while axons from RGCs of the temporal retina stay ipsilaterally. In mammalian species with less binocular vision, such as mice, only a small subpopulation of RGCs in the ventrotemporal (VT) region of the retina will project their axons ipsilaterally. Studies in *Xenopus* and mice

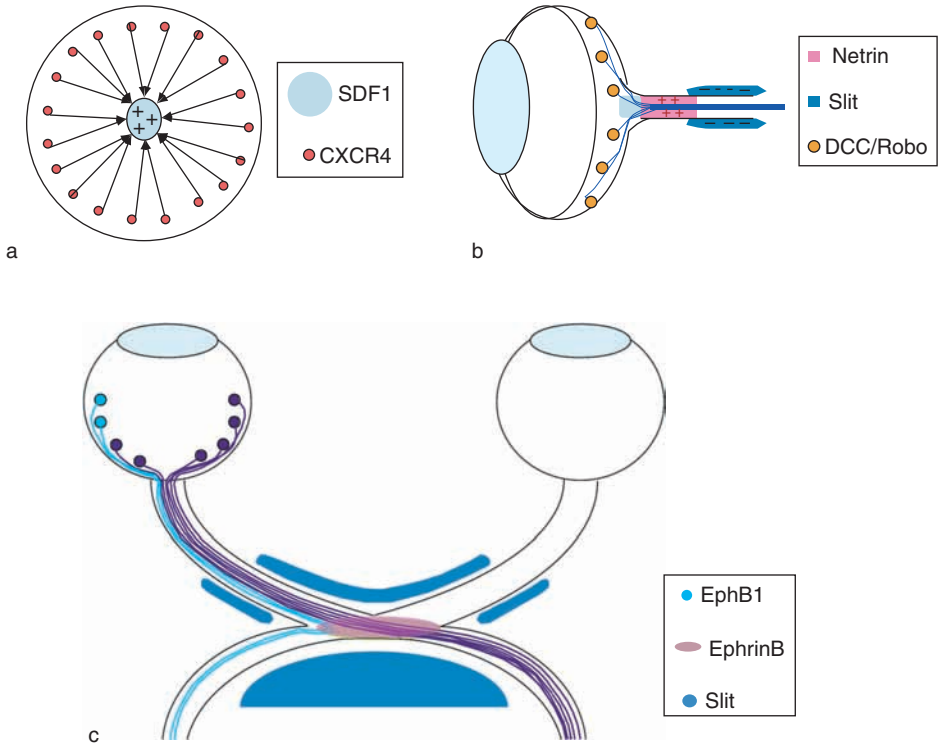

Figure 1 Molecules guide retinal ganglion cells at the optic disk, optic nerve, and optic chiasm. (a) Expressed stromal cell-derived factor-1 (SDF-1) attracts the chemokine (CXCR4)-expressing retinal ganglion cells toward the optic disk. (b) Netrin-1 attracts retinal ganglion cells into the optic nerve. Slit proteins prevent retinal ganglion cell axons from straying away. (c) Slits channel retinal ganglion cell axons toward the optic chiasm. EphrinB prevents EphB1-expressing retinal ganglion cells from crossing the optic chiasm.

have demonstrated that the repulsive ligand ephrinB is expressed at the optic chiasm. EphrinB prevents the crossing of ipsilaterally projecting RGC axons which express the receptor EphB1 (Figure 1(c)). In EphB1 null mice, ipsilateral projections are dramatically reduced and almost all RGC axons cross the midline. The expression of EphB1 in these noncrossing RGCs is likely to be regulated by the zinc-finger transcription factor Zic2 and the Lim-homeodomain factor Islet2.

Topographic mapping in the optic tectum/superior colliculus Leaving the optic chiasm, RGC axons project to several subcortical regions. The most prominent midbrain target is the optic tectum of fish, amphibians, and avians, or the SC of mammals. In the OT/SC, RGC axons form an ordered map of the visual field, the positions on the retina being topographically mapped onto the OT/SC along two orthogonally oriented axes: the temporal–nasal (T–N) axis of the retina maps along the anterior–posterior (A–P) axis of the OT/SC (Figure 2, upper panel), whereas the dorsal–ventral (D–V) axis of the retina maps along the lateral–medial (L–M) axis of the OT/SC (Figure 2, lower panel). The T–N to A–P mapping is primarily governed by the repulsive guidance of the EphA–ephrinA interactions. The EphA expressions in RGCs have a high-to-low gradient along the T–N axis, whereas the ephrinA expressions in the OT/SC show a low-to-high gradient along the A–P axis. These two countergradients together control either the sites of RGC axon termination in fish and amphibians, or the locations of interstitial collateral sprouting in avians and mammals (Figure 2, upper panel). The D–V to L–M mapping is achieved by functions of two sets of guidance cues, ephrinB and Wnt. EphrinB is expressed in a high-to-low gradient along the M–L axis in the OT/SC, while EphB is expressed in a low-to-high gradient along the V–D axis in the retina. High-level EphB signaling results in repulsion, while low-level signaling leads to attraction. EphB–ephrinB signaling results in a net attraction of all RGC axons to the medial part of the OT/SC. This is balanced by Wnt protein signaling. Wnt3 causes repulsion of RGC axons through the Ryk receptor, but attraction is through the Frizzled (Fzl) receptor. Wnt3 is expressed in a high-to-low gradient along the M–L axis in the OT/SC. Ryk is expressed in a low-to-high gradient along the V–D axis in the retina, while Fzl expression is equal along the V–D axis. The net result is that Wnt3 causes RGC axons to project toward the lateral part of the OT/SC. The combination of ephrinB and Wnt signaling is required for topographic mapping of RGC axons from the retina to the OT/SC (Figure 2, lower panel). This projection can also be refined by

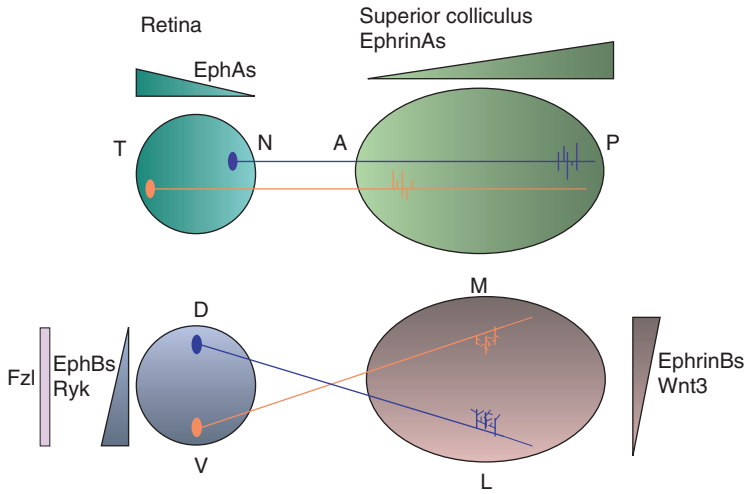

Figure 2 Topographic mapping of retinal ganglion cell projection in the superior colliculus, showing the temporal–nasal (T, N), the anterior–posterior (A, P), the dorsal–ventral (D, V), and the lateral–medial (L, M) axes. Schematic representations of graded receptors (EphA, EphB, Fzl, Ryk) and ligands (Wnt3 protein, ephrinA, ephrinB) that guide retinocollicular topographic mapping.

spontaneous correlated neural activity during a brief critical period.

Topographic mapping and segregation into eye-specific layers in LGN The lateral geniculate nucleus of the thalamus is the other main target of RGCs and is the relay station for visual input to the cortex. RGC axons again form ordered projections that map the visual field onto the LGN. Molecular mechanisms similar to those that govern the retinotopic map in the OT/SC also regulate the retinogeniculate projections. Gradients of ephrins in the LGN and countergradients of Ephs in RGCs regulate the topographic mapping in the LGN. Moreover, in mammals, input from the two eyes segregates into stereotyped eye-specific layers in the LGN. The precise pattern of segregation requires the function of EphrinAs, whereas the segregation of the eye input *per se* depends on neural activity.

From LGN to the Visual Cortex

The next connection in the visual pathway in mammals is for LGN neurons to relay the visual input that they receive from RGCs to the visual cortex. Compared with the pathfinding of RGCs, our understanding of the molecular mechanisms that guide the LGN neurons projections is quite limited. Here we summarize the current findings in this field, focusing on the guidance cues that are supported by *in vivo* evidence from mutant mice.

Getting to the cortex via thalamocortical projections LGN neurons follow precise pathways to relay information from the thalamus to the visual cortex. Their axons first project ventrally through the ventral thalamus. As they approach the dorsal border of the hypothalamus, they make a sharp turn and extend dorsolaterally through a narrow corridor, called the internal capsule (IC), into the ventral telencephalon. All thalamocortical axons (TCAs) likewise follow this path. Slit1 and Slit2 present in the hypothalamus prevent the TCAs from entering into the hypothalamus and also push TCAs away from the midline and toward the IC. Projecting into the IC requires recently identified 'corridor cells,' which express the membrane-bound form of neuregulin (Nrg1-CRD) that helps to attract and/or permit the growth of thalamic axons into the IC (Figure 3). Further extension of these axons to the cortex depends on the secreted form of neuregulin, containing an immunoglobulin-like (Ig) domain (Nrg1-Ig), which is expressed in the pallium and acts as a long-range chemoattractant (Figure 3). TCAs express the receptor Erb4 for both isoforms of neuregulin. Thalamocortical projection is severely disrupted in either the neuregulin or the Erb4 knockout mice in manners consistent with the roles of corridor cells and neuregulin in attracting TCAs. In addition, netrin-1 restricts the width of the internal capsule, and the semaphorin Sema6A is involved in the guidance of some TCAs.

Innervating and mapping in the visual cortex The molecular guidance cues that direct either the LGN axons toward the visual cortex or the subsequent axon branch projections into layer IV of the cortex are unknown at present. In terms of cellular events, axons initially accumulate and wait below the developing visual cortex in a zone called the subplate, which contains the first postmitotic neurons of the cerebral cortex, the subplate neurons. Subplate

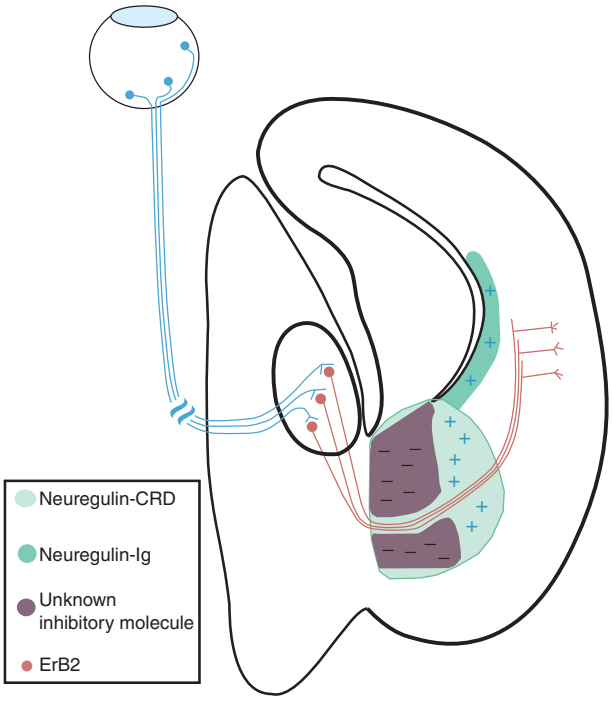

Figure 3 Thalamocortical axon projections are guided in part by neuregulin isoforms. Thalamocortical axons which express ErB2, the receptor for neuregulin, project through a narrow corridor toward cortex. The corridor is made attractive by the presence of a membrane-bound form of neuregulin (Nrg-CRD). In addition, a secreted form of neuregulin containing an immunoglobulin-like (Ig) domain (Nrg-Ig) acts as a long-range attractant for these axons.

neurons have been shown to play important roles in controlling where LGN axons wait and where they project, but the identities of the guidance molecules expressed by subplate neurons are unknown. LGN axons also form topographic maps of the visual field in the visual cortex. Here again, interactions between gradients of EphA/ephrinA direct the formation of the map, but they are not the only factors. Moreover, activity-dependent processes are likely to play a major role in shaping the visual connectivity in the cortex.

Building the Olfactory System

Odorants are detected by olfactory sensory neurons (OSNs) residing in the nose of mammals. OSNs project their axons to the olfactory bulb (OB), the first relay station in the olfactory pathway, and form a sensory map that encodes the quality of odorous chemicals. Mitral (and tufted) cells in the olfactory bulb project to multiple olfactory centers via the lateral olfactory tract (LOT). In the following sections, we review receptors and cues that have been shown to play important roles in guiding the OSN projection and the formation of LOT.

Formation of an Olfactory Sensory Map in the Olfactory Bulb

In the nasal epithelium of mammals, each OSN expresses only one odorant receptor (OR) from a family of about 1000 genes that all encode seven-transmembrane proteins. All neurons expressing the same OR, although randomly distributed in a broad circumferential zone along the dorsoventral axis in the epithelium, project convergently to a pair of glomeruli, each located at a stereotyped position on one side of the olfactory bulb. In this way, the identity of a chemical group (or groups) that is recognized by a given OR is mapped onto two anatomically conserved locations in the bulb (Figure 4). How do neurons expressing a specific OR find their target with such precision?

ORs play important roles in OSN targeting. Since the choice to express a given OR is tightly linked to the choice of the axonal projection site in the olfactory bulb, a model in which ORs dictate OSN axon targeting was proposed and supported by several lines of evidence. mRNAs of ORs are found in the OSN axons. OR proteins are concentrated not only in dendrites (to detect odorants) but also on axon termini in the glomeruli. Swapping either the entire coding sequence or the DNA sequence for a few amino acids of one OR with those of another resulted in altered OSN axonal convergence points in the bulb.

How might ORs influence where OSNs form the glomeruli? Two recent studies have shed a light on this puzzle. The first study strongly suggests that ORs signal through G-proteins to regulate intracellular cyclic adenosine monophosphate (cAMP) levels and that different cAMP concentrations determine the differential expressions of axon guidance molecules along the A–P axis. All OSNs express the olfactory-specific G-protein, G_{olf}, throughout their life span. In addition, they express the generic α subunit of G-proteins, G_s, at younger stages. ORs are coupled to these stimulatory G-proteins through a conserved DRY (Asp-Arg-Tyr) motif in their cytoplasmic loop after the third transmembrane domain. Mutating this G-protein binding site on a given OR, I7, caused the mutant ($I7_{DRY}$) OSNs to project diffusely in a broad domain in the olfactory bulb and failure to form glomeruli. This lack of axonal convergence can be partially reverted by expressing either a constitutively active G_s or a constitutively active cAMP response element-binding protein (CREB) in $I7_{DRY}$ neurons. Interestingly, co-expressing a dominant-negative protein kinase A (PKA) with the wild-type I7 results in these neurons projecting to a drastically anteriorly shifted position. In contrast, putting either a constitutively active G_s or a constitutively active PKA into

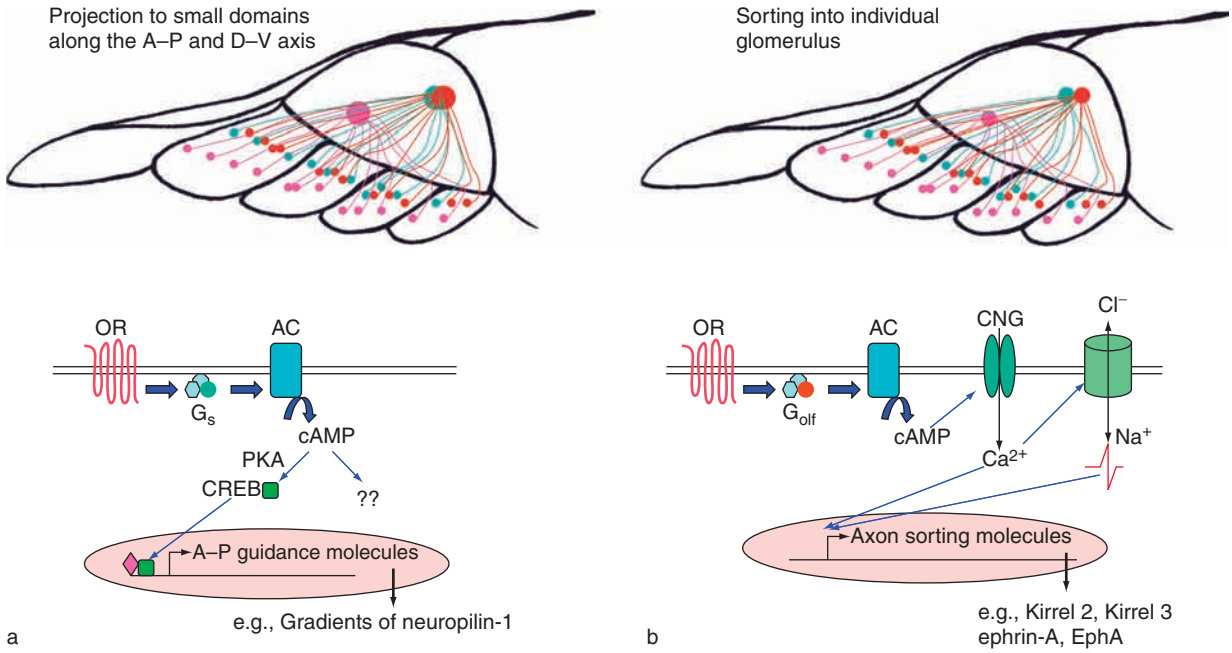

Figure 4 Formation of the olfactory sensory map is guided by odorant receptors (ORs). (a) ORs activate G-protein (G_s)signaling/cAMP signaling to guide the projection of olfactory sensory neurons along the anterior–posterior (A–P) axis and dorsal–ventral (D–V) axis to a small domain in the olfactory bulb. (b) Odorant receptors activate olfactory-specific G-protein (G_{olf}) signaling to induce calcium entry and neuronal depolarizations. These neuronal activities in turn regulate the expression of axon sorting molecules that help sort olfactory sensory axons into distinct glomeruli. AC, adenylate cyclase; CNG, cyclic nucleotide gated channel; cAMP, cyclic adenosine monophosphate; CREB, cAMP response element-binding protein; PKA, protein kinase A.

I7-OSNs, causes them to converge at a much more posterior location in the olfactory bulb. These results led the authors to propose a model in which ORs instruct OSN targeting indirectly via OR-derived cAMP signals (most likely by activating G_s) and that different amounts of cAMP regulate the expression of A–P axon guidance molecules, probably through activating CREB (Figure 4(a)). Indeed, the authors have found that expression of the neuropilin-1 receptor positively correlates with the cAMP level, and that neuropilin-1 is expressed in a low-to-high A–P gradient in the glomeruli layer of the olfactory bulb. However, this model also implies that each OR can determine a unique level of cAMP, a hypothesis that remains to be tested. Moreover, since cAMP has been shown to have profound effects on growth cone turning responses, it is conceivable that OR-derived cAMP also control OSN growth cones navigation locally.

Another study used transgenic mice in which the majority of OSNs express one particular OR. This has led to the discovery that ORs also control the expression of the homophilic adhesive molecules (such as the Ig domain containing Kirrel2/Kirrel3 proteins) and the repulsive molecules (such as ephrinA5/EphA5) in an OR-specific and activity-dependent manner (Figure 4(b)). Different ORs, through OR-evoked activity, determine the 'on' or 'off' expression of certain cell adhesion or repulsion molecules, such that different classes of OSNs have distinct combinatorial adhesive/repulsive codes. Once OSNs project to a narrow domain in the olfactory bulb, the adhesive/repulsive molecules then control the sorting of olfactory axons expressing different ORs into different glomeruli within that domain, by both homophilic and mutual repulsive interactions. At the present, it is not known how ORs control the expression of these sorting molecules through OR-dependent neuronal activity. Neural activity may be spontaneous and random, as spontaneous activity is known to be required for the formation of the olfactory map. Spontaneous and odor-evoked activities are also required for the maintenance of the olfactory sensory map.

From the Olfactory Bulb to Higher Olfactory Centers

Mitral and tufted (M/T) cells are the projection neurons in the olfactory bulb; they send axons through the lateral olfactory tract onto several structures of the olfactory cortex. M/T axons are known to wait in the LOT (for about 2 days in the mouse) before sending collateral branches to the olfactory cortex. M/T cells with dendrites that receive input from one particular glomerulus project collaterals to multiple cortical areas in a highly distributed and complex, but not random, manner. Because the organizational

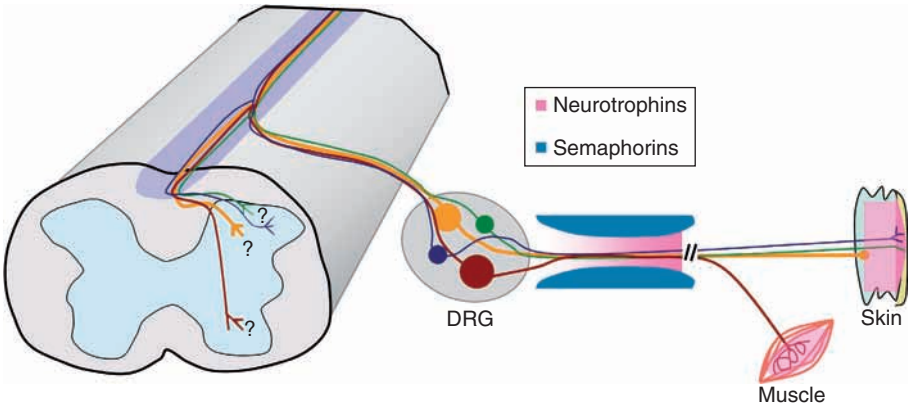

Figure 5 Peripheral projections of somatosensory neurons are guided in part by neurotrophins and semaphorins. Schematic representation of axonal projections of dorsal root ganglion (DRG) neurons. The molecules that guide the central axons into different layers of the spinal cord are unknown and are depicted as '?' in the drawing. Two classes of molecules, neurotrophins (including nerve growth factors, brain-derived neurotrophic factor, neurotrophins 3, 4, and 5, and glial-derived neuronal factors) and semaphorins, direct the peripheral axon outgrowth of sensory neurons by acting as chemoattractants and surround-repellents, respectively. Neurotrophins also promote the innervation, axon arborization, and sensory ending formation of DRG neurons in the target organs such as skin and muscle.

principle of the M/T projections in olfactory cortices is unclear, studies of this pathfinding process have largely focused on the formation of the LOT.

M/T cell projection to and along the LOT requires both short-range and long-range guidance cues. The short-ranged cues are believed to be at the surface of specialized cells called LOT cells. These cues help guide the M/T axons into the LOT and promote their elongation, although their molecular identities are unknown. M/T axonal projections in the LOT are also guided by long-range chemorepulsive molecules expressed in the septum, such that the axons avoid innervating septum regions *in vivo*. Slit1 and Slit2 together account for the repulsive ligands expressed in septum that repel M/T axons expressing Robo receptors. In Slit1 and Slit2 double-mutant embryos, M/T projection is completely disorganized, with many axons entering the septum region.

Building the Somatosensory System

The vertebrate somatic sensory system transmits to the brain information about physical stimuli that the body experiences. The stimuli can be painful, thermal, mechanical, or proprioceptive. Each of these modalities is detected by distinct types of primary sensory neurons and is processed by different central pathways. Although it is generally true that the body surface is topographically mapped onto subcortical processing centers and ultimately to the somatosensory cortex, surprisingly little is known about the molecular mechanisms that control the establishment of the somatosensory map and the neural circuit for each distinct modality.

The cell bodies of somatosensory neurons reside in a series of ganglia located outside the spinal cord (dorsal root ganglia) or the hindbrain (trigeminal ganglia), with the exception of the facial proprioceptive neurons, the cell bodies of which reside in the trigeminal mesencephalic nuclei inside the brain stem. Each sensory neuron grows two major branches stemming from a unipolar axon: a peripheral axon that innervates a specific body target, such as skin, viscera, or muscles, and a central axon that projects into the spinal cord or brain stem and forms specific synapses (Figure 5). Therefore, each neuron has two pathfinding tasks, one to the periphery and the other to the central nervous system (CNS), and the two processes must be coordinated such that the body is faithfully mapped onto the brain. The molecular mechanisms that guide the central axons of sensory neurons are largely unknown at present, but significant progress has been made in understanding the guidance of peripheral somatic sensory axons.

Peripheral Growth and Projection of Somatic Sensory Neurons Are Guided by Neurotrophins and Semaphorins

The growth of peripheral sensory axons follows stereotyped trajectories; this process is partly guided by neurotrophins (NTs) and semaphorins (Figure 5). NTs (including nerve growth factor, brain-derived neurotrophic factor, NT3, NT4, NT5, and glial-derived neuronal factor) are expressed and secreted by peripheral target tissues. They exert many biological effects on sensory neurons, acting not only as survival and differentiation factors, but also as chemoattractive factors for sensory axons. Both *in vitro* and

in vivo experiments have demonstrated that the chemotropic effects of NTs (for promoting axon elongation and target innervation) can be separated from their chemotrophic (survival) effects. For example, nerve growth factor-soaked beads, when embedded *in vivo* in embryos, can attract sensory axons growing toward them. The peripheral axons grow in fascicles, forming nerve bundles. This fasciculation is regulated by class III secreted semaphorins, which are expressed in a pattern surrounding the axon pathways, and sensory neurons express receptors that recognize these repellents (i.e., neuropilins 1 and 2 and plexins A3 and A4). Mice deficient in any of these receptors or semaphorin3A have severely defasciculated sensory axons. Besides NTs and semaphorins, little is known about the peripheral target selection and innervation process of sensory axons.

Summary

In summary, we have briefly reviewed the axon guidance molecules that work in concert to orchestrate the pathfinding processes that lead to the formation of visual, olfactory, and somatosensory systems. A general scheme is emerging, in that sensory neurons form topographic maps in their projections to the brain, such that sensory information is transformed into distinct spatial activities for the brain to process. These maps are guided first roughly by molecular gradients, and then by refinement with neuronal activity-dependent axon sorting. En route to their targets, neurons are guided to project through stereotyped pathways. These pathways are laid down by a combination of attractive and repulsive cues that usually result in a narrow channel for axons to navigate.

See also: Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Guidance Cues and Guidepost Cells; Axon Guidance by Glia; Axonal Regeneration: Role of Growth and Guidance Cues; Fovea: Primate; Olfactory Cortex: Comparative Anatomy; Olfactory Neuron Patterning and Specification; Olfactory Coding; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Optic Nerve, Optic Chiasm and Optic Tracts; Optic Tectum: Development and Plasticity; Retina: An Overview; Retinal Ganglion Cells: Receptive Fields; Retinal Development: An Overview; Retinal Development: Cell Type Specification; Somatosensory Plasticity; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging.

Further Reading

Deiner MS, Kennedy TE, Fazeli A, et al. (1997) Netrin-1 and DCC mediate axon guidance locally at the optic disc: Loss of function leads to optic nerve hypoplasia. *Neuron* 19: 575–589.

Imai T, Suzuki M, and Sakano H (2006) Odorant receptor-derived cAMP signals direct axonal targeting. *Science* 314: 657–661.

Li HS, Chen JH, Wu W, et al. (1999) Vertebrate Slit, a secreted ligand for the transmembrane protein Roundabout, is a repellent for olfactory bulb axons. *Cell* 96: 807–818.

Li Q, Shirabe K, Thisse C, et al. (2005) Chemokine signaling guides axons within the retina in zebrafish. *Journal of Neuroscience* 25: 1711–1717.

Lopez-Bendito G, Cautinat A, Sanchez JA, et al. (2006) Tangential neuronal migration controls axon guidance: A role for neuregulin-1 in thalamocortical axon navigation. *Cell* 125: 127–142.

McLaughlin T and O'Leary DD (2005) Molecular gradients and development of retinotopic maps. *Annual Review of Neuroscience* 28: 327–355.

Nguyen-Ba-Charvet KT, Plump AS, Tessier-Lavigne M, et al. (2002) Slit1 and Slit2 proteins control the development of the lateral olfactory tract. *Journal of Neuroscience* 22: 5473–5480.

Patel TD, Jackman A, Rice FL, et al. (2000) Development of sensory neurons in the absence of NGF/TrkA signaling *in vivo*. *Neuron* 25: 345–357.

Pfeiffenberger C, Yamada J, and Feldheim DA (2006) Ephrin-As and patterned retinal activity act together in the development of topographic maps in the primary visual system. *Journal of Neuroscience* 26: 12873–12884.

Plump AS, Erskine L, Sabatier C, et al. (2002) Slit1 and Slit2 cooperate to prevent premature midline crossing of retinal axons in the mouse visual system. *Neuron* 33: 219–232.

Schmitt AM, Shi J, Wolf AM, et al. (2006) Wnt-Ryk signalling mediates medial-lateral retinotectal topographic mapping. *Nature* 439: 31–37.

Serizawa S, Miyamichi K, Takeuchi H, et al. (2006) A neuronal identity code for the odorant receptor-specific and activity-dependent axon sorting. *Cell* 127: 1057–1069.

Wang F, Nemes A, Mendelsohn M, et al. (1998) Odorant receptors govern the formation of a precise topographic map. *Cell* 93: 47–60.

Williams SE, Mason CA, and Herrera E (2004) The optic chiasm as a midline choice point. *Current Opinion in Neurobiology* 14: 51–60.

Yaron A, Huang PH, Cheng HJ, et al. (2005) Differential requirement for plexin-A3 and-A4 in mediating responses of sensory and sympathetic neurons to distinct class 3 semaphorins. *Neuron* 45: 513–523.

Axon Guidance: Guidance Cues and Guidepost Cells

L Ma, University of Southern California, Los Angeles, CA, USA
K Shen, Stanford University, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Guidepost cells, also called landmark cells, were originally used to refer to specific cells in the developing insect limb that help peripheral nerve cells find their central targets. As vividly depicted by their name, these cells provide local cues to guide pioneer neurons to navigate through an unfamiliar territory in embryonic tissues. In combination with other cellular mechanisms, such as selective adhesion, they ensure the proper development of stereotypic neural connections that underlie complex behavior.

Although they have been historically studied in insects, guidepost cells have been demonstrated in many organisms as well as in both peripheral and central nerve systems. In fact, they now take a broader meaning and more appropriately refer to the intermediate targets in the course of neural development. In addition to guiding axons, they provide local guidance cues to control many other processes during nerve development. This article reviews the historical studies in insect cells, their role in midline crossing, and their emerging function in synapse formation.

Historical Studies of Guidepost Cells in Grasshoppers

The concept of 'guidepost cells' was first suggested by Australian biologist Michael Bate in his seminal study of peripheral pioneer neurons in grasshopper *Locusta migratoria*. In insects, peripheral sensory neurons are produced by a small number of epidermal cells at the body surface. Most of these cells invade their central targets by hopping along the preexisting nerve fibers that are established by pioneer neurons early in development. So how do pioneer neurons connect to the central nervous system (CNS) in the first place? To understand this, Bate followed them in the developing antenna and limbs by both light and electron microscopy. He discovered that these neurons, also with cell bodies born at the edge of each appendage, have to extend their axons through the entire length of the antenna or the limb to reach the CNS. Interestingly, through the course of their journey, not all axons follow a linear path, the shortest route. Instead, some of them, especially those in the limb, take an indirect route and change growth directions several times before reaching the central target. Based on this observation, Bate concluded that the connection path is determined by the interaction of the pioneer axons and the extrinsic cues in the limb. He suggested that these cues are provided by a group of cells, 'the signpost,' that "occur consistently at intervals along the developing appendage and seem to provide a series of stepping stones between the tip of the limb and its base."

The use of 'guidepost cells' or 'landmark cells' was adopted in subsequent studies of peripheral pioneer neurons in grasshopper limb buds, in which the simple nerve connection and the easy access for experimental manipulation provided an attractive system to further test this idea. Using newly developed staining techniques, many more cells have been found that can potentially serve as guidepost functions in the insect limb. They have some characteristic features, which were later used as major criteria to define guidepost cells in other systems: (1) they are located along the route of pioneer axonal path; (2) they are contacted by the growth cone of pioneer axons; (3) they are separated from each other but within reach of growth cone filopodia; (4) they are distinct from neighboring cells because they can be labeled by specific antisera on the surface, and they do not line up to form a continuous path for pioneer growth cones to follow; and (5) they form a special connection with axons because dye can pass from pioneer neurons to these cells. These physical and morphological attributes lead to a 'guidepost hypothesis,' which suggests that the placement of a series of distinctive cells is used to guide pioneer neurons along their trajectory from a distance.

The hypothesis was later tested in a cell ablation experiment, in which the putative guidepost cells that can be recognized immunologically were selectively killed with high-intensity light. The removal of guidepost cells in the developing grasshopper limb buds caused the pioneer neurons to wander away from their normal axonal pathway, and often resulted in the formation of multiple branches. Therefore, cells in the embryonic tissues are present to guide pioneer neurons during early development.

Guidepost Cells and Guidance Cues in Axon Guidance

Following the initial studies in the grasshopper, guidepost cells have also been described in several other invertebrates, including leech, moth, and fruit fly. Although the use of the term has been limited to these simple systems, the cellular studies of axonal pathfinding in grasshoppers have provided two important concepts that have simplified our understanding of

how complex neural pathways are connected. First, they have revealed that the long and sometimes irregular axonal trajectories can be broken down into short segments, and at each segment axonal growth is regulated by local cues at choice points. These choice points contain a cluster of individual cells as the guidepost cells in insects or often a group of functionally specialized cells, which are referred to as intermediate targets in vertebrates. Second, they have shown that complex neural networks are established in two phases, by pioneer neurons that enter the axon-free environment in early development and by later born neurons that face a more complex environment filled with intertwined nerve fibers from early projecting neurons. Whereas the pioneer neurons are guided by local guidance cues, those later developing axons can follow preexisting nerve fibers by selective fasciculation, a cellular mechanism that has also been well studied in grasshoppers.

How do guidepost cells regulate the growth of axons? What are the guidance cues they provide to change the growth direction of developing axons? Ramon y Cajal proposed a century ago that diffusible chemicals in the embryonic environment could guide axonal growth over a long range, just like the chemotaxis of single motile cells. They do so by attracting growth cones, the motile sensors at the leading edge of a neuron that have exquisite sensitivity to detect chemical cues in the embryonic tissues. *In vitro* studies of cultured neurons initially demonstrated the existence of factors secreted by intermediate targets of axons. In the past 15 years, many molecules have been identified by genetic screening and biochemical purification to serve as guidance cues. They include several families of highly conserved extracellular molecules – Netrins, Semaphorins, Slits, and Ephrins – as well as classic neurotrophins and morphogens.

The biochemical and biophysical properties of guidance cues separate them into two classes, diffusible and contact mediated, although their distinction has become less clear in recent studies. Diffusible cues are secreted by cells from a long distance and can either attract axons toward the target or repel axons away from it. Contact-mediated cues are often associated with extracellular matrix or cell surface and regulate the adhesiveness of the growth environment, so they either provide a permissive surface on which axons grow or create an unfavorable region that axons tend to avoid. As shown by embryological, tissue culture, and genetic studies, these molecules are present in developing nervous system and work together to control the growth and guidance of axons during their long journey to their eventual synaptic partners. The following sections review recent studies of midline guidance in both invertebrates and vertebrates and use it as an example to demonstrate now a simple change in axon growth direction is regulated by guidance cues at a specific choice point.

Guidepost Cells at the Midline

The midline in the animal CNS is important for establishing neural pathways used for bilateral communication. In the insect nerve cord or the vertebrate spinal cord, a group of neurons called commissures extend their axons across the midline. Once crossed, the axons turn rostrally, join the longitudinal nerve tracks, and eventually synapse on the contralateral side of the brain.

Cells in the ventral midline have been found to serve as intermediate targets to guide the commissural axons to cross the midline. In flies, there are several midline glial cells that are important for controlling the cross because mutations affecting their formation greatly reduce the fidelity of the growth of commissures crossing the midline. In the vertebrate spinal cord, the main action occurs at the floor plate, a region that contains several layers of neuroepithelial cells with distinct morphological and immunological features at the ventral midline. They are derived from the initial neural fold, but their properties are induced by a molecule, Sonic hedgehog, secreted from the notochord that lies ventrally below them. In mouse mutants of Danforth's short tail or Gli2 (a zinc-finger transcription factor) in which the floor plate is missing, commissural axons grow abnormally when they reach the ventral midline. The same has been documented in zebra fish, in which mutations affecting floor plates also result in the misguidance of spinal commissural axons. Therefore, the ventral midline cells provide a unique system for understanding axon guidance at a choice point.

Attracting to the Midline

The initial characterization in grasshoppers demonstrated that a single filopodium of the sensory growth cone makes a direct contact with guidepost cells and leads the axon in the correct direction. The distance between each pair of guidepost stops is less than 100 μm, which might be sufficient for filopodia to search in space and find the guidepost cells by differential adhesion. However, for commissural neurons, especially those from vertebrates, their targets are hundreds of micrometers away. How do they know where to find the intermediate target?

The ventral midline cells appear to secrete diffusible molecules that attract commissural growth cones

to cross midline. The presence of diffusible factors was first demonstrated in the study of vertebrate commissural neurons in culture. When the floor plate from chick or rodent embryos was cultured adjacent to the dorsal spinal cord explant in three-dimensional collagen gels, commissural axons in the explant grew out toward the floor plate. Subsequent biochemical purification identified a family of secreted protein Netrins that can stimulate commissural axon outgrowth and attract them toward COS cell aggregates that produce the protein. Netrins are secreted by the floor plate cells and bind DCC, a cell surface receptor that is exclusively expressed on commissural neuron axons. Genetic analyses in mice, fruit flies, as well as *Caenorhabditis elegans* have established that this ligand and receptor pair is required for midline attraction because mutations in these two genes lead to the failure of commissural axons growing toward the midline. However, interestingly, in the mouse netrin-1 mutant spinal cord, some axons still reach the floor plate. This is due to the presence of another molecule, Sonic hedgehog, also secreted from the floor plate.

How do soluble factors attract axons over a long distance? Target-derived factors have been thought to form a diffusible gradient that guides the growth cone, a theory proposed by Ramon y Cajal after the discovery of growth cones. Recent studies of netrin expression in chicks demonstrated that proteins secreted in the spinal cord do form a dorsal–ventral gradient, with the highest amount concentrated at the floor plate. In addition, growth cones in culture can respond to netrin gradients and turn toward the source that provides the protein. Therefore, chemoattraction provides a simple mechanism to explain the guidance of embryonic axons toward their targets.

Repelling Away from the Midline

After commissural axons are attracted to the midline, most of them do not stay there but instead leave the midline and project to the contralateral side of the spinal cord. A remarkably conservative and reproducible feature is that they cross the midline only once. How is this achieved?

The guidepost cells that provide attraction at the ventral midline also secrete another family of extracellular molecule, Slits. These molecules bind and activate cell surface receptor Robos on the commissural axons and repel them away from the midline after crossing. This was first demonstrated in a genetic screen in flies, in which a mutation in Slit caused all the commissural neurons to collapse at the midline. Interestingly however, the initial study of the Robo receptor revealed a very different phenotype. With only Robo eliminated, commissural axons freely cross and recross the midline and wrap around it to give the roundabout phenotype that is very different from Slit. This is because there are multiple Robo isoforms in flies. When another receptor Robo2 is also deleted in flies, the same Slit phenotype is observed, suggesting that both Robo1 and Robo2 can signal commissural axons to leave the midline, whereas only Robo1 is needed to prevent recrossing. Therefore, slits provide a negative guidance cue to commissural axons and directly activate Robo receptors to drive them away from the midline, thus preventing axon from recrossing. The same molecular mechanism is also used in vertebrates. When all three Slit genes are deleted from the mouse spinal cord, a considerable amount of commissurals are stalled near the floor plate.

In addition to preventing commissural axons from recrossing, the Slit proteins secreted by the midline cells serve as a guide to determine how far each interneuron axon should extend and which fascicle to join. Three Robo 1 homologs are differentially expressed on axonal fascicles, with Robo only on the medial fascicle, Robo and Robo3 on the intermediate ones, and all three (Robo, Robo2, and Robo3) on the most lateral one. The amount of Robo expressed on their surface provides a combinatorial code to determine the position of longitudinal fascicles. When the receptor level is perturbed by overexpression or knockdown, the fascicle positions along the lateral axis either shift away or move closer to the midline accordingly.

Slits are expressed in the midline at the same time as the attractive factor netrins. How do the commissural axons avoid being repelled from the midline before crossing? In flies, this is accomplished by an intracellular protein Comm, which appears to interact with Robo receptors and keeps them away from the growth cone membrane before crossing. In post-crossing axons, Comm expression is downregulated and thereby surface Robos are increased to respond to the repulsive signal from slits. Interestingly, no Comm homolog has been found in vertebrates, but a third Robo-like receptor, Robo3/Rig1, which is expressed also only on the precrossing axons, appears to serve the same function. In the Robo3/Rig1 mouse mutant, commissural neurons reach the ventral side of the spinal cord but stay away from the floor plate, mimicking the defect initially found for the Comm mutation in flies.

Adhesion at the Midline

The initial study of pioneer neuron guidance in grasshoppers suggests that cell adhesion may play an important role in controlling growth direction by

guidepost cells. Studies in the past 20 years have identified many cell adhesion molecules that form homophilic or heterophilic dimers on the cell surface. In the vertebrate floor plate, several immunoglobulin superfamily cell adhesion molecules – axonin-I, NrCAM, and NgCAM – have been shown to be involved in proper control of the behavior of commissural axons at the floor plate. If their functions are perturbed by antibodies in chick embryos, the axons reach the floor plate but fail to cross the midline and turn to the ipsilateral side of the spinal cord.

Another cell surface molecule, Fasciclin II (FasII), provides a permissive substrate for the later born neurons to join the fascicle and extend along the longitudinal axis after crossing. In fly FasII mutants, these bundles do not form tightly, even though they can turn and move away from the midline, as repelled by slits.

Guidepost Function in Synapse Formation

The concept of guidepost cells has also been used in a different developmental context. During the development of neural circuits, axons are guided to different regions of the nervous system, where they contact their synaptic partners. Synaptic target selection and synapse formation are also critical steps to achieve the precise assembly of neuronal circuits. Several studies have shown that distinct populations of guidepost cells are important at the level of synaptogenesis. The following sections summarize the discoveries of these studies.

Caja-Retzius Cells and Certain GABAergic Interneurons as Guidepost in Hippocampus

In hippocampus, the cell bodies of pyramidal neurons are located in the basal region and send out dendrites toward the pial surface. Distinct subcellular domains of pyramidal dendrites are innervated by two populations of afferents. Distal portions of pyramidal dendrites receive input from the ipsilateral entorhinal afferents, whereas the proximal dendrites form synapses with the commissural fibers. This layer structure is established during development with the help of two populations of guidepost cells. Caja-Retzius cells and a set of GABAergic interneurons are early developing neurons found in the two afferent layers. They synapse with the entorhinal afferents and the commissural axons, respectively. These synapses are transient in nature since some of the Caja-Retzius cells and GABAergic interneurons die later. The disappearance of transient synapses is accompanied by synaptogenesis of afferents with the pyramidal dendrites as the synapses are transferred from the guidepost cells to mature synaptic targets.

The significance of the guidepost cells in the development of hippocampal circuit was demonstrated by cell lesion experiments. When the Caja-Retzius cells are ablated, the laminar innervation of the entorhinal axons is impaired. It is interesting to note that the ingrowth of entorhinal axons into the hippocampus precedes the extension of pyramidal dendrites. It is conceivable that the early existence of the Caja-Retzius cells is important to hold the presynaptic terminal in place before the true postsynaptic target arrives. Since in the absence of the Caja-Retzius cells laminar projection of the entorhinal axons is impaired, this suggests that early synaptogenesis might be important for the stabilization of axon arbors. Indeed, several *in vivo* time-lapse studies have shown that synapse formation in the CNS is very dynamic, with constant synapse formation and disassembly, and branch addition and retraction. The presence of synapses on axon branches increases the stability of the branches. Therefore, the laminar distribution of the guidepost cells provides a scaffold for the presynaptic terminals on the afferent fibers, which consequently stabilizes the axonal arborization and achieves laminar innervation.

Subplate Neurons as Guidepost in the Maturation of Visual Cortical Circuit

In the mature visual system of vertebrate animals, thalamic inputs directly innervate primary visual cortex layer 4 neurons. Functional organization of the visual cortex, such as ocular dominance columns and orientation columns, emerges through specific synaptic circuit formation. Interestingly, thalamic inputs first form synapses with another population of neurons, the subplate neurons, before connecting to the layer 4 neurons. At this early time, subplate axons innervate the layer 4 neurons and relay the information from thalamus to the cortex. Later, during the critical period of cortical activity-dependent plasticity, the subplate neurons gradually die through programmed cell death. In the meantime, adult circuit forms in which thalamic inputs directly synapse onto layer 4 neurons.

Ablation of subplate neurons results in the failure of segregation of the thalamic inputs into ocular dominance columns and the formation of orientation columns. This strongly suggests that subplate neurons are essential for synaptic remodeling and maturation of neural circuit. In support of this notion, one study showed that subplate ablation prevents the upregulation of $GABA_A$ receptor expression and perturb the maturation of inhibitory circuits in the layer 4 neurons. Collectively, subplate neurons act as a relay station at early stages of cortical development and are indispensable for patterning mature synaptic circuits.

Vulval Epithelial Cells as Guidepost in the Development of the Egg-Laying Synaptic Circuit

In *C. elegans*, egg-laying behavior is controlled by a pair of motor neuron HSNs. The cell bodies of HSNs are situated just posterior to the vulva, with its axon guided ventrally and then anteriorly. The HSN axons defasciculate dorsally from the ventral nerve cords near the vulval opening and innervate the vulval muscles by forming a cluster of neuromuscular junctions onto the muscle arms. The HSN axons also form synapses onto the VC motor neurons in this region, which in turn innervate muscles. Genetic and developmental analysis revealed that the surrounding vulval epithelial cells play an important guidepost role in the formation of this egg-laying neural circuit. In the absence of these epithelial cells, HSNs form ectopic synapses, which are located more anteriorly than normal. Molecularly, an immunoglobulin superfamily protein, SYG-2, was found to perform the guidepost function. SYG-2 is expressed transiently by the guidepost epithelial cells at early stages of HSN synapse formation. SYG-2 directly binds its receptor on HSN, another immunoglobulin superfamily protein, SYG-1, and clusters SYG-1 at the segment of HSN axon near the vulva. SYG-1 induces accumulation of synaptic vesicles/presynaptic active zone components and directs the location and target selection of HSN presynaptic specialization. In the absence of functional SYG-1 or SYG-2, HSN has a reduced number of synapses formed at its normal location. Instead, the majority of synapses are formed onto adjacent body wall muscles, inappropriate synaptic targets, at anterior ectopic locations. Developmentally, the axons of HSN reach the synaptic region prior to the outgrowth of the postsynaptic VC dendrites. Presynaptic terminals can be observed in this segment of the HSN axons before the VC dendrites reach the same region. The transient synapses of HSN are likely to form directly onto the guidepost cells before being 'handed over' to the late maturing postsynaptic targets. The guidepost vulval epithelial cells have several important functions in the development of the egg-laying organ. These cells attract the migrating sex myoblast, which gives rise to the vulval muscles. They also stimulate the branching of the VC motor neurons. Therefore, guidepost cells not only spatially and temporally control the maturation of the egg-laying neurons and muscles but also regulate the assembly of the neural circuits at the synapse formation level.

The three types of synaptic guidepost cells mentioned previously are probably examples of similar cell types that have not been discovered. Guidepost cells seem to be particularly important in synapse formation where there is temporal discrepancy between axonal and dendritic development. It is conceivable that guidepost cells stabilize axons by forming transient synapses, which disappear upon the arrival of true postsynaptic dendrites.

Conclusion

The function of guidepost cells in axon guidance is well established. They are frequently the sources of axon guidance molecules that attract or repel axon growth cones. The emerging roles of guidepost cells in synapse formation and neural circuit maturation reveal previously unknown complexity during synaptic circuit assembly.

See also: Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Pathfinding: Netrins; Growth Cones.

Further Reading

Bate CM (1976) Pioneerneurones in an insect embryo. *Nature* 260: 54–56.

Bentley D and Caudy M (1983) Pioneer axons lose directed growth after selective killing of guidepost cells. *Nature* 304: 62–65.

Dickson BJ (2002) Molecular mechanisms of axon guidance. *Science* 298: 1959–1964.

Ghosh A and Shatz CJ (1992) Involvement of subplate neurons in the formation of ocular dominance columns. *Science* 255 (5050): 1441–1443.

Ho RK and Goodman CS (1982) Peripheral pathways are pioneered by an array of central and peripheral neurones in grasshopper embryos. *Nature* 297: 404–406.

Kanold PO, Kara P, Reid RC, and Shatz CJ (2003) Role of subplate neurons in functional maturation of visual cortical columns. *Science* 301(5632): 521–525.

Kanold PO and Shatz CJ (2006) Subplate neurons regulate maturation of cortical inhibition and outcome of ocular dominance plasticity. *Neuron* 51(5): 627–638.

Kidd T, Bland KS, and Goodman CS (1999) Slit is the midline repellent for the robo receptor in *Drosophila*. *Cell* 96: 785–794.

Kidd T, Brose K, Mitchell KJ, et al. (1998) Roundabout controls axon crossing of the CNS midline and defines a novel subfamily of evolutionarily conserved guidance receptors. *Cell* 92: 205–215.

Long H, Sabatier C, Ma L, et al. (2004) Conserved roles for Slit and Robo proteins in midline commissural axon guidance. *Neuron* 42: 213–223.

Palka J, Whitlock KE, and Murray MA (1992) Guidepost cells. *Current Opinion in Neurobiology* 2: 48–54.

Sabatier C, Plump AS, Ma L, et al. (2004) The divergent Robo family protein rig-1/Robo3 is a negative regulator of slit responsiveness required for midline crossing by commissural axons. *Cell* 117: 157–169.

Sanes JR and Yamagata M (1999) Formation of lamina-specific synaptic connections. *Current Opinion in Neurobiology* 9(1): 79–87.

Serafini T, Colamarino SA, Leonardo ED, et al. (1996) Netrin-1 is required for commissural axon guidance in the developing vertebrate nervous system. *Cell* 87: 1001–1014.

Serafini T, Kennedy TE, Galko MJ, Mirzayan C, Jessell TM, and Tessier-Lavigne M (1994) The netrins define a family of axon outgrowth-promoting proteins homologous to *C. elegans* UNC-6. *Cell* 78: 409–424.

Shen K and Bargmann CI (2003) The immunoglobulin superfamily protein SYG-1 determines the location of specific synapses in *C. elegans*. *Cell* 112(5): 619–630.

Shen K, Fetter RD, and Bargmann CI (2004) Synaptic specificity is generated by the synaptic guidepost protein SYG-2 and its receptor, SYG-1. *Cell* 116(6): 869–881. [Erratum in *Cell* 117 (4): 553, 2004].

Super H, Martinez A, Del Rio JA, and Soriano E (1998) Involvement of distinct pioneer neurons in the formation of layer-specific connections in the hippocampus. *Journal of Neuroscience* 18(12): 4616–4626.

Tessier-Lavigne M and Goodman CS (1996) The molecular biology of axon guidance. *Science* 274: 1123–1133.

Tessier-Lavigne M, Placzek M, Lumsden AG, Dodd J, and Jessell TM (1988) Chemotropic guidance of developing axons in the mammalian central nervous system. *Nature* 336: 775–778.

Axon Guidance: Morphogens as Chemoattractants and Chemorepellants

J B Thomas and S Yoshikawa, Salk Institute for Biological Studies, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The growth cones of developing neurons are guided to their targets by attractive and repulsive cues in the extracellular environment. Receptors on the growth cones recognize these cues and transduce signals that ultimately lead to changes in the direction of growth. Several families of molecules acting as guidance cues have been identified, including Netrins, Slits, Ephrins, and Semaphorins. In addition to these families, secreted signaling molecules from families of classical morphogens, known for their roles in controlling cell fates in a concentration-dependent manner, can act as axon guidance molecules. Sonic hedgehog (Shh) of the hedgehog family, bone morphogenetic protein 7 (BMP7) of the transforming growth factor-beta (TGF-β) family, and members of the Wnt family have all been found to function in the guidance of specific classes of neurons.

For each of the three morphogen families, the receptors and canonical signaling pathways through which they control cell fate are well studied and involve transcriptional regulation as their output. However, some members of these families can also signal through noncanonical pathways and control processes such as cell movements and cell polarity. The roles of these families in axon guidance, plus the finding that they likely act directly on the growth cone, argue for the activation of noncanonical signaling pathways that lead to cytoskeletal rearrangement.

Shh as a Chemoattractant for Commissural Neurons

Commissural neurons residing in the dorsal region of the spinal cord adjacent to the roof plate project axons through the floor plate at the ventral midline (Figure 1(a)). These commissural axons initially extend away from the roof plate and take a ventral and circumferential pathway through the dorsal spinal cord. Midway to the ventral midline, the axons change course and project ventrally and medially to the floor plate. After crossing the floor plate, they then make an abrupt turn and project anteriorly on the contralateral side of the spinal cord.

Commissural axons are guided to the ventral midline by Netrin-1, a chemoattractant secreted by cells of the floor plate and the adjacent periventricular zone. In *Netrin-1*$^{-/-}$ mutant mice, as well as in mice mutant for the Netrin receptor deleted in colorectal cancer (DCC), many commissural axons project abnormally, often failing to enter the ventral region of the spinal cord. However, some commissural axons in these mutants do manage to reach the floor plate, suggesting the existence of one or more additional factors in the floor plate. Indeed, the floor plate of *Netrin-1*$^{-/-}$ mice is still effective in reorienting commissural axons when juxtaposed to spinal cord explants, indicating the presence of another chemoattractant.

Shh was implicated as the additional factor since it is secreted by the floor plate and is known to have long-range effects within the spinal cord. In addition, Shh can cause rapid changes in the growth cone behavior of cultured retinal ganglion cells. An initial indication that Shh might be the missing floor plate factor was garnered by demonstrating the ability of Shh-expressing COS cells to reorient commissural axons in spinal cord explants. This reorienting effect of Shh is mediated through the Shh signaling component Smoothened (Smo), since the Smo inhibitor cyclopamine abolishes the effect.

Assessing the axon guidance role of Shh *in vivo* is more challenging because of its role in early patterning of the spinal cord. Shh acts as a morphogen to generate distinct neuronal cell types within the ventral spinal cord in a concentration-dependent manner. Although Shh is not directly required for generation of the dorsally located commissural neurons, these neurons do project their axons ventrally through regions that are patterned by Shh. This rules out the analysis of mice mutant for Shh, Smo, or the Shh receptor patched (Ptc) since other guidance cues in the ventral spinal cord of these mutants might be altered. However, using the Cre/loxP recombinase system and the *Wnt-1* promoter to drive expression of Cre recombinase, Smo can be selectively removed in commissural neurons developing within an otherwise normally patterned spinal cord. In these conditionally mutant mice, commissural axons project abnormally into the ventral spinal cord, sometimes failing to make the medioventral turn toward the floor plate and instead projecting along the edge of the spinal cord. Thus, Shh appears to be acting as a chemoattractant for commissural neurons. Ultimately, the commissural axons in these conditional Smo mutants do reach the ventral midline and form

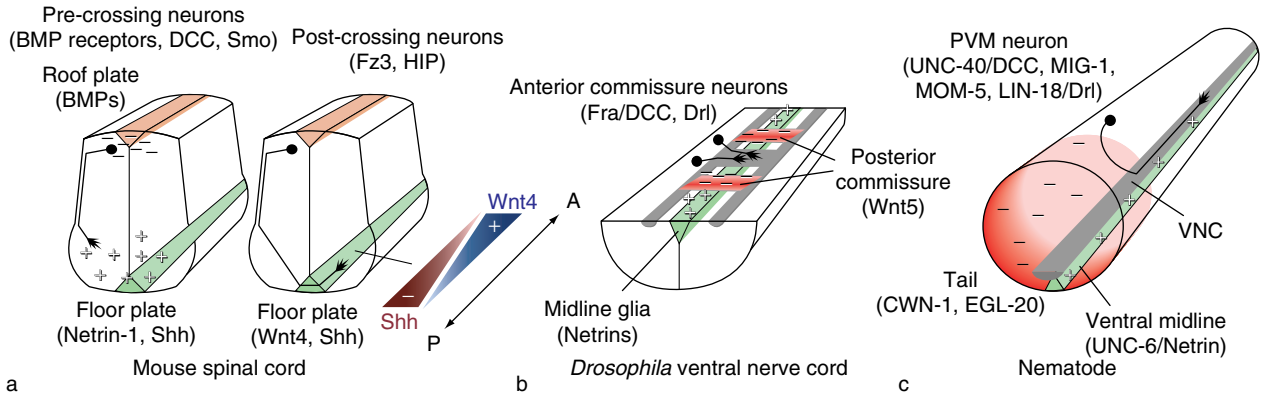

Figure 1 Morphogens as axon guidance molecules. Schematic of the axon projections of neurons in the developing mammalian spinal cord (a) and the ventral nerve cords (VNC) of *Drosophila* (b) and nematode (c). The floor plate (green) secretes Netrin-1 (+), which acts as a chemoattractant for commissural axons expressing the Netrin receptor DCC. The floor plate also secretes Shh, which functions in concert with Netrin-1 to attract commissural axons expressing the Shh signaling component Smo. The roof plate (orange) secretes BMP7 (−), which repels the initial growth cones of commissural neurons away from the dorsal midline, probably as a heterodimer with GDF7. The specific BMP receptors in commissural neurons mediating the repulsive response are not known. After crossing the floor plate, growth cones of commissural neurons turn anteriorly in response to an increasing posterior-to-anterior (P-to-A) gradient of Wnt4, plus a decreasing gradient of Shh, secreted by floor plate cells. Fz3 and HIP are thought to mediate the responses to Wnt4 and Shh, respectively. In *Drosophila*, the major axon tracts (gray) consist of the bilaterally symmetric longitudinal connectives running in the anterior/posterior axis and, connecting the two sides in each segment, an anterior and posterior commissure. Like the vertebrate floor plate, the midline glia (green) secrete Netrins (+) which attract commissural axons expressing the DCC homolog Fra. Commissural axons that project through the anterior commissure, all of which express the Drl receptor, are repelled from the posterior commissures of their segment of origin and the adjacent anterior segment by Wnt5 (−). In the nematode, migrating cells that end up at the ventral midline (green) secrete the Netrin homolog UNC-6 (+), as does the VNC (gray). Axons expressing UNC-40, the nematode DCC homolog, are attracted to the VNC. Wnt proteins CWN-1 and EGL-20 (−) are expressed by cells in the tail and act to repel PVM and other anteriorly projecting neurons expressing the Fz receptors MIG-1 and MOM-5. The Drl/Ryk homolog, LIN-18, acts redundantly in Wnt-mediated repulsion.

a normal-looking commissure, arguing that Netrin-1, in the absence of Shh, is sufficient to attract commissural axons to the floor plate. If Shh is the only additional guidance factor secreted by the floor plate, the expectation is that *Smo* mutant commissural neurons should not be able to reach the floor plate in a *Netrin-1*$^{-/-}$ mutant background.

Although a loss-of-function phenotype is the gold standard for demonstrating that a gene is required for a particular process, a phenotype does not necessarily reveal the underlying mechanism. For example, rather than acting directly on the growth cones of commissural neurons, Shh could be activating a retrograde signal within the neurons that regulates the expression of genes encoding other guidance receptors. The definitive test for a guidance molecule's direct action on growth cones is the *in vitro* growth cone turning assay. Here, the spatial and temporal application of a factor can be precisely controlled by pulsing it through a pipette, creating a gradient of purified protein to which isolated cultured neurons can respond. Importantly, in this context Shh has the ability to rapidly attract growth cones and this attraction is abolished by cyclopamine. Thus, Shh, signaling through Smo, can indeed act as a chemoattractant.

Shh as a Chemorepellant for Postcrossing Commissural Neurons

After crossing the midline, commissural axons turn anteriorly along the contralateral side of the floor plate. RNA interference (RNAi) experiments in chick embryos have implicated Shh in this guidance event. When Shh levels are reduced by injection of Shh dsRNA, commissural axons tend to either stall at the floor plate exit point or project posteriorly. A decreasing posterior-to-anterior gradient of Shh suggests that in this case Shh is acting as a chemorepellant, a notion supported by its ability to repel postcrossing commissural axons in spinal cord explants. Importantly, this repulsion is not mediated by Smo and Ptc, since neither is expressed by commissural neurons by the time they have crossed the midline. Furthermore, cyclopamine has no effect on turning behavior. Instead of Ptc, another Shh receptor, hedgehog interacting protein (HIP), has been implicated. Lowering HIP function by RNAi knockdown results in the same turning defects observed when Shh is downregulated.

A model has thus emerged that depending on the receptor employed, Shh can act as either a chemoattractant or a chemorepellant. However, in contrast

to Shh chemoattraction, it is not known whether HIP-mediated repulsion by Shh is a result of direct action on the growth cone rather than the transcriptional control of other guidance receptors, such as those involved in Wnt-mediated attraction of post-crossing commissural axons.

BMP7 as a Chemorepellant for Commissural Neurons

The early phase of commissural axon growth within the dorsal spinal cord is unaltered in either *Netrin-1*$^{-/-}$ mutants or mice lacking a floor plate, suggesting the existence of additional non-floor plate-derived factors in their guidance. The proximity of commissural neurons to the roof plate and the stereotyped projections of their axons away from the dorsal midline suggests a role of the roof plate in repulsion (Figure 1(a)). This idea is supported by experiments showing that when juxtaposed to a spinal cord explant, the roof plate can reorient commissural axons away from the side facing the roof plate. Furthermore, the reorienting activity can be mimicked by COS cells expressing BMP7, one of the BMPs secreted by the roof plate.

BMP7 functions most efficiently in the axon reorienting assay by forming heterodimers with GDF7, another BMP family member expressed by roof plate cells but which on its own has no axon reorienting activity. In *BMP7*$^{-/-}$ mutant mice, there is a significant increase over wild type in the number of commissural axons that initially project medially and dorsally instead of laterally and ventrally. Similar numbers are seen in *GDF7*$^{-/-}$ mutants. However, in each case, these defects are corrected such that later in development the projection pattern of commissural neurons resembles that of wild type. Thus, in addition to BMP7 and GDF7, other cues must be operating–repulsive ones from the roof plate and/or attractive ones from more lateral positions in the dorsal spinal cord. The former possibility is bolstered by the finding that roof plates from *BMP7*$^{-/-}$ *GDF7*$^{-/-}$ double mutants, although compromised in their ability to redirect commissural axons within spinal cord explants, still retain some residual axon-reorienting activity.

Similar to the patterning role of Shh in the ventral spinal cord, BMP family members function early in development to generate specific dorsal cell types. In fact, GDF7 is required to generate a specific subclass of commissural neurons, raising the question of whether the observed guidance effects of BMP7 are direct or indirect. Although the guidance role of BMP7 has not been cleanly separated from its inductive role, as was done for Shh using the conditional *Smo* mutation, several lines of evidence argue against an indirect mechanism involving changes in cell fate. First, *BMP7*$^{-/-}$ mutants show no obvious defects in dorsal spinal cord patterning, presumably due to redundancy with one or more of the other roof plate BMPs. Second, commissural neurons in spinal cord explants from *BMP7*$^{-/-}$ mutants are able to reorient to the same extent as those from wild type in response to a wild-type roof plate, suggesting that these neurons retain at least some of their wild-type properties, including the expression of an appropriate BMP receptor complex capable of transducing the BMP7 signal. Finally, although *in vitro* growth cone turning assays to test for repulsion have not been carried out, BMP7 alone, as well as BMP7:GDF7 heterodimers, is capable of causing rapid growth cone collapse of commissural neurons *in vitro*. Collectively, these studies advance the case for a direct role on the growth cone in axon guidance.

Wnt Proteins as Chemorepellants

Wnt5 in *Drosophila*

In addition to specifying cell fates, Wnt proteins play diverse signaling roles in the developing nervous system. These include presynaptic axon remodeling during vertebrate synaptogenesis, the maturation of the *Drosophila* neuromuscular junction, and the polarity of neurons along the anterior–posterior axis in the nematode *Caenorhabditis elegans*. As first shown in *Drosophila*, Wnt family members have also turned out to function as axon guidance molecules.

As in vertebrates, the large number of commissural neurons in the *Drosophila* embryonic ventral nerve cord (the fly counterpart to the spinal cord) project their axons across the midline to the contralateral side (Figure 1(b)). Analogous to the vertebrate floor plate, specialized cells at the midline, the midline glia, divide the two halves of the ventral nerve cord and play a critical role in axon guidance. Like the floor plate, midline glia secrete Netrins that act as chemoattractants for commissural axons expressing Frazzled (Fra), the fly homolog of the Netrin receptor DCC.

Once attracted to the midline, commissural axons do not cross randomly. Instead, in each segment they reproducibly choose one of two distinct tracts that connect the two sides, either the anterior or the posterior commissure. This choice of whether to project anteriorly or posteriorly into the appropriate commissure is controlled by one of the seven fly members of the Wnt family, Wnt5, and a receptor to which it binds, derailed (Drl). The Drl receptor is expressed on the growth cones and axons of all neurons that

project through the anterior commissure. In *drl* mutants, many of the anterior commissure axons abnormally project through the posterior commissure; conversely, when misexpressed by posterior commissure neurons, Drl switches their axonal projections to the anterior commissure. Drl is therefore not only required by anterior commissure neurons for their guidance but also sufficient to dictate the choice of the anterior commissure for crossing axons.

In *wnt5* mutants, as in *drl* mutants, anterior commissure axons project abnormally through the posterior commissure. Since Wnt5 is expressed by cells associated with the posterior commissure (and the Drl receptor by anterior commissure neurons), it is functioning as a chemorepellant to keep Drl-expressing axons from entering the posterior commissure. Such a repulsive activity was illustrated *in vivo* by misexpressing Wnt5 in the midline glia. In these embryos, anterior commissure axons are prevented from crossing the midline in a Drl-dependent manner, overriding Netrin-mediated attraction.

Although Drl binds Wnt5 and is required for Wnt5 function in guidance, it is not yet clear whether Drl transduces the Wnt5 signal within growth cones or whether it acts as a coreceptor with a member of the other family of Wnt receptors, the Frizzled (Fz) proteins. Fz receptors are good candidates for transducing such a signal since they have been shown not only to mediate canonical Wnt signaling resulting in transcriptional regulation but also to mediate noncanonical signaling involved in controlling cell movements. In support of the coreceptor hypothesis, the mammalian homolog of Drl, Ryk, can bind to the cysteine-rich domain of Fz proteins when the two are co-expressed in tissue culture cells, suggesting the ability of Fz and Ryk to form a complex. Regardless of which receptor(s) transduces the signal, it is clear that Wnt5 controls the guidance of anterior commissure axons and that the Drl receptor is essential for Wnt5 signaling.

However, does Wnt5 act directly or indirectly in these guidance events? The difficulty in purifying active Wnt proteins has hindered efforts to test Wnt5 for repellant activity in an *in vitro* growth cone turning assay. However, a key observation *in vivo* does suggest a direct role in guidance: Drl can dictate commissure choice in a Wnt5-dependent manner when misexpressed on growth cones shortly before they make the choice. This makes it less likely that a retrograde signal, followed by a round of transcription and translation, is required.

Mammalian Wnts

Wnt proteins, acting through the Ryk receptor, also control axon guidance events in the developing mammalian nervous system. Similar to Drl expression on anterior commissure axons in *Drosophila*, Ryk is expressed on axons that project across the midline within the corpus callosum. In $Ryk^{-/-}$ mutant mice, callosal axons reach and cross the midline but often fail to project away from it on the contralateral side. The Wnt ligand in this case appears to be Wnt5a, which together with Wnt5b is most closely related to *Drosophila* Wnt5. Wnt5a binds to Ryk and is expressed by midline glia known to be required for proper guidance of the callosal axons. The spatial and temporal expression pattern of Wnt5a suggests that it is acting as a chemorepellant to guide Ryk-expressing axons away from the midline, a notion supported by the finding that Wnt5a can repel cortical axons in explants from wild-type, but not $Ryk^{-/-}$, brains.

Like callosal axons, axons of corticospinal neurons also express Ryk. These axons project posteriorly from the brain through a dorsal region of the spinal cord that exhibits an increasing posterior-to-anterior gradient of a number of Wnts, including Wnt1 and Wnt5a. Anti-Ryk antibodies block both the posterior growth of corticospinal axons *in vivo* and the repellant activity of Wnts on cortical axons in cultured explants, providing good evidence that Ryk is required for posterior guidance by interpreting the Wnt gradient as a chemorepellant one. The striking similarities between mammalian Wnt-mediated corticospinal axon guidance and *Drosophila* Wnt5-mediated commissure choice argue for a deep-rooted conservation of function: both fly Wnt5 and mammalian Wnts are acting as chemorepellants that signal via the Drl/Ryk receptor to control guidance in the anterior–posterior axis.

Wnts in the Nematode

In the nematode *C. elegans*, the conserved cue Netrin (UNC-6 in nematode) guides axons to the ventral nerve cord at the midline (**Figure 1(c)**). Once they enter the ventral nerve cord, growth cones choose to project either anteriorly or posteriorly. For example, the axon of the posterior ventral microtubule cell (PVM) enters the nerve cord and projects anteriorly toward the head. Two Wnts, CWN-1 and EGL-20, are expressed in the posterior end of the embryo and act redundantly as chemorepellants to guide the PVM axon, as well as others, anteriorly. In *cwn-1:egl-20* double mutants, the PVM axon often projects posteriorly and this guidance defect is enhanced by misexpression of EGL-20 anteriorly. In contrast to *Drosophila* and mouse, Wnt repulsion of the PVM axon is mediated primarily through two Fz receptors, with LIN-18, the nematode homolog of Drl/Ryk, acting redundantly, presumably through a parallel pathway.

Wnt Proteins as Chemoattractants

After crossing the floor plate of the spinal cord, commissural axons turn anteriorly toward the brain (Figure 1a). Observations of cultured spinal cord explants provided the first indication that this guidance event might be controlled by a chemoattractant. Near the posterior cut end of spinal cord explants, commissural axons project anteriorly in a normal fashion, but near the anterior cut end, they either stall or project randomly along the anterior–posterior axis. This suggests that an increasing posterior-to-anterior gradient of a diffusible chemoattractant, which slowly escapes from the cut ends of explants, guides commissural axons anteriorly. A candidate molecule approach led to the Wnt family, particularly Wnt4, whose mRNA levels exhibit an increasing posterior-to-anterior gradient in the floor plate. Wnt4-expressing COS cells juxtaposed to the anterior cut end of explants can redirect postcrossing axons anteriorly, in effect rescuing the projection defects near the cut end. Further support for this guidance event being mediated by a Wnt protein comes from the analysis of mice deficient for Fz3, one of the Wnt receptors expressed by commissural neurons. In these mice, commissural growth cones emerge from the floor plate and tend to project randomly along the anterior–posterior axis of the spinal cord.

Similar to Wnt function in *Drosophila* and nematode, Wnt4 appears to control anterior–posterior guidance. However, in contrast to the repellant activity of Wnts in the guidance of *Drosophila* commissural axons, the nematode PVM axon, and of mammalian callosal and corticospinal axons, Wnt4 is acting as a chemoattractant in the guidance of postcrossing commissural axons. Notably, this guidance event does not involve Ryk since commissural neurons do not express the receptor. Instead, it seems that a member of the Fz family of Wnt receptors, acting in the absence of Ryk, mediates the guidance. A model similar to that for Shh has been proposed in which Wnts can act as either chemoattractants or chemorepellants depending on whether a Fz or Ryk receptor is involved. However, the relative roles of these receptors in Wnt-mediated guidance in nematode suggest that the situation may turn out to be more complicated.

Conclusion

Hedgehog, TGF-β, and Wnt family members not only have roles in cell fate specification and early embryonic patterning but also act as axon guidance molecules. For Shh and BMPs, morphogen gradients initially used to specify cell types in the spinal cord are reused later in development to attract or repel axons. There is evidence that Shh, BMP7, and Wnt proteins act directly on the growth cone rather than through canonical signaling pathways to the nucleus. The current challenge, and one which is faced generally in studies of guidance molecules, is to identify the signaling pathways engaged by the receptors and to understand how these signals are integrated within the growth cone to achieve changes in direction of growth.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Pathfinding: Netrins; Axonal Pathfinding: Guidance Activities of Sonic Hedgehog (Shh); Axonal Pathfinding: Extracellular Matrix Role; Growth Cones; Morphogens: History.

Further Reading

Bourikas D, Pekarik V, Baeriswyl T, et al. (2005) Sonic hedgehog guides commissural axons along the longitudinal axis of the spinal cord. *Nature Neuroscience* 8: 297–304.

Butler SJ and Dodd J (2003) A role for BMP heterodimers in roof plate-mediated repulsion of commissural axons. *Neuron* 38: 389–401.

Charron F, Stein E, Jeong J, McMahon AP, and Tessier-Lavigne M (2003) The morphogen sonic hedgehog is an axonal chemoattractant that collaborates with netrin-1 in midline axon guidance. *Cell* 113: 11–23.

Dickson BJ (2002) Molecular mechanisms of axon guidance. *Science* 298: 1959–1964.

Fradkin LG, Garriga G, Salinas PC, Thomas JB, Yu X, and Zou Y (2005) Wnt signaling in neural circuit development. *Journal of Neuroscience* 25: 10376–10378.

Keeble TR, Halford MM, Seaman C, et al. (2006) The Wnt receptor Ryk is required for Wnt5a-mediated axon guidance on the contralateral side of the corpus callosum. *Journal of Neuroscience* 26: 5840–5848.

Liu Y, Shi J, Lu CC, et al. (2005) Ryk-mediated Wnt repulsion regulates posterior-directed growth of corticospinal tract. *Nature Neuroscience* 8: 1151–1159.

Lu W, Yamamoto V, Ortega B, and Baltimore D (2004) Mammalian Ryk is a Wnt coreceptor required for stimulation of neurite outgrowth. *Cell* 119: 97–108.

Lyuksyutova AI, Lu CC, Milanesio N, et al. (2003) Anterior–posterior guidance of commissural axons by Wnt-frizzled signaling. *Science* 302: 1984–1988.

Pan CL, Howell JE, Clark SG, et al. (2006) Multiple Wnts and frizzled receptors regulate anteriorly directed cell and growth cone migrations in *Caenorhabditis elegans*. *Developmental Cell* 10: 367–377.

Schnorrer F and Dickson BJ (2004) Axon guidance: Morphogens show the way. *Current Biology* 14: R19–R21.

Yoshikawa S, McKinnon RD, Kokel M, and Thomas JB (2003) Wnt-mediated axon guidance via the *Drosophila* Derailed receptor. *Nature* 422: 583–588.

Axonal and Dendritic Identity and Structure: Control of

C G Dotti and A Gärtner, University of Leuven and Flanders Institute of Biotechnology, Leuven, Belgium
F Calderon de Anda, Picower Institute for Learning and Memory, MIT, Boston, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In its simplest biological application, the term polarity is used to describe the existence of an asymmetric process, whether this refers to the shape of an organism, an organ or a single cell; directed movement; the distribution of molecules and organelles; and intracellular vectorial flows, centrifugal and centripetal. The relevance of understanding the cellular and molecular basis of polarity lies in the fact that most basic biological functions are a consequence of asymmetry. For instance, during development, the formation of the main body axes, that is, rostral–caudal and ventral–dorsal, is the consequence of the establishment of polarized domains. Polarized activities also control the proper three-dimensional organization of organs and tissues and the coordinated orientation of cells in two dimensions in certain tissues, a process called planar polarity. At the single-cell level, to cite only a few examples, polarity is important for asymmetric cell division, which generates cellular diversity; for the migratory movement of cells in response to external stimuli; for the differential uptake and release of ligands; and for the reception and propagation of electrical signals through dendrites and axons in neurons.

A polarized neuron is characterized by a very particular molecular and supramolecular organization of axons and dendrites, that is, a differential distribution of both cytoplasmic and membrane components. Most simply seen, the capacity of axons to transmit signals over long distances is due to the segregation of voltage-gated sodium channels and synaptic vesicles to this domain. Dendrites, on the other hand, are capable of receiving incoming information owing to the presence of neurotransmitter receptors. At the cytoplasmic level, the microtubule binding protein MAP2 is specifically distributed in dendrites, and axons are enriched in tau. Moreover, microtubules in axons are always oriented with their plus end toward the periphery, whereas in dendrites they exhibit a mixed polarity. Neurons, however, establish a clearly polarized phenotype very early in development, before they have acquired the capacity to transmit and receive electrical signals.

Neuronal polarization becomes visible shortly after the last mitosis, through the sprouting of a number of extensions which, due to the lack of a clear axonal or dendritic identity, are called 'minor neurites.' In the typical case of neurons generated in the ventricular or subventricular zones, these initial extensions constitute the cell's migratory force for reaching its final position in the cortex. It is of great importance to understand how neurons establish their polarity. Defects in this early process will have severe consequences in the way the brain is organized into different layers, nuclei, and areas, leading to brain dysfunction. Moreover, failures in the definition of axons and dendrites at later stages will lead to defective wiring of the brain. This article describes the morphological events and mechanisms leading to the regulated generation of the primordial neurites and how they differentiate into precursors of functional axons and dendrites. The anatomical changes occurring during the development of polarity have been well described for a variety of neurons *in situ*. However, the molecular, biochemical, and cellular biological knowledge on how a neurite initially forms from the round cell body and how this neurite later acquires axonal or dendritic characteristics mainly relies on studies performed in hippocampal neurons taken from rat embryos, which spontaneously polarize under *in vitro* conditions. After a short description of the morphological phases of neuronal polarization *in situ*, this article describes how neuronal polarization is understood as a result of the large number of studies performed in hippocampal neurons in culture.

From Round to Polar: Morphological Steps

Developmental Stages *In Situ*

In situ, cortical pyramidal neurons are generated in the proliferative layer of the dorsal telencephalon and follow mainly a radial path toward the pial surface, where they occupy the most superficial layer of the developing cortical plate and populate the cortex in an inside first, outside last sequence, generating the six-layered cortex. In the mouse neocortex, neurogenesis is initiated at embryonic day (E) 10, peaks at E15, and finishes around birth. Initially, ventricular zone neuroepithelial cells give rise, through symmetric proliferative divisions, to neurons and radial glia cells (E10–E12). While the newly generated neurons begin to differentiate, radial glia cells continue to divide either symmetrically, into two neurons, or asymmetrically, into one neuron and one radial glia cell. In addition, new neurons arise from symmetric

neurogenic divisions of basal precursors in the subventricular zone. Irrespective of the place and mode by which they are generated, neurons start to migrate toward upper cortical layers through a four-phase process. In phase 1, ventricular zone-derived neurons acquire a bipolar morphology, with one neurite facing toward the ventricle and one neurite toward the opposite, pial surface and migrating to the subventricular zone. In phase 2, neurons pause for approximately 24 h, forming more neurites, making the cells appear multipolar. In phase 3, a subpopulation of neurons extend a process toward the ventricle. Finally, in phase 4, cells migrate, after again acquiring a bipolar shape with a pia-directed leading process, to their final destination, where neurons attain their final shape and polarity. Thus, neuronal polarization begins at the place where neurons are generated, through the sprouting of initial asymmetric neurites. In addition, *in vivo* morphological progression studies indicate that the first 'role' that polarization plays is to support neuronal migration – quite different from the electrical transmission role of polarity in the mature neuron.

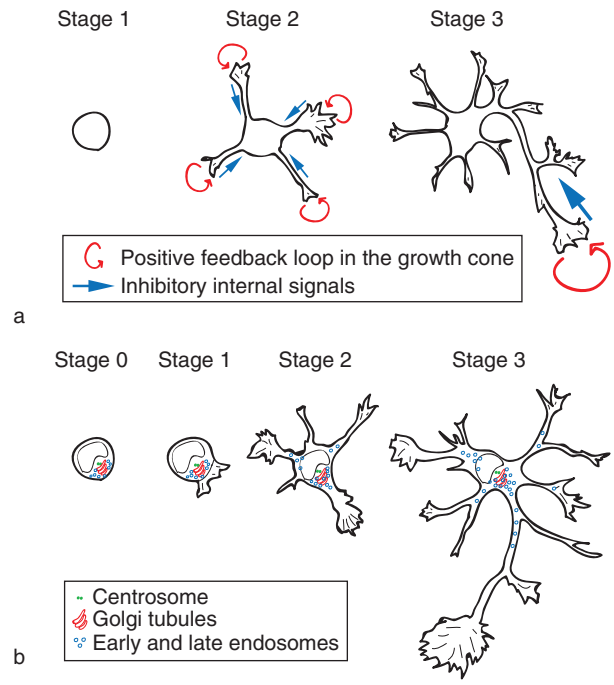

Figure 1 Establishment of polarity in dissociated hippocampal neurons. (a) This model proposes that neuronal polarization starts at the stage 2–3 transition. Positive and negative feedback regulation leads to the disruption of stage-2 symmetry. Red circles with arrows represent molecular signaling cascades that form positive feedback loops in the growth cone and promote neurite extension. Blue arrows represent a negative signal that is generated at each growth cone, propagates throughout the cell, and modulates the positive feedback loops in other neurites. The growth cone with highest positive feedback (produced probably by the environment) generates a strong negative signal, precluding other neurites from growing; meanwhile the neurite with the highest positive feedback grows as an axon. (b) The second model proposes an intrinsic polarization that is established at the immediate postmitotic stage and maintained through differentiation. Modified from Dotti CG, Sullivan CA, and Banker GA (1988) The establishment of polarity by hippocampal neurons in culture. *Journal of Neuroscience* 8: 1454–1468. From De Anda FC, Pollarolo G, Da Silva JS, Camoletto PG, Feiguin F, and Dotti CG (2005) Centrosome localization determines neuronal polarity. *Nature* 436: 704–708.

Developmental Stages *In Vitro*

Although the initial development of neurons is described well *in situ* on a morphological basis, most data describing functional and molecular aspects underlying the establishment of polarity are derived from a more easily accessible *in vitro* model system, that is, isolated rat hippocampal neurons developing in low-density culture. An early study on the polarization of these neurons demonstrated that polarization occurs in a most stereotyped and reproducible manner: Shortly after dissociation from the brain, embryonic hippocampal neurons are round (named stage 1 of development). Within the next 24 h, the round cells sprout four to six morphologically and ultrastructurally identical neurites (called stage 2). Finally, one of these neurites elongates more rapidly than the others, becoming the cell's axon. This is, according to this study, the first stage of polarity and represents developmental stage 3. A most important implication of that model and subsequent work was that polarity is established at the transition between stages 2 and 3 (see Figure 1(a)). A different perspective has arisen from more recent work showing that the neurite that appears first from the cell body of the new neuron later becomes the axon. The data from this recent work are consistent with the view that the immediate postmitotic neuron corresponds to stage 0, which represents a round neuron with a polarized cytoplasm marking the site of axonal outgrowth. The stage after the neuron grows the first sprout from that site is named stage 1. Stage-2 and stage-3 neurons in the new model correspond to stages 2 and 3 in the earlier classification (Figures 1(a) and 1(b)). The data from the newer work suggest that the establishment of polarity occurs immediately after mitosis, at the transition between stage 0 and stage 1 (see Figure 1(b)). Figure 2(a) illustrates the progression of polarity in actual neurons. The important difference in these two concepts of when polarity is first established is that they imply different mechanistic principles. Thus, from the earlier model, the random choice of a neurite for an axonal destiny implies that polarized growth is a stochastic process, in which the cross-talk between one neurite and the

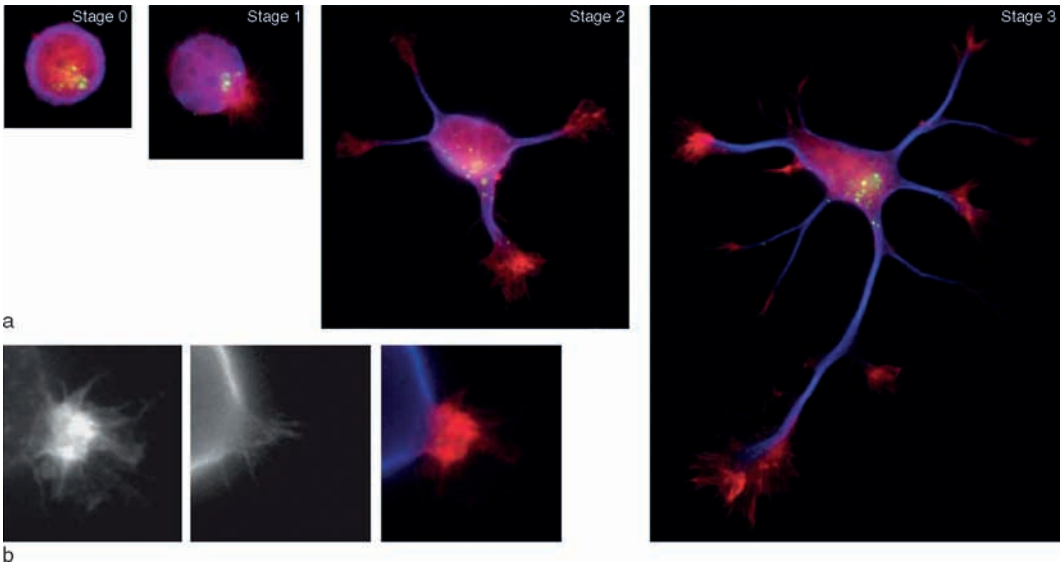

Figure 2 (a) Intracellular polarization through different stages of neuronal differentiation as proposed in the model in Figure1(b). Neurons were fixed after different times in culture and stained for tubulin (blue), F-actin (red), and endosomes (green). (b) A neuronal sprout containing unstable F-actin with microtubules entering the sprout. Left panel, actin (red), middle panel, tubulin (blue), and right panel, merging of tubulin (blue) and actin (red) signals.

most appropriate environmental cue decides final outcome (i.e., axonal outgrowth). On the contrary, the temporal hierarchy described by the second model, suggesting that the first neurite has greater likelihood of becoming an axon, does not permit environmental influence, implying that the neurite that later grows into an axon is intrinsically predisposed. As will be shown, a number of early and recent observations indicate that the models overlap and that the establishment of neuronal polarity is, in the end, the con sequence of an initial intrinsically mediated predisposition later confirmed through extrinsic signaling.

From Round to Polar: Creating a Positional 'Lead'

Irrespective of the primary cause determining which neurite later becomes axon, the fact is that before this occurs, neurites must form from the round cell body of an immediately postmitotic neuron. This we consider as a most critical stage of neuronal polarization as it is the precondition for the later development of neurites into either axons or dendrites.

The sphere is the most economical cell architecture. Thus, the occurrence of the earliest morphological change leading to the appearance of the first neurite sprout is necessarily the product of energy-consuming activities that, on the one hand, disrupt the natural sphere-sustaining activities and, on the other hand, activate expansive forces. The site where sphere-sustaining forces are disrupted and consequently where neurite-inducing forces act must be tightly oriented in space. As described in the *in situ* polarization process, the place where the first neurite emerges is a most critical event as it determines the fidelity of migration. Therefore, the following section describes the most conspicuous intracellular constituents behind the disruption of the postmitotic neuron sphere as well as those involved in generating the primordial neurite and the way in which these activities are restricted in space.

Disrupting the neuronal sphere will require, among many other activities, changes in the dynamics of microtubules and actin filaments. In the quiescent, nonmotile stage, actin filaments are organized in a tight meshwork that surrounds the entire cell, directly underneath the plasma membrane. The action of a spatially restricted signal, whether a centrifugal action from the inside of the cell (e.g., microtubules or transported membrane) or a centripetal force from an environmental signal acting on the plasma membrane, will affect the stability of the tight actin meshwork, producing a localized change in this structure. Commonly, this change is the destabilization of the tight actin structure, leading to the formation of an actin cortex 'pore' allowing for the penetration of microtubules and thus increasing their chances to directly contact the plasma membrane at a spatially restricted place (Figure 2(b)). In fact, when the tight actin cortical organization is broken, highly motile actin-containing structures appear on the surface of cells: filopodia and lamellipodia, rich in highly dynamic actin and microtubules. In a neuron these

structures precede the formation of a neurite. Neurites therefore arise as a consequence of the destabilization of the actin cortex, the polymerization and stabilization of microtubules, and the directed membrane delivery via these microtubule tracks from the center of the cell. Microtubule polymerization can work in addition as one of the centrifugal forces inducing a localized change in the cortical actin, to trigger the escape of microtubules to the periphery to make the neurite. Microtubules polymerize in a polarized fashion from the minus to the plus end. While the minus end is either embedded in the microtubule organizing center (MTOC) or protected by minus end-binding proteins, the plus end dynamically explores the cytoplasm, and its dynamics and localization (e.g., its connection with the cell cortex) are regulated by a number of plus end-binding proteins, such as cytoplasmic linker protein-170 (CLIP-170), CLIP-115, lissencephaly 1 (Lis1), and end-binding protein 1. These proteins, which collectively are called plus end tracking proteins, are key players during neuronal migration, and their dysfunction leads to migration defects and thus severe brain malformations. It is interesting that *Lis1* haploinsufficiency results in reduced filamentous actin at the leading edge of migrating neurons. To summarize, changes in the dynamics of the actin and microtubule skeleton are major events leading to the disruption of the sphere and in turn to neurite production. However, attention should be drawn to many other events, such as localized endocytosis, local translation of messenger RNAs, calcium release, and mitochondria trafficking, without which the microtubule–actin tandem would not be able to act as a mechanical force.

How does the neuron position the microtubule–actin sphere-breaking/neurite-generating activities so that a neurite is made at the right place? This is a most critical issue as *in situ* the primary neurites must arise from a precise pole in order to more efficiently interact with the cues required to guide the neuron's migration. The precise place where microtubule–actin-based activities occur can be theoretically determined by two mechanisms. The first most economically profits from the asymmetric organization of the cytoplasm of a recently generated neuron. In fact, one immediate consequence of cytokinesis is the occurrence of an asymmetrically positioned centrosome, the organelle from which microtubules are nucleated, in the daughter cells (**Figure 3(a)**). The asymmetric positioning of this organelle adds to the separation 'force,' which will result in stronger contact of microtubules with the nearby actin cortex and thus the plasma membrane (**Figure 3(b)**). Since an asymmetrically positioned centrosome is thought to play a role in other polar processes, such as the orientation of the migratory axis of motile cells, it can certainly be important in controlling the emergence of neurites in neurons.

The second mechanism that can spatially restrict the activity of cytoskeletal forces depends on the localization of certain molecules to a restricted place on the membrane. In parallel to studies on bud site selection on yeast, where the membrane of the future bud site is marked and recruits a cascade of molecules that regulate trafficking and the cytoskeleton, such a hypothetical membrane mark could also be a determining feature in neuronal polarity (**Figure 3(c)**). Good candidate molecules to exert such a role are the proteins that regulate polarity during asymmetric cell division, such as the partitioning-defective proteins (Pars). Indeed, Par3 and Par6 are described as apically localized in neuroepithelial and radial glia cells. Their sorting to one of the two cells during division provides that cell with precursor properties, while the other cell becomes a neuron. Therefore, it is quite possible that small membrane patches containing Par3/6 are sufficient to bestow asymmetry to the nascent neuron, irrespective of where the centrosome is positioned. In this scenario, one can envision that these or similar molecules are responsible for 'weakness' on the cortical actin or the plasma membrane, in turn allowing microtubules to enter and support neurite expression. Whether an asymmetric, centrosome-mastered cytoplasm creates a major positional lead for polarity or whether molecular 'polarizers,' working independently or in tandem with centrosome-emerging forces establish initial asymmetry should be tested (**Figures 3(b)–3(d)**).

Irrespective of their degree of participation or hierarchical positioning in the polarity cascade in the postmitotic neuron (i.e., whether a molecular polarizer determines the positioning of the centrosome or vice versa), both these mechanisms support the idea of a mitotically determined process, strongly indicating that the decision of polar growth axis is intrinsically controlled.

From Round to Polar: Making a Neurite Become an Axon

Once sphere-breaking, neurite-inducing forces have been positioned in the sphere, a morphological deformation will result that eventually will become a neurite. Later, from the first or another neurite (depending on the model, **Figure 1(a)** or **1(b)**), the axon will form.

The growth of all neurites, independent of whether they become the axon or a dendrite, is dependent on the directed activity of the microtubule–actin

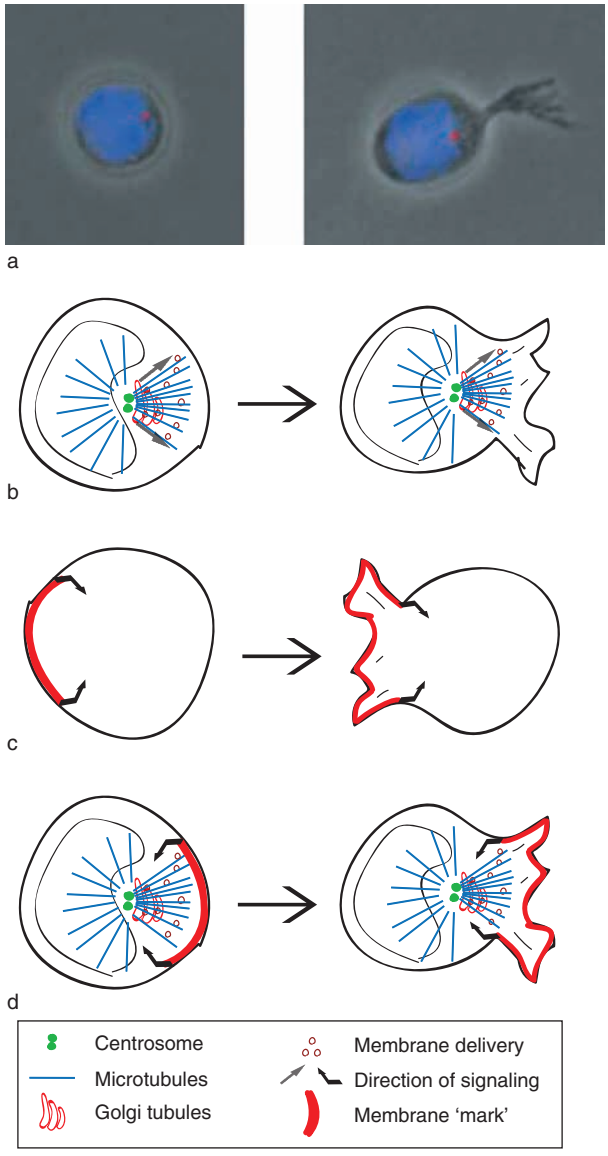

Figure 3 Ways neuronal asymmetry might be produced and might induce sprout formation. (a) The intracellular polarization of neurons coincides with the first sprout. Neurons were fixed after 1–3 h, and the centrosome was labeled in red and the nucleus in blue. (b) Forces originating from the asymmetric centrosome provide directionality and could regulate where the sphere is disrupted. (c) Disruption of symmetry might be explained by a membrane 'mark' that directs neuronal sprouting. (d) Another possibility is that inherited membrane asymmetries signal the inside of the cell to produce and maintain a cytoplasmic polarization.

cytoskeleton and selective membrane transport. The number of neurites will depend on the intrinsic program of the neuron: monopolar, bipolar, or multipolar. In fact, microtubules irradiate from the centrosome/MTOC in all directions of the sphere and are transported into all neurites. Actin is also similarly responsible for the presence of filopodia and lamellipodia in all neurites, and all neurites contain membrane carriers and mitochondria. Obviously, suppression of microtubule polymerization or membrane trafficking or stabilization of actin dynamics results in the lack of neurite outgrowth. The important issue, however, is which mechanisms neurons use to select only one of these neurites for fast growth, so that it later becomes the axon. One possible mechanism is through the simple 'passive' accumulation of 'neurite-making forces' in the place where the initial breakage was made. In this scenario, the neurite formed first will eventually become the axon because it is the place where these forces acted for the longest period and, consequently, where the growth threshold signal is surpassed first. Since the decision on the place on the sphere where the first neurite forms is defined by the positioning of the centrosome, which in turn is a consequence of the plane of cleavage of the last mitosis, this hierarchical, 'quantitative' mechanism implies, necessarily, that the decision of axon destiny is an intrinsic, mitotically predisposed process (Figure 4(a)).

The alternative mechanism is a 'qualitative' one, mediated by the 'sorting' to a single neurite, not necessarily the first, of molecules (called 'polarizers') with the specific capacity to affect microtubule–actin–membrane dynamics. This model of axon selection is illustrated in Figure 4(b). In line with such a predicted mechanism, it has been reported that certain molecules, such as the ganglioside converting enzyme PMGS or the small guanosine triphosphatases (GTPases) Rap1B and Cdc42, are preferentially transported to one neurite before it becomes the axon. Clearly, the polarization of such molecules to the tip (growth cone) of a single neurite can induce a special rate of growth (axonal) by selectively affecting the stability of the actin filaments or microtubules or by increasing membrane fusion or, most likely, by affecting all them in line or in parallel. Consistently, both PMGS and Rap1B signal to the pleiotropic PI3K, whose phosphorylating activity impinges on members of the actin and microtubule regulating proteins. Other molecules whose polarization before axonal growth can induce this growth are the Rho GTPases, a family of proteins with a very well elucidated role in the control of actin cytoskeleton dynamics. Studies of neurite elongation have shown that in general, RhoA inhibits and Cdc42 and Rac1 promote neurite elongation. RhoA activation induces stress fibers in starved fibroblast and leads to a more rigid actin cytoskeleton, whereas Cdc42 induces filopodia and Rac, membrane ruffles. If these actions apply to neurons, one would predict that Rho activity is low and Cdc42 and Rac1 activity are high in the growing

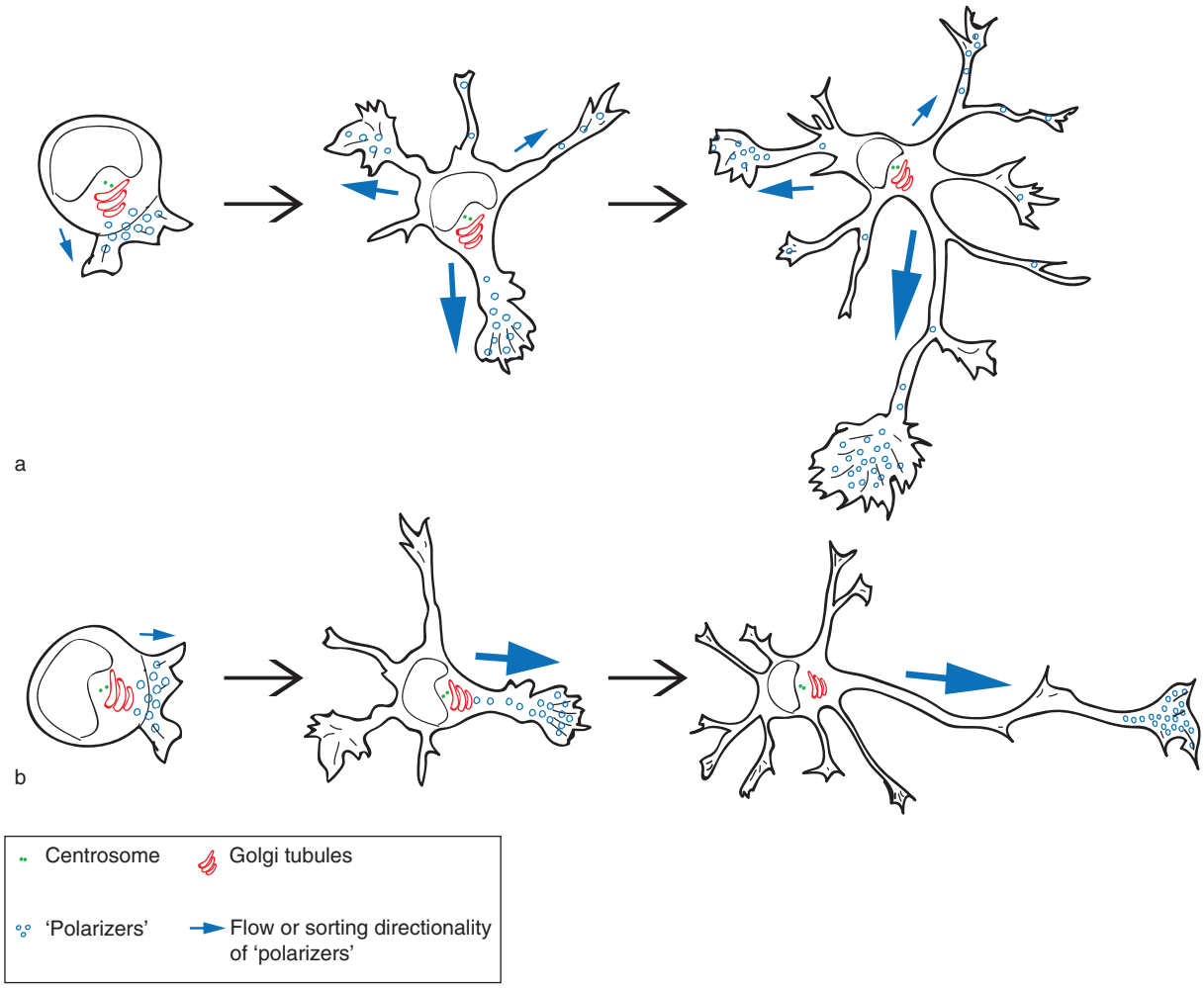

Figure 4 Sorting or cytoplasmic flow could select one neurite for sustained growth. (a) A quantitative mechanism would imply the delivery of 'polarizers' to all neurites, but with a preference for the one that began to grow first. (b) A second 'quantitative' model suggests a sorting process selectively sorting to the axon.

axon. In line with this, inhibition of RhoA has been reported to lead to the outgrowth of multiple axons, and the inhibition of Rac or Cdc42 inhibits axonal outgrowth. As mentioned already, Cdc42 has been found to be accumulated before axonal outgrowth in a single growth cone. From the latter mechanism arises a most basic question: what determines the sorting of the polarizers? One possibility is that it is the consequence of the time–hierarchy quantitative process described in the first place. This would not only reconcile quantitative and qualitative mechanisms but would also stress that the establishment of neuronal polarity is absolutely intrinsically defined. In this scenario, the extracellular environment would play a role in sustaining and guiding the already formed axon. The observation that the activity of PMGS increases locally on the addition of growth factors, an event also observed with the insulin-like growth factor receptor, is additional evidence of the importance of extracellular stimuli in growth sustenance. Moreover, there is abundant literature on the role of different types of extracellular ligands on the induction of axonal or dendritic growth.

A valid conclusion from all these findings is that a quantitative and centrifugal intracellular mechanism confers preferential, but not exclusive, growth capacity on a single neurite, with the extracellular environment playing a critical role in confirming such a predisposition. Thus, neuronal polarity is the result of the synchronous and sequential effect of intrinsic and extrinsic inputs.

Sustaining Polarized Growth

The sudden growth of the 'selected' neurite is a most important aspect of neuronal polarization, signaling to the cell that the normal one-axon, several-dendrites phenotype is finally established. A corollary

is that axonal fate is active whereas dendritic fate is passive (lack of sudden growth destines the remaining neurites of the cell to become dendrites). Obviously, final dendritic architecture requires more processes than a growth rate different from that of axons, and therefore dendritic destiny will be an active process as well. Science is only beginning to understand the mechanisms that lead to dendritic determination once the axon is formed, but the mechanisms responsible for rapid growth of the newly formed axon are fairly well understood. One important, well-described mechanism required for rapid growth is the destabilization of the actin skeleton in the growth cone of the selected neurite. Thus, neurons treated with the actin depolymerizing drug cytochalasin D extend, not one, but several rapidly growing neurites with axonlike characteristics. More important, local administration of this drug to any growth cone provides that selected neurite with axonal growth capacity. Another critical component behind rapid growth is the stability of microtubules. Indeed, the axon contains a higher proportion of acetylated, and thus more stable, tubulin, and a dephosphorylated version of microtubule binding protein tau, with a differential capacity to regulate microtubule dynamics depending on its phosphorylation state, is present in a distal–proximal gradient in axons.

The process of induction of rapid growth and of its maintenance requires the contribution of certain sustaining molecules, probably the same as the aforementioned 'polarizers.' Thus, for instance, the Par proteins described earlier are essential for axonal growth. The role of Par in axonal growth is mediated via kinases, such as the atypical protein kinase C ζ and PI3K, with a profound effect on cytoskeleton dynamics. Recently, it has been shown that the inhibition of the serine–threonine kinase GSK-3 leads to the outgrowth of supernumerous axons and thus is involved in regulating polarity. GSK-3 is interesting since among its many substrates are several microtubule-binding proteins, such as APC, tau, and Map1B. Its activity could therefore control polarization via the regulation of microtubule stability, microtubule bundling, and the dynamics at microtubule ends. In fact, phosphorylation of Crmp-2, a microtubule-binding protein whose ectopic overexpression leads to the outgrowth of extra axons, by GSK-3 reduces its microtubule-binding activity. On the other hand, GSK-3-mediated phosphorylation of APC changes the capacity of APC to bind and stabilize microtubules, influencing neuronal polarity and axon outgrowth.

Thus, any molecule with a critical role in the modulation of one or more of the critical determinants of the neurite-producing forces, whether microtubules, actin, membrane fusion/retrieval, or other events leading to large structural changes, will, to a greater or lesser degree, have an effect on neuronal polarity.

Conclusions and Summary

- *In situ* excitatory pyramidal neurons polarize in the ventricular and subventricular zone immediately after the last neurogenic division by sprouting two opposite neurites, which support migration.
- *In vitro* hippocampal neurons polarize immediately after mitosis: The neurite grown first eventually becomes the axon. Whether the first neurite *in situ* likewise becomes the axon still needs to be confirmed.
- The position from which the first neurite sprouts is marked by the presence of the centrosome and membrane organelles. It remains to be established firmly whether such spatial correlation alone determines the formation of the first sprout or whether a preexisting neurite-forming signal is needed.
- The first neurite grows as the axon because it is in the spot receiving growth information for the longest amount of time. It remains to be established whether the sorting of specific molecules primes such polarized delivery.
- A second neurite can become the axon if the first is lost. It remains to be established whether the selection of the new axon follows the same principles that govern the selection of the first.
- Dendrites begin to differentiate only after axonal destiny is conferred on one neurite.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axonal Pathfinding; Dendrite Development, Synapse Formation and Elimination; Glutamate Regulation of Dendritic Spine Form and Function; Interstitial Axon Branching/Collateral Elimination; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Vesicular Sorting to Axons and Dendrites.

Further Reading

Bradke F and Dotti CG (2000) Changes in membrane trafficking and actin dynamics during axon formation in cultured hippocampal neurons. *Microscopy Research and Technique* 48: 3–11.

Bradke F and Dotti CG (2000) Establishment of neuronal polarity: Lessons from cultured hippocampal neurons. *Current Opinion in Neurobiology* 10: 574–581.

Craig AM and Banker G (1994) Neuronal polarity. *Annual Reviews in Neuroscience* 17: 267–310.

Da Silva JS and Dotti CG (2002) Breaking the neuronal sphere: Regulation of the actin cytoskeleton in neuritogenesis. *Nature Reviews of Neuroscience* 3: 694–704.

De Anda FC, Pollarolo G, Da Silva JS, Camoletto PG, Feiguin F, and Dotti CG (2005) Centrosome localization determines neuronal polarity. *Nature* 436: 704–708.

Dent EW and Gertler FB (2003) Cytoskeletal dynamics and transport in growth cone motility and axon guidance. *Neuron* 40: 209–227.

Dotti CG and Banker GA (1987) Experimentally induced alteration in the polarity of developing neurons. *Nature* 330: 254–256.

Dotti CG, Sullivan CA, and Banker GA (1988) The establishment of polarity by hippocampal neurons in culture. *Journal of Neuroscience* 8: 1454–1468.

Goslin K and Banker G (1989) Experimental observations on the development of polarity by hippocampal neurons in culture. *Journal of Cell Biology* 108: 1507–1516.

Götz M and Huttner WB (2005) The cell biology of neurogenesis. *Nature Reviews of Molecular and Cell Biology* 6: 777–788.

Horton AC and Ehlers MD (2003) Neuronal polarity and trafficking. *Neuron* 40: 277–295.

Kriegstein AR and Noctor SC (2004) Patterns of neuronal migration in the embryonic cortex. *Trends in Neuroscience* 27: 392–399.

Noctor SC, Martinez-Cerdeno V, Ivic L, and Kriegstein AR (2004) Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. *Nature Neuroscience* 7: 136–144.

Wiggin GR, Fawcett JP, and Pawson T (2005) Polarity proteins in axon specification and synaptogenesis. *Developmental Cell* 8: 803–816.

Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons

L S B Goldstein, University of California San Diego School of Medicine, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Size Problems in Neurons

Neurons are usually the largest and most highly polarized cells in multicellular organisms. In addition to the cell body, neurons have a dendritic compartment and an axonal compartment. Focusing on the axon, some simple size calculations reveal just how substantial the burden of their size is. Consider a 1 m motor neuron in a human adult. The 1 m axon can have a diameter ranging from a few tenths of a micrometer up to several micrometers as a consequence of variability in growth in diameter. The diameter of the cell body is on the order of 20–30 μm. Thus, the axon can be as much as or more than a thousand times the volume of the cell body. In addition to extraordinary volume, neurons must move materials such as synaptic proteins and vesicles synthesized in the cell body an enormous distance to reach the synapse where they have essential functions. Since diffusion is inadequate to support distant synapses and large axons and dendrites supplied with appropriate materials, neurons use a supply system composed of microtubule tracks and two types of microtubule-dependent molecular motor proteins, called kinesins and dyneins. These supply systems physically move the required large amounts of material from the cell body to the synapse and also keep the axon and dendrite supplied with ion channels, cytoskeletal proteins, membrane proteins, organelles, and many other components needed for axonal function and viability. In addition, although neurons use electrical signaling in the form of the action potential to communicate with targets and to drive electrical networks, a great many growth factors are moved to and then secreted at axonal termini to signal forward to targets. Simultaneously, targets release growth factors, which must complex with receptors and then be moved back to cell bodies to send chemical signals. Collectively these movements define axonal and dendritic transport. Axonal transport is better understood and consists of fast and slow components. In the case of movement from cell body to synapse, axonal transport is called anterograde, whereas from synapse to cell body, it is called retrograde. Dendrites have comparable burdens but also have a more complicated organization of cytoskeletal elements and a more highly branched structure with shorter processes than the axon. As a result, less is known about transport in dendrites although most work suggests that similar systems and principles are at work.

Microtubule Organization

Microtubules have structural, kinetic, and biochemical polarity (Figure 1). 'Minus' ends in nonpolarized cells are anchored at the centrosome whereas 'plus' ends radiate out toward the cell periphery. In axons, minus ends point toward the cell body, and plus ends point toward the synapse. Dendrites are more complex, with microtubules generally having a mixed polarity orientation in proximal dendrites but having plus end distal organization in the distal dendrites that have been studied. As discussed below, molecular motor proteins have defined directions of movement with respect to microtubules. Thus, a motor protein that walks from minus to plus along a microtubule is likely to play a role in anterograde transport in axons whereas its role in dendrites may be more complex. Similarly, motors that move from plus to minus along the microtubule will have retrograde functions in the axon and, again, more complex roles in the dendrites. In fact, because of the complex organization of microtubules in dendrites, principles relating direction of movements and microtubule organization to functions in the dendrites are at present poorly understood.

Assays for Axonal Transport

Many types of assays have been used to define the characteristics of axonal transport. One type of assay relies on radioactively labeling proteins during synthesis in the cell body and then measuring the rate at which different types of radioactive proteins appear at different distances along the axon (Figure 2). This type of assay led to the discovery that axons have both fast transport, which consists of proteins associated with vesicles, and slow axonal transport, which consists largely of soluble proteins and cytoskeletal proteins such as tubulin, actin, and neurofilaments. A second type of assay relies on physically interrupting the movement of materials along the axon by crushing a nerve containing many axons or by tying a thread around the nerve (ligation) to block physical movement (Figure 3). This assay has been used most commonly with sciatic nerve axons in which cell bodies are clustered together in the vicinity of the spinal cord and axonal termini are distributed in the periphery of the animal. Thus, ligating a sciatic nerve blocks both anterograde transport and retrograde transport and

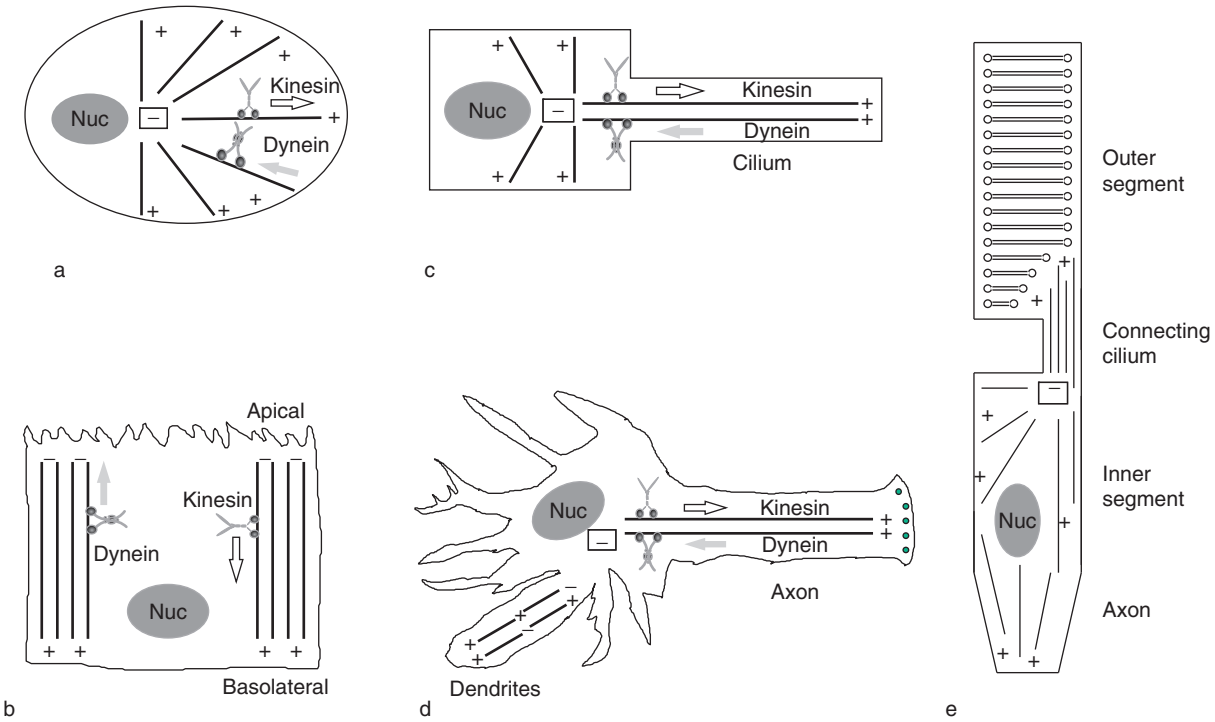

Figure 1 Schematic diagrams of microtubule organization in a variety of cell types. (a) Nonpolarized cell, (b) polarized cell, (c) ciliated cell, (d) neuron, (e) rod photoreceptor. + and −, polarity of microtubules; Nuc, nucleus. Likely directions of motor-mediated movements are indicated (arrows). Reproduced with permission from Goldstein LS and Yang Z (2000) Microtubule-based transport systems in neurons: The roles of kinesins and dyneins. *Annual Review of Neuroscience* 23: 39–71.

leads to accumulation over time of anterograde materials on the proximal side (nearest the spinal cord) and retrograde materials on the distal side (nearest the periphery) of the ligature. Accumulated materials can be assessed either by biochemical means, such as electrophoresis and Western blotting, or by microscopic methods, such as immunofluorescence for specific proteins and organelles. These types of assay have not only led to the realization that different materials are moved at different rates in axons; they have also revealed that some materials that move from cell body to axonal terminus, such as the anterograde kinesin molecular motors, do not return to the cell body and are thus likely to be degraded at the terminus. These assays can be used to characterize both normal behavior and behavior in a variety of mutations that disrupt axonal transport or cause disease by interfering with proper axonal functions. Finally, another important type of assay of axonal transport relies on direct visualization of moving vesicles and organelles in the light microscope in live cells; these methods often use specific labeling of proteins or organelles with fluorescent molecules.

Kinesins

Kinesin molecular motor proteins were originally discovered in the giant axon of the squid. Although not recognized at the time, this new type of molecular motor protein, although biochemically distinct from myosins and dyneins, is the founding member of a diverse but related collection of molecular motor proteins that were deployed for many different functions in eukaryotic cells. The founding member of this family, originally called conventional kinesin, and now called kinesin-1, is a heterotetrameric protein composed of two kinesin heavy chains (KHCs) and two kinesin light chains (KLCs) (Figure 4). The motor activity of kinesin-1 resides in KHC in an N-terminal region of approximately 350 amino acids. This protein domain has been shown to be capable of generating adenosine 5′-triphosphate (ATP)-dependent movements along microtubules *in vitro*, has been crystallized and solved structurally, and has proven to be conserved and shared among all kinesin-related motor proteins. The nonmotor portion of KHC has two structural domains. One structural domain is an alpha-helical coiled coil that mediates KHC dimerization. The other region is of unknown structure and complexes with the KLCs to form the so-called tail or cargo binding domain of kinesin-1. KLC has a short coiled coil domain at its N-terminus that may mediate the dimerization and/or association with KHC, and then has a region of six relatively well-conserved tetratricopeptide repeat (TPR) units, whose structure has been solved in other proteins and which are predicted

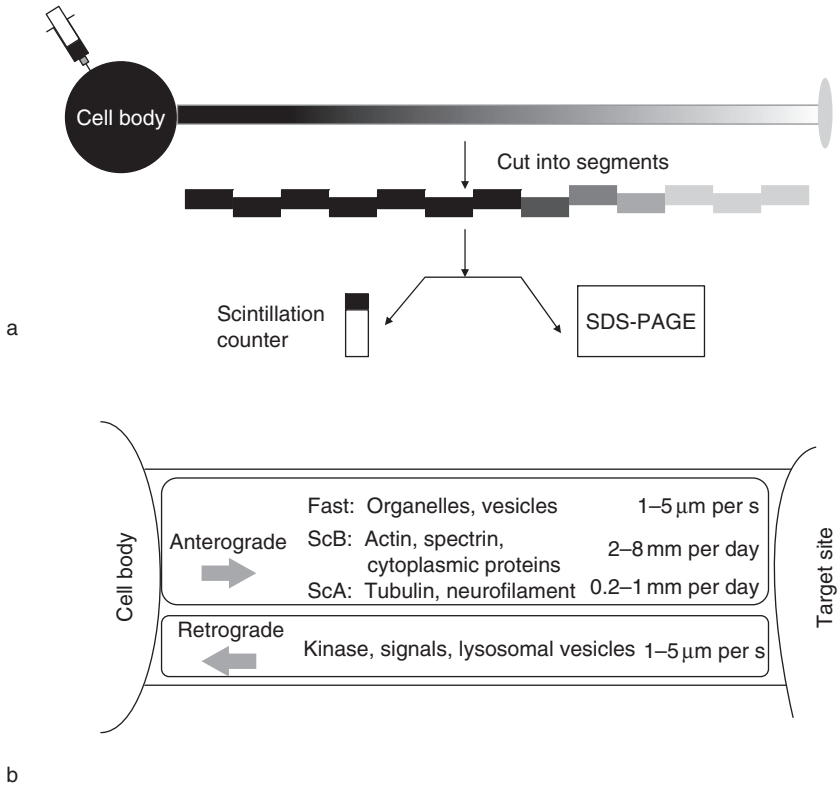

Figure 2 Typical axonal transport assays and pathways. (a) Pulse-labeling experiments are typically used for measuring axonal transport rates and character. Radiolabeled protein precursors are injected into a region containing the cell bodies of particular neurons (e.g., dorsal root ganglion). After different times, the nerves are harvested and cut into segments. The segments can be used to evaluate rates of transport using scintillation counting, sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), and autoradiography, or other analytical methods. (b) Diagram of anterograde and retrograde transport pathways indicating potential cargoes and transport rates. Reproduced with permission from Goldstein LS and Yang Z (2000) Microtubule-based transport systems in neurons: The roles of kinesins and dyneins. *Annual Review of Neuroscience* 23: 39–71.

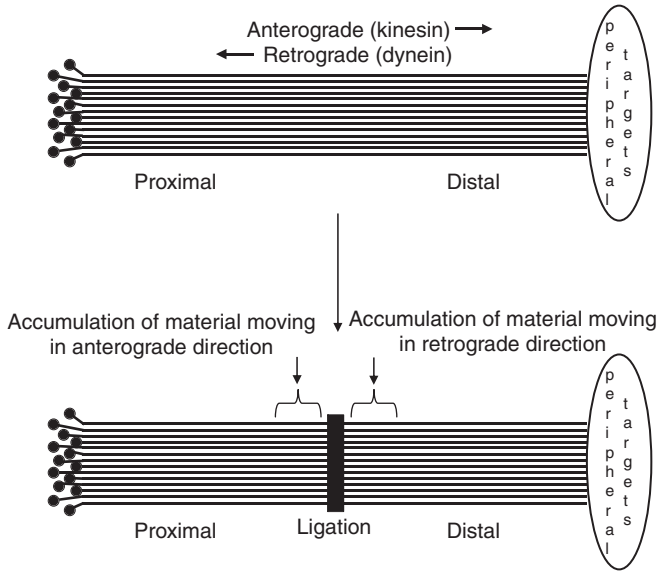

Figure 3 Schematic diagram of sciatic nerve ligation assay. Cell bodies clustered near the spinal cord define the proximal end of the nerve and are the source of materials moving in the anterograde direction. Axonal contacts with peripheral targets at the distal end of the nerve are the major source of material moving in the retrograde direction. Ligation (tying off) of the nerve results in accumulation of materials at the site of ligation.

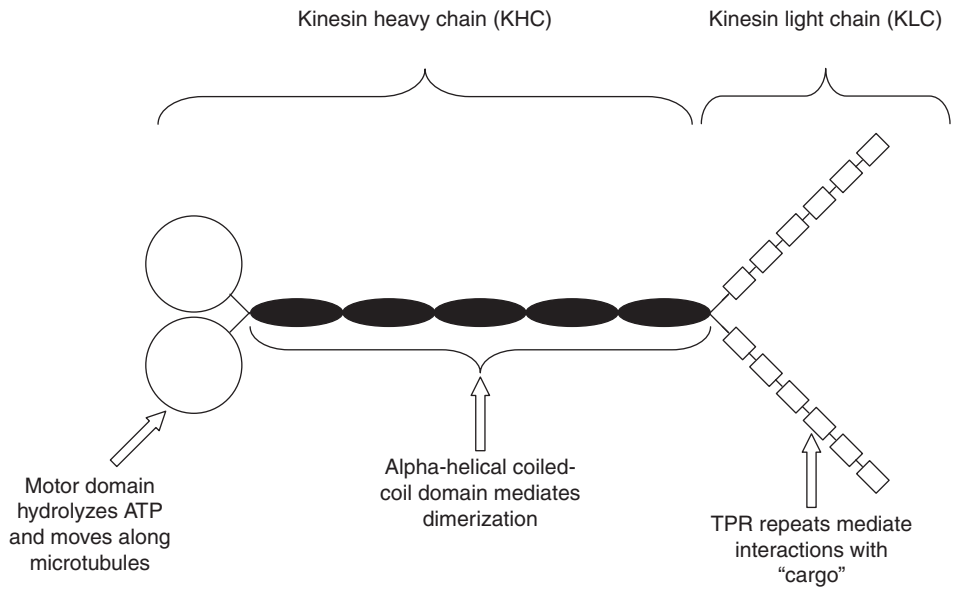

Figure 4 Schematic diagram of kinesin-1 structure. Kinesin-1 is a heterotetramer consisting of two kinesin heavy chain (KHC) subunits and two kinesin light chain (KLC) subunits. KHC contains the motor domain, the dimerization domain, and a short terminal domain that interacts with the KLC subunits. Cargo binding may be mediated by KHC and KLC together or individually. KLC also plays roles in regulating KHC motor domain activity. ATP, adenosine 5′-triphosphate; TPR, tetratricopeptide repeat.

to form a triple alpha-helical groove composed of three TPR units. Thus, KLC is predicted to have two of these triple alpha-helical grooves that may be important in cargo binding and kinesin regulation.

The sequence of the KHC motor domain led to the discovery that eukaryotic genomes encode many different kinesin motor proteins. These proteins are all predicted to share a common motor domain but to have a variety of structurally diverse 'tail' domains attached that may harness force-generating activities by kinesins to a variety of cargoes in various cellular types and functions (Figure 5). Although the initial expectation was that all these predicted motor proteins would be plus end-directed motor proteins, a surprising degree of diversity in behavior was discovered. Some of these kinesin motor proteins mediate minus end-directed movements, and some may not be motor proteins at all and may mediate the rapid depolymerization of microtubules. Both of these types of proteins nonetheless couple ATP hydrolysis to their activities. Some kinesins may have other functional activities in the motor domain, but these have not yet been clearly elucidated. While many kinesins appear to play roles in nonneuronal cells, in particular in the process of cell division, a number are shared between neurons and nonneuronal cells. Many kinesins are specific only to neurons, where they may be involved in axonal or dendritic transport processes. For example, in mammals, on the order of 10–15 of the known kinesins among the 50 or so predicted to be encoded in the genome are specialized for cell division. In *Drosophila*, because of the smaller genome, proportionally fewer are used in mitosis and in neurons. A key question is which motor proteins move which types of vesicles and organelles in each type of neuron.

In neurons, although there is a large diversity of different kinesin motor proteins and cargoes, some simple principles have begun to emerge. For example, kinesin-1 appears to play roles in the anterograde transport of a number of different types of vesicles whose contents are needed along the axonal membrane and at the synapse, such as vesicles containing the amyloid precursor protein (APP), whose processing may cause Alzheimer's disease. Kinesin-1 also appears to play a role in the transport of neurofilaments, which are required for radial growth of axons, and in the movement of mitochondria. It is interesting to note that recent work suggests that mitochondrial movement by kinesin-1 may be mediated by a special form of kinesin-1 in which KLC is replaced by another protein, whose name is Milton. A second prominent class of kinesins, referred to as kinesin-3, appears to play roles in the movement of protein components of synaptic vesicles and may also play a role in the movement of mitochondria. Other kinesins may move mRNA particles, components of postsynaptic complexes and postsynaptic receptors, as well as other organelles, including peroxisomes, elements of endoplasmic reticulum, and the Golgi apparatus.

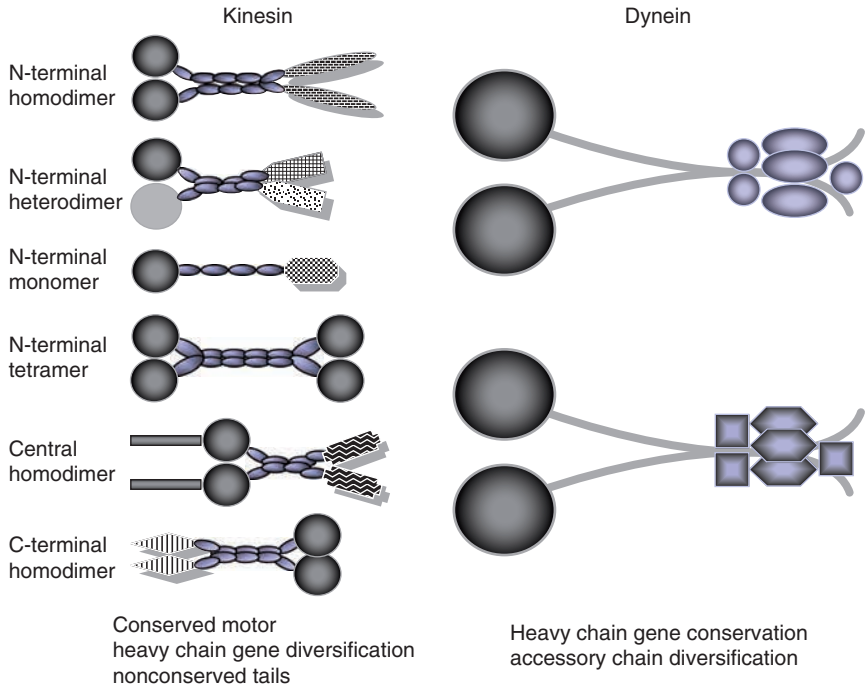

Figure 5 Principles of kinesin and dynein diversification. Kinesins appear to have evolved by gene duplication and divergence such that a conserved motor (shaded ovals) has become harnessed to a wide array of diverse tail domains. The location of the motor domain in the heavy chain polypeptide allows kinesin motors to be divided into several groups: an N-terminal group, a central group, and a C-terminal group. In addition, N-terminal kinesins have been found as homodimers, heterodimers, homotetramers, and monomers. Cytoplasmic dyneins appear to have conserved the bulk of their principal motor polypeptide and have generated far fewer divergent genes. Accessory chain diversification may be the principal mechanism for functional specialization. Reproduced with permission from Goldstein LS and Yang Z (2000) Microtubule-based transport systems in neurons: The roles of kinesins and dyneins. *Annual Review of Neuroscience* 23: 39–71.

An important and unresolved question is whether the movements of nonmembrane particles and proteins are mediated by direct interactions with the motor proteins or by binding to vesicle or organelle membranes and hitchhiking.

Another type of kinesin used for movements in some types of neurons is called kinesin-2. This kinesin is generally a heterotrimer of two nonidentical motor subunits associated with a third nonmotor domain protein, which may form part of the tail. This class of motor protein has been found to be used primarily for movement of materials in cilia and is thus necessary for the construction of cilia and perhaps for signaling mediated by cilia. In neurons in particular, this class of kinesin appears to be used most prominently in sensory neurons of various types, which often use modified cilia as their sensory endings within which receptors for odors or chemical molecules are bound and signals are transmitted. Vertebrate photoreceptors have a modified cilium as their light-sensing domain, the so-called outer segment, and this modified cilium also appears to require the use of this kinesin-2 type of motor protein for the movement of photoreceptive components such as opsin.

Dyneins

Dyneins provide a radically different example of the way in which motor proteins may be utilized to generate force and movement within neurons. In fact, most of the common principles learned from studying kinesins or myosins, which appear to have a shared evolutionary ancestor, appear not to be true for dyneins. A typical cytoplasmic dynein protein complex such as that used in neurons is many times the size of a native kinesin protein complex. The dynein complex generally consists of two copies of a dynein heavy chain and then several intermediate, light intermediate, and light chains, which together with the dynein heavy chains form the dynein complex and define its functions (Figure 5). The motor region of dynein is present in the heavy chains and is structurally and mechanistically different from kinesins and myosins, as it appears to use a mechanism most similar to that found in AAA ATPases. Cargo binding and regulatory activities appear to be conferred by the other subunits of dynein. Unlike kinesins and myosins, cytoplasmic dynein heavy chains are not highly duplicated and diversified in eukaryotic genomes.

Mammals have at most two or three cytoplasmic dynein heavy chain genes and a few each of genes encoding the other subunits. It has been argued that this small and relatively nondiverse collection of dynein heavy chain proteins carry out a large number of different functions in neurons, such as retrograde transport of compartments derived from endosomes and retrograde transport of components involved in signaling from the synapse to the nucleus. The needed diversity of dynein function and regulation is suggested to be provided by diversity in the accessory subunits. An unanswered question is why dyneins and kinesins have evolved so differently in eukaryotic organisms. It is conceivable that this divergent set of structural principles reflects an unknown principle about the regulatory needs in complex cells such as neurons.

Motor Regulation

Logic and experimental evidence suggests that molecular motor proteins must be regulated so that they do not move unless they are associated with an appropriate cargo. This issue is best understood for kinesin-1, which has been the subject of many studies of regulation and attachment. Early biochemical studies of kinesin-1 revealed that even though the protein could generate movement along microtubules *in vitro*, its ATPase activity measured in solution was much lower than could account for the physical parameters of movement. The answer was revealed in a series of studies that removed the tail domain from conventional kinesin in a series of recombinant constructs. These studies found that the kinesin motor domain with its tail domains attached has a very low ATPase activity while the kinesin motor domain with the tail domain removed has an ATPase activity that is high enough to account for the rates of movement seen *in vitro* and *in vivo*. In addition, structural studies revealed that kinesin-1 could exist in either a folded conformation, in which the tail and the head appeared to be associated, or in an extended conformation, in which the tail and the motor were separated; the latter conformation looks most appropriate for generating active movement. Although all the details have not yet been elucidated, considerable evidence now suggests that some combination of coupling kinesin-1 to an appropriate cargo combined with the action of soluble proteins leads to activation of the motor activity, perhaps by unfolding the motor. There is some evidence that phosphorylation mediated by signaling kinases may also be an important part of this regulation, although the details of how this is done are not yet clear. An intriguing possibility is that the signaling kinases might regulate motor protein association with appropriate cargoes in the cell.

Numerous biophysical studies have revealed that like other track-walking enzymes such as polymerases, motor proteins exhibit variation in a potentially regulatable property referred to as processivity. A simple definition of this property is that it refers to the number of sequential reaction cycles or steps along a microtubule or actin filament that a motor protein will undergo following a single encounter of the motor with the appropriate track. For example, a highly processive motor protein will take many steps, sometimes hundreds, along a microtubule after encountering it. A poorly processive motor protein will take at most a few or even only one step after an initial encounter. The biophysical basis of this behavior has been argued to be a result of the fraction of the motor's enzymatic/mechanochemical cycle that is occupied by the strongly bound state of the motor to the microtubule. Thus, a poorly processive motor is one for which the bulk of the reaction cycle is occupied by either a weak or a detached state of association of the motor to the microtubule whereas a highly processive motor is characterized by a large fraction of the cycle's being occupied by the strongly bound state. One consequence is that in solution, a single molecule of a poorly processive motor cannot walk very far, if at all, along the microtubule, whereas a highly processive motor can walk for great distances. Another consequence is that as the number of highly processive motors increases on moving particles, velocity is not expected to increase although the distance traveled is expected to increase. For poorly processive motors, both velocity and distance traveled are expected to increase as motor number increases. High processivity is a property that one might expect of a motor whose natural function is to move small vesicles over long distances. Indeed kinesin-1 appears to have as one of its functions the movement of small vesicles over long distances, and most work suggests that only a very few kinesin motors are found on most vesicles. Another consequence of this biophysical behavior is that processivity itself might represent a regulatory event for motor proteins. For example, as discussed below, kinesin-3 may be an example of clustering of single molecules of a poorly processive motor leading to highly processive behavior because one motor will remain bound while the others are detached. Cytoplasmic dynein in neurons also appears to exhibit this property, but the mediator of enhanced dynein processivity appears to be a protein complex called dynactin. Dynactin has been suggested to link dynein to some cargoes such as vesicles and organelles and also to enhance the processivity of the dynein motor by

providing a static microtubule binding function that tethers the dynein vesicle complex to a microtubule when the motor protein is in the detached state. Dynactin may thus be required for processive movement of vesicles that have small numbers of dynein molecules on their surfaces.

Another possible regulatory mechanism has emerged from studies of kinesin-3, which as a soluble native protein appears to be monomeric and has been generally found to be a poorly processive motor protein. This motor protein has a lipid-binding domain in its tail, and there is intriguing evidence that induced clustering of lipids can bring multiple kinesin-3 subunits together on the surface of a vesicle. This clustering may allow the kinesin-3 proteins to oligomerize and initiate processive movement through the action of multiple motor domains that become clustered and coordinated.

A final biophysical behavior that has been observed and that might be related to processivity and the ability of one or more molecules to generate movement is that some motor proteins, such as kinesin-1, track along a single protofilament of a microtubule in a highly linear movement process. Motor proteins such as dynein, however, appear to wander over the microtubule surface as they generate net force and movement. Whether this is a biophysical curiosity or is in some way related to the differing functions of these two motor protein types remains to be answered.

Mechanisms of Motor Protein Attachment to Vesicles and Organelles

The mechanisms by which dynein and kinesin motor proteins attach to vesicles, organelles, and other cargoes in neurons remain poorly understood. A number of candidate proteins have been found by use of yeast two-hybrid assays and other biochemical measures of protein interactions. In many of these cases, however, it is unclear whether the interactions that have been identified are required for attachment of motor proteins to cellular cargoes or whether these interactions reflect transient regulatory interactions that may occur on the surface of the cargo. In only a very few cases has strong evidence been reported for a role of an identified protein or lipid to link a motor protein to a cargo. For example, for kinesin-3 there is strong evidence that lipid associations can in some circumstances be sufficient to link this motor protein to a vesicle. How important the lipid interaction is for normal kinesin-3 attachment, and whether other protein factors might be involved, is unclear. For kinesin-1, a number of proteins have been suggested to be important for attachment, including APP, which is implicated in causing Alzheimer's disease. Evidence for this function is strong but not yet definitive, and there is some evidence that these interactions may also, or alternatively, play a role in kinesin-1 regulation. Another set of candidates for kinesin-1 attachment proteins is the c-Jun N-terminal kinase (JNK) scaffolding proteins, which may play roles in damage signaling in axons and dendrites and for kinesin-1 attachment to vesicles. There is some evidence that JNK scaffolding proteins are part of the mechanism by which kinesin-1 may be attached to APP vesicles. For dyneins, there is widespread acceptance that dynactin is important for attachment of dynein to vesicles and organelles, but definitive evidence for this point is lacking, and there is some evidence that is inconsistent with this view. At this time, considerably more experimental evidence is needed to unravel the relationship between attachment of motors to cargoes and their regulatory control. Novel methods may be needed to detect and demonstrate the functions of these interactions.

Conclusion and Perspectives

Although much has been learned about the general features of the systems that underlie movement in axons and dendrites of neurons, it is clear that many important details are not yet known. In particular, how motor proteins are attached to cargoes and regulated is particularly poorly understood. Understanding these issues is not only important for understanding the special functions of neurons in networks and systems; an emerging theme is that insults to these systems may be important in the development of a wide range of neurodegenerative disorders such as Alzheimer's disease and Huntington's disease.

See also: Axonal Transport Disorders; Axonal mRNA Transport and Functions; Axonal Transport and Alzheimer's Disease; Axonal Transport and Neurodegenerative Diseases; Axonal Transport Tracers; Axonal and Dendritic Identity and Structure: Control of; Microtubule Associated Proteins in Neurons; Mitochondrial Organization and Transport in Neurons; Perceptual Learning and Sensory Plasticity; Slow Axonal Transport.

Further Reading

Blasius TL, Cai D, Jih GT, et al. (2007) Two binding partners cooperate to activate the molecular motor Kinesin-1. *Journal of Cell Biology* 176: 11–17.

Cai D, Hoppe AD, Swanson JA, and Verhey KJ (2007) Kinesin-1 structural organization and conformational changes revealed by FRET stoichiometry in live cells. *Journal of Cell Biology* 176: 51–63.

Goldstein LS (2003) Do disorders of movement cause movement disorders and dementia? *Neuron* 40: 415–425.

Goldstein LS and Yang Z (2000) Microtubule-based transport systems in neurons: The roles of kinesins and dyneins. *Annual Review of Neuroscience* 23: 39–71.

Hackney DD (2007) Biochemistry: Processive motor movement. *Science* 316: 58–59.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews Neuroscience* 6: 201–214.

Lawrence CJ, Dawe RK, Christie KR, et al. (2004) A standardized kinesin nomenclature. *Journal of Cell Biology* 167: 19–22.

Miki H, Okada Y, and Hirokawa N (2005) Analysis of the kinesin superfamily: Insights into structure and function. *Trends in Cell Biology* 15: 467–476.

Nakata T and Hirokawa N (2007) Neuronal polarity and the kinesin superfamily proteins. *Science's STKE: Signal Transduction Knowledge Environment* 372: pe6.

Pfister KK, Fisher EM, Gibbons IR, et al. (2005) Cytoplasmic dynein nomenclature. *Journal of Cell Biology* 171: 411–413.

Schroer TA (2004) Dynactin. *Annual Review of Cell and Developmental Biology* 20: 759–779.

Stokin GB and Goldstein LS (2006) Axonal transport and Alzheimer's disease. *Annual Review of Biochemistry* 75: 607–627.

Vale RD (2003) The molecular motor toolbox for intracellular transport. *Cell* 112: 467–480.

Relevant Websites

http://www.proweb.org – Kinesin Home Page.

http://www.dynein.org – The Dynein Homepage, Faculty of Biological Sciences, University of Leeds, UK.

Axonal Injury in Demyelinating Disease and CNS Injury

G A Criste, G J Kidd, and B D Trapp, Cleveland Clinic Foundation, Cleveland, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Multiple sclerosis (MS) is an inflammatory neurodegenerative disease of the central nervous system (CNS). MS is the leading cause of nontraumatic neurological disability in young adults in the United States and Europe, and approximately 2.5 million people worldwide suffer from this disease. The majority of MS patients initially exhibit a relapsing-remitting (RR) disease course (RR-MS) characterized by alternating episodes of neurological disability and recovery. Patients eventually convert to a secondary-progressive disease course (SP-MS) characterized by persistent and increasing neurological disability. Classically, MS has been characterized as a disease stemming from focal immune-mediated demyelination, which causes disability through block of conduction along demyelinated axons. More recently, however, the fundamental role of axonal injury in promoting irreversible disability in MS has been recognized and efforts to reduce or prevent axon loss have become a major focus of MS research.

Evidence that axonal injury is an early and persistent event in the progression of MS pathology comes from several sources. Histopathological studies have documented axonal loss in MS lesions (Figure 1), and neuroimaging approaches, including magnetic resonance imaging (MRI) and MR spectroscopy (MRS), have indicated that axon loss is pivotal and dynamic during the evolution of MS pathology. Proposed mechanisms of axonal injury in MS, however, remain diverse and speculative. Although generally considered to be sequelae of inflammatory demyelination, the limited success of immunotherapy in treating progressive disability in MS patients has diverted attention to neuroprotective therapies. The finding that diffused, irreversible axonal transection occurs early in the course of this complex disease has underscored the need to develop axonal neuroprotective treatments.

Axonal Pathology in MS Lesions

Histological evidence of axonal pathology was reported in the early literature on MS. These reports include descriptions of axonal swellings, axonal transection, and Wallerian degeneration, as well as discussions regarding the functional consequences of such pathology. However, the extent of axonal loss was controversial. In their classical works on MS, both Charcot and Marburg described MS pathology largely in terms of demyelination and reactive gliosis, and emphasized the relative sparing of axons in the lesions. In 1936, Putnam reported a 50% loss of axons in MS lesions from 11 patients. That same year, however, Greenfield and King reported normal axon densities in more than 90% of MS lesions from 13 patients that they studied. The differences between these studies were suggested to result from more sensitive axon staining in the latter.

More recent studies using contemporary technology such as MRI, MRS, and confocal microscopy have demonstrated that axonal transection begins at disease onset and that cumulative axonal loss provides the pathological substrate for the progressive disability that most long-term MS patients experience. Moreover, postmortem studies have shown that several histopathological abnormalities, including axonal loss, can be detected in the normal-appearing white matter (NAWM) and cortical gray matter of patients with MS, suggesting a more diffuse pathology than previously thought.

Axonal Damage Occurs during the Early Stage of MS

One important clue that axonal function was disrupted in MS lesions came from studies of amyloid precursor protein (APP). APP is transported by fast axonal transport and is normally present in axons at low levels that are not detectable by immunohistochemistry. APP accumulation in axons located in active MS lesions and at the border of chronic active MS lesions suggested axonal impairment in MS lesions. In particular, some APP immunoreactive structures exhibited the morphology of terminal axonal swellings, suggesting that axonal transection had occurred. The number of APP-labeled axonal swellings correlated with the degree of inflammation in the lesions. Similar swellings were detected with neurofilament immunostaining. Using confocal microscopy and computer-based three-dimensional reconstruction, axonal ovoids were identified as terminal ends of transected axons (Figure 2(a)). In one study quantifying axonal transections in 11 patients with disease duration ranging from 2 weeks to 27 years, extensive axonal destruction was demonstrated in cerebral white matter MS lesions. Active lesions contained over 11 000 terminal ends per cubic millimeter (Figure 2(b)), the edge of chronic active lesions contained over 3000 terminal ends per cubic millimeter, and the core of chronic

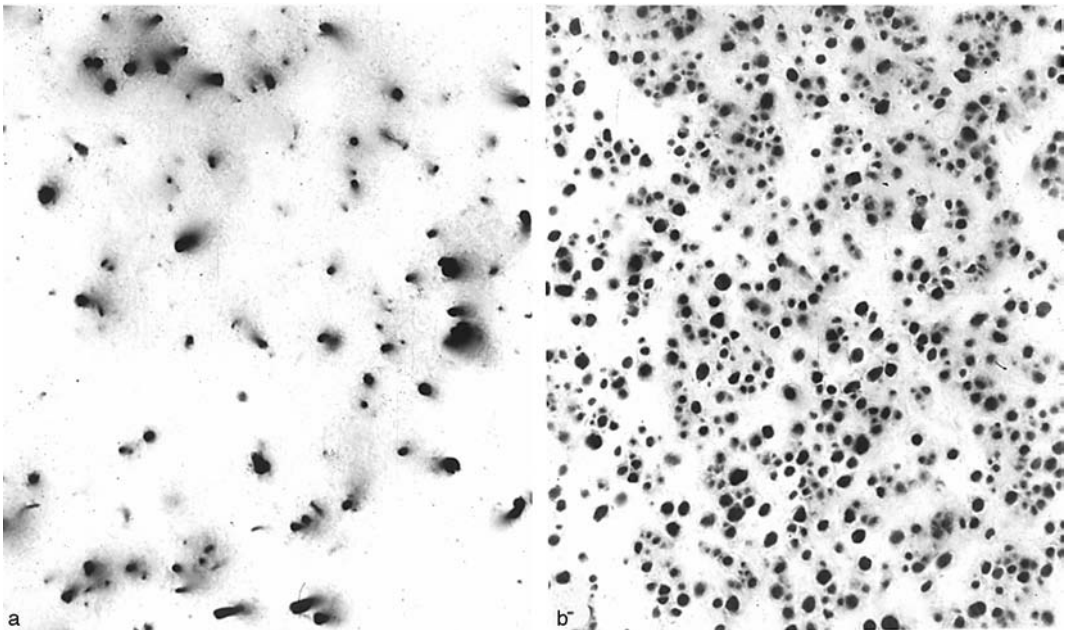

Figure 1 Axonal loss in MS patients. Axonal density is significantly decreased in a demyelinated area of MS spinal cord (a) compared to control cord (b). Reproduced from Bjartmar C, Yin X, and Trapp BD (1999) Axonal pathology in myelin disorders. *Journal of Neurocytology* 28(4–5): 383–395, with kind permission from Springer Science and Business Media.

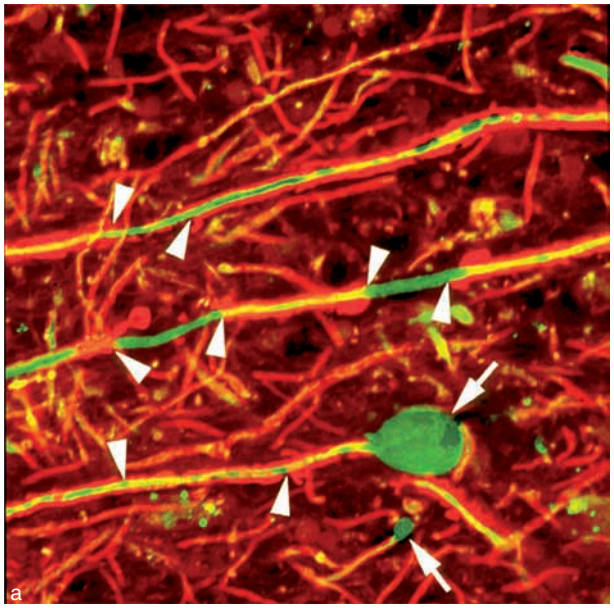

Activity of lesion	*Transected axons/mm³ (mean + sem)*
Active	11 236 + 2775*
Chronic active	
Edge	3138 + 688*
Core	875 + 246*
Nonlesion white matter	15 + 3*
Control white matter	0.7 + 0.7

b

Figure 2 Axonal transection during inflammatory demyelination. Axonal transection occurs during inflammatory demyelination and induces formation of terminal axonal ovoids (a, arrows). When quantified (b), transected axons are abundant in MS lesions and appear to correlate with inflammatory activity of the lesion (* = statistically significant at $p < 0.01$). Reproduced with permission from Trapp BD, Peterson J, Ransohoff RM, Rudick R, Mork S, and Bo L (1998) Axonal transection in the lesions of multiple sclerosis. *New England Journal of Medicine* 338: 278–285, Copyright © 1998 Massachusetts Medical Society. All rights reserved.

active lesions contained on average 875 terminal ends per cubic millimeter. In contrast, less than one transected axon was found per cubic millimeter in control white matter.

Together, these data demonstrate a positive correlation between axonal transection and degree of inflammation in cerebral white matter MS lesions undergoing demyelination. The presence of axonal

ovoids in patients with short disease duration clearly demonstrated that axonal transection begins at an early stage of MS.

Axonal Loss Also Occurs in NAWM

It is well established that axons, once severed, will undergo relatively rapid Wallerian degeneration distal to the site of transection. Unlike axons, CNS myelin can persist for a long time after transection of the proximal axon. Remaining myelin sheaths may appear as empty tubes or as degenerating ovoids by microscopy. Despite this histological pathology, the white matter may appear grossly normal and appear unremarkable by conventional neuroimaging techniques.

Immunohistochemical evidence suggestive of Wallerian degeneration, such as discontinuous staining of axonal neurofilaments and presence of terminal axonal ovoids, has been demonstrated in NAWM from MS brains. Reductions in axonal density by 19–42% in areas without plaque have been reported in the lateral corticospinal tract of MS patients with lower limb weakness. By comparing axonal density in lesions and in NAWM from the cervical spinal cords of SP-S patients, one study showed how extensive axon loss can be. The average reduction in axonal density in lesions of lateral and posterior columns was 61%. In NAWM, however, the average decrease in axonal density was as much as 57%. By taking into account both decreased axonal density and reductions in tissue volume, another study showed that total axonal loss in the corpus callosum of MS patients averaged 53%; disease duration was between 5 and 34 years. The measured reduction in axonal density was only 34% in the same material. These findings emphasize the need to consider both tissue volume and axonal density to properly assess the degree of total axonal loss. These studies also indicate that white matter may appear normal judged by immunohistochemistry for myelin, or on MRI scans, but may still exhibit a considerable axonal dropout, especially in chronic patients with long disease duration.

Wallerian degeneration in NAWM has also been described by immunohistochemistry in an MS patient with short disease duration. The patient succumbed to a fatal brain stem lesion just after a 9-month history of RR-MS with few permanent neurological signs. Demyelinated lesions were not found in the spinal cord postmortem. However, the ventral column of the spinal cord, which contains axons projecting from the brain stem lesion, exhibited a 20% axonal loss. Microscopy revealed myelin ovoids and signs of myelin degradation by activated microglia, characteristic of Wallerian degeneration. Since much of the myelin remains, these can represent 'invisible lesions' as far as MRI and immunostaining for myelin are concerned.

Neuronal Pathology Also Occurs in MS Cortex

In addition to the more commonly described white matter locations, MS lesions can also involve the gray matter. The histopathological features and the clinical significance of such lesions are not completely understood. Gray matter lesions are difficult to detect macroscopically and histologically. On conventional T2-weighted MRI images, MS lesions in the cerebral cortex are much less obvious than white matter lesions. For these reasons, the frequency of cortical lesions has often been underestimated. Recent studies have shown that the use of gadolinium enhancement resulted in an increased detection of cortical lesions on MRI scans by 140%. A total of 26% of these enhancing lesions arose within or adjacent to the cerebral cortex. These results suggested that conventional MRI under-reports the presence of cortical lesions, when compared to neuropathological analysis.

In postmortem MS cortex, neurite transection is common, and TUNEL-positive neurons were detected in ~50% of chronic lesions, indicating that apoptotic loss of neurons can be significant. Interestingly, compared to white matter lesions, inflammation is substantially reduced in these lesions. Gray matter lesions contained fewer inflammatory cells, no perivascular cuffs, and consisted mainly of reactive microglia. T cells have been proposed to take a central role in the pathogenesis of MS. While T lymphocytes are found in MS white matter lesions, fewer are detected in cortical lesions that extended through both white and gray matter. In intracortical demyelinated lesions, T cell numbers were not different to T lymphocyte density in nondemyelinated cerebral cortex within the same tissue block.

It has been hypothesized that injury to neurons in cortical or subcortical lesion may provide the biological correlate to the cognitive dysfunction many MS patients experience. In fact, executive and cognitive functional deficits arise in 40–70% of these patients. Increased knowledge regarding mechanisms of neuronal damage in cortical MS lesions will contribute to the understanding of the functional significance of such lesions.

Mechanisms of Axonal Injury in MS

The pathophysiology of axonal injury in MS is poorly understood. It seems likely that different mechanisms may contribute to axonal degeneration at different stages of the disease. Elucidating the cellular and

molecular mechanisms that promote axonal loss in MS is important to the development of future neuroprotective therapies.

Genetic Susceptibility to Axonal Injury

Current evidence indicates that interactions between multiple genes influence the outcome of MS in individual patients. One line of evidence suggesting a genetic influence on axonal damage in inflammatory demyelination comes from studies of Theiler's murine encephalomyelitis virus (TMEV) disease, a viral-induced model of inflammatory CNS demyelination. Infected animals with susceptible genetic backgrounds develop neurological impairment and pathological changes comparable to those in MS. For example, infected SJL/J mice develop chronic demyelination, neurological deficits, and extensive loss of axons in the spinal cord, while C57BL/6x129 mice are resistant to TMEV-induced disease. When the major histocompatibility complex (MHC) class I protein is absent from C57BL/6x129 mice, they develop similar demyelination to SJL/J mice after infection, but do not go on to develop functional disability. In contrast, C57BL/6x129 mice lacking MHC class II developed a chronic inflammation with neurological deficits and axonal degeneration in spinal cord white. These results indicate that MHC molecules are involved in susceptibility to demyelination and axonal damage and highlight the possible role that genetic differences can play on the development of pathology during inflammatory demyelination.

Genes encoding trophic factors that are involved in neuroprotection and repair are at least as important as the immune-related genes that are thought to promote pathology. In this context, ciliary neurotrophic factor (CNTF) is an interesting candidate to promote resistance to axonal injury in MS. One study described a correlation between CNTF null mutations and early onset of MS symptoms. This suggests that axonal loss, which is the basis of disability, may be accelerated in these individuals through lack of CNTF-mediated trophic support of neurons and oligodendrocytes. Although a second study reported no correlation between CNTF genotype and disease, the initial results accord well with observations of experimental allergic encephalomyelitis (EAE) in CNTF knockout mice. After induction of EAE with myelin oligodendrocyte glycoprotein (MOG), $CNTF^{-/-}$ mice showed a significantly earlier disease onset and a delayed recovery from relapses.

Axonal Injury and Inflammation

Current knowledge suggests that MS is a primary inflammatory demyelinating disease of the CNS, and that disease activity reflects CNS inflammation, although much of the disease may not be detected clinically. For example, most RR-MS patients exhibit progressive brain atrophy and persistent inflammation, identified as gadolinium-enhancing lesions by MRI, regardless of their clinical symptoms. Since axon pathology, including transection of axons, correlates with the degree of inflammation in MS lesions, early axonal transection may reflect vulnerability of the demyelinated axons to inflammation. Indeed, the inflammatory microenvironment contains a variety of substances that could potentially injure axons, including proteolytic enzymes, cytokines, oxidative products, and free radicals produced by activated immune and glial cells. The ability of cytotoxic $CD8^{+}$ T-cells to mediate axonal transection has been described in MS tissue, EAE mice, and *in vitro*. Treatment of EAE mice with the AMPA/kainate glutamate receptor antagonist NBQX resulted in increased oligodendrocyte survival and reduced axonal damage.This suggests that excitotoxicity mediated by glutamate is involved in tissue damage in acute lesions. Directly or indirectly, inflammation may also affect energy metabolism of axons (discussed below). Inflammatory intermediates such as nitric oxide (NO) can act directly on mitochondria, while local inflammatory edema may interfere with the blood supply and potentially induce an ischemic mechanism of axonal degeneration.

Regardless of the molecular mechanisms involved, inflammation is a major factor behind accumulating axonal pathology at early stages of MS. For that reason, aggressive anti-inflammatory treatment during RR-MS may have indirect but long-term effects in preventing axonal injury in addition to important direct effects on inflammation.

Loss of Neuroprotective Effects of Myelin

The past decade has seen a deeper understanding of the intricate interdependence between the myelin-forming cells and its associated neuron. For example, axonal neuregulin has been found to control the proliferation and particularly the survival of oligodendrocytes and Schwann cells, ensuring a good match between the axon surface area requiring myelination and the number of surviving myelinating cells. Myelin-forming cells in turn have a profound influence on axonal morphology and physiology, affecting axonal diameter, neurofilament phosphorylation, cytoskeletal organization, and axonal transport rates. It follows therefore that disease processes that directly affect myelin-forming cells might also influence the underlying axon as well.

Peripheral nervous system (PNS) dysmyelinating diseases (in which myelin forms abnormally) are clearly

associated with axonal changes. Charcot–Marie–Tooth neuropathy type 1 (CMT1) is a genetically heterogeneous group of chronic dysmyelinating peripheral neuropathies. Mutations affecting the myelin genes peripheral myelin protein 22 (PMP22), protein zero (P_0), and connexin-32 account for most CMT1 cases. While the electrophysiological consequences of dysmyelination account for some of the neurological symptoms in CMT1, the main contributor to clinical progression is axon loss, as determined by measurements of nerve conduction amplitudes and motor unit numbers. As in MS, this may be related to abnormal glial–axonal interactions. Thus, axonal degeneration may be a final common outcome in a wide variety of myelin-related diseases.

In the CNS, myelin protein defects also result in axonal pathology, even though they may cause little or no myelin abnormality. Myelin-associated glycoprotein (MAG), a member of the immunoglobulin gene superfamily, is a low-abundance myelin protein enriched in the innermost myelin membrane that lies adjacent to the axon. In MAG-deficient mice, postnatal myelination progresses as in wild-type animals. From the age of 5 weeks, however, progressive axonal atrophy including reductions in axonal caliber, neurofilament spacing, neurofilament phosphorylation, and increased Wallerian degeneration is observed. MAG inhibits neurite outgrowth and causes growth cone collapse *in vitro*, suggesting that it can modulate the axonal cytoskeleton. The findings indicate that MAG has direct or indirect long-term modulating effects on the cytoskeleton via axonal kinases or phosphatases.

The X-linked myelin proteolipid protein (PLP/DM20) is the major structural protein of compact CNS myelin. Mutations, deletions, or duplications involving the PLP gene cause Pelizaeus Merzbacher disease (PMD) and spastic paraplegia of varying severity in humans. In mice, PLP mutations result in premature oligodendrocyte death and dysmyelination. Many of these phenotypes, however, are 'gain-of-function' effects due to toxicity of misfolded proteins encoded by the mutated genes. In contrast, PLP-null mice remain competent to myelinate CNS axons of all calibers and to assemble compacted myelin sheaths. Ultrastructurally, however, the electron-dense 'intraperiod' lines in myelin remain condensed, correlating with its reduced physical stability. From the age of 6 weeks, PLP-deficient mice exhibit focal axonal swellings containing dense bodies and mitochondria in CNS regions containing mainly small-diameter axons. The accumulation and distribution of organelles and neurofilaments in axonal swellings indicate impairment of retrograde axonal transport. A length-dependent axonal degeneration in the absence of demyelination and inflammation in patients with PLP-null mutations has also been reported. Replacement of PLP with P_0 in PLP-null mice resulted in normal-appearing CNS compact myelin with PNS-like membrane ultrastructure. Although initially indistinguishable neurologically from wild-type or PLP-null mice, the P_0-CNS mice exhibited accelerated rates of axonal damage and rapid progression of neurological deficits and morbidity. These results suggest that PLP gene expression is required for axonal maintenance independent of its roles in myelin formation.

Unlike other myelin proteins, CNP is not crucial to myelin assembly but is essential for axonal maintenance. CNP is distributed throughout the cell soma and in noncompacted regions of myelin. In mice lacking CNP, myelin assembly was not visibly affected. However, at about 4 months of age, the mice developed motor deficits that progress with age. Histological analysis revealed late-onset axonal pathology characterized by abnormal axonal swellings and degeneration of many axons.

These data from myelin protein defects support dual roles of the oligodendrocytes, first, in the formation of the myelin sheath, and second, in the maintenance of the underlying axon. Understanding the signaling pathways and effector molecules, such as kinases and phosphatases, may lead to new therapeutics that promote axonal preservation in demyelinating and dysmyelinating diseases.

Mitochondrial Involvement in Axonal Injury

Recent evidence suggests a hypoxia-like metabolic injury as a pathogenetic component of axonal injury in MS. Although this model was largely derived from studies of white matter injury in models of ischemia and neurotrauma, recent observations suggest that such a mechanism operates in inflammatory brain lesions in MS as well. In this model, ischemia leads to ATP depletion. The resulting energy crisis impairs the function of ATP-dependent ion channels (e.g., Na^+–K^+ ATPase, Na^+–Ca^{2+} ATPase) leading to an elevation in intracellular Na^+ concentration. Accumulation of axoplasmic Na^+ through non-inactivating Na^+ channels, together with membrane depolarization, promotes reversal of Na^+–Ca^{2+} exchange and axonal Ca^{2+} overload. Ultimately, the pathological increase in intracellular Ca^{2+} drives Ca^{2+}-dependent enzymes to damage the axon.

Mitochondrial damage could occur through several mechanisms. Astrocytes, activated microglia, and macrophages in the CNS release substantial amount of NO in MS lesions. One mechanism of the toxic action of NO is the impairment of mitochondrial function, and NO-treated white matter has

been shown to undergo ATP depletion and irreversible damage. Mitochondrial dysfunction has also been implicated by microarray analysis of postmortem MS motor cortex. Decreases in nuclear-encoded mitochondrial genes from four of the five respiratory chain complexes have been reported, which raises the interesting possibility that inherent defects in these organelles may arise in MS and further compromise energy production capacity.

Microvascular pathology also contributes a major role in the hypoxic MS pathology, starving mitochondria of oxygen and reducing their capacity to produce ATP. Edema within inflammatory lesions leads to focal disturbance of microcirculation with subsequent ischemia. Such a mechanism is likely to have greatest impact in CNS locations where room for tissue expansion is limited, such as the spinal cord and the optic nerves. Inflammatory damage to the vessel walls can also lead to activation of the clotting cascade resulting in local microvascular thrombosis and ischemic injury to the axons similar to a stroke.

Axonal compensation after demyelination may place increased demands on axonal mitochondria. Demyelinated axons revert to continuous rather than saltatory conduction by express Na^+ channels along the length of the axolemma. However, propagation of action potential under this circumstance involves much greater ion movements and hence ATP consumption. Recent reports on the beneficial aspects of Na channel blockers such as flecainide, phenytoin, and tetrodotoxin in attenuating axonal pathology in animal models of MS may reflect reduced energy demands on the demyelinated axons. Na^+ channel blockers can also reduce neuroinflammation through actions on the microglia and macrophages in EAE and MS. Thus, in addition to direct neuroprotective effect, Na channel blockers may reduce axonal degeneration through anti-inflammatory effects.

Strategies for Axonal Protection

Therapeutic strategies aimed at preventing axonal and neuronal damage are crucial in heading off permanent disability in MS patients. Most neuroprotective strategies are based on the emerging models presented above, which have provided novel perspectives on providing axonal protection.

Anti-Inflammatory Strategies

The correlation between axonal damage and magnitude of inflammation suggests that the axons may be innocent bystanders in the surrounding inflammatory milieu during active demyelination. The clinical observation that the topographic pattern of irreversible, progressive neurological deficits in MS depends on the localization of the previous inflammatory attacks seems to favor this interpretation. Anti-inflammatory agents are most effective during the RR phase when inflammation dominates the clinical picture. Since many axons are lost during this phase, early treatment with these agents is likely to be highly beneficial.

Remyelination Strategies

Myelin contributes both as a physical barrier between the axon and surrounding milieu and by providing trophic support, and so strategies that aid in remyelination can confer axonal protection. Remyelination is common in MS but not sufficiently robust to promote full and sustained recovery. Strategies based on enhancing production of endogenous myelinating cells or supplementing them by transplantation of progenitors are being explored at present.

Interruption of the Secondary Injury Cascade in Axons

In the hypoxia-like models of axonal injury, different cellular insults result in impairment in energy production that leads to a reversal of the Na^+–Ca^{2+} exchanger. The resulting surge in Ca^{2+} levels in the axoplasm drives enzymatic processes that lead to cellular destruction. Based on these pathogenetic mechanisms, treatment with Na channel blockers and Na^+–Ca^{2+} channel blockers may prevent this Ca^+-driven autolysis of neuron. In EAE models and *in vitro* studies, Na channel blockers like phenytoin and flecainide may have succeeded in providing axonal protection through this mechanism. Bepridil, an inhibitor of Na^+–Ca^{2+} exchange has been shown to protect axons from injury caused by NO *in vitro*. These drugs are worth looking into as some of them have already been in the market for decades. Once proven effective in MS, we have an immediate addition to our arsenal against MS that has a relatively well-established safety profile.

Conclusion

Axonal loss is the major cause of irreversible disability in MS. Axonal damage, including complete transection of the axon, begins early in MS and correlates with inflammatory activity. Extensive reductions in axon number may also impact nonlesioned white and gray matter. Understanding the mechanisms underlying axonal degeneration and how these mechanisms operate at different stages of disease may ultimately produce novel therapeutic interventions that enhance

axonal survival in MS brains and delay progression of neurological disability in MS patients.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Injury: Neuronal Responses; Demyelinating Diseases; Neuropathy: Metabolically-Induced; Premenstrual Dysphoric Disorder and Postpartum Major Depression: Chronobiology ; Wallerian Degeneration.

Further Reading

Arnold DI (2005) Changes observed in multiple sclerosis using magnetic resonance imaging reflect a focal pathology distributed along axonal pathways. *Journal of Neurology* 252(supplement 5): v25–v29.

Bjartmar C, Kidd G, Mork S, Rudick R, and Trapp BD (2000) Neurological disability correlates with spinal cord axonal loss and reduce *N*-acetyl aspartate in chronic multiple sclerosis patients. *Annals of Neurology* 48: 893–901.

Bjartmar C and Trapp BD (2001) Axonal and neuronal degeneration in multiple sclerosis: Mechanisms and functional consequences. *Current Opinion in Neurology* 14: 271–278.

Bjartmar C and Trapp BD (2003) Axonal degeneration and progressive neurologic disability in multiple sclerosis. *Neurotoxicity Research* 5: 157–164.

Bjartmar C, Wujek JR, and Trapp BD (2003) Axonal loss in the pathology of MS: Consequences for understanding the progressive phase of the disease. *Journal of Neurological Sciences* 206: 165–171.

Bjartmar C, Yin X, and Trapp BD (1999) Axonal pathology in myelin disorders. *Journal of Neurocytology* 28(4–5): 383–395.

Bruck W (2005) The pathology of multiple sclerosis is the result of focal inflammatory demyelination with axonal damage. *Journal of Neurology* 252(supplement 5): v3–v9.

Evangelou N, Konz D, Esiri MM, Smith S, Palace J, and Matthews PM (2000) Regional axonal loss in the corpus callosum correlates with cerebral white matter lesion volume and distribution in multiple sclerosis. *Brain* 123(Pt. 9): 1845–1849.

Kornek B and Lassmann H (1999) Axonal pathology in multiple sclerosis. A historical note. *Brain Pathology* 9: 651–656.

Peterson JW, Kidd GJ, and Trapp BD (2005) Axonal degeneration in multiple sclerosis: The histopathological evidence. In: Waxman S (ed.) *Multiple Sclerosis as a Neuronal Disease*, pp. 165–184. New York: Elsevier.

Stys PK (2004) Axonal degeneration in multiple sclerosis: Is it time for neuroprotective strategies? *Annals of Neurology* 55: 601–603.

Trapp BD, Peterson J, Ransohoff RM, Rudick R, Mork S, and Bo L (1998) Axonal transection in the lesions of multiple sclerosis. *New England Journal of Medicine* 338: 278–285.

Relevant Website

http://www.nationalmssociety.org – The National Multiple Sclerosis Society.

Axonal Injury: Neuronal Responses

G Raivich, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Injury to neurons results in a host of many different structural, electrophysiological, and molecular responses. Some of these are associated with and may play an important role in the mounting of a successful response and the ensuing recovery of function. Others reflect the enhanced likelihood of cell death, either within hours or days of the injury, particularly in neonatal animals, or during a delayed form after a brief or lengthy regenerative phase, and/or the gradual withering away and atrophy, processes that can be addressed with neurotrophic therapy.

Axonal injury and the ensuing neuronal response can also induce a strong glial and immune reaction. These reactions can have a positive or a negative impact on neuronal regeneration. For example, cell death and the ensuing inflammation can impart a strong stimulus for the generation of new neurons. The nonneuronal responses will also provide the cellular basis for the immune surveillance of the injured brain, a key component of their overall function. The aim of this article is to present a concise overview of specific molecular, structural, and electrophysiological changes occurring in the injured neuron itself.

Early Axonal Reaction

The Initiation

The physiological and molecular signals that act as sensors of axonal injury and induce the regenerative program are still a subject of debate. Generally, three major changes appear to provide separate signals, even though they may act in a complementary and/or synergistic way:

1. Ionic flux, injury-induced discharges. There is a rapid entry of extracellular ions such as calcium and sodium through the transiently opened axonal plasma membrane, before it is resealed, resulting in depolarization and a sequence of injury-induced action potentials. The calcium influx and bystander activation of intra-axonal proteases and cytoskeleton remodeling underlie growth cone formation and may also be involved in changing intra-axonal protein synthesis.

Early changes in the cell body of axotomized neurons precede the arrival of molecules traveling by rapid axonal transport. They include the rapid elevation of calcium and cAMP, ribosomal RNA processing, and the increase of gap junction proteins such as connexin-43 in neighboring astrocytes. However, inhibition of antidromic electrical discharges does not interfere significantly with the somewhat later induction in transcription factors, glucose utilization, or the onset of regeneration.

2. The tip of the injured axon is briefly exposed to the intracellular content of neighboring axons and Schwann cells. These substances include the ciliary neurotrophic factor (CNTF), neurotrophin-3 (NT3), and fibroblast growth factors (FGFs), which are released following injury, deposited locally, and can then act on the proximal tips of the injured axons to induce a specific response. At a later time, the axons are also exposed to the extracellular environment of the inflamed neural tissue, containing a host of different proinflammatory cytokines, such as interleukin-1 and interleukin-6.

Particularly instructive is the role of CNTF and the nuclear transcription factor, signal transducer and activator of transcription-3 (STAT3), the downstream target of activated CNTF receptor. Phosphorylation and activation of STAT3 in the proximal part of injured axons clearly precede that in axotomized neurons, pointing to the retrograde transport of phosphorylated STAT3 as one of the injury-mediating signals. Neuron-specific deletion of STAT3 leads to enhanced posttraumatic cell death, suggesting that this signal is involved in promoting survival of lesioned neurons. Moreover, deletion of CNTF, a neurotrophin that is abundantly present in myelinated Schwann cells but not in or around the cell bodies of axotomized motor neurons, causes a significant delay in the appearance of the phosphorylated STAT3 and its nuclear translocation in neuronal cell bodies.

3. Block of retrograde transport. Last but not least, axonal injury also interferes with the retrograde flow of trophic signals from the normal innervation target. The data on nerve growth factor (NGF), the model neurotrophin that is retrogradely transported from periphery to neuronal cell body, are particularly instructive. In fact, analysis of its retrograde transport in the injured sciatic nerve shows a very drastic decrease immediately following axotomy. However, this initial, very strong decrease is only transient, for approximately 48 h. Nevertheless, there is a much longer, still threefold reduction that continues until the onset of regeneration. Moreover, NGF deprivation by neutralizing the endogenous activity with specific antibodies also induces axotomy-like changes

even in the intact, NGF-sensitive sensory and sympathetic neurons, particularly in the expression of transcription factors such as c-jun and a series of neuropeptides including galanin, vasoactive intestinal peptide (VIP), SP, CGRP, and NPY.

Retrograde Transfer of Injury Signal

The second, intermediate phase involves the posttraumatic generation and activation of axoplasmic proteins with the nuclear localization signal (NLS) sequences. These NLS proteins function as positive and retrogradely transported injury signals The same regulation also applies to importins, also known as karyopherins, an intermediary class of proteins that permit NLS proteins to become attached to the retrogradely transporting dynein motors and allow them to transfer injury signal(s) to the nucleus of the injured neurons. There are several ways to verify that these are indeed injury signals. For example, interference with this importin/dynein-mediated transport of NLS proteins strongly reduces the neurite outgrowth from cultured adult neurons from previously unperturbed sensory ganglia. It also blocks the much more vigorous outgrowth from cultured neurons that were conditioned by peripheral injury.

The Cell Body Response – Morphology

In the injured neurons, the axonal disconnection is followed by a complex of structural, molecular, as well as electrophysiological changes such as the long-term hyperexcitability or LTH. The structural changes are exemplified by the appearance of growth cones at the proximal tip of the lesioned axons, the swelling of the neuronal cell body associated with a strong increase in cellular metabolism, and protein synthesis.

Chromatolysis

The injured and swollen neuronal cell bodies show augmentation and regional dispersion of areas of packed rough endoplasmic reticulum (rER), the so-called Nissl bodies (NBs) (aka Tigroid substance, Nissl shoals), in neuronal cytoplasm, shown in Figure 1. Neuronal nuclei leave their typical cell body midpoint and move to a more eccentric position. On the electron

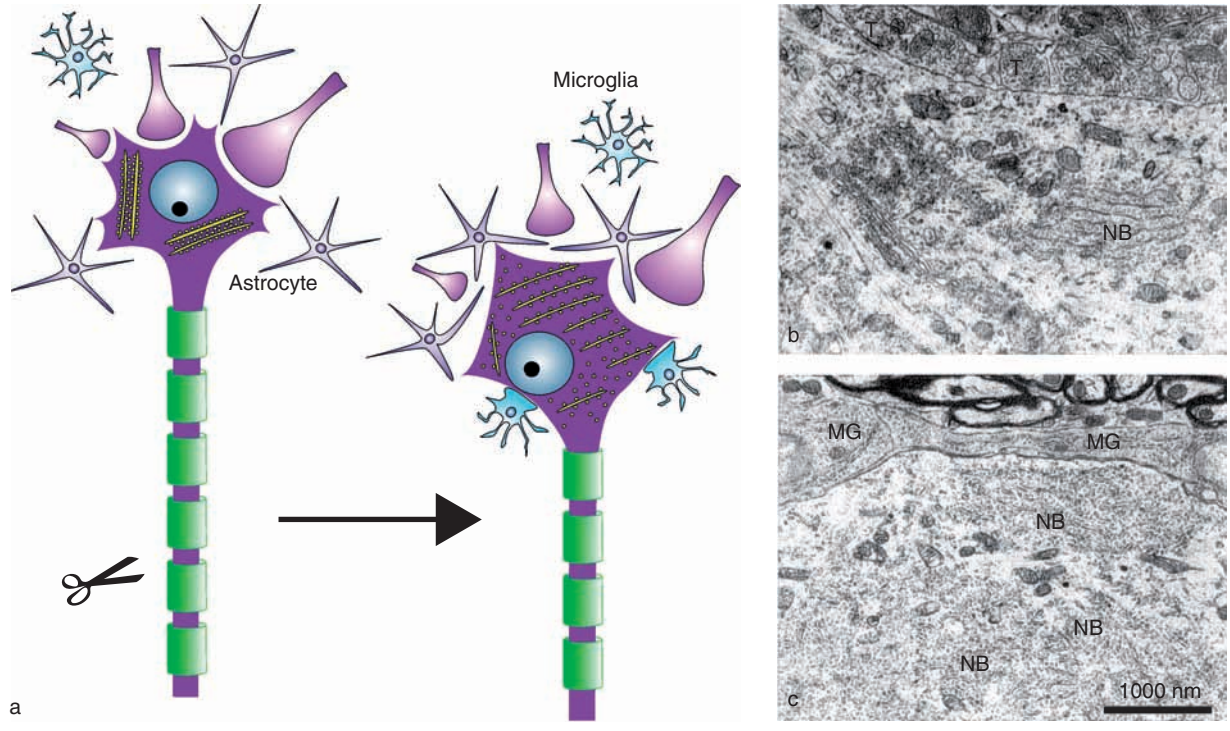

Figure 1 (a–c) Cellular and ultrastructural changes following axonal injury. (a) Schematic summary, (b, c) electron microscopy of a normal (b) and an axotomized rat facial motor neuron, 10 days after lesion (c) at higher magnification. (a) Axotomy causes a detachment of neurite terminals and filling in of empty cell surfaces with CNS glia. Inside the neuron, there is a rearrangement of rough endoplasmitic reticulum cisternae or Nissl bodies (NBs). (b, c) These NB cisternae are normally arranged in parallel (b), but take a more random and curvilinear appearance after axotomy (c). Note the enormous increase in ribosomes and enlargement of NB area. The neuronal cell surface is covered by microglial cells (MG), the neurite terminals (T) are detached. (b, c) Reproduced from Kreutzberg GW and Raivich G Neurobiology of degeneration and regeneration. In: May M and Schaitkin BM (eds.) *The Facial Nerve*, 2nd edn. New York: Thieme; 2000, pp. 67–80. Reprinted by permission of Thieme.

microscopic level, such neurons are seen to have greatly increased numbers of cellular organelles (Figures 1(b) and 1(c)). Ribosomes, either in free form or bound to the endoplasmic reticulum, are especially abundant. The organization of the rER undergoes characteristic changes. The cisternae are no longer arranged in parallel. Due to the augmentation and spreading of the rER, single NBs can no longer be seen; the cells also stain more intensively with Nissl dyes under light microscopy. The contour of the cell body's surface, which is normally smooth, becomes convex and wavy, reflecting the cellular hypertrophy and instability of the plasmalemma.

These overall changes, some of which are easily visible at the light microscopical level using basophilic dyes (eccentric nuclei, increased staining, dissolution of NBs, eccentric nuclei), were termed chromatolysis or chromatolytic reaction by Franz Nissl at the end of the nineteenth century; the term that is still used as a synonym of cellular and molecular changes in the lesioned neurons or retrograde reaction following axonal injury.

Synaptic Stripping

Structurally, these early changes are followed by retraction of neurite endings on the cell bodies of axotomized neurons. This phenomenon, called synaptic stripping, leads to a partial de-afferentation of the injured neurons and may persist for a very long time after initial injury and interfere with long-term recovery (Figure 1(c)). The emptied spaces are gradually taken over by the neighboring reactive astrocytes, which can wrap around injured neurons with multilammellar stacks of thin astrocyte processes.

Central Sprouting

In addition to the generation of new axons at the injury site, supernumerary axonal sprouts can also develop more proximally, at the nodes of Ranvier, a response first noted by Santiago Ramón y Cajal, and occasionally even at the level of the injured neuronal cell body. These latter structures have been termed dendraxons because they exhibit axonal properties and markers but frequently, though not always, grow out of the reorganized dendritic structures.

Peripheral axotomy can also induce the sprouting of central sensory processes of the affected dorsal root ganglia (DRG) neurons in the spinal cord as well as the appearance of perineuronal neurite baskets in the DRGs themselves. Together with the long-term hyperexcitability of the injured sensory neurons, both of these latter processes have been implicated to contribute to the posttraumatic neuropathic pain.

The Cell Body Response – Molecular Changes

The prompt arrival of signals for cellular injury and stress is followed by the induction of factors involved in the regulation of transcription and translation; later on, also by adhesion molecules, growth-associated proteins, and structural components required for axonal elongation. Moreover, the introduction of the different differential cDNA display and gene chip array technologies has led to a dramatic increase in the number of identified genes regulated in injured and regenerating neurons.

Translation and Signal Transduction

One of the earliest biochemical changes following axonal injury is the transient upregulation of polyamine-producing enzymes ornithine decarboxylase and transglutaminase. This has been linked to an increase in mRNA metabolism, and protein synthesis resulting in polyamines such as putrescine, spermine, and spermidine shows a strong enhancing effect on neurite outgrowth both *in vitro* and *in vivo*. The induction of polyamine-producing enzymes is followed by the upregulation of the classical, intracellular stress and growth signaling pathways – the phosphatidylinositol-3 kinase (PI3K) and the mitogen-associated protein kinase (MAPK) cascades. Both pathways are involved in stimulating neurite outgrowth. Expression of active protein kinase B, also known as Akt, a downstream target of PI3K, promotes axonal regeneration as well as neuronal survival after neonatal axonal injury. Expression of active ras, the upstream inducer of ERK1 (MAPK1) and ERK2 (MAPK2), promotes adult survival. Interference with ERK1 and ERK2 also retards axonal outgrowth.

Transcription Factors

Activation, phosphorylation, and nuclear localization of transcription factors appear as the primary signals mediating changes in gene expression in the injured neuron. The majority of transcription factors reported in the context of injury – c-Jun, JunD, ATF3, P311, Sox11, c-EBPβ, and STAT3 – go up. However, some are actually going down, for example, Islet-1, Fra-2, and ATF2. Although all of these transcriptional changes are probably involved in fine-tuning the posttraumatic and regenerative response, the data on their importance using deletion studies are now available for c-Jun, c-EBPβ, and STAT3. In the case of c-EBPβ, deletion of this transcription factor interferes with the injury-induced expression of the molecules important for the coordination of the membrane cytoskletal interface, tubulinα 1, and neuromodulin/GAP43. Deletion of neuronal STAT3 reduces the posttraumatic

survival of axotomized neurons, to the extent observed following the deletion of CNTF and LIF, pointing to STAT3 as the transcription factor that mediates these effects.

However, particularly pronounced effects are observed for neuronal deletion of c-Jun, briefly summarized in Figure 2. The absence of neuronal c-Jun (Figure 2(a)) shows the gene deletion strategy, Figure 2(b) the

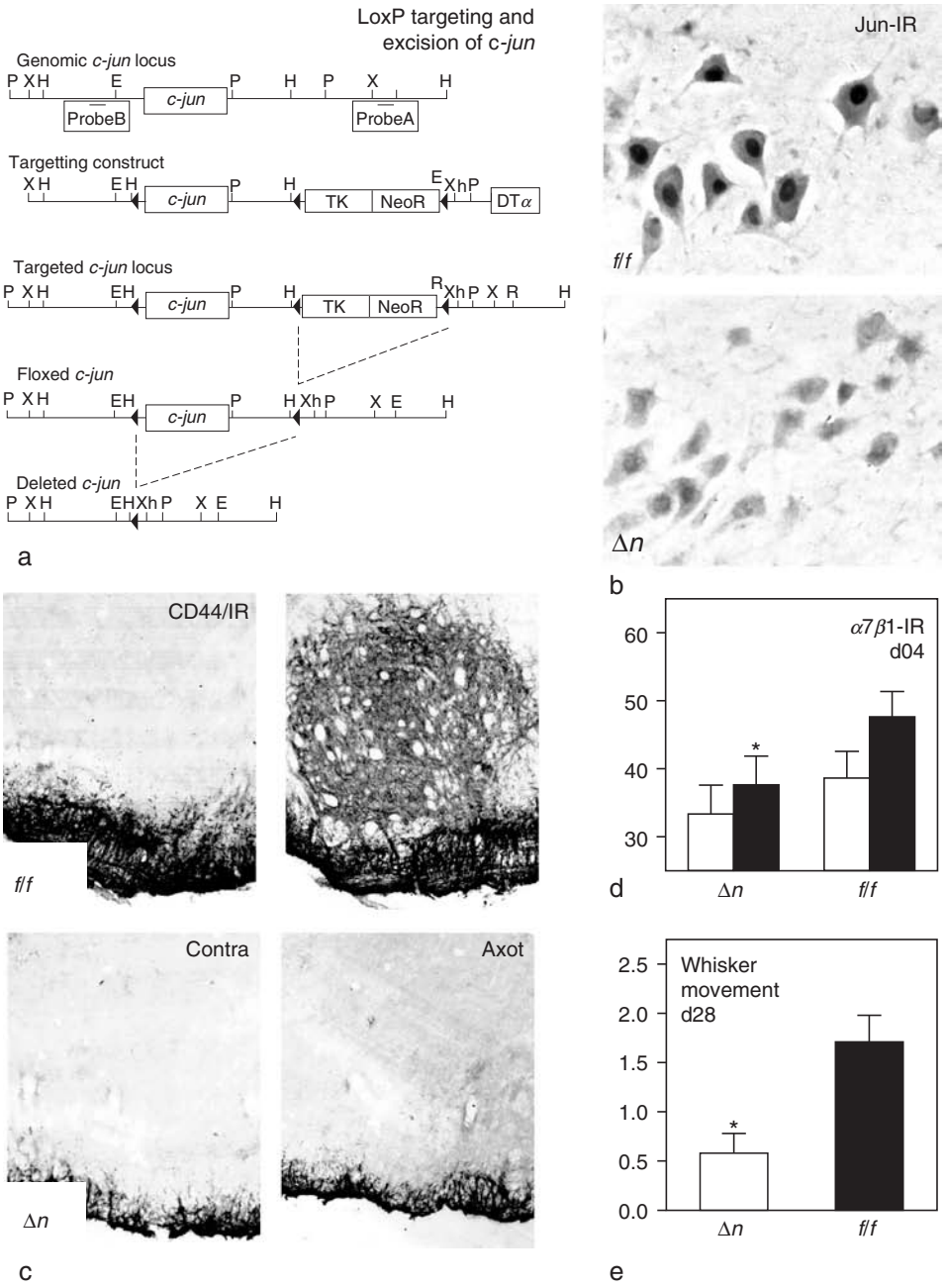

Figure 2 Deletion of the *c-jun* gene in neurons interferes with the neuronal response to injury following transection of the facial nerve. (a) A summary of the three-step gene deletion strategy: (1) by introducing the loxP sites (◄) around the *c-jun* gene and the thymidine kinase (TK) and neomycine resistance *neoR* cassette, (2) removing the floxed *TK* and *neoR* cassette, and (3) deletion of the floxed *c-jun* gene through *cre* recominase under the control of *nestin* promoter active in central neuroepithelial cells. The strategy restricts *c-jun* deletion to neural cells, to avoid embryonic lethality due to global *c-jun* gene deficiency. (b) Absence of c-Jun immunoreactivity (Jun-IR) in neurons with gene deletion (Δ*n*), but not in control animals with two copies of floxed *c-jun* gene (*f/f*), which behave as wild-type controls (not shown). (c–e) Absence of neuronal c-jun blocks the axotomy-induced appearance of neuronal hyaluronic acid receptor CD44 (c), the upregulation of the laminin receptor $\alpha 7\beta 1$ integrin (d) and interferes with normal motor recovery (e). Axotomized side: axot in (c), filled bar in (d); contralateral side: contra in (c), empty bar in (d). Reproduced from Raivich G, Bohatschek M, DaCosta C, et al. (2004) The AP-1 transcription factor c-Jun is required for efficient axonal regeneration. *Neuron* 43: 57–67.

disappearance of nuclear c-Jun immunoreactivity – interferes with the induction of most proteins expressed after injury. There is a complete block in the expression of the adhesion molecule CD44 (Figure 2(c)), which functions as a receptor for hyaluronic acid. There is also a strong decrease in the laminin receptor $\alpha 7\beta 1$ integrin (Figure 2(d)) and neuropeptide galanin that are associated with axonal regeneration, a significant reduction in the speed of reinnervation and functional recovery (Figure 2(e)), and a complete inhibition of central axonal sprouting. In addition, absence of neuronal c-Jun inhibits the recruitment of lymphocytes and activation of brain microglia, showing the importance of neuronal factors for the chemoattraction of neighboring nonneuronal cells. As *in vitro*, deletion of c-Jun interferes with posttraumatic neuronal cell death. However, it also causes severe neuronal atrophy, showing that saving neurons from cell death does not automatically translate into the maintenance of a well-nourished and functional state.

Metabolic Changes

The normal neuronal response to injury is associated with a strong increase in energy metabolism, in the uptake of glucose, and in the enzyme activity of the hexose monophophate shunt (glucose-6-phosphate dehydrogenase and 6-phosphogluconate dehydrogenase). This may be related to increased production of pentoses needed for the increased synthesis of ribonucleotides. Axotomized neurons show a massive induction in the receptors for transferrin, a carrier protein for the trivalent iron ion (Fe^{3+}) and an associated uptake of iron ions into the affected facial motor nucleus. There is also an increase in the neuronal uptake of fatty acids, which is probably involved in the provision of additional constituents to the newly formed axonal membranes.

Neurotransmission and Neurochemical Differentiation

As a general rule, proteins associated with neurotransmission are decreased after injury. This includes the acetylcholine-synthesizing enzyme cholinacetyltransferase and the degrading enzyme acetylcholinesterase in the cholinergic sympathetic and cranial and spinal motor neurons, enzymes involved in catecholamine synthesis (noradrenaline, dopamine), in adrenergic sympathetic and dopaminergic nigrostriatal neurons, or substance P and CGRP in the peripheral sensory neurons responsible for the pain modality. This downregulation is frequently accompanied by reduction in the level and activation of neurotrophin receptors that may be responsible for the neurochemical shift toward a more immature and embryonic phenotype. In sensory, sympathetic, and motor neurons, this shift is also accompanied by an upregulation of other neuropeptides, such as galanin and VIP, which are more typical of their embryonic phenotype.

Typically, the synthesis of neurofilament proteins is also reduced. However, there is strong upregulation of peripherin and the tubulin subunits T-$\alpha 1$ and T-$\beta 3$. Tubulins are major components of the microtubules which are assembled from α- and β-subunits. These microtubules are prominent axonal organelles that function as a carrier for kinesin, a molecular motor that drives axonal transport. Changes in cytoskeletal proteins thus appear important in promoting axonal regeneration. However, they may also lead to the overt morphological alterations at the level of the injured neuronal cell body.

Cell Adhesion Molecules and Cell Surface–Cytoskeletal Interactions

Increased expression or *de novo* appearance of adhesion molecules on the cell surface of the injured neuron appears to be one of the primary motors in changing the growth status of the lesioned neuron and permitting it to extend along, as well as sprout additional fibers in the injured peripheral nerve. Axotomized neurons upregulate a variety of integrin subunits that serve as receptors for extracellular matrix and cell surface molecules. They increase in receptors for the hyaluronic acid and galactoside residues such as CD44 and galectin-1 and express cell multimodal surface molecules involved in homophilic binding such as ninjurin and gicerin/CD146. The latter molecule is also a heterophilic receptor for the neurite outgrowth factor, a member of the laminin family. Another cell surface molecule strongly upregulated after axotomy is FLTR3, which contains numerous leucin-rich repeats and a fibronectin III-type extracellular domain and can form complexes with receptors for the FGF and promote FGF signaling.

Although the common assumption is that most of these molecules are involved in regeneration proper, some may also contribute to aberrant sprouting. For example, deletion of $\alpha 7$ integrin subunit is associated with a strong upregulation of its corresponding $\beta 1$counterpart, but also with a significant increase in central axonal sprouting. This may imply compensatory mechanisms involving other closely related α subunits. Central axonal sprouting on injury-induced neuronal cell surface molecules such as L1 and the close homolog of L1 (CHL1) relies on $\beta 1$-integrin function and may be involved in the effects observed in $\alpha 7$-deficient mice.

Successful axonal regeneration is accompanied by the appearance of numerous, functionally diverse families

of molecules that regulate the surface–cytoskeletal interaction. The best known so far is the GMC family that includes cytoplasmic proteins GAP43, MARCKS, and CAP23. These molecules codistribute with the phosphoinositol-4,5-diphosphate, PI(4,5)P2, at the semicrystalline, raft regions of the cell membrane. These GMC molecules are functionally exchangeable, bind exclusively to acidic phospholipids such as PI-(4,5)-P2, calcium/calmodulin, protein kinase C, and actin filaments, modify the raft recruitment of signaling molecules such as src, regulate actin cytoskeleton polymerization, organization, and disassembly, and have a formative role in the appearance of filopodia and microspikes, in addition to the process of neurite outgrowth.

Posttraumatic Neuronal Death and Cell Death Signals

Several cell surface molecules have been shown to mediate neuronal cell death following neonatal and adult axonal injury, including fas, tumor necrosis factor receptors (TNFRs) 1 and 2, and FAS. Some of these receptors carry a cytoplasmic death domain that exerts a pro-apoptotic signal through FADD; expression of dominant negative form of FADD has been shown to block the axotomy-induced cell death signal due to fas and TNFR. Recent evidence also points to glutamate excitotoxicity.

Associated downstream cytoplasmic cell death signals play a key role in promoting neuronal cell death in neonatal and immature animals that are sensitive to axonal injury, for example, noxA. Some heat shock proteins, for example, hsp27, are neuroprotective. Deletion of bax or inhibition of caspase 3 or the whole family of caspases has also been shown to prevent cell death, but with an enhanced and more persistent effect for broad caspase inhibitors than caspase 3 alone. In both cases, bax and caspases appear to act downstream of the phosphorylation of the transcription factor c-Jun and the decrease in neuronal metabolism, suggesting a sequence of events beginning with phospho-Jun, leading to atrophy, activation of bax, and ending with the initiation of the caspase cascade, resulting in neuronal cell death.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Axonal Injury in Demyelinating Disease and CNS Injury; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Neuronal Plasticity after Cortical Damage; Nogo-A: Its Role in Axon Regeneration; Peripheral Nerve Regeneration: An Overview; Sensory Re-education.

Further Reading

Benn SC and Woolf CJ (2004) Adult neuron survival strategies – Slamming on the brakes. *Nature Reviews Neuroscience* 5: 686–700.

Herdegen T and Leah JD (1998) Inducible and constitutive transcription factors in the mammalian nervous system: Control of gene expression by Jun, Fos and Krox, and CREB/ATF proteins. *Brain Research Reviews* 28: 370–490.

Kiryu-Seo S, Hirayama T, Kato R, and Kiyama H (2005) Noxa is a critical mediator of p53-dependent motor neuron death after nerve injury in adult mouse. *Journal of Neuroscience* 25: 1442–1447.

Kreutzberg GW and Raivich G (2000) Neurobiology of degeneration and regeneration. In: May M and Schaitkin BM (eds.) *The Facial Nerve*, 2nd edn., pp. 67–80. New York: Thieme.

Lieberman AR (1971) The axon reaction: A review of the principal features of perikaryal responses to axon injury. *International Review of Neurobiology* 14: 49–124.

Makwana M and Raivich G (2005) Molecular mechanisms in successful peripheral regeneration. *FEBS Journal* 272: 2628–2638.

Moran LB and Graeber MB (2004) The facial nerve axotomy model. *Brain Research Reviews* 44: 154–178.

Nissl F (1890/1892) Über die Veränderungen der Ganglienzellen am Facialiskern des Kaninchens nach Ausreissung der Nerven. *Versammlungen des Südwestdeutschen Psychiatervereins in Karlsruhe* 1890: 22 *Allgemeine Zeitschrift für Psychiatrie (Berlin)* 48: 197–198.

Perlson E, Hanz S, Medzihradszky KF, Burlingame AL, and Fainzilber M (2004) From snails to sciatic nerve: Retrograde injury signaling from axon to soma in lesioned neurons. *Journal of Neurobiology* 58: 287–294.

Raivich G, Bohatschek M, DaCosta C, *et al.* (2004) The AP-1 transcription factor c-Jun is required for efficient axonal regeneration. *Neuron* 43: 57–67.

Ramón y Cajal S (1928) *Degeneration and Regeneration of the Nervous System*. Oxford: Oxford University Press.

Sun W and Oppenheim RW (2003) Response of motoneurons to neonatal sciatic nerve axotomy in Bax-knockout mice. *Molecular and Cellular Neuroscience* 24: 875–886.

Sung YJ and Ambron RT (2004) Pathways that elicit long-term changes in gene expression in nociceptive neurons following nerve injury: Contributions to neuropathic pain. *Neurological Research* 26: 195–203.

Wiese S, Beck M, Karch C, and Sendtner M (2004) Signalling mechanisms for survival of lesioned motoneurons. *Acta Neurochirurgica Supplement* 89: 21–35.

Woolf CJ (2004) Dissecting out mechanisms responsible for peripheral neuropathic pain: Implications for diagnosis and therapy. *Life Sciences* 74: 2605–2610.

Zigmond RE, Hyatt-Sachs H, Mohney RP, *et al.* (1996) Changes in neuropeptide phenotype after axotomy of adult peripheral neurons and the role of leukemia inhibitory factor. *Perspectives on Developmental Neurobiology* 4: 75–90.

Axonal mRNA Transport and Functions

F P G van Horck and C E Holt, University of Cambridge, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons are highly polarized cells consisting of a cell body, multiple dendrites, and a single axon. During development, axons and their highly dynamic structures at the tip, the growth cones, navigate through the complex environment of the embryo to reach their synaptic targets, which are often long distances from their cell bodies. Along the pathway, growth cones encounter and respond to molecular guidance cues that are present in gradients or localized to discrete choice points. Once they arrive at their destination, axons must select the appropriate neurons with which to form synapses. The specialized functions of the axon and growth cone require a subset of proteins different from those needed in the cell body or in the dendritic compartment. Neuronal function therefore relies on an asymmetric distribution of protein and protein complexes, both during development and throughout the life of mature neurons.

Asymmetric protein distribution can be achieved by two mechanisms: axonal transport of proteins and local protein synthesis. It was long thought that the cell body was the exclusive site of protein synthesis, with newly synthesized proteins being transported into the neuronal processes by selective axonal and dendritic transport mechanisms. This view was altered by the discovery of local protein synthesis in dendrites, beginning with the detection of polyribosomes in postsynaptic structures and followed by the identification of dendritically targeted messenger RNAs (mRNAs). Local translation in dendrites is now known to play an important role in synaptic plasticity and long-term memory storage. The concept of local protein synthesis in axons has been more controversial and has been accepted only since the early 2000s. Although electron microscopy studies in the 1970s of growing axons *in vivo* revealed the presence of polyribosomes in axons, the failure to unambiguously detect components of the protein translation machinery in biochemical axonal preparations hampered further progress. Moreover, *in situ* hybridization studies revealed the presence of mRNAs in young hippocampal neurons but not in mature axons. This transient appearance of mRNAs in developing axons was thought to reflect the fact that transport mechanisms that mediate RNA sorting in mature neurons had not yet developed. However, the process of axonal translation gained more support when individual mRNAs, such as β-actin, were detected in both invertebrate and vertebrate developing axons. Evidence that these mRNAs are translated in axons was initially provided from studies in invertebrates, in which mRNAs injected into isolated axons were shown to be translated into functional proteins. Studies in invertebrates have been viewed with caution, however, since invertebrate axons have both receptive and transmissive capacities and therefore combine the functions of both dendrites and axons. Subsequent work has shown that vertebrate axons can also actively synthesize proteins. Indeed, vertebrate axons have been suggested to represent about 5% of total neuronal protein synthesis. There is, however, a relative paucity in our understanding of the molecular mechanism underlying axonal mRNA transport and translation as the field of local translation in neurons has been dominated by studies in dendrites. This article focuses on the mRNAs that have been identified in axons, the mechanism of axonal mRNA transport, and the way translation of axonal mRNAs controls critical functions in neuronal behavior.

Which mRNAs Are Transported into the Axon?

To date, the main group of mRNAs identified in axons and growth cones are those encoding cytoskeletal proteins, such as β-actin, or regulators of cytoskeletal dynamics, such as actin depolymerizing factor (ADF)/cofilin, profilin, β-thymosin, microtubule-associated protein (MAP)1b, and tau. However, mRNAs for proteins involved in cell signaling, such as RhoA and transmembrane receptors such as EphB2 have also been found, albeit to a lesser extent. Most axonal mRNAs have been identified by candidate approaches, such as by using fluorescent *in situ* hybridization on axonal cultures *in vitro*. Other approaches involve the generation of pure axonal preparations and subsequent biochemical analysis of mRNAs associated with ribosomal proteins or molecular analysis of the mRNA content. The latter approach has also led to the identification of mRNAs in regenerating adult axons. In this approach dorsal root ganglia (DRG) neurons are damaged in the animal and then allowed a period of recovery in order to allow induction of the regeneration program (see the section titled 'Axonal mRNA transport in adult axons') prior to neuronal isolation. Purified axons of these so-called injury-conditioned neurons contain, besides cytoskeletal mRNAs, those encoding heat shock proteins, resident endoplasmic reticulum (ER) proteins, proteins associated with

neurodegenerative diseases, antioxidant proteins, and metabolic proteins. This implies that axons can contain a wide range of mRNAs. The repertoire of axonal mRNAs is likely to vary depending on neuronal cell type, function, and developmental stage.

Mechanism of Axonal mRNA Transport

Targeting Elements

As has been found in nonneuronal cells, mRNA localization in axons depends on *cis*-acting sequences, called zipcodes. Zipcodes are essential for the transport of the mRNAs, and removal or mutations to them impair the targeting of the transcript. Zipcodes are usually present in the 3′ untranslated region (UTR) of the mRNA and are occasionally found in the 5′UTR or even in the coding region. Zipcodes can be short segments with defined nucleotide sequences but can also be secondary or tertiary structures such as stem loops. A well-studied example in axons is β-actin mRNA. Localization of the β-actin mRNA into growth cones is dependent on the 3′UTR. Mutation analysis has defined a 54-nucleotide sequence within the 3′UTR that is necessary and sufficient for targeting β-actin mRNA into the growth cone of chick forebrain axons. β-actin mRNA localization can be inhibited by antisense oligonucleotides directed against the 54-nucleotide sequence as well as by mutations that disrupt the secondary structure. Similarly, the mRNA for the MAP tau is targeted to the axon by an AU-rich sequence in the 3′UTR. Fusion of the tau 3′UTR axonal localization signal to a gene normally not present in axons results in axonal localization of its mRNA. In most cases, the mRNA localization signals are more complex and lack a simple consensus. Together with the fact that axonal localization signals are not highly conserved among species, it is not yet possible to predict from the primary sequence whether an mRNA is targeted into the axon.

RNA-Binding Proteins

The targeting elements in mRNAs are recognized by RNA-binding proteins (RNBPs), which are essential for mRNA transport. To date only a limited number of zipcode-binding RNBPs have been identified in axons. These include zipcode-binding proteins (ZBPs)-1 and 2, HuD, cytoplasmic polyadenylation element binding protein (CPEB), survival of motor neuron protein (SMN), and fragile X mental retardation protein (FMRP); all of which are proteins with well-established roles in mRNA localization in oocytes, embryos, and fibroblasts. RNA binding is modulated by RNA-binding domains such as RNA recognition motif, heterogeneous nuclear ribonucleoprotein K homology domains, and Arg-Gly-Gly boxes. RNBPs often harbor multiple RNA-binding modules and have the potential to interact with multiple mRNAs, as well as with other proteins involved in transport or translation. Together they form large protein complexes that are actively transported into the axon (Figure 1).

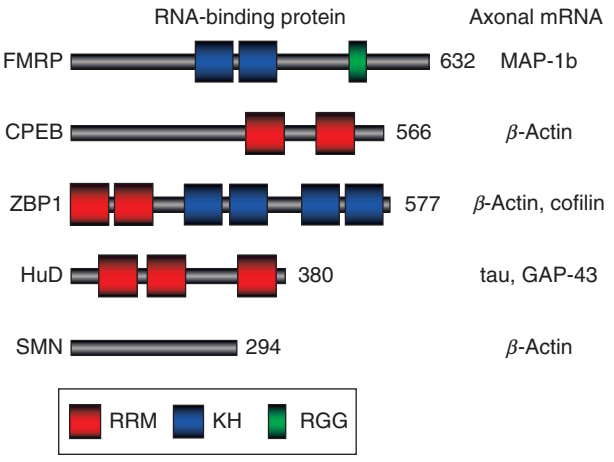

Figure 1 Examples of RNA-binding proteins (RNBPs) and messenger RNA (mRNA) complexes identified in axons and growth cones. mRNAs are depicted that have been shown to interact either directly or indirectly with the indicated RNBP in axons. RNBPs interact with the mRNAs via conserved RNA-binding modules, such as RNA recognition motif (RRM), heterogeneous nuclear ribonucleoprotein K homology (KH), and Arg-Gly-Gly (RGG) boxes. Survival of motor neuron protein (SMN) lacks any of these recognizable RNA-binding domains. Numbers refer to amino acids. FMRP, fragile X mental retardation protein; CPEB, cytoplasmic polyadenylation element binding protein; ZBP1, zipcode-binding protein 1; HuD, a zipcode-binding RNBP; MAP, microtubule-associated protein; GAP, growth-associated protein.

Active Transport

There are several mechanisms known to localize mRNA to specific subcellular regions, including mRNA diffusion and anchoring, local degradation, and active transport along cytoskeletal filaments. Of these, active transport of mRNAs has been implicated as the major localization mechanism in neurons. A consensus view of this process has come primarily from studies on dendrites, as well as from *Xenopus* and *Drosophila* early development systems. Findings on mRNA localization in axons are limited but in agreement with the current model of active transport, which encapsulates the following multistep pathway: RNA granule formation, cytoskeleton transport and release, anchoring, and translation.

RNA granule formation As soon as mRNAs are generated, they are subject to multiple dynamic interactions with RNBPs involved in mRNA processing

and nuclear export. Once in the cytoplasm, some proteins dissociate from the mRNAs whereas others remain attached. For example, both ZBP1 and 2 associate with the β-actin mRNA zipcode in the nucleus. However, while ZBP1 remains bound to β-actin mRNA in the cytoplasm, ZBP2 is a shuttling RNBP that is thought to be released for return to the nucleus. Next, proteins that are required for cytoplasmic transport, translation, degradation, and anchoring associate with the mRNA, resulting in the formation of large particles, called ribonucleoprotein (RNP) complexes or RNA granules. RNA granules can be identified by phase light microscopy, in which they appear as large, clustered aggregates. Biochemical characterization of these granules from neuronal extracts has revealed the presence of many localized mRNAs, RNBPs, ribosomal proteins, and proteins involved in the regulation of translation, as well as motor proteins that connect RNA granules to the cytoskeletal transport network. The assembly of transport particles, containing multiple mRNAs and both nuclear and cytoplasmic RNBPs, is therefore an important step in mRNA localization.

Movement along the cytoskeleton Active transport is characterized by directed movement of RNA granules at a speed that is too fast to be explained by cytoplasmic flow. Live cell imaging of mRNA or RNBP movement shows that active transport can reach velocities of 0.1–2 $\mu m\ s^{-1}$. How is such rapid transport achieved? RNA granules contain motor proteins that connect the RNP complexes with the cytoskeleton. In an analogous fashion to vesicle transport, motor proteins mediate RNP transport along cytoskeletal filaments in an adenosine triphosphate-dependent manner. In axons, the shaft consists of actin filaments and bundles of stabilized microtubules, which extend, together with some dynamic microtubules, into the growth cone central domain and are therefore an ideal track for transport into the axon.

Active transport via the cytoskeleton is mediated by specific motor proteins. Kinesin and dynein are microtubule-based motor proteins, whereas myosin is required for transport along actin filaments. In general, anterograde movement along plus end microtubules (i.e., away from the cell body, into the axon) is mediated by kinesin, whereas retrograde transport along minus end microtubules (i.e., toward the cell body, out of the axon) might involve dynein. However, the kinesin motor family comprises many family members with different expression patterns and cargo specificities, and some family members have also been shown to mediate retrograde transport. Kinesin-2 is involved in anterograde transport and is expressed in axons. One of the microtubule-based motor subunits of kinesin-2, KIF3A, interacts with HuD in HuD-tau mRNA-containing granules in axons. Moreover, inhibition of KIF3A expression abolishes tau mRNA transport into the axon, resulting in decreased levels of axonal tau protein. Since tau is important for microtubule stabilization, the decrease in expression levels causes neurite retraction. Other RNBPs have been shown to interact with several different kinesin family members, albeit not specifically in axons or growth cones. In *Drosophila* neuronal cells, FMRP granule transport depends on both kinesin-1 and cytoplasmic dynein. Knockdown of either kinesin-1 or cytoplasmic dynein is sufficient to impede both anterograde- and retrograde-directed FMRP transport, indicating that the two transport mechanisms are coupled.

The identity of the motor proteins suggests that axonal mRNAs are transported along microtubules. Indeed, treatment of neurons with inhibitors of microtubule polymerization, such as nocodazole or colchicine, reduces transport of ZBP1, SMN, and FMRP granules into the axon, showing that these granules are predominantly transported along microtubules. It should be noted that granule movement is not solely unidirectional. Detailed analysis of enhanced green fluorescent protein (eGFP)-tagged RNBP dynamics shows a variety of movements, uni- and bidirectional, as well as oscillatory movements with no net displacement. The cytoskeleton of the peripheral domain of growth cones is comprised mainly of actin filaments, which are either arranged in a network forming the lamellipodia or in f-actin bundles forming the filopodia. RNBPs such as ZBP1, SMN, and FMRP are detected in developing growth cones, suggesting that within the growth cone, transport can also take place along actin filaments. Indeed, the actin depolymerizing drug cytochalasin D blocks transport of eGFP–SMN, but it mainly inhibits short-distance-directed movements and oscillatory movements. Thus axonal mRNA transport seems to involve long-distance-directed movement of RNA granules along microtubules in the axon shaft as well as short-distance-directed movements along actin filaments in growth cones.

mRNA release and anchoring To restrict protein synthesis to a specific subcellular domain, mRNAs must not be translated until they reach their destination. This implies that a tight coupling between mRNA transport and repression of protein translation exists, and RNBPs are thought to play a key role in this regulation. ZBP1, for example, is not only essential for β-actin mRNA localization; it also represses the translation of β-actin mRNA during transport in neuroblastoma cells. Once an mRNA

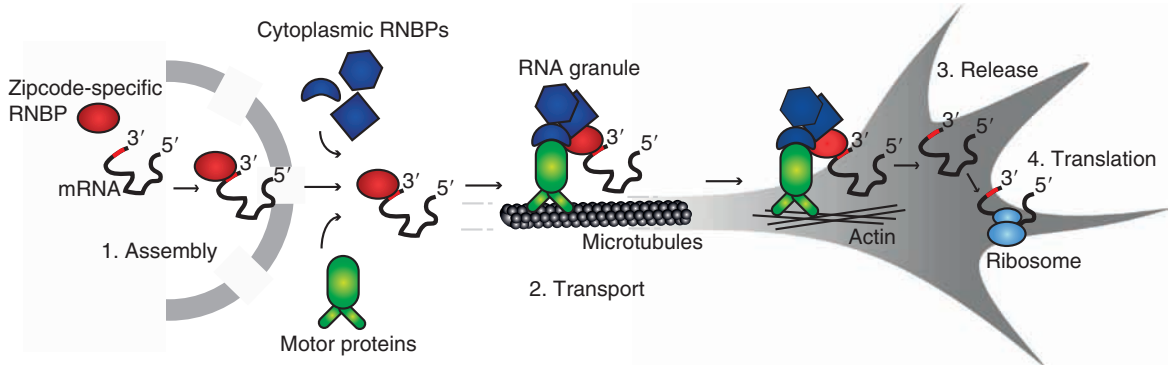

Figure 2 Model for axonal messenger RNA (mRNA) transport. Axonal mRNA localization is a multistep process, involving ribonucleoprotein (RNP) assembly in nucleus and cytoplasm, followed by active transport along microtubules and actin filaments. Once arrived at the growth cone, transported mRNAs are released and anchored until needed for translation. RNBP, RNA-binding protein.

granule arrives at its site of action, the mRNA has to be released and anchored in the correct place until needed for translation. In the case of ZBP1, this has been shown to involve a phosphorylation-dependent mechanism. The nonreceptor kinase Src phosphorylates ZBP1, which reduces the ability of ZBP1 to bind to β-actin mRNA and thereby releases the translational repression. It is interesting that this regulatory step seems to take place in the growth cones, suggesting that external cues might induce this process leading to local β-actin translation. The molecular mechanism of mRNA anchoring in neurons is not well understood. One of the components of the translation machinery, elongation factor 1a, has been implicated in anchoring β-actin mRNA to the actin filaments, where polysomes are also known to reside. Furthermore, FMRP is associated with translating polyribosomes in neuronal cells, where it is thought to regulate translation (Figure 2).

mRNA translation mRNA translation in axons is thought to involve the cap-dependent translation machinery. Cap-dependent mRNAs are abundant in growth cones, and molecular cues shown to stimulate translation in growth cones (see below) induce the rapid phosphorylation of the translation initiation factor eIF4E binding protein 1 (4EBP) and the translation initiation factor eIF4E. Hypophosphorylated 4EBP blocks translation initiation by sequestering the rate-limiting eIF4E. Hyperphosphorylation causes 4EBP to release eIF4E and allows translation initiation to proceed. In turn, eIF4E is phosphorylated by the kinase target of rapamycin (TOR, which is inhibited by rapamycin). Although rapamycin blocks translation in retinal growth cones, translation is not completely inhibited, indicating that cap-independent translation may also be involved. Which other components of the translation initiation machinery are required for axonal mRNA translation and how the process is spatiotemporally regulated are currently unknown.

Functions of Axonal mRNA Transport

Axonal mRNA transport serves to spatially restrict protein synthesis to the axon and growth cone. How is this ability to locally transport mRNAs important functionally for the axon? Axons are continuously exposed to molecular cues, which regulate their outgrowth and pathfinding. *In vitro* assays on cultured neurons have revealed that molecular cues are able to regulate mRNA transport. For example, in chick forebrain neurons, the neurotrophic factor Neurotrophin-3 (NT-3) increases β-actin mRNA localization in the growth cone within a short timescale (10 min). This increase is blocked when, prior to NT-3 stimulation, the axons are incubated with antisense oligonucleotides that interfere with the binding of ZBP1 to β-actin mRNA, indicating that NT-3 stimulates axonal mRNA transport rather than local mRNA stability. Indeed, detailed analysis of ZBP1 granule movement shows that NT-3 increases the rapid ZBP1 transport into the growth cone. Furthermore, cue-induced stimulation of mRNA transport seems to be selective since, for example, in axons from regenerating adult DRG neurons, brain-derived neurotrophic factor and nerve growth factor selectively induce the transport of a subset of mRNAs, including β-actin, vimentin, and peripherin. The precise functional significance of this finding is unknown but is presumed to reflect proteins that are required for axonal responses in both developing and adult axons.

Axonal mRNA Transport in Developing Axons

Developing axons are characterized by their ability to grow, navigate, and innervate their target region. Each of these processes requires cytoskeletal

rearrangements. Significantly, mRNAs encoding cytoskeletal proteins and cytoskeletal-associated proteins have been isolated from axons of developing neurons from a wide range of organisms. Given the importance of cytoskeletal dynamics in developing neurons, this suggests multiple roles for local translation in regulating axon and growth cone behavior, such as axon growth and pathfinding.

Axon growth Addition of protein synthesis inhibitors to cultured *Xenopus* retinal or spinal axons does not affect their rate of extension over the short term (1 h), suggesting that local translation of axonal mRNAs is not critical for short-term axon outgrowth. In contrast, disruption of the interaction between ZBP1 and β-actin mRNA by antisense oligonucleotides has been shown to impair the persistent forward movement of growth cones. This apparent discrepancy could reflect a difference in the requirement for translation and mRNA transport. Several *in vitro* studies indicate that mRNA transport is important for long-term axonal growth. Expression of a mutant ZBP1, which cannot release the β-actin mRNA, reduces neurite outgrowth in hippocampal neurons. Similarly, manipulation of HuD levels in rat cortical neurons results in significant changes in neurite outgrowth. More-direct evidence that axonal transport is important for growth comes from studies on SMN. Transport of SMN from the nucleus into the cytoplasm requires exon 7. Neurons transfected with an exon 7 deletion mutant (SMNΔ7) have shorter processes, both of the minor neurites and the incipient axon. It is important to note that this axon outgrowth defect can be rescued by redirecting SMNΔ7 into neurites by an axonal targeting sequence from growth-associated protein 43, indicating that SMN-mediated mRNA transport is critical for axon outgrowth. Further complexity is apparent from the ability of some translated mRNAs to inhibit axon outgrowth. For example, mRNA of β-thymosin, a negative regulator of actin polymerization, is accumulated in the distal neurites of *Lymnaea* neurons, and reducing the β-thymosin mRNA levels in isolated neurites increases the rate of neurite outgrowth. Thus, locally synthesized proteins can both positively and negatively regulate axonal growth.

Axon guidance As axons grow, their growth cones encounter a wide variety of signals that influence their navigation. Growth cones react to local attractive and repulsive cues in their environment by extending and retracting filopodia (fingerlike protrusions) and lamellipodia (flattened, veillike extensions), which determines the net direction of migration. Turning is caused by extension of filopodia on one side and withdrawal or collapse of filopodia on the other. Using two different chemotropic assays, the behavior of *Xenopus* retinal ganglion cell (RGC) axons *in vitro* has provided an insight into the role of local mRNA translation in growth cone guidance. In turning assays, growth cones grow toward a gradient of an attractant and turn away from a gradient of repellent over the course of 30–60 min. In collapse assays, bath application of a repellent causes complete withdrawal of growth cone filopodia and lamellipodia within 10 min. Embryonic *Xenopus* RGC growth cones are attracted to a source of netrin-1, but this attractive response is blocked by protein synthesis inhibitors. Moreover, protein synthesis inhibitors block the repulsive turning as well as the collapse response of RGCs toward the repellents Slit-2 and Semaphorin 3A (Sema3A). It is important to note that this inhibition of chemotropic responses by protein synthesis inhibitors also takes place in growth cones that have been severed from their cell bodies, showing that the translation occurs locally, in the growing axon. Indeed, netrin-1, Slit-2, and Sema3A are able to induce rapid (within 5–10 min) protein translation in the growth cone as measured by phosphorylation of eIF4E and eIF4EBP and the incorporation of labeled amino acids. It is striking that netrin-1, Slit-2, and Sema3A elicit local translation by activating different intracellular pathways involving mitogen-activated protein kinases (MAPKs). For example, netrin-1 and Slit-2, but not Sema3A, induce translation that depends on p38 MAPK. The netrin-1, Slit, and Sema3A pathways are dependent on mammalian TOR, indicating that cue-induced translation is cap-dependent.

These findings have provided evidence that growth cone turning and collapse depend on local protein synthesis, but which mRNAs are translated in these processes? Since growth cone turning involves remodeling of the microtubule and actin cytoskeleton, translation of mRNAs encoding cytoskeletal regulators are good candidates. Consistent with this hypothesis, Slit-2 has been shown to induce a protein synthesis-dependent increase in ADF/cofilin in *Xenopus* RGC growth cones. Moreover, Sema3A-mediated growth cone collapse in DRG neurons is dependent on the local translation of the small GTPase RhoA. RhoA is a key regulator of the actin cytoskeleton, and its mRNA is targeted to the growth cone by a specific sequence located in its 3′UTR. The local translation of RhoA is both sufficient and necessary for Sema3A-mediated growth cone collapse, suggesting that basal levels of RhoA are not sufficient to mediate the Sema3A responses. Thus, the local translation of cytoskeletal proteins provides the developing axon and growth cone with a mechanism to rapidly

modulate the cytoskeleton in response to extracellular cues (Figure 3).

Local mRNA translation in growth cones also plays a role in growth cone adaptation, a process whereby the growth cone desensitizes and resensitizes to a guidance cue. Adaptation is thought to be an integral part of the steering mechanism, enabling growth cones to adjust their sensitivity in a way that allows them to navigate in gradients of axon guidance cues. In *Xenopus* spinal neurons, attractive turning responses toward netrin-1 or brain-derived neurotrophic factor are attenuated if either of the chemoattractants is present in the bath. After 30 min, the desensitization disappears, and the growth cones regain their normal turning behavior. Protein synthesis inhibitors block this resensitization process, indicating that locally synthesized proteins play a role in regaining responsiveness. Collapse assays in *Xenopus* retinal growth cones show that desensitization and resensitization can even occur at a much shorter timescale (2–5 min) in response to repellents. Resensitization in response to Sema3A and netrin-1 under repulsive conditions is ligand-specific and requires new protein synthesis. Growth cone resensitization to Sema3A and netrin-1 correlates with a protein synthesis-dependent increase of their respective receptors, neuropilin and deleted in colorectal cancer, at the plasma membrane, indicating that resensitization might involve, at least partly, the translation of guidance cue receptors. In this way, growth cones can adjust their sensitivity as a function of exposure to a specific ligand.

Although functional studies on axonal mRNA transport are mostly conducted *in vitro*, local protein synthesis has also been implicated in axon navigation *in vivo*. When growing over long distances,

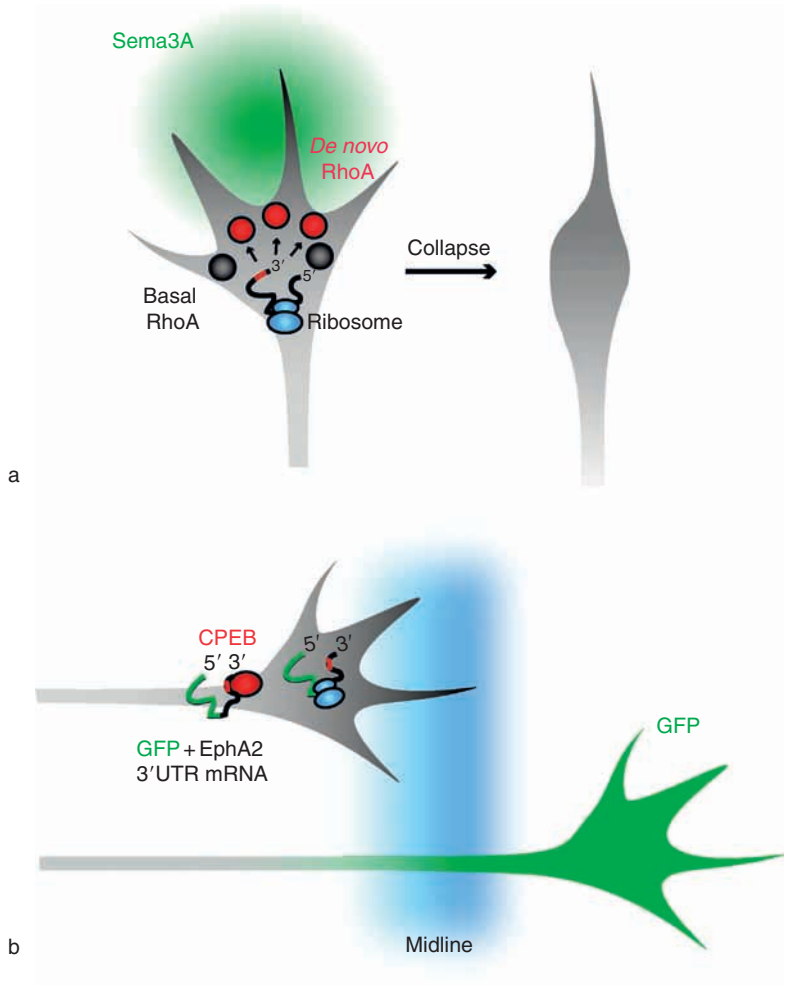

Figure 3 Diagram showing the role of axonal messenger RNA (mRNA) transport and local translation in growth cone behavior in developing axons. (a) Growth cones exposed to Sema3A induce the translation of RhoA, a small guanosine triphosphatase. The newly synthesized RhoA is required and sufficient to mediate Sema3A-induced growth cone collapse. (b) Local translation at the midline intermediate target. Translation of a reporter construct encompassing green fluorescent protein (GFP) and the 3′ untranslated region (3′UTR) of the EphA2 receptor is upregulated as growth cones enter the midline and reaches high levels as they emerge from the midline. Translation of the reporter construct is dependent on the cytoplasmic polyadenylation element sequence in the 3′UTR, suggesting that cytoplasmic polyadenylation element binding protein (CPEB) may be involved in mRNA localization or translational control.

axons use intermediate targets to which they are directed by local attractive and repulsive cues. After reaching the intermediate target, the axons must continue their journey and change responsiveness to the cues that guide them to the next target. One way to change responsiveness at an intermediate target is by changing the repertoire of guidance receptors at the plasma membrane. Studies in chick embryonic commissural neurons have suggested that this can be achieved by local translation of a receptor. Spinal commissural axons are initially attracted toward the midline floor plate, but after crossing the midline, axons lose responsiveness to the midline attractants and gain responsiveness to a new set of guidance cues. When these commissural axons are transfected with a construct encompassing the 3′UTR of the EphA2 receptor and GFP, protein expression is detected in the cell bodies but not in the growth cones. However, the protein is expressed in the growth cones of axons that have crossed the midline, suggesting that translation of the reporter construct is selectively activated after midline crossing. Translation is dependent on a cytoplasmic polyadenylation element binding element in the 3′UTR, indicating that CPEB might target the EphA2 mRNA to the growth cone.

Thus, local protein synthesis may play a crucial role in axon guidance. It has been known for decades that axons that are separated from their cell body retain their ability to grow and even make accurate guidance decisions. The ability of axons to locally translate proteins may explain this observation. By locally synthesizing proteins, the growth cone can autonomously respond to its environment. Different guidance cues along the pathway might stimulate the transport and the translation of a specific subset of mRNA needed for immediate growth cone turning, adaptation, and responses at intermediate targets. Moreover, protein synthesis may play a role presynaptically in the establishment of arbors and synapses, and studies in invertebrates indeed suggest a role for local translation in regulating synaptic connections in response to activation. It should be noted that not all cues elicit protein synthesis in the growth cones, indicating that local translation is not a ubiquitous mechanism in cue-directed growth.

Axonal mRNA Transport in Adult Axons

In contrast to developing axons, neuronal function of mature axons is not dependent on major cytoskeletal rearrangements like those involved in growth cone steering but rather on maintaining the integrity of the cytoskeleton, except when an axon is disconnected from its targets on injury. Neuronal regeneration is dependent on many factors, including severity of the injury and cell type. Peripheral nervous system neurons have regenerative capacity, whereas neurons of the central nervous system do not. Successful regeneration requires the tip of the cut axon to be remodeled and reformed into a growth cone. Experiments in cultured adult sensory and retinal axons have demonstrated that the formation of a growth cone, after axotomy, depends on local axonal protein synthesis (and degradation). It is important to note that the intrinsic regenerative capacity of the different types of neurons correlates with the levels of intra-axonal protein synthetic machinery: sensory neurons with high regenerative ability contain high levels of ribosomal proteins and activated translational initiation factors, whereas these proteins are absent in retinal axons, which have a low regenerative ability. Similar to developing axons, growth cone reformation after injury depends on activation of MAP kinases. A mechanism for the role of translation and MAP kinases has been proposed in the peripheral nervous system. Nerve injury leads to the activation of MAP kinases in the axon as well as the translation of the intermediate filament protein vimentin and the nuclear import protein β-importin. Vimentin then undergoes proteolytic cleavage, and these vimentin fragments bind to the activated MAP kinases to couple them to the importin–dynein complex for transport to the cell body. This is thought to provide a mechanism to retrogradely transport an injury signal to the nucleus to facilitate the regenerative response. Maintenance of regenerative growth, after the initiation of growth cone formation, might be regulated by extracellular stimuli. A proteomic approach has identified multiple axonally synthesized proteins in injury-conditioned DRG axons, but the growth-promoting neurotrophins selectively increase the levels of some cytoskeletal mRNAs into the axon, presumably to retain growth cone motility. Thus, axonal mRNA translation is important for multiple steps during regeneration, including growth cone formation, generation of an injury signal, and maintenance of axonal growth (Figure 4).

Axonal mRNA Transport and Neurological Diseases

Although *in vivo* data on the function of axonal mRNA transport and translation are still lacking, the importance of these processes becomes evident when the function of key regulators such as RNBPs is impaired. For example, mutations in the FMR1 gene on the X chromosome lead to the fragile X syndrome, which is the most common cause of inherited mental retardation. Mutations in the FMR1 gene lead to transcriptional silencing of the gene and a consequent lack of FMRP and FMRP-transported mRNAs. Furthermore, HuD is associated with

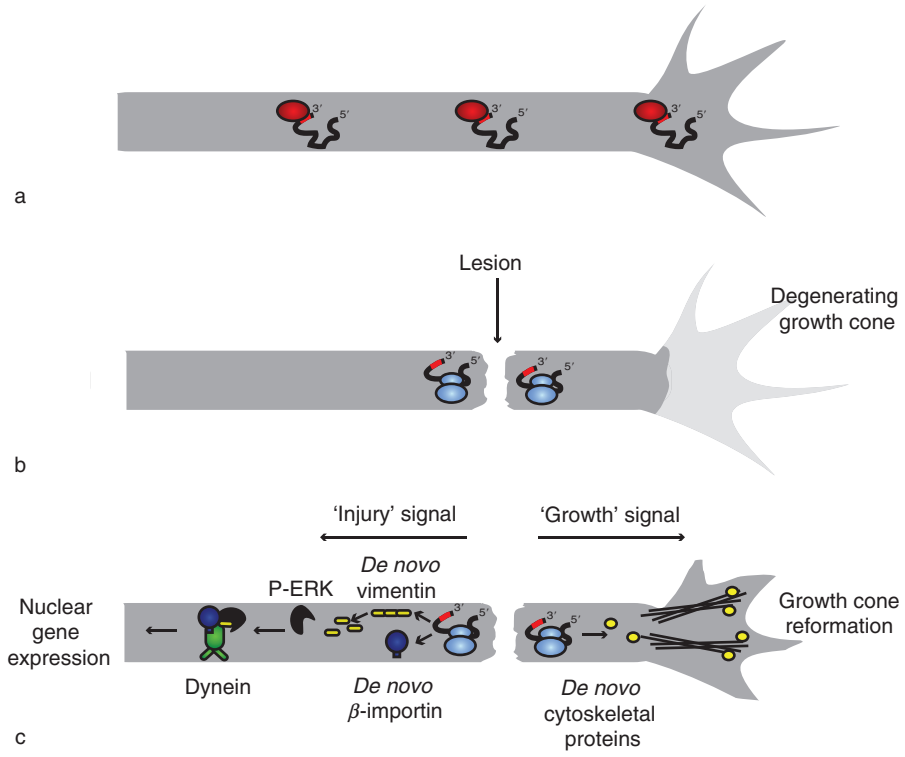

Figure 4 Local translation in injured mature axons. (a) Uninjured axons transport mRNA along the axon, presumably for cytoskeletal maintenance. (b) Once an axon is lesioned, the growth cone undergoes degeneration and loss of connections with its targets. At the site of the lesion, protein translation is induced. (c) Newly synthesized proteins include cytoskeletal proteins such as β-actin for growth cone reformation and of vimentin and β-importin for the generation of an injury signal. Vimentin protein is proteolytically cleaved, which generates fragments that link phospho-extracellular signal-regulated kinase (P-ERK) to β-importin to the dynein motor protein. This complex is retrogradely transported along microtubules into the nucleus to regulate gene expression to facilitate the regeneration process. Adapted from Willis DE and Twiss JL (2006) The evolving roles of axonally synthesized proteins in regeneration. *Current Opinion in Neurobiology* 16: 111–118.

paraneoplastic encephalomyelitis, a neurological disorder caused by an autoimmune reaction against HuD. Moreover, low levels of SMN protein result in spinal muscular atrophy, which is an autosomal recessive disorder and the leading hereditary cause of infant mortality and is characterized by a loss of α motor neurons in the spinal cord. In most cases, the causes of these neurological diseases have not been directly attributed to specific defects in axonal mRNA transport. However, knockdown of SMN in zebra fish has been shown to result in motor axon-specific outgrowth and pathfinding defects, emphasizing the importance of regulated mRNA transport for neural development. Future studies on the identification of RNP complexes and their functions might explain the cellular consequences of RNA misregulation that underlie distinct types of neurological diseases.

Conclusion

The concept of local translation at growth cones has gained increasing acceptance during the past 5 years, and it is now clear that mRNAs are transported into the axon and growth cone and can be translated in response to environmental cues. Axonal mRNA transport serves to spatially restrict protein synthesis to the axon and growth cone, providing axons with the capacity to autonomously regulate their structure and function during axonal growth, navigation, and regeneration. A general model about the molecular mechanism of neuronal mRNA transport and translation has come from studies in dendrites, and although this mechanism seems to be highly conserved, many aspects of axonal mRNA function remain completely undefined. For instance, characterization of the identity of RNA granules and determination of the ways environmental cues regulate mRNA localization within the milieu of the elongating axon are incompletely understood. One recently proposed hypothesis is that RNBPs transport mRNAs as a set that encodes proteins with a common biological function. Some dendritic RNBPs, for example, transport a coherent subset of mRNAs to specifically mediate synaptic plasticity. It will be fascinating to determine whether axonal RNA granules function in an analogous fashion to regulate

specific axon and growth cone function. Elucidating how axonal mRNA transport and translation regulate growth and pathfinding *in vivo* should provide key insights into the ways dysfunction in these processes can lead to neurological diseases.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Transport Disorders; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; Growth Cones; Myosin Transport and Neuronal Function; RNA Binding Protein Methods.

Further Reading

Bassel GJ and Kelic S (2004) Binding proteins for mRNA localization and local translation, and their dysfunction in genetic neurological disease. *Current Opinion in Neurobiology* 14: 574–581.

Brittis PA, Lu Q, and Flanagan JG (2002) Axonal protein synthesis provides a mechanism for localized regulation at an intermediate target. *Cell* 110: 223–235.

Campbell DS and Holt C (2001) Chemotropic responses of retinal growth cones mediated by rapid local protein synthesis and degradation. *Neuron* 32: 1013–1026.

Martin KC (2004) Local protein synthesis during axon guidance and synaptic plasticity. *Current Opinion in Neurobiology* 14: 305–310.

Ming GL, Wong ST, Henley J, et al. (2002) Adaptation in the chemotactic guidance of nerve growth cones. *Nature* 417: 411–418.

Piper M and Holt C (2004) RNA translation in axons. *Annual Reviews in Cell and Developmental Biology* 20: 505–523.

Piper M, Salih S, Weinl C, Holt CE, and Harris WA (2005) Endocytosis-dependent desensitization and protein synthesis-dependent resensitization in retinal growth cone adaptation. *Nature Neuroscience* 8: 179–186.

St. Johnston D (2005) Moving messages: The intracellular localization of mRNAs. *Nature Reviews in Molecular Cell Biology* 6: 363–375.

Willis DE and Twiss JL (2006) The evolving roles of axonally synthesized proteins in regeneration. *Current Opinion in Neurobiology* 16: 111–118.

Wu KY, Hengst U, Cox LJ, et al. (2005) Local translation of RhoA regulates growth cone collapse. *Nature* 436: 1020–1024.

Axonal Pathfinding

D Mortimer and G J Goodhill, The University of Queensland, St. Lucia, QLD, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Building a brain requires the precise formation of connections between vast numbers of neurons, often separated by significant distances. Axon pathfinding is thus a crucial process in the development of a functioning nervous system. A better understanding of the principles and mechanisms underlying axon pathfinding has both clinical and broader practical importance. From a clinical perspective, our ability to treat or prevent some neurological defects will be improved by a better understanding of how axon guidance can fail. Furthermore, the regeneration of damaged nerves requires axons to reconnect to their appropriate targets; hence, understanding axon guidance will be necessary for the development of therapeutic techniques. In a broader context, our understanding of axon pathfinding ties into our understanding of the nervous system as a whole: what principles underlie its formation and function? can we harness those principles in order to improve our own engineering processes, such as in the construction of self-wiring computers?

Mathematical and computational models are very useful tools for understanding the constraints on nervous system development. Ultimately, such constraints are quantitative and set by the physics of the system; hence, they must be modeled mathematically to yield the best predictive power and generate strong hypotheses. Models may also prove valuable in the development of therapies. For instance, how can axons be made to grow toward a specific target area? What additional information is needed by axons which are failing to develop correctly? Sufficiently detailed theoretical models have the potential to guide experimental research in axon pathfinding through simulations done *in silico*.

Experimental Data

Guidance Cues

Axons grow along their correct trajectories by following a molecular map consisting of spatiotemporal patterns of guidance cue molecules. Four main families of molecules have been identified based on their guidance abilities – the netrins, the Slits, the semaphorins, and the ephrins – consisting of approximately 100 distinct molecules altogether (although several other classes of molecules also provide guidance information for axons, including the neurotrophins and some classical morphogens). The netrins were first identified as attractive guidance cues which direct contralaterally projecting neurons toward the midline. The transmembrane proteins DCC and Unc-5 have been shown to act as receptors for netrin-1. Both netrins and their receptors are highly conserved between species. Netrin-1 is known to have a bifunctional role, typically attracting growth cones expressing only the DCC receptor but repelling growth cones expressing both DCC and Unc-5. The Slits and their receptors, the roundabout family (the Robos), were first identified as repellents preventing contralaterally projecting neurons from recrossing the midline. Subsequently, they were also shown to stimulate axon outgrowth and branching. The semaphorins appear to act primarily as short-range cues which repel axons from particular regions or, by forming the walls of corridors, hem axons into a preferred path. However, they have also been reported to act as long-range chemoattractants. Semaphorins are classified by their structure into eight groups. They signal through multimeric receptor complexes, with the precise structure of a complex determining its specificity for a semaphorin subgroup. Semaphorin receptor molecules include the neuropilins, the plexins, and the cell adhesion molecule L1. The ephrins are substrate-bound molecules best known for their role in the formation of topographic maps in the central nervous system. For example, the graded expression of the Eph tyrosine kinases – the ephrin receptors – in the retina combined with graded expression of the ephrins in the tectum aid in the formation of an ordered topographic mapping between the two structures. Similar strategies appear to orchestrate the formation of other topographic maps. Although ephrin/Eph signaling induces axon repulsion in these examples, under other contexts the ephrins can also act to attract axons.

The Growth Cone

Growing axons are tipped by special sensorimotor structures known as growth cones. These probe their local environment and, depending on the signals they detect, direct axon outgrowth, turning, branching, and pruning. Growth cones exhibit complex morphology, as illustrated in Figure 1. They are conceptually divided into three sections: an actin-rich peripheral region, a transitional region, and a central region containing organelles and microtubules. Fingerlike protuberances extending from the edge of the growth cone known as filopodia are supported by bundles of filamentous actin (F-actin). These appear to act as sensory devices, extending the effective

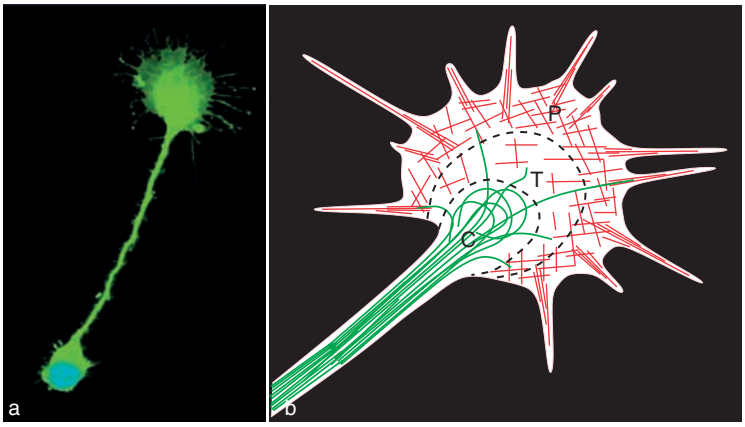

Figure 1 (a) Rat superior cervical ganglion neuron *in vitro*. The neuron was grown on a substrate of laminin for 48 h and then fixed and stained for β-tubulin. The cell body is at the bottom left, and the growth cone is at the top right. (b) Anatomy of the growth cone. This figure illustrates the division of the growth cone into three domains: transitional (T), peripheral (P), and central (C). The peripheral domain is rich in actin, which exhibits two forms of organization: loosely linked networks giving rise to lamellipodial structures and tight bundles supporting filopodia. The central domain contains organelles and microtubules which extend from the axon shaft into the growth cone. Rat SCG neuron image courtesy of Z Pujic.

sensing range of the growth cone. Structures based on more chaotic meshworks of F-actin are known as lamellipodia and have also been implicated in growth cone movement.

Growth cones undergo constant morphological change thought to be driven by remodeling of the actin network controlled by actin-binding proteins such as myosin, ARP2/3, and WASP. Growth cone motility appears to be driven by an 'actin treadmill.' Unpolymerized G-actin diffuses outward into the peripheral region and preferentially polymerizes near the cell membrane. The entire network of polymerized actin is drawn toward the central region by myosin and undergoes depolymerization in the transitional region. This constant cycle of polymerization, retrograde flow, and depolymerization is thought to generate traction when coupled to a permissive substrate through adhesion molecules.

Microtubules also play a significant, although unclear, role in growth cone motility and axon guidance. Interactions between microtubules and actin in the transitional region appear to have a strong influence on axon outgrowth and guidance.

Guidance cues influence growth cone behavior, and subsequent axon outgrowth, through cytoskeletal effectors activated or inhibited by cascades of intracellular second messengers triggered by receptor binding. Several molecules have been implicated as playing roles in this process, particularly calcium and the cyclic nucleotides cAMP and cGMP. These have been the focus of much attention, with the finding that, in some circumstances, changing the relative concentrations of these molecules within the growth cone can switch the effect of several guidance cues from attraction to repulsion or vice versa. Further intriguing findings have demonstrated that protein synthesis occurring locally within the growth cone is necessary for correct growth cone behavior.

Theoretical and Computational Models of Axon Pathfinding

Axon Extension and Branching

Experimental evidence suggests that tubulin molecules are synthesized only in the soma and then assembled into microtubules predominantly in the growth cone. This implies that axon outgrowth is limited by the rate at which microtubules can be transported to regions of active extension. A number of theoretical models have explored this idea, including various effects such as diffusive and active transport of tubulin monomers, competition between neurites for tubulin, viscoelastic stretching of axon segments, calcium-induced microtubule depolymerization, and varying intrinsic rates of tubulin polymerization and depolymerization within different growth cones. These models have been successively refined, ultimately incorporating compartment-based modeling with dynamic compartment allocation. Most strikingly, this modeling program has demonstrated that small variations in polymerization and depolymerization rates in the growth cones of different neurites can lead to sharp changes in elongation rate, including growth cone pausing and neurite retraction.

Growth cone behavior and axon extension are thought to be mediated by partially independent but related processes. This has led to modeling work

focused on characterizing the interaction between the two. The majority of this work has been phenomenological, using sophisticated mathematical machinery to better describe experimental data. Work in the mid-1980s showed that in some circumstances, axon elongation can be regarded as a one-dimensional random walk, with parameters varying with neuronal type, substrate, and chemical environment. In the mid-1990s, this was extended by performing correlation analysis which showed that axon outgrowth dynamics exhibit significant anticorrelation. Subsequent analysis showed significant correlation between the dynamics of microtubule polymerization in the central region and growth cone advance.

Growth Cone Morphology

The complex and dynamic nature of growth cone morphology further complicates our understanding of the contribution of growth cones to axon pathfinding. By statistically analyzing time-lapse images of growth cones undergoing dynamic changes in morphology, researchers have developed probabilistic rules specifying the likelihood of filopodial initiation and retraction, and also the spatial distribution of filopodia in terms of a limited number of parameters. In this model, growth cone morphology is described by the instantaneous length and angle of each filopodium, and the dynamics of the filopodia are characterized by the following parameters: the rate at which filopodia extend, the rate of retraction, the average rate at which new filopodia are initiated (modeled as a Poisson process), and the shape parameters for a gamma distribution that gives the time over which a filopodium extends before retracting. The model also specifies a simple conditionally random rule for where a filopodium initiates, and it assumes that filopodia extend radially from the center of the growth cone. Computer simulation then gives qualitatively realistic morphologies, which also satisfy quantitative constraints such as the correct average number of filopodia. By mapping the effects of external cues onto the parameters of the model, one can hope to gain some intuition as to how those cues might operate.

Other researchers have directly modeled the processes underlying actin dynamics in order to understand filopodial formation, stability, and behavior. This analysis indicates that the maximum length of a filopodium is determined by the number of bundled actin filaments in its core. For less than approximately 10 bundled actin filaments, the strain exerted on the bundle by the membrane is sufficient to cause buckling for even very short filopodia. As the number of included filaments increases, it becomes less likely that a filopodium will buckle; however, more G-actin is required for the structure to continue extending. Thus, when the number of filaments is too large, a filopodium is also unable to extend. The best tradeoff between stability and G-actin depletion is achieved with approximately 30 actin filaments. The model predicts average filopodia lengths between 1 and 10 μm, which are in agreement with experimental data. Similar work has examined the mechanisms behind lamellipodial structures.

Axon Turning and Guidance

Axons are thought to be guided by external cues through two processes: gradient-based guidance, in which the growth cone attempts to climb or descend a concentration gradient (chemoattraction or chemorepulsion), and contact-mediated guidance, in which the growth cone interacts with small regions of highly concentrated guidance cues and modulates its behavior accordingly. For gradient-based guidance, the growth cone must detect a potentially shallow gradient in the presence of noise, whereas for contact-mediated guidance the growth cone is essentially involved in a search process.

Gradient guidance Single cell chemotaxis – the attraction or repulsion of organisms such as bacteria, leukocytes, or slime molds by chemical gradients – has received a large amount of theoretical attention. Much of this can be directly applied to the case of the growth cone. Of particular relevance is a seminal contribution by Berg and Purcell, who argued that gradient detection by any small sensing device is fundamentally limited by statistical fluctuations, both due to variations in local ligand concentration and due to the inherent stochasticity of receptor binding. Growth cones are believed to sense and respond to gradients by comparing receptor binding across their spatial extent: the side of a growth cone exposed to the highest concentration of ligand will, on average, also display the largest amount of receptor binding. If growth cones do use such a spatial-sensing strategy, then in order for a growth cone to detect and reliably respond to a chemical gradient, the noise due to fluctuations in receptor binding cannot be much larger than the difference in receptor binding across its spatial extent. By modeling the physics of receptor–ligand interaction, one can estimate the limitations growth cones face when responding to chemical gradients. If the root mean squared error in a concentration measurement is given by σ_C, then the error associated with taking the difference between two such measurements is $\sigma_{\Delta C} = \sqrt{2\sigma_C}$. This gives an order of magnitude lower bound on the difference in concentration, ΔC, that the growth cone can

detect: $\Delta C_{min} \approx \sigma_{\Delta C}$. Using a simple model of receptor binding mechanics, it has been shown that for the timescale on which growth cone behavior usually occurs ($\sim$100 s), growth cones can detect gradients with a steepness of between 1% and 10%, depending on whether the guidance cue is diffusing freely or is substrate bound. However, over much longer periods (several days), experimental work has demonstrated that growth cones can respond to gradients of 0.2% or less across their spatial extent. Furthermore, at sufficiently high and low concentrations, almost all or almost no receptors are bound: this leads to further reductions in sensitivity. Experimental and theoretical work has confirmed that the highest sensitivity is achieved when approximately half of the receptors are bound, which occurs when the concentration is equal to the dissociation constant for binding.

Aside from these general constraints on gradient detection, models have also been developed which directly simulate the behavior of growth cones and axons in the presence of guidance cue gradients. Several models have focused on biochemical networks which are putatively responsible for growth cone motility and guidance. Such models are difficult to construct, partly due to the complexity of growth cone biochemistry and partly due to the lack of experimental data on important quantities such as reaction rate constants, concentrations, and interactions between molecular species. One model has focused on the Rho-GTPase signaling network, which is known to play an important role in actin-driven cell motility. Due to incomplete experimental data, the investigators took a qualitative approach, simulating the behavior of several plausible interaction networks and kinetic constants and using the results of these simulations to form hypotheses about the underlying mechanisms of growth cone motility. They found that the Rho-GTPase network undergoes a sharp transition in its dynamics when a threshold concentration of a particular signaling molecule is reached. The authors linked these two dynamic behaviors to different modes of growth cone motility, developing a model which could reproduce some experimentally observed phenomena.

Other models have placed less emphasis on specific biochemical mechanisms and have focused instead on the potential role of filopodia in axon guidance or on more general signal-processing strategies that a growth cone may implement, such as temporally or spatially averaging receptor inputs in order to reduce noise. Spatial averaging involves pooling the inputs from multiple receptors, whereas temporal averaging combines information from different time points. An interesting conclusion from this study is that spatial averaging provides the most benefit when the average is taken over approximately one-third of the growth cone's spatial extent. This optimum averaging range occurs because although spatial averaging reduces noise in a local concentration measurement, it also reduces the spatial resolution of the measurement. Because gradient detection requires concentration measurements to be made at multiple locations, the advantages gained in noise reduction are offset by the loss in resolution. Assuming growth cones use such a strategy, this has implications for the intracellular signaling network, suggesting that second messenger molecules implementing the spatial averaging process must diffuse at a rate much slower than expected for cytosolic compounds. One possibility is that spatial averaging is achieved through membrane-bound molecules.

In addition to detecting a chemical gradient, the growth cone must also amplify the possibly extremely shallow gradient in receptor binding in order to achieve a definite motile response. Understanding the mechanisms underlying this amplification has been a general focus for experimental and modeling work on microbial chemotaxis. In one influential model, amplification is achieved by coupling the external signal to a pattern formation system involving local activation and long-range inhibition. The system begins in a spatially symmetric, but unstable, steady state. Symmetry is broken by the external signal, which pushes the system into a stable, asymmetric state that reflects the direction in which the symmetry was broken. A difficulty with this approach is that the system then becomes stuck, unable to respond to new inputs such as a change in the external signal. Several additional mechanisms have been proposed to work around this, each postulating a second process which serves to reset the system to its original, unstable state. Further generalizations of this class of models suggest mechanisms for the formation of filopodia and generate testable predictions for the spatiotemporal distribution of such structures.

Contact-mediated guidance Axons are also guided by cues which are more tightly localized in space, referred to as short-range or contact-mediated cues. For example, filopodial contact with single cells expressing appropriate cues can entirely redirect an axon's trajectory. For this kind of guidance, noise is less of an issue because the signal is essentially binary: either the growth cone contacts the cue or it does not. In this situation, an appropriate theoretical framework is that of stochastic search. The question of how filopodial dynamics of the growth cone affects its ability to locate and respond to highly localized guidance cues has been addressed with the aid of the models describing growth cone morphology in terms of filopodial dynamics. The efficacy with which a growth cone is able to locate a guidance cue has been mapped against

the parameters defining the dynamics, suggesting some behaviors one might expect to observe depending on the geometry of the guidance cue distribution. This work suggests that filopodial dynamics are set, and possibly modulated, in order to increase a growth cone's ability to detect and respond to relevant cues. In a more abstract approach, growth cone movements were described by a combination of stochastic (e.g., deflection by random adhesion to the substrate) and deterministic (e.g., a tendency to move in the direction of past axon extension) motions. The growth cone was found to more effectively respond to short-range guidance signals when the two processes contributed equally. This prompted the authors to propose that growth cones modulate the relative influence of stochastic and deterministic movements depending on the importance of short-range cues at different stages in development – a suggestion consistent with experimental observations of growth cone behavior.

Axon–Axon Interactions

The models described so far consider the guidance of single neurons in isolation. However, the development of the nervous system involves the correct guidance of many axons simultaneously, and it is well established that axons use one another as additional sources of information during development. One of the earliest computer models of axon guidance attempted to explain the characteristic 'sheetlike' pattern of axon outgrowth observed in the formation of the ventral commissure of the spinal cord. Using a descriptive model of individual axon behavior, including several experimentally observed features of ventral commissural axon growth – a tendency for straight growth, for initial outgrowth to be directed ventrally, and for growing axons to extend preferentially over the surrounding matrix and not other axons – this work attempted to distinguish the most important features of individual axon behavior for the formation of axon sheets. From computer simulations, the authors concluded that initially polarized outgrowth, a suitably high density of neurons, preferential adhesivity for extension over the substrate rather than other axons, and a tendency for straight growth were sufficient to generate the observed patterns. Another model examined the possibility that growth cones secrete diffusible guidance cues in order to attract or repel one another to create or break up axon bundles.

Topographic Map Formation

A specific example of axon guidance that has been well studied theoretically is the formation of the topographic map between the retina and optic tectum/superior colliculus. In 1963, Roger Sperry first proposed that such maps could arise because gradients of molecular labels in the retina are matched to gradients of labels in the tectum. The subsequent discovery of gradients of Eph receptors and their ligands, the ephrins, in the retina and tectum confirmed this prediction. However, a large number of experiments investigating how such matching might work in detail have suggested that several other constraints are also important. Since the 1970s, numerous theoretical models of such map formation have been proposed in order to gain insight into this complexity. Some of the simplest propose sorting mechanisms, whereby an initial random map is refined by comparing the retinal origin of axons terminating at neighboring sites in the tectum. Others have hypothesized that tectal labels are at least partly induced or modified by transport of retinal labels into the tectum. Several models have highlighted the importance of competition in map formation, both between axons for tectal target space and between tectal targets for axons. Another important theme has been cooperative effects between axons, somewhat similar to the axon–axon interactions discussed previously. Increasing data on the precise role of Eph/ephrins in map formation have provided new challenges for such models, many of which are yet to be addressed.

Guidance Cue Patterning

A further area of active research aims to understand how guidance cue patterns are generated in the first place, and how effectively particular patterns can guide axons. The modeling of gradient systems in developing organisms has a long history, and gradients are thought to be a primary means for generating spatial ordering. In general, molecules expressed as gradients in order to provide spatial information are known as morphogens, and several classical morphogens have been shown to also guide axons.

A number of models have been proposed to explain how gradients of appropriate shape and stability could be set up. The simplest model assumes that the molecule of interest is diffusing away from a continuous source through a homogeneous medium. More complex models recognize the inhomogeneous nature of the medium, degradation of molecules, binding of molecules to cells, endocytosis, and active transport processes. A further complication which arises when attempting to generalize results from one experimental model to another is that of scaling: gradients form on a typical length scale, and different species have embryos of different sizes at developmental stages when axon wiring is forming. Hence, a system which works in one embryo may not work in another.

Additional constraints are placed on the formation of gradients useful for axon guidance. The ability of a

growth cone to detect and respond to a gradient varies with background concentration and gradient slope. The minimum gradient that a growth cone can detect over a particular background concentration specifies limitations on gradient-based guidance. It allows the construction of optimal gradients, in which the gradient slope is always equal to the minimal detectable gradient for the growth cone. Following this line of argument, coupled with estimates of parameters central to the model, it can be shown that the maximum distance over which growth cones can be guided by an optimal gradient is on the order of 1 cm.

Future Directions

This article provided an overview of the kinds of models which have been applied to axon guidance and how these have helped us to understand axon pathfinding. However, research of this kind is still at an early stage. Important questions remain to be answered and are the focus of active research. For instance, how sensitive can growth cones be to gradients of guidance cues? How close do they come to achieving fundamental sensitivity limits? What are the actual mechanisms they use to detect gradients? How do developing neurites integrate information from multiple guidance cues? What searching strategies do growth cones use to locate local guidance cues? What roles do axon branching and pruning play in axon guidance? How do microtubules and the F-actin cytoskeleton interact to support axon outgrowth and steering? How much of a role do axon–axon interactions play in the formation of the nervous system, and when are they important?

In addition to fresh modeling approaches, answering these questions will require significant experimental advances. A wealth of experimental data is available on axon guidance, but most studies have been aimed at identifying guidance cue molecules or intracellular molecules mediating or eliciting particular behaviors. Although such data are obviously crucial, in order to generate sufficiently constrained models, data of a more quantitative nature are needed. Recognizing this need, experimental techniques for producing well-controlled and flexible patterns of guidance cues have been developed. Ultimately, these techniques should allow us to develop better constrained models and, using them, obtain additional power to tease apart the mechanisms and principles underlying axon guidance.

Finally, the discovery that new neurons are constantly being born in adult brains opens up another area for exploration. These nascent neurons must somehow find their way to their appropriate niches and extend axons to make functional connections. Modeling axon guidance thus has a central role to play in understanding both the initial development and the normal functioning of the nervous system.

See also: Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Regeneration: Role of Growth and Guidance Cues; Cognitive Neuroscience: An Overview; Computational Methods; Growth Cones; Neuroplasticity: Computational Approaches.

Further Reading

Berg HC and Purcell EM (1977) Physics of chemoreception. *Biophysical Journal* 20: 193–219.

Buettner HM (1996) Analysis of cell-target encounter by random filopodial projections. *American Institute of Chemical Engineering Journal* 42: 1127–1138.

Dickson BJ (2002) Molecular mechanisms of axon guidance. *Science* 298: 1959–1964.

Goodhill GJ (1998) Mathematical guidance for axons. *Trends in Neurosciences* 21: 226–231.

Goodhill GJ, Gu M, and Urbach JS (2004) Predicting axonal response to molecular gradients with a computational model of filopodial dynamics. *Neural Computation* 16: 2221–2243.

Goodhill GJ and Xu J (2005) The development of retinotectal maps: A review of models based on molecular gradients. *Network: Computation in Neural Systems* 16: 5–34.

Gordon-Weeks PR (2000) *Neuronal Growth Cones*. Cambridge, UK: Cambridge University Press.

Hely TA and Willshaw DJ (1998) Short-term interactions between microtubules and actin filaments underlie long-term behaviour in neuronal growth cones. *Proceedings of the Royal Society of London, Series B* 265: 1801–1807.

Katz MJ, George EB, and Gilbert LJ (1984) Axonal elongation as a stochastic walk. *Cell Motility* 4: 351–370.

Krottje JK and van Ooyen A (2007) A mathematical framework for modeling axon guidance. *Bulletin of Mathematical Biology* 69: 3–31.

Maskery SM and Shinbrot T (2005) Deterministic and stochastic elements of axonal guidance. *Annual Review of Biomedical Engineering* 7: 187–221.

Meinhardt H (1999) Orientation of chemotactic cells and growth cones: Models and mechanisms. *Journal of Cell Science* 112: 2867–2874.

Mortimer D, Fothergill T, Pujic Z, Richards LJ and Goodhill GJ (2008) Growth cone chemotaxis. *Trends in Neurosciences* 31: 90–98.

Sakumura Y, Tsukada Y, Yamamoto N, and Ishii S (2005) A molecular model for axon guidance based on cross talk between Rho GTPases. *Biophysical Journal* 89: 812–822.

van Veen MP and van Pelt J (1994) Neuritic growth rate described by modeling microtubule dynamics. *Bulletin of Mathematical Biology* 56: 249–273.

Willshaw DJ and von der Malsburg C (1979) A marker induction mechanism for the establishment of ordered neural mappings: Its application to the retinotectal problem. *Philosophical Transactions of the Royal Society of London, Series B* 287: 203–243.

Axonal Pathfinding: Extracellular Matrix Role

P Letourneau, University of Minnesota, Minneapolis, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Normal behavior and other neural activities depend on the correct wiring of neural circuits during development. A critical step in forming neural circuits is the growth of axons from nerve cell bodies to the sites where synaptic connections are made. In spanning between neural somata and their synaptic targets, growing axons forge pathways that become the axonal tracts and peripheral nerves of the mature nervous system. The routes that axons take to reach their targets are determined by motile activities at their tips, called growth cones. Growth cones extend fine protrusions that adhere to nearby cells and surfaces. These adhesive contacts provide a toehold from which further protrusions are made. As growth cones crawl forward, they choose a path by detecting and responding to the spatial and temporal distributions of extracellular guidance molecules encountered in their local environments. Five major families of extracellular molecules – netrins, neurotrophins, semaphorins, slits, and ephrins – provide positive and negative cues that orient the migration of growth cones to their targets. These guidance molecules bind receptor proteins on growth cones and initiate cytoplasmic signals that regulate the motility and adhesive contacts that determine the advance, retreat, turning, branching, and stopping of growth cones. This article describes molecules that play a key role in axonal pathfinding by mediating the adhesive interactions necessary for growth cone migration.

Mechanism of Growth Cone Migration

Cytoskeletal Dynamics

Growth cone migration and axonal elongation involve the cytoskeletal components, microtubules and actin filaments. Axon elongation requires the advance and polymerization of microtubules, which are bundled in the axon but which spread apart in the growth cone, where individual microtubules dynamically probe forward to the front of a growth cone via polymerization and movement involving microtubule motor molecules (Figure 1). Axonal growth occurs where the main microtubule bundle and associated organelles advance in the growth cone, as determined by the positions and stabilization of these forward 'pioneering' microtubules. These dynamic microtubules project forward into an actin filament network that fills flattened dynamic projections, called lamellipodia, and fingerlike filopodia. This extensive filament system is continually remodeled, as actin filaments initiate and polymerize at the front margin and are then moved back to be fragmented and depolymerized, recycling subunits to the front. Multiple actin-binding proteins regulate this dynamic organization of actin filaments.

Growth cone migration is driven by forces produced within this actin filament domain. Actin polymerization creates protrusive forces that expand lamellipodia and elongate the tips of filopodia. Myosin motor molecules bind actin filaments and generate mechanical forces that move cargo bound to the myosin tail domains or pull on actin filaments to create tensions. Myosin II motor activity pulls actin filaments rearward, where they are depolymerized. Tensions generated by myosin II activity in the actin-rich leading margin can either direct or halt microtubule advance, depending on the situation. Myosin II-generated tensions produce the exploratory movements of lamellipodia and filopodia, whereas excessive levels of tension may sweep microtubules back in a contracting actin network that can collapse a growth cone. It is in the context of these dynamic cytoskeletal activities that adhesive interactions are critical to growth cone migration (Figure 2).

Adhesive Contacts of Growth Cones

Growth cone plasma membranes contain adhesion receptors that bind noncovalently to adhesion molecules on other cells or surfaces. Lamellipodia and filopodia initiate adhesive interactions as they explore their environment, and if these bonds persist, receptors cluster to form discrete adhesive contacts, which include intracellular adhesion complexes. Adhesion complexes remain in place or shift rearward as a growth cone advances. These adhesive complexes play two roles in growth cone migration. First, they include proteins that anchor actin filaments at adhesive sites. These links constitute a 'clutch' that stops the retrograde movement of actin filaments and permits the advance of microtubules and axonal organelles (Figure 2). Without stabilization provided by adhesive contacts growth cone migration fails, and tensions within the axonal cytoskeleton cause axonal retraction. Second, these complexes include proteins of signaling cascades, protein kinases, protein phosphatases, and Rho GTPases, which act on proteins that regulate the organization of actin filaments and microtubules. Thus, adhesive contacts provide points

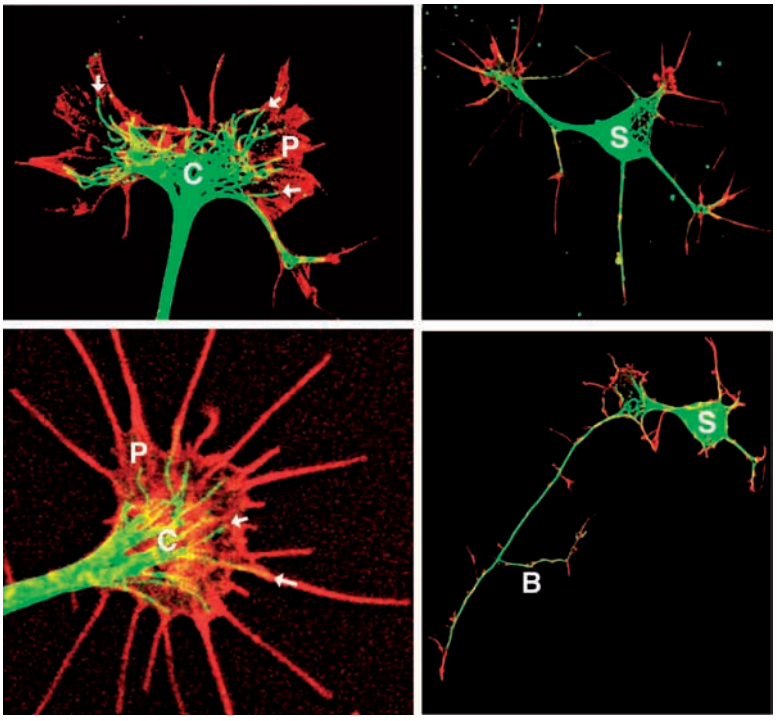

Figure 1 The distribution of microtubules and actin filaments in developing neurons and in axonal growth cones. Microtubules (green) are densely packed in the neuronal cell bodies (S) and are bundled in the axons and branches. Actin filaments are arrayed in filament networks and bundles in the peripheral domains (P) of the growth cones and along the shafts of the axons, where small areas of actin filament dynamics may give rise to collateral branches (B). In a growth cone, the microtubules from the central bundle of the central domain (C) splay apart and individual microtubules extend into the P domain and into filopodia (arrows).

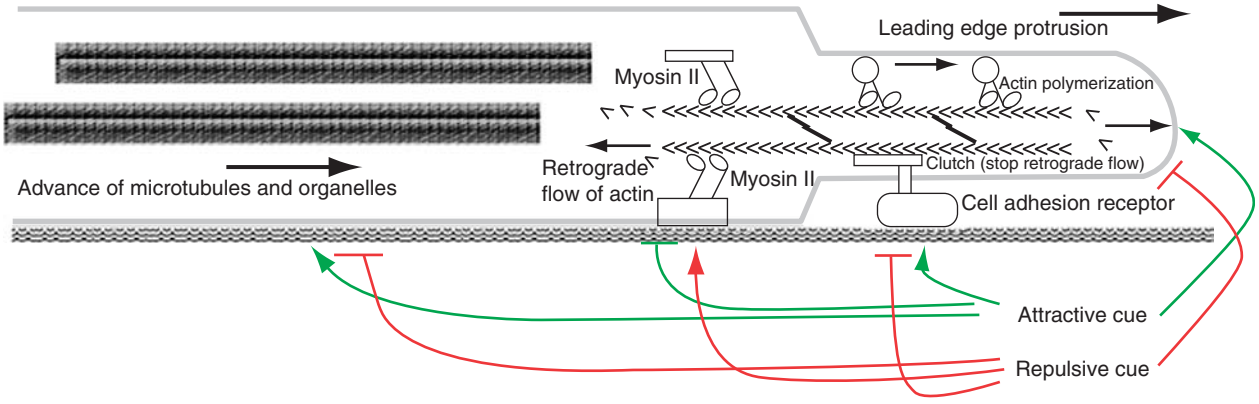

Figure 2 A model of the mechanism of growth cone migration. Actin polymerization pushes the leading margin of the growth cone forward. Forces generated by myosin II pull actin filaments backwards, where filaments are disassembled. When growth cone receptors make adhesive contacts with a surface, a 'clutch' links the adhesive contact to actin filaments of the leading edge, and the retrograde flow of actin filaments stops. This permits the advance of microtubules and organelles and promotes axonal elongation. Intracellular signaling generated by attractive and repulsive axonal guidance cues interacts with the molecular mechanisms of actin polymerization, myosin II force generation, adhesive contacts, and microtubule advance to regulate the paths of growth cone migration.

of stability that are essential to growth cone migration, and they are signaling centers from which regulatory activities promote growth cone motility.

The genetic regulation that determines neuronal phenotype also directs expression of receptors for adhesive ligands and guidance cues by the growth cones of neurons of a particular type. Extracellular positive and negative axonal guidance cues, whether surface bound or soluble, signal through their receptors to modulate an interacting set of pathways that regulate cytoskeletal and membrane dynamics. Thus, growth cone behaviors reflect a complex integration

of signaling events triggered at multiple receptors for guidance cues and adhesion molecules. By locally regulating the interplay of adhesive contacts and cytoskeletal dynamics within a growth cone, guidance cues determine the pathways of axonal growth (Figure 2).

Three major types of adhesive interactions promote growth cone navigation. Growth cones migrate within extracellular spaces that contain a complex mixture of glycoproteins, organized into an extracellular matrix (ECM) of fibers, protein aggregates, and basal laminae, which are discrete ECM layers at tissue interfaces. One major adhesive interaction of growth cones involves binding of integrin receptors to adhesive ECM proteins, especially the laminins. Two other major adhesive interactions involve growth cone contacts with cells or other axons along their pathways. These interactions involve two groups of adhesive molecules, the cadherins and the immunoglobulin superfamily of cell adhesion molecules (IgCAMs). Cadherins are expressed on all tissue types, including neurons and axons. Cadherin adhesions involve homophilic binding between like cadherin molecules on two interacting cells. Weaker heterophilic interactions between different cadherins can also occur. IgCAMs are also expressed on all tissues, including neurons. Adhesive interactions of IgCAMs can involve homophilic interactions, similar to cadherins, but also heterophilic interactions in which an IgCAM on a growth cone binds a different IgCAM on adjacent cells. Even heterophilic interactions of IgCAMs with non-IgCAMs occur.

Integrin adhesion receptors Integrin receptors are heterodimers of alpha and beta subunits. More than 20 integrin heterodimers have been identified in humans. The binding specificity for ECM components depends on the particular combination of alpha and beta subunits in a heterodimer. The 12 integrin dimers that are expressed in the mammalian nervous system include receptors for collagens, laminin-1 and laminin-5, fibronectin, tenascin, thrombospondin, vitronectin, and VCAM-1. A growth cone can express multiple integrins, allowing interactions with multiple ECM molecules. The cytoplasmic domains of integrins lack enzymatic activities, but when integrins bind adhesive ligands, conformational shifts in the cytoplasmic domains trigger formation of focal contacts that involve integrin clustering and creation of docking sites for proteins that initiate signaling and links to the cytoskeleton. When lamellipodia and filopodia of growth cones bind laminin-1, proteins that localize to the contact sites include paxillin, talin, vinculin, zyxin, and focal adhesion kinase (FAK). Vinculin and talin link actin filaments to the adhesive contacts, providing a clutch for growth cone migration (Figure 3). The presence of the adapter protein paxillin and activation of FAK initiates further protein interactions and signaling by Src family kinases, MAP kinases, and Rho GTPases. Activation of Rac1 and Cdc42 GTPases promotes actin polymerization by regulating actin-binding proteins and actin filament dynamics. Thus, when integrins on growth cones bind laminin-1, growth cone migration is stimulated by increased actin filament polymerization to protrude the leading margin and by the establishment of adhesions to stabilize these protrusions and promote the advance of microtubules.

Cadherins and IgCAMs Stimulation of N-cadherin and IgCAMs, such as NCAM and L1, by ligand binding between cells leads to activation of the FGF receptor tyrosine kinase, which triggers signals involving PLC-gamma, DAG lipase, cytoplasmic [Ca^{2+}] elevation, and activation of MAPK. IgCAMs also signal via Src kinases to activate Rac1, PI3K, and MAPK. Cadherin signaling is also reported to activate Rac1. Thus, several pathways activated by cadherins and IgCAMs promote actin filament polymerization. Adhesive binding of cadherins and IgCAMs provides anchorage for actin filaments, creating the clutch necessary for growth cone migration. The cytoplasmic tails of many cadherins, such as N-cadherin, bind catenins, which bind actin filaments and link N-cadherin adhesive sites to the actin cytoskeleton in growth cones (Figure 4). The cytoplasmic domain of L1 binds the cytoskeletal linker ankyrin, but L1–ankyrin interactions are involved in stable adhesive junctions, such as at nodes of Ranvier, and not in growth cone migration. Members of the ezrin–moesin–radixin (ERM) proteins mediate actin filament binding to membranes, and interactions of L1 (and other IgCAMs) with ERM proteins may serve as a clutch in growth cone protrusions that bind via L1 or other IgCAMs.

These adhesion receptors can be regulated in ways that are important to growth cone pathfinding. The expression levels of integrin receptors on growth cones are increased when laminin levels are low or when ECM proteins, such as proteoglycans, which interfere with laminin–integrin binding, are present. These responses would maintain growth cone adhesion and migration in environments when access to laminin is reduced. L1 is endocytosed from central regions and recycled to the leading margin of growth cones, increasing availability of L1 for adhesive contacts of lamellipodia and filopodia. The functions of adhesion receptors can also be modulated from the cytoplasm in an 'inside-out' manner or via *cis* interactions with other components of the plasma membrane. An important manner in which guidance

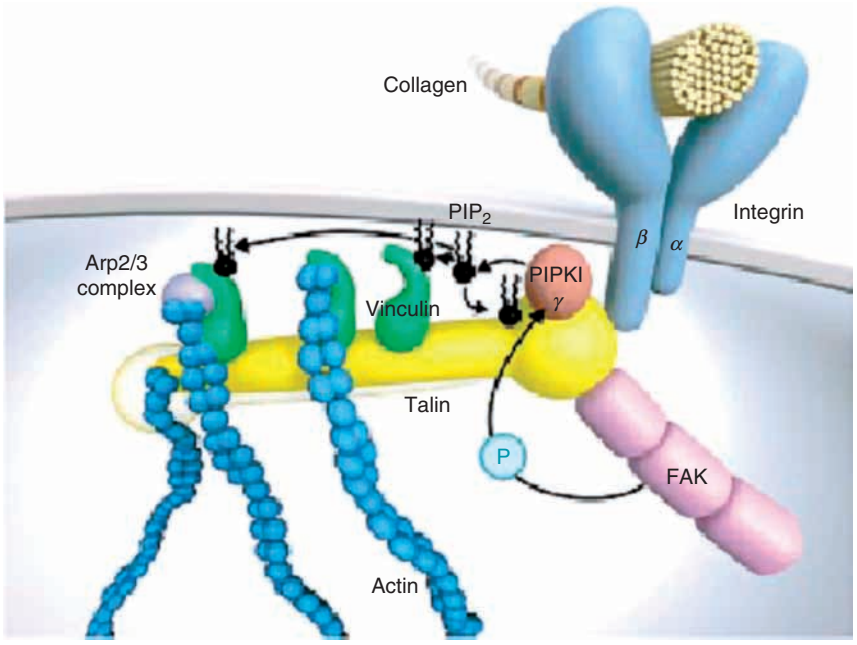

Figure 3 A model of integrin binding to ECM molecules and the formation of intracellular adhesive complexes. An alpha–beta integrin heterodimer is shown bound to a collagen fibril, and the intracellular adhesion complex is pictured, showing the proteins vinculin and talin, which are involved in linkage to actin filaments, and FAK kinase, which initiates signaling cascades. The Arp2/3 complex nucleates actin filament assembly. Reproduced from Brakebusch C and Fässler R (2003) The integrin–actin connection, an eternal love affair. *EMBO Journal* 22: 2324–2333, with permission from Nature Publishing Group.

molecules exert their positive and negative effects on growth cone pathfinding is by modulating the functions of adhesive receptors (Figure 2). For example, the negative cue semaphorin 3A may inhibit growth cone migration by blocking integrin-mediated cell adhesion. In addition, adhesion mediated by N-cadherin is inhibited by the negative guidance cue Slit protein via its receptor, Robo. Thus, the negative or repulsive effects of semaphorin 3A and Slit on growth cone pathfinding can involve these inhibitory effects on growth cone adhesion. On the other hand, the attractant netrin signals to activate the kinase FAK, which promotes integrin-mediated adhesion, suggesting that positive guidance cues activate adhesive interactions of growth cones.

Adhesion Molecules and Growth Cone Pathfinding

What are the roles of these adhesion molecules in the pathfinding behaviors of growth cones? Major pathways that are followed by many growing axons offer multiple adhesive ligands for growth cone migration, such as laminins, fibronectin and collagens in the ECM, and cadherins and IgCAMs on adjacent cells and axons. These multiple options for adhesion may provide redundancy, ensuring growth cones form sufficient adhesive contacts for effective migration. The first growth cones that 'pioneer' a pathway have limited options for binding to ECM or cell surface adhesion molecules on adjacent cells, whereas growth cones that enter an established pathway can track along previously extended axons by binding to cadherins and IgCAMs expressed on the surfaces of axonal shafts. Several *in vivo* examples of pathfinding roles of adhesion molecules are described in the following sections.

Laminins

Laminins are large adhesive glycoproteins (MW 1 000 000 Da) that consist of heterotrimers of alpha, beta, and gamma chains. Ten laminin chains are known, forming 11 known heterotrimers with widely varied expression throughout different tissues. The laminins present several domains that mediate laminin binding to several cell surface receptors and to other ECM components. The most common laminins are typically present in basal laminae, an ultrastructural ECM layer associated with epithelia, muscle cells, Schwann cells, and glia. Laminin-1, which has been studied the most, promotes axonal growth *in vitro* from virtually every type of neuron, indicating that laminins have broad roles in promoting growth cone migration. Examples of growth cone migration along basal laminae include growth cones of Rohon–Beard

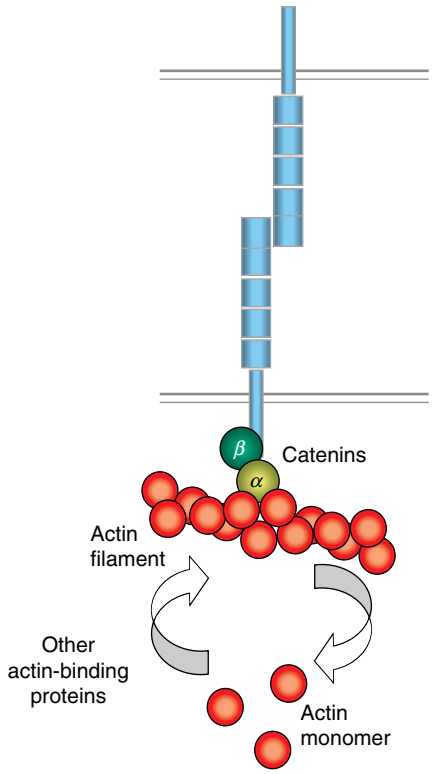

Figure 4 A model of homophilic adhesion between cadherin adhesion molecules and the intracellular binding to actin filaments. Cadherin molecules bind between cells and become linked to actin filaments by way of interactions with alpha and beta catenins. From Weis WI and Nelson WJ (2006) Re-solving the cadherin–catenin–actin conundrum. *Journal of Biological Chemistry* 281: 35593–35597.

neurons in *Xenopus*, growth cones of retinal ganglion cells in the retina and optic nerve, and pioneer axons in the grasshopper limb bud. However, in addition to basal laminae, laminin is transiently expressed in the loose cellular environments of developing tissues, including the nervous system, on cell surfaces and associated with sparse ECM fibers. The growth cones that pioneer pathways, such as the corticofugal pathway of the neocortex or the medial longitudinal fasciculus from the brain into the spinal cord, migrate within loose extracellular spaces in the wall of the immature central nervous system (CNS), where the cells are labeled in a punctate manner by laminin antibodies. The expression of laminin on these cells is transient, and eventually laminin immunoreactivity is restricted to the basal lamina at the outer boundary of the CNS wall. In the developing peripheral nervous system (PNS), laminin is expressed in basal laminae and at early stages in the mesenchyme through which motor and sensory axons extend. Schwann cells express abundant laminin, forming the basal laminae that enclose axon–Schwann cell units. This punctate cellular expression of laminin diminishes during development, although laminin remains present in basal laminae.

In view of the wide distribution of laminins and the ability of laminin-1 to promote robust axonal growth from many neuronal types, it is thought that laminins function permissively, providing adhesion that is required for growth cone migration, but not in an instructive manner to influence pathfinding decisions. Laminins and other ECM molecules may broadly promote growth cone migration along a pathway, whose boundaries are defined not by the absence of adhesive ECM molecules but, rather, by the expression in adjacent tissues of negative guidance cues, such as slits or semaphorins. This 'surround repulsion' occurs in both developing CNS and PNS. Several mutational studies have reported specific errors in pathfinding when a laminin is absent or blocked. Laminin function is essential for growth cone turning in the grasshopper limb bud, and zebra fish with mutations in the laminin-alpha-1 chain exhibit multiple axon guidance defects throughout the CNS, but not in every location. These results suggest that laminin-mediated adhesion is essential for growth cone navigation in at least some instances.

Fibronectin

Fibronectin is a large adhesive glycoprotein (MW 250 000 Da) that is widely distributed in the ECM, including within ganglia and the endoneurium of the PNS. Like the laminins, the fibronectin molecule contains multiple domains that mediate binding to other ECM components and to multiple cellular receptors, including several integrin heterodimers. During development of the PNS and CNS, fibronectin is present in a punctate distribution in loose cellular spaces of immature nervous tissue, and eventually fibronectin expression diminishes as development ends, especially in the CNS. In tissue culture studies, fibronectin promotes axonal growth, but not as vigorously as does laminin. In addition, axonal growth by PNS neurons on fibronectin surfaces exceeds the responses of CNS neurons, probably because PNS neurons express higher levels of fibronectin receptors than CNS neurons. Evidence is lacking for a requirement for fibronectin in growth cone pathfinding.

Integrins

Because neurons express multiple integrin subunits and because many ECM components, such as collagen, laminin, or fibronectin, can bind more than one integrin heterodimer, the essential roles of particular integrins in growth cone pathfinding are not clearly defined. Mouse knockouts of α_1 and α_6 integrins,

which are laminin-1 receptors, do not reveal clear defects, however, injections of anti-β_1 integrin, part of several neuronal receptors for ECM proteins, into *Xenopus* embryos disrupts retinal axonal pathfinding. Similarly, conditional knockout of β_1 integrin in sensory neurons results in deficits in innervation of skin, where sensory axons extend through the dermal ECM and along the epidermal basal lamina. The α_4ß$_1$ integrin heterodimer is specifically implicated in the growth and arborization of sympathetic axons within cardiac muscle. In *Drosophila*, mutations in the integrins α-PS1 and-PS2 lead to pathfinding errors.

Cadherins

Cadherins are characterized as single-pass transmembrane proteins that contain an ectodomain of five cadherin repeats and a conserved cytoplasmic tail. Binding of calcium ion stabilizes an extended rodlike structure of the ectodomain, which is necessary for optimal adhesion by alignment of cadherin molecules on apposing cells. There are at least 100 cadherins, and most are expressed in the developing vertebrate brain on immature cells, neurons, and glia. Their functions are numerous, including cell sorting, boundary formation, target recognition, synaptogenesis, and synapse function. Regarding axonal pathfinding, the widely expressed N-cadherin stimulates *in vitro* axonal growth from a variety of CNS and PNS neurons. *In vivo* studies involving antibody injection or genetic mutation also implicate N-cadherin in axon growth and fasciculation. These results indicate that cadherins promote growth cone migration along axons in highly populated common pathways, but it is unclear whether cadherins play a role in the pathfinding of early pioneer growth cones. In some cases, a common pathway may be shared by several classes of elongating axons, which are distinguished by the expression of different cadherins. For example, the tectofugal projections of chickens express four different cadherins among different axon fascicles. These cadherins may mediate specific pathfinding, as the formation of homophilic adhesions of growth cones to axons expressing the same cadherin directs growth cones along specific axon fascicles toward their targets. Forced expression of specific cadherins causes growth cones to abnormally follow fascicles that express the same cadherin. Finally, growth cones often share expression of specific cadherins with neurons in their particular target. Thus, cadherins also have roles in target recognition and subsequent synaptogenesis.

L1 and NCAM IgCAMs

Proteins that contain an immunoglobulin (Ig)-like domain constitute the Ig superfamily, which makes up more than 2% of human genes, constituting the largest gene family. The neuronal Ig superfamily includes a large number of molecules, which have functions in axonal pathfinding not only as cell adhesion molecules but also as axonal guidance cues and as receptors of guidance cues. This discussion is restricted to two members of this large family, L1 and NCAM. The IgCAM L1 is widely expressed on axons in the developing CNS and PNS, and tissue culture studies show that substrates coated with L1 promote homophilic adhesion and axonal growth from many neuronal types. Spontaneous human mutations in the L1 gene and mouse L1 knockout studies both indicate important roles for L1 in brain development and function. Multiple anatomical and functional deficits result from human and mouse L1 mutations, including a failure of corticospinal axons to decussate in the hindbrain pyramids. Crossing defects were not found in other tracts or were not so extensive. L1 also interacts in *cis* with receptors for other guidance cues, including the semaphorin 3A receptor, suggesting that the defect in pyramidal decussation observed in L1 mutants could be due to disrupted pathfinding responses to semaphorin 3A and other guidance cues, as well as to reduced growth cone tracking along axons.

Another prominent neuronal IgCAM is NCAM, the first neuronal IgCAM identified. NCAM is widely expressed on immature neurons and glia and also on other embryonic cells, such as myoblasts. In tissue culture studies, NCAM mediates neuronal adhesion and axon growth. In addition to homophilic adhesive interactions, NCAM also forms heterophilic adhesive interactions. Antibodies against NCAM can induce axon defasciculation *in vitro* and *in vivo*. Several isoforms of NCAM are expressed, and in some situations NCAM carries a carbohydrate polysialic acid (PSA) moiety that reduces NCAM adhesion. In NCAM-deficient mice defects in fasciculation of hippocampal axons were observed, but in general only minor defects in development or behavior were observed. Perhaps, in the absence of NCAM, other cell adhesion molecules serve the same functions.

Adhesion Molecules and Axonal Regeneration

When the pathfinding phase of circuit construction ends, as growth cones reach their targets, the expression of neural cell adhesion molecules and their receptors is downregulated. However, injury or damage to nervous tissues can disconnect neural circuits, and axons must regenerate in order to reconnect neurons. When axons are injured in the PNS, axon regeneration is often robust, leading to varying degrees of functional recovery. Schwann cells, which ensheath all PNS axons, stimulate axonal regeneration by upregulating their expression of growth factors, and

laminins, fibronectin, and cadherin, as substrates for growth cones. Axon regeneration in the PNS is also promoted by increased expression of integrin receptors by regenerating neurons. In the CNS of adult mammals, regeneration of injured axons is poor, and recent research has focused on inhibitory components of myelin and glial scars that block growth cone adhesion and trigger signals that inhibit growth cone motility. In lower vertebrates, CNS regeneration is often successful, and this involves the upregulation of expression of adhesive ligands, such as L1 and cadherins, as demonstrated in regenerating zebra fish optic nerves and spinal cords.

Several strategies for improving axonal regeneration in mammalian model systems, and eventually humans, emphasize measures to improve the adhesive environment for growth cone migration. When stem cells that express L1 are transplanted into a mammalian CNS lesion, increased regeneration of corticospinal axons occurs. Purkinje cells transfected to express L1 and GAP43 show enhanced axonal regeneration. *In vitro* regeneration of axons by adult neurons on laminin and fibronectin is improved by transfection of neurons to express increased levels of the appropriate α integrin chains. Finally, many natural and synthetic bridges have been designed that include adhesion molecules to promote axonal regeneration across lesion sites. These studies demonstrate that strategies to increase the adhesive interactions of regenerating growth cones can stimulate axonal regeneration after injuries in adults. Probably, improved axonal regeneration in adults will also require an increase in the intrinsic ability of adult neurons to sprout and grow axons. This may involve upregulation of genes for adhesion receptors, for other guidance cue receptors, and for proteins that drive the dynamic cytoskeletal functions of immature neurons.

Summary

Growth cone adhesion is integral to the mechanism of growth cone migration and pathfinding. Adhesive interactions of growth cones provide stability for lamellipodial and filopodial protrusions of growth cones and also act as signaling centers that regulate actin and microtubule dynamics and organization in a migrating growth cone. The adhesive interactions of growth cones are also a target of guidance cues that determine where growth cones turn, branch, and stop migrating. Migrating growth cones make three kinds of adhesive contacts with ECM and with other cells. These contacts involve integrin receptors, which recognize laminin and other ECM components; cadherins, which form homophilic adhesions; and IgCAMs, which can form homophilic and heterophilic adhesive interactions. Major pathways of growth cone migration during development contain one or, perhaps more typically, multiple adhesive ligands available to growth cones. The navigational decisions of growth cone pathfinding are based on local differences in adhesive stability for growth cone protrusions and in dynamic protrusive activity, as based on adhesive signaling and the integration of signaling triggered from other guidance cues.

See also: Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Pathfinding: Netrins; Axonal Pathfinding: Guidance Activities of Sonic Hedgehog (Shh); Growth Cones; Semaphorins.

Further Reading

Brakebusch C and Fässler R (2003) The integrin–actin connection, an eternal love affair. *EMBO Journal* 22: 2324–2333.

Clegg DO, Wingerd KL, Hikita ST, and Tolhurst EC (2003) Integrins in the development, function and dysfunction of the nervous system. *Frontiers in Bioscience* 8: d723–d750.

Colognato H, French-Constant C, and Feltri ML (2005) Human diseases reveal novel roles for neural laminins. *Transactions in Neuroscience* 28: 480–486.

Dent EW and Gertler FB (2003) Cytoskeletal dynamics and transport in growth cone motility and axon guidance. *Neuron* 40: 209–227.

Gordon-Weeks PR (2000) *Neuronal Growth Cones*. Cambridge, UK: Cambridge University Press.

Hortsch M (2003) Neural cell adhesion molecules – brain glue and much more! *Frontiers in Bioscience* 8: d357–d359.

Huber AB, Kolodkin AL, Ginty DD, and Cloutier JF (2003) Signaling at the growth cone: Ligand–receptor complexes and the control of axon growth and guidance. *Annual Review of Neuroscience* 28: 509–563.

Hynes RO (2002) Integrins. Bidirectional, allosteric signaling machines. *Cell* 110: 673–687.

Kamiguchi H (2003) The mechanism of axon growth: What we have learned from the cell adhesion molecule L1. *Molecular Neurobiology* 28: 219–228.

Kiryusho D, Berezin V, and Bock E (2004) Regulators of neurite outgrowth: Role of cell adhesion molecules. *Annals of the New York Academy of Sciences* 1014: 140–154.

Redies C, Treubert-Zimmermann U, and Luo J (2003) Cadherins as regulators for the emergence of neural nets from embryonic divisions. *Journal of Physiology (Paris)* 97: 5–15.

Sakisaka T and Takai Y (2005) Cell adhesion molecules in the CNS. *Journal of Cell Science* 118: 5407–5410.

Suter DM and Forscher P (2000) Substrate–cytoskeletal coupling as a mechanism for regulation of growth cone motility and guidance. *Journal of Neurobiology* 44: 97–113.

Thiery JP (2003) Cell adhesion in development: A complex signaling network. *Current Opinion in Genetics & Development* 13: 365–371.

Weis WI and Nelson WJ (2006) Re-solving the cadherin–catenin–actin conundrum. *Journal of Biological Chemistry* 281: 35593–35597.

Wen Z and Zheng JQ (2006) Directional guidance of nerve growth cones. *Current Opinion in Neurobiology* 16: 52–58.

Axonal Pathfinding: Guidance Activities of Sonic Hedgehog (Shh)

F Charron, Institut de Recherches Cliniques de Montréal (IRCM) and University of Montreal, Montreal, QC, Canada
M Tessier-Lavigne, Genentech, Inc., South San Francisco, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Over the past decade, genetic, biochemical, and molecular approaches have led to the identification of four major conserved families of guidance cues with prominent developmental effects: the netrins, slits, semaphorins, and ephrins. More recently, members from three other families of secreted signaling molecules have been shown to act as guidance cues: the wingless/Wnt, the hedgehog (Hh), and the decapentaplegic/bone morphogenetic protein/transforming growth factor-β (Dpp/BMP/TGF-β) families. In addition to their axon guidance properties, these molecules share a common characteristic of having been previously identified as morphogens controlling cell fate and tissue patterning. This discovery has opened the door to the study of an entirely new set of axon guidance cues.

This article focuses on the role of the morphogen Shh in axon guidance. After briefly introducing the role and the signaling pathway of Shh in patterning the neural tube, we discuss the emerging evidence that Shh is reused later in development to guide axons, both in the developing spinal cord and the retina. We conclude by discussing the implications for Shh signaling in axon guidance.

The Hedgehog Family in Cell Fate Specification and Tissue Patterning

Morphogens are signaling molecules produced in a restricted region of a tissue; they provide positional information by diffusing from their source to form long-range concentration gradients. A cell's program of differentiation in response to a morphogen is dictated by its position within the gradient and thus on its distance from the morphogen source. Two criteria have gained acceptance as the evidence needed to qualify a secreted signaling protein as a morphogen: it must have a concentration-dependent effect on its target cells and it must exert a direct action at a distance. In vertebrates, Sonic hedgehog (Shh) has been shown to fulfill these criteria and function as a morphogen to specify cell fate in the developing neural tube.

Hedgehog proteins are found in insects and vertebrates, but not nematodes. They are encoded by a single hedgehog gene in flies, and three in mammals: *Shh*, Indian hedgehog (*Ihh*), and Desert hedgehog (*Dhh*). In vertebrate embryos, one of the first steps in the development of the nervous system is the specification of the diverse neural cell fates of the neural tube. Shh is secreted by the notochord and by floor plate cells at the ventral midline of the neural tube, and functions as a graded signal for the generation of distinct classes of ventral neurons along the dorsoventral axis of the neural tube (Figure 1(a)). In agreement with its role as a morphogen, Shh is able to induce a range of ventral spinal cord cell fates in a concentration-dependent manner and has been shown to exert a direct action at a distance to specify neural tube cell fate.

Much evidence indicates that the cell fate specification and tissue patterning activities of Hhs are mediated by members of the Ci/Gli transcription factor family, but the signaling mechanisms that lead to activation of these transcription factors are not fully elucidated. Genetic and biochemical experiments have shown that Hhs activate signaling by binding to their receptor Patched (Ptc), which leads to the relief of Ptc-mediated inhibition of Smoothened (Smo), also a transmembrane protein, which can then activate downstream signaling (Figure 2). Smo associates directly with a Ci-containing complex, which contains the atypical kinesin Costal-2 (Cos2), the protein kinase Fused (Fu), and the Suppressor of Fused (Su(fu)). This complex constitutively suppresses pathway activity. Activation of Hh signaling reverses this regulatory effect and allows Ci to activate transcription of Hh target genes, thus specifying cell fate.

Roles of Shh in Axon Guidance

Shh Is a Chemoattractant for Commissural Axons

During spinal cord development, commissural neurons, which differentiate in the dorsal neural tube, send axons that project toward and subsequently across the floor plate, forming axon commissures (Figure 1(b)). These axons project toward the midline in part because they are attracted by netrin 1, a long-range chemoattractant secreted by the floor plate. In mice mutant for *netrin 1* or the netrin 1 receptor gene, *Dcc* (*deleted in colorectal cancer*), many commissural axon trajectories are foreshortened, fail to invade the ventral spinal cord, and are misguided. However, some of them do reach the midline, indicating that other guidance cues cooperate with netrin 1 to guide these axons. Further analysis of *netrin 1* knockout

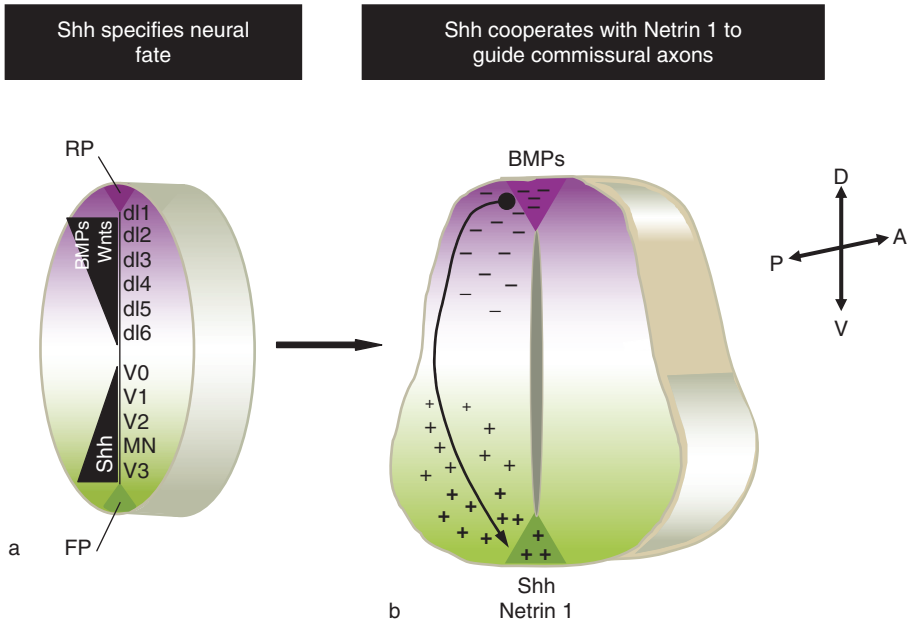

Figure 1 Neuronal cell fate specification and guidance of precrossing commissural axons by Sonic hedgehog (Shh) and netrin 1; cross-section representations of the developing neural tube. Shh is first used to pattern neural progenitors in the spinal cord and then appears to be reused as a guidance cue for commissural axons. (a) In the early neural tube, Shh – together with bone morphogenetic protein (BMP) and Wnt – protein concentration gradients act to specify neural cell fate in the ventral (and dorsal, for BMP and Wnt) spinal cord. (b) Later, the axons of differentiated commissural neurons are attracted to the ventral midline by the combined chemoattractant effects of netrin 1 and Shh. BMPs also contribute to commissural axon guidance by repelling axons from the dorsal midline. V0–V3, ventral interneuron subpopulations; dI1–dI6, dorsal interneuron subpopulations; MN, motor neurons; RP, roof plate; FP, floor plate; D, dorsal; V, ventral; P, posterior; A, anterior. Adapted from Charron F and Tessier-Lavigne M (2005) Novel brain wiring functions for classical morphogens: A role as graded positional cues in axon guidance. *Development* 132: 2251–2262.

mice suggests that the floor plate might actually express an additional diffusible attractant(s) for commissural axons.

Given its expression by the floor plate and its long-range effects in the spinal cord, Shh was a candidate for a midline-derived axonal guidance cue. Shh was indeed shown to function as an axonal chemoattractant that can mimic the netrin-1-independent chemoattractant activity of the floor plate in *in vitro* assays. The chemoattractant activity of Shh, like the chemoattractant activity of floor plate derived from *netrin 1* mutants, can be blocked by cyclopamine, which blocks the actions of Shh in cell fate determination by inhibiting the Shh signaling mediator Smo. This shows that Smo is required for Shh-mediated axon attraction and, importantly, that the netrin-1-independent chemoattractant activity of the floor plate also requires Hh signaling. Since Shh is the only Hh family member expressed in the spinal cord at this stage, these results suggest that Shh functions as a floor plate-derived chemoattractant for commissural axons.

While the reorienting effect of Shh could be due to a direct chemoattractant effect, an alternative explanation was suggested by the fact that Shh is a potent morphogen. Since in these assays commissural axon turning occurs within the spinal cord tissue explant, it seemed possible that Shh was not acting directly on the axons but rather was repatterning and altering the expression of guidance cues by cells within the explant, which then secondarily guided the axons to the Shh source. Arguing against this possibility was the finding that the spinal cord explants used to assess chemoattractant activity are at a developmental stage at which they have apparently lost the competence to be repatterned by Shh, as assessed using a battery of markers of dorsoventral patterning.

A direct action of Shh in attracting the axons was supported further by two sets of experiments. First, Shh was shown to attract the growth cones of isolated *Xenopus* spinal axons in dispersed cell culture in a cyclopamine-dependent manner, proving that Shh, acting via Smo, can function as a chemoattractant, at least for these *Xenopus* axons. A second way of providing evidence that Shh can act directly on commissural axons to guide them relied on blocking Shh signaling selectively in commissural neurons without blocking it in the terrain through which their axons course. This was achieved by conditional inactivation of a floxed allele of *Smo* using the Cre recombinase

expressed under the control of the Wnt1 promoter, which drives expression in the dorsal spinal cord (as well as in neural crest progenitors). When Cre, driven by this promoter, was used to delete a floxed *Smo* allele in the dorsal spinal cord, commissural axon trajectories were defective in the ventral spinal cord, where Cre is not expressed (Table 1). This result strongly implies that the axonal misrouting is not due to repatterning of the ventral spinal cord, and must instead reflect a guidance defect arising from loss of Smo function in commissural neurons. Taken together, these results suggest that Shh functions to guide commissural axons both *in vitro* and *in vivo* by acting directly as a chemoattractant on these axons through an Smo-dependent signaling mechanism.

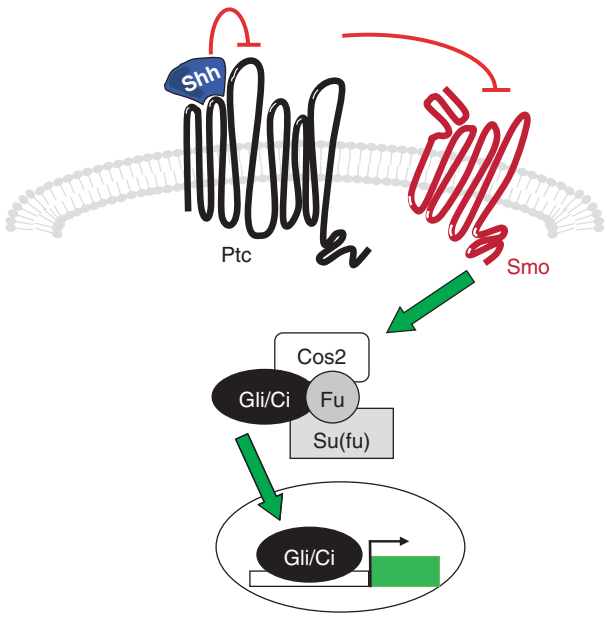

Figure 2 The Sonic hedgehog (Shh) signaling pathway. Genetic and biochemical experiments have shown that hedgehogs activate signaling by binding to their receptor Patched (Ptc; a 12-pass transmembrane protein), which leads to the relief of Ptc-mediated inhibition of Smoothened (Smo), a 7-pass transmembrane protein, which can then activate downstream signaling. Smo associates directly with a Gli/Ci-containing complex which contains the atypical kinesin Costal-2 (Cos2), the protein kinase Fused (Fu), and the Suppressor of Fused (Su(fu)). This complex constitutively suppresses pathway activity by leading to the proteolytic cleavage of Ci, which acts as a transcriptional repressor. Activation of hedgehog signaling reverses this regulatory effect and leads to the production of full-length Ci, which activates transcription of hedgehog target genes.

Boc Is a Receptor for Shh in the Guidance of Commissural Axons to the Floor Plate

Although Shh acts through Smo to attract commissural axons to the floor plate, Smo does not bind Shh, and the binding receptors acting with Smo to mediate the effect of Shh in axon guidance remain elusive. Cdon (cell adhesion molecule-related/downregulated by oncogenes) and Boc (biregional Cdon binding protein) have been implicated in enhancing muscle differentiation. They are type I transmembrane proteins consisting of four to five immunoglobulin (Ig) and two to three fibronectin type III (FNIII) repeats in the extracellular domain. This domain architecture is highly related to that of axon guidance receptors of the Robo and Dcc families. Both Cdon and Boc share a high degree of homology in their extracellular domains and are expressed during early stages of central nervous system (CNS) development. Interestingly, mice with homozygous mutations in Cdon display a microform of holoprosencephaly (HPE), a developmental defect of the forebrain and midface caused by a failure to delineate the midline. In both humans and mice, many forms of HPE are caused by disruptions in the Shh signaling pathway, suggesting that Cdon and Boc might regulate Shh signaling, a possibility supported by the finding that a Cdon/Boc homolog in *Drosophila*, iHog, is required for Hh signaling. Taken together, these observations suggested that Cdon and/or Boc might function in Shh-mediated axon guidance in vertebrates.

The ability of Boc and Cdon to bind Shh was tested and they were found to bind specifically and directly to Shh, suggesting that they could function as Shh receptors. These results are consistent with recent reports demonstrating that Cdon, Boc, and iHog bind and

Table 1 *In vivo* experiments supporting a role for Shh in axon guidance

Gene	*Species*	*Experiment*	*Phenotype*	*Reference*
Smo	Mouse	Conditional inactivation of *Smo* in commissural neurons	Commissural axons project abnormally and invade the motor columns	Charron et al. (2003)
Boc	Mouse	*Boc* inactivation	Commissural axons project abnormally and invade the motor columns	Okada et al. (2006)
Shh	Chick	Ectopic expression of Shh in the optic chiasm	Retinal axons are prevented from crossing the chiasm	Trousse et al. (2001)
Shh	Chick	Silencing of *Shh* by RNA interference	Commissural axons stall at the contralateral floor plate border, with some axons randomly turning caudally or rostrally	Bourikas et al. (2005)

signal through the Hh pathway. To explore their involvement in commissural axon guidance, the expression of Cdon and Boc was examined in the developing mouse spinal cord. Boc is expressed by differentiating commissural neurons and the Boc protein, but not Cdon, is expressed by growing commissural axons. To assess the functions of these receptors in commissural axon guidance, Boc and Cdon mutant mice were generated and sections of spinal cord from Boc and Cdon mutant embryos were analyzed. Although commissural axon projections appeared normal in Cdon mutant animals, abnormal projections of commissural axons were observed in Boc mutants: the axons were highly dispersed and invaded the ventral spinal cord, with ectopic projections extending over the motor columns. Analysis of various neural tube markers indicated that neuronal patterning is occurring normally in these animals. This phenotype is similar to that observed in mice following conditional removal of the Shh signaling mediator Smo in commissural neurons (using a Wnt1 promoter to drive Cre expression; see earlier) and suggests that Boc acts in the same pathway as Smo to guide commissural axons in response to Shh. Finally, using an *in vitro* commissural axon turning assay to test the role of Boc in Shh-mediated axon turning, it was shown that RNA interference (RNAi)-mediated knockdown of Boc impaired the ability of commissural axons to turn toward an ectopic source of Shh *in vitro*. These results indicates that Boc is required for commissural axons to respond to the chemoattractive effect of Shh. Collectively, these data suggest that Boc plays an essential role as a receptor for Shh in commissural axon guidance.

Consistent with a role for Boc in axon pathfinding, an earlier study showed defects in forebrain axon guidance in zebra fish treated with morpholinos directed against Boc. However, based on indirect evidence, these results were interpreted to reflect a role for Boc as a repulsive ligand. In the mammalian spinal cord, the aforementioned results argue that Boc functions as a receptor for Shh-mediated attraction; whether it can also function as a repulsive ligand in other contexts in mammals remains to be explored. Additionally, it will be of interest to determine whether Boc and Cdon also play a role in the guidance of other axonal tracts in the developing mammal and fish embryos.

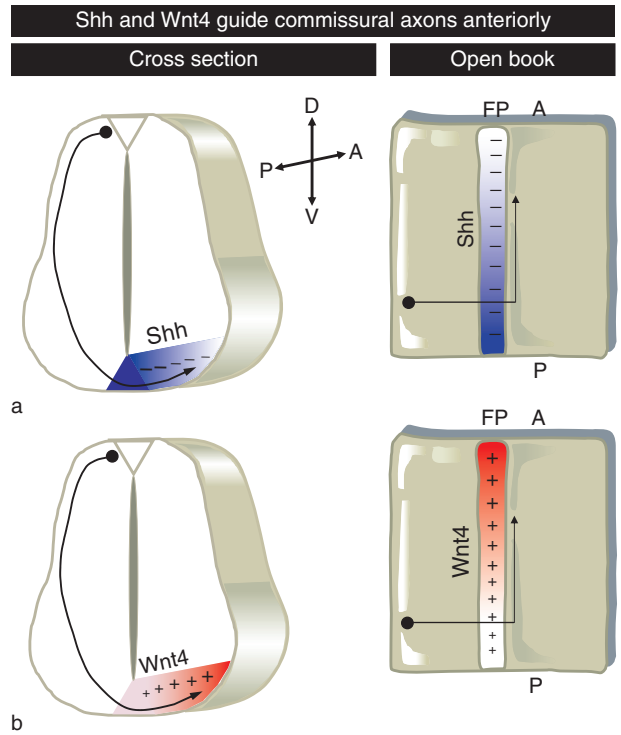

Figure 3 Guidance of postcrossing commissural axons by Sonic hedgehog (Shh) and Wnt4. (a) After crossing the floor plate, chick commissural axons are repelled from the posterior pole by an increasing anterior to posterior Shh gradient. (b) In rodents, commissural axons are attracted to the anterior pole by an increasing posterior-to-anterior Wnt4 gradient. Left panels in a and b are cross-section representations of the developing spinal cord, and right panels are open-book representations. FP, floor plate; D, dorsal; V, ventral; P, posterior; A, anterior. Adapted from Charron F and Tessier-Lavigne M (2005) Novel brain wiring functions for classical morphogens: A role as graded positional cues in axon guidance. *Development* 132: 2251–2262.

Shh Guides Commissural Axons along the Longitudinal Axis of the Spinal Cord

After commissural axons have reached and crossed the floor plate, they make a sharp anterior turn toward the brain (Figure 3). The molecules involved in the dorsoventral projection of commissural axons to and at the floor plate have been well described, but it is only recently that cues controlling anteroposterior guidance have been identified. Remarkably, the guidance of commissural axons to the floor plate is not, apparently, the only effect of Shh on commissural axons: recent evidence suggests that Shh also guides postcrossing commissural axons in the rostral direction along the longitudinal axis of the spinal cord. Using a subtractive hybridization approach to identify guidance cues responsible for the rostral turn of postcrossing commissural axons in chick embryos, Bourikas and colleagues identified differentially expressed candidates whose function they investigated by RNAi-mediated *in ovo* gene silencing. Unexpectedly, one of their candidates turned out to be Shh. In agreement with these results, silencing of the *Shh* gene by a different RNAi construct or injection of a hybridoma producing a function-blocking Shh antibody led to axon stalling at the contralateral floor plate border, with some axons

turning caudally or rostrally, apparently in a random manner. Importantly, marker analysis revealed that the patterning of the spinal cord was not apparently affected by these manipulations, suggesting that these experiments were performed after neural cell fate specification by Shh has occurred. Finally, postcrossing commissural axons were shown to avoid ectopic Shh *in vivo*. Together, these results provide strong evidence that Shh is essential for the normal guidance of commissural axons along the longitudinal axis of the spinal cord, at least in chick embryos.

An Shh gradient could guide commissural axons along the longitudinal axis directly, or could alternatively be acting only indirectly by controlling a graded distribution of a distinct guidance cue. Two lines of evidence, however, were provided for a direct role of Shh. The first came from an investigation of the receptor mechanism for this guidance. Interestingly, neither cyclopamine nor Smo RNAi interfered with the rostral turn of commissural axons along the longitudinal axis, suggesting that Smo might not be involved in this process. Instead, RNAi-mediated silencing of *Hip1*, a gene encoding an Shh-binding membrane protein transiently expressed in commissural neurons at the time when they cross the floor plate, as well as in the periventricular region, resulted in the same postcrossing phenotype as seen with Shh RNAi. These results, which contrast with the essential role of Smo in Shh-mediated attraction of commissural axons to the floor plate, suggest that Hip1 might be involved in transducing an Shh guidance signal in postcrossing commissural neurons. The relatively restricted expression of Hip1 mRNA to commissural neurons would be consistent with a direct action of Shh on these axons. A second line of evidence that supports a direct role for Shh was obtained by *in vitro* experiments, which showed that postcrossing commissural axons from spinal cord explants can be repelled by Shh beads *in vitro*. Taken together, these results suggest a model in which Shh could be functioning directly through Hip1 as a chemorepellent for postcrossing commissural axons.

Prior to the finding that Shh controls the anteroposterior guidance of commissural axons in chicks, Wnt4 was reported to play a role in this process in rodents. Using a novel *in vitro* assay, evidence was obtained that the activity responsible for the anterior guidance of postcrossing commissural axons in rodents is an increasing posterior to anterior gradient of a diffusible attractant. Although it is not yet known whether Wnt4 guides postcrossing commissural axons in chicks and whether Shh guides postcrossing commissural axons in rodents, it is nonetheless interesting to note that, if this is the case, the complementary Wnt4 and Shh gradients might act cooperatively in the rostral guidance of commissural axons. Additionally, since Boc plays a role in the guidance of precrossing commissural axons, it will be interesting to determine whether it also plays a role in the guidance of postcrossing commissural axons.

Shh Is a Negative Regulator of Retinal Ganglion Cell Axon Growth

Retinal ganglion cell (RGC) axons growing toward the diencephalic ventral midline are faced with the decision to project either contralaterally or ipsilaterally in response to guidance cues at the optic chiasm (Figure 4). Homozygous inactivation of the mouse *Pax2* gene alters the development of the optic chiasm, and RGC axons never cross the midline in these mice. Interestingly, whereas in wild-type mice Shh expression is downregulated in the chiasm as RGC axons are migrating toward this region, Shh expression is ectopically maintained along the ventral midline in *Pax2*$^{-/-}$ mice. These observations raise the possibility that the continuous expression of Shh at the ventral midline might contribute to preventing RGC axon crossing. In agreement with this idea, Trousse and colleagues found that ectopic expression of Shh in the midline region interferes with RGC axon growth and prevents them

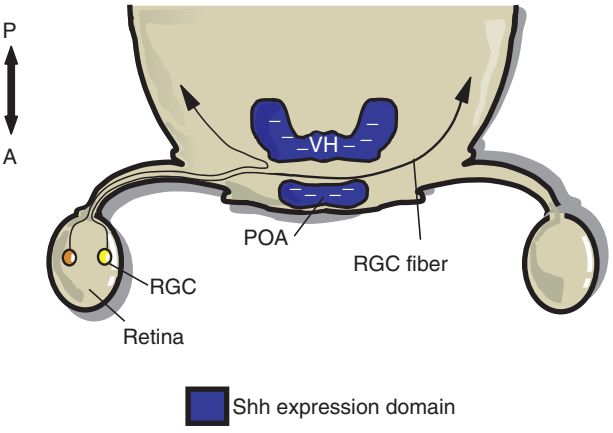

Figure 4 Sonic hedgehog (Shh) expression at the chiasm border defines a barrier within the ventral midline implicated in guiding retinal ganglion cell (RGC) axons. RGC axons growing toward the diencephalic ventral midline are faced with the decision to project either contralaterally or ipsilaterally in response to guidance cues at the optic chiasm. The Shh expression domain is shown in blue. Shh can inhibit retinal axons *in vitro*, suggesting that *in vivo* it may be acting on the axons directly rather than by altering the expression of distinct guidance cues in the chiasm, although conclusive evidence for this guidance function *in vivo* remains to be obtained. A, anterior; P, posterior; POA, preoptic area; VH, ventral hypothalamus. Adapted from Charron F and Tessier-Lavigne M (2005) Novel brain wiring functions for classical morphogens: A role as graded positional cues in axon guidance. *Development* 132: 2251–2262.

from crossing the midline (Figure 4). Consistent with the idea that Shh might be acting directly on RGC axons, it was shown that these manipulations do not affect patterning and neural differentiation in the eye. Further experiments will be required to determine whether the chiasm region is repatterned in these experiments, but *in vitro* experiments support the idea that Shh acts directly to control RGC axon migration: addition of exogenous recombinant Shh to retinal explants decreases the number and length of growing axons, without interfering with the rate of proliferation and differentiation of cells in the explant, and time-lapse analysis shows that addition of Shh to retinal explants rapidly causes growth cone arrest and subsequent retraction of RGC axons. Since the response of the growth cone to many extracellular guidance cues appears to be modulated and in some cases perhaps even mediated by intracellular cyclic nucleotide levels (cyclic adenosine monophosphate and cyclic guanosine monophosphate, cAMP and cGMP), the possibility was explored that the effect of Shh on retinal axons *in vitro* might be due to a change in cAMP levels. In agreement with this, addition of Shh to retinal growth cones was shown to decrease intracellular levels of cAMP, a finding consistent with the observation that lowering cAMP levels favors growth inhibition.

Taken together, these results provide evidence that Shh expression at the chiasm border helps define a barrier within the ventral midline that serves to guide RGC axons, and suggest that Shh may be acting on the axons directly, rather than indirectly by repatterning the chiasm. The finding that Shh acts directly to guide RGC axons is also supported by a recent study suggesting that Shh acts directly and rapidly on the growth cone of RGCs cultured *in vitro*. Proving that the effect *in vivo* is direct will, however, require additional studies, such as identifying the mechanism that mediates retinal growth cone responses to Shh and showing that cell-autonomous inhibition of this signaling pathway in the neurons results in guidance defects *in vivo*.

Molecular Mechanism underlying Shh-Mediated Axon Guidance

The aforementioned studies, summarized from several laboratories, have provided evidence that Shh plays a role in the guidance of at least two types of axons: commissural and RGC axons. In the case of precrossing commissural axons, Shh acts as a chemoattractant and its effect is mediated by Boc, a recently identified Shh receptor related to Dcc and Robo family members. In addition, the chemoattractant effect of Shh requires the canonical Shh signaling molecule Smo. After commissural axons have crossed the floor plate, evidence indicates that Shh then functions as a chemorepellent to direct their anterior migration. This later effect of Shh appears to require the Shh binding protein Hip1 and might occur independently of Smo activity. Finally, in the case of RGCs, Shh acts as a negative regulator of axonal growth, at least *in vitro*, but the receptor and the signaling molecules involved in this effect remain elusive.

Although the chemoattractant and chemorepellent effects of Shh might appear to be inconsistent with one another, many other guidance cues have been shown to be bifunctional and exert opposite effects, depending on the context. For example, extrinsic factors can convert netrin attraction to repulsion by modulating cyclic nucleotide levels. Thus, the opposite effects of Shh on pre- and postcrossing commissural and retinal axons might be due to an intrinsic or extrinsic factor that modulates cyclic nucleotide levels. Alternatively, as the molecular mechanisms underlying the effects of Shh on commissural and retinal axons are poorly understood, it is also possible that these pre- and postcrossing effects are mediated by distinct signaling pathways that result in opposite guidance effects – a possibility that also has a precedent in the case of netrins, which can attract axons by activating Dcc family receptors and repel them by activating UNC5 family receptors.

In all of the studies summarized here, Shh was shown to act rapidly (in an hour or less) to affect growth cone morphology. Although these results appear inconsistent with the model that the Shh axon guidance effect function through the transcriptional signaling pathway to the nucleus, this needs to be formally proved, as none of the studies described here has addressed this issue directly. Nonetheless, even if a transcriptional response is found to be required for their guidance effects, additional local signaling would still be required to be elicited in the growth cone in order to generate a polarized response leading to growth cone turning in a specific direction. Indeed, a purely transcriptional response consisting of a retrograde signal to the nucleus, followed by an anterograde signal back to the growth cone, cannot account for the polarized turning effect of a guidance cue. Further studies aimed at understanding the molecular mechanisms underlying growth cone turning by Shh will be necessary to identify the molecules linking Shh signaling to localized growth cone effects.

In this regard, despite many efforts, the canonical Hh signaling pathway is only beginning to be understood, and many intermediate signaling molecules remain to be identified and characterized. Thus, it

is possible that the signaling proteins eliciting the growth cone effects are simply components of the signaling pathway required for cell fate specification that are awaiting identification and further characterization to uncover their function in axon guidance.

Alternatively, Shh might be acting through entirely different signaling pathways in axon guidance and cell fate specification, including the use of a different receptor. For example, although Smo and Boc are required for Shh-mediated commissural axon guidance to the floor plate, it is not known whether Ptc, the canonical Shh-binding component of the Shh receptor, is involved. This finding contrasts with chick postcrossing commissural axon guidance, in which Smo does not appear to be required for the rostral turn away from the Shh gradient. Additional experiments on commissural and retinal axons will be required to determine the receptor components and intracellular signaling molecules mediating the effects of Shh on axon guidance.

Conclusion

The discovery that the morphogen Shh can be reused to guide axons has generated considerable excitement in the field, and it will be interesting to see to what extent its signaling components are conserved between the morphogenic and guidance responses. The characterization of the Shh chemotropic signaling pathway will open new avenues to study how guidance signals regulate the motility and steering of the growth cones, and will help elucidate the mechanisms directing the complex wiring of the nervous system.

See also: Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Pathfinding: Netrins; Growth Cones; Sonic Hedgehog and Neural Patterning; Wnt Pathway and Neural Patterning.

Further Reading

Bourikas D, Pekarik V, Baeriswyl T, et al. (2005) Sonic hedgehog guides commissural axons along the longitudinal axis of the spinal cord. *Nature Neuroscience* 8: 297–304.

Charron F, Stein E, Jeong J, et al. (2003) The morphogen sonic hedgehog is an axonal chemoattractant that collaborates with netrin-1 in midline axon guidance. *Cell* 113: 11–23.

Charron F and Tessier-Lavigne M (2005) Novel brain wiring functions for classical morphogens: A role as graded positional cues in axon guidance. *Development* 132: 2251–2262.

Connor RM, Allen CL, Devine CA, et al. (2005) BOC, brother of CDO, is a dorsoventral axon-guidance molecule in the embryonic vertebrate brain. *Journal of Comparative Neurology* 485: 32–42.

Dickson BJ (2002) Molecular mechanisms of axon guidance. *Science* 298: 1959–1964.

Fuccillo M, Joyner AL, and Fishell G (2006) Morphogen to mitogen: The multiple roles of hedgehog signalling in vertebrate neural development. *Nature Reviews Neuroscience* 7: 772–783.

Ingham PW and McMahon AP (2001) Hedgehog signaling in animal development: Paradigms and principles. *Genes and Development* 15: 3059–3087.

Kolpak A, Zhang J, and Bao ZZ (2005) Sonic hedgehog has a dual effect on the growth of retinal ganglion axons depending on its concentration. *Journal of Neuroscience* 25: 3432–3441.

Lyuksyutova AI, Lu CC, Milanesio N, et al. (2003) Anterior–posterior guidance of commissural axons by Wnt-frizzled signaling. *Science* 302: 1984–1988.

Okada A, Charron F, Morin S, et al. (2006) Boc is a receptor for sonic hedgehog in the guidance of commissural axons. *Nature* 444: 369–373.

Tessier-Lavigne M and Goodman CS (1996) The molecular biology of axon guidance. *Science* 274: 1123–1133.

Trousse F, Marti E, Gruss P, et al. (2001) Control of retinal ganglion cell axon growth: A new role for Sonic hedgehog. *Development* 128: 3927–3936.

Varjosalo M and Taipale J (2007) Hedgehog signaling. *Journal of Cell Science* 120: 3–6.

Axonal Pathfinding: Netrins

S W Moore and T E Kennedy, McGill University, Montreal, QC, Canada
M Tessier-Lavigne, Genentech, Inc., South San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Directing growing axons to their targets is an essential step toward establishing appropriate connections in the nervous system. The growth cone, located at the motile tip of an axon, senses cues within its environment to guide extending axons. Extracellular guidance cues can attract or repel axons. They may be transmembrane, glycosylphosphatidylinositol (GPI)-linked, or secreted cues. Secreted cues have the unique capacity to function at a distance by diffusing away from their source. The netrin family contains both secreted members (netrin-1–netrin-4) and GPI-linked members (netrin-G1 and netrin-G2).

The discovery of the netrin family can be traced back to observations made in the late nineteenth century by the Spanish neuroscientist Santiago Ramón y Cajal. He was the first to propose that gradients of diffusible cues might guide axons. One location where he thought this may occur is in the developing spinal cord where commissural axons extend ventrally toward the floor plate (Figure 1(a)). Specifically, he proposed that a gradient of a diffusible cue, emanating from the floor plate at the ventral midline, would function to attract growing commissural axons. In the 1980s, co-culture studies using explants of embryonic neural tube provided experimental evidence for the presence of guidance cues secreted by the floor plate. Initial experiments demonstrated that axon bundles would extend from an explant containing the cell bodies of commissural neurons, through a collagen matrix toward a floor plate explant cultured at a distance (Figure 1(b)). Subsequent experiments placed an ectopic floor plate perpendicular to the normal trajectory of commissural axons, resulting in the axons being redirected within the neural epithelium toward the ectopic floor plate (Figure 1(c)). An activity that promoted commissural axon outgrowth was then identified in lysates of embryonic chick brains. Two related proteins were purified, their corresponding cDNAs cloned, and recombinant protein shown to mimic the ability of the floor plate to attract commissural axons. Sequence analysis identified homology to UNC-6, a protein required for the circumferential guidance of cells and axons in the roundworm *Caenorhabditis elegans*. These proteins were named netrins, based on the Sanskrit word *netr*, meaning 'one who guides.'

Netrin Family Members

Four secreted netrins (netrin-1–netrin-4) have been identified in vertebrates, along with two membrane-anchored forms, netrin-G1 and netrin-G2. Thus far, orthologs of netrin-2 have only been identified in birds and fish. Mammals express secreted netrin-1, netrin-3, and netrin-4. The netrin family can be divided into three subfamilies: netrins 1–3 (sometimes called 'classical netrins'), netrin-4, and netrin-G1 and netrin-G2. Based on sequence, netrin-4 and the netrin-Gs are more similar to laminins than to netrins 1–3, which exhibit a high degree of similarity to each other (Figure 2(b)). Another compelling argument for this subdivision is based on evolutionary conservation. Orthologs of netrins 1–3 have been detected in all bilaterally symmetrical animals studied so far, while orthologs of netrin-4 and netrin-Gs have only been found in vertebrate species (Figure 2(c)). Functional differences also support this distinction. Different receptors mediate the function of netrins 1–3 and netrin-Gs. Furthermore, netrins 1–3 similarly elicit axon outgrowth from embryonic spinal commissural neurons, while netrin-Gs do not.

Netrin Structure

Despite these differences, all netrins are classified into a single family on the basis of their size, approximately 600 amino acids, the presence of two characteristic conserved N-terminal domains, domains V and VI, and a more variable C-terminal domain, domain C. Domains V and VI in netrins are homologous to domains V and VI found at the N-terminal ends of the extracellular matrix protein laminin (Figure 2(a)). Laminins are large secreted heterotrimers made up of α, β, and γ subunits. Domains V and VI of netrin-4 and netrin-Gs are most similar to β subunits of laminin, while those of netrins 1–3 are more similar to the γ subunits (Figure 2(b)).

Domain VI, at the N-terminal end of netrins and laminins, is composed of approximately 300 amino acids. In laminins, this domain interacts with heparin, cell surface receptors, and extracellular matrix (ECM) proteins, and is required for calcium-dependent multimerization that generates a larger matrix of laminin molecules. Genetic studies in *C. elegans* have demonstrated that the highly conserved sequence SXDXGXS/TW within domain VI of the netrin UNC-6 is required for both axon attraction and repulsion. The middle

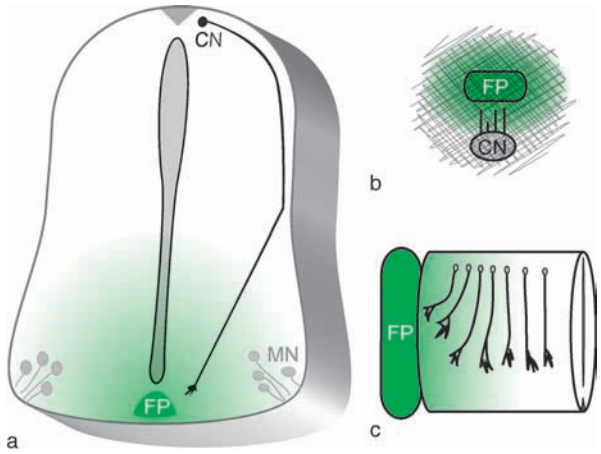

Figure 1 (a) The floor plate (FP) attracts commissural neuron (CN) axons, in part, by releasing netrin-1 (MN, motor neuron). (b) Commissural neurons will extend axon bundles from an explant of dorsal embryonic spinal cord through a collagen matrix when either floor plate or cells expressing netrin are placed within 200–300 μm. (c) An ectopic floor plate or a cell line expressing netrin-1 will attract commissural neuron axons when placed perpendicular to their trajectory.

netrin domain, domain V, is approximately 150 amino acids long and contains three cysteine-rich epidermal growth factor (EGF) repeat subdomains named V-1, V-2, and V-3. Mutation of the third EGF repeat (V-3) disrupts chemoattraction in *C. elegans*, whereas repulsion is lost following mutation of either V-2 or V-3 domain.

The C-terminal domain C of secreted netrins (netrins 1–4) exhibits relatively limited sequence similarity to domains found in the complement C3, C4, and C5 protein family (i.e., CC3, CC4, and CC5), secreted frizzled-related proteins (sFRPs), type I C-proteinase enhancer proteins (PCOLCEs), and tissue inhibitors of metalloproteinases (TIMPs). The majority of netrin protein in the embryonic or mature central nervous system (CNS) is not freely soluble. Structure–function analyses of domain C suggest that it contributes to binding netrins to cell surfaces or extracellular matrix. Domain C includes many basic amino acids that may bind negatively charged sugars associated with proteoglycans, such as heparan sulfate proteoglycans and chondroitin sulfate proteoglycans. Presentation of

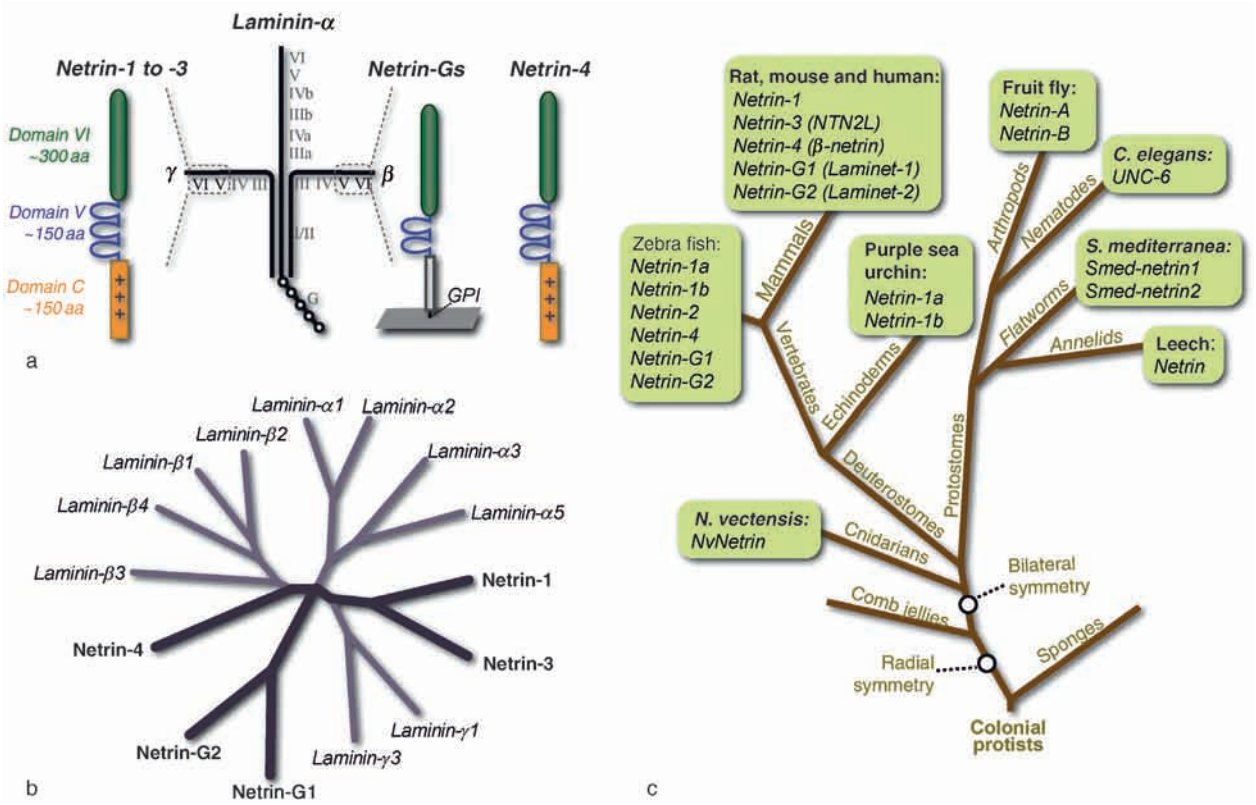

Figure 2 (a) Netrins contain N-terminal domains V and VI related to corresponding N-terminal domains of laminins. Domain V is composed of cysteine-rich epidermal growth factor repeats. Domain C in secreted netrins contains many positively charged, basic residues (GPI, glycophosphatidylinositol). (b) Phylogenetic tree based on the sequences of domains VI and V in human netrins and laminins. (c) Evolutionary tree diagram highlighting the presence of netrin homologs in a wide variety of bilaterally symmetrical organisms.

netrins bound to surfaces may be a common mode of action in the netrin family; although the domain C is not conserved in the netrin-Gs, a C-terminal GPI link anchors them to cell surfaces.

Netrin Expression and Function during Development

Orthologs of netrin-1 have a firmly established, evolutionarily conserved role as secreted cues that direct axon guidance relative to the midline of developing bilaterally symmetric nervous systems. In fact, a netrin-1 ortholog has been identified in the stellate sea anemone *Nematostella vectensis* – an organism thought to exhibit some of the earliest hallmarks of bilateral symmetry (**Figures 2(c)** and **3(g)**). Multiple lines of evidence support the conclusion that netrin-1 attracts commissural axons toward and repels subsets of motor neuron axons away from the midline. Expression of a netrin-1 ortholog at the midline early in neural development is highly conserved (**Figure 3**). Ectopic expression of netrin-1 is sufficient to alter axon extension. Furthermore, netrin-1 ejected from a micropipette functions as a chemoattractant or chemorepellent, depending on the axons stimulated. Genetic disruption of the netrin UNC-6 in *C. elegans* disrupts circumferential axon extension toward and away from the ventral midline. Loss of netrin-1 function in mice generates major disruptions in multiple axon commissures, including the spinal ventral commissure, the corpus callosum, and the hippocampal commissure.

Netrin-1 also influences axon guidance away from the midline. It is required for retinal ganglion cell axons to exit the retina, for dopaminergic axon guidance in the ventral midbrain, and for thalamocortical projections. Netrin-1 also directs neural precursor cell migration, attracting cells that will become inferior olivary, pontine, luteinizing hormone-releasing hormone (LHRH), antidiuretic hormone (ADH), and oxytocin neurons, and repelling striatal neuronal precursors, cerebellar granule cells, spinal accessory neurons, and oligodendrocyte precursor cells. Although netrin-3 can mimic the ability of netrin-1 to attract spinal commissural axons and repel trochlear motor neuron axons *in vitro,* its expression in the spinal cord begins after the initial commissural axons have pioneered the path to the floor plate. Netrin-3 may, however, influence guidance of dorsal root ganglia axons to peripheral targets in the developing peripheral nervous system (PNS).

Very little is known regarding the role of netrin-4 in the development of the nervous system. It is widely expressed in the developing olfactory bulb, retina, and dorsal root ganglia, as well as by cerebellar granule, hippocampal, and cortical neurons. A relatively low level of netrin-4 expression has been detected adjacent to floor plate cells in the developing spinal cord, but only after the first commissural axons have crossed the midline. As discussed later, netrin-Gs do not appear to have a major role in the outgrowth and guidance of axons. They are expressed primarily by neurons, with very limited expression outside the nervous system. Netrin-G1 is expressed in the dorsal thalamus, olfactory bulb, and inferior colliculus, whereas netrin-G2 is expressed in the cerebral cortex.

Netrin-Induced Signal Transduction

Research investigating the signaling mechanisms elicited by netrins has focused on netrin-1, and there is little known about signaling by other netrins. Here we review current insight into netrin-1 signal transduction during axon guidance. The DCC ('deleted in colorectal cancer') family and the UNC5 homolog family are well-established receptors for netrin-1. In vertebrate species, these include DCC and neogenin, as well as UNC5A–UNCD. Both classes of receptors are single-pass transmembrane proteins that are also members of the large immunoglobulin superfamily.

Chemoattractant Signaling in Response to Netrin-1

DCC is required for axon chemoattraction to netrin-1. Although initially identified in vertebrates as a potential tumor suppressor, the importance of DCC in axon guidance was realized shortly after the discovery of netrin-1 as a result of mutational studies done in the roundworm *C. elegans*. While mutation of the *netrin* homolog *unc-6* affects both ventrally and dorsally directed neurons, mutations of the *C. elegans* homolog of *dcc*, *unc-40*, primarily disrupt ventrally directed axons. It was quickly realized that this phenotype was consistent with a potential role for UNC-40/DCC during axon attraction to netrins/UNC-6. Loss of DCC function in mice was then shown to cause neurodevelopmental defects very similar to loss of netrin-1 function, including disruption of the spinal ventral commissure, corpus callosum, and hippocampal commissure.

Structurally, the extracellular domain of DCC contains six fibronectin type 3 (FN3) repeats and four immunoglobulin (Ig) repeats. Netrin-1 appears to bind to the FN3 repeats of DCC, though reports conflict as to the exact domain involved. The DCC intracellular domain has no known catalytic activity; rather, it contains several putative phosphorylation

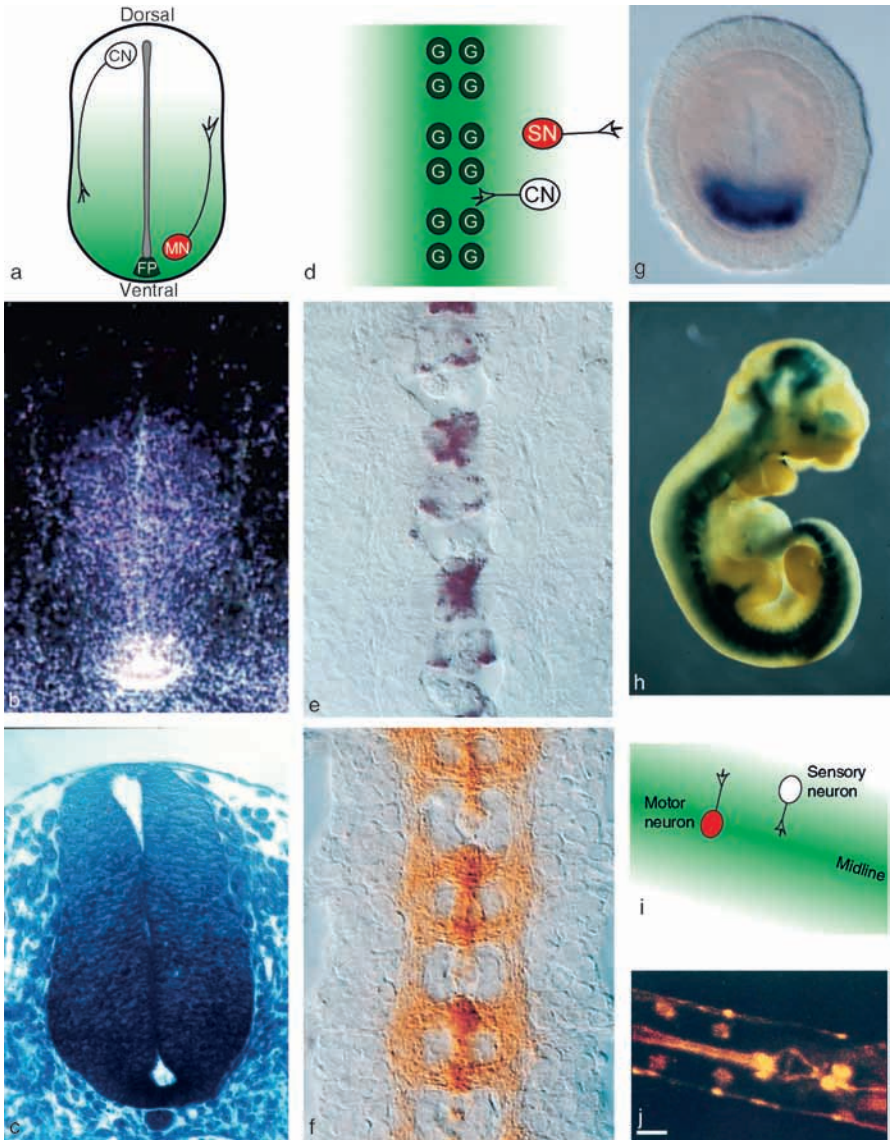

Figure 3 Midline expression of netrin homologs in a variety of organisms. The diagrams in panels (a), (d), and (i) illustrate simplified models of the distributions of netrin protein at the midline in the developing mouse spinal cord, fruit fly and *C. elegans* nematode worm, respectively. In each case, sensory neurons extend toward, while motor neurons extend away from, the source of netrin protein at the midline. *In situ* hybridization illustrates floor plate cells expressing netrin-1 in the embryonic day 9.5 mouse spinal cord (b). Whole-mount staining for expression of a β-galactosidase reporter gene shows netrin-1 expression in an entire E12.5-day-old mouse embryo, illustrating netrin-1 expression along the full rostrocaudal extent of the spinal cord, the developing brain, and the peripheral nervous system (h). *In situ* hybridization reveals netrin expression in the fruit fly (e) and the stellate sea anemme *N. vectensis* (g) during neural development. The distribution of netrin protein in the embryonic day 9.5 mouse spinal cord (c), as well as fruit fly (f) and *C. elegans* (j) embryos are also shown. CN, commissural neuron; MN, motoneuron; SN, segmental nerve. (b) Reprinted from Serafini T, Colamarino SA, Leonardo ED, et al. (1996) Netrin-1 is required for commissural axon guidance in the developing vertebrate nervous system. *Cell* 87: 1001–1014, with permission. (e, f) Reprinted from Harris R, Sabatelli LM, and Seeger MA (1996) Guidance cues at the *Drosophila* CNS midline: Identification and characterization of two *Drosophila* netrin/UNC-6 homologs. *Neuron* 17: 217–228, with permission. (g) Reprinted from Matus DQ, Pang K, Marlow H, et al. (2006) Deep evolutionary roots for bilaterality in the metazoa. *Proceedings of the National Academy of Sciences of the United States of America* 103: 11195–11200, with permission. (j) Reprinted from Wadsworth WG, Bhatt H, and Hedgecock EM (1996) Neuroglia and pioneer neurons express UNC-6 to provide global and local netrin cues for guiding migrations in *C. elegans*. *Neuron* 16: 35–46, with permission.

and binding sites for intracellular proteins. Based on particularly strong evolutionary conservation, three domains (P1, P2, and P3) within the intracellular domain of DCC have been described. The P2 domain is rich in proline residues, containing four PXXP putative SH3 domain-binding motifs, while the P3 domain contains several highly conserved possible phosphorylation sites.

Numerous studies have now described signal transduction events implicated in the response to netrin-1, but our understanding of the mechanisms underlying how a growth cone responds to netrin-1 as a chemoattractant remains fragmentary. Netrin-1-induced multimerization of DCC via its P3 intracellular domain is thought to be essential for chemoattraction. Netrin-1 binding DCC activates the Rho GTPases Rac1 and Cdc42, key intracellular coordinators of cytoskeletal and adhesive interactions. Furthermore, intracellular proteins reported to associate with the intracellular domain of DCC include the adapter protein Nck1, the tyrosine kinases Fak and Fyn, the actin-binding proteins Ena/Vasp and N-Wasp, and the serine/threonine kinase Pak, a downstream effector of Rac1 and Cdc42 activation. Netrin-1 binding DCC also triggers generation of phosphoinositides and their breakdown by phospholipase C into inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol, leading to calcium release from intracellular stores and activation of protein kinase C. Figure 4(a) outlines a speculative model of the signal transduction events that occur during chemoattraction to netrin-1.

Chemorepellent Signaling in Response to Netrin-1

Axonal repulsion in response to netrin-1 requires a member of the UNC5 protein family. Four family members are present in mammals: UNC5A, UNC5B, UNC5C, and UNC5D. In *C. elegans*, in contrast to mutation of the *dcc* homolog *unc-40* that affects ventrally extending axons, mutation of *unc-5* predominantly causes defects in dorsally directed axons (i.e., those extending away from the source of UNC-6 netrin). Misexpression of *unc-5* is sufficient to redirect them dorsally, away from the ventral source of UNC-6. The extracellular domain of UNC5 homologs contains two membrane-proximal thrombospondin repeats and distally two immunoglobulin repeats. Netrin-1 binds to both immunoglobulin repeats. The intracellular domain of UNC5 contains three identified conserved domains: a ZU5 domain, a DCC-binding (DB) domain, and a death domain (DD). Although the specific function of the ZU5 domain is not known, a homologous domain is present in the scaffolding protein zona occludens-1 found at tight junctions, and deletion of the UNC5 ZU5 domain in *Drosophila* disrupts UNC5-mediated chemorepulsion.

Interestingly, depending on the distance of the growth cone from the source of netrin, different signaling mechanisms appear to be engaged. Long-range repulsion to netrin requires expression of both UNC5 and DCC, whereas short-range repulsion requires UNC5, but DCC is not essential. One hypothesis for this difference is that DCC and UNC5 together may form a more sensitive netrin receptor complex that is able to respond to the lower concentrations of netrin present at greater distances. At long range, the cytoplasmic domains of UNC5 and DCC associate directly. At short range, genetic studies in *C. elegans* have stressed the importance of an association between UNC5 cytoplasmic ZU5 and DD domains. Several proteins that interact with the UNC5 intracellular domain have been identified. These include the tyrosine kinase Src1, the tyrosine phosphatase Shp2, the F-actin anti-capping protein Mena, the structural protein ankyrin, and the adapter protein Max1. Netrin-1-mediated growth cone repulsion triggers tyrosine phosphorylation of UNC5's intracellular domains at multiple sites. A speculative model outlining intracellular events that occur during short- and long-range netrin-mediated growth cone repulsion is illustrated in Figure 4(b).

Regulating Growth Cone Response to Netrin-1

Single growth cones have been shown to have the capacity to rapidly switch between responding to netrin as an attractant or a repellent. The mechanisms that control this shift in responsiveness have been the topic of intense investigation.

Altering UNC5 or DCC expression is one mechanism that influences how a growth cone responds to netrin. For example, the homeobox transcription factor 'even-skipped' promotes *unc5* expression in the fruit fly *Drosophila melanogaster*. Local protein synthesis within the growth cone also appears to contribute to the attraction of axons to netrin-1, though the exact mechanism of action is not clear. DCC expression can be downregulated through proteolytic degradation by either extracellular metalloproteinases or intracellularly, being ubiquitinated by an interaction with Siah-1, a RING domain-containing protein. The intracellular domains of UNC5 proteins are substrates for proteolysis by caspases.

The concentration of cyclic nucleotides within an axonal growth cone has a profound influence on how it responds to guidance cues. In particular, the intracellular concentration of cyclic adenosine monophosphate (cAMP), which activates protein kinase A (PKA), can regulate the response to netrin-1. High levels of cAMP are associated with attractant responses to netrin-1, while growth cones with low concentrations are repelled. Although cAMP and PKA activation may regulate intracellular signal transduction, a specific mechanism regulating the direction of the response made by a growth cone has not been identified. It has been shown that activating PKA leads to the recruitment of DCC from an intracellular pool of vesicles to the growth cone plasma membrane, thereby enhancing chemoattraction to netrin-1. In addition, activating protein kinase C (PKC) leads to endocytosis of UNC5 proteins inducing neurons to

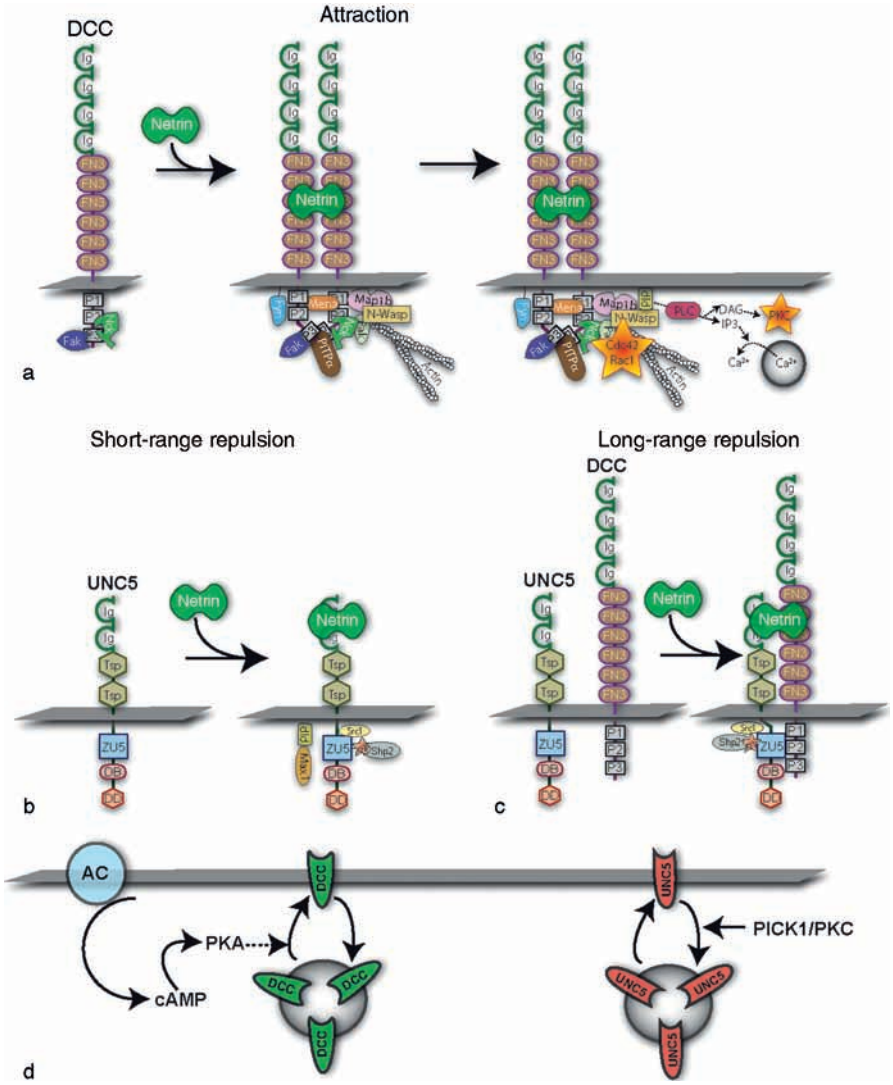

Figure 4 Speculative models outlining chemoattractive, chemorepellent, and modulatory signaling in response to netrin-1. (a) Chemoattraction to netrin-1 can be divided into three conceptually different stages: in the absence of netrin-1, during the initial response to netrin-1, and during cytoskeletal remodeling triggered by netrin-1. The adapter protein Nck1 and the tyrosine kinase Fak associate with the intracellular domain of the DCC ('deleted in colorectal cancer') protein in the absence of netrin-1. Upon netrin-1 binding, DCC multimerizes through association of its P3 domains. Phosphatidylinositol transfer protein-α (PITPα) can also bind the P3 domain of DCC and promote the generation of phosphoinositides (PIPs) by phosphatidylinositol 3-kinases (PI3Ks). PIPS can then be hydrolyzed by phospholipase C (PLC) into inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol (DAG), leading to Ca^{2+} release from intracellular stores and activation of protein kinase C (PKC). Netrin-1 binding to DCC leads to phosphorylation of the intracellular domain of DCC and association of proteins such as Fak, Fyn, and Pak. Intracellular Ca^{2+} increases can lead to activation of the Rho GTPases Cdc42 and Rac, and cause remodeling of the cytoskeleton through proteins such as the Wiskott–Aldrich syndrome protein (N-Wasp), Ena/Vasp, and Map1b. (b) Short- and long-range repellent signaling to netrin-1. In the absence of netrin-1, Mena and ankyrin link UNC5 to the cytoskeleton. Upon binding netrin-1, UNC5 is tyrosine-phosphorylated, independently of DCC, by tyrosine kinases such as Src1. Netrin-1 induces recruitment of the tyrosine phosphatase Shp2 to a phosphorylated tyrosine residue in the ZU5 domain. PIPs have been proposed to regulate interaction of Max-1 with UNC5. (c) Long-range repulsion to netrin-1 requires association between the intracellular domains of UNC5 and DCC. (d) Both DCC and UNC5 can traffic between intracellular vesicular pools and the cell surface. Protein kinase A (PKA) can recruit DCC to the plasma membrane. PKC induces endocytosis of UNC5 from the cell surface. AC, adenylate cyclase; cAMP, cyclic adenosine monophosphate; PICK1, protein interacting with C kinase-1.

switch their response to netrin-1 from repulsion to attraction. These findings suggest that neuromodulatory factors that regulate PKA or PKC may influence axon outgrowth by altering which netrin receptors are presented by the growth cone.

Other Potential Netrin Receptors

Receptors for secreted netrins, other than DCC, neogenin, and UNC5 proteins, have been suggested. The G-protein-coupled adenosine A2B receptor has been

proposed as a receptor for netrin-1. Through an influence on intracellular cAMP concentration, it is possible that A2B may influence the response to netrin-1; however, contrary to an early report, A2B is neither expressed by nor required for spinal commissural axon guidance. Integrins, best known for their role as receptors for extracellular matrix components, have also been implicated as netrin receptors. Specifically, an adhesive interaction between netrin-1 and integrins $\alpha 6\beta 4$ and $\alpha 3\beta 1$ has been suggested to contribute during development of the pancreas; however, evidence of this *in vivo* remains to be demonstrated. Interestingly, the region of netrin reported to interact with these integrins is not in domains V and VI that are homologous to laminins, which are ligands for integrins. Rather, $\alpha 6\beta 3$ and $\alpha 3\beta 1$ integrins bind to a highly charged sequence of basic amino acids at the C-terminus of netrin-1.

Netrin Expression in the Mature Nervous System

Roles for netrins beyond directing axon and cell migration during development are beginning to be identified. Netrins and their known receptors are expressed in the adult nervous system: netrin-1 is expressed by neurons, Schwann cells in the PNS, and oligodendrocytes in the CNS. Subcellular fractionation of mature CNS white matter has determined that netrin-1 protein is enriched at the interface – known as periaxonal myelin – between axons and oligodendrocytes, suggesting that a function for netrin-1 in the mature CNS may be to regulate axon oligodendroglial interactions. Interestingly, as the mammalian spinal cord matures, DCC expression falls while UNC5 homolog expression increases. This suggests that UNC5 repellent signaling may be the dominant mode of responsiveness to netrin in the adult spinal cord. Although the functional significance of netrin-1 expression in the adult CNS remains unknown, an intriguing hypothesis is that netrins contribute to maintaining appropriate connections in the intact CNS by restraining inappropriate axonal sprouting.

The expression of netrin-1 by myelinating oligodendrocytes raises the possibility that it might function as a myelin-associated inhibitor of axon growth following injury. Netrin-1 does not appear to be a major component of an injury-induced glial scar in the mature spinal cord, but essentially normal netrin-1 expression persists on either side of the injury site. Although an influence of netrin-1 on axon regeneration in the adult CNS has not been demonstrated directly, these findings suggest that netrin-1 may be a component of CNS myelin that inhibits axon regeneration by neurons expressing UNC5 following injury. Such a role may explain, in part, why increasing cAMP in neurons promotes the ability of axons to grow in the adult mammalian CNS, as increasing cAMP recruits DCC to plasma membranes of growth cones and converts netrin-mediated repulsion to attraction. Interestingly, studies carried out in lamprey, a primitive vertebrate with the capacity for substantial axon regeneration following spinal lesion, correlated poor axonal regeneration following lesion with neuronal expression of UNC5 protein.

Netrin-3, netrin-4, netrin-G1, and netrin-G2 are also expressed in the adult brain. In humans, mutation of the netrin-G1 gene is a rare cause of the childhood neurodevelopmental disorder known as Rett syndrome. This disease is characterized by normal early development followed by loss of purposeful use of the hands, distinctive hand movements, slowed brain and head growth, gait abnormalities, seizures, and mental retardation. In mice, netrin-G1 deficiency does not lead to any obvious changes in neural circuitry, but does lead to altered synaptic responses and defects in sensorimotor gating behavior. Similarly, the netrin-G2 receptor NGL-2 influences the formation of glutamanergic synapses through an interaction with the postsynaptic scaffold protein PSD-95. Together, these findings indicate that netrin-G proteins have a role in the maturation, refinement, and maintenance of synapses, rather than in the guidance of axons.

Netrin Outside of the Nervous System

Netrins are expressed in many tissues. Netrin-1 is expressed in the heart, tongue, lung, inner ear, intestine, mammary gland, and pancreas, netrin-3 is expressed in the bowel, pancreas, and muscle, and netrin-4 is expressed in the intestine, pancreas, spleen, vascular networks, kidney, ovaries, and lung. Functional roles for netrins have been demonstrated in several developing tissues. Netrin-1 has been implicated in vascular patterning, although some disagreement remains regarding whether it functions principally as a repellent, acting via UNC5B, or an attractant, acting via undefined receptors, or both (in different vascular beds). Both netrin-1 and netrin-4 shape the developing lung through an influence on branching of the epithelial endoderm. In the developing mammary gland, netrin-1 is expressed by a layer of luminal cells, and an interaction with neogenin is required for proper organization of the terminal end buds.

Conclusion

Homologs of netrin-1 are evolutionarily conserved guidance cues that function as chemoattractants and

chemorepellents that direct cell and axon migration in the developing nervous system. The receptor DCC is essential for chemoattraction to netrin-1, and UNC5 proteins are required for chemorepulsion. The netrin-G subfamily contributes to the maturation, refinement, and maintenance of synapses. In addition to these roles in the CNS, netrins also influence the development of a variety of other tissues.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Pathfinding: Extracellular Matrix Role; Cetacean Brains; Growth Cones; Rett Syndrome.

Further Reading

Barallobre MJ, Pascual M, Del Rio JA, et al. (2005) The Netrin family of guidance factors: Emphasis on Netrin-1 signalling. *Brain Research – Brain Research Reviews* 49: 22–47.

Dickson BJ (2002) Molecular mechanisms of axon guidance. *Science* 298: 1959–1964.

Fazeli A, Dickinson SL, Hermiston ML, et al. (1997) Phenotype of mice lacking functional Deleted in colorectal cancer (Dcc) gene. *Nature* 386: 796–804.

Harris R, Sabatelli LM, and Seeger MA (1996) Guidance cues at the *Drosophila* CNS midline: Identification and characterization of two *Drosophila* netrin/UNC-6 homologs. *Neuron* 17: 217–228.

Hedgecock EM, Culotti JG, and Hall DH (1990) The *unc-5*, *unc-6*, and *unc-40* genes guide circumferential migrations of pioneer axons and mesodermal cells on the epidermis in *C. elegans*. *Neuron* 4: 61–85.

Hinck L (2004) The versatile roles of 'axon guidance' cues in tissue morphogenesis. *Developmental Cell* 7: 783–793.

Huber AB, Kolodkin AL, Ginty DD, et al. (2003) Signaling at the growth cone: Ligand–receptor complexes and the control of axon growth and guidance. *Annual Review of Neuroscience* 26: 509–563.

Ishii N, Wadsworth WG, Stern BD, et al. (1992) UNC-6, a laminin-related protein, guides cell and pioneer axon migrations in *C. elegans*. *Neuron* 9: 873–881.

Keleman K and Dickson BJ (2001) Short- and long-range repulsion by the *Drosophila* Unc5 netrin receptor. *Neuron* 32: 605–617.

Kennedy TE, Serafini T, de la Torre JR, et al. (1994) Netrins are diffusible chemotropic factors for commissural axons in the embryonic spinal cord. *Cell* 78: 425–435.

Kennedy TE, Wang H, Marshall W, et al. (2006) Axon guidance by diffusible chemoattractants: A gradient of netrin protein in the developing spinal cord. *Journal of Neuroscience* 26(34): 8866–8874.

Matus DQ, Pang K, Marlow H, et al. (2006) Deep evolutionary roots for bilaterality in the metazoa. *Proceeding of the National Academy of Sciences of the United States of America* 103: 11195–11200.

Moore SW, Kennedy TE, and Tessier-Lavigne M (2007) Netrins and their receptors. In: Bagnard D (ed.) *Axon Growth and Guidance*. Austin, TX: Landes Biosciences.

Ramón y Cajal S (1899) *Texture of the Nervous System of Man and the Vertebrates*. New York: Springer.

Serafini T, Colamarino SA, Leonardo ED, et al. (1996) Netrin-1 is required for commissural axon guidance in the developing vertebrate nervous system. *Cell* 87: 1001–1014.

Serafini T, Kennedy TE, Galko MJ, et al. (1994) The netrins define a family of axon outgrowth-promoting proteins homologous to *C. elegans* UNC-6. *Cell* 78: 409–424.

Tessier-Lavigne M, Placzek M, Lumsden AG, et al. (1988) Chemotropic guidance of developing axons in the mammalian central nervous system. *Nature* 336: 775–778.

Wadsworth WG, Bhatt H, and Hedgecock EM (1996) Neuroglia and pioneer neurons express UNC-6 to provide global and local netrin cues for guiding migrations in *C. elegans*. *Neuron* 16: 35–46.

Axonal Regeneration: Role of Growth and Guidance Cues

M Bähr and P Lingor, Georg-August-University Göttingen, Göttingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Compared to other organ systems, the central nervous system (CNS) in the adult human has one of the lowest regenerative capacities following pathological insults, and within the phylogenetic tree the mammalian nervous system is the least capable to renew itself. Besides traumatic nerve injury, a great number of neurodegenerative diseases exhibit marked axonal pathologies, which often precede the histologically more impressive neuron cell death and possibly represent an early stage in disease progression. A functional restoration in a structure as complex as the nervous system therefore requires not only the preservation of cell bodies, but also the maintenance and/or reconstruction of axonal connections to their physiological target regions.

Increased life expectancy has resulted in a continuous rise in the incidence and prevalence of neurodegenerative diseases. The understanding of basic cellular and molecular mechanisms underlying the lack of CNS regeneration is thus not only of interest for the life science community but also of major socioeconomic relevance. Clearly, a deeper insight into mechanisms of axonal de- and regeneration is indispensable for the identification of future treatment strategies in the field of clinical neurosciences.

In this article, developmental mechanisms of axon guidance which regulate axonal pathfinding during embryogenesis are discussed. Interestingly, numerous molecules with developmental roles persist throughout adulthood and may take new functions in the lesioned nervous system, inhibiting or fostering axonal regeneration.

Developmental Guidance Cues: Orchestrating a Symphony of a Thousand

During development, a multitude of signals are required in order to shape naive cells to form adult tissue. Molecular gradients of the so-called morphogens have been identified to regulate cell fate by the induction of transcriptional changes, which, for example, results in cell-type specification according to a cell's position in the gradient. Cellular processes, such as axons, similarly react to molecular gradients of guidance molecules, which act to direct motile structures toward a target area. Four major groups of largely conserved developmental guidance molecules are known to date: ephrins, semaphorins, netrins, and slits. Recent evidence, however, suggests that morphogens are able to route growth cones in a similar way, in addition to their classical function in the commitment of cell fate. The versatile calcium ion influences growth cone development in a precise spatiotemporal manner, including the mediation of pathway cross-talk.

Classical Morphogens with Novel Old Functions

Morphogens are signaling molecules that during development have a locally confined expression in a specific source region from where a concentration gradient is established. The exact mechanism of gradient formation is still a matter of debate, and evidence supports diffusive mechanisms as well as repeated cycles of endo- and exocytosis. Cells positioned in morphogen gradients react to concentration differences by modification of their differentiation program inducing cell fate specification. In addition to this classical morphogen role, evidence is growing for morphogens as axon guidance molecules (Figure 1).

Sonic hedgehog (Shh) is a member of the hedgehog protein family of which two other members are known in mammals. It is produced in the notochord and in the ventral midline of the CNS – the floor plate cells, where it is involved in shaping the dorsoventral axis of the neural tube. Signaling through its receptor patched (Ptc), Shh induces the release of the G-protein-associated mediator smoothened (Smo), which eventually activates zinc finger transcription factors of the Ci/GLI family.

As a midline-derived molecule, Shh has been shown to function as a chemoattractant for commissural neurons, guiding their axons toward the midline. This could be inhibited by the inactivation of the Shh signaling mediator Smo, suggesting a direct action of Shh on the axons rather than the induction of secondary guidance cues. In contrast to commissural axons of the spinal cord, retinal ganglion cells (RGCs) are prevented from crossing the midline by continuous expression of Shh. Here, Shh signaling has been shown to induce an inhibition of cyclic AMP (cAMP) production in the cell, which is known to result in growth inhibition. The opposing effects of the same guidance cue on axonal growth of different cell types remain puzzling. It is likely to assume different signaling pathways resulting in diametrical effects or an interplay with other guidance cues present at the same time. A context-specific effect of Shh, for example, dependent on intracellular cAMP levels, is also proposed.

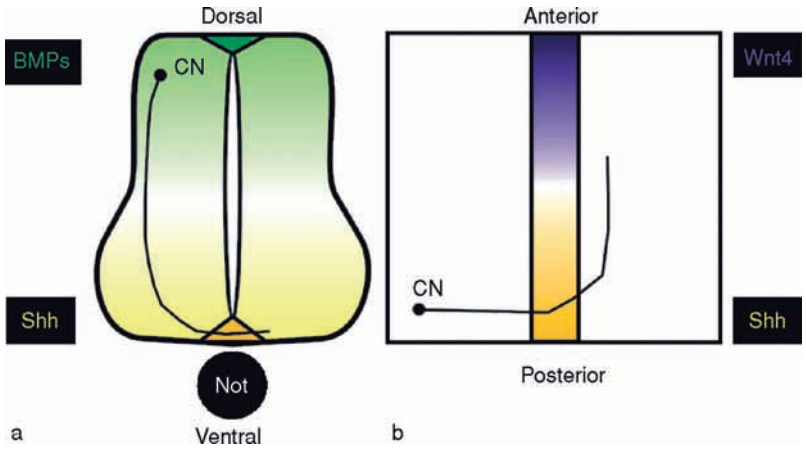

Figure 1 Morphogens involved in commissural axon guidance during embryonic development. Gradients of bone morphogenetic proteins (BMPs) and sonic hedgehog (Shh) shape the dorsoventral axis, while gradients of Shh and Wnt4 regulate the posteroanterior guidance. Cross-section of developing spinal cord (a) and 'open-book' conformation (b). CN, commissural neuron.

Similar to the regulation of the fate of ventral neurons by Shh, members of the bone morphogenetic protein (BMP) family regulate the fate of dorsal interneurons and commissural neurons. BMPs as well as the related growth and differentiation factors (GDFs) belong to the large superfamily of transforming growth factors-β (TGF-β) and signal through type I and type II TGF-β receptors. Through receptor-regulated SMADs and co-SMADs, the signal is transduced to the nucleus, where transcription is activated. BMPs have been shown to equally act as guidance cues for commissural axons after their initial role as morphogens. Heterodimers of GDF7 and BMP7 guide axon growth of commissural neurons, which could be shown by defective outgrowth of commissural axons in knockout mice lacking GDF7, BMP7, or both. A role for BMPs in the development of RGC axons has been suggested by studies in which a deletion of the BMP receptor Ib resulted in misrouting of ventral RGC axons, enabling them to enter the optic nerve head.

The Wingless/Wnt family of morphogens is expressed by roof-plate cells and participates in the developmental specification of dorsal interneurons. Binding to the Frizzled (Fz) receptors is a common denominator for members of the Wnt family, but signaling pathways that are employed downstream of Fz can be diverse: Wnt binding to the Fz receptor and its coreceptor LRP5/6 eventually results in the inhibition of glycogen synthase kinase 3β (GSK3β). In the absence of Wnt signaling, GSK3β phosphorylates β-catenin, which is then ubiquitinated and targeted for degradation by the proteasome. In this so-called canonical pathway, Wnt signaling thus induces stabilization of β-catenin. Stabilized β-catenin induces transcriptional activation of Wnt target genes via association with the lymphoid enhancer factor (Lef/Tcf). Additionally, GSK3β phosphorylates microtubule-associated proteins (MAPs), and Wnt signaling therefore may directly alter the cytoskeletal structure. Other Wnt signaling pathways involve c-Jun-N-terminal kinase (JNK) activation (the so-called planar cell polarity pathway), directing cytoskeletal organization and coordinated cell polarization as well as the release of intracellular calcium and consequent calcineurin activation (Wnt/Ca^{2+} pathway).

Similar to other morphogens, Wnt family members can act as guidance cues for developing neurons. For example, Wnt5 has been shown to repel anterior commissural axons in *Drosophila*. In mouse embryos, Wnt4 acts as a chemoattractant for postcrossing commissural axons. Again, molecules from the same family show opposing effects on axon outgrowth, which may be explained by utilization of different signaling pathways or, as in the case of Wnt5 signaling in *Drosophila*, even the employment of the unconventional receptor Derailed (Drl).

Neurotrophins

As early as in 1939 limb bud transplantation experiments performed by V Hamburger in chick embryos suggested the existence of a target-derived soluble factor that guides the developing axons of sensory neurons. The protein, now known as nerve growth factor (NGF), was later identified by S. Cohen and R. Levi-Montalcini and turned out to serve as growth and survival factor for sympathetic neurons. Besides NGF, the family of neurotrophins – named after their survival and growth-promoting effects on neurons – today comprises three other structurally related proteins in mammals: brain-derived neurotrophic factor (BDNF), neurotrophin 3, and neurotrophin 4/5. All neurotrophins share a common signaling mechanism involving binding to receptor tyrosine

kinases (trk) and the so-called low-affinity neurotrophin receptor p75NTR.

Upon neurotrophin binding, the appropriate trk receptor dimerizes and induces the activation of multiple signaling pathways, of which protein kinase A (PKA), phospholipase C-γ (PLC-γ), and phosphatidylinositol-3-kinase (PI3K) are the most important ones. Neurite elongation and axon guidance mediated by neurotrophins are likely to be transduced via the actin depolymerizing factor (ADF)/cofilin pathway and consequent f-actin rearrangement (Figure 2).

NGF is able to precisely control neurite outgrowth via local action on distinct cellular segments. Targets of sympathetic and sensory neurons secrete NGF and thus a chemoattractant function of the molecule was suggested. Later, NGF was shown to foster sensory axon elongation and arborization. In culture, NGF induces outgrowth of dorsal root ganglion (DRG) neurons.

Very similar to NGF, other members of the neurotrophin family have effects on neuronal growth, survival, and axon guidance. BDNF, for example, is able to promote the elongation of RGC axons *in vivo*. Motor neurons show an expression of all neurotrophin receptors during development, and this coincides with the expression of all neurotrophins in their muscular target tissue, promoting the hypothesis of neurotrophins as target-derived guidance cues for developing motor neurons. Interestingly, neurotrophins may act in a chameleon-like manner, changing their chemoattractant properties into a repulsive character, depending on the preconditioning of the neuron: NT-3 acted as a chemoattractant for DRG neurons, except when the culture was pretreated with NT-4/5. Preconditioning of DRG cultures with NT-3 rendered them unsusceptible toward NGF-mediated attraction. Furthermore, growth cones of DRG neurons pretreated with NGF were collapsed by local application of BDNF. The activation and cross-talk of different intracellular signaling pathways as well as a regulation of receptor expression in preconditioned cultures may be discussed as a reason for the differential effects mentioned.

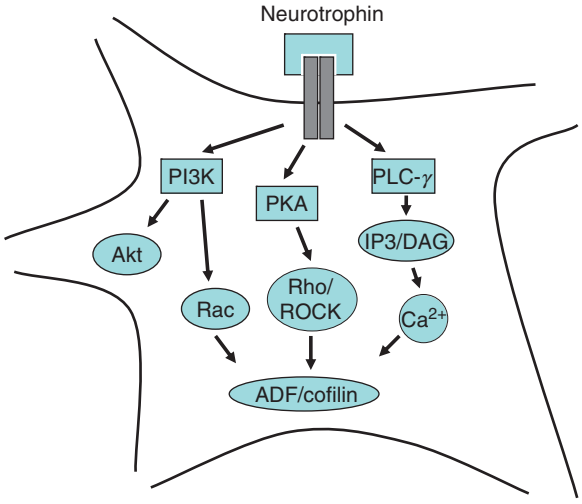

Figure 2 Signaling pathways of neurotrophins involved in axon guidance. Neurotrophin binding to the trk receptor results in activation of the phosphatidylinositol-3-kinase (PI3K), protein kinase A (PKA) or phospholipase C (PLC). Signaling through intermediate molecules results in decreased phosphorylation and thus activation of actin depolymerizing factor (ADF)/cofilin inducing neurite growth.

Besides the neurotrophin family, many other growth factors may act as guidance cues, of which only two examples are given: glial cell line-derived neurotrophic factor (GDNF), one of the most potent survival factors for dopaminergic neurons, has chemoattractive functions on peripheral neurons via activation of cyclin-dependent kinase 5 (Cdk5). Studies in rat explant cultures identified hepatocyte growth factor/scatter factor (HGF/SF) as a mesenchyme-derived chemoattractant for developing motor neurons. In addition to its function as survival factor, HGF/SF thus was able to promote axon growth into the limb bud after binding to the c-Met trk.

Ephrins

A model system for the study of axonal guidance cues in the developmental period is the retino-tectal projection. Axons of RGCs follow a precise topographical projection pattern innervating their target areas in the tectum: RGCs in the dorso-ventral axis terminate along the medio-lateral axis of the tectum, while the temporo-nasal distribution in the retina finds its tectal representation in the antero-posterior axis. The involvement of a chemical gradient derived from the target tissue which regulates the targeted outgrowth of RGC axons was suggested by stripe explant experiments. RGC axons originating from the temporal half of the retina would preferentially grow into stripes derived from the rostral part of the tectum, while stripes derived from the caudal colliculus induced growth cone collapse. A similar preference was observed for RGC axons from the dorsal retinal part, which preferred the lateral tectum stripes and vice versa. When the stripes were treated with phosphatidylinositol-specific phospholipase C (PI-PLC) this effect was no longer observed, suggesting that the guidance molecule is membrane bound via a GPI anchor. The guidance molecules responsible for this part of retinal pathfinding are now known as ephrins: two of them, ephrinA5 and ephrinA2, play a pivotal role in the establishment of the topographical map in the retinotectal projection. Studies with knockout mice for either ephrinA5 or ephrinA2 showed a

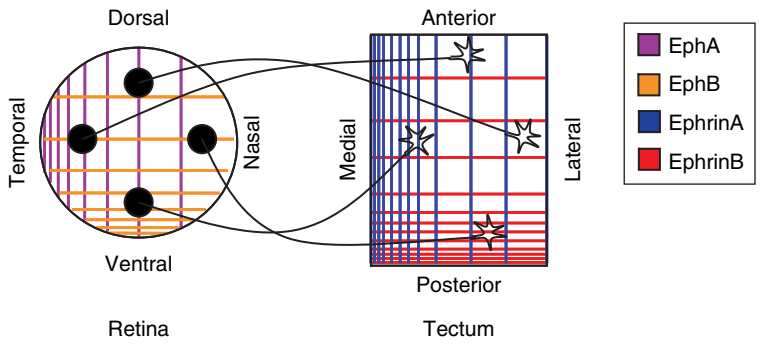

Figure 3 Establishment of retinotectal projections by gradients of Eph receptors and ephrins. Temporonasal gradients of EphA and dorsoventral gradients of EphB define RGCs projecting to the tectum. Anteroposterior gradients of ephrinB and mediolateral gradients of ephrinA result in a retinotopical termination of RGC axons in the tectum.

disorganization of the retinotectal projection, while the double knockout shows an even more prounounced disturbance of axonal wiring. Two families of ephrins are known to date: the GPI-anchored ephrinAs and the transmembrane domain-linked ephrinBs. Trks, termed EphA and EphB, respectively, are responsible for the mediation of the Ephrin signal. It is now confirmed that ephrinA/EphA gradients establish the antero-posterior tectal architecture, while ephrinB/EphB gradients determine the mediolateral projections (Figure 3).

Semaphorins

The large family of semaphorins comprises proteins that are either secreted or membrane-bound and that were initially identified for their ability to collapse growth cones of cultured neurons. Up to now, more than 20 semaphorins have been identified and classified into eight subclasses. In invertebrates, subclasses 1 and 2 are expressed, while vertebrates express subclasses 3–7. Nonneurotropic DNA viruses encode for V class semaphorins. All semaphorins are characterized by the so-called sema domain, which is located in the N-terminal region. Semaphorins of classes 3 and 4 homodimerize, which appears to be required for their function.

Neuropilins were the first receptors identified to bind class 3 semaphorins, and they establish target selectivity by formation of multimeric complexes. As is now known, neuropilins bind only to the secreted class 3 semaphorins. However, because of the shortness of their cytoplasmic domain, neuropilins alone are not capable of transducing the signal and therefore form complexes with another type of semaphorin receptors – the plexins. Four subfamilies of plexins have been identified in mammals, termed plexin-A through plexin-D. In contrast to neuropilins, plexins do not need to heteromerize for signal transduction. In their monomeric form plexins represent the receptors for membrane-bound semaphorins of classes 1, 4, 5, 6, and 7.

Additionally, the transmembrane cell adhesion molecule L1 has also been shown to form a complex with neuropilin-1 and act as a semaphorin receptor, modulating the effects of Sema3A: in its soluble form, L1 is able to block the Sema3A-induced growth cone collapse, while L1-deficient mice do not respond to Sema3A-mediated repulsion (Figure 4).

Netrins

In the search for factors responsible for the pathfinding of commissural axons in the spinal cord, the biochemical purification revealed two largely homologous proteins – Netrin-1 and -2. To date, Netrins 1–4, which are secreted, and Netrin-G, which is GPI-anchored, are known to exist in vertebrates. Mammalian netrins show homologies to the *Caenorhabditis elegans* UNC-6 protein, which in turn is related to laminin. Two protein families, DCC (deleted in colon cancer) and UNC-5, give rise to netrin receptors.

Netrins seem to have specialized in modeling bilateral symmetry via either chemoattraction (via DCC) or repulsion (via UNC-5). Netrin-1 stimulates outgrowth of commissural axons at different levels in the CNS, and *netrin-1*-deficient mice show severe defects in the establishment of the forebrain commissures. In addition, Netrin-1 is required for developmental pathfinding of RGC axons as they enter the optic nerve, and genetic ablation of Netrin-1 leads to optic nerve hypoplasia. Thalamocortical and corticospinal as well as hippocampal and cerebellar projections equally depend on Netrin-1 signaling for their proper development. One of the intriguing netrin features is its ability to change from attractant to repellent depending on developmental state and intracellular cAMP levels.

Slits

Similar to the midline-forming activity of netrins, the secreted and chemorepellent slits are involved in the

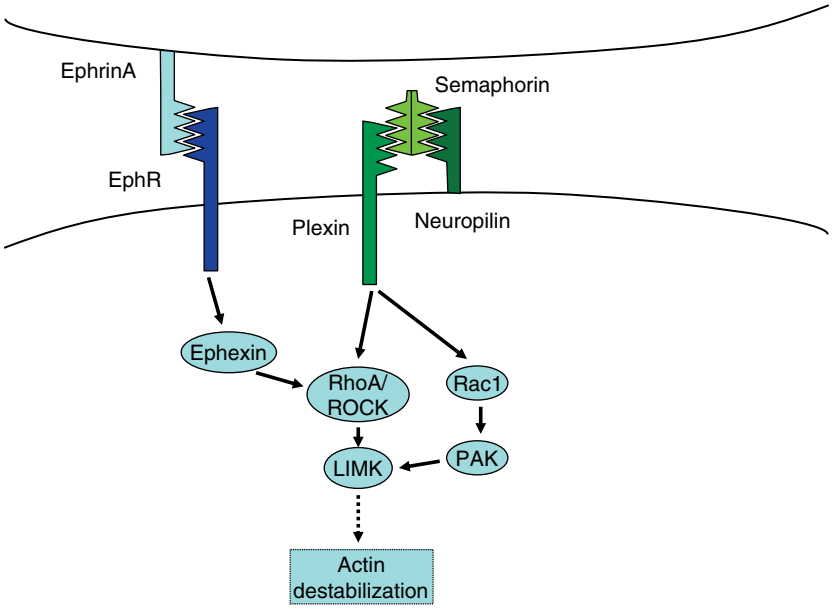

Figure 4 Guidance cue signaling activating growth inhibitory pathways. Membrane-bound ephrins bind to Eph receptors while soluble semaphorins bind to neuropilin/plexin receptors. Both receptor types can elicit growth inhibitory signaling via activation of the RhoA/ROCK cascade, finally inducing actin destabilization and growth inhibition. PAK, p21-activated kinase.

formation of symmetrical structures, such as the optic chiasm. Once bound to axonal Roundabout (Robo) family receptors, slits drive axons away from the midline. Commissural axons initially express low levels of Robo but upregulate the receptor after they successfully crossed to the contralateral side, which prevents them from recrossing the midline. Slits 1–3 have been identified in vertebrates so far, which are able to bind to all three known receptors: Robo 1, Robo 2, and Rig-1.

Left to say, that most of the molecules described above have functions outside the nervous system. Vasculogenesis represents one of the other major processes, where semaphorins, ephrins, netrins, and slits exert their function as guidance cue in development and pathology.

Calcium

The divalent calcium ion is a multifunctional player with roles as a second messenger in the transduction of intracellular signals and as mediator of the action potential. Intracellular cytoplasmic calcium levels ($[Ca^{2+}]_i$) seem to regulate growth cone motility. It has been shown that increased $[Ca^{2+}]_i$ decreases growth cone motility and a reduction of $[Ca^{2+}]_i$ promotes it. However, continuous elevation of $[Ca^{2+}]_i$ can result in adaptation of growth cone behavior, suggesting that targets of Ca^{2+} signaling are downregulated or adjust their sensitivity to the stimulus.

Growth cone motility is not only regulated by baseline cytoplasmic calcium levels, but largely responds to local fluctuations of $[Ca^{2+}]_i$, the so-called calcium transients. Such temporally and locally confined $[Ca^{2+}]_i$ shifts can be generated by calcium flux through ion channels and/or calcium release from intracellular stores. Numerous neurotransmitters are known to regulate calcium channels and thus calcium influx, which in turn can lead to repulsion or attraction of growth cones. For example, the inhibitory neurotransmitter γ-aminobutyric acid (GABA) can act as an attracting stimulus for growth cones, when binding to $GABA_A$ receptors, while exerting repulsive properties, when acting through the $GABA_B$ receptor. The precise spatial and temporal regulation of calcium levels within the growth cone appears to be crucial for the regulation of filopodial dynamics. A proposed binary response to calcium transients, however, greatly oversimplifies the complexity of the growth cone responses. For example, filopodial protrusion can be stimulated by short local Ca^{2+} elevations, but the repeated induction of calcium transients results in the opposite effect.

Downstream targets of calcium include proteins regulating actin dynamics, such as myosin family members, fodrin, α-actinin, gelsolin, and ADF/cofilin. Intracellular calcium sensor proteins, such as calmodulin (CaM), bind calcium and regulate diverse signaling cascades via Ca^{2+}/CaM-dependent phosphatases and kinases. One of the most prominent family members is the Ca^{2+}/CaM-dependent protein kinase II (CaMKII), which shows isoform-specific calcium sensitivity and controls developmental neurite extension.

Guidance Cues in the Adult and Lesioned Nervous System – Play It Again, Sema!

For a long time, the function of axonal guidance cues seemed limited to the developmental period, where neuron–target connection have yet to be established. After successful target innervation, the role of guidance cues seemed obsolete. There is, however, a growing body of evidence that molecules acting as guidance cues and that initially have been attributed classical developmental roles persist throughout adulthood. Even more importantly, experimental data suggest their involvement in the pathophysiology of the lesioned CNS. Next to commonly known inhibitory molecules derived from CNS myelin, the persistence or reexpression of repulsive guidance cues seems to be at least partially responsible for the insufficient regenerative response of the CNS. On the other hand, molecules with neurotrophic properties during embryogenesis may help to overcome regenerative inhibition and support the survival and outgrowth of lesioned neurons.

Myelin-Based Inhibitors of Axon Growth

In contrast to developing axons, neuronal processes in the adult CNS are ensheathed by oligodendrocytes allowing for fast, saltatory conduction of the action potential. Oligodendrocyte-derived myelin has been established as one of the most important inhibitors of axonal growth, while Schwann cells derived from peripheral nerves showed to be permissive toward regenerating axons.

Myelin-based inhibitors of axon growth, that is, Nogo, oligodendrocyte-myelin glycoprotein (OMgp), and myelin-associated glycoprotein (MAG), were initially proposed to account for most of the myelin-derived inhibitory activity. If this were the case, knockouts for one or several of the known myelin-based inhibitors should have markedly improved regenerative capacities in CNS lesion models. However, mutant mice lacking one or all three Nogo isoforms (Nogo-A, B, and C) showed only moderate to no increase in regeneration after spinal cord injury. A similar failure to increase regeneration was observed in MAG knockout mice.

The Nogo receptor (NgR1) transduces the Nogo signal by binding of the 66-amino-acid extracellular domain of Nogo. For signaling, NgR trimerizes with the p75NTR, a versatile receptor initially identified as the low-affinity neurotrophin receptor, and the recently identified LINGO-1, a leucine-rich repeat transmembrane protein. Besides Nogo, MAG and OMgp signal through the trimeric NgR1/p75NTR/LINGO-1 receptor complex. Similar to Nogo knockouts, NgR-deficient mice show a persistent axonal growth inhibition by myelin *in vitro*, while axons from p75NTR-deficient mice were less inhibited, suggesting a more influential role for p75NTR in the regenerative response. Nevertheless, regeneration of corticospinal tract axons after spinal hemisection was not improved either in NgR- or p75NTR-deficient mice, questioning the major role of Nogo/NgR in the mediation of regenerative inhibition. Several hypotheses have been brought forward to explain the persistent lack of regeneration in mice deficient for MAG, Nogo, p75NTR, or NgR. Myelin-based inhibition may employ pathways different from NgR/p75NTR-receptor signaling to induce inhibition. For example, the tumor necrosis factor receptor family member TROY, which is highly expressed in the adult cortex, is able to form a trimeric complex with NgR and LINGO-1 and thus substitute for p75NTR (**Figure 5**).

Next to the 'classical' myelin-based inhibitors, a number of other outgrowth-inhibiting molecules are known to be expressed in the adult CNS. Chondroitin sulfate proteoglycans (CSPGs), such as versican and brevican, are present on differentiated oligodendrocytes and participate in the inhibition of the regenerative response.

Neurocan and phosphacan are two nervous system-specific CSPGs that show high-affinity binding to the extracellular matrix protein Tenascin-C and other cell adhesion molecules, such as Ng-CAM/L1, N-CAM, and TAG-1/axonin-1. Tenascin-R is yet another extracellular matrix protein, which is expressed on oligodendrocytes of adult mice and is upregulated after lesion injury *in vivo*, strongly suggesting an inhibitory role in axonal regeneration.

Developmental Guidance Cues in the Adult

In spite of the overwhelming presence of myelin-derived regeneration inhibitors in the adult, it was equally tempting to suggest that molecules known for their repulsive function as guidance cues during developmental axonal pathfinding additionally act as inhibitory substrates in adulthood.

As outlined previously, ephrins are involved in developmental patterning and axonal guidance and exert their action after binding to transmembrane trks, the Eph receptors. *In situ* hybridization studies in mice revealed that ephrins and Eph receptors are also widely expressed in the adult CNS, although the expression is attenuated compared to the embryonic tissue. Especially, the EphA4 receptor shows a high expression pattern throughout the CNS, while ephrinBs and EphB receptors are highly expressed in regions of known plasticity, such as the olfactory bulb, the hippocampus, and the cerebellum.

A graded expression pattern of ephrinAs in the superior colliculus similar to that during development,

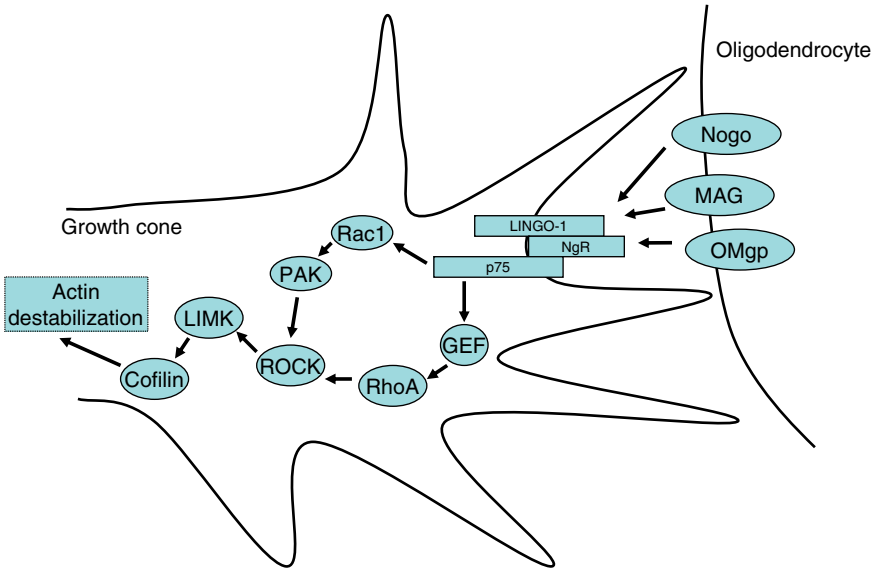

Figure 5 Inhibitory pathways triggered by myelin-based growth inhibitors. Nogo, myelin-associated glycoprotein (MAG), and oligodendrocyte-myelin glycoprotein (OMgp) signal through the trimeric NgR/LINGO-1/p75 receptor. RhoA activation via guanine nucleotide exchange factors (GEFs) induces activation of Rho-associated protein kinase (ROCK) and LIM kinase (LIMK). LIMK phosphorylates cofilin, inducing actin destabilization and growth inhibition.

was observed in the adult mouse before and after differentiation by axotomy. This suggests a requirement for topographic guidance information for functional regeneration of the retinocollicular projection.

Spinal cord hemisection induced an upregulation of EphA4 in astrocytes at the lesion site of wild-type mice. Knockout mice lacking EphA4 were shown to better regenerate corticospinal and rubrospinal axons after spinal cord hemisection than their wild-type littermates. In this paradigm, the regeneration response was improved even though the 'classical' inhibitory signaling cascade via NgR1/p75 was not affected. Myelinating oligodendrocytes in the adult spinal cord further show an expression of ephrinB3 and postnatal cortical neurons express the EphA4 receptor showing sensitivity toward ephrinB3. These data strongly suggest a persistent inhibitory function for developmental repellent molecules.

In analogy to ephrins, a permanent role for semaphorins in the adult CNS has been recently described. Oligodendocytes and astrocytes of postnatal rat optic nerves show expression of multiple semaphorin family members, and RGCs express the neuropilin-1 semaphorin receptor. Of all semaphorins tested, Sema5A had the most pronounced inhibitory effect on growth cone collapse of RGCs. When RGCs were seeded on P8 optic nerve explants, outgrowth was significantly increased after treatment with a Sema5A neutralizing antibody, suggesting an inhibitory role for Sema5A in the regeneration of adult RGC axons. Sema4D/CD100 is expressed on oligodendrocytes of adult mice and acts as a strong inhibitor for postnatal sensory and cerebellar granule cell axons. After spinal cord lesion, Sema4D expression was transiently highly upregulated in oligodendrocytes at the periphery of the lesion. Class 3 semaphorins show an upregulation in the scar tissue of spinal cord-lesioned rats, and receptors for semaphorins are expressed on lesioned cortico- and rubrospinal tract axons.

The involvement of netrins and slits in regenerative paradigms of the adult CNS is less clear. Recently, however, netrin-1 has been shown to promote outgrowth of dopaminergic neuron axons via its DCC receptor, while slit-2 repelled dopaminergic axon growth via its Robo receptor *in vitro*. Whether axonal regeneration of dopaminergic neurons *in vivo* is equally regulated by netrins and slits remains to be determined.

One common denominator of inhibitory molecules seems to be the signaling via the Rho/ROCK pathway. Similar to NgR1/p75/LINGO-1, EphA4 activates the small GTPase RhoA, which in turn leads to an activation of the Rho kinase (ROCK). Rho kinase activates LIM kinase-1 (LIMK1), which phosphorylates cofilin at its serine 3 residue and thus inactivates cofilin. As a potent regulator of actin filament dynamics, the inactivation of cofilin results in actin polymerization and reduced axonal growth.

The role of classical morphogens for axonal regeneration in the lesioned system has not been studied to a great extent yet. However, there is experimental evidence to suggest neuroprotective effects. Supranigral administration of Shh provided an increased survival of dopaminergic neurons after lesion with the selective dopaminergic neurotoxin

1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) in common marmosets and improved their locomotor activity. Adenovirally administered Shh has also been shown to protect motor neurons after axotomy of the facial nerve in adult rats.

Neurotrophic Factors Promoting Axonal Regeneration in the Adult

In contrast to the presence of inhibitory molecules, trophic support derived from neurotrophic factors seems rather limited after CNS lesions. In the mechanically lesioned spinal cord of adult rats, for example, only a marginal upregulation of BDNF and GDNF and its appropriate receptors was observed. Schwann cells of corresponding nerve roots, however, showed a much more pronounced increase in NGF and GDNF mRNA levels, indicating the better regenerative capacity of the peripheral nervous system.

The administration of neurotrophic factors in a therapeutical approach thus seems promising. In the visual system, regeneration of RGC axons can be studied in a crush model of the optic nerve, where the integrity of the nerve is conserved and the transected axons face a nonpermissive environment for regeneration. Several growth factors have been shown to promote not only the survival of RGCs after crush, but equally improve the regrowth of RGC axons into the optic nerve.

Ciliary neurotrophic factor (CNTF) has been shown to promote regeneration of RGC into peripheral nerve grafts, and axotomized RGCs express the appropriate receptors CNTF receptor α and leukemia inhibitory factor receptor (LIFR). The application of CNTF and a Nogo-neutralizing antibody was even synergistic toward promotion of the regenerative response. GDNF, a member of the TGF-β superfamily and potent survival factor for dopaminergic neurons, has been tested for proregenerative properties in spinal cord injury. Mini-guidance channels filled with GDNF and placed into a hemisection gap lesion promoted the regrowth of spinal cord axons into the implant and increased the number of propriospinal neurons.

The combination of growth factors has been the subject of several studies to promote axonal regeneration. For example, fibroblast growth factor-2 (FGF-2), Neurotrophin-3 (NT-3), and BDNF had a synergistic effect on regeneration of RGC axons following crush axotomy, being more effective than treatment with either growth factor alone. BDNF and aFGF promote regeneration of spiral ganglion cell axons into the lesioned organ of Corti. BDNF and NT-3 enhanced propriospinal axonal regeneration in the adult rat spinal cord. The effects of NT-3 and BDNF on regeneration have been shown to be at least partly mediated by enhanced polymerization of f-actin in growth cones.

Intrinsic Regenerative Capacity

The proregenerative effect of neurotrophic factors decreases with the age of neurons. Several factors have been brought forward to be responsible for age-dependent susceptibility to neurotrophic factors and decreased intrinsic regenerative capacity. Differential downregulation of neurotrophin receptors on adult neurons and the increased signaling of inhibitory pathways are likely to contribute to this. An age-dependent reaction of the nervous system toward lesion and a better restoration ability of the peripheral nervous system has been well known from clinical experience. The marked difference between regeneration responses of central and peripheral neurons in the adult CNS may also be explained by differential regulation of transcriptional programs upon lesioning. Pseudo-unipolar DRG neurons, for instance, show a pronounced regenerative response in their peripheral branch, while the ascending central branch is unable to regenerate after injury. Expression of the transcription factor c-Jun, in DRG neurons, is markedly upregulated after lesion of the peripheral branch, while central lesions induce only weak c-Jun upregulation. This correlates with the restricted regeneration potential of the central branch. However, placing a dorsal root into the hemisection site, promoting regeneration into the transplant, again induced marked c-Jun expression in DRG neurons. A very similar effect was shown on the upregulation of growth-associated protein 43 (GAP-43) expression, which is increased equally after lesion of the peripheral branch and almost nonexistent after central lesion. The upregulation of c-Jun and GAP-43 in paradigms of regeneration suggested that these proteins are involved in shaping the cell-intrinsic answer to successful regeneration. Therefore, an overexpression, for example, by viral vectors, has been proposed. Several studies, however, showed that simple overexpression of GAP-43 may not be sufficient to induce regeneration. In the current picture, GAP-43 and c-Jun are standing at the end of a transcriptional program involving the expression of a whole plethora of other so-called regeneration-associated genes (RAGs). However, c-Jun may also act as a two-sided sword in regeneration and death signaling. It has been shown that c-Jun is upregulated after optic nerve axotomy. Moreover, proximal axotomy resulted in stronger c-Jun upregulation and in fewer surviving RGCs than distal optic nerve transection. The arguments for c-Jun as an important apoptosis mediator and potential therapeutical target are supported by the fact that siRNA-mediated downregulation of c-Jun may indeed protect RGCs from axotomy-induced cell death. Thus, the reactivation of growth-associated, developmentally

regulated programs is a necessary prerequisite for axonal regeneration. However, it also renders the adult neuron more susceptible to apoptotic cell death.

In striking contrast to the mammalian nervous system, the CNS of amphibians and fish shows a vigorous regeneration reaction following traumatic lesion. Goldfish, for example, readily regrow their axons after optic nerve transection to functionally reinnervate the tectum and even restore vision. Fish RGC do not degenerate after axotomy, but show a pronounced hypertrophy with dense nucleoli, suggesting an activation of protein and RNA synthesis. Similar to mammals, trophic factors, like NGF, stimulate the outgrowth of explanted fish RGC *in vitro*. Proteins secreted by the sheathing cells of the optic nerve (axogenesis factor-1 and -2) have also been found to promote regeneration. Myelin-based inhibitors of axon growth are also present in the fish optic nerve. However, their composition seems to be different from mammalian myelin: RGC explants from fish regenerate readily on fish myelin, but fail to regenerate on rat myelin. Finally, lesion-induced intrinsic transcription programs in fish and amphibians may lead to expression of different target genes than in mammals. For example, the cell adhesion molecule Contactin 1 (Cntn1) is duplicated in fish and activated after optic nerve lesion, suggesting a regeneration-promoting role similar to L1, N-CAM, or TAG-1. The retinol-binding protein purpurin was shown to be upregulated following axotomy and promoted neurite outgrowth of retinal explants. In summary, differences in the intrinsic regenerative capacity by activation of specific transcriptional programs in fish and amphibians may be at least as important for successful regeneration as a more permissive environment.

Preconditioning lesions to the peripheral branch of DRG axons induced an increased regenerative response in the central branch of the DRG in the dorsal column. This was accompanied by an elevation of intracellular cAMP and inhibition of PKA. Increased regeneration in response to application of neurotrophic factors has also been shown to be mediated by elevated levels of cAMP. As could be demonstrated in cerebellar neurons, priming with BDNF or GDNF incurred resistance against MAG-induced outgrowth inhibition via cAMP elevation and activation of PKA. This involves inhibition of phosphodiesterase by activation of the extracellular-signal regulated kinase (ERK). Direct injection of a cell-permeable cAMP analog (dibutyryl-cAMP; db-cAMP) into the dorsal column of a lesioned spinal cord fostered regeneration. Elevated cAMP levels may modulate the response to lesion via activation of the cAMP response element binding protein (CREB) transcription factor. However, CREB activation alone may be insufficient, as was implicated in a study combining neurotrophin treatment and cAMP in a model of spinal cord injury (Figure 6).

Taken together, axonal regeneration in the adult mammalian CNS is restrained by a multitude of intrinsic and extrinsic factors: an inhibitory environment consisting of repulsive guidance cues and myelin-derived growth inhibitors paired with lack of

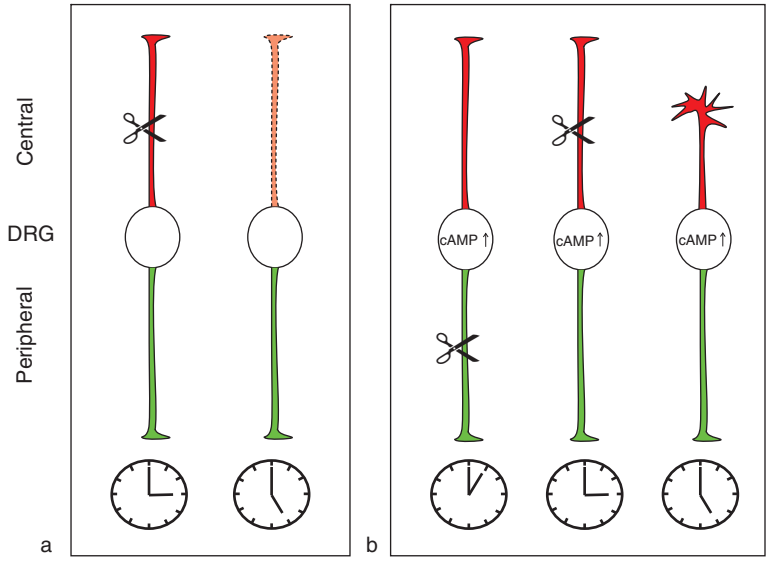

Figure 6 Preconditioning may change intrinsic growth capacity via induction of growth-associated genes. Traumatic lesion to the central branch of the dorsal root ganglion (DRG) results in its degeneration. Preconditioning lesions to the peripheral branch of the DRG and subsequent lesion of the central branch induces a regenerative response in the central branch. This is accompanied by an elevation of intracellular cyclic AMP (cAMP) levels.

trophic support and pro-regenerative transcriptional programs result in insufficient functional renewal capacity. Strategies to overcome regenerative failure will therefore most likely be based on a combination of approaches, targeting several pathways simultaneously.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axonal Injury: Neuronal Responses; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Nogo-A: Its Role in Axon Regeneration; Peripheral Nerve Regeneration: An Overview.

Further Reading

Charron F and Tessier-Lavigne M (2005) Novel brain wiring functions for classical morphogens: A role as graded positional cues in axon guidance. *Development* 132: 2251–2262.

Coleman MP and Perry VH (2002) Axon pathology in neurological disease: A neglected therapeutic target. *Trends in Neurosciences* 25: 532–537.

Gillespie LN (2003) Regulation of axonal growth and guidance by the neurotrophin family of neurotrophic factors. *Clinical and Experimental Pharmacology and Physiology* 30: 724–733.

Gomez TM and Zheng JQ (2006) The molecular basis for calcium-dependent axon pathfinding. *Nature Reviews Neuroscience* 7: 115–125.

Herdegen T, Skene P, and Bähr M (1997) The c-Jun transcription factor-bipotential mediator of neuronal death, survival and regeneration. *Trends in Neurosciences* 20: 227–231.

Klein R (2004) Eph/ephrin signaling in morphogenesis, neural development and plasticity. *Current Opinion in Cell Biology* 16: 580–589.

Luo L (2000) Rho GTPases in neuronal morphogenesis. *Nature Reviews Neuroscience* 1: 173–180.

McKerracher L and Winton MJ (2002) Nogo on the go. *Neuron* 36: 345–348.

Properzi F and Fawcett JW (2004) Proteoglycans and brain repair. *News in Physiological Sciences* 19: 33–38.

Sandvig A, Berry M, Barrett LB, Butt A, and Logan A (2004) Myelin-, reactive glia-, and scar-derived CNS axon growth inhibitors: Expression, receptor signaling, and correlation with axon regeneration. *Glia* 46: 225–251.

Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar

J Silver, K P Horn, S A Busch and A L Yonkof, Case Western Reserve University School of Medicine, Cleveland, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction to the Problem of Regeneration Failure

Unlike their counterparts in the peripheral nervous system (PNS), axons of the central nervous system (CNS) fail to regenerate following injury. With the exception of olfactory sensory neurons within the olfactory bulb and hypothalamo-neurohypophysial axons within the hypothalamus, regeneration of axonal tracts is not observed following disruption of the CNS in adult mammals. The regenerative success of these two pathways has been attributed to the specialized support cells present in these regions: olfactory ensheathing cells in the olfactory bulb and tanycytes in the mediobasal hypothalamus. Interestingly, numerous studies have been conducted using these two cell types, along with Schwann cells, the growth-promoting glia of the PNS, as bridging cells in other regions of the CNS with some success. Typically, adult CNS axons are, at best, capable of abortive sprouting, which results in the failure of these neurons to re-innervate their targets. In fact, the vast majority of functional recovery reported following CNS lesions in experimental models is thought to be due to plasticity of spared or uninjured processes and not to new growth of damaged fibers. Functional recovery of adult CNS neurons is limited because tissue damage following injury initiates a cascade of events that lead to structural and molecular changes within the lesion that are inhibitory to regeneration. There are numerous obstacles for regenerating axons to overcome, but our focus in this article is on the role of the extracellular matrix (ECM) in the glial scar.

The Glial Scar

What Is the Glial Scar?

The glial scar is a complex milieu consisting mainly of reactive astrocytes and secreted ECM molecules. In more severe injuries in which the blood–brain barrier (BBB) is broken down and the meninges are compromised, connective tissue elements, such as fibroblasts, as well as invading immunological cells, such as macrophages, may mix with astrocytes in the lesion (Figure 1). Immediately following damage to the CNS, astrocytes begin to undergo a process known as reactive gliosis, in which they become extremely hypertrophic, dramatically upregulate expression of intermediate filament proteins such as glial fibrillary acidic protein (GFAP) and vimentin, and increase secretion of ECM components. Although the response is termed gliosis, implying an increase in cell division, proliferation of astrocytes is actually rather limited and confined to the penumbra surrounding the lesion core. Recent evidence has shown that depletion of the proliferating population of astrocytes following injury prevents reestablishment of the BBB and results in increased lesion size and an exacerbated inflammatory response. Therefore, despite its devastating effect on axonal regeneration, the glial scar plays an important role in secluding the damaged site and protecting the surrounding uninjured tissue from secondary injury.

Development of the glial scar begins within hours of injury, subsequent to the introduction of non-CNS molecules and cell types into the parenchyma of the brain or spinal cord via disruption of the BBB. The BBB remains permeable to blood and serum components for up to 2 weeks, continuously bathing the CNS with numerous exogenous proteins and factors. There is a correlation between the areas of the most significant BBB breakdown, highest infiltration of activated macrophages, and maximal glial scaring. As the lesion matures, the enlarged and entangled astrocytic processes and secreted ECM components form a tenacious and rubbery scar.

Inhibitory Nature of the Glial Scar

Historically, the association of dystrophic axonal endings with dense reactive astrocytic processes within a CNS lesion led to the hypothesis that the glial scar presents a physical barrier to regenerating fibers. The distinct morphology of struggling growth cones was first described by Ramón y Cajal, and for many years these growth cones were thought to be dead endbulbs or retraction clubs and were considered sterile and incapable of growth. Recent evidence suggests that dystrophic endings retain their ability to regenerate and are capable of returning to an active state. Injured spinal cord neurons can grow into a peripheral nerve graft 4 weeks after injury, despite having stalled in the lesion for an extensive period of time. Shrunken cell bodies of injured rubrospinal tract neurons (Figure 1, E) increase in size following brain-derived neurotrophic factor (BDNF) stimulation 1 year post lesion and can extend growth-associated protein-43 (GAP-43)-positive axons into a peripheral nerve graft.

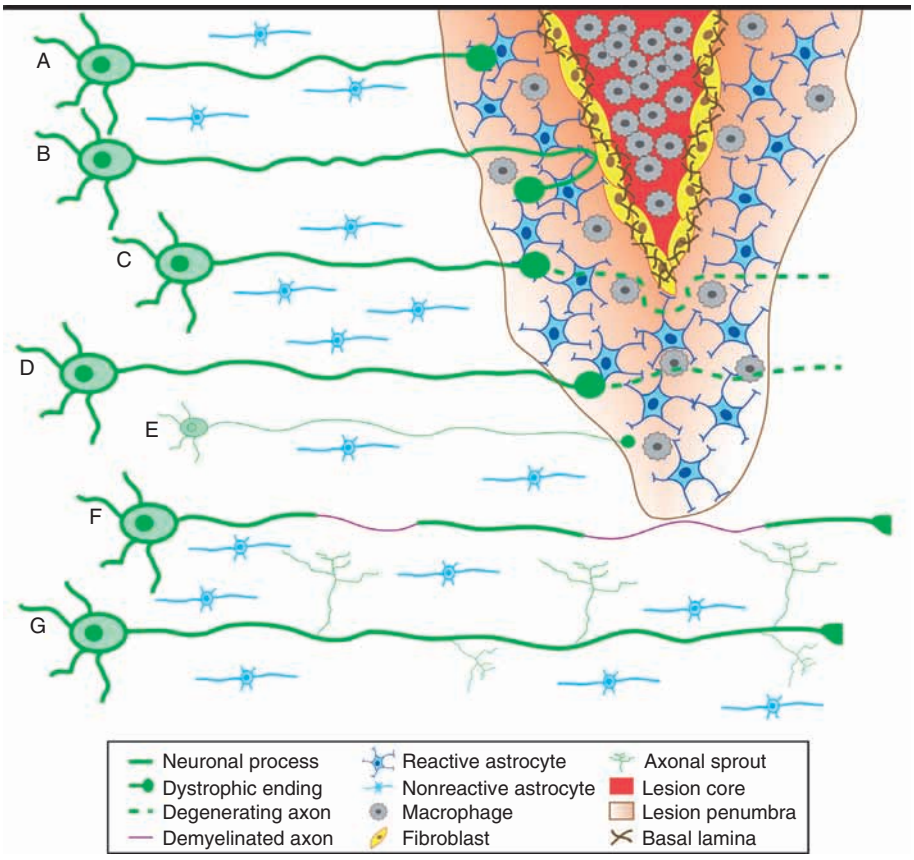

Figure 1 Cellular and molecular elements of a typical penetrating spinal cord lesion. The lesion core is devoid of neurons and astrocytes and is populated by phagocytic macrophages that can cause cavitation. Fibroblasts deposit basal lamina and form a barrier between the lesion core and lesion penumbra. Macrophages and reactive astrocytes occupy the lesion penumbra, where reactive astrocytes secrete inhibitory CSPGs in an increasing gradient toward the lesion core. Other important factors in the lesion penumbra include semaphorins, slits, and Eph/ephrins. A and B, Neurons have been axotomized from the lesion and fail to regenerate, resulting in a dystrophic club ending. B, Dystrophic neuron turns away from the intensely inhibitory lesion core and stops growing in a region of high levels of inhibitory proteoglycan. C, D, Neurons undergo Wallerian degeneration (dotted line) as a result of damage. C, Neuron has been deformed by the lesion and subsequently degenerates because of structural damage and inflammation. D, Axon dies only as a result of secondary damage due to inflammation. E, 'Ghost' neuron atrophies as a result of axotomy. The axon forms a dystrophic endbulb, and the cell body shrinks. The neuron is dormant but may be awakened by growth factors. F, Undamaged axon becomes demyelinated (purple line) both near the lesion and in regions farther away from the lesion. G, Undamaged axon sprouts to re-innervate the spinal cord. Note a contusive injury would be similar to a penetrating injury; however, it would lack the fibroblast infiltrates with their associated ECM components (basal lamina).

Even at extended time periods post lesion, the supposedly sterile endings in the corticospinal tract have been shown to be capable of sprouting.

Axonal Response to the Glial Scar

While initial studies indicated that axons retract significantly from the lesioned area after axotomy, improved tracing techniques have revealed that dystrophic axons actually persist in the vicinity of the lesion (Figure 1, A and B). This implies that even though they are stalled and do not elaborate synapses, an intrinsic mechanism is in place that allows them to maintain their viability and stability. It is possible to generate dystrophic endings on adult sensory neurons *in vitro* by culturing them on inverse gradients of the growth-promoting ECM substrate laminin and the potently inhibitory chondroitin sulfate proteoglycan (CSPG) aggrecan (see below; Figure 2). Time-lapse imaging of these dystrophic endings using this model first demonstrated that these endbulbs are, in reality, quite dynamic. The endbulbs actively rearrange their cytoskeletons, rapidly recycle their plasma membranes, and alter the location of their repertoire of integrin receptors, which bind to ECM substrates and are critical for adhesion and migration. Further work has shown that alteration of integrin expression and membrane turnover also occur *in vivo* following spinal cord injury.

It is essential to distinguish between the states of dystrophy and collapse when describing axonal responses to injury. Following direct contact with myelin components and mature oligodendrocytes, growth cones of mature neurons become shrunken,

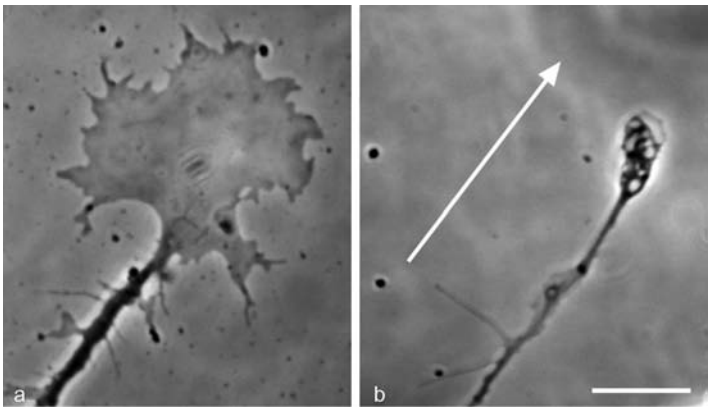

Figure 2 Normal and dystrophic growth cones. (a) Normal growth cone of adult DRG on laminin. The growth cone uses filopodia and lamellopodia to sample the environment for cues. (b) Dystrophic growth cone on a decreasing gradient of laminin plus an increasing gradient of inhibitory proteoglycan (arrow). The growth cone forms a bulbous, dystrophic ending when it encounters inhibitory CSPG, as is characteristic in the CNS lesion environment. Scale bar = 10 μm.

retracting their filopodia and lamellipodia, and enter a temporary state of quiescence. This process is known as collapse and has been well characterized *in vitro*. After some time, growth cones will begin to extend once more only to collapse again upon renewed contact with the collapsing agent. While collapse also results in stalled growth, it is very different from dystrophy, which is a highly dynamic state. Collapse is also prevalent in development and occurs frequently when pathfinding axons encounter transitions between ECM substrates. Observations of the inhibitory nature of myelin *in vitro* led to its implication as another barrier to CNS regeneration, and numerous studies of its major inhibitory element, the Nogo receptor, have used antibodies to block its binding of myelin components to enhance growth of regenerating fibers *in vivo*. While many of these studies have had some success in generating limited functional improvement, recent evidence has suggested that this is due, at least in part, to an increase in plasticity of uninjured fibers.

Immature Astrocytes Promote Regeneration while Mature Astrocytes Make Scars and Block Growth

Embryonic and early postnatal animals have a much higher capacity for regeneration following damage to the CNS than adults. While the neurons undoubtedly differ in their intrinsic growth potential (see below), there are also differences between embryonic and adult astrocyte responses to injury. When hippocampal or retinal ganglion cell (RGC) neurons are cultured on astrocyte explants from lesioned animals, neurite outgrowth is inhibited by reactive mature astrocytes and actually promoted by reactive immature astrocytes. In fact, mature reactive astrocytes can even inhibit the enhanced growth potential of embryonic neurons. Remarkably, when immature astrocytes are transplanted into the adult brain, they can alter the response of endogenous astrocytes and repress formation of the glial scar.

What Are the Molecular Triggers of Glial Scarring?

While the factor or combination of factors responsible for stimulating glial scar formation has not yet been elucidated, there are several putative candidates. Expression of both transforming growth factor beta (TGF-β) 1 and 2 is upregulated in the CNS following injury, and TGF-β2 is known to radically increase production of proteoglycans by astrocytes. The serine protease thrombin has also been implicated as a potential trigger of reactive gliosis. Thrombin, an integral element of the blood coagulation cascade, has additionally been shown to be involved in the chemotaxis and adhesion of monocytes, macrophages, and neutrophils, as well as the regulation of neurite outgrowth, induction of astrocyte proliferation, and activation of microglia in the CNS. Inflammatory cytokines appear to possibly contribute to glial scar formation as well. Members of the interleukin family may also be involved as interleukin 1, which is released by mononuclear phagocytes, has been shown to generate an inflammatory response in astrocytes. Application of interferon gamma has been shown to increase proliferation of astrocytes *in vitro* and treatment of the injured brain with it causes scar expansion. Basic fibroblast growth factor also increases proliferation of astrocytes *in vitro* and its expression increases in the CNS after injury. However, it is important to note that astrocyte proliferation is only a small part in the formation of the glial scar. It is thought that the presence of macrophages is in large part responsible for stimulating astrocytes to take on a reactive state and to construct the glial scar. When Zymosan A, a potent activator of

macrophages, is injected into the CNS, extensive migration of astrocytes away from the inflammatory site leads to cavitation, severe upregulation of proteoglycans, and substantial glial scarring. *In vitro* work has further shown that astrocyte migration away from areas of inflammation can cause axonal stretching and secondary axotomy through neuronal–glial interactions. Therefore, factors derived from blood and/or serum coupled with macrophage invasion combine to contribute to the formation of the glial scar and secondary lesion phenomena, including the demyelination of nearby otherwise healthy axons.

Proteoglycans

Proteoglycans Are ECM Molecules

There are four classes of proteoglycan: CSPG, dermatan sulfate proteoglycan, heparan sulfate proteoglycan, and keratan sulfate proteoglycan. Proteoglycans are components of the ECM that are involved in adhesion, growth, receptor binding, migration, barrier formation, and interactions with other ECM molecules. Proteoglycans consist of a protein core to which glycosaminoglycan (GAG) side chains are bound by a tetrasaccharide link (Figure 3). These GAG chains are composed of repeating disaccharide units formed from two alternating monosaccharides, usually either *N*-acetylglucosamine or *N*-acetylgalactosamine followed by uronic acid. GAG chains can vary dramatically in length, and as many as 100 can be attached to the core protein. The number of GAG chains along with their sulfation pattern and the core protein to which they are bound yield a tremendous amount of variability within and between classes.

CSPGs are inhibitory ECM molecules CSPGs include aggrecan, brevican, neurocan, NG2, phosphacan (also classified as a keratan sulfate proteoglycan), and veriscan. CSPGs are present throughout development and are crucial for axon guidance in the

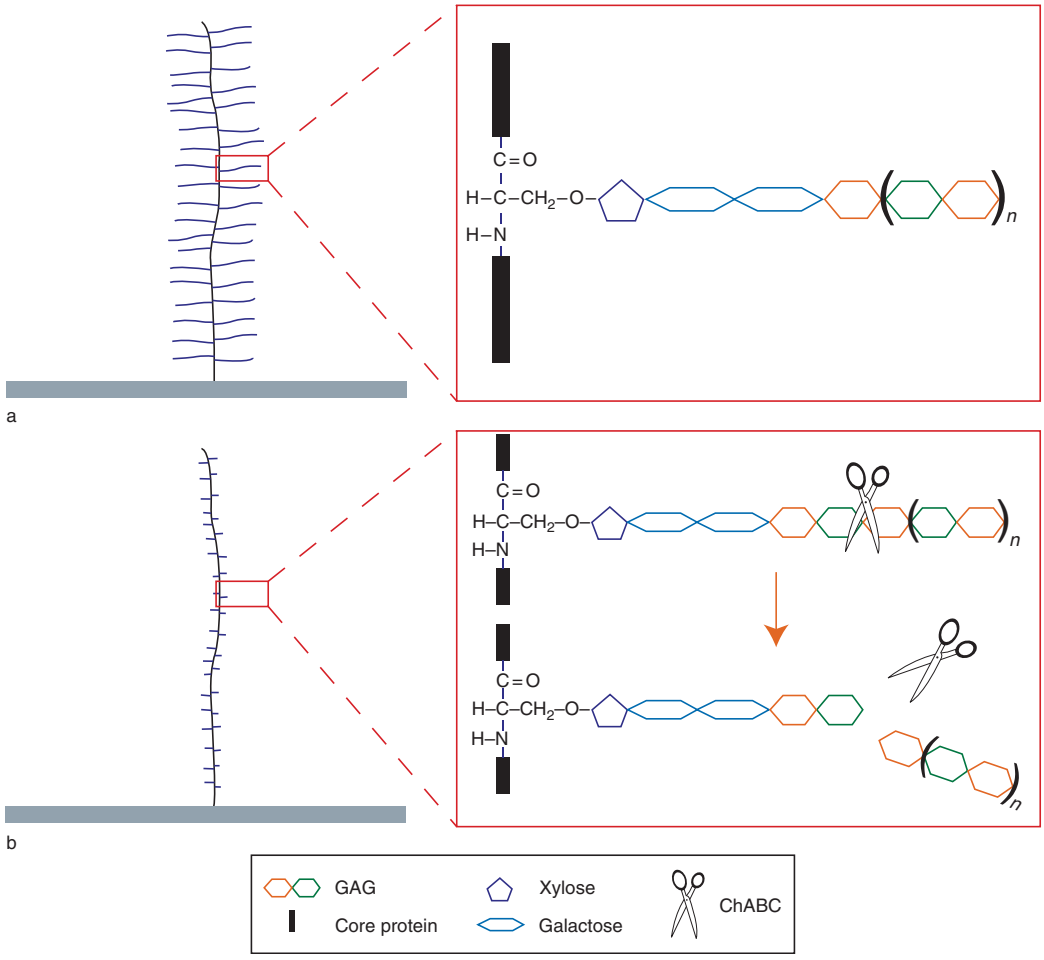

Figure 3 CSPGs and the action of chondroitinase ABC. CSPGs are often anchored to cell membranes, hyaluronan, and other ECM components (gray bar). (a) CSPGs are composed of a core protein linked to GAG side chains by a tetrasaccharide linker of xylose, glucose, and uronic acid. GAGs are disaccharides of *N*-acetylglucosamine or *N*-acetylgalactosamine and uronic acid. These disaccharides repeat to form a long chain. (b) Chondroitinase ABC (ChABC) is a bacterial enzyme that cleaves GAG chains, resulting in a stub of carbohydrates. This abolishes the majority of inhibition from CSPGs.

roof plate, the midline of the rhombencephalon and mesencephalon, the dorsal root entry zone (DRE2), the optic tract and chiasm, and the retina. An elegant study utilized the enzyme chondroitinase ABC (ChABC) from the bacterium *Proteus vulgaris* to digest the GAG chains of CSPGs in the developing retina. The absence of the CSPG GAG chains resulted in pathfinding errors of RGC neurons whose axons grew aberrantly into regions of the retina from which they are normally excluded. The expression of CSPGs changes after the critical period as the animal matures and the axons have reached their targets.

More on the inhibitory nature of CSPGs The physical nature of the glial scar barrier takes several weeks to months to fully develop, which cannot explain the inhibition to axonal growth observed acutely following injury. Interestingly, CSPG expression is robustly upregulated by mature astrocytes in the glial scar within 24 h of CNS injury in adult mammals and maintained for several months. This evidence, along with their inhibitory role in embryonic pathfinding in specific regions, implicated CSPGs as a barrier to regeneration in the CNS. In the early 1990s, it was shown that CSPGs are potently inhibitory to adult neurite outgrowth both *in vitro* and *in vivo* and play a significant role in CNS regeneration failure. CSPGs are not only potent inhibitors of axonal growth when encountered alone, but they can also block the growth-promoting effects of substrates such as laminin, fibronectin, and L1. *In vitro* adult neurites avoid stripes of laminin mixed with aggrecan when given the choice between this combination of substrates and stripes containing laminin alone. When RGCs are cultured on scar explants that are removed from the adult (but not the neonatal) brain, they can only extend long neurites if the wound tissue is pretreated with ChABC, indicating that CSPGs are potent inhibitors of neurite outgrowth *in vivo* as well.

Different neuronal populations experience varying degrees of growth retardation when exposed to CSPGs *in vitro*. RGCs are able to extend further up a stepwise gradient of aggrecan on laminin than either forebrain neurons or embryonic dorsal root ganglion cells (DRGs), although growth was significantly stalled at the transitions between steps. Immature neurons seem to be able to grow on CSPGs by upregulating integrin receptor expression until a threshold concentration of the inhibitory substrate is reached, and they can grow no further. This threshold is different for different populations of neurons. As mentioned earlier, adult DRGs grow up a smooth gradient of CSPG until the potently inhibitory rim is reached, where they stall and form dystrophic endings (Figure 2). In development, some neurons are capable of growing through proteoglycan-rich areas. For example, embryonic hippocampal neurons and thalamocortical fibers growing through the subplate are capable of extending neurites on oversulfated CSPGs. *In vitro*, when neurons are plated on low concentrations of CSPGs mixed with laminin, neurites fasciculate tightly as they would in the subplate *in vivo*. In both embryonic and adult systems *in vitro* and *in vivo*, the balance of proteoglycans and other ECM molecules is critical to the determination of growth cone behavior as well as bundling and branching patterns of axons.

Neuronal populations have different thresholds for proteoglycan inhibition when presented in a gradient *in vivo*, but all populations will eventually become dystrophic within a spinal cord lesion (Figure 1). Proteoglycan concentration is highest at the lesion epicenter and decreases in a gradient throughout the penumbra. As axons enter the lesion, they receive more inhibitory cues, and form increasing numbers of dystrophic endings as they approach the core of the developing glial scar. Axons of motor neurons are not able to grow into the lesion at all, but instead sprout in the undamaged region adjacent to the lesion. Serotonergic neurons are capable of growing to the lesion edge, and sensory axons can grow further into, but not through the lesion. Amazingly, adult sensory neurons have the ability to regenerate robustly through intact and degenerating white matter when gently microtransplanted (without creating a scar) well distal to a large lesion in the spinal cord (i.e., beyond the zone of BBB disruption). However, regeneration stops when the rapidly growing axons reach the lesion environment, indicating the potently inhibitory nature of the glial scar.

Chondroitinase ABC

ChABC digests inhibitory GAG chains on proteoglycans The use of ChABC is currently the most effective method of removing the inhibition of CSPGs (Figure 3). A single injection of ChABC into the normal uninjured brain causes a widespread reduction in the concentration of CSPGs within the ECM for at least 4 weeks. It is important to note that ChABC digestion only cleaves the GAG side chains of CSPGs and that some carbohydrate residues remain attached to the core protein. These residues can be labeled with specialized antibodies to demonstrate digestion of CSPGs in a ChABC-specific manner. While the core proteins of CSPGs are somewhat inhibitory to neurite outgrowth, the vast majority of the inhibitory nature of CSPGs lies within their GAG side chains, which is evident from the work by various laboratories showing that removal of CSPGs from the glial scar enhances regeneration and can provide some functional recovery. Administration of ChABC following a nigrostriatal pathway lesion promoted regeneration

of dopaminergic neurons to the point of re-innervating their target area in the striatum. In animals receiving a bilateral dorsal column injury, intrathecal delivery of ChABC induced growth of ascending sensory and descending motor axons into and possibly slightly past the lesion, which resulted in improved function of both proprioception and locomotion. ChABC treatment has also resulted in increased regeneration of Clarke's nucleus neurons following a hemisection lesion. Similar CSPG digestion and improved regeneration have also been observed in cats receiving hemisection and rats receiving contusion injuries. This gives hope that-continuing efforts to increase the efficacy of this technique could potentially lead to treatments for spinal cord injury in primates and hopefully in humans. Typically, ChABC is administered to the CNS through a pump or by direct injection. ChABC then remains active for 3 days *in vivo*. An alternative strategy might be to genetically engineer CNS cells to express ChABC, allowing for controlled, continuous delivery of the enzyme.

The perineuronal net and the regulation of synaptic plasticity Proteoglycans in the gray matter of the brain and spinal cord may encapsulate synapses with a lattice-like accumulation of ECM known as the perineuronal net (PNN) and prevent neuritic sprouting and synaptic plasticity. PNNs consist mostly of CSPGs produced as the animal passes through critical periods of CNS maturation. As the animal matures, PNN production increases and the potential for plasticity decreases as the synaptic connections are firmly established. Application of ChABC to the uninjured cerebellum in rats induces sprouting of unmyelinated Purkinje cell axon terminals. This CSPG-mediated inhibition of sprouting prevents changes in functional connectivity, which could lead to formation of aberrant synaptic connections. However, it may be advantageous to restore plasticity following injury. For instance, when the adult visual cortex is treated with ChABC, plasticity can be recovered. This has important applications in restoration of vision in adult strabismic animals that have been monocularly deprived since weaning and are therefore cortically blind in one eye. ChABC-induced plasticity could also be important during rehabilitation after stroke and in treating epilepsy.

Other Inhibitory ECM Molecules

In addition to CSPGs, there are many other molecules known to be upregulated following CNS injury. It is well known that axon repellants such as semaphorins, slits, and Eph/ephrins play critical roles in embryonic development, but their roles as inhibitors of regeneration in the mature CNS are just beginning to be understood.

Semaphorins The expression of semaphorin 3, a chemorepellant to growing axons during development, is upregulated by invading fibroblasts following stabbing types of injuries to the spinal cord, cortex, and lateral olfactory tract. Semaphorin 3's high-affinity receptor, neuropilin 1, is upregulated in neurons attempting to regenerate into the lesion. Regenerating axons are excluded from areas near the lesion core that contain semaphorin 3, indicating that semaphorins may be another inhibitory cue present in fibroblast-rich lesions of the CNS.

Eph/ephrins The Eph receptor and its ligand, ephrin, function in cell migration, axon guidance, and tissue patterning during development and are upregulated following CNS injury. Often, the interaction of the Eph receptor and ephrin ligand results in cell repulsion. This finding suggests that Eph/ephrin signaling might be inhibitory to axon regrowth at the lesion site, or may be important in cell migration and tissue repair. Recent work has shown a correlation between segregation of reactive astrocytes and meningeal fibroblasts and ephrinB2 and EphB2 expression, respectively, following spinal cord injury in the adult rat. EphrinB2-positive astrocytes and infiltrating meningeal fibroblasts expressing EphB2 undergo a period of intermingling, which is followed by strict segregation during formation of the glial/mesenchymal scar. EphrinB3 has been characterized as a myelin-based inhibitor of axon outgrowth and is present on myelinating oligodendrocytes. Additionally, adult corticospinal neurons express EphA4, the receptor for ephrinB3, suggesting that these axons may be inhibited by Eph/ephrin signaling in the lesion environment.

Slits Slit proteins also play important roles in cell migration and axon guidance during development and have been implicated in regeneration failure because they are upregulated after injury. The slit proteins are ligands of glycipan-1 and the Robo receptor. Cortical astrocytes express both slit and glycipan-1 mRNA following cortical injury. Robo-1 mRNA is also present on macrophages and fibroblasts following injury, which further implies that slit-mediated repulsion could be playing a role in the inhibitory environment of the lesion. While ChABC treatment removes much of the inhibition in the glial scar due to CSPGs, additional strategies must be developed to overcome inhibition resulting from other molecules present in the lesion environment.

Adult Neurons Lack a Robust, Intrinsic Regrowth Potential

Using ChABC and other treatments to remove the extrinsic inhibitory cues of the glial scar may not be

sufficient to allow long-distance regeneration because adult neurons also lack a strong intrinsic ability to regenerate following a CNS lesion. Multiple strategies have been used *in vitro* and *in vivo* to successfully amplify the growth potential of adult neurons: application of neurotrophic factors, elevation of intracellular cyclic adenosine monophosphate (cAMP) levels, and utilization of a conditioning lesion.

Neurotrophins/Growth Factors Can Increase Axonal Outgrowth and Overcome the Glial Scar

Treatment of transected dorsal column axons with nerve growth factor (NGF) or neurotrophin-3 (NT-3) increased their growth into and out of a peripheral nerve bridge as well as through the glial scar and back into host tissue. Administration of NGF into a dorsal column crush lesion resulted in increased sprouting into the lesion. Using either intrathecal or adenoviral application of NT-3 or NGF to the injured promoted regeneration of DRG axons through the DREZ and into the spinal cord, restoring nociceptive function. It is critical to understand, however, that treatment with neurotrophins can cause regenerating axons to overcome inhibition by CSPGs and other components of the glial scar, whereas ChABC treatment simply removes the source of inhibition. The increase in the capacity for outgrowth of adult axons observed following neurotrophin application is most likely due to increased expression of growth-enhancing genes in the damaged neuron.

Conditioning Lesions/cAMP Can Also Increase Axonal Outgrowth

Conditioning lesions also enhance the capability of adult neurons to regenerate and this is due at least in part to an elevated cAMP concentration within the cell body. This effect can be mimicked in adult sensory neurons by injecting cAMP into the DRG. cAMP is known to act through protein kinase A (PKA) to influence cellular sprouting and outgrowth, and it may be possible to take advantage of PKA-mediated pathways in order to overcome inhibition by the glial scar and myelin. Inhibition of the Rho GTPase increases neurite outgrowth on proteoglycan- and myelin-containing substrates, and the inhibitory activity of CSPGs within the glial scar is mediated though the Rho/ROCK signaling pathway.

Inflammation: Good or Bad?

Inflammation, although intimately linked with scar formation, may also have a positive effect on regeneration by stimulating an intrinsic growth response in neurons. Using Zymosan A (a purified yeast cell wall extract) to activate macrophages outside of the CNS compartment via intravitreal injection promotes regrowth of RGCs into and slightly past an optic nerve crush. Injection of Zymosan A into the DRG *in vivo* enhances the ability of those neurons to extend neurites on a smooth gradient of aggrecan and laminin, as well as through the DREZ and into the spinal cord following a crush injury to the dorsal root. One mediator of the Zymosan A effect is mannose, a monosaccharide. Surprisingly, treatment of neurons with mannose alone can increase their capacity to regenerate, which may be due to an increase in the overall energy reserves of the neurons. Another effector of increased axon growth, a protein secreted by macrophages, has recently been found. Thus, the combination of mannose, macrophage-derived factor, and cAMP should act synergistically to stimulate axon growth.

Combination Treatments Stimulate Regeneration Best

Recent work has combined stimulation of the intrinsic ability of adult neurons to regenerate using neurotrophins or growth factors with glial scar remodeling to decrease the inhibitory nature of the CSPGs within the glial scar. Retinal lesions, which result in denervation of the superior collicus, cause some minimal sprouting of uninjured RGC axons; however, this sprouting is not robust enough to provide functional recovery. Combining ChABC treatment of the superior colliculus with BDNF stimulation of the RGC cell bodies promoted an increase in sprouting greater than either treatment alone. Some of these sprouting fibers displayed synaptic markers indicating a real potential for some functional recovery. Perhaps the therapeutic value of ChABC treatment will be an increase in functional recovery through an enhanced sprouting response.

Conclusion

The physical and chemical barrier of the glial scar is a formidable opponent for axons attempting to regenerate. Reactive astrocytes, macrophages, fibroblasts, and a host of ECM components including inhibitory CSPGs all play a part in restricting CNS axon regeneration. Strategies to promote regeneration have had partial success, mostly resulting in sprouting of undamaged axons and limited regrowth of damaged axons. In order to overcome the inhibitory environment of the glial scar, a multipartite strategy will most likely be required that (1) removes extrinsic inhibitory factors present within the scar, (2) provides a growth-promoting bridge across or around the scar, and (3) stimulates the neurons to increase their intrinsic capacity for growth, which might allow for long-distance regeneration and functional recovery.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Pathfinding: Extracellular Matrix Role; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Glial Responses to Injury; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Nogo-A: Its Role in Axon Regeneration; Peripheral Nerve Regeneration: An Overview.

Further Reading

Bareyre FM, Kerschensteiner M, Raineteau O, et al. (2004) The injured spinal cord spontaneously forms a new intraspinal circuit in adult rats. *Nature Neuroscience* 7(3): 269–277.

Berardi N, Pizzorusso T, and Maffei L (2004) Extracellular matrix and visual cortical plasticity: Freeing the synapse. *Neuron* 44(6): 905–908.

Bray GM, Villegas-Perez MP, Vidal-Sanz M, and Aguayo AJ (1987) The use of peripheral nerve grafts to enhance neuronal survival, promote growth and permit terminal reconnections in the central nervous system of adult rats. *Journal of Experimental Biology* 132: 5–19.

Bush SA and Silver J (2007) The role of extracellular matrix in CNS regeneration. *Current Opinion in Neurobiology* 17(1): 120–127.

Corvetti L and Rossi F (2005) Degradation of chondroitin sulfate proteoglycans induces sprouting of intact Purkinje axons in the cerebellum of the adult rat. *Journal of Neuroscience* 25(31): 7150–7158.

Ellezam B, Dubreuil C, Winton M, et al. (2002) Inactivation of intracellular Rho to stimulate axon growth and regeneration. *Progress in Brain Research* 137: 371–380.

Filbin MT (2003) Myelin-associated inhibitors of axonal regeneration in the adult mammalian CNS. *Nature Reviews Neuroscience* 4(12): 703–713.

Fu SY and Gordon T (1997) The cellular and molecular basis of peripheral nerve regeneration. *Molecular Neurobiology* 14 (1–2): 67–116.

Goldshmit Y, Galea MP, Wise G, Bartlett PF, and Turnley AM (2004) Axonal regeneration and lack of astrocytic gliosis in EphA4-deficient mice. *Journal of Neuroscience* 24(45): 10064–10073.

Jone LL, Oudega M, Bunge MB, and Tuszynski MH (2001) Neurotrophic factors, cellular bridges, and gene therapy for spinal cord injury. *Journal of Physiology* 533(Pt. 1): 83–89.

Ramón y Cajal S (1928) *Degeneration and Regeneration of the Nervous System*, May RM (trans.). London: Oxford University Press.

Reier PJ and Houle JD (1988) The glial scar: Its bearing on axonal elongation and transplantation approaches to CNS repair. *Advances in Neurology* 47: 87–138.

Rhodes KE and Fawcett JW (2004) Chondroitin sulphate proteoglycans: Preventing plasticity or protecting the CNS? *Journal of Anatomy* 204: 33–48.

Sandvig A, Berry M, Barrett LB, Butt A, and Logan A (2004) Myelin-, reactive glia-, and scar-derived CNS axon growth inhibitors: Expression, receptor signaling, and correlation with axon regeneration. *Glia* 46(3): 225–251.

Schwab JM, Failli V, and Chédotal A (2005) Injury-related dynamic myelin/oligodendrocytes axon-outgrowth inhibition in the central nervous system. *Lancet* 365: 2055–2057.

Silver J and Miller JH (2004) Regeneration beyond the glial scar. *Nature Reviews Neuroscience* 5(2): 146–156.

Steinmetz MP, Horn KP, Tom VJ, et al. (2005) Chronic enhancement of the intrinsic growth capacity of sensory neurons combined with the degradation of inhibitory proteoglycans allows functional regeneration of sensory axons through the dorsal root entry zone in the mammalian spinal cord. *Journal of Neuroscience* 25(35): 8066–8076.

Tom VJ, Steinmetz MP, Miller JH, Doller CM, and Silver J (2004) Studies on the development and behavior of the dystrophic growth cone, the hallmark of regeneration failure, in an *in vitro* model of the glial scar and after spinal cord injury. *Journal of Neuroscience* 24(29): 6531–6539.

Willson CA, Miranda JD, Foster RD, Onifer SM, and Wittemore SR (2003) Transection of the adult rat spinal cord upregulates EphB3 receptor and ligand expression. *Cell Transplantation* 12(3): 279–290.

Yin Y, Cui Q, Li Y, et al. (2003) Macrophage-derived factors stimulate optic nerve regeneration. *Journal of Neuroscience* 23(6): 2284–2293.

Axonal Transport and ALS

E L F Holzbaur, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Motor neuron diseases involve the selective degeneration and death of motor neurons, leading to muscle atrophy. While some forms of motor neuron disease show an early onset consistent with developmental defects, the most prevalent forms are later in onset and are characterized by inexorable and fatal neurodegeneration.

The cellular and molecular mechanisms leading to the specific degeneration and death of motor neurons have been extensively examined. While many hypotheses have been proposed, there is a particular focus on the unique size and shape of motor neurons, which are huge cells with extended processes. Due to their morphology, motor neurons are uniquely dependent on active intracellular transport. Therefore, specific defects in transport, particularly along axons that may extend up to a meter in length, may lead to degeneration and cell death. Here, we focus on the hypothesis that defects in axonal transport may be a critical component of the neuronal degeneration observed in motor neuron diseases such as amyotrophic lateral sclerosis.

Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease, has a typical age of onset of 50–60 years, usually leading to death within 5 years due to widespread loss of motor neurons and subsequent muscle atrophy. Currently, there are no available therapies that significantly affect the course of the disease.

The worldwide prevalence of ALS is ~5 in 100 000. Most of these cases (~90%) are sporadic, with no known etiology. Proposed triggers for the disease include environmental influences, such as viral infection or exposure to toxins, or immune dysfunction. However, there are not yet sufficient epidemiological data to support any specific cause. Significantly, the course of disease in sporadic cases is indistinguishable from that observed in the majority of familial cases, consistent with a common pathway of degeneration.

Familial ALS

About 10% of ALS cases are familial, and the identification of the genetic defects involved has provided significant insights into the pathophysiology of the disease, while also raising many further questions. The most common gene linked to ALS encodes Cu/Zn superoxide dismutase 1 (SOD1); *SOD1* is a ubiquitously expressed gene involved in the conversion of superoxide. More than 100 different mutations within the *SOD1* sequence have been identified in patients with familial ALS. These mutations are not limited to the active site of the enzyme; instead they are found throughout much of the coding sequence in *SOD1*. Some but not all of the mutations inactivate the enzyme. As loss of *SOD1* expression does not cause disease in knockout mice, most of the known mutations in *SOD1* are thought to act via a toxic gain-of-function mechanism involving alterations in enzyme folding, stability, or aggregation rather than changes in enzyme activity.

Alternative genetic defects have also been identified in patients with either typical or variant forms of familial ALS (Table 1). The affected proteins include alsin, the product of the *ALS2* gene, senataxin, the product of the *SETX* ('amyotrophic lateral sclerosis 4′) gene, and vesicle-associated membrane protein (VAMP)-associated protein B (VAPB), the product of the *VAPB* ('amyotrophic lateral sclerosis 8′) gene. Two of these proteins, alsin and VAPB, are implicated in vesicle trafficking pathways in the cell. A less severe form of inherited motor neuron disease is due to a mutation in the *DCTN1* gene encoding the $p150^{Glued}$ subunit of dynactin, which is required for transport within the neuron. While neurofilaments may play a role in the pathogenesis of ALS, mutations in neurofilament genes are not a significant cause of inherited ALS.

Animal Models of ALS

The identification of mutations in the *SOD1* gene in familial ALS patients led quickly to the development of transgenic animal models for the disease. Analysis of knockout mice demonstrated that loss of *SOD1* gene expression is not sufficient to induce disease. Instead, overexpression of mutant human *SOD1* results in a mouse model with degenerative disease that mirrors the human disease in several aspects.

The first, and most widely studied model involves the overexpression of the human *SOD1* gene with a G93A mutation. This model has a fully penetrant phenotype with disease onset 3–4 months after birth; affected animals die within 4–5 months. Pathological analysis indicates that motor neurons are primarily affected, and the degeneration and death of motor neurons leads to muscle atrophy. A number of other *SOD1* mutations identified in patients with familial

Table 1 ALS-linked genes involved in transport and trafficking

Protein	*Gene*	*Position*	*Onset*	*Clinical presentation*
SOD1	*SOD1* (*ALS1*)	21q22	Adult	Amyotrophic lateral sclerosis
Alsin	*ALS2*	2q33	Juvenile	Spastic paraparesis; slowly progressive
VAPB	*VAPB* (*ALS8*)	20q13	Adult	Amyotrophic lateral sclerosis
$p150^{Glued}$	*DCTN1*	2p13	Adult	Vocal cord paralysis; slowly progressive weakness of hands, limbs

ALS have also been expressed in mouse models; these models vary in age of onset and severity of disease but show a common phenotype of motor neuron disease. Rat models have also been developed, and display pathology very similar to that seen in the mouse.

Mouse models of other genetic defects linked to ALS have also been developed. For example, knockout of the alsin (*ALS2*) gene associated with juvenile ALS have been generated, which demonstrate only subtle neuropathological defects. At the cellular level, defects in endosomal trafficking are apparent, but these defects are not sufficient to induce significant motor neuron loss or muscle atrophy. Mouse models with mutations or targeted defects in dynein or dynactin function are described in the following sections.

Mechanisms of Cell Death

Initial studies of both mouse models and patient autopsy material revealed two important observations: axonal degeneration of motor neurons and formation of protein aggregates. Based on these observations, and the relative specificity of the disease for motor neurons, a hypothesis was developed that axonal transport was blocked, potentially by the formation of aggregates of mislocalized or mutant protein, and that this in turn leads to the degeneration of the axon, followed by cell death.

Axonal transport The large size and extended processes of motor neurons make them particularly dependent on axonal transport (Figure 1). Diffusion alone is not sufficient to maintain cells of this size. Long-distance transport within the cell is dependent on microtubules, which form a polarized network extending out from the microtubule-organizing center (MTOC), positioned near the nucleus within the cell body. Transport outward from the cell body to the cell periphery is driven by members of the kinesin superfamily of molecular motors (Figure 1). This anterograde transport is required to supply the distal axon and neuromuscular junction with newly synthesized proteins and lipids, including the replenishment of synaptic vesicle components. Multiple kinesins drive anterograde transport, with some functional specification of motor for cargo, but also some redundancy in the system.

Retrograde transport, from cell periphery to cell center, is driven by the molecular motor, cytoplasmic dynein, and its activator, dynactin (Figure 1). In contrast to the extensive nature of the kinesin superfamily, retrograde transport is driven by a single major motor, cytoplasmic dynein. Retrograde transport is required to return damaged and misfolded proteins to the cell center for efficient degradation. Retrograde transport is also critical in signaling pathways, such as neurotrophic factor signaling and injury response pathways. Both the motor domain of dynein and several of the subunits of dynactin are encoded by single genes. The resulting lack of redundancy suggests that mutations in genes encoding subunits of either dynein or dynactin would have a deleterious effect on the neuron, and this prediction is consistent with observations to date.

Active transport is essential for a cell that can extend up to a meter, and thus axonal transport may be an 'Achilles heel' for the motor neuron. Therefore, it is a reasonable hypothesis that defects in axonal transport, either caused directly by mutations in motor proteins or indirectly due, for example, to the accumulation of toxic protein aggregates which might impede efficient transport, could be sufficient to induce distal axonal degeneration. Further, the key role of retrograde transport in pathways such as neurotrophic factor signaling suggests that inhibition of transport may lead more directly to cell death through the perturbation of signal transduction pathways and consequently the activation of apoptosis.

Alternative models for motor neuron death While defects in transport processes provide a plausible model for motor neuron degeneration, and are well supported by data from both genetic and cellular studies, this aspect of motor neuron biology may represent just one piece of the puzzle. Alternative models to explain motor neuron-specific cell death in ALS have also been proposed, including the formation of toxic protein aggregates, excitotoxicity, the destabilization of neuromuscular junctions, oxidative damage due to either mutant *SOD1* or dysfunctional

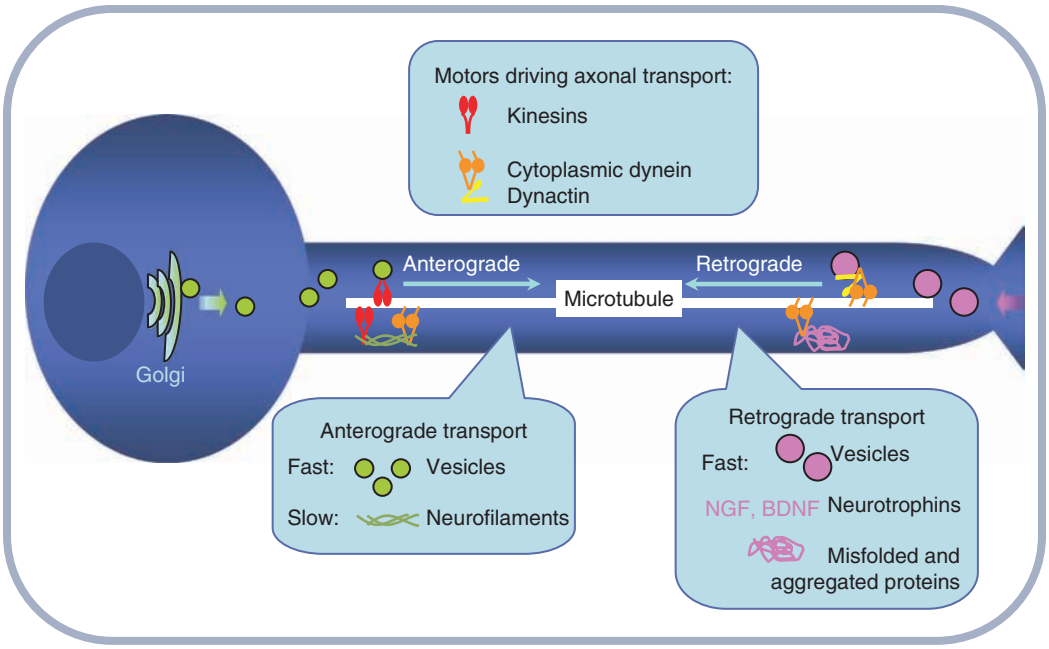

Figure 1 Axonal transport in the motor neuron is driven by microtubule motor proteins. Kinesins drive fast anterograde transport of vesicles carrying newly synthesized proteins and lipids from the Golgi outward along the axon toward the neuromuscular junction. Cytoplasmic dynein and its activator dynactin drive retrograde transport of vesicles, neurotrophins (NGF, nerve growth factor; BDNF, brain-derived neurotrophic factor), and misfolded and aggregated proteins from the cell periphery back to the cell body. Slow anterograde transport of cytoskeletal polymers such as neurofilaments involves the combined activities of both kinesin and cytoplasmic dynein.

mitochondria, and inflammatory processes sparked by activated microglia.

It is critical to note that these pathological processes are not mutually exclusive. It is actually most likely that some initial trigger, such as an accumulation of toxic protein aggregates or the loss of adequate trophic factor support, can lead to activation of cellular stress pathways, which in turn can lead to the triggering of caspase activity, leading to cell death. There is also likely to be a parallel activation of an inflammatory response, which in turn may lead to more widespread death of motor neurons. This model is consistent with recent work demonstrating that while motor neurons are the primary site of disease onset, non-cell-autonomous mechanisms contribute to disease progression.

Evidence for Defective Transport in Mouse Models of ALS

Neurofilament aggregation One of the first indications that defects in axonal transport might be linked to ALS came from the observation of neurofilament aggregates along the axons of motor neurons from affected patients; similar aggregates are observed in animal models. To test for a link between neurofilament aggregation and defects in axonal transport, transgenic mice were generated to overexpress the neurofilament subunit NF-H. Analysis of these mice demonstrated an inhibition in the transport of newly synthesized neurofilaments from cell body outward to the distal axon (Table 2). The transport of other cytoskeletal proteins, such as tubulin and actin, was also found to be inhibited.

Slow axonal transport Neurofilaments and other cytoskeletal polymers are transported down the axon at a rate of 0.2–8 mm day^{-1}, in a process known as 'slow' axonal transport. This transport is orders of magnitude slower than the transport of vesicular cargos in 'fast' axonal transport, at rates of ~200–400 mm day^{-1}. For a long time, the mechanisms driving the slow anterograde transport of neurofilaments were unclear. However, recent progress has shown that slow transport along the axon involves the same microtubule motor proteins that drive fast transport. The key difference is that for cargo such as neurofilaments that undergo slow transport, rapid excursions in either the anterograde or retrograde direction are punctuated by prolonged pauses. Both the frequent changes in direction and the extended pauses contribute to an overall net slow outward movement.

The reasons why cytoskeletal polymers are transported in such an apparently inefficient mechanism remain to be determined, but one possibility is that

Table 2 Defective axonal transport in mouse models of motor neuron disease

Model	*Experimental approach*	*Observed defect*
$Tg^{NF\text{-}H}$	Injection of [^{35}S]methionine followed by profiling of ^{35}S-labeled proteins along the sciatic nerve	Inhibition of slow anterograde transport
$SOD1^{G93A}$	Injection of [^{35}S]methionine followed by profiling of ^{35}S-labeled proteins along the ventral root	Inhibition of slow anterograde transport, selective inhibition of fast anterograde transport
	Tracking of retrograde tracer from muscle injection to motor neuron cell body	Inhibition of retrograde transport
	Electron microscopy of axon hillock and initial segment of anterior horn cells	Aberrant accumulation of neurofilaments and mitochondria
$SOD1^{G37R}$	Injection of [^{35}S]methionine followed by profiling of ^{35}S-labeled proteins along the L5 root and sciatic nerve	Inhibition of slow anterograde transport
$SOD1^{G85R}$	Injection of [^{35}S]methionine followed by profiling of ^{35}S-labeled proteins along the L5 root and sciatic nerve	Inhibition of slow anterograde transport
$Tg^{dynamitin}$	Tracking of retrograde tracer from muscle injection to motor neuron cell body	Inhibition of retrograde transport
Loa	Live imaging of a fluorescently labeled tetanus toxin fragment in cultured embryonic motor neurons	Inhibition of retrograde transport

the oppositely oriented forces, applied simultaneously by plus and minus end-directed motors, result in a continuous tension on the cargo. In the case of neurofilaments, this tension may keep the long, flexible filaments stretched out along microtubules, and therefore may counteract aggregation. Support for this hypothesis comes from the observation that knockdown of dynein expression or inhibition of dynein/dynactin function leads to the distal accumulation of neurofilaments along neuronal processes.

Axonal transport defects in mSOD1 mice – the slowing of slow transport Direct assays for anterograde transport have been performed in mouse models expressing mutant SOD1 (Table 2). Injection of [^{35}S] methionine into the spinal cord of either $SOD1^{G93A}$ or $SOD1^{G37R}$ mice, followed by the profiling of ^{35}S-labeled proteins along the axon, allows the analysis of transport rates of newly synthesized proteins in comparison to those observed in wild-type control mice. In these studies, fast anterograde transport was not generally inhibited, although there may have been some preferential inhibition of specific cargos. However, inhibition of slow axonal transport was observed, including delays in the transport of both neurofilaments and tubulin. This inhibition of transport occurred in both models prior to disease onset.

Defects in retrograde transport in mSOD1 mice Retrograde transport has also been investigated in mice expressing mutant SOD1. In studies in which a retrograde neurotracer was injected into muscle and transport to the cell bodies of motor neurons was assayed, significant inhibition of transport was observed in transgenic mice expressing $SOD1^{G93A}$ as compared to wild-type littermate controls (Table 2). The observed inhibition was at a very early point in the course of disease; inhibition was significant by 50 days after birth, well prior to the onset of clinical disease in this model. This observation suggests that disruption of retrograde transport is a very early event in the course of motor neuron degeneration and death.

Studies on axonal transport in embryonic motor neurons cultured from the $SOD1^{G93A}$ mouse also show a significant slowing of retrograde transport (Table 2). Live cell analysis of the retrograde motility of a fluorescently labeled fragment of tetanus toxin shows a shift in the speed profile, indicative of an increased frequency of pauses and oscillatory movements, as compared to the more efficient transport observed in motor neurons from wild-type mice.

The only other marker that shows such an early defect in the $SOD1^{G93A}$ model is the disruption of the Golgi apparatus, which has been observed by 30 days after birth. Disruption of the Golgi is also observed in patient tissue. This is potentially of interest, since in addition to its role in retrograde axonal transport, cytoplasmic dynein drives endoplasmic reticulum (ER)-to-Golgi trafficking in the cell body and is required to maintain an intact Golgi apparatus near the MTOC of the cell. Inhibition of dynein leads to the disruption of the Golgi apparatus.

Thus, both the observed slowing of retrograde transport in the $SOD1^{G93A}$ model and the disruption of the Golgi observed in ALS patients and in a mouse model are consistent with a functional disruption in the activity of cytoplasmic dynein within the neuron. The inhibition of slow anterograde axonal transport observed in mice expressing mutant *SOD1* may also be consistent with a defect in dynein function, as dynein interacts directly with neurofilaments as they move along the axon.

Yet to be determined is the mechanism by which the expression of mutant *SOD1* can disrupt dynein function. One possibility is that the disruption is indirect, due to the activation of stress response pathways by the accumulation of misfolded SOD1 protein. Alternatively, the mechanism may be more direct. Dynein co-localizes with SOD1 aggregates in neurons expressing mutant *SOD1*; these aggregates may deplete the pool of active motors along the axon. Further work will be required to address this question.

Defective Retrograde Transport Is Sufficient to Cause Motor Neuron Degeneration

The dynamitin transgenic mouse In order to test the hypothesis that disruption of retrograde axonal transport is sufficient to induce motor neuron degeneration, a transgenic mouse was developed with a targeted inhibition of dynein/dynactin function in motor neurons. The dynactin subunit dynamitin was expressed at high levels in motor neurons; overexpression of this subunit is sufficient to disrupt the integrity of the dynactin complex and thus effectively disrupts dynein-mediated transport. The resulting $Tg^{dynamitin}$ line exhibits late-onset, slowly progressive degeneration of motor neurons, leading to preferential loss of large-caliber motor neurons and subsequent muscle atrophy.

Loa and Cra1 mutant mice Further evidence linking defects in axonal transport to the development of motor neuron disease comes from the analysis of two lines of *N*-ethyl-*N*-nitrosourea (ENU)-mutated mice, Loa and Cra1. Both lines were found to have point mutations in the gene encoding the heavy chain of cytoplasmic dynein, which includes the motor domain of the enzyme. While neither of the point mutations maps to the motor domain of dynein, both are in domains involved in subunit interactions within the dynein holoenzyme. Therefore, the two mutations are not predicted to abrogate dynein function, but are more likely to make the motor either less stable or less effective in driving motility along the axon.

Both Loa and Cra1 heterozygous mice display phenotypes remarkably similar to the phenotype of the $Tg^{dynamitin}$ mouse. Both lines exhibit late-onset, slowly progressive degeneration of motor neurons. Importantly, while the mutations are ubiquitously expressed, the degeneration is limited to motor neurons in heterozygous mice. Consistent with an impairment but not a block in dynein function, neither mutation is lethal when heterozygous. When homozygous, both mutations are lethal within a day of birth, which may be due to an impaired ability of the pups to suckle resulting from abnormal development of facial motor neurons. However, there are no gross perturbations of the nervous system, suggesting that the mutant dynein, while functionally impaired, is sufficient to drive normal development. In contrast, a complete knockout of dynein heavy chain is lethal early in embryogenesis in the mouse.

The effects of the Loa mutation on intracellular transport have been investigated. Transport of a retrograde tracer, a fragment of tetanus toxin, was not significantly affected in embryonic motor neurons cultured from heterozygous mice, suggesting that either the impairment is subtle, or that the wild type gene is haplosufficient over this time scale. In neurons cultured from Loa/Loa mice, there was a decrease in the extent of high-speed transport of the retrograde tracer, and a marked increase in the frequency of pauses. These data indicate a significant defect in the efficiency of retrograde transport (Table 2).

A somewhat unexpected but potentially important result has come from studies examining the effects of crossing either the Loa or Cra1 mutant mice to transgenic mice overexpressing mutant *SOD1* (the $SOD1^{G93A}$ line). Rather than showing an additive or even synergistic effect on motor neuron cell death, mice expressing both the $SOD1^{G93A}$ mutant and either the Loa or Cra1 mutation showed an extension of life span (by 28% and 12%, respectively). Disease onset was also delayed. The mechanism behind this result is not yet understood. One interesting possibility is that the impairment of dynein function induced by either the Loa or Cra1 mutation delays the transmittal of a signal indicative of cellular stress, thus slowing the onset of cell death pathways.

Alternatively, it has been proposed that the expression of mutant dynein may rescue the axonal transport defect caused by the expression of mutant *SOD1*, perhaps by righting the balance between anterograde and retrograde transport and therefore restoring axonal homeostasis. Numerous studies have suggested that anterograde and retrograde transport along the axon are intimately linked, as mutations or inhibitory reagents that should selectively affect transport in a single direction instead have a bidirectional inhibitory effect. Mechanisms of motor coordination may involve either direct links between motors or between motors and adapter proteins such as dynactin, or more indirect links through coordinated regulation of motor protein activity.

Mutant Dynactin in Motor Neuron Disease

Clear demonstration of a link between the axonal transport machinery and motor neuron disease in humans came from the identification of a mutation in the gene encoding the $p150^{Glued}$ subunit of dynactin as the cause of an inherited form of motor neuron disease.

Patients with a G59S mutation in the *DCTN1* gene display an autosomal dominant, slowly progressive form of inherited motor neuron disease. Onset in early adulthood is marked by vocal fold paralysis, leading to breathing difficulties, as well as progressive weakness of the muscles in the face and hands, with distal limb atrophy developing later. Unlike typical ALS, this mutation is not uniformly fatal.

Dynactin is a required activator for most dynein functions in the cell. Dynactin enhances the processivity of dynein, meaning that the motor complex takes longer runs along the microtubule per binding encounter. Dynactin also links dynein to some intracellular cargos. The $p150^{Glued}$ mutation identified in affected patients is a G59S substitution in a critical domain of the polypeptide, known as the CAP-Gly domain, which binds directly to microtubules. The mutation affects the ability of the polypeptide to bind to microtubules, but also increases the propensity of the polypeptide to misfold and to aggregate. Thus, at the cellular level, the G59S mutation both inhibits dynein function and results in the formation of intracellular protein aggregates. Together, these detrimental effects lead to the specific death of motor neurons. Pathological protein aggregates labeled with antibodies to both dynein and dynactin have been observed in motor neurons from affected patients, consistent with this hypothesis.

More recently, four additional mutations have been identified in the coding region of the *DCTN1* gene in patients with ALS. In all cases reported to date, there are insufficient genetic data to determine if these mutations are causative, contributory, or merely allelic variants. Further analysis will be required to determine how significant dynactin mutations are to the development of full-blown ALS.

Why Are Motor Neurons So Susceptible to Defects in Axonal Transport?

Consistently, studies on mouse models of motor neuron disease demonstrate defects in retrograde transport as well as enhanced protein aggregation. Defects in anterograde motors can also lead to neurodegenerative disease. For example, one form of Charcot–Marie–Tooth disease is caused by a point mutation in a kinesin motor protein. However, motor neurons appear to be uniquely susceptible to defects in retrograde axonal transport.

Three different mechanisms may explain these observations. The first is that the health of a motor neuron depends on a continual interaction with its environment. For example, a continuous supply of neurotrophic factors may be required. Neurotrophins, secreted by target tissues, bind to receptors on the motor neuron and are then transported from the synapse back to the cell body to activate signal transduction pathways and modulate gene expression patterns. Inhibition of retrograde transport would block this continuous signaling and thus could deleteriously affect the cell. None of the models studied to date show a complete block in retrograde transport, but instead show a partial inhibition. Slowed or inefficient neurotrophin signaling over time may lead to the late-onset and progressive cell death seen in ALS patients, as well as to patients expressing mutant dynactin.

Alternatively, cell death may result from the defective clearance of misfolded or aggregated proteins. Protein aggregation is a common feature observed in both patients and model systems. Further, dynein has been implicated as a key player in the clearance of misfolded proteins through autophagy. Mutations in dynein/dynactin or alterations in the function of the complex may lead to accumulation of toxic protein aggregates along the axon, which would further impede transport. This could represent a generalized defect in clearance of misfolded proteins or a more specialized defect – for example, in the clearance of mutant SOD1 protein or of aggregated neurofilaments. This mechanism has the potential to amplify the initial defect; dynein co-localizes with aggregates of mutant SOD1 protein and this association may deplete the pool of active motor protein. Dynein also interacts with neurofilaments; loss or inhibition of dynein leads to neurofilament aggregation, a common pathological marker of motor neuron disease.

Finally, alterations in axonal transport may be secondary to the real defect. Inhibition of dynein function may lead to more critical insults to the neuron, such as the disruption of dynein-mediated trafficking (e.g., from the ER to the Golgi). Motor neurons may be uniquely sensitive to trafficking defects, given their large size and metabolic load. Or, the key defect could be in the dynein-driven motility of intracellular organelles, such as the mitochondria. Mislocalization of mitochondria may lead to mitochondrial dysfunction, which in turn may be sufficient to initiate cell death pathways. In support of this possibility, mislocalization of mitochondria has been observed in motor neurons of patients with ALS.

Defects in Transport, Trafficking, and Cytoskeletal Proteins Linked to Neurodegenerative Disease

Recent progress in identifying the genetic causes of inherited neurodegenerative diseases has turned up some intriguing results. As already noted, the identification of mutations in motor proteins has

provided strong support for the hypothesis that active axonal transport is critical to maintain the health of neurons. Further, mutations in other proteins (such as Huntingtin or amyloid precursor protein) may also affect axonal transport either directly or indirectly. Together, these data suggest that neurons are absolutely dependent on axonal transport, and are uniquely vulnerable to defects in the motors and other cytoskeletal proteins involved in this transport. A better understanding of the underlying cell biology of motor neurons will therefore be required to fully understand the development of pathology in patients with ALS, and to design rational approaches to therapeutic interventions.

See also: Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Amyotrophic Lateral Sclerosis (ALS); Animal Models of Motor and Sensory Neuron Disease; Axonal Injury: Neuronal Responses; Axonal Transport Disorders; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and Neurodegenerative Diseases; Demyelinating Diseases; Microtubule Associated Proteins in Neurons; Microtubules: Organization and Function in Neurons; Slow Axonal Transport; Vesicular Sorting to Axons and Dendrites.

Further Reading

Boillee S, Vande Velde C, and Cleveland DW (2006) ALS: A disease of motor neurons and their nonneuronal neighbors. *Neuron* 52: 39–59.

Bruijn LI, Miller TM, and Cleveland DW (2004) Unraveling the mechanisms involved in motor neuron degeneration in ALS. *Annual Review of Neuroscience* 27: 723–749.

Collard JF, Cote F, and Julien JP (1995) Defective axonal transport in a transgenic mouse model of amyotrophic lateral sclerosis. *Nature* 375: 61–64.

Hafezparast M, Klocke R, Ruhrberg C, et al. (2003) Mutations in dynein link motor neuron degeneration to defects in retrograde transport. *Science* 300: 808–812.

He Y, Francis F, Myers KA, et al. (2005) Role of cytoplasmic dynein in the axonal transport of microtubules and neurofilaments. *Journal of Cell Biology* 168: 697–703.

Kieran D, Hafezparast M, Bohnert S, et al. (2005) A mutation in dynein rescues axonal transport defects and extends the life span of ALS mice. *Journal of Cell Biology* 169: 561–567.

LaMonte BH, Wallace KE, Holloway BA, et al. (2002) Disruption of dynein/dynactin inhibits axonal transport in motor neurons causing late-onset progressive degeneration. *Neuron* 34: 715–727.

Levy JR, Sumner CJ, Caviston JP, et al. (2006) A motor neuron disease-associated mutation in $p150^{Glued}$ perturbs dynactin function and induces protein aggregation. *Journal of Cell Biology* 172: 733–745.

Ligon LA, LaMonte BH, Wallace KE, et al. (2005) Mutant superoxide dismutase disrupts cytoplasmic dynein in motor neurons. *NeuroReport* 16: 533–536.

Munch C, Sedlmeier R, Meyer T, et al. (2004) Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. *Neurology* 63: 724–726.

Pasinelli P and Brown RH (2006) Molecular biology of amyotrophic lateral sclerosis: Insights from genetics. *Nature Reviews Neuroscience* 7: 710–723.

Puls I, Jonnakuty C, LaMonte BH, et al. (2003) Mutant dynactin in motor neuron disease. *Nature Genetics* 33: 455–456.

Puls I, Oh SJ, Sumner CJ, et al. (2005) Distal spinal and bulbar muscular atrophy caused by dynactin mutation. *Annals of Neurology* 57: 687–694.

Sasaki S, Warita H, Abe K, et al. (2005) Impairment of axonal transport in the axon hillock and the initial segment of anterior horn neurons in transgenic mice with a G93A mutant SOD1 gene. *Acta Neuropathology* 110: 48–56.

Williamson TL and Cleveland DW (1999) Slowing of axonal transport is a very early event in the toxicity of ALS-linked SOD1 mutants to motor neurons. *Nature Neuroscience* 2: 50–56.

Zhang B, Tu P, Abtahian F, et al. (1997) Neurofilaments and orthograde transport are reduced in ventral root axons of transgenic mice that express human SOD1 with a G93A mutation. *Journal of Cell Biology* 139: 1307–1315.

Axonal Transport and Alzheimer's Disease

L S B Goldstein, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Basic Features of Alzheimer's Disease

The initiating symptoms or presenting symptoms of Alzheimer's disease (AD) are well known. Problems with memory, language, and cognition are all characteristic features of the behavioral symptoms that signal the initiation of this terrible neurodegenerative disorder. Most AD occurs relatively late in life, generally after 65 years of age. It afflicts 10% of people over the age of 65 and 50% of people over the age of 85. The time from diagnosis to fatality is highly variable, ranging anywhere from a few years to over a decade or more. The tragedy of AD is that not only are there substantial deficits in cognition accompanied by neurodegenerative changes in the brain, there are presently no effective drugs and no effective treatments, and the disease progresses inexorably in its course to end stage, in which a persistent vegetative state is common.

The pathological hallmarks in the brains of people who have had AD are well recognized in postmortem autopsy material. These features include amyloid beta (Aβ) plaques and neurofibrillary tangles. Amyloid plaques consist of aggregated Aβ peptides derived by proteolytic processing of the amyloid precursor protein (Figure 1). Amyloid plaques are generally associated with dystrophic neurites, which can be abnormal axons or dendrites that are characterized by substantial swelling and distension, and accumulations of vesicles and organelles. Dystrophic neurites surround and invade the plaques, but are also in regions of diseased brains that lack amyloid plaques. Neurofibrillary tangles are aggregates of a protein called tau that normally binds to and stabilizes microtubules, and which, following hyperphosphorylation and conformational change, forms characteristic paired helical filaments. Although amyloid plaques and neurofibrillary tangles are the hallmark diagnostic pathological features, their correlation with cognitive characteristics of AD is not consistent and is controversial. Many researchers have argued that loss of synapses in key areas of the diseased brain is the earliest pathological feature of AD that is most highly correlated with cognitive defects. Neuronal death is an invariant though relatively late feature of AD, as is loss of axons and white matter. The loss of axons, white matter, and synapses could be a consequence of early changes in axonal transport in AD.

Although most AD is late onset, there are well-recognized hereditary forms (familial AD, or FAD) that can begin as early as the fourth decade of life and are highly virulent in their course. Mutations in the gene encoding amyloid precursor protein (APP) and mutations in the genes encoding presenilin (PS) proteins are the major recognized causes of FAD. Duplications of the APP gene can also give rise to mid- to late-onset AD. Two other major risk factors are known. One major genetic risk factor is the allelic state in the gene (*APOE*) encoding apolipoprotein E (ApoE), which has a role in cholesterol trafficking. While the $\varepsilon 2$ variant of the *APOE* gene is protective, the $\varepsilon 4$ variant predisposes to late-onset AD. Another major risk factor is traumatic brain injury. How these risk factors lead to AD is far from clear, and indeed, what molecular defects and pathways lead to any form of AD remain controversial.

The dominant hypothesis proposed to account for the molecular and cellular changes found in AD is the amyloid cascade hypothesis, which proposes that AD is caused by toxic Aβ_{40} or Aβ_{42} peptide fragments liberated by proteolysis of APP (Figure 2). In this hypothesis, the hereditary FAD mutations are proposed to generate excess toxic forms of Aβ, and apolipoprotein E variants are proposed to affect the rates of turnover of these peptides. Most important, the amyloid cascade hypothesis and its major variants indicate that the toxic Aβ peptides initiate a biochemical and cellular cascade of pathology leading to synaptic loss, neurofibrillary tangles, and ultimately neuronal death. While there is considerable evidence for the amyloid cascade hypothesis, principally in the human APP FAD mutations, questions remain, and there are a number of alternative hypotheses that have been proposed over the years. Another class of hypothesis suggests that early deficits in axonal transport either cause, or contribute substantively to, the development of AD.

Axonal Transport Is Disrupted in AD

There has been considerable evidence for some time that axonal transport is disrupted at late stages of AD. For example, although basal fore-brain cholinergic neurons exhibit phenotypes that suggest that they lack retrograde neurotrophic input from nerve growth factor, diseased brains appear to retain high levels of

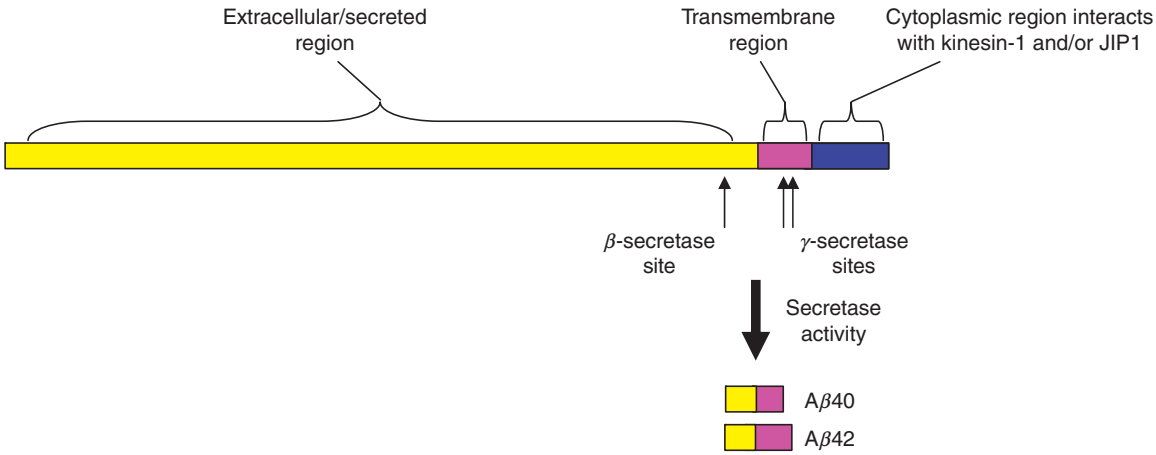

Figure 1 Schematic diagram of amyloid precursor protein. The cytoplasmic region is thought to interact with the axonal transport machinery either directly via binding to the kinesin light chain subunit of kinesin-1, or indirectly via binding to the Jun N-terminal kinase-interacting protein 1, JIP1. Sites for proteolytic activity by β- and γ-secretases are shown, which together liberate amyloid beta (Aβ) peptides that are either 40 or 42 amino acids in length.

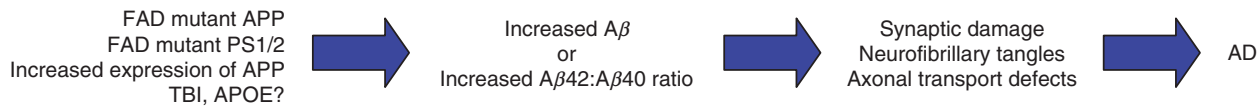

Figure 2 A standard view of the amyloid cascade hypothesis in which the various mutations, genetic variants, and environmental insults induce a series of molecular events caused by increasing levels of toxic human amyloid beta ($A\beta_{40}$ or $A\beta_{42}$), culminating in Alzheimer's disease (AD). FAD, familial Alzheimer's disease; APP, amyloid precursor protein; PS1/2, presenilins 1 and 2; TBI, traumatic brain injury; APOE, apolipoprotein E.

nerve growth factor in the cortex where basal forebrain cholinergic neurons ordinarily connect to cells that secrete nerve growth factor (Figure 3). The levels of cholinergic enzymes in the cortex appear also to be reduced in AD, even though basal forebrain cholinergic neurons appear to express normal levels of cholinergic markers even after the onset of AD.Finally, there are reports that axonal transport is defective in neurons taken from brains of AD patients postmortem, and there are long-standing observations of destruction of the microtubule cytoskeleton in axons in diseased neurons. These disruptions have for some time been thought to be related to the hyperphosphorylation and aggregation of tau protein into neurofibrillary tangles, although what events are causative and what events are consequences of each other remain unclear.

The observation that axonal transport is defective at late stages of AD has led a number for workers over the years to propose that axonal transport defects might be early, possibly causative, events in AD. However, evidence for early defects in axonal transport in AD was lacking until recently. Recent work has provided evidence for early changes in axonal transport in postmortem AD brains and in animal models of AD in which APP is overexpressed. One striking example is the finding of dystrophic axons early in AD and in AD models prior to the formation of amyloid plaques, and thus not associated with, or in the vicinity of, amyloid plaques. These dystrophic axons exhibit signs of axonal transport failure, including massive accumulations of vesicles and organelles in axons that appear to be distended by these accumulations. This type of phenotype is characteristic of mutations that disrupt the axonal transport machinery directly in *Drosophila* and in the mouse. In this context, there is some evidence that deficits in axonal transport can induce formation of, or enhance accumulation of, Aβ. A final line of evidence that raises the possibility that axonal transport deficits are relevant to the initiation of AD is the body of evidence that suggests that there are age-dependent changes and reductions in microtubules in axons. Observed changes include accumulation of APP and other transported material in axons, increases in axonal dystrophies, and some direct evidence that transport itself is reduced in aging neurons relative to younger neurons.

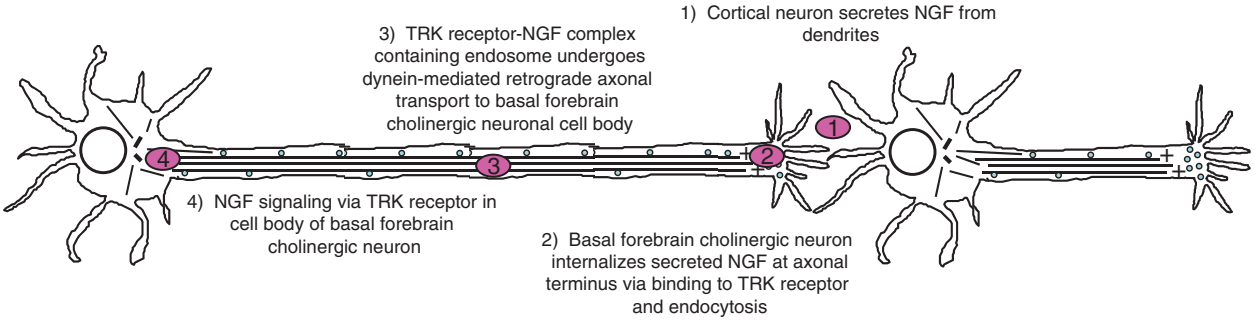

Figure 3 Pathway of nerve growth factor (NGF) signaling from cortical neurons to target cholinergic neurons in the basal forebrain. This transport and signaling pathway appears to be defective in Alzheimer's disease. Secretion of NGF by cortical neurons (1) is followed by internalization at presynaptic endings of basal forebrain cholinergic neurons (2). Retrograde transport of vesicles containing NGF and activated TrkA in postendosomal carriers (3) culminates in arrival at the basal forebrain cholinergic neuron cell body and signaling (4).

Roles and Effects of AD Risk Factors and Genes in Axonal Transport

Several reports suggest that APP and its proteolytic products may play normal roles in, and/or be able to interfere with, axonal transport. For example, deletion of the APP gene in *Drosophila* and in the mouse appears to generate mild defects in axonal transport. Overexpression of APP and its relatives dramatically interferes with axonal transport in *Drosophila* and in the mouse, and even simple duplication of the APP gene in the mouse appears to significantly inhibit some forms of axonal transport – in particular, retrograde transport of the nerve growth factor (NGF)–neurotrophin receptor signaling complex to the basal forebrain. Although this could be the result of, and often has been interpreted as such, an excess of toxic human Aβ, strong evidence in *Drosophila* and in the mouse suggests that even if human Aβ is toxic to axonal transport, other factors, perhaps the C-terminus of the APP protein, may be bigger contributors to these defects.

The mechanism of APP-induced defects in axonal transport remains unknown. One suggestion, based on biochemical evidence from a number of groups, is that the cytoplasmic tail of APP can form either a direct or indirect interaction with the kinesin light chain subunits of conventional kinesin, kinesin-1. If indirect, this interaction may occur via formation of a ternary complex with the Jun kinase scaffolding protein called JIP1 (c-Jun N-terminal kinase-interacting protein 1). There is also evidence from the squid and *Drosophila* that the C-terminus of APP interacts with the axonal transport machinery and is required for interference with axonal transport in overexpression situations. This C-terminus is also required to generate movements of APP in *Drosophila* neurons and in squid axoplasm. These findings are consistent with the proposal that overexpression of APP poisons axonal transport by competing for either JIP1 or kinesin light chain, which are needed for axonal transport of other materials (Figure 4). Although several groups have provided evidence in favor of this hypothesis, other groups do not agree. Strong additional evidence in support of this idea comes from the observation that reductions in kinesin-1 subunits in *Drosophila* or in the mouse enhance transport defects caused by overexpression of APP.

There is considerable evidence that FAD mutations and deletion mutations in the gene encoding presenilin 1 may also interfere with axonal transport. One obvious mechanism that is consistent with the amyloid cascade hypothesis is that presenilins might cause axonal transport defects by enhancing or altering Aβ production in neurons, which is then toxic to axonal transport. However, the presenilin mutations that cause familial forms of AD in humans can poison axonal transport of APP, and perhaps other proteins, in the mouse in the absence of human APP. These findings, combined with considerable evidence that kinesin-1 is required for APP transport in axons and with evidence that presenilin and β-secretase and related proteins may be in a common axonal vesicle with APP suggest that these processes may be tied together in a way that does not rely on APP processing to generate Aβ. For example, one possible mechanism of presenilin influence on the axonal transport machinery may be via control of kinesin light chain phosphorylation by glycogen synthase-3β (GSK-3β) or other kinases. A related and surprising finding is that overexpression of β-secretase, which is another protease that may generate Aβ from APP, can generate defects in axonal transport of APP, and may shift the cellular sites of Aβ production and secretion. These ideas have been hotly debated in the literature; further evidence is required to establish or refute these proposals.

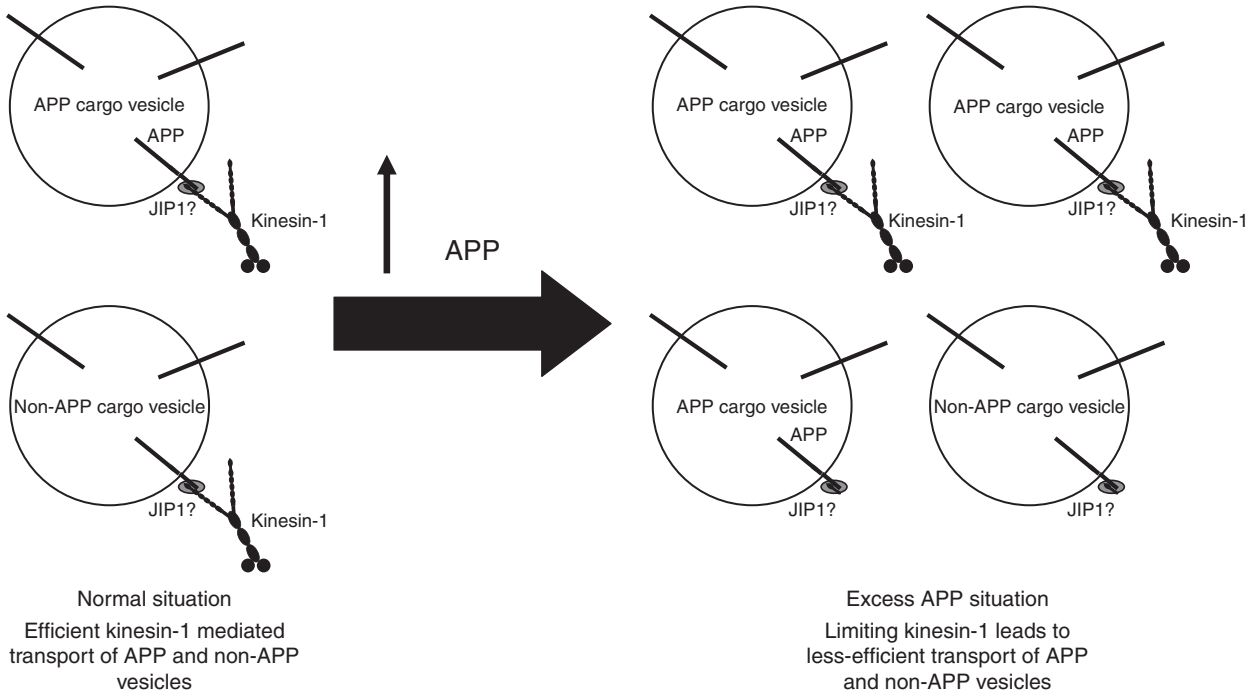

Figure 4 A possible mechanism by which excess amyloid precursor protein (APP) poisons axonal transport. In the normal situation, there is enough membrane-bound kinesin-1 to move APP and non-APP vesicular and organellar cargoes. In the presence of excess APP, competition for limiting kinesin-1 may occur so that some vesicles (intermittently) lack kinesin-1 and can stall in the axon, leading to axonal blockages and transport deficits. Although this schematic shows kinesin-1 as the limiting protein, other proteins or adapters, such as Jun N-terminal kinase-interacting protein 1 (JIP1), could be the vulnerable limiting component.

Another possible link of axonal transport to the pathogenesis of AD is via tau protein. This protein, which aggregates into neurofibrillary tangles, appears to play a normal role in the control of vesicle transport along microtubules, perhaps most prominently for vesicles transported by kinesin-1. There is considerable evidence that tau protein, which binds to microtubules, may directly regulate the movement of vesicles by kinesin-1 along microtubules. This function, in addition to the potential role of tau in initiating the polymerization of microtubules or in enhancing the stability of microtubules, may play important roles in the effects of tau protein and its mutants or of inappropriate phosphorylation variants on the axonal transport machinery. There is also evidence that overexpression of tau protein and its disease-causing mutations poison axonal transport directly.

Another important risk factor for AD is traumatic brain injury. There is evidence that injury to the brain induces so-called diffuse axonal injury, among other consequences. Diffuse axonal injury is of particular interest because it is effectively injury to the axons, which has been reported to result in large accumulations of APP and Aβ peptides at injury sites in axons. The possibility that axonal injuries cause biochemical changes in the axonal transport machinery is likely, given recent evidence that injury to axons induces activation of the damage-signaling kinase, Jun kinase (JNK). JNK is known to phosphorylate tau protein, and has been implicated in some aspects of axonal transport regulation. Thus traumatic brain injury may exert its influence on axonal transport defects via JNK activation and phosphorylation of targets such as tau and perhaps the kinesin heavy chain subunit of kinesin-1.

Another possible risk factor for AD is cholesterol. It is known that genotypic variants at the apolipoprotein E gene in humans have a large influence on the probability of developing late-onset AD. The $\varepsilon 4$ variant confers the highest risk and the $\varepsilon 2$ variant is protective. There is accumulating evidence that cholesterol and phospholipids might have important regulatory effects, not only on vesicular traffic, but also on kinesin-mediated axonal transport. One important kinesin family that is known to be essential for the transport of synaptic vesicle proteins is kinesin-3. This type of kinesin motor can bind to phospholipids in membrane bilayers, and has been reported to be activated by lipid clustering in vesicular membranes via oligomerization of the motor protein. Since cholesterol may have an impact on the ability of phospholipids to cluster, such a mechanism might affect transport directly. Another possible linkage of cholesterol to AD comes from the pediatric disease called Neimann–Pick

C type 1 (NPC1), which is a disorder of cholesterol transport and metabolism caused by defects in the *NPC1* gene. Such mutations appear to interfere with cholesterol trafficking in nonneuronal cells, possibly via the endocytic pathway. It is less clear what NPC1 mutations do in neurons, but it is clear that the mutations induce significant neurodegeneration in patients who have them, and ultimately lead to death. The striking clue is that these mutations appear to generate excess Aβ peptides and to induce the aberrantly phosphorylated forms of tau protein that lead to the formation of neurofibrillary tangles. Further work will be necessary to establish the precise molecular link between cholesterol trafficking and AD. It is intriguing, however, that defects in movement of cholesterol and related components in axons may be an important part of the problem.

Finally, Aβ peptide has been reported by a number of groups to induce phenotypes in cultured neurons that mimic those seen in neurons bearing axonal transport mutations. One group has reported that this may be a result of the effects of Aβ peptides on the state of microtubules; this influence appears to be mediated by extracellular Aβ peptides. Further work is necessary to clarify this association.

Role of Axonal Transport Failure in AD

The key issue is whether axonal transport defects are an early or late event in AD, and whether they are a consequence of Aβ toxicity, or a stimulus for Aβ generation or turnover, or both. Although further work is clearly required to resolve these issues, a testable type of hypothesis, presented in Figure 5, encapsulates some key features. Key elements are that axonal transport defects can induce synaptic defects directly, activate kinases that may induce tau hyperphosphorylation, and instigate or enhance Aβ

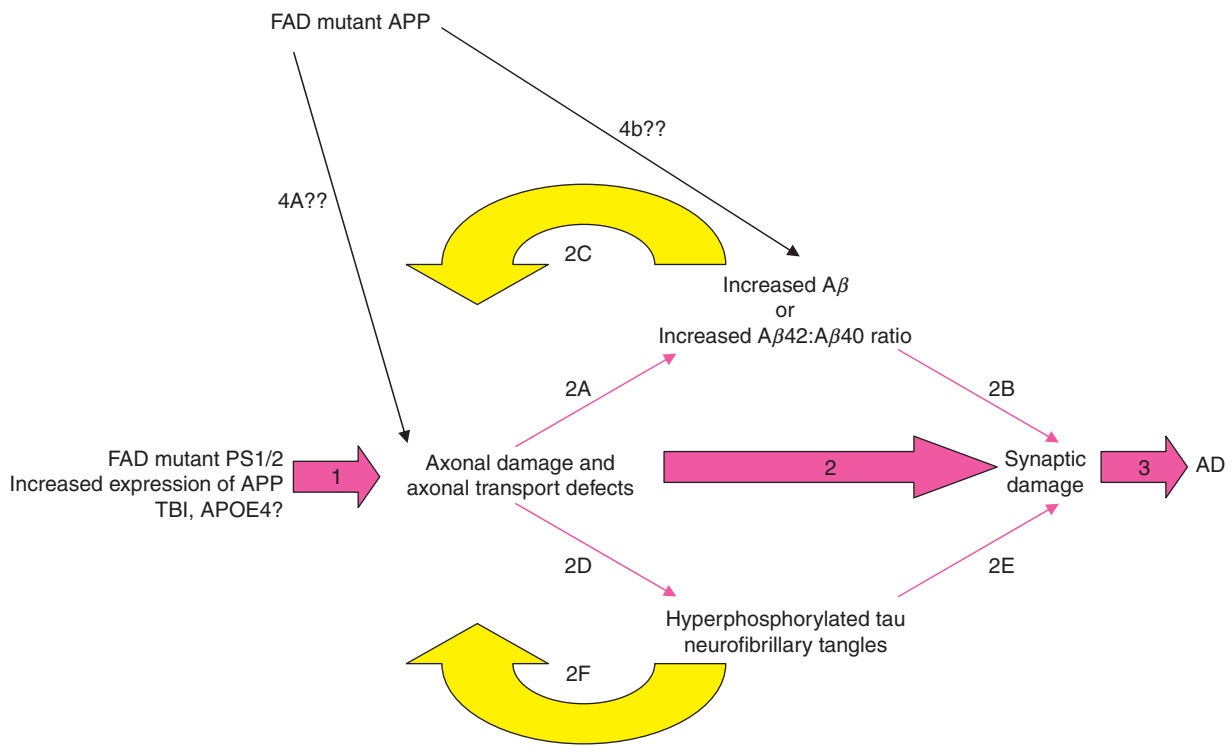

Figure 5 One possible alternative to the amyloid cascade hypothesis that incorporates axonal transport deficits into early stages of Alzheimer's disease (AD) causation or progression. (1) In familial Alzheimer's disease (FAD), mutant presenilins (PS1/2), excess amyloid precursor protein (APP), traumatic brain injury (TBI), and apolipoprotein E ε4 (APOE4) are suggested to initiate neuronal damage by interfering with axonal transport and/or causing axonal damage. (2) Axonal damage and axonal transport deficits may cause synaptic damage directly, or indirectly by the indicated alternative pathways. Axonal damage might lead to increased amyloid beta or increased ratio of the two types (Aβ_{40}, Aβ_{42}) (2A), which may cause synaptic damage directly (2B), or may lead to further axonal damage and transport defects in an autocatalytic loop (2C). Axonal damage may also lead to activation of Jun N-terminal kinase, which could initiate hyperphosphorylation of tau protein and formation of neurofibrillary tangles (2D), which may cause synaptic damage (2E) and cause further defects in axonal transport and axonal damage (2F) in an autocatalytic loop. Synaptic damage caused by all of these insults may be the primary neuronal defect causing cognitive symptoms in AD (3). How FAD mutant APP might initiate these proposed events is unknown, but could happen either by directly interfering with axonal transport or causing axonal damage (4A) or by increasing Aβ formation (4B).

formation, which may have its own toxic consequences. Unique aspects of this class of hypothesis include an autocatalytic spiral of events and multiple entry points, including mutation and environmental insults. In at least one form of hereditary AD, it can be argued strongly that this hypothesis may have validity in view of evidence that moderate increases in APP expression can induce significant axonal transport deficits in the absence of human Aβ. Although these ideas focus on axons, it is likely that comparable phenomena and behaviors exist in dendrites, for which less information is currently available.

Conclusion and Perspectives

The fundamental problem in evaluating whether defects in axonal transport can induce or significantly accelerate early stages of AD derives from the nature of the experimental systems. Only humans develop true AD, but axonal transport and biochemical defects induced by agents and mutations that can cause AD can only be rigorously studied at present in animal models of the disease. Further resolution of these issues will require noninvasive imaging suitable for studying axonal transport in living human patients, and the development of true human neuronal models of disease, perhaps using human embryonic or other types of stem cells. Finally, it is likely that truly effective therapies for AD will require a detailed and mechanistic understanding of the events leading from causative insults to the biochemistry of neuronal function and viability; axonal transport appears to be an important process in disease progression, and its role needs to be rigorously evaluated.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Aging of the Brain and Alzheimer's Disease; Animal Models of Alzheimer's Disease; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and Neurodegenerative Diseases; Dementia; Microtubule Associated Proteins in Neurons; Microtubules: Organization and Function in Neurons; Slow Axonal Transport; Stroke: Neonate vs. Adult.

Further Reading

Goldstein LS (2003) Do disorders of movement cause movement disorders and dementia? *Neuron* 40: 415–425.

King ME, Kan HM, Baas PW, et al. (2006) Tau-dependent microtubule disassembly initiated by prefibrillar beta-amyloid. *Journal of Cell Biology* 175: 541–546.

Lee EB, Zhang B, Liu K, et al. (2005) BACE overexpression alters the subcellular processing of APP and inhibits Abeta deposition *in vivo*. *Journal of Cell Biology* 168: 291–302.

Muresan Z and Muresan V (2005) c-Jun NH_2-terminal kinase-interacting protein-3 facilitates phosphorylation and controls localization of amyloid-beta precursor protein. *Journal of Neuroscience* 25: 3741–3751.

Pigino G, Morfini G, Pelsman A, et al. (2003) Alzheimer's presenilin 1 mutations impair kinesin-based axonal transport. *Journal of Neuroscience* 23: 4499–4508.

Satpute-Krishnan P, DeGiorgis JA, Conley MP, et al. (2006) A peptide zipcode sufficient for anterograde transport within amyloid precursor protein. *Proceedings of the National Academy of Sciences of the United States of America* 103: 16532–16537.

Shen J and Kelleher RJ, III (2007) The presenilin hypothesis of Alzheimer's disease: Evidence for a loss-of-function pathogenic mechanism. *Proceedings of the National Academy of Sciences of the United States of America* 104: 403–409.

Stokin GB and Goldstein LS (2006) Axonal transport and Alzheimer's disease. *Annual Review of Biochemistry* 75: 607–627.

Stokin GB, Lillo C, Falzone TL, et al. (2005) Axonopathy and transport deficits early in the pathogenesis of Alzheimer's disease. *Science* 307: 1282–1288.

Terry RD (1996) The pathogenesis of Alzheimer disease: An alternative to the amyloid hypothesis. *Journal of Neuropathology & Experimental Neurology* 55: 1023–1025.

Axonal Transport and Huntington's Disease

F Saudou and S Humbert, Institut Curie and CNRS UMR 146, Orsay, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Huntington's disease (HD) is a dominantly inherited neurodegenerative disease caused by abnormal polyglutamine (polyQ) expansion in a specific protein. HD is characterized by uncontrolled movements (chorea), personality changes, and dementia. Patients typically die 10–20 years after the appearance of the first clinical symptoms. In HD, the specific dysfunction and death of neurons in the striatum and cerebral cortex has been reported. The mutated protein is huntingtin, a large ubiquitous protein, the function of which is not fully understood. However, there is evidence that huntingtin is an antiapoptotic protein. Indeed, it is necessary for development since knockout mice die soon after gastrulation and conditional knockout mice, in which huntingtin is selectively turned off in brain and testis during adulthood, show a progressive degenerative neuronal phenotype and sterility. Moreover, overexpression of normal huntingtin protects striatal cells from a variety of apoptotic stimuli. Huntingtin becomes toxic when it contains an abnormal polyQ expansion. PolyQ-huntingtin induces the formation of neuritic and intranuclear inclusions, neuronal dysfunction, and, finally, neuronal death. The precise mechanisms underlying these phenomena are not well understood, and it is not known how increased neuronal death in the brain relates to huntingtin function and dysfunction.

Evidence suggests that huntingtin is involved in intracellular transport and that axonal transport is impaired in HD. Herein, we review studies demonstrating that huntingtin belongs to a protein complex that controls intracellular transport and that the alteration of this complex may participate in the pathogenesis of HD.

Intracellular Transport and Neurons

Neurons are highly polarized cells composed of a cell body and an extensive network of cell processes: dendrites, and a single axon. The axonal extension can reach up to 1 m in length. Most protein synthesis is restricted to the cell body. Therefore, active transport is required to supply the axon with newly synthesized materials. Conversely, external signals have to be transmitted to the cell body from the cell's extremities, where the signaling cascade is activated. Here also, active transport of receptor–ligand complexes is required for a quick and appropriate response. Due to the length of their processes, transport efficiency is of particular importance in neurons.

To allow efficient communication between cell bodies and axon termini, molecular motor proteins continuously shuttle vesicles and organelles along the microtubules and actin filaments that make up the cellular cytoskeleton. This transport process mostly involves microtubules, which are polarized structures with 'plus' (cell extremity) and 'minus' (cell body) ends. Molecular motors are considered to be unidirectional: Dynein complexes are connected to retrograde transport (from plus to minus end), whereas kinesins are connected to anterograde transport (from cell body to plasma membrane).

These two classes of motor complexes are required for efficient transport in neurons. Disruption of these complexes has dramatic consequences on neurons and leads to severe diseases.

Huntingtin and Intracellular Dynamics

There is growing evidence to suggest that huntingtin plays a role in intracellular dynamics. Cytoplasmic huntingtin co-localizes with microtubules and interacts with β-tubulin. Huntingtin is also found in neurites co-localizing with vesicles. In addition to this specific distribution, huntingtin interacts with many proteins that are implicated in intracellular trafficking. The best-described example is the case of huntingtin-associated protein 1 (HAP1), which is discussed later. Huntingtin also interacts with huntingtin-interacting protein 1 (HIP1), which binds to the clathrin light chain and regulates clathrin assembly, linking huntingtin to the control of endocytosis. A role of huntingtin in the control of endocytosis is further supported by the observation that huntingtin bound to the HAP40 protein acts as an effector of the small guanosine triphosphatase Rab5. In the pathological situation, a reduction of endosomal motility and endocytic activity in HD fibroblasts and mutant cells is observed. Finally, huntingtin associates with PACSIN1, a protein that is part of the endocytic machinery, which could act as a negative regulator.

In addition to regulating endocytosis at the plasma membrane, huntingtin may also regulate secretion and trafficking of proteins or organelles from the Golgi region. Indeed, HIP14, another huntingtin interacting protein that contains an ankyrin domain, is found at the Golgi and at vesicles. This protein is a palmitoyl

transferase involved in the sorting of proteins from the Golgi region. Also, huntingtin binds optineurin, a protein associated with Rab8 and myosin VI. This complex is found at the Golgi and in vesicles. The depletion of optineurin leads to the disruption of the Golgi ribbon structure and to the reduction in the transport of VSV-G from the Golgi complex to the plasma membrane. Finally, two other proteins that are dysregulated in HD participate in the sorting of organelles from the Golgi complex. The DNAJ-containing HSJ1b protein, possibly by regulating clathrin uncoating, promotes the processing of vesicles that contain the brain-derived neurotrophic factor (BDNF) from the Golgi to the cytoplasm. HSJ1b protein colocalizes with the transglutaminase 2 (TGase 2) enzyme that has a negative role in BDNF sorting from the Golgi. Interestingly, TGase 2 is inhibited by cystamine and by its reduced form, cysteamine, resulting in an increase in the sorting of BDNF-containing vesicles from the Golgi and in the release of this trophic factor from the cell. Treating HD mice with cysteamine results in an increase in BDNF release in the brain and in the reduction of the loss of striatal neurons in HD mutant mice. That a compound such as cysteamine, which is a Food and Drug Administration-approved drug, is neuroprotective by enhancing the intracellular processing of vesicles suggests that strategies aiming at restoring the defects in intracellular trafficking of organelles are of therapeutic interest.

Huntingtin Associates with the Molecular Motor Complex

The identification of HAP1 and its characterization with respect to huntingtin led to the hypothesis that both proteins may have an important function in the axonal transport machinery. Indeed, huntingtin and HAP1 are transported both anterogradely and retrogradely in rat sciatic nerves. Both proteins are associated with synaptic vesicles and localize with microtubules in neurites. The discovery in 1997 that HAP1 associates with the $p150^{Glued}$ subunit of dynactin further supported this possibility. Since then, a growing body of evidence has suggested that huntingtin has a major function in the control of axonal transport and that this function requires the HAP1 protein. Indeed, HAP1 binds not only to the $p150^{Glued}$ subunit of dynactin but also to the kinesin light chain 2, a subunit of the kinesin-1 complex. In agreement, depletion of HAP1 reduces kinesin-dependent transport of amyloid precursor protein vesicles. In addition to the indirect association of huntingtin to dynactin through HAP1, huntingtin also interacts with the dynein light chain of the dynein complex. Finally, huntingtin and HAP1 comigrate in the same fractions after sucrose gradient fractionation, further supporting the idea that huntingtin belongs to the molecular motor complex dynein/dynactin/kinesin-1.

Huntingtin Stimulates Axonal Transport

In addition to the previously discussed biochemical evidence, videomicroscopy approaches in cells and in *Drosophila* have further established a functional role of both HAP1 and huntingtin in axonal transport. In *Drosophila*, reduction in huntingtin protein level by the RNAi approach results in axonal transport defects. Furthermore, huntingtin directly promotes intracellular transport of vesicles along microtubules in neuroblastoma and striatal cells. Indeed, using BDNF as a marker of intracellular trafficking in cells, Gauthier and collaborators showed that expression of huntingtin enhances the velocity of BDNF-containing vesicles while reducing the percentage of time spent pausing. In support of a positive role of huntingtin in stimulating axonal transport of vesicles that contain BDNF, downregulation of huntingtin by RNAi approaches resulted in a decrease in the velocity of moving vesicles and an increase in the percentage of time spent pausing. BDNF was chosen because it is an important factor in HD. It is produced in the cortex and is transported to the striatum, the major site of degeneration in HD, where it supports neuronal differentiation and survival. It inhibits polyQ–huntingtin-induced neuronal death and its level is abnormally low in HD patients. BDNF is synthesized from the large precursor protein pre-pro-BDNF that is proteolytically processed and moves through the Golgi apparatus to the trans-Golgi network, where it is packaged into vesicles. BDNF-containing vesicles are then transported along microtubules (MTs) to the plasma membrane and subsequently released through the regulated secretory pathway. BDNF-containing vesicles are immunopositive for the classical markers of secretion and their activity-dependent release requires an intact MT network because it is blocked by nocodazole, an MT depolymerizing agent.

The huntingtin-dependent transport of BDNF vesicles is bidirectional because huntingtin stimulates both anterograde and retrograde transport in axons. It requires the HAP1 protein. Indeed, short N-terminal fragments of huntingtin that do not contain the HAP1 interacting region are unable to stimulate intracellular transport. Also, BDNF transport is reduced after downregulation of HAP1 protein and, in these conditions, huntingtin is unable to enhance BDNF trafficking. The HAP1-dependent transport is not restricted to BDNF vesicles since the movement of amyloid precursor protein vesicles is also reduced by decreasing HAP1 levels. Since it is known that the movement of amyloid precursor protein vesicles is kinesin dependent, one can infer an important role

of huntingtin and HAP1 in both anterograde and retrograde transport.

Although these studies revealed a role for huntingtin in axonal transport, the extend to which axonal transport depends on huntingtin remains unknown. Huntingtin stimulates the dynamic of BDNF-containing vesicles; however, whether other types of vesicles are regulated by huntingtin remains to be established. The observations that transport of amyloid precursor protein vesicles depends on HAP1 and that downregulation of huntingtin alters the general axonal transport in *Drosophila* strongly suggest that huntingtin regulates the transport of other small vesicles.

Milton, a *Drosophila* ortholog of HAP1, participates in the axonal transport of mitochondria, thus raising the possibility that huntingtin and HAP1 could also regulate the transport of mitochondria. However, the velocity of mitochondria is not regulated by huntingtin, and whereas HAP1 overexpression leads to the redistribution of BDNF vesicles in cells, it has no effect on mitochondria. Conversely, Milton, although known to redistribute mitochondria in cells, has no effect on BDNF vesicles. Therefore, HAP1 and Milton show specificity in the type of cargoes they are transporting.

HD Pathological Situation and Axonal Transport Defects

Several studies indicate an alteration of axonal transport in the polyQ situation. However, depending on the stage of the disease, different mechanisms may be involved.

In striatal cells and in neurons, whereas normal full-length huntingtin stimulates BDNF transport, full-length huntingtin with an abnormally expanded polyQ expansion has lost this stimulatory ability. Also, a reduction in transport of BDNF is observed in striatal cells derived from knockin mice in which a CAG expansion has been inserted in the endogenous mouse huntingtin gene. These mutant striatal cells express full-length huntingtin at endogenous levels. Similarly, polyQ-containing polypeptides encompassing amino acids 1–548 of huntingtin inhibit fast axonal transport in isolated axoplams without the formation of detectable morphological aggregates. Thus, polyQ expansion of huntingtin results in a loss of huntingtin function in transport. This phenotype is mediated by HAP1, which forms a protein complex by binding to the $p150^{Glued}$ subunit of dynactin and simultaneously binding to a domain located in the N-terminal region of huntingtin. The formation of this complex subsequently stimulates microtubule-based transport. In contrast, when huntingtin contains an abnormal polyQ expansion, it interacts more strongly with HAP1 and $p150^{Glued}$, leading to the detachment of the molecular motors from the microtubules and thus to less efficient transport of BDNF vesicles.

The presence of an abnormal polyQ expansion in huntingtin leads to a reduced anterograde and retrograde transport of BDNF-containing vesicles. In cortical neurons that deliver BDNF to striatal neurons, the reduced transport of BDNF in the pathological situation has a direct consequence on the ability of these cells to release BDNF, rendering striatal neurons particularly vulnerable. In agreement, cortical ablation of BDNF in mice induces striatal dendrite deficits followed by neuronal loss. Once released, BDNF binds to its receptor TrkB; the BDNF–TrkB complex is endocytosed in striatal neurons and transported retrogradely to the nucleus, where it activates downstream signaling pathways. The consequences of axonal transport defects in these neurons remain to be established.

A critical step in HD pathogenesis is the cleavage of polyQ–huntingtin into N-terminal fragments containing the polyQ expansion. These fragments translocate into the nucleus and induce the formation of aggregates. In the nucleus, polyQ–huntingtin fragments cause neuronal death by a gain-of-function mechanism, leading to the dysregulation of transcriptional activity. In addition, N-terminal huntingtin fragments and their aggregates accumulate in axonal processes and terminals. What are the consequences of these aggregates on axonal transport? In *Drosophila*, N-terminal huntingtin polypeptide fragments containing the polyQ expansion accumulate in axonal inclusions and cause axonal trafficking defects. This subsequently induces neuronal death and organismal death. Furthermore, mutant huntingtin aggregates alter vesicular and mitochondrial transport in mammalian neurons. These aggregation-induced defects in transport may be due to a physical blockage of vesicles but may also involve the sequestration by mutant huntingtin aggregates of motor proteins (particularly $p150^{Glued}$ and kinesin heavy chain) from other cargoes and pathways. Interestingly, this aggregation-induced alteration of general transport appears to be independent of the protein context since it has also been reported to occur in other polyQ disorders.

In summary, in early stages of the disease, the role of huntingtin in transport is lost following the alteration of the huntingtin (soluble form)/HAP1 interaction. In later stages, polyQ–huntingtin forms neuritic aggregates that contribute to a trafficking defect by a gain-of-function mechanism.

Disruption of Axonal Transport and Neurodegenerative Disorders

Transport defects have also been described in other polyQ neurodegenerative disorders. The polyQ-containing androgen receptor, the mutation of which is

responsible for spinobulbar muscular atrophy, inhibits fast axonal transport. Non-polyQ neurodegenerative diseases have also been linked to transport failure. In Alzheimer's disease (AD), mutation of the amyloid precursor protein, responsible for familial forms of AD, disrupts axonal transport. Similarly, overexpression of tau, a neuronal microtubule-associated protein that accumulates in AD, inhibits organelle trafficking. Several different neurodegenerative disorders are caused by mutation within motor proteins. For instance, Charcot–Marie–Tooth disease type 2 is caused by a loss-of-function mutation in the kinesin KIF1B, and hereditary spastic paraplegia type 10 is caused by a missense mutation in KIF5A. Finally, a mutation in the $p150^{Glued}$ subunit of dynactin has been identified in a family with an autosomal dominant form of lower motor neuron disease. Furthermore, inhibition of dynein-mediated axonal transport causes neurodegeneration in mouse models.

Conclusion

Slowing of transport might be a general phenomenon in neurodegenerative diseases. It can be either causative or contributory, depending on the disease. Neurons are highly susceptible to dysregulation of transport. Indeed, because transport in neurons occurs over very long distances, a decrease in intracellular trafficking has dramatic consequences on their viability. The development of accurate models of intracellular transport in neurons should therefore help researchers understand the molecular mechanisms underlying these diseases.

Axonal transport is a promising target for designing and testing treatments to slow or block neurodegeneration. In the case of HD, compounds that enhance transport or rescue huntingtin dysfunction might be of therapeutic interest since huntingtin directly controls transport of the pro-survival factor BDNF. This is of utmost importance because no treatment currently exists for this devastating disorder.

See also: Axonal Transport Disorders; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and Neurodegenerative Diseases; Cell Replacement Therapy for Huntington's Disease; Huntington's Disease; Huntington's Disease: Neurodegeneration; Triplicate Repeats: Huntington's disease.

Further Reading

Altar CA, Cai N, Bliven T, et al. (1997) Anterograde transport of brain-derived neurotrophic factor and its role in the brain. *Nature* 389: 856–860.

Borrell-Pages M, Canals JM, Cordelieres FP, et al. (2006) Cystamine and cysteamine increase brain levels of BDNF in Huntington disease via HSJ1b and transglutaminase. *Journal of Clinical Investigation* 116: 1410–1424.

Cattaneo E, Zuccato C, and Tartari M (2005) Normal huntingtin function: An alternative approach to Huntington's disease. *Nature Reviews Neuroscience* 6: 919–930.

Chang DT, Rintoul GL, Pandipati S, and Reynolds IJ (2006) Mutant huntingtin aggregates impair mitochondrial movement and trafficking in cortical neurons. *Neurobiology of Disease* 22: 388–400.

Gauthier LR, Charrin BC, Borrell-Pages M, et al. (2004) Huntingtin controls neurotrophic support and survival of neurons by enhancing BDNF vesicular transport along microtubules. *Cell* 118: 127–138.

Gunawardena S, Her LS, Brusch RG, et al. (2003) Disruption of axonal transport by loss of huntingtin or expression of pathogenic polyQ proteins in *Drosophila*. *Neuron* 40: 25–40.

Hafezparast M, Klocke R, Ruhrberg C, et al. (2003) Mutations in dynein link motor neuron degeneration to defects in retrograde transport. *Science* 300: 808–812.

Heerssen HM, Pazyra MF, and Segal RA (2004) Dynein motors transport activated Trks to promote survival of target-dependent neurons. *Nature Neuroscience* 7: 596–604.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews Neuroscience* 6: 201–214.

LaMonte BH, Wallace KE, Holloway BA, et al. (2002) Disruption of dynein/dynactin inhibits axonal transport in motor neurons causing late-onset progressive degeneration. *Neuron* 34: 715–727.

Lee WC, Yoshihara M, and Littleton JT (2004) Cytoplasmic aggregates trap polyglutamine-containing proteins and block axonal transport in a *Drosophila* model of Huntington's disease. *Proceedings of the National Academy of Sciences of the United States of America* 101: 3224–3229.

Li SH, Gutekunst CA, Hersch SM, and Li XJ (1998) Interaction of huntingtin-associated protein with dynactin P150Glued. *Journal of Neuroscience* 18: 1261–1269.

Li XJ and Li SH (2005) HAP1 and intracellular trafficking. *Trends in Pharmacological Sciences* 26: 1–3.

Pal A, Severin F, Lommer B, Shevchenko A, and Zerial M (2006) Huntingtin–HAP40 complex is a novel Rab5 effector that regulates early endosome motility and is upregulated in Huntington's disease. *Journal of Cell Biology* 172: 605–618.

Piccioni F, Pinton P, Simeoni S, et al. (2002) Androgen receptor with elongated polyglutamine tract forms aggregates that alter axonal trafficking and mitochondrial distribution in motor neuronal processes. *FASEB Journal* 16: 1418–1420.

Sahlender DA, Roberts RC, Arden SD, et al. (2005) Optineurin links myosin VI to the Golgi complex and is involved in Golgi organization and exocytosis. *Journal of Cell Biology* 169: 285–295.

Saudou F, Finkbeiner S, Devys D, and Greenberg ME (1998) Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. *Cell* 95: 55–66.

Szebenyi G, Morfini GA, Babcock A, et al. (2003) Neuropathogenic forms of huntingtin and androgen receptor inhibit fast axonal transport. *Neuron* 40: 41–52.

Trushina E, Dyer RB, Badger JD 2nd, et al. (2004) Mutant huntingtin impairs axonal trafficking in mammalian neurons *in vivo* and *in vitro*. *Molecular and Cellular Biology* 24: 8195–8209.

Yanai A, Huang K, Kang R, et al. (2006) Palmitoylation of huntingtin by HIP14 is essential for its trafficking and function. *Nature Neuroscience* 9: 824–831.

Axonal Transport and Neurodegenerative Diseases

S Roy, V M-Y Lee, and J Q Trojanowski, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The primary functions of the neuron are to receive, process, and transmit information. A typical neuron has multiple dendrites and a single elongated process, the axon. The dendrites and the cell bodies play a role in the collection and processing of information, and the axon is responsible for the transmission of information to other neurons via synapses. Whereas dendrites are usually in close proximity to the neuronal cell bodies and have some capacity for protein synthesis, axons can extend up to tremendous lengths, sometimes several feet, and are generally devoid of the protein synthetic machinery. As a result, axonal proteins are synthesized in the cell bodies and subsequently transported into the axons and synapses. This process is called axonal transport. Axonal transport is essential for the survival of axons and synapses, and it occurs throughout the life of the neuron.

The unique architecture of the axon combined with its dependence on the remotely located cell body for growth and maintenance makes it especially vulnerable to a variety of insults. Not surprisingly, axonal and synaptic defects are characteristic features of most neurodegenerative diseases, and it was long suspected that defects in axonal transport could play a major role in these diseases. Although correlative evidence linking axonal transport defects and neurodegenerative diseases was well known, recent experimental evidence from several laboratories suggests that many neurodegenerative diseases may be a direct consequence of altered or defective axonal transport. In this article, we first highlight the basic principles of the mechanisms of axonal transport. Then, we review some of the evidence that links axonal transport to neurodegenerative diseases, including pathogenic mutations in human genes that encode proteins involved in axonal transport and studies of animal models of disease showing impairments of axonal transport.

Basic Mechanisms of Axonal Transport

Although many specific details of axonal transport mechanisms are not fully understood, the basic principles are well established. In general, to move any cargo, two components are required – motors to move the cargo and rails to transport it. In addition, motor proteins often require adaptor/linker proteins to attach the motors to the cargo. For the rails, the neurons use the structural network of the cell, the cytoskeleton. Most long-range transport is dependent on the microtubules, but actin filaments also play a role in short-range movements. Microtubules in axons are polarized (unlike dendrites), allowing a directional bias in the movement of the motors and their cargo. For the motors, the neurons have an abundance of small molecular machines, known as kinesins and dyneins, moving the cargo mainly anterogradely (away from the cell body to the axon tip) or retrogradely (toward the cell body), respectively (Figure 1). The anterograde transport consists of components required for maintaining synapses, cytoskeletal proteins required for maintaining axonal structure and function, as well as a variety of other proteins required for neuronal function, including mitochondria, various chaperones, and glycolytic enzymes. The retrograde transport consists mainly of endosomal/lysosomal organelles that carry degraded proteins back to the cell bodies for degradation and also neurotrophic factors required for neuronal survival. In reality, however, most transport is bidirectional, with a bias either in the anterograde or in the retrograde direction.

Our understanding of the mechanisms of transport has increased dramatically during the past few decades due to advances in the development of model systems to dissect basic transport mechanisms and advances in biochemistry and genetics that have led to the identification and characterization of the molecular motors and their cellular functions. Model systems developed to visualize the transport process revealed that the bulk of proteins are transported anterogradely in two broad, discrete groups. Whereas some of the proteins are transported rapidly, at rates of 100–400 mm day^{-1} (1–5 μm s^{-1}), quickly reaching up to the tip of the axon, others move at rates that are several orders of magnitude slower, at approximately 0.2–5 mm day^{-1} (0.0002–0.05 μm s^{-1}). These two components are called fast axonal transport and slow axonal transport, respectively. Whereas fast transport mainly comprises vesicular cargo, including synaptic proteins, ion channels, and other components, slow axonal transport is primarily composed of transported cytoskeletal proteins, mainly microtubules, neurofilaments, and actin, along with many additional cytosolic proteins involved in neuronal homeostasis. Visualization of slow transport of the cytoskeleton revealed that despite moving with overall distinct dynamics, both slow and fast transport use similar

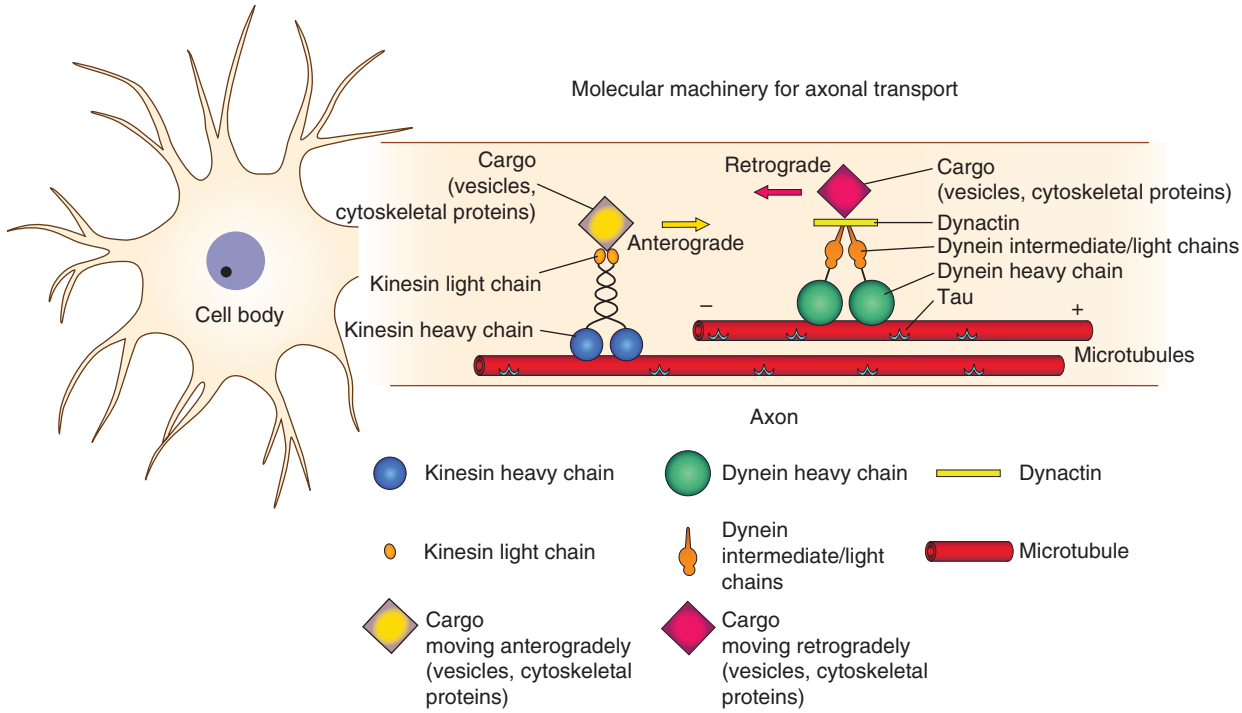

Figure 1 Basic principles of axonal transport. Cargo is transported anterogradely and retrogradely by the molecular motors kinesin and dynein, respectively. A variety of additional adaptor/linker proteins may also help in binding kinesin to the cargo (not shown). Dynactin, a multiprotein complex, links dynein to its cargo and is thought to increase the efficiency of the motor. Long-range movement occurs along cytoskeletal 'tracks,' mainly microtubules.

underlying mechanisms. These studies showed that slow cargo moves rapidly like fast cargo, but unlike fast cargo, it pauses for prolonged times during transit. Thus, the overall movement is infrequent and intermittent, leading to a slow overall rate of the transported cytoskeletal population. It has also been shown that slow and fast transport share similar molecular motors.

Progress in biochemistry and genetics has led to the identification of approximately 45 members of the kinesin superfamily (Kifs), grouped into 14 subfamilies, and two members of the cytoplasmic dynein family. Besides this large array of motor proteins, there are also a number of linker/adaptor proteins responsible for binding cargo to the motor proteins that have been identified. This heterogeneity is thought to play a major role in the recognition of specific motors to their cargo. Many excellent reviews on the subject are available.

Axonal Transport and Neurodegenerative Diseases

The following developments have dramatically highlighted the role of axonal transport disruptions in human neurodegenerative diseases: (1) the discovery of motor protein mutations in human neurodegenerative diseases, (2) axonal transport defects in animal and *in vitro* cellular models of neurodegenerative diseases, and (3) newly discovered roles in axonal transport regulation for known pathogenic proteins involved in neurodegenerative diseases. With a focus on specific diseases, we discuss some of the recent finings linking axonal transport to neurodegenerative diseases.

Kinesin Mutation in Hereditary Spastic Paraplegia

Hereditary spastic paraplegias (HSPs), also known as familial spastic paraplegias, represent an autosomal dominant inherited group of neurodegenerative diseases. Patients present in their thirties or forties with symptoms in the lower limbs, with gradual proximal spread of symptoms. Neuropathologically, a distal axonopathy is seen, with severe degeneration and gliosis of the distal corticospinal tracts and relative sparing of the tracts in the brain stem and proximal cord. Due to the peculiar distal-to-proximal 'dying back' axonopathy observed in this disease, it had been hypothesized that dysfunctions in axonal transport leading to selective damage of the distant portions of the axons may be responsible for the pathogenesis of HSPs.

Indeed, a missense mutation in one of the genes encoding a major kinesin protein (the gene for kinesin heavy chain Kif 5a) was found in a family with HSP. The same mutation was found in all affected

members of the family, as well as in some presymptomatic members. This mutation occurs within a functional motor domain of the kinesin protein and a homologous mutation in yeast has been found to decouple kinesin binding to microtubules, highlighting the functional role of the kinesin mutation in the pathology of HSPs.

Dynactin Mutations in Distal Spinal and Bulbar Muscular Atrophy

Dynactin is a large protein complex linked to the retrograde motor dynein, and it is thought to link the motor to its cargo and/or increase the processivity of the motor. Animal models disrupting the dynein/dynactin complex develop late-onset motor neuron degeneration, and missense mutations in a dynein subunit cause progressive motor neuron degeneration in mice. Due to the central role of the dynein/dynactin complex in axonal transport, and based on data from the animal studies, it was hypothesized that mutations in the dynein/dynactin complex could play a role in neurodegeneration. Indeed, mutations in the gene encoding a subunit ($p150^{Glued}$) of the dynactin complex have been reported in a family with the neurodegenerative disease distal spinal and bulbar muscle atrophy (SBMA). In familial SBMA cases, the disease is transmitted in an autosomal dominant fashion and is manifested as is a primary lower motor-type neuropathy with patients presenting in their thirties, often with breathing difficulty due to vocal fold paralysis, which later leads to weakness in the face and distal extremities. Neuron loss as well as inclusions containing dynein and dynactin were also seen in autopsy studies from one patient. The mutation reduces the binding affinity of dynactin to microtubules and also causes subtle defects in dynein function.

Kinesin Mutations in Charcot–Marie–Tooth Disease

Charcot–Marie–Tooth (CMT) disease comprises a heterogeneous group of inherited peripheral neuropathies characterized by motor and sensory deficits, often presenting in young adults as tingling, numbness, and loss of deep reflexes. The progression of the disease varies among individuals, with symptoms ranging from mild neuropathy to complete disability. Two basic forms can be recognized, with primary demyelinating (CMT1) or axonal (CMT2) types of degeneration predominating. Various genes have been implicated in CMT syndromes, including several genes known to play a role in myelination (PMP22 and MPZ) and genes for gap junction proteins (Connexin 32). However, in a remarkable series of studies, it was shown that mutations in a kinesin subunit protein (Kif 1B beta) can lead to an axonal type of CMT (CMT type 2A).

While studying animal models of kinesin knockout mice, it was found that heterozygous knockout mice for one of the kinesins, Kif 1B, developed progressive muscle weakness with normal motor nerve conduction velocities, symptoms resembling axonal type CMT, CMT type 2. Incidentally, the gene for CMT type 2A had been mapped to the same interval as the gene for Kif 1B, and genomic analysis of pedigrees with CMT type 2A revealed that these patients had a mutation in the Kif 1B gene. It was further shown that the mutant motor protein may not tightly bind to microtubules, thus suggesting a loss of function of the Kif 1B protein in patients with CMT type 2A.

Axonal Transport Defects in Alzheimer's Disease and Other Tauopathies

The story of the role of axonal transport in Alzheimer's disease (AD) is a rapidly developing one. Neuropathologically, the two hallmarks of AD are deposits of fibrillar Aβ into diffuse and neuritic plaques in the extracellular space, and filamentous accumulations of tau proteins as neurofibrillary tangles and neuropil threads within neurons and their processes. Amyloidogenic Aβ peptides are generated by proteolytic processing of the Aβ precursor proteins (APPs), conveyed by fast axonal transport. Several enzymes are involved in the proteolytic processing of APPs to Aβ, including the gamma- and beta-secretases (presenilins and BACE, respectively). Human mutations of both APPs and presenilins are seen in familial AD cases, highlighting the critical roles of these two proteins in AD. On the other hand, tau is a microtubule binding protein thought to stabilize microtubules *in vivo*.

The accumulation of tau in neuronal cell bodies and axons, as well as axonal swellings seen in AD, prompted the notion that axonal transport failure was important in the pathogenesis of this disease. Recent observations indicate that defects in axonal transport can occur long before the onset of severe symptoms in animal models of AD as well as in patients with AD. Interestingly, global reduction of axonal transport by reduction of kinesin levels can also exacerbate AD-type pathology in mouse models, further highlighting the role of diminished axonal transport in the pathogenesis of this disease.

The previous observations suggest that axonal transport defects play a role in the pathogenesis of AD. In addition, many key proteins directly involved in AD are also thought to play roles in the regulation of axonal transport. Studies of several different model systems have suggested that APP may act as a receptor for kinesin, thus proposing a direct link between a pathogenic protein and the axonal transport machinery. However, other studies have been unable to confirm

these results. It has also been proposed that APPs may be transported in a vesicular complex containing presenilins and BACE, the gamma- and beta-secretase enzymes, and altered processing of APPs within this complex can lead to exacerbation of AD. These findings suggest that misregulation of APPs, either directly from known APP mutations (as in familial AD) or indirectly via proteins associating/interacting with APPs, can lead to disruption of fast axonal transport in general, thus leading to axonal depletion of critical components and neurodegeneration.

Another interesting line of evidence highlighting the role of presenilins in the regulation of fast axonal transport in AD derives from studies of presenilin mutant mice. As mentioned previously, presenilins are proteins responsible for regulated proteolysis of APPs, and mutations in presenilins are seen in most cases of early familial AD. Several studies indicate that presenilins interact with glycogen synthase kinase-3β (GSK-3β). GSK-3β is a kinase with many different roles, including the phosphorylation of kinesin light chains, and it has been shown that GSK-3β-mediated phosphorylation of kinesin light chains led to detachment of the kinesin motor from the cargo, thus preventing further transport of cargo. By using transgenic presenilin mutant mice, it was also shown that mutant or absent presenilin increased GSK-3β levels, thereby phosphorylating kinesin light chains, detaching the kinesins from their cargo, and impairing axonal transport.

Roles in axonal transport regulation have also been assigned to tau. Because tau binds to microtubules and is thought to stabilize them, it was proposed that dysfunctions in tau can destabilize microtubules and lead to a failure of axonal transport. Indeed, human tau mutations impair the ability of tau to bind to and stabilize microtubules, and tau overexpression in cellular models can lead to defects in axonal transport. In addition, the mechanisms causing the accumulation of tau in AD are also beginning to be understood. Defective axonal transport of disease-associated mutant tau has also been demonstrated in mouse models, providing experimental evidence for a mechanism of tau accumulation.

A final line of evidence suggesting a role for axonal transport defects in AD comes from studies of ApoE4, a gene whose allelic state is associated with an increased risk for AD. Mice expressing human ApoE4 exhibit defects in axonal transport, and the receptor for ApoE4, ApoER2, binds to JIP1/2, a protein that appears to mediate the binding of APPs to kinesin I. Thus, it can be postulated that overexpression of ApoE4 protein can lead to misregulation of JIP1/2-mediated binding of kinesin to APPs, leading to defects in fast axonal transport.

Stabilization of microtubules is also being explored as a therapeutic option in AD. The proof of principle for this concept was demonstrated in a study that showed that a microtubule-stabilizing drug (paclitaxel) could ameliorate the neurodegenerative phenotype in transgenic tau mice by offsetting the loss of tau function by stabilizing the microtubules and correcting the fast axonal transport defects in these mice.

Axonal Transport and Huntington's Disease

Neuropathologically, Huntington's disease is characterized by atrophy and degeneration of striatal neurons, with aggregates of pathological polyglutamine containing the protein huntingtin. Huntingtin is a predominantly cytoplasmic protein that associates with vesicles and moves in the fast axonal transport component. Although it has been known for several years that polyglutamine repeats within the huntingtin protein cause a gain of deleterious function leading to neurodegeneration, the exact pathogenic role of the repeats is unclear. By infusing huntingtin-harboring pathological polyglutamine repeats into a model for studies of fast axonal transport, it was shown that fast axonal transport was specifically inhibited by pathologically expanded polyglutamine repeats (but not normal proteins), along with inhibition of neurite extension in cultured cells. Furthermore, disruption of the *Drosophila* huntingtin gene also caused axonal transport defects. These findings lend support to a model in which aggregates of polyglutamine repeats disrupt fast axonal transport. Whether the disruption of axonal transport is a direct effect of the polyglutamine repeats or a secondary phenomenon remains to be established.

Axonal Transport and Amyotrophic Lateral Sclerosis

The histopathologic observation of prominent neurofilament-rich inclusions in the axons of spinal motor neurons of patients with amyotrophic lateral sclerosis (ALS) led to the hypothesis that disrupted axonal transport of proteins may play a role in the pathogenesis of the disease. However, the first direct evidence that axonal transport is disrupted in ALS awaited the development of transgenic mouse models of familial ALS based on expression of mutant superoxide dismutase (SOD-1) protein mutant mice that replicate several key aspects of ALS. Studies of these SOD-1 transgenic mouse models of ALS showed that the transport of neurofilament proteins was retarded in these animals, even before the mice were symptomatic, thereby implicating impaired axonal transport as an early deficit in ALS. Mutations in the p150 subunit of dynactin have also been reported in ALS patients.

Axonal Transport and Synucleinopathies

Synucleinopathies, also known as α-synucleinopathies, are a group of neurodegenerative disorders in which the primary pathology is the intracytoplasmic accumulation of α-synuclein primarily in neurons and, in some cases, glial cells. These disorders include Parkinson's disease, dementia with Lewy bodies, the Lewy body variant of AD, multiple system atrophy, and neurodegeneration with brain iron accumulation. In familial forms of Parkinson's disease, autosomal dominant missense mutations in genes encoding for α-synuclein are seen, suggesting a role of α-synuclein in the pathogenesis of these disorders. α-Synuclein is a highly conserved protein belonging to a multigene family that includes β-synuclein and γ-synuclein. α-Synuclein is strongly expressed in neurons, highly enriched in presynaptic terminals, and transported predominantly in the slow component. Axonal transport abnormalities of α-synuclein have been proposed in synucleinopathies, based on the observation that axonal α-synuclein pathology is pronounced in the disease and also on experimental evidence suggesting that α-synuclein may play a role in transport of presynaptic vesicles. Age-related retardation in the normal transport of α-synuclein was also seen in a study. Collectively, these studies propose a model in which age-related retardation of α-synuclein transport leads to accumulations of the protein over time, predisposing to the α-synuclein pathology in axons. Although these findings are interesting, many questions remain unanswered. The physiological role of synuclein is far from clear, and much work needs to be done to identify the role of synuclein in neurodegenerative disorders.

Conclusions and Future Directions for Research

A growing body of evidence implicates axonal transport defects in the etiology and pathogenesis of neurodegenerative diseases. Although motor protein defects in neurodegenerative diseases are direct evidence for this, it is likely that many other disease proteins are directly or indirectly linked to the complicated machinery of axonal transport. Thus, studies on the molecular mechanisms of transport are necessary to facilitate our understanding of the role of axonal transport in these diseases. Since pathogenic proteins in AD have roles in the regulation of axonal transport, and these proteins have been extensively studied, the time is perhaps ripe to uncover the links between AD and impaired axonal transport. With greater understanding of the other neurodegenerative diseases, it is likely that many more links to axonal transport will be uncovered. Another exciting but largely neglected avenue of research is drug discovery efforts to counteract axonal transport impairments as therapeutic interventions for neurodegenerative diseases. Indeed, if axonal transport defects are shown to be part of a common mechanism of disease in many neurodegenerative disorders, the discovery of drugs that modulate axonal transport could result in the development of important therapeutic interventions for the treatment of these disorders.

See also: Axonal Transport Disorders; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and ALS; Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Hereditary Spastic Paraplegia; Slow Axonal Transport.

Further Reading

Brown A (2003) Axonal transport of membranous and nonmembranous cargoes: A unified perspective. *Journal of Cell Biology* 160: 817–821.

Chevalier-Larsen E and Holzbaur EL (2006) Axonal transport and neurodegenerative disease. *Biochimica et Biophysica Acta* 1762: 1094–1108.

Duncan JE and Goldstein LS (2006) The genetics of axonal transport and axonal transport disorders. *PLoS Genetics* 2: e124.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews Neuroscience* 6: 201–214.

Munch C, Rosenbohm A, Sperfeld AD, et al. (2005) Heterozygous R1101K mutation of the DCTN1 gene in a family with ALS and FTD. *Annals of Neurology* 58: 777–780.

Munch C, Sedlmeier R, Meyer T, et al. (2004) Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. *Neurology* 63: 724–726.

Puls I, Jonnakuty C, LaMonte BH, et al. (2003) Mutant dynactin in motor neuron disease. *Nature Genetics* 33: 455–456.

Puls I, Oh SJ, Sumner CJ, et al. (2005) Distal spinal and bulbar muscular atrophy caused by dynactin mutation. *Annals of Neurology* 57: 687–694.

Reid E, Kloos M, Ashley-Koch A, et al. (2002) A kinesin heavy chain (KIF5A) mutation in hereditary spastic paraplegia (SPG10). *American Journal of Human Genetics* 71: 1189–1194.

Roy S, Zhang B, Lee VM, and Trojanowski JQ (2005) Axonal transport defects: A common theme in neurodegenerative diseases. *Acta Neuropathologica (Berlin)* 109: 5–13.

Trojanowski JQ, Smith AB, Huryn D, and Lee VM (2005) Microtubule-stabilising drugs for therapy of Alzheimer's disease and other neurodegenerative disorders with axonal transport impairments. *Expert Opinion on Pharmacotherapy* 6: 683–686.

Zhao C, Takita J, Tanaka Y, et al. (2001) Charcot–Marie–Tooth disease type 2A caused by mutation in a microtubule motor KIF1Bbeta. *Cell* 105: 587–597.

Axonal Transport Disorders

P N Hoffman, P C Wong, J W Griffin, and D L Price, The Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Fast (Anterograde and Retrograde) Transport and Slow Transport

Membranous organelles, such as vesicles, are transported at rapid velocities (i.e., up to 400 mm day^{-1}) in both the anterograde (i.e., away from the cell body) and retrograde (i.e., toward the cell body) directions. For example, proteins critical for synthesis of transmitters and their uptake, vesicle trafficking, transmitter release and recycling, etc. are, for the most part, rapidly delivered to terminals where they are essential for events involved in synaptic transmission. Some protein moves bidirectionally. Trophic factor receptors (e.g., the Trk receptors) are transported as inactive (non-phosphorylated) proteins in an anterograde direction to terminals, whereas active Trk receptors (phosphorylated) and their ligands (nerve growth factor and brain-derived neurotrophic factor) are carried retrograde to cell bodies where they influence transcription and, ultimately, the biology and survival of neurons. Moreover, transported proteins can be involved in disease in different ways. APP, the amyloid precursor protein, is a type 1 transmembrane protein that is rapidly transported anterograde along β- and γ-secretases in the central nervous system and is the source, when cleaved by these secretases, of the amyloid β (Aβ) peptide. Neurotoxic Aβ peptides are released at synaptic terminals where they accumulate as Aβ oligomers (and amyloid) which interfere with synaptic communication. When mutated, dynactin, a motor protein involved in retrograde transport, is the cause of amyotrophic lateral sclerosis (ALS), a motor neuron disease (Table 1). Moreover, tetanus toxin and some viruses (poliovirus and rabies virus) enter terminals and are carried by retrograde transport to cell bodies where their activities cause diseases such as tetanus, polio, and rabies.

The Cytoskeleton and Motor Proteins

Tubulin and microtubule-associated protein, including tau, as well as specific motor complex proteins (kinesin, dynein, dynactin, etc.), are transported within axons and provide the pathways and mechanisms for the bidirectional movement of membranous organelles and other constituents (e.g., mitochondria) in axons. Microtubules, which provide the rails for much of transport, have a uniform polarity, with their plus ends directed toward axonal terminals. Similarly, dynein, dynactin, and other components of retrograde motors play critical roles in the transport of components from terminals toward cell bodies. Kinesins, dynein, dynactin, and other proteins play key roles in the movement of membranous organelles toward either the plus end (anterograde direction) or minus end (retrograde direction) of axonal microtubules. Anterograde transport relies on kinesins, which are members of a family of highly conserved motor proteins that are critical for intracellular trafficking and transport. In neurons, kinesin is the principal motor for anterograde transport, with kinesin-associated proteins mediating binding of cargo.

The cytoskeletal proteins (including actin, tubulin, and tau) move slowly down axons. Tubulin, actin, clathrin, dynein, dynactin, and glycolytic enzymes are members of slow component b (SCb) and move at a rate of 2–8 mm day^{-1}. Some tubulins and the neurofilament (NF) proteins, NF-L, -M, and -H, which play distinct roles in filament assemblies, appear to be slowly transported in the anterograde direction at velocities of 1–5 mm day^{-1}. Evidence is consistent with the idea that some of these proteins, including the NF proteins, move in the slow component because their movement, although sometimes rapid, is asynchronous and infrequent. Thus, the NF proteins pause, go, and pause – a phenomenon leading to an overall slow rate of transport. Kinesin and dynein and its partners may participate in forward and return movements of these proteins, respectively.

Axonal Injury

Axonal transection (acute axotomy) of peripheral nerves is the simplest model used to investigate the effects of injury on transport processes. Both fast anterograde and retrograde transport are interrupted, and rapidly moving membranous organelles (and bidirectionally moving mitochondria) accumulate at the cut ends of the proximal and distal nerve stumps. Subsequently, the distal stump degenerates (Wallerian degeneration (Wlds)), and regenerating sprouts extend from the proximal stump at the rate of 3–5 mm day^{-1}. Axonal regeneration requires the presence of cytoskeletal proteins and membranous elements supplied by 'slow' and 'fast' anterograde transport, respectively. After axonal transection, changes occur in the expression and transport of a variety of proteins, including those of the cytoskeleton. Elevated expression of actin and tubulin is associated with the increased delivery of these proteins to regenerating axons. At the same time, levels of NF proteins are reduced, resulting in

Table 1 Select genetic forms of ALS

Subtype	*Gene locus*	*Onset*	*Inheritance*
ALS1	*SOD1*/21q21	Adult	Autosomal dominant
ALS2	*ALS2*/2q33	Juvenile	Autosomal recessive
ALS3	18q21	Adult	Autosomal dominant
ALS4	*SNTX*/9q34	Juvenile	Autosomal dominant
ALS5	15q15	Juvenile	Autosomal recessive
ALS6	16	Adult	Autosomal dominant
ALS	*DCTN1*/2q13	Adult	Autosomal dominant
ALS8	20q13 *VAPD*	Adult	Autosomal dominant

a decrease in the amounts of NF proteins transported in the proximal stump. Because NF polymers play a major role, along with neuregulin signaling in determing axonal caliber in myelinated nerve fibers, the decrease in transport of NF proteins is associated with reductions in axonal caliber that originate in proximal axons (near the cell body) and propagate distally at rates comparable to the velocity of NF transport. This phenomenon is referred to as somatofugal axonal atrophy. After axonal injury, alterations in levels of the membrane-associated proteins transported in the fast component are more subtle with the notable exception of increased expression of GAP-43, a membrane-associated phosphoprotein involved in the guidance of growth cones. Alterations in gene expression, levels of specific proteins, and axonal transport usually return to normal after regenerating axons have reinnervated targets.

Disorders Associated with Abnormalities of Axonal Transport

Impairments of axonal transport have profound consequences for the structure and functions of nerve cells. Alterations in axonal transport can be classified by the time course (acute, subacute, or chronic), the character of the pathology (accumulations of membranous or cytoskeletal organelles, swelling or shrinkage of axons), the locations of abnormalities within axons (proximal vs. distal, as exemplified by proximal NF axonal swellings vs. distal dying back axonopathies), the regions of the nervous systems affected by disease (central nervous system or peripheral nervous system (PNS)), and types of neuron (motor, sensory, or other) vulnerable to disease.

Axonal degeneration occurs in a variety of disorders of motor and sensory neurons, including genetic disease (Charcot–Marie–Tooth disease, mutant *SOD1* and *DCTN1*-linked fALS, giant axonal neuropathy linked to mutations in the *GAN* gene, etc.), toxic neuropathies (*n*-hexane, vincristine, Taxol, and cisplatin), demyelinating diseases (Guillain–Barré syndrome), and neurodegenerative diseases such as ALS and Alzheimer's disease. Neuroaxonal dystrophy, a prototype of membranous neuropathy, is characterized by the presence of membranous organelles accumulating in the distal regions of axons. Similar pathology occurs in experimental disorders (vitamin E deficiency and intoxication with *p*-bromophenylacetylurea) and in a variety of human illnesses, including those that occur in childhood (infantile neuroaxonal dystrophy) and in aging (dystrophy of axons in the dorsal columns). It has been speculated that the accumulation of rapidly transported membranous materials may occur because of failure of local utilization, excessive anterograde delivery, or failure of recycling/removal (turnaround) by retrograde transport.

Several experimental and clinical axonopathies are associated with the pathology of the cytoskeleton, particularly the accumulation of NFs. Experimental exposure to β,β'-iminodipropionitrile (IDPN) is a prototype for this type of disorder. This toxin selectively blocks transport of NF, perhaps by altering interactions of NFs with microtubules or with anterograde motors. The transport of other cytoskeletal proteins, including actin and tubulin, is relatively unaffected, and fast anterograde and retrograde transport appear to be relatively normal. After systemic intoxication with IDPN, NF synthesis is not altered, and newly synthesized NF proteins continue to enter the axon normally. Because transport of these proteins along axons is markedly impaired, NFs accumulate and distend proximal axons (with secondary local demyelination). NF axonal pathology occurs in other experimental models, including aluminum intoxication; the overexpression of wild-type NF transgenes (altering the stoichiometry of NF protein); the expression of a mutant NF transgene which encodes a protein that fails to interact properly for assembly; and progressive motor neuronopathy, a disease in mice homozygous for mutations in *Tbce* (tubulin-specific chaperone E). Hereditary canine spinal muscular atrophy, an inherited motor neuron disorder in Brittany spaniels, shows severe accumulation of NF in proximal axons, closely resembling lesions seen in IDPN models and in ALS. Disorganization of the NF network occurs in humans with giant axonal neuropathy linked to mutations in *GAN*, which encodes gigaxonin, a member of the cytoskeletal BTB/Kelch family. Dynein and dynactin (*DCTN1*) play roles in retrograde axonal transport, and alterations in the biology of these proteins are associated with disease. For example, slowly progressive autosomal dominant ALS has been linked to a G59S mutation in the $P150^{glued}$ subunit of *DCTN1*. The dynactin complex, which includes P150 and dynamitin, provides a linker between cargo, microtubules, and dynein. In the index family, the mutation

occurs in a motif in the *P150^{glued}* subunit that mediates binding to microtubules. Modeling studies suggest that this mutation impacts on protein structure to create steric hindrance and distortion of the folding of the microtubule-binding domain. Consistent with this concept is the observation that the mutant protein binds less well to microtubules. Members of our group at Johns Hopkins have created a mutant *P150^{glued}* model of fALS with a robust motor neuron disease phenotype.

Axonal abnormalities are well documented in ALS (Lou Gehrig's disease). NF accumulates in the neuronal cell bodies and in proximal axons; the latter lesions resemble those produced by IDPN exposure and appear to reflect the impaired transfer of newly synthesized NF proteins from the cell body to the axon. Eventually, in ALS, axons degenerate and neurons die. Importantly, mutant *SOD1* fALS mice exhibit weakness and muscle atrophy associated with pathological features of human SOD1-linked ALS. These mice exhibit evidence of impaired axonal transport and develop proximal axonal swellings and more distal Wallerian degeneration followed by death of motor neurons. In a number of disorders, lesions occur more distally (NF swellings and accumulation of membranous organelles) and are often associated with distal axonal degeneration. These diseases are termed 'dying back' neuropathies because the distal axons/terminals exhibit conspicuous and early abnormalities distally and extend retrograde toward the cell bodies. Recent evidence suggests that similar processes occur in disease of the CNS.

Axon Regeneration and Protection

Intriguingly, in both experimental and human disease, regeneration and recovery can occur. The classical example of this process is regeneration after axotomy in the PNS. Usually, the distal axons of the PNS show Wallerian degeneration after transection, and eventually they regenerate successfully. Significantly, degeneration of these axons can be prevented genetically by the presence of an unusual gene. Significantly, mice (C57BL/Wlds) expressing a mutant gene (*Wlds*) exhibit slow Wallerian degeneration of axons distal to transections. The protective gene encodes an N-terminal fragment of ubiquitination factor E4B (Ube4b) fused to a nicotinamide mononucleotide adenyltransferase. The mechanism of this semidominant protection is not fully understood, and it has been suggested to be related to altered ubiquitination, changed purine nucleotide metabolism, or sirtuin-like activity. This unique genetic model provides an opportunity to discover new pathways whereby axons may be protected from degeneration.

See also: Amyotrophic Lateral Sclerosis (ALS); Axonal Injury: Neuronal Responses; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal mRNA Transport and Functions; Axonal Transport and ALS; Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Axonal Transport and Neurodegenerative Diseases; Axonal Transport Tracers; Microtubule Associated Proteins in Neurons; Slow Axonal Transport.

Further Reading

Bomont P, Cavalier L, Blondeau F, et al. (2000) The gene encoding gigaxonin, a new member of the cytoskeletal BTB/kelch repeat family, is mutated in giant axonal neuropathy. *Nature Genetics* 26(3): 370–374.

Borchelt DR, Wong PC, Becher MW, et al. (1998) Axonal transport of mutant superoxide dismutase 1 and focal axonal abnormalities in the proximal axons of transgenic mice. *Neurobiology of Disease* 5(1): 27–35.

Bruijn LI, Miller TM, and Cleveland DW (2004) Unraveling the mechanisms involved in motor neuron degeneration in ALS. *Annual Review of Neuroscience* 27: 723–749.

Carmeliet P and Tessier-Lavigne M (2005) Common mechanisms of nerve and blood vessel wiring. *Nature* 436(7048): 193–200.

Coleman MP and Perry VH (2002) Axon pathology in neurological disease: A neglected therapeutic target. *Trends in Neuroscience* 25(10): 532–537.

Cork LC, Griffin JW, Munnell JF, Lorenz MD, Adams RJ, and Price DL (1979) Hereditary canine spinal muscular atrophy. *Journal of Neuropathology & Experimental Neurology* 38: 209–221.

Ehlers MD, Kaplan DR, Price DL, and Koliatsos VE (1995) NGF-stimulated retrograde transport of trkA in the mammalian nervous system. *Journal of Cell Biology* 130(1): 149–156.

Gold BG, Griffin JW, and Price DL (1985) Slow axonal transport in acrylamide neuropathy: Different abnormalities produced by single-dose and continuous administration. *Journal of Neuroscience* 5: 1755–1768.

Griffin JW, Hoffman PN, Clark AW, Carroll PT, and Price DL (1978) Slow axonal transport of neurofilament proteins: Impairment by β,β'-iminodipropionitrile administration. *Science* 202: 633–635.

Hadano S, Hand CK, Osuga H, et al. (2001) A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nature Genetics* 29: 166–173.

Hafezparast M, Klocke R, Ruhrberg C, et al. (2003) Mutations in dynein link motor neuron degeneration to defects in retrograde transport. *Science* 300(5620): 808–812.

Hanz S, Perlson E, Willis D, et al. (2003) Axoplasmic importins enable retrograde injury signaling in lesioned nerve. *Neuron* 40(6): 1095–1104.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews Neuroscience* 6(3): 201–214.

Hoffman PN, Cleveland DW, Griffin JW, Landes PW, Cowan NJ, and Price DL (1987) Neurofilament gene expression: A major determinant of axonal caliber. *Proceedings of the National Academy of Sciences of the United States of America* 84: 3472–3476.

Hoffman PN, Griffin JW, and Price DL (1984) Control of axonal caliber by neurofilament transport. *Journal of Cell Biology* 99: 705–714.

Holzbaur ELF (2004) Motor neurons rely on motor proteins. *Trends in Cell Biology* 14(5): 233–240.

Koliatsos VE, Clatterbuck RE, Winslow JW, Cayouette MH, and Price DL (1993) Evidence that brain-derived neurotrophic factor is a trophic factor for motor neurons *in vivo*. *Neuron* 10(3): 359–367.

Lazarov O, Morfini GA, Lee EB, et al. (2005) Axonal transport, amyloid precursor protein, kinesin-1, and the processing apparatus: Revisited. *Journal of Neuroscience* 25(9): 2386–2395.

Lee MK, Marszalek JR, and Clevel and DW (1994) A mutant neurofilament subunit causes massive, selective motor neuron death: Implications for the pathogenesis of human motor neuron disease. *Neuron* 13: 975–988.

Mack TG, Reiner M, Beirowski B, et al. (2001) Wallerian degeneration of injured axons and synapses is delayed by a Ube4b/Nmnat chimeric gene. *Nature Neuroscience* 4(12): 1199–1206.

Martin N, Jaubert J, Gounon P, et al. (2002) A missense mutation in tbce causes progressive motor neuronopathy in mice. *Nature Genetics* 32(3): 443–447.

Price DL, Ackerley S, Martin LJ, Koliatsos VE, and Wong PC (2005) Motor neuron diseases. In: Brady ST, Siegel GJ, Albers RW and Price DL (eds.) *Basic Neurochemistry*, pp. 731–743. Burlington, MA: Elsevier.

Price DL, Griffin J, Young A, Peck K, and Stocks A (1975) Tetanus toxin: Direct evidence for retrograde intraaxonal transport. *Science* 188: 945–947.

Puls I, Jonnakuty C, LaMonte BH, et al. (2003) Mutant dynactin in motor neuron disease. *Nature Genetics* 33(4): 455–456.

Rowland LP and Shneider NA (2001) Amyotrophic lateral sclerosis. *New England Journal of Medicine* 344(22): 1688–1700.

Schroer TA (2004) Dynactin. *Annual Review of Cell and Developmental Biology* 20: 759–779.

Vallee RB, Williams JC, Varma D, and Barnhart LE (2004) Dynein: An ancient motor protein involved in multiple modes of transport. *Journal of Neurobiology* 58(2): 189–200.

Verpoorten N, DeJonghe P, and Timmerman V (2006) Disease mechanisms in hereditary sensory and autonomic neuropathies. *Neurobiology of Disease* 21(2): 247–255.

Williamson TL and Cleveland DW (1999) Slowing of axonal transport is a very early event in the toxicity of ALS-linked SOD1 mutants to motor neurons. *Nature Neuroscience* 2: 50–56.

Wong PC, Cai H, Borchelt DR, and Price DL (2002) Genetically engineered mouse models of neurodegenerative diseases. *Nature Neuroscience* 5(7): 633–639.

Wong PC, Pardo CA, Borchelt DR, et al. (1995) An adverse property of a familial ALS-linked SOD1 mutation causes motor neuron disease characterized by vacuolar degeneration of mitochondria. *Neuron* 14: 1105–1116.

Axonal Transport Tracers

L G Bilsland and G Schiavo, Cancer Research UK London Research Institute, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The formation of highly ordered and specialized neural networks during development is critical for the assembly of a functional nervous system. To gain a better understanding of these higher brain structures, it is necessary to correlate functional information with the intricate anatomy of specific neural circuitry, a process that becomes essential to evaluate changes in neural networks in circumstances of disease and injury. The discovery that exogenous proteins can be retrogradely transported in peripheral nerves initiated the development of neuroanatomical mapping techniques based on axonal transport tracing. The anatomical tracers now available can be classified according to their direction of axonal transport. Anterograde tracers are taken up by the cell soma or dendrites and transported down the axon to the nerve terminals. In contrast, retrograde tracers are taken up at the nerve terminals and transported toward the soma. Transport of tracers is predominantly in either an anterograde or a retrograde direction, although on closer examination it appears that most tracers are transported bidirectionally to some extent. In addition, several probes, termed transsynaptic tracers, are transferred from the original neuron to synaptically connected neurons. Neuroanatomical studies based on the detection of anterograde, retrograde, and transsynaptic tracers can provide a detailed knowledge of cellular projections and consequently morphology. Furthermore they allow investigation of neuronal connectivity that can divulge information on the organization and potential regulation of neural networks.

Early tracing studies were based on the induction of axonal degeneration. Following a brain lesion, axon degeneration was traced back to the cell soma by means of Golgi silver impregnation techniques. In the 1970s, advances in anterograde tracing included the injection of radiolabeled amino acids into specific brain regions. The anterograde transport of radiolabeled peptides and proteins was then monitored by autoradiographic techniques. This technique is still in use, although it has largely been superseded by tracers that require simpler detection methods and have improved anatomical resolution.

Anterograde and Retrograde Transport Tracers

Horseradish Peroxidase

The plant enzyme horseradish peroxidase (HRP) was one of the first retrograde tracers used. Application of HRP is generally by pressure or iontophoretic injection into the region of interest, where it is taken up by either damaged or intact axons and is endocytosed at nerve terminals via a fluid-phase mechanism. Histochemical detection of HRP is based on the production of a permanent electron-dense precipitate after HRP reacts with hydrogen peroxide and diaminobenzidine tetrahydrochloride (DAB) or tetramethylbenzidine. HRP labeling is, however, restricted to the cell soma and proximal dendrites. Conjugation of HRP with wheat germ agglutinin (WGA) significantly increases its ability to label axon terminals and allows transsynaptic transfer.

Fluorescent Dyes

The most frequently used axonal tracers are a range of fluorescent, hydrophilic dyes which include fast blue, rhodamine isothiocyanate, lucifer yellow, propidium iodide, and diamidino yellow. These dyes can be injected into the region of interest where they are taken up by fluid-phase endocytosis, or alternatively they can be microinjected directly into the cell soma. Several dyes can be used simultaneously to trace multiple cellular projections. Generally these dyes are bidirectionally transported, although some, such as fast blue and diamidino yellow, are predominantly transported in the retrograde direction. The majority of these probes are, however, not transsynaptically transferred. They are sensitive tracers, providing good resolution of cellular morphology. However, as with all fluorescence-based techniques, photobleaching can be a problem as well as saturating signals, which might obscure fine morphological details.

Fluorogold

Developed in 1986, the fluorescent retrograde tracer fluorogold has largely superseded the use of fluorescent dyes. Fluorogold can be applied by either pressure or iontophoretic injection into the region of interest, where it is taken up by damaged neurons and at nerve terminals by fluid-phase endocytosis. The advantage of this tracer lies in the availability of specific antibodies to fluorogold. Immunostaining can then be converted to a permanent electron-dense precipitate through the use of a biotinylated

secondary antibody followed by processing with an avidin–HRP conjugate and then reaction with DAB. Fluorogold provides sensitive staining of the axonal and somatodendritic compartments but does not undergo transcytosis.

Carbocyanine Dyes

This highly lipophilic family of dyes, which includes DiI (1,1′-dioctadecyl-3,3,3′,3′-tetramethylindocarbocyanine perchlorate) and DiO (3,3′-dioctadecyloxacarbocyanine perchlorate), are fluorescent membrane tracers. Carbocyanine dyes can be applied via pressure injection, placement of dye crystals in the region of interest, or biolistic methods and can also be used postfixative. These lipophilic dyes diffuse laterally ($2\,\mathrm{mm\,month^{-1}}$) through the phospholipid bilayers of the cell membrane, giving extensive labeling of cellular processes and hence detailed morphological information. Transsynaptic transfer of these dyes is not a common occurrence unless two cell membranes are very closely associated. *In vivo*, membrane fragments labeled with carbocyanine dyes can also become endocytosed into vesicles, which are then transported bidirectionally, but not transsynaptically. Thus *in vivo*, the detailed membrane labeling typical of carbocyanine dyes is replaced with a granular cytoplasmic labeling that prevents good resolution of cellular processes.

Biocytin

Extensive labeling of neuronal arborization can also be achieved with biocytin, a peptide derivative of biotin. Applied either by iontophoretic injection or as a pellet, biocytin is actively internalized and rapidly transported, predominantly in the anterograde direction, although with large injection volumes, retrograde transport is also seen. The high affinity of biocytin for avidin enables detection either with avidin–HRP followed by DAB staining or through the use of fluorescent avidin conjugates. However, biocytin is rapidly catabolized and therefore is only suitable for short-term tracing.

Fluorescent Microbeads

Introduced in the 1980s, fluorescent microbeads consist of either polystyrene or latex beads that are labeled with dyes of different colors, permitting double labeling experiments. Pressure injection of the microbeads in the brain results in a highly restricted region of uptake, the mechanism of which has not been defined, although it is likely to be due to local phagocytosis. Uptake is followed by retrograde transport of the beads. The microspheres produce stable fluorescent labeling. However, the signal is restricted to the cell soma; thus, information on cellular projections is very limited.

Dextran Amines

Fluorescent or biotinylated dextran amines were first investigated as transport tracers in the late 1980s and early 1990s. Dextrans can be applied either by pressure or iontophoretic injection or via application of dye crystals to the region of interest, where they undergo fluid-phase endocytosis. Transport then occurs in either an anterograde or retrograde direction depending on the molecular weight of the dextran. Dextrans of greater than 10 kDa are predominantly anterogradely transported, whereas dextrans of molecular weight of 3 kDa are transported retrogradely on injection in an acidic vehicle. This differentiation based on molecular weight represents a major advantage over other bidirectional tracers, allowing clearer and unequivocal results. Following standard aldehyde fixation, these hydrophilic polysaccharides are easily detected either by fluorescence or by processing of biotinylated dextran amines with an avidin–HRP conjugate followed by DAB staining to generate an electron-dense product. Dextrans are highly sensitive, nontoxic tracers that provide detailed labeling of fine cellular arborization.

Cholera Toxin B Subunit

Cholera toxin (CT) is a bacterial protein toxin produced by *Vibrio cholerae*, which binds to cellular membranes with high affinity. CT is composed of an active subunit (A), which induces toxicity by continuously activating stimulatory G-proteins (Gs), thereby increasing cyclic adenine monophosphate levels, and a homopentameric binding subunit (B). The B subunit (CTB) binds to the monosialoganglioside GM1 that is clustered in membrane microdomains on the plasma membrane of different cell types, including neurons. Once bound at the nerve terminal, CT becomes internalized in endocytic vesicles and then undergoes retrograde transport (at a rate of $102\,\mathrm{mm\,day^{-1}}$). Recombinant CTB maintains the specific binding of full-length CT and is completely atoxic, thus constituting an ideal transport probe. CTB was first used as a retrograde tracer, but recently it has been found to also be a very sensitive anterograde transport marker that gives fine definition of neuronal processes and nerve terminals. Application of CTB subunit is either by iontophoresis or pressure injection into the region of interest. Cellular labeling can then be detected with either a fluorescent or a biotinylated secondary antibody or by the use of fluorescently conjugated CTB.

The examples of anterograde and retrograde transport tracers discussed above are well established

and widely used. However, their transport, and consequently their labeling, is restricted to the neurons where their uptake occurs, which greatly limits their ability to map neuronal circuitry. Transsynaptic tracers circumvent this limitation by spreading from the initial neurons to synaptically connected second- and third-order neurons, thus providing information on neuronal connectivity.

Transsynaptic Tracers

Plant Lectins

Plant lectins are carbohydrate-binding proteins that have high binding affinity for specific sugar components of glycoproteins and glycolipids on cell membranes. One of the first applications of lectins as neuronal tracers was WGA conjugated to HRP, which significantly enhanced the tracing sensitivity of free HRP. However, *Phaseolus vulgaris* leucoagglutinin, WGA, and barley lectin (BL) are also commonly used as tracers. Application of lectins is by iontophoresis or pressure injection into the region of interest. Following binding at the plasma membrane, lectins are actively internalized and transported predominantly in an anterograde direction, although retrograde transport is also observed. Lectins label neuronal arborization and the cell soma and are also transsynaptically transferred, thus providing information on neuronal connectivity. Although specific antilectin antibodies exist, their direct conjugation with fluorescent dyes, biotin, or HRP represents a better alternative and permits their direct detection by fluorescence microscopy, immunohistochemistry, or enzymatic methods, as discussed previously.

Unfortunately, several disadvantages are associated with lectin-mediated neuronal tracing. Following injection into the brain, lectins cannot be targeted to a specific neuronal subtype, and nonspecific labeling can occur. Furthermore, lectins can be released from neurons and taken up by neighboring glial cells. The bidirectional transport of lectins can also make interpretation of neuronal circuitry data very difficult. A further disadvantage is that WGA administration can induce a severe immune response in the host and marked inflammation at the injection site. However, the use of transgenic technology in combination with plant lectin tracers can minimize some of these issues, as discussed in the section titled 'Genetic targeting with specific neuronal promoters.'

Tetanus Toxin

Tetanus toxin (TeNT) is a bacterial protein toxin produced by *Clostridium tetani*. After entering the bloodstream, TeNT binds with high affinity to a receptor complex at the neuromuscular junction. Neurospecific binding of TeNT involves the interaction with polysialogangliosides of the G1b series, such as GT1b and GD1b (Figure 1 (a)), and glycosylphosphatidylinositol-anchored proteins, including Thy-1. TeNT internalization requires the integrity of membrane microdomains, and it is preceded by the lateral sorting of TeNT from the lipid component of its receptor complex. A specialized clathrin-mediated endocytic pathway delivers TeNT into stationary early endosomes that mature into fast retrogradely transported carriers in a mechanism dependent on the small GTPases Rab5 and Rab7. Following its transcytosis into inhibitory interneurons, full-length TeNT blocks neurotransmitter release by cleaving vesicle-associated membrane protein (VAMP)/synaptobrevin, a SNARE protein localized on synaptic vesicles. This induces a spastic paralysis, which is often fatal. *In vivo*, TeNT retrograde transport and trancytosis occur not only in motor neurons/inhibitory interneurons but also in other neuronal networks, such as the pre- and postganglionic neurons innervating the superior cervical ganglion and the avian paravertebral ganglion.

TeNT consists of two associated fragments, a 50 kDa light (L) chain and a 100 kDa heavy (H) chain. The L chain has a zinc endopeptidase activity responsible for the cleavage of VAMP/synaptobrevin. The H chain consists of two domains, the N-terminal domain (H_N), which is involved in the pH-dependent membrane penetration and translocation of the catalytic domain, and the C-terminal portion (H_C) responsible for neurospecific binding. A recombinant H_C fragment (TeNT H_C) retains the high affinity for neuronal membranes and retrograde transport capacity of the native toxin but lacks toxicity. Consequently, TeNT H_C provides a useful biological tool for the investigation of both retrograde axonal transport and neuronal connectivity. Generation of hybrid proteins consisting of TeNT H_C linked to reporters such as green fluorescent protein (GFP), lacZ, or HRP permits the tracing of axonal transport.

The ability of the H_C fragment of TeNT to be endocytosed and transported in a manner similar to that of the native toxin has allowed the development of a real-time transport assay that permits the quantitative analysis of the retrograde transport of fluorescent TeNT H_C in purified primary motor neurons in culture. The speed distribution of TeNT H_C carriers *in vitro* overlaps with that seen *in vivo* ($0.8–3.6\,\mu m\,s^{-1}$). The axonal transport of TeNT H_C occurs primarily by a cytoplasmic dynein-mediated process along microtubules, although there is evidence to suggest that fast retrograde transport of TeNT H_C also involves F-actin and myosin motors,

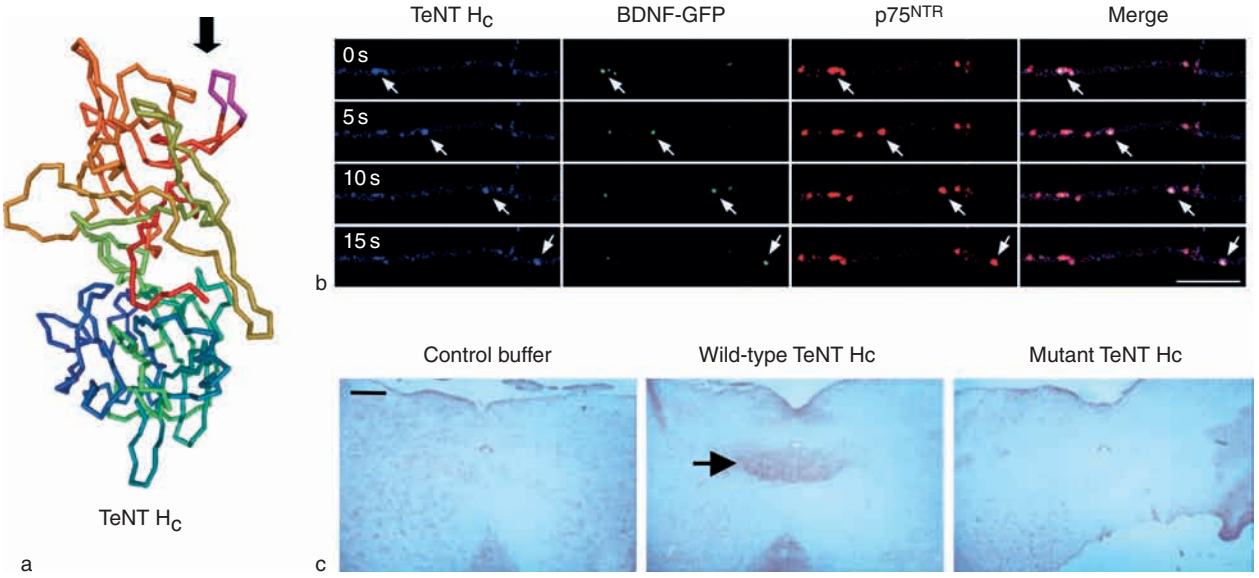

Figure 1 The binding fragment of tetanus toxin is an axonal retrograde marker *in vitro* and *in vivo*. (a) Structure of the TeNT H_C fragment. The black arrow indicates the position of one of the two ganglioside binding pockets, which are necessary for highly specific binding to the neuronal surface and trafficking. For further details, see Lalli G, Bohnert S, Deinhardt K, Verastegui C, and Schiavo G (2003) The journey of tetanus and botulinum neurotoxins in neurons. *Trends in Microbiology* 11(9): 431–437. (b) Axonal retrograde transport of fluorescently tagged. Binding fragment of tetanus toxin (TeNT H_C) in differentiated spinal cord motor neurons in culture. This probe (in blue) is cotransported in axonal carriers with brain-derived nerve growth factor fused with green fluorescent protein (BDNF-GFP) and with an antibody specific for the neurotrophin receptor $p75^{NTR}$. The motor neuron cell body is located out of view to the right. Arrows point out a TeNT H_C carrier positive for the neurotrophin and $p75^{NTR}$. (c) Retrograde labeling of the hypoglossal nucleus after intramuscular injection in the tongue of TeNT H_C. Mice were injected with control buffer, wild-type TeNT H_C, and a TeNT H_C mutant in which the ganglioside binding region highlighted in purple in (a) has been deleted. Only the specimen treated with the wild-type protein shows a strong staining in the region of the hypoglossal nucleus (arrow), which indicates that TeNT H_C undergoes retrograde transport *in vivo*. Scale bar = 5 μm (b), 167 μm (c). (b) Adapted from Deinhardt K, Salinas S, Verastegui C, et al. (2006) Rab 5 and Rab control endocytic sorting along the axonal retrograde transport pathway. *Neuron* 52: 293–305; Sinha K, Box M, Lalli G, et al. (2000) Analysis of mutants of tetanus toxin H_C fragment: Ganglioside binding, cell binding and netrograde axonal transport properties. *Molecular Neurobiology* 37: 1041–1051.

in particular myosin Va. Significantly, this assay has also identified that TeNT H_C is retrogradely transported in vesicles that also contain nerve growth factor, brain derived neurotrophic factor, and their receptors, $p75^{NTR}$ and TrkB (Figure 1(b)). Therefore, it appears that TeNT H_C exploits an essential physiological route of transport, which validates this assay as a tool to investigate endogenous retrograde transport in both health and disease. With this *in vitro* assay, a significant deficit in retrograde transport, which may contribute to disease pathogenesis, has been shown in embryonic motor neurons isolated from animal models of motor neuron disease, including the 'legs at odd angles' mouse, which has a mutation in dynein heavy chain, and the mutant $SOD1^{G93A}$ mouse.

Several studies have taken advantage of the specific retrograde transport and transsynaptic transfer of TeNT H_C for the investigation of neuronal circuitry. Early studies based on the detection of TeNT H_C labeled with colloidal gold by electron microscopy established the validity of the method. More recently, injection of either lacZ-TeNT H_C protein or complementary DNA into the tongue of mice produced strong β-galactosidase staining in the hypoglossal nucleus after retrograde transport in the hypoglossal motor neuron axons (Figure 1(c)). Furthermore, β-galactosidase staining is evident in the lateral reticular formation and several midbrain projections, such as the mesencephalic trigeminal nucleus, which are second- and third-order neurons of this network. Similarly, to investigate the regulation of locomotor activity, GFP–TeNT H_C was injected into the medial gastrocnemius muscle. Following internalization at the neuromuscular junction, GFP–TeNT H_C was retrogradely transported to the cell soma of motor neurons innervating the medial gastrocnemius muscle. Specific transsynaptic transfer allowed identification of the premotor neurons responsible for controlling the activity of gastrocnemius motor neurons. This tracing technique is not restricted to use in mammalian systems and has been shown to be effective in other organisms.

However, interpretation of circuits and connectivity revealed by the use of TeNT H_C as a neuronal tracer must take into account the dependence of TeNT H_C uptake on synaptic activity. Hence, the tracer will be

internalized rapidly at the most active synapses and more slowly at less active synapses via synaptic vesicle recycling. It is possible, therefore, that labeling of weak synaptically connected second-order neurons may coincide with the timing of labeling of strong synaptically connected third-order neurons. Furthermore, the extent of labeling is dependent on the quantity and quality of TeNT H_C initially administered.

Neurotropic Viruses

The application of neurotropic viruses as neuronal tracers exploits their infectious nature and ability to replicate within the target cell, which provides a self-amplifying signal. Furthermore, their ability to cross synapses to synaptically connected cells permits visualization of higher-order neurons and the mapping of neuronal circuits. A variety of viruses, including herpes simplex virus types 1 and 2 (HSV-1 and HSV-2), rabies and pseudorabies virus type 1, adenovirus, and adeno-associated virus, have been used for tracing motor, visual, and auditory neuronal networks *in vivo*. After injection into the brain, neurotropic viruses bind to specific receptors, are endocytosed via a variety of different pathways, and are then transported to the nucleus, where the viral genome is replicated. Viruses can be manipulated to express a capsid protein fused with a GFP variant, or else a GFP variant is translated on viral replication, thus allowing infected cells to be identified easily (Figure 2). The majority of neurotropic viruses are retrograde tracers. However, adeno-associated viruses and strain H129 of HSV 1 are predominantly anterograde tracers.

Viral neuronal tracing, however, has several limitations. As a consequence of their virulence, the infection and replication of certain viruses (e.g., herpes virus) can induce neuronal degeneration. Following replication, certain viruses can also spread to local glial cells or nonsynaptically connected neurons, thus generating nonspecific labeling. The extent of neuronal tracing is dependent on the viral strain, the dose, and the postinoculation time. However, use of a virulent strain to increase the tracing capacity may exacerbate neuronal death or increase nonspecific viral labeling. This drawback is partially avoided in case of viruses characterized by a long incubation period before the onset of symptoms (e.g., rabies). Alternatively, attenuated viruses display a limited spread to nonneuronal cells and a longer neuronal survival time postinoculation, thus extending the time available for transsynaptic transfer. It is important to note that certain viruses bind to receptors that are specific to neuronal membranes. For example, canine adenoviruses (CAVs) bind to the neuronal

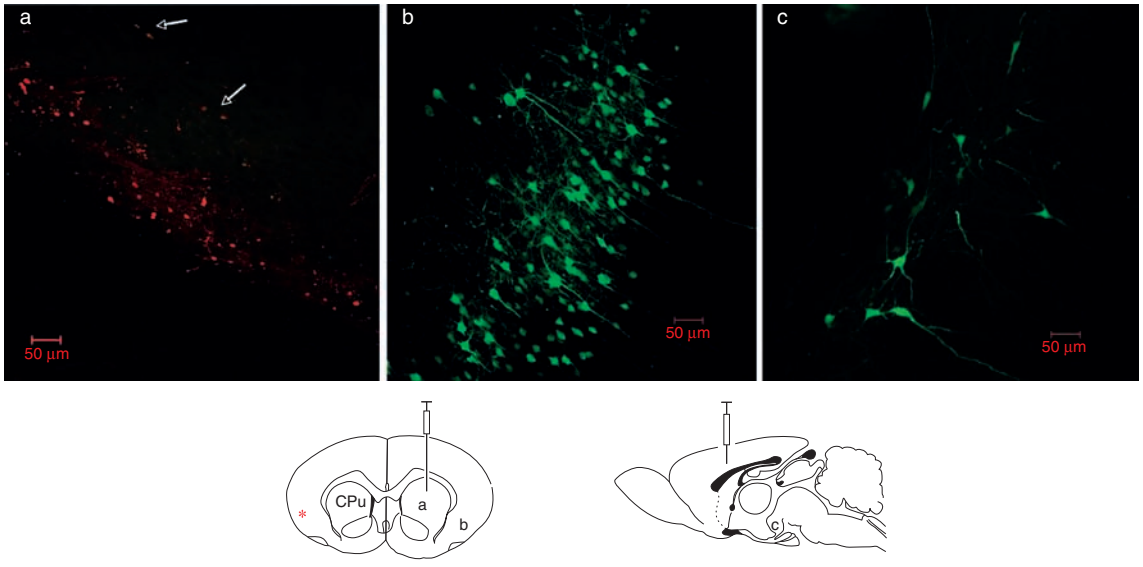

Figure 2 Retrograde transport of adenoviruses in the rat central nervous system. Both a canine adenovirus serotype 2 (CAV-2) vector expressing green fluorescent protein (CAVGFP) and a human adenovirus serotype 5 (HAd5) expressing DsRed2 (AdRFP) were stereotactically coinjected into the rat striatum (10^9 particles for each vector). Rat brains were cut rostrocaudally into 150 μm sections and were screened simultaneously for GFP and RFP expression. Each confocal image represents an approximate depth of 130 μm. At the site of injection (a), there was a dense concentration of AdRFP-transduced striatal astrocytes (~70%) and neurons (~25%) as well as some neurons transduced by both vector (yellow; white arrows). Only CAVGFP underwent significant retrograde transport from the striatum to afferent neurons in the cortex (b) and substantia nigra (c). Schematic inserts show midsagittal and frontal cross sections of the rat brain and the location of the injection site relative to the location of the transduced cells. Red asterisk represents contralateral cortex. CPU, caudate putamen. Adapted from Soudais C, Laplaec-Builhe C, Kissa K, Kremer EJ (2001) Preferential transduction of neurons by canine adenovirus vectors and their efficient retrograde transport *in vivo*. *FASEB Journal* 15: 2283–2285.

human coxsackievirus and adenovirus receptor. Consequently, CAV infection is restricted to neuronal cells in rat spinal cord cultures *in vitro*, in rat striatum *in vivo*, and in human brain biopsies *ex vivo* (Figure 2). Similarly, the rabies virus does not infect glial cells. In contrast, human adenovirus serotype 5 predominantly infects glial cells, and strains of HSV-1 and pseudorabies virus can also infect glial cells. The use of viral tracing technology therefore requires careful consideration and understanding of the benefits and limitations of each viral strain.

Genetic Tracing Technology

The development of multiple anterograde and retrograde tracers, as listed above, has provided extensive information on neuronal projections. Furthermore, the advent of transsynaptic labeling has enabled significant advances in neuroanatomical mapping. In spite of their flexibility, these tools also have severe limitations. One of the major drawbacks is their rather cumbersome targeting: The application of such tracers is normally by stereotactic injection or by administration of dye crystals into a chosen brain region. In addition, these methods do not allow the labeling of a specific neuronal population and often give rise to nonreproducible results between animals.

Recent advances in genetic technology have led to the development of novel methods for the mapping of neural networks. Specific labeling of a neuronal subset can be achieved via expression of a transgene encoding a transsynaptic tracer fused to a marker molecule, under the control of a neuronal-specific promoter. Continuous expression of the transgene enhances cellular detection and may promote a greater synaptic transfer, increasing the sensitivity of neuronal mapping. Unambiguous conclusions can therefore be made regarding the connectivity of specific neuronal subsets after monitoring axonal transport of the tracer. Genetic mapping will also prevent the nonspecific labeling that occurs after tracer injection and circumvents the need for invasive brain surgery that may damage the neural circuit under investigation. Furthermore, this technology can be used to investigate the development of neuronal circuits *in utero* and may provide functional insights since transsynaptic transfer of tracers will occur only in active circuits.

Genetic Targeting with Specific Neuronal Promoters

Transgenic expression of WGA or BL under the control of the rat olfactory marker protein gene promoter, which is expressed almost exclusively by olfactory sensory neurons in olfactory epithelium (OE) and vomeronasal epithelium (VNO), has provided reproducible mapping of olfactory circuitry. Lectin immunostaining was evident in the axons, dendrites, and cell soma of OE and VNO neurons, the first-order neurons. Furthermore, transsynaptic transfer permitted the labeling of neurons of the main and accessory olfactory bulbs, which represent the terminal fields of OE and VNO neurons (second-order), and of granule cells of the olfactory cortex (third-order), which are innervated by neurons from the main and accessory olfactory bulbs. However, immunostaining was not restricted to these areas but was also present in several neuromodulatory regions of the olfactory cortex, which indicated that retrograde transport had also occurred. BL expression has also been used to map sensory input in the olfactory cortex. Co-expression of BL with two distinct odorant receptor genes generated a transsynaptic tracer that could be monitored in localized areas of the OE, through the mitral and tufted relay neurons in the olfactory bulbs, to specific clusters of olfactory cortical neurons.

Similarly, following use of an L7 (Pcp2) promoter, WGA messenger RNA was expressed in cerebellar Purkinje cells. WGA was also detected in the synaptically connected deep cerebellar nuclei (second-order) and the red nucleus and thalamic ventrolateral nucleus, the third-order targets of cerebellar efferent axons. Furthermore, insight into neuronal connectivity in the *Drosophila* visual system was revealed by expression of WGA under the GAL4/UAS promoter, which drives expression in the R1–R6 photoreceptors in the retina. WGA immunoreactivity was present in the retina (first-order neurons), in monopolar cells (second-order targets of the photoreceptor cells), and also in the medulla, the target of the monopolar cells (third-order).

Transgenic mice expressing the GFP–TeNT H_C fusion protein have also been used to investigate neuronal circuitry. GFP–TeNT H_C under the control of a calbindin promoter, which drove the expression of the tracer in specific brain regions, has been studied. The transgenes also contained an internal ribosome entry sequence to allow co-expression of the lacZ reporter. This enabled differentiation of the neurons at the origin of the circuit, which expressed both β-galactosidase and GFP–TeNT H_C, and those receiving the protein as a consequence of transcytosis, which were only GFP–TeNT H_C-positive. Beta-galactosidase was detected in Purkinje cells, whereas GFP–TeNT H_C was also present in cerebellar granule cells, basket cells, and Golgi cells (second-order cells) due to the retrograde transport of GFP–TeNT H_C. Expression of a TeNT H_C–GFP fusion protein under the control of the orexin promoter has also been used to investigate

the neuronal regulation of orexin-expressing neurons in the lateral hypothalamic area involved in states of sleep and wakefulness. The use of TeNT H_C as a tracer has distinct advantages over lectins since it is only transported retrogradely, whereas lectins are transported bidirectionally.

Potentially viral vectors can also be genetically targeted to specific neuronal populations via insertion of a neuron-type specific promoter into the coding region of the viral glycoprotein gene. Further manipulation via insertion of additional genes, such as β-galactosidase or WGA, into the glycoprotein-coding region will induce selective protein expression. Indeed, in the olfactory system, intranasal application of an adenovirus encoding a WGA transgene under the chromogranin A promoter induced a pattern of lectin expression in the OE neurons (first-order), the main olfactory bulbs (second-order), and the main terminal fields of the olfactory cortex (third-order) similar to that obtained with transgenic expression of lectins under the olfactory marker protein promoter.

Genetic Recombination Strategies

The Cre/loxP recombination system has also been used to direct expression of transsynaptic tracers in selected neuronal populations. Breeding mice that express WGA only after excision of a loxP-flanked lacZ gene with mice expressing Cre recombinase under the control of the L7 (Pcp2) promoter generated offspring expressing WGA selectively in Purkinje cells. Anterograde transport analysis of WGA revealed connections with neurons of the deep cerebellar and vestibular nuclei (second-order) and the red nucleus of the midbrain and the thalamic ventrolateral nucleus (third-order).

This technology has also been applied in combination with viral tracing methods. A pseudorabies virus has been generated that requires Cre mediated recombination for its replication. Subsequent viral infection of transgenic mice that expressed Cre recombinase only in neuropeptide Y-expressing neurons consequently induced viral replication only in this neuronal population. Tracing of the subsequent viral spread has revealed the anatomical location of synaptic inputs regulating these neurons. More recently, a viral transcomplementation method was also employed to restrict viral expression to selected neuronal populations. A pseudotyped, GFP-expressing rabies virus, unable to spread transsynaptically, was used to infect brain slices in which relatively isolated neuronal populations had been transfected with both viral receptors and the viral glycoprotein gene required for transsynaptic spread. Following replication in transfected cells, the virus could spread transsynaptically, in a retrograde direction only, to monosynaptically connected neurons, identified by enhanced GFP expression, but no further due to the lack of glycoprotein in those neurons. Potentially, this technology may be used in combination with single-neuron electroporation to allow the electrophysiological characterization of synaptic contacts of selected neurons.

The tetracycline transactivator system provides an alternative strategy for the genetic regulation of neuronal mapping. In one example, mice transgenic for a TeNT H_C–alkaline phosphatase (AP) fusion protein under the tetracycline operator promoter were bred with mice expressing the tetracycline-controlled transactivator under the control of a cyclic phosphodiesterase splice variant, PDE102A, promoter. Expression of the fusion protein in the offspring was restricted to striatal spiny neurons, although retrograde transfer of TeNT H_C–AP revealed second-order connections with neurons of the ventral tegmental area/substantia nigra pars compacta, the dorsal raphe nucleus, and the mesopontine tegmentum.

Recombination systems can also provide greater control over the onset of transgenic tracer expression. As described above transgenic mice expressing WGA only after Cre recombinase-mediated excision of a lacZ gene have been developed. Microinjection of an adeno-associated virus expressing Cre recombinase and GFP, as a marker of viral infection, into the cerebellar cortex or visual cortex induced expression of the WGA protein in the infected areas and allowed its transsynaptic transfer to connected neurons.

Concluding Remarks

Over recent years the use of transsynaptic tracers has led to significant developments in the anatomical mapping of neural circuitry. Ongoing developments in genetic technology have further facilitated the mapping of functional interactions within selected neuronal populations, allowing their in-depth functional analysis, such as by electrophysiological recordings. The ability to identify specific connections within a given neural network *in utero* will now permit the onset of functionality in specific circuits to be determined. Ultimately, correlation of anatomical circuitry with neuronal functions will lead to a greater understanding of higher brain functions.

See also: Axonal mRNA Transport and Functions; Axonal Transport and ALS; Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Axonal Transport and Neurodegenerative Diseases; Botulinum and Tetanus Toxins; Retrograde Transsynaptic Influences; Retrograde Neurotrophic Signaling.

Further Reading

Bohnert S, Deinhardt KD, Salinas S, and Schiavo G (2006) Uptake and transport of clostridium neurotoxins. In: Alouf JE, and Popoff MR (eds.) *Bacterial Protein Toxins,* 3rd edn., pp. 390–407. London: Elsevier.

Braz JM, Rico B, and Basbaum AI (2002) Transneuronal tracing of diverse CNS circuits by Cre-mediated induction of wheat germ agglutinin in transgenic mice. *Proceedings of the National Academy of Sciences of the United States of America* 99: 15148–15153.

Callahan CA, Yoshikawa S, and Thomas JB (1998) Tracing axons. *Current Opinion in Neurobiology* 8: 582–586.

DeFalco J, Tomishima M, Liu H, et al. (2001) Virus-assisted mapping of neural inputs to a feeding center in the hypothalamus. *Science* 291: 2608–2613.

Deinhardt K, Salinas S, Verastegui C, et al. (2006) Rab 5 and Rab control endocytic sorting along the axonal retrograde transport pathway. *Neuron* 52: 293–305.

Enquist LW and Card JP (2003) Recent advances in the use of neurotropic viruses for circuit analysis. *Current Opinion in Neurobiology* 13: 603–606.

Kobbert C, Apps R, and Bechmann I (2000) Current concepts in neuroanatomical tracing. *Progress in Neurobiology* 62: 327–351.

Kremer EJ (2004) CAR chasing: Canine adenovirus vectors – all bite and no bark? *Journal of Gene Medicine* 6: S139–S151.

Lalli G, Bohnert S, Deinhardt K, Verastegui C, and Schiavo G (2003) The journey of tetanus and botulinum neurotoxins in neurons. *Trends in Microbiology* 11: 431–437.

Reiner A, Veenman CL, Medina L, et al. (2000) Pathway tracing using biotinylated dextran amines. *Journal of Neuroscience Methods* 103: 23–37.

Roux S, Colasante C, Saint cloment C, et al. (2005) Internalisation of a GFP-tetanus toxin c-terminal fragment fusion protein at mature mouse neuromuscular junctions. *Molecular and Cellular Neuroscience* 30: 572–582.

Sano H, Nagai Y, and Yokoi M (2007) Inducible expression of retrograde transsynaptic genetic tracer in mice. *Genesis* 45: 123–128.

Sinha K, Box M, Lalli G, et al. (2000) Analysis of mutants of tetanus toxin H_C fragments: Ganglioside binding, cell binding and netrograde axonal transport properties. *Molecular Neurobiology* 37: 1041–1051.

Soudais C, Laplace-Builhe C, Kissa K, Kremer EJ (2001) Preferential transduction of neurons by canine adenovirus vectors and their efficient retrograde transport *in vivo*. *FASEB Journal* 15: 2283–2285.

Ugolini G (1995) Specificity of rabies virus as a transneuronal tracer of motor networks: Transfer from hypoglossal motoneurons to connected second-order and higher order central nervous system cell groups. *Journal of Comparative Neurology* 356: 457–480.

Vercelli A, Repici M, Garbossa D, and Grimaldi A (2000) Recent techniques for tracing pathways in the central nervous system of developing and adult animals. *Brain Research Bulletins* 51: 11–28.

Wickersham IR, Lyon DC, Barnard RJO, et al. (2007) Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron* 53: 639–647.

Yoshihara Y (2002) Visualising selective neural pathways with WGA transgene: Combination of neuroanatomy with gene technology. *Neuroscience Research* 44: 133–140.

Babinski's Reflex/Sign

W M Landau, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

The sign of Babinski is a qualitative distortion of the human lower extremity flexion reflex; its occurrence is a highly sensitive correlate of functional or structural disturbance of the upper motor neuron (corticospinal/pyramidal) system. The sign has not been reported in lower primates.

The flexion reflex is a stereotyped reaction of obvious teleology: the coordinated withdrawal of the limb from a noxious stimulus is manifest when a human steps on a tack, or a frog's toe is dipped in acid, or a lobster's foot is pinched by a rock. Analysis of the mammalian flexion reflex was a major accomplishment of the Sherringtonian school. The flexion of the hind limb includes coordinated movements at foot, ankle, knee, and hip joints that bring the foot away from the stimulus toward the trunk. Note that the human muscles that dorsiflex (raise) the ankle and toes in this response are anatomically named extensors, for example, extensor digitorum longus. Conversely, plantar flexion (lowering) of the toes by flexor digitorum longus may be part of the extension support reflex that maintains upright stance.

In the normal, conscious, alert, relaxed human subject, the flexion reflex is most effectively elicited by a mildly noxious stimulus, usually a firm but dull point of pressure that moves distally in the S-1 dermatome to the ball of the foot, along the lateral plantar surface (classical Babinski site), or along the lateral side of the foot toward the fifth toe (Chaddock site). The polysynaptic spinal reflex is triggered by cutaneous nociceptive endings of small myelinated and unmyelinated afferent nerve fibers. The normal threshold response is a contraction of the short toe flexors, perhaps teleologically useful to cup the foot and move it away from the noxious stimulus. Early associated contractions, especially of tibialis anterior and extensor digitorum longus, dorsiflex the foot at the ankle; associated contractions of knee and hip flexors occur in strong responses. The coordinated movement program is generated entirely within the lumbosacral spinal cord gray matter; it is transiently depressed during the spinal shock phase evoked by acute cord transection.

The normal synergy includes plantar flexion of all the toes, due largely to activity in the short toe flexors, and in spite of the extensor digitorum longus activity that acts primarily to dorsiflex the foot at the ankle joint. In the normal subject, the extensor hallucis longus, the major dorsiflexor of the great toe, is not reactive. In clinical terminology, this normal pattern in which all toes go down is called the flexor response because the prime movers are anatomical flexors.

When the upper motor neuron pathways to the S-1 segment are impaired, structurally or functionally, the efficacy of evoked synaptic activation in the spinal segment gray matter is increased (released). This is clinically manifested by stronger and more extensive motor response, by decreased stimulus threshold, and by spread of excitable anatomical loci beyond the S-1 segment. This concept is evident in the parallel hyperactivity of monosynaptic stretch reflexes (tendon jerks) that is also elicited by upper motor neuron/tract lesions.

The useful and unique feature of the pathological plantar response is a result of this output hyperexcitability; extensor hallucis longus is recruited into contraction with its close anatomical neighbors and fellow ankle dorsiflexors, tibialis anterior and extensor digitorum longus. As a result there is hallux dorsiflexion (extension) at the metatarsophalangeal joint. Electromyographic and nerve block studies demonstrate an actual mechanical competition; the downward force of the short plantar flexors of the hallux is overcome by the counterforce extensor hallucis longus. For the plantar reflex, the only consistent manifestation of the upper motor neuron impairment is this dorsiflexion of the great toe, the extensor response of Babinski. Fanning (spreading apart) and even slight extension of the other four toes may occur with strong flexor as well as extensor responses and is not pathologically significant.

The critical movement to be observed in response to the noxious stimulus along the lateral sole of the foot is flexion or extension at the hallux metatarsophalangeal joint; with extension, the extensor hallucis longus tendon is conspicuous. In the real world most patients can relax satisfactorily, and with unrushed repetition a reliably consistent plantar reflex of one sort or another can be observed and reported, but not always. Analogous to the wide range of normal tendon jerk response, hallux flexion in the normal plantar reflex may range from near nil to near 90°. Borderline or variable responses may reflect borderline pathologic process. Whether a hallux that wiggles a little up and then down is pathologically more significant than one that wiggles down and then up is unproven. Some patients do relax poorly and display ticklish or 'voluntary withdrawal' behavior. The bottom line is that the plantar reflex is not

confined to two sharply defined up and down flavors; equivocal observations deserve verbal description with associated clinical judgment of reliability. Variations of consistency over extended periods are often important clinical data.

It is important to recognize that the plantar reflex, flexor or extensor, is an involuntary physiological measure of segmental spinal coordination pattern. Although the extensor reflex response may be associated with impaired performance of voluntary coordination and strength tasks by the cooperative patient, these phenomena are qualitatively different aspects of movement performance in the clinical neurological examination. The long-term maintenance of the extensor reflex usually marks a fixed pathological lesion like those of cerebral palsy, stroke, or destructive tumor. But its evolution and devolution may usefully mark even the hourly or daily course of transient pathological change, as with transient ischemia, trauma, hepatic, renal, and toxic encephalopathy, multiple sclerosis, subdural hematoma, meningitis, and encephalitis. There are, indeed, exceptional cases in which the plantar reflex does not reflect proven pyramidal tract lesions. Some are due to collateral lesions of spinal motor neurons, but a few are unexplained.

Whether the Babinski sign is released in the immature nervous system was long controversial. But the definitive studies of healthy infants by the Bodensteiner group found extensor responses in 100% at birth, 95.5% at 2 months, 64.8% at 6 months, 0.7% at 9 months, and 0% at 1 year. As movement behavior reflects brain growth and myelination, the pathological significance of the extensor response is a contingency of age during the first year of life.

Based upon experiments with electrical stimuli delivered to many areas over the buttock and lower extremity, Kugelberg et al. inferred concepts of wholly separate flexion and extension reflex patterns in normal subjects. Nevertheless, they concur that "pathological reflexes ... are not new reflexes ... they are the result of the breakdown of the harmonious integration of flexion and extension components in the normal reflex mechanism. Disorders of suprasegmental control impair the discriminating capacity of the reflex mechanism as regards the strength, modality and site of the sensory stimulus" (1960: 408).

In summary, the sign of Babinski is not a different reaction from the normal nociceptive polysynaptic reflex; rather, it is a hyperactive flexion reflex in which extensor hallucis longus is included by irradiation, that is, synaptic hyperexcitability within the S-1 spinal neuropil. Note that the generic name of this reflex is plantar reflex. The normal reaction with downgoing great toe is called flexor plantar reflex (response). The pathological extensor plantar reflex (response) is manifested by hallux dorsiflexion at the metatarsophalangeal joint. The eponym Babinski sign (response, or reflex) is synonymous with extensor plantar reflex (response). The term Babinski should not be used to indicate the plantar reflex or the site of stimulation; there is no such thing as a negative Babinski.

See also: Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Reflex Circuits.

Further Reading

Brain R and Wilkinson M (1959) Observations on the extension plantar reflex and its relationship to the functions of the pyramidal tract. With special reference to (1) the plantar reflex in infancy, (2) the crossed extensor plantar, and (3) extension of the great toe as an associated movement. *Brain* 82: 297–320.

Gingold MD, Jaynes ME, Bodensteiner JB, Romano JT, and Hammond I (1998) The rise and fall of the plantar reflex. *Journal of Pediatrics* 133: 568–570.

Kugelberg E (1948) Demonstration of A and C fibre components in the Babinski plantar response and the pathological flexion reflex. *Brain* 71: 304–319.

Kugelberg E, Eklund K, and Grimby L (1960) An electromyographic study of the nociceptive reflexes of the lower limb: Mechanism of the plantar responses. *Brain* 83: 294–410.

Landau WM (2003) The Babinski sign: Current views on a century-old observation. *Journal Watch Neurology* 5: 7.

Landau WM (2005) Editorial: plantar reflex amusement: Misuse, ruse, disuse, and abuse. *Neurology* 65: 1150–1151.

Landau WM (2005) Letter to the editor: The extensor plantar response in neonates is not the same as the Babinski sign. *Pediatric Neurology* 33(3): 223.

Landau WM and Clare MH (1959) The plantar reflex in man. With special reference to some conditions where the extensor response is unexpectedly absent. *Brain* 82: 321–355.

Miller TM and Johnston SC (2005) Should the Babinski sign be part of the routine neurologic examination? *Neurology* 65: 1165–1168.

Van Gijn J (1975) Babinski response: stimulus and effector. *Journal of Neurology, Neurosurgery, and Psychiatry* 38: 180–186.

Van Gijn J (1995) The Babinski reflex. *Postgraduate Medical Journal* 71: 645–648.

Van Gijn J (1996) *The Babinski Sign – A Centenary.* Utrecht, The Netherlands: Universiteit Utrecht.

Weaver RA, Landau WM, and Higgens JF (1963) Fusimotor function. Part II. Evidence of fusimotor depression in human spinal shock. *Archives of Neurology* 9: 127–132.

BAC Transgenesis: Cell-Type Specific Expression in the Nervous System

C Lai, Scripps Research Institute, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

What Is a BAC?

BAC is the acronym for bacterial artificial chromosome, a recombinant plasmid that contains a large insert of genomic DNA, typically in the 100–300 kb range. These plasmids differ from those used in routine cloning manipulations in that the BAC vectors have been derived from the fertility (F) factor, a large low-copy-number plasmid well documented for its ability to stably maintain large segments of *Escherichia coli* chromosomal DNA.

The vectors used to clone BACs were engineered to include only a small portion of the F factor, the region containing the origin of replication as well as genes that ensured its precise segregation during bacterial cell division. The low copy number (one or two per bacterium) helped to improve the stability of the inserted DNAs by reducing the probability of recombination between plasmids. The first use of an F factor-based vector to clone large DNAs was reported by Welcome Bender and colleagues (Harvard University), who assembled a 125 kb fragment of *Drosophila* genomic DNA in 1989. This vector was modified by Mel Simon and co-workers (Caltech), who used it to clone multiple large human genomic DNA fragments in 1992. These investigators introduced the term bacterial artificial chromosome, and the cloning of these human DNAs, some of which exceeded 300 kb in length, is generally considered to be the first paper on BACs.

The production of BAC genomic libraries by the laboratories of Mel Simon and Pieter de Jong (Children's Hospital Oakland Research Institute) played a key role in the genome mapping and sequencing projects. BAC libraries have larger average insert sizes than the original versions of the bacteriophage P1-derived artificial chromosome (PAC) libraries, in which the maximum insert size of 100 kb was determined by the capacity of the phage head. However, more recent versions of the PAC system that no longer use viral packaging, instead introducing DNAs via electroporation, have permitted insert sizes comparable to BACs. Despite their similarities, BACs have become the predominant platform as a result of their preferred use in the human genome project. One advantage of BACs over PACs was that their smaller vector size rendered them less costly to shotgun sequence.

Despite having very large DNA inserts that can range up to 1 Mb, the use of yeast artificial chromosomes (YACs) has been largely supplanted by BACs. BACs bear a number of advantages, including their maintenance in *E. coli*, which are easier to manipulate than yeast. Circular BAC DNAs are easier to prepare and less likely to fragment than the linear YAC DNAs. The large YAC inserts are more likely to be chimeric, that is, to be formed by the ligation of two unrelated genomic segments. In addition, the inserts in BACs are more stable and appear to undergo fewer deletion events. However, for genes that are too large for BACs, researchers may consider using YACs.

Why Are BACs Useful?

A typical BAC clone contains a segment of genomic DNA large enough to include most conventional-size mammalian genes. Because most BAC clones have inserts less than 300 kb in length, very large genes will not fit on to a single BAC. BACs that contain both a gene's coding exons and its regulatory regions are particularly useful when used as transgenes because they are likely to be expressed in a manner similar to that of the endogenous gene. The large size of a BAC helps to insulate it from the effects of transcriptional activity at the site of transgene insertion. Accordingly, mice containing one or more copies of a BAC transgene may be anticipated to express that gene in a spatially and temporally faithful manner, although at elevated levels as a result of the increase in gene dosage.

Transgenic mice produced using large BAC-based transgenes have been used to genetically complement loss-of-function mutations. This has facilitated efforts to positionally clone genes and has permitted the production of murine models of human disorders. In the earliest example of BAC transgenesis, mice generated using a 140 kb clone containing the wild-type circadian *Clock* gene were used to rescue two different lines of *Clock* mutant mice, one exhibiting a long circadian period and the other a loss-of-rhythm phenotype. Furthermore, transgenic mice bearing multiple copies of the *Clock* gene BAC exhibited a shorter period length, providing additional evidence in support of a role for this gene in the circadian pacemaking process.

BAC transgenic-mediated rescue has also been used to prove that the defect in the mutant mouse wobbler, which is characterized by motor neuron disease and defects in spermiogenesis, resided in the gene encoding the vacuolar-vesicular protein sorting factor 54 (Vps54). BAC transgenics have also been used to demonstrate that defects in the Psd-95, Dlg, and ZO1 (PDZ) domain-containing gene *whirlin* cause deafness in the whirler mouse. This gene also underlies the human autosomal deafness locus DFNB31.

Transgenic mice produced using human BACs have been used to complement defective genes in mice. In addition to providing evidence for the conservation of regulatory elements between mouse and human, this has permitted the use of modified human BACs bearing specific mutations to produce humanized murine models of specific genetic diseases, such as thalassemias. BAC transgenic mice permit altering gene dosage, and mice overexpressing the transcription factor Tbx1 have been proposed to model the condition known as the velo-cardio-facial syndrome (DiGeorge syndrome) because these animals exhibit the full spectrum of defects observed in human patients.

BAC transgenic mice have also proved useful for mapping the location of a gene's regulatory elements. This approach stems from the concept that if a line of BAC-based transgenic mice can rescue a loss-of-function mutation, then the BAC used must contain all (or nearly all) of the appropriate regulatory elements, at least for the cell type(s) analyzed. If a BAC provides only a partial rescue, then that BAC probably contains a subset of the required regulatory elements. Therefore, by using different BACs, we can infer the location of different enhancer sequences that regulate expression in different cellular subpopulations. BACs have also been used to map the regulatory regions of a gene using enhancer-trap-based strategies, in which a reporter gene and associated minimal promoter are used to search for transcriptionally active regions. In this setting, a collection of overlapping BACs containing reporters are first generated and then used to produce lines of transgenic mice. Each line is then assessed for reporter expression.

Therefore, BACs that contain a specific gene and its associated regulatory regions can serve as promoters to drive the expression of inserted genes. This is currently one of the most beneficial features of BACs because it surmounts the problems associated with finding useful promoters. In instances in which the expression of a gene marks a specific subset of cells, a BAC containing that gene may function as a cell-type specific promoter.

How Do We Identify a BAC That Contains a Gene of Interest?

As a result of the tremendous increase in genomic sequence information, identifying a BAC useful for our experiment may be as simple as a quick look at the genomic locus containing our gene of interest. If we are working with an organism whose genomic sequence has been determined, such as mouse or human, the location of specific BAC clones is included as part of the annotation provided by the web-based genomic browsers (as at the National Center for Biotechnology Information (NCBI), University of California Santa Cruz genome, and Sanger Institute websites). The BAC clones are typically displayed as sets of overlapping clones that blanket virtually any genomic region. The annotation for a given BAC can be viewed by clicking on the name of a particular clone, which typically opens a new window containing the information on that entry. One quick shortcut is to use the NCBI BAC Clone Registry and the Find clones near a genomic feature tool to enter the gene name. A set of overlapping BACs for that gene will be displayed.

For organisms for which the genomic data are not available or are incomplete, it may be necessary to use polymerase chain reaction (PCR)- or hybridization-based methods to screen for BAC clones corresponding to a candidate gene. BAC libraries can be obtained as macroarrays spotted on to nylon filters. Library screening services are also available through the BACPAC resources and through commercial vendors.

The process of identifying a BAC that contains a gene along with its required regulatory elements is currently an empirical process because there are no systematic ways of finding these regions, which may be located on either side of the first or last exons or in the introns between. Therefore, unless a specific BAC has been previously shown to contain the requisite elements, selecting a candidate BAC clone carries a measure of uncertainty. It is therefore advisable to select a BAC with considerable (50 kb or more) 5′ and 3′ flanking regions, if possible. It should be noted, however, that there are examples of genes having regulatory elements located at a considerable distance (>100 kb) from the coding exons; thus, there will be instances when BACs will simply not be large enough to include all the necessary regulatory elements for an otherwise compact-appearing gene. Of course, as the size of the regions flanking a gene increases, there is also the chance that other genes will be present on the BAC. The expression of these unwanted genes can be inactivated using homologous recombination.

How Do We Obtain BAC Clones?

A quick web-based search has suggested that there appear to be at least 300 BAC genomic libraries now available for a rapidly increasing list of species ranging from *Anopheles gambiae* (African malaria mosquito) to *Zea mays* (Indian corn). A list of registered BAC libraries and a key to BAC nomenclature can be found at the NCBI clone registry. One frequently used source for obtaining BACs is the BACPAC Resources Center at the Children's Hospital Oakland Research Institute (CHORI), which provides commonly used libraries for mouse (RPCI-23 and RPCI-24), human (RPCI-11), and other species from the Roswell Park Cancer Institute (RPCI), Lawrence Berkeley National Laboratories (LBNL), and CHORI collections. The Caltech human BAC library D (CTD) can be obtained through the commercial supplier Invitrogen. The aformentioned libraries have extensive BAC end-sequencing data, which on a practical level means that these have been precisely located on the genome and are the ones displayed by the genome browsers. The sequencing of the BAC ends of a limited set of murine and human BAC libraries was originally undertaken to facilitate the mapping and sequencing efforts underlying the genome projects.

A partial list of academic and commercial sources for BAC libraries can be found at the NCBI clone registry. One vendor, Amplicon Express (Pullman, WA), currently offers a total of 97 BAC libraries ranging from the platypus to the pinot noir grape. Therefore, before a reseracher embarks on producing a BAC library of his or her own, it is useful to perform a thorough search; it is possible that a library for the organism either has already been made or is in the works.

Unlike small plasmids, which are typically provided as purified DNAs, BAC clones are usually provided within a suitable *E. coli* host strain (usually DH10B) in the form of bacterial stab cultures. Because BAC libraries are maintained in multiwell plates, the clonal purity of the BAC supplied may not be guaranteed, and it is the recipient's responsibility to verify the identity and integrity of the clone.

How Do We Modify These Very Large DNAs?

Because BACs contain very large inserted DNA fragments, it is not feasible to use conventional restriction endonuclease-based (cutting-and-pasting) cloning methodologies to manipulate these clones. Instead, recombination-based strategies have been devised that permit the targeted insertion of DNA fragments into these large BACs. The homologous recombination is performed *in vivo*, that is, within *E. coli* bacteria (see Figure 1). There are two basic types of systems, one using the *E. coli* recombinase recA and the other using bacteriophage-derived recombinases.

The recA-based system was developed by Nathaniel Heintz (Rockefeller University) and uses recombination substrates present in a circular DNA form (see Figure 2). Using this approach, the DNA fragment

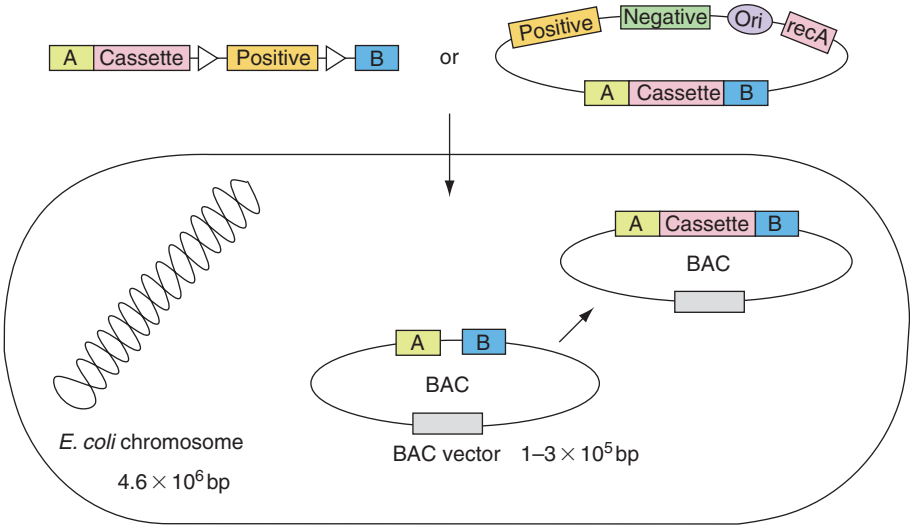

Figure 1 Overview of homologous recombination in BACs. Homologous recombination occurs within *E. coli* bacteria that contain a BAC, typically at one copy per cell. In the case of recA-mediated recombination, the recombinase activity (recA) is provided on the same plasmid as the DNA to be inserted. In the case of the phage recombinase-mediated systems, the recombinases are either provided via the introduction of a plasmid or are encoded on a defective lambda prophage engineered into the *E. coli* genome (not shown). The DNA fragments to be recombined are introduced into the bacteria by either electroporation or transformation. Note that the recombination process also requires other host-provided enzymatic activities. Occasional nonhomologous recombination events occur, with insertion occurring into the *E. coli* genome (see Figures 2 and 3). BAC, bacterial artificial chromosome; Ori, origin of replication.

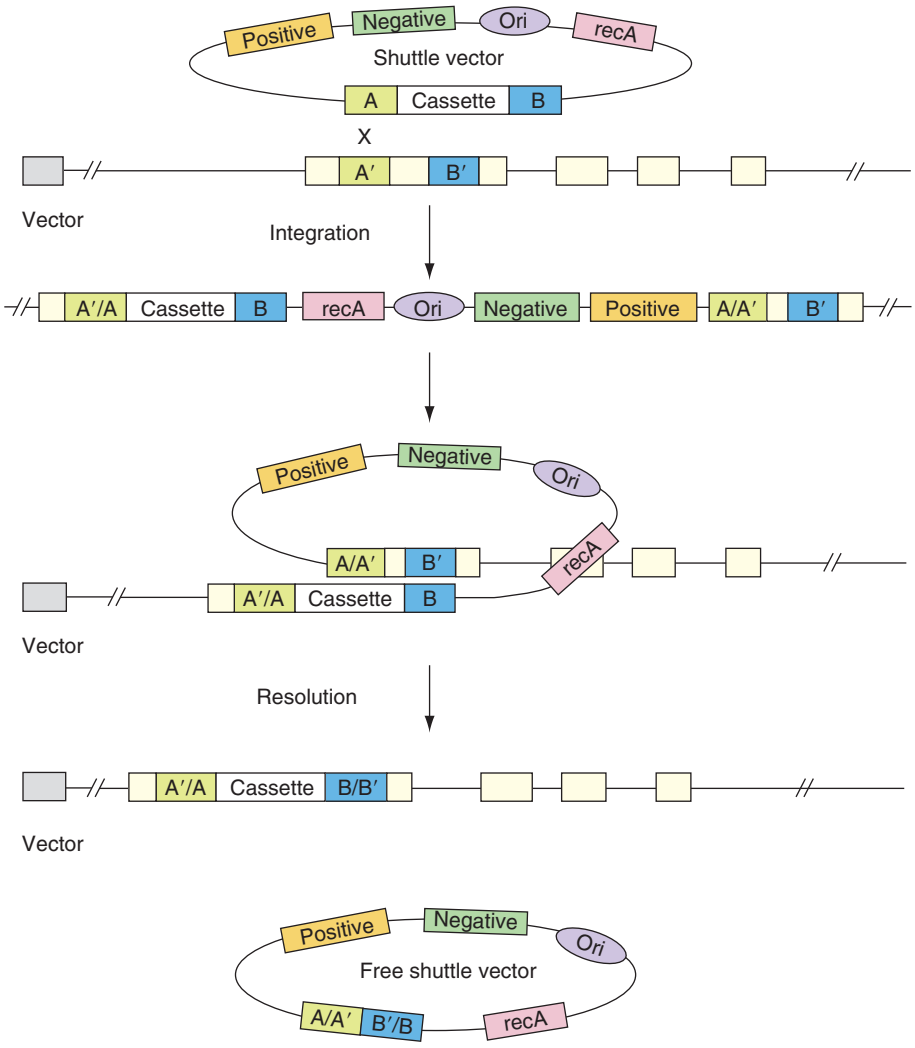

Figure 2 RecA-mediated homologous recombination in BACs. The DNA fragment to be inserted via homologous recombination (the cassette) is subcloned into a shuttle vector. The homology arms (A and B) are subcloned on each side of this cassette. In this system, the arms are typically 300–500 bp in length. The shuttle vector also contains an origin of replication, the gene encoding the recombinase recA, and markers permitting both positive and negative selection. On introduction into *E. coli* and under positive selection, the circular shuttle vector DNA may be inserted into the BAC using recombination with either homology arm A or B. Integration using arm A is depicted. The desired recombination product is obtained by selecting for the excision of the vector (negative selection) via recA-mediated recombination using homology arm B. Note that excision using arm A leads to the restoration of the original unmodified BAC. BAC, bacterial artificial chromosome; Ori, origin of replication. Reproduced from Heintz N (2001) BAC to the future: The use of BAC transgenic mice for neuroscience research. *Nature Reviews Neuroscience* 2: 861–870.

targeted for insertion into the BAC is first cloned into a shuttle vector. Two regions of homology (the homology arms) that flank this DNA fragment are also introduced. For recA-based recombination, arms of approximately 300–500 bp in length are used. The shuttle vector also contains a marker permitting positive selection and one permitting negative selection. In the first step of recombination, the circular plasmid is recombined into the BAC under positive selection. This can occur using either of the two homology arms and will lead to the insertion of the entire plasmid into the BAC. The desired recombination product is obtained under negative selection for the recA-mediated loss of the inserted plasmid, which can occur in two ways. In the first, recombination occurs using the same homology arm, returning the BAC to its initial state. In the second possibility, recombination occurs using the other homology arm, in which case the desired product remains in the BAC.

There are two systems derived from phage, one from bacteriophage lambda and the other from a lambdalike prophage, Rac. Unlike the recA-mediated system, the phage systems require a linear (as opposed to a circular) DNA as a substrate for recombination. A system using the recE and recT functions of

Rac, called ET cloning, was developed by Francis Stewart (Technische Universitat Dresden). In addition, the inclusion of the gam gene product increased the efficiency of recombination by protecting the linear DNA fragments from exonucleolytic degradation. These genes were provided on a plasmid permitting inducible expression of recE and the constitutive expression of recT and gam. The inducible expression of these activities was advantageous because there were concerns that prolonged expression of recombinase activity might lead to unwanted recombination events. Thus, this system was portable in that it could be transformed into bacteria containing a BAC. The selection for recombinants was accomplished by including selectable markers in the linear DNA to be introduced into the BAC. These markers were flanked by either loxP or frt sites, permitting their excision by either cre or flp recombinase, whose activities were introduced on plasmids. This strategy was developed to simplify the process of performing multiple independent recombination events on a BAC (see **Figure 3(a)**). For example, if a BAC contained more than one gene, it would be desirable to inactivate the other genes to avoid the unwanted expression of those gene products.

The original version of the system developed by Neal Copeland and Donald Court (NCI/NIH) used the bacteriophage lambda-encoded exo and beta gene products, which are similar in function to recE and recT. In this system, gam was also used to preserve the integrity of the linear DNA fragments. The Copeland lab system placed the expression of exo, beta, and gam under the control of a temperature-sensitive repressor, permitting inducible regulation. One advantage of this system was that it avoided the constitutive expression of gam, which reduced cell viability. These three genes were present on a defective lambda prophage, and, hence, this system was not portable. Therefore, BACs need to be introduced into these recombinogenic strains via electroporation. (Portable versions of this system have since been developed.) Two different strains were generated: one that permitted the arabinose-inducible expression of cre recombinase and one that permitted the inducible expression of flp recombinase. These activities were used to excise the selectable markers, which, as in the Stewart system, were incorporated into the linear DNA fragment. In both of the Stewart and Copeland lab systems, the linear DNAs targeted for insertion were introduced into the bacteria via electroporation.

The biggest advantage of the phage-derived systems is that the regions of homology can be quite small with useful recombination obtained with homology arms as short as 50 bp. This small size permits the introduction of these regions into the linear DNA substrates using PCR. One potential disadvantage of the phage systems has been the use of cre (or flp) recombinase to remove selectable markers because the excision events leave a single residual loxP or frt site. This is probably only an issue if the recombinant BACs are to be used for the production of transgenic mice, and then only a problem if these mice are to be mated with cre- or flp-expressing lines. In this regard, it should be noted that many of the commonly used BAC vectors also contain loxP sites; thus, BAC-based transgenic lines produced with circular BAC DNAs will probably contain these sites.

The issue of residual loxP sites has been eliminated in the most recent version of the recombination system from the Copeland lab (see **Figure 3(b)**). This approach features a strategy using a selectable marker, galactokinase (galK), for which there is both positive and negative selection. In the first step, a linear DNA fragment containing the galK gene is inserted via homologous recombination at the precise location to be modified (positive selection). In the second step, homologous recombination is used to replace the galK gene with the desired DNA element (negative selection).

The development of these homologous recombination systems has been a major advance because they have greatly facilitated the process of moving fragments of DNA from one vector to another, without the need for restriction endonucleases. Their use is not restricted to BACs; they can be used with higher-copy-number plasmids. It should be noted, however, that, if we wish to subclone very large DNA fragments, a moderate-copy-number (<20) plasmid should be selected to avoid exceeding the DNA synthetic capacity of *E. coli*. These recombination systems have proved to be of particular use in the assembly of targeting constructs used for the production of knockout or knock-in mice. Methods have also been devised that permit the introduction of point mutations; however, these are performed in the absence of a selectable marker, so an efficient screening process needs to be implemented.

The purification of BAC DNAs at both the small (miniprep) and large scale uses a modified alkaline lysis procedure. Because BACs are covalently closed supercoiled circular DNAs, they can be purified by CsCl density gradient centrifugation. Affinity-based column purification kits are also commercially available. The analysis of BAC DNAs and fragments derived by digestion with restriction endonucleases requires the use of pulsed-field gel electrophoresis (PFGE) because conventional agarose gel electrophoresis cannot easily resolve fragments >30 kb. Purified BAC DNAs can be introduced into *E. coli* using electroporation.

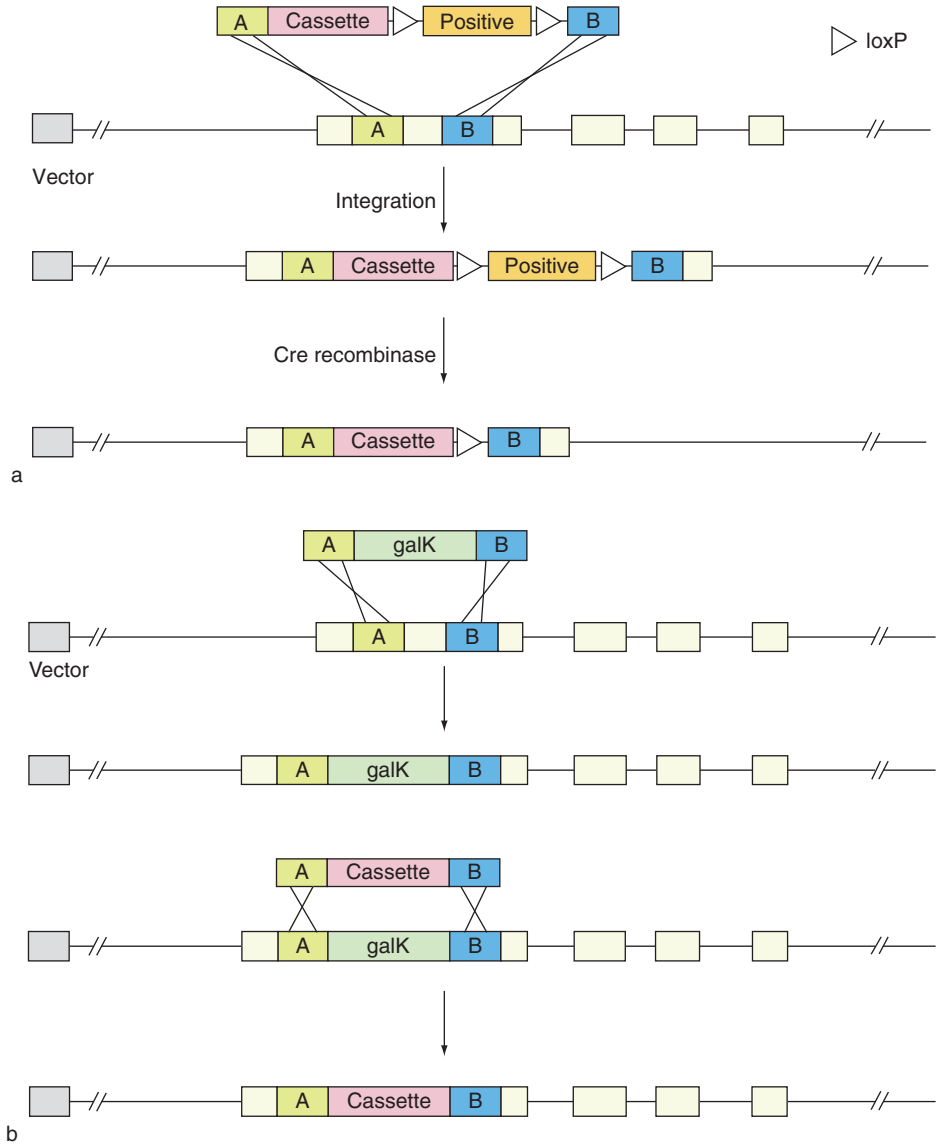

Figure 3 Two methods of homologous recombination mediated by phage recombinases: (a) homologous recombination using positive selection; (b) two-step method. In (a), the DNA fragment to be inserted (the cassette) is prepared as a linear DNA fragment. Homology arms (A and B) can be introduced using PCR as arms may be as short as 50 bp. Selection for recombinants is accomplished through the inclusion of a positive selectable marker that is flanked by either loxP or frt sites. The linear DNAs are introduced into the bacteria via electroporation. Using the Stewart lab system (ET cloning), the bacteria express recE, recT, and gam from the Rac prophage. Using the Copeland lab system, the phage lambda-encoded products exo, beta, and gam are used. Recombination of the linear DNA requires the use of both homology arms. On identification of homologous recombinants, the selectable marker can be removed by either cre or flp recombinase, giving rise to a residual loxP (or frt) site. The method shown in (b), a variant of (a) developed in the Copeland lab, eliminates the issue of the residual loxP or frt sites by using a two-step strategy. The selectable marker galactokinase (galK) is used, for which there is both positive and negative selection. In the first step, galK is inserted under positive selection. In the second step, galK is replaced by the desired cassette under negative selection.

Transgenic mice have been generated with both linear and circular BAC DNAs. Like smaller transgenes, it is thought that BAC transgenes may be present in tandem arrays in the genome. Although the insertion of such sizable segments of DNA leads to questions concerning the long-term stability of these elements, few data are available on the subject.

BAC Transfection into Mammalian Cells

BACs can be introduced into mammalian cells *in vitro*, although the efficiency of this process is hampered by the large size of the BAC DNAs. There has been moderate success in identifying condensing agents that facilitate BAC transfection, with efficiencies exceeding

50% in some cases. Because most BAC vectors do not permit replication in mammalian cells, it may be necessary to retrofit the BAC to include the appropriate origins of replication, selectable markers, and reporter genes.

BACs have been of particular use for those working with large double-stranded DNA viruses, such as herpesviruses or poxviruses. Due to their large capacity, it has been possible to clone entire viral genomes on to a single BAC and to generate altered forms using the aforementioned homologous recombination systems. The viral BAC clones can be re-introduced into mammalian cells, leading to the formation of recombinant viruses. To avoid possible interference by the BAC vector, systems have been developed that permit the cre recombinase-mediated excision of these sequences in mammalian cells. Although the use of BACs is in its early stages, it is evident that it should facilitate studies addressing viral biology, the generation of viral vectors for gene therapy, and the development of vaccines.

BACs

The nervous system contains a very large number of cell types, with estimates ranging from hundreds to thousands of different kinds of neurons and glial cells. This diversity of cell types poses a challenge for those attempting to understand the function of selected cellular subpopulations because the tools for manipulating gene expression in these subsets have, for the most part, not been generated. The use of BAC genomic clones as a source of functional promoter regions has expedited the development of transgenic lines exhibiting cell type-restricted gene expression. These lines should greatly facilitate our efforts to explore the diversity of neural cells and should provide the starting point for the development of animal models useful for the evaluation of their functional roles.

The utility of BACs for studying the nervous system has been underscored by the efforts of Nathaniel Heintz and colleagues, who have used them to produce hundreds of lines of transgenic mice expressing the enhanced green fluorescent protein (EGFP) in distinct subsets of neural cells. The primary goal of these efforts was to provide researchers with a high-resolution descriptive analysis of the cells in which specific genes are expressed in the nervous system. It was anticipated that these efforts would also lead to the identification of novel cellular subpopulations. The initial findings of this large-scale effort included the production of transgenic lines using BACs containing the gene encoding choline acetyltransferase (ChAT). The investigators had used two different ChAT BACs: one that included the entire gene plus flanking DNA and a clone that had considerable 5′ flanking DNA but that lacked the 3′ end of the gene. The BAC containing the entire ChAT gene yielded transgenic mice that appeared to express EGFP faithfully in cholinergic neurons. Although the other BAC line exhibited reporter expression in a subset of cholinergic neurons, EGFP was also detected in noncholinergic cells, suggesting that critical regulatory elements may reside in the 3′ region flanking the gene. These findings illustrate the importance of selecting the right BAC clone. Based on making several hundred BAC transgenic lines, the investigators reported that in situations in which genes were too large to fit on to a single BAC, the use of clones containing extensive 5′ flanking DNA regions (100 kb or more) was typically not sufficient to obtain proper expression of EGFP in transgenic mice, providing additional support for the concept that internal or 3′ sequences are often critical for the expression of many genes.

As additional examples of the utility of this collection of BAC transgenic mice, findings were presented for the goosecoid-like (Gscl) gene, whose expression profiles suggested that it may serve as a lineage marker for a subset of cells forming the interpeduncular nucleus. Also highlighted were the Semaphorin 3b (Sema3b)-EGFP mice, which have identified subpopulations of neurons projecting to the cerebellum, olfactory bulb, and superior colliculus. The observed patterns have helped to shape hypotheses concerning the role of this molecule in the establishment of these synaptic connections. And mice produced using a BAC containing the gene for phosphodiesterase 1C (Pde1c) appeared to contain a marked subpopulation of migrating cells that may correspond to the subpial granular layer.

The production of these mice should facilitate functional studies, such as electrophysiological analyses by permitting the direct visualization of discrete subpopulations of living cells. It is evident that the ability to mark cells should also help to identify potential synaptic partners and provide insight into their developmental origins. The data for these lines of mice are accessible through the Gene Expression Nervous System Atlas (GENSAT). These mice are being made available to other investigators, and the current status can be viewed on the GENSAT website.

The use of BACs to mark specific cell types has also been the goal of a team at Regeneron Pharmaceuticals, whose largely automated process, Velocigene, uses homologous recombination in bacteria to facilitate the rapid production of BAC-based constructs useful for recombination in embryonic stem (ES) cells. These constructs were designed to insert

the reporter gene β-galactosidase into a specific locus, leading the creation of a null allele. Although not specifically focused on the nervous system, their efforts have yielded a number of interesting patterns of gene expression in the brain.

The concept of using BACs as a source of gene-specific promoters useful for marking restricted populations of neural cells has also been exploited by other investigators. GAD67-EGFP mice have been produced and used to unravel how basket cell pinceau synapses on the Purkinje cell initial axon segment are formed. BAC transgenic glycine transporter 2 (GlyT2)-EGFP mice have also been developed and should facilitate the electrophysiological analyses of glycinergic neurons. The recent publication of regulator or G-protein-signaling-4 (RGS4)–GFP mice has also shown that, to validate that the BAC-transgenic mice exhibit faithful patterns of expression, the sites of reporter gene expression must be very carefully analyzed. Thus, one important consequence of the development of these transgenic lines is that our understanding of the sites of expression for many genes, including those considered to be well characterized, will be thoroughly reevaluated. BACs have also been used to make lines of mice expressing cre recombinase in genes expressed in Purkinje cells, retinal rod bipolar neurons, and in calmodulin-dependent protein kinase II (CaMKII)-α-expressing forebrain neurons.

A major dividend of these ongoing efforts has been the validation of specific BACs because one of the major risks when working with these clones is understanding whether a given BAC contains the regulatory elements required for the faithful expression of the gene contained therein. Thus, the identification of functionally useful BACs should facilitate the production of additional BAC-based lines. It is anticipated that these will include the production of additional cre recombinase-expressing lines, as well as those permitting temporally regulated gene expression using the tetracycline- or lac-based systems.

The development of BACs and the homologous recombination systems required to manipulate them can be viewed as key technological dividends stemming from the genome sequencing projects. Now that the genes have been identified, the goal is to understand how they function. BACs are playing a major role in this process in that they have provided a rapid way of manipulating gene expression in restricted subsets of cells, which is of particular relevance in the cell-type diverse nervous system. In addition, BACs are also being harnessed for the production of targeting constructs useful for the National Institutes of Health (NIH) Knockout Mouse Project (KOMP), whose goal is to produce null and conditional null alleles for essentially all genes. It also appears likely that BACs will continue to play a key role in defining the identity and location of a gene's regulatory elements. Thus, although BACs were first produced as a tool to expedite genomic mapping and sequencing, they are also playing a key role in the pursuit of gene function.

See also: Alzheimer's Disease: Transgenic Mouse Models; BAC Use in the Study of the CNS; Transgenic Models of Neurodegenerative Disease.

Further Reading

Ango F, di Cristo G, Higashiyama H, Bennett V, Wu P, and Huang ZJ (2004) Ankyrin-based subcellular gradient of neurofascin, an immunoglobulin family protein, directs GABAergic innervation at Purkinje axon initial segment. *Cell* 119: 257–272.

Antoch MP, Song E-J, Chang A-M, et al. (1997) Functional identification of the mouse circadian Clock gene by transgenic BAC rescue. *Cell* 89: 655–667.

Copeland NG, Jenkins NA, and Court DL (2001) Recombineering: A powerful tool for mouse functional genomics. *Nature Reviews Genetics* 2: 769–779.

Court DL, Sawitzke JA, and Thomason LC (2002) Genetic engineering using homologous recombination. *Annual Review of Genetics* 36: 361–388.

Gong S, Zheng C, Doughty ML, et al. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425: 917–925.

Heintz N (2001) BAC to the future: The use of BAC transgenic mice for neuroscience research. *Nature Reviews Neuroscience* 2: 861–870.

Lee E-C, Yu D, Martinez de Velasco J, et al. (2001) A highly efficient *Escherichia coli*-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics* 73: 56–65.

Roberts L (2001) Controversial from the start. *Science* 291: 1182–1188.

Schmitt-John T, Drepper C, Mussmann A, et al. (2005) Mutation of Vps54 causes motor neuron disease and defective spermiogenesis in the wobbler mouse. *Nature Genetics* 37: 1213–1215.

Shizuya H, Birren B, Kim UJ, et al. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America* 89: 8794–8797.

Warming S, Costantino N, Court DL, Jenkins NA, and Copeland NG (2005) Simple and highly efficient BAC recombineering using galK selection. *Nucleic Acids Research* 33: e36[online].

Valenzuela DM, Murphy AJ, Frendewey D, et al. (2003) High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nature Biotechnology* 6: 652–659.

Yang XY, Model P, and Heintz N (1997) Homologous recombination based modification in *Escherichia coli* and germline transmission in transgenic mice of a bacterial artificial chromosome. *Nature Biotechnology* 15: 859–865.

Zhang Y, Buchholz F, Muyrers JP, and Stewart AF (1998) A new logic for DNA engineering using recombination in *Escherichia coli*. *Nature Genetics* 20: 123–128.

Relevant Websites

http://www.bacpac.chori.org – BACPAC Resources at the Children's Hospital Oakland Research Institute (CHORI).

http://www.gensat.org – Gene Expression Nervous System Atlas (GENSAT).

http://www.ncbi.nlm.nih.gov – National Center for Biotechnology Information.

http://www.regeneron.com – Regeneron Pharmaceuticals (for Velocigene).

http://www.genome.ucsc.edu – University of California at Santa Cruz genome website.

http://www.sanger.ac.uk – Wellcome Trust Sanger Institute.

BAC Use in the Study of the CNS

X W Yang, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Advantages of Bacterial Artificial Chromosome Transgenic Mice

Bacterial artificial chromosome (BAC) transgenic mice are generated by the direct injection of a DNA fragment into a fertilized one-cell mouse embryo, which results in stable but random integration of multiple copies of the transgene. Unlike a gene-targeting approach in embryonic stem cells, which is aimed to manipulate the endogenous murine genes, the transgenic approach is often utilized to overexpress an exogenous transgene in order to assess the phenotypic consequences. A basic transgenic construct (Figure 1(a)) consists of a piece of genomic DNA that contains a eukaryotic promoter, associated regulatory elements (i.e., enhancers, suppressors, locus control regions, etc.), an open reading frame (ORF) that encodes the gene of interest (marker genes such as lacZ, green fluorescent protein, Cre recombinase, etc.), and a eukaryotic polyadenylation (polyA) signal. Traditional transgenic studies, using genomic DNA elements less than 20 kb, are often plagued by inconsistency due to positional effects, which refers to variable, ectopic, or frank failure of transgene expression when the same transgene is integrated at different genomic loci. Many empirical studies reveal that positional effects are due to the use of smaller genomic DNA fragments that lack critical transcriptional regulatory elements. An average mammalian gene size is about 27 kb, but the regulatory elements, which could be located in the regions flanking the gene and in the introns, could be located in genomic regions that are 100 kb or larger. Often, large genomic inserts are needed to ensure accurate, integration site-independent transgene expression *in vivo*.

The initial success with large genomic insert transgenesis was achieved with yeast artificial chromosomes (YACs), which contain an average of 500–600 kb genomic DNA and are propagated in yeast as linear artificial chromosomes. YAC transgenes provide the critical, initial proof of principle that large genomic inserts can confer integration site-independent and dosage-dependent transgene expression. However, there are several limitations, which include a high degree of chimerism (i.e., a clone contains genomic DNA from different genomic regions), clonal instability, difficulty isolating a relatively large quantity of intact YAC DNA, and a relatively high frequency of YAC transgene fragmentation, which hinders the more widely used applications of the YAC transgenic approach.

BACs are a circular, bacterial F element-based plasmid containing an average of about 180–200 kb of genomic DNA. The BAC system has several advantages over the YAC system, including clonal stability, low chimerism, and the ability to provide a large quantity of high-quality DNA. Since BACs have become the standard clones used for the majority of genome sequencing projects, the databases for the finished genomes contain not only the primary sequence and the densely annotated genes but also overlapping BAC contigs across the entire genome. To initiate an experiment involving BAC transgenesis, one can readily locate the appropriate BAC for a gene of interest by searching the public genome database and then obtaining the BAC clone from either the public or the private repositories. The ease with which one can map and obtain a BAC clone for transgenic studies greatly facilitates the use of this technology for large-scale transgenic studies.

One limitation for the BAC transgenic approach applies to the relatively small number of mammalian genes that are >200 kb in size. For these genes, the success rate of BAC transgenesis, as measured by the accuracy of transgene expression, becomes relatively low. In summary, the advantages of the BAC transgenic approach greatly outweigh its limitations, and it is for this reason that the BAC approach has become a preferred method for generating transgenic animals for neuroscience research.

Generation of the BAC Transgenic Mice

Selection of BACs for Transgenic Studies

The first step in a BAC transgenic experiment is selecting BACs that maximize the chance for accurate transgene expression *in vivo*. As a general rule, in order to recapitulate the expression pattern of a gene of interest on the BAC (termed target gene) *in vivo*, one should select BACs that contain at least 50 kb of both the 5′ and 3′ flanking genomic sequences, in addition to the genomic sequence spanning the coding region of the target gene. For mammalian genes that cover less than 100 kb, such BACs are relatively easy to locate in existing public databases such as the Genome Browsers at University of California at Santa Cruz, the Ensembl Genome Browser, or the Clonefinders at the National Center for Biotechnology. Appropriate BACs can be obtained from the public repository (BACPAC Resource Center at Oakland Children's Hospital) or from commercial sources (e.g., Invitrogen).

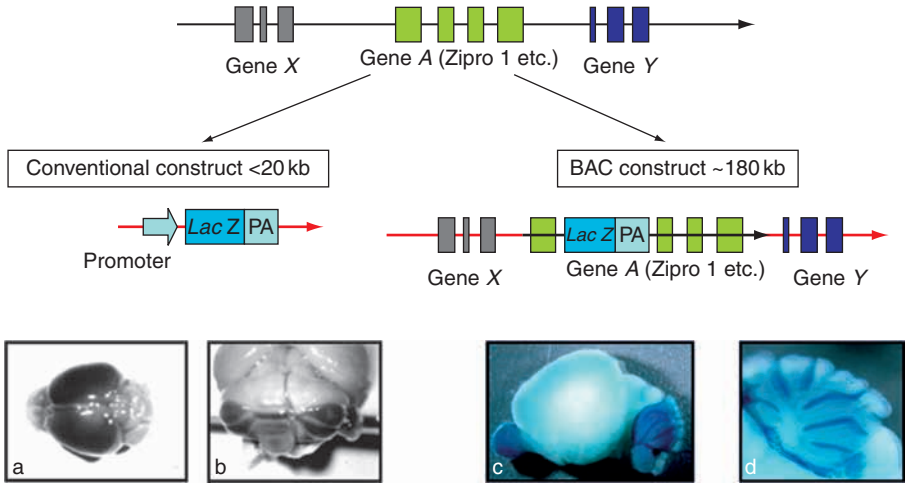

Figure 1 The advantage of BAC transgenic approach in mice. Traditional transgenic constructs, using genomic DNA elements less than 20 kb to drive the transgene expression, frequently result in low level or ectopic patterns of transgene expression (i.e., positional effects). BAC transgenes, with genomic sequence of 180–200 kb, can often confer accurate, transgene dosage-dependent, integration-site-independent transgene expression. (a, b) Examples of ectopic lacZ transgene expression in conventional transgenic studies for murine Ziprol gene. (c, d) Examples of lacZ transgene expression in a Ziprol BAC transgenic mice. Yang XW, Model P, and Heintz N (1997) Homologous recombination based modification in *Escherichia coli* and germline transmission in transgenic mice of a bacterial artificial chromosome. *Nature Biotechnology* 15: 859–865.

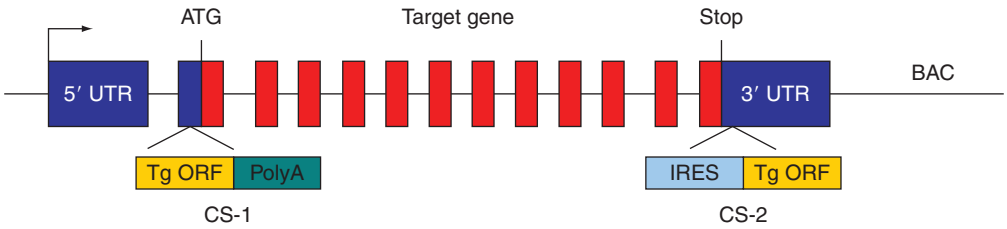

Figure 2 Two basic BAC transgenic construct designs. The first construct (CS-1) is designed to overexpress a transgenic open reading frame (Tg ORF), such as a marker gene or Cre, under the regulatory elements of the target gene on the BAC. A polyadenlyation signal (polyA) is inserted after the Tg ORF to terminate its transcription, which also ensures that the target gene itself is not overexpressed. The second construct (CS-2) is designed to overexpress the target gene and a Tg ORF, with the internal ribosomal entry site (IRES) allowing the translation of the two proteins from one messenger RNA.

Basic BAC Transgenic Construct Design

There are two basic BAC transgenic construct designs that are dependent on different utilities (Figure 2). The most common utility of BAC transgenesis is to selectively express an exogenous gene of interest (the ORF), such as the enhanced green fluorescent protein gene (EGFP) or Cre recombinase, under the regulatory elements of a target gene on the BAC. The target gene itself, however, should not be overexpressed in the transgene. In this case, the basic BAC transgenic construct consists of the ORF followed by a eukaryotic polyA signal (e.g., polyA signals from the human growth hormone or phosphoglycerate kinase), and the entire transgenic cassette is inserted in front of the translational start sequence for the target gene. Another common application for BAC transgenesis is to overexpress a target gene on the BAC and to simultaneously co-express an ORF to label the sites of overexpression. In this case, the most common BAC transgenic construct is to insert an internal ribosomal entrance sequence (IRES), followed by an ORF, after the translational stop sequence and in the last exon of the target gene. The IRES sequence permits translation of two proteins, one encoded by the target gene and one by the ORF, from the same mRNA message. These two basic BAC transgenic construct designs can satisfy the vast majority of BAC transgenic applications.

Engineering BACs by Homologous Recombination in *E. coli*

BAC transgenesis is greatly facilitated by the available set of highly efficient tools used to engineer BACs. There are two widely used homologous recombination-based methods to engineer BACs in *E. coli*: the RecA-based method and the recombineering method

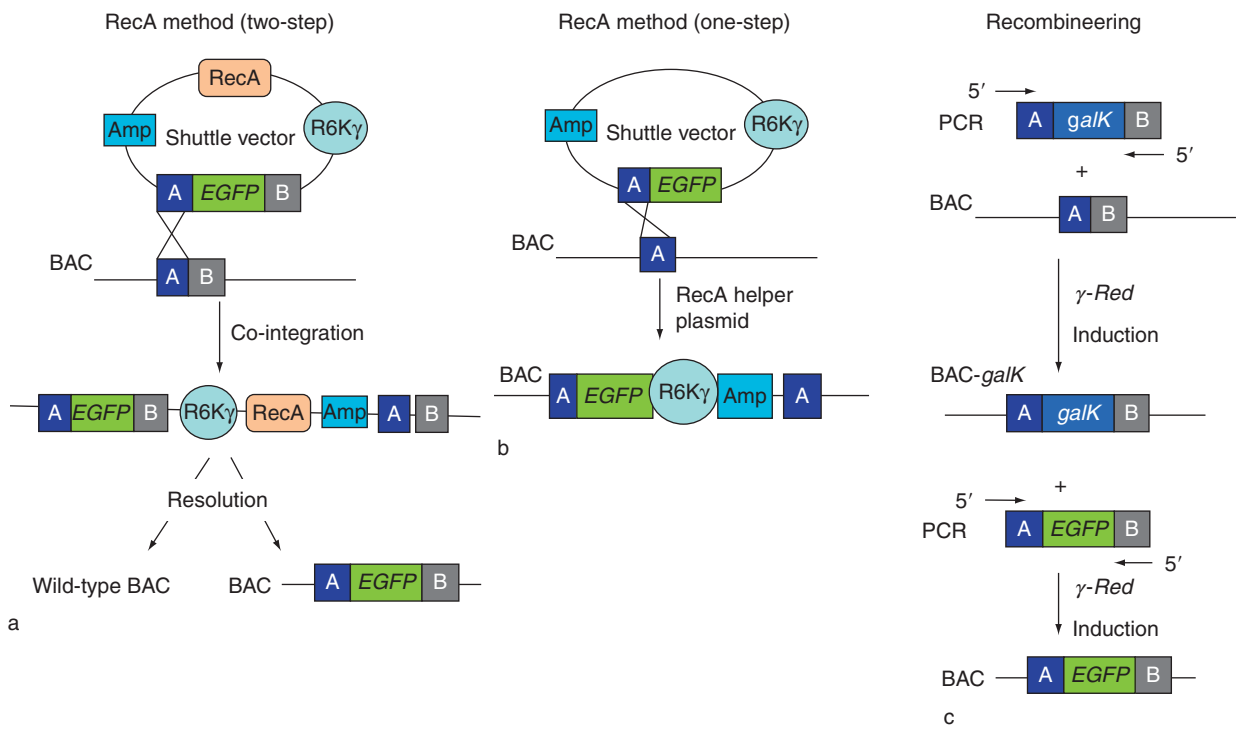

Figure 3 Commonly used methods to engineer BACs by homologous recombination in *E. coli*. (a) Two-step RecA-based homologous recombination method to precisely place a target sequence (e.g., enhanced green fluorescent protein or EGFP) in between two chosen homologous sequences (A and B) on the BAC. Two steps of recombination called co-integration and resolution are required. The shuttle vector in this method includes an R6Kγ origin of replication (cannot replicate in BAC host bacteria but can replicate in other special bacterial strains (e.g., Pir2) that supports this origin of replication), RecA gene (supporting recombination in BAC host bacteria), ampicillin-resistant gene (Amp), and the recombination cassette (i.e., A-EGFP-B). (b) One-step RecA-based method to engineer BAC. In this case, a single homology arm (A) is needed, and RecA is supplied by a helper plasmid. This method is highly efficient and has been used for large-scale BAC engineering and BAC transgenic studies (i.e., GENSAT). (c) Recombineering method of BAC engineering. Transformation of the PCR-amplified linear DNA fragments containing two small homologous sequences (A and B) and *E. coli galK* gene into specialized BAC host bacteria that can inducibly express λ-Red recombinase. Induction of the recombinase allows the insertion of *galK* gene in between A and B sequences on the BAC by homologous recombination. Subsequently, transformation of a second PCR-generated sequence containing the recombination cassette (i.e., A-EGFP-B) into the BAC host bacteria allows the precisely replacement of *galK* gene on the BAC with the EGFP marker gene. (a) Yang XW, Model P, and Heintz N (1997) Homologous recombination based modification in *Escherichia coli* and germline transmission in transgenic mice of a bacterial artificial chromosome. *Nature Biotechnology* 15: 859–865; and Gong S, Yang XW, Li C, and Heintz N (2002) Highly efficient modification of bacterial artificial chromosomes (BACs) using novel shuttle vectors containing the R6Kgamma origin of replication. *Genome Research* 12: 1992–1998. (b) Gong S, Zheng C, Doughty ML, et al. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425: 917–925. (c) Warming S, Costantino N, Court DL, Jenkins NA, and Copeland NG (2005) Simple and highly efficient BAC recombineering using galK selection. *Nucleic Acids Research* 33: e36.

(Figure 3). The first method developed to engineer BACs is the RecA-based method. The original RecA-based method, as well as its subsequent improved version, utilizes a shuttle vector that temporarily introduces the *E. coli* RecA gene into the BAC host bacteria to restore its recombination competence. In the shuttle vector, a recombination cassette is constructed so that two small stretches of homologous DNA sequences are subcloned by polymerase chain reactions (PCRs) from the BAC (each about 200–500 bp), and in between the two homologous sequences are the modifications to be introduced into the BAC (i.e., marker gene insertion, point mutation, or deletion). Once the shuttle vector is introduced into the BAC bacteria, two steps of homologous recombination can be sequentially selected to result in the precise placement of the desired modification in a chosen site on the BAC (i.e., between the two homology sequences; Figure 3(a)).

A variation of the RecA-based BAC engineering method, which is the one-step BAC modification method, permits relatively high-throughput engineering of BACs in a large-scale BAC transgenic project (Figure 3(b)). This method is designed to rapidly introduce a marker gene, such as EGFP or Cre, into many different BACs in order to study *in vivo* gene expression patterns and to label distinct neuronal types. In the one-step method, a minimized shuttle vector is constructed which contains a single homology sequence from the BAC (200–500 bp) followed

by a marker gene of choice and a eukaryotic polyA sequence. The BAC engineering step involves only one homologous recombination step, which is facilitated by RecA expressed from a helper plasmid. The modified BAC has the marker gene placed under the regulation of the target gene regulatory elements, and the extra sequence from the minimized shuttle vector does not interfere with the accuracy of transgene expression *in vivo*. The one-step BAC modification is a highly efficient procedure, with one experienced individual able to modify up to ten different BACs simultaneously over a 2–3-week period. Also this is the method used in the large-scale Gene Expression Atlas BAC transgenic project (GENSAT).

The second widely used method for BAC engineering is the recombineering method, in which the recombination is supported by the bacteriophage-encoded recombination enzymes *RecE* and *RecT* from a prophage or the *Red* genes (*gam*, *bet*, and *exo*) from λ phage. The latest and widely used version for recombineering is illustrated in Figure 3(c). First, the BAC is transformed into a specialized *E. coli* strain that can inducibly express the $\lambda-Red$ recombination enzymes and is deficient in a metabolic enzyme gene, *galK*. Second, a linear recombination cassette is generated using PCR, with two homology arms as short as 50 bp, each flanking the *galK* gene. This can serve as both a positive and a negative selection marker in an *E. coli* host that is *galK*-deficient. The cassette is then transformed into the BAC in the specialized recombination-competent bacteria, and subsequently, induction of recombination leads to precise integration of *galK* into the BAC. In the last step, a second recombination cassette containing the intended modifications (i.e., marker gene insertion, point mutation, or deletion), flanked by the same short homology arms as in the previous step, is introduced into the BAC. Induction of recombination allows the replacement of *galK* by the intended modifications, and the resulting $galK^-$ bacteria can be readily selected.

In summary, both the RecA-based and the recombineering methods are highly efficient and reliable methods to introduce mutations or marker genes into a BAC. These methods have both been applied successfully in large-scale BAC engineering projects.

Characterization and Preparation of BAC for Transgenic Microinjections

Once the BAC is correctly engineered, appropriate steps should be taken to ensure that the recombination procedures do not introduce any unwanted rearrangement or deletions. These unwanted introductions, however, should be a rather rare occurrence for the BAC modification methods described above. Analyses of modified BACS are often done by restriction enzyme fingerprinting on conventional or pulsed field gel electrophoresis.

The final critical step for generation of BAC transgenic mice is the purification of intact BAC DNA for pronuclear microinjection into fertilized mouse oocytes. There are two important issues here. First, one needs to purify intact BAC DNA away from the fragmented BAC DNA up to 100 kb, which is often present in the DNA preps. Transgenic integration of such fragmented DNA may introduce positional effects or the expression of unwanted truncated proteins that may confound the studies. Second, one must reduce impurity in the BAC DNA preps since impurity may reduce the viability of the injected oocytes and reduce the rate for generating the BAC transgenic founders. The most successful method, in terms of consistency in generating transgenic founders, is the large-scale cesium chloride centrifugation preps of BAC DNA.

Applications of the BAC Transgenic Approach in Neuroscience Research

Since the first demonstration of BAC engineering and the BAC transgenic approach by Yang, Model, and Heintz in 1997, tremendous strides have been made toward comprehensive and systematic applications of this powerful *in vivo* genetic tool in the direction of understanding the function and diseases of the mammalian brain. All these applications are based on the assertion that, for the majority of the genes in a mammalian genome (particularly for those that are $<$100 kb in genomic span), an appropriately designed BAC transgene has a very high probability of conferring accurate transgene expression *in vivo*. The confidence for such an assertion is greatly enhanced by the success of the GENSAT large-scale BAC transgenic project. In the GENSAT database, which contains several hundred BAC transgenic lines, the vast majority of them express the EGFP marker gene in the endogenous-like patterns in at least two independent mouse lines. The reproducibility and reliability of the BAC transgenic approach has inspired a myriad of *in vivo* applications affecting many facets of neuroscience, from molecular and cellular analyses to system and behavior studies and brain disease research.

Applications to Study Gene Expression and Gene Function

The BAC transgenic approach has several widely used applications centered on the analyses of expression and function of an individual mammalian gene.

Gene expression and regulation One of the most common applications of BAC transgenesis is to study the expression pattern of an individual gene in the

brain. For this purpose, a reporter gene, such as *EGFP*, *DsRed*, or *lacZ*, is introduced into a resident target gene in the BAC prior to generation of the transgenic mice. Transgene expression can be analyzed either by direct fluorescence or by immunohistochemistry. As a result, one can readily obtain high-resolution images of gene expression in the brain throughout development and into adulthood. The largest study of this kind is the GENSAT project, which has detailed analyses of central nervous system (CNS) BAC transgene expression for several hundred CNS genes. Since engineering a BAC is a very efficient process, one can sequentially introduce multicolor reporter genes into several genes on a single BAC transgene in order to simultaneously visualize the expression patterns of these genes *in vivo*.

Since the BAC contains an intact genomic locus for a gene of interest, another common application of BAC transgenesis is to dissect regulatory elements that are critical for gene expression *in vivo*, such as enhancers, suppressors, locus control regions, insulator elements, and imprinting elements. In these studies, mutant BACs (with the marker gene) carrying a mutation in the putative regulatory element are generated to assess its impact on transgene expression.

Dominant genetic studies Dominant genetic studies, such as overexpression of the wild-type or dominant-negative genes variants, can provide valuable functional insights into these genes. For example, the granule-cell-specific zinc finger transcription factor *Zipro1* has no apparent loss-of-function phenotypes, but BAC-mediated increase in *Zipro1* gene dosage reveals its critical role in controlling granule cell proliferation. One advantage of the BAC-mediated dominant genetic approach is that the transgene is expressed in an endogenous pattern and usually at relatively physiological levels (less than fivefold overexpression), thus avoiding the confounding effects of aberrant transgene expression.

BAC rescue of the loss-of-function mutants A classical application of BAC transgenesis is in the rescue of the loss-of-function mutants in mice, which facilitated the positional cloning of loss-of-function mutant genes. As a natural extension of the genetic rescue approach, one can also perform *in vivo* structure–function analyses of a target gene on the BAC by engineering specific mutations (e.g., mutating a protein domain or a phosphorylation site) prior to generating the BAC transgenic mice in order to evaluate whether the mutant transgene can rescue the null mutant phenotypes. Another potential utility for the BAC rescue approach is to study the functional differences of single-nucleotide polymorphisms (SNPs), which naturally occur among different individuals in a species and some of which may contribute to variations in brain function and/or brain disorders. To evaluate the differential functions of SNPs *in vivo*, one can generate BAC transgenes carrying the individual SNPs and then study the ability of these SNP transgenes to rescue the null phenotype. Such studies may facilitate the understanding of individual SNPs on brain function and brain diseases.

Protein tagging for *in vivo* proteomics An important venue to unravel gene function is through identification of interacting proteins. Although interacting proteins can be identified using biochemical purifications in stable cell lines or through yeast two-hybrid screening, a challenge remains to purify and identify the endogenous interacting protein complexes from the mammalian brains. Currently, such studies are limited to a few protein complexes for which excellent antibody reagents and/or purification schemes are available. With the advance in BAC transgenesis, small protein epitope tags (such as FLAG-, HA-, His-, or Myc-tags) or even larger protein tags (such as EGFP or Protein-A) can be readily engineered as an in-frame fusion to a target protein expressed from the BAC. Subsequent affinity purification of the epitope-tagged target proteins from the transgenic mouse brains can facilitate co-purification of interacting protein complexes, which can be identified by sensitive mass spectrometry techniques. Although this BAC-mediated protein tagging and *in vivo* proteomic method is still at its infancy, it represents an active area of research and may have great promise to advance understanding of CNS gene function.

Applications of BAC Transgenesis to Study Neurons and Neural Circuits

The second major area of application for BAC transgenesis is to selectively express an exogenous gene of interest (such as EGFP or Cre) in a particular neuronal type and/or a particular neural circuit defined by a driver gene on a BAC (i.e., *DAT* in dopaminergic neurons). In this case, the BAC transgene is used as a vehicle to provide genetic access to a specific cell type in the brain in order to study its roles in brain function or brain diseases. The success of the GENSAT project has already generated a rich data bank of BACs that can reproducibly target transgene expression in many different neuronal types, as defined by morphology, anatomic locations, or neurochemical characteristics. These BACs provide an immediate entry point for a variety of sophisticated *in vivo* analyses of these neurons and neural circuits.

Developing Cre/Flp recombinase mice BAC transgenesis has been widely used to develop mouse lines that express Cre or Flp recombinase in a given neuronal lineage (i.e., Purkinje cells in the cerebellum or

hair cells in the cochlea). These recombinase mice will facilitate the use of the conditional gene-targeting strategy to selectively delete or activate gene expression in a neuronal lineage in the brain.

***In vivo* electrophysiology and imaging** In many brain regions, different neuronal types are often intermixed and cannot be readily identified in live brain tissues or living organisms. BAC transgenic expression of fluorescent proteins such as EGFP facilitates detection of specific, genetically defined neuronal populations in live brain slices or living animals with an imaging window placed in the skull. Electrophysiological studies of these genetically labeled neurons in brain slices or in living animals (e.g., for striatal projection neuron subtypes and cortical parvalbumin-positive interneurons) has led to rapid advances in understanding the physiology of these neuronal types.

Recent rapid advances in genetically encoded fluorescent reporter genes have resulted in a rich repertoire of proteins that can be multicolored and can report in real-time neuronal activities, calcium signaling, receptor/channel function, and so forth. BAC transgenic delivery of these latest fluorescent reporters, combined with the latest *in vivo* imaging techniques such as two photon imaging, holds great promise in aiding the understanding of how different neuronal types function at a system level and/or in the intact behavior of animals.

Cell type-specific expression profiling A challenge in studying gene expression profiles in the mammalian brain with the microarray approach is the complexity of the brain tissues. One solution is to isolate relatively pure neuronal populations from the brain for extraction of RNA and microarray analyses. With the advent of BAC transgenesis, individual neuronal types can be labeled with fluorescent proteins while the neurons are still alive. Methods have been developed to physically dissociate and purify these genetically labeled neurons for gene expression profiling, one of which is called fluorescence-activated cell sorting (FACS)-array (Figure 4). In an example of this method, EGFP-labeled striatal projection neuron subtypes are enzymatically dissociated and subsequently purified using FACS. Only live EGFP (+) neurons are sorted since the dead neurons are labeled with propidium iodide and are sorted against. Subsequent RNA purification and microarray studies have led to identification and verification of a novel set of genes differentially enriched in these clinically important neuronal populations. Besides FACS, other approaches to purifying genetically labeled neurons, such as laser-captured microdissection or hand picking under fluorescent microscope, have been applied to perform cell type-specific gene expression profiling.

Transsynaptic genetic tracing of a neural circuit Traditional methods to map the neuronal circuitry rely on stereotactic injection of anterograde or retrograde

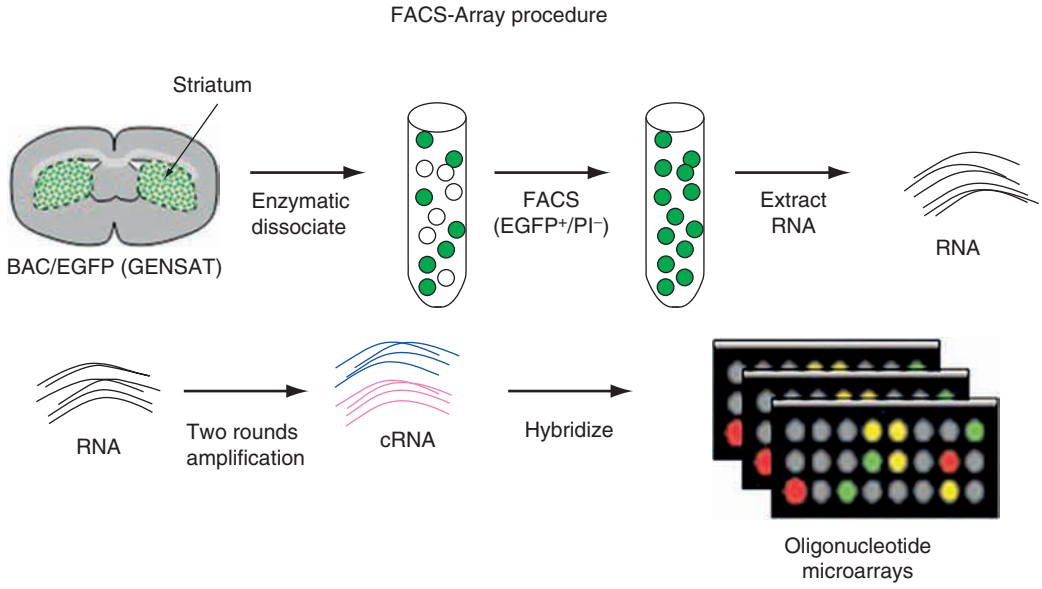

Figure 4 FACS-array procedure to profile genetically labeled neurons in the mouse brain. Striata from transgenic mice selectively expressing GFP in striatonigral or striatopallidal medium spiny neurons (MSNs) were dissected and enzymatically dissociated. The dissociated dead striatal neurons were labeled with propidium iodide (PI). The live GFP-labeled MSNs were isolated by Fluorescent Activated Cell Sorting (FACS) to select for GFP (+) and PI (−) neurons. The RNA purified from the sorted neurons was amplified *in vitro* to generate cRNA probes. The cRNA from two different MSN subtypes, labeled with either Cy-3 or Cy-5 fluorescent dyes, were hybridized onto the mouse microarrays. Adapted from Lobo MK, Karsten SL, Gray M, Geschwind DH, and Yang XW (2006) FACS-array profiling of striatal projection neuron subtypes in juvenile and adult mouse brains. *Nature Neuroscience* 9: 443–452.

chemical tracers into distinct regions in the brain. The injection sites usually have many different types of neurons, thus precluding the precise understanding of the connectivity of a single neuronal type (origin neuron) in a heterogeneous brain region. To overcome these problems, genetically encoded anterograde or retrograde transsynaptic tracers have been developed and can be delivered by BAC transgenesis to a precise neuronal type in the brain. For anterograde labeling of neurons that the origin neurons project to, one can use BAC transgenesis to express wheat germ agglutinin or barley lectin in the origin neuron. For retrograde labeling of neurons that synapse onto the origin neurons, one can express an engineered fusion protein called GFP-TTC in the origin neurons, which combines the GFP with a nontoxic, C-terminal fragment of tetanus toxin. An advantage of these genetically encoded tracers is that they can usually transfer across several synapses, thus providing an opportunity to elucidate a neural circuit either upstream or downstream of the origin neurons.

Another elegant approach has been developed to perform retrograde and transsynaptic mapping of the neuronal circuitry using a pseudorabies virus-based strategy. In this approach, BAC transgenic mice are used to express Cre in the origin neurons, and then genetically engineered pseudorabies viruses are stereotactically injected into the brain region containing the origin neurons. The virus expresses EGFP only on Cre-mediated excision of a transcription termination sequence in these neurons. Since the pseudorabies viruses can be transferred in a retrograde, transsynaptic fashion, the upstream neurons projecting to the origin neurons can be illuminated. Similar to the other approaches, the pseudorabies virus strategy can be transferred through multiple synapses to label the upstream circuits.

Inducible and/or reversible manipulation of neuronal function A powerful tool to understand how a specific neuronal type or neural circuit contributes to brain function is through inducible and/or reversible manipulations of these neurons or neural circuits, including inducible gene expression, cell ablation, or reversible activity silencing. Genetically encoded proteins have been developed to perform such sophisticated experiments. For inducible gene expression, the tetracycline-inducible system and inducible Cre or Flp recombinase systems have been developed. For cell-specific ablation, an immunotoxin-based strategy has been developed so that the interleukin-2 receptor α-subunit is selectively expressed in the target neurons. Subsequent stereotactic microinjection of an immunotoxin binding to this receptor leads to ablation of the neurons. Several genetic strategies using modified potassium channels, glycosylphosphatidylinositol anchor-tethered neurotoxins, bacterial chloride channels, and modified G-protein-coupled receptors responsive only to synthetic ligands have also been developed for ligand-induced, reversibly silencing neuronal types. Combining BAC transgenic delivery of these genetically encoded molecular switches with sophisticated physiological and behavioral studies may greatly advance understanding of how the brain functions on a system level.

Applications to Model Dominant Human Disorders and Study Disease Pathogenesis

An important application of BAC transgenesis is to develop mouse models of human disorders due to dominant genetic mechanisms. For a large number of such disorders, including several inherited neurodegenerative disorders such as familial Alzheimer's disease, familial Parkinson's disease, polyglutamine disease (i.e., Huntington's disease), and amyotrophic lateral sclerosis (ALS), transgenic mouse models have proven to be powerful tools to examine disease mechanisms and to test therapeutic strategies. Since large-insert genomic transgenes such as YACs and BACs often contain the intact genomic locus and direct accurate, endogenous-like expression of the disease-causing proteins, the phenotypes from these models are often more accurate and robust than conventional transgenic models. For example, YAC transgenic models of Huntington's and Alzheimer's disease and spinal and bulbar muscular atrophy all exhibit robust diseaselike features. As summarized above, the BAC transgenic approach is now being applied to develop human disease models. As an example, a robust new BAC transgenic mouse model of Huntington's disease (BACHD) has been developed and characterized. BACHD mice express full-length mutant human huntingtin under endogenous regulation and can rescue the lethality of the murine huntingtin loss-of-function allele. These mice develop progressive motor deficits and selective neuropathology, reminiscent of Huntington's disease patients.

Besides developing mouse models, BAC transgenesis is promising in the study of disease mechanisms and the testing of therapeutic strategies. One such potential application would be to test genetic modifiers of the disease. Potential genetic modifiers of disease pathogenesis can be readily identified in genetic model organisms suitable for a large-scale genetic screen, such as yeast, *Drosophila*, and *C. elegans*. A substantial number of these modifiers act through dominant genetic mechanisms and can therefore be tested in mice through BAC transgenic overexpression. Breeding modifier mice to the disease mouse models may reveal the disease suppression efficacy in the mammalian brain. Systematic application

of BAC transgenesis may enable rapid validation of a large number of genetic modifiers as potential therapeutic targets for brain disorders.

Summary

In summary, the BAC transgenic approach has emerged as a powerful *in vivo* genetic tool to study brain function at all levels, from the molecules and cells to the circuitry in an intact behaving animal. The advent of large-scale BAC transgenesis, particularly the GENSAT BAC transgenic project, has greatly enhanced the utility of the BAC approach for all facets of neuroscience research. The BAC approach has also been systematically applied to develop robust new animal models for human brain disorders and to develop novel therapeutic strategies. Therefore, it is safe to predict that a bright future is yet to come for the BAC transgenic approach in neuroscience research.

See also: Alzheimer's Disease: Transgenic Mouse Models; Animal Models of Motor and Sensory Neuron Disease; Animal Models of Huntington's Disease; Animal Models of Parkinson's Disease; BAC Transgenesis: Cell-Type Specific Expression in the Nervous System; Engineering Viruses for CNS studies; Microarray use for the Analysis of the CNS; Transgenic Models of Neurodegenerative Disease; Viral Vectors in the CNS.

Further Reading

Callaway EM (2005) A molecular and genetic arsenal for systems neuroscience. *Trends in Neurosciences* 28: 196–201.

Copeland NG, Jenkins NA, and Court DL (2001) Recombineering: A powerful new tool for mouse functional genomics. *Nature Reviews Genetics* 2: 769–779.

DeFalco J, Tomishima M, Liu H, et al. (2001) Virus-assisted mapping of neural inputs to a feeding center in the hypothalamus. *Science* 291: 2608–2613.

Dymecki SM and Kim JC (2007) Molecular neuroanatomy's 'Three Gs': A primer. *Neuron* 54: 17–34.

Gong S and Yang XW (2005) Modification of bacterial artificial chromosomes (BACs) and preparation of intact BAC DNA for generation of transgenic mice. *Current Protocols in Neuroscience*. Unit 5.21.

Gong S, Yang XW, Li C, and Heintz N (2002) Highly efficient modification of bacterial artificial chromosomes (BACs) using novel shuttle vectors containing the R6Kgamma origin of replication. *Genome Research* 12: 1992–1998.

Gong S, Zheng C, Doughty ML, et al. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425: 917–925.

Hadjantonakis AK, Dickinson ME, Fraser SE, and Papaioannou VE (2006) Technicolour transgenics: Imaging tools for functional genomics in the mouse. *Nature Genetics* 38: 758–769.

Heintz N (2001) BAC to the future: The use of BAC transgenic mice for neuroscience research. *Nature Reviews Neuroscience* 2: 861–870.

Kleinfeld D and Griesbeck O (2005) From art to engineering? The rise of *in vivo* mammalian electrophysiology via genetically targeted labeling and nonlinear imaging. *PLoS Biology* 3: e355.

Lee EC, Yu D, Martinez de Velasco J, et al. (2001) A highly efficient *Escherichia coli*-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics* 73: 56–65.

Lobo MK, Karsten SL, Gray M, Geschwind DH, and Yang XW (2006) FACS-array profiling of striatal projection neuron subtypes in juvenile and adult mouse brains. *Nature Neuroscience* 9: 443–452.

Warming S, Costantino N, Court DL, Jenkins NA, and Copeland NG (2005) Simple and highly efficient BAC recombineering using galK selection. *Nucleic Acids Research* 33: e36.

Watase K and Zoghbi HY (2003) Modelling brain diseases in mice: The challenges of design and analysis. *Nature Reviews Genetics* 4: 296–307.

Yang XW and Gong S (2005) An overview on generation of BAC transgenic mice for neuroscience research. *Current Protocols in Neuroscience*. Unit 5.20.

Yang XW, Model P, and Heintz N (1997) Homologous recombination based modification in *Escherichia coli* and germline transmission in transgenic mice of a bacterial artificial chromosome. *Nature Biotechnology* 15: 859–865.

Yang XW, Wynder C, Doughty ML, and Heintz N (1999) BAC-mediated gene-dosage analysis reveals a role for Zipro1 (Ru49/Zfp38) in progenitor cell proliferation in cerebellum and skin. *Nature Genetics* 22: 327–335.

Zhang Y, Buchholz F, Muyrers JP, and Stewart AF (1998) A new logic for DNA engineering using recombination in *Escherichia coli*. *Nature Genetics* 20: 123–128.

Relevant Websites

http://bacpac.chori.org – BACPAC Resources Center at Oakland Children's Hospital.

http://www.ensembl.org – Ensembl Genome Browser.

http://www.genome.ucsc.edu – Genome Browsers at University of California at Santa Cruz.

http://www.ncbi.nlm.nih.gov – GENSAT at Pubmed.

http://www.gensat.org – GENSAT at Rockefeller University.

Balance and Posture Control

L M Nashner, NeuroCom International, Inc., Clackamas, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

Balance: Definition

A person senses a state of 'balance' when perceptions of self-orientation in relation to the environment are in dynamic equilibrium. This expansive definition of balance combines conscious and automatic motor controls, conscious and unconscious brain processes, with sensory information from the visual, inner ear vestibular, and proprioceptive systems. The term balance more commonly describes the mechanisms maintaining stability during performance of daily life activities, with the majority of balance research being focused on mechanisms controlling gaze and postural stabilization. Clinical research provides a better understanding of the impact of sensory, musculoskeletal, and central nervous system diseases on gaze and posture.

Concepts for Understanding Gaze and Posture

Maintaining gaze and postural stability during daily life activities is complex and involves the coordinated actions of multiple biomechanical, sensory, motor, and brain processes. The gaze system positions the eyes relative to visual objects of interest in the environment during activities involving combinations of body, head, and object movement. The posture system positions the body center of gravity (COG) relative to the base of support during activities involving combinations of body, base of support, and visual surround movement.

Figures 1 and 2 illustrate the principal components of the gaze and posture systems, respectively. Both systems are composed of biomechanical, sensory, and motor components interconnected by afferent and efferent brain pathways. The two are organized similarly in that both combine information from multiple sensory sources, respond through the coordinated actions of multiple motor processes, and involve neural coordinating mechanisms of the brain. Both systems perform within varieties of sensory and mechanical environments by adaptively modifying how prior knowledge and current sensory information are mapped into coordinated motor actions. The two stabilizing systems are also interdependent. Gaze is difficult to stabilize unless the body is also stable. Stable gaze is necessary for accurate visual orientation information, which is in turn an important sensory component for maintaining postural stability.

Gaze System

Eye Movement Controls

To maintain visual accuracy during dynamic activities, the direction of gaze must be stabilized on visual objects of interest. Dynamic visual acuity is best when a visual object is stabilized on the fovea, a small area centered on the retina containing the highest density of cone receptor cells. Acuity degrades rapidly when the object is displaced more than a few degrees from the fovea or when the visual angle changes faster than a few degrees per second relative to the fovea. The following four movement systems are the principal components of gaze control:

1. *Vestibuloocular reflex (VOR)*. The VOR utilizes angular velocity information from the inner ear vestibular system to reflexively rotate the eyes in opposition to the rotational movements of the head. Moving the eyes in opposition to the head helps stabilize gaze when viewing a fixed object while the head moves. VOR movements commence within 15 ms of head movement and continuously move the eyes at velocities between approximately $20^\circ\,s^{-1}$ and $120^\circ\,s^{-1}$.
2. *Smooth pursuit*. The pursuit system maintains gaze on a moving visual object by driving the eyes with position feedback information derived from the conscious perception of the object. The system generates continuous movement commencing within 50 ms of object movement and maintaining the object on the fovea at velocities up to $20^\circ\,s^{-1}$. Unlike the VOR, the pursuit system requires the presence of a visual object.
3. *Saccadic*. The saccadic system generates rapid, discrete eye movements that position visual objects onto the fovea. In contrast to continuous VOR and pursuit movements, saccadic movements begin abruptly, move the eyes at velocities up to $300^\circ\,s^{-1}$, and then abruptly terminate when the object is centered on the fovea. The saccadic system can generate movements to both visualized and imagined objects, including those anticipating an object's appearance or movement.
4. *Head on trunk*. Coordinated movement of the head relative to the trunk is another important mechanism helping to maintain gaze stability. When the body moves backward, forward, and from side to side during standing and walking activities, the head rotates in the opposite direction, thereby minimizing

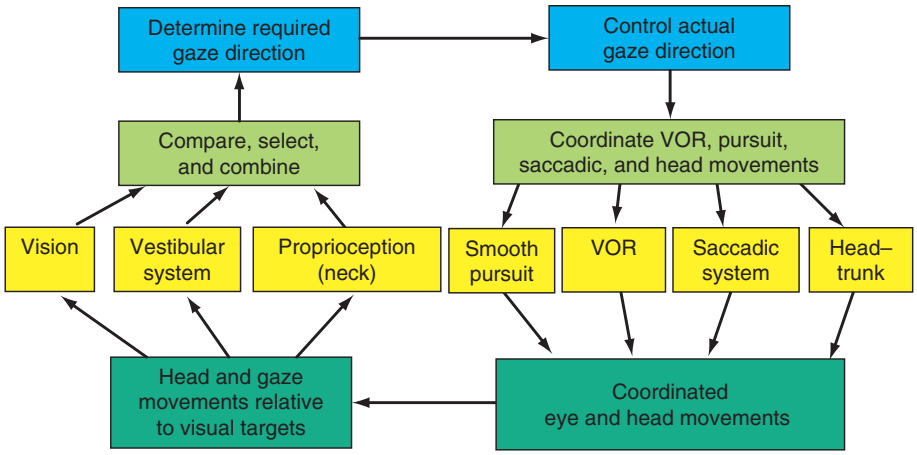

Figure 1 Principal components of the gaze stabilization system (VOR, vestibuloocular reflex).

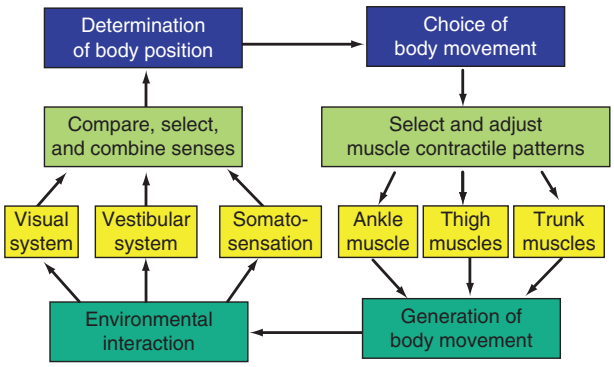

Figure 2 Principal components of the posture stabilization system.

rotational motions of the head relative to the external environment. These head movements simplify the task of gaze control when the body is moving by providing a stable platform from which to launch eye movements.

The following examples illustrate common daily life interactions among the movement systems: When seated in a moving vehicle and reading a roadside sign through the windshield, the pursuit and VOR systems work together to maintain gaze at all velocities below $120\,^{\circ}s^{-1}$. When the vehicle moves more rapidly, the saccadic system intermittently intervenes to accelerate the eyes back onto the sign. When gaze is redirected into the vehicle to read a dashboard instrument, the brain automatically suppresses the VOR component, and the pursuit system stabilizes gaze on the instrument. When standing at curbside and reading a sign on the passing vehicle, the pursuit system continuously stabilizes gaze at vehicle velocities below $20\,^{\circ}s^{-1}$, and at higher velocities the saccadic system intermittently intervenes to accelerate the eyes back onto the sign.

Posture System

Biomechanics of Posture

Erect standing and walking activities require that the body's center of gravity be maintained vertically over the base of support. The COG is a theoretical point about which the forces of gravity appear to act, a point in humans located in the lower abdominal area of the trunk. In the absence of active controls, standing and walking postures are inherently unstable due to the influences of gravity and external disturbing forces. When the COG moves outside the base of support, the person falls unless a step or stumble rapidly repositions the base of support under the COG. Controlling postural stability is similar while walking. The placement of successive steps and the forward progression of the COG are coordinated to maintain their relative vertical positions.

The base of support is the area of contact between the feet and the support surface. The term 'limits of stability' (LOS) refers to a perimeter of positions within which the vertical projection of the COG falls within the base of support. An adult standing on a flat, firm surface with feet spaced comfortably apart can position the COG approximately 12.5° back to front, and 16° left to right, and remain within an elliptically shaped LOS area. Alterations in either foot placement or surface topography alter the LOS. Heel-to-toe and tiptoe walking dramatically reduce the width and length, respectively, of the LOS and make stability significantly more difficult to control in the dimensionally reduced directions. The LOS dimensions change randomly with individual steps while walking on an irregular surface such as large stones.

The position of the COG over the support base during stance and walking is influenced by the relative positions of the ankles, knees, hips, arms, and head. A variety of postures, not just erect stance,

position the COG over the base of support. Examples of nonerect, balanced postures include the squatting position and positions in which the waist is bent forward and the ankles are rotated backward. The latter postures are commonly used when bending down to retrieve objects from the floor.

Postural Movement Patterns

The earliest movement response with a functional impact on postural stability is termed the 'automatic postural movement.' Following an external disturbance, the automatic movement commences in leg muscles at 85–95 ms latencies when recorded by electromyography, and at 120–150 ms latencies recorded as force exerted against the support surface. Automatic movements are involuntary, triggered by muscle proprioceptive inputs, and are mediated by brain and subcortical pathways.

Automatic movements are coordinated into patterns of lower leg, thigh, trunk, and neck muscle responses. However, the pattern of an automatic movement elicited by the same external disturbance differs, depending on the prior experience and expectations of the person at the time of the disturbance. Figure 3 compares ankle, thigh, and lower trunk muscle electromyogram (EMG) patterns elicited by the identical backward and forward displacements of the support surface. Panels a and b illustrate responses following practice standing on a surface longer than the length of the feet (normal). Panels c and d illustrate responses of the same person, a few minutes later, after practicing stance on a support surface shorter than the length of the feet (short). The human stick diagrams illustrate the corresponding body movements. During initial practice trials on each surface, EMG and body movement patterns resemble combinations of these two patterns.

With sufficient practice balancing on a given surface, identical forward and backward translations of the surface elicit very different EMG and body movement patterns. The responses illustrated in Figure 3 are consistent with two biomechanically distinct patterns for moving the COG of a multisegmented body forward and backward. Given the existence of the distinct patterns, however, this does not rule out the use of hybrid combinations:

1. *Ankle strategy.* On the flat surface, the COG is moved by rotating the body as an approximately rigid

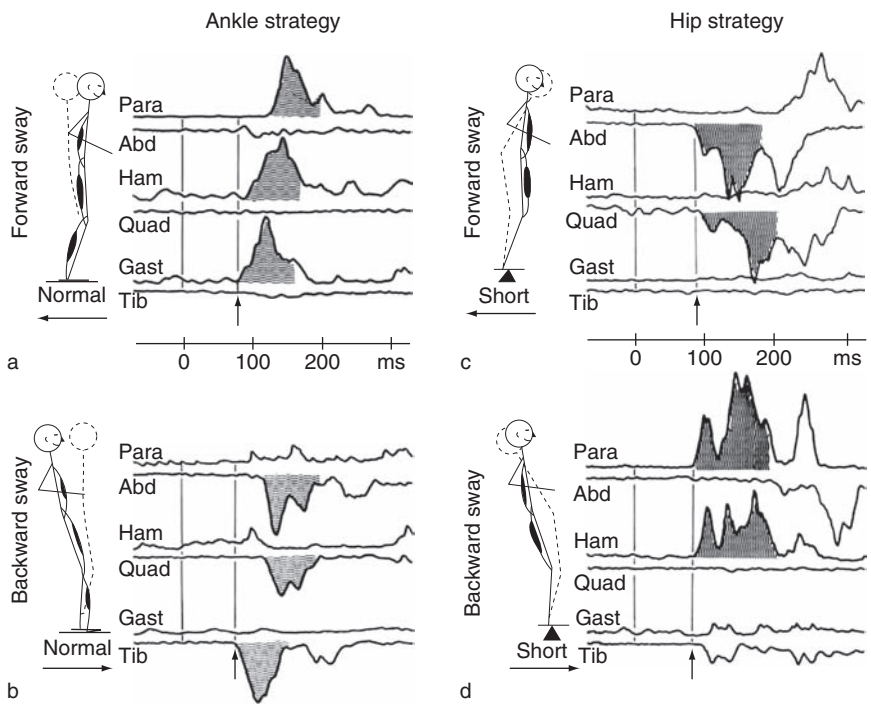

Figure 3 Postural movement strategies elicited by backward support surface translations (a, c) and by forward translations (b, d). Electromyogram traces are full-wave-rectified surface-electrode recordings of the gastrocnemius (Gast), tibialis anterior (Tib), hamstrings (Ham), quadriceps (Quad), paraspinal (Para), and abdominal (Abd) muscles of the leg and lower trunk. All electromyogram recordings were performed in the same test session with the same amplifier gains. In panels a and b, the person was well practiced at standing on a support surface longer than the length of the feet (normal). In panels (c) and (d), the person was well practiced at standing on a support surface shorter than the length of the feet (short). From Horak FB and Nashner LM (1986) Central programming of postural movements. Adaptation to altered support-surface configurations. *Journal of Neurophysiology* 55: 1369–1381.

mass about the ankle joints, hence the term 'ankle strategy.' This strategy rotates the body by exerting torque about the ankle joints while stabilizing the proximal knee and hip joints. However, because the feet are much shorter than the body height, the ability to generate torque about the ankles is limited. Hence, the ankle strategy is effective only when COG movements are relatively slow and the COG is positioned well within the LOS perimeter.

2. *Hip strategy.* On the narrow beam support surface, the COG moves by rapidly rotating the hip joints, hence the term 'hip strategy.' This strategy relies on the inertia of the trunk rapidly accelerating in one direction to generate a horizontal shear force against the support surface and move the COG in the opposite direction. Because there are no biomechanical limitations on the horizontal shear force, hip strategy is effective when standing on narrow support bases and/or the COG is positioned near the LOS perimeter. However, conditions that limit horizontal force, such as standing on ice, render the hip strategy ineffective.

Movement Control Theory

Movement control theory addresses the question: How can the brain, within less than 0.1 s, select the appropriate pattern and appropriately coordinate the actions of many muscles located throughout the body? One theory is that the brain simplifies the task by selecting in advance, from among a limited repertoire of discrete patterns, one or a combination of several patterns most appropriate to the task conditions. In other words, an automatic movement pattern is the expression of a 'postural set' reflecting recent experience with and expectations of the task. One approach to testing this theory used principal factors analysis to analyze the EMG patterns of leg muscles recorded during exposure to a variety of posture control challenges. Results supported the theory in that 80% of the EMG activity patterns were attributable to three principal factors, two resembling ankle and hip strategy patterns and a third involving vertical body movement.

Sensory Inputs to Posture

Sensing the position of the COG relative to the base of support requires combinations of visual, vestibular, and somatosensory (tactile, deep pressure, joint receptor, and muscle proprioceptive) inputs, as summarized in Table 1. When standing on firm surfaces and surrounded by fixed visible objects, information from the three senses is accurate and provides redundant information. However, postural stability cannot be maintained over the course of daily life activities using a single sense or combination of senses. Changes in environmental conditions eliminate or render inaccurate and inappropriate for postural control the orientation information derived from individual senses. This principle is illustrated by the following examples:

1. *Support surface*. Under fixed surface conditions, the somatosensory input derived from contact with the support surface provides a majority of the necessary orientation information. When the support surface is compliant, irregular, or moves, the somatosensory input is inaccurate for postural control, and visual and vestibular inputs are substituted. Dry beach sand, soft wet grass, and thick carpeting are examples of inaccurate surface conditions.

2. *Visual surround*. Objects moving within the visual surround disrupt the visual sense of orientation, because it becomes difficult to distinguish between self- and object motions. Under these conditions, visual information is de-emphasized, and greater reliance is placed on somatosensory and vestibular inputs. An example of a daily life activity is standing next to a large bus that suddenly accelerates forward, producing a momentary disorientation or unsteadiness until the brain determines that the bus is moving forward, and the body is not moving backward.

3. *Inertial–gravitational reference*. Maintaining stability is most challenging when both visual and somatosensory inputs are inaccurate or unavailable, because only the inertial–gravitational reference of the vestibular system is reliable. Under these conditions normal individuals can still maintain stability, but their sway is significantly greater. Individuals

Table 1 Utilization of the senses for balance

Property	*Sense*		
	Somatosensory system	*Vision*	*Vestibular system*
Orientation reference	Support surface	Surrounds	Gravity/inertial space
Conditions favoring use	Fixed, firm surface	Fixed visual surround	Irregular and/or moving support/visual surrounds
Conditions disrupting use	Irregular/compliant support surfaces	Moving surrounds/darkness	Unusual motion environments

with impaired vestibular function, by contrast, experience instability and frequent falls when simultaneously deprived of useful somatosensory and visual information.

Organization of Senses

The mechanisms used by the brain to organize sensory information for postural control have been studied by measuring the stability of a person exposed to conditions in which the support surface, the visual surround, or both are moved in direct relation to the spontaneous sway of the person. A technique, referred to as 'sway-referencing,' creates an orientation reference that, while perceptually correct, is functionally inappropriate for determining orientation relative to gravity. The complete protocol of six sensory conditions, including all combinations of sway-referenced conditions and eye closure, termed the sensory organization test (SOT), is illustrated in Figure 4. The person stands for three 20 s trials of each sensory condition. COG sway is recorded using a forceplate and a percentage equilibrium score calculated for each trial, with 0% indicating loss of balance and 100% indicating perfect stability. Under the first-three fixed-support SOT conditions, the person stands, eyes open in a fixed surround (condition 1), eyes closed (condition 2), and eyes open in the sway-referenced visual surround (condition 3). SOT conditions 4–6 repeat these three visual conditions in the same order, with the support surface sway-referenced.

Figure 5 illustrates the SOT results of a typical normal individual. The composite equilibrium score is a weighted average of scores for the six sensory conditions. The gray background area of the graph delineates the ranges of abnormal equilibrium scores for adults aged 20–59 years. With the support surface fixed (conditions 1–3), spontaneous COG sway motions are small and stability scores are correspondingly high. Neither elimination nor disruption of vision has significant effects on stability. With the support surface sway-referenced to eliminate useful somatosensory inputs (conditions 4–6), by contrast, all stability scores decline and vision plays a more significant role. With useful vision (condition 4), postural stability averages 80%, while closing the eyes (condition 5) reduces stability scores to 65%. Stability is similar with the eyes closed (condition 5) or open within a sway-referenced visual surround (condition 6), indicating that an inappropriate visual input is ignored.

Because the vestibular system is an inertial–gravitation reference, it is not readily manipulated by external conditions. One method used to quantify the influence of vestibular inputs and to verify that SOT condition 5 isolates vestibular control is to assess patients with clinically documented bilateral vestibular losses. As exemplified in Figure 6, a patient with compensated bilateral vestibular loss maintains balance approximately as well as do normal individuals under SOT conditions 1–4, but loses balance repeatedly under conditions 5 and 6. These results indicate that the vestibular input, while not critical for standing when functionally useful somatosensory and/or visual inputs are available, is critical when combinations of sway-referenced surface, sway-referenced visual, and eye closure (conditions 5 and 6) eliminate both alternatives.

Although the vestibular system plays a secondary role in controlling balance when accurate somatosensory and/or visual inputs are available, information from the vestibular system is essential when support surface and visual conditions unexpectedly change. Commonly occurring examples include walking on tilted, irregular, and/or compliant support surfaces, and moving in the dark and in environments containing independently moving objects such as a shopping mall or a heavily traveled street.

Accurate vestibular input is also required to coordinate 'head on trunk' movements and to perform independent head movements while maintaining postural stability, two processes critical to gaze stability. As described in the earlier section on postural movement patterns, coordination of head movements relative to the trunk stabilize the rotational orientation of the head within the environment. Vestibular loss patients, by contrast, fix the position of the head relative to the trunk rather than to the external environment, thereby limiting visual acuity during common daily life activities such as observing a moving person or vehicle while walking the opposite direction.

Sensory Organization Theory

A generally accepted theory of sensory balance control is that the brain responds to changes in the conditions of the task by adaptively reweighting the influence of individual sensory inputs to balance. Experimental evidence supporting the theory that contributions of visual and somatosensory inputs are adaptively reweighted in response to sway-referenced conditions has been provided by mathematical models simulating the performance of actual persons standing under SOT conditions.

Clinical Applications

Medical Necessity

Poor balance is a major medical problem, the principal reason for over 7 million clinic visits per year in

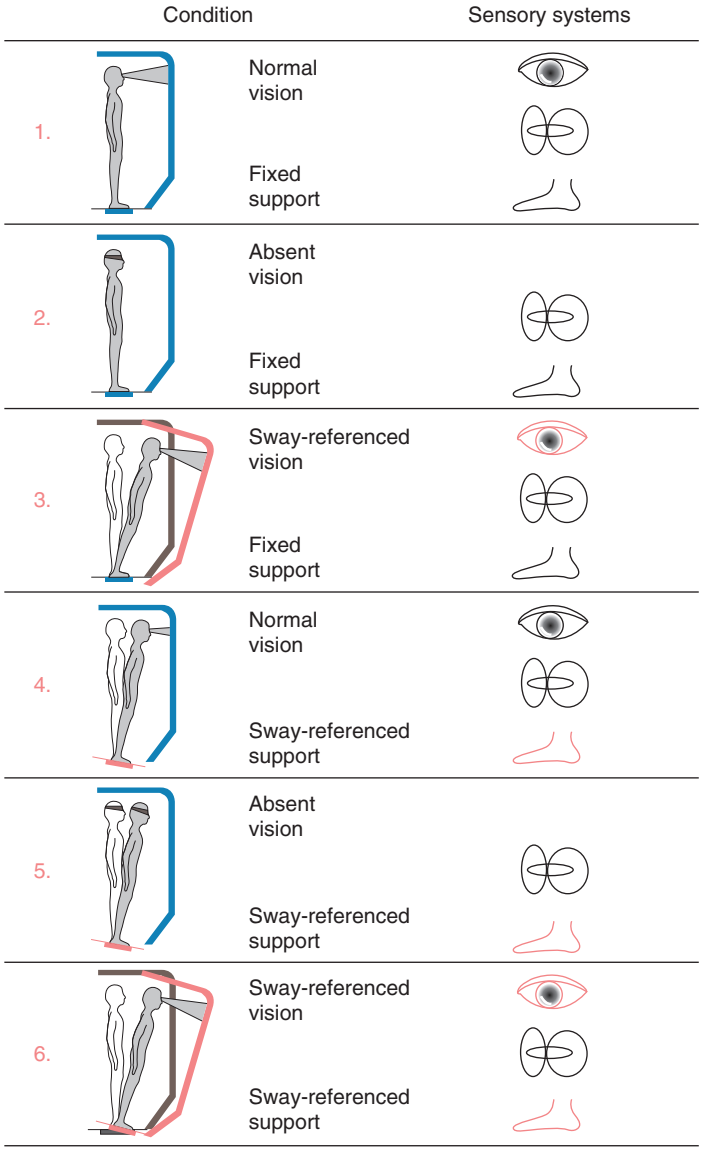

Figure 4 The six sensory organization test conditions. Anatomical symbols with each condition (eyes for vision, three overlapping circles for vestibular, and foot for lower extremity somatosensory) illustrate the senses available and appropriate for balance control under each condition.

the United States, and the reason for 3.2% of new visits to a primary care physician. Imbalance is the fourth most common neurological complaint of geriatric patients, and is the most common complaint of persons aged 85 years and older. Patients with balance disorders report a wide variety of symptoms, including inappropriate senses of self- and external environment motion (dizziness), light-headedness, postural instability, and jumpy and blurry visual images. In the elderly, the prevalence of chronic imbalance (present for greater than 1 month) is estimated to comprise between 13% and 38% of the total balance disorder population. Because good balance is critical to many daily life activities, chronic imbalance is a major cause of disability and loss of functional independence, and a major factor in increased risk for falls and fall injuries.

Historical Perspective

The traditional clinical approach to a balance disorder was to assume that a patient's problems were caused by a discrete pathology. The medical workup, including selection of diagnostic tests, was therefore structured to confirm and/or rule out the presence of the presumed localized disease. While the localized disease approach continues to be productive with many acute patients, it has proved ineffective in a majority of patients with chronic balance problems.

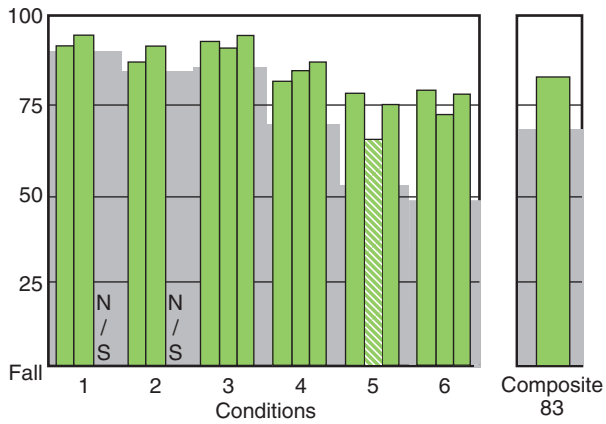

Figure 5 Sensory organization test result for a normal person. The gray background area outlines the equilibrium score falling below the range of age-matched normal persons for each sensory condition. The composite score is a weighted average of scores over the six sensory conditions. Green bars indicate individual trial scores within the normal range; N/S indicates a trial not performed.

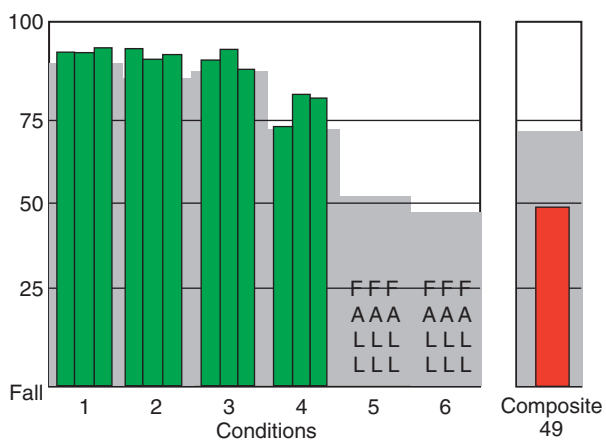

Figure 6 Sensory organization test result for a patient with bilateral loss of vestibular function. In conditions 5 and 6, 'FALL' indicates a trial during which the patient lost balance. The red bar indicates a score falling below the range of age-matched normal persons.

In the typical chronic patient, a localizing diagnosis is either not established or, when established, includes multiple pathologies that alone cannot delineate the most effective treatment approach. When a discrete disease processes is identified, traditional pharmacological and surgical treatments can sometimes stabilize the disease progression. In most chronic cases, however, medical treatments cannot cure underlying diseases or resolve associated symptoms and stability problems.

Impairment Reduction

Clinical leaders now advocate that a chronic balance disorder is effectively managed as a 'syndrome' comprising the accumulated effects of impairments within multiple systems, impairments being the functional manifestations of pathology. When a balance disorder is managed as a syndrome, impairment reduction treatments target those impairments with the biggest impact on patient function. Clinical evidence has demonstrated that rehabilitation exercise is one of the most effective impairment reduction approaches for maximizing function in patients with balance disorders.

Diagnosis

Pathology The search for pathology is always necessary and must cover a wide range of body systems. If discrete pathology is identified, the patient will frequently present with gaze and postural stability impairments that correlate poorly with the pathology. This is because impairments are influenced not only by pathology but also by interactions among system components and the adaptive responses of the brain. The pathology diagnosis only partially documents the patient's gaze and postural stabilization problems, and therefore provides only a portion of the information needed to select the most effective combination of medical and rehabilitation treatments.

Impairment An understanding of the patient's gaze and posture control impairments, the functional manifestations of pathology, provides information that is essential to treatment planning and outcome documentation, especially when an impairment reduction approach is the best treatment option. Research has provided several new, objective assessment tools for isolating and quantifying sensory, motor, and central adaptive impairments of posture and gaze control:

1. *Computerized dynamic posturography.* Computerized dynamic posturography (CDP) is a series of computerized protocols designed to isolate and quantify sensory, motor, and central adaptive impairments affecting postural stability. The sensory organization test component of CDP has been described in Figure 4. The motor control test (MCT) assesses the effectiveness of automatic postural movements. The MCT exposes the freely standing patient to sequences of unexpected, brief forward and backward displacements of the support surface, thereby displacing the patient slightly off balance and eliciting automatic postural movement responses. MCT measurements include the latency and the amplitude of movement response in each leg, and the coordination of responses between the two legs. The adaptation test (ADT) assesses ability to suppress the influence of inappropriate automatic response components. The patient is exposed to unexpected toes-up and

toes-down rotations of the support surface that elicit inappropriate automatic postural responses. ADT measures the amount of sway activity required to recover from each perturbation.

2. *Dynamic visual acuity.* Dynamic visual acuity (DVA) is a computerized protocol designed to isolate and quantify VOR impairments affecting gaze stability when the patient moves the head and views a fixed visual object. In this test, the patient identifies optotypes displayed on a computer monitor while the computer records the patient's back and forth (as if saying 'No') or up and down (as if saying 'Yes') head rotations. When the head is moving at or above a prescribed minimum velocity, an optotype is displayed for a limited duration. The test is conducted in two formats – the first with the head velocity fixed and the optotype size progressively decreased, the second with the optotype size fixed and the head velocity progressively increased. In the first format, the smallest identifiable optotype size is compared to the smallest identifiable size with the head fixed. The difference between static and dynamic optotype size is termed the 'DVA loss.' In the second format, the maximum head velocity with the optotype correctly identified is reported.

Applications of Impairment Information

Disorientation and unsteadiness are characterized by varying combinations of impairments, including (1) lack of one or more sensory inputs critical to balance, (2) misinterpretation of one or more conflicting sensory inputs, (3) delayed, weak, or poorly coordinated movement responses, (4) inappropriate movement response patterns relative to the task conditions, (5) visual acuity loss at higher head velocities, and (6) large discrepancy between head-fixed and head-moving visual acuity. The following patient examples illustrate these impairments.

Uncompensated vestibular deficit An adult patient suffered a partial vestibular loss due to vestibular neuritis. The initial symptoms of dizziness and spinning resolved within a few weeks as expected, but the patient continued to report vague feelings of being 'off balance.' The patient's SOT results are shown in Figure 7. In contrast to the bilateral vestibular loss in Figure 6, this patient maintained balance under condition 5, suggesting that the vestibular input was sufficient to maintain stability. However, the patient was unstable in the presence of the sway-referenced visual surround (condition 6), indicating that the patient failed to suppress the conflicting visual input. Failure to suppress vision conflicts helped explain why symptoms were vague and difficult to describe. Since the vestibular system was medically stable, an impairment reduction approach was selected using rehabilitation exercises to enhance visual adaptation. The patient reported a resolution of symptoms within 4 weeks, and follow-up SOT results were normal.

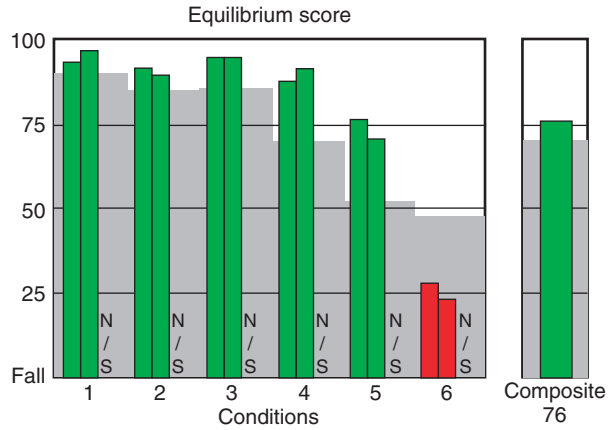

Figure 7 Sensory organization test result for a patient with an uncompensated vestibular deficit; N/S indicates a trial not performed. The red bar indicates a score falling below the range of age-matched normal persons.

Intermittent imbalance of unknown origin A construction worker self-referred, complaining of a 6-month history of intermittent dizziness, mild unsteadiness, and visual blurring. He was concerned about the safety of performing as a construction worker. Medical history and traditional site of lesion testing provided no useful information. A traditional medical workup failed to identify vestibular and central nervous system pathology. The patient's DVA results are shown in Figure 8. The patient was able to maintain normal visual acuity at velocities up to 120 $^{\circ}s^{-1}$ while moving to the left. While moving to the right, acuity was lost above 80 $^{\circ}s^{-1}$, a level of performance judged insufficient to perform complex tasks. CDP test results, by contrast, were normal. Rehabilitation treatments therefore focused on improving the patient's gaze stability and visual acuity during active and passive head movement tasks, with emphasis placed on movements to the right. Posttreatment DVA results were within normal limits. The patient reported resolution of symptoms and returned to work.

Controlled-outcome studies The effectiveness of impairment reduction approaches, such as those discussed earlier, has been demonstrated in a number of controlled clinical outcome studies. The first study included 48 chronic balance disorder patients whose standard diagnostic workup was noncontributory.

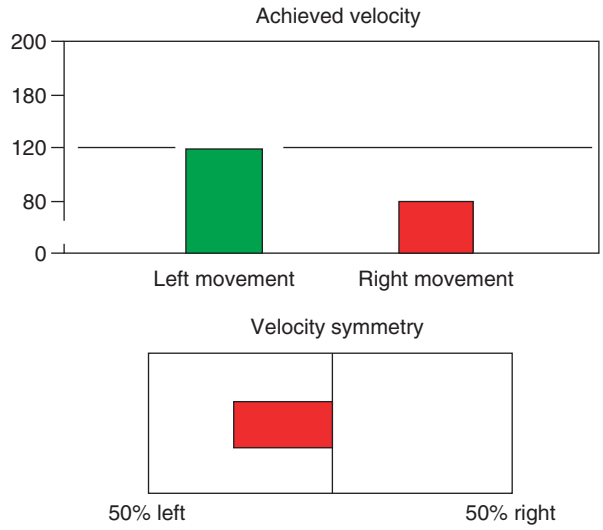

Figure 8 Gaze stabilization test result for a patient with intermittent imbalance of unknown origin. The upper graph shows the maximum achieved head velocities (degrees per second) for left and right movement directions. The lower graph shows the percentage asymmetry between the left and right achieved velocities. The green bar indicates an achieved velocity consistent with dynamic visual acuity requirements for high-performance activity. The red bars indicate achieved velocity and velocity symmetry scores below those required for high-performance activity.

The traditional discrete pathology approach would be 'watch and wait.' One-half of the patients were randomly placed in a control group and 'watch and wait' continued. The other half were placed in an 'impairment reduction' group that received vestibular rehabilitation individualized to reduce the impairments identified by the patients' CDP findings. The impairment reduction group showed significant recovery of function, while the 'watch and wait' group showed no significant changes. In another study, the impairment reduction strategy was shown to decrease the incidence of falls in elderly adults with chronic unsteadiness. A cost-effectiveness study analyzed diagnostic technologies commonly used in working up dizzy patients to identify those having the greatest impact on outcome relative to cost. This study showed that tests combining impairment information with pathological information provided by traditional diagnostic tests designed to rule out and/or verify the existence (localize) pathology provided the highest clinical value.

See also: Balance and Posture Control: Human; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Gravitational Effects on Brain and Behavior; Motion Sickness; Nystagmus; Pursuit Eye Movements; Saccadic Eye Movements; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vestibular Influences on Cognition; Vestibulo-Autonomic Responses; Vestibulo-Ocular Reflex; Vestibulospinal System and Eye–Head/Neck Movement; Visual–Vestibular Interactions.

Further Reading

Baloh RW and Halmagyi GM (eds.) (1996) *Disorders of the Vestibular System*. New York: Oxford University Press.

Black F, Angel C, Pesznecker SC, et al. (2000) Outcome analysis of individualized vestibular rehabilitation protocols. *American Journal of Otology* 21: 543–551.

Cenciarini M and Peterka RJ (2006) Stimulus-dependent changes in the vestibular contribution to human postural control. *Journal of Neurophysiology* 95: 2733–2750.

Chong R and Franklin M (2001) Initial evidence for the mixing and soft assembly of the ankle, suspensory, and hip muscle patters. *Experimental Brain Research* 136: 250–255.

Demer JL and Amjadi F (1993) Dynamic visual acuity of normal subjects during vertical optotype and head motion. *Investigative Ophthalmology and Vision Science* 34: 1894–1906.

Fife T and Baloh R (1993) Disequilibrium of unknown cause in older people. *Annals of Neurology* 34: 694–702.

Goebel J (2000) *Practical Management of the Dizzy Patient*. Philadelphia, PA: Lippincott Williams and Wilkins.

Herdman S (1999) *Vestibular Rehabilitation*. Philadelphia, PA: FA Davis.

Horak FB and Nashner LM (1986) Central programming of postural movements. Adaptation to altered support-surface configurations. *Journal of Neurophysiology* 55: 1369–1381.

Horak FB (2006) Postural orientation and equilibrium: What do we need to know about neural control of balance to prevent falls? *Age and Aging* 35(supplement 2): ii7–ii11.

Jacobson GP, Newman CW, and Kartush JM (1993) *Handbook of Balance Function Testing*. Saint Louis, MO: Mosby Year Book.

Runge CF, Shupert CL, Horak FB, et al. (1999) Ankle and hip strategies defined by joint torques. *Gait and Posture* 10: 161–170.

Tian J-R, Shubayev I, and Demer JL (2001) Dynamic visual acuity during passive and self-generated transient head rotation in normal and unilaterally vestibulopathic humans. *Experimental Brain Research* 242: 501–507.

Tinetti M, Williams C, and Gill TM (2000) Dizziness among older adults: A possible geriatric syndrome. *Annals of Internal Medicine* 132(3): 337–344.

Relevant Websites

http://www.BalanceandMobility.com – A medical information resource for balance and mobility disorders.
http://www.onbalance.com – Neurocom International, Inc.
http://www.vestibular.org – Vestibular Disorders Association.

Balance and Posture Control: Human

S M Slobounov and K M Newell, Pennsylvania State University, University Park, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Posture as a Fundamental Human Activity

Introduction

Human upright posture, traditionally defined as body segments' configuration at any given time, has been extensively studied but is still a poorly understood phenomenon. There are several reasons why existing models of posture (i.e., genetic, hierarchical, etc.) are unable to explain the enormous complexity of the mechanisms of human postural control. One reason is that there is still a strong influence of the dichotomy, similar to that of classical Cartesian dualism: "considering human posture as the genetically determined reference posture or integration of all stretch reflexes of the body in order to maintain equilibrium" versus "considering voluntary goal-directed movement as the activity of higher centers superimposed on the bodily mechanisms." There is still a debate in the literature as to whether a single central control process is responsible for both movement and its associated postures or whether there is a dual coordinated control system for voluntary movement and posture.

Human posture is a product of an extremely complex dynamical system with many degrees of freedom, which, like any other fundamental activity, undergoes dramatic changes throughout the life span. As such, the classical question of how the performer constrains the many degrees of freedom to preserve and coordinate the configuration of various body segments while performing a goal-oriented movement is still unanswered. In other words, the nature of the postural dynamic is extremely complex, since posture includes not only antigravity control (e.g., under static conditions, the center of gravity should remain within the support surface) but also the interface between perception and action via fine postural adjustments associated with suprapostural activity, stability control during locomotion, and the ability to assess the body's spatial–temporal orientation in the surrounding environment. Accordingly, a primary objective of the performer in upright stance is to explore the dynamical properties of the perceptual motor workspace and to avoid the actions that exceed its operational limitations.

Sources of Information for Postural Control

Considering posture as a complex dynamical assemblage, it is reasonable to assume that all sensory systems can contribute information relevant to the control of posture. It was hypothesized that processes within the postural control system involve organizing, integrating, and acting upon redundant visual, vestibular, and surface somatosensory inputs to provide orientation information. The contribution of sensory information to the maintenance of stable posture has been characterized as a hierarchical process in which the congruence of both support surface somatosensory and visual inputs is compared to an inertial–gravitational reference, established by vestibular reference inputs. Thus, balance is a multidimensional ability in which a range of factors such as eye–motor coordination, kinesthetic response, ampular sensitivity, and vertical semicircular canal function are crucial.

In cases of intersensory inconsistencies, inputs not congruent with vestibular reference inputs are suppressed, and greater perceptual significance is attached to those inputs in accord with vestibular sources. Overall, human upright posture is a product of a complex dynamical system that relies on integrating input from multimodal sensory sources, including (1) the visual system, (2) the vestibular system, and (3) a great number of sensory receptors embedded in the muscles, tendons, joint capsules, and skin that we commonly refer as proprioceptive systems.

Neural Basis for Multisensory Sources of Balance

A neural basis for the multisensory influences on postural motion and the whole body orientation has been proposed. Specifically, convergence of vision, audition, somatosensation, and proprioception at the cellular level in the vestibular nuclei has been shown. Also, cells have been identified in the vestibular nuclei where whole discharge rates were influenced by both body motion, as signaled by vestibular end organs, and visual motion. Recent evidence suggests that the cortex, in some way, plays a role in processing multisensory information and modifying human postural control. Moreover, it has been suggested that vestibular and somatosensory input may be integrated within a distributed cortical network, including the temporal–parietal cortex, supplementary motor area, and

prefrontal cortex, in order to process input related ego-motion and counteract a loss of balance. A more recent study with young controls has also identified functional magnetic resonance imaging (fMRI) correlates of self-perceived postural instability. Specifically, significant activation of several brain areas, including the parietal cortex, anterior cingulate, and cerebellum, was observed when young subjects visually recognized the gravitational vertical via a computer-animated model of stable versus unstable postures. Indeed, accumulated evidence from recent studies investigating the effects of human cerebral lesions on posture suggests that perception of the visual vertical involves the insula, thalamus, and superior parietal cortex. It was also shown that lesions of the temporal–parietal junction (a region of multimodal sensory integration) lead to poor equilibrium control on an unstable support.

Virtual Time to Contact

Recently, the notion of virtual time to contact (VTC), which specifies the spatial–temporal proximity of the center of pressure (COP) to the stability boundary, as an informational property in the regulation of posture was proposed. This new approach to understanding the nature of posture regulation places the emphasis on information for control not on departures from a stability point within the equilibrium region of the potential base of support, as in inverted pendulum models of posture, but rather on the temporal safety margin, as specified by the virtual time to collision with the stability boundary. A significant consequence of this approach is that the control variable for posture is defined over the organism–environment–task interaction, rather than as simply a product of the organism.

The newly created variable was called virtual time to contact, which is the instantaneous time to the functional stability boundary defined on the dynamics of each point in the time series. The word 'virtual' was used because the individual does not want to make contact with the stability boundary. Thus, VTC is an estimate of the time to the boundary should it occur, which would happen only in the case of a loss of stability, as in a fall. In this approach, a time series of the virtual time to contact can be determined that is based on the dynamics of the time to contact relative to the boundary, rather the relative position of the COP to the stability boundary. In a series of recent studies, VTC has been shown to be a more sensitive index of postural stability than other traditional COP-based measures in aged and traumatic brain-injured subjects. Clearly, further research is needed to examine the neurocognitive and physiological bases of VTC in control of human posture.

Upright Posture Control

Quiet Stance Domain

In the act of standing still, the body experiences small variations in its position. It oscillates in anterior–posterior (AP) and medial–lateral (ML) directions. This oscillation is quantified by different methods such as image recording, where whole body movement is recorded along the time dimension, or by the use of the force platform. By using a force platform one can record the time to time position of the COP. COP movements are commonly used to assess body oscillation during stance. Normally during stance the COP migrates approximately 0.4 cm in the AP direction and 0.18 cm in the ML direction, while center of mass (COM) displacements are somewhat smaller. The differences in migration of COP and COM in the AP direction have been associated with the generation of torques at the ankle joints, while ML displacements have been associated with activity of hip muscles. Several models have been suggested to explain and describe COP migration in quiet stance. Three of these models are now briefly described.

Models

Upright human posture has been frequently modeled as an inverted pendulum. Modeling the human body as an inverted pendulum is based on the assumption that joint motion occurs only at the ankle joints, body sway is very small, and the feet do not move. Since the stability of the system requires that the COM falls within the base of support, it is believed that the COM is the controlled variable of the unstable body system. COM is regulated through a continuous movement of the COP and is stabilized around a fixed reference point. In a single pendulum model, the difference between COM and COP is proportional to the acceleration of the body COM.

A two-process random-walk model was suggested in which the maintenance of vertical posture could be viewed as part of a stochastic process. This model assumed the movement of the COP as a correlated random walk and determined at which time intervals COP displacements are either positively or negatively correlated. It was found that for short time intervals (less than 1 s) COP displacements were positively correlated, that is, displacements of the COP in one particular direction were followed by displacements in the same direction. At longer timescales, negative correlations were found. Overall, it was suggested that these COP characteristics correspond to an open loop control at short timescales and a closed loop control at longer timescales.

A more recent model for the control of posture has been formulated and a new method of decomposing COP trajectories into two components, termed 'rambling' and 'trembling,' has been introduced. The decomposition first identified instant equilibrium points (IEPs). An IEP is the position of the COP at the instance when the horizontal forces are zero. At this moment the projection of the COM onto the base of support coincides with the COP position. The individual IEPs, connected through a spline fitting function, form the rambling trajectory, while the difference between the rambling and COP trajectories is called the trembling trajectory. The authors of this model suggest that the rambling trajectory describes the motion of a moving reference point with respect to which the body's equilibrium is instantly maintained, while the trembling trajectory describes body oscillation around the reference point trajectory.

Preprogrammed Reactions

A considerable body of research in human posture has focused on neural organization or muscular contraction synergy associated with postural control. Theoretical insights have suggested that a pattern of organization is preprogrammed by executive commands according to the particular motor task in order to constrain the large number of functionally related muscle synergy. Following the imposition of brief surface displacement, the automatic postural adjustments are recorded and characterized by relatively fixed temporal and structural patterns of ankle and thigh muscle activity. Preprogrammed reactions consist of a combination of muscle activation patterns specific for a given perturbation and act at a time delay of about 50–100 ms. They are different from simple reflexes in the sense that they depend on the instruction to the subject and that their magnitude can be independent of the length changes in the muscle. A vast literature exists on the regulation of posture under external perturbations and the modulation of preprogrammed reactions.

Anticipatory Posture Adjustments

It should be noted that Hess first pointed out that a proactive (i.e., anticipatory) postural stabilization is required for goal-oriented volitional movement producing an internal disturbance of equilibrium. It was shown that the ensemble of the sequential postural adjustments obeys certain rules (i.e., flexible postural synergy) that are modifiable in order to optimize the balance. The preparatory nature of postural responses is now considered to be centrally programmed and independent of any feedback from the moving limb. These preparatory (i.e., anticipatory) postural adjustments associated with volitional movement were intensively studied in relation to dynamic asymmetry and muscle forces generating movement that can perturb postural equilibrium. The disturbing effects of voluntary movement seem to be anticipated by the central nervous system (CNS). This anticipation is clearly demonstrated in studies revealing muscle activation about 50–150 ms prior to the movement initiation when no reflexes can be triggered (i.e., absence of external stimulus). Anticipatory postural adjustments have been described considering different movements performed by arm, leg, trunk, and head, including those for forearm loading and unloading tasks. It has been suggested that anticipatory posture adjustments (APAs) are generated by the CNS in a feed-forward fashion in an attempt to predict the upcoming postural perturbation and to generate approximate corrections.

Cortical Control of Human Postural Responses

The role of the cerebral cortex, in general, and the contribution of higher cortical functions in human postural control, in particular, have been less studied and are still controversial. Historically, the neuromuscular control of 'automatic' postural responses was thought to arise from the brain stem and spinal circuit with limited consideration for the role of the cerebral cortex. The original proposition that the whole righting apparatus is arranged subcortically in the brain stem, and in this way is made independent of direct voluntary influences, has persisted with time. The idea that postural responses are regulated subcortically was based upon the notion that postural responses are triggered automatically, without any voluntary intent, and therefore are initiated more quickly and with less variability than with cued voluntary movement. However, the onset of postural responses occurs at longer latencies than those of stretch reflexes, indirectly suggesting that postural responses exhibit greater potential for modulations and modifications by neural centers residing hierarchically higher along the neural axis. Moreover, unlike stretch reflexes, postural responses involve synergistic activation of muscle throughout the whole body, and these are also more context specific, flexible, and adaptable than spinal proprioceptive reflexes.

Behavioral Studies

There is considerable behavioral evidence suggesting the contribution of the cerebral cortex and higher cognitive functions in the control of upright posture. Indeed, postural response has been shown to be

modified by various cognitive-motor processes 'represented' in the cerebral cortex, including those involved in anticipatory postural adjustments. Postural responses are modified by (1) changes in cognitive load and attention when performing dual postural tasks, (2) changes in a subject's intentions to respond with a specific strategy, (3) learning and modification of postural responses with prior experience, and (4) changes in initial conditions. Considering that attention, mental calculation, and memory properties represent high-order cognitive functions, controlled by the cerebral cortex, it is reasonable to suggest direct cortical involvement in control of postural equilibrium.

Brain Imaging Studies

There are several lines of research that have directly demonstrated the involvement of the cerebral cortex in human postural control. Variation of slow negative potentials in the primary motor cortex preceding the onset of postural adjustment was observed in previous electroencephalogram (EEG) research. A more recent EEG study has also documented that the initiation of self-paced postural movement is preceded by slow negative direct current (DC) shift, similar to movement-related cortical potentials (MRCPs) accompanying voluntary goal-oriented movement. Also, a brief burst of gamma (40 Hz) EEG activity preceded the initiation of compensatory postural movement when balance was in danger. The spatial distribution of EEG patterns in postural actions approximated that previously observed during the postural perceptual tasks.

Changes in cortical excitability, revealed by slow negative DC potentials just prior to anticipated postural perturbations, have also been reported recently. Similar to DC potentials observed 1–2 s prior to a voluntary self-initiated postural movement, EEG readiness potentials with significant negative variation only in the cue condition were observed. The cue-related differences in the subjects' readiness potentials were highly correlated with the cue-related improved postural stability. Overall, this line of research supports the hypothesis that the cerebral cortex may play a role in optimization of postural responses to external perturbations thorough changes in the anticipatory central set. Therefore, the postural movement once considered to be automatic might be susceptible to voluntary cortical control.

There are several recent EEG studies examining the role of the cerebral cortex in triggered postural responses after a perturbation. These studies have shown that EEG potentials following postural perturbations (known as perturbation-evoke responses, PERs) become altered with changes in the central set, such as with changes in the predictability of a perturbation or with a secondary motor task. The PERs are thought to represent cortical processing of sensory input related to the balance disturbance that arises as an error- or conflict-related signal. Specifically, the first negative peak between 100 and 150 ms (N1) may represent sensory disturbance, and late PERs may be linked to the sensorimotor processing of balance correction, similar to the error-related negativity responses observed for incorrect responses in decision-making paradigms. These studies clearly demonstrate that the postural set influences the characteristics of the cortical contribution to the control of balance after perturbation.

Evidence suggests that the cerebellar–cortical loop is responsible for adapting postural responses based on prior experience and the basal ganglia–cortical loop is responsible for preselecting and optimizing postural responses based on current context. Thus, the cerebral cortex most likely influences longer latency postural responses directly via corticospinal loops and shorter latency postural responses indirectly via communication with the brain stem centers that harbor the synergies for postural responses, thereby providing both speed and flexibility of preselecting and modifying environmentally appropriate responses to a loss of balance.

Finally, the neural substrates for maintaining standing postures in human have been investigated recently using the mobile gantry positron emission tomography (PET) system. Interestingly, compared with the supine posture, standing upright with the feet together activated the cerebellar anterior lobe and the right visual cortex, while standing in tandem was accompanied by activation within the visual association cortex and cerebellar vermis. This finding is consistent with previous research suggesting that the human cerebellar vermis may coordinate the timing in keeping the center of gravity within the stability limits and therefore may provide control of postural scaling and central set in stance.

Concluding Comments

This article has presented a brief review of previous work with special emphasis on current research regarding the role of the cerebral cortex in human postural control. Considerable behavioral evidence clearly suggests the contribution of the cerebral cortex and higher cognitive functions in the control of upright posture. This was documented indirectly via modification of postural responses induced by various high-order cognitive functions and/or learning

that are thought to be represented in the cerebral cortex. There is also a growing body of more recent research that directly demonstrates the involvement of the cerebral cortex in postural movement, similar to many other voluntary movements under study. This was primarily documented via EEG patterns preceding self-initiated postural movement (MRCPs) and following postural perturbation (PERs). It should be noted, however, that this newly evolving understanding of the role of cortical processes in the control of posture in humans is still in its infancy. Future research implementing current advances in cellular recording and brain imaging techniques may be warranted to further explore the differential contribution and functional connectivity of various brain structures in the control of human posture.

See also: Balance and Posture Control; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Neuroimaging; Vestibular System; Vestibulo-Autonomic Responses; Vestibulospinal System and Eye–Head/Neck Movement; Visual–Vestibular Interactions.

Further Reading

Adkin AL, Quant S, Maki BE, and McIlroy WE (2006) Cortical responses associated with predictable and unpredictable compensatory balance reactions. *Experimental Brain Research* 172: 85–93.

Bernstein N (1967) *The Coordination and Regulation of Movement*. New York: Pergamon.

Calvert G, Spence C, and Stein BE (2004) *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.

Collins JJ and DeLuca CJ (1993) Open-loop and closed-loop control of posture: A random-walk analysis of center-of-pressure trajectories. *Experimental Brain Research* 95: 308–318.

DiZio PA and Lackner JR (1986) Perceived orientation, motion, and configuration of the body during viewing of an off-vertical, rotating surface. *Perception and Psychophysics* 39: 39–46.

Hess WR (1943) Teleokinetischer und ereismatisches Kraftesystem in der Biomotori. *Helver Physiologisches und Pharmacologisches Acta* C62–C52.

Horak FB, Diener HC, and Nashner LM (1989) Influence of central set on human postural responses. *Journal of Neurophysiology* 62: 841–853.

Jacobs JV, Fujiwara K, Tomita H, Furune N, Kunita K and Horak FB (2008) Changes in activity at the cerebral cortex associate with the optimization of responses to external postural perturbations when given prior warning. *Gait & Posture Supplement*.

Latash ML (1993) *Control of Human Movement*. Champaign, IL: Human Kinetics.

Magnus R (1926) Physiology of posture. *Lancet* 11: 531–585.

Mergner T, Maurer C, and Peterka RC (2003) A multisensory posture control model of human upright stance. *Progress in Brain Research* 142: 189–201.

Mori S, Stuart D, and Wiesendanger M (eds.) (1994) *Progress in Brain Research* 143. London: Elsevier.

Newell K and Corcos D (1993) *Variability and Motor Control*. Champaign, IL: Human Kinetic Publishers.

Ouchi Y, Okada H, Youshikawa E, Nobezava S, and Futatsubashi M (1999) Brain activation during maintenance of standing postures in human. *Brain* 122(2): 329–338.

Rothwell J (1994) *Control of Human Voluntary Movement,* 2nd edn. London: Chapman and Hill.

Sherrington CS (1910) Flexion-reflex of the limb, crossed extension-reflex, and reflex stepping and standing. *Journal of Physiology* 40: 28–121.

Slobounov S, Slobounov E, and Newell K (1997) Virtual time-to-collision and human postural control. *Journal of Motor Behavior* 29: 263–281.

Winter DA (1990) *Biomechanics and Motor Control of Human Movement,* 2nd edn. New York: Wiley.

Zatsiorsky VM and Duarte M (1999) Instant equilibrium point and its migration in standing tasks: Rambling and trembling components of the stabilogram. *Motor Control* 3: 28–38.

Balint Syndrome

N L Denburg and D Tranel, University of Iowa, Iowa City, IA, USA

© 2009 Elsevier Ltd. All rights reserved.

'Balint syndrome' is an acquired disturbance of the ability to perceive the visual field as a whole, and traditionally, is considered to include three components: (1) 'visual disorientation' (also known as 'simultanagnosia'), (2) 'ocular apraxia' (also known as 'psychic gaze paralysis'), and (3) 'optic ataxia.' The key constituent in the syndrome, however, is visual disorientation, and there is considerable variability in the emphasis that is placed on the other components. There is recent evidence to suggest that body schema disturbances may frequently accompany Balint syndrome (i.e, the so-called Gerstmann syndrome involving right–left disorientation, acalculia, and finger agnosia).

Visual disorientation (simultanagnosia) can be conceptualized as an inability to attend to more than a very limited sector of the visual field at any given moment. The phenomenon is subjective. Patients report that they can see clearly in only a small part of the field, the rest being 'out of focus' and in a sort of 'fog.' The sector of clear vision, moreover, is unstable, and may shift without warning in any direction, so that patients experience a literal 'jumping about' of their visual perception. Such patients are incapable of constructing a spatially coherent visual field, and they cannot follow trajectories of stimuli or place stimuli in their proper locations in space. Perception of motion is often impaired, so that patients fail to notice objects that have moved about in their visual field or fail to recognize the meaning of movements they have otherwise perceived correctly. For example, patients may fail to recognize a familiar gait or stride or fail to understand pantomime. Isolated disturbances of motion detection, however, are quite rare (one of the few such cases was described by Zihl and colleagues in 1983). Patients with visual disorientation can perceive color and shape normally, so long as objects are appreciated within a clear sector of the visual field.

An example of simultanagnosia is illustrated with the picture in Figure 1. Here, a patient with Balint syndrome was presented with a simple line drawing of a wreath (item taken from the Boston Naming Test) and asked to respond to the question, "What is this?" The patient responded "bow" rather than "wreath," indicating that her sector of vision at that moment did not include the upper portion of the line drawing. Although simultanagnosia may impair object identification, as mentioned previously, the ability to identify an object is generally superior for animate objects, in which identification is frequently based on overall shape, than for inanimate objects, often identified by distinctive features.

Ocular apraxia (psychic gaze paralysis) is a deficit of visual scanning. It consists of an inability to direct gaze voluntarily toward a stimulus located in peripheral vision, so as to bring it into central vision. Thus, patients fail to direct saccades toward stimuli that have appeared in the panorama of their visual fields, or they produce saccades that are inaccurate and miss the target. Ocular apraxia is not necessary for the development of visual disorientation, although it always occurs together with visual disorientation or optic ataxia. Nyffeler and colleagues studied the natural course of oculomotor recovery in a 53-year-old woman who suffered biparietal lesions from cardiac arrest leading to hypoxia. Thirty-seven weeks after brain damage had occurred, ocular apraxia was markedly improved, as measured by an eye-tracking device, although her ability to read a clock, for example, remained impaired, indicating lasting functional deficits.

Optic ataxia is a disturbance of visually guided reaching behavior. Patients are not able to point accurately at a target, under visual guidance, e.g., they cannot point precisely to the examiner's fingertip or to items such as a cup or coin. Interestingly, pointing to targets on their own body does not pose a problem, as this can be accomplished on the basis of somatosensory information. Also, the patients have no difficulty pointing to sound sources. Optic ataxia can occur in isolation, particularly when lesions are at the border of the occipital and parietal regions or in the parietal region exclusively.

The full Balint syndrome is associated with bilateral occipitoparietal lesions (see Figure 2, a computed tomography (CT) scan of the brain which illustrates bilateral occipital involvement with more severe damage to the left parieto-occipital region), although unilateral lesions can also produce the syndrome, especially when they are on the right. Commonly, the lesions are caused by infarcts in the border zone (watershed) between the anterior and posterior cerebral artery territories. Severe and sudden hypotension is the main cause of this type of stroke, often in elderly patients who are vulnerable to systemic blood pressure drop. The Balint syndrome can also be caused by bilateral metastases in the occipitotemporal region.

Lesions causing the full Balint syndrome typically involve both the superior visual association cortices (dorsal areas 18/19) and a portion of the anteriorly adjacent parietal region (area 7 and sometimes

Figure 1 Stimuli used to demonstrate simultanagnosia.

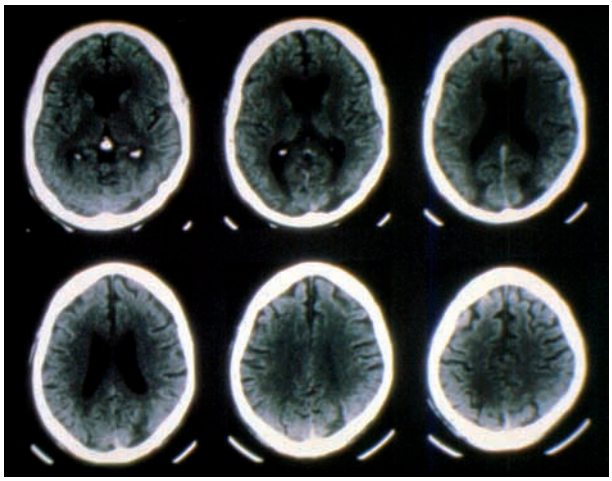

Figure 2 Illustration of the neuroanatomic findings of a patient with Balint syndrome.

area 39). When lesions are confined to the superior occipital cortices without extension into the partial region, visual disorientation is likely to occur without associated ocular apraxia or optic ataxia. The defects in motion perception that are frequent in Balint patients are probably related to damage in the lower-parietal/lateral-occipital region, in the region formed by area 39. Many patients with Balint syndrome have an impairment of stereopsis, that is, the process of recovery of depth from visual information dependent upon binocular visual interaction, although complete astereopsis is seen only in the setting of bilateral lesions.

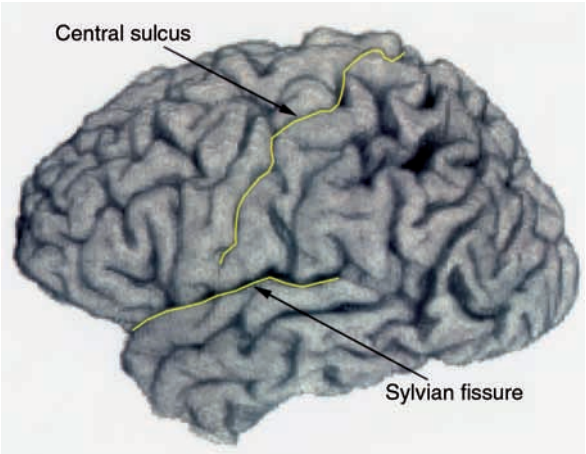

Figure 3 Illustration of the lateral view of the brain of a 50-year-old woman with probable VVAD causing Balint syndrome.

More recently, Balint syndrome has been associated with degenerative disease. Components of Balint syndrome, and particularly simultanagnosia, have been identified in patients with degeneration of the posterior cortices of the brain, referred to as visual variant of Alzheimer's disease (VVAD) or progressive visuospatial dysfunction, as well as a recent report of Balint syndrome in a patient with corticobasal ganglionic degeneration. Figure 3 illustrates the focal degeneration of the brain in a 50-year-old woman with probable VVAD.

See also: Apraxia: Sensory System; Apraxia: Disease; Attention and Eye Movements; Cross-Modal Interactions Between Vision and Touch; Eye Movement Disorders; Perception and Eye Movements; Stroke; Vision for Action and Perception; Visual Attention.

Further Reading

Damasio AR and Benton AL (1979) Impairment of hand movements under visual guidance. *Neurology* 29: 170–174.

Damasio AR, Tranel D, and Rizzo M (2000) Disorders of complex visual processing. In: Mesulam M-M (ed.) *Principles of Behavioral Neurology*, pp. 332–3728. New York: Oxford University Press.

Farah MJ (2003) Disorders of visual–spatial perception and cognition. In: Heilman KM and Valenstein E (eds.) *Clinical Neuropsychology*, 4th edn., pp. 146–160. New York: Oxford University Press.

Girotti F, Milanese C, Casazza M, Allegranza A, Corridori F, and Avanzini G (1982) Oculomotor disturbances in Balint's syndrome: Anatomical findings and electro-oculographic analysis in a case. *Cortex* 8: 603–614.

Husain M and Stein J (1988) Rezso Balint and his most celebrated case. *Archives of Neurology* 45: 89–93.

Mendez MF (2000) Corticobasal ganglionic degeneration with Balint's syndrome. *Journal of Neuropsychiatry and Clinical Neuroscience* 12: 273–275.

Mendez MF, Ghajarania M, and Perryman KM (2002) Posterior cortical atrophy: Clinical characteristics and differences

compared to Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders* 14: 33–40.

Mesulam M-M (2001) Aging, Alzheimer's disease, and dementia: Clinical and neurobiological perspectives. In: Mesulam M-M (ed.) *Principles of Behavioral and Cognitive Neurology*, pp. 439–522. New York: Oxford University Press.

Nyffeler T, Pflugshaupt T, et al. (2005) Oculomotor behaviour in simultanagnosia: A longitudinal case study. *Neuropsychologia* 43: 1591–1597.

Rafal R (2001) Balint's syndrome. In: Boller F and Grafman J (eds.) *Handbook of Neuropsychology,* vol. 4, pp. 121–141. Amsterdam: Elsevier Science.

Riddoch MJ and Humphreys GW (2004) Object identification in simultanagonosia. *Cognitive Neuropsychology* 21: 423–441.

Rizzo M (1989) Asteropsis. In: Boller F and Grafman J (eds.) *Handbook of Neuropsychology,* vol. 2, pp. 415–427. Amsterdam: Elsevier Science.

Rizzo M (1993) "Balint's syndrome" and associated visuospatial disorders. In: Hennard C (ed.) *Bailliere's International Practice and Research*, pp. 415–437. London: W.B. Saunders.

Rizzo M and Hurtig R (1987) Looking but not seeing: Attention, perception, and eye movements in simultanagnosia. *Neurology* 37: 1642–1648.

Rizzo M and Nawrot M (1993) Human visual cortex and its disorders. *Current Opinion in Ophthalmology* 4: 38–47.

Rizzo M and Vecera SP (2002) Psychoanatomical substrates of Balint's syndrome. *Journal of Neurology, Neurosurgery, and Psychiatry* 72: 162–178.

Valenza N, Murray MM, Ptak R, and Vuilleumier P (2004) The space of senses: Impaired crossmodal interactions in a patient with Balint syndrome after bilateral parietal damage. *Neuropsychologia* 42: 1737–1748.

Zihl J, Von Cramon Z, and Mai N (1983) Selective disturbances of movement vision after bilateral brain damage. *Brain* 106: 313–340.

Barrel Cortex Circuits

C C H Petersen, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

From Whisker to Cortex

While examining layer 4 of the rodent primary somatosensory cortex, Thomas Woolsey and Hendrik van der Loos discovered in 1970 that there are obvious cytoarchitectonic 'barrels' arranged identically to the whiskers on the snout (Figure 1(a)). This remarkable mapping has inspired many neuroscientists to turn their attention to the signaling pathway from whisker to cortex. Indeed, the entire sensory pathway is remarkably well delineated through stereotypical anatomical structures. The whiskers on the snout are laid out in a genetically determined pattern and the large posterior mystacial whiskers have been given a nomenclature such that all researchers can easily identify, for example, the C2 whisker (Figure 1(a)). These large posterior whiskers are actively moved at high frequencies during 'whisking' when a rodent is exploring its environment. This active movement of the whiskers is believed to enhance sensitivity and discrimination abilities of this sensory signaling pathway in the same way that humans would move their fingertips across a surface in order to gather textural information.

A whisker contact with an object during whisking or a passive deflection of a whisker evokes activity in trigeminal ganglion neurons. These sensory neurons innervate the whisker follicle and presumably express mechanogated ion channels, which have not yet been identified at a molecular level. Deflection of a whisker will induce mechanogated ionic currents, leading to action potential firing in the sensory trigeminal neurons. The sensory neurons respond to deflection of one and only one whisker, are highly selective to the direction of whisker deflection, and can be classified as rapidly adapting (RA) or slowly adapting (SA). Recently, it was proposed that different trigeminal sensory neurons might encode distinct aspects of tactile sensation, some being activated by object contact and others by whisking in free air. Quantitative analysis shows that the activity in the trigeminal ganglion neurons is extremely precise and highly reproducible, showing little trial-to-trial variability.

These action potentials in the sensory neurons accurately encoding whisker movement propagate to the brain stem, where they synaptically release glutamate onto brain stem neurons. The most important brain stem nucleus for processing of whisker information is the principal trigeminal (PrV) nucleus, which projects strongly to the ventral posterior medial (VPM) thalamus. The spinal trigeminal (SpV) complex of the brain stem also receives whisker inputs and these neurons project to the posterior medial (POM) thalamus, to the ventrolateral aspect of the ventral posterior medial (VPMvl) thalamus, and to the tectum, zona incerta, and other brain regions. Both the PrV nucleus and the SpV nuclei contain clear anatomical patterns termed 'barrelettes,' which are organized somatotopically. Data suggest that the PrV nucleus processes predominantly single whisker information, whereas the SpV nuclei process multi-whisker information.

The thalamic VPM and POM regions both receive input from the brain stem relating to whisker movement. The VPM thalamus, but not the POM thalamus, possesses somatotopically arranged 'barreloids.' Each barreloid processes information primarily related to movement of a single whisker. Within a barreloid there is a map of direction tuning such that more dorsal neurons respond better to forward and upward whisker deflection direction. The neurons in the VPM thalamus generally respond with short latency and high precision to a single whisker, in agreement with a key role of this thalamic nucleus in relaying reliable sensory information to the cortex. Much less is known about the POM thalamus, but under anesthesia it appears to be largely inhibited through inputs from the zona incerta. The physiological role of whisker information flowing through the POM region remains to be discovered.

The barreloid neurons of the VPM thalamus provide the major feed-forward excitation to the primary somatosensory barrel cortex. The principal target of innervation of the axons from a single barreloid is a barrel (Figure 1(b)). The barrels are laid out somatotopically in neocortical layer 4. The VPM neurons also project weakly to upper layer 6 of the primary somatosensory cortex. The POM neurons provide a complementary innervation of the primary somatosensory cortex, with little input to the layer 4 barrels, but with dense innervation of layer 1, layer 5A, and the layer 4 septal regions separating neighboring barrels. In contrast to the reliable encoding of sensory information at the periphery in the pattern of action potential firing of the trigeminal ganglion neurons, responses in the neocortex are highly variable, perhaps resulting from more complex context-dependent integration with spontaneous activity.

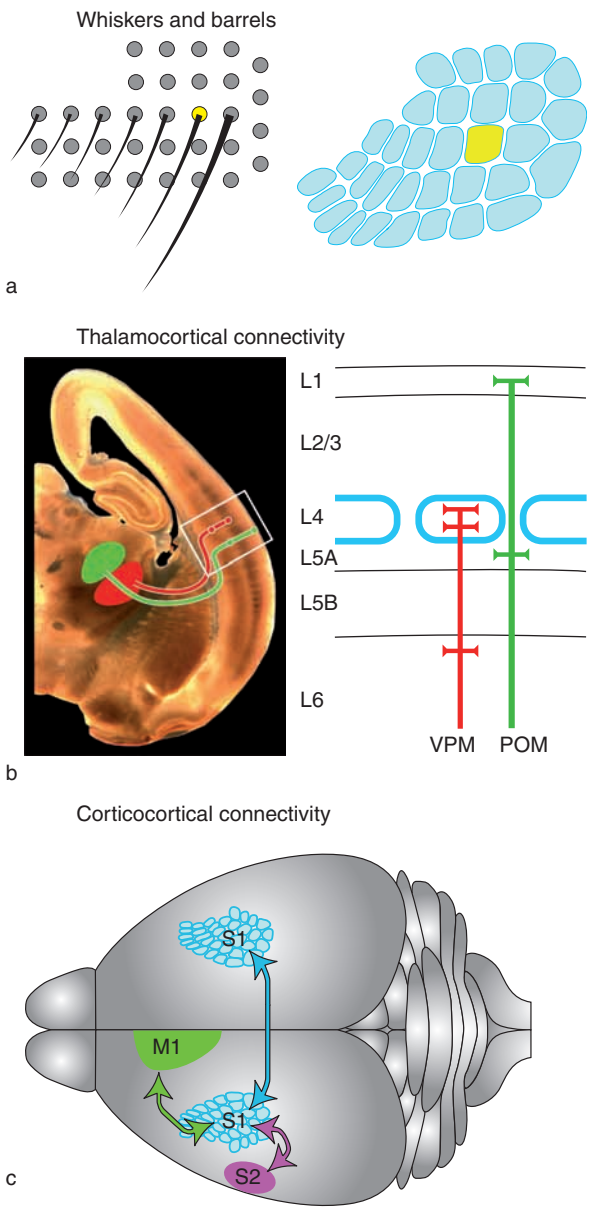

Figure 1 Long-range synaptic connections of the barrel cortex. (a) The whiskers (left) on the snout of the rodent are laid out identically to the barrels (right) in layer 4 of the somatosensory neocortex. The more posterior whiskers are longer, as indicated schematically for the C-row of whiskers. The C2 whisker is highlighted with a yellow follicle and a corresponding yellow barrel. (b) The ventral posterior medial (VPM) and the posterior medial (POM) nuclei are the two major thalamic inputs to the somatosensory neocortex. The VPM nucleus projects densely to the layer 4 (L4) barrels, with less prominent input to upper layer 6 (L6). The POM nucleus projects outside of the L4 barrels, innervating layers 1 (L1) and 5A (L5A) most densely. (c) The most prominent long-range corticocortical projections come from the contralateral primary somatosensory cortex (S1), secondary somatosensory cortex (S2), and primary motor cortex (M1).

Projections to the Primary Somatosensory Barrel Cortex

The primary somatosensory barrel cortex therefore receives major feed-forward input to the layer 4 barrels relating primarily to single whisker movement relayed by the VPM thalamic neurons. In addition to the thalamic input, which defines the primary somatosensory cortex, there are many other projections to the barrel cortex, which are likely to facilitate higher order associational and top-down processing (Figure 1(c)). The most obvious corticocortical inputs to the barrel cortex come from the surrounding dysgranular zone of the ispilateral primary somatosensory cortex, the contralateral primary somatosensory cortex (S1), the secondary somatosensory cortex (S2), and the primary motor cortex (M1).

The dysgranular zone surrounds the barrel field and is also involved in processing somatosensory information, but unfortunately little is known about response properties in this region. Anatomically it appears to provide and receive important synaptic input related to the barrel cortex. Inputs from the contralateral primary somatosensory cortex arrive via callosal axons, which target primarily this dysgranular zone and the septal regions between the barrels. Such callosal projections may signal information concerning ipsilateral multiwhisker deflection to layer 5 barrel cortex neurons.

The S2 located posterior and lateral to S1 is also extensively reciprocally synaptically connected to the barrel cortex. S2 has a more diffuse somatotopic arrangement with broader receptive fields. Little is known about the response properties of neurons in S2, how they contribute to perception, or how they interact with neurons in S1.

Input from M1 arrives primarily in layer 1 and layer 5 of barrel cortex. A surprisingly large part of M1 is involved in controlling whisker movement, although brain stem central pattern generators can evoke rhythmic whisking even in decorticate rodents. M1 is therefore more likely to be involved in fine regulation of whisker movements during tactile processing rather than in controlling each whisker cycle during whisking. The input to S1 from M1 might therefore signal fine-scale control of whisker movement, which might be important to relate to the sensory input evoked during object contacts.

In addition to these corticocortical glutamatergic inputs there are also inputs from modulatory neurotransmitter systems containing, for example,

acetylcholine, serotonin, or dopamine. The role of these neuromodulatory systems in somatosensory processing has not been clearly established.

Local Circuits in the Primary Somatosensory Barrel Cortex

A single whisker deflection will signal primarily through activity in the homologous thalamic barreloid projecting largely to a single layer 4 barrel. The layer 4 barrel neurons are therefore the first cortical neurons to become activated, but cortical processing of sensory information is distributed across the intricate synaptically connected cortical neuronal networks (Figure 2). Excitatory neurons in layer 4 can be classified as spiny stellate, star pyramidal, or pyramidal. These layer 4 neurons send their axons primarily to neurons within the same layer 4 barrel and to neurons in layer 2/3 within the same barrel column; recently, a projection to layer 5A was also described. The layer 2/3 neurons lying above a barrel receive their major input from layer 4 barrel neurons, but the layer 2 neurons lying above layer 4 septal regions receive a major input from layer 5A. In addition, all layer 2/3 neurons receive substantial intralaminar synaptic input from within layer 2/3. Since the axons of layer 2/3 pyramidal neurons can span large horizontal distances in layer 2/3, these synaptic connections are not necessarily from within the same barrel column. Layer 2/3 pyramidal neurons synapse onto layer 5 neurons, providing a major route for excitation of the infragranular layers. Excitatory layer 6 neurons are diverse and less well characterized, but projections have been reported to layer 4 and layer 5.

In addition to these excitatory neurons, there are heterogeneous populations of γ-aminobutyric acid (GABA)ergic interneurons, as found in other cortical areas. These GABAergic neurons are likely to provide a predominantly inhibitory role, preventing feed-forward explosive excitation and probably performing more subtle roles in governing excitatory neuron spike output and its precise timing.

Functional Activation of the Barrel Cortex

Whisker deflection evokes sequential action potential firing in trigeminal ganglion neurons, brain stem neurons, thalamic neurons, and cortical neurons. Sensory responses in all barrel cortex neurons consist of early depolarizing synaptic potentials, evoking action

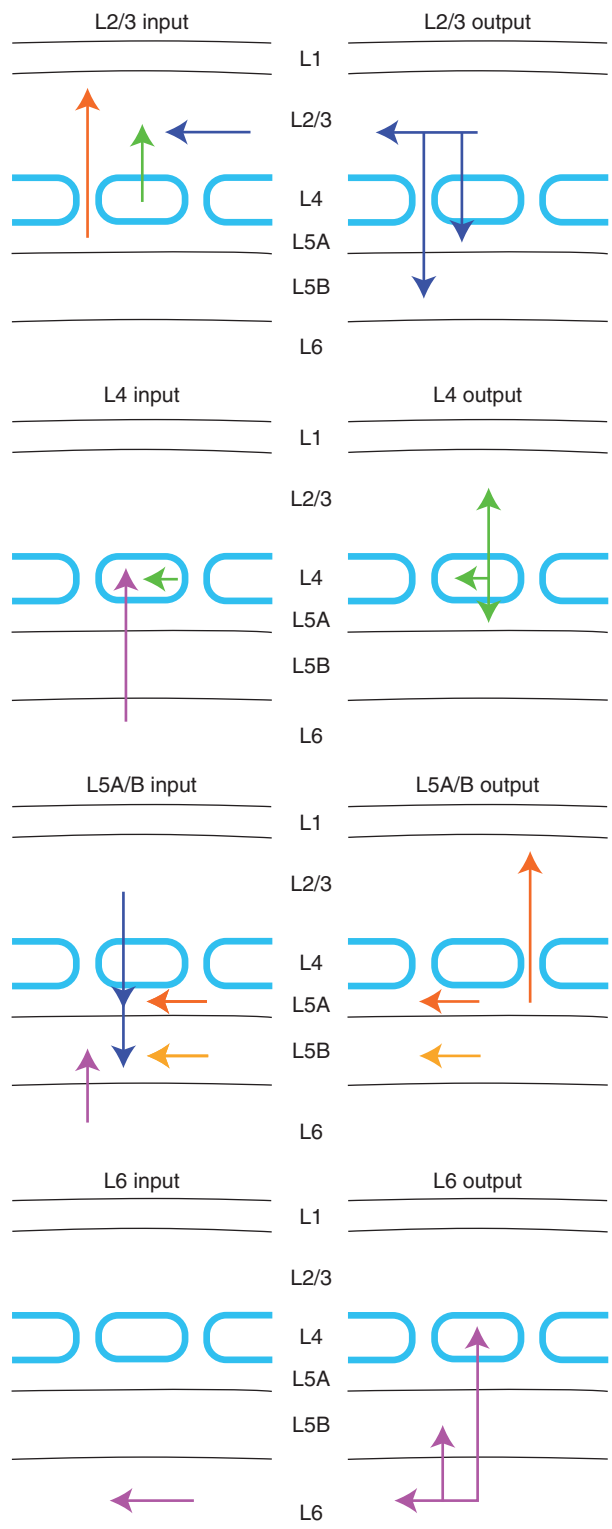

Figure 2 Local synaptic circuits of the barrel cortex. The column on the left denotes with color-coded arrows the principal local inputs to a given cortical layer (L). The column on the right shows the most prominent excitatory synaptic outputs of a given layer.

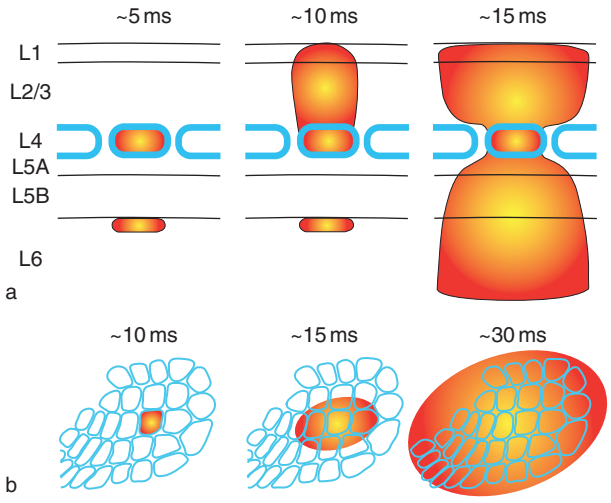

Figure 3 Functional activation of the barrel cortex. (a) Schematic drawing indicating a likely laminar sequence of excitation of the primary somatosensory barrel cortex following a brief, single whisker deflection. Thalamic input arrives predominantly in a single layer 4 (L4) barrel (cyan outlines indicate barrels). The L4 excitatory spiny stellate and pyramidal neurons depolarize, firing action potentials which propagate through their axons to excite neighboring L4 neurons and layer 2/3 (L2/3) neurons within the same barrel column. The L2/3 neurons signal across barrel columns, allowing strong multiwhisker interactions, and in addition they excite infragranular neurons. (b) Viewed across the somatotopic map of the barrel cortex, excitation evoked by deflection of a single whisker initiates in a single barrel column, and over the next tens of milliseconds the excitation spreads across the entire barrel field. Even a brief deflection of a single whisker can therefore evoke activity in a large cortical area, which may be useful for correlating the sensory information arising from the entire whisker array.

potential activity in only a subset of neurons, followed rapidly or concomitantly with GABAergic inhibition. Cortical excitation evoked by a single whisker deflection begins within the layer 4 barrel corresponding to the stimulated whisker (Figure 3). Depending on the kinetics of the stimulus parameters and perhaps the brain state, the first responses occur with a latency of ~5 ms. It is clear that the sensory pathway for tactile information to reach the neocortex is extremely rapid. In the following milliseconds sensory information reaches layer 2/3 within the same barrel column. This excitation is likely mediated by the prominent axonal projection from layer 4 to layer 2/3 described previously. Within the first ~10 ms following whisker deflection, excitation is largely confined to a single barrel column, in agreement with expectations regarding cortical columnar architecture. However, over the next milliseconds, depolarizing potentials can be recorded across large areas of the barrel cortex. This lateral spread of excitation is likely mediated by the extensive horizontal arborizations of layer 2/3 pyramidal neurons. The lateral spread of excitation is also reflected in large multiwhisker receptive fields. Excitation of infragranular neurons is more complex, since they receive some input directly from thalamocortical synapses, as well as from neurons in layer 4 and layer 2/3. Layer 5 neurons begin to depolarize ~10 ms following whisker deflection, with relatively slow time to peak perhaps reflecting the complex integration of input from multiple layers. All cortical neurons respond to deflection of many individual whiskers, with the sharpest tuning in layer 4. These broad receptive fields indicate that a key element of cortical sensory processing is therefore likely to be the integration of whisker input from across the entire whisker array and probably the integration of information from other sensory modalities and from past experiences. The growing knowledge relating to the synaptic connectivity of neurons in the barrel cortex, together with the increasingly well-defined functional activation of the barrel cortex circuits by simple whisker deflections, should allow computational studies to reconstruct the synaptic basis of simple sensory coding in anesthetized animals.

Physiological Role of the Barrel Cortex

Behavioral analysis of the whisker system has gathered increasing attention, with a focus on determining how the barrel cortex contributes to perception and motor behavior. A major obstacle to these behavioral studies is that the rodent can also obtain somatosensory information using forepaws and nose. The whiskers, however, can reach further forward and further to the side than can the paw or the nose, and current behavioral studies rely on placing the tactile stimulus out of reach to all but the long mystacial whiskers.

The work of Hutson and Masterton in 1986 defined a whisker-dependent task that required intact primary somatosensory barrel cortex and that could be performed by rodents with just a single whisker (all other whiskers being trimmed). Hutson and Masterton's gap-crossing task (Figure 4(a)) begins with the rodent on an elevated platform from which it must reach out with its whiskers to touch another remote target platform. The location of the target platform is determined by whisker touch, because this is the only way for the rodent to touch the platform and vision is typically hindered through performing the task under infrared illumination. After touching the platform with its whisker, usually many times in succession during active whisking, the animal crosses to obtain a reward. Ablation of the barrel cortex prevents the rodent from performing the gap-crossing task. The results therefore suggest

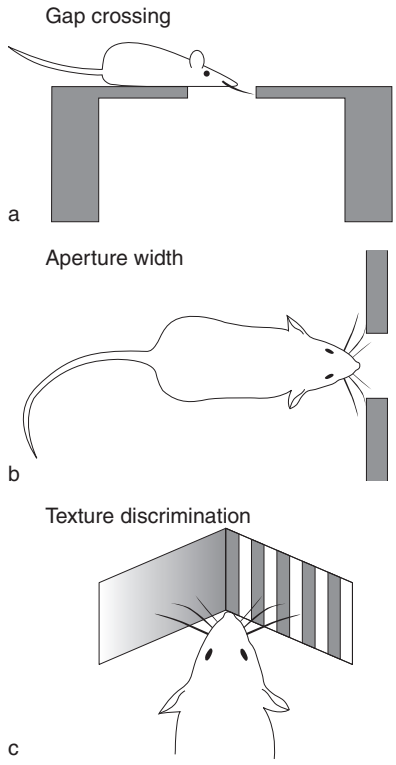

Figure 4 Whisker-dependent behaviors. (a) In the gap-crossing task, a rodent must identify a target platform, which it can palpate with its whiskers across a gap. Upon determining the presence or absence and exact location of the platform, the rodent crosses the gap to receive a reward. (b) The size of an opening is likely to be important for a rodent to determine during natural behavior. The aperture-discrimination task tests the rodent's ability to discriminate between different opening sizes by associating different reward locations with wide or narrow apertures. (c) The rodent whiskers, like human fingertips, can discriminate textures. A rat can discriminate a smooth object from a similar object with finely spaced shallow grooves.

that a single whisker alone can be used to determine object location and that the perception depends upon activity in the primary somatosensory barrel cortex.

Another interesting behavioral paradigm involves the combined processing of somatosensory whisker input from left and right whiskers. In the aperture-discrimination task (Figure 4(b)) the rodent moves its head into an opening with a variable width. The rodent learns that wider openings are associated with rewards given at a different location to rewards associated with narrower apertures. This behavioral task is attractive since it is likely to be closely related to natural behaviors, when a rodent must judge whether it can squeeze through an opening or not. From a sensory processing perspective it is also interesting, since information must be compared between left and right whisker pads, perhaps involving callosal pathways.

Texture discrimination (Figure 4(c)) can also be performed using both the short whiskers surrounding the lips and the large mystacial whiskers. Psychophysical measures of performance indicate that the discrimination abilities of the rat whiskers are similar to those of the human fingertips. Recordings of neural activity during texture discrimination and aperture discrimination are beginning to give information relating sensory processing to perception.

Perspective

The rodent barrel cortex offers the opportunity to investigate the synaptic basis of a relatively well-defined sensory system. Over the next decades we hope that increasingly quantitative data combining results from awake behaving animals, from responses to well-controlled stimuli in anesthetized preparations, and from analyses of the anatomical and functional connectivity of the synaptic circuits will further our understanding of cortical processing.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Electrical Perception and Communication; Evolution of Sensory Receptor Specializations in the Glabrous Skin; Multisensory Convergence and Integration; Neuropeptides: Sensory Systems; Seismic and Vibrational Signals in Animals; Sensory System Specializations; Somatosensory Perception; Somatosensory Cortex: Functional Architecture; Somatosensory Receptive Fields; Tactile Texture.

Further Reading

Hutson KA and Masterton RB (1986) The sensory contribution of a single vibrissa's cortical barrel. *Journal of Neurophysiology* 56: 1196–1223.

Kleinfeld D, Ahissar E, and Diamond ME (2006) Active sensation: Insights from the rodent vibrissa sensorimotor system. *Current Opinion in Neurobiology* 16: 435–444.

Petersen CCH (2003) The barrel cortex – integrating molecular, cellular and systems physiology. *Pflugers Archiv – European Journal of Physiology* 447: 126–134.

Woolsey TA and Van der Loos H (1970) The structural organisation of layer IV in the somatosensory region (SI) of the mouse cerebral cortex: The description of a cortical field composed of discrete cytoarchitectonic units. *Brain Research* 17: 205–242.

Basal Forebrain and Memory

L C P Botly, University of Toronto, Toronto, ON, Canada
M G Baxter, University of Oxford, Oxford, UK
E De Rosa, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The basal forebrain (BF) refers to an extended continuum of subcortical nuclei with projections to diverse limbic and neocortical areas implicated in various aspects of cognitive function. Damage to the BF can result in cognitive impairments. For instance, aneurysms of the anterior communicating artery leading to neuronal death in the BF result in amnesia and impairments in executive function. Furthermore, cognitive deficits in both normal aging and age-related pathological conditions have been associated with the BF. The severity of cognitive impairment observed in Alzheimer's disease is correlated with the extent of deterioration of cholinergic neurons in the BF, and a similar relationship between cognitive impairment and alterations in BF cholinergic neurons is seen in normal aging.

Given that a large proportion of BF neurons synthesize and use acetylcholine (ACh) as a neurotransmitter, cholinergic neurons have been central to most explanations of the cognitive processes attributed to the BF. Recent work on the neurochemistry of the BF has focused on identifying the specific modulatory influences BF cholinergic neurons contribute to cognition. The use of nonhuman animals has been invaluable to such investigations, and as more selective lesion methods become available for experimental studies of this region, hypotheses regarding the involvement of the BF cholinergic system in global aspects of cognitive function have been gradually revised.

Such research suggests that BF cholinergic neurons do not play a specific role in memory processing *per se*, or a general role in sustaining the function of their cortical targets. For example, performance on many behavioral tasks that are impaired after ablation of cortical neurons remains intact after removal of cholinergic input to the same cortical regions. Instead, damage limited to BF cholinergic neurons produces highly restricted impairments in particular aspects of cognition, including sensory information processing, attention, and some aspects of associative learning and memory. The aim of this article is to provide a brief summary of what is known about the anatomy of the BF, to give a historical overview of experimental studies of BF function, and finally to discuss current work on the cognitive functions of this structure.

Anatomy and Connections of the Basal Forebrain

Cholinergic BF neurons are intermingled with a substantial population of noncholinergic neurons that share similar projection patterns. Estimates of the proportion of cortically projecting BF neurons that are cholinergic vary from study to study, but generally range from 30% to 50%. Many noncholinergic neurons in the BF may be local circuit neurons, receiving cortical input and modulating activity of cortically projecting cholinergic and noncholinergic neurons.

The BF can be divided into four groups of cells: the medial septum (MS), projecting primarily to the hippocampus; the vertical limb of the diagonal band of Broca (VDB), projecting primarily to the hippocampus and cingulate cortex; the horizontal limb of the diagonal band of Broca (HDB), projecting primarily to the olfactory bulb, piriform cortex, and entorhinal cortex; and the nucleus basalis magnocellularis (NBM), or substantia innominata (SI), projecting primarily to the neocortex and amygdala. These cholinergic nuclei have also been designated Ch1–Ch4; this nomenclature approximately corresponds to the anatomical divisions. The organization of the BF is similar in the primate and in the rat, although subdivisions of the nucleus basalis can be identified reliably in the primate but not in the rat. Some areas of the BF are reciprocally connected with their targets; anatomical experiments in rats have shown that the medial septum and diagonal band project to the hippocampus and medial prefrontal cortex and receive projections back from these structures. Similarly, the nucleus basalis is reciprocally connected with the amygdala and cortex. The BF also receives inputs from the hypothalamus, as well as the midbrain and upper pons. Inhibitory projections from the nucleus accumbens to the BF may play a particular role in regulating cortical ACh release.

Both nicotinic and muscarinic ACh receptors are important for cognition. However, the density of muscarinic receptors in the neocortical target sites of the BF is much greater than that of nicotinic receptors. Unlike nicotinic receptors, which are ionotropic in nature and thus fast-acting, muscarinic receptors are metabotropic and are thus slow to respond, but they have long-lasting effects, making them likely candidates for modulating the BF's cognitive processes.

The Basal Forebrain and Memory

A central tenet of the 'cholinergic hypothesis' has been that cholinergic dysfunction specifically is partially or wholly responsible for the memory deficits seen after damage to the BF. Thus, considerable research has been devoted to the magnocellular cholinergic neurons of the BF and their function in memory. Alzheimer's disease involves loss of the magnocellular cholinergic neurons of the BF. Postmortem examinations of the brains from Alzheimer's patients have found a marked cell loss in the BF relative to the brains of healthy controls. Brains of Alzheimer's patients show a loss of the enzymes choline acetyltransferase and acetylcholinesterase in the cortical targets of the BF; choline acetyltransferase synthesizes ACh. Once released from the presynaptic terminal, ACh is degraded by acetylcholinesterase. Cognitive impairment in patients with Alzheimer's disease is correlated with the extent of cholinergic loss. These findings suggest that the loss of cholinergic transmission in the BF contributes to the memory dysfunction observed in Alzheimer's disease.

Additional support for a memory role for the BF comes from patients with amnesia resulting from an anterior communicating artery aneurysm (ACoA). Infarcts associated with an ACoA do not typically damage the traditional cerebral areas implicated in amnesia (i.e., medial temporal and diencephalic structures). ACoA amnesic patients that sustained infarcts resulting in lesions confined to the BF, with no evidence of diencephalic or medial temporal involvement, have relatively intact immediate recall and intact implicit memory, but impaired delayed recall and an increased susceptibility to proactive interference. The increased susceptibility to proactive interference, the phenomenon whereby previous learning disrupts later learning, is hypothesized to be due to deficient attentional inhibitory mechanisms. This presumed decrease in attentional processing ultimately leads to poor integration of old and new learning. Accordingly, these patients have also shown reduced information processing speed and poor divided attention ability. The occurrence of sensory information processing deficits such as these, following BF damage, begins to shed light on a role for the cholinergic BF in more attention and information-processing aspects of cognitive function (to be discussed in the second half of this article).

The Alzheimer's and ACoA data suggest that the BF contributes to memory dysfunction, motivating the development of animal models of BF damage to further test the exact role of the cholinergic BF system in mediating memory dysfunction. In these models, the BF cholinergic system or its cholinergic targets are damaged and behavioral tests are performed to determine the pattern of memory function consequent to BF damage. The validity of these models depends on the similarity of the cognitive processes being tested in animals to those in humans, as well as on the anatomical homology between the BF in animals and the BF in humans.

To confirm the cholinergic hypothesis in an animal model, two challenges must be met: (1) any memory deficits observed must be due to a specific impairment in memory processes, as opposed to disruption of other cognitive or noncognitive processes; and (2) the deficits must not be attributable to destruction of other populations of noncholinergic neurons in the BF. Most initial studies of the cognitive functions of the BF used stereotaxic placement of electrolytic or excitotoxic lesions in different regions of the BF in rats. Interpretation of these studies was made difficult by the fact that electrolytic lesions damage mixed populations of neurons as well as passing fibers, and that excitotoxic lesions (e.g., ibotenic acid, kainic acid, quisqualic acid, and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)) preserve the passing fibers but still damage mixed populations of neurons, such as γ-aminobutyric acid (GABA) neurons. Excitotoxins work by inducing cell death through hyperstimulation of glutamate receptors. The nonspecific nature of excitotoxins makes it difficult to conclude whether any cognitive impairment found is solely due to cholinergic damage or rather to the combined damage of different cell types. Ibotenate and other excitotoxin-induced lesions of the NBM disrupt learning in working memory tasks (such as delayed alternation) and reference memory tasks (measuring spatial or conditional visual discrimination learning).

Although numerous studies involving ibotenic acid lesions of the BF have demonstrated deficits in spatial and conditional learning and memory in rats, lesions of the BF made with more selective excitotoxins (quisqualic acid and AMPA) resulted in greater destruction of cortically projecting cholinergic neurons of the NBM than that seen following ibotenate lesions, but they did not induce the majority of learning and memory deficits observed after ibotenate lesions. Consequently, many of the deficits induced by ibotenate lesions of the BF cannot be attributed to a primary destruction of cholinergic neurons. Histological examinations conducted following ibotenic acid injections into the BF have revealed substantial damage to the dorsal and ventral pallidum of rats, suggesting that basal ganglia damage may have contributed to the memory deficits found.

Because excitotoxic lesion techniques are not selective for cholinergic neurons and therefore the resulting behavioral deficits may be due in part to damage in some neighboring neuronal system, explicit tests of

the cholinergic hypothesis require a neurotoxin specific for BF cholinergic neurons. The immunoglobulin G immunotoxin, 192 IgG-saporin, provides a route for directly targeting the toxin, saporin, to cholinergic neurons *in vivo* in the rat, without damaging noncholinergic neurons at the lesion site. Immunotoxic lesions restricted to specific cholinergic nuclei of the BF have generally failed to produce memory deficits, suggesting that damage to noncholinergic neurons in the BF and surrounding regions of the basal ganglia may have been responsible for the excitotoxic-induced memory impairments previously found. Rats with saporin lesions of individual BF nuclei or even those with extensive lesions of the entire BF (MS/VDB, HDB, and NBM/SI) show little or no deficits in memory. For example, spatial learning in the water maze, working memory in the radial arm maze, and contextual fear conditioning, all tasks dependent on the integrity of the hippocampus, have been reported to be normal following extensive lesions of BF cholinergic neurons. It is noteworthy that the more specific the cholinergic lesion, the less dramatic the effects on learning and memory. The studies with less specific excitotoxic lesions underscore the importance of careful validation that BF lesions intended to be specific to cholinergic neurons actually spare noncholinergic neurons at the lesion site. Importantly, many experiments have used intracerebroventricular infusions of 192 IgG-saporin, which damage cerebellar Purkinje cells in rats in addition to cholinergic BF neurons. These experiments have revealed severe impairments in many spatial learning and memory tasks, although the extent to which this represents a more complete cholinergic lesion than can be achieved by site-specific injections, a primary effect of cerebellar damage, or some interaction of the two factors remains somewhat controversial. What is clear is that extensive loss of BF cholinergic neurons – as severe or more severe than that seen in late stages of Alzheimer's disease – does not result in a profile of memory impairment in rats that resembles dementia, or even the profile of cognitive impairment seen in normal aging in rats.

The foregoing experiments in rats that suggest a role for the BF in memory function, but not a specific role for cholinergic BF neurons in memory, are supported by experiments in nonhuman primates. Like humans, nonhuman primates have a more spatially distinct NBM, making generalizations between the cognitive deficits seen in monkeys and those seen in humans more convincing than generalizations between rats and humans. Investigations have been made into the effect of ibotenic acid lesions placed stereotaxically into the BF cholinergic nuclei of Old World (macaque) monkeys. A battery of tests (delayed nonmatch-to-sample, delayed response, concurrent visual discrimination, and a visuospatial attention task) revealed no memory deficits despite the presence of an impairment in visuospatial attention. Other experiments with ibotenic acid lesions of BF nuclei in macaque monkeys report an impairment in recognition memory assessed using delayed nonmatching-to-sample, but this impairment was not delay dependent, casting doubt on whether it represented an impairment of memory or some failure of attention or encoding on the sample trial. There is some contradiction between the findings of ibotenic lesions of the BF nuclei in Old and New World monkeys. Ibotenate lesions to the NBM led to impairments in retention of preoperatively learned visual discriminations and reversal learning of visual discriminations postoperatively in New World monkeys (marmosets). This was accounted for by possible nonspecific damage caused by the lesions made in New World monkeys. Administration of the putative cholinergic toxin ME20.4 IgG-saporin to the BF nuclei of marmosets led to perceptual impairments on retention and acquisition of visual discriminations, although sparing of noncholinergic neurons at the lesion site in these monkeys was not verified. Similar experiments on damage to the VDB in marmosets revealed impairments in retention and new learning of visuospatial conditional discriminations, similar to those resulting from damage to the hippocampus. Related experiments in rats have discovered similar effects in the absence of general impairments in hippocampal-dependent learning, suggesting that this may represent a specific contribution of ACh to resolving interference in memory rather than a general role for cholinergic input in enabling hippocampal function.

Aims to increase ACh levels in the clefts of the cholinergic synapse with acetylcholinesterase inhibitors, or to directly replace ACh with cholinomimetic drugs, have largely been unsuccessful in reversing cognitive deficits in Alzheimer's patients. This may be because direct enhancement of ACh levels may disrupt any remaining intact cholinergic transmission. Another complementary, pharmacological strategy to treat the cognitive deficits consequent to cholinergic loss has been to modulate the cholinergic system by transsynaptic mechanisms – for example, by affecting cholinergic transmission with GABA. Ultimately, this approach may be more effective in alleviating the associated cognitive deficits, since this acts to amplify the endogenous physiological cholinergic signal. However, given that loss of cholinergic neurons does not seem to be a primary cause of memory impairment in these conditions, even pharmacological treatments aimed at enhancing function of remaining cholinergic neurons by transsynaptic modulation might be expected to have little beneficial effect on memory.

In summary, many aspects of learning and memory proceed normally in the absence of cholinergic input from the BF. In contrast, BF cholinergic neurons do appear to play an essential role in regulating attentional processing capacity. We turn to specific evidence for the involvement of the BF in attentional processing in the following section.

The Basal Forebrain and Attention

Mechanisms of attention are thought to enable the efficient allocation of processing resources in order to respond to important events in the environment. There are different types of attentional processes: selective attention targets attentional resources on a restricted number of sensory stimuli and excludes other sensory stimuli; sustained attention reflects a state of readiness to detect and respond to unpredictable or rare events; and divided attention refers to simultaneous monitoring of several different channels of sensory information. Attention filters what sensory information is introduced into memory. The BF appears to regulate many of these attentional functions; recent studies of lesions of the cholinergic forebrain in rats and monkeys have suggested deficits the forms of attention. For example, the five-choice serial reaction time task, based on a human test of continuous performance, measures sustained spatial attention in rats. In this task, rats are required to monitor the occurrence of a light stimulus in one of five locations. With this paradigm, both excitotoxic (e.g., ibotenate acid, quisqualic acid, and AMPA) and saporin lesions of the NBM produce impairments in performance, which are ameliorated by increasing the duration or intensity of the stimuli. In support of the involvement of the cholinergic BF, this deficit is alleviated by an acetylcholinesterase inhibitor, physostigmine, an enzyme that boosts the availability of ACh.

A role for the NBM in attention has also been demonstrated in a temporal divided attention task, where at times rats are required to divide their attention between two simultaneous stimuli: rats with excitotoxic lesions of the NBM are specifically impaired in their ability to time two stimuli simultaneously. In a cross-modal divided-attention task, infusion of 192 IgG-saporin into the BF selectively increases the response latencies in a bimodal condition, without affecting the accuracy or the retrieval and execution of the stimulus–response action in either the visual or auditory modality.

BF neurons, including the cholinergic neurons, receive direct contacts from GABAergic afferents and are inhibited by GABA. Similar to the effects of cholinergic lesions, infusions of a benzodiazepine receptor agonist into the BF impair the performance of rats tested in a sustained-attention task without altering the efficacy of perceptual processes. Infusion of the $GABA_A$ receptor agonist, muscimol, directly into the BF both decreases the activity of the cortically projecting NBM neurons and impairs accuracy on the five-choice task. Two-lever vigilance tasks present the rat with successive trials of changes in the intensity of continuously delivered auditory (e.g., white noise) or visual (e.g., illumination) stimuli. One lever measures the 'hits' and the other lever measures the 'correct rejections.' In two-lever vigilance tasks for rats, 192 IgG-saporin lesions of the NBM and intra-NBM administration of $GABA_A$ benzodiazepine receptor agonists have impaired vigilance performance.

In a visuospatial attention task, Old World monkeys with BF lesions show impairments in shifting attention. A brief cue indicates the expected location of a target stimulus. The difference in response time to a target that appears at expected and unexpected locations is used as the measure of shifting of attention. A slower response time on invalid trials, in which the target appears at an unexpected location, compared to valid trials, in which the target appears at the expected location, indicates that a respondent has paid attention to the cue that is signaling the likely location of a target. Monkeys with BF lesions showed disproportionately longer response times on invalid trials, suggestive of an impairment in disengaging attention from the invalidly cued location. This was not due to a general impairment in motor skills, since the response to the target following the initiation of a response was comparable in both control and BF lesioned monkeys. A quite similar result has also been described in rats with selective lesions of cholinergic neurons in the NBM. It is important to note that this attentional impairment was quite selective; the monkeys with BF lesions were not impaired on tests of visual learning and memory, and rats with selective cholinergic lesions of the NBM were not impaired on a variety of cognitive tasks, including tests of spatial learning and memory.

In vivo microdialysis studies provide further support for an important role for the cholinergic BF in attention, showing that tasks placing both low and high demands on sustained attention result in significant increases in ACh efflux in the frontoparietal cortex of rats, one of the cortical targets of the NBM. However, bilateral saporin lesions of the NBM only impaired performance on high attentional-demand tasks, suggesting that cortical cholinergic neurotransmission is especially needed for tasks requiring high levels of attentional processing.

Experiments that examine the regulation of attentional processing in associative learning paradigms have identified dissociable roles for rostral (septal) and caudal (NBM) regions of BF cholinergic neurons.

Increases in conditioned stimulus processing, brought about by violations of learned contingencies, require the integrity of cholinergic neurons in the NBM, but not cholinergic neurons in the MS/VDB. Decreases in conditioned stimulus processing, which occur when conditioned stimuli are preexposed in the absence of reinforcement, or are consistent predictors of another event, require the integrity of cholinergic neurons in the MS/VDB, but not the NBM. The BF has reciprocal connections with the central nucleus of the amygdala, which has potential for widespread influences on cortical processing through direct projections to lower- and higher-order sensory areas and indirectly through projections to cholinergic neurons in the BF. Stimulation of the amygdala central nucleus has been found to influence BF electroencephalogram (EEG) patterns consistent with ACh release in rats and humans. Cortical ACh release stimulated by the amygdala is thought to induce a state of cortical readiness for processing sensory information. This suggests that the BF cholinergic system may provide an output pathway for the regulation of attention and cortical information processing by the amygdala. Indeed, evidence from crossed-lesion studies suggests that projections from the central nucleus to the NBM are critical for producing increases in conditioned stimulus processing: disconnection of the NBM from the central nucleus of the amygdala by crossed unilateral lesions also disrupts the ability to increase attentional processing in response to a violation of expectancy. Similarly, such lesions impair sustained visual attention in the five-choice paradigm.

As the cholinergic hypothesis shifts gears from a purely memory focus to one that encompasses attentional processes, the question of how the BF modulates attention has come to the forefront. ACh enhances the processing of incoming sensory information from a variety of modalities. For example, following training on a vigilance task in which rats were required to press a response lever located on the right when a visual signal was presented, and to press a response lever located on the left when no signal was presented, bilateral saporin lesions of the NBM significantly impaired the ability of rats to detect signals, but not their ability to correctly reject nonsignals. It has thus been proposed that ACh modulates attention by regulating the signal-to-noise ratio (SNR) of incoming sensory information. By minimizing the processing of irrelevant sensory signals as well as previously learned information, the encoding of important incoming sensory information can occur much more efficiently. Consistent with such a proposal, both pharmacological data and lesion-based research have implicated the BF cholinergic system in the encoding of new information, but not its subsequent retrieval. Consistent with such a model, Alzheimer's patients show normal rates of forgetting, but are impaired at the initial acquisition of information, suggesting an important role for the cholinergic BF in the encoding of sensory information.

A model has been proposed of how an active BF cholinergic system is conducive to the encoding of new information, while a hypoactive BF cholinergic system is conducive to the consolidation and retrieval of previously learned information. When there are high levels of cholinergic activity in the cortex, afferent input from external sensory stimuli is enhanced, while intrinsic processing within cortices is reduced. Intrinsic processing is associated with the reactivation of neural connections representing previously learned information, and a reduction in such processing would prevent the recall of previously learned information from interfering with the encoding of new information, thereby enhancing the SNR. This has been demonstrated by studies showing that systemic muscarinic cholinergic blockade with the muscarinic receptor antagonist, scopolamine, increased proactive interference from previously learned odor- and word-pair associations during the encoding of novel associations in nonhuman and human animals, respectively. A follow-up study has revealed BF contributions, as administration of a low dose of scopolamine resulted in increased proactive interference only in rats with bilateral saporin lesions of the NBM, not in sham-lesioned controls. Although not a direct measure of cholinergic activity, functional magnetic resonance imaging (fMRI) permitted investigation of underlying cholinergic brain systems involved in proactive interference resolution *in vivo* and confirmed that this BF contribution is conserved in humans.

This role of the BF cholinergic system in reducing SNR and minimizing interference from irrelevant sensory signals and previously learned information may account for the effect of BF cholinergic lesions in memory tasks, when they appear. The involvement of these neurons in reducing proactive interference has already been noted in discrimination learning paradigms in which interference is explicitly produced. This was also suggested to account for the effect of loss of MS/VDB cholinergic neurons on visuospatial conditional learning in rats; these lesions impair an environment–spatial conditional task in which local environmental cues signal the location of a reward when the two components of the conditional problem are learned simultaneously, but not sequentially. Neurophysiological experiments have shown that place fields of hippocampal neurons are resistant to remapping when local environmental cues are changed, providing a physiological mechanism for increased susceptibility to proactive interference.

Thus far, selective cholinergic lesions in the BF system appear to have limited effect on measures of learning and memory, but rather impair attentional processing and information processing. It is assumed that attentional processes are involved in the primary stages in information processing, so it is remarkable that no dramatic effects in various learning and memory tasks have been observed following damage to BF cholinergic neurons. If attentional processes are impaired, then a performance deficit would be expected in most learning and memory tasks, but it appears that this is not the case after selective cholinergic lesions. However, the statement that "BF damage impairs attention" is an oversimplification. Instead, the effects of these lesions might be better characterized as producing an impairment in the ability to regulate attentional processing appropriately, thereby reducing the efficiency of sensory information encoding. Indeed, the impairments observed in attentional tasks following damage to the BF cholinergic system may represent one instance of a general physiological involvement of ACh in reducing cortical SNR.

Conclusions

The BF is anatomically situated to regulate information processing in cortical and limbic structures involved in a wide array of cognitive functions. Indeed, extensive damage to the BF (to both cholinergic and noncholinergic components) results in a broad array of cognitive impairments. In contrast, damage limited to BF cholinergic neurons produces more restricted impairments in the regulation of attentional processing. Hence, BF cholinergic neurons appear to play a selective role in regulating cortical information processing, rather than a generalized function in supporting the functions of their target areas. This apparently restricted function of BF cholinergic neurons suggests that a loss of these cells is probably not the only factor in producing cognitive deficits in conditions like Alzheimer's disease, or in the amnesia consequent to aneurysms in the BF.

The characterization of BF cholinergic function as purely 'attentional' may fail to capture the array of cognitive contributions from these neurons. These neurons are likely involved in other aspects of sensory processing and representational plasticity in cortical areas. A general operating principle for the role of BF cholinergic neurons in cognition remains elusive, as does a role for the BF generally in cognitive function. The BF, particularly the cholinergic component, may be more broadly involved in the encoding and processing of sensory information, as well as regulating aspects of conscious experience; hence abnormalities in BF function could contribute to a wide variety of cognitive impairments in both neurodegenerative disease and in psychopathology. These recent hypotheses represent a useful guiding principle for future studies of BF function, both in animal models and in human clinical populations.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Attentional Networks; Attentional Functions in Learning and Memory; Cholinergic Pathways in CNS; Forebrain: Early Development; Functional Amnesia; Memory Disorders; Nicotinic Acetylcholine Receptors.

Further Reading

Baxter MG and Chiba AA (1999) Cognitive functions of the basal forebrain. *Current Opinion in Neurobiology* 9: 178–183.

Blockland A (1996) Acetylcholine: A neurotransmitter for learning and memory? *Brain Research Reviews* 21: 285–300.

De Luca J and Diamond BJ (1995) Aneurysm of the anterior communicating artery: A review of neuroanatomical and neuropsychological sequelae. *Journal of Clinical Experimental Neuropsychology* 17: 100–121.

De Rosa E, Baxter MG, and Hasselmo ME (2001) Contribution of the cholinergic basal forebrain to proactive interference from stored odor memories during associative learning in rats. *Behavioral Neuroscience* 115: 314–327.

De Rosa E, Desmond JE, Anderson AK, et al. (2004) The human basal forebrain integrates the old and the new. *Neuron* 41: 825–837.

Everitt BJ and Robbins TW (1997) Central cholinergic systems and cognition. *Annual Review of Psychology* 48: 649–684.

Hasselmo ME and McGaughy J (2004) High acetylcholine levels set circuit dynamics for attention and encoding and low acetylcholine levels set dynamics for consolidation. *Progress in Brain Research* 145: 207–231.

Janisiewicz AM, Jackson O 3rd, Firoz EF, et al. (2004) Environment-spatial conditional learning in rats with selective lesions of medial septal cholinergic neurons. *Hippocampus* 14: 265–273.

Kilgard MP and Merzenich MM (1998) Cortical map reorganization enabled by nucleus basalis activity. *Science* 279: 1714–1718.

Parent MB and Baxter MG (2004) Septohippocampal acetylcholine: Involved in but not necessary for learning and memory? *Learning and Memory* 11: 9–20.

Sarter M, Hasselmo ME, Bruno JP, et al. (2005) Unraveling the attentional functions of cortical cholinergic inputs: Interactions between signal-driven and cognitive modulation of signal detection. *Brain Research Reviews* 48: 98–111.

Voytko ML (1996) Cognitive functions of the basal forebrain cholinergic system in monkeys: Memory or attention? *Behavioural Brain Research* 75: 13–25.

Vuckovich JA, Semel ME, and Baxter MG (2004) Extensive lesions of cholinergic basal forebrain neurons do not impair spatial working memory. *Learning & Memory* 11(1): 87–94.

Wainer B and Mesulam MM (1990) *Ascending Cholinergic Pathways in the Rat Brain.* Oxford: Oxford University Press.

White KG and Ruske AC (2002) Memory deficits in Alzheimer's disease: The encoding hypothesis and cholinergic function. *Psychonomic Bulletin & Review* 9: 426–437.

Záborszky L, Gaykema RP, Swanson DJ, et al. (1997) Cortical input to the basal forebrain. *Neuroscience* 79: 1051–1078.

Basal Ganglia and Oculomotor Control

O Hikosaka, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Oculomotor Control by the Basal Ganglia

The basal ganglia control eye movements. This is illustrated by a variety of movement disorders caused by dysfunction of the basal ganglia, such as Parkinson's disease and Huntington's disease. A fixed, vacant facial expression of patients with Parkinson's disease, which is often called the Parkinson's mask, is due to the paucity of movements in the face, including the paucity of eye movements. Similar symptoms were observed in MPTP-induced dopamine deficient monkeys (Figure 1(a)). Most affected among various kinds of eye movements are smooth pursuit and saccades, which require voluntary control. Parkinsonian patients are often impaired in smoothly pursuing a moving object (deficit in smooth pursuit). They are also often impaired in shifting their gaze from one position in space to another (deficit in saccade). While their relation to smooth pursuit is still unclear, the basal ganglia are shown to play a critical role in the control of voluntary saccadic eye movement.

The impairment in saccadic eye movement in patients with basal ganglia disorders has been repeatedly demonstrated using more rigorous tests with accurate measurement of eye position. Typically, the subjects are required to fixate their gaze on a spot of light (target) on the screen and, if the target steps, follow it by quickly shifting their gaze. This is called a visually guided saccade task. Compared with age-matched control subjects, saccades of parkinsonian patients tend to be small in amplitude (i.e., hypometric), slow, and delayed (i.e., long latency). Curiously, the deficit in saccades is often more severe if there is no visible object and the saccades must rely on memory. In a memory-guided saccade task, a target appears briefly while the subject is fixating at the central spot and the subject has to make a saccade after a delay to the position where the target was presented. Parkinsonian patients are more impaired or selectively impaired in memory-guided saccades than in visually guided saccades. This phenomenon may illustrate the context-dependent movement deficits in Parkinson's disease, which are widely recognized among neurologists. Selective deficits in memory-guided saccades are observed in other basal ganglia disorders, including Huntington's disease. The similarity between Parkinson's and Huntington's diseases is noteworthy because they are caused by different mechanisms, the former by a loss of neurons in the substantia nigra (SN) and the latter by a loss of neurons in the caudate nucleus (CD). This suggests that the SN and the CD work together for the control of saccadic eye movements.

Neural Circuits in the Basal Ganglia for Oculomotor Control

How the basal ganglia might control eye movements has been studied by single unit studies using animals trained on the visually guided and memory-guided saccade tasks. Electrical activity of single neurons was recorded with microelectrodes and was correlated with saccadic eye movements. Saccade-related activity has been found in various nuclei in the basal ganglia, including the SN, CD, subthalamic nucleus (STN), and globus pallidus (GP). Such saccade-related neurons are clustered in a subregion of each nucleus: dorsolateral part of the pars reticulata of the SN (SNr), central-ventral part of the CD, ventral part of the STN, and dorsal part of the external segment of the globus pallidus (GPe). Anatomical studies have shown that these saccade-related parts are connected within the basal ganglia and with saccade-related regions outside the basal ganglia (Figure 1(b)). For example, the saccade-related part of the CD receives inputs from the frontal eye field and the supplementary eye field in the frontal cerebral cortex, whereas the saccade-related part of the SNr projects their axons to the superior colliculus (SC). Note that neurons related to skeletal movements are found in different subregions in the basal ganglia, such as the putamen (equivalent to the CD) and the internal segment of the globus pallidus (GPi) (equivalent to the SNr). These facts are consistent with the idea that there are functional subdivisions within the basal ganglia as well as in larger networks including the cerebral cortex and the cerebellum, each of which may form a closed-loop functional unit. Recent studies, however, indicate that such functional segregation is not perfect.

An important feature of the basal ganglia circuits is that they use inhibitory connections as a primary means to convey signals (Figure 1(b)). Each area contains projection neurons and interneurons. While cortical inputs to the CD are excitatory and use glutamate as a transmitter, projection neurons in all areas in the basal ganglia, except the STN, are thought to be GABAergic and inhibitory. This means that the polarity of a signal is reversed each time it passes

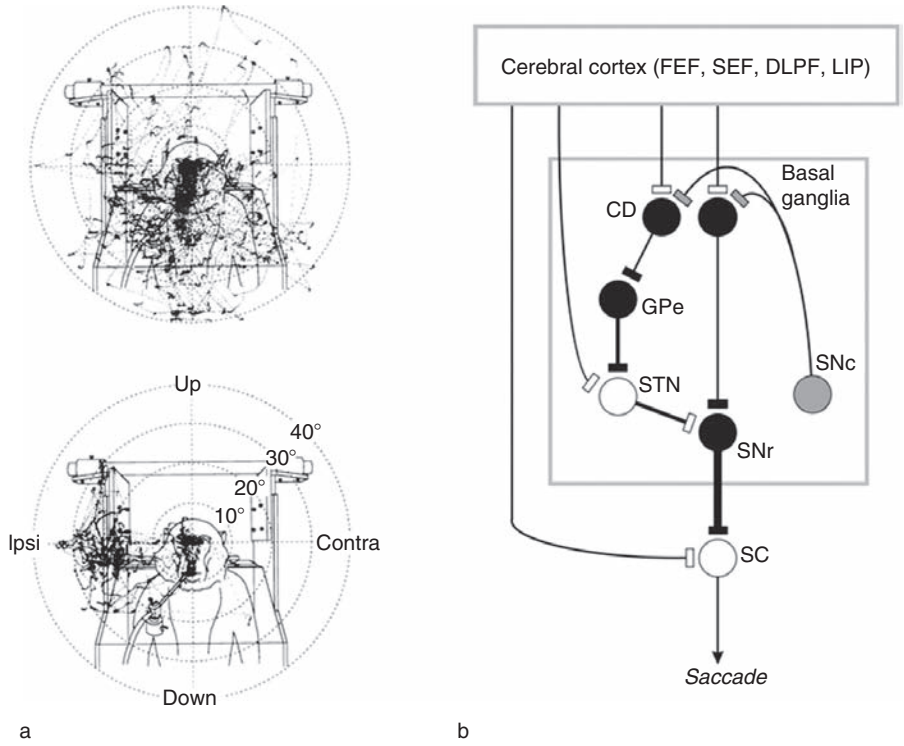

Figure 1 Abnormal eye movements in monkeys induced by basal ganglia dysfunctions. (a) Eye movements of a monkey while he was looking in a mirror before (top) and 31 days after (bottom) dopaminergic denervation of the left caudate nucleus by MPTP. Shown are the trajectories of eye movements superimposed on the view seen in the mirror by the monkey. (b) Basal ganglia neural network involved in the control of saccadic eye movement. CD, caudate nucleus; SNr, substantia nigra pars reticulata; SC, superior colliculus; SNc, substantia nigra pars compacta; GPe, globus pallidus external segment; STN, subthalamic nucleus; FEF, frontal eye field; SEF, supplementary eye field; DLPF, dorsolateral prefrontal cortex; LIP, area LIP in parietal cortex. Excitatory and inhibitory neurons and synapses are indicated by open and filled symbols, respectively. Gray symbol indicates dopaminergic neuron which exerts modulatory effects on CD neurons. The thickness of the line (axon) roughly indicates the level of spontaneous activity. The direct excitation of SC neurons by inputs from the cerebral cortex is gated by the inhibitory input from the SNr. (a) From Miyashita N, Hikosaka O, and Kato M (1995) Visual hemineglect induced by unilateral striatal dopamine deficiency in monkeys. *Neuroreport* 6: 1257–1260.

through projection neurons in one area. There are at least three parallel pathways in which saccade-related signals can be processed in the basal ganglia (Figure 1(b)): (1) direct pathway from the CD to the SNr; (2) indirect pathway from the CD to the SNr through the GPe and/or the STN; and (3) hyperdirect pathway from the STN to the SNr. Since the direct pathway consists of a series of two inhibitory connections (CD and SNr), the net effect is facilitatory. Since the indirect and the hyperdirect pathways consist of three and one inhibitory connection, respectively, the net effect is inhibitory. The basal ganglia thus could facilitate or inhibit motor processes by selectively using these pathways. The target area for saccadic eye movement is the SC.

With respect to saccadic motor control in general, the basal ganglia system is situated as a side path which has been added to the direct effect of the cerebral cortex on the SC (Figure 1(b)). Probably, the most important question is: how unique is the function of the basal ganglia, compared with the direct cortico-SC effect? The answers to this question should be found in the content of information carried by neurons in the CD and SNr.

Caudate Nucleus

The CD is a large structure in the basal ganglia and, together with the putamen, is called the striatum or the dorsal striatum. A majority of inputs to the basal ganglia is destined for the striatum which thus acts as the input station of the basal ganglia. The neural circuit in the basal ganglia related to eye movements originates in the CD and converges on the SNr which then projects to the SC (Figure 1(b)). In addition to the cortical inputs, the entire CD (together with the putamen and the ventral striatum) receives diffuse inputs from dopaminergic neurons in the SN pars compacta (SNc) and its surrounding regions. It is likely that particular combinations of these inputs create signals unique to CD neurons.

A majority of neurons comprising the CD are called medium-spiny neurons: neurons with medium-sized cell bodies and many dendrites thickly covered with spines. They are the projection neurons, neurons that project axons to the outside of the CD. They are GABAergic and inhibitory. The projection neurons are highly hyperpolarized in the resting state and emit action potentials only occasionally. A minority of neurons (less than 5%) in the CD consist of several types of interneurons. One conspicuous type is the cholinergic interneuron, which is characterized anatomically as large-aspiny neuron. They fire tonically and irregularly and are often called tonically active neurons (TANs).

Single unit studies using monkeys trained on saccade tasks have revealed that many CD projection neurons are clearly related to saccadic eye movements (Figures 2(a) and 2(b), top). Some of them respond to visual stimuli that potentially induce saccades to them. Other neurons become active before saccades. These visual-saccadic neurons have response fields which are usually centered in the contralateral field (Figure 2(c)). The responses are often highly dependent on the context. Visual responses may be enhanced if the animal attends to or memorizes the stimulus. Saccadic activity may be present only when the saccade is guided by memory or only when it is guided by visual stimuli. The neurons usually do not fire in relation to spontaneous saccades. Intermingled with such visual-saccadic neurons are found more complex neurons, such as those related to expectation of task-specific events or reward. Such complex nature of CD projection neurons appears to reflect the convergent inputs from the cortical eye fields (frontal eye field (FEF) and supplementary eye field (SEF)) and from the dorsolateral prefrontal cortex.

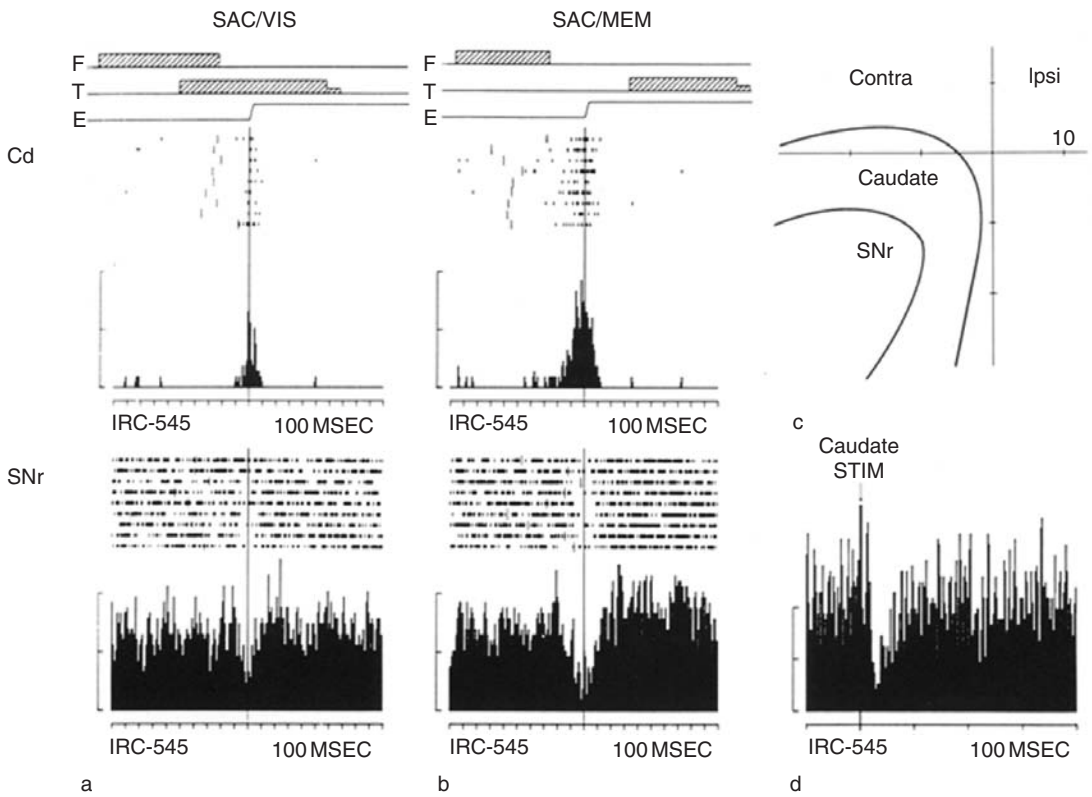

Figure 2 Oculomotor signal may be transmitted from caudate (Cd) to substantia nigra pars reticulata (SNr). Spike activities of a Cd neuron (top) and an SNr neuron were compared in two different modes of saccade: visually guided saccade (SAC/VIS) (a) and memory-guided saccade (SAC/MEM) (b). The spike activity, shown by spike rasters and histograms, were aligned on the onsets of the saccades, as indicated by the task schemes above: F, central fixation point; T, target point located with the movement fields of the neurons; E, schemetic eye position. (a) The target came on while the fixation point was still present, and the monkey was allowed to make a saccade after the fixation point went off. (b) The position of the target was indicated 2–3 s before the fixation point went off (not shown), and the monkey made a saccade to the remembered location of the cue stimulus. Short vertical bars in the spike rasters indicate the disappearence of the fixation point. (c) Movement fields of the Cd and SNr neurons. (d) Inhibition of the SNr neuron by stimulation (100 μA, single pulse) applied at the point close to the Cd neuron. Calibrations on the left of the histogram indicate 50 and 100 spikes s^{-1}. Reproduced from Experimental Brain Research, Vol. 95, 1993, pp. 457–472, Effects of Candate nucleus stimulation on subtantia nigra cell activity in monkey, figure 10, copyright Springer. With kind permission of Springer Science and Business Media.

Substantia Nigra Pars Reticulata

The SN is composed of two parts: pars reticulata (SNr), which contains GABAergic neurons, and pars compacta (SNc), which contains dopaminergic neurons. Saccade-related neurons are found in the SNr. The relation of the SNr to eye movements was first suggested by anatomical studies showing that the SNr project to the intermediate layer of the SC. The SC in mammals is known to be crucial for orienting of the eyes and head toward an object of interest. An equivalent structure, which is called the optic tectum, is universally present in other vertebrates and is under the control of the homolog of the basal ganglia. Thus, the role of the SNr (or the basal ganglia) in eye movements should be considered in a larger framework of animal behavior in which orienting is crucial for survival.

Neurons in the SNr are characterized by their rapid and tonic firing (Figure 3(a)). Their firing frequency is usually between 40 and 100 spikes s^{-1} in monkeys. Neurons in the GPi, many of which are related to skeletal movements, also fire rapidly and tonically. Thus, the high background activity is common to virtually all output neurons of the basal ganglia. Furthermore, virtually all of them are GABAergic. These facts indicate that neurons that receive outputs of the basal ganglia should be kept inhibited by them. The functional importance of the basal ganglia-induced inhibition can be demonstrated experimentally. The inhibition can be reduced or eliminated by injecting a small amount of a γ-aminobutyric acid (GABA) agonist (muscimol) into the SNr because SNr neurons have a high density of GABA receptors and therefore muscimol would inhibit their firing. After muscimol injection in the SNr, animals can no longer maintain stable eye position and make saccades continually and probably involuntarily (Figure 3(c)). This happens to all animals tested: monkeys, cats, and rats. Rats, compared with monkeys, exhibit a wider range of involuntary movements in addition to eye movements. A similar phenomenon occurs for skeletal movements when muscimol is injected in the globus pallidus internal segment. These findings may be relevant to the fact that patients with basal ganglia dysfunction usually exhibit some type of involuntary movements, such as tremor, dyskinesia, dystonia, ballism, and chorea. These involuntary movements may be caused by a reduction of the basal ganglia-induced inhibition. The reduction of inhibition could be sustained or phasic (possibly with rhythms).

However, the sustained inhibition alone can hardly be a motor control mechanism. Neurons in the SNr change (usually decrease) their firing rates in preparation for saccadic eye movement (Figures 2(a) and 2(b), bottom). Many SNr neurons stop firing in response to a visual stimulus if the animal is ready to make a saccade to it. Other neurons do so just

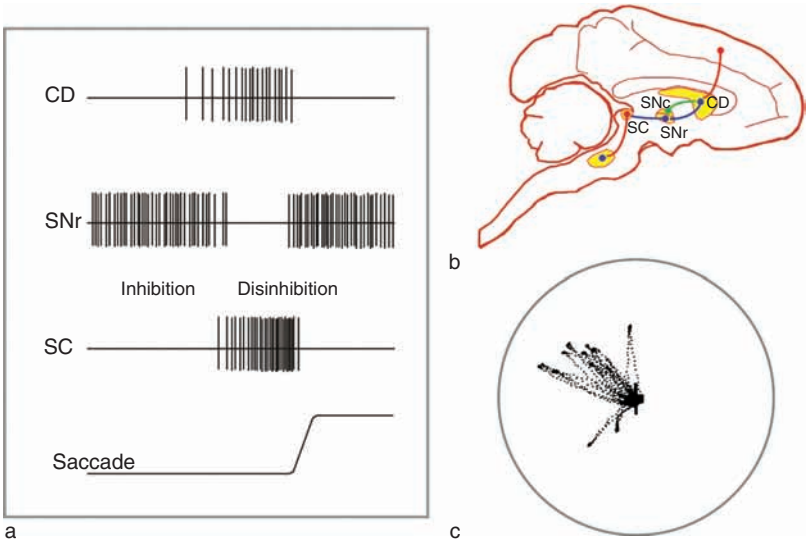

Figure 3 Basal ganglia mechanism for control of saccadic eye movement. (a) A cardinal saccade mechanism in the basal ganglia. The SC is normally inhibited by rapid firing of GABAergic neurons in the SNr. The tonic inhibition can be interrupted by GABAergic inputs from the caudate nucleus. This disinhibition, together with excitatory cortical inputs, allows SC neurons to fire in burst, which leads to a saccade to a contralateral location. (b) Simplified neural circuits in the basal ganglia for control of saccadic eye movement in a parasagittal view of the macaque brain. Red, blue, and green lines indicate excitatory, inhibitory, and modulatory connections. (c) Involuntary eye movement of a monkey after muscimol injection into right SNr, shown as trajectories of saccades during 2 s fixation periods. The monkey was unable to keep fixating at the central spot of light. From Hikosaka O and Wurtz RH (1985) Modification of saccadic eye movements by GABA-related substances. II. Effects muscimol in monkey substantia nigra pars reticulata. *Journal of Neurophysiology* 53: 292–308.

before the saccade. The inhibition of SNr neurons is caused, at least partly, by the GABAergic input from the CD. Electrical stimulation in the CD saccade-related region may elicit saccadic movements of the eye and the head to contralateral directions. If the electrical stimulation is short and weak, saccades are not evoked, but SNr neurons are inhibited and are occasionally facilitated (Figure 2(d)). The former is likely to be mediated by the direct CD–SNr inhibitory connection, while the latter is likely to be mediated by the indirect pathway through the GP (Figure 1(b)). Probably reflecting the powerful CD–SNr connection, many SNr neurons exhibit visual, saccadic, and memory-related activities with response fields centered in the contralateral field (Figure 2(c)). A group of SNr neurons are more strongly inhibited before memory-guided saccades than visually guided saccades (Figures 2(a) and 2(b), bottom).

Many of these SNr neurons project to the intermediate layer of the SC and have inhibitory synaptic contacts with saccadic burst neurons. This means that the SNr-induced tonic inhibition on SC neurons is removed or reduced before saccades (Figure 3(a)). Note that saccadic neurons in the SC receive excitatory inputs from many brain areas, especially saccade-related cortical areas: the FEF, SEF, and lateral intraparietal area (LIP). These excitatory cortical inputs, together with the SNr-induced disinhibition, would make SC neurons fire in a burst and the signal is sent to the brain stem saccade generators. Note, however, that SNr neurons may increase their activity before saccade. In such a case, the SC would be less likely to generate a signal to induce saccades.

To summarize, the SNr-induced inhibition on SC neurons acts as a gate for saccade generation (Figure 3(a)). SC neurons are constantly bombarded by excitatory inputs from many brain areas because there are so many objects that can attract our attention and gaze. However, these inputs are often incapable of inducing a burst of spikes in SC neurons due to the SNr-induced tonic inhibition. Only when the SNr-induced inhibition is reduced would SC neurons exhibit a burst of spikes reliably. This is probably a very efficient mechanism to select an appropriate action in a particular context.

Role of the Basal Ganglia in Reward-Oriented Eye Movement

Recent studies have revealed another striking feature of basal ganglia neurons: the relation to reward-oriented behavior. This feature has been investigated extensively in relation to saccadic eye movement. Typically, the amount of reward is biased depending on the direction of saccade: for example, rightward saccades are followed by a big reward and leftward saccades are followed by a small reward. This task has a strong behavioral impact: the saccade to the position associated with big reward is faster and earlier than that associated with small reward. The visual response of CD projection neurons is greatly enhanced and diminished if the saccade to the visual stimulus is expected to be followed by a bigger and smaller reward, respectively (Figure 4(a)). A minority of neurons exhibit the opposite reward modulation. The modulation is very common among visually responsive CD neurons (about 80%) and is often very strong such that the original directional tuning can be completed reversed. Similar reward-dependent modulation occurs for saccadic activity.

The relation of the CD to reward-oriented behavior is highlighted by a conspicuous group of CD projection neurons which cannot be classified as simply related to visual-saccadic processes. They exhibit growing activity while the animal is waiting for the go-signal for a saccade (Figure 5(a)). It appears to be related to saccade preparation. However, the activity occurs before the instruction is given to which position the saccade should be made. The activity is nonetheless spatially selective in that it is present only when saccades to a contralateral, rather than ipsilateral, position are followed by a big reward. If there is no positional bias in reward, the anticipatory activity is much weaker.

Further studies suggest that the reward-position-sensitive anticipatory activity is transmitted to the SC through the SNr (Figure 6). Suppose a bigger reward is associated with saccades to a right target than a left target, CD neurons on the left side would exhibit stronger anticipatory activity than those on the right side (according to the finding described above). Since a major effect of CD neurons on SNr neurons is inhibitory, SNr neurons on the left side would exhibit a stronger decrease in firing rates than those on the right side. This is actually observed experimentally. Since SNr neurons in turn inhibit SC neurons, the excitability of SC neurons on the left side would be elevated compared with those on the right side. This also has been confirmed experimentally. There is now a clear imbalance in excitability between the two sides of the SC: neurons on the left SC is more excitable than neurons on the right SC. This occurs before any instruction is given and far before a saccade is executed. This is an internal process based on the knowledge of the positional difference in reward amount. Under such a biased condition, a target that appears on the right side (i.e., associated with a big reward), which activates neurons on the left SC, would trigger a saccade more easily and more quickly than a target on the left side.

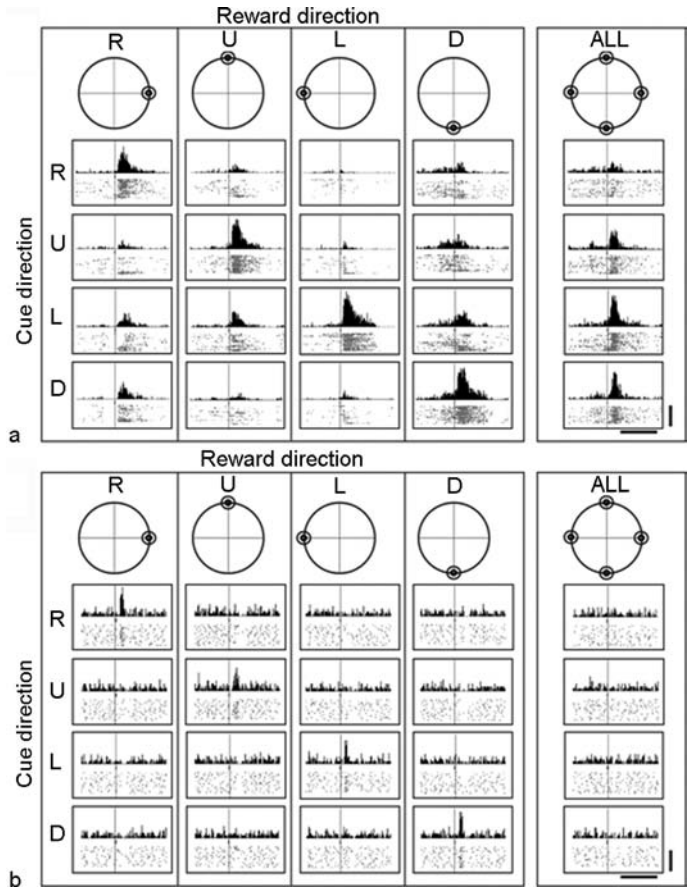

Figure 4 Responses of a right caudate neuron (a) and a left dopamine neuron (b) to visual cue in the memory-1DR and memory-ADR tasks. In one block of the memory-1DR task, only one particular direction was associated with reward. Since four target directions were used, the experiment was done in four blocks, which are shown in four left columns with different reward directions (indicated by R, U, L, D). In one block of memory-ADR task (indicated by ALL), all directions were associated with reward. Neurons were recorded from the right caudate and the left substantia nigra pars compacta (SNc). In the histogram/raster display (bin width: 20 ms), the neuronal discharge aligned on cue onset is shown separately for different cue directions; the cue directions were pseudorandomized. The reward direction is indicated by a dotted circle. The caudate neuron responded to the left (L) cue most strongly in ADR (ALL), but the response was greatly enhanced and depressed in 1DR when the cue indicated reward and no reward, respectively. The dopamine neuron showed no response to any cue in ADR. In 1DR, it responded to the cue with an excitation if it indicated an upcoming reward and with an inhibition if it indicated no reward. (a) From Kawagoe R, Takikawa Y, and Hikosaka O (1998) Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience* 1: 411–416.

In other words, the animal would make saccades more quickly to a more highly rewarded position (Figure 5(b)). Such a behavioral bias is consistently observed experimentally. These results suggest that the CD is a critical brain area for reward-oriented motivational behavior.

Dopaminergic Modulation of Information Processing in the Caudate Nucleus

Recent studies suggest that such strong reward-dependent modulation of CD neurons may be caused by dopaminergic inputs (Figure 1(b)). As mentioned above, CD projection neurons have dendrites with many spines on which both axons from cortical neurons and axons from midbrain dopaminergic neurons make synapses. A majority of cortical axons originate from the cortical eye fields and therefore are likely to carry spatial signals. In contrast, midbrain dopaminergic neurons are known to carry reward-related signals, but not spatial signals. They respond to reward which is given unexpectedly or, if reward is expected, to a sensory stimulus that predicts the reward. In the reward-biased saccade task, dopaminergic neurons respond by excitation and inhibition to sensory stimuli that predict a reward that is larger and smaller than what is expected, respectively (Figure 4(b)).

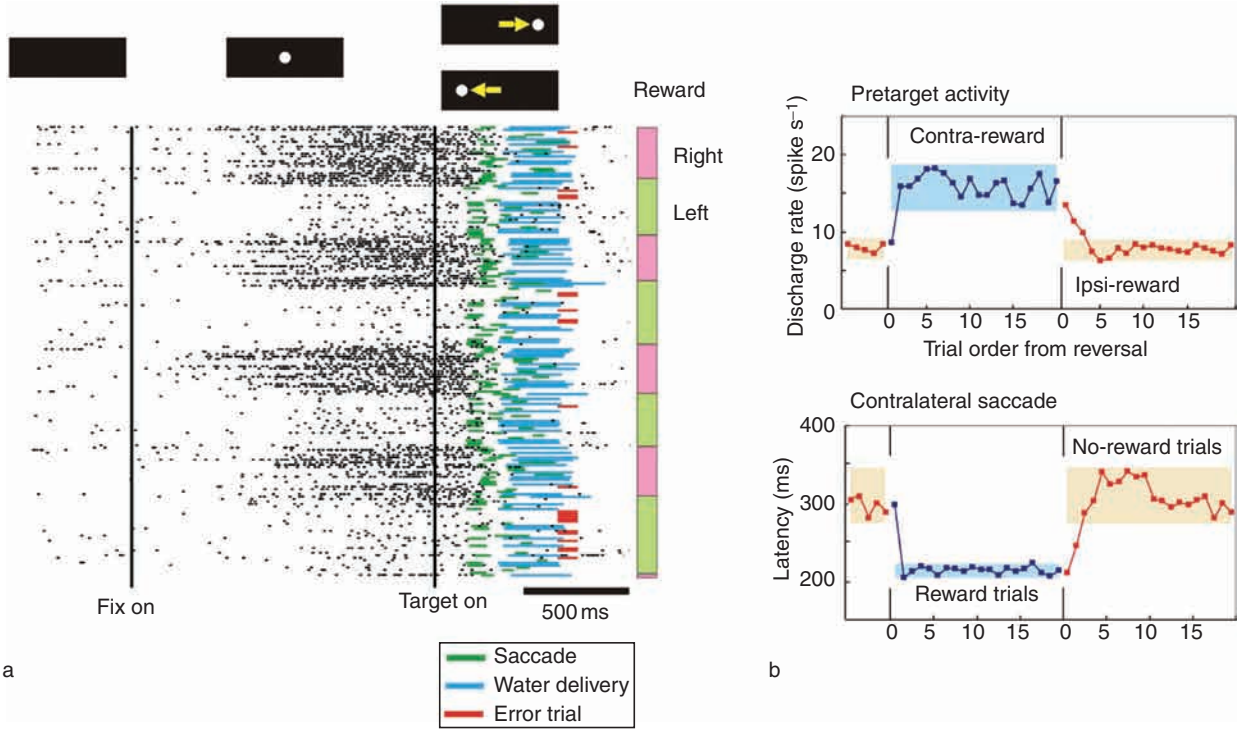

Figure 5 Pretarget activity of a caudate projection neuron selective for reward position. (a) Action potentials of a left caudate neuron (shown by dots in a raster display) while the monkey performed the visual-1DR task. The raster is aligned on the onset of the fixation point (fix on) and the onset of the target (target on). The changes in stimulus configuration are shown at top. After the fixation period of 1.5 s, a target came on at a right or left position pseudorandomly and the monkey had to make a saccade to it. In the first block of 20 trials (indicated by a pink bar), a reward (water) was given only if the target was presented on the right and the saccade was made to it. In the second block a reward was given only if the target was on the left (indicated by a green bar). These two types of blocks were repeated subsequently. The caudate neuron exhibited sustained activity before target onset while the monkey was fixating at the center spot, but the activity was much stronger when the reward was expected after saccade to the right target (which was contralateral to the neuron). (b) Correlated changes of pretarget activity of caudate neurons (top) and saccade latency (bottom). The mean pretarget activity of caudate neurons (top) was higher in the block of trials when the reward was expected on the contralateral position (indicated by blue dots) than on the ipsilateral position (red dots). The mean latency of contralateral saccades (bottom), which the caudate neurons would preferentially control, was shorter when they were preceded by higher pretarget activity and when they were followed by reward. From Lauwereyns J, Watanabe K, Coe B, and Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. *Nature* 418: 413–417.

How could dopamine influence activity of CD projection neurons? Dopamine does not exert fast excitatory or inhibitory actions, but is thought to modulate other synaptic inputs, especially glutamatergic inputs from the cerebral cortex. These findings led to the following hypothesis: the spatial signals from the cortical eye field are enhanced if dopaminergic inputs are increased (i.e., a bigger reward is expected) and depressed if dopaminergic inputs are decreased (i.e., a smaller reward is expected) (Figure 7). The interaction may occur within individual dendritic spines. This interaction might be mediated by long-term potentiation (LTP) or long-term depression (LTD). Recent studies have indicated that LTP indeed occurs if cortical inputs come in simultaneously with dopaminergic inputs and if the CD neuron fires. In support of this hypothesis, the reward-dependent bias in saccades is reduced if dopaminergic transmission in the CD is blocked by injecting dopamine D1 antagonist.

The role of interneuronal processing in the CD is less clear. TANs, which are thought to be cholinergic, respond to reward or reward predictor, similarly to dopaminergic neurons, but also respond to a sensory stimulus that predicts the absence of reward or punishment. It has been suggested that cholinergic interneurons indicate that the reward is not equal among actions to choose, but do not indicate which action is the best or the worst. The latter function would be carried out by dopaminergic neurons. Cholinergic interneurons might detect the condition in which rewards are unequal and guide dopaminergic neurons to fully operate. However, this hypothesis needs to be examined in future experiments.

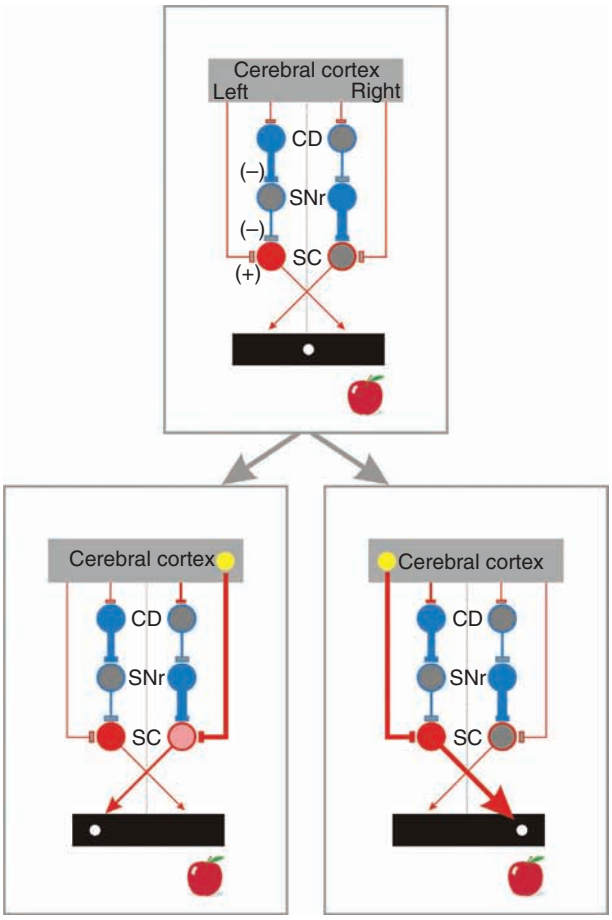

Figure 6 A hypothetical scheme showing how pretarget activity of caudate neurons might determine saccade bias toward reward position. The caudate–SNr–SC connections in the right and left hemispheres are shown to visualize the contrast in their excitability. Red circles and lines with (+): excitatory neurons. Blue circles and lines with (–): Inhibitory neurons. The thickness of the line represents the firing rate of the neuron. The color saturation of the circle represents the excitability of the neuron. (Top) Pretarget period. If a reward is expected at a right position (shown by an apple), left caudate neurons become active and right caudate neurons are inactive. This occurs while the monkey is fixating. On the left side of the brain, SNr neurons are inhibited and hence SC neurons are disinhibited; on the right side, SNr neurons remain active and hence SC neurons are inhibited. (Bottom) Posttarget period. (Bottom-left) If a target then appears on the left, cortical areas on the right side are activated (indicated by yellow) and this signal is sent to right SC neurons with excitatory connections. Since the right SC neurons have been tonically inhibited by SNr neurons, cortical inputs can only slowly activate the SC neurons and hence the saccade to the left (nonreward position) is delayed. (Bottom-right) If a target appears on the right, cortical areas on the left side are activated and this signal is sent to left SC neurons. Since the left SC neurons have been freed from the SNr-induced inhibition, cortical inputs can quickly activate the SC neurons and hence the saccade to the right (reward position) is quickened. From Hikosaka O, Nakamura K, and Nakahara H (2006) Basal ganglia orient eyes to reward. *Journal of Neurophysiology* 95: 567–584.

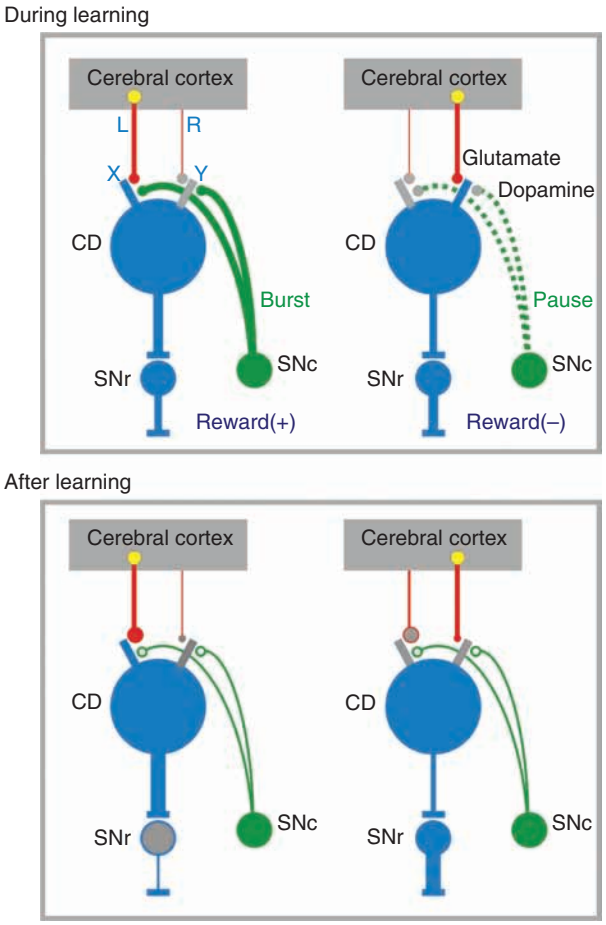

Figure 7 A hypothetical scheme showing reward-dependent plasticity of corticocaudate synapses. In one condition of the memory-1DR task (1) the caudate neuron receives two kinds of visual input (L and R) from the cerebral cortex at dendritic spines (X and Y), (2) input L represents a left visual cue signal while input R represents a right visual cue signal, (3) input L indicates an upcoming reward while input R indicates no reward, and (4) L and R inputs come in randomly one at a time. Dopamine neurons in the SNc, which are presumed to make synapses on caudate neurons nonselectively, respond to input L with a burst of spikes (top-left) and respond to input R with a pause of spikes (top-right). After several trials the efficacy of the L–X synapse increases (bottom, indicated by a larger red circle) while the efficacy of the R–Y synapse decreases (bottom, indicated by a smaller red circle). Activation of caudate neuron in response to the same input L is now enhanced (bottom-left, indicated by thick blue line), while activation in response to input R is diminished (bottom-right, indicated by thin light-blue line). The bias persists for some trials due to the plasticity even when reward is associated equally with R and L inputs (bottom, indicated by thin green lines). The bias will be reversed in the next block when input R, not input L, indicates an upcoming reward. Red lines: excitatory connections. Blue circles and lines: inhibitory neurons. Green circles and lines: dopamine neuron which is modulatory. The thickness of the line represents the firing rate of the neuron, except that the dotted green line indicates a decrease in dopamine neuron firing. From Hikosaka O, Nakamura K, and Nakahara H (2006) Basal ganglia orient eyes to reward. *Journal of Neurophysiology* 95: 567–584.

See also: Attention and Eye Movements; Basal Ganglia: Functional Models of Normal and Disease States; Cognition: Basal Ganglia Role; Eye and Head Movements; Oculomotor Control: Anatomical Pathways; Saccade–Pursuit Interactions; Saccades and Visual Search; Saccadic Eye Movements.

Further Reading

Alexander GE and Crutcher MD (1990) Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in Neurosciences* 13: 266–271.

Graybiel AM and Ragsdale CW (1979) Fiber connections of the basal ganglia. In: Cuenod M, Kreutzberg GW, and Bloom FE (eds.) *Development of Chemical Specificity of Neurons*, pp. 239–283. Amsterdam: Elsevier.

Hikosaka O, Nakamura K, and Nakahara H (2006) Basal ganglia orient eyes to reward. *Journal of Neurophysiology* 95: 567–584.

Hikosaka O, Sakamoto M, and Miyashita N (1993) Effects of caudate nucleus stimulation on substantia nigra cell activity in monkey. *Experimental Brain Research* 95: 457–472.

Hikosaka O, Takikawa Y, and Kawagoe R (2000) Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological Review* 80: 953–978.

Hikosaka O and Wurtz RH (1985) Modification of saccadic eye movements by GABA-related substances. II. Effects musicmol in monkey substantia nigra pars reticulata. *Journal of Neurophysiology* 53: 292–308.

Kawagoe R, Takikawa Y, and Hikosaka O (1998) Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience* 1: 411–416.

Kita H (1993) GABAergic circuits of the striatum. In: Arbuthnott GW and Emson PC (eds.) *Chemical Signalling in the Basal Ganglia*, pp. 51–72. Amsterdam: Elsevier.

Lauwereyns J, Watanabe K, Coe B, and Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. *Nature* 418: 413–417.

Leigh RJ and Zee DS (2006) *The Neurobiology of Eye Movements,* 4th edn., 763pp. New York: Oxford University Press.

Marín O, Smeets WJAJ, and González A (1998) Evolution of the basal ganglia in tetrapods: A new perspective based on recent studies in amphibians. *Trends in Neurosciences* 21: 487–494.

Miyashita N, Hikosaka O, and Kato M (1995) Visual hemineglect induced by unilateral striatal dopamine deficiency in monkeys. *Neuroreport* 6: 1257–1260.

Reynolds JN and Wickens JR (2002) Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks* 15: 507–521.

Schultz W (1998) Predictive reward signal of dopamine neurons. *Journal of Neurophysiology* 80: 1–27.

Basal Ganglia: Acetylcholine Interactions and Behavior

A Adler, M Joshoua, and H Bergman, Hebrew University, Jerusalem, Israel
J A Goldberg, BioControl Medical Ltd., Yehud, Israel
G Morris, Berlin Bernstein Center for Neural Computation and Charite Universtatmedizin, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction: Cholinergic Neuromodulation in the Striatum

The clinical symptoms of Parkinson's disease (PD) were first described at 1817 by the English physician James Parkinson in his 'Essay on the Shaking Palsy.' At this time physicians and scientists had no clues as to the etiology, pathophysiology, or therapeutic options for this common disease. The first effective medical treatment of PD – with belladonna, an anticholinergic agent – was accidentally discovered in the second half of the nineteenth century. For more than 100 years, anticholinergic agents were the only efficient treatment for PD, and therefore, despite their limited therapeutic effects, were frequently prescribed. The critical role of dopamine (DA) depletion in the striatum (input stage of the basal ganglia networks) in the pathophysiology of PD was revealed through a series of heroic pathological, pharmacological, and neurochemical studies throughout the twentieth century. This series of discoveries eventually ended in the seminal finding by Cotzias et al. of DA replacement therapy for PD. Thus, effective treatments of PD involve the manipulation of DA and acetylcholine (ACh) in an opposing fashion, either by elevation of the extracellular level of striatal DA or by reduction of the ACh level. These clinical opposing effects led to the DA–ACh balance hypothesis, which states that the two transmitters act antagonistically in the striatum. The superiority of DA replacement over anticholinergic treatment for PD probably led to the recent escalation in research on the basal ganglia DA system and the relative neglect of the basal ganglia cholinergic system. This article revisits the cholinergic system of the basal ganglia and discusses its role in the physiology of the basal ganglia networks as well as the pathophysiology of their related disorders (e.g., PD).

Chemical transmission in the central nervous system (CNS) is usually classified into two main categories. Fast neurotransmitters carry information from one neuron to another. Fast neurotransmitters can be excitatory, such as glutamate, increasing the likelihood that the postsynaptic cell will fire, or inhibitory, such as γ-aminobutyric acid (GABA) in the adult animal, reducing the likelihood of postsynaptic discharge. Fast neurotransmission is predominant in the specific connections between neurons and therefore is usually limited in space as well as duration. On the other hand, several brain stem nuclei (e.g., the locus coeruleus, the raphe nuclei, and the substantia nigra and ventral tegmental area) form diffuse modulatory projections throughout the CNS. The respective modulatory neurotransmitters (noradrenalin, DA, and serotonin) affect the excitability and modify the responsiveness of their target neurons; however, these neurotransmitters usually do not directly affect their target neurons' ongoing discharge. Unlike fast neurotransmission, neuromodulation is characterized by a broad span of effects in both the spatial and the temporal domains. In the CNS, ACh projections arise mainly from the nucleus basalis of Meynert (NBM) and the pedunculopontine nucleus (PPN). These diffuse cholinergic projections play a critical role in the modulation of information processing by neocortical neurons, as well as in many other neuronal structures (e.g., brain stem and spinal cord).

Given the DA–ACh balance, it is not surprising that the striatum – the input stage of the basal ganglia – has very rich cholinergic innervation. In fact, the striatum has the highest density of cholinergic markers – such as the enzymes choline acetyltransferase and acetylcholinesterase – of all CNS structures. Nevertheless, the source of this rich cholinergic innervation of the striatum is not extrinsic, that is, from the central cholinergic structures such as the NBM or the PPN. Rather, striatal ACh is secreted by striatal cholinergic interneurons, which constitute 1–2% of the striatal neuronal population. The cholinergic interneurons are large aspiny cells with extensive and dense local axons. A few dendrites arise from the soma and branch in a radiating fashion, for distances of up to 0.75 mm from the soma. In the rodent, the dendritic tree of a single cholinergic interneuron may span about a third of the striatum; however, the dendritic domain of the cholinergic neurons is smaller in the primate. The axons of the cholinergic neurons originate from the dendritic trunks and branch repetitively. The resulting axonal arborization consists of a dense plexus of fine axonal branches that fill and go beyond the region of the dendritic field. Acute mapping of cholinergic interneurons reveals that they tend to lie at or close to the striosome–matrix border, suggesting an integrative role for these striatal interneurons.

Physiological Studies of Striatal Cholinergic Cells

Extracellular recording of the spiking activity of striatal neurons in behaving animals reveals two main populations of neurons. Phasically active neurons are characterized by a short-duration action potential, very low spontaneous activity (<1 spike s^{-1}), and a phasic increase triggered by movements and other relevant behavioral events. These neurons are believed to be the spiny projection cells of the striatum (95% of all striatal cells). Tonically active neurons (TANs) have a broader action potential and a spontaneous tonic irregular firing rate of 3–12 Hz and show strong and robust responses to reward related events. Recent studies indicate that TANs are the cholinergic interneurons of the striatum. Electrophysiological studies have confirmed that TANs cannot be antidromically activated by stimulation of the globus pallidus, thus supporting the identification of TANs as striatal interneurons.

Intracellular studies have shown that the striatal cholinergic interneurons discharge autonomously even under blockade of a variety of fast synaptic transmission and neuromodulators. The striatal cholinergic interneurons are characterized by a prolonged afterhyperpolarization (AHP) following an action potential and are highly sensitive to incoming excitatory inputs: Depolarization induced by only two or three synaptic potentials can drive them to fire action potentials. The sensitivity of cholinergic cells to synaptic excitation is due to the relatively depolarized membrane potential and high input resistance of these neurons. In brain slices, these neurons display a variety of firing patterns, including rhythmic single spiking, rhythmic bursting, and irregular firing. The persistent sodium current is responsible for destabilizing the resting potential of the neuron, and the hyperpolarization-activated, cyclic nucleotide-gated, inward cation current is responsible for depolarizing the cell from the variety of AHP states that the cell exhibits. The first of these is a medium-duration AHP (mAHP, ~200 ms) that follows each action potential during rhythmic single spiking The reduced input resistance during the mAHP diminishes the efficacy of depolarizing synaptic inputs and therefore places a ceiling on the firing rate of the TANs. Calcium entry via N-type voltage-activated calcium channels during the action potential is required for activation of the small-conductance, calcium-dependent potassium current which underlies the mAHP. The mAHP is essential for the tonic firing pattern of the TANs because pharmacological treatments which reduce the amplitude of the mAHP convert single (regularly or irregularly) spiking cells into burst-firing neurons. Activation of D1 DA receptors prolongs the TANs' interspike interval by enhancing the mAHP amplitude. The second and even longer (several hundred milliseconds to seconds) AHP of the cholinergic interneuron is termed the sAHP. Calcium entry via the L-type voltage-activated calcium channels triggers calcium release from intracellular stores, which in turn activates the current underlying the sAHP. It is the buildup of the sAHP that terminates the discharge during rhythmic bursting of these neurons. Finally, cholinergic interneurons exhibit an additional calcium-independent regenerative AHP because of the inwardly rectifying potassium current, which can induce all-or-none hyperpolarizations whose time course is mostly independent of the size of the stimulus that generated it. The sAHP and regenerative hyperpolarization are likely to participate in fashioning the pause response of the TANs.

TAN Encoding of Behavioral Events

The behaviorally triggered changes in the discharge of TANs, unlike the diversity of task-related activations exhibited by the phasically active population of striatal neurons, is stereotypic in terms of both the trigger events and the pattern of change. Many studies of these neurons have emphasized the homogeneity of the TAN responses across large parts and different territories of the striatum. Other studies have revealed that as in any biological system, not all TANs are born equal. However, compared with neurons of other neuronal structures (e.g., cortical, pallidal, and nigral cells) and even with the striatal projection neurons, the TAN population is very homogeneous. The next discussion outlines the main characteristics of this unique population of cholinergic interneurons, as well as its minor variants.

The behavioral events that tend to modify the tonic background discharge of TANs include primary rewards and stimuli that have acquired a reward value during associative learning. However, as with DA neurons, the TANs do not encode the hedonic value of the rewarding event; rather their responses are modulated by predictions about stimulus timing. In classical conditioning tasks, the responses are diminished or abolished in the presence of external cues that predict the time at which the rewarding events will occur. The critical role of the TANs in reward-related learning has been confirmed by the impairment of procedural learning by selective ablation with immunotoxin-mediated cell targeting of cholinergic neurons in the striatum.

The typical response of TANs involves a transient depression of their tonic firing. This depression is often followed by rebound excitation and may be preceded (or flanked) by increases in discharge rate. The pause in firing usually occurs 100 ms after the

behavioral trigger. The mean duration of the pause responses (~250 ms) is greater than the mean interspike interval of the same neurons (~180 ms, range 60–330 ms), suggesting that the pause represents an active suppression or modified intrinsic dynamics of TAN firing rather than a resetting of the TANs' ongoing discharge. Recent studies have revealed different response patterns for aversive events and reward-related events, as well as for putamen and caudate TANs. The functional significance of these variations is still under study. Bearing in mind the DA–ACh balance hypothesis and the therapeutic effects of anticholinergic agents in PD, it is not surprising that the pause of the TAN parallels the burst discharge of DA cells.

Network Properties of TANs

The strong resemblance across the temporal patterns of neuronal responses of widely distributed TANs led to the suggestion that the TANs form a distributed network with synchronized activity. Cross-correlation studies of simultaneously recorded TANs have revealed a significant degree of synchronized activity between TANs recorded by different electrodes, at distances of a few millimeters, for example.

In the normal monkey, only very few of simultaneously recorded TANs and pallidal neurons display significant synchronization. After treatment with 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine, many of these pairs display significant synchronous periodic oscillations. The low level of correlated activity in the normal state demonstrates that TANs have only a slight effect on pallidal activity during execution of familiar behavior. The high level of oscillatory correlated activity in the parkinsonian state further suggests that coherent oscillations of the whole basal ganglia circuitry underlie the clinical features of Parkinson's disease.

The Functional Organization of TANs' Synaptic Inputs

The observation that populations of TANs can exhibit synchronous pause responses, as well as their background synchronized activity, suggests that there may be a common synaptic or neuromodulatory input to the TANs which might underlie such synchronization. The major sources of inputs to the TANs include the cortex, the central complex of the thalamus (the centromedian (CM) nucleus), and midbrain dopaminergic neurons, as well as lateral GABAergic synapses arising either from GABAergic interneurons or from collaterals of the medium-spiny projection neurons of the striatum.

Neurochemical studies have documented inhibition of striatal ACh release by the nigrostriatal DA pathway. These neurochemical studies, as well as the synchronization of the DA burst with the TAN pauses, may suggest that DA is the main driver (or the brakes) of the TANs. Indeed, lesions of the dopaminergic projection to the striatum or local application of dopaminergic antagonists results in a loss of the pause response. However, these effects can be reversed by apomorphine (a postsynaptic DA agonist that therefore restores only the background level of DA), indicating that DA may gate, but not drive, the TAN pause. Recent *in vivo* intracellular recordings have revealed that high-frequency stimulation of the substantia nigra induces potentiation of the cortically evoked excitation of striatal cholinergic interneurons and prolongs the AHP of these neurons after cortical stimulation. These findings suggest that a substantia nigra-induced AHP produces cortical stimulus-associated firing pauses in cholinergic interneurons. On the other hand, inactivation of the CM complex in primates also results in a loss of the TAN pause response. Thus, dopaminergic inputs to the TANs probably have a gating function on the pause which follows the thalamic and/or cortical excitatory inputs. However, the hypothesis that the TAN pause is generated by a mechanism independent of a previous excitation, for instance by the lateral GABAergic connections in the striatum, has not yet been explored.

TANs' Effects in the Striatum

The release of ACh in the striatum affects the striatal neuronal elements through both muscarinic and nicotinic receptors. ACh secreted by the TANs can affect the excitability of striatal projections as well as the efficacy of corticostriatal synapses. These effects are mainly mediated through muscarinic receptors. Neurochemical studies have reported a reciprocal inhibition of striatal ACh release by the nigrostriatal DA pathway and of DA release by ACh. Nicotinic receptors are predominantly located at the synaptic terminals of the dopaminergic axons. Recent studies have shown that the effect of activation of nicotinic receptors is dependent on the discharge frequency of the DA neuron. Activation of nicotinic receptors enhances the probability of DA release by a single DA spike; however, it enhances short-term depression of DA release sites. Reducing ACh activity (e.g., during a pause) suppresses DA release by a single DA spike but facilitates the DA release by a burst of DA spikes.

Summary and Conclusions

This article has summarized recent advances in knowledge of the critical role of basal ganglia ACh in behavioral control, as well as its role in basal ganglia-related

disorders, such as PD. The clinical literature and practice emphasize the role of DA in the physiology and pathophysiology of the basal ganglia. However, the opposing relationships between the typical responses of DA and striatal cholinergic neurons, between the neurochemical levels of DA and ACh in the striatum, between their effects on corticostriatal transmission, and finally between their therapeutic actions, suggest that the striatal ACh–DA balance (e.g., the DA burst vs. the TAN pause) is a key process in the computational physiology of the basal ganglia. The TAN pauses are probably triggered either by feed-forward inhibition via the striatal fast-spiking GABAergic interneurons or by supra- or subthreshold excitation from the cortex or thalamic inputs to the cholinergic interneurons that induces a rebound AHP in these neurons. Many physiological studies indicate that the pause of the cholinergic interneurons informs the projection neuron of the striatum of salient events which deserve a shift of attention, and may facilitate learning at the cellular level. The functional role of the increases in discharge of the TANs is still being debated and should be clarified in the near future.

See also: Acetylcholine Neurotransmission in CNS; Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia: Evolution; Basal Ganglia and Oculomotor Control; Basal Ganglia: Motor Functions; Basal Ganglia: Internal Organization; Basal Ganglia: Habit; Cognition: Basal Ganglia Role; MPTP Parkinsonism Model; Parkinsonian Syndromes; Procedural Learning: Striatum; Sleep–Wake State Regulation by Acetylcholine; Striatum: Internal Physiology.

Further Reading

Aosaki T, Graybiel AM, and Kimura M (1994) Effect of the nigrostriatal dopamine system on acquired neural responses in the striatum of behaving monkeys. *Science* 265: 412–415.

Apicella P (2002) Tonically active neurons in the primate striatum and their role in the processing of information about motivationally relevant events. *European Journal of Neuroscience* 16: 2017–2026.

Calabresi P, Centonze D, Gubellini P, Pisani A, and Bernardi G (2000) Acetylcholine-mediated modulation of striatal function. *Trends in Neurosciences* 23: 120–126.

Cragg SJ (2006) Meaningful silences: How dopamine listens to the ACh pause. *Trends in Neurosciences* 29: 125–131.

Doya K (2002) Metalearning and neuromodulation. *Neural Networks* 15: 495–506.

Goldberg JA and Wilson CJ (2005) Control of spontaneous firing patterns by the selective coupling of calcium currents to calcium-activated potassium currents in striatal cholinergic interneurons. *Journal of Neuroscience* 25: 10230–10238.

Graybiel AM, Aosaki T, Flaherty AW, and Kimura M (1994) The basal ganglia and adaptive motor control. *Science* 265: 1826–1831.

Kitabatake Y, Hikida T, Watanabe D, Pastan I, and Nakanishi S (2003) Impairment of reward-related learning by cholinergic cell ablation in the striatum. *Proceedings of the National Academy of Sciences of the United States of America* 100: 7965–7970.

Morris G, Arkadir D, Nevet A, Vaadia E, and Bergman H (2004) Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron* 43: 133–143.

Raz A, Feingold A, Zelanskaya V, Vaadia E, and Bergman H (1996) Neuronal synchronization of tonically active neurons in the striatum of normal and parkinsonian primates. *Journal of Neurophysiology* 76: 2083–2088.

Reynolds JN, Hyland BI, and Wickens JR (2004) Modulation of an afterhyperpolarization by the substantia nigra induces pauses in the tonic firing of striatal cholinergic interneurons. *Journal of Neuroscience* 24: 9870–9877.

Shimo Y and Hikosaka O (2001) Role of tonically active neurons in primate caudate in reward-oriented saccadic eye movement. *Journal of Neuroscience* 21: 7804–7814.

Wang Z, Kai L, Day M, et al. (2006) Dopaminergic control of corticostriatal long-term synaptic depression in medium spiny neurons is mediated by cholinergic interneurons. *Neuron* 50: 443–452.

Wilson CJ, Chang HT, and Kitai ST (1990) Firing patterns and synaptic potentials of identified giant aspiny interneurons in the rat neostriatum. *Journal of Neuroscience* 10: 508–519.

Zhang H and Sulzer D (2004) Frequency-dependent modulation of dopamine release by nicotine. *Nature Neuroscience* 7: 581–582.

Basal Ganglia: Evolution

L Medina, Universitat de Lleida, IRBLLEIDA, Catalonia, Spain

© 2009 Elsevier Ltd. All rights reserved.

Basal Ganglia Divisions and Organization in Mammals

The basal ganglia (BG) are a group of nuclei located in the basal (subpallial) part of the telencephalon and involved in control of motor behavior (including planning and execution of movement), as well as in cognitive functions such as motivation, attention, and learning. The BG were classically involved in motor control based on the obvious movement disorders associated with BG lesions or abnormal activity. BG control of movement was thought to be done mainly by a pathway to the motor cortex by way of the thalamus. Evidence accumulated during the past 15 years indicates that the BG are closely associated not only with motor and premotor areas of the cortex but also with many other areas of the cortex/pallium, including the prefrontal, orbitofrontal, and cingulate cortices, which are thought to be important for controlling many aspects of cognitive functions such as planning, working memory, sequential behavior, visuospatial perception, and attention. Importantly, the BG appear to play an important role in selecting which action to perform, which depends on learning and experience, and participate in adaptive control of motor and cognitive behaviors. Moreover, the BG are involved in reward and anticipation of forthcoming reward, and the pallial amygdalar input to BG (involved in assigning biological/emotional significance to stimuli) appears to play a key role.

The BG of mammals consist of an input station called striatum and an output station called pallidum (Figure 1). Briefly, the striatum receives sensorimotor, cognitive, and viscerolimbic/emotional information primarily from most areas of the cerebral cortex/pallium (although it also receives input from the thalamus) and transmits this information to the pallidum. The pallidum in turn projects to effector centers in the thalamus and brain stem, thus acting on thalamocortical or descending pathways and influencing motor or cognitive behaviors. Therefore, the BG can be considered as an interface that connects the cerebral cortex with brain stem neural systems that effect specific aspects of behavior. The activity of principal striatal neurons is modulated by input from dopaminergic cells of the mesodiencephalic tegmentum (including the substantia nigra and ventral tegmental area) and by local circuit neurons of the striatum. When the dopaminergic input to the striatum fails, motor and cognitive dysfunctions are observed, producing parkinsonian symptoms, as well as cognitive and emotional deficits. In order to determine whether the BG exist in nonmammalian vertebrates, and whether they show the same (homologous) divisions and cell types and similar neurochemistry, connections, and functions, it is important to briefly consider the basic characteristics of mammalian BG.

Striatum and Pallidum Develop from Distinct Histogenetic Subdivisions of the Subpallium

The two major divisions of the BG originate in different sectors of the subpallium under the influence of distinct combinations of transcription factors (Figure 2). The striatum develops from the lateral ganglionic eminence, characterized by expression of the transcription factors Dlx1/2/5/6, Ghs1/2, and Mash1, among others, but not Nkx2.1/Titf-1. On the contrary, the pallidum develops from the medial ganglionic eminence, characterized by expression of Dlx1/2, Nkx2.1/Titf-1, Lhx6, and Lhx7/8, among others. It appears that Gsh1/2 and Nkx2.1/Titf-1 are important for specification of lateral or medial ganglionic eminences, respectively (LGE or MGE; also called striatal or pallidal subdivisions), whereas Dlx, Mash, and Lhx transcription factors play a role in the differentiation of neurons. Lack of function of Gsh1/2 produces severe malformation of the striatum, which is transformed into pallium/cortex. Lack of function of Nkx2.1 (as in Nkx2.1 knockout mice) produces severe malformation of the pallidum and transformation of the MGE into LGE. On the other hand, Dlx and Mash transcription factors appear to be involved in the acquisition of the GABAergic phenotype, which is typical of BG neurons, as well as their projections.

Dorsolateral versus Ventromedial Parts of the Basal Ganglia: A Gradient from Motor to Cognitive and to Motivational/Emotional Functions

The BG contain two main divisions: (1) a dorsal part mainly involved in motor and cognitive/executive functions and (2) a ventral part involved in cognitive and limbic/emotional functions such as motivation and attention. The division between the dorsal and the ventral parts of the BG is not sharp but, rather, gradiental from dorsolateral to ventromedial. Thus, from dorsolateral to ventromedial parts of the striatum, there is a gradient from motor to executive/cognitive to motivational/emotional functions.

Each BG part consists of a striatal portion and an associated pallidal portion (Figure 1). Thus, there is a

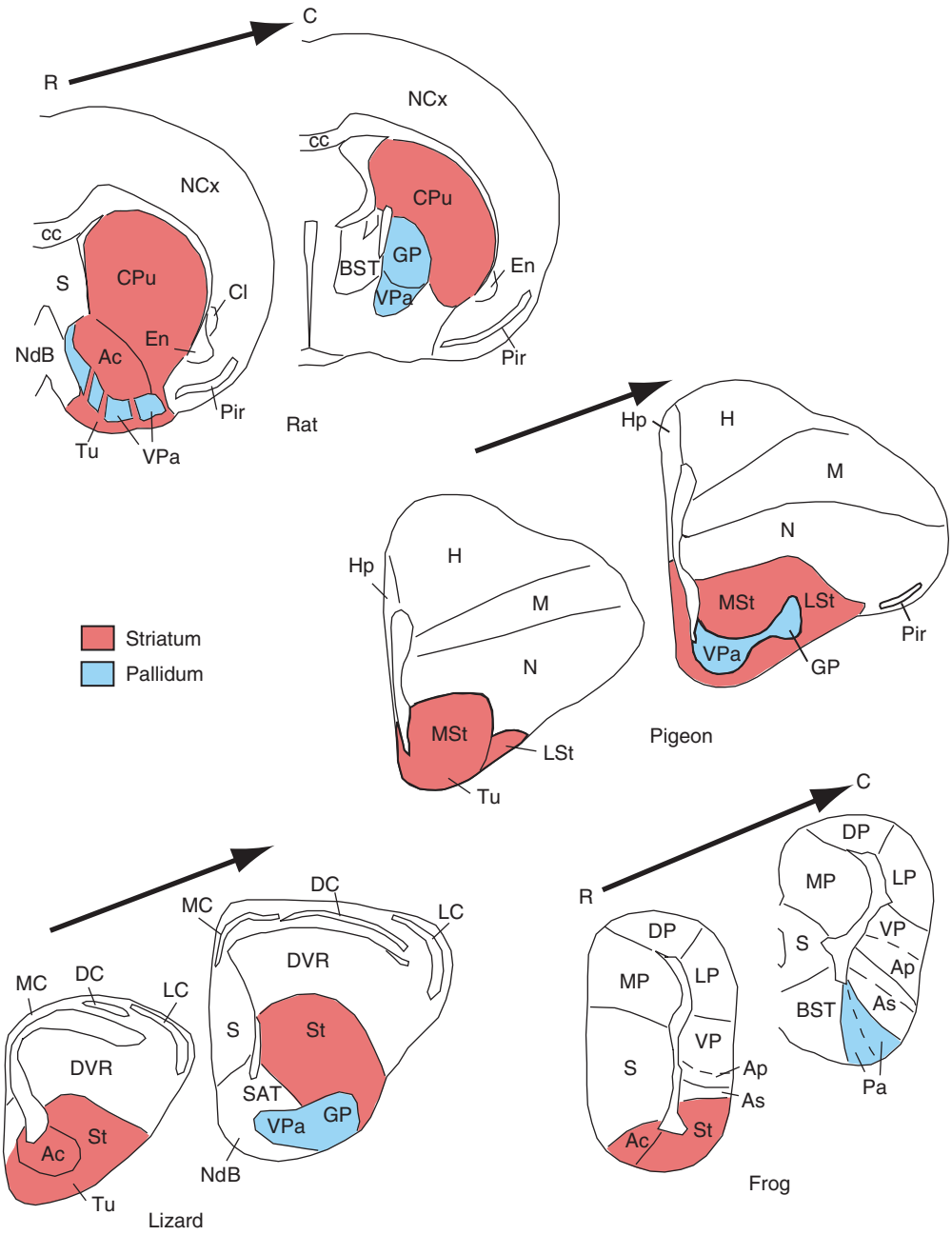

Figure 1 Schematic drawings of frontal sections through the telencephalon of a mammal (rat), a bird (pigeon), a reptile (lizard), and an amphibian (frog) showing the striatal (red) and pallidal (blue) subdivisions of the basal ganglia. Two different rostrocaudal levels are shown for each species. Ac, nucleus accumbens; Ap, pallial amygdala; As, subpallial amygdala; BST, bed nucleus of the stria terminalis; cc, corpus callosum; Cl, claustrum; CPu, caudoputamen complex (dorsal striatum); DC, dorsal cortex (reptile); DP, dorsal pallium; DVR, dorsal ventricular ridge; En, endopiriform nucleus; GP, globus pallidus; H, hyperpallium (dorsal pallium in birds); Hp, hippocampal formation; LC, lateral cortex (reptiles); LP, lateral pallium; LSt, lateral striatum (birds); M, mesopallium (lateral pallium of birds); MC, medial cortex (reptile); MP, medial pallium; MSt, medial striatum (birds); N, nidopallium (part of ventral pallium in birds); NdB, nucleus of the diagonal band (Broca); NCx, neocortex (dorsal pallium of mammals); Pa, pallidum; Pir, piriform cortex; S, septum; SAT, striato-amygdaloid transition area; St, striatum; Tu, olfactory tubercle; VP, ventral pallium; VPa, ventral pallidum.

dorsal striatum that includes the caudate nucleus and putamen in primates and carnivores (artificially separated by the fibers of the internal capsule) or simply the caudoputamen complex in rodents (CPu). The CPu projects to the globus pallidus (GP), which represents the dorsal part of the pallidum (Figure 3). In mammals, the GP shows two subdivisions called internal and external segments of the globus pallidus

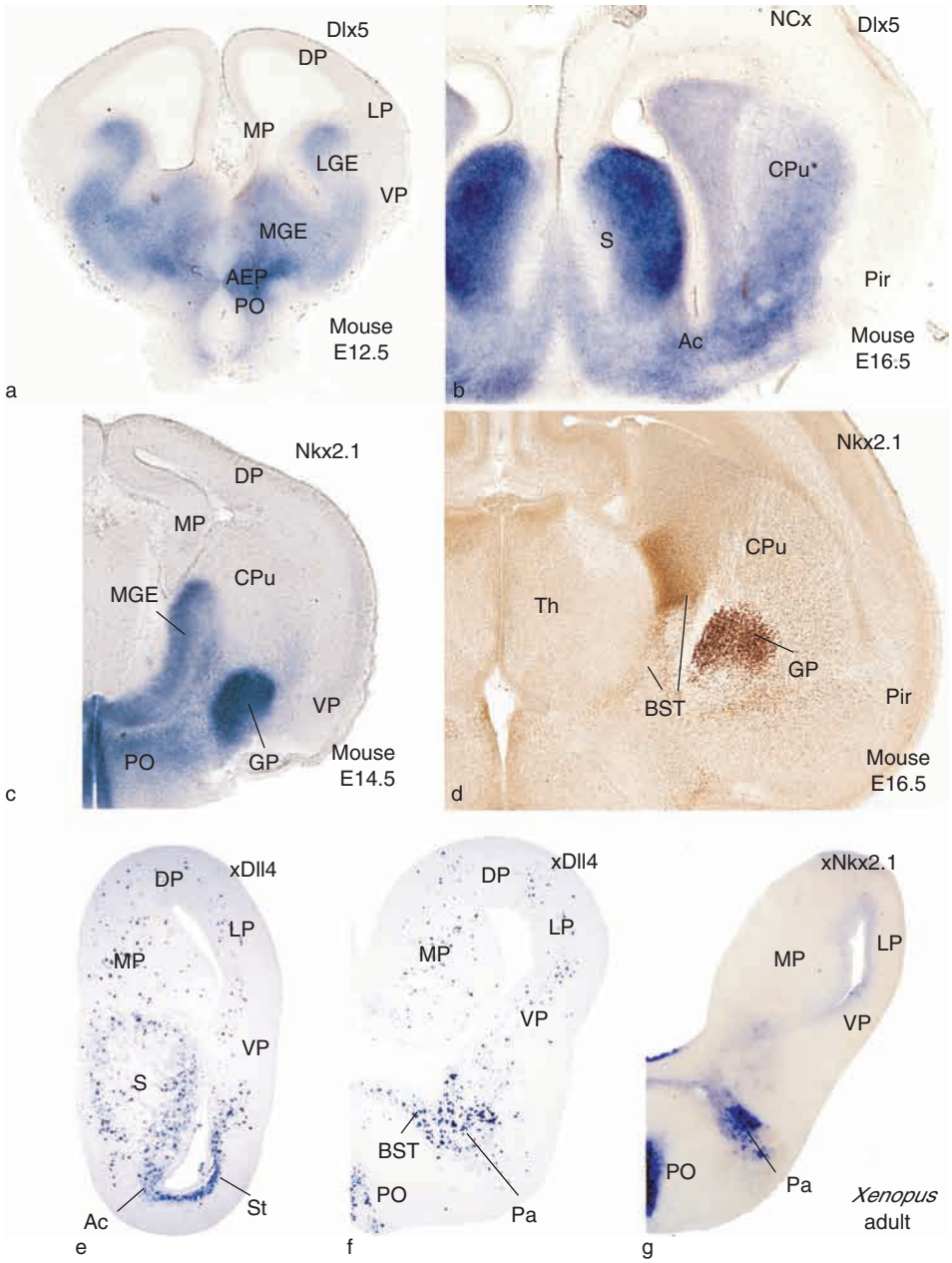

Figure 2 (a–d) Digital images of frontal sections through the embryonic telencephalon of a mouse showing the expression of the transcription factors Dlx5 (a, b) or Nkx2.1 (c, d). Dlx5 is expressed in the entire subpallium, including the lateral and medial ganglionic eminences (LGE and MGE, respectively), most of the septum, anterior peduncular area (AEP), and part of the preoptic area (PO). Nkx2.1 is expressed only in part of the subpallium, in a domain that includes MGE (which gives rise to the globus pallidus), a pallidal part of the septum, AEP, and PO. In contrast, the LGE and its striatal derivatives (including the caudoputamen complex and nucleus accumbens) do not show nested expression of Nkx2.1. Nevertheless, the striatum contains dispersed cells expressing Nkx2.1, but these are immigrant interneurons that originate in the Nkx2.1-expressing domain. (e–g) Digital images of frontal sections through the telencephalon of the frog *Xenopus laevis* (adult animals) showing the expression of Distal-less-4 (xDll4, ortholog of Dlx2 of mouse (e) or xNkx2.1 (f and g). The expression patterns of these transcription factors in the telencephalon of *Xenopus* are comparable to those of the ortholog genes in mouse. The subpallium is rich in cells expressing xDll4 and includes most of the septum and the striatal and pallidal parts of the basal ganglia. The pallidal subdivision (but not the striatum) also expresses xNkx2.1. These data strongly support the homology of the striatal and pallidal parts of the basal ganglia between *Xenopus* and mouse. Ac, nucleus accumbens; BST, bed nucleus of the stria terminalis; CPu, caudoputamen complex (dorsal striatum); DP, dorsal pallium; GP, globus pallidus; LP, lateral pallium; MP, medial pallium; Pa, pallidum; Pir, piriform cortex; S, septum; St, striatum; Th, thalamus (dorsal thalamus); VP, ventral pallium.

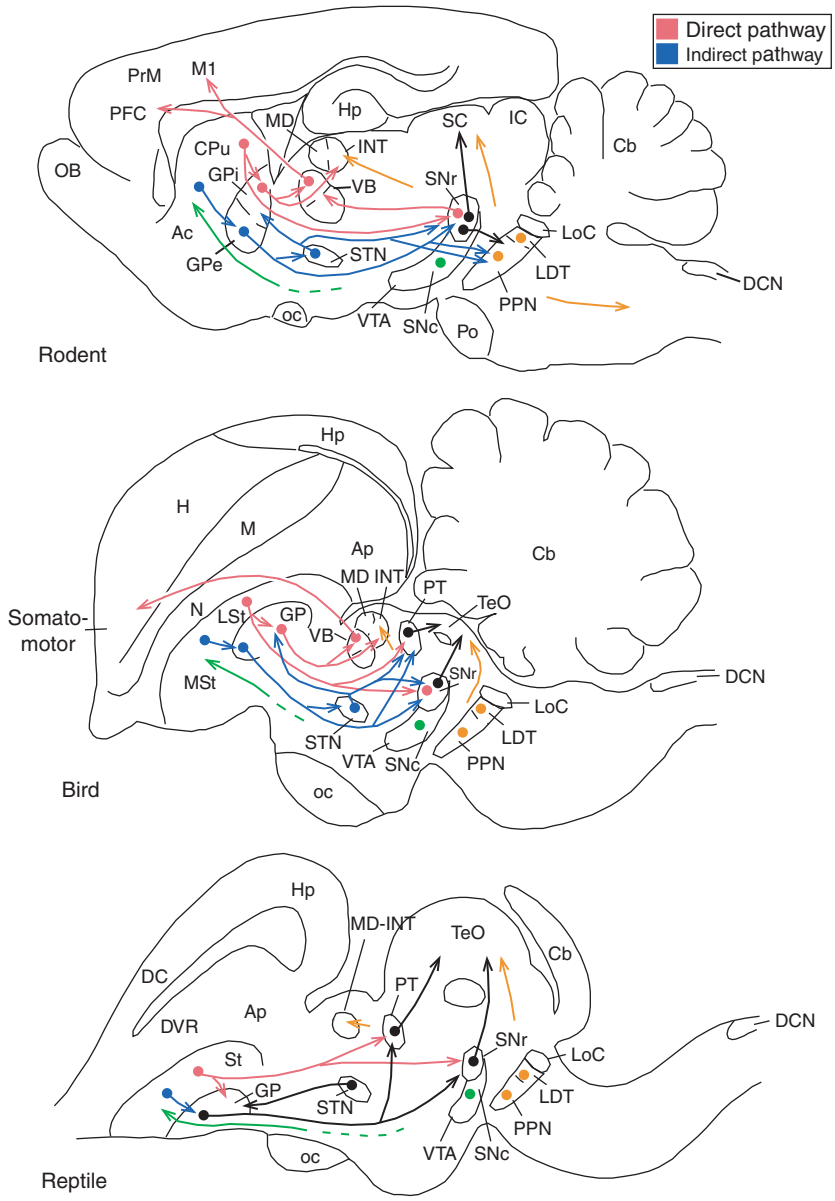

Figure 3 Schematic drawings of lateral brain views of a mammal (rodent), a bird (pigeon), and a reptile (lizard) showing the main connections of the dorsal (somatomotor) basal ganglia (BG). Note the existence of similar BG pathways to the tectum (superior colliculus) by way of the substantia nigra pars reticulata (SNr) in all groups. In addition, reptiles and birds show BG pathways to the tectum by way of the pretectum. In all groups, the activity of the dorsal striatum is modulated by dopaminergic input from the substantia nigra pars compacta (SNc; the input is represented in green). These features possibly characterized the BG of ancestral amniotes. Moreover, dorsal BG pathways to the neocortex/dorsal pallium are observed in birds and especially in mammals. These BG–cortical/pallial channels include a direct pathway (red) and an indirect pathway (blue), and the latter involves a glutamatergic center called subthalamic nucleus (STN). The dorsal pallidum does not project to the thalamus (dorsal) in reptiles, suggesting that dorsal BG–neocortical/dorsal pallial loops evolved independently in birds and mammals. In mammals, the cholinergic pedunculopontine tegmental nucleus (PPN) receives dorsal BG input and is reciprocally connected with the SNr. By way of this nucleus, the BG are able to influence descending motor pathways (via the tectum/colliculus and medullary groups) and ascending activating pathways (via midline/intralaminar thalamic nuclei) that are involved in arousal and awareness (the cholinergic pathways are depicted in orange). The PPN is present in birds and reptiles, and apparently it gives rise to ascending activating pathways via the midline/intralaminar thalamic nuclei and to descending pathways via the tectum. However, the PPN of birds and reptiles appears to receive input mainly from the ventral BG and little or no input from dorsal BG. It is unknown whether a reciprocal connection between PPN and SNr exists in birds or reptiles. Ac, nucleus accumbens; Ap, pallial amygdala; Cb, cerebellum; CPu, caudoputamen complex (dorsal striatum); DC, dorsal cortex (reptile); DCN, dorsal column nuclei; DVR, dorsal ventricular ridge; GP, globus pallidus; GPe, globus pallidus, external or lateral segment; GPi, globus pallidus, internal or medial segment; H, hyperpallium (dorsal pallium in birds); Hp, hippocampal formation; IC, inferior colliculus; INT, intralaminar thalamic nuclei; LDT, laterodorsal tegmental nucleus; LoC, locus coeruleus; LSt, lateral striatum (birds); M1, primary motor cortex; M, mesopallium (lateral

(GPi and GPe, respectively). On the other hand, there is a ventral striatum that includes the nucleus accumbens and olfactory tubercle (its striatal part) and also an associated ventral pallidum (Figures 1 and 3).

The general features of the dorsal and ventral parts of the BG are similar. For example, the striatum typically shows a high acetylcholinesterase activity and is densely innervated by dopaminergic axons, which distinguish it from the adjacent telencephalic areas (i.e., the cortex/pallium and pallidum). However, the connections of the dorsal and ventral BG parts, although following a similar pattern, are different. Thus, whereas the CPu receives its main cortical input from widespread areas of the neocortex, including motor, premotor, and prefrontal areas, the nuclear accumbens (Acc) receives its main cortical input from limbic-related cortical/pallial areas, such as the hippocampal formation and the pallial amygdala (basolateral complex), as well as from the insular/perirhinal cortex and from parts of the prefrontal and cingulate cortex. In relation to the input from the prefrontal cortex, whereas the projections to the dorsal striatum originate in the areas involved in selective attention, working memory, decision making, and goal-directed behavior (dorsolateral prefrontal cortex of primates and prelimbic cortex of rodents), the projections to the ventral striatum primarily arise in areas involved in visceromotor and emotional behavior (orbitomedial cortex of primates and infralimbic cortex of rodents). Similarly, whereas the dorsal striatum is mainly modulated by dopaminergic input from the substantia nigra pars compacta (SNc; A9 catecholamine cell group), the ventral striatum is mainly modulated by dopaminergic input from the ventral tegmental area (A10 catecholamine cell group), SNc (A9 group), and retrorubral field (A8 group) and by noradrenergic input from the locus coeruleus (Figures 3 and 4). On the other hand, whereas the projections of the globus pallidus end on the motor, lateral part of the thalamus (ventral anterior/ventral lateral nuclei) and thereby influence motor/premotor, prefrontal, and other parts of the neocortex, the projections of the ventral pallidum target the medial (midline) thalamus (mediodorsal nucleus) which in turn projects to the prefrontal cortex (primarily the orbitomedial cortex of primates and infralimbic cortex in rodents). This part of the prefrontal cortex receives input from the hippocampal formation and pallial amygdala (basolateral complex), which are important for motivational/emotional functions (including assigning an emotional significance to stimuli).

However, there is no sharp boundary between dorsal (dorsolateral) and ventral (ventromedial) BG. In fact, the connectivity patterns of dorsal versus ventral BG change following a gradient from dorsolateral to ventromedial, as also noted previously for the functions. For example, the ventromedial part of CPu shows some connections similar to those of the adjacent core part of nucleus accumbens, such as input from the amygdala, which is typical of the limbic, ventromedial striatum.

Cell Types, Neurotransmitter Organization, and Connections

Both the striatum and the pallidum are characterized by having GABAergic neurons that show inhibitory, descending projections. In the striatum, the vast majority (approximately 90–95%) of the neurons are medium-sized, spiny neurons that project to the pallidum or brain stem. There are two major types of principal or projection neurons in the striatum: those that contain GABA and the neuropeptides substance P and dynorphin (GABA+/SP+/DYN+) and those that contain GABA and the neuropeptide enkephalin (GABA+/ENK+). Both cell populations are highly segregated in the dorsal part of the striatum (each type representing half of the total amount of projection neurons, with only 4% of co-localization) but less so in the ventral striatum. In the dorsal striatum, the two major cell types project to different targets: GABA+/SP+/DYN+ striatal neurons primarily project to the GPi (and only show a minor projection to the external segment) and to the substantia nigra (pars compacta and pars reticulata), whereas GABA+/ENK+ striatal neurons project heavily to the GPe.

In addition to projection neurons, the striatum also contains four subtypes of local circuit neurons (interneurons), which represent approximately 5–10% of all striatal neurons and modulate the activity of the principal striatal neurons. The four interneuron subtypes described in the striatum are (1) cholinergic interneurons; (2) interneurons containing GABA and parvalbumin; (3) interneurons containing GABA, nitric oxide, and the neuropeptides somatostatin and neuropeptide Y; and (4) interneurons containing

pallium of birds); MD, mediodorsal thalamic nucleus (midline thalamus); MSt, medial striatum (birds); N, nidopallium (part of ventral pallium in birds); OB, olfactory bulb; oc, optic chiasm; PFC, prefrontal cortex; Po, preoptic area; PrM, premotor cortex; PT, pretectal nucleus of the posterior commissure (it receives BG input); SC, superior colliculus; St, striatum; TeO, optic tectum; VB, ventrobasal thalamus; VTA, ventral tegmental area (A10 catecholamine cell group).

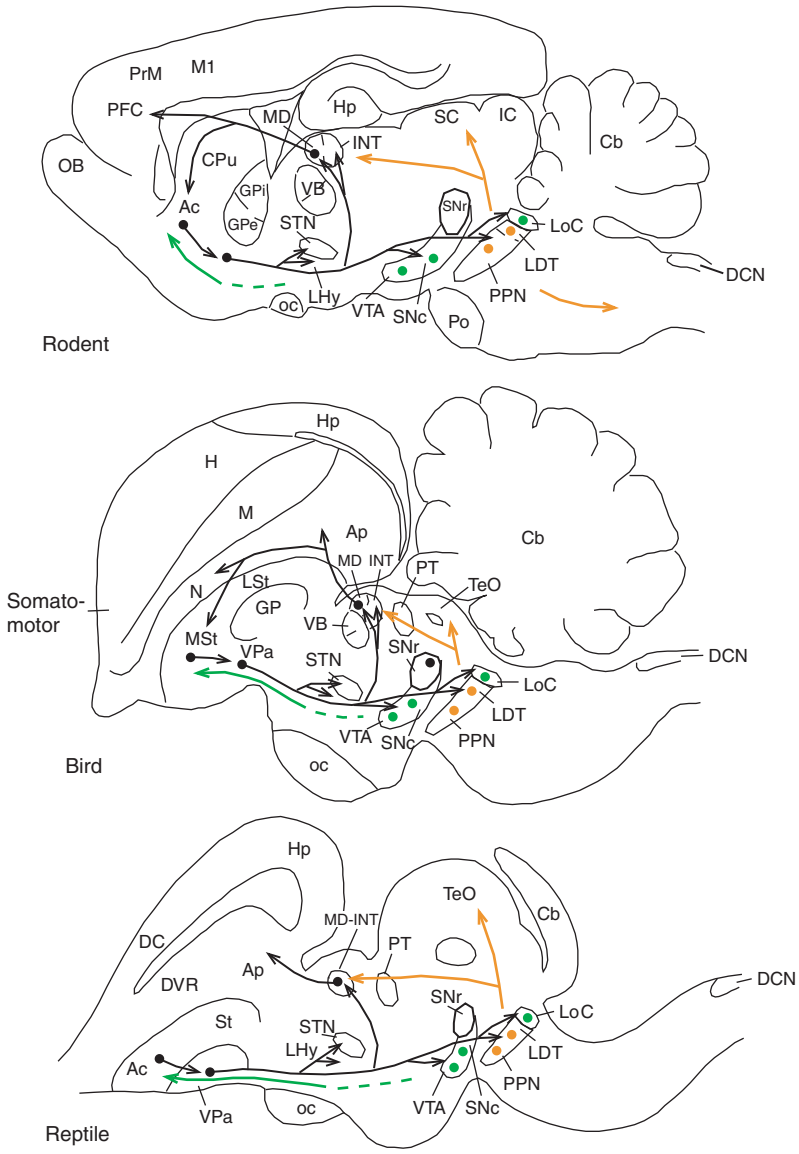

Figure 4 Schematic drawings of lateral brain views of a mammal (rodent), a bird (pigeon), and a reptile (lizard) showing the main connections of the ventral (viscerolimbic) basal ganglia (BG). Note the existence in all groups of similar ventral BG projections to the lateral hypothalamus, habenula (not shown), ventral tegmental area (VTA; A10 cell group), sustantia nigra pars compacta (SNc; A9 cell group), pedunculopontine tegmental nucleus (PPN), and locus coeruleus (LoC). In addition, the ventral BG projects to comparable midline/intralaminar thalamic nuclei, which in turn project back to the ventral BG in all groups. In birds and reptiles, the midline thalamus also projects to the ventrolateral pallium, including the pallial amygdala. These were possibly basic features for ventral BG pathways in ancestral amniotes. In mammals, the midline thalamic nucleus that receives ventral BG input shows a projection to the prefrontal cortex, and this appears to represent a newly evolved connection. In all groups, the ventral striatum is modulated by dopaminergic input from the A8–A10 cell groups and by noradrenergic input from LoC (the catecholaminergic input to the ventral striatum is depicted in green). By way of the PPN, the ventral BG in all groups may influence descending motor pathways (through the tectum) and ascending activating pathways (through the midline/intralaminar thalamus) (the cholinergic pathways are depicted in orange). Cb, cerebellum; CPu, caudo-putamen complex (dorsal striatum); DC, dorsal cortex (reptile); DCN, dorsal column nuclei; DVR, dorsal ventricular ridge; GP, globus pallidus; GPe, globus pallidus, external or lateral segment; GPi, globus pallidus, internal or medial segment; H, hyperpallium (dorsal pallium in birds); Hp, hippocampal formation; IC, inferior colliculus; INT, intralaminar thalamic nuclei; LDT, laterodorsal tegmental nucleus; LHy, lateral hypothalamus; LSt, lateral striatum (birds); M1, primary motor cortex; M, mesopallium (lateral pallium of birds); MD, mediodorsal thalamic nucleus (midline thalamus); MSt, medial striatum (birds); N, nidopallium (part of ventral pallium in birds); OB, olfactory bulb; oc, optic chiasm; PFC, prefrontal cortex; Po, preoptic area; PrM, premotor cortex; PT, pretectal nucleus of the posterior commissure (it receives dorsal BG input); SC, superior colliculus; SNr, substantia nigra pars reticulata; St, striatum; STN, subthalamic nucleus; TeO, optic tectum; VB, ventrobasal thalamus; VPa, ventral pallidum.

GABA and calretinin. It appears that the striatum of rodents and primates also contains a small subpopulation of neurons containing tyrosine hydroxylase/dopamine. This appears to represent another interneuron subtype, without clear relation to the four interneuron subtypes mentioned previously. However, although most striatal interneurons appear to originate outside the striatum (in the MGE or basal to it), it is likely that the striatal neurons containing tyrosine hydroxylase/dopamine originate in a dorsal subdivision of the lateral ganglionic eminence (dLGE), which is known to be the source of the dopaminergic interneurons of the olfactory bulb.

Pallidal neurons are typically large and the majority contain GABA, the neuropeptide LANT6 (a neurotensin-related hexapeptide), and parvalbumin. The dendrites of these cells show a typical flat-disk arrangement and are contacted by striatal axons while passing through. However, the cellular composition of the pallidum is more complex. For example, in the GPe, although the majority of the neurons (approximately two-thirds) are the typical large cells containing GABA, LANT6, and parvalbumin and primarily show descending projections to the subthalamus and substantia nigra pars reticulata (SNr), there is an important subpopulation of neurons (representing approximately one-third) that contain GABA, calbindin, and ENK and project back to the striatum (in addition to the subthalamus/SNr). Moreover, it appears that the GPe also contains a minor interneuron subpopulation containing calretinin.

The pallidum receives its major input from the striatum and projects to the brain stem and to the thalamus (influencing thalamocortical circuits). Due to the mostly segregated striatal input to GP, the GPi is characterized by dense SP+ innervation, whereas the GPe is characterized by dense ENK+ innervation. The ventral pallidum (VP) is densely innervated by both SP+ and ENK+ fibers. As noted previously, the projections of VP differ from those of GP (Figures 3 and 4). In the thalamus, while VP targets the mediodorsal nucleus (midline thalamus), GP projects to the motor part of the ventrobasal complex (ventral anterior and ventral lateral nuclei), thus influencing different thalamocortical circuits. Both VP and GP also project to intralaminar thalamic nuclei (including the parafascicular nucleus). Similarly, whereas GP projects to the subthalamic nucleus (STN), VP primarily projects to the medial aspect of STN and adjacent lateral hypothalamic area. In the tegmentum, whereas GP projects to the SNr, VP mainly projects to the ventral tegmental area, SNc, retrorubral area (A10, A9, and A8 catecholamine cell groups, respectively) and locus coeruleus. These differences in connectivity between GP and VP underlie the functional differences described for dorsal and ventral BG. On the other hand, the projections of GPe and GPi differ and are related to separate circuits that have opposite functions for BG control of movement. The organization of these circuits is explained in more detail in the following section.

Functional Circuits

In mammals, the BG influence motor, cognitive, and emotional functions by way of either BG-thalamocortical recurrent (re-entrant) pathways or descending outputs to the brain stem (e.g., the BG outputs to tectoreticulospinal pathways). The activity of BG is modulated by dopaminergic, serotonergic, and noradrenergic input.

Parallel cortico-basal ganglia-thalamocortical loops for control of motor, cognitive–executive, and emotional/motivational functions One interesting aspect of BG connections with the cortex is that they are organized in parallel, mostly segregated circuits. These were first proposed when comparing ventral versus dorsal BG, which by way of distinct connections (distinct input from the cortex and output to different thalamocortical circuits) could influence either motivational/emotional or motor behavior (Figures 3 and 4). Later, the existence of multiple parallel, segregated circuits was observed, channeled through different parts of the dorsal or ventral BG. Thus, it appears that each functional area of the cortex is related to a specific, functionally equivalent part of the BG, and that there are multiple parallel cortico-BG-thalamocortical loops or channels, each involved in control of a specific aspect of motor behavior or cognition. Each cortico-BG-thalamocortical channel reaches the same cortical area that gives rise to the circuit; that is, these are mostly closed channels. Distinct circuits influence distinct cortical areas and can control different aspects of motor, cognitive–executive, or emotional/motivational behavior. However, there is evidence that some BG-thalamocortical loops are partially open. For example, the ventral putamen of primates gives rise to a circuit that ultimately targets the primary motor cortex, which does not project to that part of the putamen. In contrast, this part of the ventral putamen receives amygdalar input, thus providing a route for limbic/emotional influence on motor output.

The lateral, motor thalamus (ventral anterior and ventral lateral nuclei) that mediates BG pathways to the cortex also receives cerebellar input, mostly mediated by the dentate nucleus. For this reason, it used to be thought that motor information from both BG and cerebellum converged on the same thalamic nuclei that project to the motor cortex. However, it is now clear that BG and cerebellar targets in the motor

thalamus are mostly segregated. Evidence indicates that although the BG and cerebellum reach the same cortical areas of the frontal or prefrontal cortex mostly by way of parallel thalamocortical channels, they target different sectors of each cortical area.

Direct, indirect, and hyperdirect pathways for release and inhibition of movement The BG-thalamocortical circuits consist of direct and indirect pathways, which have opposite roles on control of motor/cognitive functions. These are better known for the dorsal BG, which show segregated projections through the substantia nigra and primarily distinct parts of GP (Figures 3 and 5). The GPi and the SNr receive their major input from GABA+/SP+/DYN+ cells of the CPu and project to the motor thalamus (ventral anterior/ventral lateral nuclei of the ventrobasal thalamic complex). This striato-GPi/SNr-thalamic circuit constitutes the so-called direct pathway for BG control of movement (red color in Figures 3 and 5). Since the striatal, pallidal, and SNr projections are GABAergic, and the thalamocortical projection is excitatory, the activation of this pathway by the cortex or thalamus produces a disinhibition of the motor thalamus and ultimately produces activation of the motor cortex and a release of movement. This direct pathway is controlled by an indirect circuit, which originates in the GABA+/ENK+ cells of the CPu (blue color in Figures 3 and 5). The GABA+/ENK+ striatal neurons project to the GPe, which in turn projects to an excitatory hypothalamic center called the STN that projects to the GPi and SNr. Activation of this indirect pathway produces disinhibition of the STN and thus excitation of GPi and SNr. This ultimately inhibits the motor thalamus and blocks movement. It appears that the combined activation of both direct and indirect pathways releases desired movement while blocking unwanted movement.

In addition to the direct and indirect pathways for movement release and inhibition, there is a hyperdirect pathway that produces movement inhibition by way of a fast, direct motor cortical input to the STN.

Basal ganglia pathways to the brain stem In addition to the cortico-BG-thalamocortical reentrant loops, the BG also influences motor or cognitive behavior by way of descending projections to the brain stem. These projections are better known for the dorsal BG, which project to the SNr that in turn projects to the midbrain tectum (superior colliculus) (Figures 3 and 5). This pathway targets tectoreticulospinal cells located in deep layers of the tectum. The apical dendrites of these cells receive visual input from the retina, and this BG pathway is important for controlling eye, head, and neck movement and thus for control of visuomotor and orientation behavior. In addition, the SNr projection to the midbrain tectum targets ascending tectocortical cells, providing another mechanism for visuomotor control.

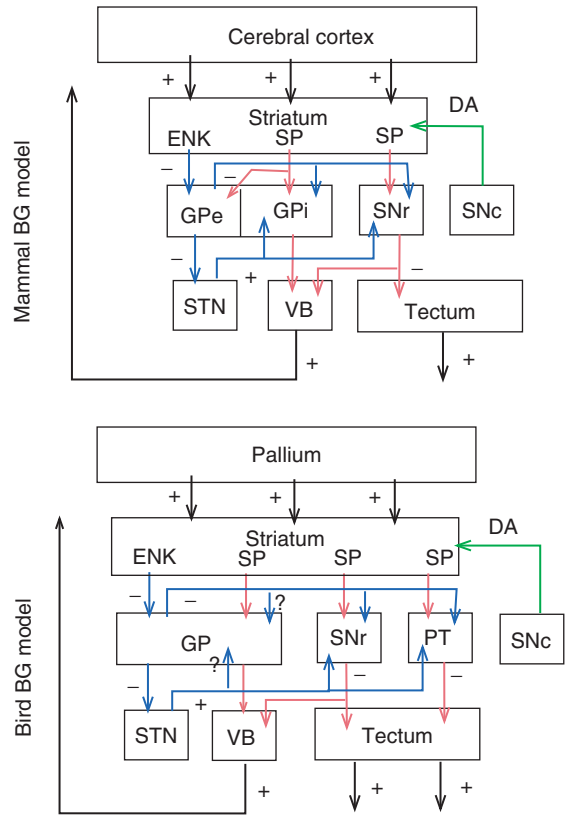

Figure 5 Diagrams representing basal ganglia (BG) functional models in mammals (top) and birds (bottom). Based on the similarity of cell types, neurotransmitters, and connections, it appears that the BG of birds and mammals use mostly similar mechanisms for control of motor behavior. These involve similar pathways to the midbrain tectum (superior colliculus), which are very prominent in birds, and similar pathways to the frontal cortex/dorsal pallium, which are extremely well developed in mammals. The BG projections involve direct and indirect pathways that are involved in either the release or inhibition of movement, respectively. The direct pathway (red) originates in SP+ cells of the striatum and is going to produce disinhibition of either thalamocortical projections or tectoreticulospinal projections, thus producing release of movement. Activation of the indirect pathway (blue) originates in ENK+ cells of the striatum and produces disinhibition of a glutamatergic center, the subthalamic nucleus (STN), that in turn projects to the direct pathway, producing the opposite effect. For simplicity, only a few centers and connections are included in the scheme (the pedunculopontine tegmental nucleus, which plays an important role in mammals, has been omitted but is discussed in the text). DA, dopamine; GP, globus pallidus; GPe, globus pallidus, external or lateral segment; GPi, globus pallidus, internal or medial segment; PT, pretectal nucleus of the posterior commissure (it receives BG input); SNc, substantia nigra pars compacta (A9 catecholamine cell group); SNr, substantia nigra pars reticulata; VB, ventrobasal thalamus.

On the other hand, the ventral BG and the adjacent parts of the dorsal BG project to the pedunculopontine tegmental nucleus (PPN) and laterodorsal tegmental nucleus (LDT) in the reticular formation of the rostral hindbrain (Figures 3–5). The PPN and LDT are cholinergic cell groups that belong to an ascending activating system that reaches widespread areas of the cortex and striatum mainly by way of the intralaminar/midline nuclei of the thalamus. Through these projections, the PPN and LDT appear to play an important role in arousal and awareness. Thus, through projections to the PPN/LDT, the BG are able to influence this important aspect of cognition. Moreover, the PPN projects to the SNc and globus pallidus, and it is reciprocally connected with the SNr and STN. The PPN is able to influence BG-thalamocortical recurrent circuits and has been proposed to be an important actor for BG control of motor behavior.

In addition, the PPN projects to the midbrain tectum (which has been related to orienting behavior and initiation of saccades) and also shows descending projections to medullary systems (Figures 3–5). These pathways through the PPN constitute additional links for BG modulation of tectal outputs and other descending systems involved in motor control.

Modulation of basal ganglia activity As noted previously, activity of striatal principal neurons is modulated by ascending projections from dopaminergic, serotonergic, and noradrenergic groups of the brain stem (A8–A10 cell groups, raphe nuclei, and locus coeruleus, respectively) (Figures 3–5). The best known of these is the modulatory input from dopaminergic cells of the ventral tegmental area (A10), the substantia nigra pars compacta (SNc, A9), and the retrorubral field (A8). In the striatum, dopamine activates GABA+/SP+/DYN+ neurons while inhibiting GABA+/ENK+ (Figure 5). Depletion of the dopaminergic innervation of the striatum (as in Parkinson's disease due to gradual death of SNc neurons) produces movement abnormalities. This releases the indirect pathway from inhibition, producing unwanted movements (tremor) at rest during initial stages. This also produces loss of excitation of the direct pathway, gradually leading to bradykinesia (slowness of movements), postural instability, akinesia, and muscular rigidity at later stages. In addition, depletion of dopaminergic innervation leads to cognitive/emotional deficits. This is also observed in Parkinson's disease, which is partially due to loss of dopaminergic innervation in the striatum and prefrontal cortex, as well as cell loss in the locus coeruleus (and consequently loss of noradrenergic innervation in the ventral striatum) and in the PPN.

Whereas the dopaminergic input to the CPu mainly arises in the SNc (A9) and is important for movement control and cognition, the input to the ventromedial striatum (including the nucleus accumbens) mostly originates in the A8–A10 groups and appears important for motivation and reward.

The Basal Ganglia in Nonmammals

Conservative Evolution of the Basal Ganglia in Tetrapods

The BG have also been described in the basal (subpallial) telencephalon of birds, reptiles, and amphibians, suggesting that they are present in all tetrapod vertebrates (a group that includes mammals, birds, reptiles, and amphibians) (Figure 6). Similar to the BG of mammals, in these vertebrates the BG show striatal and pallidal subdivisions and also similar neurochemical characteristics and basic connections (Figure 1). This suggests that many of these features were present in the first tetrapod vertebrates, and that BG evolution has been highly conservative. However, important divergence in BG evolution appears to have occurred in the transition from stem tetrapods to amniotes (including birds, reptiles, and mammals; Figure 6) and later in the lineages to birds or mammals.

Developmental evidence supports basal ganglia homology The most important evidence supporting the homology of the BG across different tetrapod

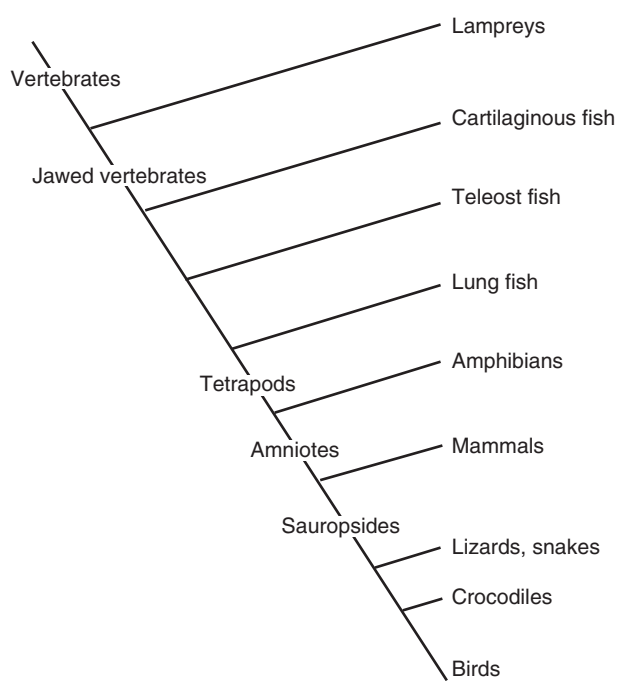

Figure 6 Cladogram showing the phylogenetic relationship between different groups in vertebrates.

groups comes from developmental studies and expression of transcription factors. First, the BG in both mammals and nonmammalian tetrapods develop from the same (homologous) histogenetic division of the telencephalon, the subpallium, characterized by expression of Dlx/Distal-less family transcription factors and lack of pallial marker genes such as Emx1/2 and Tbr1/2 (Figure 2). Second, the subpallium of mammals and nonmammalian tetrapods contains the same two major subdivisions, namely striatal and pallido-pedunculo-preoptic subdivisions (Figure 2). The striatal subdivision (called LGE in mammals) expresses Dlx/Distal-less but not Nkx2.1/Titf-1 and gives rise to the striatum. The pallido-pedunculo-preoptic subdivision expresses Dlx/Distal-less, Nkx2.1/Titf-1, Lhx6, and/or Lhx7/8 and gives rise to the pallidum proper (called MGE in mammals), a pallidal part of the septum (septal pallidum), the anterior peduncular area (AEP), and a subpallial (anterior) part of the preoptic area (POa) (AEP–POa are related to part of the extended amygdala and the cholinergic cell groups of the basal telencephalon). Thus, the BG in all tetrapods consist of the same (homologous) two subdivisions, striatum and pallidum, which are formed from homologous histogenetic fields, located in the same topological position (same relative position within the brain), under the influence of the same transcription factors. Moreover, evidence indicates that these transcription factors play the same role in BG development in mammals and nonmammals. For example, Nkx2.1/Titf-1 is involved in specification of the pallidal subdivision in both mammals and nonmammals, and lack or depletion of function of this transcription factor produces severe malformation of this territory.

Comparable cell types, neurotransmitter organization, and connections Similar to mammals, the BG of birds, reptiles, and amphibians are rich in GABAergic neurons and typically show GABAergic, inhibitory projections (in contrast to the cortex/pallium, which shows glutamatergic, excitatory projections). As in mammals, the striatum of birds, reptiles, and amphibians is characterized by high acetylcholinesterase activity and dense dopaminergic innervation that originates in comparable cell groups of the mesodiencephalic tegmentum (A8–A10 cell groups). The striatum is the major input station (receiving input from both pallium/cortex and thalamus), and the pallidum is the major BG output center to the hypothalamus/subthalamus, thalamus, pretectum, and brain stem. The striatum, especially in amphibians and also in reptiles and birds, also contributes to some of the BG outputs.

Based on their relative position in the subpallium and connections, dorsal (dorsolateral) and ventral (ventromedial) parts of the striatum have been identified in amphibians, reptiles, and birds. As in mammals, the striatum projects to the pallidum in nonmammalian tetrapods, and each striatal division has an associated pallidal part.

Similar to the CPu of mammals, the dorsolateral striatum of birds and reptiles contains two major types of projection neurons: (1) GABA+/SP+/DYN+ neurons projecting to the pallidum and substantia nigra and (2) GABA+/ENK+ neurons mainly projecting to the pallidum (red or blue, respectively, in Figure 3). Although the dorsal pallidum in nonmammals shows no subdivisions, the dorsal striatopallidal projections from either SP+ or ENK+ cells end on different pallidal cells, and the pathways remain segregated at least in birds. The ventromedial striatum of birds and reptiles, and the striatum of amphibians, also contain ENK+ and SP+ GABAergic neurons projecting to the nigra and ventral tegmental area, although the segregation of both cell types is less clear. Regarding the interneurons, the striatum of birds and reptiles possesses at least the same three major interneuron subtypes described in mammals: acetylcholine, GABA/parvalbumin, or GABA/neuropeptide Y/somatostatin (the latter also contains nitric oxide synthase and shows NADPH-diaphorase-positive activity). Cholinergic and nitrergic/somatostatin-containing cells are also present in the striatum of amphibians, suggesting the presence of at least two interneuron subtypes (which were possibly present in the striatum of stem tetrapods). As in mammals, it appears that the striatal interneurons in nonmammalian tetrapods originate outside the striatum, in the histogenetic domain expressing Nkx2.1/Titf-1 (which includes the pallidal subdivision).

Similar to mammals, the dorsolateral striatum of birds and reptiles receives widespread cortical/pallial input (including from the dorsal pallium but also from other parts of the pallium) and projects to the dorsal pallidum (globus pallidus) (Figure 3). The ventromedial striatum of birds and reptiles, on the other hand, receives input from limbic parts of the pallium (including the pallial amygdala and the medial pallium) and projects to the ventral pallidum (Figure 4). The pallial input to the striatum is much more modest in amphibians and originates in limbic areas such as the medial pallium (hippocampal-like), the ventral pallium (pallial amygdalar-like), and a rostral pallial area reciprocally connected to olfactory areas (Figure 7). It seems that the striatum does not receive input from the dorsal pallium in amphibians. As in mammals, the whole striatum of birds, reptiles, and amphibians receives midline/intralaminar-like thalamic input. In addition,

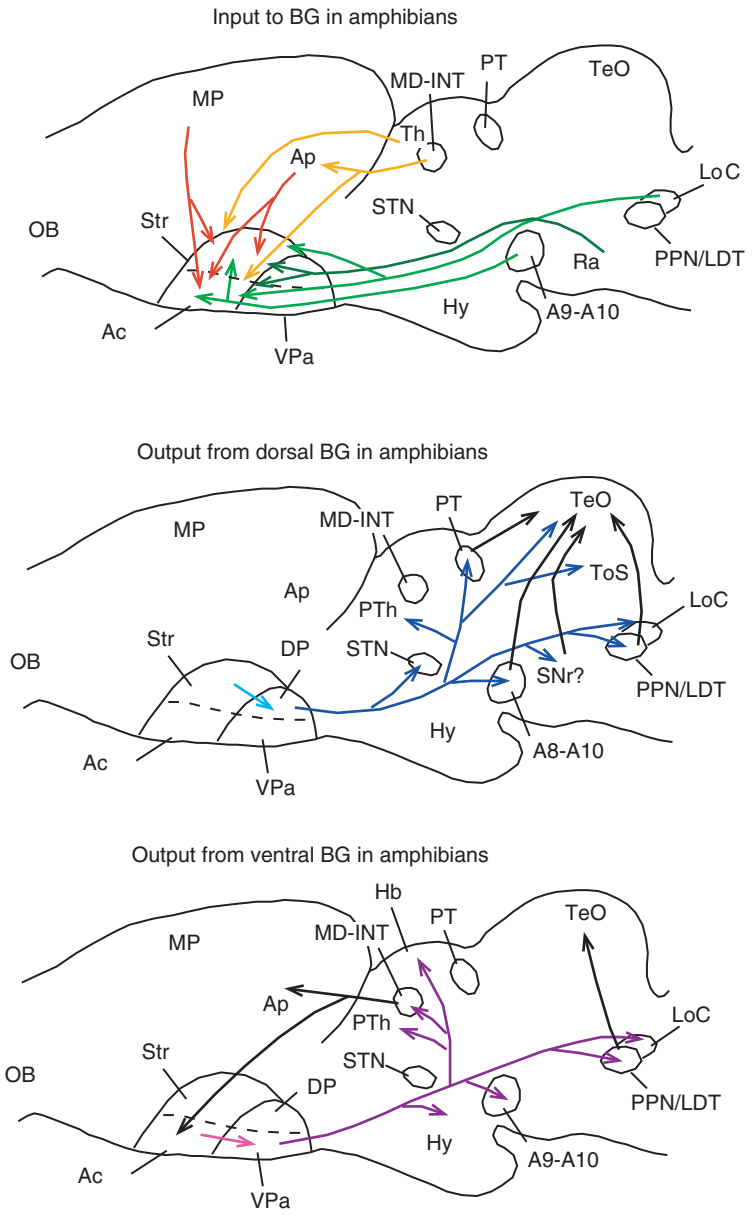

Figure 7 Schematic drawings of lateral brain views of an amphibian (a frog) showing the main afferent (top) or efferent pathways (middle and bottom) of the dorsal and ventral basal ganglia (BG). All or most of these connections are also present in reptiles, birds, and mammals and were likely present in ancestral tetrapods. Ac, nucleus accumbens; Ap, pallial amygdala; DP, dorsal pallium; Hy, hypothalamus; INT, intralaminar thalamic nuclei; LDT, laterodorsal tegmental nucleus; LoC, locus coeruleus; MD, mediodorsal thalamic nucleus (midline thalamus); MP, medial pallium; OB, olfactory bulb; PPN, pedunculopontine tegmental nucleus; PT, pretectal nucleus of the posterior commissure (it receives BG input); PTh, prethalamus (ventral thalamus); Ra, Raphe nuclei; SNr, substantia nigra pars reticulata; STN, subthalamic nucleus; Str, striatum; TeO, optic tectum; Th, thalamus (dorsal thalamus); ToS, torus semicircularis (comparable to the inferior colliculus of mammals); VPa, ventral pallidum.

the striatum of amphibians receives important input from sensory thalamic nuclei. As a whole, in amphibians the thalamic input to the striatum is more important than the pallial input (Figure 7), and this was possibly similar in stem tetrapods.

On the other hand, in birds, reptiles, and amphibians the pallidum contains GABA+/LANT6+ neurons that project to the hypothalamus/subthalamus, thalamus, and brain stem in a pattern that partially resembles that in mammals (Figures 3, 4, and 7). Thus, the dorsal pallidum of nonmammals projects to the subthalamus and SNr, which is typical of the mammalian globus pallidus (Figures 3 and 7). The ventral pallidum of nonmammals projects to the lateral hypothalamus (adjacent to the subthalamus), the A8–A10 catecholamine cell groups (including the

ventral tegmental area and SNc), the cholinergic pedunculopontine/laterodorsal tegmental nuclei, and the locus coeruleus (Figures 4 and 7).

The ventral and dorsal pallidum of birds also show projections to distinct thalamic nuclei (midline- or ventrobasal-like) that project to either limbic or somatomotor telencephalic areas, respectively (including the dorsal cortex/pallium) (Figures 3 and 4). In addition, the ventral and dorsal pallidum of birds show projections to intralaminar-like thalamic nuclei (including the dorsointermediate nucleus, comparable to the parafascicular nucleus of mammals), which in turn show widespread projections to the striatum and pallium. These pallido-thalamo-telencephalic projections of birds partially resemble those described in mammals. However, in reptiles and amphibians only the ventral pallidum, but not the dorsal pallidum, projects to the dorsomedial thalamus (Figures 3, 4, and 7). The available data suggest the existence of similar ventral BG-thalamo-striatal pathways in all tetrapods (by way of midline/intralaminar-like thalamic nuclei). However, the BG-thalamo-pallial connection appears to be different between mammals and nonmammals. The BG-thalamo-telencephalic circuits are explained in more detailed in the next section.

In addition to the projections to the thalamus (dorsal), the pallida (dorsal and ventral parts) of mammals, birds, reptiles, and amphibians do share similar projections to the prethalamus (ventral thalamus), which involves an indirect BG influence on the thalamus (dorsal thalamus) and on thalamo-telencephalic projections (Figure 7). This BG projection to the prethalamus was likely present in the origin of tetrapods.

On the other hand, the dorsal pallidum of amphibians, reptiles, and birds projects to a pretectal nucleus characterized by the presence of GABA+/ENK+ cells that project to the tectum (Figures 3 and 7). Evidence suggests that the presence of a pallido-pretectal-tectal pathway was the primitive condition in tetrapods, but this BG pathway to the tectum through the pretectum has undergone a great reduction in the mammalian lineage and is apparently absent in rodents, carnivores, and primates. It would be interesting to analyze the presence of a BG-pretecto-tectal pathway in noneutherian mammals.

Functional circuits: Evidence for conservation and divergence Based on available data on neurotransmitters and connections, the BG of mammals, birds, reptiles, and amphibians share some basic circuits and appear to be involved in similar basic motor and limbic functions. However, some differences exist, indicating divergent evolution of some BG circuits in the avian and mammalian lineages.

Data on the dorsal BG are more abundant and suggest the existence of several pathways to the midbrain tectum in nonmammals, some of which are comparable to those of mammals (Figures 3, 5, and 7). For example, the dorsal BG of mammals and nonmammals show a comparable projection to the tectum by way of the SNr. This pathway appears to be present at least in all amniotes (mammals, birds, and reptiles). In addition, the dorsal BG of birds, reptiles, and amphibians possess another pathway to the midbrain tectum by way of the pretectum (through the nucleus of the posterior commissure in reptiles or its homolog in birds and amphibians). This BG-pretecto-tectal pathway shows great development in reptiles and especially in birds. As in mammals, these pathways to the midbrain tectum of nonmammals are suggested to be involved in BG control of eyes, head, and neck movement and in orientation responses. The involvement of nonmammalian BG in motor behavior has been evidenced in birds and reptiles by the observation of movement disorders following destruction of the dopaminergic innervation of the striatum.

There is evidence in birds indicating the existence of direct and indirect pathways in the BG projection to the tectum (red or blue in Figures 3 and 5). Thus, the dorsal BG projection to both SNr and pretectum comes directly from SP+ neurons of the striatum and indirectly through the pallidum by way of the ENK+ striatopallidal projection. Since striatal, pallidal, SNr, and pretectal cells involved in these projections are GABAergic, activation of the SP+ striatal cells will likely produce disinhibition of tectoreticulospinal neurons and release of movement, whereas activation of the ENK+ cells likely produces the opposite. Moreover, birds, reptiles, and amphibians appear to have a glutamatergic subthalamic center comparable (homologous) to the subthalamic nucleus of mammals (Figures 3, 5, and 7). The subthalamic nucleus of these vertebrates receives input from the dorsal pallidum (via the ENK+ striatopallidal pathway) and shows an excitatory projection to the SNr and pretectum, thus providing another indirect mechanism for BG movement control. Moreover, the subthalamic nucleus projects back to the dorsal pallidum and, possibly as in mammals, may influence the direct pathway (i.e., subthalamic axons contact pallidal cells receiving SP+ striatal input), providing another indirect mechanism for BG movement control (Figure 5).

In addition to the previously mentioned pathways to the midbrain tectum, birds have a dorsal BG pathway to a somatomotor area of the dorsal pallium/cortex by way of a dorsal thalamic center (ventrointermediate area of birds, comparable to the motor part of the ventrobasal complex of mammals)

that resembles in many respects the well-known BG-thalamo-motor cortical circuit of mammals (Figures 3 and 5). Similar to mammals, this pathway in birds shows a direct, SP+ striatopallidal circuit and an indirect ENK+ striatopallidal circuit that controls the direct channel through a glutamatergic center of the subthalamus. As in mammals, based on their similar neurotransmitters and connections, it is likely that activation of the direct circuit produces release of movement, whereas activation of the indirect circuit produces inhibition of unwanted movements.

The striking similarity of avian and mammalian dorsal BG pathways to dorsal pallium suggests that they are homologous. In addition to the specific dorsal BG pathway to the dorsal pallium by way of a ventrobasal-like thalamic nucleus, the dorsal BG of birds and mammals show comparable projections to an intralaminar-type thalamic nucleus (dorsointermediate nucleus or DIP in birds, comparable to the parafascicular nucleus of mammals) that projects to the telencephalon (to both striatum and cortex/pallium). However, it seems that such dorsal BG pathways to the cortex/pallium by way of the thalamus do not exist in reptiles since the dorsal BG does not project to the dorsal thalamus in any reptilian species analyzed. Therefore, the most parsimonious interpretation is that mammals and birds evolved dorsal BG-thalamo-pallial pathways independently.

On the other hand, the pallido-thalamo-dorsal pallial pathway in birds is meager and likely has only a minor influence on motor control. Thus, in contrast to mammals, the BG of birds, reptiles, and amphibians exert their control on movement mainly (if not only) by way of pathways to the midbrain tectum.

Regarding the ventral BG, evidence indicates that they show many similar connections between mammals, birds, reptiles, and amphibians (Figures 4 and 7). This suggests that they are involved in similar basic functions by way of similar circuits. For example, in amphibians, reptiles, birds, and mammals, the ventral or ventromedial striatum receives input from the limbic cortex/pallium, including the medial (hippocampal) pallium and the pallial amygdala, and is reciprocally connected with the dopaminergic cells of the ventral tegmental. This suggests that, as in mammals, the ventral (ventromedial) striatum of birds, reptiles, and amphibians may be involved in functions such as motivation/reward, attention, and learning, and this has been shown in birds.

In addition, the ventral BG in birds, reptiles, amphibians, and mammals show comparable projections to the reticular formation of the rostral hindbrain, where the PPN and LDT tegmental cholinergic nuclei are located (Figures 4 and 7). Considering the high similarity in the cholinergic systems among reptiles, birds, and mammals, it is likely that the BG projection to the PPN/LDT is a gateway for general activation of the telencephalon (pallium and subpallium) by way of intralaminar/midline thalamic nuclei (Figure 4). Thus, this may represent a BG pathway for control of arousal and awareness, which appears to be present at least in all amniotes. As in mammals, the cholinergic tegmental nuclei that receive BG input in nonmammals appear to represent a link for influencing descending tectoreticulospinal pathways (Figures 4 and 7) and perhaps for controlling some aspects of visuomotor and orientation behavior in nonmammals.

On the other hand, amphibians, birds, reptiles, and mammals show ventral BG projections to medial thalamic nuclei that project to the telencephalon (Figures 4 and 7). In mammals and birds, the relay in the medial thalamus includes both midline and intralaminar nuclei (Figure 4). The midline thalamic relay (mediodorsal nucleus of mammals and dorsomedial thalamic nuclei of birds) projects to the limbic-related ventral striatum and to specific areas of the pallium involved in cognitive tasks (prefrontal cortex in mammals and nidopallium in birds). The midline thalamic relay in birds also projects to the pallial amygdala. In birds and mammals, there is a closed cortico/pallio-BG-thalamocortical/pallial loop by way of the midline thalamus. However, the loop of birds is not comparable to that of mammals since it involves a pallial region (nidopallium, which derives from the ventral pallium) different from the neocortex (a dorsal pallial derivative). On the other hand, the intralaminar-like thalamic relay (e.g., the parafascicular nucleus of mammals and its putative avian homolog, the dorsointermediate nucleus) shows relatively wider projections to both striatum and pallium/cortex in both birds and mammals.

In reptiles and amphibians, the ventral BG projects to the dorsomedial thalamus, and this dorsomedial thalamic area projects back to the ventral BG (ventral striatum) (Figures 4 and 7). Based on its position and reciprocal connection with ventral BG, the dorsomedial thalamic nucleus receiving BG input of amphibians and reptiles resemble the midline and intralaminar nuclei of mammals and birds. The dorsomedial thalamus of reptiles that receives ventral BG input also projects to the pallium, including the ventrolateral pallium (posterior dorsal ventricular ridge (DVR) and pallial amygdala), resembling the projections of the midline thalamic nuclei of birds. A similar ventral BG-thalamo-ventrolateral pallial circuit is also present in amphibians (Figure 7) and may have been present in the brain of ancestral tetrapods. It appears that this basic pathway underwent a great elaboration especially in birds (most prominently in songbirds). In mammals, the midline thalamus has

also become very complex and includes nuclei (distinct from the BG-recipient mediodorsal nucleus) that project to the pallial amygdala and hippocampal formation. Perhaps all of these midline thalamic nuclei of mammals evolved from a basic cell group present in stem amniotes or stem tetrapods, similar to those of the dorsomedial thalamus of modern reptiles or amphibians (receiving ventral BG input and projecting to the pallial amygdala). On the other hand, based on the lack of projections from the midline thalamic relay to the dorsal pallium, reptiles, amphibians, and birds do not possess a circuit comparable to the ventral BG-thalamo-prefrontocortical loop of mammals.

Divergence and Novelties of the Basal Ganglia in Tetrapods

As noted previously, the BG share many features (in terms of development, subdivisions, cell types, neurotransmitter organization, and connections) across amphibians, reptiles, birds, and mammals, indicating a highly conservative evolution in tetrapods. However, as also noted, there are some differences between groups denoting the existence of divergent evolution and the appearance of novel features in the BG in some lineages within tetrapods. The major divergent steps apparently occurred in the transition from stem tetrapods to amniotes and later in the evolutionary lineages of birds and mammals.

Expansion of thalamus and cortex/pallium drove the evolution of basal ganglia-thalamocortical pathways The transition from a semiaquatic habitat, such as that used by amphibian tetrapods, to a fully terrestrial habitat, such as that used by reptiles, was accompanied by dramatic changes in the locomotor (musculoskeletal) system, sensory organs, and related centers of the central nervous system. This apparently drove the expansion of the dorsal thalamic centers involved in sensory processing and also the expansion of their telencephalic targets, especially in the pallium. Of particular relevance was the expansion of the pallium, with the appearance of specific areas involved in sensory processing, association, and sensorimotor integration. Within this framework, the pallium/cortex displaced the subpallium as the major center for processing and integrating the complex and fast-changing information of the new habitat and as the major executing center in terms of response. This task was done in conjunction with the subpallium and was accompanied by great development and elaboration of cortico/pallio-BG projections. This is the pattern observed in modern reptiles, as well as in birds and mammals, with a large to huge expansion of the thalamus and cortex/pallium, and, with respect to amphibians, a large expansion of the thalamocortical/pallial and cortico/pallio-BG projections. The richer and more elaborated information that reaches the BG in reptiles (as well as birds and mammals) mainly through the pallium/cortex suggests that in these animals, this structure plays a more sophisticated role in motivation, learning, and motor control than in amphibians.

The thalamus, the cortex/pallium, and the associated BG, as well as their projections, have continued to expand in the evolutionary lineages of birds and mammals. This has been accompanied by the evolution of additional direct BG projections to the dorsal thalamus, by the elaboration of preexistent BG-thalamocortical/pallial pathways, and by the appearance of novel BG-cortical loops (e.g., the dorsal BG-thalamo-dorsal pallial/neocortical pathways). Since the dorsal BG pathways to the neocortex/dorsal pallium via the thalamus are not present in modern reptiles, it seems that these circuitries evolved independently in birds and mammals.

An interesting aspect of this evolution is that birds and mammals have expanded different sectors of the thalamus and associated pallium. Thus, birds show the greatest expansion in the collothalamus (part of the thalamus receiving sensory input from the colliculus/tectum) and in the ventrolateral pallium (which gives rise to a large pallial structure, the DVR, that includes the nidopallium of birds; at caudal levels it includes the pallial amygdala). In contrast, mammals show the greatest expansion in the lemnothalamus (part of the thalamus receiving sensory input directly from the lemniscus or from the retina) and associated dorsal pallium (neocortex). As a consequence, the nature of the pallial input that primarily reaches the BG differs between mammals and birds. This is explained in more detail later.

The medial striatum of songbirds and budgerigars: Specialization in song/vocalization learning and plasticity A major medial part of the medial striatum of birds (previously called lobus paraolfactorius) shows cellular and connectivity patterns that resemble the ventral (ventromedial) striatum of mammals, such as the nucleus accumbens. This includes input from the medial (hippocampal) pallium and pallial amygdala; reciprocal connection with the dopaminergic ventral tegmental area; input from noradrenergic cells of the locus coeruleus; and projections to the ventral pallidum, hypothalamus, habenula, and pedunculopontine tegmental nucleus. This part of the striatum has been shown to be involved in reward and also in learning and memory formation. This part of the avian medial striatum appears to represent the viscerolimbic striatum (together with the olfactory tubercle), and it

appears to be involved in functions similar to those described for the mammalian nucleus accumbens, including motivation, reward, and cognitive functions such as learning and memory. The most lateral part of the avian medial striatum gradually shows features more typical of the mammalian CPu and is involved in motor control. As in mammals, there is no sharp boundary between the 'limbic' striatum and the 'somatomotor' striatum, and the transition from one to the other is gradual.

In oscine songbirds and budgerigars (e.g., the parrot), the medial striatum includes a specific area that is involved in song or vocalization learning and plasticity. This area of the medial striatum is called area X in oscine songbirds and the magnocellular nucleus of the medial striatum in budgerigars. Both striatal areas are part of highly elaborated pallio-BG-thalamopallial circuits, which together with other pathways and forebrain centers specific to songbirds or budgerigars control different aspects of song or vocalization (from learning to production) (Figure 8). These circuits of oscine songbirds and budgerigars are suggested to have evolved independently. Nevertheless, these circuits and the specific brain centers may have evolved from basic, less-elaborated pathways and centers present in the origin of the avian radiation. Supporting this, a number of studies indicate that area X shows many features typical of the surrounding medial striatum, including a majority of small spiny GABAergic neurons containing SP or ENK, cholinergic, and GABAergic interneurons; glutamatergic input from the ventrolateral pallium (including a nidopallial center called LMAN); and a reciprocal connection with the ventral tegmental area (among others). An interesting aspect of area X is that it contains a pallidal-like cell type that projects to a specific nucleus the thalamus (medial part of the dorsolateral nucleus) that in turn projects to LMAN (thus closing the loop) (Figure 8). The physiological properties of this cell type are similar to those of typical pallidal cells, including being spontaneously active and capable of sustained high-frequency firing. This has led to the suggestion that area X is a mixed striatopallidal area. Interestingly, the pallidal-like cell type of area X receives input from the spiny neurons of this area (which are typically striatal and receive direct pallial input), indicating that area X contains the basic elements for a typical striato-pallido-thalamic direct pathway (the striatal and pallidal cells are depicted in pink or purple, respectively, in Figure 8). Moreover, not all pallidal-like cells of area Xproject to the thalamus. There is a subpopulation of pallidal-like cells that project to other pallidal-like neurons of area X projecting to the thalamus, thus providing an indirect BG route to the thalamus. All of this evidence raises the possibility that the pallidal-like cells of area X are truly pallidal (i.e., they originate in the pallidal subdivision of the subpallium). However, these cells do not appear to express the transcription factor Nkx2.1 (at least in adult animals, although this may be due to downregulation). On the other hand, these cells contain the neuropeptide ENK (typical in striatal cells). However, the globus pallidus in mammals also contains a subpopulation of ENK+ neurons, and this projects back to the striatum. Therefore, the exact nature of the pallidal-like cells of area X is unclear. Nevertheless, the majority of the features of area X (except the presence of the pallidal-like cells projecting to the thalamus, which represent a minority) are typically striatal and resemble the adjacent areas of the avian medial striatum. Interestingly, the medial striatum of non-songbirds, such as the chick, also contains pallidal-like cells and projects to the thalamus. Moreover, in chicks there is a ventral BG-thalamo-nidopallial loop that resembles that through area X in songbirds (Figure 4). This supports the view that area X of songbirds evolved from a striatal field showing some basic neurochemical and connectivity features that are common to most (if not all) birds.

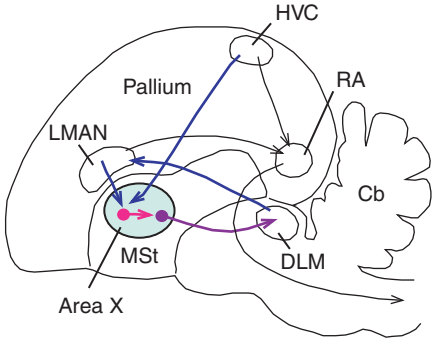

Figure 8 Schematic drawing of a lateral view of a songbird brain showing circuits related to song production and learning. Note the existence of a basal ganglia–thalamocortical/pallial loop that plays a role in song learning and plasticity, which involves a specific nucleus of the medial striatum called area X. This nucleus appears to include spiny striatal neurons (pink) as well as pallidal-like neurons (purple). Cb, cerebellum; DLM, dorsolateral thalamic nucleus, medial part; HVC, high vocal center; LMAN, nidopallial nucleus of songbirds (lateral magnocellular nucleus of the anterior nidopallium; part of the ventral pallium); MSt, medial striatum; RA, nucleus robustus of the arcopallium (part of ventral pallium).

Limbic versus motor parts of the basal ganglia: Evidence suggests that the limbic part is prevalent in amphibians, reptiles, and birds As noted previously, the BG in tetrapods includes a ventromedial portion involved in motivational/emotional (viscerolimbic-like) functions and a dorsolateral part involved in

executive and motor functions. Both parts are also involved in cognitive functions such as learning. In mammals, the transition between both parts is gradual rather than sharp, and this also appears to be the case in birds and possibly other tetrapods.

The motor/executive BG part shows the greatest development in mammals, and this is related to the development of the cerebral cortex (particularly the neocortex) and the sensorimotor lemnothalamus. In birds, however, a large part of the striatum (perhaps most of the avian medial striatum in songbirds and at least half of it in chicken or pigeon) appears to show viscerolimbic-type features and is involved in motivation, reward, and learning. This specific striatal part includes area X of songbirds, which has been proposed to be a specific part of the viscerolimbic BG specialized for song. This is related to the fact that in many birds, and especially songbirds, the ventrolateral pallium (including the nidopallium, mesopallium, and arcopallium) shows the greatest development, but the dorsal pallium (called hyperpallium or Wulst) shows a relatively moderate to small size (owls are an exception to this general rule since they show major development of the dorsal pallium). A large part of the medial striatum of birds receives input from the ventrolateral pallium (particularly from the pallial amygdala and nidopallium, which in songbirds includes LMAN and the high vocal center; **Figure 8**). In mammals, the homologous ventrolateral pallium – based on both embryological and molecular criteria – gives rise to the claustrum, pallial amygdala, piriform cortex, and perhaps the agranular part of the insular cortex, but not the neocortex. The latter derives from the dorsal pallium. The avian dorsal pallium (called hyperpallium) projects only to the lateralmost part of the medial striatum and to the lateral striatum, which appear to represent the somatomotor striatum. The lateral striatum (as the dorsal pallium) is particularly small in songbirds and budgerigars and, as a whole, the somatomotor striatum is relatively small compared to the viscerolimbic striatum. The situation may be different in owls, which possess a highly developed dorsal pallium (hyperpallium). This does not undermine the role of the BG and the pallium in birds. Birds do show a large and highly complex telencephalon which is involved in highly complex behaviors (e.g., song and vocalization learning and production; most mammals are unable to do this). The development of the ventrolateral pallium in birds has involved the appearance of numerous sensory, associative, and even motor centers without precedent in the ventral pallium of vertebrates and that cannot be functionally compared to the claustrum or pallial amygdala of mammals. This has involved the evolution of complex systems and pathways through the BG, which are involved in highly sophisticated functions such as song/vocalization learning. Birds therefore comprise a vertebrate group useful for studying BG evolution and its role in complex behaviors such as song and vocalization.

The telencephalon of modern reptiles shows an organization relatively similar to that of birds, with a large ventrolateral pallium (DVR) and a small dorsal pallium. This is also the case in amphibians, although the pallial and subpallial structures are much smaller in these animals than in reptiles. It is therefore likely that the reptilian and amphibian BG are mostly devoted to viscerolimbic functions.

The Basal Ganglia of the First Vertebrates: Evidence from Fish

Genetic embryological evidence, together with neurochemistry evidence and some connectivity patterns in adult teleost and cartilaginous fish, suggests the existence in jawed fish of BG comparable (homologous) to those of tetrapod vertebrates. The BG of jawed fish appear to contain striatal and pallidal subdivisions, expressing either only Distal-less or both Distal-less and Nkx2.1/Titf-1. The genetic evidence argues against the proposal of a mixed striatopallidal field in these animals. The striatal division is the so-called dorsal part of the ventral telencephalic area (Vd), whereas the pallidal division appears to be the ventral part of the ventral telencephalic area (Vv) (**Figure 9**). Supporting this, Vv is rich in cells containing parvalbumin (which is typical of the pallidal projection neurons in tetrapods). However, the pallidal field of jawed fish – derived from the Nkx2.1/Titf-1-expressing domain – appears to include mixed features of ventral pallidum, pallidal septum, and more basal telencephalic areas of tetrapods (all of which derive from the Nkx2.1-expressing domain).

Some neurochemical and connectivity features of the striatum of jawed fish resemble those of the tetrapod striatum. For example, the striatum of jawed fish (Vd) possesses GABAergic neurons containing SP or ENK that project to the pallidum (Vv). It also projects to pretectal and tegmental nuclei that, in turn, appear to project to the tectum. This suggests the existence of BG pathways to the tectum in jawed fish comparable to those described in tetrapods. Moreover, the striatum is reciprocally connected with a dopaminergic center of the diencephalic tegmentum comparable to a rostral part of the ventral tegmental area–substantia nigra (A10–A9 catecholamine groups) of tetrapods and also receives noradrenergic input from the locus coeruleus. This suggests that at least these basic aspects characterized the BG of the first jawed vertebrates. In addition, the striatum receives input from the pallium (from the field that appears comparable to the ventrolateral

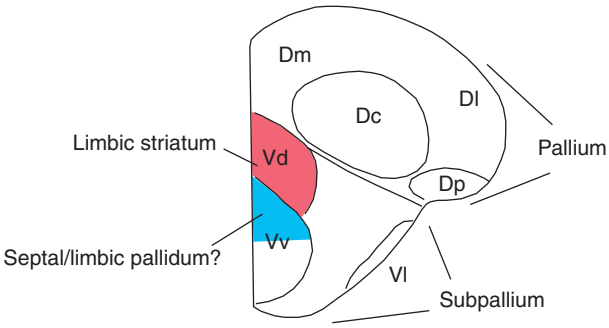

Figure 9 Schematic drawing of a frontal section through the telencephalon of zebra fish (a teleost fish) showing the striatal (Vd) and putative pallidal (Vv) subdivisions. As in tetrapods, these two divisions derive, respectively, from subpallial domains expressing either Dlx1/2 or both Dlx1/2 and Nkx2.1/Titf-1 transcription factors. Similarly to mammals, the subpallial derivatives of zebra fish are rich in GABAergic cells. The striatum (red) contains peptidergic cells that project to the pallidum, and it is reciprocally connected with a dopaminergic cell group comparable to the A9–A10 catecholamine cell groups of tetrapods (substantia nigra and ventral tegmental area). The pallidum is rich in parvalbumin-containing cells but shows mixed connections, resembling the ventral pallidum (perhaps the dorsalmost aspect of Vv; blue) and pallidal septum of tetrapods. Dc, dorsal telencephalic area (pallium), central part; Dl, dorsal telencephalic area (pallium), lateral part; Dm, dorsal telencephalic area, medial part; Dp, dorsal telencephalic area (pallium), posterior part; Vl, ventral telencephalic area (subpallium), lateral part.

pallium of tetrapods, which includes the pallial amygdala and olfactory areas), thalamus, and several sensory areas in the hypothalamus and brain stem (including gustatory and visceral areas). The pattern of connections suggests that the BG in these animals may primarily be involved in viscerolimbic functions. Nevertheless, data suggest a role in motor functions (e.g., the observation of 'parkinsonian symptoms' following lesion to the dopaminergic systems).

In addition, the striatum of jawed fish shows direct projections to the pallium, which is not typical of the tetrapod striatum. This raises doubts about the homology of this striatal field of fish with the BG striatum of tetrapods, and it opens the possibility of homology with the striatal part of the septum (known to project to the pallium in tetrapods). However, evidence indicates that the striatum in reptiles and birds shows scarce direct projections to the pallium (including the ventrolateral pallium). This raises the possibility that perhaps such direct striatopallial projections were common in ancestral jawed vertebrates but were gradually lost during evolution within the tetrapod radiation.

The pallidum: A novelty of jawed vertebrates Comparative genetic embryological evidence from zebra fish (a teleost fish) and lamprey (a jawless fish close to the origin of vertebrates) indicates a major difference with regard to subpallial development and divisions, which have important implications for understanding the organization of the BG in the first vertebrates and for understanding their evolution. The subpallium of zebra fish, like that of tetrapods, has two major subdivisions: a subdivision expressing Dlx/Distal-less family genes that gives rise to the striatum and a subdivision expressing Dlx/Distal-less and Nkx2.1/Titf-1 homeobox genes that gives rise to the pallidum proper (in addition to a pallidal part of the septum and other cell groups of the basal telencephalon). In contrast, Nkx2.1/Titf-1 is not expressed in the telencephalon in lamprey, meaning that this jawless fish lacks a pallidal subdivision in the subpallium (Figure 10). This is in agreement with the apparent absence of a pallidum in adult lampreys based on neurochemical and connectivity studies. This suggests that the BG of the first vertebrates was only striatal-like, and the pallidum appeared during or after the transition from jawless to jawed vertebrates. This was apparently related to the appearance of a novel expression domain of Nkx2.1/Titf-1 in the basal telencephalon. An important question that remains to be answered concerns what factor(s) triggered this new expression in the telencephalon, although evidence suggests that this may partially be related to sequence changes in the regulatory region (promoter/enhancer) of Nkx2.1/Titf-1.

No pallidum, no immigrant interneurons in the striatum? In tetrapods, the Nkx2.1-expresssing domain of the basal telencephalon (including the pallidal and other basal subdivisions) appears to be the source of the interneurons of the striatum. This has been shown for the striatal interneurons containing acetylcholine, parvalbumin, and nitric oxide/somatostatin/neuropeptide Y. The absence of a domain expressing Nkx2.1/Titf-1 in the basal telencephalon of lamprey suggests that the striatum in this jawless fish lacks immigrant interneurons comparable to those of tetrapods. These may have started to appear after the evolution of the Nkx2.1-expressing pallidum, involving a tangential migration of some cells from the pallidal subdivision into the striatum. However, this may have been a gradual process since, of the three major interneuron subtypes present in the striatum of amniotes (mammals, birds, and reptiles), only two subtypes are present in the striatum of amphibians (cholinergic and nitrergic/somatostatin-containing cells), and only one subtype (possibly nitrergic/somatostatin-containing cells) appears to be present in the striatum of jawed fish. In contrast to this view, one study has shown the presence of extremely few cholinergic neurons in the striatum of lamprey, raising questions

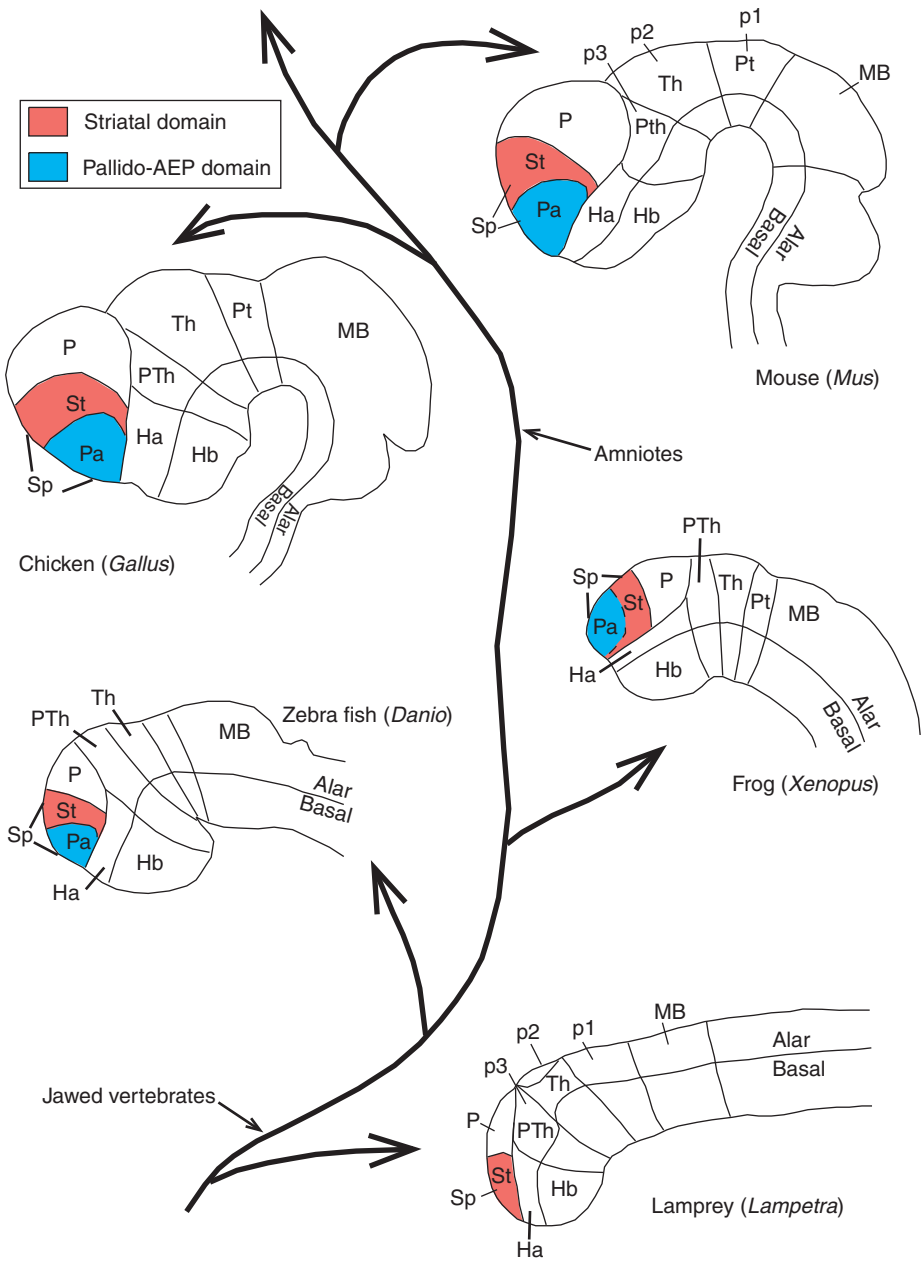

Figure 10 Schematic drawings of the embryonic brain (lateral view) of different vertebrate groups showing the striatal (St) and pallido-pedunculo-preoptic (Pa-AEP) domains in the subpallium, based on the expression of either Dlx/Distal-less family genes or both Dlx/Distal-less and Nkx2.1/Titf-1. Note that Nkx2.1/Titf-1 is not expressed in the embryonic telencephalon of lamprey, which is correlated with the apparent absence of a pallidum in adult animals. Ha, alar hypothalamus; Hb, basal hypothalamus; MB, midbrain; P, pallium; Pt, pretectum; PTh, prethalamus (ventral thalamus); Sp, subpallium; Th, thalamus (dorsal thalamus).

about the origin of these cells. On the other hand, the striatum of fish contains TH+ cells, and at least some of these may be native striatal interneurons.

The case of TH+ cells in the striatum of fish: Do they come from a dLGE-like subdivision? A striking feature of the striatum of teleost and cartilaginous fish is that it contains neurons expressing tyrosine hydroxylase (TH; the rate-limiting enzyme of catecholamines) and dopamine (DA). This is a feature that has not been observed in the striatum of amphibians, reptiles, or birds. In recent years, however, with the development of *in situ* hybridization techniques for mRNA detection, it has been possible to identify a subpopulation of TH+/DA+ cells in the striatum of rodents and primates which appear to represent interneurons. As such, it would be interesting to reinvestigate their existence in other vertebrate groups. In

mammals, the dLGE is known to produce TH+/DA+ interneurons that migrate tangentially to the olfactory bulb. It is likely that this same embryonic subdivision is the origin of the striatal cells containing TH and DA. Could the TH+/DA+ cells of the fish striatum originate in an embryonic dorsal striatal subdivision comparable to dLGE of mammals? To answer this question, it would be necessary to analyze the expression of transcription factors typical of that subdivision, such as Pax6, and to study experimentally whether this would produce the striatal TH+ neurons. In the meantime, it is worth noting that the striatum of cartilaginous fish appears to have two main subdivisions, and the dorsal one is rich in TH+ cells and shows a close relationship to the olfactory bulb, resembling the relation between dLGE and the rostral migratory stream into the olfactory bulb observed in mammals.

See also: Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia and Oculomotor Control; Basal Ganglia: Internal Organization; Birdsong Learning; Cognition: Basal Ganglia Role; Deep Brain Stimulation and Movement Disorder Treatment; Procedural Learning: Striatum; Striatum: Internal Physiology; Transcription and Reward Systems.

Further Reading

Brox A, Puelles L, Ferreiro B, and Medina L (2003) Expression of the genes GAD67 and Distal-less-4 in the forebrain of *Xenopus laevis* confirms a common pattern in tetrapods. *Journal of Comparative Neurology* 461: 370–393.

Carrillo GD and Doupe AJ (2004) Is the songbird area X striatal, pallidal, or both? An anatomical study. *Journal of Comparative Neurology* 473: 415–437.

Farries MA, Ding L, and Perkel DJ (2005) Evidence for 'direct' and 'indirect' pathways through the song system basal ganglia. *Journal of Comparative Neurology* 484: 93–104.

Folgueira M, Anadon R, and Yanez J (2004) An experimental study of the connections of the telencephalon in the rainbow trout (*Oncorhynchus mykiss*): I. Olfactory bulb and ventral area. *Journal of Comparative Neurology* 480: 180–203.

Gonzalez A, Lopez JM, Sanchez-Camacho C, and Marin O (2002) Regional expression of the homeobox gene NKX2–1 defines pallidal and interneuronal populations in the basal ganglia of amphibians. *Neuroscience* 114: 567–575.

Graybiel AM and Saka E (2004) The basal ganglia and the control of action. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences III*, 3rd edn., pp. 495–510. Cambridge, MA: MIT Press.

Marin O, Smeets WJAJ, and Gonzalez A (1998) Evolution of the basal ganglia in tetrapods: A new perspective based on recent studies in amphibians. *Trends in Neuroscience* 21: 487–494.

Mezey S and Csillag A (2002) Selective striatal connections of midbrain dopaminergic nuclei in the chick (*Gallus domesticus*). *Cell and Tissue Research* 308: 35–46.

Osorio J, Mazan S, and Retaux S (2005) Organisation of the lamprey (*Lampetra fluviatilis*) embryonic brain: Insights from LIM-homeodomain, Pax and hedgehog genes. *Developmental Biology* 288: 100–112.

Pombal MA, El Manira A, and Grillner S (1997) Organization of the lamprey striatum – Transmitters and projections. *Brain Research* 766: 249–254.

Puelles L, Kuwana E, Puelles E, et al. (2000) Pallial and subpallial derivatives in the embryonic chick and mouse telencephalon, traced by the expression of the genes Dlx-2, Emx-1, Nkx-2.1, Pax-6, and Tbr-1. *Journal of Comparative Neurology* 424: 409–438.

Reiner A, Medina L, and Veenman CL (1998) Structural and functional evolution of the basal ganglia in vertebrates. *Brain Research Brain Research Reviews* 28: 235–285.

Strick PL (2004) Basal ganglia and cerebellar circuits with the cerebral cortex. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences III*, 3rd edn., pp. 453–461. Cambridge, MA: MIT Press.

van den Akker WMR, Brox A, Puelles L, Durston AJ, and Medina L (2007) Comparative functional analysis provides evidence for a crucial role for the homeobox gene Nkx2.1/Titf-1 in forebrain evolution. *Journal of Comparative Neurology* 506: 211–223.

Wullimann MF and Mueller T (2004) Teleostean and mammalian forebrains contrasted: Evidence from genes to behavior. *Journal of Comparative Neurology* 475: 143–162.

Basal Ganglia: Functional Models of Normal and Disease States

M A Kliem and T Wichmann, Emory University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Functional Anatomy of the Basal Ganglia

General Circuit Overview

The term 'basal ganglia' encompasses the striatum, the external and internal segments of the globus pallidus (GPe and GPi, respectively), the pars compacta and the pars reticulata of the substantia nigra (SNc and SNr, respectively), and the subthalamic nucleus (STN). These structures are components of larger cortico–subcortical reentrant pathways (see Figure 1). Anatomical and physiological studies have shown that the basal ganglia, thalamus, and cortex are linked to form a system of segregated parallel circuits, which permit the basal ganglia to simultaneously participate in different functions, such as movement, reward, and procedural learning. Five major circuits have been identified, and were named after the presumed functions of the cortical areas they involve: a 'motor' circuit, centered on the precentral cortical motor fields; an 'oculomotor' circuit, originating from the frontal and supplementary eye fields; two prefrontal circuits, involving the dorsolateral prefrontal and lateral orbitofrontal cortices; and a 'limbic' circuit, centered on the anterior cingulate and medial orbitofrontal cortices. Each of these circuits may be composed of several subcircuits.

Although these circuits originate and terminate in different cortical areas, and occupy different domains in the basal ganglia and the thalamus, they share certain anatomical features, which are outlined in the simplified schemes shown in Figures 1 and 2. Each circuit arises from a separate cortical area and passes through different portions of the striatum, GPe, GPi, SNr, and STN, and parts of the ventrolateral, ventral anterior, and mediodorsal (VL, VA, and MD, respectively) thalamus. In each of these loops, the striatum and STN serve as 'input' stages of the basal ganglia, whereas GPi and SNr serve as 'output' stages, sending inhibitory projections to thalamus and brain stem. The thalamus projects back to the cortical areas of origin of the respective circuit to complete the reentrant loops.

Motor Circuit

Of the different circuits passing through the basal ganglia, the motor circuit has received the most attention, because abnormalities in this circuit may play a major role in the development of the motor abnormalities in Parkinson's disease and other movement disorders. The motor circuit arises from the precentral motor fields and engages the putamen, the 'motor' portions of GPe, STN, and GPi/SNr, as well as parts of the VL thalamus. The circuit is organized into somatotopic regions that contain groups of neurons which discharge either in conjunction with the preparation for or the execution of limb movements. The motor circuit may also play a role in motor functions, ranging from the scaling of kinematic parameters such as the amplitude or velocity of movement, the initiation or execution of internally triggered movements, the sequencing of movements, and the switching between movements or motor programs, to a role in procedural learning.

Direct and Indirect Basal Ganglia Pathways

Two major projection systems link the striatum to the output structures of the basal ganglia (Figure 2, left panel): a monosynaptic inhibitory (γ-aminobutyric acid (GABA)ergic) 'direct' pathway between striatum and GPi/SNr, and a polysynaptic 'indirect' pathway which involves GPe and STN. In the indirect pathway, the projections between the striatum and GPe, and between GPe and the STN are inhibitory (GABAergic), while the STN–GPi/SNr pathway is excitatory (glutamatergic). In addition, a reciprocal inhibitory connection between both segments of the globus pallidus which circumvents the STN is also considered as part of the indirect pathway.

The striatal neurons that give rise to the direct and indirect pathways differ in several ways which may strongly affect their functions. One difference is that in rodent experiments the two populations of striatal neurons appear to receive inputs from different populations of cortical neurons. Furthermore, direct pathway neurons receive stronger inputs from the intralaminar nuclei of the thalamus (the centromedian and parafascicular nucleus; CM/Pf), relative to indirect pathway neurons. Inputs from other thalamic nuclei (such as feedback circuits emanating from VA/VL) target direct and indirect neurons alike.

There are also biochemical differences between direct and indirect pathways. Although both groups of neurons are GABAergic, the direct pathway neurons express substance P as a cotransmitter, while indirect pathway neurons express enkephalin instead. Another important difference is that the activity of the direct and indirect pathways appears to be differentially modulated by dopamine, released from terminals of the nigrostriatal projection. Dopamine has been shown to inhibit the activity of striatal neurons

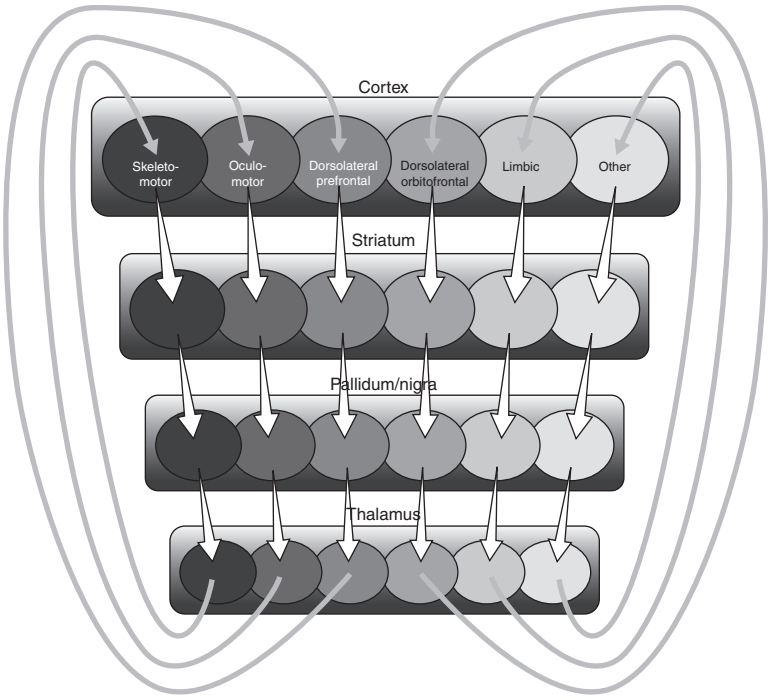

Figure 1 Schematic view of the large-scale anatomic organization of related areas of cortex, basal ganglia, and thalamus. All three structures form reentrant circuits which occupy anatomically distinct areas.

that give rise to the indirect pathway via activation of dopamine D2 receptors, and to facilitate the activity of direct pathway neurons via activation of D1 receptors.

Activation of the direct pathway by cortical inputs would inhibit the activity of the output nuclei GPi and SNr, and thereby disinhibit thalamocortical projection neurons. In contrast, activation of striatal neurons that give rise to the indirect pathway would have a net excitatory effect on GPi/SNr activity and would thereby act to inhibit thalamocortical neurons (Figure 2, left panel). Dopamine may have a substantial effect on the balance of activity between direct and indirect pathways. Through its differential effects, exerted via activation of different dopamine receptor types, a reciprocal relationship may exist between striatal dopamine levels and basal ganglia output. Increased release of dopamine may activate the direct pathway and inactivate the indirect pathway so that basal ganglia output from GPi/SNr to the thalamus is reduced. Conversely, reduced release of dopamine would result in disinhibition of the indirect pathway and reduced facilitation of the direct pathway, leading to increased basal ganglia output. There is also evidence that dopamine may be released at extrastriatal sites, including the pallidum, STN, and SNr. For instance, dopamine released from local dendrites of SNc neurons may activate D1 receptors on terminals of the striatonigral projection. This may then result in increased local GABA release in the SNr and subsequent inhibition of the activity of SNr neurons.

The anatomic model described here can be applied to generate a description of the involvement of the basal ganglia in the control of movements. According to this model, movements are initiated at the cortical level. Cortical inputs to the basal ganglia may then transiently activate both direct and indirect pathways. A phasic increase of inhibition of GPi via the monosynaptic direct pathway may lead to a brief disinhibition of thalamocortical neurons, which may then allow increased activity in thalamic neurons projecting to specific cortical areas, thus facilitating the movement. Activation of the indirect pathway may increase GPi/SNr output and may thereby have the effect of inhibiting or stabilizing the activity in the thalamus and brain stem. This may suppress potential competing movements, or terminate ongoing movements.

While such models are attractive due to their simplicity, several new findings suggest that views based purely on the anatomic connections are not accurate. For instance, while in the older models the overall effect of phasic activation of corticostriatal projection neurons would entirely depend on the interplay between the direct and indirect pathways,

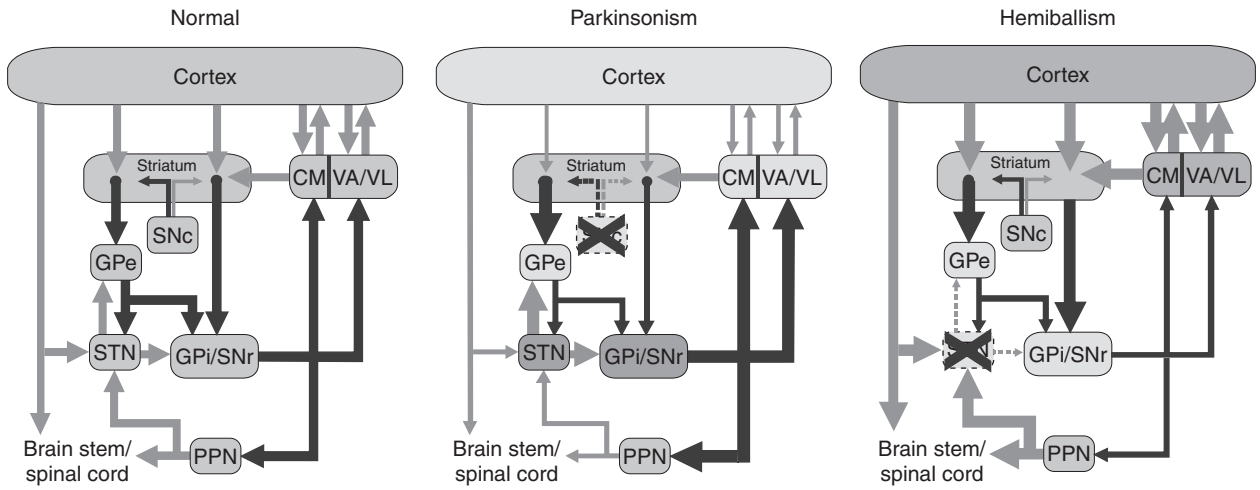

Figure 2 Activity in the basal ganglia–thalamocortical circuitry (normal) and overall activity changes occurring in parkinsonism and hemiballism. The basal ganglia circuits include striatum, external and internal globus pallidal (GPe, GPi, respectively) segments, substantia nigra pars reticulata and pars compacta (SNr, SNc, respectively), and subthalamic nucleus (STN). Cortical input reaches the basal ganglia via the corticostriatal and the cortico–subthalamic connections. Basal ganglia output is directed toward the centromedian (CM) and ventral anterior/ventral lateral (VA/VL) nucleus of thalamus, as well as the pedunculopontine nucleus (PPN). Excitatory connections are shown in gray, inhibitory pathways in black. In the parkinsonian state, loss of dopamine in the striatum results in activity changes throughout the basal ganglia (represented by the thickness of the connecting arrows), leading to increased basal ganglia output to the thalamus. In hemiballism, damage to the STN results in reduced excitatory input to the basal ganglia output nuclei, leading to a cascade of other changes throughout the entire circuitry.

it now seems that basal ganglia output may also be shaped by nonlinear effects that go beyond straight excitation or inhibition. For example, recent studies of the interaction between GPe and STN have shown that release from GABAergic inhibition does not simply allow STN neurons to return to their baseline activity, but may also result in significant transient rebound excitation of such neurons. Similarly, glutamatergic inputs to basal ganglia neurons may not only lead to activation of recipient neurons, but may also alter firing patterns, such as burst firing or oscillatory discharge. The impact of these transient changes on the overall functioning of the basal ganglia is not clear at this time.

At a yet higher level of complexity, recent work has suggested that some of the basal ganglia functions can only be understood in terms of coordinated interactions of large neuronal ensembles in cortex and basal ganglia. For instance, there is emerging evidence that ensemble activity involving large populations of basal ganglia neurons may play a prominent role in the transmission and processing of cortical oscillatory activity.

Pathophysiology of Motor Hypo- and Hyperkinetic Movement Disorders

Most movement disorders of basal ganglia origin are believed to arise from dysfunction of the 'motor' circuit, resulting in alterations of GPi and SNr output. Clinically, these disorders are dominated by either poverty of movement (hypokinetic disorders) or an excess of movement (hyperkinetic disorders). Hypokinetic disorders (e.g., Parkinson's disease) feature an impairment of movement initiation (akinesia) and reductions in the amplitude or velocity of voluntary movements (bradykinesia). By contrast, hyperkinetic disorders (e.g., chorea or ballism) are characterized by involuntary movements.

Hypokinetic Disorders

The most common hypokinetic disorder is Parkinson's disease, which is characterized by akinesia, bradykinesia, a 5 Hz tremor at rest, and muscular rigidity. The most salient pathological sign is the finding of degeneration of the SNc neurons that give rise to the dopaminergic nigrostriatal projection. The earliest and most prominent dopamine loss is found in the motor portion of the striatum, the putamen. Studies of the neuronal activity in the basal ganglia in primates which were rendered parkinsonian by treatment with the dopaminergic neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) have suggested the pathophysiological changes that are shown in Figure 2 (middle panel). According to this model, loss of dopamine in the striatum appears to have the opposite effect on the activity in the indirect and direct pathways, eventually leading to an overall increase in activation of GPi and SNr via the indirect

pathway, and reduced inhibition of GPi and SNr via the direct pathway.

Perhaps related to striatal dopamine loss, there are also morphological changes in striatal neurons in the Parkinsonian brain. These changes may further influence the balance between direct and indirect pathways. Thus, in postmortem tissue from patients with Parkinson's disease, the overall size of the dendritic tree of striatal cells and the number of specialized dendritic membrane compartments (spines) on striatal projection neurons that receive cortical and other inputs were significantly reduced. Similar changes, particularly affecting the indirect pathway, have also been documented in rodent models of the disorder.

The activity of indirect pathway neurons downstream from the striatum is also altered. For instance, GPe activity is reduced in MPTP-treated animals, whereas discharge rates in STN and GPi/SNr are increased. An important piece of evidence in favor of the hypothesis that STN overactivity directly contributes to the development of parkinsonism is that lesions of this nucleus ameliorate all of the major parkinsonian motor signs.

Imaging studies of cortical function have shown that the activity changes in the basal ganglia in parkinsonism are accompanied by reduced activity in related cortical motor areas. In addition, there is a task-dependent increase of activation in some cortical areas which are normally not activated under these circumstances. The aberrant activation of these areas may be the product of compensatory mechanisms which may add to the inefficiencies of movement generation in parkinsonism.

In addition to the rate changes, a variety of changes in discharge patterns may also play a role, ranging from increased synchronization between neighboring neurons to oscillatory burst patterns (Figure 3). Recent reviews of the pathophysiology of parkinsonism have emphasized synchronized oscillations in the basal ganglia, based largely on recordings of local field potentials (LFPs) in patients with Parkinson's disease undergoing functional neurosurgery. These recordings have suggested that abnormal beta-range oscillations develop throughout the entire basal ganglia–thalamocortical circuitry, and these may play a major role in the production of parkinsonian signs. At the same time, 'prokinetic' normal oscillations in the gamma range appear to be suppressed in the parkinsonian state. Several mechanisms have been described that may account for oscillatory and non-oscillatory changes in phasic discharge properties, including pacemaker mechanisms that result from the interaction of different nuclei (for instance, GPe and STN), membrane properties of individual basal ganglia neurons that may predispose these cells to the development of recurring bursts, or altered cortical activity patterns that may directly drive basal ganglia neurons.

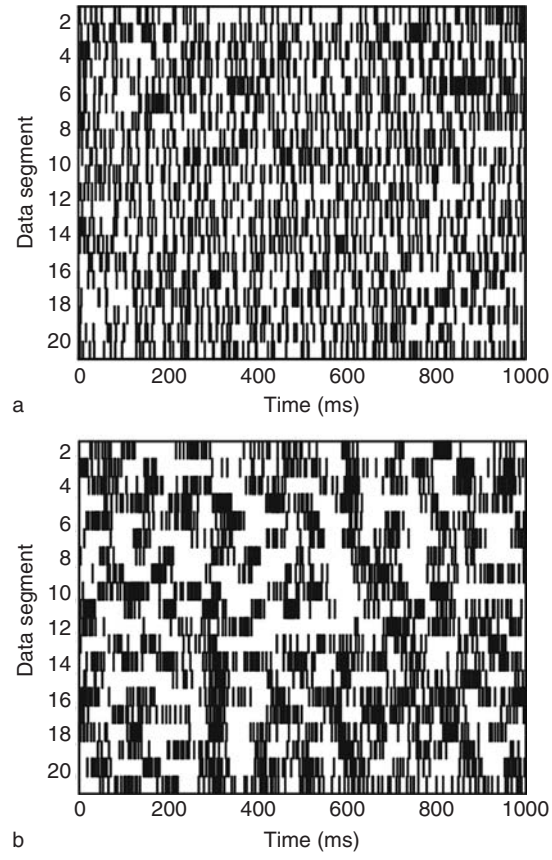

Figure 3 Raster displays of spontaneous neuronal activity recorded in the internal segment of the globus pallidus in normal (a) and Parkinsonian (b) primates. Shown are 20 consecutive 1000 ms segments of data in each panel. Each of the short vertical lines in the figure represents a single action potential. In addition to an increase in activity, there are also obvious changes in the firing patterns, with a marked prominence of bursts and oscillatory discharge patterns in the parkinsonian state.

The link between abnormal network activity in the basal ganglia–thalamocortical circuitry and parkinsonian motor signs has not been firmly established. It appears that some aspects of akinesia and tremor are due to abnormal oscillatory bursting in the basal ganglia output nuclei. Bradykinesia may result from more specific disturbances of cortical processing by excessive phasic inhibition of thalamocortical circuits. Recent evidence indicates that damage to sites which are anatomically linked to the basal ganglia, but have historically not been considered to be part of these structures (in particular, the pedunculopontine nucleus), may also play a role in the development of akinesia.

Parkinsonism is also associated with nonmotor abnormalities, such as oculomotor disturbances, autonomic dysfunction, depression, anxiety, sleep disturbances, impaired visuospatial orientation, and cognitive abnormalities. It is likely that some of these impairments are due to abnormal discharge in nonmotor circuits of the basal ganglia, which may be affected by dopamine loss in much the same way as is the motor circuit. In addition, degeneration of neurons outside of the basal ganglia has been described to occur in Parkinson's disease. Thus, in addition to nigral damage, postmortem examinations of brains from parkinsonian patients show substantial damage in pigmented brain stem nuclei, the olfactory tubercle, and cortex.

Hyperkinetic Disorders

In contrast to hypokinetic diseases, hyperkinetic disorders appear to be associated with abnormally low basal ganglia output from GPi and SNr. Reduced basal ganglia output may arise from a shift of the balance of activity between direct and indirect pathway activity toward the direct pathway. Thalamocortical neurons under such conditions may become more responsive to cortical inputs or exhibit an increased tendency to discharge spontaneously, thus leading to involuntary movements. The involuntary movements seen in ballism and Huntington's disease, as well as those induced by dopaminergic drugs, may all arise from this mechanism.

The term 'hemiballism' refers to involuntary proximal limb movements on one side of the body; these are most frequently due to damage restricted to the contralateral STN (see Figure 2, right panel). The STN damage, in turn, is thought to lower the STN's excitatory drive on GPi. These changes have been directly demonstrated in primate and rodent experiments, in which STN lesions result in a significant reduction of tonic discharge in GPi and a decrease in the phasic responses of GPi neurons to limb displacement.

In early stages of Huntington's disease, degeneration of striatal neurons that give rise to the indirect pathway is thought to result in excessive inhibition of STN neurons and abnormally low GPi activation. This, in turn, may result in proximal and distal involuntary movements. In later stages of the disease, the degeneration spreads to involve direct pathway neurons, and the associated phenotype changes in many patients are reminiscent of parkinsonism. Perhaps as a consequence of extensive neuronal degeneration of nonmotor circuits, most patients also develop cognitive and emotional disturbances.

Dystonia, a movement disorder which is often associated with basal ganglia lesions, is dominated by abnormal twisting movements and postures. The pathophysiology of dystonia is poorly understood, most likely due to the fact that dystonia can be a symptom of a variety of diseases with very different pathophysiologic backgrounds. Recordings in dystonic patients suggest that discharge rates in GPi are reduced, likely due to overactivity along the direct pathway. Similar to parkinsonism, however, the level of synchrony of spiking activity of neighboring neurons is increased in dystonia, and there is increased bursting throughout the basal ganglia. Recently, oscillatory LFP changes have also been identified in this condition.

Surgical Treatments

As a result of the better understanding of the pathophysiologic abnormalities that may underlie movement disorders such as Parkinson's disease, neurosurgical interventions in the basal ganglia have been reintroduced as effective treatments for these conditions. Initially, neurosurgeons performed almost exclusively ablative procedures as treatments for movement disorders, including procedures such as pallidotomies or thalamotomies. In the meantime, however, the focus has shifted to deep brain stimulation therapy in which electrical leads are implanted into the basal ganglia or thalamus and are used for chronic high-frequency stimulation. Conceivably, both types of procedures work through elimination of disruptive basal ganglia outputs. While therapeutic lesion effects in Parkinson's disease are perhaps understandable in terms of a reversal of the described rate changes in the basal ganglia, deep brain stimulation may target predominately firing pattern abnormalities of the stimulated circuitry.

Nonmotor Symptoms Related to Basal Ganglia Dysfunction

Pathology in the nonmotor circuits of the basal ganglia may be important not only for the development of some of the nonmotor signs accompanying Parkinson's disease and other movement disorders, but also for several unrelated behavioral and psychiatric disorders. Because anatomical and functional details regarding these circuits and the disease associated with their dysfunction are not completely worked out, they are only briefly mentioned here.

The dorsolateral prefrontal circuit originates in Brodmann areas 9 and 10 of the cortex, and its subcortical route involves the head of the caudate nucleus, the dorsomedial portion of GPi, and the rostral SNr, as well as portions of the VA and MD nuclei of the thalamus. This circuit has been implicated

broadly in 'executive functions' – for instance, the organization of behavioral responses to complex problems or the use of verbal skills in problem solving. Damage to portions of this circuit is associated with a variety of behavioral abnormalities related to these cognitive functions.

The dorsolateral orbitofrontal circuit involves the lateral prefrontal cortex, the ventromedial caudate nucleus, and associated areas of GPi/SNr and thalamus. This circuit appears to play a role in the mediation of empathetic and socially appropriate behaviors. Consequently, damage to this area is associated with irritability, emotional lability, failure to respond to social cues, and lack of empathy.

The limbic circuit arises in the anterior cingulate gyrus of cortex and involves the ventral striatum (which also receives 'limbic' input from hippocampus, amygdala, and entorhinal cortices). The ventral striatum then projects to the ventral and rostromedial GPi and rostrodorsal SNr. From there the pathway continues to neurons in the paramedian portion of the MD nucleus of the thalamus, and back to the anterior cingulate cortex. This circuit may be important in motivated behavior and may convey reinforcing stimuli to diffuse areas of the basal ganglia and cortex.

In a general sense it appears that the psychiatric disorders resulting from dysfunction of the nonmotor circuits parallel those seen in the motor circuit. One of the best studied psychiatric disorders in this regard is obsessive–compulsive disorder. Functional imaging studies have demonstrated abnormalities in the activity of limbic basal ganglia–thalamocortical projection systems in patients with this disorder. Reminiscent of the situation in Parkinson's disease, patients with obsessive–compulsive disorder appear to access 'unusual' brain regions (such as portions of the temporal cortices) when they are confronted with stimuli that call for recruitment of corticostriatal systems.

Tourette syndrome, a disorder in which obsessive–compulsive symptoms occur in combination with vocal or motor tics, is likely associated with abnormalities in the limbic and motor basal ganglia–thalamocortical circuitry. Not unexpectedly, the motor symptoms in this disease are reflected in additional changes in brain activation, particularly in the sensorimotor cortices and supplementary motor area.

Finally, the basal ganglia may also be involved in the development of depression, even in the absence of motor dysfunction. Thus, positron emission tomography studies in patients with major depression have shown reduced metabolic activity or blood flow in cortical portions of the limbic system, especially the orbitofrontal cortex and prefrontal cortices, in conjunction with similarly reduced activity in related areas of the basal ganglia, particularly the ventral striatum and caudate nucleus. In addition, stimulation of the limbic portion of the STN, or the SNr in parkinsonian patients who are treated with deep brain stimulation, can induce dramatic mood changes, ranging from depression to mania.

Conclusion

Our understanding of the role of the basal ganglia in normal movement remains very limited. Much more is known about abnormalities associated with basal ganglia disorders. By contributing to the reemergence of neurosurgical and other interventions as treatments for movement disorders, models of basal ganglia dysfunction have already had a major impact in the treatment of diseases such as Parkinson's disease or dystonia.

The models described herein continue to evolve. Gradually, the view of the basal ganglia as a set of nuclei by which cortical activity is simply relayed back to cortex is replaced by more general organizational schemes in which single-cell spiking and large-scale oscillatory and nonoscillatory activities of local or spatially distributed ensembles of neurons play a role.

See also: Animal Models of Parkinson's Disease; Basal Ganglia: Internal Organization; Cell Replacement Therapy: Parkinson's Disease; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Deep Brain Stimulation; Deep Brain Stimulation and Movement Disorder Treatment; Dopamine; Huntington's Disease; Parkinsonian Syndromes; Tourette's Disorder.

Further Reading

DeLong MR (2000) The basal ganglia. In: Kandel ER, Schwartz JH, and Jessel TM (eds.) *Principles of Neural Science*, 4th edn., pp. 853–867. New York: McGraw-Hill.

Graybiel AM (2005) The basal ganglia: Learning new tricks and loving it. *Current Opinion in Neurobiology* 15: 638–644.

Hutchison WD, Dostrovsky JO, Walters JR, et al. (2004) Neuronal oscillations in the basal ganglia and movement disorders: Evidence from whole animal and human recordings. *Journal of Neuroscience* 24: 9240–9243.

Middleton FA and Strick PL (2000) Basal ganglia and cerebellar loops: Motor and cognitive circuits. *Brain Research Reviews* 31: 236–250.

Nutt JG and Wooten GF (2005) Clinical practice. Diagnosis and initial management of Parkinson's disease. *New England Journal of Medicine* 353: 1021–1027.

Smith Y, Raju DV, Pare JF, et al. (2004) The thalamostriatal system: A highly specific network of the basal ganglia circuitry. *Trends in Neuroscience* 27: 520–527.

Basal Ganglia: Habit

A M Graybiel, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The study of habits has a long and distinguished history in the field of psychology and has now emerged as a dynamic field of interest in neuroscience. Habits are familiar to all of us as repetitive behaviors that we have acquired and that we can perform almost without thinking. Habits considered in this everyday sense are usually unobtrusive and helpful parts of our daily behavior. Examples range from how we comb our hair or put on our shoes to how we drive to work or how we interact with friends. A wide range of behaviors that are variously called mannerisms and rituals also fall into this category and are justifiably linked to the study of habits. These behaviors bring the study of habits close to work on the effects of social and emotional factors on behavior. Under many circumstances, habits can become urgently repeated, dominant parts of our behavior and as such can be undesirable, maladaptive, or even destructive. Habits related to taking drugs of abuse (typically called 'habit forming' drugs) fall into this category, and so do some of the other repetitive behaviors that we to varying degrees lack control of. In psychology, this broad set of behaviors have all been considered as reflecting a degree of automatization of behavior and have been contrasted with conscious, flexible behaviors in pursuit of goals and projects.

All these behaviors have in one or another way been linked with activity in the basal ganglia and the neural circuits of which the basal ganglia are a part. This link to the basal ganglia has in turn generated interest in the possibility that the brain circuits underlying habit formation are also implicated in the genesis of the repetitive behaviors that characterize a range of neurologic and neuropsychiatric disorders, such as Tourette's syndrome and obsessive–compulsive disorder. Before considering these circuits, it is important to consider how scientists have tried to sharpen the definition of habits.

Definitions of Habit in Cognitive Neuroscience

Most definitions of habit in cognitive neuroscience begin with habit learning. In studies of humans, a major approach has been to consider habits as behaviors that can be learned and performed with no or low levels of awareness ('implicitly'), in contrast to the facts and episodes that are acquired with conscious awareness ('explicitly'). In this view, habits result from the activity of basal ganglia-based procedural learning and memory circuits, and these are contrasted with circuits for explicit learning and memory, represented by the hippocampus and medial temporal lobe. Hippocampal damage, resulting in amnesic syndromes, leaves motor learning relatively intact, along with many automatic, habitual behaviors. By contrast, damage to the basal ganglia (most often damage to the striatum) leaves explicit learning relatively intact but makes implicit learning difficult or impossible. This contrast is based on studies of patients with focal brain damage and also on observations of patients with disorders including Alzheimer's disorder and Parkinson's disease. The famous patient H.M., with massive hippocampal/medial temporal lobe damage, lost the ability to store new 'fact' memories but was able to learn new procedures – for example rotor pursuit. Such an amnesic patient can learn a probabilistic (implicit learning) task, and yet can have absolutely no awareness of having learned it. Such a clean-cut division of labor of hippocampal and basal ganglia functions is probably overdrawn, as some studies emphasize, but nonetheless makes the point that habits are mostly below the level of high-level attention and conscious representation. Human brain imaging studies have confirmed differential activation of these two systems and have emphasized their complementary, if not competitive, functions.

Definitions of Habits in Psychology

In parallel with this work on human explicit and implicit learning systems, studies in animals have shown a rather similar dissociation between the animal's remembering explicitly where it went on a previous trial and its robotlike going to the same place regardless of the previous choice. These studies, mainly employing navigational tasks in mazes learned by rats, have suggested a similar division between behaviors dependent on the striatum and behaviors dependent on the hippocampus or temporal lobe. But these studies, like those with humans, have pointed out that the division between explicit and procedural (habit) learning may not just follow anatomic subdivisions. For example, basal ganglia (striatal) regions anatomically linked to the hippocampus exhibit some of the characteristics of hippocampal (explicit learning)

circuits, not implicit (stimulus–response; S-R) learning circuits. An influential attempt to define habits operationally in experimental animals rests on the basic idea that the animals can either act in a purposeful way, orienting their activity toward an outcome or goal, or can act without regard to the goal. If goal directed, the behavior was called action–outcome (A-O). But if the behavior became automatic to the point that the outcome was no longer guiding it, the behavior was considered to occur automatically in an S-R, reflexive way, regardless of outcome.

The advantage of this way of defining habits is that behavioral tests can be devised to determine the degree to which outcome matters. If reward is devalued (e.g., by pairing the reward with malaise-inducing treatments out of the task, or by simple satiation), a habit should continue, but an A-O behavior should be sensitive to the devaluation and decrease or cease. The reward devaluation paradigm can thus be used in lesion studies to determine which brain regions are responsible for the sensitivity to reward devaluation and which are not, the latter being candidate sites in 'habit circuits.'

Procedural Learning and Skill Learning

Work on habit learning directly intersects with the large field of work on procedural learning, which focuses on the transition from novice to expert behavior that occurs as a procedure is practiced repeatedly. In experimental studies, reaction times (simple and serial) are a principal behavioral measure: When responses are made repeatedly to a sequence of stimuli, the reaction times fall even if the participant is unaware of the repeating sequence and the sequence is learned implicitly. To the degree that skill learning produces a highly repeatable motor behavior – as in riding a bicycle or hitting a good backhand in tennis – there is much commonality with habit learning. However, would riding a bike with skill be called a habit? Probably not, unless the behavior were triggered in a context, repeatedly, as though with a goal in mind (e.g., riding around the park every morning for exercise before going to work).

The acquisition of skills and procedures engages a large set of neural circuits, including motor and premotor regions of the neocortex, the cerebellum, and the basal ganglia. Careful study of the patterns of activation of these regions during the course of skill learning has demonstrated a characteristic progression of activity from anterior to more posterior parts of these cortical networks and related regions of the striatum and cerebellum. This signature of learning bears many similarities to the patterns of changing activity observed in attempts to study habit learning, although direct comparisons are hampered by the fact that most of the habit learning studies have been performed in rodents, and with notable exceptions, most of the motor skill learning studies have been done in humans and other primates.

Emotional Habits, Social Habits, and Rituals

The field of habit learning naturally links up with the study of emotional behavior and with social and cultural practices, as so many of these are ritualized. This broad field cannot be covered here, but it is important to emphasize that the acquisition of habits occurs not only for sensorimotor behaviors, but also for behaviors that require only modest skill but engage the emotional and social sides of behavior. The customs and practices of societies, groups, clubs, and individual couples can be as dominant as idiosyncratic habits can be in an individual. Imaging studies are beginning to open up this field to neuroscience. Of interest are findings suggesting that some of the basal ganglia-based pathways engaged in habit learning are also engaged in these social contexts, and thus that the social contexts may be fundamental to the process of habit learning in many situations. This work ranges from the study of courtship and mating behaviors to the appreciation of art. At the same time, models of reinforcement learning and neuroeconomics are increasingly being applied to behaviors that are social in nature, further tying together these lines of work.

Lesion Studies and Attempts to Define a 'Habit Circuit' in the Brain

Lesion studies have long implicated the striatum in the performance of habits. Lesions that damage the striatum disrupt habit performance, but lesions that damage the hippocampus or amygdala disrupt such habits to a lesser degree or not at all. Early studies did not distinguish clearly the different parts of the striatum, but it has been established progressively that the lateral (sensorimotor) striatum is involved in the performance of habits, whereas more medial (associative) parts of the striatum are engaged in cue-based, conditional learning and performance. In the past 5 years, direct attempts have been made to use reward devaluation and related training schedules in experiments on rats to determine whether these anatomically defined striatal regions can also be distinguished in terms of differential involvement in A-O and S-R behaviors. The results suggest that they can be. Lesions of the medial striatum in rats disrupt A-O behavior. Specifically, if an animal has learned a behavior in a reward-based training paradigm and reward is then devalued, the animal's

sensitivity to the reward devaluation is reduced with medial striatal lesions (i.e., the animal repeats the habit). By contrast, lesions of the lateral striatum disrupt S-R behavior: the animal's lack of sensitivity to reward devaluation (defining habit) is decreased by lateral striatal lesions, so that now the animal does not repeat the habit.

These findings have been interpreted as indicating that the striatum, and particularly the lateral striatum, is critical for habit learning and that habits are 'stored' there. But much evidence suggests that this is not so. The basal ganglia are embedded in larger brain circuits that interconnect them with brain stem nuclei and with regions of the cerebral cortex. Even on anatomical grounds, then, it is preferable to think of the functioning of the striatum and related basal ganglia nuclei as parts of these larger circuits. For habit learning and habit memory, evidence from lesion studies supports this broader view. For example, regions of the frontal cortex of the rat modulate these A-O and S-R modes of behavior. If the infralimbic region of the frontal cortex is damaged, rats that otherwise would perform a task in the habit mode (regardless of outcome) now are sensitive to outcome. The relative dependence of behavioral performance on the striatum probably strongly depends on the particular behavior, as well as on specializations of different regions of the striatum for particular sensory, motor, and other behavioral contexts. For example, the ventral striatum is strongly and directly interconnected with regions of the limbic system, olfactory system, and parts of the hippocampus and may serve specialized functions. It is interesting that this region is critical for courtship behavior, a form of instinctual behavior closely related to learned habits and rituals. A part of the dorsolateral striatum is essential for the expression of the highly ritualistic grooming behavior called syntactic grooming. In some songbirds, a specialized part of the basal ganglia forms the core of a circuit necessary for song learning. Altogether these results suggest that the basal ganglia participate in both the learning of habits and the performance of acquired habits, even if the template for the behavior is not stored in a given part of the striatum.

Computational Framework for Studies of Habit Learning

The field of research on reinforcement learning is making a strong contribution to the study of habit learning. The fundamental concept is that an animal (or, more generally, an agent) with a goal to achieve explores the environment in pursuit of the goal receives feedback from the environment as he acts, and decides on which actions to take depending on this feedback. The animal needs to explore, but it also needs to exploit (i.e., select or avoid an action), and the balance between whether to explore or to exploit exists in the context of uncertainty, which exploring can reduce. But precipitous exploitation can be costly. The animal thus works probabilistically, estimating the values of subgoals, forming a policy to follow, and judging the value of particular actions in the context according to that policy. There is considerable interest in determining whether parameters in such reinforcement learning models – for example, sub-goal value, action value, and policy – are represented by activity in the brain, and evidence suggests that such parameters are indeed recognizable. The best examples of this congruence between experiment and theory are in relation to the firing of the dopamine-containing neurons of the midbrain and the striatal neurons that receive these projections. The firing of the dopamine-containing neurons suggests that these neurons correspond to the reward prediction error postulated by some reinforcement learning theories, notably the temporal difference algorithm. These neurons may also code uncertainty and action value. Action value and aspects of reward prediction are, in turn, congruent with the firing patterns of neurons in the striatum. These findings have prompted the view that basal ganglia circuits are active in relation to the exploration–exploitation balance in behavior and to the development of the best policy for obtaining reward. Some authors cast the development of a habit as establishing a fixed value for actions, and they contrast this aspect of basal ganglia (striatal) activity with a more flexible control of behavior in which fixed values are not assigned to behaviors. There are clear ties between these model-based views and the experimentally derived views of A-O and S-R behavior.

Addictions as Habits

When people or experimental animals are exposed repeatedly to addictive drugs, especially by intermittent exposure, many become addicted to the drug or drug class. This addiction is manifest by their seeking to take the drug repeatedly and by their extreme motivation to have and to take the drug. In humans, these behaviors can occur despite conscious awareness that the drug use is 'bad' (antisocial, dangerous, or even life-threatening) in the long or short run. The dopamine-containing neurons of the midbrain and their targets in the striatum and cortex are central to the acquisition and expression of such drug-induced habits. The induction of an addictive behavior requires the midbrain dopamine-containing neurons in the

ventral tegmental area, and the expression of the addictive behavior requires the ventral striatum (and allied circuits), principally including the nucleus accumbens. Studies have probed the relationship between these behaviors and the behaviors of nonaddicted animals performing habits. There are many similarities. A key aspect of the study of addiction is that the molecular and cellular bases of these behaviors can be studied in experiments by administering these drugs to animals or by having the animals self-administer the drugs. This work has great potential to uncover mechanistic bases of such extreme habits and perhaps of habit circuit functioning in general.

Chunking of Action Repertoires

These various views of the process of acquiring a habit – or a learned procedure or sequential behavior – have deep commonalities. One indication of this comes from studies in which chronic recordings have been made in the striatum as animals go through the process of acquiring a new behavior. This work suggests that there is a form of 'neural exploration' early in training, when the population of neurons fire in relation to virtually all aspects of the procedure, followed by a phase of 'neural exploitation,' in which the firing patterns of the neurons become tuned to selective aspects of the procedure. In the sensorimotor part of the striatum, the population of neurons eventually tend to fire the most at the beginning and end of the entire procedure, suggesting that they may help bracket or chunk the behavior after exploitation. Insofar as this part of the striatum helps to release or to inhibit actions, this chunking function could help to encode behaviors that are accepted as part of behavioral policy through a process of reinforcement-based learning. The explore–exploit framework has the added appeal that it is applicable also to the learning of behaviors such as singing in birds, in which the behavior goes through a prolonged period of babbling (exploration) before exploitation (the crystallized adult song).

See also: Addiction: Neurobiological Mechanism; Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia: Motor Functions; Basal Ganglia: Acetylcholine Interactions and Behavior; Basal Ganglia: Internal Organization; Drug Addiction: Behavioral Neurophysiology; Learning, Action, Inference and Neuromodulation; Motor Skill Learning; Procedural Learning: Striatum; Procedural Learning in Humans.

Further Reading

Brainard MS and Doupe AJ (2002) What songbirds teach us about learning. *Nature* 417: 351–358.

Daw ND and Doya K (2006) The computational neurobiology of learning and reward. *Current Opinion in Neurobiology* 16: 199–204.

Daw ND, Niv Y, and Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8: 1704–1711.

Dickinson A and Balleine BW (1994) Motivational control of goal-directed action. *Animal Learning & Behavior* 22: 1–18.

Doyon J and Benali H (2005) Reorganization and plasticity in the adult brain during learning of motor skills. *Current Opinion in Neurobiology* 15: 161–167.

Graybiel AM (1998) The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory* 70: 119–136.

Graybiel AM (2004) The basal ganglia and cognitive pattern generators. *Schizophrenia Bulletin* 23: 459–469.

Graybiel AM (2005) The basal ganglia: Learning new tricks and loving it. *Current Opinion in Neurobiology* 15: 638–644.

Graybiel AM and Saka E (2004) The basal ganglia and the control of action. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences,* 3rd edn., pp. 495–510. Cambridge, MA: MIT Press.

Hyman SE, Malenka RC, and Nestler EJ (2006) Neural mechanisms of addiction: The role of reward-related learning and memory. *Annual Review of Neuroscience* 29: 565–598.

James W (1890) Habit. In: *The Principles of Psychology*, pp. 104–127. New York: Dover.

Kalivas PW and Volkow ND (2005) The neural basis of addiction: A pathology of motivation and choice. *American Journal of Psychiatry* 162: 1403–1413.

Packard MG and Knowlton BJ (2002) Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience* 25: 563–593.

Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36: 241–263.

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Volkow ND and Li TK (2004) Drug addiction: The neurobiology of behaviour gone awry. *Nature Reviews. Neuroscience* 5: 963–970.

Yin HH and Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* 7: 464–476.

Basal Ganglia: Internal Organization

J P Bolam, M T C Brown, J Moss, and P J Magill, MRC Anatomical Neuropharmacology Unit, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The basal ganglia are a group of subcortical nuclei involved in a variety of processes, including motor, associative, cognitive, and mnemonic functions. The dorsal division of the basal ganglia consists of the striatum (divided into the caudate nucleus and putamen by the internal capsule in primates and other species), the external segment of the globus pallidus (GPe; simply globus pallidus in rodents), the internal segment of the globus pallidus (GPi; entopeduncular nucleus (EP) in several species, including rodents and cats), the subthalamic nucleus (STN), and the substantia nigra (SN). The latter structure is divided into two main parts, the dorsal pars compacta (SNc), in which the dopaminergic nigrostriatal neurons are located, and the more ventral pars reticulata (SNr). In addition to these structures that are associated with motor and associative functions of the basal ganglia, there is a ventral division of the basal ganglia (ventral striatum or nucleus accumbens; ventral pallidum, which is equivalent to GPe and GPi; and the ventral tegmental area, which is a medial continuation of the SNc). The ventral division of the basal ganglia is primarily associated with limbic functions. The nomenclature of divisions of the basal ganglia can be particularly confusing to a reader new to the field. For this reason a table (Table 1) of commonly used terms and their synonyms or homologous structures is included.

The main transmitter used by neurons of the basal ganglia is γ-aminobutyric acid (GABA); as many as 99% of all basal ganglia neurons are GABAergic. The only exceptions are glutamatergic neurons of the STN, dopamine neurons of the SNc, and one population of interneurons in the striatum that utilizes acetylcholine. The major inputs to the basal ganglia are glutamatergic and are derived from the cerebral cortex and the thalamus. The main point of entry of the cortical and thalamic information to the basal ganglia is the striatum, although there are also significant projections to the STN (Figure 1). Virtually the entire cortical mantle projects onto the striatum. The corticostriatal projection is derived from several classes of cortical pyramidal neurons that predominantly innervate the ipsilateral striatum but also the contralateral striatum. The corticosubthalamic projection is derived from more restricted regions of the cortex (see below). The corticostriatal and thalamostriatal projections are highly topographically organized and impart functionality onto the striatum and, consequently, other divisions of the basal ganglia. The main synaptic targets of the cortical and thalamic inputs to the basal ganglia are the medium-sized densely spiny projection neurons of the striatum (Figures 2 and 3).

In what is now considered the classical view of basal ganglia circuitry, the functional organization is such that cortical and thalamic information is processed within the striatum and integrated with the many other inputs that reach the basal ganglia (e.g., from amygdala, hippocampus, and dorsal raphé), which primarily innervate the striatum, and then the 'processed information' is transmitted to the output nuclei of the basal ganglia, that is, the GPi (or EP) and the SNr, by two routes, the so-called direct and indirect pathways. In the direct pathway, cortical and thalamic information is transmitted directly from the striatum to the output nuclei. In the indirect pathway, cortical and thalamic information is transmitted indirectly to the output nuclei via the complex network interconnecting the GPe and STN. The basal ganglia then influence behavior by projecting, via the output nuclei, to the thalamus (mainly the ventral anterior and ventral lateral nuclear complex, mediodorsal nucleus and the intralaminar nuclei) that in turn projects back to the cortex or by projecting to subcortical regions that are involved in movement, including the superior colliculus, the pedunculopontine nucleus (PPN), lateral habenula, and reticular formation (Figure 1).

Overlying this feed-forward organization of the basal ganglia are many feedback pathways. The major one of these is the dopaminergic projection from the SNc that predominantly innervates the striatum but also innervates the STN and GPe. This projection modulates the flow of cortical and thalamic information through the basal ganglia. Loss of these dopamine neurons in Parkinson's disease leads to an imbalance of the flow of cortical information through the basal ganglia in favor of the indirect pathway and the motor symptoms associated with this disorder. Other feedback systems include the serotonergic projection from the dorsal raphé and the projection from the GPe to striatum (see later).

Table 1 Nomenclature

Commonly used name for a particular structure	*Homologous structure and/or synonyms*
Striatum	Neostriatum, caudate-putamen
Globus pallidus (rat)	External segment of globus pallidus (GPe; primate)
	Lateral globus pallidus (GPl; primate)
Entopeduncular nucleus (rat)	Internal segment of globus pallidus (GPi; primate)
	Medial globus pallidus (GPm; primate)
Ventral striatum	Nucleus accumbens

Striatum

The principal neurons of the striatum, the medium-size densely spiny neurons (MSN; Figure 3), give rise to the direct and indirect pathways. The MSNs have a perikaryon about 15 μm in diameter and give rise to several dendrites that are initially spine free but then become densely laden with spines after the first bifurcation. They account for up to 97% of striatal neurons in rodents but a smaller proportion in primates. They utilize GABA as their major neurotransmitter and are subdivided into two major subpopulations on the basis of their projection region, pattern of axonal collateralization, and neurochemical content. One subpopulation gives rise to the direct pathway, they preferentially project to the output nuclei of the basal ganglia (but also send a collateral to the GPe), and selectively express the neuropeptides, substance P and dynorphin, and D1 dopamine receptors. The other subpopulation gives rise to the indirect pathway, they exclusively project to the GPe and express enkephalin and D2 dopamine receptors. Both populations also give rise to local axon collaterals, the principal target of which are other MSNs. Axon terminals derived from MSNs form symmetric synapses (Gray's type 2) and their morphological features are similar in all their target regions (Figure 4(a)). Their spines are the main target of corticostriatal afferents and at least part of the thalamostriatal projection; both projections form asymmetric synapses (Gray's type 1; Figure 2). An individual cortical axon gives rise to only very few synapses with an individual MSN; since MSNs possess in the region of 10 000–15 000 spines, the implication is that there is a massive convergence of cortical axons (and possibly thalamic axons) on an individual MSN. The MSNs are also the principal target of the dopaminergic feedback from the SNc (spine necks, dendrites, and perikarya; Figure 2).

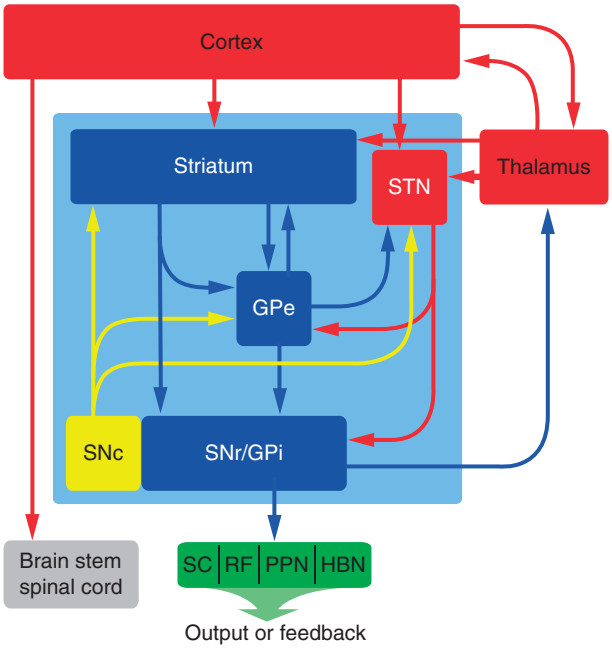

Figure 1 Simplified block diagram of the basal ganglia and their principal connections. The nuclei of the basal ganglia are included in the light blue box and consist of the striatum, the external segment of the globus pallidus (GPe), the subthalamic nucleus (STN), the substantia nigra pars reticulata and the internal segment of the globus pallidus (SNr/GPi), and the substantia nigra pars compacta (SNc). The two major inputs to the basal ganglia are from the cortex and the thalamus (mainly the intralaminar nuclei). The SNr and GPi constitute the output nuclei of the basal ganglia. The basal ganglia influence behavior by the output nuclei (SNr/GPi) projecting to the thalamus and thence back to the cortex, and projections to the superior colliculus (SC), the reticular formation (RF), the pedunculopontine nucleus (PPN), and the lateral habenula (HBN). Dopamine neurons of the SNc provide a massive feedback projection to the striatum and to the GPe and STN, which modulates the flow of cortical and thalamic information through the basal ganglia. In the now classical description of the organization of the basal ganglia, the output from the striatum to the output nuclei is referred to as the 'direct pathway' of information flow through the basal ganglia. This pathway leads to a disinhibition of basal ganglia targets and is considered to be associated with basal ganglia behavior. The 'indirect pathway' consists of the projection from the striatum to the GPe and the projection of the GPe to the STN and thence to the output nuclei. However, recent evidence suggests that the indirect pathway is much more complicated, involving a complex network interconnecting the GPe and the STN and the output nuclei. Furthermore, the STN receives input from both the thalamus and cortex. Activity in the indirect pathway is believed to underlie, at least in part, the resting inhibitory output from the GPi/SNr, and activation of the indirect pathway leads to a greater inhibition of basal ganglia targets and is considered to be associated with the attenuation of basal ganglia behavior. Dark blue indicates structures that are principally GABAergic, red indicates structures that are principally glutamatergic, yellow indicates structures that are dopaminergic, and green indicates structures with variable neurochemistry.

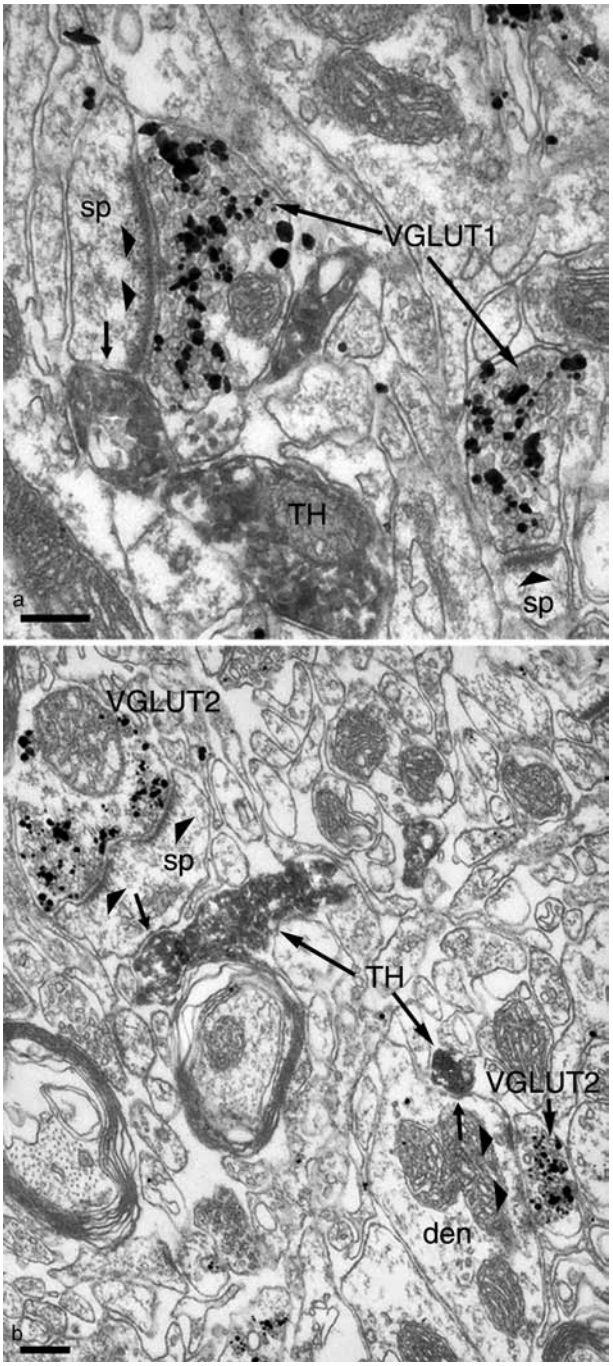

Figure 2 Synaptic connections of cortical, thalamic, and dopamine terminals in the striatum. (a) Electron micrograph of a section immunolabeled to reveal the vesicular glutamate transporter 1 (VGLUT1) as a marker of glutamatergic corticostriatal terminals and tyrosine hydroxylase (the rate-limiting enzyme in the synthesis of catecholamines; TH) as a marker of dopaminergic nigrostriatal axons and terminals. The VGLUT1 immunoreactivity was revealed by means of an immunogold method (electron-dense particles, which appear as dense black blobs), and the TH was revealed by an immunoperoxidase method (amorphous, electron-dense reaction product giving the labeled structure an overall dark appearance). The majority of cortical terminals typically form asymmetric synapses (Gray's type 1; arrowheads) with dendritic spines (sp) of the principal neuron of the striatum, the medium-size densely spiny neuron (MSN; see **Figure 3**). Note that cortical terminals make synaptic contact with the dendritic shafts of MSNs to a much lesser extent and also make synaptic contact with the dendritic shafts of interneurons. The TH-immunoreactive axon makes symmetric synaptic contact (Gray's type 2; small arrows) with the neck of the spine that receives the input from the cortical terminal. This spatial relationship between cortical terminals and nigrostriatal terminals has been considered the principal site of interaction between dopamine and glutamate in the striatum and, indeed, the basal ganglia as a whole. Note that dopaminergic terminals also contact the dendritic shafts and perikarya of MSNs and of interneurons. (b) Electron micrograph of a section of striatum immunolabeled in a manner similar to (a) except that the vesicular glutamate transporter 2 (VGLUT2) was used as a marker of thalamostriatal terminals. Axon terminals derived from the thalamus make synaptic contact with a higher proportion of dendritic shafts than do corticostriatal terminals, although the relative proportions of spines and shafts varies with the thalamic nucleus of origin. In this micrograph, one VGLUT2-positive terminal forms an asymmetric synapse (arrowheads) with a spine (sp; top left). A second VGLUT2-positive terminal forms asymmetric synaptic contact (arrowheads) with a dendritic shaft (den; lower right). Both the postsynaptic spine and the postsynaptic dendritic shaft receive synaptic input (small arrows) from TH-positive terminals, indicating that, at least in qualitative terms, the spatial relationship between glutamatergic thalamostriatal terminals and dopaminergic terminals is similar to the relationship between corticostriatal and dopaminergic terminals. Scale bar = 200 nm (a, b). From Moss J and Bolam JP (unpublished).

The striatum also contains at least three populations of GABA interneurons: (1) fast-spiking, parvalbumin (PV)-positive interneurons; (2) those that express nitric oxide synthase (NOS); and (3) those that express calretinin. The PV-positive neurons receive major inputs from the cortex, thalamus, and neurons of the GPe (see the next section) and provide an inhibitory input to the MSNs. The striatum also contains a population of large cholinergic interneurons that receive input from the cortex, thalamus and MSNs, and in turn innervate MSNs. The interneurons are also targets of the dopaminergic input from the SNc. The interneurons account for only a small proportion of striatal neurons (~3–10%, depending on species). In addition to its functional division into the direct and indirect pathways, the striatum possesses subcompartments, the striosomes (or patches), matrix, and matrisomes, which possess different neurochemical markers and have different input/output characteristics.

GPe

The main extrinsic synaptic target of MSNs that give rise to the indirect pathway are the neurons of the GPe. All GPe neurons are GABAergic and differentially express the calcium-binding proteins PV or calbindin. They are arranged in such a way as to expose

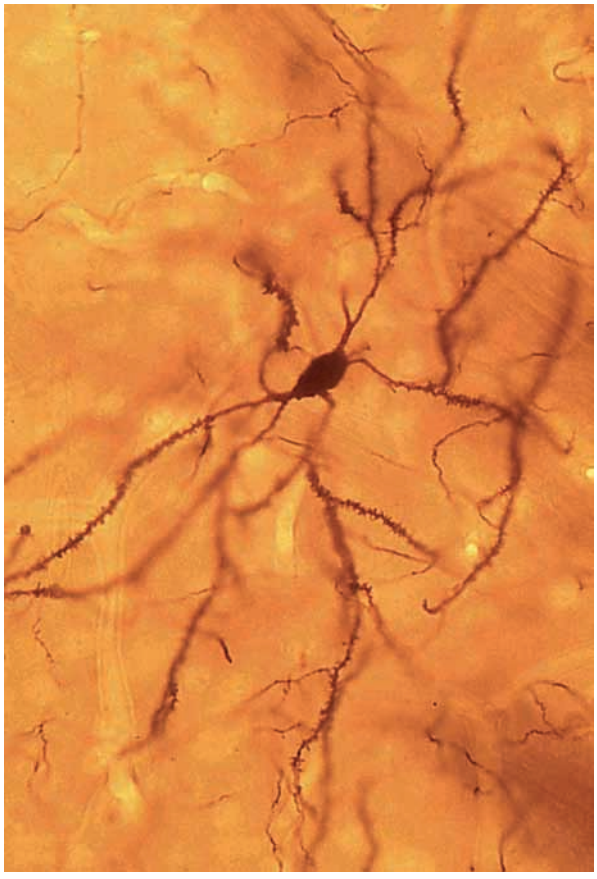

Figure 3 Golgi-impregnated medium-size densely spiny neuron (MSN) in the rat striatum. This is the principal neuron of the striatum, accounting for up to 97% of all striatal neurons. They are GABAergic, give rise to the output from the striatum, and give rise to local axon collaterals. The cell body is approximately 15 μm and in this focal plane gives rise to four primary dendrites. The dendrites are initially spine free and then become densely laden with spines, usually after the first bifurcation. An individual MSN possesses 10 000–15 000 spines, each of which receives a glutamatergic input at its head (see **Figure 2**). The axon is not visible in this micrograph, but segments of Golgi-impregnated axons (much finer structures), presumably derived from other neurons, are present in the field. At this anatomical level, the MSNs giving rise to the direct and indirect pathways are indistinguishable, as are MSNs from different species.

the maximum extent of their dendritic arbor to the incoming striatal afferents. Thus their principal input is from the striatum; pallidal neurons are ensheathed by striatal terminals in a manner similar to neurons of the GPi (**Figure 4(a)** and see below). However, GPe neurons also receive prominent inputs from the STN (**Figures 4(e) and 4(f)**) and from the local axon collaterals of other pallidal neurons. Individual GPe neurons in turn give rise to projections to functionally related territories of the STN, the output nuclei of the basal ganglia, and the SNc (**Figure 5**), where they form symmetric synapses predominantly with the cell bodies and proximal dendrites of their target neurons (**Figures 4(b), 4(c), and 6**). In addition, all GPe neurons give rise to local axon collaterals that underlie a complex and structured microcircuitry within the GPe (**Figure 5**). About a quarter of GPe neurons give rise to collaterals that innervate the striatum in addition to the projections described above (**Figure 5**). Their principal synaptic targets within the striatum are the proximal regions of the PV-positive and NOS-positive GABAergic interneurons. Because the main synaptic targets of the PV-positive GABAergic interneurons are the MSNs, and an individual interneuron can innervate several hundred MSNs, GPe neurons can theoretically control the activity of every neuron in the striatum. Thus, by virtue of their extensive axon collaterals, GPe neurons are in a position to influence the activity of neurons at every level of the basal ganglia. The strategically important placement of their synapses on the proximal regions of their target neurons (**Figure 6**), their GABAergic nature, and their high resting discharge imply a critical role in the modulation of the flow of information through the basal ganglia.

STN

The STN is a small nucleus located below the thalamus; it consists of a relatively homogeneous population of neurons that extend their dendrites in the principal plane of the nucleus. The most distinguishing characteristic of this nucleus is that, unlike the majority of neurons in the basal ganglia (99%), neurons of the STN use glutamate as a neurotransmitter and thus exert excitatory influences on their targets. Traditionally, the principal inputs of STN neurons have been considered to be derived from the GPe, and indeed these neurons provide the major inhibitory input to the STN (**Figure 4(c)**). In fact, the STN is reciprocally connected to functionally related regions of the GPe. Two other major inputs to the STN are derived from the cortex and intralaminar thalamic nuclei. The glutamatergic corticosubthalamic input is mainly derived from the motor, premotor, and prefrontal cortices and is exclusively ipsilateral and, at least in part, is derived from collaterals of cortical neurons innervating the striatum. The corticosubthalamic axon terminals form asymmetric synapses principally with small-diameter dendrites, and this pathway is the fastest route by which the cortex can influence the STN and, through the projections of STN neurons, the output neurons of the basal ganglia. The thalamic input to the STN is also glutamatergic, gives rise to asymmetric synapses with small-diameter dendrites, and is derived from those thalamic nuclei that also innervate the striatum. The STN also receives input from

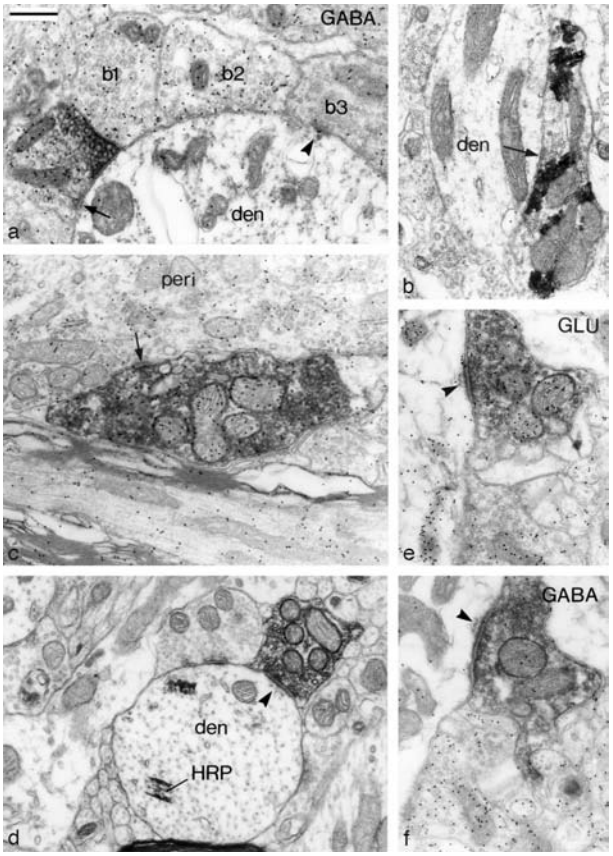

Figure 4 Characteristic features of synaptic terminals of neurons of the basal ganglia. A series of electron micrographs illustrating typical ultrastructural features, synaptic specializations, and neurochemistry of synaptic terminals derived from the striatum (a), the external segment of the globus pallidus (GPe) (b,c), and the subthalamic nucleus (STN; (d–f)). The sections illustrated in (a), (c), and (f) were labeled by the postembedding immunogold method to reveal γ-aminobutyric acid (GABA) immunoreactivity, and the section illustrated in (e) was processed to reveal glutamate (GLU) immunoreactivity. (a) A striatal terminal (farthest to the left) in the internal segment of the globus pallidus (GPi) labeled following an injection of the anterograde tracer biotinylated dextran amine (BDA) in the putamen of a squirrel monkey. The terminal is identified by the electron-dense reaction product formed by the histochemical reaction to reveal the BDA. It is in symmetric synaptic contact (arrow) with a dendrite (den) and is immunoreactive for GABA (as indicated by the high density of immunogold particles overlying it). Two neighboring boutons (b1 and b2), which also possess the morphological features of striatal terminals, are also immunoreactive for GABA. These terminals are derived from the striatal medium-size densely spiny neurons. Contrast these with bouton b3, which forms an asymmetric synapse (arrowhead) and does not display GABA immunoreactivity. This bouton possesses morphological features of a terminal derived from the STN. (b, c) Terminals derived from the GPe forming symmetric synaptic contacts (arrows) with a dendritic shaft (den) in (b) and a perikaryon (peri) in (c). The terminal in (b) is in the rat entopeduncular nucleus (equivalent to GPi) and was labeled after an injection of the anterograde tracer *Phaseolus vulgaris* leucoagglutinin in the GP of the rat. The bouton in (c) is in the STN and was labeled after an injection of BDA in the GPe of the squirrel monkey. This section was incubated to reveal GABA immunoreactivity, and the labeled bouton displays a high concentration of immunogold particles overlying it, indicating that it is enriched in GABA. Note the similarity in the morphological characteristics of the pallidal boutons despite the fact that they are from different species and in different nuclei. (d) A terminal derived from the STN. This terminal is in the rat substantia nigra pars reticulata and was anterogradely labeled following the injection of the anterograde tracer biocytin in the STN. It forms an asymmetric synapse (arrowhead) with a dendrite (den) that contains retrogradely transported horseradish peroxidase (HRP) from the ventromedial thalamic nucleus, thus identifying it as a dendrite of a basal ganglia output neuron. (e,f) Adjacent sections of the same anterogradely labeled STN bouton in the GPe of squirrel monkey that forms an asymmetric synapse (arrowheads). The bouton is enriched in glutamate immunoreactivity (in (e)) but is not immunoreactive for GABA (in (f)). Note the similarity in the morphological features of the subthalamic terminals in different species and in different nuclei. Scale bar = 0.5 μm (a–f). From Smith Y, Bevan MD, Shink E, and Bolam JP (1998) Commentary: Microcircuitry of the direct and indirect pathways of the basal ganglia. *Neuroscience* 86: 353–387.

dopamine neurons of the SNc, serotonergic neurons from the dorsal raphé, and cholinergic inputs of the PPN.

The main outputs of the STN are to functionally related regions of the output nuclei of the basal ganglia, that is, the SNr (Figure 4(d)) and GPi (or EP). Individual STN neurons also innervate the GPe (Figures 4(e) and 4(f)), the PPN, and the dopamine neurons of the SNc. In each target, the terminals of STN neurons form asymmetric synapses, and at least some have been shown to be enriched in glutamate (Figure 4(e)). The placement of STN terminals is mainly on the proximal regions of their targets. The glutamatergic nature of STN neurons, their relatively high resting discharge rate, and their responsiveness to excitatory inputs from the cortex and thalamus underlie a critical role in setting the level of activity in basal ganglia output neurons.

SNr and GPi

The SNr and the GPi (or EP in rodents) represent the major output nuclei of the basal ganglia, sending their axons to the thalamus and thence back to the cortex or to subcortical structures involved in the control of behavior, including the superior colliculus, parvicellular reticular formation, lateral habenula, and the PPN (Figure 1). The relative importance of the two output nuclei in the control of behavior varies among species of different orders. Thus, output to thalamus from the GPi in primates is probably more important than the output from the SNr whereas in rodents, the output from the SNr is probably more important than that from the EP.

Basal ganglia output neurons are large GABAergic neurons with long, infrequently branching, aspiny

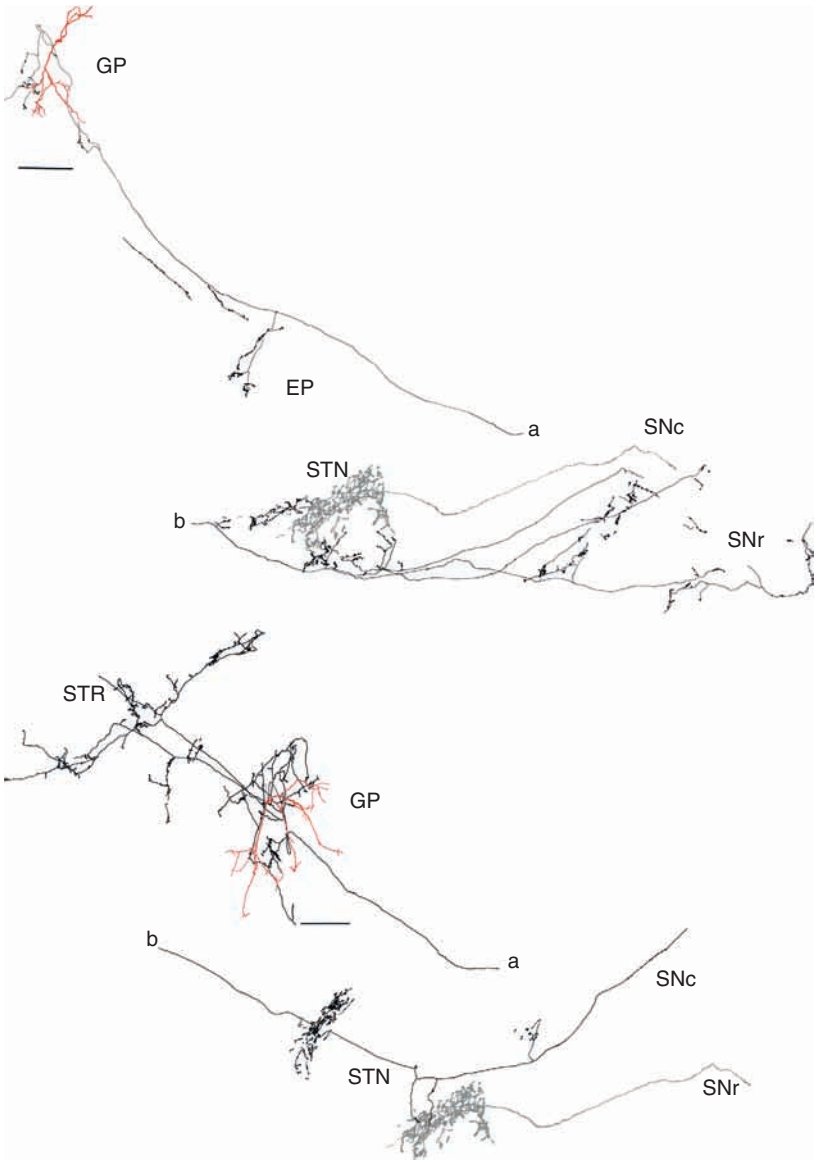

Figure 5 Individual neurons of the rat globus pallidus (GP) innervate multiple regions of the basal ganglia. Partial reconstructions of individual neurons from the rat GP that were labeled with neurobiotin *in vivo* by the juxtacellular method following electrophysiological characterization. These neurons were reconstructed from serial sections using a camera lucida drawing tube. The cell bodies and dendrites are shown in red, and the axon in black. In both cases the drawing of the axon has been cut in two; the ends of the axons marked 'a' connect with the ends marked 'b'. The upper neuron gave rise to local axon collaterals within the GP, and the main axon traveled through the internal capsule to the entopeduncular nucleus (EP), subthalamic nucleus (STN), and substantia nigra pars compacta (SNc) and pars reticulata (SNr). In addition to local axon collaterals and clear projections to the STN, SNc, and SNr, the lower neuron gives rise to two collaterals that innervate the striatum (STR). About a quarter of labeled GP neurons gave rise to collaterals that innervated the STR. Observations similar to these have been made in the external segment of the primate GP. The neuron of origin of those structures shown in gray could not be distinguished beween the two neurons. Scale bar = 300 μm. From Bevan M, Booth P, and Bolam JP (unpublished neurons).

dendrites. In the GPi at least, they are similar to GPe neurons in that their dendrites are arranged in such a way as to expose their maximum surface to the incoming striatal afferents. Their major input is from the MSNs of the striatum that give rise to the direct pathway; thus dendrites and perikarya are ensheathed in afferent terminals, most of which are derived from the striatum (Figure 4(a)). They also receive a prominent afferent input on their perikarya and proximal dendrites from the GABAergic neurons

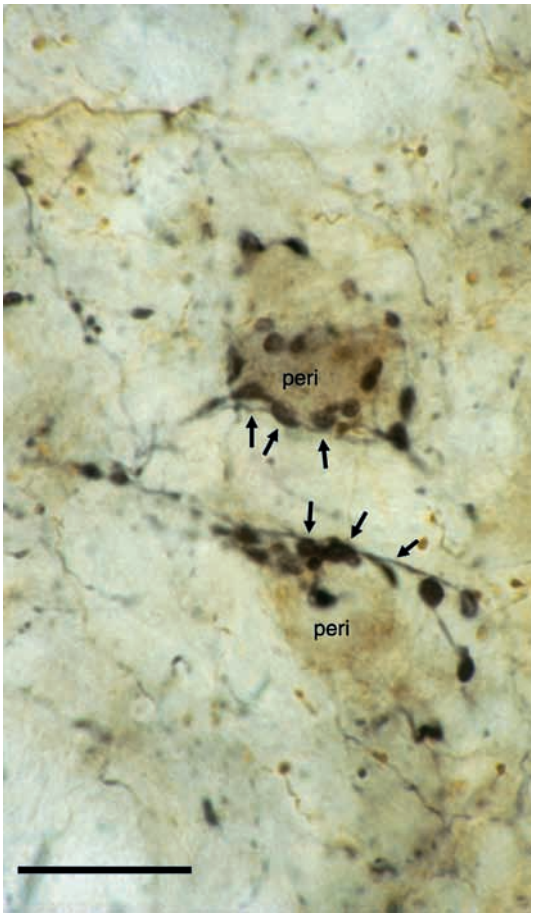

Figure 6 Globus pallidus (GP) neurons form multiple contacts with the proximal regions of their target neurons. Light micrograph of the substantia nigra pars reticulata (SNr) of a rat, in which anterograde tracer was deposited in the GP and revealed by a peroxidase method. The perikarya of two SNr neurons are present in the center of the field (peri) and are apposed by many anterogradely labeled large axonal boutons (some indicated by small arrows) derived from the GP. This 'basket-like' innervation is typical of the manner in which GP neurons innervate their target neurons. Note that this section was taken from an animal that also received an injection of anterograde tracer in the striatum; the brown-labeled structures are thus of striatal origin. Scale bar = ~20 μm. From Smith Y and Bolam JP (unpublished).

of the GPe (Figure 4(b)) and glutamatergic neurons of the STN (b3 in Figures 4(a) and the labeled boutons in Figures 4(d)–(4f)). Individual basal ganglia output neurons thus receive synaptic input from both the direct pathway and the indirect pathway. The high resting discharge rate of neurons in the GPi and SNr underlies a tonic inhibition of neurons in the target regions of the basal ganglia under resting conditions. Increased activity of striatal afferents to these neurons leads to a reduction in their firing rate and hence reduced inhibition (i.e., disinhibition) of the targets of the basal ganglia. This disinhibition is considered to be a key factor in the way the basal ganglia influence behavior.

SNc

The SNc consists principally of dopamine neurons although some evidence suggests it also has a population, albeit small, of GABAegic neurons. The dopamine neurons lie in a densely packed band with the majority of dendrites following the plane of the compacta (Figure 7). Some dendrites, particularly of those neurons located in the ventral part of the compacta, also descend into the SNr (Figure 7). Dopamine neurons receive a dense innervation from GABAergic terminals derived from the striatum and the GPe and also receive glutamatergic input from the STN, cholinergic terminals from the PPN, and serotonergic terminals from the dorsal raphé. The axon emerges from the cell body or primary dendrite (and, rarely, a secondary dendrite). Dopamine neurons do not give rise to local axon collaterals within the SN, although it is evident that the dendrites can release dopamine and thus influence neighboring neurons. The main target of the axon is the striatum, but the GPe and the STN also receive a significant dopaminergic input via collaterals of the main axon. Individual dopamine neurons give rise to a remarkable number of synapses. Estimates vary between about 250 000 and 330 000 synapses at the level of the striatum, plus any additional synapses arising from collaterals in the GP and STN. The axon terminals of dopamine neurons give rise to symmetric synapses. The major targets in the striatum are the spines, dendritic shafts, and perikarya of MSNs (Figure 2). Synaptic contact on spines generally occurs at the neck, and the spine invariably receives a glutamatergic terminal derived from the cortex or thalamus, forming an asymmetric synapse (Figure 2). This spatial relationship underlies the modulatory role of dopamine on the excitatory influence from cortex and thalamus.

Concluding Remarks

For the sake of clarity, this description of the internal organization of the basal ganglia has been simplified, omitting many of the less well understood (but probably still important) synaptic connections. Furthermore, the description relies heavily on data derived from the dorsal division of the basal ganglia. Nevertheless, the basic principles of the connections

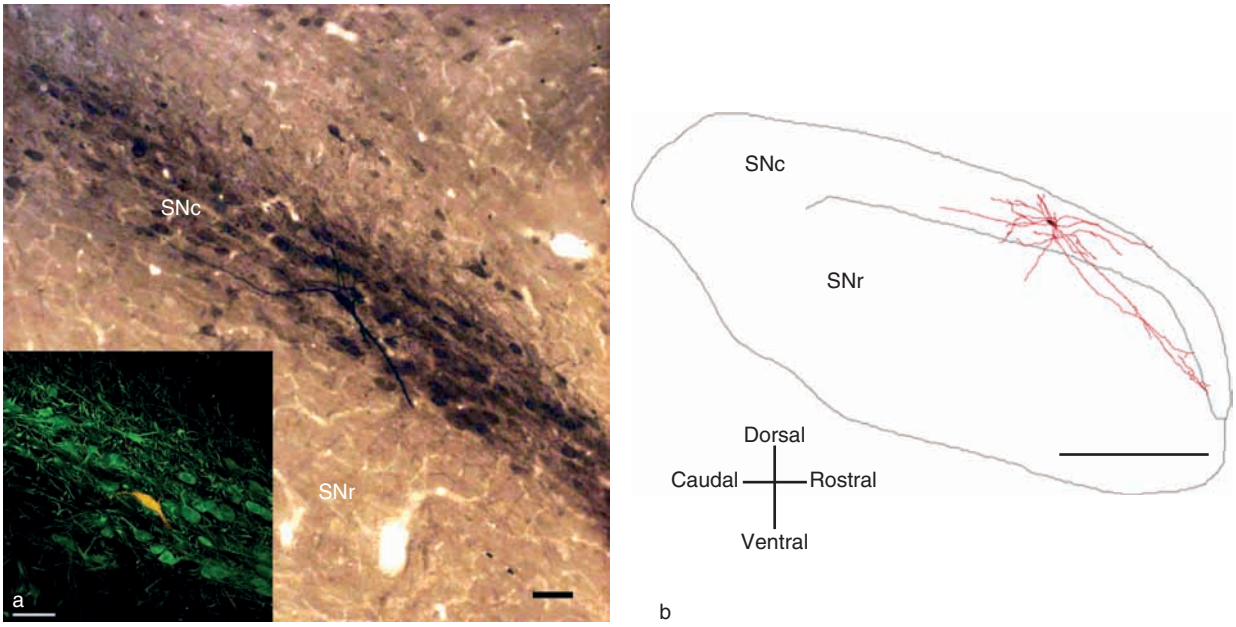

Figure 7 Micrographs and a digital reconstruction of an individual dopamine neuron in the rat substantia nigra pars compacta (SNc), recorded *in vivo* and then labeled with neurobiotin by the juxtacellular method. The neurobiotin, together with immunoreactivity for tyrosine hydroxylase (TH), were revealed by fluorescence methods ((a), inset). In this section the TH immunoreactive elements were revealed by a green fluorescent marker and the double labeling for TH and the neurobiotin by the yellow fluorescence. The labeled neuron was thus confirmed to be dopaminergic by the presence of immunoreactivity for TH. After neurochemical characterization, the neurobiotin and immunoreactivity for TH were localized by peroxidase reaction products, and the main panel in (a) shows the labeled neuron in the dense layer of TH-positive neurons in the SNc. The digital reconstruction in (b) shows the typical arrangement of the majority of dendrites of dopamine neurons that follow the plane of the SNc, with some dendrites descending into the SN pars reticulata (SNr). The axon of this neuron was followed to the globus pallidus, where is gave rise to a cluster of boutons, and then to the striatum. Scale bar = 25 μm; (a, inset) 500 μm (b). From Brown MTC, Bolam JP, and Magill PJ (unpublished).

described here also apply to the ventral division of the basal ganglia. One can thus conclude from the study of the functional organization of the basal ganglia that the complex functions of the basal ganglia are underpinned by an equally complex cellular and synaptic architecture.

See also: Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia: Motor Functions; Basal Ganglia: Acetylcholine Interactions and Behavior; Basal Ganglia: Physiological Circuits; MPTP Parkinsonism Model; Procedural Learning: Striatum; Striatum: Internal Physiology.

Further Reading

Alexander GE and Crutcher ME (1990) Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in Neurosciences* 13: 266–271.

Bevan MD, Booth PAC, Eaton SA, and Bolam JP (1998) Selective innervation of neostriatal interneurons by a sub-class of neuron in the globus pallidus of the rat. *Journal of Neuroscience* 18: 9438–9452.

Bolam JP, Bergman H, Graybiel A, et al. (2006) Molecules, microcircuits and motivated behaviour: Microcircuits in the striatum. In: Grillner S and Graybiel A (eds.) *Microcircuits: The Interface between Neurons and Global Brain Function*, pp. 165–190. Cambridge, MA: MIT Press.

Bolam JP, Booth PAC, Hanley JJ, and Bevan MD (2000) Synaptic organisation of the basal ganglia. *Journal of Anatomy* 196: 527–542.

DeLong MR (1990) Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences* 13: 281–285.

Gerfen CR and Wilson CJ (1996) The basal ganglia. In: Björklund A, Hökfelt T, and Swanson L (eds.) *Handbook of Chemical Neuroanatomy,* vol. 12, Part III, pp. 369–466. Amsterdam: Elsevier.

Graybiel AM (2005) The basal ganglia: Learning new tricks and loving it. *Current Opinion in Neurobiology* 15: 638–644.

Lacey CJ, Bolam JP, and Magill PJ (2007) Novel and distinct operational principles of intralaminar thalamic neurons and their striatal projections. *Journal of Neuroscience* 27: 4374–4384.

Sadek AR, Magill PJ, and Bolam JP (2007) A single-cell analysis of intrinsic connectivity in the rat globus pallidus. *Journal of Neuroscience* 27: 6352–6362.

Smith Y, Bevan MD, Shink E, and Bolam JP (1998) Commentary: Microcircuitry of the direct and indirect pathways of the basal ganglia. *Neuroscience* 86: 353–387.

Tepper JM, Abercrombie ED, and Bolam JP (eds.) (2007) *Special Issue: GABA and the Basal Ganglia: From Molecules to Systems. Progress in Brain Research* 160.

Tepper JM and Bolam JP (2004) Functional diversity and specificity of neostriatal interneurons. *Current Opinion in Neurobiology* 14: 685–692.

Basal Ganglia: Motor Functions

M E Anderson, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Although James Parkinson described the motor signs of the disorder that bears his name in 1817, the association between lesions of the basal ganglia and movement disorders was not really established until the clinico-pathological reports of Wilson and the Vogts in the early part of the twentieth century. Identification of this relationship led to a focus on the role of the basal ganglia in movement. In spite of recent evidence that the basal ganglia are involved in cognitive as well as motor function, abnormalities of movement are still the most visible and easily identified signs of basal ganglia dysfunction.

Anatomical Circuits: Skeletomotor and Oculomotor Systems

The basic connectivity from the cerebral cortex through the basal ganglia is shown in Figure 1. This type of circuit is repeated for many cortical areas, including the several motor areas of the frontal and parietal lobes. Cortical inputs to the basal ganglia are primarily to the striatum (both the putamen and the caudate nuclueus) and to the subthalamic nucleus. Output from the basal ganglia is entirely inhibitory, with γ-aminobutyric acid (GABA) as the neurotransmitter, and it originates from neurons in the internal segment of the globus pallidus (GPi) or the reticular portion of the substantia nigra (SNr).

The inhibitory output of GPi is largely destined for the ventrolateral and intralaminar nuclei of the thalamus and for the pedunculopontine region of the midbrain (PPN). That from SNr projects primarily to more medial and rostral areas of the ventral thalamus and caudally to the superior colliculus, as well as to the PPN. Anatomical tracing studies using viruses that are injected into various areas of cerebral cortex and then transported retrogradely and transsynaptically to the GPi or SNr showed that the basal ganglia origins of input to different cortical areas via the thalamus are segregated; that is, they are parts of parallel circuits. Furthermore, these parallel circuits form closed loops, with the basal ganglia output returning to an area of the cortex from which the cortical input is derived, via the striatum. For example, striatal neurons in the loop that connects the basal ganglia with the supplementary motor area also receive input from the supplementary motor cortex. In addition to these closed loops, a circuit may also be influenced by input from other cortical sources. Some of these are direct cortico-cortical connections (e.g., the primary motor cortex receives input from the supplementary motor area), and some are via the thalamus (e.g., from areas of thalamus that are connected to the prefrontal cortex). This allows the areas involved in the execution of motor output to be influenced by areas whose signals relate more to working memory, sequential order, or motor planning, in addition to the terminal kinematic or kinetic variables.

Some of the inhibitory output from the SNr is directed to the superior colliculus, which largely influences oculomotor and neck muscles and, hence, the eye and head position (see Figure 1). Another target of nigral inhibitory output is the pedunculopontine region of the brain stem, which especially affects posture and locomotion.

Additional elements in each circuit through the basal ganglia include the subthalamic nucleus, which provides glutamatergic excitatory synaptic input to both the pallidal segments and to the SNr (Figure 1), and the nigrostriatal dopaminergic system, which provides a modulatory action at several points in the basal ganglia, but most obviously at the putamen and caudate nucleus.

Thus, the loops through the basal ganglia that are relevant in the production of controlled movement may include those that originate from each of the areas of the frontal lobe, with individual areas contributing short-term memory, motor planning, sequencing of movement, or the speed and amount of force generation or displacement. The final consequence of signal processing in any of the loops through the basal ganglia is a change in inhibitory output for that circuit from GPi or SNr.

The Basal Ganglia Act by Varying Tonic Inhibition

Activity of Output Neurons in the Absence of Movement

Even when the animal is stationary, the inhibitory output neurons of the GPi and SNr are among the most active cells in the nervous system, with tonic discharge rates in the range of 50–100 Hz (see Figure 2). This tonic inhibition of neurons in the thalamus, superior colliculus, and PPN seems to provide stability to the system because transient suppression of activity in arm-related areas of GPi by local injection of muscimol, a $GABA_A$ receptor agonist, results in drift of the arm away from the hold position (Figure 3(b)). This drift can be corrected when something, such as a

change in the target light color, brings the error to the animal's attention. A similar instability occurs in the oculomotor system when muscimol is injected into the SNr, except that the eye movements off and back on the target are both rapid, producing what is called a saccadic jerk.

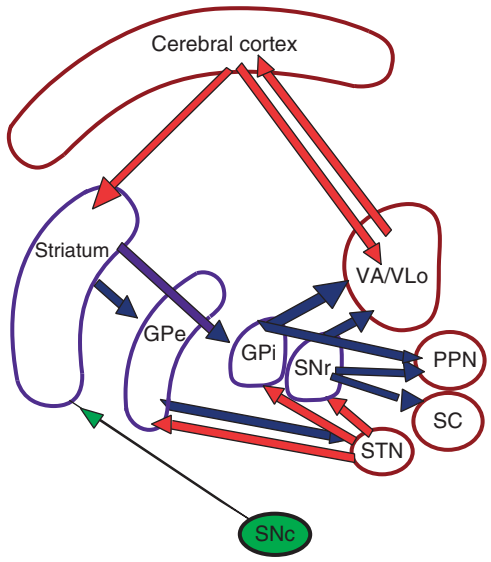

Figure 1 Basic connectivity from the cerebral cortex through the basal ganglia. GPe, external segment of the globus pallidus; GPi, internal segment of the globus pallidus; PPN, pedunculopontine region of the midbrain; SC, superior colliculus; SNc, compact portion of the substantia nigra; SNr, reticular portion of the substantia nigra; STN, subthalamic nucleus; VA/VLo, ventral anterior/ventrolateral nucleus of the thalamus.

Basal Ganglia Inhibition Changes During Movement

When a movement is to be made, activity changes in a select group of neurons in the SNr or GPi, cerebral cortex, superior colliculus, and presumably PPN. If the eyes are to produce a rapid movement (a saccade), the change is a decrease in SNr discharge, resulting in a disinhibition of the collicular neurons that are activated to produce saccade initiation. This pause in nigral activity occurs primarily when saccades are made to visual targets in the contralateral visual field and is enhanced when the target location must be remembered from a visual target presented briefly before a go cue. This rather simple reduction in inhibitory activity is adequate to facilitate movement in the oculomotor system, in which a rapid change in activity in agonist motor neurons drives the eye to the new location, but the movement is stopped by the viscoelastic properties of the eye and associated tissues.

When an arm movement is made, the changes in activity of the GPi neurons are more complicated, as is the spatiotemporal pattern of activity in the muscles used to control the multisegmented arm. When arm movements are tested in several movement directions, there is usually at least one for which the pallidal neuron's activity is at least initially reduced below the resting rate, although often followed by enhanced discharge, as shown in Figure 2. For the majority of movements, the GPi cell might actually have an increase in discharge. The reduction in GPi inhibitory activity facilitates the appropriate circuits to produce

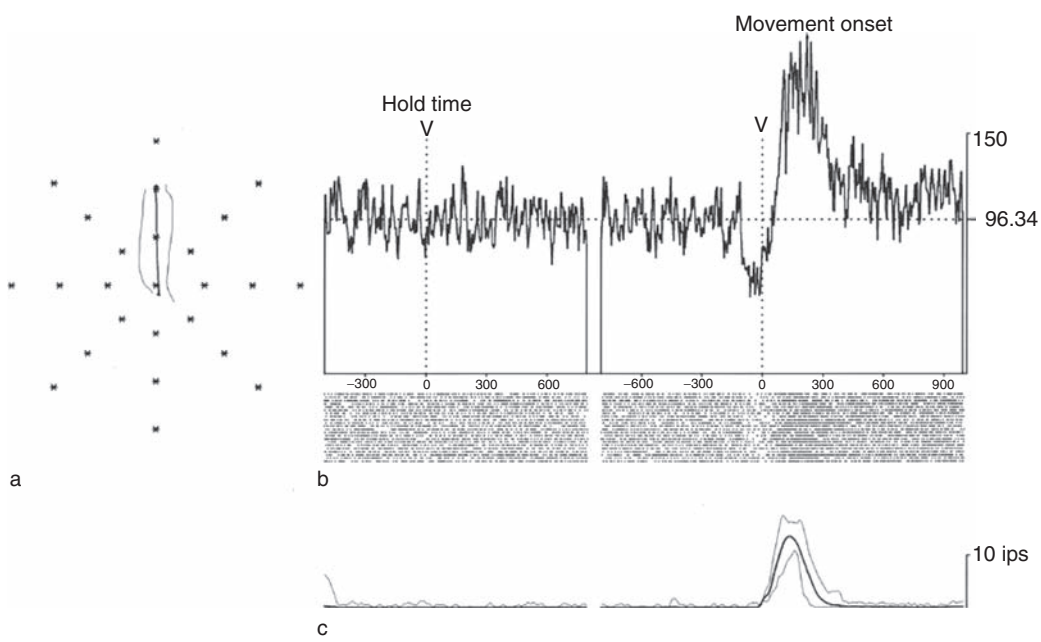

Figure 2 Pallidal discharge during hold and perimovement periods. (a) Thick line, average position; thin lines, extreme position of the hand; dots, potential targets. (b) Discharge of a neuron in GPi. Lower rasters: time of action potential (each dot) in each trial (line of dots), centered in time around the beginning of the hold time (time 0, left panel) or around the beginning of hand movement (time 0, right panel). Upper histogram, average discharge rate. (c) Hand speed; thick line, average across trials; thin line, extremes of hand speed.

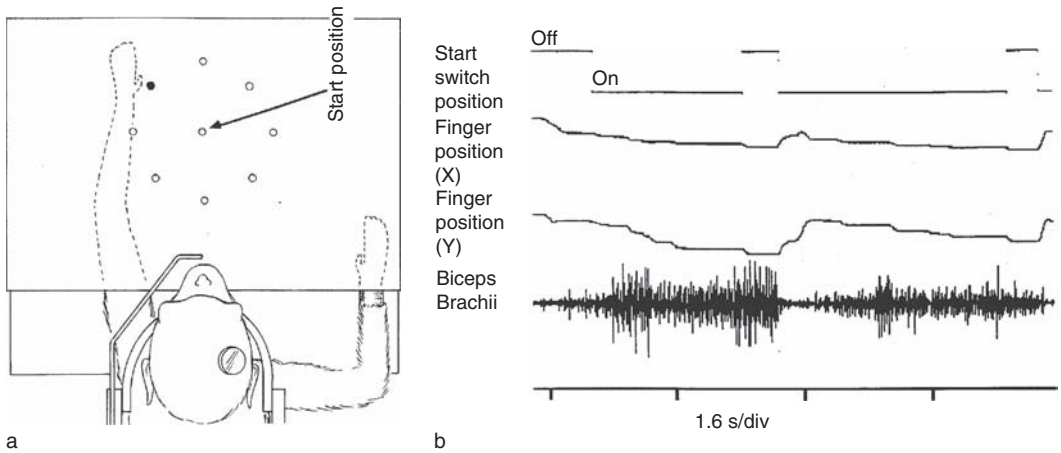

Figure 3 Instability of hand position induced by intrapallidal injection of muscimol. (a) Target hand position during the initial hold period (center, open circle, and start position) or the end of a targeted hand movement (filled circle). (b) Hand drift during hold time and active return to home when start switch color indicates error. Top, start switch state; middle two lines, x and y vectors of hand position; bottom, electromyographic activity of biceps brachii. (a) Reproduced from Turner RS, Owens J, and Anderson ME (1995) Directional variation of spatial and temporal characterics of limb movements made by monkeys in a two dimensional work space. *Journal of Neurophysiology* 74: 684–697, used with permission from The American Physiological Society. (b) Reproduced from Inase M, Buford JA, and Anderson ME (1996) Changes in the control of arm position, movement, and thalamic discharge during local inactivation in the internal globus pallidus of the monkey. *Journal of Neurophysiology* 75: 1087–1104, used with permission from The American Physiological Society.

muscle contraction in a select set of muscles, but the facilitation of GPi activity results in enhanced inhibition of thalamocortical neurons and reduced thalamic drive to selected areas of the motor cortices.

It has been proposed that the mixed pattern of suppression and activation of pallidal inhibitory output during limb movements is appropriate to produce a focused facilitation (disinhibition) of desired motor activation while suppressing undesired motor activity via the pallidal neurons whose discharge increases. This complicated change of activity in the GPi is appropriate to enhance the complicated spatiotemporal patterns of activation in agonist/antagonist groups of muscles associated with a variety of limb movements and to deal with changes in load encountered by the limbs.

Basal Ganglia Output Reflects Both Kinematic Variables and Movement Conditions

Basal Ganglia Output and Kinematic Variables

If basal ganglia output is to influence movement, we might predict a relationship between the kinematic variables of the movement, such as its direction, amplitude, or velocity, and the activity or change of activity in the basal ganglia output neurons during movements. On the other hand, changes in basal ganglia activity might reflect conditions under which the movement is made, such as with external sensory guidance, memory guidance, or a trigger for the next step in a sequence of movements.

The activity of movement-related GPi neurons, like that of neurons in motor areas of the cerebral cortex, is directionally tuned. That is, movement-related changes in activity are most pronounced for movements in a particular direction (referred to as the preferred direction). This preferred direction is consistent across movements of different speed and extent and does not vary when movements are made to remembered versus visible target locations.

Pallidal activity does vary with changes in the speed and/or extent of arm movements, especially in the preferred direction. This relation has been shown both by electrophysiological recordings of GPi activity in normal monkeys and by functional imaging studies in humans.

The importance of basal ganglia output to the control of arm movement can also be tested by examining the kinematics of movement when pallidal output neurons are temporarily inhibited with the $GABA_A$ agonist muscimol. In normal monkeys, this reversible inhibition generally results in slowed and hypometric movements. This could represent a failure both to inhibit antagonist muscles via enhanced pallidal activity in their circuit and to facilitate agonists via the temporally timed pause in activity in other circuits.

Task-Dependence of Motor-Related Changes in Basal Ganglia Activity

Beyond the kinematic aspects of movement that may be affected by the basal ganglia, it has been suggested that the basal ganglia are especially important in (1) controlling movements made under internal guidance,

such as those made to remembered target locations that are not indicated at the time the movement is to be made; (2) learning or combining movements into motor sequences; and (3) chunking activities, grouping them in a way that underlies habits or automatic behaviors.

Basal ganglia involvement in internally versus externally guided movements The magnitude of changes in pallidal activity is modified in different behavioral conditions, even if the movement direction, amplitude, and velocity are constant. When arm-related pallidal discharge is compared in normal monkeys under sensory-guided and memory-guided conditions, the activity of individual neurons is markedly influenced by the spatial or temporal memory demands of the task, although the sample as a whole may have no preferential difference in sensory- versus memory-guided conditions. This means that, although a group of neurons may have preferential changes for particular movement conditions, a range of movement conditions may be facilitated by selected groups of pallidal neurons.

The basal ganglia and movement sequences Several studies have indicated the importance of the basal ganglia in learning or implementing sequences of movements. Although other areas of the brain, including the cerebellum, may also play important roles in learning or implementing sequential movements, the basal ganglia consistently show activation during sequence learning and performance.

The locus of activation may change in the basal ganglia, however, as the sequence is learned. In functional imaging studies, the anterior dorsal (associative) portions of the striatum are active during initial explicit learning of a motor sequence, whereas the posterior ventral (sensorimotor) portion is more active when the sequence is overlearned. Furthermore, there is evidence that accuracy is positively correlated with the early signal in the anterior dorsal striatum, whereas reaction time is correlated (negatively) with the later signal in the posterior ventral putamen. In monkeys trained for varying amounts of time on different complex reaching sequences, the neurons in the caudate nucleus and putamen anterior to the anterior commissure are more activated while new motor sequences are learned, whereas neurons in the more posterior, sensorimotor putamen are more involved in the production of already learned sequences.

The basal ganglia and automatic movements or habits When movements – including complex sequences – are repeated sufficiently, they may be performed automatically. Experimentally, an automatic movement can be defined as one that is not disturbed by a concurrent cognitive or second motor task. The increase in errors produced during the concurrent performance of the second task may be used to evaluate automaticity.

Although several studies have suggested that the basal ganglia play a special role in the automatic phase of a learned task, these nuclei certainly are not uniquely active during automatic movements. Functional imaging studies, for example, show that the basal ganglia, like several other structures, are activated during both the acquisition and the automatic phases of a complex finger press task. In normal subjects, the high level of activation present in select areas of both the basal ganglia and the cortex during early stages of sequence learning may actually decrease and become more focused when task performance reaches an automatic stage. However, no new areas – cortical or subcortical – are uniquely activated when sequence performance becomes automatic.

The Influence of Dopamine on the Basal Ganglia and Movement

The role of dopamine in the basal ganglia is of special interest because it is primarily the dopaminergic nigrostriatal neurons that degenerate in Parkinson's disease. The major action of dopamine in the basal ganglia is on the striatal neurons, and the available data suggest that dopamine may influence both their intrinsic membrane properties and mechanisms of synaptic plasticity, such as long-term potentiation (LTP) and long-term depression (LTD).

In operant task conditions, dopaminergic neurons are active when animals receive an unexpected reward and briefly fall silent when an expected reward is absent. When the probability of reward differs for different targets in a task, the activity of primate midbrain dopamine neurons shifts based on the expected reward value for the target. Moreover, the gain of their stimulus-evoked responses reflects the variance of the reward value. And the activity of striatal neurons, which are known to receive dopaminergic input, also changes with target reward probability.

Pallidal neurons may show activity that is co-modulated with both movement direction and reward prediction. Furthermore, the predictive power of the pallidal activity changes dynamically during the course of a trial. It predicts reward more accurately at the beginning of the trial and is linked more closely to movement direction by the end of the trial. Such a combination of reward and kinematic signals could reflect both the cortico-basal ganglia-thalamocortical circuit in which the pallidal neuron participates and a synaptic weighting signal modified by dopamine on the basis of the reward. Thus, dopamine could

influence synaptic plasticity that couples reward or motivation to the activity of striatal neurons and, eventually, to the movement-related output of the basal ganglia.

Parkinson's Disease and Movement

The classic motor signs of parkinsonism include bradykinesia (slowed movement); a tremor at rest or during a static contraction; and rigidity, a tonic contraction of motor units, commonly resulting in flexion-biased postures of the trunk and limbs. Associated with these signs are deficits in balance, gait, and speech production. In general, patients do not demonstrate a strict weakness, nor do they report sensory loss. On more careful testing, the bradykinesia is usually linked to hypometric movements. Reduced movement amplitude, may, in part, be related to erroneous proprioception, which has been well-documented in parkinsonian subjects. Abnormal coupling of proprioceptive information with a motivational signal could lead to abnormal movement amplitude and velocity (the latter two usually covary).

The degree to which the basal ganglia are important for automaticity and habit learning also has been addressed in parkinsonian subjects, in whom usually automatic motor behaviors such as arm swing and postural adjustment commonly are depressed. Using the dual task already described to determine the degree to which sequential finger movements could be performed automatically (without disturbance from a concurrent task), parkinsonian subjects could achieve automatic performance of a four-step learned sequence. What most could not do, however, was to perform a longer sequence (12 steps) automatically, a task that most normal subjects performed successfully. Furthermore, functional imaging during automatic performance of the four-step task continued to show widespread activation in both the cortex and the basal ganglia of parkinsonian subjects, even when it was performed automatically, whereas the activation of these areas was more focused in normal subjects, who performed the four-step sequence automatically. Again, no new areas were activated in the basal ganglia of the parkinsonian subjects to achieve automaticity.

This continued widespread activation of cortical and subcortical areas, including the basal ganglia, during experimentally defined automatic motor performance of the short sequence in parkinsonian subjects is similar to that in normals during initial learning of either a short or a long task. Activity of the basal ganglia is not uniquely related to automaticity, but it is a part of a distributed network of activity that commandeers more of the brain when the task is novel and difficult and becomes more focused when the task is performed automatically.

Involuntary Movements after Levodopa Treatment

Since the early 1970s, the standard treatment for symptoms of Parkinson's disease has been oral administration of levodopa (L-DOPA), the precursor of dopamine that can cross the blood–brain barrier and be available in the brain parenchyma. Although this treatment (usually as Sinamet, in which L-DOPA is combined with a peripheral decarboxylase inhibitor to increase the central action) may be effective for several years, its effectiveness gradually diminishes and dyskinesias develop. These involuntary movements, which may sometimes appear similar to normal, puposeful movements, cannot be suppressed, and they markedly disrupt normal motor function.

During periods of dyskinesias, the tonic activity of the inhibitory basal ganglia output neurons in the GPi is unusually low and is interrupted by bursts. The absence of tonic pallidal inhibitory activity seems to leave the system unstable, with resulting uncontrollable movements.

One mechanism proposed to underlie the dyskinesias induced by chronic L-DOPA treatment is based on abnormalities in synaptic plasticity that appear in animal models after long-term treatment with L-DOPA. A rodent model of Parkinson's disease can be produced by destruction of the dopaminergic nigrostriatal system using 6-hydroxydopamine (6-OH DA). After chronic treatment with L-DOPA, abnormal involuntary movements appear in some animals but not in others. (This variability is also characteristic of L-DOPA-induced dyskinesias in parkinsonian subjects.) Slices from animals with and without dyskinesias are then used to compare synaptic plasticity in the two groups. LTP of the corticostriatal synapse, which can be produced in normals by high-frequency stimulation of the corticostriatal fibers, can be produced in slices from animals with or without dyskinesias if dopamine is present in the bathing medium. However, if the stimulus train is switched to low frequencies (1–2 Hz) in slices from animals without dyskinesias, the LTP is reversed (depotentiation), whereas in those from animals that developed dyskinesias, LTP cannot be reversed. It is as if the corticostriatal synapses in animals with dyskinesias become stuck in the potentiated state. Potentiated activation of striatal neurons, which have a net inhibitory action on the inhibitory GPi output cells, reduces their stabilizing tonic activity, and this could lead to the involuntary movements.

The failure to reverse LTP involves a D1-receptor-mediated mechanism, which is consistent with the fact that dopamine agonists that act primarily via the D2 receptor and are used to treat symptoms of Parkinson's disease do not tend to induce dyskinesias. Thus, although dopamine clearly is required to induce synaptic potentiation in the striatum, over the long term its rather nonspecific chronic action may lead to unwanted changes in synaptic plasticity and abnormal movements.

Summary

A combination of electrophysiological, anatomical, imaging, and lesion studies show that some of the circuits through the basal ganglia carry signals that are related to the kinematics of movement. Disruption of the output of these circuits leads to instability in motor control and abnormalities in the amplitude and velocity of targeted movements. Dopamine acts in the basal ganglia to modulate the signals of these circuits, and in Parkinson's disease, in which dopaminergic nigral neurons degenerate, movements are slowed and hypometric. Because dopamine release is related to the probability of success/reward, the basal ganglia are a site where motivational and kinematic signals can be integrated to scale synaptic strength, as must occur during motor learning.

See also: Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia: Evolution; Basal Ganglia and Oculomotor Control; Basal Ganglia: Acetylcholine Interactions and Behavior; Basal Ganglia: Physiological Circuits; Deep Brain Stimulation and Parkinson's Disease; Dopamine; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopaminergic Agonists and L-DOPA; Long-Term Depression: Cerebellum; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Parkinsonian Syndromes; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Arkadir D, Morris G, Vaadia E, and Bergman H (2004) Independent coding of movement direction and reward prediction by single pallidal neurons. *Journal of Neuroscience* 24: 10047–10056.

Grafton ST (2004) Contributions of functional imaging to understanding parkinsonian symptoms. *Current Opinions in Neurobiology* 14(6): 715–719.

Hikosaka O, Nakamura K, and Nakahara H (2006) Basal ganglia orient eyes to reward. *Journal of Neurophysiology* 95: 567–584.

Hikosaka O, Nakamura K, Sakai K, and Nakahara H (2002) Central mechanisms of motor skill learning. *Current Opinions in Neurobiology* 12(2): 217–222.

Inase M, Buford JA, and Anderson ME (1996) Changes in the control of arm position, movement, and thalamic discharge during local inactivation in the globus pallidus of the monkey. *Journal of Neurophysiology* 75: 1087–1104.

Lehericy S, Benali H, Van de Moortele PF, et al. (2005) Distinct basal ganglia territories are engaged in early and advanced motor sequence learning. *Proceedings of the National Academy of Sciences of the United States of America* 102: 12566–12571.

Middleton FA and Strick PL (2000) Basal ganglia and cerebellar loops: Motor and cognitive circuits. *Brain Research: Brain Research Reviews* 31: 236–250.

Mink JW (1996) The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology* 50: 381–425.

Picconi B, Pisani A, Barone I, et al. (2005) Pathological synaptic plasticity in the striatum: Implications for Parkinson's disease. *Neurotoxicology* 26: 779–783.

Turner RS and Anderson ME (1997) Pallidal discharge related to the kinematics of reaching movements in two dimensions. *Journal of Neurophysiology* 77: 1051–1074.

Turner RS, Owens JW, and Anderson ME (1995) Directional variation of spatial and temporal characteristics of limb movements made by monkeys in a two-dimensional work space. *Journal of Neurophysiology* 74: 684–697.

Basal Ganglia: Physiological Circuits

A Nambu, National Institute for Physiological Sciences, Okazaki, Japan

© 2009 Elsevier Ltd. All rights reserved.

Composition of the Basal Ganglia

The basal ganglia are composed of four nuclei (see Figure 1(a)):

1. The striatum, which comprises the caudate nucleus, putamen and ventral striatum (including the nucleus accumbens).
2. The globus pallidus, which consists of the external segment (GPe) and internal segment (GPi).
3. The subthalamic nucleus (STN).
4. The substantia nigra, which is divided into the pars reticulata (SNr) and pars compacta (SNc).

The basal ganglia receive inputs from wide areas of the cerebral cortex. Information processed in the basal ganglia returns mainly to the cerebral cortex, especially to the frontal lobe, via the thalamus (Figure 1(b)). Thus, the basal ganglia and the cerebral cortex form a loop circuitry (cortico-basal ganglia loop). Through this loop, the basal ganglia control cortical activity. Some output of the basal ganglia also descends to the brain stem. The basal ganglia can be grouped functionally into four categories:

1. Input nuclei: striatum and STN, which receive cortical inputs.
2. Output nuclei: GPi and SNr, which project outside the basal ganglia to the thalamus and the brain stem.
3. Connecting nucleus: GPe, which connects the input nuclei to the output nuclei.
4. Modulatory nucleus: SNc, which modulates the activity of the basal ganglia.

The caudate nucleus and putamen can be considered as a continuum in that they share similar cellular morphology and firing patterns. In rodents, they form a single entity. The striatum of primates and carnivores is separated by dense fiber bundles of the internal capsule into the caudate nucleus and putamen. The caudate nucleus receives inputs mainly from the prefrontal cortex, and the putamen from the motor cortices. The GPi and SNr can also be considered as a continuum separated by the internal capsule. In dolphins, the internal capsule runs more dorsally, and the GPi and SNr are fused together. In rodents and carnivores, the GPi is small and buried in the fibers of the internal capsule, thus classically named the entopeduncular nucleus, and the GPe is sometimes called simply the globus pallidus.

Circuitry of the Basal Ganglia

The striatum as the input nucleus receives excitatory glutamatergic inputs from the cerebral cortex. Striatal projection neurons can be classified into two groups on the basis of projection sites and neurotransmitters. Information in the striatum is transferred to the output nuclei, GPi/SNr, via two different pathways (Figure 2(a)): (1) the 'direct pathway,' in which striatal neurons containing γ-aminobutyric acid (GABA), substance P, and dynorphin project monosynaptically to the GPi/SNr and (2) the 'indirect pathway,' in which striatal neurons containing GABA and enkephalin project polysynaptically to the GPi/SNr by way of sequential connections with the GPe (composed of GABAergic neurons) and STN (composed of glutamatergic neurons). On the one hand, signals through the direct pathway have inhibitory effects on GPi/SNr neurons. On the other, signals through the indirect pathway have excitatory effects on GPi/SNr neurons because both striato-GPe and GPe–STN projections are inhibitory and STN–GPi/SNr projections are excitatory.

In addition to the canonical basal ganglia circuitry, the STN receives direct inputs from the cerebral cortex in a somatotopically organized manner and can thus be considered another input nucleus of the basal ganglia, suggesting a third pathway, the 'hyperdirect pathway,' in which STN neurons receiving cortical inputs project monosynaptically to the GPi/SNr (Figure 2(a)). The hyperdirect pathway conveys strong excitatory signals from the cortex to the GPi/SNr with faster conduction velocity than is possible via the direct or indirect pathway.

The output nuclei of the basal ganglia, GPi/SNr, are composed of inhibitory GABAergic neurons, and they fire at high frequency (40–100 Hz). Thus, neurons in the target nuclei, that is, the ventral anterior and ventral lateral nuclei (VA/VL) of the thalamus and brain stem nuclei, such as the pedunculopontine tegmental nucleus (PPN) and superior colliculus, are strongly tonically inhibited. VA/VL thalamic neurons project to the frontal cortex.

Dopaminergic neurons in the SNc project to the striatum and modulate differentially the activity of striatal projection neurons in the direct and indirect pathways (Figure 2(a)). Dopamine excites striatal neurons in the direct pathway through the dopamine D1 receptors, whereas it inhibits striatal neurons in the indirect pathway through dopamine D2 receptors.

The motor loop controlling movements of forelimbs and hindlimbs is well studied among the cortico-basal ganglia loops. This loop starts from the motor cortices (such as the primary motor cortex (MI), supplementary motor area (SMA), and premotor cortex (PM)), goes

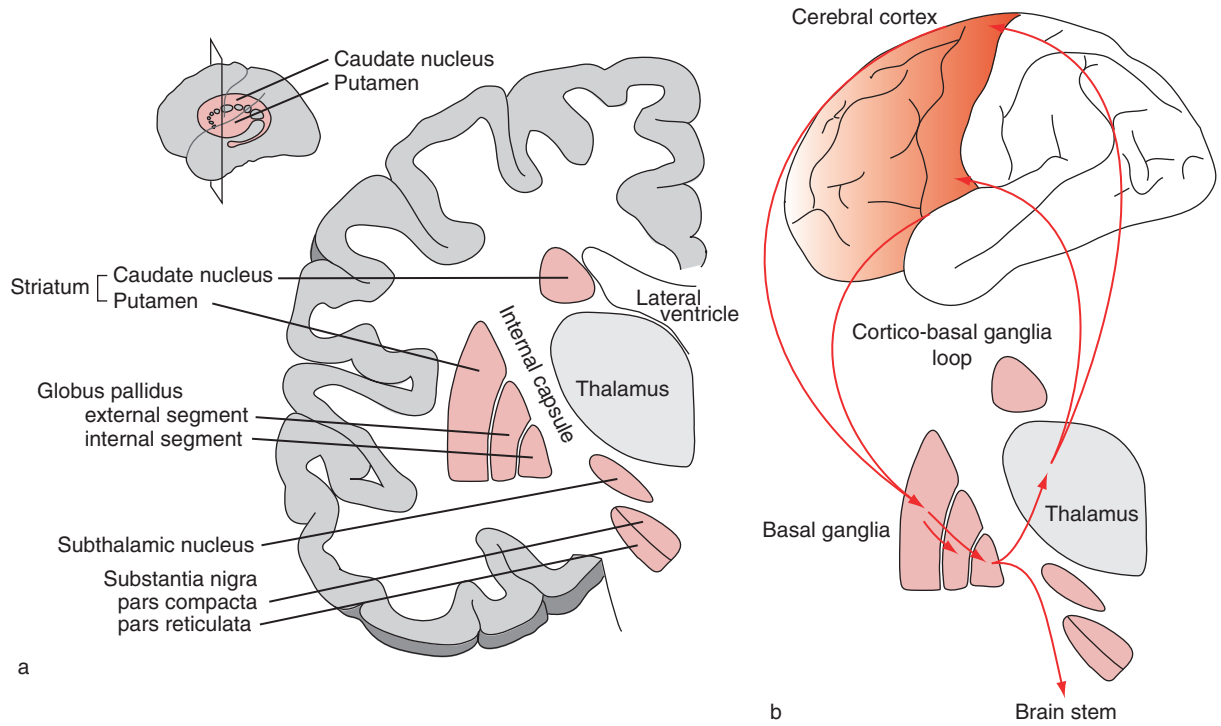

Figure 1 Basal ganglia. (a) Human basal ganglia shown in a coronal section and in a lateral view (inset). (b) The basal ganglia receive cortical inputs and return processed information to the cerebral cortex through the thalamus (cortico-basal ganglia loop) and to the brain stem.

through the somatomotor territories of the basal ganglia, and finally returns to the original cortices through the VA/VL of the thalamus. The somatomotor territories of the basal ganglia include the caudoventral putamen, the caudoventral GPe/GPi, and the dorsal STN, and each territory shows somatotopic organization (Figure 2(b)). In addition to the motor loop, the oculomotor loop, prefrontal loop (composed of dorsolateral prefrontal and lateral orbitofrontal loops), and limbic loop connect the cerebral cortical areas (the frontal eye field/supplementary eye field, prefrontal cortex and limbic cortex, respectively) with corresponding parts of the basal ganglia and thalamic nuclei and form multiple parallel, segregated, and functionally distinct but homologous loops (Figure 2(a)). Through these multiple loops, the basal ganglia control eye movements, higher brain functions, and emotions, as well as limb movements.

The aforementioned basal ganglia circuitry may be oversimplified. Other significant projections exist:

1. It is supposed that striatal neurons projecting to the GPe and GPi belong to different groups. However, tracing studies of a single neuron have shown the existence of neurons projecting to both the GPe and the GPi. It has also been reported that some striatal projection neurons express both D1 and D2 receptors. Moreover, electrophysiological experiments have failed to show the expected effect of dopamine on striatal projection neurons.
2. Thalamic neurons, such as in the VA/VL and centromedian-parafascicular complex (CM/Pf), project also to the striatum, suggesting a short circuit of the striato-GPi-thalamo-striatal loop (Figure 2(a)).
3. The cortico-basal ganglia loops are considered to be composed of several segregated and parallel loops. However, the projections overlap partially. Thus, these loops should be viewed more as a continuum than as subdivisions with strict boundaries.
4. The GPe gives rise to GABAergic projections to the GPi/SNr, striatum, and GPe itself through local axon collaterals. Thus, the GPe may be viewed as a center nucleus projecting to multiple sites of the basal ganglia.
5. The STN projects to the GPe as well as to the GPi, using axon collaterals. The STN and GPe have intimate interconnections, and interconnected groups of neurons in the GPe and STN innervate the same population of neurons in the GPi.

Striatum: One of the Input Nuclei of the Basal Ganglia

The striatum receives excitatory glutamatergic inputs from wide areas of the cerebral cortex. Cortical inputs maintain functional topography; for example,

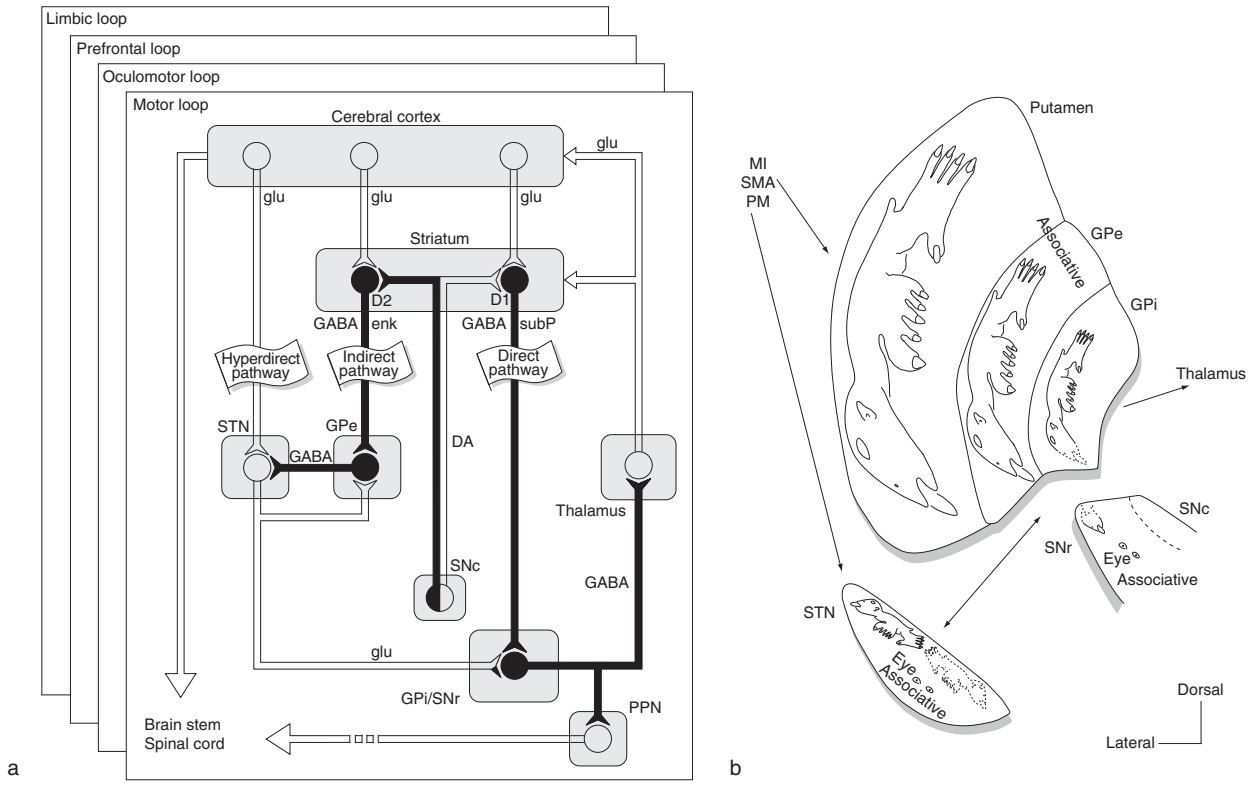

Figure 2 Circuitry of the basal ganglia. (a) Schematic diagram of the circuitry and neurotransmitters of the basal ganglia–thalamocortical circuitry. Open and filled symbols represent excitatory and inhibitory neurons, respectively. D1 and D2, dopamine receptors; DA, dopamine; enk, enkephalin; GABA, γ-aminobutyric acid; glu, glutamate; GPe, external segment of the globus pallidus; GPi, internal segment of the globus pallidus; PPN, pedunculopontine tegmental nucleus; SNc, substantia nigra pars compacta; SNr, substantia nigra pars reticulata; STN, subthalamic nucleus; subP, substance P. (b) Somatotopical organization of the basal ganglia related to the motor loop. MI, primary motor cortex; PM, premotor cortex; SMA, supplementary motor area. (a) Reproduced from Alexander GE and Crutcher MD (1990) Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in Neurosciences* 13: 266–271, with permission from Elsevier.

the motor and sensory cortices project to the caudal part of the putamen, and the prefrontal cortex projects to the caudate nucleus and the rostral part of the putamen. Hindlimb, forelimb, and orofacial areas of the MI project to the dorsal to ventral part of the caudal putamen in a somatotopically organized manner (Figure 2(b)). Corticostriatal afferents arise from small to medium-size pyramidal neurons located mainly in the upper half of layer V, with some being distributed in layer III. Some of corticostriatal afferents are axon collaterals of corticospinal neurons. During behavior, corticospinal neurons show muscle-like movement-related activity. However, corticostriatal neurons show more-selective activity to the parameters of behavioral tasks. Thus, both kinds of information are conveyed to the striatum. The striatum also receives excitatory glutamatergic inputs from the CM/Pf and VA/VL of the thalamus, dopaminergic inputs from the SNc, and serotonergic inputs from the dorsal raphe nucleus.

The striatum is composed of projection neurons that are the majority in number (80–95%) and interneurons (Figure 3). Projection neurons are inhibitory GABAergic medium-size neurons, and their dendrites are covered with numerous spines. Medium spiny projection neurons are usually silent and fire when they receive inputs, so they are described as phasically active neurons in *in vivo* unit recording. Projection neurons can be divided into two groups on the basis of the transmitters and projection sites, as described above. Interneurons lack spines and are classified into at least three groups: (1) cholinergic large aspiny neurons, (2) GABAergic aspiny neurons that contain parvalbumin, and (3) GABAergic aspiny neurons that contain somatostatin/nitric oxide synthase. Interneurons, as well as projection neurons, receive inputs from the cerebral cortex, thalamus, and SNc and finally project to projection neurons and control their activity (Figure 3). Cholinergic large aspiny neurons fire spontaneously at 2–10 Hz and are described as tonically active neurons in

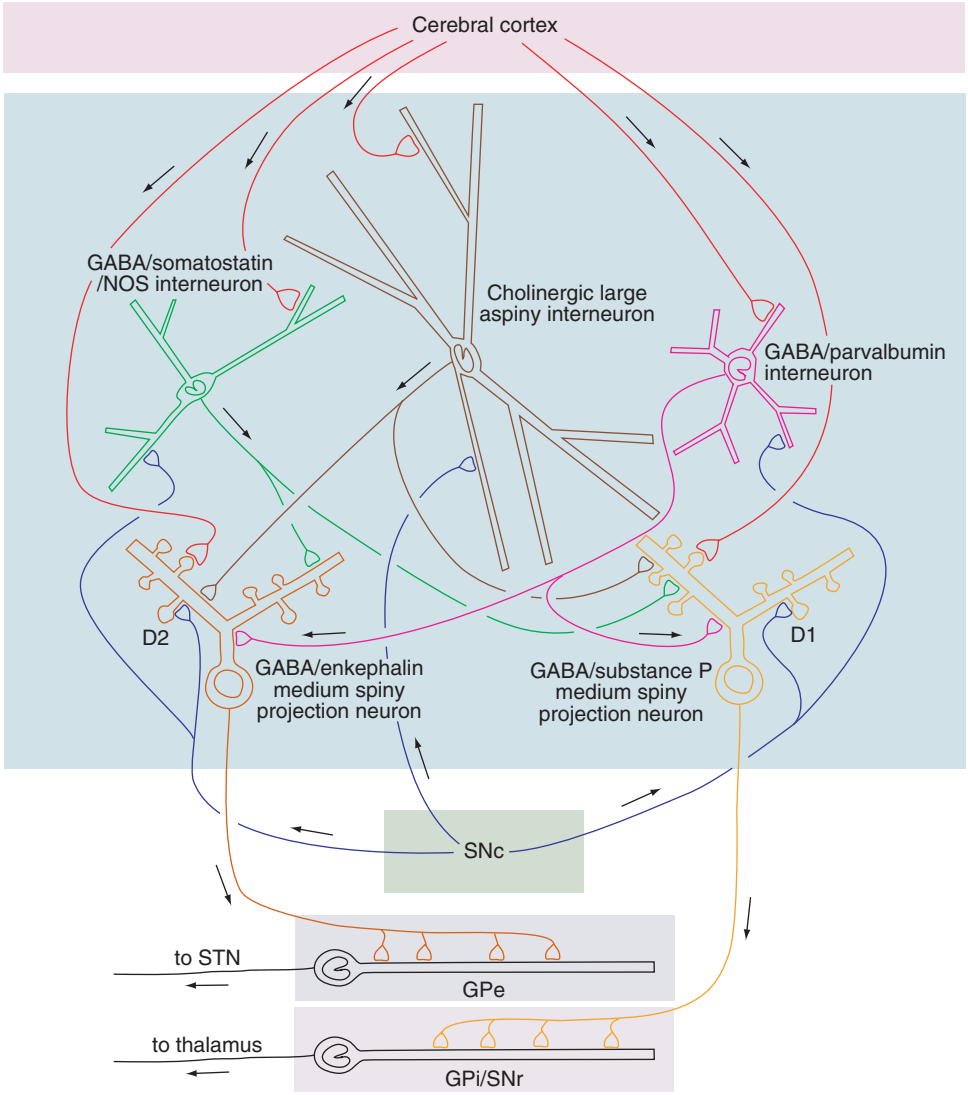

Figure 3 Circuitry within the striatum and its outputs. Spiny projection neurons receive cortical inputs directly and indirectly through interneurons and project to the external segment of the globus pallidus (GPe), internal segment of the globus pallidus (GPi), and substantia nigra pars reticulata (SNr). D1 and D2, D1 and D2 dopamine receptors; GABA, γ-aminobutyric acid; NOS, nitric oxide synthase; SNc, substantia nigra pars compacta; STN, subthalamic nucleus. Modified from Wilson CJ (2004) Basal ganglia. In Shepherd GM (ed.) *The Synaptic Organization of the Brain*, 5th edn., pp. 361–413. New York: Oxford University Press.

in vivo unit recording. The *in vivo* activity patterns of other interneurons remain to be elucidated.

Although the striatum looks very homogeneous, it can be divided anatomically into the μ-opiate receptorrich 'patch' compartment ('striosome') and the surrounding 'matrix' compartment. However, the functional significance of the compartmentalization is not fully understood.

GPe and STN: Two Strongly Interconnected Nuclei of the Basal Ganglia

Striatal neurons containing enkephalin and GABA project to the GPe (Figures 2(a) and 3). The GPe is composed of GABAergic neurons and fires spontaneously at high frequency (40–100 Hz). The GPe projects to the STN and GPi. It also projects to the GPe itself by local axon collaterals. The STN is composed of excitatory glutamatergic neurons and fires at medium rate (20–40 Hz). The STN receives excitatory inputs from the frontal lobe, including the motor cortices, in a somatotopically organized manner, and it receives inhibitory inputs from the GPe, and projects to the GPe, GPi, and SNr (Figures 2(a) and 2(b)). Thus, the STN can be considered an input nucleus of the cortico-STN–GPi/SNr hyperdirect pathway, as well as a relay nucleus of the striato-GPe–STN–GPi/SNr indirect pathway. The GPe projects to the GPi/SNr, suggesting another striato-GPe–GPi/SNr indirect pathway. Interconnections between the STN and GPe/GPi

form a fiber bundle called the subthalamic fasciculus (Figure 4).

GPi and SNr: Output Nuclei of the Basal Ganglia

Information processed in the basal ganglia reaches the GPi/SNr. The majority of output information returns to the cerebral cortex through the thalamus, while a minority descends to the brain stem. The pallidothalamic projections arise from the GPi as two fiber bundles, ansa lenticularis (from the lateral or rostral portion of the GPi) and lenticular fasciculus (from the medial or caudal portion of the GPi), merge at Forel's field H, form the thalamic fasciculus, and finally reach the thalamus (Figure 4). This projection terminates at the oral part of the ventral lateral nucleus (VLo), the parvicellular part of the ventral anterior nucleus (VApc) of the thalamus, and the lateral habenular nucleus, and gives axon collaterals also to the CM. The VLo and VApc project to the motor cortices including the MI, SMA, and PM; thus, wide areas of the motor cortices are under the control of the basal ganglia (the motor loop). The neurons projecting to the lateral habenular nucleus are located in the ventromedial border of the GPi and are different from those projecting to the thalamus. The lateral habenular nucleus projects to the raphe nuclei, ventral tegmental area, and SNc through the fasciculus retroflexus. The descending fibers depart from the ascending fibers in Forel's field H and project to the PPN (Figures 2(a) and 4). The PPN projects down to the spinal cord through the brain stem reticular formation and controls locomotion and posture.

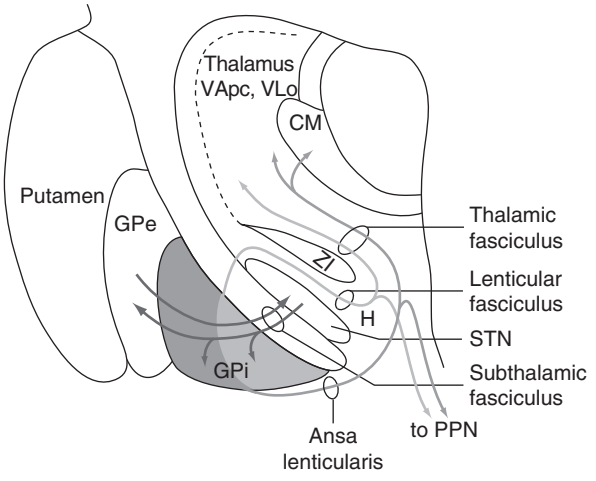

Figure 4 Output pathways of the internal segment of the globus pallidus (GPi) shown in a coronal section. CM, centromedian nucleus; GPe, external segment of the globus pallidus; H, Forel's field H; PPN, pedunculopontine tegmental nucleus; STN, subthalamic nucleus; VApc, parvicellular part of ventral anterior nucleus; VLo, oral part of ventral lateral nucleus; ZI, zona incerta. Reproduced from Parent A, Cossette M, and Lévesque M (2000) Anatomical considerations in basal ganglia surgery. In: Lozano AM (ed.) *Movement Disorder Surgery: Progress in Neurological Surgery,* vol. 15, pp. 21–30. Basel. Karger, with permission from S Karger AG.

The nigrothalamic projections terminate more rostrally than the pallidothalamic projections at the magnocellular part of the ventral anterior nucleus (VAmc), the medial part of the ventral lateral nucleus (VLm), and the paralaminar part of the medial dorsal nucleus (MDpl) of the thalamus. These thalamic nuclei project to the prefrontal cortex, frontal eye field, and supplementary eye field, forming the prefrontal and oculomotor loops. The SNr also projects to the superior colliculus (nigrotectal projection) and PPN (nigrotegmental projection).

On the one hand, ascending projections from the basal ganglia to the cerebral cortex through the thalamus control the learned movements of hands and the higher brain functions. On the other, descending projections to the motor centers in the brain stem control the innate motor functions, such as eye movements, trunk movements, locomotion, mastication, and vocalization.

The cerebellar nuclei project to the oral part of the ventral posterolateral nucleus (VPLo) and the caudal part of the ventral lateral nucleus (VLc) of the thalamus, which are more caudal than the pallidothalamic terminal fields. Thus, the SNr, GPi, and cerebellar nuclei project to different thalamic subnuclei, from rostral to caudal part of the thalamus. In the motor loop, the outputs from the GPi and cerebellar nuclei finally project to the motor cortices.

SNc: Modulatory Nucleus of the Basal Ganglia

The SNc is composed of dopaminergic neurons and projects widely to the basal ganglia and frontal cortex, especially to the striatum. Dopamine receptors are one of the G-protein-coupled receptors, identified as subtypes D1 through D5. They are classified into the D1 group (comprising D1 and D5), which activates adenylate cyclase, and the D2 group (D2, D3, and D4), which inhibits adenylate cyclase. Among striatal projection neurons, neurons in the direct pathway express mainly D1 receptors, and neurons in the indirect pathway have D2 receptors. Thus, dopamine has differential effects on these two different neuronal groups (Figures 2(a) and 3).

Functions of the Basal Ganglia

Physiological functions of the basal ganglia are not yet fully understood; the most probable hypothesis

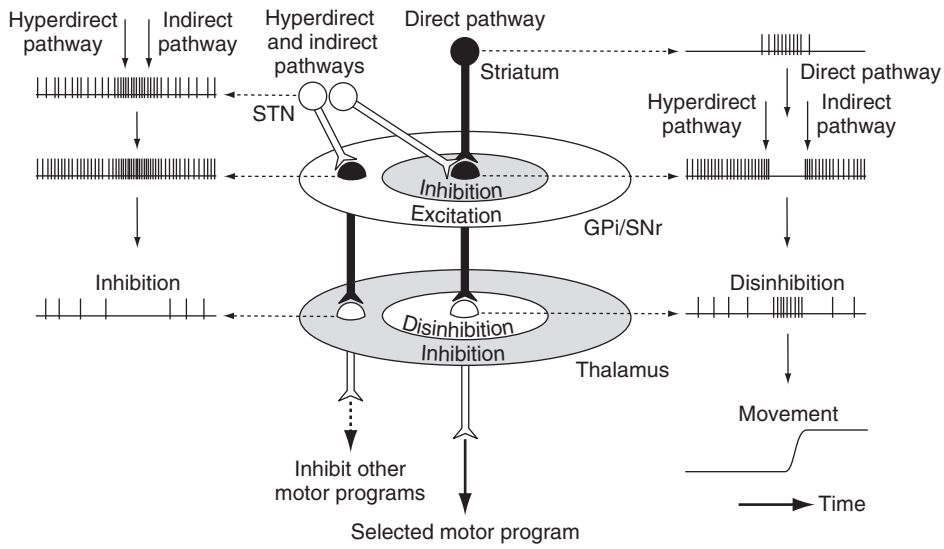

Figure 5 Spatial and temporal distributions of basal ganglia activity during behavior based on the focused selection model. Signals through the direct pathway inhibit internal segment of the globus pallidus/substantia nigra pars reticulata (GPi/SNr) neurons, activate thalamic neurons by disinhibition, and release a selected motor program. On the other hand, signals through the hyperdirect and indirect pathways have broad excitatory effects on GPi/SNr neurons in temporal and spatial domains, clearly initiate and terminate the selected motor program, and inhibit other unnecessary competing motor programs. STN, subthalamic nucleus. Reproduced from Nambu A (2007) Globus pallidus internal segment. In: Tepper JM, Abercrombie ED, and Bolam JP (eds.) *GABA and the Basal Ganglia: From Molecules to Systems: Progress in Brain Research*, vol. 160, pp. 135–150. Amsterdam: Elsevier, with permission from Elsevier.

of basal ganglia functions is the focused selection of intended movements. Neurons in the output nuclei of the basal ganglia are inhibitory GABAergic and fire at high frequency; thus, neurons in the target structures, such as the thalamus, are inhibited continuously (Figure 5). When striatal neurons are activated by cortical inputs, the striatal neurons inhibit GPi/SNr activity through the striato-GPi/SNr direct pathway. The continuous inhibition from the output nuclei to the target structures is temporally removed (disinhibition), and neurons in the target structures, including thalamic and cortical neurons, are activated, resulting in the release of a selected motor program. These mechanisms have been investigated in saccadic eye movements.

On the other hand, signals through the hyperdirect and indirect pathways have excitatory effects on the GPi/SNr and thus have inhibitory effects on thalamic neurons. Considering the conduction velocity, signals through the hyperdirect pathway first actively inhibit thalamic neurons, then those through the direct pathway disinhibit them, and finally those through the indirect pathway inhibit them again (Figure 5). Thus, signals through the hyperdirect and indirect pathways clearly control initiation and termination of the selected motor program.

In addition to this temporal aspect, the enhancement by differential inputs through the hyperdirect, direct, and indirect pathways may work in the spatial domain as well. Anatomical studies have shown that STN–GPi fibers arborize more widely and terminate on more proximal neuronal elements than do striato-GPi fibers. Signals through the hyperdirect and indirect pathways activate GPi/SNr neurons extensively, thereby inhibiting large areas of the thalamus, whereas signals through the direct pathway disinhibit thalamic neurons only in the center area (Figure 5). Thus, signals through the hyperdirect and indirect pathways inhibit thalamic neurons in the surrounding area, which are related to unnecessary, competing motor programs.

Because of the temporal and spatial inputs to the target structures through the hyperdirect, direct, and indirect pathways, only the selected motor program is executed at the selected timing, and other competing motor programs are canceled. The oculomotor, prefrontal, and limbic loops control the activity of corresponding cortical areas in a similar manner.

Summary and Conclusions

The basal ganglia circuitry can be summarized as follows: (1) The basal ganglia control not only voluntary movements but also other higher brain functions through the multiple and parallel cortico-basal ganglia loops. (2) The output of the basal ganglia is controlled differentially by the direct, indirect, and hyperdirect pathways. (3) The basal ganglia are proposed to be involved in the focused selection of intended movements.

During two decades, accumulated knowledge of the basal ganglia has revolutionized our understanding on this structure. It also has suggested new therapies for movement disorders. However, we still do not have

the complete answer to the simplest question: What is the primary function of the basal ganglia? We need more studies focused on the information flow through the cortico-basal ganglia loop to answer additional questions concerning what kind of information each pathway conveys and how such information is processed in each nucleus.

See also: Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia: Motor Functions; Basal Ganglia: Internal Organization; Deep Brain Stimulation and Movement Disorder Treatment; Dopamine Neurons: Reward and Uncertainty; Dystonia: Myoclonus–Dystonia; Dystonia: Classification, Genetics and Therapeutics; Huntington's Disease; MPTP Parkinsonism Model; Parkinsonian Syndromes; Procedural Learning: Striatum; Striatum: Internal Physiology.

Further Reading

Alexander GE and Crutcher MD (1990) Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in Neurosciences* 13: 266–271.

Bar-Gad I and Bergman H (2001) Stepping out of the box: Information processing in the neural networks of the basal ganglia. *Current Opinion in Neurobiology* 11: 689–695.

DeLong MR (1990) Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences* 13: 281–285.

Hikosaka O, Takikawa Y, and Kawagoe R (2000) Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological Reviews* 80: 953–978.

Mink JW (1996) The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology* 50: 381–425.

Nambu A, Tokuno H, and Takada M (2002) Functional significance of the cortico-subthalamo-pallidal "hyperdirect" pathway. *Neuroscience Research* 43: 111–117.

Bayesian Cortical Models

R P N Rao, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A fundamental challenge faced by the brain is to act on the basis of noisy sensory information and incomplete knowledge about the world. For example, a mouse seeking food in a natural environment must combine information about its current location estimated from sensory information such as whisker movements, sight, and odor with prior knowledge about places, paths, and potential food locations in the environment. Similarly, a tennis player must integrate prior knowledge about where her opponent tends to hit the ball with visual information about the current trajectory of the ball before deciding on her next stroke. A driver on a highway must integrate visual information about other cars and his own self-motion with prior knowledge about the dynamics of the car being driven, lane-changing rules, and typical driving patterns of other drivers. In each of these cases, the central problem that the brain needs to solve is one of inferring task-relevant attributes about the world in the presence of noisy and uncertain sensory information. Bayesian inference provides a mathematically rigorous way to combine noisy sensory information with prior knowledge, leading to revised beliefs about the current state of the world. Considerable evidence exists from psychophysical studies that the brain may indeed be employing Bayesian inference for perception and sensorimotor integration. These studies typically rely on Bayesian models formulated at an abstract mathematical level. The challenge is to understand how such models can be implemented by networks of neurons. This article reviews some recently proposed models for implementing Bayesian inference in cortical circuitry.

Bayesian Inference

Bayesian inference is based on a 240-year-old result known as Bayes' theorem, named after Reverend Thomas Bayes (1702–61), who first proposed a special case of the theorem. The theorem can be derived from the definition of joint probability of two random variables, say X and I:

$$P(X, I) = P(X|I)P(I) = P(I|X)P(X) \qquad [1]$$

The second equality implies that

$$P(X|I) = \frac{P(I|X)P(X)}{P(I)} \qquad [2]$$

which is Bayes' theorem for the two random variables X and I. To see how the theorem can be used for inference, suppose that X represents a task-relevant attribute such as the anticipated location of the ball in the tennis example above. Let I denote new visual information based on observing the opponent hit the ball. $P(X)$ is the prior distribution over locations, representing prior knowledge about where the opponent tends to hit the ball based on past experience with this opponent. With Bayes' theorem, a new distribution over possible ball locations can be inferred that takes the new visual input I into account: this is the posterior $P(X|I)$, which combines the prior $P(X)$ and the input likelihood $P(I|X)$ according to the above equation (the term $P(I)$ is a normalization constant which can be computed by integrating or summing the numerator $P(I|X)\ P(X)$ over all locations X). Bayes' theorem is useful because in most cases, it is easier to estimate $P(I|X)$ (e.g., by gathering statistics about I for a fixed X) than to estimate directly $P(X|I)$.

Bayesian inference is extremely versatile because it can be applied in myriad different situations. For example, one can generalize the example above to sequential Bayesian inference given inputs evolving over time: if I_1 is the input received at time step 1, then $P(X|I_1) = k_1 P(I_1|X)P(X)$, where k_1 is the normalization constant. Now if a new input I_2 is received at time step 2, the new posterior is given by

$$\begin{aligned} P(X|I_1, I_2) &= k_2 P(I_1, I_2|X)P(X) \\ &= k_2 P(I_2|X)P(I_1|X)P(X) \\ &= kP(I_2|X)P(X|I_1) \end{aligned} \qquad [3]$$

where it was assumed that the inputs are independent given X and where k and k_2 are normalization constants. Thus, the new posterior can be computed recursively by multiplying the new input likelihood with the posterior from the previous time step. Similarly, consider the problem of estimating a quantity X (e.g., location of a stimulus) from inputs from two different modalities (e.g., audition and vision). If I_a and I_v are the two inputs, the posterior can once again be estimated as

$$P(X|I_v, I_a) = k_3 P(I_v|X)P(I_a|X)P(X) \qquad [4]$$

assuming the inputs are independent given the source X. Note that this equation predicts a multiplicative interaction between the two input modalities and the prior distribution over X, a prediction that has received support from several psychophysical studies.

The above examples involved a few random variables and dependencies between them. The world that animal brains must cope with requires large

numbers of random variables to describe it with complex dependencies between them. Researchers in the fields of artificial intelligence and statistics have developed general algorithms for Bayesian inference over any set of random variables given 'evidence' (input information) regarding a subset of random variables. These algorithms are typically based on describing the domain of interest in terms of a 'graphical model' wherein the nodes of a graph denote random variables and the connections between the nodes represent dependencies between the random variables as characterized by conditional probabilities. A prominent algorithm for Bayesian inference is the belief propagation algorithm, which works by passing 'messages' (probability distributions) between the nodes of a graphical model. Messages received at each node are combined to yield the posterior probability distribution of the random variable associated with that node in the light of given input data or evidence. The challenge faced by computational neuroscientists is to understand how probabilities could be represented in the brain and how general Bayesian inference algorithms such as belief propagation can be implemented in neural circuits. Three classes of models addressing this issue are reviewed below.

Log Probability Models

One of the earliest suggestions regarding neurons representing probabilities was made by Barlow in 1969. Barlow proposed that neurons in the retina represent the certainty that a particular trigger feature (e.g., increase in light intensity) is present in the input stimulus. Specifically, Barlow suggested that the neurons signal the negative logarithm of the probability of the null hypothesis that the trigger feature is absent. Thus, a low response indicates a high degree of certainty that the trigger feature is present. Such a model generates responses that bear similarities to the responses of cat retinal ganglion cells.

Several recent cortical models for representing probabilities with neurons can be seen as variants of Barlow's retinal model. In these models, cortical neurons are assumed to represent the logarithm of probabilities or logarithm of ratios of probabilities. The reason for a logarithmic representation is that such a representation converts multiplication and division of probabilities to addition and subtraction, respectively, operations that can be implemented readily using conventional models of neurons (such as integrate-and-fire models).

Inspired by data from decision-making studies in humans and monkeys, Gold and Shadlen have proposed a model for neurons in cortical lateral intraparietal and frontal eye field areas that interprets their responses as representing the log likelihood ratio between two alternatives. Their model is based on the observation that the responses of neurons in these areas exhibit a behavior similar to a random walk to a fixed threshold. The neuron's response increases given evidence in favor of the neuron's preferred hypothesis and decreases when given evidence against that hypothesis, resulting in an evidence accumulation process similar to computing a log likelihood ratio over time. Gold and Shadlen developed a mathematical model to formalize this intuition. They have shown that under reasonable assumptions, the difference in responses between pools of neurons preferring opposing stimuli (e.g., leftward vs. rightward motion) is proportional to the log likelihood ratio between the two alternatives represented by the stimuli (**Figure 1**). Thus, downstream neurons, such as those in the lateral intraparietal area, need only add up these differences over time for their responses to be proportional to the log likelihood ratio between two alternatives. Decision making is accomplished by comparing the evolving neural response (denoting log likelihood ratio) to a decision threshold.

An extension of the Gold and Shadlen model to the case of spiking neurons has been proposed by Deneve. In this model, each neuron is assumed to represent the log 'odds' ratio for a preferred binary-valued state, that is, the logarithm of the probability

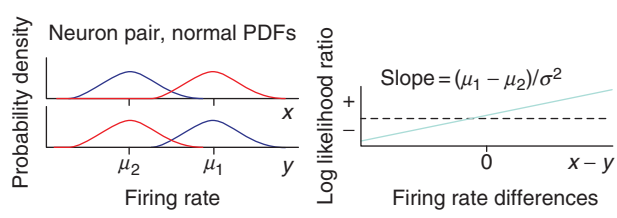

Figure 1 Relationship between neuronal response and log likelihood ratio according to the Gold–Shadlen model. The graphs illustrate that the difference in spike rates from two appropriately chosen neurons (or neural ensembles) can approximate the log likelihood ratio favoring one hypothesis over another. (Left, top plot) The red and blue curves are probability density functions (PDFs) that describe the hypothetical distributions of responses x from a sensory neuron when either hypothesis h_1 or h_2, respectively, is true (h_1, red; h_2, blue). Note that this neuron responds best when h_1 is true (red curve). The PDFs for a second neuron with responses y are shown below. This 'antineuron' responds best when h_2 is true (blue curves). The PDFs are assumed to be normal distributions with means μ_1 and μ_2 as shown and variance σ^2. (Right) The green curve represents the log likelihood ratio computed from the firing rates x and y of the neuron–antineuron pair. Note that the log likelihood ratio is proportional to the difference ($x-y$) in the firing rates. Knowledge of the PDFs is not required to interpret this difference: a positive value indicates support for h_1; a negative value indicates support for h_2. Adapted from Gold JI and Shadlen MN (2001) Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences* 5(1): 10–16.

that the preferred state is 1 over the probability that the preferred state is 0, given all inputs seen thus far. To promote efficiency, each neuron fires only when the difference between its log odds ratio and a prediction of the log odds ratio (based on the output spikes emitted thus far) reaches a certain threshold.

A potential drawback of a log probability ratio representation is that it can become cumbersome when the number of hypotheses being considered is large, given the large number of ratios, corresponding to different combinations of hypotheses, that need to be represented. An alternative approach is to explicitly represent the probability (or log probability) of a hypothesis or random variable using a single neuron. For example, if the random variable is direction of motion, a population of neurons, each preferring one particular direction of motion, could be used to represent the probability distribution over directions of motion. The firing rate of each neuron in this model is proportional to the probability (or log probability) of the direction of motion preferred by that neuron. The distribution of firing rates over the population can be regarded as a sampled representation of the probability distribution over the continuous space of motion direction. A model for Bayesian inference based on such a representation was proposed by Rao. A recurrent network of leaky integrator neurons was shown to be capable of implementing a simple form of the belief propagation algorithm over time. Recurrent connections between neurons encode the transition probabilities between the states represented by the neurons while the feedforward connections from the input neurons encode the likelihood function. Bayesian inference was approximated by implementing the belief propagation equations in the logarithmic domain; this allows the equations to be implemented by leaky integrator neurons, causing the output response of each neuron to become proportional to the log posterior probability of the state encoded by the neuron (Figure 2). This model was later extended to the case of exact inference in spiking neurons and hierarchical networks (see the section titled 'Hierarchical models').

Population Coding Models

A second class of models for representing and performing inference using probability distributions is based on the idea of population coding. Whereas the models discussed above utilize single neurons to directly represent probabilities or probability ratios, population coding models assume that probability distributions are implicitly represented by populations of neurons. For example, one class of models uses basis functions (or 'kernels') to represent probability distributions within neuronal ensembles. In this approach, a distribution $P(X)$ over stimulus X is represented implicitly in neural responses as a linear combination of basis functions:

$$P(X) = \sum_i r_i b_i(X) \qquad [5]$$

where r_i is the normalized response (firing rate) and b_i the implicit basis function associated with neuron i in

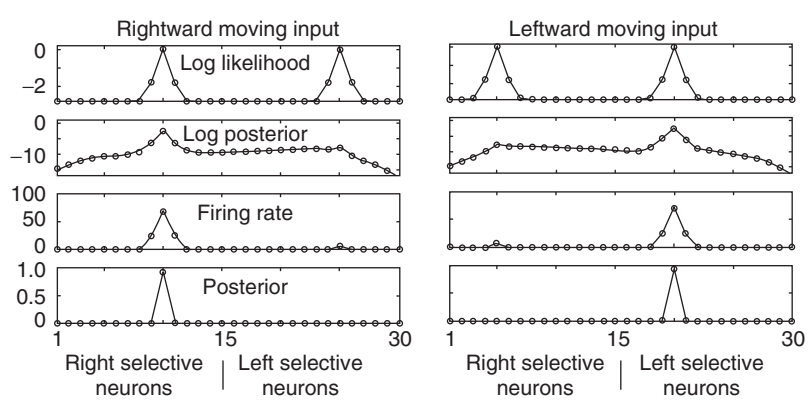

Figure 2 Responses from a log probability model for Bayesian motion detection. The plots depict responses from a log probability model for belief propagation proposed by Rao. A network of 30 neurons receives inputs from a one-dimensional model retina. Inputs consist of bars moving rightward or leftward. The receptive field of each neuron is localized to one particular spatial location on the retina using a Gaussian spatial filter. The input likelihood function for a neuron is proportional to the product of its spatial filter and the current image. The recurrent synaptic connections between neurons represent the transition probabilities of rightward and leftward motion in such a way that the first 15 neurons are selective for rightward motion and the other 15 are selective for leftward motion. Left, log likelihood, log posterior, firing rates, and posterior probabilities computed by the model network for a rightward moving stimulus when it arrives at the central image location. Note that the log likelihoods are the same for the left and right selective neurons, but their outputs correctly reflect the direction of motion due to recurrent interactions in the network. Right, the same plots for a leftward moving bar as it reaches the central location. From Rao RPN (2004) Bayesian computation in recurrent neural circuits. *Neural Computation* 16(1): 1–38.

the population. The basis function of each neuron is typically assumed to be linearly related to the tuning function $f_i(x)$ of the neuron as measured in physiological experiments. A limitation of the basis function approach above is that due to its additive nature, it cannot represent distributions that are sharper than the component distributions. A more recent model proposed by Jazayeri and Movshon combines the above idea with a log probability representation. In this model, the firings rates of neurons are assumed to be governed by an exponential distribution (such as Poisson), and the log probability distribution over stimulus X is assumed to be represented as

$$\log P(X) = \sum_i r_i \log f_i(X) \qquad [6]$$

A second class of population coding models is based on the use of a Bayesian decoding strategy to estimate the stimulus X (or its distribution), given a set of responses r_i. For example, in the distributional population coding (DPC) model of Zemel, Dayan, and Pouget, the responses are assumed to depend on general distributions over X, and a maximum *a posteriori* probability distribution is computed from possible distributions over X (Figure 3). The best estimate in this method is not a single value of X but an entire distribution over X, which is assumed to be represented by the neural population. The underlying

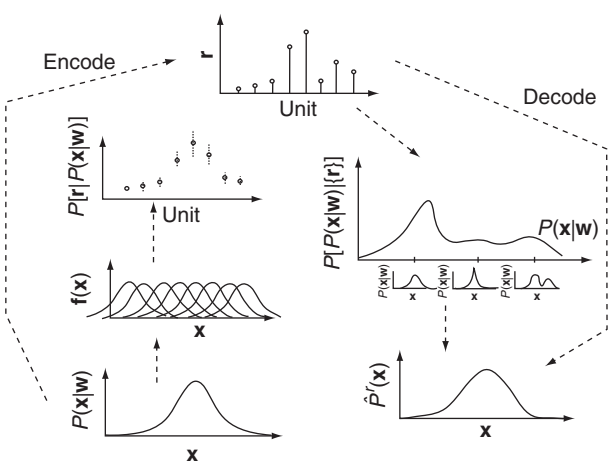

Figure 3 Distributional population coding (DPC) model. A set of firing rates **r** (topmost plot) may be interpreted as implicitly encoding a probability distribution $P(\mathbf{x}|\mathbf{w})$ over a variable **x** (the variable **w** represents any information that **x** depends on). Given a set of rates **r**, one can use Bayesian methods to obtain a probability distribution over possible implicit distributions, $P[P[\mathbf{x}|\mathbf{w}] \mid \mathbf{r}]$. The final decoded distribution $\hat{P}^r(x)$ is the maximum *a posteriori* distribution, obtained from the peak of this distribution over distributions. The error bars in the plot capture the variances in the firing rates of the units. From Zemel RS, Dayan P, and Pouget A (1998) Probabilistic interpretation of population codes. *Neural Computation* 10(2): 403–430.

goal of representing entire distributions within neural populations is common to both the DPC approach and the log probability models described above. However, the approaches differ in how they achieve this goal: the DPC method assumes prespecified tuning functions for the neurons and a sophisticated decoding operation to extract probability values whereas log probability models rely on exponential or linear decoding operations.

Many of the above models cannot disambiguate between cases in which there are multiple peaks in the decoded distribution due to uncertainty about a single stimulus or due to simultaneous multiple stimuli present in the input. A more recent model proposed by Sahani and Dayan, the doubly distributional population coding model, handles this problem by encoding probability distributions over a function $m(x)$ of the input x rather than distributions over x itself. The greater representational capacity of this method, however, comes at the expense of more-complex encoding and decoding schemes.

The distributional coding models discussed above were geared primarily toward representing probability distributions. More recent work by Zemel and colleagues has explored how distributional codes could be used for inference as well. In their approach, a recurrent network of leaky integrate-and-fire neurons maps an input population of spikes $R(t)$ to an output population of spikes $S(t)$. The network is trained to capture the probabilistic dynamics of a hidden random variable $X(t)$. This is done by adapting the weights of the neural network in such a way that the output decoding distribution $Q(X(t)|\mathbf{S}(t))$ is as close as possible to the input distribution $P(X(t)|\mathbf{R}(t))$, where $\mathbf{R}(t)$ and $\mathbf{S}(t)$ are vectors representing the input and output spike trains, respectively. The advantage of this approach is that the decoding process may allow a higher-fidelity representation of the output distribution than direct representational schemes such as the log probability models above. On the other hand, since the probability representation is implicit in the neural population, it becomes harder to map mathematically derived inference algorithms such as belief propagation to neural circuitry.

A different model by Ma and colleagues addresses the problem of inference using probabilistic population codes. In this model, the posterior distribution of a stimulus X given a population response vector of firing rates $\mathbf{r}$ is given by Bayes' theorem:

$$P(X|\mathbf{r}) = kP(\mathbf{r}|X)P(X) \qquad [7]$$

where k is a normalizing constant, $P(\mathbf{r}|X)$ is the firing rate distribution given X, and $P(X)$ is the prior distribution over X. Thus, the variability in neuronal

responses to a fixed stimulus is advantageously used to represent a posterior distribution over the stimulus. Assuming that this variability is captured by independent Poisson statistics (or other Poisson-like distributions from the exponential family), Bayesian inference can be performed by simply adding population responses on a neuron-by-neuron basis (Figure 4). For example, going back to the problem of estimating a quantity X (e.g., location of a stimulus) using inputs $\mathbf{r}_1$ and $\mathbf{r}_2$ from two different modalities (e.g., vision and audition), the posterior $P(X|\mathbf{r}_1, \mathbf{r}_2) \propto P(\mathbf{r}_1|X)\, P(\mathbf{r}_2|X)\, P(X)$ can be computed by a third population of neurons whose firing rate vector $\mathbf{r}_3$ is given by $\mathbf{r}_3 = \mathbf{r}_1 + \mathbf{r}_2$ (assuming a uniform prior distribution over X). This property, which is similar in spirit to the additive properties of log probability models, is a consequence of working with the exponential family of distributions (a related idea was explored by Hinton and Brown in their restricted Boltzmann machine model).

Hierarchical Models

There has been considerable interest in the neural implementation of hierarchical models for inference. Part of this interest stems from the fact that hierarchical models often capture the multiscale structure of input signals such as images in a very natural way (e.g., objects are composed of parts, which are in turn composed of subparts, etc.). A hierarchical

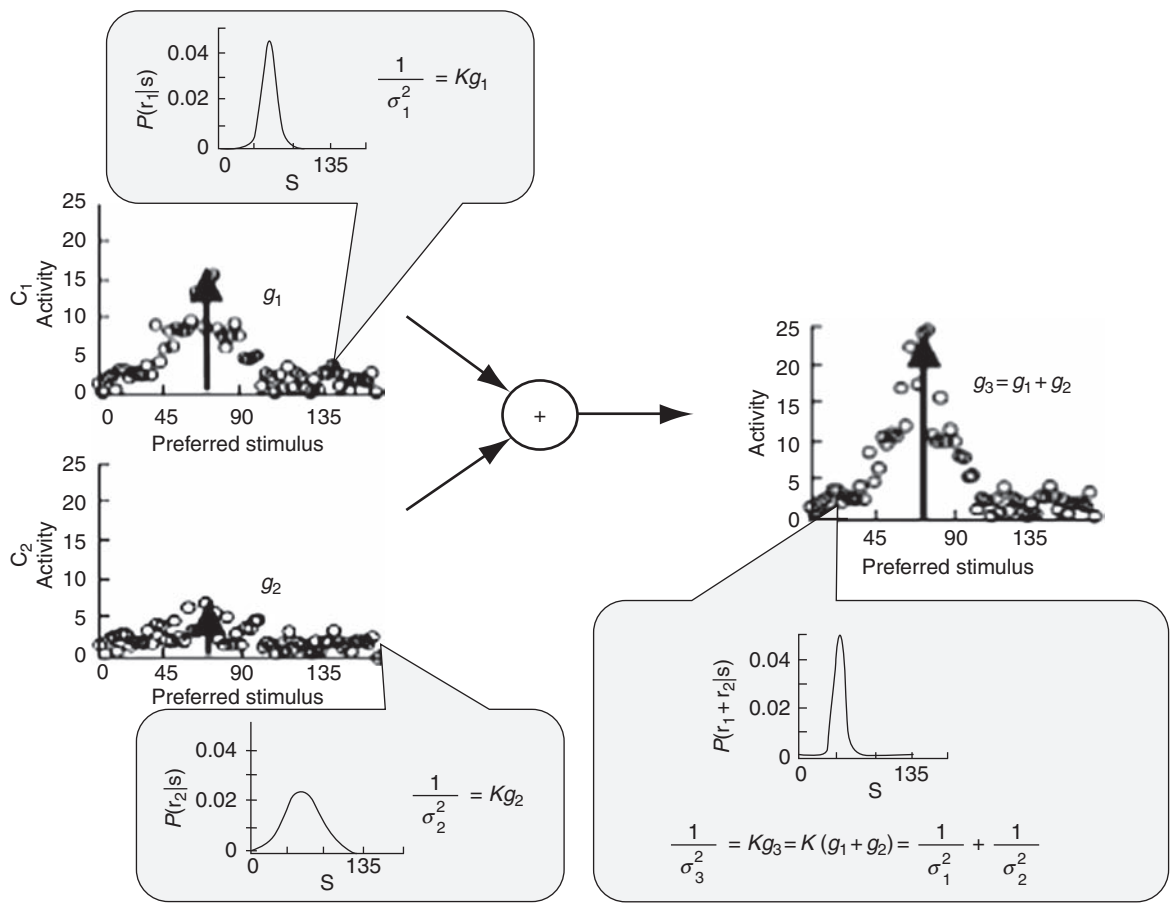

Figure 4 Probabilistic population codes. The figure depicts Bayesian inference with probabilistic population codes for Gaussian probability distributions and Poisson variability in neural responses. The left plots show population activity (firing rates) in response to two cues C_1 and C_2 (e.g., audition and vision) related to a stimulus s. Each of these implicitly encodes a Gaussian probability distribution with variances σ_1^2 and σ_2^2 inversely proportional to the gains g_1 and g_2 of the population codes (K is a constant depending on the width of the tuning curves of neurons and the number of neurons). r_1 and r_2 denote firing rates. Adding these two population codes leads to the population activity shown on the right. This output also encodes a probability distribution with a variance inversely proportional to the gain, which is equal to $g_1 + g_2$. Since g_1 and g_2 are in turn inversely proportional to σ_1^2 and σ_2^2 respectively, the inverse variance of the output code is the sum of the inverse variances associated with C_1 and C_2. This is the variance expected from optimal Bayesian inference using Gaussian distributions in eqn [4] in the text (assuming a uniform prior distribution). Thus, taking the sum of two population codes is equivalent to taking the product of their encoded distributions. From Ma WJ, Beck JM, Latham PE, and Pouget A (2006) Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9: 1432–1438.

decomposition often results in greater efficiency, both via parsimony of representation (e.g., a large number of objects can be represented by combining a small set of parts in many different ways) and by speeding up learning. A second motivation for hierarchical models has been the evidence from anatomical and physiological studies that cortical areas are reciprocally connected via feedforward and feedback connections, and many regions of the primate cortex appear to be hierarchically organized (e.g., visual cortex and motor cortex).

Dayan, Hinton, and colleagues investigated a hierarchical model network called the Helmholtz machine that uses feedback connections from higher to lower levels to instantiate a probabilistic generative model of its inputs. An interesting proposed learning method with the suggestive name of 'wake–sleep' algorithm involves learning the feedback weights during a 'wake' phase based on inputs and the feedforward weights in the 'sleep' phase based on 'fantasy' data produced by the feedback model. Although the model employs feedback connections, these are used only for bootstrapping the learning of the feedforward weights (via fantasy data). Perception involves a single feedforward pass through the network.

A hierarchical network that does employ feedback for inference was explored by Lewicki and Sejnowski. In this model, each unit encodes a binary state, and the probability that a unit's state S_i is equal to 1 depends on the states of its 'parents' (units sending connections to unit i), denoted by $pa[S_i]$, via

$$P(S_i = 1|pa[S_i]; \mathbf{W}) = h\left(\sum_j w_{ji} s_j\right) \qquad [8]$$

where $\mathbf{W}$ is the matrix of synaptic weights, w_{ji} is the synaptic weight from S_j to S_i, and h is the noisy OR function $h(x) = 1 - \exp(-x)$ $(x \geq 0)$. Bayesian inference is performed using a procedure known as Gibbs sampling, which involves sampling the value of a particular unit in the network given previously sampled values for the rest of the units and repeating this process for different units. After several iterations, these values will be typical samples from the posterior distribution of states in the network.

A different hierarchical cortical model based on the idea of predictive coding was suggested by Rao and Ballard. In this model, feedback connections convey predictions of expected activities at lower levels while the feedforward connections carry the residual errors between the predictions and actual activities; these errors are used to correct the state estimates at each level (Figure 5(a)). The state estimate at each level can be regarded as the mean of a Gaussian distribution over states at that level. The model can be viewed

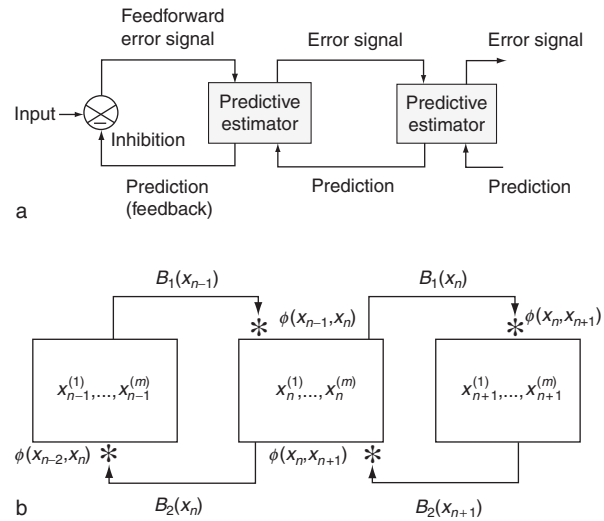

Figure 5 Hierarchical predictive coding and particle filtering models. (a) Schematic depiction of the hierarchical predictive coding model of the visual cortical pathway comprising lateral geniculate nucleus, V1, V2, ... At each hierarchical level, feedback pathways carry predictions of neural activity at the lower level, while feedforward pathways carry residual errors between the predictions and actual neural activity. These errors are used by the predictive estimator at each level to correct its current estimate of the input signal and generate the next prediction. The predictive estimators allow each level to maintain a Gaussian distribution over the features encoded by that level. (b) Schematic diagram of the hierarchical particle filtering model of the visual cortex. The $x_i^{(j)}$s denote 'particles' or sampled values for the features encoded by visual cortical area i. Each particle has a weight $w_i^{(j)}$ such that the weighted sum of all particles is a discrete approximation to the full posterior distribution over the feature x_i. B_1 and B_2 denote bottom-up and top-down 'beliefs' that reflect probabilistic information from lower and higher areas, respectively. The ϕs denote feedforward or feedback synaptic weights between cortical areas chosen to facilitate hierarchical inference. * denotes multiplication. (a) Adapted from Rao RPN and Ballard DH (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience* 2(1): 79–87. (b) Adapted from Lee TS and Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A* 20(7): 1434–1448.

as an approximate neural implementation of the Kalman filter, a well-known algorithm for state estimation used in engineering and control theory. A related hierarchical model of the visual cortex based on an approximate inference method known as particle filtering was suggested by Lee and Mumford. The particle-filtering-algorithm can be viewed as a generalization of the Kalman filter to nonlinear dynamics and non-Gaussian distributions. Inference involves message passing between different levels of the network in a manner similar to belief propagation but using sampled values for the random variables (Figure 5(b)). The actual neural implementation of particle filtering was left open. The Lee–Mumford model shares similarities with log probability models

in that both utilize sampled representations of probability distributions over continuous spaces.

A hierarchical cortical model based on belief propagation, with a possible neural implementation, was suggested by Rao. In this model, lower-level neurons represent specific image features (such as edges) at specific locations while higher-level neurons factor these features into more abstract attributes such as locations (irrespective of features) and features (irrespective of locations; Figure 6). Noisy leaky integrate-and-fire neurons are used to implement the belief propagation equations for a hierarchical network, under the assumption that the dendrites of neurons approximate a logarithmic transformation of their inputs. Under these conditions, the spiking probability of a neuron becomes proportional to the posterior probability of the feature (or location) being encoded by the neuron. The probabilistic integration of top-down prior information with bottom-up image likelihood information, as prescribed by Bayesian inference, is shown to account for several attentional effects in the visual cortex such as multiplicative modulation of tuning curves, restoration of neural responses in the presence of distracting stimuli, and influence of attention on neighboring unattended locations. A related model, proposed by Yu and Dayan, is based on a five-layer neural architecture and a log probability encoding scheme. This model explains both reaction times in a spatial attention task and multiplicative response modulation in the visual cortex.

Conclusion

Bayesian models have been used to explain a surprisingly wide range of perceptual phenomena in recent years. These models demonstrate how particular choices for the prior distribution of stimuli yield posterior distributions that can help explain a variety of psychophysical data. Most of these models are, however, formulated at an abstract mathematical level, leaving open the question of their neural implementation. Computational neuroscientists are beginning to investigate ways in which cortical circuits may implement Bayesian inference.

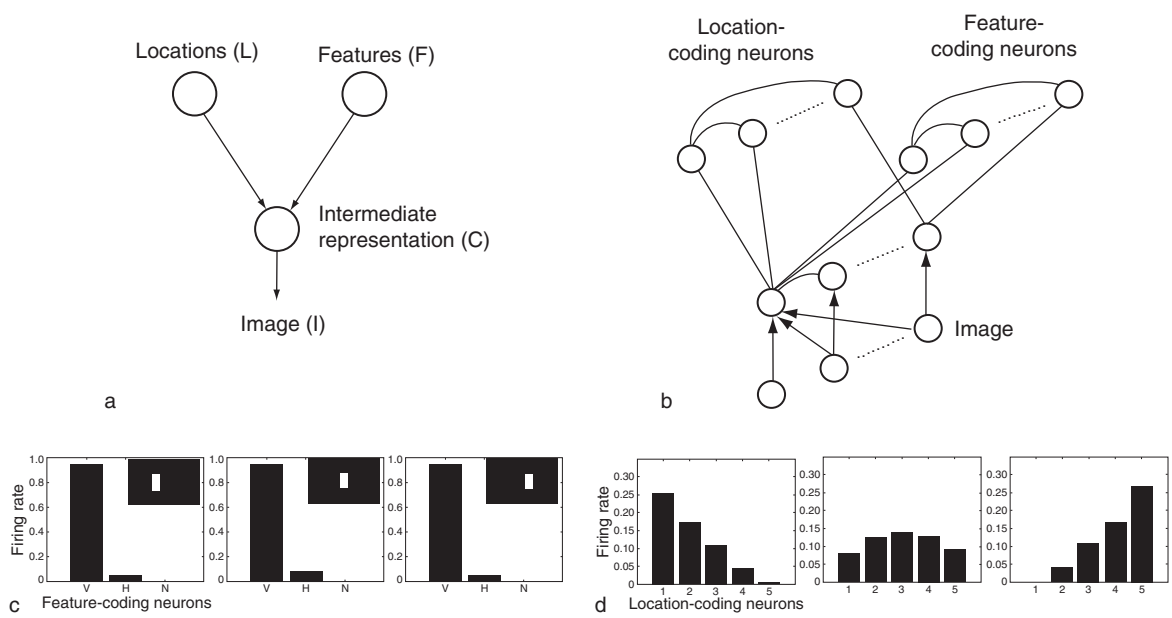

Figure 6 Hierarchical Bayesian cortical model of visual cortical areas. (a) Probabilistic generative model that factors images into spatial locations and features. (b) Three-level network for implementing Bayesian inference for the generative model in (a). Circles represent neurons, arrows represent feedforward connections, and lines represent bidirectional connections. The intermediate representation neurons are assumed to correspond to visual cortical areas V1/V2. The location-invariant feature-coding neurons (F) are assumed to reside in V4 while the feature-invariant location-coding neurons (L) are assumed to reside in areas of the parietal cortex. The location-coding neurons compute the posterior probability P(L|I) while the feature-coding neurons compute the posterior probability P(F|I). (c) Each bar graph represents the responses of three feature-coding model neurons, the first coding for vertical features (V), the second for horizontal features (H), and the third for the null (blank) stimulus (N), when the input image I contains a vertical bar flashed at three different locations (dark inset images). Note that the vertical feature-coding neuron's firing rate, denoting the posterior probability of the vertical feature, remains high and invariant to bar location. On the other hand, as shown in (d), information about the vertical bar's current location is captured by the location-coding neurons. From Rao RPN (2005) Bayesian inference and attentional modulation in the visual cortex. *NeuroReport* 16(16): 1843–1848.

This article reviewed three classes of Bayesian cortical models: log probability models, population coding models, and hierarchical models. Log probability models typically assume that a given cortical neuron encodes a particular preferred stimulus, the responses of the neuron reflecting the posterior probability of this stimulus (or, alternatively, the log posterior or log posterior ratio between the presence and absence of the stimulus). Population coding models assume that probability distributions are implicitly encoded in the responses of a population of neurons, and these models rely on a decoding function to recover the probability distribution from a given set of responses. Population coding models are typically more efficient in representing distributions than are log probability models: for example, the basis function model can represent distributions over continuous variables efficiently whereas the log probability approach can only represent such distributions as sampled distributions using a set of discrete values (preferred stimuli of the neurons). On the other hand, because they represent probabilities directly rather than implicitly, log probability schemes lend themselves more easily to implementing sophisticated Bayesian inference algorithms such as belief propagation. Finally, motivated by the hierarchical structure of the cortex, numerous models have been suggested for implementing hierarchical Bayesian inference. Whether such models can provide a unified account of the neurobiological properties observed across different cortical areas remains an interesting open question. Also intriguing is the question of whether known mechanisms of synaptic plasticity, such as Hebbian and spike timing-dependent plasticity, facilitate Bayesian inference in cortical networks.

See also: Attention: Models; Attentional Networks; Bayesian Models of Motor Control; Executive Function and Higher-Order Cognition: Computational Models; Memory: Computational Models; Prediction Errors in Neural Processing: Imaging in Humans; Psychophysics of Attention.

Further Reading

Anderson CH and Van Essen DC (1994) Neurobiological computational systems. In: Zurada JM, Marks RJ II, and Robinson CJ (eds.) *Computational Intelligence: Imitating Life*, pp. 213–222. New York: IEEE Press.

Barlow HB (1969) Pattern recognition and the responses of sensory neurons. *Annals of the New York Academy of Sciences* 156(2): 872–881.

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions* 53: 370–418.

Dayan P, Hinton G, Neal R, and Zemel R (1995) The Helmholtz machine. *Neural Computation* 7: 889–904.

Deneve S (2005) Bayesian inference in spiking neurons. In: Saul LK, Weiss Y, and Bottou L (eds.) *Advances in Neural Information Processing Systems 17*, pp. 353–360. Cambridge, MA: MIT Press.

Doya K, Ishii S, Pouget A, and Rao RPN (2006) *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT Press.

Gold JI and Shadlen MN (2001) Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences* 5(1): 10–16.

Hinton GE and Brown AD (2002) Learning to use spike timing in a restricted Boltzmann machine. In: Rao RPN, Olshausen BA, and Lewicki MS (eds.) *Probabilistic Models of the Brain: Perception and Neural Function*, pp. 285–296. Cambridge, MA: MIT Press.

Jazayeri M and Movshon JA (2006) Optimal representation of sensory information by neural populations. *Nature Neuroscience* 9: 690–696.

Knill DC and Richards W (1996) *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press.

Lee TS and Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A* 20(7): 1434–1448.

Ma WJ, Beck JM, Latham PE, and Pouget A (2006) Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9: 1432–1438.

Rao RPN (2004) Bayesian computation in recurrent neural circuits. *Neural Computation* 16(1): 1–38.

Rao RPN (2005) Bayesian inference and attentional modulation in the visual cortex. *NeuroReport* 16(16): 1843–1848.

Rao RPN and Ballard DH (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience* 2(1): 79–87.

Rao RPN, Olshausen BA, and Lewicki MS (2002) *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press.

Sahani M and Dayan P (2003) Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Computation* 15: 2255–2279.

Yu A and Dayan P (2005) Inference, attention, and decision in a Bayesian neural architecture. In: Saul LK, Weiss Y, and Bottou L (eds.) *Advances in Neural Information Processing Systems 17*, pp. 1577–1584. Cambridge, MA: MIT Press.

Zemel RS, Dayan P, and Pouget A (1998) Probabilistic interpretation of population codes. *Neural Computation* 10(2): 403–430.

Zemel RS, Huys QJM, Natarajan R, and Dayan P (2005) Probabilistic computation in spiking populations. In: Saul LK, Weiss Y, and Bottou L (eds.) *Advances in Neural Information Processing Systems 17*, pp. 1609–1616. Cambridge, MA: MIT Press.

Bayesian Models of Motor Control

M Berniker and K Kording, Northwestern University Institute of Neuroscience, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Motor control is fundamental to the nervous system: only through our movements do we interact with the world. Many models of motor control assume that of all the ways we could move, we typically choose one of the best. This implies that we can have a 'notion' of how bad one movement is in comparison to other possible movements. Within the framework of decision theory, this notion is formalized by a cost function that assigns a numerical value to every possible movement. The right movement decision seems simple: choose the action a^* that has the lowest associated cost. This decision process can be compactly described by the equation,

$$a^* = \arg\min_a (\text{cost}(\text{a})) \qquad [1]$$

For example, several studies have proposed that people move their hand from one position to another in a fashion that is as 'smooth' as possible. Thus, the cost function penalizes movements that are not smooth. It was found that many properties of human reaching movements, such as the velocity profiles, can be well predicted by such cost functions.

This framework assumes that we can deterministically choose a movement and be sure of its outcome. Yet, most motor outcomes are affected by uncertainty. To illustrate, if we aim at the center of a dartboard, we are not certain to hit the bullseye with the dart (even if we are a seasoned champion). If we continue to aim at the center and throw many times, we will end up with a distribution of dart positions. This distribution, $p(x|a)$, characterizes the likelihood of each outcome, the dart position x, given that we chose action a, the aiming point. This affects how we should choose the best action, the one that minimizes our likely cost:

$$a^* = \arg\min_a \left(\sum_{\substack{\text{possible}\\ \text{outcomes}}} \text{cost}\,(\text{outcome})\; p(\text{outcome}|a) \right) \qquad [2]$$

In other words, we should choose the action that minimizes our expected cost. In the example of playing darts, the best aiming point is a point where we will receive high scores even if we make large mistakes. Indeed, both amateur and world-class players are known to adopt a strategy that is well predicted by this approach.

While playing darts, the errors largely stem from motor noise (Figure 1(a)). However, there are many sensory sources of uncertainty as well. For example, our visual system is noisy and the location of the dartboard relative to our body is thus uncertain (Figure 1(a)). Moreover, our proprioceptive system is noisy as well; the exact configuration of our body as we release the dart is uncertain (Figure 1(a)). These sources of uncertainty combine to produce uncertainty in the movement outcome given our movement decision.

The difficulty in treating these movement decisions is that the probability of an outcome given our action decision ($p(x|a)$) is generally difficult to compute. The problem is that computing this distribution requires the specification of information that is generally uncertain. This includes information about our body as well as the external world. Bayesian inference specifies how such pieces of information should be combined. It thus predicts how pieces of information should be integrated to allow for motor control that minimizes our cost.

Bayesian Integration

Combining uncertain information to produce a coherent and accurate estimate of the world is a central problem faced in motor control. A number of psychophysics studies have analyzed how people integrate multiple sources of uncertain information to make sensorimotor decisions.

Combining Prior Knowledge with New Evidence

As an example of how we integrate uncertain information to estimate the world, consider the game of tennis. When playing tennis, it is helpful to estimate where the ball will land. Vision does not provide perfect information about the ball's trajectory or velocity, and we are thus uncertain where the ball will land. The visual system, although noisy, still provides us with an estimate, or likelihood, of where the ball will strike the court. This likelihood is the probability of having a particular sensory input for each possible location the ball may land (Figure 1(b)). This knowledge may be combined with information obtained from experience: the locations where the ball may land are not uniformly distributed over the court. The locations are usually concentrated within the confines of the court and may be highly peaked near the baselines where it is most difficult to return the ball. This distribution of positions is called the prior (Figure 1(b)). Bayes' rule defines how to combine the prior and the likelihood to make an optimal estimate of the location where the ball lands (Figure 1(b)).

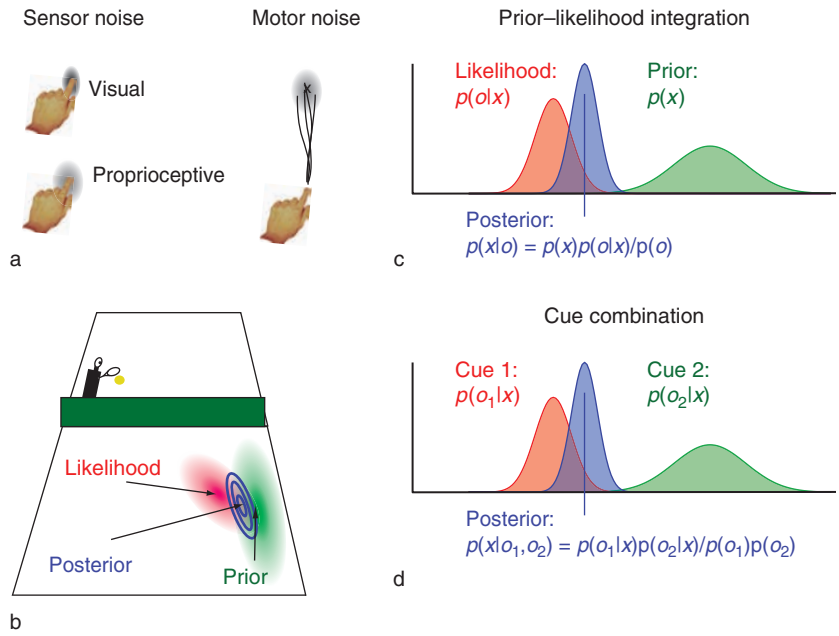

Figure 1 (a) Possible sources of noise in the motor system. Visual and proprioceptive senses contain inherent uncertainties. Motor commands also possess noise and result in uncertain movements. (b) Bayesian integration in tennis: we view a ball traveling toward the court floor and estimate the location where it will land, a likelihood (red). Experience from many such instances tells us that the ball usually lands close to the baseline, a prior (green). From these two pieces of information, we produce an updated estimate of where we believe the ball will land, a posterior (blue ellipses). (c) Bayesian integration of a prior and likelihood. The prior, denoted with the green curve, represents the probability of a state, *x*. The likelihood, denoted with the red curve, represents the probability of observing the data, *o*, given *x*. The posterior is the probability that *x* is the state, given our observation, *o*. (d) Bayes' rule applied to cue combinations.

Bayes' rule states that the probability of the ball landing at position x given our observation o is the product of the prior probability of the ball at x, with the likelihood of the observation o given the position x, divided by the probability of the observation:

$$p(x|o) = p(x)p(o|x)/p(o) \qquad [3]$$

The distribution produced by eqn [3] is shown graphically Figure 1(c). We can also interpret Bayes' rule as the 'optimal' means of combining a prior and a likelihood as it produces an estimate with the minimum uncertainty. A few studies have shown that when people combine prior and new information, their behavior is close to that predicted by Bayes' rule. Peoples' ability to integrate priors and likelihoods in a manner that is close to the optimal prediction by Bayes' rule has been shown in many modalities.

A typical study will let people use a motor task to indicate their estimate of the location of a target. In every trial, this location will be drawn from a Gaussian probability distribution (the prior). One group of subjects experience a narrow target distribution and thus have a narrow prior. A different group of subjects will experience a wide target distribution, thus having a wide prior. Bayesian statistics predicts how subjects should combine the likelihood with the prior. These predictions are then compared to human performance.

Such paradigms have been applied to numerous estimation problems. Sensorimotor integration, force estimation, timing estimations, speed estimations, the interpretation of visual scenes, and even cognitive estimates are examples. These studies jointly demonstrate that people are very good at combining prior knowledge with new evidence in a way predicted by Bayesian statistics.

Cue Combination

In many cases, it is not prior knowledge that is combined with a likelihood but, rather, two different cues that are combined into a joint estimate. For example, we may see and feel an object at the same time. We can then use what we saw and what we felt to infer the properties of the object. Bayesian statistics allows us to also solve these problems and the posterior is simply the product of the likelihoods (Figure 1(d)). Some studies have found that people are also optimal at this combination.

As a typical study, we consider an analysis of how people combine visual and auditory information to estimate the position of a target. In such a study, the precision of visual and auditory perception is

measured for each subject (unimodal). Subsequently, the precision and accuracy of perception is measured when subjects can use both modalities (bimodal). The performance in the bimodal trials can be predicted from the unimodal trials using Bayesian statistics.

As with studies investigating prior–likelihood integration, such a paradigm has been applied to numerous cue combination problems. It has been applied to the combination of visual and auditory information as well as tactile and visual information. It has also been applied to several cases in which two pieces of knowledge from the same modality are combined. For example, texture and stereo disparity are both cues for estimating the depth of a visual object. In all tested cases, cues are combined by the subjects in a manner that is close to the optimum prescribed by Bayesian statistics.

Bayesian Integration over Time

The state of the world and our information about it are continually changing. We thus constantly need to integrate this new information with our current beliefs to inform our estimates of the world. This implies that Bayesian integration should take place in a continuous manner.

This approach is extensively taken in modern applications of control theory. Many of these applications are built on the idea of Kalman filtering, which is a method for using an error in our prediction of the world to continuously update our beliefs (Figure 2(a)). This technique, although not usually described as such, is equivalent to a sophisticated application of Bayesian statistics. At every time step, a Kalman filter combines its estimate of the world's state (the prior) with a measured observation (the likelihood) to update the prediction of the world's current state, represented by a posterior (Figure 2(b)). This formalism is well developed and used in a wide range of applications from aeronautics to humanoid robotics. Indeed, even the motor control problems of two applications as disparate as controlling a jet and controlling our bodies share many computational analogies. In both cases, continuously incoming information needs to be assessed to move precisely (albeit on different timescales). Only recently has this methodology of Kalman filtering been applied to make quantitative predictions of human movement behavior.

A number of studies have examined how subjects integrate information in a continuous fashion. These studies have compared the way in which subjects move with the predictions of optimal Bayesian models. They have found close agreement between the predictions of the optimal theory and the way movements are actually executed.

As an example of this problem, we consider a recent theory of motor adaptation. The properties of our muscles change continuously throughout our lives. Over several timescales, muscles change – fatigue over the course of hours, disease over the course of weeks, and aging over many years. This means that we must estimate the strength of our muscles if we are to move precisely. The error in our estimate of the strength of our muscles will translate into a movement error. We can use this error to obtain a likelihood function characterizing the strength of the muscles. According to Bayesian statistics, we need to combine the newly obtained information, our motor error, with our prior belief. We can thus infer a new and improved estimate (Figure 2(c)). Often, adaptation is probed using a target jump paradigm. In these paradigms, an experimental manipulation is used to change subjects' perception of the strength of their limb. In these circumstances, Bayesian statistics predicts that

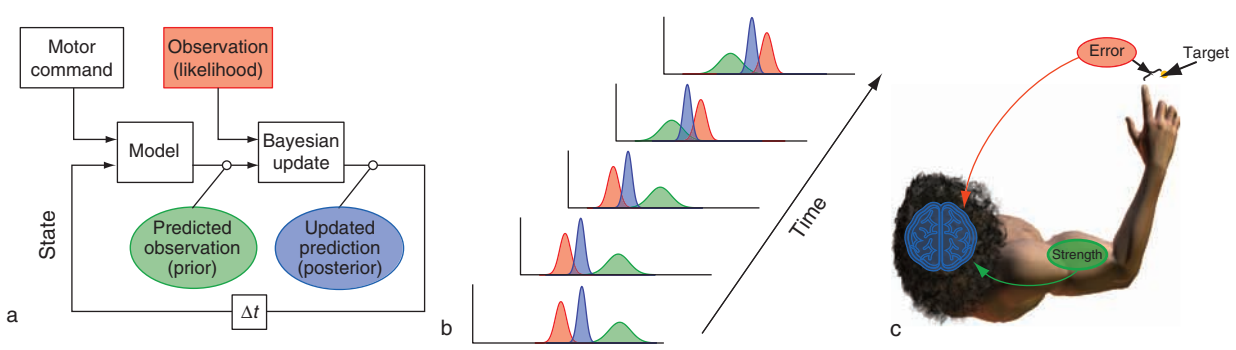

Figure 2 (a) Typical procedure for optimally estimating the world's state in modern control theory. A model of the world, combined with a motor command, is used to estimate a predicted state, or prior distribution (green). Observations of the world dictate the likelihood of a particular observation given the current world state (red). A Kalman filter is used to make a Bayesian update of our belief in the world's current state (blue). (b) This process of Bayesian inference repeats itself at each time step. (c) Motor adaptation can be framed as an analogous update procedure. Our prior belief of muscle strength (green) is integrated with our observed motor error (red) to update a belief of our muscle's strength.

subjects should adapt over time. The way subjects adapt in this setting is then compared to the adaptation profile predicted by optimal Bayesian adaptation.

It has been found that motor adaptation over time can be well understood from the predictions of Kalman filtering. It has also been found that the way people estimate the position of their hand in the dark is well predicted by the same methods. As another example, when people balance a pole on their hand, they also seem to use such strategies. These studies highlight that when people integrate information over time, they seem to do so in a manner that is consistent with the optimal Bayesian solution.

Bayesian Inference of Structure

We previously discussed how people combine two cues into a joint estimate using Bayes' rule. The application of these models implicitly assumed that we are certain that both cues have a common source: if we hear a tone while seeing a flash, we assume that both of the cues originate at the same position and have a common cause. This idea has been used very successfully for describing human behavior in cue combination experiments.

Although successful as implemented thus far, these probabilistic models have obvious shortcomings. For instance, we know from experience and experimental studies that if two cues are very dissimilar our perception of unity breaks down. If we see a flash far to our right while perceiving a tone far to our left, we perceive two independent events. We infer the structure of events; we estimate if two events have a common cause or if they just randomly co-occur. This implies that if two events are perceived as having a common cause, they should influence the perception of each other. On the other hand, if they are perceived as not having a common cause, the perception of each should be independent of the other. In some circumstances, the cues can create an illusion of a common cause, the so-called ventriloquist effect.

Traditional Bayesian models of psychophysical behavior are unable to explain this range of phenomena. Indeed, experiments that tested these models used cues that were close to coincident. For example, the position of a visual stimulus was close to the position of an auditory stimulus.

Other studies have tested subject performance in situations in which two cues are dissimilar from one another. A typical study will present a visual stimulus at a random position within the visual field and simultaneously present an auditory stimulus at a different random position. Subsequently, subjects are asked where they perceive the stimuli.

These studies have found that when two stimuli are near coincident, people tend to infer a common cause and use each cue to guide estimation of the other. With increasing disparity between the cues, subjects' belief in a common cause decreases (Figure 3). Traditional Bayesian models predict that in the case of Gaussian likelihoods the estimate of the value of one cue should be linearly influenced by the value of the other cue. In sharp contrast to these predictions, experiments have found nonlinear cue interactions. These interactions are well predicted if we assume that on each trial the subjects infer a common or distinct cause for both cues.

These effects are found in a wide range of experimental situations. They have been found in position estimation using visual and auditory stimuli, in the estimation of depth from several cues, and also in number estimation using visual, tactile, and auditory stimuli. Evidence suggests that subjects integrate not only priors and likelihoods but also the causal structure underlying a set of sensory stimuli.

The statistical problem that subjects solve in cue combination implicitly involves an inference about the causal structure of the stimuli. In these studies, people are uncertain about the identity and number of relevant variables; for example, did the flash of light cause the auditory tone, did the tone cause the flash, was there an unknown cause for them both, or were they simply coincidental? The problem faced by the nervous system is thus similar to cognitive problems that occur in the context of causal induction. Many experiments demonstrate that people interpret events in terms of cause and effect. The results presented here show that sensorimotor integration exhibits some of the factors that make human cognition difficult to understand. Carefully studying and analyzing seemingly simple problems such as cue combination may provide a fascinating way of studying the human cognitive system in a quantitative manner.

Future Directions for Bayesian Inference

The Bayesian models reviewed so far do well at predicting the results from many experimental studies of motor control. Whereas in theory Bayesian statistics' applicability is rarely limited, in practice there are many cases in which this technique is hindered.

Most of the previously discussed Bayesian models assume that noise sources are Gaussian and that the interactions between the subject and the world are linear. Recent approaches in Bayesian statistics allow the application of powerful framework well beyond these simple assumptions. For instance, the structure inference problem is a simple case of non-Gaussian

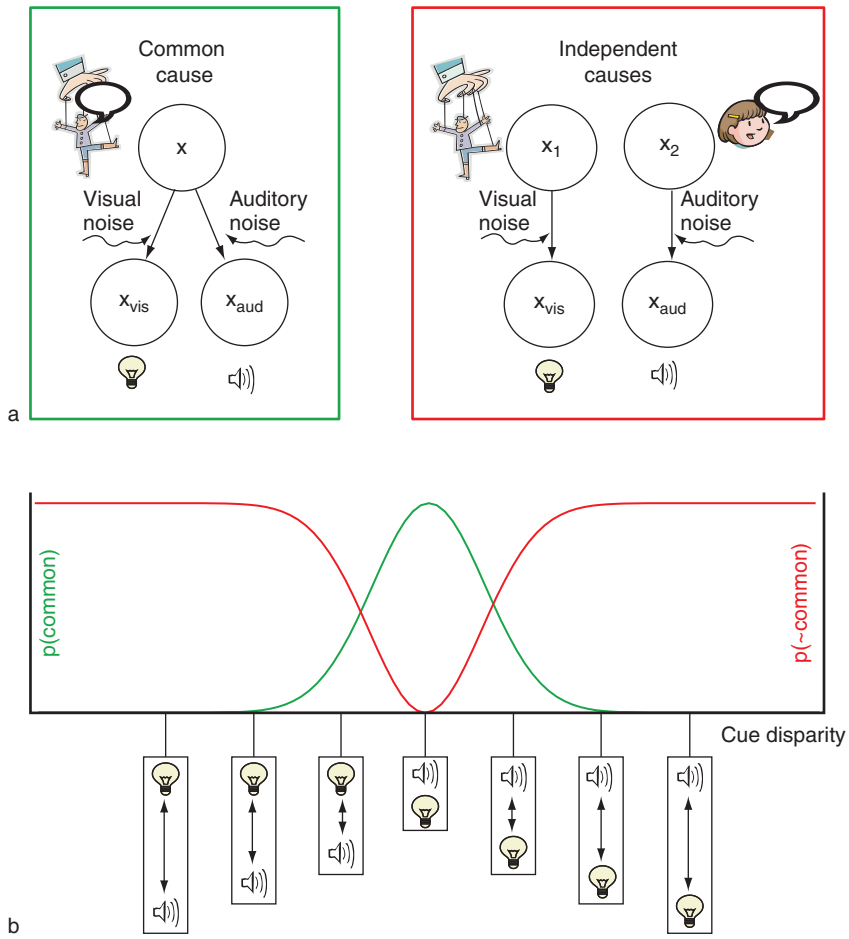

Figure 3 (a) Two structural beliefs of the world. If two cues are adequately coincident, subjects perceive them as having a common cause (green box), a phenomenon typified through ventriloquism. If the cues are disparate in time or space, subjects perceive them as having independent causes (red box). (b) Subjects' belief of common cause. Because the spatial disparity of two cues, a light flash and a tone, is experimentally controlled, the belief in a common cause can be manipulated.

probability distributions. The techniques that may be applied to such problems range from linearization techniques such as extended Kalman filters over variational techniques to particle-filtering methods. The specialized Bayesian literature deals extensively with such computational difficulties.

Bayesian inference allows us to estimate the present state of the world given all the sensory observations we have obtained from the past until now. If the problem we are faced with requires making only one decision at a single point in time (static problem), then decision theory (see eqn [2]) readily allows us to decide optimally. Yet, movement is fundamentally a sequential decision (dynamic) problem. Any motor command we produce in the present influences the situations in which we will find ourselves in the future. This implies that motor commands are not independent from each other but, rather, each command must be viewed in the context of all possible future situations it may direct us.

Optimal control theory is the systematic study of problems of this class. Often utilizing a Bayesian framework, it employs analytical and numerical techniques to solve the motor control problem. For the set of problems in which dynamics are linear, noise is Gaussian, and cost functions are quadratic, optimal control provides efficient solutions. Much research addresses ways of extending the approaches to more complicated problems. Typical applications involve the control of humanoid robots and the control of aircraft. The theory is complex and beyond the scope of this article. Advances in optimal control coupled with advances in Bayesian methods hold promise for allowing progressively more complicated movement behaviors to be analyzed and predicted.

Discussion

Much if not most of neuroscience is driven by postulates regarding the problems the nervous system is

attempting to solve. For example, researchers recording from the visual system may assume that visual features such as edges are of importance to the object recognition problem. When researchers study the motor system, they may decide that efficient movements are the intended goal. This way of thinking is called normative: a description of the problem being solved is converted into predictions of how the nervous system 'should' work. Models that predict behavior from such descriptions are especially informative and make precise quantitative predictions of experimental data.

Normative models prescribe how the nervous system should behave, as opposed to neural modeling, which describes how the nervous system behaves. To clarify the distinction, models derived from meaningful descriptions of the problem to be solved should be considered normative or prescriptive. For example, some theories of movement assume that people try to move as smoothly as possible, and these models explain a large amount of experimental data. However, there is no explicit reason why smooth movements would be better than similar but less smooth movements. Perhaps more to the point, these models do not yield insight on why subjects move the way they do. In the case of smooth movements, research demonstrated that a different cost function, clearly tied to the movement problem, could also predict the data; optimizing movements for precision naturally leads to smooth movements – a cost function that is of obvious importance.

A different way of modeling the properties of the nervous system is the so-called 'bottom-up' approach. Numerous measured or inferred properties of neurons are used to run comprehensive computer simulations of neurons within expansive networks, the results of which are meant to mimic the actual nervous system. This approach relies on the accurate knowledge of many system parameters – a formidable requirement. Furthermore, in large, complex systems such as these, small deviations in parameters can result in gross and qualitative differences from the actual system being emulated – in this case the nervous system. Even in areas where bottom-up modeling is practical, such as action potential propagation in squid giant axons, normative models may contribute to an understanding of the system. It has been conjectured that the density of sodium channels is optimally adapted to maximize the speed of action potential conductance, very possibly the result of evolutionary pressures to minimize the flight response time. This is of obvious normative reasoning because fleeing fast from approaching predators can mean the difference between life and death. Normative modeling of low-level cellular properties is possible and in close analogy with the normative 'top-down' modeling of behaviors made by decision theory.

Bayesian integration offers a computational framework for combining various sensory information, from multiple modalities and different points in time, into a joint estimate of the world. Consistent with these ideas, unimodal brain regions are known to be affected by uncertainty. For example, primary visual cortex is influenced by the statistics of a stimulus. Multimodal regions that combine information are similarly influenced by probabilities. Prefrontal, temporal, and parietal cortices are likely to exhibit similar statistical principles. Crucial in the timing of motor behaviors and motor learning, the cerebellum's function can also be explained in Bayesian terms. Interestingly, however, even cortical regions traditionally thought of as unimodal, such as primary visual and auditory cortices, have been shown to be responsive to multiple sensory modalities and receive projections from multiple sensory regions. Taken together, the functional and anatomical connectivity of the brain is consistent with the multimodal computations necessary for Bayesian inference. Considering how salient the problem of uncertainty is, numerous, if not all, brain regions can be expected to utilize Bayesian statistics to reduce uncertainty.

Bayesian statistics and normative modeling hold promise for many areas of neuroscience because they formalize the problems the nervous system needs to solve while addressing the inherent uncertainties. Normative models have few free parameters in comparison to neural models – radically simplifying their application. Moreover, normative modeling is synergistic with neural modeling; combining the two techniques will provide models that not only predict but also explain neural behavior. By quantitatively predicting behavior from these normative–neural models, we can develop a deeper understanding of the problems being solved by the nervous system. This will also allow us to ask better, more informed questions in future neuroscientific experiments.

See also: Bayesian Cortical Models; Computational Approaches to Motor Control; Motor Psychophysics; Sensorimotor Integration: Models.

Further Reading

Davidson PR and Wolpert DM (2003) Motor learning and prediction in a variable environment. *Current Opinion in Neurobiology* 13(2): 232–237.

Ernst MO (2005) A Bayesian view on multimodal cue integration. In: Knoblich G, Thornton I, Grosjean M, and Shiffrar M (eds.)

Human Body Perception from the inside Out, pp. 105–131. New York: Oxford University Press.

Gold JI and Shadlen MN (2002) Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* 36: 299–308.

Jordan MI (1998) *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Kersten D and Yuille A (2003) Bayesian models of object perception. *Current Opinion in Neurobiology* 13(2): 150–158.

Knill D and Richards W (eds.) (1996) *Perception as Bayesian Inference*. New York: Cambridge University Press.

Kording KP and Tenenbaum JB (2006) Causal inference in sensorimotor integration. *Advances in Neural Information Processing Systems* 19.

Kording KP and Wolpert DM (2006) Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* 10(7): 320–326.

MacKay DJC (2003) *Information Theory, Inference, and Learning Algorithms*. New York: Cambridge University Press.

Maloney LT, Trommershäuser J, and Landy MS (2006) Questions without words: A comparison between decision making under risk and movement planning under risk. In: Gray W (ed.) *Integrated Models of Cognitive Systems*. New York: Oxford University Press.

Niv Y, Joel D, and Dayan P (2006) A normative perspective on motivation. *Trends in Cognitive Sciences* 10(8): 375–381.

Saunders JA and Knill DC (2004) Visual feedback control of hand movements. *Journal of Neuroscience* 24(13): 3223–3234.

Scott SH (2004) Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience* 5(7): 532–546.

Shams L, Kamitani Y, and Shimojo S (2000) Illusions. What you see is what you hear. *Nature* 408(6814): 788.

Todorov E (2004) Optimality principles in sensorimotor control. *Nature Neuroscience* 7(9): 907–915.

Todorov E (2006) Optimal control theory. In: Doya K (ed.) *Bayesian Brain*. Cambridge, MA: MIT Press.

BDNF in Synaptic Plasticity and Memory

N H Woo and B Lu, National Institutes of Health, Bethesda, MD, USA

© 2009 Published by Elsevier Ltd.

Introduction

Among members of the neurotrophin family, brain-derived neurotrophic factor (BDNF) stands out for its ability to regulate synaptic plasticity and various cognitive functions of the brain. A Medline search with the terms 'BDNF' and 'synaptic' yields more than 700 research articles, mostly published in the last 7 years. Given that neurotrophins were initially defined as secretory factors that promote neuronal survival and differentiation during development, the role of BDNF in synaptic modulation was not recognized until late 1990s. A number of observations have aided in the realization that the primary function of BDNF is to regulate synaptic transmission and plasticity, rather than neuronal survival. One is that BDNF is widely distributed in many regions of the adult brain, with levels much higher than any other neurotrophins. The other is that the expression of BDNF can be rapidly enhanced by neuronal activity under conditions relevant to synaptic plasticity. Because neuronal activity is known to be crucial for synaptic plasticity, it was hypothesized that activity-dependent synaptic modulation is mediated by BDNF. Indeed, early studies demonstrated that BDNF mimics neuronal activity in altering the number and/or strength of synaptic connections. Subsequent studies revealed a much more complex and interesting picture. On the one hand, BDNF regulates various forms of synaptic plasticity, leading to changes in neuronal circuitry subserving complex behaviors. On the other hand, many aspects of BDNF cell biology, such as transcription and secretion, are tightly controlled by neuronal activity. Complex interactions between BDNF and neuronal activity may offer a plethora of means to control sophisticated cognitive functions of the mammalian brain.

Cell Biology of BDNF

All neurotrophins arise from their precursors as a result of proteolytic cleavage of the prodomain. Proneurotrophins have long been thought to be inactive precursors. However, this view was challenged a few years ago when proneurotrophins were shown to promote apoptosis via p75 neurotrophin receptor (p75NTR). This is opposite to the cell survival effect by mature neurotrophins, which act via their preferred tropomyocin-related receptor tyrosine kinase (Trk) receptors. Similarly, proBDNF and mature BDNF (mBDNF) have been shown to elicit opposite effects on synaptic plasticity (Figure 1). In recognition that proneurotrophins are biologically active, cleavage of proneurotrophins becomes an important regulatory mechanism that controls the direction of neurotrophin regulation.

Signal Transduction

The biological functions of BDNF are mediated by two receptor systems: TrkB and p75NTR. It is well established that mBDNF binds TrkB with high affinity. Upon binding, BDNF triggers TrkB dimerization resulting in tyrosine phosphorylation in its cytoplasmic domain. These autophosphorylation events recruit a series of intracellular proteins that primes subsequent activation of several signaling pathways. Three classical signaling pathways have been identified: phosphatidylinositol-3-kinase (PI3K) pathway, phospholipase C-γ (PLC-γ) pathway, and Map-Erk Kinase (MEK)-MAPK pathway (Figure 1). The majority of BDNF actions described thus far are attributed to signaling cascades activated by TrkB. In addition to cell surface signaling, BDNF induces the endocytosis of TrkB. Rather than simply inactivating TrkB, the endocytosis of BDNF-TrkB complex results in the formation of BDNF-TrkB signaling endosomes, triggering signaling cascades different from those initiated from cell surface TrkB. This process is required for translation-dependent long-term functions, and is involved in retrograde propagation of BDNF signal from axonal terminals to cell body.

All neurotrophins including proBDNF bind p75NTR, which triggers signaling events distinct from those by Trk receptors. The cytoplasmic domain of p75NTR lacks intrinsic catalytic activity. Upon ligand binding, several intracellular signal transduction cascades are activated, including nuclear factor kappa B (NF-κB), Jun kinase, and sphingomyelin hydrolysis (Figure 1). Notably, p75NTR activation is associated with the initiation of apoptosis. For many years, p75NTR was considered a 'low affinity' neurotrophin receptor. Recent studies indicate that preferred ligands for p75NTR are proneurotrophins, with binding affinities just as high as that between mature neurotrophins and Trk receptors. Current data support a model that pro- and mature neurotrophins induce very different functions through two distinct receptor-signaling systems. An added complexity is the newly discovered coreceptor sortilin. The prodomain of proneurotrophins bind sortilin, whereas the mature domain binds p75NTR. The formation of

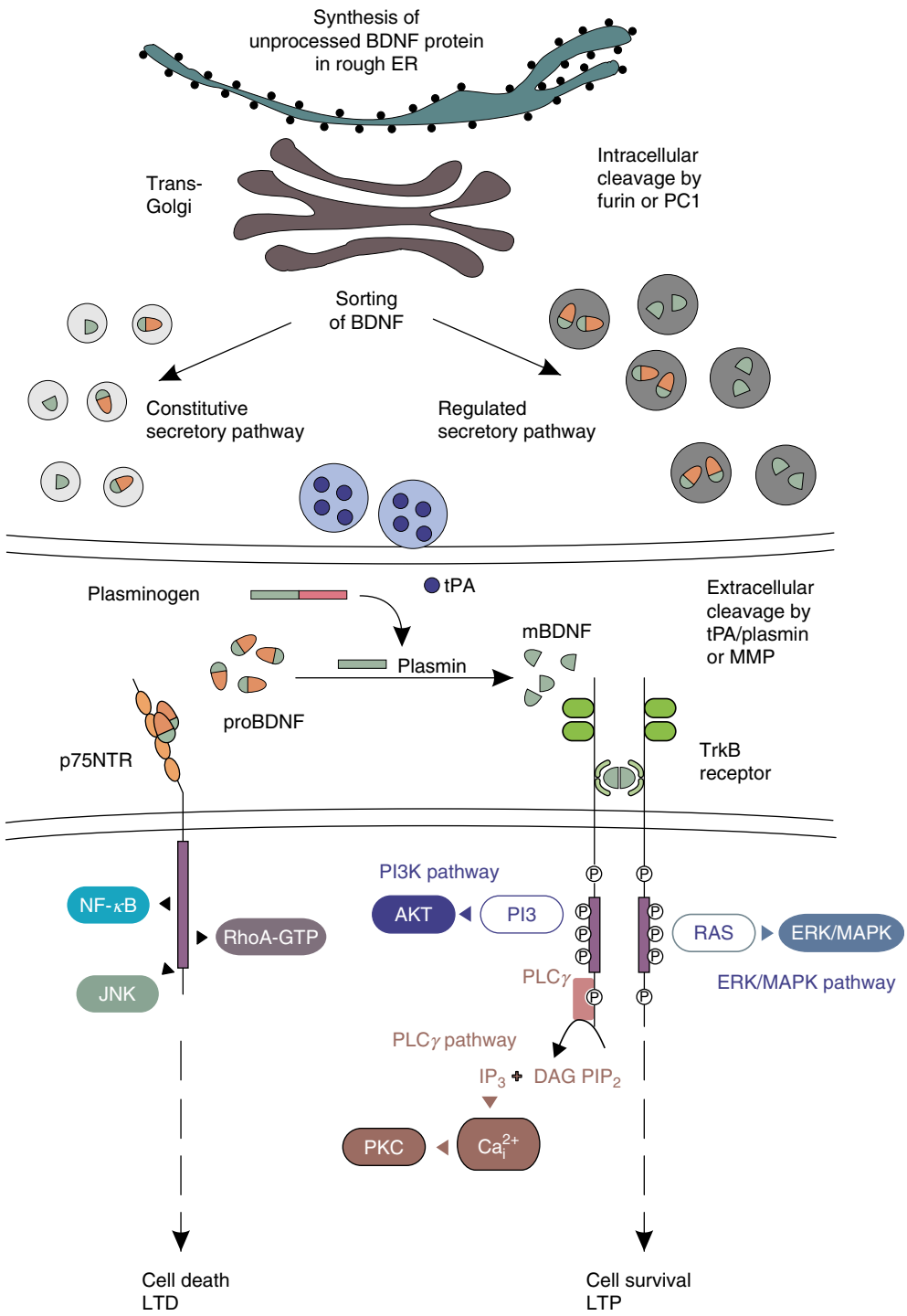

Figure 1 Synthesis, trafficking, and receptor-signaling of BDNF. Initially synthesized in the endoplasmic reticulum (ER) as a precursor protein, proBDNF is properly folded in the ER and Golgi network and packaged into secretory vesicles. Subsequently, BDNF is sorted into either the constitutive or regulated secretory pathway, and transported to the appropriate site of release. The prodomain can be cleaved intracellularly by furin or protein convertases, resulting in the secretion of mature BDNF (mBDNF). Alternatively, proBDNF can be secreted and cleaved extracellularly by the tPA/plasmin cascade or metalloproteinases to yield mBDNF. Once secreted, proBDNF and mBDNF elicit diverse and often opposing biological actions via two distinct receptor-signaling systems. mBDNF binds TrkB, leading to the autophosphorylation of tyrosine residues in the tyrosine kinase domain. Consequently, three major signaling cascades can be activated by mBDNF-TrkB, including the PI3K pathway, ERK/MAPK pathway, and PLCγ pathway. In contrast, proBDNF binds p75NTR, resulting in the activation of several signaling molecules, including NF-κB, JNK and RhoA.

the sortilin–proneurotrophin–p75NTR triplex may be necessary for p75NTR signaling.

Activity-Dependent Controls

A cardinal feature of BDNF is that its expression is regulated by neuronal activity. It is now recognized that a multitude of physiological stimuli can alter BDNF expression. For example, visual input and sensory stimulation of the whiskers control BDNF expression in the visual cortex and barrel cortex, respectively. In the superchiasmatic nucleus and amygdala, expression of BDNF is regulated by circadian rhythm and fear emotion. Remarkably, learning or exercise can also enhance BDNF expression in the hippocampus. Moreover, BDNF levels are also affected in a variety of pathological conditions associated with altered neuronal activity in the brain including seizure, Alzheimer's, depression, and stress. In addition to regulation of BDNF gene expression, a new theme emerging from recent studies is that neuronal activity also controls many cellular processes of BDNF, including intracellular trafficking, secretion of BDNF, and perhaps cleavage of proBDNF.

Transcription The genomic structure of BDNF is quite complex. In rats, there are at least four promoters controlling four short 5′ exons. Each 5′ exon is alternatively spliced onto a common 3′ exon (exon V) encoding the pre-proBDNF protein. In humans, the latest study reported seven promoters and eight exons, with exon VIII being the common exon coding for pre-proBDNF. It has long been a puzzle why nature has designed multiple BDNF transcripts that encode exactly the same protein. Cumulative evidence now indicates that these transcripts are distributed in different brain regions, different cell types, and even different parts of the cells (soma vs. dendrites). Importantly, their expression can be altered in response to different physiological stimuli. For instance, exon III transcript is detected only in cell bodies, whereas exon IV transcript is present both in cell bodies and dendritic processes of neurons in the visual cortex. During cerebellar development, thyroid hormone treatment selectively primarily enhances the expression of exon II mRNA.

Emerging evidence indicates that BDNF promoters are differentially involved in various neurological and psychiatric disorders. Promoter II-drive transcription can be suppressed by a neuronal silencer, and such suppression is removed by huntingtin, which binds and sequesters the silencer in the cytosol of cortical neurons. This is important for the survival of cortical neurons that projects to the striatum. In Huntington's disease, the mutant huntingtin can no longer bind the silencer, resulting in the translocation of the silencer into the nucleus and suppression of BDNF promoter II. Another striking example of BDNF promoter specific regulation involves MeCP2, which is a methyl-CpG-dependent transcriptional repressor that binds methylated DNA BDNF promoter III. Neuronal depolarization dissociates MeCP2 from promoter III, leading to the expression of exon III transcript in hippocampal neurons. Mutation in MeCP2, which occurs in 80% of Rett syndrome patients, abolishes this activity-dependent form of regulation. Promoter IV has been implicated in stress and is the major target of glucocortical and mineralocortical receptors. In clinical studies, a human single nucleotide polymorphism (SNP) in the exon IV has been associated with epilepsy and late-onset Alzheimer's.

Among all the promoters, promoter III has received much attention because it is by far the most effectively regulated by neuronal activity in the amygdala, hippocampus, and cortex. An increase in promoter III-driven transcription has been associated with long-term potentiation (LTP) and memory. Early work showed that BDNF gene expression was dependent on a rise in intracellular calcium and that application of high K^+ to cultured cortical neurons selectively enhanced exon III expression. Based on these observations, three elements in promoter III were characterized to be involved in Ca^{2+}-dependent expression of BDNF: the Ca^{2+} responsive sequence 1 (CaRE1) that binds Ca^{2+} responsive transcription factor (CaRF), the E-Box that binds upstream stimulatory factor (USF), and the classic cAMP responsive element (CRE) that binds cAMP responsive element binding (CREB). In addition, the transcription through promoter III is regulated by NF-κB and MeCP2. Taken together, transcription of BDNF exon III is tightly regulated by several mechanisms that couple neuronal activity with gene transcription.

Processing and trafficking Like all neurotrophins, BDNF mRNA is translated into a precursor protein, pre-proBDNF, which enters into the endoplasmic reticulum (ER) lumen through its N-terminal 'pre' sequence (signal peptide). After the removal of the pre-sequence by signal peptidases in the rough ER, the protein is folded in the trans-Golgi network and then packaged into secretory vesicles. Once folded correctly, BDNF is sorted into one of two principal pathways, the constitutive (i.e., spontaneous release) or regulated (i.e., release in response to stimuli) secretory pathway (Figure 2). The BDNF-containing vesicles are trafficked to the appropriate subcellular compartment. In neuronal dendrites and spines, BDNF appears to be stored in a special type of

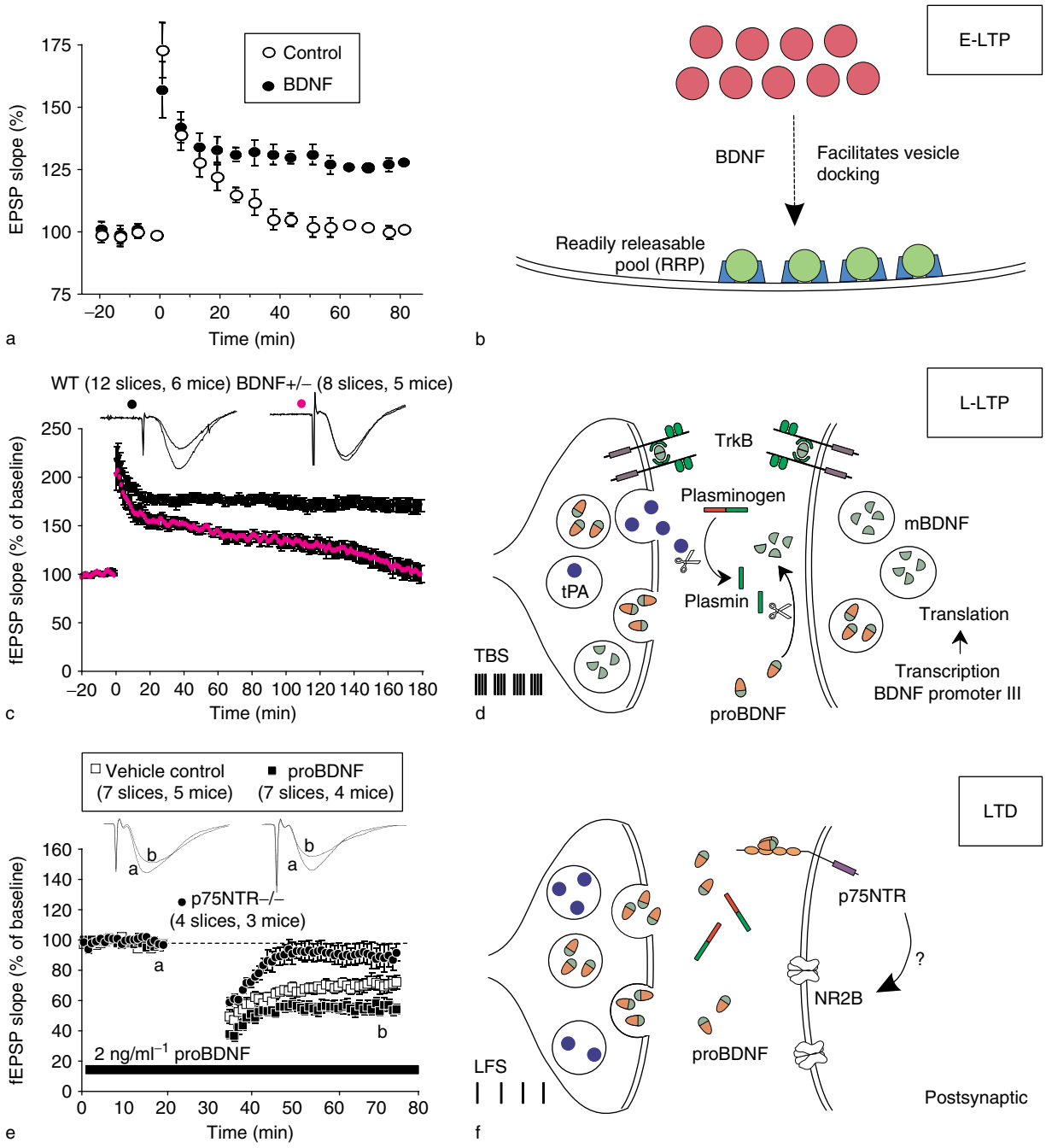

Figure 2 Regulation of distinct forms of hippocampal synaptic plasticity by BDNF. (a) mBDNF facilitates E-LTP in neonatal hippocampus in which BDNF level is low. Application of tetanic stimulation to neonatal slices (p12-p13) induces only short-term potentiation (STP), which can be converted to E-LTP by exposure to exogenous BDNF. (b) mBDNF facilitates E-LTP by promoting vesicle docking. (c) mBDNF is also involved in L-LTP. Strong theta-burst stimulation (TBS) induces robust L-LTP in wild-type (WT), but not in BDNF +/− mice hippocampal slices. (d) ProBDNF → mBDNF conversion by tPA/plasmin is required for L-LTP. Strong TBS triggers the secretion of tPA, which cleaves plasminogen to form plasmin. Plasmin subsequently cleaves proBDNF to yield mBDNF, which binds TrkB and permits L-LTP expression. (e) proBDNF promotes hippocampal LTD. Slices from p75NTR−/− mice fail to exhibit NMDA receptor-dependent LTD, whereas treatment with cleavage-resistant proBDNF enhances LTD in wild-type slice. (f) proBDNF binds to p75NTR to facilitate LTD, possibly through the regulation of NR2B, a distinct NMDA receptor subunit implicated in hippocampal LTD. EPSP, excitatory postsynaptic potential; fEPSP, field excitatory postsynaptic potential; LFS, low-frequency stimulation. (a) Reproduced from Figurov A, Pozzo-Miller L, Olafsson P, Wang T, and Lu B (1996) Regulation of synaptic responses to high-frequency stimulation and LTP by neurotrophins in the hippocampus. *Nature* 381: 706–709, with permission. (c) Reprinted with permission from Pang PT, Teng HK, Zaitsev E, et al. (2004) Cleavage of proBDNF by tPA/plasmin is essential for long-term hippocampal plasticity. *Science* 306: 487–491. Copyright 2004 AAAS. (e) Reproduced from NH Woo, Teng HK, Ciao C, et al. (2005) Activation of p75NTR by proBNF facilitates hippocampal long-term depression. *Nature Neuroscience* 8: 1069–1077, with permission.

secretory granules that lack chromogranin A (CGA), a marker for large dense core vesicles (LDCV). In contrast, conventional BDNF-containing LDCVs have been found in axons and terminals, possibly through anterograde axonal transport. Nonneuronal cells such as fibroblasts and Schwann cell secrete neurotrophins constitutively, whereas principal neurons and neuroendocrine cells secrete neurotrophins in response to depolarization and a rise in intracellular calcium.

The dendritic trafficking and synaptic localization of BDNF appear to be influenced by its own prodomain. This was implicated in a study that examined a SNP in the pro-region of the human BDNF gene. This SNP, located at nucleotide 196, produces a valine-to-methionine substitution at amino acid 66 (Val66Met). In cultured hippocampal neurons, fluorescence-tagged val-BDNF is distributed in cell body, as well as dendrites. A fraction of val-BDNF is also located at synapses, as revealed by their co-localization with synaptic markers. In marked contrast, met-BDNF is largely located in cell body and proximal dendrites. Met-BDNF was rarely localized at distal dendrites and was absent at synapses. These results suggest that the prodomain, particularly the region containing Val66, is critical for dendritic trafficking and synaptic localization of BDNF.

A long-held view is that proneurotrophins, particularly proNGF and proBDNF, are processed by intracellular proteases including the serine protease furin in the trans-Golgi network and the prohormone convertases (PC1/3) in the secretory granules. Recent studies have shown that a large fraction of BDNF in the brain is secreted in the proform, which is converted to mBDNF by extracellular proteases including plasmin or metalloproteinases (MMP3 or MMP7). Because proBDNF and mBDNF elicit distinct and often opposing biological actions through different receptors, proteolytic cleavage has now emerged as a new mechanism that determines the function of BDNF. Of particular interest is tissue plasminogen activator (tPA), an extracellular protease that converts the inactive zymogen plasminogen to plasmin. tPA is secreted from axonal terminals in response to neuronal activity. It is conceivable that neuronal activity could control proBDNF → mBDNF conversion by triggering tPA secretion.

Secretion BDNF is perhaps the only neurotrophin indisputably secreted in response to neuronal activity. In fact, majority of BDNF is sorted into the regulated, rather than the constitutive, secretory pathway. Experiments using green fluorescent protein (GFP)-tagged BDNF revealed that BDNF can be secreted from either pre- or postsynaptic sites. The amount of BDNF secretion depends on the pattern of neuronal activity. Generally, tetanus such as those used to induce LTP is more effective in inducing BDNF secretion than low-frequency stimulation. This has been demonstrated in cultured neurons as well as in slices that underwent different forms of plasticity.

Studies of Val66Met SNP have drawn attention to the role of prodomain in activity-dependent BDNF secretion. In neurons transfected with met-BDNF, depolarization-induced secretion was selectively impaired while constitutive secretion remained normal. Subsequent studies demonstrate that proBDNF is co-localized with the neurotrophin coreceptor sortilin intracellularly in secretory granules, and that sortilin interacts specifically with the prodomain in a region encompassing Val66Met. Remarkably, inhibition of the interaction between the prodomain and intracellular sortilin attenuates secretion of BDNF induced by depolarization, suggesting that this interaction is critical for regulated secretion. However, sortilin could interact with the prodomain of other neurotrophins incapable of regulated secretion, making it less likely to be a specific mechanism for sorting. On the other hand, a sorting motif was recently identified in the mature domain of BDNF, but not nerve growth factor (NGF), that interacts with a well-known sorting receptor, carboxypeptidase E (CPE). Such an interaction was deemed essential for sorting proBDNF into regulated pathway vesicles for activity-dependent secretion. Given that the prodomain promotes proper folding of neurotrophins, it is conceivable that interaction between the prodomain and sortilin may hold proBDNF in a correct configuration, exposing the mature domain to the sorting receptor CPE, which sorts proBDNF into the regulated secretory pathway.

Roles of BDNF in Synaptic Plasticity

The ability of the mammalian brain to adapt or modify itself in response to experience and/or environment depends on the plasticity of synaptic connections. Substantial evidence indicates that the number and strength of synapses is readily altered by neuronal activity. This process, known as synaptic plasticity, displays several physiological properties that substantiate its role as a cellular correlate for multiple cognitive processes, including learning and memory. These include the activity dependence and associative nature of induction as well as the input specificity of expression, all of which endow the vast storage and processing capacity of the mammalian brain. Remarkably, BDNF is involved in many of these features. An emerging theme is that BDNF plays a

critical role in regulating several forms of synaptic plasticity in distinct regions of the brain, including the hippocampus and visual cortex.

Input Specificity

Most activity-dependent forms of synaptic plasticity expressed in the brain are input specific, namely modifications at one synapse are not spread to synapses nearby. Experiments have shown that BDNF elicits its actions in a local and synapse-specific manner, with particular preference to active synapses. This unique property may be attributed to several key characteristics of BDNF signaling.

First, secretion of BDNF is activity dependent and is likely to occur at synaptic sites. The control of BDNF secretion occurs relatively fast, usually in a timescale of seconds. Imaging studies demonstrate that BDNF is often co-localized with pre- and postsynaptic markers in hippocampal neurons, suggesting the synaptic localization of BDNF. Application of high-frequency stimulation (HFS) induces a rapid decay of GFP-tagged BDNF, indicating BDNF is secreted in an activity-dependent manner at synapses. Due to its negative charge, BDNF is thought to have a limited capacity for diffusion, and therefore constrains the actions of BDNF at or near its site of secretion.

Second, there is good evidence that BDNF exon II and IV transcripts can be targeted into the dendrites of hippocampal neurons, and neuronal activity enhances such targeting. Recent studies have shown that dendritic BDNF mRNA can be translated locally into BDNF protein. Taken together, a possible scenario is that local synaptic activity triggers dendritic translation of BDNF and may serve as an alternative mechanism to ensure synapse-specific modulation by BDNF.

Finally, local synaptic activity may ensure a better response of target synapses to BDNF by regulating TrkB trafficking. High-frequency neuronal activity has been shown to promote the insertion of TrkB into the surface membrane of hippocampal neurons. This process appears to be ligand independent and requires calcium influx and activation of Ca^{2+}/calmodulin-dependent kinase II (CaMKII). BDNF secreted from active synapses/neurons recruits TrkB from extrasynaptic sites into lipid rafts, microdomains of membrane enriched at synapses. This lateral movement requires TrkB tyrosine kinase activity. Synaptic activity often induces a rise of postsynaptic cAMP. This local increase in cAMP concentration at active synapses facilitates translocation of TrkB into the postsynaptic density, and functions to gate synapse-specific effects of BDNF by controlling TrkB tyrosine phosphorylation locally. Finally, neuronal activity promotes BDNF-induced TrkB endocytosis, a signaling event important for many long-term BDNF functions. All of these could contribute to a more efficient response to BDNF at active synapses.

Early Phase Long-Term Potentiation

It is well established that BDNF plays a key role in LTP, a persistent enhancement of synaptic strength. LTP can be divided into an early phase (E-LTP) and a later phase (L-LTP). E-LTP is relatively short-lasting (1 h) and depends on protein phosphorylation, while L-LTP lasts many hours and requires new protein synthesis. Early work has focused on BDNF regulation of E-LTP in the hippocampus. In neonatal hippocampus in which BDNF levels are low, exogenous BDNF facilitates E-LTP induced by HFS (**Figure 2(a)**). In addition, exogenous BDNF facilitates LTP induced by subthreshold tetanus that normally induces weak potentiation. Conversely, in the adult hippocampus, a stage where endogenous levels of BDNF are relatively high, inhibition of BDNF activity either by function-blocking BDNF antibody or BDNF scavengers, TrkB immunoglobulin G (IgG), attenuates the expression of E-LTP.

In agreement with pharmacological studies, genetically modified mice with mutation of either BDNF or TrkB gene exhibit severe impairments in E-LTP. Interestingly, heterozygous mice (BDNF+/−) with only half of the BDNF gene dosage show a similar degree of impairment as homozygous mice (BDNF−/−), arguing that a critical level of BDNF is important for hippocampal LTP. The impairment of LTP in BDNF-mutant mice is reversed by acute application of recombinant BDNF or by virus-mediated BDNF gene transfer. Deletion of BDNF gene selectively in the adult forebrain by inducible knockout approaches confirms that the effects of BDNF on E-LTP are not due to developmental abnormalities.

The effects of BDNF on hippocampal E-LTP result primarily from alterations of presynaptic function (**Figure 2(b)**). Exogenous BDNF enhances synaptic response to HFS and paired pulse facilitation (PPF), two indicators of presynaptic function. In mice lacking BDNF, posttetanic potentiation (PTP) and PPF are significantly reduced. Electron microscopy reveals a reduction in the number of vesicles docked at presynaptic active zones in these mutant mice. Moreover, biochemical experiments using hippocampal synaptosomes indicate that BDNF modulates the levels or phosphorylation of synaptic proteins involved in vesicle docking and fusion, such as synapsin, synaptophysin, and synaptobrevin. Taken together, the presynaptic role of BDNF for the mobilization and/or docking of synaptic vesicles to presynaptic active

zones may allow hippocampal synapses to follow tetanic stimulation more effectively, resulting in the facilitation of hippocampal LTP.

It is important to note that the biological effects of BDNF are not exclusively the result of its presynaptic actions. In the dentate gyrus, the induction of LTP requires postsynaptic BDNF signaling. A series of studies have demonstrated that BDNF can exert postsynaptic modulatory effects by modulating α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type and *N*-methyl-D-aspartate (NMDA)-type glutamate receptors in neuronal cultures, and some potassium channels in hippocampal slices. In this respect, TrkB receptors have been observed to localize in the postsynaptic density of isolated synaptosomes prepared from cortical neurons. However, whether these postsynaptic modulatory effects of BDNF directly participate in LTP remains to be established.

Late Phase Long-Term Potentiation

Several early studies suggest that BDNF may also play a role in late phase long-term potentiation (L-LTP). L-LTP-inducing tetanic stimulation selectively enhances the expression of BDNF and TrkB mRNAs in the hippocampus. BDNF transcription is regulated in part by CREB, a transcription factor required for L-LTP expression. The delayed and sustained enhancement of BDNF synthesis correlates well with the time course of L-LTP, which increases 2–4 h after L-LTP induction. More direct evidence comes from electrophysiology experiments showing a significant reduction in L-LTP recorded from slices treated with TrkB-blocking antibody, or those from BDNF +/− mice (Figure 2(c)). However, BDNF only regulates L-LTP induced by theta-burst stimulation (TBS) or application of the adenylate cyclase activator forskolin, but not by four spaced trains of HFS, a more standard protocol used to induce L-LTP. These results suggest that strong tetanic stimulation may induce signaling downstream of BDNF, bypassing the requirement of BDNF in L-LTP.

Recent experiments have provided more in-depth insights as to the role of BDNF in L-LTP. In the presence of BDNF, weak tetanus that normally induces only E-LTP resulted in robust L-LTP. Moreover, perfusion of BDNF to hippocampal slices rescued L-LTP that is normally absent when protein synthesis is inhibited. It appears that BDNF is secreted largely in its precursor form (proBDNF), which is converted to mBDNF by extracellular proteases. Biochemical and genetic experiments indicate that this conversion is mediated by an enzymatic cascade that involves tPA and plasmin, two secreted proteases found at hippocampal synapses. tPA has long been implicated in L-LTP, but the precise downstream effector(s) of tPA was not established. The current data support the notion that tPA cleaves the inactive zymogen plasminogen to form plasmin, which in turn cleaves proBDNF to generate mBDNF. Application of mBDNF, but not cleavage-resistant proBDNF, completely reversed the L-LTP deficit observed in tPA and plasmin knockout mice. Thus, conversion of proBDNF to mBDNF by the tPA/plasmin system is critical for L-LTP expression (Figure 2(d)).

Two key cellular events associated with long-term synaptic plasticity are synaptic growth and *de novo* protein synthesis, both of which are regulated by BDNF. In addition to stimulating axonal growth, chronic application of BDNF to hippocampal slices increases the dendritic spine density of CA1 pyramidal neurons. This is particularly relevant since dendritic spines and protrusions are enhanced during L-LTP. In cell cultures, BDNF has been shown to increase mammalian target of rapamycin (mTOR)-dependent translation of a panel of synaptically expressed transcripts including GluR1 and homer2 mRNAs in the dendrites of hippocampal neurons. However, exogenous BDNF applied immediately after strong TBS rescues L-LTP in hippocampal slices in which protein synthesis was blocked for the entire course of the experiments. This provocative result, together with the finding that gene expression of BDNF is stimulated by L-LTP-inducing tetanic stimulation, implies BDNF is a key protein synthesis product required for long-term modifications necessary for L-LTP expression.

Long-Term Depression

In addition to its role in LTP, BDNF regulates long-term depression (LTD), a persistent reduction in synaptic strength induced by prolonged low-frequency stimulation. Expression of LTD is developmentally regulated and exists in several forms mediated by different glutamate receptors. The best-known form of LTD is the NMDA receptor (NMDAR)-dependent form, which is robustly expressed in young animals. mBDNF inhibits LTD in the visual cortex and hippocampus. Collectively, these observations point to a general theme that BDNF facilitates synaptic strengthening, but attenuates synaptic depression.

Analogous to the survival and apoptosis effects of mBDNF and proBDNF, respectively, in the periphery, an important advance is that proBDNF, if uncleaved, enhances NMDAR-dependent LTD via p75NTR in the CNS. Compared to TrkB, the role of p75NTR in synaptic plasticity has not been studied until

recently. In p75NTR mutant mice (p75NTR−/−), NMDAR-dependent LTD was completely absent while other forms of plasticity including NMDAR-dependent LTP and NMDAR-independent LTD were intact. Biochemical experiments indicate that NR2B, a specific NMDAR subunit uniquely implicated in LTD, was significantly reduced in the mutant hippocampus. Whole-cell recordings revealed a severe reduction in NR2B-mediated synaptic currents in CA1 neurons of p75NTR−/− mice. More importantly, application of cleavage-resistant proBDNF increased NR2B-mediated synaptic currents and enhanced LTD in hippocampal slices derived from wild-type mice but not in p75NTR−/− mice (Figure 2(e)). These findings suggest that proBDNF is an endogenous ligand of p75NTR during development, which acts to enhance LTD via modulating NR2B function (Figure 2(f)).

Learning, Memory, and Other Cognitive Functions

Given its central role in synaptic plasticity, numerous studies have examined how BDNF regulates the acquisition (learning) and retention (memory) of new information. Thus far, the strongest correlation is observed between BDNF and hippocampal-dependent forms of memory, which include declarative or episodic and spatial memory. During contextual learning, BDNF expression is rapidly and selectively upregulated in the hippocampus. When BDNF signaling is disrupted either by inhibitors or by genetic knockout, spatial learning is significantly impaired, as reflected by poor performance in the Morris water maze. In many cases impairments in memory were also mirrored with LTP deficits. For instance, deletion of BDNF or TrkB gene in the adult forebrain results in a significant attenuation of contextual fear or spatial memories, as well as hippocampal LTP.

A major advance came from a study on a SNP, which converts a valine (val) to a methionine (met) in the prodomain of the human BDNF gene. This SNP occurs with a frequency of approximately 19–25% in the Caucasian. Human subjects with the met allele exhibit lower hippocampal *N*-acetylaspartate (NAA), a putative measure of neuronal integrity and synaptic abundance. Functional imaging reveals an association of the met allele with abnormal hippocampal activation (Figure 3(a)). Most remarkably, subjects with the met-BDNF allele performed poorer in a hippocampal-dependent episodic memory task, but not in hippocampal-independent working memory and semantic memory tasks. In cultured neurons derived from rodent hippocampus, BDNF (val-BDNF) is packaged in secretory granules that are distributed as puncta throughout cell body and dendrites, with some localized at synapses. In contrast, significantly less met-BDNF-containing granules are localized to dendrites and synapses. Moreover, regulated secretion of met-BDNF induced by neuronal depolarization, but not constitutive secretion, is significantly reduced. Thus, impairments in trafficking, synaptic targeting, and/or regulated secretion may explain the specific memory deficits seen in human subjects with the met allele (Figure 3(b)). These results represent the first demonstration of a role for BDNF in human hippocampal function and of a single gene affecting human episodic memory.

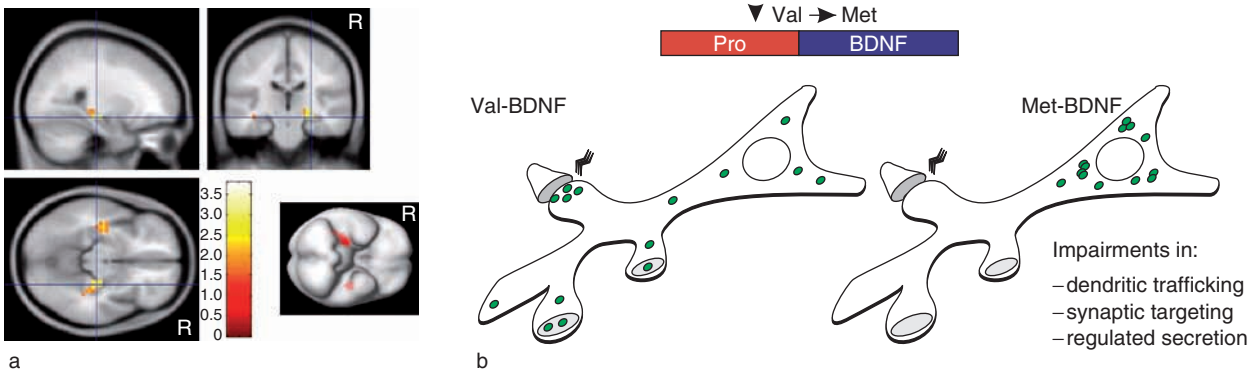

Figure 3 Impact of a SNP in the BDNF gene on cognitive brain function and intracellular trafficking of BDNF. The SNP converts a valine to a methionine in amino acid 66 located in the prodomain of BDNF (val66met). (a) Differences in fMRI responses between val/val and val/met subjects during a memory task. Subjects with val/met genotype exhibit abnormal hippocampal activation (shown in red). The met/met subjects also exhibit deficits in hippocampus-dependent episodic memory. (b) Cellular phenotypes associated with the BDNF val66met SNP. Val-BDNF is distributed throughout a typical hippocampal neuron including distal dendrites and synapses. In contrast, met-BDNF is rarely localized at distal dendrites or synapses and fails to undergo depolarization-induced secretion. Failure of intracellular trafficking and activity-dependent secretion of BDNF may underlie the cognitive deficits observed in subjects with the met allele. (a) Reproduced from Egan MF, Kojima M, Callicott JH, et al. (2003) The BDNF val66met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell* 112: 257–269, with permission from Elsevier.

Genetic analyses in mice show that genes affecting memory performance often impact other cognitive functions. It is now recognized that deficits in BDNF may contribute to neurological and psychiatric disorders. In studies of drug addiction, LTP and LTD have emerged as candidate mechanisms for drug-induced alterations in the nucleus accumbens and ventral tegmental area. Several reports have demonstrated that BDNF modulates behavioral sensitization to cocaine. Substantial evidence also points to its role in depression. There is reduced BDNF expression in the hippocampus of animal models for depression; chronic treatment with antidepressants increases its levels. However, it is unclear whether antidepressants achieve their clinical effects on depression by upregulation of BDNF.

Conclusion

Stemming from the multidisciplinary approaches used in present-day research ranging from cellular systems to behavior, BDNF is now recognized as a key regulator for synaptic circuits underlying many cognitive functions. New and unidentified role(s) of BDNF in plasticity and cognition will undoubtedly continue to surface and will provide abundant intellectual stimulation to drive future advances. Ultimately, understanding the actions of BDNF will aid in the development of therapeutic interventions that will alleviate a wide spectrum of neurological and psychiatric disorders derived from BDNF dysfunction.

See also: Enteric Nervous System: Neurotrophic Factors; Long-Term Depression: Cerebellum; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Synaptic Plasticity and Place Cell Formation.

Further Reading

Barker PA (2004) p75NTR is positively promiscuous: Novel partners and new insights. *Neuron* 42: 529–533.

Chao MV (2003) Neurotrophins and their receptors: A convergence point for many signaling pathways. *Nature Reviews Neuroscience* 4: 299–309.

Chen ZY, Ieraci A, Teng H, et al. (2005) Sortilin controls intracellular sorting of brain-derived neurotrophic factor to the regulated secretory pathway. *Journal of Neuroscience* 25: 6156–6166.

Egan MF, Kojima M, Callicott JH, et al. (2003) The BDNF val66-met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell* 112: 257–269.

Figurov A, Pozzo-Miller L, Olafsson P, Wang T, and Lu B (1996) Regulation of synaptic responses to high-frequency stimulation and LTP by neurotrophins in the hippocampus. *Nature* 381: 706–709.

Lee R, Kermani P, Teng KK, and Hempstead BL (2001) Regulation of cell survival by secreted proneurotrophins. *Science* 294: 1945–1948.

Lou H, Kim SK, Zaitsev E, et al. (2005) Sorting and activity-dependent secretion of BDNF require interaction of a specific motif with the sorting receptor carboxypeptidase E. *Neuron* 45: 245–255.

Lu B (2003) BDNF and activity-dependent synaptic modulation. *Learning and Memory* 10: 86–98.

Lu B, Pang PT, and Woo NH (2005) The yin and yang of neurotrophin action. *Nature Reviews Neuroscience* 6: 603–614.

Nagappan G and Lu B (2005) Activity-dependent modulation of the BDNF receptor TrkB: Mechanisms and implications. *Trends in Neuroscience* 28: 464–471.

Pang PT, Teng HK, Zaitsev E, et al. (2004) Cleavage of proBDNF by tPA/plasmin is essential for long-term hippocampal plasticity. *Science* 306: 487–491.

Patterson SL, Abel T, Deuel TA, et al. (1996) Recombinant BDNF rescues deficits in basal synaptic transmission and hippocampal LTP in BDNF knockout mice. *Neuron* 16: 1137–1145.

Poo MM (2001) Neurotrophins as synaptic modulators. *Nature Reviews Neuroscience* 2: 24–32.

Sun YE and Wu H (2006) The ups and downs of BDNF in Rett syndrome. *Neuron* 49: 321–323.

Woo NH, Teng HK, Siao CJ, et al. (2005) Activation of p75NTR by proBDNF facilitates hippocampal long-term depression. *Nature Neuroscience* 8: 1069–1077.

Behavioral Hierarchies

R Gillette, University of Illinois, Urbana, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Nikolaas Tinbergen wrote the following in 1951:

> It is of great importance for our understanding of instinctive behaviour as a whole to realize that the various instincts are not independent of each other. We have rejected the reflex hypothesis of behaviour and we have seen that each instinctive mechanism is constantly primed, that is to say, prepared to come into action. Such a system can only work because blocking mechanisms prevent the animal from performing continuous chaotic movements.

The term 'behavioral hierarchy' is used to describe the structural organization of behavioral expression. It reflects that certain behaviors can take precedence over others when appropriate stimuli for the multiple behaviors coincide in time. Healthy animals are generally configured for three basic functions: acquisition of nutrient resources, avoidance of debilitating injury and death, and reproduction. Each function subsumes a suite of behavioral subroutines of a size commensurate with the animal's lifestyle complexity, and the expression of the behavioral elements within and between those functions tends to be fractionated. That is, in general behaviors are expressed in a unitary way, so that conflicting behaviors are not expressed together, and the organism may concentrate its attention on achieving its goals efficiently.

The concept of the behavioral hierarchy formalizes the common observations that animals tend to devote themselves rather single-mindedly to affairs at hand that serve motives of goal satisfaction, defense, or simple avoidance of harm. In it, expression of an animal's behavioral repertory is organized to serve the competitive natures of reproduction and resource acquisition. Humans appreciate this intuitively and express it in literature and art extensively founded on interlocking themes of social dominance, reproductive success, and resource acquisition.

Tinbergen's simple and elegant scheme for the organization of behavior derived from assiduous observations of behavioral transitions and the conditions under which they occurred. It resonated with an appreciative audience of behavioral researchers in both field and lab because it had much explanatory and predictive power and because in retrospect it was so obvious. However, one of the most exciting aspects of his work was that he formulated neural hypotheses from his behavioral observations that were testable in the laboratory. While he considered his model as an overly simplified, beginning attempt at explaining the origin of behavioral hierarchies, he well realized its explanatory power.

He conceived of animal behavior as a hierarchical organization of behavioral modules, of suites in which antagonistic components lay within reciprocally inhibitory relationships (Figure 1). Thus, when one component was expressed as a result of excitation by an appropriate sign stimulus, or an internal motivational component, the incompatible behavior would be actively suppressed. His conclusions came from observations such as when the male stickleback was in reproductive condition and actively engaged in nest building, the threshold for releasing aggression against a male conspecific was markedly higher than when not so engaged. He and his contemporaries appreciated the influence of hormonal state on behavioral thresholds, auguring the present understanding of the role of plasticity in behavioral hierarchies.

Tinbergen's work happily coincided with early decades in which emerged appreciation of the nature of the nerve action potential and the integrative power of the synapse. He was also aware of contemporary views that the nervous system was organized in a somewhat hierarchical fashion, such that elements of simpler behaviors often emerged from pattern-generating 'centers' in the spinal cord and were integrated appropriately by specific sensory inputs and descending influences.

Coupled with early demonstrations that simple neural circuits could be deduced in the laboratory as pattern generators for simple behaviors in invertebrates, these proposals gave rise to the field we now recognize as comparative neuroethology, in which a basic premise has been that accurate observations of behavior can generate testable neural hypotheses regarding neural mechanisms. How have the past 60 years of neuroethological investigations refined and modified the concept of behavioral hierarchy?

Early Neural Models of Behavioral Choice

Some of the earliest investigations of arthropod model systems found that simple repetitive behaviors, such as rhythmic swimmeret beating in crayfish and lobster, could be preserved in semi-intact preparations deprived of sensory inputs. This led to the concept of the central pattern generator (CPG), a neural circuit that could generate appropriately patterned outputs by itself, and for which the patterned output could be sculpted by sensory and other inputs to adapt to specific situations; for instance, unevenness of terrain during locomotion. From this it was evident

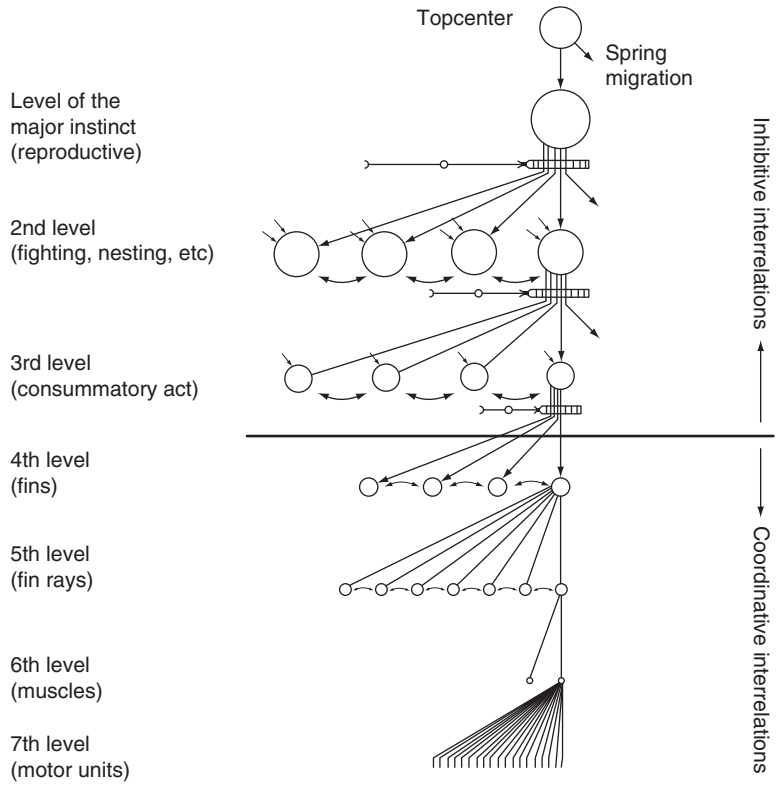

Figure 1 Tinbergen's concept of the hierarchical system of centers underlying an instinct, here shown as for reproduction in the stickleback. Centers, represented as circles at different levels, stand for behavioral suites such as the elements of fighting, nesting, and so forth at the second level. Reciprocal arrows between the centers indicate mutual inhibitory relationships such that behaviors tend to be expressed singly, without conflict. From Tinbergen N (1951) *The Study of Instinct*. Oxford, UK: Clarendon.

that basic outlines of much behavior could be generated centrally and modified by continuous sensory feedback from the environment, a concept that lent a great deal of impetus to neurophysiological inquiry into locomotion and other rhythmic behaviors in invertebrates and vertebrates.

The early probing of the swimmeret and other CPGs employed extracellular electrodes for stimulating and recording; these yielded exact documentation of the phenomenology attending motor output and control, but they could not easily extract detailed integrative mechanisms. Eventually, in a number of instances it, was found that distinct, patterned behaviors (such as swimmeret beating in semi-intact preparations), or their fictive motor output in isolated nervous systems, could be elicited by simple trains of electrical shocks to certain areas of the nerve cord. Moreover, by careful dissection, single nerve fibers could be separated which when stimulated could drive the entire behavior, but with unique effects on frequency or phase coordination that suggested their actions in combination to achieve normal behaviors. Such observations led to the concept of the command neuron, which seemed to fit in well with the putative neural organization of a behavioral hierarchy.

Evolving Perspectives of Behavioral Choice

The command neuron concept borrowed momentum from early ideas of the existence of 'centers' in the brain responsible for distinct behavioral suites. One early and attractive interpretation was that command neurons acted as integrating centers for the excitatory and inhibitory effects of sensation, internal needs, and memory, and that when behavioral thresholds were reached, their activity would command that behavior. Soon commandlike neurons in several different preparations were reliably identified and recorded with intracellular electrodes, allowing their roles in driving motor outputs and integrating sensory influences to be studied in some detail. At one point a forum of neuroethologists convened to discuss their nature, from which emerged the idea that 'command neurons' could be defined as cells "both necessary and sufficient" to drive a patterned motor output. That is, they would be active during the behavior, they could drive it they were when stimulated, and the behavior would be blocked or diminished when those cells were hyperpolarized into silence.

One type of command neuron both necessary and sufficient to its behavior was identified in a number of small rhythmic networks, including those for molluskan feeding, escape swimming in certain sea-slugs, and leech swimming. Spiking in these neurons driven by tonic intracellular current injection drove appropriate motor output of the CPG with an important sequel: rhythmic, alternating excitatory and inhibitory synaptic feedback from the oscillating CPG entrained the command neurons themselves into rhythmic activity locked to one phase of the CPG rhythm. Thus, these neurons were actually intrinsic elements of the CPGs they drove.

The finding of these commandlike neurons invited assessment of their roles in the neural organization of behavioral hierarchies. Detailed studies in mollusks and leech showed that these neurons acted as passive positive-feedback elements in the feeding CPGs, where they acted to provide the degree of excitation necessary to bring the feeding motor network over the threshold for active expression of rhythmic motor output (swimming, biting, and swallowing). The immediate relevance of these command neurons to behavioral hierarchies was that, at first glance, they seemed to be critical sites in the neuronal circuitry where specific motor expression could be suppressed, allowing expression of alternative responses to stimuli (such as active avoidance of food in the mollusk, or shortening vs. swimming in the leech).

The view of command neurons as critical sites where CPG expression could be suppressed was well in line with Tinbergen's predictions of reciprocal inhibition between incompatible behaviors. However, this view is still not the same as the notion that command neurons act as sites where **go, no-go** decisions are made by directly integrating external inputs from both sensory lines and other CPGs. Where they are presently known, those functions appear to lie presynaptic to the command neurons, within the CPGs where they both reside and act.

Neuronal Bases of Behavioral Hierarchies

For Tinbergen and his colleagues, the behavioral hierarchy of the male stickleback in reproductive season was especially clear, and this made the animal a useful model system for investigations that have spanned seven decades. The stickleback studies emphasize both the importance of the apparent roles of inhibition between temporally dominant behaviors, such that the sensory thresholds for expression of subordinate behaviors are increased, and the importance of correct routing of sensory stimuli in expression and sequencing of behavioral elements. They led to the question, how are these regulatory mechanisms actually mediated at the level of the responsible neural networks? This experimental question has been addressable in animals with simple brains and behavior.

Nutrition and reproduction are the major influences on the marked plasticity observed in behavioral hierarchies of most animals. Thus, high levels of nutrient appetite tend to be manifested in arousal and goal-directed activities. These are most notable in motile foragers, and in particular in active predators.

The medicinal leech provides a vivid example of the general types of behavioral changes attending changing hunger and reproductive states that happen in many animals. Marked hunger is manifested in heightened sensitivity to water vibrations and reduced swim thresholds. Immediate plasticity in the hierarchy is mediated by sequential reactions to sign stimuli. For instance, should oriented swimming to a vibration source lead to the lucky find of a bathing mammal, both attachment to the skin substrate and the heat of the warm-blooded animal inhibit further swimming. Crawling may lead to finding a site where warmth, appetitive odors, and thigmotaxic coziness may summate to inhibit crawling and initiate active feeding. Satiation in the leech, after imbibing up to eightfold its weight, completely changes its behavioral responses. Feeding inhibited by gut distention, the leech drops off and, if lucky enough to be back in a lake or stream, heads for the bottom and a safe niche under a rock or sunken log, completely ignoring the appetitive stimuli formerly so attractive. There, the blood meal is digested over weeks to months and converted to reproductive resources, the eggs. During this time the leech is entirely unresponsive to the appetitive influences that drove its predatory behavior, first in the short term because of the gut distension from its meal and then second probably by the reproductive hormonal influences triggered by permissive signals from the nutrient resource.

For vertebrates, detailed mechanisms of the switching mechanisms underlying behavioral hierarchies remain for the most part unknown, due to the relative difficulty of precisely specifying neuronal circuitry. However, a great deal is becoming known about the involvement of peptides and hormones. In mammals appetite and major aspects of behavioral arousal in mammals are organized around peptidergic hormones and transmitters. The satiety mechanisms in mammals are diverse. The mammalian brain keeps an account of its nutrient resources in terms of blood glucose, glycogen stores, fat stores, and digestive activity. Glucose sensor neurons abound in hypothalamic areas regulating appetite. The hormones leptin, secreted from adipose cells proportionate to mass, and ghrelin and other peptides secreted from the stomach act at the hypothalamus on the peptidergic

neuron circuits driving appetitive arousal and behavior. Neurons in the hypothalamus also produce leptin and ghrelin, among other potent peptide neuromodulators, and exert effects on neurons producing orexin and neuropeptide Y peptides that promote appetite and cortical arousal. Orexins are key peptides in stimulation of both appetitive behavior and the arousal side of sleep and wakefulness. Receiving inputs from all nutrient stores, orexin neurons couple behavioral arousal to energy balance and appetite. Behavioral decisions can therefore be made on a long-term cost–benefit basis, in which risk assessment and resource valuation can be estimated in terms of need.

Studies of simpler model systems with both uncomplicated behavioral repertories and accessible nervous systems have cast light on the interactions between motor networks for different behaviors that can organize them into specific priority sequences for expression. For active predators in foraging mode, such as the leech and the California sea-slug *Pleurobranchaea*, feeding behavior dominates most other behaviors. For instance, a sensitive withdrawal reflex is normally stimulated by a rough prod to the anterior region. However, expression of active feeding behavior in the presence of food stimuli actively suppresses the withdrawal in both species. Moreover, most ongoing behaviors can be shut down rapidly and completely by initiation of an escape or startle reaction.

Studies of behavioral hierarchies in gastropods and annelid drew schemes for the organization of behavioral priorities. For the most part, they represented an apparent central position of feeding behavior in the animals' hierarchies: feeding could suppress copulation, reflex withdrawal, and even righting responses that normally occurred when the animal was turned over. Egg laying could raise the stimulus threshold of feeding behavior to levels unlikely to be encountered during the act. However, escape responses superseded all other behaviors. Details in these and other invertebrate model systems differed in ways pertinent to the animals' particular behavioral ecologies.

Invertebrate model systems invited analysis of underlying causality by the accessibility of their nervous systems to circuit analysis. Now Tinbergen's hypothesis could be visited at the neural level to answer how the singleness of behavioral expression actually emerged. Are hierarchies organized on the basis of direct synaptic inhibition among neural networks for different behaviors, or in some other, indirect, fashion?

Initially, it seemed that the simple model would hold. The earliest studies of behavioral choice in *Pleurobranchaea* revealed that food avoidance conditioning was accompanied by synaptic inhibition of feeding command neurons in response to food stimulation. Further, it was demonstrated that active feeding expression caused suppression of reflex withdrawal due to corollary output from the feeding motor network. That corollary output acted to reduce touch stimulus-induced excitation of withdrawal motor neurons, presumably through inhibition of a presynaptic withdrawal interneuron. Similar observations were made in the sea-butterfly *Clione*, a voracious molluskan planktonic predator. Driving identified neurons, the pleural withdrawal cells, caused whole body withdrawal of the animal and inhibition of the swim neurons, which would cause the animal to crumple and drop in the water column – an avoidance behavior. In turn, the pleural withdrawal cells received inhibition when the feeding motor network was activated. In leech, suppression of swimming by activation of shortening was accompanied by inhibition of 'gating,' or commandlike neurons, the 204 cells. The dominance of one behavior over another in these models was in each case explainable by hypothetical inhibitory connections from one motor center to another.

Thus, initial observations in these diverse systems at first seemed very compatible with the concept of simple corollary inhibition of subordinate neural networks by a dominant one, from which the behavioral hierarchy emerged. However, in the absence of more-detailed information about exact neural circuitries, other configurations were not ruled out.

Further investigations in invertebrate model systems markedly refined understanding of the control principles underlying behavioral switching and choice. These involved the neural bases of decisions for alternative expression of escape, feeding, and turning in *Pleurobranchaea*, for ingestive versus egestive motor output in the sea-hare *Aplysia*, and for swimming, shortening, and crawling in the leech *Hirudo*.

Plasticity in Behavioral Hierarchy: Reconfiguration of a Feeding Motor Network

Tinbergen and his contemporaries recognized that behavioral hierarchies were more flexible than those of most human governing and military bodies, from whom the concept of hierarchy was borrowed. Rather, behavioral priorities could change with internal state, such as reproductive condition, to emphasize different goals. Plasticity in feeding behavior, closely regulated as it is by interactions of sensation, appetite, and experience, is one critical aspect that has been studied at the neural level in invertebrate model systems.

Feeding expression, a central aspect of *Pleurobranchaea*'s repertory, is regulated by motivational state (hunger), learning experience, and escape. Sensory thresholds for inducing feeding in animals either satiated or conditioned against a food stimulus with

electrical shock in a food avoidance task rise markedly. In satiated or food avoidance-trained animals, the feeding response is actively suppressed, and they actively avoid food. In analyses of the role of the feeding command neurons, the paracerebral neurons (PCNs), normally excited by food stimuli in hungry animals, showed prolonged inhibition by food stimuli in satiated and trained animals (Figure 2(a)). When the neuronal source of this inhibition was identified, it was intrinsic to the feeding motor network itself, not external in origin.

The origin of the inhibition of the command neurons was in presynaptic neurons that were themselves part of the feeding motor network. An abbreviated structure for the feeding motor network of *Pleurobranchaea* is represented in Figure 2(b). The PCNs are part of the motor network they drive, and during the cyclic biting and swallowing of feeding, their activity is rhythmically entrained to the protraction phase of the cycle of buccal mass movements. Entrainment is accomplished by excitatory feedback from the feeding oscillator during protraction, while inhibitory inputs during the retraction phase punctuate their activity. The inhibitory inputs to PCNs come from small populations of identified interneurons, the interneuron 1 (I1) group, and the interneuron 2 (I2) neurons that potently excite both the I1s and many other retraction neurons in the motor network. Thus, in the hungry, actively feeding animal, the I1s and I2s fire rhythmically during the feeding cycle. However, in satiated and food avoidance-conditioned animals, the I2s fire tonically, driving the I1s to tonically inhibit the PCNs. Through potent outputs to the retraction neurons of the feeding oscillator, tonic I2 activity also essentially locks up the network in the retraction phase of the feeding cycle (Figure 2(c)). In this instance, inhibition of feeding behavior by motivational state and learning is caused by excitation, not inhibition, of elements intrinsic to the underlying network. The mechanism of behavioral choice here is not so much inhibition of the command neurons as a reconfiguration of the feeding network to an alternate state.

The source of the excitatory input to the I2s, as yet unknown, could come from other retraction interneurons of the feeding motor network driven

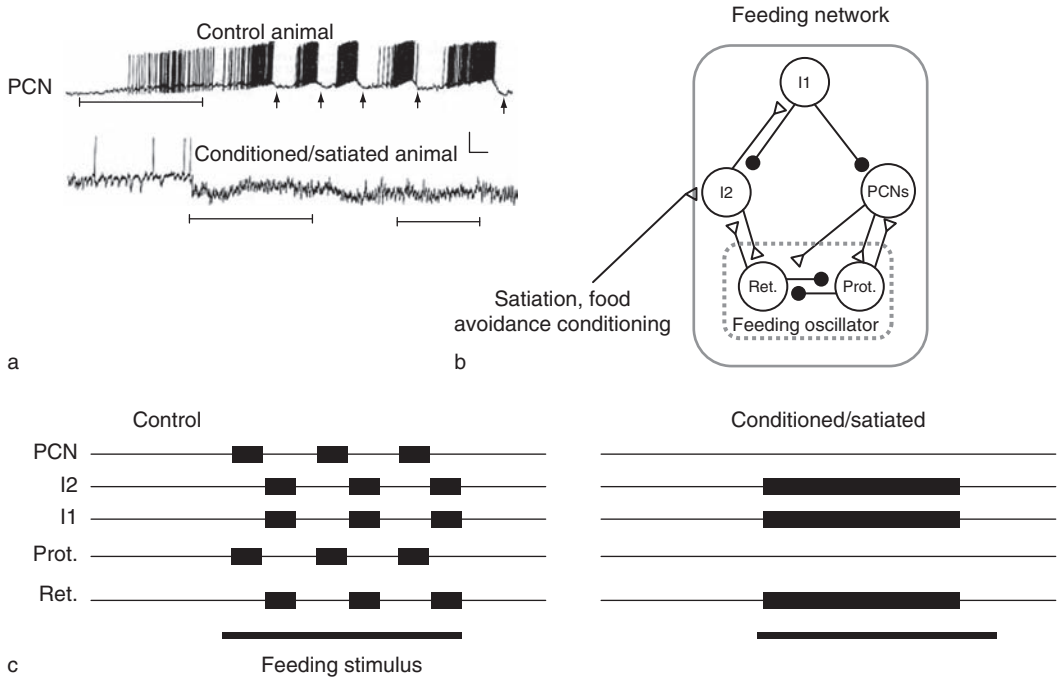

Figure 2 Suppression of feeding by satiation and food avoidance conditioning in *Pleurobranchaea*. Inhibition of feeding command neurons is mediated by inhibitory neurons intrinsic to the feeding CPG. (a) Records from paracerebral neuron (PCN) feeding command neurons during food stimulation (horizontal bars) in a naive hungry animal (top) and in an animal conditioned to avoid food by paired food–shock (bottom). Arrows indicate rhythmic bites. Calibration: 25 mV vertical, 2 s horizontal. Similar inhibitory activity in PCNs is seen in satiated animals. (b) Simplified schematic of the feeding circuit showing relations of the PCN feeding command neurons to the rest of the central pattern generator (CPG) and interneurons I1 and I2. (c) Idealized activities of the neurons of (b) during food stimulation of hungry (left) and food avoidance or satiated animals (right), when the CPG is essentially hung up in the retraction phase. Prot., protraction; Ret., retraction. (a) Reproduced from Davis WJ, Gillette R, Kovac MP, Croll RP, and Matera EM (1983) Organization of synaptic inputs to paracerebral feeding command interneurons of Pleurobranchaea californica. III. Modifications induced by experience. *Journal of Neurophysiology* 49: 1557–1572, used with permission from the American Physiological Society.

by satiation and effects of memory, and/or from collateral activity of the turn avoidance network. Avoidance turns are the alternative behavior released when feeding to appetitive stimuli is suppressed.

Suppression of a Subordinate by a Dominant Behavior: Network Reconfiguration

For inhibition of feeding by escape swimming, the mechanism of feeding inhibition is instructively different. Again, actual inhibition of feeding is excited in intrinsic inhibitory neurons of the feeding network. The dominance of escape swimming over feeding arises from actions of corollary outputs of the swim CPG (Figure 3). Inhibition of feeding is specifically exerted by the swim CPG output interneuron A-ci1 and probably yet unidentified cells. A-ci1 is driven by the pattern-generating subset of swim interneurons, A1, A10, and A3, to cause a global inhibition of the swim motor network. A-ci1 causes tonic excitation of the I1 group of the feeding CPG. Excitation of the I1s by A-ci1 interrupts the normal cyclic A1 activity during feeding and tonically inhibits the PCN feeding command neurons. However, as also shown in Figure 3(b), the I1s also provide feedback inhibition to their presynaptic I2 neurons. Activity in the I2 neurons completes a positive feedback loop in

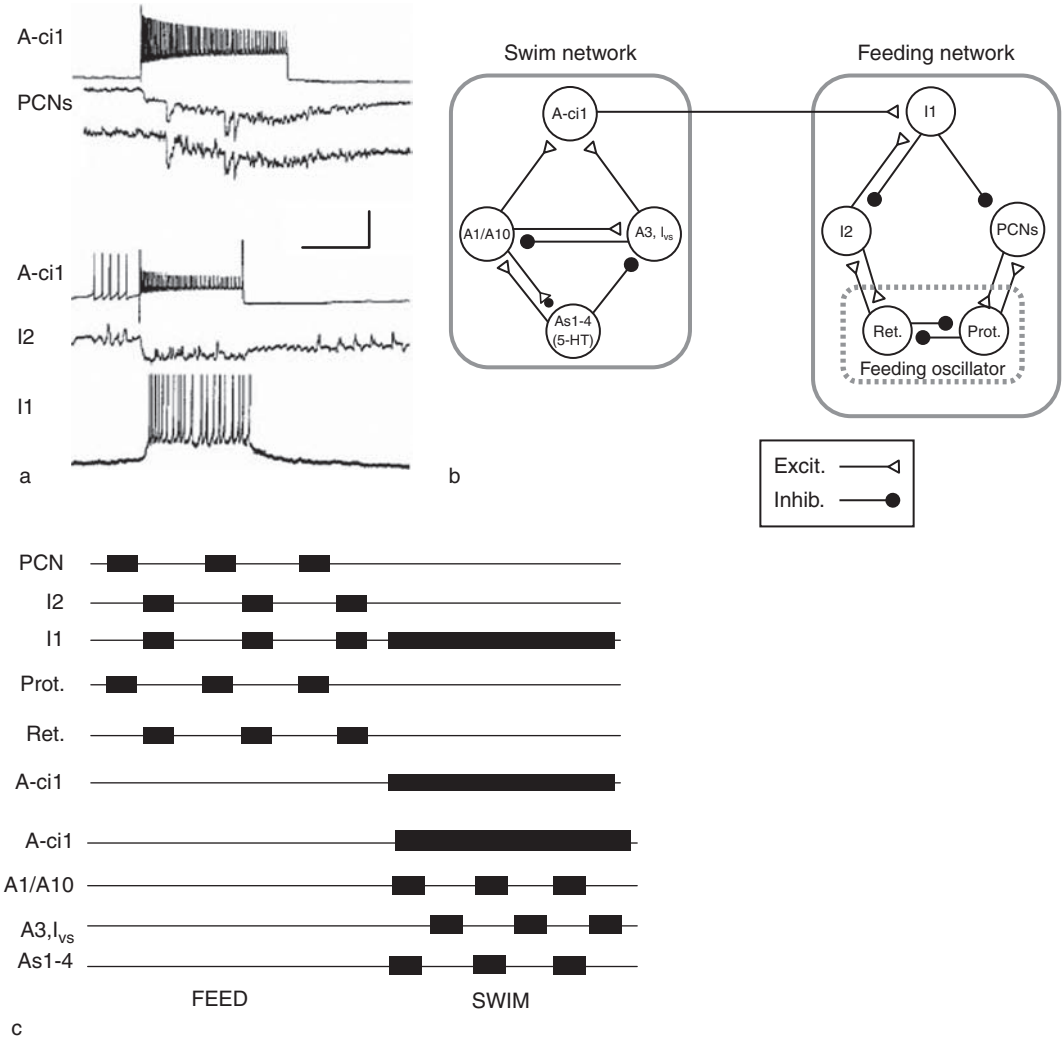

Figure 3 Suppression of feeding by swimming through corollary outputs of the swim network, swim central pattern generator (CPG) output interneurons (A-ci1), that excite inhibitory neurons, I1, in the feeding CPG that directly inhibits both the paracerebral neuron (PCN) feeding command neurons and the I2 major interneurons. (a) Intracellular records showing that driving A-ci1 excites I1 to inhibit PCNs and I2s. A-ci1s are swim network outputs. (b) Synaptic interactions within and among swimming and feeding networks. (c) Idealized activity in the feeding network and A-ci1 during interruption of ongoing feeding activity by an escape swimming. A1, A3, and A10, swim interneurons; As1–4, multifunctional serotonergic neurons; 5-HT, 5-hydroxytryptamine (serotonin). Prot., protraction; Ret., retraction. Calibration: vertical, 40 mV of all top records, 4 mV for bottom records, except for the I1 record in (b), 8 mV; horizontal 3 s. (a) Reproduced from Jing J and Gillette R (2000) Escape swim network interneurons have diverse roles in behavioral switching and putative arousal in Pleurobranchaea. *Journal of Neurophysiology* 83: 1346–1355, used with permission from the American Physiological Society.

the retraction neuron population that is necessary to normal feeding motor output. Thus, during escape swimming there is a global inhibition of the feeding oscillator network, effectively shutting down both populations of protractor and retractor neurons that compose the feeding motor oscillator and its outputs.

In the cases of suppression of feeding by either satiation/learning or escape, inhibition of the behavior is not mediated by direct synaptic inhibition from one CPG to the other. Instead, the feeding network as a whole is reconfigured through targeting strategically placed interneurons (which are not themselves command neurons but have broad effects within the network). Instead of broadly distributed inhibition from extrinsic inputs, the inputs are excitatory. The targeted neurons, activated tonically, work to change the active network configuration.

Moreover, the different modes of feeding suppression exerted from the two sources have adaptive significance. In suppression of feeding by satiation/learning, the network is locked up in one mode of its normal motor output, retraction. Increasing the appetitive sensory stimulus and providing additional excitation to the network can bring it over threshold to express a coordinated feeding attack. In contrast, the suppression caused by escape is much more complete and must ensure that the escape behavior will not be compromised by inappropriate expression of feeding.

Suppression of Turning by Swimming: Recruitment of Essential Elements

A third type of example of inhibition of one behavior by another involves quite a different mechanism: recruitment of neural elements essential to one behavior by a second, dominant behavior. This is seen in the behavioral choice between escape swimming and turning behavior in *Pleurobranchaea*. Here, elements providing the intrinsic network excitation of the pattern generator for one behavior (turning) are effectively assimilated by the pattern generator for the other (swimming). The avoidance turn swivels the animal's body through a wide angle and is followed by rapid locomotion. It is commonly triggered by noxious stimuli such as recognition of a larger, cannibal conspecific. Escape swimming can interrupt the avoidance turn to effect a faster getaway and is commonly triggered at a higher stimulus threshold.

It appears to work as follows, as outlined in Figures 4(a)–4(c): the swim network and the turn network share the four multifunctional serotonergic neurons, As1–4. At lesser intensities of noxious stimulus, As1–4 are recruited into an emergent turn CPG along with the bilaterally paired A4 turn command neurons to mediate an avoidance turn. As the turn CPG is active, the As1–4 fire tonically at a moderate frequency (3–7 Hz; Figure 4(b)), driving intense, tonic, and unilateral A4 activity that itself powers the turn. At higher stimulus intensities, the firing rates of the As1–4 more than double to activate the CPG of the escape swim. The ensuing cyclic inhibition–excitation of As1–4 interrupts and suppresses both tonic A4 activity and the avoidance turn. Thus, when the As1–4 neurons are co-opted into the swim oscillator, the A4 cells are entrained into weak cyclic activity phase locked to the swim cycle and prevented from the tonic unilateral discharge necessary to avoidance turns (Figures 4(b) and 4(c)).

The switch between the escape swim and the avoidance turn in *Pleurobranchaea* is an example of behavioral choice mediated by switching among metastable states of a neural network. Feeding networks in particular may display multiple states, among which transitions are driven by different combinations of both internal and external stimuli (Figure 4(d)). A similar mechanism mediates switching between locomotion and escape swimming in the related sea slug *Tritonia*.

Reversing Motor Output: Reconfiguring the Network

Another example of network reconfiguration underlies switches from ingestion to egestion in mollusks, where the feeding circuitry is reconfigured for opposite behavioral action. A circuit mechanism is being elucidated in the feeding circuitry of *Aplysia*, where it has been shown that the change in coordination of the feeding network from biting and swallowing to rejection is comparable to neuronal models of gait change in vertebrate locomotion.

In *Aplysia* feeding, during retraction the radula closes to grasp food material and draw it into the buccal cavity. During egestion, radula closure is shifted to happen in protraction, to move material out of the buccal cavity. Command neurons of the feeding oscillator in *Aplysia* are named cerebrobuccal interneuron 2s (CBI-2s), likely homologs of the feeding command neurons found in other gastropods. During ingestion another cerebrobuccal neuron is activated, the CBI-3. The CBI-3s are electrically coupled to CBI-2s and fire during the same protraction phase of CPG activity. They make simple synaptic connections within the CPG that reverse the phase of activity of the radula closure motor neurons, from retraction to protraction, and thereby cause egestion. A partial circuit model for switching between ingestion and egestion is shown in Figure 5. While all the circuit details of the switch are not yet known, the simple reconfiguration of the CPG to produce behaviorally opposite actions is an elegant example of multifunction and economy in neural circuit design.

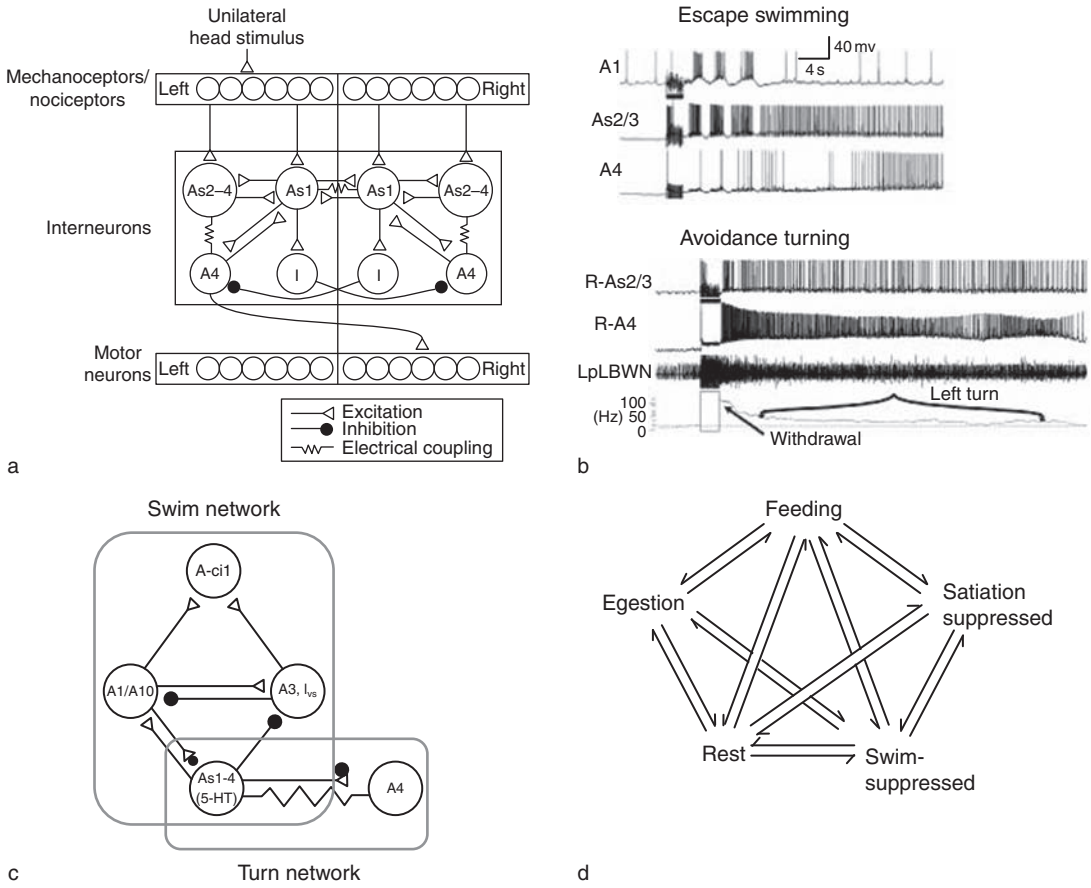

Figure 4 Suppression of the avoidance turn by escape swimming in *Pleurobranchaea* by recruitment of essential turn neurons into the swim central pattern generator. (a) A schematic of the turn network, configured for avoidance. (b) Intense bursting activity of the serotonergic multifunctional (As1–4) neurons provides excitation-sustaining activity of the escape swim network (top), while slower tonic activity supports the avoidance turn activity mediated by the turn command (A4) neurons (bottom). pLBWN, posterior lateral body wall nerve; R and L refer to right and left sides. (c) A simplified schematic representing the relationship between the escape swim and turn networks, which share the As1–4s. Excitatory connections are indicated by open triangles and inhibition by blackened circles. Lines with both symbols mean that both excitation and inhibition are produced by that pathway. (d) A diagram of the multiple states and their transitions found in a molluskan feeding network in *Pleurobranchaea californica* resulting from different combinations of sensory stimuli and internal state. A1, A3, and A10, swim interneurons; 5-HT, 5-hydroxytryptamine (serotonin). (a) From Jing J and Gillette R (2003) Directional avoidance turns encoded by single interneurons and sustained by multifunctional serotonergic cells. *Journal of Neuroscience* 23: 3039–3051.

Behavioral Switching With Large Differences in Timescale: Network Reconfiguration

For the leech the application of voltage-sensitive dyes to imaging simultaneous activity in multiple neurons powerfully complements standard microelectrode methods for identifying and characterizing CPG elements. Swimming is produced by alternating dorsal and ventral contractions of longitudinal muscles in a traveling wave that goes from anterior to posterior. Dorsoventral muscles are tonically contracted to flatten the body. During crawling, circular muscles contract to elongate the body, alternating with simultaneous contraction of dorsal and ventral longitudinal muscles to cause a segmental contraction that is propagated posteriorly. Switching between swimming and crawling can happen rapidly. By elicitation of fictive swimming and crawling in the same preparation, neurons can be identified that participate in both rhythmic behaviors. The results indicate that some neurons are multifunctional whereas others appear to be dedicated. The overlapping of the swim and crawl networks is very large, up to 93%.

All species of leech can crawl although not all swim, and it is thus likely that the swim CPG evolved from the more primitive one for crawling. Small but significant modifications may have accrued to allow the CPG to become multifunctional, capable of operating in more than a single state of coordination and in very different timescales. The results of these studies have led to the conclusion that behavioral decisions in this system are mediated by a combinatorial code of multiplexed interneurons.

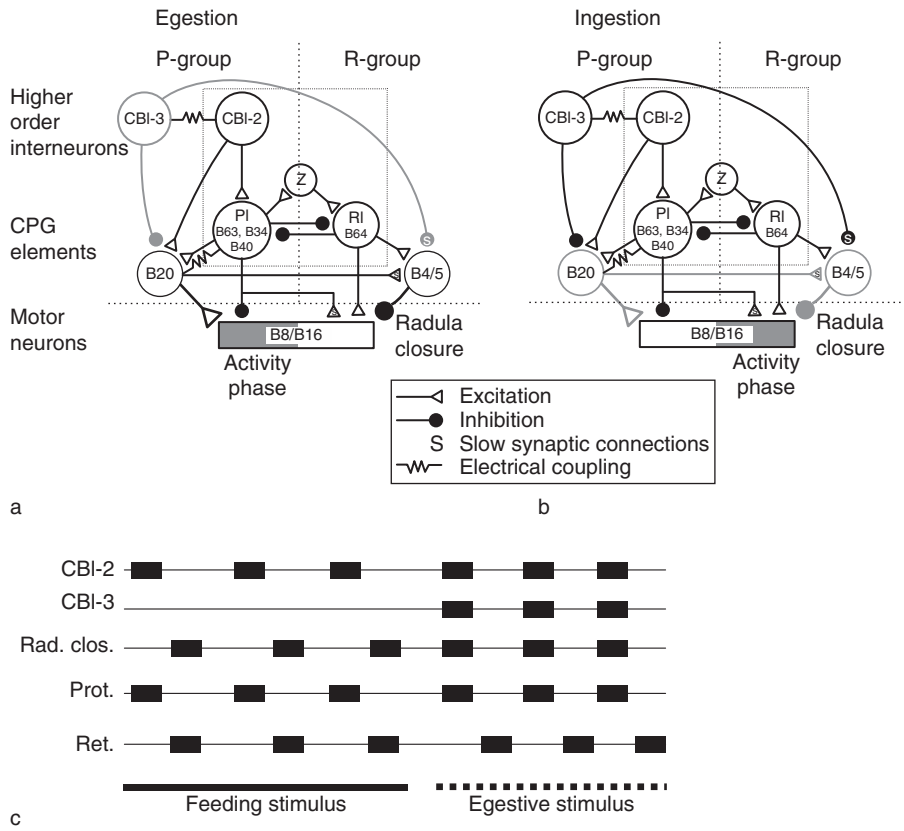

Figure 5 Switching between ingestion and egestion in *Aplysia*. During egestion (a), excitation of radula closure motor neurons during protraction is driven by strong outputs from neuron B20, while interneurons B4/5 inhibit the motor neurons during retraction. In ingestion (b), coactivation of cerebrobuccal interneuron (CBI)-3 with the command neuron CBI-2 during protraction inhibits B20 and B4/5, allowing other inputs from the central pattern generator (CPG) to invert the phase of radula motor neuron excitation to retraction, which would transport material into the buccal cavity. Feedback connections from the CPG to command neurons are not shown. (c) Idealized phases of activity of the motor network during ingestion and egestion. Prot., protraction; Rad. clos., radula closure; Ret., retraction; P- and R-groups fire during protraction and retraction, respectively; RI and PI are interneurons of the CPG, the Z neurons are hypothetical. (a, b) From Jing J and Weiss KR (2001) Neural mechanisms of motor program switching in Aplysia. *Journal of Neuroscience* 21: 7349–7362.

Invertebrate motor networks provide other instructive examples of how organization of behavioral hierarchy is shaped at the neural level by the transitions between CPG configurations and resulting outputs. This is exemplified in Figure 4(d), which diagrams the distinct states that one such motor network can assume and the transitions it can undergo. The result is reminiscent of the allosteric regulation of complex holoenzymes or ion channels. Distinct inputs direct the network from a 'resting' state into metastable, or partially stable, configurations which can transit to almost any of the others if the input configuration changes, or collapse back to rest. Such plastic configurations represent in principle how multifunctional feeding CPGs can generate both ingestion and egestion, how the locomoting mammal can switch among a variety of gaits, and how the leech can swim, crawl, or shorten using common neuron sets.

No Central Role for Command Neurons in Decision?

In those instances in which a subordinate behavior is suppressed by a dominant one and the neural mechanism is partly understood, the source of command neuron inhibition was found to be intrinsic to the CPG. Suppression of feeding in these instances, was due not so much to command neuron inhibition as to a general reconfiguration of the feeding network to nonfeeding states that incorporated silencing of the PCNs. Here, then, decision is mediated by transition of the feeding network between metastable configurations.

Indeed, two different identified command neurons were tested in the feeding network of the pond snail *Lymnaea* for their roles in actually initiating feeding. It was found that neither was activated faster than a major interneuron of the CPG core. Their actions were to influence motor neuron burst duration in one case and to set the frequency of the feeding rhythm in the other. Neither have single or a small number of neurons been found in the leech that mediate behavioral switching. Thus, combinatorial codes of activity in the different CPG elements appear to underlie the transitions from which emerge hierarchies of behavior.

A simple speculation based on the collective data is that command neurons have roles in stabilizing specific network configurations and in providing a sharp internal CPG switching function that allows vigorous behaviors like swimming or feeding to be turned on and off sharply. Thus, command neurons may not themselves integrate sensory input so much as they integrate the excitation state of their network.

Conclusion

Much remains to be learned of the circuit bases of behavioral hierarchies. However, studies of simple systems have gained several important initial insights. One is that behavioral priorities may be achieved through a dominant CPG acting to suppress competing CPGs and to promote synergistic actions. This can be effected via corollary discharge outputs that reconfigure a subordinate CPG into an inactive or synergistic configuration or through appropriation of essential elements of the subordinate CPG into its own structure or by reconfiguration of a CPG to produce an alternate output.

A second insight regards the nature of certain command neurons, whose functions do not fulfill some previously expected roles. Acting as positive feedback elements in CPGs, they do not act directly to initiate CPG activity but likely accelerate it and may stabilize certain configurations. They are not the sites where a behavioral decision is made, although the decision is strongly manifested in their activity. They are not directly excited by sensory inputs that initiate behavior, nor are they directly inhibited by inputs from dominant CPGs. Rather, when it happens, their inhibition is exerted by inhibitory neurons of their own CPG, which is restructured for a new configuration of activity. Thus, if one wished to alter expression of hierarchical relationships in behavior by some external intervention at the level of the neural circuit, the command neurons are probably not such strategic targets as the CPG neurons that may play roles normally in switching CPG configuration.

Presently it is appreciated, from a few instances in which circuit mechanisms of behavioral choice have been elucidated, that the organization of behavioral hierarchies and choice is based on the observations that CPGs may exhibit plural metastable states of coordination. Those states can be melded or parsed into other networks newly organized into activity by circumstance, where the expression of any one network may be largely dependent on the dynamic configuration of sensory and other inputs from other networks.

The contributions of Nikolaas Tinbergen's field and laboratory studies to the modern understanding of neural bases of behavior continue to strongly influence modern neuroethology. They testify to the power of behavioral observations in predicting the general nature of the underlying neural circuitry and in providing testable hypotheses to drive investigation of detail. Thanks to this, increasing knowledge of the fine details of interactions among neural circuits that organize the expression of behavioral hierarchies is providing a base to understand the neural substrates of cost–benefit decision and social behavior in simple systems, thereby providing valuable models to approach these subjects in more complex animals.

See also: Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Command Systems; Decision-Making and Neuroeconomics; Development of Behavior; Goal-Directed Behavior Theories; Reward Decision-Making.

Further Reading

Briggman KL and Kristan WB (2006) Imaging dedicated and multifunctional neural circuits generating distinct behaviors. *Journal of Neuroscience* 26(42): 10925–10933.

Davis WJ, Gillette R, Kovac MP, Croll RP, and Matera EM (1983) Organization of synaptic inputs to paracerebral feeding command interneurons of Pleurobranchaea californica. III. Modifications induced by experience. *Journal of Neurophysiology* 49: 1557–1572.

Gillette R (2006) Evolution and function in serotonergic systems. *Integrative and Comparative Biology* 46: 838–846.

Gillette R and Jing J (2001) The role of the escape swim motor network in the organization of behavioral hierarchy and arousal in *Pleurobranchaea*. *American Zoologist* 41: 983–992.

Grillner S (2003) The motor infrastructure: From ion channels to neuronal networks. *Nature Reviews Neuroscience* 4: 573–586.

Jing J and Gillette R (2000) Escape swim network interneurons have diverse roles in behavioral switching and putative arousal in Pleurobranchaea. *Journal of Neurophysiology* 83: 1346–1355.

Jing J and Gillette R (2003) Directional avoidance turns encoded by single interneurons and sustained by multifunctional serotonergic cells. *Journal of Neuroscience* 23: 3039–3051.

Jing J and Weiss KR (2001) Neural mechanisms of motor program switching in Aplysia. *Journal of Neuroscience* 21: 7349–7362.

Kristan R and Gillette R (2007) Behavioral choice. In: North G and Greenspan RJ (eds.) *Invertebrate Neurobiology*, ch. 20. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Kristan WB, Calabrese RL, and Friesen WO (2005) Neuronal control of leech behavior. *Progress in Neurobiology* 76: 279–327.

Tinbergen N (1951) *The Study of Instinct*. Oxford, UK: Clarendon.

Bell's Palsy

J J Halperin, Overlook Hospital – Atlantic Health, Summit, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Each cranial nerve has its own idiosyncratic propensity to damage, dictated at least in part by its anatomy. The fine filamentous nature of the trochlear (4th) nerve makes it susceptible in minor head injury. The long intracranial course of the abducens (6th) puts it at risk with any increase in intracranial pressure. The facial nerve (7th) has a particularly lengthy and complex course, crossing the subarachnoid space, winding through a long, narrow bony canal, exiting to cross through the parotid gland; it should not be surprising that it can be affected in a variety of circumstances. That the detailed pathophysiology of facial nerve palsy remains unclear – over 185 years after its initial description – is remarkable. However the facial nerve's peculiar anatomy must contribute to its particular susceptibility to damage.

In recent years the number of disorders clearly associated with the facial nerve has increased to include sarcoidosis, Lyme disease, HIV, and herpes (zoster and simplex) infections, yet pathophysiologic insights remain frustratingly limited. The therapeutic approach similarly has expanded, but each approach remains controversial. The following overview will attempt to summarize what is known, and elucidate these controversies.

Background

Sir Charles Bell (1774–1842), a gifted Scottish-born surgeon, anatomist, and illustrator, described the course of the facial nerve, and the peculiar 'blight' that befalls it, in 1821. Interestingly, this arguably was one of his lesser accomplishments. The first Professor of Anatomy and Surgery of the College of Surgeons in London, and the first Professor of Physiology at King's College, University of London, among his noteworthy neurologic accomplishments was introducing the concept that motor and sensory functions were segregated in the spinal roots! Also responsible for describing 'Bell's phenomenon,' he is, however, best memorialized in the eponym for a 'palsy' (perhaps a somewhat dubious honor). Over the years, many have described the clinical concomitants of Bell's palsy. Proposed therapeutic interventions have ranged from surgical decompression of all or part of the nerve, to psychological counseling, to numerous points in between.

The facial nerve's anatomy is complex (schematically shown in Figure 1) both functionally and anatomically. Originating in a nucleus in the pontine tegmentum, the motor fibers take a looping dorsal, then ventral, course around the 6th nerve nucleus, joining fibers connecting to the nucleus solitarius and lacrimal and superior salivary parasympathetic nuclei as they travel. Fibers exit the brain stem and travel adjacent to the 8th nerve in the cerebellopontine angle, cross the subarachnoid space, enter the internal auditory meatus, and then traverse the longest, most complex intraosseous course of any nerve in the body. Throughout its course, components come and go. At the level of the geniculate ganglion, parasympathetic fibers (which exit the brain stem in the anatomically distinct nervus intermedius) form the greater petrosal nerve and project to the pterygopalatine ganglion, from which postganglionic fibers pass to the lacrimal gland and to the blood vessels of the Circle of Willis. Damage to this branch results in disorders of lacrimation – either increases or decreases. Clinical interpretation of these changes is often confounded in facial nerve palsy by the equally important contribution of mechanical exposure of the eye, which in and of itself can lead either to increased tearing from irritation or to exposure-related drying.

Other parasympathetic fibers continue with the main trunk of the facial nerve, to emerge slightly more peripherally and contribute to the chorda tympani, through which fibers project to the submandibular ganglion, with postganglionic projections to the salivary nuclei. Neurons of the geniculate ganglion convey sensory input from the nasopharynx, tympanic membrane, and the ear canal. Although these fibers travel with the facial nerve, they ultimately project to the primary sensory nucleus of V and the nucleus of the spinal tract of the trigeminal nerve.

Distal to the geniculate ganglion and the greater petrosal nerve, a small motor branch exits to innervate the stapedius muscle. Damage to this results in hyperacusis. Shortly thereafter, the chorda tympani joins the nerve. This bundle, which, as mentioned, carries efferent preganglionic parasympathetic fibers destined to control salivary glands, also contains fibers of the lingual nerve, conveying taste sensation from the anterior two-thirds of the tongue. Cell bodies of these neurons also reside in the geniculate ganglion; fibers project centrally via the nervus intermedius to the nucleus solitarius. The fibers that remain distal

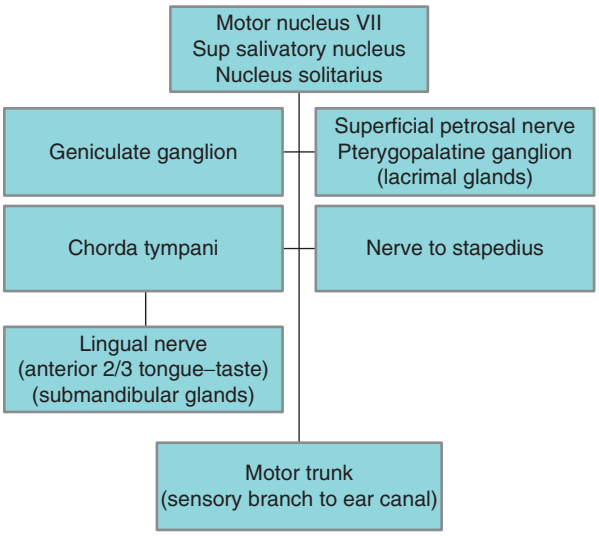

Figure 1 Schematic diagram of facial nerve.

to the chorda tympani are primarily motor, projecting to the muscles of facial expression (forehead, periorbital, and perioral muscles) and a small group of sensory fibers destined for the skin of the anterior wall of the external auditory canal.

The nerve's long course puts it at risk in multiple locations. Meningeal processes (microbial or neoplastic meningitis) can affect it in the subarachnoid space; skull base tumors can compress it within the long bony canal; inflammatory or neoplastic disease can involve it as it traverses the parotid gland (e.g., in sarcoid). As a fairly long peripheral nerve, it is at risk in peripheral neuropathies, ranging from diffuse demyelinating processes such as Guillain–Barré syndrome, to a mononeuropathy multiplex, as occurs in Lyme disease.

Clinical

Facial weakness can result not only from damage to the facial nerve, but also from damage to descending cortical projections. The clinical observation is that damage to the central descending motor pathways tends to affect movements of the mouth and cheek far more than eye closure and forehead wrinkling. This differentiation, which is generally obvious in the postacute setting, may be less apparent immediately following an acute central event (e.g., lacunar stroke), particularly if it involves deep hemispheral white matter; in such individuals some eye closure and forehead weakness may be evident acutely, improving over the first few postevent days.

The disparity in involvement between the upper and lower face is usually attributed to cortical input to the forehead being bilateral, with that to the lower face being more lateralized. An alternative hypothesis suggested more recently is that the upper face actually receives little cortical motor input from either hemisphere; this would explain clinical observations equally well.

Lesions of the facial nerve can occur anywhere along its course. Damage in the pons is often accompanied by abducens weakness, due to the close proximity of the motor fibers of 7th to the 6th nerve nucleus. As with all brain stem processes, the high density of eloquent structures tends to lead to multiple other associated findings, leading to little uncertainty as to the location of the lesion.

Damage to the nerve outside the brain stem can lead to a number of symptom combinations, depending on the location of the lesion along the nerve's long course, and also on whether damage is to all fascicles of the nerve at that level, or just some. In theory, a proximal lesion would affect all functions, resulting in abnormal lacrimation, hyperacusis, abnormal taste and salivation, weakness, and a small area of hypesthesia on the external ear. As lesions occur more and more peripherally, fewer and fewer of these elements would be evident. Curiously, when this has been studied, in at least some cohorts of patients with facial palsy, no such systematic segregation of syndromes is evident, suggesting that lesions are selective at a fascicular level within the nerve, as well as along its longitudinal course.

The other common clinical concomitant of facial nerve palsy is significant retroauricular pain ipsilateral to the palsy. This often precedes facial weakness by 1 or even 2 days, and can be quite severe. Occurring in the region of the stylomastoid foramen, where the nerve exits the skull, this is often attributed to the nerve swelling in its narrow canal. Given the small sensory component of the nerve, it could also be neuropathic referred pain, from damage to this small group of periauricular afferents. This pain typically resolves in the first few days; corticosteroids may speed this.

The weakness typically occurs fairly rapidly, evolving over about 24 h, usually maximal within 48 h, rarely progressing over as much as a week. Weakness involves all parts of the affected side of the face – forehead wrinkles are lost and the affected side of the forehead cannot be elevated. Eye closure becomes weak, often with sagging of the lower lid. The resulting exposure of the cornea can be one of the most dangerous complications of this disorder. The corner of the mouth cannot be lifted in a smile; lips cannot be sealed or the cheek puffed on the affected side. Drooling out of the affected corner of the mouth is often the patient's first hint of a significant problem. Many patients will notice hyperacusis – although they rarely volunteer this information. Similarly, they often note a change in taste

perception – in some this may be due to decreased salivation but more often it is a true dysgeusia. Proving this requires testing taste on the two sides of the anterior tongue, without allowing the patient to retract the tongue back into the mouth between tests, since diffusion of the test substance will immediately invalidate the test. Patients often describe numbness on the affected side of the face, but objective testing rarely confirms this. Most likely this represents misperception of proprioceptive input with attempted motion of the paralyzed facial muscles.

Epidemiology

The terms 'facial palsy' and 'Bell's palsy' are not synonymous – the latter term is usually restricted to patients with idiopathic acute deficits of the facial nerve.This definition is selected to exclude slowly evolving facial nerve lesions such as those due to posterior fossa or skull base tumors, as well as brain stem strokes, skull base fractures, meningitis-associated multiple cranial neuropathies, Ramsay Hunt syndrome, etc. However, this definition now becomes problematic, since there are quite suggestive data that in many instances Bell's palsy is caused by herpes simplex virus. Since these cases are no longer idiopathic, must they be removed from those considered to be Bell's palsy? What about patients with facial nerve palsies associated with diabetes or with Lyme disease? Clearly, at some point this more than semantic issue will require resolution; for now, cases associated with herpes simplex, Lyme disease, and diabetes are generally included as Bell's palsy.

The reported incidence of Bell's palsy varies somewhat, but is generally felt to be about 20–25 cases per 100 000 population per year, affecting about 1 in 60–70 individuals in their lifetime. The disorder is quite infrequent under the age of 10 years, increases by decade from ages 10 to 39 years, is fairly constant at about 35 cases per 100 000 from ages 40 to 69 years, then increases to over 50 per 100 000 (10 times the childhood incidence) in those over 70 years. There is no significant male: female or side-to-side difference. Involvement is bilateral occasionally; in about 6% of cases it can be recurrent (the recurrence occurring on either the previously affected or unaffected side). The paralysis is complete in a significant proportion of patients, estimates ranging from 45% to 70% in different series. (The former number, from a population-based study in Rochester, MN, probably reflects a number as free of selection bias as possible.) No seasonal or geographic clustering has been evident.

Bell's palsy probably occurs with increased frequency in diabetics. Hypertension has also been considered a risk factor. Although it does appear to increase the likelihood of incomplete recovery, whether the incidence is increased among hypertensives is unclear. Pregnancy has been reported as a risk factor, but results have been conflicting; the most systematic studies have failed to support this association. Incidence is increased in patients with sarcoidosis, in whom involvement can be meningeal, within the parotid, or at points in between. Similarly, incidence is increased in HIV and Lyme infection. When bilateral facial palsy occurs it may be indicative of the Guillain–Barré syndrome, sarcoidosis, or Lyme disease.

Recovery is complete in the vast majority (80–85%) of patients with Bell's palsy. The degree of improvement is commonly graded with the House and Brackmann Grading Scale (Table 1). More extensive lesions tend to predispose to incomplete recovery – presence of pain in addition to the retroauricular pain, complete facial paralysis, tearing changes, and sensory changes are all associated with incomplete recovery. Incomplete recovery is somewhat more common among those over 55 years of age. In the most extensive study of this, complete recovery occurred in 90% of those under 55 years and 79% of those 55 years and older. Although several comorbidities are linked to Bell's palsy, only hypertension is associated with a somewhat worse prognosis.

Interestingly, when all of these factors were considered in a multivariate analysis, only complete paralysis, non-ear pain, and hypertension were found to predict a somewhat worse prognosis. Other factors, such as age, and other neurologic changes apparently merely reflected the impact of these three factors. Among patients with none of these risk factors, complete recovery occurred in 96%. With an overall spontaneous recovery rate of about 85%, and 96% among low-risk patients, it is not difficult to understand why it has been so difficult to determine efficacy of different treatment strategies, and why their use has been controversial.

Pathophysiology

Understanding the mechanism underlying this mononeuropathy has been challenging. A few autopsy cases, published many years ago, were felt to demonstrate inflammatory changes in the affected nerve. Subsequent review raised the possibility that the observed changes were merely what one would expect to see as part of Wallerian degeneration in a nerve damaged more proximally. Surgical series reported swollen-appearing nerves – since it is relatively uncommon to explore normal facial nerves, subsequent authors, including Sunderland, have suggested that this was merely the nerve's normal appearance.

Table 1 House and Brackmann facial nerve grading system

Grade	*Description*	*Characteristics*
I	Normal	Normal facial function in all areas
II	Mild dysfunction	Gross: slight weakness on close inspection; slight synkinesis At rest: normal symmetry and tone Motion Forehead: moderate to good function Eye: complete closure with minimal effort Mouth: slight asymmetry
III	Moderate dysfunction	Gross: obvious but not disfiguring difference between two sides; noticeable but not severe synkinesis, contracture, and/or hemifacial spasm At rest: normal symmetry and tone Motion Forehead: slight to moderate movement Eye: complete closure with effort Mouth: slightly weak with maximum effort
IV	Moderately severe	Gross: obvious weakness and/or disfiguring asymmetry At rest: normal symmetry and tone Motion Forehead: none Eye: incomplete closure Mouth: asymmetric with maximum effort
V	Severe	Gross: only barely perceptible motion At rest: asymmetry Motion Forehead: none Eye: incomplete closure Mouth: slight movement
VI	Total paralysis	No movement

From House JW and Brackmann DE (1985) Facial nerve grading system. *Otolaryngology – Head and Neck Surgery* 93: 146–147.

Neurophysiologic observations have not been as informative as one might wish. The standard method used, electroneurography, consists of stimulating the facial nerve at the stylomastoid foramen, recording the compound muscle action potential on the face in a way designed to span several muscles (a method that can lead to unpredictable summation of the compound muscle action potentials of different muscles), and comparing the amplitude ratio of the affected side to the unaffected side. In many series, only patients with no elicitable response are sent to a neurologist or electromyographer. More sophisticated neurophysiologic studies suggest nerve demyelination in milder cases, with significant axonal damage when the paralysis is severe.

Anatomic studies have suggested that the meatal foramen is the narrowest point along the facial nerve's bony canal; intraoperative neurophysiology has suggested that this is the likely site of damage. This has led to the hypothesis that the nerve swells as a result of an inciting event and becomes entrapped at the meatal foramen. Swelling at this confined point could either produce focal demyelination and conduction block, block axoplasmic transport (resulting in nerve damage), or lead to intranerve pressures exceeding perfusion pressure (resulting in ischemic damage). The highly redundant vascular supply of the facial nerve generally makes widespread ischemia seem unlikely, but such focal ischemia is somewhat more plausible.

Recent work has implicated several specific infections in the etiology of facial nerve palsy. Lyme disease, infection with the tick-borne spirochete *Borrelia burgdorferi*, clearly is associated with facial nerve lesions, which may be bilateral in as many as 25% of such cases. The pathophysiology is unclear. This does not appear to be due to a meningeal process, even though Lyme disease can cause meningitis. Neurophysiologic studies suggest that this is due to a mononeuritis multiplex, although the precise mechanism underlying this is unclear.

One of the most interesting sets of findings in the past decade has focused on the possible association between facial nerve palsy and herpes simplex-1 infection. In one of the best known studies, endoneurial fluid was obtained – at least 15 days after onset of the lesion – at the time of 7th nerve decompression in 14 patients with unrecovering Bell's palsy. Eleven of the 14 were polymerase chain reaction (PCR) positive for HSV-1. Twelve of 13 tested patients had peripheral blood serologic evidence of past HSV-1 infection. Seropositivity and PCR positivity were significantly

more common than among controls. This has led to a number of therapeutic trials of antiviral therapy in Bell's palsy. Apparent response to this has led to a general recommendation for antiviral treatment in Bell's palsy.

The herpes hypothesis is an intriguing one, although gaps remain in our understanding. Methodologically, PCR detection of viral DNA does not necessarily indicate the virus is either viable or pathophysiologically relevant. Herpes in particular is so ubiquitous that the implications of finding it in anybody can only be understood if comparable samples are obtained from normal controls. The presence of virus so many days after the onset of symptoms is curious. In other herpes reactivation syndromes, the number of obtainable virions – weeks after the onset – tends to be small. Herpes viruses generally reside in sensory neurons. Although the facial nerve is primarily motor, it does contain sensory neurons (residing in the geniculate ganglion). It is conceivable that reactivation from the sensory neurons could result in sufficient swelling to damage motor fibers. If this were the case, though, one might expect more marked loss of sensation as part of Bell's palsy.

Diagnostic Approach

In addition to clinical assessment, electromyography (EMG) can be informative. When the lesion is incomplete, blink reflexes can demonstrate slowing indicative of demyelination. Since the lesion is typically intracranial, the nerve portion peripheral to the stylomastoid foramen may remain excitable for several days following onset, regardless of the severity of the damage to the nerve. Inexcitability after day 5 suggests a profound axonal lesion. Needle EMG can be useful in differentiating between a truly complete lesion and one sparing a few motor units, the latter having a significantly better prognosis. Observation of widespread changes of acute denervation obviously reflects a more severe lesion with a worse prognosis. On the other hand, evidence of reinnervation may be apparent on needle EMG exam long before it is clinically evident.

EMG can also be useful to demonstrate synkinesias – regenerating nerves can be misdirected, resulting in eye closure with attempted mouth movements, eye tearing with attempted facial movements ('crocodile tears'), or other peculiar combinations.

Cerebrospinal fluid (CSF) examination is useful primarily in assessing for inflammatory or neoplastic processes in the subarachnoid space. Magnetic resonance imaging (MRI) and computed tomography (CT) imaging can be helpful in addressing structural lesions of the brain stem, skull base, or nerve.

Treatment

The prognosis is generally excellent: 85% of all patients will do well with no treatment whatever. This, coupled with methodological concerns with many of the treatment trials, has led to a rather conservative approach to treatment on the part of many physicians, particularly for nonhypertensive patients with incomplete paralysis and no diffuse pain, for whom rapid and complete recovery is almost assured.

Several aspects of care are mandatory. Eye damage is one of the greatest threats to patients. They should be instructed to patch the eye, particularly when sleeping or when in a dusty environment. Lubricating drops (methylcellulose) during the day, and ointment at night, can be helpful. Weakness of the facial muscles leads to a tendency for food to 'pocket' in the cheek. Brushing the teeth after each meal can limit the risk of dental caries. Patients with complete paralysis can develop contracturing of the denervated muscles, leading to additional facial distortion and potentially limiting recovery when reinnervation occurs. Gentle massage of the paralyzed muscles for at least 10–15 min daily is said to lessen the impact of this.

For patients in whom no recovery occurs, several surgical options have been tried. Transplanting the hypoglossal nerve to facial muscles can lead to some facial muscle control on the paralyzed side (at the cost of lingual hemiatrophy and paralysis). A weight in the affected eyelid can facilitate eyelid closure and eye protection. Occasionally the eyelid must be partially sewn shut to protect the cornea.

The three areas of greatest therapeutic focus have been corticosteroids, antivirals such as acyclovir, and decompressive surgery. Unfortunately, the literature on all three approaches has been highly variable in scientific design, sample size adequacy, and interpretation. Results for each have varied from negative to compelling. The literature was recently explored in depth by a task force of the American Academy of Neurology, resulting in a practice parameter. In their meta-analysis of the available, interpretable literature regarding corticosteroids, the panel concluded that corticosteroids were useful, improving outcomes by 16%. Usual regimens consisted of 1 mg kg^{-1} per day (maximum dose 80 mg) divided into two doses per day for 4 or 5 days, then tapering over 5 to 10 days. Treated patients did not recover more quickly, but a greater percentage recovered fully. Some studies have suggested the acute retroauricular pain improves more rapidly with steroids. Patients faired best if treated in the first week. Although it is often stated that the earlier the treatment, the better the response, beyond the observation that treatment is only useful in the first week there are no compelling data that the

timing of treatment within this window affects the likelihood of recovery – a somewhat surprising observation given that the major nerve damage is thought to occur in the first 48 h. Notably, a Cochrane analysis questioned the validity of the conclusion that steroids are effective, because of methodological problems with several of the largest included studies.

Even fewer studies are available concerning use of acyclovir. Here, too, a meta-analysis led to the conclusion that treating all patients with acyclovir improved the proportion of patients recovering by 22%. Given the small number of studies available, and their sometimes conflicting conclusions, this was felt to be less compelling, but was recommended as appropriate treatment. Again, the fact that this treatment is usually given – and appears to be effective – at a time that maximal nerve damage has already occurred is theoretically puzzling.

Finally, with regard to decompressive surgery, the task force found no scientifically compelling study. While earlier studies were negative, one more recent study of patients in whom recovery was thought to be extremely unlikely by clinical and EMG criteria, who underwent a far more extensive decompression extending to the meatal foramen, provided encouraging results, with 91% of patients improving significantly (but with no controls).

Overall, the likelihood of recovery is great, apparently enhanced with early treatment with corticosteroids and/or acyclovir-like antivirals. In the appropriate setting, potentially contributing disorders such as Lyme disease, HIV, or sarcoidosis need to be considered, and treated appropriately if found Surgical treatment remains unproved, but is occasionally attempted in situations in which the likelihood of recovery seems very limited. In general, prevention of secondary complications, such as exposure keratitis, dental decay, and muscle contractures, is the most important aspect of treatment.

See also: Axonal Injury: Neuronal Responses; Brainstem and Cranial Nerves; Electromyography (EMG) and Nerve Conduction Studies; Optic Nerve, Optic Chiasm and Optic Tracts; Peripheral Nerve Regeneration: An Overview.

Further Reading

Adour KK, Ruboyianes JM, Von Doersten PG, et al. (1996) Bell's Palsy treatment with acyclovir and prednisone compared with prednisone alone: A double blind, randomized, controlled trial. *Annals of Otology, Rhinology and Laryngology* 105: 371–378.

Adour KK, Wingerd J, Bell DN, et al. (1972) Prednisone treatment for idiopathic facial paralysis. *New England Journal of Medicine* 287: 1268–1272.

Gantz BJ, Rubinstein JT, Gidley P, et al. (1999) Surgical management of Bell's palsy. *Laryngoscope* 109: 1177–1188.

Gilden DH (2004) Bell's palsy. *New England Journal of Medicine* 351: 1323–1331.

Grogan PM and Gronseth GS (2001) Practice parameter: Steroids, acyclovir, and surgery for Bell's palsy (an evidence based review). *Neurology* 56: 830–836.

Halperin JJ (2003) Facial nerve palsy associated with Lyme disease. *Muscle & Nerve* 28: 516–517.

House JW and Brackmann DE (1985) Facial nerve grading system. *Otalaryngology – Head and Neck Surgery* 93: 146–147.

Katusic SK, Beard CM, Widerholt WC, et al. (1986) Incidence, clinical features and prognosis in Bell's palsy, Rochester, Minnesota 1968–1982. *Annals of Neurology* 20: 622–627.

Murakami S, Mizobuchi M, Nakashiro Y, et al. (1996) Bell palsy and herpes simplex virus: Identification of viral DNA in endoneurial fluid and muscle. *Annals of Internal Medicine* 124: 27–30.

Salinas R, Alvarez G, and Ferreira J (2004) Corticosteroids for Bell's palsy (idiopathic facial paralysis). *Cochrane Database of Systematic Reviews* 4: CD001942.

Bergmann Glial Cells

A Verkhratsky, The University of Manchester, Manchester, UK
A Reichenbach, University of Leipzig, Leipzig, Germany

© 2009 Elsevier Ltd. All rights reserved.

The cerebellum of all vertebrates contains specialized radial astrocytes, which send long, vertically oriented processes toward the pia. These glial processes were discovered by Karl Bergmann in 1857 as cerebellar radial fibers. In 1871, Camillo Golgi recognized these structures as cells and found their somata/cell bodies, from where the radial fibers originated. Both Golgi and Ramon-y-Cajal referred to these cells as epithelial cells, and for many years they were generally known as Golgi epithelial cells. In recent times, the common name for these cerebellar radial astrocytes is Bergmann glial cells.

Morphology

The somata of Bergmann glial cells are localized in the Purkinje cell layer (Figure 1). Each Bergmann glial cell has three to six stem processes, which protrude through the molecular layer toward the pial surface. These processes contain many mitochondria and bundles of intermediate filaments, constituted by vimentin and glial fibrillary acidic protein (GFAP). They can be labeled by antibodies directed against these proteins, which reveals that they are organized in palisades, parallel to the long axis of each folium. On average, every Bergmann glial cell in rodent cerebellum has a volume of approximately 3600 μm^3, and altogether these cells occupy approximately 15–18% of the volume of the molecular layer. There are approximately eight Bergman glial cells per every Purkinje neuron; the processes of the former develop intimate relations with the dendritic arborization of the latter.

The stem processes of Bergmann glial cells end at the pia mater, with so-called endfeet, which together form a complete covering of the cerebellar surface. The subpial endfoot membranes are endowed with orthogonal arrays of membrane particles, probably representing locally enriched membrane proteins such as the water pore channel, aquaporin-4, and related channel and transporter proteins devoted to an active exchange of ions and water between the neuropile and the subarachnoidal space. On their way from soma to the pia, the stem processes extend numerous side branches, which form close contacts with synapses present on the dendrites of Purkinje neurons; on average, every Bergmann glial cell covers 2000–6000 such synapses. Morphologically, these gliosynaptic contacts are quite peculiar because they are organized in the form of morphofunctional microdomains (Figures 2 and 3). The microdomains are constructed by protrusions sent from the stem process at rather regular intervals. Each microdomain is made of a thin stalk and a complex 'head' structure, which in turn forms a lamellar perisynaptic sheath for approximately five synapses. The stalk is extremely thin (~0.3 μm) and long (up to 7 μm); the head is a complex structure with an extremely large surface area (~300 μm^2) formed by a relatively small (~16 μm^3) volume. The surfaces of the appendages may form the major part of Bergmann glial plasmalemma (overall, the surface-to-volume ratio for Bergmann glial cell varies between 13 and 17 μm^{-1}, whereas the same ratio for appendage may reach 25 μm^{-1}). The synaptic–glial interactions, therefore, are very much confined to these single microdomains; and these microdomains can be independently excited. Each microdomain also contains on average two mitochondria, which may provide a local energy supply and also serve as a barrier preventing Ca^{2+} diffusion. The coverage of synapses by glial membranes is quite specific because glial membrane tongues almost completely seal the synaptic clefts while leaving a considerable part of extrasynaptic structures uncovered.

Physiology

Membrane Conductance

The resting membrane potential of mature Bergmann glial cells is between −80 and −90 mV, and the resting input resistance is quite low (40–100 mΩ). Potassium channels are responsible for the major part of membrane permeability of Bergmann glial cells. The expression of these channels is developmentally regulated. In the early postnatal (< postnatal day 7 (P7)) period, Bergmann glial cells demonstrate a mixture of inward- and outward-rectifying potassium currents, which can be blocked by cesium, barium, and 4-aminopyrindine. Mature Bergmann glial cells (> P20) display a large passive K^+ conductance manifested by voltage- and time-independent K^+ currents, insensitive to Cs^+, Ba^+, or 4-AP. In addition, these cells have a complement of inward- and outward-rectifying K^+ channels but their density is quite low. There were some reports that Bergmann glial cells express mRNA for the voltage-gated Na^+ channel $Na_v1.6$, but functional channels were never detected electrophysiologically.

Gap Junctions

Similar to other astrocytes, Bergmann glial cells are coupled by gap junctions. The main gap junction

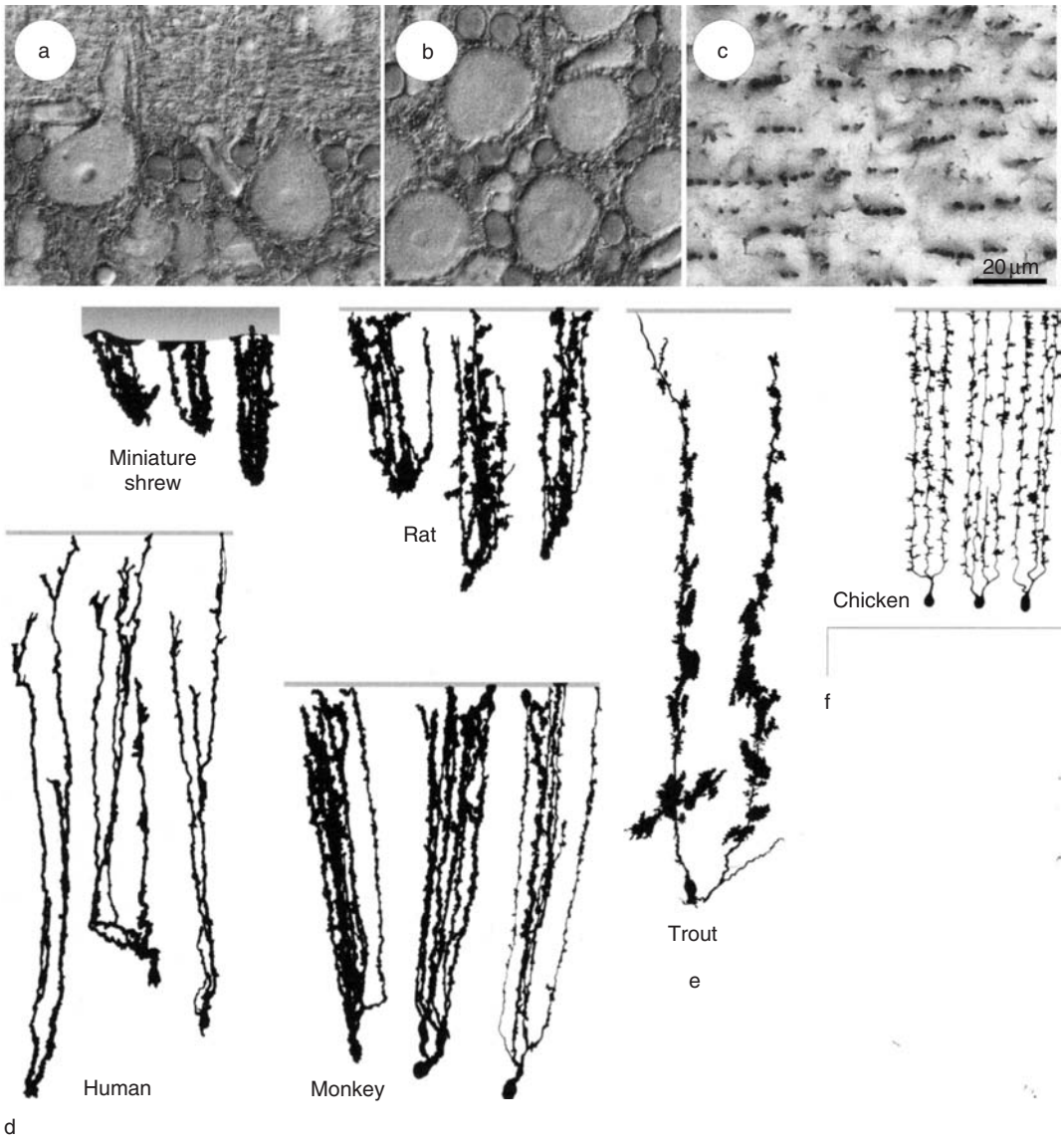

Figure 1 Morphology of Bergmann glial cells in the vertebrate cerebellum. (a–c) Bergmann glial cells of the adult rat cerebellum, immunocytochemistry. The glutamine synthetase-positive Bergmann glial cell somata (dark) surround and envelop the Purkinje cell somata (a and b), while their glial fibrillary acidic protein (GFAP)-positive stem processes run through the molecular layer where they form into rows or palisades (view from the cerebellar surface; c). (d–f) Camera lucida drawings of Golgi-stained Bergmann glial cells in the cerebella of several different mammals, including those of human (d), the trout *Salmo gairdneri* (e), and the chicken (f). Adapted from Reichenbach A and Wolburg H (2005) Astrocytes and ependymal glia. In: Kettenmann H and Ransom BR (eds.) *Neuroglia*, pp. 19–35. Oxford: Oxford University Press.

protein expressed by Bergmann glial cells is connexin-43x. The gap junction coupling (as was judged by both electrophysiological recordings and dye coupling; Figure 4) between Bergmann glial cells begins to appear at approximately P7 and is fully developed by P20. The gap junctions are preferentially localized in the distal processes of Bergmann glial cells. Dye spread between Bergmann glial cells in cerebellar slices revealed a peculiar organization of the glial syncytium: dye-coupled cells were always arranged in rows perpendicular to parallel fibers (Figure 4). This is very similar to the two-dimensional orientation of Purkinje neurons and, therefore, both neuronal and glial elements from cellular stacks, which may also have functional impact. There are also indications for the existence of gap junction coupling between Bergmann glial cells and adjacent Purkinje neurons.

Neurotransmitter Receptors

Glutamate receptors The main type of ionotropic glutamate receptors expressed by Bergmann glial cells is of the α-amino-3-hydroxy-5-methyl-4-isoxazole

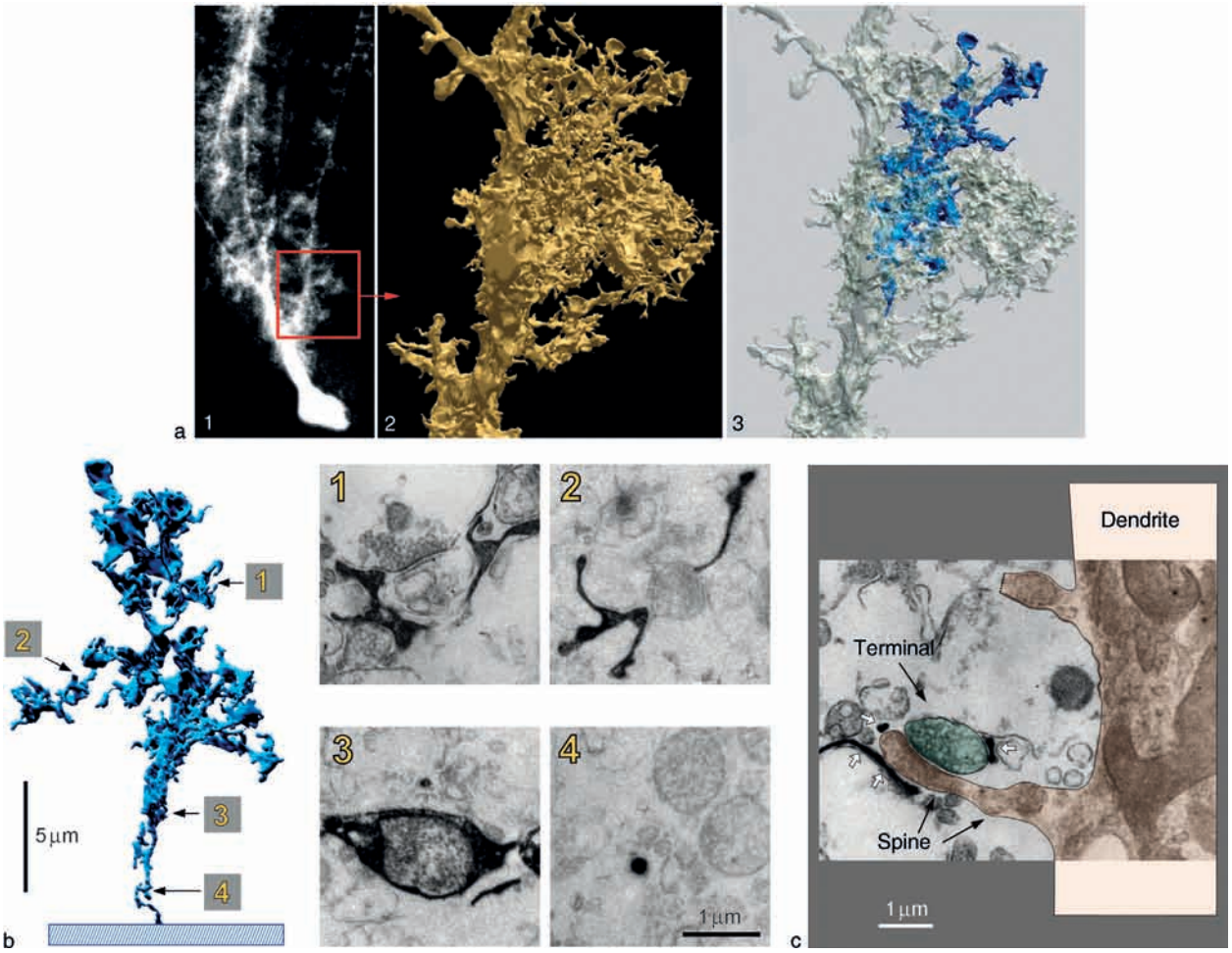

Figure 2 Bergmann glial microdomains. (a) Reconstruction of an appendage based on electron microscopic data. (1) Fluorescence light micrograph of a dye-injected Bergmann glial cell is shown; the red square (20 × 20 mm) corresponds to the portion that was reconstructed from consecutive ultrathin sections. (2) One of the lateral appendages, arising directly from fiber with all the other side branches omitted for clarity. (3) The same appendage as shown in 2, but with one of the appendages marked by blue. This labeled structure is shown in isolation and at higher magnification in (b). (b) Fine structure of appendages and relation to synapses. (Left) A small lateral appendage, arising from the reconstructed part of the glial fiber (blue in a3), is shown as a slightly turned, isolated three-dimensional reconstruction. Electron micrographs of four sections contributing to the reconstruction (designated 1–4) are shown on the right; glial compartments appear black after conversion of the injected dye. The location of these sections in the reconstruction is indicated by the labeled arrows. (1) Region directly contacting synapses. (2) Glial compartments without direct synaptic contacts. (3) Bulging glial structure containing a mitochondrion. (4) The stalk of the appendage. (c) Another example of a synapse contacted by glial compartments, appearing as black structures (white arrows); in this case, the postsynaptic element can be traced back along the spine (black arrows) to the Purkinje cell dendrite (reddish overlay; the presynaptic terminal is labeled by a green overlay). Adapted from Grosche J, Matyash V, Moller T, Verkhratsky A, Reichenbach A, and Kettenmann H (1999) Microdomains for neuron–glia interaction: Parallel fiber signaling to Bergmann glial cells. *Nature Neuroscience* 2: 139–143.

propionic acid (AMPA) subtype. Bergmann glial cells do not possess the glutamate receptor (GluR)2 subunit, making their AMPA receptors Ca^{2+} permeable. Indeed, activation of AMPA receptors by exogenous kainate triggers substantial Ca^{2+} rises in glial cells (Figure 5). There are indications that Bergmann glial cells may express mRNA for the NR2B subunit of *N*-methyl-D-aspartate (NMDA) receptors; application of endogenous NMDA was reported to trigger $[Ca^{2+}]_i$ elevation and small inward currents. These responses, however, were often blocked by tetrodotoxin, and thus most likely resulted from stimulation of neurons.

Bergmann glial cells express metabotropic glutamate receptors of the mGluR5 type; stimulation of these receptors triggers an inositol 1,4,5-triphosphate ($InsP_3$)-mediated release of Ca^{2+} from the endoplasmic reticulum (ER) intracellular Ca^{2+} stores. This mGluR5-dependent Ca^{2+} mobilization assumes the leading role in the generation of glial Ca^{2+} signals following

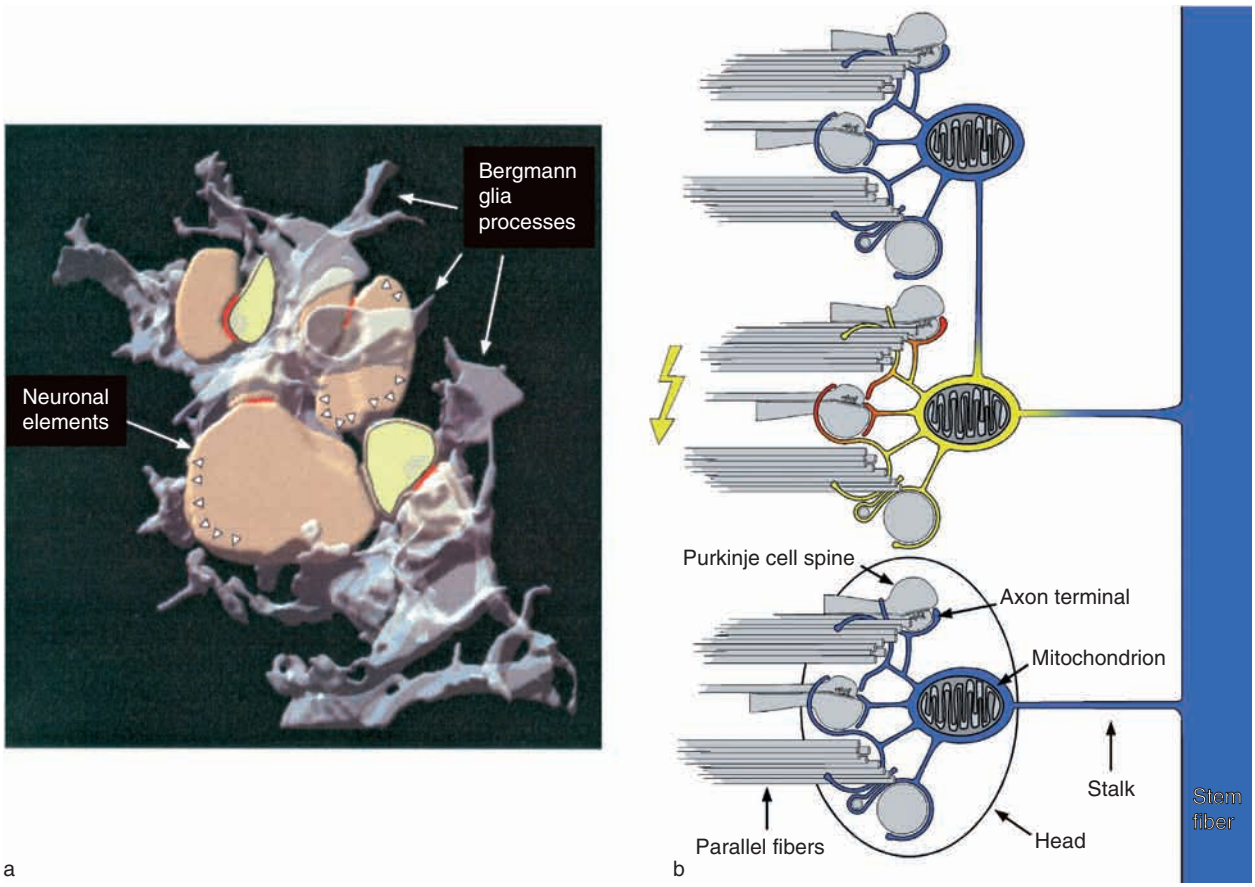

Figure 3 Glial microdomains form a substrate for glial–synaptic interactions. (a) Three-dimensional reconstruction of a group of neighboring cerebellar synapses (yellow; synaptic clefts: orange) together with the surrounding leaflets provided by the labeled Bergmann glial cell (blue-green). Arrowheads point to neuronal surfaces not covered by glial sheaths from the labeled cell. (b) Schematic representation of a Bergmann glial microdomain. The basic components of the microdomain, the stalk and the 'head,' are shown together with their relationships to the neighboring neuronal elements. Stimulation of several closely positioned parallel fibers may activate a single microdomain, inducing both membrane currents and local Ca^{2+} signals. Adapted from Grosche J, Kettenmann H, and Reichenbach A (2002) Bergmann glial cells form distinct morphological structures to interact with cerebellar neurons. *Journal of Neuroscience Research* 68: 138–149.

stimulation with glutamate. The contribution of the AMPAR-mediated Ca^{2+} influx is much smaller because of rapid desensitization of these receptors in the presence of glutamate.

Nonetheless, the AMPA receptor-mediated Ca^{2+} entry is functionally important because adrenoviral-assisted genetic introduction of the GluR2 subunit into Bergmann glial cells extinguished the Ca^{2+} permeability of these receptors and resulted in a withdrawal of glial processes from the synapses formed on Purkinje neurons. This, in turn, affected synaptic transmission by slowing the glutamate uptake, and it promoted the reinnervation of Purkinje neurons by several climbing fibers.

GABA receptors $GABA_A$ ionotropic receptors are prominently expressed in young Bergmann glial cells; application of γ-aminobutyric acid (GABA) triggers large Cl^- currents in cells from P5 to P12 animals. Immunohistochemical techniques revealed a prominent expression of α_2, α_3, and γ subunits in these cells. The $GABA_A$ receptors are predominantly located in the distal processes of the Bergmann glial cells. The expression of $GABA_A$ receptors is downregulated during postnatal development; by P20, the amplitudes of GABA-induced currents are very much diminished and only α_2 and γ_1 subunits remain detectable immunohistochemically. The $GABA_A$ receptors in mature Bergmann glial cells are primarily concentrated in the parts of glial membrane ensheathing inhibitory GABAergic synapses formed on Purkinje neurons. In addition to $GABA_A$ receptors, immunoreactivity for $GABA_{B1}$ receptors was detected in Bergmann glial cells; there are also some indications for the possible expression of $GABA_C$Rs. The functional significance of these receptor types remains to be explored.

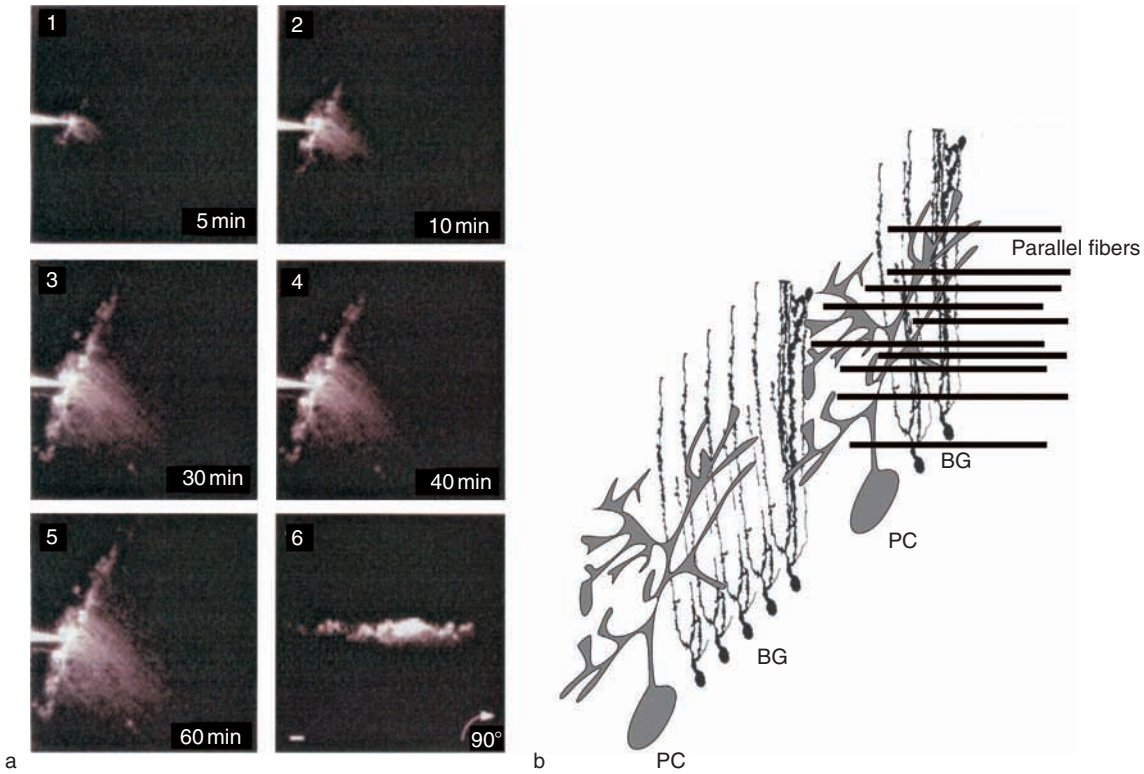

Figure 4 Gap junction coupling of Bergmann glial cells. (a) Dye coupling. Bergmann glial cell was dialyzed via the patch pipette with a solution containing Lucifer yellow. Fluorescence images were recorded a *t* times indicated times after the whole cell configuration was established using confocal microscopy. Images were three-dimensional reconstructed and illustrate the spread of dye to adjacent, coupled cells. The number of coupled cells are as follows: 1, $n = 4$; 2, $n = 8$; 3, $n = 24$; 4, $n = 26$; 5, $n = 26$. (a6) The image shown in a5 is rotated to the parasagittal orientation. The upper 25 pm of the 120 pm thick cerebellar slice is displayed. (b) Schematic image showing orientation of the coupled Bergmann glia cells. In the parasagittal sections used, the array of coupled cells extends parallel to the slice surface and perpendicular to the parallel fibers. BG, Bergmann glia cells; PC, Purkinje cell. Adapted from Muller T, Moller T, Neuhaus J, and Kettenmann H (1996) Electrical coupling among Bergmann glial cells and its modulation by glutamate receptor activation. *Glia* 17: 274–284.

Purinoreceptors Bergmann glial cells respond to extracellular application of ATP by robust cytosolic Ca^{2+} rises, which are initiated in the processes and then spread toward the soma (Figure 6). These ATP-induced Ca^{2+} responses are mediated by P2Y metabotropic purinoreceptors and subsequent $InsP_3$-induced Ca^{2+} release from the ER Ca^{2+} stores. This conclusion is based on the following observations: (1) ATP-induced Ca^{2+} signals are preserved in Ca^{2+}-free extracellular solutions; (2) applications of ATP do not trigger any measurable membrane currents; and (3) ATP-induced Ca^{2+} signaling is inhibited either by the ER calcium pump blocker, thapsigargin, or by intracellular administration of the competitive $InsP_3$ receptors antagonist, heparin. Incidentally, the P2Y purinoreceptors in Bergmann glial cells can be activated by extracellular administration of NAADP.

Adrenoreceptors Application of endogenous adrenomimetics (epinephrine and norepinephrine) induces cytosolic Ca^{2+} elevations in Bergmann glial cells. These Ca^{2+} responses are of purely metabotropic (G-protein/phospholipase C (PLC)/$InsP_3$-mediated) nature; they are unaffected by Ca^{2+} removal from extracellular milieu and blocked by thapsigargin or intracellularly administered heparin. The monoamine-induced Ca^{2+} signaling is mediated through α_1 adrenoreceptors; the Ca^{2+} responses can be mimicked by the specific α_1 adrenoreceptor agonist, phenylephrine, but not by β-adrenomimetic isoproterenol or α_2-adrenomimetic clonidine. Moreover, the monoamine-evoked Ca^{2+} signals can be almost completely blocked by the α_1 adrenoreceptor antagonist prazozin; however, they are fully resistant to the action of the α_2 adrenoreceptor antagonist yohimbine or the β adrenoreceptor antagonist propranolol.

Histamine receptors Bergmann glial cells express functional H_1 histamine receptors. Stimulation of these receptors triggers $InsP_3$-induced Ca^{2+} release from the ER. Histamine-induced Ca^{2+} signaling is

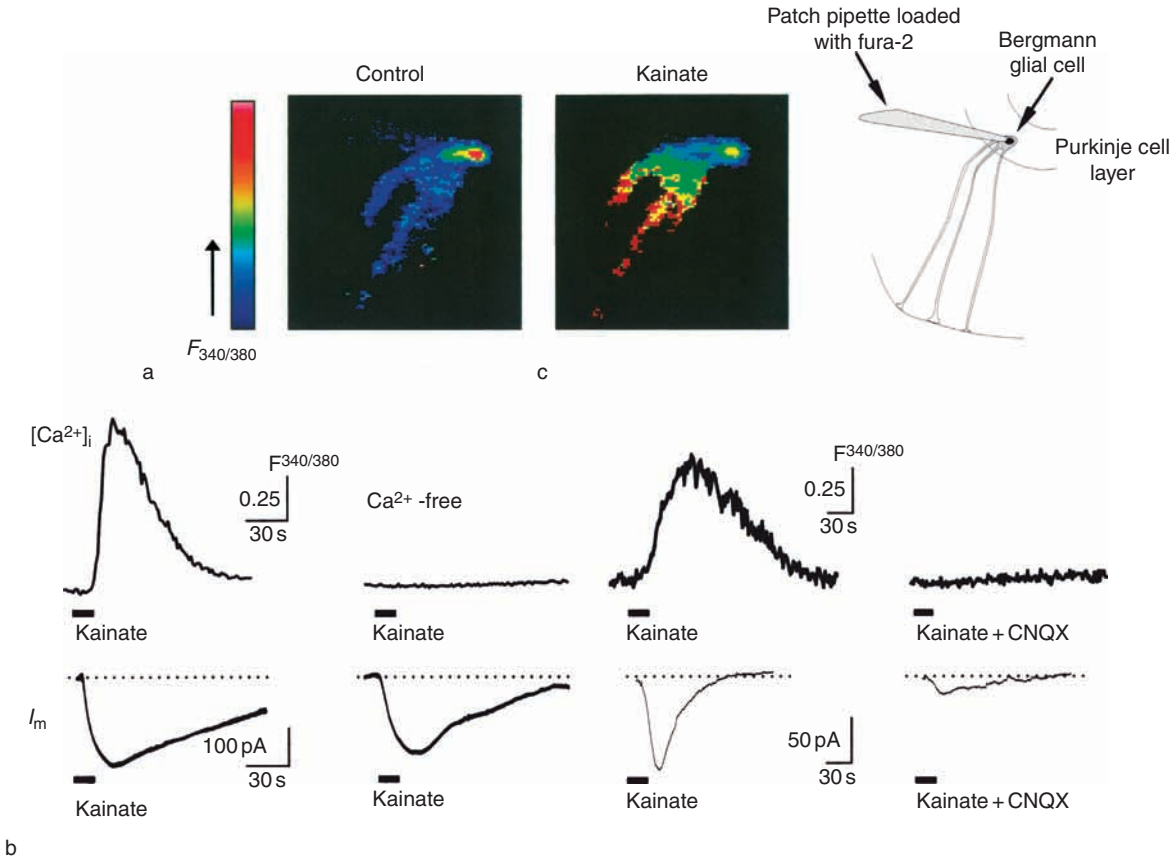

Figure 5 Stimulation of ionotropic glutamate receptors (GluRs) in Bergmann glial cell in cerebellar slices triggers Ca^{2+} influx. (a) Pseudocolor image of a fura-2-loaded Bergmann glial cell (a scheme of the experiment is shown on the right) in control conditions and upon extracellular application of 100 μM kainate. Note preferential increase in $[Ca^{2+}]_i$ in cell processes. (b, c) $[Ca^{2+}]_i$ and membrane current traces recorded from the same cell. Both removal of extracellular Ca^{2+} (b) and blockade of ionotropic GluRs by 6-cyano-7 nitroquinoxaline-2,3-dione (CNQX; 10 μM; c) inhibited kainate-induced currents and $[Ca^{2+}]_i$ elevation, suggesting a key role for Ca^{2+} influx. F340/380, fluorescence ratio at 340/380 nm. Adapted from Verkhratsky A, Orkand RK, and Kettenmann H (1998) Glial calcium: Homeostasis and signaling function. *Physiological Reviews* 78: 99–141.

very substantially inhibited by the selective H_1 antagonist chlorpheniramine.

Endothelin receptors All three forms of endothelin, ET-1 to ET-3, when applied extracellularly trigger Ca^{2+} signals in Bergmann glial cells. The receptors responsible are selectively sensitive to the ET_B receptor agonist BQ3020 and are blocked by the ER_B receptor antagonist BQ-788; the expression of ET_B receptors was further confirmed by single-cell polymerase chain reaction. Intracellularly, the stimulation of ET_B receptors results in activation of the PLC/$InsP_3$/Ca^{2+} release cascade.

Bergmann glial cells *in situ* express a defined set of neurotransmitter receptors Studies on Bergmann glial cells *in situ* (i.e., in cerebellar slices) revealed that they are always endowed with a strictly defined pattern of functional neurotransmitter receptors: all cells were sensitive to glutamate, ATP, GABA, monoamines, histamine, and endothelin. By contrast, many other neuroactive substances known to activate cultured astrocytes (e.g., bradykinin, oxytocin, substance P, adenosine, vasopressin, or PAF) did not affect the cells. Most interestingly, the sensitivity of Bergmann glial cells to neurotransmitters is very similar to that of its closest neighbor, the Purkinje neuron: in general, the modalities of receptors expressed by these two cell types are identical. Furthermore, neurotransmitter receptors expressed by the neuron and by the glial cell specifically match the neurotransmitters released in their vicinity (Figure 7), thus allowing them to sense the very same incoming information.

Neurotransmitter Transporters

Glutamate transporters Membranes of Bergmann glial cells, surrounding the excitatory glutamatergic synapses formed on the dendrites of Purkinje neurons, contain the glutamate transporter GLAST (glutamate aspartate transporter) and GLT-1 (the former being the

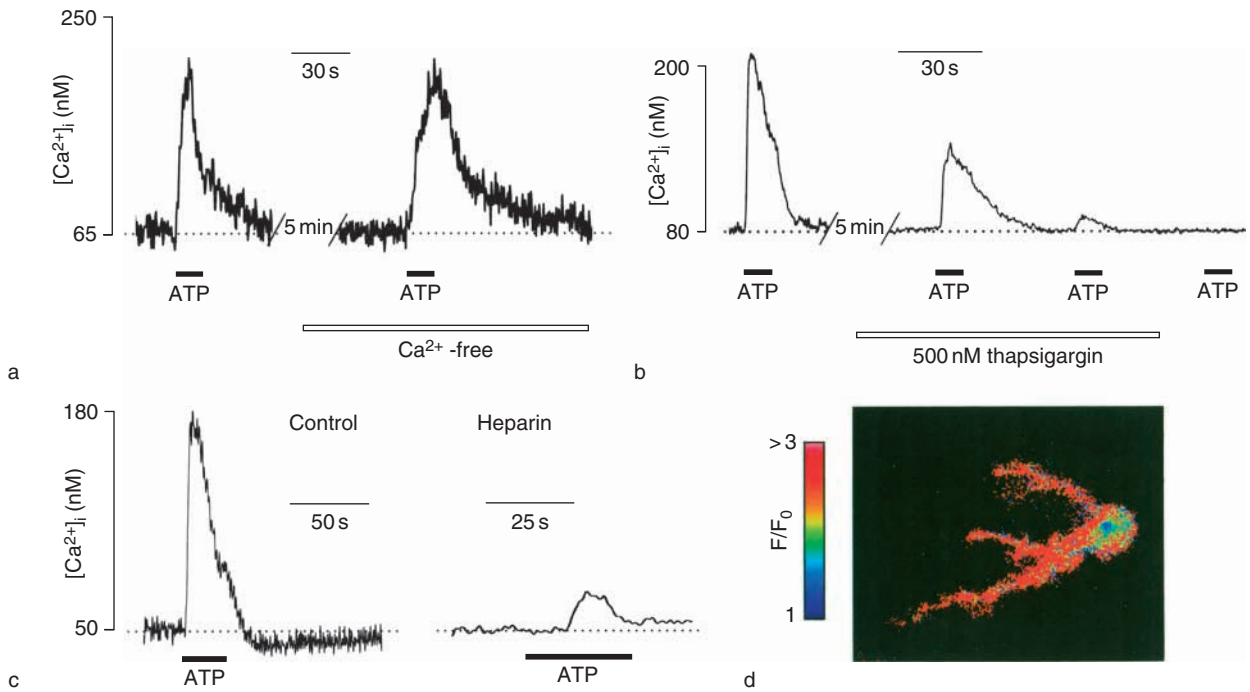

Figure 6 ATP-induced Ca^{2+} signaling in Bergmann glial cells results exclusively from inositol 1,4,5-trisphosphate ($InsP_3$)-mediated Ca^{2+} release from ER Ca^{2+} stores. (a) ATP-induced $[Ca^{2+}]_i$ transients were measured from 'bulk-loaded' Bergmann glial cells stained by incubating cerebellar slices in fura-2 acetoxymethyl ester (AM)-containing solutions. Addition of ATP triggered an increase in $[Ca^{2+}]_i$ that persisted in Ca^{2+}-free extracellular solution. (b) In a similar experiment, incubation of slice with 500 nM thapsigargin completely and irreversibly blocked ATP-induced Ca^{2+} signaling. (c) Intracellular administration of heparin via intracellular dialysis with a patch pipette inhibited $[Ca^{2+}]_i$ increase induced by ATP. Control $[Ca^{2+}]_i$ transient was recorded from fura-2 AM-loaded cells before commencing intracellular dialysis. (d) Illustration of spatial distribution of $[Ca^{2+}]_i$ at the time of maximum ATP response. Note the higher levels of $[Ca^{2+}]_i$ in Bergmann glial cell processes compared with the cell body. Adapted from Kirischuk S, Moller T, Voitenko N, Kettenmann H, and Verkhratsky A (1995) ATP-induced cytoplasmic calcium mobilization in Bergmann glial cells. *Journal of Neuroscience* 15: 7861–7871; and Verkhratsky A, Orkand RK, and Kettenmann H (1998) Glial calcium: Homeostasis and signaling function. *Physiological Reviews* 78: 99–141.

dominant type), whereas the extrasynaptic regions of the neuronal (i.e., Purkinje cell) spines are rich in neuronal excitatory amino acid transporters of the EAAT4 type. The glial and neuronal transporters play distinct roles in controlling the glutamate concentration within and around the synaptic cleft. The affinity of GLAST to glutamate is approximately 10 times lower than that of excitatory amino acid transporter 4 (EAAT4); therefore, GLAST is primarily responsible for rapid uptake of glutamate following its exocytosis. This uptake determines the time course of synaptic events and prevents an escape of high glutamate concentrations from the cleft. Inhibition of GLAST significantly increases the early spillover of glutamate, as judged, for example, by the activation of NMDA receptors on GABAergic terminals, with a subsequent increase in the strength of inhibitory input to Purkinje neurons. In this way, glutamate uptake by Bergmann glial cells can regulate the synaptic inhibition of Purkinje neurons. Furthermore, inhibition of glial glutamate transporters also affects background synaptic transmission to Purkinje neurons by increasing the frequency of spontaneous miniature excitatory postsynaptic currents (EPSCs). The EAAT4, in contrast, buffers low glutamate concentrations and prevents the delayed spillover of the neurotransmitter to neighboring synapses. According to various estimates, approximately 80% of glutamate released during physiological neurotransmission is accumulated by Bergmann glial cells and only 20% is removed by the neuronal postsynaptic compartment.

To translocate glutamate against the concentration gradient, both neuronal and glial transporters utilize transmembrane gradients of Na^+ and K^+. Uptake of a single glutamate molecule requires influx of three Na^+ ions and efflux of one K^+ ion; in addition, glutamate brings one more H^+ ion into the cell (the net cation influx underlies the electrogenic role of transporters). Therefore, the maintenance of transmembrane ion gradients is essential for an effective performance of glutamate transporters. Exposure of Bergman glial cells to glutamate triggers substantial increases in the cytoplasmic Na^+ concentration, which may impair further glutamate uptake. This

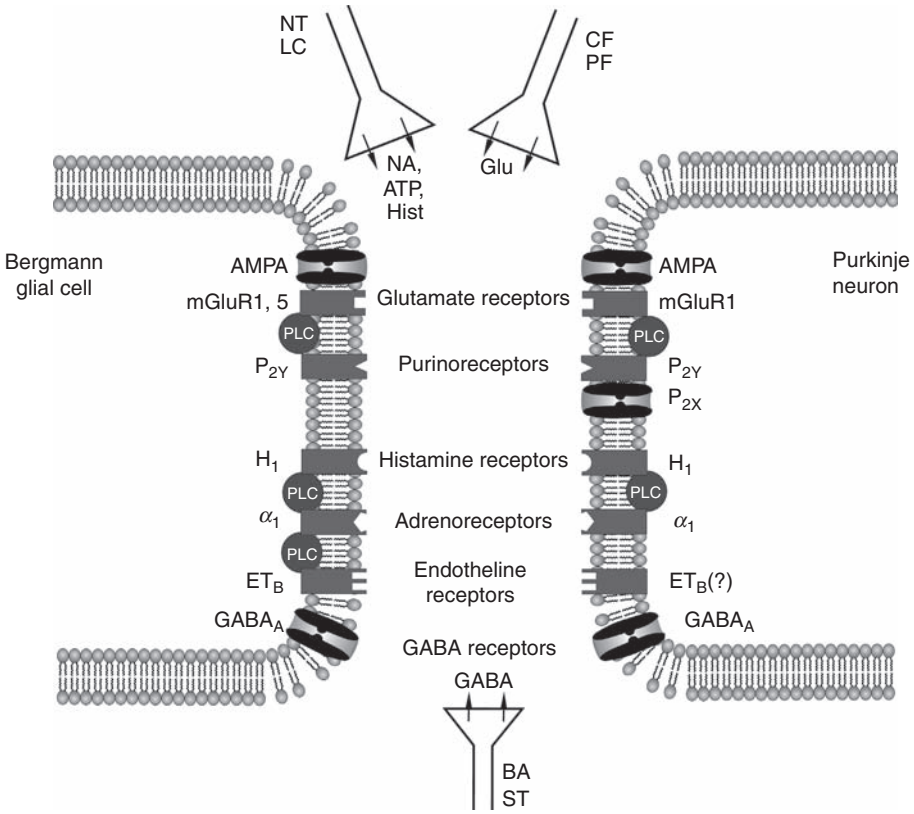

Figure 7 Bergmann glial cell and its neighbor Purkinje neuron bear a similar set of neurotransmitter receptors. NT, terminals from tuberomammillary nucleus of the posterior hypothalamus that carry histamine innervation of the cerebellar cortex; LC, terminals from locus coeruleus that utilize noradrenalin and ATP as neurotransmitters; CF and PF, climbing and parallel fibers, respectively, major neurotransmitter glutamate; BA and ST, basket and stellate cells that deliver GABA to the Purkinje neuron layer. Adapted from Verkhratsky A, Orkand RK, and Kettenmann H (1998) Glial calcium: Homeostasis and signaling function. *Physiological Reviews* 78: 99–141.

$[Na^+]_i$ increase can be counteracted by a sodium–calcium exchanger, expressed in the membrane of Bergmann glial cells; sharp cytosolic Na^+ rises turn the sodium–calcium exchanger into the reverse mode (when Na^+ is expelled in exchange for Ca^{2+}). This process may rapidly lower $[Na^+]_i$ and thus sustain the operation of the glutamate transporter. Under pathological conditions that are accompanied by cell depolarization and accumulation of glutamate and Na^+ ions in the glia, the glutamate transporter may also reverse and start to extrude glutamate from the Bergman glial cells, thus exacerbating the glutamate-dependent excitotoxic neuronal death. In addition, Bergmann glial cells express functional cysteine/glutamate transporters, which also may provide a pathway for glutamate extrusion during periods of ischemia.

GABA transporters Bergman glial cells are endowed with a specific GABA uptake system, represented by high-affinity GABA/Na^+/Cl^- transporters of the GAT-1 type. The uptake of one GABA molecule is coupled to the inward transport of two Na^+ ions and one Cl^- ion. This transporter is present in the Bergmann glial cell somata and is also concentrated in the processes enwrapping GABAergic synapses formed on Purkinje neurons. In these compartments, glial GABA uptake can be relevant for the neurotransmitter removal during normal synaptic activity. Under experimental conditions, when the Bergmann glial cells were internally perfused with solutions containing 10 mM GABA and 12.5 mM Na^+, cell membrane depolarization to −40 mV resulted in the reversal of the GABA transporter. Whether such a reversed GABA transport may be relevant *in vivo* (even under pathological conditions) remains unclear.

Glycine transporters Bergmann glial cells possess functional GlyT1 transporters that belong to the Na^+/Cl^--dependent transporter family. These transporters can work in both forward (GABA uptake) and reverse (GABA extrusion) modes, and under certain conditions they can execute glycine efflux. However, this requires a depolarization of the Bergmann glial cell to +20 mV, which is very unlikely to occur under physiological conditions.

Taurine transporters Bergmann glial cells express a Na^+/Cl^--dependent taurine transporting system, which is able to function in both forward (taurine uptake) and reverse (taurine release) modes. The reversal of the transporter requires cell depolarization (up to −50 mV) together with an increase in the intracellular Na^+ concentration above 40 mM; these conditions most likely may occur only during severe ischemia. Functionally, the taurine transporter (working in the forward mode) could be involved in taurine redistribution following a hypo-osmotic shock.

Responses of Bergmann Glial Cells to Neuronal Activity

Electrical responses Stimulation of synaptic inputs to Purkinje neurons, formed by parallel and climbing fibers, induces complex membrane current responses in Bergmann glial cells. These current responses are mediated by activation of both ionotropic AMPA-type glutamate receptors and Na^+-dependent glutamate transporters. The glial AMPA receptor-mediated responses show short-term and long-term plasticity. The former is manifested by prominent paired-pulse facilitation, whereas the latter appears as long-term depression of synaptically induced glial responses following low-frequency repetitive stimulation of parallel fibers. These plasticity patterns may allow Bergman glial cells to distinguish between synaptic inputs displaying different activities. In addition, Bergmann glial cells can also be activated by ectopic glutamate release from climbing fibers. This ectopic release predominantly activates glial AMPA receptors, and it may be important for rapid communication between nerve afferents and Bergmann glial cells.

Interestingly, direct electrical stimulation of Bergmann glial cells was reported to affect the 'background' synaptic activity between parallel fibers and Purkinje neurons. Fifty consecutive depolarizations of Bergmann glial cells (0.5 s duration, delivered at 1 Hz) substantially reduced the frequency of spontaneous postsynaptic currents recorded from the neuron.

Cytoplasmic calcium responses In parallel with membrane currents, electrical stimulation of parallel fibers triggers cytosolic calcium responses in the Bergmann glia. These Ca^{2+} responses are precisely localized, being limited to small compartments within glial cell processes (Figure 8). The size of these Ca^{2+} response sites is similar to that of morphologically distinct microdomains, which enwrap a cluster of synapses formed by parallel fibers. This striking compartmentalization of synaptically induced glial Ca^{2+} responses may be important for spatial discrimination of the incoming signals. The Ca^{2+} responses elicited in Bergmann glial cells following nerve activity result from activation of several pathways, including activation of glutamate receptors and α_1 adrenoreceptors – with subsequent Ca^{2+} release from the ER stores – as well as the NO signaling cascade, which activates plasmalemmal Ca^{2+} influx. The local Ca^{2+} signals in the Bergmann glial processes also occur spontaneously, without stimulation of synaptic inputs; most likely, this is the glial correlate of spontaneous EPSCs in nerve cells.

Development and Morphogenesis of Cerebellar Structures

During ontogenesis, Bergmann glial cells arise by transformation (i.e., loss of the ventricular cell process) from fetal radial glial cells, as well as by proliferation of immature Bergmann glial cells within the primitive molecular layer up to the second postnatal week in rat. From the beginning of their existence, Bergmann glial cells are intimately involved in cerebellar morphogenesis. First, in the early postnatal period (P4–P15) Bergmann glial cells act as a guidance system for young postmitotic granule cells, which migrate from the proliferative zone located on the surface of the cerebellar cortex (external granule cell layer) to the inner granule cell layer positioned below the Purkinje cell bodies. This migratory process involves several defined morphological stages described in detail by Ramon-y-Cajal and Pasko Rakic. First, the newborn granule neuron contacts the fiber of a Bergmann glial cell located within the deep part of the external granule layer. After establishing such a contact, the granule neuron turns first into a bipolar cell and then extends a third process that becomes the leading one, guiding the bipolar neuron along the Bergmann glial cell fiber. When this leading process elongates sufficiently, the soma (including the nucleus) of the granule neuron translocates through it, such that finally the entire neuron attains its destination.

Starting from the third postnatal week, Bergman glial cells become closely associated with Purkinje neurons. This association appears to be reciprocal because without contact with Purkinje cells, the Bergmann glia cannot develop its mature phenotype, whereas the development of Purkinje cell dendrites occurs in close alignment with Bergmann glial cell processes. These developmental neuronal–glial interactions essentially involve the Notch signaling system. Membranes of Purkinje cell dendrites express specific Delta/Notch-like EGF-related receptors (DNER), whereas the processes of Bergmann glial cells express Notch. Binding of DNER to Bergmann glial processes determines the morphological differentiation of the latter, and the inhibition of this type

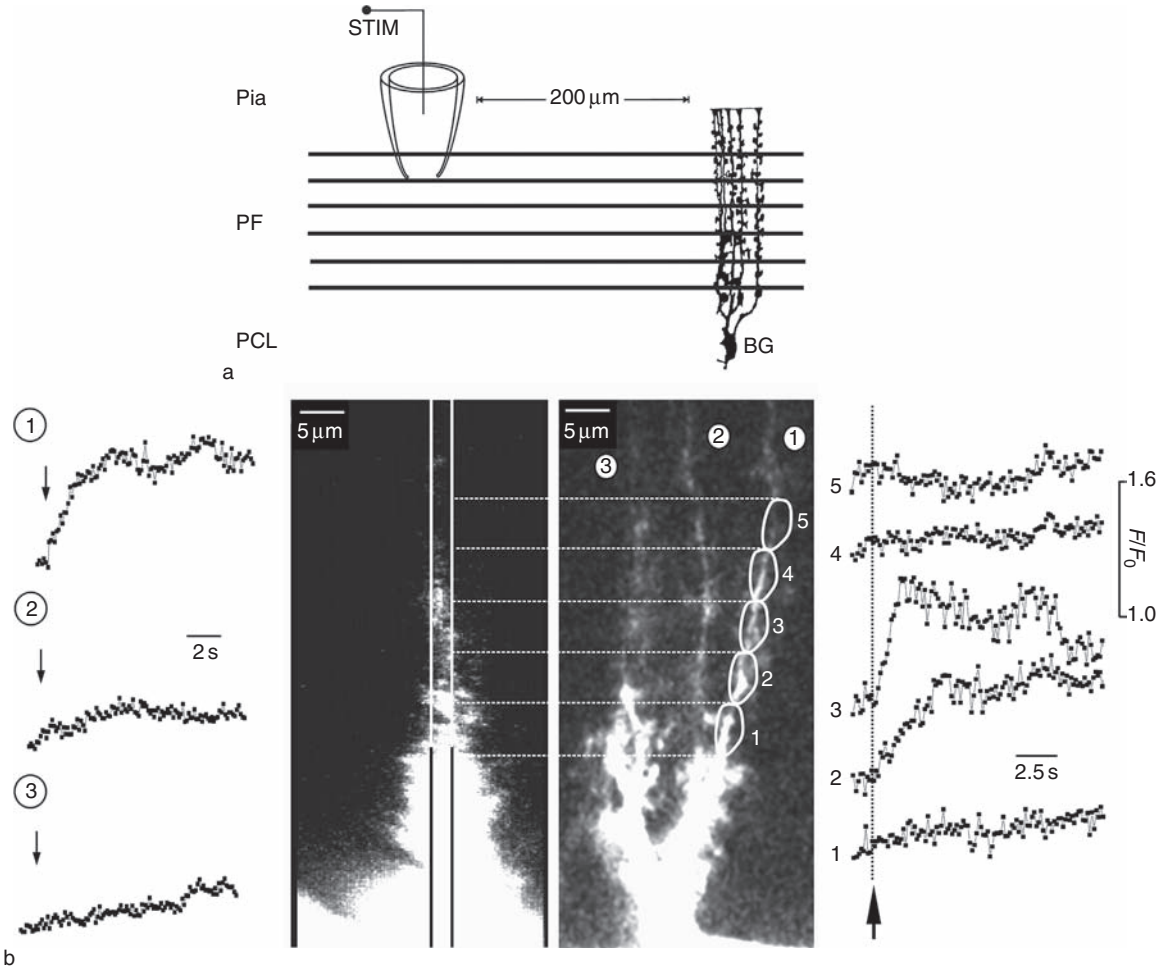

Figure 8 Stimulation of parallel fibers triggers local calcium signals in Bergmann glial cells. (a) Experimental protocol. Parallel fibers (PF) were stimulated via a pipette connected to a stimulator (STIM) while calcium-dependent fluorescence responses were recorded in a Bergmann glial cell (BG). PCL, Purkinje cell layer. (b) Confocal fluorescence intensity image of a patch-clamped Bergmann glial cell dialyzed with the calcium indicator, Oregon green 488 BAPTA-1 (center right). Three processes were distinguished (indicated as 1–3). Calcium signals in response to PF stimulation were measured independently for each process (left). Time of PF stimulation is indicated by arrows. The responding process (1) was then further subdivided into five regions of interest, in which calcium signals were measured separately (right). Time of PF stimulation is indicated by an arrow and a dotted line. The center left image was obtained from a sequence of serial sections and shows the cell in depth (turned by 90° compared to the right image). The lines indicate the focal plane used for $[Ca^{2+}]_i$ recordings. The process is thus within the volume from which recordings were obtained. Adapted from Grosche J, Matyash V, Moller T, Verkhratsky A, Reichenbach A, and Kettenmann H (1999) Microdomains for neuron–glia interaction: Parallel fiber signaling to Bergmann glial cells. *Nature Neuroscience* 2: 139–143.

of local signaling prevents normal morphogenesis of the glial cells. The intermediate filaments in the glial cell processes may also play a crucial role in these glial–neuronal interactions; in vimentin-deficient mice, many of the Purkinje neurons were found to display a paucity of spiny branchlets, a reduced number of synaptic inputs, and degenerative alterations in their cytoplasm.

Conclusions

Bergmann glial cells are specialized radial astrocytes that constitute an important cellular element of cerebellar architecture. These cells control the development of cerebellar circuits by guiding the granule cells from the external to the internal granule cell layer, and they assist in the morphogenesis of Purkinje neurons. Mature Bergmann glial cells form complex 'microdomain' structures that ensheath the synaptic contacts formed on the dendrites of Purkinje cells. Functionally, Bergmann glial cells are endowed with a well-defined set of neurotransmitter receptors, which allow them to sense the activity of synapses formed on the Purkinje neurons. Neuronal activity stimulates the receptors localized in the Bergmann glial cell membrane, resulting in an activation of both transmembrane currents and intracellular Ca^{2+} responses. The latter are highly localized within the

microdomains, thus enabling spatial discrimination of the incoming signals. Bergmann glial cells also possess numerous transporter systems that participate in the clearance of neurotransmitters released during synaptic activity. Thus, Bergmann glial cells can be regarded as active elements of information processing in the cerebellum.

See also: Astrocyte: Response to Injury; Cerebellum: Clinical Pathology; Cerebellum: Evolution and Comparative Anatomy; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Gamma-Aminobutyric Acid (GABA); Glial Cells: T Cell Interactions; Glial Glutamate Transporters.

Further Reading

Bellamy TC and Ogden D (2005) Short-term plasticity of Bergmann glial cell extrasynaptic currents during parallel fiber stimulation in rat cerebellum. *Glia* 52: 325–335.

Bordey A and Sontheimer H (2003) Modulation of glutamatergic transmission by Bergmann glial cells in rat cerebellum *in situ*. *Journal of Neurophysiology* 89: 979–988.

Burnashev N, Khodorova A, Jonas P, et al. (1992) Calcium-permeable AMPA-kainate receptors in fusiform cerebellar glial cells. *Science* 256: 1566–1570.

Clark BA and Barbour B (1997) Currents evoked in Bergmann glial cells by parallel fibre stimulation in rat cerebellar slices. *Journal of Physiology* 502: 335–350.

Grosche J, Kettenmann H, and Reichenbach A (2002) Bergmann glial cells form distinct morphological structures to interact with cerebellar neurons. *Journal of Neuroscience Research* 68: 138–149.

Grosche J, Matyash V, Moller T, Verkhratsky A, Reichenbach A, and Kettenmann H (1999) Microdomains for neuron–glia interaction: Parallel fiber signaling to Bergmann glial cells. *Nature Neuroscience* 2: 139–143.

Huang H and Bordey A (2004) Glial glutamate transporters limit spillover activation of presynaptic NMDA receptors and influence synaptic inhibition of Purkinje neurons. *Journal of Neuroscience* 24: 5659–5669.

Iino M, Goto K, Kakegawa W, et al. (2001) Glia–synapse interaction through Ca^{2+}-permeable AMPA receptors in Bergmann glia. *Science* 292: 926–929.

Kirischuk S, Kettenmann H, and Verkhratsky A (1997) Na^{+}/Ca^{2+} exchanger modulates kainate-triggered Ca^{2+} signaling in Bergmann glial cells *in situ*. *FASEB Journal* 11: 566–572.

Kirischuk S, Kirchhoff F, Matyash V, Kettenmann H, and Verkhratsky A (1999) Glutamate-triggered calcium signalling in mouse Bergmann glial cells *in situ*: Role of inositol-1,4,5-trisphosphate-mediated intracellular calcium release. *Neuroscience* 92: 1051–1059.

Kirischuk S, Moller T, Voitenko N, Kettenmann H, and Verkhratsky A (1995) ATP-induced cytoplasmic calcium mobilization in Bergmann glial cells. *Journal of Neuroscience* 15: 7861–7871.

Kirischuk S, Tuschick S, Verkhratsky A, and Kettenmann H (1996) Calcium signalling in mouse Bergmann glial cells mediated by α_1-adrenoreceptors and H_1 histamine receptors. *European Journal of Neuroscience* 8: 1198–1208.

Kulik A, Haentzsch A, Luckermann M, Reichelt W, and Ballanyi K (1999) Neuron–glia signaling via $alpha_1$ adrenoceptor-mediated Ca^{2+} release in Bergmann glial cells *in situ*. *Journal of Neuroscience* 19: 8401–8408.

Matsui K, Jahr CE, and Rubio ME (2005) High-concentration rapid transients of glutamate mediate neural–glial communication via ectopic release. *Journal of Neuroscience* 25: 7538–7547.

Muller T, Moller T, Berger T, Schnitzer J, and Kettenmann H (1992) Calcium entry through kainate receptors and resulting potassium-channel blockade in Bergmann glial cells. *Science* 256: 1563–1566.

Muller T, Moller T, Neuhaus J, and Kettenmann H (1996) Electrical coupling among Bergmann glial cells and its modulation by glutamate receptor activation. *Glia* 17: 274–284.

Rakic P (1971) Neuron–glia relationship during granule cell migration in developing cerebellar cortex. A Golgi and electron microscopic study in Macacus Rhesus. *Journal of Comparative Neurology* 141: 283–312.

Reichenbach A and Wolburg H (2005) Astrocytes and ependymal glia. In: Kettenmann H and Ranson BR (eds.) *Neuroglia*, pp. 19–35. Oxford: Oxford University Press.

Riquelme R, Miralles CP, and De Blas AL (2002) Bergmann glia $GABA_A$ receptors concentrate on the glial processes that wrap inhibitory synapses. *Journal of Neuroscience* 22: 10720–10730.

Verkhratsky A, Orkand RK, and Kettenmann H (1998) Glial calcium: Homeostasis and signaling function. *Physiological Reviews* 78: 99–141.

Warr O, Takahashi M, and Attwell D (1999) Modulation of extracellular glutamate concentration in rat brain slices by cysteine–glutamate exchange. *Journal of Physiology* 514: 783–793.

Bilingualism

K Hakuta, Stanford University, Stanford, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Bilingualism (multilingualism) refers to the coexistence of more than one language system within an individual, as contrasted to monolingualism. The question of how the two languages interact at the cognitive and behavioral levels has been of long-standing interest to psycholinguists as well as to neurologists, clinicians, and educators. There has been great anticipation that developments in cognitive neuroscience could shed further light on important fundamental questions in the understanding of bilingualism.

Bilingualism as an individual condition is nested within a distribution of broader societal circumstances that cause language contact. There are many different manifestations of this variability. Bilingualism may be the result of growing up in a bilingual community, such as a bilingual neighborhood of an immigrant community in New York. However, that is different from bilingualism that results from growing up in an officially bilingual country such as Canada, where its two official languages are separated by geographical regions. Bilingualism that is accompanied by literacy in both languages is different from bilingualism in which schooling is available in one language (the one that also carries social prestige) but not the other. Although the interest of the cognitive neuroscientist in bilingualism may be in understanding the neural bases of the distribution of the two linguistic systems in the bilingual, the reality is that research subjects and clinical patients invariably come from a sampling from the social distribution. It is thus necessary to begin an understanding of bilingualism from its social bases.

Societal Characteristics

Social scientists have studied the properties, distribution, and stability of bilingualism in many societies and settings over time. The following sections discuss some key characteristics that may influence the bilingualism of individuals growing up within them.

Social Status

Whether the two languages of the bilingual hold equal or different status within the community will have consequences on the stability of the bilingualism over time, across generations. In English-speaking areas of Canada, for example, bilingualism is commonly attained by English speakers choosing to learn French in order to gain access to the privileges of bilingualism. These communities gain a second language without compromising the first language. At the same time, there are immigrant communities in Canada for whom English is a second language, schooling is available in English, and there is no societal value associated with the home languages of these individuals. For most of these bilinguals, the native language will have a limited range of use (often the homes, neighborhoods, and religious institutions) and it is not likely to survive in future generations. The differential status of the languages will produce bilingualism with different levels of strength and stability, even though these individuals are considered bilingual. Fully developed bilingualism in the absence of social status is mostly a transitional phenomenon.

Compartmentalization

Bilingualism often exists in circumstances in which the functions of the two languages are differently distributed across social and cognitive domains, and therefore the bilingualism is compartmentalized. Schooling is a major domain in which the language of immigrant communities is not represented. In such circumstances, the language repertoire of bilinguals is asymmetrical. Academic words are not developed in the language used only outside of schooling and academic contexts. Measurements of bilingualism using academically oriented tasks, such as a standardized test or a verbal IQ test, may yield results that are questionable in validity for the language that is not used in school.

Literacy

Literacy is a significant proxy for cultural capital and socioeconomic status of an immigrant group. Whether bilinguals have education (and thus literacy) available in both languages is telling of the relative status of the two languages in their society. The size of the mental lexicon is one area of language in which growth continues through the life span, and reading is a key means through which vocabulary growth occurs. Thus, the literacy history of the bilingual individual is important not just as a proxy to socioeconomic status but also in understanding the linguistic environment of the bilingual individual. Bilinguals who come from immigrant backgrounds can differ significantly in the extent to which they may have attained literacy in their home language

prior to immigration. Limited literacy may be due to lack of access to schools in their country of origin or, in extreme cases, a home culture such as the Hmong of Southeast Asia, which does not have a developed system of literacy.

Immigrant Generation

The phenomenon of language shift refers to immigrant communities as they move in successive generations from primary dominance in their heritage language to bilingualism and then to dominance or monolingualism in the new language. This shift is rapid in most immigrant environments, and it has been well documented in the United States. With the exception of the monolinguals who are linguistically assimilated as well as some of the first-generation immigrants who remain isolated, most of the others would be considered bilingual, even though their proficiencies would vary considerably. The first generation is characterized by considerable variation in second language proficiency, which is usually predicted by their age of immigration and educational opportunities. Second-generation bilinguals usually have the highest degree of proficiency in the two languages, especially those who are recruited to serve as translators for their parents. The third generation will contain significant variation in the heritage language proficiency, with a considerable proportion of individuals having passive or limited ability in the language, often limited to the domain of their home. Because immigrant groups tend to come in successive cohorts over a period of time, any given sampling from a seemingly homogeneous group (such as an ethnic category 'Hispanic') will still contain considerable variation with respect to immigrant generation and certainly variation in language proficiency.

Historical Circumstances Other Than Immigration

In many areas of the world, bilingualism occurs for reasons other than immigration through which languages come into contact. Bilingualism also occurs through an interest in learning a language that can provide access to communication and global resources, a status that English holds today. Bilingualism can also occur through conquest or colonization – as contrasted with voluntary immigration – in which an occupying force imposes its language on the community. Such is the case with indigenous languages in the Americas or the Aboriginal languages in Australia. These are all circumstances of bilingualism that require special analysis and attention with respect to the bilingual individuals contained therein.

Individual Considerations

The bilingual literature often classifies bilingual individuals independently from their societal conditions. This does not allow for an appreciation of the fact that the sampling for bilinguals of a given characteristic at the individual level may be different depending on the societal condition from which the sample is being drawn. A bilingual who is literate in both languages is relatively easy to find in some contexts and quite difficult to find in others, and so studies conducted with bilinguals in these two different settings may have very different sample characteristics because they are sampled from different sectors of their society. The following sections discuss considerations of bilinguals at the individual level that have been prominent in the literature but are best framed by the societal conditions in which they occur.

Age of Acquisition

The chronological age at which an individual is exposed to a second language has important consequences for the level of bilingualism attained. Some individuals, known as simultaneous bilinguals, are bilingual from birth and grow up in families in which the two languages are used interchangeably. In cases of simultaneous bilingualism, one might ask the question of whether either of the languages was more dominant (in most circumstances, that would be the mother's language), although this information may be difficult to obtain objectively.

In the case of sequential bilinguals, the individual may begin learning the second language at different ages. Typically, the literature distinguishes between early bilingualism (before age 5 years), late childhood bilingualism (between ages 5 and 15 years), and adult bilingualism. However, there is considerable variability in the actual cutoff points used. Very few studies distinguish between bilingualism in the various stages of adulthood, although this may become an important distinction as bilinguals become increasingly studied in the aged population.

Age of exposure to the second language is necessarily correlated with the degree of development of the first language. Early bilinguals are considered to be learning a second language prior to the full grammatical development of the first, and therefore the two developing systems will interact more actively. Late bilinguals have an established grammar in their first language, but their literacy skills are still in the process of development, and thus schooling becomes an important mediating factor. Adult bilinguals vary considerably in their level of formal

schooling and therefore may exhibit considerable variability in their circumstances of bilingualism based on sociodemographic characteristics. All of these conditions will lead to differing levels of proficiency in the two languages.

The most commonly asked question related to the age of the learner is whether there is a biologically conditioned critical period for second-language acquisition, parallel to the critical period hypothesis for native language acquisition. The central part of the hypothesis is that there is a period of readiness during which limited input is sufficient for complete development and a corollary that no amount of input outside the period is sufficient for normal development. There appears to be an unquestionable decline in second language attainment as a function of the age of exposure to the second language, particularly in the area of phonology. However, there is no convincing evidence for a qualitative difference in the outcome of second-language acquisition that occurs within and outside of the critical period. Since most of the research conducted thus far has used varying psycholinguistic and sociological measures of language proficiency, there is great anticipation that evidence from brain localization and activity can shed light on this most interesting problem.

Language Proficiency

Bilinguals can be differentiated with respect to levels of proficiency in their languages. Grammatical proficiency can involve attainment of native-like control in the linguistic arenas of phonology, morphology, syntax, semantics, or pragmatics. Grammatical proficiency usually involves a scale in which native-like control is an absolute threshold – for example, the control of English regular and irregular past tense marking on verbs. The most commonly examined areas of grammatical proficiency are in phonology, morphology, and syntax. Caution is necessary in using native-like control as a standard when there is considerable social and dialectal variation in the standard. South Asians, for example, are often exposed to a variety of English that is known as 'World Englishes,' which has many linguistic and social characteristics that are quite different from standard English. Also, many immigrant students in the United States are exposed to nonstandard varieties of English in schools, such as African American Vernacular English or Black English.

Another arena of proficiency is in the aspects of language involving literacy. Measures of literacy typically show variability even within a native monolingual population, and many measures are constructed in order to produce a normal distribution of performance. Performance on measures of literacy in the two languages of the bilingual can produce vast discrepancies resulting from formal educational experiences.

Balanced bilinguals are individuals considered equivalent in their proficiencies in the two languages. Many researchers consider this to be an idealized state that is not attained by most bilinguals, and bilinguals need to be identified with respect to their language dominance. It should not be assumed that the native language is dominant in sequential bilinguals, especially among first- and second-generation immigrants who are schooled in their second language. An additional complication arises from the fact that dominance may depend on the domain of language that is sampled.

Semilingualism refers to a state in which proficiency in both languages is considered incomplete. Because the grammatical aspects of language are considered to be part of the innate competence of individuals, those with incomplete proficiency in these aspects in both languages are generally considered to be linguistically impaired, assuming that dialectal differences can be ruled out. In the domain of literacy, it is generally accepted that literacy attained in one language predicts literacy attained in the other language, provided that learning opportunities are provided. Thus, failure to attain age-appropriate levels of literacy in either language is a well-noted phenomenon and worthy of serious investigation.

Active versus Passive Bilingualism

Bilinguals can vary considerably in the extent to which they use their two languages, independent of proficiency. This may be an important consideration, to the extent that responses on psycholinguistic and neurolinguistic tasks of subjects depend on the effects of practice or the recency of use of the language. For example, an adult immigrant who moves to a region where there is no sizeable speech community of speakers of the native language may have a high level of proficiency but very few opportunities for use. Also, an individual who is married to someone of the same linguistic background is more likely to maintain use of the language than someone who marries outside his or her linguistic group. The study of bilingualism that does not consider the level of active use of the two languages and simply examines levels of proficiency may miss an important source of variation on performance on linguistic tasks. Sociolinguists who have studied bilingual speech communities have identified individuals who

may rarely speak the language, or who may have lost some of the language, but are nevertheless considered to be members of that speech community and self-identify as such.

Linguistic Considerations

The similarities and differences between the languages of bilinguals introduce an important set of considerations. Linguists have noted varying amounts of interference across languages that can be characterized with respect to their differences, noting, for example, the difficulty of Japanese learners of English in mastering the /*r*/ and /*l*/ sound distinction or the plural marking on nouns. Languages can vary along the lines of historical relationships, as well as at varying levels of their structural features. For example, Japanese, Turkish, and Tamil are considered similar in grammatical characteristics despite their distance with respect to historical relationships. Yet, historical relationships can be extremely important in areas such as the lexicon because linguistic borrowing due to language contact is a relatively common occurrence.

The study of second-language acquisition has also been guided by the question of linguistic transfer from the first to the second language. Second language acquisition is considered to be more efficient if the two languages are grammatically similar rather than if they are different. Although there is debate as to the comprehensive nature of language transfer, it is undisputable that linguistic similarity will determine to a large extent the rapidity of the learning process: English-speaking adults who are learning Italian, German, Spanish, or French will progress at least twice as rapidly as if they are learning Russian, Bulgarian, Arabic, or Korean under comparable conditions, and with comparable levels of aptitude in learning a second language. On the other hand, linguistic theories that emphasize the common underlying principles of human language, such as the Chomsky-inspired approaches to universal grammar, make special note of the similarities that can also be found among second-language learners despite differences in their native language.

Psycholinguistic Perspectives

The major question dominating psychological studies of bilingualism has been how the two languages are represented – the two poles being independence and interdependence. These questions have been asked in the context of the domains of language comprehension and language production, with considerably more experimental research in the area of comprehension. A related question is how organization might vary as a function of language dominance or degree of proficiency in the two languages.

A key emerging generalization appears to be that the bilingual's two languages are usually interdependent, with evidence for a shared storage system. This evidence is strongest in the area of word identification, in which considerable experimental work in lexical access suggests that both languages are activated simultaneously during word recognition. For example, investigations of cognates are revealing because these words appear in both languages. Experiments have found that bilinguals recognize cognates as being words faster than noncognate words, presumably because both lexicons are being accessed simultaneously. Studies of semantic priming in word recognition also support the notion of linguistic interdependence. In these tasks, subjects are asked to make lexical decisions of words that are immediately preceded by a semantically related or unrelated word. In mixed language presentations, there is semantic facilitation even if the priming word is in a language different from that of the test word, again suggesting that both languages are simultaneously active.

Studies of episodic memory, particularly for word lists presented in either language, suggest that bilinguals organize information principally along semantic dimensions, although they are able to remember the language of presentation especially depending on the condition of the task. Studies of semantic memory in which incongruence between the stimulus and the word is explored (such as the Stroop Test, in which subjects are asked to name the color of the ink of words signifying congruent or incongruent colors – for example, the word 'red' printed in red or green ink) show significant cross-language interference, also suggesting a principal effect of cross-language interdependence.

A key area of exploration is the conditions of the task and context that might modulate these basic findings. Computational modeling of bilingual cognition is still in the early stages, with some promising attempts to model linguistic interdependence, independence, and interference. Such models may prove to be useful for bridging theoretical models of representation with the task and context demands of bilingualism.

Differences between bilinguals and monolinguals are an area of some interest, especially by those concerned about the possible positive or negative effects of bilingualism on cognition. Developmental research suggests that bilingualism in children may result in accelerated development in the executive function areas of attention and inhibition but not

necessarily in the area of general cognition. However, there is evidence that bilinguals are slower than monolinguals in tasks such as lexical decision making, presumably because of a larger combined lexicon.

Brain Localization

Evidence of language localization in the left hemisphere, particularly Broca's and Wernicke's areas, for specific functions is supported by the accumulated aphasia literature. Some deviations have been noted, particularly a greater degree of right hemisphere involvement in the less dominant language.

Much research on bilingual aphasia has focused on recovery patterns, such as whether recovery is parallel or whether it is differential and selective. Tabulations of cases of recovery patterns indicate that approximately four in ten cases show parallel recovery, with the remainder being approximately evenly split between stronger recovery in the first language and stronger recovery in the second language. Thus, all possibilities seem to occur in roughly equal proportions. These patterns must all be interpreted with strong caveats about the varying quality of the linguistic assessment of the recovery as well as the descriptions of the bilingual history and proficiency of the subjects. Cases of nonparallel recovery are noted as indicating differentiated organization, although alternative interpretations are possible, such as a 'switch' mechanism between the languages.

Research using bilateral stimulus presentation, such as dichotic listening, that explores the question of hemispheric dominance suggests that language may be represented more bilaterally (i.e., more evidence of right hemisphere engagement) in bilinguals than in monolinguals, with the case being more so for the second language than the first language. This evidence is not conclusive.

Research using the neuroimaging techniques of positron emission tomography and functional magnetic resonance imaging with bilinguals is showing promise in advancing the question of the independence and interdependence of the two languages. Specifically, these studies examine whether there are different or overlapping areas of activation when processing the respective languages of the bilingual.

In language production studies, bilinguals are typically asked to repeat words, generate words, or, in some cases, provide narrative during the scan. The results show that there are overlapping and focused areas of activation for bilinguals with high proficiency in both languages, regardless of whether they are early or adult bilinguals, supporting the interdependence view of bilingualism. However, among adult bilinguals with low proficiency in the second language, the pattern of activation is less focused and exhibits some degree of right hemispheric engagement. The studies strongly point to the importance of language proficiency as an explanatory condition for localization as well as an important control and subject selection criterion.

Studies of language comprehension are better able to control tasks for semantic and syntactic levels of processing. These studies, although fewer in number, suggest that semantic processing (judging whether sentences are meaningful) in the two languages shows a similar pattern of results as those found in the production studies. However, in the area of syntactic processing (judging the grammatical content of sentences), the adult bilinguals showed more extended areas of activation in their second language, even for those with high proficiency. Although these results are quite preliminary, they point to the promise of neuroimaging methods in untangling what is possibly a three-way interaction between language proficiency, age of acquisition, and specific language function (semantics vs. syntax) in bilingual representation.

Conclusion

The neuroscience of bilingualism shows great promise in addressing important issues in the mental representation of bilingualism that have been identified in the psycholinguistics and cognitive science. In turn, the study of bilingualism challenges all fields to be more specific in identifying and incorporating the diversity of the subject population. Language proficiency and age of bilingualism are both key variables in the cognitive and neuroscience literature, and both of these are determined in large part by social and cultural conditions. At a minimum, these background variables must be carefully documented and examined as part of the subject screening and data reporting in order to draw appropriate conclusions from research in the neurological paradigm.

See also: Language Development; Language: Cortical Processes; Language: Aphasia; Psycholinguistics; Second Language Acquisition; Sentence Production; Sentence Comprehension; Statistical Learning of Language; Word Production; Word Recognition; Word Learning.

Further Reading

Bhatia TK and Ritchie WC (eds.) (2004) *The Handbook of Bilingualism*. Malden, MA: Blackwell.

Bialystok E (2001) *Bilingualism in Development*. Cambridge, UK: Cambridge University Press.
Fishman J (ed.) (1978) *Advances in the Study of Societal Multilingualism*. The Hague, The Netherlands: Mouton.
Hakuta K (1986) *Mirror of Language: The Debate on Bilingualism*. New York: Basic Books.
Hamers JF and Blanc MHA (2000) *Bilinguality and Bilingualism*, 2nd edn. Cambridge, UK: Cambridge University Press.
Kroll JF and De Groot A (eds.) (2005) *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford: Oxford University Press.
Lambert WE (1972) *Language, Psychology, and Culture: Essays*. Stanford, CA: Stanford University Press.
Paradis M (2004) *A Neurolinguistic Theory of Bilingualism*. Amsterdam: Benjamins.
Romaine S (1995) *Bilingualism*, 2nd edn. Malden, MA: Blackwell.
Wei L (ed.) (2000) *The Bilingualism Reader*. London: Routledge.

Binocular Rivalry

R Blake, Vanderbilt University, Nashville, TN, USA
R P O'Shea, Otago University, Dunedin, New Zealand

© 2009 Elsevier Ltd. All rights reserved.

Binocular rivalry is a form of multistable perception in which visual awareness fluctuates irregularly between different stimuli imaged to corresponding retinal regions of the two eyes. Under ordinary viewing conditions, rivalry is not patently experienced. Instead, people with normal binocular vision usually experience a single stable visual world even though objects throughout most of the visual field produce two retinal images, one in each eye. Evidently the brain manages to blend, or fuse, the two images and to do so in a manner that precisely specifies the three-dimensional relationships among those objects. The processes responsible for such binocular single vision and stereopsis, however, are disrupted when dissimilar monocular stimuli are imaged on corresponding retinal locations. Under these conditions, the eyes transmit contradictory information to the brain, specifying the presence of two different objects situated at the same spatial location at the same time. Faced with this physical impossibility, the brain lapses into an unstable state characterized by alternating periods of perceptual dominance which continue as long as the eyes view discordant stimuli; this is the phenomenon termed binocular rivalry.

During binocular rivalry, conscious visual experience is intermittently dissociated from physical stimulation; a normally visible, complex visual stimulus viewed by one eye may be erased, or suppressed, from visual awareness for several seconds at a time. For this reason, rivalry provides an effective means for identifying neural events accompanying conscious visual awareness. The study of rivalry may also shed light on neural mechanisms underlying perceptual selection and on the resolution of ambiguous sensory information.

History of Binocular Rivalry

The first published account of binocular rivalry was by Giambattista della Porta in 1593. Porta placed one book in front of one eye and another book in front of the other eye. Despite the potential conflict presented by this dichoptic viewing condition, Porta was able to read from one book at a time with no visible disturbance from the book being viewed by the other eye. To change from one book to another, Porta described willfully moving the "visual virtue" from one eye to the other. Several centuries later, more detailed descriptions of rivalry were published in French by Étienne François Dutour, who distinguished between contour rivalry and color rivalry. The first clear description of rivalry in English was provided in 1838 by Charles Wheatstone, who also invented the mirror stereoscope and discovered the geometric basis of stereopsis. Just before the turn of the twentieth century, B B Breese published an influential monograph detailing the stimulus conditions triggering rivalry and influencing the alternations of perception over time. In the 1950s and 1960s, binocular rivalry was used extensively by new look psychologists to show personality and social influences on perception. For example, they showed that, during rivalry between two complex stimuli such as pictures, the one that was meaningful for a person tended to predominate. Psychophysical interest in rivalry was revived in the 1960s, when W J M Levelt published a monograph on it. He offered a formal model of the dynamics of rivalry built around the unifying concept of stimulus strength. Over the next few decades, psychophysical studies of binocular rivalry revealed much about the spatial and temporal characteristics of the phenomenon, together with the relation of rivalry to stereopsis and to other aspects of visual perception. Possible neural concomitants of rivalry have been sought by physiological studies in which responses of single visual neurons are correlated with perceptual reports obtained from alert, behaving animals exposed to rival stimuli. Very recently, brain-imaging studies have provided tantalizing glimpses into possible brain areas involved in producing binocular rivalry in humans.

What Instigates Binocular Rivalry

Binocular rivalry is instigated by presenting dissimilar monocular stimuli to the two eyes, usually via a stereoscope. Binocular rivalry occurs between any stimuli that differ sufficiently in their spatial configurations (contour rivalry) or in their colors (color rivalry). When the images presented to the eyes differ only in their luminances, a form of rivalry called binocular luster may be experienced. When images differ only in their rates of flicker, however, no rivalry is experienced. Rivalry may be triggered by viewing simple stimuli such as lines of differing in orientation (Figure 1(a)) or by viewing complex stimuli such as pictures of a face and of a house (Figure 1(b)). It is not essential that rival stimuli be presented absolutely simultaneously; rivalry can also be observed when the two dissimilar views are presented in rapid alternation.

Figure 1 Examples of dissimilar stimuli which, when imaged separately to the two eyes, provoke binocular rivalry: (a) orthogonally oriented gratings; (b) face and house (the face image is Sir Charles Wheatstone, inventor of the stereoscope and author of the first systematic study of binocular rivalry); (c) pair of rival stimuli in which complete dominance of one entire figure (circular bull's eye) or the other (horizontal contours) requires simultaneous dominance of both eyes within different regions of the visual field; (d) superimposed, orthogonally oriented gratings differing color, a figure that yields monocular, or pattern, rivalry without segregation of the two gratings to different eyes. (c) Adapted from Diaz-Caneja E (1928) Sur l'alternance binoculaire [On binocular alternation]. *Annales d'Oculistique* 165: 721–731.

When a figure is presented to one eye and a blank field to the other, the figure is usually seen continuously. Occasionally, however, the blank field, or even the dark field of a closed eye, can become visible, with the figure temporarily suppressed from vision. It is important to distinguish this form of interocular suppression from Troxler's effect, the spontaneous fading of a binocular stimulus owing to retinal adaptation.

When a figure is presented to one eye and a blank field to the other, introducing a different stimulus onto to the blank field usually results in that stimulus being seen immediately. This abrupt disappearance of the initially dominant figure is referred to as flash suppression, and this procedure provides an effective means for controlling the timing of transitions in perceptual state.

With conventional rivalry, a given eye is exposed to a given figure during the entire rival viewing period. There are, however, other modes of rival stimulation that can produce alternations in perceptual dominance. One technique, illustrated in **Figure 1(c)**, involves splitting each stimulus into two or more parts and presenting some parts to one eye and the other parts to the other eye. Under this condition, observers occasionally experience exclusive dominance of the original stimulus; this outcome – called interocular grouping – would be impossible if one entire eye were completely dominant at any given moment. Another technique involves rapidly and repetitively swapping the two dissimilar figures between the eyes. Here, too, observers sometimes experience periods of exclusive dominance of a given stimulus for periods of time that transcend multiple eye swaps. A third technique is simply to superimpose the two stimuli to the same eye. The relative clarity of the two stimuli fluctuates in the same way as in binocular rivalry, and occasionally one disappears entirely (see **Figure 1(d)**), a phenomenon called monocular rivalry by Breese but now also known as pattern alternation. Whether all of these instances of perceptual rivalry arise from the same neural mechanisms remains a point of contention.

How to Measure Binocular Rivalry

Binocular rivalry is most commonly measured by having an observer track fluctuations in perceptual dominance by pressing one of two buttons to denote exclusive dominance of one or the other stimulus. With this tracking procedure, one can estimate the rate of alternation in perception, the average duration of dominance periods of a given stimulus, and the total duration of dominance of a given stimulus.

Instead of tracking alternations in rivalry using buttons, the observer can be required to respond to monocular probe targets briefly superimposed on a rival stimulus. Probes are more difficult to detect when that rival stimulus is suppressed than when the same rival stimulus is dominant. Thus, probe sensitivity can serve as an index of the perceptual state of a stimulus during rivalry. Moreover, a change in a rival stimulus itself can constitute the probe; an otherwise conspicuous change in a suppressed stimulus can go unnoticed for several seconds, provided that change is made without abrupt alteration in luminance or contrast. For example, an observer may fail to detect a transition from incoherent, random motion to coherent, structured motion when that transition occurs while the motion is suppressed during rivalry.

It is also possible to infer perceptual states of dominance and suppression during rivalry using any one of several alternative physiological measures. These include optokinetic nystagmus (OKN), visual evoked potentials (VER) recorded from the scalp, magnetoencephalography (MEG) which measures magnet fields produced by the brain's electrical activity, and blood-oxygen level dependent (BOLD) signals measured using functional magnetic resonance imaging (fMRI). With all these indirect measures, the amplitude (VER, BOLD, and MEG) and the velocity (OKN) of the evoked signals vary depending on the currently dominant stimulus.

What One Experiences during Binocular Rivalry

When rival targets are first presented, one may experience both simultaneously as if the two were superimposed in visual space. Called false, or abnormal fusion, this perceptual experience lasts only a fraction of a second before one stimulus becomes exclusively visible. Thereafter, alternations begin, interspersed by periods during which bits and pieces of both stimuli are visible simultaneously (mixed dominance). Transitions between perceptual states appear as waves of dominance arising locally and spreading rapidly over space, as if one stimulus were sweeping the other from conscious awareness. Ordinarily, these waves of dominance occur unpredictably, but it is possible to ignite a transition wave locally and then measure the direction and speed of its spread over the visual field.

The incidence of periods of exclusive dominance of a given stimulus during rivalry decreases with stimulus size and light level and increases with retinal eccentricity. Dominance is not strictly confined to the boundaries of a rival stimulus but, instead, extends beyond those boundaries to form a halo of dominance surrounding the stimulus. Neighboring

regions of dominance are not independent, as evidenced by complete monocular dominance of a relatively large target occurring significantly more often than would be expected based on chance alone. Likewise, multiple rival stimuli at different regions of the visual field can engage in synchronized alternations, such that all stimuli of a given configuration (e.g., all green disks) are dominant simultaneously. Moreover, monocular stimulation falling outside the region of rivalry can enhance the predominance of a rival stimulus, acting as if it reinforced the strength of the stimulus itself.

During rivalry, perception alternates unpredictably over time: the successive durations of dominance and suppression are sequentially independent random variables. Moreover, there are large individual differences in alternation rate, and there is some evidence that rivalry rate is related to other cyclic behaviors such as the nasal cycle. Rivalry alternations also vary within the visual field of the same individual, and alternations are generally slower in older individuals. An observer cannot willfully switch from suppression to dominance instantaneously, except perhaps by blinking or suddenly moving the eyes. Nor can most observers hold one rival stimulus dominant indefinitely. Still, the predominance of a stimulus during rivalry is not completely uncontrollable. Manipulations of stimulus variables such as luminance, contrast, contour density, and velocity can produce pronounced variations in stimulus predominance, defined as the total percentage of time that a stimulus is visible during an extended viewing period. This increased predominance is produced largely by decreases in the average duration of suppression periods of a given stimulus – stronger stimuli tend to stay suppressed for shorter periods of time but do not necessarily stay dominant for longer periods of time. The predominance of a given stimulus can also be promoted by repetitively turning off both rival stimuli for several seconds and then turning them back on for several seconds – the rival stimulus dominant at stimulus offset tends to be the dominant stimulus when rivalry is again initiated.

To a limited extent, the dynamics of rivalry are susceptible to top-down influences. At the onset of rivalry, for example, the initially dominant stimulus tends to be the one toward which attention was directed immediately before the onset of rivalry. During rivalry, moreover, the successive durations of dominance of a given stimulus can be significantly prolonged by sustained attention directed to that stimulus. In the nineteenth century, Hermann von Helmholtz described being able to hold one of two rival gratings in dominance by mentally counting the number of light and dark bars making up the grating. Helmholtz attributed this control to attentional volition, but his long-time adversary Ewald Hering skeptically attributed the feat to eye movements and not to attention. Contemporary work, however, has vindicated Helmholtz's claim by showing that attention can indeed modulate rivalry alternations, albeit only modestly. This effect of attention on predominance may explain why a meaningful or an emotionally salient stimulus predominates during rivalry. Mood, too, purportedly affects binocular rivalry, with rate of alternation being significantly slower in people diagnosed with bipolar disorder. Laughter is said to alter rivalry by increasing the incidence of mixed dominance wherein portions of both rival stimuli are visible simultaneously.

Relevance of Binocular Rivalry to Visual Neuroscience

Binocular rivalry remains a topic of enduring interest for multiple reasons. First, the phenomenon itself is inherently fascinating: during rivalry a complex, suprathreshold stimulus disappears from visual awareness for seconds at a time, even though that stimulus remains imaged on the retina – perceptually speaking, a suppressed stimulus seems temporarily nonexistent. Because of this profound intermittent dissociation between physical stimulation and perceptual experience, rivalry affords a paradigmatic case for the study of conscious visual awareness. One expects to find correlated fluctuations in neural activity associated with the intermittent disappearance of a stimulus; as described in the next section titled 'Neural bases of binocular rivalry,' the search for those neural correlates has intensified, using both single-cell recording from monkeys and brain imaging in humans.

Second, and in a related vein, rivalry provides an effective technique for identifying aspects of visual processing that transpire outside of awareness. To what extent does a suppressed stimulus remain visually effective? An answer to this question can shed light on the locus of rivalry relative to other visual processes. To illustrate, exposure to a picture of an object ordinarily makes it easier to name that object when it is subsequently encountered, a phenomenon called visual priming. However, priming is disrupted if an ordinarily effective picture is presented to an eye during suppression and, therefore, outside of visual awareness. Evidently the neural events promoting visual priming arise later in the visual hierarchy than those producing suppression. On the other hand, visual adaptation to translational motion can occur even when the inducing motion itself is suppressed from awareness for a substantial portion of the adaptation phase owing to binocular rivalry. In the case of motion, neural events underlying adaptation

transpire even when the evoking stimulus falls outside of awareness. In general, rivalry can be used to implement what vision scientist Bela Julesz dubbed psychoanatomy: pinpointing the relative loci of neural processes involved in different visual phenomena.

Third, perception during rivalry sometimes assumes a mosaic appearance, meaning that bits and pieces of the competing rival stimuli are perceptually dominant within different regions of the visual field. Usually, however, dominance over portions of the visual field becomes coordinated, yielding global dominance of spatially separated object features. Thus, rivalry offers an effective means for studying the determinants of global perceptual organization. Here, too, there should exist neural signatures of this phase-locked perceptual grouping among neurons representing spatially distributed features.

Fourth, binocular rivalry is widely considered to be one member of a class of visual phenomena that all involve perceptual ambiguity and multistability. According to this approach, the input to vision is inherently underspecified and, hence, ambiguous; this thus forces the visual system to deploy nonsensory information to select among alternative perceptual interpretations. Viewed in this way, multistable phenomena, rivalry included, can be fruitfully used to pinpoint brain areas involved in attention and selection. Furthermore, some people believe that all these forms of perceptual rivalry are driven by a common neuronal oscillator whose intrinsic rate varies among individuals.

Fifth, binocular rivalry can be construed as the default outcome when binocular fusion fails, making rivalry an acid test for assessing the brain's tolerance for monocular mismatches. Rivalry, in other words, reveals the limits for binocular fusion. At the same time, binocular rivalry may shed light on an important process underlying binocular single vision. It is true that left and right foveae rarely receive dissimilar stimulation for any length of time – the oculomotor system attempts to correct this situation by altering vergence angle until matching features are imaged on the two foveae. There are, however, locations on the two retinae where dissimilar monocular images strike corresponding retinal areas, an inevitable consequence of the geometry of binocular vision. Yet one ordinarily does not experience rivalry or diplopia under these conditions. Binocular single vision thus may be accomplished by two processes: binocular fusion and binocular suppression. The potent inhibitory process revealed during rivalry may contribute to normal binocular single vision.

Sixth, rivalry may have implications for understanding the medical condition called strabismic suppression – suppression of vision in one of two eyes of an individual with misaligned eyes – a condition that can lead to the permanent degradation of visual function in that eye. The stimulus conditions eliciting rivalry suppression and strabismic suppression are comparable; corresponding areas of the two eyes receive discrepant inputs. Moreover, there is some evidence suggesting that the two forms of suppression are comparable psychophysically. Inspired by this possible link, several laboratories have used visually evoked potentials to examine the development of rivalry and binocular interactions in very young infants, the motive being to relate those findings to the onset of strabismic suppression.

Neural Bases of Binocular Rivalry

Neural concomitants of binocular rivalry have been investigated in humans using brain imaging and in alert, behaving monkeys using single-cell recording techniques. Both techniques have identified neural events correlated with rivalry dominance and suppression within multiple brain areas within the visual hierarchy.

In monkeys trained to indicate fluctuations in stimulus dominance during rivalry, single-cell recordings reveal modulations in firing rates of neurons within visual areas ranging from V1 to the inferior temporal cortex (Figure 2). In the one study that examined neural responses within the lateral geniculate nucleus (LGN) of alert, behaving monkeys, no changes in firing rate were observed during rival stimulation. In V1, only a minority of neurons show responses that track the perceptual state reported by the monkey, and there is no tendency for such neurons to be dominated by one eye or the other. In contrast, within the inferotemporal (IT) visual area nearly all neurons exhibit fluctuations in spiking activity that closely tracks the monkey's perceptual state. Several neurophysiological studies have described the existence of transient increases in neural synchrony within primary visual cortex in response to the presence of a stimulus that was dominant in rivalry; during suppression phases, neural synchrony was abolished. These fluctuations in neural synchrony are not accompanied, however, by modulations in the firing rate of neurons responsive to the evoking stimulus.

Brain imaging using fMRI in humans reveals that neurally triggered BOLD signals are attenuated in amplitude when the evoking stimulus is suppressed during rivalry (Figure 3). These attenuated BOLD signals have been observed in the lateral geniculate nucleus; in the primary visual cortex; and in areas within the temporal, parietal, and frontal lobes. Within V1, modulations in BOLD signal spread

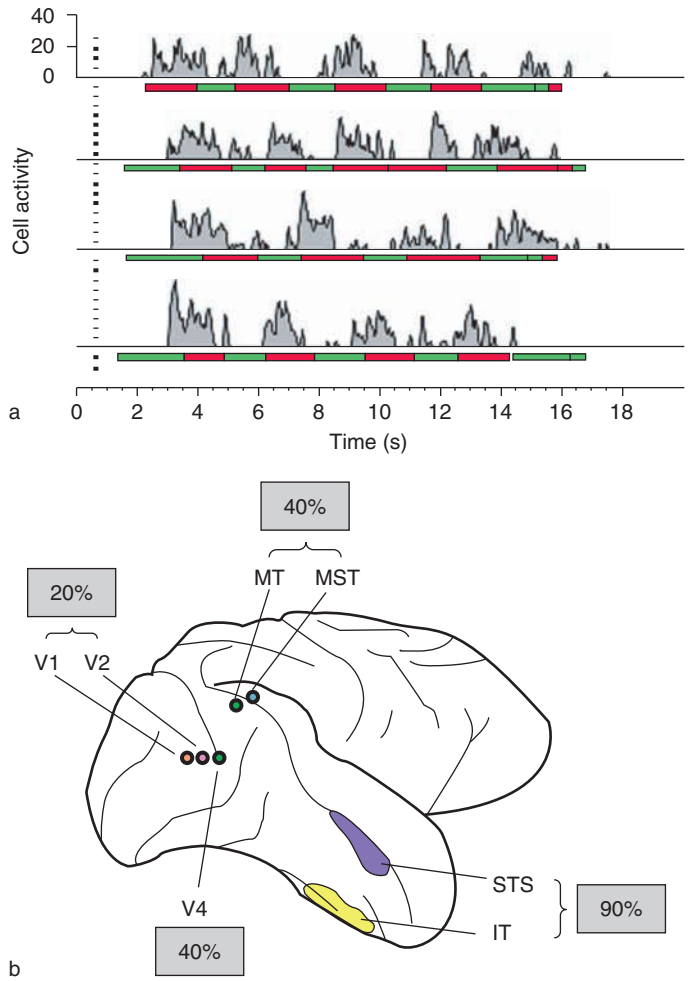

Figure 2 Neural correlates of binocular rivalry in the monkey visual system: (a) cell activity (action potentials/s) over time in a V4 visual neuron recorded while a trained monkey tracked fluctuations in perception of two rival stimuli, one being the recorded neuron's preferred stimulus; (b) percentage of neurons in different visual areas of the macaque monkey brain whose activity is modulated in synchrony with the animal's perceptual reports of dominance and suppression during rivalry. In (a), the two-tone, shaded timeline below each trace indicates dominance and suppression states of the preferred stimulus as reported by the monkey. IT, inferotemporal visual area; MST, medial superior temporal visual area; MT, middle temporal visual area; STS, superior temporal sulcus; V1, V2, V4, visual cortices. Adapted from Leopold DA and Logothetis NK (1999) Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences* 3: 254–264 (figure 2), with permission.

over time in a wavelike fashion within the retinotopic map, closely following the spatiotemporal transitions in perceptual state between one rival stimulus and the other. Within inferior parietal and prefrontal areas, transient increases in the BOLD signal are tightly correlated with switches in perceptual state. In contrast, BOLD responses from the amygdala, a lynchpin structure in the network of brain areas underlying processing of affective information, are minimally affected during rivalry. Specifically, BOLD responses in the amygdala are evoked by faces portraying emotional expressions, even when those faces are suppressed during rivalry and produce minimal activity within the putative fusiform face area.

At present, there are discrepancies between the conclusions implied by single-cell recording studies in monkeys and the conclusions being drawn from fMRI studies of humans – the fMRI results imply more pronounced effects of rivalry on neural responses within early visual areas than the single-cell results. It remains to be learned whether these discrepancies are attributable to fundamental differences in technique or to differences between human and monkey visual systems.

Neural Models of Binocular Rivalry

Over the years several neural models of binocular rivalry have been published. Although their details

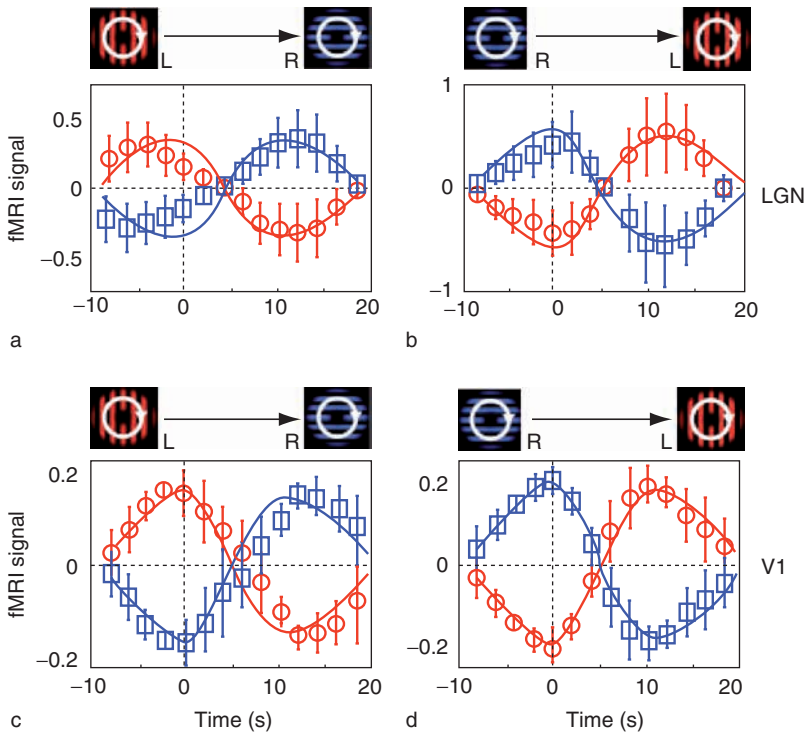

Figure 3 Neural correlates of binocular rivalry in the human visual system. Each panel shows modulations in event-related, functional magnetic resonance imaging (fMRI) BOLD signals measured from voxels dominated by left-eye stimulation (red) and by right-eye stimulation (blue). The traces in (a) and (b) were measured from voxels in the lateral geniculate nucleus (LGN) and those in (c) and (d) from voxels in primary visual cortex (V1). (a) and (c) show BOLD signal modulations accompanying transitions in perceptual dominance from left-eye/red dominant to right-eye/blue dominant; (b) and (d) show modulations accompanying transitions from right-eye/blue to left-eye red. Solid lines show responses predicted by convolving the reported perceptual alternations with a canonical hemodynamic model of the BOLD response. Reproduced from Haynes J, Deichmann R, and Rees G (2005) Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature* 438(7067): 496–499, with permission from Nature Publishing Group.

differ, nearly all these models combine reciprocal inhibition with neural adaptation to produce changes in activity patterns within neural elements representing the alternative rival stimuli (Figure 4). The most recent computational models have added to these elements the strength of excitatory drive to determine which population of neurons is active and which is suppressed. Alternations in the activity over time come about through spike-frequency adaptation, which progressively weakens the strength of excitatory neural signals produced by the dominant stimulus and, as a consequence, the strength of inhibitory influence on the neurons representing the suppressed stimulus. Some computational models implement rivalry within a relatively large ensemble of Hodgkin–Huxley type neurons, whereas others embody interactions within a simpler neural network. Rivalry models can be formulated either in terms of spike rates or conduction dynamics, which can be quite helpful when evaluating the actual physiological plausibility of best-fit parameter values. Contemporary models are formulated in terms grounded in biophysics, implying network susceptibility to levels of specific neurotransmitters (e.g., dopamine and serotonin); these models, unlike their predecessors, imply that rivalry dynamics should be affected in specific ways by experimental manipulations of the levels of those neurochemicals (e.g., through the administration of selective serotonin reuptake inhibitors). Moreover, recent models offer physiologically plausible accounts of individual differences in rivalry dynamics. Finally, these recent models, unlike earlier ones, can account for the wavelike transitions from the dominance of one stimulus to the other and for the effect of spatial configuration of rival targets on rivalry dynamics.

Eye versus Object Rivalry

During binocular rivalry, perceptual experience tells us that one object or the other is currently dominant (or that bits and pieces of both rival objects are jointly visible in the case of piecemeal rivalry). At the same time, we have no idea which eye is viewing the dominant stimulus unless we close an eye to discern the answer. Moreover, as previously mentioned, under

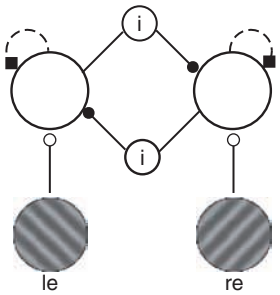

Figure 4 Generic model of binocular rivalry. Arrays of neurons responsive to dissimilar stimuli receive excitatory input (open circles) when their preferred stimuli are present. When those stimuli are imaged on corresponding areas of the two eyes, the neurons exert reciprocal inhibition on one another (interneurons i shown with filled circles). Owing to imbalance in the initial state of activation of the two arrays of neurons, one group of neurons squelches the other's activity promoting dominance of one stimulus over the other. With time, the imbalance in activation between the two groups of neurons diminishes through neural adaptation (dotted lines and filled squares), culminating in a switch in the relative activity levels and, consequently, in the perceptual status of the evoking stimuli. le, left eye; re, right eye.

some conditions people can experience slow alternations in dominance even when the two eyes' views are rapidly and repetitively exchanged during an extended viewing period. Perceptual appearance, in other words, implies that competition for dominance during rivalry transpires between alternative object representations, not between the two eyes. According to this view, the neural concomitants of rivalry competition should be found within high-level visual areas where object properties are seemingly represented. Significantly, neurons within these high-level visual areas retain no information about which of the two eyes is viewing a given object. Yet, at the same time, there is strong indirect evidence implying that binocular rivalry is quite sensitive to eye-of-origin information. Thus, for example, conspicuous changes in a rival stimulus can go completely unnoticed for several seconds when those changes are made during suppression phases of rivalry. This nonselective loss of visual awareness implies that suppression is not tailored to a given stimulus interpretation. In addition, exchanging the dominant and suppressed rival stimuli between the eyes immediately and reliably reverses the stimulus that one experiences, indicating that a given region of the eye – not a given stimulus – dominates at any given time. These and other findings imply that binocular rivalry, in fact, emerges from local competition between corresponding regions of the two eyes.

How is this conundrum to be resolved? Perhaps these seemingly incompatible modes of rivalry – eye versus object rivalry – arise from distributed neural processes, subserved by feed-forward and feedback interactions among multiple neural stages within the visual hierarchy. Indeed, this view may also resolve some of the puzzling inconsistencies among physiological results obtained using different methodologies. The next section considers this distributed view in more detail.

Binocular Rivalry as a Distributed Process

Rather than arising from a single omnibus process, binocular rivalry is more usefully considered as an aggregate of distributed processes, each with its own distinct stimulus determinants and neural concomitants and all organized in a recurrent network. According to this view, a full account of rivalry must specify the stimulus conditions that trigger rivalry in the first place and, once rivalry is triggered, what determines initial selection of one of the two alternative competing views. So, for example, is rivalry triggered only by physical differences between the left and right eyes' images, or can rivalry occur when physically similar images nonetheless give rise to alternative perceptual interpretations because, for example, of differing contexts? Once rivalry is initiated, what governs the dynamics of rivalry? In considering these questions, it is crucial to distinguish between the determinants of perceptual dominance (governing the tendency of a given figure to dominate during rivalry) and the determinants of suppression (controlling the average length of time that a given stimulus remains outside of visual awareness). The determinants of dominance and of suppression are not necessarily the same, although both influence fluctuations in perception over time. It now appears that dominance phases of rivalry are functionally equivalent to ordinary monocular viewing, which means that dominance phases during rivalry are subject to the same influences (e.g., context, prior experience, and emotional connotation) that shape vision under nonrivalry conditions. Whether suppression is influenced by those same factors remains to be determined, and an answer to that question will place important constraints on the nature and locus of neural events underlying suppression. Finally, results from psychophysical, neurophysiological, and brain imaging studies point to the following conclusions: (1) the neural events underlying suppression of a stimulus, inaugurated within early stages of visual processing, are not equivalent to the neural events associated with the physical removal of that visual stimulus, even though those two conditions of stimulation are perceptually equivalent; and (2) the neural events underlying dominance are equivalent to those associated with ordinary, monocular viewing.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Structure and Circuitry in Rodents and Felines; Attention and Eye Movements; Brainstem Control of Eye Movements; Cerebellum and Oculomotor Control; Consciousness: Neurophysiology and Visual Awareness in; Contextual Interactions in Visual Perception; Cortical Control of Eye Movements; Electromyography (EMG) and Nerve Conduction Studies; Evoked Potentials: Recording Methods; Executive Functions: Eye Movements and Neuropsychiatric Disorders; Eye Tracking and Mental Illness; Eye and Head Movements; Eye Movement Disorders; Fovea: Primate; Neural Synchrony and Feature Binding; Neuroimaging; Optokinetic Eye Movements; Perception and Eye Movements; Pursuit Eye Movements; Retinal Development: Cell Type Specification; Retinal Models; Saccadic Eye Movements; Sensory Aging: Vision; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Stereoscopic Vision; Vergence Eye Movements; Visual Attention; Visual Motion Models; Visual Motion Detection.

Further Reading

Alais D and Blake R (eds.) (2005) *Binocular Rivalry*. Cambridge, MA: MIT Press.

Blake R and Logothetis NK (2001) Visual competition. *Nature Reviews Neuroscience* 3: 13–21.

Blake R (2006) Overview of binocular rivalry, with demonstrations. http://www.psy.vanderbilt.edu/faculty/blake/Rivalry/BR.html (accessed February 2008).

Diaz-Caneja E (1928) Sur L'alternance binoculaire [On binocular alteration]. *Annales d'Oculistique* 165: 721–731.

Fox R (1991) Binocular rivalry. In: Regan DM (ed.) *Binocular Vision and Psychophysics*, pp. 93–110. London: Macmillan Press.

Freeman AW (2005) Multistage model of binocular rivalry. *Journal of Neurophysiology* 94: 4412–4420.

Haynes J, Deichmann R, and Reese G (2005) Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature* 438(7067): 496–499.

Kovacs I, Papathomas TV, Yang M, and Feher A (1996) When the brain changes its mind: Interocular grouping during binocular rivalry. *Proceedings of the National Academy of Sciences of the United States of America* 93: 15508–15511.

Laing CR and Chow CC (2002) A spiking neuron model for binocular rivalry. *Journal of Computational Neuroscience* 12: 39–53.

Leopold DA and Logothetis NK (1999) Multistable penomena: Changing views in perception. *Trends in Cognitive Sciences* 3: 254–264.

Levelt WJM (1965) *On Binocular Rivalry*. Soesterberg, Netherlands: Institute for Perception RVO-TNO.

Logothetis NK (1998) Single units and conscious vision. *Philosophical Transactions of the Royal Society of London B* 353: 1801–1818.

Mitchell JF, Stoner GR, and Reynolds JH (2004) Object-based attention determines dominance in binocular rivalry. *Nature* 429: 410–413.

O'Shea RP (1999) Translation of Dutour (1763). http://psy.otago.ac.nz/r_oshea/WebPubs/dutour63.html (accessed February 2008).

O'Shea RP (2007) Binocular rivalry bibliography. http://psy.otago.ac.nz/r_oshea/BR_bibliography/br_bibliography.html (accessed February 2008).

Pettigrew JD (2001) Searching for the switch: Neural bases for perceptual rivalry alternations. *Brain and Mind* 2: 85–118.

Polonsky A, Blake R, Braun J, and Heeger DJ (2000) Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neuroscience* 3: 1153–1159.

Rolf H (2006) Stereovision. http://axon.physik.uni-bremen.de/research/stereo/rivalry/ (accessed February 2008).

Tong F, Meng M, and Blake R (2006) Neural bases of binocular rivalry. *Trends in Cognitive Sciences* 10: 502–511.

Tong F, Nakayama K, Vaughan JT, and Kanwisher N (1998) Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* 21: 753–759.

Walker P (1978) Binocular rivalry: Central or peripheral selective processes? *Psychological Bulletin* 85: 376–389.

Wilson HR (2003) Computational evidence for a rivalry hierarchy in vision. *Proceedings of the National Academy of Sciences of the United States of America* 100: 14499–14503.

Biomechanics: Hydroskeletal

Y Yekutieli and T Flash, Weizmann Institute of Science, Rehovot, Israel
B Hochner, Hebrew University, Jerusalem, Israel

© 2009 Elsevier Ltd. All rights reserved.

Type of Skeletal Systems

The combination of muscles (force production mechanisms) and skeletons enable animals to move in a variety of ways. There are basically three types of skeletons: endoskeletons, the rigid internal skeleton of vertebrates; exoskeletons, the rigid external skeleton of arthropods; and hydrostatic skeletons, which are the focus of this article.

Hydrostatic skeletons do not posses rigid elements but rely on pressure to create stiffness. This can be done based on two types of structures. The first involves a fluid-filled cavity (FFC) which is surrounded by muscles and connective tissue. The second structure is composed mainly of muscles without any internal cavity, termed a muscular hydrostat (MH). In some of the literature, the term hydrostatic skeletons is used exclusively to denote the FFC skeletons, whereas in other publications (as here) this term refers to both types: FCCs and MHs.

FFC skeletons can be found in annelids (segmented worms, e.g., earthworms), echiurans, nematodes (roundworms), nemertean (ribbon worms), echinoderms (starfish and sea urchins), and mollusks (e.g., the foot of bivalves and gastropods). The extent and volume of the FFC is variable. Some animals have this internal cavity divided among body segments (annelids, such as earthworms and leeches), whereas in others it is continuous (echiurans and nematodes).

MH skeletons, on the other hand, are neither segmented nor have a fluid-filled extracellular space but are composed of a tightly packed three-dimensional array of muscle fibers. Examples of MH are the vertebrate tongue, the elephant trunk, the arms, the tentacles and fins of cephalopods, the foot of snails, and the feeding apparatus of some mollusks (e.g., *Aplysia*). MH skeletons have the greatest potential for localized, complex, and diverse movements – much greater than in other types of skeletons.

Hydrostatic skeletons are abundant and diverse, ranging in size from minute worms with a scale of few millimeters and grams to the giant squids reaching 20 m in overall length and giant octopuses attaining an arm span of 6 m and weighing over 40 kg. The hydrostatic skeleton of some animals, such as worms, encompasses their entire body, whereas in others it is used only for a specific organ, as in the elephant trunk. A hydrostatic skeleton may be used in combination with a rigid skeletal support (e.g., the vertebrate tongue). In this article we describe the biomechanics and movement control of hydrostatic skeletons, giving examples of the diversity and unique adaptations.

Basic Structure of Hydrostatic Skeletons

Fluid-Filled Cavity Skeletons

Animals with FFC hydrostatic skeletons usually comprise both circular and longitudinal muscle fibers, and their shapes are typically cylindrical. Diagonal elements (either as muscle fibers, connective tissue, or both) are usually an essential component that prevents kinking. We describe the structure of FFC skeletons here using three examples: the earthworm and the leech (both annelids) and *Ascaris* (a nematode).

Annelids have a segmented body in which each segment is separated by septa. Their FFC is a true coelom; that is, the cavity is completely lined with mesoderm-derived tissue. In the Hirudinean (such as the leech), the coelom is dramatically reduced and septa have been lost, making them more of a MH than a FCC skeleton (Figure 1). This is in contrast to our second annelid example, the earthworm, which has a clearly segmented structure and noticeable fluid-filled compartments (Figure 2).

An earthworm has approximately 150 segments, each with circular (circumferential) muscles responsible for radial contraction and longitudinal muscles that shorten the segment. This shortening causes radial expansion based on a mechanism described later in the article. Waves of alternating activation of the longitudinal and circular muscles produce a peristaltic wave, moving from tail to head, which is used for locomotion and burrowing.

In segmented structures, the larger the number of segments, the greater is the potential for localized control of forces and shape changes. The number and dimensions of the segments are also important in determining the leverage system – the amplification of forces or velocities.

In nematodes (the round worms, such as *Ascaris*) there are no circular muscles but, instead, fibrous cuticle that prevents shortening of the animal (Figure 3). They have longitudinal muscles, located at the body wall and arranged in four groups that can be contracted differentially to produce local shortening and bending. The whole interior of the animal, inside the fibrous cuticle, acts as a hydrostatic skeleton. They have a FFC in the form of a pseudocoelom (i.e., the body cavity is not completely lined with mesoderm), in

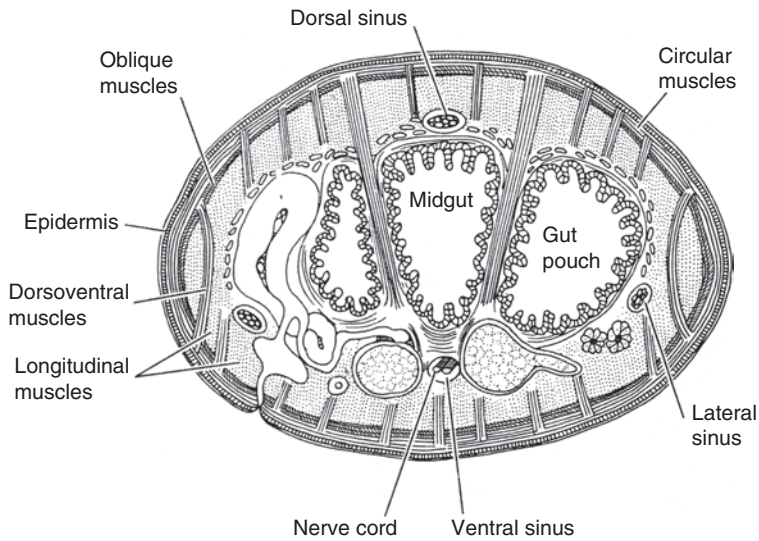

Figure 1 A cross section of a leech (*Hirudo*). The reduction of the coelom results in a body structure that is effectively aceolomate. Note the four different muscle groups: circular, oblique, dorsoventral, and longitudinal. Reproduced from Brusca RC and Brusca GJ (1990) *Invertebrates*. Sunderland, MA: Sinauer, with permission from Sinauer Associates, Inc.

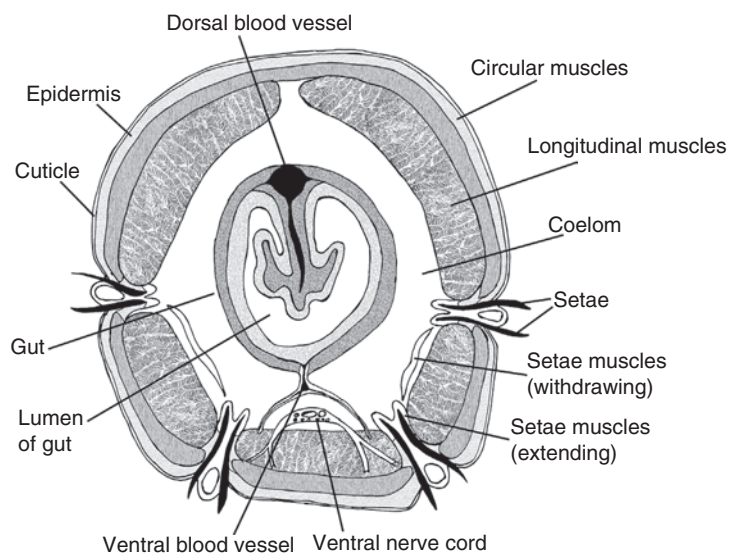

Figure 2 A cross section of an earthworm. The fluid-filled compartments (coelom) are separated by septa (not shown). Each segment contains longitudinal and circular muscles. Reproduced from Wallace RA, Sanders GP, and Ferl RJ (1990) *Biology, the Science of Life*, 3rd edn. New York: HarperCollins.

contrast to the annelids. Their muscle composition allows undulatory motion but not peristalsis, due to the lack of circular muscles.

Muscular Hydrostat Skeletons

MHs are mainly composed of tightly packed muscle fibers that are organized in one or more of three major directions: parallel to the long axis of the organ, perpendicular to the long axis, and wrapped helically or obliquely around the long axis. The different MHs show a large variability in the availability, arrangement, and positioning of the different muscle groups. In some cases, connective tissue is the major structural element in one or more of these directions.

The muscle fibers that are parallel to the long axis are termed longitudinal. Their location varies among different species, ranging from locations mainly near the inner core of the organ and up to the periphery of the organ. The perpendicular fibers may show a transverse pattern made of horizontal and vertical fibers in the core of the organ, as in the mammalian tongue, squid arm and tentacle, and octopus arm (Figure 4). A different arrangement is found in the

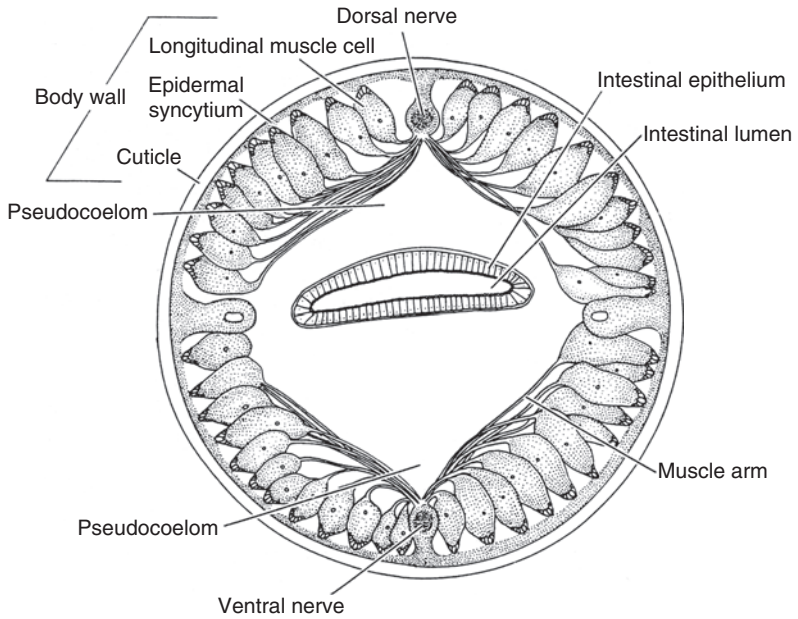

Figure 3 Stylized cross section of a nematode such as *Ascaris*. All body-wall muscles are obliquely striated longitudinal muscles arranged in four groups. The muscle cells are connected to the ventral or dorsal nerve cords by muscle arms – an extension of the muscle cells and not by neurons. Reproduced from Brusca RC and Brusca GJ (1990) *Invertebrates*. Sunderland, MA: Sinauer, with permission from Sinauer Associates, Inc.

elephant trunk and the tentacles of the chambered nautilus; there the muscles are attached to connective tissue in the center and radiate toward the periphery. Another type of perpendicular organization, that of circular muscles is found in most lizard tongues, in some mammalian tongues, and in the squid tentacles.

The third group of muscle fibers is wrapped helically around the long axis and arranged in opposite directions. This is done based on the combination of muscle fibers and connective tissue to complete the helix (in the octopus arm or the monitor tongue) or using only muscles (as in the squid tentacle).

Biomechanics of Hydrostatic Skeletons and Comparison to Rigid Skeletons

The rigid skeleton (either internal or external) serves some important functions in the generation of movement and the maintenance of posture. The skeleton provides the stiffness needed to resist external and internal forces and supports other organs. Muscles can only exert a pulling force and are unable to push themselves back into their original positions. The rigid skeleton enables the return of muscles to their working regime and the reversibility of shape changes via muscle antagonism, an arrangement of two or more muscles connected to opposing sides of a joint. The relative location of the anchoring points between muscles and bones allow for the amplification of force versus displacement and speed using leverages. In hydrostatic skeletons, these functions are carried out with no rigid elements but, instead, by muscles, connective tissue, and the incompressibility of fluid.

Constant Volume and Muscle Antagonism

The biomechanics of FFC skeletons is based on pressurizing the structure using muscle contraction and resistance of other muscles or connective tissue. Because fluids are incompressible, if the rate of flow out of the cavity is small the enclosed FFC is practically constant in volume. Any change in one dimension will change the pressure in the entire volume and will be compensated for by a change in at least one other dimension. For example, contraction of circular muscles decreases the width of the structure and, as a result of the constant volume constraint, increases its length. This in turn lengthens the longitudinal muscles, which can be later contracted to shorten and widen the structure. Thus, different muscle groups of different fiber orientations act antagonistically.

In MH skeletons, the fluid that maintains a constant volume is contained within the densely packed muscle cells and not in a distinct extracellular cavity. Nevertheless, the biomechanic principle is similar to those of FFC skeletons. Any change in one dimension will cause a compensatory change in at least one other dimension. As in FFC skeletons, the contraction of longitudinal muscles shortens and widens the structure. Contraction of the circular or transverse muscles reduces the diameter and elongates the organ.

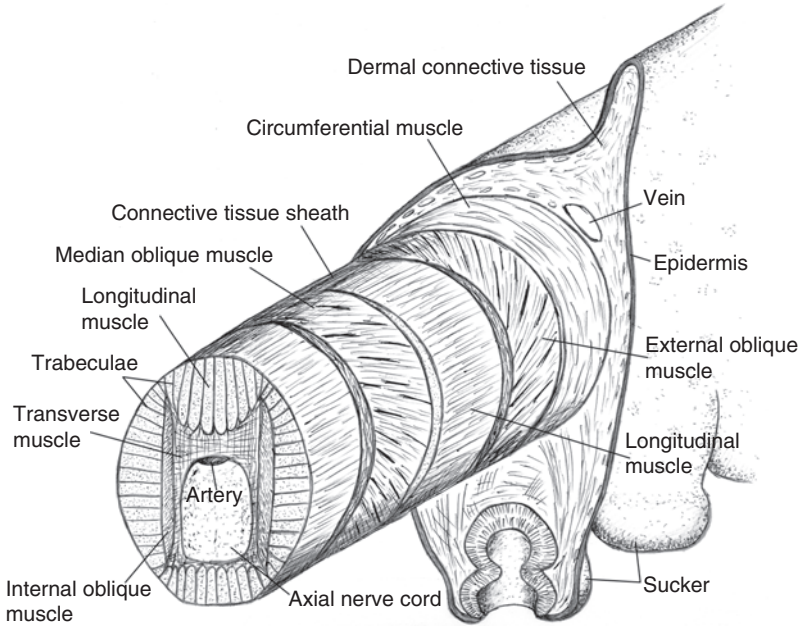

Figure 4 The musculature and connective tissue of an octopus arm as an example of a muscular hydrostat. Reprinted from Trueman ER and Clarke MR (eds.) *The Mollusca (11): Form and Function,* New York: Academic Press, Kier WM, The arrangement and function of molluscan muscle, pp. 211–252, Copyright 1988, with permission from Elsevier.

Leverages

The relation between the length and diameter of elongated MH and FFC structures enables the amplification of force or speed, resulting in a leverage system. Here we approximate these structures by a constant-volume cylinder. When the structure is long (the ratio of length to width is large), a small one-unit change in diameter creates a large change in length (Figure 5, from C to D), thus amplifying the displacement and speed. In the other extreme, when the structure is short (has a small length to width ratio), the same one-unit change in diameter results in a much smaller increase in length (Figure 5, from A to B), amplifying the force. Because the length to width ratio itself changes with length, such a system has a variable leverage, depending on the length to width ratio the movement is carried in. This is used in the squid tentacle and some lizards' tongues to protrude the organ very rapidly when these animals catch prey.

Bending

Bending in hydrostatic skeletons is the outcome of differential activation of longitudinal muscles in two sides of the organ and some resistance to compression. The location of the longitudinal fibers at the outer regions of the structure, as found in the elephant trunk, the mammalian tongue, and the octopus arm, creates larger moments. In contrast, tongues that are used for protrusion have their major longitudinal bundles in the center of the organ.

Helical Fibers

Some MH skeletons have helical fibers, and their arrangement and orientation have important implications. Naturally they function in torsional movements in which the organ is twisted along its long axis. But their contribution is not restricted to creating torsion or resisting torsional forces. Depending on their angle relative to the long axis, the contraction of the helical fibers can either elongate the structure (when the angle is less than $54^\circ 44'$) by functioning as circular muscles that decrease diameter or shorten the structure (when the angle is more than $54^\circ 44'$) by functioning as longitudinal muscles.

Stiffness

Generating stiffness is very important in providing structural stability against internal and external forces and in shaping the organ. Stiffness in MHs is achieved by co-contraction of the different muscle groups. When the longitudinal muscles exert the same force as the transverse (or circular) fibers, pressure builds up to stiffen the structure. In MHs such as the octopus arm, this activation scheme not only increases its ability to withstand internal and external forces but also straightens the activated region of the organ. Helical fibers, if present, can also contribute to stiffening of the structure by simultaneous activation of the right-handed and left-handed fibers. This does not result in torsion movement but in increased pressure and in a tendency of the structure to achieve and maintain a stable shape.

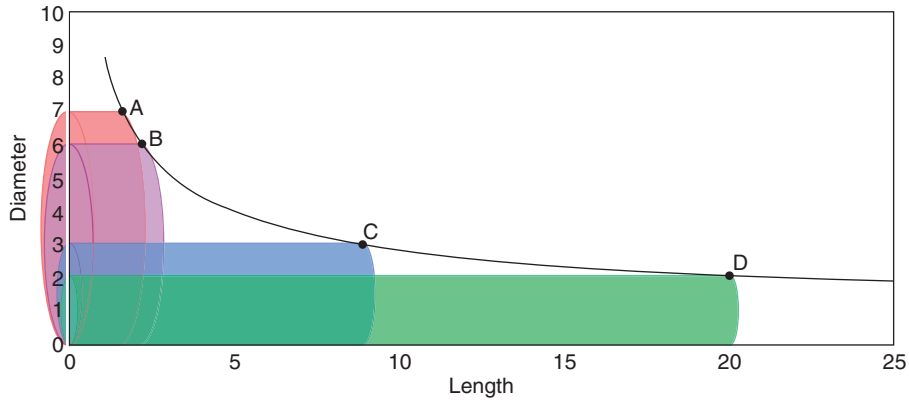

Figure 5 The relationship between diameter and length in a fixed volume cylinder is not linear. A unit decrease in diameter when the cylinder is short and wide (from A to B) results in a small increase in length of about 0.6 units. A unit length change applied on a longer and thinner cylinder (from C to D) increases length by more than 10 units. Reproduced from Kier WM and Smith KK (1985) Tongues, tentacles and trunks: The biomechanics of movement in muscular-hydrostats. *Zoological Journal of the Linnaean Society* 83: 307–324, with permission from Blackwell Publishing Ltd.

This biomechanical organizational principle may be seen as a way to simplify control; irrespective of the current shape of the organ, equal and concurrent activations of all muscle groups shape the organ to a standard shape with a degree of stiffness that is dependent on the level of activation. This standard shape is the shape in which all muscle forces are balanced, and its actual dimensions depend on the relative number of muscle fibers in each orientation and their contractile properties.

A similar mechanism can be found in FFC skeletons. Here, too, body stiffness is the result of co-contraction of antagonistic muscles. Moreover, in the leech, contracting the oblique muscles contributes to the stiffness, as well as to the achievement of an intermediate body length, similar to the effect of the helical muscles of MHs.

Interestingly, controlling stiffness by co-contraction of muscles is a known phenomenon for rigid skeleton structures. There, activating of the antagonist muscles around a joint has a stabilizing effect. But this is confined, of course, to stiffness control around a given joint. MHs and some FFCs, on the other hand, have the ability to create a dynamic skeleton by producing stiffened sections at various parts along the structure and by controlling the amount of stiffness of these sections. This is something that is impossible to achieve in rigid skeletons, in which the rigid elements are given and do not change their mechanical characteristics or location along the organ at will.

Localization of Movement

How fine is the ability to perform local movements? This depends on the organization of the biomechanical system, as well as the control system. In segmented structures such as the earthworm, the single segment is the smallest functional unit of movement generation. Naturally a larger number of segments results in a finer ability to perform localized and complex movements. Unsegmented FFCs are limited by the same factor – their one internal FFC transmits forces via its pressure to all the muscles and connective tissue surrounding it. MHs, on the other hand, have the potential for a fine-tuned local control. Their actual ability depends on the size of muscles, the spatial resolution of muscle activation, the capability of controlling muscles independently, and the sensory systems relaying information about their movements. The great degree of localization can be appreciated by considering the arm of *Octopus vulgaris*. It has 3.8×10^5 motor neurons per arm, with muscle cell dimensions of $\sim 0.01 \times 1$ mm, motor unit size of $\sim$3300 muscle cells (occupying a volume of $\sim 0.2\ \text{mm}^3$, which is $\sim$10 ppm of the entire arm) and localized mechanoreceptors. This complexity of anatomy and physiology, combined with a sophisticated motor control system, is the infrastructure for the amazing motor abilities of the octopus.

Mechanical Degrees of Freedom

Related to the issue of localization of movement and an important factor in describing the biomechanics and the motor control system is the number of degrees of freedom (DOFs) of the actuators. The number of mechanical DOFs for rigid skeletons depends on the number and type of their joints. MHs are continuous and flexible structures, and it is difficult to estimate their (obviously huge) number of DOFs or even to use this term to describe such systems. Anyhow, the potential repertoire of movements for an elephant trunk, an

octopus arm, or a tentacle of a squid is vast. Borrowing terminology from robotics, these can be considered as hyperredundant manipulators, meaning that they have many more DOFs than the minimal number needed to perform tasks in space. Planning and executing movement in hyperredundant manipulators is a real challenge in human-made machines. Studying how this is done in animals with MH skeletons may provide important insights for roboticists in the quest to advance technology and allow the production of truly flexible robots. (Later in this article, we describe some of the solutions to motion planning and control problems the octopus uses to deal with the complexity of controlling its arm.)

Variation on a Theme

Following are some examples of different behaviors and typical movements performed by flexible structures with hydrostatic skeletons with their specific adaptations.

Peristalsis and Burrowing

Peristalsis is a common mode of locomotion of many soft-bodied animals (e.g., earthworms, sea cucumbers, caterpillars, and snails). Alternating waves of muscle contraction and relaxation move along the body, anchoring the body to the substratum only in widened regions (Figure 6). In contrast to rigid skeleton animals, in this mode of locomotion there is no change in joint angles between rigid links but, instead, changes in the dimensions of body regions. A region may shorten while increasing in diameter or elongate and contract radially. The changes of body dimensions, especially in FFC skeletons, imply the flow of fluids, increasing the importance of viscous damping forces and decreasing the role of natural frequencies in shaping motion when compared to rigid articulated skeletons.

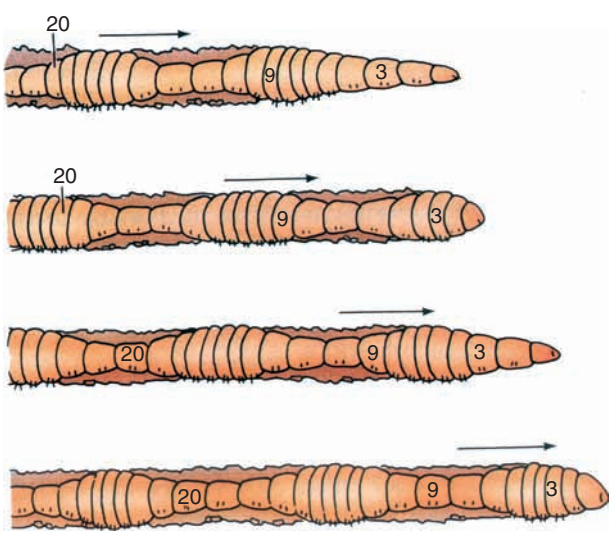

Figure 6 Peristaltic movement of an earthworm through (or over) the soil is achieved by alternating waves of muscle contraction and relaxation that move along the body. Reproduced from Purves WK, Orians GH, and Helle HC (1994) *Life: The Science of Biology*, 4th edn. Sunderland, MA: Sinauer, with permission from Sinauer Associates.

Peristaltic movements are also used in burrowing. *Urechis caupo* (of the phylum Echiura, unsegmented worms related to the annelids) is a worm specialized for burrowing that lives in U-shaped burrows in marine mud flats. It has longitudinal and circular muscles surrounding a large FFC. It uses peristaltic contractions to pump water through the tube and retrograde contractile waves to move around in its tube and to enlarge its tube by burrowing through the substratum. This body plan can be compared to the segmented body of the earthworm, another burrowing animal; segmentation isolates the effects of muscle contraction, enabling much more localized control of shape and movement and conservation of energy.

Burrowing is also a common behavior of many vertebrates, and unique among them are the caecilians. These are limbless amphibians, found throughout most of the humid tropics, that maintain their body shape during locomotion using hydrostatic pressure. They have obliquely oriented body wall muscles and a vertebral column that moves independently of the skin. This unique morphology enables the generation of a powerful forward force, approximately twice the force found in burrowing snakes of similar size. The snakes rely solely on longitudinal muscles and vertebral column for forward force production, in contrast to the caecilians, which apply force to a crossed-helical array of tendons that surrounds their body cavity, using it as a hydrostatic skeleton. Caecilians are the first vertebrates known to use the entire body as a hydrostatic system for locomotion.

Specialized and Generalized Muscular Hydrostats

Some of the elongated MH structures are generalized and some specialized. This distinction comes from considering the type of movements done by the flexible organ and its structure. The squid tentacle is a remarkable example of specialization. It is a tool specifically constructed for the rapid elongation characterizing the protrusion movement used to catch its prey (70–100% elongation in 20–40 ms, velocities of up to 2 m s^{-1}, and accelerations of up to 250 m s^{-2}). The tentacle is composed of a muscular stalk and a terminal club equipped with arrays of suckers used to grasp the prey. Various adaptations have been described. The transverse and circular muscles fibers are much more numerous than the longitudinal muscle fibers (they occupy more than

60% of the cross-sectional area, compared to 45% in the squid arm). They exhibit a cross-striated pattern typical of fast-contracting muscle cells and differ from other muscles of both squid and cephalopods in general. The tentacle muscle fibers achieve maximal velocity of shortening of 15.5 $L_0\,s^{-1}$(where L_0 is the length at which active isometric muscle force is maximal; it is usually a bit more than the relaxed muscle length), compared to 1.5 $L_0\,s^{-1}$ in arm muscle fibers. The sarcomere structure is especially built for rapid contraction, with almost 10 times shorter thick filaments. A mechanical lever that amplifies speed is also used with a length to width ratio of 20:1 in the resting squid tentacle. When both transverse and circular muscle fibers contract, the result is a rapid increase in length. Approximately 70% elongation of the tentacle is achieved with only a 23% decrease in diameter. The biomechanical adaptations for this fast-protruding mechanism result in a structure that is prone to buckling. This is compensated for by extrinsic support of two of the eight arms in the initial rapid extension phase during prey capture.

The squid protrusion movements and the adaptations allowing this fast movement were studied by various means: using high-speed filming to describe the movement kinematics, studying the muscle ultrastructure and physiology, and using mathematical modeling. The modeling showed that the unusually short myosin filaments of squid extensor muscles are necessary to create the rapid elongation behavior.

In tongues adapted for protrusion, as found in some lizards, in snakes, and in mammalian anteaters, most of the longitudinal muscles are located centrally. In contrast, MHs that are designed for manipulation, such as the elephant trunk, octopus arms, and intro-oral tongues, have smaller length to width ratios; their cross sections show similar amounts of transverse and longitudinal fibers; and their longitudinal muscles are located at the periphery of the organs, providing better leverage for bending.

The tongues of advanced snakes have an interesting subdivision of the arrangement of muscles. At the proximal portion of the tongue, the longitudinal muscles are arranged in two large bundles surrounded by four sheets of transverse muscles – an organization that facilitates protrusion. At the distal parts of the tongue, the longitudinal muscles are arranged in bundles at the periphery with the transverse at the middle – enabling bending.

Some MHs, such as the tongue of the frog *Hemisus*, are composed of muscles fibers oriented in only two directions: longitudinally and vertically. The third dimension has connective tissue but no muscles. The width of the tongue remains constant due to a transversely aligned connective tissue capsule that surrounds the dorsoventral muscle fibers. When the vertical fibers contract, the tongue protrudes with a linear increase in length compared to the geometric increase in the squid tentacle. This difference might partially explain why squid tentacles have higher acceleration ($250\,m\,s^{-2}$) than the tongues of *Hemisus* ($3.5\,m\,s^{-2}$).

Motor Control of Hydrostatic Skeletons

Next we give examples from some of the more advanced studies done on the motor control of hydrostatic skeletons in different animals. Among them are the octopus arm, the tongues of frogs, and the whole body of the leech. Many important systems are not described here due to lack of space (e.g., the control of the human tongue, the feeding mechanism in *Aplysia*, and the tongue protractions of salamanders and chameleons).

The Octopus Arm

The octopus is a solitary animal and an active predator. It uses its eight arms for various motor tasks such as locomotion (swimming and walking), object manipulation (building shelters and pulling or pushing objects), and catching and grasping prey. It can exert great force (as when opening a clamp) and shows immense dexterity (opening jar lids). Octopus movements provide an extreme example for the potential complexity of motor control because the arm can move in any direction from any point along the arm, operating well in constrained and nonpredictable environments.

The motor control of the octopus is divided between the brain and the arms. The brain is composed of 15×10^7 nerve cells, whereas the nervous system of the arms contains 4.75×10^7 neurons in each arm, of which $\sim 3.8 \times 10^5$ are motor neurons, $\sim 2.3 \times 10^6$ are sensory receptors, and 32 000 axons connect the arm and the brain. The majority of the arm nervous system ($\sim 4.5 \times 10^7$ neurons) consists of interneurons. The size of the arm nervous system, the dominance of interneurons within it, and the small number of fibers running between it and the brain suggest a high level of autonomy in the neuromuscular systems of the arms. Indeed, this is supported by the complex local motor reflexes evoked by pain, tactile, or chemical stimuli to the arm. The division of motor control between this autonomous system and the brain suggests a hierarchical organization for the motor control system.

The reaching movement of the octopus is its stereotypical way of extending an arm toward a target by propagating a bend along the arm (**Figure 7(a)**). First, a bend is created somewhere along the arm. Then the base of the arm is directed toward or slightly

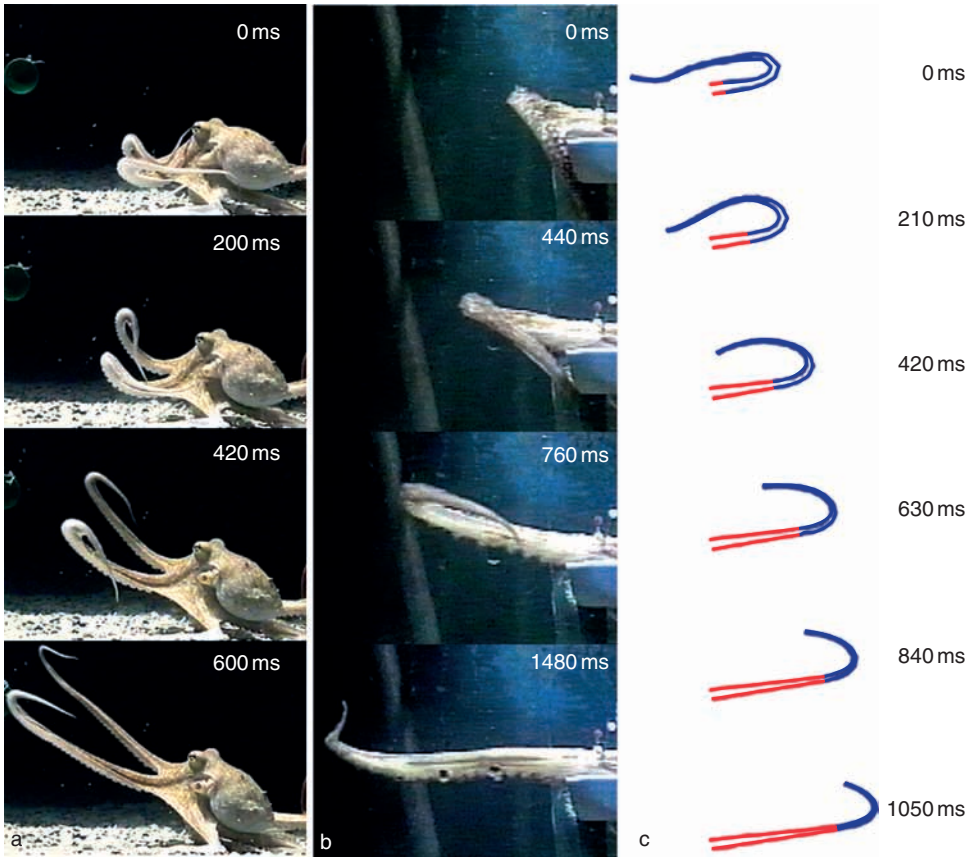

Figure 7 The reaching movement of the octopus: (a) a freely behaving octopus reaching toward a target using two of its arms; (b) a similar movement electrically evoked in an arm whose neural connections to the brain were severed; (c) sequence of images from a simulation of arm extension using a dynamic model of the octopus arm. The input to the model was a simple signal of muscle activation propagating from the base of the arm to its tip. Red represents areas of the arm in which all muscles are fully activated; blue represents areas that are not activated. Reproduced from Yekutieli Y, Sagiv-Zoahr R, Aharonov, R, et al. (2005) Dynamic model of the octopus arm, I: Biomechanics of the octopus reaching movement. *Journal of Neurophysiology* 94: 1443–1458, used with permission from The American Physiological Society.

above the target and the arm is straightened until it reaches the target (or misses). When the arm of an octopus touches a target, usually with the suckers, it wraps around the target and brings it back to the mouth (Figure 8).

Kinematic studies of the reaching movement showed that the bend follows a slightly curved planar path. Its velocity has a bell-shaped profile invariant to the direction and the duration of motion. Electromyography (EMG) studies of this movement revealed a characteristic muscle activation wave that moves along the arm and precedes the bend. Moreover, the amplitude of the EMG signal measured at the beginning of the movement was best correlated with the maximal velocity of subsequent bend propagation, meaning that it is possible to predict the kinematics of this movement based on the initial EMG signal. In a denervated arm preparation (whose nervous system was disconnected from the brain), electrical stimuli to the axial nerve cord of the arm or mechanical stimuli to the skin can evoke extension movements similar to natural reaching movements (Figure 7(b)). In the same preparation, the initial posture of the arm determined the direction of the path of the evoked movements.

The control of this movement is characterized by two important principles. The first is the reduction in complexity by the coupling among different DOFs. This is based on the observation that the reaching movement kinematics is dictated by three parameters: two angles in space to control the direction of the base of the arm and a third parameter to control the propagation of the bend along the arm. The second principle is hierarchical organization demonstrated by the division of labor between the central nervous system (presumably controlling the orientation of the base and sending commands to the arm neural network) and the periphery – the neuromuscular system of the arm that executes the movement in an autonomous way.

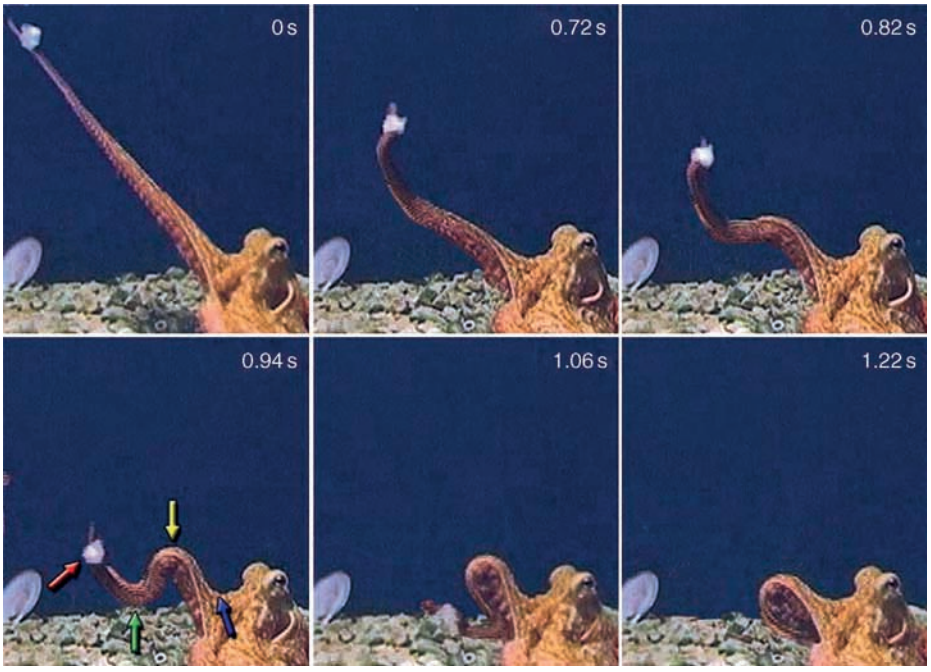

Figure 8 A quasi-articulated structure is formed during the fetching movement of the octopus. Red arrow marks the location of the food; green, yellow, and blue arrows mark the distal, medial, and proximal bends, respectively. Reprinted from *Current Biology,* 16, Sumbre G, Fiorito G, Flash T, and Hochner B. Octopuses use a human-like strategy to control precise point-to-point arm movements, pp. 767–772, Copyright 2006, with permission from Elsevier.

A forward dynamic model of the octopus arm was used to study the relation between biomechanics and control. The simulations supported the stiffening wave hypothesis, which states that a simple wave of muscle contraction, propagating distally from the base of the arm, can push forward a bend in the arm and straighten it, creating an extension movement (Figure 7(c)). It was shown that water drag forces have a major role in shaping octopus arm movements and that the extension movement is not only a way to simplify control but also an efficient way to move the arm in water. Symmetrical and simultaneous activation of all muscles of the modeled arm was shown to straighten the arm and bring it to the same standard shape irrespective of its initial configuration. Last, dynamic modeling showed that it is possible to control arm stiffness independently of arm kinematics during an extension movement. This stiffness may serve to resist external perturbing forces while performing desired movements. Whether this strategy is actually used by the octopus is not known.

Bringing food to the mouth is achieved by a fetching movement (Figure 8), in which three bends acting like joints are formed along the arm. The arm functions then quite similarly to an articulated structure, with the part of the arm holding the food and bringing it accurately to the mouth. This stereotypic movement also shows a reduction in complexity by coupling DOFs to form a small number of control parameters – the angles of the three joints. Measuring the EMG signals during this movement showed that prior to the formation of the medial joint, two waves of muscle contraction propagate toward the middle, one from the base of the arm and the other from the area holding the food. This observation provides evidence for a mechanism allowing the establishment of the position of the medial joint.

Octopuses use several modes of locomotion. Their fastest mode (also common in squids) is jet propulsion, in which the muscles on the mantle wall contract and squeeze water through their narrow funnel. Benthic species also crawl on the surface, using all or most of their arms. Recently, it was reported that some octopus species use bipedal walking, combining locomotion with the ability to remain disguised. Analyzing these movements revealed an alternating pattern of bend propagation moving along the two walking arms. Each arm in its turn is brought ahead by rotating its base forward. Pushing the octopus forward is achieved by a rotation of the base backward while the part of the arm that touches the surface propagates distally in a form of a bend.

Using muscles to create joints is not unique to elongated MHs such as the octopus and squid arms or the elephant trunk. A novel form of flexible joint mechanism was recently described for the buccal

mass of the octopus, the cuttlefish, and the squid. Their beaks are capable of complex movements using a unique muscular-hydrostatic mechanism to control the location of the pivot between the upper and lower beaks and to generate the force required for beak movements. All this is performed with no articulation of the beaks – that is, they do not have any contact in the form of a skeletal joint.

Tongue Protrusion in Frogs

Tongue protrusion in frogs is done by three main mechanisms, of which only one is based purely on hydrostatic skeleton biomechanics. We briefly describe here these mechanisms and show the relation between the hydrostatic mechanism used to protrude a tongue and other mechanisms from which it probably has evolved.

The first mechanism, and probably the most primitive one, is mechanical pulling. The tongue shortens and the genioglossus muscle (one of the major tongue muscles) contracts, pulling the tongue pad forward over the mandibular symphysis. The tongue is protruded only modestly and unless a prey item is very close, the frog has to lunge forward to catch the prey.

The second mechanism, and the most common one (in terms of the number of species using it), is inertial elongation. Angular momentum generated by the opening of the mouth is transferred to the tongue and elongates it. It requires a tight coordination of tongue and jaw movements that flips the tongue over the mandibles, extending it up to 180 % of its rest length in a very fast, ballistic movement (2.5–4 $m\,s^{-1}$ with accelerations of more than 310 $m\,s^{-2}$). The amount of elongation and precision of movement allow the animal to remain relatively stationary during feeding time. However, frogs using this mechanism are unable to correct the tongue direction during movement.

The third mechanism is hydrostatic elongation, in which the genioglossus muscle has parallel fibers (genioglossus longitudinalis) as in other frogs and also unique vertical fibers (genioglossus dorsoventralis). When the vertical fibers contract, tongue thickness decreases and, due to the constant volume, the length increases. As in inertial elongation, the tongue rotates forward over the mandibular symphysis, but it can also be aimed laterally and in elevation relative to the head. Tongue protrusion in some hydrostatic elongators (e.g., *Hemisus*) is too slow for the inertia to play a role in the movement, whereas others (e.g., *Dyscophus*) protract their tongues fast enough so that inertia may play a significant role in addition to hydrostatic elongation. There appears to be an accuracy versus speed trade-off in hydrostatic elongators.

The force production of protractor muscles in the different frog species cannot explain the differences among these three mechanisms because it differs by only 20–30%. Several other biomechanical and anatomical differences were described.

Mechanical pullers have the largest number of collagen fibers parallel to the long axis, resisting elongation. Inertial elongators have fewer fibers and hydrostatic elongators have the fewest. In *Hemisus*, a hydrostatic elongator, most of the collagen fibers are oriented perpendicularly to the long axis of the tongue.

Other differences exist with regard to the motor control of movement. Inertial elongators use fast, ballistic, movements with feed-forward control to coordinate jaw and tongue movements. They need a precise coordination of tongue and jaw to achieve accurate and successful movements. In contrast, mechanical pullers and hydrostatic elongators have slower movements and use both feed-forward and feedback control of tongue movements. They do not rely on the precise coordination of tongue and jaw. In mechanical pullers, the tip of the tongue is only a few millimeters out of the lower jaw, and in hydrostatic elongators, the tongue movements do not depend on inertia for elongators and thus are controlled independently of the jaw.

These differences are manifested anatomically. Mechanical pullers and hydrostatic elongators lack hypoglossal afferents (carrying proprioceptive information about the tongue) that modulate the phase of activity in the jaw muscles. This is in contrast to inertial elongators, in which it appears that these fibers have evolved convergently at least four times independently.

There is also a difference in the size of the motor units. In inertial elongators (*Bufo marinus* and *Rana pipiens*), approximately 250 motor neurons innervate the genioglossus muscle, compared to approximately 950 motor neurons for the same muscle in a hydrostatic elongator (*H. marmoratum*). This may serve for a more localized and diverse tongue movement and for sequential recruitment to support the tongue against gravity in slow and gradual protraction.

Leech Motor Control

Many aspects of the control of movement in the medicinal leech were investigated in great detail, including the control of its hydrostatic skeleton. The leech is a member of the annelids whose coelom is dramatically reduced and septa have been lost, making them more of a segmented MH than a FCC skeleton.

The leech has four types of muscles used for overt movements: longitudinal, circular, oblique, and dorsoventral (**Figure 1**). The four muscle activations produce

different movements: the contraction of the longitudinal muscles produces shortening, circular muscle contractions produce a reduction in cross section and elongation; contraction of the oblique muscles cause stiffening at an intermediate body length, and dorsoventral muscle contractions cause a flattening of the body and contribute to elongation. The innervation pattern of the muscles allows fine control of shape and movement. The motor neurons are either excitatory (using acetylcholine (ACh) as the neuromuscular transmitter) or inhibitory (using γ-aminobutyric acid (GABA)). Each motor neuron connects to a subset of muscle fibers of a single type in a specific region of one segment only (e.g., the longitudinal muscle fibers on the right dorsoventral side). The leech has stretch receptors embedded in the dorsal, lateral, and ventral body walls of each segment.

Figure 9 shows four distinguishable behaviors that leeches perform by combinations of lengthening, shortening, and bending using a characteristic temporal and spatial pattern of muscle contraction. As

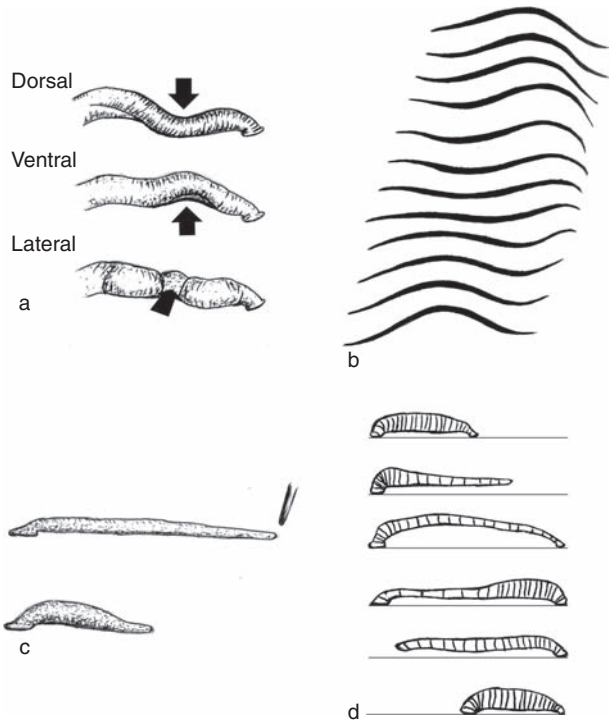

Figure 9 Example of four leech behaviors: (a) local bending as a response to light touch on different body sides; (b) swimming using undulatory movements of the whole body (swimming direction is to the left); (c) rapid shortening of the whole body as a response to touch; (d) crawling by the coordination of body elongation and shortening with sucker attaching and detaching (crawling direction is to the right). Reprinted from Kristan WB Jr., Calabrese RL, and Frissen O (2005) Neuronal control of leech behavior. *Progress in Neurobiology* 76, pp. 279–327, with permission from Elsevier.

described for other hydrostatic skeletons, the internal pressure functions as the mechanism for the transfer of forces between the different muscle groups, rendering them antagonist muscles.

Some of the behaviors described are rhythmic and produced by central pattern generators (CPGs) and can be reproduced in the isolated nervous system. These include swimming (Figure 9(b)), crawling (Figure 9(d)), and feeding. Other movements such as local bending (Figure 9(a)) and shortening (Figure 9(c)) are elicited in response to a well-defined stimulus in a reflexive way.

Local bending in response to light touch is produced by the contraction of longitudinal muscles on the side of the touch and relaxation of the other side. The response is mainly based on a feed-forward mechanism of a simple network that maps sensory locations on to motor space. This mapping resembles the computations described for other systems, such as eye movements in mammals, arm movement in primates, escape responses in crickets and cockroaches, and flight orientation in insects.

As in other hydrostatic skeletons, shortening is caused by the simultaneous contraction of all longitudinal muscles. This can be either localized to a small number of segments or span the whole body.

Swimming in the leech is an undulatory movement of a flattened body shape. The flattening is achieved by tonic contracting the dorsoventral muscles, forming a paddlelike structure that is better for propelling the body in water. Then, alternating contraction and relaxation of dorsal and ventral longitudinal muscles produce rhythmic bending of the body. An isolated nerve cord preparation is capable of producing rhythmic bursting in the motor neurons. The cycle period is longer than in the actual swimming, and the intersegmental phase lags are reduced. This indicates the role of sensory information and feedback loops. Stretch receptors in the leech body wall are the mediators of the sensory information.

The crawling of the leech involves a more complex movement than the peristaltic locomotion of the earthworm. The leech uses front and back suckers to attach to the surface, and their activity must be coordinated with the extension and contraction of the body. Elongation of each segment is produced by the contraction of the circular muscles and contraction of all longitudinal muscles within that segment. Measuring the activity of the motor neurons of a single segment during crawling reveals an alternation of bursting of the circular and longitudinal motor neurons. The same pattern of activity has been reproduced in semi-intact and isolated nervous preparations. As in swimming, this fictive motion had longer cycle periods than the intact animals.

The Role of Modeling

Modeling is an essential component of our understanding of the hydrostatic skeletons' dynamics and control. Combined with other experimental techniques, it is a powerful tool. The complexity of biomechanics and control of hydrostatic skeletons, especially of MHs, require detailed and quantitative models of these systems. Such models were used to test ideas and predict aspects of the design of the system (sarcomere of the squid); confirm assumptions about the biomechanics (stiffening wave hypothesis of the reaching movement in the octopus); show the relative significance of the biomechanical design (the forward model of frog tongue protrusion); illuminate the relative importance of the different forces (the drag forces in the reaching movement of the octopus); and allow us to investigate the interaction of dynamics, sensory information, and neuronal control (leech bending and crawling).

Modeling should be based on accurate and reliable measurements of the different properties of the system: architectural, anatomical, and physiological (e.g., tissue properties; muscle fiber orientation, composition, and forces; EMG; and neuronal signals). It must be based on kinematic and dynamic movement data (e.g., positions, velocities, and forces). This calls for the development of new tools for data acquisition and description, as well as ways to incorporate and integrate multiple-level data with modeling techniques.

See also: Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Computational Approaches to Motor Control; Development of Drosophila Neuromuscular Junctions; Kinematics and Dynamics; Mechanoreceptors; Motor Control of Feeding and Drinking; Motor Primitives; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Pattern Generation; Population Coding; Presynaptic: Mitochondria and Presynaptic Function; Reaching and Grasping; Representation of Movement; Swim Oscillator Networks; Swimming: Neural Mechanisms ; Walking in Invertebrates.

Further Reading

Brusca RC and Brusca GJ (1990) *Invertebrates*. Sunderland, MA: Sinauer.

Gutfreund Y, Flash T, Fiorito G, and Hochner B (1998) Patterns of arm muscle activation involved in octopus reaching movements. *Journal of Neuroscience* 18: 5976–5987.

Gutfreund Y, Flash T, Yarom Y, Fiorito G, et al. (1996) Organization of octopus arm movements: A model system for studying the control of flexible arms. *Journal of Neuroscience* 16: 7297–7307.

Huffard CL, Boneka F, and Full RJ (2005) Underwater bipedal locomotion by octopuses in disguise. *Science* 307: 1927.

Kier WM (1988) The arrangement and function of molluscan muscle. In: Trueman ER and Clarke MR (eds.) *The Mollusca (11): Form and Function*, pp. 211–252. New York: Academic Press.

Kier WM and Smith KK (1985) Tongues, tentacles and trunks: The biomechanics of movement in muscular-hydrostats. *Zoological Journal of the Linnaean Society* 83: 307–324.

Kier WM and Thompson JT (2003) Muscle arrangement, function and specialization in recent coleoids. *Berliner Paläobiologische Abhandlungen* 3: 141–162.

Kristan WB Jr, Calabrese RL, and Friesen WO (2005) Neuronal control of leech behavior. *Progress in Neurobiology* 76: 279–327.

Matzner H, Gutfreund Y, and Hochner B (2000) Neuromuscular system of the flexible arm of the octopus: Physiological characterization. *Journal of Neuroscience* 83: 1315–1328.

Neustadter DM, Drushel RF, Crago PE, et al. (2002) A kinematic model of swallowing in *Aplysia californica* based on radula/odontophore kinematics and *in vivo* MRI. *Journal of Experimental Biology* 205: 3177–3206.

Nishikawa KC (1999) Neuromuscular control of prey capture in frogs. *Philosophical Transactions of the Royal Society of London, Biological Sciences* 354: 941–954.

Purves WK, Orians GH, and Helle HC (1994) *Life: The Science of Biology*, 4th edn. Sunderland, MA: Sinauer.

Sumbre G, Fiorito G, Flash T, and Hochner B (2005) Motor control of flexible octopus arms. *Nature* 433: 595–596.

Sumbre G, Gutfreund Y, Flash T, Fiorito G, and Hochner B (2001) Control of octopus arm extension by a peripheral motor program. *Science* 293: 1845–1848.

Uyeno TA and Kier WM (2005) Functional morphology of the cephalopod buccal mass: A novel joint type. *Journal of Morphology* 264: 211–222.

Van Leeuwen JL, de Groot JH, and Kier WM (2000) Evolutionary mechanics of protrusible tentacles and tongues. *Netherlands Journal of Zoology* 50: 113–139.

Van Leeuwen JL and Kier WM (1997) Functional design of tentacles in squid: Linking sarcomere ultrastructure to gross morphological dynamics. *Philosophical Transactions of the Royal Society of London, Series B* 352: 551–571.

Walker ID, Dawson DM, Flash T, et al. (2005) Continuum robot arms inspired by cephalopods. In: Gerhart GR, Shoemaker CM, and Gage DW (eds.) *Proceedings of the SPIE, Vol. 5804: Unmanned Ground Vehicle Technology VII*, Orlando, FL, 29 March. pp. 303–314, Bellingham, WA: SPIE.

Wallace RA, Sanders GP, and Ferl RJ (1990) *Biology, the Science of Life*, 3rd edn. New York: HarperCollins.

Yekutieli Y, Sagiv-Zohar R, Aharonov R, et al. (2005) Dynamic model of the octopus arm, I: Biomechanics of the octopus reaching movement. *Journal of Neurophysiology* 94: 1443–1458.

Yekutieli Y, Sagiv-Zohar R, Hochner B, and Flash T (2005) Dynamic model of the octopus arm, II: Control of reaching movements. *Journal of Neurophysiology* 94: 1459–1468.

Relevant Website

http://www.wormatlas.org – WormAtlas (database of behavioral and structural anatomy of *Caenorhabditis elegans*).

Bipolar Disorder

M P DelBello, C M Adler, M A Cerullo, D E Fleck, and S M Strakowski, University of Cincinnati College of Medicine, Cincinnati, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Epidemiology and Clinical Characteristics

Bipolar Disorders

Bipolar disorder is a common psychiatric disorder that is typically characterized by recurrent episodes of mania, hypomania (a milder version of a mania that is shorter in duration and does not cause the same degree of impairment), depression, and mixed (co-occurring mania and depression) states. Bipolar I disorder is defined by having at least one manic or mixed episode. A manic episode is a distinct period of abnormally and persistently elevated, expansive or irritable mood that lasts at least 1 week (or any duration if hospitalization is necessary), during which at least three (four if the mood is only irritable) of the following symptoms persist: grandiosity, decreased need for sleep, pressured speech, flight of ideas or racing thoughts, distractibility, increased goal-directed activity or psychomotor agitation, and excessive involvement in pleasurable activity that has a potential for painful consequences. Subtypes of bipolar disorder other than bipolar I disorder are well established and have reasonable empirical data. Specifically, bipolar II disorder is characterized by periods of hypomania (manic symptoms that last for 4 days and do not cause marked impairment in social or occupational functioning) and depression with no history of a full manic episode. Cyclothymic disorder is diagnosed by a period of hypomanic and depressive symptoms (which do not meet the criteria for a major depressive) that continues for at least 2 years (1 year in children and adolescents) with no period of 2 or more months without manic or depressive symptoms. The diagnosis also requires that no major depressive, manic, or mixed episode occurs during the first 2 years of the disorder. Finally, the diagnosis of bipolar disorder not otherwise specified includes disorders with bipolar features that do not meet criteria for any specific bipolar disorder (e.g., individuals who experience manic and depressive symptoms that rapidly alternate over days but do not meet minimal duration criteria for a manic, hypomanic, or depressive episode or individuals with recurrent hypomanic episodes). Children and adolescents with bipolar disorder most commonly present with bipolar disorder not otherwise specified because youth with bipolar exhibit rapidly fluctuating mood symptoms that do not meet minimal duration criteria for a manic or hypomanic episode.

Epidemiology and Prevalence

The prevalence of bipolar I disorder is 1% or 2% in adults. Studies suggest that the prevalence of bipolar spectrum illnesses (bipolar II disorder, cyclothymia, and bipolar disorder not otherwise specified) may be as high as 5–10% in community samples of adults. The onset of bipolar disorders most commonly occurs during late adolescence or early adulthood. Bipolar I disorder and cyclothymia are equally common in men and women. In contrast, bipolar II disorder is more common in women. Bipolar disorder is the sixth leading cause of medical disability worldwide. Patients with bipolar disorder have a higher risk of suicide than patients with any other psychiatric or medical illness. Rates of suicidal ideation and attempts range between 35% and 50%, and the rate of completed suicide is approximately 20%.

Outcome

Although most individuals with bipolar disorder recover from their initial manic episode, the early course of bipolar disorder is characterized by persistent functional impairment. Moreover, following an initial manic episode 80–90% of bipolar patients will have a recurrent mood episode. Disruptions in psychosocial functioning often persist during euthymic periods (i.e., relatively symptom-free periods) and typically include deficits in areas of role function, interpersonal relations, sexual activity, and recreational enjoyment.

Vulnerability to poor outcome in bipolar disorder has been associated with several clinical variables, including the number of prior affective episodes, the presence of psychotic features (especially mood-incongruent psychosis), substance abuse, poor adherence to medications, male gender, and lower social class. Neurocognitive deficits that predate illness onset and are stable across illness course (i.e., trait or core deficits) have not been well established. However, a progression of abnormalities in cognitive domains, such as impairments in attention, memory, and executive function, may be related to the poor functional outcome.

During the past decade, research efforts to identify neurobiological markers of vulnerability to poor functional outcome have increased. Neurobiological markers may be more representative of underlying genetics than observable clinical syndrome. Advances in identifying neurocognitive, neuroanatomical, neurophysiological, and neurochemical measures that act as

markers of poor functional outcome will likely dictate improvements in diagnostic and prognostic accuracy. These types of neuroscientific characteristics will be more resistant to biases inherent to clinician-rated symptomatic, syndromic, and functional outcomes.

Neuropsychological Characteristics

Neurocognitive impairment in patients with bipolar disorder may or may not be evident in a clinical setting. However, during manic episodes, patients generally experience clinically significant cognitive dysfunction in one or more domains. This dysfunction may persist, albeit usually in an attenuated form, during periods of euthymia. However, bipolar disorder is generally not considered to involve global processing deficits, as is common in patients with schizophrenia. Unlike the relatively stable negative syndrome that characterizes the end state in schizophrenia, cognition, mood, and neurovegetative factors continually fluctuate throughout the course of bipolar disorder. In keeping with evidence that cognitive and emotional systems operate in dependent, and sometimes antagonistic, fashion in the healthy human brain, cognitive performance changes across mood states during the normal course of bipolar disorder.

Neurocognitive studies have indicated relative processing deficits in a number of specific domains during mood episodes. The most common deficits reported include decrements in sustained attention, verbal memory impairment (especially on recall and other retrieval-based measures), and deficient executive functioning. In each case, there is evidence for decreased performance during acute mood episodes with at least some normalization during euthymia. The most severe euthymic deficits are typically associated with a chronic course of illness, as indexed by an increased number of mood episodes and hospitalizations or longer mood episodes, which may be suggestive of a neurodegenerative process. There is evidence to suggest that during euthymia, accuracy measures on attention, memory, and executive tests normalize, whereas response latencies remain increased, indicating that reaction time slowing occurs independent of improvements in mood symptoms and performance accuracy (evidence also indicates that psychomotor slowing cannot be fully accounted for by the influence of medications). In general, current data do not support the view that bipolar disorder is characterized by a specific neuropsychological profile but instead suggest that cognitive deficits become more severe and diffuse during acute mood episodes, periods of psychosis, with chronic illness, and in elderly patients.

The view that bipolar disorder is related to right cerebral hemisphere dysfunction to a greater extent than the left has been largely disproved. The prevailing view is that cognitive impairment in bipolar disorder is consistent with fronto-subcortical dysfunction, particularly of the anterior limbic brain network modulating human mood and cognition. Abnormalities in typical brain/behavior relationships have been reliably identified in bipolar disorder. In adults, these abnormalities include, but are not limited to, decreased prefrontal cortical volumes related to disinhibition and poor planning, decreased striatal volumes (e.g., caudate nucleus, putamen, and globus pallidus) underlying psychomotor dysfunction, and increased amygdala volumes related to problems regulating mood and emotion. These cortical areas monitor both internal and external sensory information in order to modulate emotional and social behavior and, ultimately, maintain emotional homeostasis. Indeed, lesion studies indicate that emotional and cognitive problems often result from damage to the frontal lobes as well as to temporal limbic structures.

Beyond structural changes in brain regions involved in cognition, a number of studies have noted functional alterations in dynamic neural networks during active cognitive task engagement in bipolar patients despite a normal structural magnetic resonance imaging (MRI) (based on an individual neuroradiological interpretation). These results suggest that bipolar disorder likely does not cause brain damage *per se*, at least in the majority of cases. Instead, cognitive impairment may result from functional or chemical alterations that demonstrate a degree of plasticity relative to mood symptoms or episodes. For instance, increasing symptoms may act as a trigger for the cognitive system to recruit additional resources from functional cortical regions to compensate for decreased attentional capacity. The degree of symptom interference within the cognitive system may mediate performance across mood states, but more stable neurophysiological changes might be revealed during euthymia by processing inefficiency measures such as psychomotor slowing. Thus, the relationships among global processing deficits (e.g., overall psychomotor slowing, distractibility, and impulsivity) that cut across numerous cognitive domains (e.g., attention, memory, and executive functioning), mood states, and neuroimaging findings warrant further study in individuals with bipolar disorder.

Neuroimaging Findings

Affective experience has been linked to the actions of specific brain regions since at least 1937, when

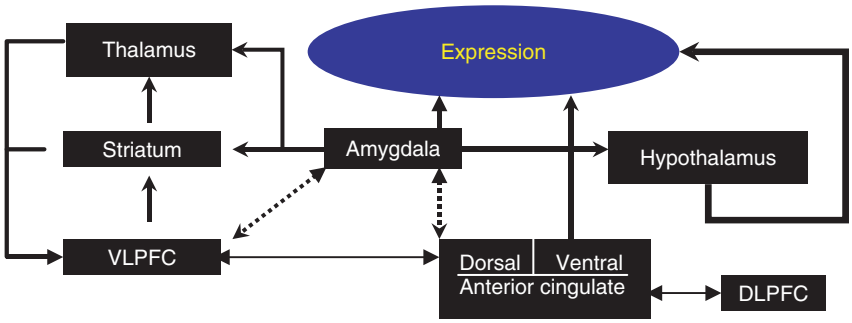

Figure 1 Model of the ALN highlighting hypothesized interactions between structures. Dotted lines indicate inhibitory interactions.

James Papez theorized the involvement of Broca's '*le grande lobe limbique*' in emotional expression. Recent research has led to the extension of the limbic network to include portions of the prefrontal cortex, anterior cingulate cortex (ACC), thalamus, amygdala, hippocampal complex, and fornix and given rise to proposals that mood and behavior are mediated by neuronal interactions within cortico-striatal-pallido-thalamic or cortico-striato-thalamic circuits. Other investigators have postulated that emotional expression is related to interactions within a larger, anterior limbic network (ALN) that encompasses a group of cytoarchitecturally similar brain regions that appear to modulate emotional and cognitive responses through a series of feedback loops between prefrontal, medial temporal, and subcortical structures (Figure 1). The ALN subsumes the extended limbic network and includes input from structures such as the cerebellum not classically associated with affective expression.

Neurostructural Findings

Several lines of evidence suggest that the affective and cognitive symptoms of bipolar disorder are related to disruption of the ALN. Prefrontal cortical volume, for instance, is decreased in bipolar patients compared with healthy controls. Decrements in volume appear to be most prominent in ventral portions of the prefrontal cortex and portions of the ACC identified with emotional regulation and include areas of glial and neuronal cell loss. Studies using voxel-based morphometry (VBM), a structural analysis technique that allows volumetric comparisons within regions and without *a priori* defined structural boundaries, similarly localize decreased prefrontal volume to the dorsolateral prefrontal cortex (DLPFC), ventrolateral prefrontal cortex (VLPFC), orbitofrontal cortex, and ACC. Region-of-interest (ROI) studies in children and adolescents largely fail to replicate these findings, but VBM studies in younger bipolar patients also observe decreased DLPFC and orbitofrontal volume. VBM studies of first-episode bipolar patients, including adolescents, and of patients early in their illness course suggest that structural changes in the VLPFC appear only over time. This interpretation is supported by other studies demonstrating age- and episode-related changes in prefrontal volume.

Portions of the temporal lobe, including anterior temporal cortex and superior temporal gyrus, also appear to be smaller in bipolar patients. Medial temporal findings, however, have been mixed. Amygdala volume has been widely reported to be increased in adult patients with bipolar disorder, whereas volumetric differences have for the most part not been observed in the hippocampus. In contrast, bipolar children and adolescents demonstrate decreased hippocampal and amygdala volume. These discrepant findings may be due to fundamental differences between early and later-onset bipolar disorder, or they may represent evidence of age- or episode-related changes in medial temporal structures.

In general, thalamic volumes have not been found to be affected in either adult or adolescent bipolar patients. However, a minority of studies report thalamic enlargement. Conversely, increased striatal volumes have been widely observed in patients with bipolar disorder, with only a minority of investigators failing to replicate these findings. Similarly, although findings in children and adolescents are also mixed, several studies report observations of increased putamen volumes. Although the cerebellum has been linked only recently with affect regulation, a small number of studies have consistently observed decrements in the cerebellar vermal size of bipolar patients.

Studies using diffusion tensor imaging, a magnetic resonance technique that utilizes measurement of water diffusion to study white matter tracts *in vivo*, suggest that abnormalities are also present in white matter tracts connecting subcortical structures to portions of the prefrontal cortex. Specifically, white matter abnormalities in the anterior fronto-occipital fascicules, which include tracts connecting the orbitofrontal cortex to the temporal and occipital lobes,

and in the anterior genu of the posterior internal capsule (corticothalamic fibers) are present in patients with bipolar spectrum disorders. Similarly, evidence of white matter pathology in prefrontal white matter tracts has been noted in first-episode bipolar patients. Furthermore, increased numbers of white matter hyperintensities (WMHs), areas of increased signal intensity on T2 MRI images that may be indicative of white matter neuropathology, are also increased in patients with bipolar disorder. These areas are concentrated in the frontal cortex and may represent early signs of demyelination. Moreover, WMHs are associated with both poor outcome and cognitive deterioration in bipolar patients.

Neurofunctional Findings

Structural abnormalities in portions of the ALN and in the white matter tracts linking subcortical structures with the prefrontal cortex are accompanied by evidence of functional changes encompassing subcortical, medial temporal, and prefrontal regions.

Functional neuroimaging studies observe changes in neuronal activity in portions of the ALN across a wide range of affective and nonaffective cognitive stimuli. In particular, the VLPFC is overactivated in euthymic bipolar patients in response to most cognitive stimuli. One exception is the Stroop, an inhibition task associated with decreased VLPFC activity across both adult and adolescent studies. Activation of the closely networked ACC is both increased and decreased across functionally disparate subsections.

Prefrontal cortex is reciprocally connected to the amygdala and hippocampus, with projections into the amygdala arising from orbital, medial prefrontal, and lateral portions of the prefrontal cortex. The basal ganglia, and in particular the striatum, also play a significant role in both cortical and emotional regulation. Multiple striatal pathways project to orbitofrontal, medial, and lateral prefrontal regions, and these prefrontal areas project to the ventromedial striatum and thalamus to create feedback loops through which the prefrontal cortex influences psychomotor response to emotion. The nucleus accumbens, which forms part of the ventral striatum, is a central structure in these interconnected limbic and striatal networks. The nucleus accumbens projects to the ventral pallidus, hypothalamus, and brain stem, and it receives abundant afferents from the prelimbic cortex, hippocampal formation, thalamus, and amygdala. The nucleus accumbens is particularly closely associated with the amygdala; the extended amygdala and portions of the nucleus accumbens may be part of a common functional anatomic continuum. Through reciprocal connections, increased ALN activity is also associated with inhibition of neuronal activity in the DLPFC in response to affective and nonaffective stimuli, interfering with performance on cognitive tasks mediated through this region. Euthymic bipolar patients viewing affective faces also showed increased activity in striatum, thalamus, and potions of the medial temporal cortex. Increased activation was similarly reported in these regions in response to nonaffective cognitive stimuli, including attention and working memory tasks. Findings in bipolar adolescents are limited but are consistent with those in adult patients. Limitations of most studies examining neurofunctional abnormalities in patients with bipolar disorder include variability in patient mood state and medication exposure.

Neurochemical Findings

Magnetic resonance spectroscopy (MRS) studies are consistent with findings of neurostructural and neurofunctional abnormalities in bipolar patients. Studies using phosphorous MRS (31P-MRS) show evidence of altered phospholipid metabolism in bipolar disorder that may be indicative of metabolic dysfunction. Decreased phosphocreatine (PCr) and phosphomonoester (PME) concentrations in euthymic bipolar patients are consistent with impaired metabolic activity. Increased concentrations of lactate, also observed in portions of the ALN of bipolar patients, suggest that these findings are associated with a shift in neuronal metabolism from oxidative phosphorylation to glycolysis, which may reflect neuronal hyperactivity as well as mitochondrial dysfunction. Intracellular abnormalities in pH are also consistent with the presence of impaired mitochondrial function, leading several investigators to hypothesize that mitochondrial dysfunction is directly involved in the pathophysiology of bipolar disorder. This hypothesis is supported by genetic studies linking bipolar disorder with specific polymorphisms and transcriptional abnormalities of mitochondrial DNA. Impaired metabolic function may in particular interfere with membrane turnover and impact secondary messenger systems related to the phosphatidylinositol cycle. In addition, mitochondrial dysfunction in patients with bipolar disorder may increase their vulnerability to excitotoxic insults. Indeed, proton MRS (1H-MRS) studies indicate that *N*-acetyl-aspartate (NAA), which is an amino acid found in high concentrations within neurons that is widely utilized as a nonspecific marker of neuronal viability and integrity, is decreased in prefrontal cortical regions of patients with bipolar disorder. This finding is consistent with more gross morphological observations of decreased prefrontal volume.

1H-MRS studies also report increased Glx (glutamate, glutamine, and γ-aminobutyric acid (GABA))

levels in prefrontal regions (e.g., the ACC and VLPFC) of bipolar adults and youth, particularly during mood episodes. The elevated Glx resonance also supports the notion of abnormal metabolism in bipolar patients since most (80–90%) of the total cortical glucose oxidation is associated with the energy required for glutamatergic neurotransmission. Although the 1H-MRS Glx resonance includes glutamate, glutamine, and small amounts of GABA, in prefrontal cortical areas Glx is predominantly glutamate. Increased concentrations of glutamate, an excitotoxic neurotransmitter, are associated with cellular death and may in turn be related to findings of decreased NAA during episodes of mania.

Manic episodes are also associated with other neurochemical evidence of abnormal neuronal metabolism, including decreased PCr and PME resonances. PCr is a high-energy phosphate formed from adenosine triphosphate (ATP) and creatine that transfers high-energy phosphate from the mitochondria when ATP is consumed. PCr transfers its phosphate to adenosine diphosphate (ADP) to replenish ATP stores, acting to buffer concentrations of the latter, energy storage molecule. 31P-MRS studies demonstrate decreased levels of PCr in bipolar patients during manic episodes, suggesting that bioenergetic impairments worsen with mania. Additionally, 31P-MRS and 1H-MRS studies of PME, choline, and myo-inositol (mI) concentrations in bipolar patients suggest the presence of alterations in membrane turnover and second messenger signal transduction systems which are also indicative of changes in cellular metabolism.

Together, functional and structural imaging findings suggest that the pathophysiology of bipolar disorder is characterized by abnormalities of the ALN. Studies of euthymic bipolar patients demonstrate a pattern of neurofunctional responses indicative of oversensitivity both in portions of the ALN involved in perception and expression of mood and in those involved in regulating and modulating this expression. During mania, there appears to be a loss of compensatory activity in modulatory portions of the ALN that is associated with the behavioral manifestations associated with bipolar disorder.

Heritability and Genetics of Bipolar Disorder

Family, Twin, and Adoption Studies

The heritability of bipolar disorder is determined through family, twin, and adoption studies. In family studies, the relatives of patients with bipolar I disorder are examined to determine if they have bipolar disorder or other mental illness. In the 17 major family studies of bipolar disorder, subjects with one immediate family member (i.e., parent, sibling, or child) with bipolar I disorder have a 2–19% chance of being diagnosed with bipolar I disorder. Thus, bipolar I disorder clearly aggregates in families. However, family studies cannot determine if this aggregation is due to shared genetics or environmental factors. For this, twin and adoption studies are needed. Twin studies compare the prevalence of bipolar disorder in monozygotic (identical) and dizygotic (fraternal) twins who share the same environment, and adoption studies examine the prevalence of bipolar disorder in parents of adopted children with bipolar disorder. In 6 studies of twins raised together, if one monozygotic twin (i.e., identical twin) has bipolar I disorder, there is a 20–75% chance the other twin will have the disorder compared to 0–7% chance of sharing the disorder in dizygotic twins. The two adoption studies of bipolar disorder also support the heritability of this disorder. Adoptive parents of bipolar children have a 2% or 3% rate of bipolar disorder compared to a 5–7% rate in the biological parents.

These three types of studies demonstrated a high level of heritability in bipolar I disorder, which may the highest of any major psychiatric disorder. Finally, family studies have also shown that bipolar I disorder also aggregates with major depressive disorder and bipolar II disorder. Eleven to 19% of subjects with an immediate family member with bipolar I disorder have major depressive disorder. In contrast, there is a low risk (1% or 2%) of bipolar I disorder in subjects with immediate family members with major depressive disorder. Bipolar I disorder is found in equal rates in family members of patients with bipolar I and II disorder. However, bipolar II disorder is more prevalent in relatives of patients with bipolar II disorder (6.1%) compared to relatives of patients with bipolar I disorder (1.8%).

Genetic Studies

A significant limitation of family, adoption, and twin studies is that they cannot identify the exact nature of the heritability found. Determining the specific genetic components involved requires linkage and association studies. Linkage studies search chromosomes for genetic markers of diseases in families with a high prevalence of bipolar disorder by looking at the association between marker genes and disease expression. These studies allow the identification of specific regions of the chromosome that are associated with disease expression. In bipolar I disorder, several chromosomal regions have been found that are associated with disease susceptibility. The following chromosomal regions have the most evidence

for their involvement in bipolar disorder: 2p, 4p, 4q, 6q, 8q, 11p, 12q, 13q, 16p, 16q, 18p, 18q, 21q, 22q, and Xq.

Association studies examine whether particular gene variants are associated with disease susceptibility. Instead of looking at families with a high prevalence of a disease, association studies examine individuals with the disease and comparison subjects without the disease and search for the presence or absence of candidate alleles. Candidate genes are generally selected based on animal models or knowledge of the disease pathology. In bipolar I disorder, association studies have found evidence of the involvement of multiple alleles, including *COMT*, *DAT*, *HTR4*, *DRD4*, *DRD2*, *HTR2A*, *5-HTT*, *G72/G30*, *DISC1*, *NCAM1*, *P2RX7*, *MAOA*, *GABRA3*, *GRIN2A*, *CHMP1.5*, *ADCY9*, and *BDNF*. Many of these genes are involved with the neurotransmitters dopamine (DRD5, DAT, DRD4, DRD2, and MAOA), serotonin (HTR4, HTR2A, 5-HTT, and MAOA), GABA (*GABRA3*), and glutamate (*GRIN2A*). Other genes are involved in neural growth (*BDNF* and *NCAM1*), structural proteins (*DISC1*), and neuronal signaling pathways (*P2RX7*, *CHMP1.5*, and *ADCY9*). There are also some genes whose functions are not clear (*G72/G30*). Among these multiple genes, the *G72/G30* gene complex, located on chromosome 13q, is the only finding that has been fully replicated. However, the specific causal mechanisms by which this gene complex or any of the other suspected genes leads to the pathology of bipolar disorder is unknown. One of the difficulties with genetic studies in bipolar disorder is that disease expression likely requires the interactions of multiple candidate genes, not any single gene. Therefore, future research is needed to examine the specific effects of the interactions among multiple genes. In addition, endophenotypes may prove useful for future heritability studies. Endophenotypes may include alterations in neurocognitive tests, brain structure, neurophysiology, or neurochemical marker. Since many of these traits can be quantified, they will allow greater statistical power in linkage studies.

Treatment

Several pharmacological agents are available to treat the acute mood (i.e., manic and depressive) episodes of bipolar disorder and to prevent further episodes (i.e., maintenance treatment). Mood-stabilizing medications commonly used to treat bipolar disorder include lithium; antiepileptic agents such as valproic acid, carbamazepine, and lamotrigine; and atypical (second-generation) antipsychotics such as olanzapine, 2-clozapine quetiapine, risperidone, aripiprazole, and ziprasidone. Although these medications are effective, large clinical trials suggest that only approximately half of any bipolar patient group will exhibit an acute antimanic treatment response to any given medication. Similarly, the acute antidepressant effect of these agents appears to be less than the antimanic response. Combining medications improves response rates by approximately 15–20%; however, this strategy results in more adverse events. Furthermore, maintenance studies report high rates of affective episode recurrences even with good medication adherence.

Although a number of clinical predictors of drug response (e.g., specific symptoms) have been hypothesized, these predictors are unreliable in individual patients. Therefore, the early treatment course of bipolar disorder often consists of repeated ineffective medication trials. Moreover, the mechanisms by which pharmacological agents exert their effects in bipolar patients remain largely unknown. Neurobiological predictors of treatment response may be more useful than clinical predictors. Neuroimaging studies may identify neurobiological predictors of treatment effects and neurobiological changes associated with treatment response. For example, lithium, which inhibits inositol monophosphatase, has been shown *in vivo* to reduce mI levels in the frontal lobe of patients with bipolar disorder. However, the mI reduction occurs at a time when the clinical state is unchanged, suggesting that the effects of lithium are more downstream in cellular processes. Other studies have demonstrated that elevated mI levels, prior to lithium administration, are associated with lithium response. Functional magnetic resonance imaging (fMRI) and MRS studies examining the neurofunctional and neurochemical effects of other medications for bipolar disorder are needed to improve the clinical management and decrease the morbidity and mortality of bipolar disorder.

Conclusion

Bipolar disorder is a common psychiatric disorder characterized by affective instability and cognitive deficits, particularly during mood episodes. Abnormalities within the ALN and related brain regions appear to be involved in the neurophysiology of bipolar disorder. Moreover, studies indicate a high level of heritability in bipolar disorder, although the involvement of specific genes has not been definitively established. Several effective pharmacological treatment options exist for patients with bipolar disorder. Future studies combining neurocognitive, neuroimaging, and genetic techniques may be useful to identify endophenotypes of bipolar disorder and ultimately to develop rationale treatment strategies and improve the outcomes of individuals with bipolar disorder.

See also: Antipsychotic Drugs; Cognitive Dysfunction in Psychiatric Disorders; Mood Stabilizers; Neuroimaging; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Psychiatric Disorders: Functional Genomics.

Further Reading

Adler CM, DelBello MP, and Strakowski SM (2006) Brain network dysfunction in bipolar disorder. *CNS Spectrums* 11: 312–320.

Bearden CE, Hoffman KM, and Cannon TD (2001) The neuropsychology and neuroanatomy of bipolar affective disorder: A critical review. *Bipolar Disorder* 3: 106–150.

Cerullo MA and Strakowski SM (2007) Biologic factors in different bipolar disorder sub-types. In: Soares JC and Young AH (eds.) *Bipolar Disorders: Basic Mechanisms and Therapeutic Implications,* 2nd edn. New York: Informa Healthcare USA.

Hayden EP and Nurnberger JI (2005) Molecular genetics of bipolar disorder. *Genes, Brain and Behavior* 5: 85–95.

Keck PE Jr. (2006) Long-term management strategies to achieve optimal function in patients with bipolar disorder. *Journal of Clinical Psychiatry* 67(supplement 9): 19–24.

Maier W (1993) The distinction of bipolar II disorder from bipolar I and recurrent unipolar depression: Results of a controlled family study. *Acta Psychiatrica Scandinavica* 87: 279–284.

Martinez-Aran A, Vieta E, Reinares M, et al. (2004) Cognitive function across manic or hypomanic, depressed, and euthymic states in bipolar disorder. *American Journal of Psychiatry* 161: 262–270.

Quraishi S and Frangou S (2002) Neuropsychology of bipolar disorder: A review. *Journal of Affective Disorders* 72: 209–226.

Sachs GS (2006) Implementing evidence-based treatment of manic and mixed episodes. *Journal of Clinical Psychiatry* 67(supplement 11): 12–17.

Savitz J, Solms M, and Ramesar R (2005) Neuropsychological dysfunction in bipolar affective disorder: A critical opinion. *Bipolar Disorder* 7: 216–235.

Shih RA, Belmonte PL, and Zandi PP (2004) A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. *International Review of Psychiatry* 16: 260–283.

Relevant Websites

http://www.neurotransmitter.net – Neurotransmitter.net See 'Bipolar disorder genetics.'

http://www.genetics.med.ed.ac.uk – Medical Genetics Sections. See 'Schizophrenia and bipolar disorder genetics blog.'

Bird Brain: Evolution

E D Jarvis, Duke University Medical Center, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

The classical century-old view of bird brain evolution was based on a theory of linear and progressive evolution, where the vertebrate cerebrum was argued to have evolved in lower to higher anatomical stages, with birds at an intermediate stage. As such, the avian cerebrum was thought to consist mostly of basal ganglia territories, and these were thought to control mostly primitive behaviors. This view, although slowly challenged, was not formally changed until 2004 and 2005, when a forum of neuroscientists that reevaluated brain terminologies and homologies for birds and other vertebrates published a new nomenclature that reflects a modern understanding of bird brain evolution. This view proposes that the avian cerebrum was inherited as a package with pallial (or cortical-like), striatal, and pallidal regions from its recent reptilian and distant stem amniote ancestors. The avian pallial domain is organized differently from the mammalian pallium or cortex in that the avian is nuclear and the mammalian is layered. However, the pallia of both groups perform similar functions and have some similar connectivity. The avian striatal and pallidal domains are organized more similar to those of mammals and, likewise, perform similar functions in both groups. This article reviews these classical and modern views for bird brain evolution in the context of vertebrate brain evolution.

The Classical View

The classical view of telencephalic evolution had its origin in the late nineteenth and early twentieth century following the impact of Darwin's theory of evolution. Inspired by Darwin's theory, Ludwig Edinger formulated an evolution-based model of brain organization from 1885 to 1908. Edinger and others combined Darwin's concept of 'evolution' with the nineteenth-century version of Aristotle's 'scala naturae,' resulting in the view that evolution was unilinear and progressive, from fish to amphibians, reptiles, birds and mammals, primates, and, finally, humans, ascending from 'lower' to 'higher' intelligence in a chronological series. They believed that the brains of extant vertebrates retained the ancestral structures and, therefore, that the origin of specific human brain subdivisions could be traced back in time by examining the brains of extant nonhuman vertebrates. Although the spinal cord, hindbrain, midbrain, thalamus, and cerebellum appeared to be well conserved across vertebrates, Edinger noted that the internal organization of the telencephala or cerebrum exhibited the most pronounced differences across vertebrates. In mammals, the outer part of the telencephalon had prominently layered gray matter (Figure 1(b), green), whereas the inner part had nuclear gray matter (Figure 1(b), purple). The inner part was located ventral to the lateral ventricle. In nonmammals, the outer and inner parts of the telencephala had mostly nuclear gray matter, and most of it was located ventral to the lateral ventricle in birds and reptiles (Figure 1(b), purple).

On the basis of these considerations, Edinger proposed that telencephalic evolution occurred in progressive stages of increasing complexity and size, culminating with the human cerebrum. He argued that the stages proceeded in a ventral-to-dorsal direction, with each new vertebrate group acquiring a more advanced cerebral subdivision. He argued that first there was the old brain, the paleoencephalon (also called the basal ganglia or subpallium), which controlled instinctive behaviors, followed by the addition of a new brain, the neoencephalon (also called the pallium or mantle at the top), which controlled learned and intelligent behaviors. The telencephalic subdivisions within each vertebrate group were further named by him, Ariëns-Kappers, and others with the prefixes 'paleo' (oldest), 'archi' (archaic), and 'neo' (new) to designate the presumed relative order of evolutionary appearance of each subdivision. To these prefixes the root word 'striatum' was added for the presumed paleoencephalic subdivisions and 'pallium' (meaning mantle) or 'cortex' (meaning bark) for the presumed neoencephalic subdivisions. The term striatum was used because a large part of the basal ganglia (i.e., paleoencephalon) in mammals, which is now commonly call caudate putamen, contains fiber bundles coursing through it, giving it a striated appearance.

The classical view that became dominant in the context of bird brain was that the bird subpallium consisted of 'paleostriatum' (old striatum) inherited originally from fish, said to be the antecedent of the human globus pallidus, and devoted to olfactory information processing. Above the paleostriatum, birds were said to have inherited an 'archistriatum' (archaic striatum) originally from amphibians, said to be the antecedent of the human amygdala. Above and adjacent to the archistriatum, birds were said to have inherited a 'neostriatum' (new striatum) from their closest relative, reptiles, said to be the antecedent of the human caudate and putamen. Reptiles were

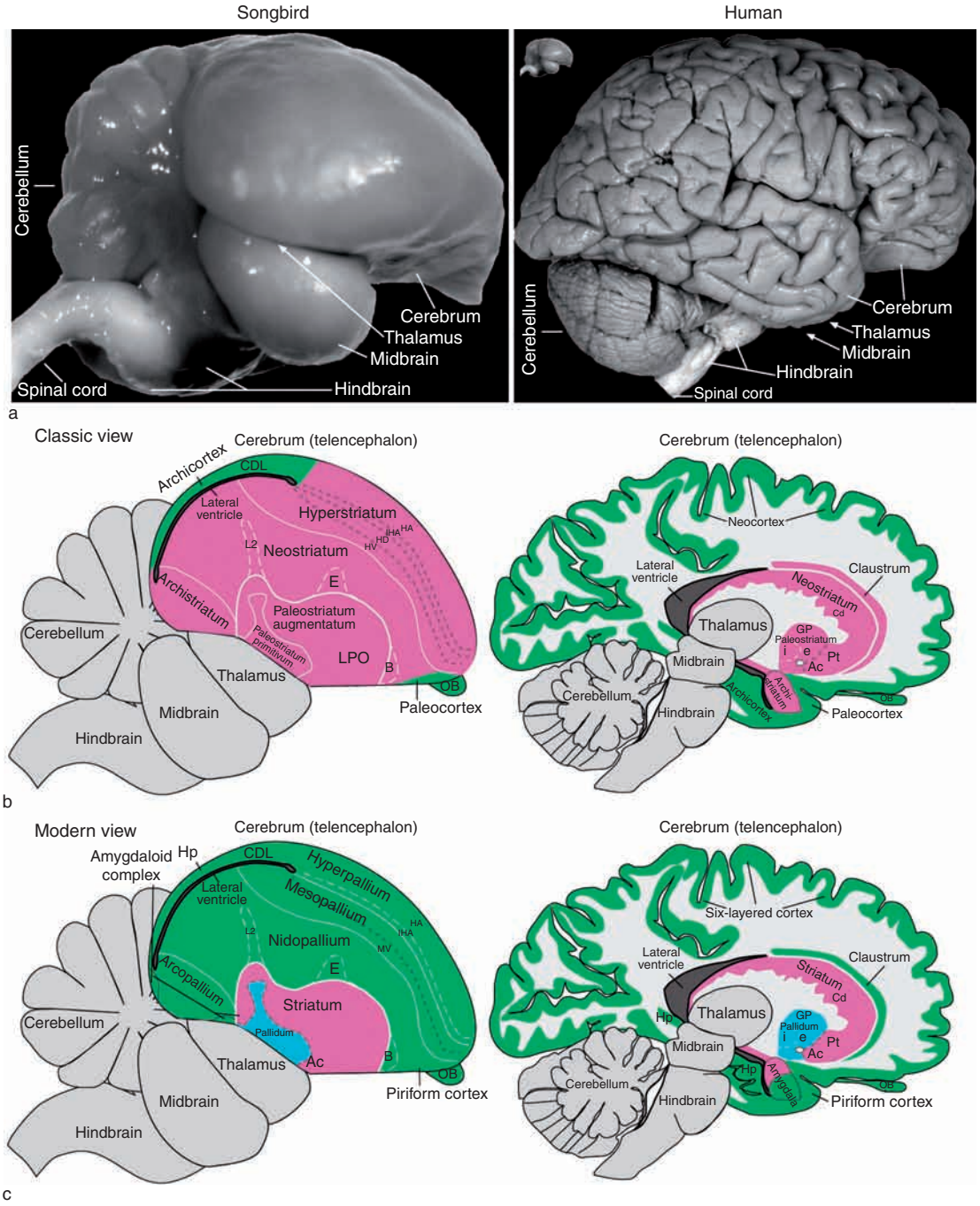

Figure 1 Avian and mammalian brain relationships. (a) Side view of a songbird (zebra finch) and human brain to represent avian and mammalian species. In this view, the songbird cerebrum covers the thalamus; the human cerebrum covers the thalamus and midbrain. Left inset next to the human brain is the zebra finch brain to scale with that of the human. Human brain image is from John W. Sundsten of the Digital Anatomist Project (http://www9.biostr.washington.edu/da.html) and used with permission. (b) Classic view of avian and mammalian brain relationships. Although past authors had differing opinions regarding which brain regions are pallium versus subpallium, the images are color-coded to the meaning of the names given. (c) Modern view of avian and mammalian brain relationships according to the conclusions of the Avian Brain Nomenclature Consortium. White solid lines are lamina, cell-sparse zones separating brain subdivisions. Large white areas in the human cerebrum are axon pathways called white matter. Dashed lines divide regions that differ by cell density or cell size; dashed white lines separate primary sensory neuron populations from adjacent regions. Classic view: Ac, accumbens; B, nucleus basalis; Cd, caudate nucleus; CDL, dorsal lateral corticoid area; E, ectostriatum; GP, globus pallidus (i, internal segment; e, external segment); HA, hyperstriatum accessorium; HD, hyperstriatum dorsale; HV, hyperstriatum ventrale; IHA, interstitial hyperstriatum accessorium; L2, field L2; LPO, lobus parolfactorius; OB, olfactory bulb; Pt, putamen. Modern view where different: B, basorostralis; E, entopallium; HA, hyperpallium apicale; Hp, hippocampus; IHA, interstitial hyperpallium apicale; MV, mesopallium ventrale. (b, c) Adapted from Jarvis ED, Gunturkun O, Bruce L, et al. (2005) Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience* 6: 151–159, with permission.

additionally thought to have elaborated the paleostriatum into an older part (the primitivum) and a newer part (the augmentatum), which was passed onto birds. Birds were then thought to have evolved a large additional basal ganglia subdivision, the 'hyperstriatum' (hypertrophied striatum), considered unique to birds.

The bird pallium was said to consist of a 'paleocortex' inherited from fish pallium and believed to be the antecedent of the human olfactory cortex. Birds were said to have inherited an 'archicortex,' also thought to be olfactory and primitive, from amphibians and believed to be the antecedent of the human hippocampus. Birds were thought not to have evolved any further pallial regions. In contrast, mammals were thought to have evolved from the paleocortex and/or archicortex, a 'neocortex.' The archicortex and/or paleocortex with their two or three cell layers were assumed to be primitive; the neocortex with its six layers was assumed to be more recently evolved and a substrate for more sophisticated behavior. This classical view was codified in the major comparative neuroanatomy text by Ariëns-Kappers, Huber, and Crosby in 1936 and became pervasive throughout neuroscience. This view is now known to be incorrect.

The Modern View

In 2004 and 2005, an international consortium of specialists in avian, mammalian, reptilian, and fish neurobiology – the Avian Brain Nomenclature Consortium – published several reports on a new nomenclature that represents the current understanding of avian telencephalic organization and homologies with mammals (Figure 1(c)). The consortium concluded that the avian telencephalon is organized into three main, developmentally distinct domains that are homologous in fish, amphibians, reptiles, birds, and mammals: pallial, striatal, and pallidal domains (Figure 1(c)). For birds, they renamed these domains and their respective subdivisions with homology-based terms or roots that allow reference to named regions in mammals. In doing so, they eliminated all phylogeny-based prefixes (paleo-, archi-, and neo-) that erroneously implied the relative age of each subdivision.

The avian paleostriatum was simply renamed striatum after its mammalian homolog (Figure 1(c)). The paleostriatum primitivum and adjacent ventral region was renamed pallidum also after its mammalian homolog. Similar to the mammalian pallidum, the avian pallidum has a sparse distribution of cells, giving the region its pale appearance and its name. The dorsal part of the avian pallidum was determined to be homologous to the mammalian globus pallidus and renamed as such, whereas the ventral part was determined to be homologous to the mammalian ventral pallidum. The dorsal pallidum, however, differs between mammals and birds. In mammals, it consists of two segments with distinct connectivity, the internal and external globus pallidus, whereas in birds neurons with both phenotypes are intermingled (Figure 1(c)).

The avian pallium was determined to be organized not into three striatal subdivisions (hyperstriatum, neostriatum, and archistriatum) but, rather, into four pallial subdivisions renamed as hyperpallium (hypertrophied pallium; upper part of old hyperstriatum), mesopallium (middle pallium; lower part of old hyperstriatum), nidopallium (nest pallium; old neostriatum), and arcopallium (arched pallium; most of old archistriatum) (Figure 1(c)). Several neuronal populations adjacent to the arcopallium and the posterior part of what had been regarded as archistriatum were considered homologous to pallial and subpallial parts of the mammalian amygdala, and they were renamed as the avian amygdaloid complex. Other regions that were widely recognized to be homologous among vertebrates – the hippocampus, olfactory (piriform) cortex, and olfactory bulb – did not require name changes.

Some Anatomical and Molecular Parallels

Some key similarities between avian and mammalian subpallium include a high enrichment of dopaminergic axon terminals in the striatum that originate from the substantia nigra pars compacta and ventral tegmental area neurons of the midbrain. Both avian and mammalian striatum contain two major classes of spiny neuron types, those with the neuropeptide substance P (SP) and those with the neuropeptide enkephalin (ENK), which project to two different neuron populations in the pallidum. In both birds and mammals, the SP neurons seem to be involved in promoting planned movement, whereas the ENK neurons seem to be involved in inhibiting unwanted movements. Both the avian and the mammalian striatum participate not only in instinctive behavior and movement but also in motor learning. Developmental studies indicate that the avian and mammalian subpallium consists of two separate histogenetic zones that express different sets of transcription factors: (1) a dorsal zone that corresponds to the lateral ganglionic eminence and that selectively expresses the transcription factors Dlx1 and Dlx2 but not Nkx2.1 and (2) a ventral zone that corresponds to the medial ganglionic eminence and selectively expresses all three transcription factors. The lateral ganglionic eminence gives rise to the striatum and the medial ganglionic eminence

gives rise to the pallidum. Similar striatal and pallidal territories have been found in reptiles.

Some similarities between avian and mammalian pallium (i.e., mammalian cortex) include direct projections of sensory visual, auditory, and somatosensory input from the thalamus. The corresponding avian brain regions subserve the same type of sensory information processing as is performed by the mammalian cortex. Likewise, the avian hyperpallium and arcopallium give rise to major descending projections to premotor and motor neurons of the brain stem and spinal cord, as do mammalian corticobulbar and corticospinal pathways. These avian brain regions subserve critical roles in motor control and sensorimotor learning, similar to the mammalian cortex. The pallial relationships between these avian and mammalian brain regions are also supported by molecular embryology studies. During development, the avian hyperpallium, mesopallium, and nidopallium and the mammalian pallium selectively express the pallium-specific transcription factors Emx1, Pax6, and Tbr1. In adult birds and mammals, comparative expression profiles of the neurotrophic factor brain-derived neurotrophic factor and the glutamate receptor mGluR2 indicate that the avian arcopallium express these pallial-specific mRNAs at high levels, as does the mammalian cortex.

A Modern View of Telencephalic Evolution

The organization of the true basal ganglia of birds relative to mammals and other vertebrates (i.e., distinct nuclear striatal and pallidal domains) is quite conserved. In contrast, the organization of their pallial domains differs to a greater degree. The nuclear type of organization of the pallium found in birds is also present in reptiles. The avian hyperpallium possesses a unique organization so far found only in birds; it consists of semilayered subdivisions, and it might have evolved more recently than the mammalian six-layered cortex since birds evolved well after mammals (by ~50–100 million years) (Figure 2). The region below the ventricle, called the dorsal ventricular ridge (DVR), contains mesopallium, nidopallium, and arcopallium, and it is a nuclear pallial organization that is unique to birds and reptiles. The six-layered cortex is a pallial organization unique to mammals. Because all major groups of living mammals (monotremes, marsupials, and placentals) have a six-layered cortex, it was presumably inherited from their common therapsid ancestor more than 200 million years ago (Figure 2). Furthermore, mammals may not have arisen from reptiles but, rather, from therapsids, because reptile and mammal lineages apparently had their last common ancestor in stem amniotes. Because all nonmammalian therapsids are extinct (Figure 2), it is difficult to trace from stem amniotes to mammals the evolutionary history of mammalian telencephalic organization – layered, nuclear, or otherwise. Thus, the reptilian nuclear pallial organization cannot be assumed to represent the ancestral condition for mammals, as it is for birds.

In general, the evidence suggests that pallial, striatal, and pallidal domains exist in most or all vertebrates. Thus, it has been hypothesized that the telencephalon of early fishes possessed all three domains, which were then inherited as a package by later vertebrates, including birds, and independently

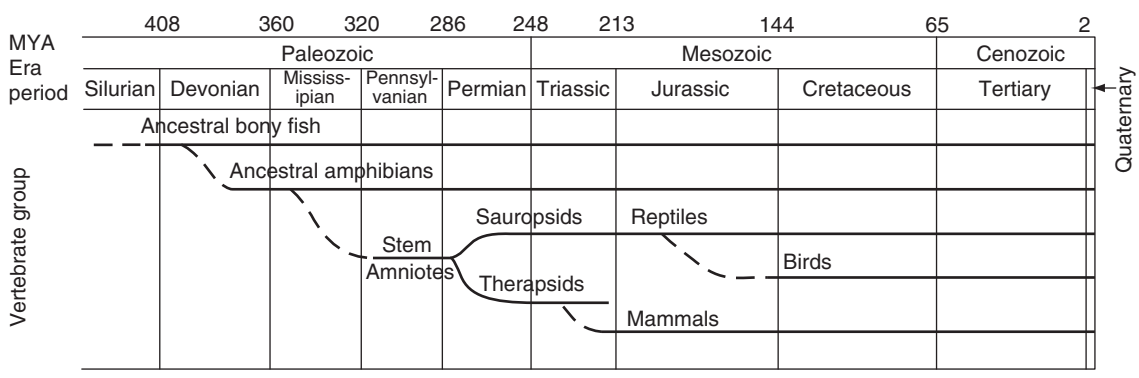

Figure 2 Simplified modern view of vertebrate evolution. The diagram begins with the fish group that contains the most recent ancestor to land vertebrates. This view differs from the classic view in that instead of giving rise to reptiles, ancestral amphibians are thought to have given rise to stem amniotes. Stem amniotes then split into at least two major groups: the sauropsids, which gave rise to all modern reptiles as we know them today, and the therapsids, which, through a series of now-extinct intermediate forms, evolved into mammals. Many sauropsids (reptiles) are currently living. Solid horizontal lines indicate temporal fossil evidence. Dashed lines indicate proposed ancestral links from other types of data. MYA, millions of years ago. Based on Carroll RL (1987) Fossils and relationships. In: *Vertebrate Paleontology and Evolution*, pp. 1–13. New York: Freeman; and Evans SE (2000) General discussion II: Amniote evolution. In: Bock GR and Cardew G (eds.) *Evolutionary Developmental Biology of the Cerebral Cortex*, no. 228, pp. 109–113. Chichester, UK: Wiley.

modified in them. The conserved organization of striatal and pallidal domains indicates that there might be constraints on how the basal ganglia can be organized. The diverse organizations of the pallial domains suggest that there are fewer constraints on how the pallium can be organized.

Detailed Views That Need Further Testing

When the modern view of bird brain evolution was developing, it was accompanied by several new hypotheses about one-to-one homologies between specific avian and mammalian pallial subdivisions. These hypotheses are under active investigation and make up a large area of research on bird brain evolution. They are classified into two main categories: nuclear-to-layered hypotheses and nuclear-to-claustrum/amygdala nuclei hypotheses.

Nuclear-to-Layered Hypotheses

First proposed by Karten, these hypotheses (Figure 3(a)) state that the similarities in connectivity between the avian hyperpallium, nidopallium and arcopallium of birds, and the six-layered cortex of mammals stem from a common origin of these structures – that is, they are homologous. Karten proposed that the common ancestor of birds, reptiles, and mammals possessed a nuclear pallium that was transformed into a laminar pallium early in the mammalian lineage, maintaining connectivity of the ancestral nuclear network. In this regard, he argued that the avian pallium is divided into three sets of serially connected types of neurons – thalamorecipient neurons (field L2, entopallium, and basorostralis), pallio-pallial neurons (nidopallium), and extratelencephalic projection neurons (arcopallium) – with cell types and interconnectivity similar to those of mammalian cortical layers IV, II and III, and V and VI, respectively (Figure 3(a)). Similar arguments were made by others for the avian upper hyperpallium (also known as the Wulst), which also has serially connected neuron types that resemble those found in the mammalian cortex. In this view, avian L2 neurons are homologous to layer IV neurons of mammalian primary auditory cortex, basorostralis neurons to layer IV of primary somatosensory cortex, entopallium neurons to layer IV of extrastriate visual cortex, and the interstitial hyperpallium apicale neurons to layer IV of striate visual cortex. Supporting this hypothesis, gene expression studies have shown that avian thalamorecipient nuclear fields (L2, entopallium, basorostralis, and interstitial hyperpallium) and the mammalian thalamorecipient layer IV of cortex selectively express some of the same genes. Avian extratelencephalic projection neurons (in the arcopallium but not in the hyperpallium) and mammalian extratelencephalic projection neurons (layer V neurons of cortex) both selectively express some of the same genes. Thus, although the avian pallium is not organized cytoarchitectonically into layers, its nuclear subdivisions bear marked similarities in connectivity and molecular profiles to different layers of the mammalian cortex.

Nuclear-to-Claustrum/Amygdala Hypotheses

These hypotheses differ from the nuclear-to-layered hypotheses with regard to mammalian homologies with the avian DVR (mesopallium, nidopallium, and arcopallium; Figure 3(b)). In an early formulation of the nuclear-to-claustrum/amygdala hypotheses, Bruce and Neary proposed that the avian DVR represents an elaboration of parts of the mammalian amygdala. Subsequently, Striedter proposed that the avian DVR represents an elaboration of the mammalian amygdala and claustrum, and that the connectivity that the DVR shares with the neocortex evolved independently. Support for this view is based on findings that both the avian DVR and mammalian claustrum/amygdala are nuclear in organization, that both the avian DVR and part of the mammalian amygdala have similar connections, and that both have conserved developmental expression patterns of regulatory genes that play key roles in brain regionalization and morphogenesis. Puelles and colleagues proposed that the common topographic expression patterns of the transcription factors Emx1 and Pax6 in the avian mesopallium and in the mammalian dorsal claustrum and basolateral amygdala indicate that these structures commonly arose from the lateral pallium (Figure 3(b)). They argued that the absence of Emx1 but the presence of other pallial genes in the avian nidopallium and in the mammalian ventral claustrum and lateral anterior amygdala indicate that these structures commonly arose from the ventral pallium. They further argued that the avian arcopallium and mammalian amygdala consist of subpallial parts derived from striatal and pallidal cell groups and, by this association, that the avian arcopallium is homologous to the mammalian amygdala, as originally proposed by Edinger.

Both categories of hypotheses have their limitations. For the nuclear-to-layer hypotheses, developmental studies have not been conducted to test whether the three types of serially connected neurons in birds arise from cell types similar to those that give rise to the cortical layers in mammals. Furthermore, not all gene expression patterns support one-to-one molecular relationships between avian pallial subdivisions and mammalian cortical layers. Neither do all findings support the nuclear-to-claustrum/amygdala nuclei hypotheses. More investigation is required. This is an active area of research and debate, but the

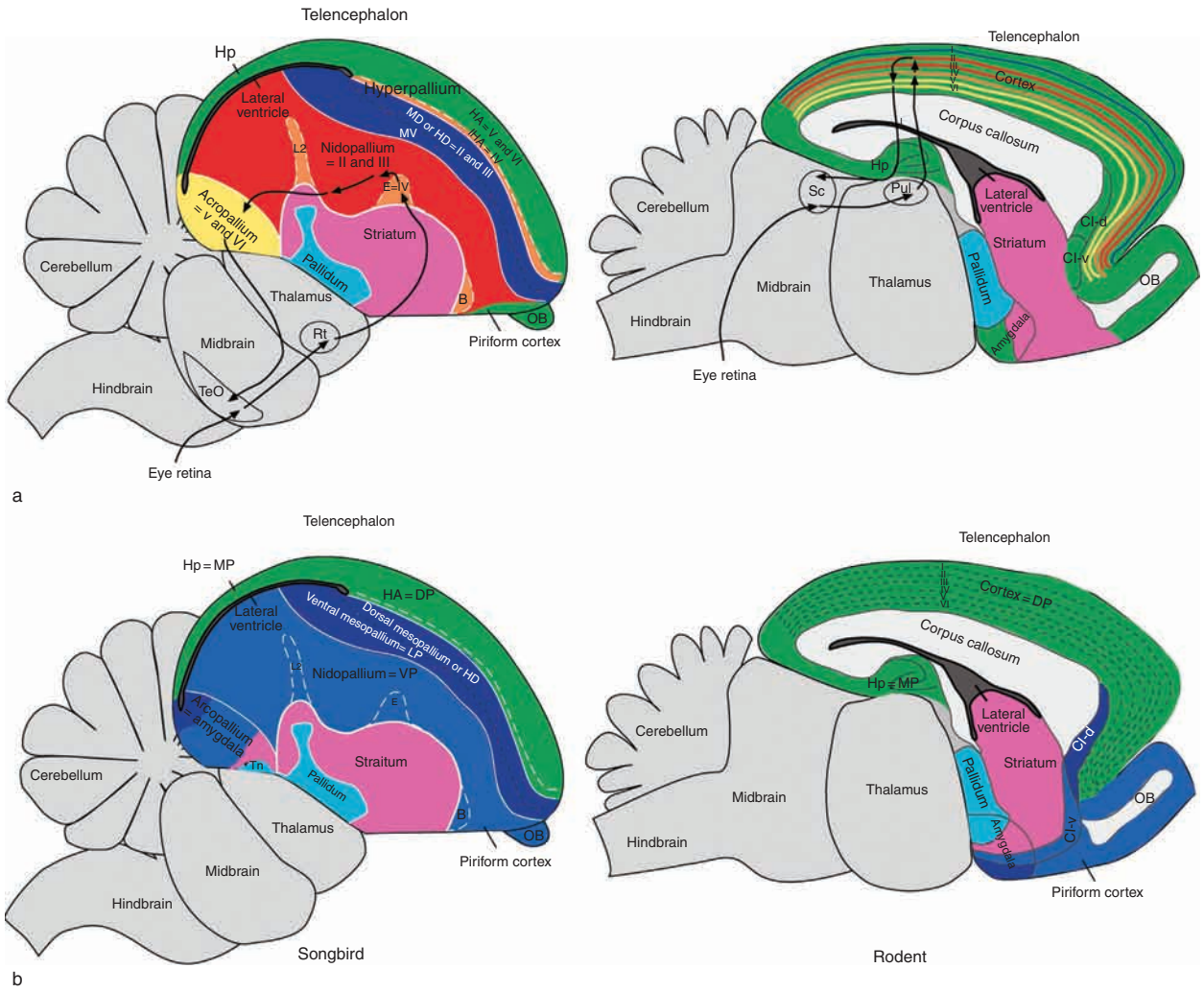

Figure 3 Working hypotheses on avian and mammalian pallial homologies in the context of the new avian brain nomenclature. (a) An example of a nuclear-to-layered hypothesis. Connectivity of tectofugal visual pathways in avian (songbird, left) and mammalian (rat, right) brains is shown. Color coding indicates proposed homologies between birds and mammals. (b) An example of a nuclear-to-claustrum/amygdala hypothesis. B, basorostralis; Cl-d, claustrum, dorsal part; Cl-v, claustrum, ventral part; DP, dorsal pallium; E, entopallium; HA, hyperpallium apicale; HD, hyperpallium densocellulare; Hp, hippocampus; L2, field L2; LP, lateral pallium; MD, dorsal mesopallium; MP, medial pallium; MV, verntral mesopallium; OB, olfactory bulb; Pul, pulvinar nucleus; Rt, nucleus rotundus; Sc, superior colliculus; TeO, optic tectum; Tn, nucleus taenia; VP, ventral pallium. Adapted from Jarvis ED, Gunturkun O, Bruce L, et al. (2005) Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience* 6: 151–159, with permission.

new avian terminology is compatible with adoption of any one-to-one homology hypothesis should future evidence be more convincing.

Avian Cognition and Brain Function

Based on the modern view, the adult avian pallium, as in mammals, comprises approximately 75% of the telencephalic volume. This realization of a relatively large and well-developed avian pallium that processes information in a similar manner to mammalian sensory and motor cortices may help explain some of the cognitive abilities of birds. As summarized by the Avian Brain Nomenclature Consortium, pigeons can memorize up to 725 different visual patterns, learn to categorically discriminate objects as 'human-made' versus 'natural,' discriminate cubistic and impressionistic styles of painting, communicate using visual symbols, rank patterns using transitive inferential logic, and occasionally 'lie.' New Caledonian crows make tools out of leaves or novel human-made material, use them appropriately to retrieve food, and are believed to pass this knowledge on to other crows through social learning. Magpies develop an understanding

of object constancy at an earlier relative age in their life span than any other organism tested and can use this skill to the same extent as humans. Scrub-jays show episodic memory, recall for events that take place at a specific time or place, once thought to be unique to humans. This same species modifies its food-storing strategy according to the possibility of future stealing by other birds and therefore displays a behavior that would qualify as theory of mind. Owls have a highly sophisticated capacity for sound localization, used for nocturnal hunting, that rivals that of humans and that is developed through learning. Parrots, hummingbirds, and oscine songbirds possess the rare skill of vocal learning. This trait is a critical substrate in humans for spoken language, and with the exceptions of cetaceans and possibly bats and elephants, it is not found in any other mammal. Parrots, in addition, can learn human words and use them to communicate reciprocally with humans. African gray parrots in particular can use human words in numerical and relational concepts – abilities once thought unique to humans. These vocal behaviors are controlled by vocal learning pathways through both pallial and subpallial regions.

In general, these cognitive functions are carried out by the telencephalon, including the six-layered cortex in mammals but by nuclear pallial areas in birds. That is, the mammalian six-layered cortical architecture is not the only neuroarchitectural solution for the generation of complex cognitive behaviors. Pallial-cortical folding is also not required. Birds apparently cannot use cortical folding because of the nuclear organization of their telencephalon; among mammals, such folding is related more to absolute brain size than to behavioral complexity. Rather, the presence of specific brain subdivisions and connections is a more important factor for the generation of behavioral complexity. This suggests that there is an avian method to perform complex behaviors and a mammalian method.

Conclusion

The avian telencephalon is thought to have been inherited from its common ancestors with reptiles and mammals as a package containing pallidal, striatal, and pallial domains. It conserved the pallidal and striatal domains but modified the pallial domains. The pallial domain still performs similar functions as the pallial regions of mammals, such as the six-layered cortex and amygdala. To determine more specific homologies between avian and mammalian pallia, additional investigations are necessary, and such investigations should help determine the basic functions of specific cell types in the vertebrate brain.

See also: Basal Ganglia: Evolution; Bird Song Systems: Evolution; Brain Asymmetry: Evolution; Brain Connectivity and Brain Size; Brain Evolution: Developmental Constraints and Relative Developmental Growth.

Further Reading

Ariëns Kappers CU, Huber CG, and Crosby EC (1936) *Comparative Anatomy of the Nervous System of Vertebrates, including Man*. New York: Hafner (Reprinted 1960).

Bruce LL and Neary TJ (1995) The limbic system of tretopods: A comparative analysis of cortical and amygdalar populations. *Brain, Behavior and Evolution* 46: 224–234.

Butler AB and Cotterill RM (2006) Mammalian and avian neuroanatomy and the question of consciousness in birds. *Biology Bulletin* 211: 106–127.

Carroll RL (1987) Fossils and relationships. In: *Vertebrate Paleontology and Evolution*, pp. 1–13. New York: Freeman.

Doupe AJ, Perkel DJ, Reiner A, and Stern EA (2005) Birdbrains could teach basal ganglia research a new song. *Trends in Neurosciences* 28: 353–363.

Evans SE (2000) General discussion II: Amniote evolution. In: Bock GR and Cardew G (eds.) *Evolutionary Developmental Biology of the Cerebral Cortex*, no. 228, pp. 109–113. Chichester, UK: Wiley.

Gunturkun O (2006) Avian cerebral asymmetries: The view from the inside. *Cortex* 42: 104–106.

Jarvis ED, Gunturkun O, Bruce L, et al. The Avian, Brain Nomenclature, Consortium (2005) Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience* 6: 151–159.

Karten HJ (1991) Homology and evolutionary origins of the 'neocortex.' *Brain, Behavior and Evolution* 38: 264–272.

Medina L and Reiner A (2000) Do birds possess homologues of mammalian primary visual, somatosensory and motor cortices? *Trends in Neurosciences* 23: 1–12.

Molnar Z and Butler AB (2002) Neuronal changes during forebrain evolution in amniotes: An evolutionary developmental perspective. *Progress in Brain Research* 136: 21–38.

Northcutt RG (2001) Changing views of brain evolution. *Brain Research Bulletin* 55: 663–674.

Puelles L, Kuwana E, Puelles E, et al. (2000) Pallial and subpallial derivatives in the embryonic chick and mouse telencephalon, traced by the expression of the genes Dlx-2, Emx-1, Nkx-2.1, Pax-6, and Tbr-1. *Journal of Comparative Neurology* 424: 409–438.

Reiner A, Perkel DJ, Bruce LL, et al. (2004) Revised nomenclature for avian telencephalon and some related brainstem nuclei. *Journal of Comparative Neurology* 473: 377–414.

Reiner A, Perkel DJ, Bruce L, et al. (2004) The Avian Brain Nomenclature Forum: A new century in comparative neuroanatomy. *Journal of Comparative Neurology* 473: E1–E6.

Reiner A, Perkel DJ, Mello CV, and Jarvis ED (2004) Songbirds and the revised avian brain nomenclature. *Annals of the New York Academy of Sciences* 1016: 77–108.

Striedter GF (1997) The telencephalon of tetrapods in evolution. *Brain, Behavior and Evolution* 49: 179–213.

Relevant Websites

http://avianbrain.org – AvianBrain.org.

http://www9.biostr.washington.edu – University of Washington, Digital Anatomist Project.

Bird Song Systems: Evolution

E D Jarvis, Duke University Medical Center, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Bird song systems refer to specific neural systems found in those species of birds that produce learned vocalizations, usually learned songs. This vocal learning trait is rare, as it is found to date in only three distantly related groups of birds (parrots, hummingbirds, and songbirds) and at least four distantly related groups of mammals (cetaceans, bats, elephants, and humans). Vocal learning is a critical behavioral substrate for spoken human language. Brain pathways for vocal learning have been studied in the three vocal learning bird groups, where each appears to have seven similar but not identical cerebral vocal nuclei distributed into two vocal pathways: a posterior vocal pathway, which in songbirds is used for producing learned song, and an anterior vocal pathway, which in songbirds is used for learning song. These vocal pathways are not found in vocal nonlearning birds but are similar to brain pathways used for other types of sensorimotor learning in mammals. An auditory pathway that includes the cerebrum is present in all birds and is used to perceive the vocalizations and possibly for auditory learning. This article presents the organization and function of these pathways in vocal learning birds and discusses hypotheses about their evolution.

Vocal Learning Behavior

The unique aspect of bird song systems is that they are involved in vocal learning. Vocal learning is the ability to acquire vocalizations through imitation rather than instinct. It is distinct from auditory learning, which is the ability to make associations with sounds heard. Auditory learning occurs when a dog learns to associate the words 'sit' (English), *sientese* (Spanish), or *osuwari* (Japanese) with the act of sitting. The dog understands the word but cannot imitate the sound of the word. Vocal learners, such as humans, parrots, and some songbirds, are able to imitate noninnate sounds such as sit, *sientese*, or *osuwari*. Most vocal learners, however, imitate sounds of their own species that have been passed on through cultural transmission. Vocal learning depends on auditory learning, but auditory learning does not depend on vocal learning. Vocal learners must hear the sounds they will later imitate. They use auditory feedback to correct their vocal output by comparing the output with auditory memories of the sounds they are trying to imitate.

Given the above definitions, most, if not all, vertebrates are capable of auditory learning, but few are capable of vocal learning. The latter has been found experimentally to date only in the three distantly related groups of birds (parrots, hummingbirds, and songbirds, Figure 1(a)), and the four distantly related groups of mammals (cetaceans, bats, elephants, and humans). There is some suggestion that sea lions are also vocal learners. Each of the vocal learning avian (Figure 1(a)) and mammalian groups has close vocal nonlearning relatives. Thus, it has been argued that vocal learning has evolved independently of a common ancestor in the three vocal learning bird groups and presumably in the vocal learning mammalian groups. The question that thus arises is whether there is something special about the brains of animals that can imitate sounds.

Vocal Learning Song Systems

Only vocal learners – such as songbirds, parrots, hummingbirds, and humans – have brain regions in their cerebrums that control the acoustic structure and syntax of their vocalizations. In birds, these regions make up what are called song systems because they control learned song. Remarkably, in all three vocal learning bird groups, their song systems consist of seven comparable vocal brain nuclei: four posteriorly located nuclei and three anteriorly located nuclei (Figure 1(b); abbreviations in Table 1). These vocal nuclei have been given different names in each bird group because of the possibility that each group evolved its vocal nuclei independent of a common ancestor with such nuclei.

The connectivity of the vocal nuclei has been studied best in songbirds (Figure 2(a)), but sufficient information exists about parrots and some about hummingbirds to make comparisons (Figure 1(b)). In all three vocal learning bird groups, the posterior nuclei form a posterior vocal pathway that projects from a nidopallial vocal nucleus (songbird HVC, parrot NLC, hummingbird VLN) to an arcopallial vocal nucleus (songbird RA, parrot AAC dorsal part, hummingbird VA), to the midbrain premotor (DM) and medulla vocal motor (nXIIts) neurons (Figures 1(b) and 2(a), black arrows). The nXIIts projects to the muscles of the syrinx, the avian vocal organ.

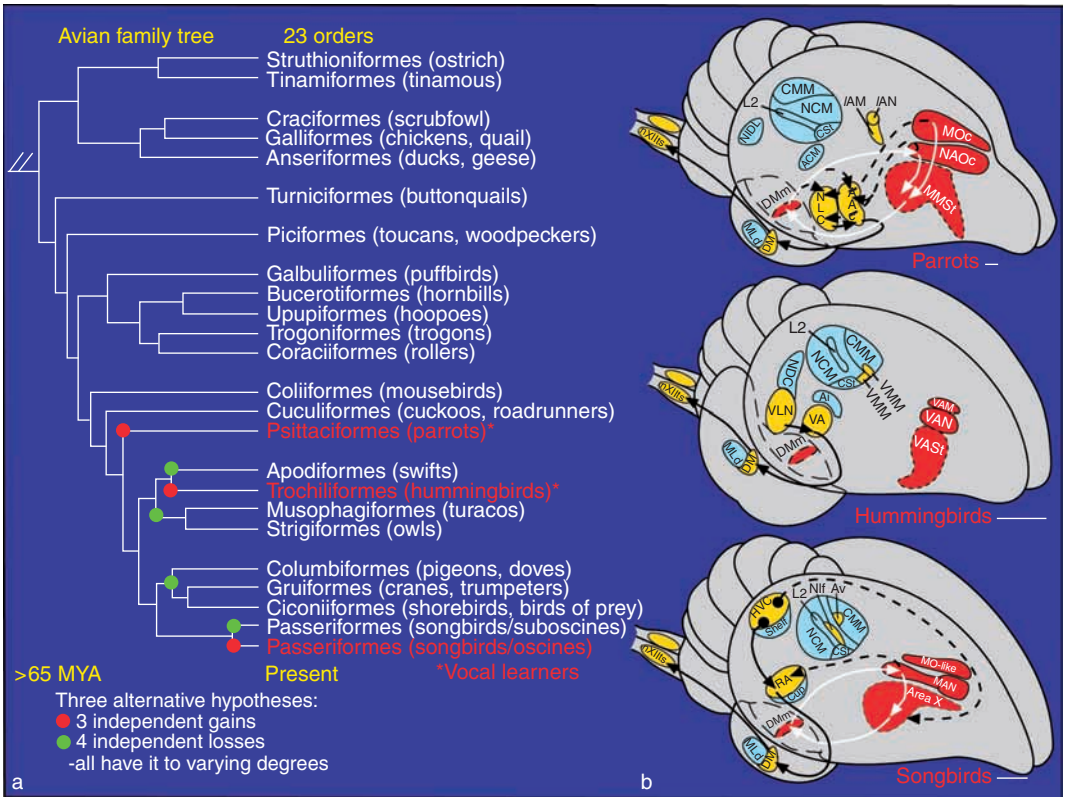

Figure 1 Phylogenetic relationship of birds and comparison of song pathways in vocal learners. (a) Family tree of living avian orders. The tree was derived from DNA–DNA hybridization analysis by Sibley CG and Ahlquist JE (1990) *Phylogeny and classification of birds: A study in molecular evolution*. New Haven, CT: Yale University Press. The Latin name of each order is given, along with examples of common species. Passeriformes are divided into two suborders: suboscine and oscine songbirds. MYA, millions of years ago. Red and green circles show the minimal ancestral nodes where vocal learning could have either evolved independently or been lost independently. Independent losses would have required at least one common vocal learning ancestor prior to parrots. The tree is not meant to present the final dogma of avian evolution, as there are differences of opinion among scientists. (b) Proposed comparable vocal and auditory brain areas among vocal learning birds. Yellow regions and black arrows indicate proposed posterior vocal pathways; red regions and white arrows indicate proposed anterior vocal pathways; dashed lines show connections between the two vocal pathways; blue indicates auditory regions. For simplification, not all connections are shown. Basal ganglia, thalamic, and midbrain regions have broken-line boundaries to indicate that they are deeper in the brain relative to the anatomical structures above them. Scale bar = ~7 mm. Abbreviations are in **Table 1**. Adapted from Jarvis ED, Ribeiro S, da Silva ML, et al. (2000) Behaviourally driven gene expression reveals song nuclei in hummingbird brain. *Nature* 406: 628–632.

The connectivity of the other two cerebral posterior vocal nuclei in hummingbirds is not known, nor has it been well studied in parrots. In songbirds, one of these nuclei (NIf) projects to HVC, and the other (Av) receives a projection from HVC. The posterior vocal pathway is responsible for production of learned song and calls (determined only in songbirds and parrots); songbird HVC is thought to generate syntax and RA the acoustic structure of syllables. Vocal nonlearning birds have DM or nXIIts, but without projections from the arcopallium, for production of innate songs and calls. The posterior vocal pathway is also called the vocal motor pathway because of its role in song production.

The anterior nuclei (connectivity examined only in songbirds and parrots) are part of an anterior vocal pathway loop, in which a pallial vocal nucleus (songbird MAN, parrot NAOc) projects to a striatal vocal nucleus (songbird Area X, parrot MMSt), the striatal vocal nucleus projects to a nucleus of the dorsal thalamus (songbird DLM, parrot DMM), and the dorsal thalamus projects back to the pallial vocal nucleus (songbird MAN, parrot NAOc; **Figures 1(b)** and **2(a)**, white arrows). The parrot pallial MO nucleus also projects to the striatal vocal nucleus (MMSt). Connectivity of the songbird MO-like analog has not yet been determined. The anterior vocal pathway is responsible for learning song (determined mostly in songbirds); lateral MAN (LMAN) is thought to generate variability and Area X to generate stereotypy. Functions of the other nuclei of the anterior vocal pathways are not yet well studied. The anterior

Table 1 Abbreviations

AAC: central nucleus of the anterior arcopallium
AACd: central nucleus of the anterior arcopallium, dorsal part
AACv: central nucleus of the anterior arcopallium, ventral part
ACM: caudal medial arcopallium
Ai: intermediate arcopallium
Area X: area X of the striatum
Av: avalanche
B: basorostralis
CM: caudal mesopallium
CMM: caudal medial mesopallium
CSt: caudal striatum
DLM: medial nucleus of dorsolateral thalamus
DM: dorsal medial nucleus of the midbrain
DMM: magnocellular nucleus of the dorsomedial thalamus
E: entopallium
HVC: (a letter-based name)
L1: field L1
L2: field L2
L3: field L3
LLD: lateral lemniscus, dorsal nucleus
LLI: lateral lemniscus, intermediate nucleus
LLV: lateral lemniscus, ventral nucleus
LMAN: lateral part of MAN
MAN: magnocellular nucleus of anterior nidopallium
MLd: mesencephalic lateral dorsal nucleus
MMAN: medial part of MAN
MMSt: magnocellular nucleus of the anterior striatum
MO: oval nucleus of the mesopallium
MOc: oval nucleus of the mesopallium complex
NAOc: oval nucleus of the anterior nidopallium complex
NCM: caudal medial nidopallium
NDC: caudal dorsal nidopallium
NIDL: intermediate dorsal lateral nidopallium
NIf: interfacial nucleus of the nidopallium
NLC: central nucleus of the lateral nidopallium
nXIIts: tracheosyringeal subdivision of the hypoglossal nucleus
Ov: nucleus ovoidalis
PAm: nucleus parambigualis
RA: robust nucleus of the arcopallium
RAm: nucleus retroambigualis
SO: superior olive
St: striatum
Uva: nucleus uvaeformis
VA: vocal nucleus of the arcopallium
VAM: vocal nucleus of the anterior mesopallium
VAN: vocal nucleus of the anterior nidopallium
VAS: vocal nucleus of the anterior striatum
VA/VL: ventral anterior/ventral lateral nuclei of the mammalian thalamus.
VLN: vocal nucleus of the lateral nidopallium
VMM: vocal nucleus of the medial mesopallium
VMN: vocal nucleus of the medial nidopallium

vocal pathway is sometimes called the vocal learning pathway but is most often called the anterior forebrain pathway (AFP) because of its anterior location.

One major difference among vocal learning birds is in the connections between the posterior and anterior vocal pathways. In songbirds, the posterior vocal pathway sends input to the anterior vocal pathway via HVC to Area X; the anterior pathway sends output to the posterior pathway via LMAN to RA and medial MAN (MMAN) to HVC (Figures 1(b) and 3(a)). In contrast, in parrots, the posterior pathway sends input into the anterior pathway via ventral AAC (AACv, parallel of songbird RA) to NAOc (parallel of songbird MAN) and MO; the anterior pathway sends output to the posterior pathway via NAOc to NLC (parallel of songbird HVC) and AAC (Figures 1(b) and 3(b)). An important distinction in this regard is that the parrot posterior pathway vocal nuclei do not appear to send projections to the striatal nucleus of the anterior pathway. Another major difference is the relative location of the posterior vocal nuclei; in songbirds, they are embedded in the auditory regions; in hummingbirds, they are situated more laterally but still adjacent to the auditory regions; in parrots, they are situated far laterally and physically separate from the auditory regions (Figure 1(b)). At a minimum, the axons connecting the vocal and auditory systems would have to take different routes in the different vocal learners.

Cerebral Auditory Pathway

Vocal learning and vocal nonlearning birds have very similar auditory pathways. These similarities include projections from ear hair cells to cochlear ganglia neurons, to auditory pontine nuclei (CN, LL), to midbrain (MLd) and thalamic (Ov) nuclei, and to primary (L2) and secondary (L2, L3, NCM, and CM) pallial areas. A descending auditory feedback pathway from the dorsal nidopallium (vocal learners: songbird HVC shelf; parrot NIDL; vocal nonlearner; pigeon Nd) to the intermediate arcopallium (vocal learners: songbird RA cup, parrot ACM; vocal nonlearner: pigeon AIVM) to shell regions around thalamic (Ov) and midbrain (MLd) auditory nuclei (Figure 2(b)). An auditory region exists in the striatum (CSt) of all birds examined, but because of lack of detailed study of the connectivity of CSt in different species, it is not yet possible to make reliable comparisons for the striatal auditory region.

The source of auditory input into the vocal pathways of vocal learning birds is unclear. Various routes have been proposed. In songbirds, these include the HVC shelf into HVC, the RA cup into RA, Ov or CM into NIf, and from NIf dendrites in L2. In parrots, these include the NLC shell into NLC, and nucleus basorostralis, L1, and L3 into the comparable NIf-like nucleus LAN. More study is necessary to determine whether any of these regions truly brings auditory input into the vocal pathways.

The cerebral auditory pathway is thought to be responsible for processing complex sounds, including species-specific songs, in a hierarchical manner. Auditory stimuli are relayed from the thalamus (Ov)

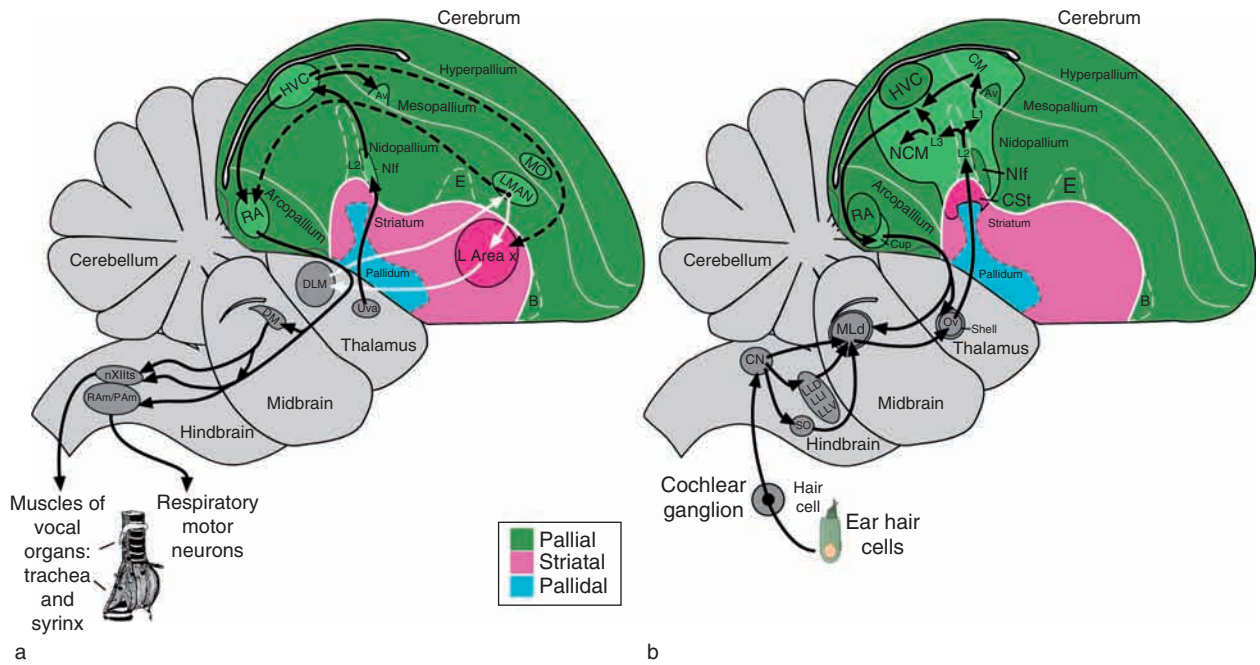

Figure 2 Comparison of vocal and auditory pathways of the songbird brain. Only the most prominent and/or most-studied connections are indicated. (a) Vocal pathway. Black arrows show connections of the posterior vocal pathway (or vocal motor pathway); white arrows, connections of the anterior vocal pathway (or pallial-basal ganglia-thalamic-pallial loop); and dashed lines, connections between the two pathways. (b) Auditory pathway. Most of the hindbrain connectivity is extrapolated from nonsongbird species. For clarity, reciprocal connections in the pallial auditory areas are not indicated. Colors denote major cerebral brain subdivisions of birds. Abbreviations are in **Table 1**. Adapted from Jarvis ED, Güntürkün O, Bruce L, et al. (2005) Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience* 6: 151–159.

to the cerebrum, where L2 first performs a simple processing of the acoustic features of sounds, then L1, L3, and NCM process more-complex aspects of sounds, such as sequencing and discriminations, and finally CM, the highest station, processes the most complex aspects, such as fine discriminations. In this regard, NCM and CM are also thought to be involved in auditory learning for sound associations and discriminations, and have been proposed to be involved in auditory memory formation of songs used for vocal learning. This hypothesis still needs verification.

Evolution of Vocal Learning Pathways

In the simplest of interpretations, there are three alternative explanations of the way similar, although not identical, vocal learning brain pathways could have evolved among birds (**Figure 1(a)**): (1) the vocal systems in the three vocal learning bird groups evolved independently of a common ancestor. If this is true, then the similarities are remarkably coincidental. (2) In the common ancestor of vocal learning birds, there was a vocal learning pathway with seven cerebral nuclei, which were then lost multiple independent times in closely related bird groups. If this were true, then the similarities are not as remarkable, but it would suggest that there is strong selection to extinguish vocal learning. (3) Most, if not all birds have vocal learning to various degrees, and songbirds, parrots, and hummingbirds independently amplified the associated brain pathways for their more highly developed vocal learning behaviors. If this were true, it would mean that many birds have at least primordial brain structures for vocal learning. So far, the evidence for this last hypothesis has been lacking.

One proposed answer is that different factors are dependent on or independent of a common ancestor. A proposed dependent factor is preexisting connectivity and function. The connections of the anterior and posterior vocal pathways resemble nonvocal motor pathways in mammals (**Figure 3(c)**), and the regions around the vocal nuclei form connections similar to those of the vocal nuclei and are active during production of nonvocal motor behavior. A proposed independent factor is that out of the proposed preexisting motor system, the vocal learning systems for song in the three bird groups evolved independently.

Specifically, the songbird and parrot posterior vocal pathways are similar in connectivity to mammalian motor corticospinal and corticobulbar

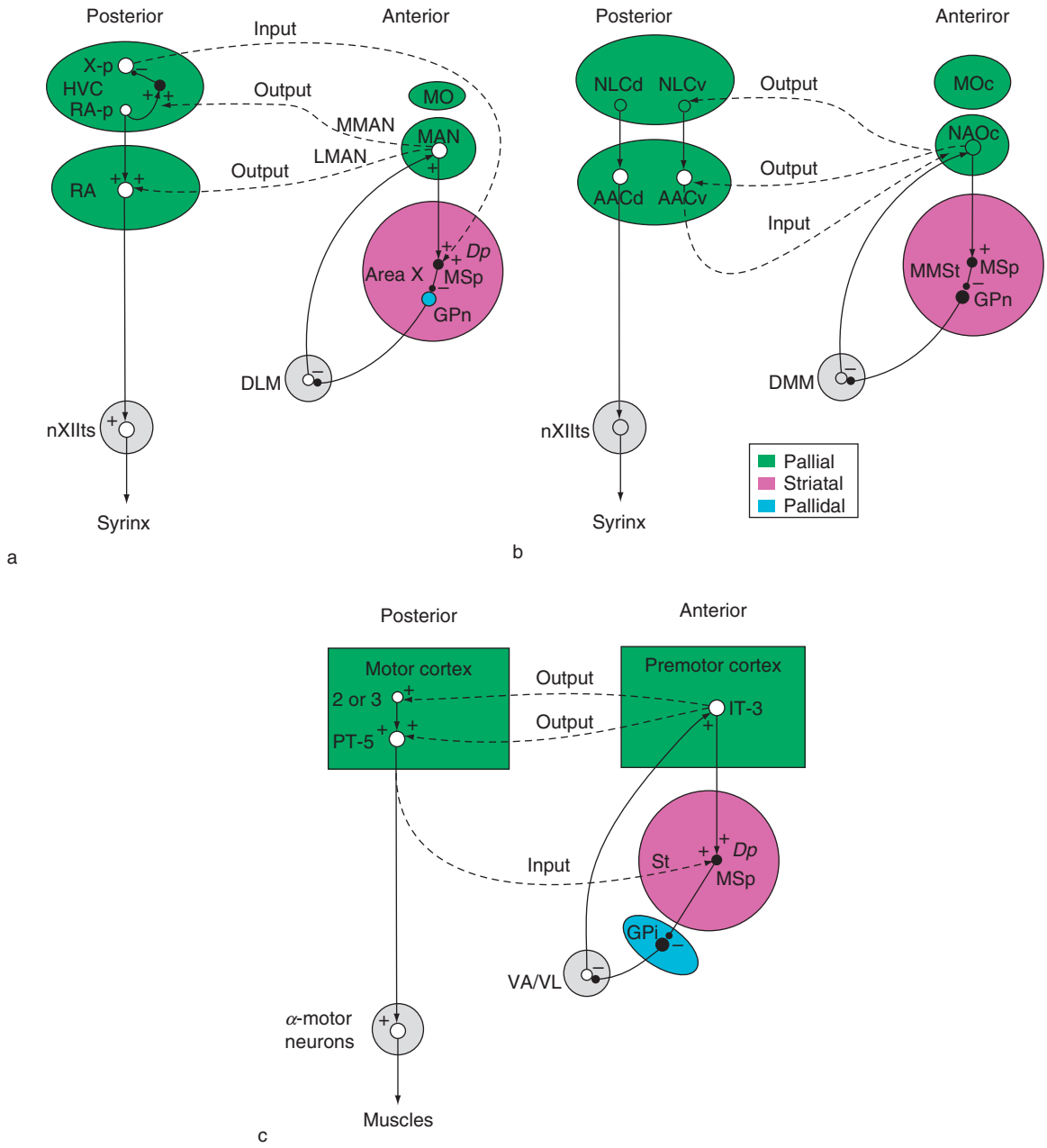

Figure 3 Comparative and simplified cellular connectivity of anterior and posterior vocal pathways in (a) songbirds and (b) parrots, and (c) motor pathways in mammals. Dashed lines: connections between anterior and posterior pathways; inputs and outputs are labeled relative to anterior pathways. Output from songbird MAN to HVC and RA is not from the same neurons; MMAN neurons project to HVC, and LMAN neurons project to RA. **O**, excitatory neurons; •, inhibitory neurons; +, excitatory glutamate neurotransmitter release; −, inhibitory GABA release; MSp, medium spiny neuron; GPn, globus pallidus-like neuron in songbird Area X and parrot MMSt. Only the direct pathway through the mammalian basal ganglia (St to GPi) is shown as this is the pathway most similar to Area X connectivity (MSp to GPn). X-p, X-projecting neuron of HVC; RA-p, RA-projecting neuron of HVC; PT-5, pyramidal tract neuron of motor cortex layer 5; IT-3, intratelencephalic projecting neuron of layer 3. Abbreviations are in **Table 1**. Adapted from Jarvis ED (2004) Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Science* 1016: 749–777.

pathways (Figures 3(a)–3(c)). Projection neurons of songbird RA and parrot AACd are similar to pyramidal tract (PT) neurons of lower layer 5 of mammalian motor cortex. The latter send long axonal projections out of the cerebrum through pyramidal tracts to synapse onto brain stem and spinal cord premotor and α-motor neurons that control muscle contraction and relaxation. The RA-projecting neurons of

songbird HVC and projection neurons of parrot NLC are similar to layer 2 and 3 neurons of mammalian motor cortex, which send intrapallial projections to layer 5 (Figures 3(a)–3(c)). Mammalian parallels to songbird NIf and Av are less clear. In mammals, sensory neurons from the thalamus project to layer 4 of cortex; motor feedback from the thalamus projects to layer 5 (in primary motor cortex) or layer 3 (in frontal cortex). NIf is directly adjacent to L2 and, like L2, similar to mammalian layer 4 neurons, it receives thalamic input from UVa, but like the mammalian thalamic input to layer 3, UVa has motor-associated activity, firing during vocalizing.

The songbird and parrot anterior vocal pathways are similar in connectivity to mammalian cortical-basal ganglia-thalamic-cortical loops (Figures 3(a)–3(c)). The projection neurons of songbird MAN and parrot NAO are similar to intratelencephalic (IT) neurons of layer 3 and upper layer 5 of mammalian premotor cortex, which send collateral projections, one to medium spiny neurons of the striatum ventral to it and the other to other cortical regions (Figures 3(a)–3(c)). Unlike mammals, the spiny neurons in both songbird Area X, and presumably parrot MMSt, project to pallidal-like cells within vocal nuclei Area X and MMSt instead of to a separate globus pallidus structure consisting only of pallidal cells. This striatal-pallidal cell intermingling may be a general trait of the anterior avian striatum. The projection of the pallidal-like cells of songbird Area X and parrot MMSt is similar to the motor pallidal projection neurons of the internal globus pallidus (GPi) of mammals, which project to the ventral lateral (VL) and ventral anterior (VA) nuclei of the dorsal thalamus (Figures 3(a)–3(c)). Mammalian VL/VA projects back to layer-3 neurons of the same premotor areas, closing parallel loops, like songbird DLM and parrot DMM projections to LMAN and NAO.

Because connections between the posterior and anterior vocal pathways differ between songbirds and parrots, comparisons between them and mammals will also differ. Input into mammalian cortical-basal ganglia-thalamic-cortical loops is from collaterals of PT–layer-5 neurons of motor cortex, one that projects to the brain stem and spinal cord and the other that projects into the striatum (Figure 3(c)). In the songbird, however, a specific cell type of HVC, called X-projecting neurons, projects to the striatum separately from those neurons (neurons of RA) that project to the medulla. Parrot vocal connectivity is even more different from mammalian connectivity. In parrots, AAC of the arcopallium has two anatomically separate neuron populations, AACd, projecting to the medulla, and AACv, projecting to anterior pallial vocal nuclei NAO and MO. Output of mammalian anterior pathways are the collaterals of IT-layer 3 and IT-upper layer 5 neurons that project to other cortical regions (Figure 3(c)).

In terms of function, recent results suggest that vocal nuclei of vocal learning birds are embedded within at least seven brain areas active during the production of limb and body movements. The same movement-associated brain areas are found in vocal learning birds, such as ring doves. These and related findings have led to a motor theory for the origin of vocal learning whereby in the avian brain, a preexisting motor system in a vocal nonlearner ancestor is proposed to have consisted of seven brain regions distributed across mesopallial, nidopallial, arcopallial, and striatal brain subdivisions and separated into two pathways: an anterior premotor pathway that forms a pallial-basal ganglia-thalamic-pallial loop and a posterior motor pathway that sends descending projections to brain stem and spinal cord premotor and α-motor neurons. Then a mutational event or events might have caused descending projections of avian arcopallium neurons that normally synapse onto nonvocal motor neurons to instead synapse onto nXIIts neurons in vocal learners. Thereafter, cerebral vocal brain regions could have developed out of adjacent motor brain regions using the preexisting connectivity. Such a mutational event would be expected to occur in genes that regulate synaptic connectivity of pallial motor neurons to α-motor neurons. This hypothesis needs to be tested with more experiments before it can be accepted with more confidence. If it is true, then this hypothesis needs to account for how auditory information enters the vocal pathway during evolution of a vocal learning system.

Evolution of the Auditory Pathway

The auditory pathway described in birds has a number of similarities with auditory pathways in reptiles and mammals (Figure 4). Birds, reptiles, and mammals have ear hair cells that synapse onto sensory neurons, which project to cochlea and lemniscal nuclei of the brain stem, which in turn project to midbrain (avian MLd, reptile torus, mammalian inferior colliculus) and thalamic (avian Ov, reptile reunions, mammalian medial geniculate) auditory nuclei. The thalamic nuclei in turn project to primary auditory cell populations in the pallium (avian L2, reptile caudal pallium, mammalian layer 4 of primary auditory cortex). For connectivity in the cerebrum, detailed studies exist only for mammals and birds. It has been proposed that avian L1 and L3 neurons are similar to mammalian layers 2 and 3 of primary auditory cortex, the

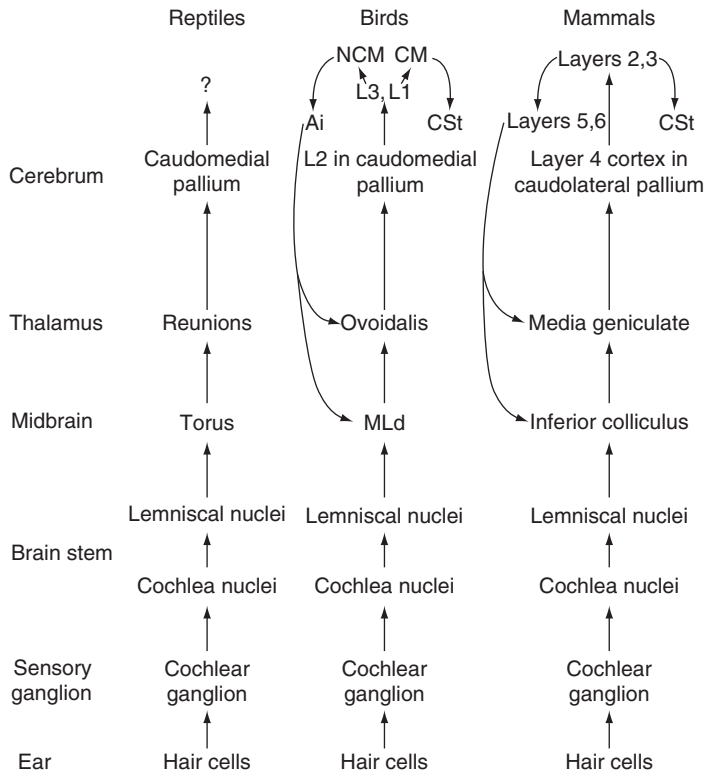

Figure 4 Comparative and simplified connectivity among auditory pathways in reptiles, birds, and mammals. The connectivity from CM to CSt in birds needs verification by retrograde tracing. Abbreviations are in **Table 1**. From Jarvis ED (2004) Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Science* 1016: 749–777.

latter of which receives input, like L2, from layer 4. Avian NCM and CM have been proposed to be similar to layers 2 and 3 in that they form reciprocal intrapallial connections with each other and receive input from L2. Neurons of the songbird RA cup, parrot ACM, and pigeon AIVM of the arcopallium are similar to mammalian layers 5 and 6, which send descending auditory feedback projections to the shell regions of mammalian thalamic and midbrain auditory nuclei. Avian CSt may be similar to an auditory region in mammals that is located caudally in the striatum. The connectivity of this region has not been verified in birds or mammals.

The similarities suggest that the auditory pathways of birds, reptiles, and mammals may be homologous. Further, since this auditory pathway exists in all birds studied, the auditory pathway in vocal learning birds was presumably inherited from their common avian ancestor, which inherited it from their common stem-amniote ancestor with other vertebrates, thought to have lived ~320 million years ago. A cerebral auditory pathway would explain why birds and nonhuman mammals, including dogs, exhibit auditory learning, including learning to understand the meaning of simple human speech, although presumably with less facility than a human has. In this regard, when a vocal learning system evolves, it presumably uses the preexisting auditory pathway to bring auditory information into it.

Evolution of Vocal Learning Behavior

The previous sections discuss neural or genetic-based factors that may have influenced the evolution of bird song systems. This last section discusses the possible epigenetic, or environmental/behavioral factors. A number of hypotheses have been proposed to explain what selects for vocal learning. These include: (1) individual identification, (2) semantic communication, (3) territory defense, (4) mate attraction, and (5) rapid adaptation to sound propagation in different environments. For individual identification, the argument is that among vocal learners, each individual can learn its own song and thus have a unique signature that differentiates it from others of its population. However, individual identification by voice is not a unique characteristic of vocal learners. As no two individuals look identical, neither do any two individuals within an avian or mammalian population sound identical, allowing vocal nonlearners to identify individual conspecifics by voice. It is true

that with vocal learning, individuals can be called by unique names. This, however, has so far been found among humans and maybe bottlenose dolphins but no other vocal learners. Thus, it is unlikely that individual identification was a primary driving force for the selection of vocal learning.

A common notion is that a selective advantage of language is semantic communication. Semantic vocal communication means that animate or inanimate objects have a specific vocalization or series of vocalizations associated with them. It contrasts with affective vocal communication, which means that vocalizations have emotional content associated with them. Despite this common notion, many vocal nonlearners, such as chickens and vervet monkeys, use innate calls to communicate semantic information, such as 'an eagle above,' 'a snake on the ground,' or 'a food source.' Vocal learning birds also use innate calls with such semantic information and only rarely have been found to use learned vocalizations for mainly semantic communication. Examples of the latter include the African Grey parrot's use of human speech words and the black-capped chickadee's use of calls to indicate predator size. Since this is not the common use of learned vocalizations among vocal learners, semantic communication was probably not a major driving force for the selection of vocal learning.

Vocal learners use their learned vocalizations most commonly in affective contexts, to defend territories and/or attract mates. These uses are found among the songbirds, parrots, and hummingbirds. Many vocal nonlearners use their innate calls and crows to defend territories and attract mates; examples include the crowing of male doves and chickens. There is, however, an additional feature of mate attraction that vocal learners use learned vocalizations for: variability. Vocal learners, but not vocal nonlearners, have the ability to produce more-varied frequency modulation (FM) and syntax, during vocal development and/or, in some species, after reaching adulthood. These variations in FM and syntax are preferred by females of the songbird species that have been examined. In addition, canaries have been shown to use their two voices, common among birds, to produce even greater FM variations, thought to stimulate estrogen production in the listening females; these are sometimes called sexy syllables or songs. In fact, in vocal nonlearners, such as ring doves, artificially increased computer-generated FM of innate sounds is more stimulatory for mating than are the innate coos. Therefore, birds with the ability to produce more vocal variety are likely to be selected for this trait. Once the ability to produce variable sounds is selected for, then the ability of vocal learning maybe selected for. In this regard, mate attraction was probably a major driving force in the selection for vocal learning.

For sound transmission, vocal nonlearners produce their vocalizations best in specific habitats, which makes their vocal behaviors less adaptable to changes in the environment. For example, a pigeon's low-frequency vocalizations travel best near the ground, whereas an eastern phoebe's higher-pitched vocalizations travel better higher in the air. In contrast, vocal learners have the ability to change voice characteristics, either during the lifetime of an individual or through several generations, presumably allowing better group communication in different environments.

If mate attraction for varied sounds and a minimal mutational event to cause motor pallial areas to synapse onto α-vocal motor neurons is what it takes to evolve vocal learning, then this still does not explain why it is so rare. One proposed explanation is that predation is a strong selection factor against vocal learning. If more-varied syntax is attractive to mates, it may also be more attractive to predators. As innate vocalizations tend to be more constant, they may be naturally habituated to more easily, potentially becoming part of the background noise. Therefore, in order for a predator not to habituate to the sounds of his prey, he would have to evolve a neural mechanism to overcome the natural habituation at times when he is hungry. If this were the case, then a species would have to overcome predatory pressure or have a relaxed predatory pressure before it could evolve vocal learning. In this regard, some mammalian vocal learners have very few if any major predators (humans, whales, and elephants). Hummingbirds, because of their rapid flight and escape behavior, are also commonly known to be fearless of many animals. Various parrot species and ravens (a corvid songbird) are prey to hawks and large owls, like other birds, but they are said to be highly effective at mobbing behavior to evade such predators. A cautionary note: it is not known how valid the distinction is between vocal learners and nonlearners in terms of predation, as no systematic study has been done on prey and predatory behaviors across vocal learning and vocal nonlearning groups.

Some findings support this association. Okanoya and colleagues have shown that Bengalese finches that have been bred in captivity, without predators, for the past 250 years and without human selection for singing behavior show more-varied syntax than their white-backed munia conspecifics still living in the wild, from which they derive. Zann has shown that zebra finches bred in captivity show more variation in the songs learned among adults of a colony than do their wild-type conspecifics. For both species, females prefer the more varied songs. Given these findings, one

would expect to find more syntax complexity selected for in the wild than currently exists. Perhaps predatory pressure is selecting against it.

In summary, the investigations into the evolution of bird song systems are an active area of study, often leading to insights and to novel hypotheses about evolution of brain systems for complex behaviors, including spoken language. A great amount of knowledge has been obtained, but a great amount more is needed before various hypotheses can be rejected or new ones generated.

See also: Auditory System: Central Pathways; Bird Brain: Evolution; Birdsong Learning: Evolutionary, Behavioral, and Hormonal Issues; Birdsong Learning; Birdsong: The Neurobiology of Avian Vocal Learning; Signal Production and Amplification in Birds; Vocal Communication in Birds.

Further Reading

Brenowitz EA (1997) Comparative approaches to the avian song system. *Journal of Neurobiology* 33: 517–531.

Catchpole CK and Slater PJB (1995) *Bird Song: Biological Themes and Variations*. Cambridge, UK: Cambridge University Press.

Doupe AJ, Perkel DJ, Reiner A, and Stern EA (2005) Birdbrains could teach basal ganglia research a new song. *Trends in Neuroscience* 28: 353–363.

Farries MA (2001) The oscine song system considered in the context of the avian brain: Lessons learned from comparative neurobiology. *Brain Behavior and Evolution* 58: 80–100.

Janik VM and Slater PJB (1997) Vocal learning in mammals. *Advances in the Study of Behavior* 26: 59–99.

Jarvis ED (2004) Brains and birdsong. In: Marker P and Slabbekoorn H (eds.) *Nature's Music: The Science of Birdsong*, pp. 226–271. San Diego, CA: Elsevier-Academic Press.

Jarvis ED (2004) Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Science* 1016: 749–777.

Jarvis ED (2006) Selection for and against vocal learning in birds and mammals. *Ornithology Science* 5: 5–14.

Jarvis ED, Güntürkün O, Bruce L, et al. (2005) Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience* 6: 151–159.

Jarvis ED, Ribeiro S, da Silva ML, et al. (2000) Behaviourally driven gene expression reveals song nuclei in hummingbird brain. *Nature* 406: 628–632.

Marler P and Slabbekoorn H (eds.) (2004) *Nature's Music: The Science of Birdsong*. San Diego, CA: Elsevier-Academic Press.

Nottebohm F (1972) The origins of vocal learning. *American Naturalist* 106: 116–140.

Okanoya K (2002) Sexual display as a syntactical vehicle: The evolution of syntax in birdsong and human language through sexual selection. In: Wray A (ed.) *The Transition to Language*, pp. 46–63. Oxford, UK: Oxford University Press.

Sibley CG and Ahlquist JE (1990) *Phylogeny and classification of birds: A study in molecular evolution*. New Haven, CT: Yale University Press.

Wild JM (1997) Neural pathways for the control of birdsong production. *Journal of Neurobiology* 33: 653–670.

Zann RA (1996) *The Zebra Finch: A Synthesis of Field and Laboratory Studies*. New York: Oxford University Press.

Zeigler HP and Marler P (eds.) (2004) *Special Issue: Behavioral Neurobiology of Birdsong. Annals of the New York Academy of Science* 1016.

Birdsong Learning

I R Fiete, University of California, Santa Barbara, and Center for Neural Systems, Pasadena, CA, USA
H S Seung, Massachusetts Institute of Technology, Boston MA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Birdsong has been compared to human speech because both are examples of animal vocalizations that are at least partially learned. Indeed, some of the neural pathways underlying these vocal behaviors may be similar in songbirds and humans. More generally, birdsong can be compared to any motor behavior that involves sequential control and learning and is an ideal playground for experimental and theoretical explorations of the neural basis for motor learning and control.

The neurons involved in song production and learning are clustered in a small number of discrete nuclei with well-defined boundaries, fairly well-characterized connectivity, and increasingly well-studied neurophysiology. The various species of oscine songbirds display a vast spectrum of complexity in syllable sequencing (syntax), improvisation, and vocal acquisition. For example, song can be as simple as repetitions of a set of syllables arranged in fixed order (e.g., zebra finches), or it can be intricately improvised, with syllable ordering governed by probabilistic transitions between different syllables (e.g., domesticated bengalese finches). Mockingbirds can acquire novel songs and learn novel nonbird sounds throughout life. Yet the basic neural song circuit is similar in all oscines. This diversity of behavior subserved by an underlying similarity of structure makes birdsong an intriguing system for future comparative studies of sophisticated motor sequencing and control.

Models have so far focused on the zebra finch, a songbird with one of the simplest of oscine song repertoires, because most recordings of neural activity have been obtained from these birds. The male zebra finch learns one song motif with fixed syllable order and sings repetitions of that motif throughout its life. Song acquisition takes place in a sensory phase, when the juvenile bird acquires a mental template of the song of a possibly unrelated adult tutor, and an overlapping sensorimotor phase, when the juvenile begins to vocalize and tries to match its vocalizations with its mental template. In the laboratory the sensory and sensorimotor phases can be separated: template acquisition is fast and can be completed and the tutor bird removed from the juvenile's presence before the juvenile has begun to practice its song. Sensorimotor learning is a slow process of self-generated trial and error and can proceed successfully in complete auditory and social isolation after template acquisition. However, auditory feedback of the bird's own vocalizations is crucial during sensorimotor learning because a bird deafened at this stage cannot learn to reproduce its mental template of the tutor song.

This article focuses on the motor and sensorimotor aspects of song production, in particular the formation of neural sequences to drive song, the learning of the motor map, and the reasons song may be encoded the way it is by songbird premotor neurons. The aim is to illustrate how theoretical work has contributed to the understanding of the song system and highlight the resulting predictions for experiment. Auditory neural coding and dynamics and the physics (acoustics and mechanics) of songbird vocal production are beyond the scope of the article.

Basic Functional Anatomy of the Song Pathway

In the oscine vocal premotor circuit, high-level vocal center nucleus (HVC) projects to nucleus robustus archistriatalis (RA). RA drives the motor neurons that innervate the syringeal and respiratory organs for song production (Figure 1). Lesioning any of these nuclei immediately disables song production, while nuclei upstream of HVC are not necessary for singing in adult zebra finches. Therefore, HVC originates the top-level premotor drive for adult zebra finch song. Although the circuitry between the premotor nuclei is feedforward, connectivity within HVC and within RA is extensively recurrent. In addition to the premotor circuit, the song system contains the anterior forebrain pathway (AFP), which connects HVC to RA through an indirect path. Lesion studies reveal that the AFP is crucial for the sensorimotor phase of song learning and long-term song maintenance but not for song production. The AFP is an avian homolog of the mammalian thalamocortical–basal ganglia pathway.

Based on anatomical, neurophysiological, and lesion data, it is believed that HVC generates the spatiotemporal premotor drive for sequential motor activation in the form of sequential neural activity. The HVC activity is 'abstract,' in the sense that it encodes only temporal ordering, not song features (Figures 2(a) and 2(b)). The synapses from HVC to RA and from RA to the motor neurons convert the sequence of HVC activity patterns into the sequence of motor commands that produces song (Figure 2).

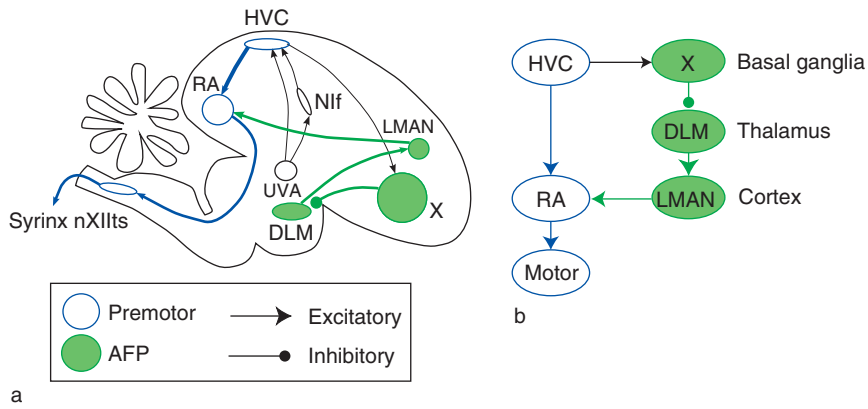

Figure 1 The neural pathways underlying song learning and song production. (a) Song learning and production involves a discrete set of nuclei. (b) The same areas, unfolded to better illustrate their connectivity. Lines terminating in arrows designate excitatory or putative excitatory connectivity. Lines terminating in circles designate inhibitory synapses. Blue nuclei are part of the premotor pathway. Green nuclei are part of the anterior forebrain pathway (AFP) and are important for sensorimotor song learning but not song production. HVC, high-level vocal center nucleus; LMAN, lateral magnocellular nucleus of the anterior nidopallium; NIf, interfacial nucleus; RA, nucleus robustus archistriatalis.

According to this perspective, there are two types of learning. First, HVC acquires the capability of generating a long sequence of neural activity patterns (Figures 2(f) and 2(g)). Second, the circuitry downstream from HVC acquires the capability of mapping HVC activity patterns to the correct sound-producing patterns of activity in RA. The HVC–RA synapses show extensive structural plasticity, with synaptic growth followed by massive retraction, precisely during the sensorimotor learning period. Electrophysiological synaptic maturation accompanies such structural remodeling. Therefore, motor map learning from HVC to the vocal outputs is likely to involve plasticity in the HVC–RA synapses. Other possible loci of plasticity for sensorimotor learning, not ruled out by experiments, include the recurrent RA–RA synapses or the RA–motor neuron connections.

Sequence Generation

The high-level premotor area HVC displays clear sequential patterns of neural activation. In the awake singing zebra finch, each recorded RA-projecting HVC neuron (also called an HVCRA neuron) fires at most one brief burst of spikes (lasting ≈6 ms) per second-long song motif (Figure 2(a)). These neurons are otherwise inactive during singing. Different RA-projecting HVC neurons fire their single bursts at different fixed times in the motif. Thus as a population, the HVCRA neurons cover time during song.

In addition to RA-projecting neurons, HVC contains inhibitory interneurons which ramify within HVC and X-projecting neurons which send their outputs to the AFP. Inhibitory interneuron activity is very different from HVCRA activity: these neurons fire tonically throughout song, with high rates and relatively little modulation in firing frequency across song (Figure 2(b)).

In the adult zebra finch, deafening causes no immediate effect on song production, suggesting that auditory feedback is not important for sequence generation. Also, lesion studies indicate that input from the higher interfacial nucleus to HVC is not necessary for singing in zebra finches. What network connectivity and neural dynamics underlie the sequence generation dynamics observed in the top-level premotor area, and how might that connectivity become established as hatchlings mature?

The Synaptic Chain Model

The 'synaptic chain' model is an old idea for how networks of neurons might produce sequential patterns of activity. (The idea was already old when it was critiqued by Lashley in 1951 under the name 'associative chaining' model.) In the model's simplest version, the neurons in a network are divided into groups, and the groups are ordered in a sequence. From each group of neurons, there are excitatory synaptic connections to the neurons in the next group of the sequence. The groups and their synaptic connections are like the links of a chain.

If the first group in the sequence is activated, it sends synaptic drive to the second group, which then becomes active. The second group in turn excites the third group to become active, and so on down the chain. (An alternative mechanism for the propagation of sequential activity in songbird HVC is based on the removal of lateral inhibition between groups of

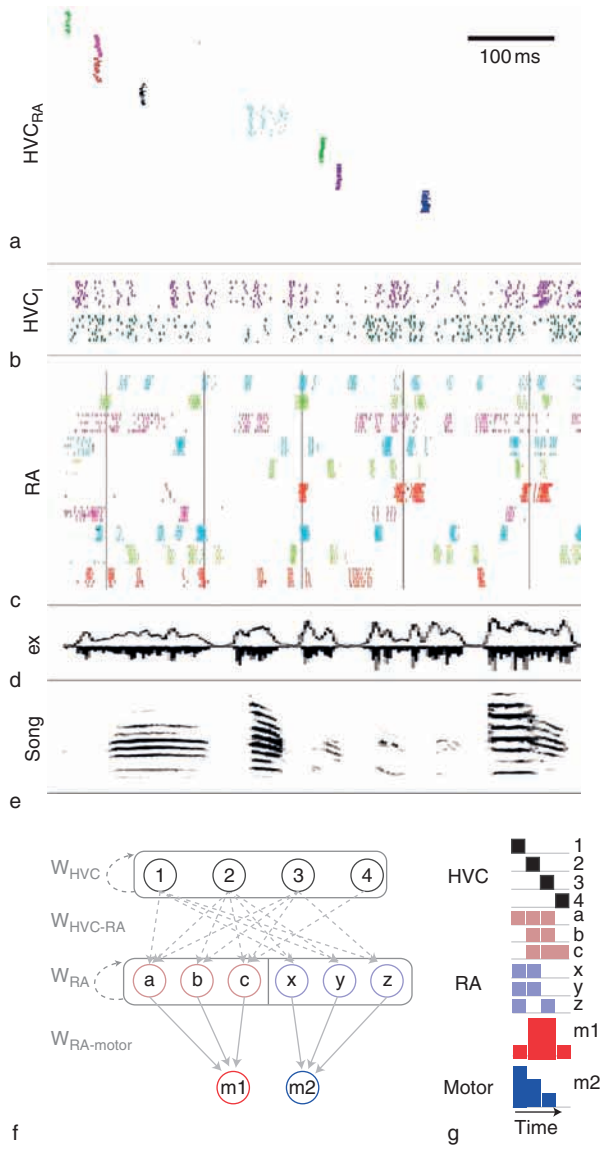

Figure 2 Summary of data and schematic model. (a)–(e) Activity in the premotor pathway. Raster plot of single-unit activity in nucleus robustus archistriatalis (RA)-projecting high-level vocal center nucleus (HVC) neurons (HVC_{RA} neurons; a), in HVC interneurons (HVC_I; b), and in RA (c). Consecutive lines of the same color are multiple recordings of the same neuron. Ticks represent single spikes. (d) Electromyogram of activity in the abdominal expiratory muscle. (e) Spectrogram of resulting song. (a) and (b) adapted from Hahnloser RH, Kozhevnikov AA, and Fee MS (2002) An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* 419(6902): 65–70; (c) from Leonardo A and Fee MS (2005) Ensemble coding of vocal control in birdsong. *Journal of Neuroscience* 25(3): 652–661; (d) and (e) from Suthers RA, Goller F, and Pytte C (1999) The neuromuscular control of birdsong. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 354(1385): 927–939; (f,g) Schematic of loci of song production and plasticity. (f) Flow of activity: HVC drives RA, which drives muscle contraction via motor neurons in a feedforward path. In addition, connectivity within HVC and RA is recurrent. Dashed lines represent weights that are likely to be plastic; solid lines with arrows designate weights assumed to be fixed. HVC, with its recurrent connectivity, produces the high-level sequential drive. Motor neurons, designated by motor neuron pools m1 and m2, sum the activity from RA ((g), middle eight traces) to generate drive for the muscles ((g), bottom two traces). Sequence learning involves plasticity in the recurrent HVC connectivity. In this simple picture, sensorimotor learning is dissociable from sequence learning and involves plasticity in the HVC–RA and perhaps RA–RA connections.

neurons; however, this mechanism does not allow for the self-propagation of sequential activity, instead requiring an external periodic drive (clocked input) to induce transitions between activity states.)

More-complex versions of the synaptic chain model include inhibitory as well as excitatory synapses or different types of synaptic currents with multiple time constants. In still other versions, precisely synchronous spiking of one group is required to activate the next group. These are known as synfire (or 'synchronous firing') chain networks.

The role of intrinsic bursting Two recent models have implemented the synaptic chain idea with excitatory conductance-based neurons to model sequence generation in the zebra finch HVC.

Recordings in the singing zebra finch show that when an RA-projecting HVC neuron is active during song, it fires a high-frequency burst of four to six spikes in a 6–10 ms interval (Figure 2(a)). Previously we discussed the question of the mechanism of sequential activation in HVC. Another basic question concerns the mechanisms that give rise to the bursting behavior. One possibility is that the neurons burst because they are driven by synaptic input that lasts for the duration of the burst. Alternatively, the neurons might burst as a result of the dynamics of their intrinsic conductances.

Li and Greenside have advanced a model that does not use intrinsic bursting. It is consistent with the fact that previous brain slice studies of HVC neurons have not revealed intrinsic bursting. The network model is able to produce sequential neural activity with multispike bursts (Figure 3(a)). However, the realism of this model is compromised by the fact that the peak firing rate within the propagating multispike burst achieved by the model (≈15 ms interspike intervals) is approximately 10 times lower than seen in the actual HVC (≈1.5 ms interspike intervals).

In contrast, Jin et al. have advanced a model in which HVC neurons are intrinsically bursting. The model generates multispike bursts at peak firing rates consistent with those seen in experiments (Figure 3(b)). For comparison, Jin et al. also simulated a synaptic chain model without intrinsic bursting. Unlike Greenside and Li, they were able to achieve higher peak firing

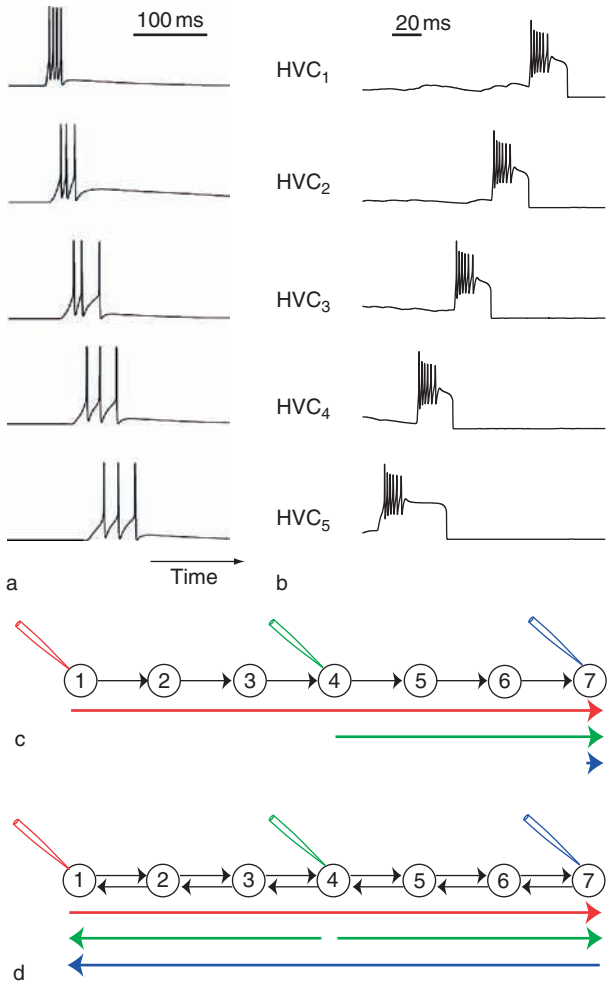

Figure 3 (a,b) Propagation of bursts in synaptic chain models of high-level vocal center nucleus (HVC). (a) Neurons in the model are not intrinsically bursting. Activity propagates due to synaptic chain connectivity, and the burst of spikes in each neuron is due to sustained synaptic drive from the network. Spike frequency during the burst is ≈10 times lower than in nucleus robustus archistriatalis (RA)-projecting HVC (HVCRA) neurons. Adapted from Li M and Greenside H (2006) Stable propagation of a burst through a one-dimensional homogeneous excitatory chain model of songbird nucleus HVC. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 74(1.1): 011918. (b) HVCRA neurons are assumed to possess an intrinsic mechanism for burst generation. A synaptic chain network can robustly produce stable sequences of multispike neural activation. Adapted from Jin DZ, Ramazanoglu F, and Seung HS (2007) Intrinsic bursting enhances the robustness of a neural network model of sequence generation by avian brain area HVC. *Journal of Computational Neuroscience* 23(3): 283–299. (c,d) Unidirectional synaptic chain versus bidirectional connectivity. (c) In a synaptic chain, connections between different groups of neurons project in a predominantly 'forward' direction. Activity initialized at group 1 (red 'electrode') will drive sequential activity up to the end of the chain, at group 7. Stimulation of the last group of the chain (blue 'electrode') will produce no propagating activity or sequence playback. Activation of a group near the middle of the chain will generate playback of a fragment of the original activity sequence. The resulting vocalization should correspond to normal song, but beginning midsong or midsyllable. (d) Bidirectional synaptic connectivity. Activity initialized at group 1 (red 'electrode') will produce a unidirectional sequence of activity propagating from beginning to end, as in the synaptic chain. Stimulation at the end of the chain (blue 'electrode') will lead to sequence playback in the reverse direction, from group 7 to 1. The resulting vocalization should be the song or a syllable sung in reverse. Stimulation of a central group in the chain will lead to activity propagation in both forward and backward directions. The resulting vocalization should be a superposition of the song (or a syllable) sung forward and in reverse, starting in the middle.

rates more consistent with experimental data, but this required careful fine-tuning of parameters like synaptic strengths. In contrast, the model with intrinsic bursting was quite robust to changes in parameters like synaptic strengths.

Superficially, the Jin et al. model may seem inconsistent with previous experiments *in vitro*, in which somatic current injection did not trigger intrinsic bursting in RA-projecting HVC neurons. However, one can imagine two scenarios in which this inconsistency is resolved. First, somatic current injection might be ineffective for revealing intrinsic bursting that is caused by a dendritic calcium spike. This depends on the strength of coupling between the soma and dendrite and on the voltage threshold for a dendritic spike. Second, the intrinsic properties of RA-projecting HVC neurons appear to be different *in vitro* and in the brain of a singing bird. For example, the peak firing rates observed *in vitro* are never as high as those observed during song. Possibly, HVC neurons *in vitro* lack some neuromodulator secreted during song, and this neuromodulator could be necessary for intrinsic bursting. Given the findings of Jin et al. regarding robustness, it seems important to look again for intrinsic bursting of HVC neurons.

Directionality of excitatory connectivity In the synaptic chain models mentioned earlier, there are excitatory synapses from each group to the next group in the chain. Unidirectionality is a fundamental feature. While the synaptic chain model is intuitively plausible, it is important to subject it to experimental testing and to contrast the model's predictions with those of competing models. For example, one can imagine a bidirectional synaptic chain in which there are excitatory synapses between adjacent groups in the chain, but the synapses go in both directions, 'forward' and 'backward' (**Figures 3(c) and 3(d)**).

In both the unidirectional and bidirectional models, it is assumed that song is initiated by activation of the first group in the chain. This causes the entire chain to be activated sequentially in the forward direction (**Figures 3(c) and 3(d)**, red 'electrode'). Both models can reproduce the sequential activity patterns observed in HVC during singing. However,

the bidirectional model can also produce other kinds of activity patterns that the unidirectional model cannot (Figures 3(c) and 3(d)), which could be used as the basis of experimental tests that distinguish between the two models.

In these tests, activity in the HVC network would be initiated by some means, and the subsequent response to stimulation would be observed. If the last group in the chain is stimulated (Figures 3(c) and 3(d), blue 'electrode'), there would be no subsequent activity in the unidirectional model. In the bidirectional model, the activity would propagate backward along the chain and would result in a sequence of motor commands that is reversed from normal song. Another experiment would be to stimulate an intermediate group in the chain (Figures 3(c) and 3(d), green 'electrode'). According to the bidirectional model, subsequent activity would simultaneously propagate both forward and backward along the chain. (Note that the playback will not continue from the forward to the backward direction, or from the backward to the forward direction, because of the existence of a refractory period for activation. This point is well known in the theory of excitable media.) In the unidirectional model, activity would propagate only forward.

While such experiments may sound straightforward, they are actually technically demanding. So far no topographic organization of RA-projecting HVC neurons has been detected. In other words, the location of a neuron in HVC does not seem to be related to the time at which it is activated during a song motif. If coactive neurons were co-located in HVC, it would be relatively straightforward to activate a hypothetical group in the synaptic chain models by stimulation through a microelectrode. Instead, the neurons of a hypothetical group appear to be scattered throughout HVC. Therefore, a group would be best identified through optical imaging of activity, and best stimulated through optical means also.

So far, no one has attempted to precisely stimulate a group of HVC neurons. However, cruder perturbation experiments have been performed. For example, HVC activity during song has been perturbed by stimulating RA, which leads to retrograde effects on HVC. Stimulation leads to skips in motif sequencing, but the subsequent vocalization is a recognizable fragment of the motif sung in the forward direction, beginning at a different position within the motif. This is more consistent with the unidirectional model than with the bidirectional model. But more-conclusive tests will require precisely controlled stimulation of activity in HVC.

A more direct way of testing the synaptic chain models is to check their predictions concerning the synaptic connectivity of HVC. Since the real HVC is no doubt 'messier' than in these idealized models, the predictions should be phrased probabilistically. Suppose A and B are two RA-projecting HVC neurons that are activated in succession. In the unidirectional model, the probability of a connection from A to B is expected to be higher than that of a connection from B to A or of a connection between two neurons chosen at random. (Alternatively, the predicted asymmetry of coupling between A and B might be reflected in synaptic strength rather than in the presence or absence of connections.) In the bidirectional model, the probability of a connection from A to B is equal to the probability of a connection from B to A.

Testing this kind of prediction could be done by comparing physiological and neuroanatomical measurements. Optical imaging could be used to find two neurons A and B that are activated in succession. Then optical stimulation could be used to see whether and how they are connected to each other. Alternatively, a completely neuroanatomical approach would be to reconstruct the entire connectivity of HVC using serial section electron microscopy, look for a chain within this connectivity, and determine its directionality.

The role of inhibition The synaptic chain models above were based on synaptic excitation, without reference to the role of synaptic inhibition. Inhibition is often included in such models to globally 'dampen' activity in a temporally nonspecific way and can enhance the stability of sequential activity in the excitatory neurons. To implement such a global inhibition effect, each inhibitory interneuron receives excitatory input from a large random set of projection neurons, and each projection neuron in return receives synapses from the interneurons. If the burst times of the projection neuron population were distributed across the entire song motif, then each interneuron would receive roughly continuous drive and would spike continuously.

Consistent with this picture, the temporal firing patterns of inhibitory interneurons in HVC are much less specific than those of the HVCRA neurons: an inhibitory interneuron is typically active throughout most of the motif. This suggests that the connectivity of inhibitory neurons is also much less specific than that of excitatory neurons.

Sequence Learning

Hand in glove with the question of what intrinsic neural properties and network architectures could underlie the stable propagation of activity in HVC is the challenge of understanding how such architectures could emerge from initially less-structured networks.

Intuitively, associative (Hebbian) learning rules such as spike-timing-dependent plasticity (STDP) seem ideally suited to produce the kinds of asymmetric connectivity characteristic of synaptic chains: if neuron A fires before B, the connection from A to B is strengthened, but that from B to A is weakened. If a naïve network is trained with this rule for synaptic plasticity on a set of input patterns presented sequentially, one would expect network weights to encode the temporal correlations in the training patterns and later to autonomously reproduce the same long, stable chains of neural activity when stimulated.

But contrary to intuition, past modeling studies show that associative learning rules are not sufficient to configure networks to autonomously generate long neural sequences. Networks trained in this way usually require finely tuned excitatory and inhibitory connectivity that is continuously renormalized to balance network activity during learning, as well as additional pacemaker inputs to propagate activity from one time step to the next. The situation is even worse if the network does not have access to training inputs that are themselves sequential. In such cases, the best response even with fine-tuning of parameters is the formation of extremely short chains (fewer than five steps) using thousands of neurons.

The reason for this is that associative learning rules do not balance activity equitably across space (neurons) and time; instead, they tend to amplify inequities. For example, a neuron that by chance receives strong synapses from its inputs will fire often. Because of its frequent firing, it will tend to rapidly form strong outgoing synaptic connections as well. As a result, activity tends to accumulate around a few overly active neural hubs, producing epileptic or synchronous burstlike network states rather than long, orderly chains. Empirically, these problems remain unsolved with the imposition of local constraints on the maximum strengths of individual synapses. For these reasons, the question of how plasticity rules that are local in both space and time can impose global constraints on activity balance over all space and all time is nontrivial.

Two independent models provide a conceptual framework for how associative learning rules must be augmented to allow initially unstructured networks to self-organize and generate stable, long sequences of neural activation. Both models use STDP as the primary activity-dependent mechanism for synaptic change. But crucially, in addition to STDP, both models rely strongly on competitively enforced synaptic bounds that are not computed on the level of single synapses but are triggered by neuronwide constraints. Although the rules governing synaptic remodeling are not local to the individual synapses, they are local to single neurons. Remarkably, these local neural bounds nevertheless serve to uniformly and globally distribute activity throughout the network in both space and time.

The qualitative nature of these synaptic bounds, and how they are triggered and enforced, is where the two models diverge. The model of Jin and Jun assumes that the formation of a finite number of strong connections from a neuron triggers axon remodeling and elimination of weak synapses. Fiete et al. propose two separate schemes based on neural resource constraints: one assumes that the production rate of synapse-building resources is limited, so potentiation in one synapse triggers a uniform heterosynaptic weakening of all other synapses at that neuron. The other scheme assumes a cap on the total resources a neuron can devote to synaptic maintenance; therefore all synapses at a given neuron undergo uniform weakening whenever STDP drives their summed synaptic strength to exceed the cap.

Both models make predictions for the rules governing synaptic plasticity within HVC and for the statistics of the resulting connectivity within HVC. The Fiete et al. model produces several sequences of different lengths, with a definite distribution of chain lengths. Determining whether this distribution matches syllable length distributions in birdsong could indicate whether songbird syllables attain their specific lengths largely due to bottom-up (low-level network) properties or top-down (behavioral, evolutionary, or respiratory) constraints.

Sensorimotor Learning of the Motor Map

The main goal-directed aspect of learning in zebra finches takes place during the sensorimotor phase, when the juvenile learns to match, through auditory feedback and template comparison, its own string of vocalizations with its memorized mental copy of the tutor song. Because the goal-directed aspect of song acquisition is so striking and because natural goal-directed learning behaviors that are easy to quantify are rare, several models have focused on this aspect of song learning. Yet the microscopic synaptic plasticity rules that underlie such learning remain experimentally undiscovered and theoretically debated. Predictions from existing theoretical or computational models can serve to greatly narrow, in a principled way, the dauntingly large number of possible plasticity paradigms to investigate in experiment, and they are now specific enough to be tested in detail in slice preparations of the song system. (Experimental testing of the proposed paradigms is contingent on

the identification of a reinforcement or critic signal, which is widely expected to exist but has not yet been found.)

Existing models of goal-directed sensorimotor song learning share the essential schema of the cartoon model (Figure 2) for the roles assigned to the different premotor nuclei: HVC produces a high-level sequence, and sensorimotor learning involves mapping that abstract sequence into the correct vocal motor sequence through plasticity in the HVC–RA synapses. The primary differences between models lie in the roles assigned to the AFP (Figure 4), in the particular rules for how activity triggers plasticity in HVC–RA synapses, and in their levels of biophysical realism and detail. The plausibility of different assumed roles for the AFP can already begin to be evaluated with existing electrophysiological and lesion results.

Early experiments showed that lesioning the terminal lateral magnocellular nucleus of the anterior nidopallium (LMAN) of the AFP, which projects back into the premotor pathway, abolishes the ability of juvenile songbirds to learn song. Moreover, the immediate and lingering effect of the LMAN lesion, besides stoppage of song learning, is to dramatically reduce the large trial-to-trial variability of juvenile song. These results suggested that the role of LMAN in song learning is possibly to drive vocal experimentation through perturbation of the premotor song pathway.

An early, and in several respects prescient, computational model is based in part on these observations. The model applied the specific algorithm of weight perturbation (with momentum) from the reinforcement learning literature to suggest that LMAN directly generates small static weight changes in the HVC–RA connections (Figure 4(a)). According to the model, these LMAN-driven static weight changes last for the duration of one song rendition, after which they are selectively retained, based on correlation with a song evaluation (reinforcement or critic) signal. The reinforcement (critic) signal is a function of the match between the actual and desired (tutor) songs. Because the weight perturbation rule has been proven by others to perform gradient climbing on the scalar reinforcement signal, it will improve song performance and do so independent of most parameter values. Numerical simulation of a small network model of the birdsong network, consisting of rate-based units or neural assemblies with time 'coarse-grained' into syllable-long chunks, has demonstrated how this may happen. Experimental studies have supported several features of the model, including the high-level schema of actor–critic reinforcement learning (Figure 5), the role of perturbations in song learning, and the role of the AFP in driving perturbations.

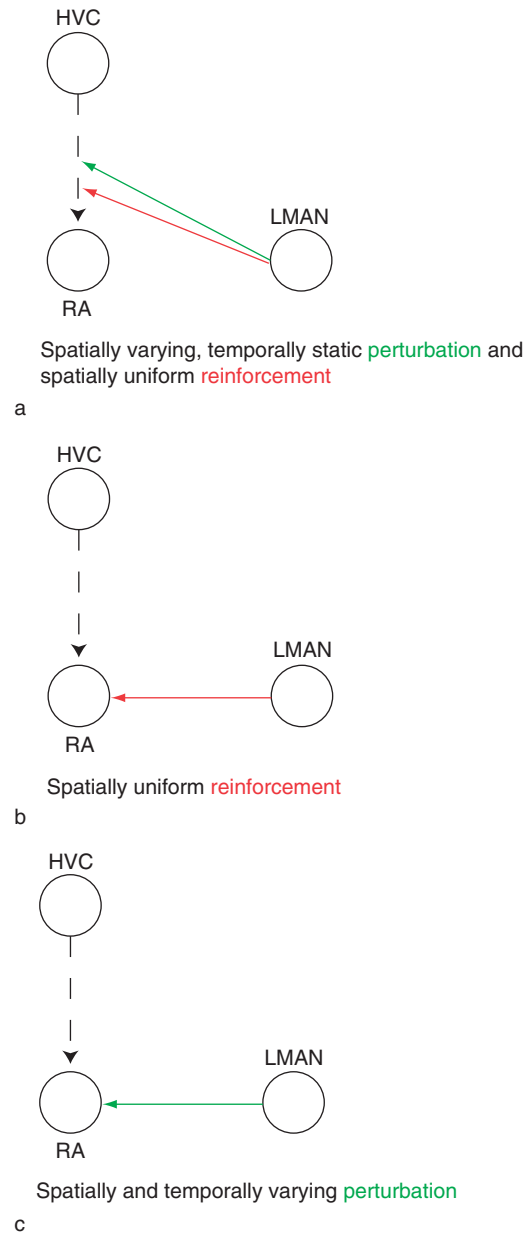

Figure 4 Proposed roles for lateral magnocellular nucleus of the anterior nidopallium (LMAN) in sensorimotor learning. The dashed line indicates a plastic synapse between high-level vocal center nucleus (HVC) and nucleus robustus archistriatalis (RA) neurons. (a) The hypothesis of Doya and Sejnowski: LMAN must produce different, independent perturbations in each HVC–RA weight originating from different HVC neurons. These perturbations must be static over the duration of one song rendition but vary across renditions. LMAN must also broadcast a separate, spatially uniform reinforcement signal to the premotor pathway. (b) The hypothesis of Troyer and Doupe. LMAN's sole role is to provide a spatially uniform reinforcement signal to the premotor pathway. Learning is associative (Hebbian). (c) The hypothesis of Fiete and colleagues: LMAN must inject different perturbations to different RA neurons. The perturbations are excitatory synaptic inputs to RA and vary in time over one song rendition. (A spatially uniform reinforcement signal is assumed to arrive at RA from elsewhere in the network.)

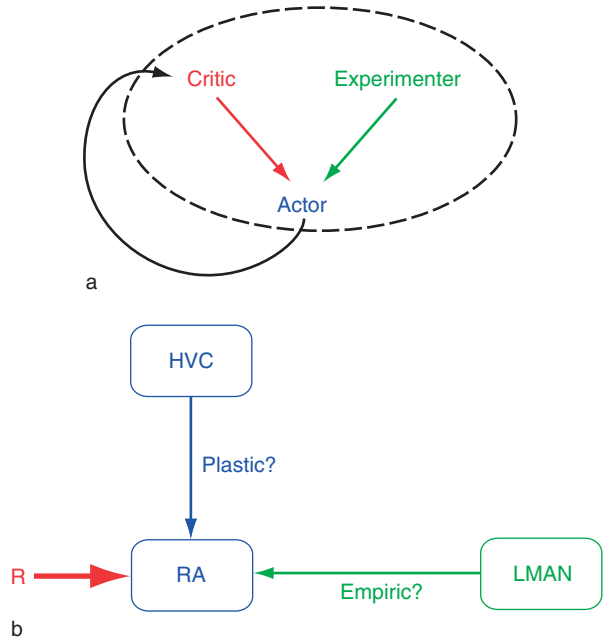

Figure 5 The empiric synapse hypothesis. (a) The empiric synapse hypothesis for goal-directed learning consists of three parts: The actor network, which is responsible for performing the desired task and has plastic synapses, and two of its inputs, one an experimenter, whose role is to drive perturbations in the actor through empiric synaptic inputs, and the other a critic, which provides through a reinforcement signal a scalar assessment of the performance of the actor on the desired task. Actor neurons are hypothesized to correlate experimenter-driven perturbations with the reinforcement signal to derive a signal for synaptic change, in the direction of the gradient of the reinforcement signal. (b) In the song system, the actor network is hypothesized to be the premotor pathway including high-level vocal center nucleus (HVC) and nucleus robustus archistriatalis (RA) and plastic HVC–RA synapses. The experimenter is hypothesized to be the AFP. A critic-generated reinforcement signal (R) is widely assumed to exist in the song system, although its neural identity is unknown. Adapted from Fiete IR, Fee MS, and Seung HS (2007) Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *Journal of Neurophysiology* 98: 2038–2057.

What are the biophysical assumptions and implications of this model? According to the model, both the song evaluation (reinforcement) signal and the perturbative drive are generated in the AFP, and delivered to the premotor pathway by LMAN. This means that LMAN must carry and convey two independent signals, for perturbation and reinforcement, to the HVC–RA synapses (Figure 4(a)).

The model is based on batch mode weight perturbation, and it assumes that LMAN can somehow directly perturb the weights of the HVC–RA synapses in a way that the perturbations vary from song rendition to rendition but are static over the course of a rendition. Moreover, for the weight perturbation scheme to explore enough of the motor space for song learning, HVC–RA synapses that originate from different HVC neurons must receive different, and independent, perturbations from LMAN. There are a few experimental inconsistencies with these requirements in the birdsong system. First, single- and multiunit activity in LMAN varies rapidly during a song motif, without an unusually large transient at the beginning. It is therefore difficult to imagine how LMAN neurons could provoke an instantaneous perturbation in the HVC–RA weights at the beginning of a song rendition, which is held constant for the rest of the rendition. (It may be possible for weight perturbations to be driven by slow neuromodulatory action of LMAN onto the HVC–RA synapses; within the model, reinforcement would then have to be mediated either by ordinary synaptic transmission from LMAN to RA or through a second neuromodulator released by LMAN neurons into RA.) Second, each RA neuron receives an estimated 1000 inputs from up to 200 HVC neurons but only approximately 20–50 inputs from LMAN. Thus, it is unlikely that synapses from different HVC neurons could receive independent perturbations to fulfill the requirements for weight perturbation.

Finally, Doya and Sejnowski found that learning with purely random inputs from LMAN was too slow, with poor convergence to the tutor song, even though the simulated network was far smaller than the songbird motor pathway and even though a song motif consisted of very few time steps (each time step represented an entire syllable). Therefore, they assumed that LMAN activity itself evolves by reinforcement learning on the same global reinforcement signal, so that its output is composed of a term proportional to the true HVC–RA weight gradient, superimposed on a random component. With this assumption, the AFP must perform at least three functions: (1) generate and carry random perturbation signals whose dimension might be as large as the number of HVCRA neurons, (2) generate and carry a true gradient estimate after correlating its random activity with reinforcement, and (3) function as a critic, generating and carrying the reinforcement signal. With these assumptions, the model indeed learns; however, the AFP is assumed to possess these capabilities *de facto*, without modeling how the necessary computations are performed or how the resulting signals are carried by LMAN and delivered to the premotor pathway, leaving crucial questions unresolved. Whether the AFP actually possesses these capabilities is an experimentally open question.

A model proposed by Troyer and Doupe in 2000 suggests that plasticity in HVC–RA synapses and RA–RA synapses may be largely associational or Hebbian in nature. According to the model, plasticity is gated by a reinforcement signal, which LMAN

delivers to the premotor network (Figure 4(b)). The associational rule also includes a subtraction term for synaptic weakening, bringing it closer to reinforcement learning rules, but the specific form used is not proven in the literature to guarantee improved performance. As a result, there are no general assurances of convergence or stability, even in networks of simplified rate-based neurons. The model requires fine-tuning by hand of several homeostatic mechanisms to maintain stability during learning. The authors demonstrated through numerical simulation in rate-based networks that under particular parameter settings, with a small set of RA assemblies, HVC–RA connections, and a few syllable-long time steps, the network can learn to match a desired output.

A strength of this model is that, unlike others in the literature, it stresses the importance of lateral connections within RA, which, when subject to the same plasticity rules as HVC–RA synapses, form assemblies that are capable of pattern completion even with partially correct inputs, to generate 'good' output syllables. The AFP's only role in the model is to generate and convey a spatially uniform reinforcement signal to RA (Figure 4(b)). However, in the zebra finch, LMAN activity is spatially inhomogeneous (neural activity is not strongly correlated across the nucleus), and connections to RA are not highly divergent, which implies that different RA neurons are likely to receive fairly uncorrelated inputs from LMAN rather than a uniform signal. As formulated, the model cannot explain why, in experiments, LMAN lesions reduce song variability. Experiments fail to reveal qualitative changes in LMAN activity when an adult bird's auditory feedback is altered, which suggests that LMAN may not primarily convey an error or match signal. (If LMAN does generate an error signal, but bases it on learned predictions about auditory feedback, it is possible that prolonged rather than transient exposure to distorted auditory feedback is necessary to produce changes in LMAN activity.)

Subsequent informal proposals, not grounded in numerical simulation or modeling, have espoused more information-rich primary roles for LMAN in song learning. These include suggestions that LMAN may provide detailed instructive or supervisory error signals to RA. Unfortunately, these suggestions lack mechanistic proposals or models of how the AFP might correctly compute and generate these information-rich, time-varying supervisory signals, leaving the bulk of the computational burden of song learning unsolved and at the door of the AFP. In reality it is possible that the AFP supplies such a supervisory signal to RA, but no experimental or computational studies exist yet to suggest that it does.

A recent model of song learning by Fiete and collaborators returns to a reinforcement learning framework and attempts to do two things: (1) propose that, and illustrate how, spiking RA neurons might perform reinforcement learning by using rapidly time-varying LMAN synaptic inputs that perturb RA membrane voltages (rather than using static synaptic perturbations of the plastic HVC–RA weights; Figure 4(c)) and (2) show as a proof-of-principle that although it is widely thought to be slow, reinforcement learning can be fast enough to explain song acquisition in a full-scale model of the birdsong network with spiking, conductance-based neurons, delayed feedback, scalar reinforcement, and purely random independent perturbations.

The model assumes that LMAN perturbs the premotor pathway by injecting independent random currents into different RA neurons (Figure 4(c)) through ordinary glutamatergic synaptic transmission (Figures 5, 6(d), and (e)). This is consistent with recent stimulation and pharmacological inactivation experiments in LMAN, which show that LMAN helps drive song variability through glutamatergic synaptic projections to RA. The model requires at most as many independent LMAN inputs as there are RA neurons. The LMAN inputs are assumed to be time varying throughout song. Reinforcement is assumed to be computed and delivered from elsewhere in the brain, with learning following the high-level, actor–critic schema. Under these conditions, the derivation of the rules for how activity should modify synaptic weights, and the rules themselves, are quite different from weight–perturbation-like algorithms (Figures 6(d)–6(f)): each HVC–RA synapse must somehow deduce an instruction for change based on time-varying fluctuations in the conductance of its postsynaptic RA neuron rather than simply decide whether to retain or discard direct static perturbations of its own weight through correlation with the reinforcement signal. The resulting rule is guaranteed to robustly (without parameter tuning) improve song performance, even in recurrent networks of conductance-based spiking neurons and with delayed reinforcement.

The model is able to produce learning in a biologically realistic network model of the song system, with spiking neurons and an impoverished, delayed reinforcement signal (Figures 6(a) and 6(b)). Learning takes place online, and activity in the network evolves on the millisecond timescale, driven by 5–6 ms bursts in the HVC inputs. Within 2000 iterations, the network output resembles the tutor song, which is a recording of an actual zebra finch song (Figure 6(c)). Fiete et al. have shown that learning time even with random LMAN input should scale well in the full-sized premotor network and may converge in far

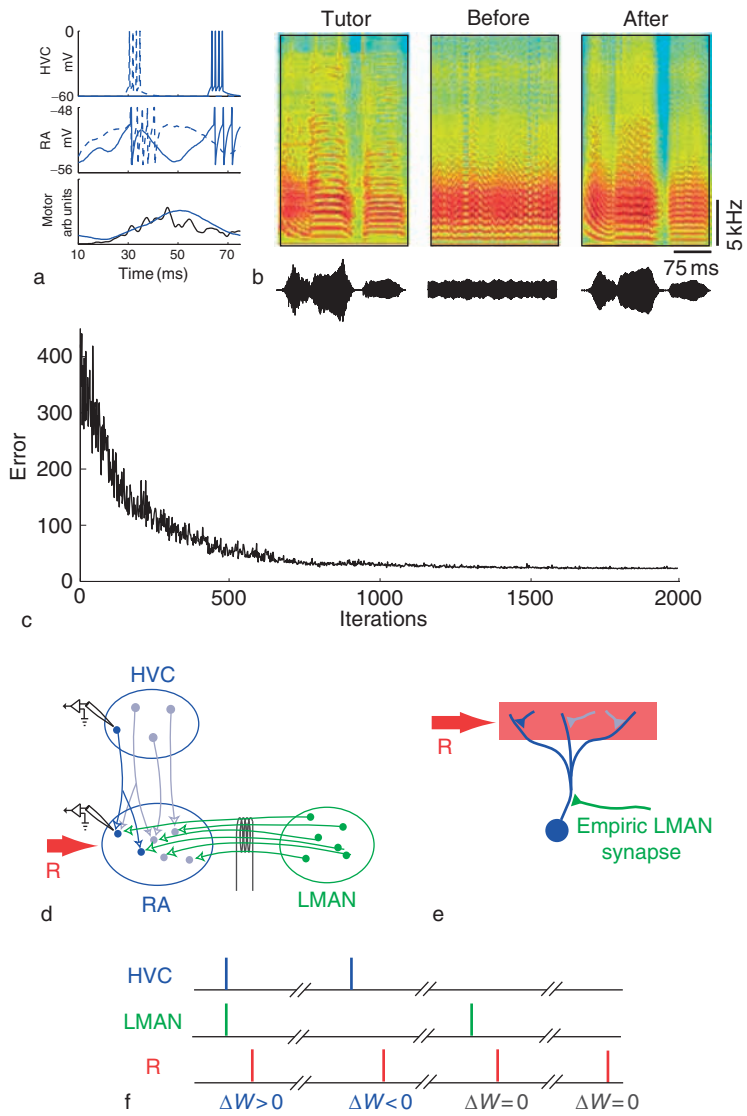

Figure 6 Empiric synapses: results and predictions. (a) Activity of two model high-level vocal center nucleus (HVC) neurons (top), two representative nucleus robustus archistriatalis (RA) neurons (middle), and one of the motor output pools (bottom; black: tutor, blue: network response) during song learning. (b) Spectrograms of the tutor song and the output of the model song network before and after 1200 iterations of learning, together with the respective sound pressure waves. (c) Mismatch between the network output and the tutor song decreases relatively rapidly. Learning has reached an approximate asymptote by 1000 iterations. (d,f) Prediction for bidirectional synaptic plasticity from the empiric synapse model. (d) Possible experimental design for testing the empiric synapse rule in the birdsong pathway involves focused stimulation in HVC, intracellular recording in RA, and fiber bundle stimulation of the lateral magnocellular nucleus of the anterior nidopallium (LMAN)–RA axons, together with control of the reinforcement signal (R). (e) Magnified schematic of a single RA neuron and its inputs. Each RA neuron receives a few empiric synapses from LMAN and several from different HVC neurons; in addition, RA receives a reinforcement signal. Blue (gray): active (quiescent) premotor inputs; green: empiric input; red: reinforcement signal. (f) Protocol for induction of bidirectional synaptic plasticity in active HVC–RA synapses: Simultaneous stimulation of an HVC neuron and the LMAN axon bundle, followed by administration of reinforcement, should lead to potentiation of all stimulated HVC synapses.

fewer iterations than sung by the typical zebra finch (≈100 000 renditions) during song acquisition.

The model's central prediction is a specific paradigm for the induction of bidirectional, heterosynaptic activity-dependent synaptic plasticity in the premotor vocal pathway (Figure 6(f)) that should be possible to test in slice experiments.

The model above is consistent with a large body of experimental data in the song system and provides concrete predictions for synaptic plasticity. However, LMAN's role in the model is to provide purely random exploration. This assumption does not contradict the data but is a limited view of what the AFP, a complex neural pathway, might be capable

of. More-sophisticated strategies for AFP-delivered perturbation may speed up learning. This avenue is wide open for discovery by theory and experiment.

More generally, models of song learning based on the AFP as a source of random or directed perturbations to the premotor system could be used as comparative tools for studying the role of the basal ganglia thalamocortical loop in mammalian sensorimotor learning.

Premotor Codes and Representations

Up to 30 000 neurons are devoted to song production and learning in the zebra finch. Together, these neurons drive only eight to ten muscles that produce song. What are some of the functional requirements and costs that determine why so many neurons are required for song production?

The ultrasparse or unary coding of song by HVCRA neurons (Figures 2(a)) exacts a large energetic and space cost in terms of the number of neurons required to encode a given set of patterns. To produce a sequence of 100 steps (L) with 200 neurons per time, unary coding requires 20 000 neurons. At the other extreme, a combinatorial code with a downstream readout capable of distinguishing between different combinatorial patterns would require far fewer neurons, scaling as log (L) instead of L; intermediate possibilities would require intermediate numbers of neurons. What are some of the properties of a unary code that might offset its capacity costs and explain why it exists in HVC?

A possible explanation for ultrasparse HVC sequences could be mechanistic: recurrent networks cannot store many dense patterns because of increasing interference between patterns. Studies show that sparsely active attractor networks can store more patterns. Nevertheless, in these studies, 'sparse' is far denser than unary, so the reason for ultrasparse coding must lie elsewhere.

Because oscine birdsong is learned, a natural place to investigate the role of unary HVC codes is in learning. There are actually two distinct roles that unary codes could play in song learning.

First, an analytical and numerical study by Fiete et al. shows that unary coding within HVC enables optimally fast learning of the feedforward motor map from abstract sequences in HVC to sequences of muscle activation. If each HVC neuron were to fire twice instead of once, at random times in a song motif of fixed length, the fastest possible speed with which the input–output map can be learned is twice as slow (Figures 7(a)–7(d)). There is an intuitive explanation of the underlying mathematical reason for this effect. Suppose an HVC neuron and its corresponding HVC–RA synapses are active twice, when the tutor song has opposing characteristics (e.g., pitch or amplitude), at those two times. The synapses will be pulled in opposing directions to best match the tutor song (Figure 7(c)) at each time. Because the learning task is realizable, there is some value of synaptic strength that achieves the desired output, and in the end this value will be reached. But the conflicting demands on synapse strength at the two times in the motif make approach to the solution with a hill-climbing rule slower, even if the parameter controlling learning step size is optimized to produce the fastest possible learning without network instability.

Yet answers to why questions, however compelling, risk being just-so stories unless they generate specific testable predictions. The work described above argued that unary codes may exist in HVC because of their utility for fast song learning. However, single-unit HVCRA activity and unary codes have been observed only in adults. Thus, the central prediction – which if false would invalidate the hypothesis – is that activity in HVCRA neurons, even if nonsparse initially (early in the sensory period), should be unary near the onset of sensorimotor learning in juvenile zebra finches instead of emerging gradually after or toward the end of sensorimotor learning.

The analysis leads to an additional prediction. To accommodate a larger song repertoire, HVC could change in two *a priori* equally likely ways: (1) keep HVC volume (number of neurons) fixed but increase the number of bursts per neuron or (2) increase HVC volume but maintain a unary code. Energetics might favor the first possibility. But according to the learning speed analysis, if zebra finches are under pressure to learn their songs quickly, then unary coding should be conserved, even at the cost of a larger HVC. This suggests that coding in HVC across zebra finches and related songbirds with somewhat larger repertoires (measured by number of unique syllables) should remain unary while HVC size scales with repertoire size. Indeed, existing experimental findings corroborate these predictions: HVC size scales with repertoire size after correcting for overall brain size, while RA, LMAN, or other song nuclei do not display similarly consistent relationships.

Second, analysis and simulation by Fiete et al. has shown that if the HVC code is unary and if, during a song rendition, performance is continuously evaluated and a reinforcement signal continuously delivered (albeit delayed), then the time taken to learn a song is independent of the length of the learned song. Because the HVC code is unary in zebra finches, this prediction can be used in experiments (where learning time is monitored for birds trained on tutor songs of

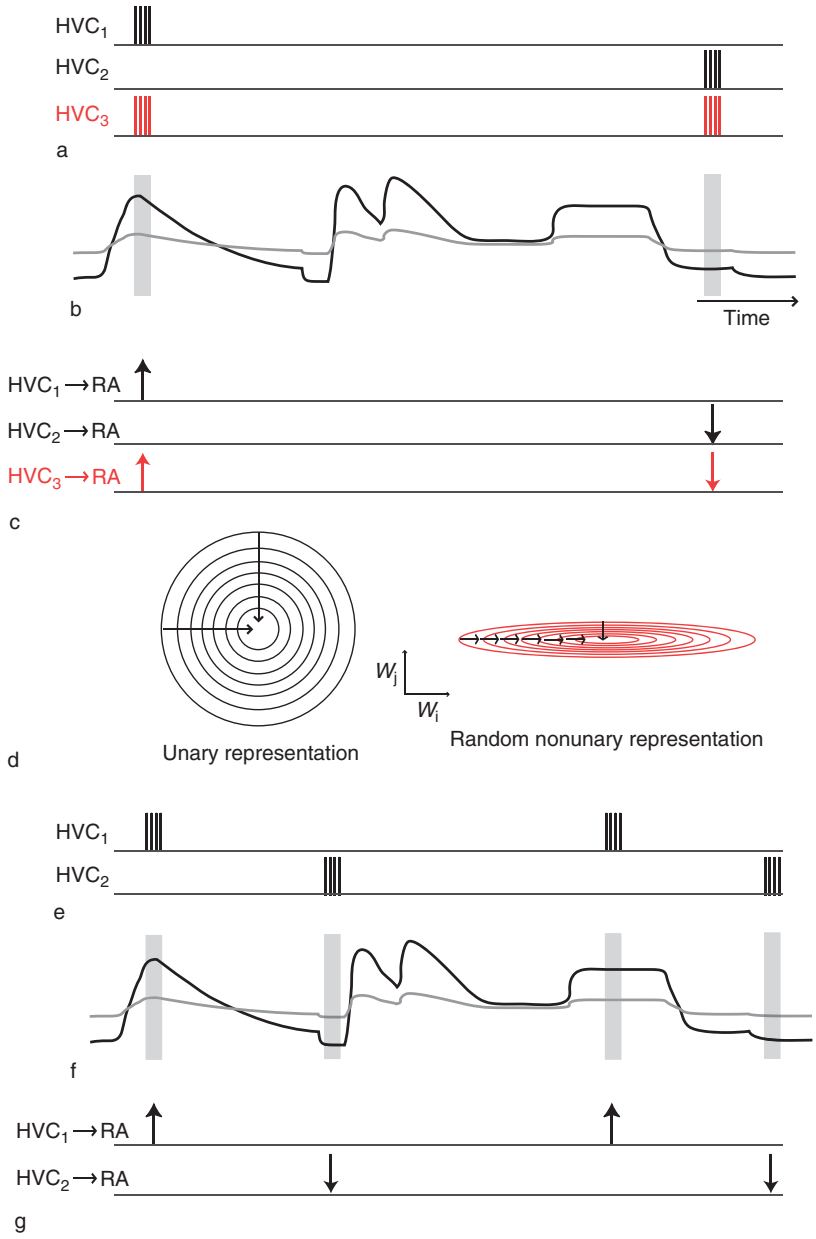

Figure 7 Unary coding in high-level vocal center nucleus (HVC) helps speed the acquisition of the motor map. (a) The activity of three hypothetical HVC neurons. The first two are each active only once in the song motif, but the third is active twice. (b) Hypothetical pitch of the tutor song (black) and the pupil network (gray). (c) Direction in which synapses should change to reduce error between tutor and pupil pitch. Synapses from HVC neuron 3 should be strengthened at the first and weakened at the second activity burst to improve the tutor–pupil match. Such conflicting demands on the synapses of neurons that are active at two or more random times in a motif cause a slowdown in learning speed. (d) Contours of iso-error in the learning surface. Learning a feedforward map in a network with unary coding in the top layer is like learning on an isotropic cost surface and can be fast. Denser coding in the input layer produces correlations and makes the learning surface anisotropic. To keep the error from diverging along the steep directions, the learning rate must be kept low. As a result, best-case learning is slower than in the isotropic case. HVC activity tuned to acoustic features may be helpful for generalizable learning. (e) If HVC neurons fired multiple bursts at selected points when the acoustic features of tutor song are similar, rather than at random times, there would be no interference in the learning update, (f). If HVC neurons acquired such a tuning to features in the tutor song, it could be easy for the bird to quickly reproduce a specific heard sound by activating the requisite sound-tuned HVC neuron. RA, nucleus robustus archistriatalis.

different lengths) as an indirect probe of whether the elusive reinforcement signal in the song pathway is delivered online or in batch mode. If the reinforcement signal is evaluated and delivered only once, at the end of song, then the time taken to learn a tutor song should scale linearly with its length. Similarly, if the HVC code at each time consisted of a random pattern of activations, with half of all neurons active per time

(dense code), learning time should scale with song length (even if reinforcement is delivered online).

Do other songbirds also display unary HVC representations of song? At present, single-unit recordings of RA-projecting HVC neurons are available only from zebra finches. Learning with a pure time code is nongeneralizable: like one piece of magnetic tape, each pattern is used to produce song at only one specific time. An entirely different set of HVC neurons and HVC–RA weights must be used and trained to produce the same sound if it recurs elsewhere in the motif. Therefore, songbirds that rapidly acquire and imitate new songs (albeit after slowly acquiring a first song) may use a different encoding strategy (e.g., tuning codes based on acoustic features) from a time-code (Figures 7(e)–7(g)). In addition, songbirds with vastly larger repertoires, regardless of the speed of song acquisition, may display different HVC codes because of capacity constraints.

Conclusions

Advances in the acquisition of behavioral and anatomical data and neurophysiological recordings from awake, singing birds are rapidly turning the birdsong circuit into an ideal system for unraveling the mechanisms underlying motor control and learning. Theoretical models, by quantitatively validating or eliminating candidate explanations of network function, have helped lay a groundwork for understanding the dynamical and coding principles that lead to such functionality. Theoretical models also provide a functionally motivated set of predictions about neural activity, connectivity, coding, and plasticity that narrow the choices of future experiments and are presently ripe for experimental testing. Future experiments and modeling in songbird species with more-flexible learning behaviors should greatly enhance understanding of generalizable motor learning and control.

See also: Bird Song Systems: Evolution; Birdsong Learning: Evolutionary, Behavioral, and Hormonal Issues; Birdsong: The Neurobiology of Avian Vocal Learning; Hebbian Plasticity; Learning, Action, Inference and Neuromodulation; Motor Sequences; Signal Production and Amplification in Birds; Spike-Timing-Dependent Plasticity Models; Spiking Neuron Models; Synaptic Mechanisms of Learning; Vocal Communication in Birds.

Further Reading

Brainard MS and Doupe AJ (2000) Interruption of a basal ganglia–forebrain circuit prevents plasticity of learned vocalizations. *Nature* 404(6779): 762–766.

Doya K and Sejnowski TH (1995) A novel reinforcement model of birdsong vocalization learning. In: Tesauro G, Touretzky DS, and Leen TK (eds.) *Advances in Neural Information Processing Systems*, pp. 101–108. Cambridge, MA: MIT Press.

Fiete IR, Burger L, Senn W, and Hahnloser RHR (2007) Enforcing global coding constraints by synaptic competition – a model for the formation of long ultrasparse sequences in the songbird. Submitted, 2007. (See also Society for Neuroscience Abstracts, 2005.)

Fiete IR, Fee MS, and Seung HS (2007) Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *Journal of Neurophysiology* 98: 2038–2057.

Fiete IR, Hahnloser RH, Fee MS, and Seung HS (2004) Temporal sparseness of the premotor drive is important for rapid learning in a neural network model of birdsong. *Journal of Neurophysiology* 92(4): 2274–2282.

Jin DZ, Ramazanoglu F, and Seung HS (2007) Intrinsic bursting enhances the robustness of a neural network model of sequence generation by avian brain area HVC. *Journal of Computational Neuroscience* 23(3): 283–299.

Jun JK and Jin DZ (2007) Development of neural circuitry for precise temporal sequences through spontaneous activity, axon remodeling, and synaptic plasticity. *PLoS ONE* 2(8): e723.

Li M and Greenside H (2006) Stable propagation of a burst through a one-dimensional homogeneous excitatory chain model of songbird nucleus HVC. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 74(1.1): 011918.

Troyer TW and Doupe AJ (2000) An associational model of birdsong sensorimotor learning I. Efference copy and the learning of song syllables. *Journal of Neurophysiology* 84(3): 1204–1223.

Birdsong Learning: Evolutionary, Behavioral, and Hormonal Issues

G F Ball, Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Birdsong: Definitions and Functions

In animal communication, a signal is classified as a 'song' generally on the basis of its perceived function. The vocalization that is used in a reproductive context to attract a mate and/or defend a territory can be thought of as the individual's song while vocalizations that function to signal the presence of predators or food, to maintain group cohesion, or to allow parents to communicate with their young are called 'calls.' The definition of birdsong is further complicated because songbirds are a taxonomic unit (i.e., the order Passeriformes, or perching birds). Thus birdsongs, strictly defined, are the species-typical vocalizations that serve the function of mate attraction or territorial defense and that are produced by avian species in this order. These vocalizations are usually quite elaborate and melodious, which is why members of the order Passeriformes were given the name 'songbird.' However, the songbird order is quite large, consisting of nearly 50% of the more than 9000 living species of birds. There is, not surprisingly, substantial variation within this group, so for a particular species, the song produced as defined by its function is not necessarily the most complex vocalization in its repertoire.

Birdsong is of interest to neuroscientists because it is a complex motor behavior that is learned via a developmental process that involves the precise guiding of motor patterns by auditory feedback. The neural circuit for this behavior has been described in some detail, so it provides neuroscientists with an excellent opportunity to attack a number of important questions relating to learning and motor production. Furthermore, song is one of a suite of traits that is needed for successful reproduction, so not surprisingly, the production, and to some extent the learning, of song is profoundly regulated by gonadal sex steroid hormones as well as, in some contexts, by adrenal glucocorticoids. This article summarizes behavioral, comparative, and hormonal aspects of birdsong learning that places neural and neuroendocrine studies of birdsong in a broader perspective.

Birdsong Learning: Behavioral and Comparative Perspectives

Experience and some sort of learning effects on vocal production are widespread among vertebrate species. What makes birdsong learning so unusual compared with vocalizations produced by many other taxa is how profoundly important auditory feedback is in order for songbirds to produce songs that are anything like the stereotypic species-typical vocalizations one hears in the field. Pioneering studies by William Thorpe at Cambridge University and Peter Marler at the University of California Berkeley and Rockefeller University of north temperate zone songbirds such as the chaffinch, the white-crowned sparrow, the song sparrow, and the swamp sparrow clearly established the basic phenomenon. Thorpe and Marler employed an experimental approach in which they raised birds in acoustic isolation from shortly after hatching and documented serious deficits in the later song that was produced when the birds reached sexual maturity. They were able to rescue these behavioral deficits if they provided the nestlings with conspecific vocal experience. They then observed how song develops through a gradual process by which the young bird matches its vocal output with an auditory memory it previously formed for conspecific song based on experiences with conspecifics. Typical of these early studies was the approach taken by Marler in his studies of white-crowned, song, and swamp sparrows, in which he carefully manipulated the types of songs that the birds heard during ontogeny. He was able to establish some basic principles based on this work. These will be summarized here. However, it is important to note at the outset that subsequent work by many investigators has shown that there are important variations on these principles.

The Behavioral Development of Song

The gradual process of song development is itself a type of evidence that song must be learned via auditory feedback. A young male songbird typically begins singing sometime after learned songs have been memorized, but the imitations are not fully formed. Instead, the young male starts with subsong, an amorphous, noisy twittering that changes gradually, first into plastic song, which also is highly variable but contains the first obvious signs of mature song structure, and finally crystallizing into the stable patterns of mature, adult song. Subsong resembles the babbling of human infants and may be important for developing the motor skills of singing and other prerequisites for song learning, such as the ability to guide the voice by the ear. Rehearsal of previously memorized song patterns begins in plastic song. Often more plastic song themes are produced than are needed for mature singing, and many are discarded when song crystallization occurs. At song crystallization, a stereotypic form of the species-typical song is produced.

The Song-Learning Process

Sensitive Periods for Song Learning; The Effects of Isolation and Deafness; Learning Preferences

There is an underlying pattern in the steps typically required in learning to sing. First is the acquisition phase, when a bird hears songs and commits some of them to memory. These are stored for a period that varies markedly from days to months among species until the bird begins to recall songs from memory and starts to produce imitations. There is thus a separation in time of the acquisition or sensory phase and the production or sensorimotor phase, ending with the production of crystallized (i.e., stereotyped), species-typical adult song.

There are often sensitive periods for song acquisition, sometimes restricted to a short period early in life and sometimes extending into adulthood. Even close relatives, such as sparrows and canaries, may differ in this respect. Several species of sparrows have a sensitive period for acquiring songs, beginning at about 20 days of age, soon after young males become independent from their parents, and ending 4 to 6 weeks later. These sensitive periods for learning are another aspect of the process that exhibits significant species variation. Some birds will keep learning throughout their lives whereas others seem to learn only during a specified time in their first year. However, proving such limits definitively can be quite difficult. Also, the boundaries of sensitive periods can vary depending on the strength of song stimulation and the influence of physiological factors, such as hormonal states, that vary with the season. If young are hatched late in the season and singing has already ceased for that year, closure of the sensitive period may be delayed until the following spring. The experimental withholding of stimulation by songs of the birds' own species can also delay closure of the learning. Species that learn song only during a sensitive period early in ontogeny are known as age-limited learners. There are also species that continue to learn new songs throughout their lives, and these are referred to as age-independent learners. Still other species fall between these extreme cases in terms of song development.

Regardless of whether they have had the chance to learn, songbirds can generally produce some aspects of the song that can be recognized as being species typical. When sparrows are raised in isolation, for example, the note structure and tonal quality of their songs are abnormal, but each species still produces some basic features of normal song organization. Studies of deaf birds' songs, which are highly abnormal and variable in structure, have been valuable in explicating the importance of auditory feedback and its timing (Figure 1). A degraded form of singing is observed both if a male becomes deaf before he experiences song and if he is deafened after hearing song but before the development of singing (Figure 1). To transform a memorized song into a produced

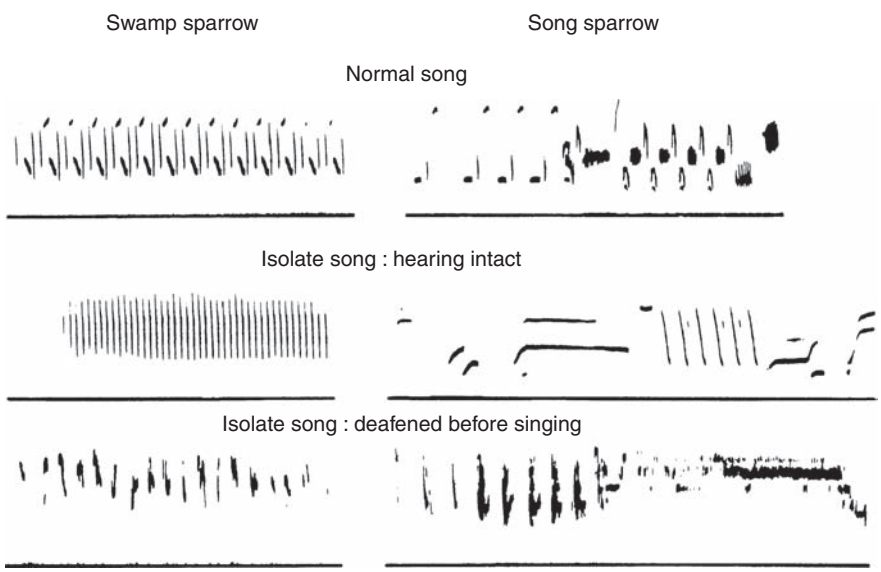

Figure 1 Sonagram illustrating the songs of two closely related songbird species, song sparrows and swamp sparrows. The top row illustrates the normal species-typical song produced by adults. The second line illustrates song produced by males that had been deafened after the sensory phase of song learning but before the onset of the sensorimotor phase. The third line illustrates sonagrams of songs produced by males that had been deafened and raised in social isolation. Note that in birds that develop song in the absence of auditory feedback, there is still evidence of species-typical structure. In birds raised in complete auditory isolation, the degree of song degradation is profound. From Marler P and Slabbekoorn H (eds.) (2004) *Nature's Music: The Science of Birdsong.* Amsterdam: Elsevier.

song, the bird must be able to hear its own voice. One can infer that there are auditory memories for song, involved in guiding song production, that must vary in their specifications from species to species.

One aspect of the process of memory formation needed for song learning is that it involves selective or biased learning. In several sparrow species, for example, it has been shown that species-typical vocalizations appear to be privileged in that they are learned preferentially. Another aspect of this memory that has been strongly advocated by Marler and colleagues is that the memory encodes species-typical patterns of vocal behavior even before it hears the song of its own species. This idea of an innate specification of species-typical vocalizations would be one way to explain why, even in sparrows reared in isolation, many species-typical attributes of their vocalizations are apparent in the abnormal songs they produce. According to this view of song learning, the formation of the auditory memory guiding song learning would result from a memorization by selective activation and attrition of innate circuits rather than via a selective instruction process. The relative validity of these two opposing models of auditory memory formation has yet to be resolved.

Species Variation in Song Learning: Evolutionary Considerations

Species variation in song learning can be approached from several different levels. First, there is variation at a high level of taxonomic classification, such as class, order, or family. Second, there is interspecific variation within a family and then intraspecific variation within a species or population. We will consider both levels of variation and discuss possible functions for this variation. Vocal learning of the sort described in songbirds has been described in only two vertebrate classes: mammals and birds. Within these classes there seems to be marked variation as to whether a taxon exhibits any type of learned vocalization based on auditory feedback. For example, in birds there is clear evidence in three of 23 extant orders of birds that vocal learning via auditory feedback matching occurs: the parrot, songbird, and hummingbird orders. Coincidently, these are among the most successful avian orders in terms of number of species. The occurrence of vocal learning in these three orders is associated with the emergence of neural specializations for vocal learning and perception, at least according to studies of immediate early gene expression associated with song production and the hearing of song. Further, more-detailed neuroanatomical studies are needed to clarify the similarities and differences among these different orders, but to a first approximation, there does seem to be a coevolutionary process between the evolution of vocal learning and specialized brain structures. This link between brain variation and behavioral variation is also apparent within the songbird order. The order Passeriformes consists of two suborders. The Tyranni suborder consists of mostly New World species, such as flycatchers, manakins, and cotingas. On the basis of a number of anatomical traits, this group can be viewed as a more primitive songbird compared with the suborder Passeri, which contains a larger number of species and is thought to be more-recently evolved. Passeri contains the traditional species one thinks of as songbirds, such as sparrows, tits, wrens, buntings, tanagers, and warblers. Studies of a very few extant species suggests that Tyrannid species do not learn their vocalizations, and there is evidence that birds in this suborder lack forebrain neural specializations, known as the song system, that are involved in the vocal learning and production so prominent in the Passeri suborder. Thus, one general principle of interest to neuroscientists is that vocal learning coevolves with neural changes. There are some suggestions that members of the Tyranni may exhibit vocal learning. It will be interesting to find out via follow-up studies whether this is the case and whether the occurrence of vocal learning is associated with the emergence of brain specializations associated with vocal learning.

A question that has been addressed by ethologists for many years is why, in terms of adaptive significance, this form of vocal learning has evolved at all. Despite long-standing interest in this question, the answer remains enigmatic. Hypotheses include the fact that variation in song gives males an advantage in mate-choice competition (females often prefer males with more-complex songs) and in male–male interactions (larger learned repertoires are often more effective in repelling conspecific competitors). Another idea, proposed 30 years ago, is that vocal learning promotes the formation of regional dialects that can act as a signal for individuals in the same population who share similar local, coadapted gene complexes. The problem is that nonlearning taxa can exhibit similar processes, and it does not seem necessary to have learning *per se* in order to have complex vocal signals that are effective in intra- and intersexual competition. The fact that many avian orders with vocal learning exhibit a large number of species does suggest that vocal learning will facilitate speciation. However, even in this case, the correlation is not perfect, as in the case of the neotropics, where nonlearning songbirds in the suborder Tyranni can have a higher species density than learning members of the suborder Passeri.

Recently there has been an increase in interest in understanding variation in learning strategies within

taxa that exhibit vocal learning. There are many variations on the theme of vocal learning. Beecher and Brenowitz have pointed out that songbird species vary in when they learn their songs (whether it is the first month or the entire first year or their entire life); in whether they learn a single songtype or an entire repertoire of songs; in whether this repertoire develops by imitating models, by improvising on songs heard from others, or even by inventing songs that apparently bear little in common with model songs; in whether they absolutely require experience with species-typical songs or whether they can learn songs even when isolated from conspecific song; and in whether they are limited to copying conspecific sounds or tend to incorporate many sounds from their environment. All this variation raises questions about mechanism (how does variation in the nervous system control this variation in learning strategies?) as well as function (what are the advantages and disadvantages of this variation?). Issues of mechanism are beyond the scope of this article. Some of the hypotheses relevant to the function of intraspecific variation in song learning overlap with hypotheses put forward for interspecific variation in the occurrence or nonoccurrence of learning. For example, repertoire size, which appears to facilitate the effectiveness of song presentation in some species in both an aggressive and a reproductive context, is enhanced by any learning strategy that enhances the maintenance of variation in song. Another variable that is often evoked is the enhancement of song matching. Song matching means that an individual sings songs that are similar to a neighbor's song. Neighbors with similar songs tend not to fight as much as neighbors with different songs. Data such as these are consistent with the 'dear enemy' hypothesis that new neighbors are much more apt to be expansionist than established neighbors. Hence one will tolerate birds with whom one shares songs because that is a sign that they are from the same local area. How can such matching come about? One way is that songs overproduced during plastic song can be selected based on the population the male settles in when he first reaches sexual maturity. Studies of field sparrows have found that interactions between a male and his neighbors can influence which of the overproduced songs will be selected for crystallization and inclusion in the final repertoire. Songs most similar to the neighbors' songs are retained, and those that are different are rejected. It is as if a process akin to operant conditioning influences which songs are selected for later crystallization among those in auditory memory. The studies that have been completed suggest that there is a rich opportunity to critically evaluate both the functional basis of existing variation in song learning and the question of whether this variation has coevolved with a particular set of neural specializations.

Hormones and Song Learning and Production

Steroid hormones play an important role in regulating the timing of song production as well as phases of the learning process. The link between sex steroid hormone action and song behavior was first made on the basis of field studies correlating seasonal changes in physiology with changes in song behavior. Many male temperate-zone songbirds sing at high rates in the spring compared with other seasons. In these species, seasonal differences in male song are positively correlated with dramatic seasonal increases and decreases in aspects of reproductive physiology such as gonadal size and hormone secretion in the plasma. These correlational studies have implicated testosterone of gonadal origin as being particularly important in promoting high rates of song. However, there is interspecific variability in the degree to which maximal rates of singing are observed outside the breeding period. For example, robins living in northern Europe sing at relatively high rates throughout the year, only pausing in July, although most other songbirds living in the same region do not. Birds that sing outside the breeding season do exhibit seasonal cycles in gonad size and endocrine secretions that are similar to other temperate-zone species. Studies of select species that exhibit territorial song production in the autumn, such as song sparrows in the western United States, mockingbirds, and the European robin, clearly suggest that song behavior in the fall can be elicited by the appropriate stimulus in the absence of gonadal testosterone. It is interesting that in the case of the song sparrow, there is evidence that this autumnal singing involves estrogen acting in the brain even though the gonad is inactive. The source of this neuroactive estrogen in this case does not appear to be from testosterone of gonadal origin that is then locally metabolized. Overall, these data indicate that there is no tight correlation between endocrine activity and song rate in all temperate-zone species.

Seasonal changes in reproductive physiology in these species may relate to changes in other aspects of song, such as repertoire size or stereotypy. For example, seasonal changes in song repertoires have been observed in European starlings and canaries, in which the number of song types and other measures of song complexity may change. In addition, seasonal changes in other measures of song, such as stereotypy, have been described in white-crowned sparrows and song sparrows. A careful consideration of behavioral data from temperate-zone songbirds suggests that although song output is positively correlated with various measures of reproductive physiology,

including hormone concentration in the plasma, there is not necessarily a strong causal relationship between the two, as is the case for sex steroids and certain reproductive behaviors such as lordosis in rats, the bow coo display in male ring doves, or male-typical copulatory behaviors in Japanese quail. Experimental studies of the effects of exogenous hormone administration or castration with hormone replacement on song have been performed on a relatively small number of species, but these studies confirm this view of the relationship between steroids and song behavior. Administering exogenous testosterone can clearly increase song rate. Several different studies of zebra finches have shown that castration greatly reduces but does not eliminate male-typical song whereas in red-winged blackbirds, castration was reported to eliminate adult song production. In the case of song sparrows in the western United States, castrated males were able to maintain fall territories and sang at high rates in response to territorial challenge in a manner that was indistinguishable from intact controls. The hormonal control of song behavior therefore appears to be a clear case of a hormone-enhanced, rather than hormone-dependent, behavior. Species-typical stimulus factors (the presence of a conspecific male and/or female as well as a nest site or a favorable environment) promote song production in males. The presence of gonadal steroids in the plasma can increase the probability and intensity of these behavioral responses to the appropriate stimulus, but this presence is not essential for behavioral activation. One should therefore not be surprised by reports of substantial song production being observed in association with low steroid hormone concentrations in some cases. The stimulus factors releasing song can be so strong in some cases that high gonadal steroid concentrations are not necessary for song production to be observed. The critical hormone acting in the brain to promote song production is not necessarily an androgen. Steroid hormone replacement studies in castrated birds indicate that both androgenic and estrogenic metabolites of testosterone are needed to completely restore high rates of singing. In zebra finches, it has been suggested that estrogenic metabolites selectively promote female-directed song. These data indicate that the two primary metabolites of testosterone may act differently on song behavior.

There is also evidence that sex steroid hormones may also play a role in the song-learning process. Castrated male songbirds are able to form the auditory memories needed for song learning, but they fail to develop the crystallized song characteristic of male birds when they reach sexual maturity in their first year. The birds will stay in a nearly constant state of plastic-song production, but when they receive exogenous testosterone, the song rapidly crystallizes. When juvenile birds are administered testosterone after the sensory phase, when they have formed an auditory memory that will guide sensorimotor

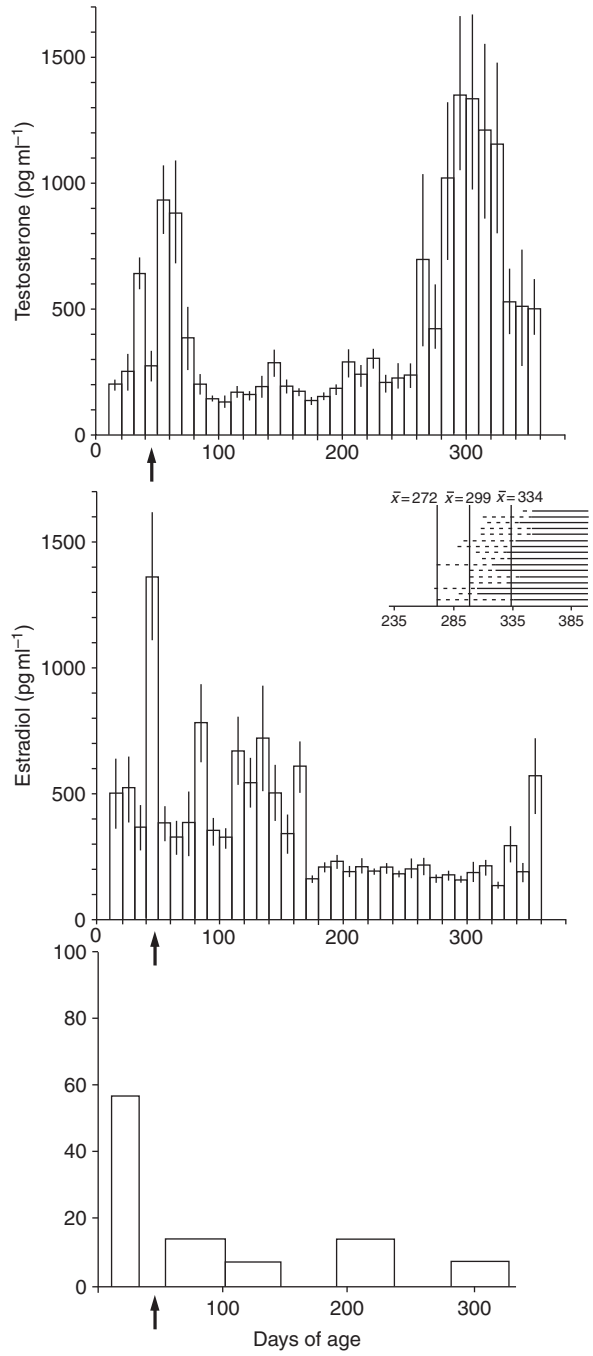

Figure 2 Changes in plasma concentrations in testosterone (top panel) and 17β-estradiol (second panel) during song development in male song sparrows. The bottom panel illustrates the timing of songs that were learned and later produced by the males when they reached sexual maturity. Note that high plasma estradiol concentrations correlate with the high rate of song acquisition prior to day 100, whereas testosterone is high during the onset of song crystallization. The transition from plastic song to crystallized song is illustrated by the inset below the top panel.

learning but before the onset of the sensorimotor-learning period, they will prematurely crystallize their song, albeit in a degraded state. Findings of correlational studies of steroid-hormone changes during development in song sparrows are consistent with the conclusions of these experimental studies (**Figure 2**). Testosterone is relatively low during the sensory phase of song learning and the period afterward, but it rises significantly in close correlation with the transition from the plastic stage of song learning to the stage of crystallized song. These studies of changes in plasma steroid hormones during song development also suggested a possible role for 17β-estradiol in song learning *per se*. It was found that estradiol is high in the plasma during the period of auditory memory formation but that it declines prior to the onset of sensorimotor learning, the time when androgens increase to establish crystallized song (**Figure 2**). Thus estrogens and androgens may play very different but complementary roles, facilitating song learning, in the case of estradiol, but then ending the period of song variability, in the case of androgen. This hypothesized role of estrogen is based on correlational studies and has not been experimentally investigated. The possible cognitive enhancing effects of estrogens on vocal learning remain an intriguing but untested hypothesis.

See also: Bird Brain: Evolution; Bird Song Systems: Evolution; Birdsong Learning; Birdsong: The Neurobiology of Avian Vocal Learning; Signal Production and Amplification in Birds; Vocal Communication in Birds.

Further Reading

Ball GF, Riters LV, and Balthazart J (2002) Neuroendocrinology of song behavior and avian brain plasticity: Multiple sites of action of steroid hormones. *Frontiers in Neuroendocrinology* 23: 137–178.

Beecher MD and Brenowitz EA (2005) Functional aspects of song learning in songbirds. *Trends in Ecology and Evolution* 20: 143–149.

Catchpole C and Slater PJB (1995) *Bird Song: Biological Themes and Variations*. Cambridge, UK: Cambridge University Press.

Kroodsma D (2005) *The Singing Life of Birds*. Boston, MA: Houghton Mifflin.

Marler P and Slabbekoorn H (eds.) (2004) *Nature's Music: The Science of Birdsong*. Amsterdam: Elsevier.

Schlinger BA and Brenowitz E (2002) Neural and hormonal control of birdsong. In: Pfaff D (ed.) *Hormones, Brain and Behavior*, vol. 2, pp. 799–839. San Diego, CA: Academic Press.

Zeigler HP and Marler P (eds.) (2004) *Behavioral Neurobiology of Birdsong*, vol. 1016. New York, NY: Academy Press.

Birdsong: The Neurobiology of Avian Vocal Learning

R Mooney, Duke University Medical Center, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Song Behavior

The Structure of Song

Birdsong and other acoustical signals can be visualized using a sound spectrograph, which plots the frequency components of the signal as a function of time. Many birdsongs are relatively tonal, with sound energy focused within a narrow frequency band centered between 1 and 5 kHz, and can sound pleasantly musical to our ears. Other birdsongs describe a broader frequency bandwidth with little or no harmonic structure, and they sound noisy and less musical to human listeners. One feature of tonal songs is that they can be easily heard above environmental sounds, such as the sound of the wind, water, and rustling leaves, which typically have spectral energy below 1 kHz. Thus, tonal birdsongs are especially effective long-distance communication signals that help individuals maintain large breeding territories during the nesting season.

The elementary unit of song is the note, which appears as a continuous trace on a sound spectrograph (Figure 1). Birds group one or more notes together to form syllables and in turn order these syllables into a stereotyped sequence referred to as a phrase or a motif. The phrase may be repeated one or more times in rapid succession during a song bout. Some adult songbirds, such as white-crowned sparrows and zebra finches, utter only a single song phrase which they maintain largely unchanged throughout life. Other adult songbirds, such as mockingbirds, can sing tens or even hundreds of different songs, which are referred to as song types. Both the fine temporal structure of individual notes, often involving rapid (millisecond range) modulations in frequency and amplitude, and the overall sequence of notes and syllables constituting the phrase are learned features of the bird's song.

The Function of Song

Oscine songbirds comprise more than half of the 9000 extant avian species, a taxonomic diversity paralleled by rich variation in song structure, the degree to which one or both sexes sing, and the social contexts in which song is used. In most species, song is a reproductive behavior uttered by adult males to communicate identity, to establish territory, and to attract females. As with many mating displays, male birdsong serves as a form of ritualized combat to minimize aggressive encounters with other males and as a sexual aphrodisiac to attract and arouse receptive females. Male and female songbirds are extremely proficient at recognizing the songs of other birds, which is not surprising given the importance of song recognition for reproductive success. For example, males largely ignore the songs of familiar neighbors but still respond aggressively when they hear an unfamiliar conspecific song. Another example of recognition is that males of some species with multiple song types will sing songs that most closely match those of their neighbors. Female songbirds recognize and choose mates based on acoustical features of male song. Indeed, perceptual studies suggest that females actually surpass males in their ability to distinguish conspecific songs from imitations of those songs produced by avian mimics, such as mockingbirds. The male song features that females find attractive are not well understood, but one idea is that they are honest indicators of fitness, as with sexually selected features in other animal species. An intriguing and unusual twist with respect to birdsong is that sexual selection acts on a learned rather than innate trait.

Learning/Sensitive Periods

Young birds learn their songs during a juvenile sensitive period marked by two phases, both of which depend on auditory experience. During the first phase, known as sensory acquisition, the young bird listens to and memorizes the song of a tutor, most often the bird's father or a neighboring conspecific male. However, young songbirds raised by a foster species mimic the songs of their foster rather than of their biological parents. Furthermore, young birds raised in isolation from any song model subsequently fail to develop species-typical songs and instead develop rudimentary 'isolate' songs.

During the second phase of song learning, known as sensorimotor learning, the pupil matches its own song to the memorized tutor model. Song produced during sensorimotor learning is highly variable in note and sequence structure, but such plastic song is gradually refined so that it becomes similar to the tutor model. Song structure loses its variable nature and becomes highly stereotyped or 'crystallized' with the onset of sexual maturity. Although juvenile birds do not need further exposure to the tutor song during sensorimotor learning, they do need to hear their own

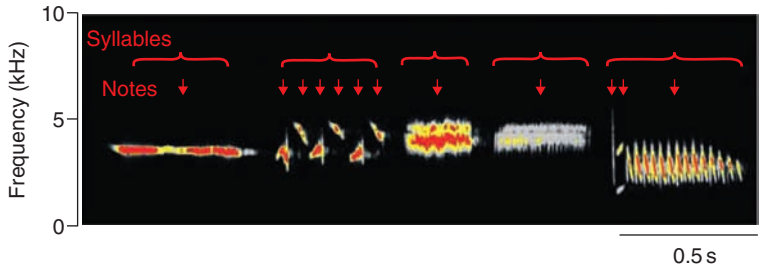

Figure 1 Song structure and nomenclature. A sound spectrograph of the song of an adult male white-crowned sparrow (*Zonotricha leucophrys*). Notes (arrows) are the elementary units of song, and they are grouped to form syllables *(brackets),* a defined sequence of which constitutes the song phrase or motif. The spectrograph, wherein hotter colors correspond to louder sounds, was generated using sound analysis software provided by Ofer Tchernichovski and Partha Mitra.

songs: juvenile birds deafened after hearing a tutor but before starting to sing subsequently develop highly aberrant songs. This suggests that the juvenile uses auditory feedback to learn its song, perhaps by evaluating differences between its own song and a 'template' created by hearing the tutor song. Interestingly, isolates still rely on auditory feedback to develop their songs, and isolate songs within a species share basic features, pointing to the existence of an innate song model that can guide rudimentary learning. Thus, song learning involves two forms of learning: auditory imprinting on the tutor song and vocal motor learning guided by auditory feedback.

Song learning exhibits many remarkable features. One is that exposure to the tutor song regulates the sensitive period duration: juveniles raised in isolation from a tutor display an extended sensitive period and will copy song models later in life than will birds with normal access to a tutor. This indicates that the sensitive period is governed in part by auditory experience rather than an innate 'clock.' Another remarkable feature is that long delays can separate sensory acquisition and sensorimotor learning: in swamp and song sparrows, as well as most other temperate zone songbird species, sensorimotor learning usually begins 8–10 months after the close of sensory acquisition. Yet another feature is that accurate memorization often requires only very limited exposure to the tutor: young nightingales accurately reproduce strings of 75–100 syllables after hearing these strings as few as 15–20 times. Finally, in many species, auditory feedback continues to play an important role in song maintenance: adult songs can decrystallize when birds are deafened or exposed to distorted auditory feedback. These behavioral aspects of song learning make the songbird's brain a rewarding place to search for the neural mechanisms that regulate sensitive periods for learning, facilitate prodigious feats of auditory memorization, and enable feedback-dependent vocal learning and maintenance.

Mechanisms of Singing and Song Learning

Peripheral Mechanisms

The avian organ of song is the syrinx, a bipartite structure located at the confluence of the trachea with the two bronchi. Elegant studies of syringeal and respiratory activity in singing birds have lent considerable insight into avian vocal gymnastics. Particularly impressive is the songbird's capacity to produce asymmetrical series of vocal gestures by operating the two halves of the syrinx independently, in some cases simultaneously generating two harmonically unrelated notes. The exact details of how the spectral features of song are controlled by the syrinx are still debated but depend in part on the bird's ability to manipulate the bronchial apertures via fine control of the syringeal muscles. In addition, upper vocal tract filtering, including effects due to dynamic changes in beak gape, regulates vocal pitch while also removing harmonics, yielding highly tonal sounds. The temporal features of song, as with speech, rely on an extremely fine temporal integration of vocal and respiratory activity. In birds, the respiratory system is highly modified to meet the metabolic demands of flight and includes an extensive set of air sacs that act as bellows to constantly perfuse the relatively inflexible lungs with oxygenated air. Songbirds exploit their superior respiratory capacity to produce virtuoso vocal displays that are distinguished by extremely rapid trills (>30 Hz) of often extended duration.

Central Mechanisms for Singing

In a manner analogous to the evolution of the peacock's exuberant tail fan, sexual selection based on learned vocalizations in songbirds has resulted in a tremendous hypertrophy of those brain areas important to singing and song learning. The songbird's brain is distinguished from those of nonlearners by a well-defined neural

circuit essential to singing referred to as the 'song system' (Figure 2). A striking neuroanatomical parallel between human speech and birdsong is that both are under the executive control of the telencephalon, in contrast to the brain stem-mediated vocalizations typical of many other vertebrates, including nonhuman primates. The amenability of the song system to detailed electrophysiological recordings makes the songbird especially attractive for probing neural mechanisms pertinent to human speech, including auditory–vocal interactions and respiratory–vocal integration executed under telencephalic influence.

The song system can be functionally divided into two separate pathways: one is obligatory for singing and the other is essential to audition-dependent vocal plasticity. The song production pathway includes a series of serially organized telencephalic nuclei, including HVC and RA, the latter of which directly innervates the brain stem vocal–respiratory network. The brain stem vocal–respiratory network includes the tracheosyringeal (XIIts) motor neurons that innervate the muscles of the syrinx, as well as the respiratory premotor neurons in the lateral medulla that control the respiratory rhythm and provide synaptic input to the XIIts motor neurons. The idea that RA strongly modulates and coordinates syringeal and respiratory activity during singing is supported by the observation that bilateral RA lesions render the bird mute for singing but not for other unlearned vocalizations. Further direct support for RA's role as a song motor area derives from studies employing chronic recordings in singing zebra finches, which show that single RA neurons display amazingly precise (<1 ms jitter) action potential discharge locked to specific acoustical features of the bird's song.

How are premotor signals transformed between HVC and RA? In contrast to single RA neurons, which fire many times per motif, the HVC neurons that innervate RA (HVC_{RA} neurons) generate an extremely sparse burst of action potentials, firing only one brief (~10 ms) burst of action potentials per motif. Ultimately, a high degree of synaptic convergence between many HVC axons and single RA neurons could transform a temporally sparse and purely time-based code in HVC to a more continuous movement-based code in RA. One functional advantage of sparse temporal activity in HVC_{RA} neurons is that it maximizes learning rates: Error signals that 'correct' the connections of any single HVC_{RA} neuron alter one rather than many time points in the motif.

Despite important insights into the patterns of HVC and RA neuronal activity during singing, and increasingly detailed information about the synaptic architecture of the song system, the song patterning mechanisms remain enigmatic. In many central pattern generators (CPGs), rhythmical activity is an emergent property of a distributed network of neurons, whereas in other CPGs, rhythmical activity arises from highly localized master oscillatory neurons. Whether the song CPG is distributed across many neurons in one or more song nuclei or localized to a smaller population of master oscillatory neurons is unknown.

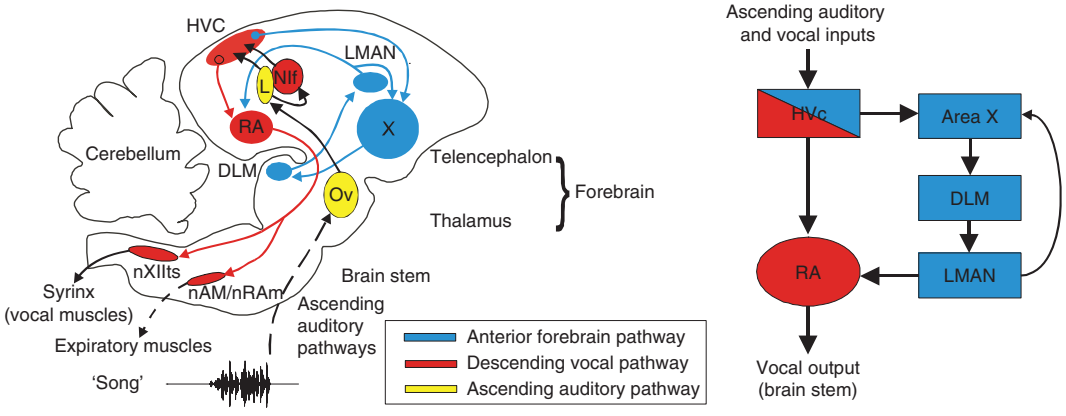

Figure 2 Neural circuitry for singing and song learning. A sagittal view of the songbird's brain (left) and an equivalent schematic (right) showing neural pathways important to learned vocalizations. HVC, RA, and RA's brain stem targets (nXIIts and nAM/nRAm) form a descending vocal motor pathway (red) essential to song, which is emitted from the bird's syrinx. HVC contains two distinct projection neuron populations – one that innervates RA give and another that innervates the anterior forebrain pathway (AFP; blue). The AFP, which is essential to audition-dependent plasticity in juvenile and adult birds, includes the basal ganglia homolog area X, the thalamic nucleus DLM, and LMAN, which influences vocal quality through its connections to RA. Auditory information to both song pathways derives from HVC, which in turn is believed to receive its auditory input from NIf and field L (the equivalent of mammalian primary auditory cortex; Ov is the avian homolog of the mammalian medial geniculate body). The names of various song nuclei are given as acronyms for simplicity sake. Dashed lines indicate that intervening synapses are not shown.

Neural Circuits for Song Learning

A major attraction of the song system is its potential for shedding light on the cellular and systems mechanisms of learning. One tantalizing clue about neural mechanisms of sensorimotor learning is that a second major part of the song system, known as the anterior forebrain pathway (AFP), must be intact for audition-dependent vocal plasticity in both juvenile and adult birds. The AFP connects HVC to RA via an interposed series of structures, including area X (a possible homolog to the mammalian basal ganglia), a thalamic nucleus (DLM), and an anterior telencephalic nucleus (LMAN), and thus resembles mammalian cortico-basal ganglia pathways in its general organization. Lesions of LMAN made in juvenile birds 'freeze' song development, even though similar lesions made in adults exerted no discernible effect on their crystallized songs. Rather remarkably, adult LMAN lesions block deafening-induced song degradation, indicating an active role for the AFP in song maintenance.

What is the synaptic basis for the AFP's effect on vocal plasticity? Intracellular studies revealed that LMAN and HVC axons provide convergent input onto the same RA premotor neurons, with LMAN axons predominantly activating postsynaptic NMDA receptors and HVC axons activating mostly AMPA receptors. That this synaptic arrangement actually drives vocal plasticity is supported by the finding that electrically stimulating LMAN can induce slight song variability in adult birds and that pharmacologically blocking LMAN input onto RA neurons can temporarily freeze song structure in juvenile birds. These observations show that LMAN is necessary for vocal plasticity and point to synaptic mechanisms for vocal plasticity. An important next step is to understand whether the AFP's role in song learning is to actively instruct the song motor system to generate a 'correct' song or simply to permit vocal plasticity.

In support of the idea that the AFP serves an instructive role, neurons in LMAN (as well as those in HVC, the source of auditory input to the AFP) respond in a highly selective manner to playback of the bird's own song (BOS) and thus appear well suited to detect variations in song performance. Although it is tantalizing to speculate that LMAN activity is influenced by auditory feedback when the bird sings its song, the singing-related activity patterns of LMAN neurons do not immediately change when adult birds are exposed to distorted auditory feedback. This suggests that much or all of the singing-related activity in the AFP is a corollary discharge of song motor activity emanating from HVC. Similar distorted auditory feedback studies are necessary in other parts of the song system, including HVC and its auditory afferent NIf, to better understand whether auditory feedback directly modulates singing-related activity in any part of the song system. One possibility is that the auditory activity that can be detected in the song system may serve functions other than conveying feedback, perhaps those related to song perception – a view supported by the finding that lesions in both HVC and the AFP impair the bird's ability to recognize the songs of other birds. Ultimately, a major goal of future research will be to detect where auditory feedback signals affect neurons during singing and how auditory information about vocal performance is translated into a signal that can correct or reinforce certain vocal patterns.

Another outstanding goal is to ascertain the mechanisms of sensory learning. Namely, where and how is the template stored? At first glance, song system neurons with selective auditory response properties might seem well suited to function as part of a template. However, numerous studies indicate that song system neurons are biased to the BOS, in contrast to a template model wherein neurons show a tutor song bias. Although preliminary, emerging evidence suggests that tutor-selective cells may localize to secondary areas of the auditory telencephalon. Whether and how such auditory information is used to influence song learning remain unknown, but the discovery of areas containing neurons with a strong bias to the tutor song could provide important clues as to where early auditory experience acts on the brain and in the future could help reveal how auditory memories literally are recalled.

Conclusions

The songbird is a master vocal mimic that relies on its keen auditory memory and auditory feedback to learn its courtship songs. The behavioral aspects of sensitive periods for song learning are fairly well understood, but we still understand relatively little about the underlying cellular mechanisms that limit such prodigious vocal learning to juvenile life. Are there molecules that regulate the duration of the sensitive period, and can these be artificially regulated to reopen the sensitive period in adult birds? Elegant studies in singing birds have revealed important features of the peripheral mechanisms that generate song and have also revealed some of the central premotor activity patterns associated with singing. Despite these important advances, our understanding of the dynamical neural processes of song patterning remains largely incomplete. Is there a localized central pattern generator for song, or is song produced by a more distributed network? Also, important steps have been

taken toward unraveling the neural mechanisms of vocal plasticity, but exactly how plasticity mechanisms are engaged in the service of sensorimotor learning is unclear. Finally, we still do not know how early auditory experience of a tutor produces lasting song memories and how auditory feedback interacts with these memories to guide sensorimotor learning. Answering these important questions will require more extensive electrophysiological recordings in freely behaving birds, including longitudinal recordings during juvenile periods of song learning, coupled with detailed intracellular analysis of the song circuitry. Some of the most exciting and technically challenging questions in the field of song learning are still awaiting answers.

See also: Bird Song Systems: Evolution; Birdsong Learning: Evolutionary, Behavioral, and Hormonal Issues; Birdsong Learning; Learning, Action, Inference and Neuromodulation; Signal Production and Amplification in Birds; Spike-Timing-Dependent Plasticity Models; Spiking Neuron Models; Synaptic Mechanisms of Learning; Vocal Communication in Birds.

Further Reading

Ashmore RC, Wild JM, and Schmidt MF (2005) Brainstem and forebrain contributions to the generation of learned motor behaviors for song. *Journal of Neuroscience* 25(37): 8543–8554.

Bottjer SW, Miesner EA, and Arnold AP (1984) Forebrain lesions disrupt development but not maintenance of song in passerine birds. *Science* 224: 901–903.

Brainard M and Doupe A (2000) Interruption of a forebrain–basal ganglia circuit prevents plasticity of learned vocalizations. *Nature* 404: 762–766.

Brenowitz EA (1991) Altered perception of species-specific song by female birds after lesions of a forebrain nucleus. *Science* 251: 303–305.

Doupe A and Kuhl P (1999) Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience* 22: 567–631.

Hahnloser R, Kozhevnikov A, and Fee M (2002) An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* 419: 65–70.

Immelmann K (1969) Song development in zebra finch and other Estrildid finches. In: Hinde RA (ed.) *Bird Vocalisations*, pp. 61–74. London: Cambridge University Press.

Kao MH, Doupe AJ, and Brainard MS (2005) Contributions of an avian basal ganglia–forebrain circuit to real-time modulation of song. *Nature* 433(7026): 638–643.

Konishi M (1965) The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift fur Tierpsychologie* 22(7): 770–783.

Leonardo A (2004) Experimental test of the birdsong error-correction model. *Proceedings of the National Academy of Sciences of the United States of America* 101(48): 16935–16940.

Margoliash D (1983) Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *Journal of Neuroscience* 3: 1039–1057.

Marler P and Peters S (1981) Sparrows learn adult song and more from memory. *Science* 213: 780–782.

Marler P and Peters S (1988) Sensitive periods for song acquisition from tape recordings and live tutors in the swamp sparrow *Melospiza georgiana*. *Ethology* 77: 76–84.

Marler P and Tamura M (1964) Culturally transmitted patterns of vocal behaviour in sparrows. *Science* 146: 1483–1486.

Mooney R (1992) Synaptic basis for developmental plasticity in a birdsong nucleus. *Journal of Neuroscience* 12(7): 2464–2477.

Mooney R (2000) Different subthreshold mechanisms underlie song selectivity in identified HVc neurons of the zebra finch. *Journal of Neuroscience* 20: 5420–5436.

Mooney R and Prather J (2005) The HVC microcircuit: The synaptic basis for interactions between song motor and vocal plasticity pathways. *Journal of Neuroscience* 25(8): 1952–1964.

Olveczky BP, Andalman AS, and Fee MS (2005) Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLoS Biology* 3(5): e153.

Phan ML, Pytte CL, and Vicario DS (2006) Early auditory experience generates long-lasting memories that may subserve vocal learning in songbirds. *Proceedings of the National Academy of Sciences of the United States of America* 103: 1088–1093.

Nottebohm F and Arnold AP (1976) Sexual dimorphism in vocal control areas of the songbird brain. *Science* 194: 211–213.

Nottebohm F, Stokes TM, and Leonard CM (1976) Central control of song in the canary, *Serinus canarius*. *Journal of Comparative Neurology* 165: 457–486.

Solis MM and Perkel DJ (2005) Rhythmic activity in a forebrain vocal control nucleus *in vitro*. *Journal of Neuroscience* 25(11): 2811–2822.

Suthers RA, Goller F, and Pytte C (1999) The neuromuscular control of birdsong. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 354(1385): 927–939.

Yu AC and Margoliash D (1996) Temporal hierarchical control of singing in birds. *Science* 273: 1871–1875.

Blindsight: Residual Vision

R L Tomsak, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, OH, USA

Published by Elsevier Ltd.

Introduction

The word blindsight is an oxymoron and the phenomenon is a paradox. I treat this subject from the perspective of a clinical neuro-ophthalmologist who has background in experimental psychology and visual neuroscience.

In the least ambiguous sense, blindsight is the discernment of a visual stimulus in the absence of conscious awareness (blindsight type 1 of Weiskrantz). However, it will soon be evident that much of the literature deals with cases in which the subject acknowledges some perception in the blind region of visual field without actually recognizing the stimulus (blindsight type 2 of Weiskrantz). Further subcategorization of blindsight has been done and will be discussed later. All cases of blindsight require some damage to the primary visual cortex (V1) with sparing of parts of visual association cortex and/or associated neural circuitry. Blindsight, in some ways, can be considered a condition in which there is unawareness of awareness, although this term could be applied to certain forms of agnosia, especially anosoagnosia, or the unawareness of a body part. Furthermore, unawareness and denial of cortical visual loss is present in Anton's syndrome. Most types of blindsight, however, are not agnosia but something more primitive, involving damage to, and sparing of, specific neural circuitry dealing with the basics of visual perception.

Background

Long before the term blindsight was coined by Weiskrantz in 1974, the effects of trauma to the visual cortex had been investigated. At the turn of the twentieth century, Inouye systematically studied soldiers who sustained traumatic injury to the visual cortex in war. Thirteen of his cases were totally blind initially (could not distinguish between light and dark), yet all recovered the ability to at least count fingers over weeks to months. Thus, even in adults with wounds resulting in total blindness, recovery of some degree of form vision was the rule rather than the exception.

A few years later, Riddoch studied soldiers wounded in similar circumstances during World War I. He detailed 10 cases and divided them into three groups. The first group (5 cases) could perceive movement in the blind areas of visual field (usually homonymous hemianopsia, complete or incomplete) but at best could perceive only light and no form. The second group (4 cases) ultimately recovered some object vision, but in all movement perception returned first and first became noticeable in the periphery of the abnormal visual field. The last single case (group 3) noted no recovery of either form or movement vision. Riddoch concluded that movement could be considered a special type of visual perception, that the appreciation of movement returned before abilities for object recognition, that improvement of movement perception began in the periphery of visual field, and that the appreciation for light – but not form – returned coincident with the ability to sense movement.

Helen the Monkey

Much of the blindsight literature describes cases of partial blindness – that is, humans or animals with complete or incomplete homonymous hemianopsias, with some form of residual vision in the blind half of the visual field. Rather than starting with the hemianopic situation, I prefer to describe in some detail the best studied case of an adolescent monkey after almost complete surgical removal of the striate cortex (V1).

The monkey was called Helen by Humphrey and co-workers. Following her death 8 years after operation, her brain was studied. Pathologic analysis showed that the striate cortex was removed bilaterally, with the exception of a small island of left calcarine cortex that would correspond to a small portion of the far peripheral right upper visual field. If this visual island was functional, it was nowhere near the part of visual field involved in high acuity form vision.

Helen was carefully studied behaviorally and psychophysically. For the initial 19 months following operation, only light perception vision was apparent. Thereafter, the first visual function to recover was the ability to turn her head or eyes toward a moving object. After training, she could reach out for objects. Ultimately, she could recognize black objects against a light background, best done if the objects were presented in her central visual field. At this time, she appeared to have tunnel vision for stationary objects, but full-field movement perception. Reaching behavior

was always preceded by fixating eye movements, and she did not reach for objects presented eccentrically without looking directly at them first. This was in contrast to normal monkeys who easily reach for eccentric objects without looking directly at them.

Approximately 5 years after operation, she was taken outdoors for walks. Initially, she collided with obstacles, but ambulatory vision improved almost daily. Soon she was recognizing trees and reaching out for branches with apparent judgment of depth. In the laboratory, her ability to recognize small objects (currants or pieces of chocolate) on a white floor became acute, but she grasped these objects with her palm rather than with her fingers. When she picked up objects, she was almost always ambulating, leading Humphrey to speculate that her judgment of distances was facilitated by self-motion and dynamic visual information rather than static visual information. She could move in and around three-dimensional objects placed around the room without difficulty. However, when a black piece of tape was placed on the floor surrounded by three-dimensional obstacles and she was allowed to roam the room, she would return to, and persist in trying to pick up, the same piece of tape even after multiple attempts.

She never learned to differentiate shapes. In other experiments, this was demonstrated by her inability to discriminate between a triangle and a circle surrounding a central black spot. Similarly, she could not discriminate between two circles of same size with only one having the central black spot. However, she could easily and accurately choose between a lone black spot and a circle with a black spot at its center. Other experiments were done with light stimuli of different luminances and color stimuli. All and all, it seemed that Helen's focal, or central, vision had been significantly altered by removal of the striate cortex with preservation to large extent of her peripheral or ambulatory vision.

D.B., G.Y., and Other Hemianopes

A large amount of literature reports experiments done with hemianopic human subjects. A number of criticisms have been directed at this experimental model for blindsight. Perhaps the most obvious is the well-documented observation that scotomatous areas can have various depths of visual insensitivity. However, the consensus of data shows that residual visual functions do exist, allowing subjects to detect and localize stimuli presented in the blind areas. What is missing is the normal ability to recognize these stimuli – an ability sometimes described as visual consciousness.

Weiskrantz and colleagues extensively studied human subject D.B., who developed hemianopsia following removal of a benign lesion in the right occipital lobe at the age of 33 years. A second man, G.Y., also has been extensively studied. G.Y. developed hemianopsia after being struck by a car at the age of 8 years. Research on these two individuals, as well as on others, revealed that given the right setting, subjects could detect stimulus orientation, location, spatial frequency, wavelength, and movement in the blind areas of visual field. Other studies found that emotional content could be recognized, and figure completion performed, when the abnormal field was stimulated. Pupillary reactions could be elicited by isoluminant visual stimuli presented in the blind field. Furthermore, priming experiments showed that a stimulus presented to the abnormal field could alter the response to a stimulus presented soon after to the normal half of vision.

In this regard, Marcel studied two right-sided hemianopes who were semantically primed. A word with ambiguous meaning (e.g., 'bank') was presented in the normal field after the abnormal half of field was primed with a word that would bias the interpretation of the other (e.g., 'money' and 'river'). The priming impressively biased the meaning of the ambiguous word in this study.

The previous discussion is not meant to imply that every one of the subjects reported had all abilities demonstrated. Indeed, most hemianopes do not have blindsight; Weiskrantz found blindsight in only 5 of 25 hemianopes tested for this ability. In this regard, my neurologist colleague, Dr. Monroe Cole, wrote an excellent personal description of the subjective effects of hemianopsia. He has never had blindsight or Riddoch phenomenon but initially was very bothered by visual hallucinations. It is interesting that the blindsight literature routinely does not comment on the presence or absence of hallucinations since these might represent confounding events and are very common in clinical practice, especially in the first few weeks following development of hemianopsia.

Types of Blindsight and Neural Pathways

As noted in the introduction, Weiskrantz described two types of blindsight. Type 1 is the ability to discriminate a visual stimulus without conscious perception of doing so, and type 2 involves an awareness of some sort of visual experience in the blind field in the absence of correct recognition of the stimulus.

Danckert and Rosetti expanded the classification to three types (Figure 1), in part based on the observation that in some people with blindsight, a motor act (e.g., accurately pointing to, or looking at, a stimulus) is the

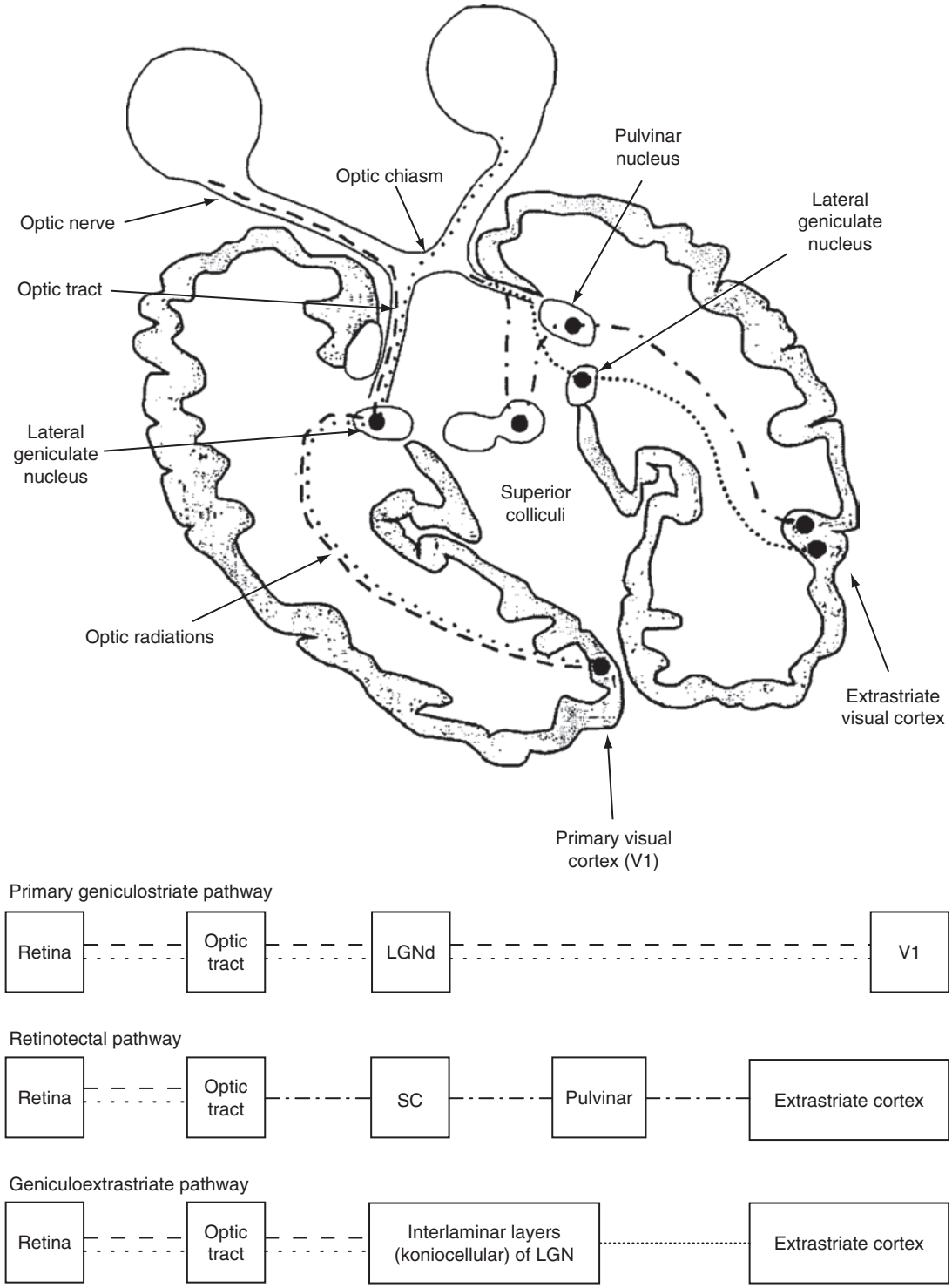

Figure 1 Schematic representation of the various pathways for visual information from the retina to striate (V1) and extrastriate cortex. The primary geniculostriate pathway is indicated by the dashed line from the temporal hemiretina to the left eye and the widely spaced dotted line from the nasal portion of the right eye. The two secondary pathways indicated are shown originating from the optic tract for clarity, with the retinotectal pathway indicated by the dashed/dotted line and the geniculostriate pathway indicated by the closely spaced dotted line. The pathways are also represented in simple box and arrow form below the schematic. Note that anatomical work in monkeys has shown direct koniocellular projections to area MT. The possibility exists for other such pathways from the interlaminar layers of the LGN to regions of extrastriate cortex other than area MT. Reproduced from Danckert J and Rossetti Y (2005) Blindsight in action: What can the different subtypes of blindsight tell us about the control of visually guided actions? *Neuroscience and Biobehavioral Reviews* 29: 1035–1046, with permission from Elsevier.

best demonstration of their residual vision: (1) action-blindsight is demonstrated by grasping, pointing, or saccades; (2) attention-blindsight is demonstrated by detecting spatial orientation, motion detection, and other tasks of discrimination; and (3) agnopsia (to not know what one sees) is the ability of form or wavelength detection and semantic priming. Implicit in this classification is the concept that different neural pathways (Figure 1) subserve each type of blindsight.

A pathway involving superior colliculus, pulvinar nucleus of the thalamus, and poster parietal cortex mediates action-blindsight. A pathway that travels through superior colliculus and pulvinar but terminates instead in the MT region of extrastriate cortex (V5) also mediates attention-blindsight. Agnopsia purportedly is mediated through the koniocellular regions of the dorsal lateral geniculate, interlaminar nuclei of the thalamus, and more ventral regions of extrastriate cortex.

The possible role of koniocellular pathways in blindsight is discussed in detail by Vakalopoulos. This K cell pathway arises in the lateral geniculate from regions between the magnocellular and parvocellular layers. In primates, the K cells project to cytochrome oxidase-rich blobs in V1, as well as to foveal regions in V2, and also to V4 and inferotemporal cortex.

Diffusion tensor imaging has been performed on hemispherectomized individuals with and without attention-blindsight. Ipsilateral and contralateral projections from superior colliculus to primary visual areas, visual association areas, precentral area/frontal eye fields, and internal capsule of the remaining hemisphere were found only in those with attention-blindsight.

Clinical Implications: Visual Restitution Training

A commercially available product (NovaVision) aims at visual restitution training (VRT) for hemianopes. Binocular visual stimulation is performed on a computer monitor within a transition zone between the intact visual field and the impaired field. Debate exists as to whether described improvements are real or are the result of adaptive strategies such as eccentric fixation or changes in saccadic behavior. Scanning laser ophthalmoscopy was used to control fixation behavior in 17 patients with hemianopic field defects studied after VRT training for 6 months; there was an insignificant reduction in the size of the field defects. However, approximately two-thirds of patients believed that the training helped from a subjective standpoint.

Poggel and colleagues studied a man who had a gunshot injury to the back of the head causing visual field loss. Spontaneous recovery of visual field occurred over a 16-month period and then stabilized. Seven months after stabilization, VRT was begun and visual field examinations were done sequentially for approximately 1 1/2 years. VRT-related improvement was found at the borders of the visual field defects, but it was noted that light stimuli of 2000 ms (per VRT protocol) were more easily perceived than stimuli presented by conventional threshold static perimetry (150 ms). Over time, the subject noted improvement in everyday functions such as riding a bicycle, shaking hands with another person, and avoiding obstacles when walking.

In another study, visual field improvement, as measured by clinical perimetry, was only 0.4° on average, but it was 4.9° by the perimetric methods of the VRT protocol (the normal horizontal dimension of visual field is approximately 170°).

Unfortunately, subjects given VRT have not been investigated systematically for blindsight.

Conclusions

Blindsight is a fascinating phenomenon whose study provides insights into the neurophysiology of normal vision. Factors complicating its study include spontaneous improvement of visual functioning after injury, varying depths of scotomatous areas, and operational issues surrounding definitions and experimental designs. Also, the role of learning is a potentially important factor, as shown especially by Helen the monkey. Currently, the efficacy of VRT is controversial and has not been directly correlated with the capacity for blindsight.

See also: Consciousness: Neurophysiology and Visual Awareness in; Vision: Light and Dark Adaptation; Visual System: Multiple Visual Areas in Monkeys; Visual Cortex in Humans; Visual Deprivation; Visual Development; Visual Cortical Models of Orientation Tuning.

Further Reading

Cole M (1999) When the left brain is not right and the right brain may be left: Report of personal experience of occipital hemianopsia. *Journal of Neurology, Neurosurgery, and Psychiatry* 67: 169–173.

Cowey A (2004) The 30th Sir Frederick Bartlett lecture: Fact, artifact, and myth about blindsight. *Quarterly Journal of Experimental Psychology* 57A: 577–609.

Danckert J and Rossetti Y (2005) Blindsight in action: What can the different subtypes of blindsight tell us about the control of visually guided actions? *Neuroscience and Biobehavioral Reviews* 29: 1035–1046.

Humphrey NK (1974) Vision in a monkey without striate cortex: A case study. *Perception* 3: 241–255.

Humphrey NK (2006) *Seeing Red: A Study in Consciousness*. Cambridge, MA: Harvard University Press.

Inouye T (2000) Visual disturbances following gunshot wounds of the cortical visual area. *Brain* 123(supplement): 1–101.

Kentridge RW, Heywood CA, and Weiskrantz L (1997) Residual vision in multiple retinal locations within a scotoma: Implications for blindsight. *Journal of Cognitive Neuroscience* 9: 191–202.

Leh SE, Johansen-Berg H, and Ptito A (2006) Unconscious vision: New insights into the neuronal correlate of blindsight using diffusion tractography. *Brain* 129: 1822–1832.

Marcel AJ (1998) Blindsight and shape perception: Deficit of visual consciousness or of visual function? *Brain* 121: 1565–1588.

Poggel DA, Kasten E, Muller-Oehring EM, Sabel BA, and Brandt SA (2001) Unusual spontaneous and training induced visual field recovery in a patient with a gunshot lesion. *Journal of Neurology, Neurosurgery, and Psychiatry* 70: 236–239.

Reinhard J, Schreiber A, Schiefer U, et al. (2005) Does visual restitution training change absolute homonymous visual field defects? A fundus controlled study. *British Journal of Ophthalmology* 89: 30–35.

Riddoch G (1917) Dissociation of visual perception due to occipital injuries, with especial reference to the appreciation of movement. *Brain* 40: 15–57.

Vakalopoulos CA (2005) Theory of blindsight – The anatomy of the unconscious: A proposal for the koniocellular projections and interlaminar thalamus. *Medical Hypotheses* 65: 1183–1190.

Weiskrantz L (1998) *Blindsight: A Case Study and Implications*. New York: Oxford University Press.

Wessinger CM, Fendrich R, and Gazzaniga MS (1999) Variability of residual vision in hemianopic subjects. *Restorative Neurology and Neuroscience* 15: 243–253.

Relevant Website

http://www.novavision.com – NovaVision.

Blood Pressure: Baroreceptors

A F Sved, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

For normal functioning of the cardiovascular system it is essential that arterial blood pressure (ABP) be highly regulated within the normal range. Usually, ABP is tightly regulated, and fluctuations in ABP on a minute-to-minute or hour-to-hour time frame are minimal. As with most homeostatic systems, the key parameter, in this case ABP, is controlled, at least in large part, by a powerful negative feedback reflex. The arterial baroreceptor reflex, the major negative feedback reflex controlling ABP, is composed of afferent nerves sensitive to changes in ABP that convey information to the central nervous system (CNS), outputs of the CNS that can impact ABP, and the central neural pathways connecting these inputs and outputs. This article describes the afferent limb of this reflex, the efferent components, and then the central neural systems connecting the inputs and outputs and providing a complex neural circuitry for the integration of these reflexes with the physiological status of the organism.

Arterial Baroreceptors

The afferent component of the baroreceptor reflex is composed of pressure-sensitive receptors associated with two major blood vessels, the aortic arch and the carotid sinus, and their projections into the CNS. The blood pressures in these two major arteries are arguably the most critical to the organism, because they represent, respectively, whole-body perfusion pressure and perfusion pressure of the brain. The afferents from the aortic arch form the aortic depressor nerve, which joins with the vagus nerve in the neck. The cell bodies of these afferents, as well as other vagal afferent fibers, are present in the nodose ganglion. The baroreceptor afferents from the carotid sinus, along with chemosensitive afferents from this region, form the carotid sinus nerve, which joins with the glossopharyngeal nerve. The cell bodies of these afferents, as well as other glossopharyngeal afferents, are present in the petrosal ganglion. These baroreceptor afferents all project into the CNS and terminate in the dorsal medial medulla in a distinct region of the nucleus tractus solitarius (NTS). A few studies have suggested the presence of arterial baroreceptors in the coronary circulation, but if they exist they do not have the profound influence on cardiovascular regulation that can be readily shown for the aortic and carotid baroreceptors.

Arterial baroreceptors are stretch receptors that are activated (and therefore fire more action potentials) when ABP increases and distends the blood vessel wall (**Figure 1**). As with stretch receptors in other tissues, the response of these afferents adapts to prolonged stretch; in the case of baroreceptors, this adaptation, or resetting, begins to occur within minutes of a sustained change in blood pressure (**Figure 1(b)**). Thus, baroreceptor afferent firing codes acute changes in pressure rather than a specific pressure. It is important to keep in mind that ABP is a pulsatile parameter, and baroreceptor afferent activity follows this pressure wave and provides the brain with a dynamic and pulsating signal.

In addition to being directly sensitive to changes in pressure, various local and circulating factors (e.g., endothelin, prostacyclins) also influence the firing rates of baroreceptor afferent nerves (**Figure 1(c)**). Thus, while changes in ABP clearly influence the baroreceptor signal provided to the brain, this signal cannot be considered as solely reflecting ABP or providing an accurate high-fidelity signal of absolute ABP. Rather, baroreceptor afferent activity might be considered as coding changes in ABP, but in a manner that is influenced by other factors as well. This issue is important when considering the role of the baroreceptor reflex in the long-term regulation of arterial pressure.

In recent years, studies have begun to address the role of specific ion channels involved in barosensitivity of the baroreceptors. Most of this research has focused on the degerin/epithelial Na^+ channel (Deg/ENaC) family of ion channels, and the current data suggest that this family of channels is responsible for the key mechanoelectrical transduction events (**Figure 1(a)**). Deg/ENaC subunits are expressed by barosensitive neurons, and the mechanosensitivity of these neurons is inhibited by the Deg/ENaC blocker, amiloride. Recent data in knockout mice also suggest a role of Deg/ENaC subunits, primarily 1 and 2, in barosensitivity. However, the Deg/ENaC family does not comprise the only channel types at the baroreceptive endings that are involved in baroreceptor activity. Transient receptor potential (TRP) channels have also been suggested to be part of the sensory transduction process, but they are likely involved more as modulators of the primary barosensitive channels. Although TRP channels impart mechanosensitivity in certain other systems, this does not appear to be the case for arterial baroreceptors.

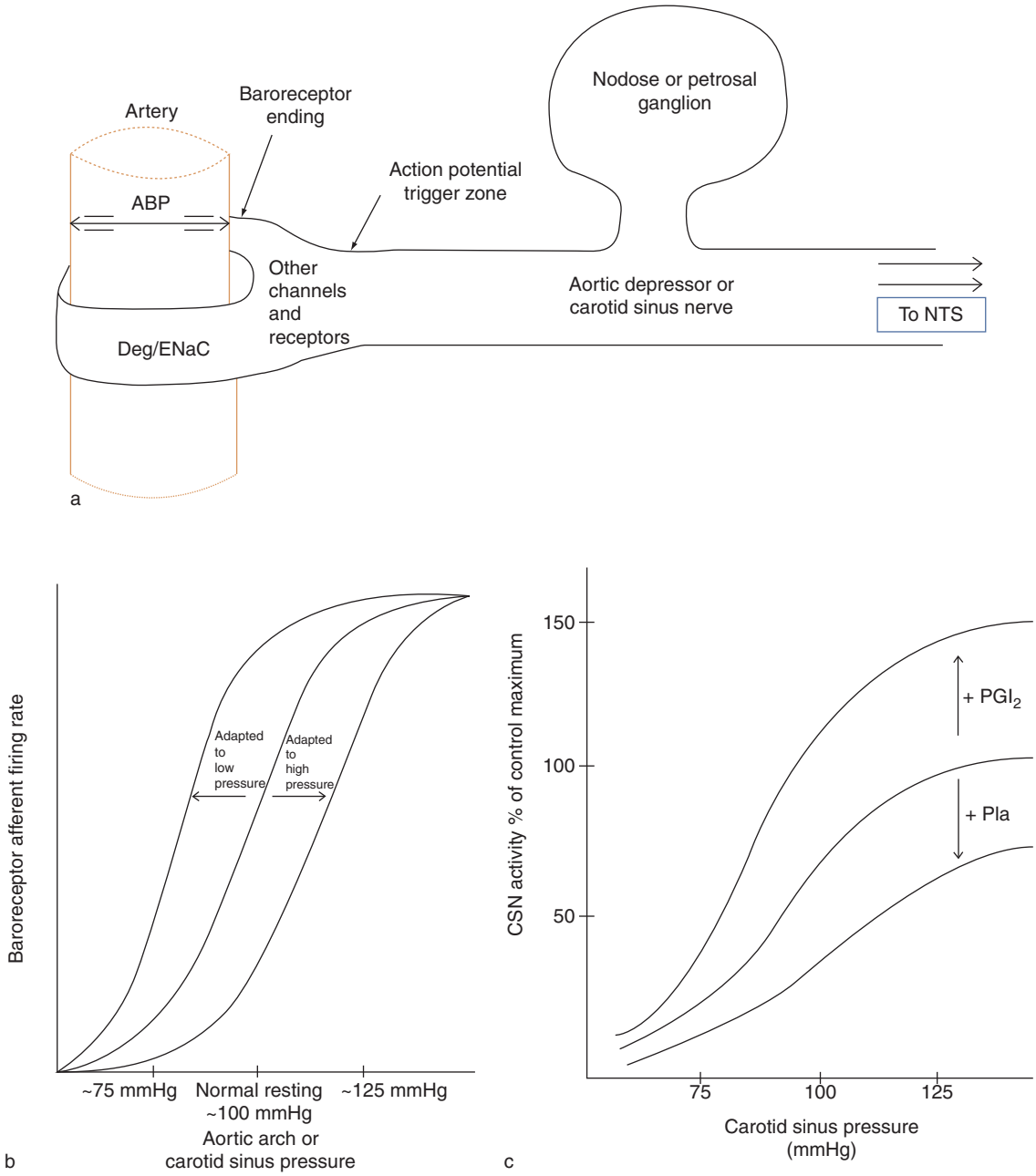

Figure 1 Schematic of a baroreceptor afferent neuron (a), showing the mechanosensitive channels (Deg/ENaC, degerin/epithelial Na^+ channel) on the endings at the blood vessel (ABP, arterial blood pressure), the generation of action potentials at the action potential trigger zone, and the transmission into the nucleus tractus solitarius (NTS). The firing rate of baroreceptor afferent nerves is increased with increasing blood pressure, but this relationship is modified by adaptation to higher or lower pressures (b) or response to chemical agents such as prostaglandin I_2 (PGI_2) or platelets (Pla) (c). (b) Adapted from Munch PA, Andresen MC, and Brown AM (1983) Rapid resetting of aortic baroreceptors *in vitro*. *American Journal of Physiology* 244: H672–H680. (c) Based on Chapleau MW, Cunningham JT, Sullivan MJ, et al. (1995) Structural versus functional modulation of the arterial baroreflex. *Hypertension* 26: 341–347.

Depolarization of baroreceptor endings resulting from stretch, and modified by a variety of other influences, results in the generation of action potentials carried along the axons of these pseudounipolar neurons that ultimately project to the NTS.

Effector Systems of Arterial Baroreceptor Reflexes

The term 'arterial baroreceptor reflex' is used to refer to a variety of physiological responses elicited by

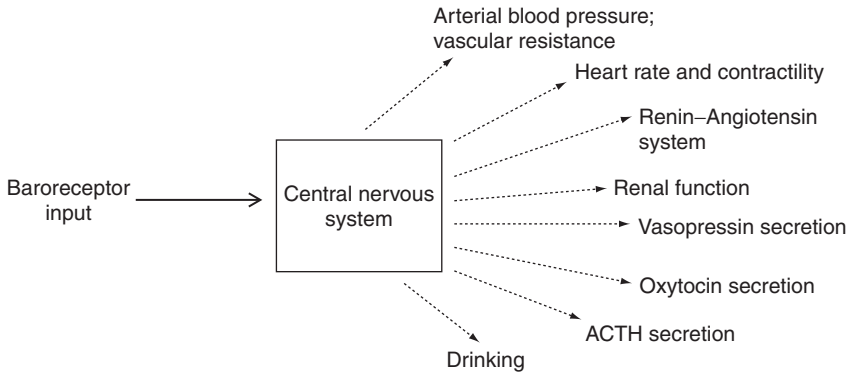

Figure 2 Baroreceptor reflex effector systems. Changes in baroreceptor afferent activity reflexively influence many outputs of the brain relevant to cardiovascular regulation. ACTH, adrenocorticotropic hormone.

changes in baroreceptor afferent activity. While the baroreceptor reflex is often used in specific reference to the reflexive change in ABP and heart rate brought about by changing autonomic outflow to the heart and vasculature, there are a large number of physiological responses resulting from changes in baroreceptor afferent input, and they would each appropriately be termed a baroreceptor reflex. For example, decreases in ABP sensed by arterial baroreceptors elicit an increase in vasopressin (antidiuretic hormone) secretion from the posterior pituitary and an increase in thirst (as reflected by increased fluid ingestion). Thus, baroreceptor reflexes can be viewed as having autonomic, endocrine, and behavioral effectors, all of which interact to promote cardiovascular homeostasis (Figure 2).

At the core of baroreceptor reflexes are the changes in sympathetic outflow, directed at the vasculature and the heart, and in parasympathetic (vagal) outflow, directed at the heart. Changes in baroreceptor afferent activity evoke reflexive changes in autonomic activity to the heart and sympathetic activity to many (but not all) vascular beds. Baroreceptor reflex control of autonomic activity to the heart provides a rapid means of adjusting cardiac output to match ABP. Imposed increases in ABP, detected by arterial baroreceptors, reflexively decrease heart rate (and cardiac output) by increasing parasympathetic activity and decreasing sympathetic activity. Conversely, decreases in ABP elicit the opposite responses (Figure 3). However, these two directions of responses should not be considered simply symmetrical responses. Rather, the decrease in cardiac output elicited by increased ABP is more related to increased parasympathetic activity, whereas the increase in cardiac output elicited by decreased ABP is more related to an increase in sympathetic activity, at least in some species. For example, in rats, the bradycardia caused by acute increases in ABP can be largely blocked by drugs that interfere with the parasympathetic nervous system.

Changes in baroreceptor afferent activity markedly alter sympathetic nerve activity directed toward certain vascular beds, particularly those impacting total peripheral resistance. Thus, whereas baroreceptors readily influence sympathetic activity directed toward renal, mesenteric, splanchnic, and muscle vascular beds, they have limited impact on the cutaneous or cerebral circulations. Increased baroreceptor afferent activity can powerfully inhibit sympathetic vasomotor outflow, whereas decreased baroreceptor afferent activity powerfully excites it.

While these influences of baroreceptor activity on autonomic outflow directed to the heart and blood vessels might be at the core of baroreceptor reflex responses, several other baroreceptor-evoked responses also contribute to cardiovascular homeostasis (Figure 2). Several of these responses are considered here to illustrate specific points.

One of the most powerful regulators of renin secretion from the juxtaglomerular cells of the kidney is the sympathetic innervation of these cells, and this sympathetic innervation is controlled in large part by baroreceptor input. Thus, decreases in ABP result in decreased arterial baroreceptor input to brain and in turn result in increases in sympathetic stimulation of the juxtaglomerular cells and increased renin secretion. The resulting generation of angiotensin II produces a variety of physiological effects, ranging from increased constriction of blood vessels to increased sodium reabsorption by the kidneys and increased fluid intake. Thus, arterial baroreceptors, by influencing renal renin secretion, have a wide influence on systemic physiology.

Arterial baroreceptors also influence secretion of posterior pituitary hormones. Decreased ABP sensed by

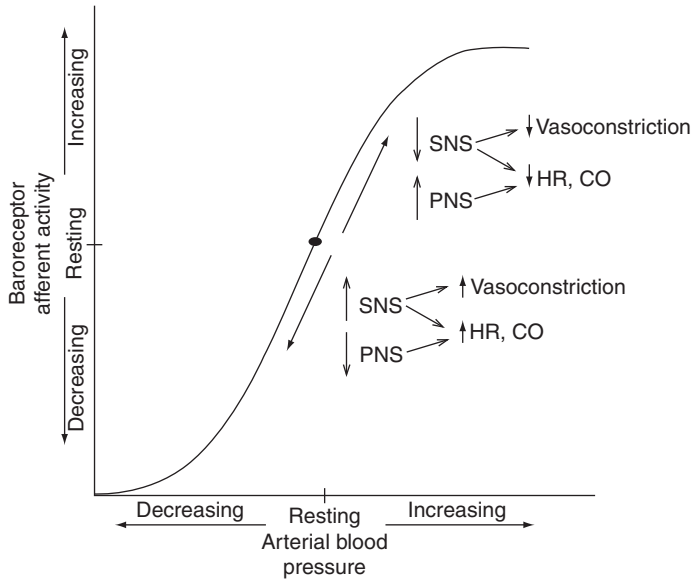

Figure 3 Relationship between arterial blood pressure, baroreceptor afferent activity, and cardiovascular autonomic outflow. SNS, sympathetic nervous system; PNS, peripheral nervous system; HR, heart rate; CO, cardiac output.

arterial baroreceptors increases vasopressin secretion from the posterior pituitary, with readily understandable influences on cardiovascular homeostasis: increased fluid retention by the kidneys and increased arterial vasoconstriction. Increases in ABP momentarily inhibit the activity of vasopressin-secreting cells (a response that has been used as a defining characteristic of magnocellular vasopressin neurons), but this effect does not appear to be sustained long enough to have any physiological impact. Arterial hypotension also causes secretion of the other major posterior pituitary hormone, oxytocin. While the physiological significance of hypotension-evoked oxytocin secretion is not clear, recent studies indicate that this hormone promotes renin secretion from the kidney, at least in rats. In contrast to vasopressin-secreting cells, increases in ABP appear to have no effect on oxytocin-secreting cells.

Changes in ABP, sensed by arterial baroreceptors, influence drinking behavior. Decreases in ABP stimulate fluid intake, an effect that appears to be mediated entirely via arterial baroreceptor-evoked renin secretion with the resulting increase in blood levels of angiotensin II acting on the brain to stimulate drinking. In contrast, increases in ABP inhibit drinking behavior and this effect appears to be mediated by baroreceptor-stimulated neuronal pathways in the brain.

In summary, changes in the afferent activity of arterial baroreceptors reflexively elicit a variety of autonomic, endocrine, and behavioral responses included in cardiovascular homeostasis. While many of these responses occur in opposite fashion to increases and decreases in ABP, the mechanisms responsible for these different effects might be quite distinct.

Central Pathways and Processing

Input to the central nervous system from arterial baroreceptors is through the NTS in the dorsomedial medulla. Within the NTS, baroreceptor afferents synapse with second-order sensory neurons, and at this stage the information begins to get integrated and relayed. The baroreceptor afferents are excitatory to the second-order neurons, likely using glutamate as the primary neurotransmitter. Neural transmission at this initial synapse can be blocked by glutamate receptor antagonists, as can baroreceptor reflexes when such drugs are applied to the NTS. The neural pathways subserving baroreceptor reflex control of parasympathetic and sympathetic responses have been well studied and this circuitry is well established (Figure 4). Neurons in the NTS, likely the second-order sensory neurons, innervate parasympathetic cardiac preganglionic neurons, most of which are present in the nucleus ambiguus. The synapse between the second-order sensory neuron and the parasympathetic preganglionic neuron is excitatory and likely to use glutamate as a neurotransmitter. The baroreceptor reflex circuitry linking the NTS with the sympathetic preganglionic neurons in the intermediolateral cell column in the thoracolumbar spinal cord is somewhat more complex and necessarily involves an inhibitory interneuron, because increases in baroreceptor input decrease sympathetic activity. Barosensitive glutamatergic NTS neurons, at least some of which are second-order sensory neurons, innervate neurons in the caudal ventrolateral medulla (CVLM) and excite γ-aminobutyric acid (GABA)ergic

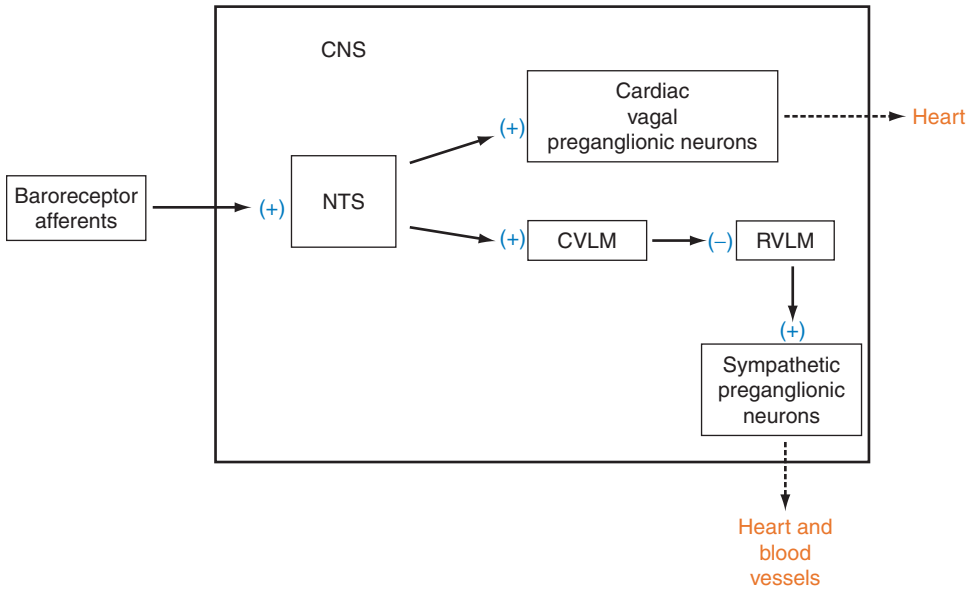

Figure 4 Central nervous system baroreceptor pathway linking baroreceptor afferents to sympathetic and parasympathetic outflow. Plus (+) and minus (–) symbols refer to excitatory synapses and inhibitory synapses, respectively.

neurons in this region. The GABAergic CVLM neurons innervate sympathoexcitatory neurons in the rostral ventrolateral medulla (RVLM) that innervate sympathetic vasomotor preganglionic neurons in the spinal cord. Thus, activation of baroreceptor afferents inhibits sympathetic vasomotor outflow by a pathway involving an excitatory projection from NTS to CVLM, an inhibitory projection from CVLM to RVLM, and an excitatory projection from RVLM to sympathetic preganglionic neurons. Available evidence supports glutamate being the primary neurotransmitter at each of the excitatory synapses in this reflex pathway and GABA being the neurotransmitter at the inhibitory link from CVLM to RVLM. At each of the sites in this pathway, inputs from other brain regions have the capacity to modify baroreceptor reflex responses, and most brain-evoked responses that influence cardiovascular function have some influence on baroreceptor reflex processing.

Central nervous system pathways mediating other baroreceptor-evoked responses are less well characterized. Hypotension-evoked secretion of vasopressin from the posterior pituitary gland involves activation of the magnocellular vasopressin neurons of the paraventricular and supraoptic hypothalamic nuclei, likely by a pathway involving the ventrolateral medulla. Presumably hypotension-evoked excitation of magnocellular oxytocin neurons is via similar pathways, but this has not been extensively studied. The inhibition of vasopressin neurons by acute increases in ABP is mediated by distinctly different pathways from the NTS to hypothalamus (again illustrating the asymmetry between responses evoked by activating vs. unloading of baroreceptors).

As for the central neural pathways involved in baroreceptor-mediated influences on drinking, hypotension-evoked drinking involves activation of the sympathetic innervation of the kidney juxtaglomerular cells, with the resulting generation of angiotensin II. This angiotensin II then acts on neurons in the subfornical organ, a forebrain region lacking a blood–brain barrier, to activate brain circuits involved in drinking. In contrast, increases in ABP inhibit drinking via central neural pathways from the NTS that have not been studied in any detail.

The extensive nature of central baroreceptor reflex circuits is readily demonstrated by studies showing numerous brain regions expressing immediate-early genes such as c-*fos* in response to arterial baroreceptor activation or unloading. All of the regions directly involved in the aforementioned baroreceptor responses express immediate-early genes under conditions in which they would be expected to be activated by baroreceptor manipulations, but many other regions, presumably involved in higher order integration of baroreceptor responses, also express immediate-early genes in these studies.

Role of Arterial Baroreceptors in Cardiovascular Regulation

Arterial baroreceptors provide a dynamic signal to the brain regarding changes in ABP and elicit reflex responses to maintain ABP within a narrow range.

Baroreceptor reflexes are essential for normal cardiovascular homeostasis and stabilize ABP at normal levels on a rapid time frame. Studying the impact of destruction of arterial baroreceptors on cardiovascular regulation has provided useful information on their function. Elimination of baroreceptor signals carried in the aortic depressor nerve and carotid sinus nerve results in a rapid increase in ABP and heart rate, reflecting the marked decrease in baroreceptor afferent signal to brain. Over the course of several days, ABP returns to a normal average level, but the variability around this normal average is markedly increased. A conclusion that can be drawn from such observations is that arterial baroreceptor reflexes are not critical for the maintenance of a normal average level of ABP, but they are essential for stabilizing ABP. These results from baroreceptor-denervated animals have also been used to support the notion that baroreceptor reflexes are not important in the long-term control of blood pressure, but this conclusion has been challenged by more recent studies showing that chronic stimulation of arterial baroreceptor afferents can have a sustained influence on ABP.

Cardiopulmonary Baroreceptors

In addition to the baroreceptors positioned to sense ABP, there are also baroreceptors in the atria, and veins leading to them, referred to as cardiopulmonary baroreceptors. These baroreceptors sense stretch in the atria, which, because of the distensible nature of the atria, and veins leading to them, reflects blood volume. Thus, hypovolemia decreases stretch on these receptors and this leads reflexively to a number of responses aimed at restoring blood volume. Responses initiated by unloading cardiopulmonary baroreceptors include increased vasopressin secretion, increased renin secretion, and increased thirst. Nonhypotensive hypovolemia, as can be caused by moderate hemorrhage or removal of isotonic plasma volume by various experimental procedures, elicits each of these responses. As noted previously, each of these responses is also characteristic of hypotension in the absence of hypovolemia, but the response to unloading of arterial and cardiopulmonary baroreceptors at the same time, as might be produced by hypotensive hemorrhage, is much greater than are responses produced by either stimulus in isolation.

See also: Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Central Cardiovascular Control; Autonomic Nervous System; Cardiovascular Function: Central Nervous System Control; Circumventricular Organs; Energy Homeostasis: Visceral Control; Heart Rate Variability: A Neurovisceral Integration Model; Parasympathetic Nervous System; Pituitary Gland (Cell Types, Mediators, Development); Sympathetic Nervous System; Vasopressin/Oxytocin and Receptors.

Further Reading

Chapleau MW, Cunningham JT, Sullivan MJ, et al. (1995) Structural versus functional modulation of the arterial baroreflex. *Hypertension* 26: 341–347.

Chapleau MW, Li Z, Meyrelles SS, et al. (2001) Mechanisms determining sensitivity of baroreceptor afferents in health and disease. *Annals of the New York Academy of Sciences* 940: 1–19.

Dampney RA, Polson JW, Potts PD, et al. (2003) Functional organization of brain pathways subserving the baroreceptor reflex: Studies in conscious animals using immediate early gene expression. *Cellular and Molecular Neurobiology* 23: 597–616.

Guyenet PG (2006) The sympathetic control of blood pressure. *Nature Reviews Neuroscience* 7: 335–346.

Lohmeier TE, Hildebrandt DA, Warren S, et al. (2005) Recent insights into the interactions between the baroreflex and the kidneys in hypertension. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 288: R828–R836.

Munch PA, Andresen MC, and Brown AM (1983) Rapid resetting of aortic baroreceptors *in vitro*. *American Journal of Physiology* 244: H672–H680.

Stricker EM and Sved AF (2002) Controls of vasopressin secretion and thirst: Similarities and dissimilarities in signals. *Physiology & Behavior* 77: 731–736.

Sved AF, Ito S, Madden CJ, et al. (2001) Excitatory inputs to the RVLM in the context of the baroreceptor reflex. *Annals of the New York Academy of Sciences* 940: 247–258.

Blood–Brain Barrier and Neurovascular Mechanisms of Neurodegeneration and Injury

B V Zlokovic, University of Rochester Medical School, Rochester, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Neurovascular System – Functional Hierarchy

The neurovascular system in the normal brain is organized at different functional levels; the different levels are integrated with each other in order to maintain a tightly controlled chemical composition of the neuronal internal microenvironment. The neurovascular system controls functioning of neuronal circuits through regulation of cerebral blood flow (CBF), in response to brain activation and brain metabolic and clearance demands, and blood–brain barrier (BBB) bidirectional transport of molecules in and out of the central nervous system (CNS). Normal vascular hierarchy in the brain is illustrated in Figure 1. This hierarchical arrangement includes three important functional segments of the neurovascular system: the brain capillary neurovascular unit, the small pial and intracerebral arteries, and the large cerebral arteries. Here, the role of the brain capillary unit in the pathogenesis of chronic neurodegenerative disorders is discussed; how neurovascular dysfunction in small and large cerebral arteries may contribute to neurodegeneration and neuronal injury in Alzheimer's disease (AD) is briefly reviewed.

Brain Capillary Unit in Healthy Brain

Normal human brain has about 100 billion capillaries, which corresponds to about 100 billion neurons. Thus, the ratio of brain capillaries to neurons in normal adult human brain is 1:1. The total length of capillaries in adult human brain is about 400 miles. The surface area available for molecular transport exchanges between blood and brain interstitial fluid (ISF) is about 12 m^2.

Blood–Brain Barrier

Brain capillary endothelial cells (BECs) represent the site of the anatomical BBB; they are normally sealed to the tight junctions, forming a continuous cellular layer that limits free transport of solutes between brain ISF and blood, and vice versa (Figure 1(b), panels ii and iii). The thickness of the cerebral endothelial membrane is 0.3–0.5 μm.

The endothelial monolayer of the BBB expresses numerous substrate-specialized transport systems, some of which are preferentially localized to the luminal, or blood-facing, side, whereas others are expressed mainly at the abluminal, or brain-facing, side. The BBB has a high metabolic rate, reflecting high energy demands for active ATP-dependent transporters, such as the sodium pump, which is localized mainly at the abluminal side of the BBB and regulates sodium and potassium homeostasis in the CNS. The BBB specific transporters for different classes of nutrients (e.g., for glucose, amino acids, vitamins, and nucleosides), and/or the BBB receptors that transport proteins (e.g., leptin receptors, transferrin, and insulin) and lipoproteins (e.g., low-density lipoprotein receptor) and metals (e.g., iron), mediate transcellular influx of their respective circulating substrates across the endothelium in the direction from blood to brain. Other transport systems at the BBB work mainly to reduce the levels of potentially toxic molecules in the CNS; such toxic components are the result of brain activity and metabolism (e.g., excitatory transmitters, amyloid beta, and metabolic waste products). The efflux systems transport toxic molecules and waste metabolic products in the direction from brain ISF to blood, and the transport process is initiated at the abluminal membrane. The multidrug transporter P-glycoprotein (P-gly) is an ATP-dependent efflux pump that has been shown to be required for the BBB to mediate rapid removal of ingested toxic lipophilic metabolites, including many drugs. The P-gly protein is encoded in humans by the *MDR1* gene and in rodents by the *mdr1a* and *mdr1b* genes.

Basement Membrane

The basement membrane separates endothelium from pericytes and astrocytes, two other cell types which regulate function of the neurovascular unit. The basement membrane is composed of different extracellular matrix structural proteins (e.g., collagen, laminin). In addition, embedded into the basement membrane are several important receptors (e.g., integrins), which play a key role in mediating endothelial signaling, cell migration, and brain capillary tube formation during angiogenic responses. Growth factors such as vascular endothelial growth factor (VEGF) are also bound to extracellular matrix proteins and are activated *in situ* by matrix metalloproteinases (MMPs).

Pericytes

Pericytes are situated on the outer side of the basement membrane. They are responsible for the development

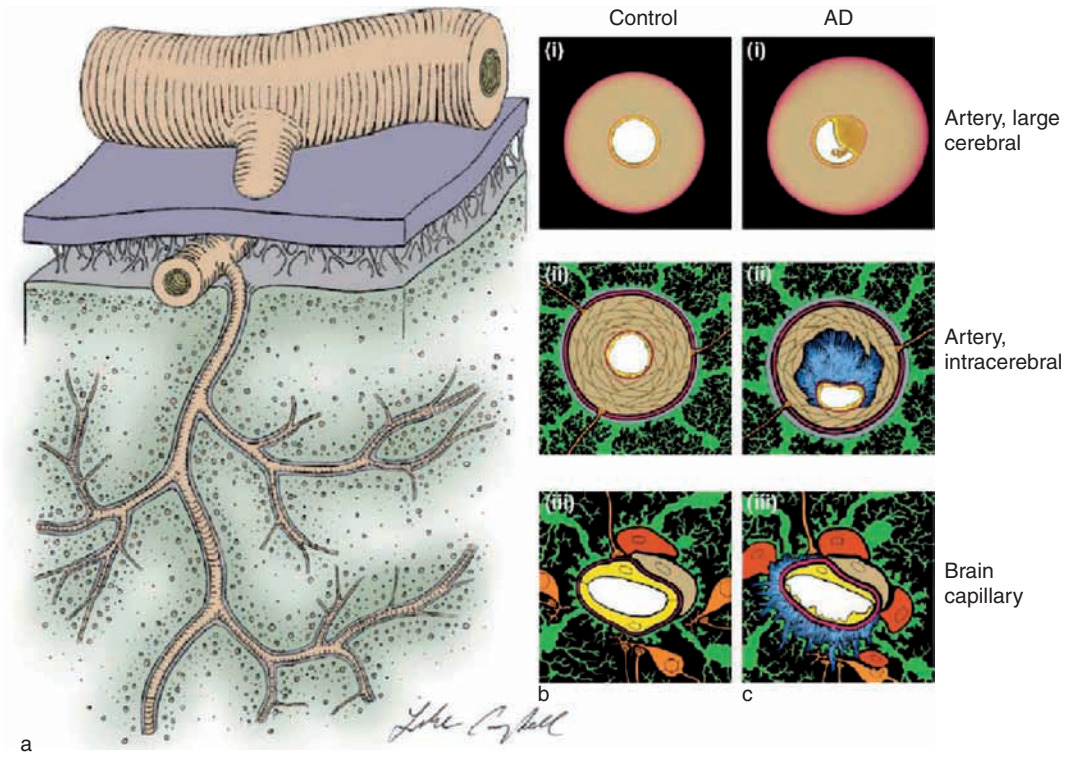

Figure 1 Neurovascular unit. (a, b) Normal vascular hierarchy (top to bottom). Cerebral arteries (a, top; b, i) branch into pial arteries running onto the surface of the brain. Penetrating intracerebral arteries are separated from the brain by the Virchow–Robin spaces. Intracerebral arteries (a, middle; b, ii) branch into numerous smaller arteries and arterioles, which gradually lose the glia-limitans astrocytic membrane (green) of the Virchow–Robin spaces, giving rise to precapillary arterioles and brain capillaries (a, bottom; b, iii). (b) The smooth muscle cell layer (ii, khaki) normally occupies about 90% of the arterial vessel wall. At the capillary level, a tightly 'zipped' brain endothelial layer (iii, yellow) forms the blood–brain barrier. The basement membrane (iii, fuschia) separates the endothelium from other cells of the neurovascular unit, including pericytes (khaki) and astrocytes (green). Astrocytic microdomains (green) ensheathe brain capillary endothelium. Occasional microglia (dark orange) are found, usually in contact with pericytes. Perivascular neurons and synaptic contacts (light orange) are in close proximity to astrocytes (green). (c) Alzheimer's disease (AD) pathology. Large cerebral arteries often display atherosclerotic plaques (a bulge inside the lumen) and may develop thrombotic disorder (i). Pial and intracerebral arteries develop amyloid lesions (ii, blue) and have a reduced smooth muscle cell layer (ii, khaki), characteristics of the cerebral amyloid angiopathy. At the capillary level (iii), there is degeneration of endothelium (yellow), thickening of the basement membrane (fuschia), amyloid lesions (blue), and activation of microglia (dark orange) and astrocytes (green). Local perivascular neurons and synaptic connections (light orange) are severely disturbed. Reproduced from Zlokovic BV (2005) Neurovascular mechanisms of Alzheimer's neurodegeneration. *Trends in Neuroscience*s 28: 202–208, with permission.

of BBB characteristics and play an important role in maintaining vascular integrity and maturation of the BBB. Perivascular pericytes release a large number of growth factors and angiogenic molecules which regulate vascular remodeling, angiogenesis, and blood flow locally within the brain microcirculation, at the capillary level.

Monocytes and Microglia

Perivascular monocytes (which originate in blood and traverse the BBB) and resident microglia play critical roles in innate and adaptive immune responses of the CNS. In certain areas of the CNS, such as in the subventricular zone (SVZ) or dentate gyrus in the hippocampus, pericytes and monocytes are often found together with a small pool of vascular and neuronal progenitor cells which participate in ongoing plasticity responses of the CNS.

Astrocytes

Astrocytes are positioned between neurons and capillary BECs and communicate with these cells through their numerous 'foot processes' (Figure 1(b), panels (ii) and (iii)). Astrocytes play a major role in inflammatory CNS response and regulation of brain water and electrolyte metabolism. Astrocytes, along with vascular smooth muscle cells in small arteries, may contribute to neurovascular coupling, which serves to synchronize neuronal activity and metabolic demands to local cerebral blood flow regulation. Astrocytes contribute

to brain communication pathways by regulating synaptic transmission and neuronal firing thresholds. Autocrine regulations and paracrine cross-talk between endothelial cells, pericytes, perivascular monocytes, microglia, astrocytes, and neurons are critical for normal neuronal functions.

Passive Transport of Molecules in Brain Fluids

Normal cytoarchitecture of brain capillary units allows for rapid diffusion of nutrients and oxygen within brain ISF in the direction from blood to neurons and neurovascular cells, and allows rapid diffusion of neurotransmitters and metabolic waste products away from synapses and neurons back to the microcirculation. The distances for diffusion of molecules between the adjacent human brain capillaries are between 20 and 40 μm.

In addition to passive diffusion, brain ISF–cerebrospinal fluid (CSF) 'bulk flow' mediates transport of molecules into the CSF at a slow rate and irrespective of the size of molecules. The CSF acts normally as a sink for potentially toxic molecules and metabolic waste products. Toxic molecules and metabolic waste products are removed from the CSF by active transport or facilitated diffusion across the choroid plexus epithelium, or by vacuolar transport across the epithelial arachnoid granulations, back into circulation. In contrast, clearance across the BBB mediates rapid removal of several neurotoxins, into the blood, via their respective specialized transport systems.

Brain Capillary Unit in Disorders of the Aging Nervous System

Normal aging is associated with reductions in brain capillary networks and changes in brain microcirculation which result in diminished brain capillary blood flow. These changes, which lead to reductions in the resting CBF, occur in different CNS regions in AD, Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), and other CNS disorders. Moderate reductions in brain capillary density and severe vessel regression may lead to moderate reductions in resting CBF (oligemia) or to severe reductions in CBF (ischemia), respectively, which in turn may cause reversible or irreversible damage of the neurovascular capillary unit.

Modest reductions in CBF are associated with diminished protein synthesis; normal protein synthesis is important for normal learning and memory. More severe reductions in CBF lead to shifts in intracellular pH and water, imbalance in electrolytes, and accumulation of glutamate in brain ISF at levels which may initiate excitotoxic damage. Finally, severe reductions in CBF, similar to those found in ischemic stroke, may lead to ischemic neuronal death mediated by the intrinsic mitochondria-initiated apoptotic pathway, by the extrinsic death-receptor-initiated apoptotic pathway, by aponecrosis, or by necrosis. In several neurodegenerative disorders, changes in the brain capillary unit, degeneration of the brain capillaries, and/or reductions in the resting CBF may be the first sign of disease process prior to neuronal changes and neurodegeneration. A few examples are discussed in the following sections.

Parkinson's Disease

In normal brain, melanin-containing neurons of the zona compacta have a very close spatial relationship with the blood supply. The neuronal vascular contact is so intimate that the capillary appears to have an almost intracellular position. In brains of PD patients, the normal contact between nigral neuron and capillary is lost, and capillaries do not have their normal shape. In early stages of disease, the most striking feature in PD is the changed relationship between capillaries and nigral cells.

Familial Amyotrophic Lateral Sclerosis

Recent work suggests that brain capillaries in familial models of ALS – for example, in superoxide dismutase 1 (SOD1) mutant mice – are affected early in the course of the disease, even during a presymptomatic stage. Reduction in vascular endothelial growth factor (VEGF) level by only 25% impairs spinal cord perfusion and results in motor neuron degeneration, reminiscent of ALS. Moreover, treatment with VEGF extends survival and improves functional outcome in SOD1 mice, whereas antiangiogenic therapy with VEGF2 receptor (VEGF2R) antagonist aggravates the course of disease by accelerating clinical expression of symptoms and by enhancing mortality.

Alzheimer's Disease

Figure 1(c) (panel (iii)) shows altered cytoarchitecture of the capillary unit in AD. Reduced microvascular density, an increased number of fragmented vessels with fewer intact branches, atrophic string vessels, increased irregularity of capillary surfaces, marked changes of vessel diameter (both enlarged and reduced), capillary basement membrane thickening, and collagen accumulation in the basement membrane have been described in AD. Diminished nutrient/oxygen transport from blood to brain cells and impaired elimination of toxic metabolic waste products caused by dysfunction in the transport clearance systems at the BBB, deregulated CBF, increased interstitial diffusional distances, and interrupted ISF–CSF bulk flow upset normal neuronal functioning.

In AD, the brain endothelial wall often looks degenerated. According to recent findings this may reflect unsuccessful aberrant angiogenesis, which is often aborted at an early stage of brain capillary tube formation. Lack of mesenchyme homeobox gene 2 (*MEOX2*) in AD BECs mediates aberrant angiogenesis in AD and abnormal responses to VEGF and other angiogenic factors, resulting in premature vessel regression. Alzheimer's neurotoxin amyloid beta (Aβ) peptide and amyloid deposits accumulate on the outer side of the basement membrane, promoting a local neuroinflammatory vascular response, including activation of brain endothelium, perivascular microglia, pericytes, and astrocytes. Activated cells of the neurovascular unit secrete proinflammatory cytokines and vasoactive substances (e.g., endothelin), further suppressing the CBF and amplifying cellular stress, which may contribute to cognitive defects. Thus, altered physiology of the capillary unit, compromised brain microcirculation, and vascular neuroinflammatory response are of major importance for the pathogenesis of AD.

Blood–Brain Barrier in Disorders of the Nervous System

The involvement of the BBB in several chronic disorders of the CNS has been well documented.

Multiple Sclerosis

The BBB mechanisms critically regulate immune responses of the CNS in conditions such as multiple sclerosis (MS) and in experimental models of MS such as experimental allergic encephalitis (EAE). It has been shown that transport of different subsets of cytotoxic T lymphocytes from blood to brain is critical for lymphocyte infiltration of the CNS, which in turn results in neuronal killing. It is believed that T helper (Th17) lymphocytes, which secrete interleukin-17, play an important role in neuronal killing. But, whether Th17 lymphocytes are activated from pre-Th17 subsets of T cells in peripheral organs and are transported across the BBB, or whether Th1, Th0, and/or other pre-Th17 lymphocytes initially penetrate the BBB followed by maturation into Th17 cells in the CNS locally, is a matter of debate. The BBB involvement in either case is important. It is obvious that the BBB molecular mechanisms which are involved in transport of different subsets of T cells into the CNS represent potentially important therapeutic targets.

Autoimmune Disease Syndrome Dementia

The BBB-regulated cellular trafficking is critical for development of CNS pathology in different neuroinflammatory conditions, including meningitis, encepalitis, and autoimmune disease syndrome (AIDS) dementia caused by human immunodeficiency virus type-1 (HIV-1). For example, HIV-1 infiltrates the CNS largely via infected monocytes and macrophages. HIV-1 infiltration is mediated by its envelope glycoprotein, gp120, which activates protein kinase C isoforms in brain endothelial cells to alter BBB permeability and allow for monocyte migration. Involvement of MMPs during HIV-1 infiltration, and in other disorders of the CNS such as MS, cerebral ischemia, migraine, and brain tumors, has been well established. Thus, transport of peripheral monocytes carrying the HIV virus across the BBB into the CNS precedes development of AIDS dementia. Better understanding of the BBB molecular mechanisms in this case is not only important for the pathogenesis of the disease, but also may represent an important therapeutic target.

Parkinson's Disease

Transport of putative neurotoxins into and out of the CNS across the BBB is considered to be a potential mechanism by which toxins reach their neuronal targets in PD. It is believed that the absorption or metabolism of putative PD toxins is influenced by genetic factors. The importance of the BBB transport in PD has been suggested by *MDR1* gene polymorphisms, which regulate the expression of the P-gly detoxification efflux system at the BBB. Studies in the Chinese population indicate that polymorphisms in the *MDR1* gene can reduce the risk for PD. Involvement of P-gly at the BBB has also been suggested by a positron emission tomography (PET) study demonstrating an increased retention of verapamil (a P-gly substrate) in the midbrain of PD patients. Thus, low levels of P-gly activity at the BBB may favor retention of putative PD toxins in the midbrain of individuals at risk for PD.

It is well known that paraquat, a widely used insecticide, can produce parkinsonism in rodents. The effect of paraquat can be blocked by competitive inhibition of the BBB large neutral amino acid transporter. This further supports the concept that entry of putative toxins into certain brain regions across the BBB could be important for the pathogenesis of PD. It is interesting to note that guinea-pigs exhibit a selective transport of dopamine across the BBB in the caudate putamen, whereas hippocampus and cortex are impermeable to dopamine, suggesting again regional differences in BBB transport. In rats made hemiparkinsonian by unilateral injections of 6-hydroxydopamine into the substantia nigra, treatment with levodopa increases endothelial proliferation and is associated with increased transport of levodopa. The endothelial proliferation and changes

in levodopa transport are limited to the basal ganglia, subventricular zone (SVZ) region, and hippocampal dentate gyrus, and are not found elsewhere in the CNS. Changes in regional capillary endothelial cells have been correlated with the development of dyskinesia, a serious side effect of levodopa therapy in PD patients. This finding suggests that disease treatments may alter BBB properties, which in turn can influence the course of the disease.

Alzheimer's Disease

Elevated levels of Aβ play a primary role in the pathogenesis of AD, and may precede intraneuronal changes and formation of neurofibrillary tangles and tau pathology in models of AD and in AD patients. Work in experimental models of AD has shown that transport exchanges across the BBB critically influence homeostasis of Aβ in the CNS, which is central to formation of neurotoxic oligomeric Aβ species and vasculotoxic forms of Aβ. Despite the fact that Aβ is produced by almost all cells in peripheral tissues and in the CNS, its physiological role remains a puzzle. Aβ normally circulates in plasma, CSF, and brain ISF as soluble Aβ_{40} > Aβ_{42}, and is bound in plasma mainly to its putative chaperone molecules, including soluble lipoprotein receptor-related protein 1 (sLRP1), which prevents reentry of Aβ into the brain. Major influx and efflux of Aβ receptors at the BBB is discussed in greater detail in the following sections.

Transvascular Transport of Aβ

Receptor for advanced glycation end products (RAGE) is a major influx transporter for Aβ across the BBB. RAGE binds soluble Aβ and mediates its pathophysiologic cellular responses. In contrast to normal BBB, which expresses relatively low levels of RAGE, accumulation of RAGE ligands in the aging brain (e.g., AGE proteins, Aβ) increases RAGE expression. Thus, in AD and in transgenic models of β-amyloidosis, RAGE expression increases in affected cerebral vessels, microglia, and even neurons. RAGE/Aβ interaction at the BBB is implicated in development of the disease process in AD through transcytosis of circulating Aβ across the BBB, inflammatory and procoagulant responses in endothelium, brain endothelial nuclear factor-kappa B (NF-κB)-dependent apoptosis, and suppression of CBF.

RAGE/Aβ interaction stimulates microglia activation and response, and produces oxidative damage to those neurons which express RAGE, thus contributing to neuronal killing directly and indirectly via microglia. RAGE inhibitors which block pathophysiolgical responses of RAGE/Aβ interaction in affected vasculature can inhibit cytokine and chemokine production, oxidant stress, and Aβ BBB transport. RAGE inhibitors are currently being tested in AD patients. Besides AD, RAGE plays a role in diabetic CNS vasculopathy and in acute ischemic brain injury after stroke.

Lipoprotein receptor-related protein 1 (LRP1) is a major efflux transporter for Aβ across the BBB. LRP1 is a member of the low-density lipoprotein (LDL) receptor family and acts as a multifunctional scavenger and signaling receptor. LRP1 is expressed in brain capillary endothelium and is localized to the abluminal side at the BBB. Its reduced expression is observed during normal aging in rodents, in nonhuman primates, and in AD patients associated with positive staining of cerebral vessels for Aβ_{40} and Aβ_{42}. Receptor-associated protein (RAP) null mice with severe functional deficiency in LRP1 at the BBB develop accumulations of Aβ when crossed with amyloid beta peptide precursor protein (APP)-overexpressing mice. On the other hand, overexpression of an LRP1 cluster-IV minigene on neurons promotes Aβ retention in the CNS *in vivo*, suggesting that LRP1 on neurons mediates retention of Aβ in brain. In contrast, LRP1 (or another lipoprotein receptor on astrocytes) mediates degradation of amyloid deposits. Therefore, astrocytes and brain vessels act in concert to export Aβ from the CNS, which ultimately requires LRP1 activity at the BBB.

Work in transgenic mice producing low levels of poorly cleared high beta-sheet vasculotropic Dutch/Iowa Aβ mutants suggests that defective LRP1-mediated vascular clearance of Aβ across the BBB leads to accelerated Aβ brain capillary deposition and formation of amyloid lesions. In contrast, mice overexpressing RAGE in the vascular system have enhanced Aβ accumulation and amyloid deposition when crossed with APP-overexpressing mice. The clearance systems for brain Aβ at the BBB have limited capacity and are impaired in AD and animal models of AD, whereas RAGE expression increases. Thus, it is possible that dense Aβ plaques in APP-overexpressing mice and in AD patients originate on blood vessels which later degenerate and/or disappear in the course of the disease. Findings in models of experimental cerebral amyloid angiopathy (CAA) and human CAA indicate that basement membrane and/or extracellular matrix components are critical for the formation of Aβ fibrils and/or seeding of vascular amyloid deposits.

It is of note that mice which lack P-gly at the BBB (knockouts for *mdr1a* and *mdr1b* genes) have reduced clearance of Aβ from the CNS and lower levels of LRP1 in brain capillaries. Crossing *mdr1a*/*mdr1b* null mice with APP-overexpressing mice accelerates accumulation of Aβ and amyloid deposition, raising a possibility that *mdr1a* and *mdr1b* genes may

influence Aβ clearance either directly through P-gly and/or indirectly through LRP1 at the BBB. The role of BBB mechanisms in maintaining Aβ CNS homeostasis is shown in Figure 2.

Small and Large Cerebral Arteries in AD

Aβ is a potent vasoconstrictor in the cerebral circulation. Early changes in APP-expressing mice are associated with impaired endothelium-dependent regulation of neocortical microcirculation and reduction in functional hyperemia. A mismatch between CBF, metabolism, and brain activity is shown in sporadic AD by magnetic resonance imaging (MRI) and PET. Abnormal responses of pial arteries to brain activation, also known as neurovascular uncoupling, have been demonstrated in AD models and AD patients.

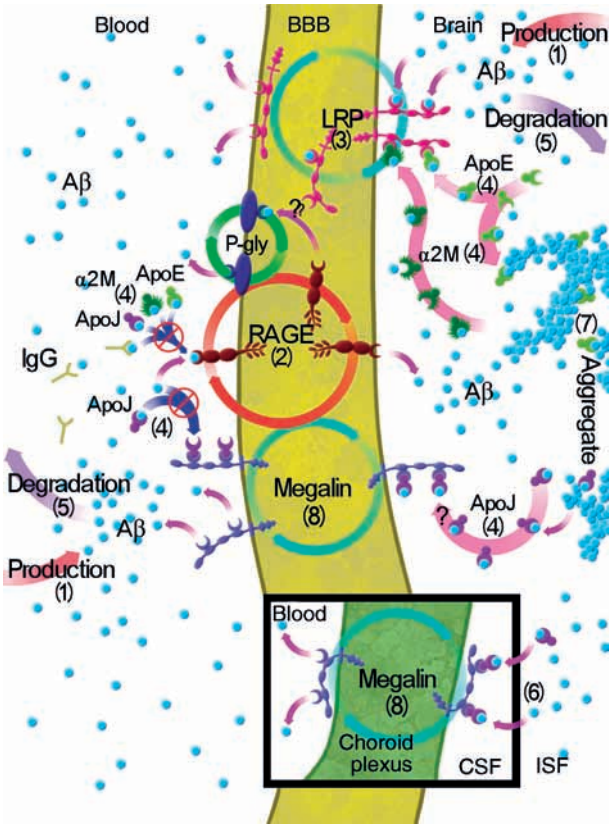

Figure 2 Regulation of brain amyloid beta (Aβ) concentration at the blood–brain barrier (BBB): (1) Rate of production from the amyloid beta peptide precursor protein in brain and periphery; (2) rate of entry into the brain from circulation via the receptor for advanced glycation end products (RAGE); (3) rate of rapid clearance of soluble Aβ from brain to blood across the BBB via lipoprotein receptor-related protein (LRP); (4) binding to proteins such as soluble lipoprotein receptor-related protein 1, apolipoproteins E and J (ApoE, ApoJ), and α_2-macroglobulin (α2M), which can influence Aβ accumulation in brain and/or promote Aβ efflux across the BBB (e.g., soluble LRP in plasma; P-gly, P-glycoprotein); (5) rate of Aβ degradation by a variety of proteases (e.g., enkephalinase, insulinase, plasmin, tissue plasminogen activator, matrix metalloproteinases); (6) rate of slow removal via the interstitial fluid (ISF) and cerebrospinal fluid (CSF) bulk flow into the bloodstream, which appears to be responsible for 10–15% of Aβ clearance from normal brain; (7) rate of oligomerization and aggregation in the central nervous system; (8) transport via other receptors (e.g., megalin). Reproduced from Zlokovic BV (2005) Neurovascular mechanisms of Alzheimer's neurodegeneration. *Trends in Neurosciences* 28: 202–208, with permission.

Pial and intracerebral arteries in AD are affected by CAA. The layer of vascular smooth muscle cells (VSMCs) is often reduced, resulting in the rupture of the vessel walls and intracerebral bleeding. Pathogenic levels of vasculotropic mutant forms of Aβ (e.g., Dutch, Iowa, Arctic, Flemish, Italian) accelerate degeneration of VSMCs, which may contribute to cerebrovascular pathology and hemorrhagic stroke, as seen in familial forms of AD.

Recent studies have pointed out that, in AD, an overexpression of two transcription factors that control VSMC cell differentiation – namely, serum response factor (SRF) and myocardin (MYOCD) – leads to a hypercontractile arterial phenotype and inability of VSMCs to effectively clear Aβ, predisposing to its accumulation in vessel walls and to development of CAA.

Large cerebral arteries in AD do not develop CAA. However, they are regularly affected by atherosclerosis, which reduces resting CBF and brain perfusion and may precipitate a chronic ischemic condition. The Rotterdam Scan Study (a population-based study of nondemented elderly) demonstrated that silent brain infarcts detected on MRI are associated with dementia in elderly people. The longitudinal Nun Study also showed that demented AD individuals with amyloid and tau pathology have numerous brain microinfarctions. The link between increased Aβ production and ischemia has been well documented in these studies.

Neurovascular Model(s) of Neurodegeneration

Accumulating evidence suggests that neurovascular mechanisms and disruption of the neurovascular unit may precede, accelerate, and/or contribute to a disease process leading to neurodegeneration in several disorders in the aging brain. Transport of brain neurotoxins from the circulation across the BBB in AD and PD, inefficient efflux from the CNS of toxic proteinaceous accumulates (e.g., Aβ) or excitotoxins (e.g., glutamate) in AD and stroke, transport of cytotoxic lymphocytes across the BBB (in MS) and of monocytes carrying HIV-1 (in AIDS dementia), and the role of VEGF in ALS are all examples illustrating disease pathways by which the brain capillary unit

controls development of pathology in the aging brain. Activation of endothelium, astrocytes, microglia, monocytes, and pericytes is common to these diseases. Moreover, aberrant angiogenesis and vessel regression at a capillary level may augment brain hypoperfusion already imposed by senescent cerebral vasculature. A chronic neurovascular 'vicious circle' leads to irreversible BBB failure, alterations in the chemical composition of brain ISF, and inflammatory responses which disrupt synaptic and neuronal functions, all leading to chronic neuronal injury and loss.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: Transgenic Mouse Models; Alzheimer's Disease: MRI Studies; Amyloid: Vascular and Parenchymal; Animal Models of Parkinson's Disease; Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Gene Expression in Normal Aging Brain; Huntington's Disease; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Vascular Issues in Neurodegeneration and Injury.

Further Reading

Begley DJ and Brightman MW (2003) Structural and functional aspects of the blood–brain barrier. *Progress in Drug Research* 61: 39–78.

Chow N, Bell RD, Deane R, et al. (2007) Serum response factor and myocardin mediate arterial hypercontractility and cerebral blood flow dysregulation in Alzheimer's phenotype. *Proceedings of the National Academy of Sciences of the United States of America* 104: 823–828.

Deane R, Wu Z, Davis J, et al. (2004) LRP/amyloid beta-peptide interaction mediates differential brain efflux of Abeta isoforms. *Neuron* 43: 333–344.

Deane R, Yan SD, Submamaryan RK, et al. (2003) RAGE mediates amyloid-beta peptide transport across the blood–brain barrier and accumulation in brain. *Nature Medicine* 9: 907–913.

Iadecola C (2004) Neurovascular regulation in the normal brain and in Alzheimer's disease. *Nature Reviews Neuroscience* 5: 347–360.

Kanmogne GD, Schall K, Leibhart J, et al. (2007) HIV-1 gp120 compromises blood–brain barrier integrity and enhance monocyte migration across blood–brain barrier: Implication for viral neuropathogenesis. *Journal of Cerebral Blood Flow and Metabolism* 27: 123–134.

Kortekaas R, Leenders KL, van Oostrom JC, et al. (2005) Blood–brain barrier dysfunction in parkinsonian midbrain *in vivo*. *Annals of Neurology* 57: 176–179.

Lambrechts D, Storkebaum E, Morimoto M, et al. (2003) VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. *Nature Genetics* 34: 383–394.

Lee CG, Tang K, Cheung YB, et al. (2004) MDR1, the blood–brain barrier transporter, is associated with Parkinson's disease in ethnic Chinese. *Journal of Medical Genetics* 41: e60.

Ransohoff RM, Kivisakk P, and Kidd G (2003) Three or more routes for leukocyte migration into the central nervous system. *Nature Reviews Immunology* 3: 569–581.

Tanzi RE, Moir RD, and Wagner SL (2004) Clearance of Alzheimer's Abeta peptide: The many roads to perdition. *Neuron* 43: 605–608.

Vermeer SE, Prins ND, den Heijer T, et al. (2003) Silent brain infarcts and the risk of dementia and cognitive decline. *New England Journal of Medicine* 348: 1215–1222.

Wu Z, Guo H, Chow N, et al. (2005) Role of the MEOX2 homeobox gene in neurovascular dysfunction in Alzheimer disease. *Nature Medicine* 11: 959–965.

Zlokovic BV (2005) Neurovascular mechanisms of Alzheimer's neurodegeneration. *Trends in Neurosciences* 28: 202–208.

Bone Morphogenetic Protein (BMP) Signaling in the Neuroectoderm

C M Mizutani and E Bier, University of California, at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The nervous system is a highly heterogeneous tissue comprising a great diversity of cell types that interconnect in complex patterns to control a myriad of conscious and unconscious behaviors. Not surprisingly, creating such an intricate system requires a series of many cellular interactions during development. Because various organisms have a wide range of different life strategies and needs, there is also a great diversity in the function and development of nervous systems across species. Notwithstanding the inherent complexity and diversity of nervous system function and development, there are remarkable parallels between the formation and function of the nervous system in organisms ranging from fruit flies and nematodes to vertebrates. In several cases, homologous gene sets play critical roles in processes such as neural induction, neurite pathfinding, synaptogenesis, action potential propagation, transmitter secretion and reception, and behavior. This high degree of conservation of basic cellular and molecular functions suggests that the common ancestor of current living metazoans had a well-formed nervous system with many of the core properties shared by diverse present-day organisms.

One of the best characterized examples of conserved pathway function in neural development is the role of bone morphogenetic protein (BMP) signaling during neural induction. During this early phase of embryonic development, BMP signaling actively represses neural cell fates in epidermal regions of the embryo. In neuroectodermal regions, BMP signaling is blocked by various BMP antagonists, which permits the default program of neural development to prevail. Because many of the pathway components required for neural induction are similarly deployed in vertebrates and invertebrates, it seems highly likely that this similarity reflects the conservation of an ancestral mechanism for specifying neural versus epidermal cell fates. BMPs also play important roles in the subsequent patterning of the nervous system along the dorsal–ventral (DV) axis. It is less clear, however, whether this latter phase of neural patterning is accomplished by homologous or convergent mechanisms. In this article, we briefly review the evidence for a conserved function of BMP signaling during neural induction and then focus on how BMPs are believed to act during neural patterning in different organisms. We propose that a unifying theme may underlie the apparent diversity of these patterning mechanisms, wherein BMPs act by a common mechanism to repress the expression of neural genes in a dose-dependent fashion. We also consider how conserved and diverse elements of neural patterning may have evolved.

Evolutionary Conservation of BMP Inhibition during Neural Induction

Nearly a century ago, Hans Spemann and Hilde Mangold showed that ventral transplantation of the dorsal lip of an amphibian embryo into a recipient embryo led to the production of a secondary neural axis. Using distinguishable host and donor embryos they demonstrated further that the dorsal lip, or Spemann organizer as it is now often called, was the source of secreted neural-inducing signals which could redirect the development of surrounding cells fated otherwise to give rise to epidermis. Since these seminal experiments, there has been great interest in isolating and understanding the function of neuralizing factors. Several neural inducers have been identified from *Xenopus* in recent years, including Noggin, Chordin (Chd), and proteins in the DAN family, which are expressed in dorsal mesodermal cells making up the Spemann organizer during late blastula and early gastrula stages. These structurally diverse neural inducers function via a common double negative mechanism by antagonizing the function of BMP signaling (**Figure 1(a)**). They bind to BMPs (BMP2/BMP4) with high affinity, preventing them from activating BMP receptors. In the nonneural ectoderm, where BMP4 is expressed at high levels, BMP signaling functions to promote epidermal fates and to repress the expression of neural genes. Similarly, in *Drosophila*, the Chd homolog known as Short Gastrulation (Sog) is expressed in the lateral neuroectoderm and blocks BMP signaling in the dorsal ectoderm. It is likely that the DV axes in vertebrate and invertebrate embryos were inverted during evolution, such that the epidermis forms ventrally in vertebrates but dorsally in invertebrates. In flies, as in vertebrates, BMP signaling represses the expression of neural genes and activates the expression of nonneural genes. It is noteworthy that in *Drosophila* significantly less BMP signaling is required to repress the expression of neural genes than to activate expression of epidermal genes. One of the genes activated by

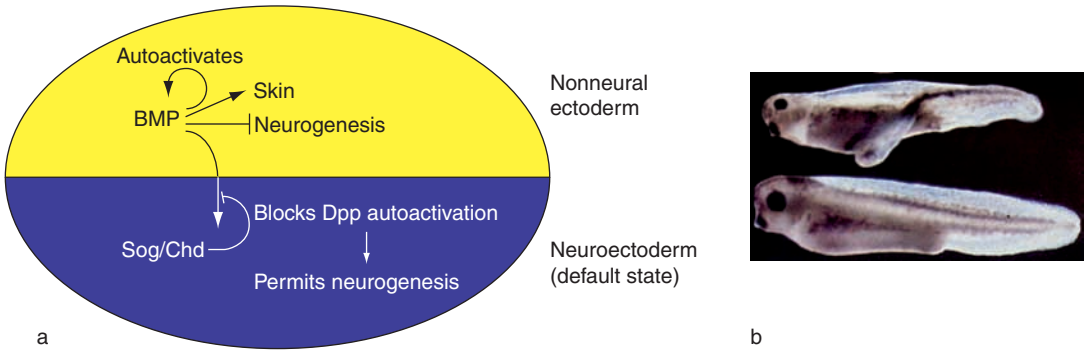

Figure 1 BMPs suppress neuronal fates in the ectoderm of vertebrates and arthropods: (a) BMP signaling in the nonneural ectoderm represses the expression of all neural genes and activates the expression of epidermal genes, including the BMP4/Dpp genes, thereby creating a positive feedback loop, referred to as autoactivation; (b) an example of the conserved nature of BMP inhibition and neural induction in *Xenopus* embryos. In (a), neural-inducing factors, such as the extracellular BMP antagonists Sog/Chd, are supplied to the neural ectoderm, where they bind to BMPs and prevent them from triggering BMP autoactivation, thereby preventing the invasive spread of BMP signaling into the neuroectoderm. This double-negative mechanism allows cells in the neuroectoderm to follow the default neural development pathway. In (b), the injection of *sog* mRNA from *Drosophila* into ventral blastomeres of early *Xenopus* embryos leads to the formation of secondary neural axes (top embryo; compare to wild-type embryo on bottom) similar to those observed in the original embryo-grafting experiments of Hilde Mangold and Hans Spemann. BMP, bone morphogenetic protein; Dpp, Decapentaplegic (the ortholog of vertebrate BMP4/2); Sog, Short gastrulation (the ortholog of vertebrate Chordin (Chd)). Adapted from Biehs B, François V, and Bier E (1996) The *Drosophila short gastrulation* gene prevents Dpp signaling from autoactivating and suppressing neurogenesis in the neuroectoderm. *Genes & Development* 10: 2922–2934; and Schmidt J, François V, Bier E, and Kimelman D (1995) The *Drosophila short gastrulation* gene induces an ectopic axis in *Xenopus*: Evidence for conserved mechanisms of dorsal-ventral patterning. *Development* 121: 4319–4328.

BMP signaling in *Drosophila* and vertebrates is the Dpp/BMP4 gene itself, which results in a positive feedback loop, referred to as autoactivation. If unopposed by BMP antagonists, BMP autoactivation can result in the invasive spread of BMP signaling into the neuroectoderm. This ectopic expression of BMP ligands leads to the repression of neural gene expression and to the activation of epidermal genes.

Conservation of the Chd/BMP signaling system extends to the functional level, as revealed in cross-species experiments. For example, injection of *Drosophila sog* mRNA into the ventral blastomeres of *Xenopus* embryos generates duplicated neural axes (Figure 1(b)) similar to those induced by injection of vertebrate Chd or by transplantation of Spemann organizer tissue. Similarly, vertebrate BMPs and BMP antagonists have the same activities in *Drosophila* as they do in vertebrate embryos. Other extracellular components of the BMP pathway identified in *Drosophila* have also been shown to play similar roles in early vertebrate embryos (Figure 2). For instance, embryos lacking Tolloid (Tld) and Twisted gastrulation (Tsg) activity have defects in BMP signaling. Tld is a metalloprotease that can cleave and inactivate Sog, whereas Tsg forms a trimeric complex with Sog/Chd and BMPs and modifies the BMP inhibitory function of Sog by binding to it and by generating alternative Tld cleavage products. Similarly, in vertebrates, the *Xenopus* counterpart of Tld, Xolloid (Xld), cleaves Chd in positions corresponding to two of the four sites in Sog that are cut by Tld, thereby reducing Chd activity. In addition, *Xenopus* and zebra fish homologs of Tsg also can form a ternary complex with Chd and BMPs to modulate BMP signaling.

Opposing Graded BMP and Hh Signals Pattern the Vertebrate Neuroectoderm

A wealth of embryological and genetic evidence in vertebrates indicates that, following their role in neural induction, BMPs play an important role as morphogens in organizing gene expression along the dorsal–ventral axis of the developing nervous system (note that morphogens are molecules distributed in a graded fashion that function in a dose-dependent fashion to activate or repress gene expression). Once the dorsal-most ectodermal region of the vertebrate embryo is specified as neuroectoderm (often referred to as the neural plate), these cells undergo a concerted set of bilaterally symmetric apical constrictions, causing them to fold inside the embryo by the process of invagination (also referred to more specifically as neurulation). BMP-expressing epidermal cells bordering the neural plate are thereby brought into juxtaposition to form a single coherent dorsal epidermal mass (Figure 3(a)). The invaginated neural plate forms a longitudinal cylinder, which then closes on itself and separates from the overlying epidermis to form the neural tube. The dorsal-most cells of the neural tube lie immediately below the BMP-expressing epidermis and are subsequently induced to express BMPs. This

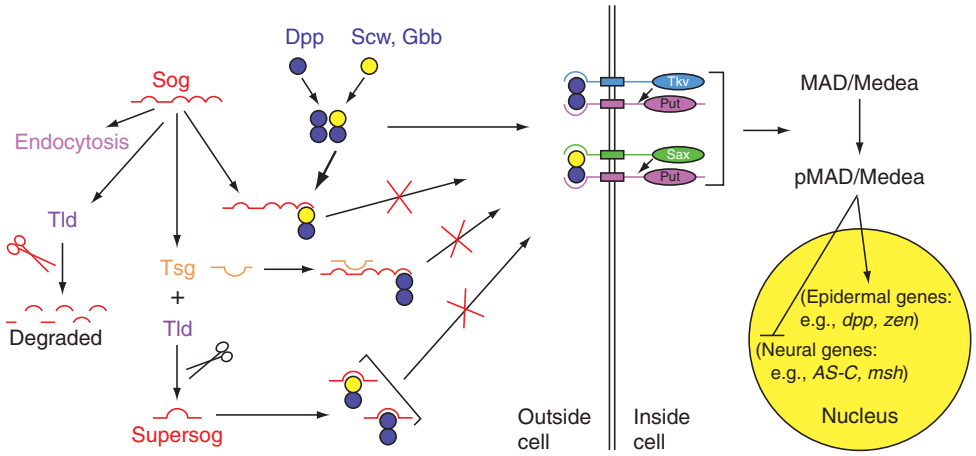

Figure 2 Extracellular regulation of BMP signaling. Diagram of BMP signaling pathway from *Drosophila,* highlighting elements that are conserved in vertebrates. BMP homodimers (e.g., Dpp-Dpp) or heterodimers (e.g., Dpp-Scw) induce the dimerization and then tetramerization of type I and type II BMP receptors. Following receptor dimerization, the type I receptor chain (e.g., Sax or Tkv receptors) phosphorylates the type II chain (e.g., the Put receptor), leading to phosphorylation of the cytoplasmic signal transducer SMAD (or MAD/Medea in *Drosophila*). Phosphorylated MAD (pMAD) then enters the nucleus, where it acts as a transcriptional cofactor to either activate gene expression (e.g., epidermal genes, including *dpp* and *zen*) or repress it (e.g., neural genes, including those of the *Achaete-Scute* complex or *msh*). Extracellular modulators of BMP signaling include Sog/Chd, which binds to Scw and inhibits peak BMP signaling mediated by the Dpp-Scw heterodimer. Tolloid is a metalloprotease that cleaves and inactivates Sog. Dpp is required as a cofactor of Tld in this cleavage reaction, both *in vitro* and *in vivo*. Tsg binds to Sog and Dpp and has been proposed to act in one of two possible ways. First, by forming a trimeric complex with Sog and Dpp-Scw heterodimers, as well as Dpp-Dpp homodimers, it broadens the BMP inhibitory range of Sog. The trimeric complex may also act as a carrier that protects BMPs from receptor-mediated endocytosis and thereby helps concentrate BMP heterodimers along the dorsal midline. The second role of Tsg is to alter the cleavage of Sog by Tld such that alternative, more broadly active forms of Sog (called Supersog) are formed. These truncated forms of Sog can bind directly to either Dpp-Dpp homodimers or Dpp-Scw heterodimers and inhibit their activity. *AS-C*, *achaete-scute* gene complex; BMP, bone morphogenetic protein; Dpp, Decapentaplegic (the ortholog of vertebrate BMP4/2); Gbb, Glass bottom boat; *msh*, *muscle specific homeobox;* Sax, Saxophone; Scw, Screw; Sog, Short gastrulation (the ortholog of vertebrate Chordin (Chd)); Tkv, Thick veins; Tld, Tolloid (the ortholog of vertebrate Xolloid (Xld)); Tsg, Twisted gastrulation; *zen*, *zerknüllt.*

dorsally restricted BMP expression is thought to lead to the formation of a BMP activity gradient, which is high dorsally and low ventrally. High levels of BMP signaling in the dorsal regions of the neural tube result in the expression of genes such as Msx1/2 and Pax7 in cells giving rise to migratory neural crest cells and sensory cells, whereas lower BMP levels result in the expression of lateral markers such as Gsh, Pax6, and Dbx1/2 in cells generating various interneurons (Figure 4(a)). It is not known whether BMPs act directly or indirectly to activate dorsal markers. In current models, BMPs are typically portrayed as having a direct positive role inducing gene expression, in part because a BMP-responsive enhancer region of the Msx1 gene has been shown to have binding sites for SMADs that are required for the activation of this cis-regulatory element. However, it is not clear that this element is responsible for Msx1 expression in dorsal cells of the neural tube because Msx1 is also expressed in ventral cells of the embryo during this same period.

In addition to the gradient of dorsally produced BMPs, the neural tube also receives ventral inductive cues provided by the Sonic Hedgehog (SHh) morphogen. As a consequence of the prior invagination of the mesoderm, cells derived from the Spemann organizer form a stiff longitudinal structure known as the notochord, which underlies the neural tube. These notochord cells secrete SHh and induce the neighboring ventral neural tube cells (called the floorplate) to acquire notochord-like properties, such as expression of the transcription factor HNF3β and SHh itself, which maintains its expression by a positive feedback mechanism (Figure 4(a)). Notochord cells also continue to express BMP inhibitors such as Noggin and Chd. SHh produced in the notochord and floorplate of the neural tube is distributed in a concentration gradient reciprocal to that of the BMP gradient (i.e., SHh is high ventrally and low dorsally). High levels of SHh result in the expression of ventral genes, such as Nkx2.2 and Nkx6.1, in cells that ultimately give rise to motor neurons, whereas lower levels of SHh lead to the expression of lateral markers.

In addition to organizing gene expression in the dorsal and ventral regions of the neural tube, BMPs and SHh also antagonize one another. For example, co-expression of BMP antagonists with limiting amounts of SHh greatly increases the ventralizing activity of SHh. Reciprocally, when BMPs are provided at levels

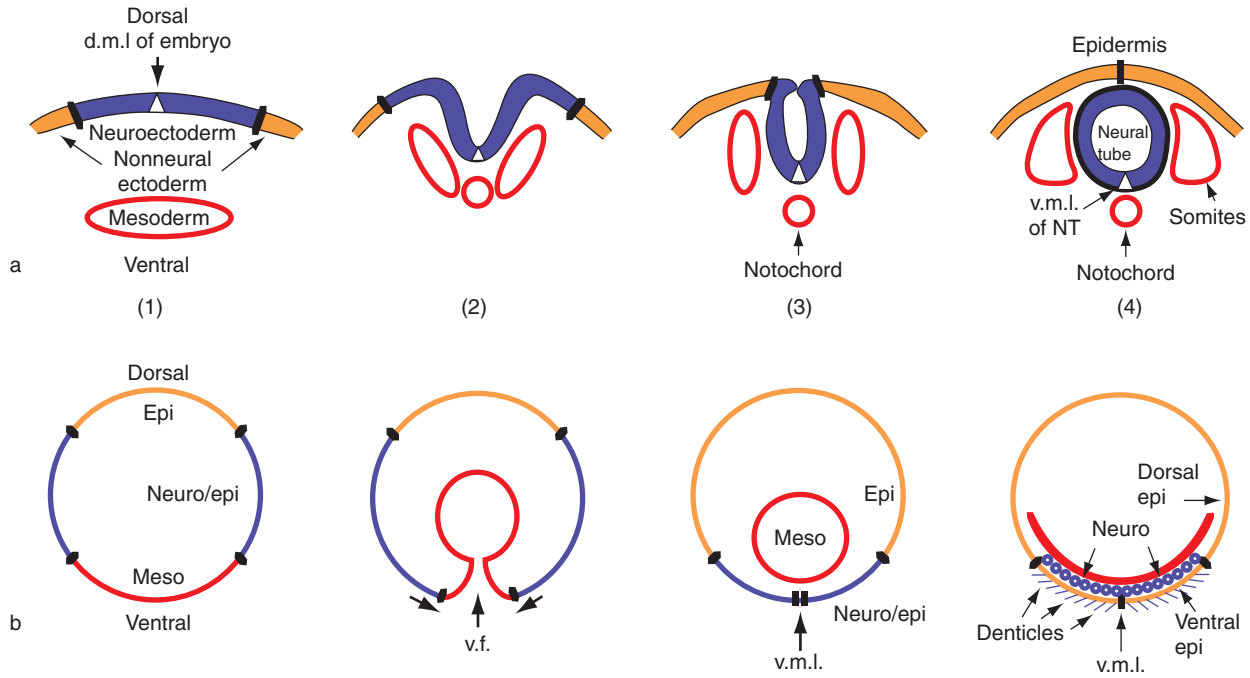

Figure 3 Early neural development in vertebrate and fly embryos, cross-sectional views of embryos: (a) neurulation in vertebrates; (b) delamination of neuroblasts in *Drosophila*. In (a), (1) is a cross section of the dorsal region of the embryo indicating the neuroectoderm (blue), or neural plate, which later invaginates along the dorsal midline. The adjacent more ventrally located nonneural ectodermal cells (orange) express high levels of BMPs and repress the expression of neural genes. Mesodermal cells (red) have already entered the interior of the embryo by this stage, by the process of involution, which begins dorsally at the blastopore and then expands ventrally. In (2) and (3), as invagination of the neural plate proceeds, its two ventral borders are brought into contact and the adjacent epithelial cells fuse into a single coherent domain as the neural ectoderm detaches to form the neural tube. In (4), The invagination of the neural plate leads to a reversal of dorsal-ventral (DV) polarity of the nervous system with respect to the primary embryonic DV axis because cells located originally at the dorsal midline of the embryo (white triangle) assume the most ventral position in the internalized neural tube. The ventral-most cells of the vertebrate neural tube, which are referred to as the floorplate, come into direct contact with a specialized mesodermal derivative known as the notochord, a rigid rodlike structure that provides support to the tadpole. Cells from the Spemann organizer give rise to the notochord. Flanking the neural tube laterally are the somites, a mesodermal tissue that gives rise to the adult bony skeleton and to muscle. In (b), (1) is a cross-section view of the early *Drosophila* embryo indicating the three germ layers along the DV axis. In (2) and (3), invagination of the mesoderm brings the left and right halves of the neural ectoderm into contact to form the ventral midline of the embryo and nervous system. In (4), neural precursor cells (neuroblasts) individually delaminate from the ectodermal epithelium and reside between the overlying epithelium and the more internal mesoderm. An important comparative point is that, due to the double inversion of the DV axis in vertebrates relative to flies, the final orientation ends up being the same. For example, the ventral midlines of vertebrates and flies are both formed by cells that were originally furthest from the source of BMPs in the nonneural ectoderm. BMP, bone morphogenetic protein; d.m.l., dorsal midline; NT, neural tube; v.f., ventral front; v.m.l., ventral midline. (See also **Figure 4**.) Adapted from Bier E (2000) *The Coiled Spring: How Life Begins*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

typical of the dorsal or lateral regions of the neural tube, they can override the induction of ventral markers by SHh.

Graded BMP-Mediated Repression of Neural Genes in *Drosophila*

Following their respective resident roles in consolidating cell fate choices within the epidermal and neural regions of the fly embryo during neural induction, Dpp and Sog play nonautonomous roles in the further subdivision of these two regions. Cells in the dorsal region of the embryo express uniform levels of *dpp* RNA and are initially equivalent because they are defined by the absence of the maternally derived Dorsal (Dl) morphogen. (Dl is a transcription factor related to mammalian nuclear factor (NF)-κB that sets up the initial DV polarity of the embryo. High levels of Dl ventrally specify the mesoderm, graded low levels of Dl define the neuroectoderm, and the absence of Dl in dorsal cells permits expression of *dpp*; see **Figure 4(b)**.) Polarity in the dorsal region is created by Sog diffusing dorsally from the lateral neuroectoderm, where it is cleaved and inactivated by the Tld protease, which is co-expressed with Dpp in dorsal cells. The adjacent ventral source of Sog and dorsal Tld sink result in the formation of a Sog protein gradient in the dorsal region, which is high

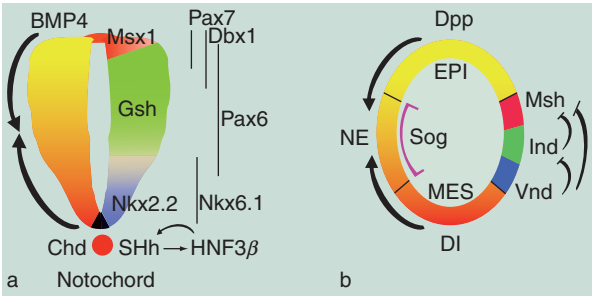

Figure 4 BMPs pattern the vertebrate and *Drosophila* neuroectoderm, cross-sectional diagrams: (a) the vertebrate neural tube, indicating the opposing gradients of BMPs and Sonic Hedgehog (SHh); (b) a blastoderm-stage *Drosophila* embryo, indicating the opposing Dpp and DI morphogen gradients. In (a), high levels of BMPs dorsally result in the expression of the neural identity gene Msx1, and lower levels specify cells expressing lateral genes such as Gsh. High levels of SHh ventrally lead to the expression of the ventral neural identity gene Nkx2.2, whereas lower levels result in the expression of more lateral markers. The expression patterns of some other transcription factors along the dorsal-ventral (DV) axis are also indicated, including Pax7, Dbx1 (and, but not shown, Dbx2, which extends more ventrally than Dbx1), Pax6 (which also plays a conserved role in eye formation), and Nkx6.1. In (b), threshold-dependent repression mediated by the Dpp gradient helps pattern the *ind* and *msh* domains of expression, whereas concentration-dependent activation by the DI gradient helps defining the *vnd/ind* border. An important mechanism contributing to the sharp, mutually exclusive neural domains is the ventral dominant cross-inhibition among the neural identity genes, wherein Vnd represses the expression of *ind* and *msh* and Ind inhibits expression of *msh*. Adapted from Mizutani CM, Meyer N, Roelink H, and Bier E (2006) Threshold-dependent BMP-mediated repression: A model for a conserved mechanism that patterns the neuroectoderm. *PLoS Biology* 4: e313 (online).

ventrally and low dorsally (Figure 5(a)). It has been proposed that this Sog gradient creates an inverse BMP activity gradient with peak levels in dorsal-most cells and lower levels in more ventral cells, which can be visualized by the *in situ* activation of the signal transducer phospho-MAD (pMAD). Sog may also carry Dpp dorsally and concentrate it along the dorsal midline. This BMP gradient results in the nested activation of a series of genes, including the transcription factors *zen*, *pannier*, and *ush*. The primary consequence of the graded activation of Dpp target genes is the subdivision of the dorsal region into two parts: a dorsal-most extra-embryonic domain (amnioserosa) and a more ventral epidermal domain. Note, however, that even the lower relative levels of BMP signaling present in the epidermal portion of the dorsal region are sufficient to repress the expression of all neural genes in those cells.

There is also evidence for a reciprocal influence of the dorsal ectoderm on patterning the lateral neuroectoderm mediated by Dpp diffusing ventrally (although diffusion of small amounts of Dpp ventrally remains to be demonstrated directly). Because Sog and a transcriptional repressor of BMP signaling known as Brinker (Brk) are expressed in the neuroectoderm, the levels of BMP signaling in neuroectodermal cells would be expected to be much lower than those in the dorsal region, where Dpp is expressed and only low levels of graded Sog are present. As a consequence of Dpp being present in limiting amounts within the neuroectoderm, its ability to repress neural gene expression becomes dosage dependent. This dosage-sensitive repression has been most conclusively studied with regard to the expression of the neural identity genes *vnd*, *ind*, and *msh*, which are required to specify the fates of the three primary rows of neuroblasts in the embryonic central nervous system (CNS) (Figures 4(b) and 5(b)). *vnd*, the homolog of vertebrate *Nkx*2.2, is expressed in the ventral-most row of the neuroblasts; *ind*, the homolog of *Gsh*, is expressed in the middle row of neuroblasts; and *msh*, the homolog of Msx1/2, is expressed in the dorsal row of neuroblasts. The fact that orthologous sets of neural identity genes are expressed in the same relative ventral-to-dorsal order with regard to BMP-expressing cells in vertebrates and flies suggests that this configuration reflects an ancestral state that has been conserved during evolution (Figure 4). (Note that, despite the fact that the primary DV axes are inverted in vertebrates and *Drosophila* embryos, the final relative order of neural identity genes ends up being the same as a consequence of the neural DV pattern being reversed with respect to the remainder of the embryo following invagination of the neural plate. Such a secondary reversal does not take place in *Drosophila*, in which neuroblasts delaminate isotopically from the epidermis to form a subepithelial layer, as indicated in Figure 3(b).)

An important regulatory feature of neural identity genes in *Drosophila* is that they cross-inhibit one another in a ventral-dominant fashion in which Vnd represses expression of both *ind* and *msh*, and Ind inhibits the expression of *msh* (Figure 4(b)). As Dpp diffuses ventrally, it represses expression of the intermediate neural identity gene *ind* more effectively than *msh*. This results in *ind*, but not *msh*, being repressed by BMP signaling in dorsal cells of the neuroectoderm, which are closest to the Dpp source. BMP-mediated repression of *ind* expression, in turn, relieves ventral-dominant repression of *msh* by Ind, resulting in *msh* expression in the dorsal-most domain of the neuroectoderm. Thus, as a consequence of the cross-inhibitory interactions among neural identity genes, sharp boundaries of neural gene expression domains are established in response to graded Dpp signaling along the neuroectoderm. Similar cross-regulation of neural identity genes has also been

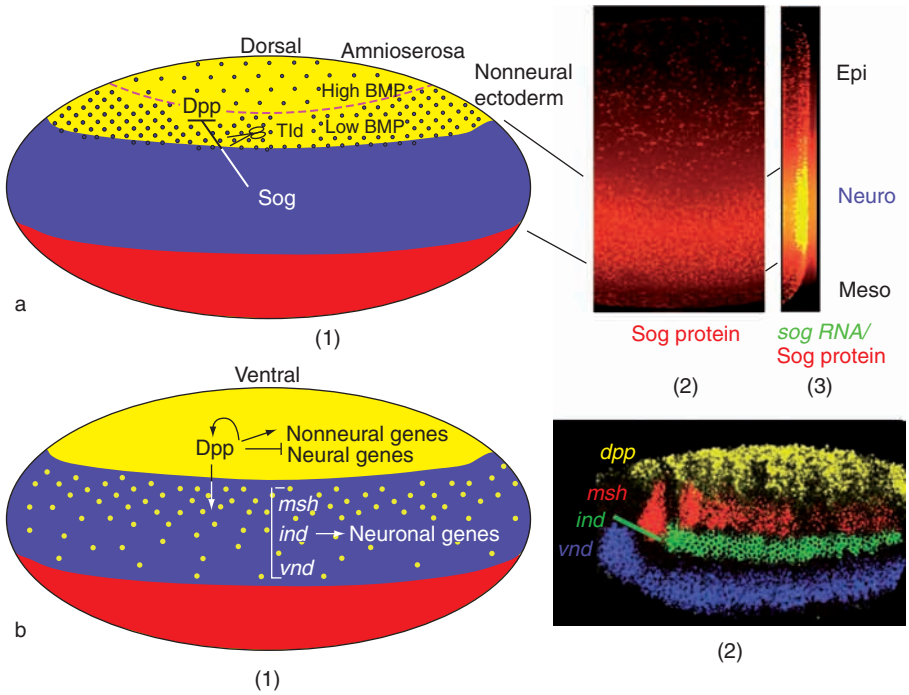

Figure 5 BMP-mediated patterning of the fly neuroectoderm: (a) reciprocal BMP gradient created by diffusion of Sog into dorsal regions and its cleavage by Tld protease in the dorsal-most cells; (b) ventral diffusion of Dpp from the dorsal ectoderm into the lateral neuroectoderm. In (a), double staining of Sog RNA and Sog protein reveals that Sog protein can be detected further dorsally forming a gradient, beyond its domain of expression within the lateral neuroectoderm ((1) and (2) show a lateral view and (3) shows a transversal view of the embryo). In (b) the ventral diffusion of Dpp forms a concentration gradient to pattern the expression domains of neural identity genes *msh* (red), *ind* (green), and *vnd* (blue) mRNA. BMP, bone morphogenetic protein; Dpp, Decapentaplegic; Sog, Short gastrulation; Tld, Tolloid. (a), (1) and (2), from Srinivasan S, Rashka KE, and Bier E (2002) Creation of a Sog morphogen gradient in the *Drosophila* embryo. *Developmental Cell* 2: 91–101. (b), (2), From Kosman D, Mizutani CM, Lemons D, Cox WG, McGinnis W, and Bier E (2004) Multiplex detection of RNA expression in Drosophila embryos. *Science* 305: 846.

observed in vertebrates, athough it remains to be determined whether they follow a ventral-dominant pattern as has been shown in *Drosophila*.

In ventral regions, it appears that the primary system involved in patterning the neuroectoderm is the opposing ventral-to-dorsal Dl gradient, which is provided maternally as previously described. Moderate levels of Dl in ventral cells activate *vnd*, whereas lower levels activate *ind*. Because Vnd represses the expression of *ind* and *msh*, the graded action of Dl results in *vnd* being expressed exclusively in ventral-most cells of the neuroectoderm and *ind* expression in the adjacent intermediate domain where the levels of Dl are too low to activate *vnd*. Although Dpp signaling can also repress the expression of *vnd* and can regulate the dorsal borders of all three neural identity genes, the border between the *vnd* and *ind* domains is established primarily by graded activation of these genes by Dl, whereas the border between *ind* and *msh* is determined primarily by the threshold-dependent repression of these genes by Dpp signaling emanating from dorsal epidermal cells.

Neural Patterning in Other Groups of Organisms

Primary insights into the mechanisms of neural induction have been provided by classical model systems such as flies, frogs, zebra fish, and mice, however, it is important to complement these studies with analyses of organisms from other phylogenetic groups. Such evo-devo studies provide two important types of information. First, cross-genome comparisons have revealed a striking degree of gene loss during the evolution of lineages that include the model systems *Drosophila* and *Caenorhabditis elegans*. Thus, finding a vertebrate gene not present in flies or other insects does not necessarily imply that the gene evolved within the vertebrate lineage following its divergence from invertebrates but, rather, that it may simply have been lost in the insect lineage. Second, one of the most interesting features of evolution is the appearance of novel structures within specific lineages which can only be understood through comparative studies using diverse organisms.

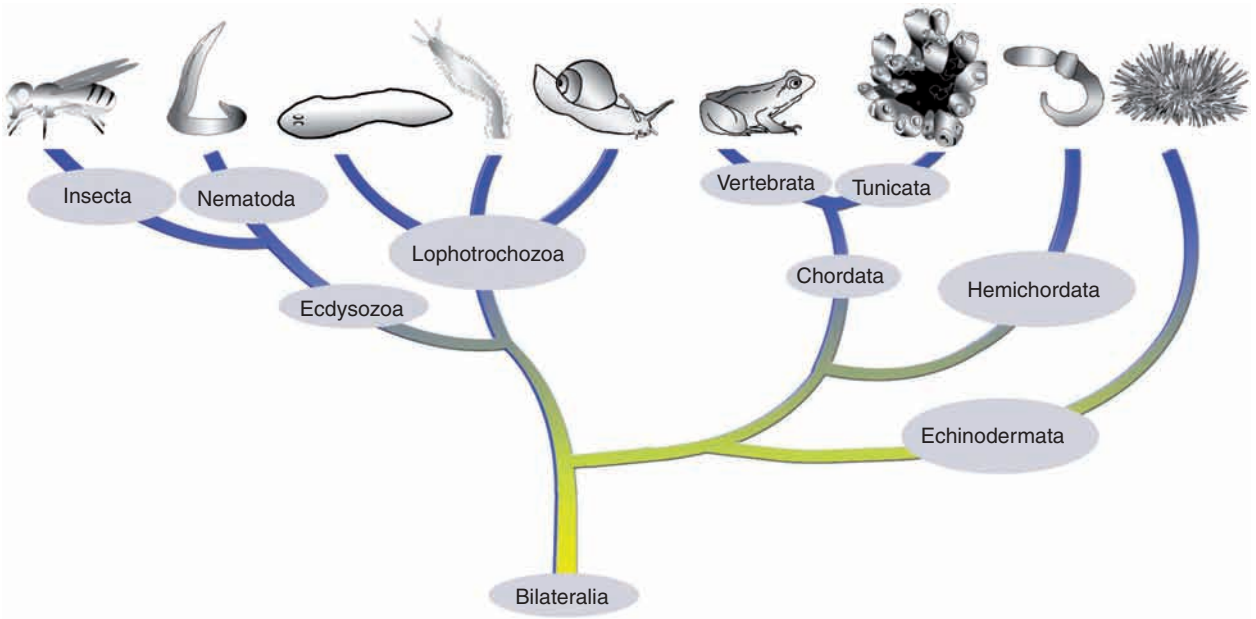

Figure 6 Simplified phylogeny of the bilateralia. Evolutionary tree indicating the relationships of the three major groups of bilateral animals. This tree is based on cladistic analysis of morphological characters and on 18S ribosomal RNA sequence-divergence data. Because of the rapid diversification of animal forms during and preceding the Cambrian radiation, the relationships at the base of the tree are not certain. The current view depicted in this tree is that the bilateralia consist of three major groups: (1) chordates (which include vertebrates and tunicates), hemichordates, and echinoderms; (2) the ecdysozoa, which includes arthropods such as insects and nematode worms such as *C. elegans*, and (3) the lophotrochozoa, which includes mollusks, flatworms, and annelid worms. The great majority of developmental studies have focused on the first two branches of the tree (chordates and ecdysozoa); there has been much less analysis of lophotrochozoan development. Further comparisons of developmental strategies and genetic pathways among these three groups will provide a much improved view of the common ancestor of the bilateralia.

Current phylogenies tentatively group metazoa in one of three major groups: deuterostomes, which include chordates (vertebrates and ascidians), hemichordates, and echinoderms, ecdysozoa (which include arthropods such as *Drosophila* and nematodes such as *C. elegans*), and lophotrochozoa (which include flatworms, annelid worms, mollusks, and other shell-enclosed organisms) (Figure 6). Conspicuously missing among the model organisms, which have been used to define developmental paradigms, are those in the large diverse group of lophotrochozoa. Studies from additional members of the ecdysozoa and chordate lineages would also provide more generality to our current views of development and should shed light on which features are truly conserved versus which independently evolved in different lineages. We briefly summarize some current evo-devo studies in other organisms that bear on the mechanistic origins of neural induction and patterning. It is important to bear in mind, however, that each of these species is also likely to have lost genes that were present in the common ancestor of bilateral animals.

In spider embryos, the DV axis is established in a very different way than in *Drosophila* or vertebrates. A small group of Dpp-expressing mesodermal cells migrates under the epidermis, leaving a linear track of overlying epidermal cells in which BMP signaling persists and which ultimately forms the dorsal midline. The spider *sog* gene is expressed in the ventral ectoderm which gives rise to the nervous system, as in other arthropods. *sog* function is required for ventral cell-fate specification, including the nervous system, because the reduction of *sog* activity by RNA interference (RNAi) results in the spread of high BMP signaling into ventral cells and the subsequent loss of ventral structures. The invasion of BMP signaling into the neuroectoderm of *sog* RNAi spiders and its suppression of neuroectodermal fates parallel the role of BMP signaling in *Drosophila* and vertebrates. Despite the difference in how BMP signaling is established in the spider embryo, the way it is employed supports the view that an ancestral role of neural inducers was to prevent BMP from spreading into the neuroectoderm and suppressing neurogenesis.

Hemichordates, which are thought to be most closely related to echinoderms, include marine worms and other sessile marine organisms that retain only a moderate degree of DV organization as a consequence of their nearly rotationally symmetric body plans. Early during development, BMP4 and Chd/Sog are expressed in opposing domains and define a DV axis in hemichordate embryos consisting of three germ

layers and distinct domains of gene expression, except for the nervous system. Unlike vertebrates and arthropods, the nervous system in hemichordates is not condensed into either dorsal or ventral ganglia but, rather, consists of dispersed neurons which send their axons into one of two major axonal bundles, one running dorsally and the other ventrally. Because neurons form around the entire circumference of these embryos, BMP signaling does not inhibit the formation of neurons in the dorsal region, nor does BMP overexpression inhibit neuron formation elsewhere. In addition, although hemichordates have recognizable counterparts of at least *vnd/Nkx2.2* and *msh/Msx*, the homologs examined so far do not display any obvious restriction in their expression along the DV axis. One possible explanation for these observations is that the neural repressive function of BMPs in vertebrates, arthropods, and spiders arose following the separation of these lineages and that the original function of BMP signaling in bilateral ancestors may have been to establish DV polarity. Alternatively, the neural suppressive function of BMP signaling may have been lost during the course of hemichordate evolution as specialization along the DV axis became greatly simplified and the animals assumed a nearly rotationally symmetric body plan. This latter view accounts for the common pattern of neural identity gene expression in vertebrates and arthropods, as well as its potential common dosage-sensitive regulation by BMP signaling. Future experiments should resolve this question, particularly by examining the expression of neural identity genes in other chordate branches and in various lineages of the lophotrochozoa, the third major branch of the metazoan evolutionary tree. When compared to the other groups, the lophotrochozoa appear to be one of the slowest evolving groups, having lost far fewer genes present in the common bilateral ancestor and typically having ventral nerve cords similar to those in arthropods, although primitive flat worms (platyhelminths) have either diffuse nervous systems or only anterior nerve nets.

In sum, the current knowledge of neural induction in diverse bilateral embryos suggests that the role of BMPs in neural induction reflects the conservation of a mechanism that evolved from a common bilateral ancestor, although it is formally possible that this may have arisen independently in several different lineages. Clearly there are species-specific aspects that have been described, but it is not clear whether this is evidence against a common origin rather than an indication that the mechanism has been lost or highly modified in various lineages. Further analysis of additional groups should resolve these issues.

A second important evolutionary question is whether the role of BMPs in patterning neural identity also originated in a common ancestor. Indeed, because vertebrates and flies share a common set of neural identity genes expressed in the same relative order with regard to a source of BMPs and because BMPs play a prominent role in patterning the dorsal region of the nervous system in both flies and vertebrates, it seems likely that neural patterning by the BMPs was a common feature of the bilateral ancestor. Clearly, other species-specific signaling pathways are also important in DV patterning of the nervous system; for example, primary morphogens involved in ventral neural patterning appear to be different in flies (i.e., Dl) and vertebrates (i.e., Hedgehog (Hh)). Nonetheless, it is tempting to speculate that BMPs once were sufficient to pattern the entire neural DV axis. According to this hypothesis, additional signaling systems were then added to buttress patterning at the low end of the BMP gradient during the divergence of the vertebrate and invertebrate lineages. Consistent with this view are experiments on DV patterning of the mouse spinal cord. When the function of the Hh signaling pathway is completely abolished (i.e., by removing both SHh and the default repressor of the Hh pathway, known as Gli3), much of the ventral pattern is restored relative to what is lost in *SHh*-single mutants. In addition, under conditions of low-level Hh signaling, the gene-expression profile in neural-plate explants can be adjusted to ventral, lateral, or dorsal levels by adding increasing doses of BMPs, indicating again that BMPs alone are able to pattern the full DV span of neural cell fates. Thus, in the early bilateral ancestors, which are believed to have been very small (less than 2-mm long), a single BMP morphogen gradient may have been sufficient to create a pattern along the entire DV axis.

DV Inversion in Vertebrates?

The fact that DV polarity of the nervous system and the circulatory system appears to be reversed in vertebrates relative to invertebrates was noted by the renowned French comparative anatomist Geoffroy St.-Hilaire, who proposed that vertebrates were essentially upside-down invertebrates (**Figure 7(a–c)**). The patterns of gene expression in vertebrates and invertebrates summarized here have led many modern evo-devo enthusiasts to support Geoffroy St.-Hilaire's hypothesis. One possible exception to the axis inversion model, however, is the head region. Comparison of gene-expression markers for eyes such as *Pax6/eyeless* (which are thought to have played an ancestral role in specifying some properties of light-sensitive organs in metazoa), as well as genes expressed in the vertebrate hypothalamus and a potentially homologous neuroendocrine organ in

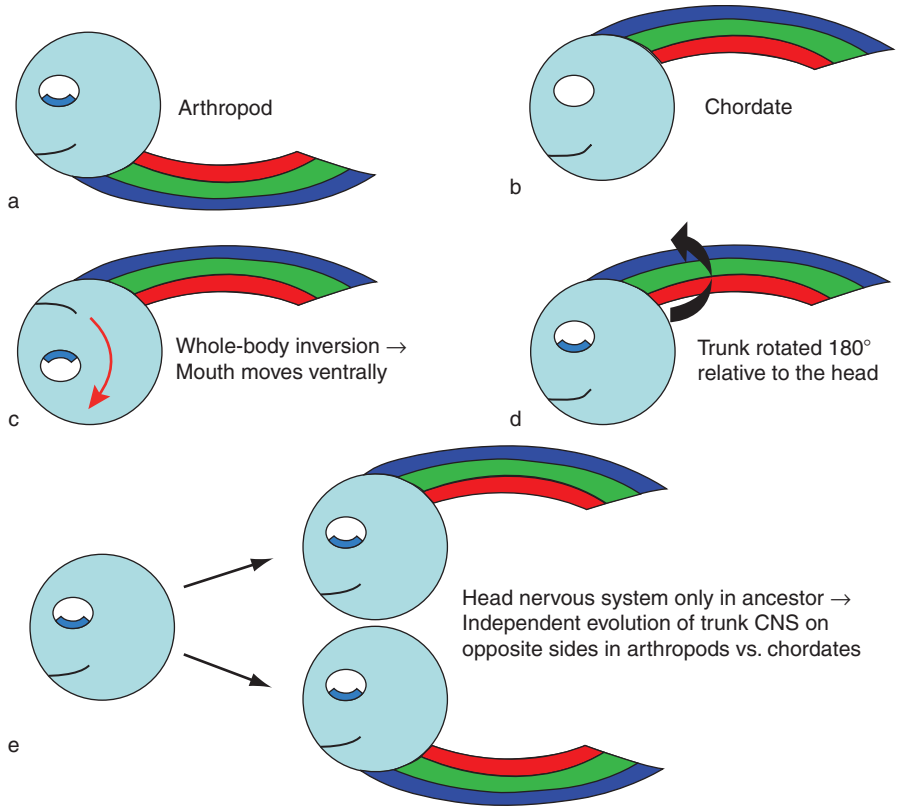

Figure 7 Possible origins of dorsal–ventral (DV) axis inversion in vertebrates: (a) invertebrate DV axis; (b) vertebrate DV axis; (c) Geoffroy St.-Hilaire model; (d) and (e), alternative models. In the early nineteenth century, Geoffroy St.-Hilaire proposed that the DV axis of vertebrates was inverted with respect to that of invertebrates, based on the opposite positions of the nervous system and heart (dorsal in flies pumping anteriorly vs. ventral in vertebrates pumping posteriorly). Several scenarios have been proposed to account for the apparent axis inversion in light of recently obtained molecular data. In the original Geoffroy St.-Hilaire model (c), the entire DV axis was inverted, followed by a ventral migration of the mouth orifice. Alternatively (d), only the trunk region was rotated by 180° with respect to the head, followed by migration of the mouth opening as well. An attractive feature of this hypothesis is that it also explains why the left and right sides of the vertebrate sensory nervous system map primarily to the opposite side of the brain. Another possibility (e) is that the last common bilateral ancestor had only a condensed anterior nervous system (or brain) and that the condensed central nervous system (CNS) trunk later evolved separately (and with opposite DV polarity). This hypothesis does not account for the similar BMP-mediated mechanisms for establishing the conserved neural patterning along the DV axis.

Drosophila, suggests that the order of DV patterning in the brain might be the same in flies and invertebrates. Thus, the relative DV patterns in the head and trunk appear to be opposite. One explanation for the apparent differences in head and trunk patterning is that the anterior brain may have evolved first from an anterior net of cells and then condensed trunk nervous systems developed later and in opposite DV orientations following the split of vertebrates and arthropods (Figure 7(e)). One argument against this model is the shared DV pattern of neural identity gene expression and the dosage-sensitive regulation of these genes by BMPs, which seems difficult to imagine having evolved twice by chance. Another possible explanation is that the inversion of neural pattern was confined to the trunk and that the body was rotated by 180° with respect to the head, which remained in a fixed DV orientation (Figure 7(d)). This hypothesis could also offer a potential explanation for an otherwise puzzling feature of the vertebrate nervous system – that the primary sensory axonal projections cross from left to right (or decussate, in the jargon). In other words, the right hand maps primarily to the left sensory cortex, as does the right eye to the left visual cortex. There is no evidence for an analogous primary cross-representation in invertebrates. For example, eyes project primarily ipsolaterally in all invertebrates examined. It is also possible that apparent differences between the head and trunk reflect a sampling bias and that further analysis of additional conserved gene sets expressed in the head will support the original Geoffroy St.-Hilaire model for the full-body axis inversion. One interesting testable prediction of the head–trunk rotation model is that genes expressed along the entire anterior–posterior (AP) axis of the nervous system in a restricted DV pattern in arthropods might have opposite split DV expression domains in the head versus trunk regions of vertebrates.

Conclusion

BMPs play a similar all-or-none role in repressing the expression of neural genes in epidermal regions in vertebrates and arthropods and then play a dosage-sensitive role in establishing a conserved pattern of neural identity gene expression during early patterning of the neuroectoderm. An important question to resolve is whether BMPs function in vertebrates as they do in flies, by using threshold-dependent repression of neural identity genes in conjunction with ventral-dominant cross-inhibition among neural identity genes. Analysis of the role on BMPs in patterning the nervous systems of other organisms will provide additional information for reconstructing the elements of neural induction present in the common bilateral ancestor. Such broadened evo-devo studies will also reveal how evolutionary novelties arise in specific lineages to give rise to the rich array of neural development and function in diverse organisms.

See also: Basal Ganglia: Internal Organization; Genetic Regulation of Circadian Rhythms in Drosophila; Learning and Memory in Invertebrates: *Drosophila*; Neural Induction in Chicks; Olfaction in Invertebrates: Drosophila; Sleep and Waking in Drosophila.

Further Reading

Arendt D (2005) Genes and homology in nervous system evolution: Comparing gene functions, expression patterns, and cell type molecular fingerprints. *Theory in Biosciences* 124: 185–197.

Arendt D and Nubler-Jung K (1999) Comparison of early nerve cord development in insects and vertebrates. *Development* 126: 2309–2325.

Biehs B, François V, and Bier E (1996). The *Drosophila short gastrulation* gene prevents Dpp signaling from autoactivating and suppressing neurogenesis in the neuroectoderm. *Genes & Development* 10: 2922–2934.

Bier E (2000) *The Coiled Spring: How Life Begins.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

De Robertis EM, Larrain J, Oelgeschlager M, and Wessely O (2000) The establishment of Spemann's organizer and patterning of the vertebrate embryo. *Nature Reviews Genetics* 1: 171–181.

Geoffroy St.-Hilaire E (1822) Considérations générales sur la vertèbre. *Memoirs du Museum National d'Histoire Naturelle* 9: 89–119.

Gilbert SF (2006) *Developmental Biology,* 8th edn. Sunderland, MA: Sinauer Associates.

Harland R (2000) Neural induction. *Current Opinion in Genetics and Development* 10: 357–362.

Kosman D, Mizutani CM, Lemons D, Cox WG, McGinnis W, and Bier E (2004) Multiplex detection of RNA expression in Drosophila embryos. *Science* 305: 846.

Lee KJ and Jessell TM (1999) The specification of dorsal cell fates in the vertebrate central nervous system. *Annual Review of Neuroscience* 22: 261–294.

Lowe CJ, Terasaki M, Wu M, et al. (2006) Dorsoventral patterning in hemichordates: Insights into early chordate evolution. *PLoS Biology* 4: e291 (online).

Mizutani CM, Meyer N, Roelink H, and Bier E (2006) Threshold-dependent BMP-mediated repression: A model for a conserved mechanism that patterns the neuroectoderm. *PLoS Biology* 4: e313 (online).

Raible F and Arendt D (2004) Metazoan evolution: Some animals are more equal than others. *Current Biology* 14: R106–R108.

Sanes DH, Reh TA, and Harris WA (2006) *Development of the Nervous System,* 2nd edn. Amsterdam: Elsevier Academic Press.

Schmidt J, François V, Bier E, and Kimelman D (1995) The *Drosophila short gastrulation* gene induces an ectopic axis in *Xenopus*: Evidence for conserved mechanisms of dorsal-ventral patterning. *Development* 121: 4319–4328.

Srinivasan S, Rashka KE, and Bier E (2002) Creation of a Sog morphogen gradient in the *Drosophila* embryo. *Developmental Cell* 2: 91–101.

Spemann H and Mangold H (1924) Über Induction von Embryonanlagen durch Implantation artfremder Organisatoren. *Wilhelm Roux' Archiv fiir Entwicklungsmechanik der Organismen* 100: 599–638.

Stathopoulos A and Levine M (2002) Dorsal gradient networks in the *Drosophila* embryo. *Developmental Biology* 246: 57–67.

Tessmar-Raible K and Arendt D (2003) Emerging systems: Between vertebrates and arthropods, the Lophotrochozoa. *Current Opinion in Genetics and Development* 13: 331–340.

Wolpert L, Beddington RS, Brockes J, Jessell TM, Lawrence P, and Meyerowitz E (2001) *Principles of Development,* 3rd edn. New York: Oxford University Press.

Relevant Websites

http://www.sdbonline.org – Atlas of *Drosophila* Development and Society for Developmental Biology.

http://www.expasy.ch – ExPASy, Swiss Institute of Bioinformatics. *Caenorhabditis elegans* worm enzymes.

http://flybase.net/ – Flybase.

http://flybrain.neurobio.arizona.edu – FlyBrain.

http://www.informatics.jax.org – Informatics. Mouse genome.

http://www.ncbi.nlm.nih.gov – National Institutes of Health. Zebra fish genome.

http://www3.ncbi.nlm.nih.gov – National Institutes of Health. Online Mendelian Inheritance in Man (OMIM). Human disease genes.

http://sdb.bio.purdue.edu – Purdue University Interactive Fly.

http://genome-www.stanford.edu – Stanford University. *Saccharomyces* yeast genome.

Borderline Personality Disorder

H W Koenigsberg and L J Siever, Mount Sinai School of Medicine, New York, NY, USA; and James J Peters VA Medical Center, Bronx, NY, USA

© 2009 Published by Elsevier Ltd.

Introduction

Borderline personality disorder (BPD) is one of a set of ten discrete personality disorders – disorders affecting the characteristic and habitual ways the individual reacts to emotional stimuli, interpersonal situations, and impulses. At one time, the personality disorders were conceptualized as arising primarily from early life experience, whereas the acute, recurring, or chronic psychiatric syndromes, such as schizophrenia, depression, or bipolar disorder, were seen as having strong biological determinants. A body of work examining personality development and temperament, inheritance of personality traits, and the neurochemistry and regional brain activity in personality disorder patients has led to recognition of the role of neurobiological factors in the personality disorders.

BPD is a prevalent disorder, estimated to occur in 1% or 2% of the population. Patients with BPD are characterized by great emotional reactivity, high degrees of impulsivity, recurring self-destructive behaviors, inappropriate levels of anger, intense and unstable relationships with others, and a sense of emptiness and confusion about their own identities. As a result, their moods shift unpredictably, their relationships are stormy and tumultuous, they often alienate others, and they hurt themselves. They repeatedly threaten or attempt suicide. In the diagnostic classification system of the American Psychiatric Association, the diagnosis is made when five of nine observable features are present (Table 1).

The patterns of feeling and behaving which characterize BPD are likely to arise from the interaction between biological tendencies and developmental life experiences. Understanding the neurobiology of BPD provides a window onto one important determinant of the disorder. It may help in identifying factors that predispose to BPD, in clarifying the relationship between BPD and so-called near-neighbor disorders such as bipolar disorder, and in guiding the development of pharmacologic treatments. This article focuses on the neurobiological aspects of BPD.

Two distinct personality traits, impulsive aggression and affective instability, which appear to have strong biological correlates, co-occur in patients with BPD. Each of these traits has been shown to run separately in families of patients with BPD, and when they coincide in a relative, that individual is likely to meet criteria for a diagnosis of BPD. During the past decade, researchers have begun to tease out some key biological correlates of these personality traits.

The Neurobiology of Impulsive Aggression in BPD

Neurotransmitter Dysregulations

Serotonin Early studies of impulsive and violent suicide attempters identified lower levels of the serotonin metabolite, 5-HIAA, in cerebrospinal fluid. Autopsy studies of such subjects found increased density of serotonin (5-HT) receptors in the frontal cortex compared to nonsuicidal subjects who had died by other violent means. This work suggested that reduced serotonin function might be associated with impulsive aggression in personality disordered patients. Taking advantage of the fact that secretion of the hormone prolactin is stimulated by the action of 5-HT at serotonin receptors in the hypothalamus, the functional state of cerebral serotonin systems may be assayed *in vivo* by challenging the subject with a serotonin agonist and measuring the release of prolactin in response. A series of studies using fenfluramine, a serotonin reuptake inhibitor and releasing agent, found a diminished prolactin response following the fenfluramine challenge in personality disorder patients with a history of impulsive aggression compared to those without, and the degree of blunting in prolactin response was correlated with the extent of impulsive aggression reported by the subjects. A blunted prolactin response to fenfluramine was also observed in BPD patients compared to patients with other personality disorders. Similar findings were observed when more specific serotonin agonists were employed, such as buspirone, a partial agonist at the serotonin 5-HT_{1A} receptor; ipsapirone, a more selective 5-HT_{1A} agonist; and *m*-chlorophenylpiperazine (*m*-CPP), a 5-HT_{2A} and 5-HT_{2C} partial agonist.

These neuroendocrine studies assess the function of serotonin systems in the brain in the region of the hypothalamus. Although this provides a window into serotonergic activity in the brain, the hypothalamus is not directly involved in behavior. Functional neuroimaging studies have made it possible to measure serotonin system activity in regions of the brain that are more directly involved in the control of behavior.

Table 1 *DSM-IV* criteria for borderline personality disorder

A pervasive pattern of instability of interpersonal relationships, self-image, and affects and marked impulsivity beginning by early adulthood and present in a variety of contexts, as indicated by five (or more) of the following:

1. Frantic efforts to avoid real or imagined abandonment. Note: do not include suicidal or self-mutilating behavior covered in criterion 5.
2. A pattern of unstable and intense interpersonal relationships characterized by alternating between extremes of idealization and devaluation.
3. Identity disturbance: markedly and persistently unstable self-image or sense of self.
4. Impulsivity in at least two areas that are potentially self-damaging (e.g., spending, sex, substance abuse, reckless driving, and binge eating). Note: do not include suicidal or self-mutilating behavior covered in criterion 5.
5. Recurrent suicidal behavior, gestures, or threats, or self-mutilating behavior.
6. Affective instability due to marked reactivity of mood (e.g., intense episodic dysphoria, irritability, or anxiety usually lasting a few hours and only rarely more than a few days).
7. Chronic feelings of emptiness.
8. Inappropriate, intense anger or difficulty controlling anger (e.g., frequent displays of temper, constant anger, and recurrent physical fights).
9. Transient stress-related paranoid ideation or severe dissociative symptoms.

From the American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC: American Psychiatric Association.

Other neurotransmitter systems Animal studies have suggested a link between aggression and levels of vasopressin in the central nervous system. A study of personality disorder patients found a positive correlation between levels of arginine vasopressin in the cerebrospinal fluid and a lifetime history of aggression.

Norepinepherine modulates arousal level and vigilance to dangerous or threat conditions. It appears to be associated with irritability and anger and may be associated with outwardly directed aggression, but there is little evidence associating norepinepherine with impulsive aggression *per se* in BPD.

The Neuroanatomy of Impulsive Aggression

Positron emission tomography (PET) studies, which use radioisotopes to image regional brain metabolic activity, have demonstrated reduced activity in prefrontal regions that inhibit behavior. A PET study, which compared regional brain metabolic activity at baseline and after fenfluramine infusions, showed that patients with a history of aggression or impulsivity had smaller increases in activity after fenfluramine than did healthy control subjects in the orbital prefrontal and adjacent ventromedial prefrontal cortex and the cingulate cortex. A similar study using *m*-CPP as the serotonergic agonist found that patients with high levels of impulsive aggression did not show *m*-CPP activation in the left anteriomedial frontal cortex or the anterior cingulate cortex but showed greater activation in the posterior cingulate cortex compared to healthy controls (Figure 1). The rate of serotonin synthesis in brain regions may be imaged by obtaining a PET scan after administration of radiolabeled α-[^{11}C]methyl-L-tryptophan (α-MTrp). α-MTrp is an analog of the serotonin precursor L-tryptophan. It is not incorporated into the protein but is trapped in the neuron as serotonin synthesis occurs. Its irreversible trapping provides an index of serotonin synthesis. Male and female BPD patients show decreased α-MTrp trapping in the anterior cingulate gyrus and in the superior temporal gyrus compared to healthy controls. In addition, both male and female borderline patients showed a negative correlation between their scores on a laboratory measure of impulsivity and α-MTrp trapping in the medial frontal gyrus, anterior cingulate gyrus, temporal cortex, and striatum. A PET study utilizing the serotonin transporter radiotracer, [^{11}C]McN5652, demonstrated decreased transporter availability in the anterior cingulate cortex in personality disorder patients with impulsive aggression, most of whom met criteria for BPD. Taken together, these studies provide convincing evidence that in impulsive aggressive BPD patients there is a decreased responsiveness of serotoninergic neurons located in regions of the cerebral cortex that provide inhibitory control of behavior.

Genetic Contributions

Although family studies, including studies comparing concordance rates of BPD among monozygotic and dizygotic twins, suggest that genetic factors play a role in borderline personality disorder, identifying the genetic polymorphisms specific to personality disorders is a difficult undertaking because personality disorders may arise only when several genes are affected simultaneously, or there may be several different genetic pathways leading to the same personality disturbance. In addition, treating the personality disorders as phenotypes may yield phenotypes that are too heterogeneous to be associated with specific gene polymorphisms. Despite these limitations, there is

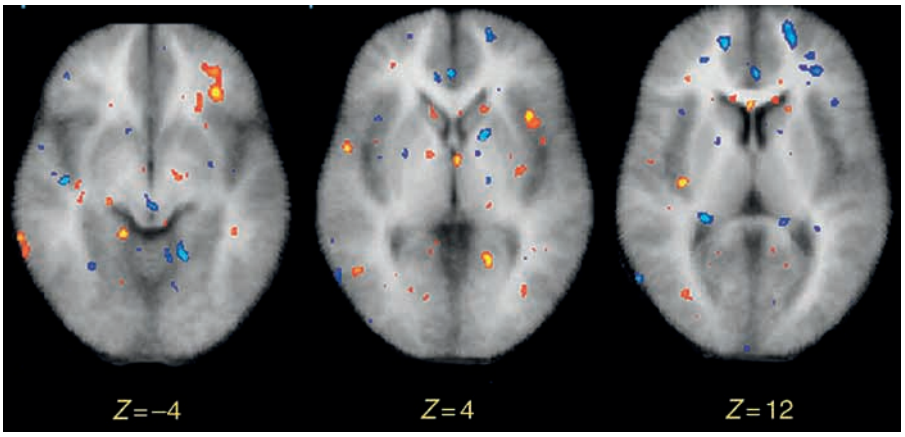

Figure 1 Horizontal slices of the brain showing regions with reduced (blue) and increased (red) metabolic activity after the administration of *m*-CPP in BPD patients compared to healthy controls. *Z* represents Talairach *z* coordinate of slice. Note the decreases in activity in the orbital frontal cortex in the patients compared to controls. Adapted from New AS, Hazlett EA, Buchsbaum MS, et al. (2002) Blunted prefrontal cortical ^{18}fluorodeoxyglucose positron emission tomography response to meta-chlorophenylpiperazine in impulsive aggression. *Archives of General Psychiatry* 59: 621–629.

suggestive evidence of associations between certain specific genes and borderline personality disorder.

Because of the role of the serotonin system in impulsive aggression, efforts to find genetic correlates of impulsive aggression have centered on candidate genes known to code for proteins involved at various stages of serotonin neurotransmission. Mice in which the gene that codes for the 5HT1B receptor has been deleted are particularly aggressive. In humans, the G allele of the 5HT1B gene has been found to be associated with a history of suicidality in a sample of Caucasian subjects with personality disorders. In another study, one particular allele of the 5-HT2A receptor gene was associated with self-mutilation in a group of personality disorder subjects and with suicide in a subgroup of males within that sample. Polymorphisms of a gene that codes for tryptophan hydroxylase, the enzyme controlling the rate-limiting step in serotonin synthesis, have also been examined for associations with aggression, but different studies have found conflicting results. Further work needs to be done to identify the genetic correlates of BPD. One promising strategy is the identification of intermediate phenotypes associated with BPD – phenotypes that may be more closely related to biological attributes of BPD than more global phenotypes such as a diagnosis of BPD or even the impulsive aggressive or affective instability phenotypes.

The Neurobiology of Affective Instability in BPD

Affective instability, the vulnerability to intense and rapid shifts in affect in response to psychosocial cues, is a defining feature of BPD and is associated with the intense anger, identity disturbance, and suicidality of borderline patients.

Neurotransmitter Dysregulations

Cholinergic systems A diverse body of evidence has implicated cholinergic systems in mood regulation. When administered to patients with depression, arecholine, an agonist at cholinergic receptors, induces increased levels of depression, anger, and hostility. The administration of physostigmine, an agent which increases cholinergic activity by inhibiting the cholinesterase enzyme, has been shown to induce depression in depressed and manic bipolar patients. Sensitivity to physostimine in healthy males, as indexed by their behavioral and cardiovascular response, correlated with the subjects' reported levels of affective lability, irritability, feelings of stress, and low life contentment. In BPD patients, infusions of physostimine induced a depressed mood, whereas it did not do so in healthy controls. Procaine, a cholinergic agonist which activates limbic structures including the amygdala, has been shown to produce a high level of dysphoria in BPD subjects. Finally, REM latency, which is under cholinergic control, is shortened and more variable in BPD patients than in healthy controls. These observations suggest that a heightened sensitivity to cholinergic activity may be associated with affective instability in BPD.

Other neurotransmitter systems The noradrenergic system modulates arousal level and vigilance to the environment. Increased noradrenergic activity could explain the heightened reactivity to environmental

stimuli seen in borderline patients. Dextroamphetamine, a releaser and reuptake inhibitor of norepinepherine as well as of dopamine, induced a level of dysphoria in a group of healthy subjects which correlated with the degree of affective lability present. A more specific assay of the functional activity of noradrenergic systems can be assessed by administering clonidine, an α_2 agonist, to subjects and measuring the level of growth hormone (GH) secretion induced, since growth hormone release is mediated by postsynaptic α_2 receptors. In patients with BPD as well as in healthy controls, but not in depressed patients, clonidine-induced GH levels correlated with subjects' ratings of irritability.

Gamma-aminobutyric acid (GABA) acts as an inhibitory neurotransmitter in the central nervous system. The amygdala, which plays a key role in the evaluation and expression of emotion, is rich in GABA receptors. These observations raise the possibility that low levels of GABA activity may be associated with affective instability. This hypothesis receives additional support from the observation that three mood stabilizers – lithium, valproate, and carbamezepine, which can reduce affective instability in BPD patients – appear to increase GABAminergic activity. However, few studies of the GABAminergic system in BPD have been carried out, so the involvement of GABA dysregulation in BPD remains conjectural.

The Neuroanatomy of Affective Instability

The amygdala plays a central role in assessing the emotional salience of stimuli and in mediating the emotional response to those stimuli. Its connections to the hippocampus enable explicit memories to trigger emotional reactions as well. These two structures may be implicated in the affective instability in BPD. Several structural imaging studies of BPD have identified decreased amygdalar and hippocampal volumes. In addition, one study found a decrease in anterior and posterior cingulate gyrus gray matter in BPD – a finding of particular interest because of the role of the cingulate in regulating emotion and impulsive action. Two studies employing functional magnetic resonance imaging to measure regional brain activity found greater activation in the amygdala and fusiform cortices in BPD patients with affective instability than in healthy controls as they viewed pictures that elicited negative emotions. These studies suggest that amygdalar dysregulation may be associated with the affective instability of BPD. It is also possible that structural or functional disturbances of brain regions upstream of the amygdala may be contributory. For example, dysfunctions of the orbital frontal cortex, a region which has inhibitory control over the amygdala, may play a role in affective instability. A pilot study of the voluntary suppression of emotional reactions to negative pictures suggested that BPD patients were less able than healthy control subjects to activate the rostral and pregenual cingulate cortex – regions involved in the regulation of emotion.

Summary

There is increasing evidence that two distinct traits which co-occur in BPD, impulsive aggression and affective instability, are associated with particular neurobiological dysregulations. Decreased serotonin activity in the orbital frontal cortex has been found in BPD patients with high impulsive aggression. Increased levels of arginine vasopressin or norepinepherine may also be implicated in impulsive aggression in BPD. Affective instability, although less well understood, may involve disturbances in cholinergic neurotransmission as well as decreased amygdalar and hippocampal volumes. Dysregulation in norepinepherine or GABA mechanisms may play a role in affective instability in BPD. The finding of decreased serotonin activity provides a neurobiologic rationale for the use of selective serotonin uptake inhibitors to reduce impulsive aggression in BPD, a treatment shown to be clinically effective. A promising next step in research on the neurobiology of BPD would be the identification of clinically valid biological intermediate phenotypes of BPD. Such intermediate phenotypes may help in identifying genetic correlates of BPD. In addition, functional imaging of BPD patients as they process affective stimuli should help us to better understand the mechanisms of affective instability in BPD. As the neurobiological correlates of BPD are better understood, we will be better able to formulate developmental models of the disorder which integrate biological and environmental factors.

See also: Aggression: Neurochemical and Molecular Mechanisms; Amygdala: Contributions to Fear; Anxiety Disorders; Cholinergic Pathways in CNS; Gamma-Aminobutyric Acid (GABA); Growth Hormone ; Hippocampus; Norepinephrine: Adrenergic Receptors; Positron Emission Tomography (PET); Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology.

Further Reading

Gurvits IG, Koenigsberg HW, and Siever LJ (2000) Neurotransmitter dysfunction in patients with borderline personality disorder. *Psychiatric Clinics of North America* 23: 27–40.

Koenigsberg HW, Harvey PD, Mitropoulou V, et al. (2002) Characterizing affective instability in borderline personality disorder. *American Journal of Psychiatry* 159(5): 784–788.

New AS, Hazlett EA, Buchsbaum MS, et al. (2002) Blunted prefrontal cortical ^{18}fluorodeoxyglucose positron emission tomography response to meta-chlorophenylpiperazine in impulsive aggression. *Archives of General Psychiatry* 59: 621–629.

Siever LJ, Buchsbaum MS, New AS, et al. (1999) D,L-fenfluramine response in impulsive personality disorder assessed with [^{18}F] fluorodeoxyglucose positron emission tomography. *Neuropsychopharmacology* 20: 413–423.

Siever LJ and Davis KL (1991) A psychobiological perspective on the personality disorders. *American Journal of Psychiatry* 148: 1647–1658.

Siever LJ, Koenigsberg HW, and Reynolds DA (2003) Neurobiology of personality disorders: Implications for a neurodevelopmental model. In: Cicchetti D and Walker E (eds.) *Neurodevelopmental Mechanisms in the Genesis and Epigenesis of Psychopathology.* Cambridge, UK: Cambridge University Press.

Siever LJ, Torgersen S, Gunderson JG, Livesley WJ, and Kendler#,32] KS (2002) The borderline diagnosis III: Identifying endophenotypes for genetic studies. *Biological Psychiatry* 51: 964–968.

Silverman JM, Pinkham L, Horvath TB, et al. (1991) Affective and impulsive personality disorder traits in the relatives of patients with borderline personality disorder. *American Journal of Psychiatry* 148: 1378–1385.

Skodol AE, Siever LJ, Livesley WJ, Gunderson JG, Pfohl B, and Widiger TA (2002) The borderline diagnosis II: Biology, genetics, and clinical course. *Biological Psychiatry* 51: 951–963.

The American Psychiatric Association (1994) *Dragnostick and Statistical Manual of Mental Disorders,* 4th edn. Washington, DC: American Psychiatric Association

Botulinum and Tetanus Toxins

T Lang, University of Bonn, Bonn, Germany
R Jahn, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

© 2009 Published by Elsevier Ltd.

Introduction

Bacteria of the genus *Clostridium* produce so-called clostridial neurotoxins (CNTs) that represent nature's most toxic proteins. They include seven botulinum neurotoxins (BoNTs; indicated with letters from A to G) and tetanus toxin (TeNT). CNTs are composed of three functional protein domains, which reflect the multistep mechanism by which they block neurotransmitter release. The three domains include a binding and a translocation domain, which are linked by a disulfide bond to an enzymatic domain with zinc-endopeptidase activity. Once released into the bloodstream, the binding domain binds with high specificity and affinity to presynaptic membranes of neuromuscular junctions. Neurospecific binding is followed by receptor-mediated endocytosis into acidifying vesicular compartments, and acidification triggers the translocation domain to expose the enzymatic domain into the reductive milieu of the cytosol in the nerve terminal. The disulfide bond is cleaved, leading to activation of the catalytic domain. In the cytosol of the nerve terminal, block of neurotransmitter release is achieved by cleavage, and thereby interfering with the function, of so-called SNARE proteins that represent the core of the synaptic vesicle fusion engine. Whereas botulinum toxins directly block the invaded nerve terminal causing muscle paralysis (botulism), the intracellular pathway of TeNT is different: after endocytosis, TeNT-containing vesicles are retrogradely transported to the dendrites of the motor neuron. Here, the toxin is released, crosses the synaptic cleft, and enters the nerve terminals of inhibitory interneurons. As a result, motor neurons are hyperexcited, resulting in muscle cramps (tetanus). Since approval by the US Food and Drug Administration in 1989, botulinum neurotoxin A has been proven to be an effective drug for treating muscle hyperactivity and spasticity disorders. It is also widely used in cosmetics to eliminate wrinkles.

Clostridial Neurotoxins

The genus *Clostridium* comprises a group of diverse gram-positive, anaerobic eubacteria that can form endospores resistant to heat and chemicals. Some strains are extremely prolific in producing neurotoxins. They include *Clostridium tetani*, which produces TeNT (causing tetanus), and certain strains of *Clostridium botulinum*, *Clostridium barati*, and *Clostridium butyricum*, which produce the seven known BoNTs, termed BoNT/A, BoNT/B, BoNT/C1, BoNT/D, BoNT/E, BoNT/F, and BoNT/G (causing botulism). Apart from neurotoxins, some *Clostridae* produce toxins causing severe infections such as gas gangrene. Animals that are killed by the toxins serve as anaerobic fermenters for the strictly anaerobic bacteria, a fact that is thought to constitute the main evolutionary advantage of the toxins for the bacteria. When the nutrients of the cadavers are consumed, *Clostridia* sporulate and the spores are dispersed into the environment. In the case of *C. tetani*, spores germinate in deep wounds or skin ruptures that are sealed from air exposure, followed by toxin production in the bacterial cytosol. The toxin is released into the bloodstream by autolysis of the bacteria and then reaches the presynaptic terminals of motor neurons. In contrast, botulism is caused only rarely by wound infection with spores (wound botulism). Instead, the major route of poisoning is consumption of food contaminated with spores of BoNT-producing bacteria. After ingestion, BoNTs evade proteolysis and denaturation in the digestive tract. This is due to the fact that unlike TeNT, BoNTs are secreted in complex with other proteins (some with hemagglutinating activity) that presumably protect the toxin during the passage of the stomach. In the intestine, the slightly alkaline pH causes dissociation of the toxin complexes and the free BoNTs enter the blood vessels through transcytosis. Finally, similar to TeNT, they bind to nerve terminals of motor neurons.

CNTs Domain Composition

All eight CNTs are homologous proteins with similar domain structure that are probably derived from a common ancestor. CNTs are composed of three functionally distinct domains, each of which performs a particular task during intoxication. Toxins are synthesized as single-chain polypeptides of approximately 150 kDa. The polypeptides are posttranslationally cleaved by a single proteolytic cut to yield dichain molecules (Figure 1) composed of a 50 kDa light chain (LC) and a 100 kDa heavy chain (HC) linked by a disulfide bond. The LC carries the catalytic activity that leads to block of neurotransmitter release, and the HC is required for binding, internalization, and transfer of the LC into the neuronal cytosol. This 'division of labor' between a catalytically active polypeptide and a separate domain that mediates internalization is shared with other bacterial toxins, such as diphtheria toxin, cholera toxin, and anthrax toxin,

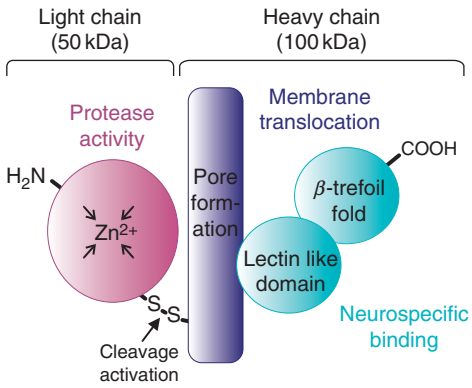

Figure 1 CNT domain composition. CNTs are composed of three major functional domains which mediate neurospecific binding to presynaptic membranes of motor neurons (cyan), membrane translocation from endocytic compartments into the cytosol of the nerve ending (dark purple), and proteolysis of proteins driving synaptic vesicle fusion (pink). The proteolytic domain is located on the 50 kDa light chain that is C-terminally attached via a disulfide bond to the 100 kDa heavy chain. The domain has a zinc-endopeptidase activity that is activated upon cleavage of the disulfide bond in the cytosol of the nerve terminal. The heavy chain contains the domains for neurospecific binding and for translocation from acidifying endocytic compartments into the cytosol of the nerve ending (in the case of TeTx, endocytic compartments migrate to the neuron dendrites where the toxins transsynaptically intoxicate interneurons). The binding domain can be further subdivided in two domains, each providing a separate binding site.

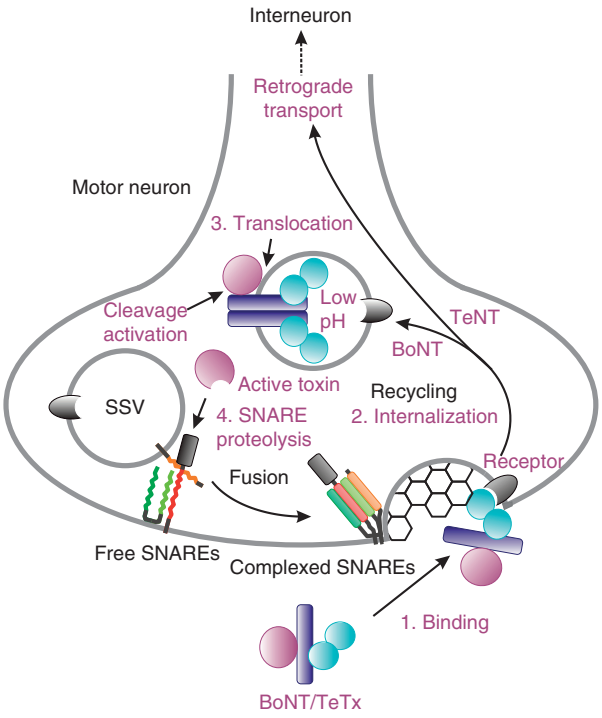

Figure 2 Mechanism of CNT intoxication. The CNT binding domain (cyan) interacts with corresponding receptors at the surface of presynaptic neurons at neuromuscular junctions. Binding is followed by endocytosis involving clathrin-coated vesicles. BoNT-containing vesicles remain in the nerve terminal, whereas vesicles with TeNT are retrogradely transported to the dendrites where they enter transsynaptically inhibitory interneurons. In any case, the endocytosed vesicular structures acidify, triggering membrane insertion of the H_NC domain (dark purple) accompanied by translocation of the light chain (LC; pink) across the membrane. The reductive milieu of the neuronal cytosol reduces the disulfide bond connecting the light and the heavy chain, activating the cleavage activity of the LC. The active toxin can diffuse to its targets (SNARE proteins) localized at the plasma or synaptic vesicle membrane. Proteolysis of SNAREs inhibits synaptic vesicle fusion. The LCs of different CNTs cleave the three different SNAREs involved in neuronal exocytosis at different positions within the SNARE motif.

each of which causes cell damage or cell death by a different enzymatic mechanism.

The HC is composed of two functionally different domains of approximately equal size. The C-terminal domain (H_CC) contains two subdomains and mediates highly selective binding to synaptic plasma membranes. Binding is followed by endocytotic uptake of the toxin, resulting in internalization. Upon acidification of the endocytotic organelle, the N-terminal domain of HC (H_NC) mediates translocation of the catalytic domain into the synaptic cytoplasm. The N-terminally localized LC possesses a catalytic center with a zinc-dependent endopeptidase activity. The zinc ion is coordinated by means of side chain interactions of a conserved zinc-binding HExxH motif that is characteristic for zinc proteases. The dichain molecule shows no enzymatic activity. For activation, the disulfide bond connecting the LC with the HC needs to be reduced, which occurs in the reductive milieu of the presynaptic cytoplasm.

Mechanisms of Intoxication

The mechanisms of synapse invasion and of intoxication are very similar for all CNTs, although the intracellular routes differ between TeNT and the botulinal toxins. For the BoNTs, the steps include: (1) neurospecific binding to the surface of presynaptic motor neurons, (2) internalization into acidifying endocytic compartments, (3) translocation from these compartments into neuronal cytosol, and (4) proteolytic activity in the nerve terminal that causes block of neurotransmitter release (Figure 2). For TeNT, additional steps are involved between steps 2 and 3. Despite the overall similarities in the intoxication mechanism, there are differences between the individual CNTs with respect to surface receptors and the intracellular proteolytic targets of the LCs.

Binding

TeNT and BoNTs bind to the surface of peripheral nerve terminals by means of their respective H_CC domains. Binding occurs at femtomolar concentrations and with very high specificity. Both

polysialogangliosides and specific neuronal surface proteins serve as receptors for toxin binding, but there are differences between the individual toxins. In the H_CC domain of TeNT, two different polysialoganglioside binding sites have been identified. By contrast, BoNT/A and BoNT/B have only a single high-affinity polysialoganglioside binding site that, in addition, differs from those present in TeNT. Furthermore, it is becoming clear that the proteins involved as toxin coreceptors are different. TeNT binds to the GPI-anchored surface protein Thy-1, but due to the high-affinity interaction with gangliosides, efficient invasion does not depend on the coreceptor. In contrast, luminal domains of synaptic vesicles have been invoked as coreceptors for BoNTs, including synaptotagmins I and II for BoNT/B and-G and SV2 for BoNT/A. Hence, a combined receptor model has been proposed for BoNTs which involves a low- and a high-affinity binding site provided by gangliosides and synaptic vesicle proteins, respectively. As part of the synaptic vesicle membrane, the surface exposure of both synaptotagmin and SV2 is increased during synaptic activity, in line with the observation that BoNTs preferentially bind to the surface of active neurons.

Internalization

Once bound to the surface of peripheral nerve terminals, CNTs are internalized by clathrin-mediated endocytosis into acidifying vesicular compartments. For this process, the H_CC domains appear to be sufficient. For some CNTs, it has been shown that high-frequency stimulation of neurotransmitter release increases toxin internalization, indicating that CNTs exploit the endocytotic pathway of synaptic vesicles. This notion is supported by the fact that both synaptotagmins and SV2s serving as coreceptors for BoNTs are synaptic vesicle proteins. However, the exact nature of the toxin-carrying endocytotic compartments and their relation to early and recycling endosomes and to recycling synaptic vesicles is unclear. Furthermore, the reasons for the different intracellular pathways taken by BoNTs and TeNT have remained elusive. BoNTs largely remain within the nerve terminal they invade, a feature that has been exploited in their clinical use as safe drugs for blocking muscle excitation in a highly localized fashion. In contrast, TeNT-containing vesicles are transported retrogradely along the axon toward the dendrites of the motor neurons. Here, the toxin is released by postsynaptic exocytosis (transcytosis) and subsequently endocytosed by the nerve terminals of glycinergic interneurons that are then inhibited. What is the reason for this difference? It is possible that TeNT and BoNTs are internalized by different organelles or differentially sorted after endocytosis. Alternatively, the toxins may share a common entry pathway but TeNT then generates a signal that triggers vesicle binding to microtubule-dependent motors involved in retrograde transport, thus separating it from BoNT-containing vesicles that would remain inside the presynaptic terminal.

Translocation

As mentioned previously, the exact nature of CNT uptake compartments is unknown, but there is general consensus that these compartments acidify by the activity of the vacuolar proton pump after internalization. Like other bacterial toxins that translocate enzymatically active domains into the cytoplasm, CNTs undergo a major pH-dependent structural rearrangement that predominantly involves the H_NC domain (translocation domain). As a result, the toxin is converted from a water-soluble conformation into a conformation in which hydrophobic segments are exposed on the surface. It is assumed that in this conformation the toxin can permeate the lipid bilayer of the endocytic compartment. Evidence for a mechanism involving the formation of hydrophilic pores has been provided by studies showing that the H_NC domains of TeNT and BoNTs form ion channels in artificial and biological membranes at low pH. These channels probably operate as protein-translocating channels that transfer the catalytic domain across the vesicle membrane into the cytosol of the nerve terminal. Alternatively, the channel-induced permeability changes could lead to osmotic lysis of the CNT-containing organelles, thereby releasing the toxin into the cytosol of the nerve terminal. In any case, the released toxin is not active unless the catalytic domain refolds in the neutral environment of the cytosol, and the interchain disulfide bond is completely reduced.

Proteolysis

The LCs of CNTs are proteolytic enzymes that are solely responsible for the toxin action in inhibiting neurotransmitter release. Unlike other neurotoxins that function as channel or receptor blockers, the enzymatic nature of CNTs allows for elimination of many targets by a single CNT molecule. It is thus conceivable that a single CNT molecule suffices to inactivate a nerve terminal, which may be one of the reasons why the CNTs are by far the most potent of all known natural toxins.

The molecular mechanism by which the CNTs inactivate neurotransmission is well understood. All CNTs are metalloproteases with an extraordinary substrate specificity. The toxins exclusively cleave

three synaptic membrane proteins termed syntaxin 1, SNAP-25, and synaptobrevin 2 (or VAMP2). These proteins belong to the SNARE superfamily of proteins that function as catalysts of membrane fusion in intracellular vesicle traffic and exocytosis. A characteristic feature of all SNAREs is a conserved stretch of 60–70 amino acids arranged in heptad repeats termed SNARE motif. Syntaxin 1 and synaptobrevin 2 contain a single SNARE motif adjacent to the C-terminal transmembrane domain. In contrast, SNAP-25 lacks a transmembrane domain and contains two SNARE motifs that are separated by a palmitoylated linker. Although many details are unclear, the overall mechanism by which SNAREs mediate membrane fusion is well understood. Prerequisite for fusion is that complementary SNAREs must reside in the membranes destined to fuse. In the synapse, synaptobrevin 2 resides in the membrane of synaptic vesicles, whereas syntaxin 1 and SNAP-25 are in the presynaptic plasma membrane (Figure 3). During exocytosis, the four SNARE motifs provided by the three SNAREs form a tight complex that pulls the opposed membranes together, thus overcoming the energy barrier separating the two membranes and finally leading to membrane fusion.

All CNTs block neurotransmitter release by executing a single proteolytic cut within the SNARE motifs, thereby arresting the process of synaptic vesicle fusion. With the exception of BoNT/C1 (which cleaves syntaxin 1 but also has weak activity for SNAP-25), each of the toxins is specific for only one of the three neuronal SNARE proteins: BoNT/A and -E cleave SNAP-25, whereas TeNT and the remaining BoNTs cleave synaptobrevin 2. Considering the common evolutionary origin of the toxin, this diversity is astounding. Even more fascinating is the fact that even the toxins specific for the same substrate cleave at different positions within the SNARE motif (with the exception of TeNT and BoNT/B, which both cleave synaptobrevin at the same site).

What is the reason for this extraordinary specificity which is highly unusual for proteases? The crystal structures of the catalytic chains of BoNTs -A, -B, and -E have revealed that overall folds and catalytic sites are largely identical, with the differences being confined to surface residues. Thus, the structures of the catalytic sites cannot explain substrate specificity. Deletion mutagenesis studies have revealed that only rather large fragments of the SNAREs are cleaved by the toxins, suggesting that substrate recognition requires extended stretches located N-terminal of the scissile bond. The crystal structure of a toxin–substrate complex (BoNT/A–SNAP-25) has been solved, revealing multiple interactions between the substrate and the toxin distal to the cleavage site (Figure 4), in line with a cooperative recognition mechanism.

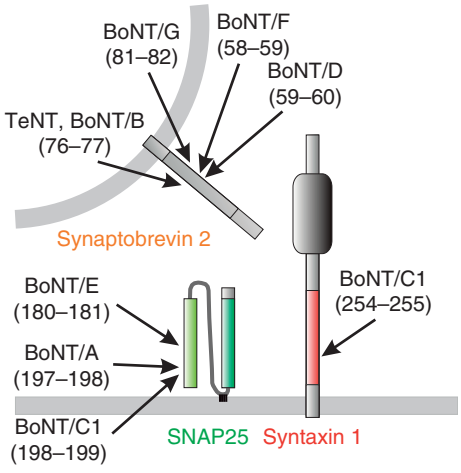

Figure 3 SNARE cleavage by CNTs. All CNTs cleave specifically neuronal SNAREs, including syntaxin 1 and SNAP-25 at the presynaptic membrane and synaptobrevin 2 at the synaptic vesicle membrane. All cleavage sites (numbers indicate amino acid positions) are located within the SNARE motifs, the parts of the SNARE proteins that are involved in the formation of SNARE complexes that pull the opposed membranes closely together, leading to fusion. As a result of CNT action, SNARE complexes cannot form properly, blocking synaptic vesicle fusion and thereby neurotransmitter release.

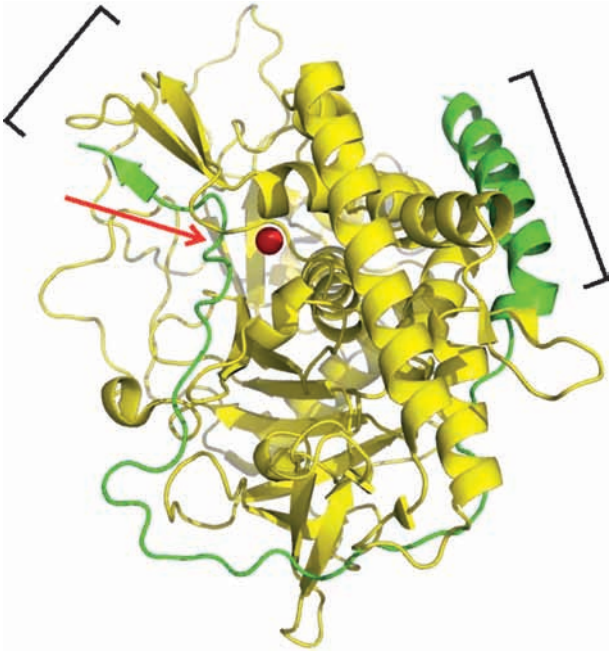

Figure 4 Crystal structure of a CNT–SNARE complex. Crystal structure of BoNT/A-LC (yellow) in complex with the C-terminal SNARE motif of SNAP-25 (green). An array of binding sites distal from the catalytic center wraps the substrate around the light chain and determines specificity (brackets indicate prominent contact points). The catalytic center contains a Zn^{2+} ion (red sphere) and cleaves SNAP-25 (red arrow) close to its C-terminal end.

The requirement of additional binding sites is also consistent with the need for allosteric activation of the toxins' catalytic activity. Hence, the cleavage mechanism involves SNARE binding to the CNTs via the corresponding binding sites on the toxin surface that are distal to the catalytic center, a substrate-induced conformational change activating the catalytic site, the fitting of the SNARE motif into the active site, and finally the cleavage of the substrate. From structural work it is evident that the substrates fit into the active sites only as random coils, which explains why SNAREs are cleavage resistant in SNARE complexes in which they adopt an α-helical conformation. In addition, crystal structures of the holotoxins have revealed that the translocation domains hinder access of the substrate to the cleavage site, in agreement with the observation that chain dissociation mediated by reduction of the interchain disulfide bonds is a prerequisite for activity.

Motor Neuronal Intoxication and Regeneration

The release of neurotransmitter from the nerve terminal is lost a few days after intoxication of a motor neuron by a sublethal dose of BoNTs. Although CNTs efficiently block the function of the neuron, they do not affect any other cell function and they do not cause morphological changes or induce apoptosis. In addition, although the poisoned neuromuscular junction remains structurally intact, some morphological changes at the endplate are observable. Thin neuritic processes emanate from intoxicated nerve endings (terminal sprouting) or from the preterminal node of Ranvier (nodal sprouting), extending and connecting to the muscle fiber. Synaptic activity reappears in the sprouting network and functional neurotransmission recovers. At this point, the parent terminals are still completely blocked. At later times (from a few weeks to several months, depending on the type of BoNTs), the parent terminal recovers and the network of sprouting axon collaterals progressively 'loses' its activity and is retracted.

There are significant differences between the BoNTs with respect to the duration of muscle paralysis. In humans, local injections of BoNT/A, -B, -C, -E, and -F cause neuromuscular paralysis for approximately 4–6, 3, 4 or 5, 1, and 2 months, respectively. Hence, BoNT/A induces a fivefold longer inhibition compared to BoNT/E, although the same SNARE is cleaved. The reasons for these discrepancies are not fully understood. For instance, it has been suggested that the truncated SNAP-25 remaining after BoNT/A cleavage persists in the synapse, thus effectively competing with newly generated intact SNAP-25, whereas the BoNT/E truncated SNAP-25 (which is smaller) may have a shorter half-life. An alternative explanation suggests that differences in the lifetime of the corresponding LCs in the neuronal cytosol may be responsible for the differences in recovery time, with the LC of BoNT/A persisting longer than the LC of BoNT/E. Finally, some observations suggest that the truncated SNAP-25 generated by BoNT/A is replaced more slowly than that generated by BoNT/E, thus increasing the half-life of the nonfunctional SNARE fragment.

Pathologies

Clostridia are not pathogenic, but the toxins they produce cause fatal diseases. The first clinical description of CNT poisoning was documented 25 centuries ago and was caused by TeNT. Hippocrates described the symptoms of a paralyzed patient with hypercontracted skeletal muscles causing spastic paralysis, a condition that he termed tetanus (in Greek, *tetanos* means contraction). Severe intoxication with TeNT usually follows wound infection with spores of *C. tetani*. The violent and persistent spasms affect muscles of the head, trunk, and limbs and can lead to body exhaustion. Death may occur due to respiratory failure or heart failure. Spores of *C. tetani* are widespread and tetanus vaccination is compulsory in many areas of the world. However, tetanus still causes death of thousands of humans annually in countries in which regular vaccination programs are not available.

The first well-documented cases of poisoning by botulinum toxins were described in 1817 by Justinus Kerner. He reported the typical symptoms as a result of an increase in fatal food poisoning occurring in Württemberg, Germany, during the Napoleonic wars. The disease was linked to the consumption of blood sausages and termed botulism (Latin for 'sausage-related condition'). The hallmark of botulism is a generalized muscular weakness, first affecting ocular and throat muscles and later extending to all skeleton muscles. In the more severe forms, a generalized flaccid paralysis develops which ultimately causes death by respiratory failure. Currently, because sterile methods of food preservation are standardized, botulism caused by contaminated food is rare. However, occasional outbreaks occur – for instance, when home-preserved, spore-contaminated vegetables and fish are consumed or when food storage is compromised. In addition, spore-contaminated honey is known to precipitate botulism of infants. So-called infant botulism was discovered in 1976. It is caused by spores that find ideal conditions for germination in the intestines of infants. Generally, artificial respiration

keeps patients suffering from botulism alive, and full recovery occurs once neurotransmission is regenerated. Although botulism is rare in humans, fatal outbreaks of botulism are common in animal populations, particularly among birds and fish.

Pharmacocosmetics Utility

The therapeutic utility of BoNTs is based on their ability to locally block involuntary muscle activity for weeks to months with high specificity and potency. Botulinum toxin A was first approved in 1989 by the US Food and Drug Administration for treatment of three disorders (strabismus, hemifacial, and blepharospasm), and since then the number of indications has increased (including numerous focal dystonias, spasticity, tremors, and migraine and tension headaches). A recent application is the immobilization of damaged joints in preoperative and postoperative care (e.g., to reduce widening of facial scars). Here, preferentially BoNT/E is applied, which causes synaptic blockades of relatively brief duration.

For treatment, highly diluted toxin is directly injected into the hyperactive muscle. One of the primary complications of BoNT/A therapy is the development of an immune response, which counteracts the efficacy of the toxin, particularly in long-term treatment regimens. BoNT/B is used as an effective alternative for patients who no longer respond to BoNT/A. Also, side effects have been described that are caused by diffusion of toxin to neighboring muscles leading to transient and sometimes debilitating ptosis. Another 'side effect' that was first observed by ophthalmologists led to widespread application of the toxins in cosmetics. When BoNTs were used to treat blepharospasm, frown lines and facial wrinkles temporarily disappeared – an observation that resulted in increased general use of the toxins in beauty salons. More than 1 million Americans annually use BoNT/A injections for the elimination of wrinkles, thus comprising the largest market for BoNTs. Finally, the cosmetic application of toxins led to the accidental discovery that local BoNT injections decrease the frequency of migraines in some cases, leading to another field of therapeutic application of BoNTs.

See also: Active Zone; Neuromuscular Junction (NMJ): A Target for Natural and Environmental Toxins in Humans; Neuromuscular Junction Plasticity in Mammals and Botulinum Toxins; Neurotoxins; Neurotoxins and their Neurotoxicology; Presynaptic Endosomes; SNAREs; Synaptic Vesicles; Vesicular Neurotransmitter Transporters.

Further Reading

Breidenbach MA and Brunger AT (2004) Substrate recognition strategy for botulinum neurotoxin serotype A. *Nature* 432: 925–929.

Brunger AT (2005) Structure and function of SNARE and SNARE-interacting proteins. *Quarterly Reviews of Biophysics* 38: 1–47.

Davletov B, Bajohrs M, and Binz T (2005) Beyond BOTOX: Advantages and limitations of individual botulinum neurotoxins. *Trends in Neurosciences* 28: 446–452.

Foran PG, Davletov B, and Meunier FA (2003) Getting muscles moving again after botulinum toxin: Novel therapeutic challenges. *Trends in Molecular Medicine* 9: 291–299.

Grumelli C, Verderio C, Pozzi D, Rossetto O, Montecucco C, and Matteoli M (2005) Internalization and mechanism of action of clostridial toxins in neurons. *Neurotoxicology* 26: 761–767.

Hippe H, Andreesen JR, and Gottschalk G (1999) The genus *Clostridium* – nonmedical. In: Dworkin M (ed.) *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community,* first electronic edn. Berlin: Springer.

Johnson R (1999) Clostridial toxins as therapeutic agents: Benefits of nature's most toxic proteins. *Annual Review of Microbiology* 53: 551–575.

Meunier FA, Schiavo G, and Molgo J (2002) Botulinum neurotoxins: From paralysis to recovery of functional neuromuscular transmission. *Journal of Physiology (Paris)* 96: 105–113.

Rossetto O, Morbiato L, Caccin P, Rigoni M, and Montecucco C (2006) Presynaptic enzymatic neurotoxins. *Journal of Neurochemistry* 97: 1534–1545.

Schiavo G, Matteoli M, and Montecucco C (2000) Neurotoxins affecting neuroexocytosis. *Physiological Reviews* 80: 717–766.

Turton K, Chaddock JA, and Acharya KR (2002) Botulinum and tetanus neurotoxins: Structure, function and therapeutic utility. *Trends in Biochemical Sciences* 27: 552–558.

Brain Adrenergic Neurons

A F Sved and J P Card, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Following initial anatomical descriptions of brain catecholamine neurons, initially based on the Falk–Hillarp fluorescence method and subsequently confirmed with immunocytochemical localization of tyrosine hydroxylase (TH) and dopamine-β-hydroxylase (DβH), it became clear that there was a distinct population of neurons that also contained the enzyme phenylethanolamine-*N*-methyltransferase (PNMT). These PNMT-containing neurons, which have been observed in many mammalian species, contain the other enzymes required for epinephrine (EPI) biosynthesis and have frequently been referred to as adrenergic neurons. In the initial description of brain stem neurons containing PNMT-like immunoreactivity, as well as the other enzymes required for EPI synthesis, these putative adrenergic neurons were shown to exist in three distinct populations in the caudal brain stem. This article focuses on three topics related to these neurons. First, the anatomical location and connections of these neurons are presented, suggesting that these neurons are situated in a manner consistent with their involvement in control of physiological systems. Second, neurochemical aspects of these neurons are considered, leading to the conclusion that it is inappropriate to consider them as 'adrenergic' neurons. Finally, the functional characteristics of these neurons are discussed. Although these PNMT-containing neurons are found in three loose clusters in the medulla, most of the information regarding these neurons is based on one of these cell groups, the C1 cells. Throughout this article, most of the information presented is based on studies in rat, by far the most studied species in this regard, although similar neurons exist in other mammalian species including humans.

Neuroanatomical Organization of Adrenergic Neurons

Location and Architecture of PNMT-Containing Neurons

PNMT-containing neurons are classically grouped according to an alphanumeric nomenclature that defines three groups of neurons – C1, C2, and C3 – spatially segregated within the caudal brain stem. This classification schema follows from a similar convention in which A1–A17 was initially used to define topographically segregated populations of noradrenergic and dopaminergic neurons extending from caudal medulla through the forebrain to the olfactory bulb.

The C1 cell group, in close relation to the noradrenergic A1 cell group, is a component of the lateral tegmental catecholamine cell column in the ventrolateral medulla. As the name implies, the catecholamine neurons that contribute to this column are found within the ventral and lateral portion of the brain stem. The disposition of C1 with respect to the A1 noradrenergic cell group is best appreciated in sagittal sections prepared for dual localization of PNMT and TH (Figure 1). In such preparations, the A1 and C1 cell groups are visible as a continuous column extending from the spinomedullary junction rostrally to the caudal border of the facial motor nucleus. There is little overlap between A1 and C1, with a relatively sharp transition occurring at a coronal plane marking the rostral extent of the dorsal motor nucleus of the vagus nerve and the C1 cell column in rat extending approximately 2 mm longitudinally from that point. The rostral C1 cells form a phenotypically distinct component of the rostroventrolateral medulla (RVLM), a region containing neurons that innervate sympathetic preganglionic neurons in the spinal cord and are involved in cardiovascular regulation. The C1 neurons are distributed throughout RVLM but are present in largest numbers in the rostral portion of the nucleus (Figures 1(b)–1(e)). C1 cells are also found extending caudal to the RVLM. The dendritic arbors of C1 neurons are extensive when viewed in the coronal plane and extend in all directions from the multipolar somata, which approximate 20 μm in widest diameter (Figures 1(c)–1(e)). Horizontal and sagittal sections through the nucleus confirm that the largest radial extent of dendritic fields of individual neurons is in the coronal plane.

The C2 and C3 cell groups are both found in the dorsal aspect of the caudal brain stem (Figures 2 and 3). C2 neurons reside within the nucleus of the solitary tract (NTS), principally at intermediate and rostral levels of this complex cell group (Figure 2(b)). The largest population of C2 neurons is found in the intermediate portion of the NTS that is adjacent to the area postrema (Figures 2(b) and 2(e)). The PNMT immunopositive neurons in this area are approximately 10 μm in widest diameter, densely concentrated in a dorsolateral subdivision of NTS, and give rise to restricted dendritic arbors that are largely confined to the area of cell aggregation. A second group of

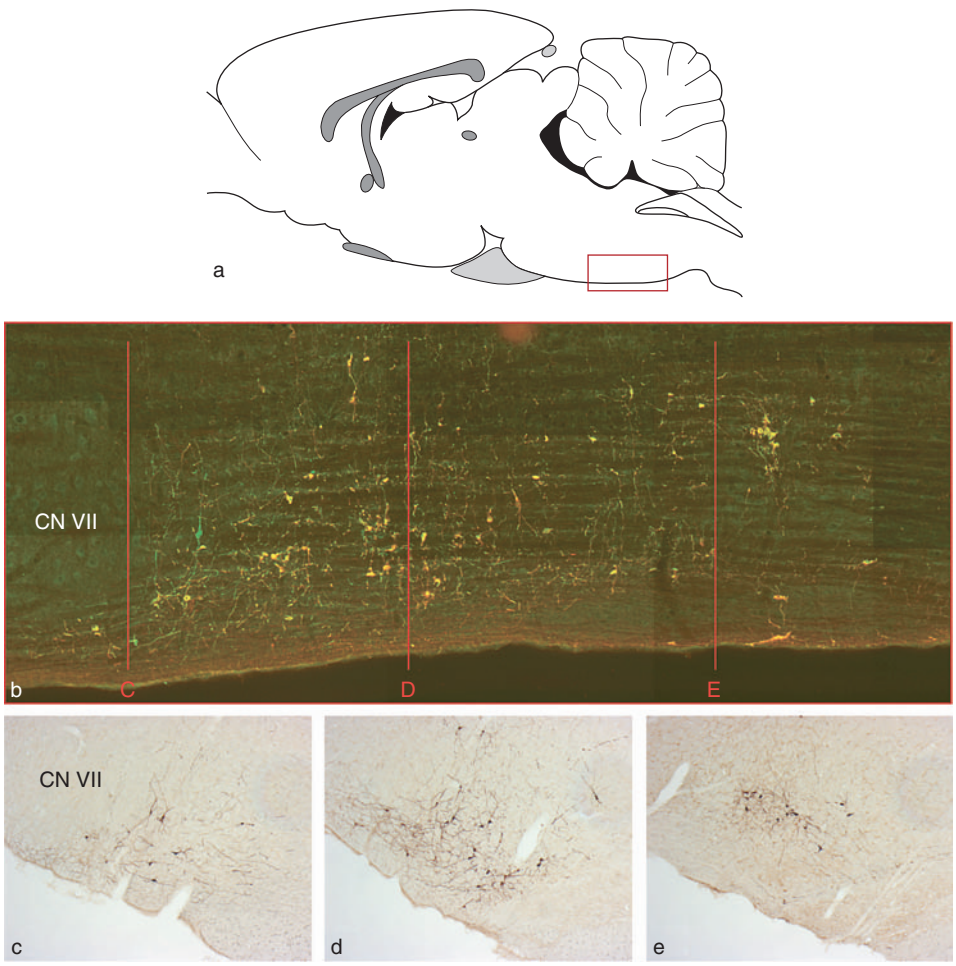

Figure 1 The location and morphology of neurons contributing to the C1 cell group are illustrated. (a) The location of C1 in the ventrolateral medulla is shown by the red boxed area in the sagittal diagram, approximately 2.0 mm lateral from the midline. (b) Dual localization of PNMT and TH in the sagittal plane reveals that C1 neurons co-localizing these enzymes (and thus appearing yellow) extend caudally from the facial nucleus (CN VII) for approximately 2 mm. (c, d) Rostrally, these neurons are largely coextensive with the cardiovascular regulatory nucleus of the rostral ventrolateral medulla. The morphology of these neurons in coronal sections through three representative planes through rostral, intermediate, and caudal levels of C1 is shown in (c)–(e). The approximate location of the coronal planes for each of these figures is indicated in (b). (a) Adapted from the Paxinos and Watson rat brain atlas.

PNMT immunopositive neurons is present in the rostral third of the NTS, rostral to the area postrema (Figures 2(b)–2(d)). Although fewer in number compared to the aforementioned neurons, these neurons are larger in size and give rise to more extensive dendritic arbors. The cell bodies of the most rostrally situated neurons are concentrated along the borders of the NTS and their dendrites project differentially into the NTS (Figures 2(b) and 2(c)). Based on location, morphology, and relative staining intensity of PNMT and TH, the caudal and rostral clusters of C2 neurons might represent distinct anatomical and functional subgroups.

C3 neurons are located medially to the more rostrally situated subset of C2 neurons and extend more rostrally (Figure 3(b)). They form a column in relation to the medial longitudinal fasiculus on either side of the midline. The C3 group is similar to that of C2 in that it is composed of two subgroups of neurons that are distinguished by their location and cytoarchitecture. One subset of C3 neurons forms a compact bundle immediately subjacent to the ependymal lining of the fourth ventricle (Figures 3(b)–3(d)). The perikarya of these neurons approximate 10 μm in widest diameter. In the coronal plane, these neurons appear to give rise to a very limited dendritic arbor (Figure 3(b)), but horizontal sections through this cell group reveal an extensive dendritic arbor with a rostrocaudal orientation (Figure 3(c)). The second subgroup of C3 neurons is characterized by larger multipolar perikarya. These neurons are widely dispersed around and below the medial longitudinal fasiculus and give rise

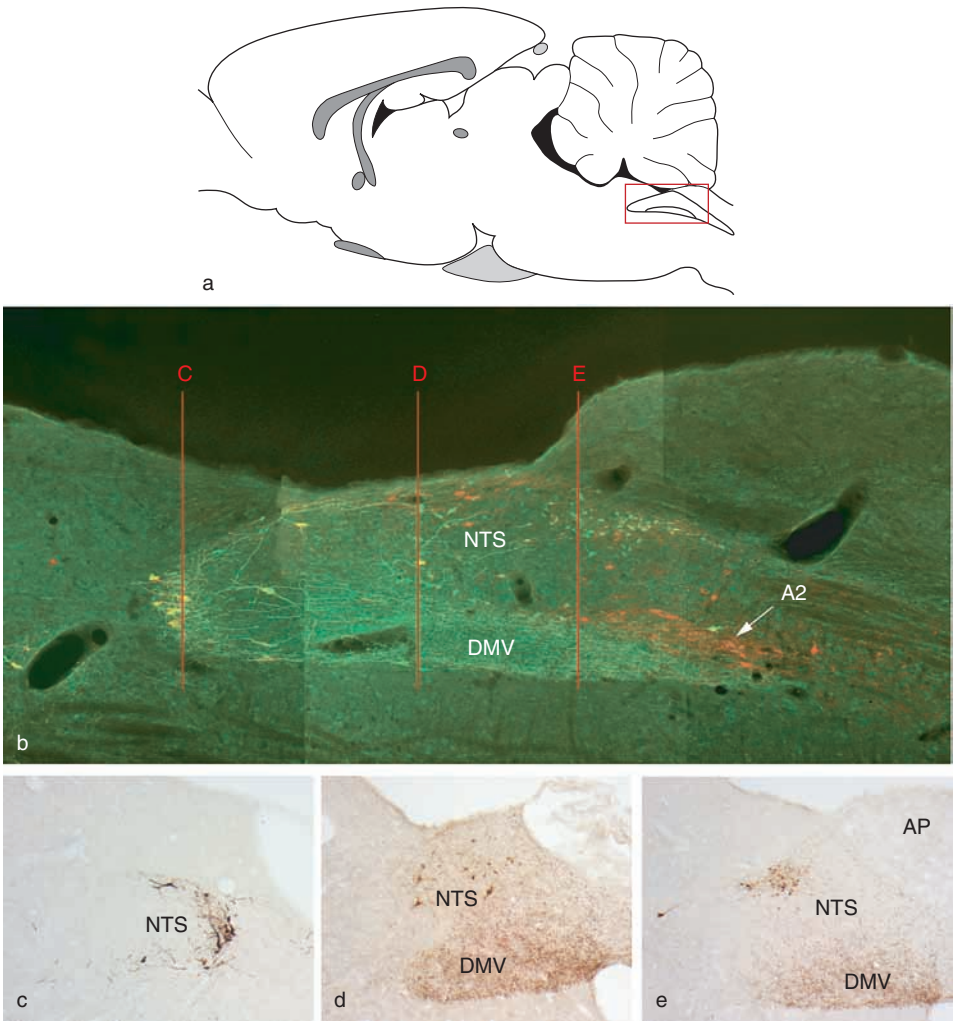

Figure 2 The location and morphology of neurons contributing to the C2 cell group are illustrated. (a) The location of C2 in the caudal and dorsal medulla is shown by the red boxed area in the sagittal diagram, approximately 1.0 mm lateral from the midline. (b) Dual localization of PNMT and TH in the sagittal plane reveals that C2 neurons co-localizing these enzymes (yellow) occupy intermediate and rostral levels of the nucleus of the solitary tract (NTS) and are spatially separate from the A2 noradrenergic cell group (red). The morphology of C2 neurons in coronal section through three representative planes through rostral, intermediate, and caudal levels of NTS is shown in (c)–(e). The approximate location of the coronal planes for each of these figures is indicated in **Figure 3(b)**. (d, e). Note the dense terminal arbor concentrated in the dorsal motor nucleus of the vagus (DMV). (a) Adapted from the Paxinos and Watson rat brain atlas.

to a large dendritic arbor that is most extensive in the coronal plane (**Figure 3(d)**).

Distribution of PNMT-Containing Afferents

Terminal arbors immunopositive for PNMT, although extending through the rostrocaudal extent of the neuraxis, are considerably more restricted than those observed with immunocytochemical localization of TH or DβH. The densest terminal arbors are found within the hypothalamus, medulla, and thoracic spinal cord. Spinal cord afferents are sequestered principally within the column of sympathetic preganglionic neurons. The densest concentrations of PNMT-containing afferents in the medulla are observed in the dorsal motor vagal complex and within the area of C1 neurons. The plexus within the dorsal vagal complex is particularly dense within the motor nucleus of the vagus (**Figures 2(d) and 2(e)**). However, a plexus of moderate density is also present at intermediate and caudal levels of the NTS, whereas the rostral NTS contains only sparse terminals (**Figures 2(c)–2(e)**). The most prominent plexus in the rostral medulla and midbrain is found within the ventrolateral periaqueductal gray matter. Within the hypothalamus, particularly dense terminal arbors are present in the perifornical subdivision of the lateral hypothalamic area, within the parvicellular subdivisions of the paraventricular nucleus, and in the periventricular region surrounding

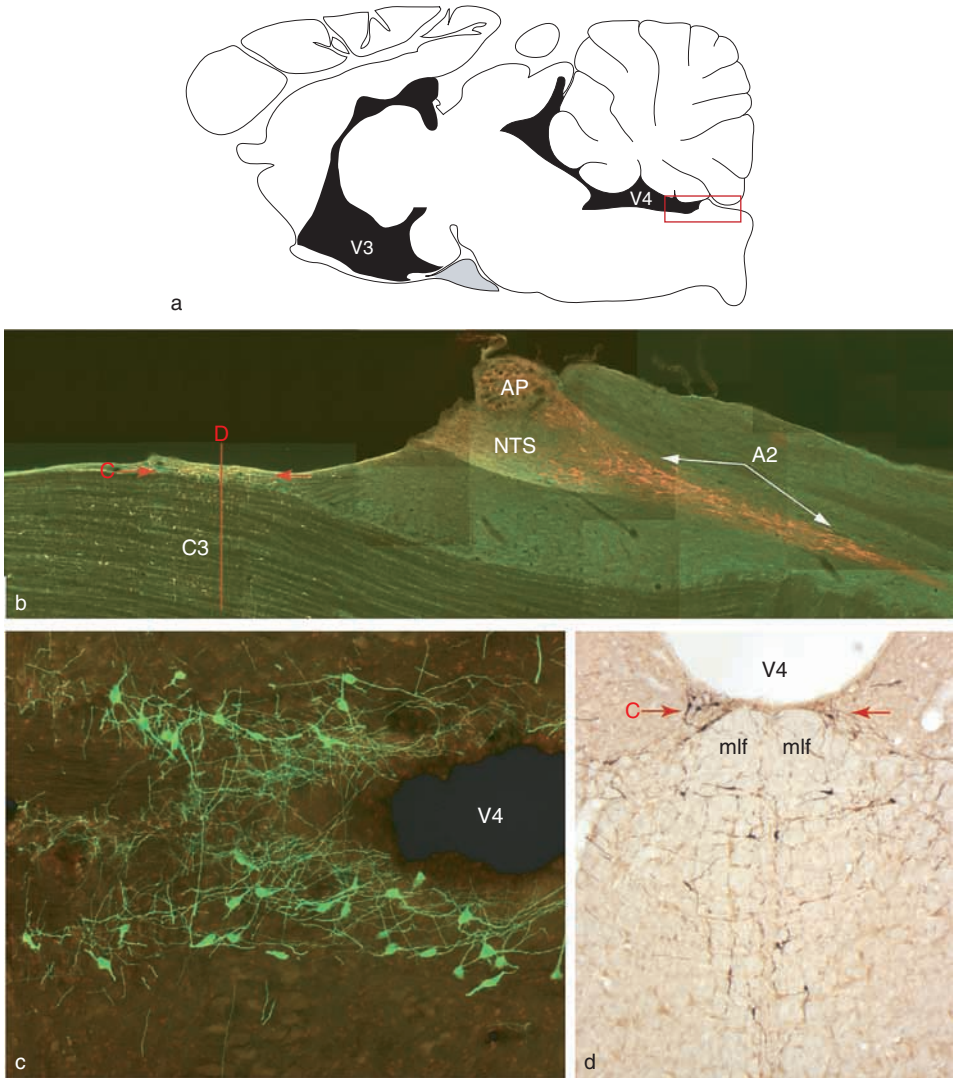

Figure 3 The location and morphology of neurons contributing to the C3 cell group are illustrated. (a) The location of C3 in the dorsal medulla rostral to the dorsal vagal complex is shown by the red boxed area in the sagittal diagram, approximately 0.1 mm lateral from the midline. (b) Dual localization of PNMT and TH in the sagittal plane reveals that C3 neurons co-localizing these enzymes (yellow) are concentrated in two distinct groups that differ in location and morphology. The A2 neurons appear red because they contain TH but not PNMT. (b, d) Neurons immediately subjacent to the ependymal lining of the fourth ventricle form a compact aggregation above the medial longitudinal fasciulus (mlf). When viewed in horizontal section (plane of section marked by red arrows in b and d), the neurons exhibit an extensive dendritic arbor (c; PNMT stained green). The second group of neurons is larger and dispersed more widely in columns parallel to the midline. These neurons are multipolar and give rise to radiating dendrites that are best viewed in the coronal plane (d). (a) Adapted from the Paxinos and Watson rat brain atlas.

the organum vasculosum of the lamina terminalis. Arbors of moderate density are present in the dorsomedial hypothalamic and arcuate nuclei. It is notable that the diencephalon also contains a plexus of moderate density in the thalamic paraventricular nucleus. The densest concentration of fibers in the limbic forebrain is found within the bed nucleus of the stria terminalis. Sparse terminal arbors are present in the central amygdaloid nucleus, the nucleus accumbens, and the anterior olfactory nuclei.

Circuit Organization of the PNMT-Containing Neuronal Cell Groups

The afferent and efferent connections of the PNMT-containing cell groups are not well-defined because the disposition of these cell groups has made them difficult to analyze with conventional tracers. Chief among the difficulties in this regard is the fact that the neurons are interposed among other heterogeneous noncatecholaminergic cell populations. Thus, demonstration of afferents to areas containing

PNMT-containing neurons cannot be equated with synaptic input to those neurons without ultrastructural analysis. This approach has been used to identify afferents to C1 and C2, but it is clear that there is much to be learned regarding the identity of the afferents that regulate the activity of these brain stem neurons.

Retrograde tracing has proven to be more informative in defining the projection targets of brain stem PNMT-containing neurons. It is well established that C1 neurons are a prominent contributor to reticulospinal projections that exert regulatory influence over sympathetic outflow, and it has also become apparent that C2 and C3 contribute to this projection. Additionally, retrograde tracing has documented ascending projections of all three PNMT-containing cell groups to the hypothalamus. Nevertheless, the contribution of the PNMT-containing cell groups to the catecholamine innervation of these projection targets, which is known to also arise from PNMT-containing cell groups, has been impossible to determine with classical anterograde tracers since they do not discriminate among phenotypically distinct populations of neurons. This technical limitation has been circumvented by the development of a lentivirus expression system in which a synthetic DβH promoter controls the expression of a green fluorescent protein reporter gene. Using this phenotypically restricted expression of a reporter that fills axons, the distribution of C1 efferents has been mapped in detail, demonstrating that these neurons project to many central sites involved in autonomic regulation. Given the spatial separation of the C2 and C3 cell populations from other catecholamine cell groups, it should also be possible to use this method to map the distribution of their efferent projections.

Cellular and Molecular Biology of Adrenergic Neurons

Since the time of the initial characterization of the brain stem PNMT-containing neurons, it was clear that they contained the other enzymes required to synthesize EPI (TH, aromatic amino acid decarboxylase, and DβH) and these neurons have often been referred to as adrenergic neurons. However, there is reason to believe that they might not utilize EPI as a neurotransmitter or even synthesize it in appreciable quantities.

The first sign that something might be amiss with the synthesis of EPI in these 'adrenergic neurons' might have been that levels of EPI are not as high as might be expected. For example, in the rat C1 area the levels of norepinephrine are more than an order of magnitude higher than the levels of EPI. This holds for the terminal fields as well, where EPI levels in terminal fields of PNMT-containing neurons are much lower than might be expected based on the density of the PNMT-containing terminals. In the hypothalamus, especially the paraventricular nucleus, there is a substantial input from brain stem PNMT-containing neurons but EPI levels are surprisingly low compared to norepinephrine levels.

Understanding the route and location of EPI synthesis in the hypothalamus is complicated by the existence of PNMT-containing cells that do not contain the other enzymes required for catecholamine synthesis. Nonetheless, EPI in the hypothalamus is dependent on ascending projections from the brain stem; transection of the brain at the level of the colliculi eliminates hypothalamic EPI levels, although it does not appreciably decrease PNMT levels. Furthermore, the catecholamine neurotoxin 6-hydroxydopamine, which does not destroy PNMT-containing catecholamine neurons because they do not express membrane catecholamine transporters, nonetheless eliminates hypothalamic EPI levels. The most straightforward explanation for hypothalamic EPI levels, as initially proposed by Ivan Mefford, is that EPI is synthesized from norepinephrine released in the hypothalamus, which is then N-methylated by the local PNMT-expressing cells with the resulting EPI then being taken up into terminals of classical noradrenergic neurons (Figure 4). This EPI could then be stored and released in a depolarization-induced, calcium-dependent manner, as is the case with sympathetic nerve terminals exposed to high circulating levels of EPI released from the adrenal gland.

In the spinal cord, the situation is somewhat clearer because all of the PNMT activity can be explained by PNMT terminals of PNMT-containing brain stem neurons. In the area of the intermediolateral cell column in the spinal cord, despite an abundant innervation by PNMT-containing nerve terminals, there is essentially no detectable EPI. In contrast, the levels of norepinephrine are at least 1000-fold higher, even though approximately half of the DβH-containing terminals also appear to contain PNMT. (Although some studies have reported low levels of EPI in rat spinal cord, specificity problems with the various assay methods could easily account for this finding.)

If nerve terminals of brain stem PNMT-containing neurons have all the enzymes necessary to synthesize EPI but still do not make EPI, one question that arises is why they do not synthesize EPI. One possibility for why EPI is not synthesized in these cells is that the

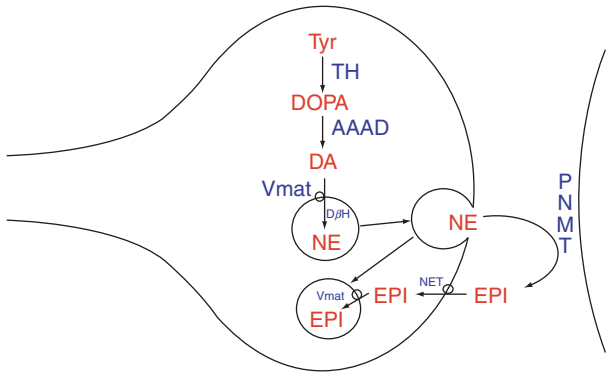

Figure 4 Putative mechanism for the presence of EPI in noradrenergic nerve terminals in hypothalamus. The putative mechanism for EPI synthesis and storage in the hypothalamus involves the synthesis of EPI from norepinephrine in the extracellular fluid, with the EPI then taken up and stored in terminals of noradrenergic neurons. Tyr, tyrosine; DA, dopamine; EPI, epinephrine; PNMT, phenylethanolamine-*N*-methyltransferase; NE, norepinephrine; TH, tyrosine hydroxylase; AAAD, aromatic amino acid decarboxylase; Vmat, vesicular monoamine transporter; DβH, dopamine-β-hydroxylase; NET, norepinephrine transporter.

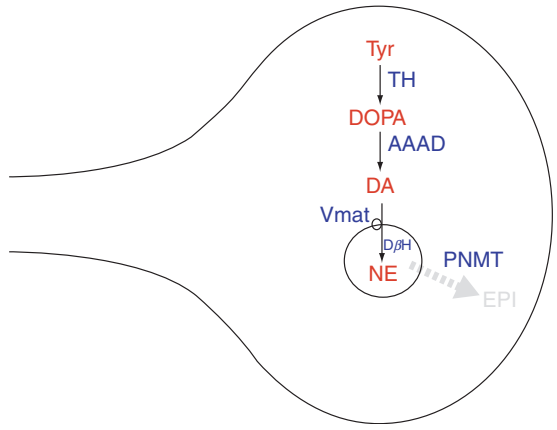

Figure 5 Putative explanation for why EPI synthesis does not occur in PNMT-containing neurons. In this model, PNMT is present in the cytoplasm and norepinephrine, which is synthesized and sequestered in synaptic vesicles, does not again access to PNMT. The gray line represents a step that is not thought to occur. Tyr, tyrosine; DA, dopamine; NE, norepinephrine; EPI, epinephrine; PNMT, phenylethanolamine-*N*-methyltransferase; TH, tyrosine hydroxylase; AAAD, aromatic amino acid decarboxylase; Vmat, vesicular monoamine transporter; DβH, dopamine-β-hydroxylase.

enzymes are compartmentalized in the neuron in such a way that PNMT does not have an opportunity to methylate norepinephrine (Figure 5). In the adrenal medulla, where EPI-synthesizing enzymes are present in great abundance, DβH is an intravesicular enzyme, whereas PNMT is cytoplasmic. This compartmentalization appears to be similar at other sites. This raises the issue of how norepinephrine, which is synthesized from dopamine inside synaptic vesicles, exits the vesicle, is N-methylated by PNMT, and then reenters the vesicle. Although mechanisms exist to transport catecholamines from the cytoplasm into vesicles (i.e., the vesicular monoamine transporter), there is no known transporter that would mediate transport in the opposite direction. This would allow norepinephrine to be made in vesicles but prevent cytoplasmic PNMT from converting it to EPI. This scenario would predict that PNMT-containing nerve terminals would contain norepinephrine, which appears to be the case.

In this scenario, how could the adrenal chromaffin cells synthesize EPI? Surprisingly, viewed in the context of this compartmentalization, it is not clear how EPI could be synthesized even in adrenal chromaffin cells, in which EPI synthesis most certainly does occur. Clearly, there is something about this well-accepted biosynthetic pathway that is yet to be understood. As noted previously, in the hypothalamus, EPI might be synthesized from norepinephrine that is released from vesicles. However, chromaffin cells lack a plasma membrane catecholamine transporter, so the norepinephrine must be generated intracellularly and gain access to the cytoplasm.

A second question that arises as a result of the PNMT-containing neurons not synthesizing EPI is why they have PNMT and the other biosynthetic enzymes. Because there are no data that bear directly on this issue, any suggestion is pure speculation. One possibility is that it is necessary for PNMT to be expressed during development, possibly as a developmental cue. In this regard, it is interesting to note that in rats (and presumably other species) PNMT-containing neurons are vastly overexpressed early in development and then are pruned to adult numbers during early postnatal maturation. Another possibility is that PNMT expression is regulated along with certain other proteins. It is interesting to note that PNMT-expressing catecholamine cells (both adrenal chromaffin cells and central nervous system neurons) lack a plasma membrane catecholamine transporter, suggesting that the presence of one might be a regulatory corequisite of the absence of the other.

As noted previously, the PNMT-containing catecholamine neurons might synthesize and store norepinephrine. Furthermore, this stored norepinephrine might be releasable upon stimulation, as suggested by a decrease in norepinephrine levels in the thoracic intermediolateral cell column in response to high-frequency electrical stimulation of the C1 area. The release of norepinephrine along with the absence of a plasma membrane catecholamine transporter in these neurons suggest that norepinephrine released from these neurons might have a longer duration action and a wider area of influence than with more classical noradrenergic neurons. However, the neurotransmitter phenotype of

these neurons is not limited to a catecholamine. The PNMT-containing catecholamine neurons all express the vesicular glutamate transporter VGLUT-2, suggesting that they are glutamatergic. Furthermore, a variety of neuropeptides, most notably neuropeptide Y, also exist in subpopulations of these PNMT-containing neurons.

Functional Aspects of Adrenergic Neurons

Functions of PNMT-containing neurons are not well elucidated, in part because of a lack of tools to study them. Initial attempts to clarify the actions of these neurons were based on the use of drugs to inhibit PNMT. However, two significant issues hampered such studies. First, all of the early PNMT inhibitors had additional potent effects, most notably blockade of α-adrenergic receptors. Second, although these drugs might block the conversion of norepinephrine to EPI, the neurons could continue to release norepinephrine, and this switch from EPI to norepinephrine may not have major consequences for the receptor. Based on the understanding that these neurons might not synthesize EPI, this approach is even less useful. The PNMT-containing cell groups are not discrete homogeneous populations, interfering with experimental approaches to manipulate these cells. Furthermore, the classical catecholamine neurotoxin 6-hydroxydopamine does not destroy these neurons because they lack a plasma membrane catecholamine transporter.

Three approaches have allowed the study of the functional involvement of these neurons. First, the anatomy of the PNMT-containing neurons has allowed suggestions of their functions. For example, C1 cells projecting to the region of sympathetic preganglionic neurons in the spinal cord are located in a region of the ventrolateral medulla involved in cardiovascular regulation. Thus, the jump to suggesting that these C1 neurons might be involved in cardiovascular regulation is a small one. Functional suggestions based on neuroanatomy received support from the use of expression of early response genes, such as c-fos, as anatomical markers of neurons activated by various stimuli. Consistent with a role of the C1 cells in cardiovascular regulation, these neurons express Fos in response to hypotension. Second, electrophysiological characterization of the response properties of the neurons helps to clarify their function. With the development of juxtacellular labeling to allow phenotypic characterization of the recorded neuron, it became possible to describe the firing properties of these neurons in response to a variety of stimuli. Finally, studies have used an immunotoxin consisting of an antibody to DβH conjugated to saporin to selectively destroy DβH-containing neurons, including those that also contain PNMT; injecting this toxin into the area of the C1 neurons selectively destroys these cells.

Using these three approaches, a functional role of C1 cells is beginning to emerge, but the more sparse C2 and C3 groups remain relatively unstudied. Here, we provide a brief review of the potential functions of C1 neurons.

C1 neurons were first suggested to be important in cardiovascular regulation because of their anatomical overlap with the region of the RVLM critical for the supraspinal drive of sympathetic vasomotor tone, their projections to the region of sympathetic preganglionic neurons, and their sensitivity to cardiovascular perturbations. These general observations have been confirmed and refined during the past 20 years since the critical importance of the RVLM in cardiovascular regulation was widely acknowledged. The rostral group of C1 cells certainly has neuroanatomical connections and response properties consistent with their involvement in cardiovascular reflexes. However, they are intermingled with neurons that also have response properties consistent with a role in cardiovascular regulation. Furthermore, the electrophysiological characteristics and spinal cord pharmacology are consistent with the non-C1 population potentially playing a more dominant role. With the development of anti-DβH saporin toxin, it has been possible to selectively destroy the C1 population in the RVLM and examine the consequences for cardiovascular regulation. Substantial depletions of C1–RVLM neurons produce a decrease in resting mean arterial pressure of approximately 10 mmHg and attenuate sympathoexcitatory cardiovascular reflexes. The response properties of the ascending C1 neurons, projecting primarily to hypothalamic regions, are generally similar to those of the spinally projecting C1 neurons and therefore may play a role in the integration of sympathetic responses with behavioral and neuroendocrine outputs. The activation of C1 cells to a variety of stressful stimuli is consistent with this notion.

Summary

Despite the initial discovery of brain stem PNMT-containing catecholaminergic neurons more than 25 years ago, surprisingly little is known about these neurons. They appear to function within the brain circuits involved in autonomic regulation and integration of autonomic, behavioral, and endocrine responses. The broad appreciation of their anatomical connections is being refined with newly developed techniques but remains consistent with a role of these neurons in autonomic regulation. Finally, the initial

notion that these PNMT-containing neurons are adrenergic neurons is being questioned as more is learned about their neurochemistry.

See also: Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Norepinephrine: CNS Pathways and Neurophysiology; Norepinephrine: Adrenergic Receptors; Sympathomimetic Drugs and Adrenergic Receptor Antagonists.

Further Reading

Card JP, Sved JC, Craig B, Raizada M, Vazquez J, and Sved AF (2006) Efferent projections of rat rostroventrolateral medulla C1 catecholamine neurons: Implications for the central control of cardiovascular regulation. *Journal of Comparative Neurology* 499: 840–859.

Hökfelt T, Johansson O, and Goldstein M (1984) Central catecholamine neurons as revealed by immunohistochemistry with special reference to adrenaline neurons. In: Bjorklund A and Hökfelt T (eds.) *Handbook of Chemical Neuroanatomy: Vol. 2. Classical Transmitters in the CNS*, pp. 157–276. Amsterdam: Elsevier.

Madden CJ and Sved AE (2003) Rostral ventrolateral medulla C1 neurons and cardiovascular regulation. *Cellular and Molecular Neurobiology* 23: 739–749.

Mefford IN (1987) Are there epinephrine neurons in rat brain? *Brain Research Reviews* 12: 383–395.

Schreihofer AM and Guyenet PG (1997) Identification of C1 presympathetic neurons in rat rostral ventrolateral medulla by juxtacellular labeling *in vivo*. *Journal of Comparative Neurology* 387: 524–536.

Stornetta RL, Sevigny CP, Schreihofer AM, Rosin DL, and Guyenet PG (2002) Vesicular glutamate transporter DNPI/VGLUT2 is expressed by both C1 adrenergic and nonaminergic presympathetic vasomotor neurons of the rat medulla. *Journal of Comparative Neurology* 444: 207–220.

Sved AF (1989) PNMT-containing catecholaminergic neurons are not necessarily adrenergic. *Brain Research* 481: 113–118.

Brain Asymmetry: Evolution

A W Toga, K L Narr, P M Thompson, and E Luders, University of California at Los Angeles School of Medicine, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

The structural and functional specialization of the two cerebral hemispheres has warranted a tremendous amount of research during the past century. A number of everyday occurrences, whether we are consciously aware of them or not, provide evidence for asymmetrical behavioral traits: the foot preference of the snowboarder, the ear preference of the cell phone user, or the inclination for most people to use their right hands for fine motor activities. Neuroimaging and lesion studies of the brain confirm hemispheric specialization (or lateralization) for specific behavioral functions and show pronounced asymmetries in the structure of some brain regions in the two hemispheres.

Behavioral Traits: Language and Handedness

The most recognized manifestations of functional lateralization are the dominance of the left hemisphere for handedness and language. Approximately 90% of the population is right-handed, with the motor control of the right hand confined to brain regions in the left hemisphere that include the precentral gyrus or primary motor cortices. Observations that language is more severely impaired in response to tumors or strokes in the left hemisphere, as first reported by nineteenth-century anatomists Broca and Wernicke, also confirm that language functions are lateralized to the left hemisphere in most individuals. Specifically, the brain regions supporting language production and some aspects of syntactic processing are localized primarily in anterior left hemisphere regions that include the opercular and triangular sections of the inferior frontal gyrus (Broca's area). Language comprehension, such as understanding spoken words, is processed in the left posterior section of the superior temporal gyrus and around the temporoparietal junction (Wernicke's area). Interestingly, language dominance and handedness are not perfectly correlated. Approximately 97% of right-handers have their speech and language localized in the left hemisphere, whereas 3% demonstrate right hemisphere lateralization or bilateral language representation. These relationships degrade to 70% (right-lateralized) versus 30% (left-lateralized) in left-handed individuals.

Structural Asymmetries

Asymmetric behavioral traits and functional lateralization appear to be accompanied by hemispheric differences in brain structure. Although cursory examination of the human brain fails to expose profound left–right differences, more careful comparisons of the two hemispheres reveal a variety of asymmetric features. Macroscopic asymmetries (e.g., of fissurization) are complemented by microscopic asymmetries (e.g., of dendritic arborization) and by neurochemical asymmetries (e.g., in dopaminergic sensitivity). The first structural hemispheric differences were described in the late 1800s. They were macroscopic and concerned regions surrounding the Sylvian fissure, a deep sulcus on the lateral surface of the brain which separates the temporal lobe from the frontal and parietal lobe. Subsequently, increasingly more features of the brain have been shown as structurally asymmetric. Findings based on simple visual inspections have been confirmed and complemented by observations from manual region-of-interest analyses and from contemporary computational image analysis methods that allow whole-brain measurements. However, the results across laboratories are not always consistent, and the functional significance of structural asymmetries is not always obvious or agreed upon.

Petalias and Related Asymmetries

Among the most prominent observations of brain asymmetry are the right frontal and left occipital protrusions of the surface of one hemisphere relative to the other (Figure 1). These protrusions also induce imprints on the inner skull surface, known as petalias. Several studies have shown that these petalias are more prominent in right-handers. A second feature, sometimes regarded as separate from the frontal and occipital protrusions, is that the right frontal region is often wider than the left, whereas the left occipital region is often wider than the right (Figure 1). Another related prominent geometric distortion of the hemispheres is known as Yakovlevian anticlockwise torque. This prenatally established pattern encompasses the features described previously and includes the frequent extension of the left occipital lobe across the midline (over the right occipital lobe), bending the interhemispheric fissure toward the right (Figure 1). As a consequence, structures surrounding the right Sylvian fissure are also torqued forward relative to their counterparts on the left.

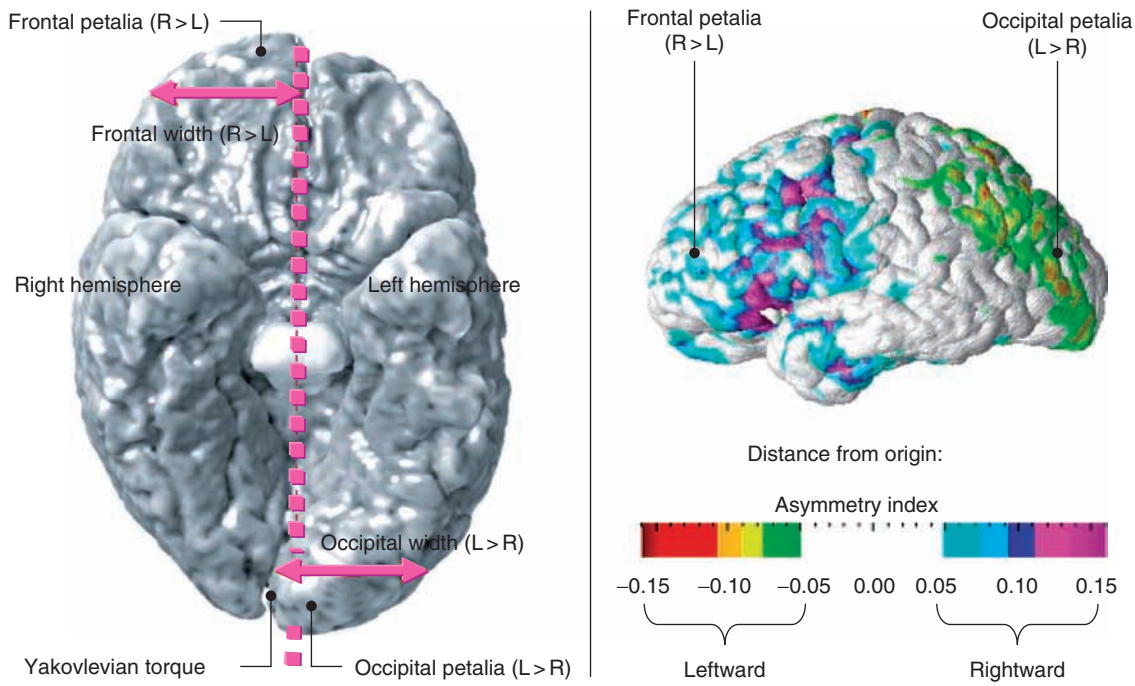

Figure 1 Petalia asymmetry. (Left) A three-dimensional rendering of the inferior surface of the human brain exaggerated to illustrate prominent asymmetries found in the gross anatomy of the two brain hemispheres. Noticeable protrusions of the hemispheres, anteriorly (R > L) and posteriorly (L > R), are observed, as well as differences in the widths of the frontal (R > L) and occipital lobes (L > R). A twisting effect is also observed, known as Yakovlevian torque, in which the left occipital lobe is splayed across midline and skews the interhemispheric fissure in a rightward direction. (Right) The magnitude and direction of hemispheric shape differences, which are estimated by measuring distances from a central point (origin) in the brain to thousands of spatially equivalent cortical surface locations in each hemisphere and by comparing these distances using an asymmetry index. The color scale illustrates anterior protrusions of hemispheric shape in the right hemisphere and posterior protrusions of hemispheric shape in the left hemisphere in one individual.

Sylvian Fissure and Related Asymmetries

At the posterior limit, the right Sylvian fissure curves upward more anteriorly than the left Sylvian fissure in the majority of brains. The Sylvian fissure also typically follows a steeper trajectory in the right hemisphere, whereas it extends further posteriorly and is longer in horizontal length in the left hemisphere (Figure 2). In addition to shape asymmetries in the Sylvian fissure, neuroscientists have noted that on the superior surface of the temporal lobe, buried within the Sylvian fissure, there is typically only one transverse gyrus in the left hemisphere, whereas there are two on the right. These transverse gyri (Heschl's gyri) constitute primary auditory cortex. In addition, the extent or area of the cortical surface posterior to the first gyrus of Heschl, a brain region known as the planum temporale (PT), exhibits a hemisphere-specific gross morphology. The PT, a structure involved in the analysis of sound amplitude and frequency, is commonly observed as larger in the left hemisphere. This leftward asymmetry appears to be related to the degree of handedness, with right-handers exhibiting stronger leftward asymmetries than left-handers. Finally, Broca's area (the pars opercularis and pars triangularis of the inferior frontal gyrus, respectively) has been shown to be larger in volume in the left hemisphere than its homolog in the right hemisphere, although opposing results exist.

Notably, the structural asymmetries of Heschl's gyri, the PT, and Broca's area – all brain regions involved in auditory processing or speech perception and production – might constitute the anatomical substrate for language lateralization. In support of this hypothesis, investigators have observed that verbal fluency and the asymmetry of the pars triangularis are correlated in subjects with above average intelligence. In addition, language lateralization appears to follow the direction of the PT asymmetry. Moreover, others have demonstrated a strong leftward PT asymmetry among subjects with left hemisphere speech representation but no consistent PT asymmetry among subjects with right hemisphere speech representation in a sample of left-handers only.

Central Sulcus and Related Asymmetries

The postcentral gyrus is shown to exhibit structural hemispheric asymmetries in the majority of individuals, where the trajectory of the postcentral sulcus

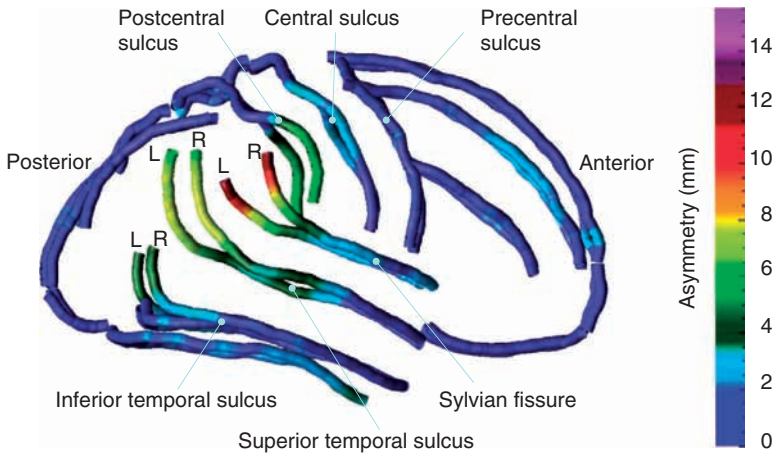

Figure 2 Sulcal asymmetry. From a lateral view, the average shape of the major sulci and fissures for 15 subjects are shown for both the left and the right hemisphere by looking through the brain. The color bar further indexes the magnitude of shape differences between matching sulci in each hemisphere. Note the typical hemispheric differences of the Sylvian fissure (longer in the left (L) and steeper and curving upward more anteriorly in the right (R)), where these asymmetries are largely mirrored in the superior and inferior temporal sulci. Finally, the trajectory of the postcentral sulcus appears to shift more anteriorly in the right hemisphere as opposed to the left hemisphere.

appears to shift more anteriorly in the right hemisphere as opposed to the left (Figure 2). This asymmetry may be associated with slope and horizontal length asymmetries of the Sylvian fissure and temporal sulci and may reflect asymmetries in the parietal operculum that complement PT asymmetries in right-handed subjects.

Whereas some neuroscientists have reported that the central sulcus is generally deeper and larger in the right hemisphere, other investigators have observed a rightward asymmetry of the central sulcus in left-handers only, or even a pronounced leftward asymmetry in right-handers. In fact, this latter macrostructural leftward asymmetry was complemented by a microstructural leftward asymmetry in neuropil volume (i.e., tissue compartment containing dendrites, axons, and synapses) in the primary motor cortex located in the precentral gyrus.

Other asymmetries have been detected in subcortical regions that are more proximal to the motor effectors. For example, the right cortical spinal tract is found to be larger than the left in 75% of subjects, and the left pyramid crosses more rostrally and is larger than the right in 82–87% of subjects. Possibly related to asymmetries in regions that supplement motor function, a rightward asymmetry in callosal regions that contain predominantly projections from the motor cortices (e.g., callosal anterior body) has been reported. Finally, anterior cerebellar volumes have been observed as larger in the right hemisphere, whereas posterior cerebellar volumes are reported as larger in the left.

Ventricular Asymmetries

Some hemispheric differences in brain structure may go unnoticed in individual subjects due to the high intersubject variability of anatomy. Population-based brain atlases overcome this problem by averaging three-dimensional models of anatomy across hundreds, or even thousands, of subjects while storing statistics on anatomic variation. Ventricular asymmetry is an example of a statistically significant effect that becomes clear in a group average brain map but is not universally apparent in individual subjects. As demonstrated in Figure 3, the left lateral ventricle is typically wider and larger than the right, and the left posterior horn is longer in normal adults. This finding is consistent with volumetric measures and may reflect rapid, asymmetric growth in the overlying language systems; it can occasionally be seen in the embryonic brain, using ultrasound, as early as 29–31 weeks postconception.

Tissue Component Asymmetries

Investigators have also examined the asymmetry of brain tissue compartments; they observed significant hemispheric differences of gray matter (GM) in frontal, temporal, parietal, and occipital regions, that include Heschl's gyrus, the PT, the amygdala, and the hippocampus (L > R), as well as in inferior and medial temporal gyrus regions, the lateral thalamus, and the anterior cingulum (R > L). Recent analyses have shown hemispheric differences with respect to the thickness of the cortex, where some regions

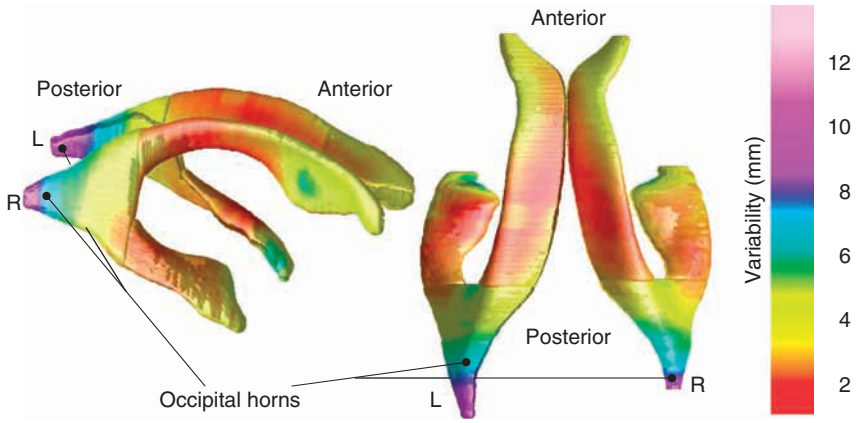

Figure 3 Ventricular asymmetry. The anatomy of the lateral ventricles is shown across subjects ($N = 40$) in three-dimensional view. These maps of average ventricular anatomy show that the left ventricle is larger than the right ventricle. The anatomic asymmetry is clearly localized to the occipital horn, which extends (on average) 5.1 mm more posteriorly on the left than the right. This is consistent with the petalia and torque effects described previously. This asymmetry may go unnoticed in individual subjects due to the high intersubject variability of anatomy.

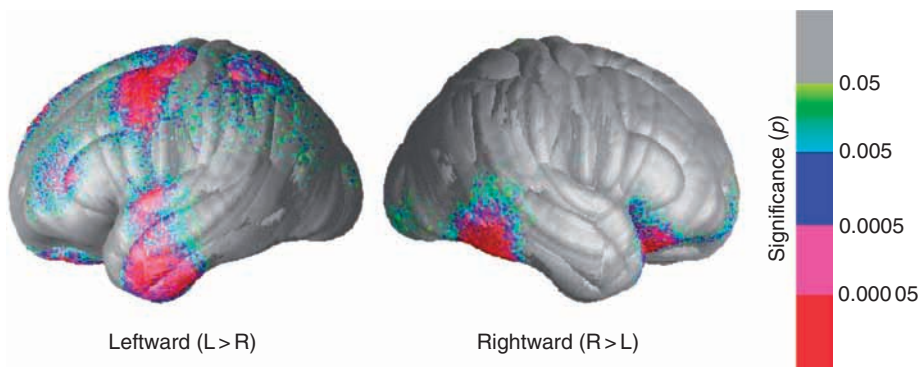

Figure 4 Cortical thickness asymmetry. Statistical maps demonstrating significant hemispheric differences of cortical thickness in a large sample of subjects ($N = 60$). The left brain demonstrates leftward (L > R) asymmetries in the anterior temporal lobe, including the inferior, middle, and superior temporal gyri and the precentral gyrus extending anteriorly to adjacent regions. Two additional larger clusters favoring the left are apparent in the middle frontal gyrus and superior parietal lobe (extending more diffusely inferiorly, covering the inferior parietal lobe and supramarginal gyrus). Smaller clusters of leftward asymmetry are evident in superior frontal regions very close to the midline extending along the longitudinal fissure and in the orbital gyrus. The right brain demonstrates significant rightward asymmetries (R > L) in the posterior inferior temporal lobe and inferior frontal gyrus (comprising the pars orbitalis, triangularis, and opercularis and extending into the extreme anterior tip of the temporal lobe) and near the frontal pole. In general, leftward asymmetries are spread over larger regions than rightward asymmetries.

exhibit pronounced thickness asymmetries that resemble GM asymmetries while other regions reveal distinct asymmetries (Figure 4).

Determining Factors of Brain Asymmetries

Heredity versus Environment

Perisylvian asymmetries are already present in children. Their magnitude, however, appears to increase throughout childhood and the teenage years, even after adjusting for developmental increases in brain volume. This suggests that there may be hemispheric differences in white matter maturation, perhaps during the many regional growth spurts in myelination that occur in childhood. It is not certain which factors determine the initiation or amplification of cerebral asymmetries. It appears unlikely that hemispheric differences are solely genetically predetermined. For example, an individual's genotype may not be the only determinant for laterality because many identical twins are discordant for handedness and differ considerably in their expression of PT asymmetry. However, genetic factors appear to influence hemispheric volumes twice as strongly in right-handed

twin pairs than in twin pairs with at least one left-hander (non-right-handers). Investigators have hypothesized that the decrement in genetic control of cerebral volumes in non-right-handed twin pairs supports the existence of a 'right-shift' genotype in the majority of the population (expressed as a right hand/left hemisphere bias) that is lost in non-right-handers. Thus, genes appear to be significant contributors to brain and behavioral asymmetries, where the degree of genetic determination also depends on brain structure. For example, GM volumes in perisylvian areas are reported to be under predominantly genetic control, whereas gyral and sulcal patterns have been demonstrated as much less heritable. A number of potential nongenetic contributors to variations in brain asymmetry, including fetal orientation, hormones, and functional adaptations, are discussed later. In addition, gender-specific asymmetries and disturbances in asymmetries as may be associated with some specific diseases are explained.

Fetal Orientation

Asymmetric influences in the prenatal environment may lead to perceptual and motor asymmetries. Two-thirds of fetuses are confined to a leftward fetal position in the third trimester, with their right side facing outward. Lateralization of language perception may result from asymmetries in their auditory experience. For example, the right ear may be better positioned to discriminate high-frequency speech sounds. It has also been argued that asymmetrical vestibular stimulation *in utero* may produce behavioral motor asymmetries later in life. Finally, the exposure to ultrasound in fetal life has been suggested to increase the chances of being left-handed by approximately 30%.

Hormones

In male rats, the right neocortex is thicker than the left. Female rats, on the other hand, display a non-significant trend toward the opposite pattern. Notably, castration at birth, which prevents the flow of androgens from the testis to the brain, blocks the formation of the normal rightward brain asymmetry in male rats. Similarly, neonatal ovariectomy reverses the female pattern to the male pattern. Interestingly, maternal environmental or nutritional stress has been demonstrated to reverse the male-typical asymmetry in fetal male rats to the female pattern by both shifting and depressing a testosterone surge that normally occurs on gestational day 18. These findings suggest that levels of androgenic and ovarian sex steroids, before and after birth, play a role in modulating brain asymmetry, at least in rodents.

In humans, the hormonal determinants for sex-specific symmetries are less understood. In their widely cited theory of cerebral lateralization, Geschwind and Galaburda suggested that elevated testosterone levels might be responsible for deviations from the normal dominance pattern (i.e., right-handedness and leftward language dominance, as well as rightward visuospatial dominance). According to this theory, if testosterone levels are higher than normal *in utero*, consequences include a smaller left hemisphere and even anomalous dominance due to a delay of left-hemispheric growth. This model has been posited to explain the different maturational rates of the sexes (with females generally maturing faster) and also to explain the relative advantage males show for right hemisphere visuospatial tasks and that females show for left hemisphere linguistic tasks. It has also been used to explain the greater incidence of left-handedness in males. Finally, hormonal influences might largely determine gender-specific structural asymmetries, as outlined next.

Gender-Specific Asymmetries

Gender differences in structural asymmetries, with larger interhemispheric differences in males compared to females, have been widely replicated where sex-dependent patterns appear to be complemented by behavioral and neuroactivational scores. To explain greater functional lateralization in men compared to women, it has been suggested that either the functions of the hemispheres are less sharply differentiated in women than in men or, alternatively, that larger commissural systems in women may act to reduce the difference in lateralized response scores between hemispheres. Sex differences in brain organization, both within and between hemispheres, are thought to underlie sex differences in motor and visuospatial skills, linguistic performance, and vulnerability to deficits following stroke and other focal lesions.

With emphasis on structural asymmetries, a larger right hemisphere volume has been identified in male fetuses, but no equivalent pattern has been reported in adults. Other studies have shown a significantly deeper central sulcus in the left hemisphere than in the right, but only in male right-handers; interhemispheric asymmetry was reported absent in female right-handers. Similarly, a larger leftward GM asymmetry in males compared to females has been observed in a region posterior to the central sulcus. Possibly related to these findings, pronounced rightward parasagittal asymmetries have been detected in the anterior callosal body in males but appear to a much lesser degree in females (Figure 5). Other

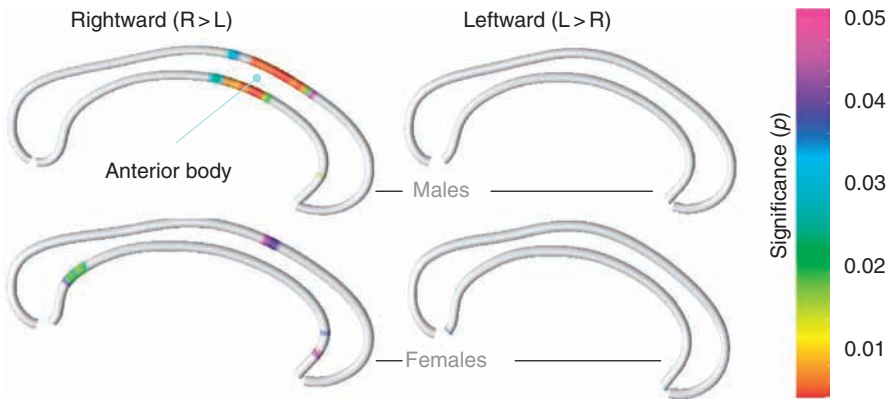

Figure 5 Callosal asymmetry. Statistical maps demonstrating significant gender-specific asymmetries in a large sample of subjects (30 men and 30 women). Differences between callosal thicknesses were measured in the left and right hemispheres several millimeters apart from the midsagittal plane. Rightward asymmetries are largely increased in men, supporting the assumption of a sexually dimorphic organization of male and female brains that involves hemispheric relations and is reflected in the organization and distribution of callosal fibers.

observations of sexually dimorphic cerebral asymmetries include pronounced rightward asymmetries of the planum parietale in right-handed men compared to right-handed women, whereas left-handed subjects demonstrate the opposite pattern. Gender-dependent asymmetries of the inferior parietal lobe and PT, with males having significantly larger leftward asymmetries and females showing reversed, diminished, or no asymmetries, have also been reported. Greater lateralization of right frontal petalias (in right-handers and left-handers) and occipital petalias (in left-handers only) in men compared to women further confirm previous findings of greater frontal and occipital asymmetries in men and reductions of the typical asymmetries in women.

Although the majority of studies suggest diminished asymmetries in female brains, a number of studies also exist that either failed to detect significant gender effects with respect to hemispheric differences or revealed even more pronounced asymmetries in females. A study that analyzed hemispheric differences with respect to the thickness of the cortex observed asymmetry profiles that were similar in both sexes. Notwithstanding, hemispheric differences appeared slightly pronounced in males compared to females, albeit a few regions also indicated greater asymmetry in females compared to males.

Functional Adaptation

Experience-dependent plasticity and asymmetric behaviors may also induce different neuronal changes in the two hemispheres. In rats, the asymmetric use of only one forelimb in the postweaning period induces an asymmetrically larger neuropil volume and lower cell packing density in the motor cortex. In mice with a hereditary asymmetry in their whisker pads, a dominant right whisker pad has been associated with left paw preference. Limb preference may therefore be associated with asymmetries in sensory input, although it is not known whether this relationship is causal.

In humans, professional right-handed keyboard players who had received intensive bimanual training from early childhood are more symmetrical in hand skill tests. Moreover, the absolute length of the central sulcus appears related to the age of commencement of musical training such that those musicians who had started early in life with musical training exhibited the longest sulcus on both sides. These findings suggest that some brain asymmetries are not necessarily genetically determined and may result from lateralized motoric activity and/or sensory stimulation.

Aberrant Asymmetries and Disease

Reduced or even inverted volume asymmetries of the PT have been reported in some subjects with reading disorders or developmental dyslexia, as well as in some people with unusual right-hemispheric dominance for speech. Analogously, functional magnetic resonance imaging studies have shown a pattern of brain activation in stutterers that is shifted toward the right in both motor and auditory language areas. This may suggest an inherent difference in the way in which normal subjects and stutterers process language.

Controversy surrounds reports of altered brain asymmetry in schizophrenia. For example, it has been proposed that altered anatomical and functional asymmetries in schizophrenia constitute a genetic and evolutionary basis for the disease that has developed in concert with hemispheric specialization for language. It has also been suggested that schizophrenia

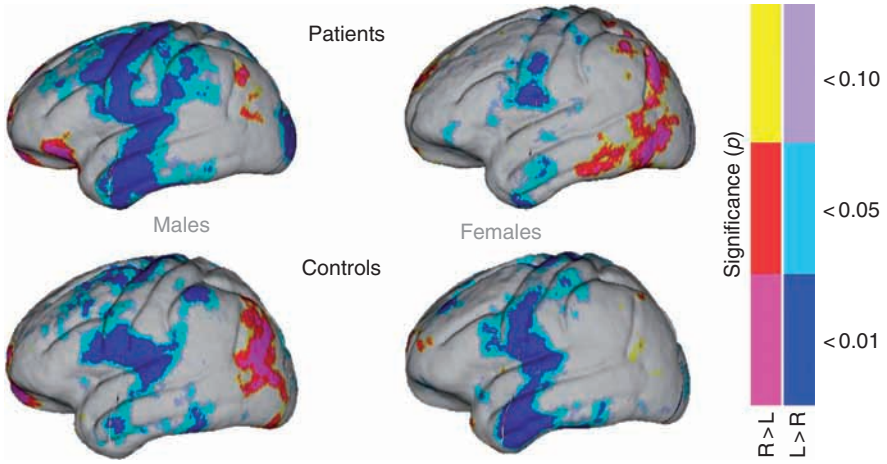

Figure 6 Cortical thickness asymmetry in schizophrenia. Statistical maps show significant hemispheric differences in cortical thickness within groups defined by sex and a diagnosis of schizophrenia ($N = 150$). The patterns of cortical thickness asymmetries appear similar in all groups (leftward asymmetries of thickness in sensorimotor and perisylvian cortices and rightward cortical thickness asymmetries in posterior temporopariteal cortices). Notably, these patterns were not shown to differ statistically in patients with schizophrenia compared to demographically similar healthy comparison subjects.

is due to an anomaly of cerebral dominance, in which an agnosic right shift gene is suggested to be a major contributor toward schizophrenia pathophysiology. Observations of altered structural asymmetries in patients with schizophrenia include the reversal of normal petalias; disproportionately reduced left hemisphere temporal lobe volumes; reduced asymmetries of the PT, the superior temporal gyrus, and the Sylvian fissure; and alterations in hemispheric gyrification indices. Findings indicating intact structural lateralization in schizophrenia, however, are not uncommon. For example, one study failed to detect schizophrenia-related changes in cortical thickness asymmetries in a large sample of schizophrenia patients compared to demographically similar healthy comparison subjects (Figure 6). Thickness asymmetries, however, were shown to vary by disease status in non-dextrals.

Some diseases appear to progress asymmetrically. Patients with semantic dementia generally show asymmetric anterolateral temporal atrophy (typically worse on the left side) with relative sparing of the hippocampal formation. In Alzheimer's disease, a spreading wave of GM loss emerges initially in entorhinal and temporal-parietal cortices, sweeping into frontal and ultimately sensorimotor territory as the disease progresses. This sequence occurs in both hemispheres, but left hemisphere regions are affected earlier and more severely. The right hemisphere follows a similar pattern approximately 2 years later (Figure 7). Furthermore, cerebrospinal fluid volumes in the Sylvian fissure appear to rise more sharply on the left than the right in patients with dementia compared to normal controls. Additional left-greater-than-right metabolic dysfunctions in a group of patients with early dementia have been observed using positron emission tomography. These disease-related asymmetries suggest either that the left hemisphere is more susceptible than the right to neurodegeneration in Alzheimer's disease or that left hemisphere pathology results in greater structural change and lobar metabolic deficits.

Evolutionary Origins of Anatomical Asymmetries

Brain Size Expansion

Increasing brain size in humans is possibly one of the driving forces in the phylogeny of hemispheric specialization. For example, it has been suggested that the massive evolutionary expansion of the brain may have resulted in a level of complexity where duplication of structures was no longer efficient. Functional incompatibility or the need for simultaneous parallel processing related to increasing cognitive capacities during the course of evolution might also constitute driving forces for the emergence of functionally separated systems. Due to competition for space within the brain, different functions might have been confined to different hemispheres (rather than to different networks within a hemisphere). As a possible consequence, the right hemisphere in humans outperforms the left in the analysis of spatial relations and also shows a tendency for global processes (e.g., with respect to attention and memory storage). The left hemisphere,

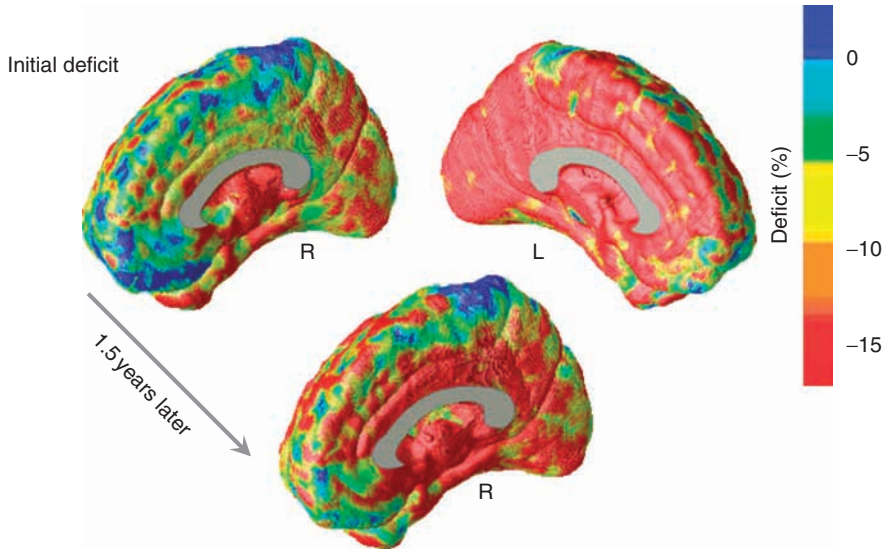

Figure 7 Asymmetrical progression of Alzheimer's disease. These maps show the average profile of GM loss in a group of patients with mild to moderate Alzheimer's disease ($N = 17$) compared to a group of healthy age- and gender-matched controls ($N = 14$). Initially, the right hemisphere (R) is much less severely affected than the left (L), but after 1.5 years the deficit progresses to encompass more of the right hemisphere.

on the other hand, is specialized in language and shows superiority in categorizing (as opposed to global processing).

Time limits associated with the transfer of information across the corpus callosum between the brain hemispheres may also favor development of unilateral networks in larger brains. It has been reported that more asymmetrical brains have a corpus callosum with a reduced midsagittal area relative to more symmetrical ones. In addition, there exists an inverse relationship between forebrain size and relative callosal size accounting for relatively smaller callosal areas in larger brains. Since smaller callosal areas may reflect fewer or thinner fibers connecting the two hemispheres, this suggests the degree of interhemispheric connectedness decreases with increased brain size.

Left-Hemispheric Dominance for Language

The evolutionary development of language in humans may have led to marked volume asymmetries in structures crucial for speech production and perception as well as for motor dominance. For example, in humans, the left PT – an extension of Wernicke's posterior receptive language area – is up to 10 times larger than its right hemisphere counterpart and is perhaps the most functionally significant human brain asymmetry. Language is commonly lateralized to the left hemisphere, and some argue that this is advantageous: (1) it avoids competition between hemispheres for control of the muscles involved in speech, and (2) it may be more efficient to transfer language information between collections of focal areas in a single hemisphere. The main pitfall in arguing that left hemisphere dominance provides an evolutionary advantage is that bilateral language representation, or rightward dominance, is also common. In addition, leftward dominance does not, in general, provide a cognitive advantage.

Left-Hemispheric Dominance for Handedness

A hypothesis first proposed by Condillac in 1746 suggests that the left hemisphere's dominance for language processing evolved from its control of the right hand. The left-hemispheric programming of skilled movement and gesture may have evolved to encompass control of the motor systems involved in speech. Broca's area, in particular, is a premotor module that sequences complex articulations that are not limited to speech. Great apes, including chimpanzees and gorillas, also have an enlarged area 44 (Broca's area). This area controls muscles of the face and vocal tract but is not as well connected with the homolog of Wernicke's area as it is in humans. It has been suggested that nonhuman primates developed a homolog of Broca's area due to a link between primate vocalization and gesture: captive apes usually gesture with the right hand as they vocalize. Research on indigenous gestural languages invented by children in Taiwan and in Nicaragua provides some evidence for the innate relation between gesture and language. Functional neuroimaging studies also suggest that deaf subjects using a gestural sign language activate many of the

systems involved in verbal language production. These congruencies in functional anatomy may support the hypothesis that verbal language evolved from gestural language as an outgrowth of the already asymmetric motor control system. It was proposed that language is a relatively recent evolutionary adaptation (not more than 200 000 years old), where the Neanderthal vocal tract was incapable of articulating the range of modern human speech sounds.

Brain Asymmetries in Nonhuman Species

Functional asymmetries in the brain were initially thought to be uniquely human, reflecting unique processing demands required to produce and comprehend language. Nonetheless, functional and structural asymmetries have been identified in nonhuman primates and many other species. For example, Japanese macaques exhibit a right ear advantage for processing auditory stimuli, passerine birds produce song primarily under left hemisphere control, and frogs utilize their left hemisphere to control clasping vocalizations. Moreover, with respect to structural asymmetries, cerebral petalias are seen in phylogenetically older primates (and other species), as evidenced by endocasts from fossilized cranial bones. Similarly, the PT asymmetry also appears in higher nonhuman primates. In most great apes, the Sylvian fissure is longer and straighter on the left side than on the right. It has also been observed that the right hemisphere is generally larger than the left in rats, mice, rabbits, and cats. Although asymmetries in nonhuman species can be strikingly similar to those in humans, asymmetries defined in animal studies may not be easy to extrapolate to humans because the precursors of language-related asymmetries in humans may not be present in other species. The mechanisms that underlie some cerebral asymmetries in humans might differ substantially from those that underpin brain asymmetry in other mammals. Notwithstanding, the existence of hemispheric differences in nonhumans indicates that brain asymmetry is neither unique to humans nor completely dependent on the development of language.

Acknowledgments

This work was supported by NIBIB/NINDS/NIMG grant P01 EB001955 and NIH/NCRR resource grant P41 RR013642. Additional support was provided by the National Institutes of Health through the NIH Roadmap for Medical Research grant U54 RR021813. Additional support was also provided by the National Institute of Mental Health grant K01 MH073990.

See also: Brain Damage: Functional Reorganization; Brain Injury: Functional Recovery After; Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Dichotic Listening Studies of Brain Asymmetry; Hemispheric Specialization and Cognition.

Further Reading

Galaburda AM and Geschwind N (1981) Anatomical asymmetries in the adult and developing brain and their implications for function. *Advances in Pediatrics* 28: 271–292.

Geschwind N and Galaburda AM (1987) *Cerebral Lateralization: Biological Mechanisms, Associations, and Pathology.* Cambridge: MIT Press.

Hellige JB (1990) Hemispheric asymmetry. *Annual Review of Psychology* 41: 55–80.

Hugdahl K and Davidson RJ (2003) *The Asymmetrical Brain.* Cambridge, MA: MIT Press.

Previc FH (1991) A general theory concerning the prenatal origins of cerebral lateralization in humans. *Psychological Review* 98: 299–334.

Toga AW and Thompson PM (2003) Mapping brain asymmetry. *Nature Reviews Neuroscience* 4: 37–48.

Zilles K, Dabringhaus A, Geyer S, et al. (1996) Structural asymmetries in the human forebrain and the forebrain of non-human primates and rats. *Neuroscience and Biobehavioral Reviews* 20: 593–605.

Relevant Websites

http://www.loni.ucla.edu – Laboratory of Neuro Imaging, UCLA.
http://nihroadmap.nih.gov – National Institutes of Health Roadmap for Medical Research.

Brain Composition: Age-Related Changes

V Prakash Reddy, University of Missouri-Rolla, Rolla, MO, USA
G Perry, University of Texas at San Antonio, San Antonio, TX, USA
M A Smith, Case Western Reserve University, Cleveland, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

At the cellular level, age-related neurological diseases including Alzheimer's disease (AD), Parkinson's disease, and Down syndrome share the characteristics of oxidative stress. Oxidation of lipids, DNA/RNA, and proteins produces a variety of toxic compounds, protein modifications, and DNA damage. AD is the leading cause of senile dementia, and there are more than four million AD patients in the United States alone. It is estimated that by 2050, fourteen million people will be inflicted by this disease. The familial form of AD worldwide accounts for 5–10% of the cases, whereas the sporadic form ranges from 90% to 95%. The brains of sporadic as well as familial cases of AD are characterized by extensive neuronal loss, deposition of extracellular senile plaques consisting of amyloid-β (Aβ) peptide aggregates, and intracellular neurofibrillary tangles (NFTs) consisting of hyperphosphorylated tau proteins.

Aβ40–42 peptides (Aβ40–42) are formed *in vivo* through the consecutive action of β-secretase and γ-secretase on Aβ protein precursor (AβPP), whose physiological function is not yet clear. Presenilin 1 and 2 are the intermembrane aspartyl proteases that process many diverse proteins in the lipid bilayers and form the active centers of the γ-secretase enzyme. Transition metal-mediated oxidative stress plays a key role in the formation of Aβ aggregates and NFTs. Further, cell cycle re-entry followed by abnormal mitotic alterations in some neuronal cells (less than one in 10 000 at a given time) leads to their apoptosis.

Protein modifications as a result of glycation and lipid peroxidations also play a major role in the regulation of oxidative stress at sites of the lesions of AD. Antioxidants such as vitamin C and α-tocopherol (vitamin E) as well as nonsteroidal anti-inflammatory drugs (NSAIDs) such as aspirin have limited protective effects against AD. In addition, metal ion chelators may also have beneficial effects for the control of the progression of neurodegenerative diseases. Among the most commonly used drugs for the treatment of the symptoms of AD are Donepezil and galantamine (also called as Razadyne; galantamine hydrobromide), which are acetylcholinesterase (AChE) inhibitors. Donepezil shows concentration-dependent inhibition of cellular proliferation in human neuroblastoma cell lines, and thus approaches in targeting cell cycle abnormalities for the treatment of AD are promising. Cholesterol and other lipids also aid in the secretion of Aβ from AβPP. Thus, cholesterol/lipid-lowering drugs are also being considered for the treatment of AD. Statins inhibit enzymes involved in the biosynthesis of cholesterol, and in addition show anti-inflammatory effects and show other beneficial effects by modulating α-secretase and β-secretase enzymes. Further, stereoisomerization of the naturally occurring L-amino acid residues of the long-lived proteins *in vivo* to D-amino acids results in the formation of proteolysis-resistant toxic peptides which may also contribute the etiology of the age-related diseases. To date, treatment effectiveness has been modest and variable among patients. While any improvement or delay of disease progression is viewed positively by families and caregivers, all drug treatments require careful monitoring and consideration of the risks and side effects.

Lipids

Lipid oxidation (also called lipid peroxidation) has consequences in aging and age-related diseases including neurological diseases such as AD and Parkinson disease. Membrane- and lipoprotein-associated polyunsaturated fatty acids upon reactive oxygen species (ROS) or reactive nitrogen species (RNS)-mediated oxidations give rise to a variety of toxic molecules including malondialdehyde, glyoxal, acrolein, *trans*-4-hydroxy-2-nonenal (HNE), and *trans*-4-oxo-2-nonenal (ONE). HNE and ONE are particularly toxic compounds to the surrounding biomacromolecules, as they can react relatively readily with cysteine-derived thiol groups and ε-amino groups of lysine units, which may result in enzyme deactivation. They can also deplete levels of intracellular glutathione, a cellular antioxidant, by Michael reactions involving the thiol group of the glutathione. Similarly, HNE deactivates glutathione peroxidase enzyme, which is responsible for detoxifying hydrogen peroxide. HNE and ONE also react with amino groups of nucleic acid-bases in DNA forming a variety of adducts, resulting in transversion mutations. HNE-derived propano adducts have been shown to exist in cerebellum and cortex of cases of AD (Figure 1).

There is strong evidence for the modulation of the activity of the γ-secretase as a function of the concentration of cholesterol and other lipids: higher cholesterol levels lead to the enhanced formation of

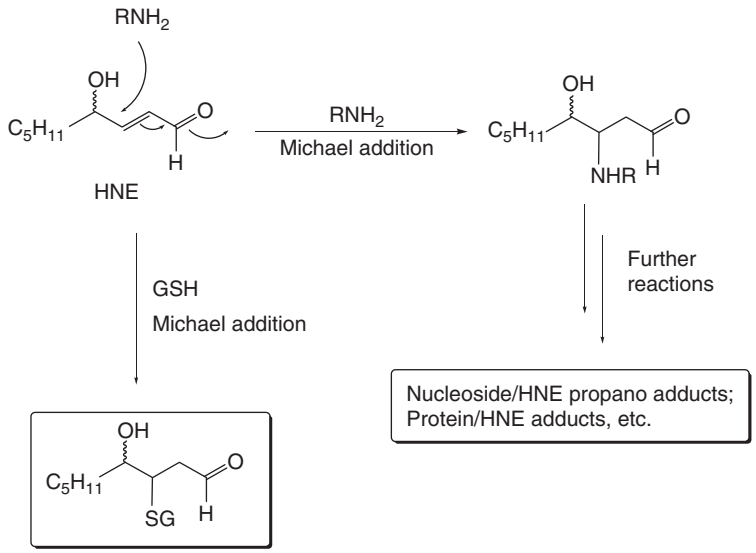

Figure 1 HNE-derived propano adducts are found in the brain of patients with AD.

the Aβ peptide and vice versa. Hypercholesterolemia is a risk factor for AD, as elevated concentrations of cholesterol are found in the brains of AD patients during the progression of AD, and thus cholesterol-lowering drugs are currently being developed for the treatment of AD. Statins (3-hydroxy-3-methylglutaryl-CoA reductase inhibitors) have well-established cholesterol-lowering properties and inhibit the Aβ-induced expression of interleukin-1β and inducible nitric oxide synthase (iNOS). Statins also have anti-inflammatory properties that are independent of their cholesterol-lowering effects. Although at present there are no reliable clinical data to show the effectiveness of statins in AD therapy, they are expected to attenuate the symptoms of AD by cholesterol-lowering as well as anti-inflammatory properties. Apolipoprotein E (ApoE) deficiency in AD brains impairs the lipid transport system, thereby altering lipid homeostasis in cases of AD. The cholesterol-lowering drug Probucol increases the concentrations of cerebrospinal fluid (CSF) ApoE levels and decreases the CSF Aβ1–42 levels.

Metals and Oxidations

Transition metal ions, for example, Fe^{2+} and Cu^{+}, although essential in maintaining the central nervous system, induce neuronal damage when present in excess quantities. Fe^{2+} and Cu^{+} are sources of ROS (e.g., hydrogen peroxide, hydroxyl radical, and superoxide radical anion) and RNS (e.g., nitrous acid, nitric oxide, peroxynitrite) through Fenton and related reactions. Metal ion-catalyzed reduction of molecular oxygen gives the superoxide radical anions, which are further reduced to hydrogen peroxide (H_2O_2). Although $O_2^{\bullet -}$ and H_2O_2 by themselves are relatively nontoxic, H_2O_2 may lead to the production of hydroxyl radicals (HO•) through metal ion-catalyzed Fenton reactions. The hydroxyl radicals are highly toxic to neurons and the surrounding medium. Aβ may be a source of the ROS, as it binds to Cu^{2+} and Fe^{3+}. In the presence of biological reducing agents, Cu^{2+} and Fe^{3+} are reduced to Cu^{+} and Fe^{2+}, which are mediators of the Fenton reaction forming hydroxyl radicals (Figure 2). In agreement with this, when Aβ is pretreated with the iron chelator desferrioxamine, neuronal toxicity is significantly reduced. In various animal models, the addition of trace amounts of copper, but not zinc or aluminum,-resulted in increased Aβ deposition. However, copper ions bound to Aβ can cause conformational changes that confer to it a superoxide dismutase activity, maintaining oxidative balance. Thus, the role of Aβ in the cause of oxidative stress in AD may be dependent on the metal ion homoeostasis in the surrounding medium. It is notable that although aggregated Aβ peptide in combination with associated metal ions can induce oxidative stress, the formation of Aβ itself is triggered by enhanced cellular oxidative stress, through activation of β-secretase and expression of AβPP by the activation of the AP-1 transcription factor.

There is abundant evidence showing that the oxidative stress mediated by ROS and RNS precedes the onset of the formation of Aβ and NFT in cases of AD. Serving as markers of oxidative damage during aging

and in age-related diseases such as AD, Parkinson's disease, and Down syndrome, 8-hydroxydeoxyguanosine (8-OHdG), 8-hydroxyguanosine (8-OHG), protein carbonyls, and nitrotyrosine are abundantly formed in neuronal cells, decades before the onset of the disease hallmarks, such as Aβ plaques or NFTs. Abnormal mitochondria are the source of the majority of the ROS produced in cells, and metal ions are intrinsically involved in their formation. Iron and copper ions have been demonstrated histochemically to be localized in Aβ plaques and NFTs, the main constituents of the brains of AD. Desferrioxamine (Fe^{3+} chelator), tetrathiomolybdate (Cu^{2+} chelator), and mercaptopropanol (Pb^{2+} and Hg^{2+} chelator) suppress the expression of AβPP, thus lowering the Aβ peptide secretion. Desferrioixamine and clioquinol, a Cu^{2+} chelator, have found limited therapeutic applications for AD (Figure 3). Oxidative stress may contribute to the irreversible protein aggregation associated with neuronal degeneration. For example, HNE binding to phosphorylated tau protein facilitates the *in vitro* formation of filaments. Also, oxidative stress and reactive oxygen species bring about the covalent cross-linking of tau filaments to form large aggregates that are resistant to proteolytic cleavage.

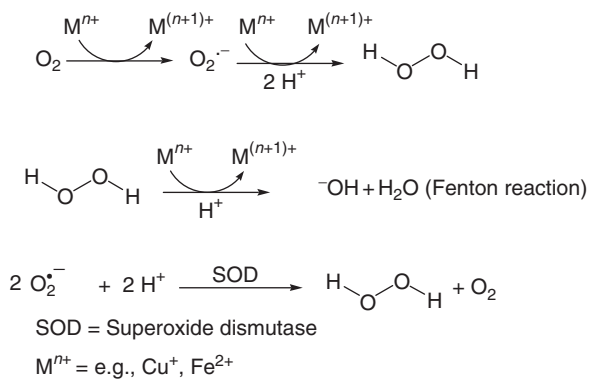

Figure 2 Reactions involving the detoxification of reactive oxygen species (ROS).

Racemization

The amino acid residues of long-lived proteins in biological tissues are progressively racemized with time, and the racemization process may be abnormally accelerated in age-related diseases. Long-lived proteins such as lens crystallin show increased amounts of D-amino acid residues, which is due to the slow racemization of the L-amino acids over time. Aspartic acid residues are especially prone to racemization. D-Asp concentrations increase with age in tissues such as dentine and may serve as estimates of the age of the individuals. In lens crystallin proteins, amino acid residues Asp151 and Asp58 are relatively more stereochemically labile. Racemization of the native L-amino acids in proteins results in the change of their conformations, resistance to proteolytic degradation, and loss of enzyme activity of the involved enzymes.

There are several reports of racemization of the Ser and Asp residues of the insoluble Aβ1–40 peptides *in vivo* during the aging process, over a 20–30 year duration. The secondary structures of the [D-Asp23] Aβ20–29 and [D-Ser26]Aβ20–29 have been shown to differ from the corresponding unracemized peptides by vibrational circulation dichroism. Purified amyloid was found to contain relatively large amounts of D-aspartate and D-serine, which may be derived from the long-lived AβPP. The Aβ peptide with racemized Ser or Asp residues, but not the unracemized peptide, is hydrolyzed by chymotrypsin-like enzymes to form soluble peptide fragments, for example, [D-Ser26]Aβ25–35, which are neurotoxic and are involved in the degeneration of hippocampal neurons in AD brains. The latter peptides are also resistant to brain proteinases.

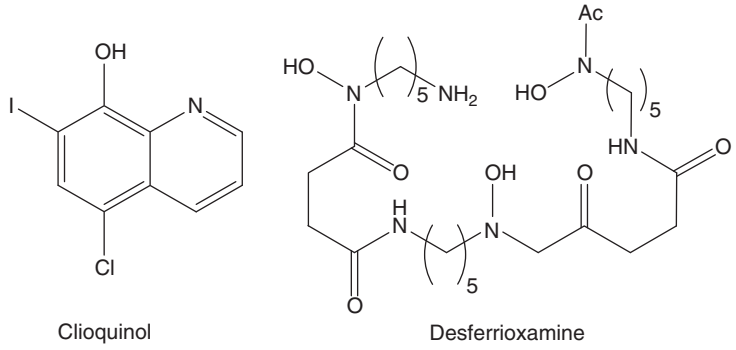

Figure 3 The metal chelators, clioquinol, and deserioxamine, have both shown promise in clinical trials for the treatment of AD.

Conclusions

Aging and age-related diseases, in particular AD, have oxidative stress as an important pathogenic factor. ROS, derived through abnormal mitochondrial metabolism involving transition metal ions, and lipoxidation-derived aldehyde products (also implicating transition metal ions in their generation) oxidatively modify cellular proteins and exert DNA/RNA damage. Oxidative stress precedes the formation of the Aβ plaques and NFTs in AD, which in turn, through associated transition metal ions, further exacerbate the oxidative stress. Further, racemization of amino acids in the long-lived proteins, for example, AβPP in AD, may result in the formation of proteolytically stable, toxic peptides. Antioxidants and agents that would sequester the ROS and lipid peroxidation endproducts may be useful targets in the identification of drugs for age-related diseases.

See also: Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Amyloid: Vascular and Parenchymal; Lipids and Membranes in Brain Aging; Metal Accumulation During Aging.

Further Reading

Bosch-Morell F, Flohe L, Marin N, and Romero FJ (1999) 4-hydroxynonenal inhibits glutathione peroxidase: Protection by glutathione. *Free Radical Biology and Medicine* 22: 1383–1387.

Friguet B, Stadtman ER, and Szweda LI (1994) Modification of glucose-6-phosphate dehydrogenase by 4-hydroxy-2-nonenal. Formation of cross-linked protein that inhibits the multicatalytic protease. *Journal of Biological Chemistry* 269: 21639–21643.

Grimm MOW, Grimm HS, Paetzold AJ, et al. (2005) Regulation of cholesterol and sphingomyelin metabolism by amyloid-β and presenilin. *Nature Cell Biology* 7: 1118–1123.

Perry G, Castellani RJ, Hirai K, and Smith MA (1998) Reactive oxygen species mediate cellular damage in Alzheimer disease. *Journal of Alzheimers Disease* 1: 45–55.

Poirier J (2005) Apolipoprotein E, cholesterol transport and synthesis in sporadic Alzheimer's disease. *Neurobiology of Aging* 26: 355–361.

Ritz-Timme S and Collins MJ (2002) Racemization of aspartic acid in human proteins. *Ageing Research Reviews* 1: 43–59.

Selkoe DJ (2004) Cell biology of protein misfolding: The examples of Alzheimer's and Parkinson's diseases. *Nature Cell Biology* 6: 1054–1061.

Shapira R, Austin GE, and Mirra SS (1988) Neuritic plaque amyloid in Alzheimer's disease is highly racemized. *Journal of Neurochemistry* 50: 69–74.

Sjoegren M, Mielke M, Gustafson D, Zandi P, and Skoog I (2006) Cholesterol and Alzheimer's disease – Is there a relation? *Mechanisms of Ageing and Development* 127: 138–147.

Sortino MA, Frasca G, Chisari M, et al. (2004) Novel neuronal targets for the acetylcholinesterase inhibitor Donepezil. *Neuropharmacology* 47: 1198–1204.

Sparks DL, Friedland R, Petanceska S, et al. (2006) Trace copper levels in the drinking water, but not zinc or aluminum influence CNS Alzheimer-like pathology. *Journal of Nutrition, Health, and Aging* 10: 247–254.

Takeda A, Smith MA, Avila J, et al. (2000) In Alzheimer's disease, heme oxygenase is coincident with Alz50: An epitope of tau induced by 4-hydroxy-2-nonenal modification. *Journal of Neurochemistry* 75: 1234–1241.

Webber KM, Raina AK, Marlatt MW, et al. (2005) The cell cycle in Alzheimer disease: A unique target for neuropharmacology. *Mechanisms of Ageing and Development* 126: 1019–1025.

Zhu X, Raina AK, Perry G, and Smith MA (2004) Alzheimer's disease: The two-hit hypothesis. *Lancet Neurology* 3: 219–226.

Brain Connectivity and Brain Size

A Schüz, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
F Sultan, Hertie Institute for Clinical Brain Research, Tübingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Types of Connectivity in the Brain

The various parts of the brain differ in their macroscopic layout and internal structure. They can have the shape of flat pieces of gray substance – such as the cerebral and the cerebellar cortex, the dentate nucleus of the cerebellum, the tectum opticum of nonmammalian species, and the vagal lobe in some fishes – or that of globular structures, such as the many thalamic nuclei or parts of the striatum. The neurons of which they are composed differ in shape, size, and patterns of ramification. Particular synaptic arrangements, such as the glomeruli in the cerebellum or the triads in the lateral geniculate body of the thalamus, can occur. In addition, the various parts of the brain differ in their internal connectivity. For example, the neurons in the cerebral cortex form a rich network among themselves, as do the neurons in the striatum. However, the two kinds of networks differ in one decisive point: the cerebral cortex forms a predominantly excitatory network, whereas the network within the mammalian striatum is mainly inhibitory. In contrast to the cortex and the striatum, most of the neurons of the main cell type in the thalamus seem not to form a network among themselves but to relay their input directly to the cortex. A similar situation is found in the cerebellar cortex: the neurons of the most numerous cell type, the granule cells, do not make any synapses with each other but relay their input to the output neurons, the Purkinje cells. These, and other characteristic features, reflect the specific function of a given part of the brain.

Brain Size and Number of Neurons

The weight of the brain of different vertebrates ranges over six orders of magnitude. In mammals, the range is five orders of magnitude: the smallest mammalian brain is that of the pygmy white-toothed shrew (*Suncus etruscus*), weighing 0.062 g, and the largest is found in the sperm whale (*Physeter macrocephalus*), weighing approximately 9500 g.

In most organs, an increase in volume is due to an increase in the number of cells, whereas the size of the cells remains nearly constant. In the brain, however, the increase in the number of neurons is much less than the increase in its volume (Figure 1). Here, the brain as an interconnected network of cells departs from the general trend of size increase in other organs. This is related to a more than proportionate increase in the fiber mass connecting the neurons in larger brains.

In this article, we elaborate on some aspects of connectivity and on their relation to brain size. We focus on two parts of the brain – the cerebral and the cerebellar cortex. A comparison of the two cortices may provide us with general rules underlying brain enlargement, and deviations thereof may hint at the specific requirements of different computational schemes.

Cerebral Cortex

Basic Connectivity

All sensory systems end up in the cerebral cortex. As in some other parts of the brain, such as the thalamus, information from different modalities enters separate regions of the cortex. However, unlike other parts of the brain, an essential feature of the cerebral cortex is its rich internal connectivity, which allows for ample mixing of the signals reaching the cortex from sensory or other systems. This is due to the fact that (1) most cortical neurons are 'interneurons' in the sense that they receive signals from cortical neurons and, in turn, relay them to other cortical neurons, and (2) most cortical neurons are excitatory and therefore capable of relaying excitation from their afferents to the neurons to which they are connected.

Two kinds of excitatory connections can be distinguished in the cortex: most of the excitatory neurons (pyramidal cells) seem to provide both – axon collaterals, which spread in the region in which the cell body is located, and distant connections to another part of the cortex via a main axon which usually runs through the white matter. Thus, there is a continuous, mainly excitatory, network consisting of axons within the cortical gray matter, as well as distant connections via the white matter.

Functional Implications

The self-connectedness of an excitatory system is unique among the major subdivisions of the brain. Functionally, it is suited for representing correlations in the outside world by correlated activity in the cortex and to incorporate these correlations (i.e., 'to learn') into the cortex by strengthening the synapses between neurons which tend to be active together, as postulated by the psychologist Hebb in 1949.

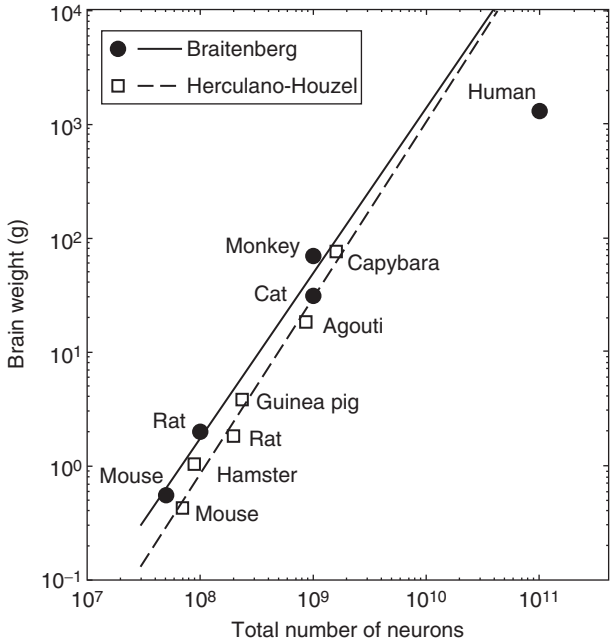

Figure 1 Brain weight as a function of total number of neurons. (All data in this article are plotted in double logarithmic coordinates.) The two data sets are based on different methods but show a similar pattern of behavior: with the exception of the human brain, the average contribution of individual neurons to the volume (weight) of the brain increases at approximately the 3/2 power of number of neurons. Circles: these data were calculated on the basis of density and volume measurements from various sources in the literature. From Braitenberg V (2001) Brain size and number of neurons: An exercise in synthetic neuroanatomy. *Journal of Computational Neuroscience* 10: 71–77. Squares: these data are based on the isotropic fractionator method. From Herculano-Houzel S, Mota B, and Lent R (2006) Cellular scaling rules for rodent brains. *Proceedings of the National Academy of Sciences of the United States of America* 103: 12138–12143.

The actual ability of a cortex to learn all possible kinds of correlations and then combine them in new ways will strongly depend on the degree of its self-connectedness, between both individual neurons and separate cortical regions. In this context, it is important to note that the ramification patterns and the distribution of synapses along the dendritic and axonal arborizations are such that individual pyramidal cells make synapses with thousands of different cortical neurons. In the mouse, the average neuron is connected to approximately 5000 other neurons. This is a small proportion of the approximately 1.6×10^7 neurons in the whole cortex of the mouse (total number in both hemispheres of neo- and allocortex), but it would theoretically suffice to connect each neuron to every other cortical neuron in two steps ($5000 \times 5000 = 2.5 \times 10^7$). In reality, a few more steps are required as has been shown by tracer injection. How is the degree of self-connectedness affected by brain size?

Number of Cortical Neurons and Connectedness

If the number of neurons contacted by an individual neuron remains constant, the fraction of neurons reached by an individual neuron decreases with an increase in the total number of neurons. In reality, this is partially compensated by an increase in the average size of the neurons with brain size – that is, an increase in the total length of axonal and dendritic ramifications and a corresponding increase in the number of synapses per neuron. The evidence for this is as follows: (1) a decrease in the density of neuronal cell bodies in the gray matter with brain size, indicating an increase in the fibers associated with each neuron; (2) no indication of a substantial change in the thickness of the fibers filling the space between the neuronal cell bodies, indicating that the length of cell processes is the main determining factor of volume; and (3) a relatively constant density of synapses in the neuropil. The exact quantitative relationships are not clear since there are few quantitative data on points 2 and 3. Point 1, the density of neurons in brains of different size, has been investigated extensively for more than 50 years but the exact relationship is still a matter of debate. However, most studies report a decrease in the density of neurons (D) with an increase in brain volume (B) of approximately $D \sim B^{-1/3}$(Figure 2). A study on rodent brains, using a method completely different from those of previous studies (Figure 2(a)), corroborates this figure. A decrease in the density of neurons with the $-1/3$ power of brain volume is compatible with an increase in the length of cell processes as the cubic root of brain volume (i.e., as the linear dimensions of the brain).

It should be mentioned, however, that the density of neurons may differ somewhat between orders or families: primates seem to have a higher density than other mammals with similar brain size (Figure 2(b)).

In the motor cortex, the average number of synapses per neuron has been found to be approximately 4.6 times higher in the macaque than in the mouse. Since the macaque has approximately 10^9 cortical neurons, this is not enough to maintain the degree of connectedness at the level of that in the mouse. After two synaptic steps, only approximately 50% of the total number of neurons could be reached from an individual neuron (in the ideal case of maximum divergence). One more synaptic step would be required in the monkey to connect each neuron to every other.

However, a constant degree of connectedness between the individual neurons in different species may not be required for a well-functioning cortex. It might be sufficient to maintain a constant degree of

connectedness between groups of neurons. An increase in the size of the groups would make this possible, as envisaged in the Braitenbergian scheme of square root compartments – a network consisting of N neurons subdivided into $\sqrt{N}$ compartments, each containing $\sqrt{N}$ neurons. Every compartment could then be connected to every other via one neuron. To maintain the same degree of connectedness between compartments in large brains, the number of neurons in each compartment would need to increase at the same rate as the number of compartments (i.e., as the square root of the total number of cortical neurons). It is interesting to note that the volume of the white matter in small and large brains is compatible with this theoretical scheme, indicating that a constant degree of connectedness

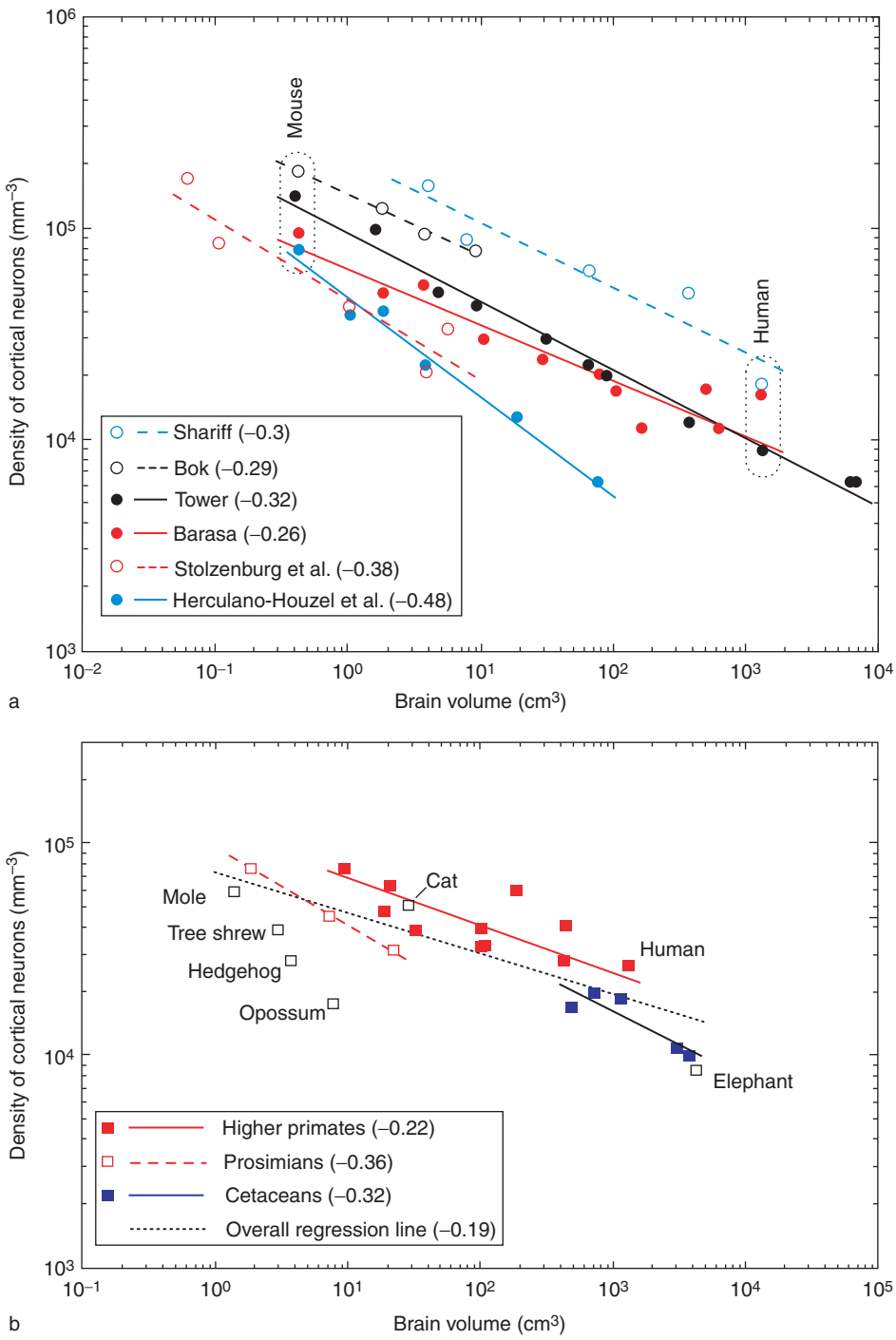

Figure 2 Continued

between groups of neurons might possibly be maintained in large brains.

Horizontal Connectivity of the Cortex and Brain Size

The increase in the volume of the cortical gray matter is mainly due to an increase in surface area. This leads to the folding of the neocortex in large brains. By contrast, the thickness of the cortex tends towards an asymptotic value of only a few millimeters with brain size. How does the large increase in cortical surface area influence the horizontal connectivity of the cortex?

In large brains, the connectivity of the cortex in its tangential plane is served by three kinds of fibers: (1) axon collaterals running tangentially within the cortical gray matter and connecting the cortex in itself up to distance of a few millimeters, (2) the U-fiber system of the white matter which follows the gyri and sulci and connects the cortex in itself up to a few centimeters, and (3) the long-distance fibers which run more deeply in the white matter and connect the various cortical lobes and the two hemispheres to each other. Figure 3 shows the approximate number of fibers in each of these three groups in the human brain. Broadly speaking, the number of fibers in each group is inversely proportionate to their average length. This suggests that they are all part of one system connecting the subdivisions of the cortical plane with each other.

Indeed, in small, lissencephalic brains, a clear-cut division into these three groups cannot be made. In some species, the distances to be bridged hardly exceed a few millimeters. In this case, much of the horizontal connectivity can be accomplished via the gray matter, and the white matter is correspondingly thin. As the number of neurons increases with brain size, both the number and the average length of fibers in the white matter increase. Correspondingly, the volume of the white matter increases considerably more than that of the gray matter with brain size (Figure 4).

It is not clear how these quantitative relations of the white substance in brains of different sizes are related to another observation, namely, that in larger

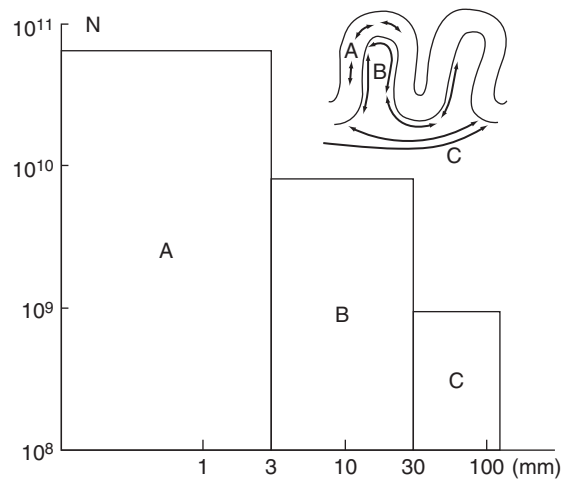

Figure 3 Distribution of the lengths of fibers connecting the human cortex in the horizontal plane. A, number of horizontal axon collaterals running within the gray matter (see inset). B, number of fibers in the peripheral white matter which follows the gyri and sulci (U-fiber system). C, cortico-cortical fibers in the remaining white matter, connecting the cortex over larger distances (including the callosal fibers). There is an inverse relationship between number of fibers in the three categories and their length. From Schüz A and Braitenberg V (2002) The human cortical white matter: Quantitative aspects of cortico-cortical long-range connectivity. In: Schüz A and Miller R (eds.) *Cortical Areas: Unity and Diversity*, pp. 377–385. London: Taylor & Francis.

Figure 2 Density of neurons as a function of brain volume. (a) Measurements by six different authors. The slopes of the individual regression lines are shown in brackets. Most regression lines show a remarkably similar slope around -1/3, despite systematic differences in density between individual studies (see data for mouse and humans). All correlations are significant. The steeper slope in the study by Herculano-Houzel et al. can be explained by the fact that the white substance was included in the density measurements. Data by Shariff refer to the motor cortex of primates, those by Bok to rodents and a rabbit, those by Stolzenburg et al. to insectivores, those by Herculano-Houzel et al. to rodents, and those by Tower and by Barasa to species from a variety of mammalian orders. Tower (from left to right): mouse, rat, guinea pig, rabbit, cat, dog, macaca, beef, man, elephant, and whale. Barasa: mouse, rat, guinea pig, rabbit, cat, dog, sheep and goat, pig, cow, horse, and man. (b) Measurements by one author (Haug) are shown for a particularly large variety of species. Regression lines are shown separately for some taxonomic groups. The overall regression line is flatter than the regression lines in a, but a slope of approximately −1/3 can also be found within cetaceans and prosimians. Higher primates show a higher density of neurons than other groups. (a) Data from Herculano-Houzel S, Mota B, and Lent R (2006) Cellular scaling rules for rodent brains. *Proceedings of the National Academy of Sciences of the United States of America* 103: 12138–12143; Shariff GA (1953) Cell counts in the primate cerebral cortex. *Journal of Comparative Neurology* 98: 381–400; Bok ST (1959) *Histonomy of the Cerebral Cortex*. Amsterdam: Elsevier; Stolzenburg JU, Reichenbach A, and Neumann M (1989) Size and density of glial and neuronal cells within the cerebral neocortex of various insectivorian species. *Glia* 2: 78–84; Tower DB (1954) Structural and functional organization of mammalian cerebral cortex: The correlation of neuron density with brain size. *Journal of Comparative Neurology* 101: 19–51; and Barasa A (1960) Forma, Grandezza e Densita dei Neuroni della Corteccia Cerebrale in Mammiferi di Grandezza Corporea Differente. *Zeitschrift für Zellforschung* 53: 69–89. Reproduced from 'Brain sizes, surfaces, and neuronal sizes of the cortex cerebri: A stereological investigation of man and his variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant),' *American Journal of Anatomy*, vol. 180, 1987, pp. 126–142. Copyright (1987, Wiley-Liss). Reprinted with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

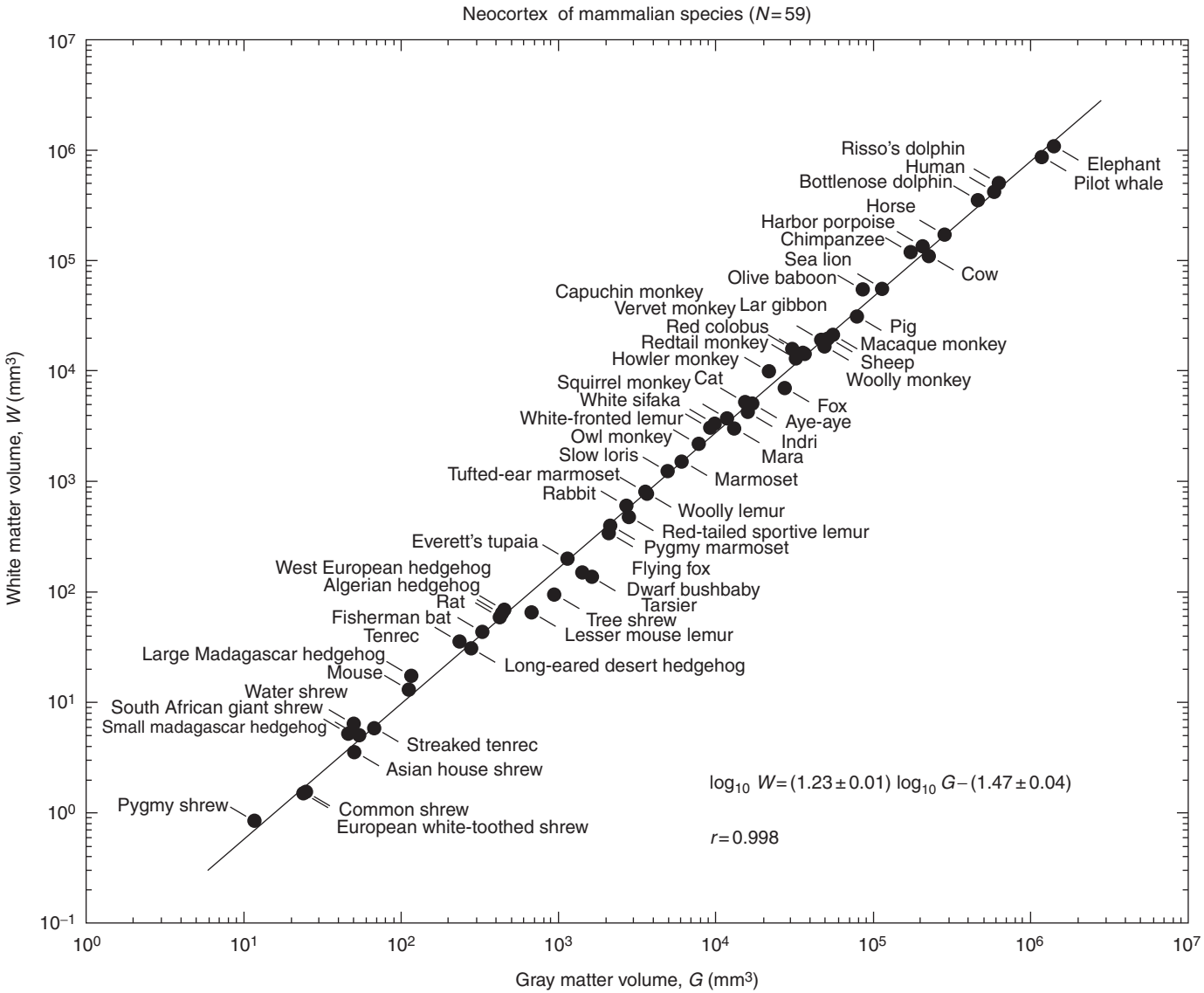

Figure 4 Volume of the cortical white matter plotted against volume of the cortical gray matter. The increase of white matter is more than proportional to that of the gray matter. From Zhang K and Sejnowski TJ (2000) A universal scaling law between gray matter and white matter of cerebral cortex. *Proceedings of the National Academy of Sciences of the United States of America* 97: 5621–5626.

brains, both the size and the number of cortical areas are increased. The number of patches of axonal ramifications observed after a localized tracer injection also seems to increase. By contrast, the size of the patches changes only slightly.

Fiber Thickness and Conduction Time

Conduction times constitute an additional point to be considered in the discussion about connectivity and brain size. Due to the fact that the diameter of the mammalian brain varies between a few millimeters and a few hundred millimeters, conduction times should differ by the same proportion, unless they were compensated by thicker axons and correspondingly higher conduction velocity. However, this is only partially the case. Observations on the corpus callosum of mouse and monkey show that some fibers are indeed several times thicker in the monkey than the thickest fibers in the mouse. However, the majority of fibers have a similar distribution of diameters in both species, with most of the fibers having diameters less than 1 μm. As shown in 1994 by Ringo and associates, keeping conduction times in large brains as short as those in small ones in all fibers is rendered impossible by the fact that thicker axons lead to a further enlargement of brain volume. This, in turn, leads to longer conduction times and the necessity of still thicker axons. According to their calculations, in order to reduce conduction times in the human corpus callosum by 50%, the volume of human brain would have to increase by 50%.

Cerebellar Cortex

Cerebellar Connectivity

The cerebellum is present in all vertebrates but undergoes a remarkable expansion in the orders of birds and mammals. Since most of our knowledge about the cerebellum is derived from mammals, we focus on this animal class. The mammalian cerebellum is composed of a strongly folded cerebellar cortex which engulfs the white matter and the deep cerebellar nuclei. The cerebellar cortex consists of three layers: the granule cell layer adjacent to the white matter, the molecular layer at the cerebellar surface, and, in between these two, a thin layer mainly consisting of a two-dimensional array of large cell bodies, the Purkinje cells. The cerebellar cortex receives two kinds of afferents, mossy and climbing fibers, which mediate a wide variety of vestibular, somatosensory, visual, and auditory sensory modalities, as well as tectal and cerebral cortical afferents. Mossy fibers contact the granule cells that send their axons to the molecular layer where they make T-shaped ramifications, the so-called parallel fibers. The parallel fibers are unmyelinated and form excitatory synapses on their postsynaptic targets. They contact the dendritic trees of the Purkinje cells and the inhibitory interneurons of the molecular layer (the so-called basket and stellate cells) up to a distance of 2 or 3 mm from the site of bifurcation. The climbing fibers, which have their origin in the inferior olive, contact the Purkinje cells directly. The Purkinje cells, which mediate the only output of the cerebellar cortex, project to the deep cerebellar nuclei and to the vestibular nuclei.

Compared to the cerebral cortex, the cerebellar cortex stands out by virtue of its particular geometry. Externally, the parallel arrangement of the folds of the cerebellum is notable, with all of them running in a latero-lateral direction. Internally, the parallel fibers take the same course, with all of them running along the folds. In contrast, the dendritic trees of the Purkinje cells and the dendrites and axons of the basket and stellate cells ramify only in the direction perpendicular to the folds.

The majority of cell types in the cerebellum are inhibitory. However, the granule cells with their parallel fiber axons are the most numerous cells in the cerebellum and are excitatory. In contrast to the excitatory pyramidal cells in the cerebral cortex, the granule cells do not form any contacts with each other. In fact, the division of the cerebellar cortex into two major layers (the granular layer and molecular layer) has the important consequence of keeping the granule cell dendrites and their axons separate (i.e., the dendrites in the granule cell layer and the axons in the molecular layer), thus preventing them from making synapses among themselves.

Functional Implications of Cerebellar Microcircuitry

The microcircuitry of the cerebellar cortex has some important functional implications. First, the cerebellar network is a feed-forward network that hinges on the activity of the afferents that reach the cerebellum. In marked contrast to the cerebral cortex, the lack of any excitatory cerebello-cerebellar connections does not allow for sustained reverberating excitatory loops within the cerebellar cortex. Second, the parallel fibers are thin and unmyelinated and, hence, slow-conducting axons (conduction velocity of 0.3–0.5 $m\,s^{-1}$). With their branches, they contact rows of hundreds of Purkinje cells in succession, with conduction delays depending on their distances from the point of entry of the granule cell axon. Third, the spread of excitation and inhibition matches the anatomical anisotropy described previously: While parallel fibers mediate excitation along the folds, the

basket and stellate cells pass on their inhibition to Purkinje cells in a direction across the folds. The axons of these inhibitory cells are also short ranged and do not exceed 1 mm. In summary, the cerebellar circuitry performs operations on its inputs that are short ranged both in space (a few millimeters) and in time (approximately 10 ms). The short range of interaction is certainly an important feature of the computations for which the cerebellum is responsible.

Scaling of the Cerebellum with Brain Size

The surface of the cerebellum increases in parallel with the surface of the cerebral cortex (Figure 5). However, whereas the volume of the cerebral neocortex (including its white matter) increases during mammalian evolution from 15% to 75% of total brain volume, the volume of the cerebellum remains relatively constant (10–20%). Despite this lower contribution to brain volume, if the number of neurons is considered, cerebellar neurons actually comprise the majority of neurons in the brain (60–80% of all neurons from mouse to human). This discrepancy is due to two factors. First is the high density of cerebellar neurons. The density of cerebellar neurons (i.e., the granule cells) is approximately 10 times higher than the density of neurons in the cerebral cortex (e.g., in a small rodent, 10^6 granule cells mm^{-3} compared to approximately 10^5 cortical neurons mm^{-3}). Furthermore, the granule cell density is remarkably similar in brains of different size (Figure 6), in contrast to the decrease in the density of cerebral cortex neurons in larger brains. Second, the relative increase in the total volume of the cerebral cortex with brain size is largely

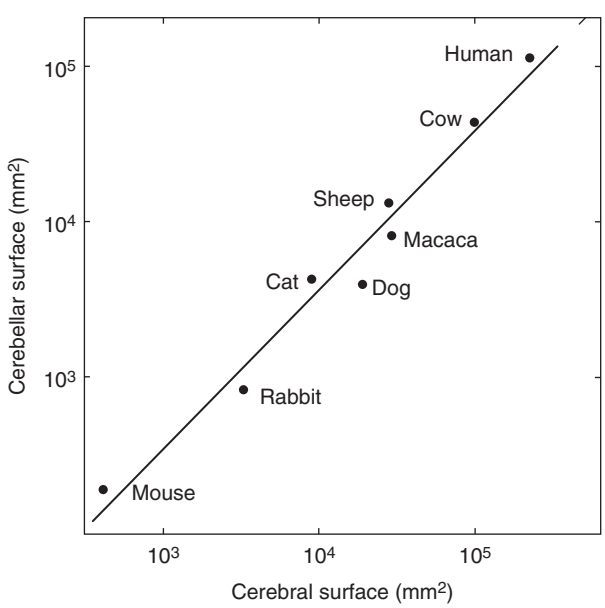

Figure 5 Proportional increase in the surfaces of the cerebellar and the cerebral cortex. The regression line has the following equation: $\log Y = 1.02 \log X - 0.52$ ($r = 0.95$, $p < 0.0003$). From Sultan F (2002) Brain evolution: Analysis of mammalian brain architecture. *Nature* 415: 133–134.

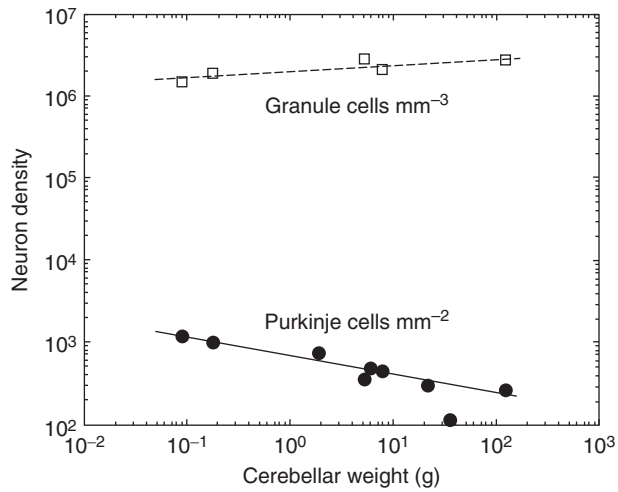

Figure 6 Density of granule and Purkinje cells in relation to cerebellar weight. The density of granule cells mm^{-3} does not show a significant dependence on brain size in these data (a slight decrease with brain size has been reported). By contrast, the density of Purkinje cells mm^{-2} shows a significant decrease with cerebellar weight, with a slope of approximately $-1/3$. The data for both kinds of neurons range from mouse to man and are taken from several sources. The densities of granule cells are from Smolyaninov VV (1971) Some special features of organization of the cerebellar cortex. In: Gelfand IM, Gurfinkel VS, and Fornin SV (eds.) *Models of the Structural–Functional Organization of Certain Biological Systems*, pp. 251–325. Cambridge, MA: MIT Press; Harvey RJ and Napper RMA (1991) Quantitative studies on the mammalian cerebellum. *Progress in Neurobiology* 36: 437–463; Palkovits M, Magyar P, and Szentagothai J (1971) Quantitative histological analysis of the cerebellar cortex in the cat: I. Number and arrangement in space of Purkinje cells. *Brain Research* 32: 1–13; and Andersen BB, Gundersen HJ, and Pakkenberg B (2003) Aging of the human cerebellum: A stereological study. *Journal of Comparative Neurology* 466: 356–365. The data for the Purkinje cells were obtained from Fan H, Favero M, and Vogel MW (2001) Elimination of Bax expression in mice increases cerebellar Purkinje cell numbers but not the number of granule cells. *Journal of Comparative Neurology* 436: 82–91; Harvey RJ and Napper RMA (1991) Quantitative studies on the mammalian cerebellum. *Progress in Neurobiology* 36: 437–463; Mwamengele GL, Mayhew TM, and Dantzer V (1993) Purkinje cell complements in mammalian cerebella and the biases incurred by counting nucleoli. *Journal of Anatomy* 183: 155–160; Mlonyeni M (1973) The number of Purkinje cells and inferior olivary neurones in the cat. *Journal of Comparative Neurology* 147: 1–10; and Fox CA, Siegesmund KA, and Dutta CR (1964) The Purkinje cell dendritic branchlets and their relation with the parallel fibres: Light and electron microscopic observations. In: Cohen MM and Snider RS (eds.) *Morphological and Biochemical Correlates of Neural Activity*, pp. 112–141. New York: Harper & Row. The data by Mwengele et al. concern total numbers of Purkinje cells. Densities were obtained by dividing these numbers by total cerebellar surface, taken from Sultan F and Braitenberg V (1993) Shapes and sizes of different mammalian cerebella. A study in quantitative comparative neuroanatomy. *Journal für Hirnforschung* 34: 79–92.

due to the strong increase in its white matter. By contrast, cerebellar white matter remains an almost constant fraction of brain volume (Figure 7).

The size and shape of granule cells are remarkably conserved across phylogeny. Granule cell soma size does not change much with brain size, in contrast to many other neuron types. The dendrites and axons of these cells are also remarkably conserved across species. The axons of the granule cells, the parallel fibers, have a constant length in different species (chicken, rat, monkey, and human) and a similar diameter. The constant size of granule cells is therefore probably an important feature related to the computations which the cerebellum performs.

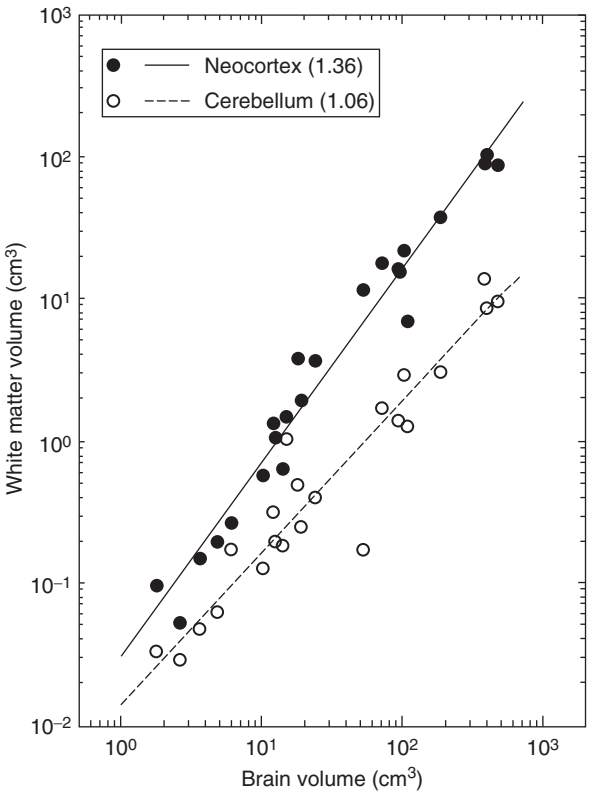

Figure 7 The white matter of neocortex and cerebellum show different scaling to brain volume. The white matter of the neocortex increases more than proportionally with brain size, probably reflecting the necessity for maintaining cortico-cortical connectivity with the increase in neuron number. In contrast, cerebellar white matter comprises solely input and output fibers, which may be expected to increase proportionally with the number of cerebellar neurons. White matter volumes were obtained from Bush EC and Allman JM (2003) The scaling of white matter to gray matter in cerebellum and neocortex. *Brain, Behavior and Evolution* 61: 1–5. Total brain volumes were obtained from Brauer K and Schober W (1970) *Katalog der Säugetiergehirne*. Jena, Germany: VEB Gustav Fischer; Hofman MA (1988) Size and shape of the cerebral cortex in mammals: II. The cortical volume. *Brain, Behavior and Evolution* 32: 17–26; and Stephan H, Frahm H, and Baron G (1981) New and revised data on volumes of brain structures in insectivores and primates. *Folia Primatologica* 35: 1–29.

In contrast to the granule cells, Purkinje cell density does scale with brain size (Figure 6). The number of Purkinje cells per unit area of cortex (and therefore, to a first approximation, also per volume of cortical gray substance) decreases with brain size, with an exponent of approximately −1/3. This is similar to the decrease in neocortical neuron density described previously. It correlates with an increase in the size of the Purkinje cells with brain size, which may have to do with the fact that the Purkinje cells are the output neurons of the cerebellar cortex (i.e., they have to project over larger distances in larger brains).

The particular microcircuitry of the cerebellar cortex is also reflected in the anisotropic shape of the unfolded cerebellar surface and in the different scaling of cerebellar width and anterior–posterior length, as shown in Figure 8. Cerebellar volume increase accompanying brain size increase is mainly effected by an increase in anterior–posterior length.

Summary of Cerebral and Cerebellar Scaling

The most significant feature that the two parts of the brain have in common is their similar increase in surface area with brain size. This may partially reflect the fact that the two cortices are connected with each other both ways (via pontine nuclei and thalamus, respectively).

The main differences between the two parts of the brain are as follows:

1. Whereas in larger brains, the cerebral cortex increases in both dimensions of its surface, the cerebellar cortex increases much more in its anterior–posterior than in its latero-lateral extension. This reflects the internal isotropy of the cerebral cortex as opposed to the internal anisotropy of the cerebellar cortex. In the case of the cerebral cortex, this makes sense in connection with the two-dimensional topographic representation of space via the sensory surfaces and equal operations on them in both dimensions. In the case of the cerebellum, the spread of activity is a one-dimensional affair, suggesting a role in time phenomena. In larger cerebellar cortices, it seems to be more important to add new excitatory beams rather than to increase their length.

2. The granule cells in the cerebellar cortex are remarkably constant in the size of their cell bodies and length of their cell processes. This indicates a roughly constant number of synapses per granule cell with brain size. By contrast, the neurons in the cerebral cortex have, on average, a larger number of synapses in larger brains.

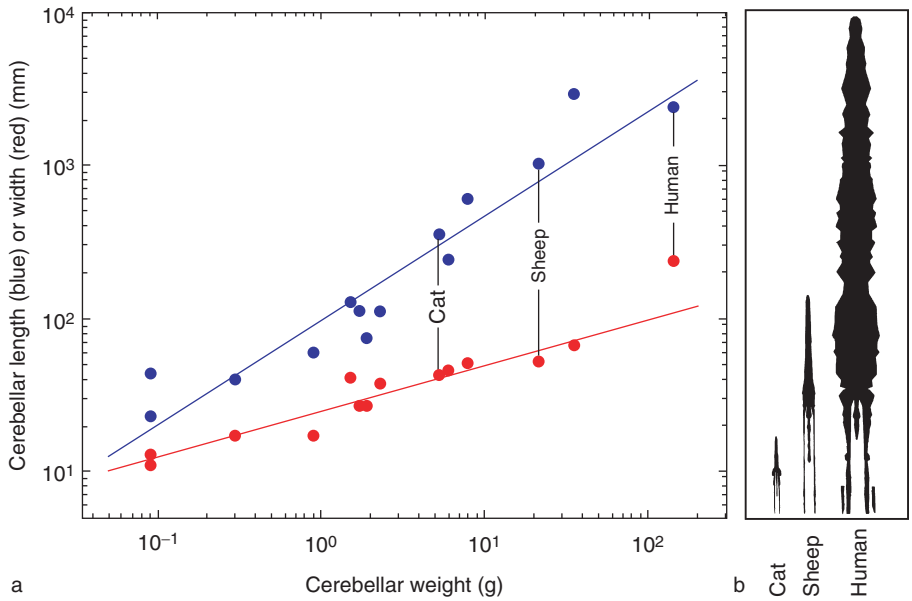

Figure 8 (a) Length and width of the unfolded cerebellar surface depend in different ways on the size of the cerebellum. Cerebellar width scales with the 1/3 power, whereas cerebellar length scales with the 2/3 power of cerebellar volume. (b) Three unfolded cerebella of a cat, sheep, and human are shown. The width of the cerebellum corresponds to the direction of the parallel fibers, whereas cerebellar length corresponds to the rostrocaudal extension of the unfolded cerebellar cortex. Data from Sultan F and Braitenberg V (1993) Shapes and sizes of different mammalian cerebella. A study in quantitative comparative neuroanatomy. *Journal für Hirnforschung* 34: 79–92.

3. One of the most striking changes with brain size is the strong relative increase in the white matter of the cerebral cortex. Although the white matter of the cerebellar cortex also increases slightly more than its gray matter with brain size, it remains a nearly constant fraction of the whole brain. This difference between the two cortices reflects the important role of cortico-cortical connections in the case of the cerebral cortex and their complete absence in the white matter of the cerebellar cortex.

See also: Allometric Analysis of Brain Size; Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brain Scaling Laws; Cerebellum: Evolution and Comparative Anatomy; Cerebral Cortex.

Further Reading

Andersen BB, Gundersen HJ, and Pakkenberg B (2003) Aging of the human cerebellum: A stereological study. *Journal of Comparative Neurology* 466: 356–365.

Barasa A (1960) Forma, Grandezza e Densita dei Neuroni della Corteccia Cerebrale in Mammiferi di Grandezza Corporea Differente. *Zeitschrift für Zellforschung* 53: 69–89.

Bok ST (1959) *Histonomy of the Cerebral Cortex*. Amsterdam: Elsevier.

Braitenberg V, Heck D, and Sultan F (1997) The detection and generation of sequences as a key to cerebellar function: Experiments and theory. *Behavioral & Brain Sciences* 20: 229–245.

Braitenberg V and Schüz A (1998) *Cortex: Statistics and Geometry of Neuronal Connectivity*, 2nd edn. Berlin: Springer.

Brauer K and Schober W (1970) Katalog der Säugetiergehirne. Jena, Germany: VEB Gustav Fischer.

Bugbee NM and Goldman-Rakic PS (1983) Columnar organization of cortico-cortical projections in squirrel and rhesus monkeys: Similarity of column width in species differing in cortical volume. *Journal of Comparative Neurology* 220: 355–364.

Bush EC and Allman JM (2003) The scaling of white matter to gray matter in cerebellum and neocortex. *Brain, Behavior and Evolution* 61: 1–5.

Cragg BG (1967) The density of synapses and neurons in the motor and visual areas of the cerebral cortex. *Journal of Anatomy* 101: 639–654.

DeFelipe J, Alonso-Nanclares L, and Arellano JI (2002) Microstructure of the neocortex: Comparative aspects. *Journal of Neurocytology* 31: 299–316.

Fan H, Favero M, and Vogel MW (2001) Elimination of Bax expression in mice increases cerebellar Purkinje cell numbers but not the number of granule cells. *Journal of Comparative Neurology* 436: 82–91.

Fox CA, Siegesmund KA, and Dutta CR (1964) The Purkinje cell dendritic branchlets and their relation with the parallel fibres: Light and electron microscopic observations. In: Cohen MM and Snider RS (eds.) *Morphological and Biochemical Correlates of Neural Activity*, pp. 112–141. New York: Harper & Row.

Frahm HD, Stephan H, and Stephan M (1982) Comparison of brain structure volumes in Insectivora and primates: I. Neocortex. *Journal für Hirnforschung* 23: 375–389.

Harvey RJ and Napper RMA (1991) Quantitative studies on the mammalian cerebellum. *Progress in Neurobiology* 36: 437–463.

Haug H (1987) Brain sizes, surfaces, and neuronal sizes of the cortex cerebri: A stereological investigation of man and his-variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant). *American Journal of Anatomy* 180: 126–142.

Herculano-Houzel S, Mota B, and Lent R (2006) Cellular scaling rules for rodent brains. *Proceedings of the National Academy of Sciences of the United States of America* 103: 12138–12143.

Hofman MA (1985) Size and shape of the cerebral cortex in mammals: II. The cortical surface. *Brain, Behavior and Evolution* 27: 28–40.

Hofman MA (1988) Size and shape of the cerebral cortex in mammals: II. The cortical volume. *Brain, Behavior and Evolution* 32: 17–26.

Hutsler JJ, Lee D-G, and Porter KK (2005) Comparative analysis of cortical layering and supragranular layer enlargement in rodent, carnivore and primate species. *Brain Research* 1052: 71–81.

Jerison HJ (1973) *Evolution of the Brain and Intelligence*. New York: Academic Press.

Jerison HJ (1991) *Brain Size and the Evolution of Mind*. New York: American Museum of Natural History.

Kaas JH (2000) Why is brain size so important: Design problems and solutions as neocortex gets bigger or smaller. *Brain and Mind* 1: 7–23.

Mlonyeni M (1973) The number of Purkinje cells and inferior olivary neurones in the cat. *Journal of Comparative Neurology* 147: 1–10.

Mwamengele GL, Mayhew TM, and Dantzer V (1993) Purkinje cell complements in mammalian cerebella and the biases incurred by counting nucleoli. *Journal of Anatomy* 183: 155–160.

Palkovits M, Magyar P, and Szentagothai J (1971) Quantitative histological analysis of the cerebellar cortex in the cat: I. Number and arrangement in space of Purkinje cells. *Brain Research* 32: 1–13.

Palm G (1986) Associative networks and cell assemblies. In: Palm G and Aertsen A (eds.) *Brain Theory*, pp. 211–228. Berlin: Springer.

Prothero J (1997) Scaling of cortical neuron density and white matter volume in mammals. *Journal of Brain Research* 38(4): 513–524.

Ringo JL, Doty RW, Demeter S, and Simard PY (1994) Time is of the essence: A conjecture that hemispheric specialization arises from interhemispheric conduction delay. *Cerebral Cortex* 4: 331–343.

Schüz A and Braitenberg V (2002) The human cortical white matter: Quantitative aspects of cortico-cortical long range connectivity. In: Schüz A and Miller R (eds.) *Cortical Areas: Unity and Diversity*, pp. 377–385. London: Taylor & Francis.

Schüz A and Demianenko GP (1995) Constancy and variability in cortical structure: A study on synapses and dendritic spines in hedgehog and monkey. *Journal of Brain Research* 36: 113–122.

Schüz A and Preißl H (1996) Basic connectivity of the cerebral cortex and some considerations on the corpus callosum. *Neuroscience and Biobehavioral Reviews* 20(4): 567–570.

Shariff GA (1953) Cell counts in the primate cerebral cortex. *Journal of Comparative Neurology* 98: 381–400.

Stephan H, Frahm H, and Baron G (1981) New and revised data on volumes of brain structures in insectivores and primates. *Folia Primatologica* 35: 1–29.

Stolzenburg JU, Reichenbach A, and Neumann M (1989) Size and density of glial and neuronal cells within the cerebral neocortex of various insectivorian species. *Glia* 2: 78–84.

Smolyaninov VV (1971) Some special features of organization of the cerebellar cortex. In: Gelfand IM, Gurfinkel VS, and Fornin SV (eds.) *Models of the Structural–Functional Organization of Certain Biological Systems*, pp. 251–325. Cambridge, MA: MIT Press.

Sultan F (2002) Brain evolution: Analysis of mammalian brain architecture. *Nature* 415: 133–134.

Sultan F and Braitenberg V (1993) Shapes and sizes of different mammalian cerebella. A study in quantitative comparative neuroanatomy. *Journal für Hirnforschung* 34: 79–92.

Tower DB (1954) Structural and functional organization of mammalian cerebral cortex; The correlation of neuron density with brain size. *Journal of Comparative Neurology* 101: 19–51.

Zhang K and Sejnowski TJ (2000) A universal scaling law between gray matter and white matter of cerebral cortex. *Proceedings of the National Academy of Sciences of the United States of America* 97: 5621–5626.

Brain Damage: Functional Reorganization

J Grafman, R Zahn, and E Wassermann, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

Recent research in cognitive neuroscience has made great strides in assigning behavioral functions to discrete areas of the brain and describing the knowledge, representational elements, and processes that are subserved by the tissue within their borders. Although much work remains to be done to provide a complete functional atlas of the thinking brain, one fact is generally accepted about how representational knowledge is stored. That is, the brain appears to be composed of modular networks within which a particular representational element is homogeneously represented. Such elements may range in complexity from the edge detectors, used in visual processing and stored in the occipital cortex, to high-level plan representations used to guide behavior and stored in the prefrontal cortex. To perform a cognitive or perceptual function, sets of modules are activated in concert. Thus, the human brain can be subdivided into functional modules, which must be combined to perform a task.

Neuroplasticity is a term used to describe structural and functional changes at the genetic, molecular, neuronal, system, and behavioral levels. Such changes may reflect new experience and learning or can occur as a reaction to brain damage. Although a great deal of scientific literature describes neuroplasticity at the molecular, structural, and system levels in animals, we focus on neuroplastic changes at the system level that can be observed in humans. Current thinking about systems neuroplasticity suggests that at least four kinds of neuroplastic changes operate at the representational module level after brain damage: (1) homologous area adaptation, (2) cross-modal reassignment, (3) potentiation of topographic representations, and (4) compensatory masquerade.

Homologous Area Adaptation

One form of neuroplasticity, homologous area adaptation, appears to prevail before children reach puberty and underlies the observation that damage to a particular brain region can be compensated for by shifting its functions to other brain areas, often to the homologous region of the opposite hemisphere. An implication of this mechanism is that the brain area accepting the new operation will be more 'crowded' with functional representations (both the new and the 'usual' representations are now stored in a region that before stored only the usual representation). This crowding leads to a sparser or 'lower resolution' representation of knowledge or sensory input within both modules and makes degraded performance or interference likely when the two functions are activated simultaneously. This form of neuroplastic change is reported less often in adults. This may be because regions of the developing brain are less 'committed' to specific cognitive processes, because they have not yet been encountered in school (e.g., mathematics) or demanded by the environment (e.g., complex social interactions).

We studied an adolescent who had incurred a severe right parietal lobe brain injury as a young child. Despite the severity and location of the injury, evaluation showed that the boy had developed essentially normal visuospatial function but had impaired arithmetic skills. We inferred that at the time of injury, the left parietal region assumed some of the functions normally stored in the right parietal lobe. Because most calculation skill is acquired in school, the injury and plastic change occurred before the age of arithmetic acquisition. Thus, we argued, spatial processes had claimed the left parietal region before arithmetic instruction, making it more difficult for the patient to learn and store mathematical facts. Functional magnetic resonance imaging (fMRI) of his brain activity during arithmetic processing indicated that he activated the left parietal lobe (among other regions). This finding implied that that the left parietal lobe was still programmed to store arithmetic facts even though it was now also committed to spatial processing, leading to the degradation of the subject's arithmetic ability.

Some investigators have claimed that the proportion of a functional region that is damaged determines the amount of homologous region adaptation that occurs. The underlying principle is as follows. Neighboring and homologous cortical regions have primary and secondary, or latent, functional assignments. Usually the secondary function is inhibited competitively by output from the nearby or contralateral brain region where that function is primary. Therefore, cortical regions are only able to express their secondary functions when competitive inhibition is removed. Given this logic, plasticity might be promoted by incurring complete rather than partial damage to a particular brain region, because when the damaged region is completely incapacitated, it cannot

inhibit the expression of its usual function elsewhere in the brain.

We studied another patient who suffered a large infarction that destroyed most of the left cerebral hemisphere, although he had some spared and functionally active islands of tissue in the left parietal and frontal areas. This patient could read words but not nonwords, and he had great difficulty calculating. On fMRI, word reading activated a broadly distributed network of areas in the right hemisphere. Attempts to read nonwords were unsuccessful and did not produce significant activation in either hemisphere. This suggested that the right hemisphere had assumed some functions of the left hemisphere. However, although the left hemisphere substrate for reading known words was destroyed, allowing the right hemisphere to assume its function, the additional ability to read nonwords (a crucial skill reflecting the ability to learn to read new words) did not transfer and was lost with the injury. This indicates that functional neuroplasticity may be more limited in adults. Interestingly, during fMRI scanning, while the patient was performing arithmetic fact verification, both right and left parietal lobe activation was found. However, the left parietal activation was found in locations around the periphery of the area in the brain devoted to this ability. Even with this pattern, the patient performed poorly, although his accuracy increased slightly after a week of training with a concomitant increase in cortical activation. We suspect that the remaining left parietal tissue was inadequate for accurate calculation but still inhibited the right parietal cortex from assuming more of a role in calculation.

Cross-Modal Reassignment

A second form of neuroplasticity, cross-modal reassignment, involves the introduction of new sensory inputs to a brain region that has been deprived of its main sensory input. For example, positron emission tomography (PET) and fMRI studies of tactile discrimination have shown that adults who became blind early in childhood have somatosensory input to area V1 of the occipital cortex, whereas sighted controls do not. Presumably, in the blind, such input succeeds in activating the representations stored in area V1, because the cognitive operations in that area are somewhat independent of the modality of input, and their function is helpful in making the discriminations between geometric patterns required for reading Braille characters. In this study, only tactile discrimination of Braille letters, as opposed to simply palpating fields of nonsense dots or performing verbal language tasks, activated V1 in the early blind. In contrast, in sighted people exposed to the same stimuli and tasks, the V1 activation was absent and there was evidence of decreased activation in area V1 (suggesting that recruiting the brain regions that usually process the tactile input (e.g., parietal cortex) ordinarily inhibits or depotentiates the visual system). The fact that visual cortex activation is important in Braille reading and may account for the superior ability of blind subjects was demonstrated in a study in which transcranial magnetic stimulation (TMS) was delivered over the occipital cortex. This stimulation, designed to interfere with cortical processing, dramatically impaired the performance of the early blind but not sighted or late-life blinded individuals. The fact that such cross-modal sensory plasticity is possible and requires an early lesion suggests that afferent connections from multiple sensory sources may be latent in cortical areas such as V1, which is considered by most clinicians to be rigidly assigned to visual processing. Presumably these secondary inputs atrophy over time under the competitive influence of the primary input. Undoubtedly, limits exist to this form of neuroplasticity. For example, color processing cells in the occipital cortex are so specialized for visual input that they would be unlikely to accept other forms of sensory input.

Potentiation of Topographic Representations

This form of neuroplasticity involves changes within brain regions devoted to a particular kind of knowledge or operation. Studies in the motor cortex indicate that the intracortical connections devoted to a processing function may become temporarily enhanced or even enlarged with skill acquisition or frequent exposure to a stimulus. This enhancement may, in turn, result in topographic expansion of the area of cortex devoted to the function or representation. In two studies using two different techniques (TMS and electroencephalography (EEG)), we demonstrated that repetition of a cued sequence of key presses caused increasing excitability of neurons in the representation of the performing hand in the early stages of learning while the pattern was still implicit (i.e., the subjects did not become aware of the pattern, but their responses were becoming faster). When learning became explicit (i.e., the subjects could produce it without cueing), the excitability of the hand representation returned to baseline. This type of change, measured as a larger EEG potential before movement and a larger muscle response to TMS of the hand representation, could result from topographic expansion of this representation by unmasking of latent connections within the cortex or from the cortex to the spinal cord. Although EEG

and TMS cannot measure motor map size, changes seen in fMRI studies during, and even weeks after, motor learning seem to bear this out. Evidence suggests that rapid enlargement of cortical maps can persist in individuals who develop or are trained in a particular skill that they use routinely. The meaning of map expansion is still unclear. One possible explanation for a change in map size is that continual input may overload the processing capacity of a local cortical network and signal it to expand by recruiting new neurons from adjacent tissue into the network to handle the increased load. Another possibility is that before enough synaptic change has occurred for the minimal necessary volume of cortex to perform the skill, an enlarged area of the network is recruited. With increasing practice, the system becomes more and more efficient, restricting activity to a few highly dedicated and stimulus specific neural elements.

Compensatory Masquerade

The fourth form of neuroplasticity, compensatory masquerade, means the novel use of an established and intact cognitive process to perform a task previously dependent on an impaired cognitive process. This can be a subtle change detectable only with fine-grained cognitive tasks. For example, there may be a single route from home to the office. One way to navigate that route may depend most upon spatial processes, is implicitly stored, and is performed rapidly. Another way to get to work would depend on verbal labeling of landmarks that are remembered explicitly and processed more deliberately. A large brain injury may affect both way-finding processes, but typically one is relatively spared. The patient may then be able to use the spared strategy to navigate the same route that was previously dependent on the impaired strategy. Unless a neuropsychologic study carefully evaluated both processes in some detail, an observer could be misled into thinking that a more fundamental form of neuroplasticity had occurred (e.g., homologous area adaptation).

Discussion

We have reviewed four major forms of neuroplasticity that can be studied and promoted in healthy human volunteers and patients after brain damage. The neurophysiologic processes that underlie plasticity continue to be studied at the level of synaptic physiology and ultrastructure after repeated stimulation.

Recently it has become clear that neurogenesis, synapse formation, and remodeling of axons and dendrites occur in adult human brains. Neurogenesis in adulthood appears to occur in both the hippocampus and subventricular zone. It is unclear what functional role new neurons play in previously existing networks. It would appear that neurogenesis and cortical rewiring would be ideal mechanisms to assist brain regions partially damaged by stroke and other lesions to recover their functions. However, because recovery of function after brain injury is rarely if ever complete, their contributions to recovery of function appear to be restricted. The neuroanatomical, physiological, and molecular bases of these restrictions to plasticity are the focus of intense investigation.

By focusing on the four forms of plasticity previously outlined, fundamental questions about how functional cooperation between brain regions is achieved can be asked. Why does the ability to shift functions between sites decline over development and aging? This may be the result of the necessity to stabilize connectivity between maturing brain areas dedicated to particular functions. The increasing degree and stability of this connectivity suggests that the networks themselves make a commitment to a particular functional representation. If functional substitution was permitted beyond the early stages of the formation of these networks, then continuous and substantial reorganization of distributed networks would be required for their efficient operation. Therefore, it is likely that the increasing connectivity required for learning is directly related to the increasing functional commitment of brain areas throughout life.

A second question about plasticity emerges from the observation of cross-modal activation of V1 in the early blind. Traditional notions about deafferentation lead one to expect atrophy and or 'denervation hypersensitivity' of a maladaptive kind in the visual areas of the blind. Instead, evidence suggests that the modality-independent processing capabilities of the visual cortex (e.g., discrimination of width, angles, feature conjunctions) can be used to process spatial information from more than one source. This is a remarkable departure from the idea that cortical functions are assigned to areas entirely based on input routing rather than processing architecture.

Clinical approaches to enhancing recovery of cognitive function in patients recovering from brain injury have historically been implemented in an atheoretical way. Early approaches relied simply on massed repetitive practice in various domains of function (e.g., memory, attention). More recent efforts at rehabilitation have begun to use treatments motivated by cognitive theory. Despite this increasing sophistication, there have been too few clinical trials that included large numbers of subjects or adequate controls, and most rehabilitation studies have compared programs of stimulation to supposedly neutral

interactions or to no intervention at all rather than investigating the validity of specific interventions on specific outcomes. Furthermore, the therapeutic strategies used in these studies were not always developed with systems and cognitive neuroscience findings in mind. Although some recent counterexamples exist, the rehabilitation field has often lagged behind other specialties in clinical neuroscience in adapting laboratory findings to the clinic.

Fortunately, many conceptual frameworks for neuroplasticity exist that could lead to new rehabilitation methods. We suggest that proven methods for inducing changes in the cortical topography at the network level and rerouting or unmasking latent input and output to the affected network should guide the development of new rehabilitation techniques. Testing of paradigms designed to cause and manipulate neuroplasticity in healthy volunteers and patients will help investigators find the optimal conditions for neuroplastic change. Such paradigms may also lead to 'challenge' studies using pharmacologic and other agents to enhance or prevent neuroplastic changes. Indeed, the future looks promising as we probe the limits of functional neuroplasticity in adult humans while answering fundamental questions about the stable and dynamic topography of information processing.

See also: Adult Cortical Plasticity; Axonal Injury: Neuronal Responses; Brain Injury: Functional Recovery After; Brain Injury: Magnetic Resonance Studies of Metabolic Aspects; Brain Trauma; Brain Asymmetry: Evolution; Neuronal Plasticity after Cortical Damage.

Further Reading

Aglioti S, Bonazzi A, and Cortese F (1994) Phantom lower limb as a perceptual marker of neural plasticity in the mature human brain. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 255(1344): 273–278.

Buonomano DV, Hickmott PW, and Merzenich MM (1997) Context-sensitive synaptic plasticity and temporal-to-spatial transformations in hippocampal slices. *Proceedings of the National Academy of Sciences of the United States of America* 94(19): 10403–10408.

Chklovskii DB, Mel BW, and Svoboda K (2004) Cortical rewiring and information storage. *Nature* 431: 782–788.

Chugani HT, Muller RA, and Chugani DC (1996) Functional brain reorganization in children. *Brain Development* 18(5): 347–356.

Cohen LG, Celnik P, Pascual-Leone A, et al. (1997) Functional relevance of cross-modal plasticity in blind humans. *Nature* 389(6647): 180–183.

Debanne D (1996) Associative synaptic plasticity in hippocampus and visual cortex: Cellular mechanisms and functional implications. *Reviews in Neuroscience* 7(1): 29–46.

Donoghue JP (1995) Plasticity of adult sensorimotor representations. *Current Opinion in Neurobiology* 5(6): 749–754.

Gilbert CD (1996) Plasticity in visual perception and physiology. *Current Opinion in Neurobiology* 6(2): 269–274.

Goldman S and Plum F (1997) Compensatory regeneration of the damaged adult human brain: Neuroplasticity in a clinical perspective. *Advances in Neurology* 73: 99–107.

Gould E, Reeves AJ, Graziano MSA, and Gross CG (1999) Neurogenesis in the neocortex of adult primates. *Science* 286: 548–552.

Grafman J and Litvan I (1999) Evidence for four forms of neuroplasticity. In: Grafman J and Christen Y (eds.) *Neuronal Plasticity: Building a Bridge from the Laboratory to the Clinic*, pp. 131–140. Berlin: Springer.

Heiss WD and Teasel RW (2006) Brain recovery and rehabilitation. *Stroke* 37: 314–316.

Ide CF, Scripter JL, Coltman BW, et al. (1996) Cellular and molecular correlates to plasticity during recovery from injury in the developing mammalian brain. *Progress in Brain Research* 108: 365–377.

Karni A, Meyer G, Rey-Hipolito C, et al. (1998) The acquisition of skilled motor performance: Fast and slow experience-driven changes in primary motor cortex. *Proceedings of the National Academy of Sciences of the United States of America* 95: 861–868.

Keyvani K and Schallert T (1998) Plasticity-associated molecular and structural events in the injured brain. *Journal of Neuropathology and Experimental Neurology* 61(10): 831–840.

Klintsova AY and Greenough WT (1998) Synaptic plasticity in cortical systems. *Current Opinion in Neurobiology* 9: 203–208.

Kolb B, Gibb R, and Robinson TE (1998) Brain plasticity and behavior. *Current Directions in Psychological Science* 12(1): 1–5.

Levin HS, Scheller J, Rickard T, et al. (1998) Dyscalculia and dyslexia after right hemisphere injury in infancy. *Archives of Neurology* 53(1): 88–96.

Leviton A, Bellinger D, Pagano M, and Rabinowitz M (1998) Models of delayed recovery. *Journal of Child Neurology* 10 (5): 385–391.

Lledo PM, Alonso M, and Grubb MS (1998) Adult neurogenesis and functional plasticity in neuronal circuits. *Nature Reviews Neuroscience* 7: 179–193.

Melzack R, Israel R, Lacroix R, and Schultz G (1998) Phantom limbs in people with congenital limb deficiency or amputation in early childhood. *Brain* 120(Pt 9): 1603–1620.

Merzenich M, Wright B, Jenkins W, et al. (1998) Cortical plasticity underlying perceptual, motor, and cognitive skill development: Implications for neurorehabilitation language comprehension in language-learning impaired children improved with acoustically modified speech. *Cold Spring Harbor Symposium on Quantitative Biology* 61(5245): 1–8.

Merzenich MM, Jenkins WM, Johnston P, et al. (1998) Temporal processing deficits of language-learning impaired children ameliorated by training. *Science* 271(5245): 77–81.

Pascual-Leone A, Grafman J, and Hallett M (1998) Modulation of cortical motor output maps during development of implicit and explicit knowledge. *Science* 263(5151): 1287–1289.

Plaut DC (1998) Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language* 52(1): 25–82.

Rauschecker JP (1998) Mechanisms of compensatory plasticity in the cerebral cortex. *Advances in Neurology* 73: 137–146.

Romero SG, Manly CF, and Grafman J (1998) Investigating cognitive neuroplasticity in single cases: Lessons learned from applying functional neuroimaging techniques to the traditional neuropsychological case study framework. *Neurocase* 8(5): 355–368.

Rosenzweig MR and Bennett EL (1998) Psychobiology of plasticity: Effects of training and experience on brain and behavior. *Behavioral Brain Research* 78(1): 57–65.

Sadato N, Pascual-Leone A, Grafman J, et al. (1998) Activation of the primary visual cortex by Braille reading in blind subjects. *Nature* 380(6574): 526–528.

Seitz RJ and Freund HJ (1998) Plasticity of the human motor cortex. *Advances in Neurology* 73: 321–333.

Taub E, Uswatte G, and Elbert T (1998) New treatments in neurorehabilitation founded on basic research. *Nature Reviews Neuroscience* 3(3): 228–236.

Ward NS (1998) Neural plasticity and recovery of function. *Progress in Brain Research* 150: 527–535.

Wood JN and Grafman J (1998) Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews Neuroscience* 4(2): 139–147.

Zhuang P, Toro C, Grafman J, et al. (1998) Event-related desynchronization (ERD) in the alpha frequency during development of implicit and explicit learning. *Journal of Electroencephalography and Clinical Neurophysiology* 102(4): 374–381.

Brain Development: The Generation of Large Brains

S Herculano-Houzel, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

© 2009 Elsevier Ltd. All rights reserved.

What Larger Brains Are Made Of

More Neurons

One century of comparative studies has led to the notion that larger mammalian brains are built with larger numbers of neurons whose average size increases only modestly as brain size increases. Although technical difficulties preclude precise estimates of average neuronal size, changes in this parameter can be inferred from changes in neuronal density. Compared to the large volumetric expansion of structures such as the cerebral cortex, the relatively small decrease in neuronal density (presumably a result of larger neuronal size) in the structure across species indicates that increasing neuronal size is a factor in the generation of large brains, but one that is less important than increasing numbers of neurons. For example, the brain of the capybara, a giant Amazonian rodent, is 180× larger than the mouse brain, contains 22× more neurons, but its density of neurons per unit of tissue is only 8× smaller than in the mouse. The discrepancy may be even greater in primates: although the human brain weighs 175× more than a marmoset brain and has a 162× heavier cerebral cortex, the neuronal density in this structure is only about 3× smaller in humans. This indicates that the increase in average neuronal size contributes to increasing cortical size, but is a minor factor compared to increased number of neurons, in making the human cortex much larger than the marmoset cortex. Therefore, the generation of larger brains, at least within mammalian orders, requires the generation of larger numbers of neurons during development.

More Surface

The larger cerebral and cerebellar cortices of larger brains are increased mostly in their surface area, rather than in thickness. For example, the gray matter of the human cerebral cortex is roughly 1000× larger in surface area than the gray matter of the mouse cortex, but only about 2× thicker than the latter. The addition of larger numbers of neurons to cortical structures during brain development must thus proceed in a way that privileges surface expansion rather than greater thickness of the cortical wall.

More Glia

In contrast to the decreasing neuronal density, and therefore presumably increasing average neuronal size, large brains have densities of nonneuronal cells that are fairly similar to those found in smaller brains. This suggests that the average size of nonneuronal cells does not vary much with brain size, although their total number must increase rapidly in larger brains. In rodents, the total nonneuronal cell density in the brain is only 2× smaller in the capybara than in the mouse, even though the 180× larger brain of the former species has 85× more nonneuronal cells than the latter, suggesting that nonneuronal cells are not much larger in larger brains. Thus, besides the generation of larger numbers of neurons of moderately larger neurons, the development of larger brains requires the formation of an even larger number of glial cells whose size is presumed to vary little.

Neurons First, Glia Later

Brain development follows a series of well-documented steps. Early in development, the progenitor cells that compose the proliferative zones that line the ventricles undergo a series of symmetrical divisions, which expand exponentially the number of progenitors that will eventually give rise to neurons and glial cells in the brain. These symmetrical divisions establish the size of the founder population of progenitor cells that will build the brain.

Asymmetrical divisions ensue, in which one progenitor gives rise, for instance, to one self-renewing progenitor and one proliferating cell, which will give rise to the neurons that migrate outward to form the cortical plate and the parenchyma of other brain structures. Regressive events also occur in development, and may drastically reduce the neuronal population. According to some estimates, the total number of cortical neurons may be cut in half during development due to cell death.

The glial population of the brain is formed only once the neuronal population is already in place, as cell-intrinsic mechanisms eventually promote a switch from neuronal to glial production by progenitor cells. Neurogenesis is mostly prenatal; gliogenesis extends into postnatal life.

How to Add More Neurons

If large brains have an increased number of neurons, their generation must involve the production of larger

numbers of neurons during development. This may be achieved in at least three ways: by increasing the number of progenitors, by increasing the number of neurons generated by each progenitor, and by decreasing the number of neurons lost to cell death. These processes necessarily take time, so it seems logical that large brains take longer to develop than small brains. Indeed, the neuronal population of the mouse cortex is generated over about 6 days, while it takes about 55 and 100 days to generate all neurons in the macaque and human cortex, respectively. Gestation time increases coordinately with brain size at birth, among various mammalian orders.

More Progenitors Expand the Cortical Surface

Clonal analysis has shown that the neurons generated from one single progenitor are often clustered radially in the cerebral cortex. Since larger cerebral cortices are increased mostly in surface rather than in thickness, they should be generated by a larger number of radial clusters of neurons. If each radial cluster is derived from one progenitor cell, it follows that, all other things being equal, a larger number of progenitor cells should give rise to a larger surface of cerebral cortex of an unchanged thickness, containing a larger number of radial clusters of neurons. This is the essence of the 'radial-unit hypothesis,' proposed by Pasko Rakic in 1985.

Since the population of progenitor cells is expanded exponentially by symmetric cell division, small increments in the number of divisions will give rise to much larger numbers of progenitors. With only $3\times$ extra rounds of cell division, for instance, a $2^3 = 8\times$ larger number of progenitors would be generated. If all other factors remain unchanged, these three extra rounds will generate a brain with $8\times$ more neurons. In this scenario, the $22\times$ larger number of neurons in the capybara brain could be achieved by an average 45 extra rounds of symmetric division of the early progenitor cells, prior to the generation of neurons.

Experimental data have shown that X-ray irradiation of monkey embryos during the period of symmetric division of progenitors results in a decrease in cortical surface, with little effect on its thickness. Moreover, genetic manipulations that increase the number of progenitors early in development do lead to animals with an abnormally large cortical surface.

However, the expansion of the progenitor symmetrical division is unlikely to be the sole factor that increases the number of neurons in brain structures. The radial-unit hypothesis is based on the proposition by Rockel, Hiorns, and Powell in 1980 that there is a constant number of close to 150 000 neurons underneath 1 mm^2 of cortical surface of any mammalian species. This proposed uniformity has been contested by several groups, who argue that the number of neurons underneath a given surface may vary by as much as $8\times$ among species. While the expansion of the cortical surface is a major factor in building larger brains, and may be achieved by mechanisms that increase the population of founder progenitor cells, other factors must also contribute.

More Neurons per Progenitor Thicken the Cortical Surface

If increased symmetrical divisions expand the number of radial clones that will compose the cortical surface, increased asymmetrical divisions should expand the number of neurons that compose each clone, and thus increase the thickness of the cortical wall. Accordingly, there is evidence that experimental manipulations that increase the number of neurons generated per neuronal progenitor cell do lead to thickening of the cortical wall within a single species. Comparative studies addressing the number of neurons generated per progenitor across species, however, are lacking in the literature.

Decreased Cell Death Enlarges the Brain

Cell death is a normal mechanism that regulates the number of neurons in the brain. Differentiated neurons are subject to cell death, which is generally expected to cut the neuronal population of the brain in half, but the number of progenitor cells has also been shown to be regulated by cell death. Experimental genetic manipulations that delete genes required for cell death lead to animals with abnormally large brains. However, these mutations are lethal, and it is not known whether more subtle, natural genetic modifications of the cell-death pathway are an actual mechanism in the development of large brains. Interestingly, chemical manipulations that diminish but do not abolish progenitor cell death lead to both, expansion of the cortical surface and thickening of the cortical wall. This shows that the regulation of cell death is a possible mechanism in the generation of brains of different sizes in evolution.

Natural Mechanisms Are Still Unknown

Overall, it is possible that cortical surface and thickness be regulated both independently and coordinately, by developmental mechanisms that seem to apply to other brain structures as well. However, despite the relative wealth of data on experimental manipulations that affect proliferation, differentiation, and cell death and lead to the development of abnormally larger brains, it is still unclear whether

and how these processes vary naturally among species of different brain size.

More Glia Follow

Larger brains tend to have an increased ratio of glial cells to every neuron, or g/n ratio, compared to smaller brains. Fin whales, for example, have a gray matter g/n ratio of 454, while that ratio in humans is 178. Since glial cells appear in development only once the neuronal population is formed, mostly postnatally, mechanisms that adjust the number of glial cells to the number of neurons already present in the cortical wall must be in place in development.

Glial proliferation can be influenced by neuronal activity. Because it has been widely believed that the relative glial expansion has trophic and metabolic meaning for the neuronal population, the increase in the g/n ratio with increasing brain size has often been hypothesized to be regulated directly by neuronal activity. This would take place during gliogenesis, in postnatal development. Accordingly, the neuronal activity-regulated generation of large numbers of glial cells would further expand brain size beyond the enlargement due to the generation of larger numbers of neurons.

However, the numeric expansion of glial cells relative to neurons seems to contradict the observation that the neuronal need for metabolic support remains similar across species. Moreover, the increased g/n ratio of larger brains is not accompanied by any systematic variation in glial density, as large brains have glial cells distributed as densely as in small brains. These discrepancies could be settled if, during development, gliogenesis were regulated not by neuronal activity, but mainly by the overall neuronal mass in structures such as the cortical wall. This is in line with the observation that the proliferation of glial progenitors is controlled by contact inhibition with glial cells. Gliogenesis would thus presumably ensue until the entire cortical wall was occupied by evenly distributed glia. The higher the number of neurons in the cortical wall, the larger the latter would become, and thus the larger would be the volume of tissue to accommodate proliferating glial cells. Since increasing neuronal size also contributes to making brains larger, the cortical volume to be invaded by newly born glial cells is presumably yet larger in large brains. A larger number of glial cells would thus be generated for every neuron, leading to the increased g/n ratio observed in large brains.

Genetic Regulation of Brain Size

As mentioned above, changes in brain size can be generated experimentally through the manipulation of several genes controlling cell proliferation, differentiation, and death during development. Although it is unclear how these processes vary among species, it is already established that there is variability in genes that regulate brain size across primate species, humans included. For example, cell cycle genes associated with microcephaly (MCPH1, ASPM, CDK5RAP2, CENPJ) and caspase genes controlling cell death have higher evolutionary rates in primates, particularly in great apes and humans, than in rodents or carnivores, and seem to have undergone positive selection. It is unclear, however, how these gene changes are related to brain size.

See also: Allometric Analysis of Brain Size; Brain Connectivity and Brain Size; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brain Scaling Laws; Neurogenesis in the Intact Adult Brain; Programmed Cell Death.

Further Reading

Caviness VS Jr., Takahashi T, and Nowakowski RS (1998) Numbers, time and neocortical neuronogenesis: A general developmental and evolutionary model. *Trends in Neuroscience* 18: 379–383.

Herculano-Houzel S, Mota B, and Lent R (2006) Cellular scaling rules for rodent brains. *Proceedings of the National Academy of Sciences of the United States of America* 103: 12138–12143.

Kingsbury M, Rehen SK, Contos JJA, Higgins CM, and Chun J (2004) Non-proliferative effects of lysophosphatidic acid enhance cortical growth and folding. *Nature Neuroscience* 6: 1292–1299.

Rockel AJ, Hiorns RW, and Powell TPS (1980) The basic uniformity in structure of the neocortex. *Brain* 103: 221–244.

Sacher GA and Staffeldt EF (1974) Relation of gestation time to brain weight for placental mammals: Implications for the theory of vertebrate growth. *American Naturalist* 108: 593–615.

Tang BL (2006) Molecular genetic determinants of human brain size. *Biochemical and Biophysical Research Communications* 345: 911–916.

Brain Evolution: Developmental Constraints and Relative Developmental Growth

B L Finlay, Cornell University, Ithaca, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Components of the brain in mammals enlarge in large part by regular power laws based on the size of the brain overall. To transform the brain of a mouse to the brain of a chimpanzee, we can imagine that the volume or cell number of each brain component (for example, midbrain or cerebellum) or cell class (such as Purkinje cells of the cerebellum or rods in the retina) is each set by its own independent, nonlinear dial. The neocortex and the cerebellum grow with 'positive allometry' with respect to the rest of the brain, so that the larger brains become in absolute size, the progressively greater the proportion of the brain constituting neocortex and cerebellum. Conversely, the proportion that is medulla or midbrain decreases (Figure 1). The direct mechanistic cause of this conserved pattern of disproportionate but predictable growth is conservation of the order of neurogenesis across species such that structures or cell groups produced last experience the greatest expansion of the numbers of stem cells in their precursor pools in species with extended developmental periods.

Whether this conserved and disproportionate pattern of expansion of brain components is a metabolic and structural load that evolving brains must bear because of a maladaptive uniform application of a developmental program, as the word 'constraint' would suggest, is very suspect. Particularly so in that variation in size is one of the major ways that brains vary! Such conservation might better be interpreted as a mechanism that permits graceful scaling of brain function with brain size. In the same way that components of the body scale nonlinearly with body mass to compensate for the functional requirements of each organ that change lawfully at different body sizes, the brain must also scale appropriately. For example, skeletal mass in terrestrial mammals increases at a greater exponent than total body mass, particularly for supporting limbs (consider the deer, cow, and elephant), not because skeleton production is trapped by a developmental constraint requiring the animal to generate useless volumes of skeleton in supporting limbs at large body sizes, but because proportionately more cross-sectional skeletal mass is required to support the body. For skeletons and body mass, the reasons for allometric predictability are reasonably self-evident; for the relative scaling of medulla versus cortex, the underlying reasons may not be so clear and are the subject of this article.

The basic phenomena of brain component scaling will be discussed first, and then the properties of conservation and variation in developmental mechanisms that underlie conservation and variation in brain component scaling. Finally, several examples of conserved processes in neuroembryology that have the property of permitting flexible and robust conformation to commonly encountered evolutionary challenges will be considered.

Brain Scaling

Brain versus Body

The brain scales regularly with the size of the body, with an allometric exponent of about 0.75, an exponent whose cause has been the source of numerous academic debates. Therefore, as the body increases in size, the brain also increases in size, but so as to become an even smaller percentage of the body mass. As any cursory examination of office information processing devices will demonstrate that the sizes of a controller and the thing controlled are not obligatorily linked, why is there such a regular relationship in the case of the brain and the body? An obvious peculiarity of biological devices compared with engineered devices is that biological devices are composed of the elemental unit of cells, whose size is limited by diffusion and intracellular transport mechanisms, and not by the size of the body they compose. As the brain's most basic function is to move the body, and the larger bodies will be composed of a larger number of muscle units, an increase in the number of neurons will be necessary to innervate larger musculature (though axonal and dendritic arbors have a considerably greater range in possible size than a cell body does, and account for some of the range of neuromuscular scaling). Similarly, if large animals attempt to maintain the absolute spatial resolution of their skin sensation at some acceptable value (which seems reasonable, as things to be detected and localized, such as insects, cannot be depended on to scale with their prey animal's body size), the number of receptors and associated neurons transmitting information to the brain should scale with body size as well. There is no need whatever, however, to scale other sensory systems, such as eyes, ears, and vestibular organs, with body size, and yet they do scale. Larger organs

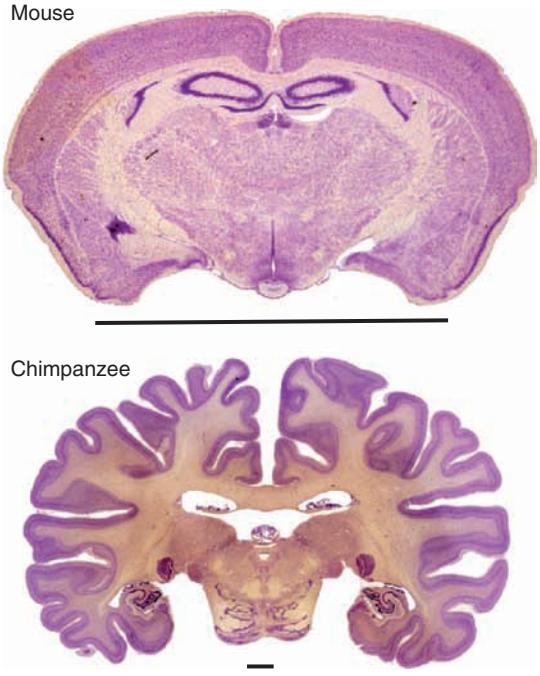

Figure 1 Comparison of the relative volumes of neocortex and brain stem in mouse (*Mus musculus*) and chimpanzee (*Pan troglodytes*) for comparable coronal brain sections. Scale bar = 0.5 cm for each. From Comparative Mammalian Brain Collections, http://brainmuseum.org.

can permit greater spatial resolution, but that causes improvement, not just maintenance, of quality. This issue will be mentioned again in the section titled 'Employment of "late equals large" to gracefully scale the retina.'

Body size does not entirely control brain size, however; at any particular body size, considering all the vertebrates, there is a possible range of about 40 in the ratio of brain to body size, and within the bony fish, birds, reptiles, and mammals, about a factor of 10 in the same ratio. While there is some arguable association of absolute brain size with behavioral complexity, in fact the two can be quite dissociable. For example, while a hummingbird could fly, collect nectar, and conduct social interactions within a space of the size of the brain case of a baleen whale, the length and complexity of the list of behaviors hummingbirds and whales might exhibit somewhat favors the hummingbird. It is relative size of the brain with respect to the body, variously measured as various ratios or statistical residual variance compared with the animal's taxon, that corresponds best to behavioral complexity or intelligence.

Brain Component Scaling

While it is relative brain size, measured best by the 'encephalization quotient,' or residual deviation from the mean brain–body size value from the appropriate taxonomic group, brain component architecture is best measured with respect to absolute brain size. This was first noted for the relative size of the cortex compared with the rest of the brain; since the cortex has predictable positive allometry with respect to the brain, the larger the brain becomes, the more it comes to be composed of cortex. As humans have the largest encephalization quotient, but not the largest brain, this means that absolutely larger brains, such as the brains of elephants and many cetaceans, are composed of proportionately more cortex than are human brains.

The cortex is not a special case of predictable scaling, however. Each identifiable brain part has its own exponent, as can be seen in the different slopes of each structure in Figure 2, which plots the slopes of six structures versus 'brain core' – medulla, pons, midbrain, and thalamus. In fact, each identifiable region of the cortex, such as visual cortex, other primary sensory cortices, and parietal or frontal cortex has its own exponent as well, which has caused a great deal of confusion about whether components of the human cortex, particularly the frontal cortex, are as large as should be 'expected.' Many would expect a simple ratio, such as the human brain as the chimpanzee brain multiplied by three, and would predict that a deviation from such a ratio should represent special selection or adaptation. A multiplicative ratio, however, is rarely the nature of the relationship between brain parts – in fact, it occurs only in the special case when the ratio of the allometric exponents of the two parts happens to be one. If the frontal cortex has a positive allometry with respect to the cortex overall, as the cortex enlarges in absolute terms, the frontal cortex will come to comprise a greater percentage of the entire cortex. Thus humans have an entirely predictable, but disproportionate, amount of frontal cortex in their brain compared with other primates.

The Limbic Factor

In the various first investigations of brain component scaling of the large data set collected by Stephan and associates, it was observed by many researchers that primates, compared with bats and insectivores, showed a 'grade shift' with respect to the amount of cortex their brains contained, producing more cortex per unit brain but relatively less olfactory bulb, olfactory cortex, and hippocampus, as well as some other components of the extended limbic system. Knowing whole brain volume captures about 94% of the variability in brain component scaling in mammals overall, and adding a second limbic component gains another 4% of the variance. Until recently, it was unclear

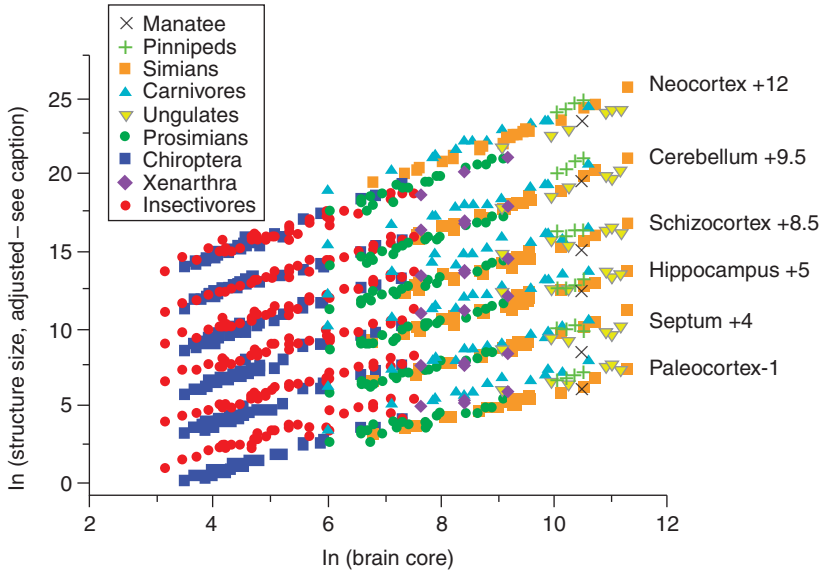

Figure 2 The sizes of six brain structures relative to brain stem and basal ganglia core size, across 160 species in nine taxonomic groups. The six structure sizes have been adjusted by the indicated arbitrary constants (listed on the right of the figure) to separate the six scatterplots visually. Both scales are natural log scales. The regression slopes are paleocortex (olfactory cortex), 0.826; septum, 0.847; hippocampus, 0.912; schizocortex, 0.963; cerebellum, 1.150; and neocortex, 1.441. Natural logarithm, ln. From Reep R, Darlington RB, and Finlay BL (2007) The limbic system in mammalian brain evolution. *Brain, Behavior and Evolution* 70: 57–70.

whether this differential allocation of tissue in the brain was specific to the primate–insectivore contrast or whether it was characteristic of mammalian brain variation more generally. With a sample that included marine mammals, known to have smaller olfactory bulbs, and also some species that would appear to be unusually dependent on chemosensation, such as anteaters, substantial variation along the limbic–cortex tissue allocation axis (Figure 3) has been found to be characteristic both between and within mammalian orders. This relationship was not an obligatory reciprocal relationship, however; it is possible for some animals (such as carnivores) to have both a relatively large neocortex and a relatively large olfactory bulb–hippocampus complex with respect to the brain core, or to be reduced in both (like bats), but variation on this dimension dominates the data. Unusually large amounts of variation on this same axis are true of individual variation within species as well, for species as diverse as swine, minks, and mice.

The Developmental Structure of the Observed Species Differences

A basic, conserved pattern of neurogenesis proved to underlie the consistency observed in mammalian brain component scaling. Those structures which became disproportionately large as brains became absolutely larger were the same structures and cell groups whose 'birthdays' were late. That is, those structures whose precursor pools had an extended time to enlarge before the first of their cells exited the precursor pool and differentiated as neurons were the same ones that became differentially large. We call this relationship 'late equals large,' and it falls directly out of the kinetics of cell proliferation in the developing embryo, which we represent in its most generalized form in Figure 4. This figure represents the empirical observation that cell division begins roughly simultaneously in all parts of the neural tube that will give rise to the mammalian brain, but neurogenesis terminates in some parts earlier than others. When neurogenesis is extended from the 21 days or so required to generate a mouse brain to the 100 or so required for a monkey, the consequences for different brain components are not uniform: the regions withdrawing neurons the latest from their precursor pools allow those regions to see more cell divisions and become substantially larger.

The strength of this relationship brings forward some essential regularities about mammalian brain development observed empirically. First, the brain acts as if it is a rate-limiting step in development overall. Early embryos are composed of a large percentage of brain by volume, and the literal bulk of somatic development occurs later, in some species predominantly postnatally. If brain weight is plotted against gestational length in days across diverse eutherian mammalian species, an extremely straight line results, as if a fixed amount of brain can be produced per day. No evidence to date points to locally variable rates of neuron or precursor development; to

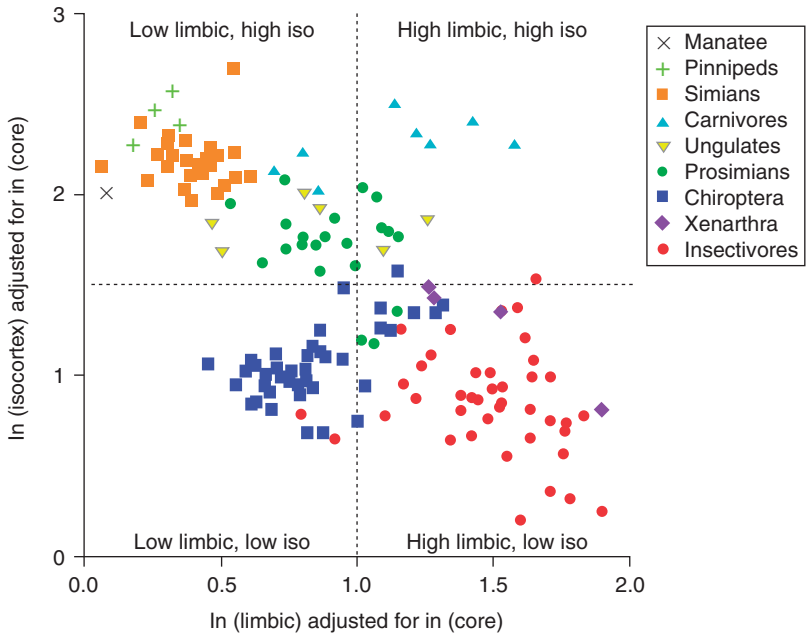

Figure 3 Relationship between limbic and isocortical (neocortical) components of the telencephalon. Plotted are residual variance in the volume of limbic and isocortical components, referenced to brain core volume, for 160 species in nine taxonomic groups. Overall there is a push–pull arrangement between the limbic and isocortical components, indicated by the preponderance of values in the upper left and lower right quadrants. However, some carnivores represent a balanced expansion of both components, and most chiropterans exhibit reduction of both components. iso, isocortex. From Reep R, Darlington RB, and Finlay BL (2007) The limbic system in mammalian brain evolution. *Brain, Behavior and Evolution* 70: 57–70.

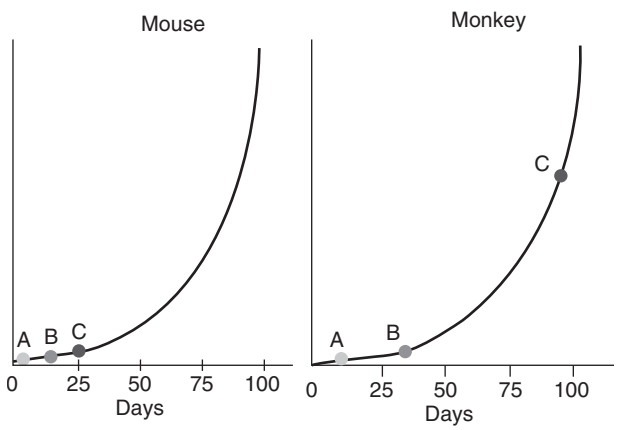

Figure 4 'Late equals Large': a schematic of the consequences for eventual size of a structure generated early (A), intermediate (B), or late (C) in the order of neurogenesis for a species with a short period of neurogenesis and a small brain, like a mouse, versus one with a long period of neurogenesis and a large brain. In the long-development species, the precursor pool for late-generated structures has a longer time to multiply and becomes disproportionately large.

produce more brain, precursors are generated, not more quickly, but over a longer time. (In marsupials, noneutherian mammals, however, the overall pace of brain production is considerably slowed compared to placental mammals, so it is possible to change rate wholesale.) Finally, there is a spatial organization, derived from the initial segmental structure of the neural tube, to the order in which structures enter terminal neurogenesis, and this spatial organization accounts for about half the variation in date of termination of neurogenesis (Figure 5). Those areas that are the most anterior in the neural tube, and most lateral with respect to the original neural tube, that is, the alar plate, are the ones that enter terminal neurogenesis the latest. In fact, for two of the three most anterior and lateral division of the neural plate, which give rise to the hippocampus, neocortex, and olfactory bulb, neurogenesis does not terminate at all.

Limbic System Reduction and Amplification

For primates, evidence exists about how the wholesale reduction of the olfactory bulb, olfactory cortex, hippocampus, and several other nuclear limbic structures, as well as the enlargement of the neocortex, is produced: if the schedule of rodent neurogenesis is appropriately transformed and lined up against the schedule of the rhesus monkey, it is obvious that the distributed limbic structures in the monkey all appear to commence terminal neurogenesis on the same day, and all of them earlier than would have been expected for the same structures in a rat (Figure 6). Conversely, all the layers of the cortex begin terminal neurogenesis later in development than would be predicted from the rat schedule. Those structures that do not differ in their expected size – midbrain, basal ganglia,

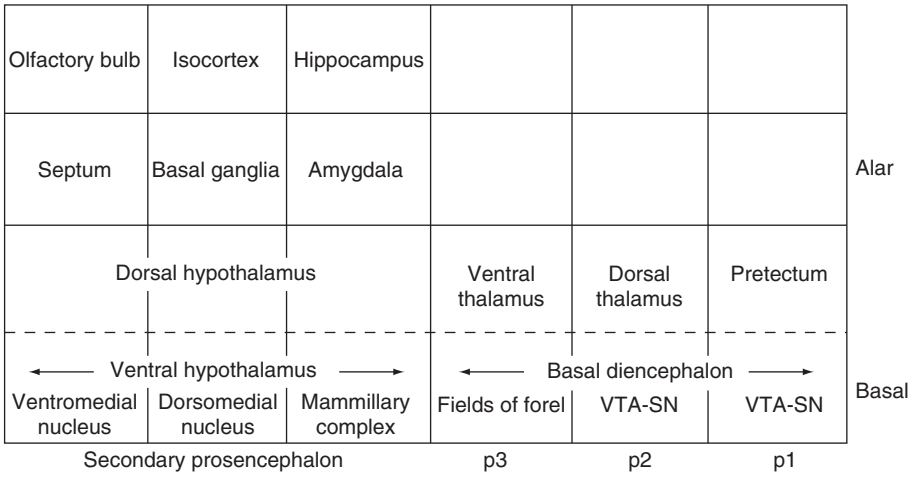

Figure 5 Prosomeric model of the forebrain, showing the embryonic spatial relationships of the areas of the brain which undergo the most protracted neurogenesis: olfactory bulb, neocortex, and hippocampus.

and cerebellum – line up at the equivalent times predicted. It is a very interesting research question as to what the genomic alteration might be in monkeys that alters the onset of neurogenesis forward and backward in this distinctive way. It is not yet known whether the other animals that differ on the same axis, like marine mammals, have altered their onsets of neurogenesis in equivalent ways.

The animals whose schedules of neurogenesis have been examined include rodents, the rhesus monkey, cat, ferret, and possum – not an extensive comparative base, but a corpus that has been examined closely to see whether any other potential units of covariation in brain might be observed in any one species or across species (for example, functional systems, such as visual or auditory systems, or parts of brain geometry such as midbrain or hindbrain). Thus far, no such covarying units that could explain more interspecific variation in brains have been identified.

Developmental Constraint, or a Useful Reservoir of Variation for Brain Evolution?

Many aspects of early brain segmentation and genetic specification are highly conserved across vertebrates. The most reasonable interpretation of the extremely conserved fundamental body plan in vertebrates, one of many examples, is not that it is a crystallized 'constraint' on vertebrate organization, but rather a robust and flexible system that by its nature permits graceful variation. The fundamental mechanisms of the specification of body plan are 'evolvable.' How might this argument be extended to a segmental pattern of neurogenesis that permits by its nature systematic variation in the duration of the genesis of precursor pools for neurons, and that has the further property of systematically and disproportionately amplifying the numbers of certain populations of precursor pools with respect to each other as neurogenesis is extended? As discussed earlier, if a mammal is to be selected for a larger brain, it must take longer to make it. In mammals, the principal way brains differ in size is across, not within species. Since this type of size variation has happened so commonly, the property of graceful scaling should be viewed as a historical filter on the kinds of developmental programs that can undergo selection and that are preserved today. It should be noted, however, that mammals are unusual with respect to most vertebrates in that most of their neurogenesis is confined to early development. In many other animals, both brain and body grow throughout life and must be scaled 'on line,' coupling scaling of the two structures in the life of an individual animal.

Employment of 'Late Equals Large' to Gracefully Scale the Retina

The retina is a case in which, in order to maintain functionality in a system of different scales, cell groups must (and do) scale with different exponents. Rods and cones must scale at different slopes with eye size in order to hold constant their particular functions. If an eye becomes twice as large in diameter, no change is necessary in the number of cones to retain the same visual acuity; since the retina is flooded with photons in diurnal vision, a single cone will have no difficulty encountering a photon in the visual angle it represents regardless of the angle the cone itself subtends. More cones could of course be added, to improve acuity, but we are discussing here what is required to maintain equivalent, not improved, function over different eye sizes.

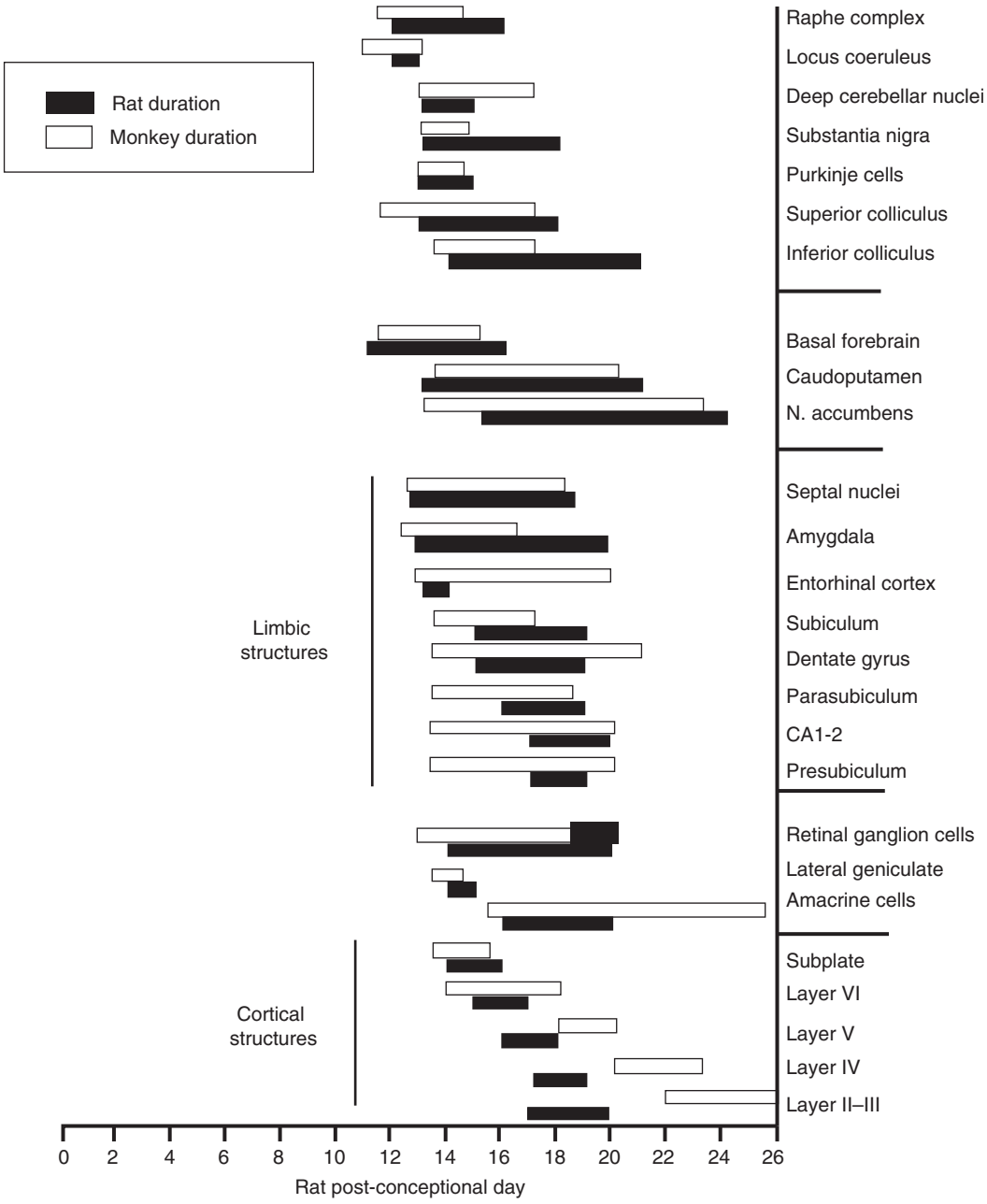

Figure 6 Using the 'translating time' functions described at http://www.translatingtime.net/ (Translating Time Across Developing Mammalian Brains), the schedule of neurogenesis of the monkey is translated to the duration of a rat's for direct comparison. Cessation of neurogenesis for limbic structures begins simultaneously and earlier than expected in the monkey, and the generation of the neocortex later.

The same solution will not work for rods. Working at low light levels and low photon numbers, a single rod located in a larger absolute retinal angle will fail to detect most photons, even allowing for biologically plausible increases in the size of a single rod. Rods must tile the surface of the retina to maintain sensitivity, increasing in number approximately at the square of change in retinal diameter. The scaling of rods and cones in diurnal primates in fact conforms closely to this functional requirement; for example, the human eye, larger than the marmoset eye, has only about 1.2 times the number of cones and retinal ganglion cells in the marmoset eye but 8 times as many rods (Figure 7). How is this consistent within- and across-species disproportional scaling necessity executed in the schedule of neurogenesis of the

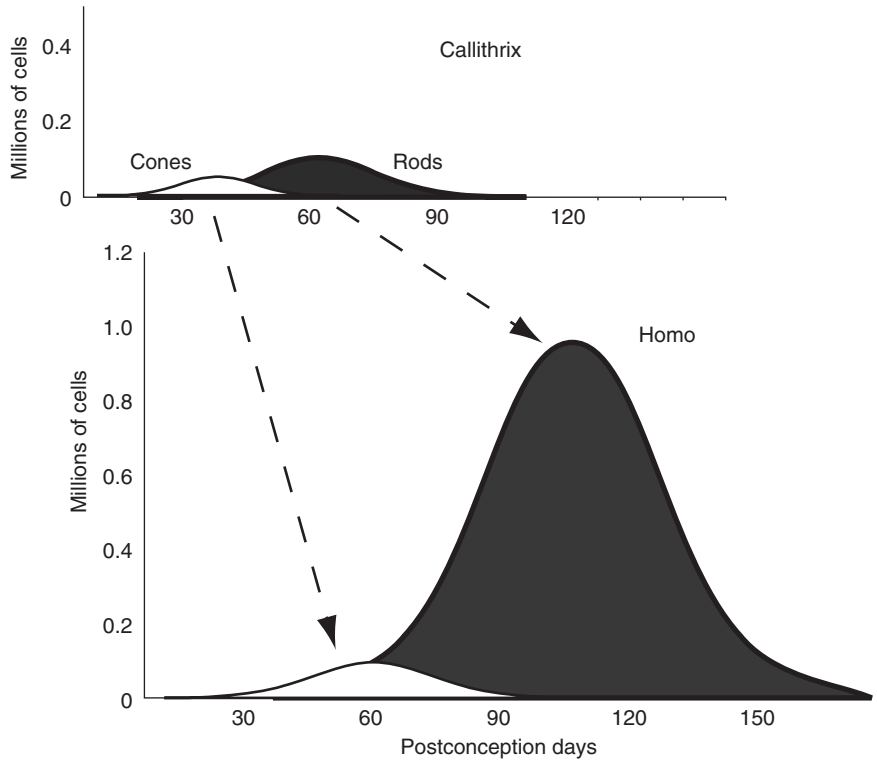

Figure 7 The hypothesized consequences of extended neurogenesis in the retina of *Homo sapiens* versus *Callithrix iacchus* for their relative numbers of cones and rods. The positioning of rod proliferation late in neurogenesis allows dispropotionate proliferation of rods in a larger retina to maintain sensitivity. Areas under the curves for rod and cone numbers are drawn to reflect empirically determined values in each species. Reproduced with permission from figure 2.5 in Finlay BL, Silveira LCL, and Reichenbach A (2005) Comparative aspects of visual system development. In: Kremers J (ed.) *The Structure, Function and Evolution of the Primate Visual System*, pp. 37–72. Chichester: Wiley, Copyright 2005, Copyright John Wiley & Sons Limited.

retina? While the precise kinetics are being worked out, the schedule of neurogenesis in the retina is arranged such that extension of the period of embryogenesis should automatically produce the desired differential scaling, in that cones go into terminal neurogenesis early, and rods late. Those cell types that must change in number with a high exponent with eye diameter, rods and their attendant bipolar cells, are located last in order of differentiation, and those that need not change are produced first. This obligatory, coordinated scaling of retinal cell classes to match functional requirements is an example of the idea of 'evolvability.' The retina offers a potential source of variability in the onset of photoreceptor differentiation by cell class, some of which might gracefully scale with eye size, and some not. Those potential ancestors with perhaps a reversed order of neurogenesis in the retina, rods first, cones last, who might have enjoyed a selective advantage at a larger brain size, but which had ineffective vision in the dark as a result, presumably would enjoy less reproductive success. Common selection points, such as variable body and brain size, will naturally gravitate to coordinate with predictable aspects of embryogenesis. The retina is a nice example of order of neurogenesis adapted to permit several contrasting functional constraints in a single sensory system.

Using Continuing Neurogenesis to Permit Rapid Memory Acquisition and Loss in the Hippocampus

The hippocampus and the neocortex present another possible example of the siting of a function with respect to a favorable location for continued neurogenesis, the most lateral aspect of the alar plate. Both structures are involved in the integration of multisensory information and in memory but are distinct in that the hippocampus is specialized for immediate short-term memory and the transfer of the short-term store to the long-term storage of the neocortex. New neurons are generated in the hippocampus throughout life, but this unusual maintenance of neurogenesis is well suited to hippocampus function: the 'noise' introduced into the connectivity of a structure whose memories are continually overwritten does not represent a problem for the maintenance of any particular memory and in fact may be useful in

memory erasure. The component of the brain important for short-term memory might function best and survive millions of bouts of selection if the appropriate circuitry and cellular processes were sited in a brain location where the pattern of protracted neurogenesis was favorable.

Why Does the Cortex Get Big and the Spinal Cord and Sensory Thalamus Stay Small? A Network Hypothesis

This argument follows the same lines as that described for the rods and cones of the retina and for the hippocampus: over evolutionary time, important functions in the brain are likely to site themselves where the pattern of neurogenesis is permissive for optimal function, in this case, optimal scaling of its network architecture. The pattern of selective enlargement of the telencephalon and constraint on spinal cord and thalamus contrasts with a type of brain evolution essentially unseen in mammals. Consider that a star-nosed mole, with its specialized end organ for palpation, does not generate a dedicated processing center of cortical proportions for its new sensory system in the spinal cord, but rather in the neocortex, where mouse whiskers, primate hands, and elephant trunks all locate themselves as well. If brain evolution could truly be mosaic, that is, any structure could be the target of differential selection on size, why are integrating structures of the size of the cortex not situated in the spinal cord, the hindbrain, and the midbrain?

The answer to the question of coordination of brain, sensory systems, and body scaling and the preferential enlargement of the cortex may be found in the nature of what is likely to be represented in a larger brain and the particular nature of networks in the brain, particularly the cortex. A common hypothesis about brain evolution is that an animal might specialize or improve function by adding a new 'module,' by which is meant a specialized processing system, relatively encapsulated, like bat echolocation, palpation of the earth by the star nose of the mole, or human language. Minimal consideration of even these systems shows that all must constantly integrate motor output with sensory input, which must then further direct muscle effectors, and even the most committed sensory system has multimodal components. Chemosensory information must be integrated with somatosensory, and in current fMRI studies in human language processing, it has been shown that generating the words for tools activates primary somatomotor cortex. So it is obvious that elaboration of sensory, cognitive, or motor capacities should be situated in locations where multimodal integration is possible.

Why body, brain, and sensory systems should covary so regularly, even though neural tissue is metabolically expensive, may have its answer in the distributed nature of neuronal connections. Particularly in the cortex, each neuron receives thousands of synapses, none of which has a 'command' nature; not only that, there are small assemblies of neurons, such as the cortical column, with high local interconnectivity and which almost certainly participate in multiple disparate computations, such as in the language example above. In modeling studies, if the size of just one component of a multicomponent system is increased differentially, the excessive input–output (I/O) requirements of the single larger component when communicating with other system components propagates itself throughout the neural net, leaving the remaining systems with far fewer computational resources than they would have if the large system had simply been deleted from the resources of the net. It is a peculiar and not much noticed fact that the primary sensory and motor nuclei are the first-generated ones in the thalamus, causing them to scale with negative allometry with respect to the rest of the thalamus. The primary sensory nuclei and their inputs act as a significant bottleneck on input to the cortex – for example, the human retina has only about 6 times as many retinal ganglion cells contributing optic nerve axons to the thalamus as a gerbil's! The primary sensory and motor areas of the cortex are also the ones that appear to be the most specified in terms of location and size in the developing brain. Further, in all the primary sensory areas, if an animal develops a specialization, such as an auditory, acoustic, or somatosensory 'fovea,' the increased allocation of cortical area to the specialization takes its place within the primary cortex and does not appear to enlarge it. Rather, the effect of the specialization is seen primarily in the proliferation of new cortical areas. It is interesting that the two large neural systems that show evidence of decoupling, the neocortical and limbic, do not directly combine primary sensory inputs in the thalamus, which may be permissive for diverging allometric scaling that these two systems show (**Figure 4**).

These anatomical observations, initial modeling studies, and hypotheses are here combined to suggest that the fundamental kinetic of cytogenesis described as 'late equals large' may be employed repeatedly as a natural substrate for graceful scaling in vertebrate development and evolution, whether to defend the functional requirements of rods versus cones in sensitivity or acuity, respectively, or the I/O requirements of an associative net with multiple separate inputs. Thus, the conservation seen in the scaling of body, sensory, and motor inputs and brain components does probably represent a constraint, although not a constraint that development places on function but rather functional constraints that exploit to best purpose the intrinsic nonlinearities of neurogenesis.

See also: Brain Asymmetry: Evolution; Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brain Evolution: The Radiator Theory; Brain Modules: Mosaic Evolution; Brain Scaling Laws.

Further Reading

Catania KC and Kaas JH (1997) Somatosensory fovea in the star-nosed mole: Behavioral use of the star in relation to innervation patterns and cortical representation. *Journal of Comparative Neurology* 387: 215–233.

Finlay BL and Brodsky PB (2006) Cortical evolution as the expression of a program for disproportionate growth and the proliferation of areas. In: Kaas JH (ed.) *Evolution of Nervous Systems*, pp. 73–96. Boston, MA: Elsevier.

Finlay BL, Darlington RD, and Nicastro N (2001) Developmental structure of brain evolution. *Behavioral and Brain Sciences* 24: 263–308.

Finlay BL, Hersman MN, and Darlington RB (1998) Patterns of vertebrate neurogenesis and the paths of vertebrate evolution. *Brain, Behavior and Evolution* 52: 232–242.

Finlay BL, Silveira LCL, and Reichenbach A (2005) Comparative aspects of visual system development. In: Kremers J (ed.) *The Structure, Function and Evolution of the Primate Visual System*, pp. 37–72. Chichester, UK: Wiley.

Gerhart J and Kirschner M (1997) *Cells, Embryos and Evolution*. Malden, MA: Blackwell.

Jerison HJ (1997) Evolution of prefrontal cortex. In: Krasnegor NA, Lyon GR, and Goldman-Rakic PS (eds.) *Development of the Prefrontal Cortex: Evolution, Neurobiology and Behavior*, pp. 9–26. Baltimore, MD: Brookes.

Pascual-Leone A, Amedi A, Fregni F, and Merabet LB (2005) The plastic human brain. *Annual Review of Neuroscience* 28: 377–401.

Pulvermuller F (2001) Brain reflections of words and their meaning. *Trends in Cognitive Sciences* 5: 517–524.

Ragsdale CW and Grove EA (2001) Patterning the mammalian cerebral cortex. *Current Opinion in Neurobiology* 11: 50–58.

Reep R, Darlington RB, and Finlay BL (2007) The limbic system in mammalian brain evolution. *Brain, Behavior and Evolution* 70: 57–70.

Schmidt-Nielsen K (1984) *Scaling: Why Is Animal Size So Important?* Cambridge, UK: Cambridge University Press.

Stephan H, Baron G, and Frahm HD (1981) New and revised data on volumes of brain structures in insectivores and primates. *Folia Primatologica* 53: 1–29.

Relevant Websites

http://brainmuseum.org – Comparative Mammalian Brain Collections.

http://www.translatingtime.net – Translating Time Across Developing Mammalian Brains.

Brain Evolution: The Radiator Theory

D Falk, Florida State University, Tallahassee, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Two remarkable facts characterized the evolution of hominins since their ancestors split from those of living chimpanzees approximately 5–7 million years ago (Ma): Unlike chimpanzees, hominins evolved and refined habitual bipedalism, and during the course of their evolution mean brain size increased to three times that expected for chimpanzees of equivalent body sizes. Although the primary selective pressures that resulted in habitual bipedalism and enlarged brains in hominins remain a matter of intense speculation among paleoanthropologists, it is generally recognized that bipedalism evolved before brains became significantly enlarged. According to the radiator hypothesis, selection for bipedalism kicked off a cascade of events that eventually resulted in the evolution of a network of cranial veins that contributed to selective cooling of brains (SBC) under conditions that stimulated hyperthemia in certain species of hominins, much as the radiators of cars prevent their engines from overheating. This vascular mechanism for regulating brain temperature released a thermal constraint that had previously kept brain size in check, thereby paving the way for the natural selection of enlarged brains (for whatever behavioral reasons) that occurred in the direct ancestors of *Homo sapiens*.

Evolutionary Changes in Posture Altered Human Vascular Systems

The hemodynamics of blood circulation in animals fluctuate in response to a variety of factors, including action of the heart, changing respiration, and gravitational (hydrostatic) pressures that alter with varying body postures. Data from arboreal snakes and long-necked giraffes suggest that animals that evolved changes in their habitual body postures did so in conjunction with adaptive changes in their circulatory systems (especially the venous system) in the face of altered hydrostatic pressures. Humans are just as subject to gravity as other animals and retain a replumbed cranial vascular system that differs from those of other primates, which they evolved in conjunction with selection for bipedalism. As a result, cranial blood flow in modern humans is highly dynamic and alters with an individual's posture, breathing, exertion, and thermal state. In monkeys, the internal jugular veins are the major pathway of cerebral venous return. In upright (but not supine) humans, however, the delivery of blood fluctuates with respiration and exertion between the internal jugular vein and a large valveless vascular network (the vertebral plexus) that extends from the cranium to the coccyx. Because of the habitual upright posture of humans, the vertebral plexus is believed to be a major pathway for cerebral venous return. Emissary veins help to deliver blood to the vertebral plexus of veins in upright humans.

Emissary Veins Facilitate Selective Brain Cooling in *Homo sapiens*

Brains are extremely metabolically active and are very sensitive to heat. Brain temperature is determined by the temperature of arterial blood supplying the brain, the rate of blood flow, and the rate of metabolic heat production of the brain. When animals exercise, the temperature of the brain may be kept below that of the body by SBC, which is facilitated in numerous carnivores and ungulates (including dogs, cats, wolves, hyenas, oxen, sheep, and lambs) by cooling arterial blood as it flows through a complex network of arteries (the *rete mirabile*, or marvelous net) that is immersed in cooler venous blood within the cavernous sinus at the base of the brain. Thus, in those animals that have them, retia mirabilia act as countercurrent heat-exchange networks that transfer heat from inside to outside their crania.

Human brains have relatively greater cooling needs than other animals because of the increased ratio of cerebral to resting metabolic rates that is associated with comparatively big bodies and makes brains more susceptible to neurological disturbance (and even death) from hyperthermia and because human brains are comparatively large and thus generate a good deal of heat. Nonetheless, the question of whether human brains are subject to SBC has been controversial because people (and other primates) lack carotid retia. However, a number of other animals, including horses, rabbits, rats, flying bats, and pigeons, have been shown to experience SBC despite their lack of carotid retia. Although it would be unethical for experimenters to measure brain temperature directly in human subjects exposed to hyperthermic conditions, mounting experimental evidence suggests that SBC occurs in humans via cooling of the airways (ventilation through the buccal cavity and nose), fanning of the face, and by evaporation of sweat from the head. The latter mechanism employs a widespread network of emissary veins that pass through various openings in the skull to form connections between the venous sinuses of the dura

inside the skull, the veins within the diploë of the cranial bones, and the veins on the exterior of the skull. The existence of SBC in humans was proposed more than 20 years ago by Michel Cabanac and Heiner Brinnel based on their experiments comparing blood flow in two large emissary veins located near the posterior top and sides of the skull (the parietal and mastoid, respectively) and a third vein at the front of the skull (the ophthalmic) that passes through a foramen to form connections between the cavernous venous sinus internally and facial vein externally (Figure 1). Direction of blood flow within these valveless veins was determined in subjects experiencing two conditions – hypothermia and exercise-induced hyperthermia. During hypothermia, the blood did not flow or flowed slowly from the brain out to the surface of the skull. During hyperthermia, on the other hand, blood that had been cooled by vasodilation in conjunction with evaporation of sweat from the scalp and forehead flowed rapidly from the surfaces of the crania into the cranial cavity. These reversals of blood flow are a consequence of reversed blood pressure gradients between the exterior and internal cranial venous networks, which are controlled by the sympathetic nervous system. The authors hypothesized that the three veins were part of a rich system of innumerable subcutaneous veins (many of them microscopic) that collects cooled blood from the vasodilated skin of the heads of hyperthermic humans and delivers it into their braincases.

Although the hypothesized existence of a widespread network of cranial emissary veins that facilitates SBC in humans initially met with skepticism, in 1996 Wolfgang Zenker and Stefan Kubik published a detailed analysis of human vascular anatomy that is capable of facilitating a bulky transfer of cooled venous blood from the skin of the head, nasal, and paranasal mucous membranes through the skull to the rich network of veins in the dura mater, which in turn may transmit temperature changes to the cerebrospinal fluid and, from there, to the brain (Figure 2). This finding lends support to (and broadens) the hypothesis that convective SBC involving an elaborate network of tiny veins that perforate the cranial vault occurs in humans when the blood in the skin is cooled by evaporation during intense sweating from the head.

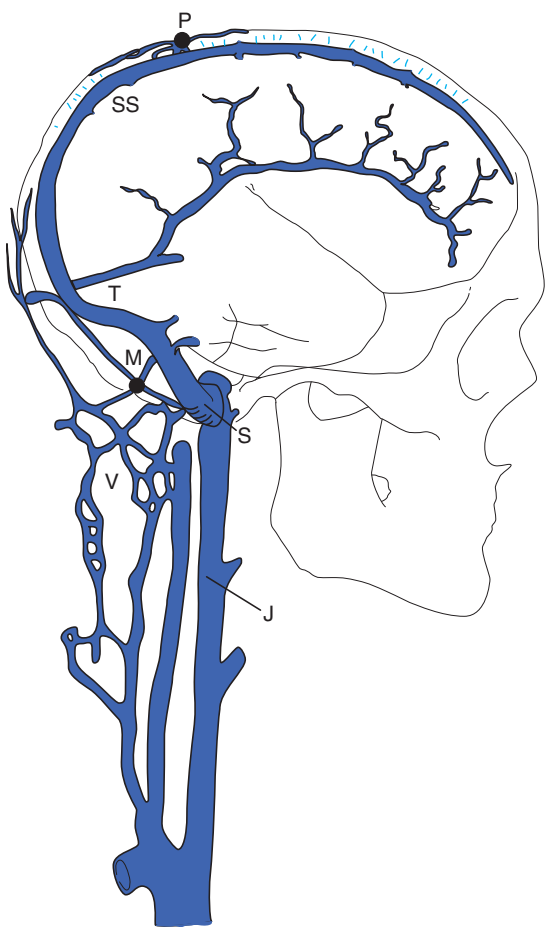

Figure 1 Venous sinuses and parietal (P) and mastoid (M) emissary veins in humans. When humans are lying down, intracranial blood flows preferentially from the sigmoid sinus to the internal jugular vein. Upright humans deliver cranial blood preferentially to the vertebral plexus. These emissary veins are part of a much more extensive network of smaller emissary veins that penetrate the skull. SS, superior sagittal sinus; T, transverse sinus; S, sigmoid sinus, which connects with the large internal jugular vein, J; V, vertebral plexus of veins. Copyright Dean Falk.

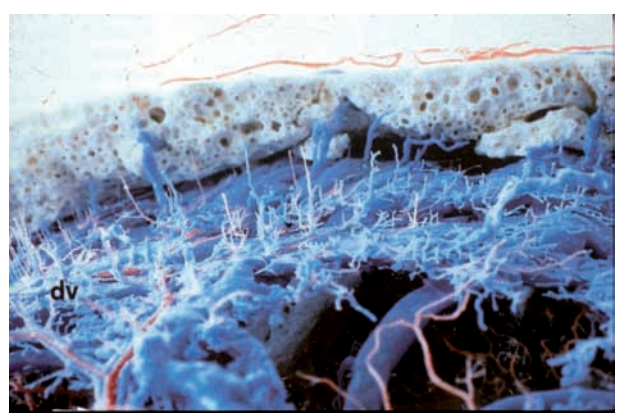

Figure 2 Corrosion cast of cranial blood vessels. Parietal bone is preserved. Note the vessels arising from the dural vascular rete and entering the bone (dv). Reproduced from *Anatomy and Embryology*, vol. 193, 1996, pp. 1–13, Brain cooling in humans – anatomical considerations, Zenker W and Kubic S, figure 4, copyright Springer-Verlag, with kind permission of Springer Science and Business Media and Zenker W.

Cranial Blood Flow in Early Hominins

Because the large named emissary veins pass through observable foramina and venous sinuses leave traces that are imprinted within the braincase, cranial blood flow has been studied in three genera of fossil hominins (*Paranthropus*, *Australopithecus*, and early *Homo*). Fossils of so-called robust australopithecines (*Paranthropus*) date from approximately 2.5 to 1.0 Ma and are known from South and East Africa. Although robust australopithecines were bipedal, they are believed to have been too specialized in many aspects of their anatomy to have been direct ancestors of *Homo*. One of these specializations involves cranial blood flow. To date, all 12 of the *Paranthropus* skulls that are complete enough to score manifest an enlarged route at the back of the skull for delivering blood from the braincase. This enlarged occipital/marginal (O/M) venous sinus (Figure 3) appears to have had connections with the vertebral plexus of veins that was presumably as important for receiving cranial blood in upright (bipedal) *Paranthropus* as it is for upright humans. It is important to note, however, that unlike humans (and like apes), *Paranthropus* appears to have lacked a high frequency of emissary veins and therefore an evolved adaptation for SBC.

In contrast to *Paranthropus*, the African great apes, *Australopithecus*, early *Homo*, and humans have very low frequencies of enlarged O/M sinuses and instead rely on the transverse/sigmoid sinuses for delivering cranial blood to their internal jugular veins. These groups differ markedly, however, in their frequencies of named emissary veins. Like robust australopithecines, the African great apes have very low frequencies of emissary foramina. Comparative studies on the frequencies of numerous named emissary foramina show that their frequencies increased dramatically during hominin evolution. Specifically, data for the parietal and mastoid foramina suggest that since branching off from *Australopithecus* prior to 1.6 Ma, *Homo* has experienced a steady increase in the vascular network of emissary veins that facilitate SBC under hyperthermic conditions. Significantly, elaboration of the emissary network of veins in the lineage leading to *Homo sapiens* correlated with an ongoing trend for enlarging brain size.

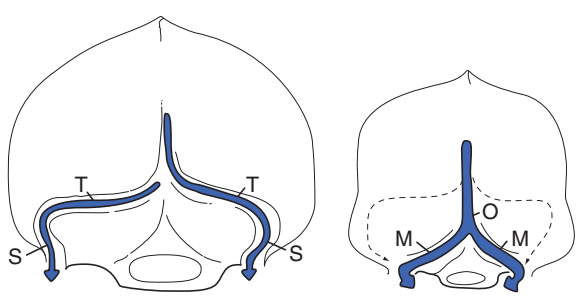

Figure 3 Occipital views of typical cranial venous sinus systems in *Australopithecus*, early *Homo*, and modern humans (left) and robust australopithecines (*Paranthropus*) (right). In the former, a major exit for cranial venous blood is through the transverse (T)/sigmoid (S) sinuses, which deliver blood to the internal jugular veins. Although this system was sometimes retained in robust australopithecines, all of them drained a significant amount of cranial blood through an enlarged occipital/marginal (O/M) sinus system that was present on one or both sides. Copyright Dean Falk.

The Evolution of Brain Size and Cranial Blood Flow

Brain size remained conservative during the course of evolution in *Paranthropus*, which evolved an enlarged O/M sinus for draining cranial blood but not an elaborate network of emissary veins entailed in SBC (Figure 4). On the other hand, accumulating evidence suggests that brain size increased in the lineage leading from *Australopithecus* to *Homo sapiens* in conjunction with elaboration of the vasculature entailed in SBC in living people (Figure 5). This is not surprising in light of the fact that larger brains have greater cooling needs (discussed previously). As noted, *Paranthropus* and *Australopithecus* both evolved habitual bipedalism that necessitated replumbing of vascular systems in response to hydrostatic pressures, similar to other animals that experience major shifts in habitual posture and locomotion. Evidence suggests that, like the earliest hominins known from the fossil record, *Paranthropus* was not adapted to thermally stressful open grasslands or lifestyles, which is consistent with a cranial vascular system that lacks cooling capabilities. By approximately 1.9 Ma, however, brain size and frequencies of emissary foramina had begun to increase in other hominins compared to contemporaneous *Paranthropus*. According to the radiator hypothesis, this happened because the particular way the cranial vascular system in the lineage leading to *Homo* was evolving benefited SBC, thereby releasing a thermal constraint that had previously kept brain size in check. What, then, were these other hominins doing to bring about such a departure?

The Impact of Environment and Behavior on SBC in Hominins

Not too long ago, paleoanthropologists believed that earliest hominins that lived in Africa were adapted to open or partially wooded grassland savannas that were characterized by high predator pressures and other challenges associated with daily travel to seek

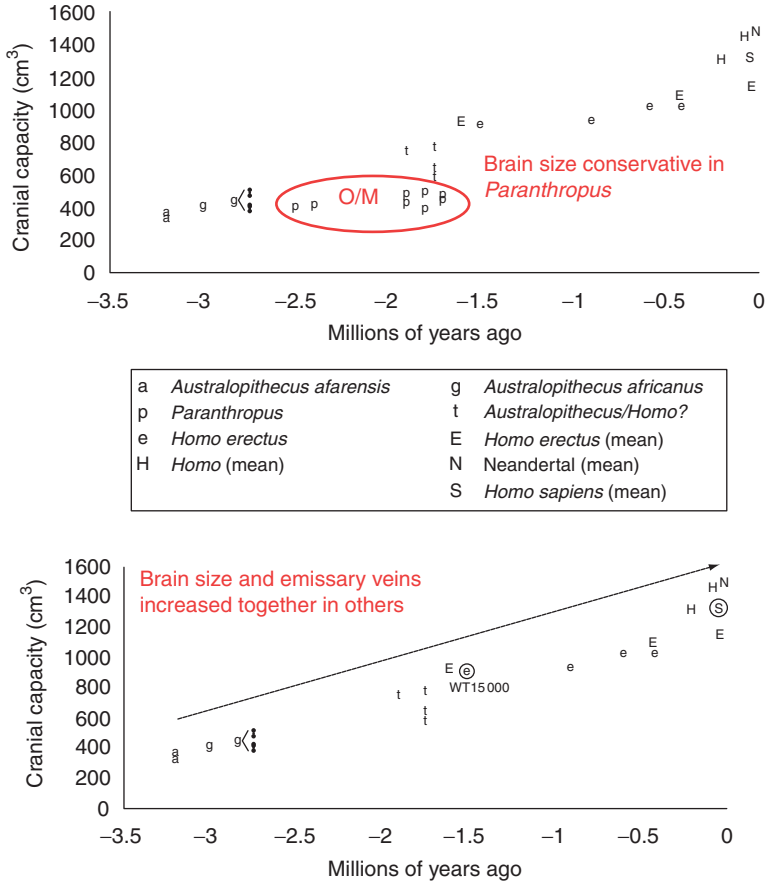

Figure 4 Cranial capacity (as a surrogate for brain size) plotted against time for a variety of fossil hominins and extant *Homo sapiens*. (Top) Brain size remained conservative over time in *Paranthropus*, which evolved an enlarged O/M sinus for draining cranial blood but not an elaborate network of emissary veins entailed in SBC. (Bottom) mounting evidence suggests that brain size increased in the lineage leading from *Australopithecus* to *Homo sapiens* in conjunction with elaboration of the vasculature entailed in SBC in living people. Copyright Dean Falk.

food, water, and sleeping sites. Such difficult savanna habitats were thought to be crucial for the initial evolution of bipedalism as well as for the eventual shaping of advanced cognitive abilities and social complexity in hominins. This view has changed, however. Relatively recent discoveries of bipedal hominins that lived approximately 7.0–6.0 Ma in Chad and Kenya were associated with forested or wooded habitats rather than open savannas, as is now recognized to be the case for australopithecines that lived in Ethiopia slightly before 3.0 Ma, including the famous fossil known as Lucy (*Australopithecus afarensis*). *Homo* is currently thought to have arisen in East Africa at some point between 2.7 and 1.7 Ma, which coincides with a climate that was growing drier and vegetation that was becoming increasingly open – circumstances that were conducive for the savanna living in East Africa once envisioned for earlier hominins. This timeframe coincided with the departure of brain size from that of *Paranthropus* in hominins that appeared to be transitional between *Australopithecus* and *Homo* (Figure 4), as well as the first instances of hominin migration out of Africa.

In contrast to *Paranthropus* and *Australopithecus*, a modern body build had evolved in early *Homo* by 1.5 Ma, as evidenced by the skeleton of a tall, slim 10- to 12-year-old boy from Kenya who had humanlike rather than apelike limb proportions (KNM-WT 15,000 or WT 15K) (Figure 6). The boy's brain volume of approximately 900 cm^3 would have been close to the adult value, and it was twice the average for both types of australopithecines and two-thirds the mean volume of 1350 cm^3 for living people. WT 15K's linear body build is ideal for dissipating heat in hot climates and remains the norm for that area of the world. In addition to walking, humans are adapted for sustained running and most of the structural bases for the latter are indicated by

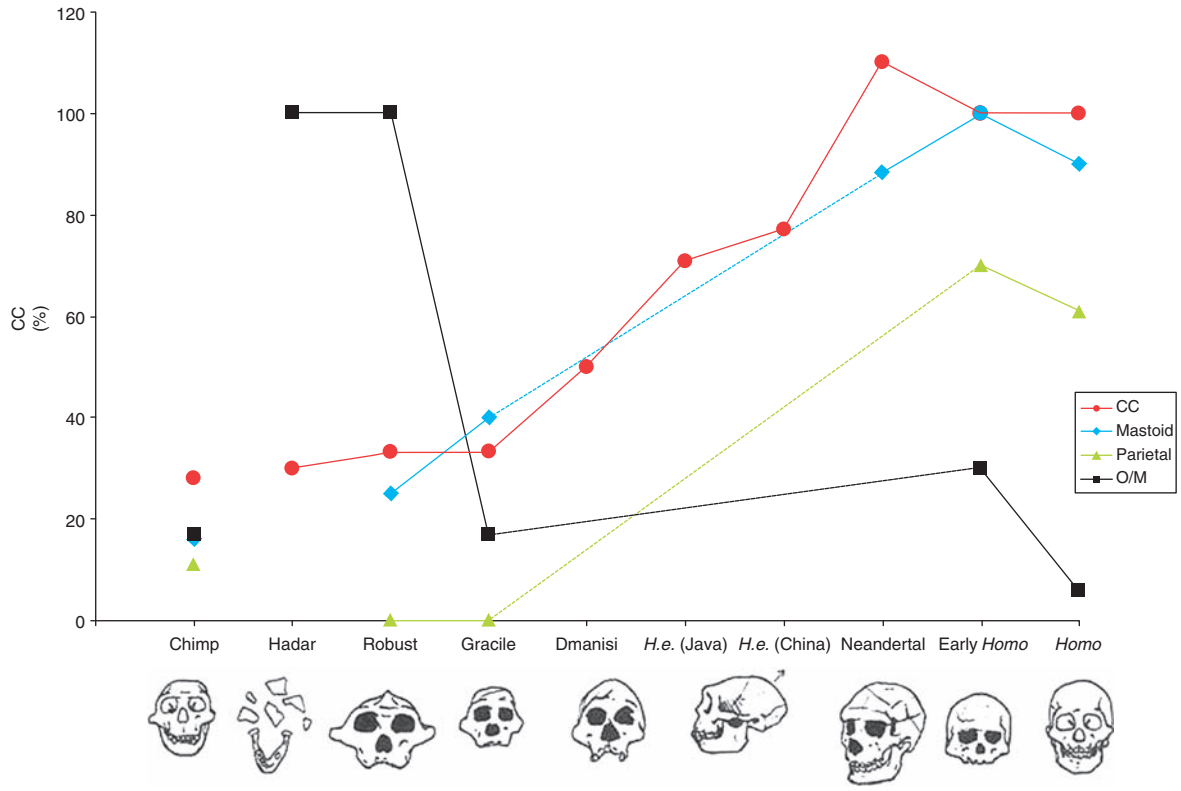

Figure 5 Frequencies of mastoid (blue diamonds) and parietal (green triangles) emissary foramina and enlarged O/M sinuses (black squares) plotted against mean cranial capacities (CC) expressed as percentages of the 1350 cm^3 mean for modern humans (red circles). Dashed lines indicate fossils that have not been scored; robust = *Paranthropus*; gracile = *Australopithecus africanus*. Chimpanzees are included for comparative purposes; hominins are arranged in approximate chronological order from left to right. In Hadar and robust specimens, O/M is fixed; in the latter, emissary foramina occur in low apelike frequencies. The reverse is true for the *Australopithecus* (gracile)–*Homo* lineage in which O/M frequencies fluctuate around those of apes, whereas those for emissary foramina increase through time in conjunction with brain size. Sketch by Brian Worthington; permission by Dean Falk.

WT 15K's skeleton. It is therefore reasonable to hypothesize that *Homo* evolved anatomical specializations that facilitated both walking long distances and endurance running in a thermally stressful climate, perhaps in conjunction with scavenging and chasing down prey (persistence hunting). The increase in emissary foramina associated with increased brain size in the direct ancestors of *Homo sapiens* strongly suggests that SBC was among these specializations.

Summary

Selection for habitual bipedalism in early hominins necessitated a reconfiguration of their cranial vascular systems in response to altered hydrostatic pressures, which was achieved in *Paranthropus* and *Australopithecus* by different adaptations. The former lived in relatively covered environments and evolved an enlarged O/M sinus system for delivering blood from the cranium which was not associated with SBC. Brain size remained conservative throughout the evolution of this now-extinct group. On the other hand, a network of emissary veins began to evolve at some point after 2.7 Ma in association with refinement of bipedalism (and possibly selection for endurance running) in the descendants of early *Australopithecus* that had shifted from relatively closed to more open and thermally stressful environments. This incipient cranial 'radiator' permitted SBC in these hominins, thereby releasing a thermal constraint that had previously kept brain size in check. Brain size continued to enlarge in this group and their *Homo* descendants in conjunction with ongoing elaboration of a venous apparatus for SBC. It is important to note that the evolution of mechanisms for SBC did not directly cause brain size to enlarge in the human lineage; rather, it released a thermal constraint that had previously constrained natural selection for larger brains.

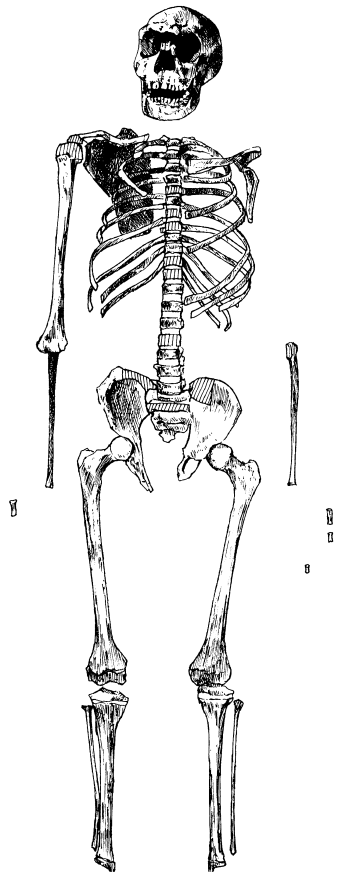

Figure 6 The skeleton of KNM-WT 15 000 is from a young *Homo erectus* boy who lived in Kenya 1.5–1.6 Ma. The 10- to 12-year-old would have grown to more than 6 feet tall and had a brain that was twice the size of the average for *Paranthropus* and *Australopithecus*. Copyright Dean Falk.

See also: Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brain Scaling Laws.

Further Reading

Aiello L and Wheeler P (1995) The expensive tissue hypothesis: The brain and digestive system in human and primate evolution. *Current Anthropology* 36: 199–221.

Baker MA (1979) A brain-cooling system in mammals. *Scientific American* 240: 130–139.

Behrensmeyer AK (2006) Climate change and human evolution. *Science* 311: 476–478.

Boyd GI (1930) The emissary foramina of the cranium in man and the anthropoids. *Journal of Anatomy* 65: 108–121.

Bramble DM and Lieberman DE (2004) Endurance running and the evolution of *Homo*. *Nature* 432: 345–352.

Cabanac M and Brinnel H (1985) Blood flow in the emissary veins of the human head during hyperthermia. *European Journal of Applied Physiology* 54: 172–176.

Caputa M (2004) Selective brain cooling: A multiple regulatory mechanism. *Journal of Thermal Biology* 29: 691–702.

Falk D (1990) Brain evolution in *Homo*: The "radiator" theory. *Behavioral and Brain Science* 13: 333–381.

Falk D (2004) *Braindance: Revised and Expanded Edition*. Gainesville: University Press Florida.

Lillywhite HB (1987) Circulatory adaptations of snakes to gravity. *American Zoologist* 27: 81–95.

Mariak Z, White MD, Lewko J, Lyson T, and Piekarski P (1999) Direct cooling of the human brain by heat loss from the upper respiratory tract. *Journal of Applied Physiology* 87: 1609–1613.

McConaghy FF, Hales JRS, and Hodgson DR (1994) Selective brain cooling in the horse during exercise. In: Milton AS (ed.) *Temperature Regulation: Recent Physiological and Pharmacological Advances*, pp. 189–193. Basel: Birkhäuser.

Nagasaka T, Brinnel H, Hales JRS, and Ogawa T (1998) Selective brain cooling in hyperthermia: The mechanisms and medical implications. *Medical Hypotheses* 50: 203–211.

Walker A and Leakey R (1993) *The Nariokotome Homo erectus Skeleton*. Cambridge, MA: Harvard University Press.

Zenker W and Kubik S (1996) Brain cooling in humans – anatomical considerations. *Anatomy and Embryology* 193: 1–13.

Brain Fossils: Endocasts

R L Holloway, Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The human brain, as any other part of our body, has an evolutionary history, no more or less remarkable than the evolutionary histories that have fashioned the nervous systems of millions of other living organisms on our planet. Yet it remains to be seen whether the trillions of potential neural states that can, through a singular human adaptation, namely, language, link with other brains of the same species and either maintain or exterminate our existence. After all, hundreds of thousands, if not millions, of species have met their terminal fates, extinction, during the past 500 or so million years. The human species, as we recognize ourselves, has had an evolutionary life span of only a few hundred thousand years, at most, whereas many animal genera have had life spans averaging 5–10 My. As a genus, we might have an ancestry of perhaps 2 My; what is left is largely up to our collective use of our brains, and another 5 My seems rather unlikely.

How did we emerge as the primate with the largest brain in absolute size and complexity? What does the fossil record tell us about the evolutionary journey our ancestors embraced during just the last 3–5 My, when we became hominids?

Evolutionary Approaches

Paleoneurology, the Direct Evidence

There are three important approaches to these questions, only one of which we will explore here in more detail. The line of evidence that is direct is paleoneurology, the study of the fossil evidence of the central nervous system, as found on the surface of endocasts. Endocasts are not fossilized brains, but rather the casts of the insides of crania of fossil animals, which in our case mean the hominids of the past 3–5 My known from Africa, Asia, and Europe. Endocasts come in three flavors: natural, artificial (or human-made), and 'virtual' (as generated by a computer from computed tomography (CT) scan data).

Some of the earliest hominid endocasts, such as those from South Africa, such as the infamous Taung child, and known as *Australopithecus africanus* and/or *robustus*, are known from natural endocasts, fashioned over thousands of years by fine sediments entering the postmortem crania, and gradually being compacted and infiltrated by calcareous solutions to become a rock-hard external replica of whatever neuroanatomical details existed on the internal table of cranial bone. Note that in this natural process, the state of the crania will determine how good the replication of surface details will be, that is, how complete the cranial portions are, and whether or not they are distorted by postmortem geological changes, fractures, and/or plastic flow. No thalami, amygdala, hippocampi, cerebellar nuclei, mirror neurons, or fiber tracts such as the corpus callosum, or arcuate fasciculus will ever be preserved. In hominids, then, and indeed, even most primates, fossil or extant, endocasts can only show the surface features of the cerebral cortex, posterior cerebellum, and some of the brain stem to the level of the foramen magnum (Figure 1).

Man-made endocasts are usually made by applying rubber latex, or silicon rubber to the internal cranial surfaces of fossil crania, and these then become the casts which are studied. More recently, because of technical advances in computer imaging and scanning techniques, 'virtual' endocasts can be made, and an endocast model literally carved out of a plastic block using the data set of scanned coordinates collected from CT scans. The same sorry set of restrictions is applicable to these endocasts. Additionally, the resolution of the scans is variable (usually between 1 and 0.5 mm); thus not all details can be collected from a 'virtual' rather than an 'actual' endocast.

So what, then, are the potential valuable insights one can derive from the direct evidence for human brain evolution if one studies endocasts?

Comparative Neurology, Indirect Evidence

Before answering this question, it is important to look at other kinds of evidence. First, there is the study of comparative neurology. It is the most important line of evidence for understanding the relationships between neural and behavioral variation. It compares the size and structures of various animal brains and attempts to relate the differences and similarities to behavior. Most importantly, however, is that the objects of study, the brains, are simply the current end products of past evolutionary events: the comparative record is not an evolutionary sequence or record. Nevertheless, monkeys, chimpanzees, rats, mice, etc., are routinely compared to modern humans as if these were in some way an evolutionary record. They are not. Nevertheless, without comparative neurology, it would be impossible to study paleoneurology (the reverse does not necessarily hold).

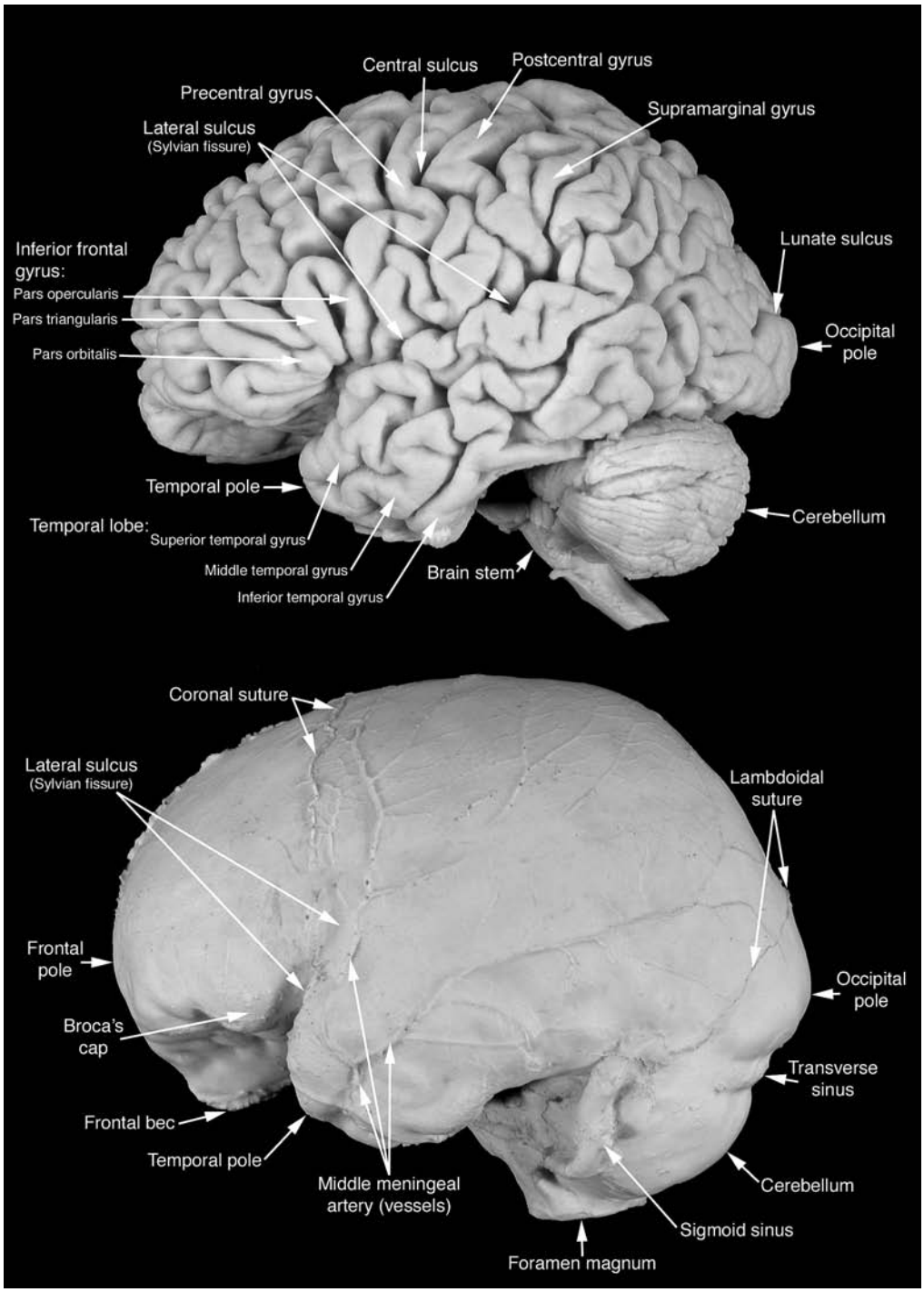

Figure 1 Lateral view of a modern human brain cast and the accompanying endocast. It is obvious that much detail is lost in the endocast.

Molecular Neurogenetics

Another source of information regarding brain evolutionary histories is so new, that little can be said, except that it could well become an important source of direct evidence regarding the evolutionary history of any living animal's brain. This source is molecular neurogenetics. Here, we are beginning to unravel not only what the genetic elements are, but which ones have been under selective pressures in the past.

Data Provided by Endocasts

First, and perhaps of most importance and interest is size, meaning the endocast's volume in cubic centimeters or milliliters. This volume is of course somewhat larger than the volume of the brain, as the endocranial volume (EV) also includes meninges, cerebrospinal fluid, and cranial nerves. The EV is about 8–12% more than the actual brain volume. It is most intriguing to realize that the modern human brain varies in size from about 900 to 2000+ g (remember the old chestnut of Anatole de France at 900 and Swift and Turgenev at about 2000+?), which represent about 1000 g (or cc) which is the total evolutionary change from early australopithecines at 400 ml, to modern human average of about 1400 ml. The average brain weight for modern *Homo sapiens* is 1330 g. Intriguing or not, we now know that there is some correlative relationship between brain size and cognitive abilities at an R of about 0.3, and the fossil evidence for complex cognitive behavior surely shows an increasing complexity through the 2.6 My of hominids' tool productions and later, art works. That said, one cannot, with any degree of certainty, establish connections between selection for brain sizes in hominids and their tools. One reason is that there is a relationship between body size and brain size, and it could well have been the case that selection favored increased body size at certain times and the brain, because of its allometric relationship with body size, simply increased in size, eventually bestowing greater cognitive ability on the hominids so endowed. While rarely the case, postcranial remains are found with crania, and these can provide biological anthropologists some estimates of body size, which in turn can lead to estimates of relative brain size, and when placed within a comparative data set (say all mammals, or all primates) can lead to the calculation of encephalization coefficients (EQs). For example, *A. africanus*, which existed about 2–3 Ma, had an average brain size of about 400–450 cc, roughly the same as chimpanzees, and less than gorillas, yet their body sizes suggest slightly higher EQs than these living apes. However, when we examine these relationships in *Homo erectus* hominids, we find that their body weights were essentially the same as our own, yet their brain sizes were on the order of 750–850 cc. Clearly, there must have been selection pressures operating in the Pleistocene to increase brain size without increasing body size, thus making it reasonably clear that human brain evolution was a mosaic affair, at least as far as brain and body sizes were concerned. Furthermore, it does appear that Neandertal hominids living 200 000 to about 28 000 years ago did have, on average, larger brains than we do, but we also believe that their body sizes were also larger, with more lean muscle mass related to adaptations to colder temperatures. Table 1 provides a listing of hominid brain sizes, by taxonomic group, and their respective EQs. EQs are a relative measure of a species' degree of encephalization calculated from some known allometric data-base. For example, Martin's EQ in Table 1 is calculated from the equation, EQ Martin = brain weight of an animal/0.0991 × (body weight)$^{0.762\ 37}$. The data comes from 88 species of primates where the allometric equation for this data set becomes the denominator of the above equation. The EQ Homo is based on the equation: EQ Homo = brain weight/1.0 (body weight)$^{0.649\ 06}$. Since

Table 1 Hominid brain sizes, by taxonomic group, and their respective EQs

Taxa	*Mean vol.* (cm^3)	*N*	*Range* (cm^3)	*Mean* (*Ma*)	*Body mass* (*kg*)	*EQ Martin*	*EQ Homo*
A. afarensis	445	5	387–550	3.11	37.00	4.87	42.79
A. africanus	462	9	400–560	2.66	35.50	5.21	45.58
P. ethiopicus	431	4	400–490	2.09	37.60	4.66	41.01
A. garhi	450	1	450.00	2.50	NA	NA	NA
H. erectus	941	20	727–1220	0.81	57.80	7.32	67.64
H. ergaster	800	2	750–848	1.74	57.50	6.25	57.72
H. habilis	610	6	510–687	1.76	34.30	7.06	61.50
H. heidelbergensis	1265	12	1150–1450	0.27	68.70	8.64	81.30
H. rudolfensis	788	2	752–825	1.87	45.60	7.35	66.08
H. neanderthalensis	1487	28	1200–1700	0.08	64.90	10.60	99.14
H. sapiens	1330	23	1250–1730	0.01	63.50	9.63	89.90
H. soloensis	1155	7	1013–1250	0.06	NA	NA	NA
P. robustus	493	3	450–530	1.50	36.10	5.49	48.11
P. boisei	515	6	475–545	1.65	41.30	5.17	46.02
P. troglodytes	405	> 100	350–450	0.01	46.00	3.75	33.75
G. gorilla	500	> 100	400–685	0.01	105.00	2.47	24.39

N, number of specimens.

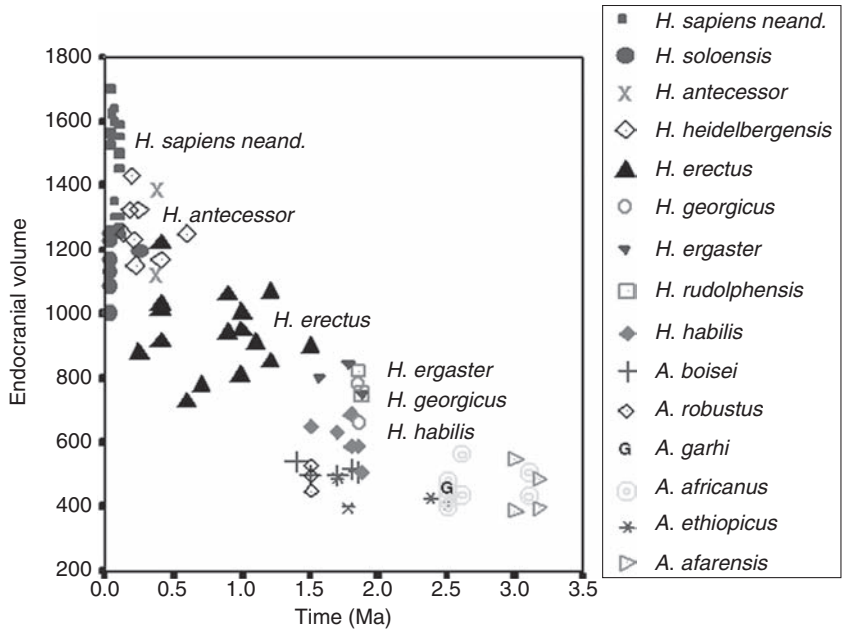

Figure 2 This chart shows the changes through time of hominid EV. It is important to realize that each point itself takes up many thousands of years, and that such a rendition can have little to suggest in terms of actual evolutionary forces and resultant brain changes.

H. sapiens has the highest EQ (whatever method or data set is used), this immediately provides an EQ for an animal as a percentage of modern *Homo*'s value, which is 1.00 or 100. (This equation is based on plotting the log 10 of modern *Homo*'s 1330 brain against his body weight of 65 000, and the intercept is 1.0, since if body weight is zero, so is the brain.)

If the brain sizes of our hominid ancestors of the last 3 My are plotted against time (Figure 2), one finds a curve that appears curvilinear, but which has so many gaps in the fossil record, that it cannot truly represent the complexity of selective forces that operated both on brain and body sizes during the Pliocene and Pleistocene. Of interest, there are data which show that brain size in modern humans has decreased over the last 10 000 years, and should the protein resources of this planet suffer severe degrading, it can be reasonably speculated that brain size would show a further decrease in those regions severely affected. What effect this would have on cognition is not certain.

Second, endocasts provide information about cerebral asymmetries. As LeMay and Galaburda *et al.* showed, cerebral petalias (projections of cerebral cortex into the surrounding cranial table of bone) correlated at roughly 90% with handedness in living modern humans. For example, a torque-like pattern of a protrusion of the left occipital cortex and a rightward greater width of the prefrontal lobe was highly concordant with right-handedness. The opposite pattern correlated with left-handedness, and both patterns could be found with ambidextrous subjects. Again, these relationships were correlations, and the relationships were clearly not 100% obligate. As we know from many neurological studies, including split-brain research, each hemisphere appears to show specializations for cognitive processing, the left hemisphere usually credited with greater analytical reasoning, the right with more holistic, gestalt-like processing. While it is true that asymmetries are found in many different animal brains, this particular pattern of a torque-growth process appears relegated to humans alone, although more research needs to be done with the chimpanzee, our closest animal relative (Figure 3).

Here is where the principle of uniformitarianism is important: if we find similar cerebral patterns on fossil hominid endocasts, are we reasonably entitled to suggest that they also had some degree of cerebral specialization and cognitive processing either very similar to, or identical to our own? Since behavior does not fossilize, it is unlikely we will ever know the full answer. Many archeologists and paleoanthropologists believe, for example, that there could not have been any hominids with language until one finds evidence for symbolic art, as found in the caves of France and Spain, dating to roughly 20 000–30 000 years ago. Others, including this author, conclude that if one finds evidence for cerebral asymmetries, including those for left/right Broca's regions, then one must consider that language could well have been present, at least by 1.8 Ma, since the fossil hominid, *Homo rudolfensis* (KNM-ER 1470, with an EV of 752 cc),

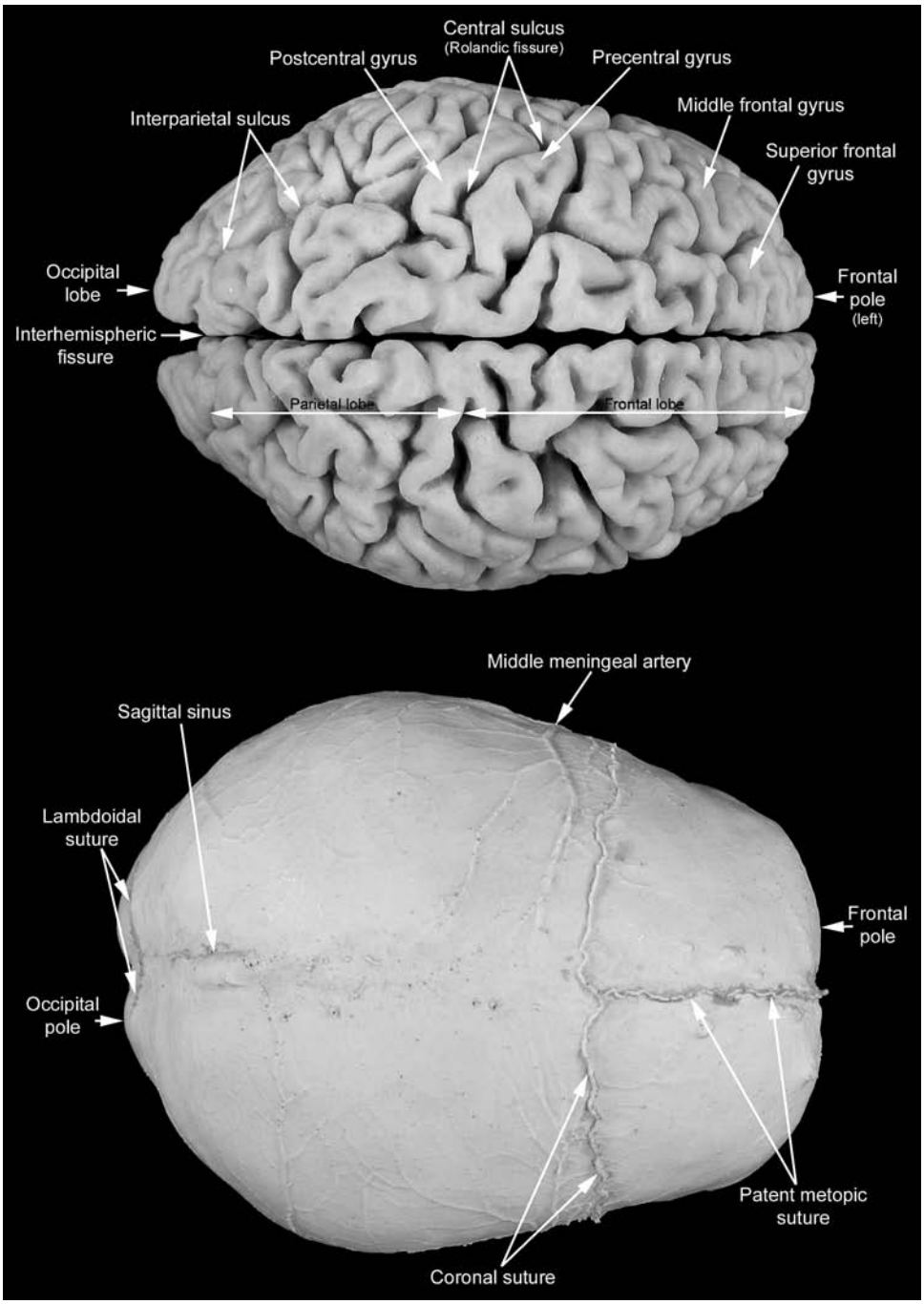

Figure 3 A dorsal view of a modern brain cast and the accompanying endocast. Note that the left occipital lobe extends further posteriorly than the right side, while the lateral width of the right frontal lobe is greater than the left. This is the typical torque-like petalial pattern associated with right-handedness.

shows such asymmetries very clearly. Similarly, such asymmetries can be found in some of the *H. erectus* specimens, and also Neandertals. No art is known for the so-called Clovis peoples of 12 000 years ago in the Americas, just beautifully fashioned spear points. Would one argue that these members of modern *H. sapiens* lacked language? Australian Aborigine stone tools are primitive, but anthropologists still struggle to understand their complex social relationships, learned by Aborigines through their languages. As the author and de LaCoste-Lareymondie showed, the petalial patterns of a combination of left-occipital and right prefrontal width were not found in a large sample of chimpanzees and gorillas, but were found in many of the fossil hominid endocasts. We conclude, that language, surely more primitive than our

own, has been around for at least a million years, if not well before that.

Third, endocasts sometimes do show cerebral morphology, that is, convolutional details that suggest reorganizational changes as well as size increases of the whole brain. For example, relevant to the above suggestions regarding Broca's region, the various endocasts of australopithecines do not show a Broca's cap region (areas 44, 45, 47) that is like our own, whereas later fossil hominids, such as KNM-ER 1470, and all subsequent fossils of the genus *Homo*, do appear to have such a region like our own, at least in terms of endocast morphology. (What the actual cytoarchitecture was is anyone's guess.)

Another important landmark, at least in primate neuroanatomical terms, is the lunate sulcus, which in great apes (as well as all anthropoids) is a sulcus delimiting the anterior extent of primary visual striate cortex (PVC), area 17 of Brodmann from adjacent parietal association cortex. As we know from allometric studies, the modern human PVC is about 120% less in volume than would be expected for a primate of its brain size, and the lateral geniculate body is about 144% less. The question is, when did this reduction take place, and what part of the cortex had to expand relatively to make up for the lost PVC volume? The answer to the latter is evident, and that is that parietal association cortex probably expanded relatively in hominids. The first question, however, has been largely controversial. When Dart published his epochal paper in *Nature* (1925), part of his reasoning for naming the small natural endocast, and fossil crania and mandible as *A. africanus*, was that the endocast showed a lunate sulcus in a relatively posterior position compared to apes, that is, more of a human-like posterior position. This led Dart, a student of Grafton Elliot Smith, the neuroanatomist who helped define the lunate sulcus, to conclude that even a couple of million years ago, the hominid brain had undergone some cerebral reorganization from an ape-like ancestor. Modern humans seldom show a complete lunate sulcus, and, when present, it is mostly fragmented, and is not a completely perfect delimiter of PVC, although it does occur in a relatively posterior position (Figures 4 and 5).

As mentioned earlier, not all convolutions impress onto the internal table of cranial bone, and reading some of the dim imprints can be very challenging, and controversial. The sulcus in question is not without ambiguity, but the placement of the lunate sulcus in more anterior positions either converts the Taung hominid to a monkey, or violates the remaining morphology. As we showed in several publications, placing the lunate in a typical chimpanzee position violates the remaining morphology. Nevertheless, these two examples, Broca's regions and the lunate sulcus, demonstrate that the hominid brain became reorganized to a human-like pattern early in hominid evolution, and well before the final enlargement of the cerebral cortex as in modern *H. sapiens*.

Fourth, one can measure the endocasts using CT scans or plain calipers, rendering large data sets with chord and arc measurements, and then apply sophisticated multivariate analyses to the data sets, in the hope that combinations of measurements might reflect some taxonomic shape differences that might

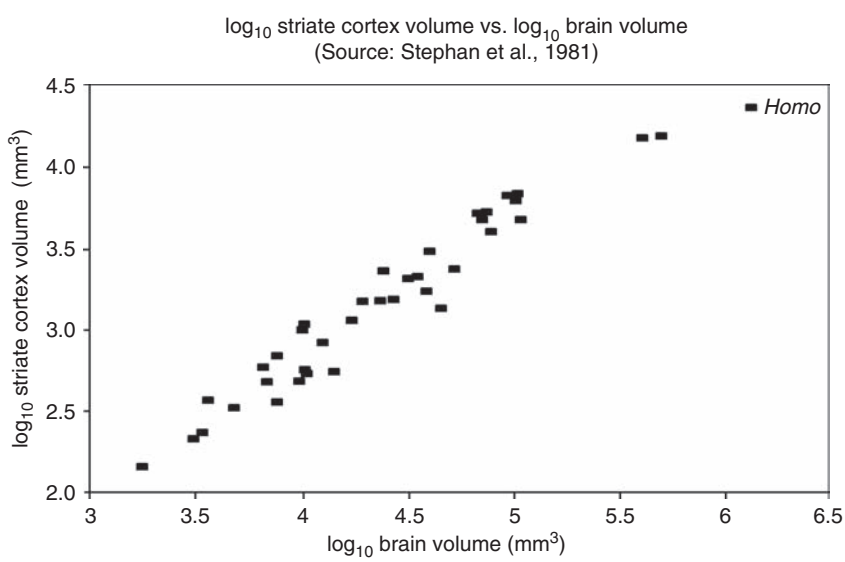

Figure 4 This shows a typical allometric relationship between total brain size and a part of the brain, primary visual striate cortex (area 17 of Brodmann). The log–log relationship is obviously linear, and the human point is about 121% lower than would be expected based on the rest of the primate data.

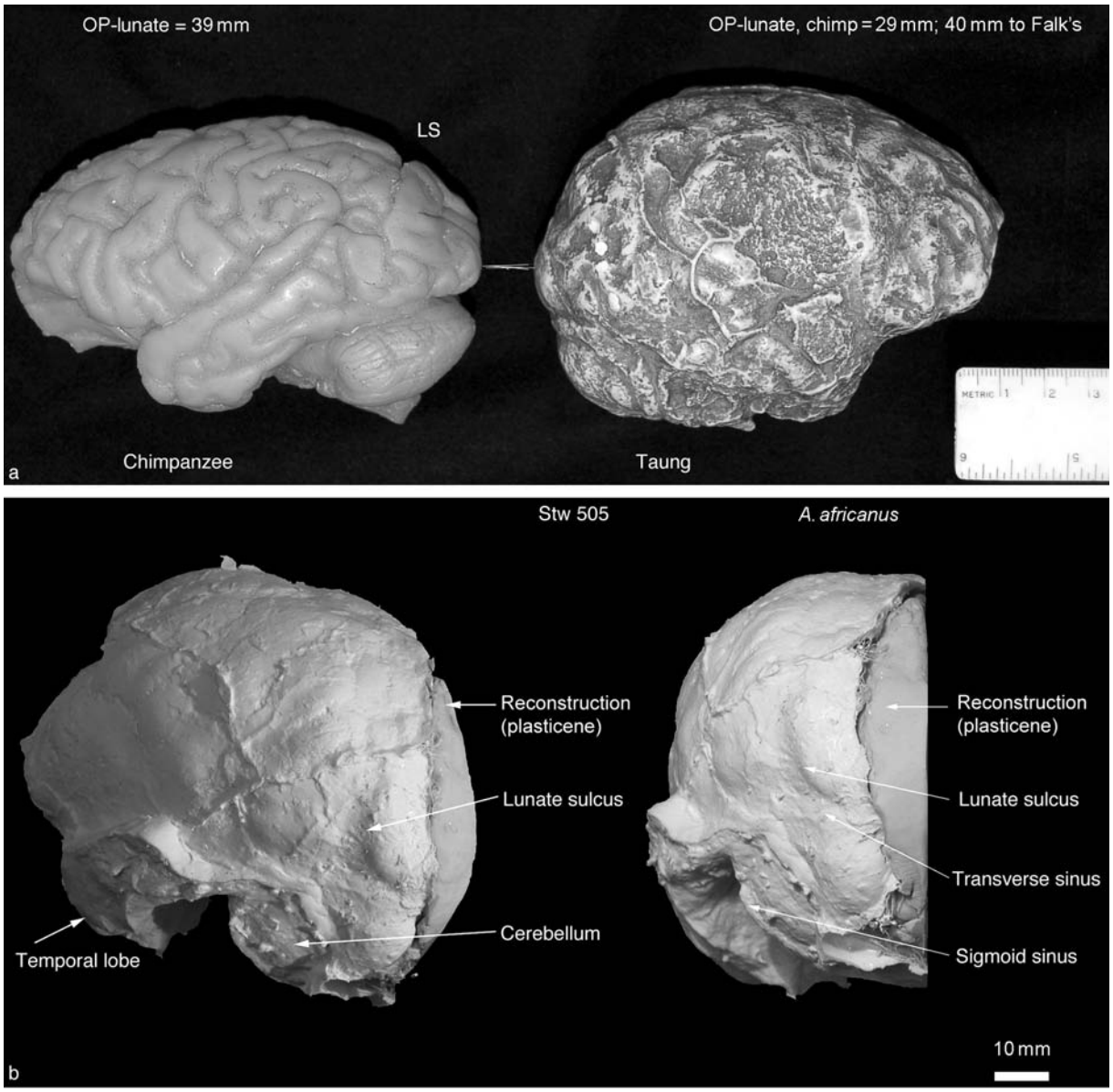

Figure 5 (a) This shows the contrast between a typical chimpanzee (*Pan troglodytes*) brain and that of the natural endocast of the Taung fossil (*Australopithecus africanus*) from South Africa. When the typical chimp position for the lunate sulcus (LS) is placed on the Taung endocast, it violates the sulcal morphology, indicating that the lunate was not in a typical ape position on the Taung child. Placing the LS more anteriorly reduces the Taung child to a monkey status, which is clearly wrong. (b) Views (oblique, occipital) of the new Stw 505 *A. africanus* endocast, showing a true lunate sulcus in a relatively posterior position, thus indicating reorganization of the cortex prior to any significant brain enlargement.

be in turn related to function. Alas, this is not the case. All of the measures and the resulting indices largely, if not completely, reflect the fact that the measurements are completely correlated to the endocasts' volumes. Still, to the extent that endocast shape might have some taxonomic value, or help with reconstructions of missing regions, these measurements do have value.

Fifth, the meningeal patterns of blood supply to the meninges through the middle meningeal artery (mostly) do show variability between hominid groups. In general, as the brain enlarged, the pattern of meningeal imprints on hominid endocasts shows an increase in complexity. Unfortunately, these patterns have no known association with behavioral variables.

These five sets of data are what one can find by studying endocasts. As animals differ more, say between mammals, reptiles, amphibians, birds, and fish, brain endocasts show very different patterns of neurological structures, and the interested reader will

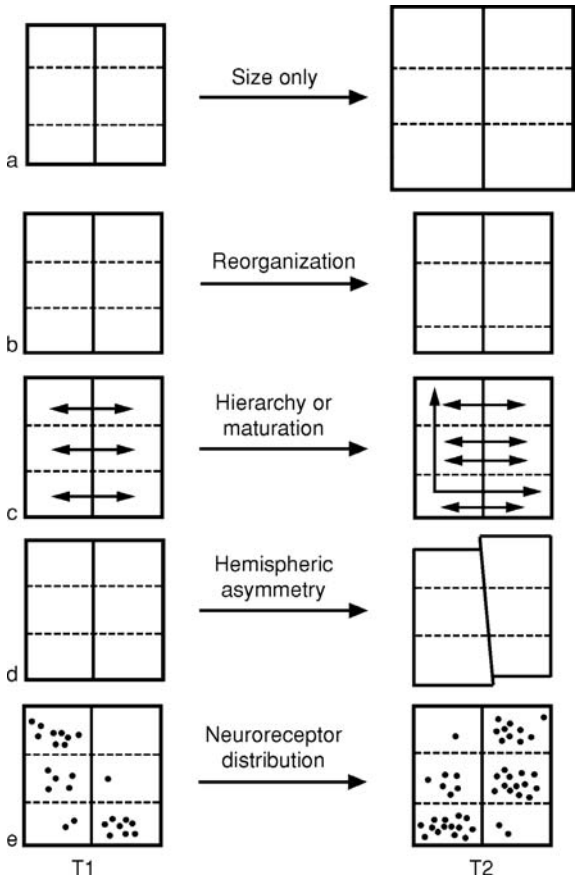

Figure 6 These boxes represent some possible scenarios of brain changes from 'time 1' to 'time 2'. Scenarios (b–e) suggest reorganizational changes that may not be visible on endocasts, although (b) and (d) are variably present on endocasts. (b), For example, symbolizes a reduction of primary visual striate cortex, with a posteriorly placed lunate sulcus, and (d) shows the typical torque-petalial pattern we associate with right-handedness. Neither (c) or (e) are likely to show details on the cortical surface, as the changes could be related to fiber connections, or to subcortical structures not appearing on endocasts.

find a rich literature of comparative neuroanatomy covering these and many invertebrate species (Figure 6).

Conclusions

Clearly, with so few fossil remains discovered thus far, and these spanning a time period of perhaps 5 My, our understanding of how our brains evolved must be incomplete, and largely speculative. In the past, and continuing today, it is brain size that appears to capture most scientists' minds. The direct paleoneurological evidence tells us, however, that our brain evolution picture had to have been far more complex, as we have clear evidence of reorganizational changes interspersed with brain size changes, some of which were absolute, and some relative, that is, dependent on body size increases or decreases. The picture, most importantly, is not a lineal one, but one in which the rates of change were not constant but shifting with selection pressures, immigration, genetic drift, mutations, and environmental challenges involving a most complex interplay of climates, ecologies, prey and predators, as well as social and material interactions within and between groups. Brain evolution in hominids was a mosaic of mosaics. Measuring the size of brain endocasts is the least difficult task in unraveling the evolutionary changes through time. Far more difficult, yet equally important, is discovering the organizational changes that occurred, and when these occurred. The picture starts out with brain sizes within the range of our closest relative, the common chimpanzee, *Pan troglodytes*, with an average brain size of about 385–400 cc. This is apparent in australopithecine hominids, starting with *Australopithecus afarensis* at about 3–3.5 Ma. However, one of these endocasts shows a lunate sulcus in a posterior position quite different than the anterior position found in apes, including the Miocene *Proconsul africanus* fossil. The frontal lobes of this australopithecine, as well as the later derived *A. africanus* group, are narrow, and the Broca's region is not human like. That reorganizational change occurs much later, being present in *Homo* lineages (such as *H. erectus*, *H. ergaster*, *H. antecessor*, *H. habilis*, *H. rudolphensis*, *H. georgicus*, *H. heidelbergensis*, and *H. sapiens neandertalensis*) at about 1.8 Ma, with a brain size of 752 cc. Thus, two critical reorganizational changes took place before the major increase in brain and body size occurred. The author believes that those changes relate to social skills including primitive language being evolved, setting a positive feedback pattern, in which behavioral complexity and brain size are inextricably linked. After ~1 Ma, brain sizes increase to modern values, while the body size remains similar to our own, and the archeological records suggest increasing social and material (stone tools, etc.) behavioral complexity occurring through this time. Perhaps other reorganizational changes in the brain, refinements of cortical structures, acquisition of specific socially relevant behavioral modules accompanied these size increases. What these were cannot at this time be gleaned from the direct endocast evidence. On these questions, the evidence is mute and perhaps forever moot.

See also: Allometric Analysis of Brain Size; Brain Asymmetry: Evolution; Brain Development: The Generation of Large Brains; Brain Modules: Mosaic Evolution; Brain Scaling Laws; Brains of Primitive Chordates; Evolution of Vertebrate Brains; Social Brain: Evolution.

Further Reading

Anderson B (2003) Brain imaging and *g*. In: Nyborg H (ed.) *The Scientific Study of General Intelligence. Tribute to Arthur R. Jensen*, pp. 29–40. Amsterdam: Pergamon.

Connolly CJ (1950) *External Morphology of the Primate Brain.* Illinois: CC Thomas.

Dart R (1925) *Australopithecus africanus*: The man-ape of South Africa. *Nature* 115: 195–199.

Edinger T (1929) Die fossilen Gehirne. *Ergebneise Anatomische und Entwicklungen Geschichte* 28: 1–221.

Edinger T (1949) Paleoneurology versus comparative brain anatomy. *Confina Neurologica Separatum* 9: 5–24.

Galaburda AM, LeMay M, Kemper TA, and Geschwind N (1978) Right–left asymmetries in the brain. *Science* 199: 852–856.

Henneberg M (1998) Evolution of the human brain: Is bigger better? *Clinical and Experimental Pharmacology and Physiology* 25: 745–749.

Hirschler P (1942) *Anthropoid and Human Endocranial Casts.* Amsterdam: N.V. Noord-Hollandsche Uitgevers Maatschappij.

Holloway RL (1967) The evolution of the human brain: Some notes toward a synthesis between neural structure and the evolution of complex behavior. *General Systems* 12: 3–19.

Holloway RL (1975) *The Role of Human Social Behavior in the Evolution of the Brain. 43rd James Arthur Lecture on the Evolution of the Human Brain.* New York: American Museum of Natural History.

Holloway RL (1985) The past, present, and future significance of the lunate sulcus in early hominid evolution. In: Tobias PVT (ed.) *Hominid Evolution: Past, Present, and Future*, pp. 47–62. New York: AR Liss.

Holloway RL (1996) Evolution of the human brain. In: Lock A and Peters C (eds.) *Handbook of Human Symbolic Evolution*, pp. 74–116. New York: Oxford University Press.

Holloway RL, Broadfield DC, and Yuan MS (2004) Schwartz JH and Tattersall I (eds.) *The Human Fossil Record, Vol. 3: The Human Fossil Record: Brain Endocasts: The Paleoneurological Evidence.* New York: Wiley-Liss.

Holloway RL and de LaCoste-Lareymondie MC (1982) Brain endocast asymmetry in pongids and hominids: Some preliminary findings on the paleontology of cerebral dominance. *American Journal of Physical Anthropology* 58: 101–110.

Jerison H (1973) *Evolution of Brain and Intelligence.* New York: Academic Press.

Kotchetkova VI (1978) *Paleoneurology.* Washington, DC: VH Winston.

LeMay M (1976) Morphological cerebral asymmetries of modern man, and nonhuman primates. *Annals of the New York Academy of Science* 280: 349–366.

Preuss TM (2000) Taking the measure of diversity: Comparative alternatives to the model-animal paradigm in cortical neuroscience. *Brain, Behavior, and Evolution* 55: 287–299.

Preuss TM, Cáceres M, Laucher J, et al. (2003) Using genomics to identify human brain specializations. *American Journal of Physical Anthropology Supplement* 36: 171.

Rilling JK (2006) Human and nonhuman primate brains: Are they allometrically scaled versions of the same design? *Evolutionary Anthropology* 15: 65–77.

Brain Glucose Metabolism: Age, Alzheimer's Disease, and ApoE Allele Effects

A Pupi, M T R De Cristofaro, B Nacmias, and S Sorbi, University of Florence, Florence, Italy
L Mosconi, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Technical Issues in Brain Glucose Metabolic Imaging in Aging and Degeneration

The cerebral metabolic rate of glucose (CMRglc) can be measured *in vivo* using positron emission tomography (PET) imaging with [^{18}F]-2D-fluoro-deoxiglucose (FDG) as the tracer. Before we present FDG-PET findings in aging individuals and Alzheimer's disease (AD) patients, some technical issues relevant to the use of PET in these populations should be described.

First, FDG is taken up in the brain at a gray matter: white matter ratio of 5:1. The cortical gray matter ribbon thickness varies between 1 and 4.5 mm, with an average of approximately 2.5 mm, which is well below the 4.25 mm spatial resolution of current state-of-the-art PET scanners. Therefore, the FDG-PET signal in the cortex spills out to the surrounding low-activity regions (white matter and cerebrospinal fluid, CSF). This effect is called partial volume effect (PVE). PVE can be thought of as a blurring of higher-activity gray matter structures with the surrounding lower-activity spaces. PVE generates an artifactual reduction of FDG-PET activity as the cortex gets thinner (Figure 1). This is relevant to elderly subjects and AD patients, whose brains may show cortical atrophy (i.e., cortical thinning due to neuronal loss) and ventricular enlargement. It is possible to correct for PVE using mathematical models and co-registration with magnetic resonance imaging (MRI), which offers a structural image of the distribution of gray and white matter and the CSF. Partial volume correction adjusts FDG uptake values as a function of the underlying tissue volume, therefore increasing CMRglc values in atrophic regions, and provides a measure of metabolic activity per gram of brain tissue. Despite the technical importance of PVE, several FDG-PET studies have shown that CMRglc reductions in AD patients compared to controls remain significant after PVE correction and therefore reflect true reductions of CMRglc.

Second, AD is associated with a phenomenon called selective neuronal vulnerability, in which specific regions are vulnerable to neurodegeneration and others are spared. CMRglc reductions are found in the brain regions most affected by AD pathology (i.e., amyloid-beta and tau pathology, and neuronal loss), for example, the medial temporal lobes (MTLs), including the hippocampus (HIP), entorhinal cortex (EC) and transentorhinal cortex, and parahippocampal gyrus; the posterior cingulate cortex (PCC); the parieto-temporal cortex (PTC); and the frontal association cortex (FC). Traditionally, FDG-PET scans in aging and AD were sampled using manually drawn regions of interest (ROI) of the brain structures known to be selectively involved, either on PET-co-registered MRI or directly on PET. However, the method is subject to high inter- and intra-operator variability. To overcome some of these issues, several methods of automatic ROI extraction and cross-modality co-registration have been developed, and their application guarantees a better reproducibility and higher anatomical accuracy of the results.

A totally different approach is based on voxel-based analysis (VBA), a family of techniques that analyzes brain images at the single voxel (the image-unitary 3D element) level to identify regions showing significant CMRglc reductions between groups of subjects by means of intrasubject or intersubject statistical comparisons and generation of statistical parametric maps. Because human brains differ in size, shape, and orientation, VBA methods adequately modify these features by virtue of warping and spatial normalization algorithms so that all images conform to a standardized brain shape in the stereotactic space (i.e., a template image) (Figure 2). After spatial normalization, statistical analysis is performed on a voxel-by-voxel basis, correcting for confounding variables known to affect CMRglc (e.g., age, gender, and education) by standard multiple linear regressions (Figure 3). Among these methods, the most widely used for FDG-PET analysis in dementia are Statistical Parametrical Mapping (SPM) (Wellcome Department of Neurology, London, UK) and NEUROSTAT (University of Washington, Seattle, WA, USA). Moreover, diagnostic VBA methods have also been developed to perform an automated comparison of a single patient's scan to a database of reference FDG-PET scans to facilitate clinical decision (see the sections titled 'Brain glucose metabolic alterations: operating characteristics in Alzheimer's disease and mild cognitive impairment' and 'FDG-PET: Work in progress').

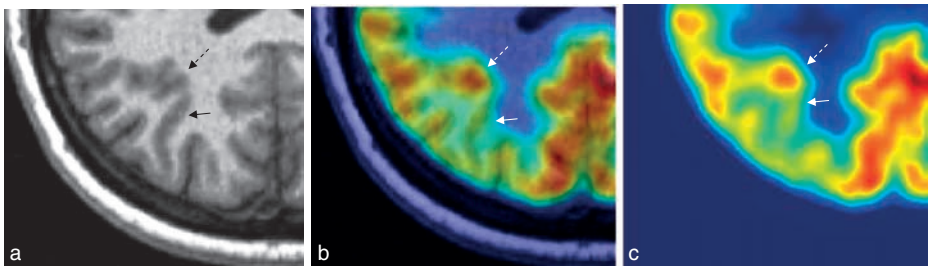

Figure 1 Visual depiction of partial volume effect in the temporal–occipital region of a representative middle-age normal subject: (a) MRI; (b) co-registered MRI and FDG-PET; (c) FDG-PET. FDG uptake is represented on a green to red color scale, with red indicating high FDG uptake (i.e., high local metabolic activity). Among other factors, visualization of FDG uptake is related to the thickness of the cortex. Thinner cortical gyri show lower signal intensities (solid arrows) than do thicker cortical gyri or areas of convergence of different gyri (dotted arrows), which show higher metabolic activity. CMRglc may, in fact, be comparable in the two regions, and the apparent difference in intensity depends on the response of the PET scanner to the different thickness of the activity source (the cortex). CMRglc, cerebral metabolic rate of glucose; FDG, [^{18}F]-fluoro-2-deoxiglucose; MRI, magnetic resonance imaging; PET, positron emission tomography.

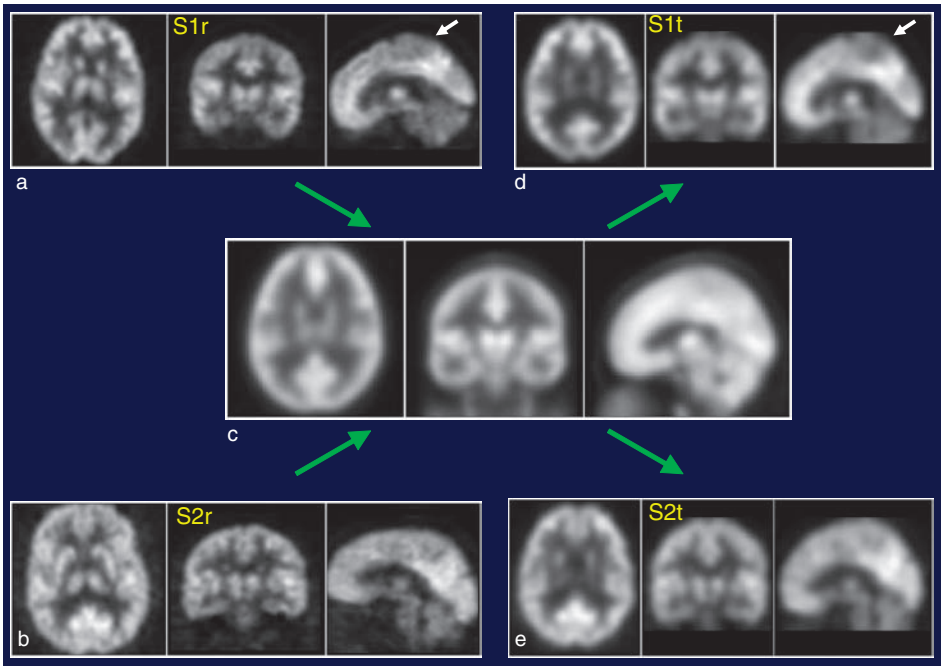

Figure 2 Spatial normalization with voxel-based analysis (VBA): (a–b) raw images from two subjects, S1 and S2; (c) FDG-PET template from the Montreal Neurological Institute; (d–e) normalized images using SPM2. Spatial normalization is the key point to apply VBA. In (a) and (b), the two brains differ in their size, shape, and orientation. The VBA software contains a mathematical algorithm (usually very complex) which resizes, reshapes, and reorients the brain, minimizing, as the cost function, the differences using a FDG-PET template (c). After normalization (d–e), the subjects' brain images have the same size, shape, and orientation, but the metabolic differences are maintained (white plain arrow in (a) and (d), showing S1's brain cortical region with low metabolism). FDG, [^{18}F]-fluoro-2-deoxiglucose; PET, positron emission tomography; S1r, subject 1 raw images; S1t, subject 1 transformed images; S2r, subject 2 raw images; S2t, subject 2 transformed images; SPM, Statistical Parametrical Mapping.

Molecular and Cellular Factors Involved with Glucose Hypometabolism in the Aging Brain and Alzheimer's Dementia

There seems to be very little or no relationship between CMRglc decreases and the density of pathological hallmarks of AD: amyloid-beta plaques and neurofibrillary tangles. Neuronal loss is thought to be the most likely cause of decreased FDG uptake, although CMRglc decrease seems to be an early event in AD, possibly occurring prior to neuronal loss. Moreover, there is a significant CMRglc decrease during normal aging, whereas age-related neuronal loss is minimal.

Another molecular factor that has been implicated in aging and AD, and their related CMRglc reductions,

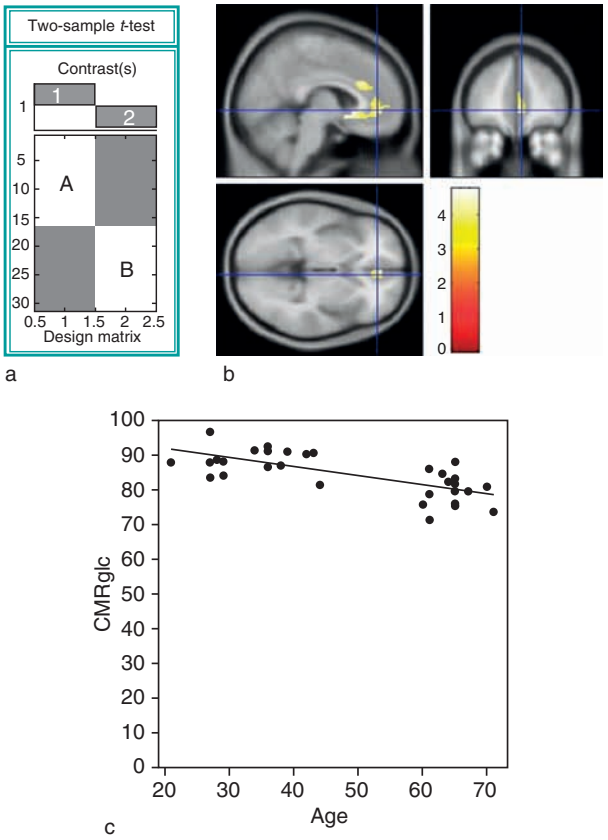

Figure 3 Statistical analysis of the effects of aging on metabolic activity: (a) the basic components of the analysis, the proposed contrast and the groups; (b) the region(s) in which the metabolic activity is significantly different in accord with the proposed contrast; (c) CMRglc in each subject's ACC, normalized to the mean cortical value, plotted versus the subject's age. The second phase of voxel-based analysis (VBA) is the statistical analysis, performed after spatial normalization. This figure presents a basic SPM analysis. The example is a comparison of two groups of normal subjects, a group A of 16 subjects with mean age 34.3 ± 6.1 (range 21–44 years) and a group B of 15 subjects with mean age 64.9 ± 3.0 (range 60–71 years). The analysis is a *t*-test. The contrast applied in the study is shown at the top of (a) by the two small gray labels (1 and 2). In the lower part of (a), the two large white columns (A and B) indicate the subjects of the two groups. The two white columns indicate the number of subjects in each group (*y*-axis). The two gray labels indicate the direction of the contrast. In this study, the regions searched for are those in which the metabolism is higher in the younger group than in the older one (upward direction of the gray label 1 and downward for label 2). In (b), the regions in which the metabolic activity is significantly different in accord with the proposed contrast are shown in a scale which reports the value of *t* (0–5 in the color bar). In this example, the anterior cingulate cortex (ACC) fulfills the contrast and is colored yellow. The colored region is represented on a magnetic resonance template. As shown in (c), the slope of the regression line is 6.9% per decade. CMRglc, cerebral metabolic rate of glucose.

is oxidative stress (OS). OS is the result of a complex of damaging events correlated with the energetic production in neurons and astrocytes, and its disruptive effects may be differently expressed depending on individual brain compensatory capacity, which in turn is determined by genetic, environmental, and lifestyle factors. Over time, OS can impair cellular energy production at the mitochondrial level. Such impairment has been supposed to be the primary cause of CMRglc decreases on FDG-PET, a hypothesis contradicted by the known activity of allosteric (due to feedback activity) mechanisms which are activated by cellular energy demand and increase cellular glucose intake. Actually, AD neurons *in vitro* show an increased, not a decreased, glycolytic consumption to compensate for decreased energy production.

Synapse dysfunction and loss are instead currently seen as the most probable cause of CMRglc reduction in aging and AD. First, synapse molecular impairment, structural changes, and loss are early events in AD, as well as, to a lesser extent, in normal aging and senescence. Hippocampal synapse loss and dysfunction have been detected before neurofibrillary tangles in human tau transgenic mice. Second, there are sound observations linking local glucose intake with the activity of glutamatergic synapses, which constitute 80% of total brain synapses. A decline in synaptic function decreases local glutamate release and may therefore reduce glucose intake. Third, synaptic loss, like FDG uptake, closely correlates with cognitive deficits in AD. Finally, animal studies with synaptophisin (i.e., a presynaptic vesicle membrane protein used as a marker of synaptic density) have shown a significant positive correlation between synapse density and regional FDG uptake. Overall, these findings strongly support the hypothesis that resting-state CMRglc measures on FDG-PET reflect integrated synaptic activity and suggest that FDG-PET imaging may be ideal tool for investigating changes in brain function in aging and AD.

Several other factors may be implicated in FDG-PET changes in aging and AD. For example, amyloid-beta is deposited in the walls of cerebral arterioles in normal individuals and AD patients, causing vascular damage that may in turn impair glucose supply. Moreover, although it is generally accepted that FDG-PET measures are an index of neuronal activity, it is still debated whether and to what extent astrocytes play a role in the observed signal. Astrocytes, or neuroglial cells, are hypothesized to be responsible for the anaerobic component of the glycolytic path, converting glucose into lactate, which is then shuttled into neurons for further metabolism. This hypothesis is currently being tested. In this case, FDG uptake in the brain is limited to astrocytes, and FDG-PET images represent astrocytic, instead of neuronal, glucose metabolic rate. The observation of a tenfold prevalence of glial cells with respect to neurons in the brain and the poor understanding of nerve versus glial energy metabolism need further investigation.

Brain Glucose Metabolism Changes with Aging

Aging is the most important risk factor for late-onset sporadic AD (LOSAD). Therefore, aging has been a point of interest since the beginning of FDG-PET studies in AD. Early FDG-PET studies that visually compared healthy elderly with young or middle-age normal subjects failed to show significant FDG-PET changes associated with age. The negative reports are probably due to the fact that age-related CMRglc changes are subtle and difficult to identify by visual inspection (Figure 4) on low-resolution PET scanners, such as those used in the early years. More recently, with improvements in PET spatial resolution and the use of more sophisticated analytical methods like VBA, age-related CMRglc changes have been detected. Cross-sectional FDG-PET studies have shown that normal elderly subjects, compared to young controls, show CMRglc decreases in the association neocortex, particularly in the FC and the anterior cingulate cortex (ACC) (Figure 3), and also in the PTC regions that are usually hypometabolic in clinical AD. On the other hand, the PCC, which is usually hypometabolic in AD, is apparently not affected by aging. These regional changes are paralleled by global CMRglc reductions of approximately 6% per decade (Figure 3), as demonstrated by quantitative CMRglc measurements. Interestingly, this rate of CMRglc decrease is consistent with the estimated rate of decrease in synaptic density measured postmortem in healthy elderly brains. This correspondence further suggests that FDG-PET measures and synaptic density are strongly related.

Longitudinal FDG-PET studies that monitored changes in cognitive performance in normal healthy subjects showed that CMRglc changes in the PTC correlate with cognitive decline over time. Moreover, there is evidence that CMRglc reductions in the EC and HIP among normal elderly accurately predict future decline from normal cognition to mild cognitive impairment (MCI), often a prodrome to AD, and dementia. These studies were based on the detection of group differences and did not examine whether individual patterns of baseline FDG-PET alterations may be predictive of clinical outcome in individual subjects. Nonetheless, longitudinal FDG-PET studies of normal elderly individuals followed for 6–14 years with serial FDG-PET exams have shown that the rate of CMRglc decline per year is highly predictive of clinical outcome. Longitudinally, in stable normal individuals the rate of hippocampal CMRglc reductions was less than 1% per year, whereas patients who declined to MCI and AD had significantly higher rates of CMRglc reductions of 2.4% and 4.4% per year, respectively. A similar effect was also found in the PCC in patients who declined to AD.

Overall, these data show that age-related CMRglc changes are found also in the brain regions showing CMRglc deficits in AD and that metabolic differences between healthy brain aging and AD may be more quantitative than topographic in nature (i.e., the magnitude rather than the site of CMRglc decrease may be important). Important issues stem

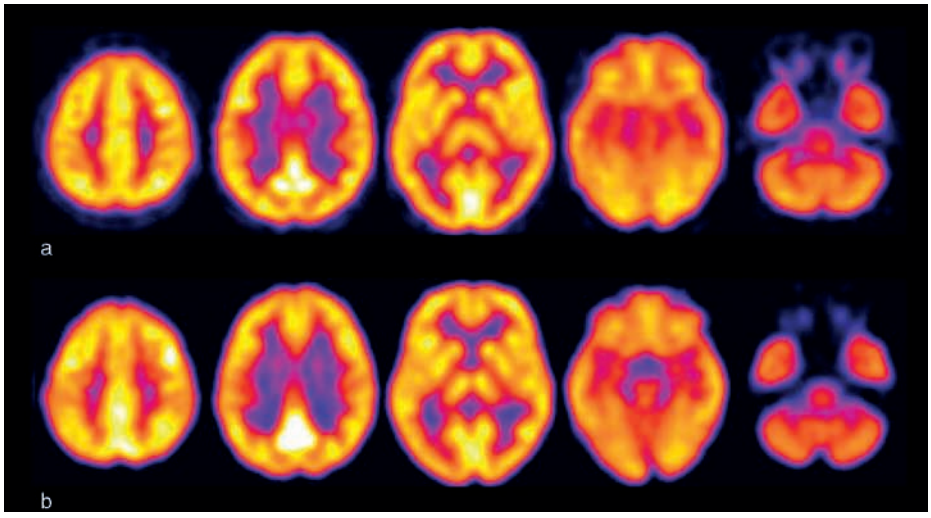

Figure 4 FDG-PET changes in the brain with age: (a) young control (35 years old); (b) elderly control (65 years old). Axial FDG-PET images of the two subjects are shown from vertex (left) to cerebellum (right). The color scale is from blue (low activity) to white (high activity.) A visual search for differences between two subjects is not able to find differences. The brains were spatially normalized with SPM2. FDG, [^{18}F]-fluoro-2-deoxiglucose; PET, positron emission tomography; SPM, Statistical Parametrical Mapping.

from this consideration. First, deviations from age-associated CMRglc declines and CMRglc declines related to MCI and AD may be determined, on an individual basis, by the imbalance between damaging events and protective capacity, a theory currently at the basis of the pathogenesis of AD. Second, in operative terms, it is necessary to take into account the effects of age in any clinical FDG-PET studies aimed at identifying AD as opposed to normal aging.

Brain Glucose Metabolic Alterations in AD: Pathophysiology

In the early 1980s, the clinical use of FDG-PET was restricted to epilepsy and brain tumors. But its potential in the study of degenerative brain diseases became clear as soon as an AD-specific distribution pattern of the CMRglc deficits was discovered.

Afterward, a more systematic approach to the examination of AD was made possible by the development and application of automated image analysis procedures, which enriched the FDG-PET evaluation of more objective criteria based on the statistical detection of CMRglc differences between AD and normal subjects. AD patients exhibit early CMRglc reductions in the PCC and in the PTC (Figure 5), followed by FC hypometabolism found in severe dementia. Quantitative measurements of CMRglc with FDG-PET have also shown that a widespread metabolic reduction underlies regional changes, partially sparing the basal ganglia and cerebellum. These signs are like a signature of AD in the brain and may

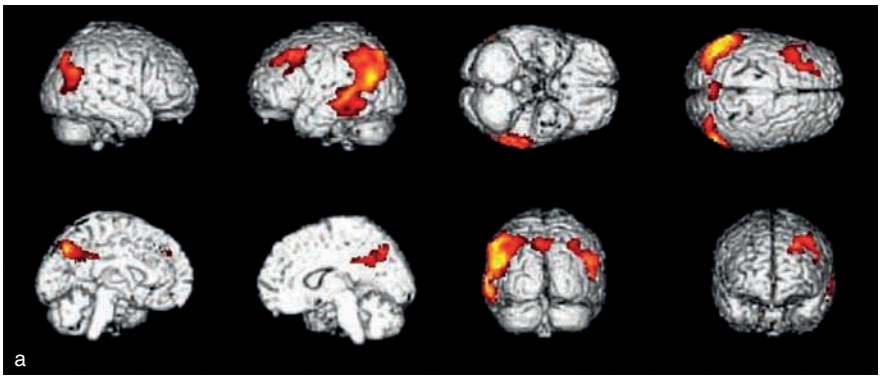

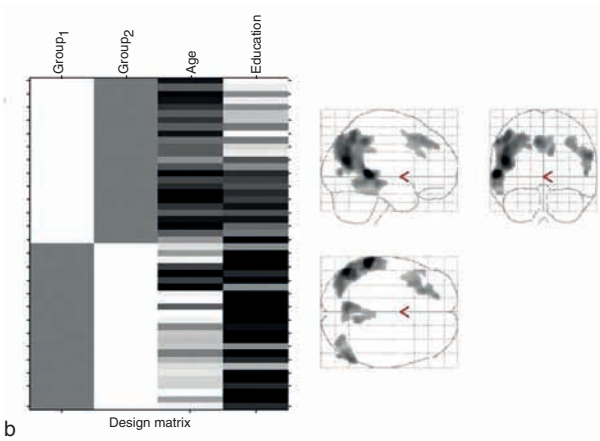

Ke	Coordinates			Z	Functional area	BA
3905	−60	−40	4	6.57	Middle temporal gyrus	22
	−48	−68	20	6.25	Middle temporal gyrus	39
	−48	−60	46	5.72	Inferior parietal lobule	40
858	−6	−66	44	5.59	Precuneus	7
	10	−54	30	5.17	Cingulate gyrus	31
	4	−38	32	4.79	Cingulate gyrus	31
826	58	−62	26	5.56	Superior temporal gyrus	39
	54	−76	6	5.19	Middle occipital gyrus	19
	32	−74	42	5.13	Precuneus	19
724	−26	28	44	5.23	Middle frontal gyrus	8
	−26	38	40	5.14	Middle frontal gyrus	8
	−36	20	46	5.13	Middle frontal gyrus	8
183	46	2	38	5.21	Middle frontal gyrus	6
	50	18	40	5.19	Middle frontal gyrus	8
89	54	−54	44	4.84	Inferior parietal lobule	40
32	60	−52	−8	4.78	Inferior temporal gyrus	37

c

Figure 5 Population study comparing brains of normal controls (NCs) and Alzheimer's disease patients (ADs), with age and education as confounding covariates: (a) significantly hypometabolic regions in ADs are depicted in red-to-yellow scale on a magnetic resonance template; (b) design matrix with age and education as confounding covariates and the classical SPM glass-brain presentation; (c) results table. Twenty-five NCs (age 61.3 ± 5.2 years, education 11.2 ± 4.1 years) ($Group_1$) and 25 ADs (age 72.2 ± 9.4 years, education 6.7 ± 3.2 years) ($Group_2$) were compared using SPM. In (a), the parietal and temporal regions are hypometabolic bilaterally. Metabolic involvement of the frontal lobe is shown on the left hemisphere. In the medial surface of the brain, severe hypometabolism in the posterior cingulate cortex is bilaterally shown. Lower left panel: In (c), the characteristics of each brain region are reported. Ke reports the extension (in voxels) of the regions. The column 'coordinates' reports (in millimeters) the peaks in the region, localized with *x* (left to right side, respectively, negative to positive values), *y* (back to front, respectively, negative to positive values), and *z* (above to below the intercommisural plane, respectively, negative to positive values) in Talairach space. The third and fourth columns are the names of the functional areas and the Broadman areas (BAs).

be used to diagnose and to differentiate AD from other dementias. In addition to differential diagnosis, the magnitude and extent of the CMRglc reductions on FDG-PET predict with high accuracy clinical progression and correlate with disease severity.

Despite the well-known early pathological involvement of the MTL in AD, only recently did clinical FDG-PET studies demonstrate CMRglc reductions in the MTL as an additional specific and consistent sign of early and advanced AD. This finding is in agreement with the finding of MTL synapse loss at a very early stage of AD. In contrast to the MTL, the PTC is affected by pathology in a later phase of AD, although PTC CMRglc decline is a relatively early event which does not seem to strictly correlate with local pathology. It has been suggested that PTC CMRglc reductions may be a distant effect of deafferentation of this region from the MTL, due to the disruption of neurons connecting the MTL to the PTC. Lesional studies in experimental animals demonstrated that the interruption of HIP and EC input/output connections can produce CMRglc decreases in the PTC. Moreover, a functional (metabolic) connectivity has been shown between the EC and PCC with FDG-PET in normal subjects and AD patients. In this model, damage to the MTL is a likely source of remote neocortical CMRglc impairment, or a sort of metabolic diaschisis due to the interruption of MTL inputs to the associative cortex. Later on, axonal transport dysfunction could participate in further decreasing cortical metabolic activity in the PTC and induce pathological changes such as synaptic and neuronal loss.

Brain Glucose Metabolic Alterations: Operating Characteristics in AD and Mild Cognitive Impairment

The clinical diagnosis of AD may be difficult to make at the early stages of disease when symptoms are not fully expressed and are hard to differentiate from physiological age-related declines. There is a great need for an accurate and early AD diagnosis because several clinical entities exist that share similar symptoms with AD but that, unlike AD, are potentially treatable (e.g., dementia syndrome of depression) or have a different response to specific treatments (e.g., frontotemporal dementia, FTD, and anticholinesterase inhibitors). The increasing availability of medications for symptomatic treatment of mild to moderate AD, which may be most effective in the early phases of disease, and the potential for protective/reparative treatments for preclinical AD phases, namely MCI, represent an additional reasons for developing early image diagnostic tools.

Several studies have put forth FDG-PET as a candidate tool for the preclinical detection of dementia. The largest published study with pathologically verified cases showed that bilateral PTC hypometabolism on FDG-PET has a sensitivity of 94% and a specificity of 73% in distinguishing AD from normal aging and other neurodegenerative diseases (Figures 6–8). Therefore, the presence of PTC CMRglc alterations appears to be relatively specific to AD. In the same study, the presence of PTC hypometabolism was shown to have prognostic value in predicting the progressive course of dementia with 93% sensitivity and 76% specificity. In 2004, these results prompted US Medicare to approve the reimbursement of FDG-PET scans performed in patients with documented cognitive decline of at least 6 months and a recently established diagnosis of dementia that meets diagnostic criteria for both AD and FTD.

For clinical and strategic reasons, it is also important to develop early diagnostic procedures to forecast on an individual basis who among MCI subjects will eventually decline to AD. Recently, FDG-PET capability to predict a decline from MCI to AD has been tested by several research groups. A clear sign predictive of decline has not been found in the general population of MCI patients. By restricting the evaluation to amnestic MCI patients, who show higher rates of decline to AD (up to 50% in 3 years compared to other MCI subtypes), it appeared clear that CMRglc reductions in the PCC and PTC are present in the future decliners at least 1–3 years prior to the clinical diagnosis of AD, with prediction accuracies between 75% and 90%.

FDG-PET can attain diagnostic power in MCI when, in addition to the typical cortical pattern already described, it is used to also evaluate the MTL metabolism, which is more visible with the use of oblique axial projections (Figures 9 and 10). When the severity of hypometabolism of the PTC and PCC is associated with that of the MTL, FDG-PET is able to differentiate MCI patients from normal subjects with a 85% sensitivity and 71% specificity and from AD patients with a 100% sensitivity and 77% specificity.

FDG-PET: Work in Progress

The most common approach to the diagnosis of AD with FDG-PET is by visual inspection, but visual analysis requires expert examiners. To obviate the need for expert readers, several automated diagnostic tools have been developed. They are based on the availability of a normal FDG-PET database and of VBA software. Thereafter, either the statistical parametric images are reviewed by the reader or the algorithm itself expresses a probabilistic attribution of the subject to a class of disease (AD, AD Lewy body variant, or FTD). The diagnostic performances of

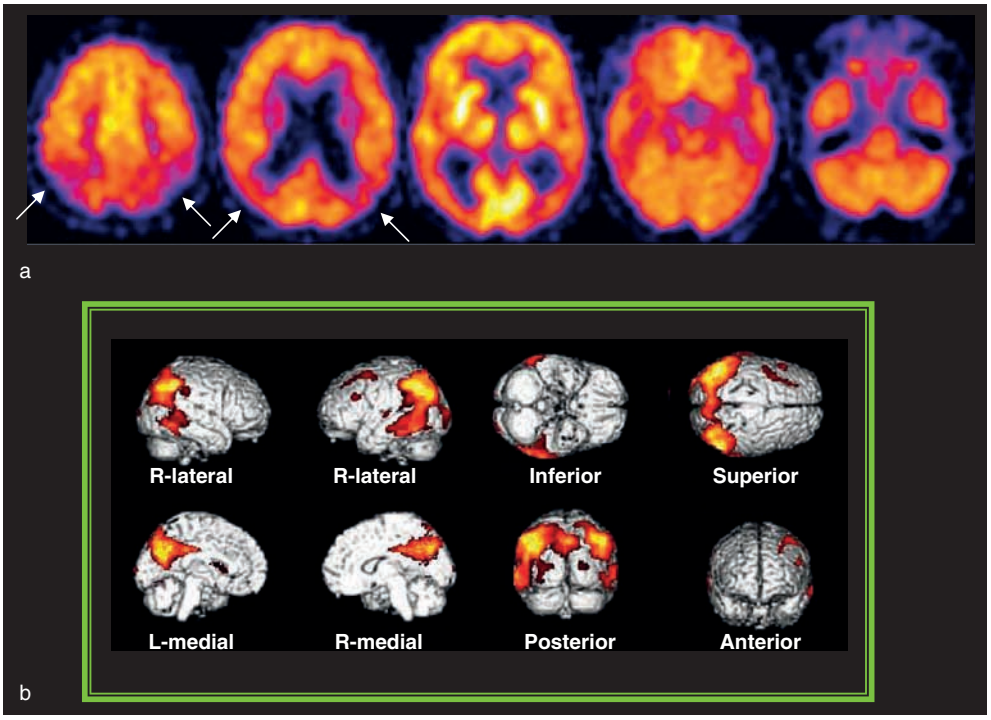

Figure 6 FDG-PET brain pattern in a patient affected by Alzheimer's disease (AD): (a) axial FDG-PET images; (b) automated analysis in which the classic SPM display shows the hypometabolic regions in comparison with a database of normal controls. The patient is a woman 63 years old, with AD onset at 57 years old, MMSE 17. The images in (a) show bilateral parieto-temporal hypometabolism (white arrows). Metabolism is relatively high in the occipital cortex and in the basal ganglia. The display in (b) confirms the presence of symmetrical parieto-temporal and posterior cingulate cortex hypometabolism. FDG, [^{18}F]-fluoro-2-deoxiglucose; L, left; MMSE, Mini-Mental State Examination; PET, positron emission tomography; R, right; SPM, Statistical Parametrical Mapping.

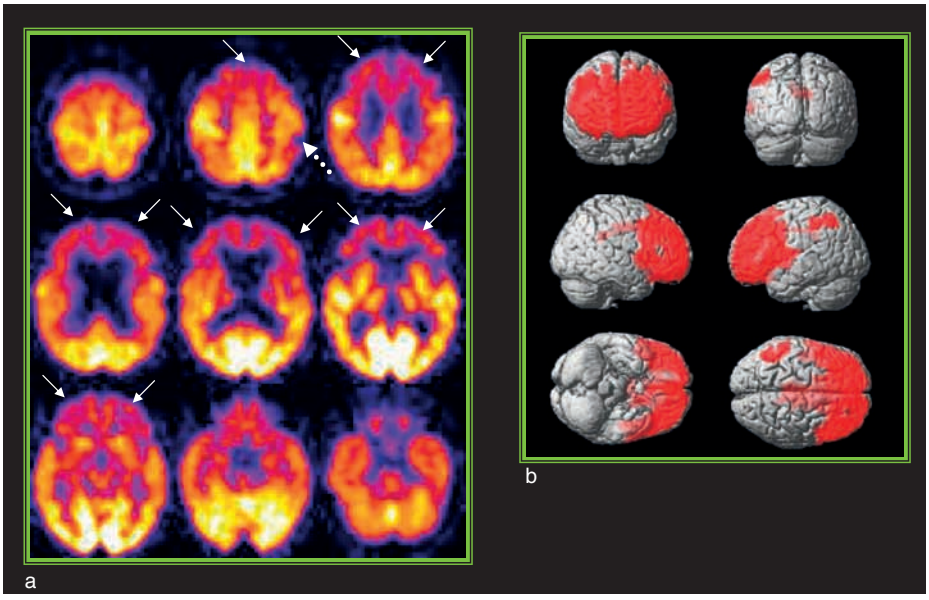

Figure 7 FDG-PET brain pattern in a patient with frontal dementia: (a) axial FDG-PET images; (b) automated analysis in which the classic SPM display shows the hypometabolic regions in comparison with a database of normal controls. The images in (a) show bilateral frontal hypometabolism (white plain arrows). The display in (b) confirms the presence of symmetrical frontal hypometabolism. Parietal hypometabolism on the left side might be due to sulcal enlargement (dotted arrow). FDG, [^{18}F]-fluoro-2-deoxiglucose; PET, positron emission tomography; SPM, Statistical Parametrical Mapping.

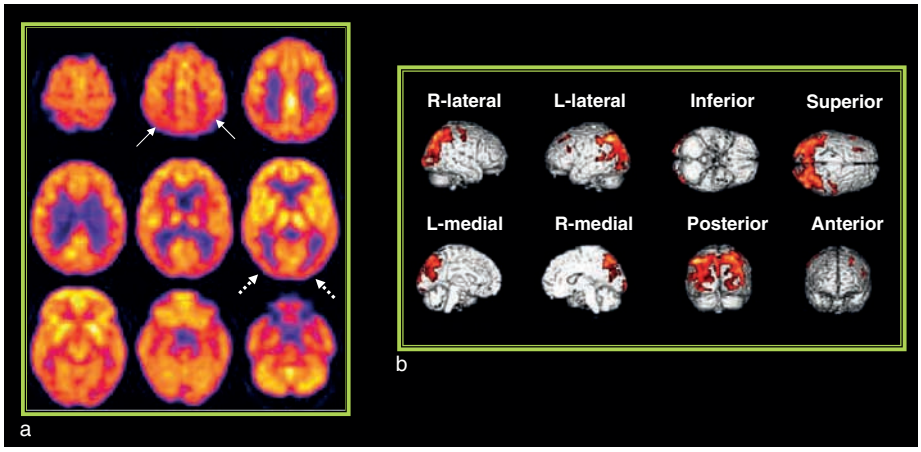

Figure 8 FDG-PET brain pattern in a patient with AD Lewy body variant. (a) axial FDG-PET images; (b) automated analysis in which the classic SPM display shows the hypometabolic regions in comparison with a database of normal controls. The images in (a) show bilateral frontal and parietal hypometabolism (white arrows). In addition, occipital hypometabolism is shown (white dotted arrows). The display in (b) confirms the presence of symmetrical parieto-temporal hypometabolism and the involvement of occipital cortex. FDG, [^{18}F]-fluoro-2-deoxiglucose; PET, positron emission tomography; SPM, Statistical Parametrical Mapping.

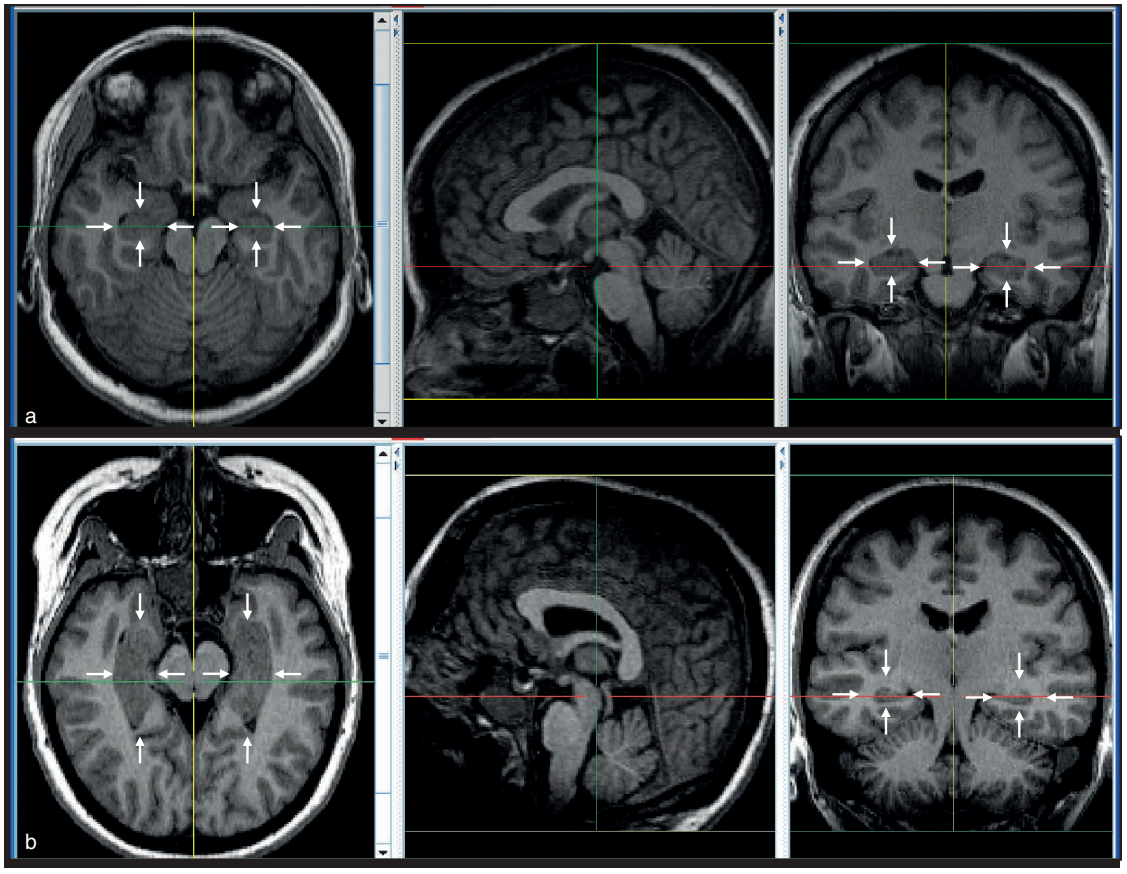

Figure 9 Hippocampal visualization in which MRI images are used to show the different orientations applied at brain volume for examining cortical metabolism and hippocampal metabolism: (a) orientation commonly used with functional images; (b) orientation used for visualizing hippocampal bodies. In (a), the horizontal plane is oriented so as to include anterior and posterior commissurare (intercommisural plane). This is called bicommisural or fronto-occipital orientation. With this orientation, hippocampal bodies, surrounded by white arrows, are cut out of their major axis plane. In (b), the sagittal image (middle) clearly shows the different horizontal plane. The hippocampal bodies are cut in the plane (white arrows). Images are prepared with MIPAV. MRI, magnetic resonance imaging.

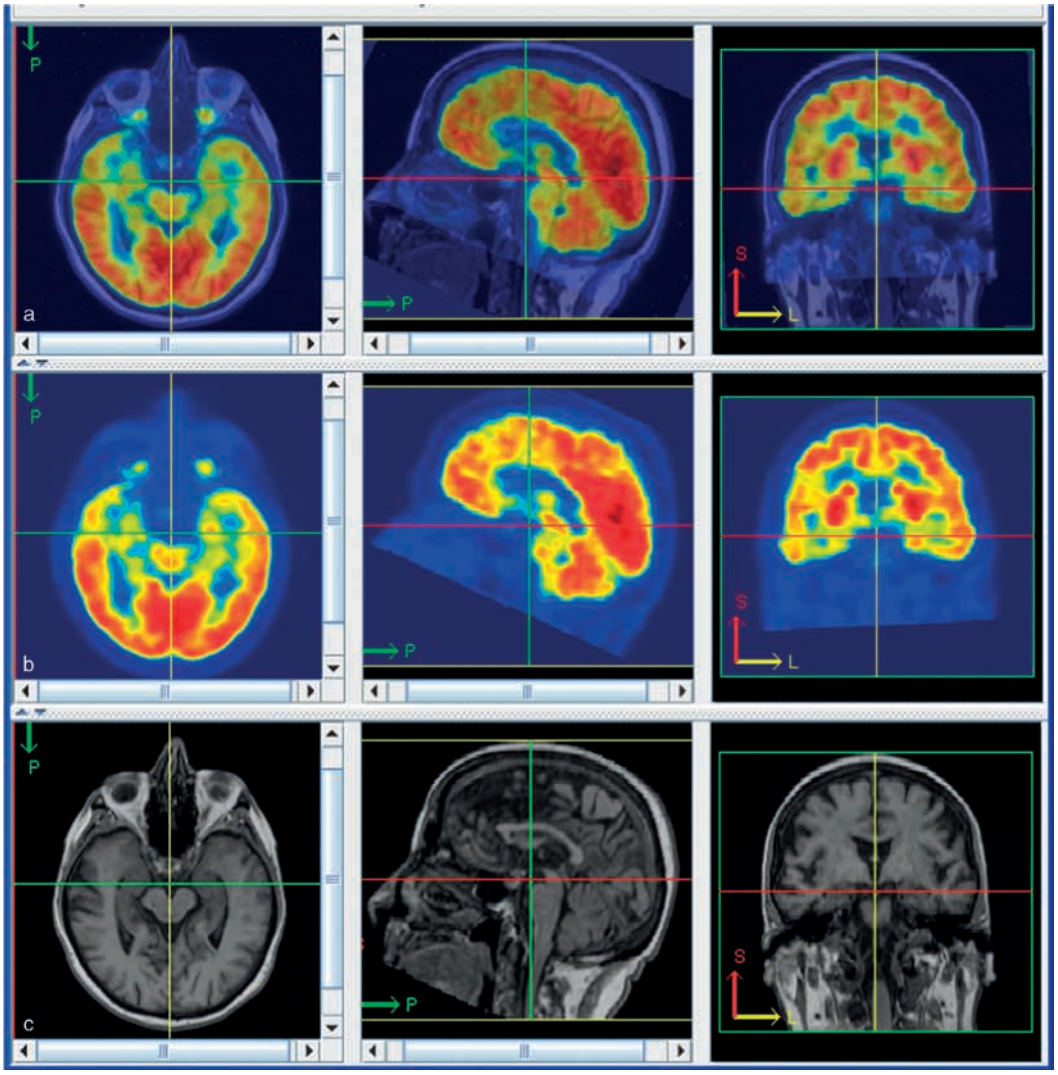

Figure 10 Co-registered MRI FDG-PET images of a normal subject, 30 years of age: (a) MRI and FDG-PET fused images; (b) FDG-PET images; (c) MRI images. Hippocampal metabolism is usually lower than cortical metabolism. Images are prepared with MIPAV. FDG, [^{18}F]-fluoro-2-deoxiglucose; L, lateral; MRI, magnetic resonance imaging; P, parietal; PET, positron emission tomography.

the automatic diagnostic systems are comparable with that of visual inspection by experts. We caution the reader that the use of automatic diagnostic systems for AD has not been approved by regulatory institutions and that clinical application is not recommended. Examples of the application of VBA to single-case studies are given in Figures 6–8.

FDG-PET has also been proposed as a surrogate endpoint in AD clinical trials. A surrogate endpoint is a biological marker, laboratory measurement, or physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions, or survives. Using FDG-PET as surrogate endpoint in clinical trials in AD could substantially reduce the sample size necessary to detect a treatment effect, the trial duration, and the expense. Other analyses have shown that the use of FDG-PET as surrogate endpoint after a treatment to slow or halt the progression of AD could allow a tenfold reduction in the number of subjects needed to enroll in a clinical trial. FDG-PET is not currently recommended as a surrogate endpoint in phase III drug trials in AD because it has not been demonstrated that halting or hindering CMRglc reductions corresponds to improving the patient's health.

Genetic Factors in Relation with Brain Metabolism Changes in Aging, Mild Cognitive Impairment, and AD

Correlative studies of CMRglc changes and genetic factors have been performed in normal individuals and in AD and MCI patients. The interest of these studies is twofold: to discover in genetically predisposed

subjects the earliest metabolic signs of AD and to better investigate the cellular and molecular factors of regional metabolic reduction.

Few cross-sectional studies with FDG-PET have been done in subjects carrying autosomal dominant mutations which determine an early-onset familial AD manifestation (EOFAD). These genes are *APP, PS1*, and *PS2*, and their proteins are involved in the production and metabolism of amyloid precursor protein. The importance of these studies is that they may be performed long before the emergence of cognitive clinical symptoms and may help to better define the metabolic continuum among normal brain, aging, senescence, MCI, and AD. In these studies, the same regions affected by CMRglc decrease in MCI and AD (i.e., the PTC and HIP) were already hypometabolic more than a decade before subject's expected time of overt clinical dementia (Figure 11). Despite possible biological differences with nonautosomal AD, the finding of this specific endophenotype in autosomal EOFAD subjects, although cognitively normal, strengthens the importance of the CMRglc impairment in the PTC and HIP as a research and diagnostic marker of the evolution from normal brain function to dementia; also it may be proof of a direct relationship between the amyloid metabolism and CMRglc decrease, an issue currently still largely debated.

FDG-PET changes in relation with the major genetic risk factor of AD, the ε4 allele of apolipoprotein E (ApoE) gene, have been investigated in AD patients at the onset of cognitive symptoms. The AD-specific hypometabolic pattern does not differ among the different ApoE genotype groups (with 0, 1, or 2 ε4 alleles). Anyway, when researchers compared the AD genetically different groups directly, hypometabolic differences were evident within the brain areas adjacent to those typically hypometabolic in AD, in a ε4-related fashion, suggesting that the metabolic damage of AD might be exacerbated by the dose of the ε4 allele. This effect of ApoE ε4 on CMRglc is consistent with the less efficient protective role of ApoE ε4 protein toward synapses in Aβ mouse models, as well as with the possibility that the ApoE ε4 protein directly damages the synapses, the main factor of FDG uptake in the brain. Moreover, the dose-dependent relationship between the number of copies of ε4 and the severity (extension) of CMRglc reduction in PTC is consistent with the known relationship of the ε4 dose and the age of onset of AD, such that ε4/ε4 subjects have an earlier onset than do heterozygous ε4 subjects.

The growing interest in the association of CMRglc and genetic factors in MCI patients derives from the compelling need to select the patients for protective/reparative interventions. Indeed, a genetic predisposition due to ApoE ε4 allele by itself is a predictor of conversion from MCI to AD with a good sensitivity (75%) but poor specificity (50%). When the genetic predisposition is considered together with FDG-PET hypometabolism in specific areas (the ACC and inferior FC), the capacity to predict conversion attains a sensitivity of 75% and a specificity of 89%.

An interesting issue is the study of brain metabolism in normal aging in association with the ApoE ε4 allele in order to select as early as possible the at-risk subjects for treatment and for protection against AD. In middle-age and older healthy subjects with normal memory performance, a single copy of the ApoE ε4 allele was associated with lowered PTC and PCC

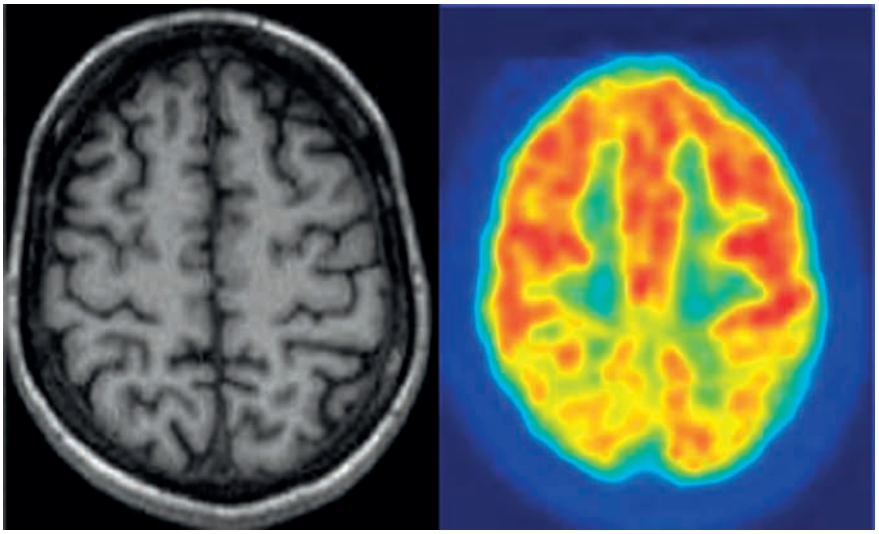

Figure 11 Metabolic impairment in a preclinical AD genetic case. Metabolic reduction in temporoparietal cortex in a subject 35 years old with a genetic mutation PS1 Leu392Val. At the time of this FDG-PET study (right) the subject was cognitively normal (MMSE 30). MRI-coregistered image (left) does not show any atrophy in the corresponding regions.

metabolism, which predicted the degree of cognitive decline after 2 years of longitudinal follow-up. This does not occur in the absence of the ApoE ε4 risk factor, representing possible proof that PTC hypometabolism might be a consequence of a genetic impairment of brain cells and, thus, might be causative of (instead deriving from) the cognitive impairment.

Although the ApoE ε4 allele by itself may account for a large fraction of genetic risk of developing AD, the role of several other susceptibility genes is suspected. Therefore, studies in AD patients have focused on LOFAD patients in which these unknown AD susceptibility genes may act in concert with epigenetic events to determine AD. In these patients (approximately one-third of all late-onset AD), there is a reduction of CMRglc in the PCC, HIP complex, and occipital regions compared with LOSAD patients, balanced for the presence of the ApoE ε4 allele. This finding suggests that these unknown susceptibility genetic factors may participate in the protective/reparative cellular functions against OS because they decrease CMRglc in brain regions in which a reduced capacity to compensate for stress and damages at the cellular level is known to be present.

See also: Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: MRI Studies; Axonal Transport and Alzheimer's Disease; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Genomics of Brain Aging: Apolipoprotein E; Glial Glutamate Transporters: Electrophysiology.

Further Reading

Mosconi L, Brys M, Glodzik-Sobanska L, et al. (2007) Early detection of Alzheimer's disease using neuroimaging. *Experimental Gerontology* 42: 129–138.

Reiman EM (2007) Linking brain imaging and genomics in the study of Alzheimer's disease and aging. *Annals of the New York Academy of Sciences* 1097: 94–113.

Silverman DH, Small GW, Chang CY, et al. (2001) Positron emission tomography in evaluation of dementia: Regional brain metabolism and long-term outcome. *Journal of the American Medical Association* 286: 2120–2127.

Relevant Websites

http://www.loni.ucla.edu – Alzheimer's Disease Neuroimaging Initiative (ADNI) research.

http://www.fda.gov – Food and Drug Administration, Peripheral and Central Nervous System Drugs Advisory Committee.

http://www.cms.hhs.gov – Medicare expansion of the coverage of FDG-PET for the diagnosis of Alzheimer's disease.

http://mipav.cit.nih.gov – MIPAV (medical image processing, analysis, and visualization) software.

http://imaging.mrc-cbu.cam.ac.uk – Montreal Neurological Institute.

http://www.uni-koeln.de – Network for Efficiency and Standardisation of Dementia Diagnosis (NEST-DD).

www.fil.ion.ucl.ac.uk – Statistical Parametrical Mapping (SPM) website.

http://128.95.65.28 – University of Washington NEUROSTAT website.

Brain Injury: Functional Recovery After

D G Stein, Emory University School of Medicine, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Conceptual Issues

Functional recovery after brain injury is generally thought of as the behavioral concept that cognitive, sensory, or motor impairments gradually diminish or are eliminated over time. This concept implies that, after a brain injury or stroke, acute or chronic behavioral and motor deficits disappear spontaneously or are reduced by pharmacological, physiological, surgical, or behavioral rehabilitation strategies. Functional recovery can also be defined by changes at the molecular level of the neuron. For instance, after an injury, neural transmission or release of transmitters at the synapse may be disrupted and 'functional recovery' can be defined as a return of neurotransmission to preinjury or even higher than preinjury levels. Indeed, many studies of plasticity and recovery of cell membrane or nuclear functions, such as the expression or inhibition of genes which control cell metabolism and growth, never examine whether the behavior of the organism is directly affected – in either an adaptive or maladaptive way. Dramatic changes in central nervous system organization and metabolic activity could be disruptive to, or even prevent, further recovery of function, so it is critical to have detailed behavioral outcome measures to correlate with physiological alterations in brain activity associated with injury and subsequent repair. For example, very high levels of glutamate, an excitatory neurotransmitter in normal circumstances, can overstimulate vulnerable neurons to the point of killing them off at sites remote from the injury, disrupting complex behaviors. Furthermore, psychological 'functions' that are initially lost and then recovered are not always clearly defined, and as a result, both function and recovery can mean different things to different people. Thus, in laboratory rats with frontal cortex lesions, cognitive functions involved in learning are said to be preserved if the animals can run through a maze to get to a goal box. No attention is paid to the possibility that there could be abnormalities in their gait or shifts in the hierarchy of sensory cues they use to remember where they are going in the maze. By ignoring the question of how the animals get to the goal box, recovery may be assessed as complete, although the underlying physiological processes mediating the behavior may be very different from those seen in intact animals. Thus, 'recovery' can be defined in terms of attaining a goal, regardless of how that goal is achieved.

In the clinic, such an ends-based definition of functional recovery is more difficult to accept. Many in the field of rehabilitation still hold to the idea that 'true' recovery of function does not exist. When brain-damaged patients succeed in activities of daily living by substituting new, different behavioral strategies to accomplish goals, as they often do, this is assumed to reflect a process of 'compensation.' The goal for rehabilitation, in this view, is to help patients learn to live more effectively with their impairments – to get around them. Although the brain-damaged patient may accomplish the same goal as an intact person, the behavioral strategies used may be very different. Without measurements of specific physiological correlates, it is circular to argue that if no deficits are observed, it is because the patient has learned to 'compensate,' the tests are insensitive, or the wrong tests have been used, rather than because any true recovery of function has occurred.

Although 'compensation' is a phenomenological construct that can be evaluated only by identifying and measuring behavioral changes, underlying changes in cerebral physiology and structure must be present to mediate the behavior. Many studies show that rehabilitation training and general learning experiences can directly modify cortical maps in dramatic and long-lasting ways. These new maps, which are basically aggregates of neurons, can respond to sensory inputs that did not activate them before the injury-induced reorganization. Extensive reorganization leading to behavioral compensation is inconsistent with the idea that specific areas of the brain are hardwired to 'control' or 'mediate' behaviors – and thus that the destruction or perturbation of a center in which a specific function is localized must lead to a permanent loss (or permanent impairment) of that function. The notion of substitution of function, or compensation, after brain damage is consistent with the idea that intact parts of the brain are structurally modifiable to take over the functions of tissue lost after injury or disease. From this point of view, remaining healthy tissue is said to modulate its physiological activities in response to damage to a specific brain region.

In restitution or sparing of function, initial behavior is intact or returns to preinjury performance. Such recovery is thought to be due to the equipotentiality or vicarious capacity of cerebral areas. Equipotentiality is the capacity of anatomically distinct, sometimes

distant, areas of the brain to mediate disparate functions. Such plasticity is thought to be more likely when the brain is developing and not yet specialized for particular activities, such as speech or patterned movement. Recovery of motor and sensory functions and sparing of language and cognition is sometimes complete when injury occurs at certain critical periods in early development.

Even the mature central nervous system has redundant or 'backup' subsystems that participate in this takeover of function. The term for this is neural vicariation. 'Parallel' and perhaps even widely (spatially) distributed 'systems' can sometimes mediate the behaviors of concern. However, the mechanisms by which area 'a' could 'take over the function' of damaged area 'b' and still perform function 'a' are usually not well described.

Injury-Induced Reorganization of Cerebral Maps

Until the development of sophisticated imaging and electrophysiological techniques, it was not possible to determine whether such equipotential, vicarious, or pluripotent systems did or did not exist, so the concept of neuronal plasticity in adults was largely dismissed until near the end of the twentieth century. Clinical researchers now use functional magnetic resonance imaging (fMRI), positron emission tomography, scanning and microelectrode stimulation, and recording from multiple brain sites to show that functional compensation is the result of extensive reorganization of activity in the damaged brain. For instance, after an infarct in the striatum of human patients, a number of subcortical areas both ipsilateral and contralateral to the injury site become activated as behavioral recovery progresses. Such activation, measured by increases in cerebral blood flow and glucose metabolism, is not observed when normal volunteers perform the same motor tasks. Thus, if 'tricks' are being used to compensate for lost functions, they may be mediated by extensive shifts in the neuronal and metabolic activity of cerebral structures that are not activated or thought to be generally involved in the behaviors (or not in the same way) in normal circumstances. This type of reorganization can occur rapidly after an injury or take weeks, months, or years, making prognoses for recovery complex and difficult.

Microelectrode recording techniques have demonstrated that in primates, certain brain areas such as the sensory motor cortex are capable of very rapid electrophysiological reorganization in response to injury. Within very short periods of time after lesions of the afferent pathways or even after cortical damage, tissue adjacent to the affected areas reorganizes or expands its receptive fields to capture and mediate at least some of the original physiological activities that were destroyed by the injury. For example, if the visual (occipital) cortex is deprived of its inputs from the optic nerves, that same cortex can now be stimulated by auditory input – the 'visual' cells are 'taken over' by the auditory nerve fibers. The auditory cortex could be said to expand into the territory previously occupied and controlled by visual inputs. Likewise, if the auditory cortex is destroyed or the auditory nerve is cut, the auditory cortex will be 'captured' by visual inputs that have expanded into the deafferented zone. However, if the visual cortex is now activated by auditory input, it may be a clear example of neuroplasticity, but is it 'functional recovery'? This demonstrates again why functional assessment of recovery is so important in a clinical and practical context. Rapid reorganization tends to be more robust during early development. This may be one reason why young subjects with brain damage recover better than older counterparts with the same injury.

Imaging and electrophysiological techniques have also been used to support the historical idea of diaschisis (injury-induced functional inhibition) proposed by Constantin von Monakow at the beginning of the twentieth century. In diaschisis, brain injury triggers a form of long-lasting depression of neural activity. Recovery of function was thought to occur as diaschisis dissipated over time, whereas permanent deficits, due especially to subcortical strokes, were said to be the result of permanent diaschisis. Although the logic appears circular, current research shows that injury to the brain does produce chronic suppression of activity associated with behavioral impairments, often in the area of the brain contralateral to the injury, showing that the process of injury involves many more structures than just the directly traumatized area.

At the very least, the notion that the 'rewiring' of neuronal pathways can be achieved by drugs or environmental stimulation needs to be approached carefully because the wiring diagram for each individual is heavily influenced by early experience and training. Standardized rehabilitation therapies might result in the formation of maladaptive connections and the prevention of functional recovery. For instance, if the current paradigm of structure–function relationships proves valid, it is difficult to imagine how the capture of auditory input by visual fibers would automatically lead to an adaptive outcome.

In conjunction with the electrophysiological experiments in primates, computer-assisted rapid fMRI studies in humans have shown that multiple systems throughout the brain become activated during selected

cognitive tasks, such as naming objects or performing mathematical calculations. Many neuroscientists now accept the view that widely distributed neural 'modules' and parallel processing are involved in even simple cognitive or sensory processing; however, there is much less agreement regarding the number of modules (brain areas) that are required to mediate even a simple behavior such as reaching for a piece of food. A corollary of this assumption is that focalized damage would cause widespread changes both serially and in parallel throughout the brain. Therefore, recovery of behavioral functions may depend on equally widespread changes in synaptic connectivity and their nuclear and metabolic consequences.

Therapies for Brain Injury

In the search for new therapies to treat traumatic brain damage, it has become clear that the injury itself is not a monolithic or isolated event. Brain injury is now seen as a very complex cascade of biochemical and morphological processes of varying duration, each of which may contribute to neuronal death and/or repair and regeneration. Until recently, head trauma research focused on what happens to neurons and their transmitters. Now it is becoming increasingly clear that the repair of the vascular system and blood–brain barrier and the posttraumatic activities of different types of glial cells, macrophages, T cells, and other inflammatory agents all contribute significantly to the prognosis for recovery.

During the past 40 years, clinical trials have failed to find a successful treatment for traumatic brain injury (TBI). For stroke victims, only one agent, tissue plasminogen activator, is available, and it is used infrequently (in only 3–5% of patients) because of its risks. Most current drug interventions for brain damage are categorized into two classes: (1) those that can be given early in the injury cascade to inhibit inflammation and cytotoxicity in vulnerable neural cells and (2) those that are given later to agonize or antagonize the synthesis, release, and reuptake of neurotransmitters and neuromodulators, which can produce a state of overexcitation and oxidative stress leading to loss of nervous tissue. Such agonist drugs which increase neurotransmitter activity or block inhibitory neurotransmitters to overcome diaschisis have been used to promote functional recovery after stroke in both animal and human subjects. However, clinical trials with these agents have had limited success. Another category of treatment, behavioral therapy, such as forcing patients to use an affected limb by restricting the use of the good limb (the forced-use paradigm), or treatment with stimulants (e.g., amphetamine) may reduce or eliminate the diaschisis caused by the injury. Do these therapies work because other brain areas take over the function of the area of stroke? Current evidence can be cautiously interpreted to suggest that this may be the case, but the issue is still the subject of much research and debate.

Neuroprotective Agents

There are three main types of therapies to increase functional recovery from brain injury. First is the use of neuroprotective compounds to reduce the initial inflammatory cascade that can destroy injured neurons. This phase is followed by necrosis of tissue in the area of the injury and then by programmed cell death (apoptosis) that can continue for weeks or months after the initial injury. Many of these changes are not limited to the injury site but, rather, take place throughout the brain. The sometimes subtle and persistent changes in personality, motivation, and daily performance seen in brain-injured patients may be due to the widespread changes in brain physiology that are not always detected with current noninvasive assay techniques.

The neuroprotective agents that reduce cytotoxicity tend to absorb damaging molecules such as free radicals. Free radicals react with many sensitive organelles in nervous tissue, including cellular membranes, DNA, and proteins, changing their conformation and affecting their function. Unfortunately, brain injury clinical trials using free radical scavengers have not had successful outcomes. Other therapies which block the hyperexcitation of the receptors for glutamate have been studied extensively in the laboratory. As with free radical scavengers, many of these compounds have also proven ineffective in improving clinical outcome after TBI.

One additional category of agents that does show promise for reducing the early stages of vascular and neuronal tissue destruction are the neurosteroids progesterone and estrogen. These steroids appear to produce beneficial effects through actions at several levels of the injury cascade. Both can reduce lipid peroxidation, excitotoxicity, and inflammatory immune reactions. Progestins and estrogens have the advantage that they are natural substances that can be injected systemically or given orally beginning immediately after the onset of injury. In laboratory investigations, progesterone was effective in reducing cerebral inflammatory factors, edema, and secondary neuronal degeneration caused by necrosis and apoptosis. Thus, as cerebral edema is reduced, neuronal sparing can also be observed in those areas of the brain that have reciprocal connections to the zone of injury. The onset of cerebral edema also appears to be related to the

production of free radical lipid peroxidation. Postinjury treatment in male rats given progesterone can markedly reduce the formation of 8-isoprostane (a prostaglandin-like substance used as a marker for peroxidation reactions) in the first 24 h after contusion of the medial frontal cortex. This change is highly correlated with behavioral recovery, reductions in the size of the lesion, and enhanced neuronal sparing, which endure long after the treatments have terminated. These data suggest that neuronal and behavioral recovery may be due to the scavenging of free radicals and reduction of inflammation in the injury zone. In one study, researchers found that males with severe brain injury have more than twice the levels of lipid peroxidation reactions to injury compared to females. This may be one reason why brain-injured females tend to recover better than males with the same extent of damage.

Although many still regard progesterone only as a female sex hormone, its use in males in the acute period after brain injury produces the same beneficial morphological and behavioral consequences as in females. One clinical trial found that postinjury administration of progesterone could decrease mortality in severely brain-injured patients by more than 50% and significantly reduce functional deficits in moderately brain-injured patients.

Drugs that can modulate γ-aminobutyric acid (GABA)ergic synapses have been implicated in the recovery process. GABA is thought to be primarily inhibitory in the cortex, and substances that increase GABA synthesis appear to block recovery from cerebral injury, especially in the somatosensory system, when administered chronically. Interestingly, progesterone stimulates the $GABA_A$ receptor to reduce excitatory mechanisms, and in this role it could prevent postinjury epilepsy and excitotoxicity during the early stages after brain injury.

The problem with managing levels of neurotransmitters as a means of promoting functional recovery is that pharmacological manipulations are extremely sensitive to the elapse of time after the injury. Therefore, increased levels of inhibition that may be needed to block excitotoxicity in the damaged area may in fact disrupt recovery if treatment is maintained throughout the course of rehabilitation.

Neural Regeneration

Another approach to treatment is chronic administration of pharmacological agents or the transplantation of various types of embryonic stem or progenitor cells directly into the brain tissue to enhance or stimulate regenerative neuronal and vascular mechanisms of repair after the injury cascade has subsided. Most of these techniques are still in the experimental stage and have not been successfully applied in large-scale clinical testing. As mentioned previously, neuroprotective agents act by protecting the brain from secondary injury in the immediate period after trauma. Research has started to focus on stimulating the regeneration of neural tissue. Such therapies consist of administration of trophic factors – agents that stimulate the repair, regeneration, elongation, and reconnection of damaged axons or dendrites – and transplantation of stem cells, which can migrate to areas of damage to replace lost neurons and glial cells. However, neuronal sprouting may not always be beneficial. For example, lesion-induced sprouting of new afferents into the hippocampus can lead to increased seizure activity, whereas some sprouting in the injured spinal cord can produce spasticity.

Until only a few years ago, glial cells or embryonic stem cells received scant attention as treatments for brain damage. Now, glia have been shown to secrete neurotrophic factors that can stimulate sprouting and regeneration, contribute to axonal repair by remyelinating damaged nerve fibers, and directly modify and contribute to neural transmission in response to brain injury. Neuroscientists have shown that progenitor cells lining the walls of the cerebral ventricles, and stem cells, can give rise to new neurons in the hippocampus, an area consistently implicated in memory processing.

Stem cells also morph into glia, which are now known to play a role in stimulating the growth of new blood vessels and capillaries – processes that are also critical to neural repair after injury. In developing new therapeutic agents to promote functional recovery, it would be advantageous to focus some research on the effects of drugs or stem cell lines on the enhancement of this glial and vascular response to injury. Such effort would be especially important if glia can be shown to contribute to neurogenesis or regeneration in the adult brain.

A class of proteins and peptides showing some promise in preclinical head injury research are neurotrophic factors produced by the brain during development and after injury. During the past decade, advances in molecular biology have made it possible to produce human trophic factors in genetically engineered bacteria. One problem with their use as therapeutic agents in human patients is that they do not appear to pass through the blood–brain barrier in sufficient quantity to be clinically effective. However, rats given intracerebral or intraventricular brain-derived trophic factor, nerve growth factor, and basic fibroblast growth factor have shown significant functional and morphological recovery after lesions of the nucleus basalis magnocellularis, the septal nucleus, the hippocampus, and the entorhinal cortex, among others. Despite good results with laboratory animals, high-quality clinical trial data with these proteins are

still lacking. Despite their potency, delivery of trophic proteins directly into the brain appears to be the limiting factor because systemic administration by injection seems to have unwanted side effects.

Another new area of research is the use of neural stem cells to repopulate the injured nervous system with tissue that has the capacity to develop into a variety of different organs, including brain cells. In one study, researchers showed that when cells from a mouse neural stem cell line were injected into cholinergic basal forebrain areas of mice with fimbria–fornix lesions, these implanted cells differentiated into a neuronal phenotype that expresses choline acetyltransferase and the p75 neurotrophic factor receptor. Other investigators have injected bone marrow stromal cells into the internal carotid artery after TBI in rats. In this study, the injected cells were able to enter the brain parenchyma at the site of injury and differentiate into both neuronal- and astroglial-like phenotypes. Rats treated with these stromal cells had improved neurological scores compared to their injured controls.

Behavioral Rehabilitation Therapies

Behavioral rehabilitation therapies combined with antidepressants, antispasmodics, anticonvulsants, or other agents are used to help brain-injured patients readapt to activities of daily living. One of the more successful rehabilitation approaches has been the 'forced-use paradigm' mentioned previously, in which stroke patients, even years after their injury, have been trained to withhold the use of their unaffected limbs while being required to manipulate aspects of their environment with their stroke-affected limb. Multicenter testing with this approach has demonstrated moderate to good recovery that appears to be sustained after the training has been terminated.

Investigators have also shown that reorganization and expansion of functional cortical maps can be induced by intensive training in normal animals. In brain-damaged subjects, specific kinds of sensory training may be beneficial as rehabilitation therapy because they directly enhance the pace and extent of reorganization and recovery of functional plasticity. Furthermore, there are apparently marked individual differences in normal individuals in the location and size of cortical receptive fields which vary in their degree of modifiability according to the subject's experiential history. An interesting example of experiential modification of the brain is found in London cabbies. Using MRI scanning, University College of London investigators found that compared to the general population, cabbies have a larger hippocampus (a part of the brain associated with long- and short-term memory storage and spatial learning). The 'extra' learning required to memorize the streets of London supposedly accounts for the additional hippocampal growth.

Conclusion

The injury process is a complex cascade of events that can be influenced by many variables, including sex, age, severity and momentum of injury, hormonal state, health history, and time of trauma. Other therapies may include treatments that do not directly affect neural tissue but instead promote recovery of function by working indirectly on the immune system to reduce inflammation or on the vasculature to increase blood flow and brain metabolism. Because brain injury in its various forms is a major public health problem, more attention is now being paid to research in this area. Meanwhile, older paradigms which held that the mature brain is incapable of inherent plasticity and repair are being replaced by a more optimistic vision of central nervous system plasticity, which will eventually translate into more effective and enduring treatments and therapies. Policy changes at the National Institutes of Health suggest that there will be more focus on 'translational' research to bring new drug therapies and better care to TBI victims, but only time will tell whether such policies prevail and prove effective.

See also: Animal Models of Stroke; Brain Damage: Functional Reorganization; Brain Injury: Magnetic Resonance Studies of Metabolic Aspects; Brain Trauma; Neuroprotection: Pharmacological Approaches; Stroke: Injury Mechanisms.

Further Reading

Fawcett JW, Rosser AE, and Dunnett SB (eds.) (2001) *Brain Damage, Brain Repair.* New York: Oxford University Press.

Matthew PM, Honey GD, and Bullmore ET (2006) Applications of fMRI in translational medicine and clinical practice. *Nature Reviews Neuroscience* 7: 732–744.

Mountz JM, Liu HG, and Deutsch G (2003) Neuroimaging in cerebrovascular disorders: Measurement of cerebral physiology after stroke and assessment of stroke recovery. *Seminars in Nuclear Medicine* 33: 56–76.

Nudo NJ (2003) Adaptive plasticity in motor cortex: Implications for rehabilitation after brain injury. *Journal of Rehabilitation Medicine* 41(supplement): 7–10.

Silver JM, McAllister TW, and Yudofsky SC (eds.) (2005) *Textbook of Traumatic Brain Injury.* Washington, DC: American Psychiatric Publishing.

Stein DG (2006) Concepts of CNS plasticity and their implications for understanding recovery after brain damage. In: Zasler ND, Katz DI, and Zafonte RD (eds.) *Brain Injury Medicine: Principles and Practice*, pp. 97–108. New York: Demos.

Wright DW, Kellermann AL, Hertzberg VS, et al. (2006) ProTECT: A randomized clinical trial of progesterone for acute traumatic brain injury. *Annals of Emergency Medicine* 49(4): 391–402.

Brain Injury: Magnetic Resonance Studies of Metabolic Aspects

R Vink, University of Adelaide, Adelaide, SA, Australia

© 2009 Elsevier Ltd. All rights reserved.

Brain injury may result in the development of irreversible tissue damage through both primary and secondary injury mechanisms. Primary injury is caused by the mechanical events associated with the initial injury to the brain, whereas secondary injury is the chemical and metabolic changes initiated by the original insult. These secondary changes then develop over minutes to days after the primary event to cause further brain injury. This gradual development of secondary injury with time provides a window of opportunity to attenuate, or even prevent, the secondary injury process and subsequently improve the outcome. However, the factors that contribute to secondary injury first need to be identified before appropriate 'antifactors' can be developed. One of the major difficulties associated with the identification of secondary injury factors has been the use of invasive physiological and biochemical techniques that by their nature destroy tissue. Moreover, such techniques are limited to single time points and do not permit the correlation of measurements of physiological and metabolic changes with the ultimate survival and neurological outcome following brain injury. One class of techniques that has gained increasing acceptance as a valuable tool in the study of brain injury is magnetic resonance (MR) spectroscopy (MRS) and magnetic resonance imaging (MRI). These MR techniques are noninvasive and permit the repeated monitoring of various metabolic and physiologic changes over time. Moreover, at an experimental level, a brain-injured animal can be allowed to recover and evaluated with regard to long-term neurological outcome and histopathology. Thus, each animal serves as its own control, and correlations can be made between early metabolic and physiological changes and chronic behavioral recovery or tissue damage.

Magnetic Resonance

The application of MR techniques to the study of biological systems was first performed in the 1950s, but it was not until the advent of surface coil techniques in 1980 that the technology was successfully applied to the study of organ metabolism *in vivo*. Since then, the investigation of brain metabolism by MR techniques has resulted in literally thousands of medical and scientific research studies. Although a number of nuclei can be evaluated with MR, most MR studies in brain injury have concentrated on either phosphorus or proton resonances. It is not the intention of this article to discuss the basic principles of MR, and interested readers can find more detailed information from some excellent online resources.

Phosphorus MRS

Phosphorus MRS allows the monitoring of metabolic events involving phosphorus metabolites. Because many of the critical energetic processes occurring in the cell are essentially phosphate transfer reactions, phosphorus MRS has grown to be the most widely used MRS technique in the study of metabolism. A typical phosphorus MRS spectrum obtained from the brain of a rat is shown in Figure 1. Several metabolites can be identified, including phosphomonoesters (PME), inorganic phosphate (Pi), phosphodiesters (PDE), phosphocreatine (PCr), and the three phosphates of adenosine triphosphate (ATP). The integrated area under each peak is directly proportional to the concentration of that metabolite, whereas the chemical shift (i.e., position) of the peak provides information about the ionic environment of that metabolite. For example, the Pi peak is commonly used to determine the intracellular pH, the chemical shift of the Pi peak being dependent on the dissociation state of HPO_4^{2-}. Similarly, intracellular free magnesium concentration (Mg^{2+}) can be determined on the basis of the chemical shift of βATP, which is dependent on the relative concentrations of Mg^{2+}-bound ATP and Mg^{2+}-free ATP. Thus, change in the position of the peaks provides valuable information about the intracellular ionic environment, whereas change in the intensity of the peaks provides an indication of the relative concentrations of the metabolites.

Proton MRS

Application of proton MRS to the study of the brain has been somewhat slower than that of phosphorus MRS because of difficulties inherent in this method. The relative concentration of the intracellular metabolites is in the millimolar range, whereas the concentration of water approaches 40 M, which is equivalent to 80 M in protons. Thus, to observe the less-concentrated metabolites, effective suppression of the dominant water peak is essential. Furthermore, the difference in chemical shifts between metabolites is far less than that observed in phosphorus

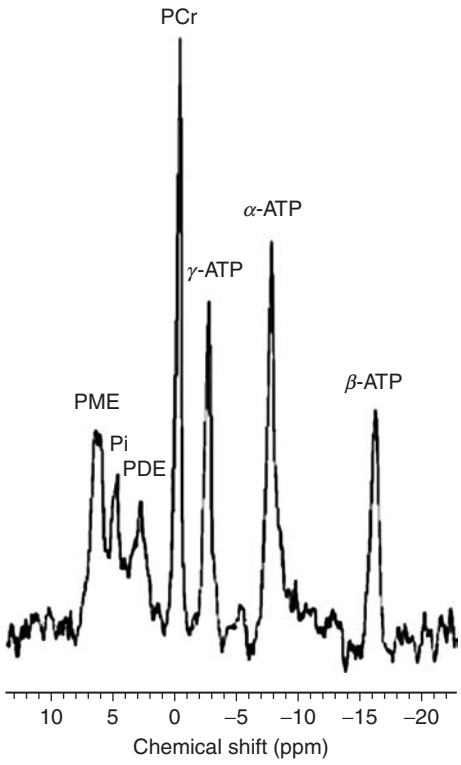

Figure 1 Typical phosphorus MRS spectrum obtained from the brain of an anesthetized rat. PME, phosphomonesters; Pi, inorganic phosphate; PDE, phosphodiesters; PCr, phosphocreatine; ATP, adenosine triphosphate (α, β, and γ peaks).

MRS. For example, the normal phosphorus spectrum extends over approximately 40 parts per million (ppm), whereas most observable proton resonances occur within a 10 ppm range. Thus, there is significant overlap between metabolites in a proton spectrum that makes identification of individual resonances somewhat difficult. Nevertheless, proton MR techniques have been developed that allow for the monitoring of certain metabolites, including lactate, *N*-acetyl aspartate (NAA), total creatine (creatine plus phosphocreatine), total choline, lipids, and amino acids. Furthermore, as higher field magnets are developed and applied to brain studies, the degree of overlap between individual resonances will be reduced providing greater chemical shift dispersion and improved detection sensitivity such that more metabolites will be visible.

Proton MRI

MRI differs from MRS in that total signal intensity is mapped against location, giving an image of the signal distribution. Most studies have mapped the proton intensity of water, whose concentration varies by 14% between tissues, thus giving excellent soft tissue contrast, particularly in brain (Figure 2). In addition to anatomical images of brain structure, the technology can also be used to obtain information on vascular anatomy, blood flow, edema development, water diffusion and perfusion, and glucose turnover. Also, metabolic information has been obtained by using an exciting MRI development termed functional imaging, which provides information on the relationships between brain structures and their function.

MRS in Brain Injury

Noninjured Brain

Numerous MRS studies have reported 'normal' values for various parameters in brain. From phosphorus studies, intracellular pH is approximately 7.1, and intracellular free Mg^{2+} concentration ranges from 0.3 to 0.6 depending on age, gender, and species. Although absolute concentration values of metabolites in brain have been reported, these values vary depending on the methods used. Therefore, it is usual in phosphorus MRS to report ratios that avoid the pitfalls of absolute quantitation. The ratios so far reported are somewhat species dependent; the PCr:Pi ratio ranges from 2.0 in humans to 4.0 in rats, and the PCr:ATP ratio ranges from 1.0 to 2.0, respectively. It should be noted that these are adult values, and ratios are lower for neonatal brain. ADP concentration has been calculated from the creatine kinase equilibrium reaction to be approximately 20–30 $\mu mol\, g^{-1}$ tissue wet weight. This calculation uses *in vitro* values for PCr and ATP and the MRS measured values for pH and free Mg^{2+}. Because this value for ADP concentration is significantly less than previous values determined using *in vitro* techniques, it is thought to represent the free cytosolic pool and not the total ADP pool.

In water-suppressed proton MR spectra of normal brain (Figure 3), NAA is the dominant metabolite, with a concentration of approximately 13 mM, although the exact value is dependent on age and, to a small degree, location (gray matter vs. white matter). Originally used as an internal standard, it is now known that the NAA concentration varies with pathophysiological state and cannot be used as a concentration standard. Indeed, it is widely accepted that NAA is a neuronal marker reflecting both neuronal integrity and neuronal metabolic activity. It is commonly expressed in terms of a ratio with other visible metabolites, including total choline (NAA:Cho) or total creatine (NAA:Cr), with the latter being the more frequently used. The normal value for NAA:Cr is approximately 1.35 and varies according to brain location and age. The degree of water suppression affects peak intensities of those metabolites closest to the water frequency, and this should be borne in

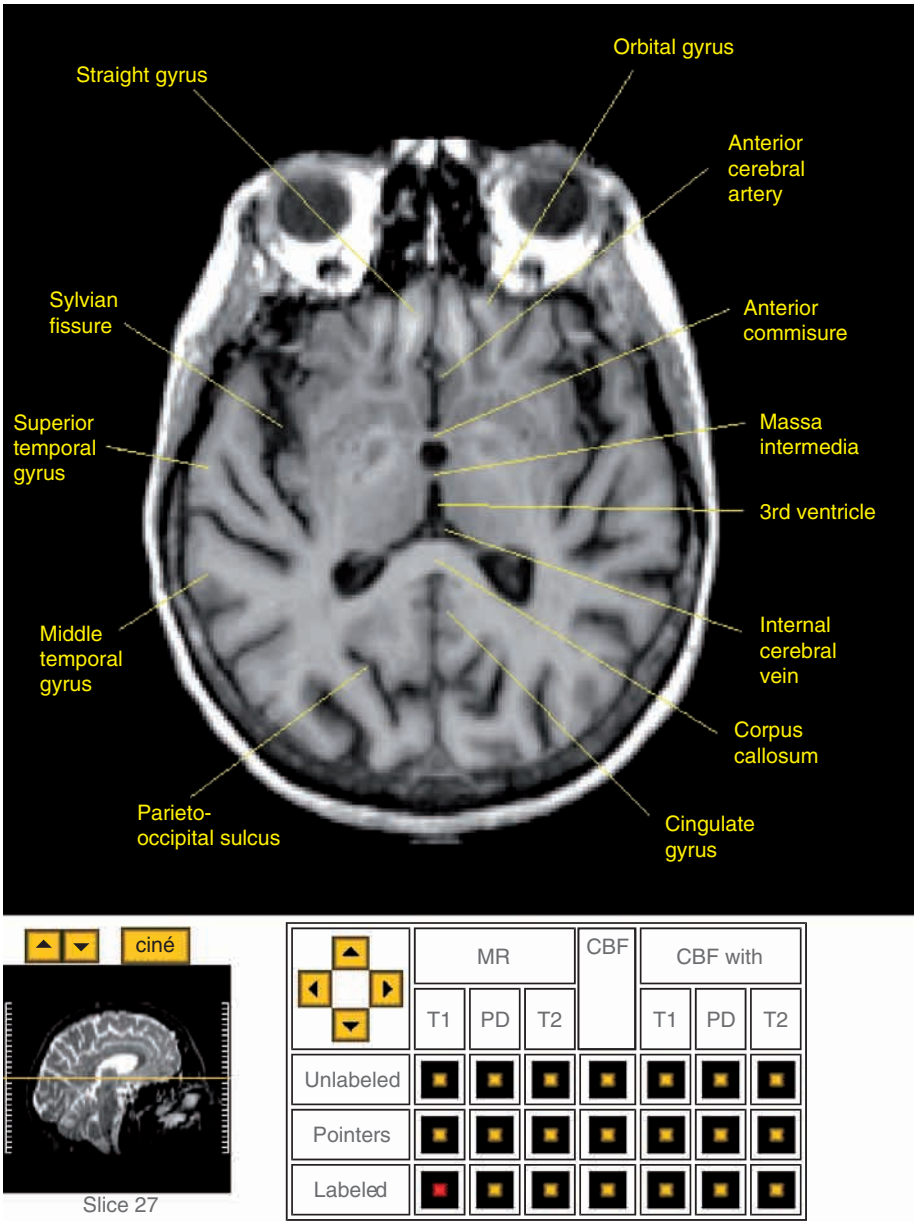

Figure 2 MRI image through the human brain.

mind when examining relative metabolite concentrations. Although very important after brain injury, lactate is normally not visible in normal brain because it is present at submillimolar concentrations.

Ischemia

Many of the early MRS studies of brain metabolism following central nervous system (CNS) injury evaluated changes caused by cerebral ischemia. Ischemia results in a loss of high-energy phosphates such as PCr and ATP and an increase in Pi. A profound decrease in intracellular pH from 7.1 to approximately 6.4 was invariably associated with loss of high-energy phosphates. Under conditions of normoglycemia, proton MRS studies have shown that the decrease in pH is linearly correlated with accumulation of lactic acid, although studies in hypoglycemic animals suggest that factors other than lactate accumulation may contribute to brain acidosis. Studies of ischemia in various animals have combined MRS with electroencephalography (EEG) and simultaneous measurement of cerebral blood flow. These studies have reported that whole brain blood flow must decrease to less than 20 ml per 100 g per min before any changes occur with respect to phosphorus metabolites, pH, lactate accumulation, or the EEG. Chronic ischemia studies have shown that although total phosphorus content of the injured tissue decreases,

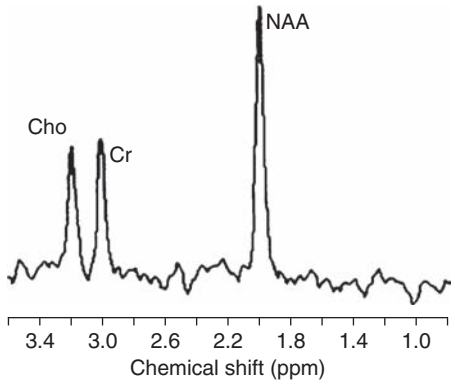

Figure 3 Typical proton MRS spectrum obtained from human brain. Cho, total choline; Cr, total creatine; NAA, *N*-acetyl-aspartate.

the pH and relative concentrations of the metabolites recover to normal values. It has been suggested that this may reflect tissue heterogeneity with necrotic tissue failing to contribute to the MRS spectrum. This loss of signal is also detected in proton spectra in which NAA has been observed to decrease in injured tissue as opposed to lactate, which shows significant increases. MRS studies of ischemia have focused on reflow phenomena and pharmacologic or physiologic intervention of secondary damage, particularly the effects of calcium antagonists, alkalinizing agents, and hypothermia. Although relatively few studies have been reported, it appears that restoration of intracellular pH and bioenergetic state is associated with improved tissue survival after ischemia.

Trauma

Changes in MRS metabolites following experimental brain trauma are very different from those observed in ischemia. For example, trauma that results in mild to severe levels of motor dysfunction does not cause loss of ATP or sustained lactic acidosis. Only when very severe levels of trauma are induced is there a depletion of ATP reserves, most likely as a result of secondary ischemia caused by the extreme nature of the insult. Nonetheless, there are significant changes in the MRS spectra following less-severe levels of traumatic brain injury (TBI). Brief acidosis is apparent in the immediate posttraumatic period, although the acidosis is slight (pH decreases to a minimum value of 6.8) and transient, recovering by 1.5 h after the injury. The development of intracellular acidosis is correlated with accumulation of lactic acid, which, as shown in ischemia studies, is linearly related to intracellular acidosis. Cellular bioenergetic state also decreases following trauma as reflected in declines in PCr:Pi ratio and calculated cytosolic phosphorylation potential. This decline in the bioenergetic state is thought to reflect an increased energy demand in response to attempts to maintain ionic gradients after trauma.

Phosphorus MRS has also shown that the brain intracellular free Mg^{2+} concentration decreases by as much as 60% after brain injury. In addition to being mandatory for all energy-producing and energy-consuming reactions, the magnesium ion is essential for membrane stability; modulation of K^+, Ca^{2+}, and *N*-methyl-D-aspartate channel activity; protein synthesis; RNA aggregation; DNA synthesis; regulation of vascular tone; and apoptosis. Thus, a decline in the free and tissue concentration of this ion may be involved in many of the reported secondary events following trauma. The evidence demonstrating that intervention which restores Mg^{2+} levels following CNS trauma improves neuronal survival and improves functional outcome suggests that Mg^{2+} may be a critical early factor leading to irreversible tissue damage after CNS trauma. Although the importance of the magnesium ion has long been recognized, MRS studies have now provided evidence, at least in trauma, stroke, migraine, and substance abuse, that free Mg^{2+} may play a central role in metabolic regulation *in vivo*.

Proton MRS studies have demonstrated that NAA declines following trauma in a manner similar to that observed in ischemia. Moreover, the decline in NAA, and particularly gray matter NAA, has been reported to predict outcome after trauma and, specifically, overall neuropsychological performance. Because NAA is restricted to neurons, it has also been proposed that the loss of this metabolite is a marker for neuronal cell death. However, observations suggest that NAA decline may also occur through metabolic depression or in response to altered osmolar state. As such, the use of NAA as an assay of neuronal numbers should be discouraged, although its use as an indicator of neuronal status seems valid. Decreased levels of NAA, or of NAA/Cr ratios, are indicative of a poor prognosis.

Choline levels following trauma have also attracted interest, with studies suggesting that higher levels of this metabolite in the frontal lobe of children may be related to arithmetic ability. Whether this relates to the fact that choline is a precursor to the neurotransmitter acetylcholine is still unknown and requires further investigation. For this reason, changes in individual metabolites are sometimes difficult to interpret, and it has been suggested that combinations of changes may better indicate specific metabolic events. For example, macrophage recruitment in the posttraumatic inflammatory response may be reflected in combined changes of lactate, alanine, glutathione, and polyamines. The difficulty in this approach is that with all the potential permutations, it may be

impossible to establish what specific event is associated with each 'metabolic fingerprint.'

MRI in Brain Injury

Numerous studies have used MRI in the examination of brain injury, particularly with respect to characterizing lesion location and size after insult, as well as overall size of the brain and its substructures. Studies in both traumatic and ischemic brain injury have demonstrated a correlation between lesion size and histopathology, with experimental and clinical studies showing that the MRI-detected lesion size is correlated to outcome. Indeed, a reduction in lesion size by pharmacological means has been noted to result in an improvement in outcome. Trauma studies have further demonstrated that it is not only the size of the lesion but also the location of the lesion that is related to posttraumatic motor outcome, cognition, and memory performance. Childhood brain injury often resulted in white matter atrophy and reductions in hippocampal volume that impacted on outcome, whereas MRI-detected white matter injury, ventriculomegaly, and intraventricular hemorrhage in brain-injured premature infants was associated with abnormal outcome. Taken together, these results clearly demonstrate a promising application of MRI in the clinical management of head injury. A novel application of MRI involves biomechanical characterization of injury; rapid MRI acquisitions permitted the estimation of tissue strains and strain rates following TBI, thus providing the foundation for the development of protective devices and measures against primary injury.

Diffusion-weighted imaging (DWI) and the related diffusion tensor imaging (DTI) have been used to characterize the time course of lesion development and tissue microstructural changes after injury. This application of DWI can detect lesion development within minutes of the insult (Figure 4), whereas DTI can provide information on the integrity of neuronal fiber tracts (fiber tracking). When used to characterize edema formation, DWI provides information on and can differentiate between cytotoxic

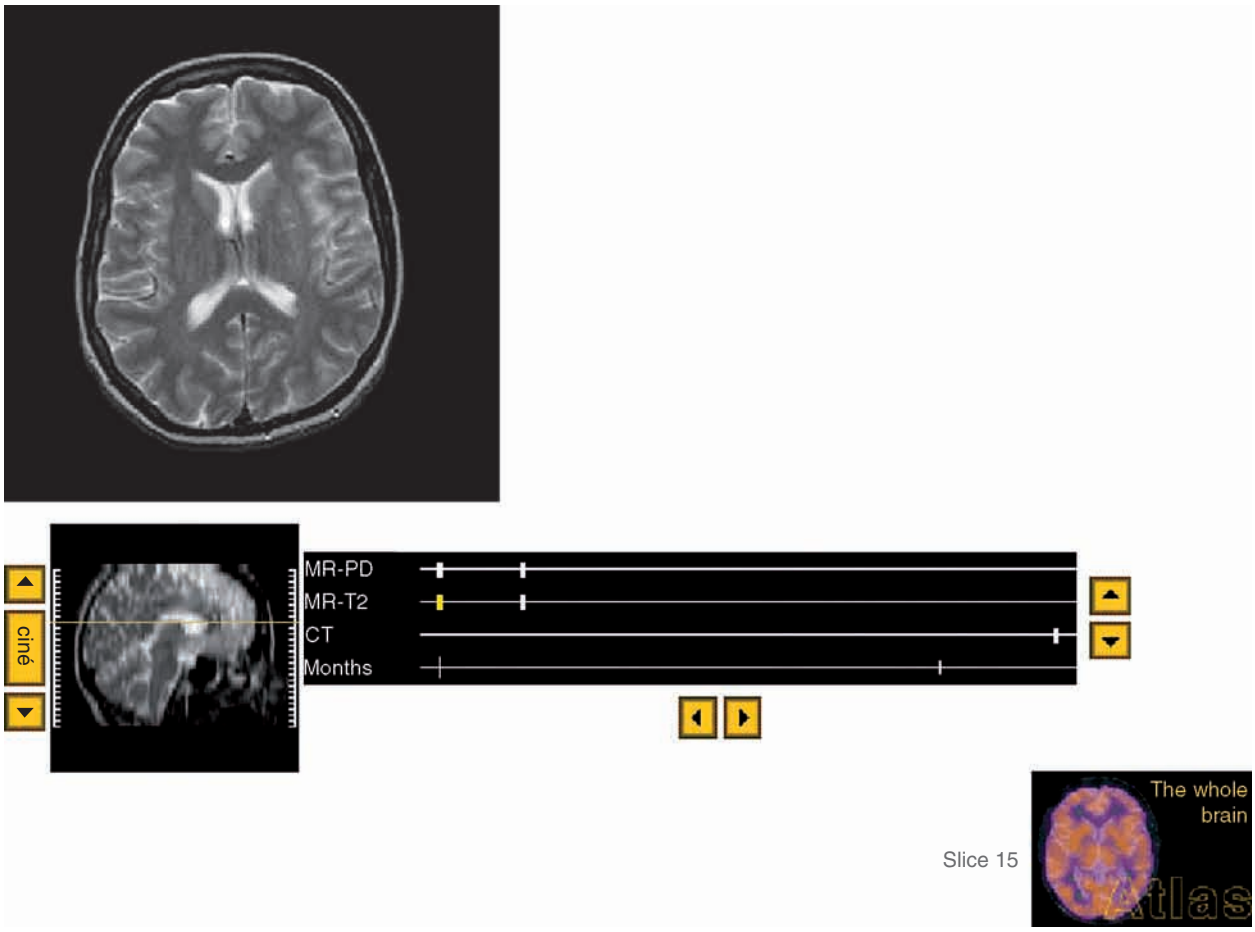

Figure 4 MRI scan of human brain showing the superior ability of diffusion-weighted MRI to detect lesions not readily visible by conventional MRI techniques.

and vasogenic edema. Findings to date have shown that cytotoxic edema is always present in ischemia, whereas vasogenic edema is present in the initial stages of mild to severe brain trauma. Although the reasons for this difference are still unclear, it has been proposed that cellular energy status may be a contributing factor. Another MRI application to brain injury is vascular angiography, which can provide information on blood vessel anatomy and restriction of flow on the basis of blood flow through the vessels. The rate of tissue perfusion can then be determined using perfusion-weighted imaging. Hemorrhage as a result of brain injury is easily detected by MRI (Figure 5), although it is the ability to detect nonhemorrhagic diffuse axonal injury that sets MRI apart from computed tomography. The presence of hemorrhage in diffuse axonal injury-type lesions is a poor prognostic sign, with lesions in the deeper brain regions being more strongly associated with poorer neuropsychological performance. These hemorrhagic diffuse axonal injury lesions are best detected with the recently developed susceptibility weighted MRI techniques, although the gradient-echo techniques are still routinely used with good success.

Although applications of functional MRI (fMRI) in brain injury have been limited, there are a number of encouraging reports suggesting that the technique may become particularly useful in monitoring the brain's ability to recover function as it attempts to repair damaged connections. For example, decreased activity of the anterior cingulated cortex in TBI patients was potentially related to neural network damage most likely associated with diffuse axonal injury. Moreover, in a separate study, the magnitude of fMRI signal change during sensorimotor tasks in TBI patients was correlated to the magnitude of subsequent clinical impairment. In perinatal stroke, functional recovery in the visual cortex was correlated with the recovery of structural integrity in major pathways of the visual system. Clearly, significantly more studies are required before fMRI techniques become routinely utilized in patient care, although the technique shows promise far beyond the original applications in the 1990s.

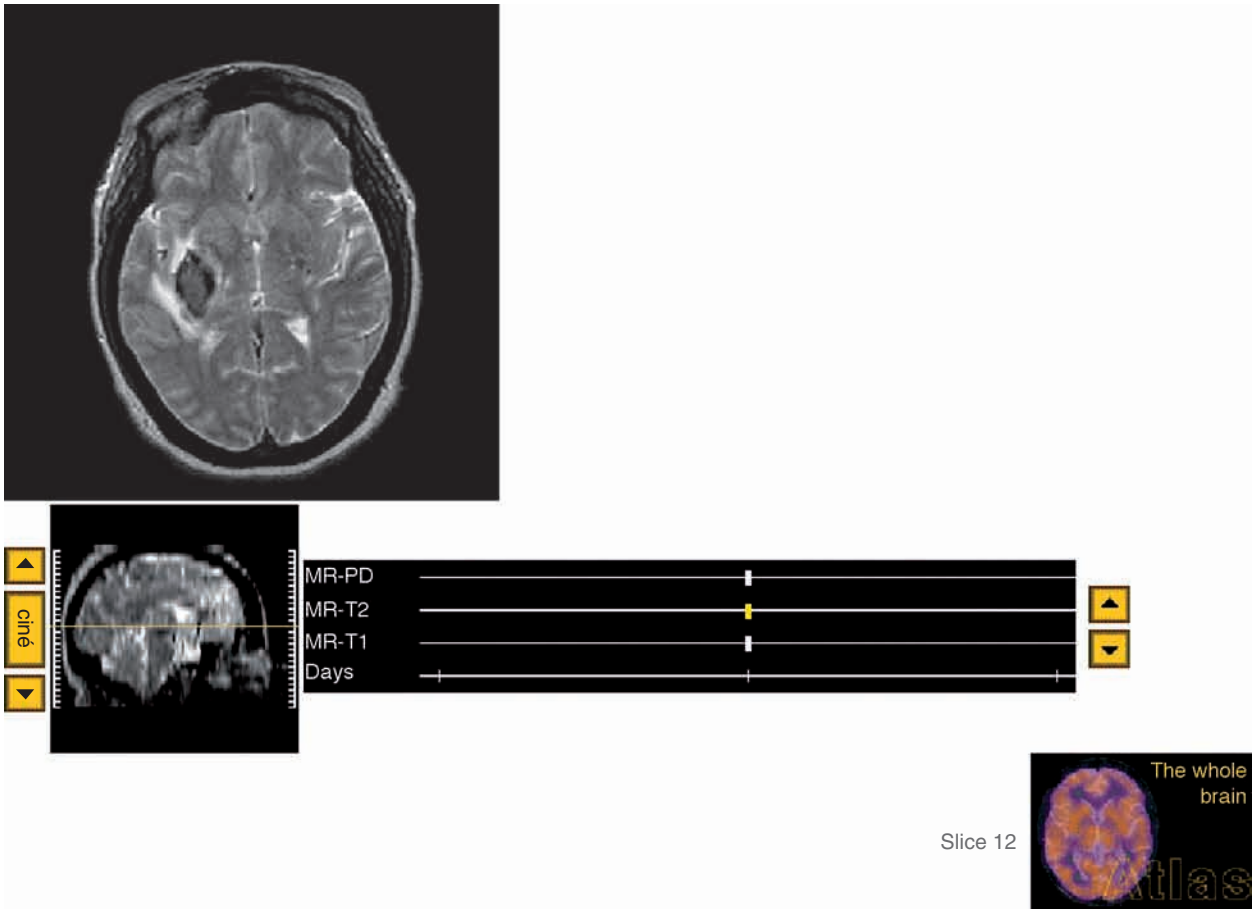

Figure 5 MRI scan of human brain illustrating the presence of hemorrhage.

Future Directions

It is clear that the unique ability of MR techniques to noninvasively monitor physiological and metabolic processes over time permits the investigation of basic mechanisms involved in brain injury and the efficacy of drug treatments to promote recovery. Furthermore, with the development of improved localization techniques, it may soon be possible to determine the selective vulnerability of specific brain regions to injury. Progress in MR technique development has also opened the way for a number of exciting studies. In particular, fMRI techniques are being increasingly utilized to illustrate disorders of brain activation, including speed of information processing, that may be correlated to posttraumatic lesion development. Reports are also appearing that describe the measurement of local oxygen consumption using MRI and MRS-determined rate of glucose consumption in human brain. Furthermore, the development of routine chemical shift imaging techniques will permit the neuroscientist to examine the distribution of a metabolite across the brain in a manner similar to how a radiologist examines anatomical proton images. It appears that future applications of MR techniques will initiate a number of exciting new studies that will further assist in the characterization of the pathophysiological events associated with brain injury.

See also: Alzheimer's Disease: MRI Studies; Brain Injury: Functional Recovery After; Brain Trauma; Cognition: An Overview of Neuroimaging Techniques; Diffusion Tensor Imaging (DTI); fMRI: BOLD Contrast; Magnetic Resonance Spectroscopy; Neuroimaging; Perfusion MRI; Stroke: Injury Mechanisms.

Further Reading

Bammer R, Skare S, Newbould R, et al. (2005) Foundations of advanced magnetic resonance imaging. *NeuroRx* 2: 167–196.

Hunter JV, Thornton RJ, Wang ZJ, et al. (2005) Late proton MR spectroscopy in children after traumatic brain injury: Correlation with cognitive outcomes. *American Journal of Neuroradiology* 26: 482–488.

Jang SH, Kwon YH, You SH, et al. (2005) Medial reorganization of motor function demonstrated by functional MRI and diffusion tensor tractography. *Restorative Neurology and Neuroscience* 23: 265–269.

Johnson KA and Becker JA (1999) *The Whole Brain Atlas*. http://www.med.harvard.edu/AANLIB (accessed in Sep. 2008).

Kobylarz EJ and Schiff ND (2004) Functional imaging of severely brain-injured patients: Progress, challenges, and limitations. *Archives of Neurology* 61: 1357–1360.

Koretsky AP (2004) New developments in magnetic resonance imaging of the brain. *NeuroRx* 1: 155–164.

Lee B and Newberg A (2005) Neuroimaging in traumatic brain injury. *NeuroRx* 2: 372–383.

Ross B and Bluml S (2001) Magnetic resonance spectroscopy of the human brain. *The Anatomical Record (The New Anatomist)* 265: 54–84.

Rutherford MA, Ward P, and Malamatentiou C (2005) Advanced MR techniques in the term-born neonate with perinatal brain injury. *Seminars in Fetal & Neonatal Medicine* 10: 445–460.

Strangman G, O'Neil-Pirozzi TM, Burke D, et al. (2005) Functional neuroimaging and cognitive rehabilitation for people with traumatic brain injury. *American Journal of Physical Medicine & Rehabilitation* 84: 62–75.

Vink R and Cernak I (2000) Regulation of brain intracellular free magnesium following traumatic injury to the central nervous system. *Frontiers in Bioscience* 5: 656–665.

Relevant Website

http://www.med.harvard.edu – The Whole Brain Atlas.

Brain Modules: Mosaic Evolution

R A Barton, University of Durham, Durham, UK

© 2009 Elsevier Ltd. All rights reserved.

Brains Evolved

The only plausible scientific explanation for the existence of what Charles Darwin termed "organs of extreme perfection and complication," such as the brain, is that they evolved over millennia by natural selection. The human brain contains millions or even billions of neurons organized into a variety of interacting but distinct neural systems. Neuroscience has predominantly approached the problem of understanding this complex organization without explicit reference to how it evolved. Yet the functional organization of a system must reflect both the adaptive problems to which it is a solution, and the combined evolutionary and developmental processes from which it emerges. The way that brains are organized is thus intrinsically related to the way that they evolved.

Using Extant Species to Study Brain Evolution

Because the internal structure of brains does not fossilize, their evolution can be studied only indirectly. One indirect method is to analyze the patterns of surface convolutions whose imprints are visible on the internal surface of fossilized crania (or on endocasts of the crania). While useful for documenting evolutionary changes in gross morphological features and certain fine details of cortical surface structure, these data are of limited use in understanding how neural organization evolved, as they reveal nothing about internal structural changes.

The second indirect method is potentially more useful in this respect. It reconstructs evolutionary changes by analyzing comparative anatomy of extant species within a phylogenetic context. A variety of phylogenetic techniques for the analysis of comparative data are now available. The basic rationale for all of these techniques is that the way a trait evolved can be reconstructed by mapping character states in extant species (e.g., the size of a particular nucleus) onto a phylogenetic tree of those species. Reconstruction of the evolution of two or more traits allows evaluation of the patterns of correlated change among the traits (see Figure 1). For example, the hypothesis that two brain nuclei are components of an integrated system implies that they underwent correlated evolution, in terms of their overall size, number of neurons, synaptic connectivity, or other neural characteristics. Whether any such correlated evolution reflects a specific neurobiological relationship between the nuclei, that is, whether it exists independently of more general evolutionary changes, such as those in overall brain size, can be tested by using statistical methods for controlling the effects of confounding variables. Hence, comparative tests of neural organization are based on analyses of the way in which brain structures evolved in relation to one another.

Why is it necessary to use information on the phylogenetic relationships among species in comparative analyses, rather than just correlating the raw species values? The answer is that species are not independent data points in analyses that aim to test hypotheses about how traits evolved. Closely related species are more similar to one another than are distantly related species, because closely related species have undergone less evolutionary divergence. If we are interested in the way in which traits (e.g., the sizes of brain nuclei) evolved in relation to one another, evolutionary changes, rather than species differences, are the most relevant data. Phylogenetic methods explicitly take into account the branching patterns and divergence times between species, and consequently allow valid inferences about trait evolution. These methods are currently little known within neuroscience, despite the fact that neuroscientists often make evolutionary inferences, either explicitly or implicitly, based on comparisons among species.

The process of reconstructing trait evolution on phylogenetic trees emphasizes that extant species do not represent ancestor–descendant relationships. Each extant species is the endpoint of both shared and independent evolutionary histories over an identical period of evolutionary time since their common ancestor. The significance of this point is that no extant species can be treated as an embodiment of the ancestral traits of any other species. Although this point has been made repeatedly, it is still possible to find either implicit or explicit reference in the neuroscience literature to evolutionary pathways between supposedly 'primitive' and 'advanced' extant species. Invariably, such interpretations of species differences are anthropocentrically biased, with phylogenetic affinity to humans being cast as a primitive–advanced continuum (the so-called *scala natura*). This way of thinking about species differences is wrong, and it distracts attention from the ways in which brains have become specialized in different kinds of ways

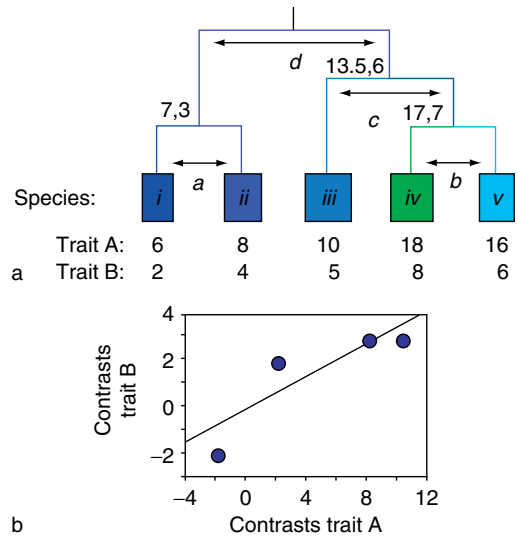

Figure 1 Basic rationale for the phylogenetic analysis of comparative data, using the method of independent contrasts. Independent contrasts are calculated as differences between pairs of species, between species values and an ancestral node, or between higher nodes. (a) Beginning at the tips of the tree, contrast *a* is calculated as the difference between species i and ii, while contrast *b* is calculated as the difference between species iv and v. Then, contrasts are calculated using higher nodes (the 'mixed' contrast *c* and internal contrast *d*; pairs of values at internal nodes are the reconstructed ancestral values of traits A and B). Contrasts are thus calculated up the tree, maximizing the number of possible contrasts while using each branch no more than once. The direction of subtraction for calculating contrasts is arbitrary but it is usual to force the independent variable to be positive. These contrasts are independent of one another and can be analyzed using standard statistical techniques. Strictly, the contrasts should be standardized according to the amount of time since evolutionary divergence. In this simplified example, however, unstandardized contrasts are plotted against one another for illustrative purposes (b). Here, traits A and B show positively correlated evolution.

in different lineages. In contrast, studying the patterns of brain evolution across numerous independently changing lineages has the potential to illuminate general aspects of brain organization.

Modularity and Brain Evolution

The different components of organisms tend to evolve in a coordinated fashion. Hence, as overall body size increases, so does the size of limbs, muscles, tendons, and internal organs, and this is necessary for the animal to function well. At the molecular level, such coordinated evolutionary change may involve regulatory genes that determine overall patterns of growth, such as the rate and duration of cellular division. However, the very fact that separate components of organisms can be differentiated implies that they have a degree of structural and functional independence. The concept of modularity refers to the idea of integration within, and relative autonomy among, the different components. Modularity may refer to the way that particular structures function, to the way in which they evolve, or to the way in which they develop during ontogeny. These types of modularity are conceptually distinct yet often related. For example, the idea that a set of components comprises a functional module implies that they are structurally integrated to perform the function. If the integrity of such a functional module is to be preserved when it is modified by natural selection, the constituent components must change together. Such coordinated evolutionary change is facilitated if there is developmental modularity, allowing relatively few genetic changes to cause a cascade of coordinated structural changes. On the other hand, developmental modularity may act as a constraint on 'evolvability': the evolutionary changes that are developmentally possible or likely.

Modularity in organism design is associated with mosaic evolution: to the extent that modules are functionally and developmentally dissociable, organisms can be expected to evolve in a mosaic fashion, with different modules changing in response to different selection pressures. For example, testes are structurally and histologically distinct tissues with the specific functions of sperm and testosterone production. While the size of the testes correlates closely with the size of other organs and overall body size, comparative studies have shown that testes of promiscuously mating species are large relative to body size, as an adaptation to sperm competition. Hence, testes evolved in part independently of changes in the size of other organs and in overall size. Such mosaic evolutionary change appears to have been ubiquitous, and other well-known examples include increased heart and lung size as an adaptation to the energetic costs of flying in bats and birds, variation in relative spleen size associated with degree of exposure to nematode infection, and relatively elongated loops of Henle in the kidneys of desert mammals.

Unlike bodily organs, individual brain components are extensively – even massively – interconnected to support the integrated processing of information and coordinated behavior, potentially limiting the importance of mosaic evolution. The extent to which modularity and mosaic evolution apply to the systems of the brain has, therefore, proved relatively contentious. Nevertheless, much evidence suggests that individual brain components are grouped within structurally, functionally, and developmentally differentiated neural systems specialized for handling particular cognitive operations. Indeed, systems neuroscience is largely predicated on such modularity.

Although there is little argument that the brain is at least partly modular, the nature, extent, and developmental causes of this modularity are still hotly debated. It has also been suggested that, once the fundamental structure of the central nervous system was established early in the evolution of the chordates, the extent of mosaic change in the brain may have been limited by developmental constraints. The fact that all major brain structures tend to show highly correlated size variation across species has been taken as evidence for this developmental constraint hypothesis. Recent evidence from comparative studies, however, provides evidence for mosaic evolution of brain structure.

Comparative Evidence for Mosaic Evolution of Brain Structure

Relative Size Change in Major Brain Structures

The extent to which brain structure appears to have evolved in a mosaic fashion depends on the anatomical level at which the analysis is done. Analyses of variation in the size of major brain subdivisions, such as the telencephalon, tectum, cerebellum, and medulla, generally indicate a pattern of strongly coordinated evolutionary change, with 90% or more of the variance in a brain structure being attributable to overall size. Even at this crude level of analysis, however, there are striking examples of variation that is independent of overall size. Among teleost fishes, specializations for different modes of foraging and sensory processing are associated with dramatic differences in the relative sizes of relevant brain structures (Figure 2). In mammals, the superior colliculus is approximately 10 times larger in ground squirrels than in rats, despite the fact that overall brain size is very similar. And also among mammals, primates have neocortices that are, on average, 5 times larger than in insectivores of similar brain size, with some comparisons again indicating tenfold differences.

These examples of relative size differences are impressive, particularly given that the structures involved connect with many other brain regions and process a variety of types of information. The mosaic evolution hypothesis implies that the differences reflect size changes in specific components within these structures. For example, the neocortex is a highly heterogenous organ, processing information from all the senses and being involved in many different aspects of sensory, motor, and cognitive processing. Concordant with the concept of mosaic evolution, differences in neocortex size are related to the evolution of specific sensory regions. Primate visual specialization provides an example. Primates exhibit a range of peripheral and central specializations for vision, including highly convergent orbits supporting binocular vision, large eyes, trichromatic vision (in Old World monkeys and apes), a concentration of ganglion cells in the central retina; a greatly expanded representation of the central field in visual cortex; a distinctive pattern of projections between eye and brain; a distinctly laminated lateral geniculate nucleus (LGN) within which information from the same hemifield of each retina is brought into visuotopic register in separate layers, before converging on numerous single binocular neurons in the visual cortex; and relatively numerous and extensive visual cortical regions. This suite of visual features was a fundamental component of the adaptive shift in the evolution of the first primates, but there is variability in most or all of these features within the order. Both variation in neocortex size within

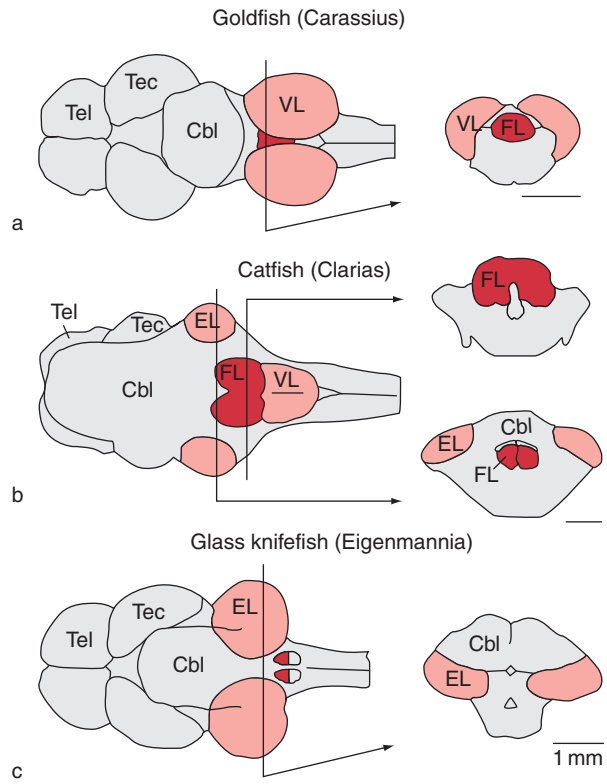

Figure 2 Mosaic brain evolution in teleost fish. Dorsal views of brains on the left, transverse sections on the right. VL, vagal lobes; FL, facial lobes; Tec, tectum; Tel, telencephalon; Cbl, cerebellum; EL, electrosensory lateral line lobe. Sensory specializations are associated with dramatic differences in the size of associated brain structures. Goldfish have enlarged vagal lobes (involved in taste and processing of food). Catfish have large facial lobes (processing gustatory information from the barbells), and relatively large electrosensory lateral line lobes, though the latter are not as large as in electric fish such as the glass knifefish. Courtesy of Striedter GF (2005) *Principles of Brain Evolution*. Sunderland, MA: Sinauer Associates.

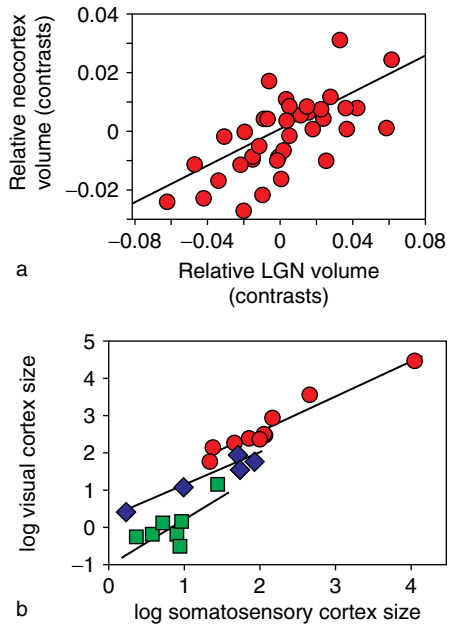

Figure 3 Visual specialization underlies the evolution of large neocortex size in primates. Graph (a) shows the results of a phylogenetic analysis indicating positively correlated evolution between the relative volumes of the whole neocortex and of the LGN, after removing the effects of variation in other brain structures. Graph (b) plots surface areas of visual cortices against those of somatosensory cortices for primates and two other taxonomic groups (red circles are primates, blue diamonds are marsupials, green squares are insectivores). (a) Data from publications by H Stephan and colleagues. (b) Data provided by P Kaskan (Kaskan PM, Franco ECS, Yamada ES, Silveira LCD, Darlington RB, and Finlay BL (2005) Peripheral variability and central constancy in mammalian visual system evolution. *Proceedings of the Royal Society of London B* 272: 91–100).

the primates and differences in neocortex size between primates and nonprimates are related specifically to size change in visual cortices relative to other cortical regions (Figure 3). The data in Figure 3(b) are plotted in log–log coordinates for analytical convenience, masking the absolute size differences. For example, tamarin monkeys have somatosensory cortices slightly smaller than do hedgehogs (22.5 vs. 28.5 mm^2, respectively), but their visual cortices are approximately 9 times larger (158 vs. 17.4 mm^2, respectively). Hence, primates are relative visual specialists, while insectivores are relative somatosensory specialists, reflected in the mosaic nature of neocortical evolution.

Correlated Evolution between Components of Functional Systems

Because neural systems are extended networks with connections that cut across the major structures of the brain, mosaic evolution should be evident in correlated evolutionary changes among connected components, independent of changes in other systems. Quantitative phylogenetic comparative analyses strongly support this prediction: anatomical and functional connections closely predict patterns of correlated evolution among nuclei in a range of neural systems, including the hippocampal complex, amygdala, olfactory system, auditory system, and motor system (Figure 4). In some cases, it has been possible to test the prediction of correlated evolution at the level of individual nuclei. In the corticocerebellar system, evolutionary changes in the size of connected nuclei in the hind-, mid-, and forebrain correlate strongly after controlling for variation in other brain structures. Of the four main vestibular nuclei and three cerebellar nuclei, the two with direct connections (the lateral vestibular and middle cerebellar nuclei) are those that show correlated evolution after controlling for the size of the other nuclei. Hence, anatomical connections predict detailed patterns of correlated evolution, substantiating the hypothesis of mosaic change at the level of neural systems.

Comparative Studies Inform Systems Neuroscience

The cases cited above are examples of testing predictions of evolutionary hypotheses derived from experimental neurobiological data. The fact that brain nuclei co-evolved according to their functional connections also allows us to carry out the reciprocal program, that of testing hypotheses about brain organization using comparative data. A recent example concerns debate about the structural and functional integrity of the amygdala. On the basis of anatomical, pharmacological, and electrophysiological data, Swanson and Petrovich questioned the existence of the amygdala as a structurally and functionally coherent unit, arguing that the term 'amygdala' refers to a collection of disparate nuclei that are really parts of different neural systems. Some other researchers, however, emphasize the dense connections between nuclei within the amygdala. The problem here is in deciding what constitutes sufficient evidence to conclude that the amygdala either does or does not exist as a structurally and functionally integrated entity, hence breaking the cycle of claim and counterclaim based on affinities among nuclei and differences between them respectively.

Accepting that, like all complex biological organs, neural systems exist because they evolved by natural selection, the question of whether particular components constitute components of such a unified system is intrinsically an evolutionary question. The crux of the matter is thus how such nuclei evolved: did they evolve together, in a coordinated fashion,

	Rest of brain	Olfactory bulb	Entorhinal cortex	CBL amygdala	Cerebellar nuclei	Visual cortex
Olfactory cortex	0.18 n.s.	0.31 **	0.09 n.s.	0.32 *	0.23 n.s.	
Hippocampus	0.37 n.s.	–0.06 n.s.	0.61 ***	–0.12 –0.77	0.14 n.s.	
CM amygdala	–0.36 n.s.	0.12 n.s.	–0.30 n.s.	0.91 ***	0.65 *	
Vestibular nuclei	–0.38 n.s.	0.13 n.s.	–0.11 n.s.	0.35 *	0.95 ****	
Lateral geniculate	1.04 n.s.	–0.52 *	0.49 n.s.	–0.40 n.s.	–0.70 n.s.	0.83 *

Figure 4 Correlated evolution among primate brain structures is predicted by functional and anatomical connections. Each cell in the figure contains the results of a phylogenetic analysis of the volumes of the corresponding structures. The figures are partial regression coefficients for pairs of structures, after controlling for evolutionary change in all of the other structures in the table. Asterisks indicate statistical significance (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$; n.s., nonsignificant). Predicted correlations are in the diagonal line of blue boxes. Negative correlations between visual and olfactory systems reflect evolutionary differentiation into nocturnal and diurnal niches. The positive correlation between amygdala and olfactory cortices is unsurprising on anatomical and functional grounds (see below). Hence, there are only two correlations that would not be predicted (both between the amygdala and the cerebellar–vestibular system). Overall, these analyses and similar ones for insectivores indicate a clear pattern of strongly and positively correlated evolution among structures within systems, and a lack of such correlated evolution among structures in different systems. Reprinted by permission from Macmillan Publishers Ltd.: *Nature* (Barton RA and Harvey PH (2000) Mosaic evolution of brain structure in mammals. *Nature* 405: 1055–1058), copyright (2000).

and in a way that cannot be attributed merely to their integration within a larger, more global system (e.g., the limbic system, or even the brain as a whole)? Phylogenetic analysis of comparative data in primates and insectivores clearly shows that the amygdala coheres as an evolutionary unit. After controlling for variation in a range of other brain structures, including other limbic structures, separate groups of nuclei in the amygdala show significantly correlated evolution, hence refuting the claim that they are parts of entirely different neural systems (Figure 5).

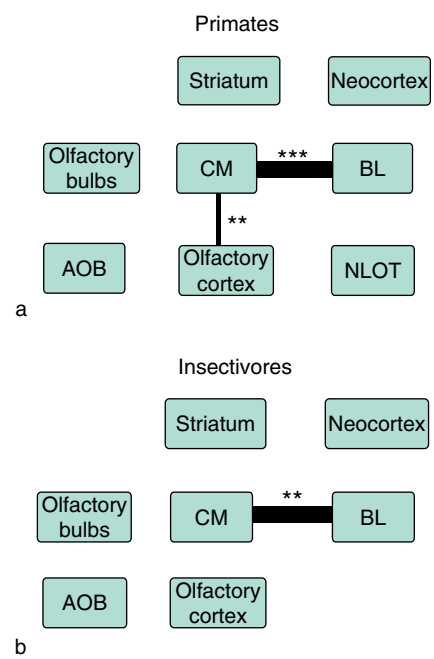

Figure 5 Evolutionary coherence of the amygdala in two mammalian groups. Phylogenetic comparative analysis indicates strongly correlated evolution between separate amygdala components, even after removing the effects of a variety of other limbic and nonlimbic structures. Bars between structures represent significant partial regression coefficients, with bar width reflecting strength of the correlation (**, $p < 0.01$; ***, $p < 0.001$). AOB, accessory olfactory bulbs; BL, basolateral nuclei; CM, centromedial nuclei; NLOT, nucleus of lateral olfactory tract. Summarized from analyses in Barton RA, Aggleton J, and Grenyer R (2003) Evolutionary coherence of the mammalian amygdala. *Proceedings of the Royal Society of London B* 270, 539–544.

Correlated Evolution between Brain Structure and Ecology

If, as suggested, neural system evolution is a result of natural selection on modular information-processing capacities, brain structure should correlate with relevant aspects of animals' ecology and behavioral specialization. Comparative studies have begun to reveal such correlates of mosaic brain evolution. In birds, the relative sizes of the HVC nucleus and of the hippocampus correlate with song repertoire size and spatial memory demands (food caching), respectively. In primates, visual and olfactory system evolution are both associated with ecology, though in different ways: diurnal habits correlate with increased relative volume and cell number in parvocellular, but not magnocellular layers of the LGN, and with decreased

olfactory system size. The evolutionary dissociation between parvo- and magnocellular pathways accords with experimental data on functional segregation within the visual system and on behavioral attributes of different species. Parvocellular projections to visual cortex mediate fine-grained, acute, and color vision, abilities that are characteristic of diurnal primates. Relative parvocellular LGN size and visual cortex size also correlate with foraging behavior and with social group size, independently of the correlation with activity period, suggesting that visual system evolution may have been affected by several aspects of lifestyles. The size of the accessory olfactory bulbs, which mediate pheromonal processing, are correlated with social variables, such as social group size (in New World monkeys) and spatial dispersion (in nocturnal lemurs and lorises), whereas the main olfactory bulbs correlate more strongly with nonsocial variables such as diet. In carnivores, relative olfactory bulb size also correlates with ecological variables. In bats, wing area and habitat complexity correlate with the relative size of the auditory inferior colliculi (among echolocating species) and with relative hippocampus size.

Neural System Evolution and Brain Size

The adaptive and cognitive significance of brain size has been much debated. If brains are mosaics of independently evolving neural systems, it should be possible to document at least broad types of information processing that have been enhanced during brain expansion (or compromised during brain size reduction). Primates' visual specialization again provides an example. Evolutionary changes in the volumes of both the LGN and primary visual cortices correlate positively with brain size relative to body size after removing the effects of evolutionary changes in other structures. One specific visual factor implicated in brain size evolution is stereopsis, because the degree of convergence of the orbits also correlates with brain size. Thus, brain size evolution in primates is specifically associated with modifications of the visual system. Similar evidence is emerging for mammals (in which the size of the optical input to the brain correlates with brain size) and birds (in which eye size correlates with brain size). These recent results provide the first explanations of brain size variation in terms of specific neural systems. Broader understanding of mammalian brain size evolution will, however, require investigation of mosaic size change across a variety of neural systems and phylogenetic lineages.

Conclusions

The components of any adaptive complex, such as the brain, by definition undergo coordinated evolution. At the same time, functional differentiation of neural systems within the mammalian brain potentially creates scope for a mosaic pattern of evolutionary change overlying the allometric covariation of individual components. Comparative analyses unequivocally show that specific systems and structures have evolved in part independently of change in overall brain size, and that such system-specific change correlates with behavioral ecology. The brains of different species evolved under different selection pressures and mediate different behavioral adaptations. These differences caution against making sweeping inferences about 'the brain' based on study of any one model species, and provide strong justification for comparative studies of brain structure informed by evolutionary principles. By studying how natural selection modified brains, we can derive insights into their functional organization.

See also: Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brain Scaling Laws.

Further Reading

Allman JM (1999) *Evolving Brains*. New York: Scientific American Publications.

Barton RA (2004) Binocularity and brain evolution in primates. *Proceedings of the National Academy of Sciences of the United States of America* 101: 10113–10115.

Barton RA and Harvey PH (2000) Mosaic evolution of brain structure in mammals. *Nature* 405: 1055–1058.

Breuker CJ, Debat V, and Klingenberg CP (2006) Functional evo-devo. *Trends in Ecology and Evolution* 21: 488–492.

Finlay BL and Darlington RB (1995) Linked regularities in the development and evolution of mammalian brains. *Science* 268: 1578–1584.

Harvey PH and Pagel MD (1991) *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Jerison HJ (1973) *Evolution of the Brain and Intelligence*. New York: Academic Press.

Nunn C and Barton RA (2001) Comparative methods for studying primate adaptation and allometry. *Evolutionary Anthropology* 10: 81–98.

Striedter GF (2005) *Principles of Brain Evolution*. Sunderland, MA: Sinauer Associates.

Swanson LW and Petrovich GD (1998) What is the amygdala? *Trends in Neurosciences* 21: 323–331.

Brain Na,K-ATPase

J-D Horisberger and K Geering, University of Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Na,K-ATPase, the Na^+ pump, is a transmembrane protein belonging to the P-type ATPase family. Its primary physiological role is the maintenance of large gradients, inward for sodium (Na^+) and outward for potassium (K^+), across the plasma membrane of all animal cells. Na^+ and K^+ gradients are necessary for the control of resting membrane potential, cell volume, intracellular pH, and calcium concentration, and for import or export of many solutes by means of various coupled transport systems. Na,K-ATPase is also the target of a group of drugs and toxins known as 'cardiac steroids' (e.g., digoxin, ouabain) because of their structure and use in the treatment of congestive heart failure.

Na,K-ATPase: Genes, Subunits, and Isoforms

A functional Na,K-ATPase results from the association of two subunits, α and β (see Figure 1). The main α subunit (~100 kDa), also called the catalytic subunit, contains the cation- and nucleotide-binding sites and the catalytic site responsible for the hydrolysis of ATP and the transfer of the γ phosphate of ATP to an aspartic residue of the α subunit (the phosphorylation site). The β subunit a smaller (~30–40 kDa) glycoprotein, is essential for the maturation and the expression of a stable protein at the plasma membrane. In mammals, there are four known isoforms of the α subunit ($\alpha1$, $\alpha2$, $\alpha3$, and $\alpha4$) and three isoforms of the β subunit ($\beta1$, $\beta2$, and $\beta3$). Except for the $\alpha1$ and $\beta1$ isoforms, which are ubiquitous, the expression of the other isoforms is tissue specific.

There is a good understanding of the structure of the Na,K-ATPase α subunit (see Figure 1), thanks to determination of the high-resolution structure of sarco/endoplasmic reticulum Ca^{2+}-ATPase (SERCA) by X-ray crystallography. The sequences of these two proteins are sufficiently similar to allow reliable molecular structure predictions by homology. In contrast, the three-dimensional structure of the β subunit is not known.

In many tissues, the α/β complex is found associated with a third small protein of about 14 kDa, called FXYD protein because of the presence of a conserved motif FXYD (phenylalanine, any amino acid, tyrosine, and aspartic acid) in its extracellular domain. Seven different FXYD proteins are known (FXYD1–FXYD7). They are also expressed in a tissue-specific manner and they appear to modulate the functional characteristics of Na,K-ATPase.

Na,K-ATPase Isoform Distribution in the Central Nervous System

α Subunit

The central nervous system expresses $\alpha1$, $\alpha2$, and $\alpha3$ Na,K-ATPase isoforms. In adult brain, all three α isoforms are present in all brain regions. $\alpha1$ levels are highest in cerebrum and cerebellum relative to the two other isoforms. On the whole, there are only subtle differences in the macroscopic distribution of the three α isoforms in the brain. The expression pattern in the eye has received particular attention. In the lens epithelium, which is thought to conduct Na^+ and K^+ transport for the entire lens cell mass, $\alpha1$ and $\alpha2$ isoforms are present, whereas in lens fibers, which have a poor ion transport capacity, only the $\alpha1$ isoform is expressed. On the other hand, in the rat retina, the $\alpha1$ and $\alpha3$ isoforms account for greater than 97% of the Na,K-ATPase mRNA. Finally, in the ocular ciliary epithelium, which contributes to the production of aqueous humor, the $\alpha2$ isoform is expressed at high levels in the nonpigmented epithelium, and $\alpha1$ is expressed in the pigmented epithelium. Significantly, Na,K-ATPase α subunits are located in the apical membrane of the retinal pigment epithelium and the choroid plexus, in contrast to the typical basolateral expression of Na,K-ATPase in kidney and intestinal epithelia.

The expression pattern of the different α isoforms in neurons and in glial cells is complex. It is mostly accepted that the $\alpha1$ isoform is expressed in neurons and glia, $\alpha2$ is expressed in glia, and $\alpha3$ isoform is restricted to neurons. However, this does not exclude that some neurons predominantly express the $\alpha3$ isoform and others express $\alpha1$ and $\alpha3$ isoforms. Moreover, despite a preponderance of the $\alpha2$ isoform in astrocytes, some neurons may express $\alpha2$ as well.

α Isoform expression is regulated by various mechanisms. The expression of all α isoforms increases during development, reflecting the increased specific Na,K-ATPase activity observed during brain development. Neuron-specific expression of $\alpha3$ is mediated by a silencer and another *cis*-acting element in the $\alpha3$ subunit promoter. $\alpha3$ expression is only observed in fully differentiated neurons and not in undifferentiated

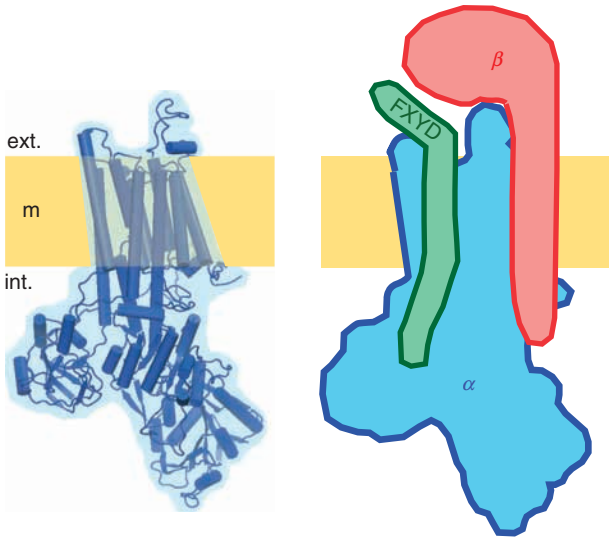

Figure 1 Na,K-ATPase structure. The left-hand panel shows the structure of the α subunit of Na,K-ATPase, obtained by homology modeling with sarco/endoplasmic reticulum Ca^{2+}-ATPase (SERCA; Protein Database code 1SU4). The solid cylinders represent α-helical structures. The plasma membrane (m) is shown in yellow. The transmembrane part of the protein is formed by 10 transmembrane helices. The right-hand panel is a scheme of the two constitutive subunits, α and β, together with the associated small FXYD (phenylalanine, any amino acid, tyrosine, and aspartic acid) protein. The respective position of the subunits is still hypothetical.

pluripotent embryonic stem cells, which express only α1 isoforms. In mature neurons, the α3 isoform contributes to about one-third of the total Na,K-ATPase activity. Its appearance coincides temporally with essential neuronal functions such as outgrowth of axonal and dendritic processes, synaptic events, and responses to neurotransmitters, suggesting that this isoform has a specialized function in neuronal physiology. Polarized as well as nonpolarized distributions of α1 and α3 isoforms in neurons have been described, suggesting that cellular localization of the isoforms may not be fixed but may vary with specific functional demands. As to the glial expression of the α2 isoform, it is thought to be dependent on contact with neurons. Unlike in the adult brain, where α2 is mainly present in glial cells, late in the gestation period of mice, α2 is widely expressed in neurons. This result indicates that α2 may be important in the modulation of neuronal activity in the neonate and could explain why mice lacking α2 isoforms die immediately after birth from lack of breathing. In conclusion, the expression patterns and the contributions of α isoforms to total Na,K-ATPase activity in the brain are likely to be under complex control by neurotransmitters, hormones, and developmental processes.

β Subunit

β1, β2, and low levels of β3 isoforms are expressed in mammalian brain. The β2 isoform has in particular been identified in choroid plexus, retina, vestibular dark cells, and ampullary crests of the inner ear's semicircular ducts, whereas the β3 isoform is expressed in the retinal photoreceptor. In general, the β1 isoform predominates in neurons and β2 predominates in astrocytes, making α1–β1 and α2–β2 Na,K-ATPase isoenzymes the most abundant isoenzymes in neurons and glial cells, respectively. The presence of β3 isoforms has also been observed in oligodendrocytes. β1 isoforms in neurons and in astrocytes show no developmental regulation. The β2 isoform, on the other hand, is observed in cellular processes of astroglia and in soma of neurons in newborn mice, but its expression decreases during maturation and is absent in neurons of adult animals. The differential expression and cellular localization of these isoforms suggest a complex regulation of their gene expression and their cellular distribution. Moreover, the distribution pattern of α and β isoforms is consistent with the possibility that all combinations between different α and β isoforms may exist and may produce different Na,K-ATPase isoenzymes with different functional properties.

The picture is even more complicated by the fact that β2 isoforms in the brain may have two functions. β2 isoforms are partner subunits of α2 subunits, but β2 isoforms are also adhesion molecules on glia (AMOG), which are specifically involved in neuron–astrocyte interactions and promote neurite outgrowth. It is not known whether the β2 isoform exerts two independent functions or whether there is a link between cell adhesion and modulation of the ionic milieu.

FXYD Proteins

Of the auxiliary subunits of Na,K-ATPase, the FXYD proteins phospholemman (FXYD1), phosphohippolin (FXYD6), and FXYD7 are expressed in brain. Both absence and presence of the γ subunit (FXYD2) has been reported, whereas Mat-8 (FXYD3), CHIF (FXYD4), and Ric (FXYD5) are absent from the brain. FXYD1 is mostly expressed in heart and skeletal muscle, but moderate levels of FXYD1 mRNA levels are found in the brain. The FXYD1 protein is particularly enriched in choroid plexus and in cerebellum, where it is expressed in both neurons and glial cells and is associated with all three Na,K-ATPase α isoenzymes. On the other hand, FXYD6 is highly expressed in brain compared to other tissues. FXYD6 is widely distributed in different brain regions and is exclusively expressed in

neurons. Expression of FXYD6 is higher during development than in adult brain and is differentially regulated in different brain regions during development. Finally, FXYD7 is exclusively expressed in brain neurons and glial cells, where it is associated specifically with Na,K-ATPase $\alpha 1$–β isoenzymes. FXYD7 protein expression is observed in several regions of the brain, with lowest expression in the hypothalamus. Overall, the expression of several FXYD proteins in brain, which are associated with Na,K-ATPase, and potentially can modulate its transport properties, adds to the possibility for an extremely fine and complex control of Na^+ and K^+ transmembrane gradients by Na,K-ATPase in the central nervous system.

Functional Role of Na,K-ATPase in the Brain

Role in Neurons and Glia

The neuronal action potential is due to the time-controlled flows of Na^+ and K^+ ions along their electrochemical gradients, and the activity of Na,K-ATPase in neurons is absolutely essential to any neuronal activity because it is the sole transport mechanism responsible for the maintenance of these Na^+ and K^+ gradients across the neuronal membrane. The resting membrane potential, a key determinant of neuronal excitability, is also mainly due to the K^+ gradient across the neuronal membrane. In addition, because Na,K-ATPase works by an asymmetrical exchange of three Na^+ ions for two K^+ ions, it carries a small outward current that tends to hyperpolarize the cell membrane. This effect is small (a few millivolts) under steady-state conditions, but can reach a physiologically significant level (e.g., after a train of action potentials that results in Na^+ loading of the neuron and an increase of the extracellular K^+ concentration).

During a period of high neuronal activity with long trains of action potentials, large amounts of K^+ flow out of the neurons during the repolarization phase of the action potentials. Excess K^+ must then be cleared from the narrow intercellular spaces because a high extracellular K^+ concentration would depolarize the neuronal resting membrane potential. K^+ clearance is performed not only by Na,K-ATPase in the neuronal membrane but also by Na,K-ATPase in the plasma membrane of glial cells surrounding the neuronal membrane.

In neurons and in glial cells, the gradients of Na^+ and K^+ are also essential for the function of secondary active transport systems that allow the import of many organic molecules, including the precursors of neurotransmitters, and particularly for the recapture of neurotransmitters released at synapses, mostly by astrocytes. In fact, a significant part of the energy generated by oxidative metabolism in astrocytes is used to restore the Na^+ and K^+ gradients that are 'consumed' by the neurotransmitter reuptake coupled-transport systems. The phenomenon of glial reuptake of the excitatory neurotransmitter L-glutamate has been studied in great detail (see Figure 2). After its release in the synaptic cleft, L-glutamate can be taken up by several coupled transport systems, in particular by the excitatory amino acid transporters EAAT1 (GLAST) and EAAT2 (GLT-1), both expressed in the astrocyte plasma membrane. These coupled-transport systems have a complex stoichiometry: 1 glutamate + 2(or 3) Na^+ and 1 H^+ entering the cell in exchange for 1 K^+. The reuptake of L-glutamate thus depends on both the Na^+ and K^+ electrochemical gradients and is therefore doubly dependent on adequate function of Na,K-ATPase. Failure of glutamate clearance and its persistence in the synaptic cleft result in excessive glutamate receptor stimulation, a pathological phenomenon named excitotoxicity that has the potential to lead to neuronal damage and death.

The increased workload on Na,K-ATPase following reuptake of L-glutamate triggers glucose uptake by the astrocytes, its metabolic processing, and release of lactate, providing an energy metabolite for neurons. Thus Na,K-ATPase not only maintains the cation gradients but is also part of a metabolic signaling pathway between glia and neurons.

Gradients of Na^+ and K^+ are also important to many other cellular functions not specific to neurons

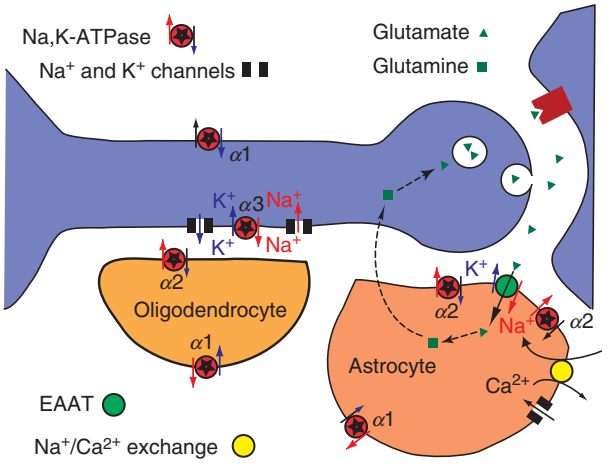

Figure 2 Physiological role of Na,K-ATPase in neurons and glia. This schematic illustrates the role of Na,K-ATPase in the extrusion of Na^+ out of the neuronal and glial cells, in the uptake of K^+ from the intercellular spaces, and in the regeneration of the Na^+ and K^+ gradients used by coupled transport systems such as the Na^+/Ca^{2+} exchanger or the excitatory amino acid transporter (EAAT) (see text for details).

but nevertheless essential to neuronal physiology; these include the maintenance of intracellular pH and Ca^{2+} concentration, thanks to secondary active transporters (Na^+/H^+ exchangers, Na^+/Ca^{2+} exchanger) and the maintenance of cell volume. The α2 isoform appears to have a specially important role in the maintenance of intracellular Ca^{2+}, probably because the Na^+/Ca^{2+} exchanger and the Na,K-ATPase isoforms, including the α2 isoform, may be present together in restricted subdomains of the plasma membrane. This close interaction between the α2 isoform and the Na^+/Ca^{2+} exchanger has been extensively studied in heart muscle, but seems also to be true in astrocytes.

Thus, a reduced function of Na,K-ATPase – due to insufficient energy (ATP) supply because of anoxia or ischemia, inhibition by toxic substances or drugs, or defect of genetic origin – is expected to result in multiple neuronal dysfunction: depolarization of the resting potential, increased excitability, slower conduction velocity, intracellular acidosis or calcium overload, or insufficient reuptake of neurotransmitter (which, in the case of excitatory amino acid neurotransmitters such as L-glutamate, may lead to excitotoxicity). Failure of the transport function of Na,K-ATPase may thus be extremely detrimental for neuronal function and survival.

Pharmacological inhibition of Na,K-ATPase by cardiac steroid drugs results in a wide variety of neuropsychological symptoms when excessive dosage or reduced clearance result in toxic blood levels of these drugs. Common symptoms are visual disturbances (mostly altered color vision due to retinal toxicity), depression, mental confusion, hallucinations, and sometimes delirium. Although these toxic effects are clearly due to Na,K-ATPase inhibition-related neuronal dysfunction, their exact mechanisms are not known in detail. In contrast, some neuroprotective effects of cardiac steroids have been reported in experimental animal ischemia models; in this case the protective effect has been attributed to a reduced energy expenditure.

Information from Animals Models

The first observation of a phenotype resulting from a genetic alteration of Na,K-ATPase was described in *Drosophila melanogaster*. Insertion of a P element in the Na,K-ATPase α subunit locus resulted in a 'bang' phenotype, which is a short-term (a few seconds) paralysis upon exposure to a loud noise. It was shown that these flies had a reduced expression of Na,K-ATPase and also displayed an increased sensitivity to Na,K-pump inhibition by ouabain. More recently, several mutations in the α subunit locus have been described to be associated either with a gain of function or a loss of function of Na,K-ATPase and to result in various phenotypes of temporary paralysis upon exposure to high temperature or mechanical shock, and also age-dependent neurodegeneration. Similar phenotypes can be observed after pharmacological inhibition of Na,K-ATPase activity with ouabain. Taken together, all of these data show that a normal and well-regulated Na,K-ATPase function is required for neuronal function and survival, especially under stressful conditions.

Considering the central role of Na,K-ATPase in cell physiology, it is not surprising that homozygous gene inactivation for the ubiquitously expressed α1 gene is embryonically lethal. Homozygous inactivation of the α2 isoforms prove also to be lethal: $\alpha2^{-/-}$ mice die soon after birth. The cause of death appears to be a deficiency in the control of respiratory movements by the brain stem breathing center, a defect probably due to neuronal development alterations. Mice heterozygous for α1 or α2 gene inactivation display mostly cardiac and skeletal muscle phenotypes, and it is mostly their cardiac functions that have been investigated in detail, but neither heterozygosity shows obvious neuronal phenotypes. Gene inactivation of the α3 and α4 isoforms has not been reported.

Concerning the β subunit, targeted deletion of the β2 isoform results in severe and progressive neurological defects starting at about 2 weeks of age, with tremor and limb paralysis, leading to death within a few days after the onset of symptoms. Pathological alterations are mainly found in astrocytes of the brain stem, with swelling of astrocytic end-feet. In these mice, there is also an accelerated rate of cell death of photoreceptor cells in the retina. Remarkably, replacement of β2 protein by the β1 isoform (β2/β1 knock in) prevents at least a large part of the alterations observed in the $\beta2^{-/-}$ mice, allowing a normal life span in these animals, but the rate of apoptotic death of photoreceptor cells is still increased when compared to wild-type mice.

Human Diseases Associated with Genetic Alterations of Na,K-ATPase Genes

Genetic alterations in Na,K-ATPase function have long been incriminated in cardiovascular diseases such as hypertension, but it is only recently that mutations in Na,K-ATPase genes could be demonstrated to be responsible for human genetic diseases. Mutations in two Na,K-ATPase α subunit genes, encoding α2 and α3, were found linked to essentially neurological phenotypes. Outside the neurobiological domain, mutations in the *FXYD2* gene (Online Mendelian Inheritance in Man (OMIM) database entry #154020) have been shown to lead to hypomagnesemia due to urinary

magnesium loss, probably associated with a dysfunction of Na,K-ATPase in Henle's loop.

Familial Hemiplegic Migraine Type 2

Familial hemiplegic migraine type 2 (*MHP2* (*FHM2*); OMIM #602481) is a rare genetic disease transmitted in the autosomal dominant mode and characterized by the typical symptoms of a migraine with aura; this aura often includes hemiparesia. While other symptomatically similar forms of migraine can be caused by mutations in a gene encoding a Ca^{2+} channel subunit (*CACNA1A* (*FHM1*)) or in a gene encoding a voltage-gated sodium channel (*SCN1A* (*FHM3*)), *MHP2* (*FHM2*) is associated with a number of (at least ten different) mutations in the gene (*ATP1A2*) encoding the $\alpha 2$ isoform of Na,K-ATPase. Several of these mutations have been shown to result in strong decrease or abolishment of Na,K-ATPase transport function. Some mutations result in a failure of stable expression of the protein at the cell surface, in some cases probably because of a defective association with the β subunit, and some other mutations result in expressed but functionally dead mutants. Although the physiopathological mechanism linking the Na,K-ATPase functional alteration to the phenotype is not fully understood, the main hypothesis involves either a defect in the reuptake of excitatory neurotransmitters or a failure of the control of intracellular Ca^{2+} (this last hypothesis being supported by the similar phenotype in *CACNA1A*, linked to gain of function mutations in a calcium channel subunit). Both mechanisms can be causally linked to reduced Na^+ (and K^+) gradients across the plasma membrane.

Rapid Onset Dystonia Parkinsonism

This autosomal-dominant genetic disease, also called dystonia 12 (*DYT12*; OMIM #128235), is characterized by the very rapid onset (over a few hours), in previously asymptomatic patients, of dystonia with signs of parkinsonism. This sudden onset often follows a trauma, a febrile episode, or another kind of stressful triggering event. Dystonia 12 has recently been shown to be linked with several missense mutations in the gene (*ATP1A3*) encoding the $\alpha 3$ isoform of Na,K-ATPase. Each of these mutations results in the modification of residues that are highly conserved in Na,K-ATPase genes throughout the animal kingdom, suggesting that they are functionally important residues. The essential functional role of these residues has been confirmed by showing that each of the *DYT12*-associated mutants is unable to sustain life in a ouabain-resistance test.

See also: Action Potential Initiation and Conduction in Axons; Adenosine Triphosphate (ATP); Excitotoxicity in Neurodegenerative Disease; Glial Glutamate Transporters: Electrophysiology; Glial Glutamate Transporters; Ion Channel Localization in Axons; Parkinsonian Syndromes; Potassium Homeostasis in Glia; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies.

Further Reading

Béguin P, Crambert G, Monnet-Tschudi F, et al. (2002) FXYD7 is a brain-specific regulator of Na,K-ATPase alpha 1-beta isozymes. *EMBO Journal* 13: 3264–3273.

Blanco G and Mercer RW (1998) Isozymes of the Na-K-ATPase: Heterogeneity in structure, diversity in function. *American Journal of Physiology – Renal Fluid & Electrolyte Physiology* 5: F633–F650.

de CarvalhoAguiar P, Sweadner KJ, Penniston JT, et al. (2004) Mutations in the Na^+/K^+-ATPase alpha 3 gene ATP1A3 are associated with rapid-onset dystonia parkinsonism. *Neuron* 2: 169–175.

DeFusco M, Marconi R, Silvestri L, et al. (2003) Haploinsufficiency of ATP1A2 encoding the Na^+/K^+ pump alpha 2 subunit associated with familial hemiplegic migraine type 2. *Nature Genetics* 2: 192–196.

Geering K (2006) FXYD proteins: New regulators of Na, K-ATPase. *American Journal of Physiology – Regulatory Integrative & Comparative Physiology* 2: F241–F250.

Gloor S, Antonicek H, Sweadner KJ, et al. (1990) The adhesion molecule on glia (AMOG) is a homologue of the β subunit of the Na,K-ATPase. *Journal of Cell Biology* 110: 165–174.

Magistretti PJ (2006) Neuron–glia metabolic coupling and plasticity. *Journal of Experimental Biology* 12: 2304–2311.

Peng L, Martin-Vasallo P, and Sweadner KJ (1997) Isoforms of Na,K-ATPase alpha and beta subunits in the rat cerebellum and in granule cell cultures. *Journal of Neuroscience* 10: 3488–3502.

Sweadner KJ (1992) Overlapping and diverse distribution of Na-K ATPase isozymes in neurons and glia. *Canadian Journal of Physiology & Pharmacology* 70: S255–S259.

Weber P, Bartsch U, Schachner M, et al. (1998) Na,K-ATPase subunit beta-1 knock-in prevents lethality of beta-2 deficiency in mice. *Journal of Neuroscience* 22: 9192–9203.

Brain Scaling Laws

M A Changizi, Caltech, Pasadena, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The mammalian brain varies over five orders of magnitude in volume from shrew to whale, and large mammalian brains do not look like small mammalian brains. For example, large neocortices have lower neuron density, are more convoluted, have disproportionately more white matter, and possess a greater number of specialized cortical areas. Knowing the manner in which brains change as a function of size is important in our attempt to understand the fundamental principles governing the brain. In particular, the hope is to find fundamental brain properties that are kept invariant as a function of brain size, the idea being that the brain changes as it does because it has been selected to maintain important invariants. Here, many of the broad ways the neocortex changes as a function of brain size are reviewed, and the principles that may explain these scaling relationships are discussed.

Before proceeding, note that bigger mammals appear to require bigger brains (brain mass increases as the 3/4 power of body mass), whether or not the animal is any more intelligent. (There is currently no accepted explanation for this brain–body scaling relationship.) Larger but similarly intelligent mammals have bigger brains, and their brains will differ from those found in smaller but similarly intelligent mammals, and therefore these differences are not due to differences in intelligence. The brain scaling relationships discussed here are, for this reason, likely to be explained by fundamental dilemmas encountered by trying to build a larger, highly interconnected brain, not explained by principles for building a 'smarter' brain.

Number of Neurons and Neural Connectivity

A first question one might reasonably ask is, how fast does the number of neocortical neurons increase as a function of brain size? The number of neurons increases disproportionately slowly as a function of brain volume, increasing approximately proportional to the 2/3 power of brain volume (i.e., number of neurons $\propto V_{\text{brain}}^{2/3}$) or, equivalently, neuron density decreases as the 1/3 power. In other words, brain volume (and also gray matter volume) increases disproportionately quickly as a function of the number of neurons, namely as the 3/2 power of the number of neurons, so that quadrupling the number of neurons requires an eightfold increase in brain size. (See Figure 1(a) for the approximate scaling relationships.)

How does the number of synapses per neuron vary with brain size? It is known that synapse density appears approximately constant as a function of brain size (on the order of $10^9\,\text{mm}^{-3}$). Bigger brains, then, have lower neuron density but approximately the same synapse density. Because every synapse, of course, belongs to a neuron, it immediately follows that the number of synapses per neuron is larger in larger brains. In particular, the total number of synapses in the brain increases proportional to brain volume (because synapse density is invariant), and the total number of neurons increases proportional to the 2/3 power of brain volume, from which it follows that the average number of synapses per neuron is proportional to $V_{\text{brain}}/(V_{\text{brain}}^{2/3}) = V_{\text{brain}}^{1/3}$. That is, it follows that the number of synapses per neuron increases approximately as the 1/3 power of brain volume (Figure 1(a)).

Because the number of neurons increases as $V_{\text{brain}}^{2/3}$ and the number of synapses per neuron increases as $V_{\text{brain}}^{1/3}$, it follows that the number of synapses per neuron, δ, increases as the square root of the number of neurons, N; that is, $\delta \propto \text{N}^{1/2}$. Assuming that the neocortex is small world (i.e., having sufficiently many axons connecting otherwise separated parts of neocortex that the network diameter can be approximated as that of a random network), the network diameter (i.e., the average number of axons that must be traversed to get from any neuron to another) may be approximated as $\Lambda \approx (\log N)/(\log \delta)$, which can be manipulated into $\Lambda \approx 2 \times [1 + (\log c)/(\log N)]$, where c is the proportionality constant in the equation $\delta = c\text{N}^{1/2}$. For sufficiently large brains, then, the network diameter approaches 2. Also, estimates of c are on the order of 1, meaning that perhaps network diameter may be approximately 2 in actual neocortices, despite the number of synapses per neuron scaling up much more slowly than the total number of neurons (namely as a square root of the total number of neurons) (Figure 1(a)).

We have just reported the approximate empirical scaling relationship for the number of neurons, the number of synapses per neuron, and the network diameter but have not given any hint as to why the brain might conform to these scaling relationships. This will be discussed later, where it will be shown that increasing the number of synapses per neuron as the square root of the number of neurons is the slowest rate of increase possible while still allowing high intra-area connectivity and high area–area connectivity.

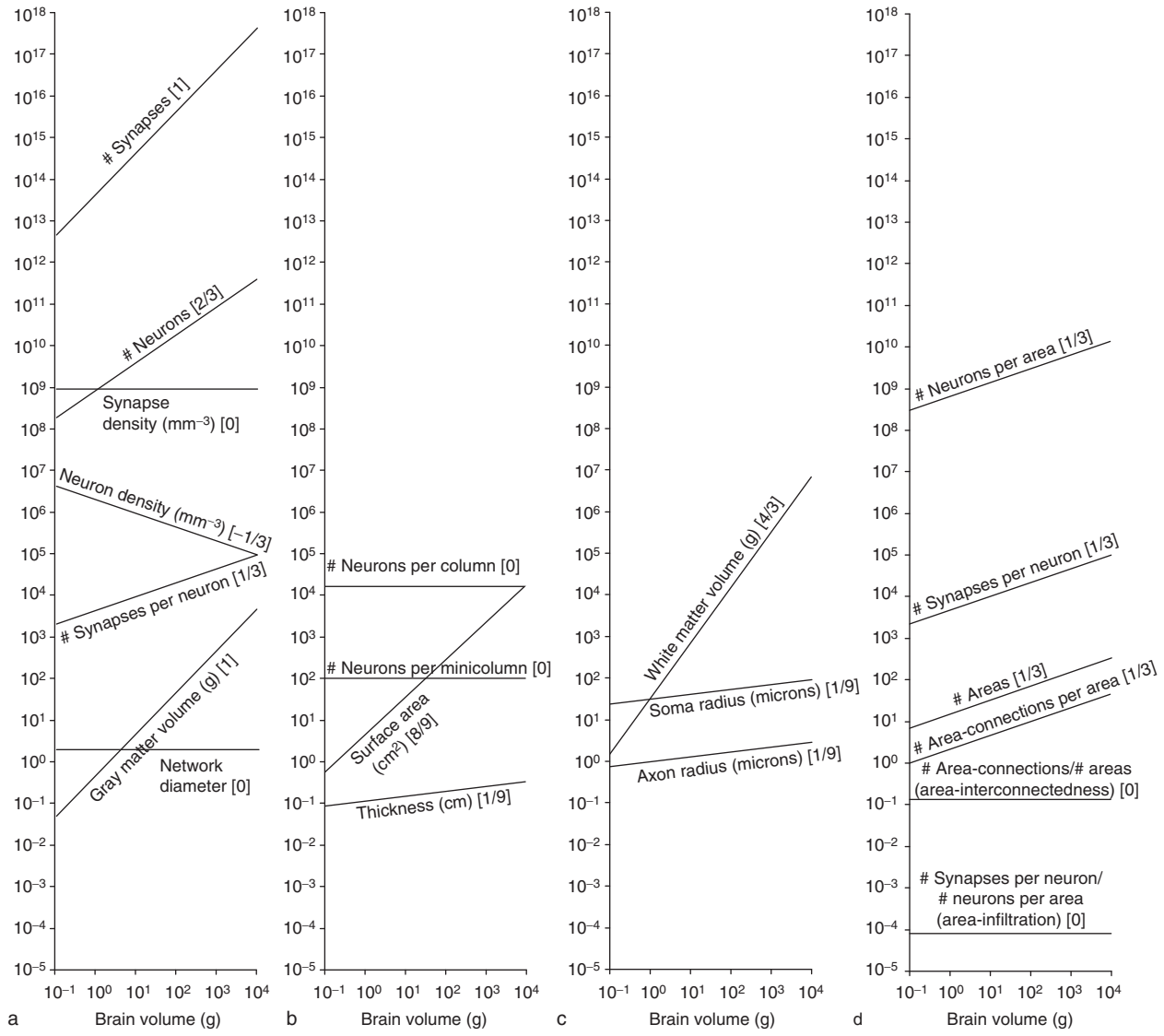

Figure 1 Various neocortical features as a function of brain mass. The number in the square bracket indicates the scaling exponent in each case; for example, for # neurons in (a), the '2/3' in square brackets indicates that the number of neurons varies with brain volume to the power of 2/3. The four plots respectively show the scaling relationships concerned with (a) number of neurons and neural connectivity, (b) gray matter convolutedness and thickness, (c) white matter volume and axon caliber, and (d) specialized areas and area–area connectivity.

Convolutedness and Thickness

One of the most salient differences between small and large brains is that the neocortical surface is smooth in small brains and convoluted in large brains. In fact, the surface area of the gray matter increases as approximately the 8/9 power of brain volume, more quickly than the 2/3 power expected for geometrically similar objects. This difference between 8/9 and 2/3 serves to quantify the rate at which larger brains become increasingly convoluted (Figure 1(b)).

Given that gray matter surface area increases as the 8/9 of brain volume, and given that gray matter volume obviously equals its surface area multiplied by its thickness, it follows that thickness increases as the 1/9 of brain volume, or very slowly (Figure 1(b)).

From this slow increase in gray matter thickness, it is possible to derive an interesting invariant feature of mammalian brains. Recall that (volumetric) neuron density decreases as the 1/3 power of brain volume. Because linear dimensions are the cube root of volumes, the linear neuron density decreases as the 1/3 power of the volumetric neuron density. Therefore, it follows that the linear neuron density decreases as the 1/9 power of brain volume. In other words, the linear dimensions of some fixed number of neurons along a line will increase as the 1/9 power.

In this light, one can see that the slow increase in gray matter thickness is explained entirely by the neuron density decrease in larger brains, meaning that the number of neurons through the cortical thickness (from pia to white matter) is approximately invariant as a function of brain size. Minicolumns are thin neuroanatomical structures consisting of neurons extending from pia to white matter, and we can conclude – and there is evidence – that the number of neurons per minicolumn does not vary as a function of brain size (**Figure 1(b)**). (There is also evidence for another invariant-sized computational unit, the module, which includes larger structures such as column, blobs, barrels, and stripes.)

This suggests an important potential principle underlying why surface area and thickness scale as they do. Namely, gray matter appears to be built out of minicolumns which are repeating computational units of invariant size – that is, computational units possessing a number of neurons that does not vary as a function of brain size – extending through the thickness of the neocortex gray matter. From this principle of invariant-size computational units, and from the neuron density decrease discussed previously, it follows that thickness increases as the 1/9 power of brain volume, entailing, in turn, that surface area scales as the 8/9 power. This explanation for why brains become increasingly convoluted at the rate that they do relies also on the neuron density decrease, and we have not yet provided an explanation for it; this explanation will complete the explanation for why surface area and thickness scale as they do.

Axon Caliber and White Matter

For nearly all biological trees, arbors with more leaves possess thicker trunks. The same is true for neurons: neurons with more synapses tend to have thicker early arbor segments (e.g., the soma or the near-soma axon segment). We have previously seen that larger brains have neurons with a greater number of synapses, and we should not be surprised to find thicker trunked neurons in bigger brains. Although there is large variability in axon caliber, evidence supports that axon caliber tends to slowly increase with brain volume, namely approximately as the 1/9 power (e.g., for white matter axons and for spinal motor neuron somas) (**Figure 1(c)**).

White matter consists of axonal wiring connecting disparate parts of neocortex, and its volume depends, in part, on the caliber of the axons. White matter scales disproportionately quickly as a function of brain size, namely approximately with a scaling exponent of approximately 4/3 (**Figure 1(c)**), although sometimes it has been measured to be as low as approximately 1.2.

What explains these scaling relationships for axon caliber and white matter volume? Let us start with axon caliber; why an exponent of 1/9, rather than some other positive exponent? This exponent is consistent with a law that appears to apply to neural arbors concerning how the diameter of a parent segment relates to the diameter of the daughter segments downstream. In particular, evidence suggests that the cube of the diameter of the parent is equal to the sum of the cubes of the daughter segments (e.g., $R_{\text{trunk}}^3 = R_1^3 + R_2^3$), which is called Murray's law (and has been found to apply to a variety of kinds of natural arbor), and is the power-efficient diameter setting for laminar flow through pipes and also the optimal setting for simultaneously optimizing volume and transmission speed. The 1/9 exponent is derivable from Murray's law as follows: each of the terms in the summation of Murray's law is for one daughter segment and can be replaced by the sum of the cubes of that daughter's daughters, and so on, until reaching the synapses so that $R_{\text{trunk}}^3 = \delta \times R_{\text{synapse}}^3$, where δ is the number of synapses and R_{synapse} is the linear dimensions of synapses. Because synapse density is invariant, R_{synapse} must be invariant, and therefore the equation becomes the proportionality $R_{\text{trunk}}^3 \propto \delta$. However, recall that the number of synapses per neuron scales as the 1/3 power of brain volume, and thus $R_{\text{trunk}} \propto V_{\text{brain}}^{1/9}$.

The 1/9 scaling law for axon caliber is key for understanding white matter volume scaling. The volume of white matter is proportional to the product of the number of white matter-projecting neurons, N_{white}, the length of the average connection, L, and the square of the axon caliber, R; that is, $V_{\text{white}} \propto N_{\text{white}} \times L \times R^2$. Supposing that the number of white matter-projecting neurons scales proportionally with the total number of neurons, and treating the average connection length as proportional to the linear dimensions of the white matter (and recalling that axon caliber scales as the 1/9 power), one may derive that the volume of white matter increases as the 4/3 power of brain volume. If, instead, the average connection length is proportional to the linear dimensions of the brain (rather than to the white matter), white matter volume scales as the 11/9 power of brain volume. White matter volume, then, scales up disproportionately quickly because white matter axon caliber increases as it does, and this, in turn, is due to the increasing number of synapses per neuron and due to Murray's law. If white matter axon caliber were constant (which would require that the number of synapses per neuron be constant), then white matter volume would scale proportionally with brain size. Although we now understand why white matter volume scales the way it does given that the number of synapses per neuron scales as it does, we have not

yet addressed why the number of synapses per neuron scales the way it does, something crucial to brain scaling which we discuss later.

Areas and Area–Area Connectivity

The neocortex is parceled into specialized areas, and the number of areas increases approximately as the 1/3 power of brain volume. Recalling that the number of neocortical neurons increases as the 2/3 power of brain volume, it follows that the number of areas increases as the square root of the number of neocortical neurons. It immediately follows that the average number of neurons per area increases as the square root of the number of neocortical neurons, or as the 1/3 power of brain volume (Figure 1(d)).

We are now in position to notice an important invariance. Recall that the number of synapses per neuron increases as the 1/3 power of brain volume, and therefore the number of synapses per neuron scales proportionally with the number of neurons per area. Therefore, when a neuron connects to an area (either to its own area or to another), it is able to connect to a certain fraction of the neurons there (approximately on the order of 10^{-4}), and this fraction does not appear to vary as a function of brain size (Figure 1(d)). This invariance property is referred to as invariant area infiltration because, independent of brain size, neurons have, on average, sufficiently many synapses to 'infiltrate' an invariant fraction of the neurons in an area. Thus, there appears to be strong selective pressure across mammalian neocortices for the average number of synapses per neuron to scale up just fast enough that neuron interconnectedness within areas can remain invariant and that, when an area connects to another area, the average neuron making the connection has sufficiently many synapses to infiltrate an invariant fraction of the neurons in the area.

Concerning area–area connectivity, evidence suggests that the average number of areas to which an area connects scales proportionally with the 1/3 power of brain volume or, again, as the square root of the number of neurons (Figure 1(d)).

An important invariance principle follows from this. Recall that the number of areas scales as the 1/3 power of brain volume, and therefore the average number of areas to which an area connects scales proportionally with the number of areas. That is, no matter the size of the brain, each area is able to connect to a certain fraction of the areas in that brain (probably approximately 30%), and that fraction does not appear to vary as a function of brain size (Figure 1(d)). This invariance property is referred to as invariant area interconnectedness. There appears, then, to be strong selective pressure across mammalian neocortices for the number of area connections per area to scale up just fast enough that area networks can maintain an invariant level of area interconnectedness. Equivalently, there is selective pressure for the total number of area–area connections to scale as the square of the total number of areas.

We have seen that the number of areas and the number of area connections per area each scale as the 1/3 power of brain volume, but we have not yet provided an explanation. Recall that we have also not yet provided any explanation for why neuron density and the number of synapses per neuron scale as they do. The next section discusses how these are connected and how they may be explained.

Economical Well-Connectedness

Previously, we described that neuron density decreases, and the number of synapses per neuron increases, in larger brains; this fact was crucial for understanding why brains become increasingly convoluted (and the principle of invariant computational units was also key), and it was also crucial for comprehending why white matter volume increases disproportionately quickly (and Murray's law was also key). We have not yet provided an explanation for the neuron density and number of synapses per neuron scaling, nor have we provided an explanation for why the number of areas and the number of area connections per area increase.

In the previous section, we noted that the mammalian neocortex maintains two important invariants – invariant area infiltration (or it could have been named invariant area intraconnectedness) and invariant area interconnectedness. The first states that the number of synapses per neuron increases just fast enough that neurons can connect to an invariant fraction of the neurons within a cortical area. In particular, both quantities increase in larger brains, namely as the 1/3 power of brain volume. This means that no matter the brain size (and no matter the size of the areas), each area maintains constant intra-area connectivity; also, when an area sends an axon to communicate to another area, it is able to synapse to some constant fraction of the neurons there. Invariant area interconnectedness, on the other hand, states that the number of area connections per area increases just fast enough that areas can connect to an invariant fraction of the areas in the neocortex. In particular, both quantities increase as the 1/3 power of brain volume. This means that no matter the brain size (and no matter the total number of areas), the neocortex area network maintains a constant area–area connectivity or, equivalently, the number of area–area edges scales as the square of the number of areas.

That is, the neocortex appears to take a two-tiered approach to high interconnectivity: invariant neural connectivity for the neural network within areas and invariant area connectivity for the entire neocortical area network. It is not understood why the neocortex takes a two-tiered approach of this kind. Why not a one-tiered design in which the number of synapses per neuron scales proportionally with the number of neurons? A one-tiered design would lead to brain size increasing much faster as a function of the number of neurons – and thus a brain of any given size would have to have many fewer neurons – but such a brain would be much more fully connected. It is difficult to state why a one-tiered design would be worse than a two-tiered design, for how does one balance the costs of having fewer neurons against the benefit of having them much more highly interconnected? Alternatively, why not a take a three-tiered design? Again, the advantages of a two-tiered design are not well understood.

Given that the neocortex follows a two-tiered design, is it possible to explain some of the yet unexplained scaling exponents discussed previously (namely why neuron density decreases and why the number of areas increases as it does)? Invariant area infiltration tells us that $\delta \propto W$, where δ is the number of synapses per neuron and W is the number of neurons per area. Invariant area interconnectedness tells us that $D \propto A$, where D is the number of area connections per area and A is the number of neocortical areas. The problem is that these two proportionalities do not tell us how W relates to A. If we knew that, then we could compute how these variables relate to the total number of neurons (which is the product of W and A). Thus, the answer is "no," the two invariance principles are not sufficient to explain everything. The reason is that there are many ways of implementing a two-tiered design. For example, a two-tiered design could just possess two areas no matter the brain size, in which case the number of neurons per area would increase proportional to the total number of neurons. Alternatively, a two-tiered design could try to keep the number of neurons per area low by compartmentalizing as fast as possible.

Although there is a continuum of two-tiered designs (parameterized by how quickly the number of neurons per area increases), some designs are more economical than others in terms of the amount of wire required. If the neocortex requires conformance to invariant area infiltration – as it appears to – then the lower the number of neurons per area, the lower the number of synapses per neuron. Fewer synapses per neuron means less arborization for each neuron and thinner wires, which in turn leads to lower distances for the neurons to travel because the brain can be more compactly organized. That is, the lower the number of neurons per area, the less wiring volume is required. Such a 'save wire' desideratum has turned out to be very successful at predicting nervous system organization at many levels, and appears critical for comprehending why the neocortex scales as it does. Recall that invariant well connectedness referred to the pair of invariants: invariant area infiltration and invariant area interconnectedness. Consider now the hypothesis that the neocortex satisfies invariant well connectedness economically, which we may call economical well connectedness for short. As previously mentioned, this hypothesis expects the number of neurons per area to scale up as slowly as possible, which in turn is to say that the number of areas should increase as quickly as possible. The number of areas cannot scale up faster than the number of neurons per area because if it did, then there would not be enough neurons in an area to connect up to an invariant fraction of all the areas, violating invariant area interconnectedness. Thus, the fastest that the number of areas can increase is proportional to the number of neurons per area, or $A \propto W$. However, because $N = W \times A$, where N is the total number of neurons, it immediately follows that W and A each scale as the square root of the number of neurons. But then the number of synapses per neuron, δ, and the number of area connections per area, D, also scale as the square root of the number of neurons. That is, $\delta \propto W \propto A \propto D \propto N^{1/2}$. We see, then, how the area and area connection relationships discussed previously are explained. Also, given that synapse density is invariant, gray matter volume (and brain volume) scales as the total number of synapses, or $\propto \delta \times N$, and this is proportional to $N^{3/2}$, or $N \propto V_{\text{brain}}^{2/3}$. That is, this explains the neural density (and number of synapses per neuron) scaling relationships discussed previously – relationships that were also part of the story in understanding convolutedness, axon caliber, and white matter volume.

Conclusion

Despite the large variety of scaling relationships in Figure 1, they can be understood in terms of a relatively small number of principles. Economical well connectedness is the principle that the neocortex maintains an invariant two-tiered design (satisfying invariant area infiltration and invariant area interconnectedness) and does so economically. We saw that from this principle the scaling relationships can be derived for the number of neurons, the number of synapses per neuron, the number of areas, the number of neurons per area, and the number of area connections per area. The result for the total number of

neurons was crucial for understanding why bigger brains become increasingly convoluted, but also crucial was the principle of invariant computational units (stating that the neocortical gray matter consists of invariant-sized minicolumns), without which the convolutedness (or surface area) scaling relationship would not have followed. Finally, in order to understand why axon caliber and white matter volume scale as they do, it was necessary to understand Murray's law, which governs the rate at which neural arbors become thinner as they near the synapses.

See also: Allometric Analysis of Brain Size; Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brain Fossils: Endocasts; Neocortex: Origins.

Further Reading

Braitenberg V (2001) Brain size and number of neurons: An exercise in synthetic neuroanatomy. *Journal of Computational Neuroscience* 10: 71–77.

Changizi MA (2003) *The Brain from 25,000 Feet: High Level Explorations of Brain Complexity, Perception, Induction and Vagueness*. Dordrecht, The Netherlands: Kluwer Academic.

Changizi MA (2007) Scaling the brain and its connections. In: Kaas JH (ed.) *Evolution of Nervous Systems*. New York: Elsevier.

Haug H (1987) Brain sizes, surfaces and neuronal sizes of the cortex cerebri: A stereological investigation of man and his variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant). *American Journal of Anatomy* 180: 126–142.

Hofman MA (1985) Size and shape of the cerebral cortex in mammals. I: The cortical surface. *Brain Behavior and Evolution* 27: 28–40.

Jerison HJ (1973) *Evolution of the Brain and Intelligence*. New York: Academic Press.

Kaas JH (2000) Why is brain size so important: Design problems and solutions as neocortex gets bigger or smaller. *Brain and Mind* 1: 7–23.

Prothero J (1997) Cortical scaling in mammals: A repeating units model. *Journal of Brain Research* 38: 195–207.

Prothero J (1997) Scaling of cortical neuron density and white matter volume in mammals. *Journal of Brain Research* 38: 513–524.

Rockel AJ, Hiorns RW, and Powell TPS (1980) The basic uniformity in structure of the neocortex. *Brain* 103: 221–244.

Tower DB (1954) Structural and functional organization of mammalian cerebral cortex: The correlation of neurone density with brain size. *Journal of Comparative Neurology* 101: 9–52.

Brain Trauma

I F Dunn, A H Kim, and W B Gormley, Harvard Medical School, Boston, MA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Traumatic brain injury (TBI) is responsible for nearly 2 million emergency room admissions annually in the United States, affecting 2% of the population per year. These injuries range from mild concussions necessitating only observation to large hemorrhages requiring emergent surgery. Approximately 250 000 of these patients require hospitalization, and nearly half of these will undergo surgery for evacuation of a hematoma. Despite significant advances in our understanding of the pathophysiology involved and improved therapy, mortality remains high, at 20%, with significant long-term neurological deficits in 35% of survivors. As a result of this significant morbidity and mortality, a joint task force of the Brain Trauma Foundation and the American Association of Neurological Surgeons (AANS) has developed evidence-based guidelines in an attempt at systematizing the care of patients with TBI. The results of collaborative efforts such as these have focused management emphasis on both the primary injury – that which occurs from the initial traumatic insult – and the secondary injuries, the pathophysiologic events occurring in subsequent minutes, hours, and days after the primary insult.

In this article, we review current concepts in TBI. We first address the areas of classifications and assessment of TBI. Second, we focus on the pathophysiology and management strategies of both primary and secondary brain injury. Common themes across different types of head injuries are emphasized.

Assessment and Classification of Traumatic Brain Injury

Glasgow Coma Score

Developed in 1974 by Jennett and Teasdale, the Glasgow Coma Score (GCS) (Table 1) is the most widely used method of determining the severity of TBI. Included in the assessment are a patient's eye opening, verbal response, and motor response, giving a general gauge of the level of consciousness. Although it has been widely criticized as being excessively simplistic and not accurately reflecting the complete clinical picture of an acutely injured patient, the GCS remains the primary assessment tool in the TBI population. Its simplicity is what has made it so valuable, allowing accurate communication between observers with different levels of expertise in the field. In our experience, the combination of the GCS with a quick but systematic assessment of brain stem function, particularly pupillary reactivity, gives an accurate clinical picture of the neurologically injured patient.

The GCS has been used to help classify injuries according to their clinical severity. Each GCS criterion is assigned a point value. Two broad categories include minor head injury, encompassing injuries that yield total GCS scores from 9 to 15, and severe head injury, with total GCS scores from 3 to 8. Each category of head injury is treated differently. A simple guide to the management of minor head injuries is shown in Figure 1 (severe head injuries are discussed in greater detail later). Minor head injuries may be further divided into minimal, mild, or moderate head injury. With minimal injuries (GCS = 15), patients are neurologically intact and have suffered no loss of consciousness or amnesia. These patients usually do not require computed tomography (CT) scanning except in select cases (see Figure 1 for CT indications) and are discharged home with instructions to return should signs or symptoms of brain injury develop (including headache, nausea/vomiting, seizure). With GCS =14, patients have mild brain injury and may be associated with a brief (<5 min) loss of consciousness or memory impairment. Our institution routinely obtains CT scans on these patients; in one of the largest retrospective studies of mildly head-injured patients, 16% of patients with a GCS of 14 had abnormal CT scans, with 1.6% of patients requiring surgery according to Culotta's study. A GCS score of 9–13 indicates moderate injury wherein patients are able to follow commands at the time of examination but have had longer periods of loss of consciousness (>5 min) and neurologic findings on exam. This group of patients requires careful monitoring, as over 30% of patients in this subgroup have abnormal CT scans, with Stein's group reporting 8% of patients requiring surgery. For this reason, patients should have emergent CT scans with prompt neurosurgical consultation should findings be abnormal (Figure 1). GCS scores from 3 to 8 indicate severe TBI with worse outcomes. Nearly 80% of patients with an initial hospital GCS of 3–5 have an eventual outcome of death, severe disability, or vegetative state; patients with an initial GCS of 3 have a 65% mortality rate. Management of severe brain injury is discussed in greater detail in subsequent sections.

Primary and Secondary Traumatic Brain Injury

Significant morbidity and mortality results not only from the original injury sustained during TBI, but also from ensuing pathophysiologic processes evolving over a period of hours to days. This has prompted the differentiation between what has been termed 'primary' brain injury and 'secondary' brain injury.

Table 1 Glasgow Coma Scale[a]

Patient response	*Score*
Eye opening (E)	
Spontaneous	4
To speech	3
To pain	2
Not open	1
Verbal response (V)	
Conversant	5
Confused	4
Nonsense	3
Sounds	2
Silence	1
Intubated	1T
Motor response (M)	
To command	6
To pain:	
Localizing	5
Withdrawal	4
Arm flexion	3
Arm extension	2
No response	1

[a]GCS score = E + V + M; range from 3 (or 3T) to 15.

Primary Brain Injury

Primary brain injury, referring to the clinical sequelae resulting directly from the initial injury, can be broadly divided into focal and diffuse injuries (Table 2). Focal injuries are related to a narrow force vector delivered to a reduced intracranial region. We see this with epidural hematomas located under a skull fracture caused by a force vector delivered usually in one plane. This type of injury tends to produce intracranial hemorrhages such as epidural or subdural hematomas, or cerebral contusions and intracerebral hematomas. Diffuse injuries are more often related to wider vector and higher acceleration/deceleration injuries, as seen with motor vehicle accidents. Diffuse injuries comprise the clinical spectrum from concussion to diffuse axonal injury (DAI) with widespread long tract signs and posttraumatic coma.

The first critical step in the management of traumatic brain injury is to determine the severity of the primary injury using clinical exam and radiographic evaluation. It is during this assessment of the primary injury that the decision for emergent surgical intervention is required. Once this decision is made and the primary injury is addressed, then the patient moves into the phase of prevention of secondary injury. Management of TBI may thus be considered in two phases: identification and management of primary brain injury followed by the prevention of secondary brain injury. We consider these phases in turn.

Surgical intervention in TBI is unequivocally indicated in a number of clinical scenarios and is primarily indicated in patients with intracranial hemorrhage of sufficient size to compromise neurologic status. Indeed, Pickard and colleagues observed that craniotomy

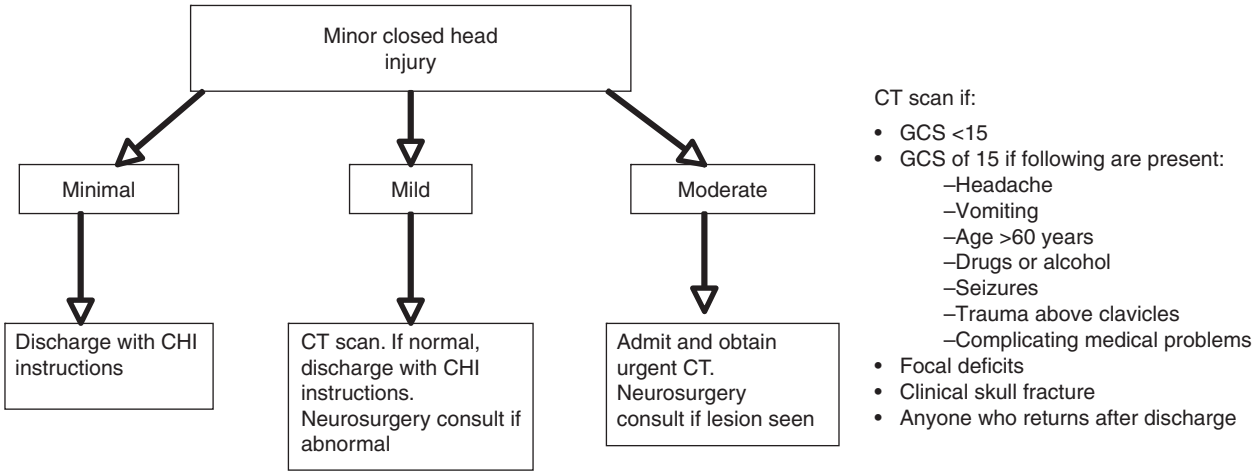

Figure 1 Minor closed head injury (CHI) management: recommendations. CT, computed tomography; GCS, Glasgow Coma Scale. Modified from Ingebrigtsen T, Rise IR, Wester K, et al. (2000) Scandinavian guidelines for management of minimal, mild and moderate head injuries (translation of title; article in Norwegian). *Tidsskr Nor Laegeforen* 120: 1985–1990.

Table 2 Types of structural primary brain injuries

Type of injury	*Characteristic*
Focal	Hematoma
	Epidural
	Subdural
	Intracerebral
	Contusion
	Concussion
	Lacerations
Diffuse	Concussion
	Multifocal contusion
	Diffuse axonal injury

for acute epidural hematoma is one of the most cost-effective interventions among all surgeries, based upon calculations of long-term disability expenses averted and years of productive life gained after surgery. Surgical indications in TBI and traumatic hemorrhages and skull fractures are briefly addressed in the following sections.

Hemorrhages Acute blood can be readily identified on noncontrast CT of the head, an important component in the initial evaluation of a patient with TBI. The main types of traumatic brain hemorrhage requiring surgery are epidural, subdural, and intracerebral. When surgically appropriate, acute epidural hematomas are typically removed by craniotomy due to the gelatinous consistency of an acute blood clot, making it difficult to manage without complete exposure of the area of bleeding. Characteristics of traumatic blood on CT can help guide indications for surgery, although always in the context of a patient's neurologic status. In a group of 158 patients followed prospectively, Servadei et al. reported that over 90% of patients with epidural hematoma with thickness greater than 15 mm and midline shift greater than 5 mm underwent operative intervention with good outcomes. Another study by Bezircioglu et al. analyzed 80 patients with epidural hematoma and observed that patients exhibiting hematoma of volume less than 30 cm^3, hematoma thickness less than 20 mm, and no neurological deficit could be managed nonoperatively in the vast majority of cases (94%), but with close monitoring. Expectant management can be a reasonable option if the mass effect is minimal and the patient's neurological status is normal, especially in the pediatric population. However, the cost of conservative management is that the patient must be monitored carefully in the intensive care unit for several days.

Accumulation of blood in the subdural space can occur in an acute or chronic manner, and the surgical management of each subgroup differs considerably, owing mostly to the consistency of evolving blood. The acute subdural hematoma can be surgically treated by craniotomy or craniectomy, and in a subset of individuals can be allowed to liquefy to a chronic state and then drained by burrhole evacuation in a delayed fashion. The advantage of the burrhole is the lower surgical risk, as some cases may even be performed under local anesthesia at the bedside. But depending on the acuity and severity of the neurologic compromise due to the subdural blood, delayed or expectant management may not be appropriate. In distinction to the epidural hematoma, predominantly a disease of younger patients (mean age, 20–30 years old), a strong correlation exists between poor outcome and age in patients with subdural hematomas. Indeed, in Howard's study of 23 patients 66 years old or greater with a GCS of less than 9, 74% of patients died. As with epidural hematoma, decreasing the time to surgery improves outcome; Haselsberger and colleagues noted that those patients whose subdural hematomas were surgically evacuated within 2 h of onset of coma had a substantially lower mortality (47 vs. 80%) compared with those operated on 2 h after coma.

An entity that is observed with increasing frequency in the era of modern medicine is the acute-on-chronic subdural hematoma. Often the consequence of anticoagulation for a heart condition such as atrial fibrillation, this group of subdural hematomas can have a significant and sudden acute component on top of an otherwise indolent and inconsequential chronic subdural in the setting of even trivial trauma. The first step in the management of these patients is to stop ongoing bleeding by discontinuing the anticoagulation and correcting the coagulopathy. Although infusion of fresh frozen plasma is a reasonable means of achieving the latter goal, such patients often have comorbidities, including arrhythmias and heart failure, that decrease their tolerance for fluid loads. An alternative that is still undergoing evaluation for traumatic intracerebral hemorrhage is recombinant factor VII, a treatment that Mayer et al. demonstrated as having clear benefits with in limiting the growth of spontaneous intracerebral hemorrhage. The low-volume bolus afforded by factor VII and its efficacy in correction of the coagulation profile make its use plausible in the setting of anticoagulated patients with traumatic intracranial blood, but this issue has thus far been studied in a only a limited number of case reports.

Intraparenchymal hemorrhages occur commonly in TBI, and most are treated nonoperatively. Often these intraparenchymal hemorrhages present as contusions on the inferior frontal lobe surface and the pterional region of the anterior temporal region, sites where the

parenchyma strikes the bony margins upon cerebral displacement. They can initially be small to moderate-sized contusions, which over the course of the ensuing 48–72 h may increase dramatically in size, creating significant mass lesions where only small lesions were present. Consequently, small contusions seen on initial CT scans should be followed closely for the first 3 days, as this represents their period of maximal swelling and increase in size. Even after this period, contusions tend to continue to grow in size, but at a substantially slower rate. This problem is made much more dramatic in elderly patients on traditional anticoagulation or even on antiplatelet agents or in patients with coagulopathy secondary to liver disease or other systemic disorders. When considering surgery, the patient's neurological status, comorbidities, and response to medical therapy are considered, as no clear radiological thresholds have been defined for determining surgical intervention with contusions. Dominant among these issues in considering surgery are the patient's preoperative clinical status and the amount of intracranial mass effect, shift, or herniation that exists. Surgical approaches include craniotomy or craniectomy with evacuation of clot and/or lobectomy.

In all of the aforementioned hemorrhage types, subsequent brain edema and high intracranial pressure (ICP) may not respond to standard medical measures. In these cases, decompressive craniectomy may be attempted during the initial surgical intervention as a measure to assist in the subsequent management of postoperative cerebral edema and potential secondary injury.

Skull fractures Skull fractures are often associated with intracranial hemorrhage and are an important component of the patient evaluation in TBI. Open skull fractures – those that have an overlying scalp laceration – have traditionally been managed with expedited operative debridement with or without removal of bone fragments. Indeed, Jennet and Miller reported that operative debridement of open skull fractures within 48 h was associated with an infection rate of 4.6%, whereas the infection rate jumped to 36.5% with an operative delay greater than 48 h. In such cases, immediate bone replacement may obviate the need for a second surgery (cranioplasty) but, based on first principles, may be predicted to increase infection risk. However, retrospective studies do not support an increased risk of infection with primary bone replacement unless gross contamination is present. A fracture with depression greater than the thickness of the skull has also been argued to be an indication for surgical management, which includes bone elevation or removal. Although on cosmetic grounds this logic is sound, elevation of closed depressed fractures does not necessarily alter the clinical course. For instance, the risk of posttraumatic epilepsy between elevated and nonelevated skull fractures is indistinguishable in Jennett's large retrospective series.

Secondary Brain Injury

Surgery typically is undertaken in response to the primary brain injury. Secondary brain injury, however, refers to cellular processes which take place over hours to days after the initial brain injury and which compound the effects of the initial injury. Secondary brain injury results not only from the delayed effects of primary brain injury, but also from aggravating factors such as hypotension, hypoxia, intracranial hypertension, and inadequate cerebral perfusion pressure (CPP). Untreated, the ultimate results are intracranial hypertension and cerebral ischemia, which can significantly compound the effects of the initial injury.

Intracranial hypertension

Pathophysiology Intracranial hypertension is a common complication of severe head injury; along with the prevention of cerebral ischemia, its management is essential to the arrest of secondary injury. Nearly 50% of head-injured patients with intracranial mass lesions and 33% of patients with DAI have persistently elevated ICP. ICP is determined by the contents of the intracranial vault, the physiology of which was codified by the Monro–Kellie doctrine: the total volume available for intracranial contents is constant and an augmentation in one compartment must be accompanied by a diminution in the other. An uncompensated increase in one or more compartments results in ICP above the normal range of 5–15 mmHg. In the absence of a mass lesion, increased ICP in severe TBI results from alterations in cerebrospinal fluid (CSF) flow, cerebral blood volume (CBV), and the development of vasogenic and cytotoxic edema. Physiologic volume-buffering mechanisms exist and involve alterations of blood and CSF volume and include arteriolar vasoconstriction, increasing cerebral venous outflow, and downward CSF displacement through the foramen magnum or into expanded root sleeves. Once these mechanisms are exhausted, ICP rises in an exponential fashion. The very step nature of the volume–pressure curve complicates ICP management, since once a rise in ICP is noted, it signifies a significant reduction in intracranial compliance (**Figure 2**). Identification of a measure of intracranial compliance that would assist as a predictor of changes in ICP has been a subject of much interest. Sustained ICP readings of greater than 20 mmHg are abnormal. ICP readings above

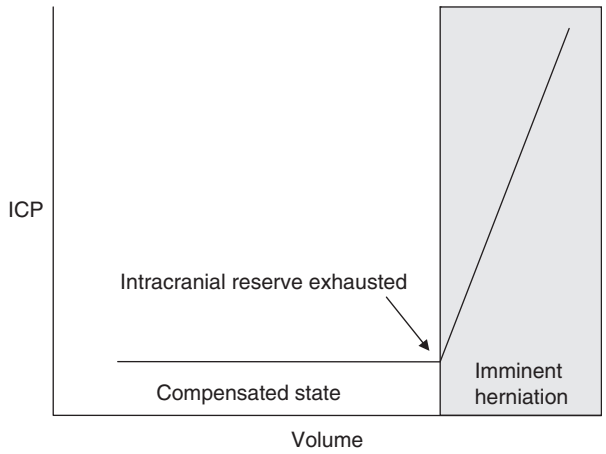

Figure 2 Intracranial pressure (ICP) and volume relationships. Classic pressure–volume curve depicts the exponential rise in ICP once compensatory mechanisms of handling additional intracranial volume are exhausted. In the severely head-injured patient, additional volume may take the form of increased cerebrospinal fluid from poor absorption; increased cerebral blood volume, vasogenic edema from a compromised blood–brain barrier, and cytotoxic edema.

20 mmHg are considered elevated; elevations above 30 mmHg are considered severe in our institution and are treated aggressively. Extensive research has shown that at these pressure levels the patients can readily enter into plateau waves, during which pressures can rise to between 30 and 60 mmHg, creating significant intracranial ischemia and severe secondary injury. ICP readings greater than 40 mmHg represent severe preterminal intracranial hypertension.

The two principal consequences of elevated ICP are cerebral herniation and impaired cerebral perfusion, as denoted by the formula CPP = mean arterial pressure (MAP) – ICP. The two principal forms of cerebral herniation are uncal herniation and central diencephalic herniation. Uncal herniation is produced by medial displacement of the temporal lobe by middle fossa lesions in which the ipsilateral oculomotor nerve, cerebral peduncle, reticular activating system, and, with severe compression, contralateral cerebral peduncle (Kernohan's phenomenon) are compromised. The resulting clinical syndrome involves coma, ipsilateral fixed and dilated pupil, and contralateral hemiparesis. Should the contralateral peduncle be compressed against the contralateral tentorial edge, ipsilateral motor function may also be compromised. Central or downward herniation may result from globally increased ICP and is defined by progressive caudal displacement of the diencephalon through the tentorial incisura with progressive deformation and dysfunction of the diencephalons, progressing to the midbrain and pons. The sequence of events in this progression was eloquently described by Plum and Posner and leads to coma from involvement of the reticular activating system. Basilar artery perforators are stretched and often cause hemorrhages (Duret's hemorrhages).

Cerebral ischemia The avoidance of cerebral ischemia is one of the pillars in the modern management of severe head injury. Management principles center on maintaining an adequate CPP and avoiding systemic hypotension, hypoxia, and anemia so as to ensure adequate cerebral blood flow (CBF). Important in discussing CBF in TBI is CPP, defined as the MAP minus the ICP. Normally, CBF is relatively constant in a range of CPPs between 40 and 150 mmHg. A dynamic system of arterial vasoconstriction and dilation exists to preserve CBF and oxygen delivery. At low perfusion pressures within this range, the arteriolar bed dilates to reduce cerebral vascular resistance (CVR) and increase blood flow to normal levels. At higher perfusion pressures, arteriolar constriction occurs to protect the vascular bed from overperfusion. At perfusion pressures below the autoregulatory threshold, arterioles passively collapse and CBF falls. At pressures above the autoregulatory limits, vessels are maximally dilated and CBF increases.

Pressure autoregulation is lost in many patients following TBI. Bouma and colleagues reported in a series of 117 patients that autoregulation was lost in 49% of patients. Even when pressure autoregulation is intact, the parameters at which autoregulation occurs were often shifted, with significant therapeutic implications. The lower limit of autoregulation may increase from the 40–50 mmHg range to the 70–90 mmHg range, meaning that normally adequate perfusion pressures may cause ischemia in these patients, even when autoregulation is intact.

Given these findings, the clinician must assume that in patients with severe TBI, CBF has a linear relationship to CPP. To prevent cerebral ischemia and maximize outcome, CPP must be closely monitored and maintained. Prospective studies have demonstrated a mean mortality rate of 21% in patients with severe TBI (GCS of 3–7) when CPPs are maintained above 70 mmHg, as compared with a 40% mortality rate among patients with a GCS of 3–7 whose CPPs were not managed as such (the Traumatic Coma Data Bank series). While targets may differ among various researchers, there is consensus that CPP less than 60 mmHg is harmful. The Brain Trauma Foundation currently supports maintaining CPP above 60 mmHg, while other centers advocate a CPP above 70 mmHg.

Aside from the fact that increased ICP can decrease CPP, it is also true that drops in CPP can increase ICP. As CPP decreases, pial arterioles vasodilate and accommodate larger CBVs, diagrammed in Rosner's vasodilatory cascade (Figure 3). This increased blood volume can increase ICP. When CPP is restored, pial arterioles can constrict and ICP will often decrease. Furthermore, adequate cerebral oxygenation is threatened by systemic hypotension and hypoxia. Miller and colleagues showed a 150% increased risk of mortality in patients with severe TBI who had at least one episode of hypotension as defined by systolic blood pressure (SBP) <90 mmHg when compared with patients without hypotension. Similarly, Stochetti and colleagues showed that severe TBI patients with hypoxic episodes prior to intubation had a 77% mortality rate, and 100% of survivors were severely disabled.

Anemia may also decrease tissue oxygenation, and the hematocrit deserves particular attention in the head-injured patient. While an optimal value has not been established, Kee has suggested that brain tissue oxygen delivery begins to decrease below a hematocrit of 33%. Most centers transfuse packed red blood cells in patients with hematocrits below 30%. As with all trauma patients, a hematocrit refractory to blood transfusion should prompt a search for extracranial bleeding sources. Several methods exist to assess oxygenation of brain and blood in trauma patients.

Monitoring

Consensus has emerged regarding indications for placing an ICP monitor in patients with TBI, based on the identification of groups at risk for developing intracranial hypertension. At highest risk are patients with a GCS <8 and an abnormal CT scan; up to 60% of these patients develop elevated ICP readings according to Eisenberg and colleagues. While patients with a GCS <8 with normal-appearing CT scans on admission have a 10–15% chance of developing elevated ICPs, there is a subgroup with a 60% chance of intracranial hypertension: those with age >40 years, SBP <90 mmHg, and unilateral or bilateral motor posturing. We place ICP monitoring devices in all patients with a GCS <8, unless there is a coexisting coagulopathy. We also consider placement in some patients with a GCS >8 who are undergoing extensive anesthesia for extracranial surgery. Ventriculostomy permits drainage of CSF in addition to ICP monitoring, and is especially indicated in patients with hydrocephalus and intraventricular hemorrhage (IVH), or whenever intracranial hypertension is expected to be a significant problem.

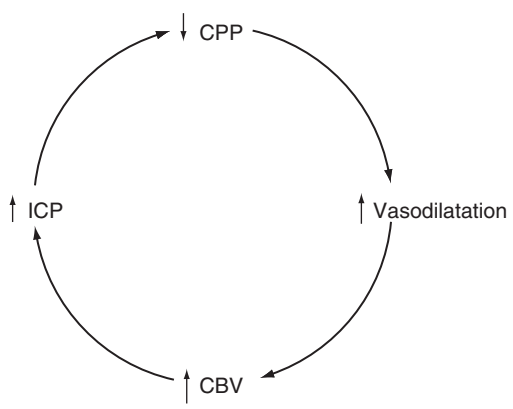

Figure 3 Cerebral perfusion pressure (CPP), intracranial pressure (ICP), cerebral blood volume (CBV), and vasodilatation relationship in preserved autoregulation.

Management of Raised Intracranial Pressure

The goals of intracranial hypertension management are the maintenance of satisfactory CPP and the avoidance of intracranial herniation. The clinical justification for ICP reduction is that elevated ICP is an independent predictor of poor neurologic outcome in patients with severe head injury. While this management principle is well supported, the particular threshold at which treatment is initiated varies. More recently, Marshall's group found that initial or sustained ICP readings above 20 mmHg were effective predictors of poor neurologic outcome among patients with severe head injury. The Brain Trauma Foundation supports initiating treatment at ICP readings between 20 and 25 mmHg, while preserving adequate CBF. A caveat in the rigid adherence to these oft-quoted values is that herniation may occur at ICP readings below 20 mmHg, particularly with temporal lesions. On the other hand, a therapeutic CPP target may be reached with ICP readings as high as 25 mmHg. At our institution, we initiate treatment for ICP readings greater than 25 mmHg, unless CPP is compromised at lower levels. We employ a stepwise approach, beginning with techniques to optimize venous outflow from the brain (head elevation), followed by pharmacotherapy and CSF drainage (Figure 4).

Physical Positioning

We advocate elevating the head of the bed to 30° and maintaining a neutral head position in order to maximize venous return from the brain and thereby

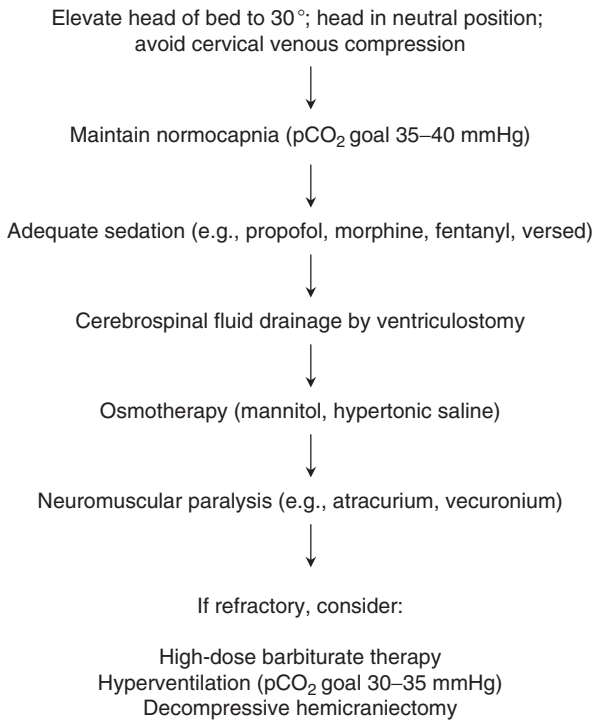

Figure 4 Stepwise protocol for intracranial pressure management.

reduce CBV. While some advocate keeping the head of the bed flat to maximize CPP, Feldan et al. have shown that CPP and CBF may be preserved and ICP reduced by elevating the head of the bed to 30°.

Avoidance of Hypoventilation: CO_2 Reactivity

Changes in arterial carbon dioxide (CO_2) affect CBF and vascular caliber. Hypoventilation leads to an increase in arterial partial pressure of CO_2 ($PaCO_2$), resulting in vasodilatation and an increase in CBF and ICP. While pressure autoregulation may be deranged in severe head injury, carbon dioxide reactivity is preserved. The routine measurement of $PaCO_2$ in head-injured patients is strongly recommended. Hypoventilation raises $PaCO_2$ and ICP, while $PaCO_2$ levels below 35 mmHg may lower ICP but compromise CBF. The ideal $PaCO_2$ level in head-injured patients is normocapnia, to avoid elevations in ICP from hypoventilation and decreased CBF and ischemia from hyperventilation. The only circumstance in which mild chronic hyperventilation may be considered is when cerebral hyperemia – more prevalent in the pediatric population – is demonstrated by either jugular bulb venous saturation ($SjVO_2$) or CBF measurements. Even in these circumstances, however, treatment with chronic hyperventilation is hazardous and should be carried out with the aforementioned monitoring and using extreme caution, as hypoperfusion can be devastating in these patients.

Adequate Sedation and Analgesia; Neuromuscular Paralysis

Sedation and analgesia may blunt the sympathetic stimulation provided by pain from traumatic injury or agitation from brain injury, endotracheal irritation, or routine nursing care, all of which if untreated may increase ICP. The short-acting sedative–anesthetic propofol is being used with increasing frequency because its short half-life permits frequent neurologic examinations and because it appears to have superior effects on ICP reduction compared to narcotic-based regimens. Propofol is a potent vasodilator and careful attention must be paid to MAP and CPP during its administration. When intracranial hypertension becomes severe, patients may require pharmacologic paralysis by neuromuscular blockade for maximum muscle relaxation.

CSF Drainage by Ventriculostomy

CSF drainage via an external ventricular drain (EVD) or ventriculostomy is a highly effective and physiologic method of ICP reduction in head-injured patients. The utility of the ventriculostomy catheter, however, may be compromised in the presence of considerable brain edema and subsequent ventricular collapse. When a ventriculostomy is in place, we use it to drain CSF to its maximal capacity for increased ICP control prior to starting osmotherapy. When an intraparenchymal monitor is in place, however, osmotherapy may be instituted first. In our experience, however, we have found that even when patients present with small ventricles secondary to significant cerebral edema, that a ventricular catheter can be successfully placed. In these circumstances, it becomes invaluable in managing elevated ICP. We, therefore, recommend the use of a ventricular catheter, wherever possible, prior to the use of osmotherapy and paralysis, and certainly before the institution of barbiturate coma.

Osmotherapy

Mannitol is a mainstay in the control of elevated ICP. An osmotic diuretic, mannitol was commonly thought to reduce ICP by reducing intracranial water, but, as first shown by Barry and others, is now recognized to also expand plasma volume and decrease viscosity. Canine and human studies have demonstrated that mannitol increases CBF and decreases ICP by autoregulatory vasoconstriction. Furosemide, which can inhibit the production of CSF, may potentiate the effects of mannitol. When given together, a greater and more sustained decrease in ICP has been observed.

When using mannitol, intravascular volume status, serum osmolarity, sodium, and blood urea nitrogen (BUN) and creatinine must be monitored for possible adverse effects. When the serum osmolarity is raised above 320 Osm, mannitol, particularly in large doses, may precipitate acute tubular necrosis. With prolonged use, mannitol can also disrupt the blood–brain barrier, and it may pass into the brain parenchyma and cause a rebound effect, with subsequent increases in ICP.

Hypertonic Saline Solutions

Hypertonic saline solutions may be used as an adjunct or alternative to mannitol. One advantage of these agents compared to mannitol is a lower risk of rebound intracranial hypertension and renal failure. In hyperosmolar concentrations, sodium chloride creates a gradient which drives water from its interstitial and intracellular compartments into the intravascular space, reducing brain water and ICP. More recent work by Christou and colleagues have suggested that hypertonic saline also has antiinflammatory properties.

Data on the effectiveness of hypertonic saline solutions in the management of raised ICP in head-injured patients have been gathered only in small case series and controlled groups. Quereshi's group reported that 23.4% saline in doses of 30–60 ml over 20 min can markedly reduce ICP when standard treatments fail. Hartl et al. demonstrated that 7.5% hypertonic saline was particularly effective in ICP reduction and CPP elevation in patients with severe TBI whose ICPs were refractory to standard therapy. Hypernatremia is commonly observed but usually resolves as renal free water clearance is reduced. Potential adverse consequences include central pontine myelinolysis and seizures, as well as hypernatremia, congestive heart failure from fluid shifts, and coagulopathy.

Intracranial Hypertension Refractory to Standard Treatment: Other Measures to Consider

The measures described in preceding sections are routinely used to manage intracranial hypertension. When ICP begins to increase despite maximal treatment, other measures are considered, such as hyperventilation with the use of appropriate CBF monitoring, decompressive craniectomy, or barbiturate-induced burst suppression. Prolonged hyperventilation can actually be harmful, and barbiturate-induced burst suppression has not demonstrated to improve outcomes, despite its ability to control reduced cerebral metabolic rates for oxygen ($CMRO_2$). The role of decompressive hemicraniectomy has not been fully defined, as initial studies showed little benefit, but a recent revival in the popularity of the procedure purports to show improved outcomes. Therefore, we use these measures in select situations only, after careful consideration and discussion with family members regarding the advantages and disadvantages of each.

Barbiturates

Barbiturates potentiate the effects of γ-aminobutyric acid (GABA) at GABA receptors, and reduce $CMRO_2$ by globally inhibiting neuronal transmission. Barbiturates may also decrease ICP by limiting free-radical-mediated lipid peroxidation and limiting resultant cerebral edema. The routine use of barbiturates as the prophylactic treatment of ICP is not indicated. In one study, over half of the treated group developed clinically significant hypotension during pentobarbital administration. Nevertheless, barbiturates decreased ICP in two early studies. In a prospective randomized trial, Eisenberg et al. showed that pentobarbital coma in patients with refractory intracranial hypertension had a 2:1 benefit in the control of ICP. Plasma and CSF pentobarbital levels correlate poorly with physiologic effects, and pentobarbital dosing and effects are best followed by electroencephalogram (EEG) monitoring, with dose titration to achieve burst suppression. The most frequently cited and clinically important side effects of barbiturate administration are hypotension and global oligemic hypoxia, followed by hypokalemia, respiratory depression, hepatic dysfunction, and renal dysfunction. For these reasons, this treatment should only be used in conjunction with a Swan–Ganz catheter.

Hyperventilation

Hyperventilation was at one time routinely used as a prophylactic means of reducing intracranial hypertension. However, ICP reduction by hyperventilatory vasoconstriction carries with it the significant risk of inducing cerebral ischemia through inadequate CBF, particularly in the first 24 h after injury, when CBF is already severely compromised. These physiologic effects have been tied to outcome, with Muizelaar et al. reporting an improved outcome at 3 and 6 months when prophylactic hyperventilation was avoided for 5 days after injury. The routine use of hyperventilation should therefore generally avoided, but it remains a rapid method of reducing ICP in emergency situations with acute neurologic deterioration and impending herniation, In such situations, hyperventilation should be maintained only as long as necessary to initiate other treatments.

Decompressive Hemicraniectomy

Delayed decompressive hemicraniectomy, where the bone is removed and the dura is opened to increase

the size of the intracranial compartment, can lower ICP. The first large series of such intervention was the bifrontal decompressive craniotomy reported by Kjellberg and Prieto in 1971; outcomes, however, were generally poor. Gaab et al. reported their prospective results on the use of bifrontal or hemicraniectomy with more stringent inclusion criteria, in which they excluded patients over 40 years of age and those with GCS <5. They reported a 14% mortality rate and a 38% rate of 'full restitution.' Jane's group reported favorable outcomes in 60% of patients undergoing a modified Kjellberg procedure within 48 h of injury, with a GCS between 3 and 8 and intractable hypertension, but with ICP less than 40 mmHg. A recent Cochrane Systematic Review of decompressive craniectomy for refractory ICP identified only one small, randomized, controlled trial addressing this issue in the pediatric population (<18 years old). In this trial containing 27 patients, Taylor and colleagues demonstrated a risk ratio reduction of 0.54 for death, as well as for death, vegetative state, and severe disability 6–12 months after the event in those patients who underwent bitemporal decompressive craniectomy for high ICP. Other studies without direct comparisons with patients that did not have the procedure have also noted favorable outcomes. The particular timing and indications for intervention have yet to be codified.

Management Paradigms

The principal goal of severe TBI management is the limitation of secondary brain injury from cerebral ischemia and intracranial hypertension after any primary injury has been addressed. The avoidance of intracranial hypertension is essential to this goal and has as its physiologic basis the prevention of brain herniation and reduced CPP. This ICP-based approach to patients with severe TBI centers on the maintenance of a normal ICP, with stepwise measures taken as shown in Figure 4.

Others have stressed the importance of CPP-based management, in which maintenance of a particular CPP is central to management, rather than the avoidance of a particular ICP. The guiding physiologic underpinnings of this approach are that a decrease in CPP – whether by an increase in ICP or a decrease in MAP – leads to compensatory vasodilatation and an increase in CBV, which further increases ICP and, in turn, even further compromises CPP (Figure 3). Initial favorable outcomes have led to the widespread adoption of CPP preservation in the management algorithms of severe TBI.

Given these considerations, the three cornerstones of an integrated management paradigm which monitors and controls both ICP and CPP include (1) rapid identification and potential management of intra- or extra-axial mass lesions, (2) ICP monitoring and treatment of intracranial hypertension, and (3) CPP preservation and avoidance of cerebral ischemia (Figure 5). Also crucial in management schemes is meticulous systemic medical management of these patients, including appropriate seizure prophylaxis.

Hypothermia, Neuroprotection, and Restorative Therapies

Future directions in brain injury management include the formulation of strategies to 'protect' the head-injured brain from further secondary injury and to repair injured neural tissue. An early attempt at neuroprotection was hypothermia, which produces proportionate reductions in energy production, use,

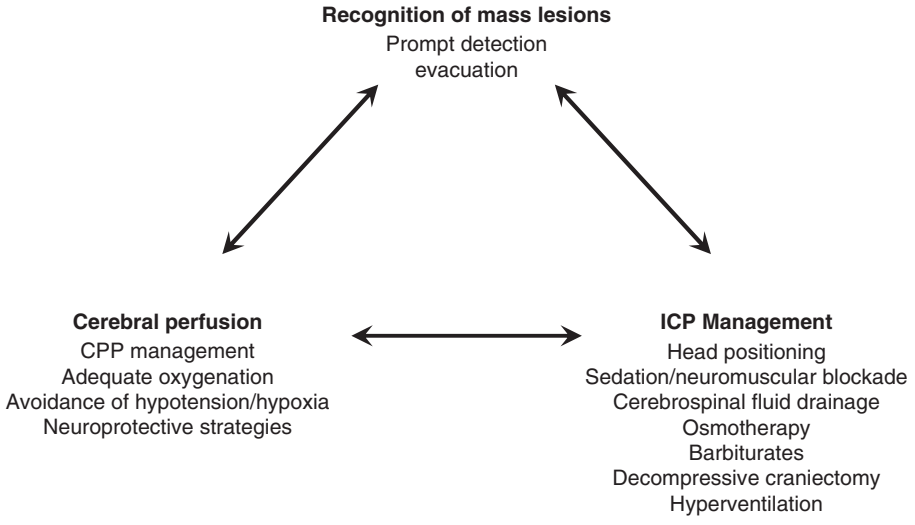

Figure 5 Cornerstones of severe traumatic brain injury management. CPP, cerebral perfusion pressure; ICP, intracranial pressure.

$CMRO_2$, and CBF. While early pilot studies demonstrated favorable outcomes in head-injured patients, a multicenter prospective randomized-controlled trial by Clifton and colleagues failed to show improvement in outcomes. However, a follow-up study in a subgroup of patients that may have benefited is currently in progress.

Therapeutic strategies have also been developed to antagonize the glutamate excitotoxic pathway at the level of the *N*-methyl-D-aspartate (NMDA) receptor to reduce calcium entry into injured neurons, but no convincing effect on outcome has been demonstrated to date. Tirilazad, a 21-aminosteroid that inhibits lipid peroxidation, and the free radical scavenger polyethylene glycol–superoxide dismutase were both studied in large clinical trials, with no benefit to outcome reported.

Another exciting possibility is the stimulation of endogenous neurogenesis or engraftment of exogenous neural stem cells to repair damaged brain by cell replacement or growth factor release. In the coming years, current TBI management paradigms will surely be complemented by highly selective neuroprotective and neurorestorative strategies to enhance the likelihood of meaningful functional recovery.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Injury in Demyelinating Disease and CNS Injury; Blood-Brain Barrier and Neurovascular Mechanisms of Neurodegeneration and Injury; Brain Damage: Functional Reorganization; Brain Injury: Functional Recovery After; Brain Injury: Magnetic Resonance Studies of Metabolic Aspects; Cerebellum: Clinical Pathology; Coma and Other Pathological Disorders of Consciousness; Coma; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Neuronal Plasticity after Cortical Damage; Vascular Issues in Neurodegeneration and Injury.

Further Reading

Brain Trauma, Foundation. The American, Association of Neurological, Surgeons. The Joint, Section on Neurotrauma, Critical Care (2000) Trauma systems. *Journal of Neurotrauma* 17: 457–462.

Bullock R, Chesnut R, Clifton GL, et al. (1996) Guidelines for the management of severe head injury. Brain Trauma Foundation, American Association of Neurological Surgeons, Joint Section on Neurotrauma and Critical Care. *Journal of Neurotrauma* 13: 641–734.

Chesnut RM, Marshall LF, Klauber MR, et al. (1993) The role of secondary brain injury in determining outcome from severe head injury. *Journal of Trauma* 34: 216–222.

Ingebrigtsen T, Rise IR, Wester K, et al. (2000) Scandinavian guidelines for management of minimal, mild and moderate head injuries (translation of title; article in Norwegian). *Tidsskr Nor Laegeforen* 120: 1985–1990.

Juul N, Morris GF, Marshall SB, et al. (2000) Intracranial hypertension and cerebral perfusion pressure: Influence on neurological deterioration and outcome in severe head injury. The Executive Committee of the International Selfotel Trial. *Journal of Neurosurgery* 92: 1–6.

Marshall LF, Smith RW, and Shapiro HM (1979) The outcome with aggressive treatment in severe head injuries. Part I: The significance of intracranial pressure monitoring. *Journal of Neurosurgery* 50: 20–25.

Mayer SA, Brun NC, Broderick J, et al. (2005) Safety and feasibility of recombinant factor VIIa for acute intracerebral hemorrhage. *Stroke* 36: 74–79.

Muizelaar JP, Lutz HA 3rd., and Becker DP (1984) Effect of mannitol on ICP and CBF and correlation with pressure autoregulation in severely head-injured patients. *Journal of Neurosurgery* 61: 700–706.

Narayan RK, Greenberg RP, Miller JD, et al. (1981) Improved confidence of outcome prediction in severe head injury. A comparative analysis of the clinical examination, multimodality evoked potentials, CT scanning, and intracranial pressure. *Journal of Neurosurgery* 54: 751–762.

Plum F and Posner J (1982) *Diagnosis of Stupor and Coma,* 3rd edn., New York: Oxford University Press.

Rosner MJ (1995) Introduction to cerebral perfusion pressure management. *Neurosurgery Clinics of North America* 6: 761–773.

Servadei F, Ciucci G, Piazza G, et al. (1989) A prospective clinical and epidemiological study of head injuries in northern Italy: The commune of Ravenna. *Neurosurgical Review* 12(supplement 1): 429–435.

Stein SC and Ross SE (1992) Moderate head injury: A guide to initial management. *Journal of Neurosurgery* 77: 562–564.

Suarez JI, Qureshi AI, Bhardwaj A, et al. (1998) Treatment of refractory intracranial hypertension with 23.4% saline. *Critical Care Medicine* 26: 1118–1122.

Teasdale G and Jennett B (1974) Assessment of coma and impaired consciousness. A practical scale. *Lancet* 2: 81–84.

Brain Volume: Age-Related Changes

G Bartzokis and P H Lu, University of California School of Medicine, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The popular concept of postnatal human development separated into three stages (development/maturation, adulthood/stability, and aging/degeneration) may be appropriate to human physical characteristics such as height but has distorted the understanding of human brain biology. Over the life span, the human brain is in a continuous state of change roughly defined as periods of development/maturation followed by degeneration/repair. The popular concept of a stable, unchanging 'adult' life stage for the brain is misleading and biologically invalid (Figures 1, 2, and 3). The concept of a stable adult stage (defined as ages 18–55) is valid only when whole brain volume is considered (Figure 1(c)) but fails to capture the underlying infrastructural changes to our brain's 'Internet' (see section titled 'The myelin model of the human brain' and Figures 1(b), 2, 3, and 4(b)). Our very nature as a species depends on optimizing the performance of our brain's Internet, and this performance is dependent on the extensive and protracted process of myelination (see next section and Figure 5).

Magnetic resonance imaging (MRI) can provide images with appropriate resolution and tissue contrasts to quantify *in vivo* the brain's developmental and subsequent degenerative changes. MRI images can measure white matter volumes. When myelin-sensitive inversion-recovery and heavily T1-weighted MRI sequence images are used, they likely include some of the deeper layers of cortical gray matter as myelination spreads into the cortex (lower panels of Figure 2). Thus, these imaging methods measure white matter (WM) volumes that can be better conceptualized as 'myelinated WM volume,' which includes deeper layers of cortex. Not surprisingly, these imaging methods have replicated and extended postmortem observations using myelin stains (Figures 2 and 3). In other words, such sequences are the MRI equivalent of myelin stains, which is the likely reason MRI studies using these methods correspond so closely with postmortem myelin stain data (Figures 2 and 3). Using such images, MRI can detect the shift of the gray matter–myelinated WM boundary into cortical gray matter caused by the process of 'intracortical' myelination (lower panels of Figure 2). This intracortical myelination, together with continued myelination of unmyelinated subcortical fibers (Figure 6(a)) and addition of myelin lamellae in the subcortical myelinated fibers, form the myelinated WM volume. The term 'myelinated WM volume' may therefore be a more appropriate description of the volume measured with inversion-recovery and heavily T1-weighted sequences than simply 'WM volume.' Although this way of conceptualizing MRI data helps explain the similarities between MRI data and postmortem studies, the usual term of 'WM volume' will be used throughout this article except for specific discussion of the impact of intracortical myelination (e.g., Figure 2).

In humans, the disproportional contributor to the enlargement of the late-maturing structures such as the frontal and temporal lobes is their higher WM and myelin content compared to other species (Figure 5). Going against popular belief, recent studies show that compared to primates, adult humans have neither disproportionately enlarged frontal lobes for our overall brain size nor disproportionate increase in gray matter. Rather, human brain has disproportionately more (approximately 20%) WM. These findings are consistent with the overall evolutionary trend to have increasing cognitive abilities associated with disproportionate increases in glia rather than neuron numbers. During brain evolution, glial cell numbers have increased disproportionately compared to neurons; from 10–20% of all brain cells in the nematode to 25% in the fly, 65% in the rodent, and 90% in human. The higher number of oligodendrocytes results in the disproportionately higher percentage of brain dry weight accounted for by myelin in humans (30% higher), compared with rodents.

The Myelin Model of the Human Brain

The human brain is unique in its high myelin content as well as its long developmental (myelination) phase. On average, the myelination phase continues until approximately age 50–60 (Figures 1 and 3) and thus spans our entire evolved life span (before modern medicine, few individuals lived beyond that age). The popular analogy of the brain to a computer is misleading as it ignores the WM (axons, or 'wires,' plus their myelin, or 'insulation') that composes approximately half the adult human brain. A more useful analogy for the human brain is thus the Internet, where speed, bandwidth, and an online connection are the paramount attributes of optimal function. All these attributes of the brain's Internet depend on myelin.

Axon myelination results in saltatory conduction of action potentials that markedly increases (>tenfold) signal transmission speed. Speed makes it

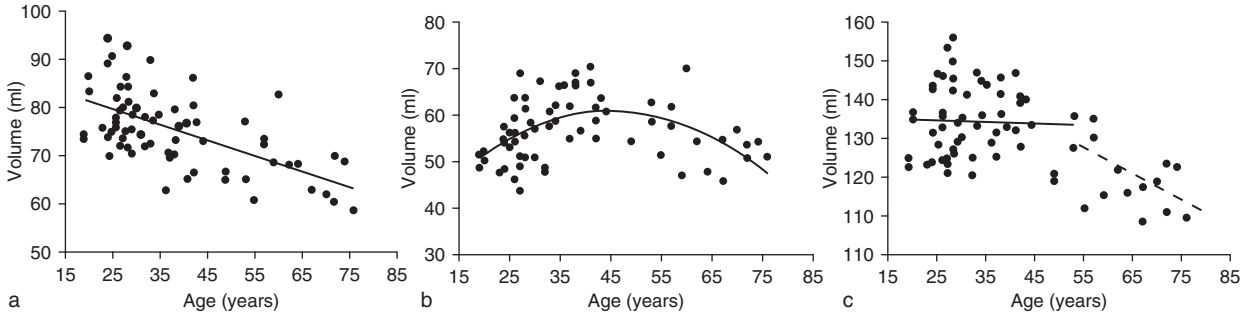

Figure 1 MRI volumes in frontal plus temporal lobes of healthy individuals across the life span. (a) Gray matter. Linear: $r = -0.62$, $p < 0.0001$; Quadratic: not significant (NS), $p > 0.94$. (b) White matter. Linear: NS, $p > 0.84$; Quadratic: $t = 4.56$, $p < 0.0001$, Peak = 45 years. (c) Gray plus white matter. Linear correlations for adults (18–54) and older age (55–76), confirming stable adult brain volume (gray plus white matter) observed in postmortem studies. See Miller AK, Alston RL, and Corsellis JA (1980) Variation with age in the volumes of grey and white matter in the cerebral hemispheres of man: Measurements with an image analyser. *Neuropathology and Applied Neurobiology* 6: 119–132. For details on methods, see Bartzokis G, Beckson M, Lu PH, Nuechterlein KH, Edwards N, and Mintz J (2001) Age-related changes in frontal and temporal lobe volumes in men: A magnetic resonance imaging study. *Archives of General Psychiatry* 58: 461–465.

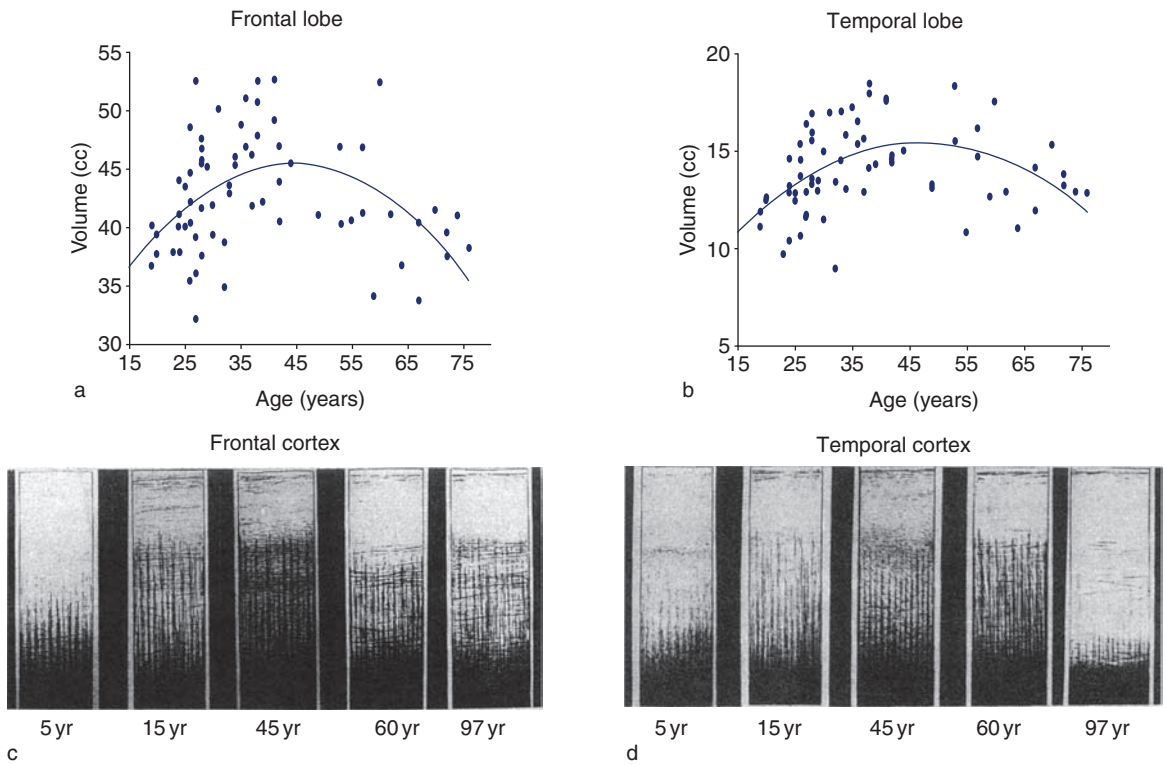

Figure 2 Quadratic (inverted U) myelination trajectories of human brain over the life span. Myelination (*y*-axis) vs. age (*x*-axis) in (a) frontal and (b) temporal lobes of normal individuals. Figures (a) and (b) are *in vivo* MRI data reproduced from Bartzokis G, Beckson M, Lu PH, Nuechterlein KH, Edwards N, and Mintz J (2001) Age-related changes in frontal and temporal lobe volumes in men: A magnetic resonance imaging study. *Archives of General Psychiatry* 58: 461–465, copyright ©(2001) American Medical Association. Figures (c) and (d) are, respectively, frontal and temporal cortex postmortem intracortical myelin stain data from Kaes, 1907; adapted and reproduced in Kemper T (1994) Neuroanatomical and neuropathological changes during aging and dementia. In: Albert M and Knoefel J (eds.) *Clinical Neurology of Aging*, 2nd edn., pp. 3–67. New York: Oxford University Press, with permission from Oxford University Press. The data depicted in (a) and (b) were acquired some 100 years later than the data depicted in (c) and (d), yet the two samples of normal individuals show remarkably similar myelination trajectories in the two regions. Note that different brain regions have significantly different myelination trajectories even when the regions are similar, as is the case with these two association regions. Frontal lobe peak myelination is reached at age 45, but peak myelination is reached even later in the temporal lobe.

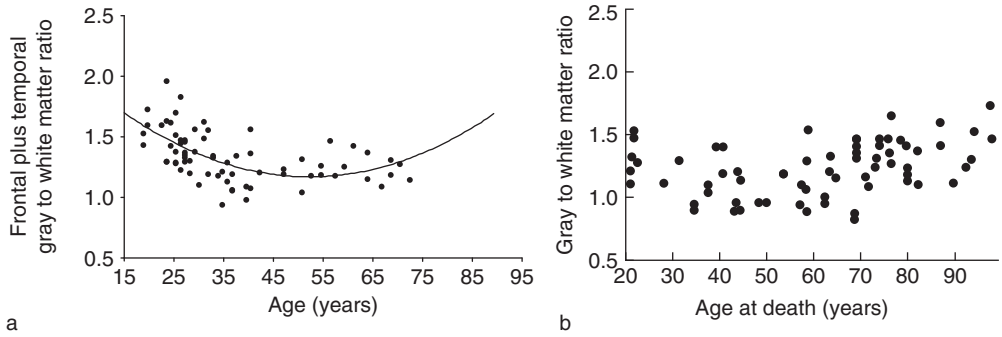

Figure 3 (a) *In vivo* evaluation of frontal plus temporal gray to white matter ratio in healthy individuals. (b) Postmortem evaluation of whole brain gray to white matter ratio. Adapted from Miller AK, Alston RL, and Corsellis JA (1980) Variation with age in the volumes of grey and white matter in the cerebral hemispheres of man: Measurements with an image analyser. *Neuropathology and Applied Neurobiology* 6: 119–132, with permission from Blackwell Publishing.

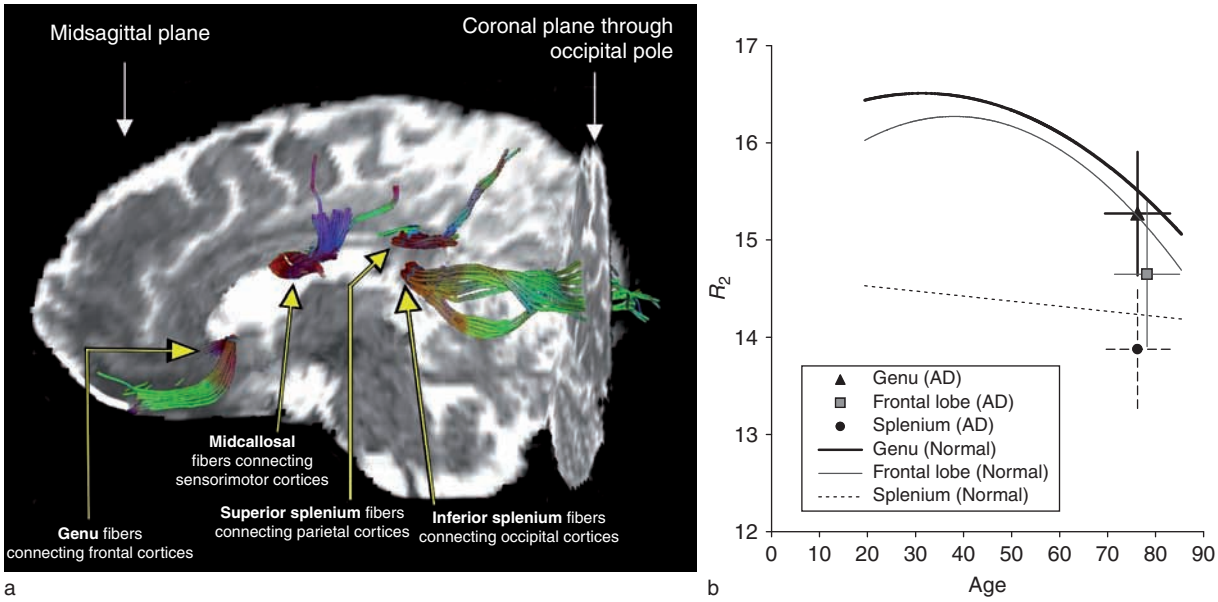

Figure 4 (a) Diffusion tensor imaging showing the trajectory and direction of fibers passing through different regions of the corpus callosum. The directions are coded as colors, with brown representing fibers traveling directly out of (perpendicular to) the page (where the arrows are pointing), green representing the direction that is parallel and horizontal to the plane of the page, and purple representing the direction of perpendicular and vertical to the page. (b) Transverse relaxation rate (R_2) vs. age in healthy and Alzheimer's disease subjects. Normal, healthy controls ($n = 252$, regression lines depicted); AD, Alzheimer's disease patients ($n = 36$, means and standard deviations depicted).

possible to integrate information across the highly spatially distributed neural networks that underlie our higher cognitive functions. Myelination also decreases the refractory time (time needed for repolarization before a new action potential can be supported by the axon) by as much as 34-fold. Thus, if the brain were to be compared to the Internet, myelination not only increases its speed (e.g., transforming it from a telephone line system to a T1 line system) but also its 'bandwidth,' increasing the actual amount of information that can be transmitted per unit of time. This allows myelinated axons to support 'high-frequency bursts' of signals, on which processing speed and long-term potentiation depend, and further facilitates information processing by allowing for precise temporal coding of neuronal activity. However, until a network or circuit is fully myelinated (from neuron body to neuron body), it will not be 'online' to immediately affect ongoing cognition and behavior. Figure 2 demonstrates that from a biological standpoint, becoming a healthy middle-aged adult involves intracortical myelination all the way to the neuron body that brings networks or circuits online. Thus, humans' extensive process of myelination markedly increases cognitive

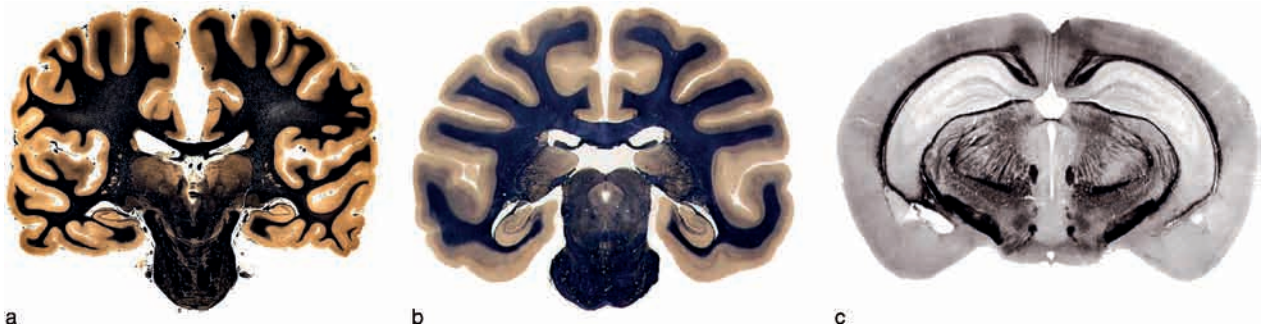

Figure 5 Coronal myelin stains of (a) human, (b) chimpanzee, and (c) rodent brains. Brains are not to scale. Chimpanzee and rodent brain are enlarged to approximate human brain in order to more easily demonstrate the striking differences in the proportion of myelin (stained black) in each brain. Human brain has approximately 20% more white matter than the chimpanzee's.

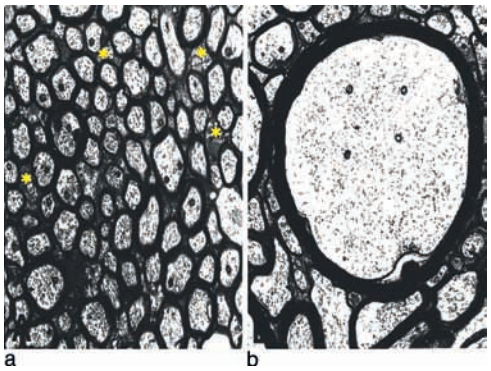

Figure 6 Corpus callosum regional differences in size and myelination of axons. (a) Genu: many small, thinly myelinated axons and clusters of unmyelinated axons (20–30% of axons in this structure connecting late-myelinating frontal regions). (b) Larger, more heavily myelinated axons and far fewer unmyelinated axons (<7%; none in this image) observed in regions like ventral splenium which connect primary sensory (visual) early myelinating regions. Reproduced from Lamantia AS and Rakic P (1990) Cytological and quantitative characteristics of four cerebral commissures in the rhesus monkey. *Journal of Comparitive Neurology* 291: 520–537, with permission from John Wiley & Sons, Inc.

information-processing speed and underlies many of our brain's unique capabilities, such as language, memory, inhibitory control, and higher cognitive functions. This interpretation suggests that the brain could experience neurodevelopmental arrest, even in adulthood, if pathological states (e.g., brain trauma, schizophrenia, severe stress, substance abuse, etc.) alter the normal age-related pattern of structural changes.

The widely distributed neural networks on which the brain's organization is based make speed of processing paramount for higher cognitive functions. The impact of myelin on information processing speed makes the production and maintenance of myelin essential for normal brain function. The vulnerability of the oligodendrocytes and the myelin they produce to genetic and environmental insults likely also underlies our brain's unique susceptibility to highly prevalent and often unique human disorders of development, ranging from autism and learning disabilities to schizophrenia. In older age, such vulnerabilities likely also underlie the exponential increase in degenerative disorders such as Alzheimer's and Parkinson's diseases, whose prevalence is such that they may be considered 'normal' (i.e., affect the majority of the population above the age of 90).

Functional Implications of the Myelin Model

The view of development and aging as a continuum of interactive structural and functional developmental trajectories fosters the examination of multiple factors that promote or inhibit myelination throughout the protracted process of human brain development. Quadratic-like patterns of change over the life span similar to the myelination trajectory have been demonstrated in neurophysiologic and cognitive models of brain function and are also observed in studies of clinical symptomatology in a variety of neuropsychiatric disorders.

Such quadratic-like life-span trajectories have been repeatedly documented in age-related changes of cognitive speed that increase throughout childhood and early adulthood and then decrease with an accelerating rate in old age. Salthouse delineated the hypothesis that a basic parameter such as speed is directly related to biological factors and is essential for higher-order cognitive processing. Similar quadratic trajectories are also apparent in paradigms that test inhibitory control. These quadratic trajectories also resemble the trajectory of some neurophysiologic (electroencephalogram) parameters. The myelin model described above would suggest that the principal underlying biological factor to consider as an explanation for all such life-span trajectories is the process of myelination (Figures 2, 3, and 4).

In the context of the myelin model, the occurrence of all these changes in the structure and function of

the normal human brain peaking at roughly the same midlife stage can be viewed as indirect evidence that after infancy, myelination is the central biologic process underlying human aging-related changes. It is important to point out that the age-related cognitive improvements in processing speed and inhibitory control peak in adulthood. Thus, these developmental trajectories do not temporally 'fit' (by more than a decade) the regressive processes of neuronal, synaptic, axonal, and dendritic pruning and elimination that occur primarily before mid-adolescence.

Brain Development during Childhood to Adolescence

The focus on myelin and myelination is not intended to diminish the importance of other early brain developmental processes such as neurogenesis (primarily intrauterine but whose disruption can cause catastrophic abnormalities that are usually evident at birth or in early infancy). The focus on myelin is especially not intended to dismiss the drastic processes of neuronal, synaptic, axonal, and dendritic pruning and elimination that appear to reduce the 'connectivity' of the child's brain by as much as two-thirds as it develops into adulthood. These processes, which have been hypothesized to affect several childhood disorders, are in large part completed by puberty or mid-adolescence.

These regressive pruning processes occur in concert with a precisely regulated but highly vulnerable and sometimes ignored process of myelination. The regressive pruning processes occur in a heterochronous pattern in primary process areas (motor, sensory) before association areas (frontal, temporal, and parietal lobes; see example in Figure 2). This pattern of heterochronous, volume-reducing, regressive pruning matches the pattern of the heterochronous, volume-increasing, developmental process of myelination (Figure 1). The human skull becomes rigid in late childhood, eliminating the possibility of further expansion of the brain. Thus, it can be argued that the regressive pruning processes of childhood described above occur in large part to provide the 'room' (intracranial volume) and possibly other resources necessary to support the crucial process of myelination.

How can it be that a process that produces such tremendous losses of synapses and axons does not result in dementia in this young age range? To return to the Internet analogy, the pruning of the neurons, and synaptic loss is equivalent to the reduction in capability of the computers of the Internet (e.g., reduced random access memory). The dramatic gains in processing speed (transmission speed, bandwidth, and online connectivity) and other possible benefits, such as reduced energy consumption, provided by myelination likely more than compensate (functionally) for the regressive pruning processes. Instead of reducing 'connectivity' between different parts of the brain suggested by the regressive pruning processes, the myelination process markedly increases the functional potential (connectivity) of the remaining neural networks and circuits. It is thus possible to continue human cognitive, emotional, and behavioral development without the loss, and in fact with the addition, of functional capacity that in middle age culminates in healthy adults with better emotional regulation and inhibitory control underlying what is commonly referred to as wisdom.

Studies examining brain development in children and adolescents have demonstrated increasing WM volume and decreasing gray matter volume in later childhood and adolescence. These findings were recently confirmed in prospective MRI studies of brain development which demonstrated that WM volume increases linearly between ages 4 and 20, a period characterized by increased myelination. Unlike the WM changes, cortical gray matter changes were quadratic rather than linear and exhibited region-specific patterns of change (heterochronous) within this age span. Gray matter reached maximum volume at about age 12 in the frontal lobes but not until age 16 in the temporal lobes. The myelination process continues this regional developmental trajectory, resulting in the same regional pattern of maximal myelination reached 30 years later in the frontal lobe (in the mid-40s) and in the temporal lobe (in the late 40s; Figure 2).

Brain Development Continues during Adulthood until Middle Age

The history of research on age-related changes in adult human brain structure has been immersed in confusion and controversy. Much of the confusion stemmed from the long-standing dogma that in normal adulthood (18–55 years of age), large reductions in cortical neuron numbers occur. This erroneous conclusion was based on data showing that neuronal density in young adults was higher than in middle-aged and older adults. The age-related reductions in cortical neuronal density were subsequently shown to be caused by an age-related tissue fixation shrinkage artifact that caused cortices of younger individuals to shrink more than those of older ones. Together with the assumption that neuronal density was equivalent to number, the dogma took hold that beginning in young adulthood, neuronal numbers decreased with

age. Recent postmortem data based on unbiased stereologic assessments have demonstrated that there is minimal, if any, neuronal cell loss or cortical thickness decrement before the age of 55.

The confusion was also promoted by the underappreciation of the fact that in primate brain, postnatal development is primarily driven by the protracted process of myelination. Unlike other brain cell lines, oligodendrocyte numbers continue to increase into old age to support the process of subcortical and intracortical myelination. The excess cortical shrinkage of younger brains that occurs on fixation is likely due to the lower levels of intracortical myelin at younger ages (Figure 2). Intracortical myelination increases during adulthood to peak in the mid- to late fifth decade of life in association regions such as the frontal and temporal lobes (Figure 2), and the presence of lipid-rich myelin reduces the amount of dehydration-related shrinkage that occurs in older compared to younger brains. The lower the myelin content of any cortex, the more shrinkage would be expected on fixation and the higher the neuronal density (as opposed to numbers) would be.

With the advent of MRI, the dogma was reinforced that in healthy aging, intracortical neuronal numbers and presumably gray matter thickness decline beginning in the teens. This occurred because the vast majority of imaging studies observed an apparent reduction in gray matter volume in adulthood (ages 18–55). The two largely erroneous conclusions (the excess cortical shrinkage of younger brains that occurs with fixation and the imaging findings of an apparent cortical volume loss with age) are likely both due to the process of increasing intracortical myelination. Thus, as Figure 1 demonstrates, the apparent decrease in gray matter volume from ages 18 to 55 (Figure 1(a)) is offset by the increase in myelinated WM volume (Figure 1(b)), to result in a stable brain volume in this age range (Figure 1(c)), as established by postmortem studies. The shift in the gray–white border created by intracortical myelination (Figures 1 and 2) can also help explain why some studies of structure–function relationships in healthy developing individuals report improving function with faster rates of gray matter volume decline. From the framework of the myelin model, such findings can best be interpreted as improved function with increasing intracortical myelination.

With few exceptions, however, most of the early MRI studies reported that WM volume remains constant into the seventh decade, as if it were unaffected by the aging process. This is in contrast to later imaging studies using improved methods to demonstrate continual WM expansion into middle age. The failure of early imaging studies to detect age-related WM volume increase in adulthood may be accounted for by many methodological issues. These include the use of axial images aimed at studying the brain in its entirety rather than focusing on regions such as the frontal and temporal lobes, which complete maturational changes later (in middle age) than the primary sensorimotor and occipital areas (Figures 7(a) and 7(b)). The age-related trajectory of structural changes can also depend on the age range examined because different regions differ in their age-related trajectory of myelination (Figures 2 and 7(a)). Finally, the use of automated procedures for gray and WM segmentation and suboptimal contrast can contribute to misclassification of tissues, and slice thickness greater than 3 mm can increase partial volume effects.

The difference in gray–white contrast obtained with various imaging parameters deserves special focus as it can help clarify discrepant imaging results. This focus can also highlight how consistent imaging and postmortem results can be when appropriate imaging contrasts are employed. When myelin tissue stains or myelin-sensitive brain imaging sequences are used to assess whole brain changes with age, the process of myelination (including intracortical myelination) results in similar quadratic (inverted U) age-related changes in the gray to WM ratio in the adult age range (Figure 3). As outlined above, the process of intracortical myelination can alter both the amount of fixation shrinkage of the cortex and the location of the gray–white border detected on brain imaging studies (Figure 2). Since most imaging sequences are sensitive (in varying degrees) to myelin and water content, they will detect an apparent age-related decrease in cortical volume as the process of age-related intracortical myelination causes the border between gray and WM to 'track' the process of intracortical myelination and be placed higher into the cortex (Figure 2). The slope of the apparent gray matter volume decline will depend on the region being evaluated (e.g., the average of the heterogeneous cortical areas included in the measure – see Figure 7), and the age ranges examined as different regions differ in their age-related pattern of myelination. Thus, many factors can influence imaging findings, but in general, inversion recovery and heavily T1-weighted images are most strongly influenced by the myelin content of the tissues.

Figure 3 demonstrates how myelin-tracking postmortem and *in vivo* imaging of two samples of normal individuals over the life span produce consistent results. Regression lines from the 1980 'gold standard' myelin stain study of AK Miller and colleagues evaluating whole brain gray/WM ratio (Figure 3(b)) estimate a value of approximately 1.3 for the gray/WM ratio of healthy individuals at the age of 20. In MRI measures that were limited to the evaluation of frontal plus temporal regions, a higher ratio would

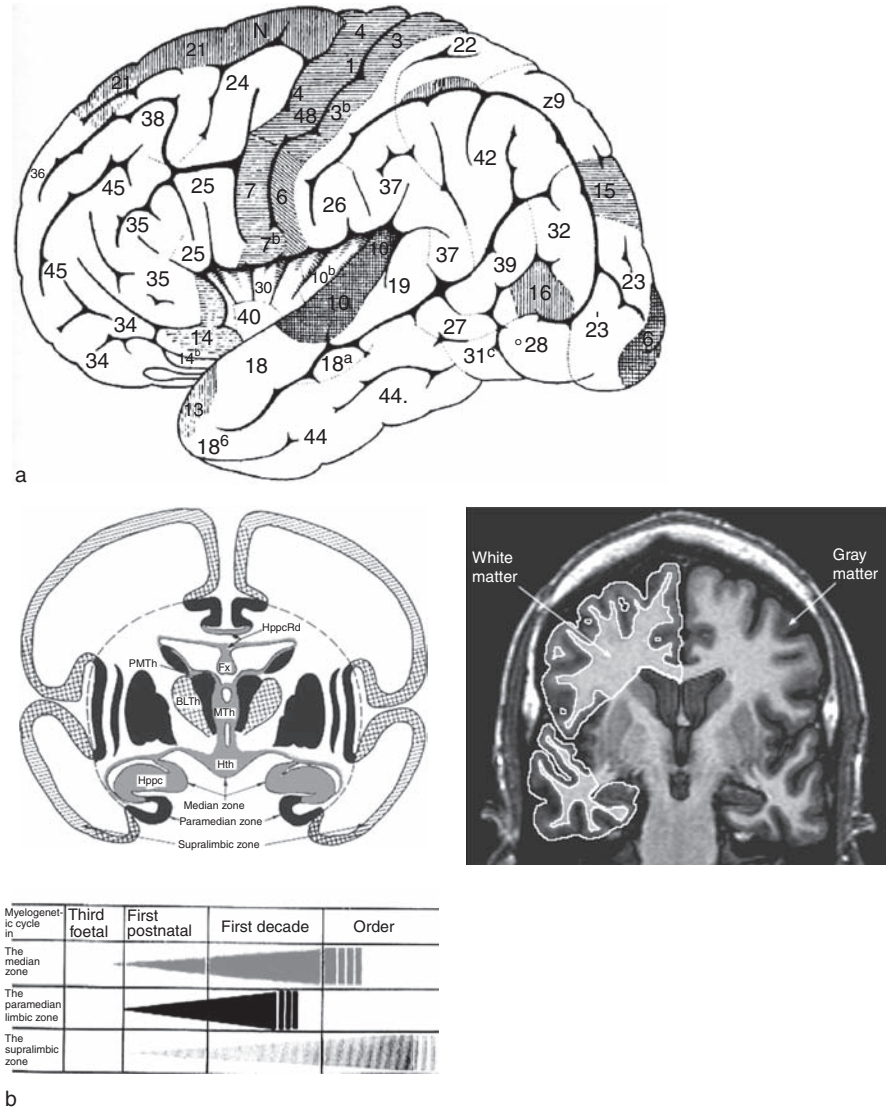

Figure 7 (a) The myelogenetic fields numbered according to the time of myelination. From Flechsig, 1920; reproduced in Meyer A (1981) Paul Flechsig's system of myelogenetic cortical localization in the light of recent research in neuroanatomy and neurophysiology: Part II. *Canadian Journal of Neurological Sciences* 8: 95–104. This is one of the few works attempting to map the sequence of myelination. It illustrates the developmental heterogeneity (heterochronic timing) of brain myelination, with sensory and motor regions (shaded) myelinating first (in very early childhood) and temporal and frontal regions myelinating last – confirmed by Yakovlev PI and Lecours AR (1967) *Regional Development of the Brain in Early Life*. Boston, MA: Blackwell Scientific. (b, left) The image shows three zones of cerebral myelination which differ (are heterochronic) by age at myelination. Medial zone, light gray; paramedial zone, dark; dorsolateral zone, opercular and paralimbic areas – cross hatching; supralimbic association areas – diagonal hatching. (Left) Adapted from Yakovlev PI and Lecours AR (1967) *Regional Development of the Brain in Early Life*. Boston, MA: Blackwell Scientific, with permission from Blackwell Publishing. Showing the 'supralimbic association areas' (diagonal hatching of the frontal and temporal cortices) as these regions continue myelinating into 'older age.' This is graphically depicted in the lower left graph where the cortical supralimbic zones (lowest horizontal graph) continue myelinating well beyond the first decade. On the right, MRI image used to measure myelinated white matter volume in the same frontal and temporal 'supralimbic association areas', demonstrating myelination continues late into the fifth decade (also see **Figure 2**). In the MRI image the outer border of the brain and 'islands' of lighter material seen in the frontal gray matter are cerebrospinal fluid areas that are eliminated from computation using a coregistered T2-weighted image. Inversion recovery MRI reproduced from Bartzokis G, Beckson M, Lu PH, Nuechterlein KH, Edwards N, and Mintz J (2001) Age-related changes in frontal and temporal lobe volumes in men: A magnetic resonance imaging study. *Archives of General Psychiatry* 58: 461–465, copyright © (2001) American Medical Association.

be expected since, as described in myelin stain studies (**Figures 5 and 7**), myelination is continuing in these lobes beyond age 20 although it is complete in many other primary brain regions (e.g., motor, visual). From inversion recovery MRI images, a gray/WM ratio of 1.55 for the frontal plus temporal lobes at age 20 (estimated from regression lines – **Figure 3(a)**) is thus consistent with the gold standard myelin stain post-mortem gray/WM ratio results (**Figure 3(b)**). In addition, the minimum (trough) gray/WM ratio is reached

at age 50 in the postmortem whole brain data and at age 54.6 for *in vivo* data. The trough is again strikingly consistent given that the brain frontal and temporal lobe regions continue to develop in adulthood, which would tend to push the trough to older ages.

Thus, during healthy postadolescent development and maturation of the prefrontal and association areas, apparent gray matter volume reduction observed on imaging studies occurs in concert with an expansion in myelinated WM volume that continues into middle age (**Figure 2**), when peak myelin brain levels are reached. Myelin-sensitive imaging sequences produce data that are strikingly similar to postmortem myelin stain data (**Figures 2 and 3**). The myelinated WM expansion volumetrically cancels out the apparent gray matter reduction; thus, both imaging and postmortem studies show minimal, if any, changes in total brain (gray plus WM) volume in normal individuals during adulthood (18–55 years of age) (**Figure 1(c)**). In older age, the cortex undergoes additional changes consisting primarily of shrinkage of large neurons and increase in the proportion of small neurons. What effect such changes, combined with the increased number of oligodendrocytes and intracortical myelin growth and eventual loss, have on imaging results remains to be fully elucidated.

The biology of brain myelination shows that the life span can be appropriately conceptualized as dynamic, roughly quadratic (inverted U) trajectories that differ (are heterochronic – have different developmental/age trajectories) by brain region (**Figure 2**) and function. The trajectory of intracortical myelination with a shift in the gray/WM separation can help explain the quadratic trajectories in life span WM volumes (**Figure 2**). It can also help explain the complementary, U-shaped gray matter life span trajectories that emerge when people in their ninth and tenth decades are included in aging studies. At advanced ages, the loss of intracortical myelin with a shift of the gray–white border toward the WM could make the cortex appear to be getting 'thicker' or maintaining its volume with age while the overall WM volume continues to decline. This trend will also create a U-shaped curve in a gray–white ratio analysis (**Figure 3**), where the trough represents the point at which loss in myelinated WM volume surpasses the change in gray matter volume produced (at least in part) by the loss of intracortical myelin.

Aging and Myelin Breakdown

In addition to gross volume changes in myelinated WM, the loss of myelin integrity in older age has profound effects on brain function and age-related cognitive decline in human and nonhuman primates. The breakdown in the structural integrity of myelin sheaths can be measured indirectly *in vivo* with MRI using multiple methods. One such method uses transverse relaxation rate (R_2) measures (R_2 is directly associated with transverse relaxation times [T_2] by the formula $R_2 = 1/T_2 \times 1000$). These measures are markedly sensitive to small changes in the amount of tissue water. Ultrastructural electron microscopy studies demonstrate that age-related myelin breakdown results in microvacuolations consisting of splits of myelin sheath layers that create microscopic fluid-filled spaces and increase MRI-'visible' water, thus decreasing R_2. These microvacuolations are ultrastructurally very similar to reversible myelinopathies produced by certain toxins. Animal studies have confirmed that this type of myelin breakdown can be detected with MRI in circumscribed susceptible regions and that the histopathologic changes produced by toxins as well as the recovery process can be tracked by changes in MRI relaxation rates. In asymptomatic individuals, focal WM regions of R_2 decrease (T_2 increase) have also been associated with myelin pallor on histological stains (**Figure 8**). Although R_2 has not been directly correlated with myelin breakdown due to normal aging (as opposed to the reversible toxin-induced myelin breakdown and focal regions described above), in humans and primates, healthy aging is not associated with neuronal loss. On the other hand, the process of myelin breakdown and loss has been repeatedly demonstrated. In this article, the term 'myelin breakdown' will therefore be used to refer to the R_2 findings.

In addition to being heterochronic, the brain's myelination process is highly heterogeneous (**Figures 2 and 7**). For example, the myelination process for primary sensory and motor regions such as the visual system occurs primarily in infancy and consists of large axons myelinated very heavily by a single oligodendrocyte per myelin segment (**Figures 6(b) and 4(a)**). In the late-myelinating association regions such as frontal lobes and corpus callosum regions that connect them (such as the genu – **Figures 6(a) and 4(a)**), the myelination process continues into early middle age (**Figure 4(b)**) and consists of myelination of much smaller fibers by thinner myelin sheaths (**Figure 6(a)**). In these late-myelinating regions, one oligodendrocyte myelinates multiple axons (as many as 50–60 in the latest myelinating regions in the frontal and temporal lobes).

Peters and others have shown in both primate animal models and human postmortem and *in vivo* data that myelination is followed by a degenerative process consisting of progressive myelin breakdown and loss (**Figures 2 and 4**). Published observations from a primate model of aging show that myelin breakdown begins to occur while active myelination

and myelin volume expansion is still continuing. The same sequence has been shown to occur in human aging. Analysis of data for 68 normal men for whom both frontal lobe WM (FWM) volume (Figure 2) and relaxation rate (a measure sensitive to changes in tissue integrity which produce increased tissue water – Figure 4(b) were available showed that, as would be expected, the beginning of WM integrity decline in Fwm (at age 38) (Figure 4(b)) precedes the age at which Fwm volume decreases begin (at age 45; Figure 2).

These imaging data are also consistent with postmortem data from human brain that show a 45% loss of myelinated fiber length during aging, affecting primarily thinner and less heavily myelinated fibers such as the ones seen in the frontal lobe WM and genu of the corpus callosum (Figure 6(a)). These declines are also consistent with age-related ferritin iron declines in these WM regions. It is postulated that age-related myelin breakdown and oligodendrocyte loss (Figure 8) release the considerable iron stores of oligodendrocytes and their myelin, which are then redistributed and contribute to the age-related increase in brain iron levels observed in heavily-myelinated subcortical gray matter regions.

Functional Implications of Age-Related Myelin Breakdown

Neuroimaging and postmortem data have demonstrated that the process of myelination peaks in middle age, followed by myelin breakdown and loss (Figures 2, 3, and 6(b)). Age-related slowing in speed of information processing, particularly after middle age, has been repeatedly documented in the neuropsychological literature and is a well-accepted clinical phenomenon. Salthouse and others have demonstrated that the slowing in cognitive processing speed (CPS) with age likely underlies the age-related decline observed in most, if not, all human cognitive functions.

Investigations of CPS performance across the life span reveal a trajectory that parallels myelination and subsequent myelin breakdown. The similarity in these trajectories supports the hypothesis that in older age, myelin breakdown may be the cellular mechanism underlying the age-related decline in cognitive speed. Thus, the uniquely extensive myelination of the human brain makes myelin maintenance and repair especially critical for sustaining our high CPS. Myelin breakdown may result in a progressive 'disconnection' of widely distributed neural networks that eventually contributes to age-related cognitive decline and Alzheimer's disease (AD).

Apolipoprotein E (ApoE) 4 genotype, coded on chromosome 19, is second only to advancing age in its influence as a risk factor for developing AD. Epidemiological studies have demonstrated that ApoE4 genotype shifts the age of onset for AD by more than a decade, thus accounting for the vast majority of observed cases of AD. Conversely, the onset age is delayed in carriers of the ApoE2 allele.

The physiologic role of ApoE suggests that myelin may be a central biological link between the presence of this gene and AD. In nonhuman primates, myelin

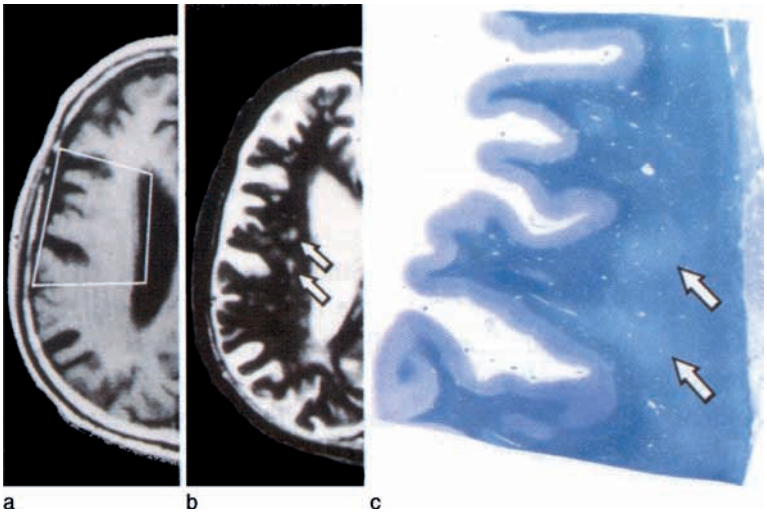

Figure 8 White matter 'hyperintensities' on MRI are often the center of much larger areas of diffuse myelin breakdown. Small patchy hyperintense white matter lesions on T2-weighted MRI (arrows, (b)) are often the very center of much more widespread and diffuse regions of myelin breakdown, as shown in the postmortem myelin stain of (c) – see arrows. The panels are different MRI images through the same slice of tissue; (a) is a T1-weighted image, (b) is a T2-weighted image, and (c) is a histologic section (luxol fast blue staining for myelin) through the same slice of the same brain (showing the boxed region in left panel). Adapted from Takao M, Koto A, Tanahashi N, Fukuuchi Y, Takagi M, and Morinaga S (1999) Pathologic findings of silent hyperintense white matter lesions on MRI. *Journal of the Neurological Sciences* 167: 127–131, with permission from Elsevier.

repair and remyelination is an essential process that continues into old age. ApoE is the primary transporter of endogenously produced lipids such as cholesterol and sulfatide, entirely produced by oligodendrocytes and essential for myelin production and function. ApoE coordinates mobilization and transport of such lipids for uptake and use in repair, growth, and maintenance of myelin as well as other membranes, including synapses and dendrites essential for brain plasticity and learning. ApoE is essential to the extensive process of 'recycling' lipids such as cholesterol released when damaged myelin is degraded and supplying these lipids for rapid membrane repair and biogenesis. This process of 'remyelination' has been demonstrated in the primate model. The quantity of ApoE molecules necessary for this recycling process is lowest in ApoE4 carriers and highest in individuals with the ApoE2 allele. Thus, it has been hypothesized and imaging data have supported that the disruption of myelin repair secondary to insufficient ApoE molecules results in greater age-related myelin breakdown and contributes to the association between ApoE4 and AD as well as the associated influence of ApoE genotype on age-related cognitive decline.

Studies testing the myelin model have demonstrated that healthy older individuals possessing the ApoE4 genotype displayed more myelin breakdown with age. This relationship was absent for participants with the ApoE2 genotype, and ApoE3 individuals displayed an intermediate slope of decline. This genetic effect was specific to late myelinating regions and was not observed in early myelinating regions such as the splenium of the corpus callosum (**Figures 6 and 4**). Furthermore, CPS performance (as assessed by Trails A and Digit Symbol neurocognitive tasks) was significantly associated with R_2 in late myelinating regions (frontal lobe WM, $p < 0.0001$, and genu, $p < 0.004$) but not in the early myelinating splenium region ($p > 0.2$). These data suggest that myelin breakdown in healthy older individuals underlies the age-related cognitive decline that ultimately progresses to AD.

Conclusions

Postmortem and *in vivo* imaging evidence can be integrated to suggest that an understanding of brain aging and the aging-related processes of myelin production and subsequent breakdown may be directly relevant to deepening our understanding of the biology and function of the human brain. The lifelong quadratic trajectories of myelination and eventual myelin breakdown and loss of different networks, circuits, and regions can help explain the lifelong trajectories of cognitive, behavioral, and emotional changes that define our everyday lives, from the impulsiveness of youth to the wisdom of older ages. More important, the focus on the myelination trajectories of the human brain may help explain many of the unique but highly prevalent and devastating diseases that plague our species over its entire life span.

This myelin model of the human brain posits that the uniquely pervasive and protracted myelination process creates a continuum of vulnerability among the myelin-producing oligodendrocytes and their myelin that culminates in the latest developing oligodendrocytes. Genetic and/or environmental effects that affect myelin development and breakdown will manifest as risk factors (or protective factors) for both developmental disorders such as autism and schizophrenia and aging-related disorders such as AD and Parkinson's disease.

The creation and maintenance of myelin is at the core of the earliest changes involved in both postnatal brain development and, in older age, the aging process itself and the subsequent emergence of degenerative disorders such as AD. Myelin deficits disrupt brain functions that depend on highly synchronized timing of high-frequency bursts of action potentials and eventually result in functional 'disconnections' of association cortical regions. The eventual progression to frank neuronal loss results in permanent deficits. This model suggests that many pathological states (e.g., genetic and hormonal states, head trauma, hypertension, hypercholesterolemia, and substance abuse) can affect the normal age-related pattern of myelin development, resulting in developmental disorders that at older ages can also affect the pattern of myelin breakdown and the manifestations of degenerative disorders. This life-span model of a seamless transition from brain development and its dynamic interaction with degenerative processes can be directly tested using prospective *in vivo* imaging studies.

The focus on myelination provides a conceptual framework for considering and developing novel myelin-centered treatments with wide spectra of efficacy that likely will extend to many disorders that share derangements in myelin health. The model suggests that the current focus on neuronal neurotransmitter 'imbalances' that much of neuropsychopharmacology is attempting to correct may be too narrow. It suggests that expanding the research focus to include interventions that affect vulnerable structural developmental processes and specifically the process of myelination may provide opportunities for novel and powerful interventions. Interventions that affect myelination may be effective in strengthening inhibitory controls of cognition in developmental disorders of younger ages as well as providing preventive opportunities for age-related degenerative disorders. As briefly outlined

above, the tools (imaging, genetic, molecular, clinical, etc.) to perform such medication development work directly in humans are either already available or being developed at a rapid pace. The model suggests that interceding early in dysregulated developmental and accelerated degenerative trajectories may increase the effectiveness of treatments, decrease the need for more aggressive interventions later, and thus be accomplished with reduced risk to patients.

See also: Aging of the Brain; Aging: Extracellular Space; Alzheimer's Disease: MRI Studies; Amyloid: Vascular and Parenchymal; Cognition in Aging and Age-Related Disease; Demyelinating Diseases; Functional Neuroimaging Studies of Aging; Lipids and Membranes in Brain Aging.

Further Reading

Bartzokis G (2004) Age-related myelin breakdown: A developmental model of cognitive decline and Alzheimer's disease. *Neurobiology of Aging* 25: 5–18.

Bartzokis G (2004) Quadratic trajectories of brain myelin content: Unifying construct for neuropsychiatric disorders. *Neurobiology of Aging* 25: 49–62.

Bartzokis G, Beckson M, Lu PH, Nuechterlein KH, Edwards N, and Mintz J (2001) Age-related changes in frontal and temporal lobe volumes in men: A magnetic resonance imaging study. *Archives of General Psychiatry* 58: 461–465.

Bartzokis G, Minztz J, Marx P, et al. (1993) Reliability of *in vivo* volume measures of hippocampus and other brain structures using MRI. *Magnetic Resonance Imaging* 11: 993–1006.

Bartzokis G, Sultzer D, Lu PH, Nuechterlein KH, Mintz J, and Cummings J (2004) Heterogeneous age-related breakdown of white matter structural integrity: Implications for cortical 'Disconnection' in aging and Alzheimer's disease. *Neurobiology of Aging* 25: 843–851.

Connor JR and Menzies SL (1996) Relationship of iron to oligodendrocytes and myelination. *Glia* 17: 83–93.

Giedd JN, Blumenthal J, Jeffries NO, et al. (1999) Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience* 2: 861–863.

Jernigan TL and Gamst AC (2005) Changes in volume with age-consistency and interpretation of observed effects. *Neurobiology of Aging* 26: 1271–1274.

Kemper T (1994) Neuroanatomical and neuropathological changes during aging and dementia. In: Albert M and Knoefel J (eds.) *Clinical Neurology of Aging,* 2nd edn., pp. 3–67. New York: Oxford University Press.

Kretschmann H-J, Tafesse U, and Herrmann A (1982) Different volume changes of cerebral cortex and white matter during biological histological preparation. *Microscopica Acta* 86: 13–24.

Lintl P and Braak H (1983) Loss of intracortical myelinated fibers: A distinctive age-related alteration in the human striate area. *Acta Neuropathologica* 61: 178–182.

Marner L, Nyengaard JR, Tang Y, and Pakkenberg B (2003) Marked loss of myelinated nerve fibers in the human brain with age. *Journal of Comparative Neurology* 462: 144–152.

Miller AK, Alston RL, and Corsellis JA (1980) Variation with age in the volumes of grey and white matter in the cerebral hemispheres of man: Measurements with an image analyser. *Neuropathology and Applied Neurobiology* 6: 119–132.

Oldendorf WH and Oldendorf W Jr. (1988) *Basics of Magnetic Resonance Imaging*. Boston: Martinus Nijhof.

Peters A, Morrison JH, Rosene DL, and Hyman BT (1998) Are neurons lost from the primate cerebral cortex during normal aging? *Cerebral Cortex* 8: 295–300.

Peters A and Sethares C (2002) Aging and the myelinated fibers in prefrontal cortex and corpus callosum of the monkey. *Journal of Comparative Neurology* 442: 277–291.

Salthouse TA (2000) Aging and measures of processing speed. *Biological Psychology* 54: 35–54.

Schaie KW (2005) What can we learn from longitudinal studies of adult development? *Research in Human Development* 2: 133–158.

Takao M, Koto A, Tanahashi N, Fukuuchi Y, Takagi M, and Morinaga S (1999) Pathologic findings of silent hyperintense white matter lesions on MRI. *Journal of the Neurological Sciences* 167: 127–131.

Yakovlev PI and Lecours AR (1967) *Regional Development of the Brain in Early Life*. Boston, MA: Blackwell Scientific.

Brain–Computer Interface

J R Wolpaw, New York State Department of Health and State University of New York, Albany, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the nearly 80 years since Hans Berger first recorded electroencephalographic activity from the scalp, the electroencephalograph (EEG) has been used primarily for clinical diagnosis, for exploring brain function, and to a very limited extent for therapy. At the same time, many people have speculated that the EEG or other reflections of brain activity might be useful for another purpose as well: to serve as an alternative method for the brain to send messages and commands to the outside world. While the brain's normal communication and control capabilities depend on nerves and muscles, the existence of easily recordable brain signals, such as the EEG implied the possibility of establishing nonmuscular communication and control based on brain–computer interfaces (BCIs).

In spite of recurring scientific and popular interest in the idea of BCIs, and despite a few encouraging initial efforts, it is only in the past 20 years that sustained research has begun, and only in the past 12 years that a recognizable field of BCI research and development has emerged. The field is now populated by a large and rapidly growing number of research groups throughout the world. This new surge of interest and activity is due largely to the combination of four important elements.

The first element is greater appreciation of the needs and the abilities of people paralyzed by disorders such as cerebral palsy, spinal cord injury, brain stem stroke, amyotrophic lateral sclerosis (ALS), and muscular dystrophies. Life-support technology (e.g., home ventilators) now enables even the most severely disabled people to survive for many years. Furthermore, it is now clear that even people who have little or no voluntary muscle control, who may be nearly 'locked-in' their bodies, unable to communicate in any way, can have lives that they consider enjoyable and productive if they can be given even the most basic means of communication and control.

The second element is the greater knowledge of the nature and functional correlates of the EEG and other measures of brain activity that has come from animal and human research. Along with this new understanding have come improved methods for recording these signals, in both the short-term and the long-term. This increased knowledge and improved technology are directing and enabling more sophisticated and productive BCI research.

The third element is the easy availability of powerful, inexpensive computer hardware that supports the complex real-time analyses of brain activity important for successful BCI operation. Until quite recently, much of the online signal processing used in current BCIs was impossible or extremely expensive.

The fourth element contributing to the rapid growth of BCI research is new appreciation of the brain's remarkable adaptive capacities, both in normal life and in reaction to trauma or disease. The growing recognition of these adaptive capacities generates enormous excitement and interest in the potential for using them to create novel interactions between the brain and computer-based devices, interactions that can substitute for or even augment the brain's normal neuromuscular interactions with its external and internal environments.

Definition of a BCI

A BCI creates a nonmuscular output channel for the brain. Instead of being executed through peripheral nerves and muscles, the user's wishes are conveyed by brain signals (such as those detected by an EEG), and these brain signals do not depend for their generation on neuromuscular activity. (Thus, for example, a device that uses visual-evoked potentials in the EEG to determine eye-gaze direction is not a true BCI because it relies on muscular control of eye position and simply uses the EEG as a measure of that position.)

Like other communication and control systems, a BCI establishes a real-time interaction between the user and the outside world. The user gets feedback on the results of the BCI's operation, and that feedback influences the user's intent and the brain signals that encode that intent. For example, if a person uses a BCI to control the movements of a robotic arm, the arm's position after each movement affects the person's intent for the succeeding movement and the brain signals that convey that intent. Thus, a system that only records and analyzes brain signals, without providing the outcome of that analysis to the user in a real-time interactive fashion, is not a BCI. Figure 1 shows the design and operation of any BCI.

Popular speculation and some scientific efforts have been based on the belief that BCIs are 'mind-reading' or 'wire-tapping' technology, devices that listen in on the brain, detect its intent, and then execute that intent. This is a misconception that ignores a central feature of the brain's interactions with the external world. The motor skills that accomplish a person's intent, whether it be to walk across a room, speak specific words, or play a particular piece on the

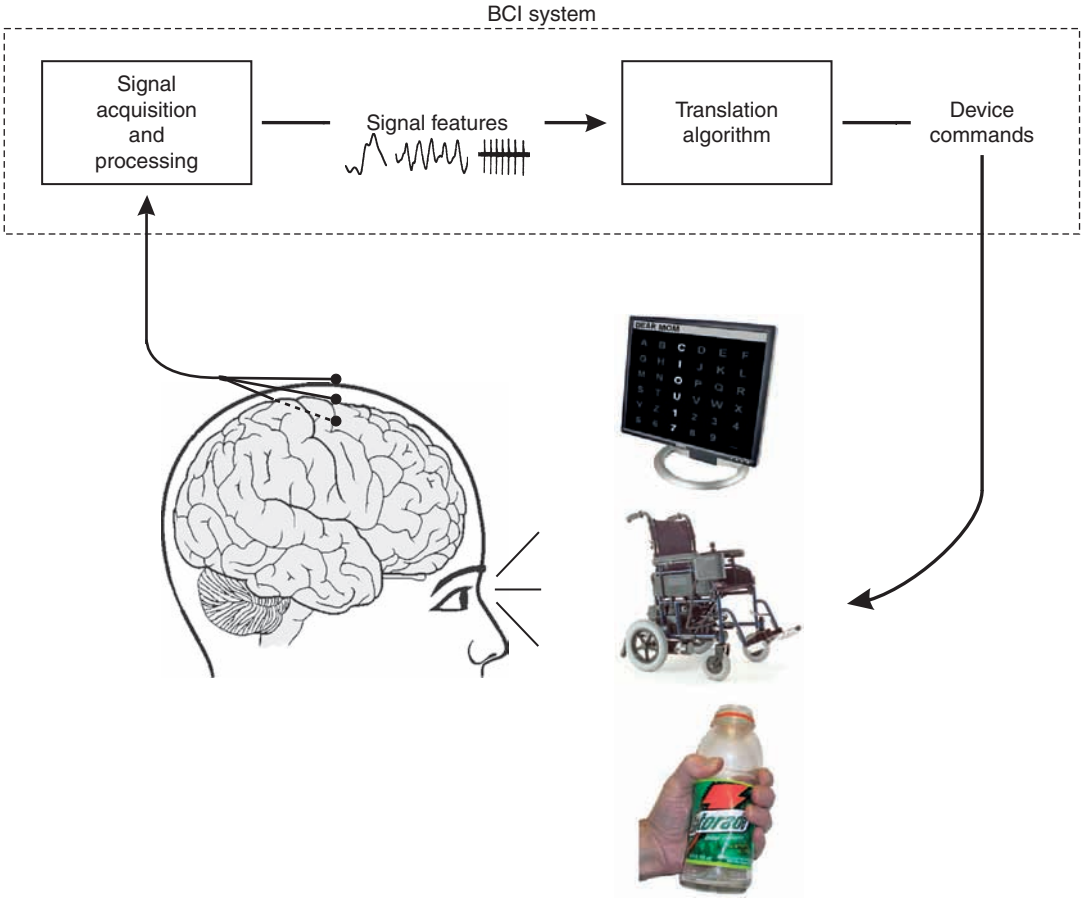

Figure 1 The basic design of any brain–computer interface (BCI) system. Signals reflecting brain activity are acquired from the scalp, from the cortical surface, or from within the brain and are analyzed to measure signal features (such as amplitudes of evoked potentials or electroencephalograph rhythms or firing rates of single neurons) that reflect the user's intent. These features are translated into commands that operate a device such as a word-processing program, a wheelchair, or a neuroprosthesis. Adapted from Wolpaw JR and Birbaumer N (2006) Brain–computer interfaces for communication and control. In: Selzer ME, Clarke S, Cohen LG, Duncan P, and Gage FH (eds.) *Textbook of Neural Repair and Rehabilitation; Neural Repair and Plasticity,* pp. 602–614. Cambridge: Cambridge University Press, with permission from Cambridge University Press.

piano, are mastered and maintained by initial and continuing 'adaptive changes' in brain function. In early development and throughout later life, central nervous system (CNS) neurons and synapses change continually to master new skills and to maintain those already mastered. This adaptive plasticity is responsible for basic skills such as walking and talking and for more specialized skills such as ballet, and it is guided by the results produced. Thus, as muscle strength, limb length, and body weight change with growth and aging, the CNS modifies its outputs so as to maintain its motor skills.

This requirement for initial and continuing adaptation exists whether the person's intent is carried out naturally, that is, by peripheral nerves and muscles, or through an artificial interface, a BCI, that uses brain signals instead of nerves and muscles. Successful BCI operation requires the effective interaction of two adaptive controllers: the user, who must generate brain signals that encode intent; and the BCI system, that must translate these signals into commands that accomplish the user's intent. Thus, BCI usage is essentially a skill that user and system work together to acquire and maintain. The user encodes intent in signal features that the BCI system can measure, and the system measures these features and translates them into device commands. This initial and continuing dependence on the mutual adaptation of user to system and system to user is a fundamental principle of BCI operation; its effective management is one of the principal challenges for BCI research and development.

Brain Signals That Could Be Used in a BCI

A variety of different methodologies can detect brain activity. These include methods for recording electrical or magnetic fields, functional magnetic resonance

imaging (fMRI), positron emission tomography (PET), and functional near-infrared (fNIR) imaging. However, magnetoencephalography, fMRI, and PET are not currently suited for everyday use due to their intricate technical demands, high expense, and/or limited real-time capabilities. Only electrical field recording (and possibly fNIR imaging) are likely to be of value for practical applications in the near future.

The electrical fields produced by brain activity can be recorded at the scalp (electroencephalographic activity (EEG)), at the cortical surface (electrocorticographic activity (ECoG)), or within the brain (local field potentials (LFPs)) or neuronal action potentials (spikes). Figure 2 shows these three recording alternatives. Each method has advantages and disadvantages. EEG recording is easy and noninvasive, but it has limited topographical resolution and frequency range and can be contaminated by electromyographic (EMG) activity from cranial muscles or electrooculographic (EOG) activity. ECoG has better topographical resolution and frequency range, but it requires implantation of electrode arrays on the cortical surface, which has been done as yet for only brief periods (a few days or weeks) in people. Intracortical recording, or recording within other brain areas, yields signals with the highest resolution, but it requires insertion of multielectrode arrays within brain tissue, and it faces as yet unresolved problems in reducing tissue damage and scarring and in achieving long-term recording stability.

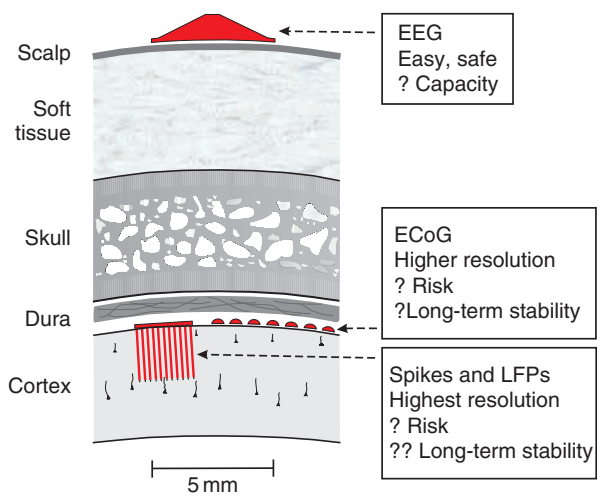

Figure 2 Recording sites for electrophysiological signals used by brain–computer interface (BCI) systems. Electroencephalographic activity (EEG) is recorded by electrodes on the scalp. Electrocorticographic activity (ECoG) is recorded by electrodes on the cortical surface. Neuronal action potentials (spikes) or local field potentials (LFPs) are recorded by electrode arrays inserted into the cortex. Each recording method has advantages and disadvantages. Adapted from Wolpaw JR and Birbaumer N (2006) Brain–computer interfaces for communication and control. In: Selzer ME, Clarke S, Cohen LG, Duncan P, and Gage FH (eds.) *Textbook of Neural Repair and Rehabilitation; Neural Repair and Plasticity*, pp. 602–614. Cambridge, UK: Cambridge University Press, with permission from Cambridge University Press.

The ultimate usefulness of each of these methods will hinge on the range of communication and control applications it can support and on the extent to which its disadvantages can be overcome. The question of the relative value of noninvasive methods (i.e., EEG), moderately invasive methods (e.g., ECoG), and more-invasive methods (e.g., intracortical recording) remains unanswered. It is conceivable that practical, stable, and safe techniques for long-term recording within the brain will prove relatively easy to develop. On the other hand, the information transfer rates possible with intracortical methods may turn out to be no greater than those achievable with less-invasive methods (e.g., ECoG, or even EEG). It is quite likely that different recording methods will prove useful for different applications and for different individuals. Comprehensive evaluations of the characteristics and capacities of each recording method are needed.

Present-Day BCIs

Human BCI experience to date consists largely of noninvasive EEG-based research. A few short-term ECoG studies have been reported, and limited data are available from a few people with intracortical implants. Most intracortical BCI data come from animals, primarily monkeys. EEG-based BCI methods can certainly support simple applications and seem to be able to support more complex ones. Invasive methods may be able to support more complex applications, but the issues of risk and long-term performance are not yet resolved.

Three different kinds of EEG-based BCIs have been evaluated in people. They are distinguished by the particular EEG features they use to derive the user's intent. Figure 3(a) illustrates a P300-based BCI. It focuses on the P300 component of the event-related brain potential, which appears in the EEG over central areas about 300 ms after a salient, or attended stimulus. In most P300-based BCIs described to date the stimulus is visual. Typically, letters, numbers, and/or other possible choices are arranged in a matrix, and the rows and columns of the matrix flash in rapid succession. Only the row and column that contain the item that the user wants to select produce P300 potentials. By detecting these P300 potentials, the BCI system can determine the user's selection. This BCI method is able to support operation of a simple word-processing program that enables users to

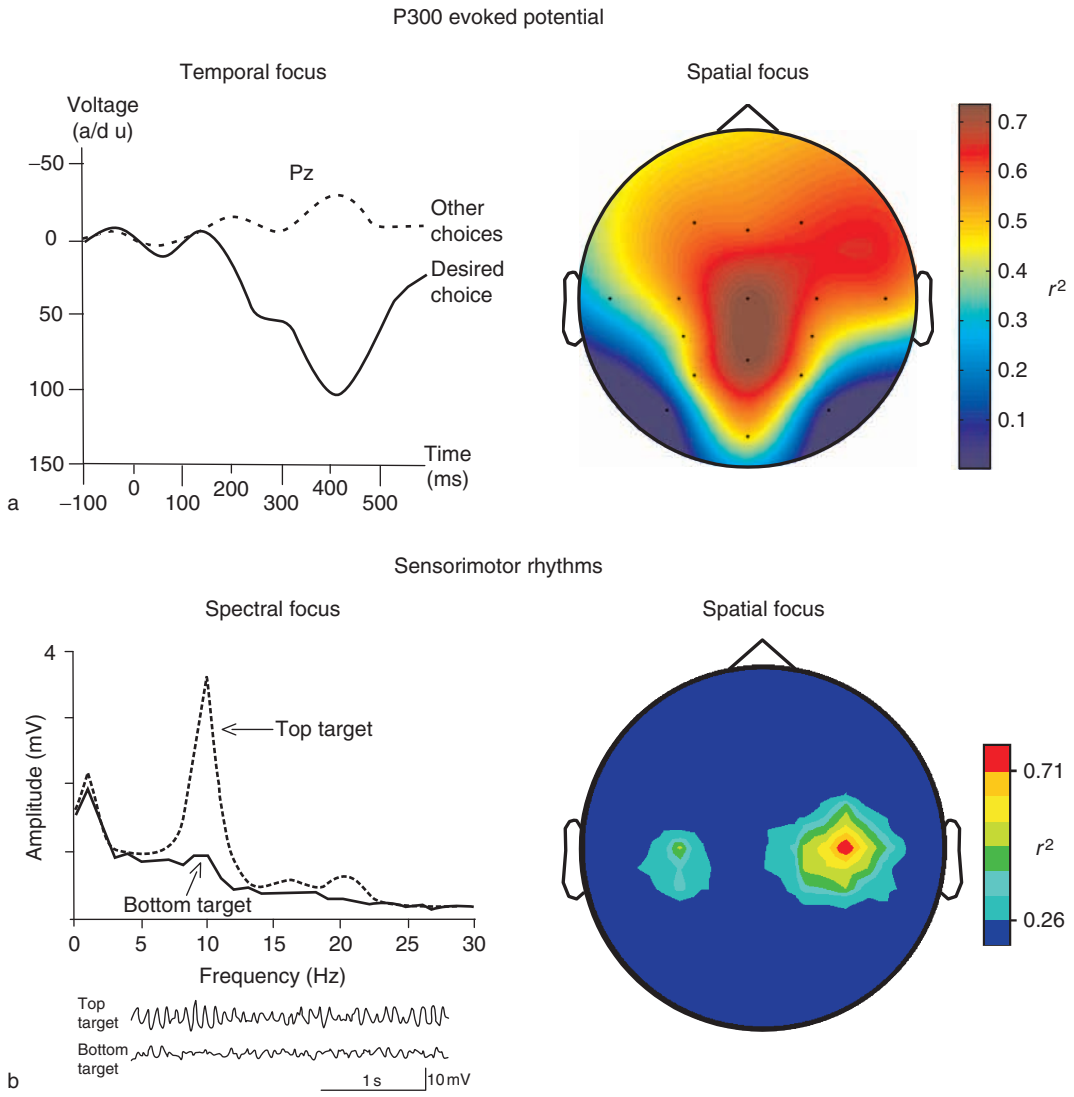

Figure 3 Noninvasive EEG-based BCI methods that use EEG recorded from the scalp. (a) P300 evoked potential brain–computer interface (BCI). A matrix of possible selections is shown on a screen. Scalp electroencephalographic activity (EEG) is recorded over centroparietal cortex (EEG recording location P_Z) while these selections flash in succession. Only the selection desired by the user evokes a large P300 potential (i.e., a positive potential about 300 ms after the flash). r^2 is the coefficient of variation; a/d u are the analog to digital conversion units. (b) Sensorimotor rhythm BCI. Scalp EEG is recorded over sensorimotor cortex. Users control the amplitudes of one or more 8–12 Hz mu rhythms or 18–26 Hz beta rhythms to move a cursor to a desired target located somewhere on a computer screen. Frequency spectra (top) for top and bottom targets indicate that this user's control of vertical cursor movement is clearly focused in the mu-rhythm frequency band. Sample EEG traces (bottom) also show that the mu rhythm is prominent with the top target and minimal with the bottom target. Trained users can also control movement in two dimensions. Adapted from Kübler A, Kotchoubey B, Kaiser J, Wolpaw JR, and Birbaumer N (2001) Brain–computer communication: Unlocking the locked-in. *Psychological Bulletin* 127: 358–375, with permission from American Psychological Association.

communicate at rates up to several words per minute. Improvements in signal analysis may substantially increase its capacities.

Figure 3(b) illustrates a BCI based on sensorimotor rhythms. Sensorimotor rhythms are 8–12 Hz (mu) and 18–26 Hz (beta) oscillations in the EEG recorded over sensorimotor cortices. Normally, changes in mu and/or beta rhythm amplitudes accompany movement and sensation, and motor imagery as well. BCI studies show that people can learn to control mu or beta rhythm amplitudes in the absence of any movement or sensation and that they can use this control to move a cursor to select letters or icons on a screen or to operate a simple orthotic device. Both one- and two-dimensional control are achievable. Sensorimotor rhythm-based BCIs, like P300-based BCIs, can support basic word-processing or other simple functions. They might also support multidimensional control of the movements of a neuroprosthesis or a device such as a robotic arm.

A BCI can also use slow cortical potentials (SCPs) in the EEG, which last from 300 ms to several seconds. In normal brain function, negative SCPs accompany preparatory depolarization of the underlying cortical network, and positive SCPs are thought to reflect cortical disfacilitation or inhibition. With appropriate training, people can learn to control SCPs to produce positive or negative shifts. With this control, they can perform basic word-processing and other simple control tasks such as accessing the Internet.

Current BCIs depend mainly on visual stimuli and visual feedback. However, people who are severely disabled may lack the vision or eye movements needed to perceive visual stimuli, especially if the stimuli change rapidly. Thus, BCI systems that employ auditory rather than visual stimuli would also be valuable – and are under investigation.

Figure 4(a) illustrates a BCI that uses sensorimotor rhythms in ECoG recorded by electrode arrays on the cortical surface. ECoG signals are much higher in amplitude than are scalp-recorded EEG signals, they have much higher spatial and temporal resolution, and they are much less susceptible to contamination by EMG or EOG. ECoG encompasses not only mu and beta rhythms but also higher-frequency gamma (>30 Hz) rhythms, which are very small or entirely lacking in EEG. With appropriate interelectrode spacing, ECoG can resolve activity limited to a few square millimeters of cortical surface. To date, ECoG studies have been confined to short-term experiments in people temporarily implanted with electrode arrays prior to epilepsy surgery. These studies reveal sharply focused ECoG activity associated with movement and sensation and with motor imagery. Furthermore, with only a few minutes of training, people can learn to use motor imagery to control cursor movement.

The speed of this learning, which appears to be faster than that usually found with sensorimotor rhythms in scalp-recorded EEG, together with ECoG's high topographical resolution, wide spectral range, and freedom from contamination, suggests that ECoG-based BCIs might provide communication and control superior to that possible with EEG-based BCIs. Widespread clinical use of ECoG-based BCIs will require development of fully implanted systems (i.e., systems that use telemetry and thus do not have wires passing through the skin) and clear evidence that they provide safe and stable recording for years.

Figure 4(b) shows a multielectrode array for intracortical recording and the locations of its implantation in human motor cortex. Intracortical BCI studies conducted in monkeys and to a limited extent in humans have shown that single-neuron activity recorded by such arrays can be used to control movement of a cursor in one, two, or even three dimensions. It appears that LFPs, which can be recorded by the same electrode arrays and reflect nearby synaptic and neuronal activity, might provide similar control. In these intracortical single-neuron and LFP studies, the usual strategy has been to define the neuronal activity associated with standardized limb movements, then to apply this activity to simultaneously control comparable cursor movements, and finally to demonstrate that the neuronal activity alone can continue to control cursor movements in the absence of actual limb movements. As Figure 4(b) shows, the relationships between neuronal activity and cursor movements change over time. Ideally, neuronal activity adapts over sessions so as to improve cursor control. This adaptation, like the adaptations seen with EEG- and ECoG-based BCIs, illustrates the need for initial and continuing adaptation of system to user and user to system.

The major questions that must be answered prior to clinical use of intracortical BCIs include their long-term safety, the stability and persistence of their signals in the face of cortical tissue reactions to the implanted electrodes, the long-term usefulness of these signals, and to what extent their capabilities in actual practical applications (e.g., in neuroprosthesis control) significantly exceed those of less-invasive BCIs. The first two websites listed provide videos that illustrate that a noninvasive EEG-based BCI using sensorimotor rhythms can provide cursor control comparable in speed and accuracy to that achieved to date with intracortical methods.

Signal Processing

A BCI records brain signals and analyzes them to derive device commands. This signal processing has two parts. The first part is feature extraction, the measurement of those features of the signals that encode the user's intent. These features can be relatively simple measures such as the amplitudes or latencies of particular evoked potentials (e.g., P300), the amplitudes or frequencies of particular rhythms (e.g., sensorimotor rhythms), or the firing rates of individual cortical neurons, or they can be more complex measures such as spectral coherences or weighted combinations of simple measures. To provide effective BCI performance, the feature-extraction part of signal processing needs to focus on features that actually do encode the user's intent, and it needs to extract those features accurately.

The second part of BCI signal processing is a translation algorithm that translates these features into device commands. Features such as rhythm

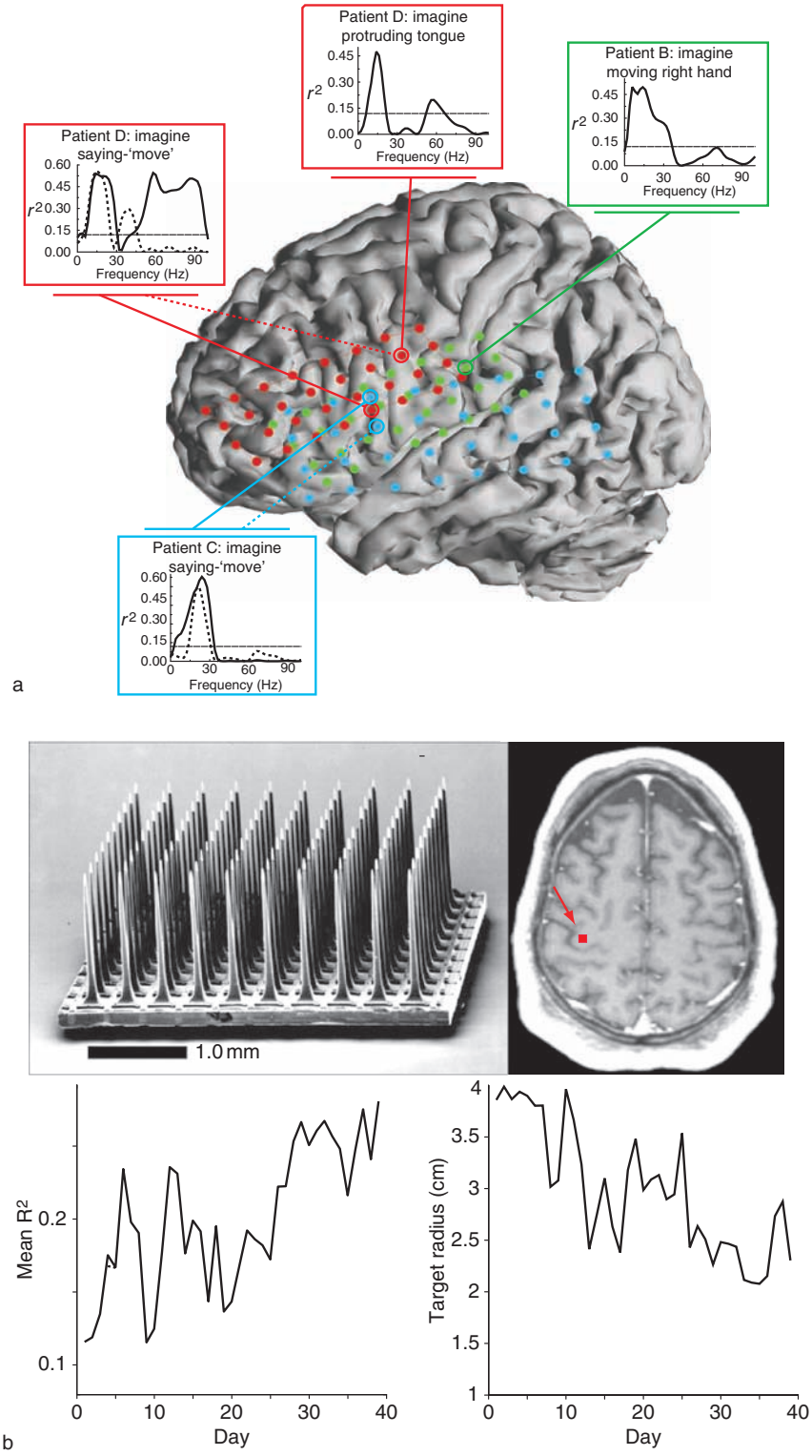

Figure 4 Invasive brain–computer interface (BCI) methods. (a) Human electrocorticographic (ECoG) control of vertical cursor movement using specific motor imagery to move the cursor up and using rest (i.e., no imagery) to move it down. The electrodes used for online control are circled, and the spectral correlations of their ECoG activity with target location (i.e., top or bottom of screen) are displayed. Electrode arrays for Patients B, C, and D are green, blue, and red, respectively. The specific imagined actions used are indicated. The substantial levels of control achieved with different kinds of imagery are apparent. (The dashed lines indicate significance at the 0.01 level.) For Patients C and D, the solid and dotted r^2 spectra correspond to the sites indicated by the dotted and solid line locators, respectively. Reproduced from Leuthardt EC, Schalk G, Wolpaw JR, Ojemann JG, and Moran DW (2004) A brain–computer interface using electrocorticographic signals in humans. *Journal of Neural Engineering* 1: 63–71, with permission from IOP Publishing Limited. (b) Top left: Array

amplitudes or neuronal firing rates are translated into commands that specify outputs such as cursor movements, icon selection, or prosthesis operation. Translation algorithms can be simple (e.g., linear equations) or complex (e.g., neural networks, support vector machines).

A successful translation algorithm ensures that the user's range of control of the chosen features supports selection of the full range of device commands. Suppose, for example, that the feature is the amplitude of a 21–24 Hz beta rhythm in the EEG over left sensorimotor cortex, that the user can vary this feature over a range of 1–5 μV, and that the application is horizontal cursor movement. In this case, the translation algorithm must ensure that the 1–5 μV range permits the user to move the cursor to both the right and left edges of the screen and at a rate consistent with the rapidity and maximum duration of the user's beta control. In addition, the algorithm must accommodate spontaneous variations in the user's range of control (i.e., variations due to diurnal change, fatigue, or other factors). Finally, the translation algorithm should be able to adapt to at least accommodate, and at best encourage, improvements in the user's control. Thus, if the user's range of control improves from 1–5 to 1–8 μV, the translation algorithm should take advantage of this improvement in order to increase the speed and/or precision of cursor control.

The need for ongoing adaptation of the translation algorithm to accommodate spontaneous and other changes in the signal features reflects the continuing importance of system–user and user–system adaptation and has important implications. First, it means that new algorithms cannot be evaluated adequately by off-line analyses alone. They must also be evaluated online, so that the effects of their adaptive interactions with the user can be determined. This online evaluation should be long-term as well as short-term since important adaptive interactions often develop gradually. Second, the need for ongoing adaptation means that simpler algorithms, for which adaptation is typically easier and more effective, have an inherent advantage. Simple algorithms (e.g., linear equations) should be replaced by more complex alternatives (e.g., neural networks) only after online as well as off-line evaluations demonstrate that the more complex alternatives give superior long-term performance without continual and laborious recalibration procedures.

Potential Users

In their current early stage of development, BCIs are likely to be of significant practical value primarily for people with the most severe neuromuscular disabilities, people for whom conventional assistive communication technologies, all of which require some measure of consistent voluntary muscle control, are not satisfactory options. These include people with ALS who decide to accept artificial ventilation (rather than to die) as their disease advances, children and adults with severe cerebral palsy who do not have useful muscle control, patients with brain stem strokes who have only minimal eye movement control, individuals with severe muscular dystrophies or peripheral neuropathies, and possibly people with acute disorders associated with extensive paralysis (such as Landry-Guillain–Barré syndrome). People with slightly less severe disabilities, such as those with high-cervical spinal cord injuries, may also find BCI technology preferable to conventional assistive communication methods that co-opt their remaining voluntary muscle control (e.g., methods that depend on gaze direction or EMG of facial muscles). The extent to which future BCIs prove useful to those with much less-severe motor disabilities will depend on the speed and precision of the control the BCI systems can provide and on the reliability and convenience of their use.

People with disabilities of different kinds may differ in the BCI methods that are most useful to them. For some, the CNS deficits responsible for their disability may impair their control of certain brain signals and not others. For example, the motor cortex damage that can accompany ALS or the subcortical damage of severe cerebral palsy could conceivably impair generation or control of sensorimotor rhythms or single neuron activity. In these individuals, other brain signals (e.g., P300 potentials or neuronal activity from other brain areas) might be good alternatives.

of 100 microelectrodes chronically implanted in human motor cortex to record neuronal action potentials and local field potentials to control a cursor or other device. Top right: Position of array in human motor cortex. Reprinted by permission from Macmillan Publishers Ltd: *Nature,* Hochberg LR, Serruya MD, Friehs GM, et al. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. 442: 164–171; copyright 2006. Bottom: Control of three-dimensional cursor movements by single neurons in motor cortex of a monkey. The left graph shows the improvement over daily training sessions of the average correlation (r^2) between the firing rate of a single cortical neuron and target direction. The right graph shows the resulting improvement in performance (assessed as the mean target radius required to maintain a 70% target hit rate). As the firing rates of the neurons that are controlling cursor movement become more strongly correlated with target direction, the size of the target can be reduced. Adapted from Taylor DM, Helms Tillery SI, and Schwartz AB (2003) Information conveyed through brain-control: Cursor versus robot. *IEEE Transactions on Neural Systems Rehabilitation and Engineering* 11: 195–199 (© 2003 IEEE), with permission from IEEE.

In this regard, it is encouraging that people severely disabled by ALS appear to retain the ability to control sensorimotor rhythms or single-neuron activity in sensorimotor cortex.

Apparently trivial and prosaic factors are likely to have substantial impact on the practical success of BCI applications. Factors such as the steps required for donning and doffing electrodes or for accessing a BCI application, or the user's appearance while operating the BCI, may greatly affect the numbers and characteristics of the people who adopt the system and the degree to which they actually use it in their daily lives.

Applications

BCIs have an extensive range of possible practical applications, from very simple to very complex. Simple BCI applications have already been demonstrated in the laboratory and in limited clinical use. They include systems for answering Yes/No questions, managing basic environmental control (e.g., lights, temperature), adjusting a television, or opening and closing a hand orthosis. Such simple systems can be configured for basic word-processing, sending e-mail, or accessing the Internet. For people who are totally paralyzed (i.e., locked-in), these simple BCI applications could make it possible to lead lives that are pleasant and even productive. Indeed, many recent studies indicate that, with good supportive care and the capacity for basic communication, severely paralyzed people can enjoy what they consider to be a reasonable quality of life and are hardly more likely to be depressed than are people without physical disabilities. Thus, simple BCI applications seem to have a significant future in their capacity to improve the daily lives of those who are most severely disabled. Indeed, a few severely disabled people are already using EEG-based BCI systems for important purposes in their daily lives.

More-complex BCI applications could control devices such as a motorized wheelchair, a robotic arm, or a neuroprosthesis that provides multidimensional movement to a paralyzed limb. Although present efforts focus on development of invasive BCI systems for such applications, noninvasive EEG-based BCIs also offer the possibility of such control. The eventual practical importance of such BCI applications will hinge on their capacities, practicality, and reliability, on their acceptance by specific kinds of users, and on the extent to which they have important advantages over conventional methodologies.

Establishment of the clinical value and practicality of BCI applications will require thorough evaluation to demonstrate their long-term reliability, to show that people actually use the applications, and to document that this usage has beneficial effects on factors such as mood, quality of life, and productivity. Particularly in the first stages of their development, it will frequently be necessary to configure applications that match each user's unique needs, desires, and physical and social environments. Although the cost of BCI equipment is relatively modest, current systems need substantial and continuing expert oversight, which is extremely expensive and now obtainable only from a few research labs. As a result, these systems are not available to most potential users. Thus, the widespread dissemination of BCIs will also depend on the degree to which the need for such ongoing technical support can be minimized. BCI systems must be easy to set up, easy to use, and easy to maintain if they are to have significant practical impact on improving the lives of people with severe disabilities.

Nature and Needs of BCI Research and Development

BCI research and development is a multidisciplinary effort. It requires neuroscience, engineering, applied mathematics, computer science, psychology, and rehabilitation. The need to select useful brain signals, to record them reliably, to analyze them appropriately in real-time, to control devices that provide functions valuable for those with severe disabilities, to manage the intricate short-term and long-term adaptive interactions between user and system, and to integrate BCI applications into the daily lives of their users means that the expertise and efforts of all these disciplines are essential for success. Thus, each BCI research group must incorporate all the essential disciplines, or groups with different expertise must collaborate closely. Collaborative studies are being facilitated by the widespread adoption of the general-purpose BCI software platform BCI2000, which can readily accommodate a wide variety of different signals, processing methods, applications, operating protocols, and hardware (see third website address). Productive collaborations have also been encouraged by recent meetings drawing BCI researchers from all relevant disciplines and from all over the world, by numerous symposia and collections of BCI research presentations at larger general meetings, and by publication of extensive sets of peer-reviewed BCI articles.

See also: Amyotrophic Lateral Sclerosis (ALS); Axonal Transport and ALS; Corticomotoneuronal System; Corticospinal Development; Electroencephalography (EEG); Evoked Potentials: Recording Methods; Evoked Potentials: Clinical; Map Plasticity and Recovery from Stroke; Synaptic Mechanisms of Learning.

Further Reading

Birbaumer N, Ghanayim N, Hinterberger T, et al. (1999) A spelling device for the paralyzed. *Nature* 398: 297–298.

Donchin E, Spencer KM, and Wijesinghe R (2000) The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering* 8: 174–179.

Hochberg LR, Serruya MD, Friehs GM, et al. (2006) Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* 442: 164–171.

Kübler A, Kotchoubey B, Kaiser J, Wolpaw JR, and Birbaumer N (2001) Brain–computer communication: Unlocking the locked in. *Psychological Bulletin* 127: 358–375.

Kübler A, Nijboer F, Mellinger J, et al. (2005) Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface. *Neurology* 64: 1775–1777.

Leuthardt EC, Schalk G, Wolpaw JR, Ojemann JG, and Moran DW (2004) A brain–computer interface using electrocorticographic signals in humans. *Journal of Neural Engineering* 1: 63–71.

Pfurtscheller G and Lopes de Silva F (2005) Event-related desynchronization (ERD) and event related synchronization (ERS). In: Niedermeyer E and Lopes daSilva FH (eds.) *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, pp. 1003–1016. Baltimore: Williams and Wilkins.

Robbins RA, Simmons Z, Bremer BA, Walsh SM, and Fischer S (2001) Quality of life in ALS is maintained as physical function declines. *Neurology* 56: 442–444.

Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, and Wolpaw JR (2004) BCI2000: A general-purpose brain–computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering* 51: 1034–1043.

Taylor DM, Helms Tillery SI, and Schwartz AB (2003) Information conveyed through brain-control: Cursor versus robot. *IEEE Transaction on Neural System Rehabilitation and Engineering*. 11: 195–199.

Taylor DM, Tillery SI, and Schwartz AB (2002) Direct cortical control of 3D neuroprosthetic devices. *Science* 296: 829–832.

Vaughan TM and Wolpaw JR (eds.) (2006) *The Third International Meeting on Brain-Computer Interface Technology: Making a difference IEEE Transactions on Neural Systems Rehabilitation Engineering* 14(2): 126–127.

Wolpaw JR and Birbaumer N (2006) Brain–computer interfaces for communication and control. In: Selzer ME, Clarke S, Cohen LG, Duncan P, and Gage FH (eds.) *Textbook of Neural Repair and Rehabilitation; Neural Repair and Plasticity*, pp. 602–614. Cambridge, UK: Cambridge University Press.

Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, and Vaughan TM (2002) Brain–computer interfaces for communication and control. *Clinical Neurophysiology* 113: 767–791.

Wolpaw JR and McFarland DJ (2004) Control of a two-dimensional movement signal by a noninvasive brain–computer interface in humans. *Proceedings of the National Academy of Sciences of the United States of America* 101: 17849–17854.

Relevant Websites

http://www.bci2000.org – BCI 2000 (brain–computer interface research project).

http://www.nature.com – *Nature* Supplementary Information (Supplementary Video 1 download).

http://www.bciresearch.org – QuickTime video ('Two-Dimensional Cursor Control With Scalp-Recorded Sensorimotor Rhythms') from the BCI Group, Wadsworth Center, Albany, NY.

Brains of Primitive Chordates

J C Glover, University of Oslo, Oslo, Norway
B Fritzsch, Creighton University, Omaha, NE, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Craniates (which include the sister taxa vertebrata and hyperotreti, or hagfishes) represent the most complex organisms in the chordate phylum, particularly with respect to the organization and function of the central nervous system. How brain complexity has arisen during evolution is one of the most fascinating questions facing modern science, and it speaks directly to the more philosophical question of what makes us human. Considerable interest has therefore been directed toward understanding the genetic and developmental underpinnings of nervous system organization in our more 'primitive' chordate relatives, in the search for the origins of the vertebrate brain in a common chordate ancestor. We describe some of the principal features of the central nervous systems of hemichordates, cephalochordates, and urochordates – taxa of increasing genetic relatedness to craniates. The central nervous systems of these taxa show a variety of forms, from a hardly specialized basiepithelial nerve plexus (hemichordates) to a few small ganglia with a tail nerve cord (urochordates), a swimming spinal cord with a barely recognizable cerebral vesicle (cephalochordates), and a fully developed brain and spinal cord (craniates). It is important to emphasize that although the term primitive may be construed as implying a lower rank in evolutionary sequence, we use it strictly with respect to extant brain complexity. Each of these groups has in fact existed for substantial periods of evolutionary time and represents highly adapted life-forms in their own right. We focus here on shared aspects of central nervous system structure and development that may relate to a common evolutionary origin but also indicate some of the specialized structural features that bear witness to the degree of divergence and range of derived characters that have arisen in the several hundred million years since these groups split from that origin.

Phylogenetic Relationships

Originally constructed solely through comparative analysis of morphological traits, phylogenetic trees are being increasingly tested and revised on the basis of gene sequence comparisons. Such analysis, although providing a more direct link to the evolutionary clock, is nevertheless hampered by differing rates of evolution, both among species and among genes, and a still largely deficient fossil record. Until recently, it was widely accepted, both on morphological and molecular grounds, that cephalochordates and craniates were sister taxons, with urochordates being more distant craniate relatives and with hemichordates being more closely related to echinoderms (Figure 1(a)). The molecular data only weakly supported a coherent chordate taxon, however, indicating that apparent morphological similarities among chordates are imposed on deep divisions among the extant deuterostome taxa. Recent analysis of a substantially larger number of genes has reversed the positions of cephalochordates and urochordates, promoting the latter to the most closely related craniate relatives (Figure 1(b)).

Comparative Appearance of Brains, Spinal Cords, and Nerves

Primitive Craniates (Cyclostomes)

The jawless cyclostomes, lampreys and hagfish, constitute a basic clade of craniates that shows the major features of brain organization also found in gnathostomes (jawed vertebrates). Externally, the brain of extant cyclostomes consists of a bilaterally symmetric rostral enlargement, the forebrain (telencephalon), each side of which is composed of an olfactory bulb and a cerebral hemisphere (Figure 2). Each hemisphere is connected to the single, bipartite diencephalon. The diencephalon gives rise to a number of neural appendages. Ventrolaterally are located the paired optic nerves leading to the lateral eyes. Dorsally lies a single enlargement, the habenula, that in lampreys, but not hagfish, has an attached pineal organ. Ventrally is located the pituitary gland. The central ventricle of the diencephalon is continuous with the ventricle in the midbrain, the hindbrain, and the central canal of the spinal cord, all of which are greatly reduced in hagfish. Lampreys have a dorsal central opening in the midbrain that is covered by a choroid plexus, a uniquely derived feature of lampreys not shared by other craniates. Lampreys, like gnathostomes, have an oculomotor nerve leaving the midbrain ventrally and a trochlear nerve exiting lateral to the cerebellum, neither of which do hagfish possess. Caudal to the small cerebellum in lampreys is the rhombencephalon (hindbrain) with a large choroid plexus covering the IVth ventricle. A small ventricle is also found in hagfish, but there is no trace of a

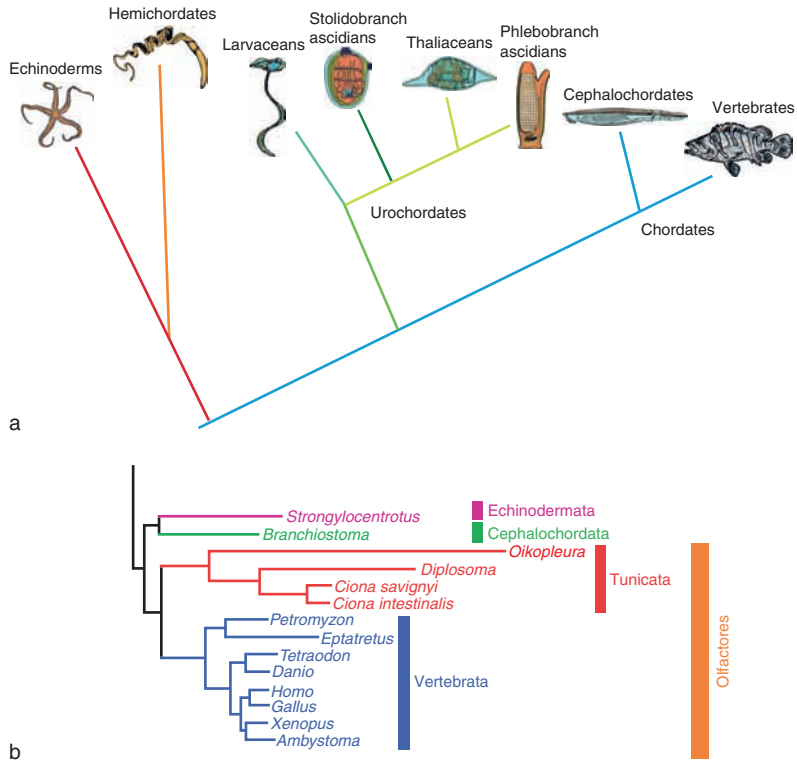

Figure 1 (a) Conventional scheme of phylogenetic relationships among chordates and related taxa. (b) A revised scheme based on more recent gene sequence analysis. Adapted from original image by Dr. Billie Swalla, University of Washington.

choroid plexus and the presence of a cerebellum has been questioned.

Lampreys have all the hindbrain nerves found in gnathostomes (trigeminal, abducens, facial, otic or stato-acoustic, glossopharyngeus, and vagus), with the possible exception of the hypoglossal nerve. The relative positions and fiber compositions of these nerves are very similar but not identical in cyclostomes and gnathostomes. For example, the abducens nerve root is almost integrated into the trigeminal nerve root in lampreys, whereas it is always a separate ventral nerve root at a more caudal level in gnathostomes. Moreover, gnathostomes have three distinct motor neuron populations in the brain stem, each innervating a different type of peripheral target: the somatic motor neurons that innervate somitomere-derived musculature, the branchial motor neurons that innervate branchial arch-derived musculature, and the visceral motor neurons that innervate neural crest-derived parasympathetic ganglia of the head and body. In contrast, cyclostomes as a group lack the somatic motor neurons of the hypoglossal nucleus and have no visceral motor neurons since no cranial parasympathetic ganglia are known to exist. However, the hindbrain of lampreys contains so-called branchiomotor neurons, the axons of which exit dorsally and innervate branchial arch-derived muscle. To emphasize differences between gnathostomes and cyclostomes, the hagfish hindbrain, while recognizable as such, is unusually shaped, which relates to differences in its internal organization. The organization of cranial nerves in hagfish, other than the apparent absence of the entire extraocular muscle-related nerves, shows a number of deviations from gnathostome vertebrates. Hagfish have three completely segregated parts of the trigeminal nerve, two otic (stato-acoustic) nerves, no recognizable vagal ganglion, and a facial nerve that exits dorsal to the otic nerves.

The spinal cord in craniates is a continuous extension of the neural tissue of the hindbrain. Adult cyclostomes have unusually shaped spinal cords that are dorsolaterally flattened. Lampreys, like gnathostomes, have separate dorsal and ventral roots, but unlike in gnathostomes, these do not converge to form mixed spinal nerves and are asymmetrically distributed between the left and right side. A similar but more pronounced asymmetry of spinal nerves is exhibited by cephalochordates and has been suggested to be primitive for chordates because an asymmetric pattern is also found in hagfish and some gnathostomes. In hagfish, as in gnathostomes, the dorsal and ventral nerve roots converge, except in

Saccoglossus cambrensis

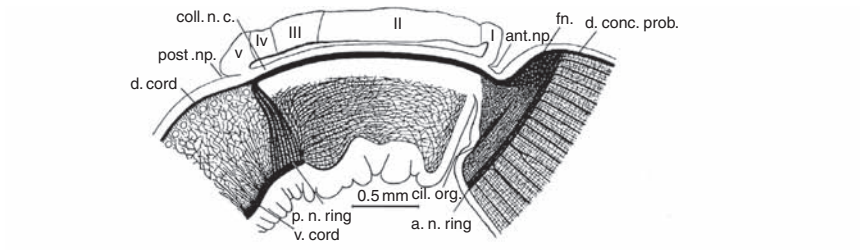

Branchiostoma floridae

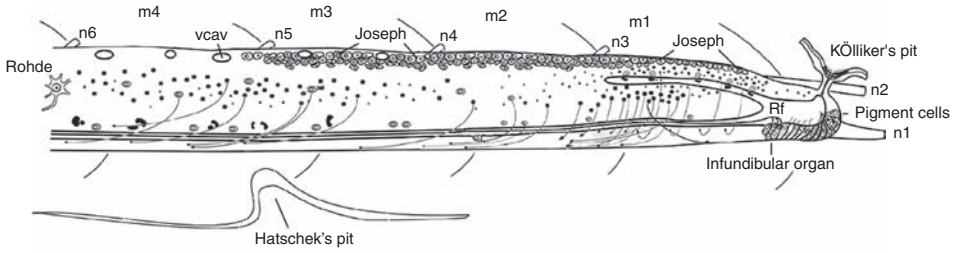

Ciona intestinalis

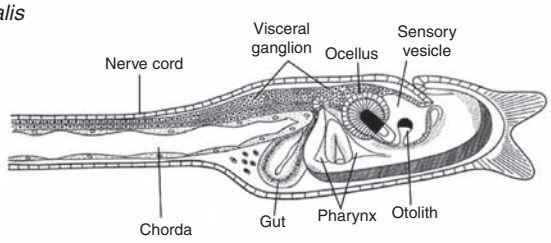

Oikopleura dioica

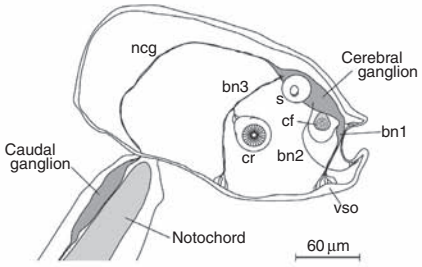

Myxine glutinosa

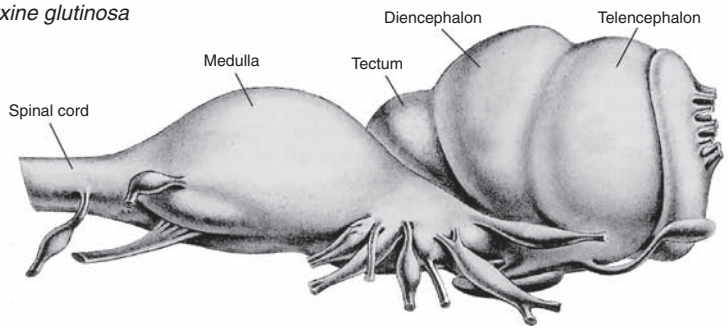

Figure 2 Continued

the tail region. It is believed that convergent dorsal and ventral roots is a derived feature of gnathostomes and that the superficial similarity of this feature in hagfish is independently derived. The fiber composition of spinal nerves is highly variable among cyclostomes and fits neither the cephalochordate nor the basal gnathostome composition. Consequently, their relationship to the nerves of cephalochordates and urochordates is problematic.

As in craniates, somatic motor neurons are distributed along the length of the spinal cord in cycolostomes and provide independent innervation of each segment of body musculature (**Figure 3**). Cyclostomes do not have autonomic ganglia and lack spinal visceral motor neurons that innervate such ganglia (**Figure 4**). It appears that autonomic ganglia arose with gnathostomes through the advent of migratory neural crest cells. This innovation has led to the formation in higher craniates of preganglionic parasympathetic motor neurons in the brain stem and the sacral part of the spinal cord and preganglionic sympathetic motor neurons in the thoracic and upper lumbar spinal cord (**Figure 4**).

Cyclostomes have dorsally located intramedullary sensory neurons (dorsal cells), which appear to be homologous to the Rohan–Beard cells of anamniote craniates, as well as sensory neurons in dorsal root ganglia. Dorsal root ganglia and proximal cranial ganglia are derived from migratory neural crest in craniates; in cyclostomes, spinal dorsal root ganglia are generated by neural crest, whereas cranial ganglia derive from placodes. In contrast to their craniate counterparts, the dorsal root ganglion cells in cyclostomes express modulatory neurotransmitters such as serotonin and neuropeptides and may therefore have more derived functions, although they do appear to transmit nociceptive signals in lampreys.

In many chordates, large neurons with descending axons have been identified in the rostral region of the neuraxis. Using cytological criteria, one can identify certain large reticulospinal neurons in the rhombencephalon of lampreys (the Müller and Mauthner cells) that have long descending axons that extend the length of the spinal cord. Similar neurons are found in hagfish, but Mauthner cells cannot be identified among them, and whether the other large cells are homologous to Müller cells is unclear.

Cephalochordates

The central nervous system of cephalochordates is a simple tube that does not show any obvious enlargement at the rostral pole comparable to the brain

Figure 2 A comparison of the basic anatomical structure of the hemichordate, cephalochordate, urochordate, and craniate central nervous systems. Enteropneust hemichordates (represented by *Saccoglossus cambrensis*) have an epidermal nerve network that shows condensations in certain areas. At the base of the proboscis is an anterior nerve ring (a.n. ring) that is next to the ciliary organ (cil.org.), which is adjacent to the oral opening (asterisk). The collar region has a collar nerve cord (coll.n.c.), an invaginated part of the epidermis with anterior and posterior neuropores (ant.np., post. np.) that lies dorsal to the buccal cavity. At the third body division, the metasome, the collar nerve cord becomes confluent with the dorsal nerve cord (d.cord) and, through the posterior nerve ring (p.n.ring), with the larger ventral cord. Neither true nerves nor major sensory organs are apparent in this simple epithelial nerve net. Cephalochordates (represented by an adult specimen of *Branchiostoma floridae*) have a hollow nerve cord that is fairly uniform in diameter throughout its length. There is a recognizable vesicular expansion at the anterior end, with a number of specialized appendages and neuron groups that provide indications of cephalic elaboration. Among these are Kölliker's pit (believed by some to be an olfactory organ), the pigmented cells of the single frontal eye, the infundibular organ (which gives rise to Reissner's fiber; Rf) and preinfundibular zone (possibly containing an organ of balance), and a dorsal column of putative photoreceptors called Joseph cells. The first nerve (n1) is ventral, and all other nerves are dorsal. Muscle segments (m) are located along the entire length of the nervous system. Urochordates have small nervous systems with a recognizable tripartite organization, consisting of a rostral ganglion (termed the cerebral ganglion in appendicularians and the sensory vesicle in ascidians), a caudal ganglion (termed the visceral ganglion in ascidians), and a caudal nerve cord. This organization is maintained throughout life in appendicularians (exemplified here by *Oikopleura dioica*), whereas ascidians (represented here by *Ciona intestinalis*) lose this organization at metamorphosis from larva to adult, in line with the transition from pelagic to sessile lifestyle. The rostral ganglion is intimately associated with a statocyst (s, containing an otolith) and in ascidian larvae also an ocellus. In appendicularians, there is also a ciliated funnel (cf), an outgrowth of the cerebral ganglion believed to be homologous to the pituitary gland. Several brain nerves (bn) connect the rostral ganglion to peripheral sensory and motor structures. The rostral ganglion is connected to the caudal ganglion by a nerve trunk (ncg) in appendicularians and by a cellular 'neck' region in ascidians. The nerve cord of ascidians is evidently nearly aneuronal, consisting predominantly of ependymal cells, whereas in appendicularians it contains neurons, including motor neurons. Craniates (shown is the hagfish, *Myxine*) have a brain that develops from invaginated ectoderm that becomes completely transformed into nervous tissue but remains confined within the former epithelial basement membrane. Major subdivisions that can be identified on the basis of external and internal features include (in anteroposterior order, right to left in the figure) the telencephalon (forebrain), diencephalon, mesencephalon (midbrain), and rhombencephalon (pons and medulla oblongata, or the hindbrain). Several distinct cranial nerves pass through the basement membrane to connect the brain with various sensory organs that provide chemical (olfaction and taste), mechanical (touch, proprioceptive, auditory, and vestibular), and visual (eyes) input. The brain integrates this input with information about internal states and uses this to generate motor output that is effected via motor neurons in the brain stem and spinal cord. The image of amphioxus is reproduced from Wicht H and Lacalli T (2005) The nervous system of amphioxus: Structure, development, and evolutionary significance. *Canadian Journal of Zoology* 83: 122–150. The image of *Saccoglossus Cambrensis* is reproduced from Bullock TH and Horridge GA (1965) *Structure and Function in the Nervous Systems of Invertebrates*. San Francisco, CA: W. H. Freeman. The image of *Myxine glutinosa* is reproduced from Nieuwenhuys R, Donlekaar HJT, and Nicholson C (1998) *The Central Nervous System of Invertebrates*. New York: Springer.

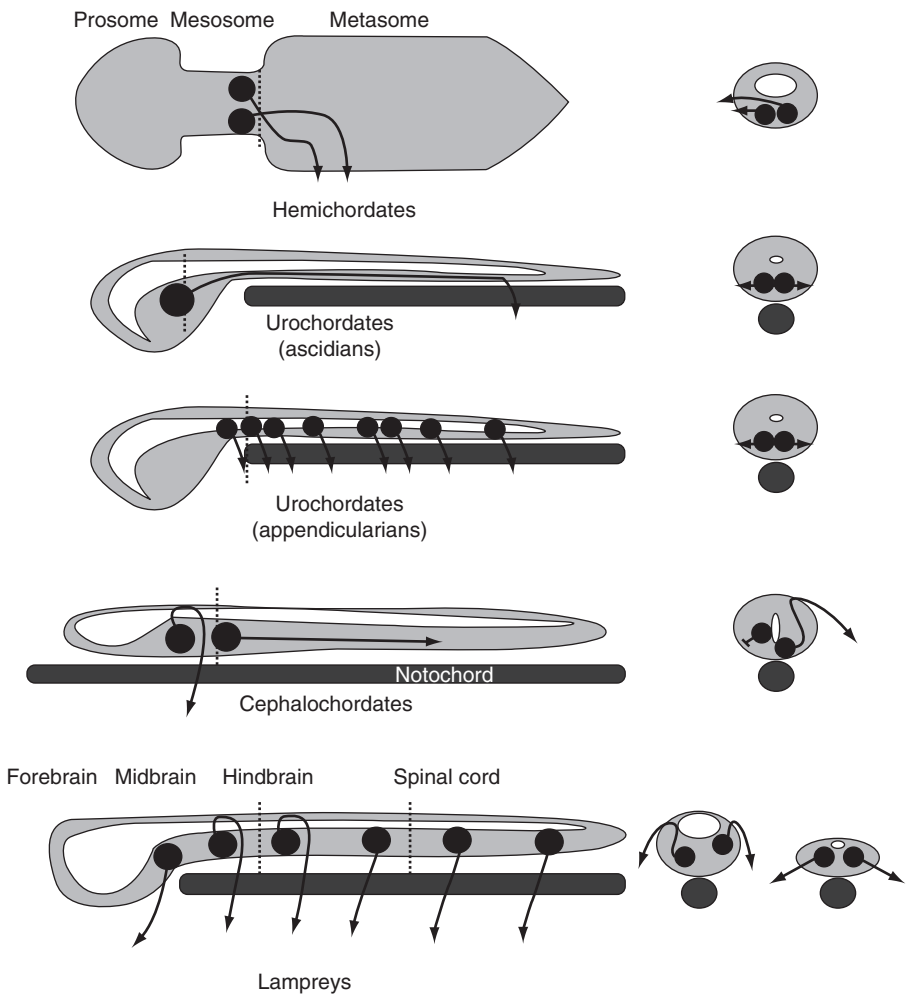

Figure 3 The distribution of motor neurons is shown from the dorsal aspect (hemichordates) or lateral aspect (others) as well as in transverse projections viewed in the caudal direction from the levels indicated by the dotted lines. Motor neurons in hemichordates are predominantly located in the caudal half of the collar cord, and their axons typically cross the basal membrane individually to reach muscle fibers. In ascidian urochordate larvae, all motor neurons are concentrated in the caudal ganglion and project caudally through the nerve cord to reach muscle fibers of the tail. In appendicularian urochordates, motor neurons are also present in the nerve cord and project directly to adjacent muscles. Cephalochordates have three types of motor neurons (only two are shown). One type is the somatic motor neuron that extends an axon along the spinal cord margin to innervate muscular processes that form synapses abutting the cord. The second type is the visceral motor neuron that projects through the dorsal root to innervate the pterygial muscle. The third type innervates muscle cells of the notochord at the ventral spinal cord margin. Lampreys have only somatic motor neurons in the spinal cord, whose axons project out through ventral roots, and in their hindbrain only dorsal exiting motor neurons, referred to as branchiomotor neurons because they innervate muscle derived from branchial arches. Lampreys also have three populations of ocular motor neurons in the midbrain/hindbrain region, some of which have the appearance of somatic motor neurons.

of craniates, hence the alternate name 'acrania' for this taxon. The hollow tube has a central canal that shows a vesicular enlargement at the anterior pole, into which the processes of frontal eye sensory cells extend. Like craniates, cephalochordates (as well as urochordates) have a unique structure that extends throughout the central canal, Reissner's fiber, which may provide a topological landmark for comparing different brain regions (Figure 5). Another structure of similar utility is the notochord, which shows a striking difference in its relationship to the brain in cephalochordates and craniates. In both taxa, the notochord lies subjacent to the entire length of the spinal cord. In craniates, it extends rostrally to end at the transition from hindbrain to midbrain, whereas in cephalochordates it extends beyond the rostral aspect of the neural tube. The notochord is known to have a major inductive influence on the developing brain and spinal cord of craniates through the release of diffusible signaling proteins. This influence and the resultant regional patterning is therefore likely to be shifted rostrally along the neuraxis of cephalochordates relative to craniates, potentially contributing to the acraniate nature of the cephalochordates.

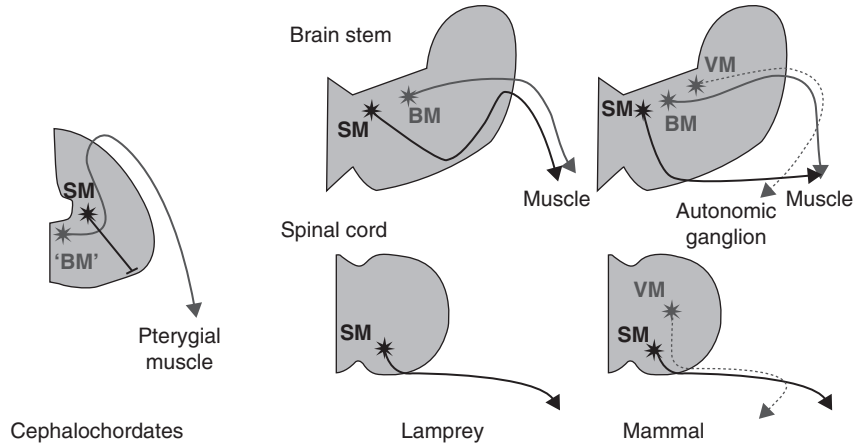

Figure 4 Motor neurons and their projections are compared in craniates and cephalochordates. Two motor neuron types have been characterized in cephalochordates – the branchiomotor neurons, which innervate the pterygial muscle and exit through dorsal roots, and the somatic motor neurons, which do not exit the neural tube but rather form synapses with muscle processes at the lateral margin of the spinal cord. The lamprey brain stem has mainly branchiomotor neurons (BM) that project through dorsal roots. Whether the abducens is a somatic motor nucleus in lampreys as indicated (SM) is unclear because its motor neurons also project through a dorsal root. The lamprey spinal cord has only somatic motor neurons that exit through the ventral root. Mammals have evolved visceral motor neurons (VM) in the spinal cord that migrate into a distinct position and project to autonomic ganglia. In mammals, there is also the addition of visceral motor neurons in several cranial nerves that project to the parasympathetic ganglia. These visceral motor neurons can be regarded as special branchiomotor neurons.

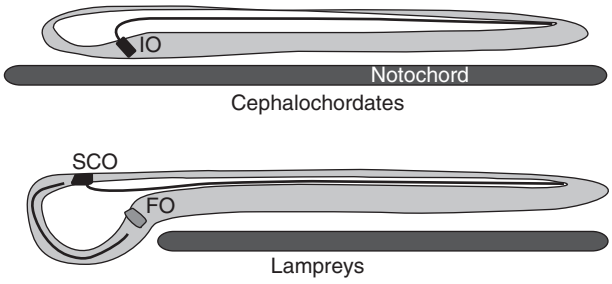

Figure 5 In one proposal for the evolutionary origin of the forebrain, the flexural organ (FO) and the subcommisural organ (SCO) represent a split of a single original cell population, represented by the infundibular organ (IO) in cephalochordates, thereby intercalating a novel population (black line). If this is accepted, then the entire area between the SCO and FO of craniates could be viewed as a craniate neomorph, in which case the entire telencephalon and large areas of the thalamus of craniates would have no equivalent in cephalochordates.

The cranial and spinal nerves of cephalochordates are difficult to compare with those of craniates. First, they are distributed too far rostrally, leaving virtually no space for what is typically considered the brain. There are no true ventral roots in the sense of motor neuron axons leaving the spinal cord. Rather, the somatic muscles extend processes that contact the ventrolateral margin of the spinal cord where they appose synapses from central processes of motor neurons (Figures 3 and 4). Moreover, the notochord is composed of muscle fibers that similarly contact the ventral spinal cord where they receive synapses from ventromedial motor axons. Motor neurons are found throughout the spinal cord and as far rostral as the primary motor center between the second and third dorsal nerves. An additional set of motor neurons are the so-called visceral motor neurons. These are the most ventral neurons in the nervous system, situated virtually at the floor plate in the anterior portion of the nerve cord. The axons of these cells exit through the dorsal roots and supply the pterygeal muscle with cholinergic endings. There is in fact only one true ventral nerve, the first (most rostral) nerve. All other nerves are dorsal and become increasingly more caudal on the right than on the left side – an asymmetry that by far exceeds anything found in other chordates.

There are no dorsal root ganglia anywhere along the neuraxis of cephalochordates, in line with the apparent lack of a migratory neural crest population. Sensory projections arise from dorsally located intramedullary sensory neurons, similar to those found in cyclostomes. In addition, there is evidence for the presence of peripheral sensory neurons located in the epithelial lining of the buccal and atrial body cavities that project centripetal axons into the nerve cord. Paired sensory organs are conspicuously absent in cephalochordates. However, there is an unpaired frontal eye, putative balance and olfactory organs, and certain primordia of mechanosensors and chemosensors may be present among the numerous single or multicellular sensory organs. Among these are small encapsulated clusters of sensory cells located in branches of the first two brain nerves, the corpuscles of Quatrefages, believed to have a mechanosensory function.

Other features of the cephalochordate anterior neuraxis are the conspicuous lamellar body, believed to be

homologous to the vertebrate pineal gland but present only in larvae, the infundibular organ, whose name implies a proposed homology to the ventral diencephalic region, and the dorsally located Joseph cells, which are putative photoreceptors. Conspicuously pigmented individual ocelli are distributed along the length of the spinal cord. How these various features relate to the internal divisions of the craniate brain and spinal cord is unclear, leaving open the question of the existence of brain subdivisions, although gene expression studies have begun to shed some light on this issue.

In the rostral neuraxis of juvenile cephalochordates, two pairs of larger interneurons with descending axons have been identified and termed 'ventral giant cells of the primary motor center,' potentially homologous to the large reticulospinal neurons of cyclostomes. Another system of large neurons, the Rohde cells, are situated dorsally in the anterior region, starting at what is considered to be the transition from brain to spinal cord, and have descending axons in the spinal cord. A separate population of Rohde cells is located in the caudal spinal cord.

Urochordates

The urochordate subphylum includes the ascidians, the thaliaceans, and the appendicularians. Nearly all ascidians, some thaliaceans, and all appendicularians have a free-swimming larval stage with a notochord and a small central nervous system that appear homologous to the same structures in craniates. In contrast to the rather uniform appearance of the neuraxis of cephalochordates, the larval ascidian and larval and adult appendicularian central nervous systems are organized into an obvious tripartite structure (Figure 2). The three divisions are (1) a rostral ganglion (totaling approximately 215 cells in the ascidian *Ciona* and approximately 75 cells in the appendicularian *Oikopleura*), which contains sensory receptor structures, including an ocellus and/or an otolith, followed by (2) a caudal ganglion (containing approximately 45 cells in *Ciona* and approximately 25 neurons in *Oikopleura*), from which extends (3) a caudal nerve cord (containing approximately 65 cells, mostly ependymal, in *Ciona*, and approximately 30 neurons and 25 support cells in *Oikopleura*). In addition, the rostral and caudal ganglia are connected by a nerve trunk in appendicularians or a slender 'neck' region containing nerve cells in ascidians. Other urochordate taxa seem to exhibit primarily variations in size but not in structure. The ganglia of urochordates have the organization of an invertebrate ganglion, with cell bodies at the periphery and a neuropil in the center. Several nerves that vary considerably between species have been traced from adult ganglia, reaching up to 75 nerves in certain salps. These nerves appear to be mixed sensory and motor nerves and are asymmetric in several species, potentially related to the overall body asymmetry.

The organization of motor projections varies among urochordates. In ascidian larvae, all the motor neurons (three to five pairs in different species) are located in the caudal ganglion and project along the predominantly aneural nerve cord before exiting to innervate the peripheral muscle. In appendicularians, both the caudal ganglion and the nerve cord contain motor neurons, whose axons project laterally to the peripheral muscle either directly or after extending some distance in the cord.

Comparisons between urochordates and cephalochordates claim similarities in sensory organs within the central nervous system, in particular the infundibular sensory cells and otolith. The ciliated funnel, which extends to one side of the rostral ganglion, is believed to be homologous to the pituitary. However, the interpretation of homologies for most sensory cells and sensory organs outside the central nervous system is contended. Some adult urochordates have fairly complex eyes attached to the cerebral ganglion. There is a marked asymmetry in the organization of central and peripheral sensory structures.

No large neurons with descending axons that might be homologous to the large reticulospinal neurons of craniates have been described in urochordates, although there are some large cells within the rostral ganglion that could subserve a similar function.

Hemichordates

Hemichordates are not grouped within the chordates but share a number of chordate features, including a rudimentary notochord-like structure, the stomochord. Whether the simple nervous system of hemichordates should be referred to as a central nervous system is unclear. The overall organization is that of a basiepithelial plexus that shows regional concentrations in the three parts of the body, the protosome (the preoral proboscis), the mesosome (the postoral collar), and the metasome (or trunk, with rostral gill slits). The intraepithelial nerve plexus is well developed on the basement membrane but remains epithelial even in the invaginated collar region, which is hollow and opens through two neuropores (Figure 2). Concentrations of longitudinal strands of cells and fibers also exist on the trunk, where they form a dorsal and a larger ventral cord. The proboscis has a well-developed nerve plexus and numerous sensory cells. Except for the preoral ciliary organ with its abundance of sensory cells, there are no specialized sense organs. Concerning nerves, it appears that muscles are supplied by nerve fibers that cross the basement membrane singly and diffusely without forming obvious peripheral nerves.

Large neurons have been described in hemichordates, clustered in the caudal part of the collar cord and also scattered in more rostral and caudal areas. Some of these neurons have uncrossed or crossed axons that extend toward the ventrolateral longitudinal muscles and therefore may be motor neurons. Others have been compared to the Mauthner and Müller cells of craniates.

Molecular Patterning of the Neuraxis during Development

Despite their overall anatomical differences, the central nervous systems of the primitive chordates exhibit similarities to craniates with respect to the molecular mechanisms responsible for regional patterning during development, along both anteroposterior and dorsoventral axes. Shared by all chordates and the hemichordates is the expression of *otx* genes in the most rostral region of the central nervous system and the expression of *hox* genes starting more caudally; in all these taxa except cephalochordates, the *engrailed* gene (*en*) is expressed in the region between *otx* and *hox* genes, and in all save the hemichordates *pax* genes of the Pax-2/5/8 class are expressed in the very most rostral region (Figure 6). These elements alone provide support for proposed homologies between specific portions of the primitive chordate nervous

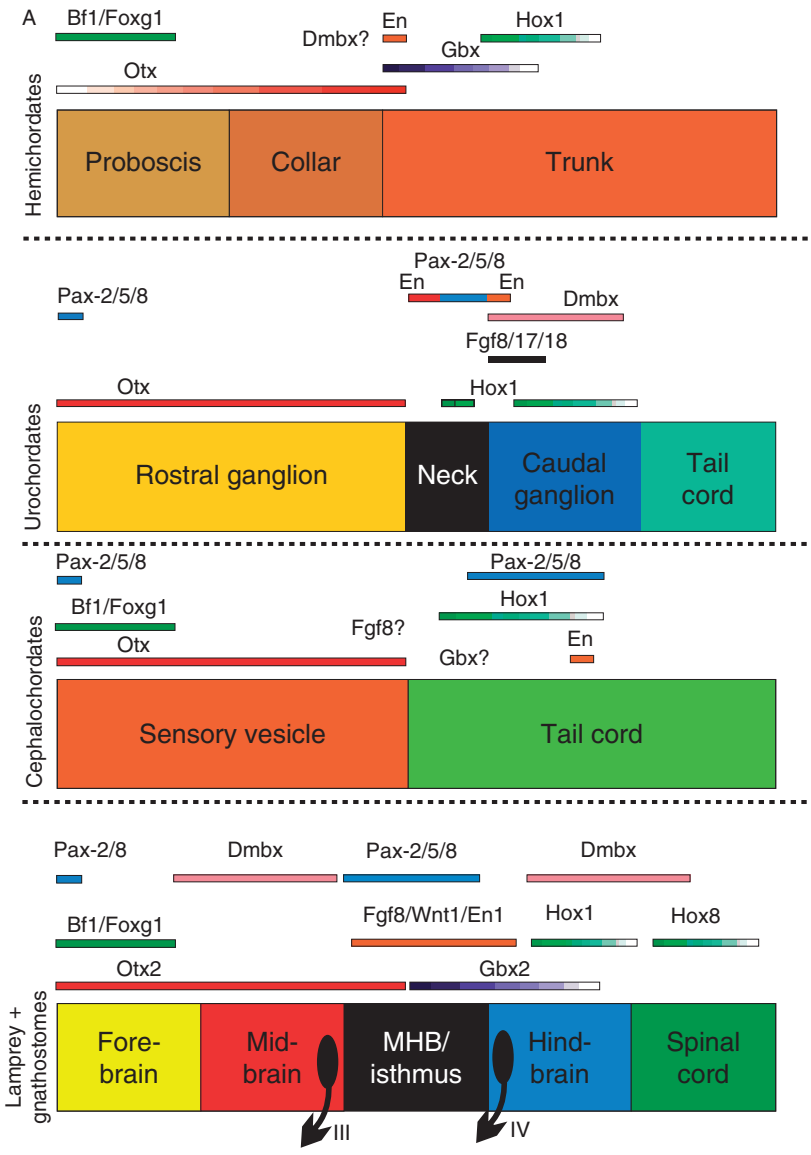

Figure 6 Gene expression patterns define regions along the anteroposterior axis. Several relationships are highly conserved, including the rostral expression of *pax2/5/8* and *otx* genes and a more caudal expression of *hox* and *en* (*engrailed*) genes. These gene expression patterns provide support for defining putative forebrain and hindbrain homologs in cephalochordates and urochordates. Intermediate gene expression patterns exhibit too many variants to support a clear homology with the midbrain and midbrain/hindbrain boundary (MHB), however.

systems and the brain subdivisions in craniates. On this basis, both urochordates and cephalochordates would appear to have homologs of the craniate forebrain, hindbrain, and spinal cord. Less compelling is the presence of a region homologous to the midbrain and the midbrain/hindbrain boundary (MHB), the latter of which is an important inductive organizer region in the craniate brain. Although certain genes that delineate the MHB in craniates, such as Pax2/5/8, are expressed in the general vicinity of the expected MHB location, the specific gene combination criteria that define the MHB in craniates are not met in any of the primitive chordates. This may mean that the definitive MHB and midbrain are later evolutionary elaborations of the neuraxis.

These genes, whose expression regionalizes the anteroposterior neuraxis (Figure 6), may be ancestrally related to the overall rostrocaudal patterning of the deuterostome body, and the degree of pattern conservation may reflect the fact that the neural tube evidently evolved only once in ancestral chordates. The presence of a similar anteroposterior pattern in the basiepithelial nerve plexus of hemichordates suggests that this anteroposterior regionalization predates the invagination of the neural tube from external ectoderm. A similar anteroposterior sequential pattern of some of the key genes in the insect nervous system lends support to this idea.

An alternative approach to deducing anteroposterior homologies between craniate and noncraniate nervous systems is to assess the location of specific neuron types known to bear a specific topographic relationship to given craniate brain subdivisions. Catecholaminergic neurons, which express the enzyme tyrosine hydroxylase (TH), provide a good example because these have very focal distributions within the craniate brain. In cephalochordates, TH- and catecholamine-positive neurons are found near the anterior end of the neural tube and near the first two dorsal nerves. This has been interpreted to indicate similarities with the di-, mes-, and rhombencephalic catecholaminergic neuron groups of craniates. TH-positive neurons are also present within the rostral ganglion of urochordates, providing support for a diencephalic or mesencephalic character there. Catecholamine-positive neuron populations have not been described in hemichordates.

Neural Crest, Placodes, and Their Derivatives

One of the fundamental evolutionary advances that characterizes craniates is the appearance of the neural crest and ectodermal placodes, from which are derived a large spectrum of peripheral structures and cell types, notably through the process of epithelial delamination that generates migratory mesenchyme. Restricting our focus to neural structures, the neural crest gives rise to all peripheral ganglia in the trunk, including sensory, autonomic, and enteric ganglia, and together with specific placodes, all of the cranial autonomic ganglia. It also gives rise to Schwann cells, which provide myelination of peripheral nerves. The placodes, which are local thickenings of the embryonic cranial ectoderm, invaginate to form sensory structures and associated neural epithelia and ganglia, in addition to contributing to some autonomic ganglia. The elaboration of sensory and motor function in the cranial region provided by the advent of neural crest and placodes clearly facilitated a more rapid acquisition of the predatory behavioral repertoire exhibited by craniates.

The substantial divergence in peripheral sensory and motor structures and limitations in the embryonic data have made the search for neural crest and placode precursors in noncraniates a difficult field. In hagfish, neural crest and placodes are clearly present at embryonic stages, but there is little direct evidence that they generate migratory cell populations, possibly limiting the scope and spatial range of the derivatives they can produce. Definitive migratory neural crest appears to be lacking in cephalochordates and urochordates. In particular, neural crest-specific genes do not appear to be expressed in the lateral margins of the neural tubes of these taxa as they are in craniates. Nevertheless, there are small populations of migratory neural cells in both taxa that have been proposed to have an evolutionary relationship to the neural crest. Possible placode-like structures have been identified both in amphioxus and in urochordates, and the expression of some placode-specific genes in these structures strengthens the argument that they may represent the early origins of certain craniate placodes. The first brain nerves of appendicularians, for example, express a gene profile and possess cell types consistent with a homology to the craniate olfactory placode, and similar evidence provides support for a homology between the appendicularian ciliary funnel and the adenohypophysis. In amphioxus, Hatschek's pit has also been proposed as an adenohypophyseal homolog, but here the gene expression data are weaker. A number of other small peripheral sensory structures in appendicularians express placode-specific genes, but their relationship to craniate placodes is less clear. It has been suggested that specific mechanoreceptors may represent precursors of the lateral line placode derivatives. Beyond these proposals, the evidence for placodal homologies is currently weak, and even the placodal nature of these best documented cases is still controversial.

See also: Allometric Analysis of Brain Size; Brain Development: The Generation of Large Brains; Brain Fossils: Endocasts.

Further Reading

Bassham S and Postlethwaite J (2005) The evolutionary history of placodes: A molecular genetic investigation of the larvacean urochordate *Oikopleura dioica*. *Development* 132: 4259–4272.

Bullock TH and Horridge GA (1965) *Structure and Function in the Nervous Systems of Invertebrates*. San Francisco, CA: W. H. Freeman.

Canestro C, Bassham S, and Postlethwaite J (2005) Development of the central nervous system in the larvacean *Oikopleura dioica* and the evolution of the chordate brain. *Developmental Biology* 285: 298–315.

Delsuc F, Brinkmann H, Chourrout D, and Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439: 965–968.

Fritzsch B and Northcutt RG (1993) Cranial and spinal nerve organization in amphioxus and lampreys: Evidence for an ancestral craniate pattern. *Acta Anatomica* 148: 96–109.

Ghysen A (2003) The origin and evolution of the nervous system. *International Journal of Developmental Biology* 47: 555–562.

Hirth F, Kammermeier L, Frei E, Walldorf U, Noll M, and Reichert H (2003) An urbilaterian origin of the tripartite brain: Developmental genetic insights from *Drosophila*. *Development* 130: 2365–2373.

Holland ND (2003) Early central nervous system evolution: An era of skin brains? *Nature Reviews Neuroscience* 4: 617–627.

Lowe CJ, Wu M, Salic A, et al. (2003) Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* 113: 853–865.

Meinertzhagen IA, Lemaire P, and Okamura Y (2004) The neurobiology of the ascidian tadpole larva: Recent developments in an ancient chordate. *Annual Review of Neuroscience* 27: 453–485.

Nieuwenhuys R, Donkelaar HJT, and Nicholson C (1998) *The Central Nervous System of Vertebrates*. New York: Springer.

Schlosser G (2005) Evolutionary origins of vertebrate placodes: Insights from developmental studies and from comparisons with other deuterostomes. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 304: 347–399.

Shimeld SM and Holland ND (2005) Amphioxus molecular biology: Insights into vertebrate evolution and developmental mechanisms. *Canadian Journal of Zoology* 83: 90–100.

Søviknes AM, Chourrout D, and Glover JC (2005) Development of putative GABAergic neurons in the appendicularian urochordate *Oikopleura dioica*. *Journal of Comparative Neurology* 490: 12–28.

Søviknes AM, Chourrout D, and Glover JC (2007) Development of the tail nerve cord of the appendicularian urochordate *Oikopleura dioica*. Journal of Comparative Neurology 503: 224–243.

Wicht H and Lacalli TC (2005) The nervous system of amphioxus: Structure, development, and evolutionary significance. *Canadian Journal of Zoology* 83: 122–150.

Brainstem and Cranial Nerves

C Matesz and G Szekely, University of Debrecen, Debrecen, Hungary

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The history of cranial nerve classification dates back to the twelfth century, and distinction of the 12 pairs of cranial nerves is associated with the publication by Soemmering in 1778. In that time, the great epoch-making explorations were achieved in the field of comparative neuroanatomy, and the recognition of gradually increasing organismal complexity in the more advanced vertebrate phyla led to the conclusion that new levels in complexity arose by differentiation of existing structures in evolution. This view was the Kopftheorie ('head theory') of JW Goethe in 1790, who stated that the cranium was composed of modified vertebrae. Consequently, new structures in the head and neck region were homologized with structures in the trunk accordingly, and the homologization of spinal nerves with nerves of the brain stem produced a number of contradictions and a peculiar terminology.

In our present endeavor we want to consider the following characteristics of motor neuron pools innervating different structures: the specific functions performed by the target muscles of motor neurons, the morphology of individual neurons constituting the neuron pool, and the genetic profile of neurons in that pool. This complex approach reveals a number of minor details which help make a distinction among motor neurons classified in a common group, but which control different muscles with a variety of actions. In taking this approach, problems sharply emerge in assessing transitions between different species of vertebrae when a new function appears and the smooth transition between classes is missing, or extinct. For example, some fishes have neither a tongue nor a hypoglossal nerve, but their 'evolutionary neighbors,' the amphibians, possess a beautiful tongue with a rather sophisticated innervation. Or consider the capacity of mammals to grind and chew food, which depends on the evolution of a 'chewing organ' innervated by a group of motor neurons; this feature does not exist in nonmammalian vertebrates. Another example would be the auditory muscles and their innervation; the course of their evolution is still an exciting problem, not only of neurobiology but of evolution as well.

We think that with our method of assessing the characteristics of motor neuron pools innervating different structures we can demonstrate the beautiful parallelism in the evolution of the peripheral target and the central neuronal structure. We can show, from literary data, differences in the genetic profiles of motor neurons controlling different actions; by pursuing this line of studies, we may eventually find the genetic background in transition from one class to the next during the course of evolution. In order to demonstrate the differences in organization of cranial nerve nuclei of different species, we have to start with the organization of motor neurons in the spinal cord. We try to group the motor neurons according to the specification they control.

Efferent System of the Spinal Cord

Function, Position, and Form

According to the action performed by muscle groups in the trunk region, we can distinguish motor neurons that innervate epaxial muscles, hypaxial muscles, limb muscles, and others (Figure 1). Epaxial motor neurons innervate muscles that originate from, and are attached to, the axial skeleton. They occupy the ventromedial part of the ventral horn and their fusiform perikaryon emits a strong lateral and medial richly arborizing dendritic tree. The medial dendrites cross the midline and reach the contralateral lateral white column. The position of motor neurons and the bilateral expanding dendritic arbor receiving inputs on both sides provide an excellent model for the sinuous movements of the spine.

Hypaxial motor neurons innervate muscles that surround body cavities. These motor neurons are found in the lateral part of the ventral horn. Two dendrite stems originate in the irregular perikaryon. The lateral dendrite arborizes in the lateral white column, and the medial abuts on the midline and arborizes mainly in the gray matter. The hypaxial muscles probably contribute to the sinuous movements of the trunk.

Unlike the two former groups, neurons that innervate limb muscles are confined exclusively at the level of limb segments of the spinal cord, and only these neurons receive monosynaptic innervation form the dorsal roots. Their perikarya cluster in the enlarged dorsolateral part of the ventral horn. Of the three dendrites, a laterally oriented branch expands in the lateral white column. A dorsal branch extends toward the dorsal root and receives the bulk of dorsal root fibers. The third branch is directed medially and terminates on the midline. The presence of these neurons is an absolute prerequisite for the control of coordinated limb movements.

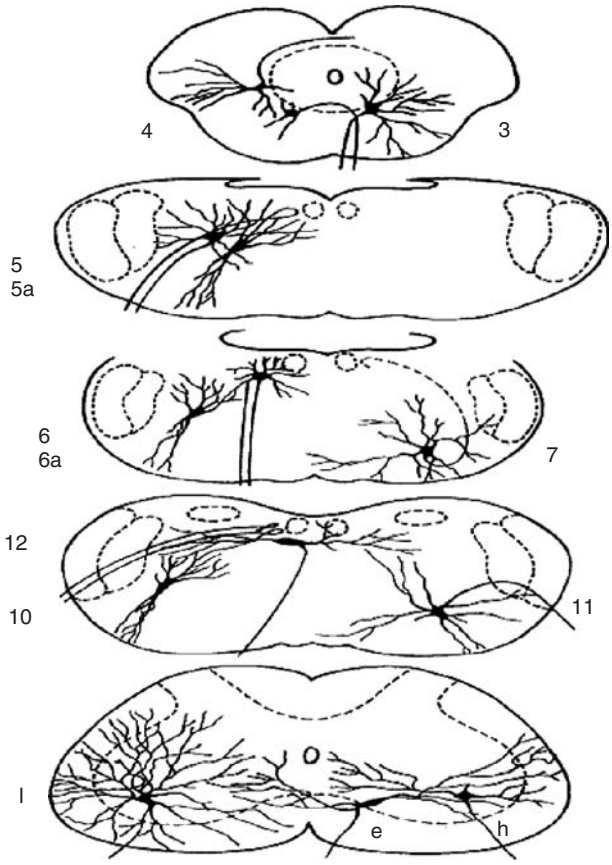

Figure 1 Camera lucida drawings from cobalt-stained specimens, showing cross sections of the brain stem and the spinal cord. Numerals refer to cranial nerve motor nuclei. In the mesencephalon (top section), an oculomotor nucleus with crossing axon is shown on the left side. The nucleus of the accessory nerve nucleus is pushed up into the lower medulla in order to save space. In the spinal cord (bottom section), a limb muscle (l), an epaxial muscle (e), and a hypaxial muscle (h) neuron are shown. Szekely G and Matez C (2004) Cranial nerves – a new perspective and classification. In Adelman G and Smith BH (eds.) The Encyclopedia of Neuroscience, 3rd edn. Oxford: Elsevier.

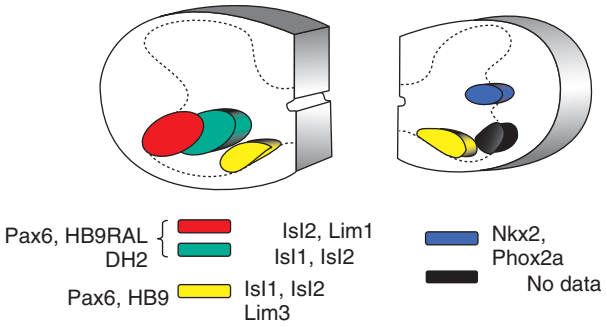

Figure 2 Schematic drawing, color-coded to indicate the genetic profile of the different spinal motor neurons (the gene-related color code is shown on the bottom): yellow, motor neurons with crossing dendrite innervating epaxial limb muscles; black, motor neurons innervating hypaxial muscles; red, the lateral subgroup of motor neurons innervating limb muscles; green, the medial subgroup of motor neurons innervating limb muscles; blue, sympathetic visceromotor neurons. Szekely G and Matez C (2004) Cranial nerves – a new perspective and classification. In Adelman G and Smith BH (eds.) The Encyclopedia of Neuroscience, 3rd edn. Oxford: Elsevier.

In the thoracic section of the spinal cord a group of small neurons forms a continuous column, and the neuronal axons continue their route to the sympathetic ganglia. In the sacral section of the spinal cord another group of small neurons can be found at the dorsal aspect of the ventral horn; these neurons supply the perineal muscles. In mammalian vertebrates there is a group of motor neurons filling up the ventral horn at the midcervical level of the spinal cord. These motor neurons, by their longitudinal dendritic expansion, innervate the diaphragm.

Genetic Profile

Differentiation of progenitor neuroblasts is achieved through the actions of a cascade of transcription factors. In the ventral part of the medullary tube, cells expressing *Pax6* homeobox genes respond to a graded signaling activity of the Sonic hedgehog (Shh) protein and acquire a distinct 'ventral' identity, and begin to express numbers of the LIM homeodomain (HD) family of transcription factors (Figure 2). Progenitors with Lim3 (Lhx3) transcription factor acquire a motor neuron fate by the expression of the NMR2 HDs, and the transient expression of HB9 HD determines the motor neuron fate. Progenitors which leave the cell cycle earliest give rise to the so-called primary motor neurons. They form the medial motor column of type 1 neurons and innervate the epaxial muscles. Longitudinal differentiation of the medullary tube is controlled by the *Hox* homeobox gene family, expressed by the paraxial mesoderm. The Hoxc8 HD identifies the branchial segments, and the *retinaldehyde dehydrogenase 2* (*RALDH2*) gene (current nomenclature, *aldehyde dehydrogenase 1 family, member A2; ALDH1A2*) is the genetic marker of neurons to form the type 3 motor neuron which innervates limb muscles. Following establishment with the periphery, distinct pools are defined by the expression of *PEA3* and *ER81* genes, members of the *ETT* homeobox family. Neurons expressing the corresponding HD proteins encoded by *PEA3* and *ER81* innervate muscles that derive respectively, from the posterior and anterior part of the limb bud. There are no data for the differentiation of type 2 neurons innervating hypaxial muscles.

In the intermediate zone of the medullary tube, the expression of the Phox HD protein marks the identity

of the sympathetic visceromotor neurons at thoracic levels. The genetic history of the sacral parasympathetic neuron group is not known. Unfortunately, motor neurons that are present only in mammals and innervate the diaphragm have not been investigated.

Efferent System of the Brain Stem

Traditionally, two discontinuous motor neuron columns containing the nuclei of cranial nerves are discerned in the brain stem: the somatomotor column, containing hypoglossal (12th), abducens (6th), trochlear (4th), and oculomotor (3rd) nuclei, and the specific visceromotor column, containing accessory (11th), vagus (10th), glossopharyngeal (9th), facial (7th), and trigeminal (5th) motor nuclei (Figure 3). In the forthcoming description, we call these columns the dorsomedial column and the ventrolateral motor column, respectively. This terminology also indicates the positions of the columns in the cross section of the brain stem. Whereas the dorsomedial column is very conservative, disclosing virtually no changes in phylogenesis, the ventrolateral column greatly differs in the nonmammalian and mammalian brain stems. In addition, some satellite neurons and vegetative parasympathetic motor neurons are among the efferent neurons of the brain stem. We describe the cranial nerve motor nuclei according to their function.

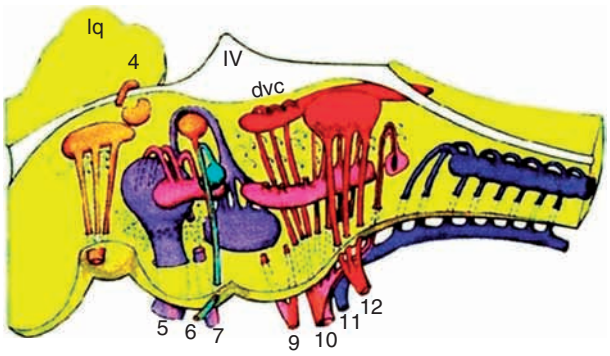

Figure 3 Median-sagittal section of the brain stem, showing the position of cranial nerve motor nuclei in the rat. Some findings obtained from the lizard brain are also incorporated in the scheme. Numbers refer to the corresponding nuclei and nerves. The accessory abducens nucleus is colored light blue; the broken line demarcates the site of the Edinger–Westphal nucleus, and dots with curved-line tails indicate the probable sites of preganglionic neurons of the ciliary ganglion. Similar dots and lines represent the scattered preganglionic parasympathetic neurons in the lateral visceromotor area of the medulla. Small rings in front of the m7th nucleus and ventral to the m5th nucleus represent the neurons of the auditory muscles. A section of the facial root, which would cover these neurons, is removed from this site. In the caudal part of the nucleus ambiguus, a big oval spot represents a neuron which sends the axon through the caudalmost rootlet of the vagus nerve. This rootlet is 'captured' by the ascending root of the accessory nerve, and vagal and accessory fibers travel together for a short while. Where the two nerves (10th and 11th) transiently fuse, all ambiguus fibers join the common trunk of the vagus. A disk is cut out of the nucleus ambiguus in order to reveal the course of the axon. The cerebellum is not shown. dvc, dorsal visceral column of the medulla; Iq, lamina quadrigemia; IV, fourth ventricle. Szekely G and Matez C (2004) Cranial nerves – a new perspective and classification. In Adelman G and Smith BH (eds.) The Encyclopedia of Neuroscience, 3rd edn. Oxford: Elsevier.

Dorsomedial Column

Movement of the tongue: The hypoglossal nucleus (12th nucleus) The motor neurons have spindle-shaped perikarya which emit a lateral and medial dendrite stem (Figure 1). The medial dendrite extends into the contralateral side; the axon exits next to the midline on the ventral surface of the medulla. This general shape can be found in the frog, lizard, and rat and is reminiscent of the medially positioned motor neurons innervating epaxial muscles in the spinal cord. As shown in Figure 4, the genetic profiles of the two types of neurons are also similar.

Coordinated movements of the eyeball (6th, 4th, and 3rd nuclei) The eyeball is moved by the same six nuclei in all of the vertebrate phyla, showing negligible variation from fish to apes. In most vertebrates

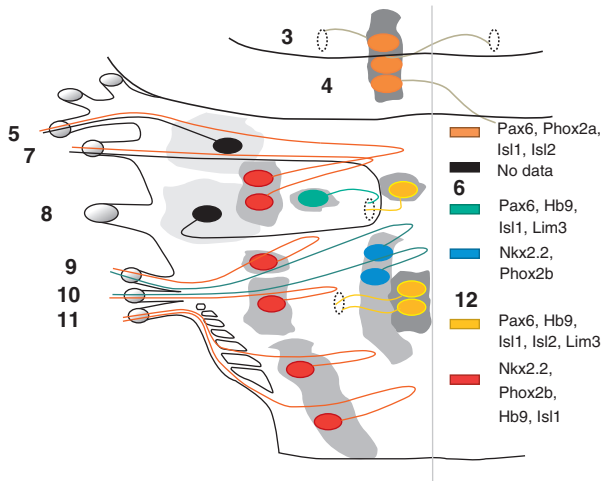

Figure 4 Schematic drawing of the brain stem, showing the genetic profile of the motor nuclei of cranial nerves. Shaded gray areas with decreasing color intensity indicate the approximate extent and position of the respective nuclei labeled with numbers. The darkest gray areas next to the midline indicate the dorsomedial column of the medulla; the next grade of shading is the dorsal visceromotor nucleus, followed by the accessory abducens nucleus; the next is the branchiomotor column, and the two lightest gray areas indicate the main trigeminus and facialis nuclei. Solid circles in the gray areas represent motor neurons. Narrow, dotted-line oval shapes indicate ventral axon exit areas. The gene-related color code is shown on the right. Szekely G and Matez C (2004) Cranial nerves – a new perspective and classification. In Adelman G and Smith BH (eds.) The Encyclopedia of Neuroscience, 3rd edn. Oxford: Elsevier.

there is an additional group of retractor muscles, which are innervated by the accessory abducens nucleus and withdraw the eye into the socket. The abducens nerve originates in two nuclei (Figures 1 and 3). The main nucleus, which innervates the lateral rectus muscle, lies in line with the 12th nucleus. The general shape of the neurons resembles the hypoglossal neurons but there are no crossing dendrites. The axons exit close to the midline on ventral surface. The accessory nucleus (a6) lies lateral and caudal to the main 6th nucleus in lateral reticular formation. The polygonal perikarya give origin to a lateral broomlike and a dorsal fanlike dendritic arborization; the axon arises from the dorsal part of neuron, runs about the main 6th nucleus, and joins the 6th root.

The 4th and 3rd nuclei form a continuous column in the mesencephalic tegmentum (Figure 3). Their dendrites arborize in a fanlike manner (Figure 1). There is a very precise musculotopic organization within this column of neurons. Axons of the rostral part of the column originate in ventral part of the neurons and exit at the ventral surface of the mesencephalon. In the caudal pole of the 3rd nucleus, axons originate in the dorsal part of the neuron, follow a curved course, cross to the other side, and join the contralateral 3rd root. They innervate the medial rectus muscle. Axons in the 4th nucleus also originate in the dorsal aspect of the neurons, follow a curious dorsalward route, and appear on the contralateral side of the dorsal surface of the mesencephalon. These fibers innervate the superior oblique muscle. Note that these two muscles co-contract with the contralateral lateral rectus muscles in a lateral gaze.

Genetic profile The morphology and genetic profile of the six main neurons and the 12th neuron closely resemble each other, but both of these parameters differ in the a6 neurons (Figure 4). Although the course of the axon suggests a later migration of the a6 nucleus from the main 6th nucleus, the differences in morphology and genetic profile do not support a common origin of the two kinds of abducens neurons. In fat, the neurons differentiate in separate loci in the medulla. The retractor muscles and their nerves are absent from Chiroptera (bats) and apes.

After leaving the cell cycle, neurons of the 4th and 3rd nuclei take a different route by expressing the *Phox2a* gene. The Isl2 HD protein is thought to induce the secretion of a protein that guides the outgrowing axon in the ventral direction. One would expect another factor which controls the decussation of axons in the caudal part of the column of the 3rd and 4th nuclei, and even drives one group of axons to the dorsal side of the mesencephalon. The possible candidate responsible for the navigation of the trochlear axons is the Fgf8 protein.

Ventrolateral (Branchiomotor) Column

As shown in Figure 1, the general features of the morphology of motor neurons can be summarized in the following manner: the dendritic arbor displays a dorsoventral orientation, and the ventral array arborizes in a broomlike manner, whereas the dorsal array expands like a fan. The course of the axon is very characteristic: it originates in dorsal aspect of the neuron, takes a dorsomedial course, then turns back, sometimes sharply, as a hairpin, and, passing the ventral aspect of the nucleus of origin, exits in the ventrolateral sulcus of the medulla. Disregarding size differences, this description applies to all motor neurons, including the parasympathetic visceromotor neurons; therefore, this form is regarded as the basic representation of motor neurons in this column.

Control of head movements: The accessory (11th) nucleus Although the alleged precursors of this nerve indicate that it is only an accessory branch of the vagus–glossopharyngeal complex in fishes, the nucleus and the nerve comprise a fully independent structure, innervating a variety of different muscles associated with the skull in different species. The nucleus is at the medullospinal transition level (Figures 1 and 3). The scattered neurons do not form a compact nucleus, but they can be easily recognized by their relatively large size and characteristic form. In mammals, the thin column of the accessory neurons is evident as far as the 6th cervical spinal cord segments.

Deglution and phonation: The nucleus ambiguus (9th and 10th nuclei) The nucleus ambiguus is the common nucleus of the glossopharyngeus and vagus nerve. Its name originates from the difficulty in identifying the location of the components, which are quite variable in different species. We have to deal separately with a central structure controlling phonation. As is well known, laryngeal sound can be produced only by mammals (bird songs are generated in the syrinx, innervated by the 12th nerve). Only the mammalian larynx is provided with sophisticated set of muscles; the nonmammalian larynx possesses an insignificant constrictor muscle. This difference is reflected in the structure of the 10th nucleus (Figures 1 and 3). In submammalian species, the nucleus ambiguus emerges in the reticular formation and terminates almost unnoticed. In mammals, neurons in the caudal part of the nucleus can be easily distinguished by the fact that they grow bigger than

adjacent neurons of reticular formation. Their axons form the most caudal rootlets of the 10th nerve, they transiently fuse with rostralmost rootlets of the 11th nerve, and then separate again and form the recurrent laryngeal nerve. They innervate all laryngeal muscles except one: the cricothyroid muscle. This muscle is supplied from rostral part of the 10th nucleus, which innervates the pharyngeal constrictor muscles. Considering its position and the source of innervation, the cricothyroid muscle is not a real laryngeal muscle. Laryngeal vocalization is therefore a new acquisition in phylogenesis.

Control of jaw movements: The nonmammalian (accessory) facial (7th) and trigeminal (5th) nuclei The spectacular phylogenetic difference between nonmammals and mammals is prominent in the region of the head – the primary mandible joint undergoes very profound transformation during phylogenesis; unfortunately, we have no data about possible transitional forms and can therefore describe only the beginning process and end result. The primary mandible joint acts as the hinge of a two-armed lever. The structural restriction permits only opening and closing the mouth. The participating muscles are the adductor mandible and depressor mandible. The first derives from the primordial constrictor muscle at the level of the first branchial arch, and is innervated by the mandibular branch of the 5th nerve. At the level of second branchial arch, the constrictor splits into a deep and superficial sheet. The deep sheet gives rise to the depressor mandibulae and the sphincter colli. The facial nerve innervates both muscles.

Neurons of the 5th and 7th nerves form a common group in some animals and are separated in others (Figures 1 and 3). The caudal part of the common nucleus is composed of a morphologically homogeneous population of neurons that innervate the opener muscle. The rostral part of the nucleus is more voluminous, containing more neurons; their dendritic arborization does not present as regularly as do the characteristic morphological features of branchiomotor neurons, as can be seen in the facial nucleus.

Control of mastication and facial expression: The mammalian (main) trigeminal (5th) and facial (7th) nuclei The mammalian 5th and 7th nuclei contain two types of neurons (Figure 1). One type is represented by the accessory a5 and a7 neuron types – neurons that innervate the depressor mandibulae in nonmammalian vertebrates. These neurons have preserved their original form, the dendritic arborization pattern, and their course of axons, but only the most rostral neurons form a conglomerate; many of them are scattered along a ventrolateral line between the well-defined neuron groups of the mammalian 5th and 7th nuclei (Figure 3). The rostralmost of these neurons send their axons via the trigeminal motor root; the axons of neurons in the caudal part of this line of neurons join the facial nerve and they innervate the suprahyoid muscles.

The secondary mandibular joint began to evolve probably with the appearance of carnivorous mammal-like reptiles. It has been conventionally assumed that the muscles derive from muscle mass that gives rise to the adductor mandibulae in lower vertebrates, but it has now been observed that the radical remodeling that occurs in mammals makes impossible the homologization of individual mammalian masticatory muscles to the derivatives of the primordial adductor muscle mass. Parallel to the transformation of the joints, the teeth are fastened via gomphosis in the alveoli of the upper and lower jaw. The complex arrangements of the muscles and their new type of innervation have rendered it possible to grind and masticate food. Joints, muscles, and nerve are novel acquisition, in evolution.

Muscles of facial expression are also mammalian attributes. Their evolutionary course is not known. In the old literature, the constrictor colli is regarded as the primordium of the facial muscles. this is improbable, because the constrictor is innervated by accessory (a7) motor neurons. Large, polygonal neurons of the facial muscles form a well-defined nucleus in the ventrolateral column. The axons originate in the ventral aspect of the perikaryon; after a short run they turn dorsal, and at the lateral aspect of the 6th nucleus they form a whirl-like structure. This tangle of fibers extends rostrally and dives down to the exit site, forming the enigmatic genu of the facial nerve. The facial muscles and nerve can also be regarded as novel structures.

Satellite neurons to the facial–trigeminal nuclear complex In the frog, there is a muscle sheet, the orbital muscle, which separates the orbit from the snout cavity. It is innervated by a group of neurons in the upper pole of the 5th nucleus, and the axons reach the muscle by way of the ophthalmic and maxillary nerves. In the mammalian brain stem a loosely arranged group of small neurons can be found ventral to the 5th and 7th nuclei (Figure 3). Their axons join either the 5th nerve (musculus tensor tympani) or the 7th nerve (musculus stapedius). These tensor tympani and stapedius muscles probably evolved together with the transformation of the mandibular joint.

Vegetative parasympathetic column A group of small cells form a continuous column lateral to the

dorsomedial nuclei from the caudal end of the 7th nucleus till the rostral part of spinal cord. This group of cells is the dorsal nucleus of vagus (Figure 3). Axons emerging from the dorsal aspect of the small perikarya run in the direction of longitudinal medial fascicle, then after sharp turn they join the roots of the 7th, 8th, and 9th nerves. At the level of the ambiguus nucleus, there are a number of middle-sized neurons scattered in lateral part of the reticular formation. Their axons take a similar course before joining the 9th and 10th nerve roots. This group of neurons is called the medullary vegetative column.

Genetic Profile

The basic difference between the dorsomedial and ventrolateral column is that the *Pax6* gene interprets the induction of the Shh HD protein in the first column, whereas we find the Nkx2 homeobox family in the second column (Figure 4). The Phox2 protein determines the route of axons, and a subsequent HB9 effect defines their motor neuron identity. The LIM HD family is represented by the Isl1 HD. It is interesting that the axons of all branchiomotor and visceromotor neurons take the characteristic hairpin loop course and exit dorsolaterally. This may suggest that the initial differentiation of the two classes of neurons controlled by the *Nkx2.2* and *Phox2b* genes determines this axonal route as well. Regarding the most variable morphology and function controlled by the branchiomotor column, one would expect some genetic differences in the backgrounds of these morphological and functional variations.

Unfortunately, there are no data about the 'new' neurons in the class of branchiomotor neurons. It would be especially interesting to follow the phylogenetic appearance of the main trigeminal nucleus and the transformation of the primary mandibular joint, and how it is associated with the development of the middle ear, its muscles, and their innervation. The same consideration applies to the 'new' facial nucleus, which is different in its structure from the other motor nuclei, and to the muscles of facial expression, which are also different from other skeletal muscles.

In summary, the foregoing samples present a number of motor neuron types that can be classified on the basis of morphological characteristics. Well-defined functions can be assigned to major classes in a number of cases. The capacity to perform some of these functions is determined in early embryonic life. The *RALDH2* (*ALDH1A2*) gene distinguishes motor neurons in the lateral part of the ventral horn, and the variously expressed Lim proteins continue to subdivide this pool of neurons. The *Hox* homeobox genes identify the limb segments of the spinal cord, and presumably trigger the sequence of gene expression that determines the 'limb' fate of motor neurons, which are alone capable of moving the limb in a coordinated manner. HDs of the *ETT* homeobox family supposedly guide motor fibers to the appropriate muscle group. The instructive contribution of the somatodendritic morphology to the neural implementation of patterned movements (e.g., limb, jaw) verifies the interest in their differentiation and raises the question of the extent to which this differentiation is controlled by genetic and epigenetic mechanisms. Recently, the instrumental role of dendritic geometry was emphasized in the organization of neuronal interactions. In this context, motor neurons that can be identified in the embryo provide excellent material for the study of the development of neural morphology, and they may help answer questions concerning the role of genetic and epigenetic factors in the shaping of neural organization.

Afferent System of the Brain Stem

Sensory fibers are carried only by the branchial cranial nerves (Figure 5). The trigeminal nerve subserves the exteroceptive innervation of most of the head. Its sensory nuclei constitute a long column that continues into the dorsal horn of the spinal cord. This is called the general somatic afferent column in the

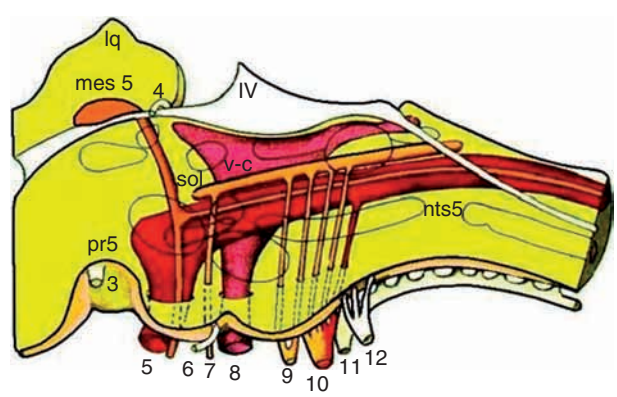

Figure 5 Cranial nerve sensory nuclei are shown in the same outline of the brain stem as in Figure 3. For a better orientation, the motor nuclei are indicated with thin lines. lq, lamina quadrigemia; mes5, the mesencephalic sensory nucleus of the 5th nerve; IV, fourth ventricle; pr5, the principal sensory nucleus of the 5th nerve; nts5, the nucleus of the spinal tract of the 5th nerve; sol, the solitary tract and nuclear complex; v-c, the vestibulocochlear complex. Szekely G and Matez C (2004) Cranial nerves – a new perspective and classification. In Adelman G and Smith BH (eds.) The Encyclopedia of Neuroscience, 3rd edn. Oxford: Elsevier.

classic scheme. However, this nerve supplies the mucous membranes of the mouth, tongue, and nasal cavity, and the pulp of the teeth; this is a general viscerosensory function in the classic scheme. Furthermore, from its mesencephalic nucleus, this nerve provides the proprioceptive innervation to masticatory muscles (which belong to the specific visceromotor group), to eye muscles (which are innervated by another type of motor nuclei), and to the periodontal connective tissue. In many animals, the snout and the whiskers function as real sense organs. They are also innervated by the trigeminal nerve, with specific thalamic and cortical representation. These fibers should be classified in a common group with the 8th nerve, which is the sensory nerve of another sense organ in the hindbrain region, and should be referred to as specific somatosensory fibers. Thus the trigeminal nerve contains all possible types of sensory fibers. The classic scheme breaks down in the trigeminal sensory field as well.

Sensory fibers of the 7th, 9th, and 10th nerves terminate in a general visceral afferent column and in a specific visceral afferent column, according to the classic scheme. The general visceral afferent column corresponds to the solitary tract and its nuclear complex (**Figure 5**). The specific visceral column, the alleged center of gustatory sensation, supposedly lies medial to the solitary complex. Modern tracing techniques have failed to detect this specific visceral afferent center, and the area medial to the solitary complex seems to be involved in vestibulocerebellar connections. Specific and general visceral sensations, therefore, converge onto a common center. There is also an extensive interchange of fibers between the trigeminal and solitary sensory nuclei in the frog, lizard, and rat; this suggests that interoceptive and exteroceptive impulses are jointly processed in the trigeminal and solitary nuclear complexes. The two kinds of afferent impulses are mixed to an unknown extent in these centers. The most lateral part of the hindbrain harbors sense organ nuclei for hearing and equilibration, and for the lateral line organ in aquatic vertebrates.

See also: Brain Asymmetry: Evolution; Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brainstem Control of Eye Movements; Corticospinal Development; Forebrain: Early Development; Midbrain Patterning; Neural Patterning: Midbrain–Hindbrain Boundary; Primate Interneurons; Single Photon Emission Computed Tomography (SPECT); Spinal Motor Neurons: Properties; Vestibular Influences on Cognition; Vestibular System.

Further Reading

Arber S, Han B, Mendelsohn M, et al. (1999) Requirement for the homeobox gene Hb9 in the consolidation of motor neuron identity. *Neuron* 23: 659–674.

Briscoe J, Sussel L, Serup P, et al. (1999) Homeobox gene Nkx2.2 and specification of neuronal identity by graded Sonic hedgehog signaling. *Nature* 398: 622–627.

Cordes SP (2001) Molecular genetics of cranial nerve development in mouse. *Nature Reviews Neuroscience* 9: 611–623.

Jurata LW, Thomas JB, and Pfaff SL (2000) Transcriptional mechanisms in the development of motor control. *Current Opinion in Neurobiology* 10: 72–79.

Kury P, Gale N, Connor R, et al. (2000) Eph receptors and ephrin expression in cranial motor neurons and the branchial arches of the chick embryo. *Molecular and Cellular Neuroscience* 15: 123–140.

Lance-Jones C, Omelchenko N, Bailis A, et al. (2001) Hoxd10 induction and regionalization in the developing lumbosacral spinal cord. *Development* 128: 2255–2268.

Matesz C and Szekely G (1996) The somatotopic and cytoarchitectonic organization of the ambiguus nucleus in the frog. *Journal of Comparative Neurology* 371: 258–269.

Matesz C, Birinyi A, Kothalawala DS, et al. (1995) Investigation of the dendritic geometry of brain stem motoneurons with different functions using multivariant multistatistical techniques in the frog. *Neuroscience* 65: 1129–1144.

Pattyn A, Hirsch M, Goridis C, et al. (2000) Control of hindbrain motor neuron differentiation by the homeobox gene Phox2b. *Development* 127: 1349–1358.

Starck D (1982) *Vergleichende Anatomie der Wirbeltiere auf evolutionsbiologischer Grundlage*, vol. 3. Heidelberg: Springer.

Straka H, Baker R, and Gilland E (2006) Preservation of segmental hindbrain organization in adult frogs. *Journal of Comparative Neurology* 494: 228–245.

Székely G (1976) The morphology of motoneurons and dorsal root fibers in the frog's spinal cord. *Brain Research* 103: 275–290.

Székely G (2001) An approach of the complexity of the brain. *Brain Research Bulletin* 55: 11–28.

Székely G and Matesz C (1993) The efferent system of cranial nerve nuclei: A comparative neuromorphological study. *Advances in Anatomy Embryology and Cell Biology* 128: 1–92.

Tsuchida T, Ensini M, Morton SB, et al. (1994) Topographic organization of embryonic motor neurons defined by expression of LIM homeobox genes. *Cell* 79: 957–970.

Brainstem Control of Eye Movements

E L Keller, Smith-Kettlewell Eye Research Institute, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Saccadic Eye Movements and Fixation

Saccades are rapid, steplike shifts of the orientation of the eyes that serve to reposition the line of sight on a new object of interest. They are among the most rapid and precisely controlled form of motor behavior in animals with foveal or heightened central visual acuity. Because visual acuity is depressed during saccades, the neural circuits that control these eye movements have evolved to generate extremely high rotational velocities (up to $800^\circ\,s^{-1}$ or more in primates) and accurate terminations that place the fovea on or near the new target of interest. These characteristics of saccades produce movements of short duration (e.g., 50–100 ms) in which duration is a fixed function of saccade amplitude. Once a saccadic movement ends, an active fixation system is engaged that prevents further movement of the eyes for a short period of time as the visual system inspects the newly acquired target. The saccadic system, therefore, operates in a discontinuous fashion, alternating between periods of fixation and high-speed movement. Saccades that overshoot or undershoot the target, that drift onto the target, or that occur with degraded speeds or drifts in eye position between saccades are sensitive indicators of pathology in the saccadic and fixation control system. Neurologists have used such measures of ocular motility as indicators of saccadic system malfunction and as indicators of the probable location of lesions in the brain that produced the malfunctions.

Organization of Brain Stem Circuits Involved in Saccades

The precision with which saccadic eye movements are controlled suggests that feedback about the ongoing eye movement is used to control the trajectory and endpoint of these movements. However, visual feedback from the retina is too slow to affect the movement accuracy during saccades due to their very short duration. A major advance in understanding the saccadic system was provided by neurophysiological studies that showed that saccades are not ballistically guided but, indeed, are controlled by feedback during the movement. This feedback is derived, not from the visual system but, instead, from an efference copy of the neural commands controlling the motor output. Because this source of feedback is derived from the motor system output, it has extremely short latency and can thus control the accuracy of saccades during the movement in almost real time. The ability of the saccadic system to use efference signals to control saccadic accuracy stems from one of its unusual characteristics when compared to other motor systems. The load moved by the system is the eyeball and other orbital tissues. This load is nearly invariant over all sizes and directions of movement, and saccades are simple rotations about a single, relatively fixed point. In addition, proprioceptive feedback signals from stretch receptors in the extraocular muscles do not directly control saccades, but these signals may be involved in longer-term adaptive control of eye movements. Neural control signals directly from receptors in the eye muscles to oculomotor neurons, which is called afferent feedback, is much more ubiquitous in other skeletal motor systems. Proprioceptive input from the eye muscles may also be used to calibrate the efferent copy during development.

The concept that saccades are controlled by neural feedback signals originating from the motor (efferent) output has led to the general model of saccadic system organization shown in Figure 1(a). This model is a simple, one-dimensional representation of the negative feedback system that produces horizontal rotations of the eyes. Signal development and processing are represented with single input/output operators that summarize the neural mechanisms involved. Descending control signals from sources outside the brain stem provide the desired goal of the saccade. This signal is indicated as E_d in Figure 1(a). The new eye position or goal signal originates from the cerebral cortex, cerebellum, and superior colliculus (SC). Although the SC is anatomically part of the brain stem, its operation is not discussed here, except to note that it forms part of the distributed input to the brain stem circuits generating eye-movement control. The spatial frame of reference of E_d and how it is coded are still a matter of discussion among investigators of the saccadic system. Frame of reference means whether the desired rotation of the eyes is specified in retinotopic or spatial coordinates. Methods of coding means whether E_d is generated by a topographically organized activation of a population of neurons (place coding) or by a frequency code in which the rate of neural discharge codes the desired goal or saccadic end point (rate coding). Studies of the cortical and collicular sources of E_d suggest that this input signal is retinotopically coded by a distributed place-coded ensemble of activity. The cerebellar

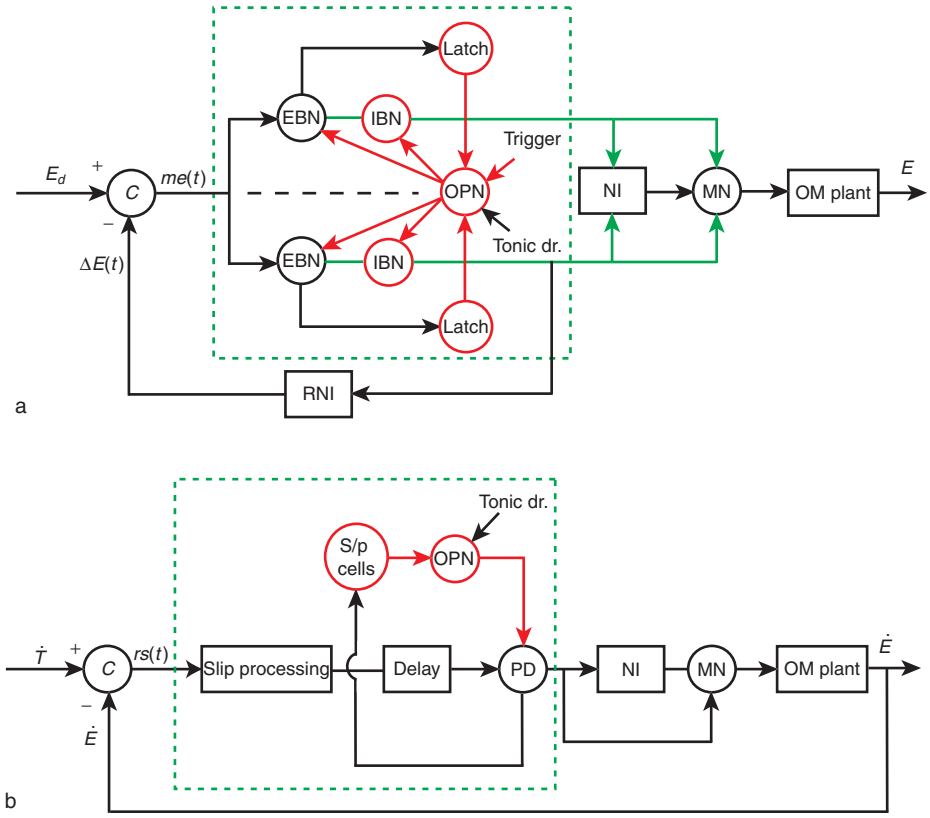

Figure 1 Schematic view of the brain stem organization of (a) the saccadic system and (b) the smooth pursuit system. Black circles, rectangles, and pathways indicate excitatory operations. Red circles and pathways indicate inhibitory operations. Arrowheads indicate the direction of control signal flow. In (a), the efferent control of saccade trajectory is generated by the lower feedback pathway. The green dotted-line box is the saccadic burst generator. The dashed horizontal line indicates that the system is divided into a push/pull fashion for left/right directed horizontal saccades or into up/down vertical saccades. In (b), the afferent control of smooth pursuit eye movements is accomplished by the generation of a retinal slip signal, the motion difference signal between the moving target and the moving eye and attached sensor (the retina). The system operates continuously to minimize retinal slip. The green dotted-line box is the pursuit generator. *C*, comparator; *E*, eye position; $\dot{E}$, eye velocity; E_d, desired shift in eye position; $\Delta E(t)$, current eye displacement during a saccade; EBN, excitatory burst neuron; IBN, inhibitory burst neuron; *me*(*t*), dynamic motor error signal; MN, motor neurons; NI, neural integrator; OM plant, orbital mechanical system; OPN, omnipause neurons; PD, pursuit drive neurons; RNI, resettable neural integrator; *rs*(*t*), dynamic retinal slip signal; S/p cells, saccade/pursuit neurons; $\dot{T}$, target velocity; Tonic dr, tonic drive.

contribution to E_d may be a mixed rate/place-coded signal. The distributed input from the SC connects to the brain stem burst generator (the green dashed box in Figure 1(a)), whose cells are rate-coded. Thus, a spatial-to-temporal transform must occur at this level of the system. However, the SC connects to burst-generator cells with a pattern of variable synaptic strengths (weights). Those inputs from the rostral SC discharge for small saccades and connect to the burst-generator cells with small weights, whereas those from the caudal SC discharge for large saccades and connect with large weights. This arrangement may produce the spatial-to-temporal transform required in the saccadic system, in which the same ensemble input from the rostral or caudal SC produces a smaller or larger rate of discharge in burst-generator cells for smaller or larger saccades, respectively.

The Neural Comparator

The efferent feedback signal coding the current displacement of the eyes (with respect to their starting position) during a saccade is subtracted from the desired goal (E_d) of the movement to generate a dynamic motor error signal (*me*(*t*)). In Figure 1(a), this computational step is symbolized with standard linear system operators as a lumped subtractive operator or comparator (*C*). The use of this linear system operator or comparator suggests that neurons carrying a discharge coding eye displacement ($\Delta E(t)$) inhibit a single group of neurons that also receive an excitatory input coding the desired goal. The resulting output of the group of neurons is the motor error signal. It is now clear that the actual operation of this part of the system is much more complicated and that the comparator merely serves as a short-hand symbol

to represent the theoretical operations that the saccadic system must perform to insure the accuracy of the saccadic endpoint through feedback control. In the first place, the descending input to the comparator is a distributed signal representing the goal with a spatial population code, not a rate-coded signal. Furthermore, saccades are actually three-dimensional rotations of the eye; however, attempts to extend the model shown in Figure 1(a) to three-dimensional rotations (by adding a second circuit that represents vertical movements) have not been successful because of the kinematic laws governing three-dimension rotations. The results of these complications means that a literal rate-coded motor error signal may not exist in any one group of brain stem neurons but that the driving signal representing error may instead be a highly distributed signal having both place and rate coding. Nevertheless, the system depicted in Figure 1(a) remains a viable way of framing the discussion of the neurophysiology and connectivity of the various types of brain stem neurons that control saccades.

Saccadic Burst Generator

The saccadic burst generator (SBG) is a brain stem circuit consisting of anatomically separated groups of neurons with different firing characteristics that collectively control saccade direction and amplitude. Because the orbital mechanical system (OM plant) is heavily dominated by viscous forces, the muscle forces that generate saccades must provide a large pulse of force to the rotating eyeball for the duration of the movement. The high-velocity ocular rotations reached during saccades amplifies the need for the large pulse of muscle force, and the essence of saccadic control is to remove this pulse of force just before the eyes reach the desired saccadic goal (i.e., when $me(t) = 0$). This pulse of force is generated by a burst of motor neuron activity in the agonist muscles that leads to saccade movement onset and lasts for the duration of the saccade. The motor neuron activity supplying the antagonist muscles shows the reciprocal pattern of activity – a complete inhibition of discharge during the saccade. This reciprocally organized burst/pause of motor neuron activity during saccades is generated by the output of the SBG. The reciprocal nature of the functional organization of the SBG, which involves ipsilateral excitatory pathways and inhibitory pathways that cross the midline of the brain stem is simplified in Figure 1(a) with the use of the solid green lines into a single group of motor neurons.

The SBG consists of burst neurons and omnipause neurons (OPNs). The former are further differentiated into excitatory burst neurons (EBNs) and inhibitory burst neurons (IBNs). Anatomically, EBNs are located in two regions, the paramedian pontine reticular formation (PPRF), just rostral to the abducens nucleus (AN), and a region of the mesencephalic reticular formation called the rostral interstitial nucleus of the medial longitudinal fasciculus (riMLF). The IBNs are located in the medullary reticular formation just caudal to the AN and are intermixed with EBNs in the riMLF. The caudal groups of burst neurons are responsible for the horizontal component of saccades, whereas the rostral burst neurons subserve the vertical component of saccades. Both EBNs and IBNs produce a high-frequency burst of activity that that slightly leads (by ~10 ms) the onset of saccades, and this burst of activity ends before saccade offset. The discharge rate of the burst neurons is related nonlinearly to instantaneous saccadic velocity through a saturating exponential function. The more caudal group of EBNs project monosynaptically to the ipsilateral AN, whereas the caudal IBNs project across the midline to the opposite AN. This pattern of projections ensures the reciprocally organized discharge of ANs for horizontal saccades; it is indicated by the green solid lines in Figure 1(a). The intermixed rostral EBNS and IBNs project to the oculomotor and trochlear nuclei. The timing of the burst in burst neurons determines the timing and duration of the burst or pause in motor neuron discharge through their direct connections to the motor neurons (green arrows). As their name implies, burst neurons do not discharge during fixation.

The other functionally distinct group of neurons in the SBG is the OPNs. These cells are located in thin bilaterally placed nuclei located on the midline of the brain stem, just rostral to the AN. They discharge tonically at a high constant rate (~100 sps) during fixation, but they pause completely before and during saccades in all directions. The OPNs project monosynaptically to and inhibit all the groups of burst neurons, both EBNs and IBNs. The high discharge rate of OPNs and their inhibitory connections to burst neurons silences these latter neurons during fixation, thus preventing unwanted saccades. Earlier studies suggested that the OPNs precisely gated the timing of the burst in EBNs and IBNs, a control function which would determine saccade amplitude. More recent studies showed that the resumption of discharge in OPNs near the saccade end may not be precise enough to terminate the burst in EBNs and IBNs by itself. These studies suggested, instead, that other neurons, perhaps IBNs, may play a role in the endpoint control of saccades.

Intracelluar studies in OPNs during saccades show that the pause in OPN discharge is caused by

a pulselike hyperpolarization of the OPN membrane just before saccade onset that rapidly silences their spike discharge. Following this transient sharp hyperpolarization of the OPNs, the trajectory of the membrane hyperpolarization closely parallels the dynamic shape of the ongoing saccadic eye velocity. These observations suggest that OPNs receive two signals during saccades: a short presaccadic inhibitory input, called the trigger, that silences them and a more prolonged inhibitory input, called the latch, that holds off OPN discharge for the duration of the saccade. The trigger signal initiates saccades by turning off OPNs and, thus, disinhibiting burst neurons, whereas the latch signal maintains the pause in the OPNs for the duration of the saccade. Microelectrode recordings have not yet identified the cell groups or the locations of the hypothesized trigger and latch neurons. Finally, because OPNs are active when the eyes are fixed between saccades, there must be an input tonic drive signal to these cells. Alternatively, the spontaneous discharge of OPNs could be an endogenous property of this type of cell. Most likely, the tonic discharge of OPNs stems from an exogenous input because the high-frequency discharge in these cells ceases when monkeys become less alert.

The Neural Integrator

During fixation, oculomotor neurons discharge at a steady rate proportional to eye position (e.g., AN motor neurons increase their discharge as a function of the abducted position of the innervated eye). The neural circuit that generates this eye position signal in motor neurons is called the neural integrator (NI). The input signal to the integrator comes from the EBNs, neurons which discharge with a signal proportional to eye velocity during saccades. The function of the integrator is to integrate the eye velocity signal from EBNs, in a mathematical sense, to provide an appropriate eye-position signal to the motor neurons. The eye-position signal in motor neurons generates muscle force to hold the eye steady during fixation against elastic restoring forces in the orbit that would otherwise return the eye to the straight-ahead position between saccades. The NI is located in several brain stem nuclei, including the vestibular nuclei and the nucleus prepositus hypoglossi, but circuits through the cerebellum also contribute to its operation.

The Resettable NI

The feedback pathway to the comparator consists of a highly distributed set of pathways that include connections from the burst neurons to the SC and the cerebellum. Because both the latter structures also contribute to the feed-forward signal, E_d, it is clear that the saccadic control system is not a standard negative feedback system that directly computes motor error as depicted in Figure 1(a). Instead, this schematic serves as an organizing concept to guide and clarify anatomical and neurophysiological data and to emphasize that saccades are not generated as ballistic movements but are controlled by internal feedback. Overall, the signal generated by the feedback pathways is a dynamically changing representation of current eye placement (change in eye position), not eye position, which is the output of the NI. This is schematically depicted in Figure 1(a) by the inclusion of a separate resettable NI in the feedback pathway, which generates an eye displacement signal. Studies have shown the time constant for the reset of this NI must be extremely short. The cells and pathways that contribute to the generation of the eye displacement signal have not been determined. Furthermore, eye velocity signals from efference copy may also contribute to the overall feedback pathways.

Smooth Pursuit Eye Movements

The brain stem mechanisms that control smooth pursuit eye movements have not been specified with the clarity that has emerged for saccades. The control of pursuit movements is fundamentally different from saccades because they continue over a lengthy time course that allows the feedback of afferent signals from the retina to contribute to the control of the movement. Furthermore, the early phases of pursuit tracking of a target that suddenly starts to move are open loop (not controlled by feedback at all). The neural elements located near the pursuit system output involve the cerebellum and the vestibular nuclei, but also involve some of the same brain stem elements that were traditionally assigned exclusively to the saccadic system. Figure 1(b) shows the brain stem organization of the pursuit system.

Target Motion Inputs and the Formation of Retinal Slip Signals

The cortical signals specifying target motion ($\dot{T}$) are directed through the pontine nuclei and the nucleus reticularis tegmenti pontus, both brain stem structures. The colliculus also contributes to these distributed inputs, although its signals may specify a position motor error, not a motion signal. The same structures contribute to the formation of a retinal slip (*rs*) signal, which is the error signal driving steady-state smooth tracking eye movements. Slip signals are derived from information about target motion and the feedback of the physical motion of the eye ($\dot{E}$) (because the sensor, the retina, is attached to the moving eye). Just as is the case for the saccadic

system, these observations mean that the pursuit system is not actually a standard negative feedback system and that the signal comparator shown in Figure 1(b) is actually a highly distributed operation involving many brain structures.

The Pursuit Generator

The pursuit system is subject to the mathematical laws of feedback control even if the feedback is highly distributed among many brain structures. There is a long delay (on the order of 100 ms) in the pathways that generate the slip signals and the forward pathways in the pursuit generator. This delay and the high gain present in the forward pathways means that this feedback-controlled system would be hopelessly unstable without compensatory signal processing pathways. Eye velocity would oscillate with growing excursions about the actual velocity of the intended target. Although small, damped oscillations are often observed in the transient pursuit response to a suddenly moving target, the steady-state eye velocity response after a few hundred milliseconds is very stable and closely matches target velocity. This requires that the pursuit generator contain mechanisms that overcome the potential delay-caused instability. These mechanisms may include the use of internal (efferent) feedback about eye velocity (as in the saccadic system), the use of anticipation and other cognitive signals derived from raw target motion signals, and the use of slip acceleration signals in the feedback pathways. Many of these hypothesized higher-level signals involve cortical and cerebellar processing, but the brain stem may also make a contribution to their generation. This contribution is shown schematically in Figure 1(b) as the box labeled slip processing.

Brain Stem Elements in the Pursuit System

The same NI used in the saccadic system to generate steady eye position may be shared by the pursuit system. However, some studies suggest that these NI circuits may be partially separated in the brain stem. The burst neurons, the EBNs and IBNs in the SBG, do not participate in pursuit eye movements.

It was originally assumed that OPNs also did not participate in pursuit movements, but recently it has been determined that OPNs do participate in pursuit movements. Instead of being fully on during fixation and completely off during eye movements (as they are for saccades), OPNs modulate their discharge downward during pursuit and then resume normal levels of activity at the end of the pursuit movement. Because OPNs project inhibition monosynaptically to both the rostral and caudal groups of burst neurons, the downward modulation in OPN discharge significantly lowers the ongoing level of inhibition in these burst neurons at the onset of pursuit. However, they do not begin to discharge because they do not have an error signal present as they do at the onset of saccades (see Figure 1(a)). That the lowered level of OPN discharge present during pursuit movements contributes functionally to the generation of higher pursuit eye velocity has been shown with electrical microstimulation of OPNs. A train of stimulation pulses delivered to the OPNs during ongoing steady-state pursuit turns the OPNs back on fully, which produces an immediate reduction in the pursuit velocity that lasts for the duration of the stimulation train. During stimulation, pursuit velocity did not fall to zero, indicating that OPN modulation only acts as a gain control on pursuit, not as an on/off switch as it does for saccades.

The observation that OPNs are not fully turned off at pursuit onset but, instead, are more gradually turned down in a fashion that resembles increasing eye velocity, suggests that during pursuit no inhibitory trigger signal exists in the pursuit system. In the saccadic system, this signal to OPNs serves to completely and sharply turn off these cells at saccade onset.

Saccade/Pursuit Neurons

Saccade/pursuit neurons (S/p cells) are found near and sometimes intermixed with EBNs. These cells have weaker bursts during saccades than EBNs and produce a low level of discharge that is sustained for the duration of the pursuit. Some also show a gradual buildup of discharge before the onset of both saccades and pursuit, in contrast to EBNs, which show an abrupt initiation of high-frequency discharge just before saccade onset. In this feature, they resemble long-lead saccadic burst neurons that are found scattered throughout both the pontine and mesencephalic reticular formation.

The trajectory of the discharge modulation in S/p cells for pursuit movements roughly parallels the increasing eye velocity during pursuit initiation. Thus, S/p cells have a discharge during pursuit that is the mirror image of that of OPNs. To explain this relationship, Figure 1(b) suggests that S/p cells inhibit OPNs and are the source that turns down, but does not totally silence, OPN discharge during pursuit. OPNs, in turn, inhibit both saccadic burst cells and pursuit driving cells (PDs), the latter found in the vestibular nuclei and the cerebellum. Thus, S/p cells have many of the same functional connections as the latch neurons of the saccadic system and may be shared by the two systems.

This brief summary of the brain stem circuits involved in pursuit movements suggests that many

of the same types of neurons are shared by the pursuit and saccadic systems. This claim is supported by other studies that show that some input signals to the brain stem from the rostral SC specify only that a position motor error exists. Whether this error signal generates a saccade, a pursuit movement, or a combination of both is determined by decision processes at the level of the cerebellum and brain stem.

See also: Brainstem and Cranial Nerves; Cerebellum and Oculomotor Control; Cortical Control of Eye Movements; Eye Movement Disorders; Oculomotor Control: Anatomical Pathways; Oculomotor System: Models; Pursuit Eye Movements; Saccade–Pursuit Interactions; Saccadic Eye Movements; Superior Colliculus; Target Selection for Pursuit and Saccades.

Further Reading

Heinen SJ and Keller EL (2004) Smooth pursuit eye movements: Recent advances. In: Chulupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 1402–1414. Cambridge, MA: MIT Press.

Hepp K, Henn V, Vilis T, and Cohen B (1989) Brainstem regions related to saccade generation. In: Wurtz RH and Goldberg ME (eds.) *The Neurobiology of Saccadic Eye Movements*, pp. 205–212. Amsterdam: Elsevier.

Keller EL (1991) The brainstem. In: Carpenter RHS (ed.) *Vision and Visual Dysfunction*, pp. 200–223. London: Macmillan.

Keller EL (2004) Distributed neural processing in the saccadic system. In: Hall WC and Moschovakis AK (eds.) *The Superior Colliculus: New Approaches for Studying Sensorimotor Integration*, pp. 277–301. Boca Raton, FL: CRC Press.

Krauzlis RJ (2004) Recasting the smooth pursuit eye movement system. *Journal of Neurophysiology* 91: 591–603.

Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University Press.

Lisberger SG, Morris E, and Tychsen L (1987) Visual motion processing and sensory–motor integration for smooth pursuit eye movements. *Annual Review of Neuroscience* 10: 97–129.

Missal M and Keller EL (2003) Shared brainstem pathways for saccades and smooth-pursuit eye movements. *Annals of New York Academy of Sciences* 1004: 29–39.

Scudder CA, Kaneko CRS, and Fuchs AF (2002) The brainstem burst generator for saccadic eye movements: A modern synthesis. *Experimental Brain Research* 142: 439–462.

VanGisbergen JAM, Robinson DA, and Gielen S (1981) A quantitative analysis of generation of saccadic eye movements by burst neurons. *Journal of Neurophysiology* 45: 417–442.

Brainstem Respiratory Circuits

J L Feldman, University of California at Los Angeles, Los Angeles, CA, USA
D R McCrimmon, Northwestern University, Chicago, IL, USA
G S Mitchell, University of Wisconsin, Madison, WI, USA
C A Del Negro, College of William and Mary, Williamsburg, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Breathing, the process of moving air through the mouth, nose, and airways and into and out of the lung, is an essential behavior for life. Breathing plays a critical role in several fundamental domains of mammalian nervous system function, including regulatory, volitional, and emotional domains.

Volition

Breathing is a highly automated behavior but with significant voluntary control. At least for short periods, we can alter our breathing by holding our breath or by consciously breathing faster or slower. The precise control of expiratory airflow over the vocal folds underlies phonation, including speech, in which both cognitive and emotional mental processes are integrated in a complex motor pattern.

Emotion

Many emotional behaviors, such as laughing, sighing, and crying, involve stereotyped breathing movements. These are evoked independently of volitional movements and use neural pathways of an emotional motor system that are located more laterally in the brain stem than those effecting voluntary breathing movements. Extensive lesions in the ventromedial pons result in a locked-in syndrome in which patients are quadriplegic and speechless but regulatory breathing movements appear normal because the essential neural circuits, located more caudally in the medulla and spinal cord, are typically unaffected. Their breathing is not altered by acknowledged requests to hold their breath or breathe faster, but breathing patterns can show a marked change in response to a (funny) joke, consistent with the stereotyped 'laughing' breathing pattern, that is, episodic and prolonged expiratory efforts. Breathing disturbances can accompany mental states with strong emotional components, such as during panic attacks. In patients with the genetic disorder Rett's syndrome, pronounced disturbances in breathing occur during stress, and these disturbances are thought to be of cortical origin.

Regulation

The most essential role of the respiratory control system is to maintain homeostasis, regulating breathing to optimize oxygen and carbon dioxide exchange. Breathing is sometimes erroneously referred to as an 'autonomic' function, that is, a regulatory act mediated by smooth muscle and glands. Although airway caliber can be modulated by airway smooth muscle, the essential role of skeletal muscles in the respiratory pump excludes breathing from this categorization. Aerobic metabolism consumes O_2 and produces CO_2. In the lung, these gases are exchanged by passive diffusion at the alveolar–capillary interface, which in humans has a surface area of $\sim 70\,m^2$. To move O_2 in and CO_2 out, the lung expands and contracts, driven by skeletal muscles of the respiratory pump. The critical central circuitry for the regulatory control of breathing is in the brain stem and spinal cord. Consequently, decerebration rostral to the pons results in a normal breathing rhythm capable of responding to many homeostatic challenges. Also included in the regulatory component are neural circuits underlying the processing of peripheral chemosensory and mechanosensory signals relevant to breathing, as well as reflexes that protect the airways and lung, including cough, sneeze, sigh, and gag.

Automatic breathing in mammals is both stable and labile (Table 1). As metabolism increases or decreases, ventilation must change accordingly. As one goes from rest to mild exercise, O_2 consumption can easily triple. Because there is only a modest reservoir of O_2 in the blood (about 4 min worth at rest), ventilation must change rapidly and accurately.

In healthy humans at rest, normal breathing during wakefulness and quiet sleep (referred to as eupnea) is characterized by alternating inspiratory and expiratory airflow at about 12–15 breaths per minute. Mammals are unique among vertebrates in having a diaphragm, an upward facing dome-shaped sheet of muscle separating the thoracic and abdominal viscera. In an adult, modest contraction of the diaphragm ($\sim$1.5 cm downward deflection) with a patent airway expands the lung by $\sim$500 ml. Expiration at rest does not require active effort in humans, as the elasticity of an inflated lung and expanded ribcage cause a recoil back to their resting volume. This end-expiratory volume is referred to as the functional residual capacity of the lung. The precise rhythmic muscle activation pattern for breathing is optimized to minimize energy expenditure ($<5\%$ of total body

Table 1 Breathing is both robust and labile

Robust	*Labile*
Works at birth	Rapid responses
Works continuously	Exercise
100 years!	Postural changes
Hard to fatigue	Sleep–wake
Failure is not an option	Hypoxia
0.5×10 cycles	Hypercapnea
	Slower responses
	Altitude
	Illness
	Development/aging
	Integrates with
	Exercise
	Phonation/speech
	Chewing/swallowing

metabolism at rest, increasing to 10% during heavy exercise), as breathing is an essentially continuous behavior throughout life. Over an evolutionary timescale, inefficiencies represent survival disadvantages, and on a shorter timescale, inefficiencies due to disease or injury lead to serious breathing discomfort, or breathlessness (dyspnea). Dyspnea is an essential signal, providing feedback that breathing is somehow compromised and that changes in behavior may be warranted.

Acute changes in breathing pattern underlie important reflexes. Cough, a large inspiration followed by a strong expiratory effort initially against a closed, then open, glottis, produces a high-velocity, turbulent expiratory airflow that moves mucus and foreign objects away from the lung, up the airways, and toward the mouth. Gag is a powerful coordinated reflex aimed at dislodging objects occluding the airways. Sneeze represents a rapid expiratory effort preceded by a large inspiration. Sigh is a very large inspiratory effort that can reinflate portions of the lung that have collapsed due to pulmonary disease such as pneumonia; sighs can also be induced by certain emotional states, and the relationship between the mechanisms of reflex and emotional sighs is unknown.

Breathing is an automatic function and, as such, is one of the few behaviors that can be studied *in vivo* in intact awake, anesthetized, or decerebrate animals. It may also be examined in highly reduced *in vitro* preparations, typically from neonatal rodents. Such preparations have allowed a broad range of investigations that at present support an in-depth understanding of the basic neural mechanisms that underlie breathing in mammals: (1) the basic rhythm is generated in the brain stem, (2) the rhythm is transformed into a pattern of motorneuronal activity by premotor circuits and motor neurons in the medulla and spinal cord, and (3) the rhythm and pattern are modulated by various afferents: Sources within the brain stem include 'central chemoreceptors,' raphe serotonergic and various brain stem noradrenergic and dopaminergic neurons, and the rostral pons. Peripheral sources, such as chemoreceptors in the carotid bodies and mechanoreceptors of the lung and respiratory muscles, also modulate breathing.

Most recently, studies of respiratory rhythm and pattern generation have been dominated by studies using highly reduced *in vitro* and *in situ* preparations that generate a rhythmic respiratory-related motor pattern. Key early discoveries have included the following: (1) Respiratory rhythm persists in thin (250–700 μm) transverse brain stem slices that encompass the preBötzinger complex (preBötC). (2) Rhythm persists in these preparations after attenuation of postsynaptic inhibition, suggesting the possibility that network interactions are not necessary for respiratory rhythm generation.

One, Perhaps Two, Oscillators for Generation of Respiratory Rhythm

One key medullary structure for breathing is the preBötC of the ventral respiratory column. Based on several observations, most investigators agree that the preBötC contains microcircuits critical for rhythm generation: (1) Medullary slices containing the preBötC generate a respiratory-related rhythm (one must be careful here and elsewhere in recognizing that behavior observed *in vitro* often differs distinctly from respiratory behaviors in intact mammals for multiple reasons, including absence of peripheral and many central inputs, lower temperature, and hypo- or hyperoxygenation of the bathing medium in the *in vitro* chamber). (2) Moderate lesions selectively targeting preBötC neurons that express the neurokinin-1 receptor in intact adult rats (unanesthetized and forebrain intact) induce a disturbed breathing pattern during sleep; with more extensive lesions, ataxic breathing occurs during wakefulness, with apneas (complete cessation of breathing) during sleep. (3) Neonatal mice lacking the transcription factor *MafB* show a marked reduction in preBötC neurons accompanied by markedly abnormal breathing patterns. (4) Juvenile rats with a transection of the brain stem just rostral to the preBötC continue to generate rhythmic inspiratory-dominated breathing patterns expressed in the spinal motor output to pump muscles.

A second proposed site for rhythm generation is the ventral parafacial respiratory group (given its proximity to the retrotrapezoid nucleus (RTN), and because of the as yet to be determined distinction between these two groups of neurons, this region may be referred to as the RTN/pFRG). Another

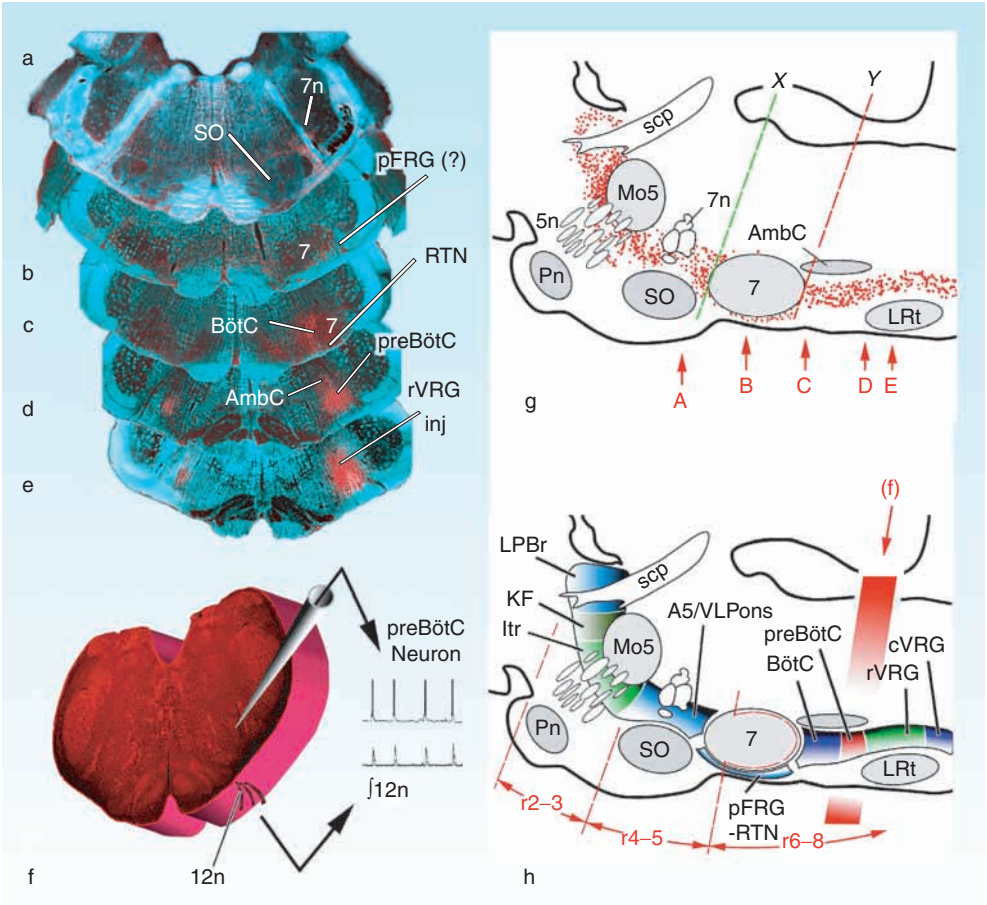

Figure 1 Respiratory regions in the lateral rhombencephalon. Expiratory and inspiratory oscillators (parafacial respiratory group (pFRG) and pre-Bötzinger complex (preBötC), respectively) are part of an essentially continuous column of respiratory-related neurons in the lateral rhombencephalon (see (a)–(e) and (g)). While the pFRG is well documented in the neonatal and adolescent rat brain, its exact location has yet to be directly determined in the adult. Hence, retrogradely labeled cells in (b) are tentatively identified as the pFRG(?). Relative to respiration, this lateral pontomedullary column is usually subdivided along its length into compartments based on distinct roles of the component neurons relative to rhythm and pattern generation (in (h), respiratory compartments are color-coded while nearby landmark brain stem regions are coded gray). The posterior part of the rostral part of ventral respiratory group (rVRG) and the caudal part of ventral respiratory group (cVRG) are characterized by inspiratory and expiratory bulbospinal premotor neurons, respectively, while the preBötC and anterior part of the rVRG contain inspiratory interneurons. The BötC contains mainly inhibitory expiratory interneurons with a subpopulation also projecting to the spinal cord. The pFRG consists of 'preinspiratory' interneurons neurons regulating expiratory motor output. The retrotrapezoid nucleus (RTN) originally designated all the neurons beneath the facial nucleus that project to the respiratory column in ventrolateral medulla. However, it has been suggested more recently that this term might be usefully limited to a group of nonphasic chemosensory neurons located at the caudal pole of the facial nucleus. Sections (a)–(e) are darkfield images of tranverse sections from an adult rat brain in which retrograde neuronal labeling of respiratory-related neurons is depicted in red after a tracer injection in the rostral part of the ventral respiratory group (rVRG in (e)). The approximate levels of sections (a)–(e) are shown by the arrows in (g). Sections (g) and (h) are schematic sagittal sections of the rat brain stem depicting retrograde labeling in the lateral brain stem after a tracer injection in the preBötC (g) and the division of this cell column into functionally distinct respiratory-related compartments (h). Cutting planes *x* and *y* in (g) indicate levels at which more-rostral brain stem removal does not interfere with expiratory motor output (plane *x*) or eliminates expiratory motor output (plane *y*). In (h), the correspondence of adult rhombencephalic regions to their origins in embryonic rhombomeres (r) is indicated by the labeled red arc and radial dashed dividing lines. The dividing line between rhombomeres 4–5 and rhombomere 6 has been displaced to include the facial nucleus (7) within r4–5. In the mammalian embryo, facial motor neurons originate in r4–5, then migrate dorsally, caudally, and ventrally to rhombomere 6. (f) A transverse section representative of neonatal brain stem slices encompassing the preBötC that are used for *in vitro* recording of neurons in the preBötC that fire in phase with continuous integrated inspiratory-like activity (∫12n) on the hypoglossal nerve (12n). The brain stem location of the slice shown in (f) is indicated by the red bar in (h). 5n, trigeminal nerve; 7, facial nucleus; 7n, facial nerve; 12n, hypoglossal nerve; A5, A5 noradrenergic neurons; AmbC, nucleus Ambiguus, compact part; BötC, Bötzinger complex; Itr, intertrigeminal nucleus; KF, Kölliker-Fuse nucleus; LPBr, lateral parabrachial area; LRt, lateral reticular nucleus; Mo5, motor nucleus of the trigeminal nerve; pFRG, parafacial respiratory group; preBötC, preBötzinger complex; PN, basilar pontine nuclei; scp, superior cerebellar peduncle; SO, superior olive; VLPons, ventrolateral pons. From George F Alheid and Donald R McCrimmon, Department of Physiology, Northwestern University.

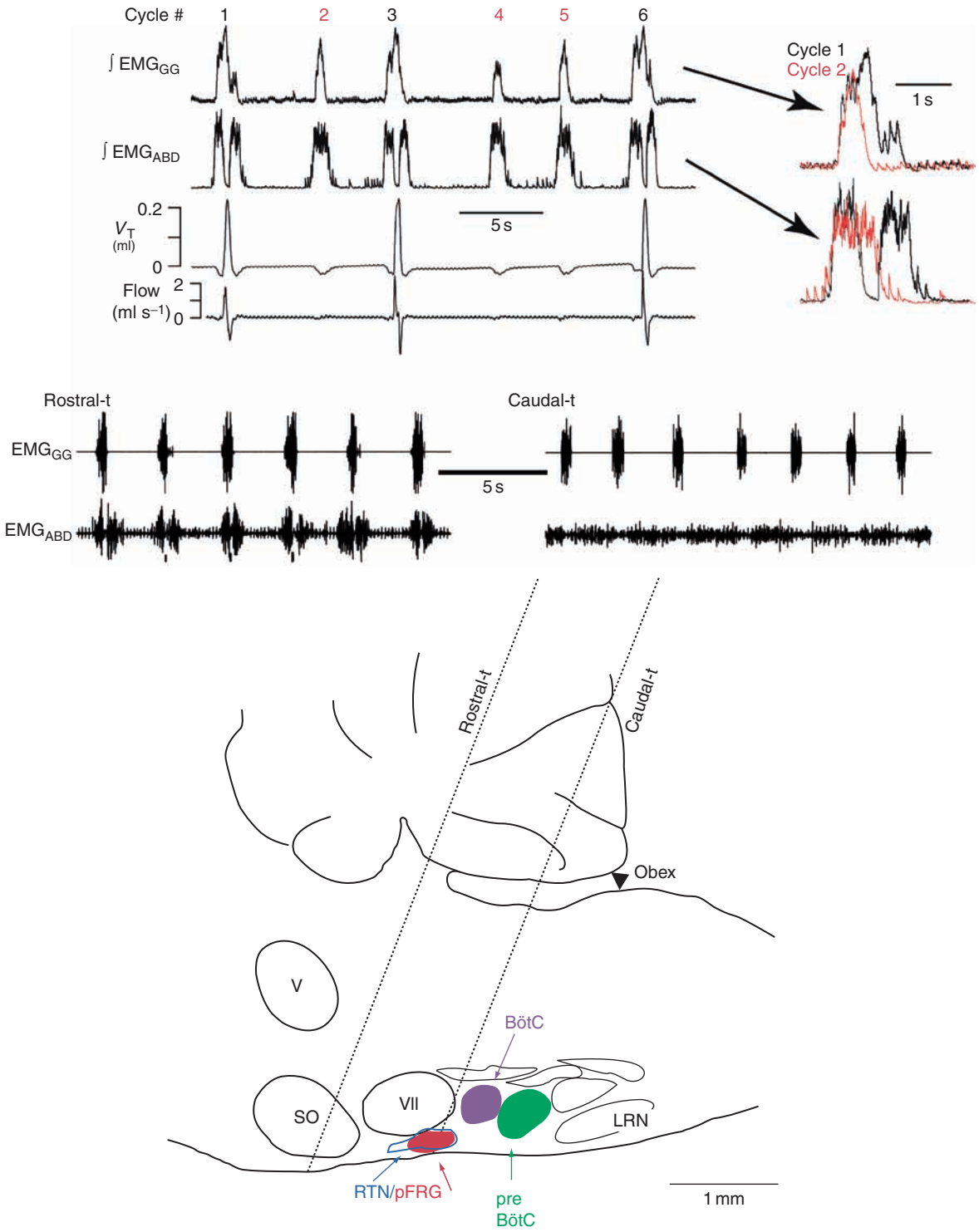

Figure 2 Inspiratory and expiratory rhythm generators in juvenile rat *in vivo*. (Top left) Typical recording with normal cycles (1, 3, 6) interspersed with cycles without inspiratory activity (2, 4, 5). Traces from top: integrated genioglossus (GG) muscle electromyogram (EMG), integrated abdominal muscle (ABD) EMG, tidal volume (V_T), and air flow (inspiratory air flow is up). (Top right) Superposition of $\int$EMG$_{GG}$ and $\int$EMG$_{ABD}$ in a normal cycle (black) compared with an inspiratory skipped cycle (red). Note the marked differences, especially in the $\int$EMG$_{ABD}$ burst. (Bottom) Alternating inspiratory and active expiratory activity is evident in EMG$_{GG}$ and EMG$_{ABD}$ following complete brain stem transection in the plane marked rostral-t. Following a more caudal transection (caudal-t), effectively removing the RTN/pFRG, only active inspiratory activity is present. BötC, Bötzinger complex; LRN, lateral ventricular nucleus; pFRG, parafacial respiratory group; preBötC, preBötzinger complex; RTN, retrotrapezoid nucleus; SO, superior olive. From Janczewski WA and Feldman JL (2005) Distinct rhythm generators for inspiration and expiration in the juvenile rat. *Journal of Physiology* 570(Pt 2): 407–420.

collection of neurons related to rhythmogenesis in the respiratory network is located just adjacent to the ventral surface of the facial nucleus. Some RTN/pFRG neurons project to the respiratory circuits of the ventrolateral medulla. RTN/pFRG neurons were initially investigated concerning their possible function as central chemosensors owing to their close proximity to proposed chemosensitive sites on the ventral surface of the medulla. Neurons in this general region, and especially those located laterally in the RTN, also contribute to rhythm generation, at least under some conditions. Key observations include the following: (1) A cluster of rhythmically active neurons is concentrated in the RTN/pFRG in *en bloc* preparations. When isolated from the preBötC by intramedullary transection in these preparations, a block of medullary tissue containing the RTN/pFRG can continue to generate a rhythm. (2) Inspiratory activity in slices which contain the preBötC but not the RTN/pFRG slow down continuously when exposed to μ-opiate agonists, whereas in *en bloc* preparations containing both structures, inspiratory activity slows discontinuously ('quantal slowing'; Figure 1). *In vivo*, opiates induce quantal slowing of inspiratory activity, which would be expected if RTN/pFRG neurons play some role in generating or influencing the underlying inspiratory rhythm, but expiratory activity remains unaffected (Figure 1). These latter observations suggest that the preBötC and RTN/pFN both have the intrinsic capacity to generate a respiratory rhythm but that they are normally entrained to produce a single respiratory rhythm. However, under some conditions, such as opiate receptor activation *in vitro*, the coupling is weakened and the RTN/pFRG is unencumbered and produces a predominantly expiratory-related motor rhythm reflecting its intrinsic oscillatory properties.

Thus, while many structures can modulate respiratory pattern, the working hypothesis of many in the field is that there are (at least) two coupled oscillators, the preBötC and the RTN/pFRG. An additional hypothesis is that inspiratory rhythm is generated by the preBötC (Figure 2) and active expiration by the RTN/pFRG.

The neural circuitry for breathing in mammals is the end result of an evolutionary lineage of rhythmic movements essential for gas exchange in vertebrate lung whose origins are closely related to the development of gas exchange in the gills of fish and the transformation of the swim bladder into the lung. Comparative analyses of breathing in nonmammalian vertebrate species, including reptiles and amphibians, are consequently of interest in understanding the origin of the oscillator(s) for terrestrial breathing. Tadpoles have two distinct breathing patterns (buccal and lung) that can act separately or in ensemble; in the latter case, they are phase-locked. It is interesting that the lung rhythm is opiate sensitive, whereas the buccal rhythm is not. Wilson and colleagues have speculated that these oscillators are homologs of the preBötC and the RTN/pFRG, respectively. An interesting distinction of mammalian breathing is that it is primarily driven by the diaphragm, a powerful inspiratory muscle not present in nonmammalian vertebrates. The relative role of the two oscillators (Figure 3) may differ between mammals and nonmammals since the power stroke for breathing is primarily inspiratory and diaphragm driven in mammals, whereas in most reptiles and amphibians, the power stroke is dominated more by expiratory effort.

While we do not understand why any respiratory oscillator would be opiate sensitive (or insensitive), this property clearly has consequences at birth, when there is a significant surge of maternally derived opiates in the newborn mammal. Such an opiate surge would depress the opiate-sensitive preBötC, revealing

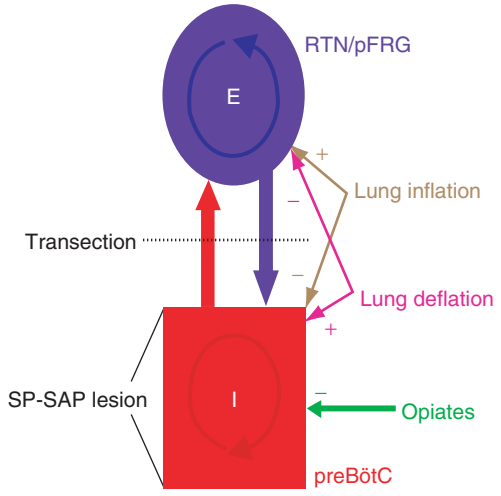

Figure 3 Summary of gross organization of respiratory rhythmogenesis in the brain stem of mammals. In this view, two oscillators are differentially affected by various inputs, such as opiates and lung inflation/deflation. The more rostral oscillator is located in the region of the retrotrapezoid nucleus/parafacial respiratory group (RTN/pFRG) and appears to drive active expiratory (E) activity (and may not be rhythmic at rest in many mammals when there is little or no active expiration), and the more caudal oscillator is located in the pre-Bötzinger complex (preBötC) and appears to drive inspiratory (I) activity. SP-SAP lesion of preBötC neurokinin-1 receptor (NK1R) neurons disrupts breathing. Transection between the two oscillators disrupts expiratory motor outflow although inspiratory activity continues unabated. Lung inflation enhances the activity of the expiratory oscillator and depresses the inspiratory oscillator, which serves ultimately to reduce lung volume, whereas lung deflation has the opposite effect on the inspiratory and expiratory oscillators. It is further speculated that the mechanism of rhythmogenesis in the preBötC involves a group pacemaker. What is unresolved in this model is why lesions of the preBötC NK1R neurons in adult rats disrupt rhythm and render breathing ineffective. From Feldman and Del Negro, *Nature Reviews Neuroscience*, 2006.

activity in the opiate-insensitive RTN-vpRFG, which then functions as what Champagnat and colleagues have suggested is an 'antiapnea center.'

While breathing is a continuous movement in mammals over the timescale of minutes, it can be interrupted by such transient behaviors as speech, emesis, defecation, or breath holding. Moreover, during sleep or unconsciousness, breathing can stop for brief periods, and such sleep apneas are common even in otherwise healthy humans. The preBötC may play a role in apneas in which the rhythm of breathing is interrupted (central sleep apneas). Thus, focal lesions specifically targeting a small subset of preBötC neurons (those expressing the neurokinin-1 receptor; NK1R) lead to significant disruption of, or even pathological, breathing in adult mice (**Figure 4**). Destruction of <50% of preBötC NK1R neurons does not appear to markedly affect the breathing pattern. At ~50% destruction, breathing during sleep is disturbed, exhibiting a marked increase in apneas. At ~80% destruction, a normal rhythmic breathing pattern is no longer seen even during wakefulness, and sleep invariably induces prolonged apneas that cease only on awakening in the form of abrupt arousal.

Serious disruption of breathing during sleep, including central apneas, is seen in at least three human neurodegenerative diseases: Parkinson's disease (PD), multiple systems atrophy (MSA), and amyotrophic lateral sclerosis (ALS). In the human ventrolateral brain stem, there is a marked decrease in NK1R neurons, presumably including those in the preBötC, in individuals who died at late stages of PD and MSA. A similar loss might be expected in individuals at the late stages of ALS, who often suffer from central sleep apneas. Recurring apneas, each inducing a period of brain hypoxia, could initiate a vicious cycle that accelerates the loss of preBötC neurons, leading to further and more profound breathing disruptions and ultimately death by asphyxia or from an apnea-induced heart attack or stroke.

Cellular and Network Bases for Generation of Respiratory Rhythm

The simplest explanation for the principal mechanism underlying generation of respiratory rhythm is autonomous bursting (pacemaking) in a subset of neurons, presumably in the preBötC and/or RTN/pFRG. The persuasive, but now debated, evidence in this case came from recordings of medullary slices that contain the preBötC and generate an inspiratory rhythm. That this rhythm continues after blockade of postsynaptic inhibition led directly to the hypothesis that pacemaker neurons generate the respiratory rhythm.

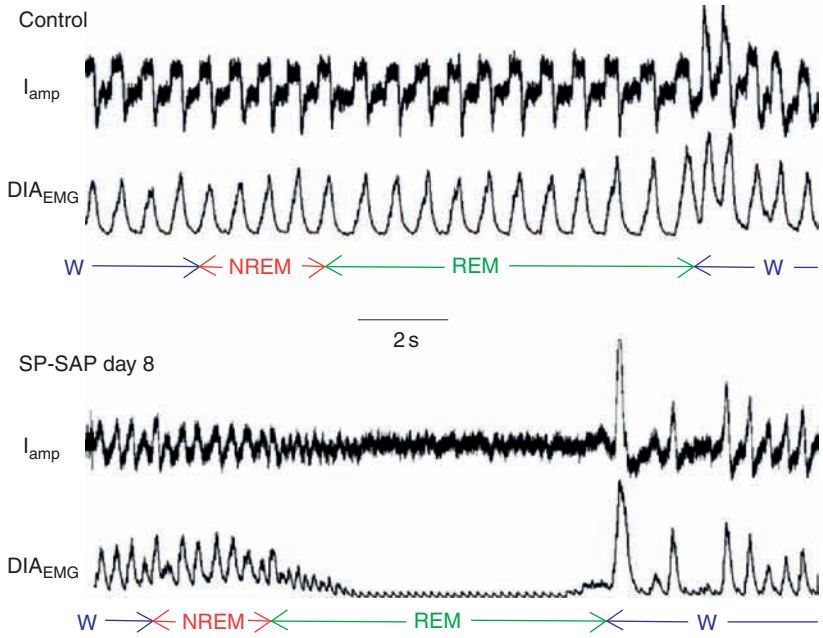

Figure 4 Lesions of the pre-Bötzinger complex (preBötC) significantly disturb breathing pattern in an adult unrestrained, unanesthetized rat. SP-SAP was injected bilaterally into the preBötC, which eventually killed more than 80% of preBötC neurokinin-1 receptor neurons on (or shortly after) postinjection day 10. Under control conditions, breathing is continuous throughout the sleep–wake (W) cycle, as determined from electroencephalogram and neck electromyogram (EMG) signals (not shown). However, on day 8, breathing pattern is normal except during rapid eye movement (REM) sleep, when apnea develops and ventilation is restored only on waking; this cycle repeats. I_{amp}, inspiratory amplitude; DIA_{EMG}, integrated diaphragmatic EMG; NREM, non-REM. From McKay LC, Janczewski WA, and Feldman JL (2005) Sleep-disordered breathing following targeted ablation of preBötzinger complex. *Nature Neuroscience* 8: 1142–1144.

Based solely on *in vitro* studies in neonatal rodents (Figure 5), neurons with bursting-pacemaker properties represent a subpopulation (5–25%) of preBötC neurons. There are two types of preBötC pacemaker neurons. One type, present at birth, is voltage dependent and driven by a persistent Na conductance (I_{NaP}). The second type appears to be extraordinarily sparse before postnatal day 7 in rodents and is dependent on Ca^{2+} and/or a Ca^{2+}-activated nonspecific cation current (I_{CAN}).

While the pacemaker hypothesis is intuitively appealing, there is no evidence to show that pacemaker activity is essential for respiratory rhythm generation at this time. In fact, there are data suggesting that such activity is not essential, at least *in vitro*. Thus, rhythm persists in brain stem slices after blocking the key pacemaker currents in the two populations of preBötC neurons that support pacemaker activity. While an obligatory role of pacemaker activity is disputed, there is a critical role in rhythmogenesis for intrinsic I_{NaP} and I_{CAN} currents that appear to be present in all preBötC neurons, pacemaker or otherwise, and contribute to burst generation in the context of amplifying inspiratory-related synaptic input.

Regardless of the cellular properties that underlie rhythm generation, network properties undoubtedly contribute in important ways. As mentioned above, rhythmogenesis appears unaffected unless >50% of preBötC NK1R neurons are destroyed. Unfortunately, very little is known in detail about the local connectivity

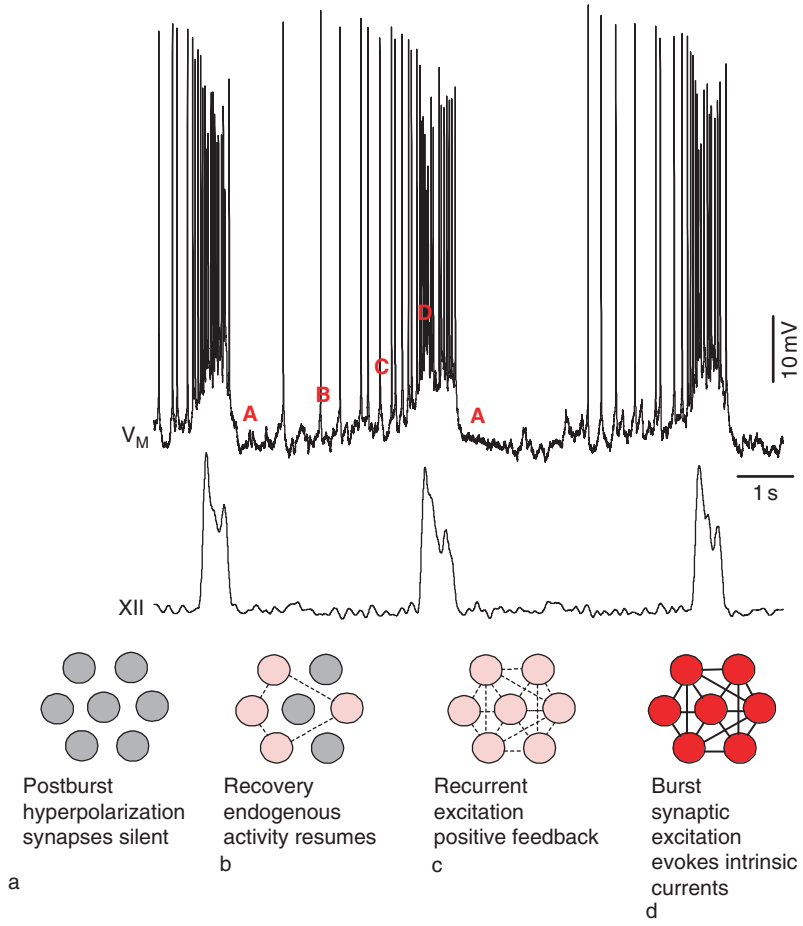

Figure 5 Group-pacemaker hypothesis of respiratory rhythm generation. The membrane potential of a rhythm-generating neuron is shown (V_M) with network activity, represented by XII motor output. Images at the bottom depict neuronal activity at different stages of the cycle. **A** is the refractory state following inspiration in which activity-dependent outward currents depress membrane potential and excitatory synapses in the network are inactive. During epoch **B**, the most excitable neurons recover from postburst hyperpolarization and begin to spike at low rate. In **C**, these highly excitable cells begin to synaptically activate other neurons, leading to aggregation of network activity, itself due to recurrent synaptic excitation – a positive-feedback process. The inspiratory burst ensues in **D** once a critical number of cells in the network are activated by recurrent excitation. In this final step, synaptic inputs recruit burst-generating intrinsic currents such as calcium-activated nonspecific cationic current and persistent sodium current, which give rise to large inspiratory burst potentials with high-frequency spike activity. Inspiratory bursts terminate due to intrinsic properties of cells that can reverse the positive feedback process, including Ca-dependent K currents and electrogenic pumps, which are recruited by cationic influx during inspiration. From Feldman and Del Negro, *Nature Reviews Neuroscience*, 2006.

of the preBötC or of RTN/pFRG neurons and scarcely more about their brain stem interconnections.

Cellular and Network Basis of Burst Pattern Formation

Once the respiratory timing has been established by the process of respiratory rhythm generation, the oscillating neuronal activity must be distributed appropriately to the different respiratory muscles to generate rhythmic and coordinated breaths. Little is known about the details of burst pattern formation, but activity from the preBötC is relayed to premotor neurons and interneurons of the ventral respiratory column. These premotor neurons relay information to respiratory motor neurons with inspiratory and expiratory activity, including motor outputs to the respiratory pump muscles and muscles regulating upper airway resistance. Sensory information and modulatory inputs are integrated at each step in this translation of respiratory rhythm into the detailed motor output to each respiratory muscle. Although apneas of central origin most likely arise from abnormalities in the process of rhythm generation, obstructive apneas and other forms of motor dyscoordination represent abnormalities in burst pattern formation.

Summary

Breathing in mammals consists of thoracic and upper airway movements that move gas into and out of the lungs for respiratory homeostasis. Breathing adjusts to changing physiological demands such as exercise yet also accommodates episodic behaviors like coughing and emesis. The underlying respiratory rhythm most likely originates in the preBötC and may also involve the RTN/pFRG, which is hypothesized to be critical for active expiration. Rhythms in preBötC neurons depend on excitatory network interactions as well as intrinsic membrane properties.

See also: Autonomic Nervous System: Respiratory Control; Autonomic Nervous System: Central Respiratory Control; Brainstem and Cranial Nerves; Evolution of Vertebrate Respiratory Control; Respiration.

Further Reading

Barabasi A-L (2002) *Linked: The New Science of Networks*. Cambridge, MA: Perseus.

Benarroch EE, Schmeichel AM, Low PA, and Parisi JE (2003) Depletion of ventromedullary NK-1 receptor-immunoreactive neurons in multiple system atrophy. *Brain* 126: 2183–2190.

Blanchi B and Sieweke MH (2005) Mutations of brainstem transcription factors and central respiratory disorders. *Trends in Molecular Medicine* 11: 23–30.

Del Negro CA, Koshiya N, Butera RJ Jr., and Smith JC (2002) Persistent sodium current, membrane properties and bursting behavior of pre-Bötzinger complex inspiratory neurons in vitro. *Journal of Neurophysiology* 88: 2242–2250.

Del Negro CA, Morgado-Valle C, and Feldman JL (2002) Respiratory rhythm: An emergent network property? *Neuron* 34: 821–830.

Del Negro CA, Morgado-Valle C, Hayes JA, et al. (2005) Sodium and calcium dependent pacemaker neurons and respiratory rhythm generation. *Journal of Neuroscience* 25: 446–453.

Feldman JL, Mitchell GS, and Nattie EE (2003) Breathing: Rhythmicity, plasticity, chemosensitivity. *Annual Review of Neuroscience* 26: 239–266.

Fortin G, Borday C, Germon I, and Champagnat J (2004) Breathing at birth: Influence of early developmental events. *Advances in Experimental Medicine and Biology* 551: 143–148.

Gray PA, Janczewski WA, Mellen N, McCrimmon DR, and Feldman JL (2001) Normal breathing requires preBötzinger complex neurokinin-1 receptor-expressing neurons. *Nature Neuroscience* 4: 927–930.

Gray PA, Rekling JC, Bocchiaro CM, and Feldman JL (1999) Modulation of respiratory frequency by peptidergic input to rhythmogenic neurons in the preBötzinger complex. *Science* 286: 1566–1568.

Janczewski WA and Feldman JL (2005) Distinct rhythm generators for inspiration and expiration in the juvenile rat. *Journal of Physiology* 570(Pt 2): 407–420.

Koshiya N and Smith JC (1999) Neuronal pacemaker for breathing visualized *in vitro*. *Nature* 400: 360–363.

McCrimmon DR, Alheid GF, Jiang M, Calandriello T, and Topgi A (2004) Converging functional and anatomical evidence for novel brainstem respiratory compartments in the rat. *Advances in Experimental Medicine and Biology* 551: 101–105.

McKay LC, Janczewski WA, and Feldman JL (2005) Sleep-disordered breathing following targeted ablation of preBötzinger complex. *Nature Neuroscience* 8: 1142–1144.

Mellen NM, Janczewski WA, Bocchiaro CM, and Feldman JL (2003) Opioid-induced quantal slowing reveals dual networks for respiratory rhythm generation. *Neuron* 37: 821–826.

Onimaru H and Homma I (2003) A novel functional neuron group for respiratory rhythm generation in the ventral medulla. *Journal of Neuroscience* 23: 1478–1486.

Pagliardini S, Ren J, and Greer JJ (2003) Ontogeny of the pre-Bötzinger complex in perinatal rats. *Journal of Neuroscience* 23: 9575–9584.

Rekling JC and Feldman JL (1998) PreBötzinger complex and pacemaker neurons: Hypothesized site and kernel for respiratory rhythm generation. *Annual Review of Physiology* 60: 385–405.

Smith JC, Butera RJ, Koshiya N, Del Negro C, Wilson CG, and Johnson SM (2000) Respiratory rhythm generation in neonatal and adult mammals: The hybrid pacemaker-network model. *Respiration Physiology* 122: 131–147.

Smith JC, Ellenberger HH, Ballanyi K, Richter DW, and Feldman JL (1991) Pre-Bötzinger complex: A brainstem region that may generate respiratory rhythm in mammals. *Science* 254: 726–729.

Vasilakos K, Wilson RJ, Kimura N, and Remmers JE (2004) Ancient gill and lung oscillators may generate the respiratory rhythm of frogs and rats. *Journal of Neurobiology* 62(3): 369–385.

Broca's Area: Evolution

W D Hopkins, Yerkes National Primate Research Center, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

In 1861, Paul Broca described the speech and linguistic production deficits of a patient named 'Tan Tan.' Upon death, Broca obtained the postmortem brain and described a large lesion in the left inferior frontal gyrus (IFG). From his subsequent clinical observations, Broca concluded that the left inferior frontal lobe was important for linguistic functions in humans. Thirteen years later, Wernicke described his famous cases which constituted patients with lesions in the posterior portion of the left temporal lobe. Wernicke's patients had deficits in speech and linguistic comprehension. Since the time of Broca and Wernicke, these two brain regions have received an enormous amount of scientific attention and empirical investigation. Behavioral as well as neurophysiolocal imaging technologies have consistently implicated these two regions in higher order linguistic, cognitive, as well as motor functions in the human brain. In terms of morphology, there is also evidence of a left hemisphere bias in humans, although this remains somewhat debated. For example, using standard region-of-interest approaches to measuring Broca's area, left hemisphere asymmetries have been reported in the pars operculum and pars triangularis, but some voxel-based analyses from magnetic resonance imaging (MRI) have failed to reveal consistent evidence of leftward asymmetries within the inferior frontal region.

Almost since the earliest observations of Broca, there has been parallel interest in whether animals, and particularly our closest living relative the great apes, possess similar lateralized cortical systems that might serve as antecedent conditions to the emergence of language and speech in modern humans. Scientific interest in whether apes and other nonhuman primates exhibited neuroanatomical asymmetries was undoubtedly influenced and stimulated by the behavioral demonstrations of sophisticated cognitive abilities such as self-recognition, symbol acquisition, and tool use in nonhuman primates, particularly in great apes. However, until recently, it has been widely held that only humans exhibit brain asymmetries at the population level; however, this conclusion was largely based on very little data in great apes and relatively simple and descriptive studies in other primate species involving few numbers of subjects. With the advent of modern imaging technologies as well as an increasing interest in preserving and examining postmortem brains, the question of whether apes and other nonhuman primates exhibit population-level asymmetries has become increasingly evident.

This article describes and summarizes the comparative evidence of Broca's area neuroanatomy and function in nonhuman primates. Integrating these findings may provide important clues to the evolution of brain asymmetry and functions as it potentially relates to the evolution of cognitive and behavioral functions in primates, including humans.

Morphology and Cytoarchitectonics

Broca's area is composed of Brodmann's area 44 and 45, and these cell types have been described in a variety of nonhuman primate species. The most commonly studied genera has been *Macaca*, and this has included cytoarchitectonic and histological studies of areas 44 and 45. In macaques, area 44 and 45 cells have been identified in regions surrounding the inferior arcuate sulcus, inferior principal dimple, and sulcus principalis, although there is debate about the characterization of the ventrolateral region of the macaque brain. Specifically, previous studies had not adopted the same cytoarchitectonic criteria as that employed in human brain and therefore some confusion has arisen over the exact region and function of cells within this morphological region. However, it has been shown that the cytoarchitecture and granular structure of areas 44, 45, and 6 are significantly different in the macaque brain. Specifically, within area 45, there are unusually large neurons in layer III compared to layer V. Second, area 45 has a well-developed granular area, whereas it is rudimentary and irregular in area 44 and absent in agranular area 6V. Thus, within the ventrolateral region of the frontal lobe of the macaque, there are distinct layering patterns reflective of areas 6, 44, and 45, respectively.

With respect to chimpanzees and orangutans, early studies initially mapped areas 44 and 45 in the chimpanzee brain. According to Von Bonin, predominantly area 44 cells were found in the IFG in the region lying between the precentral inferior sulcus (PCI) and the fronto-orbital (FO) sulcus. In contrast, area 45 cells were found in the frontal cortex, just dorsal to the superior tip of the FO sulcus. Sherwood and colleagues performed a similar cytoarchitectonic analysis of the IFG of chimpanzees and found a somewhat different pattern of results from those of Von Bonin and colleagues. According to Sherwood and colleagues, both area 44 and area 45 cells, as well as

area 6 (primary motor cortex), can be found in regions corresponding to the IFG originally described by Von Bonin. Moreover, Sherwood and colleagues argue that there are considerable individual differences in the sulci making up the inferior frontal region, and the presence of inconsistent structural landmarks makes mapping the cellular to the morphological regions difficult. Inconsistencies between the cytoarchitectonic and sulcal border of regions 44 and 45 have similarly been claimed in postmortem human brains.

Lateralization of Broca's Area

As noted previously, clinical and experimental data clearly suggest that the left inferior frontal cortex is involved in speech production, although other motor systems, such as the basal ganglia and cerebellum, play a role in language production. With respect to morphology, data from humans have largely shown that inferior frontal cortex is larger in the left compared to the right hemisphere. For example, measuring brain asymmetries from MRI, both the pars operculum and the pars triangularis have been reported to be larger in the left compared to the right hemisphere, particularly among right-handed individuals. Less clear from the extant data in humans is the evidence of mircrostructural asymmetries (or cellular asymmetries). In one study, postmortem measurements of the cytoarchitecture of Brodmann's regions 44 and 45 in ten individuals showed a leftward asymmetry for area 44 but not area 45. Moreover, cytoarchitectonic borders of areas 44 and 45 did not strongly coincide with sulcal contours used to define the regions. In a follow-up study, in a sample of 34 brains ranging in age from 3.5 months to 75 years of age, evidence of asymmetries in Broca's area was reported but for Brodmann's area 45, not area 44, as in the original report. Using a different approach, others have reported that there are significantly more acetylcholinesterase-rich neurons located in layer III within Brodmann's area 45 in the left compared to the right hemisphere in a sample of six human brains. In none of the postmortem studies were handedness data on the subjects available.

With respect to the comparative data on asymmetries in the frontal lobe, there are few studies in monkeys and most have directly measured sulci form postmortem brain or endocasts. For example, Falk and colleagues examined asymmetries of 11 sulci in the frontal lobe in a sample of 403 macaque endocasts. Population-level rightward asymmetries were found for the length of the rectus and lateral orbital frontal sulcus. A significant leftward asymmetry was found for the length of the central sulcus. Falk and colleagues also reported a right frontal petalia asymmetry in the macaques, results consistent with findings in great apes. In summary, although there is evidence of population-level asymmetries in macaque brains, asymmetries in regions or sulci demarking Broca's area have not been reported.

In great apes, the bank of tissue lying between FO and PCI was measured from MRIs in a sample of 27 great apes, including 20 chimpanzees, five bonobos, and two gorillas (**Figure 1**). For the measurement of the IFG, MRI scans were aligned in the sagittal plane and cut into 1 mm slices. Consecutive parasagittal slices were traced from the lateral surface of the IFG to the most medial slice immediately preceding the insula. The IFG was defined as the portion of cortical surface bounded by the FO and the PCI sulci. Both FO and PCI could be clearly seen in parasagittal (1 mm thick) MRI slices. Overall, a significant leftward asymmetry was found for the IFG of the great apes.

MRI scans have been obtained in 49 more chimpanzees and the addition of these specimens has confirmed the previous observations of a left hemisphere asymmetry in the frontal operculum of chimpanzees (**Figure 2**). A comparison of asymmetry values between the original sample and the subsequent 49 individuals found no significant differences between them. If the frontal operculum is divided into lateral and medial portions, the leftward asymmetry is particularly robust for the medial compared to the lateral portion of this region (**Figure 2**).

Some have suggested that there is too much variability in IFG sulci, particularly in the PCI which comprise the landmarks of the frontal operculum in great apes, to obtain a reliable assessment of asymmetry in this region. Although variability is observed in the sulci comprising the frontal operculum of apes, others have argued that consistent criteria can be applied to this region to obtain reliable measurements. Specifically, the majority of cases of PCI bifurcation observed in the sample of chimpanzee brains follow a typical pattern in which the posterior ramus of the bifurcation (1) terminated at a more dorsal level than the anterior one and (2) tended to run more horizontally than the anterior ramus, which in turn ran more vertically and parallel to the central sulcus. Notably, Sherwood and colleagues gave a similar account for this usual pattern of PCI bifurcation in their cadaver specimens. Thus, most instances of PCI bifurcation do not present any serious ambiguity as to determining the posterior boundary of the pars opercularis. Because (1) PCI is usually seen as running vertically and roughly parallel to the central sulcus in the common sulcal pattern of the IFG of great apes and (2) in many instances it is likely that the posterior ramus of

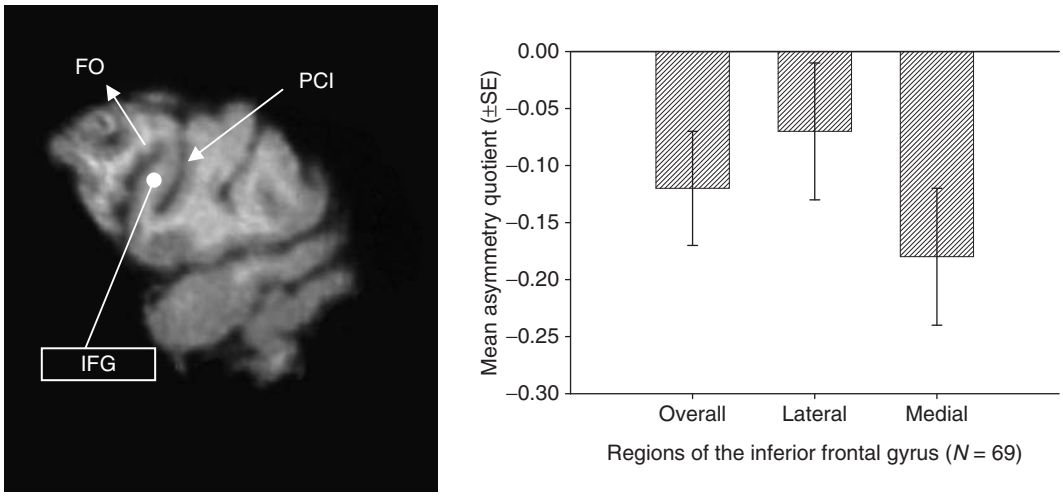

a b

Figure 1 (a) The frontal operculum or inferior frontal gyrus (IFG) of a chimpanzee brain. The tissue lies between the precentral inferior sulcus (PCI) and fronto-orbital (FO) sulcus as defined by Cantalupo and Hopkins. (b) Mean asymmetry quotients (± SE) for the overall region of the IFG as well as divided into the lateral and medial portions of the gyrus. Asymmetry quotients (AQ) values are derived following the formula $AQ = (R - L)/((R + L) \times 0.5))$, where R and L refer to the summed area measures for the left and right cerebral hemispheres, respectively.

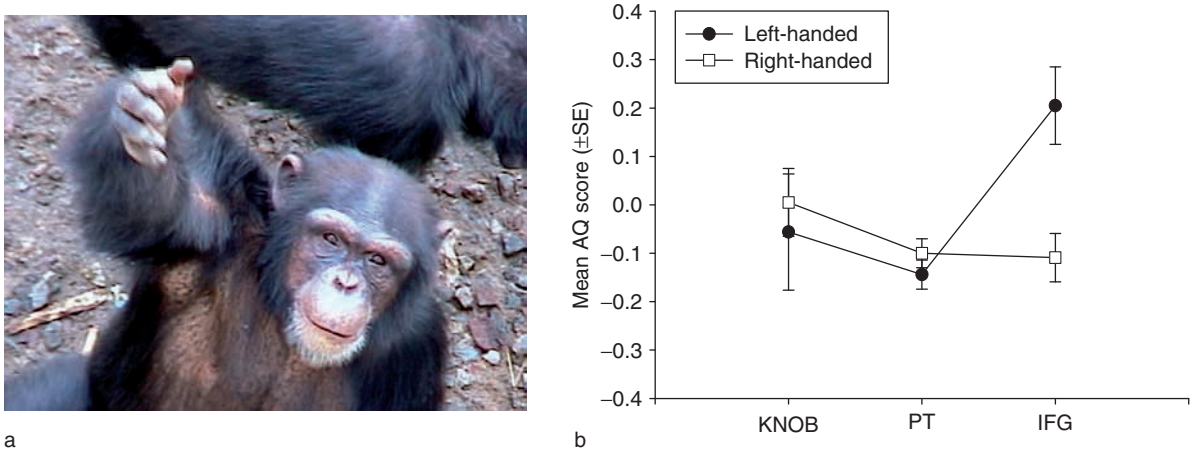

a b

Figure 2 (a) Chimpanzee manually gesturing toward a human when begging for food. (b) Mean asymmetry quotients (±SE) for three brain regions measured in chimpanzees, including the inferior frontal gyrus (IFG), planum temporale (PT), and motor–hand area of the precentral gyrus (KNOB). Chimpanzees that prefer to gesture with their right hand have a more leftward asymmetry in the IFG compared to chimpanzees who do not prefer to use the right hand. No significant differences in PT and KNOB are found for right- and non-right-handed chimpanzees.

a PCI bifurcation is, in fact, the subcentral anterior sulcus joining PCI, one can consistently adopt the criterion of choosing the anterior ramus of a PCI bifurcation as the posterior boundary for the pars opercularis. In fact, this criterion was consistently employed in previous reports. Lastly, individual differences in bifurcation patterns of FO are less frequent than in PCI; thus, my colleagues and I measured FO from MRI scans in the left and right cerebral hemispheres in a sample of 66 chimpanzees. A significant leftward asymmetry was found for FO, and asymmetries in the length of FO positively correlated with asymmetries in the entire region of the IFG ($r = 0.588$, $df = 67$, $p < 0.01$), suggesting that FO is a good estimate of lateralization in the inferior frontal region.

In summary, leftward morphological asymmetries, whether quantified for the tissue lying between PCI

and FO or by measuring the length of FO on successive 1 mm slices, are present in chimpanzees. The extent to which the morphological regions map to individual differences in cytoarchitectonics remains unknown but clearly warrants further investigation.

Potential Evolutionary and Functional Correlates of Broca's Area in Primates

The implication of Broca's area in speech and language functions in humans has theoretically stimulated a significant amount of work on the role of the homologous regions to Broca's in nonhuman primates. Inherent in many of these arguments is the question of whether human speech evolved directly from neural substrates associated with the production of vocalizations in closely related species or whether language and speech evolved from an earlier, manual gestural system of communication. Discussing different evolutionary theories of language evolution is beyond the scope of this article, but none of the existing data on the function of Broca's area in nonhuman primates strongly support any one theory or another; however, there are at least some data hinting at the function of Broca's area.

The homolog to Broca's area in monkeys does not appear to have any role in the production of species-specific vocalizations because lesions to this region do not disrupt either the frequency or the structure of sound production. Indeed, many have argued that the production of vocalizations by nonhuman primates is not under cortical control. However, a single cell recording study in macaques showed that stimulation of area 44 neurons from the frontal lobe simultaneously induced both digit and oral–facial movements. Similarly, early research performed in chimpanzees and recent work in macaques has shown that stimulation of the ventral portion of the motor cortex, including regions within the IFG, induces orofacial movements as well as movement of the larynx. Thus, the inferior frontal region, particularly area 44, seems to be involved in orofacial movements, movements of the larynx, and possibly hand or digit movements.

Rather than focusing on orofacial movement and its potential link to language and speech evolution, a substantial body of work has accumulated during the past 10 years on the role of the inferior frontal regions, notably area F5 in the macaque, in the perception and action of motor movements. Notably, so-called mirror neurons have been identified in area F5 in macaques (as well as other regions). Single cell recording studies in monkeys have shown that mirror neurons located in F5 respond when monkeys execute a grasping action or perceive another animal (or human) engaging in the same grasping action. In humans, a number of functional imaging studies have implicated Broca's area in production and recognition of simple grasping movements, orofacial movements, and imitation of gestures and mimed actions. It has been hypothesized that mirror neurons may have played a vital role in the evolution of brain systems underlying language and speech. Rather than focusing entirely on direct selection for motor systems associated with speech, it has been argued that the presence of mirror neurons in F5 (macaques) or Broca's area (humans) provides a neural substrate for the emergence of necessary cognitive links between recognition and production of complex motor actions, including social learning and imitation.

In chimpanzees, studies have examined the association between asymmetries in the inferior frontal region and lateralization in hand use for manual gestures compared to hand use for noncommunicative actions. Specifically, in a sample of 56 chimpanzees, asymmetries in the IFG were compared in subjects classified as right- or non-right-handed for manual gestures, and right-handed apes had a significantly greater leftward asymmetry than non-right-handed individuals. When the handedness of the same sample of chimpanzees was characterized for a noncommunicative motor action, simple reaching, no significant differences in asymmetry values for the IFG were found. Thus, differences in the asymmetry quotient scores of the IFG were specific to hand use for manual gestures.

Conclusions

Based on cellular anatomy, regions homologous to Broca's area are evident in apes and more distantly related primate species. In humans and great apes, the inferior frontal lobe is lateralized, with a significant majority of subjects exhibiting a left hemisphere asymmetry. It is unclear whether more distantly related primates show asymmetries in the frontal operculum, particularly in regions homologous to areas 44 and 45, but this should be investigated further. With the development of quantification of cytoarchitectonic regions using gray-level indices, the question of asymmetry in macaques and apes seems feasible. In addition, there are few studies in humans and virtually none in nonhuman primates examining whether asymmetries at the cellular level correlate with either morphological asymmetries in the same regions or behavioral asymmetries. Similarly, this issue also warrants further investigation, not just on asymmetries in the inferior frontal region but also in other parts of the cortex. The comparative data further suggest that asymmetries in the IFG of chimpanzees are associated with variation in hand use for gestural communication. In principle, these

data support theories proposing that lateralization for human speech evolved from an earlier, extant gestural communication system present in early hominids. Although possible, the results do not rule out other theoretical explanations for left hemisphere lateralization in human speech and other motor functions (i.e., handedness). Moreover, certainly other functional asymmetries could be linked to these regions. For example, there is evidence of population-level orofacial asymmetries in marmosets, rhesus monkeys, and chimpanzees, and perhaps these asymmetries might be associated with frontal operculum asymmetries. Lastly, additional studies are needed on neural connectivity between Broca's area and other cortical regions of the brain. The multimodal function of Broca's area seen in humans clearly implies connectivity between many other neural regions involved in multiple functions. It may be that differences in connectivity between Broca's and other brain regions account for the complexity in structure and function of this brain region in closely and more distantly related primates, including humans.

See also: Brain Asymmetry: Evolution; Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Orofacial Motor Control.

Further Reading

Amunts K, Schleicher A, Burgel U, et al. (1999) Broca's region revisited: Cytoarchitecture and intersubject variability. *Journal of Comparative Neurology* 412: 319–341.

Arbib MA (2005) From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences* 28: 105–167.

Cantalupo C and Hopkins WD (2001) Asymmetric Broca's area in great apes. *Nature* 414: 505.

Corballis MC (2002) *From Hand to Mouth: The Origins of Language*. Princeton, NJ: Princeton University Press.

Foundas AL (1998) MRI asymmetries of Broca's area: The pars triangularis and pars opercularis. *Brain and Language* 64: 282–296.

Hopkins WD, Pilcher D, and Cantalupo C (2003) Neural basis of cognition, communication and handedness in nonhuman primates. In: Maiestripieri D (ed.) *Primate Psychology*. Cambridge, MA: Harvard University Press.

Hopkins WD, Russell J, Freeman H, Buehler N, Reynolds E, and Schapiro SJ (2005) The distribution and development of handedness for manual gestures in captive chimpanzees. *Psychological Science* 16: 487–493.

Jurgens J (2002) Neural pathways underlying vocal control. *Neuroscience and Biobehavioral Reviews* 26: 235–258.

Lieberman P (2002) On the nature and evolution of the neural bases of human language. *Yearbook of Physical Anthropology* 45: 36–62.

Petrides M, Cadoret G, and MacKey S (2005) Orofacial somatomotor responses in the macaque monkey homologue of Broca's area. *Nature* 435: 1235–1239.

Rizzolatti G and Arbib MA (1998) Language within our grasp. *Trends in Neurosciences* 21: 188–194.

Rogers LJ and Andrew JR (2002) *Comparative Vertebrate Lateralization*. Cambridge, UK: Cambridge University Press.

Sherwood CC, Broadfield DC, Holloway RL, Gannon PJ, and Hof PR (2003) Variability of Broca's area homologue in African great apes: Implications for language evolution. *Anatomical Record* 271A: 276–285.

Toga AW and Thompson PM (2003) Mapping brain asymmetry. *Nature Reviews Neuroscience* 4: 37–48.

Watkins KE, Paus T, Lerch JP, et al. (2001) Structural asymmetries in the human brain: A voxel-based statistical analysis of 142 MRI scans. *Cerebral Cortex* 11: 868–877.

Brodmann's Areas

E G Jones, University of California, Davis, CA, USA

© 2009 Published by Elsevier Ltd.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

The concept of the cortical area or cortical field has its origins in comparative cytoarchitectonic studies of the cerebral cortex made in the first decade of the twentieth century by a number of anatomists, particularly Bolton, Lewis, Campbell, Brodmann, C Vogt, and O Vogt. Among these names, that of Brodmann is perhaps the best remembered since the numbers he assigned to the cortical fields he identified have entered into the common vocabulary of cortical anatomy and physiology. The numbers simply indicate the order in which Brodmann studied them (Figure 1).

Cytoarchitectonics, and myeloarchitectonics, the complementary technique that followed it, have as their basis the recognition that different regions of the cerebral cortex have reproducible differences in cell and fiber layering and packing density. Architectonics is the art of determining where the boundaries fall between different areas. The early impetus to this type of study came first from phrenology, then from demonstrations of the electrical excitability of the motor cortex, and from the recognition of the dependence of vision on posterior regions of the cerebral cortex. The identification of cytoarchitectonic areas and their borders thus became a search for areas of the cerebral cortex with distinct functions. The visual, motor, and somatic sensory areas were the first identified and later auditory, associational, and many further subdivisions of the original areas have been described. In the human brain Brodmann recognized nearly 50 areas, Campbell less than 20, C Vogt and O Vogt more than 200. Progressively fewer were found in the cortices of lower mammals, though Brodmann argued that the sensory and motor areas could be identified in all. At the moment, most neuroscientists would probably agree that Brodmann's divisions come somewhere near what seems reasonable, though after a period of the converse, the trend now appears to be going in the direction of a finer and finer parcellation, akin to that of the Vogts, as functional subdivisions are recognized within areas by new techniques and as imaging studies reveal new functional localizations. Brodmann's numbered fields have been given names in other nomenclatural schemes, and individual areas in modern experimental studies have sometimes been given specific names, but in general there is a remarkable correspondence between schemes. Where new fields have been described, they commonly represent further parcellations of larger regions of cortex identified as single fields by Brodmann.

The persistence of the architectonic field concept and its resurgence after the attacks on it of the 1950s and 1960s, stem from the facts that structural variation can be seen across the cortex and that many of these structural variations can be correlated with distinct patterns of input–output connections, with distinct physiological properties, and often with definable aspects of sensation, perception, cognition, and motor control.

A modern definition of a cortical field would probably encompass the following points: (1) It is an area of singular cellular and fiber architecture often with boundaries that are particularly sharp and recognizable even by the inexperienced eye. The border between areas 17 and 18 in the visual cortex is probably the most clear-cut. Borders between other areas are not always so sharp and one area may appear to give place to another over a gradient of architectonic change. This may represent a truly blurred border or be imposed by stretching of the cortex in the floor of a sulcus or on the crown of a gyrus and then present a trap for the unwary who may see a distinct field where one does not exist. (2) It is an area that receives afferent fibers from the major population of cells in a defined relay nucleus of the thalamus. This input stops abruptly at an architectonic border and often serves to demonstrate the position of a sharp border when architectonics alone suggests a gradient of change. There appears to be no overlap between inputs from different specific thalamic nuclei to adjacent fields. Fibers arising from so-called nonspecific thalamic nuclei and from subsidiary populations of cells in a relay nucleus, however, often disregard these borders. (3) It receives a set of corticocortical and commissural axons from a limited, defined, and constant set of other cortical areas. These tend to respect the same boundaries delimited by architectonics and by the distribution of thalamic afferents. (4) It gives rise to a constant set of output connections to a limited and constant set of cortical and subcortical targets, including returning fibers to those thalamic nuclei (specific and nonspecific) from which it receives inputs. (5) A topographic map of a receptive periphery such as the visual field, auditory space, the body surface, or of the body's musculature can be detected in certain cortical fields by evoked potential or microelectrode mapping or by electrical stimulation

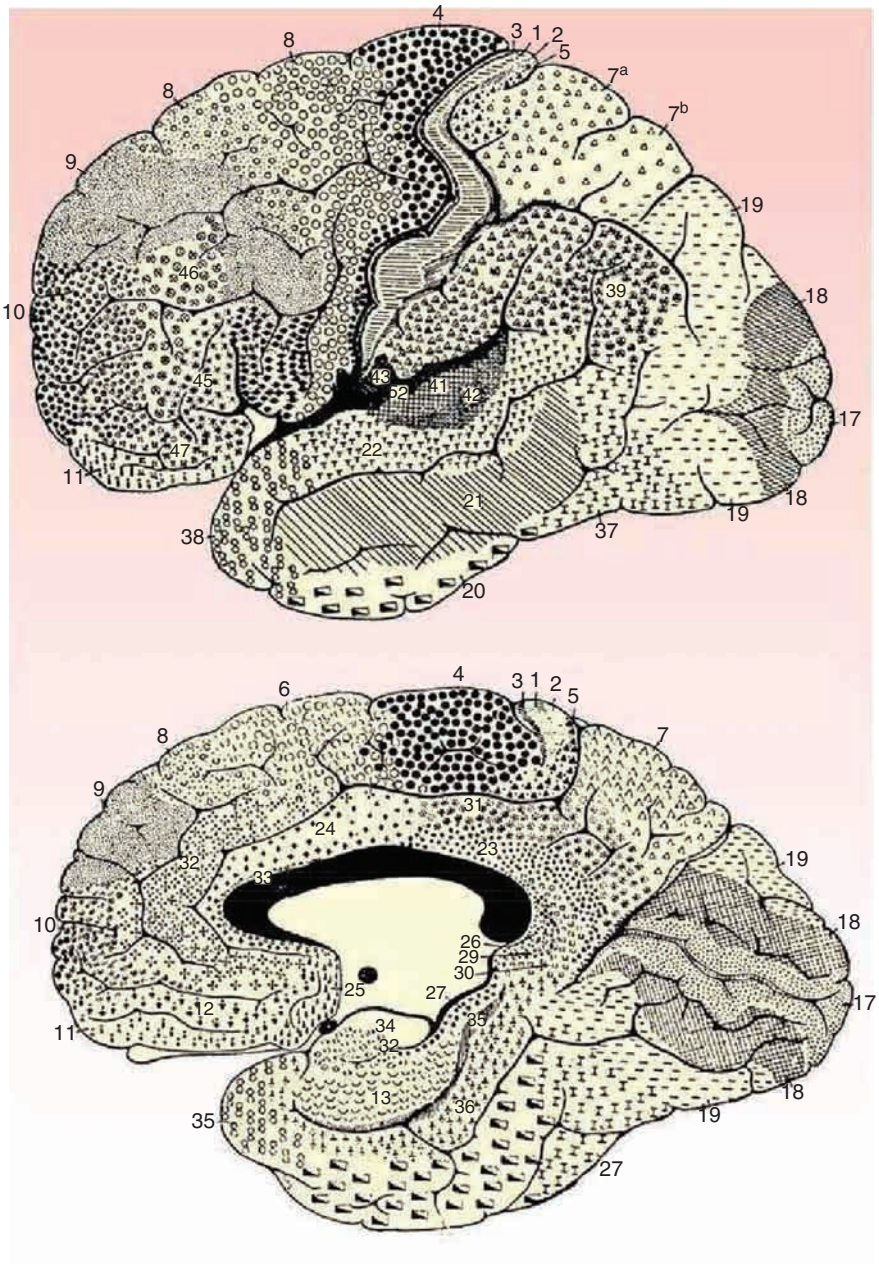

Figure 1 Brodmann's map of the areas in the human cerebral cortex. (Top) lateral view; (bottom) medial view. From Brodmann K (1910) Feinere Anatomie des Grosshirns. In: Lewandowsky M (ed.) *Handbuch der Neurologie, Vol. 1: Allgemeinere Neurologie*, pp. 206–307. Berlin: Springer.

of the field. In association with these maps, the neurons in the field may display physiological properties that distinguish them from those in other fields. Aggregation of subpopulations of cells with different functional properties can occur in a field and sometimes, as in some of the visually related areas, these aggregations may have definable anatomical characteristics. These characteristics may then serve as a basis for further subdividing the cortical field. (6) Deactivation of a field by ablation, cooling, or injection of drugs may lead to loss of a particular sensory or motor function or of other behaviors. (7) The field may show a characteristic pattern of histochemical or immunocytochemical staining different from that of its neighbors. The so-called chemoarchitectonic borders defined in this manner commonly coincide with those defined by cell and fiber staining.

Obviously, not every cortical field can be demonstrated to meet all these criteria, primarily because of

the difficulty in eliciting physiological responses in many of them or in defining the nature of a deficit resulting from their deactivation. Similarly, arguments invariably ensue about the sharpness of certain architectonic borders and about whether particular architectonic changes indicate the existence of new fields or simply variation within a field, a variation perhaps resulting from a local concentration of cells projecting to the same site or with a particular physiological property. The connectional basis of the cortical field, especially that defined by connections with a particular principal thalamic nucleus, probably remains the most reliable anatomical guide for deciding when a region of cortex should be called a cortical area or field independent of its neighbors and for determining its boundaries. It is to be hoped that eventually molecular-genetic markers will be discovered that determine this reproducible parcellation of the cerebral cortex on the basis of structural and functional properties.

See also: Cerebral Cortex; Memory Consolidation: Cerebral Cortex; Topographic Maps: Molecular Mechanisms.

Further Reading

Bailey P and von Bonin G (1951) *The Isocortex of Man*. Urbana, IL: University of Illinois Press.

Brodmann K (1909) *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig: Barth.

Brodmann K (1910) Feinere Anatomie des Grosshirns. In: Lewandowsky M (ed.) *Handbuch der Neurologie, Vol. 1: Allgemeinere Neurologie*, pp. 206–307. Berlin: Springer.

Campbell AW (1905) *Histological Studies on the Localization of Cerebral Function*. London: Cambridge University Press.

Flechsig P (1920) *Anatomie des menschlichen Gehirns und Rückenmarks auf myelogenetischer Grundlage*. Leipzig: Thieme.

Jones EG (1985) *The Thalamus*. New York: Plenum.

Jones EG and Burton H (1976) Areal differences in the laminar distribution of thalamic afferents in cortical fields of the insular, parietal and temporal regions of primates. *Journal of Comparative Neurology* 168: 197–248.

Lewis JW and Van Essen DC (2000) Mapping of architectonic subdivisions in the macaque monkey, with emphasis on parieto-occipital cortex. *Journal of Comparative Neurology* 428: 79–111.

Peters A and Jones EG (eds.) (1984–1994) *Cerebral Cortex*, vols. 1–10. New York: Plenum.

von Economo C and Koskinas CN (1925) *Die Cytoarchitektonik der Hirnrinde des erwachsenen Menschen*. Berlin: Springer.

CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy)

A Joutel, INSERM U740, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Arteries are specified into different calibers and types of vessels to perform different functions. Schematically, the major arteries of the trunk are elastic arteries of large diameter and low resistance that provide the resilience these arteries need to absorb the hemodynamic stress of cardiac systole, and to release this energy in the form of sustained blood pressure during diastole. Conversely, distal arteries are muscular arteries of small diameter and high resistance that are critically important for the development of resting vascular tone, preventing peripheral vascular collapse, and for local regulation of blood flow, particularly in the brain. Pathological changes of distal arteries in the brain produce small deep infarcts or hemorrhages and widespread white-matter changes. Cerebral small-artery diseases are recognized as being the origin of a major proportion of abnormal conditions. They are responsible for 20–30% of all strokes and are the leading cause of vascular dementia. The molecular and cellular bases of small-artery diseases of the brain are poorly understood, although three major risk factors have been identified, including aging, arterial hypertension, and diabetes. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is a recently identified autosomal dominant small-artery disease, caused by mutations in the Notch 3 receptor. The fact that CADASIL is a prototypical small-artery disease renders it a paradigm for studying the pathobiology of related diseases. In this article, the core findings on CADASIL and the Notch 3 signaling pathway are reviewed.

CADASIL

The story of this disease began back in the early 1990s, with the clinical, neuropathological, and molecular characterization of a large affected French pedigree. Since then, CADASIL has been diagnosed in several hundred pedigrees worldwide and now emerges as the most common form of hereditary stroke and vascular dementia. The onset of symptoms can occur from ages 25 to >60 years, but most commonly begins during the fourth to fifth decade. Notably, the disease manifests irrespective of vascular risk factors. Affected patients exhibit a variety of symptoms, including transient ischemic attacks and strokes (85%), cognitive deficits (60%), migraine with aura (30%), psychiatric disorders (30%), and epilepsy (up to 10%). The overall course of the disease is highly variable but gradually leads to cognitive decline, dementia, and premature death of the patient about 15–20 years after clinical onset. On magnetic resonance imaging (MRI) of the brain, there is evidence of widespread white-matter hyperintensities on T_2-weighted images, subcortical lacunar infarctions on T_1-weighted images, and old cerebral microhemorrhages on T_2*-weighted images or gradient-echo images. Virtually no therapies are available to modulate the progression of the disease.

On neuropathological examination, CADASIL brains show a diffuse myelin loss and multiple, small, deep infarcts located within the basal ganglia, the deep white matter, and the brain stem. The underlying vascular lesion is a unique nonatherosclerotic, nonamyloid arteriopathy involving primarily the small penetrating brain arteries. The arteriopathy is characterized by prominent alterations of vascular smooth muscle cells, which ultimately degenerate, and the thickening of arterial walls by the accumulation of various extracellular matrix components. The defining pathological hallmark is the deposit of granular osmiophilic material (GOM), of yet uncertain composition, within the basement membrane of smooth muscle cells. Similar arterial changes, particularly GOM deposits, are detected in the small arteries of many other organs, indicating that CADASIL is actually a systemic arteriopathy, although the symptoms are restricted to the central nervous system (Figure 1).

The mutated gene was identified through a positional cloning approach. The CADASIL locus was first mapped to the short arm of chromosome 19 by linkage analysis in a large affected family. Analysis of additional families revealed that CADASIL is a genetically homogeneous disorder. The critical region was narrowed to an 800 kb interval by identification of cross-over events in these families. CADASIL patients were shown to exhibit deleterious mutations in the *Notch3* gene contained within this critical interval. These mutations were absent in control individuals and cosegregated with the disease phenotype in affected families, demonstrating that *Notch3* is the causative gene of CADASIL.

The Notch 3 Receptor

Structure of Notch 3

NOTCH3 belongs to the *NOTCH* gene family, which encodes evolutionarily conserved single-pass

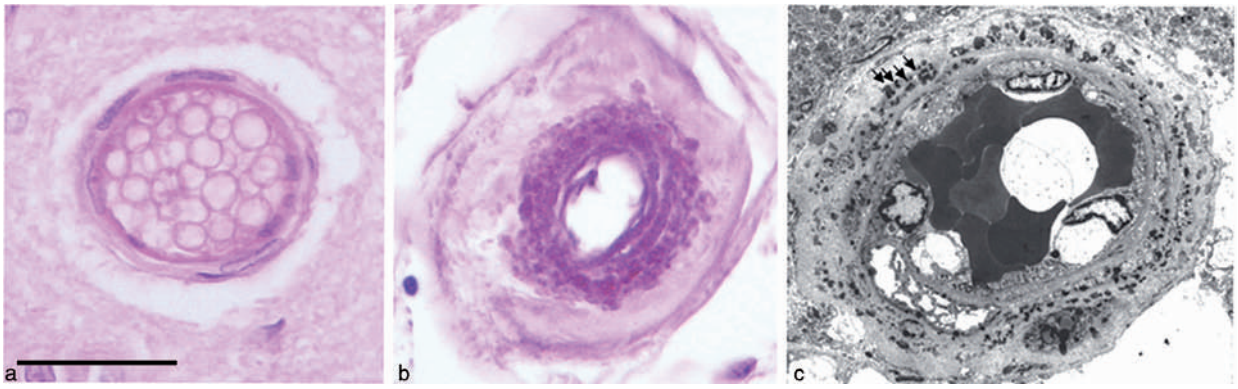

Figure 1 The characteristic arterial lesions in CADASIL patients. Section of a brain artery from a control individual (a) and an elderly CADASIL patient (b) (hematoxylin and eosin staining). The CADASIL patient's artery exhibits prominent thickening of the vessel wall, with accumulation of eosinophilic material, and disappearance of smooth muscle cells. Electron micrograph (right) of a CADASIL brain artery shows degeneration of smooth muscle cells and numerous deposits of granular osmiophilic material (arrows) located within the basement membrane of smooth muscle cells. Scale bar 30 μm = (a, b). Micrograph courtesy of MM Ruchoux.

transmembrane receptors. The family includes the *Drosophila melanogaster Notch* gene and four mammalian *Notch* genes (*Notch1–4*). Notch receptors share a common structural organization: a large extracellular domain containing a characteristic tandem array of 29–36 epidermal growth factor-like repeats (EGFRs) and three cysteine-rich Notch/Lin12 repeats, followed by a short transmembrane domain and an intracellular domain containing the RAM domain, six ankyrin repeats flanked by two nuclear localization signals. Notch 3 receptor is a 2321-amino-acid-long protein; it contains all typical Notch motifs – namely, an extracellular domain with 34 EGFRs. Notch 3, as the other Notch receptors, is synthesized as a precursor protein which is cleaved (S1) in two during the transport to the cell surface, forming a large ~210 kDa ectodomain ($Notch^{ECD}$) and a ~97 kDa membrane-tethered intracellular domain ($Notch^{TMIC}$). These products remain associated at the cell surface as a heterodimer through noncovalent dependent interactions (Figure 2).

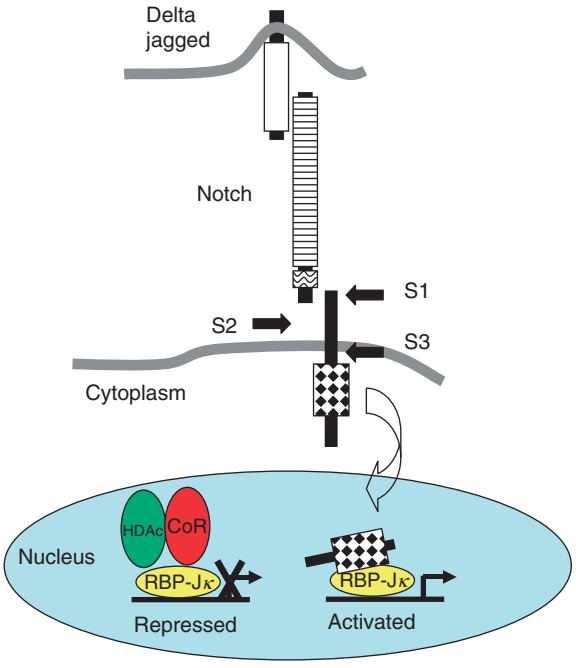

Figure 2 Schematic diagram of Notch signaling in mammals, showing S1, S2, and S3 cleavage steps and the RBP-Jκ proteins (see text for discussion). HDAc, histone deacetylase; CoR, corepressor.

Functional Role of Notch 3 in the Vessel

The Notch signaling pathway plays a central role in the development and maturation of most vertebrate organs, with pleiotropic effects depending on dose and context. In most cases Notch signaling blocks differentiation toward a primary differentiation fate in a cell and instead directs the cell to a second, alternative differentiation program or forces the cell to remain in an undifferentiated state. In addition to the classical role of 'gatekeeper against differentiation,' recent studies demonstrated that Notch signaling may also be instructive – that is, Notch signaling can actively promote a particular cell fate.

Analysis of the Notch 3 expression pattern in late embryos and adult tissues provided the first insight into its role. Examination of sections of various organs, including the brain, heart, skeletal muscle, lung, kidney, urinary bladder, and intestine, using *in situ* hybridization and immunohistochemistry, revealed that Notch 3 is predominantly expressed in blood vessels, with a preferential expression in arteries. Within the arteries, expression of Notch 3 is confined to smooth muscle cells, being absent in endothelial cells and adventitial fibroblasts. Notably, Notch 3 expression in not detected in the media (muscular wall) of large, elastic arteries. Moreover, Notch 3 is not expressed in visceral smooth muscle cells.

Careful analysis of mice completely lacking *Notch3* (*Notch3*$^{-/-}$) established a critical role for Notch 3 in the structural and functional integrity of small arteries, particularly of brain arteries. *Notch3* null mice are viable and fertile, demonstrating that the *Notch3* gene is not essential for embryonic development or fertility in mice. Adult *Notch3*$^{-/-}$ mice show no overt phenotype. However, upon detailed analysis, *Notch3*$^{-/-}$ mice were found to exhibit marked arterial defects affecting the distal arteries, but, remarkably, sparing the large elastic arteries. Mutant arteries are enlarged, with a thinner muscular coat and prominent alterations in the shape and size of smooth muscle cells. At the functional level, arterial myogenic responses are strongly impaired. In arteries of the cerebral circulation, myogenic responses underlie the autoregulation of cerebral blood flow (i.e., the maintenance of constant blood flow during variations in systemic blood pressure). Specifically, in response to angiotensin II or phenylephrine injections, *Notch3*$^{-/-}$ mice show normal arterial blood pressure elevations but strongly impaired cerebral blood flow (CBF) autoregulation: the slight increase in CBF elicited by acute induced hypertension in wild-type mice is greatly exacerbated in *Notch3*$^{-/-}$ mice.

Further investigations established that Notch 3 cell autonomously regulates the arterial differentiation and postnatal maturation of vascular smooth muscle cells. Specifically, in *Notch3* null mice, the postnatal maturation stage of smooth muscle cells, that lends arterial vessels their final shape, is deficient, and there is strong evidence that smooth muscle cells in mutant arteries lack an arterial fate and are converted to a venous fate. Remarkably, endothelial cells in mutant arteries retain an arterial fate. Until recently, it was thought that the genetic program specifying artery and vein identity was restricted to the endothelial cells. The current findings indicate that arteries are genetically distinct from veins also at the level of smooth muscle cells. Moreover, the signaling pathways controlling arterial identity of these two vascular cell types are regulated independently. Finally, while it has been shown that arterial specification of endothelial cells is required during early embryonic vascular development, to maintain boundaries between arteries and veins, these data indicate that arterial differentiation of smooth muscle cells is required to ensure integrity of the structure and function of small arteries in the adult (Figure 3).

How *Notch3* expression regulates the postnatal arterial maturation and differentiation of smooth muscle cells is not fully understood. *In vitro* studies suggest a direct effect of Notch 3 on the actin cytoskeletal dynamics of smooth muscle cells. Cultured smooth muscle cells expressing a constitutively active Notch 3 receptor exhibit dramatic cell shape changes, with a triangular or polygonal shape and very few cytoplasmic expansions and protrusions as compared to control cells. Moreover, smooth muscle cells expressing the constitutively active Notch 3 receptor show prominent actin stress fibers, an increased steady-state level of polymerized actin, and higher resistance to actin-depolymerizing drugs.

The Notch 3 Signaling Pathway

Activation of Notch receptors is triggered by interactions with ligands of the Delta/Jagged family expressed on juxtaposing cells. Ligand binding to Notch proteins triggers a tumor necrosis factor-α converting enzyme/disintegrin metalloproteinase (TACE/ADAM)-mediated cleavage (S2) within the extracellular juxtamembranous region, followed by a presenilin/γ-secretase-dependent proteolytic (S3) release of the Notch intracellular domain (NICD) from the membrane. NICD translocates to the nucleus and forms a transcriptional activation complex with a DNA-binding protein termed CBF1/RBP-Jκ, which regulates the transcription of target genes by displacement of corepressors and simultaneous recruitment of coactivators. A fairly limited number of Notch/RBP-Jκ target genes have been identified, and among these are the *Hes* and *Hrt* gene families that comprise in the mouse six and four members, respectively. *In situ* hybridization and real-time reverse transcriptase–polymerase chain reaction (RT-PCR) analysis showed that the expression level of these genes was unaltered in arteries of *Notch3*$^{-/-}$ mice, strongly suggesting that *Notch3* effects in arteries are not mediated by these genes. Thus the Notch 3 direct target genes in small vessels remain to be identified.

CADASIL: A Disordered Notch 3 Signaling Pathway?

Mutation Spectrum

Mutational analysis in a very large number of CADASIL families has revealed several key features. First, all mutations are located within the 23 exons that encode for the 34 EGFRs. Second, the mutations are strongly clustered within EGFR2–EGFR5, repeats that have no recognized specific function. Third, the mutations are highly stereotyped: the vast majority are missense mutations and a very few are splice site mutations or small in-frame deletions, but remarkably, all of the mutations lead to an odd number of cysteine residues. These data provided the basis for setting up a robust and reliable diagnostic test of CADASIL.

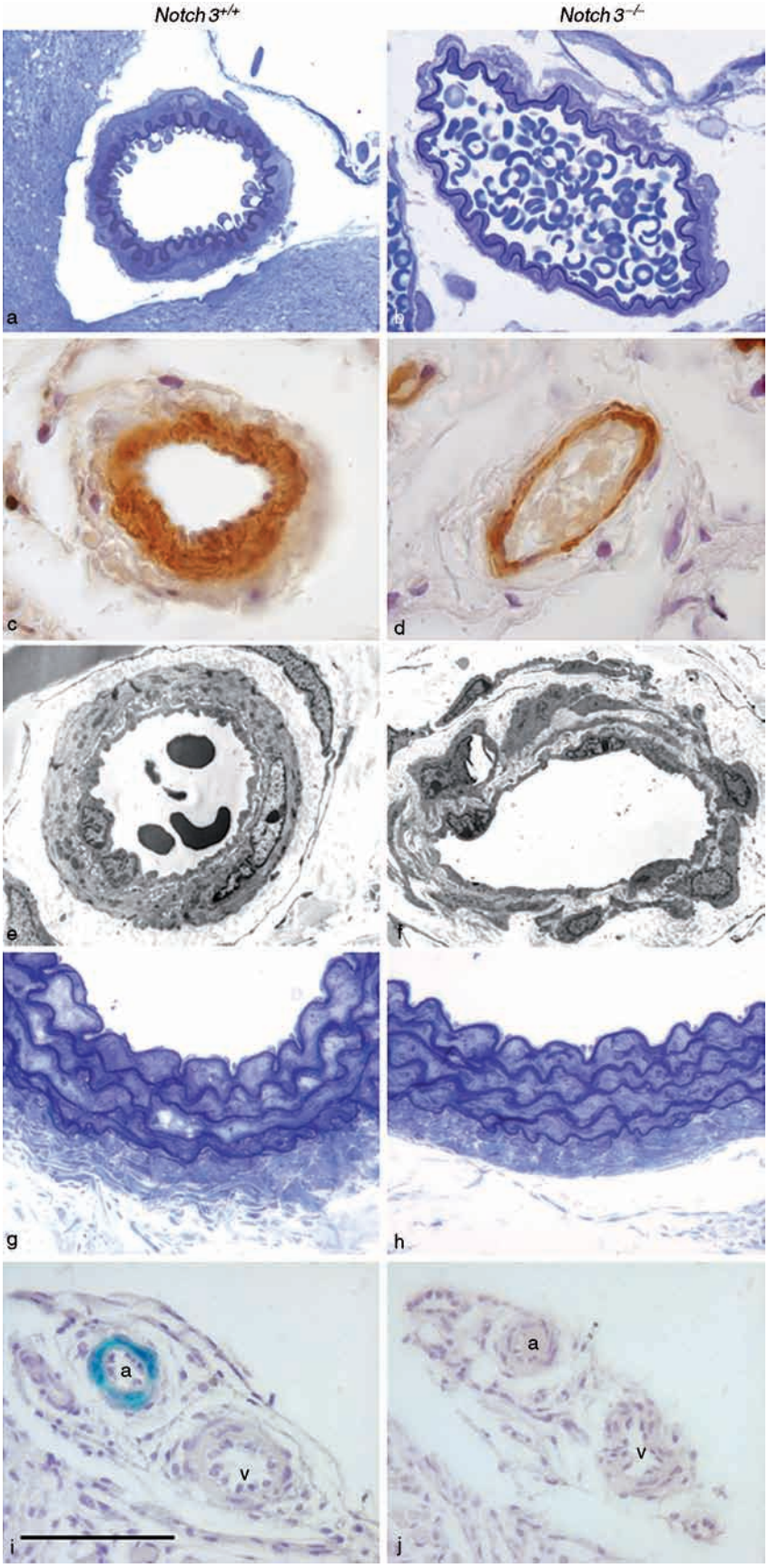

Figure 3 Continued

Accumulation of Aggregated Notch 3^{ECD} Protein

The highly stereotyped nature of pathogenic mutations detected in CADASIL patients suggest that an aberrant homo- or heterodimerization of Notch 3, due to abnormal disulfide bridging, may be involved in the pathogenesis of this condition. In tissues from CADASIL patients, immunohistochemical and biochemical studies show an abnormal accumulation of aggregated Notch 3 protein within the brain vessels as well as in the vessels of many other organs (e.g., kidney, skeletal muscle, skin). Importantly, there is a selective accumulation of the ectodomain of Notch 3 (Notch 3^{ECD}) without any accumulation of the Notch 3^{TMIC}. Immunoelectron microscopy revealed that Notch 3^{ECD} accumulates at the plasma membrane of smooth muscle cells. However, whether Notch 3^{ECD} is a component of GOM is a matter of debate. Amounts of Notch 3 transcripts are essentially identical in both CADASIL and control individuals, indicating that Notch 3 accumulation is unlikely to result from an increased synthesis of Notch 3 mRNA but rather from an impaired clearance of the protein.

Notch 3 and GOM aggregates are pathological hallmarks of CADASIL; early in the pathogenesis, they can be detected in the blood vessels before the appearance of clinical symptoms and brain MRI abnormalities. However, the question remains as to how they participate in the pathogenic process. The notion that ultrastructural alterations of smooth muscle cells can be detected prior to GOM and Notch 3^{ECD} deposits in CADASIL patients (or mouse models recapitulating the CADASIL arterial lesions) suggests that the deposits are an outcome rather than a primary causative event in the CADASIL pathogenic process (Figure 4).

Impact of CADASIL Mutations on Notch 3 Receptor Activity

The fact that all CADASIL mutations result in an odd number of cysteine residues within a given EGFR has suggested that CADASIL disease is associated with a gain-of-function activity of mutant protein. Gain-of-function mutations in principle include (1) hypermorphic mutations which elevate gene activity, (2) dominant-negative mutations which reduce wild-type gene activity, and (3) neomorphic mutations in which mutant protein has additional activity. However, it remains possible that CADASIL mutations are hypomorphic mutations that lower activity of the receptor, since there is ample evidence that Notch function is sensitive to gene dosage. Additionally, loss of *Notch3* in the mouse results in smooth muscle cell alterations, although they do not appear to phenocopy those seen in human CADASIL patients.

Effects of CADASIL mutations on Notch 3 signaling have been investigated in cultured cells. Contrasting with the highly stereotyped nature of the mutations, these *in vitro* studies failed to reveal an underlying common mechanism. On the one hand, a number of mutations, including those located in the mutational hot spot region (EGFR2–EGFR5), appear to retain the ability to activate RBP-Jκ in response to Delta/Jagged binding, although some mutations exhibit impaired maturation and decreased cell surface expression of the receptor. On the other hand, a few mutations have been identified within EGFR10–EGFR11, a domain required for ligand binding. *In vitro*, these mutations show strongly reduced signaling activity. Interpretation of these data is complicated by the fact that these experiments do not mimic the effect of *Notch3* mutations in human arteries, since mutant Notch 3 is vastly overexpressed, and, importantly, Notch 3^{ECD} accumulation at the smooth muscle cell plasma membrane, which could titrate the ligand and inhibit the normal pathway, is not recapitulated in these assays. Additionally, wild-type Notch 3 expressed in cultured cells may have confounded the interpretation of the data. Moreover, only the response to membrane-bound ligands of the Delta/Jagged families was investigated, and additional ligands have been now identified.

Conclusion and Future Prospects

CADASIL, an inherited small-vessel disease, emerges as the most prevalent heritable condition leading to stroke and vascular dementia. The causative gene encodes a transmembrane receptor, Notch 3, primarily expressed in smooth muscle cells of small arteries.

Figure 3 Specific arterial defects in *Notch3*$^{-/-}$ mice. (a, b) Semithin sections of the middle cerebral artery from *Notch3*$^{+/+}$ and *Notch3*$^{-/-}$ mice; the mutant artery is enlarged, with a flattened elastica lamina. (c, d) α-Smooth muscle actin immunostaining on wild-type and *Notch3*$^{-/-}$ arteries, showing that, in the mutant artery, smooth muscle cells are present but have an abnormally thin profile. (e, f) Electron micrographs of *Notch3*$^{+/+}$ and *Notch3*$^{-/-}$ arteries show the marked alteration in the shape and size of smooth muscle cells in the *Notch3*$^{-/-}$ artery. (g, h) Semithin sections of a wild-type and mutant aorta show no defect in the *Notch3*$^{-/-}$ sample. (i, j) *Notch3*$^{+/+}$ and *Notch3*$^{-/-}$ mice, heterozygous for the arterial smooth muscle-specific *SM22α-lacZ* transgene, were assayed for smooth muscle cell arterial identity. X-Gal staining is restricted to arterial smooth muscle cells in the *Notch3*$^{+/+}$ sample and is markedly reduced in the *Notch3*$^{-/-}$ artery, indicating a defective arterial fate of smooth muscle cells. a, artery; v, vein. Scale bars = 45 μm (a, b), 34 μm (c, d), 8 μm (e), 14 μm (f), 34 μm (g, h), 85 μm (i, j).

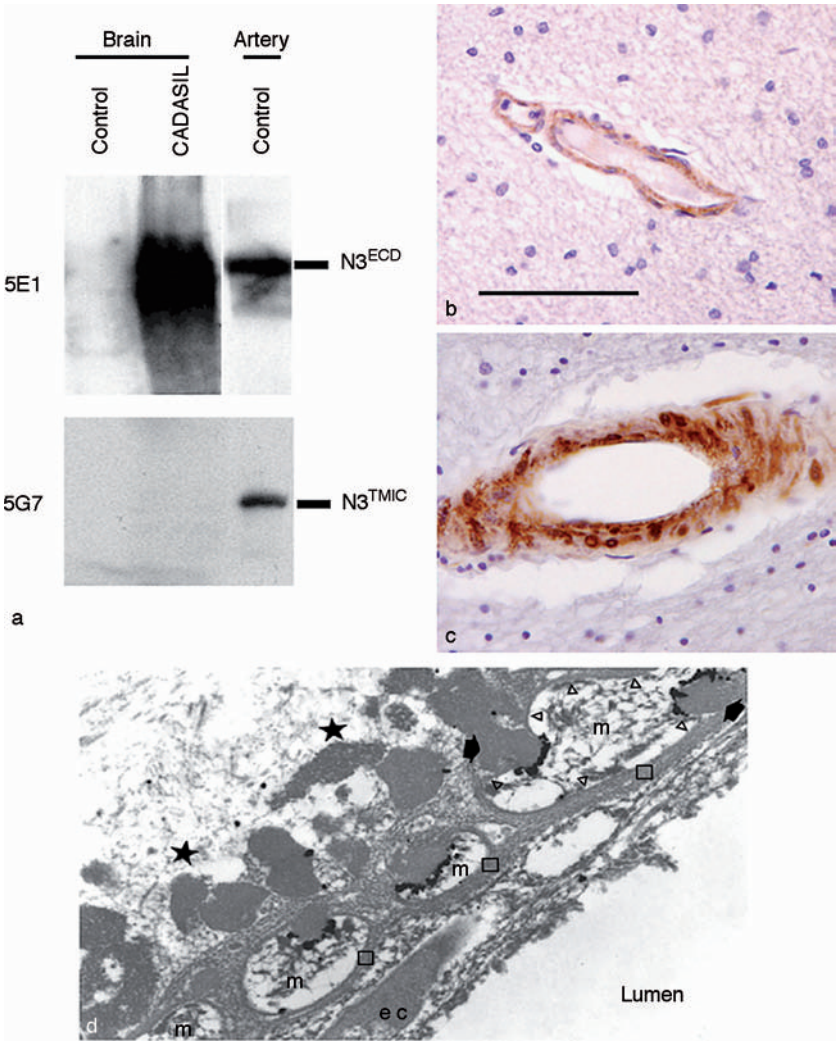

Figure 4 Notch 3^{ECD} ($N3^{ECD}$) abnormally accumulates at the plasma membrane of smooth muscle cells in CADASIL patients. (a) Protein extracts prepared from CADASIL and control human brains and from of a control human artery were immunoblotted with the anti-Notch 3^{IC} monoclonal antibody 5G7 (bottom) and the anti-Notch 3^{ECD} 5E1 monoclonal antibody (top). The 210-kDa Notch 3^{ECD} and the 97-kDa Notch 3^{TMIC} ($N3^{TMIC}$) are detected in the control artery but not in the control brain, owing to the very low expression level of Notch 3 and to its expression being restricted to smooth muscle cells. In the CADASIL brain, there is a prominent accumulation of the Notch 3^{ECD} fragment, but no accumulation of the 97-kDa Notch 3^{TMIC} fragment. (b, c) Brain paraffin sections from a control individual (b) and a CADASIL patient (c) were immunolabeled with the anti-Notch 3^{ECD} monoclonal antibody 1E4. In the control brain, Notch 3 antibody labels the arterial wall, with a weak and homogeneous staining. By contrast, the CADASIL brain artery exhibits a characteristic strong and granular staining. Scale bar = 85 μm. (d) CADASIL brain sample was processed for immunoelectron microsocopy analysis using the 1E4 anti-Notch 3^{ECD} antibody. Numerous granular osmiophilic material deposits (large arrow) are observed within the basal lamina (square), other granular osmiophilic material deposits spread far away, within the vessel wall (star). Notch 3 immunolabeling (black dots) is clustered at the cell surface of smooth muscle cells, closely arranged around the granular osmiophilic material. ec, endothelial cell; Lu, lumen) (ultrathin section stained with uranyl acetate and lead citrate; ×20 000).

Recent data document a critical role for Notch 3 in the elaboration and maintenance of small arteries. However, whether CADASIL arises from an alteration of the Notch 3 signaling pathway remains a fundamental yet unresolved issue. Notch 3 signaling components are attractive candidates for their involvement in other small-vessel diseases and as potential therapeutic targets in small-vessel diseases.

See also: Animal Models of Stroke; Notch Pathway: Lateral Inhibition; Notch Signal Transduction: Molecular and Cellular Mechanisms; Stroke; Stroke: Injury Mechanisms.

Further Reading

Artavanis-Tsakonas S, Rand MD, and Lake RJ (1999) Notch signaling: Cell fate control and signal integration in development. *Science* 284(5415): 770–776.

Chabriat H, Joutel A, Vahedi K, et al. (1998) CADASIL (Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy). In: Barnett HJM, Mohr JP, Stein BM, et al. (eds.) *Stroke: Pathophysiology, Diagnosis, and Management,* 3rd edn., pp. 933–943. New York: Churchill Livingstone.

Domenga V, Fardoux P, Lacombe P, et al. (2004) Notch3 is required for arterial identity and maturation of vascular smooth muscle cells. *Genes & Development* 18(22): 2730–2735.

Joutel A, Andreux F, Gaulis S, et al. (2000) The ectodomain of the Notch3 receptor accumulates within the cerebrovasculature of CADASIL patients. *The Journal of Clinical Investigation* 105(5): 597–605.

Joutel A, Monet M, Domenga V, et al. (2004) Pathogenic mutations associated with cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy differently affect Jagged1 binding and Notch3 activity via the RBP/JK signaling Pathway. *The American Journal of Human Genetics* 74(2): 338–347.

Joutel A, Vahedi K, Corpechot C, et al. (1997) Strong clustering and stereotyped nature of Notch3 mutations in CADASIL patients. *Lancet* 350(9090): 1511–1515.

Ruchoux MM and Maurage CA (1997) CADASIL: Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy. *Journal of Neuropathology & Experimental Neurology* 56(9): 947–964.

Schweisguth F (2004) Notch signaling activity. *Current Biology* 14(3): R129–R138.

Cadherins and Synapse Organization

S-H Lee and L F Reichardt, University of California at San Francisco, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Cadherins constitute a large superfamily of cell-to-cell adhesion molecules defined by the presence of multiple copies of a unique domain termed the cadherin motif in their extracellular domains. Cadherins mediate the formation of many types of cell–cell junctions, including adherens junctions and desmosomes in epithelia, adherens junctions and complexus adherens in endothelia, intercalated disks in cardiac myocytes, and synaptic junctions between axons and their target cells. The vast majority of the cadherins are transmembrane proteins with functionally important interactions mediated by both extracellular and cytoplasmic domains. Some cadherins also have functionally significant *cis* interactions that control their oligomerization and the strength of cell-to-cell adhesion.

Most of the more than 80 mammalian cadherins are expressed in the nervous system and many have been shown to be present at synapses and in some instances to regulate synapse development or function. Synapses are formed through a series of reciprocal inductive events in which contact between axons and their targets (muscle at the neuromuscular junction and dendrites in the central nervous system (CNS)) is followed by assembly of pre- and postsynaptic constituents. Many cadherin members function in synapse formation and plasticity (Table 1).

The classical cadherins were identified initially as Ca^{2+}-dependent cell adhesion molecules that mediated primarily homophilic adhesion. One of the cadherin's motifs present in a cadherin contain hydrophobic residues including one or two tryptophan residues mediating interactions with the same cadherin present on other cells. Gene sequence analysis has enlarged the cadherin superfamily membership, which can be subdivided into several subfamilies based on homology, structure, and function. The classical cadherins have been divided into type I and type II subfamilies, distinguished by the presence of conserved amino acid residues in the most N-terminal cadherin motif which results in differences in the structures of their interaction interfaces. All classical cadherins in vertebrates are single-pass transmembrane proteins that contain five tandem repeats of the cadherin motif in their extracellular domains and a conserved cytoplasmic domain that mediates through linker proteins named catenins association with the F-actin cytoskeleton. Invertebrates contain similar cadherins that also mediate interactions through the catenins with the F-actin cytoskeleton but have different numbers of extracellular cadherin motifs. The desmosomal cadherins, the desmogleins and desmocollins, constitute a distinct family of cadherins that mediate through their cytoplasmic domains interactions through the plakophilins and desmoplakins with intermediate filament networks. Because they are not found at synapses, they are not discussed further in this article. Additional cadherin subgroups include the seven-pass transmembrane or Flamingo cadherins, and the protocadherins. In addition, there are a number of proteins containing the cadherin motif that do not fit into these subfamilies, including the receptor tyrosine kinase c-ret, Fat, Ksp-cadherin, and LI-cadherin. c-ret is stimulated by the GDNF family of neurotrophic factors and is not discussed further in this article. Atypical glycosyl phosphatidylinositol-linked cadherin, T-cadherin, is expressed in neurons and regulates neurite outgrowth.

Classical Cadherins

Each member of the classical cadherins mediates primarily homophilic interactions with the same cadherin on neighboring cells. Of particular interest, several members have been shown to be associated with networks of interconnected neurons within the brain and to be present at synapses, suggesting a role in establishing specific synaptic connections. Type I cadherins, such as N-, E-, P-, and R-cadherins, and many different members of type II cadherins, such as MN-cadherin and cadherin-6, -8, and -11, are observed at the synapses.

At synapses, cadherins are typically found in puncta adherens-like structures within or surrounding synapses, suggesting they play a role in organizing synaptic membrane domains similar to their roles in organizing adherens junctions and surrounding membrane domains in epithelial cells. In epithelial cells, cadherins typically collaborate to form adherens junctions with a subfamily of the immunoglobulin superfamily, named nectins. Nectins promote cell adhesion but exhibit strongest interactions with related, but not identical, family members. In contrast to epithelia, synapses are not symmetric and it is not clear how cadherins functioning alone can ensure that axons form synapses with dendrites as opposed to other axons. Work has demonstrated that differential transport of different nectins to axonal versus dendritic

Table 1 The synaptic phenotypes upon perturbation and suggested function of main cadherin components

Class	*Molecules*	*System*	*Mechanism of inhibition and synaptic phenotypes*	*Function*
Classical cadherin	N-cadherin	Chick optic tectum *in vivo*	Ab inhibition: overshooting of retinal axons that fail to stop in normal target layers	Target recognition
		Thalamic slice culture	Ab inhibition: overshooting of thalamic afferent beyond layer IV to the pial surface	Target recognition
		Drosophila visual system	Genetic mutation: disruption of target selection by R1–R8 cell axons in the lamina and medulla	Target recognition
		Cultured hippocampal neurons	Dominant negative inhibition: longer dendritic protrusions; smaller spine head width; decreased presynaptic puncta; inhibition of late-phase LTP	Synaptic maturation
	Cadherin-11	Mouse hippocampus	Genetic mutation: increased LTP in CA1 region; altered fear- and anxiety-related responses	Synaptic plasticity
	β-Catenin	Hippocampal pyramidal neurons *in vivo* and *in vitro*	Genetic mutation: disruption in the localization of synaptic vesicles and synaptic transmission	Maturation of synapse
		Cultured hippocampal neurons	Phosphorylation mutant (Y654F): increased spine β-catenin levels; increased PSD size and mEPSC frequency	Synaptic plasticity?
	αN-catenin	Cultured hippocampal neurons	Genetic mutation: increased dendritic spine length and decreased spine stability Overexpression: increased spine density and stabilization of spines	Maturation of synapse
	p120 catenin	Hippocampal pyramidal neurons *in vivo* and *in vitro*	Genetic mutation: reduced dendritic spine and synapse densities *in vivo* and *in vitro*; misregulated Rho family GTPase activity	Maturation of synapse
	δ-Catenin	Hippocampal pyramidal neurons *in vivo*	Genetic mutation: cognitive defect; impaired short-term and long-term plasticity	Maturation of synapse; synaptic plasticity
Protocadherin	Protocadherin-γ	Mouse spinal cord	Genetic mutation: decreased spinal cord synapse density, reduced synaptic current amplitude, sensorimotor defect	Induction or maturation of synapse
Seven-pass transmembrane cadherin	Flamingo	*Drosophila* visual system	Genetic mutation: inappropriate target selection by R1–R6 cell axons in the lamina; disruption of local pattern of synaptic terminals in the medulla	Target recognition

compartments promotes the preferential interactions observed by axons with dendrites. In this manner, the nectins may help ensure that only appropriate synaptic contacts are formed or stabilized within the CNS. Mutations of several cadherin genes or perturbation of cadherin function impair normal synapse development and function in vertebrates and invertebrates.

Cadherins appear to play roles in several stages of synapse development. Synaptic adhesion molecules, including the cadherins, are important in directing the localization and specificity of interactions between presynaptic axons with postsynaptic targets. In many instances, cadherins are required for this target selection process. In the *Xenopus* retina, for example, impairment of N-cadherin function slows axon growth, thereby preventing axons from reaching their normal targets in the optic tectum (**Figure 1**). In elegant experiments analyzing the pattern of synaptic connections formed by photoreceptor axons in *Drosophila*, N-cadherin appears to mediate homophilic and attractive interactions between the growth cones of photoreceptors and their targets that precede

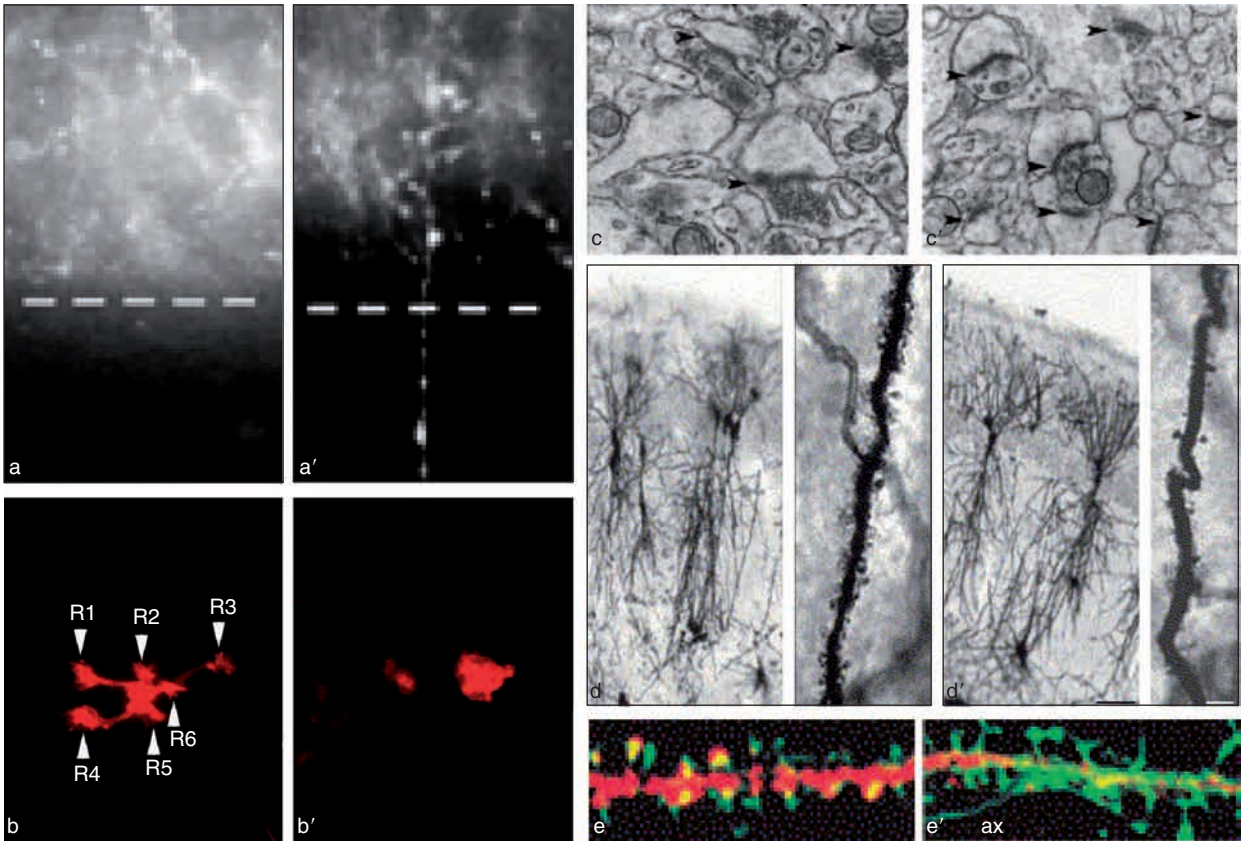

Figure 1 Importance of cadherin complex in synapse development. Examples of synaptic phenotypes resulting from perturbations of cadherin and catenin functions are shown. (a, a′) N-cadherin regulates lamina-specific connectivity in the brain. Injection of N-cadherin monoclonal antibody-producing hybridoma into chick optic tectum results in extension of some retinal axons past their normal target in laminae B and D in the optic tectum *in vivo*. Retinal axons were visualized by DiI injection. (b, b′) In *Drosophila* N-cadherin is required for R cell axon target connectivity. In cadherin mutant animals, the normal projection patterning by R1–R6 axons is disrupted and R cell growth cones fail to extend to their normal targets. DiI was injected into single ommatidia. (c, c′) β-Catenin plays a role in synaptic vesicle localization and presynaptic assembly. Deletion of β-catenin results in a reduction of vesicle number per synapse in the stratum radiatum of the CA1 hippocampus. (d, d′) p120catenin regulates dendritic spine development *in vivo*. Pyramidal neurons from the CA1 hippocampus which lacks p120 catenin exhibit reduced spine number compare to control neurons. Pyramidal neurons were visualized by Golgi staining. (e, e′) N-cadherin regulates dendritic spine morphogenesis of hippocampal pyramidal neurons. Neurons expressing cN390Δ, an N-cadherin mutant which lacks its extracellular domain, form immature and elongated filopodial-like spines and spines with bifurcated heads. Cells were stained with F-actin (green). Transfected cells are labeled in red. (a, a′) Inoue A and Sanes JR (1997) Lamina-specific connectivity in the brain: Regulation by N-cadherin, neurotrophins, and glycoconjugates. *Science* 276: 1428–1431. (b, b′) Dara provided by Thomas R Clandinini. (e, e′) Reproduced from Togashi H, Abe K, Mizoguchi A, et al. (2002) Cadherin regulates dendritic spine morphogenesis. *Neuron* 35: 77–89, Reprinted with permission from American Association for the Advancement of Science.

synaptic partner selection. Inhibition of N-cadherin function disrupts laminar-specific connectivity in chick optic tectum and results in overshooting of thalamic afferents beyond layer IV to as far as the pial surface in cortical–thalamic slice cultures. Cadherins are also required for synapse maturation. For presynaptic maturation, active zone proteins and synaptic vesicles are concentrated at synaptic contacts. During CNS development, presynaptic compartments are known to be assembled from preassembled packets of active zone proteins, synaptic vesicles, and associated synaptic proteins, and they are trapped at the sites of cell–cell contacts. The N-cadherin complex plays an important role in localizing synaptic vesicles at nascent synapses *in vitro*. In the CNS, many excitatory synapses are made onto dendritic spines, actin filament-based protrusions. N-cadherin-mediated contact stabilization and modulation of the actin cytoskeleton through interacting signaling molecules are essential for proper spine maturation *in vitro*. Finally, cadherins have a role in synaptic efficacy *in vivo*. The deletion of *cadherin-11* alters mouse behavior and synaptic plasticity. Taken together, these results indicate that cadherins have roles in synaptic development, specificity, and function. It seems almost certain that mutations in the many additional genes encoding classical cadherins will also prove to have interesting synaptic and

behavioral phenotypes, especially when mutations are combined to reduce functional redundancy. The large number of neuronally expressed classic cadherins complicates studies on their neural functions because of the potential for compensation.

In considering the roles of cadherins in directing synaptic development and function, it is important to consider the proteins that are associated with cadherins (Figure 2). Through these proteins, cadherin ligation regulates several distinct signaling pathways which impinge on neuronal development and synapse development and function. Two major classes of proteins associate directly with the classic cadherins: members of the β-catenin (β-catenin and plakoglobin) and p120catenin (p120ctn) (p120ctn, δ-catenin, p0071, and armadillo repeat protein deleted in velocardio-facial syndrome (ARVCF)) families. Members of each family contain a central domain consisting of several armadillo (arm) repeats and flanking N- and C-terminal domains. The cadherin cytoplasmic domains bind to the arm repeats of these proteins.

β-Catenin controls cell adhesion and cytoskeletal organization through interactions with α-catenin and with Rho family GTPase regulators, such as

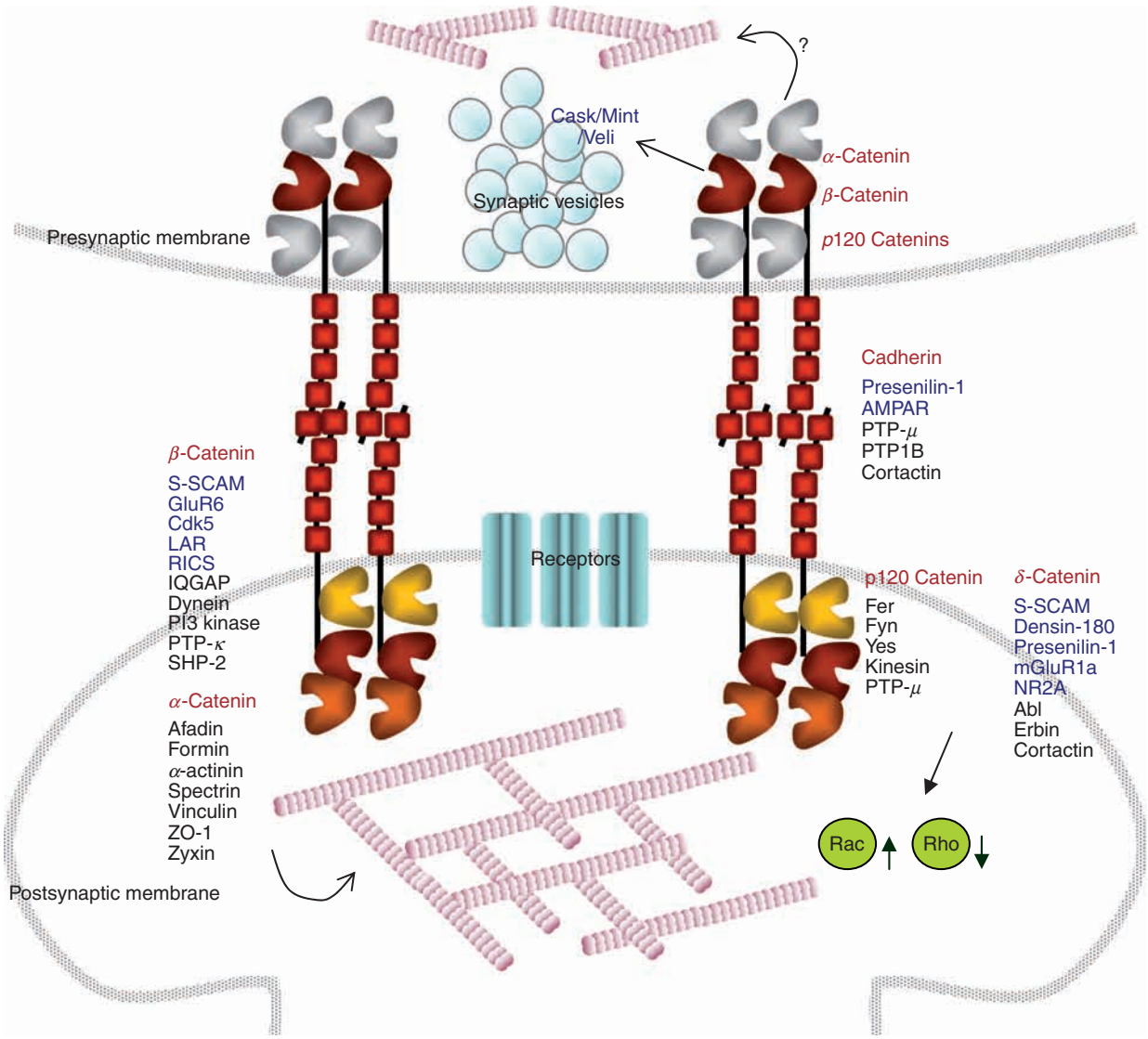

Figure 2 Protein constituents and interactions in cadherin adhesion complexes. Classical cadherins interact with catenins. Both cadherins and catenins also interact with diverse additional proteins, thereby regulating many intracellular signaling pathways. Reported interacting partners of cadherin, β-catenin, α-catenin, p120 catenin, and δ-catenin are listed. The proteins which have been shown to interact with cadherin components in the synapse are shown in blue. Cadherin signaling regulates the actin cytoskeleton through p120-catenin family-mediated regulation of Rho family GTPases and cortactin. Through association with cadherins mediated through β-catenin, α-catenin provides a physical link to actin-binding proteins and the F-actin cytoskeleton.

Ras-GAP-like protein with IQ motifs (IQGAP) and Rho GAP involved in the β-catenin-*N*-cadherin and NMDA receptor signaling (RICS). Some of these interactions are known to be competitive with cadherin binding. In addition, the presence of β-catenin is required for activation of the classical wingless-type MMTV integration site (Wnt) signaling pathway, and the role of β-catenin in this pathway is mediated through interactions with transcription factors such as the T-cell factor (TCFs), negative regulators such as ICAT (inhibitor of β-catenin and TCF), and regulators of protein stability such as adenomatosis polyposis coli (APC), axin, and glycogen synthetase kinase 3-β (GSK3β). Phosphorylation of cadherins and β-catenin regulates their interactions, functions, and turnover, and many protein kinases and phosphatases have been shown to control the stability of the cadherin–β-catenin complex following extracellular stimuli. In the synapse, for example, neural activity induces dephosphorylation and redistribution of β-catenin into spines, where it interacts with cadherins to influence synaptic size and strength. Brain-derived neurotropic/factor (BDNF) has been shown to regulate tyrosine phosphorylation and interaction of β-catenin with cadherin, thereby modulating synaptic vesicle localization and synapse density *in vitro*. Within cadherin-centered junctions, a major role of β-catenin is to provide a link from cadherins to α-catenin, which in turn promotes interactions with the F-actin cytoskeleton. The presence of β-catenin and α-catenin is required to promote strong adhesion between cells. The association between these two catenins is mediated through the N-terminal region of β-catenin, a domain containing multiple serine/threonine phosphorylation sites that regulate separately the association with α-catenin and β-catenin degradation. Dendritic spines on αN-catenin-deficient neurons are unstable and highly mobile. Loss of αN-catenin reduces the stability and size of synaptic contacts, possibly by disruption of the cadherin–F-actin linkage or perturbation of cadherin-mediated regulation of F-actin organization.

β-Catenin and most p120ctn family proteins include C-terminal post synaptic density-disc large-Z01 protein (PDZ) domain interaction motifs. Through these, cadherins can associate indirectly with the PDZ domain-containing scaffold proteins know to be crucial for synapse organization. The β-catenin C-terminus binds the Cask-Mint-Veli complex, the neuronal isoform of MAGI (MAGUKS (membrane-associated guanylate kinases) with inverted domain structure), S-SCAM (synaptic scaffold molecule), and additional PDZ domain-containing proteins, some of which modulate β-catenin's transcriptional functions. Through its C-terminus, β-catenin functions as a scaffold that promotes localization of the Cask-Mint-Veli complex and synaptic vesicles to synapses. β-Catenin has also been shown to promote dendrite growth and arborization through pathways independent of Wnt signaling. Through its interaction with MAGI-1b, β-catenin has been shown in nonneuronal cells to recruit the phosphatase phosphatase and tensin homlog (PTEN) to adhesion sites where it promotes their stability through localized suppression of Src-induced invasiveness in cancer cells. Src suppression is likely to promote adhesion site stability through control of Rho family guanine nucleotide exchange factors (GEFs) and possibly Rho family effectors such as myosin II. Through these interactions, PTEN may also regulate synapse stability. Another MAGUK family member which interacts with β-catenin, lp-dlg/KIAA0583, provides a potential alternative link through vinexin and vinculin to the F-actin cytoskeleton. Finally, there are poorly understood intramolecular interactions between the N- and C-terminal and arm domains in β-catenin that regulate the interactions of these domains with other proteins.

β-Catenin is also an essential intermediate in the classical signaling pathway activated by Wnt proteins, but largely separate pools of β-catenin appear to be involved in its cadherin-associated and Wnt pathway-mediated functions. Wnt signaling results in stabilization of β-catenin and then it interacts with TCF family and other transcription factors to regulate gene expression. As a result of its essential role in this major Wnt signaling pathway, β-catenin has many additional roles within the nervous system, including regulation of neuroepithelial precursor proliferation, forebrain patterning, and sensory neuron specification. Wnt7a/Dishevelled-1 signaling promotes presynaptic assembly and neurotransmitter release in mossy fiber–granular cell synapse in the cerebellum, but it is not clear whether this is mediated by β-catenin-dependent canonical or noncanonical signaling pathway. Finally, plakoglobin, which is primarily expressed in epithelia, is a homologue of β-catenin that functions similarly in promoting cell–cell adhesion. However, plakoglobin cannot replace β-catenin in the Wnt signaling pathway. Interest in plakoglobin has increased due to the observation that it is enriched in some synapses.

Members of the p120ctn family interact with the membrane-proximal portion of the cytoplasmic domains of the classical cadherins. All members of this protein family have ten putative arm repeats with high sequence identity (more than 45%) among members and flanking N- and C-terminal polypeptides with less homology. Except for p120ctn, these proteins include PDZ domain interaction motifs at their C-terminals capable of interacting with several PDZ proteins, including synaptic scaffold proteins.

p120ctn and δ-catenin are strongly expressed in the nervous system. Mice lacking δ-catenin are viable but have many behavioral deficits that have been attributed to perturbations of synaptic function. In the hippocampus of mutant animals, short-term plasticity, as assessed by paired-pulse facilitation, is impaired, whereas long-term plasticity, as assessed by measurements of long-term potentiation, is elevated following high-frequency stimulation but reduced following low-frequency stimulation. Long-term depression is also impaired at hippocampal synapses in these animals. The absence of p120ctn results in early mouse embryo lethality. Mice in which p120ctn is eliminated solely in the forebrain are viable and have severely perturbed neuronal morphologies. In particular, the dendrites of neurons form with few spines and the density of synapses is severely reduced. This phenotype can be recapitulated in cell culture and rescue experiments indicate that p120ctn enhances spine density through inhibition of Rho activation while it promotes spine maturation through interactions with cadherins, acting either upstream or downstream of the cadherins.

p120ctn family members have been shown to regulate cadherin turnover, intracellular signaling, and cytoskeletal organization in several cell types. First, the presence of p120ctn family members is important for stabilizing cadherins on the cell surface *in vitro* and *in vivo*. In their absence, cadherins are transported normally to the surface but are rapidly endocytosed and degraded. p120ctn family members also play important roles in the signaling functions of adherent junctions. Cadherin-mediated adhesive junctions recruit and activate numerous signaling proteins, including PI3 kinase, Rac, Rho, and Arp2/3. The p120ctn family members promote assembly of these signaling centers through recruitment of the tyrosine kinases Fer, Fyn, and Yes and phosphatases including SHP-1, PTP-μ, and DEP-1 and have been shown to substrates of several kinases including receptor tyrosine kinases and Src family kinases. In the nervous system, δ-catenin also interacts with neurotransmitter receptors such as mGluR1a and NR2A and also PDZ-containing scaffold proteins such as S-SCAM. p120ctn family members regulate actin filament organization through regulation of Rho GTPases, with p120ctn promoting activation of Rac and inhibition of Rho following cell adhesion. The C-terminal region of δ-catenin also interacts in a phosphorylation-dependent manner with cortactin, an activator of the Arp2/3 complex, thereby providing a second pathway for regulation of the F-actin cytoskeleton. In addition, p120ctn family members also interact with the microtubule network directly as well as indirectly through kinesin.

Somewhat analogously to the involvement of β-catenin in Wnt signaling, the arm repeats of p120ctn associate with the transcriptional repressor Kaiso and thereby control expression of Wnt11 and *Xenopus* morphogenesis. Finally, deletion of p120ctn from keratinocytes results in elevated Rho activity, which in turn promotes activation of nuclear factor kappa B (NF-κB) and a proinflammatory cascade. Deletion in acinar epithelial cells prevents normal gland development and results in delamination of epithelial cells, almost certainly as a consequence of decreased surface E-cadherin and/or NF-κB.

In summary, the classical cadherins have been shown to regulate synapse formation, function, and plasticity. Many of the functions of the cadherins are controlled through binding to two families of catenins, which in turn control cadherin complex signaling, association with the cytoskeleton, and association with molecules important in synaptic function, including transmitter receptors and synaptic vesicles. In the future, it will be important to understand how these interactions are regulated by signaling pathways activated through activity and neurotransmitter receptors because these seem likely to control alterations in synaptic efficiency, in part, through regulation of synaptic cell adhesion molecule function.

Seven-Pass (Flamingo-like) Transmembrane Cadherins

The prototypic member of the seven-pass transmembrane cadherin family, Flamingo, was identified as a protein that mediates interactions between epithelial cells that control planar cell polarity. In the planar cell polarity pathway, Flamingo acts through modulation of Wnt/frizzled signaling. Flamingo has nine cadherin repeats, two laminin G domains, five epidermal growth factor (EGF)-like domains, and seven transmembrane domains and has been shown to promote cell aggregation *in vitro*. Cadherin, EGF-like, LAG-like, seven pass receptor (Celsr)1, Celsr2, and Celsr3 are the mammalian orthologs of Flamingo and are expressed in brain and epithelial cells, exhibiting distinct expression patterns.

Flamingo has several actions that influence early steps in neuronal differentiation that precede synapse formation. First, in *Drosophila* sensory neurons, it acts autonomously to promote axon growth and prevent premature axon branching. In addition, Flamingo controls the timing and length of dorsal dendrites of these neurons. Interestingly, it functions to delay their initiation and thereby limit their extent of growth. Flamingo is expressed on both the axons and the dendrites of these neurons and some of the mutants that affect its function also affect its

localization, suggesting that proper localization is required for its function. It has been suggested that Flamingo engages in both homophilic and heterophilic interactions, the latter with unknown ligands, to control the extent and patterning of dendrite growth. Flamingo is also expressed in the *Drosophila* visual system, where its presence is required for normal targeting of R1–R6 photoreceptor axons to appropriate cartridges in the lamina and of R8 axons in the medulla. Flamingo is expressed on the surfaces of these growth cones during their invasion of the brain. Intriguingly, other genes known to be required in the planar cell polarity pathway are not required for normal target selection by *Drosophila* photoreceptor axons, so Flamingo appears to be acting through a distinct pathway to regulate axon guidance. However, the genetic pathway through which *Celsr3* controls axon tract development in mouse brain development seems likely to involve the planar cell polarity pathway because mice mutants in Celsr3 and frizzled (Fzd)-3, a member of the Frizzled family, have similar phenotypes.

Protocadherins

Protocadherins constitute a very large family of proteins containing up to seven cadherin-like motifs that are distinct from those present in classical cadherins, a single transmembrane domain, and distinct cytoplasmic domains. In mammals, many of these proteins are highly expressed in the nervous system, where several have been localized to synapses. Their expression patterns in the brain frequently correlate with functional CNS subdivisions, and it has been proposed that they have important roles in establishing and maintaining synaptic specificity. Four clusters (α, β, γ, and δ) of genes encoding these proteins have been identified. Products of the protocadherin-α cluster are also named the CNRs (cadherin-like neuronal receptors). In the α, β, and γ clusters, single exons, each of which is coupled to a separate promoter, encode the variable extracellular domains. In the α and γ clusters, the mRNA product derived from each exon is spliced to an exon encoding the transmembrane and cytoplasmic domain, resulting in expression of several distinct extracellular domains, each fused to a constant cytoplasmic domain. β Cluster expressed transcripts contain variable extracellular domains together with a transmembrane and cytoplasmic domain. They lack the shared, constant exon which is observed in the α and γ clusters. Genes in the fourth cluster, the δ-protocadherin genes, are distributed on different chromosomes and exhibit similar alternative splicing patterns to those observed in the α and γ clusters.

Studies have shown that, in addition to biallelic expression, there is monoallelic but combinatorial expression of variable exons of the protocadherin-α and -γ cluster in single Purkinje cells – that is, only one of the two chromosomes can be used as a template for transcription of a single variable exon, but more than one variable exon can be expressed within a single neuron. In principle, this is a powerful mechanism to generate neuronal diversity and identity. In contrast to the classical cadherins, however, there is comparatively little evidence that these proteins function through homophilic adhesion. Some reports suggest that they may interact with each other, recognize arginine-glycine-aspartic acid (RGD)-binding β_1 integrins, and interact with reelin. It has also been suggested that protocadherins control homophilic binding by classical cadherins. There is little information on the cytoplasmic signaling pathways regulated by the protocadherins. Fyn is reported to interact with CNR cytoplasmic domain. Pcdh18 has been shown to associate with Dab1, which has been shown to mediate reeling signaling upon phosphorylation by an Src family kinase such as Fyn.

Members of the α- and γ-protocadherin proteins have been localized to synapses and have been proposed to control synapse formation, specificity, or maintenance. Analysis of mice in which the entire protocadherin-γ had been deleted found that it resulted in early lethality with specific loss through apoptosis of interneurons in the spinal cord. In addition, the spinal cords of these animals had a lower density of synapses. Analysis of this mutant in a *Bax* gene mutant background, which prevents apoptosis, revealed a striking deficit in the density of both excitatory and inhibitory synapses. Synaptic function was also compromised in neurons cultured from these animals. The results provide direct evidence that this family of protocadherins promotes synaptic development and/or maturation. Currently, there is no evidence that they control synaptic specificity, but future studies on this and other protocadherins seem likely to be informative. Mechanisms through which these proteins control synapse development remain mysterious.

Fat-Related and Other Members of Cadherin Superfamily

The Fat cadherins are single-transmembrane pass proteins that typically contain in their extracellular domains 34 tandemly arrayed cadherin repeats, plus in the membrane proximal region one or two laminin A subunit G domains and several EGF motifs. Involvement of Fat cadherins in synapse formation or function is purely speculative. There are two and four Fat cadherin genes in *Drosophila* and mice,

respectively. Mouse Fat2 mediates homophilic binding and is detected solely in cerebellar granule cells, where it is localized to the axons of these cells, the parallel fibers. Its expression level is elevated during the time period of synapse formation by these cells and remains high at later times. Its localization suggests that it helps organize the parallel fibers, which form synapses primarily on the dendritic spines of Purkinje cells. Fat3 expression is more broadly expressed in the embryonic CNS, including the spinal cord, but there is no information on its CNS functions.

Cadherin 23 is a novel single-pass transmembrane cadherin with 27 extracellular tandemly organized cadherin repeats that is expressed in both hair cells in the inner ear and photoreceptors in the retina. Cadherin 23 is a component of the tip link in hair cell stereocilia and forms a complex with each of the other proteins whose genes are mutated in Usher syndrome, a syndrome that results in deafness and blindness resulting from stereoclia disorganization. Through a C-terminal PDZ interaction motif, cadherin 23 has been shown to associate with the PDZ domain-containing scaffold protein harmonin. Harmonin is an F-actin bundling protein that also associates protocadherin-15, myosin VIIa, and a scaffold protein called Sans which contains several ankyrin repeats plus a sterile alpha-motif (SAM) domain. In addition, several transmembrane proteins, including usherin (a putative cell adhesion molecule), Usher's Syndrome type 2B (USH2B; a putative ion cotransporter), the very large G-protein-coupled receptor USH2C, and the four-pass transmembrane protein clarin, also associate with this complex. Mutations in each of the genes encoding these proteins also result in characteristic forms of Usher syndrome. In addition to their roles in organization of stereocilia, several of these proteins are expressed at photoreceptor synapses, suggesting that they have additional functions at these synapses.

The cytoplasmic domain of mammalian Fat1 has been shown to bind Ena/Vasp proteins and to regulate actin dynamics at cell–cell contacts. In *Drosophila*, the absence of Fat, but not of Fat-like, results in hyperproliferation of cells in imaginal disks. Fat is also involved in regulation of planar cell polarity, where it appears to modulate the strength of Wnt-Frizzled signaling, although the mechanisms through which it interacts with the planar polarity pathway are not understood. It seems likely that future studies will reveal diverse functions for these proteins, including roles at the synapse, but currently these are not clear.

Conclusion

This article summarized the current knowledge on the regulation and function of cadherin superfamily proteins known to be expressed at synapses or involved in synapse formation. Many of the mechanistic studies have been performed in nonneuronal cell types and it will be fascinating to determine whether these mechanisms govern their synaptic functions. Future studies are also expected to reveal the functions of individual genes and proteins in different types of synapse at different stages in their development.

See also: Axonal Pathfinding: Extracellular Matrix Role; Cell Adhesion Molecules at Synapses; Dendrite Development, Synapse Formation and Elimination; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Glia and Synapse Formation: An Overview; Postsynaptic Development: Neuronal Molecular Scaffolds; Postsynaptic Specialization Assembly; Presynaptic Development and Active Zones.

Further Reading

Bamji SX, Shimazu K, Kimes N, et al. (2003) Role of beta-catenin in synaptic vesicle localization and presynaptic assembly. *Neuron* 40: 719–731.

Elia LP, Yamamoto M, Zang K, and Reichardt LF (2006) p120 catenin regulates dendritic spine and synapse development through Rho-family GTPases and cadherins. *Neuron* 51: 43–56.

Hirayama T and Yagi T (2006) The role and expression of the protocadherin-alpha clusters in the CNS. *Current Opinion in Neurobiology* 16: 336–342.

Inoue A and Sanes JR (1997) Lamina-specific connectivity in the brain: Regulation by N-cadherin, neurotrophins, and glycoconjugates. *Science* 276: 1428–1431.

Kimura H, Usui T, Tsubouchi A, and Uemura T (2006) Potential dual molecular interaction of the *Drosophila* 7-pass transmembrane cadherin Flamingo in dendritic morphogenesis. *Journal of Cell Science* 119: 1118–1129.

Lilien J and Balsamo J (2005) The regulation of cadherin-mediated adhesion by tyrosine phosphorylation/dephosphorylation of beta-catenin. *Current Opinion in Cell Biology* 17: 459–465.

Prakash S, Caldwell JC, Eberl DF, and Clandinin TR (2005) *Drosophila* N-cadherin mediates an attractive interaction between photoreceptor axons and their targets. *Nature Neuroscience* 8: 443–450.

Reiners J, Nagel-Wolfrum K, Jurgens K, Marker T, and Wolfrum U (2006) Molecular basis of human Usher syndrome: Deciphering the meshes of the Usher protein network provides insights into the pathomechanisms of the Usher disease. *Experimental Eye Research* 83: 97–119.

Takeichi M and Abe K (2005) Synaptic contact dynamics controlled by cadherin and catenins. *Trends in Cell Biology* 15: 216–221.

Togashi H, Abe K, Mizoguchi A, et al. (2002) Cadherin regulates dendritic spine morphogenesis. *Neuron* 35: 77–89.

Cajal's Place in the History of Neuroscience

J DeFelipe, Instituto Cajal (CSIC), Madrid, Spain

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Santiago Ramón y Cajal was born on 1 May 1852, in Petilla de Aragón (Figure 1), a small village in Navarra (Northern Spain), and he died on 17 October 1934, in Madrid. He studied medicine at the Faculty of Medicine in Zaragoza and was professor of anatomy and histology at the universities of Valencia, Barcelona, and Madrid (Figure 2). He was also the director of the Laboratorio de Investigaciones Biológicas (Biological Research Laboratory) for more than 20 years, a facility later to be incorporated into a research institute that was inaugurated in 1932 and, through the kindness of His Majesty Alfonso XIII, was named the Instituto Cajal.

Cajal published almost 300 articles and several books of great importance (see later) and he received numerous awards and distinctions (Figure 3); among these were some of the most prestigious awards of his time, including the Moscow Award (1900) and the Helmholtz Gold Medal (1905); he also received the Nobel Prize for Physiology or Medicine (1906), which he shared with Camillo Golgi (1843–1926). He was also an important figure in the development of science and culture in Spain. For example, he not only published several nonscientific books (e.g., 1905; *Cuentos de Vacaciones*, Madrid: Fortanet) but he also founded the journals *Revista Trimestral de Histología Normal y Patológica* in 1888 and *Revista Trimestral Micrográfica* in 1896 (later called *Trabajos del Laboratorio de Investigaciones Biológicas de la Universidad de Madrid*). He was also a pioneer in the development of color photography (1912; *La Fotografía de los Colores. Fundamentos Científicos y Reglas Prácticas*, Madrid: Moya) and served as the director of several institutions, the most important being the Junta para la Ampliación de Estudios, founded in 1907. This institution was created with the idea of promoting science, culture, and education in Spain, and Cajal was its first president. During the Spanish civil war the activities of the Junta practically ceased, all its centers and laboratories were closed, and many of its members went into exile. In 1939, Franco's new regime converted the centers and laboratories of the Junta into the Consejo Superior de Investigaciones Científicas, currently one of the most important Spanish scientific research institutions.

The emergence of Cajal's work in the field of neuroscience produced a dramatic change in the course of neuroscience history. However, his scientific career did not experience an easy beginning, as, unlike most scientists of his era, he was not introduced to the field under the direction of an acknowledged researcher; Cajal became a prominent neurohistologist essentially on his own. Also, in contrast to many other great researchers, Cajal did not make a single, unique great discovery, but instead made numerous diverse and important contributions to our understanding of the entire structure and function of the nervous system. His observations regarding degeneration and regeneration, together with his ideas about the development and plasticity of the nervous system, not only had a deep influence on the scientists of his time, but they also constitute a virtually inexhaustible source of inspiration at present. This is thanks to Cajal's correct interpretation of the microscopic images obtained first with Golgi's 'reazione nera' (black reaction) impregnation technique, and later, the other methods that he developed.

It is clear today that our knowledge of the nervous system is certainly the result of the collective work of a large number of scientists. Nevertheless, the research of Cajal contributed more than that of anybody else in creating the scientific atmosphere necessary for the birth of modern neuroscience. The genius of Cajal resided in his extraordinary capacity of observation and his interpretation of the microscopic images; he was able to extract from these images the fundamental features of neuron and glial morphology. This talent was coupled to his ability to produce schemes summarizing his observations and to show the possible functional implications of the neuronal circuits based on connections by contacts ('neuron doctrine'), later called synapses. Indeed, all scientists of Cajal's era had the same microscopes and produced similar histological preparations, but it was Cajal who clearly saw details that were unnoticed by others or that were erroneously interpreted. An interesting example is Cajal's discovery of the dendritic spines using the Golgi method, in 1888 (see later). These were elements that he considered fundamental to the nervous system, while others, like Golgi, considered them to be mere histological artifacts.

Summary of the Scientific Career of Cajal

Cajal's career can be divided into three major phases. The first of these began in 1877 with the completion

Figure 1 Photographs taken in 2001 of the village of Petilla de Aragón and the house where Cajal was born.

Figure 2 Portrait of Cajal and his collaborator Juan Bartual Moret at Cajal's laboratory in Valencia (around 1885). Copyright by Herederos de Santiago Ramón y Cajal.

of his doctoral thesis, 'Patogenia de la Inflamación,' and extended until 1887 when he was introduced to Golgi's method of impregnation. During this period he published various histological and microbiological studies (some of which were signed with the pseudonym Dr. Bacteria) that were generally of relatively little significance.

In the second phase of his career (1887–1903) Cajal's research was very productive, and it was the period in which he exploited the Golgi method to describe almost every part of the nervous system in detail. During this period of his life he was to formulate important theories, the most popular being the neuron doctrine and the law of dynamic polarization of the neuron. The histological descriptions were so accurate that his classic book summarizing these studies, *Textura del Sistema Nervioso del Hombre y de los Vertebrados*, is still a title of reference in many neuroscience laboratories. Indeed, if each cell type that Cajal discovered was given his name, there would be so many types of Cajal cells that it would make neuroscientist's life impractical. Cajal also published several classic articles during the first few years of this second phase, including La rétine des vertébrés (1892; *La Cellule* 9: 121–133), La fine structure des centres nerveux (1894; *Proceedings of the Royal Society of London* 55: 444–468), and his masterworks on the human cerebral cortex. Furthermore, Cajal discovered and named the axonal growth cone ('cono de crecimiento') and devised the hypothesis of chemotaxis or chemotactism, later to be called neurotropism. At present, these early contributions represent two of the most exciting fields of research in neuronal development. In addition, Cajal made important contributions to general pathology, and to cancer research in particular. These contributions are evident in the articles he published and especially in his book *Manual de Anatomía Patológica General*, the first edition of which appeared in 1890

Figure 3 A plaque and some medals that Cajal received in recognition of his studies. The images of the medals were kindly supplied by Dr. Pere Berbel.

(Barcelona: Imprenta de la Casa Provincial de Caridad) and that he updated in successive editions until the end of his career.

In the third phase of Cajal's career, which began in 1903 and ended with his death in 1934, he experimented with a new technique, the reduced silver nitrate method. This period was devoted mainly to research into degeneration and regeneration in the nervous system. He published numerous articles on this subject that were of great relevance and that were summarized in another classic book, *Estudios sobre la Degeneración y Regeneración del Sistema Nervioso*. During this phase, Cajal also published important papers on the structure of the retina and optic centers of invertebrates, and the classic review article on the neuron theory, 'Neuronismo o reticularismo.' Therefore, rather than attempting to present here an exhaustive review of the work and influence of Cajal in neuroscience, the focus instead is on some of Cajal's first studies with the Golgi method, which represent the starting point of

the birth of modern neuroscience. In particular, two theories are discussed: the neuron doctrine and the law of dynamic polarization of neurons.

Cajal and the Golgi Method: The Neuron Doctrine and the Commencement of a New Era in Neuroscience

Before the studies of Golgi and Cajal, very little was known about the microanatomy and functioning of the nervous system, mainly because there were no histological techniques available to visualize an entire neuron (cell body, dendrites, and axon), or to follow the course of the axons up to its terminal arbor (Figure 4). Thus, it was not possible to address the critical issue of neuronal connectivity. At that time, the most common view regarding the organization of the nervous system was the reticular theory, developed principally by Joseph von Gerlach (1820–96). von Gerlach was considered to be the father of this theory, which states that the nervous system consists of a diffuse network that is formed by the anastomosing branches of the nerve cell processes (dendrites and axons), the cell somata principally playing a role in nourishment.

The structure and functioning of the nervous system truly began to be understood when Cajal started to try to unravel its organization using the Golgi method. Without this method, Cajal would probably have passed into the history of neuroscience as just another of many good scientists but one whose work was of no exceptional relevance. However, as we will see, in the end it is likely that without Cajal neuroscience would have suffered an important delay in its development. Accordingly, the Nobel Prize that Cajal and Golgi shared in 1906 is absolutely justified. Certainly, Cajal's studies on degeneration and regeneration were also of extraordinary importance and they were performed with a different method, reduced silver staining (see later). Furthermore, he used a variety of other methods to reveal a number of important findings. Nevertheless, the intellectual skeleton necessary for the interpretation of the new experiments and techniques was constructed thanks to the ideas he obtained from his studies with the Golgi method. However, Golgi was not the only scientist responsible

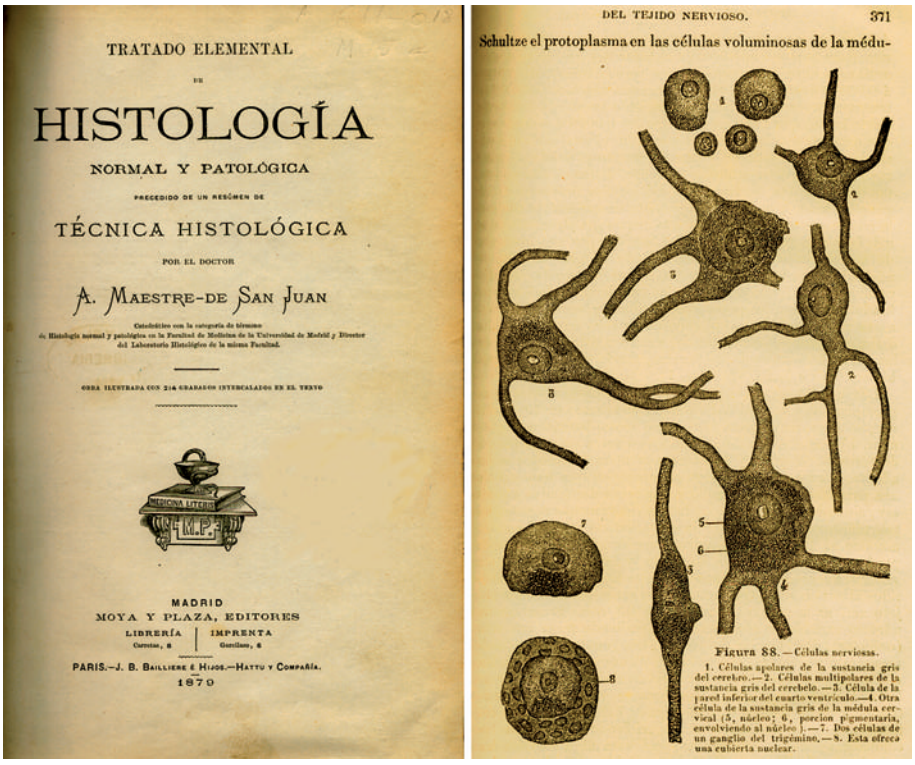
TRATADO ELEMENTAL
DE
HISTOLOGÍA
NORMAL Y PATOLÓGICA
PRECEDIDO DE UN RESÚMEN DE
TÉCNICA HISTOLÓGICA
POR EL DOCTOR
A. MAESTRE-DE SAN JUAN
MADRID
MOYA Y PLAZA, EDITORES
LIBRERÍA | IMPRENTA
PARIS.—J. B. BAILLIERE É HIJOS.—HATTU Y COMPAÑÍA.
1879

DEL TEJIDO NERVIOSO. 371
Schultze el protoplasma en las células voluminosas de la médu-
FIGURA 88.—Células nerviosas.
1. Células apolares de la sustancia gris del cerebro.—2. Células multipolares de la sustancia gris del cerebelo.—3. Célula de la pared inferior del cuarto ventrículo.—4. Otra célula de la sustancia gris de la médula cervical (5, núcleo; 6, porcion pigmentaria, envolviendo al núcleo).—7. Dos células de un ganglio del trigémino.—8. Esta ofrece una cubierta nuclear.

Figure 4 Pages from a nineteenth-century text; the right-hand page depicts morphological types of nerve cells. When Cajal was studying for his doctorate and attended the doctoral course of normal and pathological histology in 1877, one of the principal texts used was the *Tratado de Histología Normal y Patológica* (*Treatise of Normal and Pathological Histology*) by Aureliano Maestre de San Juan, professor of histology at the Faculty of Medicine in Madrid. The description of the structure of the nervous system was rather simple in these early publications. For example, Cajal would have read the following description regarding the morphological types of nerve cells in the *Tratado de Histología*: "The nerve cell may or may not present prolongations of its protoplasm, and from that come their division in 1°., apolar, that is to say, without any protoplasmic prolongation . . .; and 2°., others [nerve cells] that show a variable number of prolongations."

for the success of Cajal in his scientific career. There were also another three scientists who were fundamental to the development of his career: Aureliano Maestre de San Juan (1828–90), and, more particularly, Luis Simarro (1851–1921) and Albrecht Kölliker (1817–1905). While it was Maestre who introduced Cajal to the world of the microscope, Simarro introduced him to the Golgi method. In contrast, Kölliker 'discovered' Cajal, which was fundamental in disseminating Cajal's observations, concepts, and theories throughout the scientific community.

Interestingly, when Golgi examined the silver-impregnated preparations, he concluded that the reticular theory enunciated by Gerlach was wrong, since Golgi thought that dendrites ended freely and that only the axons and their collaterals were seen to anastomose. Therefore, he suggested that the nervous system consisted of a "rete nervosa diffusa" (diffuse nervous network), an idea that he supported even during his Nobel Prize lecture. It was in this scientific atmosphere that Simarro first showed Cajal a Golgi-impregnated preparation in 1887. Beautifully retained even in the translations of several of Cajal's writings is the historical moment when Cajal discovered the properties of the Golgi method (see Figures 5 and 6):

> In summary, a method was necessary to selectively stain an element, or at most a small number of elements, that would appear to be isolated among the remaining invisible elements. Could the dream of such a technique truly become reality, in which the microscope becomes a scalpel and histology a fine [tool for] anatomical dissection? A piece of nervous tissue was left hardening for several days in Müller's pure liquid [potassium dichromate] or in a mixture of this [fixative] with osmic acid. Whether it was the distraction of the histologist or the curiosity of the scientist, the tissue was then immersed in a bath of silver nitrate. The appearance of gleaming needles with shimmering gold reflections soon attracted the attention. The tissue was cut, and the sections were dehydrated, cleared, and then examined [with the microscope]. What an unexpected spectacle! On the perfectly translucent yellow background sparse black filaments appeared that were smooth and thin or thorny and thick, as well as black triangular, stellate or fusiform bodies! One would have thought that they were designs in Chinese ink on transparent Japanese paper. The eye was disconcerted, accustomed as it was to the inextricable network [observed] in the sections stained with carmine and hematoxylin where the indecision of the mind has to be reinforced by its capacity to criticize and interpret. Here everything was simple, clear and unconfused. It was no longer necessary to interpret [microscopically] the findings to verify that the cell has multiple branches covered with 'frost,' embracing an amazingly large space with their undulations. A slender fibre that originated from the cell, elongated over enormous distances and suddenly opened out in a spray of innumerable sprouting fibres. A corpuscle confined to the surface of a ventricle where it sends out a shaft, which is branched at the surface of the [brain], and other cells [appeared] like comatulids or phalangidas. The amazed eye could not be removed from this contemplation. The dream technique is a reality! The metallic impregnation has unexpectedly achieved this fine dissection. This is the method of Golgi! [...] whose clear and decisive images enables us to cast off the famous net of Gerlach, [as well as] of the [dendritic] arms of Valentin and Wagner, and of so many other fanciful hypothesis. (Cajal Histologic du systéme nerveux del'homme et des vertébrés)

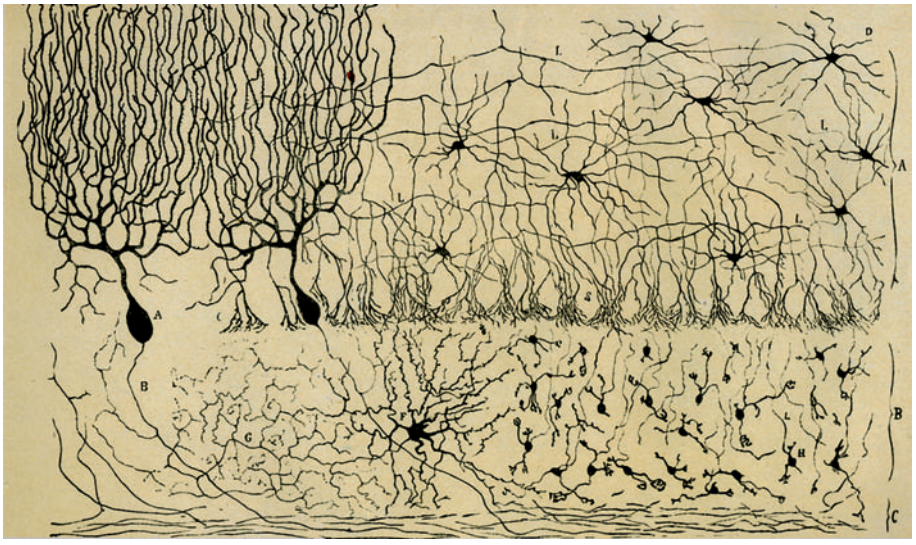

Figure 5 First illustration by Cajal from a Golgi-impregnated preparation of the nervous system. Cajal provided the caption: "Vertical section of a cerebellar convolution of a hen. Impregnation by the Golgi method. A, represents the molecular zone, B, designates the granular layer and C, the white matter." From Cajal SR (1888) Estructura de los centros nerviosos de las aves. *Revista Trimestral de Histolología Normal y Patológica* 1: 1–10. Copyright by Herederos de Santiago Ramón y Cajal.

Figure 6 Cajal's drawing of Golgi-impregnated neuroglial cells. The figure caption provided by Cajal stated: "Neuroglia of the superficial layers of the cerebrum: child of two months. Golgi Method. *A*, *B*, *[C]*, *D*, neuroglial cells of the plexiform layer; *E*, *F*, [*G*, *H*, *K*], *R*, neuroglial cells of the second and third layers; *V*, blood vessel; *I*, *J*, neuroglial cells with vascular [pedicles]." This figure was reproduced as figure 697 in the book *Textura del Sistema Nervioso del Hombre y de los Vertebrados*. Copyright by Herederos de Santiago Ramón y Cajal.

Cajal's typical vivid writing style and enthusiasm, appreciated here, were especially represented in the French translation of the *Textura*. (Note: The comatulids of which Cajal wrote are marine crinoid invertebrates such as sea lilies and feather stars. Phalangidas (or opiliones), also known as water harvestmen, are arachnids that superficially resemble true spiders but have small, oval-shaped bodies and long legs. In the preceding extract, Cajal is probably referring to some neuroglial cells that, when stained with the Golgi method, have a morphology resembling that of these invertebrates; see Figure 6).

Cajal was captivated by this marvelous staining and immediately started to use it to analyze practically the entire nervous system in several species. One year later after the meeting with Simarro, Cajal had published his first important article based on results obtained with this method in the avian cerebellum (Figure 5). In this study, entitled 'Estructura de los centros nerviosos de las aves,' published in 1888, Cajal made two great contributions. First, he described for the first time the existence of dendritic spines (which he also named), structures that currently generate particular interest, as they are highly plastic, and the main postsynaptic site of excitatory synapses in the cerebral cortex. Second, Cajal confirmed Golgi's conclusion that dendrites end freely, but in contrast to Golgi, Cajal added the critical conclusion that this also applies to axons and their branches:

> We have carried out detailed studies to investigate the course and connections of the nerve fibers in the cerebral and cerebellar convolutions of the human, monkey, dog, etc. We have not been able to see an anastomosis between the ramifications of two different nervous prolongations, nor between the filaments emanating from the same expansion of Deiters [axons]. While the fibers are interlaced in a very complicated manner, engendering an intricate and dense plexus, they never form a net [...] it could be said that each [nerve cell] is an absolutely autonomous physiological canton [unit]. (Cajal, translation – *Revista Trimestral de Histolología Normal y Patológica* 1, 1888)

Cajal and the Neuron Doctrine

During the next few years, Cajal provided many examples from throughout the nervous system to support his observation that dendrites and axons end freely (Figure 7). Indeed, the early studies that Cajal performed with the Golgi method were so decisive that they represented the main core of the review published in 1891 in the journal *Deutsche Medizinische Wochenschrift* (vol. 17) by Wilhelm von Waldeyer-Hartz (1836–1921). In this article, the term 'neuron' was introduced to denominate the nerve cells, and the so-called neuron doctrine become popular. By the end of the nineteenth century this theory was the most accepted theory to explain the organization of the nervous system, in which the neuron was considered as the anatomical, physiological, genetic, and metabolic unit of the nervous system. However, there were some notable exceptions, since in addition to the reticular ideas of Golgi, some prestigious researchers, such as Stephen von Apáthy (1863–1922), Hans Held (1866–1942), and Albert Bethe (1872–1954), were ardent supporters of the reticular theory. In particular, Bethe started to publish several influential experimental papers in 1901 on the regeneration of peripheral nerve fibers. In these writings he claimed that axons that had separated from their trophic centers were capable of autoregeneration, supporting the polygenetic theory of axon growth and regeneration claimed by earlier physiologists and pathologists. This theory states that the regenerated axons could result from the fusion of numerous axon segments due to the differentiation and transformation of the ensheathing cells of the old nerve fibers, a theory not compatible with the neuron doctrine. This was one of the main reasons why Cajal started another major project to study degeneration and regeneration in the nervous system. In this study, he mainly used the reduced silver nitrate method, a new technique that he had developed in 1903, which enabled him to examine the regenerated axons in greater detail than was possible with any other staining technique available at that time. From the beginning of these studies, it was clear to Cajal that the new fibers, which appeared in the peripheral stump of a cut nerve, originated from the sprouting of axons of the central stump, and with this the neuron doctrine once again became

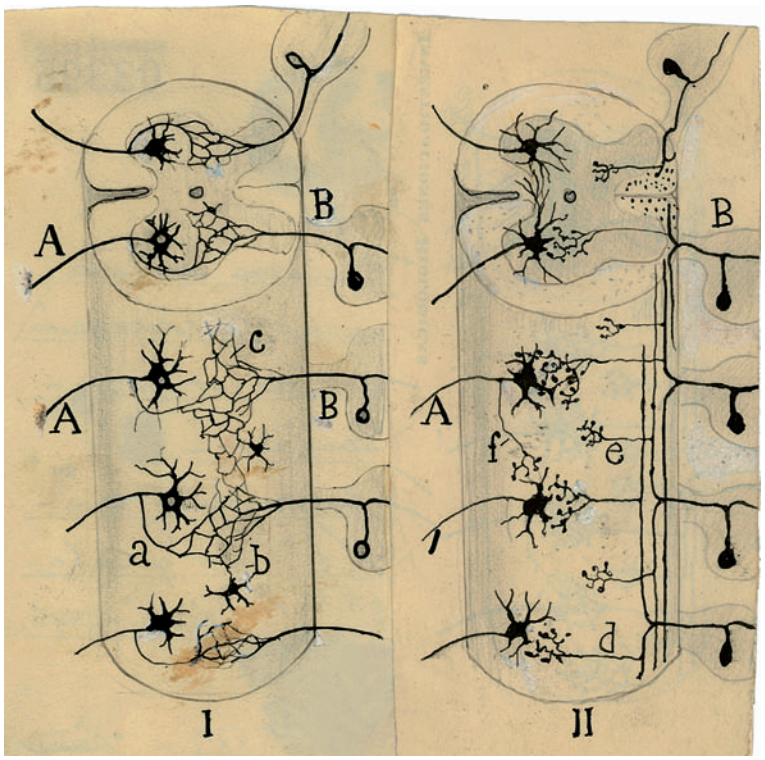

Figure 7 Cajal's drawing to explain the differences between the 'neuron theory' and the 'reticular theory.' The figure caption provided by Cajal stated: "Scheme to compare the concept of Golgi regarding the sensory–motor connections of the spinal cord (I) with the result of my investigations (II). *A*, anterior roots; *B*, posterior roots; *a*, collateral of a motor root; *b*, cells with short axon which, according to Golgi, would intervene in the formation of the network; *c*, diffuse interstitial network; *d*, our long collaterals in contact with the motor cells; *e*, short collaterals." This figure was reproduced as figure 9 in *Recuerdos de mi Vida*. Copyright by Herederos de Santiago Ramón y Cajal.

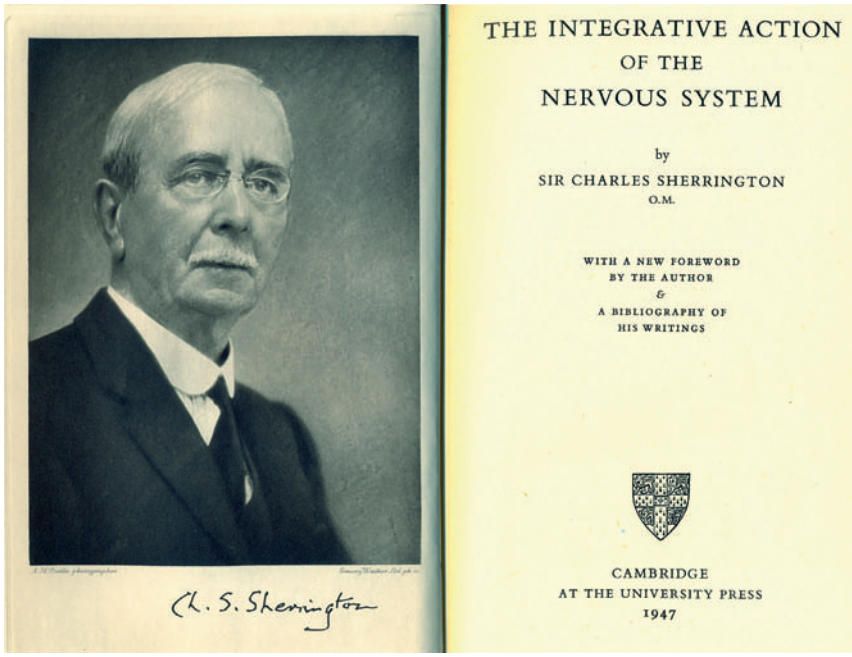

Figure 8 Portrait of Sherrington included in the 1947 edition of his book, *The Integrative Action of the Nervous System*.

widely accepted. The many, fundamental contributions of Cajal to the neuron doctrine were summarized by himself in several articles and books, and especially in the article 'Neuronismo o reticularismo,' published in 1933.

Nevertheless, in spite of the wealth of experimental data (physiological and Wallerian secondary degeneration studies) and histological observations supporting the neuron doctrine, the crucial observation that there was a physical separation between the neurons took a long time to emerge, since the limited resolving power of light microscopy impaired making this distinction. Thus, the neuron doctrine was generally accepted but with certain doubts and exceptions up to the middle of the twentieth century. This situation was beautifully summarized by Charles Sherrington (1857–1952) in his classic book, *The Integrative Action of the Nervous System* (see Figure 8). Sherrington, who in 1897 introduced the term 'synapse' for the hypothetical one-way contact between axon terminals and somata or dendrites, wrote the following with regard to this subject in 1947:

> As to the existence or non-existence of a surface of separation or membrane between neurone and neurone, that is a structural question on which histology might be competent to give valuable information. In certain cases, especially in Invertebrata, observation (Apáthy, Bethe, etc.) indicates that many nerve-cells are actually continuous one with another. It is noteworthy that in several of these cases the irreversibility of direction of conduction which is characteristic of spinal reflex-arcs is not demonstrable [...]. But in the neurone-chains on the gray-centred system of vertebrates, histology on the whole furnishes evidence that a surface of separation does exist between neurone and neurone. [...] It seems therefore likely that the nexus between neurone and neurone in the reflex-arc, at least in the spinal arc of the vertebrate, involves a surface of separation between neurone and neurone; and this as a transverse membrane across the conductor must be an important element in intercellular conduction. [...] In view, therefore, of the probable importance physiologically of this mode of nexus between neurone and neurone it is convenient to have a term for it. The term introduced has been *synapse*. (Foster and Sherrington, *The Integrative Action of the Nervous System*, 1897)

To resolve this structural question posed by Sherrington, the advent of electron microscopy in the 1950s was necessary to confirm that there is a separation between the pre- and postsynaptic elements of approximately 10–20 nm, the so-called synaptic cleft.

Law of Dynamic Polarization of Nerve Cells

Obviously, it is not the same to think that nerve currents flow through a continuous rather than a discontinuous network of neuronal processes. Thus, the new ideas about connections between neurons led to novel theories on the relationship between neuronal circuits and brain function. Cajal wrote in the

Textura: "The cell bodies [dendrites and axons], terminate freely but nevertheless, the flow of currents is not impeded in such an infinitely interrupted, fragmented nervous system. How can such currents flow? There can be only one answer, by contact, in much the same way that electric current crosses a splice between two wires."

An important consequence of the neuron doctrine was Cajal's theory of the law of dynamic polarization of nerve cells. At that time, the direction of impulse activity within the neurons remained a puzzling question to researchers, as nicely expounded by Cajal in 1917 in his autobiography *Recuerdos de mi Vida*: "In what direction does the nervous impulse travel within the neuron? Does it spread in all directions, like sound or light, or does it pass constantly in one direction like water in the watermill?"

It was commonly thought that dendrites played a nourishing role, whereas axons transmitted nervous impulses in a cellulifugal direction (a generalization based mostly on the cellulifugal conduction in the axons of the spinal motor neurons). However, there was no general consensus or a clear idea about the role of dendrites in the processing of information. In 1889, Cajal thought that at least in certain cases it was clear that the dendrites play a role in the reception of currents, and 2 years later, he tried to generalize this idea with the Law of Dynamic Polarization. This law was based on the direction followed by the impulses in regions of the nervous system where, through their activity, it was apparently clear what anatomical routes the impulses followed, such as in the visual and olfactory systems (**Figure 9**): "If in such inquiry, the [dendritic] arborization is always shown as a *receptor* apparatus and the [axonal arborization] as an apparatus for the *application* of the [impulses], by analogy we would have attained a rule to judge the direction of the currents in the [nerve cells within the central nervous system]."

Cajal proposed that neurons could be divided into three functionally distinct regions: a receptor apparatus (formed by the dendrites and soma), an emission apparatus (the axon), and a distribution apparatus (terminal axonal arborization). Later he realized that the soma does not always intervene in the conduction of the impulses and that, sometimes, impulse activity goes directly from the dendrites to the axon (**Figure 10**). Thus, the Law of Dynamic Polarization was changed

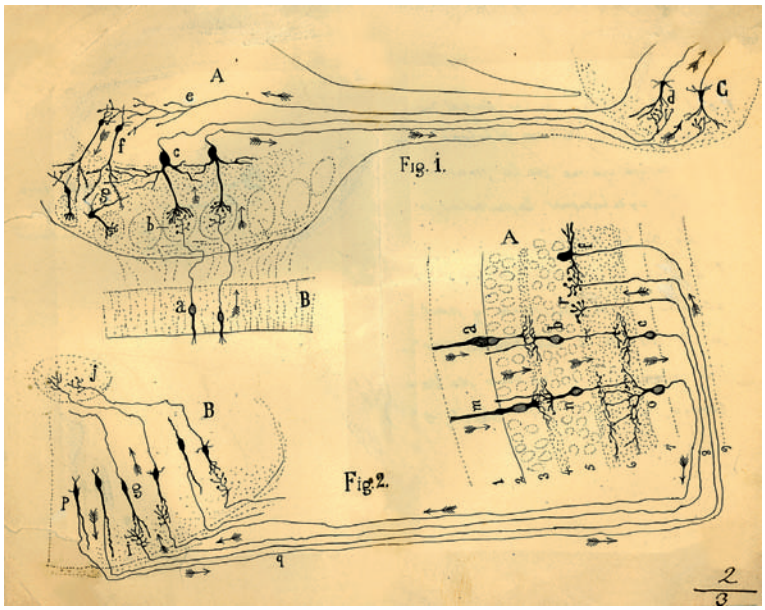

Figure 9 Cajal's scheme showing the current flow in the visual and olfactory systems. The figure caption provided by Cajal stated: "*Fig. 1*. Scheme of cellular connections of the olfactory mucosa, olfactory bulb, tractus, and olfactory lobe of the cerebrum. The arrows indicate the direction of the currents. *A*, olfactory bulb; *B*, mucosa; *C*, olfactory lobe. *a, b, c, d*. One-way or centripetal pathway through which sensory or olfactory excitation passes. *e, f, g*, Centrifugal pathway through which the [nervous] centers can act on the elements of the bulb, granules and nerve cells, whose protoplasmic processes penetrate in the glomeruli. *Fig. 2*. Scheme of the pathway of visual excitation through the retina, optic nerve and optic lobe of the birds. *A*, retina; *B*, optic lobe. *a, b, c*, represent a cone, a bipolar cell and a ganglion cell of the retina, respectively, the order through which visual excitation travels. *m, n, o,* parallel current emanating from the rod also involves bipolar and ganglion cells. *g*, cells of the optic lobe that receive the visual excitation and transfer it to *j*, the central ganglion. *p, q, r,* centrifugal currents that start in certain fusiform cells of the optic lobe and terminate in *r*, in the retina at the level of the spongioblasts; *f*, a spongioblast." From Cajal SR (1891) Significación fisiológica de las expansiones protoplásmicas y nerviosas de las células de la substancia gris. *Revista de Ciencias Médicas, Barcelona* 22: 673–679; 715–723. Copyright by Herederos de Santiago Ramón y Cajal.

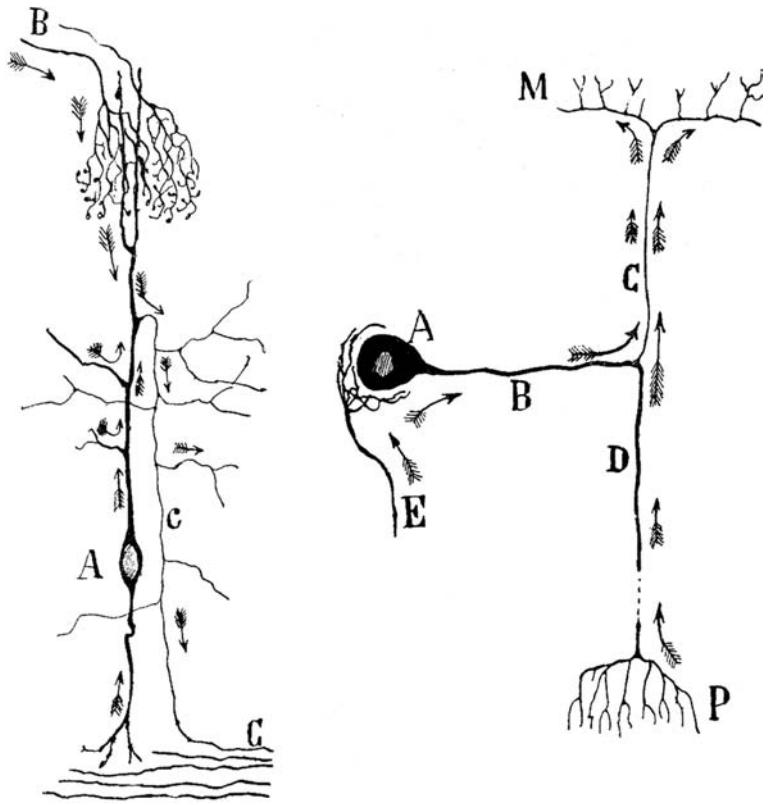

Figure 10 Cajal's captions for this drawing: (Left) "Crosier cell of the optic lobe of the sparrow. *A*, soma; *B*, fibers arriving from the retina; *c*, central white matter; *C*, axon. Arrows indicate the direction of the current [flow]." (Right) "Scheme showing the current flow in a mammalian [spinal] ganglion sensory cell. *A*, soma; *B*, shaft, *D*, axipetal or peripheral process which supplies the current [arising in the periphery]; *C*, axon who carries the [impulses] towards the spinal cord; *E*, fiber that gives rise to the peri-cellular arborisation; *M*, spinal cord." From Cajal SR (1897) Leyes de la morfología y dinamismo de las células nerviosas. *Revista Trimestral Micrográfica* 2: 1–28. Copyright by Herederos de Santiago Ramón y Cajal.

to the Theory of Axipetal Polarization: "The soma and dendrites display axipetal conduction, whereby they transmit the nervous waves towards the axon. Conversely, the axon or cylinder-axis has somatofugal or dendrifugal conduction, propagating the impulses received by the soma or dendrites towards the terminal axonal arborizations. [...]. This formula can be applied universally without exception, both in the vertebrates and invertebrates."

These studies of Cajal had a deep impact on researchers at that time, and his observations and theories were essentially confirmed. However, the introduction and development of electron microscopy, together with the appearance of new techniques to study the nervous system at the anatomical, physiological, and molecular level, have shown over the years that there are many exceptions that challenge the neuron doctrine and the Law of Dynamic (or Axipetal) Polarization. For example, axons have been found to form synapses with other axons (axo-axonal synapses), and presynaptic elements are not necessarily axonal but they can be dendrites or somata. Thus, a variety of synaptic relationships have been observed, including dendro-dendritic, somato-somatic, somato-dendritic, dendro-somatic, dendro-axonic, and somato-axonic synapses. Furthermore, neurons are not only connected by point-to-point chemical synapses, but they may also be coupled electrically, while the direction of transmission may be bidirectional through gap junctions. The plasma membranes of adjacent neurons at the gap junctions are separated by a gap of about 2 nm, but they contain small channels that connect the cytoplasm of the adjoining neurons. These channels permit the diffusion of small molecules and the flow of electric current. Moreover, the transmitter released at synaptic or nonsynaptic sites may diffuse a certain distance and act on other synaptic contacts, or on extrasynaptic receptors. Finally, neuroglial cells are not only key components of the nervous system for their numerous structural, metabolic, and protective functions, but it has been shown that astrocytes are involved in information processing through bidirectional signaling between these cells and neurons. These different features represent further complexities in the system, making it even more difficult to establish the information flow and wiring of neural circuits. Nevertheless,

axo-dendritic synapses are by far the most common type of synaptic contact (followed by axo-somatic synapses). Other types of synaptic contacts are not found in all regions of the nervous system, and when they are present, they are usually established only between certain types of neurons. Therefore, we agree with the conclusion of Gordon Shepherd that the idea of the nerve cell as an independent cellular unit, with an overall polarization to mediate its input–output functions, continues to be one of the foundations on which our concepts of nervous function are based.

Coming back to Cajal, if some of the recent discoveries mentioned herein had been made in his time, it would probably not have been a surprise to him. In fact, he had indeed described the existence of anastomosis in certain instances, which he sometimes called interprotoplasmic or internervous bridges. However, Cajal was convinced that the vast majority of connections between neurons involved "contact or contiguity." In his article 'Neuronismo o reticularismo,' once again Cajal was ahead of his time when he drew the following conclusion:

> We are neither exclusive nor dogmatic. We are proud of retaining a mental flexibility that is not embarrassed by the need to rectify. Thus exceptions may arise to the neuronal discontinuity that is so evident in innumerable examples. We ourselves have reported [some cases of anastomosis] [...]. For this reason, we have pointed out, and we insist in this, when dealing with the morphology and connections of neurons, we must abide by the law of the greatest numbers, that is to rigorously statistical criteria. (Cajal, *Archivos de Neurobiología* 13, 1933)

See also: Cellular Dynamics Revealed by Digital Holographic Microscopy; Cognitive Neuroscience: An Overview; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Fluorescent Biomarkers in Neurons; History of Neuroscience: Early Neuroscience; Neuron Doctrine: Historical Background; Neurophysiology: Past and Present.

Further Reading

Araque A, Carmignoto G, and Haydon PG (2001) Dynamic signaling between astrocytes and neurons. *Annual Review of Physiology* 63: 795–813.

Azmitia E, DeFelipe J, Jones EG, et al. (eds.) (2002) *Changing Views of Cajal's Neuron. Progress in Brain Research,* vol. 136. Amsterdam: Elsevier.

Bennett MVL (1977) Electrical transmission: A functional analysis and comparison with chemical transmission. In: Kandel (ed.) *Handbook of Physiology, Section 1, The Nervous System,* vol. 1, pp. 357–416. Baltimore: Williams and Wilkins.

Cajal SR (1899, 1904) *Textura del Sistema Nervioso del Hombre y de los Vertebrados*. Madrid: Moya. This book was revised and extended in the French version *Histologie du Système Nerveux de l'Homme et des Vertébrés* (translated by Azoulay L.), Paris: Maloine (1909, 1911). English translations: *Histology of the Nervous System of Man and Vertebrates* (translated by Swanson N and Swanson LW), New York: Oxford University Press (1995); *Texture of the Nervous System of Man and the Vertebrates* (an annotated and edited translation of the original Spanish text with the additions of the French version by Pasik P and Pasik T), New York: Springer (2000/2001).

Cajal SR (1913–1914) *Estudios Sobre la Degeneración y Regeneración del Sistema Nervioso*. Madrid: Moya. English translations: *Degeneration and Regeneration of the Nervous System* (translated by May RM), London, Oxford University Press (1928); reprinted and edited with additional translations by DeFelipe J and Jones EG: *Cajal's Degeneration and Regeneration of the Nervous System*. New York: Oxford University Press (1991).

Cajal SR (1917) *Recuerdos de Mi Vida, Volume 2. Historia de Mi Labor Científica*. Madrid: Nicolás Moya.

Cajal SR (1933) ¿Neuronismo o reticularismo? Las pruebas objetivas de la unidad anatómica de las células nerviosas. *Archivos de Neurobiología* 13: 217–291; 579–646. English translation: Neuron theory or reticular theory? Objective evidence of the anatomical unity of nerve cells (translated by Ubeda-Purkiss M. and Fox CA). Madrid: Consejo Superior de Investigaciones Científicas (1954).

DeFelipe J and Jones EG (1988) *Cajal on the Cerebral Cortex*. New York: Oxford University Press.

DeFelipe J and Jones EG (1992) Santiago Ramón y Cajal and methods in neurohistology. *Trends in Neuroscience* 15: 237–246.

DeFelipe J (2002) Sesquicentennial of the birthday of Santiago Ramón y Cajal (1852–2002), the father of modern neuroscience. *Trends in Neuroscience* 25: 481–484.

DeFelipe J (2006) Brain plasticity and mental processes: Cajal again. *Nature Reviews in Neuroscience* 7: 811–817.

Jones EG (1994) The neuron doctrine. *Journal of History of Neuroscience* 3: 3–20.

Peters A, Palay SL, and Webster H de F (1991) *The Fine Structure of the Nervous System. Neurons and Their Supporting Cells*. New York: Oxford University Press.

Shepherd GM (1991) *Foundations of the Neuron Doctrine*. New York: Oxford University Press.

Zoli M, Jansson A, and Sykova E (1999) Volume transmission in the CNS and its relevance for neuropsychopharmacology. *Trends in Pharmacological Sciences* 20: 142–150.

Relevant Website

http://www.ibro.org – International Brain Research Organization: Neuroscience History on the Web.

Calcitonin Gene-Related Peptide (CGRP) and Receptors

W Ma, J-G Chabot, and R Quirion, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Calcitonin gene-related peptide (CGRP) was discovered in the early 1980s as a gene product of the alternative splicing of the calcitonin gene. It is an important member of the CGRP family, which also includes amylin, adrenomedullin (AM), intermedin, and the calcitonin receptor-stimulating peptide (Figure 1). The broad distribution of CGRP in both central and peripheral nervous systems (CNS and PNS) suggests its extensive biological effects, including the modulation of the senses and autonomic functions, vasodilation, cardiovascular protection, muscular motility, learning and memory, pain and inflammation, as well as the regulation of water and food intake. CGRP exerts its biological effects by acting on specific CGRP receptors. The CGRP1 receptor belongs to the type II or B family of the G-protein-coupled receptor (GPCR) superfamily. The CGRP receptor is rather unique compared to other GPCRs. In addition to sharing a seven-transmembrane domain protein known as calcitonin receptor-like receptor (CLR) with AM receptors (AM1 and AM2), CGRP and AM receptors require other accessory proteins to be functional. These accessory proteins include the receptor-activity-modifying proteins (RAMPs), which determine receptor specificity, and the receptor complement protein (RCP), which facilitates signal transduction coupling. This article reviews the anatomical distribution of CGRP and its receptors, as well as some of its biological functions.

Discovery of CGRP

CGRP is composed of 37 amino acids and was discovered by a molecular biological approach in 1982 by Amara and coworkers. Alternative RNA splicing is an important means to increase proteomic complexity in eukaryotic organisms. Through this process, the majority of human premessenger RNAs (mRNAs) generate more than one mRNA, which often leads to the production of multiple polypeptides from a single gene. Tissue-specific alternative RNA processing plays an important role in regulating gene expression. The human calcitonin/CGRP gene is a good example of such regulation. The mRNA encoding CGRP was identified as a result of alternative splicing of RNA transcripts from the calcitonin gene. This gene is predominantly expressed in two tissues, thyroid C cells and a subset of primary sensory neurons.

The isolation and sequence analysis of the calcitonin genomic DNA and calcitonin/CGRP complementary DNA (cDNA) demonstrated that both CGRP and calcitonin mRNAs are generated by differential processing from a single genomic locus. The calcitonin/CGRP gene consists of six exons (Figure 2). The calcitonin/CGRP pre-mRNA is alternatively spliced in a tissue-specific manner, leading to the production of calcitonin mRNA containing exons 1–4 in thyroid C cells and CGRP mRNA containing exons 1–3, 5, and 6 in neurons.

The possible existence of other gene products related to CGRP was investigated by screening the library of chimeric plasmids containing inserts complementary to mRNAs from rat medullary thyroid carcinomas with a clone of CGRPα cDNA probe. The analysis resulted in the identification of a novel mRNA, called CGRPβ, which is closely related to CGRPα mRNA. The CGRPβ gene does not contain sequences corresponding to the calcitonin-coding sequences of the calcitonin/CGRPα gene. Thus, only CGRPβ can be derived from this second gene. Human CGRPβ gene was identified in 1985 by Steenbergh and collaborators and differs from human CGRPα by three amino acids. CGRP is a highly conserved peptide between mammalian species, with 26 homologous amino acids. Figure 3 depicts the primary structures of CGRP and related peptides.

In general, the transcripts for CGRPα and CGRPβ have a similar distribution in CNS and PNS. However, some evidence suggests that they can be differentially distributed and regulated in certain tissues, such as primary sensory neurons and intrinsic enteric neurons, suggesting selective regulation of these CGRP genes. The expression of the calcitonin/CGRP gene is regulated by glucocorticoids and cyclic adenosine monophosphate. Alternative RNA splicing is subject to complex control, often involving multiple sequence elements and protein factors. Studies to date have identified a large number of *cis*-acting regulatory elements located in the exons and introns as well as *trans*-acting factors to regulate alternative RNA splicing. Studies on alternative splicing factors have indicated that they are often RNA binding proteins capable of binding to specific *cis*-acting RNA elements, thus affecting RNA processing choice through interactions with the basal splicing or polyadenylation machinery. For example, it has been shown that

TIAR, a recently characterized splicing regulator which is related to TLA, plays a role in regulating the calcitonin/CGRP gene.

Neuroanatomical Localization

CGRP is broadly distributed in the nervous system of both vertebrates and invertebrates. The distribution of CGRP mRNA and CGRP-like immunoreactivity (CGRP-IR) has been reported in the CNS and PNS of various species, including humans. In Table 1, the distribution of CGRP mRNA, CGRP-IR, and CGRP receptor binding sites in the rat CNS and PNS is illustrated, Figures 4 and 5 depict CGRP-IR projections and CGRP receptor binding sites, respectively, in the rat brain.

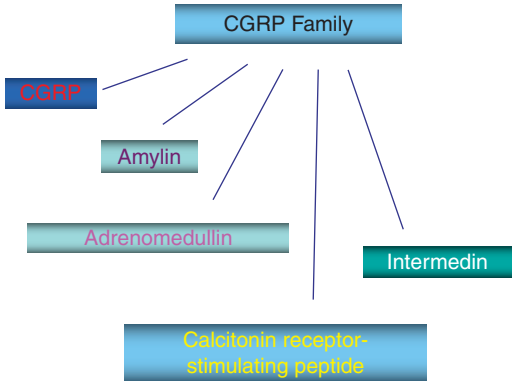

Figure 1 Diagram depicting members of the calcitonin and calcitonin gene-related peptide (CGRP) family: CGRP, amylin, adrenomedullin, intermedin, and calcitonin receptor-stimulating peptide.

CGRP mRNA-Containing Neurons

Using *in situ* hybridization histochemistry, the distribution of CGRP mRNA was examined in the rat CNS (Table 1). Generally, the highest densities of CGRP mRNA-expressing neurons are mainly seen in the lateral hypothalamic area, various thalamic nuclei, brain stem cranial motor neurons, lateral superior olivary nucleus, parabrachial nuclei, and the majority of α-motor neurons of the spinal cord, with about 50% of neurons in the cranial trigeminal and dorsal root ganglia expressing CGRP.

Moderate levels of CGRP mRNA-positive neurons are observed in the medial division of the bed nucleus of the stria terminalis, the nucleus of the lateral olfactory tract, various hypothalamic and thalamic nuclei, the nucleus of the lateral lemniscus, the reticular paragigantocellular nucleus, the cuneate nucleus, and various hypoglossal nuclei.

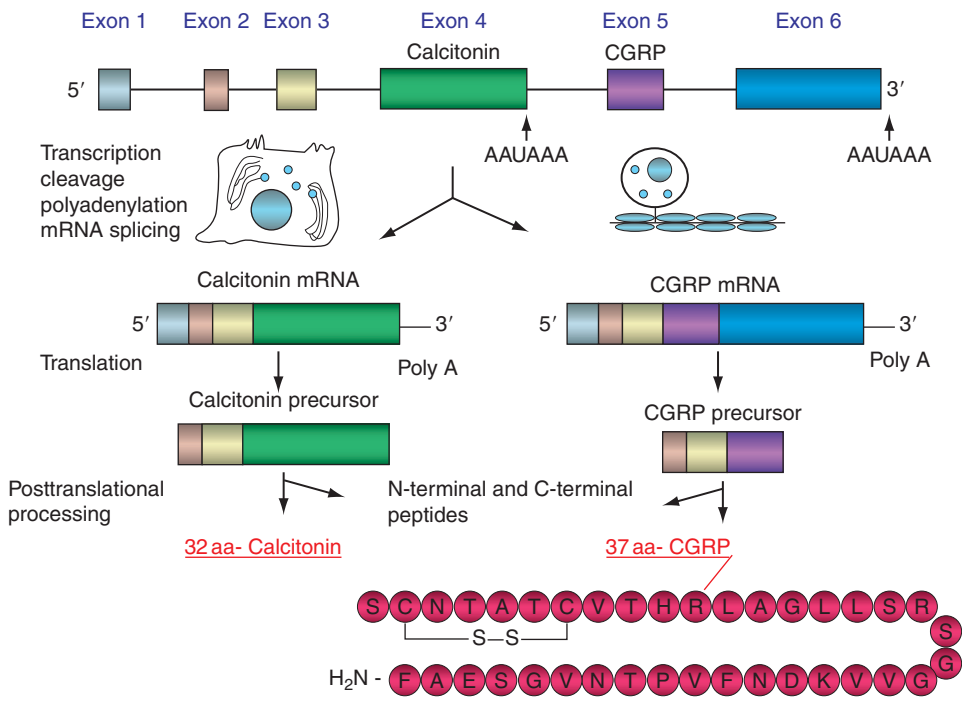

Figure 2 Model of tissue-specific neuropeptide production from the calcitonin gene. Alternative RNA processing pathways generate calcitonin or calcitonin gene-related peptide (CGRP) in a tissue-specific manner. mRNA, messenger RNA; aa, amino acid; Poly A, polyadenosinc. Adapted from Jacques D, Dumont Y, van Rossum D, et al. (2000) Calcitonin gene-related peptide (CGRP), amylin and adrenomedullin: Anatomical localization and biological functions in the mammalian and human brains. In: Quirion R, Bjorklund A, and Hökfelt T (eds.) *Peptide Receptors, Part I*, pp. 301–374. Amsterdam: Elsevier.

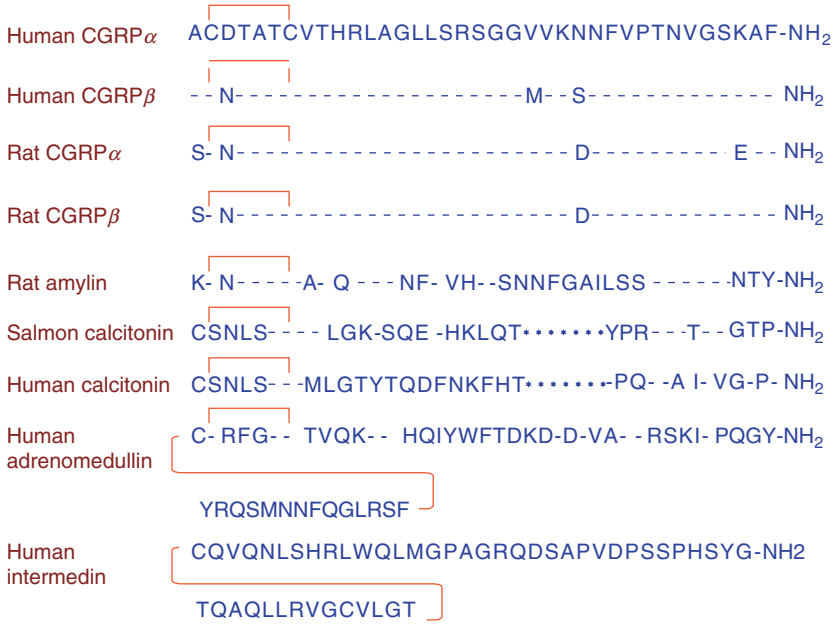

Figure 3 Amino acid sequence of CGRP and related peptides. Amino acids in common with human CGRPα are indicated by -. Spaces inserted to allow for sequence comparison are indicated by *. Adapted from Jacques D, Dumont Y, van Rossum D, et al. (2000) Calcitonin gene-related peptide (CGRP), amylin and adrenomedullin: Anatomical localization and biological functions in the mammalian and human brains. In: Quirion R, Bjorklund A, and Hökfelt T (eds.) *Peptide Receptors, Part I*, pp. 301–374. Amsterdam: Elsevier.

Low levels of CGRP mRNA-expressing neurons are seen in the lateral division of the bed nucleus of the stria terminalis, the dorsal hippocampus, various hypothalamic and thalamic nuclei, the mesencephalic cuneiform nucleus, and the pontine and medullary reticular nuclei.

CGRP-IR Neurons and Axons

Through immunohistochemistry, the distribution of CGRP-IR in the rat CNS has been described in detail. In accordance with the distribution of CGRP mRNA, CGRP-IR neurons and axons are widely but unevenly distributed in the rat brain. As shown in Table 1, high amounts of CGRP-IR neurons are seen in various parabrachial nuclei and the Kölliker-Fuse nucleus. Moderate levels of CGRP-IR neurons are observed in various amygdaloid, hypothalamic, and thalamic nuclei, cranial and spinal motor neurons, and the lateral superior olive. Low levels of CGRP-IR neurons are observed in the primary olfactory cortex, bed nucleus of the stria terminalis, various hypothalamic and thalamic nuclei, arcuate nucleus, central periaqueductal gray, cuneiform nucleus, tegmental nuclei, inferior colliculus, lemniscus nuclei, pontine nuclei, central gray of the pons, cuneate nucleus, nucleus of Roller, and raphe obscurus nucleus.

High levels of CGRP-IR axons are seen in various hypothalamic nuclei, including the medial preoptic area, periventricular area, dorsomedial nucleus, medial eminence, and medial forebrain bundle area, as well as the superficial layers of the trigeminal dorsal horn. Moderate levels of CGRP-IR axons are found in various thalamic nuclei, septal nuclei, bed nucleus of the stria terminalis, central amygdaloid nucleus, striatum, periaqueductal central gray, medial geniculate body, parabigeminal nucleus, lateral lemniscus, interpeduncular nucleus, parabrachial nucleus, superior olive, cochleal nucleus, nucleus of the solitary tract, and vestibular nuclei. Low levels of CGRP-IR axons were noted in the olfactory bulb, medial prefrontal cortex, most neocortical areas, and anterior pituitary.

High densities of CGRP-IR axons are also found in laminae I and II of the spinal dorsal horn. CGRP-containing axons constitute about 30% of primary afferent axons of the Lissauer's tract, the major input to superficial laminae of the dorsal horn. More than 80% of spinal afferent neurons supplying visceral structures including urogenital and upper gastrointestinal tracts as well as blood vessels are CGRP expressing. Low densities of CGRP-IR axons are seen in deeper layers of the dorsal horn as well as in the ventral horn and areas surrounding the central canal.

CGRP Receptor Binding Sites

CGRP receptor binding sites are discretely distributed in the CNS of various species, including rat and human. As shown in Table 1 and Figure 5, high densities of CGRP binding sites are seen in the olfactory

Table 1 The distribution of CGRP mRNA, CGRP-IR, and CGRP receptor-binding sites in rat CNS and PNS

Region	*CGRP mRNA*	*CGRP-IR*	*CGRP receptor binding sites*
Telencephalon			
Medial bed nucleus of the stria terminalis	+++	+++ (axons)	
Preoptic subnucleus	+++		
Nucleus of lateral olfactory tract	+++		
Lateral bed nucleus of the stria terminalis	+	+ (cells), +++ (axons)	
Dorsal hippocampus	+	+++ (cells)	
Medial amygdaloid nucleus		+++ (cells)	+→+++
Central amygdaloid nucleus		+++ (axons)	+++++
Amygdalhippocampal area		+++ (cells)	
Postmedial cortical amygdaloid nucleus		+++ (cells)	
CA2 field of Ammon's horn			+
CA1 and CA4 field of Ammon's horn		+ (cells)	+
Primary olfactory cortex		+ (cells)	
Medial corticohypothalamic tract		+ (cells)	
Medial forebrain bundle area		+++++ (axons)	
Septal area		+++ (axons)	+→+++
Olfactory bulb		+ (axons)	+++
Medial prefrontal cortex		+ (axons)	
Most neocortical areas		+ (axons)	
Medial and lateral olfactory nuclei			+++
Cingulate cortex			+++
Anterior bed nucleus of the stria terminalis			+++++
Tenia tecta			+++
Islands of Calleja			+++
Olfactory tubercle			+++
Anterior nucleus accumbens			+++
Vertical limb of the diagonal band			+++
Fundus striatum			+++
Ventral pallidum			+++
Basolateral amygdaloid nucleus			+++++
Lateral amygdaloid nucleus			+++++
Cingulate, piriform, temporal and entorhinal cortices			+→+++++
Subfornical organ			+++
Subiculum			+
Diencephalon			
Lateral hypothalamic area	+++++	+++ (cells)	
Posterior thalamic nucleus	+++++	+ (cells), +++ (axons)	
Peripeduncular nucleus	+++++	+++ (cells)	
Subparafascicular nucleus	+++++		
Posterior intralaminar thalamic nucleus	+++++	+ (cells), +++ (axons)	
Caudal lateral hypothalamic area	+++		
Arcuate hypothalamic nucleus	+++	+ (cells)	+++
Parvocellular part of the thalamic ventroposterior nucleus	+++	+++ (axons)	
Parafascicular thalamic nucleus	+++	+++ (axons)	
Medial preoptic nucleus	+	+ (cells), +++++ (axons)	+++++
Zona Incerta	+	+ (cells)	
Ventromedial thalamic nucleus	+	+ (cells), +++ (axons)	
Dorsal hypothalamic area	+		+++
Medial paracentral thalamic nucleus	+	+++ (axons)	+++
Periventricular hypothalamic nucleus		+++ (cells)	+++
Dorsomedial hypothalamic nucleus		+++ (cells)	+++
Perifornical nucleus		+++ (cells)	
Premammillary nucleus		+++ (cells)	
Lateral subparafascicular thalamic nucleus		+++ (cells), +++ (axons)	
Anterior hypothalamic nucleus		+ (cells)	+
Parvicellular part of the paraventricular hypothalamic nucleus		+ (cells)	

Continued

Table 1 Continued

Region	*CGRP mRNA*	*CGRP-IR*	*CGRP receptor binding sites*
Dorsal thalamic area		+ (cells), +++ (axons)	
Paraventricular area		+++++ (axons)	+
Dorsomedial nucleus		+++++ (axons)	
Median eminence		+++++ (axons)	
Dorsal part of interpeduncular nucleus		+++ (axons)	
Ventral striatum		+++ (axons)	
Preoptic suprachiasmatic nucleus			+
Supramammillary and lateral mammillary nuclei			+++
Medial mammillary nucleus			+++++
Medial habenular nucleus			+++
Metencephalon			
Parabigeminal nucleus	+++++	+++ (axons)	
Oculomotor nucleus	+++++	+++ (cells)	+
Trochlear nucleus	+++++	+++ (cells)	+
Dorsal and ventral nucleus of the lateral lemniscus	+++	+ (cells), +++ (axons)	
Rostral and medial reticular paragigantocellular nucleus	+++		
Mesencephalic cuneiform nucleus	+		
Dorsal marginal region of pontine nuclei	+	+ (cells)	
Pontine central gray	+	+ (cells)	
Medullary reticular nucleus	+		
Central periaqueductal gray		+ (cells), +++ (axons)	
Ventral and dorsal tegmental		+ (cells)	+++++
Central nucleus of inferior colliculus		+ (cells)	+++++
Medial geniculate nucleus		+++ (axons)	+→ +++
Peripeduncular nucleus			+++++
Substantia nigra			+→+++
Superficial colliculus			+++++
Darkschewitsch nucleus			+
Dorsal raphe			+
Interfascicular nuclei			+
Hindbrain			
Parabrachial nucleus	+++++	+++++ (cells), +++ (axons)	+→+++
Kölliker-Fuse nucleus	+++++	+++++ (cells)	
Facial nucleus	+++++	+++ (cells)	
Ambiguous nucleus	+++++	+++ (cells)	
Hypoglossal nucleus	+++++	+++ (cells)	
Perihypoglossal nucleus of Roller	+++++	+ (cells)	+++++
Trigeminal motor nucleus	+++++	+++ (cells)	+++++
Abduducens nucleus	+++++	+++ (cells)	
Lateral superior olivary nucleus	+++++	+++ (cells), +++ (axons)	
External cuneate nucleus	+++	+ (cells)	
Perihypoglossal nucleus	+++		+++++
Raphe obscurus nucleus		+ (cells)	+++++
Superficial layers of sensory trigeminal nucleus		+++++ (axons)	+++++
Cochlear nucleus		+++ (axons)	+→+++
Vestibular nucleus		+++ (axons)	+++++
Nucleus solitary tractus		+++ (axons)	+++++
Locus coeruleus			+++++
Central gray of the pons			+++++
Mesencephalic tract of trigeminal nerve			+→+++
Gracile nucleus			+++++
Inferior olive			+++++
Cuneate nucleus			+++++
Area postrema			+++++
Molecular layer of cerebellum			+++++
Purkinje cell layer of cerebellum			+→+++

Continued

Table 1 Continued

Region	*CGRP mRNA*	*CGRP-IR*	*CGRP receptor binding sites*
Cerebellar nuclei			+→+++
Spinal cord			
Motor neurons in laminae VII-IX	+++++		
Superficial laminae of dorsal horn (LI-LII)		+++++ (axons)	+++++
Medial portion of laminae III-IV			+++++
Lamina X			+++++
Lamina II and ventral horn			+
Primary sensory ganglion			
Trigeminal ganglion	+++++ (50% neurons	+++++ (30% neurons)	
Dorsal root ganglion (50% neurons)	+++++ (50% neurons)	+++++ (30% neurons)	

CA1, CA2, and CA4, area 1, 2, and 4 in the Cornus Ammon's; CGRP, calcitonin gene-related peptide; mRNA, messenger RNA; CGRP-IR, CGRP-like immunoreactivity; CNS, central nervous system; PNS, peripheral nervous system; +, low; +++, moderate; +++++, high; +→+++, low to moderate; +→+++++, low to high.

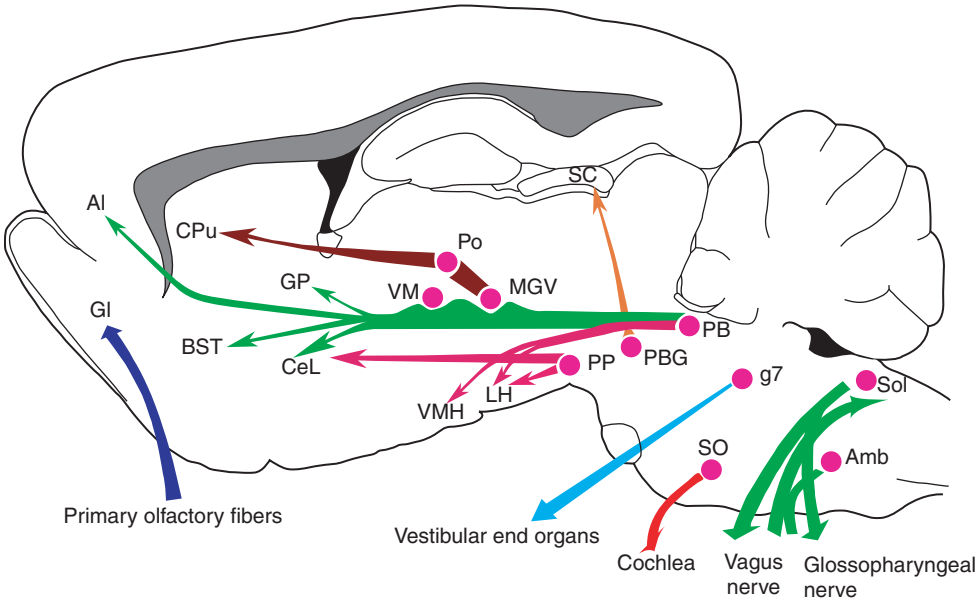

Figure 4 Summary of the main CGRP-containing pathways shown on a sagittal section diagram of the rat brain. AI, agranular insular cortex; Amb, ambiguus nucleus; BST, bed nucleus of the stria terminalis; CeL, central amygdaloid nucleus, lateral division; CPu, caudate–putamen (striatum); g7, genu of the facial nerve; GI, glomerular layer of the olfactory bulb; GP, globus pallidus; LH, lateral hypothalamic area; MGV, medial geniculate nucleus, ventral part; PB, parabrachial nucleus, PBG, parabigeminal nucleus; Po, posterior thalamic nuclear group; PP, peripeduncular nucleus; SC, superior colliculus; SO, superior olive nucleus, Sol, nucleus of the solitary tract; VM, ventromedial thalamic nucleus; VMH, ventromedial hypothalamic nucleus. Adapted from Jacques D, Dumont Y, van Rossum D, et al. (2000) Calcitonin gene-related peptide (CGRP), amylin and adrenomedullin: Anatomical localization and biological functions in the mammalian and human brains. In: Quirion R, Bjorklund A, and Hökfelt T (eds.) *Peptide Receptors, Part I*, pp. 301–374. Amsterdam: Elsevier Science.

bulb, olfactory nuclei, nucleus accumbens, amygdaloid nuclei, septum nuclei, cingulate, piriform, temporal, and entorhinal cortices, hippocampus, various regions of the hypothalamus, preoptic area, periventricular nuclei, various mammillary nuclei, various thalamic nuclei, habenular nuclei, peripeduncular nucleus, tegmental nuclei, inferior and superior colliculi, locus coeruleus, vestibular nucleus, cochlear nucleus, mesencephalic trigeminal nucleus, molecular layer of the cerebellum, hypoglossus nucleus, vestibular nuclei, solitary nucleus, vagus nucleus, gracile nucleus, inferior olive, medullar reticular nuclei, Roller nucleus, raphe magnus nucleus, raphe obscurus nucleus, spinal trigeminal nucleus, area postrema, and cuneate nucleus. In the spinal cord, high densities of CGRP binding sites are found at all segmental levels in lamina X and in intermediolateral and intermediomedial nuclei of the thoracic segments, lamina I, as well as the medial portion of laminae III–V.

Moderate levels of CGRP binding sites are seen in the substantia nigra and medial geniculate nucleus. Low to moderate levels of CGRP binding sites were

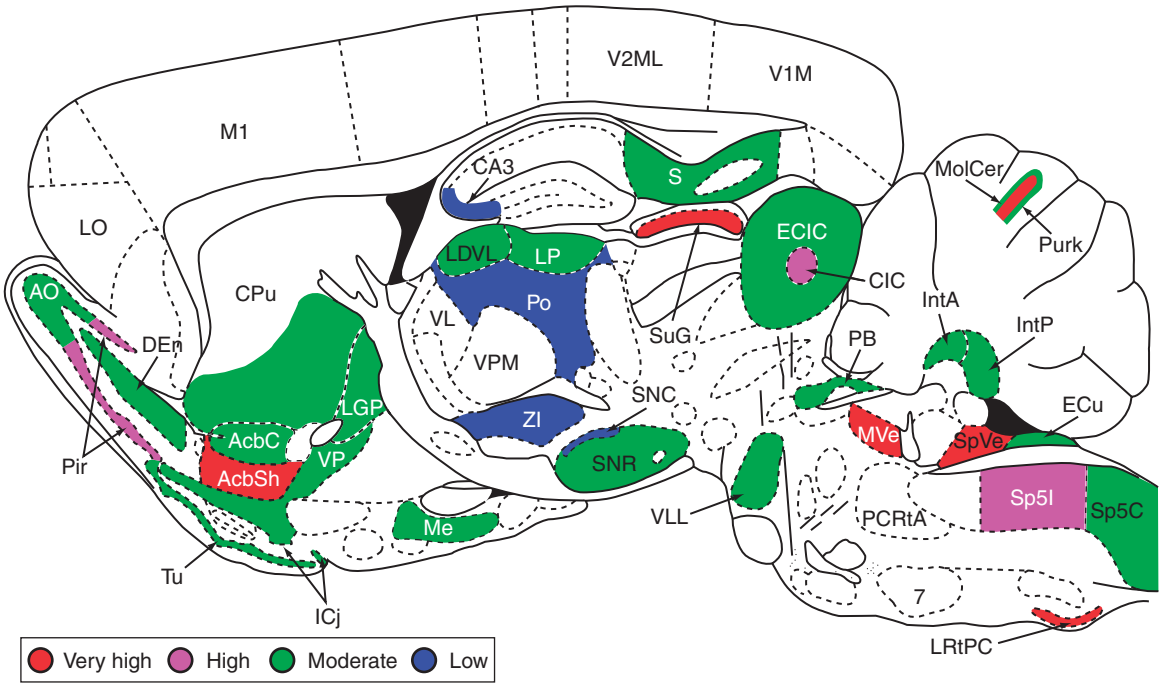

Figure 5 Schematic representation of a sagittal section of rat brain according to the atlas of Paxinos and Watson (1998). The distribution and semiquantitative rating (low, moderate, high, and very high) of calcitonin gene-related peptide (CGRP) binding sites ([^{125}I]hCGRP) are indicated with different colors on the figure. 7, facial nucleus; AcbC, accumbens nucleus, core; AcbSh, accumbens nucleus, shell; AO, anterior olfactory nucleus; CA3, field CA3 of hippocampus; CPu, caudate–putamen (striatum); CIC, central nucleus of the inferior colliculus; DEn, dorsal endopiriform nucleus; ECIC, external cortex of the inferior colliculus; ECu, external cuneate nucleus; ICj, islands of Calleja; IntA, interposed cerebellar nucleus, anterior part; IntP, interposed cerebellar nucleus, posterior part; LDVL, laterodorsal thalamic nucleus, ventrolateral part; LGP, lateral globus pallidus; LO, lateral orbital cortex; LP, lateral posterior thalamic nucleus; LRtPC, lateral reticular nucleus, parvicellular part; M1, primary motor cortex; Me, medial amygdaloid nucleus; MolCer, molecular layer of the cerebellum; Mve, medial vestibular nucleus; PB, parabrachial nucleus; PCRtA, parvicellular reticular nucleus, alpha part; Pir, piriform cortex; Po, posterior thalamic nuclear group; Purk, Purkinje cell layer of the cerebellum S, subiculum; SNC, substantia nigra, compact part; SNR, substantia nigra, reticular part; Sp5C, spinal trigeminal nucleus, caudal part; Sp5I, spinal trigeminal nucleus, interpolar part; SpVe, spinal vestibular nucleus; SuG, superficial gray layer of the superior colliculus; Tu, olfactory tubercle; V1M, primary visual cortex, monocular area; V2ML, secondary visual cortex, mediolateral area; VL, ventrolateral thalamic nucleus; VLL, ventral nucleus of the lateral lemniscus; VP, ventral pallidum; VPM, ventral posteromedial thalamic nucleus; ZI, zona incerta.

observed in the Darkschewitsch nucleus, oculomotor and trochlear nuclei, dorsal raphe nucleus, interfascicular nucleus, parabrachial nucleus, Purkinje cell layer of cerebellum, and deeper cerebellar nuclei. Low levels of binding sites were seen in lamina II and the ventral horn.

Recent Progress in CGRP Receptor Research

Early on, based on the differential biological activities of various CGRP analogs, CGRP receptors have been classified into the CGRP1 and CGRP2 subtypes. Pharmacologically, these two subtypes have been described based on properties of the antagonist CGRP8–37 and the agonist (Cys(ACM^{2,7}))CGRP. Structure-activity studies of CGRP analogs showed that the C- and N-terminal regions of the peptide interact independently with their receptors. While C-terminal peptides such as CGRP8–37 behave as preferential CGRP1 receptor antagonists, N-terminal peptides such as CGRP1–12 act as weak agonists. Structural modifications of CGRP28–37 have yielded micromolar to nanomolar affinity ligands. The CGRP1 subtype has been cloned, and it is a complex macromolecular structure (see below), but the existence of molecularly distinct CGRP2 receptors is yet to be confirmed.

A GPCR belonging to the secretin receptor family (B family of GPCR) has been described and is known as CLR. As shown in Figure 6, CLR is not a functional receptor by itself and requires accessory proteins to function. These proteins are known as RCP which couple CLR to signaling pathways and RAMP, a single membrane-spanning protein, which acts as a chaperone protein and confers pharmacological specificity to CLR. This complex behaves as a CGRP1 receptor if CLR is associated with RAMP1, while AM receptors (AM1 and AM2) are generated by the association of CLR with RAMP2 and RAMP3,

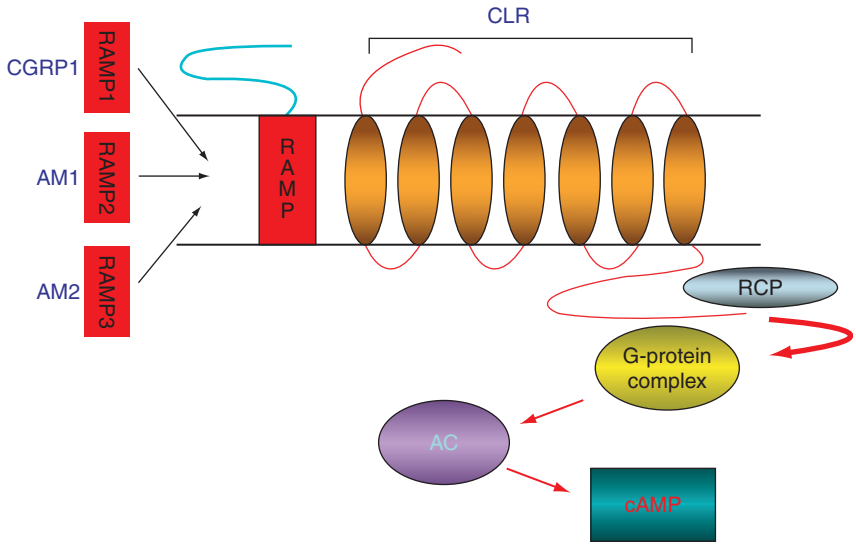

Figure 6 Receptor model for calcitonin gene-related peptide (CGRP) and adrenomedullin (AM). CGRP1, AM1, and AM2 share the calcitonin receptor-like receptor (CLR), a G-protein-coupled seven transmembrane domain. Three receptor activity-modifying proteins (RAMPs) are single transmembrane domain protein and determine the pharmacological specificity of the receptor. The receptor complement protein (RCP) is suggested to play a role in the coupling of the receptor to its subsequent signal transduction pathway via adenylate cyclase (AC), leading to increased production of cAMP. A functional receptor is optimally composed of CLR, RAMP, and RCP, with RAMP1 in CGRP and RAMP2 and 3 in AM1 and AM2, respectively. Adapted from Brain SD and Grant AD (2004) Vascular actions of calcitonin gene-related peptide and adrenomedullin. *Physiological Reviews* 84: 903–934.

respectively. Recently, the second intracellular loop of CLR was shown to provide the molecular determinants for signal transduction and cell surface expression. Human, rat, and porcine CLRs have been cloned and characterized. Currently, a major focus is on the identification of potent and specific nonpeptide antagonists for this receptor in order to establish its physiological and pathophysiological roles. BIBN4096BS is the current prototype, and this molecule has been reported to be effective in the treatment of migraine in a clinical trial.

RCP-like immunoreactivity has been examined in the rat CNS and PNS. Its distribution is broad, closely matching with that of CGRP-IR and CGRP binding sites. Cells expressing CLR and RAMP1, 2, and 3 have recently been shown to be present in the dorsal root ganglia (DRG) and superficial dorsal horn of the rat spinal cord.

CLR knockout mice were recently generated using homologous recombination. Similar to AM knockout mice, homozygous CLR knockout mice exhibited severe cardiovascular defects and died at the embryonic stage, suggesting that CLR is a key protein involved in normal embryonic development.

Biological Activities

The wide distribution of CGRP-IR and receptors suggests a broad involvement in various brain functions and more specifically in sensory, motor, and integrative systems. The main CGRP pathways are described in the following sections and summarized in Figure 4.

CGRP and Olfaction

As mentioned before, high levels of CGRP-IR axons and CGRP receptor binding sites are distributed in the olfactory system, suggesting that CGRP may play a role in olfaction. When applied to cultured mouse olfactory bulbs, CGRP was shown to increase the number of tyrosine hydroxylase/dopamine-expressing neurons *in vitro*. It is still unclear if CGRP is directly involved in the differentiation of dopaminergic olfactory bulb neurons, since CGRP was hardly detectable within most olfactory neurons during development. CGRP-containing axons in the olfactory epithelium were also shown to exhibit a region-specific distribution, with some CGRP-IR axons penetrating through the epithelium to reach the lumen of the nasal cavity. CGRP could either be transported from afferent axons or synthesized locally.

CGRP and Audition

CGRP-containing neurons and axons are observed in audition-related regions such as the superior olive nucleus, medial geniculate body, and lateral lemniscus, suggesting a role for CGRP in the processing of auditory information, although no direct functional data are yet available.

CGRP-containing neurons in the superior olive nucleus were shown to project axons to the ipsilateral cochlea, known to amplify and sharpen mechanical responses to sound. These axons were found to terminate in the inner spinal bundle underneath the inner hair cells and to form synaptic contacts with afferent terminals on these cells. These synapses are in turn controlled and/or regulated by efferents containing neuroactive substances. A cholinergic/CGRP vestibular efferent system could feedback on primary sensory axons and receptor hair cells in the cochlea, but it is not yet clear whether CGRP and acetylcholine are functionally interacting to induce such effects.

CGRP and Vision

CGRP-containing neurons are present throughout the parabigeminal nucleus, which projects essentially to the superior colliculus. The presence of high densities of CGRP binding sites in the superior colliculus and its lower levels in the lateral geniculate nucleus suggests a role for CGRP in the visual system. CGRP-containing cells were also found in the accessory optic system of adult rats, suggesting that CGRP may play a neuromodulatory role in vision and eye movements. CGRP, present in rat corneal nerves, is derived from sensory, sympathetic, and parasympathetic origins. A role for CGRP in vision certainly deserves further investigation.

CGRP and Taste/Feeding

CGRP is expressed in various neurons associated with taste, including sensory axonal endings in taste buds, as well as in central projections of these fibers terminating in the rostral part of the nucleus of the solitary tract. CGRP was also shown to be expressed in the relay system originating from the parabrachial nucleus and projecting to thalamic nuclei. Moreover, most motor neurons of the hypoglossal nucleus are CGRP expressing. Similar to CGRP-IR, CGRP binding sites are noted in the multisynaptic gustatory pathway, including the nucleus of the solitary tract, parabrachial nucleus, central amygdaloid nucleus, ventromedial posterior nucleus of the thalamus, and insular cortex. The presence of both CGRP and CGRP binding sites in olfactory and gustatory systems suggests that CGRP may have a key role in taste and ingestive behaviors.

A role for CGRP in taste and feeding is also supported by functional data. For example, intracerebroventricular (icv) administration of CGRP decreases food intake, and its receptors in the forebrain may mediate the anorexic effect. Lesion of the lateral portion of the parabrachial area, which contains numerous CGRP-positive neurons, induces hyperphagia and obesity whereas its stimulation causes hyperglycemia. Moreover, icv injection of CGRP also suppresses gastric secretion, which is influenced by the lateral and ventromedial regions of the hypothalamus and the central amygdaloid nucleus, two areas enriched with CGRP-IR. CGRP is also present in capsaicin-sensitive primary afferents, supporting its role in hyperemic and acid response. Hence, CGRP is likely to be a significant player in taste and feeding behaviors.

CGRP and Learning Behavior

Abundant CGRP-containing axons and a few CGRP-positive neurons are present in the caudate-putamen and a few CGRP-containing neurons in the thalamus project to various central amygdaloid subnuclei. Most of CGRP-containing axons in the central amygdaloid subnuclei originate from the parabrachial nucleus of the pons. It has been suggested that the overlap between acoustic and somesthetic inputs in the posterior thalamus, and the projection from this region to the amygdala and the striatum, is the anatomical substrate for the formation of learned associations between acoustic stimuli and pain. Bilateral destruction of the medial geniculate nucleus or of the amygdala and striatum prevents changes in autonomic activity and behaviors elicited by acoustic conditioned stimuli. CGRP has also been shown to be involved in the facilitation of learning and memory tasks. The effect of CGRP on learning could be blocked by pretreatment with either a serotonergic, β-adrenergic, or opiate receptor antagonist, suggesting that CGRP modulates several systems that are involved in learning behaviors. CGRP was also shown to attenuate learning impairments induced by MK-801 via a sigma receptor-related mechanism. Finally, CGRP may act as a neurotransmitter in the expression of fear in aversive learning tests, and CGRP released within the amygdala is likely to be involved in auditory fear conditioning.

CGRP and Autonomic Functions

Abundant CGRP-containing neurons are present in the parabrachial nucleus, an important relay center for processing autonomic function-related information, including cardiovascular, respiratory, and sleep regulatory processes. CGRP-containing neurons in the parabrachial nucleus project to a variety of forebrain structures, including the central nucleus of the amygdala and bed nucleus of the stria terminalis. In line with its anatomical distribution, microinjections of CGRP into the central nucleus of the amygdala have elicited increases in arterial blood pressure and heart rate in the rat. After exposure to stressful conditions, the majority of CGRP-containing neurons in the central nucleus of the amygdala exhibited fos immunoreactivity, suggesting that

projection pathways from the parabrachial nucleus to this nucleus and to the medial preoptic area are activated by stress.

CGRP-containing neurons in the parabrachial nucleus and in the caudal and intermediate part of the nucleus of the solitary tract also project to various areas, including the lateral hypothalamus, central nucleus of the amygdala, caudal parts of the caudate-putamen, globus pallidus, lateral septal nucleus, and bed nucleus of the stria terminalis, as well as to layer III of the agranular insular and perirhinal cortices. Accordingly, CGRP may be involved in the relay of visceral sensory information to these autonomic function-related regions. In addition, CGRP contained in both visceral afferent and efferent axons innervate the heart and branchial muscles in the pharynx, providing support to the findings that CGRP modulates a variety of autonomic functions.

CGRP and Motor Neurons, Development and Functions

Virtually all motor pools contain CGRP-IR motor neurons. CGRP is present in both embryonic and adult motor neurons. CGRP may play an active role in the formation of a functional synapse at the neuromuscular junction. CGRP receptor binding sites are also present in myotubes in culture and skeletal muscle membrane preparations. The possible effects of CGRP on striated muscle are (1) inducing increased levels of surface acetylcholine receptors in cultured chick myotubes by enhancing the rate of synthesis and insertion of new receptors into the plasma membrane, (2) increasing the rate of desensitization of nicotinic receptors by reducing acetylcholine-activated channel opening frequency, (3) regulating the phosphorylation of nicotinic acetylcholine receptors in rat myotubes, and (4) inhibiting insulin-stimulated synthesis of glycogen while stimulating glycolysis in mammalian skeletal muscles.

CGRP modulates the biochemical, physiological, and metabolic properties of skeletal muscles through the activation of specific CGRP receptors. Several trophic factors have been shown to regulate the content of CGRP in motor neurons. For example, CGRP-IR is upregulated in mouse motor endplates in which nerve terminal sprouting was induced by the exogenous application of ciliary neurotrophic factor. In addition, basic fibroblast growth factor (bFGF), insulin-like growth factor 1 (IGF-1), and IGF-2 were shown to have important regulatory effects on the expression of CGRPα in rat motor neurons. CGRP levels in motor neurons may also be influenced by axotomy. For example, sciatic nerve section or facial nerve crushing in rat induces a transient increase in CGRP (mRNA and protein) in axotomized motor neurons. Blockade of axonal flow also increases the expression of CGRPα mRNA in motor neurons. CGRP levels in motor neurons are sensitive to supraspinal inputs since thoracic spinal cord transection resulted in decreased levels of CGRP in motor neurons. It is interesting to note that lower CGRP levels were observed in motor neurons of Wobbler mice, a model for amyotrophic lateral sclerosis and infantile spinal muscular atrophy.

CGRP and Sensory Neurons

CGRP-containing cells constitute 40–50% of DRG and trigeminal ganglion neurons. CGRP-containing primary afferents terminate in the superficial layers of the dorsal horn or the trigeminal nucleus. While CGRP-positive cells mainly belong to unmyelinated C and thinly myelinated Aσ axons, only 10–15% of CGRP-containing cells are identified as Aα/β axons. CGRP is frequently shown to coexist with other peptides in primary sensory afferents, including substance P, cholecystokinin, and dynorphin.

The localization of CGRP in small DRG neurons as well as in axonal terminals in the pain-related area of the dorsal horn suggests that CGRP may participate in nociceptive transmission. CGRP is one of the best examples of a neuromodulator that could exert rather limited effects by itself and yet significantly potentiates the nociceptive effects of other substances, such as substance P. For example, CGRP has been shown to inhibit the degradation of substance P and to facilitate the release of both substance P and excitatory amino acids from primary sensory afferents. Intrathecal administration of CGRP was shown to facilitate spinal nociceptive response and to produce a slower onset but prolonged excitation of nociceptive dorsal horn neurons *in vivo*. *In vitro* studies also showed that CGRP can produce a slow depolarization by a direct action on nociceptive dorsal horn neurons.

It has recently been shown that strain dependence of CGRP expression results in variable responses to noxious heat, which could be abolished by peripheral injection of CGRP, blockade of cutaneous and spinal CGRP receptors, or long-term inactivation of CGRP with a CGRP-binding Spiegelmer. Linkage mapping has shown that variable behavioral pain responses were due to differential expression of the *Calca* gene that encodes for CGRPα.

Chronic exposure to opioids such as morphine is shown to result in a marked decrease in its antinociceptive effects, physical dependence, and parallel increases in CGRP content in DRG neurons. Antagonists such as CGRP8–37 and BIBN4096BS are able to prevent or reverse the development of morphine

tolerance and physical dependence, suggesting their potential usefulness in the maintenance of the antinociceptive properties of morphine and derivatives.

It is also well known that CGRP expression in primary sensory neurons is modulated by nerve growth factor, which is required for the survival of CGRP-expressing cells during development and needed for the maintenance of the phenotype during adulthood. Most CGRP-containing nociceptors also co-express transient receptor potential vanilloid receptor 1, a nonselective Ca^{2+} channel which can be activated by pungent capsaicin, heat, and prion. Neonatal capsaicin treatment is able to eliminate almost all CGRP-containing DRG neurons, while capsaicin is able only to deplete CGRP in these cells in adulthood. CGRP is upregulated in DRG neurons during inflammation while downregulated in injured DRG neurons following nerve injury. CGRP also plays an important role in the pathogenesis of migraine and cluster headache. Most interesting, the CGRP antagonist BIBN4096BS has been shown to be effective to treat migraine in a recent clinical trial.

CGRP Effects on Inflammation

CGRP-IR is broadly expressed within the cardiovascular system, consisting of a dense peripheral sensory network innervating the arteries, veins, and heart of various species, including human. CGRP receptor binding sites are also present in the heart and blood vessels. CGRP is one of the most potent vasorelaxing substances known thus far. CGRP-induced vasorelaxation is mediated via CGRP1 receptors, coupled to Gsα protein. CGRP-induced vasodilation is possibly mediated through both nitric oxide- and endothelium-independent and dependent pathways.

Inflammation can be categorized as immunogenic and neurogenic. Immunogenic inflammation is mediated by the activation of specific immune cells, whereas neurogenic inflammation is mediated directly by the activation of nociceptive and thermal-sensitive endings in various tissues. CGRP is actively involved in both immunogenic and neurogenic inflammation. During neurogenic inflammation, tissue damage is able to activate nociceptive axonal endings to cause CGRP-containing nociceptors to release CGRP into inflamed tissues. Released CGRP can, in turn, potentiate the local effects of inflammatory agents including histamine, leukotriene, bradykinin, and substance P, essentially through its vasodilating properties. CGRP can also modulate responsiveness of immune cells such as T-helper cells, B cells, neutrophils, and macrophages. CGRP was recently shown to be produced in lymphocytes and monocytes, as well as in invading macrophages in inflamed areas, and is involved in the production of some proinflammatory chemokines such as monocyte chemoattractant protein 1 and cytokines such as interleukin-1β, tumor necrosis factor-α, and interleukin-6. Hence, CGRP contributes to the inflammatory processes by promoting vasodilation and modulating the activity of various types of immune cells. Therefore, developing more-selective and -potent nonpeptide CGRP receptor antagonists will have clinical significance in the treatment of neurogenic and/or immunogenic inflammation-related diseases such as diabetes, asthma, migraine, nerve injury-associated neuropathic pain, and rheumatoid arthritis.

CGRPα Knockout Mice

Homozygous CGRPα knockout mice with normal calcitonin expression were obtained through target disruption of exon 5 of the calcitonin/CGRPα gene or by the introduction of a stop codon upstream of the CGRPα encoding sequence in exon 5. These two lines of knockout mice were viable and fertile and did not display obvious abnormalities. Calcitonin/CGRPα double knockout mice also developed normally, without obvious defects, but the number of viable fetuses in uterus of homozygous mothers was reduced compared to heterozygous knockout or wild-type mothers.

In CGRPα knockout mice derived from an introduced translational stop codon in exon 5, the cardiovascular system is apparently not affected. But in CGRPα knockout mice generated by targeted disruption of exon 5, mean arterial pressure, heart rate, peripheral vascular resistance, and sympathetic nervous activity are elevated. In calcitonin/CGRPα double-knockout mice, the systemic and mean arterial pressure are also increased. It is clear that the generation of CGRPα-β double-knockout mice is required since CGRPβ is normally expressed in CGRPα knockout mice and could hence have a compensatory effect.

The contribution of CGRPα in pain perception was also examined in CGRPα knockout mice with the selectively disrupted exon 5 of the calcitonin gene, and in calcitonin/CGRPα double knockout mice. Acute pain sensation in CGRPα knockout mice was not affected, but these mice exhibited reduced edema formation and nociception to chemical pain associated with inflammation. Calcitonin/CGRPα double-knockout mice also did not display abnormality in acute pain sensation. Unlike normal mice, these double-knockout mice exhibited no secondary hyperalgesia after arthritic inflammation. These observations suggest that endogenous CGRPα is involved in the production and transmission of pain associated with both neurogenic and arthritic inflammation.

Conclusion and Perspectives

CGRP is widely distributed in both CNS and PNS and possesses a wide variety of biological effects. Although our understanding of CGRP receptors has progressed over the past decade, limited information is available on molecular characteristics and topography of possible subtypes. The fact that receptor recognition is dependent on the level of expression of associated RAMPs adds another level of complexity. Hence, a clear understanding of the role of the various RAMPs in CGRP receptor pharmacology and signaling is required in order to make significant progress on the role of CGRP, related peptides, and their receptors in various diseases.

See also: Motor Neuron Specification in Vertebrates; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Discovery; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Superior Colliculus.

Further Reading

Aiyar N, Rand K, Elshourbagy NA, et al. (1996) A cDNA encoding the calcitonin gene-related peptide type 1 receptor. *Journal of Biochemical Chemistry* 271: 11325–11329.

Amara SG, Jonas V, Rosenfeld MG, et al. (1982) Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products. *Nature* 298: 240–244.

Born W, Fischer JA, and Muff R (2002) Receptors for calcitonin gene-related peptide, adrenomedullin, and amylin: The contributions of novel receptor-activity-modifying proteins. *Receptor Channels* 8: 201–209.

Brain SD and Grant AD (2004) Vascular actions of calcitonin gene-related peptide and adrenomedullin. *Physiological Reviews* 84: 903–934.

Bunton DC, Petrie MC, Hillier C, et al. (2004) The clinical relevance of adrenomedullin: A promising profile? *Pharmacology & Therapeutics* 103: 179–201.

Edvinsson L (2005) Clinical data on the CGRP antagonist BIBN4096BS for treatment of migraine attacks. *CNS Drug Reviews* 11: 69–76.

Jacques D, Dumont Y, van Rossum D, et al. (2000) Calcitonin gene-related peptide (CGRP), amylin and adrenomedullin: Anatomical localization and biological functions in the mammalian and human brains. In: Quirion R, Bjorklund A, and Hokfelt T (eds.) *Peptide Receptors, Part I*, pp. 301–374. Amsterdam: Elsevier.

Juaneda C, Dumont Y, and Quirion R (2000) The molecular pharmacology of CGRP and related peptide receptor subtypes. *Trends in Pharmacological Sciences* 21: 432–438.

Kress A, Jacobowitz DM, and Skofitsch G (1995) Detailed mapping of CGRP mRNA expression in the rat central nervous system: Comparison with previous immunocytochemical findings. *Brain Research Bulletin* 36(3): 261–274.

Luebke AE, Dahl GP, Roos BA, et al. (1996) Identification of a protein that confers calcitonin gene-related peptide responsiveness to oocytes by using a cystic fibrosis transmembrane conductance regulator assay. *Proceedings of National Academy of Science USA* 93: 3455–3460.

McLatchie LM, Fraser NJ, Main MJ, et al. (1998) RAMPs regulate the transport and ligand specificity of the calcitonin-receptor-like receptor. *Nature* 393: 333–339.

Mogil JS, Miermeister F, Seifert F, et al. (2005) Variable sensitivity to noxious heat is mediated by differential expression of the CGRP gene. *Proceedings of National Academy of Science USA* 102: 12938–12943.

Paxinos G and Watson C (1997) *The Rat Brain in Stereotanic Coordinates Compact*, 3rd edn. New York: Academic Press.

Poyner DR (1995) Pharmacology of receptors for calcitonin gene-related peptide and amylin. *Trends in Pharmacological Sciences* 16: 424–428.

Poyner DR, Sexton PM, Marshall I, et al. (2002) International Union of Pharmacology. XXXII. The mammalian calcitonin gene-related peptides, adrenomedullin, amylin, and calcitonin receptors. *Pharmacology Reviews* 54: 233–246.

Rosenfeld MG, Mermod JJ, Amara SG, et al. (1983) Production of a novel neuropeptide encoded by the calcitonin gene via tissue-specific RNA processing. *Nature* 304: 129–135.

Calcium and Signal Transduction

M Baudry, University of Southern California, Los Angeles, CA, USA
X Bi, Western University of Health Sciences, Pomona, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Calcium Regulation

In the course of evolution, calcium has become a fundamental component of intracellular signaling; Ca^{2+}-mediated signaling is present in almost all eukaryotic organisms. The signaling properties of calcium originate from the dynamic characteristics of the mineral: cytoplasmic concentrations of free Ca^{2+} can change very rapidly from the resting levels of 50–100 nM to the millimolar range, and can return to baseline very quickly, too. In other words, Ca^{2+} signaling depends on transient changes of intracellular free Ca^{2+} concentration. Evolution has chosen calcium as a major intracellular signaling messenger, and it also selected a set of mechanisms that tightly control intracellular Ca^{2+} concentration.

Under resting conditions, intracellular Ca^{2+} concentration is very low, 50–100 nM, while extracellular concentration is in the millimolar range (2.5 mM). There is therefore a 10 000-fold difference in Ca^{2+} concentration across the plasma membrane and endo(sarco)plasmic reticulum. The energy contained in this gradient allows rapid influx of Ca^{2+} once the 'gates/channels' open. Maintaining such a steep Ca^{2+} gradient also requires high energy expenditure and depends on the coordinated functions of pumps/transporters located in the plasma membrane and various subcellular organelles, and on a number of Ca^{2+}-binding proteins (Figure 1).

Plasma Membrane Regulation of Ca^{2+} Levels

Plasma membrane Ca^{2+} pump and exchanger Low intracellular Ca^{2+} concentration is achieved in part by the activation of various proteins in different subcellular locations. At the plasma membrane level, two systems are involved in the extrusion of Ca^{2+}: a low-affinity, high-capacity, and electrochemically driven Na^+/Ca^{2+} exchanger (NCX), and a high-affinity, low-capacity Ca^{2+}-ATPase. The plasma membrane Ca^{2+}-ATPase (PMCA) has a $K_d < 0.5\,\mu M$ and is assumed to play a dominant role in cells when cytosolic concentration is around 0.1–0.2 μM, while the exchanger ($K_d \sim 10$–$15\,\mu M$) functions best when Ca^{2+} concentration rises to 1–10 μM. PMCA is activated by acidic phospholipids, with the mono- and bis-phosphorylated derivatives of phosphatidylinositol being the most effective ones. PMCA is also regulated by calmodulin and other Ca^{2+}-activated enzymes. For instance, calpain truncation of its C-terminal domain makes the pump constitutively active, while phosphorylation by protein kinase A (PKA) and protein kinase C (PKC) increases its Ca^{2+} affinity. In humans, four different isoforms of PMCA, which are products of more than 26 transcripts of four genes, have been identified. While PMCA1 and PMCA4 are ubiquitously distributed, PMCA2 and PMCA3 are mainly found in brain. Interestingly, expression levels of PMCA are regulated by Ca^{2+}.

While PMCA is the dominant Ca^{2+} ejector at the plasma membrane of nonexcitable cells, the NCX plays a fundamental role in excitable tissues. The capacity of the plasma membrane Na^+/Ca^{2+} exchanger to pump calcium is 10 times greater than that of PMCA. Unlike PMCA, the NCX is not an enzyme; it is a transporter mediating the import of three Na^+ ions for the extrusion of one Ca^{2+}. The extrusion of Ca^{2+} against its large electrochemical gradient depends on the energy of the Na^+ gradient generated by the Na^+ pump. Depending on the thermodynamic values of Ca^{2+} and Na^+ and on the membrane potential, the direction of the NCX-mediated transport of Na^+ and Ca^{2+} can be reversed – that is, Ca^{2+} may enter cells in exchange for extruded Na^+. The activity of NCX is also regulated by intracellular pH, ATP levels, other ions, and redox agents. In addition, phsophorylation of NCX by PKA and PKC has been shown to increase its activity. To date, three different genes encoding three NCX isoforms (NCX1–NCX3) have been cloned in mammals. The distribution of these NCX isoforms has also been analyzed: NCX1 is expressed in most tissues, including brain and heart, while NCX2 and NCX3 are mainly detected in neuronal and skeletal muscle tissues. Analyses of messenger RNA and protein levels have revealed that all three isoforms of NCX are highly expressed in most central nervous system (CNS) structures, although in some brain areas they are differentially expressed. Interestingly, the NCX1 isoform exists at both pre- and postsynaptic sites.

Plasma membrane Ca^{2+} channels Besides pumps and exchangers, the plasma membrane also contains Ca^{2+} channels. Ca^{2+} channels are transmembrane proteins that function as a 'gate.' These gates are generally closed in resting cells, but when the gates open, Ca^{2+} rapidly flows into the cytoplasm down its electrochemical gradient. There are three types of Ca^{2+}

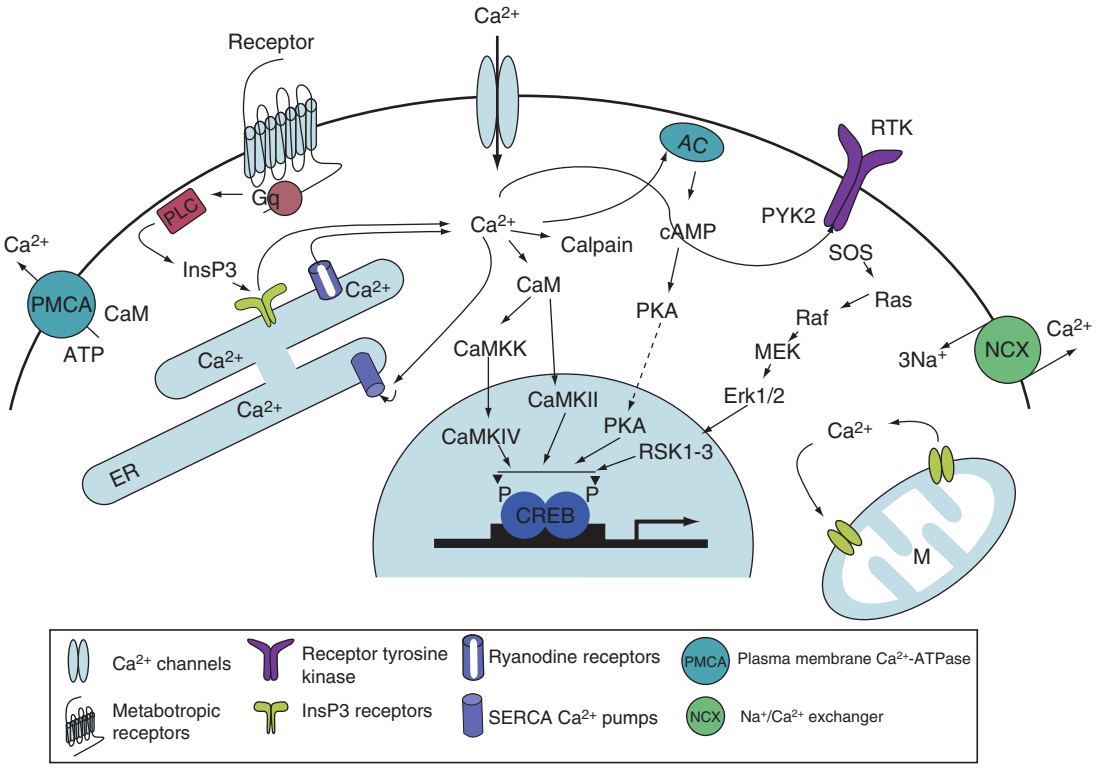

Figure 1 Schematic representation of the major mechanisms involved in Ca^{2+} regulation and signaling, illustrating the various mechanisms involved in Ca^{2+} influx, transport, extrusion, and storage present in various cellular compartments. Also illustrated are several mechanisms involved in Ca^{2+} signaling, including phosphorylation processes, gene regulation, and protein degradation. As discussed in the text, age-related alterations in several of these mechanisms have been identified and could be responsible for age-related decline in cognitive function and increased susceptibility to diseases and insults. AC, adenylate cyclase; CaM, calmodulin; CaMKK, CaMKII, CaMKIV, calmodulin-dependent protein kinases; cAMP, cyclic adenosine monophosphate; CREB, cyclic adenosine monophosphate (cAMP) response element-binding protein; ER, endoplasmic reticulum; Erk1/2, extracellular signal-related kinase 1/2; InsP3, inositol triphosphate; M, mitochondrion; MEK, mitogen-activated protein extracellular kinase; NCX, Na^+/Ca^{2+} exchanger; PKA, protein kinase A; PLC, phospholipase C; PMCA, plasma membrane Ca^{2+}-ATPase; PYK2, proline-rich tyrosine kinase 2; RSK1-3, ribosomal S6 kinases 1–3; RTK, receptor tyrosine kinase; SERCA, sarco(endo)plasmic reticulum Ca^{2+}-ATPase; SOS, son of sevenless exchange factor.

channels, defined by the mechanisms controlling their opening and closing: (1) voltage-operated channels (VOCs), the function of which depends on membrane potential, (2) receptor-operated channels (ROCs) that are gated by binding of ligands to their receptors, and (3) store-operated channels (SOCs) that are controlled by Ca^{2+} levels in the endoplasmic reticulum (ER). VOCs are the best characterized Ca^{2+} channels and have high Ca^{2+} selectivity. VOCs are widely distributed, not only in excitable tissues, but also in non-excitable tissues. At the single-channel level, these channels are subdivided into L, N, P, Q, R, and T types based on their physiological and pharmacological characteristics. Recent molecular work reveals that, structurally, VOCs are large complexes consisting of either four or five subunits that are encoded by five separated genes ($\alpha 1$, $\alpha 2$, β, γ, and δ); a1, the product of the $\alpha 1$ gene is the largest and can form a channel, while the other subunits are regulators.

A number of Ca^{2+} channels belong to the ROC family, including both ionotropic and metabotropic glutamate receptor families. According to their ligand selectivity, ionotropic glutamate receptors are further divided into α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-, *N*-methyl-D-aspartate (NMDA)-, and kainic acid (KA)-type receptors. Metabotropic glutamate receptors are receptors coupled with G-proteins. Genes encoding all glutamate receptors have been cloned and their structures and functions have been characterized. Among these receptors, NMDA receptors have the highest Ca^{2+} permeability. Interestingly, activation of NMDA receptors also depends on membrane potential, which means that Ca^{2+} flux through this receptor channel is also voltage gated. Traditionally, AMPA receptors are considered Ca^{2+} impermeable. Following the cloning of this family of receptors, it became clear that Ca^{2+} impermeability was due to the presence of one of the

subunits, GluR2. In rare cases in which AMPA receptor complexes do not contain this subunit, they are Ca^{2+} permeable. Activation of ROCs is also subjected to regulation by proteins kinases and proteases.

SOCs have been best studied in neuroendocrine cells, in which depletion of hormones and Ca^{2+} induces activation of SOCs. SOCs have also been found in other cell types, including mast cells, basophils, and neurons. However, their physiological and pharmacological properties, molecular structures, and regulation are still largely unknown.

The ER

The ER is the best characterized Ca^{2+} signaling organelle. It has large Ca^{2+} buffering capacity and functions as the main intracellular Ca^{2+} storage compartment. The ER Ca^{2+} buffering and storage properties originate from the concerted action of a number of proteins. Ca^{2+} uptake into the ER lumen results from the activation of Ca^{2+} pumps, belonging to the sarco(endo)plasmic reticulum Ca^{2+}-ATPase (SERCA) family. Release of Ca^{2+} from the ER is controlled by inositol triphosphate (InsP3)-gated channels and ryanodine receptors (RyRs), also know as Ca^{2+}-gated Ca^{2+} channels. Ca^{2+}-binding proteins, such as calsequestrin and calreticulin, also contribute to Ca^{2+} storage in the ER. Ca^{2+} concentration in the ER has been estimated to be in the millimolar range, which generates a large electrochemical gradient for Ca^{2+} across the ER membrane. RyRs are generally activated by Ca^{2+} influx at the plasma membrane, induced by depolarization or binding of ligands to ROCs. InsP3-gated channels, also know as InsP3 receptors, are activated by the binding of InsP3. InsP3 is a product of phospholipase C (PLC)-mediated hydrolysis of the membrane phospholipids, phosphatidylinositol 4,5-bisphosphate (PtdInsP2), generally as a result of activation of G-protein-coupled receptors in the membrane. In addition, InsP3 receptors are regulated by Ca^{2+}, ATP, and phosphorylation by PKA. The positive feedback stimulation of RyRs and InsP3 receptors by Ca^{2+} leads to a regenerative opening of the release channels and contributes to the formation and propagation of Ca^{2+} waves.

Mitochondria and Ca^{2+} Regulation

In addition to the ER, mitochondria also contribute to Ca^{2+} homeostasis in the cytoplasm; this function of mitochondria becomes more important in nerve terminals in the presence of high levels of electrical activity. In particular, increasing evidence indicates that mitochondria limit the rise in intracellular calcium concentration ($[Ca^{2+}]_i$) that occurs during periods of intense activity. This function might be particularly critical under conditions of reduced energy supply (ischemia, hypoglycemia) or following injury, as well as during aging. Alternatively, it has been proposed that Ca^{2+} uptake by mitochondria serves to adjust the concentration of free Ca^{2+} inside the mitochondria, which in turn controls the rate of energy metabolism by regulating the activity of a number of Ca^{2+}-dependent intramitochondrial enzymes. Another critical role of the mitochondria in apoptotic cell death has been identified: the mitochondrial permeability transition pore (mPTP) is a complex of proteins in the mitochondrial membrane that is activated by a number of factors, and in particular by calcium when its concentration reaches a certain threshold inside the mitochondria. Following opening of the mPTP, several factors, including apoptotic-inducing factor and cytochrome *c*, are released from the mitochondria and initiate apoptotic cell death.

Ca^{2+} Signal Transduction

Ca^{2+} signaling is so ubiquitous that it is involved in almost every single function of a living cell. In neurons, Ca^{2+} signaling controls neurogenesis, neuronal differentiation, axonal growth, dendritic maturation, and synaptic plasticity. Ca^{2+} also contributes critically to neuronal death under various conditions. At the molecular level, Ca^{2+} signaling regulates transcription and posttranscriptional modifications of mRNA, translation and posttranslational regulation of proteins, and protein phosphorylation and degradation. Among the diverse properties of Ca^{2+} signaling, one feature that needs to be emphasized is the so-called coupling function. Ca^{2+} signaling couples extracellular signals (neurotransmitters, growth factors, cytokines, etc.) to intracellular responses (enzymatic reactions, gene expression, cytoskeleton rearrangement, release of neurotransmitters, etc.).

Ca^{2+} and Transcription Regulation

Ca^{2+} signaling control of gene expression has been known for over a decade. In the CNS, Ca^{2+}-dependent transcriptional regulation is involved in dendritic morphogenesis, path finding, and activity-dependent synaptic modifications leading to adaptive properties of brain function. The best characterized signaling pathway that conveys Ca^{2+} signals to gene regulation is the calmodulin and Ca^{2+}/calmodulin-dependent protein kinase (CaMK) system (Figure 1). Elevated Ca^{2+} levels, resulting from the activation of channels located on plasma membrane or from Ca^{2+} released from ER, activate CaMKs indirectly via binding to calmodulin.

CaMKs, especially CaMKIV and CaMKII, then phosphorylate a transcription factor, the cyclic adenosine monophosphate (cAMP) response element-binding protein (CREB), at residue Ser133, which is critical for factor activation. Recent work has shown that nuclear Ca^{2+} can also increase transcriptional activity by recruiting a coactivator, the CREB-binding protein (CBP). Interestingly, phosphorylation of CREB depends on the entry of Ca^{2+}: Ca^{2+} influx through L-type voltage-gated Ca^{2+} channels (VGCCs) leads to sustained CREB Ser133 phosphorylation, while Ca^{2+} entry through NMDA receptors triggers only transient phosphorylation of CREB at Ser133, at least in cultured embryonic neurons. The transcriptional activity of CREB is also regulated by other Ca^{2+}-dependent signaling pathways, such as activation of PKA and the Ras/mitogen-activated protein kinase (MAPK) pathway. In fact, CREB was first reported as a transcriptional factor that responded to elevation of cellular cAMP levels induced by activation of PKA. Activation of CREB through the Ras/MAPK pathway involves the recruitment of nuclear RSK kinases (including RSK2) that phosphorylate CREB at Ser133. Brain-derived neurotrophic factor (BDNF) plays critical roles in development and synaptic plasticity by regulating CREB in an activity-dependent manner. In addition to protein kinases, protein phosphatases such as calcineurin also contribute to Ca^{2+}-induced gene regulation. Finally, Ca^{2+} can also influence gene expression by directly binding to transcriptional factors.

Ca^{2+}-Mediated Regulation of Protein Phosphorylation and Dephosphorylation

Protein phosphorylation is an enzymatic reaction that modifies either serine/threonine or tyrosine residues. Numerous Ser/Thr-type protein kinases and one Ser/Thr protein phosphatase are regulated by Ca^{2+}. In general, Ca^{2+} regulates these kinases indirectly by binding to calmodulin, with the exception of PKC, in which case Ca^{2+} and phospholipids combine to regulate the activity of this kinase. Interestingly, CaMKII is abundantly present in essentially all neuronal compartments. While CaMKIV plays a predominant role in CREB phosphorylation and gene regulation, CaMKII is enriched in postsynaptic densities and is critically involved in synaptic function and plasticity. CaMKII is also subject to modulation by autophosphorylation at Thr286, a process that has been proposed to participate in hippocampal CA1 region long-term potentiation (LTP) and in hippocampus-dependent behaviors such as spatial learning and memory. CaMKII phosphorylates a large number of synaptic proteins, including AMPA and NMDA receptors. Recent work suggests that CaMKII enhances synaptic function by phosphorylation of AMPA receptor GluR1 subunits at Ser831 and by promoting insertion of new receptors into synapses. Other kinases, such as PKA and PKC, have also been implicated in synaptic plasticity. In a recent study, Sacktor and colleagues showed that an inhibitor of an atypical PKC isoform (PKCζ) not only reversed LTP maintenance but also erased existing memories.

Ca^{2+}-Dependent Protein Degradation

Proteins are generally degraded by proteases in the cytosol or in specialized organelles, such as lysosomes. A large number of proteases are under the regulation of Ca^{2+}-dependent signaling cascades. Among these, Ca^{2+}-dependent proteases (calpains) and caspases are the best characterized. A few studies have also reported that the ubiquitin–proteasome system is under the regulation of Ca^{2+}, although the reported results were not conclusive. Calpains, a 16-member family of proteases, have attracted tremendous interest due to their diverse roles under physiological as well as pathological conditions. Structurally, calpains are heterodimers composed of a large catalytic subunit and a small subunit. The function of the small subunit is still not clear. The crystal structure of this subunit reveals the existence of Ca^{2+} binding motifs, the binding ability of which has been experimentally confirmed. The large subunit varies in different calpain isoforms. The most common forms of calpains are μ-calpain, activated *in vitro* by micromolar calcium concentrations, and m-calpain, activated at millimolar calcium concentrations. The unique property that distinguishes calpain from other proteases is that it does not completely degrade its substrates into inactive fragments; instead, it produces a partial proteolysis that generally changes the function of the proteins. To date more than 100 proteins have been characterized as calpain substrates. These proteins belong to several classes of proteins, including transcription factors, structural proteins, receptors or channels, protein kinases and phosphatases, and other proteases. The variety of calpain substrates further stresses the diversity of functions in which this family of proteases is implicated. In the CNS, calpain has been implicated in cell migration, axonal growth, synaptic plasticity, and, most importantly, neurodegeneration. In addition to Ca^{2+}, calpain activity is regulated by multiple factors, including phosphorylation and association with membrane phospholipids and with DNA. There is also a highly selective endogenous calpain inhibitor, calpastatin. Calpastatin is a ubiquitously expressed protein of 110 kDa. Interestingly, binding of calpain to its inhibitor is also Ca^{2+} dependent.

Ca^{2+} Regulation and Signal Transduction in Aged Brains

The 'calcium hypothesis of aging,' proposed by Zaven Khachaturian and Gibson and Peterson more than 20 years ago, postulated that a slow disruption of Ca^{2+} homeostasis accompanied the aging process and was responsible for many of the pathological consequences of aging. The words that Khachaturian wrote then, "it is clear that any significant and long lasting change in the normal functioning of Ca^{2+} transport systems, pumps, buffers or storage systems that help maintain the homeostasis could influence the delicate balance of Ca^{2+} concentration in the cytosol, with serious consequences," are still echoed today in most studies of Ca^{2+} regulation and aging. The initial idea that basal Ca^{2+} concentration was increased in aged cells, and in particular in neurons, has been largely disproved. The current consensus is that the major difference between neurons from young and old animals consists in the rate of $[Ca^{2+}]_i$ recovery following stimulation-induced increase in intracellular Ca^{2+} level. The lack of changes in basal Ca^{2+} concentrations indicates that none of the mechanisms involved in setting this level is significantly altered with aging, and in particular, that the metabolic activity of aged neurons is sufficiently normal to provide the energy required for establishing this set point. On the other hand, the slower recovery of Ca^{2+} following increased Ca^{2+} influx suggests that one or several processes that are engaged following a rapid increase in Ca^{2+} concentrations are altered with aging. Among these, it has been clearly demonstrated that the activity of the plasma membrane Ca^{2+}-ATPase is decreased with aging. One of the hypotheses proposed to account for such decrease is a modification of Ca^{2+}-ATPase regulation by calmodulin, possibly due to increased oxidation of this regulatory protein, as well as a large decrease in the levels of calmodulin. Activity of the SERCA is also decreased with aging, resulting in changes in endoplasmic Ca^{2+} homeostasis. However, it is worth mentioning that age-related changes in SERCA are quite mild, in contrast to what is observed under pathological conditions, in which changes in SERCA play a more important role. On the other hand, no age-related changes in the Na^+/Ca^{2+} exchanger have been reported in neurons. A hallmark of aged neurons is an increase in the number of L-type voltage-gated calcium channels. As will be discussed in the following section, this effect has important functional consequences for normal synaptic physiology and for synaptic plasticity.

As discussed earlier, an important regulator of Ca^{2+} homeostasis consists in the pool of mitochondria. Numerous morphological studies have indicated that while the number of mitochondria might decrease with aging, their size increases. Functionally, aged mitochondria also exhibit chronic depolarization, which could be due to a decreased activity of the electron transport chain. Such decreased activity of the electron transport chain and the resulting membrane depolarization might account for the reported decrease in Ca^{2+} uptake by aged mitochondria. More importantly, alterations in mitochondria with aging result in their decreased effectiveness to respond to any increase in energy demand occurring during periods of increased neuronal activity. Similarly, any fluctuations in the supply of energy substrates or in oxygen will have important consequences for the formation of ATP and consequently for the functions of the multiple ATP-dependent enzymes involved in Ca^{2+} homeostasis. The consensus that is therefore emerging is that alterations in Ca^{2+} homeostasis with aging result from relatively subtle changes in some elements contributing to Ca^{2+} regulation, together with alterations in the metabolic status of the cells due to increased mitochondrial dysfunction.

Functional Consequences of Age-Related Changes in Ca^{2+} Signaling

As previously mentioned, Ca^{2+} is an important signaling molecule, and any alteration in Ca^{2+} homeostasis will have profound functional consequences at all levels of neuronal function, and this is one of the reasons involved in formulating the 'calcium hypothesis of aging.' One of the immediate consequences of the increased number of L-type voltage-gated calcium channels is the increase in the afterhyperpolarization that follows action potentials due to increased activity of Ca^{2+}-dependent potassium channels. This phenomenon has been well described by many groups and is responsible for the decreased excitability found in aged neurons and, more importantly, for decreased synaptic plasticity, in particular in LTP, as it makes summation of depolarization more difficult and therefore decreases the activation of NMDA receptors, a critical step in LTP induction. Interestingly, blockade of the L-type channels or of some of the Ca^{2+}-activated potassium channels reestablishes normal LTP and learning in aged animals.

Another consequence of the alterations in Ca^{2+} homeostasis is that Ca^{2+} signals following intense neuronal activity are elevated and prolonged, thus producing excessive activation of all of the targets and signaling pathways normally activated by intracellular Ca^{2+}. One of the pathways activated by Ca^{2+} leads to regulation of gene expression, and it has

recently been shown that aging is associated with much wider changes in gene expression than initially thought. In particular, expression of pathways related to inflammation, mitochondrial function, oxidative stress responses, and synaptic plasticity is significantly increased in aged animals. Interestingly, calpain activation has also been shown to be increased in several brain regions and cellular locations in aging, as increased levels of calpain-mediated spectrin degradation products are present in cerebral cortex and cerebellum from aged brains. Furthermore, activated calpain 1 has also been found in white matter and in activated microglia in aged monkeys. Considering the numerous cross-talks between protein phosphorylation and protein degradation, it is apparent that alterations in Ca^{2+} regulation will have complex effects on the extent and duration of changes in protein phosphorylation, susceptibility to protease degradation, and, ultimately, functional modifications. Another critical factor, mentioned previously, relates to the age-related increase in oxidized proteins due to increased formation of reactive oxygen species. In the case of calmodulin, this results in its decreased ability to stimulate PMCA and other calmodulin-dependent enzymes.

Conclusions

All the data accumulated since the original formulation of the 'calcium hypothesis of aging' have led to revision of the details of the hypothesis while maintaining and strengthening its overall validity. It is now clear that multiple factors participating in Ca^{2+} regulation and signaling are altered during aging and contribute to alterations in Ca^{2+} homeostasis and signaling functions. Some of these processes have been directly linked to age-related cognitive decline, while others have been shown to be involved in increased neuronal vulnerability to age-related diseases and insults. While we do know much more about the mechanisms involved in Ca^{2+} homeostasis and signaling and their alterations in aging, we are far from being able to propose solutions that would readily counteract these changes and provide the long-sought élixir de jouvence.

See also: Calcium Channels; Calcium Channel Subtypes Involved in Neurotransmitter Release; Calcium Channels and SNARE Proteins; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Calcium Homeostasis in Glia; Large Conductance Calcium-Activated Potassium Channels; Neuromodulation of Calcium Channels; Voltage-Gated Calcium Channels.

Further Reading

Annunziato L, Pignataro G, and Di Renzo GF (2004) Pharmacology of brain Na^+/Ca^{2+} exchanger: From molecular biology to therapeutic perspectives. *Pharmacological Reviews* 56: 633–654.

Bertoni-Freddari C, Fattoretti P, Giorgetti B, et al. (2003) Role of mitochondrial deterioration in physiological and pathological brain aging. *Gerontology* 50: 187–192.

Canitano A, Papa M, Boscia F, et al. (2002) Brain distribution of the Na^+/Ca^{2+} exchanger-encoding genes *NCX1, NCX2*, and *NCX3* and their related proteins in the central nervous system. *Annals of the New York Academy of Sciences* 976: 394–404.

Carafoli E, Santella L, Branca D, et al. (2001) Generation, control, and processing of cellular calcium signals. *Critical Reviews in Biochemistry and Molecular Biology* 36: 107–260.

Chan DC (2006) Mitochondria: Dynamic organelles in disease, aging, and development. *Cell* 125: 1241–1252.

Crompton M (2004) Mitochondria and aging: A role for the permeability transition? *Aging Cell* 1: 3–6.

Ermak G and Davies KJ (2002) Calcium and oxidative stress: From cell signaling to cell death. *Molecular Immunology* 38: 713–721.

Gareri P, Mattace R, Nava F, et al. (1995) Role of calcium in brain aging. *General Pharmacology* 26: 1651–1657.

Gibson GE and Peterson C (1987) Calcium and the aging nervous system. *Neurobiology of Aging* 8: 329–343.

Khachaturian ZS (1987) Hypothesis on the regulation of cytosol calcium concentration and the aging brain. *Neurobiology of Aging* 8: 345–346.

Landfield PW (1996) Aging-related increase in hippocampal calcium channels. *Life Sciences* 59: 399–404.

Michaelis ML, Bigelow DJ, Schoneich C, et al. (1996) Decreased plasma membrane calcium transport activity in aging brain. *Life Sciences* 59: 405–412.

Montminy MR and Bilezikjian LM (1987) Binding of a nuclear protein to the cyclic-AMP response element of the somatostatin gene. *Nature* 328: 175–178.

Nicholls DG (2002) Mitochondrial function and dysfunction in the cell: Its relevance to aging and aging-related disease. *International Journal of Biochemistry & Cell Biology* 34: 1372–1381.

Nixon RA (2003) The calpains in aging and aging-related diseases. *Ageing Research Reviews* 2: 407–418.

Philipson KD and Nicoll DA (2000) Sodium–calcium exchange: A molecular perspective. *Annual Review of Physiology* 62: 111–133.

Squier TC and Bigelow DJ (2000) Protein oxidation and age-dependent alterations in calcium homeostasis. *Frontiers in Bioscience* 5: 504–526.

Toescu EC, Verkhratsky A, and Landfield PW (2004) Ca^{2+} regulation and gene expression in normal brain aging. *Trends in Neurosciences* 27: 614–620.

Verkhratsky A (2005) Physiology and pathophysiology of the calcium store in the endoplasmic reticulum of neurons. *Physiological Reviews* 85: 201–279.

West AE, Chen WG, Dalva MB, et al. (2001) Calcium regulation of neuronal gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11024–11031.

Calcium Channel and Calcium-Activated Potassium Channel Coupling

N V Marrion, University of Bristol, Bristol, UK

© 2009 Elsevier Ltd. All rights reserved.

Activation of calcium (Ca^{2+})-activated potassium (K^+) channels by influx of Ca^{2+} ions through voltage-dependent Ca^{2+} channels is critical for regulating cell excitability in many cell types. The activation of Ca^{2+}-dependent K^+ channels can hasten the repolarization of an action potential or modulate the frequency of action potential firing. In early years, it was of paramount importance to identify the subtype of Ca^{2+}-activated channel and resolve the physiological role of its activation. The origin of intracellular Ca^{2+} was of little significance, and it was generally assumed that all possible sources of Ca^{2+} entry were involved and that Ca^{2+}-activated channels simply responded to global changes of intracellular Ca^{2+}. In recent years, the pharmacology of blockers of Ca^{2+} channels has improved dramatically, permitting some investigation into the identity of voltage-dependent Ca^{2+} channels that mediate the entry of Ca^{2+} to cause activation of Ca^{2+}-activated K^+ channels. There is now considerable interest in whether there is selective pairing of Ca^{2+} and Ca^{2+}-activated channels.

Regulating Membrane Excitability by Activation of Ca^{2+}-Activated Potassium Channels

Activation of Ca^{2+}-activated potassium channels results in membrane hyperpolarization that is crucial in regulating action potential duration and firing frequency. For example, activation of large conductance Ca^{2+}-activated potassium (BK, $K_{Ca}1.1$) channels hastens the repolarization of the action potential in hippocampal pyramidal neurons and underlies generation of the fast afterhyperpolarization (AHP) that controls the frequency of action potential firing at the onset of a burst. In contrast, a train of action potentials in hippocampal CA1 pyramidal neurons is terminated by an AHP that displays two main time components. The AHP results, at least in part, from activation of small conductance Ca^{2+}-activated potassium (SK, $K_{Ca}2$) channels. AHPs can be broadly classified into two groups – those that are sensitive to the bee venom toxin apamin and those that are not. Most examples of AHPs in excitable cells are apamin sensitive, with only a small number of cell types (e. g., cortical and hippocampal pyramidal cells) exhibiting apamin-insensitive slow AHPs.

BK ($K_{Ca}1.1$) channels are activated by the concerted action of membrane depolarization and a rise of intracellular Ca^{2+}, with the concentration of Ca^{2+} required to open a BK channel being less at depolarized levels. In general, BK channels are considered to exhibit a low affinity for Ca^{2+} binding, requiring submembrane concentrations of Ca^{2+} in excess of $1\,\mu M$. In contrast, activation of SK ($K_{Ca}2$) channels is voltage independent and the channels display a higher affinity for Ca^{2+} binding, being activated by submicromolar concentrations of intracellular Ca^{2+}. The different sensitivities of BK and SK channels to intracellular Ca^{2+} have obvious implications for where the Ca^{2+}-activated channel may be located relative to the source of Ca^{2+}. In general, the concentration of intracellular Ca^{2+} required to activate either BK or SK channels has not been observed during their activation. This suggests that neither BK nor SK channels are responding to global changes of intracellular Ca^{2+} and that a local submembrane rise of Ca^{2+} is more likely to be responsible for the activation of both of these channels. The opening of a voltage-gated Ca^{2+} channel will provide a submembrane microdomain of raised Ca^{2+} that will last for several hundred microseconds. The concentration of Ca^{2+} reached in the microdomain will be dependent on the amount of Ca^{2+} that enters during a single Ca^{2+} opening, which will depend on the channel conductance and amount of time that it is open. Because the Ca^{2+} concentration will decrease exponentially with distance from the microdomain core, Ca^{2+}-activated channels can be placed at particular distances from the Ca^{2+} channel to experience the appropriate concentration of Ca^{2+}. The lower sensitivity of BK channels to Ca^{2+} compared with SK channels strongly suggests that BK channels reside very close to or possibly in direct contact with voltage-gated Ca^{2+} channels. In contrast, it is likely that SK channels are located at some distance from Ca^{2+} channels (Figure 1) because they require less Ca^{2+} to be activated.

The pharmacology of voltage-dependent Ca^{2+} channels has improved in recent years, with compounds available to distinguish between native T-, L-, N-, P/Q-, and R-type calcium channels. In most cases, functional coupling has been inferred from the effects of block of a Ca^{2+} channel subtype on the

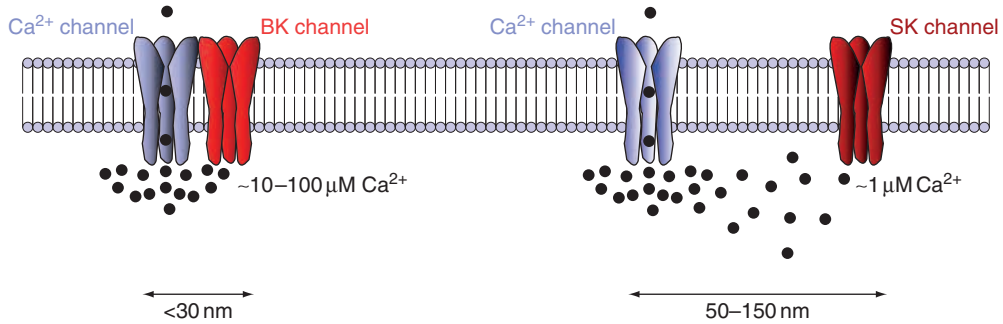

Figure 1 Schematic representation of possible locations of voltage-dependent Ca^{2+} channels and the Ca^{2+}-activated BK and SK channels. The combination of measured Ca^{2+} sensitivities of BK and SK channels and experimental data suggests that a clear difference of relative locations of these Ca^{2+}-activated K^+ channels to Ca^{2+} channels exists.

activation of a Ca^{2+}-activated potential or membrane current. These types of data fail to distinguish between functional coupling resulting from a specific interaction between channel subtypes and that arising because channels share the same subcellular location.

Pharmacological Determination of Ca^{2+} Channel Subtypes Underlying Activation of Ca^{2+}-Dependent Hyperpolarizations in Central Neurons

Identity of the Ca^{2+} Channel Subtype Activating Apamin-Sensitive AHPs

The cloning of SK channel subunits revealed three subtypes termed K_{Ca}2.1–2.3. Apamin-sensitive AHPs are likely derived from either K_{Ca}2.2 or -2.3 channel subunits. Although each cloned subunit displays characteristic sensitivity to the bee venom toxin, most brain slice experiments have used a concentration of apamin (>10 nM) that is too high to permit discrimination of the identity of the apamin-sensitive SK channel. Most of this work has been performed using current clamp recording from central neurons and the effect of Ca^{2+} channel subtype-specific blockers on Ca^{2+}-dependent hyperpolarizations has been assessed, with most examples of apamin-sensitive AHPs decaying with a time constant of approximately 70–100 ms.

Coupling of Ca^{2+} Channel Subtypes to Activation of Apamin-Sensitive AHPs

Ca_V1.2 and -1.3 L-type Ca^{2+} channels Both Ca_V1.2 and -1.3 L-type Ca^{2+} channels are blocked by dihydropyridine antagonists. There are two examples of apamin-sensitive AHPs with slow decay reported to result from Ca^{2+} entry through L-type Ca^{2+} channels. The example reported from gonadotropin-releasing hormone neurons was shown to be generated from Ca^{2+} entry through L-, N-, Q-, and R-type Ca^{2+} channels.

Ca_V2.2 N-type Ca^{2+} channels N-type Ca^{2+} channels are blocked selectively by ω-conotoxin GVIA. It has been observed that a large component of the apamin-sensitive AHP found in both nucleus basalis cholinergic neurons and subthalamic nucleus neurons is sensitive to the conotoxin.

Ca_V3 T-type Ca^{2+} channels A minor contribution of Ca^{2+} entry through T-type Ca^{2+} channels underlies generation of an apamin-sensitive AHP in nucleus basalis cholinergic neurons. In contrast, activation of the apamin-sensitive AHP in dopaminergic midbrain neurons is selectively coupled to T-type Ca^{2+} channels.

Ca_V2.1 P/Q-type Ca^{2+} channels P- and Q-type Ca^{2+} channels are both derived from the Ca_V2.1 subunit but differ in their apparent sensitivity to the toxin ω-agatoxin IVA. P-type Ca^{2+} currents are blocked by low concentrations of ω-agatoxin IVA ($IC_{50} = 1$–3 nM), whereas Q-type currents require higher concentrations ($IC_{50} = 90$ nM). In contrast, both P- and Q-type (and N-type) channels are blocked by ω-conotoxin MVIIC. In practice, it is difficult to distinguish between P- and Q-type currents using this pharmacology. However, the apamin-sensitive AHP in neocortical neurons is blocked by approximately 20% by 25 nM ω-agatoxin IVA. This suggests that Ca^{2+} entry through P-type channels activates approximately 20% of this AHP. It appears that a component of Q-type current may also contribute to the generation of the apamin-sensitive AHP in neocortical neurons because addition of ω-conotoxin MVIIC after ω-agatoxin IVA blocked another 20% of the AHP. In contrast, ω-conotoxin GVIA alone had no effect. A similar incomplete block by ω-conotoxin MVIIC of an apamin-sensitive AHP is observed in cerebellar Purkinje cells.

It is clear that there is not a common pairing between the Ca^{2+}-activated SK and Ca^{2+} channel to generate apamin-sensitive AHPs in different neurons. In addition, there are several examples of the apamin-sensitive AHP being generated from Ca^{2+} entry through

multiple Ca^{2+} channel subtypes. Clearly, there are two interpretations of these data. It is possible that the apamin-sensitive SK channel is associated with a number of different Ca^{2+} channel subtypes in those neurons. This seems unlikely and it probably suggests that apamin-sensitive SK channels are not functionally coupled to Ca^{2+} entry through a particular Ca^{2+} channel subtype. The current underlying the apamin-sensitive AHP is rapid in onset and decays with an exponential time course that is similar to that of the SK channel following the rise of intracellular Ca^{2+} from uncoupled sources.

Identity of the Ca^{2+} Channel Subtypes Activating Apamin-Insensitive AHPs

An apamin-insensitive AHP is seen in a small number of cell types, such as hippocampal and cortical pyramidal neurons. It exhibits a very unusual time course when present, displaying both a slow rise (peaking approximately 0.5–1 s after the end of the action potential train) and decay phase (exponential decay of time constant of 1.5–3 s). There has been considerable discussion regarding the mechanism(s) underlying this unusual time course, with hypotheses ranging from a large distance separating Ca^{2+} and Ca^{2+}-activated channel to a slowly activating Ca^{2+}-activated potassium channel or delayed facilitation of coupled L-type Ca^{2+} channels dictating the time course. Each proposal has its merits, but all are hindered by the fact that the identity of the Ca^{2+}-activated potassium channel that underlies the slow apamin-insensitive AHP is unknown.

Coupling of Ca^{2+} Channel Subtypes to Activation of Apamin-Insensitive AHPs

$Ca_V1.2$ and -1.3 L-type Ca^{2+} channels Dihydropyridine L-type Ca^{2+}channel antagonists significantly block the slow apamin-insensitive AHP in hippocampal pyramidal neurons. This is shown in Figure 2, in which the selective effects of the L-type channel antagonists are seen together with the lack of effect of blockers of other Ca^{2+} channel subtypes. The block of the slow AHP in hippocampal CA1 neurons in a slice preparation by the dihydropyridine L-type Ca^{2+} channel antagonist was incomplete (Figure 2), but the degree of block was very similar to the partial sensitivity to these channel antagonists observed for cloned $Ca_V1.3$ L-type current. These data suggest that L-type Ca^{2+} channels are functionally coupled to activation of the Ca^{2+}-dependent potassium channel that underlies the slow apamin-insensitive AHP and that the L-type channel is $Ca_V1.3$.

$Ca_V2.2$ N- and $Ca_V2.1$ P/Q-type Ca^{2+} channels The apamin-insensitive AHP in neocortical pyramidal neurons was blocked completely by the combination of ω-conotoxin GVIA and ω-agatoxin IVA,

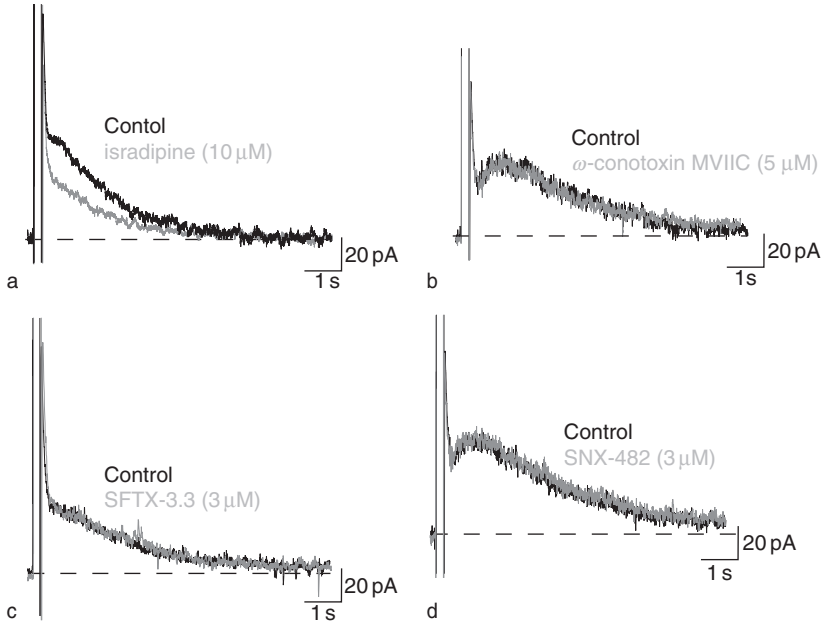

Figure 2 Selective coupling of Ca^{2+} channel subtype activation and generation of I_{sAHP}. (a) I_{sAHP} following a step depolarization to +10 mV (100 ms duration) from a holding potential of −50 mV in the absence and presence of isradipine (10 μM; gray trace). The slow current was partially blocked by the L-type Ca^{2+} channel blocker. (b–d) No effects on I_{sAHP} were observed for toxin blockers for N- and P/Q-type (ω-conotoxin MVIIC; 5 μM), T-type (SFTX-3.3; 3 μM), or R-type (SNX-482; 3 μM) Ca^{2+} channel subtypes.

with each blocker inhibiting approximately 50% of the AHP.

The different proposals concerning the origin of the slow time course of the apamin-insensitive AHP all imply that a single Ca^{2+} channel subtype is responsible for its activation. Different Ca^{2+} channels exhibit various subcellular locations. If a large distance separated Ca^{2+} and Ca^{2+}-activated channel or the AHP was generated by a slowly activating Ca^{2+}-activated potassium channel, it would be expected that the time course of the AHP would change if one Ca^{2+} channel subtype was blocked. No change in time course was reported when Ca^{2+}entry was through one of the two Ca^{2+} channel subtypes that generated the AHP in neocortical neurons. Therefore, it is difficult to reconcile these data with any proposed mechanism of what may underlie the slow time course of this AHP.

Pharmacological Determination of Ca^{2+} Channel Subtypes Underlying Activation of BK ($K_{Ca}1.1$) Channels in Central Neurons

Activation of BK channels hastens action potential repolarization and underlies generation of the fast AHP in some central neurons, such as hippocampal and cortical pyramidal neurons. Therefore, the relationship between Ca^{2+} and BK channels is based mostly on measurement of these parameters.

$Ca_V2.2$ N-type Ca^{2+} Channels

Action potential duration is prolonged in hippocampal slice CA1 neurons after block of N-type Ca^{2+} channels (Figure 3). This effect was selective for N-type Ca^{2+} channels, with no effect of L-type Ca^{2+} channel antagonists. Pharmacological block of either channel subtype has effects on synaptic release of glutamate in the hippocampus, which is in accord with N-type Ca^{2+} and BK channels being present on synaptic terminals.

$Ca_V1.2/1.3$ L- and $Ca_V2.2$ N-type Ca^{2+} Channels

Ca^{2+} entry through both L- and N-type Ca^{2+} channels underlies the activation of BK channel current in voltage-clamped acutely dissociated neocortical pyramidal neurons. In addition, it has been demonstrated that block of BK channels prolongs action potential repolarization. However, an effect of Ca^{2+} channel blockers on action potential duration has not been observed. Some suggestion of selectivity has been reported, with Ca^{2+} entry through N-type channels activating a transient BK channel current, whereas L-type Ca^{2+} channel activation generates both inactivating and sustained BK channel currents.

$Ca_V2.1$ P/Q-type Ca^{2+} Channels

A loose coupling between P/Q-type Ca^{2+} and BK channels has been reported for cerebellar Purkinje cells. BK channel activation was only observed after

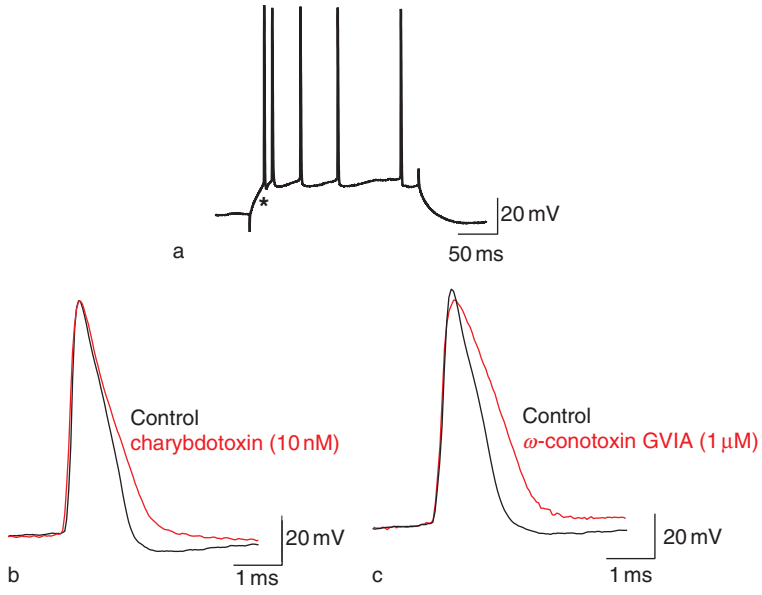

Figure 3 Selective block of N-type Ca^{2+} channels prolongs action potential duration. (a) An action potential train of five spikes in a CA1 pyramidal neuron from a hippocampal slice evoked by a 200 ms depolarizing current injection. The asterisk indicates the fast AHP. (b) Example action potentials illustrating the slowing of action potential repolarization and block of the fast AHP after addition of the BK channel blocker charybdotoxin (10 nM). (c) Block of N-type Ca^{2+} channels by ω-conotoxin GVIA (1 μM) slowed the lower half of action potential repolarization and abolished the fast AHP, like that observed with direct block of BK channels by charybdotoxin.

increasing single Ca^{2+} channel current through P/Q-type channels by increasing extracellular Ca^{2+}. Interestingly, although this coupling suggests distance between Ca^{2+} and BK channels, the effect was selective for P/Q-type channels.

Direct Measurements of Channel Association in Brain

All the presented examples rely on the effect of Ca^{2+} channel blockers on the Ca^{2+}-dependent hyperpolarizations (or outward current) that are attributable to activation of the Ca^{2+}-activated channels, BK ($K_{Ca}1.1$) and SK ($K_{Ca}2$). As mentioned previously, this approach fails to distinguish between coupling arising because of a common subcellular location and that arising from a specific interaction between channel subtypes. A number of studies have attempted to identify the direct association of channel subtypes. These studies have used a range of electrophysiological and biochemical techniques to resolve specific interactions between Ca^{2+} and Ca^{2+}-activated channels.

Ca^{2+} Channels and SK ($K_{Ca}2$) Channels

Single channel recordings from acutely dissociated hippocampal CA1 pyramidal neurons showed a selective coupling between Ca^{2+} and Ca^{2+}-activated channels, with L-type Ca^{2+} channels selectively activating SK channels. This coupling exhibited a delay between the closing of the L-type channel and the opening of the SK channel. This delay and estimated open probability of the SK channel opening suggested that the SK channel experienced approximately 1 μM Ca^{2+} during activation. This in turn suggested that a distance of 50–150 nm separated the channel subtypes. Therefore, it would not be expected that the two channel subtypes be directly associated, but it is more likely that they are assembled within a complex that forms a bridge between the two channels. Interestingly, this has been reported in cardiac cells, in which $Ca_V1.3$ L-type Ca^{2+} and $K_{Ca}2.2$ channels are associated by a common interaction with the actin-binding protein α-actinin 2.

Unfortunately, there are few immunovisualization data available to confirm a co-localization of Ca^{2+} and SK channels in brain. For example, the proposed functional coupling between P/Q-type Ca^{2+} channels and an apamin-sensitive SK channel in neocortical neurons is not borne out by immunohistochemical localization data. $Ca_V2.1$ P/Q-type channels are located in the dendrites of layer V, whereas the only apamin-sensitive SK channel present ($K_{Ca}2.2$ (SK2)) is in the soma of layer V neurons. In contrast, $Ca_V2.1$ and $K_{Ca}2.3$ (SK3) channel subunits have been shown to be on mossy fiber terminals innervating hippocampal CA3 pyramidal neurons, although no functional data are available to confirm functional coupling between the two channel subtypes. The identity of the SK channel underlying the slow AHP in hippocampal and cortical neurons is unknown. However, the only candidate in the past has been the $K_{Ca}2.1$ or SK1 channel. Co-localization of $Ca_V1.3$ L-type and $K_{Ca}2.1$ (SK1) channels has been observed in the soma and proximal dendrites of hippocampal neurons. This co-localization at the light level is consistent with the pharmacological block of L-type Ca^{2+} channels inhibiting the slow AHP in hippocampal neurons.

Ca^{2+} Channels and BK ($K_{Ca}1.1$) Channels

There is reasonable correlation between examples of functional coupling of particular Ca^{2+} channel subtypes to the activation of BK channels and immunolocalization of channel subunits.

$Ca_V2.2$ N-type Ca^{2+} channels and BK channels A direct functional coupling between N-type Ca^{2+} and BK channels was observed using single channel recording from acutely dissociated hippocampal CA1 pyramidal neurons. This coupling is supported by immunohistochemical data that show that N-type Ca^{2+} channels and BK channels are both predominantly located in the dendrites of CA1 hippocampal neurons. Immunovisualization of both N-type and BK channels in cortex has shown both channels to be present in outer neocortical layers II and III.

$Ca_V1.2$ and -1.3 L-type Ca^{2+} channels and BK channels There are few immunohistochemical data to suggest that L-type and BK channels reside in the same subcellular compartment within the hippocampus. However, both $Ca_V1.2$ and BK channels are present in all cortical layers, which supports the pharmacological data from neocortical cells of a functional coupling of L-type Ca^{2+} and BK channels. The confirmation of identical subcellular locations for each channel subtype is unavailable.

$Ca_V2.1$ P/Q-type Ca^{2+} channels and BK channels There is no evidence that the reported loose functional coupling between these two channel subtypes is reflected in channel co-localization at the light level, with BK channels shown to be present in the soma of cerebellar Purkinje cells and the $Ca_V2.1$ (P/Q-type) subunit on terminals. However, both channel subtypes are present on terminals within the cerebellum, which supports the finding that both channel types affect transmitter release.

Biochemical determination from rat brain of interaction between channel subunits has shown that $Ca_V1.2$, -1.3, -2.1, and -2.2 subunits can all be immunoprecipitated with $K_{Ca}1.1$ (BK) channels. This undoubtedly supports the reported examples of functional coupling. This type of data indicates that Ca^{2+} and BK channels are closely associated but falls short of demonstrating that the interactions are direct. However, experiments using the intracellular application of the fast Ca^{2+} chelator BAPTA have indicated that $Ca_V2.2$ and $K_{Ca}1.1$ channels are within 30 nm of each other. Considering that the diameter of an α subunit of a voltage-dependent potassium channel (Kv1.2) has been estimated from crystallization data to be approximately 10 nm, it is likely that Ca^{2+} and BK channels directly associate through their α subunits. This makes mechanistic sense, with rapid coupling between these channels underlying the repolarization of the hippocampal action potential.

Conclusion

In the approximately 20 years since the discovery of the slow AHP in hippocampal neurons, our understanding of the relationship between Ca^{2+} entry and the activation of Ca^{2+}-activated potassium channels has become more sophisticated. No longer do we assume that any subcellular location of Ca^{2+} entry is sufficient for the activation of either BK or SK channels. This is entirely consistent with the cytotoxic effect of accumulating a high concentration of intracellular Ca^{2+}. Now, we can consider that localized Ca^{2+} entry has its effect within nano- or microdomains and is buffered or extruded before the cell's bulk experiences a significant accumulation of the divalent ion. The obvious extension to the acceptance of co-localized Ca^{2+} and Ca^{2+}-activated channels is whether the association can be regulated in space and time. If the subcellular location of co-localized channels were to change, the integration of action potentials between the soma and dendrites would be affected. Both soma and dendrites of central neurons possess a complement of Ca^{2+} and Ca^{2+}-activated potassium channels that control firing patterns. This could substantially change if the subcellular co-localization were to be regulated. Finally, it seems likely that the number of co-localized channels can be regulated over time or with the onset of disease states. The amplitude of the slow AHP increases with age, which correlates with a deficit in learning a simple associative task. L-type Ca^{2+} channels increase in number with age, and the increase in the slow AHP likely results from an increased association of L-type and SK channels. It is possible that targeting of co-localization of L-type Ca^{2+} and SK channels to reduce the age-evoked increase in the slow AHP would produce improvements in cognition that may be beneficial in certain disease states that exhibit symptoms of dementia.

See also: Calcium Channels; Calcium Channel Subtypes Involved in Neurotransmitter Release; Calcium Channels and SNARE Proteins; Dendritic Signal Integration; Ion Channel Localization in Cell Bodies and Dendrites; Neuromodulation of Calcium Channels; Potassium Channel Regulation; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Calcium Channels; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Berkefeld H, Sailer CA, Bildl W, et al. (2006) BK_{Ca}–Ca_V channel complexes mediate rapid and localized Ca^{2+}-activated K^+ signalling. *Science* 314: 615–620.

Catterall WA, Perez-Reyes E, Snutch TP, and Striessnig J (2005) International Union of Pharmacology. XLVIII: Nomenclature and structure–function relationships of voltage-gated calcium channels. *Pharmacological Reviews* 57: 411–425.

Grunnet M and Kaufmann WA (2004) Coassembly of big conductance Ca^{2+}-activated K^+ channels and L-type voltage-gated Ca^{2+} channels in rat brain. *Journal of Biological Chemistry* 279: 36445–36453.

Hallworth NE, Wilson CJ, and Bevan MD (2003) Apamin-sensitive small conductance calcium-activated potassium channels, through their selective coupling to voltage-gated calcium channels, are critical determinants of the precision, pace, and pattern of action potential generation in rat subthalamic nucleus neurons *in vitro*. *Journal of Neuroscience* 23: 7525–7542.

Kato M, Tanaka N, Usui S, and Sakuma Y (2006) SK channel blocker apamin inhibits slow afterhyperpolarization currents in rat gonadotropin-releasing hormone neurons. *Journal of Physiology* 574: 431–442.

Lima PA and Marrion NV (2007) Mechanisms underlying activation of the slow AHP in rat hippocampal neurons. *Brain Research* 1150: 74–82.

Loane DJ, Lima PA, and Marrion NV (2007) Co-assembly of N-type Ca^{2+} and BK channels underlies functional coupling in rat brain. *Journal of Cell Science* 120: 985–995.

Lu L, Zhang Q, Timofeyev V, et al. (2006) Molecular coupling of a Ca^{2+}-activated K^+ channel to L-type Ca^{2+} channels via α-actinin2. *Circulation Research* 100: 112–120.

Marrion NV and Tavalin SJ (1998) Selective activation of Ca^{2+}-activated K^+ channels by co-localized Ca^{2+} channels in hippocampal neurons. *Nature* 395: 900–905.

Misonou H, Menegola M, Buchwalder L, et al. (2006) Immunolocalization of the Ca^{2+}-activated channel Slo1 in axons and nerve terminals of mammalian brain and cultured neurons. *Journal of Comparative Neurobiology* 496: 289–302.

Pineda JC, Waters RS, and Foehring RC (1998) Specificity in the interaction of HVA Ca^{2+} channel types with Ca^{2+}-dependent AHPs and firing behavior in neocortical pyramidal neurons. *Journal of Neurophysiology* 79: 2522–2534.

Power JM, Wu WW, Sametsky E, Oh MM, and Disterhoft JF (2002) Age-related enhancement of the slow outward

calcium-activated potassium current in hippocampal CA1 pyramidal neurons *in vitro*. *Journal of Neuroscience* 22: 7234–7243.

Salkoff L, Butler A, Ferreira G, Santi C, and Wei A (2006) High-conductance potassium channels of the SLO family. *Nature Reviews Neuroscience* 7: 921–931.

Stocker M (2004) Ca^{2+}-activated K^+ channels: Molecular determinants and function of the SK family. *Nature Reviews Neuroscience* 5: 758–770.

X, Gu XQ, and Haddad GG (2003) Calcium influx via L- and N-type calcium channels activates a transient large-conductance Ca^{2+}-activated K^+ current in mouse neocortical pyramidal neurons. *Journal of Neuroscience* 23: 3639–3648.

Tanabe M, Gähwiler BH, and Gerber U (1998) L-type Ca^{2+} channels mediate the slow Ca^{2+}-dependent afterpolarization current in rat CA3 pyramidal cells *in vitro*. *Journal of Neurophysiology* 80: 2268–2273.

Vergara C, Latorre R, Marrion NV, and Adelman JP (1998) Calcium-activated potassium channels. *Current Opinion in Neurobiology* 8: 321–329.

Williams S, Serafin M, Mühlethaler M, and Bernheim L (1997) Distinct contributions of high- and low-voltage-activated calcium currents to afterhyperpolarizations in cholinergic nucleus basalis neurons of the guinea pig. *Journal of Neuroscience* 17: 7307–7315.

Wolfart J and Roeper J (2002) Selective coupling of T-type calcium channels to SK potassium channels prevents intrinsic bursting in dopaminergic midbrain neurons. *Journal of Neuroscience* 22: 3404–3413.

Womack MD, Chevez C, and Khodakhah K (2004) Calcium-activated potassium channels are selectively coupled to P/Q-type calcium channels in cerebellar Purkinje neurons. *Journal of Neuroscience* 24: 8818–8822.

Calcium Channel Subtypes Involved in Neurotransmitter Release

R W Tsien, Stanford University Medical Center, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Ca^{2+} Channels in Excitation–Response Coupling

Voltage-gated calcium channels (Ca^{2+} channels) are members of a superfamily of voltage-gated ion channels which also includes sodium and potassium channels (Na^+ and K^+ channels, respectively). Expressed in a wide variety of cell types, Ca^{2+} channels transduce changes in membrane potential (depolarization) to increased intracellular [Ca^{2+}], which in turn triggers diverse cell biological processes, including secretion, contraction, migration, proliferation, and transcription. Such 'excitation–response coupling' is exemplified by the link between presynaptic depolarization and the triggering of neurotransmitter release, the main topic of this article. Voltage-gated Ca^{2+} channels offer many advantages for excitation–response coupling. The opening of a voltage-gated Ca^{2+} channel is a powerful mechanism for delivering the second messenger (Ca^{2+}) very quickly. There is no lack of Ca^{2+} in the external milieu, and membrane transport mechanisms operate to keep cytosolic Ca^{2+} low, providing an approximately 10 000-fold concentration gradient for supporting a large net influx. Because basal levels of free Ca^{2+} concentration are low (typically 10^{-7} M), influx can cause a rapid and large percentage change in intracellular Ca^{2+}. Unlike Ca^{2+} channels in membranes that surround intracellular Ca^{2+} stores, Ca^{2+} channels in the plasma membrane have direct access to the transmembrane voltage, a global indicator of cellular activity. A rapid, voltage-dependent closing of the channel allows for dissipation of the message by diffusion and strong, rapid local buffering thereby achieving spatiotemporally precise signaling.

To capitalize on all of these advantages, Ca^{2+} channels must satisfy a number of structural and biochemical requirements. First, they should be highly voltage dependent, making them quick to both open and close in response to changes in membrane potential (e.g., action potentials and excitatory postsynaptic potentials). Second, they must be highly permeable to Ca^{2+} but not to other ions, such as Na^+ or K^+. Calcium channels must also be properly localized near the relevant targets of Ca^{2+} regulation, allowing for local increases in Ca^{2+} concentration. Finally, they must be subject to modulation and diverse enough to allow particular classes of calcium channels to be selectively regulated. All of these requirements seem to have been fulfilled by the Ca^{2+} channels found in excitable cells in general and those in nerve terminals in particular. Given the wide range of downstream effects mediated by Ca^{2+} entry, it is perhaps not surprising that a rich diversity of Ca^{2+} channels has evolved, with individual channel types specialized for, but not exclusively limited to, specific biological roles.

Steps in Coupling Excitation to Neurotransmitter Release

Excitation–secretion coupling at typical glutamatergic nerve terminals can be described by the following steps (Figure 1, top):

1. Action potential in axon leads to electrotonic depolarization of axonal bouton.
2. Depolarization causes opening of Na^+ channels and action potential in bouton.
3. Action potential depolarization causes opening of voltage-gated Ca^{2+} channels.
4. K channels open, repolarizing voltage and increasing the driving force for Ca^{2+} to pass through Ca^{2+} channels.
5. Ca^{2+} rises sharply in a microdomain around the mouth of the Ca^{2+} channels.
6. Ca^{2+} binds to Ca^{2+} sensor for vesicle fusion, synaptotagmin triggering vesicle to fuse to surface membrane.
7. Fusion pore opens, providing passage between one or more vesicles containing the neurotransmitter glutamate and the synaptic cleft.
8. Neurotransmitter (e.g., glutamate) passes through fusion pore and diffuses in the cleft.
9. Neurotransmitter binds to multiple sites on postsynaptic receptors (e.g., AMPA receptors), causing their channels to open.
10. Excitatory postsynaptic current ensues.

This provides a general cell physiological framework for understanding the role of specific types of Ca^{2+} channels. The process of neurosecretion relies on two distinct cycles – the classical Hodgkin cycle for nerve excitation (Figure 1, bottom left) and the Heuser–Reese cycle for the clathrin-mediated retrieval of vesicle membrane components (Figure 1, bottom right). Opening of gates in the Hodgkin cycle controls Ca^{2+} entry via Ca^{2+} channels, which in turn regulates vesicle priming and fusion and possibly endocytosis as well.

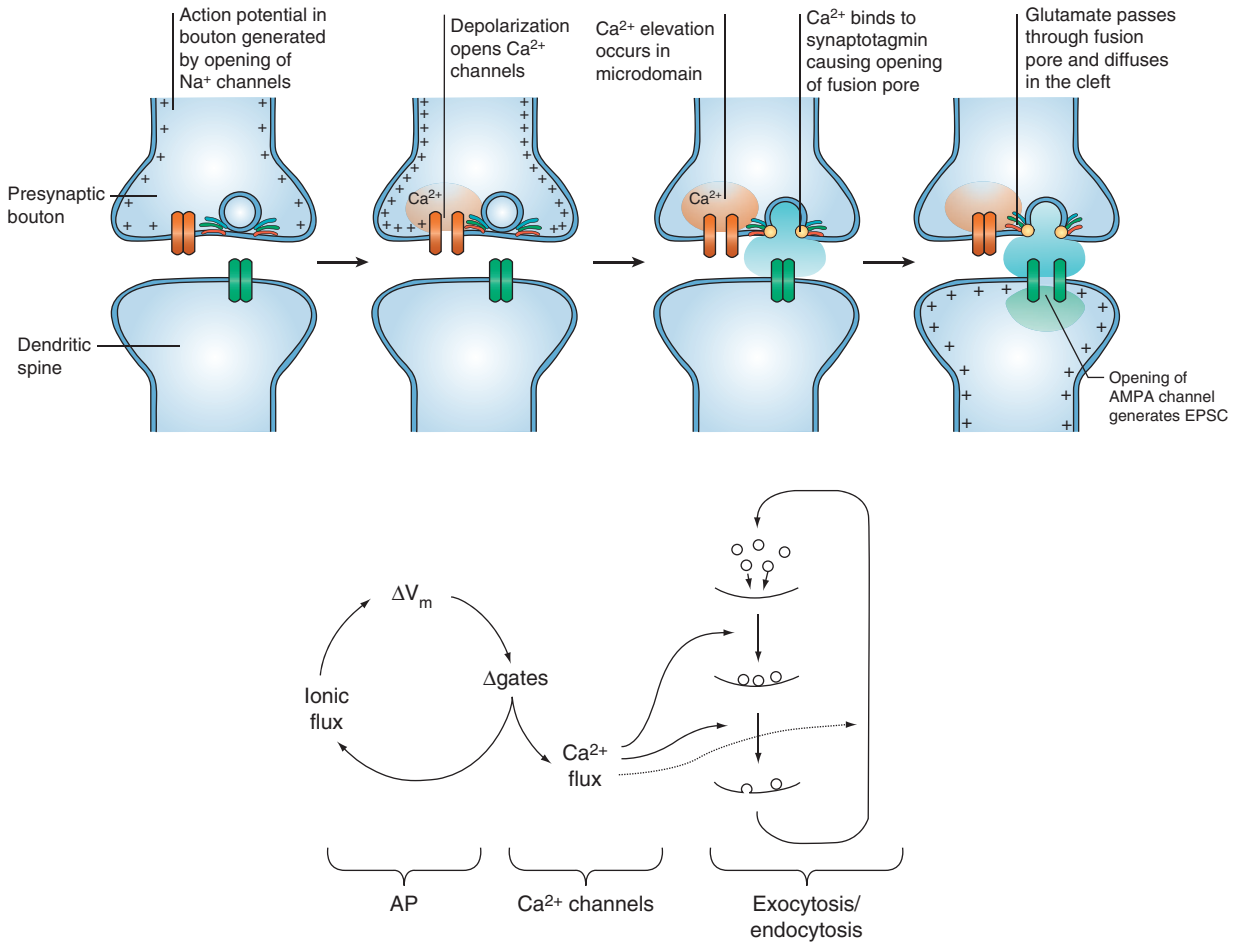

Figure 1 (Top) Schematic showing steps in the process of chemical synaptic transmission, exemplified by an excitatory synapse with postsynaptic receptors for the fast neurotransmitter glutamate. (Bottom) Ca^{2+} forms the link between the classical Hodgkin cycle of excitability and the synaptic vesicle cycle of the kind envisioned by Heuser and Reese. AP, action potential at the presynaptic terminal; V_m, membrane potential. (Top) Reproduced from Lisman JE, Raghavachari S, and Tsien RW (2007) The sequence of events that underlie quantal transmission at central glutamatergic synapses. *Nature Reviews Neuroscience* 8: 597–609, with permission. (Bottom) Reproduced from Tsien RW and Barrett CF (2004) A brief history of calcium channel discovery. In: Zamponi G (ed.) *Voltage-Gated Calcium Channels*. New York: Kluwer/Plenum, with permission.

Discovery of Multiple Ca^{2+} Channel Types

Ca^{2+} channel types were initially classified according to their voltage threshold for activation through the work of Susumu Hagiwara and others. High-voltage-activated (HVA) channels require strong depolarization, whereas low-voltage-activated (LVA) channels activate with relatively little change in membrane potential. This distinction was reinforced by the finding of Armstrong and colleagues that LVA channels deactivate (shut off) much more slowly with hyperpolarization than HVA channels. Beginning in 1985 and for approximately the next 10 years, understanding of the multiple types of Ca^{2+} channels was significantly advanced by analysis of their biophysical, pharmacological, and molecular properties in combination. It is now generally agreed that the HVA channels most important for mammalian neurotransmission (N-, P/Q-, R-type) are different from the HVA channels that dominate excitation–contraction or excitation–transcription coupling (L-type) or the LVA channels that help govern pacemaker activity (T-type).

The presynaptic Ca^{2+} channels are exemplified by N-type channels, first discovered in recordings from cell bodies of sensory neurons and later found to support neurotransmitter release from sympathetic neurons or cholinergic nerve terminals in the ciliary ganglion. P/Q- and R-type channels were discovered later and also contribute to varying degrees to neurotransmission at various central nervous system (CNS) synapses, including large and relatively accessible synapses discussed later. P/Q-, N-, and R-type channels are generated by pore-forming subunits known as $Ca_V2.1$, $Ca_V2.2$, and $Ca_V2.3$, respectively; each is

the specific target of a blocking toxin (ω-AgaIVA, ω-CTx-GVIA, and SNX-482, respectively). Auxiliary subunits (α_2-δ and β) contribute to proper cellular trafficking and function of the pore-forming α_1 subunits.

The single-letter terminology for Ca^{2+} channel types (L-, N-, and T-type) reflects initial distinctions based on patch clamp recordings of unitary channel activity. The observed pattern of unitary channel properties – their slope conductances and activation and inactivation properties – did not fit neatly into the LVA/HVA categorization of Hagiwara. Closer examination of whole cell and single channel recordings made it clear that three channel types could be distinguished, then called T-type (LVA) and N- and L-type (both HVA). The L-type channels were so named because they had a large unitary conductance to Ba^{2+} and supported a long-lasting Ba^{2+} current (different properties were found with Ca^{2+} as the charge carrier, but the Ba^{2+} currents provided the most distinctive channel profiles). Importantly, the L-type channel activity was similar to that found in heart cells and could be distinguished by a profound enhancement with the dihydropyridine drug Bay K8644. The T-type channels generated tiny unitary Ba^{2+} currents, gave rise to a transient average current, showed characteristically slow deactivation following sudden repolarization, and were also found in heart cells. Finally, the N-type channels were largely specific to neurons, had an intermediate conductance to Ba^{2+}, and although they required negative holding potentials to be available for opening, they were activated at high voltages, indicating that they were neither T- nor L-type. The diverse nature of the Ca^{2+} channel, advanced on the basis of patch clamp recordings, has received ample corroboration from molecular biology. Sequence analysis has confirmed and extended the original tripartite classification (Figure 2). The identification of N-type channels presaged the discovery of the Ca_V2 subfamily of α_1 subunits, now known to be critical for neurotransmission.

Further experiments in the early 1990s resulted in the identification of two additional members of the subfamily exemplified by N-type channels. Llinás and colleagues discovered a noninactivating, dihydropyridine-insensitive current in the cell bodies of cerebellar Purkinje neurons, henceforth called P-type. Others demonstrated that P-type channels were exquisitely sensitive to the funnel web spider toxin peptides (e.g., ω-Aga-IVA). Cerebellar granule displayed a prominently inactivating current with somewhat different properties, labeled Q-type. Multiple lines of evidence obtained with antibodies, antisense techniques, and gene knockout have firmly established that P- and Q-type currents both arise from the same pore-forming subunit, now known as $Ca_V2.1$ or α_{1A}. For example, both current types are completely eliminated in cerebellar Purkinje and granule neurons of α_{1A}−/−mice. The original distinctions between the P and Q components have been mostly accounted for by splice variation in the α_{1A} gene, hence the currently accepted terminology of 'P/Q-type' channel.

The actions and specificity of available toxins and pharmacological compounds are sufficient to allow a separation of current components based primarily on pharmacology. A combination of nimodipine (L-type-specific), ω-conotoxin-GVIA (N-type-specific), and ω-Aga-IVA (P/Q-type-specific) can be used to tease apart the various Ca^{2+} current types in cerebellar granule neurons. Application of inhibitors

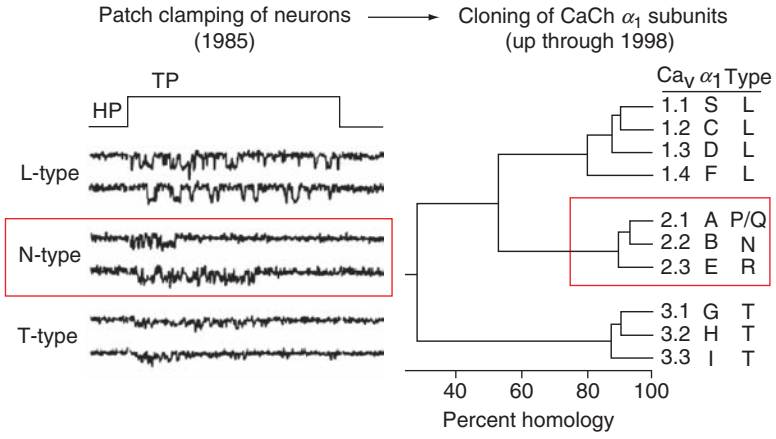

Figure 2 Agreement between channel biophysics studied by patch clamping neurons and molecular cloning of α_1 subunits in tripartite Ca^{2+} channel classification. (Left) Distinctions between three types of Ca^{2+} channels based on recordings of unitary channel activity. N-type channel activity was separated from L- or T-type activity on the basis of its requirement for strongly negative holding potential (HP) and strongly depolarized test potential (TP), as well as its unresponsiveness to dihydropyridines such as Bay K8644 (not shown). (Right) Ensuing molecular cloning revealed three major subfamilies of pore-forming α_1 subunit.

of the three known channel types revealed a fourth component with kinetic and pharmacological properties distinct from L-, N-, and P/Q-type channels. Because this residual current was resistant to the known inhibitors, it was tentatively named R-type. This current showed unique properties, including very rapid decay and an unusual sensitivity to block by Ni^{2+}. Various approaches indicate that R-type current is supported by channels encoded by $Ca_V2.3$. In several neuronal systems, R-type current can be blocked by a specific inhibitor – a spider toxin peptide known as SNX-482.

In General, Multiple Types of Ca^{2+} Channels Contribute to Central Nervous System Neurotransmission

Ca^{2+} channels from the Ca_V2 subfamily are the primary types responsible for excitation–secretion coupling at fast excitatory and inhibitory synapses in the brain. At most of these CNS synapses, the neurotransmitter release is accomplished through synergistic actions of more than one type of channel, most prominently P/Q- and N-type channels. This seems to be the case for typical, approximately 1 μm CNS nerve terminals as well as large presynaptic structures (e.g., the calyx of Held of the auditory brain stem and the mossy fiber synapse of the hippocampus). R-type channels also participate in triggering exocytosis in those instances in which its potential contribution has been carefully explored. Thus, all three members of the Ca_V2 subfamily may play a significant role in fast central neurotransmission. The biophysical properties of these channels (e.g., their rapid activation upon strong depolarization and brisk deactivation following repolarization) are well suited to responding to typical presynaptic action potentials, which reach strongly positive overshoots but have a brief duration on the order of 1 ms.

Intimate Relationship between Ca^{2+} Entry and Neurotransmitter Release

Only Ca^{2+} Channels Close to the Release Machinery Matter

As Ca^{2+} enters, it can act rapidly because only Ca^{2+} elevation near the Ca^{2+} channels is required, rather than a change in the bulk cytoplasm. This is because the Ca^{2+} sensors that trigger vesicle release are strategically located within a 'microdomain' near the Ca^{2+} channels. The first support for this concept came from work on the giant synapse of the squid and indicated that Ca^{2+} could rise to 200 μM in the microdomain. An important advance occurred with the development of structural methods, including electron tomography, that helped localize the intramembranous particles thought to be Ca^{2+} channels. At the neuromuscular junction, it has been estimated that vesicles are positioned within 20 nm of the Ca^{2+} channels – close enough that the vesicles may even hinder the diffusion of Ca^{2+} away from the channel.

Transmitter Release is Triggered by Ca^{2+} Entry, whereas Depolarization *per se* Exerts No Direct Effect

The effects of a rise in presynaptic Ca^{2+} and membrane depolarization have been dissected apart by using light-induced release of caged Ca^{2+} to drive vesicle fusion. In these circumstances, strong depolarizations comparable to presynaptic action potentials failed to increase synaptic responses. This excludes the idea that neurotransmission is jointly dependent on both $[Ca^{2+}]$ and presynaptic membrane potential, although it remains possible that presynaptic membrane voltage can have important modulatory effects, for example, on signaling by G-proteins.

The Ca^{2+} Sensor for Rapid Vesicle Release Is Synaptotagmin

There is substantial evidence that the Ca^{2+} sensor for fast neurotransmission is a Ca^{2+}-binding protein known generically as synaptotagmin. Various isotypes of synaptotagmin, most generally synaptotagmin 1 but also synaptotagmins 2 and 9, can act as a trigger for rapid transmission, as indicated by experiments using knockout mice. There are two main theories about how synaptotagmin works: by unleashing the SNARE machinery by removal of a tonic inhibitory brake or by adding a fusion-promoting influence, derived from Ca^{2+}/syt 1 binding to membrane phospholipids, to that already provided by preformed SNARE complexes.

Transmitter Release is Related to Ca^{2+} Entry by a Steep Power-Law Function

In a classical study on the frog neuromuscular junction, Dodge and Rahamimoff showed that neurotransmitter release was a steep function of external Ca^{2+} concentration and, by extension, Ca^{2+} influx. Later investigators extended the analysis to direct measurements of presynaptic Ca^{2+} currents, and cytosolic Ca^{2+} changes, converging on the idea that the Ca^{2+} sensor for secretion behaves as if it were cooperatively activated by the binding of four or five Ca^{2+} ions. The basis of this cooperativity remains unclear. Synaptotagmin has five Ca^{2+} binding sites, suggesting that the cooperative properties of single synaptotagmin molecules might be sufficient to

account for the Ca^{2+} cooperativity of release. Consistent with this, mutations in synaptotagmin that affect Ca^{2+} binding reduce this cooperativity. However, additional factors might influence the observed Ca^{2+} cooperativity. Vesicles contain several synaptotagmin molecules, possibly working together to produce release; cooperativity might be influenced by other molecules and the combined effects of multiple SNARE complexes.

Multiple Ca^{2+} Channels (and Channel Types) Provide Input to a Common Ca^{2+} Sensor

Pharmacological studies are consistent with regard to a scenario in which a mixed population of Ca^{2+} channels coexist at single release sites and contribute jointly to the local Ca^{2+} transient that operates on the Ca^{2+} sensor to trigger vesicular fusion. The convergence of Ca^{2+} delivered by different types of channels can be manifested as synergism or redundancy, depending on the experimental conditions (Figure 3). With normal action potentials as the stimulus, block of either P/Q- or N-type (but not L-type) Ca^{2+} channels can cause a significant diminution of transmitter release (Figure 3(a)). In the illustrated example, the percentage inhibition of neurotransmission for block of P/Q-type channels (~90%) and for block of N-type channels (~40%) totals more than 100%. This makes sense because both Ca^{2+} channel types contribute to Ca^{2+} influx at individual release sites and thus provide additive fluxes to take advantage of the steep power law for neurotransmission. Another way to establish that both channel types coexist at individual release sites is to examine the consequences of a depolarization imposed by application of K^+-rich solution, a much stronger stimulus for channel activation than a single action potential. In this case, both types of channels must be blocked to appreciably inhibit transmission (Figure 3(b)). The explanation for this redundancy is simple: Provided that the depolarizing stimulus is strong and sustained, Ca^{2+} influx

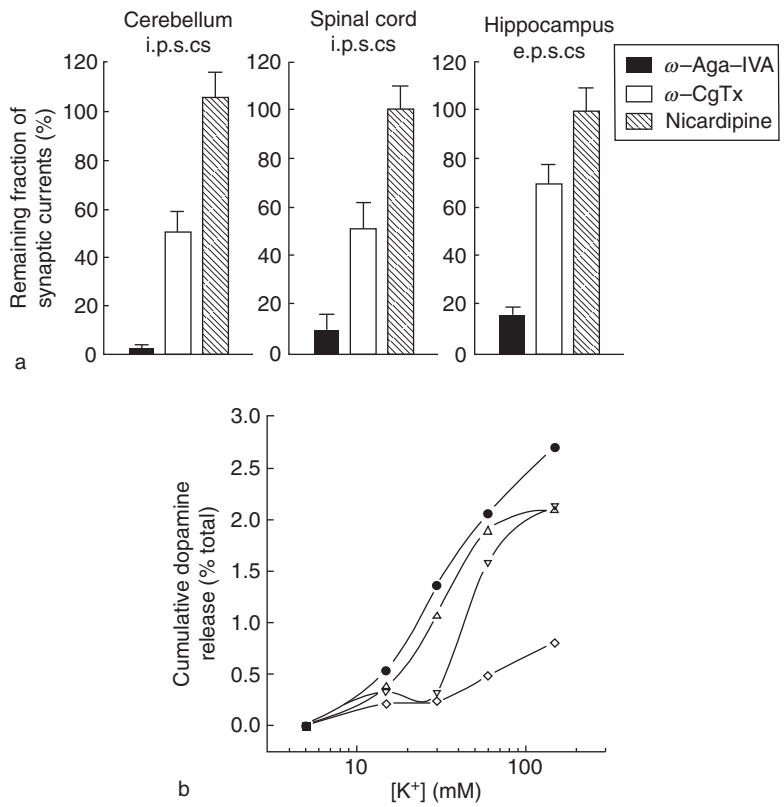

Figure 3 Differing perspectives on transmitter release and the role of N- and P/Q-type Ca^{2+} channels. (a) Contributions of various Ca^{2+} channels at CNS synapses when the stimulus for Ca^{2+} is a single action potential. Shown are the proportions of synaptic currents remaining after blockade of P/Q-type channels (ω-Aga-IVA present), with N-type channels blocked (ω-CgTx present), or with L-type channels blocked (nicardipine). Reproduced from Takahashi T and Momiyama A (1993) Different types of calcium channels mediate central synaptic transmission. *Nature* 366: 156–158, with permission. (b) Effect of blockade of P/Q- or N-type channels, individually or in combination, on dopamine release from rat striatal synaptosomes. Note that with strong depolarization (external potassium concentration >100 mM), both kinds of channels must be blocked in order to strongly inhibit neurosecretion. Circles, no channel inhibition; triangles, N-type channels blocked; upside down triangles, P/Q-type channels blocked; diamonds, both N- and P/Q-type channels blocked. Reproduced from Turner TJ, Adams ME, and Dunlap K (1993) Multiple Ca^{2+} channel types coexist to regulate synaptosomal neurotransmitter release. *Proceedings of the National Academy of Sciences of the United States of America* 90(20): 9518–9522, with permission.

through either type of channel is sufficient to drive secretion, and the other type can be regarded as 'spare channels.' The synergistic or redundant actions of multiple channel types can be observed in the same experiment (Figure 4). The switch between cooperative and redundant behavior can be accomplished simply by prolongation of presynaptic action potentials with the K^+ channel blocker 4-aminopyridine. Prolongation of the action potential causes more activation of each type of Ca^{2+} channel, favoring its self-sufficiency.

Exceptions to Generalizations about Multiple Ca^{2+} Channel Types

The detailed composition of Ca^{2+} channel subtypes that mediate transmission can vary widely among individual nerve terminals.

Not always Multiple Channel Types

Although neurosecretion is supported by more than one type of channel at most CNS synapses, at some synapses transmission may be heavily dominated by a single channel type. Examples include some preparations historically important for discovery of basic principles of neurotransmission such as mammalian neuromuscular junction and squid giant synapse, in which P- or P/Q-type channels are overwhelmingly prevalent. At some hippocampal inhibitory synapses, GABA release is mediated entirely by P/Q-type or by N-type Ca^{2+} channels. Neurotransmission at sympathetic neuron–effector junctions is governed almost completely by N-type channels. Differences in the reliance on various Ca_V2 family members take on additional significance when synaptic circuits are the target of pharmacological and possibly therapeutic intervention. For example, in the dorsal horn of the spinal cord, the main component in the circuitry of pain transmission/modulation, inhibitory synaptic transmission is dominated by P/Q-type Ca^{2+} channels; in contrast, N-type channels are more closely associated with glutamate release from the peptidergic primary sensory neurons that convey nociceptive information to the dorsal horn.

Not always Ca_V2 Channels

It is generally believed that Ca^{2+} channels of the Ca_V2 class are specialized to support transmitter release. This is true, but it is worth pointing out that in some cases, Ca_V1 and even Ca_V3 channels can also participate. For example, ribbon synapses in the retina and cochlea rely on $Ca_V1.3$ channels to trigger transmitter

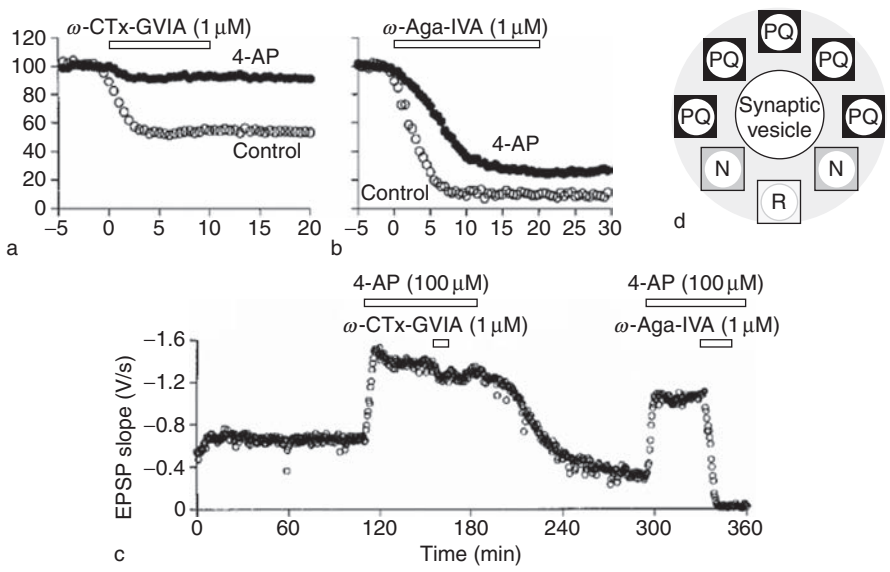

Figure 4 Concerted or redundant actions of P/Q- and N-type channels, depending on strength of depolarizing stimulus. (a) Dependence of synaptic transmission on N-type Ca^{2+} channels changes markedly upon prolongation of the presynaptic action potential with 4-AP. Control graph (open circles) shows approximately 50% inhibition of basal transmission by ω-CTx-GVIA. In 4-AP (solid circles), the effect of N-type channel blockade is barely evident. (b) 4-AP also reduces the degree of inhibition produced by P/Q-type channel inhibition with ω-Aga-IVA. (c) Example of an experiment in which simultaneous blockade of N-type and P/Q-type Ca^{2+} channels with ω-CTx-GVIA and ω-Aga-IVA abolished transmission completely, even in the presence of 4-AP. (d) Schematic representation of a release site surrounded by multiple types of Ca^{2+} channels. (a–c) Reproduced from Wheeler DB, Randall A, and Tsien RW (1996) Changes in action potential duration alter reliance of excitatory synaptic transmission on multiple types of Ca^{2+} channels in rat hippocampus. *Journal of Neuroscience* 16: 2226–2237, with permission. (d) Reproduced from Cao Y, Piedras-Rentería ES, Smith GB, Chen G, Harata NC, and Tsien RW (2004) Presynaptic Ca^{2+} channels compete for channel type-preferring slots in altered neurotransmission arising from Ca^{2+} channelopathy. *Neuron* 43: 387–400.

release. Furthermore, Ca_V2 channels can perform important neuronal functions distinct from excitation–secretion coupling, such as excitation–transcription coupling.

Not always the Same Type of Ca^{2+} Channel throughout Development

At some synapses, N-type channels develop early but are gradually supplanted by late-blooming P/Q-type channels. Mammalian development might be echoing a phylogenetic progression seen at the neuromuscular junction, where lizards and frogs use L- and N-type channels, respectively, but mammals deploy P/Q-type channels.

What Are Multiple Ca^{2+} Channels Good for?

Why are there different kinds of presynaptic Ca channels? Almost certainly it is not because of evolutionary pressure for different biophysical properties. Indeed, basic features such as extremely high Ca^{2+} selectivity, high open channel flux rate, and steeply voltage-dependent channel gating are very similar for P/Q-, N-, and R-type currents. In all cases, selectivity and permeation are conferred by a pore that is continually occupied by at least one Ca^{2+} ion, which prevents more abundant Na^+ ions from rushing through. Despite overall similarities, there are substantial differences in channel regulation, and this is likely to be the main reason for channel diversity (Table 1). For example, P/Q- and N-type channels appear to differ in their susceptibility to modulation by G-proteins. N-type channels show a greater degree of modulation. On the other hand, P/Q-type channels are more susceptible to progressive increases with use, a phenomenon known as Ca^{2+} channel facilitation. Because of the steep power-law relationship between Ca^{2+} entry and secretion, even small increases in P/Q Ca^{2+} entry can have a profound effect on the strength of transmission. The presynaptic Ca^{2+} channels constitute a key convergence point for regulation of neurosecretion, either by neurohumoral modulation or by activity-dependent regulation. Much is known about the structural motifs in α_1 subunits that support various forms of regulation, for example, by G-proteins and protein kinases.

Table 1 Differences between P/Q- and N-type channels with regard to their properties and contributions to synaptic transmission

	P/Q-type channels	*N-type channels*
Specific blocker	ω-AgaIVA	ω-CTx-GVIA, ω-CTx-MVIIA = *Prialt*
Genetic diseases	Migraine, ataxia, epilepsy	No channelopathies found
Development	Late blooming	Early prominence
Voltage gating	Similar	Similar
Modulation by G-protein-coupled receptors	Weaker	Stronger
Facilitation by repeated stimuli	Stronger	Weaker if at all

See also: Calcium Channels; Calcium Channels and SNARE Proteins; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Dendritic Signal Integration; Exocytosis: Ca^{2+}-Sensitivity; Invertebrate Neurohormone GPCRs; Ion Channel Localization in Cell Bodies and Dendrites; Neuromodulation of Calcium Channels; Voltage-Gated Calcium Channels.

Further Reading

Bollmann JH and Sakmann B (2005) Control of synaptic strength and timing by the release-site Ca^{2+} signal. *Nature Neuroscience* 8: 426–434.

Cao Y, Piedras-Rentería ES, Smith GB, Chen G, Harata NC, and Tsien RW (2004) Presynaptic Ca^{2+} channels compete for channel type-preferring slots in altered neurotransmission arising from Ca^{2+} channelopathy. *Neuron* 43: 387–400.

Catterall WA (1996) Molecular properties of sodium and calcium channels. *Journal of Bioenergetics and Biomembranes* 28: 219–230.

Dolphin AC (2006) A short history of voltage-gated calcium channels. *British Journal of Pharmacology* 147(supplement 1): S56–S62.

Dunlap K, Luebke JI, and Turner TJ (1995) Exocytotic Ca^{2+} channels in mammalian central neurons. *Trends in Neurosciences* 18: 89–98.

Ertel EA, Campbell KP, Harpold MM, et al. (2000) Nomenclature of voltage-gated calcium channels. *Neuron* 25(3): 533–535.

Lisman JE, Raghavachari S, and Tsien RW (2007) The sequence of events that underlie quantal transmission at central glutamatergic synapses. *Nature Reviews Neuroscience* 8: 597–609.

Llinás R, Sugimori M, Hillman DE, and Cherksey B (1992) Distribution and functional significance of the P-type, voltage-dependent Ca^{2+} channels in the mammalian central nervous system. *Trends in Neurosciences* 15: 351–355.

Miljanich GP (2004) Ziconotide: Neuronal calcium channel blocker for treating severe chronic pain. *Current Medicinal Chemistry* 11: 3029–3040.

Nowycky MC, Fox AP, and Tsien RW (1985) Three types of neuronal calcium channel with different calcium agonist sensitivity. *Nature* 316: 440–443.

Olivera BM, Miljanich GP, Ramachandran J, and Adams ME (1994) Calcium channel diversity and neurotransmitter release: The ω-conotoxins and ω-agatoxins. *Annual Review of Biochemistry* 63: 823–867.

Piedras-Rentería ES, Barrett CF, Cao Y-Q, and Tsien RW (2007) Voltage-gated calcium channels, calcium signaling and channelopathies. In: Krebs J and Michalak M (eds.) *Calcium: A Matter of Life or Death*, pp. 127–166. Elsevier: New York.

Schneggenburger R and Neher E (2005) Presynaptic calcium and control of vesicle fusion. *Current Opinion in Neurobiology* 15: 266–274.

Takahashi T and Momiyama A (1993) Different types of calcium channels mediate central synaptic transmission. *Nature* 366: 156–158.

Tsien RW and Barrett CF (2004) A brief history of calcium channel discovery. In: Zamponi G (ed.) *Voltage-Gated Calcium Channels*. New York: Kluwer/Plenum.

Turner TJ, Adams ME, and Dunlap K (1993) Multiple Ca^{2+} channel types coexist to regulate synaptosomal neurotransmitter release. *Proceedings of the National Academy of Sciences of the United States of America* 90(20): 9518–9522.

Wheeler DB, Randall A, and Tsien RW (1996) Changes in action potential duration alter reliance of excitatory synaptic transmission on multiple types of Ca^{2+} channels in rat hippocampus. *Journal of Neuroscience* 16: 2226–2237.

Calcium Channels

W A Catterall, University of Washington, Seattle, WA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Voltage-gated calcium channels mediate calcium influx in response to membrane depolarization and regulate intracellular processes such as contraction, secretion, neurotransmission, and gene expression in many different cell types. Their activity is essential to couple electrical signals in the cell surface to physiological events in cells. This article describes their biochemical, molecular, and genetic properties; physiological roles; and pharmacological significance.

Calcium Currents in Excitable Cells

Ionic currents conducted by calcium ions in different cell types have diverse physiological roles and pharmacological properties, and an alphabetical nomenclature has evolved for the distinct classes of calcium currents (Table 1). L-type calcium currents typically require a strong depolarization for activation, have slow voltage-dependent inactivation, and are blocked by the organic L-type calcium channel antagonists, including dihydropyridines, phenylalkylamines, and benzothiazepines. They are the main calcium currents in muscle and endocrine cells, where they initiate contraction and secretion, respectively. L-type currents are prominent in neurons, in which they are important for integration of synaptic inputs in dendrites, regulation of many intracellular events, and initiation of excitation–transcription coupling, which regulates expression of many neural genes. L-type calcium currents also trigger synaptic transmission in specialized ribbon synapses in photoreceptors in the retina and hair cells in the cochlea.

N-type, P/Q-type, and R-type calcium currents also require strong depolarization for activation. They are relatively unaffected by L-type calcium channel antagonist drugs but are blocked by specific polypeptide toxins from snail and spider venoms (Table 1). They are expressed primarily in neurons, in which they initiate neurotransmission at most fast synapses and also mediate calcium entry into cell bodies and dendrites.

T-type calcium currents are activated by weak depolarization and are transient due to rapid voltage-dependent inactivation. They are resistant to both organic antagonists and most snake and spider toxins used to define the N- and P/Q-type calcium currents (Table 1). They are expressed in a wide variety of cell types, in which they are involved in shaping the action potential and controlling patterns of repetitive firing. They have particularly prominent roles in repetitively firing cells such as the pacemaker cells in the sinoatrial node of the heart and the relay neurons in the thalamus that control sleep and wakefulness by rhythmic firing of action potentials.

Calcium Channel Subunits

The calcium channels that have been characterized biochemically are complex proteins composed of four or five distinct subunits, which are encoded by multiple genes (Figure 1). The α_1 subunit of 190–250 kDa is the largest subunit, and it incorporates the conduction pore, the voltage sensors and gating apparatus, and most of the known sites of channel regulation by second messengers, drugs, and toxins. The α_1 subunit is composed of approximately 2000 amino acid residues organized in four homologous domains (I–IV) with six transmembrane segments (S1–S6) in each. The S1–S4 segments serve as the voltage sensor module, whereas transmembrane segments S5 and S6 in each domain and the P loop between them form the pore module.

The α_1 subunits are associated with four distinct auxiliary subunits encoded by three different genes (Figure 1). The intracellular β subunit is a hydrophilic protein of 50–65 kDa. The transmembrane disulfide-linked $\alpha_2\delta$ subunit complex is encoded by a single gene, but the resulting pre-polypeptide is posttranslationally cleaved and disulfide bonded to yield the mature covalently linked α_2 and δ subunits. A transmembrane γ subunit having four transmembrane segments is a component of skeletal muscle calcium channels, and related subunits are expressed in heart and brain. Although these auxiliary subunits modulate the functional properties of the calcium channel complex, the pharmacological and electrophysiological diversity of calcium channels arises primarily from the existence of multiple α_1 subunits.

Calcium Channel Genes

Calcium channel α_1 subunits are encoded by 10 distinct genes in mammals. These calcium channels are named using the chemical symbol of the principal permeating ion (Ca), with the principal physiological regulator (voltage) indicated as a subscript (Ca_v). The numerical identifier corresponds to the Ca_v channel α_1 subunit gene subfamily (1 to 3) and the order of discovery of the α_1 subunit within that subfamily (1 through n).

Table 1 Physiological function and pharmacology of calcium channels

Channel	*Current*	*Localization*	*Specific antagonists*	*Cellular functions*
$Ca_v1.1$	L	Skeletal muscle, transverse tubules	Dihydropyridines, phenylalkylamines, benzothiazepines	Excitation–contraction coupling
$Ca_v1.2$	L	Cardiac myocytes; smooth muscle myocytes; endocrine cells; neuronal cell bodies and proximal dendrites	Dihydropyridines, phenylalkylamines, benzothiazepines	Excitation–contraction coupling, hormone release, regulation of transcription, synaptic integration
$Ca_v1.3$	L	Endocrine cells, neuronal cell bodies and dendrites, cardiac atrial myocytes and pacemaker cells, cochlear hair cells	Dihydropyridines, phenylalkylamines, benzothiazepines	Hormone release, regulation of transcription, cardiac pacemaking, synaptic integration hearing – neurotransmitter release from ribbon synapses in auditory hair cells
$Ca_v1.4$	L	Retinal rod and bipolar cells, lymphocytes	Not established	Neurotransmitter release from ribbon synapses in photoreceptors
$Ca_v2.1$	P/Q	Nerve terminals and dendrites, neuroendocrine cells	ω-Agatoxin IVA	Neurotransmitter release, dendritic Ca^{2+} transients, hormone release
$Ca_v2.2$	N	Nerve terminals and dendrites, neuroendocrine cells	ω-CTx-GVIA	Neurotransmitter release, dendritic Ca^{2+} transients, hormone release
$Ca_v2.3$	R	Neuronal cell bodies and dendrites	SNX-482	Repetitive firing, dendritic calcium transients
$Ca_v3.1$	T	Neuronal cell bodies and dendrites, cardiac and smooth muscle myocytes	None	Pacemaking, repetitive firing
$Ca_v3.2$	T	Neuronal cell bodies and dendrites, cardiac and smooth muscle myocytes	None	Pacemaking, repetitive firing
$Ca_v3.3$	T	Neuronal cell bodies and dendrites	None	Pacemaking, repetitive firing

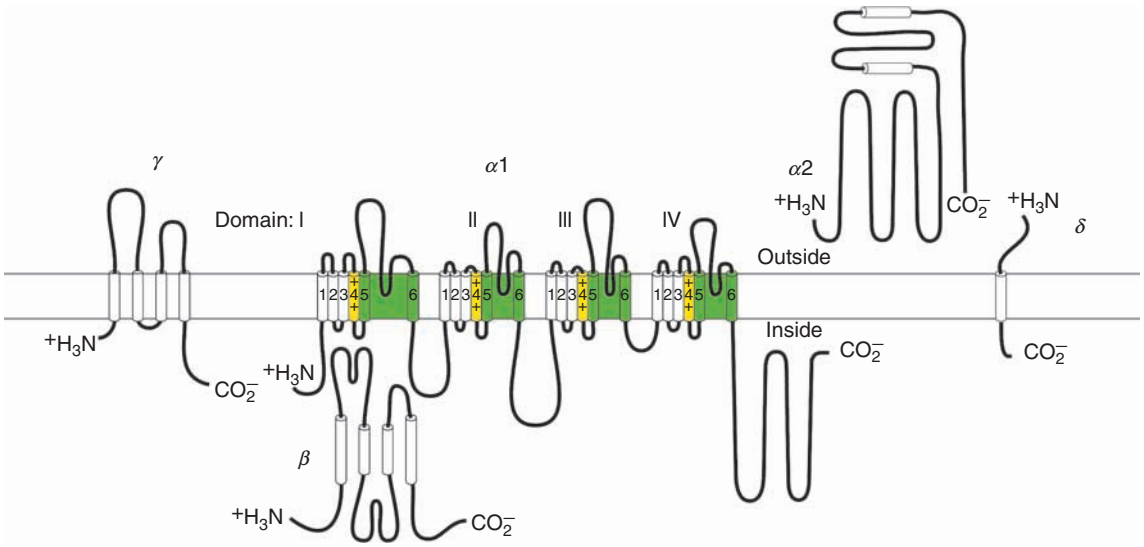

Figure 1 Subunit structure of Ca_v1 channels. The subunit composition and structure of calcium channels purified from skeletal muscle are illustrated. Predicted helices are depicted as cylinders. The lengths of lines correspond approximately to the lengths of the polypeptide segments represented. Pore forming module, green; S4 voltage sensor, yellow.

The Ca_v1 subfamily ($Ca_v1.1$–$Ca_v1.4$) mediates L-type Ca^{2+} currents (Table 1). The Ca_v2 subfamily ($Ca_v2.1$–$Ca_v2.3$) mediates P/Q-type, N-type, and R-type Ca^{2+} currents, respectively (Table 1). The Ca_v3 subfamily ($Ca_v3.1$–$Ca_v3.3$) mediates T-type Ca^{2+} currents. The complete amino acid sequences of these α_1 subunits are more than 70% identical within a subfamily but less than 40% identical among the three subfamilies. Division of calcium channels into these three subfamilies is phylogenetically ancient because representatives of each are found in the *Caenorhabditis elegans* genome. Consequently, the genes for the different α_1 subunits have become widely dispersed in the genome and even the most closely related members of the family are not clustered on single chromosomes in mammals.

Each of the auxiliary subunits of calcium channels is encoded by a small gene family. Four genes encode $\alpha_2\delta$ subunits, four genes encode β subunits, and eight genes encode γ subunits. The specificity of assembly of calcium channel auxiliary subunits with the principal, pore-forming α_1 subunits is not fully determined. However, in neurons, it appears that multiple subtypes of auxiliary subunits can assemble successfully with α_1 subunits. Therefore, it is likely that control of expression and localization of the principal and auxiliary subunits determines the subunit partners present in calcium channel complexes. Although the $Ca_v\gamma_1$ subunit is associated specifically with skeletal muscle $Ca_v1.1$ channels, other $Ca_v\gamma$ subunits interact with other calcium channels, glutamate receptors, and possibly other membrane signaling proteins. Thus, the γ subunits discovered as components of calcium channels apparently have a more widespread role in assembly and cell surface expression of other membrane signaling proteins.

Calcium Channel Structure and Function

Intensive studies of the structure and function of the related pore-forming subunits of sodium, calcium, and potassium channels have led to identification of their principal functional domains. Each domain of the principal α_1 subunits consists of six transmembrane α helices (S1–S6) and a membrane-associated P loop between S5 and S6. The S4 segments of each homologous domain serve as the voltage sensors for activation, moving outward and rotating under the influence of the electric field and initiating a conformational change that opens the pore. The transmembrane pore is formed in the midst of a square array of the four homologous domains. The S5 and S6 segments and the membrane-associated P loop between them form the pore lining of the voltage-gated ion channels. The narrow external pore is lined by the P loop, which contains a glutamate residue in each domain that is required for calcium selectivity. Remarkably, substitution of only three amino acid residues in the pore loops between the S5 and S6 segments in domains II–IV of sodium channels is sufficient to confer Ca^{2+} selectivity. The inner pore is lined by the S6 segments, which form the receptor sites for the pore-blocking calcium antagonist drugs specific for L-type calcium channels. All Ca^{2+} channels share these general structural features, but the amino acid residues that confer high affinity for the organic Ca^{2+} antagonists used in therapy of cardiovascular diseases are present only in the Ca_v1 family of Ca^{2+} channels, which conduct L-type Ca^{2+} currents.

The auxiliary subunits of calcium channels have an important influence on their function. $Ca_v\beta$ subunits greatly enhance cell surface expression of α_1 subunits and shift the kinetics and voltage dependence of activation and inactivation. Most $Ca_v\beta$ subunits accelerate activation and inactivation and shift the voltage dependence of gating to more negative membrane potentials, but the $Ca_v\beta_{2a}$ subunit has the opposite effect in most cases. The $\alpha_2\delta$ subunits also enhance cell surface expression of α_1 subunits, but they have smaller and less-consistent effects on the kinetics and voltage dependence of gating. The γ subunits do not increase cell surface expression of Ca_V channels, and in some cases they reduce it substantially. The functional role of these subunits of calcium channels is the least well defined.

Three-Dimensional Structure of Calcium Channels

The three-dimensional structure of calcium channels is unknown, but it is expected to resemble the structures of the closely related voltage-gated sodium and potassium channels. Low-resolution structural models have been developed from image reconstruction analysis of $Ca_v1.1$ channels purified from skeletal muscle membranes, and some of the structural features have been associated with the α_1 and $\alpha_2\delta$ subunits (**Figure 2(a)**). Further high-resolution structural analysis will be required to confirm these initial structural models.

The three-dimensional structure of the $Ca_v\beta$ subunits has been determined at high resolution by X-ray crystallography (**Figure 2(b)**). These subunits contain conserved SH3 and guanylate kinase domains like the MAGUK family of scaffolding proteins. These two domains are arrayed side by side in the $Ca_v\beta$ subunit (**Figure 2(b)**). The $Ca_v\beta$ subunits bind to a single site in the α_1 subunits (the α interaction domain (AID)), which is located in the first half of the intracellular loop connecting domains I and II. AID forms an α helix that is bound tightly to a groove in the guanylate kinase domain of the $Ca_v\beta$ subunit. This tight, multipoint binding interaction likely sustains the association between calcium channel α_1 and β subunits throughout the lifetime of the calcium channel complex at the cell surface membrane. MAGUK proteins often bind more than one protein partner, so $Ca_v\beta$ subunits may also interact with other intracellular proteins.

Physiological Roles of Calcium Channels

Calcium channels have two distinct classes of functional roles: They contribute to electrical events during action potential generation and conduction at the cell surface, and they conduct calcium into the cell to initiate many intracellular events.

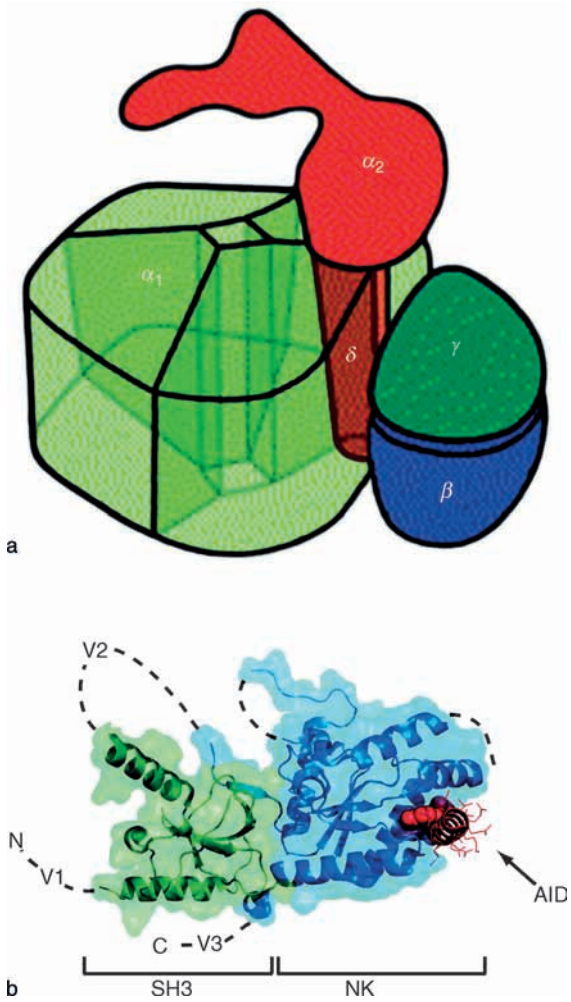

Figure 2 Three-dimensional architecture of calcium channels. (a) Illustration of the skeletal muscle $Ca_v1.1$ channel based on cryo-electron microscopy. This drawing assumes pseudo-fourfold symmetry of the α_1 subunit. The view shows the extracellular side with the α_2 subunit. The α_1, γ, and δ subunits are embedded into the lipid membrane (not shown), which separates the extracellular α_2 subunit from the cytosol. α_2 is anchored via the disulfide-linked δ subunit within the α_1 subunit. The proposed model allows for a tight interaction between α_1 and δ as well as between α_1 and γ. (b) Structure of the $Ca_v\beta$ complex with the α interaction domain (AID). Coordinates are for the $Ca_v\beta_2$a–$Ca_v1.2$ AID complex, with SH3 (green) and NK (blue) domains indicated. V1–V3 show the locations of the three variable domains that are absent from the structure. AID (red) binds to a deep groove in the NK domain. AID residues tyrosine, tryptophan, and isoleucine are shown as CPK. The remaining residues are shown as lines.

Depolarization of neurons by calcium ions entering through voltage-gated calcium channels helps to shape the action potential and has an important role in determining the frequency of firing of action potentials. Depolarization of neurons by activation of sodium channels during the action potential activates Ca_v1 and Ca_v2 channels, and the further depolarization caused by calcium current flowing inward through them broadens the action potential in dendrites, cell bodies, and nerve terminals and modulates the shape of the repolarizing phase of the action potential. Ca_v3 channels conduct T-type calcium currents that activate in the same potential range or even more negatively than sodium channels and thereby influence the threshold and frequency of firing of neurons. These calcium channels are prominent in repetitively firing cells, in which they are activated following repolarization from one action potential and boost the membrane potential to the threshold for activating sodium channels and firing the next action potential. This function of Ca_v3 channels is important in the thalamus, in which repetitive firing of neurons is required for determining sleep and wakefulness, as well as controlling other aspects of cortical function.

In addition to their role in electrical signaling, calcium channels are also the key transducers of electrical signals. They convert depolarization of the cell surface membrane into an intracellular transient of calcium, which serves as a second messenger to initiate many physiological events. In neurons, three of the key physiological events initiated by calcium channels are synaptic integration, synaptic transmission, and gene regulation. All three subfamilies of calcium channels participate in calcium signaling in dendrites and cell bodies and in the integration of synaptic signals. In contrast, Ca_v1 and Ca_v2 channels are specialized for gene regulation and synaptic transmission, respectively.

Calcium entering neurons through Ca_v1 channels is particularly effective in activating expression of neuronal genes in response to electrical activity, in comparison to calcium entering through other permeability pathways. The molecular basis for this preferential effect of Ca_v1 channels on gene expression is not known, but one idea is that proteins bound to the Ca_v1 channels, or even the C-terminus of the Ca_v1 channels, serve as a calcium-regulated molecular signal to control gene expression.

Calcium entering through Ca_v2 channels is primarily responsible for initiating synaptic transmission at fast synapses. $Ca_v2.2$ channels, which conduct N-type calcium current, are most important at synapses formed by neurons of the peripheral nervous system. In contrast, $Ca_v2.1$ channels, which conduct P/Q-type calcium currents, are most important at synapses formed by neurons of the central nervous system. Both $Ca_v2.2$ and $Ca_v2.3$ channels also have a significant role in initiating transmitter release at central synapses. In some cases, including a subset of inhibitory interneurons of the hippocampus, $Ca_v2.2$ channels are predominant in neurotransmitter release. On the other hand, $Ca_v2.3$ channels seem to always play a secondary role at central synapses compared to $Ca_v2.1$ and/or $Ca_v2.2$ channels.

The molecular basis for the prominent role of Ca_v2 channels in neurotransmission is not completely understood, but one factor is their direct interactions with the neurotransmitter release machinery in calcium channel signaling complexes.

Calcium Channel Signaling Complexes

Calcium entering neurons through calcium channels forms a transient microdomain of high calcium concentration. Intracellular proteins that must bind calcium to initiate essential physiological processes must be located near calcium channels in order to receive the calcium signal. In many cases, this close localization is achieved by direct interaction with the intracellular domains of calcium channels, which serve as signal transduction platforms for cytosolic calcium signaling. Similarly, proteins that regulate calcium channels in response to second messenger signaling are also often directly associated with calcium channels.

Ca_v1 channels are regulated by cAMP-dependent protein phosphorylation, initiated by neurotransmitters and hormones acting at many G-protein-coupled receptors. Regulation of these channels by local cAMP concentrations is achieved by binding of cAMP-dependent protein kinase to A-kinase anchoring proteins (AKAPs) that in turn bind directly to calcium channels. AKAP15 binds directly to Ca_v1 channels and tethers protein kinase A (PKA) to the channels through a modified leucine zipper interaction with a site in the C-terminal domain (Figure 3(a)). The C-terminal domain is proteolytically processed (Figure 3(a)), and the distal C-terminal domain containing the site of interaction of AKAP15 and PKA remains bound to the channel and regulates its function. Calcium/calmodulin-dependent protein kinase II binds directly to the proximal C-terminal domain of $Ca_v1.2$ channels and regulates channel function (Figure 3(a)). These direct interactions position the kinases near the sites of phosphorylation that regulate channel activity and thereby enhance speed and specificity of channel regulation.

Ca_v2 channels bind a wide range of signaling proteins (Figure 3(b)). The SNARE proteins involved in docking and release of neurotransmitters bind to a synaptic protein interaction ('synprint') site in the intracellular loop connecting domains II and III. This interaction is important to position docked synaptic vesicles near the source of entering calcium in order to ensure rapid and efficient release of neurotransmitters. Moreover, this interaction regulates calcium channel activity so that channels with nearby docked synaptic vesicles are more likely to open in response to membrane depolarization. Ca_v2 channels also bind G-protein β_γ subunits that inhibit channel activation in response to neurotransmitters, which build up in the synaptic cleft and bind to G-protein-coupled receptors located in presynaptic nerve terminals. Multiple sites of interaction of G-protein β_γ subunits have been identified, with the intracellular loop connecting domains I and II providing a primary point of channel regulation (Figure 3(b)). Ca_v2 channels bind calmodulin and related neurospecific calcium sensor proteins to a site in the C-terminal domain. These proteins mediate short-term facilitation and inactivation of calcium channel activity, and these regulatory effects in turn contribute to short-term facilitation and depression of synaptic transmission during trains of repetitive action potentials. The C-terminal domain of Ca_v2 channels also binds calcium/calmodulin-dependent protein kinase II, which positions this key enzyme to receive calcium signals in nerve terminals and regulate nearby presynaptic proteins by phosphorylation.

Calcium Channel Molecular Pharmacology

The pharmacology of the three subfamilies of calcium channels is quite different, and it relates directly to their distinct physiological roles. The Ca_v1 channels are the molecular targets of the organic calcium channel blockers used widely in the therapy of cardiovascular diseases. These drugs are thought to act at three separate, but allosterically coupled, receptor sites (Table 1). Phenylalkylamines are intracellular pore blockers, which are thought to enter the pore from the cytoplasmic side of the channel and block it. Their receptor site is formed by amino acid residues in the S6 segments in domains III and IV, in close analogy to the local anesthetic receptor site on sodium channels (Figure 4). Dihydropyridines can be channel activators or inhibitors and therefore are thought to act allosterically to shift the channel toward the open or closed state, rather than by occluding the pore. Their receptor site includes amino acid residues in the S6 segments of domains III and IV and the S5 segment of domain III. The dihydropyridine receptor site is closely apposed to the phenylalkylamine receptor site and shares some common amino acid residues (Figure 4). Diltiazem and related benzothiazepines are thought to bind to a third receptor site, but the amino acid residues that are required for their binding overlap extensively with those required for phenylalkylamine binding.

The Ca_v2 subfamily of calcium channels is relatively insensitive to dihydropyridines and other organic calcium channel blockers, but these calcium channels are specifically blocked with high affinity by peptide toxins from spiders and marine snails (Table 1). The $Ca_v2.1$

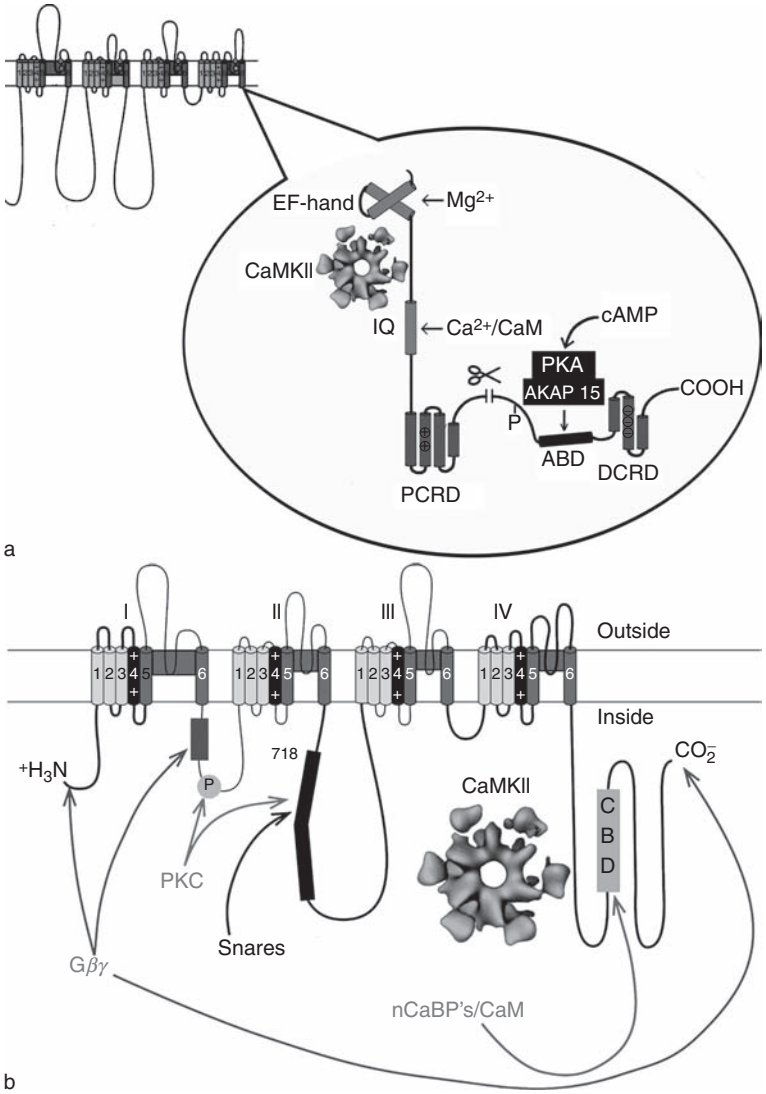

Figure 3 Calcium channel signaling complexes. (a) The $Ca_v1.2$ channel signaling complex. The C-terminal domain of the $Ca_v1.2$ channels is shown in expanded presentation to illustrate its regulatory interactions clearly. ABD, AKAP15-binding domain; DCRD, distal C-terminal regulatory domain; PCRD, proximal C-terminal regulatory domain; scissors, site of proteolytic processing. (b) The presynaptic calcium channel signaling complex. A presynaptic calcium channel α_1 subunit is illustrated as a transmembrane folding diagram as in **Figure 1**. Sites of interaction of SNARE proteins (the synprint site), $G\beta\gamma$ subunits, protein kinase C (PKC), calmodulin, and neuronal calcium-binding proteins (CBD), and calcium/calmodulin-dependent protein kinase II are illustrated.

channels are blocked specifically by ω-agatoxin IVA from funnel web spider venom. The $Ca_v2.2$ channels are blocked specifically by ω-conotoxin GVIA and related cone snail toxins. The $Ca_v2.3$ channels are blocked specifically by the synthetic peptide toxin SNX-482 derived from tarantula venom. These peptide toxins are potent blockers of synaptic transmission because of their specific effects on the Ca_v2 family of calcium channels. Block of $Ca_v2.2$ channels by a synthetic analog of a conotoxin, ziconitide, provides relief from chronic pain in animal models and in humans, and a new generation of organic calcium channel blockers with specificity for $Ca_v2.2$ channels is being evaluated in clinical trials.

The Ca_v3 subfamily of calcium channels is insensitive to both the dihydropyridines that block Ca_v1 channels and the spider and cone snail toxins that block the Ca_v2 channels. There are no widely used pharmacological agents that block T-type calcium currents specifically with high affinity. The organic calcium channel blocker mibefradil is somewhat selective for T-type versus L-type calcium currents (three- to fivefold). High concentrations of ethosuximide block T-type calcium currents specifically, and this effect of the drug may be responsible for its effectiveness in treatment of absence seizures. The peptide kurtoxin inhibits the activation gating of $Ca_v3.1$ and $Ca_v3.2$ channels. Development of more

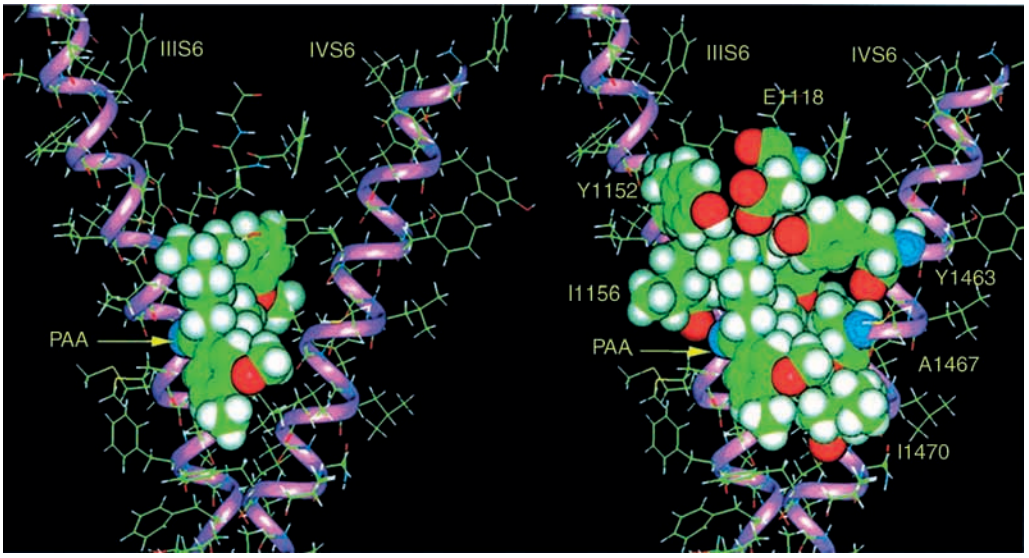

Figure 4 Receptor site for calcium channel blocking drugs. (Left) A phenylalkylamine positioned in the inner area. The high-affinity phenylalkylamine calcium channel blocker desmethoxyverapamil is illustrated in its bound position in a ribbon model of the inner pore, without the side chains of the amino acids so that the position of the drug molecule can be seen. (Right) The phenylalkylamine receptor site. Binding of desmethoxyverapamil (D888) in the half-folded conformation inside of the pore and the interface crevice between domains III and IV is illustrated. The aromatic rings on the ends of D888 interact with the side chains of Tyr1463 and Ile1470, correspondingly. The central amine interacts directly with the carboxylate of the side chain of Glu1118 of the selectivity filter, thereby blocking permeation.

specific and high-affinity blockers of the Ca_v3 subfamily of calcium channels would be useful for more detailed analysis of their physiological roles and potentially for therapy of pain, epilepsy, sleep disorders, and cardiac arrhythmias.

Calcium Channelopathies

Mutations in calcium channels cause multiple inherited disorders in humans. Mutations of skeletal muscle $Ca_v1.1$ channels cause hypokalemic periodic paralysis, which is characterized by flaccid paralysis of voluntary muscles correlated with low serum potassium concentration. Mutations in the $Ca_v1.2$ channels cause Timothy syndrome. This multifaceted disorder includes cardiac arrhythmia, developmental abnormalities, and an autism-like mental retardation. Mutations in $Ca_v1.4$ channels, which are highly expressed in the retina, cause stationary night blindness. Mutations in $Ca_v2.1$ channels cause spinocerebellar ataxia and familial hemiplegic migraine. The mutations in these different disorders all cause important changes in the functional properties of calcium channels, but the precise relationships between changes in the physiological functions of calcium channels and the pathophysiology of these inherited diseases is unresolved and is a very active area of research.

See also: Calcium Channel Subtypes Involved in Neurotransmitter Release; Calcium Channels and SNARE Proteins; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Calcium Homeostasis in Glia; Cyclic Nucleotide-Gated and Hyperpolarization-Activated Channels; Hemiplegic Migraine and Channel Disorders; Ion Channel Localization in Cell Bodies and Dendrites; Large Conductance Calcium-Activated Potassium Channels; Neuromodulation of Calcium Channels; Sodium Channels; Transient Receptor Potential (TRP) Channels; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Calcium Channels; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Catterall WA (2000) Structure and regulation of voltage-gated Ca^{2+} channels. *Annual Review of Cell and Developmental Biology* 16: 521–555.

Catterall WA, Perez-Reyes E, Snutch TP, and Striessnig J (2005) International Union of Pharmacology: XLVIII. Nomenclature and structure–function relationships of voltage-gated calcium channels. *Pharmacological Review* 257: 411–425.

Ertel EA, Campbell KP, Harpold MM, et al. (2000) Nomenclature of voltage-gated calcium channels. *Neuron* 25: 533–535.

Hockerman GH, Peterson BZ, Johnson BD, and Catterall WA (1997) Molecular determinants of drug binding and action on L-type calcium channels. *Annual Review of Pharmacology and Toxicology* 37: 361–396.

Hofmann F, Lacinová L, and Klugbauer N (1999) Voltage-dependent calcium channels: From structure to function. *Reviews of Physiology, Biochemistry and Pharmacology* 139: 33–87.

Lipkind GM and Fozzard HA (2003) Molecular modeling of interactions of dihyodropyridines and phenylalkylamines with the inner pore of the L-type Ca^{2+} channel. *Molecular Pharmacology* 63: 499–511.

Miljanich GP and Ramachandran J (1995) Antagonists of neuronal calcium channels: Structure, function, and therapeutic implications. *Annual Reviews of Pharmacology and Toxicology* 35: 707–734.

Perez-Reyes E (2003) Molecular physiology of low-voltage-activated t-type calcium channels. *Physiological Review* 83: 117–161.

Snutch TP (2005) Targeting chronic and neuropathic pain: The N-type calcium channel comes of age. *NeuroRx* 2: 662–670.

Striessnig J (1999) Pharmacology, structure and function of cardiac L-type calcium channels. *Cellular Physiology and Biochemistry* 9: 242–269.

Triggle DJ (2006) L-type calcium channels. *Current Pharmaceutical Design* 12: 443–457.

Tsien RW, Lipscombe D, Madison D, Bley K, and Fox A (1995) Reflections on calcium channel diversity, 1988–1994. *Trends in Neurosciences* 18: 52–54.

Van Petegem F and Minor DL Jr. (2006) The structural biology of voltage-gated calcium channel function and regulation. *Biochemical Society Transactions* 34: 887–893.

Yu FH and Catterall WA (2004 October 5) The VGL-chanome: A protein superfamily specialized for electrical signaling and ionic homeostasis. *Science STKE* 2004(253): re15.

Calcium Channels and SNARE Proteins

Z-H Sheng and A G M Leenders, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

Ca^{2+}-triggered synaptic vesicle (SV) exocytosis depends on the assembly of the basic fusion machinery, the soluble *N*-ethylmaleimide-sensitive factor attachment protein (SNAP) receptor (SNARE) complex, in which the vesicle-associated v-SNARE protein synaptobrevin (VAMP) interacts with two plasma target membrane-associated SNARE (t-SNARE) proteins, SNAP-25 and syntaxin-1, to form a stable heterotrimeric SNARE complex. Maturation into a release-ready SV requires synaptotagmin I, an integral Ca^{2+}-binding protein of the synaptic vesicle membrane, which provides the Ca^{2+}-dependent regulation of the fusion machinery. Calcium influx into the presynaptic terminal triggers the calcium sensor of the fusion machinery, upon which the SNARE complex binding progresses from a *trans* to a *cis* state, resulting in the fusion of apposing membranes and the release of neurotransmitter.

Neurotransmitter release is initiated within 200 μs after the arrival of the action potential and critically depends on Ca^{2+} influx through high-voltage Ca^{2+} channels. Given that synaptic vesicle exocytosis is a highly Ca^{2+}-regulated process and that the intracellular Ca^{2+} concentration falls off steeply as a function of the distance away from the Ca^{2+} channels, the brief rise in the Ca^{2+} concentration to the level necessary for exocytosis is thought to occur only in close proximity to the presynaptic Ca^{2+} channels. Thus, the tight, spatially localized rise in intracellular Ca^{2+} and the subsequent activation of synaptic vesicle fusion are likely achieved by the direct physical link between the Ca^{2+} channels and the synaptic vesicle release machinery.

Ca^{2+} channels are multisubunit complexes, composed of a major pore-forming transmembrane subunit $\alpha1$ and smaller auxiliary subunits that include β and disulfide-linked $\alpha2\delta$ subunits. Of the many types of voltage-gated Ca^{2+} channels (Ca_v1, Ca_v2, and Ca_v3) expressed in neurons, members of the Ca_v2 family, in particular $Ca_v2.1$ (P/Q-type) and $Ca_v2.2$ (N-type) channels, are the predominant subtypes supporting neurotransmitter release in central synapses. New evidence has emerged indicating that both of these subtypes interact directly with SNARE proteins through a region on the intracellular loop connecting domains II–III of the $\alpha1$ subunit; this region is called the synaptic protein interaction site, or, for short, the synprint site. The stability of these interactions is critically dependent on the intracellular Ca^{2+} concentration, protein kinase C (PKC), and Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) phosphorylation of the synprint site, thus allowing the channel–SNARE complexes to be dynamically regulated in response to Ca^{2+} influx and activation of signal transduction pathways. Disruption of these interactions in presynaptic neurons impairs synaptic transmission by reducing the amplitude of excitatory postsynaptic potentials (EPSPs) and the sensitivity of neurotransmitter release to external Ca^{2+}. Studies support the hypothesis that the synprint site tethers the readily releasable pool of SVs and the release machinery near presynaptic Ca^{2+} channels for effective fast exocytosis. In addition to this functional role in the anterograde process of synaptic transmission, SNARE proteins also retrogradely modulate the voltage dependence of steady-state inactivation on Ca_v2 channels, thus providing a feedback mechanism for fine-tuning Ca^{2+} influx and, hence, synaptic transmission.

The SNARE Protein Interaction Site on Presynaptic Ca_v2 Channels

About a decade ago, immunocytochemical studies in neurons indicated that both N- and P/Q-type Ca_v2 channels are densely localized with syntaxin-1 at presynaptic nerve terminals and are crucial components in neuronal excitation–secretion coupling. Furthermore, these channels can be isolated as a complex with SNARE proteins from brain homogenates, suggesting that both types of Ca_v2 channels are part of the synaptic vesicle fusion complexes. These initial findings triggered an intensive investigation into what molecular determinant allows these channels to bind SNARE proteins and how this physical interaction regulates synaptic transmission and channel function.

The direct interaction between the presynaptic Ca^{2+} channel and syntaxin-1A was first described in 1994 by Sheng and colleagues based on *in vitro* biochemical analysis of the recombinant SNARE proteins associated with fusion proteins expressing various cytoplasmic loops of the N-type $Ca_v2.2$ channel. The SNARE proteins syntaxin-1A and SNAP-25, but not VAMP2, specifically interact with an amino acid sequence extending from residues 718 to 963 within the cytoplasmic loop between domains II and III (LII–LIII) of the N-type Ca^{2+} channel $\alpha1B$ subunit, the synprint site (Figure 1(a)). The synprint peptide specifically

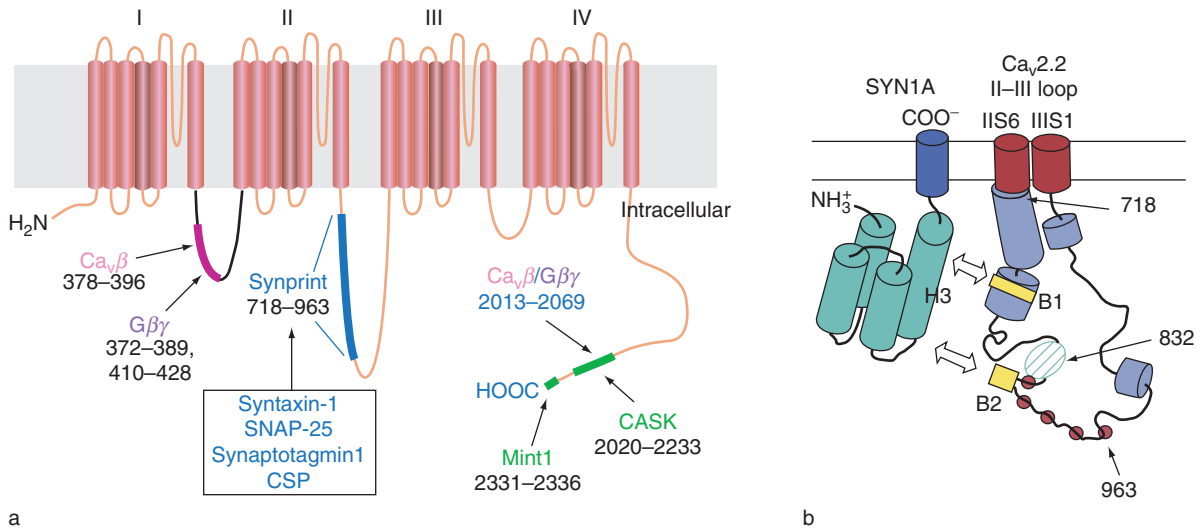

Figure 1 Structural model of the $Ca_v2.2$ synprint domain. (a) The synprint site of the $Ca_v2.2$ channels. The α1 subunit is composed of four homologous transmembrane domains, which are formed by six membrane-spanning helices and connected by large cytoplasmic loops. The domain II–III loop contains the synprint site, which binds syntaxin-1, SNAP-25, cysteine string protein (CSP), and synaptotagmin I. (b) Working model of the interaction between the $Ca_v2.2$ synprint with syntaxin-1A (SYN1A). Based on a Chou–Fasman secondary structural prediction algorithm, an ~80-amino-acid-long bundle with high probability of forming a helical domain is depicted extending from IIS6 into the II–III loop . This helical segment is interrupted by a short random coil that precedes a second α helix containing the first binding site (B1; yellow box). Following this helical segment, a coil domain with collagen-like sequence homology (amino acids 823–848; hatched oval) is adjacent to the second binding site (B2; yellow box). Red balls mark the arginine- and histidine-rich clusters, which may serve as membrane surface interaction sites. The synprint B1 and B2 segments provide potential binding contact sites for the C-terminal coiled-coil (H3) of syntaxin-1 in a closed conformation (white arrows). (a) Adapted from Spafford JD and Zamponi GW (2003) Functional interactions between presynaptic calcium channels and the neurotransmitter release machinery. *Current Opinion in Neurobiology* 13: 308–314, with permission from Elsevier. (b) Adapted from Yokoyama C, Myers SJ, Fu J, *et al.* (2005) Mechanism of SNARE protein binding and regulation of Ca_v2 channels by phosphorylation of the synprint protein interaction site. *Molecular and Cellular Neuroscience* 28: 1–17, with permission from Elsevier.

interferes with the co-immunoprecipitation of native N-type Ca^{2+} channels with syntaxin-1, indicating that this binding site is required for the stable interaction of these two proteins. This interaction was shown to be Ca^{2+} dependent, with maximal binding at 20 μM Ca^{2+} and reduced binding at resting Ca^{2+} concentrations and at concentrations above 50 μM. Although the physiological significance of this biphasic Ca^{2+}-dependent interaction is not clear, reduced binding affinity at high Ca^{2+} levels may imply that there is a sequential step in the dissociation of syntaxin-1A and SNAP-25 from the channels, in order for fusion to proceed.

A recent study by Yokoyama and colleagues defined the molecular determinants of syntaxin-1A and SNAP-25 binding in the synprint site more precisely. Two different peptide segments of $Ca_v2.2$, amino acids 773–786, and 872–901, are independently required to bind syntaxin-1A and SNAP-25 to each half of the synprint site. Based on structural analysis using the Basic Local Alignment Search Tool for proteins (BLAST-P), the two halves of the synprint site are separated by a linker sequence that contains a collagen-like strand with a number of proline and glycine residues (Figure 1(b)). This motif is predicted to form an extended and flexible linker which connects the two SNARE-binding segments. Furthermore, the synprint sequence exhibits a high charge density, with a charged residue found, on average, every 2.5 amino acids. This hydrophilic structure of the synprint site allows for facile interactions with the hydrophilic surface of the amphipathic α helices of SNARE proteins. In addition, repeated arginine/histidine-rich sequences in the synprint site may provide contacts with phospholipids in the plasma membrane or in the docked synaptic vesicles, thus mediating a structural tethering of synaptic vesicles near presynaptic $Ca_v2.2$ channels.

$Ca_v2.1$ (P/Q-type) channels also contain a synprint site in their LII–LIII intracellular loop. However, unlike the N-type, two isoforms of the α1A subunit of $Ca_v2.1$ channels, BI and rbA, exhibit different patterns of interaction with synaptic proteins and different Ca^{2+} dependencies. While the BI isoform interacts with syntaxin-1A and SNAP-25 independent of Ca^{2+} concentration, the rbA isoform interacts with SNAP-25 but not with syntaxin-1. Given that these Ca^{2+} channels are differentially distributed in the synapses of the central nervous system, the observed differences in the interactions between the channels and the SNARE proteins may reflect subtle variations in the efficacy of the initiation of transmitter release by Ca^{2+}

influx through these different channels, and may confer specialized regulatory properties that contribute to synaptic modulation. The synprint site is considered unique to vertebrate $Ca_v2.1$ (P/Q-type) and $Ca_v2.2$ (N-type) channels, as other Ca^{2+} channel subtypes, such as $Ca_v1.2$ (L-type, for neuroendocrine secretion) and $Ca_v2.3$ (R-type, for synaptic transmission), lack the classical synprint site in the corresponding regions. However, the functional interaction of both $Ca_v1.2$ and $Ca_v2.3$ channels with SNARE proteins is observed, possibly controlled by a mechanism that does not involve the synprint site.

Both $Ca_v2.1$ and $Ca_v2.2$ channels interact with the C-terminus of syntaxin-1A (amino acids 181–288), suggesting that Ca^{2+} channels bind to syntaxin-1A at a C-terminal site near the intracellular surface of the plasma membrane (**Figure 1(b)**). In addition, a recent study suggests that the N-terminal sequence (amino acids 1–69) of syntaxin-1A is also able to bind synprint *in vitro*.

Besides the SNARE proteins syntaxin-1A and SNAP-25, one other abundant synaptic vesicle protein, synaptotagmin, has been shown to directly interact with the synprint site of both N- and P/Q-type channels. Synaptotagmins are a family of Ca^{2+}-binding proteins containing two homologous C2 domains, the importance of which in exocytosis has been well established both genetically and biochemically. Synprint directly binds to synaptotagmin-I through the same region in the C2B domain required for synaptotagmin oligomerization and for adapter protein (AP-2) binding. Within this region, two adjacent lysine residues (K326 and K327) are critical for synaptotagmin oligomerization and AP-2 and synprint binding. Moreover, syntaxin interacts exclusively with either synprint or synaptotagmin in a Ca^{2+}-dependent manner, such that at low Ca^{2+} concentration syntaxin-1 binds synprint, whereas at higher concentrations (>30 μM) its association with synaptotagmin increases, thus providing a potential biochemical correlate for the sequence of events occurring at the synprint site of presynaptic Ca^{2+} channels during synaptic vesicle exocytosis. Collectively, these studies provide a molecular basis for a physical link of the neuronal N-type and P/Q-type Ca^{2+} channels with the synaptic vesicle fusion machinery, enabling tight structural and functional coupling of Ca^{2+} entry sites and neurotransmitter release sites.

The Synprint Site Is Important for Efficient Release of Neurotransmitter

When the action potential invades the nerve terminal, the Ca^{2+} influx through presynaptic Ca^{2+} channels rapidly elevates the Ca^{2+} concentration to 200–300 μM in microdomains of approximately 0.3 μm^2 at presynaptic terminal. This transient Ca^{2+} influx triggers synaptic vesicle exocytosis through binding to the Ca^{2+} sensor of the release machinery. Since diffusion and sequestration rapidly decrease the Ca^{2+} concentration, the Ca^{2+} concentration at the nerve terminal falls off steeply as a function of the distance from the Ca^{2+} entry sites. A physical anchoring of the release machinery to the source of the Ca^{2+} influx would be an attractive model for the docked synaptic vesicles to be spatially localized within the Ca^{2+} microdomains. Thus, whether binding of the SNARE proteins to the synprint site contributes to the efficiency of neurotransmitter release has been the subject of considerable investigation.

The first functional study by Mochida and coworkers demonstrated that injection of the synprint peptides into presynaptic superior cervical ganglion neurons (SCGNs) in culture significantly depressed the excitatory postsynaptic response, presumably by competitive uncoupling of the endogenous Ca^{2+} channels–SNARE interaction. In this case, rapid synchronous synaptic transmission is inhibited, while late asynchronous release and paired-pulse facilitation are increased, consistent with the hypothesis that synaptic vesicles are shifted from a pool for synchronous release to a pool not optimally primed or positioned for synchronous release. This study indicates that in addition to mediating Ca^{2+} entry, the N-type Ca^{2+} channels may have an effect on the transmitter release via its interaction with the SNARE proteins. This study also supports the notion that a tight physical interaction between the presynaptic Ca^{2+} channels and the release apparatus is required for rapid, synchronous neurotransmitter release.

To determine whether the inhibition of neurotransmitter release by the synprint peptides is due to the displacement of the docked vesicles from the N-type Ca^{2+} channels, Rettig and colleagues used *Xenopus* nerve muscle co-cultures, in which synaptic transmission is mainly dependent on the N-type Ca^{2+} channels. Injection of the synprint peptides into embryonic *Xenopus* spinal neurons shifted the Ca^{2+} dependence of release to a higher external concentration, even though Ca^{2+} influx was not altered. Transmitter release was reduced by approximately 50% when cells were stimulated in an extracellular solution containing physiological Ca^{2+} concentration. High external Ca^{2+} concentration rescued this inhibition, implying that the Ca^{2+} channels are not optimally located for efficient release, or are detached from release sites following the addition of the synprint peptides.

To determine whether the synprint deletion mutant could support vesicle fusion in secretory cells, Harkins and colleagues transiently expressed the wild-type or

the synprint-deleted mutant of $Ca_v2.2$ channels in mouse pheochromocytoma (MPC) 9/3L, a cell line containing the release machinery but lacking the voltage-dependent Ca^{2+} channels. When compared with the cells expressing the wild-type channels, expression of the mutant channel in the cells significantly reduces both total secretion and the rate of secretion as determined by capacitance measurements, although Ca^{2+} currents are similar regardless of whether synprint is deleted. A similar study by Mochida and colleagues showed that the deletion of the synprint site on the $Ca_v2.1$ P/Q-type channel results in a 50% reduction in release in cultured SCG neurons. This effect is correlated with a reduced localization of the exogenously expressed channels to presynaptic terminals. Transfer of the synprint sequence from $Ca_v2.1$ to $Ca_v1.2$ channels, which conduct L-type Ca^{2+} currents and are ineffective in supporting synaptic transmission, is sufficient to initiate synaptic transmission. These results consistently support the notion that the synprint site is necessary for the efficient functional coupling of Ca^{2+} influx through the $Ca_v2.1$ and $Ca_v2.2$ channels and vesicle exocytosis (Figure 2).

Given that t-SNARE proteins are widely distributed in neurons, including the dendrites and soma, the interaction between the synprint site and the t-SNARE proteins is likely a permissive event for presynaptic localization of the Ca^{2+} channels. Additional primary targeting signals required for nerve terminal localization of presynaptic Ca^{2+} channels remain to be identified. A recent study suggested that the adapter proteins Mint and CASK play a critical role in the targeting of N-type Ca^{2+} channels via a direct interaction with the C-terminal tail of the channels (Figure 1(a)).

While the synprint–SNAREs interaction critically affects the efficiency of neurotransmitter release in vertebrate neurons, recent evidence suggests that this interaction might not be absolutely required for synaptic vesicle exocytosis. First, invertebrate Ca^{2+} channels, representative of the Ca_v2 channels, lack a synprint site, although these channels participate in synaptic transmission. Second, splice variants of the human N-type channels lacking the synprint region have been identified. As originally noted by Zamponi, these findings raise the question as to how these

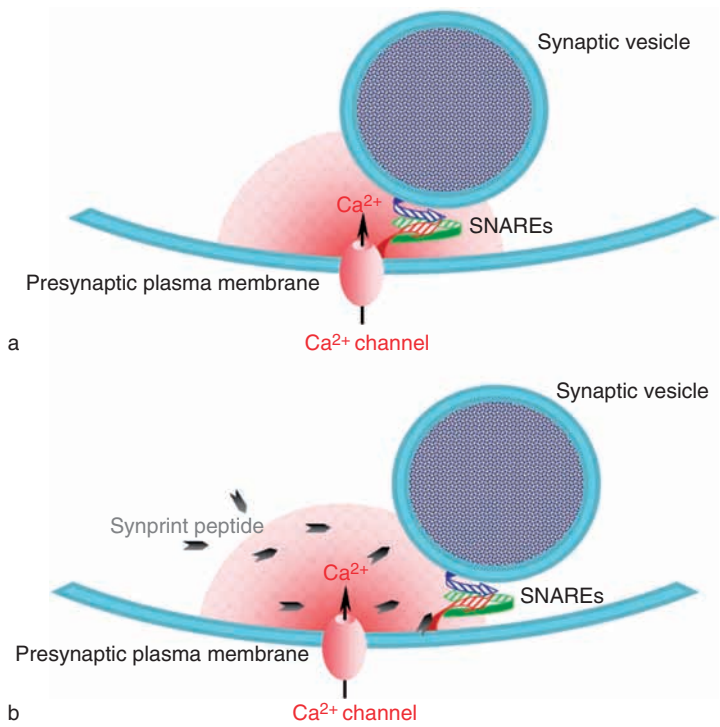

Figure 2 The proposed inhibitory role of the synprint peptide on the efficiency of neurotransmitter release. (a) The physical and functional coupling of presynaptic Ca_v2 channels and synaptic vesicle fusion machinery. As the intracellular Ca^{2+} concentration falls off steeply as a function of distance away from the Ca^{2+} entry sites, fast exocytosis likely occurs only in proximity to the Ca^{2+} channels. Thus, binding of the SNARE proteins to the synprint site contributes to the efficiency of neurotransmitter release. (b) The synprint peptide (gray chevrons) competitively uncouples the physical link between the Ca^{2+} channels and the SNARE proteins. This decreases the efficiency by shifting Ca^{2+} dependence to higher values. Rapid, synchronous synaptic transmission is inhibited, while late, asynchronous release and paired-pulse facilitation are increased, consistent with the hypothesis that synaptic vesicles are shifted from a pool primed for synchronous release near the Ca^{2+} entry sites to a pool not optimally primed or positioned for synchronous release.

channels couple to the synaptic vesicle fusion machinery. In addition, the deletion of the synprint site on both $Ca_v2.1$ and $Ca_v2.2$ channels did not completely abolish synaptic transmission and vesicle exocytosis. Therefore, further investigation is necessary to determine whether synprint serves as a primary binding site for the SNARE proteins or just as a SNARE anchoring domain to maintain the structural integrity of the fusion complex at nerve terminals.

Regulation of Presynaptic Ca^{2+} Channels by SNARE Proteins

A number of studies have documented that syntaxin-1A and SNAP-25 not only bind to presynaptic Ca^{2+} channels in order to closely couple the channels and the vesicle release machinery, but they are also capable of regulating channel activity, thus acting as a negative feedback mechanism for the modulation of synaptic transmission. As initially reported by Bezprozvanny and colleagues and shortly after by Wiser and colleagues, co-expression of syntaxin-1A with $Ca_v2.1$ or $Ca_v2.2$ channels in *Xenopus* oocytes results in a reduced availability of the channels to open and inhibits Ca^{2+} channel activity by shifting the voltage dependence of steady-state inactivation under long, depolarizing prepulses toward more negative membrane potentials. However, strong hyperpolarizing pulses can overcome this effect, implying that syntaxin-1A does not simply block or reduce expression of the Ca^{2+} channels. Interestingly, Zhong and colleagues demonstrated that the inhibitory effects of syntaxin could be relieved by co-expressing synaptotagmin, suggesting a dynamic process involving multiprotein interactions or the structural reorganization of these protein complexes.

Like syntaxin-1A, SNAP-25 reduces calcium channel-opening probability when co-expressed with N- and P/Q-type channels. However, as reported by Javis and Zamponi, in the presence of both syntaxin-1 and SNAP-25, the shift toward inactivation is ablated and the opening probability of the channels is restored to normal. Thus, depending on the protein(s) associated, the channels differ in their ability to contribute to Ca^{2+} influx. When in the presence of only syntaxin-1A or SNAP-25, the synaptic vesicles are unable to dock at the plasma membrane or near the channel, and the negative shift in inactivation potential reduces Ca^{2+} entry through this channel, thus preventing Ca^{2+} overload. In contrast, the formation of the SNARE complex mediates synaptic vesicles docking at release sites and concomitantly increases the channel availability for opening, thus favoring Ca^{2+} influx through the channels localized at the docked synaptic vesicles. In principle, such a mechanism could provide selective Ca^{2+} entry through channels bound to docked vesicles, whereas channels associated with one SNARE protein remain relatively inactive. This mechanism allows the modulatory interaction for the channels to be switched on and off at appropriate recycling stages of the SNARE proteins and synaptic vesicles. In addition, the inhibitory role of syntaxin-1 on channel opening is abolished in the presence of neuronal Sec-1 (nSec-1), a protein that blocks the assembly of the SNARE complex by locking syntaxin-1 in a closed conformation. Altogether, these findings support a model for the SNARE-mediated modulation of presynaptic Ca^{2+} channels: the channel activity is inhibited when it is associated with syntaxin-1 in a closed conformation and in the absence of the other SNARE proteins.

Stanley and Mirotznik revealed that syntaxin-1A regulates presynaptic calcium channels in intact neuronal terminals via its susceptibility to G-protein inhibition. Botulinum toxin, which cleaves syntaxin-1A, prevents G-protein modulation of presynaptic Ca^{2+} channels in chick calyces, suggesting a cross-talk between syntaxin-1A and G-protein effects on N-type Ca^{2+} channels. Subsequent studies by Jarvis and colleagues showed that the synprint site and $G\beta$ bind to distinct portions of syntaxin-1A and that expression of syntaxin-1A appears to be a prerequisite for tonic G-protein inhibition of the N-type Ca^{2+} channels. These results support a hypothesis that syntaxin-1A acts as a chaperone by tethering $G\beta$ to the N-type Ca^{2+} channels in synapses, thus facilitating G-protein regulation.

By combining site-directed mutagenesis with functional expression in *Xenopus* oocytes, Bezprozvanny and colleagues addressed an important question as to whether the anchoring and modulatory interactions between syntaxin-1 and Ca^{2+} channels are mediated by the same molecular determinant. Their studies showed that both the transmembrane region and a short segment within the H3 helical cytoplasmic domain of syntaxin-1A are critical for channel modulation but not for binding to the synprint site of the N-type Ca^{2+} channels. Given that this short segment is buried within the core of the four-helix structure of the SNARE complex, the modulatory action of syntaxin-1A can be enabled only after disassembly of the SNARE complex following synaptic vesicle exocytosis. The synprint site binds to the whole H3 region of syntaxin-1A, confirming the original biochemical findings of Sheng and colleagues. However, deletion of the synprint site weakened the modulation of the channels by syntaxin-1A, but did not abolish it, arguing that the synprint sequence acts as an important anchor in facilitating channel modulation but is not essential for the modulatory action. While the

precise site on the channels for syntaxin-mediated modulation remains to be determined, the identification of separate anchoring and modulation sites on the channels provides molecular and structural insights into how Ca^{2+} entry into the nerve terminals is tightly correlated to the recycling state of the release machinery.

Although the SNARE complex, together with synaptotagmin, is thought to be the minimal essential machinery required for Ca^{2+}-dependent synaptic vesicle fusion, a multitude of presynaptic proteins have been identified over the past decade, most of which have been shown to directly participate in or regulate the synaptic vesicle cycle via interactions with the SNARE proteins, thus having the potential to modulate the SNARE–synprint interaction. A direct interaction between Munc18 and the synprint site was reported, although the significance of this interaction in synaptic transmission remains to be elucidated. Rab-interacting molecule (RIM), an active-zone protein implicated in short- and long-term synaptic plasticity, has also been shown to interact with the synprint region *in vitro*. Since RIM interacts with multiple active-zone components and synaptic vesicle proteins, it is unclear how RIM modulates sequential steps in synaptic vesicle exocytosis and how these biochemical interactions translate into physiological functions. Cysteine string protein (CSP) was first identified as a synaptic vesicle-associated protein with the ability to modify Ca^{2+} influx in *Drosophila* and *Torpedo*. Subsequent studies demonstrated that CSP interacts with the synprint site of both P/Q- and N-type Ca^{2+} channels and that this interaction results in a robust tonic inhibition of channel activity by G-protein $\beta\gamma$ subunits. The ability of CSP to enhance G-protein signaling suggests that a channel physically coupled to a docked vesicle is more prone to G-protein inhibition than one uncoupled to a docked vesicle. These studies further support the notion that the functional activity of the presynaptic Ca^{2+} channels is regulated by the SNAREs and SNARE-binding proteins.

Modulation of the SNARE–Synprint Interaction

Second-messenger-activated regulation of neurotransmitter release via modulation of protein interactions within the exocytotic apparatus has a potentially important role in synaptic plasticity. Several protein kinases, including PKA, PKC, PKG, and CaMKII, are expressed in presynaptic terminals. Ca^{2+} channels and the SNARE proteins are phosphorylated by one or more of these kinases. Original *in vitro* biochemical studies by Yokoyama and colleagues revealed that phosphorylation of the synprint peptide with PKC and CaMKII strongly inhibits its binding to recombinant syntaxin-1A and SNAP-25 and also prevents its association with the native SNARE complex isolated from rat brain homogenates. Moreover, subsequent functional studies in both the Tsien and Zamponi laboratories consistently showed that PKC phosphorylation blocks the negative shift of steady-state inactivation of $Ca_v2.2$ channels, possibly by a reduced binding of syntaxin-1 to the synprint site. These studies suggest that phosphorylation of the synprint site by PKC or CaMKII may serve as a biochemical switch for the SNARE–synprint interactions, and thus phosphorylation is a candidate presynaptic mechanism for the regulation of neurotransmitter release.

Yokoyama and colleagues recently reported a more detailed analysis of these phosphorylation sites. The synprint site of $Ca_v2.2$ channels consists of two separate microdomains, each of which binds syntaxin-1 and SNAP-25 *in vitro* and is regulated by PKC phosphorylation at serines 774 and 898 and by CaMKII phosphorylation at serines 784 and 896. Each phosphorylation site can control syntaxin-1 and SNAP-25 binding to half of the synprint site, suggesting a bipartite structure for the synprint domain. However, phosphorylation of the intact $Ca_v2.2$ channels before exposure to syntaxin-1A reduces the binding affinity to syntaxin-1A, but does not completely prevent it or cause its dissociation. These observations are consistent with the functional study, in which PKC activation in the transfected tsA-201 cells does not completely dissociate syntaxin-1A from the channels, but does completely reverse the negative shift of steady-state inactivation caused by syntaxin-1A. Altogether, these findings support a notion that while PKC phosphorylation of $Ca_v2.2$ channels is sufficient to block the syntaxin-mediated modulation of channel function, it destabilizes but not completely disrupts its physical interaction with syntaxin-1A. As PKC phosphorylation reverses the syntaxin-1A effect on channel inactivation, this mechanism might contribute to increased Ca^{2+} entry at sites near docked vesicles, thus enhancing the effectiveness of Ca^{2+} influx during neurotransmission.

Conclusion

The interactions between the presynaptic Ca^{2+} channel synprint site and SNARE proteins not only provide a molecular basis for a physical and functional coupling between Ca^{2+} influx and subsequent exocytosis, but also provide a model for sequential protein–protein interactions at or near the Ca^{2+} channels – interactions that are either essential or modulatory steps in the release of synaptic vesicles. Further regulation is

imposed by interplay between the synprint site interactions and second-messenger modulation of calcium channels through phosphorylation and G-proteins. Altogether, the complexity of the interactions at and near the synprint site of the Ca^{2+} channels suggests that the mammalian brain has devised numerous pathways for the selective and precise control of sequential events, resulting in efficient vesicle release in the presynaptic nerve terminals.

See also: Calcium Channels; Calcium Channel Subtypes Involved in Neurotransmitter Release; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Dendritic Signal Integration; Exocytosis: Ca^{2+}-Sensitivity; Invertebrate Neurohormone GPCRs; Ion Channel Localization in Cell Bodies and Dendrites; Neuromodulation of Calcium Channels; SNAREs; Voltage-Gated Calcium Channels.

Further Reading

Bezprozvanny I, Scheller RH, and Tsien RW (1995) Functional impact of syntaxin on gating of N-type and Q-type calcium channels. *Nature* 378: 623–626.

Bezprozvanny I, Zhong P, Scheller RH, et al. (2000) Molecular determinants of the functional interaction between syntaxin and N-type Ca^{2+} channel gating. *Proceedings of the National Academy of Sciences of the United States of America* 97: 13943–13948.

Catterall WA (2000) Structure and regulation of voltage-gated Ca^{2+} channels. *Annual Review of Cell and Developmental Biology* 16: 521–555.

Harkins AB, Cahill AL, Powers JF, et al. (2004) Deletion of the synaptic protein interaction site of the N-type ($Ca_V2.2$) calcium channel inhibits secretion in mouse pheochromocytoma cells. *Proceedings of the National Academy of Sciences of the United States of America* 101: 15219–15224.

Jarvis SE and Zamponi GW (2003) Functional interactions between presynaptic calcium channels and the neurotransmitter release machinery. *Current Opinion in Neurobiology* 13: 308–314.

Mochida S, Sheng ZH, Baker C, et al. (1996) Inhibition of neurotransmission by peptides containing the synaptic protein interaction site of N-type Ca^{2+} channels. *Neuron* 17: 781–788.

Mochida S, Westenbroek RE, Yokoyama CT, et al. (2003) Requirement for the synaptic protein interaction site for reconstitution of synaptic transmission by P/Q-type calcium channels. *Proceedings of the National Academy of Sciences of the United States of America* 100: 2819–2824.

Rettig J, Sheng ZH, Kim DK, et al. (1996) Isoform-specific interaction of the alpha1A subunits of brain Ca^{2+} channels with the presynaptic proteins syntaxin and SNAP-25. *Proceedings of the National Academy of Sciences of the United States of America* 93: 7363–7368.

Rettig J, Heinemann C, Ashery U, et al. (1997) Alteration of Ca^{2+} dependence of neurotransmitter release by disruption of Ca^{2+} channel/syntaxin interaction. *Journal of Neuroscience* 17: 6647–6656.

Sheng ZH, Rettig J, Takahashi M, et al. (1994) Identification of a syntaxin-binding site on N-type calcium channels. *Neuron* 13: 1303–1313.

Sheng ZH, Rettig J, Cook T, et al. (1996) Calcium-dependent interaction of N-type calcium channels with the synaptic core complex. *Nature* 379: 451–454.

Sheng ZH, Westenbroek RE, and Catterall WA (1998) Physical link and functional coupling of calcium channels and the synaptic vesicle docking fusion machinery. *Journal of Bioenergetics and Biomembranes* 30: 335–345.

Spafford JD and Zamponi GW (2003) Functional interactions between presynaptic calcium channels and the neurotransmitter release machinery. *Current Opinion in Neurobiology* 13: 308–314.

Stanley EF and Mirotznik RR (1997) Cleavage of syntaxin prevents G-protein regulation of presynaptic calcium channels. *Nature* 385: 340–343.

Yokoyama CT, Myers SJ, Fu J, et al. (2005) Mechanism of SNARE protein binding and regulation of Ca_v2 channels by phosphorylation of the synaptic protein interaction site. *Molecular and Cellular Neuroscience* 28: 1–17.

Zamponi GW (2003) Regulation of presynaptic calcium channels by synaptic proteins. *Journal of Pharmacological Sciences* 92: 79–83.

Calcium Homeostasis in Glia

A C Charles and K C Brennan, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

All cells use changes in intracellular calcium concentration as a signal to regulate their physiological function. For glial cells, the study of calcium signaling has led to significant new insights into their functional roles in the nervous system. The technique of calcium imaging using fluorescent calcium indicators has been a driving force behind the increasing recognition that glial cells are active participants in most functions of neural tissues. Knowledge of the specific mechanisms of calcium signaling in glial cells represents an important foundation for an understanding of the unique functions of these cells in the physiological and pathological processes of the central nervous system.

Glia cells (astrocytes, oligodendrocytes, and microglia) all show increases in the concentration of cytoplasmic free calcium ($[Ca^{2+}]_i$) with multiple spatial and temporal patterns. Ca^{2+} signaling in astrocytes has been most commonly studied (see Figure 1). Increases in $[Ca^{2+}]_i$ may be localized in subcellular microdomains, propagated as waves within individual cells, or communicated from cell to cell as intercellular waves. Ca^{2+} increases may occur as single transients or as repetitive oscillations. While many of the patterns of glial Ca^{2+} signaling were originally observed in cell culture or *in vitro* in brain slice preparations, recent *in vivo* studies have confirmed that glial Ca^{2+} responses occur under physiological conditions in the intact nervous system.

Sources of Calcium in Glial Cells

Glial-cell Ca^{2+} signals may be generated by release of Ca^{2+} from intracellular stores, influx of Ca^{2+}, or a combination of the two. An individual Ca^{2+} response involves a complex interplay between different sources and sinks for Ca^{2+}. Because glial cells are 'nonexcitable cells' – that is, most glial cells don't generate action potentials – their Ca^{2+} signaling typically relies on release of Ca^{2+} from intracellular stores to a greater extent than in neurons, where Ca^{2+} signaling depends more on Ca^{2+} influx that is related to changes in membrane potential. A typical glial Ca^{2+} response is therefore slower and more sustained than that of a neuron.

Intracellular Calcium Release

Most glial-cell Ca^{2+} signaling responses involve release of Ca^{2+} from intracellular stores. This can be demonstrated by the occurrence of Ca^{2+} responses in the absence of extracellular $Ca^{2+}$and by the inhibition of Ca^{2+} responses by depletion of intracellular Ca^{2+} stores by the sarcoplasmic/endoplasmic reticulum adenosine triphosphatase (SERCA) Ca^{2+} pump inhibitors such as thapsigargin or cyclopiazonic acid. Ca^{2+} stores that are released by binding of intracellular messenger inositol triphosphate (IP_3) to IP_3 receptors on the endoplasmic reticulum appear to be the primary source for intracellular Ca^{2+} release in glial cells. Uncaging of IP_3 evokes Ca^{2+} responses in astrocytes similar to those observed in response to other stimuli. Conversely, inhibition of IP_3-mediated signaling by blocking the IP_3 receptor (IP_3R) with xestospongin or heparin or by buffering IP_3 blocks astrocyte Ca^{2+} responses. The role of Ca^{2+} stores released by ryanodine receptors is not as well defined. Astrocytes in culture have been shown to express ryanodine receptors associated with the endoplasmic reticulum, and caffeine and other ryanodine receptor agonists have been reported to evoke a Ca^{2+} response in some preparations but not in others. In general, however, pretreatment with ryanodine does not affect spontaneous or evoked Ca^{2+} signaling in glia, suggesting the ryanodine receptors do not play a primary role in most types of astrocyte Ca^{2+} signaling. Cells in the oligodendrocyte lineage also express ryanodine receptors which have been implicated in the generation of highly localized calcium release events, known as calcium 'sparks.' Ryanodine receptors may also co-localize with other proteins involved in calcium signaling in specific microdomains in oligodendrocyte progenitors. As in astrocytes, however, the function of ryanodine receptors and their corresponding calcium stores in oligodendrocytes remains uncertain.

SERCAs pump Ca^{2+} from the cytoplasm into membrane-bound stores. The function of these pumps has been investigated primarily with compounds that inhibit them – thapsigargin and cyclopiazonic acid. Application of these SERCA inhibitors to glial cells typically evokes a sustained increase in $[Ca^{2+}]_i$ and depletes intracellular Ca^{2+} stores, making the cell incapable of responding to subsequent stimuli that normally evoke Ca^{2+} release. Fluorescently tagged thapsigargin indicates that SERCAs are distributed throughout the cell in association with intracellular membranes and that in oligodendrocytes, SERCAs appear to be localized in proximity to mitochondria. Other studies indicate that SERCAs and IP_3Rs are

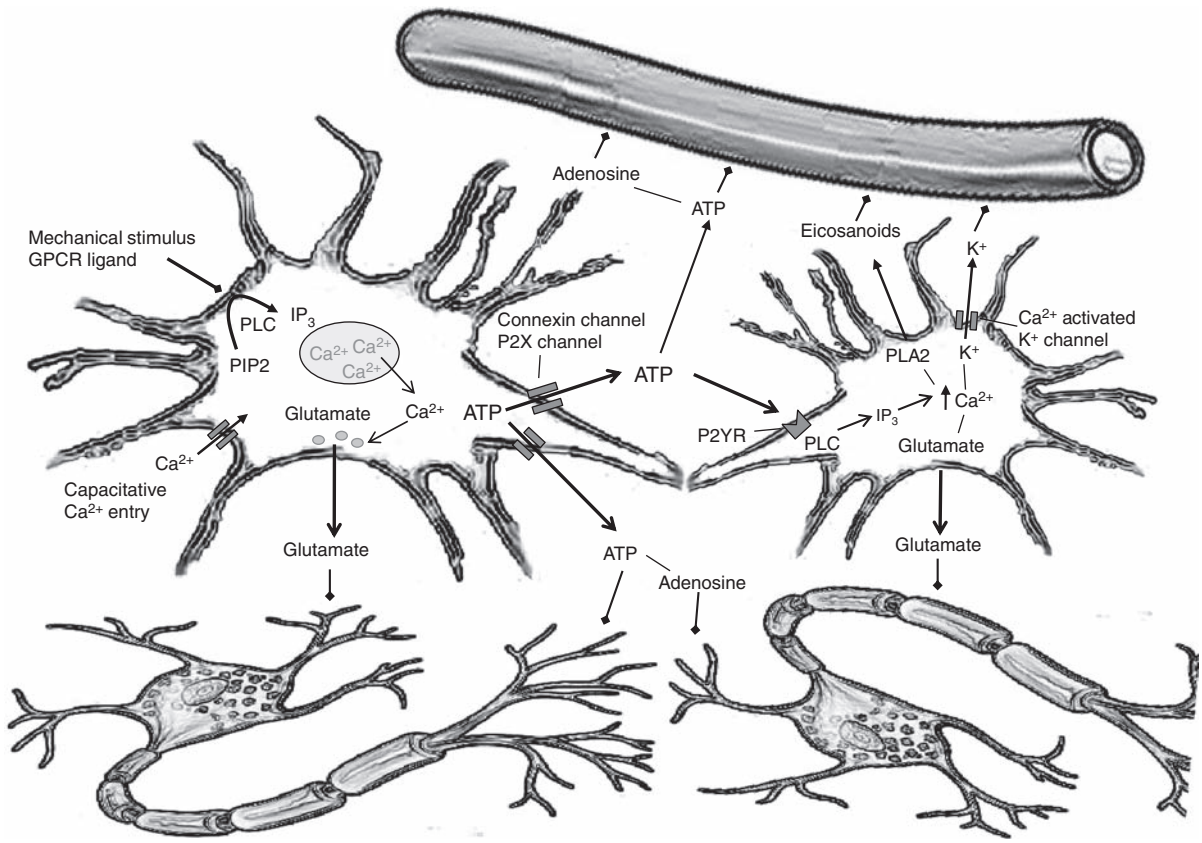

Figure 1 Astrocyte calcium signaling. Schematic showing some of the primary components of intra- and intercellular calcium signaling in astrocytes and their interaction with neurons and blood vessels. ATP, adenosine triphosphate; GPCR, G-protein-coupled receptor; PLC, phospholipase C; PIP2, phosphatidylinositol bisphosphate; IP_3, inositol triphosphate; P2X channel, ionotropic purinergic receptor channel; P2YR, purinergic receptor; PLA2, phospholipase A2.

present in complexes of plasma membrane and endoplasmic reticulum, forming functional units for Ca^{2+} signaling that coordinate Ca^{2+} influx and Ca^{2+} release.

Capacitative Calcium Entry

Release of Ca^{2+} from intracellular stores in glial cells activates influx of Ca^{2+} through channels known as 'store-operated channels.' This Ca^{2+} influx, also known as 'capacitative Ca^{2+} entry,' is believed to be important in the refilling of depleted intracellular Ca^{2+} stores. Consistent with this role, capacitative Ca^{2+} entry has been observed to occur locally at sites where the endoplasmic reticulum is in proximity to the plasma membrane. Capacitative Ca^{2+} entry was initially believed to primarily involve channels in the class of transient potential channels. Subsequent studies, however, indicated that most store-operated channels are likely distinct from transient potential channels. The identity of these store-operated channels in glia, as in other cells, remains uncertain. Store-operated channels may be inhibited by dihydropyridines and activated by BayK 8644, compounds that had previously been considered to be specific to voltage-gated calcium channels (VGCCs). The lack of specificity of these compounds and the absence of consistent and specific inhibitors of capacitative calcium entry have been problematic in determination of the functional significance of capacitative calcium entry in glia. It is clear, however, that capacitative calcium entry can contribute substantially to the duration of a Ca^{2+} response that is triggered by an initial release of intracellular Ca^{2+}. Capacitative Ca^{2+} entry in astrocytes has been found to be inhibited by both arachidonic acid and zinc, suggesting that it may be actively modulated as a component of intercellular signaling. It has also been reported to be increased under hypoxic conditions, suggesting that it may play a role in the response of astrocytes to injury.

VGCCs

All glial cells express different types of VGCCs, and in each cell type, the number and type of channels expressed may vary with development and differentiation. Although they do not fire action

potentials, glial cells can show depolarization of membrane potential to a level that activates opening of voltage-gated channels. Astrocytes in culture express L-, N-, R-, and T-type calcium channels, but not P/Q-type channels. In cultured astrocytes, depolarization can evoke increases in $[Ca^{2+}]_i$ that are enhanced by BayK 8644 and inhibited by dihydropyridines. However, as mentioned above, these compounds may also modulate capacitative Ca^{2+} entry. In slice preparations, depolarization-evoked increases in astrocytic Ca^{2+} were found to be entirely generated by intracellular Ca^{2+} release activated by metabotropic glutamate receptors rather than by calcium influx through voltage-gated channels. These findings suggest that VGCCs do not contribute substantially to astrocyte calcium signaling under physiological conditions. However, different subtypes of Ca^{2+} channels have been reported to be upregulated in reactive astrocytes following brain injury, as well as pilocarpine-induced status epilepticus, suggesting that changes in the 'state' of astrocytes may result in an increased role for VGCCs.

Oligodendrocytes also express different types of VGCCs, whose level of expression may vary considerably with the stage of development and differentiation. Transient upregulation of R-type channels in oligodendrocytes has been reported during myelinogenesis. It has been suggested that VGCCs are important in the functional response of oligodendrocytes to axonal activity, and they may also be important in the setting of axonal injury.

Transient Receptor Potential Channels

Transient receptor potential (TRP) channels are calcium-permeable cation channels with polymodal activation properties. Although they have been widely studied in neurons and other cell types, there have been few reports of TRP channel expression in glia. One study implicated TRP channels in spontaneous Ca^{2+} oscillations in astrocytes. Others found a specific expression of a TRP channel subtype in oligodendrocytes and microglia. At this stage, however, the potential contributions of TRP channels to glial Ca^{2+} signaling, and their functional significance, have not been extensively characterized.

Mitochondria

Mitochondria may serve as both a buffer for increases in $[Ca^{2+}]_i$ in astrocytes and a source that contributes to Ca^{2+} signals. Studies using the mitochondria selective indicator rhod-2 indicate that mitochondrial Ca^{2+} increases and oscillates in synchrony with cytoplasmic Ca^{2+} in astrocytes stimulated with different ligands. Consistent with mitochondrial Ca^{2+} uptake, intercellular Ca^{2+} waves in astrocytes are associated with waves of mitochondrial depolarization measured using the indicator tetramethylrhodamine ethyl ester. Disrupting mitochondrial function with mitochondrial uncouplers has been reported to increase the rate of propagation of intercellular Ca^{2+} waves, suggesting that mitochondrial uptake serves as a buffering mechanism to reduce the propagation of increases in $Ca^{2+}{}_i$ concentration. In other conditions, disruption of mitochondrial function with uncouplers may decrease the Ca^{2+} response to specific ligands, indicating that mitochondrial Ca^{2+} release contributes to the response. Decreased mitochondrial buffering of Ca^{2+} may be associated with increased Ca^{2+} signaling in astrocytes from older animals, an effect that could be involved in changes in the function of astrocytes with aging.

Ligand-Evoked Calcium Signaling in Glial Cells

A wide variety of neurotransmitters, neuromodulators, growth factors, and cytokines can evoke Ca^{2+} signaling in glial cells. Most, if not all, transmitter receptors expressed on neurons are also expressed on some type of glial cell. However, glial cell Ca^{2+} signaling is predominantly evoked by metabotropic receptors (G-protein-coupled receptors that activate the formation of intracellular messengers) as opposed to ionotropic receptors (ligand-gated ion channels). While ionotropic receptors play specialized roles in some glial cell types, the majority of transmitter-evoked signaling responses identified thus far involve release of intracellular Ca^{2+} as opposed to Ca^{2+} influx. Of these, signaling by purinergic receptors (receptors for adenosine triphosphate (ATP), adenosine diphosphate, and adenosine) and glutamate receptors has received the most attention, in part because these receptors appear to play key roles in glial–glial, glial–neuronal, and glial–vascular signaling. But many other ligands also evoke Ca^{2+} signaling in glia through multiple distinct receptor systems.

The primary mechanisms for ligand-evoked Ca^{2+} signaling in glia are activation of phospholipase C (PLC) by Gq-protein-coupled receptors, phosphorylation of PLC by receptor tyrosine kinases, and calcium influx through ligand-gated channels. As described earlier, binding of ligands to metabotropic receptors, such as binding of ATP to purinergic receptors or glutamate to metabotropic glutamate receptors on glial cells, activates the G-protein subtype Gq, which in turn activates PLC, leading to the formation of IP_3. IP_3 then binds to IP_3Rs on intracellular membranes, releasing Ca^{2+} from membrane-bound

internal stores. Activation of G-protein-coupled receptors may evoke a single peak of increased $[Ca^{2+}]_i$ or sustained oscillations in $[Ca^{2+}]_i$. It has been suggested that the frequency and amplitude of these Ca^{2+} oscillations could enable frequency and amplitude modulation of intracellular and intercellular signaling. Localized Ca^{2+} responses to receptor activation in an individual cell can also be propagated throughout the cell as intercellular waves (see the section titled 'Intercellular calcium waves').

In addition to activation by G-proteins, PLC can be activated by phosphorylation by growth factors such as receptor tyrosine kinases. While most studies have focused on their downstream activation of nuclear transcription factors, Ca^{2+} signaling evoked by receptor tyrosine kinases may also be critical to generating a cell's functional response. The response of glioma cells to epidermal growth factor is a good example of this type of signaling. Binding of epidermal growth factor to its receptor results in dimerization of the receptor, which in turn leads to phosphorylation of PLC, causing formation of IP_3 and intracellular Ca^{2+} release. Unlike G-protein-coupled receptors, which evoke an immediate Ca^{2+} response, activation of receptor tyrosine kinases results in a $[Ca^{2+}]_i$ increase that may be delayed by up to a minute after exposure to the ligand. There may also be a sustained monophasic or oscillatory Ca^{2+} increase that continues for minutes after removal of the ligand, as well as a sustained increase in the extent of propagation of intercellular Ca^{2+} waves.

In glial cells, Ca^{2+} influx through ligand-gated ion channels such as *N*-methyl-D-aspartate (NMDA) receptors, or through voltage-gated Ca^{2+} channels in response to ligand-evoked depolarization, has been reported primarily in cells of the oligodendrocyte lineage. NMDA receptors have been found in oligodendrocyte progenitors and mature oligodendrocytes, as well as in intact myelin, where they are involved in oligodendrocyte Ca^{2+} responses to acute injury. Ca^{2+} influx through kainate receptors has been reported in Bergmann glia and may represent a mechanism by which they respond to the activity of Purkinje cells. Astrocytes and microglia express ionotropic purinergic (P2X) receptors, but in the case of astrocytes, attention regarding these receptors has been focused more on their potential role as a pathway for ATP release than as a pathway for Ca^{2+} influx (see the section titled 'Intercellular calcium waves').

Patterns of Calcium Signaling in Individual Cells

Ca^{2+} release and/or influx may occur at discrete sites within individual cells, resulting in 'microdomains' of increased $[Ca^{2+}]_i$. Localized increases in $[Ca^{2+}]_i$ may allow spatial discrimination of particular Ca^{2+}-dependent functions. There is growing evidence that multiple components of Ca^{2+} signaling machinery may be co-localized in discrete units with the cell. Independent Ca^{2+} signaling in individual processes has been demonstrated in astrocyte processes in Bergmann glia and hippocampal astrocytes in slice preparations. In hippocampal slices, Ca^{2+} transients in individual astrocyte processes were found to be independent of neuronal activity, whereas in Bergmann glia, they were evoked by activity of neighboring neurons. Similar localized Ca^{2+} responses may occur in cells of the oligodendrocyte lineage in response to axonal release of glutamate. Astrocyte end-feet also show highly localized Ca^{2+} signaling in response to neuronal activity or release of caged IP_3. Localized Ca^{2+} signaling in glial cells could represent a mechanism by which they can participate in cell-to-cell signaling with multiple spatially separate cells independently, with the potential for a coordinated response if the Ca^{2+} signal spreads throughout the entire cell. For example, Ca^{2+} signaling in individual processes of a single glial cell could serve to integrate input from multiple individual synapses or axons or, conversely, to convey signals to multiple neighboring neurons or vascular cells. Localized intracellular Ca^{2+} release within a single glial cell may be propagated as an intracellular wave throughout the cell. These intracellular Ca^{2+} waves are believed to be generated by diffusion of IP_3 and sequential release of Ca^{2+} from neighboring intracellular Ca^{2+} stores. Intracellular Ca^{2+} waves are commonly observed in glial cells in culture, as well as in brain slice preparations.

Intercellular Calcium Waves

In addition to Ca^{2+} changes occurring within individual cells, glial cells are capable of widespread intercellular communication of Ca^{2+} signals in the form of intercellular 'waves.' Intercellular Ca^{2+} waves in astrocytes were originally described as a delayed response to application of glutamate to astrocytes in culture. Glutamate evoked an initially simultaneous increase in $[Ca^{2+}]_i$ in a microscopic field of astrocytes, followed by multifocal intercellular Ca^{2+} waves that occurred throughout the field. Other studies showed that intercellular Ca^{2+} waves in astrocytes could be evoked by transient mechanical perturbation of the membrane of a single cell. Because it reliably evokes intercellular Ca^{2+} waves from a point source in a single cell, mechanical stimulation has been widely used as a method for evoking Ca^{2+} waves in glial cells and a variety of other cell types. Other conditions that evoke Ca^{2+} waves include local application of transmitters, electrical stimulation, or removal of

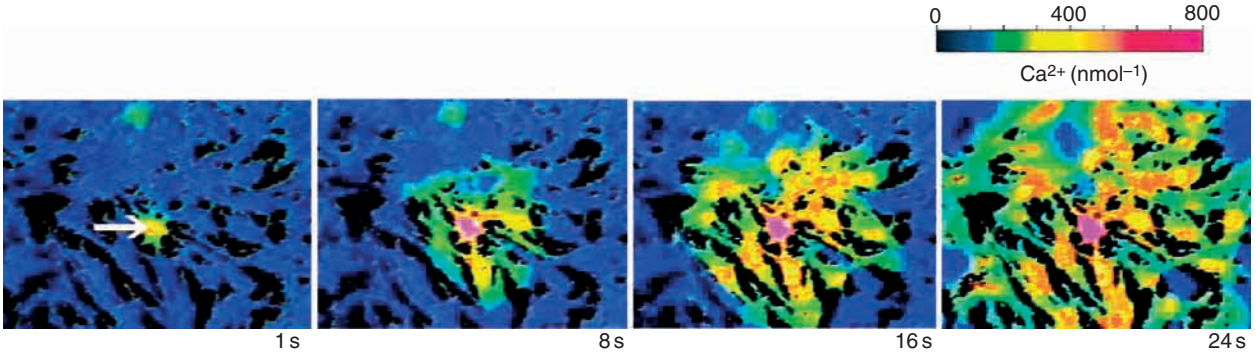

Figure 2 Intercellular calcium wave in glial cells. Image sequence showing intercellular calcium wave evoked by mechanical stimulation of a single cell in a mixed glial culture. Pseudocolor scale represents intracellular calcium concentration as measured by ratiometric fluorescence imaging with the indicator fura-2. Brief deformation of the membrane of a single cell (arrow) with a glass micropipette results in an increase in intracellular calcium concentration that is propagated as a wave to multiple surrounding cells.

extracellular Ca^{2+}. Limited intercellular Ca^{2+} waves may also occur spontaneously in purified glial cultures and to a much greater extent in glial cells grown in mixed culture with neurons, as well as in organotypic slice preparations. In organotypic hippocampal slice cultures, Ca^{2+} waves involving three to ten astrocytes occur spontaneously, whereas excitatory stimulation evokes broad curvilinear Ca^{2+} waves involving up to 100 cells. These Ca^{2+} wave fronts collide and annihilate and can form spiral patterns consistent with an excitable medium.

Initial observations suggested that calcium waves in glial cells are propagated from cell to cell via gap junctions (intercellular channels connecting the cytoplasm of adjacent cells). Calcium waves typically spread from cell to cell at specific sites of contact between neighboring cells, and the extent of intercellular spread of calcium waves was found to be correlated with the expression of connexin 43, the primary gap junction protein in astrocytes. However, subsequent studies indicated that intercellular calcium waves in astrocytes involve release of an extracellular messenger and activation of membrane receptors on neighboring cells. The primary messenger for intercellular Ca^{2+} waves has been found to be ATP. The apparently paradoxical involvement of connexins in intercellular Ca^{2+} waves mediated by release of ATP can be resolved by the observation that undocked connexin channels in individual cells, or 'connexin hemichannels,' can serve as a pathway for release of ATP from astrocytes. Other mechanisms of ATP release from astrocytes mediating intercellular Ca^{2+} waves, including P2X receptors, have been identified, suggesting that there may be multiple pathways for ATP release that are important for intercellular Ca^{2+} signaling. Connexin hemichannels, P2X receptors, and other potential pathways of ATP release are blocked by physiological concentrations of extracellular Ca^{2+} and Mg^{2+}. This can explain the observation that intercellular Ca^{2+} waves can be evoked by reduction of extracellular Ca^{2+}. The rate and extent of intercellular Ca^{2+} waves in astrocytes may depend, not only on the release of ATP, but also on the number and subtype of purinergic receptors that are expressed. See Figure 2.

Potential Functions of Glial Intercellular Ca^{2+} Waves

In a general sense, intercellular Ca^{2+} waves in glial cells enable a coordinated or spatially amplified multicellular response to a localized stimulus. Ca^{2+} waves may signal to neighboring glial cells of the same type, to different glial cell types, to neurons, or to vascular cells. One example is the potential role of intercellular Ca^{2+} waves in astrocytes in organizing a spatially propagated metabolic response. Astrocyte Ca^{2+} waves have been reported to drive parallel waves of increased intracellular Na^{+}. These Na^{+} waves, via glutamate-mediated glucose transport, in turn may trigger production and release of lactate, a primary metabolic substrate for neurons. This has been proposed as a mechanism by which a localized increase in neuronal activity could lead to a coordinated increase in delivery of metabolic substrates to surrounding cells. Another setting in which glial intercellular Ca^{2+} waves may be important is in the developing central nervous system. In the embryonic brain cortex, Ca^{2+} waves in radial glia have been reported to modulate neuronal migration. Intercellular Ca^{2+} waves in the retina may modulate the neuronal response to light, as well as mediate the vascular response to neuronal activity.

The extent to which intercellular Ca^{2+} waves occur in astrocytes *in vivo* is uncertain. Intercellular Ca^{2+} waves occur much more commonly in cell culture

than they do in brain slices. 'Physiological' stimulation in brain slices typically does not evoke cell-to-cell communication of Ca^{2+} responses in astrocytes, whereas Ca^{2+} waves are consistently observed in conjunction with the phenomenon of spreading depression (see the section titled 'Calcium signaling in cortical spreading depression and migraine') in brain slice preparations. It is therefore possible that under physiological conditions, astrocyte Ca^{2+} signaling involves very limited if any intercellular spread, whereas more-extensive propagation of Ca^{2+} waves may occur if a certain stimulus threshold is crossed.

Release of Signaling Molecules by Glial Calcium Signaling

One important function of Ca^{2+} signaling in glial cells is to trigger release of molecules that signal to other glial cells, neurons, or vascular cells. Calcium-dependent release of glutamate, ATP, D-serine (a modulator of NMDA receptor signaling), cytokines, and eicosanoids by astrocytes has been reported. While vesicular release of transmitters in the central nervous system had previously been considered to be exclusively a neuronal function, there is now extensive evidence for Ca^{2+}-dependent exocytotic vesicular release of glutamate by astrocytes. Astrocytes express all the necessary machinery for regulated vesicular release, including soluble N-ethylmaleimide-sensitive-factor attachment protein receptor proteins and vesicular transporters. Calcium-dependent plasma membrane fusion of vesicles has been visualized with total internal reflectance microscopy. A variety of studies indicate that vesicular release of glutamate may be a key mechanism for astrocyte–neuron signaling.

Mechanisms of Ca^{2+}-dependent release of ATP and D-serine by astrocytes have not been as clearly defined. As described earlier, release of ATP and glutamate through connexin channels, P2X receptor channels, or anion channels may be independent of changes in intracellular Ca^{2+} and may in fact depend to a greater extent on modulation of ATP release pathways by extracellular calcium than on changes in intracellular calcium. However, it is also possible that ATP and D-serine are released in vesicles, although this mechanism has not been as extensively characterized as it has for glutamate. In the case of eicosanoids, Ca^{2+} appears to trigger release of these molecules by activation of the phospholipase A2 pathway. Molecules produced by this pathway then have the capacity to diffuse directly through the plasma membrane into the extracellular space.

Ca^{2+} signaling may also be important for the release of signaling molecules by microglia. Ca^{2+}-dependent release of purines, cytokines, and nitric oxide by microglia has been demonstrated. Ca^{2+} may thereby serve as a critical trigger for the functional responses of activated microglia in the setting of inflammation and neuropathic pain.

Calcium Signaling in Astrocyte–Neuron Interaction

Ca^{2+} signaling plays a critical role in bidirectional glial neuronal signaling. Glial cells respond with increases in $[Ca^{2+}]_i$ to neuronal activity via neuronally released transmitters or possibly through changes in extracellular ionic concentrations associated with excessive neuronal activity. Conversely, increases in $[Ca^{2+}]_i$n in glial cells may evoke release of neurotransmitters, neuromodulators, or metabolic substrates that regulate neuronal synaptic strength, synchronization, or metabolic function. The pathway for Ca^{2+}-mediated glia–neuron signaling that has been most commonly reported in multiple preparations is astrocytic glutamate release activating synaptic or extrasynaptic neuronal glutamate receptors. But it is clear that other signaling molecules whose release is triggered by Ca^{2+} may also modulate neuronal activity. It is also possible that changes in extracellular K^+ evoked by Ca^{2+} increases in astrocytes may modulate the activity of neurons, as has been shown for vascular cells.

Calcium Signaling in Astrocyte–Vascular Interaction

The close proximity of astrocyte processes to both synapses and vascular cells, by virtue of their end-feet that ensheathe blood vessels, puts them in an ideal position to mediate changes in vascular tone that occur in response to neuronal activity. Multiple studies have now shown that calcium signaling in astrocytes evoked by either specific stimulation of astrocytes or by stimulation of adjacent neurons can result in either vasodilation or vasoconstriction. Intracellular Ca^{2+} waves along individual astrocyte processes leading to end-feet can be evoked by neuronal activity. This raises the possibility that a Ca^{2+} response initiated at a part of an astrocyte adjacent to a synapse could be spatially propagated as an intracellular wave to a different process contacting a blood vessel. In this way astrocytes could translate information about synaptic activity into changes in vascular tone, thus serving as a pathway for neurovascular coupling. Modulation of vascular tone may occur by multiple mechanisms, including calcium-dependent release of vasoactive messengers including eicosanoids, ATP, and nitric oxide, as well as calcium-activated K^+ channels that in turn control vascular tone by regulating local concentrations of extracellular K^+.

Calcium Signaling in Cortical Spreading Depression and Migraine

Spreading depression (SD) is a propagated electrophysiological wave that was described by Leão in 1944 as a direct current potential shift and depression of electroencephalogram (EEG) activity that spreads across the cortex at rates of 3–5 mm min^{-1}. In humans, SD has been demonstrated in the setting of brain injury. SD occurs in all gray matter regions of the central nervous system, although it is more readily provoked in some areas (e.g., cortex, CA1 region of hippocampus) than others. It can be evoked by mechanical stimulation, by electrical stimulation, or by local application of K^+, glutamate, or in some areas, acetylcholine. The temporal and spatial characteristics of intercellular Ca^{2+} waves in astrocytes are remarkably similar to those of spreading depression. In brain slice preparations, intercellular Ca^{2+} waves in astrocytes consistently occur in association with SD measured with either optical or electrophysiological approaches. In some reports, astrocyte Ca^{2+} waves preceded SD, whereas in others, Ca^{2+} waves occurred initially in conjunction with SD but extended beyond the point that SD decayed. In all reports, glial Ca^{2+} waves and SD could be pharmacologically dissociated, although inhibiting one affected the velocity and distance of propagation of the other. Thus, although pharmacological inhibition of one phenomenon does not exclude occurrence of the other, astrocyte Ca^{2+} waves and SD consistently occur together in the absence of pharmacological manipulation. One possibility is that they share common mechanisms for initiation and/or propagation but that additional factors specific to each individual process are required for their independent occurrence. And although they may not have a requisite interdependence, there are multiple potential mechanisms by which the two phenomena could interact with clinically important consequences.

Propagated waves of cortical activity and blood flow have been observed with positron emission tomography and magnetic resonance imaging techniques in patients with migraine. These waves have temporal and spatial characteristics that are similar to those of both astrocyte Ca^{2+} waves and spreading depression. A long-standing hypothesis is that the propagated changes in brain activity leading to the clinical migraine aura are caused by cortical SD (CSD). However, the classical EEG changes of CSD have never been directly demonstrated in migraine (possibly because of the lack of sensitivity of surface EEG). In addition, cortical waves of changes in blood flow in migraine patients can be seen in the absence of significant clinical symptomatology other than headache, an observation that is difficult to reconcile with a profound inhibition of neuronal activity, as occurs with CSD. One hypothesis is that intercellular Ca^{2+} waves in astrocytes, in the absence of the classical electrophysiological phenomenon of CSD, could underlie the propagated waves of cortical activity and blood flow in migraine.

Summary

The investigation of calcium signaling in glial cells has led to dramatic new information about the dynamic functions of glia, and in particular their active, bidirectional communication with neurons and vascular cells and resulting modulation of synaptic activity and vascular tone. The advance of *in vivo* imaging techniques promises to yield even more significant information about calcium signaling in glia and its roles in these and other physiological and pathological processes in the nervous system.

See also: Astrocyte: Calcium Signaling; Calcium and Signal Transduction; Calcium Waves in Glia; Calcium Waves: Purinergic Regulation; Gap Junction Communication; Glial Ion Homeostasis: A Fluorescence Microscopy Approach; Intracellular Calcium and Neuronal Death; Ionic Channels in Glia; Neurotransmitter Release from Astrocytes; Potassium Homeostasis in Glia.

Further Reading

Bernardinelli Y, Magistretti PJ, and Chatton J-Y (2004) Astrocytes generate Na^+-mediated metabolic waves. *Proceedings of the National Academy of Sciences of the United States of America* 101: 14937–14942.

Butt AM (2006) Neurotransmitter-mediated calcium signalling in oligodendrocyte physiology and pathology. *Glia* 54: 666–675.

Charles A (1998) Intercellular calcium waves in glia. *Glia* 24: 39–49.

Charles A (2005) Reaching out beyond the synapse: Glial intercellular waves coordinate metabolism. *Science's STKE: Signal Transduction Knowledge Environment* 270: pe6.

Färber K and Kettenmann H (2006) Functional role of calcium signals for microglial function. *Glia* 54: 656–665.

Fiacco TA and McCarthy KD (2006) Astrocyte calcium elevations: Properties, propagation, and effects on brain signaling. *Glia* 54: 676–690.

Haydon PG and Carmignoto G (2006) Astrocyte control of synaptic transmission and neurovascular coupling. *Physiological Reviews* 86: 1009–1031.

Metea MR and Newman EA (2006) Calcium signaling in specialized glial cells. *Glia* 54: 650–655.

Takano T, Tian GF, Peng W, et al. (2006) Astrocyte-mediated control of cerebral blood flow. *Nature Neuroscience* 9: 260–267.

Wang X, Lou N, Xu Q, et al. (2006) Astrocytic Ca^{2+} signaling evoked by sensory stimulation *in vivo*. *Nature Neuroscience* 9: 816–823.

Calcium Waves in Glia

B Haas and H Kettenmann, Max-Delbrück-Center for Molecular Medicine, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Astrocytes can communicate with each other by waves of Ca^{2+} increases, which propagate with a speed of 10–20 μm s^{-1} from cell to cell and can extend for distances up to 500 μm. This phenomenon was first noted in 1990 while studying cultured hippocampal astrocytes in response to application of glutamate. In parallel with the Ca^{2+} wave propagation, ATP and glutamate concentrations can increase in the extracellular space. In the years following the initial astrocyte study, the astrocyte wave has been intensely studied in cell culture. This article provides a review of the initiation of Ca^{2+} waves in astrocytes (different cultures, and organotypic and acute slices) and the intra- and intercellular mechanisms controlling the wave. It also summarizes the role of Ca^{2+} waves in diseases, the influence of astrocyte Ca^{2+} activity on other cells, and the impact during development. Some mathematical models for describing mechanisms of Ca^{2+} waves are introduced.

Since some issues remain controversial we find it important to point out that different labs have used different preparations in their studies. Astrocytes can vary with respect to brain area, development, or simply culture condition. We feel that it is likely that there is diversity with respect to functions and properties of astrocytes.

Ca^{2+} Waves Can Be Elicited in Astrocyte Populations by Several Types of Stimuli

Different stimuli are applicable to trigger a Ca^{2+} wave. Ca^{2+} waves spread in response to glutamate application, and a variety of other stimuli have been reported. Waves can be triggered by electrical or mechanical stimulation (e.g., by poking an astrocyte with a patch clamp pipette). Such procedures will even work in Ca^{2+}-free solution, although the stimulus might trigger a decrease in Ca^{2+} in the stimulated cell. The lowering of extracellular Ca^{2+} by itself can be effective to initiate waves. During pathologic events, extracellular Ca^{2+} can decrease by as much as tenfold. Also, a Ca^{2+} increase seems to be sufficient to trigger waves, since application of the Ca^{2+} ionophore, ionomycin, can elicit a Ca^{2+} wave. Moreover, there are several substances which can induce a propagating wave if locally applied. These include ATP and norepinephrine, as shown in rat cortical astrocytes. Albumin applied to rat cortical astrocytes triggers oscillation in individual cells which are initially unsynchronized but become synchronized after several minutes of application, and oscillations are observed to spread through the astrocyte network as a wave. The proapoptotic protein Bax evokes Ca^{2+} signals and wave propagation among cells. These Ca^{2+} changes are speculated to correlate with proapoptotic activity of Bax. Also, diffusible substances such as nitric oxide, a potent vasodilator, can induce waves when applied locally.

Mechanism of Propagation: Involvement of Purinergic Signaling

The initial observation that isolated astrocytes which had no physical contact with other astrocytes in cell culture participated in the Ca^{2+} wave indicated that a substance in the extracellular space could mediate the wave. Ca^{2+} waves can propagate across cell-free lanes up to 120 μm apart (Figure 1). Suramin, an antagonist for purinergic receptors, was found to reduce astrocytic Ca^{2+} signaling, thus the activating substance was soon identified as ATP. Moreover, it was observed that apyrase, an ATP-degrading ectoenzyme, attenuates astrocyte Ca^{2+} signaling. Culture medium collected from astrocyte cultures during, but not before, Ca^{2+} wave propagation, contained ATP and could evoke Ca^{2+} waves when applied to another culture. This led to the concept that release of ATP, with subsequent activation of purinergic receptors, determines the propagating mechanism. As studied for cultured spinal cord astrocytes, activation of P2Y1 purinoreceptors is sufficient for the propagation of the wave, but a second metabotropic purinergic receptor is also involved. While real-time bioluminescence imaging in rat cortical astrocytes points to a restricted, brief, and abrupt point source of ATP, which challenges the regenerative ATP release, studies with labeled adenine nucleotide pools in murine astrocyte cultures indicate that astrocytes are capable of ATP-induced ATP release, which supports a role for regenerative ATP release in Ca^{2+} wave propagation (Figure 2).

Mechanism of Propagation: Involvement of Gap Junctions

Astrocytes in culture are commonly coupled by gap junctions, and therefore astrocytes form a syncytium of cells. The gap junction blockers halothane and

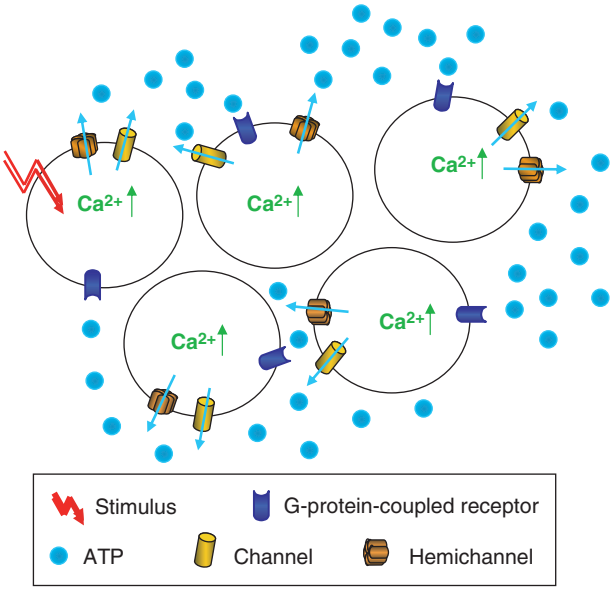

Figure 1 Model for Ca^{2+} waves due to regenerative ATP release. Release of ATP from the stimulated cell (red bolt) leads to activation of purinergic receptors on neighboring cells. These G-protein-coupled receptors stimulate via phospholipase C-activation inositol-3-phosphate production, which in turn triggers additional ATP release. ATP release can occur through hemichannels, P2X7 receptors, or anionic channels.

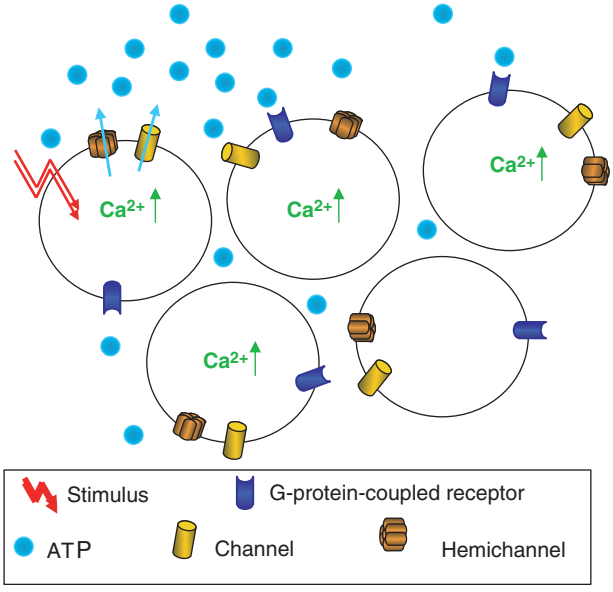

Figure 2 Model for Ca^{2+} waves due to a point source ATP release. ATP release occurs only from the stimulated cell. Passive diffusion activates purinergic receptors and thereby Ca^{2+} increases in neighboring cells.

octanol prevent intracellular waves from propagating between rat hippocampal astrocytes, but do not affect the velocity of Ca^{2+} spread within a given cell. Studies from cultured rat striatal astrocytes have led to the conclusion that the propagating mechanism is due to

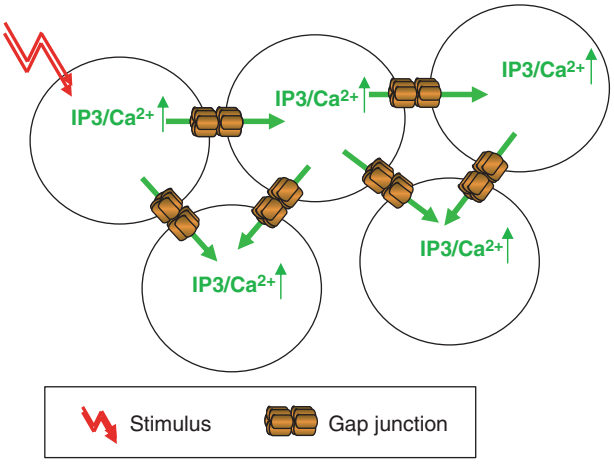

Figure 3 Model for wave propagation via gap junctions. Increased inositol-3-phosphate (IP3)/Ca^{2+} spreads from the stimulated cell via gap junctions to (coupled) neighboring cells, where it triggers regenerative mobilization of IP3/Ca^{2+}.

production of inositol phosphate and its propagation through gap junctions (Figure 3).

Interaction between Purinergic Signaling and Gap Junctions

Several studies indicate complex interactions between purinergic signaling and gap junctional communication in astrocytes. In cultured spinal cord astrocytes, genetic deletion of the major astrocyte gap junction protein, connexin 43, does not significantly affect the propagation of Ca^{2+} waves. Moreover, deficiency in connexin 43 markedly potentiates responses to ATP and UTP, indicating an autocrine communication between functional gap junctions and the purinergic signaling system. This change in sensitivity after deletion of connexin 43 is due to changes in the subtype composition of metabotropic purinergic receptors, shifting from an adenine-sensitive P2Y1 to a uridine-sensitive P2Y4 receptor.

Connexin expression could also influence the release of ATP. There are several hypotheses concerning how astrocytes release neurotransmitter. Besides the established release pathway via Ca^{2+}-dependent exocytosis, one possibility is the release through so-called hemichannels, which represent half of a gap junction. These hemichannels are not pathways between cells, but rather between cells and the extracellular space, similar to conventional ion channels. Hemichannels are thought to be activated by a decrease in extracellular Ca^{2+} concentration. In a study of murine astrocytes, a Ca^{2+} wave triggered by mechanical stimulation was inhibited by the connexin channel blocker flufenamic acid. Since the hemichannel activator quinine evoked

ATP release and Ca^{2+} signaling, the authors concluded that ATP is released through connexin hemichannels. In contrast, with cultured spinal cord astrocytes, in a study based on analyzing connexin 43-deficient and P2X7-deficient mice, ATP was not released via hemichannels, but via a large purinergic receptor, the P2X7 receptor. Moreover, the commonly used gap junction blockers carbenoxolone, octanole, and flufenamic acid may also affect P2X7 receptor signaling.

Nitric Oxide Signaling Can Mediate Ca^{2+} Waves

Based on studies of mixed glia–neuronal and hippocampal slice cultures, not only ATP but also nitric oxide (NO) release could be an important mediator for propagating Ca^{2+} waves. Application of the NO donor molsidomine, or puffing aqueous NO, evokes a local intercellular Ca^{2+} wave, even in the absence of extracellular Ca^{2+}. In contrast to waves triggered by mechanical or electrical stimulation, the NO-triggered wave depends on ryanodine-sensitive stores, indicating that Ca^{2+} originates from a different storage compartment, as compared to the previously described waves.

Intracellular Mechanisms Controlling the Wave

As already mentioned, Ca^{2+} waves can propagate in the absence of extracellular Ca^{2+}. The intracellular Ca^{2+} increase is important for wave propagation. By using different Ca^{2+} chelators with different Ca^{2+} affinities, the wave is attenuated in relation to the strength of the chelator. 2-Bis(*o*-aminophenoxy) ethane-*N*,*N*,*N*′,*N*′-tetraacetic acid (BAPTA), a chelator with high Ca^{2+} affinity, almost completely blocks the occurrence of Ca^{2+} waves. Most studies indicate that Ca^{2+} is released from the endoplasmic reticulum by activation of inositol-3-phosphate (IP3) receptors. Indeed, there is a very good correlation between the distribution of the IP3 receptors and sites with higher Ca^{2+} release kinetics (i.e., responses with higher amplitude and steeper rate of rise). The stores are filled by the thapsigargin-sensitive sarco/endoplasmic reticulum Ca^{2+}-ATPase (SERCA) pump and there exists a strong co-localization between IP3 receptors, SERCA pumps, and Ca^{2+} release sites. IP3-sensitive Ca^{2+} signaling involves the activation of protein kinase C and diacylglycerol (DAG), since Ca^{2+} dynamics have been shown to be linked to dynamics of protein kinase C and DAG translocation.

While most studies have confirmed that the thapsigargin-sensitive endoplasmic reticulum stores are the source for the cytoplasmic Ca^{2+} increase, the wave observed after application of the proapoptotic protein Bax is due to release of Ca^{2+} from mitochondria, and the NO-mediated wave is linked to ryanodine-sensitive stores.

Ca^{2+} Waves in Intact Tissue

All of the previously described experiments were done in culture. Cultured astrocytes are not a perfect model for healthy brain astrocytes. The increased levels of glial fibrillary acidic protein (GFAP) as compared to *in vivo* levels indicate that cultured astrocytes resemble reactive astrocytes as they occur after pathologic conditions. This is not much of a surprise, since culture conditions do not mimic the healthy brain. Moreover, culture conditions can influence the phenotype of astrocytes, and there is increasing evidence that astrocytes comprise a heterogeneous population within the intact brain. It is, therefore, essential to test and probe all of the different mechanisms in a more intact system, either brain slices or intact animals. First attempts to come closer to the *in vivo* situation involved the use of hippocampus slice cultures. In organotypic cultures from neonatal mice, Ca^{2+} waves involving hundreds of cells were observed to propagate in complex patterns, such as spiral wave fronts in the CA1 and CA2 regions. Similar to that reported for cultures, these waves were sensitive to thapsigargin and did not require the presence of extracellular Ca^{2+}. Studies in freshly isolated retinas from rats show that Ca^{2+} waves can be elicited by application of ATP and by electrical and mechanical stimulation, similar to the response in culture. Moreover, the wave propagates through a network of two glial cell types – namely, astrocytes and Müller cells. Interestingly, propagation in the astrocyte network employs a different mechanism, compared to propagation from the astrocyte to the Müller cell. While astrocyte-to-Müller cell signaling is very sensitive to purinergic antagonists or to the ATP-degrading enzyme apyrase, propagation within the astrocytes is much less affected. This observation has led to the conclusion that Ca^{2+} waves propagate in retina glial cells by two distinct mechanisms: while the wave propagating through astrocytes is dominated by diffusion of an internal messenger and requires gap junctions, the astrocyte-to-Müller signaling is primarily due to the release of ATP and activation of purinergic receptors.

Findings from two studies using slices from brain indicate diversity between astrocyte populations with respect to Ca^{2+} signaling. Using the corpus callosum as a model for the white-matter tract, it was demonstrated that the propagation of the wave among glial cells depends on a diffusible substance, since the

extent of propagation and the direction strongly depend on the perfusion of the slice. The ability of Reactive Blue 2 to block the response and the lack of effect of gap junction blockers indicate that the wave propagates due to ATP release and subsequent activation of purinergic receptors (Figure 4). This result is in contrast to findings from a study using cortex: in the cortical preparation, the propagation of the wave is strongly affected by gap junction blockers, but only weakly by purinergic receptor antagonists. The release of ATP could be monitored with 'sniffer cells': by seeding glioma cells with high-affinity purinergic receptors on the surface of the slice, the propagation of the ATP increase could be recorded as a Ca^{2+} signal. The involvement of gap junctions for the propagation of the wave is further supported by using mice deficient for the major astrocytic gap junction proteins, connexins 43 and 30. In this mouse line only local Ca^{2+} responses could be elicited in cortex and the wave fails to propagate over larger distances. Nevertheless, the ATP release monitored by sniffer cells was unaffected. This difference in Ca^{2+} signaling also finds its explanation in the distinct extent of electrical coupling between astrocytes in corpus callosum and cortex. When comparing biocytin injections in these two brain regions it became evident that astrocytes in cortex form large syncytia involving tens to hundreds of cells, while in the corpus callosum coupling can be absent or is restricted to a few cells (Figure 5). The extent of coupling does not necessarily predict the propagation mechanism: in hippocampus, where coupled and uncoupled astrocytes can be found, the propagation depends on a diffusible substance, since it is unaffected in mice deficient in connexins 30 and 43. It is, thus, possible that the conflicting data obtained in tissue culture may reflect the diversity of mechanisms expressed in the intact tissue. Recently, first reports of Ca^{2+} signaling in cortical astrocytes *in vivo* have been obtained. By imaging Ca^{2+} responses in astrocytes of superficial cortical layers using two-photon laser scanning microscopy, it was demonstrated that coordinated Ca^{2+} events can be recorded in anesthetized rats. These findings indicate that this Ca^{2+} activity also occurs in the live animal.

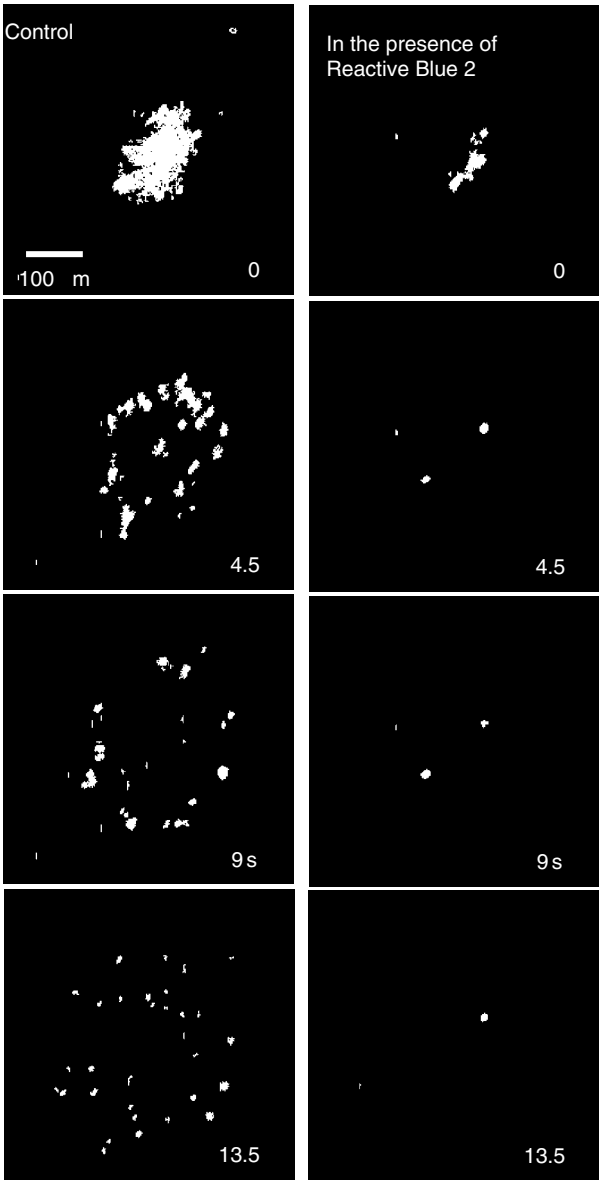

Figure 4 Propagation of an astrocytic Ca^{2+} wave in the corpus callosum acute slice. To illustrate the propagation of the Ca^{2+} wave, only those cells are displayed which were newly recruited to the population of Ca^{2+}-responsive cells. This was achieved by subtracting the previous image from the one which was displayed. The white areas represent the population of responsive cells at the times indicated. The first images show the cells responding close to the pipette briefly after stimulation (0 s). The left panel shows consecutive images, displaying the spread of the wave under control conditions. In the right panel only a local response was observed after applying the purinergic receptor antagonist Reactive Blue 2. Adapted from Schipke CG, Boucsein C, Ohlemeyer C, et al. (2002) Astrocyte Ca^{2+} waves trigger responses in microglial cells in brain slices. *FASEB Journal* 16: 255–257.

Control of the Wave by Receptors

A number of studies have indicated that the extent of propagation of the Ca^{2+} wave can be modulated by activating membrane receptors. Astrocytes have been shown to express essentially all known receptors for neurotransmitters and neuromodulators. This includes classical transmitter receptors and neuropeptide receptors, but also receptors for cytokines. The glucocorticoids methylprednisolone and dexamethasone potentiate long-range Ca^{2+} waves. Glucocorticoid treatment increases the resting cytosolic Ca^{2+} level and also doubles the extent and the amplitude of

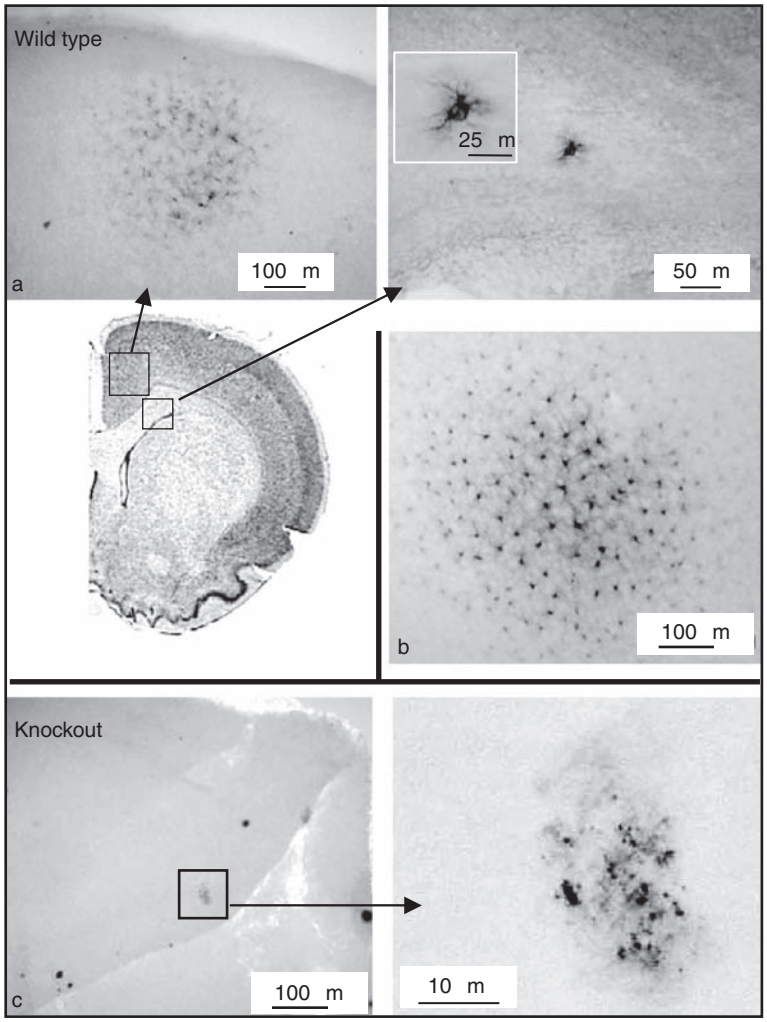

Figure 5 Astrocytic coupling in the neocortex and in the corpus callosum. (a) Biocytin was injected into a cell of the neocortex and a cell in the corpus callosum within the same slice from a glial fibrillary acidic protein/enhanced green fluorescent protein (GFAP/EGFP) transgenic mouse. Astrocytes were selected using their fluorescence at 488 nm. Left: In the neocortex, the biocytin injection resulted in the labeling of a large population of astrocytes. Right: In the corpus callosum, only a single cell is stained with biocytin. The insert on the right is a magnified image of the single injected cell. Images are from a 150 μm-thick slice that was processed as a whole after injection of the two cells. (b) One cell in the cortex of a slice from a GFAP/EGFP transgenic mouse was injected with biocytin. After fixation, 50 μm cryosections were obtained from this 300 μm-thick slice. This procedure resulted in an improved biocytin visualization as compared to processing the entire 150 μm-thick slice. (c) Similarly as described in panel a, biocytin was injected into an astrocyte identified by its membrane currents in a slice from a Cx30(−/−)/Cx43(fl/fl):GFAP-cre(+) mouse. The right image is a magnified view of the overview on the left. No biocytin-positive cells could be detected, but the processes of the patched cell can be visualized. The soma of the filled cell is not visible, probably because it is destroyed or removed when removing the patch pipette. From Haas B, Schipke CG, Peters O, et al. (2006) Activity-dependent ATP-waves in the mouse neocortex are independent from astrocytic Ca^{2+} waves. *Cerebral Cortex* 16: 237–246.

Ca^{2+} waves in whole brain astrocyte cultures as compared with untreated controls. In response to an inflammatory condition, interleukin-1β (IL-1β) is produced both by intrinsic glial cells and by invading cells from the immune system. This cytokine influences the spread of Ca^{2+} waves in primary human fetal astrocytes. Treating these cultures with IL-1β for 24 h leads to a potentiation of intracellular Ca^{2+} waves. Moreover, a shift from gap junctional communication to P2 receptor-mediated pathways is observed after the treatment. Anandamide, an endogenous arachidonic acid derivative, is seen to block Ca^{2+} waves in striatal astrocytes that propagate via gap junctions.

The encephalic and telencephalic astrocytes express melatonin receptors which are involved in circadian and seasonal rhythms. In studies of these astrocytes, a substantial difference between mouse and avian astrocytes is found – namely, in avian

astrocytes the wave is more restricted. Application of 10 μm melatonin causes a 32% increase in the area in which the wave spreads. In contrast, in telencephalic astrocytes, melatonin has no effect. All of the studies described herein have been performed in cell culture and may reflect a heterogeneity of potential responses *in vivo*.

Disease

Spreading depression is a phenomenon associated with the aura of migraine. It occurs in humans and animals. Spreading depression can be measured by imaging an intrinsic optical signal. It is correlated to a Ca^{2+} increase in astrocytes, which can be simultaneously recorded. In the hippocampus, it has been demonstrated that spreading depression is accompanied by an astrocyte Ca^{2+} wave and that the astrocyte Ca^{2+} wave can be abolished in Ca^{2+}-free, ethylene glycol tetraacetic acid (EGTA)-containing solution, while the spreading depression still occurs. This indicates that spreading depression triggers the astrocyte Ca^{2+} wave. The wave associated with spreading depression is distinct from the intrinsically propagating wave, since there is a clear difference in the speed of the propagation. Spreading depression propagates at about 30–50 $\mu m\,s^{-1}$, while the pure astrocyte wave propagates at about half that speed. Interestingly, in the cortex, spreading depression is also associated with the Ca^{2+} wave but often it is observed that the astrocyte Ca^{2+} wave propagates even further. When the astrocyte wave is no longer associated with cortical spreading depression, it drops in speed, to a value similar to that recorded for the intrinsic wave in cell culture or in the acute slice.

Alzheimer's disease is associated with a deposition of amyloid beta peptide, which is an important constituent of the plaques. Exposure of astrocytes to the amyloid beta peptide results in an increase in the amplitude and velocity of evoked Ca^{2+} waves and increases the area in which the wave spreads. While control waves require extracellular ATP for wave propagation, the increase induced by amyloid beta peptide is completely sensitive to an uncoupler of gap junctions, indicating that amyloid beta peptide modulates gap junctions and thereby influences the astrocyte Ca^{2+} wave.

Interaction with Other Cells

Astrocytic Ca^{2+} waves spread not only between astrocytes, but also lead to responses in other cell types of the nervous system, including microglia and oligodendrocytes, and also neurons, endothelial, and meningeal cells (Figure 6). As discussed previously, the astrocyte Ca^{2+} wave depends both on gap junctions and on ATP signaling. Both mechanisms are not confined to astrocytes. Actually, most cells in the brain express purinergic system receptors and thus should sense ATP released from astrocytes, which makes it likely that in the brain the astrocyte wave involves other cell types. Moreover, in parallel with the propagation, not only ATP, but also glutamate, can be released into the extracellular space. While ATP release seems Ca^{2+} independent, glutamate release depends on an increases in Ca^{2+}.

In co-cultures of neurons and glial cells, the astrocyte wave triggers a Ca^{2+} response in those neurons which are in close contact to astrocytes. One lab has reported that this neuron glia signaling is sensitive to the gap junction blocker octanol, indicating either gap junctions between neurons and astrocytes or an octanol effect on astrocytic signaling. In another study using hippocampal neurons grown on hippocampal or cortical astrocytes, the response of the neurons to the astrocyte Ca^{2+} signal was shown to be blocked by antagonists against *N*-methyl-D-aspartate (NMDA) and non-NMDA receptors, indicating that glutamate serves as a signaling substance.

In a study using slices containing the corpus callosum, it was evident that microglial cells and cells of the oligodendrocyte lineage also respond to an electrically elicited Ca^{2+} wave. While the response in the oligodendrocyte lineage cells could be recorded as a Ca^{2+} signal, this was not possible in microglial cells, since these cells do not take up ester dyes in slices. Therefore, the patch clamp technique was employed and a current response could be concommitantly recorded with the Ca^{2+} wave. This response consisted of a potassium conductance increase and was similar to the response of microglial cells to ATP.

Astrocytes are in an important strategic position concerning the coupling of neuronal activity to blood flow, since they enwrap blood vessels with their endfeet. Using co-cultures of primary astrocytes and endothelial cell lines as a model, it was demonstrated that Ca^{2+} waves induced by mechanical stimulation in a given cell can propagate from astrocytes to endothelial cells, and vice versa. The mechanism is similar to that shown for retina and corpus callosum – namely, release of ATP and subsequent activation of purinergic receptors on either astrocytes or endothelial cells. A similar interaction has been found in co-cultures of astrocytes with pia arachnoid cells, whereby Ca^{2+} activity is transmitted.

Ca^{2+} Waves during Development

Ca^{2+} waves have also been observed in radial glial cells during early development. These waves can

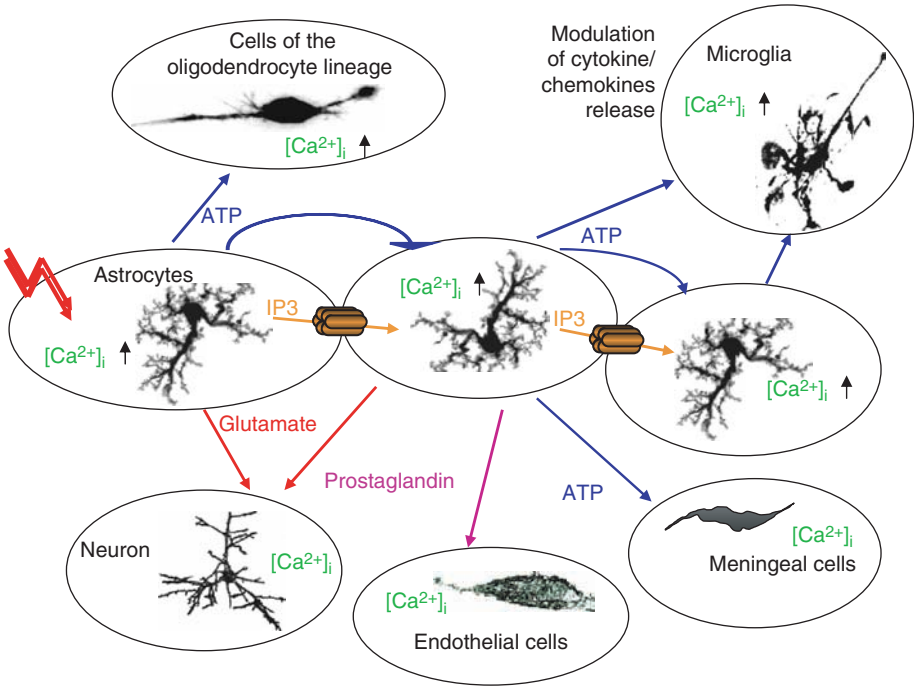

Figure 6 Schematic diagram illustrating intercellular communication between astrocytes and cells of the oligodendrocyte lineage, microglial cells, neurons, and endothelial and meningeal cells. Astrocytes can communicate intracellular Ca^{2+} signals to neighboring astrocytes via diffusion of the second messenger inositol-3-phosphate (IP3) through gap junctions or via the release of ATP. The communication between astrocytes and cells of the oligodendrocyte lineage and microglia is due to ATP release, which activates purinergic receptors. Cytokine release from microglia is modulated by ATP. As indicated, also astrocytes can release a number of substances in response to an elevation of $[Ca^{2+}]_i$. Modified from Kettenmann H and Schipke C (2004) Calcium signaling in glia. In: Hatton G and Parpura V (eds.) *Glia–Neuronal Signaling*, p. 303. Boston: Kluwer Academic Publishers.

occur spontaneously and depend on connexin hemichannels, purinergic receptors, and IP3 release. Disrupting the wave results in a decrease of cortical ventricular zone proliferation.

Modeling Astrocyte Ca^{2+} Waves

Several models have been developed for a mathematical description of astrocyte Ca^{2+} waves. One model is based on activation of purinergic receptors by ATP, which triggers the release of ATP in an autocrine manner, resulting in increasing intracellular Ca^{2+}. This model predicts the dependence of velocity on the type of P2Y receptor and the delay of the Ca^{2+} wave to the ATP wave. Another model relies on the spread via gap junctions as the essential mechanism based on observations in rat striatal astrocytes. This model predicts that the spatial range of the wave strongly depends on both signal initiation (i.e., activation of IP3) and regeneration of the IP3 signal, and accordingly the term 'limited regenerative signaling' is proposed. The crucial permissive factor is the gap junction permeability for IP3, and hence heterogeneity in gap junction composition can account for preferential pathways of wave propagation. The refractory period strongly depends on the refilling kinetics of the Ca^{2+} stores. In a third model, specialized areas which serve as amplification loci are proposed. They are considered to be areas of endoplasmatic reticulum containing higher density of IP3 receptors.

See also: Activity-Dependent Metabolism in Glia and Neurons; Astrocyte: Calcium Signaling; Astrocyte: Neurotransmitter and Hormone Receptors; Calcium Waves: Purinergic Regulation; Cell–Cell Communication Through the Extracellular Space; Gap Junctions and Hemichannels in Glia; Glia Control of Blood Flow; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging.

Further Reading

Arcuino G, Lin JH, Takano T, et al. (2002) Intercellular Ca^{2+} signaling mediated by point-source burst release of ATP. *Proceedings of the National Academy of Sciences of the United States of America* 99: 9840–9845.

Basarsky TA, Duffy SN, Andrew RD, et al. (1998) Imaging spreading depression and associated intracellular calcium waves in brain slices. *Journal of Neuroscience* 18: 7189–7199.

Charles A (2005) Reaching out beyond the synapse: Glial intercellular waves coordinate metabolism. *Science's Signal Transduction Knowledge Environment* 2005(270): pe6.

Cornell-Bell AH, Finkbeiner SM, Cooper MS, et al. (1990) Glutamate induces Ca^{2+} waves in cultured astrocytes: Long-range glial signaling. *Science* 247: 470–473.

Cotrina ML, Lin JH, and Nedergaard M (1998) Cytoskeletal assembly and ATP release regulate astrocytic calcium signaling. *Journal of Neuroscience* 18: 8794–8804.

Finkbeiner S (1992) Ca^{2+} waves in astrocytes – filling in the gaps. *Neuron* 8: 1101–1108.

Guthrie PB, Knappenberger J, Segal M, et al. (1999) ATP released from astrocytes mediates glial calcium waves. *Journal of Neuroscience* 19: 520–528.

Haas B, Schipke CG, Peters O, et al. (2006) Activity-dependent ATP-waves in the mouse neocortex are independent from astrocytic Ca^{2+} waves. *Cerebral Cortex* 16: 237–246.

Hatton GI and Parpura V (eds.) (2004) *Glia–Neuronal Signaling*. Boston: Kluwer Academic Publishers.

Hirase H, Qian L, Bartho P, et al. (2004) Calcium dynamics of cortical astrocytic networks *in vivo*. *PLoS Biology* 2(4): e96 [online; doi:10.1371/journal.pbio.0020096].

Kettenmann H, and Ransom BR (eds.) (2005) *Neuroglia*. New York: Oxford University Press.

Nedergaard M (1994) Direct signaling from astrocytes to neurons in cultures of mammalian brain cells. *Science* 263: 1768–1771.

Newman EA (2001) Propagation of intercellular Ca^{2+} waves in retinal astrocytes and Müller cells. *Journal of Neuroscience* 21: 2215–2223.

Perea G and Araque A (2005) Glial Ca^{2+} signaling and neuron–glia communication. *Cell Calcium* 38(3–4): 375–382.

Peters O, Schipke CG, Hashimoto Y, et al. (2003) Different mechanisms promote astrocyte Ca^{2+} waves and spreading depression in the mouse neocortex. *Journal of Neuroscience* 23: 9888–9896.

Scemes E, Suadicani SO, and Spray DC (2000) Intercellular communication in spinal cord astrocytes: Fine tuning between gap junctions and P2 nucleotide receptors in Ca^{2+} wave propagation. *Journal of Neuroscience* 20: 1435–1445.

Schipke CG, Boucsein C, Ohlemeyer C, et al. (2002) Astrocyte Ca^{2+} waves trigger responses in microglial cells in brain slices. *FASEB Journal* 16: 255–257.

Venance L, Piomelli D, Glowinski J, et al. (1995) Inhibition by anandamide of gap junctions and intercellular calcium signalling in striatal astrocytes. *Nature* 376: 590–594.

Volterra A, Magistretti PJ, and Haydon PG (eds.) (2002) *The Tripartite Synapse*. Oxford: Oxford University Press.

Calcium Waves: Purinergic Regulation

M L Cotrina and M Nedergaard, University of Rochester, Rochester, NY, USA

© 2009 Published by Elsevier Ltd.

Introduction

Nonelectrical cells rely on Ca^{2+} signaling as the key mode of communication with neighboring cells. This signaling mechanism is most relevant in astrocytic cells, the most abundant glial cell in the nervous system. Using Ca^{2+} waves, astrocytes can affect neuronal activity, and findings from numerous studies point to an important role for Ca^{2+} waves in regulating synaptic function. This form of communication occurs when Ca^{2+} signals that originate in one cell can spread over long distances and affect groups of cells, evoking intercellular Ca^{2+} waves. In cultured astrocytes, Ca^{2+} waves can be evoked by receptor activation (glutamate, ATP, or endothelin-1), by lowering of extracellular Ca^{2+}, or by intense neuronal firing in cultured organotypic tissue. A general feature of Ca^{2+} waves is that they propagate with a velocity of $\sim 20\,\mu m\ s^{-1}$ and expand over a maximal radius of $\sim$300–400 μm, thereby engaging 30–60 cells in a single wave. Despite controversies with regard to the mechanism of propagation, it is clear that Ca^{2+} waves require activation of phospholipase C (PLC), which produces inositol 1,4,5-trisphosphate (IP_3). In turn, IP_3 mobilizes Ca^{2+} from intracellular stores, resulting in further activation of PLC.

Current Models for the Propagation of Astrocytic Ca^{2+} Waves

Astrocytic Ca^{2+} signaling was originally proposed to depend on IP_3 and/or Ca^{2+} diffusion through gap junction channels, intercellular channels formed by proteins called connexins (model 1). Those signaling molecules were thought to promote subsequent Ca^{2+} increases in neighboring cells. This idea was supported by the experimental observation that C6 glioma cells, which lack gap junction proteins, allow the propagation of long-distance Ca^{2+} waves (>200 μm) only after transfection and overexpression of connexin 43 (Cx43). Overexpression of Cx32 in C6 glioma and HeLa cells later confirmed these observations.

This model was questioned by a study reporting that astrocytes with no physical contact with other astrocytes could also propagate Ca^{2+} waves, suggesting the existence of an extracellular mechanism for their propagation. This diffusible messenger was later identified as ATP, and parallel studies showed that ATP can mediate Ca^{2+} waves not only in astrocytes, but also in osteoblasts, liver, epithelial cells, and heart. This model for Ca^{2+} wave propagation (model 2) states that Ca^{2+} waves are regenerative and mediated by ATP release. Later evidence showed that astrocytes can, indeed, exhibit ATP-induced ATP release, but in a Ca^{2+}-independent manner, supporting the idea that Ca^{2+} waves can be regenerative, but not Ca^{2+} dependent for ATP release.

A third model challenged the idea of regenerative ATP release because Ca^{2+} waves never expand more than 300–400 μm beyond the point of origin and are therefore not truly regenerative. In fact, real-time bioluminescence imaging of extracellular ATP failed to support the concept of regenerative ATP release during Ca^{2+} wave propagation. So, model 3 states instead, that ATP is consistently released by a single cell, or from a point source only. A Ca^{2+} wave would thus be propagated by diffusion of the initially released ATP, rather than by sequential ATP release from other cells. Neighboring cells would, in turn, increase their intracellular Ca^{2+} in response to P2Y receptor activation, but would not actively participate in wave propagation with the Ca^{2+}-dependent release of additional ATP. This is consistent with studies performed with simultaneous monitoring of ATP and intracellular Ca^{2+} from astrocytic cultures. In this system, it has been observed that, upon mechanical stimulation, the Ca^{2+} wave is synchronized with the released ATP. However, extracellular ATP is the main mediator of intercellular Ca^{2+} wave propagation in a Ca^{2+}-independent manner, as abolishing Ca^{2+} increases does not preclude ATP release.

Mechanisms of Purinergic Release: Exocytosis versus Connexin Hemichannels

Several groups have confirmed that long-distance Ca^{2+} waves are mediated by purinergic signaling, but the mechanism by which ATP is released is still controversial. It has been speculated that connexin hemichannels could mediate ATP release, because several cell lines expressing Cx43, Cx32, or Cx26 all release five- to 20-fold more ATP than do mock-transfected, Cx-deficient control cells. In addition, gap junction hemichannels have a pore diameter of 10–12 Å and are permeable to large anions, including ADP and ATP. So, connexin proteins seem to regulate the propagation of Ca^{2+} waves not by allowing the passage of an intracellular messenger, but rather by potentiating ATP release. Later studies have documented that ATP release occurs concomitantly with an

increase in whole-cell current, supporting the concept of channel-mediated ATP release. The challenge, therefore, has been to test whether open connexin hemichannels can act as direct conduits for the efflux of cytosolic ATP.

It has been reported that astrocytes in culture express Ca^{2+}-activated currents that are sensitive to flufenamic acid (FFA), an agent that inhibits gap junction permeability. Mechanical stimulation that initiates Ca^{2+} waves promotes dye uptake and dye release, as well as ATP release, an effect that is also inhibited by FFA and potentiated by quinine, a hemichannel activator. However, an important criticism for some of the evidence in support of hemichannel-mediated release of ATP is that hemichannel opening probability would be small because an open channel this large would be deleterious to cells. In addition, most studies have tried to disclose hemichannel opening in the absence of Ca^{2+}, or in very low Ca^{2+} concentrations, conditions that are difficult to reconcile with normal cell function. So, what is the relevance of connexin hemichannels in normal physiology?

Some answers have come from the use of small mimetic peptides that correspond to portions of the connexin proteins. One of these peptides will only bind extracellular portions of the connexin protein when freely exposed and not engaged in the formation of a gap junction channel, and will, in addition, block hemichannel permeability. Using this tool together with dye uptake, it has been demonstrated that, in the retinal pigmental epithelium (RPE), ATP released by the spontaneous opening of gap junctional hemichannels promotes the propagation of Ca^{2+} waves and controls proliferation in the underlying retina. Most importantly, this study showed that calcium waves do not require ATP-induced ATP release, supporting the idea that ATP is released through connexin hemichannels in the trigger cell alone and diffuses passively to purinergic receptors in neighboring cells.

The use of specific peptides that block hemichannels but not proper gap junctions has demonstrated that increases in intracellular Ca^{2+} up to 500 nM can also trigger ATP release in a connexin-dependent manner. These studies are the first to demonstrate that ATP release through connexin hemichannels can happen under physiological conditions.

Data from cellular fractionation experiments have suggested that ATP is released by Ca^{2+}-dependent exocytosis of ATP-containing vesicles. It seems that these vesicles may be distinct from glutamate-containing vesicles, indicating that there may be separate exocytotic pathways for ATP and glutamate release, which may be differentially regulated. It has been also found that glia from the sciatic nerve exhibit ATP-mediated ATP release via exocytotic vesicles, release that can be regulated by the cystic fibrosis transmembrane conductance regulator.

It is possible that astrocytes can release ATP by several mechanisms, but additional studies are needed to establish the coexistence of these mechanisms, which are not necessarily mutually exclusive, and their implications for Ca^{2+} wave signaling.

The P2X7 Pore-Forming Receptor and mCa^{2+} Signaling

The family of purinergic receptors is composed of two types, ionotropic (P2X) receptors and metabotropic (P2Y) receptors. It is known that astrocytes express P2Y1, P2Y2, and P2Y4 and P2X7 receptors, and it is likely that several purinergic receptors contribute to wave propagation in astrocytes.

Astrocytes depend on P2Y1 and P2Y2 receptors for the propagation of Ca^{2+} waves. Overexpression studies have shown that the properties of intercellular Ca^{2+} waves depend on the particular P2Y receptor subtype involved, differentially affecting the efficacy, velocity, amplitude, range, and 'saltatory' nature of the Ca^{2+} wave propagation.

Among P2X receptors, P2X7 is particular in that it forms large ion channels in response to ATP or other ligands and is present in many cell types, including astrocytes. P2X7 receptors can mediate the release of excitatory amino acids, like glutamate and aspartate, disclosing a novel, ligand-stimulated, nonvesicular astrocytic route for glutamate release. P2X7 receptors are activated by high concentrations of ATP, do not desensitize, and, more importantly, depend on extracellular Ca^{2+} ions.

As P2X7 activation can promote Ca^{2+} signals in optic nerve glia and Ca^{2+} waves among bone cells, an alternative hypothesis that could explain ATP-mediated astrocytic Ca^{2+} waves is that P2X7 receptors, and not connexin hemichannels, are the main pathway for the release of ATP. Using spinal cord astrocytes from Cx43 and/or P2X7 receptor-deficient mice, it has been recently found that, contrary to previous ideas, P2X7 receptors, and not connexin proteins, could be responsible for intercellular Ca^{2+} wave amplification when cells are exposed to solutions with low concentrations of divalent cations. Interestingly, the same study reported that many of the compounds that block connexin hemichannels, and that have been used to document the involvement of hemichannels as ATP-releasing pathways, are also antagonists of the P2X7 receptor, raising the question of the true contribution of connexin hemichannels in ATP release and Ca^{2+} wave propagation. Nonetheless, it is worth noting that the absence of P2X7

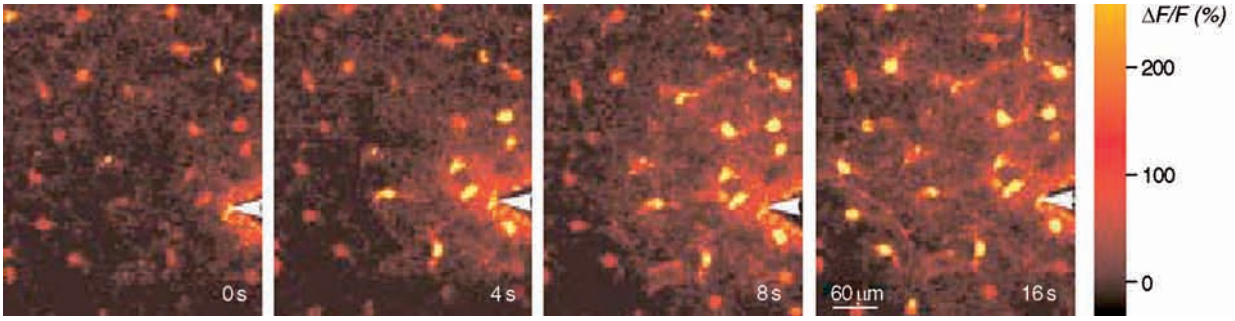

Figure 1 An astrocytic Ca^{2+} wave evoked by focal electrical field stimulation in a hippocampal slice. The white arrowhead indicates the position of the stimulating electrode. The slice was loaded with the fluorescent Ca^{2+} indicator dye Fluo-4/AM for 1 h before two-photon laser scanning microscopy. The electrical stimulation evoked a Ca^{2+} wave that engaged multiple neighboring astrocytes expanding over a radius of 150–250 µm ($\Delta F/F$, fluorescence ratio).

receptor in the model used in this study did not prevent the occurrence of astrocytic Ca^{2+} waves that were similar, in radius and number of cells engaged in the wave, to what was seen under normal conditions (Figure 1). So, the development of agents that can be used specifically to block one pathway versus the other is clearly needed to elucidate the exclusive or overlapping roles of these ATP-releasing pathways in astrocytes.

As P2X7 receptors are not easily activated, their expression is low or absent in some brain regions, and low Ca^{2+} conditions are mostly attained under pathological conditions, it has been suggested that these receptors could be major contributors of Ca^{2+} signaling during inflammation, ischemia, or other pathological states. Support for this idea comes from different fronts. First, ATP release and astrocytic Ca^{2+} signaling are both triggered by traumatic injury. Second, injury is also associated with a decrease in extracellular Ca^{2+}, which enhances ATP release and Ca^{2+} signaling and increases the affinity of the P2X7 receptor. Third, P2X7 receptor can mediate cell death. On the basis of these findings, the contribution of the P2X7 receptors in the amplification of cell damage has been analyzed in a model of spinal cord injury. Using a novel bioluminescence technique in the intact, live animal, it has been found that the peritraumatic area after acute spinal cord injury exhibits unusually high levels of ATP release, which is accompanied by the subsequent cell death of spinal neurons that are highly immunoreactive for P2X7 receptors. Further studies in this model are needed to establish the source of ATP release during the injury period.

Purinergic Signaling and Ca^{2+} Waves in Neuronal Physiology

Spontaneous occurrence of ATP-dependent Ca^{2+} waves has been observed in radial glia of brain slices during early cortical development. Disruption of these waves altered cell proliferation in the ventricular zone, indicating that glial Ca^{2+} waves could control neuronal production in early development. It has been subsequently shown that glutamate release from neurons induces ATP release from astrocytes. This release causes homo- and heterosynaptic depression in co-cultures of hippocampal cells and in CA1 neurons of hippocampal slices. This signaling pathway seems to be sensitive to membrane ectonucleotidases, is dependent on Ca^{2+} and soluble *N*-ethylmaleimide-sensitive attachment factor receptor (SNARE) activity, suppresses excitatory synaptic transmission, and modulates synaptic plasticity. An important observation is the concomitant, adenosine-mediated heterosynaptic depression of neighboring, unstimulated pathways, placing astrocytes at a central position in the coordination of synaptic networks.

Astrocytic Ca^{2+} elevations that influence extrasynaptic receptors through the release of glutamate had been previously described. In this case, spontaneous, Ca^{2+}-dependent glutamate release from a single astrocyte was able to promote *N*-methyl-D-aspartate (NMDA) receptor responses in a synchronized manner in multiple neurons, indicating that an astrocyte can simultaneously affect the activity of many neurons that may not be directly connected through chemical synapses. Astrocytic Ca^{2+} events in response to glutamate have been recently described for first time *in vivo* upon sensory stimulation.

In summary, it is clear that, in the adult brain, nonsynaptic transmitter release originates preferentially from astrocytes, is Ca^{2+} dependent, can be spontaneously detected, has slow dynamics, and has differential effects, depending on the nature of the transmitter. In the case of ATP/adenosine, spillover of synaptic glutamate or extrasynaptic glutamate seems to initiate a negative feedback loop through a purinergic pathway. In the case of glutamate, a single

astrocyte has the ability to simultaneously affect the activity of many neurons that may or may not be part of the same neuronal circuit.

Involvement of Adenosine as Purinergic Mediator of Astrocytic Ca^{2+} Signaling

Adenosine has been traditionally considered an inhibitor of neuronal activity and a regulator of cerebral blood flow (Figure 2). The cellular origin and mechanisms of adenosine release have not been established, but it is clear that astrocytes can release adenosine, since activation of astrocytes is responsible for heterosynaptic depression in hippocampal slices. Massive release of purines, including adenosine, occurs after metabolic stress and trauma. It has therefore been hypothesized that ATP/adenosine may be more important during pathological states, because there is more energy demand (as during seizures) or because there is loss of metabolites (as during ischemia), conditions that promote a reduction in the ATP levels and an increase in adenosine and other adenine nucleotide levels. In support of this hypothesis, it has been found that astrocytes release ATP/adenosine under ischemic conditions.

Because extracellular ATP does not accumulate *in vivo* for long periods, as it is rapidly degraded by membrane ectonucleotidases to produce adenosine, it is plausible that purinergic Ca^{2+} signaling plays a critical role during pathological events through the actions of adenosine. Additional studies are needed to clearly elucidate the contribution of ATP and adenosine in both normal and pathophysiological events.

Conclusions

The existence of ATP-mediated astrocytic Ca^{2+} waves is well documented. However, the mechanisms for ATP release are still highly controversial and the proposed pathways of ATP release include hemichannels, vesicular release, P2X7 receptors, and ABC transporters. However, it is important to keep in mind that astrocytes are multifunctional cells, and therefore more than one mechanism of ATP release may exist. For example, the important task of potassium buffering is undertaken by several potassium channels expressed by astrocytes, including KIR4.1 and rSlo K (Ca), but also by the K^+–Na^+–Cl^- cotransporter. It is possible that different types of stimulation might activate different pathways of ATP release. Regional and developmental differences are also likely to play a role.

Future advances in the field depend on the development of pharmacological agents with minimal

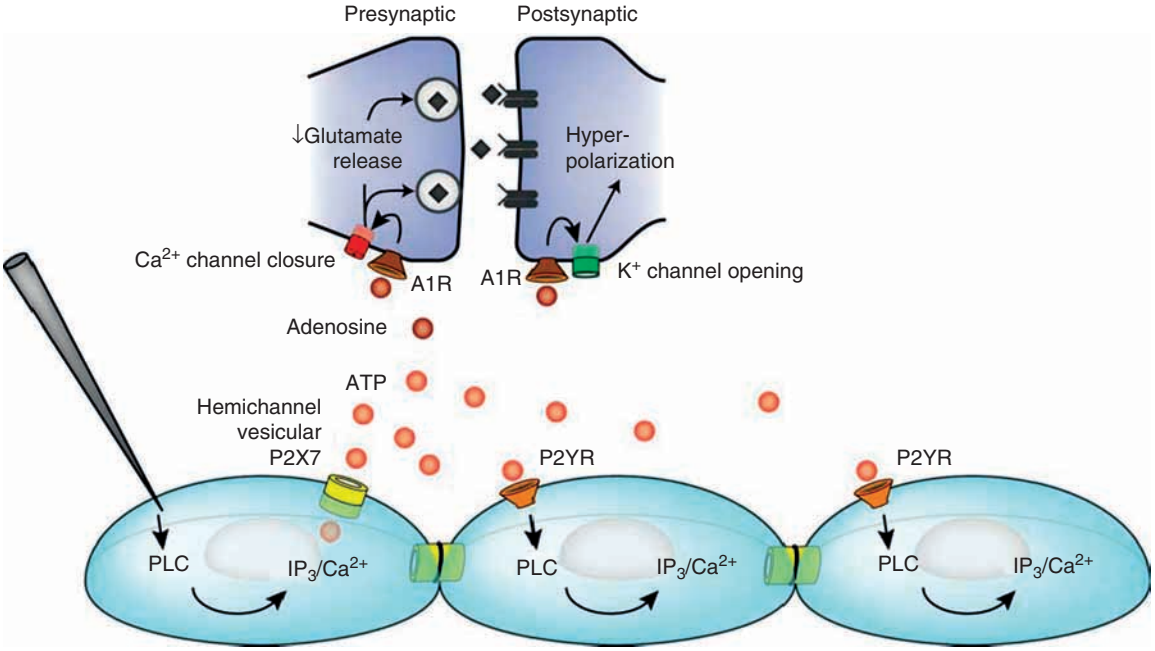

Figure 2 Current model for the mechanism and actions of astrocytic purinergic Ca^{2+} signaling. During the propagation of a Ca^{2+} wave (bottom), a single astrocyte releases ATP by a mechanism that could involve connexin hemichannels, exocytotic vesicles, or the pore-forming P2X7 receptor. ATP then propagates by diffusion to neighboring astrocytes, connected by gap junctions. These astrocytes will, in turn, increase their Ca^{2+} levels through the action of the P2Y receptor (P2YR) but do not release ATP. Astrocytic adenosine, either directly released during Ca^{2+} wave propagation or converted from ATP through the action of ectonucleotidases, can affect surrounding neurons (top) by binding to A1 receptors (A1R). This action can cause heterosynaptic depression that suppresses excitatory synaptic transmission by decreasing glutamate release and activating hyperpolarizing K^+ channels, establishing a negative feedback loop on the neurons through a purinergic pathway (PLC, phospholipase C; IP_3, inositol 1,4,5-trisphosphate).

cross-reactivity among the different release pathways to determine their relative contributions. To achieve this goal, it will also be necessary to use imaging techniques that allow the visualization of ATP dynamics and electrophysiological recordings of channel conductance, as the sensitivity of biochemical analysis versus. imaging may also lead to dissimilar results. Another issue is that cultured astrocytes represent a poor model system for defining the signaling mechanisms and functional roles of astrocytes in intact brain. Astrocytes lose their polarized phenotype *in vitro* and may activate, in cultures, pathways of ATP release that do not exist *in vivo* or have a minor role. For example, culturing of astrocytes has been associated with a marked upregulation of SNARE proteins that are involved in vesicular cycling.

Future research will also need to define the involvement of purinergic signaling in both physiological and pathological conditions. An extensive literature documents the neuroprotective actions of adenosine and it will be interesting to define the contribution of astrocytes to neuronal survival in the setting of ischemia. Astrocytic purinergic signaling may represent a powerful target for future drug development that may decrease secondary stroke injury.

See also: Astrocyte: Calcium Signaling; Calcium Homeostasis in Glia; Calcium Waves in Glia; Gap Junctions and Hemichannels in Glia; Gap Junctions and Electrical Synapses; Gap Junction Abnormalities and Disorders of the Nervous System; Gap Junction Communication; Gap Junctions: Metabolic Exchanges; Glial Ion Homeostasis: A Fluorescence Microscopy Approach; Ionic Channels in Glia; Neurotransmitter Release from Astrocytes; Purinergic Receptors; Purines and Purinoceptors: Molecular Biology Overview.

Further Reading

Arcuino G, Lin JH, Takano T, et al. (2002) Intercellular calcium signaling mediated by point-source burst release of ATP. *Proceedings of the National Academy of Sciences of the United States of America* 99: 9840–9845.

Bennett MV, Contreras JE, Bukauskas FF, et al. (2003) New roles for astrocytes: Gap junction hemichannels have something to communicate. *Trends in Neurosciences* 26: 610–616.

Cotrina ML, Lin JH, Alves-Rodrigues A, et al. (1998) Connexins regulate calcium signaling by controlling ATP release. *Proceedings of the National Academy of Sciences of the United States of America* 95: 15735–15740.

Dunwiddie TV and Masino SA (2001) The role and regulation of adenosine in the central nervous system. *Annual Review of Neuroscience* 24: 31–55.

Fellin T, Pascual O, Gobbo S, et al. (2004) Neuronal synchrony mediated by astrocytic glutamate through activation of extrasynaptic NMDA receptors. *Neuron* 43: 729–743.

Hassinger TD, Guthrie PB, Atkinson PB, et al. (1996) An extracellular signaling component in propagation of astrocytic calcium waves. *Proceedings of the National Academy of Sciences of the United States of America* 93: 13268–13273.

James G and Butt AM (2002) P2Y and P2X purinoceptor mediated Ca^{2+} signaling in glial cell pathology in the central nervous system. *European Journal of Pharmacology* 447: 247–260.

Kang J, Jiang L, Goldman SA, et al. (1998) Astrocyte-mediated potentiation of inhibitory synaptic transmission. *Nature Neuroscience* 1: 683–692.

Nedergaard M (1994) Direct signaling from astrocytes to neurons in cultures of mammalian brain cells. *Science* 263: 1768–1771.

Parpura V, Basarsky TA, Liu F, et al. (1994) Glutamate-mediated astrocyte-neuron signaling. *Nature* 369: 744–747.

Pascual O, Casper KB, Kubera C, et al. (2005) Astrocytic purinergic signaling coordinates synaptic networks. *Science* 310: 113–116.

Pearson RA, Dale N, Llaudet E, et al. (2005) ATP released via gap junction hemichannels from the pigment epithelium regulates neural retinal progenitor proliferation. *Neuron* 46: 731–744.

Sanderson MJ, Charles AC, Boitano S, et al. (1994) Mechanisms and function of intercellular calcium signaling. *Molecular and Cellular Endocrinology* 98: 173–187.

Suadicani SO, Brosnan CF, and Scemes E (2006) P2X7 receptors mediate ATP release and amplification of astrocytic intercellular Ca^{2+} signaling. *Journal of Neuroscience* 26: 1378–1385.

Wang X, Arcuino G, Takano T, et al. (2004) P2X7 receptor inhibition improves recovery after spinal cord injury. *Nature Medicine* 10: 821–827.

Wang X, Lou N, Xu Q, et al. (2006) Astrocytic Ca^{2+} signaling evoked by sensory stimulation *in vivo*. *Nature Neuroscience* 9: 816–823.

Zhang J, Wang H, Ye C, et al. (2003) ATP released by astrocytes mediates glutamatergic activity-dependent heterosynaptic suppression. *Neuron* 40: 971–982.

Calcium-Calmodulin Kinase II (CaMKII) in Learning and Memory

M N Waxham, The University of Texas Medical School, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Calcium is a simple divalent metal ion that has the distinction of directly or indirectly regulating almost every process in neurons. The highly reactive nature of calcium necessitates that cells maintain low levels (50–100 nM) of this toxic ion, and pumps, exchangers, and buffers all subserve this function. The opening of *N*-methyl-D-aspartate (NMDA) receptors or the opening of voltage-gated calcium channels leads to a rapid influx of calcium due to the large chemical gradient for this ion. While elevations in calcium can be quite varied, local concentrations reaching 10–100 μM have been estimated. Calcium can also be released from intracellular stores through activation of inositol 1,4,5-trisphosphate (IP_3) receptors or ryanodine receptors. The resulting temporal and spatial profiles of the calcium signal can thus be quite complex. These signals are decoded by the intracellular biochemical machinery into biological effects (neurotransmitter release, regulation of neurotransmitter biosynthesis, regulation of receptor and channel properties, and induction of synaptic plasticity). Many of the second messenger effects of calcium are mediated through the ubiquitous protein calmodulin. Calmodulin binds four calcium ions, inducing a conformational change that permits the protein to interact with and activate target enzymes. As such, calmodulin has no enzymatic activity; its function is to integrate the calcium signal and transduce it to other downstream enzymes.

This article focuses on one of the major classes of calcium/calmodulin-activated enzymes, the calcium/calmodulin-activated protein kinases (CaM-kinases). A discussion of the general characteristics of these enzymes is presented, followed by more detailed descriptions of individual members. One useful classification of the members of the CaM-kinase family is to consider whether they can phosphorylate a wide range of protein substrates (multifunctional) or whether they have a single or narrow set of protein substrates (substrate specific). This classification is adopted in this article with a discussion of the multifunctional kinases CaM-kinase kinase (CaMKK), CaM-kinase I (CaMKI), CaM-kinase IV (CaMKIV), and CaM-kinase II (CaMKII)) first, followed by a discussion of the substrate-specific kinases (CaMKIII and myosin light-chain kinase (MLCK)).

General Characteristics of Calcium/CaM-Kinases

All members of the CaM-kinase family catalyze the covalent transfer of a phosphate group from ATP to Ser or Thr residues in protein substrates, and, as the name implies, their phosphotransferase activity is initiated by the binding of calcium/calmodulin. The overall domain structure is very similar among family members (Figure 1), and their general characteristics are summarized in Table 1. The phosphotransferase reaction requires that ATP and protein substrates are bound simultaneously and that the amino acids involved in catalyzing the phosphotransferase reaction are properly oriented. Available data suggest that the catalytic domains of all members of the CaM-kinase family are architecturally similar. Two globular domains are evident; one that binds ATP and another that binds protein substrates, with a crevice in between the domains serving as the site for the phosphotransferase reaction (see Figure 2). Inactive CaM-kinases exist in a quiescent state through an autoinhibitory mechanism whereby a portion of the molecule interacts with the catalytic domain. In some cases, the autoinhibitory domain lies across the site for substrate binding and blocks access to the catalytic domain. In other cases, the autoinhibitory domain interacts with amino acids that prevent the proper formation of a catalytically competent active site for the phosphotransferase reaction. Sometimes a combination of these two mechanisms is evident. In either case, calcium/calmodulin binding relieves this inhibition, activating the enzyme. Several members of the CaM-kinase family are substrates for Ser/Thr phosphorylation and, in some cases, this phosphorylation is a prerequisite for activity (CaMKI and CaMKIV). In other cases, phosphorylation of the enzyme can alter its biological properties. For example, phosphorylation of CaMKII is critical for the enzyme's role in synaptic plasticity and learning and memory.

Multifunctional CaM-Kinases

CaM-Kinase Kinase

CaMKK exists in neurons and cells as a monomer of approximately 55–65 kDa and as two isoforms, α and β. CaMKK is held in an inactive state by its autoinhibitory domain, and binding of calcium/calmodulin relieves this inhibition, leading to phosphorylation of its substrates. The two best known substrates for CaMKK are CaMKI and CaMKIV, and phosphorylation of these substrates is necessary for their activation. Binding of calcium/calmodulin is

Figure 1 Line drawings identifying the domain structure of calcium/calmodulin (CaM)-kinases. The catalytic domains are shown in blue, the regulatory domains are shown in red, and the calmodulin-binding domains are shown in green. CaM-kinase II (CaMKII) is an oligomer and contains a domain necessary for oligomerization (yellow). CaMKIII (eEF-2K; elongation factor-2 kinase) has a recognition domain for eEF-2 in its C-terminal domain (orange), and its catalytic domain (hatched blue) bears little resemblance to other CaM-kinases. The actin-binding domain of smooth muscle myosin light-chain kinase (smMLCK) is shown as a white box. CAMKK, CaM-kinase kinase; skMLCK, skeletal muscle myosin light-chain kinase. Redrawn from Robison J and Colbran RJ (2004) Calcium/calmodulin-dependent protein kinases. In: Lennarz WJ and Lane MD (eds.) *Encyclopedia of Biological Chemistry*, vol. 1, pp. 281–286. Oxford: Elsevier.

Table 1 Physicochemical properties of the CaM-kinase family

Enzyme	*Structure*	*Subunit mol. mass (kDa)*	*Substrate specificity*
CaMKI	Monomer	42	Multifunctional
CaMKII	Dodecamer	50–68	Multifunctional
CaMKIII (eEF-2K)	Monomer	100	Restricted
CaMKIV	Monomer	65–67	Multifunctional
CaMKK	Monomer	55–65	Multifunctional
skMLCK	Monomer	67–150	Restricted
smMLCK	Monomer	130–210	Restricted

an important prerequisite for calcium/calmodulin-activated CaMKK to phosphorylate CaMKI or CaMKIV. The site on CaMKI and CaMKIV for CaMKK phosphorylation is either inaccessible or not in the appropriate conformation in the absence of calcium/calmodulin. CaMKK is also a substrate for phosphorylation. Protein kinase A (PKA) phosphorylates CaMKK, thereby inhibiting its activity. This inhibitory phosphorylation site is in CaMKK's calmodulin-binding domain and phosphorylation prevents calmodulin binding. CaMKK lies at the beginning of a signaling cascade, so inhibition of its activity through PKA-mediated phosphorylation also leads to suppression of CaMKI and CaMKIV activity.

CaM-Kinase I

CaMKI is a monomeric kinase of approximately 42 kDa (Table 1), and several isoforms of CaMKI have been described. It was the first of only two members of the CaM-kinase family whose crystal structure was determined (CaMKII was the second; see Figure 3). The crystallographic data permit a detailed understanding of how interactions between the autoinhibitory domain and the catalytic domain lead to suppression of activity. The autoinhibitory domain makes contacts with amino acid residues in both the ATP- and substrate-binding lobes, preventing access of these reactants to the catalytic site. Calcium/calmodulin binds to a domain adjacent to the autoinhibitory domain and in so doing removes the autoinhibitory domain from the catalytic site, permitting substrate binding. Calcium/calmodulin binding is also necessary to expose a specific Thr residue in the activation loop that serves as the substrate for CaMKK phosphorylation. When CaMKI is phosphorylated on this site, its activity rises by 10- to 20-fold but remains entirely calcium/calmodulin – dependent, even when phosphorylated.

CaMKI is found in both the cytosol and the nucleus and presumably has diverse functions. One of its best characterized substrates is synapsin 1. Synapsins are a family of vesicle-associated proteins of which the proposed function is to tether synaptic vesicles to the actin cytoskeleton within the presynaptic terminal, and several kinases have been identified that phosphorylate synapsin 1. Phosphorylation of synapsin 1 by CaMKI causes a decrease in the affinity of synapsin 1 for the vesicle, leading to increased mobility of the vesicle in the presynaptic terminal (Figure 4). CaMKI was also shown to modulate growth cone motility, and myosin II was identified as a substrate for the enzyme. Activation of CaMKI is also involved in activation of the extracellular signal-regulated kinase (ERK) signaling pathway that can lead to ERK-dependent long-term potentiation (LTP). Activity-dependent nuclear localization has also been shown for the δ isoform of CaMKI, further expanding the potential roles of CaMKI to include the regulation of gene transcription.

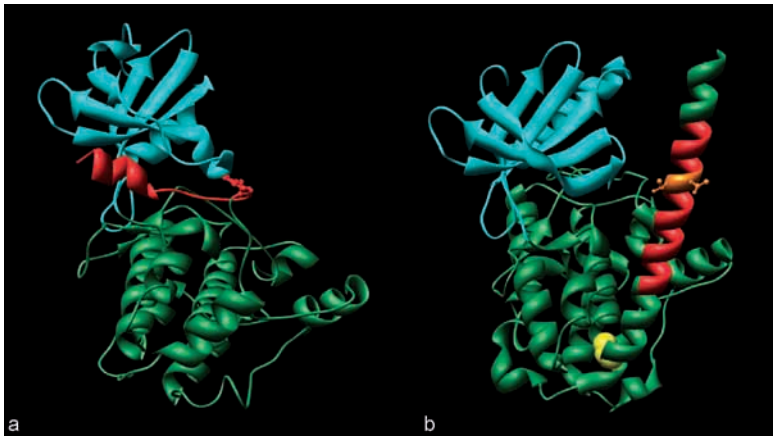

Figure 2 Ribbon diagrams showing the structure of the catalytic and regulatory domains of CaMKI and CaMKII. There is significant (>60%) amino acid homology between CaMKI and CaMKII across these domains. (a) The bilobed structure of CaMKI. The ATP-binding site is formed by the cyan domain, and the substrate-binding site is formed by the green domain. The catalytic cleft where the phosphotransferase reaction occurs lies between the two domains. The red domain represents the CaM-binding site, which lies across the catalytic cleft. Interactions of residues surrounding the CaM-binding domain with those in the two lobes of the catalytic domain maintain CaMKI in the inactive state. (b) The bilobed structure of the CaMKII catalytic domain. The colors of the domains represent features analogous to those described for CaMKI in panel (a). The autophosphorylation sites of CaMKII are also highlighted. The location of Thr286 is highlighted in yellow near the bottom of the structure. The location of Thr305 and Thr306 is shown in orange within the CaM-binding domain (in red). The ribbon diagrams were created in Chimera software and the protein database (PDB) files were obtained from the National Center for Biotechnology Information database.

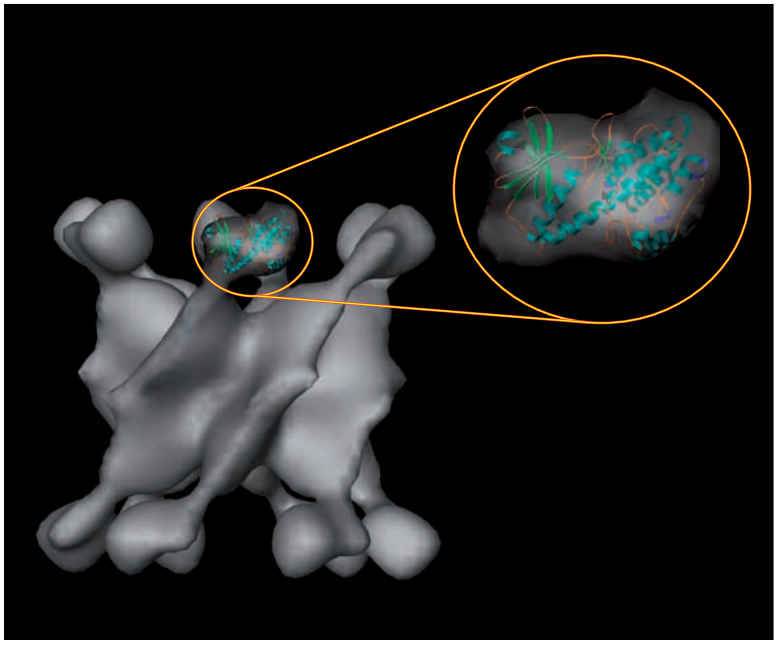

Figure 3 Structure of CaMKII. CaMKII is a multisubunit complex that is unique in structure among the CaM-kinase family. There is an association domain located in the C-terminus of the subunit (see **Figure 1**) that directs the formation of a 12-subunit complex. The association domains form a central core and 12 foot-shaped catalytic domains can be seen extending away from the core on stalklike features. The architectural arrangement of the subunits was determined using cryo-electron microscopy and 3-D particle reconstructions and is visualized as a surface-rendered structure. A ribbon diagram of the crystal structure of the CaMKII catalytic domain is docked into one of the feet. A zoomed-in side view of one of the feet, showing the predicted orientation of the crystal structure within the envelope, is shown on the right side of the figure. The figure was rendered in Explorer software and the crystal structure was obtained from the National Center for Biotechnology Information database.

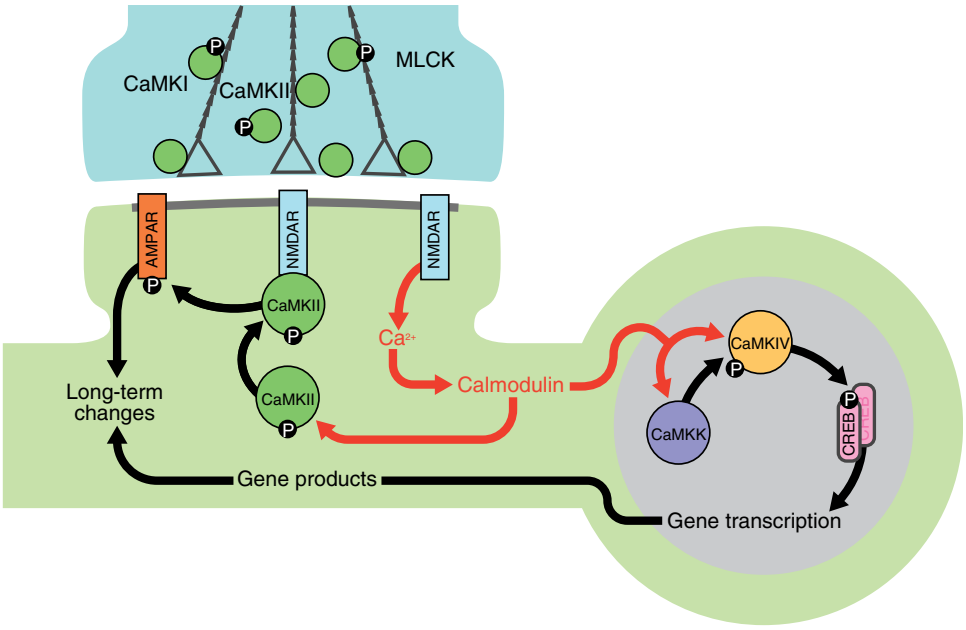

Figure 4 CaM-kinase modulation of neuronal functions. A presynaptic terminal contains transmitter-filled vesicles (green circles), some of which are associated with the actin cytoskeleton. Increased calcium, via voltage-activated calcium channels, and activation of the calmodulin pathway occur in the presynaptic terminal (not shown), leading to the activation of CaM-kinases and the regulation of tethering and transport of synaptic vesicles to the actin cytoskeleton. Phosphorylation of the synaptic vesicle-associated synapsin 1 (not shown) by CaMKI and CaMKII can release the vesicle from the actin cytoskeleton, making it available for exocytosis. Phosphorylation of myosin by myosin light-chain kinase (MLCK) is thought to facilitate the transport of vesicles within the presynaptic terminal. The membrane of the postsynaptic cell contains *N*-methyl-D-aspartate (NMDA) and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) subtypes of glutamate receptors. Calcium influx through calcium channels, such as the NMDA receptors, leads to the activation of calmodulin with the subsequent activation of the CaM-kinases. In the postsynaptic spine, CaMKII can autophosphorylate and can also phosphorylate AMPA receptors, potentiating their response to released glutamate from the presynaptic terminal. Increased calcium-bound calmodulin can similarly lead to the activation of CaMKK to phosphorylate and activate CaMKIV in the nucleus, which can subsequently phosphorylate substrates such as the transcription factor cAMP-response element-binding protein (CREB), leading to increased gene transcription. Redrawn and modified from Robison J and Colbran RJ (2004) Calcium/calmodulin-dependent protein kinases. In: Lennarz WJ and Lane MD (eds.) *Encyclopedia of Biological Chemistry*, vol. 1, pp. 281–286. Oxford: Elsevier.

CaM-Kinase IV

CaMKIV is found in neurons as a monomer of approximately 65 kDa, and its basic domain structure is similar to the other CaM-kinases (Table 1 and Figure 1). Analogous to CaMKI, activation of CaMKIV requires the binding of calcium/calmodulin and the phosphorylation of its activation loop by CaMKK. Once activated, CaMKIV undergoes an autophosphorylation reaction in a Ser/Thr-rich region of its N-terminus, and this intrasubunit autophosphorylation reaction is required for maximal activity. Once autophosphorylated, CaMKIV exhibits sustained activity even if calcium/calmodulin is removed from the enzyme. Thus, CaMKIV can exist in an activated state that outlives the calcium signal that initiated the autophosphorylation reaction. CaMKIV also autophosphorylates within its calmodulin-binding domain and this prevents calmodulin from rebinding to the enzyme. This autophosphorylation event would desensitize CaMKIV to subsequent activation by calcium/calmodulin until phosphatase activity removed phosphate from this site.

Activation of CaMKIV is linked to changes in gene transcription. CaMKIV is highly enriched in the nucleus and its best characterized substrates are nuclear transcription factors. These include cAMP response element-binding protein (CREB), serum response factor (SRF), and CREB-binding protein (CBP), among others. The phosphorylation of CREB by CaMKIV is the best characterized event in the nucleus and this occurs on Ser133, a site known to activate the transcriptional properties of CREB. The activation of CaMKIV is considered an important link between the transitions of short-term plasticity in neurons and long-term plasticity (Figure 4). The latter requires gene transcription and protein synthesis, and the activation of CaMKIV is thought to play an important step in linking the increased calcium signal to increased gene transcription. The sustained activation of CaMKIV, possibly mediated

through its calcium/calmodulin-independent activity, appears to be necessary for upregulation of gene transcription.

CaM-Kinase II

There are four genes in mammals that encode four distinct isoforms of CaMKII, α, β, γ, and δ which range in molecular mass from 50 to 68 kDa (Table 1). CaMKII is unique in structure among the family of CaM-kinases in that it exists in neurons as an oligomeric complex. The oligomer contains 12 catalytic domains that extend out from a central core (Figure 3) assembled through interactions between association domains that reside in the C-terminus of each subunit (Figure 1). The formation of oligomeric complexes relates to the high probability of CaMKII undergoing an important autophosphorylation reaction that regulates its properties (discussed later).

CaMKII holoenzyme assembly can occur between the different subunit isoforms, and the composition of the oligomer can dictate the enzyme's biological behavior. For example, the β subunit, and not the α subunit, contains an actin-binding domain that tethers CaMKII onto the actin cytoskeleton. If the β subunit is in a holoenzyme composed mostly of α subunits, the β subunit directs that assembly to the actin cytoskeleton. The association of βCaMKII with the actin cytoskeleton is decreased by binding of calcium/calmodulin. Thus, intracellular calcium increases lead to release of CaMKII into the cytoplasm by freeing it from the actin cytoskeleton and making it available to phosphorylate other protein substrates.

When activated by calcium/calmodulin, CaMKII subunits within the holoenzyme can phosphorylate each other in a process called autophosphorylation. There are parallels between constraints for the CaMKK phosphorylation of CaMKI or CaMKIV and this intersubunit phosphorylation of CaMKII. To be phosphorylated by CaMKK, CaMKI and CaMKIV must be in the calcium/calmodulin-bound form. This is also true for the autophosphorylation of subunits within the CaMKII oligomer; to be an effective substrate for autophosphorylation, the subunit within the CaMKII oligomer must be calcium/calmodulin bound. The subunit responsible for catalyzing the phosphorylation event must also be active through calcium/calmodulin binding. Because the subunits are physically linked within the holoenzyme, the probability of this autophosphorylation event happening is much greater than in the case of CaMKK and CaMKI or CaMKIV, which are constrained by their access to each other and diffusion-limited reaction steps.

One particular autophosphorylation site on the α isoform, Thr286, has received wide attention and, when autophosphorylated, has profound impacts on CaMKII activity. The location of Thr286 is identified in yellow in the ribbon diagram of the catalytic domain shown in Figure 2(b). Autophosphorylation increases the binding affinity of calcium/calmodulin by a factor of over 1000. This affinity is so tight that the process is referred to in the literature as 'calmodulin-trapping' to reflect the fact that dissociation of calmodulin slows significantly. Autophosphorylation also leads to a state of the kinase that continues to phosphorylate substrates even when calmodulin is removed. This is referred to as 'calcium/calmodulin-independent' or 'autonomous' activity. CaM trapping and autonomous activity sustain the activity of autophosphorylated CaMKII well beyond the duration of the calcium increase that triggered autophosphorylation. There are other sites of autophosphorylation on CaMKII. Another of significance is a site containing two adjacent Thr residues within the calmodulin-binding domain (Thr305 and Thr306 of αCaMKII) that, when phosphorylated, prevents calmodulin from binding. These sites are identified as orange residues in the ribbon diagram shown in Figure 2(b). This is analogous to the autophosphorylation that occurs in the calmodulin-binding domain of CaMKIV that also prevents calmodulin binding. Because the subunit no longer binds calmodulin, it cannot be active and phosphorylate other subunits or serve as a substrate for the autophosphorylation reaction. This sequence of events has been called 'inhibitory autophosphorylation.' However, if a subunit is phosphorylated at Thr286 and then at Thr306/306, it remains active because of the autonomous activity expressed when Thr286 is phosphorylated.

These autophosphorylation-induced changes in CaMKII properties have led to several hypotheses regarding CaMKII's role in neurons. The most prominent incorporate autophosphorylation and include the ability of CaMKII to decode calcium oscillations and sustain its activity well beyond the calcium signal. Because autophosphorylation of CaMKII requires that two calmodulin molecules are bound to adjacent subunits (one to activate kinase and one to expose substrate), during a brief calcium increase there is only a small probability that this arrangement would happen. As calcium oscillations increase in frequency, the probability of calcium/calmodulin binding to two adjacent subunits increases and autophosphorylation occurs. Recall that once a subunit is phosphorylated at Thr286, it 'traps' calmodulin because of the decreased off-rate, leading to an even greater likelihood of adjacent subunits being calcium/calmodulin bound and additional autophosphorylation.

The importance of CaMKII autophosphorylation in modifying neuronal processes has been studied extensively. Some of the most compelling data on

this topic come from studies in mice in which point mutations that were introduced into CaMKII prevent autophosphorylation of Thr286 (by replacing Thr286 with an Ala residue in the mouse germ line). These mice appear remarkably normal in the laboratory setting, but they show significant defects in forms of plasticity, such as LTP, and are impaired in a variety of learning and memory tasks. A similar approach was also taken to evaluate the role of the inhibitory phosphorylation sites (Thr305 and Thr306) that reside in the calmodulin-binding domain. Mice were created in which Thr residues were replaced with Val and Ala, which cannot be phosphorylated, or in which Thr306 was replaced with an Asp residue to 'mimic' phosphorylation through its negative charge. Blocking inhibitory autophosphorylation increased the amount of CaMKII associated with the postsynaptic density and lowered the threshold for inducing LTP. The Asp306 mice showed decreased CaMKII in the postsynaptic density, and both LTP and learning were blocked in these mice.

CaMKII is remarkably concentrated in neurons and can constitute up to a couple of percent of total brain protein, leading to the hypothesis it might serve a structural role. In fact, a picture is emerging for CaMKII as a scaffolding protein that serves in part to help organize proteins at neuronal synapses into molecular assemblies. CaMKII has been shown to interact directly with subunits of the NMDA receptor; α-actinin (an actin-binding protein); densin-180, a synaptic junction protein; and others. Because of CaMKII's oligomeric structure, one holoenzyme can bind to a number of different proteins simultaneously, providing the enzyme with a scaffolding function. Binding in a number of cases is modulated either through calcium/calmodulin activation of the kinase or through the subsequent process of autophosphorylation. These protein–protein interactions are thus driven through neuronal activity that leads to changes in intracellular calcium. These processes can be visualized in live neurons by tracking the activity-dependent translocation of CaMKII tagged with green fluorescent protein, which has been shown to drive CaMKII into synapses both in neuronal culture models and in living animals.

CaMKII phosphorylates a wide variety of substrates, including the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor, the NMDA receptor, nitric oxide synthase, and SynGAP, a synaptic GTPase-activating protein that modulates Ras activity. One of particular interest is the AMPA subtype of glutamate receptor, because AMPA receptors are responsible for producing currents that underlie glutamate-mediated synaptic transmission. Phosphorylation of AMPA receptors by CaMKII can be detected following stimuli that produce LTP, and this phosphorylation increases the single-channel conductance of AMPA receptors. Phosphorylation of a diverse family of protein substrates ties CaMKII activation directly into the regulation of the network of signaling proteins that control synaptic plasticity and neuronal excitability.

Substrate-Specific CaM-Kinases

CaM-Kinase III

CaM-kinase III (CaMKIII) is better known as elongation factor-2 kinase (eEF-2K), since elongation factor-2 is the only known substrate for this kinase. When phosphorylated by CaMKIII, elongation factor-2 activity is suppressed, inhibiting the growth of polypeptide chains during translation by dissociating elongation factor-2 from the ribosome. Thus, activation of CaMKIII through calcium/calmodulin binding is a means of coupling calcium increase in the cytosol to the suppression of protein synthesis.

CaMKIII is a monomer of approximately 100 kDa and is a cytosolic enzyme (Table 1). The catalytic domain of CaMKIII bears little resemblance to the catalytic domain of the other CaM-kinases. Following calcium/calmodulin binding, CaMKIII phosphorylates itself and, like CaMKII and CaMKIV, this autophosphorylation leads to calcium/calmodulin-independent activity. In addition, PKA can also phosphorylate CaMKIII and similarly induces calcium/calmodulin-dependent activity. Therefore, in the cellular setting, CaMKIII can respond to both calcium levels (through calmodulin) and cAMP levels (through PKA) in the downregulation of protein synthesis through the phosphorylation of elongation factor-2. CaMKIII is also phosphorylated and regulated through the rapamycin-sensitive mTOR (mammalian target of rapamycin) pathway by an as yet unidentified protein kinase.

Myosin Light-Chain Kinase

Isoforms of MLCK exist in smooth and skeletal muscle and in other tissues, including the brain. They all exist as monomers that can be quite variable in mass. The smooth muscle forms range in size from approximately 130–210 kDa and the skeletal muscle forms range from 67 to 150 kDa. This diversity in size is the result of mRNA splicing and/or alternative promoter usage in the gene. The skeletal muscle isoform is simpler in structure compared to the isoform from smooth muscle (Figure 1). MLCK contains the core catalytic domain, autoinhibitory domain, and calmodulin-binding domain; smooth muscle MLCK also contains an actin-binding domain, a fibronectin-binding

domain, and several other binding domains that regulate its localization. MLCKs of both types are maintained in an inactive state through the binding of an autoinhibitory domain to the catalytic domain. Specifically, the autoinhibitory domain binds to the protein substrate binding site but does not prevent binding of ATP. The substrate for MLCK is the regulatory light chain of myosin, and in muscle this is specifically associated with myosin II. In smooth muscle, the phosphorylation of the light chains associated with myosin dictates contraction by stimulating myosin ATPase activity. In skeletal muscle, phosphorylation of the regulatory light chain potentiates but is not fundamentally responsible for contraction.

Like other members of the CaM-kinase family, the MLCKs can serve as substrates for phosphorylation. The best studied effect is phosphorylation of a site within the calmodulin-binding domain that significantly decreases the affinity for calmodulin-binding. The kinase best characterized for mediating this phosphorylation in cells is CaMKII, but other kinases can phosphorylate this site *in vitro*. So, as calcium levels rise and calcium binds to calmodulin, calcium/calmodulin initially binds to MLCK, leading to phosphorylation of regulatory light chains and contraction. As the calcium levels increase further or are maintained, calcium/calmodulin activates CaMKII, leading to phosphorylation and desensitization of MLCK, decreased phosphorylation of regulatory light chains, and decreased contraction.

MLCK plays numerous roles in regulating neuronal function based on the multifaceted role of myosin in cell biology. Myosin-based transport in axons and nerve terminals is regulated through MLCK activity, and calcium-mediated recruitment of vesicles in presynaptic terminals is also regulated by MLCK activity (Figure 4). MLCK and the phosphorylation of myosin II are also found in postsynaptic spines, where they can modify receptor function via regulation of the actin–myosin cytoskeleton.

Summary

Calcium plays a fundamentally important role in regulating neuronal functions ranging from neurotransmitter release to synaptic plasticity to gene transcription. The calcium-binding protein calmodulin plays the role of transducing increased cytoplasmic calcium to the activation of downstream enzymes such as the family of CaM-kinases. This family of enzymes phosphorylates Ser/Thr residues in substrate proteins, altering their function. Many of the CaM-kinases are themselves substrates for phosphorylation, and, in some cases, this phosphorylation is essential for their activity or for their biological function. Thus this family of enzymes forms an enzymatic network that leads to temporal integration of the calcium signal. Certain members of the CaM-kinase family can phosphorylate a variety of substrates, providing the biochemical network a means of integrating well-coordinated responses. Importantly, several of these phosphorylation events lead to activity that outlives the calcium signal that initiated phosphorylation, providing a theoretical mechanism for short- and long-term plasticity.

See also: Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Long-Term Potentiation (LTP): NMDA Receptor Role; Postsynaptic Density/Architecture at Excitatory Synapses.

Further Reading

De Koninck P and Schulman H (1998) Sensitivity of CaM kinase II to the frequency of Ca^{2+} oscillations. *Science* 279: 227–230.

Elgersma Y, Fedorov NB, Ikonen S, et al. (2002) Inhibitory autophosphorylation of CaMKII controls PSD association, plasticity, and learning. *Neuron* 36: 493–505.

Gaertner TR, Kolodziej SJ, Wang D, et al. (2004) Comparative analysis of the 3-dimensional structures and enzymatic properties of the α, β, γ, and δ isoforms of Ca^{2+}/calmodulin-dependent protein kinase II. *Journal of Biology and Chemistry* 279: 12484–12494.

Giese KP, Fedorov NB, Filipkowski RK, et al. (1998) Autophosphorylation at Thr286 of the alpha calcium-calmodulin kinase II in LTP and learning. *Science* 279: 867–869.

Goldberg J, Nairn AC, and Kuriyan J (1996) Structural basis for the autoinhibition of calcium/calmodulin-dependent protein kinase I. *Cell* 84: 875–887.

Hudmon A and Schulman H (2004) Calcium/calmodulin-dependent protein kinase II. In: Lennarz WJ and Lane MD (eds.) *Encyclopedia of Biological Chemistry,* vol. 1, pp. 274–280. Oxford: Elsevier.

Kolodziej SJ, Hudmon A, Waxham MN, et al. (2000) Three-dimensional reconstructions of CaM-kinase IIα and truncated CaM kinase IIα reveal an unique organization for its structural core and functional domains. *Journal of Biology and Chemistry* 275: 14354–14359.

Robison J and Colbran RJ (2004) Calcium/calmodulin-dependent protein kinases. In: Lennarz WJ and Lane MD (eds.) *Encyclopedia of Biological Chemistry,* vol. 1, pp. 281–286. Oxford: Elsevier.

Rosenberg OS, Deindl S, Sung R-J, et al. (2005) Structure of the autoinhibited kinase domain of CaMKII and SAXS analysis of the holoenzyme. *Cell* 123: 849–860.

Soderling TR (2000) CaM-kinases: Modulators of synaptic plasticity. *Current Opinion in Neurobiology* 10: 375–380.

Soderling TR and Stull JT (2001) Structure and regulation of calcium/calmodulin-dependent protein kinases. *Chemical Reviews* 101: 2341–2351.

Relevant Websites

http://www.cgl.ucsf.edu/chimera/ – Computer Graphics Lab, University of California, San Francisco.
http://www.ncbi.nlm.nih.gov/entrez/'Protein' – Entrez Protein.
http://www.ncbi.nlm.nih.gov/entrez/'Structure' – Entrez Structure.
http://www.ncbi.nlm.nih.gov/entrez/ – National Center for Biotechnology Information.
http://www.cgl.ucsf.edu/chimera/ – UCSF CHIMERA an Extensible Molecular Modeling System.

Canal–Otolith Interactions

D Merfeld, Massachusetts Eye & Ear Infirmary, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The vestibular system, described in detail elsewhere, consists of two different types of end organs: the semicircular canals and the otolith organs. The three semicircular canals in each inner ear are nearly orthogonal; the afferent responses from these canals encode three-dimensional head rotation. The saccule and the utricle are otolith organs. The signals carried by the neurons innervating these two otolith organs encode gravity and linear acceleration in all three dimensions.

The nervous system faces several challenges when it tries to interpret the incoming sensory information from these vestibular end organs. The semicircular canals have high-pass filter dynamics, which make the canals insensitive to low-frequency rotations. The designers of amusement park rides (e.g., tilt-a-whirl) often use this characteristic, providing low-frequency rotations so that the semicircular canals provide information that is not veridical. The afferent signals from the otolith organs pose a different problem. Like all linear accelerometers, the otolith organ response is governed by Einstein's equivalence principle, which is based on the experimental finding that inertial mass and gravitational mass cannot be distinguished. Therefore, the otolith organs cannot distinguish inertial force due to linear acceleration from gravitational force. Hence, all linear accelerometers, including the otolith organs, respond to specific gravito-inertial force ($\boldsymbol{f}$), which is the vector sum of gravity ($\boldsymbol{g}$) minus linear acceleration ($\boldsymbol{a}$) and can be written as $\boldsymbol{f} = \boldsymbol{g} - \boldsymbol{a}$. The designers of other amusement park rides (e.g., gravitron) often use gravito-inertial equivalence to evoke illusory perceptions of motion or tilt.

Given these limitations, neither the semicircular canals nor the otolith organs are capable of providing unambiguous information encoding head motion and orientation to the brain. However, computational models with internal models as the key attribute have demonstrated that the information available from these two sensory systems can be combined to yield more accurate estimates of motion and orientation than is available from either sensory modality alone; experimental findings have confirmed many of these theoretical predictions. Specifically, information provided by the otolith organs influences and evokes rotational responses, even though the semicircular canals are specialized to respond to head rotation. Analogously, rotation information from the semicircular canals is used to help estimate how much of the otolith signal is due to gravity and how much is due to linear acceleration. Other sensory systems (e.g., vision) as well as efferent copy (or corollary discharge) mechanisms contribute to these perceptual and reflexive responses, but this article focuses exclusively on the sensory interactions between the semicircular canals and otolith organs.

In this article, we first provide necessary background information, including a number of definitions, before presenting experimental findings. When describing the experimental findings, we primarily focus on perceptual responses because perception is easier for many people to understand than theoretical modeling predictions and eye-movement recordings. We do this even though many experimental findings were forecast via theoretical predictions and even though some seminal results were first obtained via vestibulo-ocular reflex (VOR) measurements.

Background Definitions

A brief description of the predominant characteristics of the semicircular canals and otolith organs is essential to comprehend how the signals from these sensory systems interact. For similar reasons, we also describe the basic physics that underlies the physical cues measured by the vestibular system and the interaction of these physical cues.

We begin by defining a right-handed, head-fixed Cartesian coordinate system [x y z]. The x axis aligns with the naso-occipital axis and points toward the nose. The y axis aligns with the interaural axis and points toward the left ear. And the z axis is orthogonal to both the x and y axes and points toward the top of the head.

The predominant characteristics of the semicircular canals are that they respond to rotation and that their response dynamics to an angular velocity input are accurately represented by a first-order high-pass filter with a cut-off frequency of around 0.03 Hz or, equivalently, a time constant of approximately 5 s. Because the semicircular canals have high-pass filter dynamics, a step change in angular velocity yields an abrupt change in canal activation followed by a gradual exponential-like decay back to the baseline established when the head was stationary; in primates the afferent response decay has a time constant of around 5 s. If an abrupt deceleration is later applied

to bring the rotation to a stop, the high-pass dynamics yields a response in the opposite direction that again decays back to the baseline. The fact that the canals eventually signal no angular velocity during constant-velocity rotation is often used experimentally.

The predominant characteristic of the otolith organs is that they equivalently transduce gravity and linear acceleration. This concept can be clearly and concisely defined and explained using five standard facts or definitions:

1. Inertial mass is the same as gravitational mass ($m_i = m_g$), so a single mass ($m = m_i = m_g$) is defined. This experimental finding is reflected in Einstein's Equivalence principle, which states that gravitational forces cannot be distinguished from inertial forces due to linear acceleration.
2. Linear acceleration, $\boldsymbol{a}$, is the second derivative of position ($\boldsymbol{p}$) with respect to time ($\boldsymbol{a} = \mathrm{d}^2\boldsymbol{p}/\mathrm{d}t^2$). For example, when an object is dropped, gravity accelerates the object downward; this downward acceleration is gravitational acceleration. Because the otolith organs are fixed with respect to the head, they measure linear acceleration in head-fixed coordinates. Therefore, $\boldsymbol{a}$ is defined as a vector that represents linear acceleration of the head in head-fixed coordinates. This is the linear acceleration that would be measured by any linear accelerometer attached to the head.
3. On Earth, gravitational force per unit mass has a magnitude of approximately $9.8\,\mathrm{m\,s^{-2}}$ and pulls toward Earth's center; $\boldsymbol{g}$ is a vector that represents the magnitude and direction of gravity in head-fixed coordinates. This is the gravitational vector that would be measured by any linear accelerometer attached to the head.
4. Inertial forces were first defined by D'Alembert when he recognized that Newton's second law ($\sum \boldsymbol{F}_{\mathrm{est}} = m\boldsymbol{a}$) can be applied in rotating or accelerating environments when inertial forces (per unit mass) are defined as equal to the negative of linear acceleration ($\boldsymbol{i}_f = -\boldsymbol{a}$). Common examples of inertial forces include centrifugal force, which is the negative of centripetal acceleration; tangential force, which is the negative of tangential acceleration; and coriolis force, which is the negative of coriolis acceleration.
5. Gravito-inertial force per unit mass ($\boldsymbol{f}$) is defined as the vector sum of gravitational force per unit mass and inertial force per unit mass, $\boldsymbol{f} = \boldsymbol{g} + \boldsymbol{i}_f$.

Combining these facts and definitions generates the result that gravito-inertial force (per unit mass) is $\boldsymbol{f} = \boldsymbol{g} - \boldsymbol{a}$. This relationship is referenced repeatedly in this article because afferent responses from neurons innervating the otolith organs encode gravito-inertial force.

To better understand the interactions between rotational and gravitational cues, consider that rotation about an axis aligned with gravity (e.g., shaking one's head to signal no) does not influence the orientation of gravity relative to the head. But, whenever head rotation includes a component that is not aligned with gravity (e.g., nodding one's head to signal yes), the relative orientation of gravity with respect to the head changes, which is captured by the equation:

$$\boldsymbol{g} = -\int (\boldsymbol{\omega} \times \boldsymbol{g})\mathrm{d}t$$

where $\boldsymbol{\omega}$ represents the angular velocity of the head measured in head fixed coordinates. We can use this equation to keep track of the orientation of gravity with respect to the head, if we know the gravitational vector, $\boldsymbol{g}$, at just one point in time (initial condition) and measure or know the angular velocity of the head as a function of time. The gravity vector ($\boldsymbol{g}$) on the right-hand side of the equation can be thought of as the orientation of gravity now, whereas the gravity vector on the left-hand side can be thought of as later orientation of gravity due to angular rotation (e.g., the change in the orientation of gravity during head nodding).

Because the nervous system has access only to measurements of physical variables, not to the physical variables *per se,* it is crucial that neural representations be clearly delineated from the physical quantities they represent. The equations $\boldsymbol{g} = -\int (\boldsymbol{\omega} \times \boldsymbol{g})\mathrm{d}t$ and $\boldsymbol{f} = \boldsymbol{g} - \boldsymbol{a}$ represent real-world physics. Internal models of these physical effects can be represented by $\hat{\boldsymbol{f}} = \hat{\boldsymbol{g}} - \hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{g}} = -\int (\hat{\boldsymbol{\omega}} \times \hat{\boldsymbol{g}})\mathrm{d}t$, where $\hat{\boldsymbol{f}}$, $\hat{\boldsymbol{g}}$, $\hat{\boldsymbol{a}}$, and $\hat{\boldsymbol{\omega}}$ are neural estimates of the physical variables ($\boldsymbol{f}$, $\boldsymbol{g}$, $\boldsymbol{a}$, and $\hat{\boldsymbol{\omega}}$) and where internal models are defined as neural systems that mimic physical relationships.

Canal Influences on Tilt Responses

Signals from the semicircular canals influence tilt responses. To demonstrate this, we describe responses during several different motion paradigms. When upright and stationary in the dark, normal humans accurately estimate and report their perceived orientation with respect to gravity. Because patients with absent otolith function do not perform this task with the same accuracy as normals, this demonstrates that the otolith organs are a contributor to our ability to estimate head orientation with respect to gravity. If the same normals are maintained upright with respect to gravity and rotated on a centrifuge, their otolith organs measure a tilted gravito-inertial force (Figure 1(a)). Initially, the perception of vertical does not deviate far from alignment with gravity (Figure 1(b)). However,

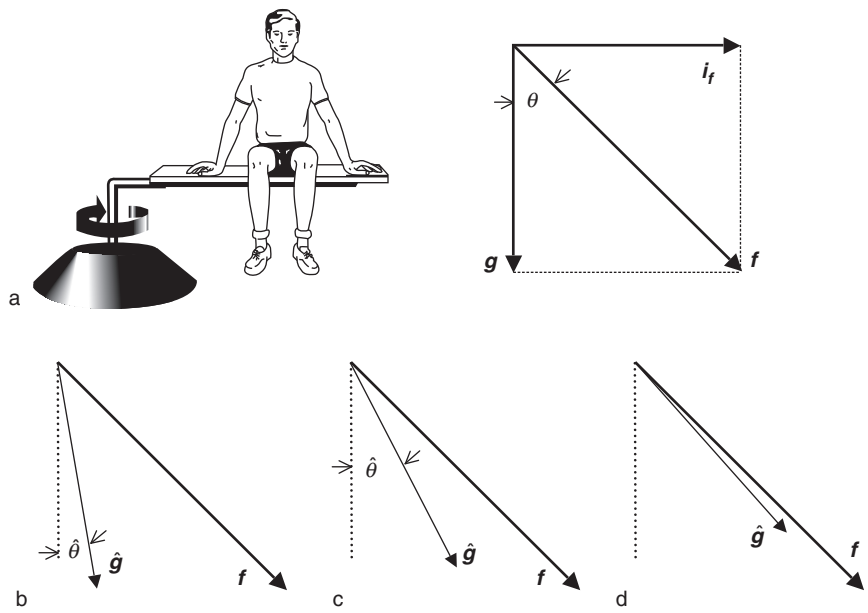

Figure 1 (a) Centrifugation physics: (b) small perceived roll tilt; (c) moderate perceived roll tilt; (d) large perceived roll tilts. When a subject is rotated at an off-center position as shown, an inertial force **(i_f)**, called centrifugal force, is present. This force sums with gravity **(g)** to yield the total gravito-inertial force **(f)** measured by the otolith organs, which is tilted with respect to gravity (θ). Perceived orientation of gravity **($\hat{g}$)** is tilted with respect to true gravity by an angle $\hat{\theta}$.

if this constant-velocity centrifugation is maintained, the perception of vertical shifts toward alignment with gravito-inertial force (Figures 1(c) and 1(d)).

The time course with which the perception of vertical aligns with gravito-inertial force is altered by the canal cues signaling rotation during centrifugation. In one experiment, subjects indicated their roll tilt perception (Figure 2, magenta) when seated at a fixed distance from the rotation axis and accelerated to a constant angular velocity. In another experiment, subjects were accelerated to a constant velocity while seated at the center of rotation. Later, when the canals no longer signaled angular velocity, the subjects indicated their roll tilt perception (Figure 2, cyan) while they were translated away from the center such that the forces measured by the otolith organs were nearly identical to those experienced during the fixed-radius trials. The primary difference between the signals from the canals and otoliths during these two motion paradigms was the presence (fixed-radius) or absence (variable-radius) of canal cues indicating yaw rotation. This difference in the canal cues led to a substantial difference in the time course of the reported roll tilt perception as was evident for both clockwise and counterclockwise rotations.

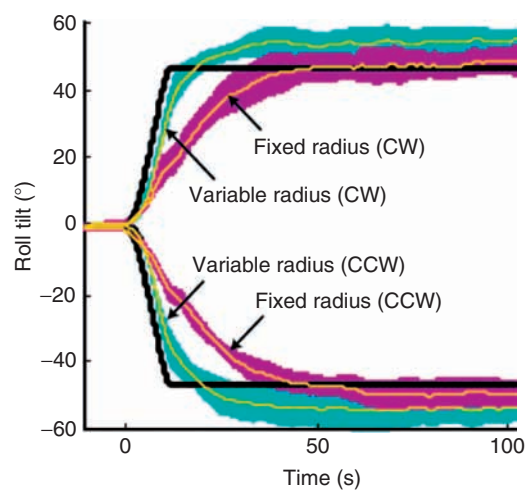

Figure 2 Perceived roll tilt during fixed-radius and variable-radius centrifugation versus time. Black lines show tilt angle (θ in Figure 1) of gravito-inertial force during centrifugation; yellow lines show average angle ($\hat{\theta}$ in Figure 1) of perceived tilt. Magenta and cyan indicate standard deviations for the fixed-radius and variable-radius conditions, respectively. CCW, counterclockwise; CW, clockwise. Reproduced from Merfeld D and Zupan L (2003) Influence of rotational cues on the neural processing of gravitoinertial force. In: Harris L and Jenkin L (eds.) *Levels of Perception,* pp. 341–373. New York: Springer, with permission from Springer-Verlag Heidelberg.

Roll tilt motion paradigms, paradigms that tilt subjects toward an ear-down position (Figure 3(a)), also demonstrate the influence of the semicircular canals on tilt responses. When humans are passively tilted in roll with their head located near the tilt axis, the otolith organs record variations in the gravitational forces relative to the head. In fact, because the head is not translating, but just tilting, there is no head linear

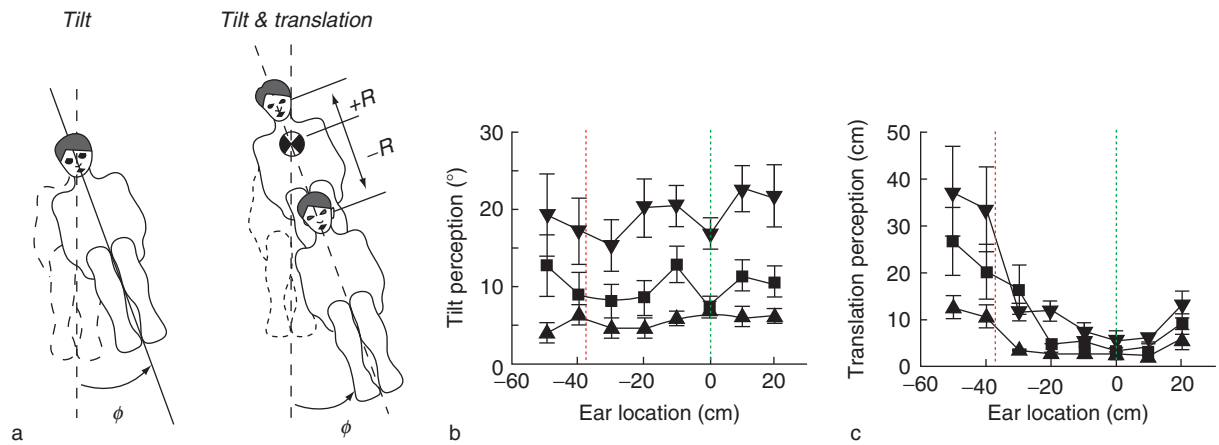

Figure 3 *Tilt* and *Tilt & Translation* tests: (a) schematic representations of the *Tilt* and the *Tilt & Translation* test paradigms; (b) verbally reported tilt angles; (c) verbally reported translation displacements. In (a), stimuli were provided by roll tilting the subjects. For the *Tilt* paradigm, the head was placed at the center of rotation ($R = 0$). For the *Tilt* & *Translation* paradigm, the head was placed above ($+R$) or below ($-R$) the tilt axis (⊗). The results in (b) and (c) are for sinusoidal roll tilts at 0.8 Hz with tilt amplitudes of 5° (▲), 10° (■), and 20° (▼). For 0.8 Hz sinusoids, the interaural component of gravity and the tangential acceleration null one another at approximately −38 cm, which is indicated by the red vertical lines. Vertical green lines indicate the head-centered (*Tilt* only) position. In (b), note that tilt perception remained accurate for all tested head positions and tilt angles, even at when no interaural otolith cues were available. In (c), note that significant translation was reported even when little or no interaural force was transduced by the otolith organs (red lines). Also note that accurate reports of little or no translation were provided with the head at the center of rotation (green lines). Error bars represent the standard error of the mean. Reproduced from Merfeld DM, Park S, Gianna-Poulin C, Black FO, and Wood S (2005) Vestibular perception and action employ qualitatively different mechanisms II: VOR and perceptual responses during combined *Tilt & Translation*. *Journal of Neurophysiology* 94: 199–205, with permission from the American Psychological Association.

acceleration. Therefore, the interaural component of the gravito-inertial force has a magnitude equal to the sine of the tilt angle $|f_y| = |g_y| = \|\boldsymbol{g} \sin \boldsymbol{\phi}\|$. When sinusoidal roll tilts having an amplitude less than 20° are performed at physiological frequencies (*c.* 0.5–1 Hz) in the dark, normals more or less accurately report how much they are tilted in roll. This is not surprising because the otolith organs accurately record head orientation with respect to gravity.

However, if the same tilt motion is repeated but now with the head located some distance below the axis of rotation, the head translates back and forth, like the bottom of a pendulum (Figure 3(a)). In this position, a sinusoidal interaural linear acceleration of the head is present at the frequency of the tilt motion; this tangential acceleration is in phase with the interaural component of gravity. For sinusoidal motion at small tilt angles ($\theta_{\max} < 20°$), there is one specific distance ($|R| \approx 9.8/(2\pi f)^2$) below the tilt axis where this tangential acceleration effectively cancels the interaural component of gravity, $g_y - a_y \approx 0$. When the head is located at this specific distance below the tilt axis, alterations in the interaural gravito-inertial force, which normally signal alterations in roll tilt, are subthreshold. Yet humans remain able to estimate their head tilt quite accurately even when tested using this motion paradigm (Figure 3(b)).

This is possible because the semicircular canals contribute to tilt responses. Specifically, as discussed earlier, head rotations that are not aligned with gravity yield changes in the relative orientation of gravity with respect to the head, $\boldsymbol{g} = -\int (\boldsymbol{\omega} \times \boldsymbol{g})\mathrm{d}t$. Studies show that the brain has implemented a distributed neural process that mimics this physical law, $\hat{\boldsymbol{g}} = -\int (\hat{\boldsymbol{\omega}} \times \hat{\boldsymbol{g}})\mathrm{d}t$. Such neural calculations are called internal models because they provide neural models that are internal to the brain of external relationships. The brain uses rotational cues from the canals via this internal model to help continually estimate the relative orientation of gravity. Neural recordings made in the vestibular nuclei of rhesus monkeys support these conclusions drawn from the perceptual data; these data demonstrate that rotational cues from the semicircular canals converge on neurons that predominantly respond to tilt stimuli.

In the presence of noise and other flaws, feed-forward estimates such as $\hat{\boldsymbol{g}} = -\int (\hat{\boldsymbol{\omega}} \times \hat{\boldsymbol{g}})\mathrm{d}t$ are imperfect. Therefore, such feed-forward neural calculations must be coupled with feedback measurements, such as those provided by the otolith organs and vision. Although motion paradigms can be defined that fool the brain into making erroneous estimates of the relative orientation of gravity, such feed-forward and feedback calculations are generally performed quite accurately. Therefore, within the normal physiological

range, our perception of vertical approximately aligns with gravity, $\hat{g} \approx g$. Furthermore, many neurons innervating the otolith organs accurately record gravito-inertial force, which can be represented as $\hat{f} \approx f$. Because it has been hypothesized that an internal model of gravito-inertial force ($\hat{f} = \hat{g} - \hat{a}$) is used to estimate linear acceleration, this suggests that for the roll tilt motion paradigms, linear acceleration is estimated as $\hat{a} = \hat{g} - \hat{f} \approx g - f = a$. These calculations suggest that, if the nervous system uses the internal models outlined here, the ability to estimate translation and/or respond to linear acceleration during the same tilt motion paradigms is also fairly accurate ($\hat{a} \approx a$). These predictions are evaluated in the next section.

Canal Influences on Translation Responses

The canals have also been shown to contribute to translation responses in both human and nonhuman primates. Illustrative reports of translation perception have been made during the passive roll tilt paradigms previously described. In these studies, humans accurately report the absence of head translation during sinusoidal whole-body-and-head roll tilts at frequencies around 0.5–1.0 Hz when a subject's head is located at or near the center of rotation such that the head does not translate. This is known to be true for peak roll tilts of up to 20°, which yield peak interaural forces of $3.4\,\mathrm{m\,s^{-2}}$ that are approximately two orders of magnitude above the human threshold for detecting linear translation ($\sim 0.05\,\mathrm{m\,s^{-2}}$) during pure translational motion. If humans simply used the presence of interaural otolith signals to indicate that they were translating, this gravitational signal would be incorrectly interpreted as being due to linear acceleration. How can humans accurately determine that they are not translating when tilted in roll with their heads at the center of rotation? This is possible because the subjects used their roll canal cues to help them keep track of the relative orientation of gravity with fairly high accuracy using internal models, as discussed in the previous section. For tilt motion stimuli with the head at the rotation axis, the force measured by the otolith organs is due to gravity alone ($f = g$). Because the subjects correctly attribute the interaural force measured by the otolith organs to the presence of gravity ($\hat{f} = \hat{g}$), this leaves no unaccounted accelerations to be transduced by the otolith organs ($\hat{a} = \hat{g} - \hat{f} = 0$). Therefore, when head tilt occurs with the head at the center of rotation, humans correctly estimate the absence of head translation because they correctly attribute the forces measured by the otolith organs to gravity.

But what happens when the head is placed some distance below the tilt rotation axis? As previously described, such motion results in head translation by an amount that depends on the distance from the center of rotation and the maximum tilt angle. In this case, subjects report the presence of head translation that is consistent with the actual head translation. The mechanisms previously outlined explain these findings. First, the canals again help the subjects to accurately estimate the relative orientation of gravity ($\hat{g} = g$). However, the gravito-inertial force measured by the otolith organs does not equal that due to gravity alone because the head is accelerating linearly during this motion paradigm ($\hat{f} = f = g - a$). Given an otolith measurement of gravito-inertial force ($\hat{f}$) and a relatively accurate estimate of gravity ($\hat{g}$), the remaining force component signaled by the otolith organs is correctly interpreted as being due to linear acceleration ($\hat{a} = \hat{g} - \hat{f}$), which approximately equals the true linear acceleration $\hat{a} \approx a$. This ability to report translational motion remained even when the head was located at the distance below the center of rotation where the interaural tangential acceleration canceled the interaural gravitational component due to the tilt via the motion paradigms already described (Figure 3(c)). In other words, because the roll canal cues contributed to the processing of the ambiguous gravito-inertial force transduced by the otolith organs, the subjects correctly reported interaural translational motion even when the interaural force cues measured by the otolith organs were subthreshold.

These human translation perception studies were inspired by earlier studies that reported similar findings via measurements of the translational vestibulo-ocular reflex (Figure 4), which is a reflexive rotation of the eye that enhances visual acuity during head motion by helping to compensate for head translation. The earlier VOR studies showed (1) that a horizontal translational VOR was evoked by sinusoidal interaural translation, (2) that no translational VOR was evoked by a sinusoidal roll tilt paradigm that provided nearly identical interaural forces, and (3) that a horizontal VOR was evoked by a motion paradigm that combined tilt and translation such that the interaural components of gravity and linear acceleration canceled one another. In other words, analogous to earlier translation perception descriptions, the monkeys responded with an appropriate horizontal VOR in response to the actual interaural translation even when the otolith organs did not register interaural force.

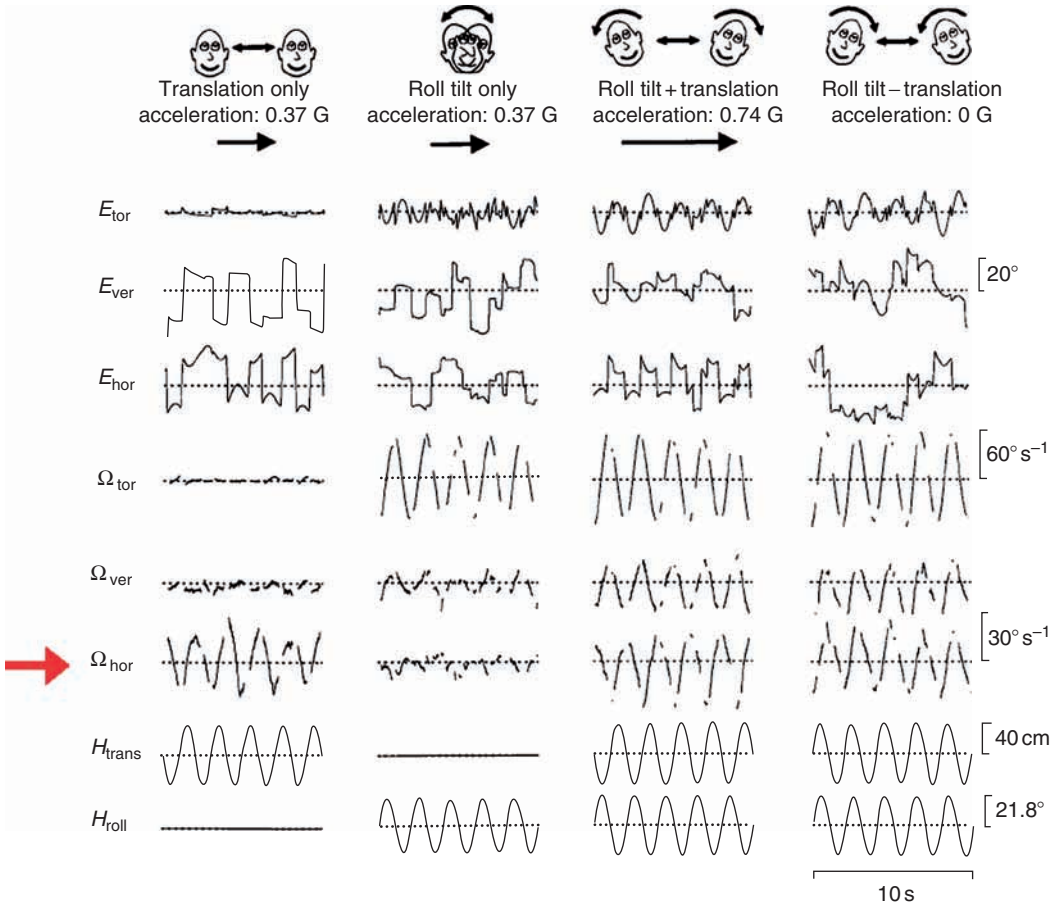

Figure 4 VOR recorded from a rhesus monkey. Torsional, vertical, and horizontal components of eye position (E_{tor}, E_{ver}, and E_{hor}, respectively) and slow phase eye velocity (Ω_{tor}, Ω_{ver}, and Ω_{hor}, respectively) during interaural translation and/or roll tilt at 0.5 Hz in complete darkness. The stimuli consisted of 'translation only,' 'roll tilt only,' 'roll tilt + translation motion,' and 'roll tilt − translation motion'. The stimulus traces (bottom) show sled position (H_{trans}) and roll tilt position (H_{roll}). Note that the horizontal VOR remained relatively constant for the three conditions that included translational motion. Red arrow indicates the row showing the horizontal slow-phase VOR response. Reproduced from Angelaki D, McHenry M, Dickman JD, Newlands S, and Hess B (1999) Computation of inertial motion: Neural strategies to resolve ambiguous otolith information. *Journal of Neuroscience* 19: 316–327, with permission from the Society for Neuroscience.

Furthermore, when the semicircular canal contributions were eliminated via a surgical procedure called canal plugging, the compensatory translational VOR evoked during the combined tilt and translation motion paradigm disappeared and a noncompensatory translational VOR was evoked during roll tilts alone. These findings emphasize the contributions of the semicircular canals to this translational VOR response because the responses changed from being compensatory to noncompensatory when the canal signals were eliminated.

Neural recordings from the flocculus and vestibular nuclei during similar motion paradigms have demonstrated similar influences of the semicircular canals on the processing of gravito-inertial force measured by the otolith organs. The canal–otolith interactions that are evident in the neural recordings, translational VOR, and human perception have all been modeled using the internal model mechanisms previously described.

Otolith Influences on Rotation Responses

Signals from the otolith organs representing gravity have long been known to influence behavioral responses evoked by rotation. Illustrative examples follow. As background, we first recall that step changes in angular velocity yield abrupt changes in canal activation followed by gradual exponential-like decays having a time constant of approximately 5 s and that rapid decelerations lead to similar responses in the opposite direction. Human rotation perception and angular VOR responses demonstrate dynamic response patterns that are qualitatively similar to

the canal afferent response, but these behavioral responses decay back to baseline more gradually than the canal signals. For Earth-vertical axis rotations, the decay time constant for these behavioral responses might be as long as 10–20 s. This prolongation of the behavioral responses via central processing is often called velocity storage.

The afferent canal signals are the same whether the rotation axis is aligned with gravity or is tilted with the respect to gravity. On the other hand, the behavioral responses evoked depend on the orientation of the rotation axis with respect to gravity. For example, if the subjects are brought to a stop following constant-velocity rotation about an Earth-horizontal axis (barbecue-spit rotation), both the postrotary perception of rotation and the postrotary reflexive eye responses (**Figure 5**) demonstrate a decay with a time constant around 5–10 s, more rapid than the 10–20 s decay time constant when the rotation axis was aligned with gravity. This influence has been attributed to the fact that otolithic and other graviceptor information contradicts the incoming rotational signal from the semicircular canals. After all, if a rotation were actually occurring about a tilted axis, the orientation of gravity would not remain static with respect to the head as is measured by the otolith organs.

Another influence of graviceptor cues on rotational responses is demonstrated most clearly by measuring the vestibulo-ocular reflexes of nonhuman primates, but, although much smaller, it can also be observed in humans. When nonhuman primates are upright in the dark and are decelerated to a stop following rotation about an Earth-vertical axis, the eye response is almost exclusively horizontal (i.e., left–right); the eye response includes little or no vertical (i.e., up-down) or torsional (i.e., rotations about the line of sight) response components (**Figure 6**). This is expected because the postrotatory canal cues signal a yaw rotation, which elicits an eye response that is primarily horizontal. However, when the subjects are tilted immediately after they have been brought to stop, the resulting eye response exhibits vertical and/or torsional response components in addition to the canal-evoked horizontal response component. These vertical and/or torsional response components contribute to shift the eye rotation axis toward alignment with gravity (**Figure 6**). To understand this influence, it is

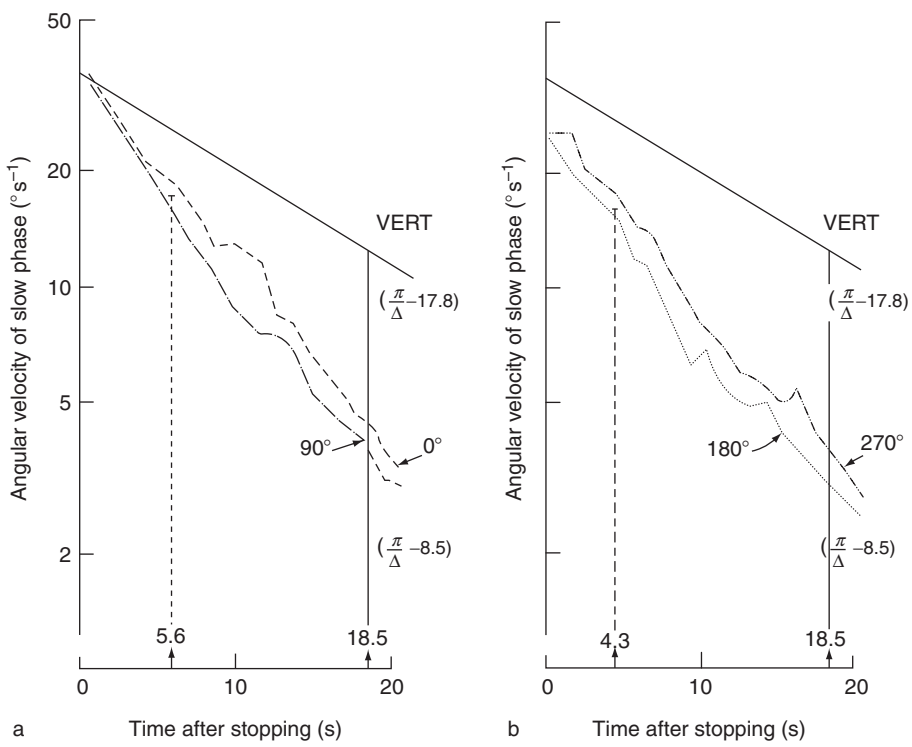

Figure 5 Comparison of angular VOR following rotation (postrotatory) when the rotation axis was aligned with gravity (VERT) or Earth horizontal: (a) 0° and 90°; (b) 180° and 270°. In (a) 0° indicates that the subject was stopped in the supine orientation and 90° indicates right side down; in (b) 180° indicates the subject was prone and 270° indicates left side down. Numbers that appear just above abscissa (e.g., 5.6) accompanied by vertical lines indicate the mean duration of the rotational sensation in seconds following deceleration to a stop. Reproduced from Benson A and Bodin M (1966) Effect of orientation to the gravitational vertical on nystagmus following rotation about a horizontal axis. *Acta Otolaryngologica* (Stockholm) 61: 517–526, with permission from the Taylor & Francis Group.

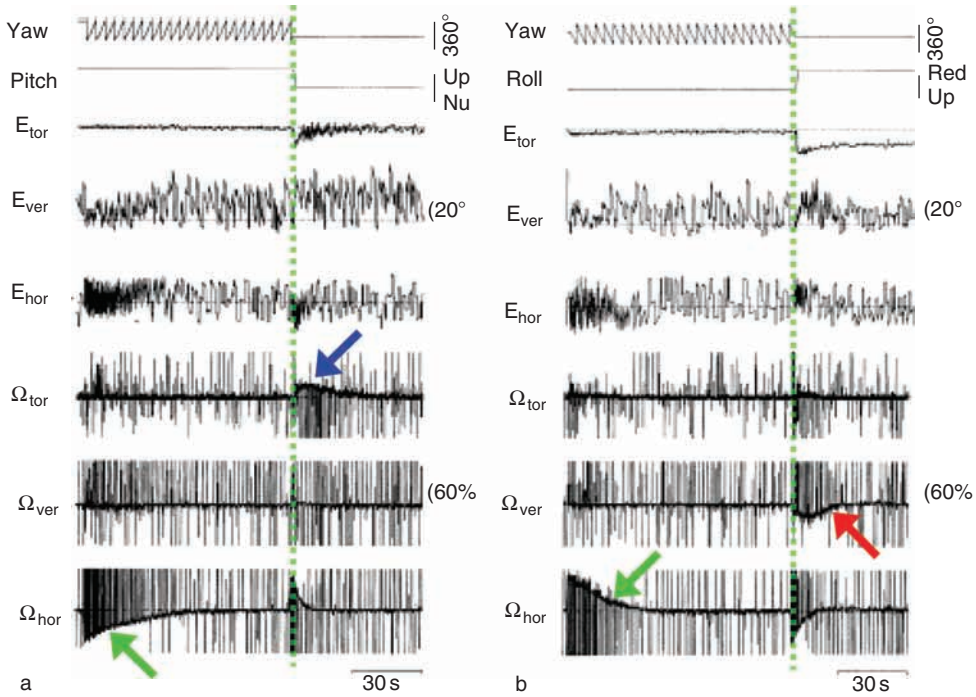

Figure 6 VOR recorded from a rhesus monkey during constant-velocity rotation about an Earth-vertical axis and, following deceleration to a stop, immediately followed by a pitch or roll tilt: (a) pitch; (b) roll tilt. Top trace shows yaw rotation position; the second trace shows pitch or roll tilt. Torsional, vertical, and horizontal components of eye position (E_{tor}, E_{ver}, and E_{hor}, respectively) and slow phase eye velocity (Ω_{tor}, Ω_{ver}, and Ω_{hor}, respectively) are shown. Green lines indicate time of deceleration to a stop and tilt with respect to gravity. Data to left of green lines show horizontal eye response (indicated with green arrows) evoked by semicircular canals. Blue arrow shows torsional VOR that shifted the axis of eye rotation toward alignment with gravity following a pitch tilt. Red arrow shows the vertical VOR that shifted the axis of eye rotation toward alignment with gravity following a roll tilt. Reproduced from Angelaki D and Hess B (1994) Inertial representation of angular motion in the vestibular system of rhesus monkeys I: Vestibuloocular reflex. *Journal of Neurophysiology* 71: 1222–1249, with permission from Springer-Verlag Heidelberg.

important to recall that the available rotational signals provided by the canals and gravitational signals provided by the otolith organs are inconsistent with one another following a postrotatory tilt. This response axis shift has been attributed to an attempt by the brain to resolve the inconsistent information it is receiving from the vestibular organs, by aligning the rotation axis with gravity.

A third influence of graviceptor cues on rotational responses is measured when subjects are rotated at a constant velocity about a rotation axis that is tilted with respect to gravity. It has long been known that the VOR response evoked by such off-vertical axis rotations (OVARs) are prolonged relative to the VOR evoked by an Earth-vertical rotation (Figure 7). Indeed, instead of decaying to zero, the presence of the rotating gravity vector, as measured by the otolith organs, evokes a constant VOR response, sometimes called the bias, long after the canal cues have decayed to zero. In addition, the VOR also exhibits a sinusoidal response component that modulates at the frequency with which the subject is rotating with respect to gravity. Several different models, including leaky-integrators and the internal model approach presented previously for rotational influences on both tilt and translation responses, mimic these physiological findings.

Summary

The afferent signals from both the semicircular canals and otolith organs include ambiguities. The semicircular canals respond to head rotation but have a limited dynamic range that often results in afferent canal signals that are not veridical, especially at low rotation frequencies. The otolith organs equivalently respond to both gravity and linear acceleration. Helping resolve these ambiguities, the otolith signals contribute to rotational responses, and the semicircular canals contribute to both tilt and translation responses. The combined information from the semicircular canals and otolith organs yields more accurate estimates of motion and orientation than is available from either sensory modality alone.

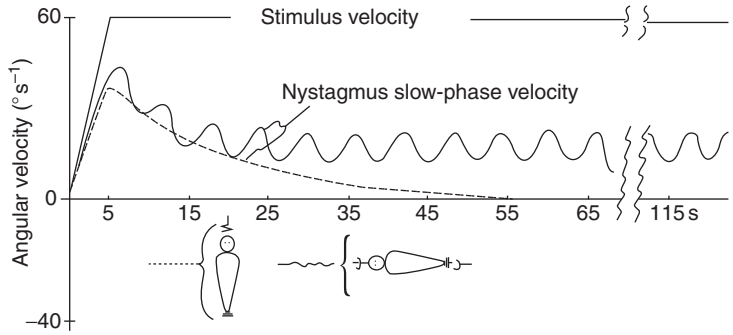

Figure 7 Schematic illustration of horizontal VOR evoked by a rotation about an Earth-vertical axis (dashed line) and an Earth-horizontal axis (solid line). Reproduced from Guedry F (1974) Psychophysics of vestibular sensation. In: Kornhuber H (ed.) *Handbook of Sensory Physiology,* pp. 1–154. New York: Springer, with permission from the American Psychological Association.

See also: Balance and Posture Control; Gravitational Effects on Brain and Behavior; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vestibular System; Vestibulo-Ocular Reflex; Visual–Vestibular Interactions.

Further Reading

Angelaki D and Hess B (1994) Inertial representation of angular motion in the vestibular system of rhesus monkeys I: Vestibuloocular reflex. *Journal of Neurophysiology* 71: 1222–1249.

Angelaki D, McHenry M, Dickman JD, Newlands S, and Hess B (1999) Computation of inertial motion: Neural strategies to resolve ambiguous otolith information. *Journal of Neuroscience* 19: 316–327.

Angelaki DE, Shaikh AG, Green AM, and Dickman JD (2004) Neurons compute internal models of the physical laws of motion. *Nature* 430: 560–564.

Benson A (1974) Modification of the response to angular accelerations by linear accelerations. In: Kornhuber H (ed.) *Handbook of Sensory Physiology, Vol. 6: Vestibular System 2: Psychophysics, Applied Aspects and General Interpretations*, pp. 281–320. Berlin: Springer.

Benson A and Bodin M (1966) Effect of orientation to the gravitational vertical on nystagmus following rotation about a horizontal axis. *Acta Otolaryngologica (Stockholm)* 61: 517–526.

Fernandez C and Goldberg J (1971) Physiology of peripheral neurons innervating semicircular canals of the squirrel monkey II: Response to sinusoidal stimulation and dynamics of peripheral vestibular system. *Journal of Neurophysiology* 34: 661–675.

Fernandez C and Goldberg J (1976) Physiology of peripheral neurons innervating the otolith organs of the squirrel monkey I: Response to static tilts and to long-duration centrifugal force. *Journal of Neurophysiology* 39: 970–984.

Goldberg J and Fernandez C (1971) Physiology of peripheral neurons innervating semicircular canals of the squirrel monkey I: Resting discharge and response to constant angular accelerations. *Journal of Neurophysiology* 34: 635–660.

Graybiel A and Brown R (1951) The delay in visual reorientation following exposure to a change in direction of resultant force on a human centrifuge. *Journal of General Psychology* 45: 143–150.

Guedry F (1974) Psychophysics of vestibular sensation. In: Kornhuber H (ed.) *Handbook of Sensory Physiology*, pp. 1–154. New York: Springer.

Merfeld DM, Park S, Gianna-Poulin C, Black FO, and Wood S (2005) Vestibular perception and action employ qualitatively different mechanisms I: Frequency response of VOR and perceptual responses during translation and tilt. *Journal of Neurophysiology* 94: 186–198.

Merfeld DM, Park S, Gianna-Poulin C, Black FO, and Wood S (2005) Vestibular perception and action employ qualitatively different mechanisms II: VOR and perceptual responses during combined tilt & translation. *Journal of Neurophysiology* 94: 199–205.

Merfeld DM, Young L, Oman C, and Shelhamer M (1993) A multidimensional model of the effect of gravity on the spatial orientation of the monkey. *Journal of Vestibular Research* 3: 141–161.

Merfeld DM, Young L, Paige G, and Tomko D (1993) Three dimensional eye movements of squirrel monkeys following post-rotatory tilt & translation. *Journal of Vestibular Research* 3: 123–139.

Merfeld DM and Zupan LH (2002) Neural processing of gravitoinertial cues in humans III: Modeling tilt and translation responses. *Journal of Neurophysiology* 87: 819–833.

Merfeld D and Zupan L (2003) Influence of rotational cues on the neural processing of gravitoinertial force. In: Harris L and Jenkin L (eds.) *Levels of Perception*, pp. 341–373. New York: Springer.

Merfeld DM, Zupan LH, and Gifford CA (2001) Neural processing of gravito-inertial cues in humans II: Influence of the semicircular canals during eccentric rotation. *Journal of Neurophysiology* 85: 1648–1660.

Raphan T and Cohen B (2002) The vestibulo-ocular reflex in three dimensions. *Experimental Brain Research* 145: 1–27.

Cardiovascular Function: Central Nervous System Control

A F Sved, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The basic goal of the cardiovascular system is to maintain appropriate blood perfusion of all tissues in the body. To attain this goal the cardiovascular system must be able to adapt rapidly to a variety of circumstances, ranging from rest to exercise to hemorrhage. Even rather routine physiological events require alterations in blood flow, such as increased gastrointestinal blood flow during the digestion of a meal or constriction of blood vessels in the legs while changing from a supine to a standing position. Regulation of blood flow to the various tissues requires regulation of mean arterial pressure and the degree of arteriolar constriction in each vascular bed. Mean arterial pressure is a function of cardiac output and the amount of resistance provided by the blood vessels. Each of these parameters is controlled, at least in part, by the nervous system.

The nervous system clearly can influence the cardiovascular system. While the autonomic nervous system plays the major role, other endocrine and behavioral responses arising from the central nervous system (CNS) also contribute. Neural control of the cardiovascular system cannot be fully appreciated by studying each of these control processes in isolation. Rather, cardiovascular regulation by the CNS must be considered as an integrated response involving neural, endocrine, and behavioral outputs (Figure 1). Thus, for example, in response to hypovolemic hypotension, such as occurs with hemorrhage, there is an increase in sympathetic neural activity to promote vasoconstriction and thereby increase blood pressure, an increase in vasopressin release to cause vasoconstriction and renal water conservation, and an increase in water and salt intake to restore the fluid deficit. This multiplicity of outputs allows compensatory backup mechanisms to operate when one or more of them are compromised, providing redundancy in the control processes of this critical physiological system.

Autonomic Control of the Cardiovascular System

In addition to local factors and hormonal influences, neural innervation of the blood vessels and heart plays a key role in cardiovascular regulation. Innervation of the heart is supplied by both the sympathetic and parasympathetic divisions of the autonomic nervous system, with parasympathetic (vagal) input slowing heart rate and sympathetic innervation increasing heart rate and contractile force. When viewed independently these actions seem to be antagonistic, but they are jointly regulated to act in concert; that is, one increases while the other decreases. In contrast, the arterioles, which are the key blood vessels involved in regulation of vascular resistance, receive only sympathetic innervation. While this sympathetic innervation of arterioles is typically vasoconstrictor via an action on α-adrenergic receptors, some blood vessels also contain β-adrenergic receptors that provide vasodilatory responses under certain conditions. Although arterioles receive only sympathetic innervation, it is important to note that these sympathetic vasoconstrictor nerves are active at rest and thus can be either excited or inhibited to produce vasoconstriction or vasodilatation, respectively. The sympathetic nervous system can differentially influence regional vascular resistances and thereby region blood flows by the selective regulation of sympathetic vasoconstrictor nerves that project to different tissues. Although the sympathetic nervous system has classically been considered to be regulated as a unit (i.e., the 'fight or flight response'), discrete and selective regulation of individual sympathetic nerves is now understood to be more the rule.

Parasympathetic innervation of the heart arises in the CNS from preganglionic parasympathetic neurons in the nucleus ambiguus and the dorsal motor nucleus of the vagus in the caudal brain stem. These cardiac parasympathetic preganglionic neurons are normally driven by a variety of excitatory synaptic inputs, though the brain circuits determining their activity at rest are still rather poorly delineated. Sympathetic innervation of the heart and vasculature arises from preganglionic sympathetic neurons in specific thoracic and lumbar spinal cord segments. These preganglionic neurons receive both spinal and supraspinal input, but in the absence of supraspinal input their activity is markedly limited. For example, high spinal cord transection reduces mean arterial blood pressure from approximately 100 mm Hg to 60 mm Hg, consistent with complete removal of resting sympathetic vasomotor activity. Such an observation, initially made back in the mid-nineteenth century, highlights the critical nature of supraspinal influences in the maintenance of arterial blood pressure and raises the question of what generates resting sympathetic vasomotor tone; this issue will be examined in more detail in the next section.

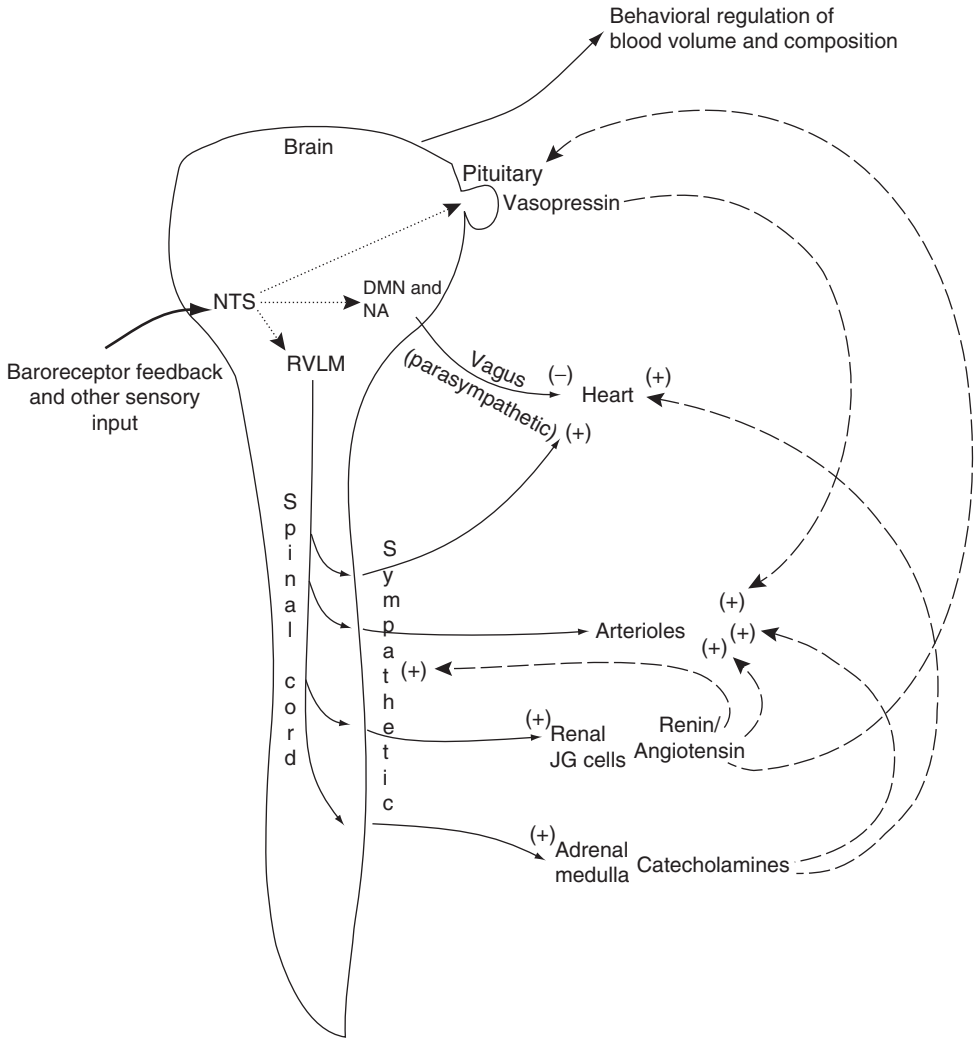

Figure 1 Overview of integrated brain control of the cardiovascular system. This schematic summarizes the CNS outputs contributing to cardiovascular regulation: parasympathetic (vagal) innervation of the heart; sympathetic innervation of the heart; sympathetic innervation of the vasculature; sympathetic innervation of the juxtaglomerular (JG) cells of the kidney that secrete renin, thereby generating angiotensin; sympathetic preganglionic innervation of the adrenal medulla controlling catecholamine secretion in to blood; vasospressin (antidiuretic hormone) secretion into blood; and behavioral responses, such as those involved in the regulation of blood volume and osmolality. Solid and dotted lines represent neural pathways; dashed lines represent hormonal influences. NTS, nucleus tractus solitarius; DMN, dorsal motor nucleus of the vagus; NA, nucleus ambiguus; RVLM, rostral ventrolateral medulla.

In addition to direct innervation of the heart and vasculature, the sympathetic nervous system also influences the cardiovascular system via a blood-borne, endocrine route. Catecholamine-secreting cells of the adrenal medulla are innervated by preganglionic sympathetic neurons. Catecholamines released into the blood from the adrenal gland have access to sympathetic neuroeffector junctions and are capable of eliciting sympathetic-like responses. Because the adrenal medulla typically secretes more epinephrine than norepinephrine, and β-adrenergic receptors are more sensitive to epinephrine than to norepinephrine, sympathetically evoked responses that are mediated by β-adrenergic receptors are more sensitive to adrenal regulation than are α-adrenergic receptor-mediated responses. Thus, since the heart and arterioles in skeletal muscle contain β-adrenergic receptors, whereas arterioles in other tissues have primarily α-adrenergic receptors, stimulation of the adrenal medulla more potently increases heart rate and skeletal muscle vasodilatation than it does vasoconstriction.

The sympathetic nervous system influences several other endocrine systems that have direct effects on the cardiovascular system. Renin release from the kidney juxtaglomerular cells into the circulation occurs in response to several factors, including increased

sympathetic outflow acting on β-adrenergic receptors. Renin released into the blood then initiates the generation of vasoactive peptides such as angiotensin II. Angiotensin II produces vasoconstriction by a direct action on angiotensin receptors located on blood vessels and also promotes vasoconstriction by acting on postganglionic sympathetic nerve terminals to enhance norepinephrine release. Circulating angiotensin also acts at brain sites lacking a blood–brain barrier (e.g., area postrema and subfornical organ) to further increase sympathetic vasomotor outflow and other responses consistent with increasing blood pressure and blood volume. Thus, angiotensin acts in a variety ways to increase arterial pressure, and the circulating levels of this peptide are controlled in large part by the brain via the sympathetic nervous system.

Sympathetic Vasomotor Tone Is Tonic and Driven from Supraspinal Sites

Following upon the observation that transections placed between the brain and spinal cord in experimental animals reduce arterial pressure to the same extent as does removing the sympathetic nervous system, experiments performed in Carl Ludwig's laboratory in the 1870s demonstrated that blood pressure was maintained with transections made in the brain at successively more caudal levels, until the transections were made in the medulla. These observations indicated that the minimal neural circuitry necessary for the maintenance of basal sympathetic vasomotor tone resides completely in the caudal brain stem. Multiple lines of evidence now support the notion that the neurons providing the key tonically active input to the sympathetic preganglionic vasomotor neurons reside in a discrete region of the rostral ventrolateral medulla (RVLM). The region near the ventrolateral surface of the medulla just caudal to the facial motor nucleus has neurons with anatomical projections to sympathetic preganglionic neurons that have firing properties consistent with them driving sympathetic vasomotor outflow. Stimulation of this region markedly increases blood pressure and sympathetic nerve activity whereas inhibiting neural activity in this region decreases arterial pressure to the same extent as does transection of the cervical spinal cord. Within the RVLM, at least two classes of spinally projecting sympathoexcitatory neurons can be distinguished based on neurochemical phenotype and electrophysiological criteria; while these two groups of neurons each contribute to cardiovascular regulation, the relative importance of these different cell types is unclear at present. Furthermore (as described in more detail later), the RVLM is not simply involved in the basal supraspinal drive of vasomotor tone. Rather, it appears to be a site where numerous and varied influences on sympathetic vasomotor tone are integrated.

The RVLM is not the only supraspinal site that provides direct innervation of sympathetic preganglionic neurons, but it clearly is the only one where acute inhibition of the neurons causes a marked decrease in arterial pressure to the same degree as does autonomic blockade. Nonetheless, other regions that innervate sympathetic preganglionic neurons, including the hypothalamic paraventricular nucleus, the pontine A5 area, and raphe pallidus, no doubt contribute to the supraspinal control of sympathetic vasomotor tone. Furthermore, all of these regions, in addition to giving rise to projections to the region of sympathetic preganglionic neurons in the spinal cord, also project to other regions involved in autonomic regulation, including the RVLM. For example, activation of the hypothalamic paraventricular nucleus increases sympathetic vasomotor tone and arterial pressure, and a portion of this response is mediated via activation of the RVLM. Thus, whereas the evidence suggests that resting sympathetic vasomotor tone is driven by the RVLM, the brain pathways involved in sympathetic vasomotor regulation are complex and interwoven.

Neural Control of the Heart Is a Balance between Sympathetic and Parasympathetic Influences

Heart rate is established by spontaneously depolarizing pacemaker cells in the sinoatrial node, and the rate of depolarization of these pacemaker cells is modulated by parasympathetic and sympathetic neural activity. Under basal resting conditions, heart rate reflects the balance of sympathetic and parasympathetic tone to the heart; in humans at rest, parasympathetic tone predominates and therefore resting heart rate is lower than is the intrinsic rate of the sinoatrial node. Due to the basal tonic autonomic control of the heart, heart rate can be increased by reducing parasympathetic activity and/or increasing sympathetic activity; the opposite responses slow the heart. These autonomic systems function in an integrated manner, and opposite changes in sympathetic and parasympathetic activity typically occur simultaneously to modulate heart rate.

The basal drive of cardiac sympathetic outflow, like sympathetic outflow to the vasculature, emanates principally from the RVLM, although the factors imparting this tonic discharge of RVLM neurons

are not fully understood. Cardiac parasympathetic preganglionic neurons in the brain stem are tonically active *in vivo*, largely reflecting input from the cardiovascular and respiratory systems. Indeed, heart rate fluctuates in phase with breathing, as a result of changes in cardiac vagal activity.

Heart rate can also be powerfully influenced by higher levels of the neuraxis. The importance of these inputs is highlighted by the ability to elicit lethal cardiac arrhythmias by stimulation of certain diencephalic and telencephalic structures that influence autonomic nerves.

Cardiovascular Homeostasis Results Primarily from a Negative Feedback Reflex

Under normal conditions, arterial pressure is maintained within very narrow limits, and this regulation is accomplished by a powerful negative feedback reflex, the baroreceptor reflex. As blood pressure begins to deviate from its regulated set point, the baroreceptor reflex rapidly compensates for this change, thereby maintaining arterial pressure within a narrow normal range. The afferent component of the baroreceptor reflex is derived from pressure-sensitive receptors associated with the aortic arch and the carotid sinus. These baroreceptor afferent nerves all project into the CNS via the IXth and Xth cranial nerves and terminate in a distinct region of the nucleus tractus solitarius (NTS). These afferents impact the activity of cardiovascular autonomic outflow, and much is known about the central neural pathways that connect the baroreceptor afferent signal projecting to the NTS to the preganglionic sympathetic vasomotor neurons in the intermediolateral (IML) and vagal cardiomotor neurons. The pathway mediating baroreceptor-evoked changes in sympathetic preganglionic neurons involves a projection from the NTS to neurons in the caudal ventrolateral medulla (CVLM). These cells in the CVLM provide inhibitory innervation of RVLM neurons that innervate sympathetic preganglionic neurons. Thus, the RVLM normally is involved both in the tonic maintenance of arterial pressure and its reflexive regulation. NTS baroreceptive neurons excite vagal preganglionic parasympathetic cardiomotor neurons in the nucleus ambiguus and dorsal motor nucleus, and this input is likely monosynaptic.

A considerable body of evidence suggests that glutamate and γ-aminobutyric acid (GABA) play dominant neurotransmitter roles in the baroreceptor reflex. For example, excitatory amino acid antagonists administered directly into the NTS, CVLM, or spinal cord completely eliminate baroreceptor reflex-evoked sympathetic vasomotor responses. Similarly, GABA antagonists injected into the RVLM block such responses. Still, many other inputs and neurotransmitters at these sites have been shown to modulate baroreceptor reflex function.

In addition to the direct autonomic control of the heart and vasculature, afferent inputs from arterial baroreceptors influence other brain outputs relevant to cardiovascular regulation.

Other Reflexes Influence Cardiovascular Homeostasis

Feedback from arterial pressure is not the only reflex involved in the short-term control of arterial pressure. Whereas it certainly is important to maintain stability of arterial pressure, a variety of conditions require arterial pressure or regional blood flow to change in an appropriate manner. For example, decreases in blood oxygen levels require circulatory adjustments to preserve oxygen delivery to the brain. Decreased blood oxygen content activates arterial chemoreceptors, located primarily in the carotid sinus and in the brain stem, which triggers reflex sympathetically mediated increases in vasoconstriction in most tissues, redirecting blood flow to the brain. The brain regions involved in this chemoreceptor reflex include the NTS as the site of termination of chemoreceptor afferent nerves and the RVLM as the site of the neurons that innervate sympathetic preganglionic neurons. Chemoreceptor and baroreceptor inputs typically do not influence the same NTS neurons, but convergence of these inputs at the level of the RVLM is extensive.

Skeletal muscle use during exercise also places demands on the cardiovascular system that must be integrated with overall cardiovascular regulation. Exercise is accompanied by increased cardiac output and a redistribution of blood flow to the contracting muscles and away from the kidneys, gut, skin, and resting muscle. This patterned cardiovascular response has been attributed to sensory feedback from the contracting muscles, central neural command that is initiated along with the cerebral signals to contract the muscles, or, most likely, a complex combination of these two mechanisms. Abundant data support the ability of muscle afferent nerves innervating the spinal cord to elicit reflexive cardiovascular responses similar to the responses produced by exercise. For example, in an anesthetized cat preparation in which the ventral roots can be stimulated to elicit isometric muscle contraction, blood pressure and heart rate will increase in proportion to muscle contraction, a response that is eliminated by section of the dorsal roots. However, other experiments point to central neural drive initiating the cardiovascular

response to voluntary muscle contraction. For example, in anesthetized cats, stimulation of the subthalamic locomotor region elicits a pattern of skeletal muscle contraction similar to walking and also elicits cardiovascular responses similar to those accompanying muscle use, even when the muscles are pharmacologically paralyzed. In humans, when the perceived exertion is increased by partially paralyzing the muscle, blood pressure and heart rate increase in proportion to effort rather than to muscle contraction. Such observations have led to the conclusion that under normal conditions either central command or muscle feedback is capable of eliciting the appropriate cardiovascular response, and that central drive predominates when there is a mismatch of signals.

Not all stimuli that evoke cardiovascular compensations are interoceptive. The behavioral alerting or defense reaction to a perceived threat or stressful stimulus serves as a good example of how cardiovascular compensations mediated by the CNS are integrated with other responses. When an organism is confronted with a threatening stimulus there is a generalized activation of the sympathetic nervous system. This sympathetic activation occurs along with an activation of the hypothalamic–pituitary–adrenal cortical axis and behavioral responses, to produce a highly integrated and coordinated defense reaction. The entire collection of responses, including autonomic, behavioral, and endocrine aspects, can be evoked by stimulation of discrete sites in the brain, such as the perifornical hypothalamus. The cardiovascular component of the defense reaction is highly patterned, including vasodilatation in skeletal muscle, vasoconstriction of most other vascular beds, and tachycardia. Two aspects of the central mechanisms mediating this cardiovascular response deserve mention. First, the response includes an inhibition of baroreceptor reflex function, mediated at least in part by inhibition of baroreceptor-sensitive neurons in the NTS. This ensures that the evoked increase in arterial pressure will not be counteracted by the baroreceptor reflex. Second, the patterned cardiovascular response is mediated through the RVLM. Thus, like other reflexively driven cardiovascular adjustments, the control of the sympathetic nervous system is exerted via spinally projecting neurons arising from the RVLM.

A Set Point for Arterial Pressure

Each of the responses noted in the preceding discussions, and short-term cardiovascular adjustments in general, represent changes in blood pressure or distribution of blood flow from some normal resting level or set point. But what sets the resting level of cardiovascular autonomic outflow upon which these reflexive or evoked changes occur? While the answer to this question is unclear at present, it appears to be determined by some critical input or inputs or else it is the result of some pacemaker or network property that can generate this baseline level of autonomic activity in the absence of any input. As an example of the former possibility, it might be that brain stem blood flow is sensed in some manner, and the 'set point' is determined, in the absence of other inputs, by maintaining brain stem blood flow at some level. In contrast, the brain might be capable of intrinsic generation of baseline cardiovascular autonomic outflow. While reports on pacemaker properties of sympathoexcitatory RVLM neurons are consistent with such a notion, current data suggest that these RVLM neurons *in vivo* do not generate such pacemaker potentials, but rather that all action potentials are preceded by excitatory synaptic potentials. Still, this does not rule out the possibility that the pacemaker neurons could be located elsewhere or that the baseline activity is an emergent property of a neural network.

The Nervous System and the Long-Term Control of the Cardiovascular System

In contrast to the established role of the brain in controlling and coordinating cardiovascular responses to acute stimuli, its role in the long-term control of blood pressure is less certain. Indeed, the autonomic nervous system is not essential for the maintenance of life under controlled conditions, and resting arterial pressure is relatively normal in the chronic absence of autonomic nervous system function. Furthermore, systems-level analyses of cardiovascular function highlight the blood volume-regulating capacity of the kidneys as the critical system involved in long-term cardiovascular control. In the models pioneered by Guyton and colleagues, in response to increased renal perfusion pressure resulting from increased arterial pressure, an increase in water and sodium excretion decreases blood volume and therefore blood pressure. Conversely, when blood pressure decreases, the reduction in glomerular filtration helps to stabilize blood volume. Since this pressure–volume regulatory capacity of the kidney is an intrinsic property of the kidney and has an infinite gain, it can precisely (although not rapidly) compensate for any change in arterial pressure. Nonetheless, the sympathetic neural innervation of the kidney modifies the pressure–volume regulatory capacity of the kidney by promoting sodium and water reabsorption. Thus, the long-term control of arterial pressure may normally depend upon sympathetic control of renal function, consistent with observations that renal denervation can produce a long-term decrease in arterial pressure.

There are additional data implicating the brain in the long-term control of arterial pressure, as manipulations in the CNS can result in chronic alterations in blood pressure.

Hypertension and the Central Neural Control of Cardiovascular Function

Chronically elevated arterial pressure is one of the most important risk factors for cardiovascular disease. Epidemiological data indicate that hypertension, typically defined in the clinic as systolic pressure >140 mm Hg or diastolic pressure >90 mm Hg, increases the risk of myocardial infarction (heart attack), stroke, congestive heart failure, and renal disease. It is estimated that more than 50 million adults in the United States, or approximately a quarter of the adult population, has hypertension.

Despite the widespread incidence of hypertension, its causes remain to be established in most instances. Indeed, no specific cause can be identified in more than 90% of cases of clinical hypertension; this is referred to as primary, or 'essential,' hypertension. Several lines of evidence suggest that the CNS may be important in the pathogenesis of hypertension. Possibly the most convincing are data that certain drugs used to treat hypertension target brain control of arterial pressure. Indeed, the understanding that some of the early drugs used to treat hypertension (e.g., clonidine and α-methyldopa) acted in the CNS to lower arterial pressure served as a major impetus for studies on the central neural control of cardiovascular function. This line of thought has also led to the hypothesis that hypertension, at least in some instances, may result from a central neural dysfunction. In addition to the fact that hypertension can be treated by drugs that act within the CNS, there are several reasons why such a hypothesis is attractive. First, the brain exerts powerful control over the cardiovascular system, and experimental manipulations within the brain can produce profound increases in arterial pressure. For example, destruction of the NTS or CVLM produces acute and severe hypertension in experimental animals. These responses can also be elicited by more selective pharmacological alteration of neurotransmitter mechanisms in these regions. Second, certain forms of experimental hypertension have been associated with changes in the synthesis, release, and receptor binding of specific neurotransmitters in specific CNS regions. For example, catecholamine synthesis is altered in several brain sites in models of hypertension in rats. Recent studies using viral vector transfection techniques to manipulate specific genes in specific brain regions provide compelling evidence that the brain is involved in the pathogenesis of hypertension in some experimental models.

See also: Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Central Cardiovascular Control; Autonomic Nervous System; Blood Pressure: Baroreceptors; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Dampney RA, Horiuchi J, Tagawa T, et al. (2003) Medullary and supramedullary mechanisms regulating sympathetic vasomotor tone. *Acta Physiologica Scandinavica* 177: 209–218.

Guyenet PG (2006) The sympathetic control of blood pressure. *Nature Reviews Neuroscience* 7: 335–346.

Janig W (2006) *The Integrative Action of the Autonomic Nervous System*. Cambridge, UK: Cambridge University Press.

Loewy AD and Spyer KM (eds.) (1990) *Central Regulation of Autonomic Functions*. New York: Oxford University Press.

Madden CJ and Sved AF (2003) Rostral ventrolateral medulla C1 neurons and cardiovascular regulation. *Cellular and Molecular Neurobiology* 23: 739–749.

Osborn JW, Jacob F, and Guzman P (2005) A neural set point for the long-term control of arterial pressure: Beyond the arterial baroreceptor reflex. *American Journal of Physiology – Regulatory Integrative and Comparative Physiology* 288: R846–R855.

Sved AF, Ito S, and Sved JC (2003) Brainstem mechanisms of hypertension: Role of the rostral ventrolateral medulla. *Current Hypertension Reports* 5: 262–268.

Cataplexy

S Nishino, Stanford University School of Medicine, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Cataplexy (one of the symptoms of narcolepsy) is a sudden drop of antigravity muscle tone triggered most often by positive emotional factors (such as laughter) and less frequently by negative emotional factors (such as anger or frustration).

Narcolepsy is a chronic sleep disorder that affects 0.02–0.05% of the general population and has also been reported in dogs (Figure 1(a)) and, more recently, in (genetically engineered) mice and rats. Human narcolepsy is caused by an interplay of genetic susceptibility and environmental factors, and the disease is tightly associated with human leukocyte antigen (HLA) DR2 and DQ6 (DQB1*0602). Although many autoimmune diseases are associated with certain HLA types, an involvement of autoimmune mechanisms with narcolepsy has not been demonstrated.

The symptoms of narcolepsy are often classified as the narcolepsy tetrad (excessive daytime sleepiness, cataplexy, hypnagogic hallucination, and sleep paralysis) or pentad (narcolepsy tetrad plus insomnia). Among these symptoms, cataplexy is pathognomonic for narcolepsy, and it occurs almost exclusively in idiopathic narcolepsy and in rare cases of symptomatic (or secondary) narcolepsy (i.e., narcolepsy associated with other neurological conditions such as a brain tumor or stroke). Sleep paralysis and hypnagogic hallucinations also occur in other sleep disorders such as sleep apnea (or even in healthy subjects, especially when their sleep–wake pattern is disturbed). Sixty percent of narcoleptic subjects diagnosed by the current International Classification of Sleep Disorders (ICSD-2) diagnostic criteria developed cataplexy, usually around the onset of sleepiness or within 3–5 years (narcolepsy with cataplexy). The other 40% of narcoleptic subjects do not exhibit cataplexy, but have rapid eye movement (REM) sleep abnormalities (i.e., sleep-onset REM periods) together with excessive daytime sleepiness (classified as narcolepsy without cataplexy). Thus, narcolepsy without cataplexy is also clinically common.

The French physician Gélineau (in 1880) first coined the term narcolepsy and described two cases of excessive daytime sleepiness, discussing how emotions influenced the onset of sleep attacks (a fall astasias was also accompanied with sleep attacks in one of these two cases). However, Lowenfeld (in 1902) is usually considered the first author to characterize cataplexy as part of the narcoleptic syndrome.

Clinical Features of Cataplexy

Severe cataplexy episodes produce a bilateral, generalized weakness of the antigravity muscles sufficient to cause a fall (although usually without injury; a video of cataplexy is available at the Stanford website). Other episodes may be partial, affecting only the face, vocal cords, or limb(s). Consciousness is not impaired unless the patient subsequently falls asleep (a transition to REM sleep often occurs). Calling the person's name or shaking his or her body can interrupt cataplexy, and patients can usually recover without any confusion. Patients can also recall events happening before and during cataplexy if they recover before falling asleep. During cataplectic attacks, the cortical electroencephalogram (EEG) is desynchronized and the EEG pattern is similar to that during waking and REM sleep (Figure 1(b) shows the EEG patterns during cataplexy in narcoleptic dogs). In contrast to epileptic seizures, no paroxysmal spikes or burst EEG patterns are observed. Neurological examinations performed at the time of an attack show a suppression of the platellar reflex. The duration of cataplexy varies from a split second to several minutes. Its frequency varies according to the time and individual subject – from less than one episode per year to several episodes per day. Cataplexy is currently treated pharmacologically, most often with tricyclic antidepressants (Table 1), but the antiepileptics are ineffective. The anticataplectic effect of tricyclic antidepressants (imipramine) was coincidentally found in the clinical practices in the 1950s. Tricyclic antidepressants are also effective for reducing sleep paralysis and hypnagogic hallucinations, but have little effect on excessive daytime sleepiness. Therefore, most narcolepsy–cataplexy patients need to take additional wake-promoting compounds such as an amphetamine, methylphenidate, or modafinil. Conversely, these wake-promoting compounds used for the treatment of excessive sleepiness have no (modafinil) or only moderate (amphetamine and amphetamine-like compounds) anticataplectic effects.

Cataplexy often worsens with poor sleep and fatigue. Patients may also experience status cataplecticus; this rare manifestation of narcolepsy is characterized by subintrant cataplexy that lasts several hours per day and confines the subject to bed. It can occur spontaneously or more often on withdrawal from anticataplectic drugs.

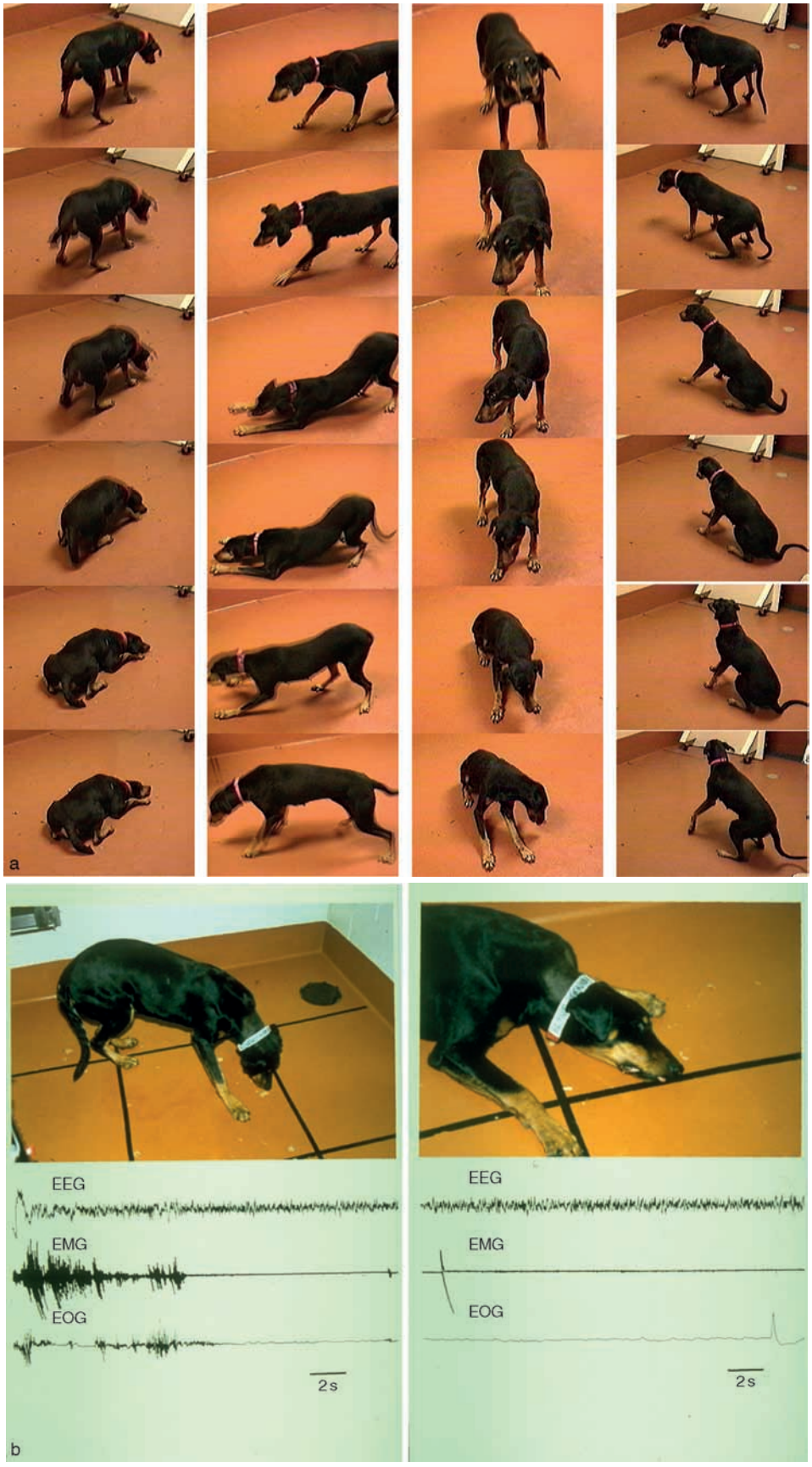

Figure 1 Continued

Table 1 Drugs commonly used for treatment of cataplexy

Drug	*Dosage*	*Side effects and comments*
Imipramine[a]	10–100 mg	Dry mouth, anorexia, sweating, constipation, drowsiness
Protriptyline[a]	5–60 mg	Dry mouth, sweating, anxiety, disturbed nighttime sleep
Clomipramine[a]	10–150 mg	Dry mouth, sweating
Fluoxetine[b]	20–80 mg	Few side effects, nausea, dry mouth
Venlafaxine[b]	75–150 mg	Few side effects, nausea
Sodium oxybate[c]	1.5–4.5 g bedtime and 2–3 h later	Morning sedation, nausea, dizziness, urinary incontinence

[a]Imipramine, protriptyline, and clomipramine are classical tricyclics.
[b]Fluoxetine is a serotonin selective reuptake inhibitor (SSRI) and venlafaxine is serotonin/norepinephrine selective reuptake inhibitor, which have less side effects.
[c]Sodium oxybate is not an antidepressant; it is a short-acting sedative. Taking sodium oxybate at night improves sleepiness and cataplexy the following day.

Physiological and Pathophysiological Aspects of Cataplexy

In humans, the first REM sleep episode occurs approximately 90 min after sleep onset and reappears every 90 min (every 30 min in dogs and cats). A total of four to five REM sleep episodes may occur during the night. Shortly after the discovery of REM sleep, it was found that in narcolepsy REM sleep often occurs at sleep-onset (sleep-onset REM periods) or even during short daytime naps (Figure 2). Ever since, cataplexy, sleep paralysis, and hypnagogic hallucinations have been often considered a disassociated manifestation of REM sleep; the occurrence of these symptoms is often explained by an abnormal rapid transition to REM sleep during active wake or sleep-onset. During REM sleep, complete and tonic inhibition of muscle tonus, together with phasic bursts of REMs and swift muscle twitching, occurs physiologically, and the amplitude recorded in the electromyogram (typically monitored in neck or chin muscles, combined with sleep EEG recordings) is the lowest. It is, however, the activity of pyramidal tract neurons in the motor cortex that mediate the limb movements, which are as high during REM sleep as during active waking. It is thought that the tonic inhibitory signal for muscle tonus originating in the dorsal pons overcomes this pyramidal motor activation system during REM sleep and that this results in complete muscle atonia during this sleep state. This notion is supported by brain lesion studies in animals. When the lesions are made in the dorsal pons, the phenomenon called REM sleep without atonia is observed, and the animals move their limbs during REM sleep. Similar phenomena (i.e., REM sleep behavior disorders) are also reported in humans associated with various neurological conditions (including stroke in the pons).

A series of experiments suggested that an executive mechanism of atonia during cataplexy shares a common pathway with atonia during REM sleep.

1. Electrophysiological experiments using familial narcoleptic Dobermans demonstrated that a population of cell groups in the brain stem that are responsible for loss of muscle tone during REM sleep are also active during cataplexy.
2. Experiments have also recently demonstrated that REM-off cells in the adrenergic locus coeruleus (LC) in the pons cease to discharge during cataplexy.
3. Experiments have shown that H-reflex activity (an electronically stimulated muscle response with monosynaptic latency due to excitation of Ia afferents in the spinal cord) profoundly diminishes or disappears during both REM sleep and cataplexy in humans.
4. The major mechanism of action of anticataplectics is thought to be the suppression of REM sleep by activating the central adrenergic system.

Therefore, the motor inhibitory components of REM sleep at the level of the lower brain stem and the spinal cord are also operative during cataplexy.

In contrast, mechanisms of the induction of cataplexy are not well understood. However, several

Figure 1 Cataplectic attacks in Doberman pinschers: (a) photos of attacks; (b) polygraphic recordings of cataplectic attacks observed in a narcoleptic Dobermans. Emotional excitations, appetizing food, or playing easily elicits multiple cataplectic attacks in these animals, and most attacks are bilateral (97.9%). Atonia was initiated partially in the hind legs (79.8%), front legs (7.8%), neck/face (6.2%), or whole-body/complete attacks (6.2%), and a progression of attacks was also seen (49% of all attacks). As shown in (b), a rapid transition from the waking state (desynchronized electroencephalogram (EEG) with high electromyogram (EMG)) to cataplexy (sudden reduction of EMG) is observed (left). During cataplexy (right), the EEG consists of low-voltage fast-wave, occasionally accompanied by rapid eye movement. EOG, electrooculogram.

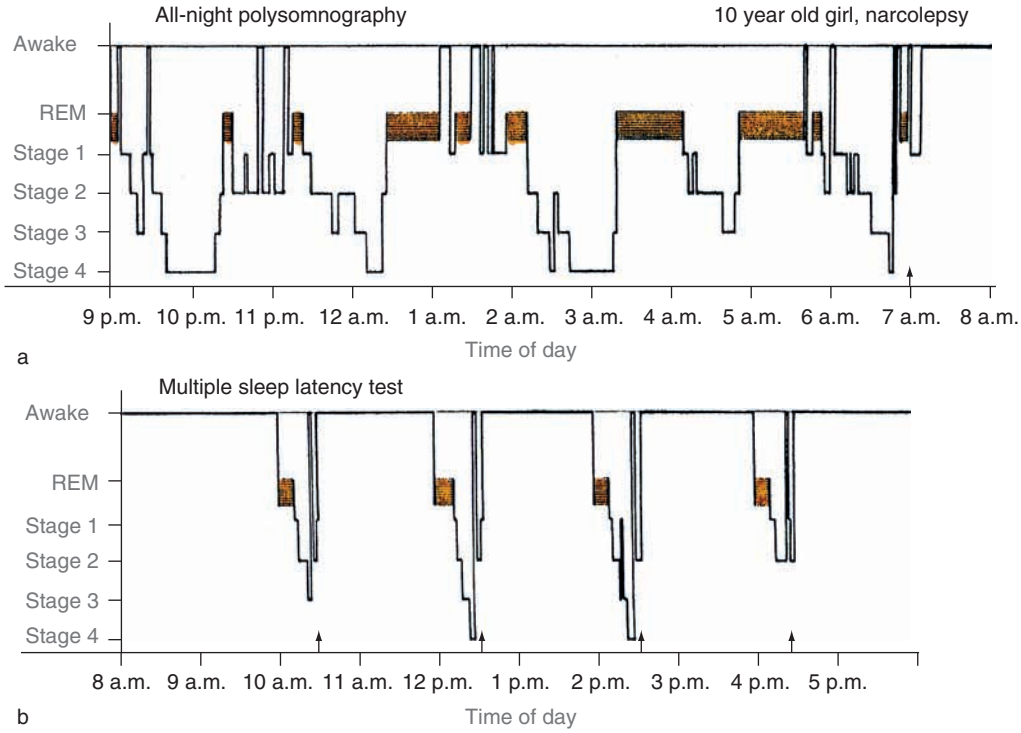

Figure 2 Test results for a 10-year-old female narcoleptic patient: (a) all-night polysomnogram; (b) multiple sleep latency test. In (a), note the appearance of a sleep-onset REM period in the night recordings and marked sleep tendency. In (b), note a sleep-onset REM period at each daytime nap test. Vertical arrows indicate a verbal calls to awaken the patient at the end of each nap test. REM, rapid eye movement. Adapted from Honda Y (1988) Clinical features of narcolepsy: Japanese experiences. In: Honda Y and Juji T (eds.) *HLA in Narcolepsy*, pp. 24–57. Berlin: Springer.

experiments using narcoleptic Dobermans suggested that mechanisms for induction of cataplexy are different from those for REM sleep.

1. Cataplexy in narcoleptic dogs can be elicited anytime by emotional excitation, whereas occurrences of REM sleep in these animals are regulated by a normal 30 min cyclicity.
2. Injection of cholinergic agonist (carbachol) into the basal forebrain (anterior hypothalamus) induces long-lasting cataplexy-like atonia in narcoleptic dogs, but this manipulation does not affect the 30 min cyclicity of the occurrence of the phasic REM sleep phenomena (i.e., REMs).
3. Dopamine D2/3 receptor agonists potently aggravate and D2/3 receptor antagonists potently inhibit cataplexy in narcoleptic dogs, but these classes of compounds have little effect on the occurrences of REM sleep.

Many normal individuals report slight muscle weakness with intense emotion, and during laughter, the H-reflex is markedly reduced in controls and also in narcoleptic subjects. Cataplexy may thus result from excessive activation of the same descending motor inhibitory pathways that are active during strong emotions, but these conceptual (emotion-triggered) inhibitory pathways and mechanisms are not yet identified.

A series of pharmacological experiments using canine narcolepsy pointed out the importance of the cholinergic system and monoaminergic system for the regulation of cataplexy. Various manipulations and pharmacological compounds reduce cataplexy, but some may be nonspecific; if animals are afraid or do not feel comfortable, cataplexy is generally reduced. In contrast, only a limited class of compounds aggravates cataplexy. These include compounds that activate cholinergic neurotransmission (especially by cholinergic M2/3 receptor agonists) and compounds that reduce monoaminergic neurotransmission (especially by adrenergic α_2 agonists (presynaptic), α_1 antagonists (postsynaptic), and dopamine D2/3 agonists (presynaptic)). Thus, there is a similarity between the pharmacological control of cataplexy and that of REM sleep, but some discrepancies have also been noted with D2/3 compounds.

Cataplexy and Hypocretin/Orexin Deficiency

Cataplexy is also a hallmark symptom of narcolepsy in animals, such as dogs and mice. The identification

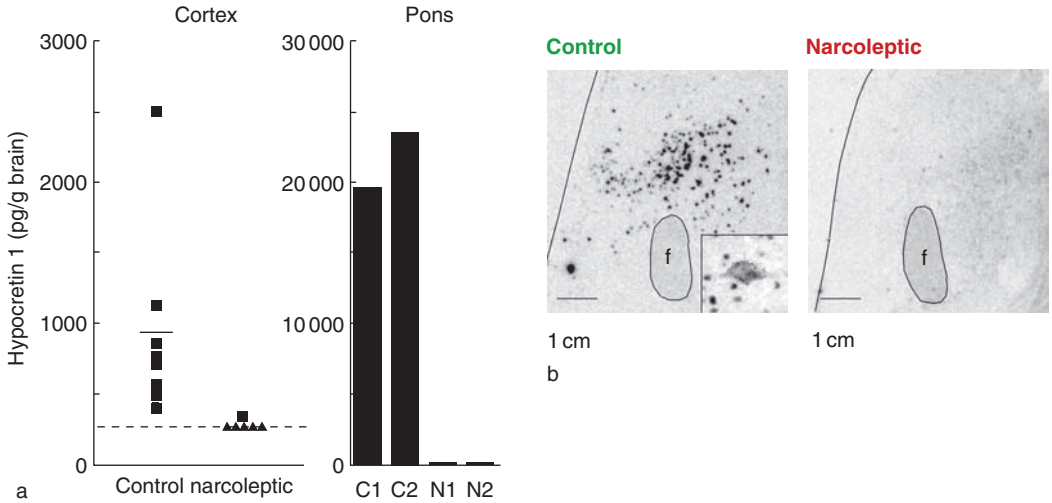

Figure 3 Absence of hypocretin signals and peptide in the hypothalamus of narcoleptic patients: (a) brain hypocretin levels in narcoleptic and control subjects; (b) hypocretin mRNA signal in the lateral hypothalamic area of control and narcoleptic subjects. In (a), cortical hypocretin-1 and hypocretin-2 (data not shown) concentrations were detectable in all controls (— indicates mean) and undetectable (— indicates the detection limit, 332 pg/g brain) in five of the six narcoleptic subjects. The difference in hypocretin concentrations (two narcoleptic subjects vs. two controls) was more pronounced in the pons where high concentrations are observed in the control subjects. In (b) preprohypocretin transcripts are detected in the hypothalamus of control (left) but not narcoleptic (right) subjects. C1/2, controls; f, fornix; N1/2, narcopleptic patients. Adapted from Peyron C, Faraco J, Rogers W, et al. (2000) A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nature Medicine* 6: 991–997.

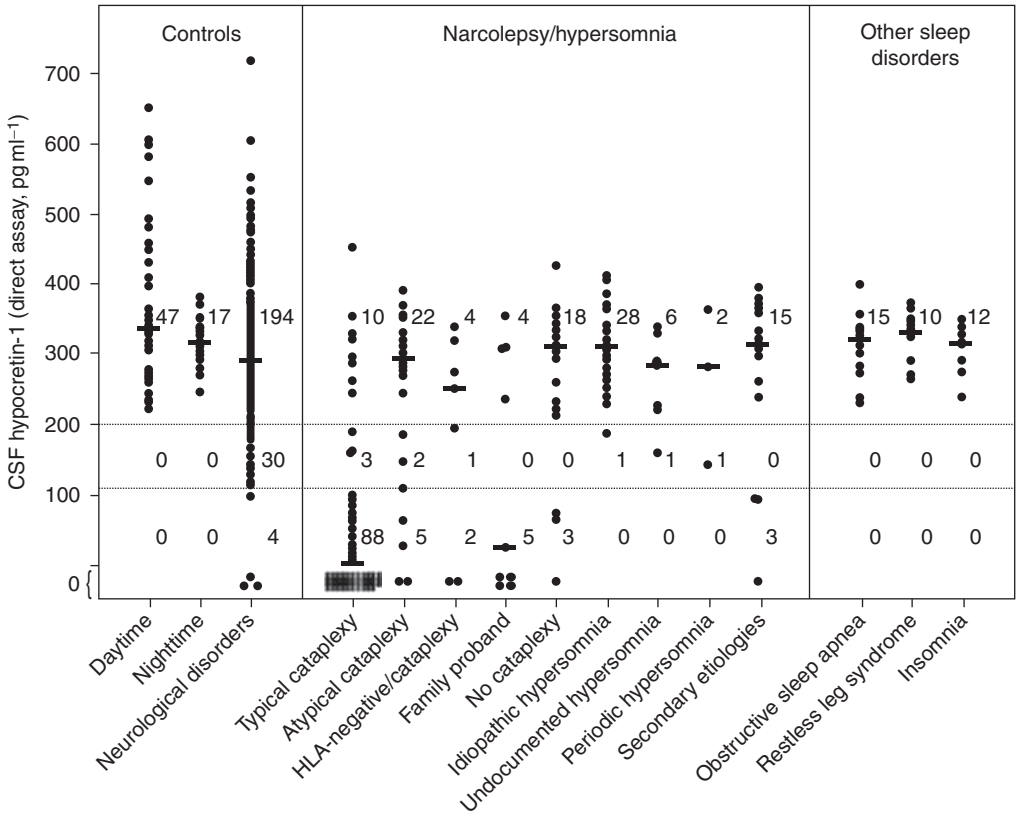

Figure 4 Cerebrospinal fluid hypocretin-1 levels across various disease categories. Each dot represents a single patient. Hypocretin-1 values below 110 pg ml^{-1} were determined to be diagnostic for narcolepsy. Concentrations above 200 pg ml^{-1} best determined healthy control values. The number of patients with hypocretin values below or equal to 110 pg ml^{-1}, above 200 pg ml^{-1}, and between these two values is indicated for each category. CSF, cerebrospinal fluid; HLA, human leukocyte antigen. Adapted from Mignot E, Lammers GJ, Ripley B, et al. (2002) The role of cerebrospinal fluid hypocretin measurement in the diagnosis of narcolepsy and other hypersomnias. *Archives of Neurology* 59: 1553–1562.

of cataplexy in these animals led to the discovery of narcolepsy in these animal species. Studies in these animals (familial canine narcolepsy in narcoleptic Dobermans and gene-knockout mice) led to the discovery of the etiology of narcolepsy in animals. Mutation of one of the two hypocretin/orexin receptors and knockout of the hypocretin/orexin gene (preprohypocretin) causes narcolepsy in dogs and mice, respectively. The hypocretin/orexin system is a recently discovered hypothalamic neuropeptidic system. Hypocretin neurons are exclusively located in the lateral hypothalmic area, but project to most brain areas. Subsequent human studies revealed that the mutation in the hypocretin-related gene is extremely rare, and only one case of early-onset narcolepsy to date was identified as having a mutation in the preprohypocretin gene. However, cerebrospinal fluid (CSF) measures (and a small number of postmortem brain studies) found hypocretin ligand deficiency in most idiopathic narcolepsy–cataplexy cases (~90%) (Figures 3 and 4).

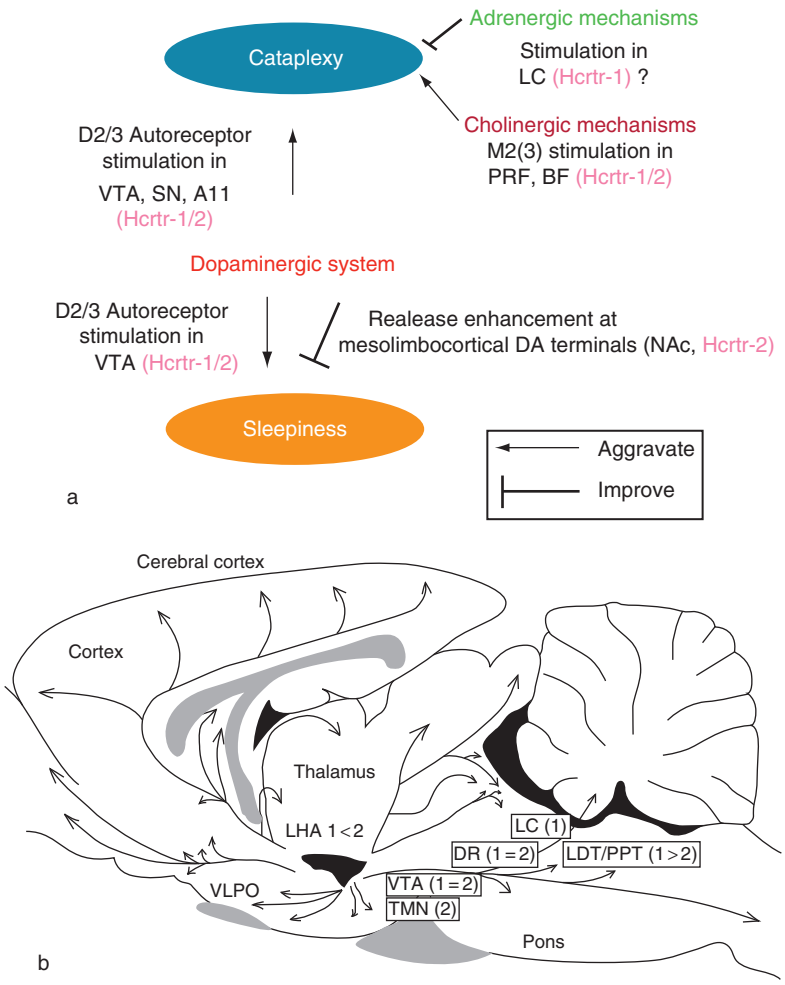

Figure 5 Monoaminergic and cholinergic control of sleepiness and cataplexy in relation to hypocretin input: (a) schematic; (b) hypocretin projections. As shown in (a), the stimulation of adrenergic transmission by adrenergic uptake inhibitors potently reduces cataplexy; this pharmacological property is probably involved in the mode of action of currently used anticataplectic agents (e.g., tricyclic antidepressants). The fact that both the presynaptic $\alpha 2$ autoreceptor stimulation and postsynaptic $\alpha 1$ blockade aggravate cataplexy is consistent with an inhibitory role of adrenergic transmission in the control of cataplexy. D2/3 autoreceptor stimulation aggravates both cataplexy and sleepiness, and amphetamines stimulate DA release at DA terminals and mediate EEG arousal. Muscarinic M2 stimulation induces behavioral wakefulness and cortical desynchrony in control dogs, but it induces cataplexy in narcoleptic dogs. As shown in (b), hypocretin-containing neurons project to these previously identified monoaminergic and cholinergic regions where hypocretin receptors (Hcrtr1/2) are enriched. Impairments of hypocretin input may result in cholinergic and monoaminergic imbalance and the generation of narcoleptic symptoms. BF, basal forebrain; D2/3, dopamine receptors; DA, dopamine; DR, dorsal raphe; EEG, electroencephalogram; LC, locus coeruleus; LDT, laterodorsal tegmentum; LHA, lateral hypothalamic area; M2, muscarinic receptor; NAc, nucleus accumbens; PPT, peduncle pontine tegmentum; PRF, pontine reticular formation; SN, substantia nigra; TMN, tuberomamillary nucleus; VLPO, ventrolateral preoptic area; VTA, ventral tegmental area.

Extended human studies further suggested that hypocretin deficiency is tightly associated with the occurrence of cataplexy and that most patients with narcolepsy without cataplexy and idiopathic hypersomnia and secondary excessive daytime sleep disorders (such as obstructive sleep apnea) have normal CSF hypocretin levels (Figure 4). A positive reading for HLA is also tightly associated with hypocretin deficiency. Hypocretin replacements are expected to be a new treatment for hypocretin-deficient narcolepsy, but this is not available due to low penetration (of the brain) of hypocretin peptides; the development of synthetic agonists are likely to be needed.

A significant degree of hypocretin deficiency already occurs around the disease onset and, in some case, before the onset of cataplexy. Moderate hypocretin deficiency is also reported in cataplexy associated with some secondary narcolepsy cases with brain tumors or vascular diseases. In these conditions, the degree of hypocretin deficiency is modest compared to most idiopathic cases (undetectably low), but projection and projection sites of the hypocretin system may also be damaged in these secondary cases. Thus, cataplexy may now appear to be a hypocretin-deficient pathological phenomenon. However, it is not known if impairments of the hypocretin neurotransmission instantly lead to occurrences of cataplexy or if any secondary changes are required to develop cataplexy. If some of these secondary changes are irreversible, hypocretin supplements (when they become available) may not be effective.

Cataplexy is also reported to occur in rare neurological conditions, such as the Nieman-pick type C (without sleepiness) disease, Prader–Willi syndrome, and myotonic dystrophy (with sleepiness). The involvements of the hypocretin system in cataplexy in these conditions have not been systematically studied.

Hypocretins are, in general, excitatory neuropeptides, and they project to most monoaminergic systems (such as the LC, dopaminergic ventral tegmental area and substantial nigra, serotonergic raphe, and histaminergic tuberomammillary nucleus) and cholinergic systems (such as the basal forebrain and brain stem cholinergic nuclei) and their terminal areas, and modulate their neurotransmissions. Hypocretin deficiency is thus likely to induce abnormalities in monoaminergic and cholinergic neurotransmissions (Figure 5). Indeed catecholamine contents (dopamine and norepinephrine) are increased in several brain regions, and this may be a compensatory increase due to the lack of excitatory inputs to these systems. This upregulation alone is not likely to prevent the occurrence of the symptoms of narcolepsy, and compounds that enhance the catecholaminergic neurotransmission are needed to control sleepiness and cataplexy. Considering the fact that compounds which shut down the catecholaminergic system (autoreceptor agonists, α_2 and D2/3 agonists) are very potent for inducing cataplexy in the canine model, a sudden decline of catecholamine tonus may be one of the most important mechanisms for triggering cataplexy; however, the precise mechanisms for this are still not understood. We can also assume that the disinhibition of executive components of REM sleep atonia and/or disfacilitation of tonic muscle activity due to the lack of the hypocretin signaling are likely to be involved in the occurrences of cataplexy. The link between this mechanism and the emotional component is still not known.

See also: Acetylcholine Neurotransmission in CNS; Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine; Autonomic Dysregulation During REM Sleep; Narcolepsy; Sleep Apnea; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Honda Y (1998) Clinical features of narcolepsy: Japanese experiences. In: Honda Y and Juji T (eds.) *HLA in Narcolepsy* pp. 24–57. Berlin: Springer.

Lammers GJ, Overeem S, Tijssen MA, and van Dijk JG (2000) Effects of startle and laughter in cataplectic subjects: A neurophysiological study between attacks. *Clinical Neurophysiology* 111: 1276–1281.

Mignot E, Lammers GJ, Ripley B, et al. (2002) The role of cerebrospinal fluid hypocretin measurement in the diagnosis of narcolepsy and other hypersomnias. *Archives of Neurology* 59: 1553–1562.

Nishino S and Mignot E (1997) Pharmacological aspects of human and canine narcolepsy. *Progress in Neurobiology* 52(1): 27–78.

Nishino S, Riehl J, Hong J, Kwan M, Reid M, and Mignot E (2000) Is narcolepsy a REM sleep disorder? Analysis of sleep abnormalities in narcoleptic Dobermans. *Neuroscience Research* 38(4): 437–446.

Nishino S and Sakurai T (eds.) (2005) *The Orexin/Hypocretin System: Its Physiology and Pathophysiology.* Totowa, NJ: Humana Press.

Peyron C, Faraco J, Rogers W, et al. (2000) A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nature Medicine* 6: 991–997.

Sakurai T (2005) Roles of orexin/hypocretin in regulation of sleep/wakefulness and energy homeostasis. *Sleep Medicine Reviews* 9(4): 231–241.

Scammell TE (2003) The neurobiology, diagnosis, and treatment of narcolepsy. *Annals of Neurology* 53(2): 154–166.

Relevant Website

http://www-med.stanford.edu – Stanford University School of Medical (animal cataleptic attacks video).

CCK/Gastrin and Receptors

M C Beinfeld, Tufts University School of Medicine, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Discovery of Cholecystokinin

The biological activity that would eventually be called cholecystokinin (CCK) was discovered by Bayliss and Starling early in the twentieth century. They showed that intestinal acidification (which releases CCK) caused gall bladder contraction. Porcine intestinal CCK 33 was identified by Mutt and Jorpes in the 1960s. The name CCK/gastrin arose in 1975 with the discovery by Vanderhaeghen of CCK-like immunoreactivity in the brain with a gastrin antiserum which cross-reacted with CCK. It soon became clear based on its isolation by Dockray that this material was not gastrin, but CCK 8. Small amounts of gastrin are found in the brain and pituitary of several species.

Distribution of CCK

CCK is one of the most abundant neuropeptides in the brain, where it is widely distributed. It is also expressed in nerves and in endocrine cells in the intestine. Gastrin is highly homologous to CCK and may have arisen from CCK through gene duplication during evolution. CCK and gastrin share the same four C-terminal amino acids. Gastrin is expressed mainly in the stomach, with lesser amounts in the intestine, colon, and pancreas. Levels of gastrin in circulation are higher than CCK levels. Both peptides contain a sulfated tyrosine residue; CCK is completely sulfated, while some percentage of gastrin is not sulfated.

As mentioned, CCK is abundant and widely distributed in the brain. CCK-positive cell bodies are found in the cerebral cortex, amygdala (Amy), thalamus, hippocampus (Hi), olfactory bulb, and hypothalamus, and in various nuclei in the mesencephalon, ventral tegmental area (VTA), and substantia nigra (SN). The distribution of CCK cell bodies and their projections is shown in Figure 1.

CCK Biosynthesis

CCK is produced by a number of specific modifications of procholecystokinin (pro CCK) (Figure 2) as it passes through the secretory pathway. The sequence and structure of pro CCK and gastrin are similar and they both display tissue-specific processing. CCK 8 is the major form in brain, while intestine has larger forms.

The signal sequence of pre-pro CCK allows it to be inserted into the lumen of the endoplasmic reticulum (ER), where it is removed by the signalase enzyme located on the luminal surface of the ER membrane. Pro CCK is then folded and transported through the secretory pathway. In the trans-Golgi network (TGN), three out of four tyrosine residues in pro CCK are sulfated by protein tyrosine sulfotransferase, and pro CCK is sorted into secretory granules with the enzymes responsible for its processing. The processing of pro CCK occurs mainly at single Arg or Lys residues. The most likely candidate enzymes responsible for the endoproteolytic cleavages in CCK during its processing are the prohormone convertases (PCs), most notably PC1, PC2, and PC5. They are widely co-localized with CCK in the brain, and the inactivation of PC1 and PC2 causes tissue-specific decreases in CCK levels. After pro CCK is cleaved by PC-like enzymes, the extra arginine residues on the C-terminus of CCK 8 are removed by a carboxypeptidase-like enzyme, leaving a terminal glycine residue. The last step in CCK biosynthesis is the amidation of glycine-extended CCK peptides by peptidyglycine-α-amidating monooxygenase.

CCK Receptor Discovery and Distribution

The actions of CCK and gastrin are mediated through two distinct CCK receptors, called 1 and 2. The CCK 1 subtype requires the seven-amino-acid C-terminal peptide with its sulfated tyrosine residue. Unsulfated CCK 8, CCK 4, and gastrin are two to three orders of magnitude less potent. For the CCK 2 subtype, which detects gastrin equally and is sometimes called the gastrin receptor, only the four C-terminal amino acids are required. Both of these receptors have been cloned, and they share considerable homology. They are both classical seven-membrane-spanning domain receptors that interact with G-proteins after agonist binding.

Hundreds of CCK receptor antagonists have been synthesized; these differentiate between the CCK receptors and have allowed a detailed analysis of their pharmacology. Some selective agonists have also been made. The CCK 2 receptor is more abundant and widely distributed in the brain than is the CCK 1 receptor, although distinct populations of the CCK 1 receptor are present. The CCK 1 receptor is present on soma and fibers in the SN, VTA, nucleus accumbens (Acb), and dorsal striatum (DS), but not in the prefrontal cerebral cortex (PFC), Amy, or Hi. The CCK 2 receptor is present on soma and fibers in the SN, VTA, PFC, Acb, DS, Hi, and Amy.

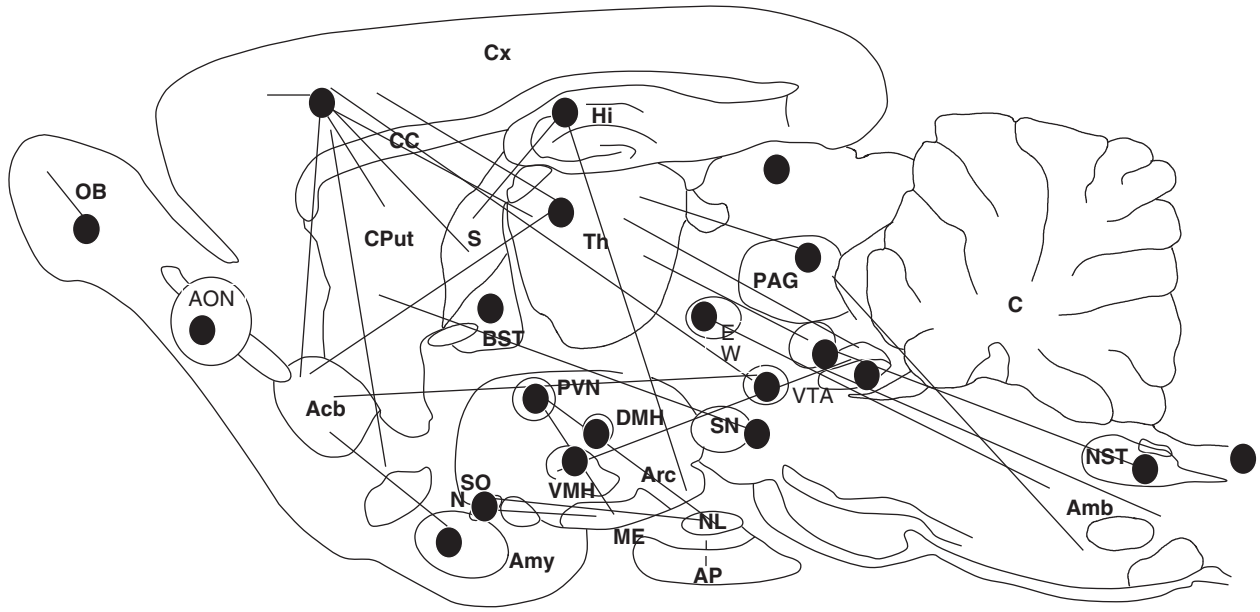

Figure 1 Distribution of CCK neurons in the rat brain and pituitary. The solid circles represent CCK cell bodies and the straight lines represent their projections. Acb, nucleus accumbens; Amb, nucleus ambiguous; Amy, amygdala; AON, anterior olfactory nucleus; AP, anterior pituitary; Arc, arcuate nucleus of the hypothalamus; BST, bed nucleus of the stria terminalis; C, cerebellum; CC, corpus callosum; CPut, caudate-putamen; Cx, cerebral cortex; DMH, dorsal medial nucleus of the hypothalamus; EW, Eddinger–Westphal nucleus; Hi, hippocampus; ME, median eminence; NL, neural lobe of the pituitary; NST, nucleus of the stria terminalis; OB, olfactory bulb; PAG, periaqueductal gray; PVN, paraventricular nucleus of the hypothalamus; S, septum; SN, substantia nigra; SON, supraoptic nucleus of the hypothalamus; Th, thalamus; VMH, ventromedial nucleus of the hypothalamus; VTA, ventral tegmental area.

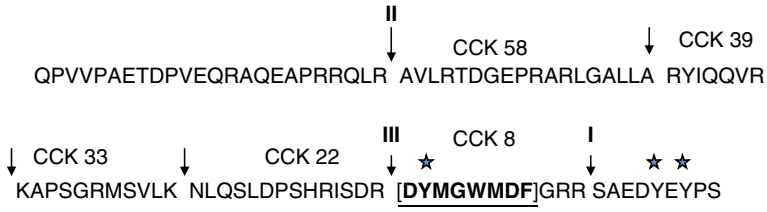

Figure 2 Sequence of mouse pro CCK. The cleavage sites are indicated by spaces and downward arrows; the CCK 8 sequence is underlined. CCK peptides are numbered backwards from the C-terminus of CCK 8. Stars mark the locations of the sulfated tyrosine residues. The order of cleavages producing CCK 8 is indicated by the roman numerals.

In the antrum of the stomach, CCK 2 receptors are present on D cells, while in the fundus, they are found on enterochromaffin-like cells, D cells, and parietal cells. CCK 1 receptors are found in the fundus on chief cells and in antral D cells and on pancreatic acinar cells in rodents but not humans. CCK 2 receptors are found in the pancreas on glucagon-producing cells in the islets.

The 'classical' biologically active processed forms of CCK and gastrin have a C-terminal amide moiety; however, glycine-extended CCK and gastrin peptides are also present. Both amidated and glycine-extended forms of CCK and gastrin are thought to promote growth of the gastrointestinal tract and may be involved in development and maintenance of gastrointestinal cancers. To better understand the trophic activity of glycine-extended CCK, effort has been expended in trying to clone a specific CCKGly or gastrinGly receptor. To date this has not succeeded.

CCK Receptor Structure and Function

The structure and activity of CCK receptors have been studied in great detail. Covalent chemical linkage of the peptide with the receptors followed by determination of where it binds has indicated that the two CCK receptors use different parts of their structure to bind the same ligands. The existence of receptors that are active in the absence of the peptide ligands (constitutively active) has been demonstrated and it appears that these receptors may contribute to gastrointestinal cancers.

Both the CCK 1 and the CCK 2 receptors have been shown to dimerize with each other. This dimerization

of the CCK 1 receptor was disrupted by binding of agonist, while CCK 2 receptor was not. The CCK 1 receptor has also been shown to dimerize with the CCK 2 receptor and to bind agonists normally. The receptor dimers displayed enhanced agonist-induced cell signaling and slower agonist-stimulated internalization. Cells expressing both receptors displayed enhanced growth, suggesting a possible role in cancer.

Intracellular Mediators of CCK Receptor Activation

The model that has been studied in the most detail is the rodent pancreatic acinar cell, which expresses mainly CCK 1 receptors. The AR42J pancreatic cell line and adrenal chromaffin cells which express the CCK 2 receptor have also been studied. In pancreatic acinar cells, CCK receptor activation recruits the heterotrimeric G-proteins G_q and G_{11} and results in increased levels of inositol trisphosphate, increased intracellular calcium, and activation of protein kinase C. CCK at higher concentrations can activate adenylate cyclase, increasing cyclic adenosine monophosphate (cAMP) levels. CCK also activates three mitogen-activated protein kinase (MAPK) cascades leading to extracellular signal-regulated kinase (ERK), c-Jun N-terminal kinase (JNK), and p38 MAPK. CCK also activates the nuclear factor-B (NF-B) pathway and the phosphatidylinositol 3-kinase/protein kinase B/molecular target of rapamycin (PI3K/PKB/m-TOR) pathway, which regulates protein synthesis. Activation of the CCK 2 receptor in gastric cancer cells activates protein kinase D2 and in human adrenal chromaffin cells it increases adenylate cyclase and cAMP.

Biological Activity of CCK

The actions of CCK in the brain are diverse and have been the subject of intense investigation. The development of specific agonists and antagonists and the use of rodent models that lack the receptors or the peptide have facilitated progress in elucidation of the functions of CCK. CCK released into circulation from the intestine following ingestion of food binds to CCK 1 receptors on the vagus nerve in the stomach and sends a signal to the brain to terminate the meal. CCK and leptin are thought to act together to reduce food intake.

CCK acting through the CCK 2 receptor on neurons in the rostral ventromedial medulla (RVM) has antiopiate activity. It inhibits morphine analgesia by inhibiting activation of pain-inhibiting output neurons of the RVM. CCK given into the RVM causes hypersensitivity that is antagonized by CCK 2 antagonist or by lesioning the dorsolateral funiculus. Continuous morphine produces sustained hypersensitivity that is reversed by CCK 2 antagonist and resulted in a fivefold increase in basal CCK levels in the RVM.

CCK is co-localized with dopamine (DA) in the VTA, and these neurons project to medial and posterior subdivisions (shell) of the Acb, the olfactory tubercle, and prefrontal cortex. CCK 1 receptors in the Acb shell display DA agonist-like activities (increased DA release and enhancement of DA-mediated behaviors), while in lateral portions of the Acb (core), CCK 2 receptors display DA antagonist-like activities (decreased DA release and inhibition of DA-mediated behaviors).

CCK acting through the CCK 2 receptor in the Hi is involved in spatial memory. CCK improves spatial memory in elderly rats and in rats in which DA has been depleted. CCK acting through the CCK 1 receptor is involved in maternal behavior and in infant–maternal recognition in sheep and in appetitive learning in infant rats. CCK released by a meal activates peripheral CCK 1 receptors to initiate these behaviors. CCK interacting with estrogen and endogenous opiates in the hypothalamus is involved in regulating sexual behavior in female rodents. A large body of literature has described the ability of CCK to trigger panic attacks in humans and to produce anxiety-related behaviors in rodents and primates. This is mediated through the CCK 2 receptor.

See also: Autonomic Nervous System: Gastrointestinal Control; Mammalian Neuropeptide Families; Neuropeptide Receptors – Drug Development; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Discovery.

Further Reading

Dong M, Liu G, Pinon DI, et al. (2005) Differential docking of high-affinity peptide ligands to type A and B cholecystokinin receptors demonstrated by photoaffinity labeling. *Biochemistry* 44: 6693–6700.

Fourmy D, Escrieut C, Archer E, et al. (2002) Structure of cholecystokinin receptor binding sites and mechanism of activation/inactivation by agonists/antagonists. *Pharmacology and Toxicology* 91: 313–320.

Kopin AS, Lee Y-M, McBride EW, et al. (1992) Expression cloning and characterization of the canine parietal cell gastrin receptor. *Proceedings of the National Academy of Sciences of the United States of America* 89: 3605–3609.

Noble F, Wank SA, Crawley JN, et al. (1999) International Union of Pharmacology. XXI. Structure, distribution, and functions of cholecystokinin receptors. *Pharmacological Reviews* 51: 745–781.

Wank SA, Harkins R, Jensen RT, et al. (1992) Purification, molecular cloning, and functional expression of the cholecystokinin receptor from rat pancreas. *Proceedings of the National Academy of Sciences of the United States of America* 89: 3125–3129.

Cell Adhesion Molecules at Synapses

L F Reichardt and S-H Lee, University of California at San Francisco, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The structures and major intracellular interactions of several different synaptic adhesion molecules believed to play important roles at synapses are shown in Figure 1. One large class is the cadherin superfamily with more than 100 members which are defined by the presence of one or more ~110-amino-acid cadherin repeats in their extracellular domains. The cadherin superfamily is subdivided into classical or type I cadherins, atypical or type II cadherins, desmosomal cadherins, protocadherins, seven-pass transmembrane cadherins, Fat cadherins, etc. A second large group of adhesion molecules is the immunoglobulin (Ig) superfamily, members of which contain varying number of extracellular Ig domains. In some Ig superfamily members fibronectin type III (FNIII) repeats are inserted between the Ig and transmembrane domains. Neural cell adhesion molecule (NCAM), synaptic cell adhesion molecule (SynCAM), L1, the nectins, Side kicks (Sdk), synaptogenesis (SYG)/nephrin, dendrite arborization and synaptic maturation1 (Dasm1), synaptic adhesion-like molecule (SALM), leukocyte common antigen-related-receptor proteintyrosine phosphatases (LAR-RPTPs), and many other members of this family have been reported to play roles in synapse development in both vertebrates and invertebrates (Table 1). The neuroligins and neurexins are heterophilic adhesion molecules characterized by the presence of an acetylcholine esterase-like domain in the former and 1–6 LNS (laminin A, neurexin, sex hormone-binding protein) domains in the latter. Multiple genes and extensive splicing generate many isoforms of these proteins that promote differentiation of both excitatory and inhibitory synapses. Interactions mediated through constituents of the extracellular matrix (ECM) are essential for synapse development at the neuromuscular junction (NMJ), and also appear to be important modulators of central nervous system (CNS) synapse development and function. Both classical ECM receptors (including integrins and the dystroglycan complex) and novel receptors (such as the receptor tyrosine kinase MUSK) mediate synapse development through ligation to ECM constituents. Virtually all of the synaptic adhesion molecules described above interact with scaffolds and other proteins within the cytoplasm. These interactions function to localize synaptic adhesion molecules at synapses and to permit efficient activation of intracellular signaling cascades after synaptic adhesion molecule ligation. Many synaptic adhesion molecules contain PDZ (PSD-95/Dlg/ZO-1) domain-binding, polyproline and other motifs which interact with PDZ, SH3, and other domains of scaffold proteins. Scaffold proteins act as key organizers to control localization of SNAREs, receptors, other synaptic proteins and organelles on both sides of the synapse.

Classical Cadherins

Classic cadherins are a family of calcium-dependent, homophilic, cell–cell adhesion molecules that have been implicated in many developmental processes. Several studies indicate that N-cadherin functions as a synaptic recognition molecule. In *Drosophila*, absence of N-cadherin prevents synapse formation by the axons of photoreceptors in the medulla. The axons from the individual photoreceptors fail to establish stable connections with their target neurons in the medulla. Although analyzed in less detail, inhibition of cadherin function in the chick optic tectum also disrupts synapse formation by the axons of retinal ganglion cells. Extrapolating from these observations, it seems highly likely that a large number of cadherin members expressed in the vertebrate CNS also affect synapse formation with many having overlapping and redundant functions. Mutation of cadherin-11 has been shown to result in enhanced synaptic plasticity (long-term potentiation (LTP)) in the hippocampus through unknown mechanisms.

Much of the progress in understanding the roles of cadherins in forming adhesive junctions has been made in studies of epithelial adherens junctions that share many molecular components with neuronal synapses. These discoveries have provided important insights into the pathways regulating synapse formation. The 'core' cadherin complex is composed of a cadherin bound through its cytoplasmic domain to β-catenin which is bound in turn to α-catenin. In addition, the membrane-proximal portion of the cadherin cytoplasmic domain interacts less strongly with members of the p120 catenin family (p120 catenin, δ-catenin, p0071, and ARVCF). p120 catenins have been shown to regulate cadherin surface stability and clustering as well as regulate Rho-family GTPases. Cadherin adhesion regulates organization of the actin cytoskeleton via several pathways through (1) direct interaction of α-catenin with actin or actin-binding proteins, including α-actinin, vinculin, spectrin, afadin,

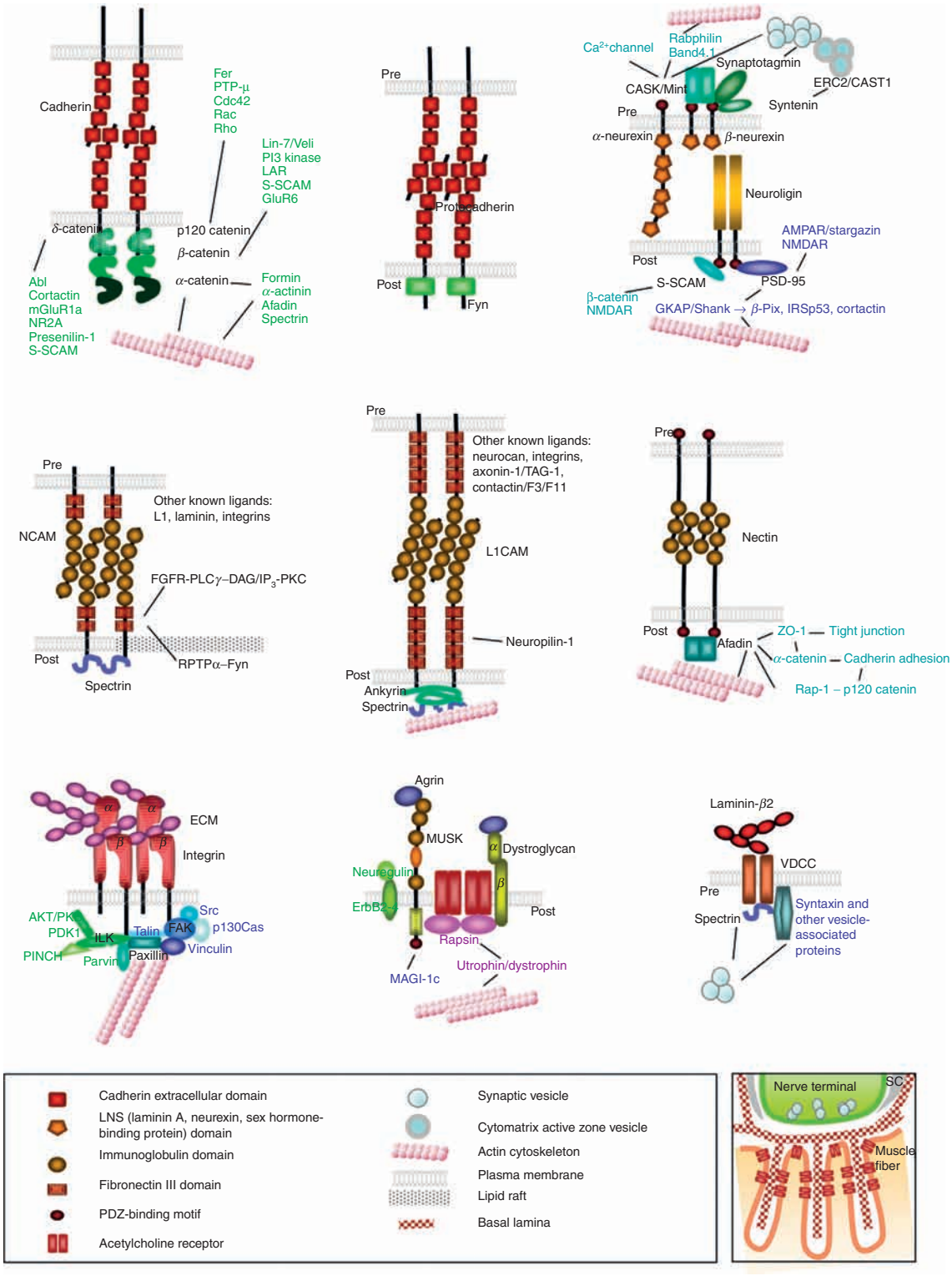

Figure 1 Structures and interactions of synaptic adhesion molecules. Top row of three diagrams illustrate interactions of the classical cadherin complex (left), protocadherin complex (middle), and neuroligin/neurexin complex (right). The second row depicts complexes formed by N(neural)CAM (left), L1-family CAMs (middle), and nectins (right). The bottom row illustrates interactions mediated as a result of integrin–ECM interactions (left), agrin/MUSK/dystroglycan interactions (middle), and laminin $\beta 2$–calcium channel association (right). A lower power schematic of the NMJ is depicted at the lower right corner of this figure. Agrin, laminin $\beta 2$, and laminin $\alpha 4$ chain-containing laminin trimers reside in the basal lamina depicted in this figure, while the agrin receptor MUSK is localized to the muscle plasma membrane. The three EGF-like repeats are not depicted in the β-neurexin structure. Pre, presynaptic compartment; Post, postsynaptic compartment; VDCC, voltage-dependent calcium channel; SC, Schwann cells. See text for details.

Table 1 Roles of adhesion molecules at the synapse

Function	*Molecule*	*Class*	*Place of action*	*Comment*
Target recognition	N-cadherin	Cadherin	Chick optic tectum, *Drosophila* visual system	Synaptic recognition
	Protocadherins	Cadherin	CNS?, spinal cord	Determination of synaptic connectivity
	Flamingo	Cadherin	*Drosophila* visual system	Interactions of R-cell growth cones with target neurons
	Neurofascin	Ig	Mouse GABAergic cerebellar Purkinje cells	Subcellular specification of GABAergic synapses
	Sidekick	Ig	Mouse retina	Lamina-specific synaptic connectivity in inner plexiform layer
	Nectin	Ig	Hippocampal pyramidal neurons	Regulation of axon–dendrite interactions
Induction and maturation of synapse	SynCAM	Ig	Hippocampal pyramidal neurons	Induction of synaptic differentiation
	Dasm, SALM	Ig	Hippocampal pyramidal neurons	Promotion of surface AMPA receptor expression and synaptic maturation
	LAR-RPTP	Ig	Hippocampal pyramidal neurons	Promotion of surface AMPA receptor expression and synaptic maturation
	Neuroligin/ β-neurexin		Hippocampal pyramidal neurons	Induction of synaptic differentiation
	Laminin β2 Laminin β4	ECM	Neuromuscular junction (NMJ) in *Drosophila* and mouse	Organization of active zone, alignment and maintenance of NMJ
	Reelin	ECM	Mouse cortex and cerebellum	Maturation of CNS synapses
	Syndecan-2	Proteoglycan	*Drosophila* NMJ, hippocampal pyramidal neurons	Postsynaptic maturation
	Versican	ECM proteoglycan	Chick optic tectum	Lamina-specific cue for presynaptic maturation of retinal ganglion axons
	CPG15 (neuritin)	GPI-linked surface protein	Motor neurons	Regulation of axonal branching and synapse maturation
Synaptic plasticity	Cadherin-11	Cadherin	Hippocampal pyramidal neurons	Regulation of hippocampal LTP
	NCAM, FasII, apCAM	Ig	Vertebrate CNS, *Drosophila* NMJ and *Aplysia* sensory neurons	Regulation of new synapse formation and structural synaptic plasticity
	Integrins	ECM receptor	Central excitatory synapse	Modulation of NMDA-mediated synaptic currents and synaptic plasticity

ZO-1, and formin-1 and through (2) regulation by p120 catenins of Rho GTPases and cortactin. Numerous kinases and phosphatases regulate the properties, stability, and interactions of the cadherin complex, including the protein kinases Cdk5, casein kinase I, Fyn, Yes, and Fer, and the protein tyrosine phosphatases PTP-μ, LAR, SHP1/2, and PTP1B. In addition to regulators of the actin cytoskeleton, the proteins in the cadherin complex interact with many additional proteins that control intracellular signaling pathways, including transcriptional regulators like Kaiso and T-cell factors (TCFs), scaffold proteins (such as A-kinase anchoring protein (AKAP)), and cytoplasmic enzymes including PI3 kinase, Cdk5, and receptor tyrosine kinases, plus cytoplasmic motor proteins and other proteins including presenillin. In addition, β-catenin and δ-catenin interact with PDZ-containing scaffold proteins that also mediate their synaptic functions.

Several of the catenins appear to function as effectors of cadherins in controlling synapse development. β-Catenin functions as a scaffold to localize synaptic vesicle pools to synaptic sites. The presence of α-N-catenin is required for normal spine maturation and stability. The absence of either δ-catenin or p120 catenin severely perturbs central synapses. Absence of p120 catenin reduces dramatically the density of dendritic spines and synapses. Absence of δ-catenin results in gross abnormalities in synaptic function and plasticity. In part, these proteins may function through regulation of the small G-proteins: Rac and Rho. They may also function through stabilization of surface cadherins or through regulation of tyrosine phosphorylation and control of cortactin activity.

Protocadherins

Protocadherins are characterized by the presence of four to seven extracellular cadherin (EC) motifs, which are distinguishable from those present in other cadherin members, and absence of the conserved

motifs in the classic cadherins that mediate interactions with β-catenin and p120 catenin family members. About 80 different protocadherins have been identified and genes of nearly 60 members are tandemly arrayed in three clusters: the α-, β-, γ-protocadherin genes (*Pcdh*-α, -β, and -γ) on a single chromosome in mammals. (α-Protocadherins are also known as CNRs, cadherin-related neuronal receptors.) Recent phylogenetic analysis has characterized several conserved motifs shared by an additional subfamily of protocadherins, the δ-protocadherins, whose genes are located on several different chromosomes. These include Pch1, Pch7, Pch8 (arcardlin), Pch9, and Pch11. The *Pcdh*-α and *Pch*-γ clusters contain two large genomic regions, with tandem arrays of large exons expressing the variable extracellular domains followed by a set of common exons, each encoding a small portion of the shared C-terminal cytoplasmic domain. *Pch*-β lacks a shared constant region. Protocadherins are highly expressed in the CNS with individual members differentially expressed in subsets of neurons, suggesting that they help generate specific synaptic connections within the CNS. Genetic deletion of the *Pch*-γ locus results in a reduction of inhibitory synapse density in the spinal cord. The mechanisms through which the protocadherins regulate synaptogenesis are not well understood. The strength of homophilic adhesion mediated by the extracellular domains of the protocadherins appears to be weaker than adhesion mediated by classic cadherin interactions. Recent evidence suggests that some protocadherins function through regulation of classic cadherin adhesion. Protocadherins may also have noncadherin proteins as ligands. For example, at least one of the Pch-α isoforms, Pch-α4, interacts through its RGD motif with β1 integrins. Protocadherin-7 interacts with PP1, a protein tyrosine phosphatase that dephosphorylates AMPA receptors and CAM kinase II.

NCAM

NCAM was the first Ig superfamily member reported to function at the synapse. NCAM is encoded by a single gene with three isoforms of 180, 140, and 120 kDa produced by alternative splicing. The two larger isoforms are transmembrane proteins, differing in the presence of a differentially spliced exon that encodes an insert in the cytoplasmic domain, whereas NCAM-120 is a glycophosphatidyl inositol (GPI)-linked protein that lacks a transmembrane and cytoplasmic domain. All isoforms have five Ig domains followed by two FNIII repeats. The NCAMs are believed to mediate both homophilic and heterophilic interactions. Recently identified interacting partners include the Ig-family cell adhesion molecule L1, the fibroblast growth factor (FGF) receptor, laminin, and some integrins. Intracellular partners include the protein tyrosine phosphatase RPTPα, the tyrosine kinase p59Fyn, and spectrin. The most remarkable feature of NCAM is that it is virtually the sole acceptor for long linear homopolymers of α2,8-linked sialic acid residues (polysialic acid, PSA). The major phenotypes of the NCAM knockout can be mimicked by the enzymatic removal of PSA, indicating that the major essential function of NCAM is to provide a substrate for PSA addition, which is highly charged, thereby increasing the volume of extracellular space and inhibiting adhesive interactions mediated through other cell adhesion molecules.

In vivo and *in vitro* results have demonstrated that NCAM regulates synaptic maturation at the NMJ and hippocampal synaptic plasticity (LTP), in part through associated PSA homopolymers. The NCAM homologs in invertebrates, *Drosophila* fasciclin II (FasII) and *Aplysia* apCAM, are involved in synapse growth, stability, and plasticity, and the expression of each is regulated by neuronal activity. Of particular interest, fasciclin II has been shown to function as both a permissive and restrictive regulator of synapse growth at the *Drosophila* NMJ, depending upon the comparative pre- and postsynaptic levels of this protein. The ability of fasciclin II to promote new synapse formation requires the presence of the *Drosophila* orthologs of amyloid precursor protein and the scaffold protein Mint. Fasciclin II forms a complex with each of these proteins, the latter through its PDZ-interaction motif. In *Aplysia* sensory neurons, downregulation of the NCAM homolog, apCAM, is required for the serotonin-induced enhancement of synaptic varicosity formation that accompanies synaptic facilitation. Normally, apCAM is downregulated following synaptic facilitation as a result of phosphorylation at a putative mitogen activated protein (MAP) kinase site. Overexpression of apCAM in sensory, but not motor neurons, completely suppresses the formation of new varicosities without interfering with the activation by serotonin of CREB. Thus, studies in both *Drosophila* and *Aplysia* indicate that NCAM homologs are important effectors of signaling pathways that control synaptic structural plasticity.

L1-Family CAMs

The L1-type cell adhesion molecules are glycoproteins with six Ig, five FNIII, a single transmembrane, and a cytoplasmic domain. Four members (L1-CAM, Neurofascin, Nr-CAM, and CHL-1) have been identified in vertebrates while *Drosophila* has only one homolog, Neuroglian. L1-type CAMs mediate several aspects of neuronal development, including neuronal migration, neurite extension, axon

pathfinding, myelination, fasciculation, synapse targeting, and LTP. Among these family members, the role and function of L1 have been particularly intensely studied. In addition to binding L1 molecules present on other cells, L1 associates with additional ligands, including the proteoglycan neurocan, several RGD-binding integrins, and the CAMs axonin-1/TAG, and contactin/F3/F11. L1 also interacts with the axon guidance receptor neurophilin in both *cis* and *trans* forms to form a receptor for Semaphorin 3A.

Of particular interest, localization of neurofascin at the initial axon segment of cerebellar Purkinje neurons results in specific localization of GABAergic synapses to this same region of the axon. At early stages of development both GABAergic synapses and neurofascin are less well localized. Refinement of neurofascin localization precedes and is required for localization of the GABAergic synapses. Neurofascin binds with high affinity to an isoform of ankyrin, an intracellular adaptor protein which is localized also to the initial axon segment and provides a scaffold for association with β-spectrin and the F-actin cytoskeleton. These interactions are required to localize both neurofascin and the GABAergic synapses. Observations in other CNS regions indicate that other L1 CAM family members are also preferentially localized to subdomains of neurons. It seems likely that they also help localize synapse formation. A number of human brain disorders have been linked to mutations in the L1 gene and L1 knockout mice reproduce several aspects of these phenotypes. It will be interesting to determine whether any aspects of the phenotypes associated with these disorders result from mis-targeting of synapses.

Other Synaptic Ig Superfamily Members

SYG-1 is a member of the Ig superfamily identified in a genetic screen for synapse mutants in *Caenorhabditis elegans* that contains four Ig's, a single transmembrane, and a cytoplasmic domain with a PDZ interaction motif. SYG-1 regulates the position of presynaptic terminal differentiation by HSNL axons which innervate the vulval muscles. It is related to *Drosophila* Duf/Kirre and the three vertebrate NEPH proteins that are important for organization of the slit diaphragm in the kidney glomerulus. SYG-1 binds to SYG-2, which has six Ig domains and one FNIII domain and is expressed transiently by primary vulval epithelial cells during synapse formation. SYG-2 provides an intermediate target that focuses accumulation of SYG-1 and synaptic vesicles. The mammalian homolog of SYG-2 is nephrin whose presence in podocytes is also essential for organization of the slit diaphragm in the kidney. The roles of the NEPH family and nephrin in vertebrate synapses are not yet known, but recent work, stimulated by the observations in *C. elegans*, has documented their expression patterns in the CNS. In addition, the NEPH proteins have been shown to interact with CASK, a constituent of the CASK/MINT/Veli complex, which is thought to participate in pre- and postsynaptic organization. Since both nephrin and the NEPH proteins are expressed in the CNS, it will not be surprising if they prove to have important functions there.

The Sdk CAMS, Sdk-1 and Sdk-2, contain six Ig domains, 13 FNIII repeats, a single transmembrane domain, and a cytoplasmic domain terminated by a PDZ domain-binding motif. Interestingly, in the retina the Sdk-1 and Sdk-2 proteins are found in nonoverlapping subsets of presynaptic (amacrine and bipolar) and postsynaptic (ganglion) cells that project to common inner plexiform (synaptic) sublaminae. Sdk-positive synapses are confined to specific sublaminae, and ectopic expression of Sdk in Sdk-negative cells redirects their arbors to Sdk-positive sublaminae, implying that the Sdk's may control laminar specificity within the retina. The interactions and signaling mechanisms controlled by the Sdks that regulate synapse location are not known.

Dasm1 is a recently reported Ig family member with five Ig-like domains, two FNIII domains, and a cytoplasmic domain with a C-terminal PDZ domain-binding motif. Dasm has been shown to promote dendritic elongation and branching as well as the maturation of excitatory glutamatergic synapses. Dasm interacts with PDZ-containing scaffold proteins, including Shank and S-SCAM. In its absence, α-amino-3-hydroxy-5-methylsoxazole propionate (AMPA) receptor-mediated, but not *N*-methyl-D-aspartate (NMDA) receptor-mediated, synaptic transmission is severely impaired in cultured hippocampal neurons. A Dasm mutant lacking the PDZ-interaction motif does not promote AMPA receptor-mediated synaptic transmission. The extracellular ligands for Dasm have not been identified.

The SALMs are another recently characterized family of five synaptic adhesion molecules (SALM1–5) that contain six leucine-rich repeats plus single Ig and FNIII repeats in their extracellular domains. The SALMs are transmembrane proteins with cytoplasmic domains terminated by a PDZ domain-binding motif. Some members of the SALM family have been shown to promote excitatory synapse formation through interactions with PSD-95 and other cytoplasmic scaffold proteins. Members have also been shown to interact with NMDA and AMPA receptors. It is not clear whether they function as typical cell adhesion molecules by mediating adhesion with other cells.

The four nectins (nectin 1–4) are transmembrane proteins containing three extracellular Ig-like repeats and a cytoplasmic domain terminated by a PDZ domain-binding motif. While each of the nectins mediates homophilic adhesion, they promote stronger heterophilic adhesion with other family members. Several nectins also interact with the integrin $\alpha_V\beta_3$. Each of the nectins interacts with the F-actin-binding protein afadin, thereby anchoring the nectins directly to the F-actin cytoskeleton. Through afadin-mediated interactions with α-catenin, the nectins recruit and collaborate with the cadherins to promote formation of adherens junctions in epithelial cells. Afadin also binds the small G-protein Rap1. The Rap1–Afadin complex has been shown to interact with p120 catenin and through p120 catenin to activate cadherin-mediated adhesion. Active (GTP-bound) Rap promotes maturation of dendritic spines in the CNS and it seems possible that this is, in part, through regulation of the afadin–nectin–cadherin complex. Nectin-1 is preferentially expressed in axons. Preferential engagement of heterophilic adhesion between axonal nectin1 and dendritic nectin3 may provide a basis for the preferential interactions of axons with dendrites versus other axons.

Members of the LAR family of receptor protein tyrosine phosphatases are transmembrane proteins with each containing three Ig and eight FNIII domains plus two cytoplasmic tyrosine phosphatase motifs. The cytoplasmic domains of these proteins interact with liprin-α/SYD-2, which binds to the PDZ-containing scaffold protein GRIP. This interaction promotes the surface expression and clustering of receptors at postsynaptic sites. In *Drosophila* and *C. elegans*, LAR and liprin-α have been also implicated in presynaptic terminal differentiation. In *Drosophila*, LAR-mediated synaptic differentiation is regulated by interactions of LAR with the cell surface-associated proteoglycans syndecan and dally-like. In cultured hippocampal neurons, LAR is concentrated at mature excitatory synapses where it promotes their development and maintenance through enhanced spine formation and surface AMPA receptor expression.

The four members of the SynCAM family (also known as nectin-like proteins) constitute a small subset of the Ig superfamily, each member of which contains three extracellular Ig domains and a PDZ-interacting motif in its short cytoplasmic domain. They were identified in a genomic search for members of the Ig superfamily capable of interacting with PDZ-domain-containing scaffolds. Similar to many other Ig family members, the SynCAMs mediate homophilic cell–cell adhesion. SynCAM1 is highly enriched in both presynaptic and postsynaptic membranes of neurons, whereas SynCAM3 is found in nonjunctional contact sites of presynaptic nerve terminals, axons, and glia cell processes. Clustering of SynCAM1 is sufficient to trigger differentiation of functional presynaptic terminals. SynCAM1 overexpression also increases synaptic function in immature excitatory neurons. Interactions mediated through association with PDZ-domain-containing scaffolds are essential for its synapse-inducing activities. The PDZ domain-binding motif in SynCAM binds to the PDZ domains present in CASK and syntenin.

Finally, another family of Ig superfamily members, named the netrin-G ligands (NGLs), have recently been implicated as regulators of synapse formation. The structure of the NGLs includes leucine-rich repeats plus an Ig domain in their extracellular domains, a single-pass transmembrane domain and a cytoplasmic domain terminated by a PDZ domain-binding motif. Three genes encode members of this family whose diversity is further increased by extensive differential splicing. The NGLs interact with the two members of the netrin-G family, proteins associated with the cell membrane through GPI motifs that are structurally related to, but functionally distinct from, the classical netrins that regulate axon guidance. The NGLs bind PSD-95, and NGL2 forms a complex in the brain with PSD-95 and NMDA receptors. Clustering of NGL2 has been shown to promote postsynaptic differentiation of excitatory, but not inhibitory synapses. NGL2 also induces presynaptic differentiation by contacting axons *in vitro* and it seems likely, although not proved, that differentiation is mediated through binding to netrin-Gs. Thus, interactions of the NGLs with nectin-G-proteins appear to be able to promote both pre- and postsynaptic differentiation *in vitro* and are strong candidates to control both the formation and specificity of synapse formation *in vivo*.

Neurexins and Neuroligins

Neuroligins are composed of a cholinesterase-like domain, an *O*-glycosylation region, a single transmembrane domain and a short C-terminal cytoplasmic tail containing a PDZ-binding motif. Three genes encoding neuroligins are present in the rat and mouse while five are present in the human genome. The first neurexin to be characterized was identified as a receptor for the spider venom constituent α-latrotoxin. There are three neurexin genes, each of which encodes two neurexin isoforms with one promoter initiating transcription of a long mRNA encoding α-neurexin isoforms while a second promoter located within an intron initiates transcription of a shorter mRNA encoding β-neurexin isoforms. The extracellular domain of α-neurexin contains three

epidermal growth factor (EGF)-like domains, each of which is flanked by a pair of LNS domains. Instead of six LNS domains, only one is present in the β-neurexins, which lack all of the EGF repeats. The extracellular domains of both α- and β-neurexins are followed by a transmembrane domain and a cytoplasmic domain containing a C-terminal PDZ-interacting site. Differential splicing of the transcripts from each of the neurexin and neuroligin genes generates a tremendous number of variants of each protein family. Recent results have demonstrated that splicing controls the specificity of neurexin–neuroligin interactions as well as their activities in promoting excitatory and inhibitory synapse formation. In addition to binding neuroligins, some, but not all isoforms of neurexins interact with dystroglycan.

On cultured hippocampal neurons, clustering of the β-neurexins or neuroligins triggers pre- or postsynaptic differentiation, respectively. In contrast to SynCAM, however, neuroligin overexpression does not result in enhanced synaptic transmission in immature excitatory neurons. Similar to the SynCAMs, the synapse-promoting activities of these proteins are dependent on their interactions with PDZ-containing scaffold proteins. The PDZ-interacting motif in neuroligin interacts with PSD-95 (which recruits AMPA and NMDA receptors) as well as with S-SCAM (which interacts with AMPA receptors and β-catenin). β-Neurexin interacts with the CASK/Mint/Veli complex, which is proposed to act as a scaffold for synaptic vesicle transport and localization. β-Neurexin also interacts with syntenin, which has recently been shown to interact with the presynaptic cytomatrix active zone protein ELKS/ERC2/CAST1, which binds in turn several additional active zone proteins, including piccolo, bassoon, and RIM. Interestingly, neuroligin1 is mainly localized to excitatory, whereas neuroligin2 is localized at inhibitory synapses. Overexpression and knockdown experiments *in vitro* suggest that the expression levels and localization of different neuroligins may play a crucial role in controlling the ratio of excitatory and inhibitory synapses. Mutations in human neuroligin genes are associated with a small percentage of autistic individuals.

ECM-Mediated Interactions

ECM molecules regulate synapse development in both the central and peripheral nervous systems. At the NMJ, several constituents of the basal lamina have been shown to be essential for normal synapse formation. Most importantly, the presence of nerve-derived agrin in the basal lamina and its receptor, the tyrosine kinase MUSK, in muscle are absolutely necessary for synapse formation and maintenance. In addition, several subunits of laminin also play important roles. The presence of the laminin $\beta 2$ subunit in the laminin trimer within the synaptic basal lamina appears to play two distinct roles. First, this subunit prevents invasion of the synaptic cleft by Schwann cells, thereby preserving communication between the motor axon terminal and the muscle surface. Second, this subunit directly binds to N-type calcium channels. In its absence few active zones are seen at NMJs, even at sites where Schwann cells have not invaded the synaptic cleft, and synaptic transmission is severely impaired. In contrast, although synaptic transmission is not impaired in its absence, the laminin $\alpha 4$ subunit functions to ensure proper alignment of active zones in the nerve terminal with folds in the postsynaptic muscle membrane. In laminin $\alpha 4$ mutants, the active zones and folds are no longer aligned with each other. Laminin $\alpha 4$ is largely excluded from the small regions of basal lamina aligned with the active zones and muscle folds, suggesting that it may inhibit the formation of these specializations. In addition to MUSK, the integrins and dystroglycan complex function as receptors for basal lamina proteins at the NMJ. Absence of $\beta 1$ integrins severely impairs NMJ. Mild deficits are observed in the absence of some constituents of the dystroglycan complex.

Although there have been many studies, there are less obvious roles for ECM constituents at synapses in the CNS. The ECM glycoprotein reelin, best known for its role in controlling neuronal migration in the cortex and cerebellum, also promotes maturation of many CNS synapses through the same signaling pathway important for its role in cell migration – Src-family kinase-mediated phosphorylation of Dab, following engagement of the reelin receptors VLDLR and apoER2. Several additional ECM receptors also affect synapse maturation. For example, the cell surface proteoglycan syndecan-2 is enriched in the mammalian hippocampus and promotes morphological maturation of dendritic spines in cooperation with the receptor tyrosine kinase EphB2. Syndecan-2 is a transmembrane glycoprotein with a cytoplasmic PDZ-interacting motif, which interacts with CASK, syntenin, and synbindin. As described previously, syndecan regulates LAR in *Drosophila*.

In *Drosophila*, elegant studies have demonstrated a crucial role for integrins in memory formation in the mushroom body. In hippocampal neuron cultures, several studies have demonstrated important roles for integrins in promoting synapse maturation. *In vivo* studies have shown that inhibition of RGD binding or β_1 integrin function through genetic or pharmacological techniques impairs long-term plasticity. Ligand engagement by integrins results in activation of several intracellular signaling cascades including those

initiated by Src-family, focal adhesion kinase, and integrin-linked kinases, plus recruitment of scaffolds and adaptor proteins, such as talin and paxillin that lead to reorganization of actin cytoskeleton.

Concluding Remarks

Much exciting evidence demonstrating important roles of synaptic adhesion molecules has been accumulated recently in both *in vivo* and *in vitro* model systems. We anticipate that future studies focusing on the crosstalk between different adhesion molecules will further illuminate our understand of synapse development and function.

See also: Axonal Pathfinding: Extracellular Matrix Role; Cadherins and Synapse Organization; Dendrite Development, Synapse Formation and Elimination; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Lamina-Specific Neuronal Connections; Neural Cell Adhesion Molecules and Synapse Regulation; Neurexins; Neuroligins; Neuromuscular Junction (NMJ): Mammalian Development; Postsynaptic Development: Neuronal Molecular Scaffolds; Postsynaptic Specialization Assembly; Presynaptic Development and Active Zones; Retrograde Transsynaptic Influences; SynCAMs.

Further Reading

Berardi N, Pizzorusso T, and Maffei L (2004) Extracellular matrix and visual cortical plasticity: Freeing the synapse. *Neuron* 44: 905–908.

Chih B, Engelman H, and Scheiffele P (2005) Control of excitatory and inhibitory synapse formation by neuroligins. *Science* 307: 1324–1328.

Davis GW (2006) Homeostatis control of neural activity: From phenomenology to molecular design. *Annual Review of Neuroscience* 29: 307–323.

Dean C and Dresbach T (2006) Neuroligins and neurexins: Linking cell adhesion, synapse formation and cognitive function. *Trends in Neurosciences* 29: 21–29.

Essner JJ, Chen E, and Ekker SC (2006) Syndecan-2. *International Journal of Biochemistry and Cell Biology* 38: 152–156.

Gerrow K and El-Husseini A (2006) Cell adhesion molecules at the synapse. *Frontiers in Bioscience* 11: 2400–2419.

Kim S, Burette A, Chung HS, Kwon S-K, et al. (2006) NGL family PSD-95-interacting adhesion molecules regulate excitatory synapse formation. *Nature Neuroscience* 9: 1294–1301.

Montgomery JM, Zamorano PL, and Garner CC (2004) MAGUKs in synapse assembly and function: An emerging view. *Cellular and Molecular Life Sciences* 61: 911–929.

Rougon G and Hobert O (2003) New insights into the diversity and function of neuronal immunoglobulin superfamily molecules. *Annual Review of Neuroscience* 26: 207–238.

Salinas PC and Price SR (2005) Cadherins and catenins in synapse development. *Current Opinion in Neurobiology* 15: 73–80.

Sanes JR and Lichtman JW (1999) Development of the vertebrate neuromuscular junction. *Annual Review of Neuroscience* 22: 389–442.

Takeichi M and Abe K (2005) Synaptic contact dynamics controlled by cadherin and catenins. *Trends in Cell Biology* 15: 216–221.

VanVactor D, Wall DP, and Johnson KG (2006) Heparan sulfate proteoglycans and the emergence of neuronal connectivity. *Current Opinion in Neurobiology* 16: 40–51.

Waites CL, Craig AM, and Garner CC (2005) Mechanisms of vertebrate synaptogenesis. *Annual Review of Neuroscience* 28: 251–274.

Yamagata M, Sanes JR, and Weiner JA (2003) Synaptic adhesion molecules. *Current Opinion in Cell Biology* 15: 621–632.

Cell Culture: Autonomic and Enteric Neurons

P J Lein, A D Fryer, and D Higgins[†], Oregon Health and Science University, Portland, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The autonomic nervous system, which is responsible for maintaining homeostasis, is divided into the sympathetic and parasympathetic nervous systems. Each system is defined anatomically by preganglionic neurons located in the central nervous system, the axons of which synapse on postganglionic neurons located in the periphery, which in turn directly innervate target organs. While there are published examples of the isolation and maintenance of preganglionic autonomic neurons in culture, it is the peripheral postganglionic neurons that have principally been isolated for tissue culture, and these are the focus of this article. Also discussed is tissue culture of the enteric nervous system, which is distinct from the autonomic nervous system, but, like the sympathetic and parasympathetic nervous systems, is derived from the neural crest.

The culture of autonomic neurons, specifically postganglionic sympathetic neurons, has a long and rich history in neurobiological experimentation. Intact sympathetic ganglia maintained in culture as explants were first described by Lewis and Lewis in 1912, and the observation that a soluble factor produced by a mouse sarcoma stimulated profuse outgrowth of neurites from sympathetic explants in culture led to the discovery of nerve growth factor (NGF) by Levi-Montalcini and colleagues. Since this seminal discovery, sympathetic neuronal cultures have proved to be a powerful tool for identifying additional neurotrophic factors and for elucidating the molecular mechanisms by which NGF and these other factors regulate neuronal survival, differentiation, and function. Beyond this application, primary cultures of sympathetic neurons continue to be a popular and robust model system for elucidating basic cell and molecular mechanisms of neuronal development and differentiation, and for studying cell and molecular mechanisms underlying sympathetic physiology and pharmacology. With the exception of the chick ciliary ganglion, primary cultures of parasympathetic and enteric neurons have not been as widely used, and studies employing these culture models have focused largely on the electrophysiological and pharmacological properties of these neuronal cell types.

Neural Crest Cells

All the cells of the sympathetic, parasympathetic, and enteric nervous systems are derived from the neural crest. The neural crest also gives rise to cell types other than peripheral neurons and glia, and because of this vast repertoire and the fact that differentiation of neural crest cells is strongly influenced by environmental factors, neural crest cell differentiation has emerged as a powerful model system for identifying basic developmental principles governing the generation of diverse cell types in not only the peripheral but also the central nervous system. Neural crest cells have been cultured successfully from both rodent and avian species, and detailed descriptions of techniques for isolating and maintaining neural crest cells from these sources have been published. The development of neural crest cell culture proved critical for confirming that premigratory neural crest cells are pluripotent, and clonal analyses of cultured neural crest cells have yielded much of the experimental evidence upon which is based our current understanding of neural crest cell fate restriction as a function of developmental timing and location along the migratory pathway.

Cultured neural crest cells remain a popular model for identifying environmental cues that influence the migratory behavior and differentiation of neural crest cells, including those progenitor cells that give rise to the sympathetic and enteric nervous systems. For example, the isolation and culture of a sympathoadrenal progenitor from migrating neural crest cells was instrumental in identifying the roles of fibroblast growth factor (FGF) and glucocorticoids in determining whether these progenitor cells differentiate into sympathetic neurons or adrenal chromaffin cells. Studies of cultured sympathoadrenal progenitor cells were also critical for demonstrating that FGF induced NGF responsiveness in these cells, providing some of the first evidence that differentiation signals involve a cascade of responses to multiple agents. Culture of crest-derived cells immunoselected from within the bowel provided critical confirmation of *in vivo* observations suggesting that laminin promoted neurogenesis of enteric precursor cells. This model system has since been used to identify additional factors that promote differentiation of both enteric neurons and glia. Other applications of crest cell cultures include determination of the intracellular signaling pathways that mediate the effects of environmental cues on migration and differentiation of sympathetic and enteric precursors, and

[†]Deceased

the identification of genomic and proteomic profiles associated with specific states of determination.

Sympathetic Neurons

Sympathetic neurons have been cultured from prevertebral and paravertebral ganglia isolated from diverse species, ranging from bullfrogs to humans, at varying developmental stages, including early embryonic development of the ganglia, when a significant percentage of ganglionic neurons are still undergoing cell division, as well the mature and even aging organism. However, cultures of sympathetic neurons are most commonly derived from prenatal (19–21 days) or early postnatal (1–14 days) rat superior cervical ganglia or embryonic chick chain ganglia. The easiest type of culture to prepare is an explant in which the intact ganglion is maintained in culture. Because of the profuse outgrowth of neurites from explanted ganglia, this preparation has been used extensively as a bioassay for neurotrophic factors and for studies of axonal growth cone behavior, including the identification of attractive and repulsive axonal guidance cues. Dissociated cultures, however, allow visualization of neurons in their entirety, which facilitates electrophysiological recording and microinjection and enables studies of sympathetic structure and function at the subcellular level. A third culture paradigm that has been used with sympathetic neurons is the Campenot chamber, in which the media bathing distal axons is physically separated from that bathing neuronal cell bodies and proximal processes (see Figure 1). This culture system consists of attaching a compartmented Teflon insert to a collagen-coated substrate using vacuum grease. Neurons are plated into one compartment and axons grow under the grease barriers into adjacent compartments. Axons are encouraged to grow through the grease barrier by scoring parallel tracks in the collagen substrate prior to attaching the Teflon insert, such that the tracks run from the left compartment into the right compartment. These tracks expose the tissue culture plastic that the axons will not grow across. When assembled correctly, minimal exchange of fluids or solutes takes place between compartments, which allows various factors to be presented selectively to neuronal cell bodies or distal axons and permits selective biochemical analyses of individual compartments. Data collected using these compartmented cultures have significantly advanced our understanding of the molecular mechanisms of anterograde and retrograde signaling. Campenot chambers have also been used successfully with sensory neurons, but to date there are no published articles reporting its adaptation to central neurons. (Detailed protocols for harvesting, plating, and maintaining sympathetic neurons using each of these three culture paradigms are provided in the literature listed under 'Further reading.')

Hundreds of articles have been published describing the behavior of sympathetic neurons in tissue culture. This system has proved popular for several reasons. First, these neurons are relatively easy to isolate, since sympathetic ganglia are prominent, anatomically discrete entities that are accessible at varying life stages, ranging from early ganglionic development (embryonic days 14 and 15 in the mouse and rat, respectively) through adulthood. Second, this preparation yields one of the most homogeneous neuronal cell populations available in vertebrates. Sympathetic ganglia contain a large population of principal neurons (>20 000 in the rat superior cervical ganglion), a small population (approximately 300 cells) of interneurons or small intensely fluorescent cells (SIF), and ganglionic glial cells. Ganglionic glial cells and other nonneuronal cells can be eliminated in culture using antimitotic agents, and interneurons do not survive in

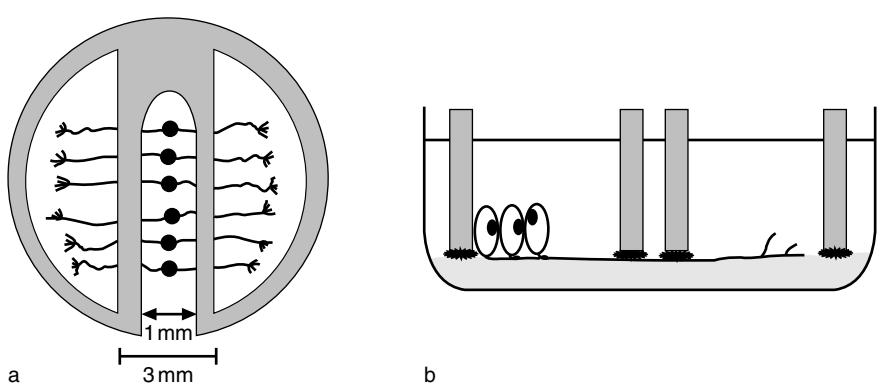

Figure 1 Schematic of Campenot compartmented chambers. (a) Top view, showing neurons plated in the center chamber of the classic Campenot Teflon insert (Camp10, available from Tyler Research Corporation, Edmonton, Alberta, Canada). Neurons typically extend axons under the barriers into the adjacent side chambers within 7 days after plating. (b) Side view, showing neurons plated in the far left chamber that have extended axons through the center chamber and into the far right chamber.

dissociated cell culture in standard serum-free culture media. Thus, changes in these cultures in response to an experimental manipulation tend to be uniform throughout the cell population. Third, the growth factor requirements of sympathetic neurons are well established, which allows the use of defined medium, and a variety of defined media and substrates work well for growing and maintaining sympathetic neurons. These neurons will grow in either the absence or presence of glia or target cells and can be maintained for up to 6 months in culture. Most importantly, when grown under the appropriate *in vitro* conditions, cultured sympathetic neurons express many of the properties expected of mature sympathetic neurons *in situ*, including synthesis and release of norepinephrine, normal electrophysiological properties and responsiveness to appropriate neurotransmitters, and a dendritic morphology that approximates that seen *in situ*. Finally, a wealth of data in the peer-reviewed literature facilitates correlation of *in vitro* and *in situ* sympathetic development, function, and plasticity. The most significant limitations of the model include the limited amounts of RNA and protein available from the sympathetic ganglia, which may restrict biochemical studies, and difficulty in manipulating gene expression. Recent advances in techniques for expressing cDNA in primary neuronal cell cultures are rapidly overcoming this latter disadvantage. There are publications reporting the successful application of lipid-based delivery systems (specifically, Lipofectamine 2000; Invitrogen), nucleofection (Amaxa), and adenoviral vectors to cultured sympathetic neurons. The reported transfection efficiencies are relatively low (ranging from 10% to 20%, with lipofection and nucleofection up to 30% to 50% with viral infection), which is adequate for endpoints based on individual cell analyses, such as morphological analyses, but may be problematic for many biochemical endpoints.

Sympathetic cultures offer particular advantages for the study of neuronal development and plasticity. The development of sympathetic neurons is critically dependent on their environment, and this plasticity is retained in tissue culture, where mechanisms can be more readily studied. It should be noted, however, that the plasticity exhibited by sympathetic neurons in culture has some practical disadvantages. The morphological, electrophysiological, and pharmacological properties of these neurons can be affected by basic parameters of the culture system, including cell density, serum, substratum, and media components such as insulin and potassium. Thus it is necessary to be certain that the culture conditions chosen are appropriate for a particular study. Nonetheless, studies of cultured sympathetic neurons have provided significant insights in terms of the identification and mechanistic characterization of environmental factors that regulate some of the most essential properties of these neurons, such as their morphology, receptor profile, and neurotransmitter phenotype. A significant advantage of sympathetic cultures with respect to studies of neuronal morphogenesis is that, unlike cultured central neurons, it is possible to control when dendritic growth occurs. Thus, sympathetic neurons can be grown under conditions in which <5% of the neuronal cell population extends dendrites, but the addition of the appropriate stimulus (bone morphogenetic proteins or Matrigel) will trigger dendritic growth in >95% of the neuronal cell population (**Figure 2**). Other widespread applications of sympathetic neuronal cultures include studies of (1) growth cone behavior, structure, and biochemical composition, (2) synaptogenesis, (3) the cytoskeleton in a variety of cellular functions, including axonal and dendritic outgrowth, axonal transport, and subcellular trafficking, (4) factors that regulate transmitter release, (5) signal transduction downstream of receptor activation, and (6) molecular mechanisms of apoptosis. A well-established application that is gaining in popularity is the use of sympathetic neurons co-cultured with target cells to study neuron–target interactions. Historically, sympathetic neuron and target cell co-cultures were used to study the influence of target on neurotransmitter phenotype. More recently, sympathetic neurons co-cultured with target tissue are being used to study the role of target-derived trophic factors on the formation and function of sympathetic synapses on heart; sympathetic control of energy homeostasis via modulation of adipose tissue function, and sympathetic regulation of immune responses via innervation of the spleen and lymph nodes.

Parasympathetic Neurons

Parasympathetic ganglia are considerably smaller than sympathetic ganglia. They range in size from as few as 2 or 3 cells to as many as 30 cells. Because of their location, usually buried within end organs, and their diffuse nature, they have received less attention as *in vitro* models than have sympathetic and sensory ganglia. Professor Geoffrey Burnstock first maintained parasympathetic nerves in cell culture in the early 1980s, initially by explant and later by dissociation of tissues. Parasympathetic nerves have now been isolated from ganglia in many different organs, including lungs, heart, bladder, and submandibular glands (enteric nerves will be considered separately).

In vitro studies on parasympathetic neurons indicate that factors affecting growth of sympathetic neurons do not automatically transfer to the parasympathetic cultures. For example NGF is required

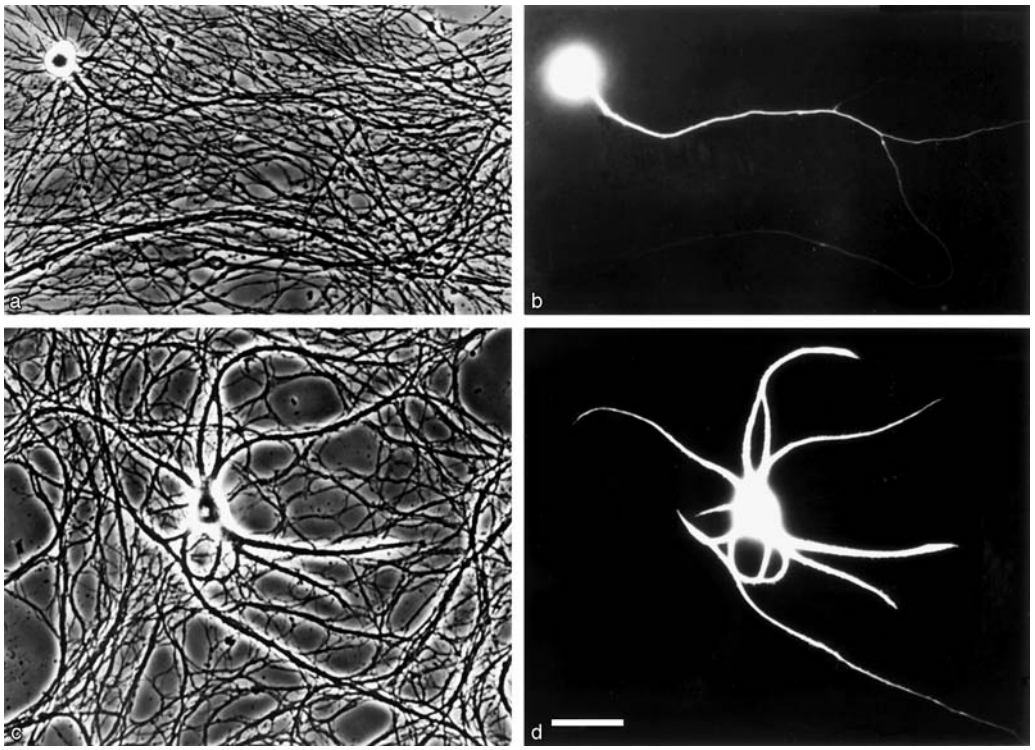

Figure 2 Sympathetic neurons do not extend dendrites in culture unless provided the appropriate environmental stimulus. Phase-contrast (a, c) and fluorescence (b, d) micrographs of rat sympathetic neurons injected with Lucifer yellow during the third week *in vitro*. Neurons grown in control cultures in the absence of nonneuronal cells and serum (a, b) typically extend only one process, which is axonal in nature. Neurons exposed to bone morphogenetic protein-7 (50 ng ml^{-1}) for 7 days were multipolar, having several tapered dendrites and one axon. Scale bar = 100 μm.

to maintain sympathetic neurons in culture. While NGF is commonly added to parasympathetic cultures, many labs do not use it, thus parasympathetic neurons do not require NGF to survive. Cell lines also do not mimic parasympathetic function. Neuroblastoma cells release acetylcholine, but regulation of receptor expression is mediated by different pathways in neuroblastoma and parasympathetic neurons. To identify the physiology and pharmacology of this branch of the autonomic nervous system, it is necessary to isolate and study the parasympathetic neurons. Since isolation of parasympathetic neurons is not yet routinely done, the technique will be described here.

Most studies with parasympathetic neurons isolate them from rats, rabbits, or guinea pigs. It is not necessary to isolate parasympathetic nerves from fetal animals. Although most tissues are derived from young or newborn animals, there appears to be little difference between cells isolated from young versus older animals. We have recently isolated and maintained parasympathetic neurons from adult human trachea.

Isolation of parasympathetic ganglia involves dissecting out the muscle containing the ganglia followed by mechanical removal of as many nonneuronal cells as possible. For example, airway epithelial cells are wiped away with sterile cotton swabs. The tissue is gently minced, and then digested enzymatically. This usually requires incubation with trypsin (0.25%) and collagenase (0.2%) or protease (0.2%) at 37 °C for 60–90 min. Thicker tissues can be incubated with the enzymes at 4 °C for 1–18 hours to allow enzyme penetration of tissues, followed by incubation at 37 °C to activate the enzymes. Enzymatic digestion is usually accompanied by gentle agitation or pipetting to disperse the cells. This procedure results in suspension of neurons, glial cells, smooth muscle cells, fibroblasts, and, depending upon the tissue of origin, endothelial or epithelial cells.

Cultures can be enriched for neurons by preplating the cell suspension on plastic dishes to remove fibroblasts and muscle cells. In addition, the antimitotic agent cytosine arabinoside (5 μM) can be added to cultures a day or two after plating to kill dividing cells. Over several days, the cultures consist predominantly of neurons. A major caveat of working with parasympathetic ganglia is that, unlike sympathetic ganglia, cultures derived from parasympathetic ganglia are heterogeneous. This is less an issue in electrophysiological, immunocytochemical, or transmitter release studies, but is a major factor when measuring protein or mRNA by Western blot or polymerase chain reaction

(PCR). In these latter experiments, controls must be added for the presence of nonneuronal cells.

Parasympathetic neurons grow best on extracellular matrix-coated slides or culture dishes. Some laboratories maintain them in Rose chambers (two coverslips, separated by a plastic washer, that are clamped together, which can be purchased from Bellco Biotechnology, product 1943–11111). Cells are usually grown in standard medium, such as medium 199, Ham's F12, or L-15, that is supplemented with 10% fetal calf serum, 0.5% glucose, and NGF (50 ng ml^{-1}), or in serum-free medium supplemented with transferrin (10 μg ml^{-1} avian transferrin or 20 μg ml^{-1} rat transferrin), sodium selenite (5 ng ml^{-1}), glutamine (20 μg ml^{-1}), insulin (5 μg ml^{-1}), NGF (20–50 ng ml^{-1}), and bovine serum albumin (1 mg ml^{-1}). Some laboratories additionally include penicillin and streptomycin in the culture medium.

Parasympathetic neurons originally appear as spherical cells that settle onto the substrate over the 48 h after plating. Neurites initially appear as small commas, and extend over the first 3 days. By 2–3 days after plating, neurons are clearly visible. Parasympathetic neurons grow in dense clusters with outwardly radiating and fasciculated neurites (Figure 3). The choice of plating matrix can influence aggregation of the cells: they form dense clusters if grown on polylysine (this substrate yields fewer cells overall), and smaller clusters (with increased cell survival) when grown on collagen, laminin, or Matrigel. Parasympathetic neurons can be maintained in culture for 7–10 days. It is rare that cultures survive in excess of 14 days, which is another difference from sympathetic cultures.

Parasympathetic neuronal cultures have been used to describe receptor distribution and neurotransmitter content (as determined primarily by immunocytochemistry), and, more recently, to characterize the molecular mechanisms underlying interactions between parasympathetic nerves and inflammatory cells. Acetylcholine release in response to electrical field stimulation or nicotine has also been measured. However, it cannot be assumed that all parasympathetic neurons behave the same way. For example the distribution of purinoceptors for adenosine 5′-triphosphate (ATP) is different in parasympathetic neurons originating from heart and lung versus those derived from submandibular or otic ganglia. Much is still unknown about the basic cell biology of parasympathetic nerves. No studies have been carried out that characterize the growth and differentiation of parasympathetic nerves; indeed, even the characterization of neurites as axons or dendrites is lacking in this neuronal cell type. Recent work demonstrating that parasympathetic nerves can be transfected (Figure 4) suggests the exciting possibilities of future studies of the cell and molecular mechanisms regulating the

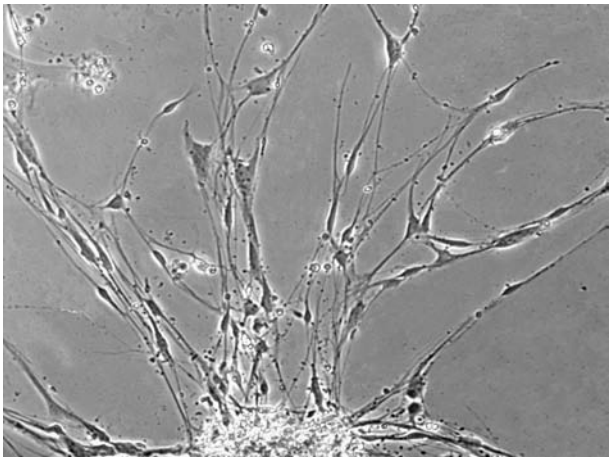

Figure 3 Phase-contrast image of parasympathetic neurons from guinea pig trachea, plated on Matrigel and maintained in serum-free medium for 7 days.

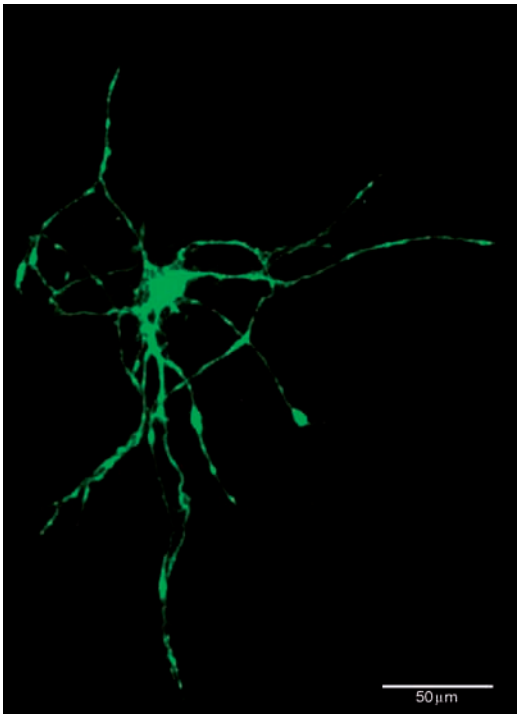

Figure 4 Primary cultures of parasympathetic neurons were plated on Matrigel and maintained in serum-free medium for 4–5 days. Cells were transfected with cDNA encoding enhanced green fluorescent protein (eGFP) using the transfection reagent Optifect (Invitrogen). An enhanced yellow fluorescent protein (eYFP) plasmid (3 μg) was incubated with Optifect (Invitrogen, 10 μl) in 200 μl of OptiMEM at room temperature for 5 min. Cells were incubated with 200 μl of eYFP/Optifect solution in 1 ml of culture medium at 37 °C in 5% CO_2 for 6 h, then rinsed with phosphate-buffered saline and fixed with 3.7% paraformaldehyde.

development, growth, and plasticity of parasympathetic neurons.

Cardiac Ganglia

Cardiac neurons appear in culture singly or as clusters, with neurites extending for hundreds of micrometers. When isolated from rats, the soma are 25–35 μm in diameter, depending upon the degree of flattening. Nuclei are approximately 5×10 μm. All cells from rat and guinea pig cardiac ganglia express acetylcholinesterase, suggesting that they are all cholinergic, and have functional M1 and M2 muscarinic receptors evenly distributed along cell bodies and neurites. However, they are actually a heterogeneous population, with three distinct electrophysiological phenotypes, and a subpopulation immunoreactive for dopamine β-hydroxylase or tyrosine hydroxylase, suggesting the presence of SIF cells. Half of all neurons in these cultures are immunoreactive for neuropeptide Y and a smaller percentage express vasoactive intestinal peptide, but tachykinin-expressing cells are not detected. Thus, cardiac parasympathetic ganglia in cell culture maintain a considerable degree of neuronal specialization, as is seen *in vivo*.

Tracheal Ganglia

Parasympathetic ganglia from trachea were initially isolated from young rats, but cultures from mature guinea pigs and adult human organ donors have been isolated and maintained in cell culture for several weeks (see reports from laboratories of Professors Burnstock and Fryer). Tracheal ganglia behave similarly to nerves *in vivo* as they release acetylcholine, express nitric oxide synthase, and, depending upon the presence of NGF, express substance P. M2 muscarinic receptors are evenly distributed along on cell bodies and neurites and function to inhibit acetylcholine release. Similar to nerves *in vivo*, inflammatory cells adhere to parasympathetic nerves in culture, and this effect is regulated by expression of chemotactic cytokines and adhesion molecules by the parasympathetic nerves. Parasympathetic nerves also release some noradrenaline and have functional α_1 adrenoceptors and γ-aminobutyric acid ($GABA_A$) receptors. As with cardiac nerves, cultured ganglia from the trachea appear to maintain all characteristics of their *in vivo* counterparts.

Chick Ciliary Ganglion Neurons

One parasympathetic ganglion that has been well studied in cell culture is the avian ciliary ganglion, because it is relatively large, discrete, and easily located outside of the eye. It should be noted that, in sharp contrast to the bird, rodent ciliary ganglia are small and dispersed, thus ciliary ganglia isolated from chick may not be an accurate model of the physiology of this ganglion in mammals. The chick ciliary ganglion contains two distinct neuronal cell populations. Ciliary neurons are large, myelinated cholinergic neurons that innervate striated muscle in the iris and ciliary muscle. It is the choroid neurons, small, unmyelinated cholinergic neurons that supply the arterial vasculature of the choroid layer, that are the parasympathetic component of these ganglia. Ciliary neurons require neurotrophic factors, such as ciliary neurotrophic factor, to support their survival in culture. Neurotransmitter expression in the choroid cells is controlled by target tissues, in this case the eye, and this interaction may not be representative of parasympathetic neurons as a class. Despite these drawbacks, the ciliary ganglion has been a popular system for neuronal cell culture, and a great deal is known about the growth requirements and behavior of this class of chick parasympathetic neurons.

Enteric Neurons

The enteric nervous system is considerably more complex than is other peripheral neural tissue. Neuronal cells bodies reside in the submucosal and myenteric ganglia, and consist of at least 14 different neuronal cell types, including interneurons, afferent neurons, and both excitatory and inhibitory motor neurons, that collectively express diverse neurotransmitters in addition to acetylcholine, including substance P, enkephalins, nitric oxide, ATP, and GABA. The enteric nervous system is highly autonomous: nerves form extensive synaptic contacts with one another and are capable of integrated reflex activity in the absence of central connections. Thus, in both its structure and its function, the enteric nervous system bears a greater resemblance to the central nervous system than to other peripheral neural tissues.

Isolation of enteric neurons is similar to that of parasympathetic neurons and involves enzymatic digestion of tissue containing the ganglia with subsequent plating on matrix gels or polylysine. Enteric neurons can survive in culture up to 3 weeks and the diversity of neuronal cell types found *in situ* is represented in culture. Individual enteric neurons are characterized by markedly diverse morphology and they retain the variety of neurotransmitters found *in situ*. Their electrophysiological properties appear to be similar to those of enteric nerves in the intact animal. Cultured enteric ganglionic glial cells

continue to synthesize intracellular glial proteins expressed by their *in situ* counterparts, including glial fibrillary acidic protein and glutamine synthetase, suggesting the feasibility of using enteric nerve cultures to study nerve–glial cell interactions.

A complication of interpreting any study involving cultured enteric neurons is the paucity of information regarding differential effects of environmental factors on different neuronal cell types within the heterogeneous population of neurons that comprise the enteric culture system. Few attempts have been made to discriminate between effects on axons versus dendrites within any individual neuron, although a number of studies have examined factors that promote neurite outgrowth. Glial cells are not required for neurite development in culture, but when they are present, enteric nerves will extend processes along them. Neurite length is enhanced by NGF, consistent with observations that enteric neurons express both the low-affinity NGF receptor (p75) as well as tyrosine kinase (Trk) receptors. Neurite outgrowth is also promoted by cyclic adenosine monophosphate (cAMP) via complementary but independent pathways, as evidenced by observations that inhibitors of cAMP have no effect on NGF-induced neurite growth, and that cAMP, but not NGF, enhances neurite branching in the myenteric nerves. Neurite length is also enhanced by ciliary neurotrophic factor, insulin-like growth factors, leukemia inhibitory factor, and interleukin-6 (IL-6). The greatest enhancement of neurite length, however, is provided by glial-derived neuronal factor (GDNF), which also induces larger clusters of cell bodies and increases neuronal survival in culture. IL-6 induces a greater number of neurites, but GDNF promotes increased neurite length.

While much research on enteric neurons in culture has demonstrated that these neurons retain the characteristics of cells found *in situ*, there is evidence that at least some neuronal cell types retain the plasticity that has been observed in other neural crest derivatives. Thus, exposure to heart-cell conditioned medium induces substantial, but not uniform, changes in both their electrophysiological and transmitter-related properties (increasing the percentage of neurons that express 5-hydroxytryptamine, while decreasing the percentage of neurons expressing vasoactive intestinal peptide and substance P).

Conclusions

Almost everything we know about the development and physiology of autonomic nerves has been derived from studies of sympathetic neurons, in large part because of the relative ease with which this neuronal cell type can be isolated and maintained in culture. Recent advances in methods for manipulating gene expression in cultured sympathetic neurons and for culturing sympathetic neurons in the presence of target tissues or ganglionic glia ensure the continued widespread use and popularity of this *in vitro* model system for discovery-based and mechanistic research. The recent development of techniques for the isolation and maintenance of parasympathetic and enteric neurons in cell culture will enable a more detailed understanding of the basic cell biology of these other peripheral neural systems, thereby refining and extending current paradigms of peripheral nervous system development, physiology, and pharmacology.

See also: Autoimmune Autonomic Neuropathy; Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Nervous System; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Allen TGJ and Burnstock G (1987) Intracellular studies of the electrophysiological properties of cultured intracardiac neurons of the guinea pig. *Journal of Physiology (London)* 388: 349–366.

Allen TGJ and Burnstock G (1990) A voltage-clamp study of the electrophysiological characteristics of the intramural neurones of the rat trachea. *Journal of Physiology (London)* 423: 593–614.

Campenot RB (1992) Compartmented culture analysis of nerve growth. In: Stevenson BR, Gallin WJ, and Paul DL (eds.) *Cell–Cell Interactions: A Practical Approach*, pp. 275–298. Oxford: IRL Press.

Chamley JH, Goller I, and Burnstock G (1973) Selective growth of sympathetic nerve fibers to explants of normally densely innervated autonomic effector organs in tissue culture. *Developmental Biology* 31: 362–379.

Fryer AD, Stein LH, Nie Z, et al. (2006) Neuronal eotaxin and the effects of CCR3 antagonist on airway hyperreactivity and M2 receptor dysfunction. *Journal of Clinical Investigation* 116: 228–236.

Higgins D, Lein PJ, Osterhout DJ, et al. (1991) Tissue culture of mammalian autonomic neurons. In: Banker G and Goslin K (eds.) *Culturing Nerve Cells,* 1st edn., pp. 177–205. Cambridge, MA: MIT Press.

James S, Bailey DJ, and Burnstock G (1990) Autoradiographic visualization of muscarinic receptors on rat paratracheal neurons in dissociated cell culture. *Brain Research* 513: 74–80.

Jessen KR, McConnell JD, Purves RD, et al. (1978) Tissue culture of mammalian enteric neurons. *Brain Research* 152: 573–579.

Jessen KR, Saffrey MJ, Bauk P, et al. (1983) The enteric nervous system in tissue culture. III: Studies on neuronal survival and the retention of biochemical and morphological differentiation. *Brain Research* 262: 17–35.

Lein PJ, Johnson M, Guo X, et al. (1995) Osteogenic protein-1 (OP-1) induces dendritic growth in rat sympathetic neurons. *Neuron* 15: 597–605.

Levi-Montalcini R (1987) The nerve growth factor 35 years later. *Science* 237: 1154–1162.

Lewis WH and Lewis MR (1912) The cultivation of sympathetic nerves from the intestine of chick embryos in saline solutions. *The Anatomical Record* 6: 7–31.

Smith CL (1998) Cultures from chick peripheral ganglia. In: Banker G and Goslin K (eds.) *Culturing Nerve Cells*, 2nd edn., pp. 261–287. Cambridge, MA: MIT Press.

Stemple DL and Anderson DJ (1992) Isolation of a stem cell for neurons and glia from the mammalian neural crest. *Cell* 71: 973–985.

Ziller C (1986) Development of neuronal properties in neural crest cells cultured *in vitro*. *Current Topics in Developmental Biology* 20: 177–193.

Cell Culture: Primary Neural Cells

R G Giffard and Y-B Ouyang, Stanford University School of Medicine, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neural tissue is composed of many different cell types: neurons, astroglia, oligodendroglia, microglia, endothelial cells, and other perivascular cells. Within these broad classes are many subsets, particularly of neurons and astrocytes. This complexity makes it difficult for neuroscientists to determine which cells are affected by a particular drug or manipulation and whether effects are direct or indirect actions on the cells of interest. The major advantages of using dissociated-cell culture to study the nervous system are (1) reduced cell complexity, (2) the ability to carry out biochemical analysis of individual identified cell types, (3) the ability to completely control the cellular environment, and (4) manipulation of gene expression in individual cell types. Cell cultures are advantageous for methods such as individual cell imaging and electrophysiology (Figure 1). The major disadvantages of primary neural cell culture are the loss of normal synaptic pathways, the loss of normal anatomic relations, and changes in gene expression. However, studies indicate that most neural cell types express similar properties *in vitro* and *in vivo*. For example, astroglia produce vimentin, their typical intermediate filament protein glial fibrillary acidic protein (GFAP; Figure 1, middle panel), and neurotrophic factors. Oligodendroglia and Schwann cells can make myelin and neurons in culture, form synapses, express appropriate neurotransmitter receptors, and conduct action potentials, as well as express neurofilaments and growth cones.

Primary Cultures versus Cell Lines

Primary cultures are prepared directly from brains isolated from animals. The tissue is dissociated by enzymatic and mechanical means into a suspension of single cells and plated onto a culture dish or other surface. Some types of cells, especially those isolated from embryonic brain, may undergo a period of proliferation. In the case of astroglia, once the individual cells approach confluence and contact one another, they stop dividing and begin to differentiate. Differentiation may be incomplete because of lack of normal cell partners or differentiation signals. Primary cultures have been used extensively to determine mitogenic and differentiation signals for individual cell types, as in studies of stem cells and the fates of their progeny cells. In this way, many factors that determine the differentiation fates of brain progenitor cells have been demonstrated. Once the cells are taken up from the dish and replated, they are termed secondary cultures, though some workers still use the term 'primary culture' to refer to cells isolated from tissue even after one or two passages. If normal cells are kept in culture and allowed to divide and expand in number, either they become senescent, if they are normal cells, or mutations occur that allow the cells to continue to divide and be passaged. At that point a cell line has been generated. Another method is to take a primary cell and transform or immortalize it with an oncogene, often of viral origin. A third method is to use cells derived from a tumor of the cell type of interest.

Cell lines are convenient in that they can be maintained indefinitely in culture and generate large numbers of cells, but they have significant drawbacks in that they are not normal cells, and after long-term passage, they accumulate genetic alterations and often, when given differentiation cues, do not fully replicate the patterns of gene expression characteristic of either primary cultures or brain cells *in vivo*. Despite this, some lines express a reasonably stable phenotype and have been used extensively for study.

One problem with mature neurons is that they are nondividing postmitotic cells and cannot be subcultured. Second, though there are cell lines that express some of the individual characteristics of differentiated neurons such as neurotransmitters, ion channels, receptors, and other neuron-specific proteins, they generally do not replicate the full phenotype of a specific neuronal subtype. PC12 cells derived from a pheochromocytoma, which have been studied extensively and are used to represent a specific population of nerve cells, only partially replicate the phenotype of the sympathetic neurons they otherwise resemble.

Glial cells can be subcultured, at least for several generations. However, the properties of the cultured cells often change gradually with passaging, as more rapidly dividing cell populations come to predominate and more 'differentiated' cells, which divide more slowly, are lost. Although many cell lines derived from glial tumors have been developed, they are used less commonly than are primary astrocyte cultures as astrocyte cultures are relatively easy to prepare and generally can be prepared in adequate quantity. In the case of one cell line said to resemble oligodendrocytes, the cells were not competent to form myelin.

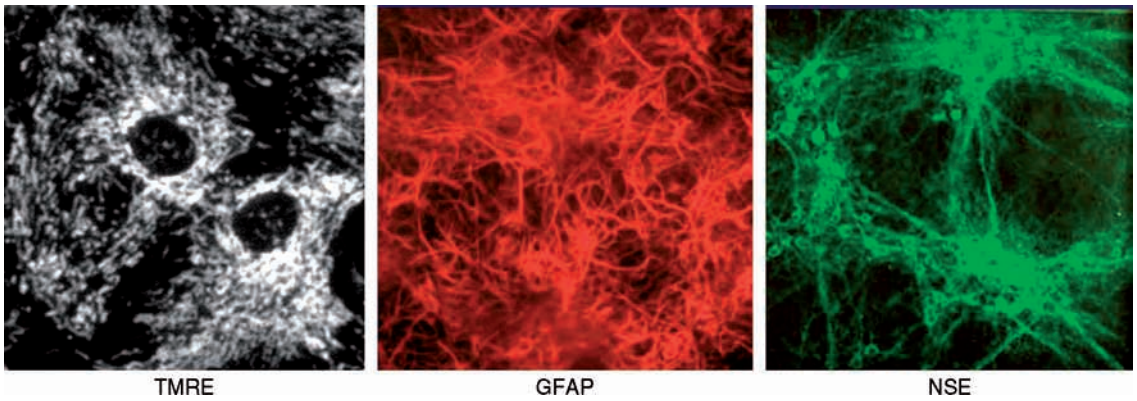

Figure 1 Left panel: Photomicrograph of a fluorescence image from tetramethylrhodamine ethylester (TMRE), a potentiometric fluorescent dye that incorporates into mitochondria in a mitochondrial membrane potential-dependent manner, demonstrating punctate perinuclear mitochondria in cultured cortical astrocytes. Middle panel: Glial fibrillary acidic protein (GFAP) immunostaining of cortical astrocytes. Right panel: Neuron-specific enolase (NSE) staining of cortical neurons.

Different Types of Primary Neural Cell Culture

Regardless of the type of cells to be grown, all methods for preparing primary cell cultures share several common features: Start with the brain region containing the cells of interest, dissect the region free of surrounding tissue, and dissociate the tissue. The single-cell suspension that is obtained is then placed on a suitable substrate and in an appropriate medium inside a temperature- and gas-controlled incubator.

The appropriate developmental age should be chosen to prepare specific populations of cells. For example, cortical neurons are generally prepared from rat or mouse embryos at day 14 (E14) or 15 (E15), hippocampal neurons from slightly older embryos (E18 to E19), and glia from rat or mouse pups 0 to 2 days after birth. Neurons are much less susceptible to damage during dissociation before they have developed extensive axonal and dendritic arbors and have become highly innervated. They may also be less dependent on their target cells for trophic support at early stages in their development. Glial cultures, which derive predominately from populations of progenitor cells, are prepared most easily when the number of these progenitors is near its peak, during the early postnatal period.

Neurons

There are many subtypes of neurons. When plated onto an appropriate substrate, primary neurons begin to extend processes within several hours, finally developing distinct axons and dendrites, forming synapses with one another, expressing the receptors and ion channels characteristic of the corresponding cell type *in situ*, and demonstrating spontaneous electrical activity, including producing action potentials.

Because neurons in primary cultures are essentially postmitotic and committed to differentiation at the time they are introduced into culture, the morphological and physiological properties of the cell population in culture correspond closely to the characteristics of the cell population present in the tissue of origin. Motor neurons express properties reminiscent of motor neurons, Purkinje cells of Purkinje cells, and so on.

One limitation of primary neuronal cultures is that they often contain a heterogeneous population of cells. Choice of tissue to culture is an important determinant of the cells in culture, and superimposed on the cell types present in the initial tissue is the ability of each cell type to survive under the conditions chosen for the culture. Choice of conditions can bias cell survival – some cell types do better in high serum, some in low, and some in fully defined culture medium. In some brain regions there is a predominant cell type, such as pyramidal neurons in hippocampus or granule cells in cerebellum. Alternatively, unwanted cells may be killed, either by specific cytotoxins or by complement-mediated lysis using antibodies against cell antigens. Another important strategy for isolating specific cell types from heterogeneous regions is called immunopanning. In this method, surfaces are coated with antibodies to surface determinants specific to an individual cell type, and the dissociated cells are then allowed to settle onto the coated surface, and nonadherent cells are removed. Often several rounds of panning are used to remove undesired cell types as well as to select desired cell types. This technique has been used successfully to isolate individual cell types from optic nerve, as well as from other starting material. The principal disadvantage of this approach is that it is difficult to isolate large numbers of cells this way, but the purity of the cell type is high.

Brief Protocol: Cortical Neurons

Primary cortical neuronal cultures are prepared from fetal mice at 15 days' gestation. Neocortexes are dissected free of meninges, treated with 0.25% trypsin for 20 min, mechanically dissociated and plated as a single-cell suspension in minimal essential media plus 5% fetal bovine serum (Hyclone) and 5% equine serum at 3.5 hemispheres per 10 ml medium. After 24 h *in vitro*, 60% of the plating medium is replaced with glial conditioned medium and B-27 serum-free supplement (2%; Gibco, Grand Island, NY) is added. Glial cell proliferation is then inhibited by addition of cytosine arabinoside (3 mmol l^{-1}). Cells are incubated in a humidified incubator at 37 °C with a 5% CO_2 atmosphere without further medium changes until use. These cultures contain <1% astrocytes at all durations in culture studied as determined by immunostaining for GFAP and neuronal nuclei and neuron-specific enolase.

Astrocytes

Astrocytes are the most numerous cell type in the brain. For decades, astrocytes were considered non-excitable support cells. This view has changed radically during the past 20 years. Astrocytes are now appreciated to perform multiple essential functions in normal brain: They are dynamically involved in synaptic transmission, metabolic and ionic homeostasis, the inflammatory response, antioxidant defense, trophic support of neurons, as well as the establishment and maintenance of the BBB. Astrocytes respond to signals from and signal actively to neurons, endothelium, and microglia. Because of the heterogeneity of nervous tissue, the small-size cell bodies, and the diffuse distribution of glial cells, it has been difficult to investigate the function of glia including astroglia *in situ*. In primary culture it is relatively easy to obtain pure preparations of astrocytes with greater than 95% homogeneity. The main contaminant is microglia, and additional steps can be taken to reduce the number of microglia to less than 1%. Some use dibuturyl cyclic adenosine monophosphate to induce morphological differentiation of astrocytes into more stellate shapes than are observed without this treatment. Within a week or two, immature astrocytes isolated from newborn brain will divide and form a confluent monolayer when plated at low density or, when plated at higher density, a layer of cells that may be a few cells thick in places. Conditions are chosen to preclude the survival of neurons.

While it is generally difficult to produce large amounts of cells from neuronal or microglial primary cultures for biochemical analysis, this is much less true of astrocyte cultures. In contrast, since astrocytes divide readily in culture, relatively large amounts of cells suitable for biochemical as well as morphological or physiological analysis can be produced.

Brief Protocol: Cortical Astrocytes

Primary astrocyte cultures are prepared from 1-to-2-day-old postnatal mice. Neocortexes are dissected free of meninges, treated with 0.25% trypsin for 20 min, mechanically dissociated and plated as a single-cell suspension onto either Primaria tissue culture plates or flasks (Becton Dickinson), poly-D-lysine coated tissue culture plastic containers, or glass coverslips at a density of two hemispheres per 10 ml. Plating medium consists of Eagle's minimal essential medium supplemented with 10% fetal bovine serum and 10% equine serum, 21 mmol l^{-1} (final concentration) glucose (Sigma Chemical, St. Louis, MO), and 10 ng ml^{-1} epidermal growth factor. Cultures are maintained at 37 °C in a 5% CO_2 incubator. Astrocyte cultures are postmitotic by days 7 to 14 in culture and generally contain <5% microglia and insignificant numbers of neurons or oligodendrocytes. The cultures increase expression of GFAP between 7 and 15 days *in vitro*, similar to what has been reported *in vivo*.

Microglia

Primary cell cultures are also used for the study of microglia, the population of macrophages resident in the central nervous system (CNS). Microglia are present in all brain areas but not uniformly distributed, being enriched in grey matter and with a relatively distinct perivascular population. The morphology of microglia *in vivo* varies according to their location and microenvironmental factors and ranges from compact amoeboid cells to ramified, branched cells, with considerable variation in the length and complexity of branching of the processes. Their plasticity is emphasized by their rapid morphological changes in response to a variety of stimuli, including mechanical or pathological insults. In culture, microglia can adopt three distinct morphologies: amoeboid, bipolar, and ramified, similar to what is seen *in vivo*. Microglia often appear as loosely attached cells atop monolayers of type I astrocytes, and can be harvested by shaking to produce relatively pure preparations.

Brief Protocol: Microglia

Cultures are prepared by shaking off from mixed glial cultures from mice at postnatal days 1 to 3. Whole cerebral hemispheres are removed and freed of meninges. Minced pooled tissue is treated with 0.25% trypsin, centrifuged, and dissociated into a

single-cell suspension by gentle trituration. Cells are plated at a density of two hemispheres per 10 ml in Eagle's minimal essential medium supplemented with 10% equine serum, 10% fetal bovine serum, 21 mmol l^{-1} (final concentration) glucose in flasks. Cells are incubated at 37 °C in a 5% CO_2 incubator. Medium is changed after 2 to 3 days *in vitro*. Microglia are harvested from mixed glial cultures between 7 and 14 days *in vitro*. After shaking at 160 rpm for 30 min on an orbital shaker, the medium is collected and centrifuged at 800 *g* for 10 min to collect the cells. Plates are coated with Collagen I (50 μg ml^{-1}; Sigma Chemical) for 10 min and washed; then microglia are plated at a density of 3×10^5 cells ml^{-1} in Eagle's minimal essential medium supplemented with 5% equine serum, 5% fetal bovine serum, and 2 mmol l^{-1} glutamine. After 1 h, wells are washed with medium to remove unattached astrocytes. Macrophage colony-stimulating factor (100 ng ml^{-1}; Sigma Chemical) is then added. This method yields cultures that contain >99% microglia as determined by histochemical staining with *Griffonia simplicifolia* B4-isolectin (IB-4; Sigma Chemical).

Oligodendrocytes

Oligodendrocytes, the myelinating cells of the CNS, develop from precursor cells known as O-2A progenitors. Although oligodendrocytes traditionally have been defined as myelinating cells, a number of recent reports suggest that these cells play additional roles. For example, oligodendrocytes in culture express a variety of trophic molecules.

Culturing oligodendrocytes is somewhat more difficult than culturing the other cell types discussed so far. In some cases, mixed glial cultures are the starting point. In other cases, white matter areas are isolated as starting material. When dissociated cells from cerebral hemispheres are cultured, they develop into a monolayer consisting mainly of astrocytes, with oligodendrocyte lineage cells on top of the monolayer. Similar to the case for microglia, cells enriched in oligodendrocyte lineage cells can be obtained from mixed glial cultures by shaking them off the monolayer and replating after centrifugation. Another method for obtaining purer cultures of oligodendrocytes or oligodendrocyte precursors is immunopanning. In this technique, dishes are coated with antibodies against a particular cell surface antigen, the cell suspension is added to the dish to allow selective binding of cells that express that antigen, and unbound cells are removed. Often several panning steps are done sequentially.

Brief Protocol 1: Forebrain Oligodendrocytes

Forebrains are dissected from newborn rats and cultured in nutrient medium-15 (NM-15) containing Minimum Essential Medium with Earle's salts and L-glutamine, heat-inactivated fetal calf serum (15%), glucose (6 mg ml^{-1}), and penicillin–streptomycin (0.5 U ml^{-1} and 0.5 μg ml^{-1}, respectively; Invitrogen, Rockville, MD) for 7 to 10 days. To eliminate microglia, glial cultures are shaken at 300 rpm for 10 min, and the supernatant is discarded. After recovery in NM-15, the oligodendrocyte type-2 progenitor cells and oligodendrocytes are separated from underlying astrocytes by shaking at 250 rpm for 12 to 16 h. Oligodendrocyte lineage cells are plated onto 35 mm poly-D-lysine-precoated dishes for immunocytochemical study or onto aminoalkysilane-precoated slides (Sigma, St. Louis, MO) for other studies in NM-15 medium at a density of 3×10^4 cells cm^{-2}. After 24 h in oligodendrocyte medium, NM-15 medium is aspirated and replaced with chemically defined medium consisting of a 1-to-1 mixture of Ham's F12 and Basal Medium Eagle containing 6 mg ml^{-1} glucose and 6.6 mmol l^{-1} glutamine (Invitrogen) and 25 μg ml^{-1} insulin, 20 nmol l^{-1} progesterone, 60 μmol l^{-1} putrescine, 30 nmol l^{-1} selenium, 100 μg ml^{-1} transferrin, 0.5 μmol l^{-1} thyroxin, 0.08 μg ml^{-1} triiodothyronine, 6.6 mmol l^{-1} glutamine, and penicillin–streptomycin (0.5 U ml^{-1} and 0.5 μg ml^{-1}, respectively; Sigma). Cultures are maintained for 96 h in oligodendrocyte medium, after which >90% of the cells in culture are oligodendrocyte lineage cells, either A2B5-positive or myelin basic protein-positive by immunocytochemical staining. The remaining nonoligodendrocyte cells are OX-42-positive or ED-1-positive microglia or GFAP-positive astrocytes.

Brief Protocol 2: Immunopanning

Cells are plated on 100 mm petri dishes that were sequentially treated with biotinylated antimouse immunoglobulin G (1:100, 12 h, 4 °C; Vector Laboratories, Burlingame, CA), A2B5 antibody with 5% bovine serum albumin (BSA) (12 h, 4 °C), and 5% BSA (4 h, room temperature; Sigma). Cells are incubated in these dishes for 30 min at 37 °C. After three culture washes in NM-15, bound cells are scraped, centrifuged (at 500 rpm for 5 min), resuspended in NM-15, and plated.

Brain Endothelium

Endothelial cells that line the microvasculature of the CNS differ fundamentally from other vascular

endothelia in their capacity to regulate the passage of molecules and cells to and from the neural parenchyma and constitute the BBB. This selectivity resides in specialized features unique to CNS endothelia, including expression of tight intercellular junctions that markedly limit paracellular permeability, as well as a unique pattern of receptors, transporters, and nonselective drug export pumps that protect the CNS from a large variety of potentially harmful hydrophobic compounds.

Although much has been learned about the unique properties of brain endothelium from *in vivo* studies, the limitations imposed by whole-animal experiments have led to the development of techniques to isolate and culture CNS-derived endothelial cells from various species. Unfortunately, when cultured alone these endothelial cells usually fail to develop functional tight junctions. To address this limitation, coculture systems have been developed in which brain endothelial cells, especially of bovine origin because of the availability of large amounts of material, are grown on microporous filter inserts in the presence of primary culture astrocytes in order to mimic the anatomical and functional relationship between brain endothelium and surrounding astrocyte foot processes. These coculture systems exhibit a much better correlation to *in vivo* permeability of standard molecules and high levels of electrical resistance.

Brief Protocol: Brain Endothelium

Primary cerebral microvascular endothelial cell cultures are prepared from young mouse brains. Gray matter from 2- to 3-week-old mice is homogenized in calcium and magnesium free Hank's balanced salt solution containing 0.35% collagenase and 0.2% BSA for 2.5 h at 37 °C. Tissue is resuspended in 25% BSA, then centrifuged at 4 °C for 20 min at 528 *g*. The pellet containing brain capillaries is incubated in 0.1% collagenase/dispase in bicarbonate containing M199 media (Gibco-BRL 11150-059) for 2 h at 37 °C to remove pericytes. The cells are pelleted, resuspended, then centrifuged on a Percoll gradient (11 ml Percoll in 21 ml Earle's balanced salt solution) at 733 ***g*** for 10 min at 4 °C. The capillary microvessel endothelial cell layer is collected, washed in M199 media, and plated on collagen-coated plates at 10^6 cells ml^{-1} in E-STIM endothelial cell culture media (Becton, Dickinson, Franklin Lakes, NJ) supplemented with 10 ng ml^{-1} epidermal growth factor plus penicillin–streptomycin (100 U ml^{-1}). Cultures are fed every 3 days.

See also: Astrocyte: Identification Methods; Astrocyte: Response to Injury; Cell Culture: Autonomic and Enteric Neurons; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Cells: Microglia During Normal Brain Aging; Microglia identification Methods; Microglia Properties; Oligodendrocyte and Schwann Cell Identification Methods; Oligodendrocyte Morphology.

Further Reading

Banker G and Goslin K (eds.) (1998) *Culturing Nerve Cells,* 2nd edn. Cambridge, MA: MIT Press.

Cohen W (ed.) (1996) *Neural Cell Culture: A Practical Approach.* Oxford, UK: Oxford University Press.

Dichter MA (1978) Rat cortical neurons in cell culture: Culture methods, cell morphology, electrophysiology, and synapse formation. *Brain Research* 149(2): 279–293.

Federoff S and Richardson A (eds.) (2001) *Protocols for Neural Cell Culture,* 3rd edn. Totowa, NJ: Humana Press.

Freshney RI (ed.) (2000) *Culture of Animal Cells: A Manual of Basic Techniques,* 4th edn. New York: Wiley.

Hayes LW (ed.) (1999) *The Neuron in Tissue Culture.* New York: Wiley.

Shahar A, de Vellis J, Vernadakis A, and Haken B (eds.) (1989) *A Dissection and Tissue Culture Manual of the Nervous System.* New York: Liss.

Cell Replacement Therapy for Huntington's Disease

O Isacson, McLean Hospital/Harvard Medical School, Belmont, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Why Neuronal Regeneration Therapeutic Strategies for Huntington's Disease?

Although new genetic information, including identification of specific markers and development of genetic models, is available for several neurological diseases characterized by regional neuronal loss including Huntington's disease (HD), these genetic studies have not yet resulted in advances in treatments for patients, except better diagnosis. Even if new pharmacological methods are developed to retard the genetic component of cell death, it is likely that cell regeneration therapies would also be required when HD brain cell death is delayed (even with more than 20 years delay since the average age of onset is 44 years) or is already apparent as the disease presents with symptoms. Moreover, since cell transfer or genesis may be possible prior to neuronal dysfunction, selectively vulnerable neurons and glia can be replaced prior to breakthrough symptoms. Consequently, there is a need for research that defines the potential and basic cellular mechanisms for neuronal replacement. In HD, while there is a genetically predetermined process causing neuronal death of neurons within the patient's caudate-putamen, implanted neural tissue (neurons and glia) lacking the mutated gene may substitute the disease-prone neurons and reconstitute functional neuronal connections. Such transfer of neurons and glia may also contribute to a slowing down of the neurotoxic processes evident in HD striatum since neurotrophic factors are produced by progenitor cell or fetal neural grafts. Furthermore, the demonstrated reconnection of host afferents with implanted neurons, and presence of normal glia, may offer resistance to cell loss caused by the progressive pathology of HD. An effective cell regeneration therapy for HD will likely involve anatomical precision and feedback control of local neurotransmitter release. In the future, neural replacement could perhaps be combined with other factors that promote regeneration and recovery of function.

Cell Transfer and Stem Cell Approaches for HD

Experimental progenitor cell transplantation HD paradigms have revealed data about structural brain repair. Several studies have confirmed and extended the initial discovery in rodent models of functional effects of striatal neuronal replacement in a primate model of HD. This underscores the validity of the approach as an example of the innovative nonpharmacological therapeutic research to broaden the horizon for HD patients. Data indicate that regenerating striatal neurons can replace neurons lost in the primate striatum to form novel axonal connections and synapses that can improve function. Fetal brain cells are an unlikely cell source for large-scale clinical and application unless derived via stem cells; however, they are *a priori* experimentally useful cells to test hypotheses and mechanisms to replace damaged circuitry, because fetal neural cells inherently possess cellular feedback systems for neurotransmitter, neurotrophic, and synaptic interactions. In future clinical cellular regeneration trials, donor cells may be derived from stem cells, genetically engineered cells, or redifferentiated tumor cell lines. Success with such cell sources is more reasonable and likely after mechanistic and detailed investigations can demonstrate safe and effective functional repair of structural brain damage using phenotypically normal striatal progenitor brain cells, serving as an exploratory standard of cellular regeneration therapy for neurodegenerative conditions.

Immunology and Inflammation Issues Specific to Neural Intracerebral Cell Transfer

The presumption that the brain is an entirely immunologically privileged transplantation site has been challenged. This concept was originally based on studies of discordant graft survival in such neural sites as the cornea, the anterior and posterior chambers of the eye, and the brain as compared to other implantation sites in the body. The relatively limited host immune response within the brain was thought to depend on a combination of factors, including the near absence of lymphatic vessels within the brain and the existence of the blood–brain barrier (BBB) that prevented immediate T cell and humoral immune access to graft cells within the brain. The transplantation of neural tissue itself (except for its blood vessels) into brain cells was even less immunogenic than other transplanted tissues, given the very low levels of MHCI expressed on neural tissue (especially neurons). However, experience with xenotransplantation to the rat brain demonstrated that graft survival is low (but not zero) without systemic immune suppression. Thus, the brain location has

immune surveillance. The same principles apply for immune suppression, since the overall frequency of xenograft rejection in the rodent brain can be greatly reduced with the aid of cyclosporine A. This immune suppression (as for FK506) is probably primarily through blocking the proliferation of T-cytotoxic cells (CD8+) by inhibiting interleukin-2 (IL-2) production from T-helper cells (CD4+), thereby markedly reducing cell-mediated cytotoxic rejections of xenotransplants mediated by T cell recognition of foreign MHC. These successful cell transfers to brain between species have been demonstrated in studies using several fetal donor tissue species. Research results suggested that the major mechanisms involved in xenograft rejection in the adult brain were T cell mediated (and MHCI dependent) and were therefore controllable using standard immune suppression procedures employed in allografting. Numerous studies of neural grafting within species, including the rat brain and human brain, had demonstrated that fetal donor allografts survive in the absence of immune suppression. However, the rodent host brain does not present a full model of primate immunology since hyperacute rejection (complement cascade leading to cell lysis after attachment of preformed antibodies found in blood to xenoantigens on foreign tissue) is rarely seen. Such species-specific humoral factors can participate in early and sometimes also in graft rejection processes. After the BBB is sealed following surgery, sensitization of the host to the xenograft by peripheral skin grafts causes high serum titers of antibodies recognizing the foreign donor cells, but this postengraftment sensitization does not usually produce significant graft rejection. In contrast, presensitization in rodent models, created by first transplanting a skin graft peripherally and then compromising the BBB by implanting fetal neural cells to the brain (in allo- or xenograft combinations) will lead to significant, although partial, rejection of the grafts. These results using presensitization in rodents are probably analogous to the normal status of natural inherited complement-mediated hyperacute rejection seen in primates and humans, thus making the human brain less hospitable to xenotransplantation of new cells for repair and regeneration than several other species.

Deriving and Examining Stem Cells for Functional Experimental Repair in HD Transgenic Models

Designing or selecting cells for neurodegenerative therapies is a new discipline. Genetic engineering of embryonic stem (ES) cells is possible, since ES cells are continuously dividing in their native uncommitted state and can therefore be made to adopt or modify genes under selection that become stable genomic components. Such work signifies an opportunity to design cells by insertion, deletion, or tagging of specific transcription factors or proteins, not only for Parkinson's disease (PD), but possibly also for HD, amyotrophic lateral sclerosis (ALS), ischemia, and general spinal cord damage. Stem cell biology and cell design may eventually get on an equal therapeutic footing with drug-related design processes. This would be of value for regenerative medicine by providing neurons and cells to treat patients with diseases that are currently inadequately served by pharmacological medicine. For HD, the cells most vulnerable to the disease process and consequently most needed for cell replacement or regeneration originate during embryological development in the striatum (caudate and putamen) region (Figure 1). At that time the lateral ganglionic eminence (LGE) of the ventricular wall grows to adulthood by generating a population of γ-aminobutyric acid (GABA) ergic projection neurons (~95% of striatal neurons), which eventually connect with distant synaptic targets in the pallidum and substantia nigra. A smaller (~5%) interneuron population is also present in the striatum (Figure 1). The powerful influence of the GABAergic projection neurons on motor initiation is revealed by their dysfunction and degeneration in HD, where during some stages of the disease involuntary movements are particularly dramatic. The neural mechanisms responsible for such involuntary movements involve overactive thalamocortical circuitry that has lost gating motor control provided by the striatal neurons linked up in these motor loops. Striatal projection neurons are defined during development by transcription factor activations that specifically include Dlx-1 and Dlx-2 as well as Pax6 and Gsh2 (Figure 1).

Several interneuronal types are interspersed in the local circuitry, but these neurons are relatively spared in this disease process. While attempts to prevent the effects of the disease-provoking HD gene have been underway for almost a decade, currently there are no treatments or a cure. Given that HD, like most neurodegenerative diseases, even if slowed by years, will eventually cause major loss of select neurons and their synapses, it is reasonable to also develop synapse repair and replacement strategies. As for PD, there have been some basic research and initial proof of principle in HD patients, showing that appropriate cell and neuronal replacement is possible. Nonetheless, the technical and conceptual hurdles in applying these scientific insights to clinical trials have proved

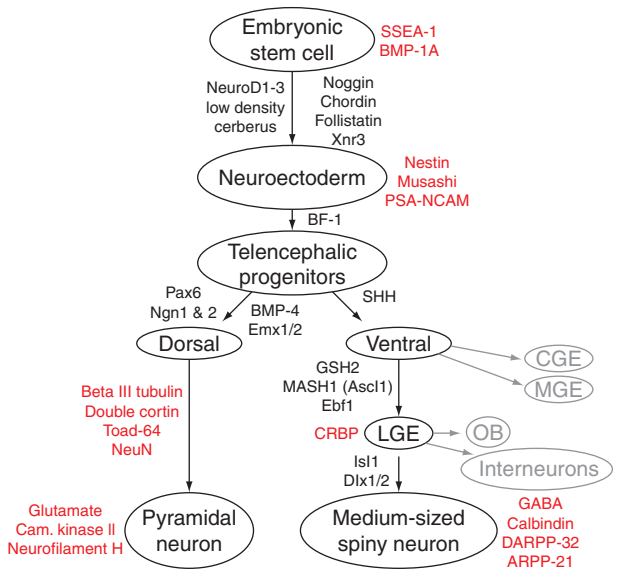

Figure 1 Diagram illustrating the specification of striatal medium-sized spiny neurons (MSNs) and cortical pyramidal neurons by induction factors (black) through phenotypic markers (red) from an embryonic stem cell. The MSNs are GABAergic projection neurons that comprise the predominant cell type in the striatum (>90%). In addition, the MSNs are the susceptible cell type that degenerates in HD. The glutamatergic pyramidal neurons project to the MSNs in the striatum and may contribute to MSN degeneration in HD. SSEA-1, stage-specific embryonic antigen 1; BMP-1A, bone morphogenic protein 1A; Xnr3, *Xenopus*-related nodal 3; BF-1, brain factor 1; PSA-NCAM, polysialylated neural cell adhesion molecule; Ngn1&2, neurogenin 1&2; BMP-4, bone morphogenic protein 4; SHH, sonic hedgehog; NeuN, neuronal nuclei; Cam. kinase II, calmodulin-dependent protein kinase II; MASH1(Ascl1), aschaete-scute complex-like 1; Ebf1(Olf1), early B cell factor 1; CRBP, cellular retinol binding protein; Isl1, islet 1; Dlx1/2, Distalless related 1/2; GABA, gamma aminobutyric acid; DARPP-32, dopamine and cyclic AMP-regulated phospho-protein, Mr = 32 000; ARPP-21, cyclic AMP-regulated phospho-protein, Mr = 21 000; LGE, lateral ganglionic eminence; CGE, caudal ganglionic eminence; MGE, medial ganglionic eminence; OB, olfactory bulb. Adapted from Isacson O (2003) The production and use of cells as therapeutic agents in neurodegenerative diseases. *Lancet Neurology* 2: 417–424.

to be greater than expected, and the absence of reliable and reproducible donor cell sources has retarded the progress of the new cell therapies. The generation of a reliable stem cell generated striatal projection neurons in a cell culture system, or by induction of transplanted genetically engineered ES or progenitor neurons *in situ* by temporally controlled activation of transcription factors controlling cell fate, are scientific avenues providing data and insights for potential future therapeutic neuronal and synaptic repair for HD.

Clinical Exploratory Trials of Transplanting Fetal Striatal Cell and Neurons to Patients with HD

A few pilot clinical trials have been conducted using striatal fetal neurons of the human LGE (USA and France). One patient from a trial died 18 months after transplantation from cardiovascular disease, and postmortem histological analysis demonstrated surviving transplanted cells with typical morphology of cells derived from the developing striatum. Selective markers of both striatal projection and interneurons showed positive transplant regions clearly innervated by host TH fibers. There was no histological evidence of immune rejection. Notably, neuronal protein aggregates of mutated huntingtin, which is typical HD neuropathology, were not found within the transplanted fetal tissue. These data support the idea that implanted neural cells lacking the mutant HD gene may be able to replace damaged host neurons and reconstitute damaged neuronal connections even in patients in whom there is a genetically predetermined process causing neuronal death within the HD striatum. This was the first demonstration that grafts derived from human progenitor striatal tissue can survive, develop, and are unaffected by the disease process, at least for 18 months, after transplantation into a patient with HD. A recent clinical study of five patients demonstrated some improvements in 60% of the patients. Most plateaued 2 years after cell implantation. Progression of motor disability continued despite the implanted cells, and by 4–6 years after the surgery, the clinical improvements were no longer apparent. Although initial exploratory neurotransplantation of fetal cells has advanced to the clinic, the importance of independent research in HD animal models is more apparent than ever (note experience from PD). It is vital to fine-tune knowledge of cell selection, immune responses, graft survival, donor–host cell interactions, synapse formation, and behavioral consequences in order to safely and successfully apply these new procedures.

See also: Axonal Transport and Huntington's Disease; Cell Replacement Therapy: Parkinson's Disease; Cell Replacement Therapy: Mechanisms of Functional Recovery; Neurogenesis in the Intact Adult Brain; Neuronal Plasticity after Cortical Damage.

Further Reading

Anderson SA, Qiu M, Bulfone A, et al. (1997) Mutations of the homeobox genes Dlx-1 and Dlx-2 disrupt the striatal subventricular zone and differentiation of late born striatal neurons. *Neuron* 19: 27–37.

Bachoud-Levi A-C, Gaura V, Brugieres P, et al. (2006) Effect of fetal neural transplants in patients with Huntington's disease 6 years after surgery: A long-term follow-up study. *Lancet Neurology* 5: 303–309.

Bachoud-Levi AC, Remy P, Nguyen JP, et al. (2000) Motor and cognitive improvements in patients with Huntington's disease after neural transplantation. *Lancet* 356: 1975–1979.

Bjorklund LM, Sanchez-Pernaute R, Chung S, et al. (2002) Embryonic stem cells develop into functional dopaminergic neurons after transplantation in a Parkinson rat model. *Proceedings of the National Academy of Sciences of the United States of America* 99: 2344–2349.

Cattaneo E, Rigamonti D, Goffredo D, Zuccato C, Squitieri F, and Sipione S (2001) Loss of normal huntingtin function: New developments in Huntington's disease research. *Trends in Neuroscience* 24: 182–188.

Corbin JG, Gaiano N, Machold RP, Langston A, and Fishell G (2000) The Gsh2 homeodomain gene controls multiple aspects of telencephalic development. *Development* 127: 5007–5020.

Deacon T, Dinsmore J, Costantini LC, Ratliff J, and Isacson O (1998) Blastula-derived stem cells can differentiate into dopaminergic and serotonergic neurons after transplantation. *Experimental Neurology* 149: 28–41.

Dunnett SB, Carter RJ, Watts C, et al. (1998) Striatal transplantation in a transgenic mouse model of Huntington's disease. *Experimental Neurology* 154: 31–40.

Emerich DF, Winn SR, Hantraye PM, et al. (1997) Protective effect of encapsulated cells producing neurotrophic factor CNTF in a monkey model of Huntington's disease. *Nature* 386: 395–399.

Freeman TB, Cicchetti F, Hauser RA, et al. (2000) Transplanted fetal striatum in Huntington's disease: Phenotypic development and lack of pathology. *Proceedings of the National Academy of Sciences of the United States of America* 97: 13877–13882.

Freeman TB, Sanberg PR, and Isacson O (1995) Development of the human striatum: Implications for fetal striatal transplantation. *Cell Transplantation* 4: 539–545.

Frim DM, Uhler TA, Galpern WR, Beal MF, Breakefield XO, and Isacson O (1994) Implanted fibroblasts genetically engineered to produce brain-derived neurotrophic factor prevent 1-methyl-4-phenylpyridinium toxicity to dopaminergic neurons in rat. *Proceedings of the National Academy of Sciences of the United States of America* 91: 5104–5108.

Hantraye P, Riche D, Maziere M, and Isacson O (1992) Intrastriatal transplantation of cross-species fetal striatal cells reduces abnormal movements in a primate model of Huntington disease. *Proceedings of the National Academy of Sciences of the United States of America* 89: 4187–4191.

Isacson O (2003) The production and use of cells as therapeutic agents in neurodegenerative diseases. *Lancet Neurology* 2: 417–424.

Isacson O and Deacon T (1997) Neural transplantation studies reveal the brain's capacity for continuous reconstruction. *Trends in Neurosciences* 20: 477–482.

Isacson O, Deacon TW, Pakzaban P, Galpern WR, Dinsmore J, and Burns LH (1995) Transplanted xenogeneic neural cells in neurodegenerative disease models exhibit remarkable axonal target specificity and distinct growth patterns of glial and axonal fibres. *Nature Medicine* 1: 1189–1194.

Kendall AL, Rayment FD, Torres EM, Baker HF, Ridley RM, and Dunnett SB (1998) Functional integration of striatal allografts in a primate model of Huntington's disease. *Nature Medicine* 4: 727–729.

Kim KS, Kim CH, Hwang DY, et al. (2003) The orphan nuclear receptor Nurr1 directly transactivates the promotor of the tyrosine hydroxylase gene, but not the dopamine b-hydroxylase gene, in a cell-specific manner. *Journal of Neurochemistry* 85: 622–634.

Marder K, Sandler S, Lechich A, Klager J, and Albert SM (2002) Relationship between CAG repeat length and late-stage outcomes in Huntington's disease. *Neurology* 59: 1622–1624.

Munoz-Sanjuan I and Brivanlou AH (2002) Neural induction, the default model and embryonic stem cells. *Nature Reviews in Neuroscience* 3: 271–280.

Nery S, Fishell G, and Corbin JG (2002) The caudal ganglionic eminence is a source of distinct cortical and subcortical cell populations. *Nature Neuroscience* 5: 1279–1287.

Pakzaban P, Deacon T, Burns L, and Isacson O (1993) Increased proportion of AChE-rich zones and improved morphologic integration in host striatum of fetal grafts derived from the lateral but not the medial ganglionic eminence. *Experimental Brain Research* 97: 13–22.

Pearlman SH, Levivier M, Collier TJ, Sladek JR Jr., and Gash DM (1991) Striatal implants protect the host striatum against quinolinic acid toxicity. *Experimental Brain Research* 84: 303–310.

Schumacher JM, Short MP, Hyman BT, Breakefield XO, and Isacson O (1991) Intracerebral implantation of nerve growth factor-producing fibroblasts protects striatum against neurotoxic levels of excitatory amino acids. *Neuroscience* 45: 561–570.

St. Hillaire M, Shannon K, Schumacher J, et al. (1998) Transplantation of fetal porcine striatal cells in Huntington's disease: Preliminary safety and efficacy results. *Neurology* 50: A80.

Sussel L, Marin O, Kimura S, and Rubenstein JL (1999) Loss of Nkx2.1 homeobox gene function results in a ventral to dorsal molecular respecification within the basal telencephalon: Evidence for a transformation of the pallidum into the striatum. *Development* 126: 3359–3370.

Toresson H, Potter SS, and Campbell K (2000) Genetic control of dorsal–ventral identity in the telencephalon: Opposing roles for Pax6 and Gsh2. *Development* 127: 4361–4371.

Tulipan N, Luo SQ, Allen GS, and Whetsell WO (1988) Striatal grafts provide sustained protection from kainic and quinolinic acid-induced damage. *Experimental Neurology* 102: 322–325.

Woods WA, Papas TS, Hirumi H, and Chirigos MA (1973) Antigenic and biochemical characterization of the C-type particle of the stable porcine kidney cell line PK-15. *Journal of Virology* 12: 1184–1186.

Relevant Website

http://www.neuroregeneration.org – Center for Neurogeneration Research.

Cell Replacement Therapy: Mechanisms of Functional Recovery

S B Dunnett, Cardiff University, Cardiff, UK

© 2009 Elsevier Ltd. All rights reserved.

Functional Recovery

The first clear demonstrations of functional recovery after transplanting embryonic nerve cells in the brain were published in 1979. Two parallel studies showed recovery of motor asymmetries in hemiparkinsonian rats following implantation of embryonic dopamine cells to replace lost dopamine cells destroyed by neurotoxic lesions.

Since then, functional recovery has been observed in many different model systems: in sensory, motor, cognitive, motivational, and neuroendocrine functions; in rats, mice, hamsters, cats, dogs, monkeys, and humans; implanted in diverse sites in the brain, spinal cord, and peripheral nervous systems; following lesions made by traumatic, ischemic, excitotoxic, mechanical, or genetic lesions, and in aged animals; and involving transplantation of primary embryonic neurons, glia, engineered cells, stem cells, and other artificial matrix and cell-based systems. As a phenomenon, functional recovery takes place following transplantation of appropriate cells, and a considerable amount has been learned in the last 25 years about the techniques and protocols to achieve favorable outcomes.

Over the past 15 years, similar procedures have been applied to patients with neurodegenerative disease, first in Parkinson's disease (PD), and subsequently in Huntington's disease (HD), Alzheimer's disease (AD), stroke, motor neuron disease (amyotrophic lateral sclerosis, ALS), spinal cord injury, retinal degeneration, and multiple sclerosis, although not always with the degree of therapeutic benefit as had been hoped for based on the preceding experimental and preclinical work. Thus, increasingly, understanding the mechanisms of graft function is seen as an imperative not just for better neurobiological understanding of transplantation biology but in order to develop more effective cell-based therapies for patients based on rational design rather than empiricism alone.

Mechanisms of Functional Recovery

The first studies of nigral grafts in hemiparkinsonian rats observed recovery of sensorimotor deficits and motor rotation asymmetries, the magnitude of which correlated with the extent of dopamine replacement measured biochemically, and the extent of dopamine fiber outgrowth from the implanted dopamine cells measured histologically. Moreover, fiber outgrowth, reformation of synaptic connections, dopamine replacement, and functional recovery, all developed over a similar time course. It was therefore natural to assume that the recovery was attributable to the implanted cells providing a specific dopamine replacement and reinnervation in the host brain. However, subsequent experience has demonstrated that although this may be a factor, a variety of alternative more or less specific mechanisms can be hypothesized (see Table 1), and in fact there is evidence for each of the alternatives operating in different circumstances.

Nonspecific Mechanisms

Neurosurgical lesion of the caudate nucleus was proposed as a potential treatment to controlling selected symptoms of PD in the 1950s. More recently, we have seen the widespread development of thalamic, pallidal, and subthalamic lesions and stimulation for similar purposes, with growing evidence that 'deep brain' subthalamic stimulation can be particularly effective in alleviating disease symptoms. It has therefore been proposed that grafts might exert their effects not by a specific influence of the implanted cells but as a result of collateral damage caused by the neurosurgical operation.

Although some nonspecific contribution cannot be ruled out, it seems unlikely to provide a general explanation for the recovery observed either in patients or in animal models. Thus, in PD, recovery is slowly progressive with growth and integration of the grafts, rather than acute as would be expected of a nonspecific effect; recovery is dependent upon specific cell types being used to replace the particular cells lost to the disease (e.g., nigral dopamine cells for PD) and is not observed with implantation of other unrelated cell types; if the grafts do not survive, any recovery is marginal; and removal of the graft either by rejection or explicit surgical removal reinstates the original deficits. Perhaps most dramatically, in another model system, when circadian rhythms are reinstated following transplantation of embryonic basal hypothalamic cells in hamsters with focal suprachiasmatic nucleus lesions, transplanting between normal and mutant hamsters with shortened circadian period, the period of the restored rhythm follows that of the donor, not that of the host animal prior to lesion. In other situations, recovery that was at first thought to be attributable to nonspecific effects of the surgery has turned out to be attributable to specific but

Table 1 Alternative mechanisms of functional recovery

Mechanism	*Description*
Nonspecific	Surgical damage recruits intrinsic processes of compensation or redundancy
Pharmacological	The graft secretes neuroactive molecules to restore disturbed neurochemical function
Trophic	The graft secretes neurotrophic molecules to promote intrinsic plastic potential of host neurons
Glial support	The graft replaces lost glia, e.g., to promote remyelination or toxic buffering
Pathway repair	The graft provides substrates for axon growth to allow circuit regeneration in the adult CNS
Neuronal reinnervation	Grafted neurons reinnervate host brain to restore lost synaptic innervation
Circuit reconstruction	Grafted neurons establish afferent and efferent connections with host neurons so as to integrate functionally into host neuronal circuitry
Full network repair	The grafts provide full structure repair of neuronal circuits of the host brain

unanticipated mechanisms. Thus, in early clinical trials of adrenal medulla, the modest recovery observed in patients in whom the chromaffin cells themselves were seen not to survive turned out to be attributable to trophic influences of the grafts on the host brain (see below).

Pharmacological Mechanisms

A more plausible mechanism for recovery observed in a variety of model systems is pharmacological replacement of essential neurochemical signals lost to the disease process. Dopamine replacement by L-3,4-dihydroxyphenylalanine (L-dopa) can alleviate many symptoms of PD; cholinesterase inhibitors promote cholinergic transmission in the cortex and can alleviate some of the memory and cognitive symptoms associated with aging and dementia; hormone replacement can alleviate hypothalamic neurohormonal deficiencies such as diabetes mellitus or sexual dysfunction. Consequently, it is plausible that the grafts that are seen to alleviate symptoms in animal model systems of each disease are having their effect by replacement of deficient dopamine, acetylcholine, vasopressin, or gonadotrophin, respectively. The mechanism by which the grafted cells exert their functional effects could be considered to be, essentially, pharmacological, albeit utilizing a delivery system that allows stable, long-term delivery at physiological concentrations into precisely defined targets. Support for this position comes from the effectiveness of grafts that are unequivocally secretory, such as following encapsulation of neurons within semipermeable polymer capsules that allow diffuse release across the membrane but not cellular contacts between graft and host.

When the implanted neurons are seen to reinnervate the brain in each case, which in most cases they do, it can be experimentally difficult to resolve the extent to which the actual mechanism of recovery is due to pharmacological replacement, such that the reinnervation is essentially an epiphenomenon. In fact, there are several reasons to believe that for many grafts reinnervation is important. Nigral grafts that do reinnervate the dopamine-depleted host striatum have a different time course and more complete profile of functional recovery than explicit dopamine replacement into the same sites, whether delivered via minipumps, encapsulated cells, or either *in vivo* or *ex vivo* gene therapy, and the time course and extent of recovery correlates higher with outgrowth and synapse formation than with the numbers of cells surviving in the grafts. Similarly, intraventricular hypothalamic grafts only restore circadian rhythms and sexual function when associated with fiber reinnervation, not with cell survival within self-contained implants. Nevertheless, many developments involving encapsulation, gene therapy, and stem cell alternatives are predicated upon a pharmacological mechanism being sufficient; for some applications this is possible, but it must be borne in mind that optimal recovery in other applications will require the implanted cells to establish explicit connections with the host brain (see below).

Trophic Mechanisms

The role of the trophic environment to promote graft survival, growth, and integration has long been recognized, and conversely there is growing recognition that grafts can provide a source of trophic molecules to promote intrinsic processes of plasticity and repair in the damaged brain. This was first observed with adrenal grafts in the dopamine depleted brains of parkinsonian mice in which the intrinsic dopamine systems had been lesioned with the peripherally administered toxin MPTP. Although the catecholamine-secreting chromaffin cells themselves did not survive in the grafts, the host brain exhibited significant sprouting of spared dopamine fibers in the ventral striatum which extended to reinnervate the areas of extensive denervation in the dorsal striatum, and provided a more plausible mechanism for the observed recovery.

A range of trophic molecules have now been identified that promote the survival, growth, and differentiation of dopamine and other neurons in the developing embryo; that similarly promote cell survival *in vitro*;

and, critically for the present discussion, that can provide neuroprotection of host neurons to various toxic, traumatic, and disease-related assault. Thus, to take three examples, GDNF can protect dopamine neurons, CNTF can protect striatal and motor neurons, and NGF can protect cholinergic neurons from toxic lesions *in vitro*, leading to the proposal that treatment with these factors might promote survival, retard degeneration, and slow disease progress in PD, HD and ALS, and AD, respectively. In each case, the technical challenge is to achieve targeted delivery across the blood–brain barrier to the precise sites where neuroprotection is required and without the side effects that follow systemic administration. Identification of cells that spontaneously secrete relevant factors (such as the NGF-secreting mouse submaxillary gland) or *ex vivo* genetic manipulation of neutral cells to provide high levels of secretion of specific relevant trophic factors can provide an effective source of cells for transplantation and neuroprotection, and is now a major topic of research effort, not just in experimental studies but also in developing clinical trials for a variety of diseases.

Glial Support

Other diseases, such as multiple sclerosis, involve loss of the glial oligodendrocytes that myelinate axons and enable sultatory conduction within the neuronal pathways of the brain. Endogenous remyelination can be promoted by glial replacement, which is achieved most effectively in experimental models of focal demyelination by implantation of the oligodendrocyte precursors. There are several key issues associated with translating this into an effective clinical strategy: first, the sporadic progression of the disease with recurrent development of further inflammatory plaques needs to be halted, to which end significant advances are now being achieved by treatment with anti-inflammatory agents such as the Campath antibodies; second, whereas in some situations demyelination might be relatively focal and hence easily targeted, new strategies are required for reconstructing relatively dispersed foci of demyelination, perhaps by recruiting the capacity of precursor cells for migration; third, after prolonged demyelination in advanced disease, the integrity of demyelinated axons can itself be compromised, so effective treatments must be associated with strategic timing and additional neuroprotection.

In other situations, glial and mixed neuronal-glial grafts may provide stabilization of the central nervous system (CNS) environment. Such a strategy has been applied in spinal cord injury where embryonic cell grafts can occupy the cavity and prevent the progression of syrinx expansion; such a strategy may be anticipated to provide maximal benefit when the syrinx is in the process of extending rostrally to invade critical cervical levels of the spinal cord.

Pathway Repair

In contrast to the considerable regenerative potential of the peripheral nervous system (PNS), the mature CNS does not typically support long-distance axon growth through nontarget areas to distant targets, even when the neurons themselves retain the capacity for extensive elongation of neurites. Various lines of evidence suggest this is at least in part related to the surface expression by glia of inhibitory molecules, such as NoGo. To overcome this constraint, peripheral nerve grafts or tracks of permissive glial cells, such as Schwann cells, can be implanted to bridge sites of pathway transection and provide a glial substrate to promote and guide long-distance axon growth in the adult CNS. For example, oblique glial tracks can promote regeneration of axons across the nigrostriatal pathway, whether from nigral grafts implanted homotopically in the substantia nigra or indeed from intrinsic nigral dopamine neurons disconnected from striatal targets by a knife-cut axotomy of the projection.

The use of bridge grafts to promote pathway reconstruction has been explored in most detail in the spinal cord, where transection of ascending sensory and descending motor projections can result in major clinical disability. Considerable advances have been achieved in combining glial bridges alongside treatment with neuroprotective and anti-inhibitory molecules in bridging complete spinal transections such that host axons can sprout and grow across the site of injury and back into the distal spinal cord. However, further elongation of the axons in the distal cord and targeted reinnervation of their specific targets, whether in the brain stem and thalamus, or onto motor neurons in the ventral horns, remains extremely limited. Indeed, there is growing evidence that a greater degree of functional benefit may be achieved, at least with presently available technologies, by transplantation of regulatory neurons involved in less precisely organized projections, such as raphé grafts in the transected spinal cord, which although not restoring voluntary locomotion can reinnervate and restore important sympathetic functions such as bladder control.

Neuronal Reinnervation

As outlined above, many plastic neuronal populations (such as those derived from the developing

embryonic brain) can be grafted into the mature nervous system, and give rise to extensive neuronal innervation. Moreover, such grafts typically do make morphologically normal synaptic contacts with their appropriate targets. The fact that developing neurites can follow abnormal routes to reach appropriate targets suggests that even in the adult brain at least some of the developmental signals directing axonal elongation and targeting remain present. However, whether such grafts are functional depends critically upon the normal organization of the intrinsic pathway. In essence, regulatory systems such as many hypothalamic cell populations or monoaminergic and cholinergic neurons of the brain's isodendritic core can routinely be seen to alleviate the functional consequences of specific loss of the same populations of cells in diverse experimental models. Conversely, there is much poorer evidence that grafts in more precisely topographic systems (such as within the cortex, corticothalamic or thalamocortical systems, or within visual pathways) can exhibit any comparable recovery. This suggests that in the former case relatively diffuse reinnervation can still achieve functional reactivation or inhibition of the targets required for regulatory or 'gating' control, whereas in the latter case the developmental signals necessary to reestablish the 'point-to-point' connectivity, and required to restore relay of precisely patterned information, are no longer operative in the adult brain, at least within the relatively crude circumstances involved in cell transplantation. As a consequence, there is no simple principle that applies equally in all circumstances; rather, the functional efficacy of a graft will depend critically on the underlying organization of the normal system that is to be replaced or repaired.

Circuit Reconstruction

In most model systems, there is little evidence that grafted cells become fully integrated into the circuitry of the host nervous system sufficient as to restore integrated neuronal information processing. However, evidence of such a level of reciprocal graft–host integration is emerging in one particular model system: transplantation of striatal neurons in animals with lesions of intrinsic striatal circuits that mimick the core pathology seen in HD. Not only do striatal tissue grafts survive in the excitotoxically lesioned striatum, but they also have the capacity both to attract cortical, thalamic, and brain stem reinnervation by the normal population of host afferents that have lost their intrinsic targets, and to give rise to efferent projections to the normal proximal targets in the globus pallidus and in many cases more distally to the substantia nigra. Grafted medium spiny cells thereby become established into a connected host–graft–host neuronal circuitry that recapitulates the normal corticobasal ganglia circuit. Higher cognitive and executive functions mediated at the cortical level, and known to be dependent upon the integrity of this connected corticostriatal circuit, are restored following striatal transplantation that reconnects the circuit. Moreover, more detailed behavioral analysis has shown restitution of motor learning and habit formation believed to be mediated at the striatal level, and parallel recording studies have established long-term plasticity (involving both potentiation and depression) at the reformed host-graft synapses, which suggest a physiological substrate for the reestablished capacity for motor learning in the grafted brain.

Thus, at least in this one model system, there is growing evidence that under appropriate (albeit infrequent and perhaps fortuitous) circumstances, grafted neurons can truly integrate into the neuronal circuitry of the host brain so as to restore intrinsic information-processing abilities lost to an experimental lesion. The extent to which a similar strategy will find application in the related human disease HD is under current evaluation, and firm conclusions are still premature. Nevertheless, the experimental studies on the requirements for relearning in grafted animals suggest that it will not be sufficient simply to reimplant the lost neurons to restore the damaged circuitry; rather, a systematic program of rehabilitation will be required to recapitulate the learning of motor skills and habits that we routinely first acquire in normal childhood.

Full Network Repair

Notwithstanding the dramatic advances for repair and recovery so far achieved in animal models of spinal cord injury, stroke, or HD, no grafts approach full reconstruction of the normal circuitry. Even in the striatal graft model, the reformation of connections is incomplete and does not approach normal density, nor complex refinement of connections, and no model shows full functional recovery on all aspects of the underlying neurodegeneration or disease. Thus, full repair is not yet possible with available technologies in any neuronal system of the adult brain. It remains a matter of speculation whether the limits of current technologies are simply a matter of refinement. It is more likely that we have not yet identified all principles necessary to achieve complete reconstruction of the damaged nervous system.

Matching Donor Cells, Mechanisms, and Applications

The prospects of developing cell replacement strategies for CNS damage and neurodegenerative disease

remain optimistic. Nevertheless, consideration of the mechanisms by which grafts exert their functional effects in the host brain indicates that a common strategy is unlikely to fit all applications, and different solutions will be required for different diseases. To date, most studies have focused simply on identifying the cells that are lost and seeking alternative cell sources and surgical strategies for their replacement. The outcome has typically been that recovery is partial, some but not all symptoms show some alleviation, and the best recovery is seen on somewhat artificial measures of neuronal dysfunction rather than on the symptoms that are likely to be of most concern to patients. Amphetamine-induced rotation in the unilateral lesion model of parkinsonism is a powerful surrogate variable for evaluating experimentally the levels of underlying nigrostriatal loss and reinnervation, but hardly has face validity to the akinesia, rigidity, tremor, and dyskinesias experienced by PD patients.

To improve on this mismatch between the clinical demands and the animal models, the first requirement is to embed transplantation studies within the context of a detailed functional analysis of the underlying neuronal organization. What exactly is the level of repair required to achieve functional recovery on the symptoms of greatest practical concern? This analysis will constrain the options available for transplantation, since different sources of cells and different transplantation strategies will be required if the need is to restore functional neuronal connections, than if a pharmacological, trophic, or neuroprotective strategy can be effective.

Similarly, with the contemporary focus on developing alternative sources of stem cells or *ex vivo* gene therapy, the new therapies will find different applications depending on the functional demands placed upon them. Encapsulation of cells engineered to express particular trophic molecules may be highly effective for neuroprotection early in disease, and may promote intrinsic regeneration later in the disease course, but can never be effective in replacing a critical neuronal circuit whose component cells are already lost. Neuronal transplantation can be considered to be coming of age to the extent that the development of new strategies for therapy and repair is based on rational design rather than simple empirical cell replacement.

See also: Brain Injury: Functional Recovery After; Cell Replacement Therapy: Parkinson's Disease; Gene Therapy: Genetically Modified Cells; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Regeneration and Functional Recovery: Strategies.

Further Reading

Björklund A, Lindvall O, Isacson O, et al. (1987) Mechanisms of action of intracerebral neural implants – studies on nigral and striatal grafts to the lesioned striatum. *Trends in Neurosciences* 10: 509–516.

Bohn MC, Cupit L, Marciano F, and Gash DM (1987) Adrenal grafts enhance recovery of striatal dopaminergic fibers. *Science* 237: 913–916.

Brasted PJ, Watts C, Robbins TW, and Dunnett SB (1999) Associative plasticity in striatal transplants. *Proceedings of the National Academy of Sciences of the United States of America* 96: 10524–10529.

Cheng H, Cao YH, and Olson L (1996) Spinal cord repair in adult paraplegic rats – partial restoration of hind-limb function. *Science* 273: 510–513.

David S and Aguayo AJ (1981) Axonal elongation into peripheral nervous system 'bridges' after central nervous system injury in adult rats. *Science* 214: 931–933.

Dunnett SB and Björklund A (1994) *Functional Neural Transplantation*. New York: Raven Press.

Dunnett SB and Björklund A (2000) *Functional Neural Transplantation II. Novel Cell Therapies for CNS Disorders*. Amsterdam: Elsevier Science.

Dunnett SB, Isacson O, Sirinathsinghji DJS, Clarke DJ, and Björklund A (1988) Striatal grafts in rats with unilateral neostriatal lesions. III. Recovery from dopamine-dependent motor asymmetry and deficits in skilled paw reaching. *Neuroscience* 24: 813–820.

Fawcett JW, Rosser AE, and Dunnett SB (2001) *Brain Damage, Brain Repair*. Oxford, NY: Oxford University Press.

Freeman TB and Widner H (1998) *Cell Transplantation for Neurological Disorders*. Totowa, NJ: Humana.

Krieger DT, Perlow MJ, Gibson MJ, et al. (1982) Brain grafts reverse hypogonadism of gonadotropin releasing hormone deficiency. *Nature* 298: 468–471.

Perlow MJ, Freed WJ, Hoffer BJ, et al. (1979) Brain grafts reduce motor abnormalities produced by destruction of nigrostriatal dopamine system. *Science* 204: 643–647.

Pluchino S, Furlan R, and Martino G (2004) Cell-based remyelinating therapies in multiple sclerosis: Evidence from experimental studies. *Current Opinion in Neurology* 17: 247–255.

Ralph MR, Foster RG, Davis FC, and Menaker M (1990) Transplanted suprachiasmatic nucleus determines circadian period. *Science* 247: 975–978.

Sanberg PR, Wictorin K, and Isacson O (1994) *Cell Transplantation for Huntington's Disease*. Austin, TX: R.G. Landes.

Shihabuddin LS, Palmer TD, and Gage FH (1999) The search for neural progenitor cells: Prospects for the therapy of neurodegenerative disease. *Molecular Medicine Today* 5: 474–480.

Tan SA and Aebischer P (1996) The problems of delivering neuroactive molecules to the CNS. *Ciba Foundation Symposia* 196: 211–236.

Winkler C, Kirik D, and Björklund A (2005) Cell transplantation in Parkinson's disease: How can we make it work? *Trends in Neurosciences* 28: 86–92.

Cell Replacement Therapy: Parkinson's Disease

P Brundin and A Björklund, Wallenberg Neuroscience Center, Lund, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Cell replacement therapy for Parkinson's disease (PD) is based on the idea that implanted immature dopamine neurons may be able to substitute for the lost nigrostriatal neurons. The principal limitations of this approach are the practical and ethical problems associated with the use of tissue derived from aborted human embryos or fetuses and also the large numbers of donors needed to obtain good therapeutic effects. To date, differentiated neuroblasts and young postmitotic dopamine neurons have been used for transplantation. However, progenitors taken at earlier stages of development might prove more effective in the future. Efforts are under way to expand multipotent neural stem or progenitor cells *in vitro* and control their phenotypic differentiation into a dopaminergic neuronal fate. Initial results suggest that *in vitro* expanded cells can survive and function after transplantation to the striatum in the rat PD model, but the number of surviving dopamine neurons has in most cases been very low. With further development, expanded progenitors or dopamine neuron precursors, possibly modified by cell engineering techniques, may offer new sources of cells for replacement therapy in PD.

History

The first studies on the transplantability of brain catecholaminergic (including dopaminergic) neurons were conducted by Seiger and Olson in the early 1970s. These studies in rats showed that immature, developing dopamine neurons from the ventral midbrain can survive transplantation to the anterior eye chamber and extend an extensive axonal terminal network on the underlying iris. The experiments showed that dopamine neurons have to be obtained at embryonic or fetal stages of development, ideally at the time of development when they are generated. Inspired by these early observations, in 1979 Björklund and Stenevi and Perlow et al. showed that fetal dopamine neurons can survive, grow, and function after transplantation to the brain in rats subjected to a 6-hydroxydopamine (6-OHDA) lesion of the intrinsic nigrostriatal dopamine system. Extensive animal experiments in rodent and primate models of PD during the subsequent decades have shown that immature dopamine neurons and neuroblasts can survive and reinnervate the denervated striatum, as well as restore dopamine synthesis and release in the area. Based on these animal studies, the first clinical trials using cells obtained from embryonic human ventral mesencephalon (VM) were carried out in 1987–89. The observations from these initial trials were quite encouraging. To date, several hundred patients have received dopamine cell transplants in open-label and double-blind, placebo-controlled trials with mixed results.

Studies in Rodents and Primates

The dopamine neurons used for transplantation are harvested from the VM (i.e., the region containing the developing substantia nigra and the ventral tegmental area) at the time when the dopamine neurons are born – that is, from 13- to 15-day-old fetuses in the rat and 6- to 8-week-old human aborted embryos. In most cases, the cells are transplanted into the striatum (i.e., close to the denervated target cells). Thereby they can extend a new terminal network in the host striatum and establish functional synaptic contacts with the striatal target neurons. Only approximately 10% of the dopamine neurons that are implanted survive the graft surgery, and various neuroprotective agents and growth factors have been employed to boost this survival rate to some extent. Coadministration of growth factors such as glial cell line-derived neurotrophic factor can promote axonal outgrowth from the transplants.

Experiments performed in rodents and primates in which the nigrostriatal dopamine system is destroyed either by the 6-OHDA toxin (in rats) or by MPTP (in monkeys) have shown that grafted dopaminergic neuroblasts reverse some, but not all, motor symptoms associated with the loss of the intrinsic nigrostriatal system. The functional effect develops gradually: the first transplant-induced effects on motor behavior are seen after 3–6 weeks with rodent cells and after a few months with human dopamine neurons. Functional recovery is correlated with growth of axonal processes from the grafted cells and establishment of a new dopamine-containing terminal network in the host striatum. Rodent studies have shown that the grafted dopamine neurons are spontaneously active and can restore a tonic release of dopamine to near-normal levels within the reinnervated area. Several observations clearly indicate that the functional recovery is dependent on a specific interaction between dopamine neurons and the host striatum: (1) the recovered function is maintained only with continued

survival of the grafted neurons, (2) dopamine neurons transplanted outside the basal ganglia induce no improvement, and (3) non-dopaminergic neurons grafted to otherwise effective striatal sites exert no effect.

In rats with 6-OHDA lesions, intrastriatal VM grafts containing between 1000 and 2000 surviving dopamine neurons (approximately 10–15% of the number of dopamine neurons normally present in the adult substantia nigra) are known to be sufficient to reverse some behavioral deficits. They can restore striatal dopamine levels to approximately 10–30% of normal, *in vivo* striatal dopamine synthesis to approximately 20–30% of normal levels, and striatal dopamine release (as measured by microdialysis in the area densely innervated by the grafted neurons) up to 40–100% of the levels seen in the intact striatum. Using multiple graft placements, it has been possible to obtain a striatal dopamine innervation density in the range of 40–75% of normal (as assessed by tyrosine hydroxylase immunoreactivity). Even with such large grafts, the recovery of lesion-induced motor deficits is incomplete and varies depending on the type of test. Thus, motor asymmetry assessed by drug-induced turning behavior is completely reversed, whereas several aspects of spontaneous motor behavior, such as stepping, forelimb use, and postural balance, are only partly restored.

There are several possible reasons why graft-induced functional recovery is limited to certain basic aspects of motor function, whereas impairments in more complex motor behavior are influenced to a lesser extent, if at all. The transplants are placed in an ectopic location (i.e., in striatum rather than substantia nigra) in order to obtain extensive reinnervation of the striatum from the grafted neurons. This may impose limitations on their functionality because the grafted dopamine neurons are likely to lack crucial afferent regulatory inputs. In the intact brain, baseline striatal dopamine neurotransmission is normally maintained by tonic synaptic and nonsynaptic dopamine release, which is largely independent of changes in neuronal impulse flow in the nigrostriatal pathway. In grafted animals, the establishment of a new dopamine-storing terminal network in the host striatum is closely correlated with the onset of graft-derived dopamine release and reversal of dopamine receptor supersensitivity. This suggests that the grafted neurons are spontaneously active and release dopamine at both synaptic and nonsynaptic sites at a near-normal rate. The grafted neurons are probably not entirely devoid of host afferent control, but these contacts might not be fully appropriate for normal function. In 1991, Fisher et al. showed that more than 50% of the grafted neurons respond to stimulation of either frontal cortex or striatum, and burst firing, which normally depends on a functional cortical input, is present in a portion of them. Neuroanatomical data also indicate that some afferent inputs (e.g., from premotor cortex and/or striatum) may contribute to the functional efficacy of intrastriatal dopamine neuron transplants. An alternative explanation for incomplete functional recovery seen in animals with intrastriatal dopamine neuron transplants is that the graft-derived innervation is confined to the striatum, whereas the denervating lesion also involves other nonstriatal areas. The inability of intrastriatal VM grafts to provide complete functional recovery may be due to the fact that limbic and cortical areas, as well as other parts of the basal ganglia, remain denervated and thus unaffected by the transplanted cells. In support of this view, in 2001 Kirik et al. showed that graft-reduced functional recovery is more pronounced in animals with partial lesions of the nigrostriatal pathway, in which the denervation is confined to the striatum, than in animals with lesions which involve both striatal and nonstriatal areas. This indicates that it may be functionally important that part of the nonstriatal dopamine projections remains intact. These spared portions of the host dopamine system, particularly those innervating nonstriatal forebrain areas, may be necessary for the VM grafts to exert optimal functional effects.

Clinical Trials

In 1990, for the first time, neural transplants were reported to survive and reduce neurological symptoms in a PD patient. The donor tissue was obtained from embryos at routine abortions. The trials were preceded by an ethical debate. In many countries, it was considered justifiable to use donor tissue from abortions if clearly separated from the transplantation surgery and the women undergoing abortions had given their informed consent. In the first trial, tissue was dissected from the VM of multiple 6- to 8-week-old human embryos, mechanically dissociated, and then injected into the striatum using stereotactic surgery on one side of the brain in a patient receiving immunosuppression. After 4 or 5 months, the patient started to exhibit a progressive reduction in bradykinesia and rigidity, whereas the tremor remained largely unaffected. Using positron emission tomography, it was possible to monitor an increase in fluorodopa uptake in the graft site, indicating that dopaminergic neurons in the implant had survived. Several subsequent reports confirmed that grafted embryonic neurons can alleviate symptoms and replace the dopaminergic neurotransmission that is lost in PD. In extreme cases, the fluorodopa uptake

approached that of a normal striatum, comparable to the 10–30% of normal levels seen in the PD patients prior to surgery or in unoperated parts of their brains. Ten years after surgery, which is the longest reported follow-up, transplanted neurons have been shown to still be alive and capable of releasing dopamine. Functional positron emission tomography studies have indicated that the grafted cells integrate with basal ganglia – cortical circuits that participate in fine motor control. This functional integration seems to take several months longer than for the grafted neurons to start producing dopamine and appears to coincide with further improvements of symptoms.

The brains of a few patients who died several years postoperatively (unrelated to the graft surgery) have been examined microscopically. These have revealed that approximately 5–10% of the grafted dopaminergic neurons survive and extend a dense network of axons into the host brain. Interestingly, some of the grafts are infiltrated with immunocompetent cells. Although the brain is an immunologically privileged transplantation site (i.e., histoincompatible grafts survive longer in the brain than in peripheral sites), it is clear that transplant rejection can take place in the brain. The level of immunosuppressive treatment that is required to avoid rejection with certainty is not known. Consequently, different clinical trials have employed widely different strategies to deal with the risk of immune rejection.

Approximately 300 patients have undergone cell transplantation surgery in several small clinical trials. However, not all patients benefited to the same extent. Approximately one-third of patients demonstrated remarkable recovery and were able to stop taking anti-parkinsonian medication, one-third were moderately improved, and one-third exhibited no clear lasting benefit from the surgery. Variation in outcome was evident even among patients operated at the same surgical center. The small trials were conducted in independent centers and therefore not coordinated with each other. As a consequence, there were several potentially important variations in the experimental protocols regarding, for example, disease stage and age of patients, donor tissue age, numbers of donors, graft storage and preparation technique, surgical methodology, and the use of immunosuppression. Therefore, it is difficult to compare the results from the different centers and devise a protocol that would have a good chance of reproducible success. Another shortcoming of the initial neural transplantation trials was that they were open-label (i.e., both the patient and the investigator were aware that the patient was receiving a graft). This meant that there was a risk that the reported beneficial effects of the transplants were due to a placebo effect or observer bias.

In the mid-1990s, the US National Institutes of Health (NIH) sponsored two large randomized, double-blind, placebo-controlled clinical trials with neural transplants in PD. The outcomes of these trials, published in 2001 and 2003, were largely disappointing to patients and researchers. In brief, there was no significant symptomatic relief in the grafted patients compared to the control subjects, according to primary outcome parameters. The positron emission tomography scans and postmortem morphological examinations showed that grafted dopamine neurons had survived, despite the lack of clear beneficial effects of the surgery. Although individual patients in these two trials may have experienced graft-related improvement, these findings clearly question whether cell transplantation is a viable strategy for PD. It is important to note, however, that there are several outstanding issues regarding how the trials were performed. First, the donor tissue was stored prior to surgery for between 2 days and 4 weeks, which may have added variability in the quality of the implants. Second, in one of the trials an unconventional frontal (as opposed to dorsal) stereotactic approach was used to target the putamen. The impact of this surgical technique on graft function is not known. Third, in one of the trials no immunusuppression was given, and in the other it was limited to a 6-month administration of cyclosporine. Fourth, there is an indication that younger PD patients who are still responsive to anti-parkinsonian medication respond better to the surgery than aged patients, and selection of relatively old patients may have affected the overall outcome. Taken together, the two NIH-sponsored trials provided very valuable information but clearly showed that neural grafts do not reliably provide clinically valuable effects with the particular protocols they employed.

Aside from the lack of graft-induced benefit, the first of the two NIH-sponsored trials also reported that some patients (approximately 15%) experienced involuntary movements (dyskinesias) that were caused by the implants. This finding was unexpected and it was particularly disturbing that the dyskinesias did not improve when the anti-parkinsonian medication was terminated. Some of the patients from open-label trials and the patients in the second NIH-sponsored trial were also found to have developed graft-induced dyskinesias of varying severity. The reports of graft-induced dyskinesias and problems in obtaining suitable donor tissue in a reproducible manner brought clinical transplantation trials in PD to a standstill in the early 2000s. The pathogenetic mechanisms underlying the involuntary movements

are not understood. Several possible explanations have been proposed, including that there is too much or unregulated dopamine release from the implants in certain striatal regions, that non-dopaminergic neurons in the implants influence the striatum, or that inflammation due to an immune response triggers the movements. Concerted efforts in experimental animals are currently under way to elucidate why neural transplants can sometimes elicit this undesirable side effect. It is of vital importance to understand when, how, and why graft dyskinesias occur because they could conceivably also develop following transplantation of stem cell-derived neurons.

Future Cell Replacement Strategies

It is important to emphasize that so far no systematic stem cell transplantation trials have been performed for PD. The cells used in clinical trials have all been postmitotic immature neurons that have already been committed to the dopaminergic neuron fate. Stem cells are characterized by their ability to self-renew and to differentiate into many different forms of cells. There are various types of stem cells related to the age of the organism in which they are found and their anatomical location. Research is ongoing in experimental models of PD aimed at developing a therapy based on neurons derived from stem cells.

An existing proposal is to utilize endogenous neurogenesis in the adult brain as a source of new dopamine neurons. Under normal conditions, neurogenesis takes place in the adult mammalian brain in at least two regions, namely the dentate gyrus of the hippocampus and the subventricular zone, which supplies the olfactory bulb with new neurons. Claims that new dopaminergic neurons can be generated in the substantia nigra of adult rodents and that lesion-induced motor deficits can be reversed are highly controversial, with several studies disputing these findings. Extensive basic research is needed before adult neurogenesis can be proposed as a strategy for brain repair in PD.

Currently, it seems more feasible that it will be possible to transplant neurons derived from stem cells in PD. These stem cells are essentially one of two transplantable types – adult-derived or embryonic stem (ES) cells. Concerning stem cells obtained from adults, several investigators have suggested that it may be possible to transplant dopamine neurons derived from stem cells located outside the central nervous system. These neurons would be the result of a phenomenon called transdifferentiation; that is, the stem cells differentiate into a cell type that is not associated with the anatomical structure in which they are located. For example, it has been suggested that neurons suitable for transplantation can be generated from stem cells residing in bone marrow or skin. These ideas have been questioned, and it is currently unclear if neurons can be generated by transdifferentiation. In some experimental paradigms in which transdifferentiation was initially suggested to have taken place, the host neurons appear to have undergone fusion with injected stem cells. Another option is to harvest stem cells from the developing or adult nervous system and coax them by specific culture media or genetic modification into becoming dopamine neurons. Although there exist several cell culture protocols that allow propagation of central nervous system stem cells *in vitro*, it is difficult to manipulate these to differentiate into dopamine-producing neurons that can survive grafting. Therefore, new experimental strategies need to be developed before neural stem cells can provide a reliable source of dopamine neurons for transplantation in PD.

The most promising option for a stem cell therapy in PD is to employ ES cells. The ES cells are derived from the inner cell mass of a blastocyst and can be propagated in large numbers *in vitro*. Mouse ES cells can be induced to differentiate into dopamine-producing neurons. These neurons have successfully been grafted to rodent models of PD, in which they innervate the brain and reverse lesion-induced motor deficits. Dopamine neurons have also been derived from nonhuman primate ES cells and successfully grafted to monkeys. In addition, research has demonstrated that it is possible to generate dopaminergic neurons from human ES cells in culture. Unfortunately, the success rate has been low when transplanting dopamine neurons derived from human ES cells into rat models of PD. This problem needs to be resolved before translation into a clinical protocol can continue. Possible explanations are that the dopaminergic neurons die because they are particularly sensitive to the grafting procedure and/or that they downregulate some key phenotypic or regulatory proteins after implantation into the adult brain.

At least three other important issues need to be resolved before dopaminergic neurons derived from human ES cells can be tested in PD patients. First, ES cells are pluripotent and therefore they can form teratomas (tumors containing tissues from all three germ layers) upon grafting. A clinical protocol would require that all pluripotent stem cells are purged from the transplant preparation. Even a single remaining pluripotent or otherwise tumorigenic stem cell could have devastating consequences for a PD patient receiving an intracerebral implant. Second, most current cell culture protocols for human ES cells include animal-derived products (culture media, feeder cells, etc.). To eliminate the risk of transfer of potentially dangerous zoonooses (e.g., species-specific

retroviruses) and to reduce the risk of graft rejection due to transfer of animal-derived molecules to the graft cells, the ES cells should be grown without any animal products present. Third, further research is required to analyze the type of dopaminergic neurons that are generated from ES cells. It is important that the dopamine neurons used for cell replacement therapy have a midbrain-specific phenotype. To be fully functional in the striatum, it is likely that they need to possess the properties of substantia nigra dopaminergic neurons because the specific physiological features of nigral neurons differ from those of dopaminergic neurons from other brain regions. Advances in transcriptional control have provided information that can help devise protocols that generate large numbers of midbrain dopamine neurons from human ES cells. Overexpression of key regulatory transcription factors, for example, can be used to achieve a high yield of dopamine neurons from mouse ES cells. Future research will reveal if that is also the case for human ES cells.

See also: Animal Models of Parkinson's Disease; Cell Replacement Therapy: Mechanisms of Functional Recovery; Cell Replacement Therapy for Huntington's Disease; Gene Therapy: Genetically Modified Cells; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Neuronal Plasticity after Cortical Damage; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Alvarez-Dolado M, Pardal R, Garcia-Verdugo JM, et al. (2003) Fusion of bone-marrow-derived cells with Purkinje neurons, cardiomyocytes and hepatocytes. *Nature* 425: 968–973.

Annett L (1994) *Functional Studies of Neural Grafts in Parkinsonian Primates*. New York: Raven Press.

Bjorklund A, Dunnett SB, Brundin P, et al. (2003) Neural transplantation for the treatment of Parkinson's disease. *Lancet Neurology* 2: 437–445.

Bjorklund A and Stenevi U (1979) Reconstruction of the nigrostriatal dopamine pathway by intracerebral nigral transplants. *Brain Research* 177: 555–560.

Brundin P, Duan WM, and Sauer H (1994) *Functional Effects of Mesencephalic Dopamine Neurons and Adrenal Chrafffin Cells Grafted to the Rodent Striatum*. New York: Raven Press.

Brundin P, Karlsson J, Emgard M, et al. (2000) Improving the survival of grafted dopaminergic neurons: A review over current approaches. *Cell Transplantation* 9: 179–195.

Cenci MA and Hagell P (2006) Dyskinesias and neural grafting in Parkinson's disease. In: Brundin P and Olanow CW (eds.) *Restorative Therapies in Parkinson's Disease*, pp. 184–224. New York: Springer.

Doucet G, Murata Y, Brundin P, et al. (1989) Host afferents into intrastriatal transplants of fetal ventral mesencephalon. *Experimental Neurology* 106: 1–19.

Fisher LJ, Young SJ, Tepper JM, Groves PM, and Gage FH (1991) Electrophysiological characteristics of cells within mesencephalon suspension grafts. *Neuroscience* 40: 109–122.

Freed CR, Breeze RE, Rosenberg NL, et al. (1992) Survival of implanted fetal dopamine cells and neurologic improvement 12 to 46 months after transplantation for Parkinson's disease. *New England Journal of Medicine* 327: 1549–1555.

Freed CR, Greene PE, Breeze RE, et al. (2001) Transplantation of embryonic dopamine neurons for severe Parkinson's disease. *New England Journal of Medicine* 344: 710–719.

Frielingsdorf H, Schwarz K, Brundin P, and Mohapel P (2004) No evidence for new dopaminergic neurons in the adult mammalian substantia nigra. *Proceedings of the National Academy of Sciences of the United States of America* 101: 10177–10182.

Kim JH, Auerbach JM, Rodriguez-Gomez JA, et al. (2002) Dopamine neurons derived from embryonic stem cells function in an animal model of Parkinson's disease. *Nature* 418: 50–56.

Kirik D, Winkler C, and Bjorklund A (2001) Growth and functional efficacy of intrastriatal nigral transplants depend on the extent of nigrostriatal degeneration. *Journal of Neuroscience* 21: 2889–2896.

Kordower JH, Freeman TB, Snow BJ, et al. (1995) Neuropathological evidence of graft survival and striatal reinnervation after the transplantation of fetal mesencephalic tissue in a patient with Parkinson's disease. *New England Journal of Medicine* 332: 1118–1124.

Kordower JH, Styren S, Clarke M, DeKosky ST, Olanow CW, and Freeman TB (1997) Fetal grafting for Parkinson's disease: Expression of immune markers in two patients with functional fetal nigral implants. *Cell Transplantation* 6: 213–219.

Lindvall O, Brundin P, Widner H, et al. (1990) Grafts of fetal dopamine neurons survive and improve motor function in Parkinson's disease. *Science* 247: 574–577.

Olanow CW, Goetz CG, Kordower JH, et al. (2003) A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson's disease. *Annals of Neurology* 54: 403–414.

Olson L and Seiger A (1972) Brain tissue transplanted to the anterior chamber of the eye: 1. Fluorescence histochemistry of immature catecholamine and 5-hydroxytryptamine neurons reinnervating the rat iris. *Zeitschrift fur Zellforschung und Mikroskopische Anatomie* 135: 175–194.

Perlow MJ, Freed WJ, Hoffer BJ, Seiger A, Olson L, and Wyatt RJ (1979) Brain grafts reduce motor abnormalities produced by destruction of nigrostriatal dopamine system. *Science* 204: 643–647.

Perrier AL, Tabar V, Barberi T, et al. (2004) Derivation of midbrain dopamine neurons from human embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* 101: 12543–12548.

Piccini P, Brooks DJ, Bjorklund A, et al. (1999) Dopamine release from nigral transplants visualized *in vivo* in a Parkinson's patient. *Nature Neuroscience* 2: 1137–1140.

Piccini P, Lindvall O, Bjorklund A, et al. (2000) Delayed recovery of movement-related cortical function in Parkinson's disease after striatal dopaminergic grafts. *Annals of Neurology* 48: 689–695.

Rosenblad C, Martinez-Serrano A, and Bjorklund A (1996) Glial cell line-derived neurotrophic factor increases survival, growth and function of intrastriatal fetal nigral dopaminergic grafts. *Neuroscience* 75: 979–985.

Takagi Y, Takahashi J, Saiki H, et al. (2005) Dopaminergic neurons generated from monkey embryonic stem cells function in a Parkinson primate model. *Journal of Clinical Investigation* 115: 102–109.

Cell–Cell Communication through the Extracellular Space

K Fuxe, Karolinska Institutet, Stockholm, Sweden
L F Agnati, University of Modena, Modena, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In any computational system (network), three aspects are of fundamental importance: modalities according to which computation is carried out in its computing elements (nodes), communication pathways (channels) among its nodes, and processes that allow signal transmission along the channels. Sometimes these aspects are not easily separated, as seems to be the case for the central nervous system (CNS). This article briefly discusses these aspects and some views on their biochemical and biophysical substrates.

Two competing theories on interneuronal communication have dominated neuroscience from the dawn of the modern discipline in the 1890s: the neuron doctrine and the theory of the diffuse nerve network, supported by Cajal and Golgi, respectively. The article provides a brief account of a view on interneuronal communication in the brain that reconciles these apparently opposing views. This unifying theory considers two main classes of intercellular communication which can be distinguished on the basis of the communication channels connecting communicating cells in the CNS: volume transmission (VT), which uses the extracellular space as a communication channel, and wiring transmission (WT), which uses cellular substrates (e.g., axons and synaptic contacts) as a communication channel. Thus, WT mode is characterized by physical contact (gap junctions) or quasi-contact (chemical synapses) between the cells of the net.

Furthermore, a biochemical mechanism, the receptor–receptor interaction (RRI), is discussed since it suggests a mechanism capable, *inter alia*, of integrating WT and VT signals at the plasma membrane level. In the early 1980s, Agnati and Fuxe introduced the concept of direct physical interactions between receptors at the membrane level and called the phenomenon RRIs. During the 1980s and after, they provided the first experimental observations for the existence of such direct physical interactions between receptors at the membrane level. Their proposal was supported by biochemical binding analysis in cell membrane preparations and by indirect evidence from functional studies in animal models. From the beginning, RRIs have been considered as an important case of protein–protein interactions. In 2002, the concept of WT and VT was extended to molecular networks, and RRIs have been considered examples of a 'wired' communication between computing 'nodes' in a molecular network, whereas the intracellular diffusion of second messengers between molecular nodes is an example of VT in molecular networks. In view of the existence of an organized extracellular matrix which is a molecular dynamic structure capable of storing and releasing signals (e.g., trophic factors) and interacting with receptors on the cell membranes (e.g., integrins), we have proposed the existence of a global molecular network (GMN) enmeshing the entire CNS. Because of the heuristic importance of this theory and, in particular, its potential value as a unitary frame for the CNS integrative action, a brief account of this theory is given.

Communication Processes in the Brain: WT versus VT

WT and VT in the Cellular Networks

The integrative actions of the CNS depend on all cell types present in it. Not only neuronal networks but also mainly complex cellular networks should be considered. A complex cellular network (CCN) can be defined as the set of cells of any type which exchanges signals in a certain volume of brain tissue and, due to this cross-talk, is capable of integrating inputs to give appropriate outputs. Let us consider some main aspects of VT versus WT (Figure 1).

Characteristics of the channel allow VT to be distinguished from WT VT is characterized by a channel with a poorly defined physical substrate and signal transmission takes place by diffusion and/or convection in the medium interposed between nodes. 'Transmission privacy,' when present, is usually given by the chemical code (e.g., the specificity of the chemical signal/receptor recognition process).

WT is characterized by the transmission of the signal along a channel with a well-defined physical substrate; thus, a 'wire' links the source node with the target node. Transmission privacy, which is usually present, is given by the fact that the channel is a private channel. Classically, in the case of CCN the WT channel between neurons is formed by axons and synapses.

Main features of VT and WT It is sometimes possible to summarize the main features of VT and WT in absolute terms but more often by means of a comparison between them:

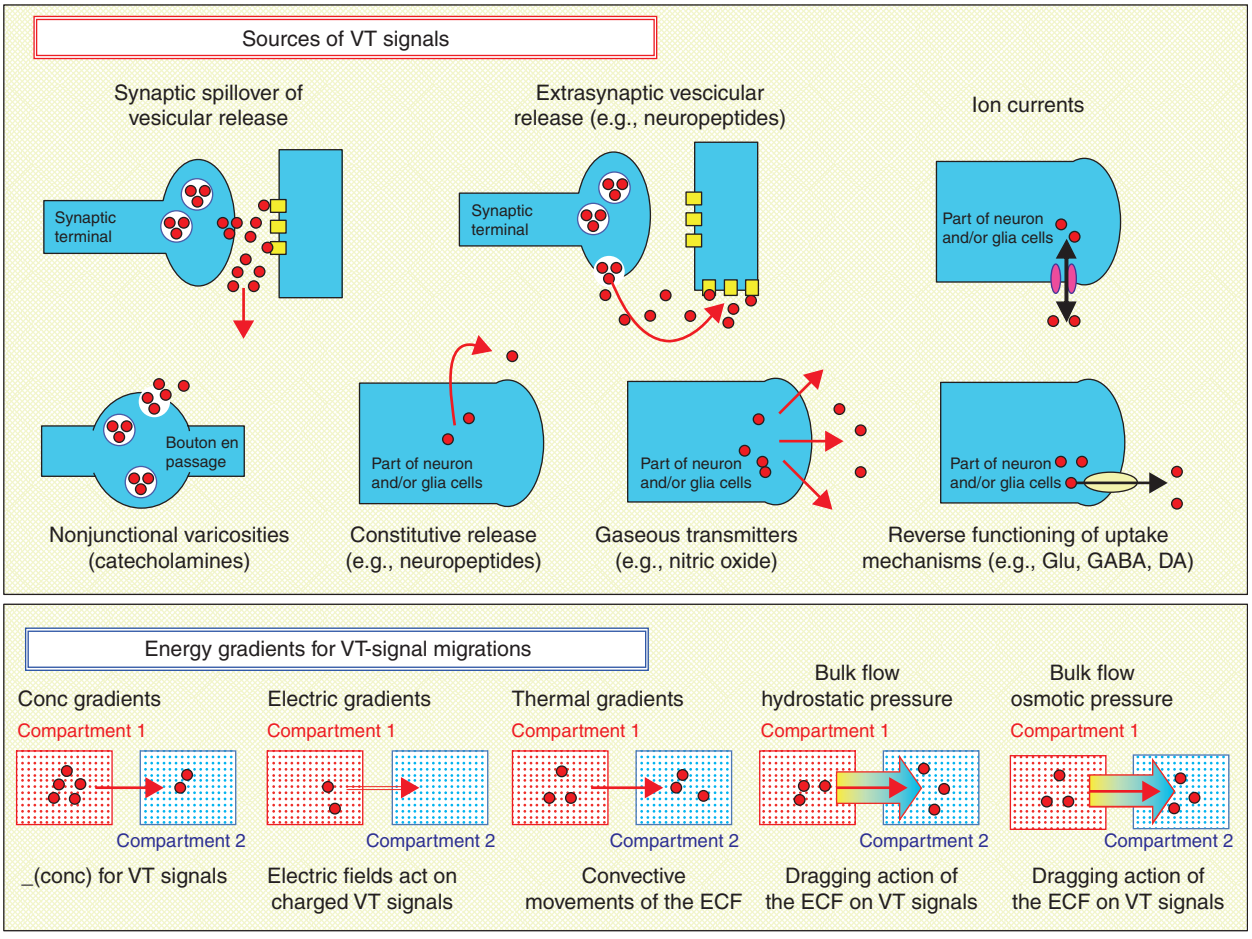

Figure 1 Schematic representation of the main characteristics of VT in the cellular networks of the brain. The main sources of VT signals are illustrated, as well as the energy gradients making their migration in the extracellular space of the brain possible. DA, dopamine; ECF, extracellular fluid; GABA, γ-aminobutyric acid; VT, volume transmission; WT, wiring transmission.

- *Main characteristic of VT.* A structurally poorly defined channel, made by the medium which pervades the network, is always interposed between the source node and the target node.
- *Main characteristic of WT.* A structurally well-defined channel connects the source node with the target node when the two nodes are not in direct physical contact.

Besides the type of channel, additional criteria help to differentiate VT from WT and also to classify different types of VT and WT, namely the distance between the source and the target of the signal and the type of signal. The source–target distance (in the case of the absence of a structurally well-defined channel) provided indirect evidence for the existence of VT based on the demonstration of the so-called 'transmitter–receptor mismatches' – that is, the spatial uncoupling between the release sites of a transmitter and the receptors capable of its recognition and decoding.

Transmitter–receptor mismatches have been especially studied in the nucleus accumbens, in which dopamine (DA)-terminal/DA D1 receptor mismatches have been observed. This mismatch can be linked to a possible role of uncoupling proteins (UCPs) in VT via their generation of thermal gradients, enhancing migration of DA signals by microconvection.

Functional implications of VT and WT for brain integrative functions VT and WT have different characteristics and thus are employed for different functions. From a physiopathological standpoint, it should be noted that in some instances, VT can play a vicarious role in the presence of a WT deficit.

VT occurs mainly in cells that communicate at a distance longer than a synaptic cleft by means of chemical signals or physical signals. These signals diffuse in the extracellular space (ECS) medium interposed

between the source and the target cell where the signal is decoded.

Thus, the following features can be distinguished: the type of VT signals, modalities of VT signal release, pathways for VT signal migration, energy gradients for VT signal migration, and decoding systems for VT signals. Diffusion of chemical VT signals in the brain is made possible by concentration gradients, and convection of VT signals is favored by thermal gradients (produced by both localized high neuronal firing and UCPs) and by the heart-produced pressure waves within the skull (tide hypothesis). It is clear that VT is very well suited for functions that do not require prompt activation of effector systems but, rather, make possible their long-lasting activation, as observed for hunger, sexual arousal, mood, and sleep. These functions are carried out especially at the level of the so-called visceral or limbic brain, where Nieuwenhuys, on the basis of his neurochemical mapping studies, suggested that a paracrine-like transmission was in operation.

WT occurs mainly in chemical synapses, gap junctions, and specialized close membrane juxtapositions in which any movement of substances, such as ions, metabolites, and transmitters, can affect juxtaposed processes of other cells. It is clear that WT is very well suited for functions that need prompt activation of effector systems and/or prompt change in the level of their activation or even their inhibition. Functions under main WT control are motor control, face recognition, and language. These functions can be carried out especially at the level of the phylogenetically most recent brain, the neopallium.

The Concept of RRIs and Receptor Mosaics

As indicated previously, these two concepts are strictly interrelated since the physical interactions among receptors in a receptor cluster are at the core of the concept of receptor mosaic (RM). Thus, the development of the two concepts and their experimental demonstration are strictly linked to each other.

Emergence of the Concept of RRIs and RMs

The hypothesis that some elaboration of information could occur at the membrane level was included in the electrophysiological concept of integration of the local currents at the level of the axon hillock leading to the triggering of or failure to trigger the action potential by the nerve cell. Even more closely linked to the concept of RRI is the hypothesis of Hodgkin and Huxley that ion fluxes via an ionotropic receptor can cause changes in the membrane potential, which in turn can change the conformation of membrane proteins such as ion channels. In fact, based on this proposal it was reasonable to deduce that the affected proteins could also be other receptors.

Cross-talk between receptors: The classical view It is therefore possible to maintain that the first proposal on RRIs could be made as an obvious consequence of Hodgkin and Huxley's hypothesis on the gating of ion channels (e.g., changes in conformation of proteins) in response to changes in the membrane potential. However, the first real demonstration of an indirect RRI was provided by Greengard in 1976, who showed that metabotropic receptors could alter the conformation of other membrane proteins, such as receptors, via a process of phosphorylation/dephosphorylation. During the same time period, Lefkowitz and Limbird demonstrated a site–site interaction between β-adrenergic receptors in erythrocyte membrane preparations (Figure 2).

RRIs: A new integrative mechanism at the membrane level In the 1980s, a new paradigm was proposed that one receptor (e.g., for peptides) could change conformation of another receptor for a completely different ligand (e.g., monoamines) via a direct physical interaction at the intramembrane level, and this interaction could also occur within a cluster of topographically ordered receptors of different classes called RMs. Furthermore, in 1987, Fuxe and Agnati introduced a proposal that can be viewed as an *ante litteram* program of proteomics: "Very likely we will find that some sophisticated elaborations of information are performed at membrane level, via interactions within and among different classes of macromolecules (such as receptors, ion pumps, ion channels. . . ."

RRIs make possible a vast miniaturization of the computational circuits and provide the cell with a first microcircuit capable of integrating and filtering extracellular signals and that can be subtly modulated by the intracellular biochemical machinery in terms of changes in number, subtype, and topology of the receptors in the RM (Figure 3).

Hence, the concept of the RM is important since it means that one can produce markedly different molecular aggregates of receptors by putting together similar building blocks to construct functionally different assemblies. These can store information and at the synaptic level change the synaptic weight, and thus be involved in at least some forms of learning.

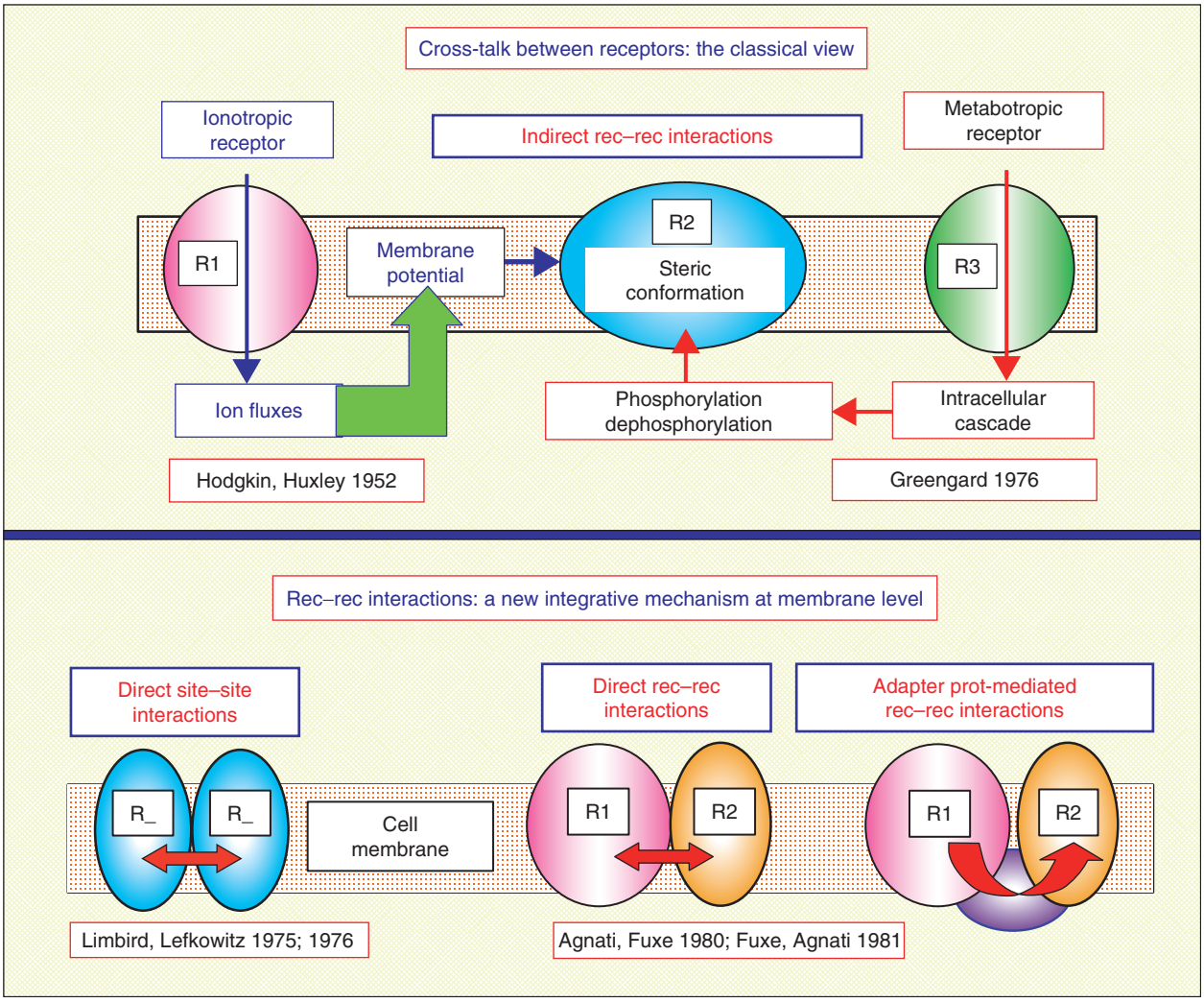

Figure 2 Schematic of the development of the concept of receptor–receptor interactions. Two classes are indicated, namely the indirect interactions between receptors involving a bioelectrical mechanism (change of membrane polarization) or an enzymatic mechanism (phosphorylation/dephosphorylation) and direct interactions based on intramembrane receptor–receptor interactions in heteromers or homomers with or without the involvement of adapter proteins.

Biochemical mechanisms involved in the assemblage of RMs have been partially clarified, and classification of RMs based on the absence or presence of classical cooperativity has been performed.

Implications of the Concept of RRIs and RMs

A new way to look at the decoding of extracellular signals arose from these two innovative concepts. Thus, special microdomains are present in the membrane and molecular networks are present in these specialized domains (Figure 3). Based on these concepts and data of our group and other groups, new characteristics of the cell biochemical machinery have been discovered and new possibilities for handling the huge number of chemical and electrical signals impinging on the cell have been described.

Biochemical implications

- Via the intramembrane regulation of receptor signaling, the biochemical pathways in the cytoplasm toward especially the nucleus can work on 'conditioned' signals. The regulation by RRIs involves modulation of receptor recognition (K_D and B_{max} values) and of G-protein coupling leading to modulatory actions in the activation or inhibition of multiple effector systems associated with the plasma membrane, such as ion channels and enzymes (e.g., adenylate cyclase).
- The RRI may make possible the appearance of novel receptor subtypes such as the $GABA_B$ receptor. Also, A1R/P2Y1R heteromerization leads to the appearance of an A1 receptor with a binding site showing P2Y agonist-like pharmacology. Thus, the

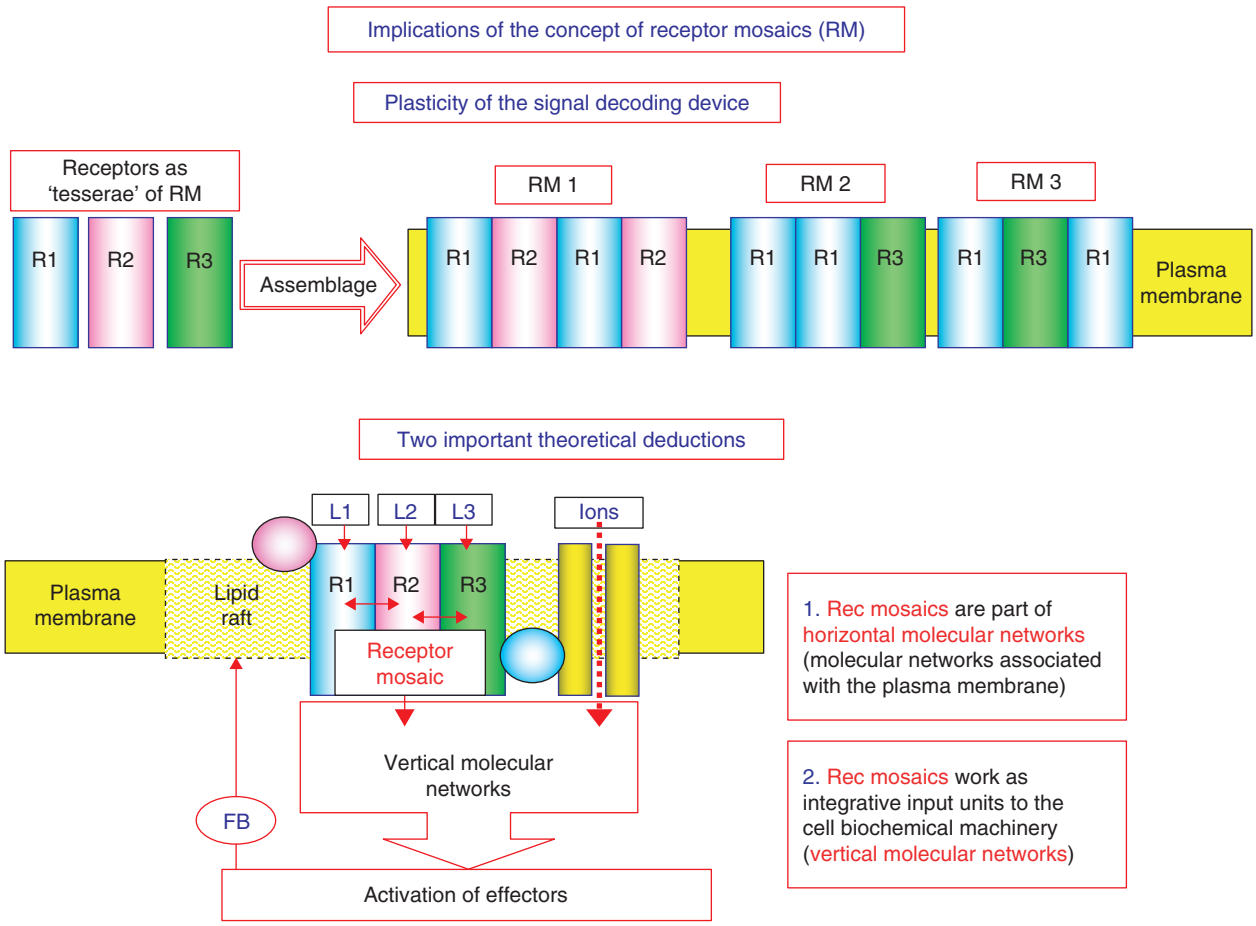

Figure 3 (Top) The importance of receptor multiplicity and topology in the assemblage of the receptor mosaics (RM). (Bottom) Schematic illustration of the RM (more than two receptors) as an integrative unit before sending the information into the vertical molecular networks involving multiple effectors and as part of horizontal molecular networks in the lipid rafts of the plasma membrane.

pharmacology of the binding pockets in the receptors participating in the formation of the heteromer may become markedly modulated.

- The conformational changes in the receptors caused by the RRI may lead to the formation of novel interactions with other membrane proteins, especially other receptors, including ligand-gated ion channels and different types of G-proteins. Thus, novel recognition-decoding systems may be formed and others may disappear, with the RM having novel interactions with other membrane-associated proteins such as scaffolding and adapter proteins.
- The RRIs also have a major impact on receptor co-trafficking, receptor maturation, cell surface expression, and internalization. Such events are of major relevance for sensitization and desensitization of receptors and especially for their cross-modulation.

RRIs between synaptic and extrasynaptic receptors integrate WT and VT signals The direct D5–$GABA_A$ and D1–NMDA receptor interactions are examples of paramount importance of such integrations since in the first case the GABA WT can be integrated with the DA VT and in the second case the Glu WT can be integrated with the DA VT.

VT integration with WT allows *inter alia* the combination of the main functional characteristics of the two communication modes. For example, it allows the prompt WT activation of the effector systems and their long-lasting activation via VT. This combination of the two modes can optimize the response of the organism to internal and/or external demands. An example is the 'fight-or-flight' reaction, in which a prompt response is needed that has to be backed up by sustained activation of the CNS alert systems and the cardiovascular and motor apparatuses.

RRIs: Relevance for Molecular Neuropsychopharmacology

Data from our group and other groups have demonstrated the existence of D2/A2A/mGluR5 mosaics at the level of the basal ganglia. On the basis of this evidence, new pathogenetic mechanisms and

therapeutic approaches have been proposed for Parkinson's disease and schizophrenic syndromes. Furthermore, our group has demonstrated GalR/5HT 1A mosaics in the tel- and diencephalon. Thus, it has been suggested that this RM may be a target to investigate pathogenetic mechanisms of depression and for the development of new antidepressant treatments.

Pathological Implications and Novel Treatments of Parkinson's Disease, Schizophrenia, and Depression

Development of A2A receptor antagonists in the treatment of Parkinson's disease based on the A2A/D2 receptor interaction in the dorsal striatopallidal GABA pathway Antagonists for A2A receptors may represent novel antiparkinsonian drugs targeting the A2A/D2 heteromer in the striatopallidal GABAergic neurons innervating the external segment of globus pallidus, where the antagonistic A2A/D2 receptor interaction occurs, leading to reduced D2-mediated inhibition of the dorsal striatopallidal GABA pathway which causes motor inhibition.

Development of mGluR5 antagonists for treatment of Parkinson's disease based on multiple mGluR5/A2A/D2 receptor interactions in the dorsal striatopallidal GABA pathway A series of reports indicate the usefulness of employing not only A2A antagonists but also mGluR5 antagonists and their combinations in the treatment of Parkinson's disease by increasing D2 signaling in A2A/mGluR5/D2 RMs located mainly in perisynaptic regions of glutamate and DA synapses on dendritic spines of striatopallidal GABA neurons and also by reducing glutamate release from corticostriatal glutamate terminals, where A2A and mGluR5 can interact synergistically in A2A/mGluR5/D2 RMs to increase glutamate release.

Development of agonists with combined A2A agonist/mGluR5 agonist properties for the treatment of schizophrenia based on multiple mGluR5/A2A/D2 receptor interactions in the ventral striatopallidal GABA pathway Experimental evidence suggests that synergistic interactions between A2A and mGluR5 receptors based on the existence of A2A/mGluR5 heteromeric complexes in postulated extrasynaptic mGluR5/A2A/D2 RMs of the ventral striatopallidal GABA neurons play a major role in increasing activity of this pathway and removing it from D2-mediated inhibition. This has been demonstrated in dual-probe microdialysis studies on this GABA system using co-perfusion with A2A and mGluR5 agonists.

Thus, we have advanced the proposal based on the DA hypothesis of schizophrenia that A2A agonists may be novel antipsychotic drugs by antagonizing D2 receptor signaling via an A2A/D2 intramembrane RRI in the ventral striatopallidal GABA system. The A2A agonist CGS21680 was shown to have an atypical antipsychotic profile by reducing the amphetamine and PCP-induced locomotor activity in doses failing to cause catalepsy. Furthermore, injection of the A2A agonist into the nucleus accumbens reversed prepulse inhibition by apomorphine, and combined treatment with subthreshold doses of a D2 antagonist and an A2A agonist led to activation of the ventral striatopallidal GABA pathway. The A2A agonist demonstrates antipsychotic-like activity in *Cebus apella* monkeys without production of extrapyramidal side effects, highlighting the development of novel A2A agonists as a strategy for treatment of schizophrenia based on their atypical antipsychotic profile.

Development of galanin receptor antagonists for treatment of depressive illness based on galR/5-HT1A receptor interactions In 1988, galanin was shown to reduce the affinity of 5-HT1A receptors in the ventral limbic cortex, providing the first evidence for the existence of antagonistic intramembrane GalR/5-HT 1A receptor interactions. In 1991, the reciprocal interaction was demonstrated with evidence that 5-HT1A receptor activation leads to an increase in the affinity of galanin receptors in various regions of the tel- and diencephalons. This increase in galanin recognition may be part of an intramembrane inhibitory feedback mechanism to reduce overactivation of 5-HT 1A signaling that occurs via the interface in a postulated GalR/5-HT 1A heteromeric complex. Also in 1991, the relevance of this RRI for depression was discussed in the context of the 5-HT hypothesis of depression, and the proposal was made that galanin receptor antagonists, by enhancing postjunctional 5-HT1A-mediated transmission in the forebrain, may represent novel antidepressant drugs. Galanin receptor antagonists may also produce antidepressant effects by blocking galanin receptors in the dorsal raphe, which inhibit the 5-HT releasing activity and firing of the ascending 5-HT pathways to the tel- and diencephalon. The galanin receptor on the HT cell bodies interacts with the 5-HT1A autoreceptor, but the functional outcome of this interaction remains to be clarified with regard to 5-HT1A autoreceptor signaling. An increased density of galanin receptor agonist binding sites has been found in the dorsal raphe in a genetic model of depression, leading to a feedback inhibition of galanin synthesis with reductions of galanin immunoreactivity in the dorsal raphe.

WT and VT in the Molecular Networks: Horizontal versus Vertical Molecular Networks

The RMs are located at the plasma membrane level and are part of horizontal molecular networks (HMNs) made by membrane-associated or membrane-integral proteins. Thus, HMNs represent the molecules located within a membrane microdomain, or associated with the plasma membrane microdomain, that are physically connected (WT) or communicate via a signal released and diffusing within the membrane (VT) (e.g., diacyl glycerol).

Vertical molecular networks (VMNs) go deep into the cytoplasm, often reaching into the nucleus to phosphorylate transcription factors. Here, WT is the transmission of signals along assembled protein modules that work as a wire. The best example is the Raf-MEK1/2-ERK1/2 cascade operating via protein–protein interactions to enter the nucleus and phosphorylate transcription factors such as Elk1. The main instance of VT in the VMN is the release of second messengers, especially calcium ions, leading to waves of calcium forming intracellularly.

HMNs are two-sided input/output networks – that is, they can receive signals from the extracellular as well as the intracellular environment of the cell and can give responses aimed at maintaining cell homeostasis, which affect both the extracellular and the intracellular environment of the cell. Thus, HMNs work as complex interfaces between the extracellular and intracellular environments and have been indicated as a target of electronic and chemical signals present at the local circuit level. Lipid rafts have emerged as specialized membrane platforms for HMNs. Our proposal of HMNs has similarities with Lisanti's proposal of signalosomes the view of the RM as an integrated input unit to the molecular networks, and with Bockaert's proposal of G-protein-coupled receptors (GPCRs) organized in a functional unit.

HMNs are formed by different classes of molecules, and some of them are targets for other cytoplasmic and extracellular molecules. Of obvious importance are those acting on the GPCRs (and hence on RMs) because they regulate the level of inputs to the cell biochemical machinery (i.e., to the VMNs). Among these are molecules controlling GPCR function (e.g., regulators of G-protein signaling), molecules controlling GPCR internalization and/or having scaffolding and adapter functions (e.g., β-arrestins and caveolins), or molecules controlling GPCR desensitization (e.g., GPCR kinases, which are key modulators of GPCR signaling can also interact with caveolins). These are some of the molecules that likely control the inputs that reach the VMNs and address the signal impinging on the cell only to some of the VMNs among all those that are potentially interconnected with the HMNs.

The Extracellular and Intracellular Molecular Networks: A GMN Enmeshes the Entire CNS

Molecular networks are composed mainly of proteins and carbohydrates, which fill up the ECS and the interior of the cells and interact with each other at the boundaries of these compartments, namely at the plasma membrane levels, to form a GMN. Thus, the GMN pervades the intra- as well as the extracellular environment of the entire CNS (Figure 4). From a functional standpoint, proteins are likely to be the most important building blocks for all the molecular networks. For those located in the ECS (the extracellular part of the GMN), carbohydrate-associated proteins also have an important scaffolding role.

A GMN is made by the extracellular matrix molecular net (the extracellular part of the GMN) and by the intracellular molecular nets. The intracellular part of the GMN communicates with the extracellular part of the GMN mainly via specialized 'windows' (e.g., the lipid rafts), where, for example, an integrin receptor permits a link between the extracellular matrix (ECM) network and the cytoskeleton, especially the microfilaments (twisted pairs of actin filaments). The extracellular part of the GMN is made by elements that can carry out a scaffolding and a computational role. Signals can be produced, elaborated, and/or released to be transmitted by the extracellular part of the GMN. Among these signals are several growth factors and hyaluronan and hyaluronan fragments.

Structure and Function of the Extracellular Part of the GMN

Hyaluronan is the backbone of several ECM molecular complexes and highly suited for space invasion. Furthermore, several ECM molecules tend to occupy a lot of space in view of the mutual repulsion of their numerous negatively charged residues. The hyaluronan molecule in solution can be viewed as a highly organized extended dynamic coil that can undergo sharp kinks, bends, and folds while maintaining distinctive hydrogen bond conformations. The characteristic chain reorganization time would be within a nanosecond timescale. This dynamic flexibility would give it excellent space-invading properties and, at the same time, promote very rapid recovery times after mechanical perturbation and distortion. Thus, hyaluronan is one of the fundamental blocks of

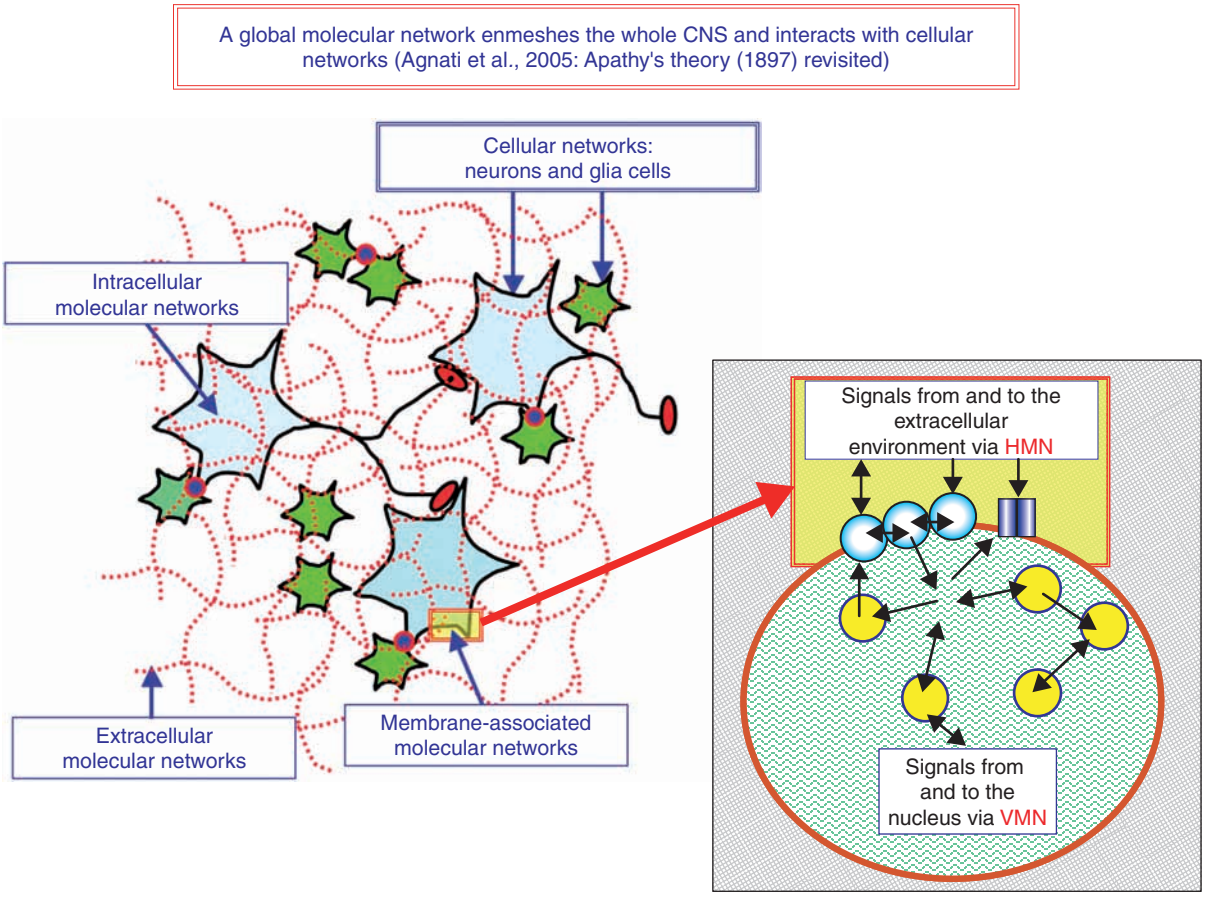

Figure 4 Illustration of the intracellular and intercellular molecular networks and how they can form a global molecular network (GMN) via interactions with the membrane-associated molecular networks (horizontal molecular networks).The GMN partially overlaps and interacts with the cellular networks composed of nerve and glial cells, giving a new model of the brain networks.

the extracellular part of the GMN which has been modeled in a simple way by Yamaguchi using the hyaluronan–lectican–tenascin R model of the brain ECM (Figure 5).

The extracellular part of the GMN, as well as the intracellular nets (microtubules, neurofilaments, and microfilaments) are made by building blocks that create, via protein–protein and protein–carbohydrate interactions, a high-order supramolecular structure. The extracellular part of the GMN has both a scaffolding role in enabling the appropriate location of the CNS components and a functional role in cooperating in the maintenance of the microenvironment around cells and in the creation of preferential diffusion pathways. The extracellular part of the GMN also affects the communication processes in the cellular circuits. Intercellular WT (i.e., synaptic contacts, but also gap junctions and specialized membrane juxtapositions) is regulated via the control of the cellular circuit geometry, as well as the control of the cell-to-cell connection efficacy. Intercellular VT (i.e., signal diffusion) is regulated via the control of the net tightness and the ECS medium fluidity, altering tortuosity in the diffusion pathways where VT signals released by a cell migrate in the ECS to reach the target cells (Figure 5).

In turn, the CCNs affect the GMN molecular composition and topology, both its extracellular part and its intracellular part. Thus, there is a complex interplay between the molecular and topological organization of the GMN and the intercellular WT and VT, and these interactions lead to optimal tuning of the integrative actions of the CNS.

A New Model of Brain Networks: Partially Overlapping Cellular and GMNs with a Nested Organization

According to Apathy's theory and as described in our own work, we have suggested the existence of a GMN enmeshing the entire CNS (Figure 4). Thus, the view of the brain as a system of neuronal (or also 'cellular') networks or even the view of the brain as a system of 'nested networks of networks'

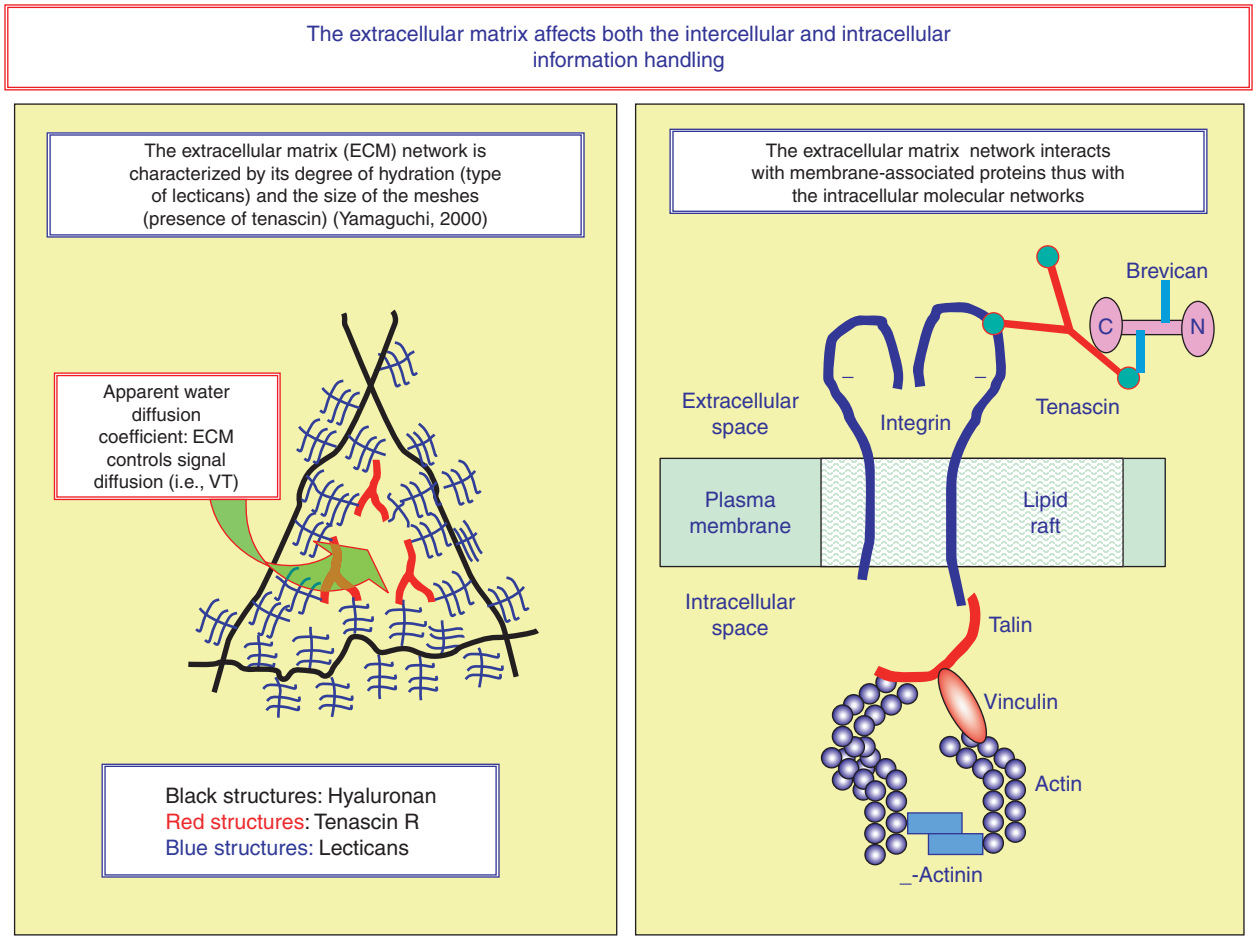

Figure 5 Illustration of the role of the extracellular matrix – built *inter alia* of hyaluronan, tenascin, and lecticans – in intercellular and intracellular information handling. The size of the meshes and the degree of hydration will control the diffusion pathways in the extracellular space and thus VT. The tenascin interaction with the integrin receptor illustrates how, via WT, the extracellular molecular network can interact with the intracellular molecular networks, in this case the actin component of the cytoskeleton.

organized as Russian dolls are no longer tenable. Thus, we propose a new model of brain networks, namely as a system of nested but partially overlapping networks. According to this proposal, the GMN is superimposed on the cellular networks, and these networks sometimes operate in parallel and sometimes in series. Communication processes within these networks and between them occur via VT and WT.

This model allows a different approach to brain function and thus to the physiology and pathology of the brain. In particular, learning processes can occur simply via changes in the communication channels between nodes – that is, between cells in the case of CCN or between macromolecules in the case of molecular networks as described by the 'Turing B-type unorganized machine.' Furthermore, 'temporal difference learning' can be transferred to the molecular level if we accept that also at this level VT and WT can be in operation.

In addition, besides the classical models of learning based on WT, it may be proposed that also VT and a proper mix of VT and WT can allow very effective learning processes. It should also be noted that according to our old RM hypothesis of the engram, learning at the molecular level (RM formation and consolidation) can be the result of a common operation between molecular and cellular networks, and in this process the extracellular molecular network can also contribute to the RM assemblage. Hence, future investigations likely will establish that a continuous interchange of information and reciprocal plastic adjustments between CCN and GMN occur during learning and that membrane-associated molecular networks (i.e., the HMN and especially the RM) have a special role in such a fundamental process, as hypothesized approximately 25 years ago.

Pathological and Therapeutic Implications of the Current View

This new model of the brain networks communicating via VT and WT opens up a new understanding of

brain physiology and pathology both in neurodegenerative diseases and in mental diseases such as schizophrenia, depression, and addiction, as well as in brain tumor development. The intramembrane RRIs taking place via heterodimers and RMs represent a new principle in molecular medicine allowing integration of signals already at the level of the plasma membrane. They open up new targets for treatment of receptor dysfunction known to occur *inter alia* in neurological and mental disorders.

See also: Aging: Extracellular Space; Brain Volume: Age-Related Changes; Cajal's Place in the History of Neuroscience; Neuron Doctrine: Historical Background; Regulation of Cell Volume in Neural Cells.

Further Reading

Agnati LF, Ferre S, Lluis C, Franco R, and Fuxe K (2003) Molecular mechanisms and therapeutical implications of intramembrane receptor/receptor interactions among heptahelical receptors with examples from the striatopallidal GABA neurons. *Pharmacological Reviews* 55: 509–550.

Agnati LF, Fuxe K, Nicholson C, and Sykova E (eds.) (2000) *Progress in Brain Research 125: Volume Transmission Revisited*. Amsterdam: Elsevier.

Agnati LF, Fuxe K, Zoli M, et al. (1986) A correlation analysis of the regional distribution of central enkephalin and beta-endorphin immunoreactive terminals and of opiate receptors in adult and old male rats. Evidence for the existence of two main types of communication in the central nervous system: The volume transmission and the wiring transmission. *Acta Physiologica Scandia* 128: 201–207.

Agnati LF, Tarakanov AO, Ferre S, Fuxe K, and Guidolin D (2005) Receptor–receptor interactions, receptor mosaics, and basic principles of molecular network organization: Possible implications for drug development. *Journal of Molecular Neuroscience* 26(2–3): 193–208.

Agnati LF, Zunarelli E, Genedani S, and Fuxe K (2006) On the existence of a global molecular network enmeshing the whole central nervous system: Physiological and pathological implications. *Current Protein & Peptide Science* 7: 3–15.

Bockaert J, Marin P, Dumuis A, and Fagni L (2003) The 'magic tail' of G protein-coupled receptors: An anchorage for functional protein networks. *FEBS Letters* 546: 65–72.

Descarries L and Mechawar N (2000) Ultrastructural evidence for diffuse transmission by monoamine and acetylcholine neurons of the central nervous system. *Progress in Brain Research* 125: 27–47.

Franco R, Casado V, Mallol J, et al. (2005) Dimer-based model for heptaspanning membrane receptors. *Trends in Biochemical Sciences* 30: 360–366.

Fuxe K and Agnati LF (eds.) (1991) *Volume Transmission in the Brain*. New York: Raven Press.

Fuxe K, Canals M, Torvinen M, et al. (2007) Intramembrane receptor–receptor interactions. A novel principle in molecular medicine. *Journal of Neural Transmission* 114: 49–75.

Fuxe K, Ferre S, Canals M, et al. (2005) Adenosine A2A and dopamine D2 heteromeric receptor complexes and their function. *Journal of Molecular Neurosciences* 26(2–3): 209–220.

Fuxe K, Rivera A, Jacobsen KX, et al. (2005) Dynamics of volume transmission in the brain. Focus on catecholamine and opioid peptide communication and the role of uncoupling protein 2. *Journal of Neural Transmission* 112: 65–76.

Gozes I (ed.) (2005) Wenner–Gren international symposium on receptor–receptor interactions among heptaspanning membrane receptors: From structure to function. *Journal of Molecular Neurosciences* 26(2–3).

Jansson A, Descarries L, Cornea-Hérbert V, et al. (2002) Transmitter–receptor mismatches in central dopamine, serotonin, and neuropeptide systems. Further evidence for volume transmission. In: Walz W (ed.) *The Neuronal Environment: Brain Homeostasis in Health and Disease*, pp. 83–101. Totowa, NJ: Humana Press.

Milligan G (2004) G protein-coupled receptor dimerization: Function and ligand pharmacology. *Molecular Pharmacology* 66: 1–7.

Sykova E (2005) Glia and volume transmission during physiological and pathological states. *Journal of Neural Transmission* 112: 137–147.

Terrillon S and Bouvier M (2004) Roles of G-protein-coupled receptor dimerization. *EMBO Reports* 5: 30–34.

Cells: 5-Hydroxytryptamine Receptors

E J Mylecharane, The University of Sydney, Sydney, NSW, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction: Background and Scope

5-Hydroxytryptamine (5-HT; enteramine; serotonin) has a long history in mammalian species studies. In the latter part of the nineteenth century, its existence was recognized as a vasoconstricting factor in serum, and its effects were able to be negated by passage through the lungs; its source was subsequently found to be platelets. In the 1930s–1940s, Erspamer and his colleagues in Italy knew it as the main hormone from enterochromaffin cells in the gastrointestinal tract mucosa, and named it enteramine. At the Cleveland Clinic in the United States, Rapport, Green, and Page isolated the serum vasoconstricting factor in 1948, and named it serotonin. Rapport determined its structure as an indolealkylamine, 5-hydroxytryptamine, in 1949 (Figure 1). Pharmacological studies in the early 1950s, by Reid and Rand in Australia, and in Erspamer's laboratory, soon confirmed that serotonin and enteramine were in fact 5-HT.

About 99% of the 5-HT in mammals is synthesized in enterochromaffin cells; approximately 90% is stored in the enterochromaffin cells, and 9% is taken up by platelets. The remaining 1% of the 5-HT in mammals is synthesized and stored in neurons. Recognition of 5-HT as an important neurotransmitter began with its discovery in the mammalian brain, by Twarog in 1953 in Cleveland, and in Gaddum's laboratory in 1954 in Britain. Most of the neuronal 5-HT is present in the central nervous system (CNS), where it has complex roles in the regulation of a host of functions, including sensory perception (e.g., pain, vision, audition), body temperature, neuroendocrine secretion, mood and emotional behavior, cognition, appetite, sexual activity, and sleep, as well as contributing to the regulation of the autonomic nervous system (ANS). A small fraction of the neuronal 5-HT is located in some peripheral parts of the ANS.

Receptors mediating 5-HT actions were first identified in Gaddum's laboratory in 1954, and were named D and M receptors. In the late 1970s, radiolabeled ligand binding sites for 5-HT in the brain were identified by Peroutka and Snyder, and named 5-HT_1 and 5-HT_2. D receptors seemed to correlate with 5-HT_2 binding sites, but M receptors and 5-HT_1 binding sites did not. The surge in development of a number of 5-HT receptor agonists and antagonists in the early 1980s led Bradley and colleagues to propose a rational 5-HT receptor classification scheme in 1986, comprising three types, 5-HT_1-like, 5-HT_2, and 5-HT_3 (the latter two types broadly encompassing the D and M receptors, respectively). The possibility of at least three 5-HT_1-like subtypes was recognized.

In 1994, Hoyer and colleagues proposed an expanded 5-HT receptor classification incorporating seven types (three of which included subtypes), based on many newly developed selective 5-HT receptor agonists and antagonists, improved understanding of receptor-coupling ('transduction') mechanisms, and rapid progress in using recombinant molecular biology techniques to determine receptor structure (as protein amino acid sequences). This classification, one of the first adopted by the International Union of Pharmacology (IUPHAR) in accordance with agreed criteria and principles, remains intact (subject to some subsequent minor adjustments – e.g., human primacy where species differences occur, and status based on further evidence): 5-HT_{1A}, 5-HT_{1B}, 5-HT_{1D}, 5-ht_{1e}, 5-HT_{1F}; 5-HT_{2A}, 5-HT_{2B}, 5-HT_{2C}; 5-HT_3; 5-HT_4; 5-ht_{5a}, 5-ht_{5b}; 5-HT_6; and 5-HT_7 (lower case letters are used for recombinant receptors not yet found in a native tissue, or not yet pharmacologically evaluated). The 5-HT_3 receptor is a ligand-gated ion channel receptor. All the others are G-protein-coupled receptors with seven transmembrane-spanning regions. The current detailed IUPHAR classification is accessible via http://www.iuphar-db.org/GPCR/index.html.

This article provides a pharmacological and physiological survey of the 5-HT receptors located on the ANS neurons: sympathetic and parasympathetic neurons; the enteric nervous system (ENS; an important part of the ANS); and the sensory afferent and CNS neurons contributing to ANS regulation. 5-HT receptors on autonomic neuroeffector cells are also surveyed. In considering the extensive 5-HT receptor-mediated actions in the ANS, an ambiguity is readily apparent – there is relatively little 5-HT neurotransmission in the ANS, its main location being within the ENS (which comprises sensory neurons, interneurons, and motor neurons, and functions as a complex and independent autonomic regulatory system in the gastrointestinal tract). A few vasculatures are innervated by 5-HT-containing neurons, and 5-HT is present in some sympathetic ganglia within small intensely fluorescent cells (which may function as interneurons). 5-HT also serves as a neurotransmitter in some CNS pathways regulating the ANS (as CNS interneurons, or as direct innervation

to preganglionic ANS neurons). Nevertheless, the peripheral components of the ANS are able to be extensively affected by 5-HT released from enterochromaffin cells and platelets, acting as an autacoid or paracrine mediator, with physiological regulatory or homoeostatic roles, and defense or pathological roles.

5-HT receptor agonists and antagonists are used therapeutically in relatively few disorders related to ANS function, mainly for gastrointestinal disturbances (e.g., gastroesophageal reflux, gastroparesis, irritable bowel syndrome, and vomiting and nausea in cancer chemotherapy and radiotherapy), the cranial vascular disturbances of migraine, and control of 5-HT-induced systemic disturbances in carcinoid syndrome (enterochromaffin cell tumors). In contrast, drugs affecting 5-HT actions (i.e., 5-HT receptor agonists and antagonists, and drugs that alter 5-HT synthesis, storage, release, reuptake, and metabolic degradation) have many therapeutic uses for disorders such as migraine, depression, anxiety, schizophrenia, cognitive dysfunction, drug dependence and abuse, eating disorders, and sleep disturbances, largely on the basis of their CNS effects. Adverse effects of these drugs include many ANS disturbances in cardiovascular, bronchial, gastrointestinal, urogenital, and body temperature regulation; thus, an understanding of 5-HT actions on the ANS may contribute to better therapeutic outcomes. An important aspect is acute 5-HT toxicity (serotonin syndrome) due to overdosage or combinations of 5-HT activity-enhancing therapeutic and illicit drugs. Four principal features of serotonin syndrome are diarrhea, hypertension or hypotension, profuse sweating, and rapid hyperthermia, clearly involving major ANS dysfunction (the others are confusion, agitation, tremor, hyperreflexia, and clonus); hyperthermia is the most likely to lead to a fatal outcome.

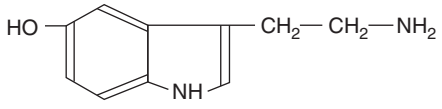

Figure 1 5-Hydroxytryptamine.

Sympathetic Neuron 5-HT Receptors

Preganglionic

Sympathetic preganglionic neuron cell bodies are located in the CNS, in thoracolumbar spinal nuclei. Their relatively short axonal projections innervate the paired paravertebral sympathetic ganglia, located bilaterally along the vertebral column (including the superior and middle cervical ganglia, the stellate ganglia, the thoracic ganglia, the lumbar ganglia, and some sacral ganglia), and the midline prevertebral sympathetic ganglia (celiac, aortico-renal, and the superior and inferior mesenteric ganglia). Some longer axonal projections innervate terminal sympathetic ganglia closer to their target structures (e.g., the adrenal medulla and the hypogastric plexus). Table 1 summarizes 5-HT receptor-mediated actions in sympathetic preganglionic neurons.

Sympathetic preganglionic cell bodies are regulated by various CNS neurotransmitters, including 5-HT, from supraspinal CNS presympathetic inputs,

Table 1 5-HT receptor-mediated effects on sympathetic neurons[a]

Receptor location	*5-HT receptor type/subtype*												
	1[b]	*1A*	*1B*	*1D*	*1F*	*2*[b]	*2A*	*2B*	*2C*	*3*	*4*	*6*	*7*
Preganglionic													
Cell body	E?					E?							E?
Axon										E			
Terminal	I?									E?			
Postganglionic													
Cell body										E			
Axon										E			
Terminal													
Heart[c]	I		I	I						E			
Vasculature[d]		I	I	I									
Vas deferens	I												
Ovarian follicle	I												

[a]E, neuroexcitation; I, neuroinhibition; ?, possible attribution based on limited evidence.
[b]Receptor subtype not determined.
[c]Neuroinhibition in dogs and guinea pigs is mediated by a 5-HT_1 receptor (subtype not determined); in rats, both rodent 5-HT_{1B} receptors and 5-HT_{1D} receptors contribute to neuroinhibition.
[d]Neuroinhibition mediated by 5-HT_{1B} receptors in various species (including rats), and also 5-HT_{1A} and 5-HT_{1D} receptors only in rats.

including the rostral and caudal ventrolateral medulla (RVLM and CVLM, respectively; see later). Electrophysiological studies have consistently demonstrated neuroexcitatory effects of 5-HT; however, 5-HT-induced neuroinhibition has been reported in a small percentage of cells in some studies. The 5-HT receptors mediating neuroexcitation have not yet been defined with appropriate selective agonists and antagonists; some studies suggest 5-HT_2 receptor-mediated neuroexcitation, but 5-HT_1 and 5-HT_7 receptors cannot be ruled out. No information is available regarding 5-HT receptor-mediated neuroinhibition in these cells.

5-HT_3 receptor-mediated neuroexcitation has been directly identified in sympathetic preganglionic axons. A 5-HT_3 receptor may also mediate neuroexcitation in the terminals of these preganglionic neurons, but this remains to be confirmed with selective antagonists. Presynaptic neuroinhibition resulting in decreased acetylcholine release from sympathetic preganglionic terminals is more difficult to investigate, because changes in postsynaptic response alone are insufficient to distinguish presynaptic from postsynaptic sites of action. 5-HT has a presynaptic neuroinhibitory effect on acetylcholine release in cervical, stellate, and splanchnic ganglia, but the receptor involved has not yet been defined; a 5-HT_1 receptor subtype (possibly 5-HT_{1B} or 5-HT_{1D}) is most likely.

Postganglionic

Table 1 also summarizes 5-HT receptor-mediated actions in sympathetic preganglionic neurons, which have relatively long axonal projections to their target cells, tissues, and organs. Electrophysiological studies have confirmed that a 5-HT_3 receptor mediates 5-HT-induced neuroexcitation, resulting in rapid depolarization, in superior cervical, celiac, and inferior mesenteric ganglia cell bodies. 5-HT also elicits a slow depolarization in a substantial proportion of cell bodies tested in the celiac and inferior mesenteric ganglia. The 5-HT receptor responsible for this slow depolarization is not a 5-HT_3 receptor; its identity remains unknown, but it has been suggested that it might be the so-called orphan 5-HT_{1P} receptor, which is located in the ENS (see later) and has not yet been incorporated into the IUPHAR classification of 5-HT receptors. In addition, 5-HT elicits neuroinhibition in cervical ganglion cell bodies, via an as yet unidentified 5-HT receptor; a 5-HT_3 receptor is not involved.

5-HT_3 receptor-mediated neuroexcitation has been identified in sympathetic postganglionic axons (recorded as depolarization in superior cervical ganglia) and in terminals (recorded as increased noradrenaline release from terminals in isolated heart preparations, accompanied by increased heart rate). 5-HT-induced presynaptic neuroinhibition resulting in decreased noradrenaline release from sympathetic postganglionic terminals in vascular tissues was first demonstrated by McGrath in the mid-1970s, using dog saphenous vein preparations, and recording decreases in smooth muscle contraction and radiolabeled noradrenaline release, in response to electrical stimulation. The role of presynaptic neuroinhibition of noradrenaline release from vascular sympathetic postganglionic terminals in 5-HT-induced vasodilation *in vivo* was clarified in the mid-1980s in the author's laboratory. Although 5-HT-induced direct vascular smooth muscle relaxation (see later) and release of endothelium-derived relaxant factor (EDRF; generally nitric oxide) contribute to *in vivo* vasodilation, presynaptic neuroinhibition of vascular sympathetic tone is a major factor in some regional circulations (e.g., in dog femoral and common carotid arterial circulations). In these studies, vasoconstrictor tone was established with normal vascular sympathetic tone, or with a direct vasoconstrictor agent after abolition of vascular sympathetic tone; the 5-HT-induced vasodilation was entirely dependent on vascular sympathetic tone. 5-HT-induced presynaptic neuroinhibition of vasoconstrictor responses to electrical stimulation of vascular sympathetic postganglionic terminals *in vivo* has also been demonstrated in experiments analogous to those *in vitro*. Other vasculatures in which presynaptic neuroinhibition of sympathetic postganglionic terminals has been identified include the vena cava, and coronary arterial, renal, and systemic (in pithed rats) circulations. A 5-HT_{1B} receptor is generally responsible for these effects. In rats, the rodent 5-HT_{1B} receptor is involved; this receptor is a species homolog of the human 5-HT_{1B} receptor, although its pharmacology is very different from that of the human homolog. There is evidence that in rats, both 5-HT_{1A} and 5-HT_{1D} receptors also contribute to the 5-HT-induced presynaptic neuroinhibition at vascular sympathetic postganglionic terminals. 5-HT-induced presynaptic neuroinhibition of noradrenaline release from sympathetic postganglionic terminals has also been demonstrated in some nonvascular tissues: dog, guinea pig, and rat heart; rat vas deferens; and bovine ovarian follicle. 5-HT_1 receptors (subtypes not yet identified) are responsible in all except rat heart, in which the rodent 5-HT_{1B} receptor and the 5-HT_{1D} receptor contribute to this presynaptic neuroinhibition.

Parasympathetic Neuron 5-HT Receptors

Preganglionic

Parasympathetic preganglionic neurons have relatively long axons, which project to ganglia located close to their target organs and tissues. Their cell

Table 2 5-HT receptor-mediated effects on parasympathetic neurons[a]

Receptor location	*5-HT receptor type/subtype*												
	1[b]	*1A*	*1B*	*1D*	*1F*	*2*[b]	*2A*	*2B*	*2C*	*3*	*4*	*6*	*7*
Preganglionic													
Cell body													
Vagal (heart)[c, d]		I?											
Sacral segment (bladder)[c]													
Sacral segment (erectile)									E				
Terminal													
Oculomotor nerve (eye)											I?		
Pelvic nerve (bladder)		I											
Postganglionic													
Cell body													
Ciliary ganglion (eye)										E			
Vesical ganglion (bladder)										E			
Terminal													
Vagal (heart)										E			
Vagal (bronchi)													
Vesical ganglion (bladder)[e]		I	E ?								E		E

[a]E, neuroexcitation; I, neuroinhibition; ?, possible attribution based on limited evidence.
[b]Receptor subtype not determined.
[c]Neuroexcitation mediated by 5-HT_{1A} receptors is probably indirect.
[d]Neuroexcitation mediated by 5-HT_3 receptors is probably indirect.
[e]Neuroexcitation possibly mediated by rodent 5-HT_{1B} receptors only in mice.

bodies are located in the CNS, in midbrain, brain stem, and sacral spinal nuclei. No parasympathetic preganglionic neuron roles for 5-HT receptors have yet been found in the facial and glossopharyngeal cranial nerves, or in vagal nerves projecting to lungs, kidneys, and gastrointestinal tract as far as the beginning of the distal colon. As summarized in Table 2, the main nuclei in which 5-HT receptors may have a parasympathetic preganglionic neuron role include the dorsal motor vagus (DMV) and the nucleus ambiguus in the medulla (providing vagal nerve outflow to the heart), and sacral spinal nuclei (providing pelvic nerve outflow to the urinary bladder and erectile tissue). DMV and nucleus ambiguus neuroinhibition may be 5-HT_{1A} receptor mediated, as it can be directly elicited by selective 5-HT_{1A} receptor agonists, but this remains to be confirmed with investigations involving selective antagonists and the elimination of indirect mechanisms. 5-HT_{1A} receptor mediated and 5-HT_3 receptor-mediated neuroexcitation of these CNS nuclei cell bodies are probably indirect actions, because the 5-HT-induced neuroexcitation is blocked by both 5-HT_3 and glutamate receptor antagonists in the DMV, and the invariable neuroinhibition mediated by 5-HT_{1A} receptors may result in disinhibition of neuroinhibitory input of γ-aminobutyric acid (GABA) neurons regulating DMV and nucleus ambiguus neurons. The only confirmed 5-HT receptor-mediated action on cell bodies in these nuclei is 5-HT_{2C} receptor-mediated neuroexcitation in sacral spinal nuclei, producing increased outflow to erectile tissue. Parasympathetic preganglionic regulation of urinary bladder function is known to involve 5-HT_{1A} receptor-mediated neuroinhibition, and 5-HT_{2C}, 5-HT_3, and possibly 5-HT_7 receptor-mediated neuroexcitation, but the CNS locations of these supraspinal and spinal nuclei are unknown (see later). A 5-HT_{1A} receptor-mediated neuroexcitation in sacral spinal segments is probably indirect, similar to that in DMV and nucleus ambiguus neurons.

At parasympathetic preganglionic terminals of the oculomotor cranial nerve in the ciliary ganglion, 5-HT decreases acetylcholine release via presynaptic neuroinhibition, possibly mediated by a 5-HT_4 receptor, but this remains to be confirmed with selective 5-HT_4 receptor agonists and antagonists. In the vesical ganglion, which regulates urinary bladder function, 5-HT_{1A} receptors mediate a presynaptic neuroinhibition of acetylcholine release from the preganglionic neuron terminals of the pelvic nerve.

Postganglionic

Table 2 also summarizes 5-HT receptor-mediated actions on parasympathetic postganglionic neurons, except for the very few vagal postganglionic neurons in the gastrointestinal tract myenteric plexus; these postganglionic neurons are regarded as an integral part of the ENS (see later). Electrophysiological studies can determine the location of receptors on cell

bodies, or on terminals remaining in cardiac and smooth muscle preparations stripped of neuronal plexuses, but this is precluded when postganglionic neuron function is monitored indirectly (e.g., via neurotransmitter efflux or neuroeffector cell response) because the postganglionic axons are very short.

5-HT_3 receptor-mediated neuroexcitation has been confirmed in oculomotor and pelvic nerve parasympathetic postganglionic cell bodies in the ciliary and vesical ganglia, respectively. In support of the latter conclusion, indirect studies show that 5-HT_3 receptor-mediated neuroexcitation of pelvic nerve parasympathetic postganglionic neurons increases acetylcholine release and urinary bladder smooth muscle contraction.

In cardiac parasympathetic postganglionic terminals, 5-HT_3 receptor-mediated neuroexcitation is responsible for increased acetylcholine release and a consequent decrease in heart rate. 5-HT enhances electrically stimulated bronchial smooth muscle contraction by a mainly presynaptic neuroexcitation, resulting in increased acetylcholine release from vagal parasympathetic postganglionic terminals, via a 5-HT_3 receptor. Studies on pelvic parasympathetic postganglionic terminals in human urinary bladder detrusor smooth muscle preparations have confirmed that 5-HT_4 and 5-HT_7 receptor-mediated presynaptic neuroexcitation results in increased acetylcholine release, and 5-HT_{1A} receptor-mediated neuroinhibition decreases acetylcholine release. 5-HT_3 receptors are not present on these terminals (in contrast to their presence on the postganglionic cell bodies). In mouse urinary bladder detrusor muscle, a rodent 5-HT_{1B} receptor-mediated neuroexcitation may be responsible for increased acetylcholine release from pelvic parasympathetic postganglionic terminals, but confirmation with more selective agonists and antagonists is needed.

ENS Neuron 5-HT Receptors

An understanding of ENS neuroanatomy and physiology provides some insight into the ENS roles of 5-HT and 5-HT receptors. Most of the peripheral neuronal 5-HT is in ENS neurons. ENS 5-HT interacts with 5-HT receptors on particular ENS neurons: submucosal and myenteric intrinsic primary afferent neurons (cholinergic, with cotransmitters calcitonin gene-related peptide in submucosal cells, and substance P in myenteric cells), which regulate gastrointestinal smooth muscle peristaltic and secretory reflexes via other interneurons; and extrinsic primary afferent neurons (which provide afferent information to the CNS via vagal and spinal afferent neurons, with cell bodies outside the gastrointestinal tract). Little is known about 5-HT receptor-mediated regulation of the very few parasympathetic postganglionic cholinergic motor neurons in the ENS which innervate smooth muscle (~1000 in the entire human small intestine, compared with approximately 10^8 ENS neurons). 5-HT from enterochromaffin cells in the mucosa can also interact with ENS 5-HT receptors.

Table 3 summarizes the 5-HT receptor-mediated actions in ENS neurons; the 5-HT_{1A} receptor-mediated action is neuroinhibitory, whereas those mediated by 5-HT_3, 5-HT_4, and 5-HT_7 receptors, as well as the so-called orphan 5-HT_{1P} receptor, are

Table 3 5-HT receptor-mediated effects on ENS neurons[a]

Receptor location	*5-HT receptor type/subtype*													
	1[b]	*1A*	*1B*	*1D*	*1F*	*1P*[c]	*2*[b]	*2A*	*2B*	*2C*	*3*	*4*	*6*	*7*
Submucosal intrinsic primary afferent														
Cell body						E					E			
Terminal												E		
Myenteric intrinsic primary afferent														
Cell body		I				E					E			E?
Terminal		I									E	E		
Extrinsic primary afferent														
Afferent axon														
Vagal											E			
Spinal											E			
Cell body											E			

[a]E, neuroexcitation; I, neuroinhibition; ?, possible attribution based on limited evidence.
[b]Receptor subtype not determined.
[c]The so-called orphan 5-HT_{1P} receptor, found in the gastrointestinal tract but not in the CNS, has not yet been incorporated into the IUPHAR classification of 5-HT receptors; some hydroxylated indalpines are selective agonists, and the dipeptide *N*-acetyl-5-hydroxytryptophyl-5-hydroxytryptophan amide is a selective antagonist.

all neuroexcitatory. Electrophysiological studies with selective 5-HT receptor agonists and antagonists have confirmed all of the 5-HT receptor types and subtypes responsible for these ENS actions, except for the possible involvement of 5-HT_7 receptors in myenteric intrinsic primary afferent cell neuroexcitation.

5-HT_{1A} receptor-mediated neuroinhibition has been found only in myenteric intrinsic primary afferent neurons. 5-HT_3 receptor-mediated neuroexcitation elicits fast depolarization in submucosal and myenteric intrinsic primary afferent cell bodies, and in myenteric terminals. Gastrointestinal stimuli (via mechanosensitive and chemosensitive receptors) activate 5-HT_3 receptor-mediated neuroexcitation in the vagal and spinal axons of extrinsic primary afferent neurons, and 5-HT_3 receptor-mediated neuroexcitation can also be elicited in the vagal afferent cell bodies in nodose ganglia. The presynaptic 5-HT_4 receptor-mediated neuroexcitation in submucosal and myenteric intrinsic primary afferent terminals enhances fast depolarizations and increased release of acetylcholine and its cotransmitters. 5-HT_{1P} receptor-mediated neuroexcitation produces slow depolarization in submucosal and myenteric primary afferent cell bodies. Recent studies in guinea pigs have shown an extensive presence of 5-HT_7 immunoreactivity in myenteric cells. A selective 5-HT_7 receptor antagonist enhanced peristaltic activity in these studies; the blockade of a 5-HT_7 receptor-mediated neuroexcitatory action of 5-HT in myenteric cells might allow the normal release of nitric oxide from these cells to make a greater contribution to gastrointestinal smooth muscle relaxation (in tandem with 5-HT_7 receptor-mediated smooth muscle relaxation; see later).

ANS Sensory Afferent Neuron 5-HT Receptors

In addition to the ENS extrinsic primary afferent neurons; see earlier, the other important ANS afferent neuron inputs regulated by 5-HT receptors are the heart and lung vagal afferents and the carotid body glossopharyngeal afferents (i.e., parasympathetic afferents, with cell bodies in nodose-petrosal ganglia). The so-called sympathetic afferent neurons convey ANS information to the CNS via sympathetic nerve trunks, dorsal roots, and the spinal cord, but little is known about the role of 5-HT receptors in these afferents. 5-HT_3 receptors are the principal 5-HT receptors with a confirmed role in ANS sensory afferent neurons, as summarized in Table 4.

A 5-HT_{2A} receptor-mediated neuroexcitation has been identified in rat cervical vagal afferents. Intravenous injection of 5-HT_3 receptor agonists produces neuroexcitation in unmyelinated vagal afferents in the heart (in the ventricular walls), which then elicits a rapid-onset decrease in heart rate and blood pressure (the so-called Bezold–Jarisch-like effect that mimics the reflex response to increased left ventricular or atrial pressure). In the lungs, 5-HT_3 receptor-mediated neuroexcitation of unmyelinated vagal afferents in collagenous alveolar tissue produces a protective nociceptive reflex which includes ANS disturbances (bronchoconstriction, and decreased heart rate and blood pressure), as well as rapid shallow breathing, dyspnea, apnea, and pain. 5-HT_3 receptor-mediated neuroexcitation of carotid body chemoreceptors results in reflex vasoconstriction, adrenal medulla catecholamine release, increased blood pressure, and

Table 4 5-HT receptor-mediated effects on ANS sensory afferent neurons[a]

Receptor location	*5-HT receptor type/subtype*												
	1[b]	*1A*	*1B*	*1D*	*1F*	*2*[b]	*2A*	*2B*	*2C*	*3*	*4*	*6*	*7*
Cervical vagal afferent													
Axon							E						
Heart													
Axon (unmyelinated)													
Vagal										E			
Lungs													
Axon (unmyelinated)													
Vagal										E			
Carotid body arterial chemoreceptors													
Axon													
Glossopharyngeal							E?			E			
Nodose-petrosal ganglia													
Cell body										E			

[a]E, neuroexcitation; I, neuroinhibition; ?, possible attribution based on limited evidence.
[b]Receptor subtype not determined.

decreased heart rate (similar to the chemoreceptor reflex response to decreased blood flow, low O_2 levels, high CO_2 levels, and acidemia); a 5-HT_{2A} receptor-mediated neuroexcitation may also contribute to the activation of this carotid chemoreceptor response. Neuroexcitatory 5-HT_3 receptors mediating a fast depolarization have been identified in nodose-petrosal ganglia cell bodies.

5-HT Receptors on CNS Neurons Regulating the ANS

Most of the vagal and glossopharyngeal sensory afferent neuron input from the abdominal and thoracic viscera projects to the nucleus tractus solitarius (NTS) in the medulla; some vagal afferents terminate in the DMV, the area postrema, and the supramedullary reticular formation. Ascending spinal sympathetic and lumbosacral parasympathetic afferents also project to the NTS. Sensory input is relayed by the NTS to medullary nuclei, which regulate efferent ANS pathways (the RVLM and CVLM, the DMV, and the nucleus ambiguus). The NTS also innervates medullary raphe nuclei (the main location of CNS 5-HT neurons). Supramedullary contributions to central ANS regulation include those from the reticular formation, the hypothalamus, and the thalamus. Table 5 summarizes 5-HT receptor-mediated actions in the main regulatory nuclei; the roles of 5-HT in the other ANS-regulating CNS regions are yet to be explored.

Little is known about the role of 5-HT in the NTS. 5-HT_3 receptor-mediated neuroexcitation releases NTS glutamate (neuroexcitatory), but the role of these 5-HT_3 receptors in regulating ANS reflexes (e.g., baroreceptor, chemoreceptor, Bezold–Jarisch) remains uncertain, because the outcomes are mainly assessed indirectly (i.e., as ANS sympathetic neuron firing, or neuroeffector events such as cardiovascular parameters, rather than neuronal activity in complex regulatory nuclei). A 5-HT_3 receptor-mediated presynaptic release of glutamate and GABA (neuroinhibitory) from vagal afferent terminals in the NTS may indirectly activate or inhibit particular NTS interneurons or second-order neurons projecting to ANS-regulating medullary nuclei.

In studies utilizing direct drug administration, 5-HT_{1A} receptor-mediated effects on RVLM neurons (which innervate thoracolumbar sympathetic preganglionic neurons) are involved in inhibition of vascular sympathetic tone in splanchnic, renal, and lumbar sympathetic neurons, but cardiac sympathetic tone is unaffected. Whether this inhibitory effect on RVLM neurons is a direct neuroinhibition, or an

Table 5 5-HT receptor-mediated effects on CNS neurons regulating the ANS[a]

Receptor location	*5-HT receptor type/subtype*												
	1[b]	*1A*	*1B*	*1D*	*1F*	*2*[b]	*2A*	*2B*	*2C*	*3*	*4*	*6*	*7*
NTS[c]													
Cell body										E?			
RVLM[d]													
Cell body		I?				E?							
CVLM[d]													
Cell body						E?							
DMV[e]													
Cell body													
Vagal (heart)[f, g]		I?											
Nucleus ambiguus[e]													
Cell body													
Vagal (heart)[f]		I?											
Spinal nuclei[e]													
Cell body													
Sacral segment (bladder)[f]													
Sacral segment (erectile)									E				

[a]E, neuroexcitation; I, neuroinhibition; ?, possible attribution based on limited evidence; NTS, nucleus tractus solitarius; RVLM, rostral ventrolateral medulla; CVLM, caudal ventrolateral medulla; DMV, dorsal motor vagus nucleus.
[b]Receptor subtype not determined.
[c]Regulating RVLM, CVLM, DMV, and nucleus ambiguus.
[d]Regulating sympathetic preganglionic neurons.
[e]Contains vagal parasympathetic preganglionic neuron cell bodies; see Table 2.
[f]Neuroexcitation mediated by 5-HT_{1A} receptors is probably indirect.
[g]Neuroexcitation mediated by 5-HT_3 receptors is probably indirect.

indirect effect via neuroinhibition of an excitatory interneuron synapsing with RVLM neurons, remains to be determined. 5-HT_{2A} receptors mediate CNS neuroexcitation, and intracerebroventricular injection of 5-HT_{2A} receptor agonists increases cardiac and vascular sympathetic tone. In direct administration studies in the RVLM, however, there is a 5-HT_{2A} receptor-mediated increase in vascular, but not cardiac, sympathetic tone; thus, the CVLM may be responsible for regulation of cardiac sympathetic tone.

As noted earlier, cell bodies of vagal and sacral segment parasympathetic preganglionic neurons are, respectively, located in the DMV and the nucleus ambiguus (innervating the heart), and in spinal nuclei (innervating urinary bladder, and erectile tissue). The relevant information on 5-HT receptor-mediated actions in these preganglionic cell bodies is included in Table 2, and also in Table 5 for convenience.

Numerous 5-HT receptor-mediated actions contribute to the CNS regulation of ANS functions in micturition. Mechanisms involving 5-HT_{1A} receptor-mediated neuroinhibition, and 5-HT_{2C}, 5-HT_3, and possibly 5-HT_7 receptor-mediated neuroexcitation, have been identified, but the specific CNS nuclei involved have not yet been clearly confirmed. The study of these mechanisms has been limited to comparisons between supraspinal and intrathecal drug administration, so that sites in the brain can be distinguished from sacral spinal sites. These studies have not utilized direct drug administration to particular CNS nuclei; intracerebroventricular administration has often been used, but this is not sufficient to identify the CNS nuclei involved.

Autonomic Neuroeffector Cell 5-HT Receptors

Table 6 summarizes 5-HT receptor-mediated actions on neuroeffector cells associated with the ANS. The only ANS neuronal source of 5-HT available to act on autonomic neuroeffector cells is limited to 5-HT-containing neurons in sympathetic nerves innervating some cerebral, basilar, vertebral, and mesenteric blood vessels. Enterochromaffin cell 5-HT activates ENS neurons (see earlier), but there is no evidence that 5-HT-containing ENS interneurons act directly on gastrointestinal smooth muscle. 5-HT receptors mediate EDRF release from endothelial cells in many vasculatures, but there is no conclusive evidence that 5-HT releases EDRF in association with ANS-induced vasodilation in the few vasculatures with cholinergic postganglionic innervation (erectile tissue, salivary, and sweat glands).

5-HT_{2A} receptors on ANS neuroeffector cells are the most widely distributed, mediating smooth muscle contraction throughout the vascular system, and in bronchial, intestinal, urinary bladder, and uterine smooth muscle, as well as increasing atrial rate and contractile force in rats (but not other species). Contraction mediated by 5-HT_{2B} receptors is limited to stomach smooth muscle. 5-HT_{1B} receptor-mediated smooth muscle contraction is principally confined to cranial and coronary arteries. Increases in cardiac muscle rate and contractile force are mediated by 5-HT_4 receptors (humans, pigs), and by 5-HT_3 and 5-HT_7 receptors in other species. 5-HT_4 receptor-mediated smooth muscle relaxation occurs in particular gastrointestinal tract regions and in pulmonary veins; 5-HT_4

Table 6 5-HT receptor-mediated effects on autonomic neuroeffector cells[a]

Receptor location	*5-HT receptor type/subtype*												
	1[b]	*1A*	*1B*	*1D*	*1F*	*2*[b]	*2A*	*2B*	*2C*	*3*	*4*	*6*	*7*
Cardiac muscle													
Atrial rate and force[c]							+			+	+		+
Smooth muscle													
Vascular[d]			C				C				R		R
Bronchial							C						
Gastrointestinal[e]							C	C			C, R		R
Bladder							C						
Uterine							C						

[a]+, increased; C, contraction; R, relaxation.
[b]Receptor subtype not determined.
[c]Species differences: 5-HT_{2A} in rat; 5-HT_3 in dog, guinea pig; 5-HT_4 in human pig; 5-HT_7 in cat.
[d]Contraction mediated by 5-HT_{1B} receptors in cranial, coronary, pulmonary and umbilical arteries, and saphenous vein, and by 5-HT_{2A} receptors widely distributed in arterial and venous circulation; relaxation mediated by 5-HT_4 receptors in pulmonary vein, and by 5-HT_7 receptors widely distributed in arterial and venous circulation.
[e]Regional differences: contraction mediated by 5-HT_{2A} receptors in intestine, 5-HT_{2B} in stomach, and 5-HT_4 in esophageal sphincter, ileum, and colon (longitudinal); relaxation mediated by 5-HT_4 receptors in esophagus and colon (circular), and by 5-HT_7 in intestine.

receptors contract smooth muscle in other parts of the gastrointestinal tract. 5-HT_7 receptors mediate vascular and intestinal smooth muscle relaxation.

Conclusion

Almost all of the functional 5-HT receptor types and subtypes in the IUPHAR 5-HT receptor classification are known to be present on ANS neurons, the exceptions being 5-HT_{1F}, 5-HT_{2B}, and 5-HT_6 receptors. The most widely distributed receptors are the neuroinhibitory 5-HT_{1A} and neuroexcitatory 5-HT_3 receptors, located on sympathetic, parasympathetic, and ENS neurons; neuroexcitatory 5-HT_3 receptors are also present on ANS sensory afferent neurons. Neuroinhibitory 5-HT_{1B} and 5-HT_{1D} receptors are confined to sympathetic postganglionic neurons; 5-HT_{2A} and 5-HT_{2C} neuroexcitatory receptors are found only on ANS sensory afferent and parasympathetic preganglionic neurons, respectively; neuroexcitatory 5-HT_4 receptors are present on parasympathetic postganglionic and ENS neurons; and neuroexcitatory 5-HT_7 receptors are located on parasympathetic postganglionic neurons. 5-HT-induced effects in central ANS regulatory neurons may be mediated by 5-HT_{1A}, 5-HT_2, and 5-HT_3 receptors, but are not yet confirmed. 5-HT_{2A} receptors are the most widely distributed receptors on ANS cardiac and smooth muscle neuroeffector cells. 5-HT_4 and 5-HT_7 receptors are present on cardiac, vascular, and gastrointestinal neuroeffector cells. The remaining locations are on vascular (5-HT_{1B}), gastrointestinal (5-HT_{2B}), and cardiac (5-HT_3) cells. The neuroexcitatory orphan 5-HT_{1P} receptor is confined to the ENS.

The principal therapeutic usage of 5-HT receptor agonists and antagonists in disorders related to ANS function is in gastrointestinal disorders, based mainly on 5-HT_4 receptor-mediated neuroexcitation in the ENS, and gastrointestinal smooth muscle contraction and relaxation, as well as 5-HT_3 receptor-mediated neuroexcitation in the ENS. Metoclopramide, a 5-HT_3 receptor antagonist and 5-HT_4 receptor agonist, is effective in the treatment of nausea and vomiting, gastroparesis, and gastroesophageal reflux disease. 5-HT_4 receptor partial agonists (cisapride, tegaserod) are effective in gastroparesis and irritable bowel syndrome, respectively. 5-HT_3 receptor antagonists (e.g., granisetron, tropisetron, ondansetron, dolasetron) are highly effective in controlling severe nausea and vomiting produced by cancer chemotherapy and radiotherapy, as well as postoperative nausea and vomiting.

Therapeutic usage of 5-$HT_{1B/1D}$ receptor agonists (e.g., sumatriptan, zolmitriptan, naratriptan) in migraine is based on selective cranial arterial vasoconstriction, but in some cases their efficacy may also involve CNS actions. 5-$HT_{2A/2B/2D}$ receptor antagonists (e.g., pizotifen, methysergide) can prevent migraine attacks; the mechanisms are uncertain, but may include regulation of cranial arteriolar vasculature. 5-HT_2 receptor antagonists are used in carcinoid syndrome to suppress excessive ANS neuroeffector cell responses to enterochromaffin cell 5-HT, especially in gastrointestinal, vascular, and bronchial smooth muscle. The wide distribution of 5-HT receptors in the ANS may be exploited in the future as new therapeutic targets; disorders of urinary bladder function may be a promising field.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Nervous System; Parasympathetic Nervous System; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors; Sympathetic Nervous System.

Further Reading

Akasu T (1992) Synaptic transmission and modulation in parasympathetic ganglia. *Japanese Journal of Physiology* 42: 839–864.

Galligan JJ (1996) Electrophysiological studies of 5-hydroxytryptamine receptors on enteric neurons. *Behavioural Brain Research* 73: 199–201.

Gershon MD (2004) Review article: Serotonin receptors and transporters – their roles in normal and abnormal gastrointestinal motility. *Alimentary Pharmacology and Therapeutics* 20(supplement 7): 3–14.

Hoyer D, Clarke DE, Fozard JR, et al. (1994) VII. International Union of Pharmacology classification of receptors for 5-hydroxytryptamine (serotonin). *Pharmacological Reviews* 46: 157–203.

Hoyer D, Hannon JP, and Martin GR (2002) Molecular, pharmacological and functional diversity of 5-HT receptors. *Pharmacology, Biochemistry and Behaviour* 71: 533–554.

Jordan D (2005) Vagal control of the heart: Central serotonergic (5-HT) mechanisms. *Experimental Physiology* 90: 175–181.

McCall RB (1990) Role of neurotransmitters in the central regulation of the cardiovascular system. *Progress in Drug Research* 35: 25–84.

McQueen DS and Mir AK (1989) 5-Hydroxytryptamine and cardiopulmonary and carotid body reflex mechanisms. In: Fozard JR (ed.) *The Peripheral Actions of 5-Hydroxytryptamine*, pp. 301–326. Oxford: Oxford University Press.

Mylecharane EJ and Phillips CA (1989) Mechanisms of 5-hydroxytryptamine-induced vasodilatation. In: Fozard JR (ed.) *The Peripheral Actions of 5-Hydroxytryptamine*, pp. 147–181. Oxford: Oxford University Press.

Peroutka SJ and Snyder SH (1979) Multiple serotonin receptors: Differential binding of [^{3}H]5-hydroxytryptamine, [^{3}H]lysergic acid diethylamide and [^{3}H]spiroperidol. *Molecular Pharmacology* 16: 687–699.

Phillips CA, Mylecharane EJ, and Shaw J (1985) Mechanisms involved in the vasodilator action of 5-hydroxytryptamine in

the dog femoral arterial circulation *in vivo*. *European Journal of Pharmacology* 113: 325–334.

Ramage AG (2001) Central cardiovascular regulation and 5-hydroxytryptamine receptors. *Brain Research Bulletin* 56: 425–439.

Ramage AG (2006) The role of central 5-hydroxytryptamine (5-HT, serotonin) receptors in the control of micturition. *British Journal of Pharmacology* 147: S120–S131.

Villalón CM, Sánchez-López A, Centurión D, et al. (2001) Unravelling the pharmacological profile of the canine external carotid vasodilator '5-HT_1-like' receptors: Coexistence of -sympatho-inhibitory 5-HT_{1B} and postjunctional 5-HT_7 receptors. *Naunyn-Schmiedeberg's Archives of Pharmacology* 363: 73–80.

Wallis DI (1989) Interaction of 5-hydroxytryptamine with autonomic and sensory neurones. In: Fozard JR (ed.) *The Peripheral Actions of 5-Hydroxytryptamine*, pp. 220–246. Oxford: Oxford University Press.

Relevant Website

http://www.iuphar-db.org – IUPHAR Receptor Nomenclature and Drug Classification database.

Cellular Dynamics Revealed by Digital Holographic Microscopy

P Marquet, Centre des Neurosciences Psychiatriques, Lausanne, Switzerland
C Depeursinge, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Modern microscopy techniques have revolutionized the analysis of living cells. With the development of new optical technologies (e.g., laser technology and confocal scanning microscopy), together with the availability of powerful numerical image processing (deconvolution techniques and computational adaptive optics), optical microscopy has moved beyond the traditional two-dimensional imaging toward the reproduction of high-resolution, three-dimensional (3-D) objects. Moreover, the development of video and digital image processing is allowing visualization of cellular movement and behavior. However, obtaining in real time noninvasive, high-resolution 3-D images, which are critical for elucidating the relationship between the dynamics of cell morphology and physiological or pathological processes, remains a technical challenge, especially since most biological cells are usually transparent (i.e., they differ only slightly from their surroundings in terms of optical properties, including absorbance and reflectance) and thus preclude the generation of adequate contrast required to obtain high-resolution video microscopy.

Consequently, many modes of contrast-enhancing techniques have been developed to overcome this limitation. Among these, fluorescence microscopy in a confocal configuration (confocal fluorescent laser scanning microscopy (CLSM)) and its extension to multiphoton fluorescent excitation is a powerful and widely used cellular imaging technique in biology.

CLSM has permitted in particular, in combination with the utilization of specific fluorescent cellular probes, the observation of dynamic processes in living cells ranging from the fusion of vesicles to the concentration changes of specific intracellular signaling molecules. The CLSM optical sectioning capacity, limited to approximately 500–800 nm in axial direction, provides a first step in the 3-D location of intracellular dynamic processes. The 3-D representation of cellular morphology is obtained by the processing of the diffraction-limited image sets, consisting of serial optical sections, using 3-D reconstruction and rendering algorithms.

However, in the field of fluorescence imaging, several pioneering super-resolution technologies, termed 4 Pi microscopy, I^5M microscopy, and stimulated emission depletion (STED) microscopy, are currently taking the resolution beyond the classical Abbe's resolution limit. Specifically, STED microscopy, utilizing reversible saturable optical fluorescence transitions, is a very promising super-resolution technology which can, in principle, break the diffraction barrier.

Nevertheless, the use of fluorescent probes, although they are attractive as chemically specific contrasting agents, is associated with undesirable side effects such as photobleaching, which could ultimately compromise attempts to improve resolution, and photodynamic damage to biological samples, particularly in the confocal observation mode, which concentrates high optical energies. The acquisition rate is also intrinsically limited by the serial acquisition of the image pixels, which precludes the possibility of simultaneously obtaining real-time and high-resolution 3-D images.

On the other hand, phase contrast (PhC), initially proposed by F Zernicke for image contrast, as well as Normarski's differential interference contrast (DIC) are two other widely used contrast-generating techniques available for high-resolution light microscopy. In contrast to fluorescence techniques, PhC and DIC allow the visualization of transparent specimens, particularly the fine subcellular structural organization, without using any staining contrast agent. These two noninvasive contrast-generating techniques, PhC and DIC, result from their capability of transforming in detectable intensity modulation, the minute relative phase shift that transparent microscopic objects, including living cells, induce between the transmitted wave light and the undeviated background wave (PhC) and between two orthogonally polarized transmitted waves (DIC). However, PhC and DIC do not allow the direct and quantitative measurement of phase shift or optical path length. Indeed, PhC suffers from well-known optical artifacts, namely the halo and shading off, making the intensity modulation at any pixel not directly proportional to the phase change due to the corresponding area of the cell. Regarding DIC, the intensity modulation is proportional to the spatial directional gradient along the shear direction of the intracellular refractive index, resulting in a spurious asymmetry into the image. Consequently, the interpretation of minute DIC or PhC signal variations in terms of the corresponding cellular morphometry or shape changes remains elusive.

In contrast, by using the coherence property of light (i.e., the light property that maintains a well-defined wavefront over a certain propagation length), interferometric techniques have the capability to provide a direct measurement of the optical path length based on an interference between the light wave passing the

specimen, called the object wave, and a reference wave. Specifically, the optical path length difference, or the corresponding relative phase shift between the object and reference waves, is coded in the resulting interferogram as an intensity modulation that can be recorded by any photosensitive media, which respond to the light intensity only.

Basically, interferometric detection makes it possible to quantitatively measure a phase shift corresponding to a fraction of the wavelength of the used coherent light wave (e.g., a few nanometers). However, because phase shifts are very sensitive to experimental artifacts, including lens defects and noise originating from vibrations or thermal drift, temporal phase shifting interferometric techniques require very demanding and costly optomechanical designs. Accordingly, their applications have been limited to a few fields, such as metrology and astrophysics. In addition, classical interferometric techniques are based on phase measuring procedures that require multiple interferogram acquisitions and phase modulation devices. Because of these technical constraints, very few attempts to use interferometric phase measurement techniques have been reported in biology for real-time living cell imaging.

Digital Holographic Microscopy

Holography Principle

Holography, initially proposed by Gabor in the late 1940s and which enjoyed a revolution in the early 1960s due to emerging laser technology, is a particularly attractive two-step interferometric imaging technique because it provides the opportunity to capture, on a single frame, both the amplitude and the phase data of the object wave (i.e., the coherent light wave diffracted by the observed specimen) (Figure 1).

The first step consists of photographically recording the interference pattern resulting from the interference of the object wave with the reference wave (Figure 1(a)). The resulting pattern was called a hologram by Gabor, in reference to the Greek word 'holos,' meaning 'whole.' The second step is the reconstruction of the object wave. This is achieved through the diffraction process of a coherent wave by the developed hologram, which can be realized either as a reflective or a transparent medium (Figure 1(b)). Specifically, from the diffraction process results a reconstructed object wave, which is an exact replica of the original wave, diffracted by the specimen. This means that you would then see the reconstructed image in its perfect three-dimensionality, exactly as if the original object were before you actually generating the diffracted wave.

However, despite the outstanding opportunities for capturing and displaying 3-D scenes, classical holographic microscopy suffers from technical constraints similar to those of other interferometric techniques.

In contrast, digital holography is achieved by using an electronic camera to record digital holograms and by applying fast numerical procedures to propagate and therefore reconstruct the object wave. Basically, the whole diffraction process, in which the coherent wave re-illuminates the hologram (Figure 1(b)), is numerically achieved by a computer according to the diffraction theory. This numerical approach of holography was proposed for the first time more than 30 years ago. However, only the development of a solid-state image sensor, with high spatial resolution

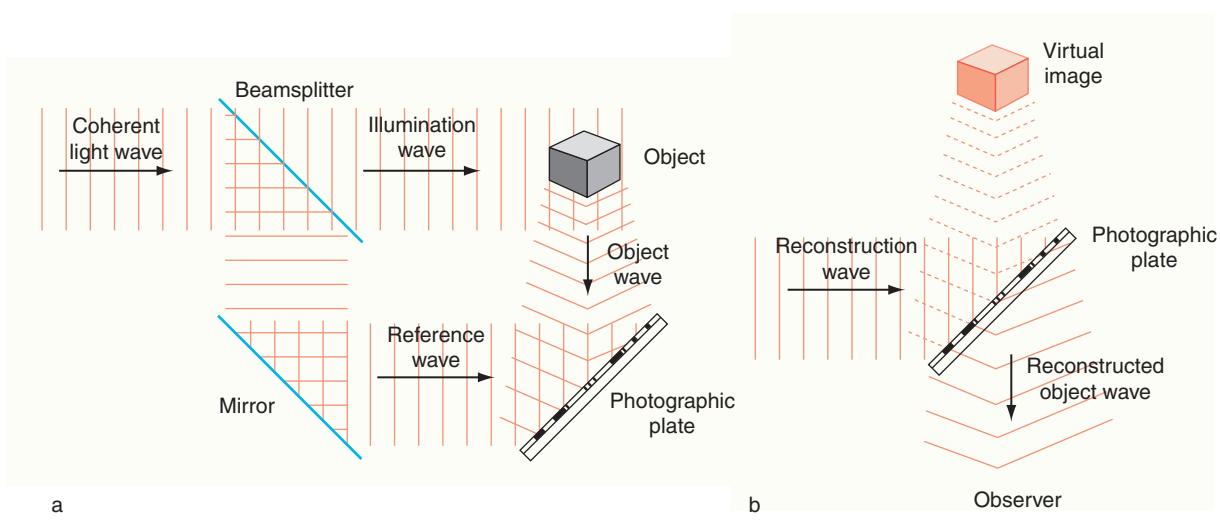

Figure 1 (a) Hologram formation and recording on a photographic plate. (b) Hologram reconstruction by re-illumination of the photographic plate by a coherent wave.

and extended dynamic range (including CCDs or complementary metal oxide semiconductor (CMOS)), combined with current computer power makes the numerical reconstruction of digitally recorded holograms efficient and has stimulated the reassessment of digital holography as a powerful 3-D imaging tool.

Within the framework of microscopy, numerical processing of holograms presents the unique advantage of offering not only the means to reconstruct an exact replica of the object wavefront but also the means to reshape it. Such a numerical reshaping allows correcting experimental artifacts, including lens defects and aberrations. Moreover, compensating for noise is also an issue. Specifically, the perturbing effects of time drifts, erratic fluctuations, or vibrations can also be eliminated to a large extent by numerical treatment of holograms, providing the temporal stability required to observe dynamics processes in particular. Moreover, it is also possible to simulate the effect of optical components (lenses, filters, etc.) on the reconstructed wavefront. Consequently, *a posteriori* numerical remodeling of the reconstructed object wavefront can provide the stigmatism required to form a high-resolution image, which is usually achieved by high-precision lenses such as microscope objectives.

Digital holographic microscopy (DHM) is thus an interferometric technique in microscopy that is considerably simpler to implement than classical interference microscopy while providing a reliable and quantitative phase mapping of the observed specimen, with an accuracy corresponding to optical path length variations of a few nanometers.

DHM Setup

The experimental setup of DHM presented in Figure 2(a) is basically a Mach–Zehnder interferometer. Light transmitted by the specimen and collected by a microscope objective (MO) forms the object wave O, which interferes with a reference wave R to produce the hologram intensity I_H recorded on a standard black-and-white CCD camera. A typical interference fringe pattern is shown in Figure 2(b).

The hologram acquisition rate and time are camera limited (typically 25 holograms per second with a standard 512×512 CCD camera, decreasing to approximately $10\,s^{-1}$ for a higher-resolution camera (1024×1024)). The hologram reconstruction rate is approximately 15 images per second for a 512×512 pixel hologram with a standard personal computer.

The numerical processing of hologram constitutes a radically novel approach in the field of digital microscopy in the sense that it performs a numerical reconstruction and optional reshaping of complex wavefronts. The need for complex optical adjustments is therefore avoided. For instance, the task of orienting a mirror or a beam splitter with translational or rotational degrees of freedom – which has to be performed very accurately in classical interferometry – is simply replaced by the digital adjustment of specific parameters involved in the hologram numerical reconstruction. Even more relevant is the digital correction of the wavefront deformation induced by the MO to form a magnified image, which scrambles the wavefront deformation produced by the specimen.

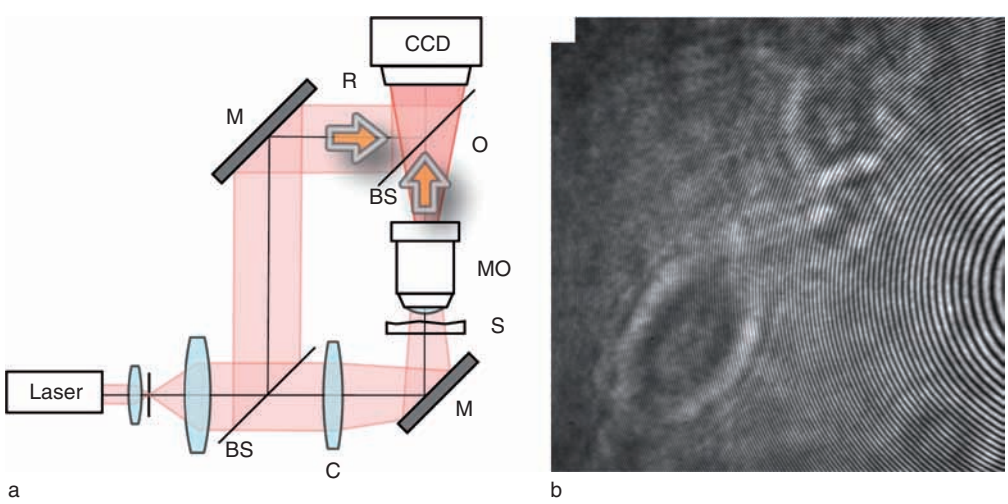

Figure 2 (a) Basic configuration for digital holographic microscopy (DHM). A laser diode produces the coherent light ($\lambda = 658$ nm), which is divided by a beam splitter (BS). The specimen (S) is illuminated by one beam through a condenser (C). A microscope objective (MO) collects the transmitted light and forms the object wave (O) which interferes with a reference beam (R) to produce the hologram recorded by the digital CCD camera. (b) Typical example of a hologram acquired by the camera (1024×1024 pixels).

To practically illustrate the benefit of performing numerically the adjustment of reconstruction parameters in digital holography, let us consider its experimental counterpart in interferometric microscopy: the well-known Linnik interference microscopy. Close matching of the optics in the two separate interferometer arms requires a mechanical design providing exceptional precision and stability. Specifically, the introduction in the reference arm of a second identical microscope objective is needed, and this must be aligned with exceptional mechanical precision. In contrast, numerical hologram processing adds unmatched flexibility: it permits to correct the phase aberrations introduced by the MO without physically adding a second identical MO in the reference arm, thus providing a mean to significantly simplify the optical design of DHM. Moreover, it allows to compute the reconstructed object wave at any distances from the MO exit pupil and to correct the associated optical aberrations. This feature provides a means of refocusing purely numerically the out-of-focus planes and thereby extending the depth of focus without any mechanical action. Thus, image information, even for an optical system with a high numerical aperture and high magnification, is no longer limited to a single plane.

Furthermore, processing and reconstructing numerical holograms *a posteriori* provides an effective means of significantly improving the phase temporal stability: for each pixel of the reconstructed image, a standard deviation, over 60 min periods, as low as 0.7° and 0.2° for a 16-frame average can be achieved. This result is equivalent to an axial sensitivity of $\lambda/1800$. In comparison, the transverse resolution of the reconstructed phase distribution is diffraction limited and thus depends on the NA and the wavelength of the coherent light source used. For example, at the wavelength of $\lambda = 658$ and for a 60×, 1.4 NA microscope objective, the transverse resolution is *c*. 0.3 μm.

Cellular Imaging

As presented in Figure 3(a), the cell culture is enclosed in a perfusion chamber and transilluminated. The object wavefront recording on the CCD encodes the phase shift related to neurons, as well as the wavefront deformation induced by the MO that is numerically corrected during the reconstruction process. A typical reconstructed quantitative phase image of a living cell in culture, presenting in color-coded 3-D perspective, is shown in Figure 3(b).

It is important to note that the interpretation of this phase information in terms of height data is not straightforward. Indeed, the images obtained give the planar distribution of phase delays due to the higher speed of light in the perfusion medium (physiological solution) compared to the intracellular compartment. This is generally addressed by using the notion of refractive index ($n = c/v$), defined as the speed of light in vacuum (c) divided by the speed of light in the medium (v), and by expressing the phase information (ϕ) in terms of optical path length ($\mathrm{OPL} = \phi \cdot \lambda/2\pi$), a physical quantity defined by the product between an average refractive index and

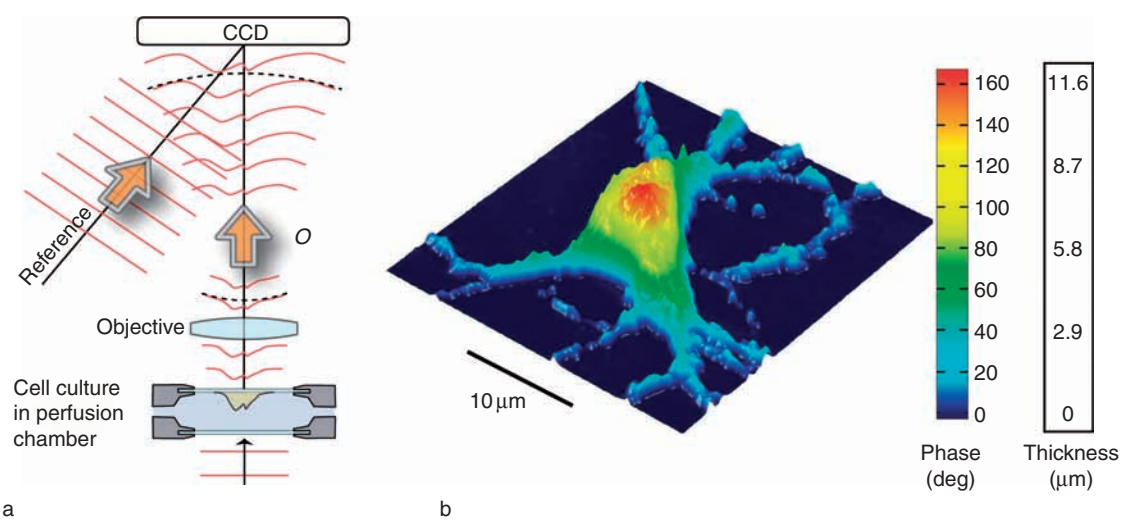

Figure 3 Digital holographic microscopy (DHM) of living mouse cortical neurons in culture. (a) Schematic representation of cultured cells mounted in a closed perfusion chamber and transilluminated. (b) Three-dimensional perspective image in false colors of living neurons in culture. Each pixel represents a quantitative measurement of the cellular optical path length (OPL), with a sensitivity corresponding to a few tens of nanometers. By using the measured mean value of the neuronal cell body refractive index, resulting from the decoupling procedure, scales (right) which relate OPL (deg) to morphology in the *z*-axis (μm) can be constructed.

the thickness of the sample. In the current case, for each pixel i of the absolute phase image resulting from the wavefront reconstruction, we have

$$\mathrm{OPL}_i = (\bar{n}_{c,i} - n_m)d_i + n_m D \quad [1]$$

where d_i is the cellular thickness corresponding to the pixel i of the reconstructed image, $\bar{n}_{c,i}$ is the intracellular mean refractive index along the corresponding cellular thickness d_i, n_m is the constant refractive index of the surrounding medium, and D is the thickness of the flow chamber. The product $n_m D$ is a reference value that can be measured anywhere outside the cell. Monitoring this reference value is important because it enables the compensation of mechanical or thermal instabilities and drifts of the setup during the experiment. The cellular quantitative phase contrast, explicitly given by $(\bar{n}_{c,i} - n_m)d_i$ for each pixel i, results from the fact that the intracellular refractive index is different from the refractive index of the perfusion solution n_m. It has been noted that cellular material, including proteins, lipoproteins, and nucleic acids, contributes to increase the intracellular refractive index to values significantly larger than n_m, in direct proportion to the amount of non-aqueous material present (dry mass), whether in the dissolved phase or not. However, the intracellular refractive index value is mainly determined by the concentration of proteins within the cell because their concentration, in terms of weight per volume, is considerably greater than that of the other solutes. Thus, the OPL contrast is an intrinsic contrast directly proportional to the amount of cellular material in the light path, sometimes referred as the 'dry mass' of the cell.

Practically, the refractive index of the perfused solution $n_m = 1.3336 \pm 0.0002$ (at $\lambda = 658$ nm) was measured with an Abbe refractometer. For cells in culture with a mean intracellular refractive index $\bar{n}_c$ of approximately 1.375 according to the rare value found in the literature and the measurements achieved with DHM (the decoupling procedure), the phase measurement pixel accuracy of 0.2°, dictated by overall experimental noise and resulting from the average over 16 frames, corresponds to a phase sensitivity of approximately 10 nm. Basically, the similar values of $\bar{n}_c$ and n_m contribute to decrease the phase sensitivity, equivalent to $\lambda/1800$ in a reflection configuration, by a factor of approximately 25. Cells in culture induce a typical OPL contrast within the range of 50–200° corresponding to a signal-to-noise ratio of approximately 1000–250. Most important, these well-contrasted quantitative phase images are obtained without contrasting agents and at very low light level illumination intensity. The typical measured irradiance at the specimen plane was approximately 200 μW cm^{-2}, which is several orders of magnitude lower than the power used in classical CLSM.

In order to adequately interpret the quantitative phase images, an experimental protocol called the decoupling procedure can be applied to calculate separately the intracellular refractive index and the corresponding cellular morphometry (i.e., the cellular thickness). This decoupling procedure, performed in a fast exchange time perfusion chamber, consists of perfusing, consecutively with the standard perfusion solution, a second solution with the same osmolarity but with an altered refractive index, $n'_m = n_m + \delta n$, and recording the two corresponding holograms. The two resultant reconstructed phase distributions provide, according to eqn [1], a system of two equations that hold for each pixel i, allowing to calculate $\bar{n}_{c,i}$ and the corresponding cellular thickness d_i. This decoupling procedure provides a thickness of 1–3 μm for the neuronal processes and approximately 11 μm for the cell body (Figure 3(b)).

Real-Time Visualization of Cellular Dynamics

DHM has the capability to address specific biological questions concerning living cells and involving cellular morphological changes induced by agonist-mediated activation of specific membrane receptors. To illustrate how the information contained in the receptor-mediated phase signal changes is related to the underlying cellular process, the DHM visualization of the well-known acute Na^+- and Cl^--dependent excitotoxicity neuronal swelling, mediated by sustained application of the excitatory neurotransmitter glutamate at high concentration, is presented. Indeed, glutamate activates postsynaptic receptors, including the ionotropic *N*-methyl-D-aspartate (NMDA), 2-amino-3-(3-hydroxy-5-methylisoxazol-4-yl) proprionate (AMPA), and kainate receptors. Upon their activation, their associated ion channel opens to allow the influx of Ca^{2+} and Na^+ ions. Consequently, perturbations in systems using glutamate result in excessive activation of the previously mentioned receptors leading to excitotoxicity characterized by two main components: (1) an acute Na^+- and Cl^--dependent component marked by immediate cell swelling and (2) delayed cell degeneration related to the influx of Ca^{2+}, whose mechanisms are not clearly elucidated.

The neuronal swelling associated with the excitotoxic effect of glutamate has been induced by prolonged application of the neurotransmitter (50 μM for 5 min) on primary culture of mouse cortical neurons. As indicated in Figure 4, the typical response

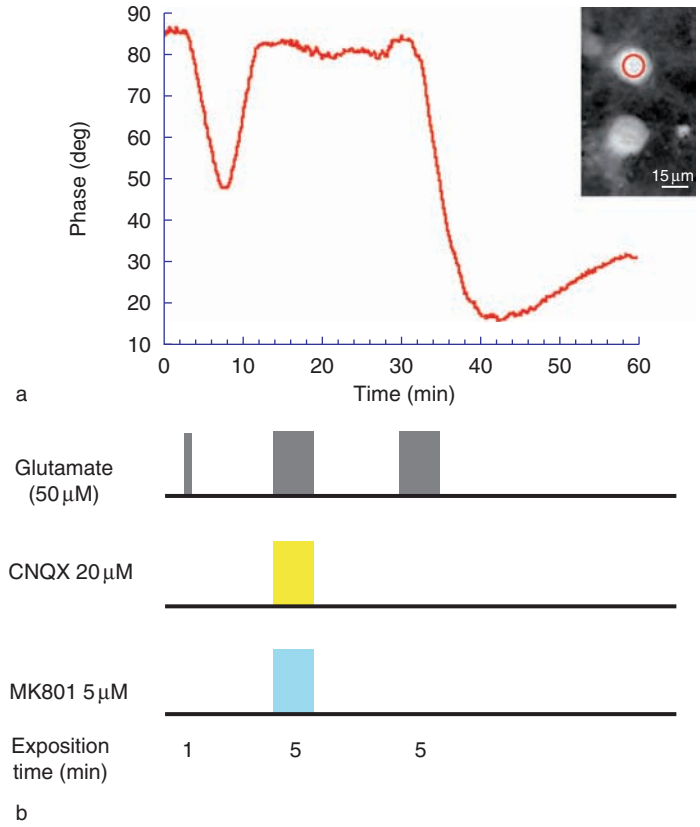

Figure 4 Real-time observations of optical path length (OPL) changes produced by prolonged glutamate applications (a). Neurons were observed in a control solution, glutamate, and pharmacological agents were applied as indicated in (b). Prolonged glutamate application induces a large nonreversible OPL decrease of approximately 70°. Such large OPL responses can be completely inhibited by co-application of 20 μM CNQX, a specific AMPA/kainate ionotropic glutamate receptor antagonist, and MK-801 (5 μM), a specific NMDA receptor antagonist. (a) The curves represent the evolution of the OPL averaged in the domain delimited by a red circle in the inset, representing the image of the OPL spatial distribution.

of such prolonged glutamate applications is large OPL decreases of several tens of degrees. These large OPL decreases quickly become irreversible as the glutamate concentration and application time increase. The receptor-specific nature of this response is demonstrated by the suppression of OPL response inhibition resulting from the co-application of MK801 and CNQX, two specific antagonists of NMDA and AMPA/kainate ionotropic receptors, respectively.

Figure 5 illustrates the typical phase distribution changes of two neuronal bodies (Figure 4 inset) induced by glutamate perfusion. The two images on the left ('standard') were acquired just before glutamate application. The images on the right ('excitotoxic') were acquired when the phase stabilized approximately 10 min after the onset of glutamate perfusion. This representation of the phase distribution shows a global phase decrease accompanied by cellular body enlargement.

In contrast to the cellular surface enlargement consistent with neuronal swelling, the global phase decrease is difficult to interpret as a cellular swelling. However, the decoupling procedure providing an image of the surface morphometry, in which each image pixel represents the cellular thickness, allows the clear assessment of the expected excitotoxic neuronal swelling (Figure 5(b)).

From such morphometry images (Figure 5(b)), absolute cellular volume, as well as relative volume variations can be calculated. The neuronal shape variations can also be quantitatively observed and indicate that excitotoxic neuronal body swelling is a spatially heterogeneous and anisotropic process presenting a cellular thickness decrease in some central regions of the neuronal body.

On the other hand, the glutamate-mediated excitotoxicity induces an intracellular refractive index decrease. The typical values are 1.387 before

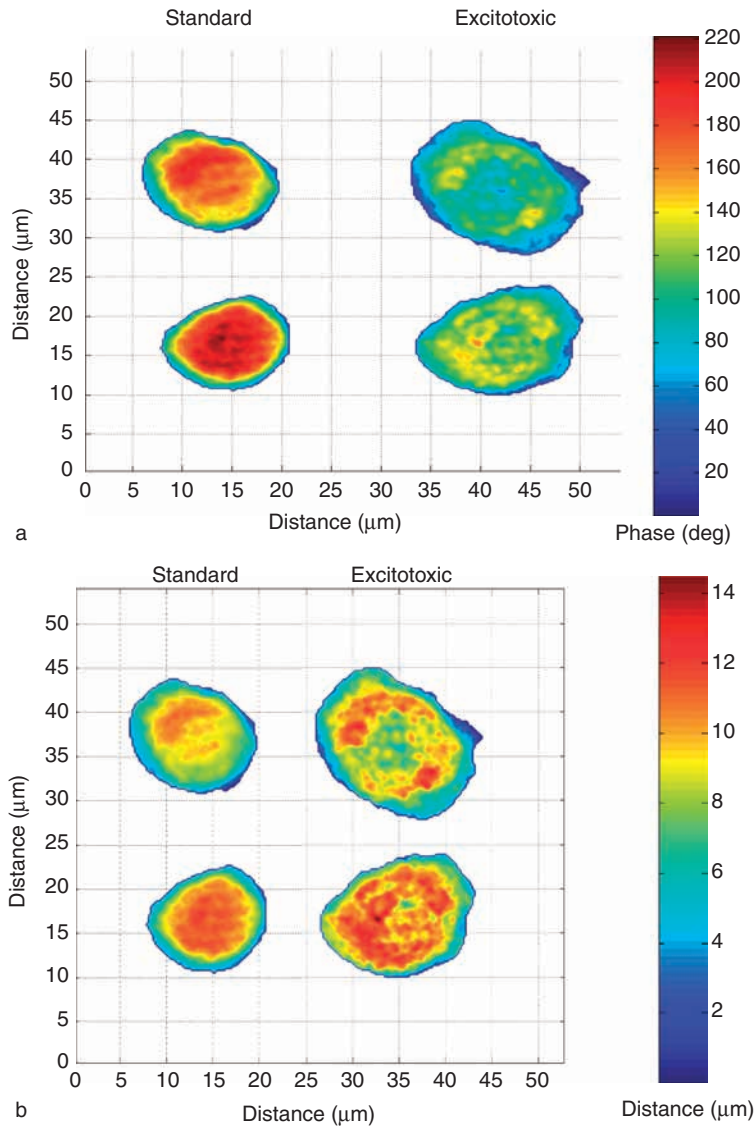

Figure 5 (a) Color-coded quantitative phase images of two neuronal cell bodies (Figure 4 inset) before ('standard') and approximately 10 min after the onset of the glutamate perfusion ('excitotoxic'). (b) Morphometry of the two cell bodies corresponding to standard and excitotoxic conditions. The *z*-axis (cellular thickness) is expressed in micrometers. Neuronal cell body boundaries were identified by a gradient-based edge detection algorithm.

glutamate application and 1.373 after neuronal swelling. This refractive index decrease is consistent with a water influx accompanying ion entry through the glutamate ionotropic receptors, resulting in a dilution of the intracellular protein concentration, the cellular component, which largely determines the intracellular refractive index value. Consequently, the *a priori* paradoxical phase signal decline results from the intracellular refractive index decrease leading to a decreasing of the product $(\bar{n}_{c,i} - n_m)h_i$ in eqn [1] despite the increased neuronal thickness h_i.

The temporal dynamics of both the 3-D cellular morphometry and the intracellular refractive index during excitotoxic neuronal swelling can be studied simultaneously.

Nanoscale Cellular Fluctuations

The DHM high axial sensibility allows noninvasive observation of cellular fluctuations at the nanoscale. As an illustration, the dynamics of red blood cells (RBCs), known to exhibit natural membrane fluctuations involving cytoskeleton, is presented. RBC

natural fluctuations are assessed by comparing the phase temporal fluctuations corresponding to RBC regions with the phase fluctuations of reference regions in which there is no biological material, corresponding to the experimental noise. Typical RBC and reference phase fluctuations are presented in Figure 6.

DHM can also perform a complete spatial mapping (Figure 7) of the RBC phase fluctuations (i.e., phase fluctuations in any specific cellular microdomains), providing a more comprehensive visualization of such a dynamic cellular process. As presented in Figures 7(a′) and 7(b′), mapping of the phase standard deviation allows clear discrimination between the RBC phase fluctuations and the random reference phase fluctuations resulting from the experimental noise. Specifically, this is expressed by the appearance of high standard deviation regions accurately corresponding to the body of living RBCs.

On the other hand, the decoupling procedure allows the calculation of the RBC intracellular refractive index values, as well as its temporal fluctuations. By considering the measurements of the experimental phase stability, the intracellular refractive index and its temporal variations, the cellular RBC membrane fluctuations can be calculated from the RBC phase signal fluctuations (Figure 6, red line). A typical intracellular refractive index value of approximately 1.41 ($\lambda = 658$ nm) was measured for living RBCs. Concerning the RBC membrane fluctuations, typical amplitudes of a few tens of nanometers with spatial variations have been obtained. It is also important to note that the temporal phase signal fluctuations provided by DHM permit a direct frequency analysis of the cellular fluctuations.

Conclusion

DHM is an optical imaging technique in which quantitative phase contrast images of transparent living cells can be obtained from a single recorded hologram, without the use of any staining agent. The interferometric principle applied to the phase detection, as well as the numerical hologram reconstruction, allows monitoring cell dynamics with a nanometer axial sensitivity. These dynamic phase images, containing information about both the cell morphometry and the intracellular refractive index, can be unambiguously interpreted due to the decoupling procedure. The capability of DHM to monitor noninvasively receptor-specific biological processes in living cells is illustrated by the visualization of the well-established

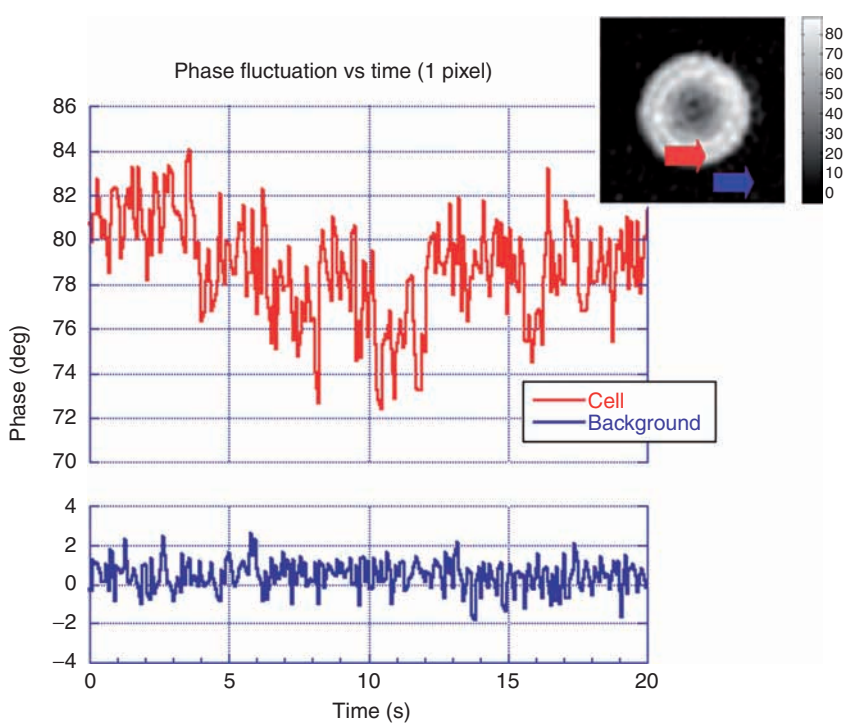

Figure 6 Real-time observation of red blood cells (RBCs) and reference phase temporal fluctuations. The red line shows the pixel phase temporal fluctuation corresponding to an RBC region indicated by the red arrowhead in the inset, representing a quantitative phase image of an RBC. Similarly, the blue curve represents the pixel phase temporal fluctuations associated with a reference region outside the RBC indicated by the blue arrow. Living RBCs were observed in high-efficiency particulate air (HEPA) buffer.

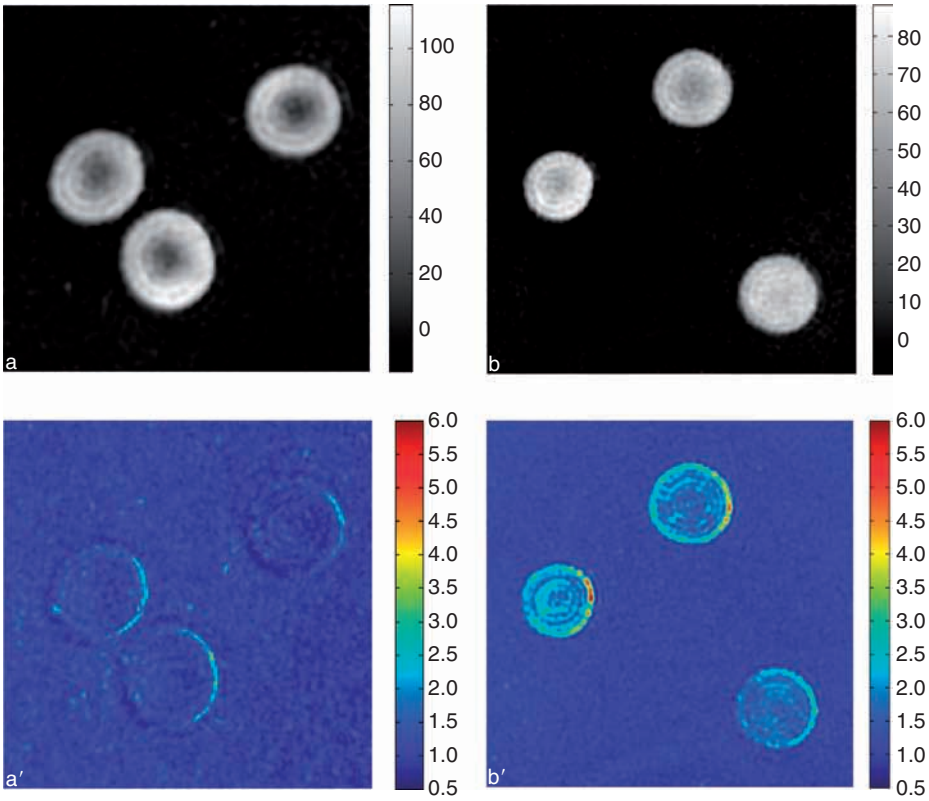

Figure 7 (a, b) Quantitative phase images of fixed and living RBCs, respectively. Each pixel value of the corresponding images (a′ and b′) is the standard deviation of the phase signal fluctuations evaluated during a 10 s interval.

glutamate receptor-mediated excitotoxic neuronal swelling. Specifically, the study of the cell shape dynamics, including membrane nano-movements, with a subwavelength axial accuracy can be addressed. In addition, the high-accuracy measurement of the intracellular refractive index, a poorly documented parameter, provides information on cell content. Because the intracellular refractive index is related to the composition of the intracellular medium, particularly proteins, spatial refractive index mapping represents a promising way to monitor the local modifications of intracellular protein distribution that occur during cellular processes.

See also: Excitotoxicity in Neurodegenerative Disease; Fluorescence Microscopy in the Neurosciences; Glial Glutamate Transporters: Electrophysiology; Glial Glutamate Transporters; Glutamate.

Further Reading

Arundine M and Tymianski M (2003) Molecular mechanisms of calcium-dependent neurodegeneration in excitotoxicity. *Cell Calcium* 34: 325–337.

Bitler A, Barbul A, and Korenstein R (1998) Detection of movement at the erythrocyte's edge by scanning phase contrast microscopy. *Journal of Microscopy* 193: 171–178.

Carl D, Kemper B, Wernicke G, and von Bally G (2004) Parameter-optimized digital holographic microscope for high-resolution living-cell analysis. *Applied Optics* 43: 6536–6544.

Colomb T, Cuche E, Charrière F, et al. (2006) Automatic procedure for aberration compensation in digital holographic microscopy and application to specimen shape compensation. *Applied Optics* 45: 851–863.

Cuche E, Marquet P, and Depeursinge C (1999) Simultaneous amplitude and quantitative phase contrast microscopy by numerical reconstruction of Fresnel off-axis holograms. *Applied Optics* 38: 6994–7001.

Davis HG and Wilkins MHF (1952) Interference microscopy and mass determination. *Nature* 169: 541.

Gale DM, Pether MI, and Dainty JC (1996) Linnik microscope imaging of integrated circuit structures. *Applied Optics* 35: 131–148.

Hell SW (2003) Toward fluorescence nanoscopy. *Nature Biotechnology* 21: 1347–1355.

Marquet P, Rappaz B, Magistretti PJ, et al. (2005) Digital holographic microscopy, a non-invasive contrast imaging technique allowing quantitative visualization of living cells with a sub-wavelength axial accuracy. *Optics Letters* 30: 468–470.

Pawley JB (2006) *Handbook of Biological Confocal Microscopy*, 3rd edn. New York: Plenum.

Pluta M (1988) *Advanced Light Microscopy: Specialized Method*, vol. 2. New York: Elsevier.

Rappaz B, Marquet P, Cuche E, Emery Y, Depeursinge C, and Magistretti P (2005) Real-time 3D visualization by digital holography microscopy of cellular dynamics with nanoscale resolution. *Optics Express* 13: 9362–9373.

Central Gustatory System and Ingestive Behavior

A C Spector, Florida State University, Tallahassee, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

There is no mystery concerning the primary role of the gustatory system. Its main function is to provide information about substances that will be potentially ingested and, in this sense, its input to the central nervous system helps guide feeding and drinking. The chemical response properties of the taste receptor cells coupled with their location at the opening of the alimentary tract leave little doubt. Taste receptor cells are found in teardrop-shaped clusters called taste buds. These specialized epithelial cells serve as chemosensory contact receptors for chemicals that are in solution. The adequate stimuli for these receptors tend to consist of compounds that are not exceptionally volatile and thus do not optimally stimulate the olfactory system. Most nutritive chemicals in nature, however, are found in mixtures and so chemosensory guidance of feeding and drinking involves an interaction between the taste, smell, touch, and temperature of a given food or fluid. The perceptual consequence of the integration of these signals is referred to as flavor.

Peripheral Taste Nerves

Rodents have been the animal model of choice for molecular, anatomical, and physiological examinations of the gustatory system, and thus much of the information in this article is based on such findings. In rodents (and humans), taste buds are distributed in relatively distinct fields in the oral cavity, each innervated by different branches of cranial nerves (CNs). The taste buds of the anterior tongue are innervated by the chorda tympani nerve and those in the palate are innervated by the greater superficial petrosal nerve, both of which are branches of the facial nerve (CN VII). Taste buds in the posterior tongue papillae (special epithelial structures that house taste buds) are innervated by the glossopharyngeal nerve (CN IX) and those found in the laryngeal epithelium are innervated by the superior laryngeal branch of the vagus nerve (CN X). In teleost fish, the taste buds innervated by the facial nerve, which in some species such as the channel catfish are distributed throughout the body surface, are critical in food search, whereas those innervated by the glossopharyngeal and vagus nerves are critical in food-elicited oropharyngeal motor activity. Although in rodents there is evidence of some degree of functional specialization regarding the sensory role of the gustatory nerves, it appears to be more complex than what has been observed in fish and remains to be fully understood.

Central Taste Pathways

The axons of the gustatory nerves in rodents synapse with second-order neurons in the rostral portion of the nucleus of the solitary tract (NST) in a rough topographic manner with the facial nerve fibers terminating most anterior and the vagal nerve fibers terminating most posterior, but there is significant overlap in these terminal fields (Figure 1). Neurons responsive to taste stimulation in the rostral NST intermingle with those responsive to mechanical and thermal stimulation of the oral cavity. Indeed, some neurons are responsive to all three types of stimulation. Vagal sensory axons arising from the major organs of the body terminate in the caudal portion of the NST in a viscerotopic manner. Thus, taste-responsive neurons are intermixed with neurons responsive to oral somatosensory stimuli and adjacent to neurons responsive to visceral stimuli. This general anatomical relationship is maintained throughout much of the central gustatory system and provides a structural basis for the integration of these sensory signals in the service of regulating feeding and drinking.

Neurons in the rostral NST project to local brain stem nuclei involved with controlling salivation and oral motor behavior. In rodents and lagomorphs, the main ascending gustatory pathway starts with the ipsilateral projection from taste-responsive neurons of the NST to the parabrachial nucleus (PBN) – an area of cells in the dorsal pons that surrounds the brachium conjuctivum (thus the name 'parabrachial'). The axons from taste-responsive neurons in the PBN form two major ascending pathways. In addition, a subset of these cells projects caudally to the premotor neurons of the medullary parvocellular reticular formation involved in the control of oral movements. One of the pathways destined for the forebrain terminates in the parvicellular subdivision of the ventral posteromedial nucleus of the thalamus, which in turn projects to insular cortex. In general, thalamocortical pathways are the conventional route of transmission of sensory input to cortex. The second forebrain taste pathway emerging from the PBN consists of axons terminating in several ventral forebrain structures, including the lateral hypothalamus, bed nucleus of the stria terminalis, central nucleus of the amygdala, and

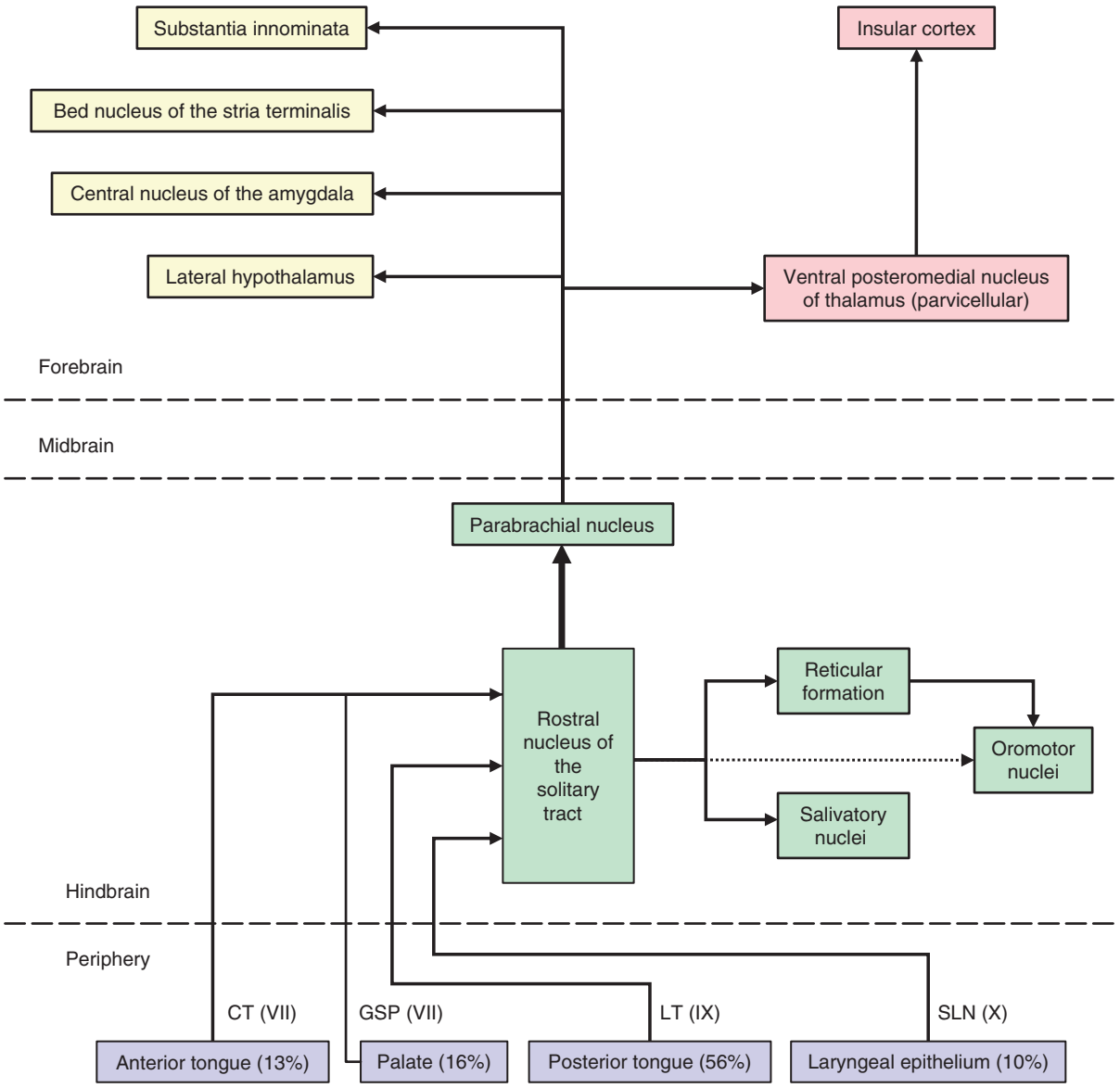

Figure 1 Schematic of the major pathways of the ascending gustatory system in rodents. The percentages in parentheses represent the approximate percentage of total taste buds located in the indicated peripheral fields (blue-shaded boxes) based on a rat model; the remaining 5% are distributed in other locations in the oropharyngeal cavity. Green-shaded boxes represent hindbrain taste nuclei and some associated local reflex structures receiving input from the NST. Red-shaded boxes represent the thalamocortical taste pathway and yellow-shaded boxes represent ventral forebrain structures receiving projections from the gustatory zone of the PBN. For simplicity, descending connections are not shown. CT, chorda tympani nerve; GSP, greater superficial petrosal nerve; LT, lingual-tonsilar branch of the glossopharyngeal nerve; SLN, superior laryngeal nerve. Roman numerals in parentheses represent the indicated CN. See text for more detail. Reproduced from Spector AC and Travers SP (2005) The representation of taste quality in the mammalian nervous system. *Behavioral and Cognitive Neuroscience Reviews* 4: 143–191, with permission.

substantia innominata – all known to be involved with the regulation of motivated behavior. It has been proposed that the thalamocortical pathway plays a privileged role in sensory discriminative aspects of gustatory processing associated with stimulus identification, whereas the ventral forebrain pathway is more involved with processing the hedonic characteristics of taste stimuli. Although there are findings in the literature that provide support for some features of this theoretical organization, evidence for this strict anatomical segregation of function remains to be found. Many of these forebrain sites are anatomically interconnected. Moreover, the insular cortex and the taste-responsive ventral forebrain structures send projections back to the PBN. Indeed, electrical stimulation of some of these forebrain sites has been shown to influence the response properties of PBN neurons during oral taste stimulation.

In rodents, the PBN appears to represent a critical anatomical hub in the gustatory system. In the

primates examined, however, gustatory projections from the NST bypass the PBN and synapse directly with thalamic neurons, which in turn project to the insular and opercular cortex. The orbitofrontal cortex, which receives some other sensory inputs, is considered a secondary cortical area for taste. Limbic forebrain structures in primates apparently derive their taste input from these cortical areas. The significance of the anatomical differences in the central gustatory pathways between rodents and primates remains to be understood.

Functional Contributions of Taste to Ingestive Behavior

The contribution of taste signals to the control of ingestive behavior is manifest in at least three ways – stimulus identification, ingestive motivation, and digestive preparation. Stimulus identification refers to the process by which animals innately or conditionally associate a taste stimulus with particular external or, more often, internal environmental events. The distinguishable qualities of taste compounds, such as the sweetness, saltiness, sourness, and bitterness that humans report, allow animals to discriminate between sapid stimuli to adjust their behavior accordingly. Although most researchers agree that there are a limited number of taste qualities, the exact number is debated and species differences are evident; for example, cats do not respond to 'sweet-tasting' compounds in the same manner that rodents and humans do. This appears to be due to a genetic abnormality that prevents the synthesis of the T1R2 protein which is part of the heterodimeric receptor that binds with sweet-tasting ligands. Because cats are carnivores, however, an absence of carbohydrate detection would not be expected to have a serious consequence on the biological fitness of these animals.

Ingestive motivation refers to the characteristics of a taste compound that either promote or discourage ingestion. Many animals will work to obtain taste stimulation. For example, in behavioral testing chambers rats will press levers in order to receive small aliquots of sucrose solution. Animals will also avoid ingesting certain compounds such as toxic plant alkaloids that are bitter tasting to humans. Taste stimuli are also capable of triggering oral motor reflexes (called taste reactivity) that facilitate either ingestion or rejection of substances in the mouth. Thus, taste stimuli possess a hedonic valence ranging from negative (e.g., aversion) to positive (e.g., reward), and this is commonly referred to as palatability. One area of the brain that receives direct or indirect input from some of the ventral forebrain taste structures and is implicated in the processing of rewarding stimuli, including palatable taste solutions, is the nucleus accumbens (NAc) shell. It has been shown in rats that dopamine release in the NAc shell increases during sham feeding of sucrose in a concentration-dependent manner. In sham feeding, the ingested contents drain out of a surgically implanted gastric cannula that is left open during the test, and thus postingestive stimulation is minimized if not entirely absent.

The palatability of a taste stimulus can be altered by conditioning. Anyone who has tasted a relatively new food or fluid and then soon afterwards has had the misfortune of becoming nauseated and vomiting knows how powerful such learned associations are, as that same food or fluid will be subsequently avoided sometimes for a lifetime. This type of robust learning, referred to as conditioned taste aversion, is very pervasive in mammalian species and has obvious adaptive value. On the other hand, the palatability of a taste stimulus can also be enhanced through conditioning by pairing its ingestion with nutritive postingestive consequences. It is important to recognize that stimulus identification and ingestive motivation, although frequently related, are nonetheless dissociable processes; two taste stimuli can be equally avoided or preferred (i.e., have equivalent palatability) but possess discriminable taste qualities.

Finally, digestive preparation refers to the physiological reflexes such as salivation that are stimulated by taste stimuli and facilitate the digestion and assimilation of ingested substances.

Neural Coding of Taste Quality

With the previous functional framework in mind, efforts have been made to try to understand the mechanics of the neural circuitry underlying these various taste processes. One long-standing goal of researchers has been to discover how the nervous system represents the qualitative features of a taste stimulus that subserve identification. Two theories of the neural coding of taste quality have dominated the literature – the labeled line theory and the across-neuron pattern theory. The labeled line theory postulates that a given taste quality arises from the necessary and sufficient activity in a specific class of neurons defined by its electrophysiologically assessed response properties. For example, the activity in S-units, so called because these neurons respond optimally to sucrose and other sugars, would give rise to a sensation of sweetness, and activity in N-units, which are maximally responsive to NaCl, would give rise to saltiness, and so on. The across-neuron pattern theory postulates that the pattern of activity from the entire population of taste-responsive neurons serves as the code for taste quality.

Proponents of the across-neuron pattern argue that there are too many neurons in the central gustatory

system that are broadly tuned, responding to more than one qualitative class of taste stimuli (e.g., sucrose and NaCl), for a putative labeled line to unambiguously represent a single taste quality. Although it is true that the breadth of tuning varies widely across taste-responsive neurons in the nervous system likely arising from converging peripheral inputs, some neurons are rather narrowly tuned and it remains unclear to which functions a given taste-responsive neuron contributes. Theoretically, narrowly tuned units could potentially be involved in labeled line coding for taste quality and more broadly tuned units could be involved with processing some other stimulus dimension such as hedonic valence. Complicating matters further is the suggestion that the temporal pattern of activity could possess information related to taste quality over and beyond the accepted view that a general increase in overall firing rate is related to increased intensity of the stimulus. Indeed, some researchers are employing computational models that incorporate both spatial and temporal coding features to examine the activity in ensembles of taste-responsive neurons. The neural coding problem has been challenging to solve partly because variations in taste quality do not ostensibly map onto a continuous physical dimension of the stimulus in a manner such as the wavelength of light does for color or the frequency of sound does for pitch, rendering the experimental search for orderly principles of representation difficult.

Hindbrain versus Forebrain Taste Function

Insight into some fundamental aspects of the functional organization of the central gustatory system has come from experiments with decerebrate rats. In this experimental preparation, the forebrain is neurally disconnected from the rest of the nervous system. These animals do not voluntarily eat or drink, they have impaired thermoregulation, and they are hyperresponsive to innocuous stimuli. Indeed, animals that survive the surgery can be maintained for months if they are fed by gavage and their core temperature is maintained by external adjustments. Remarkably, intraoral infusions of taste stimuli through chronically implanted cannulae in decerebrate rats elicit the full range of concentration-dependent oral motor responses in a manner similar to that in intact control rats.

These taste-triggered reflexes, collectively referred to as taste reactivity, are divided into two classes of behavior. The ingestive class, which in rodents includes tongue protrusions, mouth movements, and paw licks, is observed upon stimulation with normally preferred chemical compounds such as sucrose and is associated with ingestion. The aversive class, which in rodents includes gapes, chin rubs, forelimb flails, and head shakes, is observed upon stimulation with normally avoided compounds such as quinine and is associated with rejection. Various forms of taste reactivity can be seen across mammalian species and are evident in human neonates. Indeed, there is evidence that infants born with severe anatomical abnormalities of the forebrain also display taste reactivity. Thus, these findings suggest that basic taste-elicited reflexes are hardwired and that the caudal brain stem is sufficient for these responses to be maintained in the absence of a forebrain.

Intact rats that have a single pairing of sucrose with visceral malaise (induced by a drug or other chemical) will subsequently display aversive, instead of ingestive, taste reactivity to the sugar. Such conditioned taste aversions cannot be acquired by decerebrate rats even after multiple pairings. Thus, the forebrain appears to be necessary for the hedonic valence of a taste stimulus to be changed through conditioning, at least from positive to negative.

Similar impairments in conditioned taste aversion acquisition can be obtained in rats that have more focal lesions centered in the gustatory zone of the PBN bilaterally. Interestingly, rats with PBN lesions can express a presurgically learned taste aversion and can postsurgically learn to avoid fluid stimuli with some nontaste flavor cues. It would appear, then, that the deficit is not attributable to an inability to (1) sense visceral malaise, (2) detect the presence of taste stimuli, (3) recall a taste–malaise association, or (4) display avoidance behavior. Rather, the failure of rats with PBN lesions to acquire conditioned taste aversions appears to be due to an inability to integrate taste and visceral signals. Consistent with this view, rats with gustatory PBN lesions are also unable to learn a conditioned taste preference when ingestion of a specific taste solution is paired with intragastric infusions of a nutritive substance. Because a lesion in this obligate forebrain taste relay deprives downstream brain structures of taste-evoked input, it is difficult to determine the critical brain site(s) involved in any resulting impairments. It could be the PBN itself or any of the forebrain sites to which it projects. Indeed, lesions in some of those forebrain sites can also produce deficits in taste aversion learning, and the experiments with decerebrate rats, in which the PBN was presumably intact, implicate the forebrain as playing a major role in taste-viscerosensory integration.

Taste and Sodium Appetite

A superb example of a taste-guided behavior that involves an interplay between stimulus identification

and ingestive motivation in an ecological context is the phenomenon of sodium appetite. When omnivorous and herbivorous mammals are depleted of body sodium, these animals display heightened search behavior for the electrolyte coupled with increased ingestion of the mineral when found. Deer congregating around salt licks in the winter, when food sources are scarce, provide one illustration in the wild. In the laboratory, the potentiated licking elicited by saline solution in Na^+-depleted rats is relatively chemospecific and is generally not observed when other cations, such as potassium and calcium, are offered, with the exception of lithium. The ability of rats to distinguish sodium from other cations appears to arise from a special ion channel thought to be a member of the epithelial sodium channel (ENaC)/degenerin family, expressed in some taste receptor cells, although the exact nature of the channel is debated. Actually, ENaCs are found in a variety of sodium absorbing tissues including the distal tubule of the kidney, where the hormone aldosterone, released from the adrenal cortex under conditions of body Na^+ deficiency, upregulates the expression of ENaCs, which in part promotes the reabsorption of sodium.

Neurons that respond rather selectively to orally applied sodium salts can be found in both the peripheral and the central gustatory system. These narrowly tuned sodium-responsive units no longer respond to the cation if the tongue is treated with the diuretic drug amiloride, which blocks sodium entry through ENaCs. This effect appears to be limited to the neurons that respond more selectively to sodium because under amiloride blockade, sodium can still stimulate other taste neurons that respond to a wider range of electrolytes. Amiloride, which is tasteless to rats, completely eliminates the selective licking responses toward sodium salts induced by sodium depletion. Thus, at least in rats, natural selection seems to have favored similar mechanisms for the conservation of body sodium to prevent further loss and for the identification of external sources of sodium to facilitate repletion when the animal is challenged with a deficiency of this critical cation.

Although when depleted of sodium intact rats display an enhancement of ingestive taste reactivity and increase intake in response to intraorally delivered sodium chloride, decerebrate rats do not. Likewise, rats with bilateral lesions of the PBN do not display potentiated ingestive responses to sodium chloride. Thus, expression of sodium appetite appears to require the forebrain. Once again, the forebrain is implicated in the integration of taste and visceral signals.

Conclusion

This article provides only a brief overview of some fundamental features regarding the functional neuroanatomy of the central gustatory system within the context of the control of ingestive behavior. There is still much to learn. The advent of modern techniques to measure and manipulate brain function offers promise in efforts to understand the neurobiology underlying the sense of taste.

See also: Conditioned Taste Aversion; Food and Water Intake: Regulation; Salt Appetite; Sensory Aging: Chemical Senses; Taste: Vertebrate Taste Bud Physiology.

Further Reading

Caprio J, Brand JG, Teeter JH, et al. (1993) The taste system of the channel catfish: From biophysics to behavior. *Trends in Neuroscience* 16: 192–197.

Chandrashekar J, Hoon MA, Ryba NJP, and Zuker CS (2006) The receptors and cells for mammalian taste. *Nature* 444: 288–294.

Doty RL (1995) *Handbook of Olfaction and Gustation,* 2nd edn. New York: Dekker.

Grigson PS, Reilly S, Shimura T, and Norgren R (1998) Ibotenic acid lesions of the parabrachial nucleus and conditioned taste aversion: Further evidence for an associative deficit in rats. *Behavioral Neuroscience* 112: 160–171.

Grill HJ and Norgren R (1978) Chronic decerebrate rats demonstrated satiation, but not bait shyness. *Science* 201: 267–269.

Grill HJ and Norgren R (1978) Chronically decerebrate rats demonstrate satiation, but not bait shyness. *Brain Research* 143: 281–297.

Hajnal A, Smith GP, and Norgren R (2004) Oral sucrose stimulation increases accumbens dopamine in the rat. *American Journal of Physiology (Regulatory Integrative Comparative Physiology)* 286: R31–R37.

Lundy RF and Norgren R (2004) Gustatory system. In: Paxinos G (ed.) *The Rat Nervous System,* 3rd edn., pp. 891–921. San Diego: Academic Press.

Nishijo H and Norgren R (1997) Parabrachial neural coding of taste stimuli in awake rats. *Journal of Neurophysiology* 78: 2254–2268.

Norgren R and Leonard CM (1973) Ascending central gustatory pathways. *Journal of Comparative Neurology* 150: 217–238.

Pritchard TS and Norgren R (2004) Gustatory system. In: Paxinos G and Mai J (eds.) *The Human Nervous System*, pp. 1171–1196. San Diego: Academic Press.

Schulkin J (1991) *Sodium Hunger: The Search for a Salty Taste.* Cambridge, UK: Cambridge University Press.

Scott TR and Giza BK (1990) Coding channels in the taste system of the rat. *Science* 249: 1585–1587.

Spector AC (2000) Linking gustatory neurobiology to behavior in vertebrates. *Neuroscience and Biobehavioral Reviews* 24: 391–416.

Spector AC and Travers SP (2005) The representation of taste quality in the mammalian nervous system. *Behavioral and Cognitive Neuroscience Reviews* 4: 143–191.

St. John SJ and Smith DV (2000) Neural representation of salts in the rat solitary nucleus: Brain stem correlates of taste discrimination. *Journal of Neurophysiology* 84: 628–638.

Central Pattern Generators

D Bucher, Whitney Laboratory, University of Florida, St. Augustine, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Repetitive (cyclic) movements underlie crucial behaviors like locomotion (walking, swimming, crawling, and flying), as well as essential vegetative functions like respiration, heartbeat in some invertebrates, chewing, and gut movements. Although both sensory feedback and neuromuscular dynamics play an important role in shaping rhythmic motor output, the basic rhythmic activity patterns are generated by central circuits called central pattern generators (CPGs). These CPGs can, when properly activated, produce rhythmic network activity in the absence of external timing cues, that is, without rhythmic sensory feedback or rhythmic activation by descending neurons.

The most unambiguous demonstration of this principle comes from experiments with isolated parts of the nervous system that produce rhythmic activity *in vitro*, either spontaneously or in response to pharmacological activation. The overall patterning and phasing of such activity very often is surprisingly similar to the basic intact motor behavior and is therefore referred to as a fictive motor pattern or a fictive behavior (e.g., fictive locomotion). Among the first successful demonstrations of the existence of CPGs were experiments with locusts. Those experiments showed that the rhythmic motor output to flight muscles can be generated in the deafferented thoracic nerve cord in an immobilized animal (Figure 1(a)). Possibly the most robust mammalian fictive motor pattern is the respiratory rhythm that can be recorded from slice preparations of the brain stem. This activity occurs spontaneously, without electrical stimulation or pharmacological activation (Figure 1(b)).

Much effort has gone into understanding the basic cellular and network mechanisms of rhythm generation, with the ultimate goal of linking circuit operation to basic behaviors. Invertebrate CPGs have been invaluable in this endeavor, predominantly because many invertebrate circuits consist of a limited number of individually identifiable neurons with large cell bodies. This makes them tractable systems for establishing circuit connectivity and allows repeated measurements of intrinsic membrane and synaptic properties in unambiguously identified neurons across many different preparations.

In vertebrates, identifying neuron types and characterizing their properties and connections on the basis of electrophysiological and anatomical criteria have been more difficult but have been somewhat successful in the spinal cord of lamprey, amphibians, and cat. Recently, progress has been made in applying new, sophisticated molecular techniques to the study of motor pattern generation, particularly in mice and zebra fish. Such approaches may allow the use of markers for transcription factors to identify groups of neurons and establish the circuit architecture underlying motor pattern generation.

In the following discussion, general principles are introduced that have emerged from the study of a wide range of invertebrate and vertebrate preparations. Mechanisms governing the function of CPGs, the dynamics of their cellular components and their network operations, as well as context-dependent activation and coordination are discussed.

Cellular and Network Mechanisms of Rhythm Generation

Rhythmic motor patterns almost always consist of sequential or alternating activation of antagonistic groups of muscles, and this sequence is seen in the timing of bursts of action potentials between groups of neurons in the underlying CPG. How this rhythmic activity is generated and how bursting and specific relative timing between different neurons is achieved depend on the circuit architecture, the properties of synaptic connections, and the intrinsic membrane properties of CPG neurons.

Synaptic Properties

Even though a wide variety of synapse types may be found in any given CPG, synchronization of neurons that fire in the same phase of the pattern is often achieved or aided by electrical coupling through gap junctions between these neurons, whereas alternating activity is often based on reciprocal inhibitory connections.

Many chemical synaptic interactions in CPGs are based on spike-mediated mechanisms; that is, transmitter is released in response to presynaptic action potentials. However, local interneurons often also rely on graded synaptic transmission; that is, they release transmitter as a graded function of presynaptic membrane potential. In general, it is not clear what the functional significance of graded versus spike-mediated transmission is in the specific context of rhythm generation. However, in a given circuit, the dynamics of these two types of transmission and their interaction with intrinsic neuronal membrane

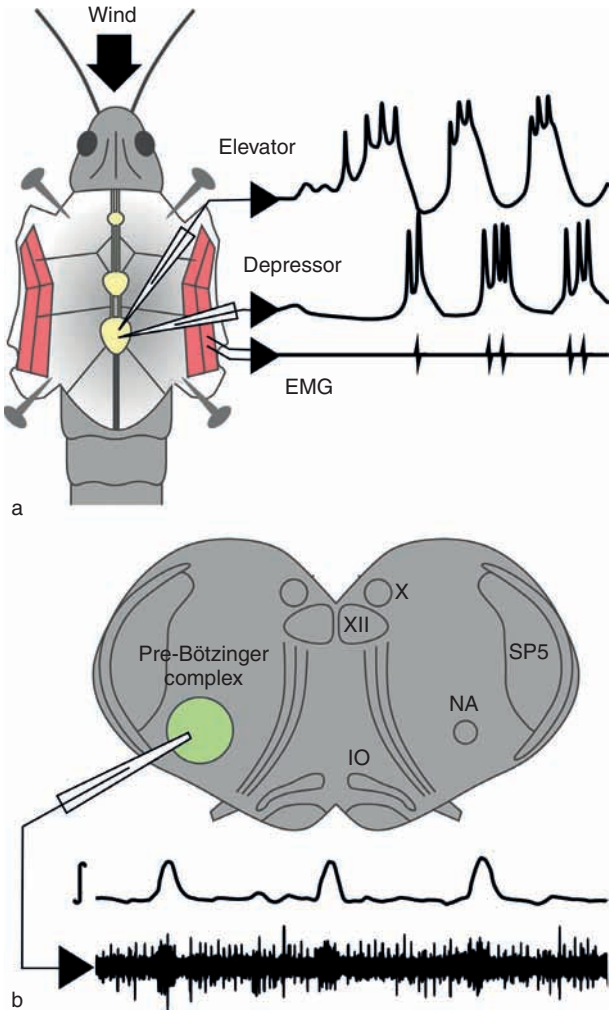

Figure 1 Generation of motor patterns in the absence of sensory feedback. (a) In the locust, rhythmic motor output to flight muscles can be generated in the deafferented thoracic nerve cord in an immobilized animal with wind stimulation to the head. Alternating activity between elevator and depressor motor neurons is generated even though no movement is produced and no sensory feedback reaches the central nervous system. (b) Transverse brain stem slice containing the central pattern generating neurons for respiration in the pre-Bötzinger complex. Population activity is recorded extracellularly (lower trace), and integrated (upper trace) to show rhythmic activity that corresponds to inspiration. XII, hypoglossal nucleus; X, dorsal motor nucleus of vagus; NA, nucleus ambiguus; SP5, spinal trigeminal nucleus; IO, inferior olive. (a) Adapted from Robertson RM and Pearson KG (1982) A preparation for the intracellular analysis of neuronal activity during flight in the locust. *Journal of Comparative Physiology* A 146: 311–320. (b) Adapted from Lieske SP, Thoby-Brisson M, Telqkamp P, and Ramirez JM (2000) Reconfiguration of the neural network controlling multiple breathing patterns: Eupnea, sighs and gasps. *Nature Neuroscience* 3: 600–607.

properties can be quite different, and CPGs exploit the possibilities of both for dynamic circuit operation.

Many synapses show considerable short-term plasticity such as depression, facilitation, or mixtures of both. In CPGs, synaptic dynamics can strongly influence the frequency of rhythmic network activity and the relative timing between different groups of neurons. This is because in rhythmic systems, depression and facilitation cause the strength of a synaptic connection to be dependent on rhythm frequency and on the duty cycle, the number of spikes, or the spike frequency of the presynaptic neuron. In the leech heartbeat CPG, the synaptic strength of connections from a single presynaptic neuron onto several postsynaptic neurons peaks at different times during a single presynaptic burst, as a result of different synaptic dynamics at these connections. Because of synaptic depression, some synaptic connections in the crustacean stomatogastric ganglion become weaker and peak later the faster the rhythm is, and this can help maintain stable relative timing between neurons at different rhythm frequencies.

Intrinsic Neuronal Properties

As in all neural networks, activity does not depend only on synaptic connections but also on a potentially rich complement of voltage-gated ion channels that determine a neuron's intrinsic membrane properties and therefore how it behaves in isolation and in response to synaptic input. A number of intrinsic properties are commonly found in CPG neurons (Figure 2). Among them are endogenous bursting properties, postinhibitory rebound, plateau properties, spike frequency adaptation, and restorative sag potentials.

Endogenous bursting means that a neuron can generate membrane oscillations in the absence of synaptic inputs (Figure 2(a)). This is an important mechanism of rhythm generation in a number of systems (see the section titled 'Pacemaker and network-based mechanisms').

In other neurons, bursting can be generated on rebound from inhibitory synaptic input (Figure 2(b)). This is an important mechanism for the production of alternating activity, and the time course of this rebound is an important determinant of relative timing between alternating rhythmic neurons.

The term plateau potential, or bistable membrane behavior, describes the ability of neurons to generate sustained firing in response to brief and relatively weak depolarizing input (Figure 2(c)). If an input crosses a certain threshold, firing activity outlasts the stimulus significantly and can either self-terminate eventually or be shut off by (again, relatively weak) hyperpolarizing input. Plateau properties can therefore uncouple neuronal firing from the duration and strength of excitatory input.

Many CPG neurons show spike frequency adaptation; that is, their firing slows down or ceases during sustained depolarization (Figure 2(d)). This can be an important mechanism to ensure that network activity

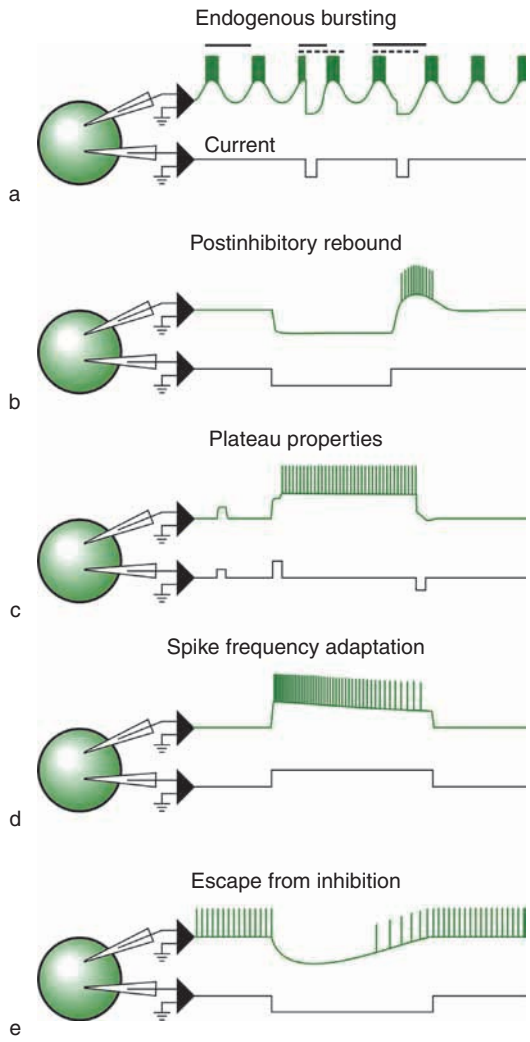

Figure 2 Intrinsic neuronal membrane properties in central pattern generators. Neuronal activity is critically dependent on intrinsic membrane properties that determine how a cell responds to synaptic input. In the absence of synaptic input, some commonly found membrane properties can be demonstrated by assessing the voltage responses to current injection. (a) Endogenous oscillators produce rhythmic activity in the absence of synaptic drive. Rhythmic activity in such neurons can be reset by current injection. Solid lines above the recording trace indicate the cycle duration, and broken lines indicate the unperturbed cycle duration. Whether the following cycle of bursting is advanced or delayed depends on both the sign and the phase of the current injection. The dependency on phase is an important characteristic of oscillators because it determines how synaptic input affects rhythmic activity. (b) Postinhibitory rebound is an overshooting response following hyperpolarizing input that can trigger a burst of action potentials. (c) Plateau potentials are persistent depolarizing responses to brief depolarizing input of sufficient amplitude. The depolarizing response either eventually self-terminates or can be terminated by brief and relatively weak hyperpolarizing input. (d) Spike frequency adaptation is a waning response to continuous depolarization of constant amplitude. Spiking activity can cease before the depolarizing input is shut off. (e) Escape from inhibition is a depolarizing response to hyperpolarizing input and can enable the neuron to eventually start firing action potentials again, despite continuing hyperpolarizing input.

does not get locked into the firing phase of neurons that inhibit functional antagonists.

Neurons that have hyperpolarization-activated inward currents can produce restorative sag potentials in response to inhibitory input; that is, they slowly start to depolarize during sustained hyperpolarization (Figure 2(e)). This is an important mechanism for neurons to escape from inhibition.

The ionic basis and the dynamics of the voltage-activated ion channels that are responsible for these properties are relatively well understood in a number of systems, at least to the point that most of the complement of ionic conductances have been identified, characterized, and attributed to certain functions. Some properties are mediated by similar conductances across many cell types and systems, whereas others can be based on a variety of different ionic mechanisms. Even though it is a reasonable approach to attribute certain aspects of a neuron's behavior to specific ionic conductances, especially if pharmacological block of this conductance abolishes this behavior, some caution is advisable. The influence that a specific ionic conductance has on a cell's behavior also depends on the overall complement of other conductances and the relative densities between them. In realistic computer models of neurons, changing the balance between the same set of conductances can change the activity patterns and responses to synaptic input radically.

Pacemaker and Network-Based Mechanisms

Where does the actual rhythmic activity originate? In principle, rhythm generation can be based on intrinsic oscillatory cellular properties or can emerge from synaptic interactions of neurons that are not intrinsically capable of generating rhythmic activity (Figure 3). In the first case, rhythmic network activity is based on the endogenous bursting properties of a core group of neurons that drive bursting in other cells (Figure 3(a)). Such neurons are referred to as endogenous bursters, oscillators, or pacemaker cells. These neurons synapse onto follower cells that burst as a consequence of rhythmic synaptic input but are tonically active or silent when synaptic input is removed experimentally. Follower neurons can have feedback synapses onto the pacemaker cells and can have rich synaptic interactions with other follower neurons (Figure 3(b)). Feedback synapses onto the pacemaker can play an important role in setting both frequency and the phasing of the pattern.

In other networks, rhythmic activity is generated in the absence of any cell-autonomous burst mechanisms and is the result of network connectivity. What may be the simplest form of network-based

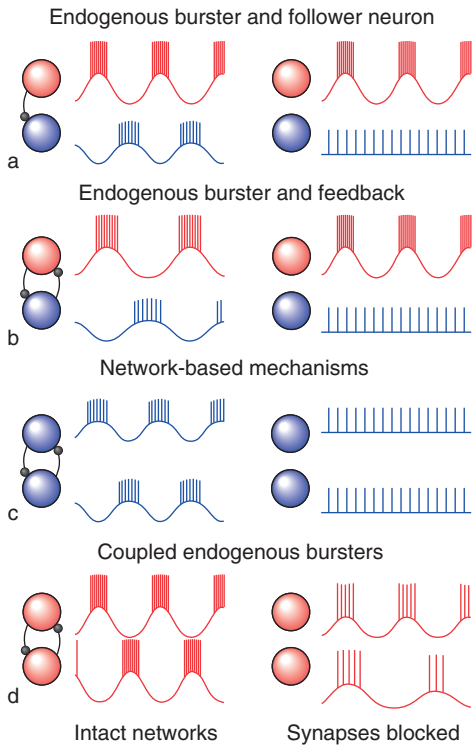

Figure 3 Network mechanisms of rhythm generation. The core mechanism of rhythm generation in a central pattern generator can be based either on endogenous bursting properties or on synaptic interactions of neurons that are not intrinsically capable of generating rhythmic activity. (a) Endogenous oscillators provide synaptic input to follower cells that allows the latter to burst rhythmically, in this case in rebound from inhibition. When synaptic interactions are blocked, the endogenous oscillator continues to cycle, whereas the follower neuron is either silent or fires tonically. (b) Feedback synapses from follower neurons to endogenous bursters can have significant influence on the pattern frequency and phasing. (c) A commonly found network-based mechanism of rhythm generation is reciprocal inhibition. This mechanism is often referred to as half-center oscillation. When synapses are blocked, both neurons are silent or fire tonically. (d) Endogenous bursters can be synaptically coupled to coordinate their bursting and/or stabilize regular rhythmic activity.

mechanisms of rhythm generation is the so-called half-center oscillation (Figure 3(c)). Here, reciprocal inhibition leads to alternating patterns of activity between neurons. The transitions from firing in one neuron to firing of the other neuron can be due to such things as release from inhibition when the active neuron shows spike frequency adaptation. Alternatively, the inactive neuron's intrinsic membrane properties can allow it to escape from inhibition, reach threshold, and in turn inhibit the other neuron.

It is not always straightforward to distinguish between cell-autonomous and network-based mechanisms. Some neurons have conditional oscillatory properties; their ability to generate rhythmic activity in synaptic isolation depends on the presence of neuromodulators and, in the intact system, can be dependent on the way the CPG is activated. In addition, pairs of neurons may have strong reciprocal synaptic connections that are essential for establishing robust and regular bursting with specific frequencies and relative timing but may also show endogenous bursting in synaptic isolation, even if this bursting is less robust (Figure 3(d)). In fact, a number of circuits show characteristics of both pacemaker-driven and network-based bursting. The circuit motif of reciprocal inhibition, often thought to be a hallmark of network-based rhythm generation, is also widespread in systems that are dominated by pacemaker neurons. In general, which mechanisms dominate may depend on context and how the circuit is activated. That being said, pacemaker-driven rhythmic networks control the mammalian respiratory rhythm and the pyloric rhythm of the crustacean foregut, two motor patterns that are continuously active. In contrast, network-based mechanisms appear to play the most important role in a range of circuits regulating locomotion, a behavior that is executed only in specific contexts.

Activation and Neuromodulation of CPGs

Most rhythmic motor behaviors can occur over a wide range of frequencies and with different durations and relative timing of activity between the muscles involved. For example, movement of a joint can be fast or slow, occur over different angles, and require different relative muscle activation for flexion and extension, dependent on the specific behavioral context. This flexibility is at least partly due to neuromodulators that potentially shape CPG activity into many different forms. The presence of neuromodulators in some cases is required even for the basic operation of a CPG, but in all cases it provides fine-tuning of rhythmic activity. Understanding CPG operation, therefore, requires understanding of not only circuit architecture and the intrinsic dynamics of neurons and synapses; in addition, it requires knowledge of the complement of neuromodulatory substances that affect a given circuit and the consequences that the presence of these substances has for circuit operation. This includes identifying which neurons in a CPG each substance targets and which receptors and ion channels in each cell type are affected. In addition, it is crucial to identify the source of modulatory substances and to know when and how they are released.

Complement and Source of Neuromodulators

The substances that act as neuromodulators can be very diverse and include classical transmitters such as γ-aminobutyric acid and glutamate, biogenic amines

such as dopamine and serotonin, a wide variety of peptides, and even gases such as nitric oxide. It is tempting to attribute certain global functions to a single modulator, but the idea that in a specific behavioral context, one modulator is released that changes the motor pattern in the desired way is unlikely to be true. In all systems where the effort has been made, a multitude of substances have been found, and it is more likely that in any given context, different combinations of these substances control CPG operation.

One of the most extensively studied systems in that respect is the stomatogastric ganglion (STG) of crustaceans, which contains CPGs that control the movement of different parts of the stomach. Here, almost 30 different substances have been identified that affect circuit operation (Figure 4(a)). These substances reach the STG through the circulatory system as neurohormones or are released into the neuropil by descending modulatory neurons or sensory afferents. Some of the substances are present in both delivery systems, presumably with different consequences because of differences in concentration and spatial distribution. In general, neurohormones affect circuits globally because they have access to the entire nervous system. In contrast, release from neuromodulatory terminals is locally restricted. That being said, paracrine release can lead to a fairly extensive diffusion of these substances, and many modulator actions are likely to be mediated through extrasynaptic receptors.

In some systems, even the CPG neurons themselves can have modulatory cotransmitters. This intrinsic modulation means that the properties of the circuit can be changed as a function of its own activity. In general, cotransmission is a widely found strategy in the modulation of CPGs. In the STG, different identified projection neurons that release the neuropeptide proctolin can elicit different forms of CPG activity. This is partly because they have different complements of cotransmitters that are released at the same time as proctolin.

Divergence and Convergence of Neuromodulator Action

Each modulator has a specific effect on circuit operation because it targets a specific subset of cells in the circuit. It may have different subcellular actions in each cell. In addition, different cells have different combinations of receptors to neuromodulators. Modulators can have a variety of effects in a single neuron (divergence), and different modulators can act on the same subcellular target in a given cell (convergence). In the pyloric circuit of the STG, all these mechanisms have been described.

Dopamine changes the activity of pyloric neurons, particularly the phase relationships, that is, the relative timing between them (Figure 4(b)). These effects are based on a wide variety of actions on synaptic and intrinsic neuronal properties (Figures 4(c) and 4(d)). Dopamine enhances the strengths of some chemical synapses and reduces the strength of others. In extreme cases, it can unmask synaptic connections that are not measurable in the absence of dopamine, and it can completely abolish others. In addition, it can alter the strength of electrical coupling between neurons.

Dopamine also affects a number of different voltage-dependent ion channels in pyloric neurons. It has several targets in each neuron, and these sets of targets differ across neurons. In addition, it can have a different effect on the same type of current in different neurons and either enhance or reduce it. In contrast to this divergence of dopamine actions, other substances converge onto the same target. Receptors for many different neuromodulatory substances are found in each pyloric neuron. A large portion of these, particularly the neuropeptides, act on the same voltage-dependent, nonselective, cation channel. In this case, the specificity of neuromodulator action results from different neurons in the circuit having different subsets of receptors.

Descending Commands and the Activation of Specific Patterns

What are the mechanisms of the decision-making process necessary to determine whether a specific motor pattern is elicited or not? This is particularly important in motor behaviors that are elicited only in specific behavioral contexts, but it is also true for specific different forms of continuously active motor output. In some cases, activating a single neuron, or a network of a few, can be sufficient to elicit an entire motor program.

The CPG for swimming in *Tritonia* can be activated by a pair of interneurons called dorsal ramp interneurons (DRI). A swim pattern is maintained as long as DRI is active. DRI is necessary and sufficient to elicit the swim pattern and therefore satisfies the rigorous definition of what is commonly called a command neuron. It receives rhythmic feedback from the CPG and therefore fires rhythmically during fictive swimming. However, it is not part of the CPG, as hyperpolarizing it during a swim pattern that is elicited by electrical stimulation of the CPG does not disrupt CPG activity. It receives excitatory input from sensory neurons and a trigger neuron, Tr1 (Figure 5(a)). A brief stimulation of this trigger neuron is sufficient to elicit a swim pattern that outlasts the stimulus substantially. However, DRI can also elicit the swim pattern when Tr1 is hyperpolarized, so Tr1 is sufficient but not necessary in triggering swimming.

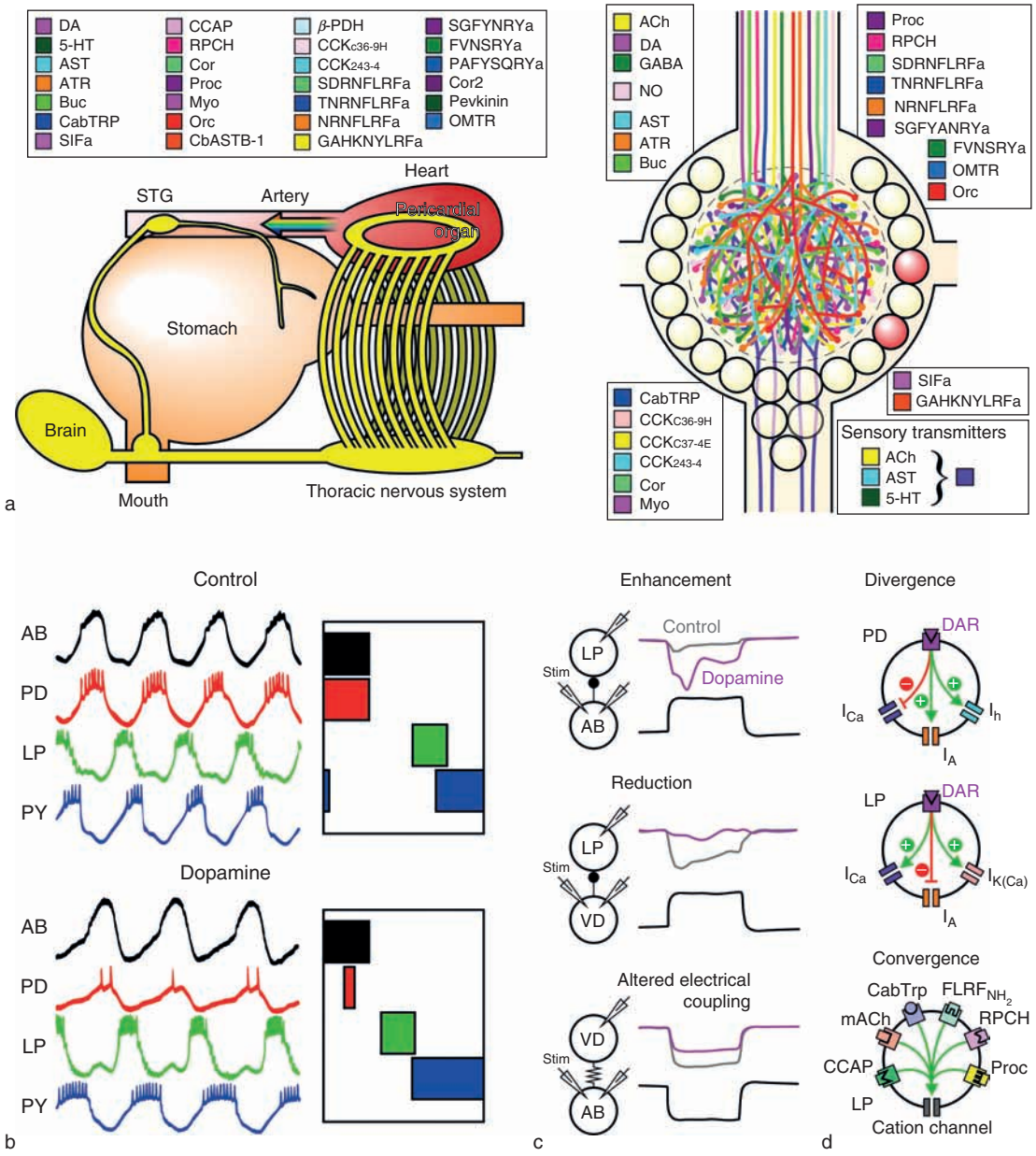

Figure 4 Neuromodulation of central pattern generators (CPGs). (a) The stomatogastric ganglion is modulated by many substances. These substances are either released into the circulatory system (mostly by the pericardial organs) and reach the stomatogastric ganglion (STG) as neurohormones or are released into the neuropil by descending modulatory neurons. (b) Dopamine changes the activity and the phase relationships (relative timing) of CPG neurons in the pyloric circuit of the stomatogastric ganglion. Intracellular recordings show the activity of neurons representative for different phases in the triphasic pyloric rhythm. The bar plots show the activity phases of these neurons in one normalized cycle. (c) Dopamine enhances or reduces chemical synaptic connections and also alters electrical coupling between pyloric neurons. The presynaptic neuron was electrically stimulated and the voltage response in pre- and postsynaptic neuron recorded. Note that chemical synapses are graded. (d) Convergence and divergence of neuromodulator action. Dopamine acts on various voltage-gated ion channels in pyloric neurons. It can act on different subsets of these channels in different neurons and affect the same channel type in different ways in different neurons. In contrast, many modulators, predominantly neuropeptides, activate the same voltage-dependent nonselective cation channel in pyloric neurons. AB, anterior burster neuron; ACh, acetylcholine; AST, allatostatin; ATR, allatotropin; Buc, buccalin; CabTRP, cancer borealis tachykinin-related peptide; CbASTB-1, cancer borealis allatostatin B-1; CCAP, crustacean cardioactive peptide; CCK, cholecystokinin; CCK, cholecystokinin; DA, dopamine; DAR, dopamine receptor; 5-HT, serotonin; GABA, gamma-aminobutyric acid; IA, A-type potassium current; ICa, calcium current; Ih, hyperpolarization-activated inward current; ICa, calcium current; LP, lateral pyloric neuron; Myo, myomodulin; OMTR, orcomyotropin; Orc, orcokinin; NO, nitric oxide; PD, pyloric dilator neuron; PDH, pigment-dispersing hormone; PY, pyloric neuron; VD, ventral dilator neuron; xProc, proctolin; xRPCH, red pigment concentrating hormone; XXa, [peptide sequence]amid. (a) Adapted from Marder E, Bucher D, Schulz DJ, and Taylor AL (2005) Invertebrate central pattern generation moves along. *Current Biology* 15: R685–R699. (b) Adapted from Harris-Warrick RM, Coniglio LM, Levini RM, Gueron S, and Guckenheimer J (1995) Dopamine modulation of two subthreshold currents produces phase shifts in activity of an identified motoneuron. *Journal of*

Such a clear role for descending neurons as in the case of DRI has been found in a number of systems, but it is most likely more common that CPGs are controlled by networks of higher-order neurons and modulatory projections. In that case, different forms of network output can be generated dependent on the relative contribution of different higher-order neurons. In *Aplysia*, the feeding CPG can generate output that is coordinated in two essentially different ways (Figure 5(b)). One pattern corresponds to a motor pattern that is used during the ingestion of food. Here, the radula, the mouth part apparatus of mollusks, is protracted while it is open and retracted after a bite, while it is closed. The other pattern corresponds to the rejection of food, when the radula is protracted while it is closed, and retracted while it is open. The choice between these patterns is biased by the relative activity of descending neurons. In the example shown in Figure 5(b), an ingestive pattern is generated when two interneurons, CBI-2 and CBI-3, are activated simultaneously. An egestive (rejection) pattern is generated when CBI-2 is activated while CBI-3 is silent.

In vertebrates, spinal CPGs are subject to modulation by a variety of substances that activate and shape the output patterns they produce. These inputs stem from descending neurons and intraspinal systems. In addition, activation of spinal CPGs is strongly controlled by inhibition from higher centers. An important role in this is played by the pallidum, the output layer of the basal ganglia. The pallidum in turn receives inhibition from the striatum, which is activated by cortical and thalamic inputs. Removal of inhibition in general is an important mechanism of motor program selection. However, to what extent this is a direct influence on the CPGs or rather affects the command circuits that activate CPGs is not clear in most cases.

Coordination of CPGs

Modular Organization of Rhythmic Motor Behaviors

How many CPGs are there? Rather than having one dedicated network for a relatively complex motor behavior like swimming or walking, central pattern generation is inherently modular. 'Unit oscillators' control antagonistic muscle activation for the movement of a single segment or a single joint in an appendage. These rhythm-generating networks need to be coordinated in specific ways to produce specific types of behaviors. For example, swimming movements in lamprey and leech are based on right-left alternation of activity at the segmental level, and this activity can be elicited in reduced preparations with only one or a few segments of the isolated nervous system intact. However, the undulatory movement of the whole body depends on coordination of these segmental patterns along the body axis, in these cases a fixed anterior–posterior phase lag that ensures a sine wave-like movement with a wavelength that matches the body length (Figure 6(a)). In the lamprey, this coordination is largely due to central interneuronal coupling so that even the isolated spinal cord displays an intersegmental phase lag similar to that found in the intact animal.

Coordination through Central Coupling or Sensory Feedback

In principle, several mechanisms can be involved in coordinating unit CPGs (Figure 6(b)). Apart from central interneuronal coordination, movement generated in response to motor commands in turn produces sensory (e.g., proprioceptive) feedback that can aid in coordinating adjacent oscillators in several ways. In some cases, sensory feedback from one part of the body has access to CPGs controlling adjacent parts of the body. In other cases, movement of one part of the body is transferred via mechanical coupling to other parts of the body, which in turn provide sensory feedback to the local oscillators. In the leech, the intersegmental phase lag during fictive swimming in the isolated nerve cord is reduced in comparison with intact animals. In the intact leech, sensory feedback has been shown to play an additional role in coordinating segmental oscillators. Coordination is even found across segments between which the nervous system had been severed, so that no direct neuronal communication was possible. Therefore, this interaction must have been provided through mechanical coupling.

Sensory feedback plays a particularly important role in shaping motor output and coordinating unit CPGs in behaviors that are heavily dependent on cycle-by-cycle corrections. Terrestrial locomotion takes place in a considerably less-constant medium than do swimming or flying. Therefore, it is no

Neurophysiology 74: 1404–1420. (c) Adapted from Johnson BR, Peck JH, and Harris-Warrick RM (1995) Distributed amine modulation of graded chemical transmission in the pyloric network of the lobster stomatogastric ganglion. *Journal of Neurophysiology* 74: 437–452; and Johnson BR, Peck JH, and Harris-Warrick RM (1993) Amine modulation of electrical coupling in the pyloric network of the lobster stomatogastric ganglion. *Journal of comparative Physiology* 172: 715–732. (d) Adapted from Marder E and Bucher D (2007) Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. *Annual Review of Physiology* 69: 291–316.

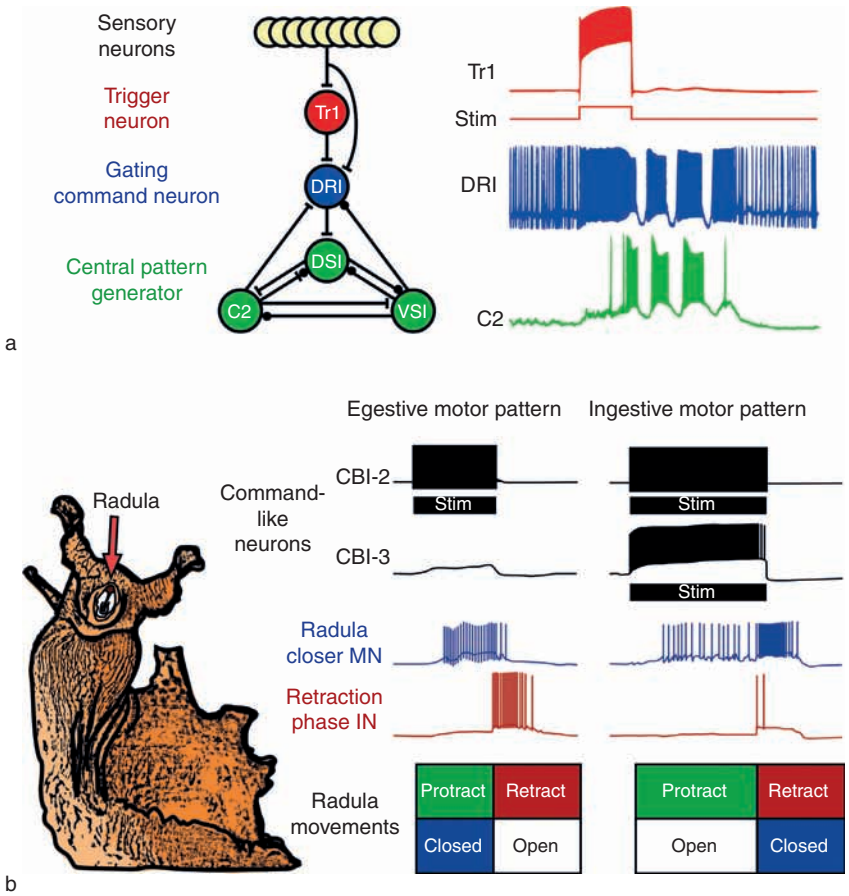

Figure 5 Motor pattern selection. (a) In the *Tritonia* swim system, activation of dorsal ramp interneurons (DRI) is necessary and sufficient to elicit a swim motor pattern. DRI receives excitatory input from sensory neurons and the trigger neuron Tr1. Transient stimulation of Tr1 elicits a swim pattern that outlasts Tr1 activity. (b) In the *Aplysia* feeding system, different relative firing activity of commandlike neurons biases the central pattern generator to express either an egestive or an ingestive motor pattern. These motor patterns correspond to the rejection of food, when the radula is closed during protraction and open during retraction, and to the ingestion of food, when the radula is open during protraction and closed during retraction. When the interneurons CBI-2 and CBI-3 are activated simultaneously, an ingestive motor pattern is elicited. When only CBI-2 is active, an egestive pattern is elicited. (a) Adapted from Frost WN, Hoppe TA, Wang J, and Tain L-M (2001) Swim initiation neurons in *Tritonia diomedea*. *American Zoologist* 41: 952–961. (b) Adapted from Morgan PT, Jing J, Vilim FS, and Weiss KR (2002) Interneuronal and peptidergic control of motor pattern switching in *Aplysia*. *Journal of Neurophysiology* 87: 49–61.

coincidence that much effort has gone into understanding the sensory control of walking. In addition, walking requires a rather complex set of coordinating mechanisms (Figure 6(c)). In a single leg joint, antagonistic muscles need to be coordinated to produce alternating contractions. This alternating activity has to be coordinated across several joints to produce alternating stance and swing, that is, a phase in which the leg is on the ground and propels the body, and a phase in which it is off the ground and moves forward to initiate the next stance. Each step sequence has to be coordinated with the step sequences of the contralateral leg and legs in other segments to produce a specific gait. In the stick insect, CPGs controlling different leg joints are operating largely independently in the absence of sensory feedback. Proprioceptors that measure movement in one joint, and mechanoreceptors that measure load or cuticle strain, can entrain and reset CPGs controlling adjacent joints. In contrast, the motor output of the deafferented lumbar spinal cord of the cat resembles stepping movements of the hindleg, with the correct sequence of activation of motor pools controlling hip, knee, ankle, and foot. However, this motor pattern is drastically modified by sensory feedback from predominantly muscular sense organs measuring movement and force.

In addition to modular CPGs that are coordinated for a specific behavioral task, many behaviors involve movements of different body parts that are not part of the same cyclic repetition of sequential muscle activations and run at different speeds. Such coordination

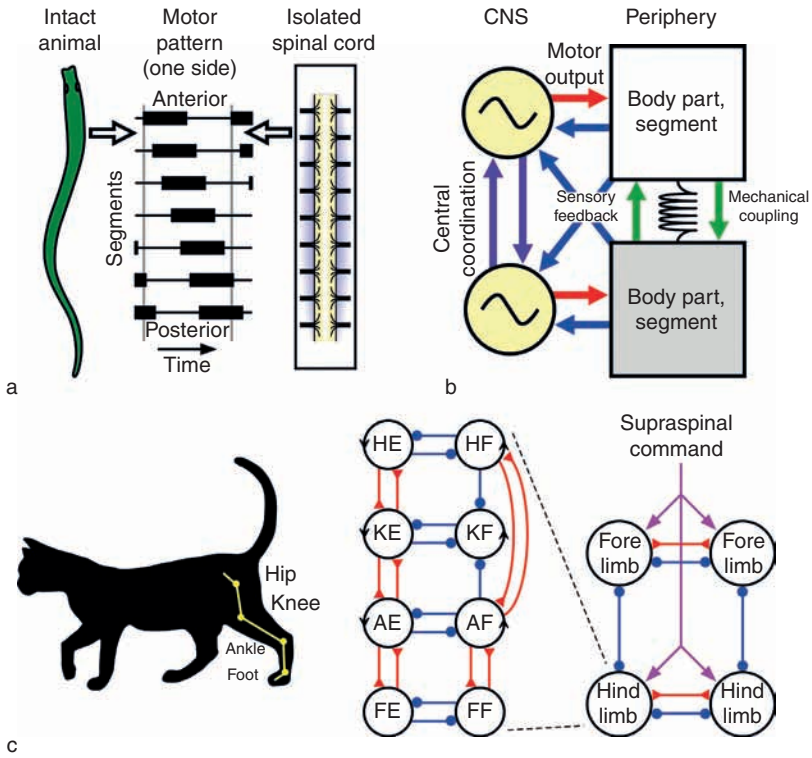

Figure 6 Coordination of motor patterns in different parts of the body. (a) Lamprey swimming consists of a mechanical wave that is transmitted along the anterior–posterior body axis. Alternating right–left activity is expressed with a fixed phase lag between adjacent segments. This pattern of coordination is also seen in the isolated spinal cord during fictive swimming. (b) Schematic of possible coordinating influences between local oscillators. CNS, central nervous system. (c) Coordination of joints and limbs during walking. Alternating extensor (E) and flexor (F) activity is produced at each leg joint (H, hip; K, knee; A, ankle; F, foot), and all joints are coordinated to produce stance and swing phase of the single leg. All limbs are coordinated to produce different gates when activated by supraspinal commands. For example, walking involves mutual inhibitory interaction between right and left, whereas excitatory interaction underlies galloping. (a, c) Adapted from Grillner S (2006) Biological pattern generation: The cellular and computational logic of networks in motion. *Neuron* 52: 751–766.

may be important to avoid mechanical interference between different parts of the body or to tie related behaviors into the same context of operation. For example, the mammalian respiratory rhythm can be coupled to the locomotor pattern during walking or running so that the duration of a respiratory cycle is an integer multiple of the locomotor cycle.

Flexibility of Motor Pattern Coordination

What is the advantage of a distributed, modular arrangement of CPGs? The absence of a dedicated circuitry for a complete behavior means that unit pattern generators can be coordinated in different ways to produce different motor output. We can walk, run, jump, and dance, behaviors that all require alternating activity of antagonistic muscles controlling our leg joints, but with different temporal relationships between different joints and legs. Mammals can walk or gallop, the former requiring alternating activity between left and right and the latter requiring synchronous activation. In the lamprey, the intersegmental phase lag shown in Figure 6(a) represents the appropriate coordination for forward swimming. This pattern can be reversed to allow backward swimming. It is interesting that this reversal can also be observed in the isolated spinal cord when the more posterior segments are stimulated appropriately.

Multifunctional Neurons and Circuits

The concept of pattern-generating modules seems to imply that the circuitry that is responsible for controlling local movements is exclusively dedicated to this one specific task. However, there is evidence from a number of systems that shows that this is not true. In the leech nervous system, almost all neurons that are active rhythmically during fictive swimming are also activated during fictive crawling. These motor patterns are mutually exclusive, and the activity phase of particular neurons in the locomotor cycle can be very different between both motor patterns. In the brain stem, pattern generation for breathing, gasping, and possibly other related behaviors like swallowing and coughing share some of the neuronal

elements. In the crustacean stomatogastric ganglion, CPG neurons can simultaneously express frequency components of different rhythms that control different parts of the stomach and run at different speeds. Some of these neurons can even completely shift coalitions and, dependent on conditions, can be integral parts of different CPGs at different times.

CPGs as Test Beds for General Principles of Circuit Function and Dynamics

Obviously, analysis of CPG function plays an important role in understanding the neural basis of many motor behaviors and has important implications for the understanding of motor function-related disorders and central nervous system injuries. In addition, studies of CPGs have also elucidated important general principles of network function. Rhythmic firing is a common, if not the prevalent, form of network activity in the brain, spanning systems functioning in a wide range of contexts, from sensory processing to cognitive tasks. Many of the principles that govern these brain functions may be shared with those found to underlie rhythmic motor pattern generation. Furthermore, much of our understanding of circuit dynamics comes from studies of CPGs. This includes both general principles of neuromodulation and important concepts of long-term homeostatic regulation of neuron and network properties. One reason for the successful use of CPGs in studies of network dynamics is that regular rhythmic motor output is relatively easy to measure and quantify. Therefore, CPGs are tractable systems for analyzing the consequences of pharmacological, genetic, or other experimental manipulations on network function.

See also: Autonomic Nervous System: Respiratory Control; Autonomic Nervous System: Central Respiratory Control; Behavioral Hierarchies; Brainstem Respiratory Circuits; Central Pattern Generators: Sensory Feedback; Command Systems; Hormones and Behavior; Neuromodulation; Pattern Generation; Respiration; Walking in Invertebrates.

Further Reading

Büschges A (2005) Sensory control and organization of neural networks mediating coordination of multisegmental organs for locomotion. *Journal of Neurophysiology* 93: 1127–1135.

Dickinson PS (2006) Neuromodulation of central pattern generators in invertebrates and vertebrates. *Current Opinion in Neurobiology* 16: 604–614.

Feldman JM and Del Negro CA (2006) Looking for inspiration: New perspectives on respiratory rhythm. *Nature Reviews Neuroscience* 7: 232–242.

Frost WN, Hoppe TA, Wang J, and Tian L-M (2001) Swin initiation neurons in. *Tritonia diomedea*. *American Zoologist* 41: 952–961.

Grillner S (2003) The motor infrastructure: From ion channels to neuronal networks. *Nature Reviews Neuroscience* 4: 573–586.

Grillner S (2006) Biological pattern generation: The cellular and computational logic of networks in motion. *Neuron* 52: 751–766.

Harris-Warrick RM, Coniglio LM, Levini RM, Gueron S, and Guckenheimer JE (1995) Dopamine modulation of two subthreshold currents produces phase shifts in activity of an identified motoneuron. *Journal of Neurophysiology* 74: 1404–1420.

Johnson BR, Peck JM, and Harris-Warrick RM (1993) Amine modulation of electrical coupling in the pyloric network of the lobster stomatogastric ganglion. *Journal of Comparative Physiology A* 172: 715–732.

Johnson BR, Peck JH, and Harris-Warrick RM (1995) Distributed amine modulation of graded chemical transmission in the pyloric network of the lobster stomatogastric ganglion. *Journal of Neurophysiology* 74: 437–452.

Kiehn O (2006) Locomotor circuits in the mammalian spinal cord. *Annual Review of Neuroscience* 29: 279–306.

Kristan WB Jr., Calabrese RL, and Friesen WO (2005) Neuronal control of leech behavior. *Progress in Neurobiology* 76: 279–327.

Kupfermann I and Weiss KR (2001) Motor program selection in simple model systems. *Current Opinion in Neurobiology* 11: 673–677.

Lieske SP, Thoby-Brisson M, Telqkamp P, and Raminez JM (2000) Reconfiguration of the neural network controlling multiple breathing patterns: Eupnea, sights and gasps. *Nature Neuroscience* 3: 600–607.

Marder E and Bucher D (2001) Central pattern generators and the control of rhythmic movements. *Current Biology* 11: R986–R996.

Marder E and Bucher D (2007) Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. *Annual Review of Physiology* 69: 291–316.

Marder E, Bucher D, Schulz DJ, and Taylor AL (2005) Invertebrate central pattern generation moves along. *Current Biology* 15: R685–R699.

Morgan PT, Jing J, Vilim FS, and Weiss KR (2002) Interneuronal and peptidergic control of motor pattern switching in *Aplysia*. *Journal of Neurophysiology* 87: 49–61.

Nusbaum MP and Beenhakker MP (2002) A small-systems approach to motor pattern generation. *Nature* 417: 343–350.

Pearson KG (2004) Generating the walking gait: Role of sensory feedback. *Progress in Brain Research* 143: 123–129.

Ramirez JM and Viemari JC (2005) Determinants of inspiratory activity. *Respiratory Physiology & Neurobiology* 147: 145–157.

Robertson RH and Pearson KG (1982) A preparation for the intracellular analysis of neuronal activity during flight in the locust. *Journal of Comparative Physiology A* 146: 311–320.

Central Pattern Generators: Sensory Feedback

W O Friesen, University of Virginia, Charlottesville, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Originating near the beginning of the twentieth century, two seemingly incompatible theories informed much of the debate concerning the neural origins of rhythmic movements in animals. One theory, put forth most clearly by Charles Sherrington, held that locomotory patterns, although coordinated by the central nervous system (CNS), arise from sensory feedback and reflexes. Although counterexamples were offered by several investigators, this theory was widely accepted before the 1960s. The second theory, ably proposed by T Graham Brown, held that central pattern generators (CPGs) formed by inhibitory neuronal circuits within the CNS are the source of rhythmic movement patterns. The latter theory gained renewed support from the work of Donald Wilson on locust flight in 1961 and finally became dogma during subsequent decades as more and more CPG circuits were identified in both vertebrates and invertebrates. Near the end of the twentieth century, the importance of sensory feedback for the expression of efficient locomotion received new recognition, leading to a convergence of the two theories. The consensus is that the neuronal circuits underlying locomotion comprise a distributed system that includes both central oscillators and sensory feedback components.

The reemerging recognition that sensory feedback plays an essential role in animal locomotion (and other rhythmic movements) includes the realization that sensory neurons exhibit many of the characteristics that define elements of a CPG. Thus it has now been demonstrated that sensory neurons (1) express rhythmic activity during locomotion, (2) make synaptic contacts with CPG interneurons, and (3) can entrain the CPG when they are stimulated rhythmically. Systems that provide examples of these characteristics include locust flight, swimming in leeches and lamprey, crayfish swimmeret beating, and walking in stick insects and crayfish.

What Is Sensory Feedback?

The neuronal activity patterns that underlie rhythmic movements in a wide variety of animal species are generated by complex neuronal circuits, the CPGs, located within the CNS. In addition, a multitude of sensory receptors, found in the joints, muscles, and other structures of the periphery, detect internal joint positions and movements as well as environmental factors. The overt movements that constitute animal behavior arise from the interplay between the CPGs and sensory feedback from these body structures. Because animals live in environments that are characterized by both temporal change and spatial unpredictability, the instinctive, inherited CPG control circuits cannot be expected to, and do not in fact, generate movement patterns that ensure adaptive behavior. Consequently, CPGs underlying all animal movements are subject to inputs via receptors that monitor the execution of motor commands in relationship to the environment and then provide inputs to motor neurons and to the central neuronal circuits that form the essential core of CPGs. Such inputs modify the neuronal activity patterns arising from the CPG with critical effects on movement amplitude, phase relationships, and the cycle period (or, equivalently, the frequency) of movements. The focus here is on the relationship of CPGs and sensory feedback in a few well-studied invertebrates and a primitive fish, the lamprey. A schematic overview of one such system, which controls the swimming movement in the medicinal leech, illustrates the general properties of the interactions between CPGs, motor output, and sensory feedback (Figure 1).

Why CPGs Are Subject to Sensory Feedback

In many animal preparations, the output of the CPG can be observed in greatly reduced preparations, which often include only parts of the CNS. Comparisons of CPG neuronal motor output with activity patterns observed in intact, behaving animals provide the basis for the concept of a CPG, and, more germane here, the comparisons have revealed deficits in motor patterns that arise when sensory feedback is absent. Most prominent, and most important, among such deficits are that neuronal activity patterns generated by CPGs in isolation have increased cycle periods, reduced activity amplitudes and suboptimal phase relationships. Moreover, in the absence of sensory cues, or internal changes in state, motor output tends to represent rather simple movements, such as straight-ahead locomotion, rather than the more complex movements observed in the intact animal.

Unpredictable development There are numerous factors that mandate sensory feedback if animals are to behave adaptively, including body growth during development and maturation. Amplitude and frequency of movements alter with growth; hence, one major consideration is the matching of CPG output to

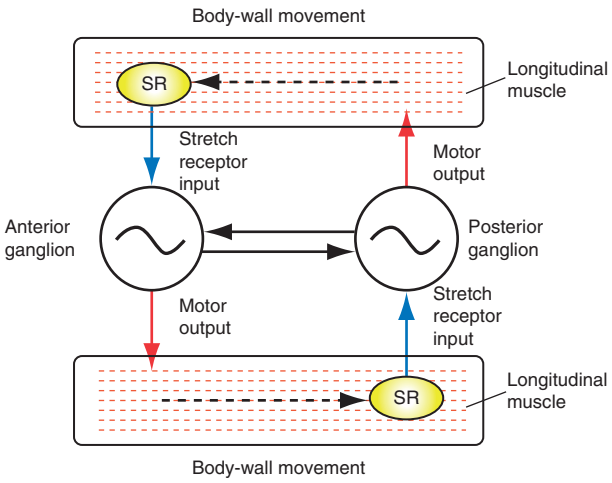

Figure 1 Schematic model for the neuronal control of swimming in the leech. Two segmental central pattern generators (CPGs) within ventral nerve cord ganglia are represented by unit oscillators (∼). These unit oscillators are interconnected and mutually entrained by intersegmental axons that run through nerve cord connectives. The cycle period of the CPG is greater than in the intact animal, whereas intersegmental phase lags are smaller. Segmental motor neurons driven by the CPGs control muscle contraction in their individual segments (red arrows from CPG). When swimming movements are generated by the intact animal, tensions in the body wall, arising from local motor neuron output, mechanical coupling between body wall segments (dashed arrows), and the environment provide sensory feedback (blue arrows) to the CPG. This rhythmic input acts to decrease cycle period and to increase intersegmental phase lags in the intact leech. Symbols: SR, stretch receptor; ∼, CPG. Reprinted from Yu X and Friesen WO (2004) Entrainment of leech swimming activity by the ventral stretch receptor. *Journal of Comparative Physiology* 190: 939–949, figure 6, © 2004 Springer-Verlag, with kind permission of Springer Science and Business Media.

the natural, resonant frequency of limb or body. Although development changes in the CPG and the body structure may be coordinately regulated by genetics, maturation processes depend also on the environment. With the latter not fully predictable, sensory feedback that modifies CPG frequency, phase, and amplitude provides a means of ensuring that body movements are efficient at all stages of growth.

Unpredictable environment Animals live in a wide variety of habitats. Some, like fish or flying insects and birds, move in nearly homogeneous media; nevertheless, obstacles, turbulence, or temperature clines are unpredictable, setting up the essential requirement for sensory input to the CPG, even for swimming and flight. Terrestrial environments exhibit a very high degree of unpredictability; hence it is not surprising that CPGs underlying walking and stepping are even more strongly controlled by sensory inputs.

Compensation for morphological alterations Further requirements for sensory feedback arise from alterations in an animal's form through feeding, accident, or predation. A particularly telling example comes from walking locomotion in insects, which modify their gait immediately when a leg is experimentally amputated. Other examples include seasonal size and weight alterations that require a reprogramming of the motor system but cannot be expected to alter the CPG. Although hormones can reprogram the nervous system to match an animal's activity to its overall condition, only sensory feedback can provide cycle-by-cycle feedback from the environment.

Components of Motor Systems

Motor systems, whether generating the neuronal substrates of locomotion or other rhythmic movements, comprise several elements. Central to these systems are one or more oscillator circuits that generate the rudiments, in period and phase, of the expressed motor activity. Excitatory inputs from higher centers, from sensory input, or from within the circuits themselves stimulate CPG output to pass from a 'resting' stable state, that is, without oscillations, to one that is unstable and hence oscillates with a frequency that may or may not match that observed in the intact animal. Most CPGs are formed entirely, or nearly so, of interneurons. However, motor neurons may constitute most of the CPG, as in the stomatogastric system of crustacea; they may be weakly involved in pattern generation, as in the leech swim CPG; or they may merely serve as followers to drive muscle contractions, as in the lamprey swim circuits. Sensory feedback arises from stretch receptors in muscles, proprioceptors, sensory hairs as in crustacean swimmerets, cuticle receptors that signal load in insects, and stretch receptors in the CNS itself, as in lamprey. These receptors may be spiking or nonspiking, but their function is to provide position, load, velocity, and angle information to the CPG – afferent information that informs the CPG whether the movement commanded is appropriately executed. Alternatively, sensory input modifies CPG function more drastically, as during animal walking, in which sensory processes control phase relationships among limbs and the intensity of motor neuron output and may advance, retard, initiate, or prevent phase transitions.

Experimental Approaches: Techniques and Animal Preparations

Experiments to determine the nature of the sensory receptors and their role in CPG modulation have

encompassed a wide range of animal preparations and experimental approaches.

Experimental Approaches and Techniques

Comparison of CPG output and movements of intact animals One methodology employed nearly universally to assess the relative importance of CPG circuits and sensory feedback in the control of animal movements is to compare the neuronal activity patterns expressed in reduced preparations with those of intact, behaving animals. Such comparisons reveal four major aspects of rhythmic activity that are modified by sensory feedback: cycle period, activity amplitude, phase relationships among local outputs, and intersegmental phase relationship. There is a wide range in the relative strength of sensory feedback effects, as illustrated in Figure 2, which shows that cycle periods of fictive (isolated CNS) swimming activity in both leech and lamprey are significantly greater than in the intact animal. However, phase relationships remain unchanged by removal of the periphery in the lamprey whereas intersegmental phase lags are reduced by nearly 50% when sensory feedback is eliminated in leeches.

Sensory entrainment A second widely used technique for determining the nature and strength of sensory feedback is to perform entrainment experiments. With this approach, either the body itself (in semi-intact preparations) or sensory receptor membrane potential is modulated rhythmically by the experimenter while the CPG is generating rhythmic output. For many systems, movements or direct current injection into sensory neurons can force the CPG to generate cycle periods that are up to ±50% of the unperturbed values. Figure 3 illustrates the entrainment technique as it is applied to the swimming rhythm in lamprey. Here physical movements of the rostral or caudal ends of a spinal cord/notochord preparation entrain the CPGs as the movement frequency is set to values both below and above those detected when movements are absent. Such experiments demonstrate that sensory structures, in this case edge cells, have access to the CPG and can modify its properties. Related, alternative approaches are to impede limb or body movements; place a limb into a particular, fixed position; or assess the effects of brief current injection into sensory neurons. Like the entrainment procedure, these alternative methods provide useful information about both the nature and the strength of sensory input in the expression of rhythmic movements.

Animal Preparations

Six preparations and behaviors are described here to illustrate current procedures of, and results from, studies on the relationships between CPG-generated rhythms and sensory inputs during the expression of rhythmic movements. These behaviors are flying in locusts, swimming in leeches and lamprey, swimmeret beating in crayfish, and walking in stick insects and crustacea.

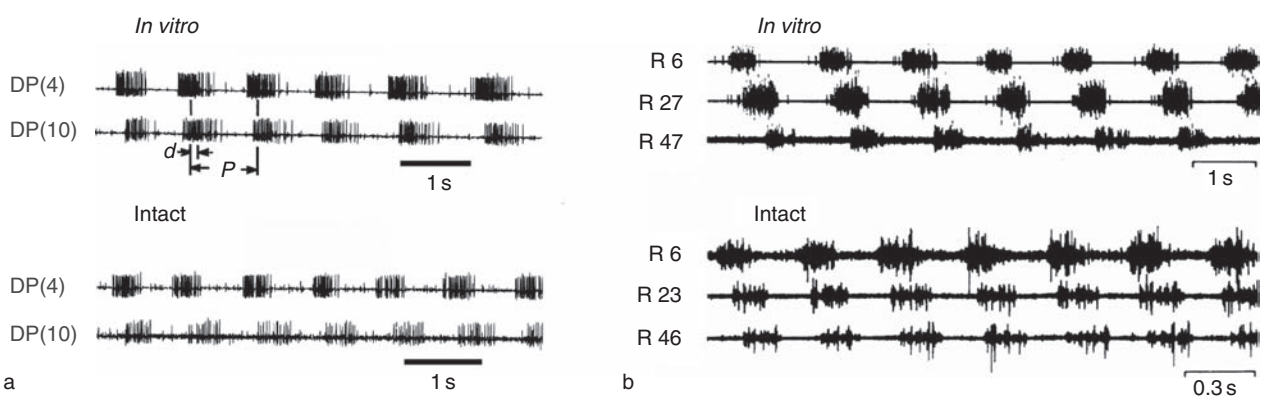

Figure 2 Activity patterns during swimming locomotion: patterns generated by CPGs compared with those expressed by nearly intact animals. (a) Leech swimming. The upper traces (*in vitro*) were obtained from a preparation consisting of the isolated leech ventral nerve cord with electrodes attached to the dorsal posterior (DP) nerves of segments 4 and 10. The lower traces (intact) were obtained from electrodes implanted onto the DP nerves of segments 4 and 10 during swimming in a nearly intact leech. The bursts in these records are axon impulses of cell DE-3, a motor neuron that innervates dorsal longitudinal muscles. P is the cycle period, and d is the intersegmental delay. (b) Lamprey swimming. The upper traces (*in vitro*) depict ventral root activity during fictive swimming. The lower traces (intact) show electromyograph records obtained from an intact swimming lamprey. Letters R and numerals indicate that recordings are from the right side at the segments indicated. Fictive swimming was elicited by bath application of 0.5 mmol l^{-1} D-glutamate. Note the different time calibrations. (a) Reproduced from Pearce RA and Friesen WO (1984) Intersegmental coordination of leech swimming: Comparison of *in situ* and isolated nerve cord activity with body wall movement. *Brain Research* 299: 363–366, with permission from Elsevier. (b) Reprinted from Wallén P and Williams TL (1984) Fictive locomotion in the lamprey spinal cord *in vitro* compared with swimming in the intact and spinal animal. *Journal of Physiology* 347: 225–239, with permission from Blackwell Publishing.

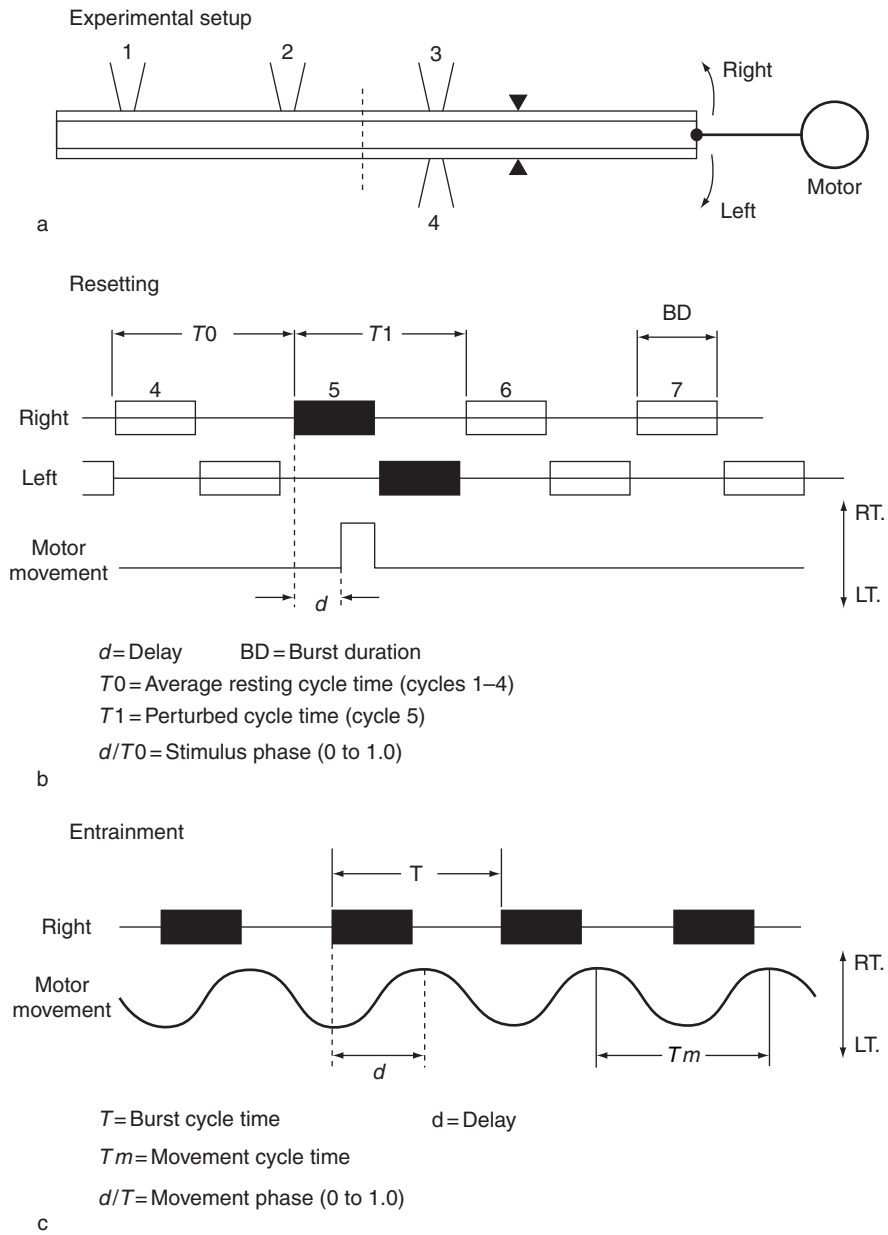

Figure 3 Schematic diagram for entrainment experiments in lamprey. (a) Diagram of the *in vitro* notochord/spinal cord preparation. The caudal (right (RT.)) end of the preparation was rhythmically deflected with a pen motor while the preparation was immobilized to the left (LT.) of the arrowheads. (b) Step inputs. Brief movements (300 ms) of the caudal end at selected phases in the swim cycle were used to determine phase changes (as a fraction of the cycle period) to generate phase response curves (not shown). (c) Entrainment methodology. Sinusoidal movements with a range of cycle periods (T) were applied to the preparation; cycle period and the phase relationship (d/T) between the imposed signal and the fictive motor output were measured. Reprinted from McClellan AD and Jang W (1993) Mechanosensory inputs to the central pattern generators for locomotion in the lamprey spinal cord: Resetting, entrainment, and computer modeling. *Journal of Neurophysiology* 70: 2442–2454; used with permission.

Locust flight Locust flight was the first behavior for which the central role of a CPG was widely accepted. Ironically, locust flight is also among the first clear examples of sensory feedback coordinating animal movements in a well-defined system. Comparisons of flying in nearly intact but tethered animals with the output of the flight CPG in the absence of sensory input reveal the following differences. Wingbeat frequency in tethered flying locust is approximately 15–20 Hz, whereas deafferented preparations exhibit frequencies about half of those values. Moreover, nearly intact tethered locusts stimulated with air currents in a wind tunnel fly continuously, whereas reduced preparations lacking sensory feedback soon terminate the CPG-generated pattern. Stimulation of forewing stretch receptors alters the expressed

rhythm in several ways. First, repeated stimulation phase-locked to the flight pattern increases wingbeat frequency. Second, such stimulation prolongs the expression of the motor pattern. These effects are phase-dependent, indicating that they function on a cycle-by-cycle basis rather than through general excitation. Third, stimulation of both forewing stretch receptors at frequencies near the wingbeat frequency entrains the flight rhythm, provided that the frequency of the entraining signal is within 20% of the natural frequency. Single stimuli applied to the stretch receptors can, if appropriately timed, reset the phase of the flight rhythm. The rhythm can also be entrained by sinusoidal depression and elevation of a forewing while the locust is flying in a wind tunnel. The entrainment range for this more natural stimulation is somewhat smaller, with stable entrainment only when the entraining frequency is within 15% of control.

More than modifying the CPG rhythm, forewing stretch receptors appear to be important components of the flight oscillator in locusts. Intracellular recordings from flight CPG interneurons reveal short-latency excitatory input from forewing stretch receptors. Because these receptors are phasically active during flight, can entrain and reset the flight rhythm, and appear to have monosynaptic inputs to CPG interneurons, they are considered to be important elements in the system generating the normal flight rhythm during wing elevation (Figure 4). Effects of these receptors and of the tegulae also increase the precision of coordination among the wings. The conclusion is that the forewing stretch receptors and the tegulae (activated by wing depression) are peripheral elements of a complex oscillator system that comprises the CPG and peripheral sensory feedback loops.

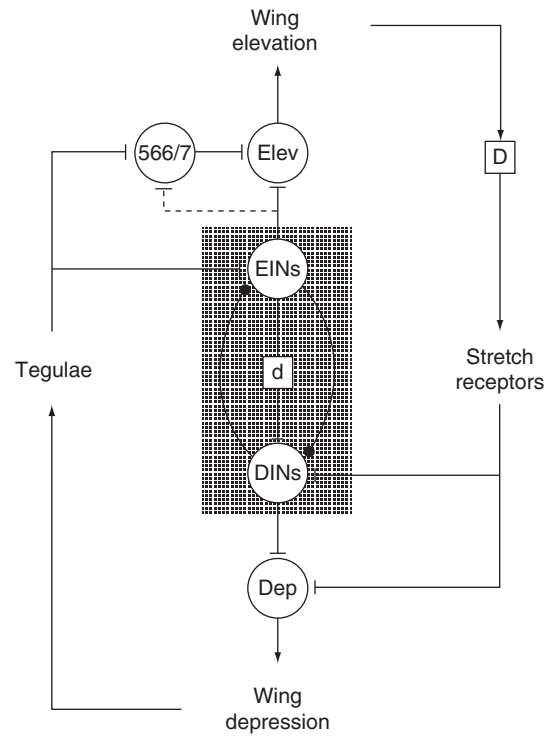

Figure 4 Diagram illustrating identified interactions between sensory afferents and the locust flight central pattern generator (CPG). EIN, elevator interneurons; DIN, depressor interneurons; 566 identified interneurons that participate in the generation of the flight rhythm only when they receive phasic input from tegulae during flight. Elev, elevator motor neurons; Dep, depressor motor neurons; d, delayed excitation; D, long delay, with input arriving at the depressor neurons one cycle subsequent to one that generated the sensory signal. With permission from Pearson KG and Ramirez J-M (1997) Sensory modulation of pattern-generating circuits. In: Stein PSG, Griller S, Selverston AI, and Stuart DG (eds.) *Neurons, Networks, and Motor Behavior*, pp. 225–235. Cambridge, MA: The MIT Press. © 1997 Massachusetts Institute of Technology.

Leech swimming Medicinal leeches are elongated, segmental worms 5–15 cm in length when fully extended. During the prologue to swimming, the leech body flattens and extends in length and width, providing a nearly uniform, flattened profile that is efficient for swimming locomotion. Leeches swim by whole-body undulations that are seen in side-view as a traveling rostrocaudal sinusoidal wave with a cycle period range of 0.3–1 s. The most efficient waveform for this whole-body mode of swimming is a quasi-sinusoidal body wave comprising one wavelength. Commensurate with this theoretical expectation, at any instant, about a single wave length is expressed by swimming leeches. The rhythmic movements of the body are generated by locally antiphasic shortening–lengthening cycles in dorsal and ventral longitudinal muscles. These contraction rhythms, which act against the elevated internal body pressure generated by activity of the flattener muscles, exhibit rostrocaudal phase delays and thereby generate the observed traveling body wave. With some 18–20 body segments actively participating in generating the swim undulations, the intersegmental phase lags in swimming animals are about 20° per segment.

Like other rhythmic movements, the activity cycles, cycle period, and intersegmental phase lags expressed by swimming leeches find a rudimentary origin in a set of interneurons that constitute the CPG. The unit elements of the CPG are found in each midbody ganglion of the leech ventral cord. Although single ganglia are capable of brief, weak swim oscillations when exposed to 50 μmol l^{-1} serotonin, at least two segments are required to generate strong oscillations and intersegmental phase delays. The leech swim CPG comprises more than 13 interneurons per segment, with some participation by sets of dorsal and ventral inhibitory motor neurons. The central oscillator network does not, however, generate the full

rhythm expressed in the intact animal. The cycle period generated by the CPG ranges from 0.6 to 2 s, barely overlapping the period range observed in intact animal (0.3–1 s), and intersegmental phase lags expressed by the CPG (about 10° per segment) are only 50% of these of intact animals (Figure 2(a)). Thus the isolated CPG does not produce neuronal activity patterns for fast, efficient locomotion.

The sensory inputs that, together with CPG circuits, generate neuronal activity patterns appropriate for efficient swimming arise, at least in part, from segmental stretch receptors, whose somata are associated with dorsal and ventral longitudinal muscles. Stretching the leech body wall induces a hyperpolarization in these receptors, which have peripherally located somata and dendrites and whose giant axons conduct this signal to the CNS electronically.

The importance of sensory mechanisms in leech swimming locomotion was recognized from the beginning of systematic studies of this behavior. For example, leeches were found to swim at a lower frequency in high viscosity media. Moreover, leeches whose movement was restricted mechanically so that they were unable to execute the normal swim undulation altered their movements or aborted ongoing traveling waves. More recently, several types of experiments demonstrated that the stretch receptors can modify the basic activity patterns generated by the CPG. First, in nerve cord–body wall preparations, these receptors encode muscle tension information via membrane potential oscillations, with amplitudes up to 10 mV recorded as the axons enter the ganglia. Second, rhythmic current injection into these receptors can alter intersegmental phase lags by ±5° in a phase-dependent manner. Third, experimentally induced stretch receptor membrane potential oscillations entrain the CPG rhythm if the imposed rhythm is within about 50% of the free-running CPG rhythm. Fourth, brief (several seconds long) current pulses injected into a stretch receptor shift the phase of the ventral nerve cord swim oscillations. Finally, there is direct electrical coupling between at least one of the stretch receptors and a CPG interneuron (Figure 5). Hence, as for locust flight, the circuits that generate the swimming rhythm in the leech comprise both the CPG and the peripheral stretch.

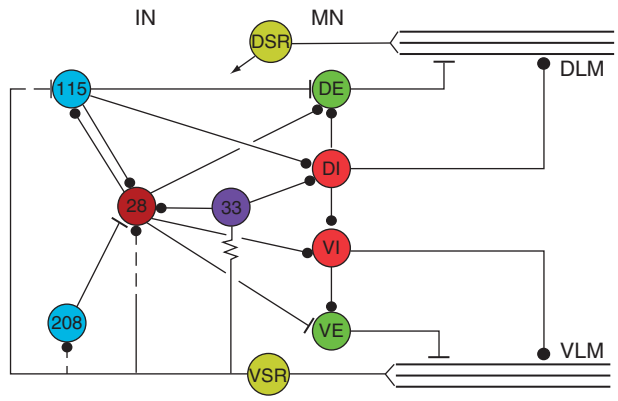

Figure 5 Interactions between stretch receptors and swim-related neurons in leeches. Only a subset of central pattern generator (CPG) neurons (cells 115, 208, 28, and 33) and the phasically active motor neurons are depicted. The ventral stretch receptor (VSR) conveys tension information from the ventral longitudinal muscle (VLM) directly to at least one CPG interneuron, cell 33. IN, interneurons. Motor neuron (MN) designations: DE, dorsal excitor; DI, dorsal inhibitor; VI, ventral inhibitor; VE, ventral excitor; interaction symbols: filled circle, chemical inhibition; 'T,' excitation; resistor, nonrectifying electrical interaction. Dashed interactions appear to be indirect. Interactions between the DSR and the CPG remain unexplored. Reprinted from Cang J, Yu X, and Friesen WO (2001) Sensory modification of leech swimming: Interactions between ventral stretch receptors and swim-related neurons. *Journal of Comparative Physiology* 187: 569–579, figure 1(b), © 2001 Springer-Verlag, publisher. With kind permission of Springer Science and Business Media.

Lamprey swimming Lampreys are primitive fish that swim via anguilliform (eellike) undulations, that, as in leeches, comprise about one wavelength at any moment. The traveling wave results from antiphasic contraction of segmental muscles and from rostra-caudal delays of 1% of the cycle period per segment for a 100-segment animal. Intact lampreys swim with cycle periods ranging from about 0.13 s to more than 0.6 s. The spinal cord of lampreys includes a series of local CPGs. Comparisons of swimming movements in intact animals and motor neuron activity patterns generated within the stationary spinal cord (fictive swimming) show similar patterns (Figure 2(b)). Although the range of cycle periods obtained by altering the concentration of the swim-inducing excitatory amino acids (0.7–2 s) in fictive swimming overlaps that observed in the intact animal, the CPG-generated period is normally about twice that of the intact animal. Unlike results from leeches, the intersegmental phase lags in lampreys do not differ markedly in isolated preparations compared with those of intact animals (Figure 2(b)). Thus sensory feedback appears to be essential for setting the cycle period but not for phase control during undisturbed swimming in the lamprey.

Sensory feedback during the side-to-side undulations of the swimming lamprey arises, in part, from segmental 'edge cells' within the spinal cord. These cells make synaptic contacts with CPG neurons and with motor neurons. Interactions between the CPG and the edge receptors in the intact animal may, like the CPG itself, comprise a functional oscillator. Experiments on spinal cord–notochord preparations demonstrated that the CPG rhythm can be entrained

by mechanical movement of either the rostral or the caudal ends. That is, sensory input arising from these movements could alter the 0.7 Hz 'natural' frequency by a maximum of 50% above and no more than 25% below this control frequency. The range of entrainment depended on the length of the preparation. Short preparations (25 segments) yielded an entrainment range that was nearly twice the range for longer preparations (50 segments). Relative entrainment was also observed with input-to-expressed frequency ratios of 1:2 and 2:1.

Brief movement pulses applied to the spinal cord, as well as current injection into edge cells, can reset the phase of the swimming rhythm of lampreys. Movement pulses, used to generate phase response curves for the spinal cord–notochord preparations, showed phase advances during two-thirds of the cycle, with the largest phase shifts near mid-cycle. Again, it seems likely that, as for locusts and leeches, normal locomotion in lamprey arises both from segmental CPGs and from sensory feedback circuits (Figure 6).

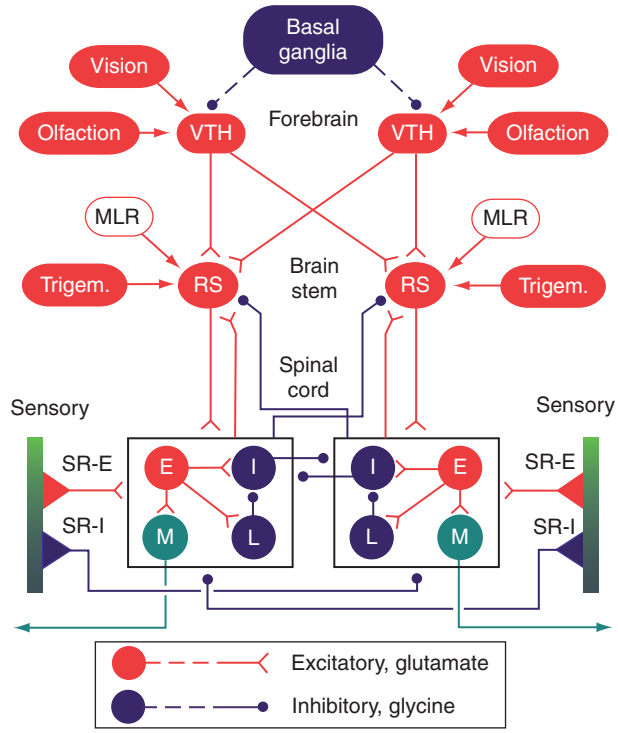

Figure 6 Circuits underlying swimming movements in lampreys. A population of reticulospinal neurons (RS) excites the spinal interneurons and motor neurons. Excitatory interneurons (E), in turn, excite other spinal neurons, including inhibitory interneurons (I) that cross the midline to inhibit neurons on the contralateral side, and lateral interneurons (L) that inhibit ipsilateral interneurons and the motor neurons (M). Spinal stretch receptor neurons (edge cells) include two types; excitatory neurons (SR-E) that excite ipsilateral neurons and inhibitory neurons (SR-I) that inhibit contralateral neurons. Symbols denote populations of cells, not single neurons. Additional control of this system appears to arise from the ventral thalamus (VTH), the trigeminal nuclei (Trigem.), and the mesencephalic locomotor region (MLR). Reproduced from Grillner S and Wallén P (2002) Cellular bases of a vertebrate locomotor system: Steering, intersegmental and segmental co-ordination and sensory control. *Brain Research Reviews* 40: 92–106, with permission from Elsevier.

Crustacean swimmeret beating Crayfish, lobsters, and other decapod crustaceans have abdominal appendages, swimmerets, that beat with metachronal rhythms. The paired segmental CPGs in this system, unlike those described above, provide a basic rhythm that is nearly identical in both cycle period and intersegmental phase lag, about a quarter of the cycle period, to those observed in the intact animal. Although the need for sensory feedback in undisturbed beating is not obvious, execution of swimmeret movements is nonetheless monitored via a set of sensory receptors, including proprioceptors that provide phasic feedback. When, for example, the movement of a swimmeret is experimentally impeded, by fixing the limb at either the retracted or protracted phase in its movement cycle, the central rhythm sometimes stops, sometimes decelerates. In addition to the proprioceptors, lobsters have hairs whose deflection in response to water movements engendered by swimmeret beating can provide rhythmic afferent information. Also, there are two types of hypodermal mechanoreceptors that are likely to be activated when swimmerets beat rhythmically. It appears that swimmeret mechanoreceptor activity, although not required for generating an appropriate activity pattern for swimmeret movement, contributes to the movement of these appendages when normal movement is somehow impeded.

Insect and crayfish walking The neuronal activity patterns that underlie animal walking movements are considerably more complex than the neuronal patterns that underlie flying, swimming, or swimmeret beating. Most animals that utilize walking locomotion have multiple limbs, with each limb comprising multiple segments. Added to those factors is the daunting task of generating effective limb placement on a highly unpredictable substrate. It is therefore not surprising that sensory receptors are closely tied to CPG elements associated with this form of locomotion.

Many forms of animal locomotion can be viewed as the rhythmic alteration of two phases, a power stroke, in which the animal exerts forces on its environment that propel it forward, and a return stroke that returns the limb to the original position for the next power stroke. A major focus for studies of sensory input to the CPG circuits that underlie the walking cycle is the control of transitions between two phases in walking locomotion, the stance and

the swing phases. Two additional foci are the control of intra- and interjoint phase angles.

One of the best-studied walking behaviors is that of the stick insect. Each leg in the stick insect has three main joints that are responsible for movements in the horizontal and vertical planes. Muscles associated with the thoracocoxal (TC-) joint drive the leg back and forth, whereas the coxatrochanteral (CTr-) and femur–tibia (FTi-) joints (Figure 7(a)) are involved in placing the tarsal segment firmly onto the substrate (during the stance phase) or raising it into the air (during the swing phase). With appropriate stimulation, deafferented thoracic ganglia can be induced to generate the antagonistic motor neuron activity patterns that cause rhythmic joint movements. Because there is little or no coordination among groups of motor neurons associated with different joints in such a preparation, it appears that a unit CPG drives muscles at each joint. The specific neuronal circuits constituting these CPGs remain largely unidentified; however, it is clear that activity in these CPGs is closely coordinated by sensory signals arising from leg proprioceptors and cuticular strain receptors. Coordination of the movements of individual joints occurs by a cascade of sensory information whereby signals from one joint in a leg reset the phase of the CPG that drives a second joint in that leg. These signals include two temporally related sensory modalities: leg weight load and joint position. The effects of such sensory signals are weighted within the CNS to generate locomotory movements.

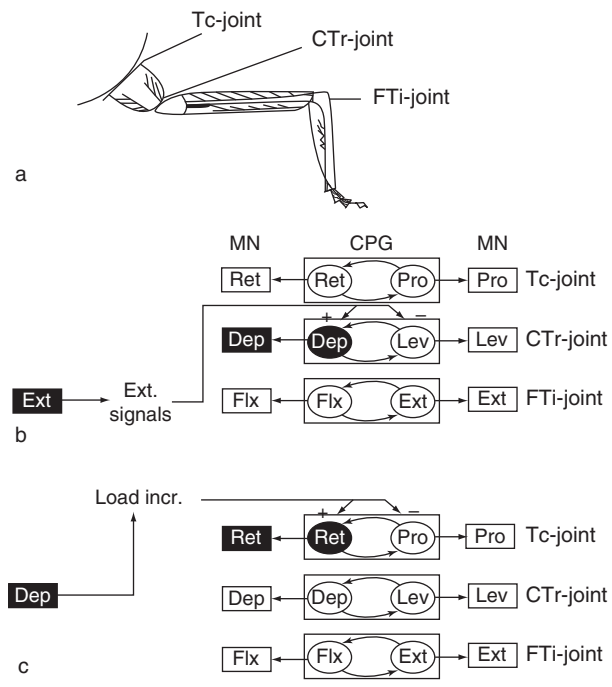

Figure 7 Schematic of sensory influences on the movements of the middle leg of the stick insect. (a) Cartoon of the stick insect leg showing the three main leg joints. (b) Extension position signals (ext. signals) from the FTi-joint switch the phase of the central pattern generator (CPG; Lev to Dep) in the adjacent joint (CTr-joint). The CPGs are denoted by boxes with two interacting neurons, with the active phase indicated by a filled circle. Arrows indicate interactions, but not specific synapses, between hypothesized CPG elements and the antagonistic sets of motor neurons (MNs; rectangles; filled elements are active) they control. (c) Load signals (load increase (incr.)) from the CTr-joint switch the state of the TC-joint CPG (Pro to Ret). TC-joint, –thoracocoxal joint; CTr-joint, –coxatrochanteral joint; FTi-joint, femur–tibia joint; Ret, retraction/retractor; Pro, protraction/protractor; Dep, depression/depressor; Lev, levation/levator; Flx, flexion/flexor; Ext, extension/extensor. Reprinted from Büschges A (2005) Sensory control and organization of neural networks mediating coordination of multisegmental organs for locomotion. *Journal of Neurophysiology* 93: 1127–1135, used with permission.

The specific roles of sensory influences on the CPGs that control the main joints in the stick insect are now known for the phase transitions that occur during step cycles. Two such transitions are illustrated in Figures 7(b) and 7(c). In Figures 7(b), movement-induced signals from FTi-joint receptors (extension signals, lower left) cause a transition from swing to stance phase. This occurs because this sensory input switches the CPG of the CTr-joint leg to the depressor phase, activating the depressor motor neuron, thereby lowering the leg to begin the stance phase. Next, in Figure 7(c), load receptor activation in the CTr-joint (load increase, left) induced by initiation of the stance phase forces a phase transition from protraction to retraction onto the TC-joint CPG, thereby activating retractor motor neurons that act at the TC-joint and begin the forward propulsion phase of the leg step cycle. Sensory receptor input often has two effects, such as when an increased load increases the activity in flexor motor neurons and hence to muscles that support the animal's weight while simultaneously inhibiting the antagonistic extensors. There appear to be two general coordinating systems: one that initiates transitions between phases of the walking cycle, and one that determines its direction. Computer simulations have demonstrated that the known sensory inputs to control limb motions can, when implemented in a simulated neuronal controller, generate leg stepping movements. Here again, rhythmic locomotory movements arise from combined CPG and sensory interactions. For the stick insect, the role of sensory input influences not only period and phase but the very transitions of the CPG between phases of the activity cycle.

As in stick insects, walking locomotion in the crayfish also arises from CPGs strongly influenced by sensory input. However, unlike the nearly independent joint CPGs of the stick insect, walking CPGs express a coordinated walking rhythm in deafferented preparations. That is, the crayfish CPGs

include effective interactions within the CNS while still requiring sensory input to ensure appropriate cycle periods and stable phase relationships. Although the CPGs associated with the eight walking legs in crayfish remain unidentified, at least one sensory structure, the thoracocoxal muscle receptor organ (TCMRO), is known to interact with the CPGs. Rhythmic stretching and releasing of the TCMRO can entrain the CPG, not only locally, but across other legs. The data show that the interactions between sensory input and the CPGs are strong, for the phase relationship between an imposed stimulus rhythm, and the entrained walking cycles is nearly constant. The suggestion is that the walking rhythm is not entrained but actually driven by sensory input. Relative coordination also occurs, with stimulus–response periods of 2:1 or 1:2 when the stimulus frequency lies outside the entrainment range. Direct current injection into the TCMRO afferents also was effective. Overall, sensory input, as in the stick insect, appears to trigger transitions between phases of the movement cycle, and hence sensory input provides essential information to the locomotory CPG in crayfish.

Conclusion: Beyond Sensory Modulation

Just as motor neurons sometimes are integral members of CPGs, even if they are not strictly limited to the CNS, so sensory neurons and muscles may be essential components of circuits that generate rhythmic movements. Sensory input varies in importance among the highly diverse circuits that generate wing movements in locust flight, swimming undulations in the medicinal leech and lampreys, and swimmeret beating and walking locomotion in insects and crayfish. For understanding the physiological bases of rhythmic movements, it appears to be unhelpful to construct artificial divisions between CPGs, motor neurons, muscles, and sensory feedback. A more useful approach is to identify oscillatory subcircuits wherever they occur, whether within the CNS or not; and whether comprising interneurons, motor neurons, effector structures, or sensory receptors. Such subcircuits, together with their mutual interconnections, generate the highly variable and adaptive oscillating patterns that underlie rhythmic animal movement.

See also: Behavioral Hierarchies; Central Pattern Generators; Command Systems; Hormones and Behavior; Neuromodulation; Pattern Generation; Swim Oscillator Networks; Swimming: Neural Mechanisms ; Walking in Invertebrates.

Further Reading

Büschges A (2005) Sensory control and organization of neural networks mediating coordination of multisegmental organs for locomotion. *Journal of Neurophysiology* 93: 1127–1135.

Cang J, Yu X, and Friesen WO (2001) Sensory modification of leech swimming: Interactions between ventral stretch receptors and swim-related neurons. *Journal of Comparative Physiology* 187: 569–579.

Delcomyn F (1980) Neural basis of rhythmic behavior in animals. *Science* 210: 492–498.

Elson RC, Sillar KT, and Bush BMH (1992) Identified proprioceptive afferents and motor rhythm entrainment in the crayfish walking system. *Journal of Neurophysiology* 67: 530–546.

Friesen WO and Cang J (2001) Sensory and central mechanisms control intersegmental coordination. *Current Opinion in Neurobiology* 11: 678–683.

Grillner S and Wallén P (2002) Cellular bases of a vertebrate locomotor system: Steering, intersegmental, and segmental co-ordination and sensory control. *Brain Research Reviews* 40: 92–106.

Hooper SL and DiCaprio RA (2004) Crustacean motor pattern generator networks. *Neurosignals* 13: 50–69.

Kristan WB Jr, Calabrese RL, and Friesen WO (2005) Neuronal control of leech behavior. *Progress in Neurobiology* 76: 279–327.

McClellan AD and Jang W (1993) Mechanosensory inputs to the central pattern generators for locomotion in the lamprey spinal cord: Resetting, entrainment, and computer modeling. *Journal of Neurophysiology* 70: 2442–2454.

Pearce RA and Friesen WO (1984) Intersegmental coordination of leech swimming: Comparison of *in situ* and isolated nerve cord activity with body wall movement. *Brain Research* 299: 363–366.

Pearson KG and Ramirez J-M (1997) Sensory modulation of pattern-generating circuits. In: Stein PSG, Griller S, Selverston AI, and Stuart DG (eds.) *Neurons, Networks, and Motor Behavior*, pp. 225–235. Cambridge, MA: MIT Press.

Rossignol S, Dubuc R, and Gossard J-P (2006) Dynamic sensorimotor interactions in locomotion. *Physiological Reviews* 86: 89–154.

Sherrington CS (1947) *The Integrative Action of the Nervous System*, 2nd edn. New Haven, CT: Yale University Press.

Wallén P and Williams TL (1984) Fictive locomotion in the lamprey spinal cord *in vitro* compared with swimming in the intact and spinal animal. *Journal of Physiology* 347: 225–239.

Wilson DM (1961) The central nervous control of flight in a locust. *Journal of Experimental Biology* 38: 471–490.

Yu X and Friesen WO (2004) Entrainment of leech swimming activity by the ventral stretch receptor. *Journal of Comparative Physiology* 190: 939–949.

Cerebellar Deep Nuclei

T W Thach

© 2009 Elsevier Ltd. All rights reserved.

Anatomic pathway tracing in the macaque clarifies the connections and topographic organization of the deep cerebellar nuclei. The output of the cerebellum is generated by the deep cerebellar nuclei (along with vestibular and reticular nuclei in the brain stem, which also receive direct Purkinje cell inputs). These project to and excite a target area within the thalamus that is sufficiently free of other inputs as to be called the 'cerebellar thalamus' (Figure 1). The target area includes VLc, VLps, VPLo, VLo, and X. The basal ganglia input to thalamus arrives more anterior in VLo and VA, and the lemniscal arrives more posterior in VPL. This cerebello-thalamic target area projects to cerebral cortical areas including the motor, premotor, supplementary motor, oculomotor, prefrontal, cingulated, and parietal regions.

Dentate, interposed, and probably fastigial and vestibular nuclei each project in overlapping fashion to the whole width (coronal) of the contralateral thalamic receiving area. The fastigius projection is sparse, bilateral, and restricted and appears not to project to X or VLo. This itself suggests that at the level of cerebellar outflow, control of the thalamic target is multiple and is repeated across each nucleus. (The projections of the three nuclei arrive in register at a macrolevel only; at the cellular level, they appear to interdigitate in patches (interpositus) and rostrocaudal rods (dentate)).

From the known somatotopic mapping in cerebellar thalamus, and from the topographic projection of each nucleus onto the common thalamic target, somatotopic mapping may be inferred within each of the cerebellar nuclei. In cerebellar thalamus, head is medial, tail lateral, trunk dorsal, and extremities ventral. In cerebellar nuclei, head is then caudal, tail rostral, trunk lateral, and extremities medial. This scheme has been supported by neural recording and inactivation studies during movement and is consistent with the electroanatomical input mapping studies in anesthetized macaques.

This is not to say that the cerebellar output is exclusively or even mainly directed to thalamus and cerebral cortex. Dentate projects to parvocellular red nucleus, the reticular nucleus of the pontine tegmentum, and the inferior olive (principal olive (PO)); interpositus projects to the magnocellular red nucleus, the reticular nucleus of the pontine tegmentum, to the inferior olive (dorsal accessory olive (DAO)), and to the spinal cord intermediate gray; and fastigius projects to lateral and descending vestibular nuclei, to n. reticularis tegmenti pontis and prepositus hypoglossi, to the reticular gray of the midbrain, to the inferior olive (medial accessory olive (MAO)), and to contralateral motor neurons in the spinal cord.

Lesion Studies: What Do the Multiple Maps in the Cerebellum Control?

If there is a separate body representation within each nucleus, it seems likely that each should control some different aspect of bodily function. In one study, unit activity was recorded within the cerebellar nuclei as the monkeys performed five different trained movements of the wrist. The hypotheses to be tested were the following: (1) each nucleus controls some parameter, within or across tasks; (2) each nucleus simply modulates to provide gain control in downstream target structures; and (3) each nucleus controls some type or 'mode' of movement and not others. The job of the animal was to flex or extend the wrist to line up the cursor within target window on an oscilloscope (Figure 2(a)) and to maintain the alignment as the window moved. The animals performed five tasks designed to dissociate hypothetical functions of the nuclei: (1) 'jerk', a prompt, visual-triggered move away from hold position; (2) 'jump', identical to the preceding, except requiring a stop within a visual target area; (3) 'pert', a return to the hold position after perturbation by torque step; (4) 'ramp', tracking of a slowly moving visual target; and (5) 'RAM', self-paced rapidly alternating movements. All tasks were performed in two directions, under two loads.

The earliest changes were seen in dentate on the jerk and jump tasks. Interpositus showed the earliest changes on the pert task, and alone showed modulation in relation to tremor on the ramp tracking. Interpositus and dentate showed equal modulation on the RAM task. Dentate and interposed discharge only occasionally related to the direction of movement and load (position was not dissociated from direction). Fastigial neurons occasionally showed modulations, but without the trial by trial or temporal consistency seen for the other two nuclei. Both dentate and interpositus seemed to contribute to all tasks, but at different times and intensities. The lack of consistent movement correlations in fastigius raised questions about whether it was taking any part in these movements.

To see whether the parameters of movement or movement stability depended on the neural discharge with which they were correlated, the different nuclei were inactivated with muscimol (temporary) or

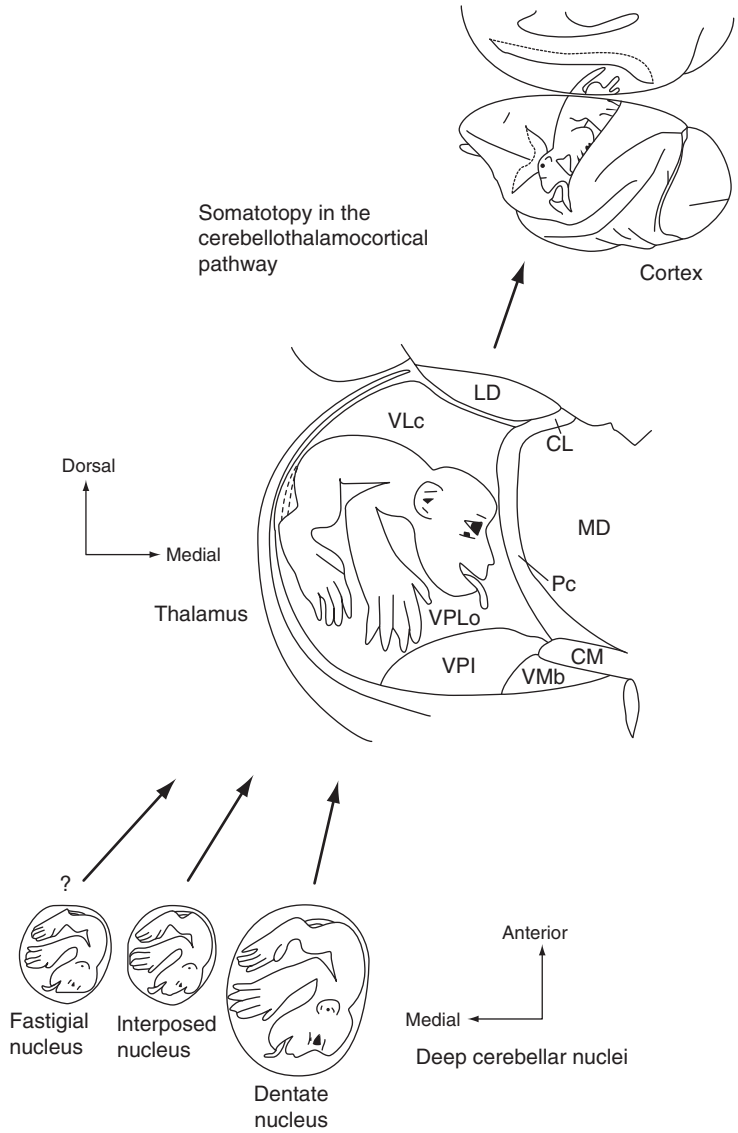

Figure 1 Projection of cerebellar nuclear somatotopic maps onto a common map in cerebellar thalamus and then to motor cortex.

kainic acid (permanent) while the proficiency of performance was monitored. Any impairment was documented by comparing the pre- and postinjection movement traces. Several variables were examined, including onset and termination times, movement time, direction, velocity, amplitude, tremor, and EMG patterns.

For dentate, there was a slight (47 ms) delay in reaction time on jerks and jumps only. There was no overshoot of final target position. There was a tremor during ramps. Tremor was transient, even after permanent kainate ablation. Perts and RAMs (except for superimposition of tremor) were normal. For interpositus, there was a slight (10–20 ms) delay in the reaction time on perts after one muscimol injection. There was a tremor at the end of jumps and perts and during ramps. The tremor persisted. There was no consistent change in frequency, amplitude, or regularity of RAMs, except the superimposition of tremor. For fastigius there was no detectable abnormality in the performance of any of the five trained wrist movements. Similar small delays and irregularities of velocity and force in single joint tracking and in isometric force have been seen in studies in cerebellum-damaged patients.

To summarize: in no instance did inactivation abolish the task performance. The seemingly contradictory set of findings of (1) a good correlation of discharge with the behavior under study and (2) little or no deficit in that behavior after ablation again raises the perplexing question of what it is that the cerebellum does.

The question of whether the cerebellum independently controls the muscles that operate each joint, or

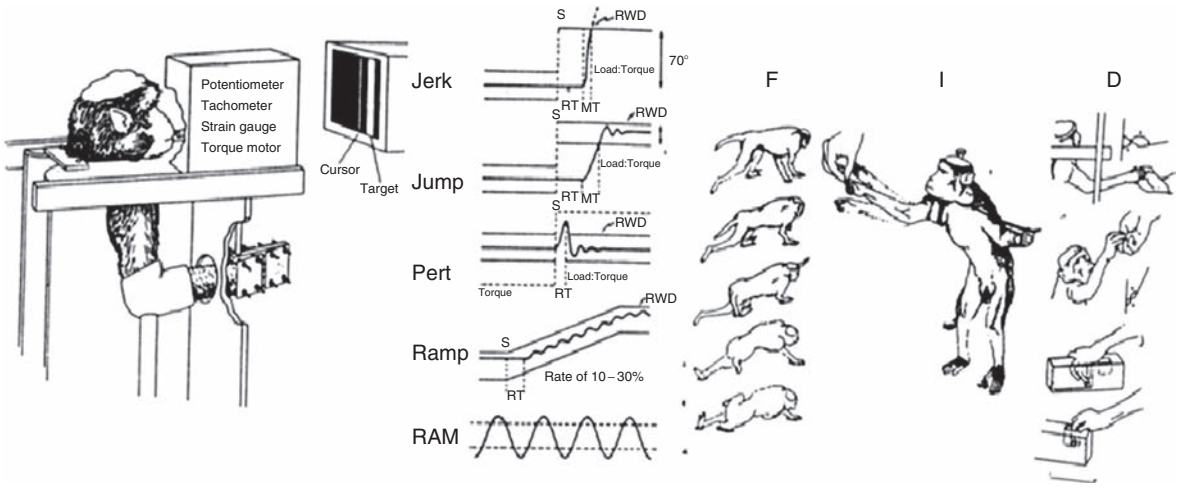

Figure 2 Cerebellar lesion impairs compound movements more than single-joint movement. Fastigial: falls to side of lesion; wide base. Interposed: antagonist delay; action tremor. Dentate: reach overshoot; x-y error; and finger incoordination. From Thach WT, Goodkin HP, and Keating JG (1992) The cerebellum and the adaptive coordination of movement. *Annual Review of Neuroscience* 15: 403–442.

whether it has a specific role and mechanism for combining the many muscles that operate many joints in a multijoint movement has previously been raised.

The preceding study was of the effect of nuclear inactivation across nuclei on the performance of five trained single-joint tasks. In the same animals in the same sessions, the performance of five untrained tasks was also observed. The tasks consisted of sitting, standing, walking, reaching out for bits of food, and picking small bits of food out of deep narrow food wells with a precision pinch of the fingers. Stance, sitting, and walking were evaluated immediately after the animal was released from the primate chair after the injections. Reaching involved coordinated movement of shoulder and elbow; reaching was performed while sitting and standing on the floor and while sitting in the primate chair. The chair reduced the requirements for active postural adjustments of trunk and other limbs and permitted easier assessment of the relative contributions of shoulder and elbow. The precision pinch task required coordinated movements of the hand and fingers, including the pinch *per se* of thumb and forefinger, and 'tea-cup' posturings of the other digits to keep them out of the way of the edges of the food well.

In sum, each nuclear inactivation produced an incapacitating impairment of bodily movement, but for each nucleus, the type of deficit was unique (Figure 2(b)). Fastigial inactivation prevented sitting, standing, and walking, with frequent falls to the side of the lesion. Interposed inactivation caused a severe action tremor of 3–5 Hz during reaching, but not during sitting, standing, or walking. Dentate inactivation caused excess angulation of shoulder and elbow in reaching, which resulted in overshoot of the target, and an increased use of single-digit strategies in attempting to pinch and pick up food morsels, from which an incoordination of compound finger movements was inferred. These tests showed differences across the three nuclei, and indicated task and movement mode specificity for each. They were qualitatively more severe than any deficits seen in the simple, single-jointed wrist movements.

Parallel Fibers and Purkinje Cell Beams Link Deep Nuclear Cells as a Mechanism of Coordination

The relationship between cerebellar somatotopic maps and parallel fiber morphology is consistent with a cerebellar role in movement coordination (Figure 3). In each body representation within the deep nuclei, the rostrocaudal axis of the body is mapped onto the sagittal axis of the nucleus. The hindlimbs are represented anteriorly, the head posteriorly, distal parts medially, and proximal parts laterally. This orientation suggests that the myotomes, which are arranged orthogonal to the rostrocaudal axis of the body, are represented primarily in the coronal dimension of the cerebellum and thus roughly parallel to the trajectory of the parallel fibers. Because the parallel fibers are connected to the deep nuclear cells by Purkinje cells, a coronal 'beam' of parallel fibers would control the nuclear cells that influence the synergistic muscles in a myotome. In this way, the parallel fiber would be a single neural element spanning and coordinating the activities of multiple synergistic muscles and joints.

Parallel fibers in the monkey are about 6 mm long, spanning a stretch of cerebellar cortex that projects

Figure 3 Parallel fiber beams: a mechanism for coordination. Parallel fibers are long enough to link together Purkinje cells projecting to different body parts within one nuclear body map, and to connect the different modalities of maps in adjacent nuclei. Within a parallel fiber beam, Purkinje cell discharge is synchronized. Cutting a beam prevents coordination of movements controlled by maps in adjacent nuclei.

across the width of one deep nucleus. Thus, a strip of Purkinje cells under the influence of a set of parallel fibers of the same origin and length control a strip of nuclear cells across an entire nucleus. Depending on which portion of the somatotopic map is involved, that nuclear strip may influence synergistic muscles across several joints in a limb or the muscles of the eyes, head, neck, and arm, for example.

A study in trained reaching rats supports this hypothesis. Simple spikes of pairs of Purkinje cells were recorded that, with respect to each other, were either aligned on a beam of shared parallel fibers or were located off beam. Firing rates of simple spikes firing in both on-beam and off-beam Purkinje cell pairs commonly showed large variations during reaching behavior. But with respect to timing, on-beam Purkinje cell pairs had simple spikes that were tightly time-locked to each other and to movement, despite the variability in rate. By contrast, off-beam Purkinje cell pairs had simple spikes that were not time-locked to each other.

What might be the consequence of the on-beam synchronicity? Since parallel fibers in the cerebellar cortex run roughly parallel to the myotomes of the body maps in the deep nuclei, the on-beam parallel fiber activity could thus recruit Purkinje cell inhibition to modulate coordination and timing of the natural myomeric synergists mapped within the deep nuclei. In addition, parallel fibers would be long enough to span (via Purkinje cell projections) two or more adjacent nuclei, fastigius controlling stance and gait, interpositus agonist/antagonist coordination at a single joint, and dentate eye–limb coordination in reaching and of digits in multidigit movements.

The role of the parallel fiber in movement coordination has been further supported in a human study. Five children who had surgical midline division of the posterior cerebellar vermis to remove fourth ventrical tumors could subsequently walk with little impairment. Further, they could hop in place on one leg as well as normal controls. Yet these same individuals could not do tandem heel-to-toe walking along a straight line without falling to one side or the other. These observations suggested that vestibular information could be used to control one leg, but that the two legs could not be thus coordinated. The division of parallel fibers crossing the midline seemed best able to explain this dissociation.

What Does the Cerebellum Contribute to Cognitive Processes?

There is now little question that the cerebellum projects to many areas of cerebral cortex outside of primary motor cortex, and that many of these areas have cognitive as well as motor functions. Does the cerebellum then control cognition?

Human brain imaging studies indicate that the cerebellum participates in some cognitive functions. One study revealed cerebellar activation when subjects imagined themselves moving or passively observed movement without moving themselves. In another study, the lateral cerebellum was activated in a language task in which subjects were asked to generate appropriate verbs for visually presented nouns. Interestingly, once the verb generation task had been well practiced, cerebellar activation diminished, suggesting a cerebellar role primarily in the learning of the noun–verb association. The cerebellum was not activated when the nouns were simply read, nor when nouns were matched with nonmotor adjectives. A third study showed that the dentate was activated bilaterally when subjects worked to solve a pegboard puzzle, and this activation was three to four times greater in magnitude than during simple peg movements. Overall, these findings indicate that the lateral cerebellum and dentate are activated in ways that would not be expected to account for movement alone.

Behavioral observations of humans with cerebellar damage have also provided evidence that the cerebellum participates in nonmotor aspects of cognition. In one case study, a cerebellar patient was unable to detect errors in a verb-generating task and did not exhibit normal practice-related learning, although the patient performed normally on standard intelligence and memory tests. In another study, cerebellar patients were deficient in both the production and the perception of a timing task. Lateral cerebellar lesions interfered with the perception of time intervals, while

medial cerebellar lesions affected the timing of implemented responses.

Some clinical studies have proposed even greater cerebellar control of mental activities. One such study has proposed a 'cognitive affective cerebellar syndrome'; however, the variety of mental deficits and the varied locations of the cerebellar lesions raised questions about the consistency of the suggested correlation. Other studies of infantile autism document attention deficits as well as the focal cerebellar atrophy. However, subjects with autism have abnormalities of movement in addition to abnormalities of mentation, while focal cerebellar lesions in previously normal individuals have not produced such a syndrome. This is also true for work on schizophrenia, which has documented a common association of focal cerebellar atrophy, yet schizophrenic patients also have disorders of movement, and focal cerebellar lesion in a previously normal individual has not been observed to cause a schizophrenic syndrome.

See also: Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Cerebellum: Clinical Pathology; Cerebellum: Evolution and Comparative Anatomy; Cerebellum: Models; Cognition: Cerebellum Role; Olivocerebellar System; Procedural Learning: Cerebellum Models.

Further Reading

Babinski J (1988) De l'asynergie cerebelleuse. *Revue Neurologique* 7: 806–816.

Babinski J (1906) Asynergie et inertie cerebelleuses. *Revue Neurologique* 14: 685–686.

Bastian AJ, Mink JW, Kaufman BA, and Thach WT (1998) Posterior vermal split syndrome. *Annals of Neurology* 44: 601–610.

Courchesne E, Akshoomoff NA, Townsend J, and Saitoh O (1975) A model system for the study of attention and the cerebellum: Infantile autism. *Electroencephalography and Clinical Neurophysiology* 44(supplement): 315–325.

Decety J, Sjöholm H, Ryding E, Stenberg G, and Ingvar DH (1990) The cerebellum participates in mental activity: Tomographic measurements of regional cerebral blood flow. *Brain Research* 535(2): 313–317.

Dum RP, Li C, and Strick PL (2002) Motor and nonmotor domains in the monkey dentate. *Annals of the New York Academy of Science* 978: 289–301.

Dum RP and Strick RL (2003) An unfolded map of the cerebellar dentate nucleus and its projections to the cerebral cortex. *Journal of Neurophysiology* 89(1): 634–639.

Gebhart AG, Petersen SE, and Thach WT (2002) The role of the posterolateral cerebellum in language. In: Highstein SM and Thomas Thach W (eds.) *The Cerebellum: Recent Developments in Cerebellar Research*, pp. 318–333. New York: New York Academy of Sciences.

Heck DH, Thach WT, and Keating JG (2007) On-beam synchrony in the cerebellum as the mechanism for the timing and coordination of movement. *Proceedings of the National Academy of Sciences of the United States of America* 104(18): 7658–7663.

Kim SG, Ugurbil K, and Strick PL (1994) Activation of a cerebellar output nucleus during cognitive processing. *Science* 265(5174): 949–951.

Mugnaini E (1983) The length of cerebellar parallel fibers in chicken and rhesus monkey. *Journal of Computational Neurology* 220: 7–15.

Okugawa G, Sedvall G, Nordstrom M, et al. (2002) Selective reduction of the posterior superior vermis in men with chronic schizophrenia. *Schizophrenia Research* 55: 61–67.

Schahamann JD and Sherman JC (1998) The cerebellar cognitive affect syndrome. *Brain* 121: 561–579.

Strick PL (2003) Cerebellar loops with motor cortex, prefrontal cortex of a nonhuman primate. *Journal of Neuroscience* 23(23): 8432–8444.

Thach WT (1996) On the specific role of the cerebellum in motor learning and cognition: Clues from pet activation, lesion studies in humans. *Behavioural Brain Research* 19: 411–431.

Thach WT, Goodkin HP, and Keating JG (1992) The cerebellum and the adaptive coordination of movement. *Annual Review of Neuroscience* 15: 403–442.

Thach WT, Perry JG, Kane SA, and Goodkin HP (1993) Cerebellar nuclei: Rapid alternating movement, motor somatotopy, a mechanism for the control of muscle synergy. *Revue Neurologique* 149(11): 607–628.

Thach WT, Perry JG, Kane SA, and Goodkin HP (1993) Cerebellar nuclei: Rapid alternating movement, somatotopy, the control of muscle synergy. *Revue Neurologique* 11: 607–628.

Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement

S M Morton, University of Maryland School of Medicine, Baltimore, MD, USA
A J Bastian, Kennedy Krieger Institute and The Johns Hopkins University School of Medicine, Baltimore, MD, USA

© 2009 Published by Elsevier Ltd.

Cerebellar Motor Disorders

The most striking and debilitating effect of damage to the cerebellum is ataxia, which comes from the Greek and translates literally to 'without order.' Ataxia is typically used to mean incoordination of movement. Cerebellar ataxia can result from damage to the cerebellum itself or the pathways to and from it. Damage is typically caused by stroke, degenerative disease, or tumor. Alcoholism is known to cause atrophy especially of the anterior lobe of the cerebellum. Multiple sclerosis often affects the cerebellum as well. Many other neurodegenerative diseases, such as multisystem atrophy and the family of spinocerebellar ataxias, can damage the cerebellum. If there is damage to only one side of the cerebellum, the resultant symptoms manifest on the same side of the body as the damage.

There are many signs of ataxia, and different individuals can have some or all of them, depending on the location of damage. Since ataxia is a very general term, it is often used in reference to movement of a specific body part (e.g., arm ataxia) or type of movement (e.g., gait ataxia). There are many common manifestations of ataxia, including dysmetria (poor distance scaling), dysdiadochokinesia (incoordination between agonist–antagonist muscle pairs during rapid alternating movements), hypotonia (reduced muscle tone), dyssynergia (poor control of multijoint movements), and action tremor (oscillating limb movements about a joint during active movement).

One notable feature of the cerebellum is its distinct functional longitudinal 'zones.' Each cerebellar zone consists of a region of cerebellar cortex and its own set of deep cerebellar nuclei. Each zone has projections to and from different areas of the brain and spinal cord. Thus, each is uniquely positioned to control certain aspects of movement but not others. The middle zone consists of the midline structure, the vermis, and the fastigial nuclei. Patients with damage to the midline zone typically present with some (possibly transient) loss of postural tone, impaired upright posture and balance, gait ataxia, and eye movement deficits. The intermediate zone is made up of the intermediate hemispheres and the globose and emboliform nuclei. Damage to this region often results in action tremor, dysdiadochokinesia, and dysmetria of the limbs. The largest region of the cerebellum is the lateral zone, which contains the two lateral hemispheres and the dentate nuclei. Patients with damage in this region have dyssynergia, dysmetria, and difficulty planning complex limb movements, especially those that are visually guided. Because each of the cerebellar arteries supplies multiple longitudinal zones and most degenerative disorders are pancerebellar, patients typically have cerebellar lesions that span multiple longitudinal zones. Patients therefore usually show a variety of cerebellar signs.

Limb Control

One of the most obvious signs of cerebellar damage is limb ataxia. Unlike the smooth and accurate movements healthy people make, individuals with cerebellar lesions have arm and leg movements that are characteristically choppy, segmented, slow, and inaccurate. The impairments become more pronounced the faster the movements are made.

Several studies have shown that cerebellar patients are often hypermetric when they are asked to make rapid step-tracking movements of a single joint. Some abnormalities commonly observed are reduced but prolonged agonist muscle activity and delayed antagonist activity. When cerebellar patients are asked to make rapid reversals of movement (e.g., test for dysdiadochokinesia), they frequently have difficulty ceasing antagonist activity so that the agonist can begin the second phase of the movement. If a mass is added to the hand (e.g., carrying a load), the hypermetria increases. This is thought to be due to an inability to increase antagonist muscle activity to decelerate and stop the movement appropriately. Other specific deficits in muscle activation patterns differ with changing movement contexts. Nevertheless, it can be said that cerebellar damage results in inappropriate amplitude scaling and timing of the onset and cessation of agonist and antagonist muscles during single joint movements.

Though there are deficits in single joint movements, the motor abnormalities in more natural multijoint movements are much greater. From a mechanical perspective, the task of controlling a multijoint movement is much more complex than that of controlling each of the elemental single joint movements alone.

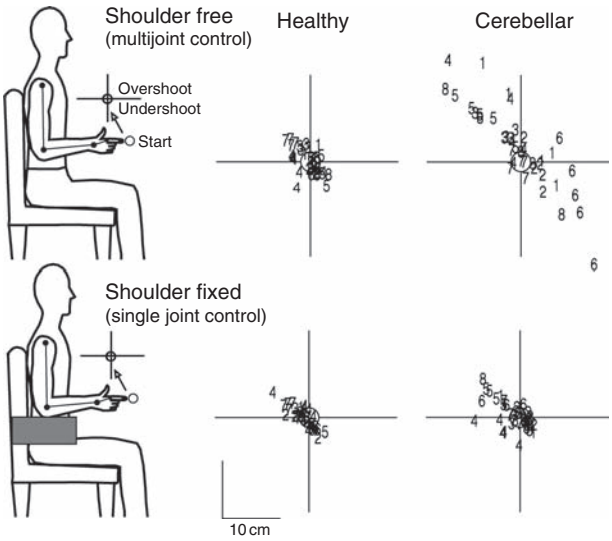

Figure 1 Finger endpoint locations for a group of healthy (left) and cerebellar participants (right) performing upward arm reaches (see experimental setup, far left) that are multijoint (top, labeled 'shoulder free') or single joint (bottom, labeled 'shoulder fixed'). Several trials from participants in each group are plotted (note: participants are given different numbers). Healthy participants are highly accurate in both conditions. Cerebellar patients are markedly inaccurate in the multijoint condition but show accuracy similar to that of healthy participants in the single-joint condition. Reproduced from Bastian AJ, Zackowski KM, and Thach WT (2000) Cerebellar ataxia: Torque deficiency or torque mismatch between joints? *Journal of Neurophysiology* 83: 3019–3030, used with permission from the American Physiological Society.

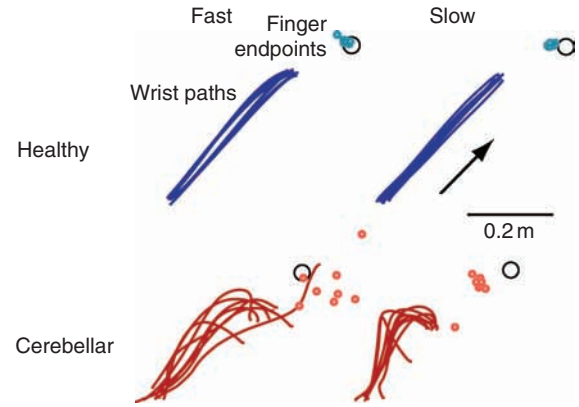

Figure 2 Wrist paths and finger endpoint locations for a typical healthy participant (top) and a cerebellar patient (bottom) performing forward arm reaches either as fast as possible (left, labeled 'fast') or slow and accurate (right, labeled 'slow'). Single reaching trials are overlaid for each type of reach for each participant. The target location is indicated by the large open circle; reach paths begin at the lower left and move upward and rightward. The healthy participant is nearly as accurate in the fast condition as in the slow (see finger endpoints) and shows very straight trajectories throughout (see wrist paths). The cerebellar patient is inaccurate during both types of movements, but noticeably worse in the fast than in the slow condition. During fast movements, hypermetria is predominant; during slow movements, hypometria predominates. Reproduced from Bastian AJ, Martin TA, Keating JG, and Thach WT (1996) Cerebellar ataxia: Abnormal control of interaction torques across multiple joints. *Journal of Neurophysiology* 76: 492–509, used with permission from the American Physiological Society.

Movements of one joint cause passive intersegmental interaction torques to occur at the other linked joints. For example, when the arm reaches forward, the active movement of the forearm will produce interaction torques that cause passive movement of the upper arm and vice versa. These torques are velocity and acceleration dependent and will therefore increase in magnitude with faster movements. Healthy individuals exploit interaction torques that assist with the intended movement and counter them when they oppose the intended movement. Because the muscle activity needs to be produced at the same time as the interaction torques, the descending commands must be generated before the movement occurs. Therefore, the nervous system must have some mechanism for predicting (rather than simply reacting to) these intersegmental dynamics (see Figure 1).

Studies of individuals with cerebellar lesions now indicate that limb ataxia may result from an inability to account for (either exploit or counter) intersegmental interaction torques normally. When cerebellar patients move their limbs fast, the torque produced by their muscles often does not counter the interaction torque appropriately, and the interaction torque contributes excessively to the generation of the movement. This results in an incoordinated reaching pattern, with the elbow and shoulder joints moving at inappropriate rates relative to one another, and movement hypermetria. On the other hand, when cerebellar patients move slowly, they tend to use a decomposition strategy. That is, they reach in a series of shoulder and then elbow movements individually, and the resulting movement is often hypometric (i.e., they undershoot the target; see Figure 2).

Balance and Posture

Patients with cerebellar damage frequently show abnormalities of posture and balance. The typical standing posture of a person with cerebellar damage is wide-based, often with the legs outwardly rotated to maximize the base of support. Also commonly seen is titubation, which is an abnormal bobbing-type motion especially involving the head, neck, and trunk. Standing postural sway is generally increased and can worsen under conditions of reduced or no vision. Specific patterns of postural sway vary depending on the location of the cerebellar lesion. If the anterior lobe is the primary

structure affected, sway is typically increased mostly in the anterior–posterior dimension, whereas with midline (the vermis and fastigial nuclei) cerebellar damage, postural sway is increased in all directions. Lateral cerebellar lesions tend to result in very minimal postural instability (see Figure 3).

Another characteristic balance deficit caused by cerebellar damage is abnormal responses to external perturbations. When the standing surface is quickly and unexpectedly displaced backward, cerebellar patients show excessive surface reactive extensor torque responses and an increased duration of the muscle activity. The behavior produced is one of reduced forward sway (caused by the movement of the support surface), followed by severely hypermetric reactive backward sway (caused by the muscle response to the perturbation).

Similarly, cerebellar patients also show impaired balance in response to perturbations generated by movements of their own body parts. Normally, when the arm is moved rapidly, muscles in the trunk and legs become active concurrently and even before the arm movement, to counter the inertial effects of the arm and maintain upright stance while minimizing postural sway and disequilibrium. Patients with cerebellar damage produce abnormally timed sequences of preparatory and concurrent postural muscle activity relative to the arm raise. Specifically, they often generate leg and trunk muscle activity that occurs too early to be effective in countering the inertial effects of the arm movement.

One idea that may explain the postural instability seen during self-generated movements is that cerebellar damage disrupts the ability to predict the intersegmental dynamics of the limbs and how the limb movements will produce torques at the trunk and other body segments that are not moving.

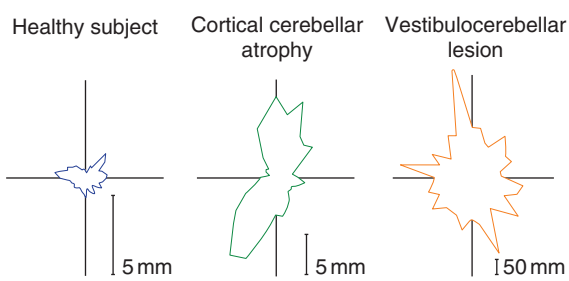

Figure 3 Sway position (center of pressure) histograms (bin width 10° polar; sampling frequency 30 ms; time 60 s) from a typical healthy participant (left), a cerebellar patient with anterior lobe cortical atrophy (middle), and a cerebellar patient with a vestibulocerebellar medulloblastoma (right). Note the scale changes across plots. The cerebellar patient with anterior lobe atrophy shows moderate instability in the anterior–posterior dimension. The cerebellar patient with vestibulocerebellar damage shows severe instability that is omnidirectional. From Dichgans J and Mauritz KH (1983) Patterns and mechanisms of postural instability in patients with cerebellar lesions. *Advances in Neurology* 39: 633–643.

Gait

Gait ataxia is probably the most common manifestation of cerebellar damage; it is often the first clinical sign of damage to the cerebellum and is also one of the most troublesome symptoms for patients. Cerebellar gait ataxia is characterized by a wide base of support and a walking path that does not follow a straight line but instead veers in different directions, giving the appearance of stumbling or drunkenness.

Quantitative studies of cerebellar gait have shown that coordination across the joints of the leg (intralimb coordination) and across the two legs (interlimb coordination) is impaired. Movements of each of the individual leg segments are timed incorrectly relative to one another and the magnitude scaled inappropriately, resulting in poor intralimb coordination. This sometimes manifests in a high-steppage gait pattern, where the foot elevation during swing phase is much higher than required to clear the foot. Inappropriate timing and scaling deficits also affect the position of one leg relative to the other, causing impaired interlimb coordination. The steps become irregular (excessively variable) in terms of the placement and timing of footfalls on one side relative to the other.

Another common feature of cerebellar gait is decomposition of movement. Decomposition is thought to reflect a compensatory strategy (rather than a primary motor deficit) to reduce the number of degrees of freedom and thereby minimize the complexity of the task. It is accomplished by breaking down a multijoint movement into a sequence of single joint movements. A common form of decomposition seen in patients with gait ataxia is a stiffening of the knee joint, probably produced through cocontraction of the muscles around that joint. In that way, the interaction torques that would be produced at other joints by virtue of knee motions are reduced.

Some aspects of intralimb leg incoordination during walking are due to deficits in dealing with interaction torques (e.g., stepping over obstacles). However, most of the observed deficits in walking relates closely to dynamic balance impairments, that is, deficits in coordinating the trunk and the two legs to maintain the body's center of mass over the feet. It has been shown that cerebellar patients with relatively minor balance impairments but relatively severe leg coordination impairments do not exhibit many of the features of gait ataxia. However, cerebellar patients with fairly minor leg coordination impairments but significant balance impairments show nearly all the signs of gait ataxia (see Figure 4).

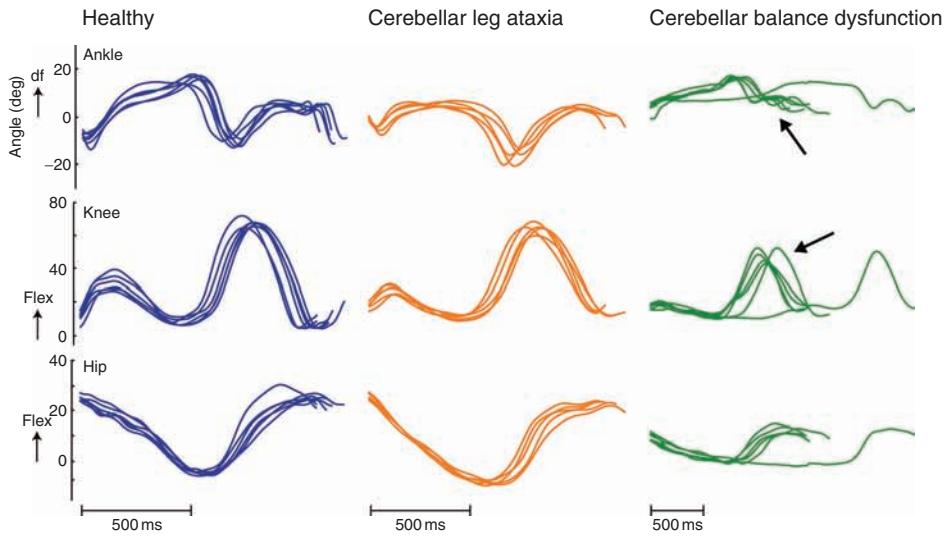

Figure 4 Joint angles plotted against time for the ankle, knee, and hip during level walking. Examples are shown from a typical healthy participant (left), a cerebellar patient with predominantly ataxia of the limbs and minimal balance impairments (middle), and a cerebellar patient with predominantly balance impairments and minimal limb ataxia. Several single strides are overlaid for each participant, and a stride is from foot contact to the next foot contact on that leg. Note the timescale changes across plots. The limb movement-impaired cerebellar patient (who has no balance deficit) walks similarly to the healthy participant. The balance-impaired cerebellar patient (who has no voluntary leg coordination deficit) shows many features of gait ataxia, including reduced angular excursions, joint–joint decomposition, increased stride-to-stride variability, and reduced speed (see arrows). Reproduced from Morton SM and Bastian AJ (2003) Relative contributions of balance and voluntary leg-coordination deficits to cerebellar gait ataxia. *Journal of Neurophysiology* 89: 1844–1856, used with permission from the American Physiological Society.

Motor Learning

The cerebellum is also essential for motor adaptation, which is an error-driven learning process used to adjust movements to new sensory and/or motor demands through practice. Adaptation occurs relatively quickly, over minutes to hours, rather than days or weeks. Changes clearly occur in the earliest component of the movement, indicating that the adaptation affects feed-forward or predictive commands to the body. The new sensorimotor calibrations are stored, causing aftereffects that persist when the new demands are removed. Adaptation is important for keeping motor commands appropriately adjusted for changes in the environment or physical aspects of the body (e.g., growth, strength). Without it, an individual's ability to move smoothly and accurately suffers.

People with cerebellar damage show impairments in or loss of adaptation of movements to altered situations or novel demands. An example of this occurs when wearing laterally deviating prism glasses, which bend light and shift viewed objects to the right or the left. Movements toward the object are initially misdirected in the same direction as the prism shift. Healthy participants can adapt with practice to regain accuracy, and they show aftereffects in the opposite direction when the prisms are removed. People with cerebellar damage are slow or unable to adapt their movements to the prism shift and show reduced or absent aftereffects. Similar adaptation deficits occur during walking and reaching under altered dynamic or visual demands. This form of learning is therefore thought to be cerebellum dependent. Such deficits in adaptation raise the important question of whether rehabilitation treatments aimed at altering movement patterns can be learned by people with cerebellar dysfunction. Longer-term training studies have not been tested and might provide important insights into whether the cerebellum is needed for very slow motor learning processes (see Figure 5).

A Common Mechanism?

The fundamental cerebellar operation is still not understood. It has been proposed to be a movement combiner, an internal model of limb dynamics, a timer, a prediction mechanism, and/or a learning machine. Several of these ideas are based on descriptions of the deficits in one or a few movement types following cerebellar damage. Some theories are more clearly formalized, others less so, and none are mutually exclusive.

One idea is that the cerebellum helps combine simple movement elements together to produce a new, single, coordinated motion. This idea is consistent with the fact that patients have more difficulty

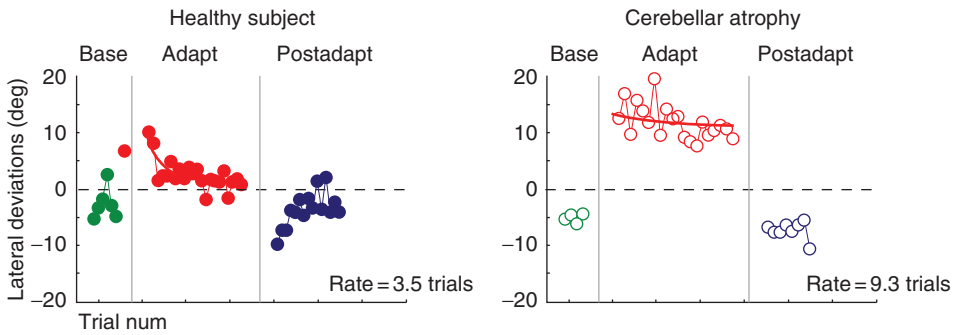

Figure 5 Maximal lateral deviations of the walking trajectory for a typical healthy participant (left) and a patient with cerebellar atrophy (right) plotted for each walking trial before (baseline), during (adaptation), and after (postadaptation) prism exposure while walking. The cerebellar patient adapts more slowly and to a lesser extent and shows a diminished aftereffect in the postadaptation phase. The adaptation rate (in numbers of trials) for the two participants is indicated on each plot. Adapt, adaptation; postadapt, postadaptation. Reproduced from Morton SM and Bastian AJ (2004) Prism adaptation during walking generalizes to reaching and requires the cerebellum. *Journal of Neurophysiology* 92: 2497–2509, used with permission from the American Physiological Society.

controlling multijoint than single joint movements and compensate by decomposing or breaking motions down into elemental components. One way that the cerebellum could contribute to this process is by computing the necessary adjustments to a motor command to account for mechanical interactions between body segments. This type of internal model would be critical for coordinated control of many parts of the body.

Another theory is that the cerebellum acts as a general timing mechanism, clocking short intervals in the millisecond to second range. This theory is consistent with the fact that cerebellar damage can cause abnormal timing of muscle activity and that patients can even have difficulty with timing perception. Precise timing control would be important for computing appropriate movement commands, regardless of how many joints or body parts were being controlled. An open question is whether timing deficits can explain the finding that more complex movements are so much more impaired than simple ones.

The cerebellum may also function as a prediction mechanism. It may function by helping to estimate the optimal motor commands, upcoming body states (i.e., position, velocity), or both. The prediction hypothesis is supported by work showing that cerebellar damage produces greater impairment of movements requiring predictive as opposed to reactive control in many different types of movement. Predictive control is critical for fast and ballistic movements, but impaired prediction can also affect slower movements, due to increased reliance on time-delayed feedback signals.

These three ideas are inherently interrelated, making it difficult to distinguish them behaviorally. Whether the cerebellum acts as a combiner, a timer, and/or a predictor, it is clear that cerebellar control must be learned and maintained through error-dependent learning mechanisms.

Summary

The cerebellum is a subcortical structure that is important for controlling and learning movement. Damage to the cerebellum results in incoordination, or ataxia, but not loss of movement, or paralysis. Different longitudinal cerebellar zones are used for control of different types of movement based on their connectivity with other brain areas. There are distinct signs of limb, balance, and gait ataxia, which are exacerbated by moving multiple joints together and by moving quickly. Thus, some cerebellar deficits seem to result from a failure of motor commands to account for mechanical interactions between moving body parts. Cerebellar patients also show deficits in timing control and in generating feed-forward or predictive motor commands. The fundamental cerebellar operation is not known, but it may act as a movement combiner, a timer, and/or a prediction mechanism. These types of computations are inherently different but could conceivably result in similar-looking deficits. More work is required to clarify the cerebellar operation. What is clear is that optimal cerebellar control is maintained through adaptive learning processes. These processes appear to operate on relatively short timescales to keep movement well calibrated for varying environmental demands.

See also: Balance and Posture Control; Balance and Posture Control: Human; Cerebellum and Oculomotor Control; Cerebellum: Clinical Pathology; Cerebellum: Models; Corticospinal Development; Fragile X

Associated Tremor Ataxia Syndrome; Friedreich Ataxia; Long-Term Depression: Cerebellum; Procedural Learning: Cerebellum Models.

Further Reading

Bastian AJ, Martin TA, Keating JG, and Thach WT (1996) Cerebellar ataxia: Abnormal control of interaction torques across multiple joints. *Journal of Neurophysiology* 76: 492–509.

Bastian AJ, Zackowski KM, and Thach WT (2000) Cerebellar ataxia: Torque deficiency or torque mismatch between joints? *Journal of Neurophysiology* 83: 3019–3030.

Boyden ES, Katoh A, and Raymond JL (2004) Cerebellum-dependent learning: The role of multiple plasticity mechanisms. *Annual Review of Neuroscience* 27: 581–609.

Dichgans J and Mauritz KH (1983) Patterns and mechanisms of postural instability in patients with cerebellar lesions. *Advances in Neurology* 39: 633–643.

Doya K (2000) Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology* 10(6): 732–739.

Ito M (2005) Bases and implications of learning in the cerebellum - adaptive control and internal model mechanism. *Progress in Brain Research* 148: 95–109.

Ivry RB, Spencer RM, Zelaznik HN, and Diedrichsen J (2002) The cerebellum and event timing. *Annals of the New York Academy of Sciences* 978: 302–317.

Kawato M (1999) Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* 9(6): 718–727.

Miall RC (1998) The cerebellum, predictive control and motor coordination. *Novartis Foundation Symposium* 218: 272–284.

Morton SM and Bastian AJ (2003) Relative contributions of balance and voluntary leg-coordination deficits to cerebellar gait ataxia. *Journal of Neurophysiology* 89: 1844–1856.

Morton SM and Bastian AJ (2004) Cerebellar control of balance and locomotion. *Neuroscientist* 10(3): 247–259.

Morton SM and Bastian AJ (2004) Prism adaptation during walking generalizes to reaching and requires the cerebellum. *Journal of Neurophysiology* 92: 2497–2509.

Ohyama T, Nores WL, Murphy M, and Mauk MD (2003) What the cerebellum computes. *Trends in Neurosciences* 26(4): 222–227.

Thach WT, Goodkin HP, and Keating JG (1992) The cerebellum and the adaptive coordination of movement. *Annual Review of Neuroscience* 15: 403–442.

Wolpert DM, Miall RC, and Kawato M (1998) Internal models in the cerebellum. *Trends in Cognitive Sciences* 2(9): 338–347.

Cerebellar Microcircuitry

M Ito, RIKEN Brain Science Institute, Wako, Japan

© 2009 Elsevier Ltd. All rights reserved.

Neuronal Elements Constituting Cerebellar Microcircuits

The cortical sheet covering the cerebellar surface has a laminar structure composed of the most superficial, molecular layer; the intermediate Purkinje cell layer; and the deepest, granular layer. These layers contain seven major neuronal types, namely, granule cells, Purkinje cells, basket cells, stellate cells, Golgi cells, Lugaro cells, and unipolar brush cells, as illustrated in Figure 1. In addition, the cerebellar cortex receives mossy fiber afferents, climbing fiber afferents, and the third type of afferent through the white matter underlying the granular layer. Purkinje cell axons provide the sole output of the cerebellar cortex and project through the white matter toward the cerebellar and vestibular nuclei. The following sections provide brief descriptions of neuronal elements of cerebellar microcircuits.

Mossy Fiber Afferents

These arise from the brain stem and spinal cord. Some of them arise directly from peripheral nerves, but most are relayed by various precerebellar nuclei. A mossy fiber afferent terminates in the granular layer, forming a rosette that is the core of the characteristic glomerulus structure. Branches of one mossy fiber supply excitatory synapses to 400–600 granule cells.

Granule Cells

These are individually the smallest (5–8 μm in diameter) but the most numerous neurons (2–3 million mm^{-3}, for a total of 10^{10}–10^{11} in humans) in the brain. Each granule cell extends four to five short dendrites, each of which receives an excitatory synapse from a mossy fiber terminal. Granule cell axons ascend through the Purkinje cell layer and molecular layer perpendicularly to the cortical surface and bifurcate to parallel fibers that extend 4–6 mm across the T-junction in the molecular layer. Granule cell axons supply excitatory synapses to other neurons in the cerebellar cortex.

Purkinje Cells

They are large inhibitory neurons lying in one layer at a density of 300–500 mm^{-2}. The human cerebellum contains 15 million Purkinje cells. Each Purkinje cell extends planar dendrites, through which numerous parallel fibers pass. Parallel fibers supply a Purkinje cell with as many as 175 000 excitatory synapses. The ascending segments of granule cell axons supply about 20% of the excitatory synapses that dendrites of each Purkinje cell receive from granule cells. Purkinje cells also receive excitatory synapses from climbing fibers and the third type of afferent, and inhibitory synapses from basket and stellate cells. Purkinje cell axons extend through the white matter to subcortical structures.

Basket Cells

Basket cells are located in the molecular layer near Purkinje cells at a density sixfold that of Purkinje cells. Basket cells receive excitatory synapses from parallel fibers and in turn supply γ-aminobutyric acid (GABA)-mediated inhibitory synapses to the bottleneck of a Purkinje cell soma, forming a unique complex structure called a 'pinceau.' One basket cell may inhibit 50 Purkinje cells, whereas one Purkinje cell may receive synapses from 20–30 basket cells. Basket cells also receive excitatory synapses from collaterals of climbing fibers.

Stellate Cells

Stellate cells are located in the superficial two-thirds of the molecular layer. There are 16–17.5 fold more stellate cells than Purkinje cells. Distinct from basket cells, stellate cells supply GABA-mediated inhibitory synapses only to dendrites of Purkinje cells.

Golgi Cells

In the granular layer, there is one Golgi cell for every 15 Purkinje cells. Each Golgi cell receives ~4788 excitatory synapses from parallel fibers on its apical dendrites extending to the molecular layer and also 228 mossy fiber terminals on their descending dendrites. A Golgi cell in turn extends a broadly branching axon to as many as 5700 granule cells and supplies GABA-mediated inhibitory synapses.

Lugaro Cells

Lugaro cells are located in or slightly below the Purkinje cell layer. Their axons supply GABA-mediated inhibitory synapses to apical dendrites of Golgi cells. The cerebellar cortex contains one Lugaro cell per Golgi cell. Ten Lugaro cells converge onto only one Golgi cell, whereas the axon of one Lugaro cell diverges onto 150 Golgi cells. Lugaro cells receive serotonergic synapses, and in the presence of serotonin, they discharge spikes regularly at 5–15 Hz.

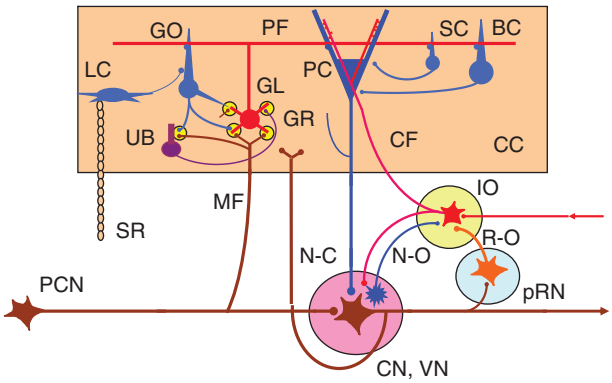

Figure 1 Wiring diagram for a cerebellar corticonuclear microcomplex. BC, basket cell; CC, cerebellar cortical microzone; CF, climbing fiber; CN, cerebellar nucleus; GR, granule cell; GL, glomeruli; GO, Golgi cell.; IO, inferior olive; LC, Lugaro cell; MF, mossy fiber; N-C, nucleocortical mossy fiber projection; N-O, nucleoolivary inhibitory projection; PC, Purkinje cell; PCN, precerebellar neuron; PF, parallel fiber; pRN, parvicellular red nucleus; R-O, rubroolivary excitatory projection; SC, stellate cell; SR, serotonergic fiber; UB, unipolar brush cell; and VN, vestibular nucleus. From Ito M (2006) Cerebellar circuitry as a neuronal machine. *Progress in Neurobiology* 78(3–5): 272–303.

Unipolar Brush Cells

Unipolar brush cells are located primarily in the granular layer of the vestibulocerebellum, which is the portion of the cerebellum receiving primary vestibular afferents. A unipolar brush cell receives an excitatory synapse on its dendritic brush from a single mossy fiber terminal, thus forming a giant excitatory synapse. Axons of unipolar brush cells branch and give rise to large terminals that, within gromeruli, synapse with both granule cell and unipolar brush cell dendrites. In this sense, unipolar brush cells are the intracortical source of mossy fibers.

Climbing Fibers

Climbing fibers are a unique structure of the cerebellum with no homologs elsewhere in the central nervous system. They originate from the inferior olive in the medulla oblongata. Each Purkinje cell is innervated by one climbing fiber, which forms numerous synaptic contacts (e.g., 26 000 in rat) with dendrites of a single Purkinje cell. This arrangement provides particularly large excitatory postsynaptic potentials superimposed with Ca^{2+} spikes.

Climbing fibers are now considered to convey error signals instrumental to synaptic plasticity and learning in the cerebellar cortex (see the section titled 'Synaptic plasticity'). Climbing fiber signals are triggered in many cases by sensory stimuli, but whether climbing fibers convey sensory or motor information (motor information may be converted from sensory information) remains controversial.

Third Type of Afferent

Afferents of the third type contain various amines (e.g., serotonin, dopamine, acetylcholine, norepinephrine, and histamine) or peptides (e.g., orexin, angiotensin II, and corticotropin-releasing factor). This third type of afferent generally extends fine varicose fibers sparsely throughout the granular and molecular layers and forms direct contacts with Purkinje cells and other cerebellar neurons. Serotonin-containing fibers arise from raphe nuclei, whereas norepinephrine-containing fibers originate from the nucleus locus coeruleus. Histamine-containing fibers and many neuropeptide-containing fibers arise from the hypothalamus. It has long been suggested, but with scarce evidence, that the third type of afferent plays roles in emotion and behavior.

Cerebellar Nuclear Neurons

A major group of neurons in cerebellar nuclei are the target of GABA-mediated inhibitory synapses supplied by Purkinje cells. Purkinje cells provide 73% of the total synapses to nuclear neurons, including almost all the somatic synapses of nuclear neurons. Cerebellar nuclear neurons also receive excitatory synapses from collaterals of mossy fiber and climbing fiber afferents. Cerebellar nuclear neurons send their axons to extracerebellar structures, but their collaterals pass through the cerebellar cortex as mossy fibers. Cerebellar nuclei also contain small neurons whose axons supply GABA-mediated inhibitory synapses to inferior olive neurons. Some neurons in the vestibular nuclei (e.g., Deiters neurons and vestibuloocular relay neurons) also receive inhibitory synapses from Purkinje cells.

Parvicellular Red Nucleus Neurons

Neurons of the parvicellular red nucleus project to the inferior olive, in contrast to neurons of the magnocellular red nucleus, which project to the spinal cord. The parvicellular red nucleus–inferior olive pathway develops in association with phylogenetically newer parts of the cerebellum. Neurons in the parvicellular red nucleus receive excitatory synapses from cerebellar nuclear neurons and in turn supply excitatory synapses to inferior olive neurons that extend climbing fiber afferents.

Local Microcircuit Mechanisms

The above-introduced neuronal elements are interconnected to form various prototypes of local microcircuit.

Two-Stepped Divergence and Convergence

The major pathway across the cerebellar cortex consists of mossy fiber afferents, granule cells, and Purkinje cells, which are connected forwardly by the repetition of divergence and convergence. This is the fundamental substrate of the multilayered computational networks of the cerebellum as hypothesized by David Marr and James Albus around the end of the 1960s. Information received by the first layer, composed of mossy fibers, is spread in the 'hyperspace' formed by a large number of granule cells in the second layer. Purkinje cells in the third layer then receive signals from selected granule cells.

Synaptic Plasticity

About 200 000 parallel fibers and one climbing fiber converge onto one Purkinje cell. Around 1970, theorists speculated the meaning of this unusually asymmetric convergence of two inputs. Because climbing fiber signals inevitably activate Purkinje cells through the close contact between a climbing fiber and Purkinje cell dendrites, the conjunctive activation of parallel fibers and the climbing fiber leads to the Hebbian condition for synaptic plasticity. That is, the potentiation of synaptic transmission occurs when presynaptic and postsynaptic elements are conjointly activated. Because the actually discovered synaptic plasticity is long-term depression (LTD) and because climbing fibers convey error signals, it has been postulated that climbing fiber signals eliminate those parallel fiber synapses involved in an erroneous performance of a cerebellar circuit, in the manner of error-driven learning.

Cerebellar microcircuits involve also other types of synaptic plasticity. For example, long-term potentiation (LTP) occurs in parallel fiber–Purkinje cell synapses stimulated without conjunctive climbing fiber activation. This LTP balances LTD evoked by conjunction. Inversely, the conjunctive stimulation of a parallel fiber and a climbing fiber induces LTP in parallel fiber synapses with basket or stellate cells, and LTD is induced in them when only parallel fibers are activated. The synaptic plasticity in basket and stellate cells therefore operates oppositely to that in Purkinje cells and would have a synergistic effect on the latter. LTP occurs in the mossy fiber extending to granule cell synapses and would have the effect of adjusting the size of the granule cell hyperspace. A weak LTD occurs in tetanized climbing fiber–Purkinje cell synapses.

Feed-Forward Inhibition

Parallel fiber impulses are transmitted to basket and stellate cells, whose inhibitory signals are further transmitted to Purkinje cells. This pathway provides a feed-forward inhibition that parallels the excitatory pathway from granule cells to Purkinje cells. When the cerebellar surface is viewed from above, basket and stellate cells extend axons perpendicularly to parallel fibers. Hence, excitation propagating along a parallel fiber bundle will be flanked by inhibition mediated by basket or stellate cells. This is a spatial pattern of lateral inhibition. However, because such a pattern has not been observed to arise spontaneously, it could be a result of the electrical stimulation of a parallel fiber bundle.

Another potentially important function of feed-forward inhibition is to curtail sharply the parallel fiber-mediated excitatory synaptic effect by a slightly (1–4 ms) delayed inhibitory synaptic effect. By this mechanism, feed-forward inhibition could increase temporal precision and thereby minimize the randomness of circuit operation. This temporal fidelity justifies the use of a network model, with its binary input and output, for modeling the operation of the cerebellar cortex.

Feedback Inhibition

The loop through which granule cells excite Golgi cells that, in turn, inhibit granule cells represents a typical microcircuit for feedback inhibition. Computer simulation predicts that this feedback inhibition from Golgi to granule cells induces 10–50 Hz oscillations in spike discharges from the latter. Another possibility is that the circuit involving mossy fibers, granule cells, parallel fibers, and Golgi cells constitutes a phase converter, whose function is to generate a set of multiphase versions of a mossy fiber input. This phase converter (an 'alike' device) is essential in Masahiko Fujita's adaptive filter model of the cerebellum. Recently, another unique model has been proposed. It features a randomly connected granule cell–Golgi cell loop pathway, which operates as a clock for generating granule cell discharge in an ever changing ensemble of patterns. The patterns do not repeat unless reset by a large mossy fiber input to the pathway.

Reciprocal Inhibition

Basket and stellate cells in the molecular layer are reciprocally connected via inhibitory synapses. They are also linked to each other through electrical synapses. A computer simulation suggests that reciprocal inhibition causes an oscillation of the activity of basket and stellate cells at 100–250 Hz. This is a substantially higher rate of oscillation than that induced by Lugaro cells (5–15 Hz). Reciprocal inhibition also occurs among Purkinje cells via their recurrent collaterals that extend to neighboring Purkinje

cells within 300 μm of the parent cell. Mutual inhibition also occurs between Purkinje cells and basket cells, suggesting a possibility that Purkinje cells are involved in a mixed reciprocally inhibitory network containing Purkinje cells and basket cells.

Noise Generator

Spontaneous discharges of Purkinje cells are characteristically irregular. These irregular discharges are not caused by excitatory synaptic drives but, unexpectedly, by tonic influences from basket or stellate cells. The pharmacological blockade of ionotropic and metabotropic glutamate receptors does not affect spontaneous Purkinje cell discharges, whereas the blockade of $GABA_A$ receptors increases the rate of such discharges and makes the discharges regular. Similar irregular discharges and their dependence on tonic inhibition were also observed in basket and stellate cells that receive tonic inhibition from themselves. How noise is generated in basket and stellate cells has not been clarified. A modeling study suggests that noise helps Purkinje cells and basket and stellate cells respond rapidly, sensitively, and linearly to external inputs.

Modular Structure: Microzone and Microcomplex

The microcircuit structure is uniform throughout the cerebellum. However, the cerebellum is divided by transverse grooves into ten lobules (I–X) and also by various markers into seven longitudinal bands (A, B, C1, C2, C3, D1, and D2). Each division of the cerebellar cortex is further divided into microzones. A microzone combines with a set of subcortical structures including small groups of neurons in (1) a cerebellar or vestibular nucleus, (2) the inferior olive, and (3) the parvicellular red nucleus in phylogenetically newer parts of the cerebellum. These conjointly constitute a corticonuclear microcomplex.

A microcomplex represents a unitary neuronal machine of the cerebellum devoted to a specific function. The presence of a microcomplex in the intermediate part of the cerebellar hemisphere has been confirmed by a combination of zebrin immunohistochemical analysis and tracer techniques. A typical microcomplex structure including the flocculus and vestibular nuclei has been dissected in detail in association with horizontal, vertical, and rotatory canal–ocular reflexes. A microcomplex involved in the horizontal canal–ocular reflex was systematically labeled by retrograde transneuronal tracing with the rabies virus from a medial rectus muscle of guinea pigs. The microzonal structure was studied in detail in the C3 zone of the anterior lobe, which receives mossy fiber and climbing fiber inputs from receptive fields in the forelimb skin. This area contains 30–40 longitudinal microzones lying side by side, each of which is 50–150 mm wide. In each microzone, Purkinje cells and basket or stellate cells receive climbing fiber inputs from the same cutaneous receptive field. Climbing fibers in adjacent microzones are activated by stimulation of adjacent skin areas. Adjacent microzones in turn innervate adjacent cell groups in the anterior interpositus nucleus, which controls movement components that have specific relationships with the location of climbing fiber receptive fields. Assuming that a microzone is 10 mm wide, it may be calculated that the human cerebellum contains on the order of 5000 microcomplexes, each of which plays a specific role via its connections with functional systems in the spinal cord, brain stem, and cerebral cortex.

Each microcomplex receives mossy fiber inputs and sends out outputs of cerebellar or vestibular nuclear neurons. Learning can modify the input–output relationship of a microcomplex with reference to error signals conveyed by climbing fiber inputs. The roles of the third type of afferent have not been clarified, but it is probable that these afferents set the functional status of a microcomplex at different levels or in different modes. A microcomplex thus appears to be a unit machine representing a multipurpose adaptive computer chip.

Incorporation of Microcomplex into Segmental Circuits

Cerebellar microcomplexes are interconnected with various reflex centers in the spinal cord and brain stem. For example, the vestibuloocular reflex, optokinetic eye movement response, ocular following response, eyeblink conditioning, and baroreceptor reflex for arterial blood pressure have been studied as simple motor or autonomic systems representing functions of the cerebellum. The cerebellum is also connected to segmental circuits for saccadic eye movement, interlimb coordination during locomotion, and the operant conditioning of the H-reflex.

These microcomplexes interconnected with segmental microcircuits subserve adaptive mechanisms required for acquiring and maintaining precise motor or autonomic control. When a microcomplex changes its dynamic characteristics according to error signals conveyed by climbing fibers, this learning will lead to the optimal tuning of dynamic characteristics of the segmental motor or autonomic centers. This tuning should minimize error signals of climbing fibers.

Incorporation of Microcomplex in Cerebral Cortical Circuits: Internal Model for Voluntary Movements

The involvement of the cerebellum in voluntary movements has been studied using various experimental paradigms since Tom Thach's pioneering work that started in the 1960s: for example, wrist flexion in response to a sudden load perturbation, the operant conditioning of wrist movements to lift a lever in response to visual stimuli, throwing while looking through prisms, reaching, the hand manipulation of a computer mouse, and finger gripping. The smooth pursuit eye movement is also a type of voluntary movement.

A characteristic circuit arrangement for the cerebellar control of voluntary movements is the cerebrocerebellar communication loop. In this loop, the motor cortex projects mossy fiber afferents to the intermediate part of the cerebellar hemisphere via the pontine nucleus, and cerebellar outputs from the interpositus nucleus pass back to the motor cortex via the thalamus. A recent study using the transneuronal transport of neurotropic viruses by Peter Strick's group confirmed that neurons in the arm area of the motor cortex receive inputs from Purkinje cells located primarily in lobules IV–VI of the cerebellar cortex. In turn, neurons in the arm area project back via the pons to granule cells primarily in the same three cerebellar lobules.

It has been hypothesized that, through this communication loop, the cerebellum provides an internal feedback that may replace the external feedback via a sensory system. If the cerebellum contains an accurate model of the lower motor system including a motor apparatus, the internal feedback through the model should enable us to perform a precise movement even without the external feedback. This hypothesis explains how one can move correctly even with the eyes closed and how dysmetria characterizes patients with an impaired cerebellum. The internal model subserving the internal feedback is called the forward model. Motor learning can be considered to be a process by which the forward model is formed and reformed in the cerebellum through LTD-based error-driven learning.

Another important circuit connects the cerebral cortex and cerebellum in parallel. For example, neurons in the supplementary motor cortex and premotor cortex project to the motor cortex, and at the same time, they project to the cerebellum via the pontine nucleus. A cerebellar microcomplex can then be incorporated parallel to the motor cortex. When the cerebellar cortex operates in parallel with the motor cortex, it forms another type of internal model that bears a transfer function reciprocally equal to the dynamics of the controlled object. The inverse model proposed by Mitsuo Kawato's group acts as a feed-forward controller that replaces the motor cortex serving as a feedback controller. The forward and inverse models and their combination have been applied successfully to the construction of a robot that is able to learn movement skills.

Incorporation of Microcomplex in Cerebral Cortical Circuits: Internal Model for Mental Functions

The most lateral part of the human cerebellum evolved in parallel to the expansion of cerebral association cortices. Recent virus mapping in monkeys has revealed that the projections to the motor and premotor cortices originate from the dorsal portions of the dentate nucleus, whereas projections to the prefrontal and posterior parietal areas of the cortex originate from the ventral portions of the dentate nucleus. The anterior intraparietal area of the posterior parietal cortex also receives projections from a broad area of the dentate nucleus. On the other hand, Purkinje cells located primarily in crus II of the ansiform lobule receive inputs from prefrontal area 46 and project to the same area 46. Therefore, a closed cerebrocerebellar communication loop is also formed between the prefrontal cortex and cerebellar hemisphere.

The presence of the cerebrocerebellar communication loop between the prefrontal cortex and cerebellar hemisphere supports the extension of the above-mentioned forward model hypothesis for voluntary movements to mental functions of the cerebral cortex. The internal model hypothesis has been extended conceptually to apply to mental functions such as thought. To do this, it is essential to assume that thought is the manipulation of mental models that are small-scale models of reality used by the mind to anticipate events requiring reason and an explanation. One may suppose that, in thought, a mental model of an image, idea, or concept is formed in the temporoparietal association cortex. The prefrontal cortex as the controller manipulates the mental model as the controlled object, just as the motor cortex controls the voluntary movements of body parts. Then, the cerebellum copies a mental model to form an internal model. Thus, thought proceeds with the internal models of the mental models as the controlled objects.

Because processes occurring in the cerebellum do not reach the level of awareness, the internal model hypothesis explains how we become able to conduct movements and thoughts unconsciously after

repeated exercises. For example, an idea pops out even without an obvious conscious effort to think it out! This hypothesis also explains diverse psychological and pathological mental experiences such as delusion, hallucination in schizophrenia, and the lack of sympathy for other persons as being caused by an impairment of a cerebellar internal model.

The inverse model postulated above for voluntary movements may also apply to thought. In this hypothesis, an inverse model substitutes the executive function of the prefrontal cortex. If both the forward and inverse models are formed, the entire process of thought could be performed within the cerebellum; the thought could then proceed in complete unconsciousness. This possibility may be relevant to the current intriguing question of whether the mental force we feel in initiating a voluntary movement is illusory. It has been argued that the conscious self feels the mental force as an experience of a movement initiated by unconscious processes, which could proceed within the cerebellum.

Conclusion

The elaborate structure of cerebellar microcircuits and their organization in a microcomplex as a unit machine have been well analyzed. The mechanism by which cerebellar microcircuits subserve various types of neural control has been clarified in terms of adaptive control and internal models. However, dynamic mechanisms operating in cerebellar microcircuits have to be investigated further, and in this respect, a realistic modeling of cerebellar microcircuits becomes increasingly important. A special interest in the field of neuroscience is the mechanism by which information not only of a physical nature in movements but also of a cognitive nature in mental activities is processed and stored in cerebellar microcircuits.

See also: Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Cerebellar Deep Nuclei; Cerebellum: Evolution and Comparative Anatomy; Cerebellum: Models; Eyeblink Conditioning; Olivocerebellar System; Red Nucleus.

Further Reading

Barlow JS (2002) *The Cerebellum and Adaptive Control*. Cambridge, UK: Cambridge University Press.

Blakemore SJ and Sirigu A (2003) Action prediction in the cerebellum and in the parietal lobe. *Experimental Brain Research* 153: 239–245.

Broussard DM and Karrardjian CD (2004) Learning in a simple motor system. *Learning & Memory* 11: 127–136.

De Schutter E (1999) Using realistic models to study synaptic integration in cerebellar Purkinje cells. *Reviews in the Neurosciences* 10: 233–245.

Eccles JC, Ito M, and Szentagothai J (1967) *The Cerebellum as Neuronal Machine*. Heidelberg: Springer.

Ekerot C-F and Jorntell H (2003) Parallel fiber receptive fields: A key to understanding cerebellar operation and learning. *Cerebellum* 2: 101–109.

Hansel C, Linden D, and D'Angelo E (2000) Beyond parallel fiber LTD: The diversity of synaptic and non-synaptic plasticity in the cerebellum. *Nature Neuroscience* 4: 467–475.

Ito M (1984) *The Cerebellum and Neural Control*. New York: Raven Press.

Ito M (2005) Bases and implications of learning in the cerebellum: Adaptive control and internal model mechanism. In: De Zeeuw C and Cicirate F (eds.) *Creating Coordination in the Cerebellum*, pp. 95–109. Amsterdam: Elsevier.

Ito M (2006) Cerebellar circuitry as a neuronal machine. *Progress in Neurobiology* 78(3–5): 272–303.

Kelly RM and Strich PL (2003) Cerebellar loops with motor cortex and prefrontal cortex of a nonhuman primate. *Journal of Neurophysiology* 73: 1329–1340.

Schmahmann JD (ed.) (1997) *The Cerebellum and Recognition*. New York: Academic Press.

Schweighofer N, Doya K, and Kuroda S (2004) Cerebellar aminergic neuromodulation: Toward a functional understanding. *Brain Research Reviews* 44: 103–116.

Wegner DM (2004) Précis of the illusion of conscious will. *Behavioral & Brain Sciences* 27: 649–659 (discussion: 659–692).

Wolpert DM, Doya K, and Kawato M (2003) A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 358: 593–602.

Cerebellum and Oculomotor Control

D S Zee and M Walker, The Johns Hopkins Hospital, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The cerebellum plays a pivotal role in the control of all types of eye movements. Its essential function is to optimize ocular motor performance so that images of objects of interest are brought to the fovea – where visual acuity is best – and then kept there quietly to give the brain time to analyze and interpret the visual scene, and so decide what to do next. The cerebellum has both immediate, on-line functions (as reflected in immediate appearance of abnormalities when the cerebellum is damaged), and long-term, adaptive functions that ensure that the ocular motor responses are calibrated correctly to the stimuli that drive them.

Anatomical Approach to the Cerebellum and Ocular Motor Control

The approach here will be based upon structural–functional correlations: cerebellar influences on eye movements will be discussed by anatomical regions (see Table 1 and Figure 1). Since the cerebellum has an important role in motor learning that permits adaptation and compensation for lesions, a rigid compartmentalization of function inferred from enduring effects of lesions within the cerebellum may not strictly apply to the contribution of the cerebellum to the control of eye movements in intact subjects. With this caveat, the ocular motor functions of the cerebellum will be interpreted using the combination of the effects of acute and of chronic lesions, and the results from physiological studies. Here, the emphasis will be on the areas within the cerebellum about which most is known: the flocculus/paraflocculus, the nodulus/ventral uvula (lobules IX and X of the cerebellar vermis), and the dorsal oculomotor vermis (OMV) (lobules V–VII) and its target in the posterior portion of the underlying fastigial nucleus called the fastigial oculomotor region (FOR).

The cerebellum has a distinctive anatomical organization. Purkinje cells of the cerebellar cortex primarily project to and inhibit cells within the underlying deep cerebellar nuclei; the vestibulocerebellum (flocculus, paraflocculus, nodulus, and uvula) also projects to the vestibular nuclei some of which can be thought of as 'displaced' deep cerebellar nuclei. There is a rich innervation to the cerebellar cortex from the inferior olive, as climbing fibers directly to Purkinje cells, and from all the premotor brain stem nuclei concerned with eye movements (and from the vestibular nerve), as mossy fibers to cerebellar granule cells, and then onto Purkinje cells via parallel fibers. There are also direct climbing fiber and mossy fiber inputs to the underlying cerebellar nuclei, so that the cerebellar cortex can be thought of as a parallel pathway that can influence (via the projections of its Purkinje cells) the direct flow of information to and from the brain stem through the deep cerebellar nuclei.

Vestibulocerebellum: The Flocculus/Paraflocculus

The flocculus and paraflocculus together with the caudal portions of the cerebellar vermis (nodulus and uvula) are part of the oldest portion of the cerebellum – the archicerebellum, also called the vestibulocerebellum (Figure 1). Neurons within the flocculus/paraflocculus discharge in relation to various features of the stimulus (e.g., head motion or visual target motion), to the position of the eye in the orbit during fixation, and to the movements of the eye during smooth pursuit, vergence, and vestibular response to head motion (the vestibulo-ocular reflex or VOR).

Pursuit With lesions of the flocculus/paraflocculus smooth tracking of a moving target is impaired, either when the head is still (smooth pursuit) or moving (VOR cancellation). Both the initiation of pursuit and the response during sustained tracking can be affected. Complete lesions of the flocculus/paraflocculus lead to a considerably decreased steady-state gain (eye velocity/target velocity) of pursuit. There is considerable recovery of pursuit function, even with relatively complete floccular/paraflocular lesions. This is likely mediated by other parts of the cerebellum though other parts of the brain might also be involved.

Gaze holding A second cardinal feature of lesions of the flocculus/paraflocculus is impaired gaze holding: after eccentric horizontal eye movements, the eyes drift centripetally (Figure 2). This finding indicates that the flocculus/paraflocculus functions under the control of the brain stem circuits that convert (integrate) velocity into position commands for all types of conjugate eye movements, and so permits steady fixation.

Spontaneous nystagmus A third distinctive feature of lesions in the flocculus/paraflocculus is downbeat nystagmus, in which the eyes drift up (slow phase) and are brought back to the fixation target by a

Table 1 Anatomical localization of ocular motor functions within the cerebellum

Structure	*Function*
Flocculus/paraflocculus	Gaze holding Smooth pursuit and VOR cancellation Control amplitude and direction of rotational VOR Match pulse step to prevent postsaccadic drift
Nodulus/ventral uvula	Modulate velocity-storage mechanism within vestibular nuclei to influence low-frequency rotational VOR • Amplitude and direction (relative to gravity) of rotational VOR • Habituation of rotational VOR • Tilt suppression of postrotatory VOR
Dorsal vermis/fastigial oculomotor region (FOR)	Saccade amplitude and direction Pursuit initiation (open-loop acceleration) Horizontal eye alignment

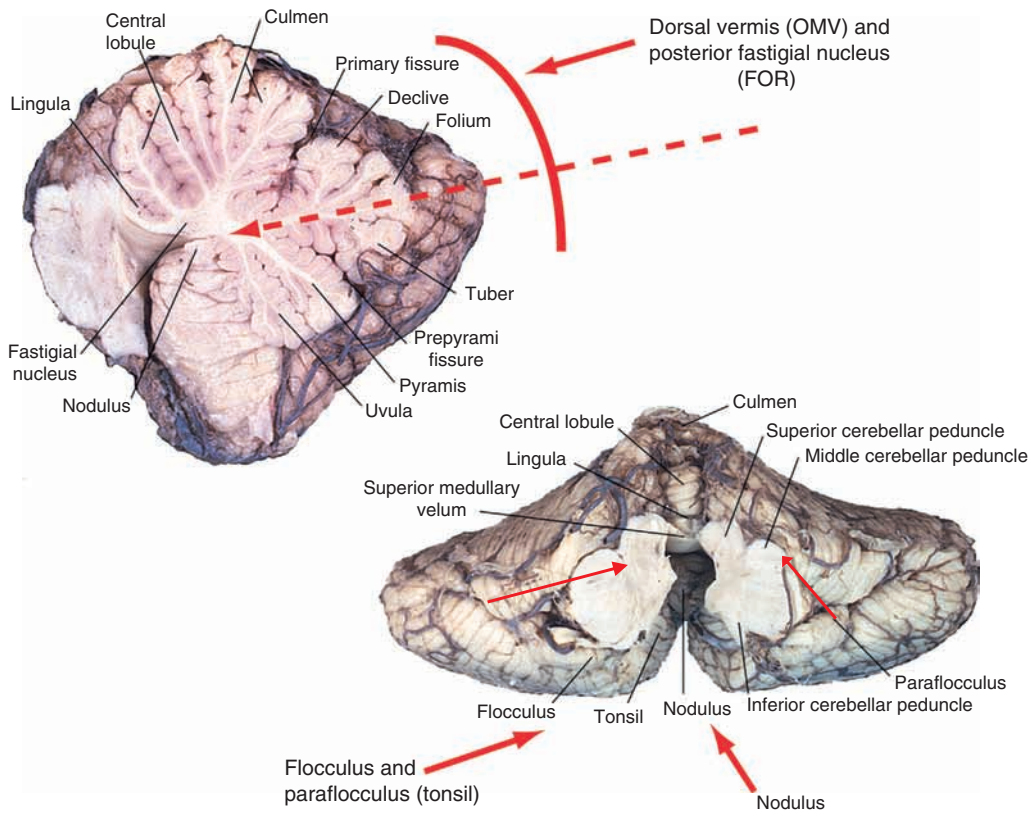

Figure 1 Cerebellar structures important for eye movement control. Modified from Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University.

corrective downward saccade (quick phase). This nystagmus likely arises from a loss of inhibition of Purkinje cells in the flocculus/paraflocculus upon pathways within the brain stem that mediate upward slow phases. This effect could be upon pathways the mediate the rotational VOR, the translational VOR, pursuit, or at some point where these signals are already combined, for example, in the brain stem circuits that mediate neural integration for gaze holding. An important feature of the upward drift is that it can have either a decreasing or an increasing slow-phase velocity waveform. These different velocity profiles suggest differences in the effect of the lesion on the vertical neural integrator. If the integrator is 'leaky' (its discharge decays with time), the velocity of the slow phase will be decreasing. If the integrator is 'unstable' (its discharge increases with time), the velocity of the slow phase will be increasing. Some monkeys with experimental floccular and parafloccular lesions show downbeat nystagmus with a decreasing velocity waveform, others an increasing velocity waveform.

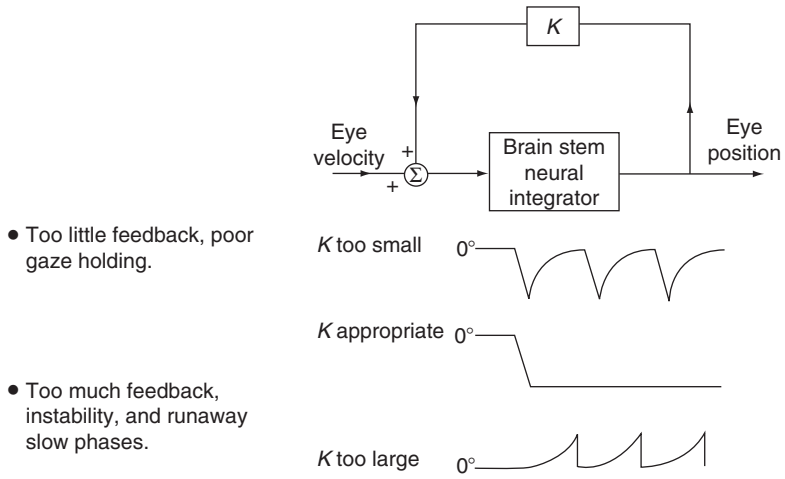

Figure 2 A simple hypothesis to explain how the cerebellar flocculus might influence gaze-holding. An inherently poor gaze-holding network resides in the brain stem. By virtue of positive feedback, the flocculus can modulate the ability of the brain stem neural integrator to generate a steady tonic eye position command. If feedback is inadequate, gaze holding is impaired. If feedback is excessive, gaze is unstable. Modified from Zee DS, Yamazaki A, Butler PH and Gucer G (1981) Effects of ablation of flocculus and paraflocculus of eye movements in primate. *Journal of Neurophysiology* 46: 878–899.

A simple hypothesis to explain the effects of lesions of the flocculus/paraflocculus on gaze holding is based upon the idea that there is an inherently poor, 'leaky' ocular motor integrator within the brain stem that sustains the discharge of neurons in the absence of an ongoing external input (Figure 2). The neurons in the paramedian tracts (PMTs) and in the medial vestibular nucleus and the nucleus prepositus hypoglossi are probably important constituents of this integrator network, and the superior vestibular nucleus and the interstitial nucleus of Cajal in the midbrain may also play a role in vertical neural integration. In this scheme, the cerebellum improves the performance of this brain stem neural integrator by acting upon a positive feedback loop around it. The flocculus/paraflocculus would then modulate the gain of this feedback loop. For the horizontal system, monkeys with floccular/paraflocculus lesions always develop a gaze-evoked nystagmus with a velocity-decaying waveform, indicating that the flocculus/paraflocculus is critical for improving the performance of the inherently 'leaky' horizontal brain stem neural integrator. The variable effect on the waveform of upward slow phases suggests that for the vertical integrator the flocculus/paraflocculus has a more subtle, modulatory role that is based upon an individual animal's inherent capabilities, which in turn depends upon its genetic makeup and its own ocular motor history (e.g., trauma or disease).

Postsaccadic drift Another feature of the floccular/paraflocculus syndrome is a brief drift of the eyes, lasting several hundreds of milliseconds, following each saccade. This postsaccadic drift reflects a mismatch between the pulse (phasic) and the step (tonic) of innervation for saccades and is due to abnormal amplitude of the step output of the neural integrator relative to its velocity input, the pulse. Normally, these two premotor signals are matched precisely, so the eye abruptly stops at the end of the saccade. In monkeys with lesions of the flocculus/paraflocculus, the direction of the postsaccadic drift, onward or backward, is variable from animal to animal, again in keeping with a modulatory role of the cerebellum upon the premotor brain stem circuits generating eye movements.

Vestibular–ocular reflexes The flocculus and paraflocculus are not critical for the VOR response to head rotations (r-VOR) in the sense that the r-VOR is still present in animals with lesions of the flocculus/paraflocculus or even after complete ablation of the cerebellum. The amplitude, however, may be normal, increased or decreased, and it may be misdirected, that is, the axes of eye rotation and head rotation become misaligned.

During rotation of the head around an earth-vertical axis, patients with diffuse cerebellar lesions may show a dynamic upward bias so that the eyes move up as well as horizontally, producing a 'cross-coupled' VOR. There are also inappropriate torsional components and the responses in the two eyes are disconjugate, with relatively more torsion in one eye and more vertical rotation in the other. A release of inhibition upon anterior semicircular circuits within the brain stem (which produce upward slow phases)

is a plausible explanation. These results implicate the cerebellum, and likely the flocculus/paraflocculus, in generating movements of each eye that have the correct amplitude and direction for perfect VOR compensation.

The cerebellum also plays a role in the translational VOR (t-VOR), and the flocculus/paraflocculus may be involved in its generation though this is not yet proven. The t-VOR can be measured alone (e.g., pure side-to-side, fore-and-aft, or up–down linear motion), or with a rotation of the head when the axis is in front of or behind the labyrinths, which leads to a combined rotation and translation. Patients with cerebellar lesions can have a profoundly impaired t-VOR even when the r-VOR is intact. They also have defects in adjusting the r-VOR for the distance of the target of interest and for the displacement of the orbits away from the axis of head rotation. These are necessary requirements for adequate gaze stabilization during head rotation, since virtually all head rotations are accompanied by some translation of the orbits. The flocculus/paraflocculus have neurons that discharge in relation to vergence and it is possible that this information is used to adjust the VOR for target distance, a necessary requirement whenever translation is a component of head motion. Other parts of the cerebellum including the nodulus and ventral uvula, the dorsal (oculomotor) vermis, and the underlying fastigial nuclei may also play a role in the t-VOR.

Vestibulocerebellum: The Nodulus/Ventral Uvula

The nodulus and adjacent ventral uvula are the most caudal aspects of the cerebellar vermis (Figure 1). The nodulus/ventral uvula acts upon the 'low-frequency' components of the VOR via its projections to a 'velocity-storage' mechanism within the vestibular nuclei. With a constant-velocity rotation of the head, the velocity-storage mechanism extends the duration of the VOR response beyond that expected from the mechanical properties of the cupula–endolymph system within the semicircular canals. This perseverating action slows the decay of nystagmus during a constant-velocity rotation in the dark and improves the performance of the VOR at low frequencies of rotation. During sustained 'off-vertical axis' rotation of the head, when there is an imposed change in linear acceleration due to the continuous reorientation of the head relative to the pull of gravity, the velocity-storage mechanism modulates the direction of compensatory slow phases, reorienting the axis of eye rotation relative to the axis of head rotation.

Lesions of the nodulus result in an increase in the duration of vestibular responses to a constant-velocity input (the VOR time constant is increased). There is also a loss of habituation to repetitive stimulation (the VOR time constant no longer decreases with repetitive stimulation). Tilt suppression of postrotary nystagmus, a phenomenon by which the decay of postrotary nystagmus is hastened by pitching the head down immediately following the end of a constant-velocity rotation, no longer occurs. With nodulus lesions, the VOR no longer shows the reorientation of the axis of eye rotation that normally occurs with an imposed linear acceleration. Patients with diffuse cerebellar disease may show abnormally directed slow phases, or 'cross-coupling' during low-frequency head rotation around an earth-vertical axis, during sustained optokinetic stimulation with the head upright, and after horizontal head shaking. These abnormalities of the direction of the VOR in response to low-frequency stimuli likely reflect dysfunction of the nodulus/ventral uvula. The nodulus/uvula also may play a role in other aspects of otolith–canal interaction and in enabling the brain to distinguish tilt from translation, since the otoliths signal linear acceleration regardless of its source, and in themselves cannot distinguish between a dynamic translation of the head and a static reorientation of the head with respect to gravity.

Periodic alternating nystagmus (PAN), a horizontal jerk nystagmus that changes direction every few minutes, may appear following lesions of the nodulus. PAN reflects the combined actions of a (1) disinhibited brain stem vestibular velocity-storage mechanism (due to loss of inhibition from Purkinje cells in the nodulus that project to vestibular nuclei) and (2) an intact adaptive mechanism that acts to null any sustained unidirectional nystagmus, thus allowing PAN to change direction. Because Purkinje cell inhibition is mediated through $GABA_B$ receptors, treatment with baclofen (a $GABA_B$ agonist) disengages the velocity-storage mechanism and stops PAN (gamma-aminobutyric acid, GABA).

The nodulus and uvula may have other ocular motor functions; lesions there also alter smooth pursuit and optokinetic nystagmus, but their exact contribution to these visual-following reflexes is unclear.

Functional Differences between the Flocculus/Paraflocculus and the Nodulus/Uvula

The vestibulocerebellum has many functions and lesions produce a variety of abnormalities. Are there overriding principles about the functions of these two areas of the cerebellum? One simple hypothesis is that the flocculus and paraflocculus are more concerned with relatively immediate and fast-acting, ocular motor functions that relate to the needs of the fovea, for example, smooth pursuit, the t-VOR and steady gaze following saccades, all of which relate to holding images steady on the fovea. The nodulus and ventral

uvula would be more concerned with generating eye movements that determine the duration and axis of eye rotation in response to low-frequency vestibular stimuli and so determine the orientation of images on the entire retina. This idea remains speculative, and even within the flocculus/paraflocculus and nodulus/uvula a further segregation of function is likely.

Dorsal Cerebellar Vermis and the Posterior Fastigial Nucleus

The dorsal vermis (lobules V–VII, called the oculomotor vermis (OMV)) and the underlying posterior fastigial nucleus (called the fastigial oculomotor region, FOR) are important in the control of saccades and pursuit (Figure 1).

Saccades Lesions in the OMV lead to changes in the accuracy, latency, and dynamic properties of saccades, as well the ability to undergo saccade adaptation. Physiological studies in monkeys reveal that Purkinje cells in the OMV discharge before saccades, and stimulation of this same area can elicit saccades. In humans, too, artificial stimulation using transcranial magnetic stimulation (TCM) and functional magnetic resonance imaging (fMRI) scanning, while subjects make saccades, confirms the participation of the OMV in the generation of saccades (Figure 3). Neurons in the FOR also discharge in relation to saccades. They supply a presaccadic burst for contraversive saccades (i.e., the right FOR bursts before leftward saccades) and a 'braking' discharge, late during the saccade, for ipsiversive saccades. Thus, each FOR acts to facilitate contraversive saccades and contributes to the termination of ipsiversive saccades. Purkinje cells in the OMV behave in a similar way to those of the FOR, though, as predicted from their inhibitory nature, their 'sign' is opposite. Thus, each side of the vermis acts to facilitate ipsiversive saccades and contributes to the termination of the contralateral saccades. Consequently, lesions in the FOR cause ipsiversive saccadic hypermetria (overshoot) and contraversive hypometria (undershoot). Bilateral FOR lesions cause bilateral hypermetria. Since Purkinje cells of the OMV inhibit the ipsilateral FOR, the effects of OMV lesions are opposite to those of FOR lesions: hypometric ipsiversive and hypermetric contraversive saccades. Bilateral lesions in OMV cause hypometric saccades in both horizontal directions. The role of the cerebellum in generation of vertical saccades is not as well localized, and other areas, such as the interposed nucleus and paraflocculus, may also be important.

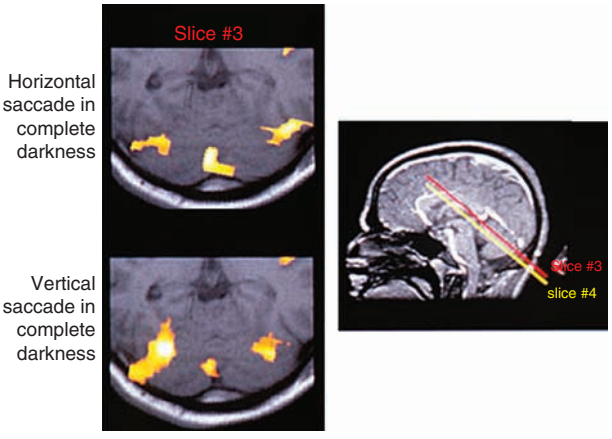

Figure 3 fMRI of the cerebellum during voluntary saccades made in darkness, showing relative activation in the regions of the dorsal vermis and also the lateral cerebellar hemispheres. Courtesy of Dr. Manuba Honda of Kyoto Japan, from Honda M, Zee DS, and Hallet M (1987) Cerebellar control of voluntary saccades in humans. *Society of Neuroscience Abstracts* 15: 1189.

How does the cerebellum influence the circuits within the brain stem that generate saccades? There are a number of possible targets by which the FOR could influence saccades. The early influence of the FOR on the initiation of saccades could be mediated through the excitatory burst neurons (EBN) within the brain stem reticular formation. The late influence of the FOR on stopping saccades could be through the inhibitory burst neurons (IBNs) in the brain stem reticular formation. These neurons could act to 'choke' or 'brake' the saccade by strongly inhibiting the abducens nucleus that had been driving the agonist muscles during the saccade. Projections of the FOR to omnipause neurons (OPNs), which inhibit premotor saccadic EBN – either directly or via projections to OPN from the fixation zone of the rostral pole of the superior colliculus – are other routes by which the FOR could help bring the saccade to an end. Modulation of the trajectories of saccades, 'on-line', has also been suggested as a function of the FOR, with the OMV receiving information about performance during saccades and adjusting its inhibition upon the FOR to assure saccade accuracy. In this scheme, the OMV would also recognize contexts (e.g., the position of the eye in the orbit) and adjust the output of the FOR to fit the particular circumstances in which the saccade must be generated. Though attractive, these speculations about cerebellar function in the control of saccades await experimental confirmation.

Pursuit Like saccades, there is considerable evidence that the OMV and FOR participate in the generation of pursuit eye movements. Lesions in the OMV lead to changes in the smooth pursuit as well the ability to undergo pursuit adaptation. Electrical

stimulation of the OMV can enhance contraversive or impair ipsiversive pursuit in monkeys, and TCM of the posterior cerebellum in humans can also evoke pursuit eye movements in the same pattern. There are neurons in both the OMV and the FOR that discharge in relation to pursuit. In the case of the FOR, neurons discharge early during contraversive pursuit and late for ipsilateral pursuit, analogous to activity associated with saccades. Thus, each FOR can facilitate contraversive pursuit and can contribute to the termination of ipsiversive pursuit. Purkinje cells in the OMV probably behave in a similar way to those of the FOR, though, as predicted from their inhibitory nature, their 'sign' is opposite. Each side of the vermis would act to facilitate ipsiversive pursuit and contribute to the termination of contralateral pursuit.

The pursuit deficits that have been reported after lesions in the OMV and the FOR are largely in accord with the physiological findings. As expected, with a lesion in the FOR, contralateral pursuit is impaired. Bilateral FOR lesions would be expected to produce reasonably symmetrical and relatively intact pursuit and this appears to be the case, though a detailed analysis of the open-loop behavior (i.e., the first 100 ms during which pursuit is initiated) of such animals is not yet available. On the other hand, bilateral lesions of the OMV produce bilateral pursuit deficits.

One other remarkable finding after OMV lesions in monkeys is that there are quantitative similarities in the effects of the lesions on saccade amplitude, and on pursuit velocity at the end of the 'open-loop' initial period of pursuit tracking (the first 100 ms of tracking after a target has started moving or has changed its speed). This similarity suggests that the OMV functions in the control of the preprogrammed, 'open-loop' component of eye movements, and furthermore that saccades and the 'open-loop' portion of pursuit share neural circuitry, not only in the cerebellum but also in the brain stem and superior colliculus. Compatible with this idea is the finding of neurons within the OMV that discharge for both pursuit and saccades, and that stimulation of the OMV can alter ongoing saccades or pursuit, depending upon the parameters of the stimulus.

Before leaving the FOR and OMV, one must ask why there are multiple areas within the cerebellum that seem related to pursuit and saccades. For example, the uvula and the interposed nucleus, the flocculus/parafloccus, and the FOR/OMV complexes, all have neurons that discharge with pursuit eye movements. Lesions in the uvula and within the lateral cerebellar hemispheres also have been associated with pursuit deficits. One dichotomy to consider is that the OMV/FOR is more concerned with the initiation and termination of the preprogrammed, open-loop, initial portion of pursuit, and the flocculus/paraflocculus is more concerned with pursuit during sustained tracking. Such a strict segregation of function, however, is not yet proven. Furthermore, after partial lesions in the cerebellar cortex, there is a considerable recovery of function, suggesting a capability for one part of the cerebellum to take over for another though other parts of the brain, too, may be involved in this recovery.

For saccades, too, there are neurons in areas of the cerebellum apart from the FOR/OMV that show saccade-related activity. They include the posterior portion of the interposed nucleus, the basal interstitial nucleus, and the lateral cerebellar hemispheres. Both pursuit and saccades are commonly associated with other motor behaviors, head movements, arm movements, and walking or running, for example, and also occur in different contexts (e.g., reflexive or self-generated). Thus, one can envision a number of areas within the cerebellum associated with different motor behaviors that might want to 'know' or use information about associated eye movements to help optimize their own performance.

Other Abnormalities of Ocular Motor Control Associated with Cerebellar Dysfunction

Abnormalities of eye torsion The torsional orientation of the eye during fixation is precisely controlled and is described by Donders' and Listing's laws. Donders' law states that during steady fixation, the torsional orientation of the eyes is fixed for a given eccentric (horizontal and vertical) eye position. Listing's law describes the way in which Donders' law is implemented; it defines the kinematic features of eye rotations that fix the torsional orientation of the globe for a given horizontal and vertical position of the eyes in the orbit. Listing's law is not perfectly obeyed in patients with cerebellar lesions. Which part of the cerebellum serves this function is not yet known.

The Cerebellum and Binocular Ocular Motor Control

Most of the findings relating the cerebellum to ocular motor control have dealt with conjugate eye movements. Less is known about the role of the cerebellum in disconjugate ocular motor control when the eyes must rotate by different amounts, either to alter the relative positions of the eyes during fixation (static alignment), or to readjust alignment during movement (dynamic alignment or yoking of the eyes).

Disorders of alignment with cerebellar lesions Patients with cerebellar damage sometimes show a

skew deviation, a vertical misalignment of the eyes that cannot be attributed to oculomotor or trochlear nerve palsy. The misalignment often changes its sense with horizontal eye position; most commonly, the abducting eye is higher. Patients with cerebellar abnormalities also may have an esotropia (a convergent squint in which the eyes turn inward), sometimes attributed to a divergence paralysis since the esodeviation is usually greater at distance. Patients with caudal cerebellar lesions may also have a divergence-beating nystagmus (convergent slow phases with divergent quick phases). Experimentally, lesions of the FOR produce convergence failure while lesions in the OMV produce an inward turning of the eyes. These types of abnormalities hint at an excess of convergence tone with some cerebellar lesions.

The VOR must be able to produce disconjugate compensatory responses, especially those associated with translation of the head. Patients with diffuse cerebellar atrophy produce disconjugate vestibular responses and also show disconjugate saccades in which the movements of the two eyes during and immediately after saccades are not yoked. Taken together with anatomical and physiological studies, there is compelling evidence for a role for the cerebellum in the disconjugate control of eye movements, though much needs to be learned about these particular aspects of cerebellar function.

Ocular Motor Learning and the Cerebellum

Perhaps the most intriguing aspect of cerebellar function is its role in motor learning. The flocculus/paraflocculus has a clear role in ocular motor learning including the amplitude (gain) and the direction (relative to head motion) of the r-VOR, the pulse-step ratio for saccades, and the gain of smooth pursuit. The OMV, too, has a role in ocular motor learning, especially in modification of the initial acceleration of pursuit and of the amplitude of saccades. As expected, the FOR also plays a role in saccade (and presumably pursuit) adaptive control. Disconjugate ocular motor control is also subject to motor learning and under cerebellar influence.

The effect of experimental lesions within the cerebellum on ocular motor function often varies from animal to animal. While differences in the extent of lesions may explain some of this variability, another interpretation relates to the central role of the cerebellum in long-term ocular motor learning. In this formulation, the cerebellum acts to compensate for differences in motor performance that are idiosyncratic from animal to animal, depending upon the particular genetic makeup of each animal, its 'life history' of disease and trauma, and the interaction between inherent capabilities and environmental influences during development and aging.

The exact mechanisms of ocular motor learning are unclear, and their elucidation is a major focus of contemporary neuroscience. Just to point out a few of the more topical issues: What are the error signals that trigger motor learning? What are the relative roles of the cerebellar cortex, and of the deep cerebellar nuclei (or in the case of adaptation of the VOR, the vestibular nuclei) in acquisition versus maintenance of motor learning? What is the role of the cerebellum in short (minutes to hours) versus long-term (days to weeks) motor learning? What are the differences in mechanisms between relatively simple motor learning (e.g., adjustment of the amplitude of the VOR for all head movements) and relatively complex, context-specific motor learning (e.g., adjustment of the amplitude of the horizontal VOR as a function of eye position or viewing distance)? What is the role of the cerebellum in developing the internal models of the effectors of movements themselves (e.g., the oculomotor muscles and orbital structures, the 'plant') that are required for the brain to calibrate motor responses in the long term? Perhaps the most challenging questions relate to the molecular bases for learning within the cerebellum, be it in alterations of synaptic transmission, changes in connectivity with neural circuits, or changes in the properties of neuronal membranes themselves. And finally, can learning be promoted with medications or therapy?

See also: Basal Ganglia and Oculomotor Control; Brainstem Control of Eye Movements; Cortical Control of Eye Movements; Executive Functions: Eye Movements and Neuropsychiatric Disorders; Eye and Head Movements; Eye Movement Disorders; Oculomotor Control: Anatomical Pathways; Optokinetic Eye Movements; Thalamus and Oculomotor Control; Vergence Eye Movements; Vestibulo-Ocular Reflex.

Further Reading

Blazquez PM, Hirata Y, and Highstein SM (2004) The vestibulo-ocular reflex as a model system for motor learning, what is the role of the cerebellum? *Cerebellum* 3: 188–192.

Catz N, Dicke PW, and Thier P (2005) Cerebellar complex spike firing is suitable to induce as well as to stabilize motor learning. *Current Biology* 15: 2179–2189.

De Zeeuw CI and Yeo CH (2005) Time and tide in cerebellar memory formation. *Current Opinion in Neurobiology* 15: 667–674.

Faulstich M, van Alphen AM, Luo C, Du LS, and De Zeeuw CI (2006) Oculomotor plasticity during vestibular compensation does not depend on cerebellar LTD. *Journal of Neurophysiology* 96: 1187–1195.

Highstein SM and Thach WT (2002) The Cerebellum: Recent developments in cerebellar research. *Annals of the New York Academy of Sciences* 978.

Honda M, Zee DS, and Hallet M (1987) Cerebellar control of voluntary saccades in humans. *Society of Neuroscience Abstracts* 15: 1189.

Ito M (2005) Bases and implications of learning in the cerebellum – adaptive control and internal model mechanism. *Progress in Brain Research* 148: 95–109.

Kassardjian CD, Tan YF, Chung JY, Heskin R, Peterson MJ, and Broussard DM (2005) The site of a motor memory shifts with consolidation. *Journal of Neuroscience* 25: 7979–7985.

Krauzlis RJ and Miles FA (1998) Role of the oculomotor vermis in generating pursuit and saccades: Effects of microstimulation. *Journal of Neurophysiology* 80: 2046–2062.

Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*. New York: Oxford University.

Manzoni D (2005) The cerebellum may implement the appropriate coupling of sensory inputs and motor responses, evidence from vestibular physiology. *Cerebellum* 4: 178–188.

Optican LM (2005) Sensorimotor transformation for visually guided saccades. *Annals of the New York Academy of Sciences* 1039: 132–148.

Rambold H, Churchland A, Selig Y, Jasmin L, and Lisberger SG (2002) Partial ablations of the flocculus and ventral paraflocculus in monkeys cause linked deficits in smooth pursuit eye movements and adaptive modification of the VOR. *Journal of Neurophysiology* 87: 912–924.

Robinson FR, Fuchs AF, and Noto CT (2002) Cerebellar influences on saccade plasticity. *Annals of the New York Academy of Sciences* 956: 155–163.

Shutoh F, Ohki M, Kitazawa H, Itohara S, and Nagao S (2006) Memory trace of motor learning shifts transsynaptically from cerebellar cortex to nuclei for consolidation. *Neuroscience* 139: 767–777.

Takagi M, Tamargo R, and Zee DS (2003) Effects of lesions of the cerebellar oculomotor vermis on eye movements in primate: binocular control. *Progress in Brain Research* 142: 19–33.

Takagi M, Zee DS, and Tamargo RJ (2000) Effects of lesions of the oculomotor cerebellar vermis on eye movements in primate: Smooth pursuit. *Journal of Neurophysiology* 83: 2047–2062.

Takeichi N, Kaneko CR, and Fuchs AF (2005) Discharge of monkey nucleus reticularis tegmenti pontis neurons changes during saccade adaptation. *Journal of Neurophysiology* 94: 1938–1951.

Thier P, Dicke PW, Haas R, Thielert CD, and Catz N (2002) The role of the oculomotor vermis in the control of saccadic eye movements. *Annals of the New York Academy of Sciences* 978: 50–62.

Voogd J and Barmack NH (2005) Oculomotor cerebellum. *Progress in Brain Research* 151: 231–268.

Walker MF and Zee DS (2005) Cerebellar disease alters the axis of the high-acceleration vestibuloocular reflex. *Journal of Neurophysiology* 94: 3417–3429.

Zee DS and Walker MF (2003) Cerebellar control of eye movements. In: Chalupa L and Werner J (eds.) *The Visual Neurosciences*, pp. 1485–1498. Cambridge, MA: MIT Press.

Zee DS, Yamazaki A, Butler PH, and Gucer G (1981) Effects of ablation of flocculus and paraflocculus of eye movements in primate. *Journal of Neurophysiology* 46: 878–899.

Cerebellum: Clinical Pathology

M Molinari, Santa Lucia Foundation, Rome, Italy
M G Leggio, University of Rome 'La Sapienza,' Rome, Italy

Published by Elsevier Ltd.

Introduction

When clinicians examine a patient for signs of cerebellar disease, they follow a procedure that developed from the observations of Luigi Luciani (1840–1919), Joseph Babinski (1857–1932), Andre Thomas (1867–1961), and Gordon Holmes (1876–1965). The notion of the pivotal function of the cerebellum in motor coordination and fine-tuning of movements originated from their pioneering work. The motor aspects of cerebellar functioning dominated cerebellar studies almost completely until the 1980s, when the importance of nonmotor cerebellar functions was recognized. As the spectrum of cerebellar functions enlarged, classical theories of cerebellar functioning were questioned so thoroughly that now there is little consensus over the basic rules controlling cerebellar physiology and the influence of this cerebral structure on other circuits in the central nervous system (CNS). To complicate matters even more, cerebellar symptoms, although similar independently of the etiology, may differ in their pattern of onset and recovery. Differences have been evidenced between focal acute and chronic degenerative pathologies as well as between developmental and adult lesions. Namely, acute vascular lesions have an abrupt, even dramatic, onset often followed over weeks or months by a progressive lessening of symptoms and even to quite complete clinical recovery. Degenerative pathologies behave in the opposite way, with rather slow onset and progressive development into a disabled state. Cerebellitis and cerebellar localization of more diffuse CNS pathologies, such as multiple sclerosis, tend to present acute onset and little if any recovery, even after stabilization of the primary pathology. During development, a very different clinical picture emerges from lack of development of the cerebellum or from acquired damage. Given this complex framework, our comprehension of the pathophysiology of cerebellar motor symptoms is still poor. However, cerebellar signs and symptoms are remarkably similar no matter what the disease process and, at least for the motor domain, are well defined and substantially the same as those listed by Holmes in 1939. In the present article, motor symptoms and related clinical signs, together with cognitive and behavioral impairments related to cerebellar diseases, are presented.

Common symptoms in the case of acute onset include headache, vertigo, nausea, and vomiting. However, these symptoms can depend on an increase in intracranial pressure resulting from expanding mass lesions or on secondary hydrocephalus due to a block of cerebrospinal fluid (CSF) flux in the aqueduct. Therefore, they cannot be considered specific to damage of the cerebellum *per se*. Since signs and symptoms related to cerebellar damage are remarkably the same no matter what the disease process, in this article they are not related to differences in pathological processes affecting the cerebellum. Different clinical scales have been developed to provide reliable and valid measures of the severity of cerebellar symptoms. Diseases specifically involving the cerebellum are listed in Table 1.

Motor Symptoms

In the motor domain, cerebellar symptoms are ipsilateral to the side of the cerebellar lesion and can be grouped as follows: (1) modified muscle tone, (2) impaired motor coordination, (3) tremor, (4) oculomotor disturbances, (5) stance and gait disturbances, and (6) speech disorders.

Muscle Tone

Some researchers believe that a reduction in muscle tone is linked to direct cerebellar influence on the spinal gamma loop; however, this interpretation has been questioned. More recently, the absence of facilitatory cerebellar influence over the contralateral motor cortex has been cited to explain cerebellar hypotonia. Although hypotonia is often intense in children, in adults it is usually associated with severe cerebellar damage or occurs in the acute stage of the disease. The following clinical tests are used diagnostically.

Gordon Holmes sign Patients place their elbows on a table and keep their forearms and hands extended upward. If hypotonia is present the hands slowly close, drooping down toward the arms, because the patients are unable to hold the antigravity posture.

Pendulousness When an extremity is displaced passively, so that it swings freely, there is a regular, precise pendulous motion that diminishes in a steady, even manner. Upper extremities are tested by moving the standing patient's shoulders back and forth. Lower extremities are tested with the patient sitting on the edge of a bed with their legs hanging freely. The examiner briskly flips one leg and lets it swing.

Table 1 The most common examples of cerebellar pathology

Pathology	*Description*
Developmental	Complete or partial absence of the cerebellum (agenesis), Dandy–Walker syndrome
Degenerative	Genetic disorders including spinocerebellar atrophies, Friedreich ataxia, ataxia telangiectasia, Refsum syndrome, amaurotic familial idiocy, multiple sclerosis, Wilson's disease, tuberous sclerosis, carcinomatous diffuse corticocerebellar degeneration
Infectious	Bacterial (abscesses, tuberculosis, syphilis, whooping cough), fungal infections, parasitic infestations (cysticercosis, malaria), viral infections (mumps, viral encephalitis, poliomyelitis, kuru, Creutzfeldt–Jakob disease, looping ill)
Toxic and metabolic	Ethanol, dilantin, carbon tetrachloride, anoxia, hypoglycemia, vitamin deficiencies (B_1, B_{12}), hypothyroidism, heat stroke
Vascular diseases	Thrombosis, embolism of superior cerebellar, anterior inferior cerebellar, or posterior inferior cerebellar branches of the basilar or vertebral arteries, transient ischemic episodes in these arterial territories, hemorrhage
Trauma	Restricted to the cerebellum; usually results from penetrating wounds and typically involves vessels and parenchyma
Tumors	Metastatic from primary sites elsewhere in the body; astrocytoma (slow growing, cystic, common in children); medulla blastoma (invasive, rapidly growing, most commonly occurs in children); ependymomas, vascular malformations, and hemangioblastoma; acoustic neuroma

Modified from Dow RS (1987) Cerebellum, pathology: Symptoms and signs. In: Adelman G (ed.) *Encyclopedia of Neuroscience*, pp. 203–206. Boston: Birkhäuser.

In patients with cerebellar disease, pendulousness is increased.

Postural asymmetries In unilateral lesions, side differences in muscle tone can induce postural asymmetries.

Pendular tendon reflexes Especially for patellar and triceps reflexes, reflex responses are characterized by limb oscillation around the position at rest.

Impairment of Motor Coordination

Impaired synchrony of agonist and antagonist muscle interplay is considered to be the neurophysiological basis of the motor coordination deficits observed in cerebellar pathologies classically grouped under the term 'ataxia.' Ataxia, from the Greek ατάξια (i.e., lack of order), refers to a lack of motor coordination and can be evidenced by several clinical tests that explore limb functions.

Heel–knee–shin/toe square-drawing tests The patient, lying supine on a bed or sitting on a chair, flexes one leg at the hip, knee, and ankle so that the heel traces a path lightly along the top of the opposite shin and then extends to all joints to retrace the line. In normal persons the path is straight and the motion continuous, with smooth and accurate termination and reversal at the endpoint. In patients with ipsilateral damage of the cerebellum the moving limb commonly overshoots the endpoint (knee, ankle) and the path oscillates from side to side. The oscillation is rough, as if several rhythms were intermixed; the turnarounds look distinctly like 'bounces.' Although lateral deviations are the most conspicuous variations, to-and-fro variations also occur; up-and-down variations are damped by contact with the shin. In drawing a square, the patient is asked to follow the examiner's finger and/or to trace an imaginary square in the air with the big toe. Here again, in cerebellar disease one sees overshooting of the endpoint, with side-to-side oscillation that characteristically increases as the target is approached.

Finger–nose–finger test The patient reaches out and uses his extended forefinger to touch the examiner's forefinger and then touches his own nose, repeating the process. The abnormalities to be looked for are (1) tremor at 3–6 Hz, which is primarily side to side and up and down, (2) overshooting the target position, which may then be corrected by a movement, and (3) in chronic cases, a tendency to 'decompose' the movement (moving first one joint then another seriatim).

Rebound The examiner holds the cerebellar patient's wrists and asks her to pull against the resistance. Without warning the examiner suddenly releases his hold. The patient is unable to cease agonist and initiate antagonist muscle actions rapidly enough, and the patient's arms fly toward her and may even hit her in the face. In a similar test, the cerebellar patient is asked to hold outstretched a leg or an arm. The examiner then briskly taps a distal spot on the limb from a variety of different directions. The normal response is to deviate minimally from the initial position and return to it promptly, using a rapid increase first in agonist and then in antagonist muscle activity, to return to and hold the position without overshooting

or oscillation. In cerebellar patients, a greater displacement occurs; the return overshoots the initial position, followed by one or more beats of oscillation.

Finger-to-finger and finger-to-thumb tests The patient is asked to tap the pad of his thumb on the pads of each of the digits from 2 to 5 and then return from 5 to 2 in rapid succession. Cerebellar disease is characterized by slowing down, a lack of fixed rhythm and imprecise placing of the thumb on the fingertips. Actually, it is rare that these signs are greatly affected by cerebellar disease, possibly due to the surviving ability of the motor cortex to control single digits normally (the thumb may be held fixed while the fingers move at the metacarpophalangeal (MCP) joint only). A more sensitive test was devised by C M Fisher in which the nail of the index finger is tapped rapidly and repeatedly on the skin-crease at the top joint of the thumb. What is observed in cerebellar disease is slowing, irregularity in the amplitude of the excursion and the rhythm of subsequent excursions, and great variation in the position of the index 'hits' to either side of the thumb-crease target. The first test uses the stereotyped mechanisms of single-digit movement built into the motor cortex. The endpoint is primarily controlled by digit length. This test is more specific for lesions of the motor cortex than for lesions of the cerebellum. The second test requires coordinated action across proximal interphalangeal (PIP) and MCP joints (and lumbricals and long flexors) to an arbitrary endpoint and tests the integrity of the dentate nucleus in the lateral cerebellum.

Rapid, alternating movements The capacity to develop alternating movements is tested by pronation–supination movements. The patient is asked to keep his forearms vertical and to perform pronation–supination movements with his hands. This function can also be tested by asking the patient to slap the palm and the back of one hand alternately on the palm of the other hand or on the thigh. Irregular rhythm and amplitude of excursion with slowing and frequent repeated 'wrong moves' are looked for, especially since the task performance is prolonged.

Tremor

Although the pathophysiology of cerebellar tremor has been addressed by numerous experimental and clinical studies, it is still a matter of debate. In general, it is thought to be generated within central motor loops.

Intention or kinetic tremor This type of tremor is tested using the finger–nose and heel–knee test. It may be present throughout the entire movement or only at the beginning, but it is particularly evident at the end of the movement. Intention tremor tends to worsen in the days or weeks after acute cerebellar lesion.

Postural tremor This type of tremor, defined also as 'action tremor,' appears during performance of postural tasks that require precision. It is tested by asking the patient either to keep her arm outstretched or, in a supine position, her legs held fixed against gravity.

Palatal tremor This type of tremor is characterized by short, rhythmic, involuntary movements of the soft palate often associated with hypertrophy of the contralateral inferior olive.

Oculomotor Disturbances

In cerebellar pathologies, oculomotor deficits are usually related to involvement of the dorsal vermis or the fastigial nucleus, the folliculus and paraflocculus, and/or the uvula and the nodulus in the pathological process. Disorders of fixation, pursuit, and saccades, together with the development of nystagmus and skew deviation, are often observed. During fixation, brief oscillations (flutter), usually horizontally conjugated and present in both the light and the dark, can be observed. Nonsmooth pursuit, characterized by a stepwise decomposition of movement, is frequently observed in cerebellar diseases, but with low specificity. Also, in saccadic movements, the characteristic overshooting observed in limb movements (i.e., ocular hypermetria) can be present.

Equilibrium and Stance

The patient is in a standing position with his eyes first open and then shut. This test was originally devised by Moritz Romberg (1795–1873) to distinguish diseases of peripheral nerves or spinal afferent somatosensory pathways from diseases of the cerebellum. Patients with somatosensory disturbances can stand well with their eyes open but may fall when their eyes are closed. Patients with cerebellar disease may be unstable with their eyes open as well as closed; unlike normal persons, they cannot use somesthesis, vestibular information, or vision for equilibrium and stance. With acute unilateral cerebellar damage, the patient falls toward the side of the lesion. Equilibrium has been clinically tested in patients with cerebellar lesions by measuring postural sway on a force-sensing platform. Lesions of the flocculo-nodular lobe and posterior vermis (which projects heavily to the lateral vestibular nucleus) cause an irregular course sway (back and forth and side to side), as does damage to the membranous labyrinths, which is indistinguishable

from 'vestibular ataxia.' Damage to the anterior lobe cortex (alcoholic cerebellar degeneration; late-life atrophy) causes back-and-forth oscillation of the trunk.

Gait

Gait is tested by observing the patient walking in a straight line, in tandem, and backwards. Gait is impaired if ataxia is present. The latter produces irregular steps, often with a broad base, staggering with clumsy corrections, altered rhythm, and irregular, often reduced, speed. Locomotion increases the tendency to fall to the side of the cerebellar lesion, especially in tandem walking. When mild unilateral damage is present, patients start to fall to the side of the lesion, sense the impending fall (vestibular sense is intact), and voluntarily correct themselves with a wider step to that side. The spinal cord contains many of the mechanisms necessary for stance and locomotion; spinal cats can make rudimentary walking movements. Results of animal studies show that stance and gait are specifically controlled by vestibular and fastigial nuclei: cells fire during walking; lesions cause falls to that side. Neurophysiological data in patients suggest that the cerebellum intervenes to adjust movements related to many joints, especially to accommodate external constraints.

Speech

Cerebellar damage affects both motor and cognitive aspects of speech. Dysarthria, cerebellar mutism, agrammatism, and reduced verbal fluency are symptoms of impaired speech reported after cerebellar damage.

Dysarthria Patients with cerebellar disease can have what is called a scanning dysarthria. This impairment is characterized by difficulty in articulating sounds into words. The problem is not at the language level – that is, in choosing and saying words correctly or in making sentences and paragraphs out of them. The patient is asked to use lips (mi-mi-mi), tongue (la-la-la), pharynx (go-go-go), and larynx (kay-kay-kay) to make sounds involving relatively smaller groups of muscles, together with those moving the thorax and diaphragm. These utterances may sound normal or may be unevenly spaced across time. Saying a sample sentence is relatively more impaired. The utterance is monotonous, lacking normally modulated inflexion of volume, and is slow and slightly slurred. The impaired patient sounds illiterate, painfully and slowly reading from a text one word at a time (hence, scanning). The topography of cerebellar control over motor aspects of speech is not clear. Holmes found this symptom mostly with midline lesions. Working in the 1970s, Richard Lechtenberg and Sid Gilman also found local lesions to be very close to the midline, with a somewhat larger percentage in the left intermediate than in the right zone. This was somewhat surprising since language is located predominantly in the left cerebral hemisphere and thus one would expect its articulation to be located in the known crossed projection zone in the right cerebellar hemisphere. There are also recent reports of patients with lesions limited to the right cerebellar hemisphere.

Cerebellar mutism The term 'mutism' refers to the inability of an alert patient to produce verbal output. The cerebellar form of mutism has been attributed to lesions of the cerebellar vermis or the cerebellar hemispheres and refers to a complete but transient loss of speech followed by dysarthria, usually after surgery for posterior fossa tumors, cerebellar hemorrhage, or trauma, or in infectious or postinfectious cerebellar syndromes. Cerebellar mutism has often been reported in children and adolescents, but rarely in adults; although it is believed to be due to lost coordination of orofacial movements, the issue is still being debated.

Agrammatism Agrammatic deficits have been reported in patients with right cerebellar lesions, and the role of the cerebellum in grammar has been recently reviewed. In general, all authors report agrammatism in spontaneous speech with or without involvement of sentence comprehension. In agreement with current theories on agrammatism, it was previously found that cerebellar processing did not affect syntactic knowledge but rather affected sentence construction.

Verbal fluency Verbal fluency is the ability to generate lists of words according to a given rule. The latter can be either a letter of the alphabet (e.g., retrieval of words that begin with the letter F) or a semantic category (e.g., retrieval of words from the semantic category of 'animals'). There are consistent reports that verbal fluency is reduced after cerebellar damage; also, functional neuroimaging data demonstrate cerebellar activation during verbal fluency tasks. In particular, verbal fluency is impaired in a modality-specific manner, with reduced phonemic rule performance and spared semantic rule performance.

Cognition and Emotion

As early as the nineteenth century there were speculations that the cerebellum might play a role in higher mental processes. However, the notion that the cerebellum might have a nonmotor function only became influential about 50 years ago. Subsequently, the advent of functional neuroimaging techniques, demonstrating

cerebellar activation in a variety of cognitive, nonmotor tasks, and the discovery of reciprocal connections between the lateral cerebral hemispheres and the contralateral frontal and parietal cerebral association cortex provided strong support for clinical evidence of cognitive defects in cerebellar pathologies. Today, the number of studies addressing the cerebellar role in emotion and cognition has increased enormously and clinical data and neuroimaging evidence converge in supporting cerebellar influence on quite disparate functions, such as attention, cognitive flexibility, linguistic processing, mental imaging, sensory discrimination, visuospatial processing, timing, and verbal memory, as well as working memory.

Cerebellar Role in Motor Learning

Animal and human studies have shown that the cerebellum is involved in and needed for a variety of different motor tasks, including modification of vestibuloocular reflex gain, adaptation of saccade endpoint to target, associative conditioning of a blink reflex, and gaze–arm calibration in prism-adapted reaching and throwing. Neurophysiological and functional neuroimaging observations in humans indicate the importance of the cerebellum in implicit/procedural learning, and lesion studies indicate procedural learning impairment in cerebellar-lesioned subjects. Although these observation have not yet proved to be of any clinical relevance, knowledge of the characteristics of motor and procedural learning impairment is of clinical relevance, especially in rehabilitation settings.

Localization: Body Maps and Motor Modes

Holmes found little evidence for localization of functions in the cerebellum. Although clinical symptoms are not very useful for a lobular localization of cerebellar damage, there are general indications that medial lesions affect posture and gait while more lateral lesions affect voluntary limb movements. Animal studies have conclusively shown that the flocculus and the nodulus (vestibulocerebellum) control antigravity stance and eye and head movements in response to vestibular information; the vermis and fastigial nucleus control antigravity stance and saccadic eye movements; the intermediate zone cerebellar cortex and interposed nuclei control agonist–antagonist muscle coordination; and the lateral hemisphere cortex and dentate nuclei control combined activities of synergistic muscles in multijointed reaching toward and grasping of objects in space. Further, animal studies strongly support a scheme of localization of functions and of body parts. In macaques, a somatotopic map of muscle actions is seen within each of the fastigial, interposed, and dentate nuclei, with anterior lower extremity, posterior head, medial distal, and lateral proximal parts; the upper extremity is represented in the midportion. The advent of functional neuroimaging and the availability of a detailed atlas of human cerebellar folia have made it possible to collect localization data. Consequently, there is growing evidence of different maps in the cerebellum.

Sensorimotor Maps

Motor and kinesthetic tasks activate the rostral and caudal spinocerebellum with two distinct homuncular representations. One is located upside down in the superior cerebellum and the other, smaller one is in the inferior cerebellum. The somatotopic map closely resembles that observed in macaques. Correlations between clinical findings and magnetic resonance imaging (MRI)-based cerebellar lesions provide support for a localization data diagnosis based on clinical evidence.

Localization of Cognitive Functions

Activation studies consistently demonstrate cerebellar activation in different cognitive functions and provide support for the development of topographic maps of the cognitive cerebellum, which evidence the following findings: explicit memory retrieval in vermal lobules IV–VIIA and in the left hemisphere, mainly in Crus I; sequence learning bilaterally in Crus I and in vermal lobule VI; language and working memory functions in vermal lobules IV–VIIA and Crus I of the right hemisphere; and verbal working memory in HVIIB of the right hemisphere. This functional topography has not yet been confirmed by clinically relevant data coupling cerebellar damage with cognitive symptoms.

See also: Cerebellar Deep Nuclei; Cerebellum and Oculomotor Control; Cerebellum: Evolution and Comparative Anatomy; Cerebellum: Models; Cognition: Cerebellum Role; Friedreich Ataxia; Olivocerebellar System; Sporadic Degenerative Ataxias and the Dominantly Inherited Spinocerebellar Ataxias.

Further Reading

Akshoomoff NA, Courchesne E, and Townsend J (1997) Attention coordination and anticipatory control. *International Review of Neurobiology* 41: 575–598.

Bower JM and Parsons LM (2003) Rethinking the 'lesser brain'. *Scientific American* 289(2): 50–57.

Desmond JE and Fiez JA (1998) Neuroimaging studies of the cerebellum: Language, learning and memory. *Trends in Cognitive Sciences* 2(9): 355–362.

Dow RS (1987) Cerebellum, pathology: Symptoms and signs. In: Adelman G (ed.) *Encyclopedia of Neuroscience*, pp. 203–206. Boston: Birkhäuser.

Dow RS and Moruzzi G (1958) *The Physiology and Pathology of the Cerebellum*. Minneapolis, MN: University of Minnesota Press.

Grodd W, Hulsmann E, and Ackermann H (2005) Functional MRI localizing in the cerebellum. *Neurosurgery Clinics of North America* 16(1): 77–99.

Holmes G (1939) The cerebellum of man. *Brain* 62: 1–30.

Ito M (2005) Bases and implications of learning in the cerebellum – adaptive control and internal model mechanism. *Progress in Brain Research* 148: 95–109.

Justus T (2004) The cerebellum and English grammatical morphology: Evidence from production, comprehension, and grammaticality judgments. *Journal of Cognitive Neuroscience* 16(7): 1115–1130.

Leggio MG, Silveri MC, Petrosini L, et al. (2000) Phonological grouping is specifically affected in cerebellar patients: A verbal fluency study. *Journal Neurology, Neurosurgery & Psychiatry* 69: 102–106.

Leiner HC, Leiner AL, and Dow RS (1986) Does the cerebellum contribute to mental skills? *Behavioral Neuroscience* 100(4): 443–454.

Manto MU and Pandolfo M (2002) *The Cerebellum and Its Disorder*. Cambridge, UK: Cambridge University Press.

Middleton FA and Strick PL (1997) Dentate output channels: Motor and cognitive components. *Progress in Brain Research* 114: 553–566.

Molinari M, Filippini V, and Leggio MG (2002) Neuronal plasticity of interrelated cerebellar and cortical networks. *Neuroscience* 111(4): 863–870.

Molinari M, Leggio MG, Solida A, et al. (1997) Cerebellum and procedural learning: Evidence from focal cerebellar lesions. *Brain* 120(10): 1753–1762.

Ramnani N (2006) The primate cortico-cerebellar system: Anatomy and function. *Nature Reviews Neuroscience* 7(7): 511–522.

Restuccia D, Marca GD, Valeriani M, et al. (2007) Cerebellar damage impairs detection of somatosensory input changes. A somatosensory mismatch-negativity study. *Brain* 130: 276–287.

Schmahmann JD (1997) The cerebellum and cognition. *International Review of Neurobiology* 41: 31–60.

Schmahmann JD, Doyon J, Toga AW, et al. (2000) *MRI Atlas of the Human Cerebellum*. San Diego: Academic Press.

Silveri MC, Leggio MG, and Molinari M (1994) The cerebellum contributes to linguistic production: A case of agrammatic speech following a right cerebellar lesion. *Neurology* 44(11): 2047–2050.

Thach WT and Bastian A (2004) Role of the cerebellum in the control and adaptation of gait in health and disease. *Progress in Brain Research* 143: 353–366.

Cerebellum: Evolution and Comparative Anatomy

M Glickstein, University College London, London, UK
J Voogd, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

This article discusses the structure and function of the mammalian cerebellum from an evolutionary point of view. This approach is only tentative since there are only a few clues from endocasts of fossil skulls; the fossil record can give only limited evidence about the evolution of soft tissues. Since there are few data on the relative size of the entire cerebellum or its parts that can be derived from fossil evidence, and there are no guaranteed living primitive forms, we focus on insights from comparative anatomy and development. Comparative anatomy reveals the relative size of the cerebellum and its subdivisions and the variations upon the common plan among vertebrates; hence, it may provide clues as to its evolution. Studies on the connectivity of the cerebellum give indications about its functions and may provide clues regarding adaptations in cerebellar structure during evolution. Cerebellar morphogenesis and its genetic control provides information on the crucial stages at which mutations make these adaptations possible. Unfortunately, the link between structure and function, so evident for many other parts of the central nervous system, is lacking for the cerebellum. This aspect, the missing link in cerebellar evolution, is discussed in the final section of this article.

The comparative anatomy of the vertebrate cerebellum was reviewed by Llinas, Larsell et al., and Nieuwenhuys et al. The evolution of the cerebellum was considered from a comparative anatomical and developmental perspective by Glickstein et al. Its developmental biology was reviewed by Sotelo. Most of the material and many of the illustrations in this article are drawn from these works.

The cerebellum, the 'small brain,' is located in the rostral roof of the fourth ventricle. Mammalian and avian cerebella are characterized by many transverse fissures, which divide it into lobes, lobules, and folia, but this lobular pattern is not present in all vertebrate cerebella. The cerebellum processes information from widely different sources, including the spinal cord, the brain stem, and the telencephalon, and projects to many different centers in the brain involved in postural adaptations and the generation of movements. The cerebellum is covered by a cortex which is typically three layered. A layer formed by the perikarya of the large Purkinje cells separates the superficial, cell-poor molecular layer from the deep layer of the granule cells. Purkinje cells form the main output element of the cerebellar cortex. The granular and molecular layers both contain several types of interneurons.

The cerebellum is characterized by a lattice structure of its cortex. The flattened dendritic trees of the Purkinje cells, which extend into the molecular layer, are oriented perpendicular to the parallel fibers, the axons of the granule cells (Figure 1). This configuration was discovered more than a century ago by Cajal in the avian cerebellum.

Histology of the Mammalian Cerebellar Cortex

We take the structure of the mammalian cerebellar cortex as our starting point (Figures 1(a)–1(c)). The cerebellar cortex has six major cell types. The dominant cell, first described by Purkinje, the nineteenth-century Czech physiologist/anatomist, is called the Purkinje cell. Purkinje cells extend as a single layer of neurons throughout the entire extent of the cerebellar cortex. Below the layer of Purkinje cells there is a dense concentration of small granule cells. In addition to the granule cells, this layer contains the Golgi cells and the unipolar brush cells. Above the layer of Purkinje cells there is a cell-sparse zone, the molecular layer, which contains two other cell types – the basket cell and the outer stellate cells.

The Purkinje cell has a large, pear-shaped cell body whose dendrites spread out broadly within the molecular layer in an orientation which is transverse to the cerebellar folia. When viewed in sections cut parallel to the fissures of the cerebellum, the dendritic tree of the Purkinje cell extends over a much narrower distance. The dendritic tree of the Purkinje cell can be divided into proximal smooth and distal spiny branches. In mammals, the Purkinje cell axon terminates in the cerebellar nuclei, which are located in the cerebellar white matter, and in the vestibular nuclei. Purkinje cells are inhibitory neurons using GABA as their transmitter.

Granule cells are small, measuring approximately 7 μm in diameter. The axons of granule cells ascend to the molecular layer where they bifurcate in a T junction. The two branches of the axon extend within the molecular layer parallel to the direction of the folium. Because of their orientation, granule cell axons are called parallel fibers. They terminate on the distal, spiny branchlets of the Purkinje cells,

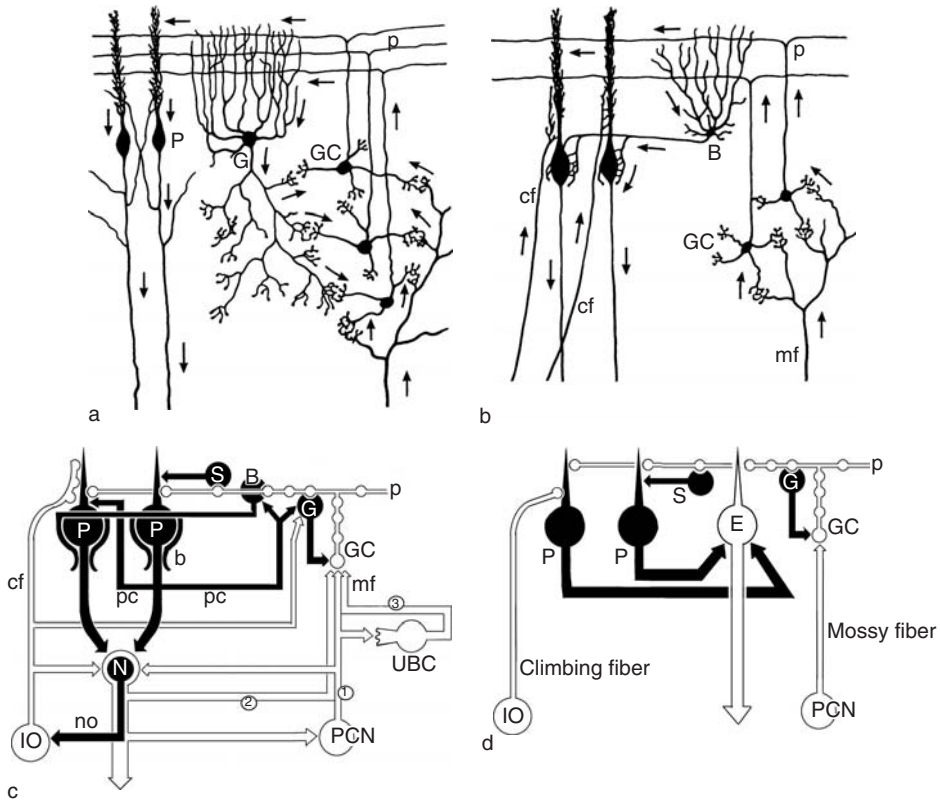

Figure 1 (a) Diagram showing the main mossy fiber–granule cell–Purkinje cell circuit and the innervation of the granule cells by the axonal plexus of the Golgi cells in the mammalian cerebellum. (b) Diagram of the climbing fiber innervation of the Purkinje cells and the basket cells with their axonal baskets surrounding the Purkinje cell perikarya in the mammalian cerebellum. (c) Diagram of the cerebellar circuitry in the mammalian cerebellum. (d) Diagram of the cerebellar circuitry in the teleost cerebellum. Inhibitory neurons are indicated in black. B, basket cell; E, eurydendroid cell; cf, climbing fiber; G, Golgi cell; GC, granule cell; IO, inferior olive; mf, mossy fiber; no, nucleo-olivary axons; pc, recurrent Purkinje cell axon collaterals; P, Purkinje cell; PCN, precerebellar nuclei; pf, parallel fiber; pi, pinceau of basket cell axons; S, stellate cell; UBC, unipolar brush cell; 1, extracerebellar mossy fiber; 2, nucleocortical mossy fiber; 3, mossy fiber collateral of unipolar brush cell. (a, b) Redrawn from Cajal SRy (1972) *Histologie du Système Nerveux de l'homme et des Vertebrés*. Madrid: Consejo Superior de Investigaciones Cientificas.

which they meet on their way, and on dendrites of Golgi, basket, and stellate cells in the molecular layer. Granule cells are excitatory, glutamatergic neurons.

Golgi cells are found below the Purkinje cells, at the top of the granule cell layer. Golgi cell bodies are approximately the same size as those of the Purkinje cell, but there are only approximately one-tenth as many Golgi cells as there are Purkinje cells. The pattern of branching of the dendrites and axons of Golgi cells differs from the pattern of those of the Purkinje cells. Golgi cell dendrites branch in a cylindrical extent within the molecular layer, and their axons remain confined within the granular layer of the cerebellar cortex. Golgi cells are a heterogeneous population of glycinergic inhibitory neurons. Their axons terminate on granule cell dendrites in the granular layer.

Basket and outer stellate cell bodies are located in the molecular layer. Their axons extend in a direction perpendicular to the folia. Basket cell axons extend along a series of Purkinje cells giving off dense basket-like collaterals which surround the Purkinje cell body and terminate at its base. Outer stellate cells have their cell body more superficially located within the granular layer. They too connect to the Purkinje dendrites. Basket and outer stellate cells are inhibitory GABAergic neurons.

Unipolar brush cells are excitatory neurons with one sturdy dendrite, located in the granular layer. Their axons terminate as mossy fibers on granule cell dendrites. Unipolar brush cells are excitatory neurons.

Two extracerebellar afferent fiber systems, mossy and climbing fibers, connect to the cerebellar cortex. Mossy fibers constitute the vast majority of the afferents to the cerebellum. They arise from many sources, including spinal cord brain stem and pontine nuclei, and synapse on the dendrites of the granule cells, with

extensive synapses on the single dendrite of unipolar brush cells. The other class of afferent fibers, the climbing fibers, derive from only a single source, the contralateral inferior olivary nucleus. Unlike the mossy fibers, which are indirectly linked to Purkinje cells by way of granule cells, climbing fibers ascend through the molecular layer and twine directly around the proximal smooth dendrites of the Purkinje cells. Climbing fibers branch sparingly in the cerebellar cortex. Each climbing fiber innervates a single Purkinje cell. Both mossy and climbing fibers are glutamatergic and excitatory.

The principal circuit of the cerebellar cortex consists of mossy fiber afferents, which connect with granule cells in complex synapses (glomeruli). Granule cell axons, the parallel fibers, connect with Purkinje cells. These cells inhibit neurons of the cerebellar or vestibular nuclei, which receive excitatory input from mossy and climbing fiber collaterals. Unipolar brush cells function as boosters for the mossy fiber input. Golgi cells provide feedback inhibition at the level of the glomeruli; stellate and basket cells provide feed-forward inhibition of the Purkinje cells (Figure 1(c)).

One of the main properties of the cerebellar circuitry is its remarkable plasticity. Long-term depression (LTD) and long-term potentiation (LTP) of its synaptic connections are the processes involved. LTD at the parallel fiber–Purkinje cell synapse is the best studied example. Each Purkinje cell receives more than 100 000 synapses from the parallel fibers which traverse its dendritic tree. At any moment, however, only a small fraction of these synaptic connections are available. Most of these synaptic connections are totally or partially depressed by LTD, a process which affects the ionotropic glutamate receptors responsible for the transmission at the parallel fiber–Purkinje cell synapse. LTD depends on the simultaneous activation of climbing fiber and parallel fiber afferents of a Purkinje cell. LTD sculpts the cerebellar cortical network on the basis of previous experience and, most likely, forms the basis of the main function of the cerebellum (i.e., the optimalization of movements). The flattening of the Purkinje cell dendritic trees guarantees maximal access of parallel fibers to the Purkinje cells.

Cerebellar Cortex in Vertebrates

The principal circuit of the cerebellar cortex can be recognized in the cerebellum of all vertebrate classes (Figure 2), with the exception of amphioxus, the lampreys, and the myxinoids, which lack a cerebellum. Flattening of the Purkinje cell dendritic trees and the orthogonal orientation of granule cell axons, the parallel fibers, with respect to the Purkinje cells are general characteristics, although the precision of this arrangement may differ. Cortical interneurons have been identified. Climbing and mossy fibers and a homolog of the inferior olive are universally present. Differences concern the spatial position of Purkinje and granule cells; that is, the presence of a three-layered cortex, the shape of the Purkinje cell dendritic tree, the morphology of the climbing fibers, and the typology of the interneurons.

In cartilaginous fishes, teleosts, amphibians, and reptiles, the Purkinje cell dendritic tree is characterized by the presence of long and straight distal branches, which ascend to the meningeal surface. These distal branches are covered with spines and thus correspond to the spiny branchlets of the mammalian Purkinje cells. The proximal smooth dendrites are largely restricted to the Purkinje cell layer or the deepest part of the molecular layer. Climbing fibers terminate on these smooth dendrites. They do not climb into the dendritic tree but, rather, remain confined to the region of the Purkinje cell layer. The configuration in the avian cerebellum is intermediate between reptiles and mammals. Smooth dendritic branches ascend halfway up into the molecular layer, spiny branchlets reach up to the meningeal surface, and climbing fibers are found in the inner half of the molecular layer.

Golgi and stellate cells have been identified in most, if not all, vertebrate cerebella. However, basket cells, which can be considered as a differentiation of the stellate cells, appear to be absent in cartilaginous and teleost fishes and amphibians.

The position of the Purkinje cell as the sole output element of the cerebellar cortex is maintained in all vertebrates. In teleost fishes, however, the cerebellar nuclei, the main target of the Purkinje cell axons, are absent. Purkinje cell axons terminate on so-called eurydendroid cells, which are located within the Purkinje cell layer (Figures 1(d) and (2)). Unlike Purkinje cells, their dendrites lack spines. The axons of these cells, like those of the cerebellar nuclei, project outside the cerebellum to terminate in the brain stem. A similar configuration may be present in cartilaginous fishes. Projections of Purkinje cells to the vestibular nuclei have been identified in all vertebrates.

Gross Anatomy of the Vertebrate Cerebellum

Although cerebellar circuitry is stereotyped in all vertebrates, the gross morphology and the distribution of granule and Purkinje cells vary greatly in different vertebrate classes. In cartilaginous fishes (Figure 3), the cerebellum consists of an evaginated corpus

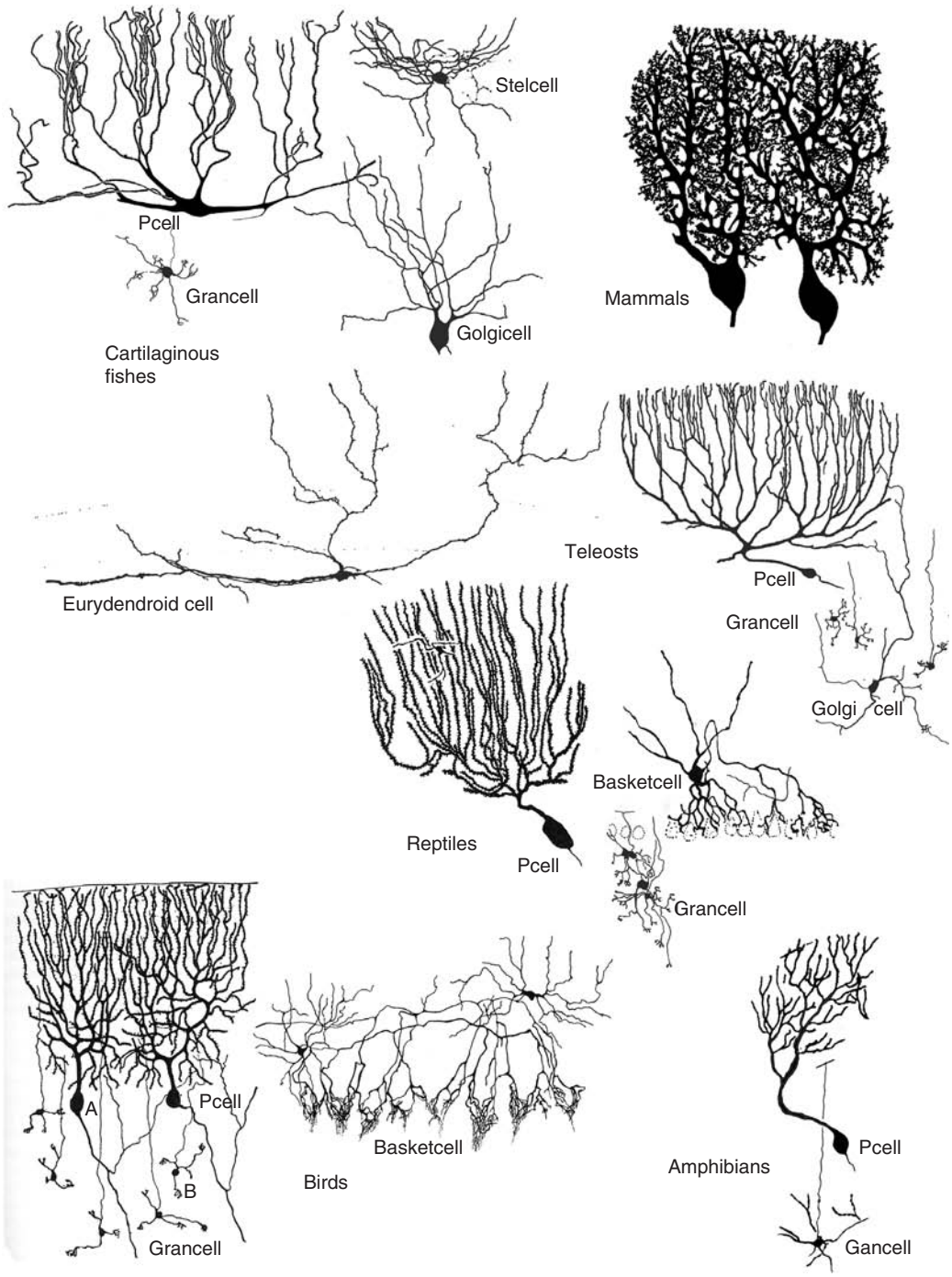

Figure 2 Neurons of the cerebellar cortex in different vertebrates. Mammalian Purkinje cells from Cajal SRy (1972) *Histologie du Système Nerveux de l'homme et des Vertebrés*. Madrid: Consejo Superior de Investigaciones Cientificas; cells from the cerebellum of cartilaginous fishes redrawn from Nicholson C, Llinas R, and Precht W (1969) Neural elements of the cerebellum in elasmobranch fishes; Structural and functional characteristics. In: Llinas R (ed.) *Neurobiology of Cerebellar Evolution and Development*, pp. 215–243. Chicago: American Medical Association; neurons of the frog redrawn from Hillman DE (1969) Neuronal organization of the cerebellar cortex in amphibia and reptilia. In: Llinas R (ed.) *Neurobiology of Cerebellar Evolution and Development*, pp. 279–325. Chicago: American Medical Association; eurydendroid cell redrawn from Pouwels E (1978) On the development of the cerebellum of the trout *Salmo gairdneri*: IV. Development of the pattern of connectivity. *Anatomy and Embryology (Berlin)* 153: 55–65; the others redrawn from Larsell O (1967) *The Comparative Anatomy and Histology of the Cerebellum: Vol. 1. From Myxinoids through Birds*. Minneapolis: University of Minnesota Press.

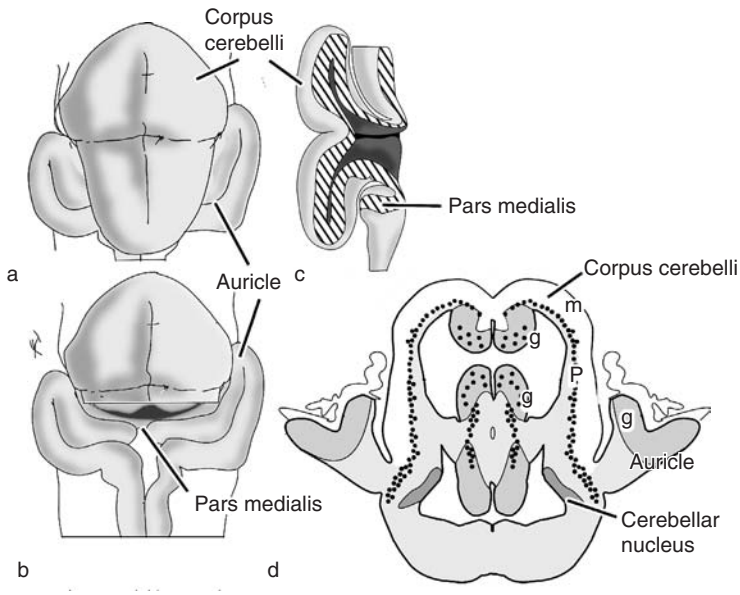

Figure 3 The cerebellum of the shark (*Squalus acanthias*). (a) Dorsal view. (b) Dorsal view; the caudal portion of the corpus cerebelli has been removed to show the connection between the auricles (the pars medialis or pars caudalis). (c) Sagittal section. (d) Transverse section through the cerebellum. The molecular layer (m) is shown in white, the granular layer (g) in medium gray, cerebellar nuclei in dark gray, and the white matter in light gray. Note granule cell ridges with dispersed Purkinje cells (P) and granular layer of the auricles. Redrawn from Larsell O (1967) *The Comparative Anatomy and Histology of the Cerebellum: Vol. 1. From Myxinoids through Birds*. Minneapolis: University of Minnesota Press.

cerebelli that is smooth in small but foliated in large species. Laterally, two auricles are present; the auricles are continuous across the midline by the pars medialis (or lobus caudalis), which caudally adjoins the corpus cerebelli. Granule cells form ridges at the ventricular surface of the corpus cerebelli and occupy the auricles. Purkinje cells are scattered among the granule cells of the ridges and are located in an irregular layer in the lateral corpus cerebelli, where a molecular layer can be identified. Purkinje cells are absent from the auricles (Figure 3). The cerebellar nucleus, identified in some species, receives Purkinje cell axons. The corpus cerebelli receives information from somatosensory and visual centers. The auricles and the lobus caudalis are connected with the lateral line organs and the vestibular centers.

In teleost fish, an evaginated corpus cerebelli, an eminentia granularis (corresponding to the auricle), and a lobus caudalis can be identified. Rostrally, the valvula cerebelli is invaginated into the mesencephalic ventricle. An irregular layer of Purkinje cells and a molecular layer cover the corpus cerebelli and the valvula; scattered Purkinje cells are present in the eminentia granularis. Cerebellar nuclei are absent. The valvula attains enormous proportions in mormyrid fish, in which it pushes the tectum aside and covers the telencephalon (Figure 4). Mormyrids are actively electroreceptive teleosts with an electric organ and electroreceptors involved in electrocommunication and electrolocation. Electroreceptive centers provide the main afferent mossy fiber input to the valvula.

In chondrostean fish, the cerebellum is invaginated into the ventricle. A rostrally projecting valvula, covered by the anterior medullary velum and two auricles connected by the lobus caudalis, can be identified (Figures 5 and 6). Granule cells form multiple ridges in the corpus cerebelli and the auricles (Figures 6(b) and 6(d)). Purkinje cells are scattered among the granule cells and in central portions of the corpus cerebelli. A molecular layer is present in the auricles.

In amphibians, the cerebellum is small and leaflike (Figure 7). Laterally, the auricles are located in the wall of the lateral recess of the fourth ventricle. A multilayer of Purkinje cells and a molecular layer cover the cerebellum. Cerebellar nuclei are present.

The cerebellum in reptiles is larger and compact and covered by a three-layered cortex (Figure 8). Cerebellar nuclei are situated centrally in the cerebellum.

The avian cerebellum is foliated (Figure 9). Parallel cerebellar folia are arranged like the pages of an opened book. Larsell distinguished ten folia in the avian cerebellum, numbered from I to X. The caudal-most

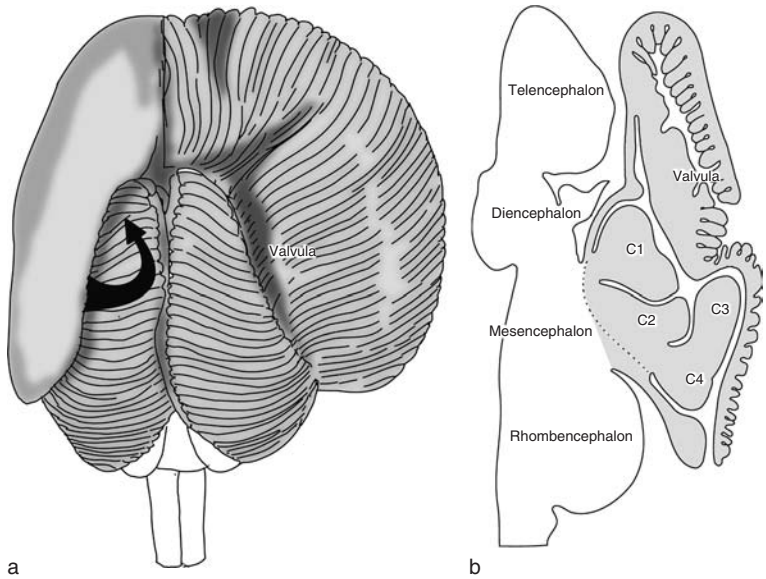

Figure 4 The mormyrid cerebellum. The large size of the cerebellum is due to the valvula. This lobule is folded over the telencephalon. Its surface forms numerous transverse ridges. (a) Dorsal view: the right valvula is unfolded. (b) Diagram of a sagittal section. C1–C4 are subdivisions of the corpus cerebelli. Redrawn from Nieuwenhuys R and Nicholson C (1969) A survey of the general morphology, the fiber connections, and the possible functional significance of the gigantocerebellum of mormyrid fishes. In: Llinas R (ed.) *Neurobiology of Cerebellar Evolution and Development*, pp. 107–134. Chicago: American Medical Association; and Nieuwenhuys R, Ten Donkelaar HJ, and Nicholson C (1998) *The Central Nervous System of Vertebrates*. Heidelberg: Springer.

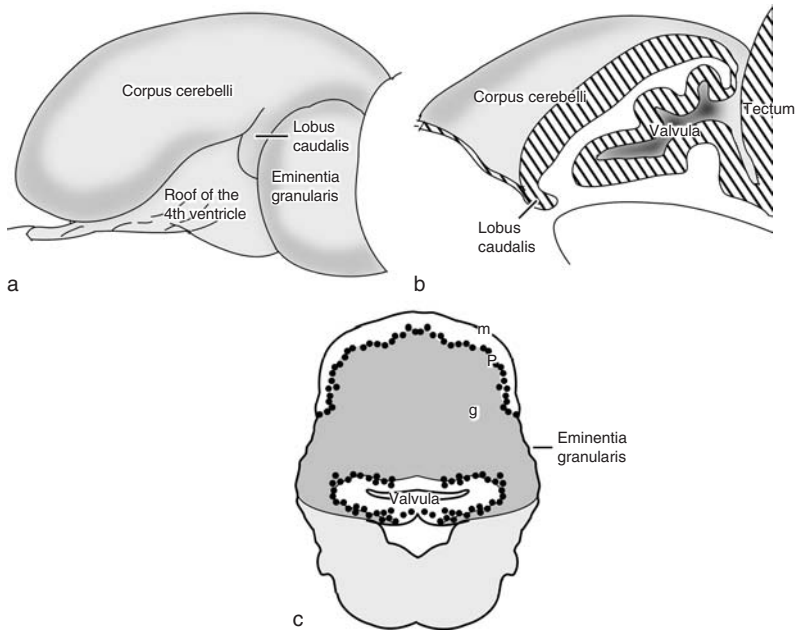

Figure 5 The cerebellum of the trout (*Salmo gairdueri*). (a) Lateral view. (b) Sagittal section. Note the invagination of the valvula – the most rostral part of the teleost cerebellum. (c) Transverse section through the cerebellum. The granule cell layer (g) is shown in medium gray and the molecular layer (m) in white. P, Purkinje cells. Redrawn from Nieuwenhuys R, Ten Donkelaar HJ, and Nicholson C (1998) *The Central Nervous System of Vertebrates*. Heidelberg: Springer.

folia laterally form the auricle, the homolog of the mammalian flocculus. The cerebellum is covered by a three-layered cortex; laterally, this cortex is unfoliated. Cerebellar nuclei occupy the central white matter.

The mammalian cerebellum is foliated and can be subdivided into a median vermis (Latin for 'worm') and two hemispheres (Figure 10). The mammalian cerebellum is divisible into three major lobes which

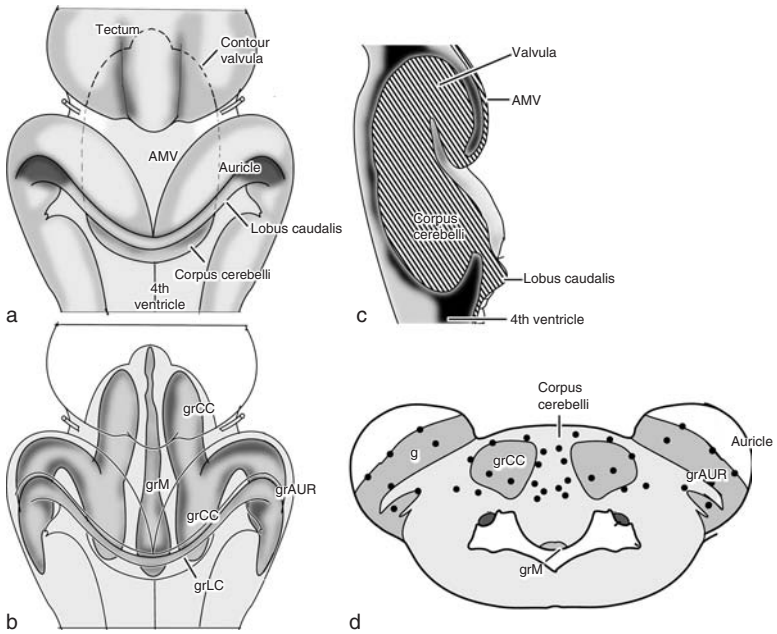

Figure 6 The cerebellum of the shovelnose sturgeon *Scaphirhynchus platorynchus*. (a) Dorsal view of the invaginated corpus cerebelli; only the caudal rim is visible after removal of the roof of the fourth ventricle. (b) Dorsal view of the translucent cerebellum showing the granule cell ridges. (c) Sagittal section of the invaginated cerebellum; the valvula is covered by the anterior medullary velum (AMV). (d) Transverse section through the cerebellum. Granule cell ridges are shown in gray. g, granule cells; grAUR, granule cell ridges of the auricle; grCC, granule cell ridges of the corpus cerebelli; grM, granule cell ridge of the midline. Redrawn from Nieuwenhuys R, Ten Donkelaar HJ, and Nicholson C (1998) *The Central Nervous System of Vertebrates*. Heidelberg: Springer.

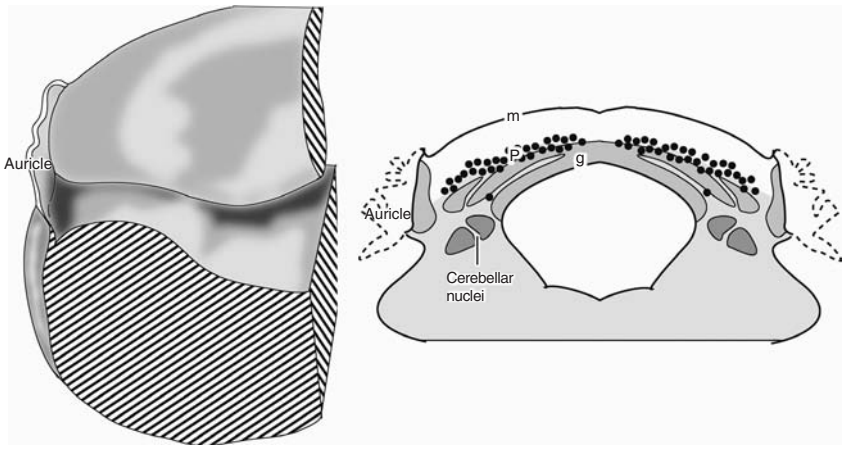

Figure 7 The cerebellum of the frog (*Rana esculenta*). (Left) Rostral aspect of the left half of the leaflike cerebellum. (Right) Transverse section through the cerebellum. The granule cell layer (g) is shown in medium gray, the cerebellar nuclei in dark gray, and the white matter in light gray. The molecular layer (m) is unlabeled. P, Purkinje cells. Redrawn from Larsell O (1967) *The Comparative Anatomy and Histology of the Cerebellum: Vol. 1. From Myxinoids through Birds*. Minneapolis: University of Minnesota Press.

are bounded by two fissures. In all mammals, the primary fissure marks the dividing line between the anterior lobe and the posterior lobe of the cerebellum, and it is particularly deep. Near the caudal end of the cerebellum, the posterolateral fissure separates the posterior lobe from the flocculus and the nodulus (flocculonodular lobe). Early anatomists suggested that the mammalian vermis and the flocculus (the 'paleocerebellum') are identical to the entire avian cerebellum, and that the hemispheres are new acquisitions (the 'neocerebellum'). However, this proposition is probably mistaken. The lateral unfoliated

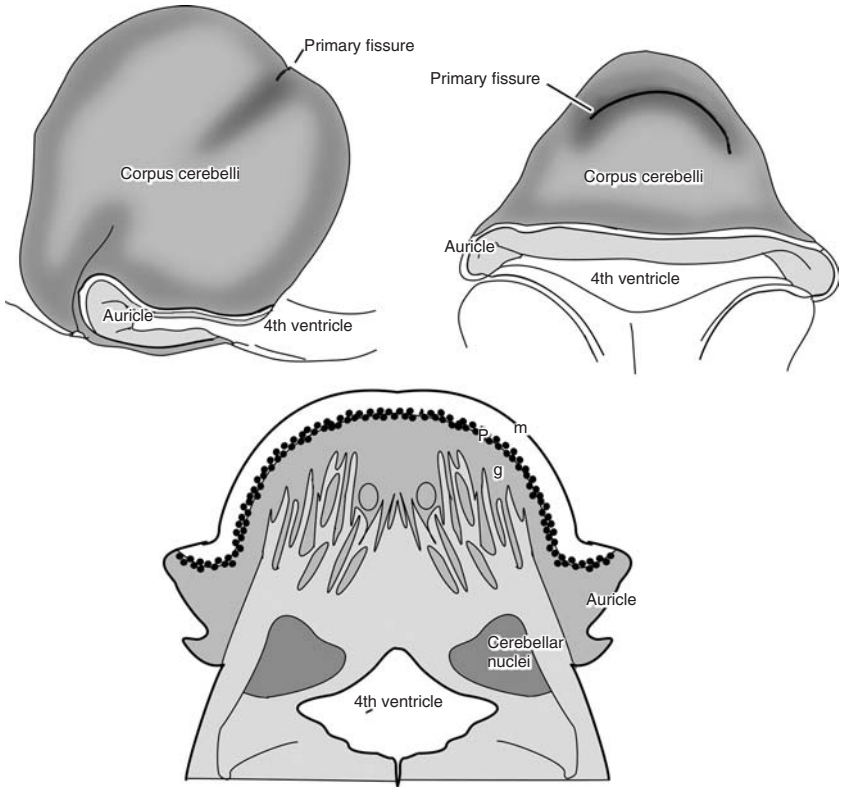

Figure 8 The cerebellum of the alligator (*Alligator missippiensis*). (a) Lateral view. (b) Caudal view. (c) Section through the cerebellum. The molecular layer (m) is unlabeled, granule cell layer (g) is shown in medium gray, cerebellar nuclei in dark gray, and white matter in light gray. P, Purkinje cells. Redrawn from Larsell O (1967) *The Comparative Anatomy and Histology of the Cerebellum: Vol. 1. From Myxinoids through Birds*. Minneapolis: University of Minnesota Press.

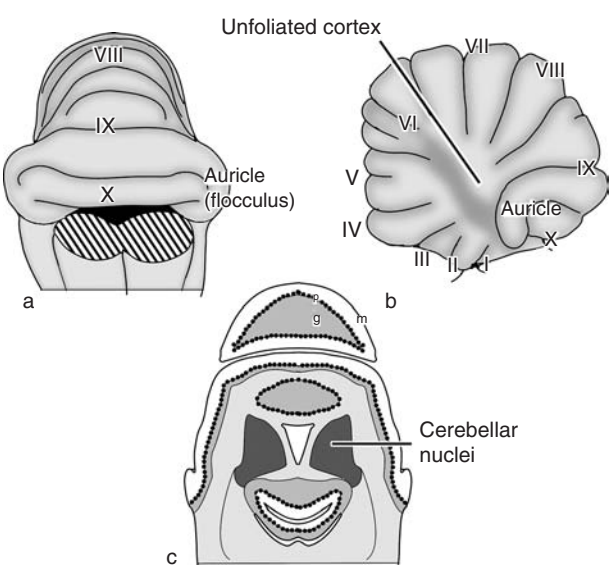

Figure 9 The brain of the pigeon (*Columba livia*). (a) Caudal view. (b) Lateral view. (c) Section through the cerebellum. The molecular layer (m) is unlabeled, the granular layer (g) is shown in medium gray, the cerebellar nuclei in dark gray, and the white matter in light gray. P, Purkinje cells; I–X, lobules I–X of Larsell. Redrawn from Larsell O (1967) *The Comparative Anatomy and Histology of the Cerebellum: Vol. 1. From Myxinoids through Birds*. Minneapolis: University of Minnesota Press.

cortex of the avian cerebellum is the likely homolog of the mammalian hemisphere and the auricle represents the flocculus.

The Dutch anatomist Lodewijk Bolk made the fundamental observation that although the cerebellum appears to be complexly folded, there is a fundamental plan to its organization. The anterior lobe and the first lobule of the posterior lobe, which Bolk called lobulus simplex, constitute a single unbranched chain of cerebellar folia. Just behind the lobulus simplex, the hemispheres become separate from the midline, frequently bending outward and forming two successive loops. Bolk called the first of these loops the ansiform lobe (**Figure 10(c)**). The second of these loops constitutes the paraflocculus and flocculus. The folia first are directed away from the midline in crus I of the ansiform lobe and then return toward the midline as crus II. The chain of folia then continues along the vermis as the paramedian lobule before looping outward again as parafloculus and flocculus. Thus, there is a continuous chain of folia from the most anterior lobule of the anterior lobe to nodulus and flocculus.

Bolk speculated that the pattern of cortical organization reflects a functional division between a midline

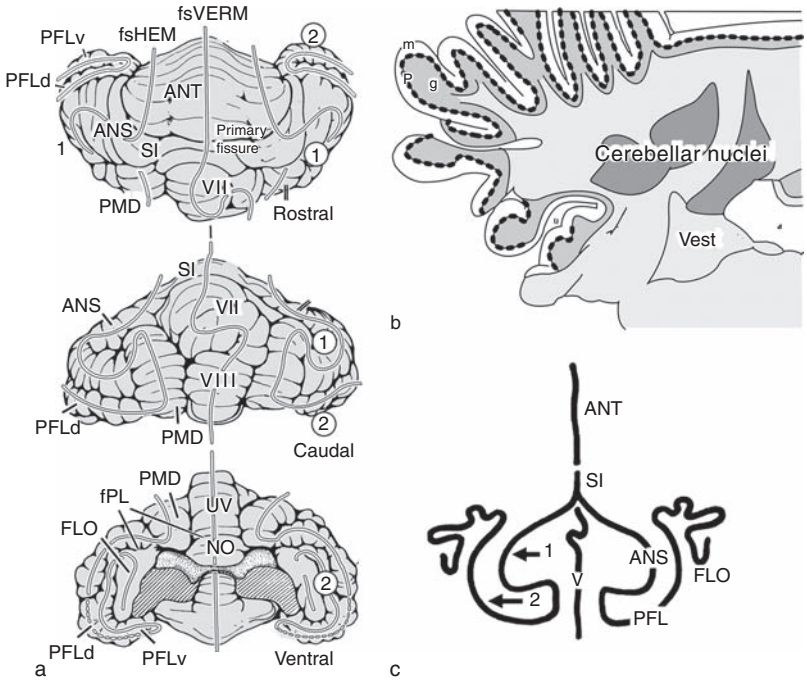

Figure 10 The cerebellum of the cat (*Felis domestica*). (a) Dorsal, caudal, and ventral views of the cerebellum. The direction of the folial chains of vermis and hemispheres is indicated by open lines. The two main loops in the folial chain of the hemisphere are located in the ansiform lobule (1) and in the paraflocculus (2). (b) Tranverse section through the cerebellum. The molecular layer (m) is unlabeled, the granular layer (g) is shown in medium gray, the cerebellar nuclei in dark gray, and the white matter in light gray. P, Purkinje cells. (c) Bolk's wire diagram of the fundamental structure of the mammalian cerebellum. In the anterior cerebellum (i.e., the anterior lobe and the lobulus simplex (SI)), vermis and hemispheres are difficult to distinguish. More caudally, the cerebellum divides into the folial chain of the vermis (v) and the two folial chains of the hemispheres. Two loops in the folial chain of the hemisphere are indicated: one in the ansiform lobule (1) and the other in the paraflocculus (PFL; 2). The last section of the folial chain of the hemisphere turns back on itself (Bolk's terminal hook, or the flocculus (FLO)). ANS, ansiform lobule; ANT, anterior lobe; fPL, posterolateral fissure; fsHEM, folial chain of the hemisphere; fsVERM, folial chain of the vermis; NO, nodulusa; PFLd, dorsal paraflocculus; PFLv, ventral paraflocculus; PMD, lobulus paramedianus; UV, uvula; VII, lobule VII of Larsell; VIII, pyramis.

cerebellum, which is responsible for the regulation of structures whose muscles require collaboration across the midline for effective motor control. For example, consider the necessary collaboration between muscles of the neck that are required for smooth turning of the head. Bolk suggested that the hemispheres are responsible for the control of distal muscles which do not require as close a collaboration between the two sides of the body. Bolk also made the assumption that there is a single map of the body's muscles in the cerebellum. Although the suggestion of a single map is wrong, Bolk's insights on the differential role of midline and hemispheric cerebellum seem valid in light of clinical and experimental evidence.

Figure 11 shows flattened maps of cerebellar cortex of a number of species of mammals, with the relative surface areas preserved. The organization of the cerebellar cortex in all mammals is essentially similar and corresponds to Bolk's fundamental plan. The anterior lobe and lobulus simplex, the first lobule of the posterior lobe, are continuous across the cerebellum from its left to its right side. Behind the lobulus simplex, the cerebellar hemispheres are no longer continuous with the vermis. They extend lateral to the vermis and are often complexly folded.

Longitudinal Modular Organization of the Cerebellar Cortex and Its Efferent Connections

At first glance, the structure of the cortex is uniform and homogeneous over the entire cerebellum. However, electrophysiological, anatomical, developmental, and neurochemical studies have shown that the Purkinje cells of the mammalian and avian cerebellar cortex are arranged in a number of parasagittal, parallel, symmetrically disposed zones. This zonal pattern can be appreciated from a reconstruction of the cerebellum of the rat (Figure 12). The cerebellum was stained with a Purkinje cell-specific antibody, zebrin II. Zebrin-positive and zebrin-negative bands are distributed in alternating longitudinal bands. It has been shown that the zebrin pattern is correlated with the distribution of many different substances

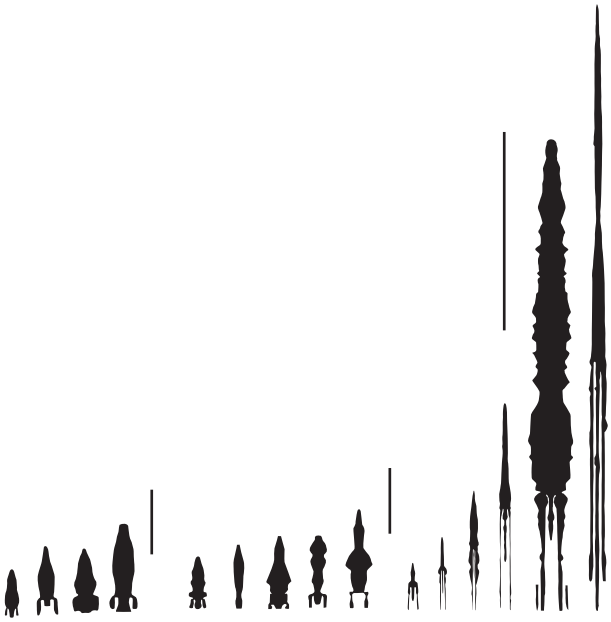

Figure 11 Outlines of the shapes of the cerebellar cortices obtained by connecting the ends of the most prominent folia. The scale is the same in the laterolateral and anteroposterior direction. Note the division of posterior cerebellum in the folial chains of vermis and hemispheres and the relative widths and lengths of these chains in different species. 1, mouse; 2, bat; 3, flying fox; 4, guinea pig; 5, rabbit; 6, pigeon; 7, hare, 8, chinchilla; 9, squirrel; 10, dog; 11, cat; 12, macaque; 13, sheep; 14, human; 15, bovine. Magnifications differ for diagrams 1–4, 5–9, and 10–15. Adapted from Sultan F and Braitenberg V (1993) Shapes and sizes of different mammalian cerebella. A study in quantitative comparative neuroanatomy. *Journal für Hirnforsch* 34: 79–92.

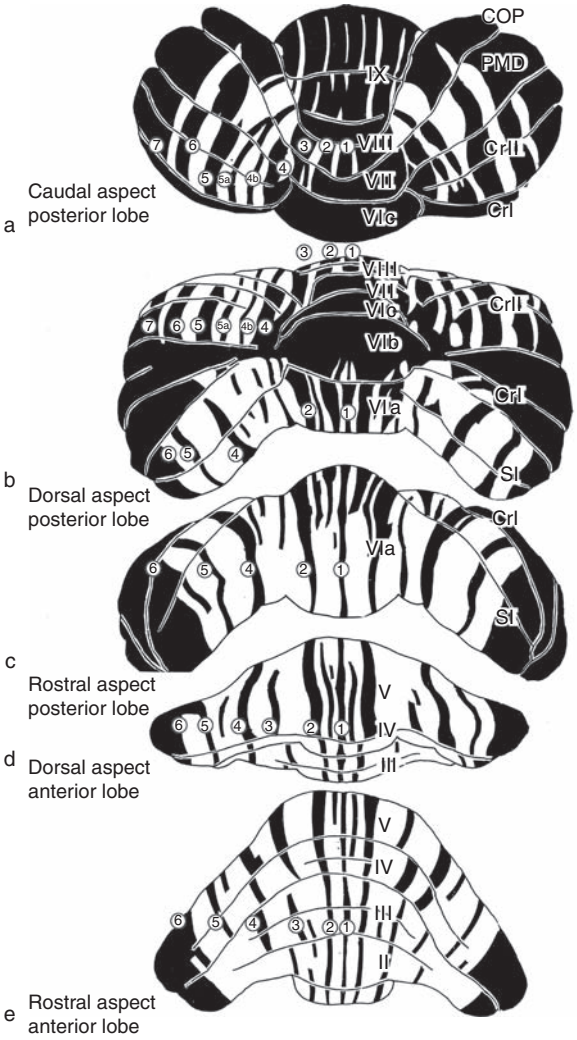

Figure 12 Zebrin-positive and -negative zones in the cerebellum of the rat. Caudal, dorsal, and rostral aspects of the posterior lobe (a–c) and dorsal and rostral aspects of the anterior lobe (d, e) are illustrated. Numbers indicate zebrin-positive Purkinje cell zones PI–P7. COP, copula pyramidis; CrI and CrII, crus 1 and II of the ansiform lobule; PMD, paramedian lobule; S1, simple lobule; I–X, lobules I–X. Reprinted from Hawkes R and Leclerc N (1987) Antigenic map of the rat cerebellar cortex: The distribution of parasagittal bands as revealed by monoclonal anti-Purkinje cell antibody mapQ113. *Journal of Comparative Neurology* 256: 29–41, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

in Purkinje cells. The zebrin pattern can be identified in most mammals investigated as well as in the avian cerebellum. Moreover, Purkinje cells of different zebrin-positive and -negative bands receive their climbing fibers from different subnuclei of the inferior olive and project to particular cerebellar or vestibular nuclei. On the afferent side, mossy fibers terminate in similar longitudinal parcellations of mossy fiber terminals. This modular organization is fundamental to the structure of the cerebellum. It can be recognized in mammals and birds. In reptiles, the distribution of Purkinje cells follows a similar longitudinal organization (Figure 13). The cerebellar or vestibular target nuclei of the modules project to motor centers in the brain stem. In the mesencephalon, they terminate in the red nucleus and the tectum, whereas in the diencephalon they terminate in the motor thalamus. Caudally, they innervate the preoculomotor centers in the vestibular nuclei and the reticular formation and neurons giving rise to descending pathways to the spinal cord.

Variations of the zonal pattern among mammals concern specific lobules (lobule VII, the 'visual vermis,' the ansiform lobule and the paraflocculus) and certain zones (the C2 zone, or zebrin zone 5 (Figure 12), with its projection to the posterior interposed nucleus, and the lateral D1 and D2 zones, zebrin zones 6 and 7, with projections to the dentate nucleus). These variations are related to the evolution of the motor and premotor cortex and the cortical visuomotor system in primates.

Development of the Cerebellum

The cerebellum is derived from an embryonic territory at the junction of the developing midbrain and

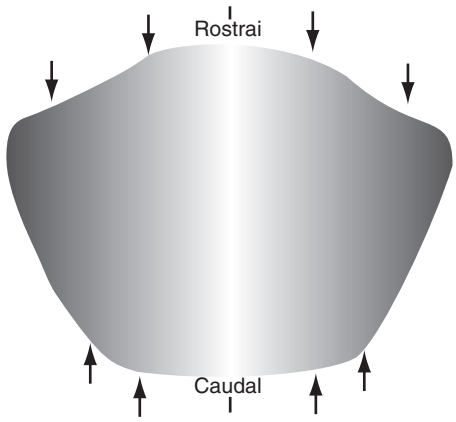

Figure 13 Map of the Purkinje cells of the cerebellum of the turtle *Pseudemys scripta elegans*. Three longitudinal Purkinje cell zones can be distinguished on the basis of cell-poor ridges and differences in the packing of the cells. Reprinted from Bangma GC, ten Donkelaar HJ, and Pellegrino A (1983) Cerebellar corticonuclear projections in the red-eared turtle *Pseudemys scripta elegans*. *Journal of Comparative Neurology* 215: 258–274, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

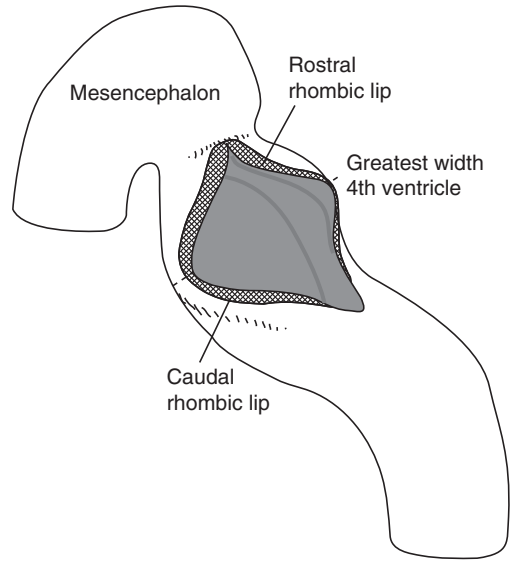

Figure 14 Diagram of the brain stem of a mammalian embryo. The rhombic lip borders on the fourth ventricle. Rostral and caudal parts of the rhombic lip can be distinguished. The rostral rhombic lip and the adjacent ventricular epithelium give rise to the cerebellum; the caudal rhombic lip gives rise to precerebellar nuclei such as the inferior olive and the pons.

hindbrain, the so-called mid-hindbrain boundary (MHB). Chick/quail transplantation experiments have revealed an organizing influence of a band of cells at the MHB. This region directs the formation of the midbrain, cerebellum, and anterior hindbrain, and it can induce midbrain and cerebellum tissue when placed ectopically. Genetic analysis in mice has shown that the position of the isthmic organizer is fixed at the boundary between the domains of expression of two homeodomain transcription factors, Otx2 and Gbx2. Otx2 is typically expressed throughout the embryonic forebrain and midbrain with a caudal border at the MHB; Gbx2 is expressed in the anterior hindbrain with a rostral limit at the MHB.

The cerebellum develops from the rostral rhombic lip and the adjacent ventricular epithelium. The rhombic lip is the dorsal rim of the alar plate which gives attachment to the thin roof plate of the fourth ventricle. The border between the rostral and caudal rhombic lip is located at the greatest width of the fourth ventricle (Figure 14). The caudal rhombic lip gives rise to neurons of precerebellar nuclei, such as the inferior olive and the pontine nuclei.

Both temporal and spatial factors appear to determine the fate of neurons produced by the rostral rhombic lip. Neurons of the cerebellar nuclei are the first cerebellar neurons to be produced (Figure 15), followed by the Purkinje cells. At a later stage, cells produced near the attachment of the roof of the fourth ventricle to the cerebellar anlage migrate rostrally over its meningeal surface. This superficial migration is known as the external granular layer (EGL). During the late prenatal and early postnatal period, the cells of the EGL proliferate and produce granule cells. These cells emit parallel fibers, and their cell bodies migrate through the Purkinje cell layer to take up a position in the internal granular layer. Interaction between parallel fibers and Purkinje cells results in flattening of their dendritic tree. Inhibitory interneurons of the cerebellar cortex arise from a separate population of cells in the ventricular epithelium. These cells migrate into the cerebellar white matter, where they proliferate and differentiate into different types of inhibitory interneurons during late prenatal stages.

An important refinement of this temporal sequence in the production of different cerebellar cell types was published by Hashimoto and Mikoshiba. They showed that Purkinje cells of the mouse cerebellum, born on different dates, settle in different longitudinal zones. However, the birth date of these cells is not the only factor which determines their zonal localization. The importance of positional information is evident from a study by Sgaier et al. Early in its development, the rostral rhombic lip changes in position from a mainly rostrocaudal orientation to a mediolateral one, a transformation illustrated by His in the nineteenth century. This change in position may be caused by mechanical factors and is related to the development of the pontine flexure and/or differential growth within the cerebellar primordium. Sgaier et al. drew attention to this change in position in their interpretations of their fate maps of the cerebellar anlage, and they suggested that the rostral rhombic lip ultimately

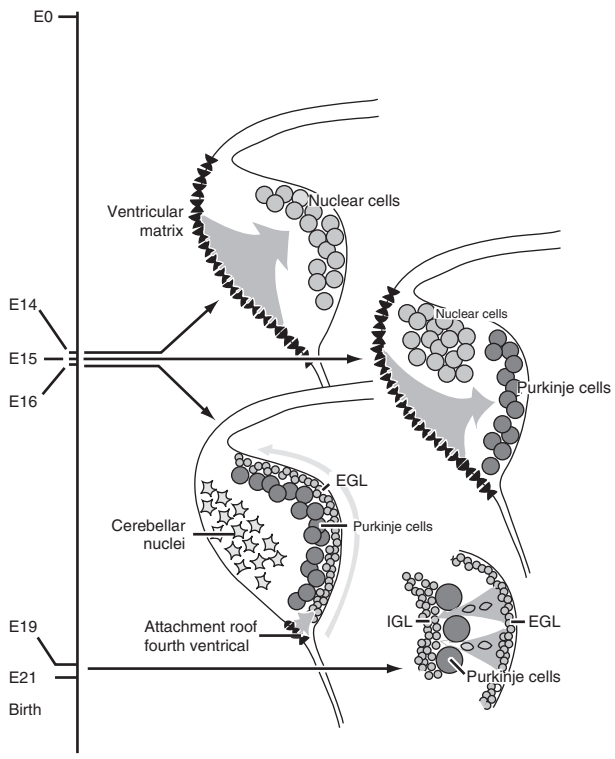

Figure 15 Diagram of the development of the cerebellum of the rat. The earliest neurons to be born in the ventricular matrix are those of the future cerebellar nuclei. They are followed by the Purkinje cells. Cells of the external granular layer (EGL) take their origin from the attachment of the thin roof of the fourth ventricle. These cells migrate rostrally to cover the entire meningeal surface of the cerebellar anlage. Later, the cells of the EGL produce the granule cells, which move inwards through the Purkinje cells, which form a monolayer to take up a position in the inner granular layer (IGL).

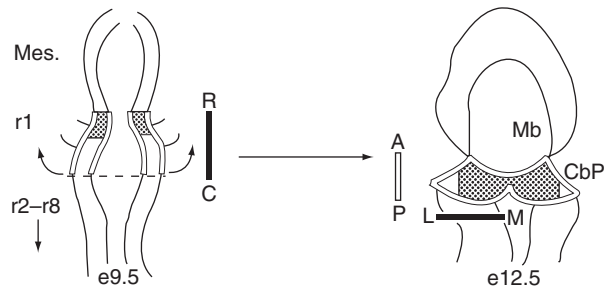

Figure 16 Orthogonal rotation of the anterior–posterior (A–P) axis of the rhombic lip gives rise to the mediolateral orientation of cerebellar zones. Growth of the brain and the cerebellar primordium in mouse between embryonic day (e) 9.5 and 12.5 results in a 90° rotation of the longitudinal axis of the rhombic lip. The darkly shaded region indicates a zone of cerebellar cell clones marked at e7.5 using an En2 promoter-based genetic fate mapping strategy. The open contour line indicates the rhombic lip and the cerebellar primordium. Note that the boundaries of the marked region change from a primarily A–P orientation to a mediolateral one. A and P, AP axis; CbP, cerebellar primordium; M and L, mediolateral axis; Mb, midbrain; Mes, mesencephalon; R and C, rostral and caudal, respectively. Modified from Sgaier SK, Millet S, Villanueva MP, Berenshteyn F, Song C, and Joyner AL (2005) Morphogenetic and cellular movements that shape the mouse cerebellum: Insights from genetic fate mapping. *Neuron* 45: 27–40.

corresponds to the medial cerebellum, and posterior rhombic lip corresponds to the lateral cerebellum. They showed that Purkinje cells in the medial cerebellum originate from the rostral (late medial) rhombic lip and Purkinje cells of the hemisphere from its caudal (late lateral) part (Figure 16).

Many other factors determining cell adhesion, cell recognition, and the expression of their final neurochemical identity, reviewed by Sotelo, are responsible for the morphogenesis of the adult modular organization of the cerebellum. Studies suggest that the division of the cerebellum into zones occurs by mechanisms that are intrinsic to that tissue and genetically encoded. If this is the case, then it is likely that functional differences among the species could be added by slight modifications of the genetic program underlying this pattern, resulting in expansions or contractions of zonal dimensions and/or cell numbers as the need arises. Mutations in genes determining the position of the cerebellum and the production and differentiation of cells in the rostral and caudal rhombic lip, which are responsible for the setting up of the temporal and spatial gradients in the production of different cell types expressing specific recognition and repulsion molecules, which would subserve the future patterning in their connections, are a possible substrate for natural selection.

Motor Control and Motor Learning: The Link between Structure and Function

Preservation of structures during evolution implies that these structures serve a definite purpose. In the case of the cerebellum, it is the circuitry of the cerebellar cortex, with its remarkable degree of plasticity, which has been preserved. This circuitry is characterized by the perpendicular relationship between the parallel fibers and the Purkinje cells and its dual mossy fiber and climbing fiber afferent connectivity. This arrangement guarantees maximal access of mossy fiber–parallel fiber afferents to the Purkinje cell. Moreover, this arrangement is repeated in the longitudinal zonal structure of the cerebellar cortex in reptiles, birds, and mammals. Climbing fibers and cerebellar interneurons enable the selection of a relevant mossy fiber–parallel fiber input in this network. Climbing fibers are generally considered to carry information on errors in motor performance. When climbing fibers and parallel fibers are simultaneously active, the parallel fiber–Purkinje cell synapses are blocked by LTD. Purkinje cell inhibition of the

pertinent motor circuits subsequently corrects the error. Whether this type of motor learning is solely dependent on the cerebellum seems unlikely, but an essential role of the cerebellar circuitry is well established.

The cerebellum is required for direct control of certain classes of movement. Following loss of the cerebellum, monkeys and humans cannot make smooth pursuit eye movements. Saccadic eye movements may be preserved, but they become dysmetric. The role of the cerebellum in motor learning has been exemplified in long-term adaptation of the vestibulo-ocular reflex, the adaptation of saccades, and in the role of the cerebellum in eyelid conditioning.

In general, there are two classes of reflex – open and closed. A closed reflex is arranged so that the reflex acts directly on the part of the body receiving the stimulus. Tapping the knee elongates the tendons connected to the quadriceps muscle. The resultant stretch of the muscle activates the sensory fibers within the spindles of the same muscle. The output of the spindles is connected by way of a single synapse to the motor neurons of the same muscle. Similarly, a painful stimulus to the hand or the foot leads to a flexion and, hence, a withdrawal of the limb from the painful stimulus. Stretch and flexion are examples of closed loop reflexes. The vestibulo-ocular reflex (VOR) is a classic example of an open loop reflex in which the reflex is not fed back onto the input. If the head is turned to the left, there is a reflex tendency for the eyes to move in an equal and opposite direction to the right, thus stabilizing the gaze. Open loop reflexes such as the VOR must be continuously calibrated in order to remain precise and adaptive. The cerebellum is necessary for such calibration. Consider the problem if a person or monkey were to wear a pair of spectacles that magnified the visual image. If the head turned to the left, the VOR would be too small to stabilize the gaze. Similarly, if the image were made smaller, the VOR would be too large. There is clear evidence that if such spectacles are worn for a few days, the amplitude of the VOR changes in the appropriate direction, and that it is the cerebellum which regulates that change using an error signal on retinal slip carried by the climbing fibers and information on vestibular input, eye position, and movement carried by the mossy fibers.

In addition to the role of the cerebellum in the direct control of some movements and the calibration of open loop reflexes, there is also evidence of a role for the cerebellum in classical conditioning. If a rabbit or person is presented with a neutral stimulus such as a tone or a light and if the neutral stimulus is followed after a very brief interval by a puff of air to the cornea, over a series of such paired conditioning trials a conditioned response will develop to the previously neutral stimulus. The sound or light will now cause blinking. Lesions of a particular region of the cerebellum in rabbits or humans result in a deficit in acquiring or retaining a conditioned response. Information on the unconditional stimulus is carried by the climbing fibers. The cerebellar cortex offers the opportunity to pair it with various conditional stimuli, carried by the mossy fibers.

See also: Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Cerebellar Microcircuitry; Cerebellum and Oculomotor Control; Cerebellum: Clinical Pathology; Cerebellum: Models; Cognition: Cerebellum Role; Evolution of Vertebrate Brains; Olivocerebellar System; Procedural Learning: Cerebellum Models; Synaptic Plasticity: Cerebellum; Vestibulo-Ocular Reflex.

Further Reading

Bangma GC, ten Donkelaar HJ, and Pellegrino A (1983) Cerebellar corticonuclear projections in the red-eared turtle *Pseudemys scripta elegans*. *Journal of Comparative Neurology* 215: 258–274.

Bolk L (1906) *Das Cerebellum der Säugetiere*. Haarlem: Fischer.

Bracha V (2004) Role of the cerebellum in eyeblink conditioning. *Progress in Brain Research* 143: 331–339.

Cajal SRy (1888) Estructura de los centros nerviosos de las aves: I. Cerebelo. *Revist trimestr de Histol norm y patol* 1.

Cajal SRy (1972) *Histologie du Système Nerveux de l'homme et des Vertebrés*. Madrid: Consejo Superior de Investigaciones Cientificas.

Ekerot CF and Jorntell H (2003) Parallel fiber receptive fields: A key to understanding cerebellar operation and learning. *Cerebellum* 2: 101–109.

Glickstein M, Voogd J, and Oberdick J (2006) *The Evolution of the Cerebellum. Evolution of Nervous Systems*. Amsterdam: Elsevier.

Hashimoto M and Mikoshiba K (2003) Mediolateral compartmentalization of the cerebellum is determined on the 'birth date' of Purkinje cells. *Journal of Neuroscience* 23: 11342–11351.

Hawkes R and Herrup K (1995) Aldolase C/zebrin II and the regionalization of the cerebellum. *Journal of Molecular Neuroscience* 6: 147–158.

Hawkes R and Leclerc N (1987) Antigenic map of the rat cerebellar cortex: The distribution of parasagittal bands as revealed by monoclonal anti-Purkinje cell antibody mapQ113. *Journal of Comparative Neurology* 256: 29–41.

Hillman DE (1969) Neuronal organization of the cerebellar cortex in amphibia and reptilia. In: Llinas R (ed.) *Neurobiology of Cerebellar Evolution and Development*, pp. 279–325. Chicago: American Medical Association.

Ito M (2006) Cerebellar circuitry as a neuronal machine. *Progress in Neurobiology* 78: 272–303.

Joyner AL, Liu A, and Millet S (2000) Otx2, Gbx2 and Fgf8 interact to position and maintain a mid-hindbrain organizer. *Current Opinion in Cell Biology* 12: 736–741.

Larsell O (1967) *The Comparative Anatomy and Histology of the Cerebellum: Vol. 1. From Myxinoids through Birds*. Minneapolis: University of Minnesota Press.

Larsell O (1970) *The Comparative Anatomy and Histology of the Cerebellum: Vol. 2. From Monotremes through Apes*. Minneapolis: University of Minnesota Press.

Llinas R (1969) *Neurobiology of Cerebellar Evolution and Development*. Chicago: American Medical Association.

McCormick DA, Lavond DG, Clark GA, Kettner RE, Rising CE, and Thompson RF (1981) The engram found? Role of the cerebellum in classical conditioning of the nictitating membrane and eyelid response. *Bulletin of the Psychonomic Society* 18: 103–105.

Miles FA, Braitman DJ, and Dow BM (1980) Long-term adaptive changes in primate vestibuloocular reflex: IV. Electrophysiological observations in flocculus of adapted monkeys. *Journal of Neurophysiology* 43: 1477–1493.

Nagata I, Ono K, Kawana A, and Kimura-Kuroda J (2006) Aligned neurite bundles of granule cells regulate orientation of Purkinje cell dendrites by perpendicular contact guidance in two-dimensional and three-dimensional mouse cerebellar cultures. *Journal of Comparative Neurology* 499: 274–289.

Nicholson C, Llinas R, and Precht W (1969) Neural elements of the cerebellum in elasmobranch fishes: Structural and functional characteristics. In: Llinas R (ed.) *Neurobiology of Cerebellar Evolution and Development*, pp. 215–243. Chicago: American Medical Association.

Nieuwenhuys R and Nicholson C (1969) A survey of the general morphology, the fiber connections, and the possible functional significance of the gigantocerebellum of mormyrid fishes. In: Llinas R (ed.) *Neurobiology of Cerebellar Evolution and Development*, pp. 107–134. Chicago: American Medical Association.

Nieuwenhuys R, Ten Donkelaar HJ, and Nicholson C (1998) *The Central Nervous System of Vertebrates*. Heidelberg: Springer.

Pouwels E (1978) On the development of the cerebellum of the trout *Salmo gairdneri*: IV. Development of the pattern of connectivity. *Anatomy and Embryology (Berlin)* 153: 55–65.

Robinson DA (1976) Adaptive gain control of vestibuloocular reflex. *Journal of Neurophysiology* 39: 954–969.

Sgaier SK, Millet S, Villanueva MP, Berenshteyn F, Song C, and Joyner AL (2005) Morphogenetic and cellular movements that shape the mouse cerebellum: Insights from genetic fate mapping. *Neuron* 45: 27–40.

Sotelo C (2004) Cellular and genetic regulation of the development of the cerebellar system. *Progress in Neurobiology* 72: 295–339.

Sultan F and Braitenberg V (1993) Shapes and sizes of different mammalian cerebella. A study in quantitative comparative neuroanatomy. *Journal für Hirnforsch* 34: 79–92.

Yeo CH, Hardiman MJ, and Glickstein M (1985) Classical conditioning of the nictitating membrane response of the rabbit: II. Lesions of the cerebellar cortex. *Experimental Brain Research* 60: 99–113.

Zhang L and Goldman JE (1996) Generation of cerebellar interneurons from dividing progenitors in white matter. *Neuron* 16: 47–54.

Cerebellum: Models

M Kawato, ATR Computational Neuroscience Laboratories, Kyoto, Japan

© 2009 Elsevier Ltd. All rights reserved.

Cerebellar Learning Models

The cerebellum is one of the best- and earliest-studied parts of the brain. In the past 50 years, a number of conceptual models have proposed various functions, such as error correction, coordination, timing processing, and rhythm. Among these, cerebellar learning models have always been the most attractive, versatile, and influential. They are also the most explicit class of theory from the computational viewpoint. A large number of experiments and theoretical studies have been conducted to resolve the controversies and disputes surrounding them. The learning models are based on unique structures of the cerebellar cortex. The Purkinje cell, the only output neuron in the cerebellar cortex, receives two major excitatory inputs from climbing fibers and parallel fibers. A single climbing fiber axon makes multiple synapses on a single Purkinje cell and induces very strong excitatory postsynaptic potentials, which lead to complex spikes that consist of a few Ca^{2+} spikes. By contrast, 200 000 synapses from parallel fibers, which are the axons of 10^{10-11} granule cells, generate simple spikes, each of which is a single Na^{+} spike.

Brindley first proposed that parallel fiber–Purkinje cell synapses undergo plastic changes if the climbing fiber input is concomitantly activated. Around 1970, Marr, Albus, and Ito postulated the first learning theory of the cerebellar neural circuits. Albus correctly predicted that synaptic efficacy decreases if the parallel and climbing fibers are simultaneously activated, which was later physiologically demonstrated as long-term depression (LTD) by Ito and colleagues. Ito proposed that the climbing fiber input carries an error signal rather than a teacher (desired output) signal and explained adaptation of the vestibulo-ocular reflex by an LTD-based learning model.

A general model of cerebellar learning, which is an extension of the model proposed by Ito to explain general movements with many degrees of freedom and a simplification of the model proposed later by Kawato, is shown in Figure 1. Mossy fiber inputs carry information on the goal of a movement, sensory feedback about the current state of the body and the external world, and the efference copy of motor commands to granule cells. Purkinje cells, mainly activated by parallel fiber inputs from granule cells, project to cerebellar nuclei, which further transmit information to downstream motor systems such as the oculomotor system, spinal cord, red nucleus, and cerebral cortex via the thalamus (depicted together as the postcerebellar network in Figure 1). Motor commands generated by this network induce movements, which are measured by sensory organs. Subtracting this measured state feedback to the brain from the desired state gives the sensory error, which is further transformed into the appropriate coordinates for Purkinje cell outputs. The inferior olive neurons send this transformed error back to Purkinje cells via the climbing fibers.

Cerebellar Purkinje cell learning can be modeled by the following simple supervised learning based on the known synaptic plasticity of parallel fiber–Purkinje cell synapses, including LTD and long-term potentiation (LTP), and the rebound potentiation of inhibitory synapses on Purkinje cells, as well as LTP between parallel fibers and inhibitory interneurons. For simplicity, the firing rate $-y$ of a Purkinje cell is assumed to be the weighted summation of 200 000 parallel fiber inputs x_i:

$$-y = \sum_{i=1}^{200\,000} w_i x_i$$

Then, the inhibitory action on deep cerebellar nuclei is negative of the Purkinje output, y. If the desired output for the sign-inverted Purkinje cell's simple spike is denoted as y_d, then the objective function to be minimized for the supervised learning, in which the synaptic weight w_i is learned so that the actual output y best approximates the desired output y_d, is given as

$$E = 1/2(y_d - y)^2$$

The steepest descent direction of this error function can be obtained by the following partial derivative (called the Widrow–Hoff rule):

$$\Delta w_i, \propto -\frac{\partial E}{\partial w_i} = -\left(-\frac{\partial y}{\partial w_i}\right)(y_d - y) = -x_i(y_d - y)$$
$$= -x_i(C - \bar{C})$$

Here, $(C - \bar{C})$ is the difference between the climbing fiber firing rate and its temporal average (background firing activity), and it is assumed to encode the error signal in the last equation.

When both parallel and climbing fibers are activated, x_i and $(C - \bar{C})$ are positive, and thus Δw_i is negative (LTD). In contrast, when the parallel fiber is activated and the climbing fiber is not, then

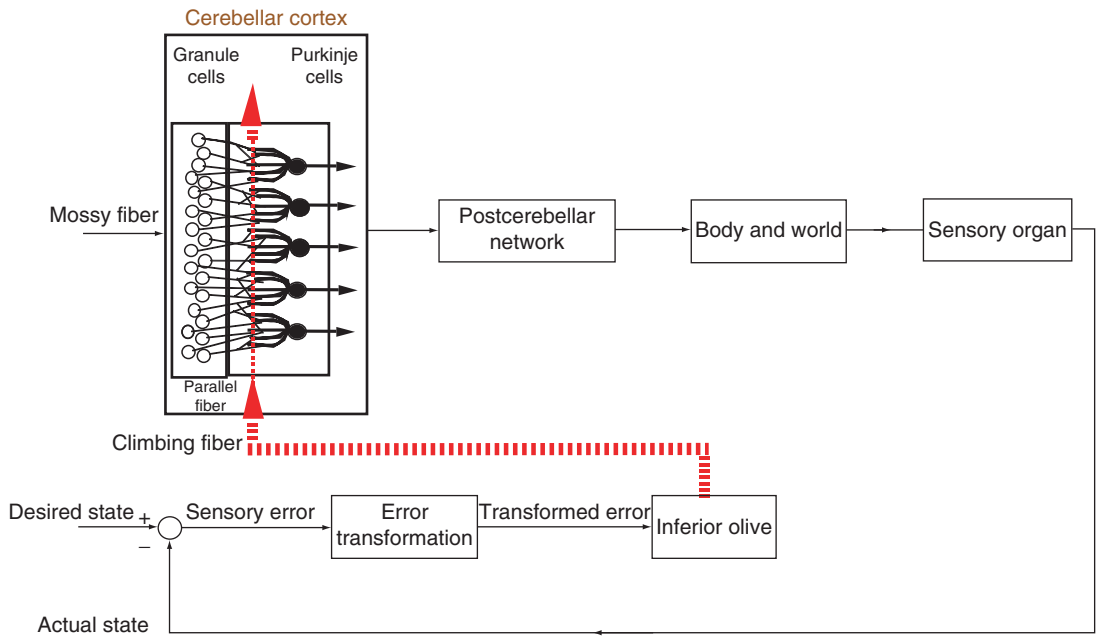

Figure 1 A cerebellar learning model.

x_i is positive but $(C-\bar{C})$ is negative, and thus Δw_i is positive (LTP). When both the climbing fiber and the inhibitory interneuron synapse on a Purkinje cell are activated, x_i and $(C-\bar{C})$ are positive, and thus Δw_i is negative, leading to an increase in the amplitude of w_i, which is negative (LTP of inhibitory pathway).

Cerebellar learning theory predicts that a memory of learning is stored at parallel fiber–Purkinje cell synapses; that LTD is an elementary cellular process of cerebellar learning; and that climbing fibers carry an error signal with some precision in the output coordinates of Purkinje cells. These predictions have all stimulated a vast amount of experimental examination and controversy, but they have been quite well supported by recent data and modeling four decades after they were first proposed. These controversial issues are discussed below.

It is well established that when the cerebellum as a whole is lesioned, for some movements, learning capability is lost, stored memory of learning is destroyed, and movements are not well corrected or coordinated. Controversies have centered mainly on the question of whether the cerebellar cortex (Purkinje cell synapses) or the cerebellar nucleus is the major site of memory. Recent systematic lesion studies have shown that the cerebellar cortex is a memory site for several days at least and that the cerebellar nuclei can contain even longer memory. Interfering with LTD pharmacologically (e.g., by administration of hemoglobin, which is a scavenger of NO, essential for LTD) or by genetic manipulation (knockout mice and their pharmacological rescue) are more recent and sophisticated ways of resolving this controversy. Abolishment of LTD was shown to destroy various motor learning capabilities, such as adaptation of walking, vestibulo-ocular reflex, optokinetic eye movements, and precise timing in eyeblink conditioning.

Computational Models of Cerebellar LTD

If LTD of parallel fiber–Purkinje cell synapses is the elementary cellular process of supervised learning, as postulated in the cerebellar learning theory, it must satisfy the following three requirements. First, LTD should be associative. That is, LTD should be induced only when both parallel fiber inputs and climbing fiber inputs are conjunctively activated. Second, LTD should be maximally induced when climbing fiber activation is delayed 50–200 ms with respect to parallel fiber activation. This is because the time window of LTD should compensate for the delay of the reafferent climbing fiber inputs to feedback the consequences of the parallel fiber–simple spike-induced movements. Third, LTD should be synapse specific; only a parallel fiber synapse that is conjointly activated with a climbing fiber input should be depressed, but not other synapses. However, a considerable number of studies have reported violations of these three requirements under some experimental conditions; consequently, LTD has been questioned several times as the elementary process of cerebellar supervised learning and has been postulated for other functions, such as the

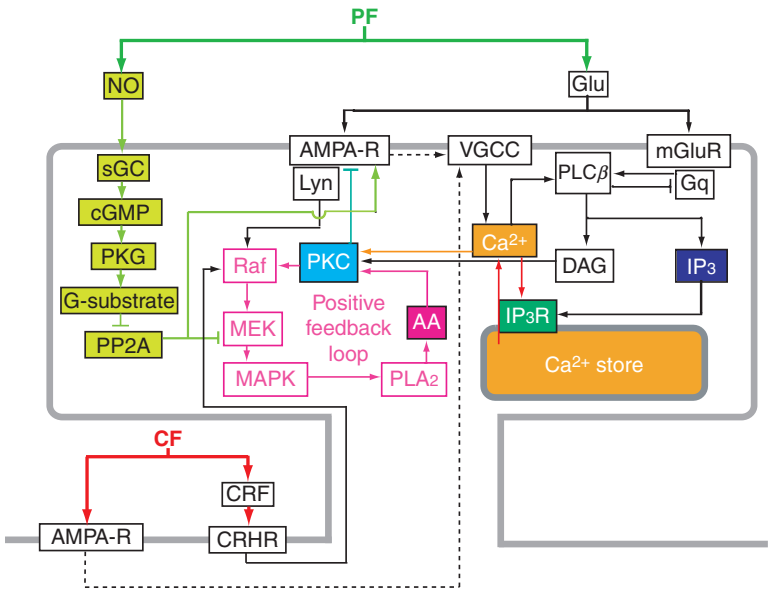

Figure 2 Signal transduction model of cerebellar long-term depression. AA, arachidonic acid; AMPA-R, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; CF, climbing fibers; cGMP, cyclic guanosine monophosphate; CRF, corticotropin releasing factor; CRHR, the human corticotropin releasing factor; DAG, diacylglycerol; Glu, glutamate; Gq, G-proteins; IP_3, inositol 1,4,5-triphosphate; IP_3Rs, IP_3 receptors; MAPK, mitogen-activated protein kinase; MEK, MAP/extracellular signal-regulated kinase; mGluR, metabotropic glutamate receptor; NO, nitric oxide; PF, parallel fibers; PKC, protein kinase C; PKG, cGMP-dependent protein kinase; PP2A, protein phosphatase 2A; PLA_2, cytosolic phospholipase A_2; PLCβ, phospholipase Cβ; sGC, soluble guanylyl cyclase; VGCC, voltage-gated calcium channel.

normalization of total synaptic efficacies on a single Purkinje cell. The kinetic simulation model of LTD signal transduction pathways in a Purkinje cell dendritic spine (Figure 2) provides a coherent perspective on these confusing experimental data and competing theoretical hypotheses, and it is described below.

Climbing fiber inputs activate α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors (AMPA-Rs) on dendritic shafts and induce depolarization, which opens the voltage-gated calcium channels (VGCCs), as indicated by the broken line in Figure 2, leading to Ca^{2+} influx from the extracellular space into the spine head, reaching a concentration of about $1\,\mu mol\,l^{-1}$. Parallel fiber inputs also induce AMPA-R activation on dendritic spines, just as a climbing fiber input does, but parallel fiber inputs produce smaller depolarization and Ca^{2+} influx. In addition, they simultaneously activate metabotropic glutamate receptors (mGluRs). Furthermore, they presynaptically synthesize nitric oxide (NO), which diffuses into the spine. In the metabotropic pathway, activated mGluRs induce production of inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol via G-proteins and phospholipase Cβ at the postsynaptic density. IP_3 then diffuses into the cytosol and binds to IP_3 receptors (IP_3Rs), which are IP_3-gated Ca^{2+} channels on the intracellular Ca^{2+} stores, such as endoplasmic reticulum (ER). Cytosolic Ca^{2+} above some threshold triggers a regenerative cycle of IP_3-dependent Ca^{2+}-induced Ca^{2+} release (depicted as a positive feedback loop between Ca^{2+} and IP_3Rs in Figure 2) from the ER via IP_3Rs opened by binding of both IP_3 and Ca^{2+}. This means that the open probability of IP_3Rs increases with Ca^{2+} concentration, and the Ca^{2+} threshold of this regenerative cycle decreases with IP_3 concentration. An increase in cytosolic IP_3 concentration was simulated to be maximal at $3\,\mu mol\,l^{-1}$ about 100 ms after the parallel fiber input. This slow rate of increase in IP_3 concentration is due to the slow dynamics of the metabotropic pathway. Because Ca^{2+} influx via VGCCs due to climbing fiber activation is an immediate event, when the climbing fiber input is delayed about 100 ms with respect to the parallel fiber input, Ca^{2+} increase coincides with IP_3 increase, and they together trigger the Ca^{2+}-induced Ca^{2+} release of $10\,\mu mol\,l^{-1}$. If either climbing fiber input or parallel fiber input alone is activated, this large Ca^{2+} release does not occur because either IP_3 or Ca^{2+}, respectively, does not reach the threshold for the regenerative cycle. Consequently, the Ca^{2+} nonlinear excitable dynamics model by Doi (right third of Figure 2) reproduced the associative nature and the spike timing-dependent plasticity (STDP) of LTD.

In the model proposed by Kuroda, the large Ca^{2+} increase, which is caused by the conjunctive parallel fiber input and the delayed climbing fiber input, induces

the activation of linear cascades of phosphorylation of protein kinase C (PKC), followed by phosphorylation and then internalization of AMPA-R, which explains the early phase of LTD up to about 10 min. On the other hand, the intermediate phase of LTD, up to several times longer, is mediated by a mitogen-activated protein (MAP) kinase-dependent positive feedback loop, which consists of PKC, Raf, MAP/extracellular signal-regulated kinase (MEK), MAP kinase, cytosolic phospholipase A_2, arachidonic acid, and back to PKC (Figure 2). This positive feedback loop has two stable equilibrium points separated by a saddle point that determines the threshold of LTD. At the low equilibrium point, the MAP kinase positive feedback loop is not activated, but if strong and long Ca^{2+} elevation occurs, the state jumps to the higher equilibrium over the saddle point, and AMPA-Rs continue to be phosphorylated and internalized during this prolonged activation of the positive feedback loop. Consequently, this MAP kinase positive feedback model (middle third of Figure 2) explains the long-term nature of LTD and is supported by George Augustine, Keiko Tanaka, and collaborators.

In the Kuroda model (left third of Figure 2), NO activates soluble guanylyl cyclase (sGC), and the activated sGC catalyzes the conversion of guanosine triphosphate into cyclic guanosine monophosphate (cGMP), which activates cGMP-dependent protein kinase (PKG). PKG phosphorylates its substrate, G-substrate, and phosphorylated G-substrate preferentially inhibits protein phosphatase 2A (PP2A). Activated PP2A dephosphorylates MEK and AMPA-R. Inhibition of the dephosphorylation of MEK was required for the activation of the MAP kinase positive feedback loop. Therefore, by ultimately inhibiting PP2A and activating the positive feedback loop, NO plays a permissive role in the induction of LTD.

The Ca^{2+} dynamics model coherently reproduces both the associative and the nonassociative natures of LTD and the spike timing dependency of associative LTD as follows. Within the physiological ranges of input strength, the conjunctive stimulation of the parallel fiber input and the delayed climbing fiber input is essential to induce a large, regenerative Ca^{2+} increase, which is necessary for crossing the threshold of the MAP kinase positive feedback loop and inducing LTD. However, if stimulation of parallel fiber bundles is strong enough, or if uncaged Ca^{2+} or IP_3 is abundant enough, these stimuli alone can increase the Ca^{2+} concentration so that it crosses the Ca^{2+} dynamics threshold, resulting in a large enough Ca^{2+}-induced Ca^{2+} release from the ER to reproduce nonassociative LTD. In the hippocampal and cerebral pyramidal neurons, backpropagation of somatic action potentials releases the magnesium block on *N*-methyl-D-aspartate (NMDA) receptors and is the essential mechanism for millisecond-order STDP there. Backpropagation of somatic action potentials requires activity of postsynaptic cells, which is the essential condition of Hebbian learning. Neither backpropagation nor the NMDA receptors exist in Purkinje cells, suggesting cellular mechanisms for the clear distinction of Purkinje cell-supervised learning, with 100-ms STDP, from cerebral Hebbian learning, with STDP of tens of milliseconds.

The excitable dynamics of the Ca^{2+} model and bistable dynamics of the MAP kinase positive feedback loop model may provide a possible molecular mechanism to resolve the 'plasticity–stability' dilemma of memory stability with environmental sensitivity. Each spine is equipped with a surprisingly small number of molecules (e.g., only tens of AMPA-Rs). With a cascade of excitable and bistable dynamics starting from Ca^{2+} dynamics, followed by the MAP kinase positive feedback loop, internalization of AMPA receptors, a possible change in cytoskeleton and membrane proteins, and finally a morphological change in the spine, LTD can be induced by very delicate and low-energy inputs but acquired memory can be maintained over a prolonged period of days, even under stochasticity due to the small numbers of molecules.

The synapse specificity of LTD is also well explained by the NO model combined with an electrical cable model of Purkinje cell dendrite and spines, the Ca^{2+} dynamics model, and the MAP kinase positive feedback loop model. When no nearby parallel fiber is activated and NO is low, even a conjunction of the parallel fiber and climbing fiber inputs cannot induce LTD. If too many surrounding parallel fibers are stimulated and NO is very high, even the synapse, which is not stimulated, exhibits LTD only by climbing fiber stimulation. In this case, a spread of LTD occurs, and the synapse specificity of LTD is lost. Only when a modest number of parallel fibers are activated and NO is within an intermediate range can synapse-specific LTD occur; the contextual information for storing different motor skills might be conveyed by this gating mechanism. This strongly predicts the sparseness of parallel fiber coding *in vivo*, which has recently gained much experimental support.

Altogether, the LTD kinetic simulation models provide a comprehensive account of several seemingly diverse, conflicting, and unrelated experimental data within the framework of complicated nonlinear dynamics regulated by Ca^{2+}, IP_3, and NO concentrations. We must note that because of the nonlinear summation effects of these signaling molecules, nonphysiological stimuli *in vitro*, such as parallel fiber bundle stimulation, uncaging of Ca^{2+}, and IP_3, might induce characteristics qualitatively different

from those induced *in vivo* LTD in terms of its associative nature, spike timing dependency, and synapse specificity.

Rhythmic versus Chaotic Firing of Inferior Olive Neurons

The most influential model of the cerebellum other than the learning theory is probably the timing, rhythm, and synchronization hypothesis. The inferior olive nucleus, which sends out climbing fibers to Purkinje cells, possesses the highest density of gap junctions in the mammalian brain. Inferior olive neurons have several unique ion channels distributed on the somatic and dendritic membrane, as shown in Figure 3(a), and these exhibit rhythmic activities in slice preparation under some conditions, especially with harmaline. However, in awake monkeys at rest, no rhythmicity has been found for complex spikes. When the effective electrical conductance of gap junctions is increased by picrotoxin application to the inferior olive, which blocks inhibitory synapses on glomeruli of inferior olive neurons from cerebellar nuclei, the inferior olive neurons tend to be spatially synchronized in their firing.

None of these unique and well-studied anatomical and physiological characteristics of inferior olive

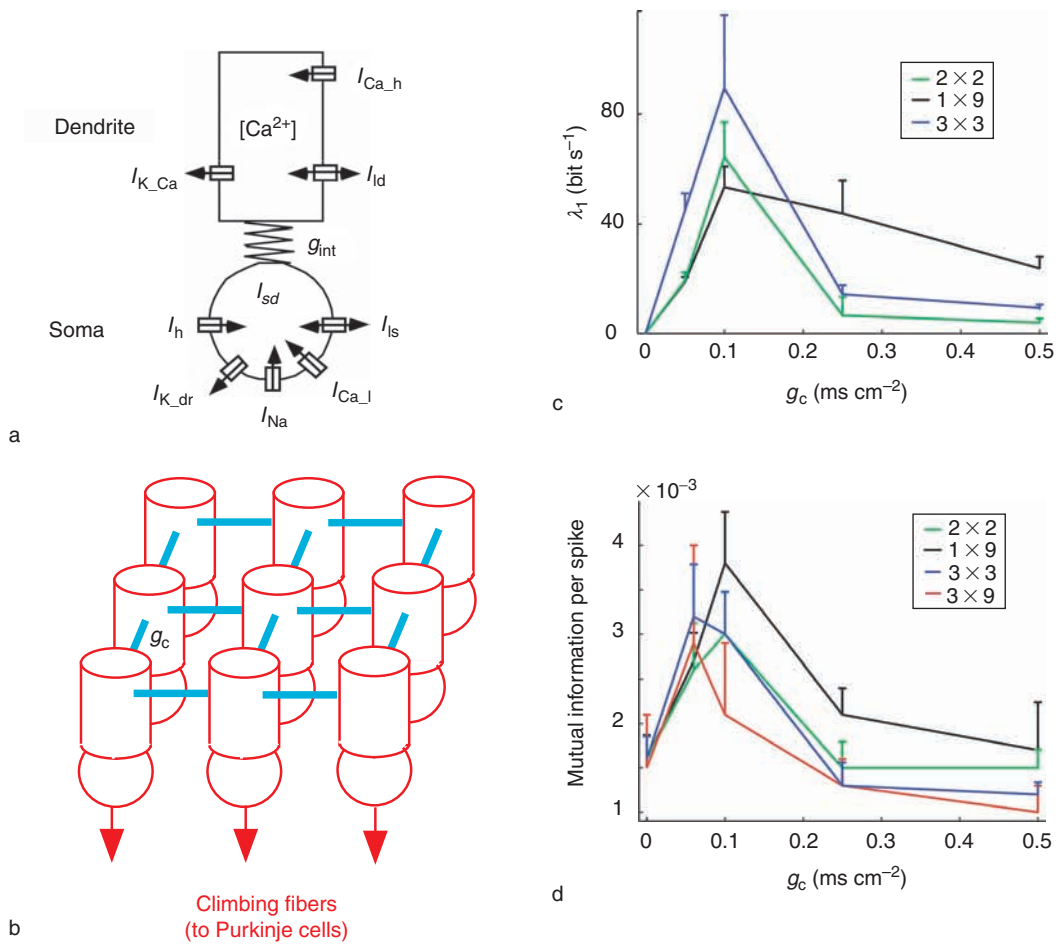

Figure 3 A network model of inferior olive (IO) neurons electrically connected by gap junctions. (a) Two-compartment biophysical model of an IO cell. The somatic ionic currents include a low-threshold calcium inward current (I_{Ca_l}), an anomalous inward rectifier current (I_h), the Hodgkin–Huxley type inward sodium (I_{Na}) and outward delayed rectifier potassium (I_{K_dr}) currents, and a leakage current (I_{ls}). The dendritic currents include a high-threshold inward calcium current (I_{Ca_h}), an outward calcium-dependent potassium current (I_{K_Ca}), and a leakage current (I_{ld}). As the dendritic potassium current I_{K_Ca} is calcium dependent, Ca^{2+} concentration dynamics is included in the model. g_{int} is the coupling conductance between the two compartments and I_{sd} is the current flowing through this conductance. The electronic coupling between this cell and multiple other cells is not shown. (b) Between 4 and 27 IO cells are connected on a lattice by electrical conductance via gap junctions; g_c, coupling strength. (c) Largest Lyapunov exponent as a function of g_c for different networks of nonidentical cells. Error bars are standard deviations obtained from 10 different simulations. When the largest exponent was large, the network exhibited marked chaotic behavior. When the largest exponent was close to zero, firing was almost rhythmic. (d) Network mutual information per spike as a function of g_c for different networks of nonidentical cells. Adapted from Schweighofer N, Doya K, Fukai H, Chiron JV, Furukawa T, and Kawato M (2004) Chaos may enhance information transmission in the inferior olive. *Proceedings of the National Academy of Sciences of the United States of America* 101: 4655–4660, with permission from National Academy of Sciences, USA.

neurons has been incorporated into or explained by the learning theory until recently. Schweighofer and colleagues developed a two-compartment model of an inferior olive neuron, as shown in Figure 3(a), with known spatial distribution of ion channels. These neurons are connected by gap junctions within a planar sheet, as shown in Figure 3(b), and the firing characteristics, as well as the input and output information transfer characteristics, have been examined. The network model reproduced rhythmic firing, as well as synchronized oscillation, when the coupling conductance was high. Figure 3(c) plots the largest Lyapunov exponent of the network dynamics as a function of the coupling conductance. The small Lyapunov exponent at large conductance indicates that the firing is synchronized and rhythmic when electrical coupling is strong, corresponding to picrotoxin application. With small conductance, each neuron periodically fires at its own frequency without much interaction; thus, the Lyapunov exponent is again close to zero, indicating an almost periodic solution. However, with medium conductance, the Lyapunov exponent is much higher than zero, indicating that neurons fire asynchronously and aperiodically. This is actually chaotic and desynchronized firing, and for this range of dynamics, input and output information transmission becomes the largest, as shown in Figure 3(d). These theoretical predictions were recently supported experimentally. Consequently, climbing fiber inputs can convey error signals most efficiently, even with their low firing frequency (typically 1 spike s^{-1}), when the conductance is medium and chaos is induced. Low-frequency firing usually implies poor representation of the error signal. Chaotic and desynchronized firing of inferior olive neurons might be a basic neural mechanism for insuring high-fidelity error-signal transmission even with a low firing rate.

The feedback error learning theory of the cerebellum explained in the next section theoretically requires low firing rates of climbing fiber inputs. This is because climbing fiber inputs convey time-delayed feedback motor commands, and if they were to be strongly manifested in Purkinje cell output, smooth and sophisticated feed-forward motor control would be ruined. Hirano and his colleagues provided experimental support for this prediction in AMPA-R delta 2 knockout mouse. Climbing fiber inputs occupy a much larger proportion of Purkinje cell output in this knockout mouse compared with a normal mouse, and optokinetic eye movements are much more delayed than is control performance.

Feedback Error Learning Model

Although early learning models proposed that the cerebellum conducts pattern recognition, probably influenced by a 'perceptron,' they are not consistent with physiological data showing that simple spike firing rates of Purkinje cells temporally change and encode dynamic and kinematic features of movements. Recent studies suggest that the human cerebellum is important for sensory and cognitive functions as well as for motor control. What general computation might be realized by the cerebellar cortex that would cover the sensory, cognitive, and motor domains; would be compatible with existing hypotheses, such as timing, coordination, and error correction; and could be learned by supervised learning such as that shown in Figure 1? Internal models seem to provide the most versatile and compatible computational entity satisfying all these requirements. Internal models are neural networks inside the brain ('internal') that can mimic ('model') the input–output characteristics of some dynamic process outside the brain. In a motor control context, a forward model of a body can predict sensory consequences of movement from the efference copy of a motor command. This is because the controlled objects in movements generate a trajectory from motor commands. By contrast, the inverse model of a body and the external world can compute the necessary motor commands from a desired movement pattern. Internal models are useful for predictive computations in sensory, cognitive, and motor functions. If an inverse dynamics model for a controlled object possessing dynamics and kinematics with multiple degrees of freedom is destroyed, the movement becomes clumsy, slow, and not well corrected, because neither feed-forward nor feedback control (both of which are dependent on the inverse model) is available. Thus, control must rely on poor and crude feedback control alone. Internal models can be learned based on supervised learning.

Around 1990, Kawato proposed that different parts of the cerebellum contain either forward or inverse models. He also proposed a computational scheme called feedback error learning to acquire an inverse model in the supervised learning scheme, as shown in Figure 4(a). If supervised learning takes place in the brain for motor control, the following difficult computational problem called 'distal teacher,' formulated by Jordan, should be resolved. As shown in Figure 1, the error between desired and actual movement patterns can be measured by sensory organs, but these movement errors could be very different in spatial coordinates as well as in temporal dynamics from the necessary error signal for the motor command. Furthermore, it should be emphasized that there exists no teaching signal for motor commands in supervised motor learning, because if it were to exist, it would be directly utilized for motor control, and there would be no need for learning. The distal teacher problem is

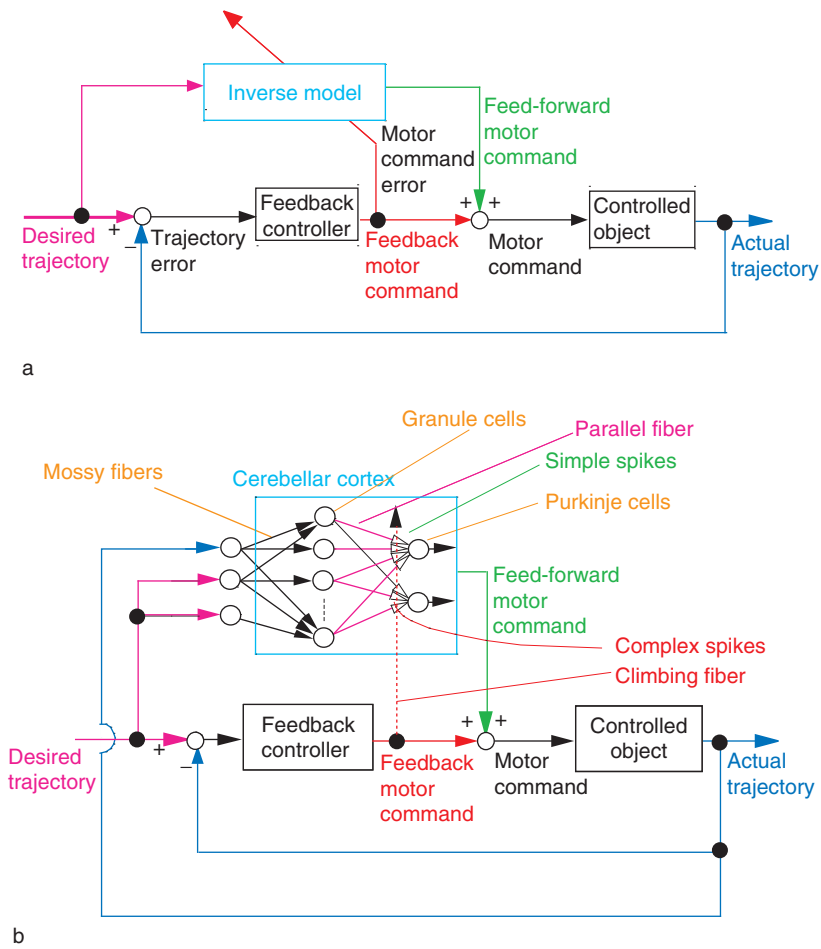

Figure 4 (a) The general feedback error learning model. (b) The cerebellar feedback error learning model. The 'controlled object' is a physical entity, such as the eyes, hands, legs, or torso, that needs to be controlled by the central nervous system. Different colors are for correspondence between (a) and (b).

almost absent in simple gain control with one degree of freedom, such as vestibulo-ocular reflex; thus, it motivated the following more recent models for general many-degrees-of-freedom movements with dynamics and kinematics. Several computational models have been proposed, such as direct inverse modeling, back-propagation through a forward model, or decorrelation learning, but the feedback error learning model, which transforms the sensory error into a motor error by a preexisting feedback controller in the central nervous system (Figure 4(a)), is probably the most biologically plausible and best supported.

The feedback error learning model depicted in Figure 4(a) is the same as the basic model shown in Figure 1 except for the following four points. First, the inverse model transforms a desired trajectory into a feed-forward motor command. Second, a crude feedback controller generates feedback motor command from the sensory error. Third, the summation of the feed-forward and feedback motor commands is sent to the controlled object. Finally, and most important, the feedback motor command is used as the error signal for motor commands in supervised learning for acquisition of the inverse model (cf. Figures 1 and 4(a)). Stability and convergence of the feedback error learning algorithm have been mathematically proved as well as successfully applied to several robotic demonstrations. Figure 4(b) adapts the basic feedback error learning model to the cerebellar neural circuit. Here, simple spikes represent feed-forward motor commands, and parallel fiber inputs represent the desired trajectory as well as the sensory feedback of the current status of the controlled object. A microzone of the cerebellar cortex constitutes (a part of) an inverse model of a specific controlled object such as the eye or arm. Most importantly, climbing fiber inputs are assumed to carry a copy of the feedback motor commands generated by the feedback control circuit. Thus, the complex spikes are predicted to be trajectory error signals already expressed in motor command coordinates. Although

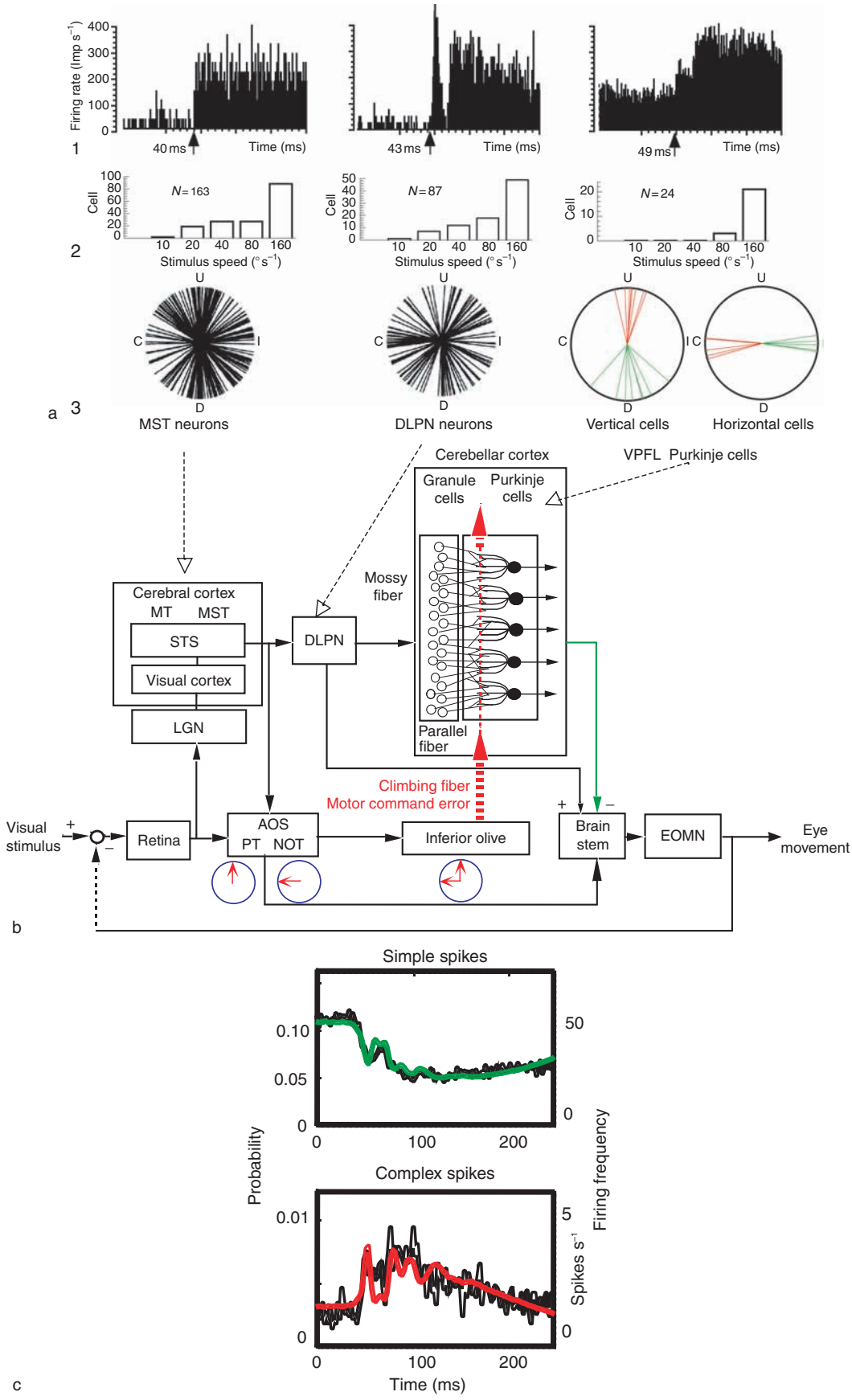

Figure 5 Continued

climbing fiber inputs are delayed by 100 ms with respect to the responsible parallel fiber inputs that caused the movement and the resulting error, spike timing dependency of LTD (see the section titled 'Computational models of cerebellar LTD') selectively changes only the responsible parallel fiber synapses.

The Ventral Paraflocculus and Ocular Following Responses

Kawano and his colleagues supported the cerebellar feedback error learning model by neurophysiological studies in the ventral paraflocculus of monkey cerebellum during ocular following responses. Ocular following responses are tracking movements of the eyes, evoked by movements of a large visual scene, that are thought to be important for the visual stabilization of gaze. In the neural circuit controlling this response Figure 5(b), the phylogenetically older, crude feedback circuit comprises the retina, the accessory optic system, and the brain stem. The phylogenetically newer and more sophisticated feed-forward pathway and the inverse dynamics model correspond to the cerebral and cerebellar cortical pathway and the cerebellum, respectively. In Figure 5(b), the sensory error signal is computed as the image motion on the retina (retinal slip). This is different from the schemes of Figures 1 and 4 in which some neural mechanisms explicitly compute the difference (error) between the desired and actual states. However, in the case of eye movement control such as ocular following responses or vestibulo-ocular reflexes, the desired state is that the eye movement is made so that the retinal image does not move, and thus the optical computation (retinal slip detection) can replace subtraction of the desired and actual states. On the other hand, the desired trajectory information transmitted to the flocculus for the vestibulo-ocular reflex is the head motion signal detected by semicircular canals whereas for the ocular following responses, it is the image motion signal conveyed by the visual areas in the cerebral cortex.

During ocular following responses, the time courses of simple spike firing frequency show complicated patterns (Figure 5(c) top, thin line). However, they were quite accurately reconstructed by using an inverse dynamics representation of eye movement (thick line). The model fit was good for the majority of the neurons studied under a wide range of visual stimulus conditions, different eye movements, and different firing patterns. The same inverse dynamics analysis for the medial superior temporal area (MST) of the cerebral cortex and dorsolateral pontine nucleus, which provide visual mossy fiber inputs, revealed that their firing time courses were not well reconstructed from motor commands but rather that they were well reconstructed from visual retinal slip. Taken together, these data suggest that the ventral paraflocculus is the major site of the inverse dynamics model of the eye for ocular following responses.

Motor commands, conveyed by simple spikes, should be directly modified and acquired by LTD while guided by motor command errors, which are conveyed by climbing fiber inputs. For this to work, the climbing fiber inputs need to convey comparable details of temporal and spatial information to the motor commands, but the ultralow discharge rates of the latter would appear to rule out this possibility. This apparently discrete and sporadic nature of climbing fiber inputs was characterized as an unexpected event detector, which suggested a reinforcement learning type theory. However, if thousands of

Figure 5 Change of neural codes and learning of inverse dynamics model in the cerebellum for ocular following responses (OFR). (a) Firing characteristics of medial superior temporal area (MST), dorsolateral pontine nucleus (DLPN), and ventral paraflocculus (VPFL) neurons. (b) Schematic neural circuit for OFR. (c) Temporal firing patterns of VPFL Purkinje cells in upward eye movements induced by upward visual motion. (a) The left, middle, and right columns are for MST, DLPN, and VPFL neurons, respectively. (1) Poststimulus-time histograms of the firing rates of a typical neuron in each of the three areas. The origin of time is the onset of visual stimulus motion. (2) Histograms of a number of cells within a given range of the optimum stimulus speeds. (3) Polar plots of optimum stimulus directions. U, C, D, and I, upward, contralateral, downward, and ipsilateral, respectively. VPFL Purkinje cells were classified into two groups: vertical cells and horizontal cells, based on simple spike (a dotted line) and complex spike (actual line) optimum directions. (b) The circuit can be divided into two main pathways. The upper part shows the corticocortical (the cerebral cortex to the cerebellar cortex) pathway, which corresponds to the feed-forward arc of feedback error learning. The lower part shows the phylogenetically older feedback pathway containing the accessory optic system (AOS), which corresponds to a crude feedback controller in the feedback error learning scheme. (c) Temporal firing patterns of nine accumulated Purkinje cells (thin curves) and their reconstruction based on inverse dynamics model (bold curves). The model predicts that the temporal firing patterns of simple spikes (upper pane) and complex spikes (lower pane) should be mirror images of each other, and this was confirmed experimentally. PT, pretectum; NOT, nucleus of optic tract; MT, middle temporal area; STS, superior temporal sulcus; LGN, lateral geniculate nucleus; EOMN, extraocular motor neurons. Adapted from Kawano K (1999) Ocular tracking: Behavior and neurophysiology. *Current Opinion in Neurobiology* 9: 467–473; Takemura A, Inoue Y, Gomi H, Kawato M, and Kawano K (2001) Change in neuronal firing patterns in the process of motor command generation for the ocular following response. *Journal of Neurophysiology* 86: 1750–1763; Kobayashi Y, Kawano K, Takemura A, et al. (1998) Temporal firing patterns of Purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkeys. II. Complex spikes. *Journal of Neurophysiology* 80: 832–848. Reproduced from Kawato M (1999) Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* 9: 718–727, with permission from Elsevier.

trials are averaged, the firing rates actually conveyed very accurate and reliable time courses of motor command error within a few hundreds of milliseconds (Figure 5(c), bottom). More specifically, complex spike time courses are better reconstructed by retinal slip than by eye movements but in the motor command coordinates rather than in visual coordinates, as detailed below. Because LTD is a rather slow process, taking several tens of minutes, it can conduct the averaging over many trials. Consequently, the firing probability of climbing fiber inputs can convey high-frequency temporal information matching information of the dynamic command signal.

The preferred directions of MST and pontine neurons were evenly distributed over 360°, as shown in Figure 5(a)(3). Thus, the visual coordinates are uniformly distributed over all possible directions. On the other hand, projections of the three-dimensional spatial coordinates of the extraocular muscles were in either a horizontal or vertical direction, and they were entirely different from the visual coordinates. Preferred directions of simple spikes were either downward or ipsiversive (Figure 5(a)(3)), and at the site of each recording, electrical stimulation of a Purkinje cell elicited eye movement toward the preferred direction of the simple spike. These data indicate that the simple spike coordinate framework is already that of the motor commands. Thus, at the parallel fiber–Purkinje cell synapse, a drastic visuomotor coordinate transformation occurs. The neural representation dramatically changes from population coding in the MST and pontine nucleus to firing rate coding of Purkinje cells at the parallel fiber–Purkinje cell synapse (Figure 5(a)). What, then, is the origin of this drastic transformation? The theory proposes that the complex spikes, and ultimately the accessory optic system, are the source of this motor command spatial framework. The preferred directions of pretectum neurons are upward, and those of the contralateral nucleus of optic tract neurons are ipsiversive (thus contraversive for the cerebellum). Their signals are conveyed to the inferior olive and further to the cerebellum to produce the complex spikes. In summary, the coordinate frame of the climbing fibers was vertical and horizontal, forming the motor coordinates, not the visual coordinates of MST or pontine nucleus. All these experimental findings were successfully simulated by the network model shown in Figure 5(b). The coordinates of semicircular canals and the extraocular muscles are very similar, and it is rather difficult to evaluate coordinate transformation in the vestibulo-ocular reflex and flocculus compared with ocular following responses. However, recent physiological studies indicate that climbing fiber inputs encode motor error rather than sensory error. This argues against a class of model that assumes sensory–motor transformation downstream of Purkinje cells, including decorrelation models.

Toward a General Model of the Cerebellum

We have a very solid computational model (feedback error learning and inverse model) and a rich set of supporting experimental data for the ventral paraflocculus with ocular following responses. The model has also been well supported for other slow eye movements, such as vestibulo-ocular reflex, optokinetic eye movements, and smooth pursuit, for the flocculus and paraflocculus. Can we extend this to other parts of the cerebellum, other types of movements, and other computational models? Many theorists believe so since the neural circuit is generally uniform except for the newly found unipolar brush cells, found primarily in the flocculonodular lobe. Furthermore, Kitazawa and colleagues have demonstrated that complex spikes recorded from cerebellar hemispheres encode reaching errors at goal points during monkey whole-arm reaching, whereas simple spikes represent movement dynamics such as an electromyogram.

What can we generalize about cerebellar functions from previous studies on specific functions and specific areas? First, synaptic plasticity is unique and ubiquitous, and the learning is undoubtedly supervised learning and not Hebbian or reinforcement learning. Second, the cerebellar 100 ms STDP, in contrast to cerebral and hippocampal millisecond-order STDP, strongly suggests that the cerebellum is better suited for approximating some nonlinear dynamical system outside the cerebellum as 100 ms is exactly the time required for sensory signal feedback from the external world. Finally, the major computation is feed-forward, with less emphasis on recurrent feedback, because spatial resources are allocated more to neurons than to their connections in the cerebellum, as is evident from the number of neurons: more than that of the cerebrum in a volume only one-tenth that of the cerebrum. Consequently, internal models are the most promising computational functions of the cerebellum. Forward models have been theoretically postulated several times and supported by neuroimaging experiments. Cancellation of a self-generated electrical field by a forward model is the function of the electrosensory lobe (cerebellum-like structure) of the weakly electric fish and might be interpreted as another support. But it is actually a different story because NMDA receptors play the essential role in anti-Hebbian learning there. Thus, supporting evidence

for forward models is much more circumstantial than is the evidence for inverse models, and we need more-decisive data. Computational models for how multiple internal models are integrated and used for motor control and cognition have been proposed (e.g., the modular selection and identification for control model), but innovative experimental methodology to further explore these phylogenetically newer functions of the cerebellum seems necessary.

See also: Cerebellar Microcircuitry; Cerebellar Deep Nuclei; Cerebellum and Oculomotor Control; Cerebellum: Clinical Pathology; Cerebellum: Evolution and Comparative Anatomy; Olivocerebellar System; Procedural Learning: Cerebellum Models; Sporadic Degenerative Ataxias and the Dominantly Inherited Spinocerebellar Ataxias.

Further Reading

Albus JS (1971) A theory of cerebellar function. *Mathematical Biosciences* 10: 25–61.

Daniel H, Levenes C, and Crepel F (1998) Cellular mechanisms of cerebellar LTD. *Trends in Neurosciences* 21: 401–407.

Dean P, Porrill J, and Stone JV (2002) Decorrelation control by the cerebellum achieves oculomotor plant compensation in simulated vestibulo-ocular reflex. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 269: 1895–1904.

De Schutter E (1995) Cerebellar long-term depression might normalize excitation of Purkinje cells: A hypothesis. *Trends in Neurosciences* 18: 291–295.

De Schutter E (1999) Using realistic models to study synaptic integration in cerebellar Purkinje cells. *Reviews in the Neurosciences* 10: 233–245.

Doi T, Kuroda S, Michikawa T, and Kawato M (2005) Inositol 1,4,5-trisphosphate-dependent Ca^{2+} threshold dynamics detect spike timing in cerebellar Purkinje cells. *Journal of Neuroscience* 25: 950–961.

Fujita M (1982) Adaptive filter model of the cerebellum. *Biological Cybernetics* 45: 195–206.

Gomi H, Shidara M, Takemura A, et al. (1998) Temporal firing patterns of Purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkeys. I. Simple spikes. *Journal of Neurophysiology* 80: 818–831.

Imamizu H, Miyauchi S, Tamada T, et al. (2000) Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature* 403: 192–195.

Ito M (1970) Neurophysiological basis of the cerebellar motor control system. *International Journal of Neurology* 7: 162–176.

Ito M (1984) *The Cerebellum and Neural Control*. New York: Raven Press.

Ito M (2006) Cerebellar circuitry as a neuronal machine. *Progress in Neurobiology* 78: 272–303.

Ivry RB and Spencer RM (2004) The neural representation of time. *Current Opinion in Neurobiology* 14: 225–232.

Jordan MI and Rumelhart DE (1992) Forward models: Supervised learning with a distal teacher. *Cognitive Science* 16: 307–354.

Kawano K (1999) Ocular tracking: Behavior and neurophysiology. *Current Opinion in Neurobiology* 9: 467–473.

Kawato M (1999) Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* 9: 718–727.

Kawato M, Furukawa K, and Suzuki R (1987) A hierarchical neural-network model for control and learning of voluntary movement. *Biological Cybernetics* 57: 169–185.

Kawato M and Gomi H (1992) The cerebellum and VOR/OKR learning models. *Trends in Neurosciences* 15: 445–453.

Keating JG and Thach WT (1995) Nonclock behavior of inferior olive neurons: Interspike interval of Purkinje cell complex spike discharge in the awake behaving monkey is random. *Journal of Neurophysiology* 73: 1329–1340.

Kitazawa S and Wolpert DM (2005) Rhythmicity, randomness and synchrony in climbing fiber signals. *Trends in Neurosciences* 28: 611–619.

Kitazawa S, Kimura T, and Yin PB (1998) Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature* 392: 494–497.

Kobayashi Y, Kawano K, Takemura A, et al. (1998) Temporal firing patterns of Purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkeys. II. Complex spikes. *Journal of Neurophysiology* 80: 832–848.

Kuroda S, Schweighofer N, and Kawato M (2001) Exploration of signal transduction pathways in cerebellar long-term depression by kinetic simulation. *Journal of Neuroscience* 21: 5693–5702.

Lang EJ, Sugihara I, and Llinas R (1996) GABAergic modulation of complex spike activity by the cerebellar nucleoolivary pathway in rat. *Journal of Neurophysiology* 76: 255–275.

Llinas R, Lang EJ, and Welsh JP (1997) The cerebellum, LTD, and memory: Alternative views. *Learning & Memory* 3: 445–455.

Llinas R and Yarom Y (1986) Oscillatory properties of guinea-pig inferior olivary neurones and their pharmacological modulation: An *in vitro* study. *Journal of Physiology* 376: 163–182.

Marr D (1969) A theory of cerebellar cortex. *Journal of Physiology (London)* 202: 437–470.

Miall RC, Weir DJ, Wolpert DM, and Stein JF (1993) Is the cerebellum a Smith Predictor? *Journal of Motor Behavior* 25: 203–216.

Ogasawara H, Doi T, Doya K, and Kawato M (2007) Nitric oxide regulates input specificity of long-term depression and context dependence of cerebellar learning. *PLoS Computational Biology* 3(1): e179 doi:10.1371/journal.pcbi.0020179.

Schweighofer N, Doya K, Fukai H, Chiron JV, Furukawa T, and Kawato M (2004) Chaos may enhance information transmission in the inferior olive. *Proceedings of the National Academy of Sciences of the United States of America* 101: 4655–4660.

Shidara M, Kawano K, Gomi H, and Kawato M (1993) Inverse-dynamics model eye movement control by Purkinje cells in the cerebellum. *Nature* 365: 50–52.

Shutoh F, Ohki M, Kitazawa H, Itohara S, and Nagao S (2006) Memory trace of motor learning shifts transsynaptically from cerebellar cortex to nuclei for consolidation. *Neuroscience* 139: 767–777.

Takemura A, Inoue Y, Gomi H, Kawato M, and Kawano K (2001) Change in neuronal firing patterns in the process of motor command generation for the ocular following response. *Journal of Neurophysiology* 86: 1750–1763.

Yamamoto K, Kobayashi Y, Takemura A, Kawano K, and Kawato M (2002) Computational studies on acquisition and adaptation of ocular following responses based on cerebellar synaptic plasticity. *Journal of Neurophysiology* 87: 1554–1571.

Cerebral Cortex

E G Jones, University of California, Davis, CA, USA

© 2009 Published by Elsevier Ltd.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

The mammalian cerebral cortex is an enormous sheet of cells covering in the human brain approximately 250 000 mm^2 and containing 1×10^9 or more nerve cells in a volume of about 300 cm^3. The human cerebral cortex is by no means the largest – it is exceeded in area by that of the whales – but it is probably the most highly differentiated. Differentiation, in this sense, means the subdivision of the cortex into the largest number of histologically distinct subareas each of which has a known or suspected function.

The fundamental divisions of the cerebral cortex in mammals are the hippocampal formation (or archicortex), the olfactory cortex and associated areas such as the entorhinal and periamygdaloid area in the piriform lobule (paleocortex), and the remainder (neocortex). The three parenthetical names indicate the phylogenetic assumptions of past generations of neuroscientists who regarded the neocortex as the hallmark of mammalian evolution. In nonmammalian brains, a hippocampal formation and an olfactory-related cortex can readily be identified, whereas a neocortex is probably absent or if present, is extremely small. The remainder of this article deals only with mammals and only with the neocortex.

The neocortex develops on the surface of the cerebral hemispheres as a series of waves of newly generated cells many of which migrate radially outward from the proliferative ventricular and subventricular zones lining the primitive lateral ventricles while others invade by migration from the region of the developing basal ganglia. The earliest arriving cells form a primordial cortex which is invaded by successive waves of other cells. Each successive wave pushes through its predecessors and comes to lie external to them. In this is the beginnings of one of the overt histological features of the neocortex: lamination. Cells in a particular lamina are born during the same limited time window and aggregate together in the cortex. In all but the most superficial layer of the cortex of a fully developed mammal there are two major neuronal classes (and, of course, neuroglial cells). The two neuronal types are called pyramidal and nonpyramidal. Among the latter are several subtypes. The names of the cells reflect their morphology as seen with stains such as that of Golgi which reveal the morphology of the whole cell. The lamination of the cortex is, at the routine histological level seen with the Nissl stain (which stains only cell somata), a lamination of cell bodies. Stains such as that of Golgi show that processes of the cells can extend across laminae.

In most parts of the mammalian cortex, a six-layered pattern first introduced by the anatomist Brodmann and based upon Nissl staining, has come to be accepted as the standard, though five- and seven-layered patterns have also been in vogue. Layer I is the layer immediately beneath the pia mater and contains very few cell bodies, belonging mainly to neuroglial cells and to a few nonpyramidal neurons – hence its name, 'molecular layer'. Layer II consists of small, tightly packed cell bodies, most of which are small pyramidal neurons, but nevertheless is called the 'external granular layer'. Layer III is a thick layer in which pyramidal cell somata increase progressively in size from superficial to deep. It is referred to as the pyramidal or 'external pyramidal layer'. Layer IV is thin but densely packed with the somata of small pyramidal and nonpyramidal cells. It is called the 'internal granular layer'. Layer V consists primarily of large, loosely dispersed pyramidal cell somata. It is referred to as the ganglionic or 'internal pyramidal layer'. Layer VI is composed of relatively tightly packed, small, round cell somata that belong mostly to a modified form of pyramidal cell. It is called the 'multiform' or 'polymorph layer'. In certain areas, some of these layers can be reduplicated or further subdivided.

The features of the Nissl-stained laminar pattern are superficial. Each layer defined in the terms given, apart from layer I, contains pyramidal and nonpyramidal cells of varying types. All these cells have dendrites and axons that can extend into or through other layers. The majority of the pyramidal cells-with somata in any layer send their apical dendrites through all the supervening layers to end, commonly in a tuft of branched dendrites, in layer I. When stained with the Golgi method, therefore, the lamination of the cortex is far less clear than in a Nissl-stained preparation.

The cerebral cortex is also laminated in terms of the distribution of afferent axons and of some of its intrinsic axonal systems. Traditionally, these layers have been described from myelin-stained preparations, but they can also be discerned in well-stained Golgi and reduced silver preparations which stain axons and dendrites. In these preparations, a tangential fiber plexus is revealed in layer I which is consequently called the plexiform layer. A condensation of fibers in the deepest part of layer I is called the stria of Kaes or Bechterew (Figure 1). Horizontal plexuses of fibers in layer IV and deep in layer V are called

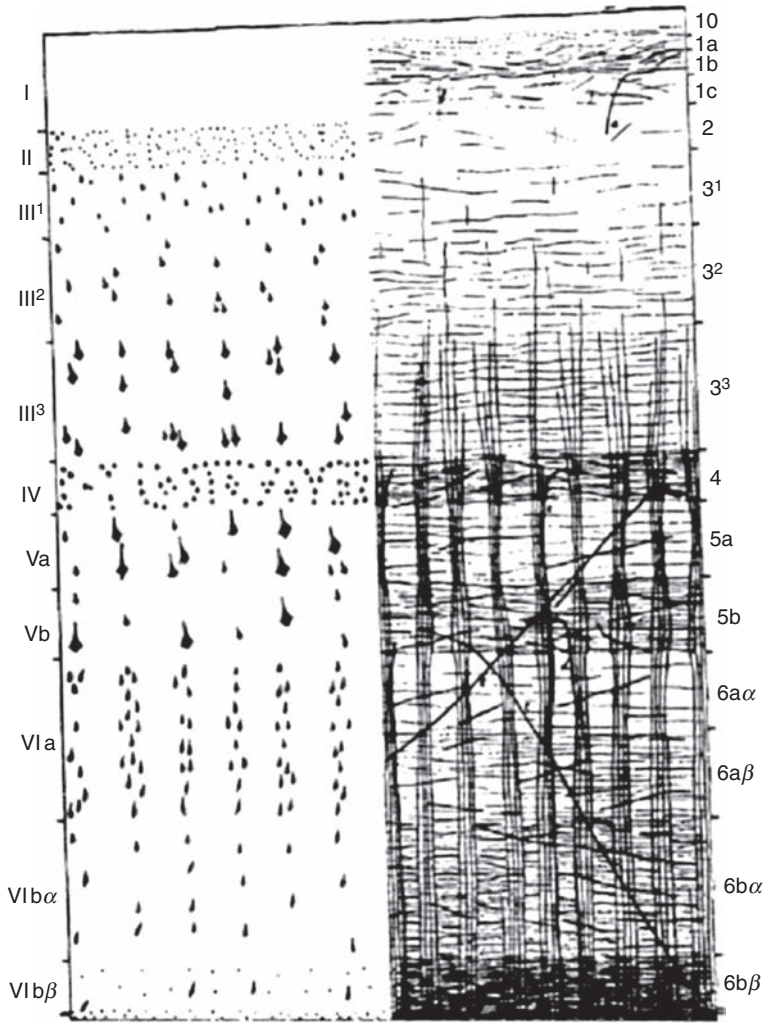

Figure 1 Vogt's scheme of the fundamental plan of cellular (left) and fiber (right) layering in the cerebral cortex. The cellular layers (I through VI) are those given in the text and Vogt divides them into sublayers. Fiber layers 4 and 5b are the bands of Baillarger; the stria of KaesBechterew, when present, is in layer 1c. From Vogt C and Vogt O (1919) Allgemeinere Ergebnisse unserer Hirnforschung. *Journal of Psychology and Neurology* 25: 279–462.

the inner and outer bands of Baillarger. These are probably formed primarily by intrinsic axons of the cortex (i.e., axons of some nonpyramidal cells and collateral branches of axons of pyramidal cells). The outer band of Baillarger is traditionally thought to be formed by the terminal ramifications of thalamocortical axons, but this has been disputed in the visual cortex where the band is referred to as the stria of Gennari and consists mainly of axons arising from intrinsic cortical sources.

Vertical bundles of myelinated axons extend through layers III to VI of the cortex, becoming thicker as they descend and as more axons are added. The bundles are referred to as radial fasciculi and contain axons of pyramidal neurons as they proceed to exit the cortex, and the axons of certain nonpyramidal neurons. Afferent axons coming into the cortex can usually be identified in myelin preparations as thick, obliquely running fibers separate from the radial fasciculi (**Figure 1**).

The laminar patterns seen in Nissl- and myelin-stained preparations vary from mammal to mammal and across the surface of the cortex of an individual mammalian species. The art of identifying and determining the boundaries of different cortical areas is known as cytoarchitectonics or myeloarchitectonics, depending on whether Nissl or myelin stains are used as the basis for the delineations made. Cortical areas, many with known functional characteristics (visual, auditory, motor, etc.) have cyto- and myeloarchitectonic features and common differences in histochemical staining as well, that enable their borders to be defined and a distinction to be made between them and adjacent areas. The basic types of neocortex are:

'homotypical', found in the frontal, temporal, and parietal lobes; 'heterotypical', divided further into granular and agranular cortex. 'Granular' or 'koniocortex' is found in the sensory areas and in it there is a predominance of small cell bodies in layers II–IV, resulting in a blurring of the borders between the layers; in 'agranular cortex', found in the premotor and motor areas, layer IV disappears during development, and layers III and V form a continuum of increasing pyramidal cell somal size. In conjunction with the changes in the Nissl staining pattern, there are changes in the myelin staining pattern. The homotypical cortex shows all the laminated plexuses, in granular cortex the bands of Baillarger predominate, and in agranular cortex thick radial fasciculi predominate.

With one or two exceptions, mainly in the visual cortex, the pyramidal cells are the output cells of the cerebral cortex, sending their axons to subcortical targets and to other cortical areas on the same and opposite sides of the brain (Figure 2). To a large extent the pyramidal cells with somata in a particular cortical layer send their axons to a defined set of targets, and the laminar localization of cells projecting to the same target(s) does not vary greatly from area to area. The modified pyramidal cells of layer VI send their axons to the principal thalamic nucleus providing the input to the cortical area in which they lie, to the claustrum, and to cortical areas linked functionally with their own. Layer V pyramidal cells provide connections to all other subcortical structures: corticospinal, corticobulbar, corticopontine, corticorubral, corticotectal, corticostriatal, and corticothalamic fibers to nonspecific thalamic nuclei. There are conflicting reports about the extent to which axons to one site arise as branches of those directed to another and there may be species differences in the degree to which this occurs. Pyramidal cells of layers II and III provide most of the axons of the corpus callosum and most of those forming ipsilateral corticocortical connections, though some callosal and ipsilateral corticocortical axons can arise from pyramidal cells with somata in other layers as well. The precise constellation of cortical and subcortical connections emanating from a cortical area varies, with some types of connection and thus the parent cell category being absent from some areas. Corticorubral fibers, for example, arise only in the motor and premotor areas. Others, such as the corticostriatal and corticopontine, probably arise in all areas. It is probably the variation in pyramidal cell size and shape related to these differential connections that helps determine the architectonic characteristics of the individual cortical areas.

Afferent axons entering a cortical area arise from a variety of sites. Every cortical area receives specific afferent fibers from a particular thalamic relay nucleus or nuclei and from a particular set of ipsilateral and contralateral cortical areas. These three sets of fibers release the excitatory amino acid transmitter glutamate. All the inputs to the sensory areas, at least, are highly ordered, topographically organized sets of connections that form an essential element in the mapping of the external world and surface of the

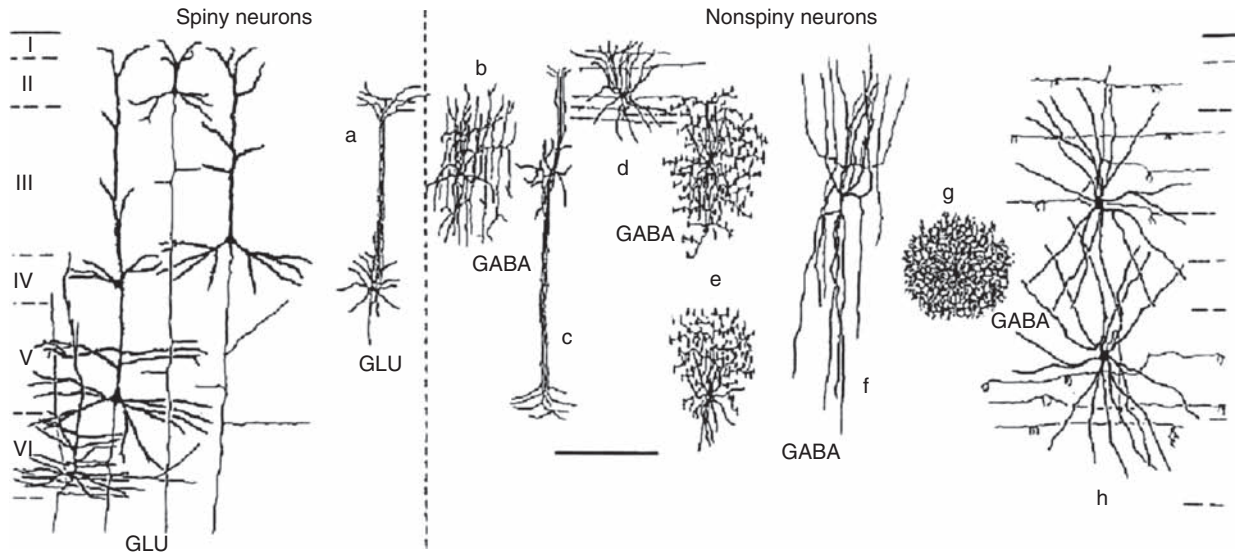

Figure 2 Basic cell types in the monkey cerebral cortex. (Left) Spiny neurons which include pyramidal cells and spiny stellate cells (A), a form of nonpyramidal neuron. Spiny neurons utilize the neurotransmitter, glutamate (Glu). (Right) Nonpyramidal cells all of which use the neurotransmitter, GABA. B, cell with local axon arcades; C, double bouquet cell; D, H, basket cells; E, chandelier cells; F, bi-tufted, usually peptide-containing cell; G, neurogliaform cell. Modified from Jones EG (1981) Anatomy of cerebral cortex: Columnar input-output organization. In: Schmitt FO, Worden FG, and Dennis SG (eds.) *The Cerebral Cortex,* pp. 199–235. Cambridge, MA: MIT Press.

body on to the cortex. Other, more diffusely organized afferent fibers whose cortical distribution does not respect cytoarchitectonic boundaries arise in so-called nonspecific nuclei of the thalamus, in the noradrenergic cells of the locus coeruleus, serotoninergic cells of the midbrain, cholinergic cells of the basal forebrain, and, for many cortical areas, in the dopaminergic cells of the midbrain. It is customarily assumed that some or all of these nonspecific afferent systems are concerned with controlling levels of cortical excitability, arousal and the conscious state. The specific and nonspecific afferent fiber systems terminate in particular cortical layers that tend to vary from area to area and from species to species. In general, specific thalamic fibers end predominantly in the middle layers (III and IV) and corticocortical fibers in layers I–IV. Many nonspecific systems end predominantly in layers I and/or VI.

The three specific afferent fiber systems terminate on both pyramidal and nonpyramidal cells. The further, internal, elaborative circuitry of the cortex is then effected by the axons of nonpyramidal cells and by the intracortical collaterals of pyramidal cell axons. Depending on their morphological types the nonpyramidal cells' axons have rather precise targets, mainly on different levels of the soma-dendritic membranes of other cells. The collaterals of pyramidal cell axons tend to be preferentially distributed to certain layers depending on the target of the principal axon of the pyramidal cell. Most output cells, seem to be affected monosynaptically by specific afferent fibers and polysynaptically by these fibers after relays through one or more other cortical neurons.

Nonpyramidal neurons belong to several morphological categories, each of which has been given a name that usually reflects some aspect of its dendritic or axonal arborization pattern (**Figure 2**). The distributions of their axons are the best determinants of cell type and determine in turn the range of action over which these intrinsic cortical interneurons act. Those with local axons probably serve to modulate local synaptic actions between afferent fibers and efferent cells; those with vertical axons effect interlaminar connections (ascending and descending) and are probably the main determinants of the columnar organization of the cortex such as is manifested in the visual cortex. Those with horizontal axons effect intralaminar and intercolumnar actions. While the division of cortical neurons into pyramidal and nonpyramidal types reflects their division into extrinsically projecting and intrinsic neurons, a further division of equal functional importance is into those which possess substantial populations of dendritic spines and those that are essentially nonspiny. This reflects a division into neurons that produce the excitatory amino acid transmitter glutamate, and those that produce the inhibitory neurotransmitter, γ-aminobutyric acid (GABA). All pyramidal cells are spiny and glutamatergic. Long horizontal collaterals of the axons of layers III and V pyramidal cells are the main sources of horizontally spreading excitation in the cortex. Nonpyramidal cells are divided into spiny and non-spiny types. Spiny nonpyramidal cells (spiny stellate cells) dominate layer IV, are the major recipients of thalamic synapses, are glutamatergic, and their axons provide the main route for information to flow out of layer IV into other layers. The majority of nonpyramidal neurons lack significant populations of dendritic spines, are found in all layers, and use GABA. They account for up to 30% of the total population of cortical neurons. They can be further divided into subpopulations based upon their production of one or other of three calcium-binding proteins: 28KD calbindin, 29KD calretinin, or parvalbumin. The chemical profile of each type of cortical interneuron tends to be associated with the possession of a particular set of membrane conductances that confers different discharge properties upon the cells. A small subpopulation also contains one or more neuropeptides. The combination of inhibitory and excitatory effects mediated by the various types of cortical cell are thought to be critical in shaping the receptive fields of neurons recorded in different layers of the cortex, in maintaining columnar organization in the cortex, and in setting up afferent inhibition between columns.

See also: Brodmann's Areas; Cerebral Cortex: Inhibitory Cells; Hippocampus; Memory Consolidation: Cerebral Cortex; Neocortex: Origins; Olfactory Cortex: Comparative Anatomy; Olfactory Cortex Physiology.

Further Reading

Anderson JC, Martin KAC, and Whitteridge D (1993) Form, function, and intracortical projections of neurons in the striate cortex of the monkey *Macacus nemestrinus*. *Cerebral Cortex* 3: 412–420.

Anderson SA, Kaznowski CE, Horn C, Rubenstein JLR, and McConnell SK (2002) Distinct origins of neocortical projection neurons and interneurons *in vivo*. *Cerebral Cortex* 12: 702–709.

Gilbert CD (1983) Microcircuitry of the visual cortex. *Annual Review of Neuroscience* 6: 217–248.

Gilbert CD (1993) Circuitry, architecture, and functional dynamics of visual cortex. *Cerebral Cortex* 3: 373–386.

Gupta A, Wang Y, and Markram H (2000) Organizing principles for a diversity of GABAergic interneurons and synapses in the neocortex. *Science* 287: 273–278.

Hendry SHC, Schwark HD, Jones EG, and Yan J (1987) Numbers and proportions of GABA-immunoreactive neurons in different areas of monkey cerebral cortex. *Journal of Neuroscience* 7: 1503–1519.

Hendry SHC, Jones EG, Emson PC, Lawson DEM, Heizmann CW, and Streit P (1989) Two classes of cortical GABA neurons

defined by differential calcium binding protein immunoreactivities. *Experimental Brain Research* 76: 467–472.

Jones EG (1981) Anatomy of cerebral cortex: Columnar input-output organization. In: Schmitt FO, Worden FG, and Dennis SG (eds.) *The Cerebral Cortex*, pp. 199–235. Cambridge, MA: MIT Press.

Jones EG (1984) The columnar basis of cortical circuitry. In: Rosenberg RN (ed.) *The Clinical Neurosciences,* Vol. 5. *Neurobiology*, pp. 357–385. London: Churchill Livingstone.

Jones EG (1993) GABAergic neurons and their role in cortical plasticity in primates. *Cerebral Cortex* 3: 361–372.

Kawaguchi Y and Kubota Y (1993) Correlation of physiological subgroupings of nonpyramidal cells with parvalbumin- and calbindinD28k-immunoreactive neurons in layer V of rat frontal cortex. *Journal of Neurophysiology* 70: 387–396.

Kawaguchi Y and Kubota Y (1997) GABAergic cell subtypes and their synaptic connections in rat frontal cortex. *Cerebral Cortex* 7: 476–486.

Kawaguchi Y (2001) Distinct firing patterns of neuronal subtypes in cortical synchronized activities. *Journal of Neuroscience* 21: 7261–7272.

Kisvarday ZF, Cowey A, Smith AD, and Somogyi P (1989) Interlaminar and lateral excitatory amino acid connections in the striate cortex of monkey. *Journal of Neuroscience* 9: 667–682.

Lund JA (1990) Excitatory and inhibitory circuitry and laminar mapping strategies in the primary visual cortex of the monkey. In: Edelman GM, Gall WE, and Cowan WM (eds.) *Signal and Sense: Local and Global Order in Perceptual Maps*, pp. 51–82. New York: Wiley-Liss.

Peters A and Jones EG (eds.) (1984–1994) *Cerebral Cortex,* vols. 1–10. New York: Plenum.

Vogt C and Vogt O (1919) Allgemeinere Ergebnisse unserer Hirnforschung. *Journal of Psychology and Neurology* 25: 279–462.

Cerebral Cortex: Inhibitory Cells

Y Kawaguchi, National Institute for Physiological Sciences, Okazaki, Japan
F Karube, Debrecen University, Debrecen, Hungary

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Different brain regions have specialized neural circuits with characteristic structures that presumably facilitate appropriate information processing to influence behavior. Consistent with its participation in higher cognitive functions, the cerebral cortex is composed of extremely complicated neuronal circuits, the properties of which are still largely unknown. To understand the operation of the neocortex it is necessary to uncover both the structural composition of the circuit (neural circuit analysis) and to reveal the way information is represented and propagated through the circuit (electrical signal analysis). Therefore it is essential to identify, structurally and physiologically, the functional classes of cortical neurons. For convenience, models of the cortical circuit have tended to minimize the number of cortical cell types. However, experimental data demonstrate that cortical neurons are extremely diverse in their morphological and physiological properties.

Morphologies of cortical neurons were systematically described first by Santiago Ramón y Cajal (1852–1934) in Spain. Using newly developed Golgi staining, Cajal observed the details of dendritic and axonal elongations and revealed the morphological composition of cortical neurons. While much subsequent work examined the morphological and physiological properties of the larger pyramidal neurons, our understanding of the less numerous and smaller nonpyramidal cells lagged behind. In comparison with the analysis of information representation on the cortical surface by recording neuron spikes in correlation with the sensory stimulation or behavioral patterns, the structural analysis of local circuitry proceeded slowly. However, recent advances allowing experimental access to individual neurons in the neocortex have facilitated the characterization of nonpyramidal neurons and revealed that these local-circuit neurons are extremely diverse in structure, connectivity, and physiological properties.

Two Basic Neuron Types in Cortex

Cortical neurons are divided into two broad types, pyramidal and nonpyramidal cells. Pyramidal cells comprise ~80% of cortical neurons, and utilize glutamate as a fast excitatory neurotransmitter. They are projection neurons that send axons to extracortical targets as well as to other, sometimes distant, cortical regions. Pyramidal cells have a relatively distinct and uniform morphology, with a single apical dendrite extending toward the pia, basal dendrites extending locally around the soma, and the main axon projecting into the white matter. Most pyramidal cells have small, spinelike protrusions from their dendrites, but the spine density on individual dendrites can be variable among different pyramidal cells. The apical dendrites display variable extension patterns. Physiologically, layer V pyramidal cells can be divided into several groups.

In contrast to pyramidal cells, nonpyramidal cells are extremely heterogeneous in both their dendritic and their axonal arborizations (Figure 1), and contain both excitatory and inhibitory cells. Layer IV in primary sensory cortical regions contains spiny nonpyramidal cells (spiny stellate cells) that utilize glutamate as an excitatory neurotransmitter. They induce excitatory postsynaptic potentials (EPSPs) in other spiny stellate cells in layer IV and in pyramidal cells in layers II/III. Most other nonpyramidal neurons in the neocortex are inhibitory and utilize γ-aminobutyric acid (GABA) as their primary transmitter. Although the specifics of cell classification are still intensely debated, extensive research by many laboratories over the past decade has defined several basic cellular subtypes. These subtypes have been identified by a combination of anatomical, physiological, and chemical characteristics, as well as by the timing and location of their cellular differentiation. In the following section we introduce the organization of GABAergic nonpyramidal cells.

Morphological Diversity of Cortical GABA Cells

The morphological details of individual neurons were initially investigated by Golgi staining, but modern studies have employed intracellular dyes and genetically linked dye expression to stain specific subsets of cortical interneurons. In nonpyramidal cells, dendritic and axonal arborizations can be used for their broad classification (Figure 1), and most classification systems to date have relied on differences in local axonal arborization patterns to describe five main nonpyramidal cell types:

1. Basket cells making multiple synaptic boutons onto the somata of other neurons.

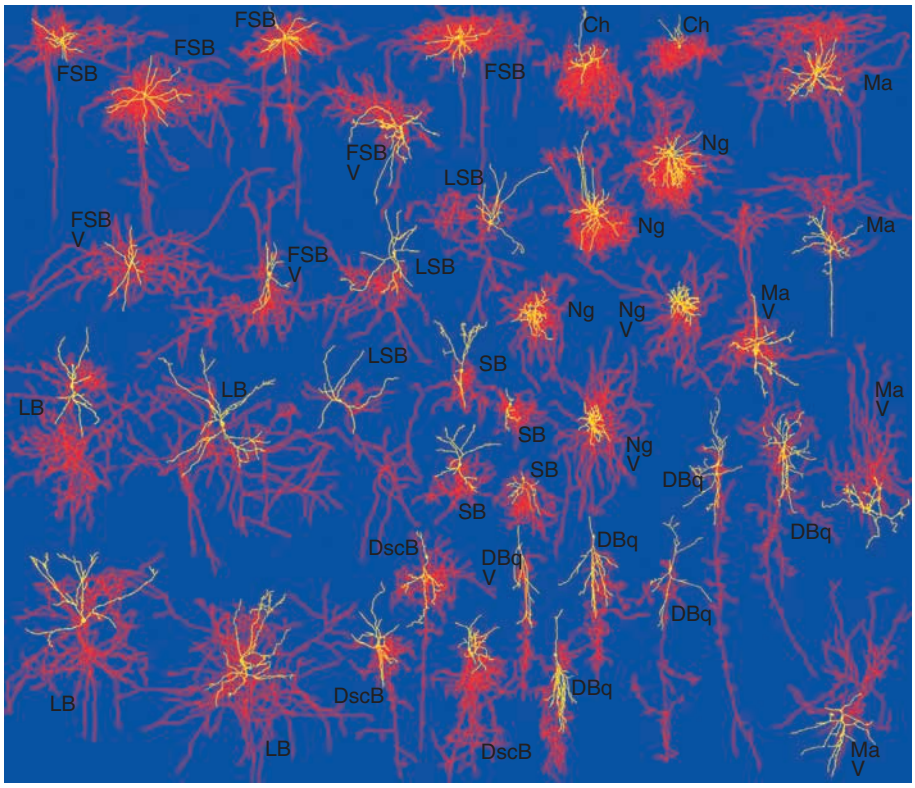

Figure 1 Examples of cortical nonpyramidal cells. Somata and dendrites are depicted in yellow, and axons in red. FSB, fast-spiking basket cell; Ch, fast-spiking chandelier cell; Ng, late-spiking neurogliaform cell; LSB, late-spiking basket cell; LB, non-fast-spiking large basket cell; DscB, non-fast-spiking descending basket cell; SB, non-fast-spiking small basket cell; DBq, double-bouquet cell; Ma, Martinotti cell. Layer V cells are indicated (V); other cells belong to layers II/III. Adapted from Karube F, Kubota Y, and Kawaguchi Y (2004) Axon branching and synaptic bouton phenotypes in GABAergic nonpyramidal cell subtypes. *Journal of Neuroscience* 24: 2853–2865, with permission.

2. Chandelier cells displaying vertical arrays of several boutons, like a chandelier.
3. Martinotti cells extending axon collaterals ascending to layer I.
4. Double-bouquet cells issuing descending axonal bundles.
5. Neurogliaform cells making dense axonal arbors resembling spider webs.

Dendritic patterns of these neurons are heterogeneous and include distinct morphologies, such as a multipolar pattern in which dendrites are projected in all directions from the soma, or a bitufted pattern in which fewer dendrites arise from the soma and tend to project orthogonally toward the pia and white matter. Although GABAergic nonpyramidal cells tend to have a much lower dendritic spine density than do pyramidal cells, the expression of spines in nonpyramidal cells is diverse, ranging from sparsely spiny to completely aspiny. Because of the diversity in these morphological properties, and the lack of correlation between specific dendritic and axonal patterns, morphological appearances alone do not suffice for the identification of individual nonpyramidal subtypes.

Expression Patterns of Specific Chemical Substances

GABA is synthesized by the enzyme glutamic acid decarboxylase (GAD), which has two isoforms identified by their molecular size, 65 and 67 kDa. Immunohistochemistry has revealed that most nonpyramidal cells, except spiny stellate cells in layer IV, express GAD and package the transmitter GABA in their synaptic terminals. Most cortical GABA cells express both isoforms of GAD. The 67 kDa GAD is expressed in both somata and axon terminals, but the 65 kDa one is found mostly in the terminals, more preferentially involved in GABAergic synaptic transmission. Further, nonpyramidal neurons tend to selectively express several peptides that are less common in pyramidal cells. These peptides include somatostatin, cholecystokinin (CCK; also expressed in some pyramidal neurons), vasoactive intestinal polypeptide (VIP), and corticotropin-releasing factor (CRF). Additionally, several types of calcium-binding proteins are also preferentially expressed in nonpyramidal cells: parvalbumin, calretinin, and calbindin. Calbindin is also expressed in some pyramidal cells. While we do not

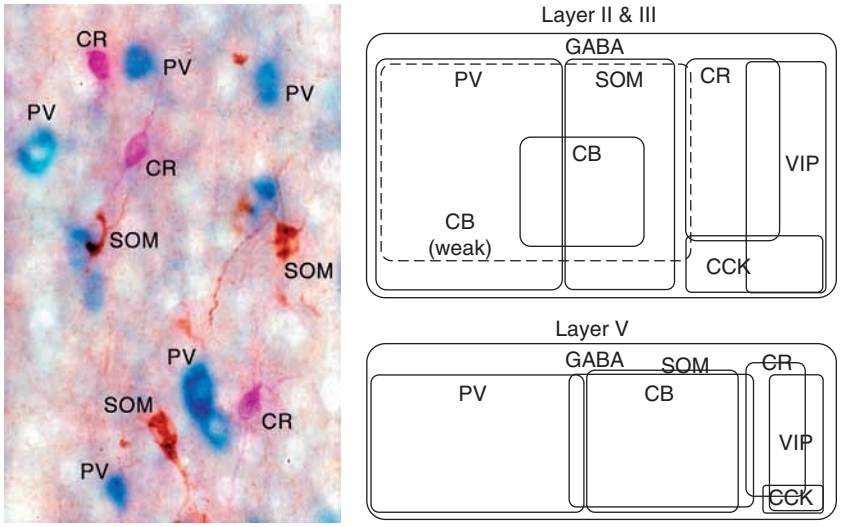

Figure 2 (Left) Three classes of GABA cells immunohistochemically stained with different chromogens in the rat frontal cortex. Somatostatin (SOM) cells are demonstrated using a rat monoclonal antibody (brown); parvalbumin (PV) cells, using a mouse monoclonal antibody (blue); and calretinin (CR) cells, using a rabbit antiserum antibody (red). These three cell classes belong to separate populations of cortical GABA cells. (Right) Schematic showing shared chemical composition of GABA cells (CB, calbindin D28k; CCK, cholecystokinin; CR, calretinin; PV, parvalbumin; SOM, somatostatin; VIP, vasoactive intestinal polypeptide). In addition to pyramidal cells of layers II/III, calbindin D28k is expressed either weakly or strongly in GABA cells immunoreactive for somatostatin or parvalbumin. In layer V, calbindin D28k is strongly expressed mainly in somatostatin cells. The relative number of cells immunoreactive for a substance is proportional to the size of the box. Micrograph of stained cells adapted from Kawaguchi Y and Kondo S (2002) Parvalbumin, somatostatin and cholecystokinin as chemical markers for specific GABAergic interneuron types in the rat frontal cortex. *Journal of Neurocytology* 31: 277–287, with permission.

as yet understand the functional significance of these peptides and calcium-binding proteins, these expression patterns are not uniform, but rather are restricted to the subpopulations of GABA cells. The uneven distribution of these peptides and proteins in GABAergic neurons stimulated researchers to use immunohistochemistry for multiple substances to reveal classes of interneurons that co-express discrete combinations of these markers (Figure 2). The findings from these studies demonstrate that parvalbumin, somatostatin, and VIP cells are distinct, nonoverlapping groups. Further, calretinin cells are negative for parvalbumin and somatostatin, but partially overlap with VIP cells. As a result, we can identify five major interneuron groups: (1) parvalbumin-positive cells, (2) somatostatin-positive cells, (3) VIP- and calretinin-positive cells, (4) VIP-positive, calretinin-negative cells, and (5) VIP-negative, calretinin-positive cells. CCK is expressed in VIP cells, but also in cells negative for VIP, somatostatin, calretinin, or parvalbumin. CRF is differentially expressed between layers II/III and V/VI, and, in layers II/III, most CRF cells are positive for VIP. Calbindin is expressed mostly in somatostatin cells, but also weakly in layer II/III parvalbumin cells. Nitric oxide synthase (NOS), the enzyme generating NO, is also found in some GABA cells expressing somatostatin and neuropeptide Y. While the expression patterns for these markers are complex and their physiological roles still largely unknown, the distinctive expression patterns in neocortical interneurons have greatly aided the identification and classification of these neurons.

Firing Pattern Diversity

Firing characteristics of cortical neurons have been investigated in both *in vivo* and *in vitro* preparations. Response patterns to artificial current injections are heterogeneous among cortical cells. Differences in spike duration, spike afterhyperpolarization, and repetitive firing pattern responses to step depolarizations are used in the physiological classification of cortical neurons. Although pyramidal cells exhibit a range of firing properties, physiologically they are a much more homogeneous group than are neocortical interneurons. The physiological characteristics of interneurons, however, are even more heterogeneous than are their dendritic and axonal morphologies. Even so, some distinctions can be made between cell types. Here we introduce the most simplified scheme for the physiological classification of nonpyramidal cells (Figure 3):

1. *Fast-spiking (FS) cells.* These neurons have relatively hyperpolarized resting membrane potentials and have lower input resistances and shorter duration spikes than do other nonpyramidal cells.

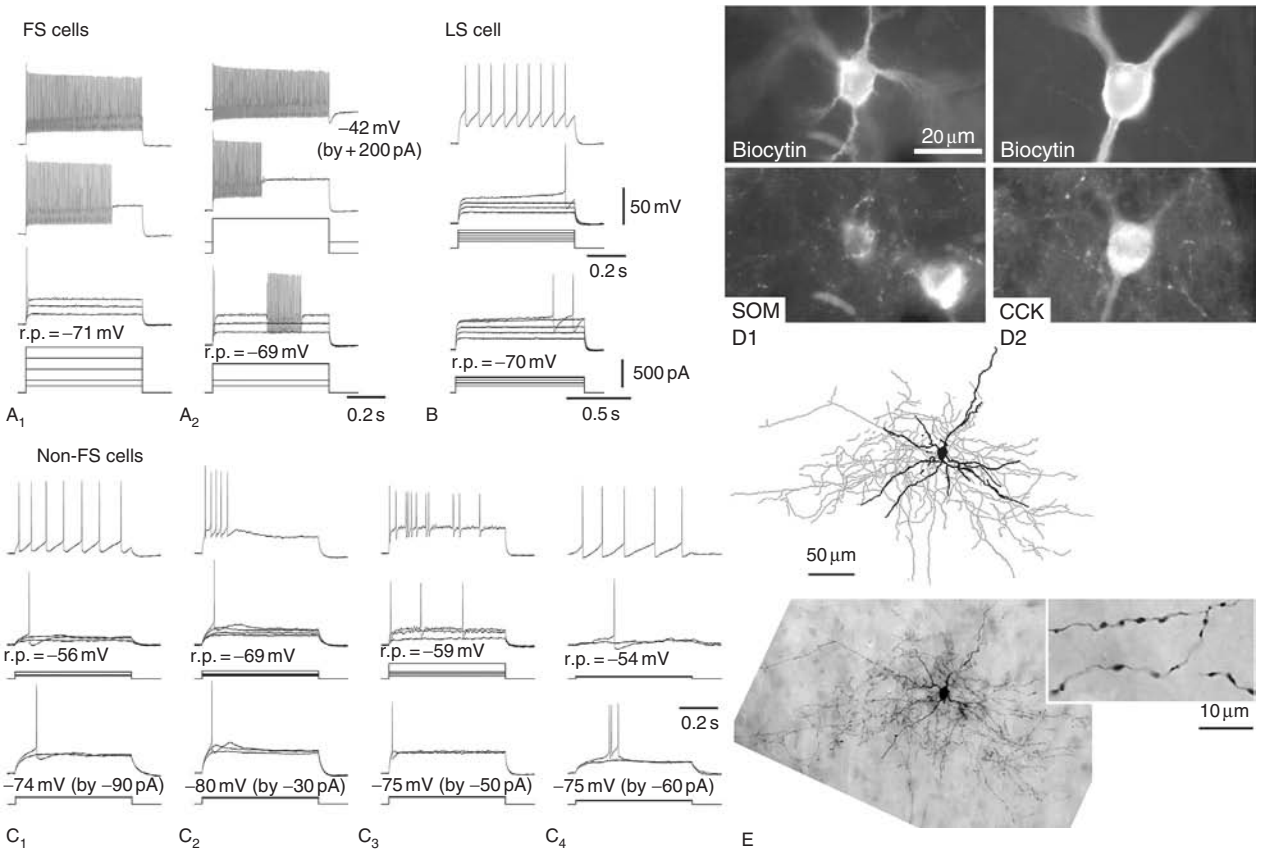

Figure 3 (Left) Firing responses of nonpyramidal fast-spiking (FS; A_1, A_2), late-spiking (LS; B), and non-FS (C_1, C_2, C_3, C_4) cells induced by depolarizing current pulses (r.p., resting membrane potential). Membrane potential changes by direct current somatic current injection (shown in parentheses) are indicated for the traces. (Right: cells D_1 and D_2) Immunohistochemical identification of physiologically characterized and biocytin-injected cells as positive for somatostatin (SOM) or cholecystokinin (CCK). (Right: cell E) The diagram above the micrograph depicts a reconstruction of dendrites (thick lines) and axon collaterals (thin lines) of an intracellularly stained fast-spiking cell. Micrograph of the stained cell shows that soma, dendrites, and axon collaterals are stained, as are axonal boutons (inset).

These characteristics suggest that these neurons need stronger excitatory input for action potential initiation. As their name suggests, sustained depolarization can induce nonadapting, repetitive firing at frequencies higher than 40 Hz.

2. *Late-spiking (LS) cells.* Given just-suprathreshold current injections, these neurons display ramp depolarizations before spike discharge. Above the threshold, nonadapting firing is observed in response to current pulses.

3. *Non-FS cells.* These neurons are neither FS nor LS cells, and are diverse in firing patterns. Some cells show a low-threshold spike (LTS) following depolarization from hyperpolarized potentials, and some occasionally generate multiple spikes following depolarization (i.e., burst-spiking nonpyramidal (BSNP) cells). Other cells, known as regular-spiking nonpyramidal (RSNP) cells, do not show LTSs, and fire continuously with variable adaptation patterns in response to current pulses. Some RSNP cells also have rebound depolarization after release from hyperpolarizing current steps that can sometimes generate a single spike, perhaps due to hyperpolarization-activated cation currents activated during hyperpolarization. Some BSNP or RSNP neurons have irregular interspike intervals; these irregular-spiking (IS) cells sometimes exhibit an initial burst of spikes. Some neurons of this group, phasic firing cells, fire several spikes or a burst discharge only during the initial depolarization in response to current pulses, displaying strong spike adaptation.

Nonpyramidal Cell Subtypes Identified by a Combination of Anatomical, Physiological, and Chemical Characteristics

Cortical nonpyramidal cells seem highly heterogeneous in their individual characteristics, anatomically, physiologically, and chemically. However, comparison of all three types of characteristics reveals

several basic cellular subtypes that can be defined as shown in Table 1 although the specifics of cell classification are still intensely debated.

FS/Parvalbumin Cell Group

FS cells are mostly positive for parvalbumin and negative for somatostatin, VIP, CCK, and calretinin. FS cells have very few spiny dendrites. This group is generally composed of basket and chandelier cells. The axon collaterals of both of these cell types do not extend into layer I. A few dendritic branches of FS basket cells arborize with in layer I, whereas FS chandelier cells extend dendrites into layer I freely. FS chandelier cells are found mostly in layers II/III (Figure 4). However, there are also non-FS chandelier cells without dendritic extensions into layer I, supposed to be negative for parvalbumin.

FS basket cells in layer V are larger in their dendritic field than are those in layers II/III. Layer V FS basket cells include large basket cells. In addition to basket and chandelier cells, there are a few FS cells issuing axon collaterals without basket terminals, some having wider and sparser axonal fields relative to other FS cells.

Table 1 Major groups of cortical GABAergic cells

Fast-spiking parvalbumin cells
Basket cells, including large basket cells
Wide arbor cells, including cells without basket terminals
Chandelier cells
Late-spiking neurogliaform cells
Non-fast-spiking cells
Somatostatin cells
Martinotti cells
Wide arbor cells
VIP/CCK/calretinin cells[a]
Double-bouquet cells
Horizontal/descending arbor cells
Small basket cells
Descending basket cells
Arcade cells
CCK large basket cells

[a]VIP, Vasoactive intestinal polypeptide; CCK, cholecystokinin.

LS Neurogliaform Cell Group

LS cells are found in layers II/III and I, and include a few neurons in layer V. LS cells in layers II/III are mostly neurogliaform. They are also called spiderweb cells based on their axonal morphologies. In neurogliaform cells, the LS characteristic is observed from both resting potential and hyperpolarized potentials. LS neurogliaform cells have a dendritic field spanning 100–200 μm, and axonal arbors about twice as wide. Their axon collaterals run horizontally, but turning at the horizontal edge of the axon territory, seemingly to observe the restricted axonal field. GABAergic nonpyramidal cells are generally divided

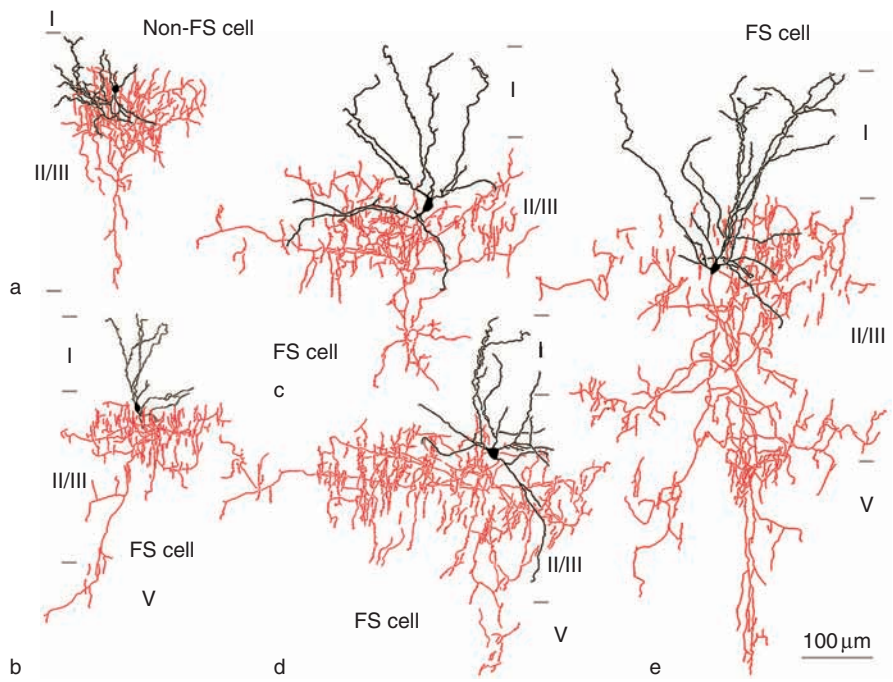

Figure 4 Reconstructions of five chandelier cells, demonstrating characteristic vertical arrays of terminal boutons. Four cells (b–e) were FS cells. Two cells (c, d) were also immunohistochemically identified as parvalbumin positive. Layers I, II/III, and V are indicated. Adapted from Kawaguchi Y and Kondo S (2002) Parvalbumin, somatostatin and cholecystokinin as chemical markers for specific GABAergic interneuron types in the rat frontal cortex. *Journal of Neurocytology* 31: 277–287.

into basket and non-basket cells based on their innervation preference for somatic domains, but innervation patterns of LS nuerogliaform cells are more variable in thier tendency to choose somatic targets. The peptides and calcium-binding proteins previously described have not been detected in LS neurogliaform cells, but alpha-actinin-2, an action-binding protein is expressed in them.

A few LS cells with dendritic and axonal morphologies different from the neuroglia form cells have basket terminals, and these may belong to the non-FS basket cell group described in the following section.

Somatostatin Cell Group

Somatostatin is expressed only in non-FS cells. Somatostatin cells in layers II/III are mostly RSNP cells, but layer V somatostatin cells include both RSNP and BSNP cells. In contrast to FS cells, these neurons have sparsely spiny dendrites, although the spine density is much lower than that found in pyramidal cells. This group contains two morphological classes: Martinotti and wide arbor cells. Martinotti cells were originally defined as deep layer cells with ascending axonal arbors. We now classify layer II/III somatostatin cells with axonal arbors ascending into layer I as Martinotti cells. Wide arbor somatostatin cells typically have axon collaterals that are ascending, but that also project in descending and horizontal directions. The dendritic fields of Martinotti cells tend to be larger in layer V cells than in cells in layers II/III. Additionally, layer I innervation is more robust in layer II/III Martinotti cells than in cells located in layer V.

VIP/CCK/Calretinin Cell Group

The cells of the VIP/CCK/calretinin cell group are more diverse in their morphological, physiological and chemical characteristics than are the aforementioned groups. Physiologically they are non-FS cells and include both BSNP and RSNP cells, as well as IS cells and phasic firing cells. They can be subdivided into six cell groups: double-bouquet, horizontal/descending arbor, small basket, descending basket, arcade, and CCK large basket cells.

Double-bouquet cells In primates, double-bouquet cells are bitufted and multipolar cells with descending axon collaterals restricted in their horizontal span. In this group we include non-FS cells with mainly descending axon collaterals, even if they do not descend together in a seemingly 'horsetail-like' fascicle. Axons of double-bouquet cells rarely make basketlike terminals. These neurons express VIP or calretinin, with some expressing both CRF and VIP. Although they are expressed in some overlapping population of GABA cells, VIP- and calretinin-expressing double bouquet cells may have different connection patterns.

Horizontal/descending arbor cells VIP- or calretinin-positive cells typically have axonal arbors that extend horizontally in layers II/III, with a few axon collaterals that descend into layer V.

Small basket cells These are non-FS cells with small somata that make basketlike terminals onto the somata of nearby neurons. They express VIP or calretinin, and some express VIP and CRF, or VIP, CRF, and CCK. They are found in layers II/III, and have axons that are generally restricted to those layers.

Descending basket cells Like the double-bouquet cells, these neurons have descending axonal arbors that innervate deeper layers. However, their axons tend to make basketlike terminals on their target neurons. The expression pattern of chemical markers in these neurons is similar to that of the small basket cells.

Arcade cells VIP cells in layer V have short ascending main axons that branch into local collaterals that send additional branches descending into the dendritic field. This axon arborization pattern looks like arched roofs (thus the name 'arcade cell'). The terminals of some VIP arcade cells make multiple boutons onto postsynaptic somata, and therefore may be classified as a type of basket cell.

CCK large basket cells Large CCK-positive cells in layers II/III, negative for VIP and CRF, are basket cells producing multiple boutons onto the somata of other cortical neurons. Thus, CCK cells include large basket cells negative for VIP, and small basket cells positive for VIP. The neocortex has both FS and non-FS basket cells, the latter being composed of small and large neurons, descending basket cells, and arcade cells.

Comparison of FS Basket Cells with Other Classes

To better understand the functional relevance of the different kinds of cortical interneurons, it is useful to compare several subtypes with FS basket cells, because they are the most abundant GABAergic neurons and have been most frequently investigated in their details, compared with the other subtypes. Despite their name, only about 20% of the axonal terminals of FS basket cells innervate somata. Most FS basket cell boutons form synapses onto dendritic shafts and spines. Therefore, basket cells innervate both somatic and dendritic domains of postsynaptic

pyramidal cells. FS basket cells generally inhibit postsynaptic cells via activation of $GABA_A$-type receptors.

Chandelier cells make multiple synaptic contacts almost exclusively onto the axonal initial segments of pyramidal cells. Given that the axonal initial segment is the primary site of action potential generation in pyramidal neurons, it has been widely suggested that chandelier cells provide a powerful inhibitory input to block the spike initiation in pyramidal cells. However, new data suggest that these GABAergic synapses may actually promote spike generation in pyramidal neurons by generating depolarizing responses in the axonal initial segment. Because single chandelier cells synapse onto up to 1000 axonal initial segments, these neurons likely control the output of large groups of cortical pyramidal neurons.

LS neurogliaform and FS basket cells both show nonadapting spike discharges compared to other nonpyramidal cell classes, but have different spike frequencies in response to depolarizing current pulses. Both cell types have multipolar aspiny dendrites. Layer II/III basket cells mostly do not extend dendritic or axonal collaterals into layer I, whereas neurogliaform cells often do. Axonal arborizations extend more horizontally in FS basket cells, but are more restricted in neurogliaform cells. Neurogliaform cells induce both $GABA_A$- and $GABA_B$-receptor mediated inhibitory postsynaptic potentials (IPSPs) in target neurons. FS basket cells are connected with each other through electrical coupling via gap junctions in their dendrites, while neurogliaform cells are electrically coupled not only with each other but also with other nonpyramidal cell types.

Somatostatin Martinotti cells make symmetrical synaptic contacts onto thin dendritic branches and spines, rather than onto somata. FS basket cells have mostly aspiny dendrites, whereas sparely spiny Martinotti cells have the highest dendritic spine density among the nonpyramidal cell subtypes. In addition to their density, the spines of FS basket and Martinotti cells also differ in their shape. Individual spines are longer in Martinotti cells than in FS basket cells. Mushroom spines, spines with a larger head compared with the neck, and multihead spines are commonly found in Martinotti cell dendrites but are very rarely observed in FS basket cells. Like FS basket cells, somatostatin cells are electrically coupled with each other. However, these two cell types process excitatory input differentially. Successive excitatory synaptic transmissions between pyramidal cells and FS cells show short-term depression in response amplitude, while responses in Martinotti cells show facilitation. Additionally, thalamic fibers from the somatosensory relay nuclei preferentially innervate FS cells rather than somatostatin cells in layer IV of somatosensory cortex, suggesting that FS cells participate in feed-forward inhibition to a greater extent than do somatostatin cells.

Basket cells innervating somata are heterogeneous and contain several subtypes of nonpyramidal cells. Non-FS basket cells include small, large, and descending cells. It remains to be investigated why multiple types of basket cells exist in the cortex, but FS and non-FS (often VIP- or CCK-positive) basket cells are distinct in their chemical modulation. FS cell terminals express muscarinic receptors, whereas CCK basket cells have cannabinoid receptors. The activation of these receptors suppresses the release of GABA in both neuronal subtypes, suggesting that GABA release is differentially regulated in these two basket cell subtypes.

Comparison with Hippocampal GABAergic Cells

Compared with inhibitory interneurons in the neocortex, those in the hippocampus, especially the CA1 region, have been investigated in more detail; comparison of GABA interneuron organization will be helpful for understanding the fundamental principles. Basically, the chemical and morphological characteristics are correlated in hippocampal interneurons, but one chemical marker is usually found in more than one class, as in neocortex. As found in neocortical basket cells, basket cells in the CA1 region of the hippocampus are heterogeneous and express (1) parvalbumin alone, (2) CCK alone, or (3) both CCK and VIP. Hippocampal interneurons that innervate dendrites express somatostatin, but also some of them express parvalbumin in addition to somatostatin. The calbindin expression pattern is also different from that in the neocortex. This suggests that the expression patterns of the various peptides and calcium-binding proteins are dependent on the local cortical area in which the neurons reside.

There are also neurogliaform cells in the hippocampus. Among calretinin- or VIP-expressing GABA cells, there are interneurons that selectively innervate other GABAergic cells; these are called 'interneuron-selective interneurons.' It is possible that a similar type of nonpyramidal cell also exists in the neocortex.

Hippocampal interneurons are profoundly involved within the hippocampus in generating network oscillations, such as theta rhythm (4–10 Hz) and high-frequency 'ripples' (120–200 Hz). Interneurons belonging to different connectivity classes are suggested to fire preferentially at distinct time points during a given oscillation. In contrast, the neocortex displays different types of oscillations, such as slow oscillation (about 1 Hz) and gamma oscillation (30–80 Hz).

The differences in interneuron organization between hippocampus and neocortex may be related to differences in these oscillation patterns.

Developmental Differentiation of GABAergic Cells

Pyramidal and GABAergic nonpyramidal cells are derived from distinct precursor cells. Pyramidal cells derive from precursor cells the dorsal telencephalon, but most cortical GABAergic neurons originate from stem cells in the ventral telencephalon (the ganglionic eminence) and migrate into the cortex tangentially. Furthermore, the subtype of GABAergic neuron originating in the ganglionic eminence depends not only on the subregion of the ganglionic eminence from which it derives, but also on the developmental stage at which it starts to migrate into the cortex. The medial ganglionic eminence generates FS cells and somatostatin cells, whereas the caudal ganglionic eminence produces calretinin cells. Therefore, the physiological, chemical, and morphological characteristics in individual subtypes are to a large extent genetically predetermined before interneurons finish their migration into the cortex, suggesting that various transcriptional factors control the fates of cortical interneurons. Indeed, some transcriptional factors involved in cortical GABA cell differentiation have already been identified.

Neuron Types and Circuit Organization in the Cortex

Recent research has revealed that there are several archetypes of cortical interneurons. Here we have classified them according to their morphological, chemical, and physiological properties. However, there are additional characteristics that may further differentiate this already heterogeneous collection of neurons. For example, the firing patterns of interneurons may differ depending on the state of ongoing cortical activity (often correlated with the behavioral state of the animal). Other differences include the expression of ion channels (ligand-gated receptors and voltage-gated channels). For instance, voltage-gated potassium channels are selectively expressed in some interneuron subtypes. Potassium channel Kv3.1 proteins are found in parvalbumin cells, while nicotinic or 5-hydroxytryptamine ($5\text{-}HT_3$) receptors are selectively expressed in CCK- and VIP-expressing neurons.

Modern molecular genetic technologies have produced several lines of transgenic mice in which a subpopulation of GABA cells is selectively labeled with fluorescent material. The ability to selectively study these genetically distinct interneurons will allow for more detailed characterization of different subclasses of interneuron. The expression patterns of multiple genes have been recently examined in identified subpopulations of interneurons, including parvalbumin- and somatostatin-expressing cells. The results indicate that while cortical interneurons are highly heterogeneous, each subtype has a specific expression pattern of genes, including transcriptional factors. It is hoped that this method of molecular taxonomy will generate a consensus regarding neuronal organization of the cortex, facilitating our understanding of how the various subclasses of neurons interact together to form functional cortical circuits.

See also: Cerebral Cortex; Gamma-Aminobutyric Acid (GABA); Memory Consolidation: Cerebral Cortex; Somatostatin and Receptors.

Further Reading

Butt SJ, Fuccillo M, Nery S, et al. (2005) The temporal and spatial origins of cortical interneurons predict their physiological subtype. *Neuron* 48: 591–604.

Cauli B, Porter JT, Tsuzuki K, et al. (2000) Classification of fusiform neocortical interneurons based on unsupervised clustering. *Proceedings of the National Academy of Sciences of the United States of America* 97: 6144–6149.

Douglas RJ and Martin KA (2004) Neuronal circuits of the neocortex. *Annual Review of Neuroscience* 27: 419–451.

Fairén A, DeFelipe J, and Regidor J (1984) Nonpyramidal neurons: General account. In: Peters A and Jones EG (eds.) *Cerebral Cortex, Volume 1, Cellular Components of the Cerebral Cortex*, pp. 201–253. New York: Plenum.

Galarreta M and Hestrin S (1999) A network of fast-spiking cells in the neocortex connected by electrical synapses. *Nature* 402: 72–75.

Gibson JR, Beierlein M, and Connors BW (1999) Two networks of electrically coupled inhibitory neurons in neocortex. *Nature* 402: 75–79.

Jones EG (1975) Varieties and distribution of non-pyramidal cells in the somatic sensory cortex of the squirrel monkey. *Journal of Comparative Neurology* 160: 205–267.

Karube F, Kubota Y, and Kawaguchi Y (2004) Axon branching and synaptic bouton phenotypes in GABAergic nonpyramidal cell subtypes. *Journal of Neuroscience* 24: 2853–2865.

Kawaguchi Y, Karube F, and Kubota Y (2006) Dendritic branch typing and spine expression patterns in cortical nonpyramidal cells. *Cerebral Cortex* 16: 696–711.

Kawaguchi Y and Kondo S (2002) Parvalbumin, somatostatin and cholecystokinin as chemical markers for specific GABAergic interneuron types in the rat frontal cortex. *Journal of Neurocytology* 31: 277–287.

Kawaguchi Y and Kubota Y (1997) GABAergic cell subtypes and their synaptic connections in rat frontal cortex. *Cerebral Cortex* 7: 476–486.

Klausberger T, Magill PJ, Marton LF, et al. (2003) Brain-state- and cell-type-specific firing of hippocampal interneurons *in vivo*. *Nature* 421: 844–848.

Kubota Y, Hattori R, and Yui Y (1994) Three distinct subpopulations of GABAergic neurons in rat frontal agranular cortex. *Brain Research* 649: 159–173.

Markram H, Toledo-Rodriguez M, Wang Y, et al. (2004) Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience* 5: 793–807.

Ramón y and Cajal S (1911) *Histology of the Nervous System*. vol. 2, Swanson N and Swanson LW (trans.). New York: Oxford University Press.

Somogyi P and Klausberger T (2005) Defined types of cortical interneurone structure space and spike timing in the hippocampus. *Journal of Physiology* 562: 9–26.

Somogyi P, Tamás G, Luján R, et al. (1998) Salient features of synaptic organisation in the cerebral cortex. *Brain Research Reviews* 26: 113–135.

Sugino K, Hempel CM, Miller MN, et al. (2006) Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature Neuroscience* 9: 99–107.

Szabadics J, Varga C, Molnár G, et al. (2006) Excitatory effect of GABAergic axo-axonic cells in cortical microcircuits. *Science* 311(5758): 233–235.

Tamás G, Lorincz A, Simon A, et al. (2003) Identified sources and targets of slow inhibition in the neocortex. *Science* 299: 1902–1905.

Cerebral Cortex: Symmetric vs. Asymmetric Cell Division

G Fishell and C Hanashima, New York University Medical Center, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The human cerebral cortex contains approximately 10^{10} cells that exhibit a wide array of physiological properties; these cells are wired together by synaptic connections that show remarkable specificity. This highly ordered structure is organized along two axes. Vertically, the cerebral cortex can be divided into six discrete laminae, each of which displays distinct afferent and efferent connections. Horizontally, cortical columns within each region of the cortex subserve specific aspects of cortical function, such as processing somatosensory or visual information or coordinating motor outputs. Despite this impressive architecture, cortical neurons can be grossly subdivided into two major classes, the excitatory pyramidal or principal neurons (80%) that are glutamatergic and provide the major structural inputs and outputs of the cortex, and the inhibitory GABAergic interneurons (~20%) that regulate cortical function through local connections. Pyramidal neurons with characteristic axonal connections, dendritic morphologies, and physiological properties are segregated in each of the different cortical layers. For instance, morphologically, deep-layer (V/VI) neurons have a bigger soma compared to the superficial-layer (II/III) neurons, which typically contain smaller cell bodies. Moreover, a majority of layer V/VI pyramidal neurons project to subcortical structures, including the thalamus, basal ganglia, and spinal cord, whereas layer II/III neurons project intracortically to other cortical areas. Similarly, cortical interneurons at least in some cases are enriched within specific cortical laminae. While the cell bodies of the Martinotti interneurons are found predominantly in the deep layers of the cortex, the neuroglioform interneurons are preferentially found in the superficial cortex. A similar subdivision of both of these populations occurs within the horizontal plane, as best shown by the expansion of input cortical layers (layers I and IV) in sensory regions such as visual cortex and enlargement of output layers of the cortex (layers V and VI) in motor regions. Understanding how such exquisite organization is established during development is one of the major challenges in developmental neurobiology.

Efforts to understand how cortical cellular diversity arises dates back to the late ninteenth century. The first systematic attempts to explore the organization of the developing brain came from the study of early neurogenesis, a time when epithelial cells span their fibers across the entire cortical wall. These epithelial cells, which appear to be equivalent to what we today call 'radial glial cells,' were referred to by His as 'spongioblasts' based on their elongated spongelike appearance and were thought to give rise solely to glial cells. The other 'cell type' observed by His was the rounded 'germinal cells' that reside at the apical surface of the neural tube, which he proposed were the mitotically active neuronal progenitors. This view was challenged by Magini, who observed dramatic morphological variations within the spongioblasts, and suggested that the spongioblasts and germs cells were simply the same population at different points in the cell cycle, an idea that was later confirmed by Sauer. Sauer reached this conclusion through the observation that cell nuclei at the basal side of the neural tube contain an increased complement of DNA (i.e., S phase), while those adjacent to the ventricle were tetraploid (i.e., M phase). This suggested that neural progenitors undergo 'interkinetic nuclear migration' as they progress through the cell cycle and concurrently undergo dynamic changes in their cellular morphology and nuclear position. That this occurred was definitively shown in 1959 where the movements of dividing cells were followed using *post hoc* analysis of cells previously labeled with [^{3}H]thymidine. This autoradiographic method revealed that cortical progenitor cells indeed undergo DNA replication at a distance from the ventricular surface, and then subsequently translocate to the apical surface, where the cells round up and undergo mitosis. From these divisions arise daughter cells that either terminally differentiate into neurons and migrate to the pia or undergo a second cycle of proliferation. Although nuclei prior to neurogenesis translocate across the entire width of the neural tube, subsequent to the onset of neurogenesis, interkinetic migration is confined to a region adjacent to the lateral ventricle. The geographical restriction of cells undergoing interkinetic nuclear migration led to this region being designated as the 'ventricular zone.'

Within the cortex, postmitotic cells exit the ventricular zone and migrate along radial glia toward the pial surface. Surprisingly, the structural role of radial glial cells in guiding migration in the cortex was established long before it was appreciated that the radial glia were themselves the progenitor population that gave rise to newborn neurons. Indeed, for the three decades after radial glial cells were known to support neuronal migration, the radial glia and cortical neuronal progenitors were considered to be two distinct cell lineages, with the former solely

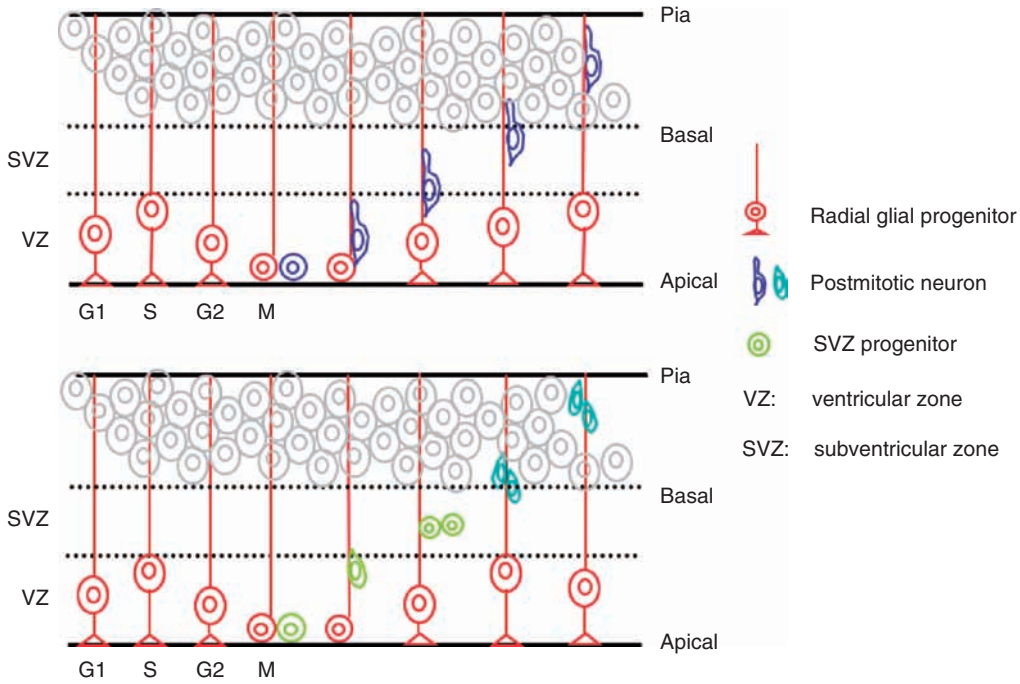

Figure 1 Asymmetric and symmetric cell division in the cerebral cortex. Top panel: Radial glial progenitors undergo interkinetic nuclear migration during the cell cycle. One asymmetric cell division results in the production of two daughter cells, one of which remains in the VZ and inherits the radial fiber and the other of which translocates outward toward the pia along the radial fiber to become a postmitotic neuron. Bottom panel: Alternatively, SVZ progenitors arise from the VZ and subsequently undergo symmetric divisions to produce two postmitotic neurons. Adapted from Noctor SC, Martinez-Cerdeno V, Ivic L, et al. (2004) Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. *Nature Neuroscience* 7: 136–144.

producing astrocytes. However, the seminal paper by Noctor and Kriegstein used time-lapse imaging of retrovirally infected cortical progenitors to demonstrate that radial glial cells, in addition to supporting the migration of nascent neurons, gave rise to them. These studies revealed that neurons assembled in close radial units are not only clonally related, but also that radial glial cells can simultaneously give rise to both differentiated neurons and progenitor cells (Figure 1). This once and for all established that radial glial cells are the major and perhaps sole progenitor for cortical pyramidal neurons. Moreover, although previous analysis suggested that asymmetric cell divisions occurred in the cerebral cortex, this analysis provided the first definitive evidence establishing the existence of this phenomenon.

Cell-Type Specification in the Neocortex

Temporal Determinants: Birthdate and Laminar Fate

How is the diversity of distinct layer neuronal subtypes created from a restricted progenitor pool within the VZ? The VZ appears to utilize temporal cues to sequentially produce different subclasses of neurons, rather than simultaneously generating multiple progenitor pools, each of which gives rise to particular cell type. Studies examining the origin of distinct neuronal subtypes indicate specific neuronal populations arise at precise times during development, presumably in response to changing temporal determinants. The first evidence that birthdate was a strong predictor of neuronal fate in the cerebral cortex came from [^{3}H] thymidine studies in rodents; it was revealed that the administration of a tritiated thymidine injection at distinct times during neurogenesis resulted in the labeling of neurons within particular cortical laminae. Moreover, this occurred in an inside-out fashion, where early-born neurons occupy the deepest position within the cerebral cortex, while later-born cells migrate past their ascendants, to take up more superficial positions. Hence, in the mature cortex, layer VI neurons are born first and layer II neurons are born last. The only exception to this rule is the earliest-born neurons, the Cajal–Retzius (RC) cells that transiently reside in layer I. The inside-out relation of laminar position to birthdate, although evident in other mammals, is more precise in higher species, such as primates. This has been interpreted as suggesting that the progenitors within the ventricular zone undergo multiple cell cycle divisions, resulting in the production of neurons destined for sequentially more superficial positions.

Direct evidence for the sequential production of cortical neurons came from both static and live-imaging clonal analysis of ventricular progenitors using retroviruses. These studies revealed that individual cortical progenitors produced multiple subtypes over a prolonged period during cortical development. These clonal and birthdating studies raised the question as to how neural progenitors in the ventricular zone change their distinct laminar fate according to temporal cues. One possibility is that extrinsic cues surrounding the ventricular zone change over time to influence the fate of the progenitor cells. Alternatively, it was proposed that progenitors cells might be intrinsically programmed to sequentially produce distinct neuron types at particular developmental timepoints. To test the relative contribution of extrinsic and intrinsic factors to temporal cortical cell-fate determination, classical transplantation experiments in ferrets were carried out to heterochronically challenge ventricular zone progenitors by transplanting them reciprocally to earlier or later developmental timepoints. Specifically, in these experiments early-born progenitors fated for deep layers of cortex (layer VI) were transplanted to periods of development when late-born neurons within layers II/III were being generated. Conversely, cortical progenitors that would normally produce layers II/III were transplanted into host cortical ventricular zone at a time when layer V was being generated. The results revealed a forward ratcheting in the potential of progenitors as development progressed. While early progenitors were found to be competent to adopt later fates when transplanted to older hosts (i.e., progenitors normally fated to produce layer V neurons upon transplantation to later stages of development could produce layer II/III neurons), the converse was not true (i.e., progenitors from later timepoints in development did not produce earlier-born populations of neurons even when transplanted to host animals where earlier populations of neurons are being produced). These results highlight that the cell fate in the cerebral cortex is largely established by intrinsic determinants and that progenitor potential becomes progressively restricted as development progresses. Interestingly, while the molecular mechanism governing these fate restrictions is not well understood, timed transplants of early-born cells to late donors revealed that the fate of neurons is set within 6 h of their final S phase.

Asymmetric and Symmetric Division of Cortical Progenitors Function in the Generation of Cortical Cell Diversity

What is the mechanism by which a common pool of cortical progenitors within the germinal zone gives rise to a broad diversity of neurons over a prolonged period of time? It has been postulated that the mode of cellular division utilized by the progenitor is central to this process. During early periods of neurogenesis, a majority of cortical progenitors undergo symmetric divisions that produce two progenitor cells, in order to expand the cortical progenitor pool. However, that a common progenitor produces distinct types of neurons over multiple cell divisions also implies that later there is a second process by which two daughter cells can assume distinct cell fates. Given that postmitotic daughter cells are produced sequentially, it suggests the existence of asymmetric divisions where one daughter cell reenters the cell cycle, while the other exits to differentiate into a mature neuron. That this can occur hints that during an asymmetric division the two daughter cells adopt fundamentally distinct intrinsic cell-fate programs, where asymmetric divisions unequally bestow cell-fate determinants into only one of the two daughter cells. Considerable efforts have been made to identify these 'determinants.' *Drosophila* neurogenesis has provided a perfect context to study this question, as neuroblasts in the central nervous system (CNS) as well as sensory organ precursor cells in the PNS undergo asymmetric divisions, depending on the precise orientation of their cleavage plane. Notably, the protein Numb, together with Prospero and the adaptor proteins Miranda and Pon, are asymmetrically localized to the basal surface of neuroblasts during metaphase. Therefore, when the orientation of the cleavage plane is horizontal to the apical surface, the daughter cell on the basal side inherits the majority of determinants, such as Numb protein. Thus, asymmetrical cell division requires two events, establishing the cell polarity by regulating the orientation of the mitotic spindle along the apical–basal axis, and targeting Numb and associated proteins to the basal side. Identification of asymmetric cell-fate determinants has come from analysis of the aberrations in the neural lineages of mutant flies, where one of the critical determinants, such as Numb or Prospero, is lost. The loss of these results in a randomized spindle orientation and a corresponding perturbed cell fate. Although the precise function of asymmetric localization of Numb is not clear, it is likely that Numb functions as an inhibitor of Notch signaling, which is required for maintaining neuroblasts in a progenitor state.

In the cerebral cortex, it has been suggested that the plane of cell cleavage may also determine the fate of the resulting daughter cells. Examination of the ferret cortex has shown that around 15% of the mitosis in the ventricular zone has a horizontally oriented cleavage plane, where, following division, one daughter cell retains the apical contact and remains in the ventricular zone, while the basal daughter cell migrates away from the ventricular

zone and becomes a postmitotic neuron. Interestingly, the mouse homolog of *Drosophila* Numb, m-Numb, is asymmetrically localized to the apical membrane of the dividing progenitors in the ventricular zone, suggesting a conserved function of Numb in determining progenitor outcome in mammals. However, a majority of the cell divisions exhibit a vertical or randomly oriented cleavage plane. Thus, the cleavage orientation is not clearly bifurcated into distinct division modes as is the case in flies. Hence, as the 15% of the horizontal cleavage cannot account for all neuronal production during peak neurogenesis period, the relationship between the cleavage plane and asymmetric division, and its contribution to developmental fate in the cerebral cortex, need to be clarified.

What, then, is the function of Numb and Notch proteins in the cerebral cortex? In mammals, the interpretation of loss-of-function of either of the genes is not straightforward, as there are multiple homologs of both genes that appear to functionally compensate for each other. In mouse two *numb* homologs exist, mouse *numb* (*m-numb*) and *numblike* (*nbl*). Mice that lack *m-numb* have highly perturbed development and die prior to E11.5. Although the $nbl^{-/-}$ mutant alone does not exhibit an apparent neural developmental phenotype, this is likely a result of functional compensation between these two genes, as double mutants demonstrate a more severe phenotype and die around E8.5. To circumvent this early lethality, a conditional allele of *m-numb* was generated to test the requirement of this gene during the neurogenic period. The outcome of these experiments suggested that Numb has pleiotropic roles during the corticogenesis period, as distinct phenotypes were observed, depending on when the gene was removed. Conditionally ablating *m-numb* and *nbl* with Cre-recombinase driven under the Nestin promoter, which specifically removes genes starting around E8.5, resulted in an early depletion of the cortical progenitor pool. Similarly, compound removal of *m-numb* and *nbl* during later corticogenesis using a late Cre driver, D6-Cre (which becomes active only after E10.5), also suggested that both genes are required to maintain the progenitor cells in an undifferentiated state. However, removal of *m-numb/nbl* gene function using an Emx1-Cre (which becomes active at E9.5 and results in complete recombination within the cerebral cortex by E12.5) resulted in hyperproliferation of the cortical progenitors as well as inhibition of neuronal differentiation, suggesting that in this context Numb acts to direct progenitors to a neuronal fate. Taken together it is apparent that further work is needed to assess why the function of Numb is diametrically opposite, depending on the context of its removal.

More clear-cut is the function of Notch signaling during neurogenesis. Loss of Notch signaling results in the precocious differentiation of neural progenitors, while overexpressing a constitutively active form of Notch (caNotch) results in inhibition of neurogenesis and promotion of a radial glia cell fate. It is notable that the transient ectopic activation of Notch signaling early in neurogenesis during the period when deep-layer neurons are produced does not disrupt the normal temporal progression of neurogenesis. Upon removal of the ectopic Notch activation, upper-layer corticogenesis resumes normally. Hence even though the transient activation of Notch prevents progenitors from differentiating, on removal of Notch signaling the normal temporal progression of neurogenesis resumes. This implies that the primary function of Notch in cortical development is to maintain the progenitors in an undifferentiated state, but it does so without disrupting the developmental clock that controls the competence of progenitors. In the future, better understanding of the precise roles of Numb and Notch in cortical development, as well as their genetic and biochemical interactions, should prove informative. In addition, whether asymmetrical cell division determines cell fate or merely controls the length of corticogenesis as a means of expanding the number of neural progenitors needs to be clarified.

Does the Segregation of Progenitors to the Ventricular Zone and Subventricular Zone Represent a Bifurcation of Cell Fates?

The types of neurons produced in the cerebral cortex are clearly restricted by temporal cues within the ventricular zone; however, the mechanisms that determine layer neuron specification may not solely depend on temporal changes in progenitor competence. Recent studies have suggested that the restriction in neuronal class production can occur by a spatial segregation of cortical progenitors. In this regard, the emergence of the secondary germinal zone in the neocortex, the subventricular zone (SVZ), may play an important role in determining the identity of upper-layer neurons in the cerebral cortex. The SVZ is a transient zone that is juxtapositioned between the ventricular zone and the intermediate zone. It was traditionally defined by those progenitors that do not undergo interkinetic nuclear migration as they transit through the cell cycle. The SVZ is also distinguishable from the ventricular zone by its lack of pseudostratified epithelial morphology.

It has been postulated that since SVZ develops later than the VZ, SVZ progenitors predominantly give rise to the glial cells at the end of corticogenesis. However, since the peak neurogenesis period of upper-layer (layer II–IV) neurons overlaps with the

expansion of SVZ progenitors, the alternative view is that the spatial segregation between SVZ and VZ may directly restrict the fate of distinct layer neurons. This hypothesis is further supported by the strong correlation in genes expressed in both SVZ and upper-layer neurons. Two genes, *Svet1* and *Cux2*, specifically label mitotic cells in the SVZ. Notably, the level of expression in the SVZ population declines as the progenitors in this pool are depleted. In turn, *Svet1* and *Cux2* are expressed in postmitotic neurons that comprise neurons in layers II–IV. Beyond this simple correlation in gene expression, mutants that have abnormal development in SVZ also show abnormalities in the differentiation of upper layers of cortex. In *Pax6* mutants, the expression of *Svet1* and *Cux2* is downregulated in SVZ progenitors, and the upper-layer neurons fail to be properly established. Furthermore, expression of both *Svet1* and *Cux2* is inverted in *reeler* mutants, suggesting that these genes define the intrinsic properties of upper-layer neurons rather than their laminar position or migratory patterns. It is notable that progenitors change their mode of cell division as they transit from the VZ to the SVZ. During corticogenesis, a majority of VZ progenitors undergo asymmetric divisions, in which the two daughter cells adopt distinct fates to remain a progenitor or become a postmitotic neuron (**Figure 1**, top panel). However, SVZ progenitors are more likely to follow a symmetric pattern of division, where two daughter cells either become a pair of postmitotic neurons or together remain as progenitors (**Figure 1**, bottom panel). It is worth mentioning that not all the progenitors undergo mitosis in the SVZ, even during the peak neurogenesis of upper-layer neurons, and it remains possible that within the VZ and the SVZ there exists greater heterogeneity of progenitors than has yet to be appreciated. In this regard it is notable that in the primate neocortex studies of neurogenesis have subdivided the SVZ into inner and outer zones, the latter of which has been described to be another major source of cortical neurons. The emergence of SVZ during evolution may be the consequence of an evolutional strategy of expanding the diversity or numbers of cortical neurons that can be produced at a given time. Hence during evolution, in addition to the size and complexity of the architecture of the cerebral cortex, the cellular diversity may also have increased.

Intrinsic Factors of Cortical Cell-Fate Specification

Fundamental to cortical neurogenesis is the prolonged production of different siblings from a common pool of progenitors within the VZ and SVZ. This almost certainly depends on intrinsic cues to direct distinct neuronal fates. This process allows the progenitor pool to expand and simultaneously produce different cell types at precise times during neural development. What, then, is the identity of temporal determinants that alter the fate of the neuron types produced? As previously mentioned, cortical progenitors appear to utilize a ratcheting mechanism by which the neuronal types produced over time become progressively restricted. Moreover, later-born progenitors retain a limited ability to differentiate into neuron types prior to their own birthdate. In rodents, the molecular identity of this progressive restriction is best understood in the switch from generating the earliest-born CR cells to deep-layer projection neurons. This process is regulated by the transcriptional repressor *Foxg1*. In the cerebral cortex, *Foxg1* is expressed in the majority of the cortical projection neurons, with the exception of layer I CR neurons. Moreover, in the absence of *Foxg1*, cortical progenitors fail to generate later-born neurons and instead continue to produce the earliest-born CR neurons. In order to determine the cell-autonomous function of *Foxg1* in cell-fate specification, *Foxg1* was conditionally inactivated in deep-layer progenitors. Interestingly, removal of *Foxg1* at E13 (the birthdate of deep-layer neurons) results in the resumption of CR cell production in the cerebral cortex. By labeling the progenitors that were born subsequent to the removal of *Foxg1* with BrdU, it was confirmed that the cells that were normally destined to become deep-layer neurons instead adopted an earliest-cell fate. This work has recently been replicated *in vitro*, where the fate of individual neuroblasts can be directly followed. These studies suggest the early cell-fate specification in the mammalian cerebral cortex is determined by an active mechanism whereby a transcriptional repressor prevents the later-born neurons from adopting an earlier fate.

How is the cortical cell fate established during the subsequent corticogenesis period? It has been reported that a zinc-finger-containing transcription factor, *Fezl*, which is specifically expressed in corticospinal motor neurons in layer V, directs both the specification and identity of these cell types. Loss of *Fezl* results in both the loss of corticospinal neurons and the failure of subcortical projections. Furthermore, overexpression of *Fezl* near the end of layer V genesis can partially override the transition of cortical progenitors to produce neurons that migrate to layer IV, which instead send a projection through the internal capsule. It remains to be explored whether the competence window of cortical progenitors responding to either *Foxg1* or *Fezl* is limited. Beyond this, it seems clear that a great deal

more remains to be learned concerning the intrinsic mechanisms by which temporal competence is controlled within the cerebral cortex.

A Cascade of Intrinsic Determinants Controls the Production of Neurons in *Drosophila*

The progressive restriction of progenitors appears to be central for the specification of distinct cell types in a variety of tissues, such as in the retina and hematopoietic systems. This mechanism appears to be evolutionarily conserved, as similar rules appear to underlie the specification of cells in the nerve cord of invertebrates. In particular, work in *Drosophila* has begun to elucidate the molecular determinants controlling the precisely orchestrated production of different neuronal subtypes during development. In flies, the cell-lineage relationships of distinct neuron types have been well characterized and transcription factors that are specifically expressed in neuroblasts at distinct times have been shown to be required for the acquisition of specific neuronal identities. For instance, the well-characterized NB7–1 lineage consists of 30+ neurons, including 5 Eve+ neurons (U1–U5) and their siblings, and subsequently 20+ interneurons. Analogous to the mammalian cortical neurons, the earliest-born U1 neurons reside in the deepest position and project a longest distance to their target muscle, whereas the latest-born U5 neurons reside superficially and send their processes only proximally. The neuroblasts that produce the 5 Eve+ distinct types express Hunchback (Hb), Hb/Kruppel (Kr), Kr, Pdm, and Pdm/Castor (Cas). It has been demonstrated by removing the function or overexpressing these genes that both Hb and Kr are necessary and sufficient to determine the first-born and second-born cell fates, respectively. Furthermore, overexpression of Hb induces Kr, and similarly expression of Kr induces Pdm, showing that sequential expression of these transcription cascades is tightly regulated to produce these distinct neuron types in an invariant order. It has also been demonstrated that the sequential expression of fate-determining genes is largely cell intrinsic, as the expression occurs in isolated neuroblasts in culture and even can partially progress in the absence of cell cycle progression (Kr-Pdm-Cas). Although corresponding mammalian homologs of Hb (Ikaros family genes), Pdm (SCIP/Oct-6), and Cas (Casz1) exist, the contribution of these factors to temporal lineage progression is not equivalent (SCIP) or has not been addressed (Ikaros, Casz1). It is also noteworthy that in *Drosophila*, the temporal progression in neuroblast identity is dependent on transcriptional factors that positively promote neuroblast identity. This is in contrast with the model in the mammalian cerebral cortex, where at least during the earliest corticogenesis, cell fate is regulated by a repressive mechanism by which a transcriptional repressor, *Foxg1*, prevents deep-layer neurons from adopting an earliest-born CR cell fate. The extent to which future analysis will reveal further similarities between mammalian cortical neurogenesis and that seen in the *Drosophila* nerve cord is presently unclear. Regardless of the precise details, it seems likely that the study of *Drosophila* neuroblasts will prove informative in helping us understand how cortical progenitors generate the diversity seen in the mature cortex. Furthermore, it seems inevitable that considerable progress will be forthcoming concerning the specific transcription cascades involved in generating cerebral cortex neuron diversity, as well as with regard to the mechanisms that ensure symmetric and asymmetric cell divisions that allow the production of the appropriate numbers of neurons at the appropriate times.

See also: Cerebral Cortex: Inhibitory Cells; Cerebral Cortex; Memory Consolidation: Cerebral Cortex; Neocortex: Origins; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Radial Glial Cells: Brain Functions.

Further Reading

Chenn A and McConnell SK (1995) Cleavage orientation and the asymmetric inheritance of Notch1 immunoreactivity in mammalian neurogenesis. *Cell* 82: 631–641.

Desai AR and McConnell SK (2000) Progressive restriction in fate potential by neural progenitors during cerebral cortical development. *Development* 127: 2863–2872.

Fishell G and Kriegstein AR (2003) Neurons from radial glia: The consequences of asymmetric inheritance. *Current Opinion in Neurobiology* 13: 34–41.

Hanashima C, Li SC, Shen L, et al. (2004) Foxg1 suppresses early cortical cell fate. *Science* 303: 56–59.

Lu B, Jan L, and Jan YN (2000) Control of cell divisions in the nervous system: Symmetry and asymmetry. *Annual Review of Neuroscience* 23: 531–556.

McConnell SK and Kaznowski CE (1991) Cell cycle dependence of laminar determination in developing neocortex. *Science* 254: 282–285.

Mizutani K and Saito T (2005) Progenitors resume generating neurons after temporary inhibition of neurogenesis by Notch activation in the mammalian cerebral cortex. *Development* 132: 1295–1304.

Molyneaux BJ, Arlotta P, Hirata T, et al. (2005) Fez1 is required for the birth and specification of corticospinal motor neurons. *Neuron* 47: 817–831.

Noctor SC, Flint AC, Weissman TA, et al. (2001) Neurons derived from radial glial cells establish radial units in neocortex. *Nature* 409: 714–720.

Noctor SC, Martinez-Cerdeno V, Ivic L, et al. (2004) Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. *Nature Neuroscience* 7: 136–144.

Pearson BJ and Doe CQ (2004) Specification of temporal identity in the developing nervous system. *Annual Review of Cell and Developmental Biology* 20: 619–647.

Reid CB, Tavazoie SF, and Walsh CA (1997) Clonal dispersion and evidence for asymmetric cell division in ferret cortex. *Development* 124: 2441–2450.

Tarabykin V, Stoykova A, Usman N, et al. (2001) Cortical upper layer neurons derive from the subventricular zone as indicated by Svet1 gene expression. *Development* 128: 1983–1993.

Zhong W, Feder JN, Jiang MM, et al. (1996) Asymmetric localization of a mammalian numb homolog during mouse cortical neurogenesis. *Neuron* 17: 43–53.

Zimmer C, Tiveron MC, Bodmer R, et al. (2004) Dynamics of Cux2 expression suggests that an early pool of SVZ precursors is fated to become upper cortical layer neurons. *Cerebral Cortex* 14(12): 1408–1420.

Cerebral Fissure Patterns

J H Kaas, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The cerebral hemispheres of all mammals are smooth in early development, but as they enlarge, they become folded in many species. The folding of neocortex tends to increase with the size of the brain during development within a species and with the size of the mature brain across taxa. The folds are called fissures or sulci, whereas the crests of folds on the surface of cortex are called gyri. Each species with fissured cortex has a characteristic pattern of fissures, although there is some variability, and elements of this pattern are apparent across groups of related species. These observations can be made without modern equipment or methods of investigation, so they have intrigued neuroanatomists for some time. The explanations for the folding, and the folding patterns, have been of two types: mechanical and functional. Mechanical explanations focus on the growth of the cortical sheet relative to its subcortical core, which seems to require folding as brains enlarge, and other factors that could influence the positions of sulci, such as the overall shape of the brain and the constraints on directions of growth imposed by parts of the brain and by the shape of the intracranial cavity. The functional explanations are not truly functional since brain functions depend on neurons and their interconnections, and not on the folds of the cortical sheet. Such explanations relate to the fact that the cortical sheet is not uniform in function but is divided into a variable number (across taxa) of cortical areas that interrelate in different ways. The functional differences from one cortical area or region to another depend on structural differences in neuron types, numbers, and connections. These differences affect the differential growth of areas in surface area and thickness. All such structural distinctions that stem from functional specializations potentially impact on the way the cortex folds to form sulci and gyri. Even small perturbations of the growing cortical sheet, in principle, can have large effects on the locations of convolutions. Functional theories of folding focus on the process of regionalization of cortex and how this process relates to the folding pattern.

Both mechanical and functional theories appear to explain some of the fissure patterns observed in brains. Fissure patterns explained by mechanical theories tell us little about the functional organization of the brain, whereas fissure patterns that are a consequence of the regionalization of the brain have the potential to tell us something about how the cortex is organized. In both instances, the major fissure patterns have a relatively fixed relationship to the locations of functional areas in the brain of any particular species and can be used as a guide to where functional areas are located. Thus, the positions of sulci are useful guides in physiological and anatomical studies of cortex and also in neurosurgery and other clinical applications. However, only sulci related to brain regionalization can tell us something about the functional organization of brains where no experimental or clinical study is possible, as is the case with the brains of extinct species viewed as skull endocasts. In particular, those of us curious about our own evolutionary history need to consider that ours is the only surviving species of the radiation that followed the divergence of our hominin ancestors from those leading to modern-day chimpanzees and bonobos approximately 5–6 Ma. We know from the endocasts of the skulls of extinct hominin species that brains increased tremendously in size over that time. Yet, we know almost nothing about the course of the functional changes that occurred in the evolution of the human brain during those 5–6 My. However, it seems likely that we can and will learn something of functional significance from the positions of fissures in the brains of our long extinct ancestors since these fissures are sometimes apparent in the endocasts of fossil skulls.

Changing Patterns of Fissures with Cortical Development and Evolution

The cerebral hemispheres are initially smooth as they emerge in development, and they gradually develop those fissures that characterize the mature brain. The human brain starts to develop fissures approximately 4 months after conception and a nearly adult pattern is present at birth (Figure 1). After birth, the brain more than doubles in size, fissures increase in depth, and small additional fissures may appear. However, the proportion of cortex hidden in fissures is highest at birth, as a small, gradual decrease occurs during the next 20 years. No further decrease occurs with subsequent aging. In contrast, the newborn ferret, which is extremely immature at birth, has none of the fissures of the adult. Fissures appear largely between postnatal days 6 and 14. Other examples of the development of fissures in various species of mammals are illustrated by Welker.

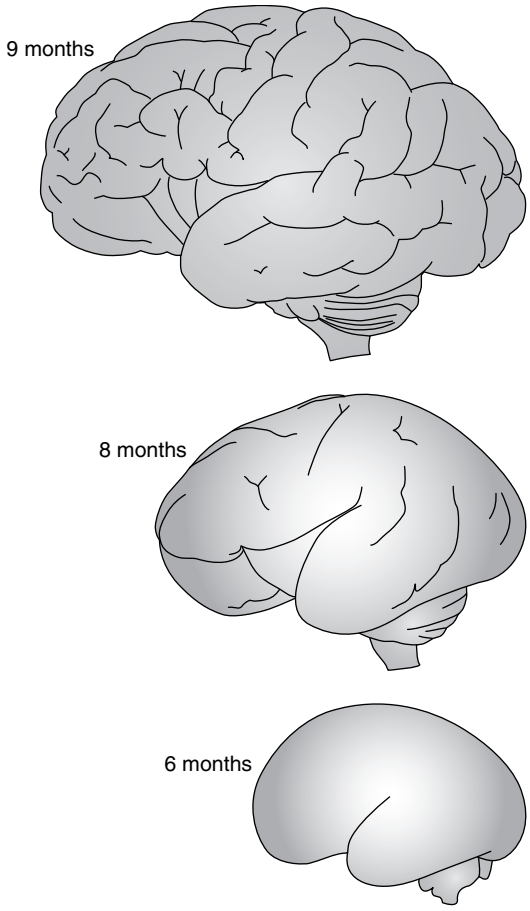

Figure 1 The development of fissures in the human brain from 6 months of gestation until birth (9 months), when all the major fissures are present.

Some folding appears to occur before the arrival of thalamocortical connections, and certainly before the development of cortico-cortical connections. The progressive emergence of fissures continues during the time when these connections are made and mature.

The brains of members of the same species have very similar patterns of fissures, although there is some variation. This variation is greater with minor, shallow fissures and in large brains with many fissures. Furthermore, variation in the depth of a fissure may be offset by countervariations in the depths of adjacent fissures – a counterbalancing that has been termed compensation. Because of the different specializations of the two hemispheres of the human brain for language and other functions, the patterns of fissures are not identical. Most notably, the lateral fissure, near and including cortex subserving audition and language, has a different course and extent in the two hemispheres. Additionally, fissure patterns in either the right or the left cerebral hemisphere of humans are somewhat variable from individual to individual, although less so in monozygotic (identical) twins than unrelated individuals. In macaque monkeys, there is evidence that both brain size and sulcus length are significantly heritable.

Fissure Patterns and Brain Size

Related species of mammals of different brain sizes may have some fissures in common, but the species with larger brains have more fissures. For example, the small-brained weasel or ferret has only a few fissures; the larger domestic cat has more, whereas the large-brained lion has the most (Figure 2). Thus, all these carnivore brains are characterized by an inverted U of a suprasylvian sulcus, enclosing cortex that is divided by a pseudosylvian sulcus (called pseudosylvian because it is the comparable position of the Sylvian or lateral sulcus of primates, but it is not the same (homologous) sulcus). Primates vary from the almost smooth brain of the small-brained marmosets to the extremely fissured brains of humans, but all primates have a lateral fissure and a calcarine fissure. The main conclusions from such observations are that within taxa of related mammals with fissures, a basic pattern of fissures is often evident, but species with larger brains have more fissures. However, given this general relationship between numbers of fissures and size of the brain, it is important to note that there is considerable variation within this relationship. Thus, the rather large cerebral hemispheres (brains are approximately 400 g) of the Florida manatees have only a lateral fissure, a splenial sulcus, and a few shallow clefts. Overall, the brain is largely smooth (lissencephalic). Beavers provide another example since they are rather large rodents with a large (for a rodent) smooth brain, whereas other rodents with the same size brains have fissures. Likewise, the brains of Prosimian primates have fewer and less deep fissures than the brains of comparable size of Anthropoid primates. On the other hand, the extremely small carnivore brain of the least weasel has all of the fissures of the brains of larger species of weasels (Figure 2(a)), although these fissures are very shallow. These differences indicate that there are other factors, in addition to brain size, that influence the formation of fissures.

In general, brains acquired more fissures as mammals with larger brains evolved in various lines of descent. We are able to identify the positions of brain fissures in the endocasts of the skulls of extinct mammals because the skull cavity conforms closely to the shape of the brain. Small mammals with small brains were generally the first representatives of the major orders of the mammalian radiation, and larger members of the orders with larger brains came later. For example, as horses evolved from rather small

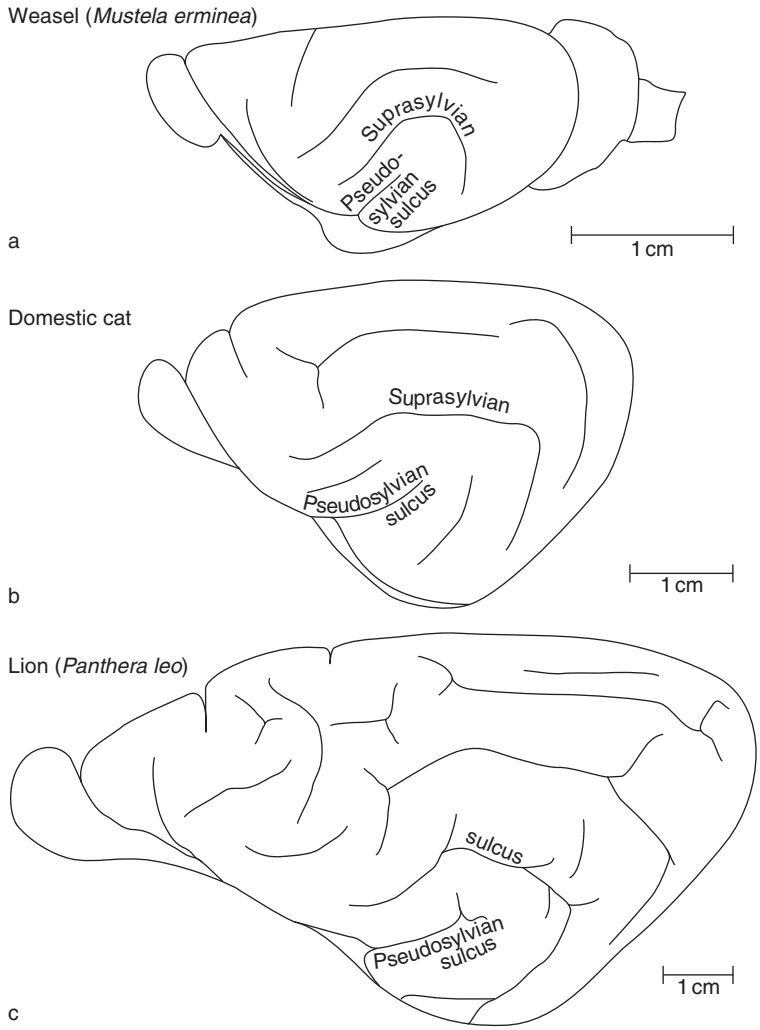

Figure 2 Dorsolateral views of the left cerebral hemispheres of three species of carnivores with brains of different sizes. All three brains contain some fissures that are characteristic of carnivore brains, including the suprasylvian and pseudosylvian sulci identified here, but the larger brains have more fissures. (a) The small-brained weasel has a cerebral cortex with only a few fissures. (b) The larger brain of the domestic cat has a few more fissures. (c) The much larger brain of the lion has even more fissures, and these fissures often have side branches.

ancestors, the course of evolution of these ungulates to present-day horses was characterized by the transformation of an almost smooth small brain to a large one with a complex pattern of many fissures. Similar increases in number of fissures occurred with the evolution of large brains in carnivores and primates.

Names of Fissures and Possible Homologies

The naming of sulci and gyri by various investigators has been generally based on physical features, such as location, orientation, and shape. A few have been named after investigators, such as Sylvian or lateral sulcus, and the Rolandic or central fissure. Fissures of the same relative locations in groups of related species are usually given the same name, implying that they are homologous (i.e., that they were present in the common ancestor of all species with those fissures, and that the fissures were retained in subsequent generations). Some investigators have attempted to homologize fissures that are similar in location across members of major taxonomic groups, such as between dogs and Old World macaque monkeys. However, such proclamations are not valid because the fossil record clearly indicates that the rather distant common ancestor of carnivores and primates was smooth brained (lissencephalic). Thus, the later emergence of fissures in both taxa cannot be homologous in the usual sense of the term (i.e., retained from a common ancestor). In a similar manner, New World cebus monkeys have large brains with a pattern of fissures that closely matches the pattern found in Old World macaque monkeys. Yet, the common ancestor of New

and Old World monkeys presumably had few fissures in cortex, probably only the lateral and calcarine fissures. However, there could be retained features of brain organization that led to the evolution of similarly located fissures as separate lines of evolution led to bigger brains. In this hypothetical case (functional theories of fissure formation are discussed later), the separate emergence of fissures in comparable locations would not be completely independent and could reflect shared patterns of gene expression in the brain or a 'deep' homology.

Experimental Manipulation of Fissure Formation

Attempts to experimentally alter the formation of sulci in developing brains have involved placing lesions in cortex or subcortical structures at various stages of fetal development or, recently, bilateral enucleation in fetal stages, which alters the development of visual cortex. In an early attempt to experimentally evaluate theories of fissure formation, Barron lesioned cortical and subcortical structures in the developing brains of fetal sheep before fissure formation and examined the brains at birth. Because a reduction in the volume of brain substance did not reduce the amount and general appearance of fissures in the remaining cortex, Barron concluded that the validity of one theory, that the skull restrains the brain and initiates the cortical folding, is unlikely. Because fissures also formed after the thalamus, corpus callosum, and basal ganglia were substantially damaged, Barron further concluded that cortical folding was not primarily dependent on a disproportionate growth between subcortical and cortical structures nor on the connections between subcortical and cortical structures. Thus, factors within the cortical sheet were the only remaining primary source of fissure formation. Later, Welker reported the results of placing cortical and subcortical lesions in the brains of neonatal cats and raccoons at approximately the time when fissures were starting to form and found that subcortical damage had a more limited effect. Although cortical lesions did not prevent the formation of fissures, they did alter their form. The remaining cortex grew and shifted toward the lesion site, and the fissure pattern became altered so that sulci tended to reorient toward the defect. These results were similar to the effects of cortical lesions in fetal macaque monkeys, again before fissures and cortical connections emerged. After such lesions, fissures did develop. However, there were abnormal sulci, especially near the lesion sites but also in more distant cortex and to a small extent in the other hemisphere. Thus, cortical lesions appear to have more effect on the development of cortical fissures than suggested by Barron. There may be both mechanical effects that distort the fissure pattern as cortex grows into the region of the lesion and other more distant effects that may relate to the altered connections in cortex as the result of the loss of some cortical tissue.

Another way to alter the course of cortical development and effect the formation of fissures has been to damage the developing visual system by removing the eyes in fetal macaque monkeys. The lack of retinal inputs to the lateral geniculate nucleus and the superior colliculus reduces their growth, and the lack of a relay of activation from the retina to the cortex via the lateral geniculate nucleus reduces the growth of primary visual cortex. Thus, the primary visual cortex at birth was smaller than normal, and the connection pattern of this reduced cortex with other areas of visual cortex undoubtedly was altered as well. The alterations in the visual system produced a deepening of minor dimples in the cortex of the occipital lobe, the formation of supplementary sulci, and an overall increase in the amount of convolution. Thus, changes in the sizes of cortical areas and their connection patterns can alter cortical fissure patterns.

These experimental findings are, to some extent, supported by the effects of abnormal brain development of fissure patterns in humans and other mammals. Although the causes of most of the cases of abnormal development, such as failure of the corpus callosum to form or the failure of the brain to grow to normal size, are unknown, such defects likely involve alterations in the genesis, migration, and connections of neurons. In human cases of microgyria, in which large numbers of shallow fissures form, cells fail to migrate to the deeper cortical layers, whereas in lissencephaly, in which only primary fissures form, cells in the more superficial layers are missing and cortical growth is reduced. In cases of abnormal brain development, whatever the cause, fissure patterns are typically abnormal.

Mechanical Theories of Fissure Formation

Many investigators agree that the fundamental reason fissures often form as brains grow is that the cortex remains relatively thin and grows as a sheet, increasing more in surface area while the underlying subcortical core of nuclei, ventricles, and white matter grows as a volume. In addition, the cortex is generated later and it grows disproportionately more with brain size than subcortical structures. As a result, the surface of the cortex expands faster than the surface of the core so that the cortex ends up 'shrink-wrapped' and folded over the core. Cortical

surface area varies by a factor of approximately 1000 across mammalian species, but the thickness of cortex varies by only a factor of approximately 5. The constraints on cortical thickness may be functional, stemming from the design of the local microcircuitry of the six-layered cortex, or related to metabolic requirements and a surface-based blood supply. Constraints may also be mechanical, such as the radial tensions resulting along the apical dendrites of cortical pyramidal cells as proposed by Van Essen. Other radial tensions in cortex result from radial glial and axonal fibers. Assuming cortical elasticity and plasticity along with radial tensions, the formation of cortical fissures can be modeled in a way that results in a fairly fixed number of convolutions of similar depth, with cortex being thicker in gyri than sulci. For sulci to vary in depth and placement in the model, additional factors need to be added, such as geometric asymmetries associated with the initial shape of cortex, local differences in cortical structure, and asymmetries in growth. For example, small asymmetries in the shape of the rapidly growing cortex would induce the development of convolutions at the point of distortion and trigger a wave of even more distant convolutions. There is some support from comparative studies for this prediction of the model. In many species in which the initial shape of cortex is more like a rostrocaudally elongated tube, fissures develop with a rostrocaudal orientation, whereas brains that are short and wide tend to develop transverse fissures. Thus, differences in the initial shape of the cerebral hemispheres influence the orientations of the fissures that develop. Another prediction of the model is that structural asymmetries, introduced by steps or gradients in the elasticity of cortex, tend to introduce fissures at or near the borders of change. As noted previously, cortex is not uniform in function or structure, and these local steps or gradients in cortical structure could trigger the formation of sulci. Because this mechanical effect is related to function, this possibility is discussed further in the next section on functional theories of fissure formation. Growth asymmetries introduced in the model also affect the formation of fissures in that convolutions emerge in the region of greater growth. This factor also is related to the functional regionalization of cortex and is discussed further later.

Several specific mechanical factors have been discussed as possibly contributing to the formation of fissures. Although cortical thickness does not vary much, a thicker cortex could be more resistant to bending so that fewer but deeper fissures form. Brains with thicker cortex would also require less cortical surface area for the same volume of cortex, and thereby fewer fissures. Perhaps the best support for these suppositions is that manatees have a thick cortex of approximately 4 mm, far more than most mammals, while having a highly lissencephalic brain. Also claustrocortex, a region of cortex where the deepest layer is closely approximated by the claustrum, is effectively thicker than other cortex as the claustrum acts as a seventh layer. Across species, deep folding in claustrocortex is absent. As another factor, Le Gros Clark and others have postulated that the constraining shape of the skull influences the formation of fissures. This conjecture has been widely dismissed because it appears that the growth of bones of the cranial vault is responsive to the underlying brain growth, and partial removal of the skull during development does not have a dramatic effect on the fissure pattern. Yet, some constraints, such as those on the relative width and length of the brain or the bending of the brain, likely exist and have some influence. Le Gros Clark also proposed that pressure from subcortical parts of the brain induces the formation of fissures. For example, Le Gros Clark suggested that the calcarine fissure in visual cortex of primates is induced by pressure from the underlying tectum of the midbrain causing a depression. Although it is difficult to find support for this view, it may be a factor. Le Gros Clark also argued that the joining of the two hemispheres by the formation of the corpus callosum provided a mechanical stress during cortical growth that would induce longitudinal folds to emerge in parallel to the corpus callosum and perpendicular to the lines of compression, especially near the corpus callosum. Several observations support this conclusion. Most notably, a corpus callosum does not develop in marsupials, and longitudinal fissures, especially on the medial wall of the cerebral hemisphere, are uncommon. In addition, in such rare cases in which the corpus callosum fails to develop in the human brain, sulci on the medial surface of the hemisphere tend to be vertical rather than longitudinal.

In summary, the sulci of the cortex are the result of mechanical factors that perturb the developing, ballooning cerebral hemisphere. Such factors as the constraints of the skull, subcortical structures, and the corpus callosum are considered here as mechanical causes that are largely or completely unrelated to the functional organization of the cortex. Mechanical causes that reflect the functional organization of cortex are considered next.

Functional–Mechanical Theories of Fissure Formation

Three features of the developing functional organization of the neocortex appear to have mechanical consequences that influence the formation of fissures.

First, the large functional units of cortex, the cortical areas, differ from each other in how they connect to other areas and to subcortical structures. As Van Essen forcefully argued, the axonal connections between areas generate mechanical tension, a force for inducing cortical folds. Since cortical areas differ in connections, and the cortical sheet is not uniform in connections, the regional specialization of this sheet impacts on the location and form of cortical fissures. Second, because of the regionalization of cortex, growth is uneven. Some cortical areas become thicker or more extensive than others do as they disproportionately grow in development, whereas other areas develop laminar specializations that may have mechanical consequences. These regional factors all have the potential to influence fissure formation. Third, cortical areas sometimes vary internally in structural organization. For example, primary visual cortex, area 17, is thinner in the portion that represents monocular vision than it is in the part that represents binocular vision, and greater cortical thickness resists fissure formation. In addition, many cortical areas are modularly organized, and this provides a potential influence as types of modules differ in cellular structure and in connections. Finally, cortical areas with an orderly representation of a sensory surface may have sharp disjunctions where nonadjacent parts of the receptor sheet are represented next to each other. In such instances, a segregating septum often emerges between the two parts of the representation that has different mechanical properties and thereby initiates the formation of a sulcus. The locations of such fissures that have mechanical causes, while reflecting the functional organization of cortex, can tell us something about cortical organization in those species for which experimental evidence is missing, and they can guide experimental studies where the relationship of fissures to functional areas has been established.

In general, regional differences in cortical structure have not been considered as important in the formation of fissures because cortex has been assumed to not vary much in thickness, cell density, and cell type. However, motor cortex is often thicker than visual cortex, visual cortex sometimes has a broad layer 4 with clear subdivisions, and motor cortex typically lacks an identifiable layer 4. Junctions between areas of different thicknesses are considered to be a favorable place for bends in the cortex. Bends are also likely in the middle of thinner areas that are anchored by thicker areas. Also, the lateral growth of superficial cortical layers relative to deep cortical layers has been postulated as a factor in the production of bulges and bends. Developmental and evolutionary changes in the sizes of some cortical areas relative to others would also likely alter fissure patterns. As previously noted, primary visual cortex (area 17) does not grow to a normal size after early prenatal loss of eyes in monkeys, resulting in an altered pattern of fissures in occipital cortex. Because the size of area 17 in humans and chimpanzees is approximately the same, although the size of the cortical sheet in humans is much bigger, this proportional difference likely accounts for some of the differences in occipital fissures in the two taxa. The expansion of prefrontal cortex relative to other parts of the frontal lobe in the evolution of the human brain also likely had consequences.

The argument that the connection patterns that form between cortical areas during development have a major effect on the locations of fissures has been most effectively made by Van Essen. His 'tension-based' hypothesis of cortical folding holds that especially the long-distance connections (as opposed to intrinsic connections) that emerge in developing cortex generate mechanical tension that resists the expansion of the cortical sheet and causes buckling that results in gyri. Thus, if two adjoining areas become strongly interconnected, the connections will resist spatial separation of the areas as cortex grows, and a buckle will form near the border between the areas, forming a gyrus so that the areas become closely approximated along opposite banks of the gyrus. A fortuitous consequence of this arrangement is that the connections between the two areas end up much shorter than in flat or otherwise folded cortex, and the conduction times for axons and the axon lengths (hence, axon volume) become less. Thus, computations are faster and less brain tissue is needed – both clear advantages. Possibly such advantages have influenced the evolution of brain factors that initiate fissures that reduce important connection pathways, but there is no evidence of this, and the tension theory does not require it. As a key example of a fold that is consistent with the tension theory, the lunate sulcus that forms in many of the anthropoid primates folds the second visual area, V2, under the primary visual area, V1 or area 17. These two areas are densely interconnected, and the folding shortens the lengths of these connections. In contrast, the tension theory predicts that fissures develop between areas with few interconnections. Because there would be little resistance to folding between two areas with few interconnections, depressions and then fissures tend to develop and separate such areas. The key example for this in the larger brained anthropoid primates is the central sulcus, which separates primary motor cortex, M1, from primary somatosensory cortex, S1 or area 3b. These two areas have very few interconnections because most somatosensory inputs to M1 are from higher order somatosensory fields. Thus, there are few connections to create binding tensions between M1 and S1, and a fold stretches those few connections even

further. Even though the consequences of such tensions from axons on the development of fissures have not been explored to the point of allowing a careful evaluation of the theory, the implications are clear. Many, and perhaps most, of the cortical fissures reflect the functional organization of the cortex and can be used as guides to determining the functional organization.

Finally, a specific type of cortical fissure develops within primary somatosensory cortex of some mammals and perhaps in other sensory areas as well. The skin, the receptor sheet for the cutaneous receptors of the somatosensory system, differs from the receptor sheets for vision (retina) and audition (cochlea) in that there are obvious splits, tears, and protrusions in the sheet. Thus, the upper lip is separated by the mouth from the lower lip, as are the digits from each other. In addition, the protruding limbs separate the receptors in the hand and foot from those on the body. Yet, the central representations of the cutaneous receptors have a compact form, especially in the cortex of primates, in which the representation of the contralateral skin receptors is in the form of a long mediolateral strip. This brings representations of digits next to each other and the representation of the thumb (D1) next to that of the lower face. In general, the intrinsic connections within a cortical area are short, but they are most dense along representations of uninterrupted receptor populations, and they are sparse between adjoining representations of separate receptor populations. Thus, there are few connections between the adjoining representations of the hand and lower face in primates and denser connections along representations of the phalanges of the same digit. Moreover, narrow zones of cortex between adjoining representations of separated populations of receptors receive few thalamic inputs and have smaller and perhaps fewer neurons. These septal zones are visible histologically and appear to emerge because of the segregation of thalamocortical afferents according to their source of activation as they compete for cortical space. We call these zones segregating septa because they separate parts of representations that have little interaction. Because these septa are structural perturbations in the cortical sheet, one would expect segregating septa to sometimes initiate the formation of sulci, and they do. Those segregating sulci have been called limiting sulci by Welker, and they are illustrated and described extensively in his monograph. In raccoons, which have a large representation of the hand in S1, small sulci separate the representations of the digits from each other, whereas deeper sulci separate the representation of the hand from the face and the hand from the adjoining representation of the foot. In other carnivores, fissures separating representations of the forepaw from the face, and the forepaw from the foot, are common. Similar segregating sulci occur in S1 of some marsupials, primates, ungulates, and rodents. If we knew more about the organization of S1 in other taxa, undoubtedly there would be more examples. However, these examples firmly establish the principle. Segregating septa within cortical representations are structurally different from adjoining cortex within the area, and this structural difference can induce a fold. Consistent with the tension-based theory of Van Essen, segregating fissures are between regions of cortex with poor interconnectivity.

It is unknown if cortical areas other than S1 have segregating septa that lead to fissures. The star-nosed mole, with its 22 appendages of its nose, has 11 septa for the 11 rays on the contralateral face in S1 and the second area, S2, of somatosensory cortex, but its small brain is smooth. However, the obvious segregating septa in S2 indicate that segregating sulci may be identified in cortical areas other than S1. In the visual system, the lack of photoreceptors in the location of the nerve head (optic disk) of the retina results in a structural modification of a central part of primary visual cortex in which nearby neurons represent receptors on opposite sites of the optic disk. This structural change in cortex is similar to a segregating septum, but a role in the formation of fissures has not been established. Yet, primary visual cortex in primates does fold down the middle, through the structurally modified zone reflecting the optic disk, and the structural difference together with a reduced local pattern of horizontal intrinsic connections could be initiating factors for the fissure.

Fissure Patterns in the Cortex of Extinct Mammals

There are a number of inferences that can be made about the functional organization of the brains of extinct mammals from the position of sulci on skull endocasts. One possibility is to use the positions of sulci to estimate the sizes of lobes of the brain. For example, the central sulcus is well developed in most anthropoid primates, and this sulcus consistently separates somatosensory from motor cortex, thereby separating frontal from parietal cortex (the frontal lobe from the parietal lobe). Since the frontal lobe contains motor and premotor areas, as well as the prefrontal cortex (important for foresight and planning in social primates) and orbital frontal and cingulate limbic cortex (important for evaluating stimuli and situations), the relative sizes of the frontal lobe in our hominin ancestors provide an important clue about the level of social cognition as well as other abilities. However, the frontal lobe is not proportionally larger overall in humans

than in apes, possibly because more of the human frontal lobe is devoted to prefrontal cortex than motor areas. Nevertheless, the relative proportions of the frontal lobe in extinct taxa of hominins remain a potential source of comparative information on frontal lobe functions. In a similar manner, the position of the lunate sulcus, which is located near the junction of primary and secondary visual cortex in some extant anthropoid primates, has been used (with some disagreement) to locate the rostral border of primary visual cortex as similar to that of modern humans in the brains of the earliest hominins.

Another possibility is to search for asymmetries in the patterns of sulci in the two hemispheres to evaluate when certain brain specializations may have emerged. As mentioned previously, in humans, the lateral sulcus is typically higher in the right cerebral hemisphere than in the left, and this difference has been attributed to the disproportionate expansion of auditory and cognitive areas in the left hemisphere that are associated with language. There are a few suggestions of this asymmetry in fossil hominins, possibly for language but more likely for the cognitive skills that eventually made language possible. Convolutions and protrusions have also been used to estimate the size of Broca's area, a portion of the frontal lobe that is critical in the production of speech. Because the region is larger in the left than the right hemisphere of modern humans, it is interesting that a Broca's region asymmetry with the larger size on the left has been noted in fossil hominin endocasts. Again, the asymmetry does not indicate that language was present but, rather, that neural systems compatible with language possibly existed. Overall, these hints about the functional organization of the brains of early hominins suggest an organization that was already more humanlike than the brains of modern-day chimpanzees.

There is also something to be learned from the positions of the sulci that are likely to have been segregating fissures in fossil brains. These are the fissures in somatosensory cortex that separate adjoining representations of skin surfaces that are partially or largely separate on the body. Evidence of segregating fissures in endocasts of fossil brains may be difficult to obtain because these fissures in extant mammals are often shallow. However, early efforts by Radinsky provided evidence that some extinct carnivores had enlarged representations of the face (whiskers) or the forepaws, suggesting different specializations of the somatosensory system. Further efforts to identify limiting sulci in extant and extinct mammals would be useful.

See also: Brain Asymmetry: Evolution; Brain Development: The Generation of Large Brains; Brain Modules: Mosaic Evolution; Brain Scaling Laws; Broca's Area: Evolution; Cerebral Cortex; Hemispheric Specialization and Cognition.

Further Reading

Armstrong E, Schleicher A, Omran H, Curtis M, and Zilles K (1995) The ontogeny of human gyrification. *Cerebral Cortex* 1: 56–63.

Baron DH (1950) An experimental analysis of some factors involved in the development of the fissure pattern of the cerebral cortex. *Journal of Experimental Zoology* 113: 553–573.

de Sousa A and Wood B (2007) The hominin fossil record and the emergence of the modern human central nervous system. In: Kaas JH and Preuss TM (eds.) *Evolution of Nervous Systems, Primates,* vol. 4, pp. 291–336. London: Elsevier.

Dehay C, Giround P, Berland M, Killackey H, and Kennedy H (1996) Contribution of thalamic input to the specification of cytoarchitectonic cortical fields in the primate: Effects of bilateral enucleation in the fetal monkey on the boundaries, dimensions, and gyrification of striate and extrastriate cortex. *Journal of Comparative Neurology* 367: 70–89.

Galaburda AM, LeMay M, Kemper TL, and Geschwind N (1978) Right–left asymmetries in the brain. *Science* 199: 852–856.

Goldman-Rakic PS (1980) Morphological consequences of prenatal injury to the primate brain. *Progress in Brain Research* 53: 3–19.

Le Gros Clark WE (1945) Deformation patterns on the cereberal cortex. In: Le Gros clark WE and Medawar PB (eds.) *Essays on Growth and Form*, pp. 1–23. Oxford: Oxford University Press

Radinsky L (1976) Cerebral clues. *Natural History* 85: 54–59.

Rakic P, Suner I, and Williams RW (1991) A novel cytoarchitectonic area induced experimentally within the primate visual cortex. *Proceedings of the National Academy of Sciences of the United States of the America* 88: 2083–2087.

Reep RL and O'Shea TJ (1990) Regional brain morphometry and lissencephaly in the Sirenia. *Brain, Behavior and Evolution* 35: 185–194.

Richman DP, Stewart RM, Hutchinson JW, and Caviness VJJ (1975) Mechanical model of brain convolutional development. *Science* 189: 18–21.

Todd PH (1982) A geometric model for the cortical folding pattern of simple folded brains. *Journal of Theoretical Biology* 97: 529–538.

Toro R and Burnod Y (2005) A morphogenetic model for the development of cortical convolutions. *Cerebral Cortex* 15: 1900–1913.

Van Essen DC (2007) Cerebral cortical folding patterns in primates: Why they vary and what they signify. In: Kaas JH and Preuss TM (eds.) *Evolution of Primate Nervous Systems, Primates,* vol. 4, pp. 216–276. London: Elsevier.

Welker WI (1990) Why does cerebral cortex fissure and fold? In: Jones EG and Peters A (eds.) *Cerebral Cortex,* pp. 3–136. New York: Plenum.

Cerebrovascular Disease

C S Ogilvy and A P Mitha, Massachusetts General Hospital, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Occlusive Diseases and Stroke

Of patients who present to the emergency room with an acute onset of a focal neurologic deficit, typically 95% have problems caused by a vascular syndrome. Of these, 15% are hemorrhagic and may be caused by intracerebral hemorrhage, subarachnoid hemorrhage, or subdural hemorrhage, while 85% are cerebral infarction. Cerebral infarction in this setting is defined as an area of the brain not receiving adequate blood flow and therefore resulting in neurologic function and ultimate cell death. Of the patients who present with infarction, 41% have an unknown etiology; 21% are lacunar in nature, or small artery or arteriolar occlusions of the perforating vessels in the deep white matter or diencephalon; 16% are cardiogenic in origin with an embolus originating in the heart; 11% are caused by large artery (typically the carotid artery) atherosclerotic disease; and 11% are found to have tandem arterial pathology with atherosclerotic lesions of the carotid artery, vertebral arteries, or intracranial arteries.

Lacunar strokes are small infarcts in deep noncortical white matter and brain stem that result in the occlusion of penetrating branches of cerebral arteries. The term 'lacune' refers to the crescent- or moon-shaped tiny stroke that results from occlusion of these vessels. These lesions were seen on pathologic evaluation and now can be seen on magnetic resonance imaging (MRI) studies after infarction. These may be due to lipohyalinosis secondary to hypertension of small arterioles (on the order of 200 μm) and may also be the source of intracranial hemorrhage when the vessel weakens and the presenting diagnosis may be that of an intraparenchymal brain hemorrhage. Lacunar strokes tend to have a standard acute presentation with fairly rapid resolution depending on the exact size of the infarction. This tends to be a diagnosis of exclusion based on the neurologic syndrome and ruling out other sources of intracranial stroke.

In addition to thromboembolic occlusions, vessel dissection may be the etiology of stroke. Dissection occurs when the intimal layer of a blood vessel is damaged, allowing blood to collect between the intima and media, resulting in narrowing of the arterial lumen. This can result from trauma or neck manipulation, but may also arise spontaneously. In these patients, vertebral and carotid dissections may lead to distal emboli and cerebral infarction. Treatment of these patients typically involves anticoagulation with heparin, followed by longer-term anticoagulation with antiplatelet therapy or warfarin. The vessel injury tends to stabilize over time and in many individuals will in fact heal itself. There are situations in which vessel dissection and occlusion can occur, and urgent revascularization may need to be considered in these patients with either an endovascular or a surgical approach.

Patients with atherosclerotic disease of the internal carotid artery (ICA) may experience a transient ischemic attack (TIA), with or without amaurosis fugax, or a stroke. In many cases, these patients can be identified without symptoms. Once a carotid lesion has been identified, careful consideration should be given to whether it would be indicated to perform carotid endarterectomy or carotid artery stenting to accommodate revascularization.

Moya-Moya syndrome is a poorly understood condition in which there is a progressive arteriopathy with narrowing of the proximal intracranial vessels, and, in response to this, the distal vessels proliferate in density and give the appearance on cerebral arteriogram of a 'puff of smoke' or Moya-Moya vessels. This can occur in children who present more frequently with cerebral infarction, and in adults who present more frequently with intraparenchymal brain hemorrhage. For patients who present with TIA or infarction, surgical revascularization may be considered with either an extracranial-to-intracranial bypass graft or an encephalodural angiosynosis procedure (EDAS), in which the superficial temporal artery is placed on the brain's surface and as a result small vessels sprout to supply the brain tissue with vessels over time.

In addition to arterial causes of infarction, venous sinus thrombosis is a situation of hyperperfusion of an area of a brain, resulting in edema or hemorrhage. This can be seen with an increased incidence in certain disease states or with certain medications such as oral contraceptives. Patients with venous sinus thrombosis are typically managed with anticoagulation and, at times when extensive thrombosis is present, may require endovascular clot lysis.

For patients who experience acute onset of neurologic deficit and are found to have arterial occlusion in the brain, intravenous tissue plasminogen activator (tPA) may be indicated in an attempt to dissolve the clot. An additional strategy is to use endovascular techniques to directly manipulate and remove the clot and provide for revascularization. In both these situations, time is of the essence, as a longer interval

of occlusion results in a greater chance of cerebral infarction. For patients with arterial infarctions, it has been well defined that there is a 23% mortality rate, with 26% of patients discharged to a nursing home. Three percent of patients were discharged to a rehabilitation unit, and 48% of patients were able to return home with some degree of neurologic deficit or a complete recovery.

When patients present to the emergency room with a new neurologic deficit, a computed tomography (CT) scan should be obtained urgently. If there is no evidence of hemorrhage, then heparin therapy should be strongly considered. With current imaging techniques, it is also possible to perform computed tomographic angiography (CTA) to evaluate the patency of the vessels. In some institutions, MRI study is performed emergently as well in order to establish perfusion of brain tissue, as well as early changes in the parenchymal water content indicative of the extent of infarcted cerebral tissue.

Risk factors associated with cerebral infarction include hypertension, hypercholesterolemia, smoking history, diabetes, and family history. In patients with TIAs or infarction, risk modification should be accomplished, as well as cholesterol lowering and blood pressure management with medications.

Cardiac source of brain embolism accounts for approximately 20% of all cerebral infarction. While there is no specific neurologic feature that can diagnose these patients, a complete cardiac evaluation is indicated, including an echocardiogram, electrocardiogram, and Holter monitoring to evaluate for any arrhythmias. Depending on results of the evaluation, anticoagulation of these patients may be indicated, especially if an arrhythmia of atrial fibrillation is discovered.

Intracranial Vascular Diseases

Intracranial Aneurysms

Intracranial aneurysms tend to occur at division points of major arterial branches of the brain blood vessels. They are typically referred to as 'berry' aneurysms, based on their saccular nature. Fusiform aneurysms may occur, in which the entire diameter of the arterial wall is diseased and produces a diffuse widening of the artery. Saccular aneurysms can present with rupture, producing subarachnoid or intraparenchymal hemorrhage. Unruptured aneurysms may be discovered as incidental findings or as a result of mass effect-causing symptoms by pressing on sensitive neurological structures.

Ruptured aneurysms For ruptured aneurysms, early treatment (within 24–48 h) is usually indicated. At the time of a subarachnoid hemorrhage from an intracranial aneurysm, patients may experience 'the worst headache of their life,' have focal neurologic deficits, or present in coma. Approximately 15% of patients with subarachnoid hemorrhage die at the time of the hemorrhage, with a total of 45–50% mortality from ruptured aneurysms within 2–3 weeks of the initial hemorrhage. Of the survivors, one-third of the patients make a good to excellent recovery, while one-third of the patients remain impaired and the final third of patients have varying degrees of neurologic deficit but may return to normal levels of activity.

Once an aneurysm ruptures, there is a threat of rerupture. This occurs with an incidence of 4% in the first day after the initial hemorrhage, and approximately 1.5% per day thereafter up to 2 weeks, yielding a 26.5% chance of a second hemorrhage over this interval. To prevent rehemorrhage, early reobliteration with surgical techniques to clip the aneurysm and seal off the base of the aneurysm or endovascular techniques to place a coil device or other agents inside the aneurysm are typically utilized (**Figure 1**).

Following subarachnoid hemorrhage, vasospasm may occur. It is believed that the blood distributed in the subarachnoid space at the time of hemorrhage induces inflammation and constriction of the cerebral vessels. Vasospasm typically occurs 5–7 days after the subarachnoid hemorrhage and carries with it a risk of potential serious neurologic morbidity from infarction. With reduced diameter of vessels, low tissue perfusion may ensue, resulting in infarction. To treat vasospasm once the aneurysm is obliterated, blood pressure may be elevated to enhance cerebral perfusion. In addition, central venous hypervolemia may be utilized to assist with maximizing cardiac output in an attempt to maximize cerebral blood flow. Patients are typically hemodiluted to a hematocrit of 32% in an attempt to maximize blood flow by minimizing the viscosity of blood without impairing the oxygen-carrying capacity.

A final problem that may develop after subarachnoid hemorrhage is hydrocephalus. This can occur acutely and be managed with ventricular drainage at the time of presentation. Hydrocephalus may also occur in subacute or chronic form, which may require ventricular drainage or permanent ventricular shunting.

Unruptured aneurysms When patients are discovered to have aneurysms that have not ruptured, there is time for discussion regarding whether or not the lesion should be treated. This decision is often based on the size of the aneurysm (typically, lesions smaller than 4–5 mm are not treated), as well as the age and overall health of the patient. Endovascular and surgical techniques are used to obliterate

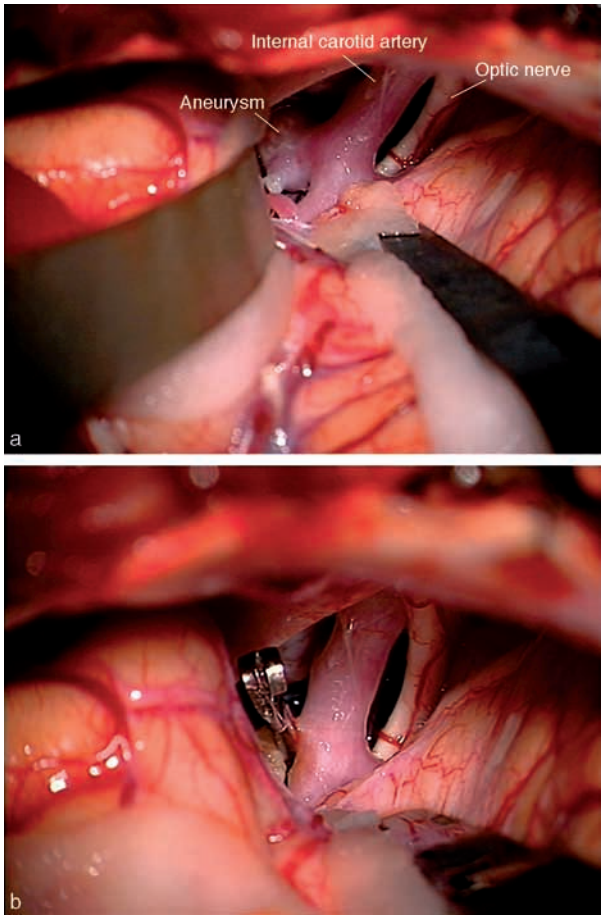

Figure 1 (a) Intra-operative view of an aneurysm arising at the junction of the left internal carotid artery and anterior choroidal artery. (b) The aneurysm is obliterated by placement of a surgical clip.

Figure 2 Angiogram (lateral view, left vertebral artery injection) of large occipital lobe arteriovenous malformation appearing as a tangle of vessels.

unruptured aneurysms; which technique is chosen is often based on the age of the patient, the location of the lesion, and the predicted risk of treatment. Aneurysms are typically given a size range of small (less than 8–9 mm), medium (10–15 mm), large (15–25 mm), or giant (greater than 25 mm). The risk of treating intracranial aneurysms does correlate to some extent with the size of the lesion as well as the age of the patient.

Vascular Malformations

Arteriovenous malformations Vascular malformations of the central nervous system are typically subdivided based on their histopathology and pathophysiology. True arteriovenous malformations (AVMs) of the brain are thought to be developmental in origin. These lesions are typically in the brain parenchyma and involve a tangle of arteries and veins with intervening brain tissue between these vessels (Figure 2). The vessels are not normal in their arterial and venous structures, and there is great variability in the histology of these structures. The brain tissue between these vessels may or may not be functional brain tissue, and this has to be carefully considered prior to recommending any form of treatment. AVMs typically present when patients have a hemorrhage from the lesion or a seizure. The lesions are also found incidentally. Once an AVM has been identified, careful consideration is given to treatment, based on balancing the risk of treatment against the risks of the natural history of the disease. Treatment modalities of open surgery, embolization followed by open surgery, and radiosurgical techniques are utilized to treat AVMs. The natural history of AVMs is such that once a hemorrhage from an AVM has occurred, the rehemorrhage rate is 6% in the first year after that hemorrhage and drops to approximately 4% per year thereafter. For patients who have not had a prior hemorrhage, it is felt that the hemorrhage rate is somewhat lower but occurs at a rate of somewhere between 2 and 4% per year for that individual. Therefore, the decision as to whether or not to treat an AVM must be made based on balancing these rupture rates against the risk of intervening to remove or radiate the lesion. As noted earlier, treatment modalities for AVMs include transarterial embolization, surgical resection, or radiosurgical techniques. Radiosurgery is a technique in which a focused beam of radiation is used to deliver a high dose of radiation to the AVM, thereby inducing an injury response that occurs over several years. Over this interval the vessels thicken and eventually obliterate on angiography.

Dural arteriovenous malformations Dural AVMs are thought to be acquired lesions. It is believed that they occur in response to a certain degree of venous thrombosis. Once a venous occlusion has occurred in or near one of the major draining sinuses of the brain, it is thought that angiogenic factors are produced which allow for development or opening of naturally occurring shunts from the arterial to the venous system on the dura. The dura is the leather-like structure that lines the inside of the skull and covers the brain. Dural AVMs can occur anywhere near a major sinus. A dural AVM may cause a bruit or may cause intracranial hemorrhage. Hemorrhage can occur when the venous drainage from the AVM occurs in a retrograde fashion through the normal cortical venous drainage and, therefore, exposes the normal cortical vein to high pressure. Such lesions often require urgent treatment in order to prevent further hemorrhage.

Dural AVMs are typically treated with transarterial or transvenous embolization with or without additional surgery to excise the lesion. If the dural AVM occurs in such a fashion that there is no cortical venous drainage, then they are thought to be a much more benign type of lesion and can be followed clinically or managed based on symptoms, such as to relieve a subjective bruit.

Cavernous malformations Cavernous malformations are lesions that can occur anywhere in the brain and spinal cord. They are best imaged by MRI and are typically occult on standard angiography (Figure 3). They may be familial in certain cases, and there have been several genes associated with familial cases of cavernous malformation (CCM 1, 2, and 3). Cavernous malformations are collections of sinusoidal vessels that are neither artery nor vein. The vessel lumens may thrombose over time or, because of the thin-walled vessels, may hemorrhage. The hemorrhages tend to be more benign than those from true AVMs since these are low-flow lesions under low pressure. While cavernous malformations can be excised surgically, it is believed that supratentorial cavernous malformations have a fairly low hemorrhage rate, and therefore older individuals are often managed expectantly. Cavernous malformations typically present with hemorrhage or seizure. Lesions in the thalamus and brain stem carry a much higher risk of surgical intervention but also may have a higher hemorrhage rate, or at least a higher risk of symptom-producing hemorrhage. This is probably true for the spinal cord lesions as well, and surgical resection must be carefully balanced against the risk of developing a neurologic deficit secondary to surgical intervention. While radiosurgery has been utilized with cavernous malformations, the end point is difficult to evaluate because the lesions do not typically obliterate after radiosurgical treatment. There have been several reports demonstrating a high incidence (25%) of complications after stereotactic radiation.

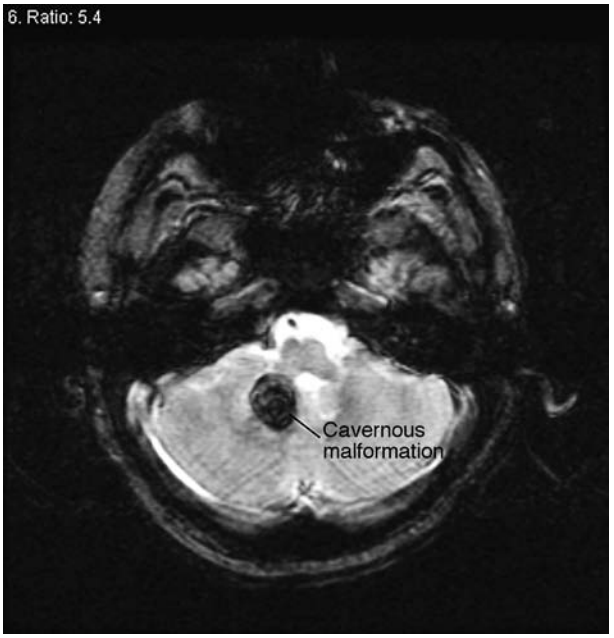

Figure 3 Magnetic resonance imaging gradient-echo sequence showing a right cerebellar cavernous malformation.

Developmental venous anomalies Approximately 15% of cavernous malformations occur in conjunction with developmental venous anomalies. Developmental venous anomalies are a collection of normal veins that coalesce to form one large draining vein but are not under any arterial shunting. These are thought to be fairly common in the general population (on the order of 2–5%). Developmental venous anomalies in and of themselves are believed not be at risk for rupture and may occur adjacent to cavernous malformations in what has been termed a 'collision lesion.' Typically, developmental venous anomalies should not be treated, as this removes the venous drainage from the area of the brain involved and may result in a venous infarction.

Capillary telangiectasias Capillary telangiectasias are small collections of dilated capillaries that can occur throughout the diencephalon and brain stem. Some authors believe that these lesions may be related to cavernous malformations but seem to have little, if any, potential for hemorrhage and often do not require any treatment. In addition, they are typically deep in the white matter, and therefore treatment would carry with it a substantial risk of neurologic deficit.

Spinal Vascular Diseases

Arteriovenous Malformations of the Spinal Cord

Dural arteriovenous fistulas Vascular malformations can also affect the spinal cord. The most commonly encountered type of vascular malformation of the spinal cord is the dural arteriovenous fistula. As with dural lesions in the intracranial compartment, dural arteriovenous fistulas in the spine are thought to be acquired lesions. They are typically small AVMs on the dural root sleeve that can occur throughout the thoracic and lumbar spinal roots. The venous drainage from these lesions is intradural and, therefore, exposes the spinal cord to elevated venous pressures. It is the elevated venous pressure that is thought to be responsible for the progressive neurologic dysfunction that occurs with these lesions. The extreme example of this is Foix-Alajouanine syndrome, in which there is progressive venous infarction and edema of the entire spinal cord, which in certain cases has resulted in complete tetraplegia and death. When these lesions are discovered early, they may be treated with embolization or, more typically, with surgical occlusion of the AVM on the dural root sleeve in a fairly straightforward fashion to obliterate the venous hypertension.

Glomus type More typical parenchymal AVMs may also occur in the spinal cord. The glomus type of AVM is a true arteriovenous malformation of the spinal cord parenchyma. These lesions often present with hemorrhage or progressive deficit and are very difficult to treat from the standpoint of surgical or radiation techniques. They tend to be high risk, and the incidence of postoperative deficit is high.

Juvenile type A juvenile type of spinal cord AVM is one that appears to be metameric in origin. It is thought to develop concurrently with the involved segment of spinal cord and associated body tissue and, therefore, involves not only the spinal cord but also the surrounding soft tissue and bone. These are typically diagnosed in children and can be quite extensive and difficult to treat.

Perimedullary The final type of AVM of the spinal cord itself is what is termed a perimedullary AVM. These can be direct arteriovenous fistulas on the surface of the spinal cord or small AVMs immediately on the spinal cord's surface. These may be treated with open surgery or endovascular embolization, depending on the relationship to the artery of Adamkiewicz, which typically occurs in the thoracic region and is one of the major feeding vessels to the spinal cord parenchyma.

Traumatic Vascular Diseases

Traumatic Aneurysms

Intracranial aneurysms may also arise following significant open or closed head injury. These are most commonly pseudoaneurysms, meaning that the vessel wall is ruptured and the hematoma is confined in the extravascular space by surrounding cerebral structures. Typical locations for these aneurysms include the distal anterior cerebral artery and the ICA at the skull base. Traumatic aneurysms can be discovered during neuroimaging following the trauma or may present with delayed hemorrhage. Diagnosis can be made with CTA, magnetic resonance angiography (MRA), or conventional angiography. Since the rupture rate is high, surgical or endovascular management is usually indicated.

Carotid-Cavernous Fistulas

Carotid-cavernous fistulas are a type of arteriovenous fistula between the carotid artery and the cavernous sinus. They can occur posttraumatically or spontaneously and are subclassified according to whether the fistula is directly between the cavernous sinus and the ICA or indirectly between the cavernous sinus and meningeal branches of the internal and/or external carotid artery. Symptoms are generally due to venous hypertension, and patients may present with a bruit, ocular chemosis, or pulsating exophthalmos. Diagnosis is usually confirmed by conventional angiography. Progressive symptoms, especially visual deterioration, are an indication for urgent intervention; however, some low-flow lesions may thrombose spontaneously. Treatment is commonly performed endovascularly by transarterial or transvenous embolization.

Spontaneous Intracerebral Hemorrhage

Spontaneous intracerebral hemorrhage without an identifiable structural cause is most commonly caused by hypertensive vasculopathy or cerebral amyloid angiopathy. It tends to affect people over the age of 60 and usually presents as a sudden onset of a focal neurological deficit with a change in level of consciousness. Diagnosis is made by plain CT in the emergency room setting (Figure 4). Hypertensive bleeds tend to be deep and involve the striatum, thalamus, and cerebellum. Spontaneous intracerebral hemorrhages caused by amyloid angiopathy, on the other hand, tend to be lobar. Anticoagulation is a significant risk factor, among others, for intracerebral hemorrhage in the elderly population. Treatment typically involves some combination of anticoagulation reversal, medical management to prevent secondary brain injury, and surgical evacuation of the hematoma.

Figure 4 Plain computed tomography scan demonstrating an acute right frontal intracerebral hemorrhage with extension into the ventricular system.

See also: Animal Models of Stroke; Axonal Injury in Demyelinating Disease and CNS Injury; Spinal Cord Injuries; Stroke; Stroke: Injury Mechanisms; Stroke: Neonate vs. Adult.

Further Reading

Auer RN and Sutherland GR (2005) Primary intracerebral hemorrhage: Pathophysiology. *The Canadian Journal of Neurological Sciences* 32(supplement 2): S3–S12.

Feigin VL and Findlay M (2006) Advances in subarachnoid hemorrhage. *Stroke* 37: 305–308.

Markus HS and Alberts MJ (2006) Update on genetics of stroke and cerebrovascular disease 2005. *Stroke* 37: 288–290.

Molyneux AJ, Kerr RS, Yu LM, et al. (2005) International subarachnoid aneurysm trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: A randomised comparison of effects on survival, dependency, seizures, rebleeding, subgroups, and aneurysm occlusion. *Lancet* 366: 809–817.

Spetzler RF, Detwiler PW, Howard AR, and Porter RW (2002) Modified classification of spinal cord vascular lesions. *Journal of Neurosurgery (Spine 2)* 96: 145–156.

Suarez JI, Tarr RW, and Selman WR (2006) Aneurysmal subarachnoid hemorrhage. *New England Journal of Medicine* 354: 387–396.

Wiebers DO, Whisnant JP, Huston J 3rd, et al. (2003) Unruptured intracranial aneurysms: Natural history, clinical outcome, and risks of surgical and endovascular treatment. *Lancet* 362: 103–110.

Cetacean Brains

L Marino, Emory University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Cetaceans are an order of fully aquatic mammals consisting of two modern suborders. These two modern suborders, collectively known as neocetes, include the odontocetes (toothed whales, dolphins, and porpoises) and the mysticetes (baleen and rorqual whales). There are 67 species of odontocetes and 11 mysticete species. The first cetacean suborder, the archaeocetes ('ancient whales') began as fully terrestrial hoofed animals derived from an ungulate ancestor. Archaeocetes later became fully aquatic before they became extinct. Archaeocetes survived from approximately 55 to 33 Ma.

Cetacean brains have evolved substantially over the past 55 My during their adaptation from a terrestrial lifestyle to a semiaquatic existence and finally, approximately 43 Ma, to a fully aquatic lifestyle. Modern cetacean brains are characterized by several features. These include large mass (both in absolute terms and in relation to body mass), changed proportions of structures related to various domains of sensory processing, a high degree of cortical gyrification, an atypical arrangement of cortical surface projection zones known as cortical adjacency, a high degree of hemispheric independence, and a highly unusual neocortical cytoarchitecture.

Gross Structure of the Cetacean Brain

The adult cetacean brain exhibits a number of unusual gross morphological characteristics. As compared with terrestrial mammals the dolphin brain is much more spherical. The brain is greatly widened bitemporally and heightened and stands in contrast to the elongated brains of primates and carnivores. In cetaceans the cranium is telescoped along the beak–fluke axis and the forebrain is foreshortened, leading to a rounded and high brain shape. The brain's natural orientation in the cranial cavity is also unusual. The brain exhibits pronounced mesencephalic, pontine, and cervical flexures, leading to a foreshortening of the forebrain and a rotation of the brain along the body axis almost 90°. This configuration has led many authors to liken the shape of the cetacean brain to that of a boxing glove. The flexures, which are maintained throughout adulthood in cetaceans, are reminiscent of those observed in the embryonic state of terrestrial mammals (see Figure 1).

Cerebral enlargement in cetaceans occurred most exuberantly in the parietal and temporal regions. Whereas primate brains feature large frontal lobes, no homologous frontal lobe region in the cetacean brain has been identified, leading many investigators to substitute the term 'orbital lobe' for frontal lobe when referring to these modest, ventrally oriented hemispheric regions (see Figure 1).

Brain Size

Modern cetacean brains are among the largest of all mammals in both absolute mass and in relation to body size. One of the hallmarks of cetacean brain evolution is the substantial increase in relative brain size and expansion of the cerebral hemispheres. Average brain size for the extinct suborder of cetaceans, the archaeocetes, was 750 g and varied from about 290 to 2240 g. Modern cetacean brain size, which averages approximately 785 g, ranges from 220 g for the adult Franciscana dolphin (*Pontoporia blainvillei*) to 8000 g for the adult sperm whale (*Physeter macrocephalus*), the largest brain on earth.

Brain size is perhaps more meaningfully expressed as an 'encephalization quotient' (EQ), which is a value that represents how large or small the brain of a given species is compared with other species of the same average body weight. When examined relative to body size within the context of EQ, the great enlargement in the brain in cetaceans becomes even more evident. Archaeocete brains were smaller than expected for their body size. Almost all modern odontocetes, on the other hand, possess above-average encephalization levels compared with other mammals and therefore have brains larger than expected on the basis of brain–body allometry. In order to understand the meaning of EQ it is helpful to note that the EQ for modern humans is 7.0. That is, our brains are about seven times larger than one would expect for our body size. Turning to cetaceans, numerous odontocete species possess EQs in the range of 4–5, that is, they possess brains 4–5 times larger than one would expect for their body size. Many of these odontocete values are second only to those of modern humans and are significantly higher than any of the nonhuman anthropoid primates (highest EQ $\sim$ 3.3). We now know that a substantial increase in EQ during cetacean evolution occurred approximately 35 Ma, when average EQ jumped from 0.5 for archaeocetes to 2.0 for the

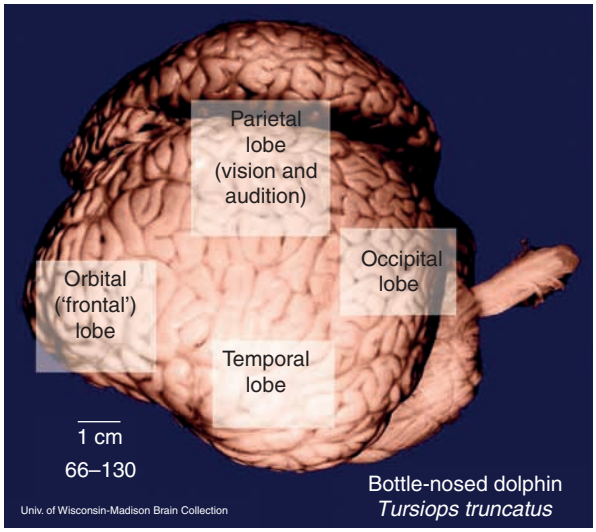

Figure 1 Dorsolateral view of bottlenose dolphin brain showing the location of the parietal, occipital, temporal, and orbital lobes. Image courtesy of the Comparative Mammalian Brain Collection, The National Museum of Health and Medicine, Armed Forces Institute of Pathology, Washington, DC.

first neocetes (again, early odontocetes and mysticetes). Therefore, although many archaeocetes possessed large brains because they had large bodies, a major increase in the size of the brain relative to the body did not occur until the emergence of the neocetes.

EQs of mysticetes are all below a value of 1, that is, below average, because of an uncoupling of brain size and body size in very large (and aquatic) animals. Nevertheless, the large absolute size of the mysticete brain (and other features, described later) demonstrates that the mysticete brain underwent evolutionary changes similar to those of odontocete brains.

Surface Features

The cetacean telencephalon is arranged into three concentric tiers of limbic, paralimbic, and supralimbic tissue and accounts for much of the total brain volume (e.g., 63% in the Franciscana, *Po. blainvillei*, and 87% in the sperm whale, *Ph. macrocephalus*). The high degree of cortical gyrification and resulting surface area in modern cetacean brains is unsurpassed among mammals, including humans. Endocasts from fossil material indicate that the surface of the archaeocete brain, in contrast, was relatively smooth, providing further evidence that modern cetacean brains have expanded substantially over time. In the bottlenose dolphin (*Tursiops truncatus*) and common dolphin (*Delphinus delphis*), neocortical surface areas have been measured, respectively, as 3745 and 2136 cm^2 and average total brain weights as 1587 and 802 g. These values exceed values for modern humans, with neocortical surface area being approximately 2275 cm^2 and average total brain weight at 1300 g.

Electrophysiological mapping studies of cetacean cortex places primary visual cortex on the vertex of the hemisphere in the lateral gyrus and the primary auditory cortex lateral and directly adjacent to it in the suprasylvian gyrus. Secondary auditory cortex lies lateral to the primary auditory field in the medial ectosylvian gyrus and somatosensory and motor cortices lie immediately adjacent and rostral to the visual and auditory regions (see **Figure 1**). This is a highly unusual arrangement for large mammal brains. This configuration, termed 'cortical adjacency,' means that all of the sensory–motor projection zones of the cetacean brain are confined to one contiguous region of the overall surface, with a vast expanse of non-projection cortical tissue (that can be tentatively described as 'integrative' or 'associative') in the remaining portion of the hemispheres.

Reproportioning of Central Sensory Structures

Auditory Structures

Cetaceans use audition in an extremely complex and multifaceted manner. Therefore, auditory structures are well developed in all cetaceans. Furthermore, auditory structures are relatively larger in odontocetes than in mysticetes, because of the ability of odontocetes to echolocate. The vestibulocochlear nerve in cetaceans is immense in diameter and is composed of relatively more auditory than vestibular fibers. The ventral cochlear nucleus, trapezoid bodies, lateral lemniscus, and inferior colliculi (auditory tectum) are all greatly enlarged in comparison with terrestrial mammals. Whereas in mysticetes the inferior colliculus can be larger or smaller than the superior colliculus (visual tectum), in odontocetes the inferior colliculus is always larger than the visual tectum and can be at least four times the size of the superior colliculus. The auditory tectum projects to a large medial geniculate nucleus in the massive thalamus. Likewise, the primary and secondary auditory projection zones on the cerebral surface are extensive.

It is important to note that auditory structures in the cetacean brain have not expanded at the expense of visual structures. With the exception of the blind river dolphins, visual structures in the cetacean brain are well developed.

Olfactory Structures

One of the hallmarks of cetacean brain evolution that is found to varying degrees in other marine mammals is the reduction of olfactory structures. Olfaction has been completely lost in adult odontocetes. Fetal odontocetes possess small olfactory structures that regress completely by birth. Interestingly, a short olfactory peduncle sometimes remains in adult sperm whales (*Ph. macrocephalus*) and northern bottlenosed whales (*Hyperoodon ampullatus*), which may be consistent with their phylogenetic status in relation to other cetacea. Adult mysticetes, while possessing a vastly reduced olfactory system, have maintained small olfactory bulbs, a thin olfactory peduncle, and an olfactory tubercle.

The reduction of olfactory input has possibly resulted in a reproportioning of the limbic system in cetaceans. Notably, the cetacean hippocampus (archicortex), fornix, and mammillary bodies are all highly reduced while the amygdala is large and well developed. These proportions may reflect the maintenance of substantial nonolfactory sources of input to the amygdala.

Hemispheric Connectivity

As in other mammals, the two hemispheres of the cetacean brain are connected by anterior and posterior commissures and the corpus callosum. In cetaceans, corpus callosum mass is considerably smaller in relation to brain mass than in other mammals. This is partly due to the general allometric rule that larger brains maintain less interhemispheric connectivity. Therefore, because most cetacean brains are quite large they possess small corpora callosa. For example, the human and killer whale (*Orca orcinus*) corpora callosa possess the same cross-sectional area despite the fact that the killer whale brain is over five times the mass of the human brain.

In addition to a relatively small corpus callosum, cetacean brains evince other aspects of hemispheric independence. In most mammals, each eye projects the majority of its nerve fibers to the contralateral hemisphere but also shares some fibers with the ipsilateral side. Cetaceans, on the other hand, exhibit the highly atypical arrangement of total decussation of the optic nerves to the contralateral hemisphere. Furthermore, each hemisphere is connected to a separate blood supply.

Perhaps the best-known behavioral evidence for a high degree of hemispheric independence in cetaceans is unihemispheric sleep. Many investigators have suggested that the relatively weaker interhemispheric connections in cetacean brains may facilitate the unusual unihemispheric sleep patterns exhibited by cetaceans. Early stages of sleep in cetaceans may occur unilaterally or bilaterally. Stage 3 (slow-wave sleep), however, occurs in only one hemisphere at a time and is correlated with unilateral changes in slow-wave electroencephalogram (EEG) activity. The awake hemisphere retains a low level of alertness. A typical bout of sleep in cetaceans involves an alternating pattern of sleep in one hemisphere and then the other. This pattern continues for 2–3 h.

Cortical Cytoarchitecture

The cetacean neocortex is relatively thin, with a width between 1.3 and 1.8 mm, as compared with 3.0 mm in humans. The thinness of the neocortex likely allows for the high degree of cortical gyrification observed in cetaceans. Cetacean neocortex is characterized by a general absence of granularity (leading to an absence of layer IV), a very prominent thick layer I, the presence of large atypical neurons in the dense layer II, and very large pyramidal neurons frequently forming clusters at the border between layers III and V. The general lack of layer IV in cetaceans has important implications for afferentation patterns. In primates and other mammals, major thalamocortical afferents synapse directly on layer IV. In cetaceans, the majority of afferents from the thalamus appear to go through the very thick layer I to synapse on extraverted neurons of layer II. There is some evidence that a small portion of afferents go to layers III and V as well.

For many decades the cetacean cortex was thought to be homogeneous and relatively nondifferentiated in character. However, the most recent findings show that the cytoarchitectural patterns in cetacean cortex are far more varied and complex than generally thought, with a degree of regional differentiation on par with other large-brained mammals. Although areas of less distinct lamination do exist there are various neocortical regions that display considerable degrees of cytoarchitectural differentiation. For instance, despite the modest gross morphological appearance of the frontal lobes, this region is distinctly laminated and comprises several cortical fields. Also, highly distinct vertical modules are observed in the insular cortex and other areas. Therefore, there is a rich diversity of cytoarchitectural fields in most of the lobes. However, the organization that underlies this diversity is different from that in other mammals and likely represents a different functional strategy for cortical processing, compared to other mammals.

Conclusions

The cetacean brain is a highly elaborated brain that represents an intriguing and very different path toward functional complexity, compared to other mammals. Cetacean brains exhibit a unique combination of characteristics that include an encephalization level second only to that of modern humans, a degree of cortical gyrification unmatched by other mammals (including humans), a high degree of anatomical and functional hemispheric independence, and a highly atypical but complex cortical cytoarchitecture. These characteristics are consistent with the long-standing evidence for cognitive and behavioral complexity in cetaceans.

See also: Allometric Analysis of Brain Size; Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brain Scaling Laws; Molecular Anatomy of the Mammalian Brain; Social Communication in Whales and Dolphins.

Further Reading

Hof PR, Chanis R, and Marino L (2005) Cortical complexity in cetacean brains. *The Anatomical Record* 287A: 1142–1152.

Hof PR, Glezer II, Nimchinsky EA, et al. (2000) Neurochemical and cellular specializations in the mammalian neocortex reflect phylogenetic relationships: Evidence from primates, cetaceans, and artiodactyls. *Brain Behavior and Evolution* 55: 300–310.

Marino L (1998) A comparison of encephalization between odontocete cetaceans and anthropoid primates. *Brain, Behavior and Evolution* 51: 230–238.

Marino L (2002) Brain size evolution. In: Perrin W, Wursig B, and Thewissen H (eds.) *Encyclopedia of Marine Mammals*, pp. 158–162. San Diego: Academic Press.

Marino L (2007) Evolution of the brain and cognition in cetaceans. In: Platek SM, Keenan JP, and Shackelford TK (eds.) *Evolutionary Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Marino L, McShea D, and Uhen MD (2004) The origin and evolution of large brains in toothed whales. *The Anatomical Record* 281A: 1247–1255.

Morgane PJ, Glezer II, and Jacobs MS (1990) Comparative and evolutionary anatomy of the visual cortex of the dolphin. In: Jones EG and Peters A (eds.) *Cerebral Cortex*, pp. 215–262. New York: Plenum.

Morgane PJ, Jacobs MS, and MacFarland WL (1980) The anatomy of the brain of the bottlenose dolphin (*Tursiops truncatus*): Surface configurations of the telencephalon of the bottlenose dolphin with comparative anatomical observations in four other cetacean species. *Brain Research Bulletin* 5: 1–107.

Oelschlager HA and Oelschlager JS (2002) Brains. In: Perrin WF, Wursig H, and Thewissen H (eds.) *Encyclopedia of Marine Mammals*, pp. 133–158. San Diego: Academic Press.

Ridgway SH (2000) The auditory central nervous system of dolphins. In: Au W, Popper A, and Fay R (eds.) *Hearing in Whales and Dolphins*, pp. 273–293. New York: Springer.

Chemical Senses: Overview

G M Shepherd, Yale University School of Medicine, New Haven, CT, USA

© 2009 Published by Elsevier Ltd.

A fundamental understanding of the chemical senses begins with the common chemical sense, defined as the ability of cells to detect and react to chemical stimuli. This applies to single cell organisms – bacteria and protozoa – sensing the composition of their environment, as well as to cells of the body specialized for sensing particular signals of the *milieu interieur* (oxygen tension, for example, by the carotid body, or the signals of the immune system), and to cells everywhere in their body organs as they interact with other cells in their different functions.

Among the special chemical senses, taste is our most obvious. Every time we put something in our mouth, we detect and discriminate between the basic tastes – sweet, salt, sour, bitter, and, especially in Asian countries, a meaty taste called umami. Molecular receptors for these qualities are being identified. Research on taste mechanisms is being carried out at several levels: the receptors and second messenger mechanisms in the cells of the taste buds in the tongue and pharyngeal cavity, and the neural pathway from the taste buds to the brain stem and on to the thalamus and neocortex. These neural mechanisms give rise to the sensation of taste, which is being studied by traditional psychophysical methods combined with new methods of genetic analysis.

The articles on smell represent the growing general interest in this sense. In invertebrates, several species are emerging as excellent models. The opportunities to identify olfactory receptors, compare and contrast them with those in the vertebrate, and correlate them with processing mechanisms in the olfactory pathway are beginning to be exploited. Earlier work established that the antennal lobe is equivalent to the vertebrate olfactory bulb as the first processing stage in the olfactory pathway. In both, the modules called glomeruli are expressions of a universal principle in processing odor information. It is becoming clear that principles revealed in one model may apply generally. These models include *Manduca*, in which there are receptors, glomeruli, and pathways for both ordinary odors and pheromones. *Drosophila* has emerged as another leading model, particularly amenable to gene manipulations. Honeybees are especially attractive for the correlations between odor maps in the glomerular layer, equivalent to the odor maps in vertebrates, neural processing mechanisms in the antennal lobe and mushroom bodies, and behavior.

In the vertebrates, olfactory research is being carried out over a wide range of levels and disciplines, beginning with the receptors, first identified by Linda Buck and Richard Axel in 1991, and providing a great impetus to redefining olfaction as a molecular sense. The receptors are believed to be the largest family in the genome. Starting as orphan receptors, the variably broad ligand specificities are gradually being identified and correlated with the specificities of the olfactory glomeruli to which their receptor cells project. Microarray techniques hold particular potential for dealing with this large gene family. The receptor cells are distributed within a complex mosaic within the olfactory epithelium in the nose, which is being analyzed by both electrophysiology and cell imaging techniques. This work will help to define at the receptor level the properties of odor detection, adaptation, and discrimination that are the basis for further processing in the olfactory bulb and beyond.

Among the advances provided by the olfactory receptors, their role in the targeting of subsets of receptor cells to individual glomeruli has been among the most insightful. It has enabled the merging of the molecular biology of the receptors with earlier work on the olfactory bulb as a model for the synaptic organization of a cortical structure. Three articles document these advances. Anatomical studies are providing increasing detail on the basic levels of synaptic organization in the olfactory bulb: within the glomerulus, between glomeruli, and at the level of mitral cell–granule cell interconnections. Physiological studies are revealing the functional interactions involved in processing odor stimuli at each of these levels. The distribution of inputs to the olfactory glomeruli from the olfactory receptor cells, as contained in odor maps, establishes the spatial and temporal nature of the processing mechanisms.

In parallel with these studies in the olfactory bulb, studies are being pursued at the level of the olfactory cortex which receives the olfactory bulb output. Here, in contrast to the stereotyped patterns of the bulb, the interconnections and interactions have appeared to be diffuse and random. However, recent studies are revealing new insights into the complexity of organization within the cortex and between the cortex and the olfactory bulb. Physiological studies are showing how the cortex functions as a system for odor discrimination and memory, expressing many properties of higher-level association cortex

in other sensory systems. Taken together, these studies provide insights into the neural mechanisms of odor processing. Despite these advances, study of the neural basis of smell perception faces the difficult task of establishing baseline data for thresholds for different odors in different species. The biggest challenge is obtaining a consensus on this data in humans.

These advances in the biology of smell have given hope that the principles discovered can be used to guide development of effective devices for detecting and discriminating volatile compounds, which has potential applications in a variety of contexts, from quality control in food production to detection of land mines. The biological and physical principles are providing the basis for the development of an overall theory of olfactory processing, equivalent to that for other sensory systems.

An important recent development is the realization that the importance of smell, particularly to humans, may be derived not so much from the smells we sense when we sniff in, such as perfumes, as from the smells we sense when we sniff out, such as during eating. Via this route, smell is combined with the other senses, especially taste and touch, to give the combined sense of flavor. This new research on retronasal smell is described, and the physiological studies of flavor neurons at the highest levels of reception and modulation, in the orbitofrontal cortex and related areas, are described.

A subject in the sense of smell of increasing interest is the role of pheromones, molecules that carry specific social information between members of a species. Classically it has been believed that in vertebrates this information is carried exclusively by the vomeronasal organ projecting to the accessory olfactory bulb, but recent studies have shown that it is carried in parallel by both the main and accessory systems. Current studies of the vomeronasal system are summarized elsewhere in the encyclopedia. An overview of the current state of knowledge about the processing of pheromones by the two systems is given elsewhere in the encyclopedia. The special case of human pheromones has attracted much attention; despite the heavily conditioned nature of human responses to olfactory stimulation, there is increasing evidence for the ability of human body odors, acting as pheromones, to influence mood, hormonal changes, and mate preference.

In summary, work on the chemical senses is being pursued vigorously in many species, with many innovative technical approaches, and at many different levels of organization. This section of the encyclopedia should enable the reader to appreciate the scope of this work and gain a new perspective on the importance of the chemical senses for animal life in general, and humans in particular.

See also: Chemical Senses: Protozoa; Electronic Nose; Flavor Physiology; Olfaction in Invertebrates: Manduca; Olfaction in Invertebrates: Drosophila; Olfaction in Invertebrates: Honeybee; Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Olfactory Higher Centers Anatomy; Olfactory System Theory; Pheromones in Mammals; Pheromones in Humans and Social Chemosignals; Retronasal Olfaction; Taste: Vertebrate Taste Bud Physiology; Taste: Vertebrate Central Pathways; Taste: Vertebrate Psychophysics; Vomeronasal Accessory System.

Chemical Senses: Protozoa

J L Van Houten, University of Vermont, Burlington, VT, USA
W E Bell, Virginia Military Institute, Lexington, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Protozoans, while solitary cells, have a variety of sensory systems, among which are those that detect and respond to chemicals in their environment. The chemicals may derive from food sources or from other similar protozoan cells that communicate about mating or differentiation via chemical signals. Highlighted here are four different sensory systems: three in free-living ciliates and one in an amoeboid cell.

Ciliated protozoa are excitable cells that swim, propelled by the beating of their cilia. The most extensively studied of the ciliates for chemosensation are the free-living ciliates, such as *Paramecium* sp., *Tetrahymena*, and *Euplotes*. Studied to a lesser extent are other ciliates, such as *Blepharisma* and *Oxytricha* (*Sterkiella*). These organisms can mate, which previously made them popular for studies of maternal inheritance and nuclear development. They have been studied using microscopy for their swimming behavior, and using electrophysiological techniques for their electrical properties, over many decades, but the advent of molecular technologies, genome projects, and RNA interference (RNAi) techniques has propelled forward the study of these organisms for their chemosensory behavior and other properties. *Paramecium* is discussed here because of the deep understanding of the physiology underlying its sensory responses that result in changes in ciliary motility. *Tetrahymena* is discussed as a species that responds with motility changes similar to those of *Paramecium*, but shows differences in signaling mechanisms, including the end result of secretion rather than locomotion. *Euplotes* has an entirely different sensory system for detecting self- and non-self for growth and mating.

Like the ciliates, amoeboid protozoa also respond to chemical stimuli with a motility response, but, lacking cilia, these cells crawl up the chemical gradients in their search for food (such as bacteria). When the food is exhausted, they begin to signal to each other to congregate into groups in preparation for producing spores to survive the adverse environmental conditions. *Dictyostelium discoideum* is discussed here as one well-studied example of amoeboid chemotaxis.

Paramecium Chemosensory Behavior

Paramecia are free-living ciliated cells that mostly dwell in ponds and streams, but some species can be found in more extreme aquatic environments. *Paramecium* swimming behavior has intrigued scientists, reportedly since the invention of the first microscope, and certainly since the writings of Jennings in the early 1900s; Jennings described the many external stimuli to which *Paramecium* responds, including temperature, pH, inorganic ions, organic compounds, gravity, and touch. Here the focus is on the responses of *Paramecium* to chemical cues. The relevant stimuli differ depending on the species under investigation, but for *Paramecium tetraurelia*, the attractant stimuli appear to signify the presence of food (bacteria) and the repellents signal less than optimal conditions (e.g., pH and ionic strength).

Chemical stimuli that cause populations of cells to congregate or disperse do so by affecting swimming behavior. Cells do not show chemotaxis (directed motility up or down a gradient), but rather show a biased random walk, which can be a very efficient mechanism for accumulating or dispersing small organisms. The cells swim fast and smoothly when moving up a gradient of attractant stimulus, and they swim more slowly and turn frequently when encountering a repellent or leaving an area of attractant. Turning and speed regulation are essential for attraction and repulsion; adaptation of the cell, so that the cell has an immediate response upon entering a new environment, sharpens the response.

Studies of *Paramecium* chemosensory behavior have built upon years of studies of *Paramecium* physiology, and benefit from the genetic and physiological research on behavioral mutants of *Paramecium*.

Physiological Basis of Chemosensory Behavior

The abrupt turn in the swimming path of a cell is caused by a calcium action potential from the opening of ciliary voltage-gated channels. Each action potential raises ciliary calcium above 10^{-6} M, and causes the axonemal structures inside the cilia to change the power part of the stroke, making the cell stop and back up for a fraction of a second. The cell then pivots in place and takes off again, swimming forward in an almost random new direction. This is the turn due to the Ca^{2+} action potential. As a cell depolarizes, it reaches threshold for action potential and thus depolarizations can initiate turns. Elegant studies of membrane electrophysiology by Kaneko, Naitoh, and Eckert showed the genesis of

the ciliary power stroke reversal with the calcium action potential and the control of the voltage-gated calcium channels (Figure 1). These channels are probably exclusively in the cilia, and deciliated cells regain the action potentials after the cilia regrow a fraction of their length.

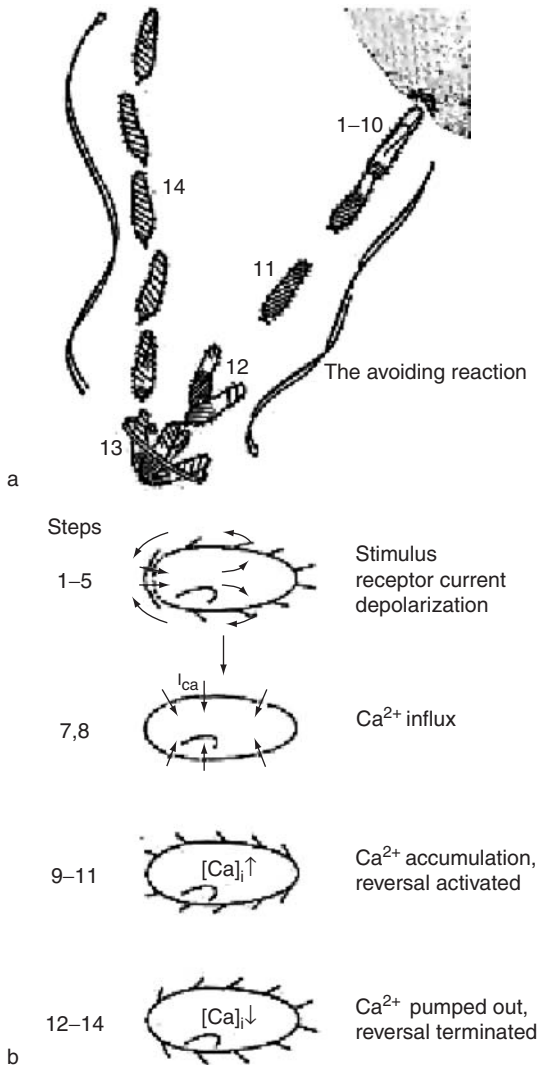

Figure 1 Physiological steps in the avoiding reaction or turn. (a) Retreat from the stimulus, and resumption of forward locomotion. (b) The sequence of steps corresponding to the numbers in diagram (a). Step 1, stretch of anterior membrane upon collision with obstacle; step 2, local increase in membrane conductance; and step 3, inward receptor current through stimulated membrane. Step 4, electrotonic spread of receptor current, produces step 5, outward current through rest of membrane (arrows show current flow). Step 6, depolarization of cell membrane (receptor potential), produces step 7, increase in calcium conductance. Step 8, inward Ca^{2+} current; step 9, rise in intracellular Ca^{2+} concentration; step 10, cilia reverse beat; and step 11, cell swims backward. Step 12, Ca^{2+} pumped out; step 13, intracellular concentration of Ca^{2+} drops, cilia resume normal orientation; and step 14, cell swims forward. (a) Adapted from Parducz B (1967) Ciliary movement and coordination in ciliates *International Review of cytology* 21: 91–128 and Grell K (1968) Protozoologie. Berlin: Springer. (a, b) Reproduced from Figure 5 Eckert R (1972) Bioelectric control of ciliary activity. *Science* 176: 473–481.

The speed of swimming correlates with the frequency of ciliary beating, and the frequency of beat is controlled by the membrane potential. Machemer's work demonstrated that as the membrane potential is hyperpolarized from resting potential, the ciliary beating increases, and as it is depolarized, the beat frequency decreases, until finally the beat direction appears to reverse because the power stroke is changed. Depolarization beyond this point results in an increased beat frequency, but in the 'reverse' direction.

In the presence of chemical attractant stimuli, cells swim fast and smoothly with few turns, due to hyperpolarization of the membrane potential away from the threshold for action potentials. The resulting behavior sends the cells up the gradient due to the biased random walk. Conversely, cells swim more slowly and turn frequently due to action potentials because they are depolarized if they happen to move away from an attractant stimulus or toward a repellent. Thus, *Paramecium* chemosensory behavior is consistent with the membrane potential control of ciliary beating.

The genetic approach to dissect the excitable membrane of *P. tetraurelia* taken by Kung (and later by Takahashi for *P. caudatum*) led to a valuable set of behavioral mutants. Some mutants cannot reverse their ciliary beat and cannot generate action potentials, whereas some swim forward rapidly under some conditions and others mostly swim backward. Some of these mutants map to the gene for calmodulin and can be cured by injected calmodulin protein or DNA. These mutants have shown us that the calmodulin molecule distinguishes among the channels that it activates, and that mutants in the N-terminal and C-terminal lobes have diametrically opposed phenotypes. For example, one group of 'Cam' mutants with mutations in the C-terminus have defects in calmodulin-activated K-conductances, which should short-circuit the calcium action potential and normally either keep the cell from turning or abbreviate the turn. Without this K-conductance, cells swim backward for long periods of time, because the membrane potential remains depolarized, and show different Ca channel kinetics due to more intracellular Ca^{2+}. Another group of mutants has changes in amino acids in the N-terminus lobe of calmodulin and they can no longer activate a calmodulin-regulated Na channel. Therefore, when cells are in Na solutions and are depolarized, the Na channels do not open and the response to Na, which should be active turning, is curtailed. There are a host of allelic mutants with different changes in the calmodulin gene sequence and at least six conductances that are affected by

calmodulin mutations. Behavioral mutants have made it possible to use the connection between ciliary motion and ion channels to produce easily identifiable phenotypes and to amplify the effects of very subtle mutations so that they can be selected.

Behavioral mutants have shown us that turning is essential for chemoresponse and that a Ca^{2+}/calmodulin-activated K-conductance is involved in the initiation of the hyperpolarization for some stimuli. The mutants have allowed for the elimination of the involvement of other conductances in the onset of the attractant-induced hyperpolarization.

Paramecium Membrane Receptors

Biochemical studies, bioinformatics, and antisense techniques have revealed that the *P. tetraurelia* cell surface receptors for attraction include 7-transmembrane receptors (e.g., for extracellular cyclic adenosine monophosphate (cAMP)) and loosely tethered receptors (e.g., for folate, the receptor of which is anchored to the cell surface through a glyco-lipid addition to the receptor). The mechanisms by which these two very different kinds of receptors send messages inside the cell are not yet known, although there is evidence for tripartite GTP-binding proteins (G-proteins) and small G-proteins. Both classes of G-proteins have been shown to participate in signal transduction for 7-transmembrane receptors and glyco-lipid-anchored receptors in other systems. Additional attractants, such as glutamate, are characterized for their binding kinetics and structure–activity, but the protein receptors have not yet been isolated. All of the receptors appear to activate K^+ channels to initiate hyperpolarization and, therefore, the signal transduction pathways must couple receptors to downstream channels. There is growing evidence for coupling to plasma membrane calcium pumps to sustain the hyperpolarization and for congregation in special lipid domains of the surface membranes, which may serve as platforms for signaling molecules as in other systems.

There is one stimulus for *P. tetraurelia* that does not attract the cells by first binding to a surface receptor. Ammonium chloride diffuses across the cell membrane as ammonia, and once inside the cell, makes the cytoplasm more alkaline. The result is a hyperpolarization, that causes fast, smooth swimming, and attraction, but the mechanism of hyperpolarization is not yet understood.

Ciliary Motion

There is an effort to correlate the physiological findings about hyperpolarization and depolarization with the workings of the axoneme, the cytoskeletal structure that makes the cilium bend and beat. For example, hyperpolarization by reduction in extracellular K has long been known to hyperpolarize the cell (just as it would a neuron) and also leads to an increase in intracellular cAMP. An adenylyl cyclase that has both the characteristics of a cyclase and K channel has been cloned and shown to be very similar to adenylyl cyclases in parasitic protozoa and *Tetrahymena*. The annotated draft genome shows that *P. tetraurelia* has 15 genes that probably code for adenylyl cyclases and all have a K channel subunit.

Cyclic nucleotides cAMP and cGMP have been shown to affect ciliary beat frequency in cells made permeant to extracellular molecules by exposure to detergent. cAMP increases frequency of beating, and cyclic GMP has the converse effect, showing us one link between the hyperpolarization of the cell and the resultant behavioral change. Substrates for cAMP-activated kinase (PKA) have been identified in *P. tetraurelia* and phosphorylation of at least one substrate, associated with an axonemal dynein, correlates with increased beat velocity and swim speed. At least one chemoattractant, glutamate, utilizes cAMP and PKA as a second messenger system that ultimately increases the ciliary beat. Other stimuli appear to utilize other second messengers, such as Ca^{2+}/calmodulin, to accelerate the beat.

Other Ca^{2+}-Dependent Processes in *Paramecium* Chemical Sensing

Nucleotide triphosphates ATP and GTP cause cells to be repelled by depolarizations that arise from Ca^{2+} conductances triggering Ca^{2+}/calmodulin-dependent Mg^{2+} and Na^+ currents. This mechanism is distinctly different from the mechanism of repulsion from high levels of K^+ or Na^+, inosine monophosphate (IMP), or quinine, which cause depolarizations but do not cause oscillations of membrane potential.

Tetrahymena

Tetrahymena thermophila responds to different arrays of chemicals compared with *Paramecium,* possibly reflecting the different niches that the two ciliates inhabit in nature. Both can live in ponds and fresh water, but *T. thermophila* can also live in a wider range of conditions, and has been found in brackish water, thermal springs, and even soil and sewage. Peptides and small proteins attract *T. thermophila,* perhaps aiding in their search for food sources. Comprehensive studies of attractants and repellents are not available for *P. tetraurelia* or *T. thermophila*, but in general the chemicals that elicit attraction for *P. tetraurelia* can be classified as soluble small molecules that are bacterial metabolites, probably signifying

their food source, and attractants for *T. thermophila* are peptides and proteins. Naturally occurring repellents for *T. thermophila* include extremes of pH and ionic strength, IMP, ATP, and GTP, with GTP the most potent; *T. thermophila* shares with *P. tetraurelia* repulsion from GTP and IMP.

A receptor for one of the peptide attractants of *T. thermophila* has been recently identified. It has structural similarity and shared epitopes with the human insulin receptor β-subunit, particularly in the tyrosine kinase domain. Other receptors that have similarity to mammalian proteins based on their binding or physiological response include those that bind β-endorphins and concanavalin A. This is not surprising, since numerous mammalian proteins, including insulin, endothelin, serotonin, histamine, platelet-derived growth factor (PDGF), and epidermal growth factor (EGF), have been identified as stimulators of *Tetrahymena* behavior. Recently, bradykinin has been added to the list of chemoattractant peptides. It is important to note that many chemoattractant peptides are also important for cell survival in *Tetrahymena* sp. Several authors have observed that the nutritional status of the culture studied is critical in determining the response of *Tetrahymena* to stimulus. Under starvation conditions, some of the above-mentioned proteins are chemotatic, whereas in well-fed cultures, these proteins are mitogenic.

Repellents such as GTP cause rapid transient depolarizations and frequent turning, while ATP causes a more sustained depolarization. Pharmacological evidence collected from both the ATP-mediated response and a lysozyme/pituitary adenylate cyclase-activating peptide (PACAP) chemorepellent response implicates a variety of pathways in *Tetrahymena* avoidance reactions, including pathways that signal through G proteins and ultimately enzymes such as phospholipase C and those activated by cAMP.

As for *Paramecium*, many of the intracellular signaling pathways remain unresolved, but the availability of an annotated genome and RNAi technology should reveal many of these control mechanisms.

Euplotes

Euplotes species have a highly developed mating system, and depending on the species, there can be an almost large number of mating types. The mating types and the Mendelian genetics of mating have long been worked out for *Euplotes raikovi* and *E. octocarinotus*. However, it took a dedicated biochemical approach in work on *E. raikovi* to identify in this species the soluble molecules secreted by one mating type that make a complementary mating type receptive to mating. Subsequently, the 'pheromones' and their receptors were identified for both species. Most recently, it has become apparent that the peptide pheromone is a subset of the amino acids of the receptor, with the pheromone secreted into the surrounding medium while the receptor, with an additional transmembrane domain, remains on the plasma membrane. The pheromone can bind to its own receptor or to a different receptor, elaborated on the surface of a complementary mating-type cell. Each cell expresses only one receptor and its cognate pheromone. When the pheromone binds to its cognate receptor, the cells are stimulated to go through the cell cycle and grow, essentially an autocrine response. When receptors are bound in an autocrine manner, intracellular cAMP levels are depressed. Disruption of this autocrine binding results in an increase in basal cAMP levels and suppression of growth. When the pheromone binds to a heterologous receptor, that cell is arrested in the cell cycle and becomes mating reactive. Since both 'self' and 'non-self' pheromone may be present simultaneously, the ratio of bound 'self' versus 'non-self' is critical in determining the response. Ultimately, ion currents and thus behavior are modified, enhancing the opportunity for a cell to contact a member of another mating type. Perhaps this is a mechanism to keep cells from mating when their nuclei are not in the proper stage to enter meiosis prior to exchange of haploid nuclei with a partner, or from expending unnecessary energy searching for mates when they are not available. Pheromones from *E. octocarinatus* for 10 mating types drive a similar orientation of cells to congregate in close proximity on the substrate because these pheromones also act as chemoattractants.

Interestingly, as in *Tetrahymena*, *E. raikovi* pheromone receptors can bind to a mammalian protein. Rather than insulin or other growth factors, the cytokine interleukin-2 (IL-2) can cross-react with the Er-1 pheromone receptor. IL-2 blocks Er-1-mediated response, and, conversely, Er-1 can associate with a mammalian IL-2 receptor, although it cannot replace IL-2 in a functional manner; whereas very little sequence homology can be noted by comparing sequence data, structural similarities between receptors exist.

Further study of the relationships between *Tetrahymena* and *Euplotes* species and the mammalian signaling molecules that stimulate these ciliates should provide valuable insight into the origin of many metazoan signaling pathways.

Dictyostelium

Many parallels have been drawn between *Dictyostelium* amoebae and neutrophils because both show chemotaxis (directed movement) of polarized cells

up very shallow gradients of stimuli. These amoeboid cells can respond to gradients that change as little as 2% from front to back of the cell, regardless of absolute concentration. These cell types share many of the same mechanisms for polarized movement, and even share a stimulus. *Dictyostelium discoideum* amoebae are attracted to cAMP, nutrients like folic acid, and platelet-activating factor (PAF). Neutrophils respond to bacterial cues like the formyl peptide fMLP, chemokines, complement factor C5a, and PAF. Upon loss of their bacterial food source, *D. discoideum* cells begin to respond to a gradient of cAMP from pulses of the stimuli given off by other cells, and eventually they move together to form a multicellular slug. Neutrophils respond to gradients that indicate the presence of bacteria and infection.

The cells that encounter a gradient of attractant develop a distinct front (pseudopod) and back (uropod) and proceed to crawl up the gradient. Actin polymerization at the anterior leading front edge of the cell, retraction of the back (uropod) through a myosin filament formation, and suppression of pseudopods between the leading edge and back are all necessary for efficient chemotaxis, that is, directed crawling up the gradient. The attractant stimuli bind to 7-transmembrane G-protein-coupled receptors, which then activate their associated tripartite G-proteins (α-, β-, and γ-subunits). The G β- and γ-subunits are crucial to carry on the signaling for chemotaxis. Curiously, the first components of the signaling pathways (receptors and G-proteins) are not redistributed as a cell encounters a gradient of attractant, and yet the cell must reorganize its shape (cytoskeleton) and somehow a front and back of the cell must form. This appears to occur due to an uneven distribution of PtdIns(3,4,5)P_3 (PIP_3), with more PIP_3 at the front. If the gradient changes, there are receptors along the entire surface to detect this change and redirect more PIP_3 to the new leading edge. To generate this uneven distribution of PIP_3, PI3K that phosphorylates PIP_2 to create PIP_3 translocates from the cytoplasm to the plasma membrane at the leading edge, and phosphatase and tensin homolog deleted on chromosome 10 (PTEN), the phosphatase that degrades PIP_3, moves to the back of the cell. This polarized distribution of these enzymes results in the concentration of PIP_3 at the leading edge, where proteins that bind to PIP_3 through plekstrin homology (PH) domains are recruited. Cytoplasmic regulator of adenylyl cyclase (CRAC) is among the PH proteins, and is required for receptor activation of the adenylyl cyclase (**Figure 2**).

Upon attractant stimulation, filamentous actin extends into the cell's leading edge, enhancing pseudopod formation and orienting the cell to crawl toward the attractant stimulus source. The sculpting of the actin cytoskeleton to form actin filament-filled pseudopods for forward motility is perhaps better known in neutrophils, but *D. discoideum* and neutrophils share molecular mechanisms here as well. Actin polymerization and branching occur in the leading edge to fill out the pseudopod and make forward chemotaxis movement possible. The localized polymerization depends on a cascade of events that begin with small G-proteins. The actin polymerization depends on a protein complex (Arp2/3) that is activated by two proteins (Wiskott–Aldrich syndrome protein (WASP) and the related WAVE/SCAR protein). These proteins, in turn, require Rho-family GTPases (small G-protein) Rac and Cdc42 for their activation. Where active Rac is found, the actin polymerization necessary for pseudopod formation will occur, and guanine nucleotide exchange factors

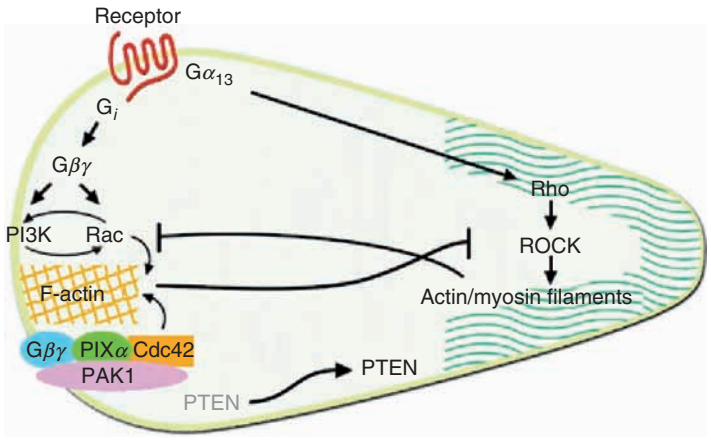

Figure 2 Localization of signaling events in a chemotaxing cell. The uropod is to the right and leading edge to the left. Note the uneven distribution of components for controlling the contractile actin/myosin filaments regulated by Rho in the uropod and the F-actin polymerization for forward motility controlled by PI3K and Rac in the leading edge. Also note the negative feedback regulation from front to back. Meili R and Firtel RA (2003) Two poles and a compass. *Cell* 114: 153–156.

(GEFs) are responsible for activating G-proteins like Rac. The connection of active Rac to the leading edge and its elevated levels of PIP_3 is that PIP_3 can activate some GEFs as can activated receptors. Also, once Rac is activated at the leading edge, it appears to feed back to recruit more PIP_3.

Just as important for successful chemotaxis as actin filament formation in the leading edge is myosin filament formation concentrated at the back of the cell. Myosin II provides power to retract the back end of the migrating cell and also to suppress pseudopod formation along the sides of the cell. Myosin filament formation in the uropod is regulated by the enzyme Akt/PKB, one of the PH domain proteins that is recruited to the PIP_3 of the leading edge, and by cGMP-binding proteins. Akt/PKB phosphorylates p21-activated kinase, which in turn regulates assembly of nonmuscle myosin II in the uropod. The second messenger cGMP has been known for some time to increase upon chemotaxis stimulation, and it appears that its role is both to induce myosin filament formation in the uropod and to activate myosin heavy chain kinase so as to inhibit filament formation in the leading edge.

The *D. discoideum* cells that are moving together by chemotaxis are able to 'count' their cell numbers to form the optimal aggregate size by secreting and detecting counting factor (CF). CF feeds into the signal transduction pathway via the cAMP receptor to effect cytoskeletal alterations that underlie directional motility. CF intersects the previously described pathways at the level of Akt/PKB translocation to the leading edge of the stimulated cell, to control how many cells aggregate together.

Conservation of the signaling pathways involved in the cytoskeletal rearrangements underlying directed motility between *D. discoideum* and neutrophils is remarkable, but there are obvious parallels and common molecular players in the change of cell shape in early developing neurons and in growth cone motility. Studies of *D. discoideum* and neutrophil chemotaxis also highlight the role of adaptation in efficient directional motility and the desensitization/resensitization of receptors.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Calcium Channels; Chemical Senses: Overview; Cyclic AMP (cAMP) Role in Learning and Memory; Pheromones and other Chemical Communication in Animals.

Further Reading

Chisholm RL and Firtel RA (2004) Insights into morphogenesis from a simple developmental system. *Nature Reviews Molecular Cell Biology* 5: 531–541.

Christensen ST, Guerra CF, Awan A, et al. (2003) Insulin receptor-like proteins in *Tetrahymena thermophila* ciliary membranes. *Current Biology* 13: r50–r52.

Christensen ST, Leick V, Rasmussen L, et al. (1998) Signaling in unicellular eukaryotes. *International Review of Cytology* 177: 181–253.

Eckert R (1972) Bioelectric control of ciliary activity. *Science* 176: 473–481.

Hassenzahl DL, Yorgey NK, Keedy MD, et al. (2001) Chemorepellent signaling through the PACAP/lysozyme receptor is mediated through cAMP and PKC in *Tetrahymena thermophila*. *Journal of Comparative Physiology A* 187: 171–176.

Jennings HS (1962) *Behavior of the Lower Organisms*. Bloomington, IN: Indiana University Press.

Jin M, Guan C-b, Jiang Y-A, et al. (2005) Ca^{2+}-dependent regulation of Rho GTPases triggers turning of nerve growth cones. *Journal of Neuroscience* 25: 2338–2347.

Kohidai L, Kovacs K, and Csaba G (2002) Direct chemotactic effect of bradykinin and related peptides – significance of amino- and carboxyterminal character of oligopeptides in chemotaxis of *Tetrahymena pyriformis*. *Cell Biology International* 26: 55–62.

Kuhlmann HW, Brunen-Nieweler C, and Heckmann K (1997) Pheromones of the ciliate *Euplotes octocarinatus* not only induce conjugation but also function as chemoattractants. *Journal of Experimental Zoology* 277: 38–48.

Kung C, Chang S-Y, Satow Y, et al. (1975) Genetic dissection of behavior in *Paramecium*. *Science* 188: 898–904.

Kung C, Preston RR, Maley ME, et al. (1992) *In vivo Paramecium* mutants show that calmodulin orchestrates membrane response to stimuli. *Cell Calcium* 13: 413–425.

Meili R and Firtel RA (2003) Two poles and a compars. *Cell* 114: 153–156.

Morange M (2006) What history tells us IV. Ciliates as models. . . of what? *Journal of Biosciences* 31: 27–30.

Ortenzi C, Alimenti C, Vallesi A, et al. (2000) The autocrine mitogenic loop of the ciliate *Euplotes raikovi*: The pheromone membrane-bound forms are the cell binding sites and potential signaling receptors of soluble pheromones. *Molecular Biology of the Cell* 11: 1445–1455.

Plattner H and Klauke N (2001) Calcium in ciliated protozoa: Sources, regulation, and calcium-regulated cell functions. *International Review of Cytology* 201: 115–208.

Pollard TD (2003) The cytoskeleton, cellular motlity and the reductionist agenda. *Nature* 422: 739–745.

Preston RR, Kink JA, Hinrichsen RD, et al. (1991) Calmodulin mutants and Ca^{2+}-dependent channels in *Paramecium*. *Annual Review of Physiology* 53: 309–319.

Stock C, Kruppel T, Key G, et al. (1999) Sexual behaviour in *Euplotes raikovi* is accompanied by pheromone-induced modifications of ionic currents. *Journal of Experimental Biology* 202: 475–483.

Stradal TEB, Rottner K, Disanza A, et al. (2004) Regulation of actin dynamics by WASP and WAVE family proteins. *Trends in Cell Biology* 14: 303–311.

Vallesi A, Giuli G, Ghiara P, et al. (1998) Structure-function relationships of pheromones of the ciliate *Euplotes raikovi* with mammalian growth factors: Cross-reactivity between Er-1 and Interleukin-2 systems. *Experimental Cell Research* 241: 253–259.

VanHaastert PJM and Devreotes PN (2004) Chemotaxis: Signalling the way forward. *Nature Reviews Molecular Cell Biology* 5: 626–634.

VanHouten J (1998) Chemosensory transduction in *Paramecium*. *European Journal of Protistology* 34: 301–307.

Chemoaffinity Hypothesis: Development of Topographic Axonal Projections

T McLaughlin and D D M O'Leary, The Salk Institute, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A critical function of the vertebrate nervous system is to interpret the environment through the connections of various sensory organs. To accomplish this task, the incoming information must be organized in an efficient manner. Perhaps the most efficient organization is achieved through the use of topographic maps that are present throughout the organism to process sensory information. A topographic map is a projection from one set of neurons to another wherein the receiving set of cells reflects the neighbor relationships of the projecting set. In the nervous system of higher vertebrates, topographic maps are common and include sensory body maps, tonotopic maps, and maps of visual space. Furthermore, topographic maps persist in some form throughout the circuitry, from first-order to higher-order connections.

Axon guidance is critical to the formation of appropriate neural connections. However, once the primary growth cone reaches the tissue that it will innervate, the problem switches from finding the correct linear path to finding the correct point in a two- (or three-) dimensional space. This latter problem, known as target recognition, has long been postulated to be under the control of gradients of axon guidance molecules. The chemoaffinity hypothesis, formally proposed half a century ago by Roger Sperry, outlines a conceptual framework by which neural connections can develop a topographic map. The predominant model for studying map development has been the primary visual projection formed by the axons of retinal ganglion cells (RGCs) to their major midbrain target, the optic tectum (OT) of fish, amphibians, and chick, or the superior colliculus (SC) of mammals. The representation of the retina onto the OT or SC can be simplified to the mapping of two sets of orthogonally oriented axes: the temporal–nasal (TN) axis of the retina along the anterior–posterior (AP) axis of the OT or SC, and the dorsal–ventral (DV) axis of the retina along the lateral–medial (LM) axis of the OT or SC (corresponding to the ventral–dorsal OT axis in nonmammalian vertebrates).

Studies of amphibians and fish, as well as invertebrates, have provided much insight into mechanisms of topographic map development and will certainly continue to have substantial impact on the field. However, this article will focus on chicks and mice because of their similarities in map development, their closer relationship to human biology, and the greater availability of relevant genetic experimentation.

The Chemoaffinity Hypothesis

The chemoaffinity hypothesis, formally proposed by Roger Sperry in 1963, presaged the dawning of the era of molecular mechanisms of map development. Sperry expressed his hypothesis in the following passage:

> I still go back to my initial interpretation proposing an orderly cytochemical mapping in terms of two or more gradients of embryonic differentiation that spread across and through each other with their axes roughly perpendicular. These separate gradients successively superimposed on the retinal and tectal fields and surroundings would stamp each cell with its appropriate latitude and longitude expressed in a kind of chemical code with matching values between the retinal and tectal maps.

His critical suggestion was that these molecular tags were associated with the neural tissue (e.g., membrane bound) and might be distributed in complementary gradients that mark corresponding points in both the sensory (e.g., retina) and target structures (e.g., OT or SC). Although Sperry based this hypothesis on his studies of regeneration of the retinotectal projection in newts and frogs, it provided the fledgling field of map development with direction. The basic foundation of the chemoaffinity hypothesis has largely been borne out, but subsequent experimental studies and mathematical models have refined it substantially to add multiple gradients of attractants and repellents and to more accurately account for the sequential phases of complex behaviors exhibited by RGC axons during map development in the OT or SC of higher vertebrates.

Based on the chemoaffinity hypothesis, each point in the OT or SC would have a unique molecular address determined by the graded distribution of topographic guidance molecules along the two tectal axes, and similarly each RGC would have a unique profile of receptors for those molecules that would result in a position-dependent, differential response to them by RGC axons. Over several decades following Sperry's hypothesis, the specificity of the projections of RGC axons to tectal cells was investigated further by the tracing of axonal projections following experimental manipulations, first in the regenerating retinotectal system of fish and amphibians, and later in the developing retinotectal or retinocollicular

projections with the development of high-resolution axon tracing techniques that work effectively in chicks and mammals. This body of evidence supported the basic hypothesis that the establishment of topographic projections involves the positional marking of RGC axons and their target sites in the OT or SC.

Prior to the discovery of the first topographic guidance molecules, the ephrins, the most compelling evidence for their existence came from the work of Friedrich Bonhoeffer and colleagues, who developed and used elegant *in vitro* assays, most prominently the membrane stripe assay. These assays were the first demonstrations of the presence of membrane-associated molecules that meet the criteria for topographic guidance molecules, and they were instrumental in these molecules' eventual identification.

Discovery of Graded Molecular Guidance Cues

Many labs using numerous approaches have carried out searches for topographic guidance molecules. Several cell surface molecules, such as TRAP and TOP_{AP}, with graded or restricted patterns in the retina, OT, or SC consistent with a role in mapping, were identified prior to the mid-1990s, but functional studies have yet to show such a role. The first description of graded molecules that proved to have properties of topographic guidance molecules came only in the mid-1990s with the cloning of two related genes, ephrinA2 (originally called Eph Ligand Family-1, or ELF-1) and ephrinA5 (originally called repulsive axon guidance signal, or RAGS), both of which are ligands of the receptor tyrosine kinase, EphA3 (originally named MEK4), expressed in a graded pattern by RGCs.

Eph receptors and ephrin ligands were not only the first topographic guidance molecules to be discovered, but they also have been defined to have prominent roles in both AP and DV mapping, and many family members are involved. The Eph family is the largest known family of receptor tyrosine kinases, comprising 14 Ephs and eight ephrins in mouse and 15 Ephs and nine ephrins in chick. Signaling through Ephs and ephrins has been implicated in a wide variety of biological processes. Ephs and ephrins are separated into two subclasses based on homology, the EphA/ephrinAs and EphB/ephrinBs, within which receptor–ligand binding and activation are promiscuous; some minimal cross-talk occurs between subclasses. EphAs and EphBs, as well as ephrinBs, are transmembrane proteins; ephrinAs are glycosylphosphatidylinositol (GPI)-linked to the cell membrane.

Ephs and ephrins can signal bidirectionally into either the Eph- or ephrin-expressing axon, and in some instances, Eph and ephrin signaling can be bifunctional, resulting in opposing axonal responses (i.e., attraction or repulsion), depending on the context. As described below, EphAs and ephrinAs act bidirectionally in AP mapping, whereas EphBs and ephrinBs act both bidirectionally and bifunctionally in DV mapping. A few other receptor–ligand pairings have also been implicated in mapping; most notably, Wnt3 and its receptors Ryk and Frizzled have been implicated in DV mapping.

An interesting historical note is the cloning of repulsive guidance molecule (RGM). RGM is the molecular activity that Bonhoeffer and colleagues believed to be the repellent membrane-associated molecule enriched in posterior membranes defined using the membrane stripe assay, as well as a growth cone collapse assay. In fact, the search for RGM resulted in the discovery of RAGS (ephrinA5), a molecule with very similar properties (i.e., GPI linked, similar size, repellent for RGC axons, and expressed in a similar gradient in the OT or SC). However, RGM knockout mice do not exhibit mapping defects, although some findings using ectopic expression in chick retina of an RGM receptor, neogenin, have suggested a role for it in AP mapping.

Mechanisms of Map Development

Determining the process by which RGCs establish topographic connections is critical for defining the roles of graded guidance molecules in map development and creating conceptual or computational models of the process. Frogs, fish, chicks, and rodents have been the vertebrates predominantly studied as models for development of retinotopic maps (Figure 1). These species exhibit important differences in the development of the visual system and retinotopic maps, as well as substantial differences in the absolute size of the OT or SC; for example, the two-dimensional area of the chick OT is more than a thousand times greater than that in frog and fish during the initial events in map development. Although each species has unique features that can be exploited, they have substantial differences in mechanisms employed by RGC axons to target their correct termination zone (TZ) and therefore also in the actual roles of topographic guidance molecules in controlling these axon behaviors.

Multiple Phases in the Development of a Topographic Map

Development of retinotectal topography in chicks and rodents is a multistep process (Figure 1(b)). RGC axons enter their midbrain target at its anterior edge and extend posteriorly parallel to the AP axis of the OT or SC. In birds and mammals, essentially all RGC axons extend well posterior to the topographically appropriate location of their future TZ. Thus,

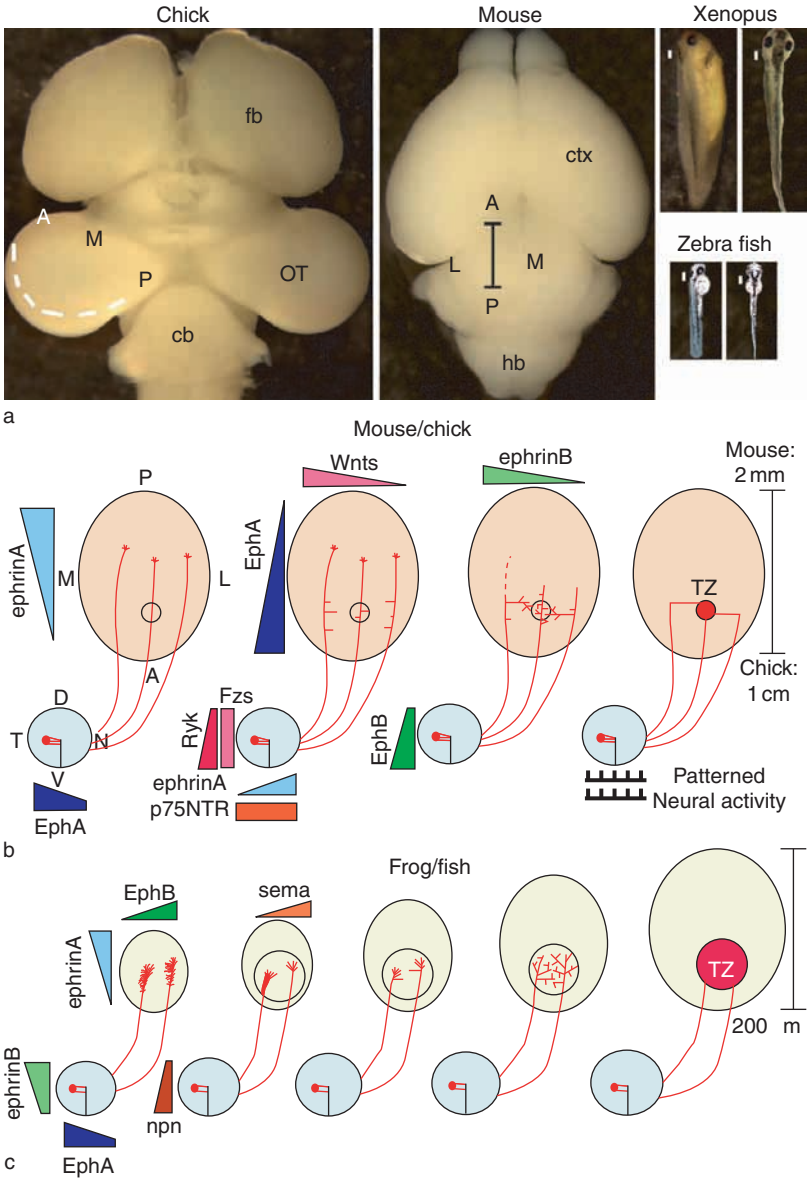

Figure 1 (a) Photographs, at critical ages during topographic map formation, of the primary model systems currently used for the examination of retinocollicular and retinotectal map development. All photographs are at the same scale. The mouse superior colliculus (SC), on the dorsal surface of the midbrain (bracket, 2 mm), is approximately as large as an entire developing zebra fish, and the chick optic tectum (OT) is approximately the size of an entire Xenopus tadpole at these ages. The Xenopus and zebra fish OT sizes are denoted by the small white bars adjacent to the brain. The mouse and zebra fish are used commonly for the potential genetic manipulations available in these organisms. Chick and Xenopus are especially useful for overexpression, mis-expression, and knockdown studies. Retinal ganglion cells from these organisms can be cultured and manipulated *in vitro*. (b, c) Mechanisms and molecules controlling retinotopic mapping in (b) mouse and chick and (c) frog and zebra fish. The names and distributions of molecules and activities demonstrated *in vivo* to control, in part, the mapping mechanisms used at each stage are listed. The gradients may represent the consensus distribution for a combination of related molecules (e.g., EphAs), which are not listed individually due to distinctions in the individual members expressed and the precise distributions between species. Molecules and activities other than those listed are likely to participate in topographic mapping. (b) In mouse and chick, retinal ganglion cell (RGC) axons enter the SC and OT, respectively, over a broad LM extent and significantly overshoot their future termination zone (TZ, circle). Interstitial branches form *de novo* from the axon shaft around the anterior–posterior (AP) position of the future TZ. Branches are directed along the lateral-medial (LM) axis of the OT or SC toward the position of the future TZ, where they arborize in a domain encompassing the forming TZ. The broad, loose array of arbors is refined to a dense TZ in the topographically appropriate location. (c) In frog and fish, the RGC growth cone extends to the TZ, where it forms a terminal arbor through a process of backbranching from the area immediately behind the growth cone. The tectum expands as terminal arborizations elaborate and refine into a mature TZ. cb, cerebellum; ctx, cortex; D, dorsal; fb, forebrain; hb, hindbrain; N, nasal; T, temporal; V, ventral. Modified from McLaughlin T, Hindges R, and O'Leary DDM (2003) Regulation of axial patterning of the retina and its topographic mapping in the brain. *Current Opinion in Neurobiology* 13: 57–69.

the growth cones of RGC axons in birds and mammals do not target their future TZ but instead extend a millimeter or more posterior to it. In addition, RGC axons from the same retinal location enter and grow across the OT or SC with an aberrantly broad distribution over its LM axis. Branches form *de novo* from the axon shaft hundreds of micrometers or even millimeters behind the growth cone. These interstitial branches form roughly perpendicular to the primary axon and preferentially extend along the LM axis toward their future TZ.

In summary, the topographically ordered connections of the mature projection are established by primary branches of RGC axons that form *de novo* interstitially along the length of RGC axons. These branches form along the shaft of the primary axon with a topographic bias for the correct position along the AP axis, which is the first indication of a topographic response. They subsequently extend along the LM axis to the appropriate site of the future TZ, where they go on to arborize.

AP Mapping

Historically, models of topographic mapping have been based on the action of molecular activities that promote axon growth. However, with the identification in the late 1980s by Bonhoeffer and colleagues of repellent activities concentrated in posterior chick OT that preferentially affect temporal RGC axons, it became clear that topographic mapping was controlled in a fashion different from that previously recognized. These findings eventually resulted in the incorporation of repellent activities into models of topographic mapping, but until recently these models focused on guiding RGC axons to their correct TZ. These models were based on a gradient-determined repulsion, in which an RGC axon would stop its growth when it reached a threshold level of repulsion found at the AP position of the future TZ. Although such models could explain retinotectal map development in amphibians and fish (**Figure 1(c)**), they can not account for aspects of map development along the AP axis in chick OT and rodent SC, including the initial AP overshoot of RGC axons and the subsequent topographic branching interstitially along the axon shaft.

Models for AP mapping in rodent SC and chick OT must explain the topographic bias in the formation of interstitial branches, which is arguably the key feature of AP map development. Models that in principle can account for topographic branching have been proposed recently, including models based on parallel gradients of a branch-repellent (or inhibitory) activity and a branch-promoting activity. However, the field has gravitated toward a model based on opposing gradients along the AP axis, each of which inhibits branching. This model is consistent with the gradients and actions of topographic guidance activities defined over the past few years and, it is important to note, can account for the phases exhibited by RGC axons during map development. As described below, the opposing gradients are based on opposing graded expression of EphAs and ephrinAs along both the TN retinal axis and the AP OT or SC axis, with one repellent gradient caused by EphA forward signaling and the other by ephrinA reverse signaling.

RGC axons respond differentially, in a manner relating to their origin along the TN retinal axis, to opposing gradients of repellent activities along the AP axis of the OT or SC. The low-to-high AP gradient of repellent activity has been shown by many complementary studies to be due to EphA forward signaling, controlled by a high-to-low TN gradient of EphA receptors on RGC axons and a low-to-high AP gradient of ephrinAs in the OT or SC (**Figure 2**). These studies include manipulating expression of EphA receptors by RGCs and of ephrinAs in the target. Recent experimental data have provided evidence that the opposing gradient of repellent activity is caused by ephrinA reverse signaling controlled by a low-to-high TN gradient of ephrinAs on RGC axons and a high-to-low AP gradient of EphAs expressed in the SC. The critical findings are that retinal axons are repelled by EphA7 in the "Bonhoeffer" stripe assay, a repulsion blocked by ephrinA5-fc that blocks EphA function, and that EphA7 knockout mice have retinocollicular mapping defects consistent with the loss of the high-to-low AP graded expression of EphA7 in the SC (EphA7 is not expressed by RGCs). Because ephrinAs are anchored to the cell membrane by a GPI linkage and lack an intracellular domain, to reverse signal they must associate with transmembrane proteins capable of activating intracellular signaling pathways. Recent studies have shown that the p75 neurotrophin receptor mediates the repellent effect of ephrinA reverse signaling on RGC axons on binding EphAs, and is required for appropriate development of the retinotopic map in the SC.

DV Mapping

Topographic specificity along the LM axis emerges through the bidirectional guidance of branches that form along RGC axons with an AP bias as described above. Branches that extend from RGC axons located lateral to their future TZ grow medially whereas branches that extend from RGC axons located medial to their future TZ grow laterally. Branches that reach

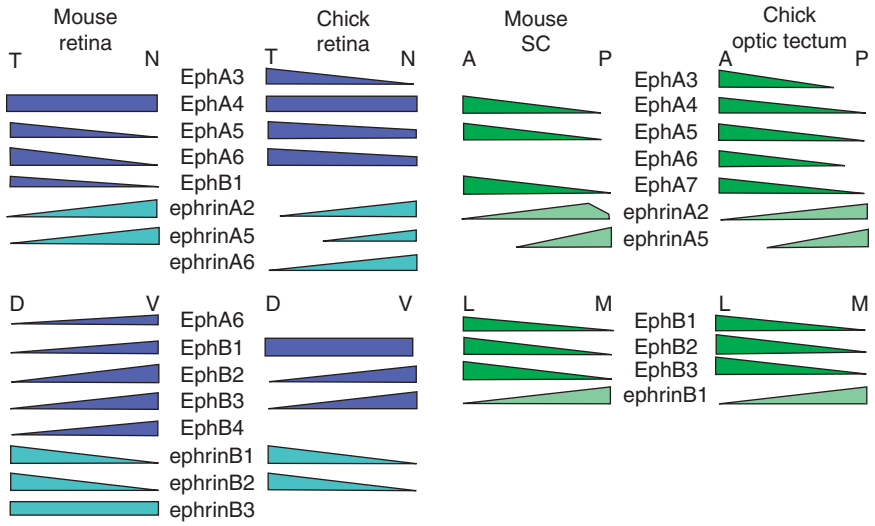

Figure 2 Distribution of Ephs and ephrins in the retina and optic tectum (OT) and superior colliculus (SC) of chick and mouse, respectively. The charts represent generalized patterns for individual Ephs and ephrins in the retina and OT or SC during critical stages in the development of the retinotopic map (E15–P7 in mouse and E6–E14 in chick). The charts are likely incomplete, and the distributions for several molecules are dynamic during development. The sizes and shapes of the gradients are generalized, and relative expression levels should not be inferred. A, anterior; D, dorsal; L, lateral; M, medial; N, nasal; P, posterior; T, temporal; V, ventral. Modified from McLaughlin T and O'Leary DDM (2005) Molecular gradients and development of retinotopic maps. *Annual Review of Neuroscience* 28: 327–355.

the area of the nascent TZ selectively form complex arbors. Therefore, not only is this guidance of interstitial branches a critical feature for DV retinotopic mapping, but the molecular mechanisms that control it must account for the bidirectional guidance along the LM axis of the OT or SC of branches of axons that have the same retinal origin and presumably express the same set of receptors at similar levels.

Studies of the molecular control of DV mapping have implicated EphB and ephrinBs, exhibiting both bidirectional signaling and bifunctional action, as well as Wnts and their receptors. In retina, EphB receptors are expressed by RGCs during map development in an overall low-to-high DV gradient, complemented by an overall high-to-low DV gradient of ephrinBs. In both chick OT and mouse SC, ephrinB1 is expressed in a low-to-high LM gradient, complemented by an overall high-to-low LM EphB gradient (Figure 2). Analyses of EphB2 and EphB3 mutant mice, with and without reverse signaling intact, show aberrant LM mapping due to defects in the guidance of interstitial branches and show that ephrinB1 acts as a branch attractant via EphB2/B3 forward signaling. Electroporation studies in chick corroborate these findings, and in addition show that high levels of ephrinB1 repel branches. Taken together, these studies show that in mice and chicks, ephrinB1 can act through EphB forward signaling as both an attractant and a repellent: a branch located lateral to its nascent TZ is attracted up the gradient of ephrinB1 toward it, whereas a branch located medial to its nascent TZ is repelled down the ephrinB1 gradient toward it. The primary axon is not influenced by ephrinB1, but instead the effects of eprhin-B1 are specific to interstitial branches and are context dependent: The location of the branch on the ephrinB1 gradient in relation to the location of its future TZ and its EphB level determine its response. In frogs, ephrinB reverse signaling has also been implicated in retinotopic mapping, but such a role has yet to be shown in mice and chicks (Figure 1(c)).

In principle, EphBs and ephrinB1 could account for DV mapping through their actions as bifunctional (i.e., attractant and repellent) and bidirectional (i.e., EphBs and ephrinBs both can act as receptors) guidance molecules. However, recent evidence suggests that Wnt signaling is also involved in DV mapping. Wnt3 is expressed in a high-to-low medial-lateral gradient in chick OT and mouse SC, and Wnt receptors, Ryk and Frizzled family members, are expressed in an overall low-to-high DV gradient by RGCs. Functional studies show that Ryk mediates RGC axon repulsion by higher levels of Wnt3 whereas Frizzled receptors mediate an attractant effect of lower levels of Wnt3 on dorsal RGC axons. Thus, the 'morphogen' Wnt3 appears to act as a topographic guidance molecule in the OT and to cooperate with EphB-ephrinB1 to regulate DV mapping. In zebra fish, Sema3D has also been implicated in DV retinotopic mapping by repelling ventral RGC axons from ventral OT.

Refinement of the Retinotopic Map

As described above, in mice and chicks, all arbors are formed by interstitial branches that preferentially arborize at, or in the vicinity of, the topographically appropriate location of the nascent TZ. In frogs, refinement of individual arbors involves a combination of directed branch dynamics; further, the continued but disparate growth of the retina and OT requires a continuous small-scale remodeling throughout life. These processes act to shape and refine developing arbors and are likely dependent on trkB/brain-derived neurotrophic factor (BDNF) interactions and neural activity. Compared to map remodeling in chicks or mice, refinement in fish and frogs is a precise shaping of arbors rather than major remodeling of a topographically diffuse projection that involves the elimination of substantial lengths of a primary axon and entire arbors that it might have formed along the eliminated segment.

In chicks and mice, the initial collection of arbors is loosely organized around the topographically appropriate position of the future TZ and requires a substantial degree of remodeling to develop the precise connections evident in the mature retinotopic map. Remodeling requires the removal of inappropriately located branches and elimination of overshooting portions of RGC axons. In mice, the remodeling of the retinocollicular projection occurs before the onset of vision but is coincident with a period of correlated spontaneous waves of activity that propagate across the retina. These waves are generated by networks of connections between RGCs and amacrine cells, a type of retinal interneuron, and correlate the activity of neighboring RGCs, thereby relating an RGC's position to its pattern of activity. These amacrine cell–RGC circuits are cholinergic and can be selectively disrupted by deletion of the $\beta 2$ subunit of the nicotinic cholinergic receptor. Analysis of $\beta 2$ knockout mice, which lack retinal waves but retain wild-type levels of RGC activity, show that correlated patterns of RGC activity are required for the large-scale remodeling of the retinocollicular projection into a refined map. Further, these studies demonstrate that a later resumption of correlated activity does not lead to proper map refinement, thereby revealing a brief early critical period for retinotopic map remodeling in mice.

Concluding Statements

From studies to date, it is clear that other receptor–ligand pairings are required to generate proper retinotopic maps. Several groups have carried out forward- and reverse-genetic screens to identify additional candidate genes involved in retinotopic mapping, or screens that could produce candidate genes as a by-product. For example, the large-scale Tubingen genetic screen in mutagenized zebra fish was designed in part to identify genes involved in RGC axon pathfinding and mapping. This near-saturation screen has resulted in the identification of about a hundred mutants, representing scores of genes; a subset of these mutants have defects in DV or AP mapping in the retinotectal projection. These mutants are being investigated further. In addition, microarray screens have identified many known and unknown genes expressed in gradients or restricted patterns along the TN or DV axes of the developing mouse retina, some of which will be found to be involved in retinotopic mapping.

It is interesting that Sperry's chemoaffinity hypothesis was, in part, a response to contemporaneous views that axonal projections could self-organize. Though it is clear that topographic guidance molecules are critical for mapping, the role that a system plays in its own development is an intriguing and understudied facet of mapping. It seems likely in many systems that the projecting axons themselves affect the development of the projection. For example, RGC axons may alter the expression levels and distributions of guidance molecules as they elaborate arbors in the target due in part to the presence of guidance molecules on RGC axons, including multiple ephrins and Ephs on RGC axons. Thus, guidance information itself is dynamic, and patterned activity has a role in mapping – therefore the progressive development of the topographic map may feed back on itself in multiple ways and influence its own development.

See also: Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Lamina-Specific Neuronal Connections; Spontaneous Patterned Activity in Developing Neural Circuits; Topographic Maps: Molecular Mechanisms.

Further Reading

Brown A, Yates PA, Burrola P, et al. (2000) Topographic mapping from the retina to the midbrain is controlled by relative but not absolute levels of EphA receptor signaling. *Cell* 102: 77–88.

Cheng HJ, Nakamoto M, Bergemann AD, and Flanagan JG (1995) Complementary gradients in expression and binding of ELF-1 and Mek4 in development of the topographic retinotectal projection map. *Cell* 82: 371–381.

Drescher U, Kremoser C, Handwerker C, et al. (1995) In vitro guidance of retinal ganglion cell axons by RAGS, a 25 kDa tectal protein related to ligands for Eph receptor tyrosine kinases. *Cell* 82: 359–370.

Feldheim DA, Kim YI, Bergemann AD, et al. (2000) Genetic analysis of ephrin-A2 and ephrin-A5 shows their requirement in multiple aspects of retinocollicular mapping. *Neuron* 25: 563–574.

Feldheim DA, Nakamoto M, Osterfield M, et al. (2004) Loss-of-function analysis of EphA receptors in retinotectal mapping. *Journal of Neuroscience* 24: 2542–2550.

Flanagan JG (2006) Neural map specification by gradients. *Current Opinion in Neurobiology* 16: 59–66.

Frisen J, Yates PA, McLaughlin T, et al. (1998) Ephrin-A5 (AL-1/RAGS) is essential for proper retinal axon guidance and topographic mapping in the mammalian visual system. *Neuron* 20: 235–243.

Hindges R, McLaughlin T, Genoud N, Henkemeyer M, and O'Leary DDM (2002) EphB forward signaling controls directional branch extension and arborization required for dorsal ventral retinotopic mapping. *Neuron* 35: 475–487.

Liu Y, Berndt J, Su F, et al. (2004) Semaphorin3D guides retinal axons along the dorsoventral axis of the tectum. *Journal of Neuroscience* 24: 310–318.

Mann F, Ray S, Harris WA, and Holt CE (2002) Topographic mapping in dorsoventral axis of the *Xenopus* retinotectal system depends on signaling through ephrin-B ligands. *Neuron* 35: 461–473.

McLaughlin T, Hindges R, Yates PA, and O'Leary DDM (2003) Bifunctional action of ephrin-B1 as a repellent and attractant to control bidirectional branch extension in dorsal ventral retinotopic mapping. *Development* 130: 2407–2418.

McLaughlin T and O'Leary DDM (2005) Molecular gradients and development of retinotopic maps. *Annual Review of Neuroscience* 28: 327–355.

McLaughlin T, Torborg CL, Feller MB, and O'Leary DDM (2003) Retinotopic map refinement requires spontaneous retinal waves during a brief critical period of development. *Neuron* 40: 1147–1160.

Nakamoto M, Cheng HJ, Friedman GC, et al. (1996) Topographically specific effects of ELF 1 on retinal axon guidance *in vitro* and retinal axon mapping *in vivo*. *Cell* 86: 755–766.

Rashid T, Upton AL, Blentic A, et al. (2005) Opposing gradients of ephrin-As and EphA7 in the superior colliculus are essential for topographic mapping in the mammalian visual system. *Neuron* 47: 57–69.

Roskies AL and O'Leary DDM (1994) Control of topographic retinal axon branching by inhibitory membrane-bound molecules. *Science* 265: 799–803.

Schmitt AM, Shi J, Wolf AM, et al. (2006) Wnt-Ryk signaling mediates medial-lateral retinotectal topographic mapping. *Nature* 439: 31–37.

Simon DK and O'Leary DDM (1992) Development of topographic order in the mammalian retinocollicular projection. *Journal of Neuroscience* 12: 1212–1232.

Simon DK and O'Leary DDM (1992) Responses of retinal axons *in vivo* and *in vitro* to position-encoding molecules in the embryonic superior colliculus. *Neuron* 9: 977–989.

Sperry R (1963) Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proceedings of the National Academy of Sciences of the United States of America* 50: 703–710.

Trowe T, Klostermann S, Baier H, et al. (1996) Mutations disrupting the ordering and topographic mapping of axons in the retinotectal projection of the zebrafish, Danio rerio. *Development* 123: 439–450.

Walter J, Kern-Veits B, Huf J, Stolze B, and Bonhoeffer F (1987) Recognition of position specific properties of tectal cell membranes by retinal axons *in vitro*. *Development* 101: 685–696.

Wong RO, Chernjavsky A, Smith SJ, and Shatz CJ (1995) Early functional neural networks in the developing retina. *Nature* 374: 716–718.

Yates PA, Roskies AR, McLaughlin T, and O'Leary DDM (2001) Topographic specific axon branching controlled by ephrin-As is the critical event in retinotectal map development. *Journal of Neuroscience* 21: 8548–8563.

Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System

T C Westfall, St. Louis University School of Medicine, St. Louis, MO, USA

© 2009 Published by Elsevier Ltd.

Introduction

Acetylcholine is the principal neurotransmitter at all autonomic preganglionic nerve terminals (sympathetic and parasympathetic), all postganglionic parasympathetic neuroeffector junctions, a few postganglionic sympathetic neuroeffector junctions (e.g., eccrine sweat glands), the somatic motor nerve neuromuscular junction, and cholinergic synapses in the central nervous system (CNS). The synthesis, storage, release, and inactivation of acetylcholine are similar at all these structures. This is probably also true for nonneuronal cholinergic systems in various parts of the body (e.g., human airway). The neurochemical events involved in the life cycle of acetylcholine are depicted in Figure 1.

Synthesis of Acetylcholine

Choline and Choline Transport

Critical to the synthesis of acetylcholine is the availability of choline. There is little *de novo* synthesis of choline in cholinergic neurons; therefore, choline is provided mostly from the diet. Choline is a dietary component essential for normal function of all cells, not just cholinergic neurons. Choline or its metabolites ensure the structural integrity and signaling functions of all cell membranes; it is the major source of methyl groups in the diet, and it obviously directly affects cholinergic neurotransmission. Cholinergic neurons have especially large requirements for choline because they use choline not only for membrane synthesis but also for acetylcholine synthesis. Choline is taken up from the extracellular space by two transport systems. First is the ubiquitous low-affinity, sodium-independent transport system that is inhibited by hemicholinium 3 with a K_i of about $50\,\mu mol\,l^{-1}$. Second is the high-affinity, sodium- and chloride-dependent, hemicholinium 3-sensitive ($K_i = 10–100\,nmol\,l^{-1}$) choline transport system. The latter form is found predominantly in cholinergic neurons and is responsible for providing choline for acetylcholine synthesis. Once acetylcholine is released from cholinergic neurons following the arrival of action potentials, acetylcholine is hydrolyzed by acetylcholinesterase (AChE; see the section titled 'Enzymatic hydrolysis: AChE') to free acetate and choline. Choline is recycled after reuptake into the nerve terminal of cholinergic cells and utilized for acetylcholine synthesis. Under many circumstances, this reuptake and availability of choline appear to serve as the rate-limiting step in acetylcholine synthesis. Acetylcholine synthesis and release are sustained during persistent and prolonged neuronal stimulation, as long as choline is available and the transporter is functional. The low-affinity choline uptake transporter is used to synthesize membrane phospholipids. These lipids are reservoirs from which choline can also be used by choline neurons for acetylcholine synthesis.

Despite the fact that the high-affinity choline uptake system was first described in the1960s, the gene for the high-affinity choline transporter (CHT_1) has only recently been cloned from a variety of species: *Caenorhabditis elegans*, mouse, *Torpedo*, *Limulus*, and humans. It is interesting that CHT_1 is not homologous to neurotransmitter transporters but is homologous to the Na^+-dependent glucose transporter family. With the molecular identification of the CHT_1, it was discovered that it does not predominantly reside at the nerve terminal plasma membrane but rather in intracellular vesicular structures. It has been observed that CHT_1 co-localizes with synaptic vesicle markers such as vesicle-associated membrane protein 2 and vesicular acetylcholine transporter (VAChT), at least when expressed in a cholinergic cell line. Although information is still incomplete, it has been hypothesized that arrival of action potentiates in nerve terminals results in exocytosis of synaptic vesicles, release of acetylcholine into the synaptic cleft, and increased trafficking of CHT_1 to the plasma membrane, where it functions to take up choline after hydrolysis of acetylcholine. Because choline transport is rate-limiting for acetylcholine synthesis, increased availability of choline via its transport by CHT_1 would favor an increase in acetylcholine stores to maintain high levels of transmitter release during neuronal stimulation. This also suggests that the availability of CHT_1 at the cell surface is dynamically regulated in a manner very similar to the regulation of the exocytosis of synaptic vesicles. Much recent information has been obtained on how trafficking of CHT_1 between the cell surface and subcellular compartments can regulate choline uptake activity. Nevertheless, the precise mechanisms involved in maintaining the distribution of CHT_1 predominantly in intracellular vesicles rather than at the terminal surface like other neurotransmitter transporters is unclear.

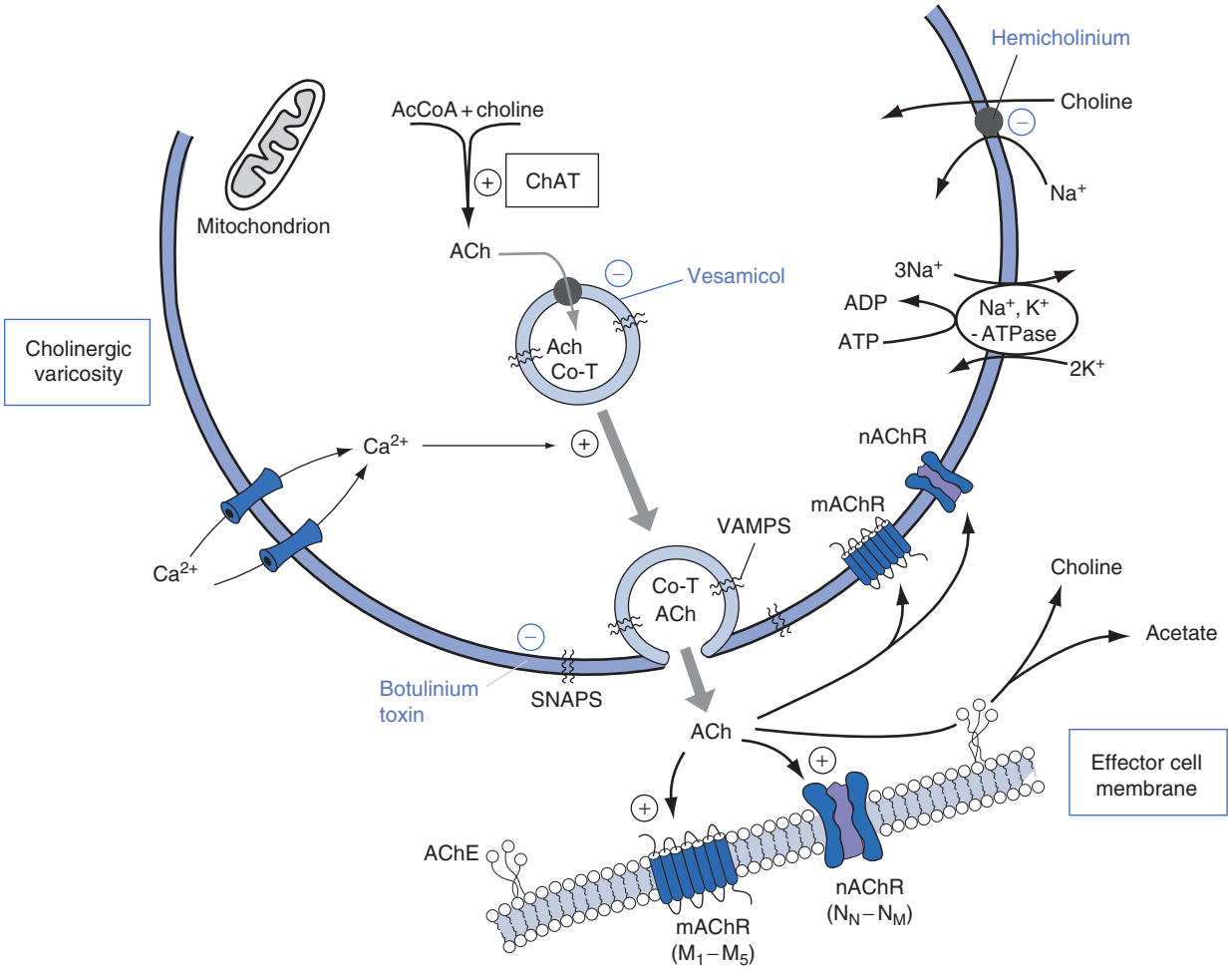

Figure 1 Schematic representation of a cholinergic neuroeffector junction showing features of the synthesis, storage, and release of acetylcholine (ACh) and receptors on which ACh acts. The synthesis of ACh in the varicosity depends on the uptake of choline via a sodium-dependent carrier. This uptake can be blocked by hemicholinium. Choline and the acetyl moiety of acetyl coenzyme A (AcCoA), derived from mitochrondria, form ACh, a process catalyzed by the enzyme choline acetyltransferase (ChAT). ACh is transported into the storage vesicle by another carrier that can be inhibited by vesamicol. ACh is stored in vesicles along with other potential cotransmitters (Co-T) such as adenosine triphosphate (ATP) and vasoactive intestinal polypeptide at certain neuroeffector junctions. Release of ACh and the Co-T occurs on depolarization of the varicosity, which allows the entry of Ca^{2+} through voltage-dependent Ca^{2+} channels. Elevated $[Ca^{2+}]_{in}$ promotes fusion of the vesicular membrane with the cell membrane, and exocytosis of the transmitters occurs. This fusion process involves the interaction of specialized proteins associated with the vesicular membrane, called vesicle-associated membrane proteins (VAMPs) and the membrane of the varicosity, called synaptosome-associated proteins (SNAPs). The exocytotic release of ACh can be blocked by botulinum toxin. Once released, ACh can interact with the muscarinic receptors (mAChRs), which are G-protein-coupled receptors, or nicotinic receptors (nAChRs), which are ligand-gated ion channels, to produce the characteristic response of the effector. ACh can also act on presynaptic mAChR or nAChR to modify its own release. The action of ACh is terminated by metabolism to choline and acetate by the enzyme acetylcholinesterase (AChE), which is associated with synaptic membranes. ADP, adenosine diphosphate; ATPase, adenosine triphosphatase. From Westfall TC and Westfall DP (2006) Neurotransmission: The autonomic and somatic motor nervous systems. In: Brunton LL, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's the Pharmacological Basis of Therapeutics*, 11th edn., pp. 137–182. New York: McGraw-Hill.

Choline Acetyltransferase

The synthesis of acetylcholine occurs within the cytoplasm of cholinergic nerve terminals by a reaction in which choline is acetylated with acetyl coenzyme A (CoA) by the enzyme choline acetyltransferase. Acetyl CoA is derived mainly from glycolysis and is ultimately produced by the enzyme pyruvate dehydrogenase. The synthesis of acetyl CoA occurs at the inner membrane of mitochondria, and it is transported by citrate to the cytoplasm, where citrate is freed by citrate lyase.

Choline acetyltransferase is synthesized in the perkaryon and transported along the length of the axon to the terminal by axoplasmic flow.

Storage of Acetylcholine

Following its synthesis in the cytoplasm of the nerve terminal, acetylcholine is transported into synaptic vesicles by VAChT, which uses a proton electrochemical gradient to move acetylcholine to the inside of the organelle. Hydrophobicity algorithm prediction of secondary structure of VAChT suggests a protein comprising 12 transmembrane domains, with hydrophilic N- and C-termini in the cytoplasm. Sequence homology places VAChT into a family of transporter proteins that includes two vesicular monoamine transporters. An adenosine triphosphatase that pumps protons into the vesicle provides the energy necessary for the transport. Transport of protons out of the vesicle is coupled to uptake of acetylcholine into the vesicle and against a concentration gradient via an acetylcholine antiporter. Two types of vesicles appear to exist in cholinergic terminals: an electron-luscent vesicle (40–50 nm in diameter) and dense core vesicles (80–150 nm). The vesicle core contains both acetylcholine and adenosine triphosphate (ATP), which are dissolved in the fluid phase with metal ions (Ca^{2+} and Mg^{2+}) and a proteoglycan vesiculin. In some cholinergic terminals, there are also peptides such as vasoactive intestinal polypeptide (VIP) that act together with ATP and acetylcholine as cotransmitters. The peptides are thought to reside in the larger dense-core vesicles. Vesicular membranes are rich in lipids (cholesterol and phospholipids) as well as proteins.

The VAChT allows for the transport of acetylcholine agonist, which has a considerable concentration gradient and is saturable and adenosine triphosphatase-dependent. Vesamicol is a drug which inhibits the transport of acetylcholine via targeting the acetylcholine H^+ antiporter channel. Vesamicol leads to a reduction in acetylcholine storage and interferes with neurosecretion. Although not precisely known, the content of synaptic vesicles is estimated to range from 1000 to more than 50 000 molecules per vesicle. There may also be large amounts of acetylcholine in the extravesicular cytoplasm.

Release of Acetylcholine

Release of acetylcholine and cotransmitters (e.g., ATP and VIP) occurs on depolarization of the nerve terminals and takes place by exocytosis. Depolarization of the terminals allows the entry of Ca^{2+} through voltage-gated Ca^{2+} channels. Elevated Ca^{2+} concentration promotes fusion of the vesicular membrane with the plasma membrane, allowing exocytosis to occur. The molecular mechanisms involved in the release and regulation of release have been the subject of intense study but are still not completely understood. As mentioned, acetylcholine, like other neurotransmitters, is stored in vesicles located at special release sites, close to presynaptic membranes and ready for release on the appropriate stimulus. The vesicles initially dock and are primed for release. A multiprotein complex appears to form and attach the vesicle to the plasma membrane close to other signaling elements. The complex involves proteins from the vesicular membrane and the presynaptic neuronal membrane, as well as other components that help link them together. It has been established that these various synaptic proteins, including the plasma membrane protein syntaxin and synaptosomal protein 25 kDa (SNAP-25), and the vesicular membrane protein, or synaptobrevin, form a complex. This complex interacts in an ATP-dependent manner with soluble *N*-ethylmalemide-sensitive fusion protein and soluble SNAPs. The ability of synaptobrevin, syntaxin, and SNAP-25 to bind SNAPs has led to their designation as SNAP regulators (SNARES). It has been hypothesized that most, if not all, intracellular fusion events are mediated by SNARE interactions. Important evidence supporting the involvement of SNARE proteins in transmitter release comes from the fact that botulinum neurotoxins and tetanus toxin, which block neurotransmitter release, proteolyze these three proteins.

Two pools of acetylcholine appear to exist. One pool, the 'depot' or 'readily releasable' pool, consists of vesicles located near the plasma membrane of the nerve terminals and is filled with newly synthesized transmitter. Depolarization of the terminals causes these vesicles to release acetylcholine rapidly or readily. The other pool, the 'reserve' pool, seems to replenish the readily releasable pool and may be required to sustain acetylcholine release during periods of sustained or intense nerve stimulation.

Regulation of Acetylcholine Neurotransmission

It is now well accepted that receptors are located on perkaryon, dendrites, and axons of neurons when they may respond to neurotransmitters or neuromodulators released from the same neurons (autoreceptors) or adjacent neurons or cells (heteroreceptors). Cholinergic nerve terminals also contain autoreceptors and heteroreceptors.

Acetylcholine neurotransmission is therefore subject to complex regulation by mediators including acetylcholine itself acting on M_2 and M_4 autoinhibitory receptors and nicotinic acetylcholine excitatory receptors. Acetylcholine-mediated inhibition of acetylcholine release via muscarinic cholinergic receptors M_2 and M_4 is thought to represent a physiological

negative feedback control mechanism. Acetylcholine released from cholinergic neurons may also alter the release of other neurotransmitters. For instance, at the neuroeffector junction in the myenteric plexus in the gastrointestinal tract or sinoatrial node of the heart, postganglionic sympathetic and parasympathetic terminals lie in juxtaposition to each other. The release of acetylcholine not only produces autoinhibition of its own release but may attenuate the release of norepinephrine via an inhibitory action of muscarinic heteroreceptors located on sympathetic terminals. The reverse is also true: after its release of sympathetic nerve terminals, norepinephrine can inhibit the release of acetylcholine by an action of α_{2A} and α_{2C} heteroreceptors on parasympathetic nerve terminals. Muscarinic and nicotinic auto- and heteroreceptors also represent drug targets, or targets for locally formed autacoids, hormones, or transmitters.

Inactivation of Acetylcholine

Enzymatic Hydrolysis: AChE

Subsequent to its release from cholinergic neurons, acetylcholine is subjected to a number of inactivation processes, including (1) diffusion from the site of release, (2) dilution in extracellular fluids, (3) binding to nonspecific sites, and (4) enzymatic destruction. The most important of these by far is the enzymatic destruction of acetylcholine to its hydrolysis products, acetic acid and choline. The reaction is catalyzed by AChE. While AChE is found in cholinergic neurons (dendrites, perikarya, and axons), it is distributed more widely than cholinergic synapses. It is highly concentrated at postsynaptic sites, especially at the neuromuscular junction. AChE exists in red blood cells and the placenta as well as nervous tissue. A second enzyme is butyrylcholinesterase (BuChE). BuChE exists in intestine and skin and abundantly in plasma. AChE can be distinguished from BuChE by selectivity of acetylcholine over butyrylcholine hydrolysis.

BuChE shows a wider substrate capacity than AChE does. Almost all the pharmacological effects of AChE drugs are due to the inhibition of AChE, with a consequent accumulation of endogenous acetylcholine in the vicinity of the nerve terminal. Distinct but single genes encode AChE and BuChE in mammals, and the diversity of molecular structure of AChE arises from alternate messenger RNA processing. Complementary DNAs for several species, including *Torpedo*, *Drosophila*, nematode, rat, mouse, and human AChE, as well as mouse, rabbit, and human BuAChE, are known.

The contact between acetylcholine and the surface of AChE occurs at two places on the protein, One is the anionic site, which is negatively charged and stereospecific. It attracts the positively charged nitrogen atom of acetylcholine and binds the attracted methyl groups by Van der Waals forces. The other site is the 'esteratic' site, which combines the electrophilic carboxy carbon atom of the acetyl group and the serine hydroxyl group of the esteratic site (the nucleophilicity of which is enhanced by the histadine imidazole group).

The three stages for the hydrolysis are as follows: (1) equilibrium is established between the enzyme and acetylcholine, forming an enzyme substrate complex; (2) the complex reacts to release choline and leaves an acetylated enzyme; and (3) the acetylated enzyme reacts with water to give acetic acid and regenerated enzyme with a $t_{½}$ of approximately 40 μs.

Although the principal function of AChE is to inactivate acetylcholine at cholinergic neuronal junctions, it has been reported to have multiple biological functions that are not obvious. These include neritogenesis, cell adhesion, synaptogenesis, amyloid fiber assembly, and activation of dopamine receptors, hematopoiesis, and thrombopoiesis, to name but a few.

Cholinergic Receptors

Once released from cholinergic neurons, acetylcholine produces its physiological or pharmacological effects by first interacting with specific receptors. The various responses are mediated by the interaction of the transmitter with different types of receptors located in different organs, tissues, or cells. There are two general types of cholinergic receptors: muscarinic and nicotinic. The original classification was based on acetylcholine's mimicking the effect of the alkaloid muscarine (muscarinic) or nicotine (nicotinic). Multiple subtypes are now known to exist. Originally the subtypes of receptors were defined by the rank order of potency of agonists and the specificity of antagonists (e.g., atropine as an antagonist of muscarinic receptors and hexamethonium or curare as antagonists for nicotinic acetylcholine receptors (nAChRs)). More recently, the various subtypes have been identified by molecular cloning of the genes.

nAChRs

The nAChRs are members of a superfamily of ligand-gated ion channels (ionotropic receptors). They primarily exist at the neuromuscular junction, autonomic ganglia and adrenal medulla, parasympathetic and sympathetic nerve terminals, and central nervous system. These receptors are the physiologic target of the neurotransmitter acetylcholine as well as naturally occurring alkaloids and synthetic drugs. The nAChR

Table 1 Characteristics of subtypes of nicotinic acetylcholine receptors

Receptor (primary receptor subtype)[a]	*Main synaptic location*	*Membrane response*	*Molecular mechanism*	*Agonists*	*Antagonists*
Skeletal muscle (N_m) $(\alpha_1)_2\ \beta_1\ \varepsilon\ \delta$ Adult $(\alpha_1)_2\ \beta_1\ \gamma\ \delta$ Fetal	Skeletal neuromuscular junction (postjunctional)	Excitatory; endplate depolarization; skeletal muscle contraction	Increased cation permeability (Na^+; K^+)	ACh Nicotine Succinylcholine	Atracurium Vecuronium d-Tubocurarine Pancuronum α-conotoxin α-bungarotoxin
Peripheral neuronal (N_N) $(\alpha_3)_2\ (\beta_4)_3$	Autonomic ganglia; adrenal medulla	Excitatory; depolarization firing of postganglion neuron; depolarization and secretion of catecholamines	Increased cation permeability (Na^+; K^+)	ACh Nicotine Epibatidine Dimethylphenylpiper-azinum	Trimethaphan Mecamylamine
Central neuronal (CNS) $(\alpha_4)_2\ (\beta_4)_3$ (α-btox-insensitive)	CNS; pre- and postjunctional	Pre- and postsynaptic excitation Prejunctional control of transmitter release	Increased cation permeability (Na^+; K^+)	Cytisine, epibatidine Anatoxin A	Mecamylamine Dihydro-β-erythrodine Erysodine Lophotoxin
$(\alpha_7)_5$ (α-btox-sensitive)	CNS; pre- and postsynaptic	Pre- and postsynaptic excitation Prejunctional control of transmitter release	Increased permeability (Ca^{2+})	Anatoxin A	Methllycaconitine α-Bungarotoxin α-Conotoxin IMI

[a]Nine individual subunits have been identified and cloned in human brain which combine in various confirmations to form individual receptor subtypes. The structure of individual receptors and the subtype composition is incompletely understood, however. Only a finite number of naturally occurring functional nAChRs constructs have been identified to date. α-btox, α-bungarotoxin. Adapted from Westfall TC and Westfall DP (2006) Adrenergic agonists and antagonists. In: Brunton LL, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's the Pharmacological Basis of Therapeutics*, 11th edn., pp. 237–296. New York: McGraw-Hill.

Table 2 Characteristics of muscarinic acetylcholine receptor subtypes

Receptor	*Size: chromosome location*	*Cellular and tissue location*[a]	*Cellular response*[b]	*Functional response*[c]
M_1	460 aa 11q 12–13	CNS; most abundant in cerebral cortex, hippocampus, and striatum Autonomic ganglia Glands (gastric and salivary) Enteric nerves	Activation of PLC; ↑ IP_3 and ↑ DAG → ↑ Ca^{2+} and PKC Depolarization and excitation (↑ sEPSP) Activation of PLD_2 ↑ AA Couples via $G_{q/11}$	Increased cognitive function (learning and memory) Increased seizure activity Decreased dopamine release and locomotion Increased depolarization of autonomic ganglia Increased secretions
M_2	466 aa 7q 35–36	Widely expressed in CNS, heart, smooth muscle, autonomic nerve terminals	Inhibition of adenylyl cyclase ↓ cAMP Activation of inwardly rectifying K^+ channels Inhibition of voltage-gated Ca^{2+} channels Hyperpolarization and inhibition Couples via G_i/G_o (PTX sensitive)	Heart: SA node: slowed spontaneous depolarization; hyperpolarization ↓HR AV node: decrease in conduction velocity Atrium: ↓ refractory period, ↓ contraction Ventricle: slight ↓ contraction *Smooth muscle:* ↑*Contraction* *Peripheral nerves:* Neural inhibition via autoreceptors and heteroreceptor) ↓ Ganglionic transmission. *CNS:* Neural inhibition ↑ Tremors; hypothermia; analgesia
M_3	590 aa Iq 43–44	Widely expressed in CNS (< than other mAChRs) Abundant in smooth muscle and glands Heart	Activation of PLC; ↑ IP_3 and ↑ DAG → ↑ Ca^{2+} and PKC Depolarization and excitation (↑ sEPSP) Activation of PLD_2 ; PLA_2; ↑ AA Couples via $G_{q/11}$	Smooth muscle: ↑ contraction (predominant in some, e.g., bladder) Glands: ↑ secretion (predominant in salivary gland) Increases food intake, body weight fat deposits. Inhibits dopamine release Synthesis of NO
M_4	479 aa 11p 12–11.2	Preferentially expressed in CNS, particularly forebrain	Inhibition of adenylyl cyclase ↓cAMP Activation of inwardly rectifying K^+ channels Inhibition of voltage-gated Ca^{2+} channels Hyperpolarization and inhibition Couples via G_i/G_o (PTX sensitive)	Autoreceptor and heteroreceptor mediated inhibition of transmitter release in CNS and periphery. Analgesia; cataleptic activity Facilitation of dopamine release
M_5	532 aa 15q 26	Expressed in low levels in CNS and periphery. Predominant mAchR in dopamine neurons in VTA and substantia nigra	Activation of PLC; ↑ IP_3 and ↑ DAG → ↑ Ca^{2+} and PKC Depolarization and excitation (↑ sEPSP) Activation of PLD_2, PLA_2; ↑ AA Couples via $G_{q/11}$	Mediator of dilation in cerebral arteries and arterioles (?) Facilitates dopamine release Augmentation of drug-seeking behavior and reward (e.g., opiates, cocaine)

[a]Most organs, tissues, and cells express multiple mAChRs.

[b]M_1, M_3, and M_5 mAChRs appear to couple to the same G-proteins and signal through similar pathways. Likewise, M_2 and M_4 mAChRs couple through similar G-proteins and signal through similar pathways.

[c]Although multiple subtypes of mAChRs coexist in many tissues, organs, and cells, one subtype may predominate in producing a particular function, or there may be equal predominance.

mAChR, muscarinic acetylcholine receptor; PLC, phospholipase C; IP_3, inositol-1,4,5-triphosphate; DAG, diacylglycerol; PKC, protein kinase C; EPSP, excitatory postsynaptic potential; PLD_2, phospholipase D; AA, arachidonic acid; CNS, central nervous system; PLA, phospholipase A; cAMP, cyclic adenosine monophosphate; SA node, sinoatrial node; AV node, atrioventricular node; HR, heart rate; PTX, pertussis toxin; VTA, ventral tegmental area.

Table 3 Summary of peripheral cholinergic receptors and responses

Organ	*Response*	*Receptor*
Eye		
Sphincter (circular) muscle of iris	Contraction (miosis)	M_3, M_2
Iris	Contraction (accommodation)	M_3, M_2
Ciliary muscle		
Heart		
SA node	Decrease in rate (negative chronotropic)	$M_2 \gg M_3$
Atria	Decrease in contractile strength (negative inotropic)	$M_2 \gg M_3$
	Decrease in refractive period	
AV node	Decrease in conduction velocity (negative dromotropic effect)	$M_2 \gg M_3$
	Increase in refractory period	$M_2 \gg M_3$
Ventricle	Small decrease in contractile strength	$M_2 \gg M_3$
Blood vessels[a]		
Arteries	Dilation (via nitric oxide from endothelial cells)	M_3
Veins	Dilation (via nitric oxide from endothelial cells)	M_3
Lung		
Bronchial muscle	Contraction (bronchoconstriction)	$M_2 = M_3$
Bronchial glands	Stimulation	M_3, M_2
Gastrointestinal tract		
Motility	Increase	M_3 (M_2)
Sphincter	Relaxation	M, M_2
Secretion	Stimulation	M_3, M_2
Urinary bladder		
Detrusor muscle	Contraction	$M_3 > M_2$
Trigone and sphincter	Relaxation	$M_3 > M_2$
Glands		
Sweat, salivary, lacrimal, bronchial, etc. Nasopharyngeal	Secretion	M_3
Nerves		
Parasympathetic terminal	Decrease in release of acetylcholine	M_2, M_4
Sympathetic terminal	Decrease in release of norepinephrine	M_2, M_4
Autonomic ganglia	Stimulation	N_N or N_1
Skeletal muscle		
(Neuromuscular junction)	Contraction	N_M or N_2

[a]Most blood vessels are not innervated by parasympathetic nerves but have M_3 receptors on endothelial cells.

is a pentamer consisting of five subunits around a central ion channel pore. Several subunits have been identified: α, β, γ, δ, and ε. All these subunits share 35–50% homology with one another. Depending on the location of the receptor, it has a different combination of subunits. At the neuromuscular junction (N_1 or N_m), the AChR is composed of two α (α_1) subunits; one β (β_1); one δ; and in embryonic or denervated muscle, one γ subunit. In the adult innervated muscle, the γ is replaced with one ε subunit. Functional neuronal nAChRs (CNS N_n) also exist as pentamers composed of two α and three β subunits. Eight human α subunits (α_2–α_7, α_9, and α_{10}) and three β subunits (β_2–β_4) have been cloned. Both the muscle and the neuronal nAChR share structural and functional properties with other ionotropic receptors, such as γ-aminobutyric acid A receptor, 5-hydroxytryptamine type 3, and glycine receptors. Autonomic ganglia (N_2) and the adrenal medulla form homomeric α_7 and heteromeric α_3/β_4, with $(\alpha_3)_2$ $(\beta_4)_3$ being the most prevalent.

The α subunits are responsible for binding acetylcholine, and the conformational change in the α subunits is responsible for permitting ion flow through the central pore of the receptor. The pentameric structure of the CNS–neuronal receptor and the large molecular diversity of its subunits (α_2–α_8, β_2–β_4) suggest there could be a large number of nAChRs with different physiological properties. They may subserve discrete functions and obviously represent multiple drug targets. The stoichiometry of the nAChR in brain is the subject of intense study. Selective ligands are becoming more and more prominent, but it is not yet possible to make a pharmacological classification based on subtypes. The characteristics of subtypes of nAChR are described in Table 1.

Muscarinic Acetylcholine Receptors

There are five distinct subtypes of muscarinic acetylcholine receptors (mAChRs), M_1–M_5, each produced by a distinct gene and all members of the family of G-protein-coupled receptors (metabotropic receptors). Like the different nAChRs, M_1–M_5 have distinct anatomical locations in the periphery and CNS and different chemical specificities. The mAChRs have been shown to be present in virtually all organs, tissues, and cell types, but various subtypes predominate in various tissues and organs (e.g., M_2 in the heart, M_3 in the detrusor muscle of the bladder). The mAChRs are located in smooth muscle, exocrine and endocrine glands, and the myocardium. They mediate the actions of acetylcholine in organs and tissues innervated by postganglionic parasympathetic nerves and postganglionic sympathetic nerves innervating eccrine sweat glands. They are also present at sites that lack parasympathetic nerves (e.g., most blood vessels). Table 2 summarizes the effects produced by activation of mAChRs. In the CNS, mAChRs are involved in regulating a large number of cognitive, behavioral, motor, and autonomic functions although some of these are also influenced by nAChRs.

Effects produced by activating mAChR are coupled through G-protein-induced changes in membrane-bound effectors and production of second-messenger molecules. The M_1, M_2, and M_3 receptor subtypes couple through the pertussis toxin-insensitive G_{11} and G_{13}, resulting in stimulation of phospholipase C activity. This results in the hydrolysis of membrane phosphatidylinositol 4,5 diphosphate to form inositol triphosphate (IP_3) and diaceylglycerol. IP_3 causes release of intracellular Ca^{2+} from the endoplasmic reticulum, which results in smooth muscle contraction as well as secretion from the appropriate cell. Diaceylglycerol activates protein kinase C, resulting in phosphorylation of numerous proteins and various physiological responses. Activation of M_1, M_2, and M_3 receptors also results in the activation of phospholipase A_2, leading to the release of arachidonic acid and consequent eicosanoid synthesis.

Stimulation of the M_2/M_4 receptors leads to interaction with G_i and G_o proteins, which results in inhibition of adenylyl cyclase and the resultant decrease in cyclic adenosine monophosphate, activation of inwardly rectifying K^+ channels, and inhibition of voltage-gated Ca^{2+} channels. Functional consequences are hyperpolarization and inhibition of excitable membranes.

Physiological Responses Produced Following Increases in Cholinergic Neurotransmission

Table 3 depicts the physiological responses due to increases in cholinergic neurotransmission or stimulation of parasympathetic neurons.

See also: Acetylcholine Neurotransmission in CNS; Acetylcholinesterase Inhibitors and Alzheimer's Disease; Acetylcholinesterase; Basal Ganglia: Acetylcholine Interactions and Behavior; Cholinergic System; Cholinergic Pathways in CNS; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Muscarinic Receptors: Autonomic Neurons; Neuromuscular Junction (NMJ): Acetylcholinesterases; Neutrotransmission and Neuromodulation: Acetylcholine; Nicotinic Acetylcholine Receptors.

Further Reading

DeBiasi M (2002) Nicotinic mechanisms in the autonomic control of organ systems. *Journal of Neurobiology* 53: 568–579.

Dhein S, van Koppen CJ, and Brodde OE (2001) Muscarinic receptors in the mammalian heart. *Pharmacological Reviews* 44: 161–182.

Ferguson SM and Blakely RD (2004) The choline transporter resurfaces: New roles for synaptic vesicles? *Molecular Interventions* 4: 22–37.

Jahn R, Lang T, and Südhof T (2003) Membrane fusion. *Cell* 112: 519–533.

Lindstrom JM (2000) Acetylcholine receptors and myasthenia. *Muscle & Nerve* 23: 453–477.

Ribeiro FM, Black SA, Prado VF, Rylett RJ, Ferguson SS, and Prado MA (2006) The "ins" and "outs" of the high-affinity choline transporter CHT1. *Journal of Neurochemistry* 97: 1–12.

Soreq H and Seidman S (2001) Acetylcholinesterase: New roles for an old actor. *Nature Reviews Neuroscience* 2: 294–302.

Taylor P (2006) Anticholinesterase agents. In: Brunton LL, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th edn., pp. 201–216. New York: McGraw-Hill.

van Koppen CJ and Kaiser B (2003) Regulation of muscarinic acetylcholine receptor signaling. *Pharmacology & Therapeutics* 98: 197–220.

Wang H and Sun X (2005) Desensitized nicotinic receptors in brain. *Brain Research. Brain Research Reviews* 48: 420–437.

Wess J (2004) Muscarinic acetylcholine receptor knockout mice: Novel phenotypes and clinical implications. *Annual Review of Pharmacology and Toxicology* 44: 423–450.

Westfall TC and Westfall DP (2006) Neurotransmission: The autonomic and somatic motor nervous systems. In: Brunton LL, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th edn., pp. 137–182. New York: McGraw-Hill.

Cholinergic Pathways in CNS

A C Cuello, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Acetylcholine as a Central Nervous System Transmitter

Historically, acetylcholine (ACh) was considered to be the most important central nervous system (CNS) transmitter. As early as 1936, Quastel demonstrated that ACh was synthesized in the brain. Indeed, ACh was the first substance to be proposed to act as a CNS transmitter in the pioneering work of Marthe Vogt and Feldberg in the 1940s and 1950s, largely based on prior lessons obtained with this substance in the peripheral nervous system by a number of outstanding pharmacologists, ranging from Otto Loewi to Sir Henry Dale. Later, Fatt, Katz, and collaborators provided evidence for a quantal release of ACh in the neuromuscular junction, laying the foundations for the idea that this transmitter substance was somehow 'packaged' in nerve terminals. The realization that this was the case came about with the ultrastructural finding in 1955 of synaptic vesicles in CNS nerve terminals by De Robertis and Bennett, and also Palay and Palade, and the compelling demonstration that isolated synaptic vesicles obtained from subcellular fractionation were enriched in ACh content, as shown by De Robertis in Buenos Aires and Whitaker in Cambridge. These studies not only were important for the realization that ACh was a bona fide CNS transmitter but they also provided strong foundations to the chemical synaptic transmission hypothesis, a concept currently applied to a wide variety of so-called classical (e.g., amines, aminoacids) and nonclassical transmitters (e.g., peptides).

CNS Cholinergic Synapses

Contrary to historical expectations, current methods have demonstrated that cholinergic synapses do not account for the majority of transmitter-specific synapses in the CNS. In the cerebral cortex, for example, cholinergic synapses account for about 5–7% of the total synaptic population. However, considerable variations in the number of cholinergic synapses across the CNS should be expected, given the degree of differential concentrations of the ACh-synthesizing enzyme (choline acetyltransferase, ChAT) in diverse brain areas. ChAT is localized mainly in the cytosolic compartment. The CNS cholinergic synapses are endowed with the expected set of organelles (e.g., mitochondria, synaptic vesicles) and molecules (e.g., classical synaptic proteins), but in addition they contain 'signature' proteins such as the high-affinity choline transporter (CHT1) and the vesicular acetylcholine transporter (VAChT). These two proteins endow cholinergic synapses with their physiological specifications. CHT1 plays an essential role in maintaining the required level of ACh synthesis, as it is the availability of choline rather than the ChAT enzymatic activity that is the rate-limiting factor for the synthesis of ACh. As expected, CHT1 is localized in synaptic membranes to facilitate the incorporation of the choline precursor.

Dietary levels of choline have been shown to have an important impact on the replenishment of ACh stores in the CNS. Besides its predicted synaptic membrane localization, CHT1 is abundantly present in synaptic vesicles. This is somewhat perplexing, as ChAT is primarily a cytosolic enzyme and the bulk of the ACh synthesis is known to occur in the cytoplasm. This location might be a consequence of the retrieval of synapses after exocytosis, but the vesicular CHT1 might also play a role in mobilizing choline from vesicles to cytoplasm and vice versa. The VAChT is heavily localized in membranes of synaptic vesicles of cholinergic presynaptic endings. Its function is the straightforward uptake of newly synthesized cytoplasmatic ACh into the vesicular stores. The ACh-enriched vesicles are thus ready for release on demand. The newly synthesized ACh appears to be the first to be released, probably from vesicles closer to the synaptic membranes, while more distant vesicles containing 'older' ACh pools are seemingly located in the cholinergic presynaptic terminal further away from the point of synaptic contact. The cholinergic synaptic vesicles, from the early De Robertis descriptions, are recognized as being small, clear, and roundish. Figure 1 represents the main features of CNS cholinergic synapses.

After the identification of the CNS cholinergic neurons and their projections (see later), it was proposed that their terminations at target sites, particularly in the cerebral cortex, were nonsynaptic in nature. In other words, it was assumed that CNS cholinergic presynaptic boutons release ACh diffusely in a 'cloud-like' modality. The idea of 'volume' or 'extrasynaptic' transmission of acetylcholine in the CNS was born out of the prevalent notion that CNS terminations containing monoamines are apparently nonsynaptic, in a manner reminiscent of the peripheral autonomic axonal terminations. The cholinergic cortical terminations were, therefore, included in the category of nonsynaptic neurons due to the lack of evidence for

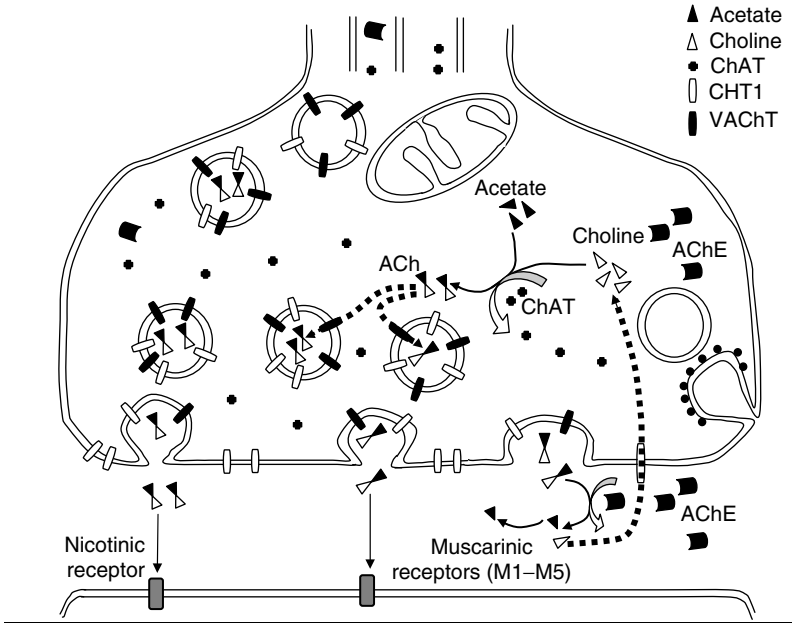

Figure 1 Schematic representation of typical CNS cholinergic synapses, as described in the text. ACh, acetylcholine; ChAT, choline acetyltransferase; CHT1, choline transporter 1; VAChT, vesicular acetylcholine transporter; AChE, acetylcholinesterase; M1–M5, muscarinic receptors; M2, presynaptic muscarinic receptor 2; N, nicotinic receptors.

synaptic specialization in ChAT-immunoreactive (IR) boutons when observed by electron microscopy. However, Mesulam and collaborators, using ChAT antibodies, noted that cholinergic boutons in the cerebral cortex of the human brain mainly displayed symmetric synaptic contacts. A more comprehensive ultrastructural study in the rat neocortex, applying improved protocols for tissue preservation and utilizing antibodies against VAChT, revealed that the cortical cholinergic presynaptic boutons in their majority establish classical synaptic contacts with symmetric membrane specializations (parallel membranes of equal thickness in the pre- and postsynaptic sides of contacts). These classical synaptic contacts occur preferentially in dendritic shafts and differ in their termination pattern from the bulk of 'noncholinergic' presynaptic boutons. Cholinergic presynaptic boutons are seldom observed contacting neuronal cell bodies and only occasionally contacting dendritic spines. These latter synapses are exclusively of an asymmetric nature (i.e., the postsynaptic membrane is notably thicker than the opposing presynaptic membrane). Figure 2 illustrates the typical ultrastructural characteristics of cholinergic (VAChT-IR) presynaptic sites (boutons) in layer V of the rat parietal cortex.

Main CNS Cholinergic Pathways

Although the existence of CNS cholinergic neurons was well accepted, it took nearly half a century for their unequivocal identification. Early attempts to define CNS systems operating with ACh applied immunohistochemical methods to reveal the presence (activity) of acetylcholinesterase (AChE), the ACh-inactivating enzyme, a procedure which was pioneered by Shute and Lewis at Cambridge University. This method revealed cell groups that later proved to be indeed cholinergic in nature, but it had the intrinsic problem that it demonstrated both 'cholinergic' and 'cholinoceptive' neurons. This approach was modified later by Butcher and collaborators when they introduced the so-called pharmaco-histochemical technique. This technique involved the inhibition of AChE, followed, after few hours, by the AChE histochemical reaction, to demonstrate newly generated AChE in cell body groups, thus avoiding the confounding images of AChE-positive neurites in the CNS tissue. While this approach was of some value, the problem remained that it also revealed several AChE-positive cells which were not cholinergic in nature. We now know that there is not a 100% correlation between ChAT-IR and AChE-positive histochemical reactions in CNS neurons. Notable examples are the conspicuous presence of AChE-positive material in dopaminergic substantia nigra and noradrenergic locus coeruleus neurons in the absence of ChAT immunoreaction. The chemical isolation of ChAT allowed the generation of reliable polyclonal and monoclonal antibodies, which opened the door for the reliable demonstration of the main groups of CNS cholinernergic neurons. Thus, a number

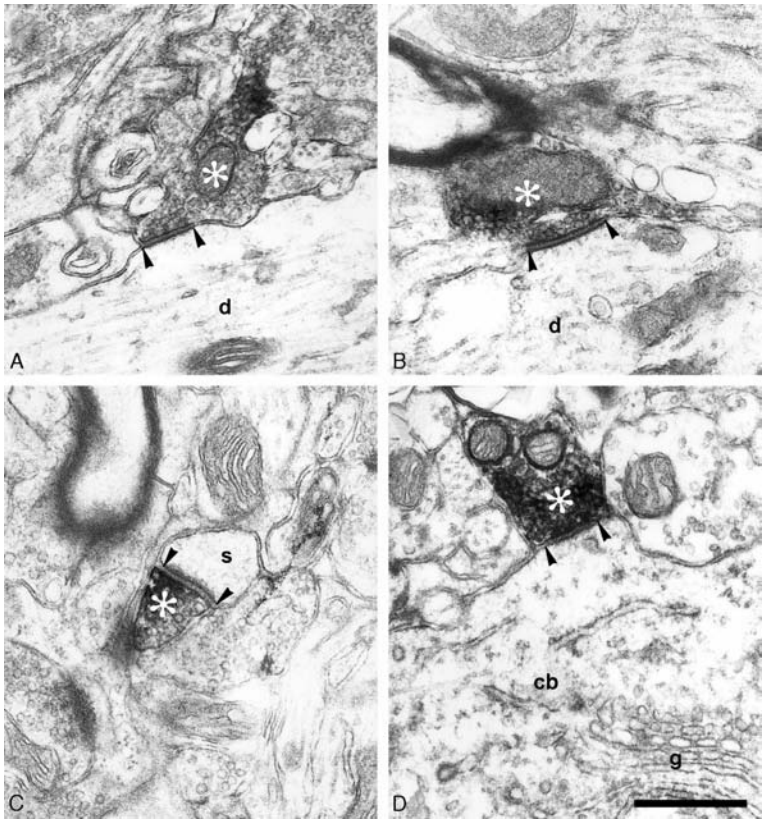

Figure 2 Electron micrograph images of VAChT-IR boutons displaying synaptic specializations in layer V of the rat parietal cortex. (A) and (B) VAChT-IR boutons (asterisks) establishing symmetric synaptic contacts with dendritic branches (d). Note the aggregation of synaptic vesicles adjacent to the presynaptic membranes. (C) Illustration of a small VAChT-IR varicosity which is occasionally observed in a synaptic contact with a dendritic spine (s); note the synapse is of the asymmetric type. (D) Micrograph of one of the few VAChT-IR boutons establishing a synaptic contact with a cell body (cb), probably belonging to a nonpyramidal neuron. Synaptic contacts are indicated by two arrowheads; g, Golgi complex. Scale bar (for all micrographs) = 0.25 μm. Reprinted from Turrini P, Casu MA, Wong TP, et al. (2001) Cholinergic nerve terminals establish classical synapses in the rat cerebral cortex: Synaptic pattern and age-related atrophy. *Neuroscience* 105(2): 277–285, with permission from Elsevier.

of reports arose during the early 1980s, applying immunohistochemistry to depict the localization of CNS ChAT-IR cell bodies and fiber tracts. At the time, the Swedish histochemists Fuxe and Dahlstrom introduced a letter and number nomenclature to define the catecholaminergic pathway. This proved popular and was widely accepted. Mesulam and collaborators, following the same principles, applied a Ch and number nomenclature to define CNS cholinergic nuclear organizations, as revealed with anti-ChAT monoclonal antibodies. These studies were the most exacting and comprehensive and provided a useful guide for further investigations, although the Ch classification is not universally applied. Following these developments, a number of lesion and track-tracing experiments revealed the existence of a variety of CNS pathways and local circuit cholinergic neurons. The main cholinergic cell groups and pathways are represented in Figure 3 in a parasagittal view of the rat CNS, where the prominent and extensively studied basalocortical and septo-hippocampal pathways are highlighted in gray.

In summary, these immunohistochemical investigations demonstrated that the ventral telencephalon displays a continuous stream of cholinergic cell bodies, notably in the olfactory tubercle, islands of Calleja, all components of the diagonal band, medial septum, nucleus preopticus magnocellularis, and the innermost region of the globus pallidus. These neurons are usually referred to (as an ensemble) as the basal forebrain cholinergic system. In rodents, the ChAT immunoreactive neurons distributed loosely in the innermost third of the globus pallidus represent the nucleus basalis, which is responsible for the large majority of cholinergic input to the cerebral cortex. The rodent cerebral cortex contains in addition a few homogenously dispersed, local circuit, fusiform cholinergic cell bodies. The organization of this 'basalocortical' pathway is very different from that of the

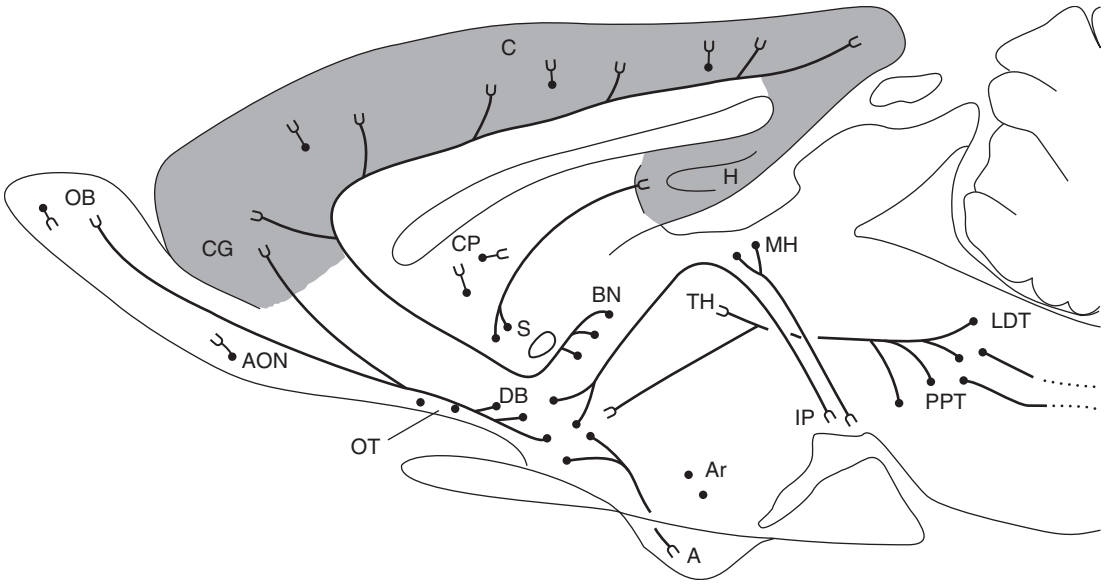

Figure 3 Schematic sagittal view of the rat CNS, representing the main cholinergic pathways, as discussed in the text. OB, olfactory bulb; AON, anterior olfactory nucleus; DB, diagonal band; S, septum (cholinergic neurons restricted to medial division); CG, cingulated cortex; CP, large interneurons in the caudate putamen; H, hippocampus; BN, nucleus basalis; A, amygdala; TH, thalamus; Ar, arcuate nucleus; PPT, peduculo-pontine tegmental nucleus; LDT, lateral doral tegmental nucleus; C, cortex; IP, nucleus interpenducularis; MH, medial habenula; OT, olfactory tubercle. For clarity, cholinergic motor neurons and cholinergic preganglionar neurons are not represented. The gray bands represent the major forebrain basalocortical and septal–hippocampal cholinergic pathways. Adapted from Cuello AC and Sofroniew MV (1984) The anatomy of the CNS cholinergic neurons. *Trends in Neurosciences* 7: 74–78.

human brain. In the human brain, the equivalent cholinergic group is composed of large (ergo 'magnocellularis') multipolar neurons located in a flattened nuclear organization lying below the anterior commissure. This organization would correspond to the neuronal group, which is nowadays referred to as the nucleus basalis magnocellularis of Meynert, or simply the nucleus of Meynert. However, it has to be stressed that this nomenclature is only applicable to the human brain and to subhuman primate brains. The rodent structure should be referred to simply as the nucleus basalis. These nuclei, both in rodents and humans, correspond to Ch4 in Mesulam's classification. A possible explanation of this differential organization of the nucleus basalis in primates and rodents might reside in the fact that in rodents the globus pallidus is split where the medial division of the globus pallidus becomes the entopenducular nucleus (also containing cholinergic neurons), while the external division becomes the globus pallidus. In consequence, in rodents, this component of the basal forebrain cholinergic neurons does not find the obstacle of the anterior commissure barrier, invading the innermost third of the globus pallidus in a diffuse manner. In the rodent, the most caudal portions of the nucleus basalis become a flattened neuronal group located between the optic tract and the last portion of the corpus striatum. Much attention has been paid to this nuclear organization, as it has been reported to be compromised in Alzheimer's disease (AD), a disease that, in advanced states, is accompanied by the loss of cortical cholinergic markers (see the section entitled 'CNS cholinergic neurons and AD').

It is well documented that the nucleus basalis innervates all areas of the neocortex in a topographic-specific manner, that is, more rostral regions of the nucleus basalis will supply cholinergic fibers to the most rostral cortical regions, and the most caudal to the most caudal. While the 'basalocortical cholinergic pathway' supplies fibers to the entire neocortex, the cingulate and entorhinal cortex most probably is innervated by cholinergic septal and diagonal band neurons. Figure 4 illustrates the appearance and relative density of cortical cholinergic presynaptic boutons as compared to the overall density of presynaptic boutons as revealed by synaptophysin.

The other prominent basal forebrain cholinergic nuclear group is located in two arching bands of neurons meeting at the midline at dorsal aspects of the septal nucleus. This group of neurons is part of a continuum of cholinergic neurons from more rostrally located cholinergic neurons of the diagonal band of Broca. These septal cholinergic neurons project heavily to all regions of the hippocampal formation via the fimbria-fornix; this is referred to as the 'septal–hippocampal' cholinergic pathway.

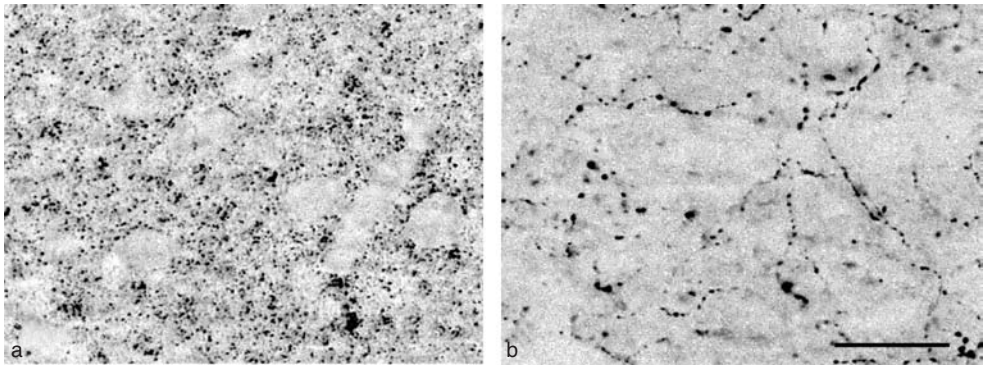

Figure 4 Immunocytochemical demonstration of cholinergic presynaptic boutons in the cerebral cortex of the rat (parietal cortex, lamina (b) as revealed with anti-VAChT antibodies. The pattern and density of the total population of cortical presynaptic boutons, as revealed with antisynaptophysin antibodies, are illustrated in (a). Scale bar = 20 μm.

These two cholinergic nuclei (basalis and medial septum) are frequently referred to in the literature as 'the' basal forebrain cholinergic neurons; however, the basal forebrain cholinergic system includes other nuclei forming a continuum in that region of the brain (as previously described). These two cholinergic neuronal groups (basolocortical and septo-hippocampal) are heavily involved in higher CNS functions such as attention, learning, and memory, and are compromised in aging and in AD (see the section entitled 'CNS cholinergic neurons and AD').

The largest CNS concentration of cholinergic fibers occurs, without any doubt, in the nucleus caudatus in humans and in the caudoputamen complex in rodents. These fibers and nerve terminals originate in very large local circuit neurons distributed throughout this region of the basal ganglia. Indeed, the highest CNS concentration of ACh and cholinergic markers in general occurs in the basal ganglia. The cholinergic system here plays an important role in mechanisms implicated in movement control and there is good physiological and pharmacological evidence for a dopaminergic/cholinergic balance in this region.

Other less conspicuous cholinergic cell groups are found in the thalamus (midline, intralaminar), epithalamus (medial habenula), and hypothalamus (posterior region and arcuate nucleus), with less defined functions than those of the forebrain cholinergic neurons. The medial habenula nuclei project heavily to the nucleus interpeduncularis via the so-called fasciculus retroflexus. In the brain stem, the motor and autonomic preganglionar neurons of cranial nerves are all of cholinergic nature, as expected from classical physiological studies preceding the advent of immunocytochemistry. In addition to these motor neurons, there is a collection of ChAT-IR neurons distributed in various regions of the brain stem. Thus, there is a diffuse reticular system (lateral reticular nucleus, parvicellular) and several nuclei, including the pedunculo-pontine tegmental (PPT), lateral dorsotegmental (LDT), parabrachial, nucleus trapezoid, superior olive, and the raphe system (nucleus magnus and obscurus). PPT and LDT assembly of cholinergic neurons projects mainly to the thalamus, epithalamus, and tectum, but also sends axons more ventrally to the basal forebrain, lateral hypothalamus, and substantia nigra. Descending branches from the PPT and LDT cholinergic nuclei reach diverse nuclei in the lower brain stem, and long projections from these neurons can reach the spinal cord. In the spinal cord, occasional small cholinergic neurons can be observed in the dorsal horn in addition to the preganglionar nuclei of the lateral horn and the large motor neurons of the ventral horn.

Although a number of cholinergic neuronal pathways have been described, there is still an incomplete account of the physiological roles for the diverse CNS cholinergic projections. The PPT–LDT ascending cholinergic pathways, along with forebrain neurons, appear to play an important role in sleep control mechanisms, whereas there is scant information regarding the possible functions of the brain stem–spinal cord descending cholinergic projections. The latter projections are suspected to participate in pain-control mechanisms.

The Functions and Neurobiology of Basal Forebrain Cholinergic Neurons

For several decades, the CNS cholinergic system was suspected to play an important role in memory. The most influential early studies pointing out such CNS cholinergic functions were carried out by Drachman and collaborators in young human subjects in which the muscarinic antagonist scopolamine, in low doses, brought about memory deficits similar to those observed in aged individuals. On the other hand, the application of the muscarinic agonists in young

volunteers improved the recall of learned verbal material. These observations on the participation of the cholinergic system in memory tasks were corroborated in subhuman primates by Bartus and collaborators, who elaborated "the cholinergic hypothesis of geriatry memory dysfunction." Numerous further investigations in rodents supported the concept of a cholinergic involvement in memory and learning. These functions, attributed initially to the so-called ascending reticular pathway, elaborated in the 1940s by Moruzzi and Magoun, had an important impact on developing physiological–anatomical theories for higher CNS functions. This anatomically ill-defined system was described as responsible for the control of wakefulness, sleep patterns, alertness, and memory. The anatomical–neurochemical identification in the brain stem and basal forebrain of neuronal cell groups and ascending pathways of catecholaminergic, serotoninergic, and cholinergic nature in the 1970s and 1980s allowed a more precise and specific understanding of the contribution of these transmitter systems to these functions. Today, the basal forebrain cholinergic system can be considered to be the rostralmost representation of such a system. There is solid experimental evidence that the basal forebrain cholinergic neurons do participate importantly in attentional, learning, and memory functions. These investigations constitute a very rich assembly of pharmacological, anatomical, and selective lesion studies combined with a wide variety of behavioral tasks. In more recent years, some emphasis has been placed on the attentional functions of cortical cholinergic inputs, thus conditioning learning and memory functions. Nevertheless, a dramatic argument in favor of the idea that ACh is necessary for learning and memory has been provided by the restoration of these functions with the cortical grafting of genetically modified cells producing ACh in rats with learning and memory impairments as a consequence of nucleus basalis lesions. The forebrain cholinergic input to cortex and hippocampus might have important modulatory roles in defining activity-dependent synaptic maps. The most provoking evidence in that direction was provided by Kilgard and Merzenich in 1998. This study demonstrated that episodic electrical stimulation of the nucleus basalis, paired with an auditory stimulus, resulted in a progressive reorganization of the primary auditory cortex in the adult rat, similar to that with behavioral training. While these findings suggest that the basal forebrain plays a role in modifying cortical synaptic sensory representation, additional pharmacological data would be necessary to firmly relate this to a cholinergic function.

Trophic Factor Dependency of Forebrain Cholinergic Neurons

The embryonic development of the basal forebrain cholinergic system is highly dependent on the expression of nerve growth factor (NGF) and the high-affinity NGF receptor TrkA and, to a lesser extent, on the expression of low-affinity $p75^{LNTR}$ receptors. The low- and high-affinity $p75^{LNTR}$ neurotrophin receptors appear to function, depending on the physiological and/or pathological circumstances, in a cooperative or competitive manner. In early postnatal stages, relatively high levels of NGF are expressed, but this expression substantially decreases after birth. In early ontogenic stages, forebrain cholinergic neurons remain highly sensitive to exogenously administered NGF. The expression and release of NGF are very low after neuronal differentiation in adulthood. However, the NGF dependency of cholinergic neurons in mature and fully differentiated CNS remains. The experimental evidence indicates that in the adult CNS, NGF maintains cholinergic neuronal phenotype. The main supply of NGF is thought to be 'target derived,' that is, from the cerebral cortex for cholinergic nucleus basalis neurons, and from the hippocampus for medial septum neurons. Thus, target ablation of the neocortex or hippocampus provokes cholinergic neuronal cell shrinkage but not death. However, transection of the fimbria-fornix can provoke a mixed situation of neuronal atrophy and cell death of cholinergic neurons in the medial septum. The application of exogenous NGF can restore the cholinergic phenotype in both lesion models. Interestingly, NGF application after limited cortical strokelike lesions is sufficient to restore the size and biochemistry of nucleus basalis cholinergic neurons as well as preventing behavioral memory deficits, even when applying NGF after lesions (see **Figure 5**).

Exogenous application of NGF in cortically lesioned adult rats also induces cholinergic synaptogenesis in the remaining, intact neocortex. On the other hand, the blocking of endogenously produced NGF by applying NGF-immunoneutralizing monoclonal antibodies in the cerebral cortex or synthetic TrkA receptor antagonists was shown to remove preexisting cortical cholinergic synapses *in vivo*. In other words, these observations would indicate that the production of small, baseline endogenous NGF in the cerebral cortex regulates the steady-state number of cholinergic synaptic boutons in the adult CNS. As trophic factors have been shown to be produced and liberated in an activity-dependent fashion, it would be reasonable to assume that the number of cortical

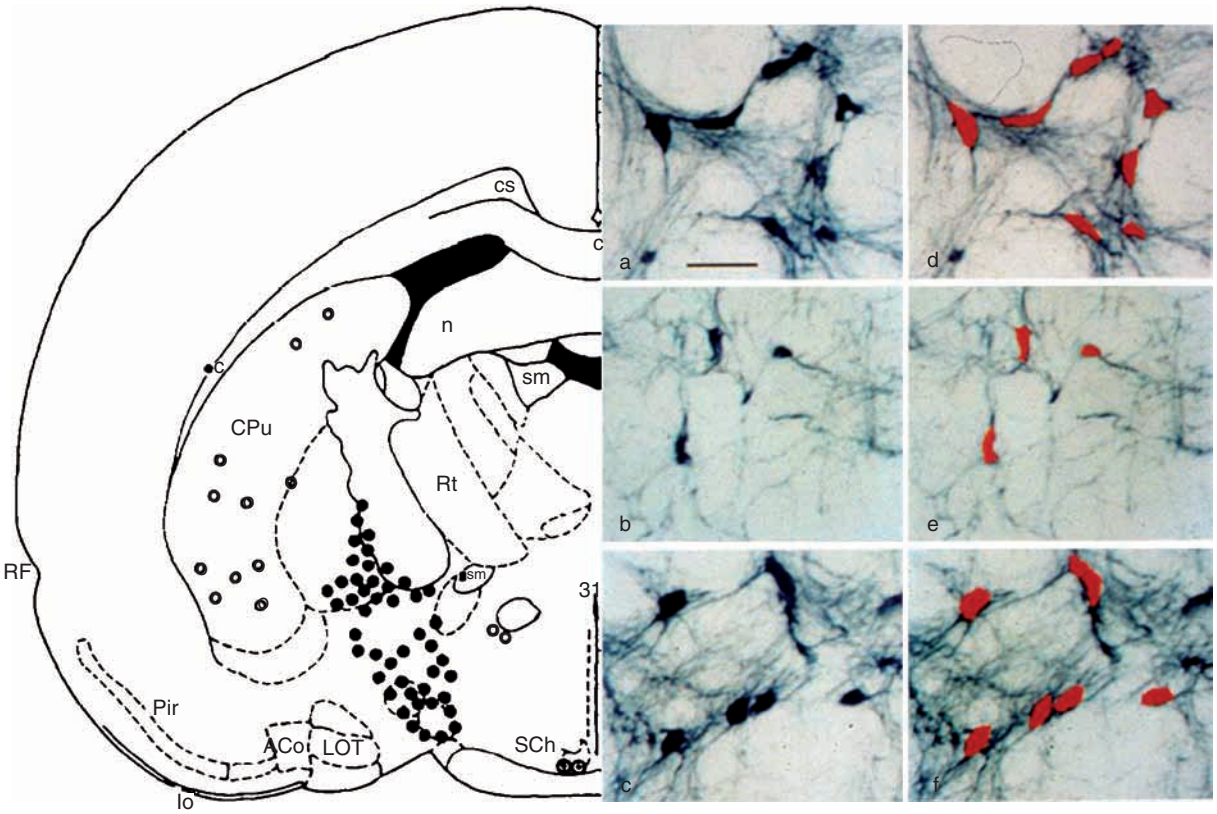

Figure 5 The schematic drawing on the left represents the distribution of abundant forebrain neurons intensively immunoreactive to p75 low-affinity NGF receptor, represented as filled circles. The cells in the inner portion of the globus pallidus correspond to cholinergic neurons of the nucleus basalis. Open circles indicate moderately p75 immunoreactive large neurons in the caudate putamen (cholinergic interneurons). The micrographs on the right (a) indicate cholinergic neurons of the nucleus basalis, as revealed by ChAT immunocytochemistry. Note the rich dendritic branching of the cholinergic neurons of the nucleus basalis. The extreme right column illustrates the computer-assisted definition of the cell soma cross-sectional area. The micrographs illustrate the efficacy of NGF to rescue retrogradely shrunken neurons following cortical lesions: (a) and (d) show ChAT-IR nucleus basalis neurons from naive control adult rats; (b) and (e) show gross cell shrinkage of the same neurons after cortical lesions; and (c) and (f) show prevention of cell shrinkage and maintenance of normal phenotype with the administration of microgram amounts of NGF. ACo, anterior cortical olfactory nucleus; c, corpus callosum; CPu, caudate putamen; cs, cingulate striae; lo, lateral olfactory tract; LOT, nucleus of the olfactory tract; n, fimbria fornix; Pir, piriform cortex; RF, rhinal fissure; Rt, reticular nucleus of the thalamus; SCh, nucleus suprachiasmaticus; sm, stria medullaris; 3, third ventricle. Reproduced from Cuello AC (2006) Cholinergic synaptic terminations in the cerebral cortex, trophic factor dependency, and vulnerability to aging and Alzheimer's pathology. In: Giacobini E and Penney JB (eds.) *The Brain Cholinergic System in Health and Disease*, figure 3.2, pp. 33–46. Oxon, UK: Informa Healthcare, with permission from Taylor & Francis.

cholinergic synapses changes constantly in the human brain, based on experience and brain activity. Such a concept would be in line with the classical tenet of Hebb that the strength of synaptic connections was linked to a growth process which takes place with synaptic efficacy.

CNS Cholinergic Neurons and AD

It is possible that CNS cholinergic neurons are somehow involved in a variety of neurodegenerative disorders. However, up to the present, none of these conditions appears as striking as in the case of Alzheimer's disease. In the 1970s, Davis and Maloney and Bowen et al. reported evidence of marked deficits in cortical cholinergic markers (particularly loss of ChAT levels) in AD brain. After the identification of the main CNS cholinergic groups, Whitehouse and collaborators reported losses of 'magnocellular' neurons in the the nucleus basalis of Meynert. These findings were coincidental with "the cholinergic hypothesis of geriatry memory dysfunction" (as discussed previously). At the time, evidence that the loss of substantia nigra dopaminergic neurons was mainly responsible for Parkinsonian symptoms, along with the evidence that the administration of dopamine

precursors could be an effective transmitter-based therapy, brought about by analogy 'the cholinergic hypothesis of Alzheimer's disease.' This proposal provoked a great deal of experimental and clinical research. We now know that the cholinergic involvement in AD is a component of the neuropathology, most likely secondary to the Aβ burden in the cerebral cortex and hippocampus. It is now also well established that the cholinergic depletion occurs at advanced stages of the AD pathology, while at the AD-prodromic stage of mild cognitive impairment (MCI) an upregulation of cholinergic markers takes place instead. Furthermore, transgenic animal models of the amyloid pathology can replicate features of the cholinergic involvement, suggesting a secondary rather than a primary involvement in the disease process. However, as discussed previously, the forebrain cholinergic system is so intimately related to memory processes that it was perceived as a viable therapeutic target. Early attempts at using muscarinic agonists and choline supplements – in the latter case to stimulate ACh synthesis, given (as discussed previously) that the high-affinity choline uptake system is the *de facto* rate-limiting factor – were without substantive effect on memory or other cognitive parameters. Instead, the application of cholinesterase inhibitors (capable of blocking acetylcholinesterases or butyrylcholinesterases in diverse degrees) to block ACh degradation has been used widely as symptomatic therapy in AD. This approach is by no means capable of halting the pathological process (disease modifying), but has proved effective in delaying cognitive decay in a good number of patients, although the efficacy wanes after a year or two of treatment. The cholinergic system is, however, involved in the disease process. The activation of M1 and M3 receptors has been shown to shift amyloid precursor protein (APP) metabolism to a nonamyloidogenic modality, an aspect that has been confirmed repeatedly in experimental models and in the clinic with the application of anticholinesterases. These observations have provided renewed interest in the development of newer, safer, and more efficaceous muscarinic agonists. Knowing the great trophic dependency of forebrain cholinergic neurons on NGF support, clinical attempts have involved applying mouse NGF in the brain of Alzheimer's patients, but with disapponting results. In these patients, the large doses applied and the diffusion of NGF to undesirable targets (e.g., sensory nociceptive fibers and hypothalamic nuclei) produced pain and marked weight loss. As NGF can act on the nucleus basalis also in the somatodendritic region in a paracrine fashion (in contrast with target derived), there is current interest in applying grafts of genetically modified NGF-producing cells, implanted in nucleus basalis neighboring areas of Alzheimer's patients, and, long range, of regulable lentoviral vectors, allowing the *in situ* expression of NGF. New therapeutic approaches to correct the cholinergic deficit might emerge from the realization that this is most probably due to the dysregulation of the protease cascade, which is responsible for the maturation of the NGF-precursor protein (ProNGF) into mature and biologically active NGF and for its degradation.

See also: Acetylcholine Neurotransmission in CNS; Acetylcholinesterase Inhibitors and Alzheimer's Disease; Acetylcholinesterase; Alzheimer's Disease: An Overview; Basal Forebrain and Memory; Basal Ganglia: Acetylcholine Interactions and Behavior; Cholinergic System; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Neuromuscular Junction (NMJ): Acetylcholinesterases; Nicotinic Acetylcholine Receptors; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Bartus RT, Dean RL, Beer B, et al. (1982) The cholinergic hypothesis of geriatric memory dysfunction. *Science* 217: 408–417.

Butcher LL and Woolf NJ (2004) Cholinergic neurons and networks revisited. In: Paxinos G (ed.) *The Rat Nervous System*, pp. 1257–1268. London: Elsevier.

Cuello AC (1996) Effects of trophic factors on the CNS cholinergic phenotype. *Progress in Brain Research* 109: 347–358.

Cuello AC (2006) Cholinergic synaptic terminations in the cerebral cortex, trophic factor dependency, and vulnerability to aging and Alzheimer's pathology. In: Giacobini E and Penney JB (eds.) *The Brain Cholinergic System in Health and Disease*, pp. 33–46. Oxon, UK: Informa Healthcare.

Cuello AC and Sofroniew MV (1984) The anatomy of the CNS cholinergic neurons. *Trends in Neurosciences* 7: 74–78.

Debeir T, Saragovi HU, and Cuello AC (1999) A nerve growth factor mimetic TrkA antagonist causes withdrawal of cortical cholinergic boutons in the adult rat. *Proceedings of the National Academy of Sciences of the United States of America* 96(7): 4067–4072.

Kilgard MP and Merzenich MM (1998) Cortical map reorganization enabled by nucleus basalis activity. *Science* 279(5357): 1714–1718.

Mesulam MM (1990) Human brain cholinergic pathways. *Progress in Brain Research* 84: 231–241.

Mesulam MM, Mufson EJ, Wainer BH, et al. (1983) Central cholinergic pathways in the rat: An overview based on an alternative nomenclature (Ch1–Ch6). *Neuroscience* 10(4): 1185–1201.

Ribeiro FM, Black SA, Prado VF, et al. (2006) The 'ins' and 'outs' of the high-affinity choline transporter CHT1. *Journal of Neurochemistry* 97(1): 1–12.

Sofroniew MV, Eckenstein F, Thoenen H, et al. (1982) Topography of choline acetyltransferase containing neurons in the forebrain of the rat. *Neuroscience Letters* 33: 7–12.

Sofroniew MV, Galletly NP, Isacson O, et al. (1990) Survival of adult basal forebrain cholinergic neurons after loss of target neurons. *Science* 247: 338–342.

Turrini P, Casu MA, Wong TP, et al. (2001) Cholinergic nerve terminals establish classical synapses in the rat cerebral cortex: Synaptic pattern and age-related atrophy. *Neuroscience* 105(2): 277–285.

Whitehouse PJ, Price DL, Struble RG, et al. (1982) Alzheimer's disease and senile dementia: Loss of neurons in the basal forebrain. *Science* 215: 1237–1239.

Winkler J, Suhr ST, Gage FH, et al. (1995) Essential role of neocortical acetylcholine in spatial memory. *Nature* 375(6531): 484–487.

Cholinergic System

E J Mufson and S E Counts, Rush University Medical Center, Chicago, IL, USA
S D Ginsberg, Nathan Kline Institute, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

For over a century the neurophysiological activity of acetylcholine (ACh) has been well documented. Although Otto Lowei, in the 1920s, identified ACh as the cardioactive substance released by the vagus nerve, it took many additional years of research to establish that ACh was also a neurotransmitter of the central nervous system (CNS). ACh is synthesized in the neuronal cytoplasm from acetyl coenzyme A and choline by the enzyme choline acetyltransferase (ChAT) (see Figure 1). Most ACh is stored in vesicles located in axon terminals and is released to the synaptic cleft upon depolarization in a process mediated by an elevation in intracellular Ca^{2+}. Following release, ACh is hydrolyzed into choline and acetate by acetylcholinesterase (AChE), which is found on both the pre- and postsynaptic membranes. Two classes of receptors are involved in cholinergic transmission: nicotinic ion channels, which mediate fast postsynaptic transmission, and the muscarinic G-protein-coupled receptors, which play a modulatory role. The nicotinic ACh ionotropic receptor (nAChR) is a pentameric membrane protein composed of four polypeptide chains, designated nAChRs α, β, δ, and γ, which are located throughout the CNS (Figure 2). The muscarinic ACh receptor (mAChR) family includes five members, M_1, M_2, M_3, M_4, and M_5, which also exhibit a subtype-specific topographic distribution within the CNS (Figure 3). The basic differences between these two receptor classes are listed in Table 1.

In the CNS, ACh facilitates many functions, including learning, memory, attention, and motor control, mainly through an extensive network of cholinergic projection pathways originating in the basal forebrain. Dysfunction of these cholinergic pathways plays a role in age-related memory impairments and in the onset of Alzheimer's disease (AD). The following discussions survey the central cholinergic projection neurons and their relation to neurotransmission, and the observations from human brain suggesting that alterations in neurotrophic support of these neurons play a central role in their dysfunction in aging and AD.

Anatomy of the Cholinergic Basal Forebrain

A diverse group of telencephalic structures situated on the medial and ventral aspects of the cerebral hemispheres compose the basal forebrain in the primate brain. Brain tissue stained for cell bodies reveals clusters of neurons with a distinct magnocellular appearance (Figure 4(a)).

In the early 1980s, the Ch1–Ch4 nomenclature was introduced to designate the cholinergic (ChAT-containing) neurons within the basal forebrain. Studies using the nonspecific cholinergic marker AChE and specific markers – ChAT (Figure 4(i)), the vesicular acetylcholine transporter (VAChT), and the high-affinity choline transporter – delimited a continuum of cholinergic neurons within the basal forebrain (Figures 5(a)–5(d)). Cholinergic neurons (25–30 × 25–30 μm) of the medial septum (Ch1) comprise 10% of this region. The diagonal band of Broca comprises the Ch2 (vertical limb) and Ch3 (horizontal limb) subgroups. Ch2 neurons are fusiform in shape, 20–25 × 30–40 μm in size, and approximately 70% of these perikarya are cholinergic. Ch3 neurons have the most diffuse boundaries in the primate brain and are fusiform-shaped (15–20 × 40–50 μm) neurons of which approximately 1–2% are cholinergic. The largest group of cholinergic neurons (90%) constitutes the Ch4 cells of the nucleus basalis (NB), which are fusiform to multipolar in shape and 40–50 × 60–70 μm in size. Topographically, the Ch4 group can be separated into anterior (Ch4a), intermediate (Ch4i), and posterior (Ch4p) subfields. The total number of cholinergic neurons located within the primate NB is approximately 200 000 per hemisphere. The term 'nucleus basalis' may be employed to designate the cholinergic components (Figures 4(b)–4(d) and 4(f)) as well as the noncholinergic components – for example, calbindin (CB; Figures 4(b) and 4(c)), glutamatergic (Figure 4(d)), γ-aminobutyric acid (GABA-)ergic, and galaninergic profiles (Figures 4(g) and 4(h)). The more restrictive Ch4 nomenclature is reserved for the cholinergic neurons located in this region. These cholinergic basal forebrain (CBF) neurons provide the major cholinergic innervation to the hippocampus (Ch1–Ch2; Figure 6), olfactory system (Ch3), amygdala, and cortical mantle (Ch4; Figure 6).

Interestingly, galanin, an inhibitory neuropeptide, co-localizes within the cholinergic neurons of the NB in rodents and lesser apes (rhesus monkeys, gibbons). By contrast, in the great apes (chimpanzees, gorillas, and humans), cholinergic NB neurons are galanin

negative (Figure 7). These cholinergic neurons are innervated by a separate population of galaninergic interneurons (Figure 4(g)) and fibers from extra forebrain sources (e.g., locus coeruleus). These findings suggest that during the evolution of the human brain the galaninergic systems diverged from rodents and lesser apes. Perhaps galanin plays a unique role in the regulation of CBF neuronal tone required for higher cortical function. The unique role of galanin in the human brain is further underscored by the observation that in the late stages of AD, galanin fibers hypertrophy upon surviving cholinergic neurons (Figure 4(h)). It has been suggested that galanin plays a crucial role in cholinergic cell survival and that galanin overexpression is a compensatory mechanism in an attempt to attenuate the progression of cholinergic cell death in AD. Available data also indicate a species difference in the localization of estrogen receptor alpha (ERα)-containing nuclei within the CBF between rodents and some primates. Numerous histochemical and gene expression studies in the rodent indicate that the CBF neurons co-express estrogen receptors. In contrast to the rodent, immunohistochemical studies indicate that ERα does not co-localize with ChAT-containing CBF neurons in the monkey. However, ERα-containing nuclei as well as minor cytoplasmic staining were found within other forebrain areas, including the lateral septum, amygdala, the bed nucleus of the stria terminalis, the subfornical organ, medial and lateral hypothalamic nuclei, and the amygdalo-hippocampal transition

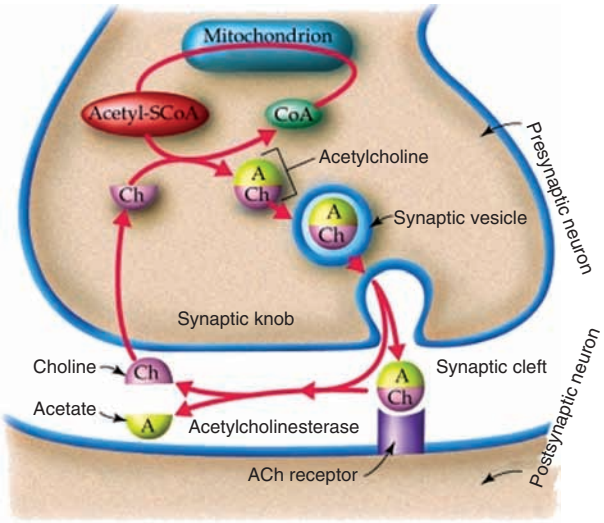

Figure 1 Schematic diagram of a cholinergic synapse illustrating the pre- and postsynaptic molecular entities involved in the synthesis, storage, release, reuptake, and signaling of acetylcholine (ACh). Reproduced from http://abdellab.sunderland.ac.uk/.

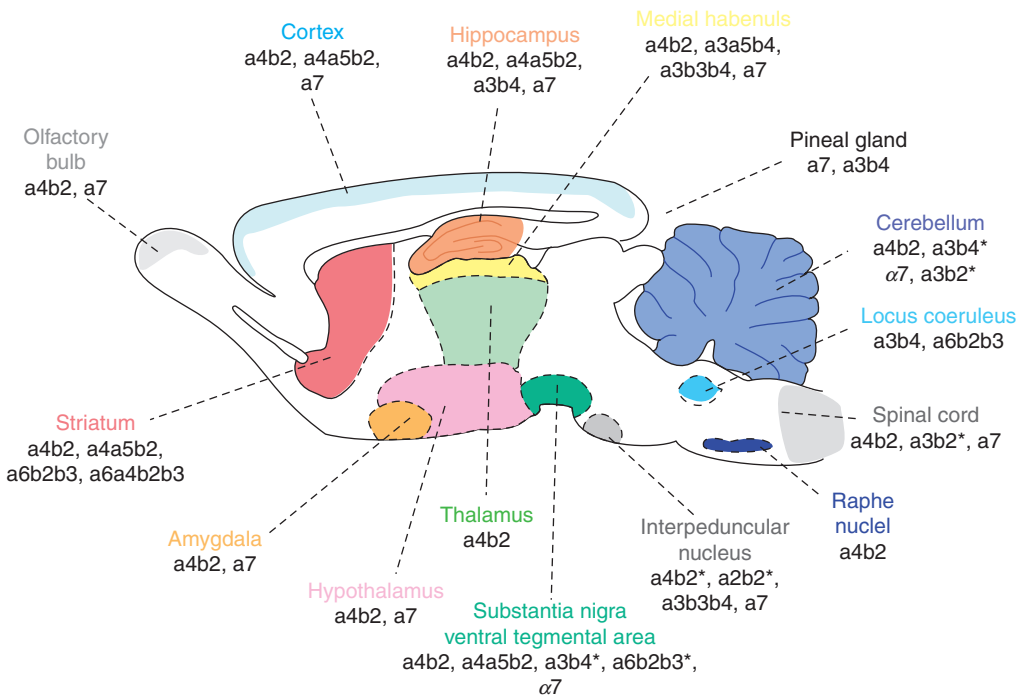

Figure 2 Schematic diagram shows the regional distribution of the main nicotinic acetylcholine receptor subtypes within the rodent central nervous system. The subtypes present in the cortex, cerebellum, hippocampus, interpeduncular nucleus, medial habenular, and pineal gland were identified by binding, immunoprecipitation, and/or immunopurification assays in tissue from rats and/or wild-type and/or receptor subunit knockout mice. Asterisk indicates possible presence of unidentified subunits. Adapted from Gotti C, Zoli M, and Clementi F (2006) Brain nicotinic acetylcholine receptors: Native subtypes and their relevance. *Trends in Pharmacological Sciences* 27: 482–491, with permission from Elsevier.

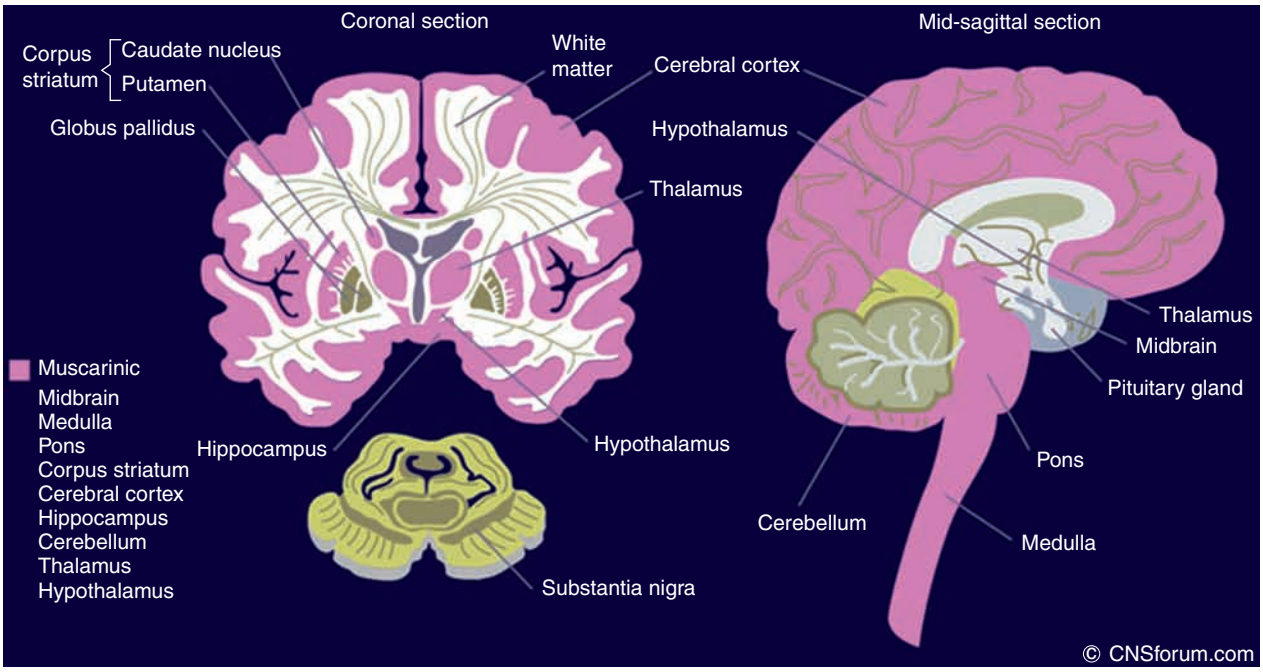

Figure 3 Schematic diagrams, coronal and sagittal, show the regional distribution of the muscarinic receptors in the primate central nervous system. Reproduced from http://www.CNSforum.com.

Table 1 Characteristics of muscarinic and nicotinic receptors

Characteristic	*Muscarinic*	*Nicotinic*
Response time to respective agonist (muscarine, nicotine)	Slow (100–250 ms)	Fast (<10 ms)
Receptor type	Metabotropic	Ionotropic
Receptor subtype	M_1, M_3, $M_5 = G_q$-coupled M_2, $M_4 = G_i$ coupled	Nine subtypes
Excitatory/inhibitory	Mixed	Excitatory
Typical antagonists	Atropine, scopolamine	Tubocurarine
Distribution	Parasympathetically innervated cardiac and smooth muscles (heart, iris, stomach, bronchitis bladder, etc.) Salivary, tear, and sweat glands CNS	All striated muscles Postganglionic parasympathetic and sympathetic neurons CNS

zone in the rhesus monkey. Studies performed on human brain tissue revealed ERα- and ERβ-positive neurons (Figure 4(e)) within the NB. However, these reports did not co-localize these receptors with any CBF neuronal marker. Further studies are needed to determine the chemical phenotype of ER-containing neurons within the primate brain and whether cholinergic neurons are regulated via genomic or epigenetic mechanisms.

Anatomy of the Thalamic and Brain Stem Cholinergic Subgroups

Within the upper brain stem a continuous band of ChAT-containing neurons extends from the rostral midbrain to the midpontine levels (Figure 8). These perikarya can be subdivided into two major fields, Ch5 and Ch6. The cholinergic neurons of Ch5 remain mostly within the pedunculopontine nucleus (PPN); this location is analogous to that of the Ch5 group identified in rodents. At the level of the PPN this sector exhibits a compact lateral component which abuts on the lateral lemniscus and a more diffuse medial portion which is interdigitated with the superior cerebellar peduncle, the medial longitudinal fasciculus, and the central tegmental tract (Figure 8). As the Ch5 sector gradually disappears, the more caudal Ch6 subgroup occupies a lateral position in the periventricular gray at a location which corresponds to the lateral dorsal tegmental nucleus. These Ch6

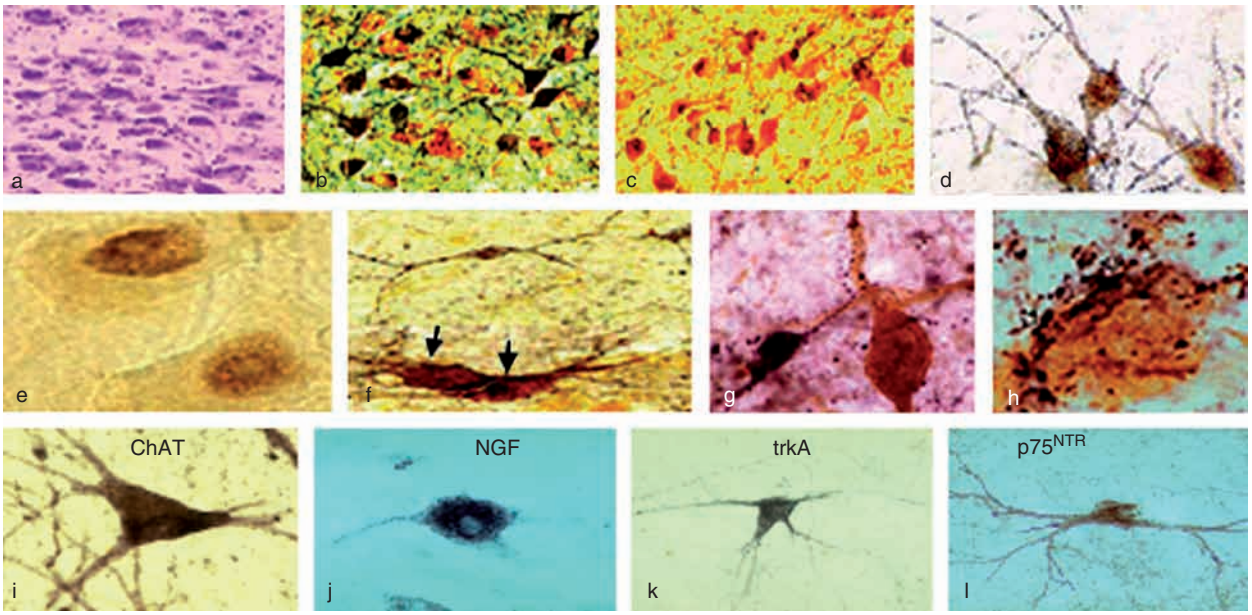

Figure 4 Photomicrographs show the diversity of neurons within the nucleus basalis. (a) Human nucleus basalis Nissl-stained neurons showing the magnocellular phenotype of cholinergic basal forebrain perikarya. (b) Concurrent immunostaining of choline acetyltransferase (ChAT; brown reaction product) and calbindin (granular blue-black reaction product) in the anteromedial Ch4 neurons in a 50-year-old. Note that most Ch4 anteromedial neurons contain ChAT and calbindin. (c) Tissue from a 93-year-old showing preservation of Ch4 ChAT-positive neurons and reduced calbindin staining. (d) Ch4 posterior neurons dual stained for GluR1 and $p75^{NTR}$. (e) Estrogen-beta-immunoreactive nuclei within Ch4 neurons. (f) Section from Ch4 neurons dual stained for muscarinic M2 receptor (black) and $p75^{NTR}$. Note the M2 single and the double (M2/$p75^{NTR}$; arrows) neurons. (g) Ch4 section dual stained for galanin (dark blue cell) and $p75^{NTR}$ (brown neuron) from a human brain. (h) Hyperinnervation of galanin (black) fibers upon a cholinergic (light brown) Ch4 neuron in Alzheimer's disease. (i–l) Nucleus basalis neurons immunoreacted for ChAT, retrogradely transported nerve growth factor (NGF), TrkA, and $p75^{NTR}$, respectively.

neurons provide cholinergic innervation to the thalamus (e.g., parafascicular nucleus). The medial habenula, a component of the epithalamus, contains densely packed oval-shaped cholinergic Ch7 neurons, which innervate the interpeduncular nucleus. A compact cholinergic cell group, Ch8, lies along the lateral margin of the mesopontine region within the parabigeminal nucleus; these neurons innervate the superior colliculus. A summary schematic diagram of the central cholinergic projection neurons (Ch1–Ch8) is shown in Figure 9. Other cholinergic neurons include the interneurons of the striatal complex, cranial nerve nuclei, reticular neurons, lateral hypothalamus, olfactory tubercle, islands of Calleja, and scattered intercortical cells seen only in the rodent.

CBF Neuron Survival

A growing body of data supports a role for the neurotrophic substance nerve growth factor (NGF) in the maintenance and survival of CBF neurons. NGF is derived from its precursor molecule proNGF (Figure 10), which is the primary form of NGF in the human brain. *In vitro* studies using dissociated rat CBF cultures or organotypic slices show that NGF treatment prevents the degeneration of cholinergic neurons observed in untreated preparations. In a similar vein, the atrophy and loss of these neurons *in vivo*, following neocortical infarction in monkeys or excitotoxic neocortical lesions in rats, are reversed with exogenous NGF treatment. In addition, infusion of NGF prevents loss of cholinergic phenotypic markers as well as septal cholinergic degeneration following septohippocampal axotomy via fimbria–fornix transection. Finally, transgenic mice that express anti-NGF antibodies in adulthood display an age-dependent loss of cholinergic forebrain neurons.

CBF Neurons Express NGF Receptors

The biological actions of NGF are mediated in part by two NGF receptors, the NGF-specific receptor tyrosine kinase TrkA and the pan-neurotrophin $p75^{NTR}$ receptor, which are produced in the CBF neurons (Figures 4(j)–4(l)) and anterogradely transported to the innervation sites of these cell bodies. These receptors mediate the binding, internalization, and retrograde transport of target-derived NGF to CBF perikarya. TrkA is a receptor tyrosine kinase that binds NGF with relatively high affinity ($K_d = 10^{-11}$) when co-expressed with $p75^{NTR}$. After binding of

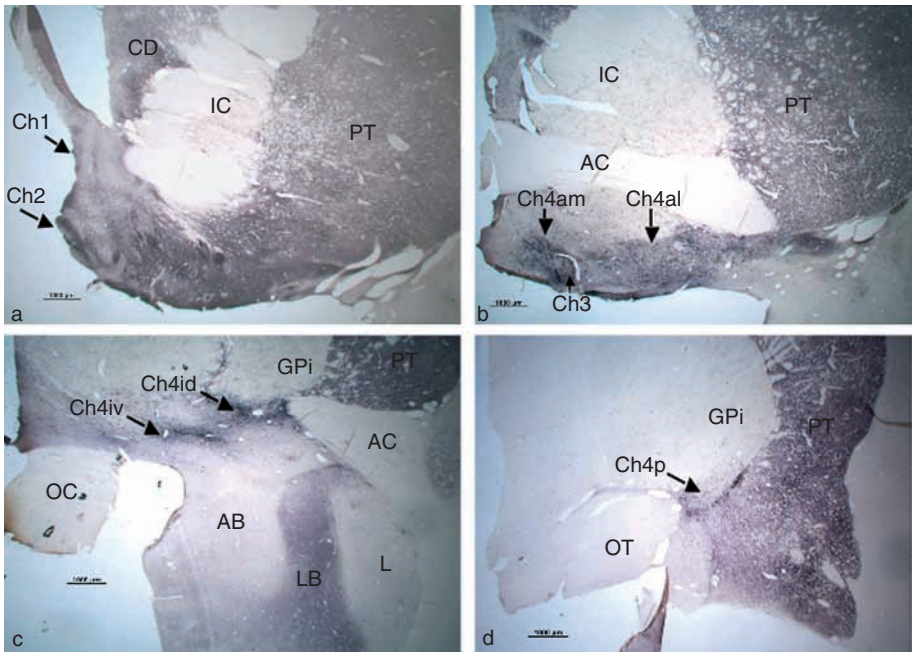

Figure 5 Sections stained for the immunohistochemical demonstration of choline acetyltransferase-positive neurons and fibers within the monkey forebrain. The sections are from progressively more caudal levels of the brain. Medial is to the left, dorsal toward the top. AB, accessory basal nucleus of the amygdala; AC, anterior commissure; CD, caudate; Ch1, cholinergic cells located within the medial septum; Ch2, cholinergic cells located within the vertical limb of the diagonal band (Ch3 cells not shown, but correspond to cholinergic cells located with the horizontal limb of the diagonal band); Ch4am/al/id/iv, anteromedial, anterolateral, intermediate dorsal, and intermediate ventral cholinergic cells; Chp, cholinergic neurons (posterior) in the major subdivision of the nucleus basalis; IC, internal capsule; GPi, internal segment of the globus pallidus; L, lateral nucleus of the amygdala; LB, lateral basal nucleus of the amygdala; OC, optic chiasma; OT, optic tract; PT, putamen.

NGF to TrkA, the receptor activates signal transduction cascades, including mitogen-activated protein kinases (MAPKs), phosphatidylinositol 3-kinase (PI3K), and phospholipase C-γ (PLC-γ) pathways that promote survival and sustain the cholinergic phenotype of CBF neurons by activation of transcription factors that alter gene expression. $p75^{NTR}$ is a member of the tumor necrosis factor receptor family and binds NGF with relatively low affinity ($K_d = 10^{-9}$). Although this pan-neurotrophin receptor lacks intrinsic catalytic activity, it interacts with cytosolic adapter proteins to activate downstream signaling molecules, including c-Jun, N-terminal Jun kinase (JNK), and nuclear factor-κB (NF-κB). Activation of these pathways through the $p75^{NTR}$ receptor *in vitro* elicits either a pro- or antiapoptotic signal that likely depends on the cellular milieu (e.g., cell type, presence or absence of specific binding partners and/or chaperones). The binding partners of $p75^{NTR}$ include sortilin, Nogo receptor, and lingo-1, in addition to TrkA.

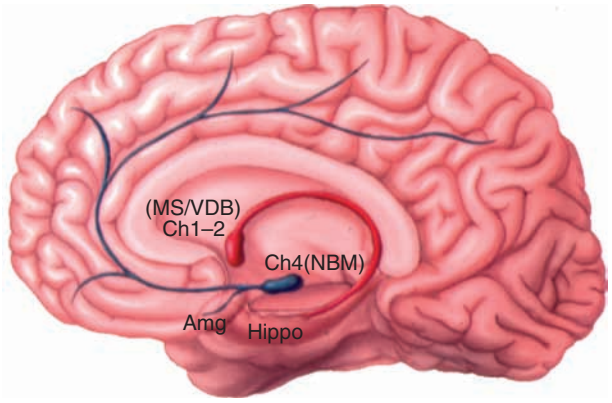

Figure 6 Artist's drawing showing the cholinergic innervation from the medial septal/vertical limb of the diagonal band (MS/VDB; Ch1–2) to the hippocampus (Hippo; red) and the nucleus basalis (Ch4) to the cortex and amygdala (Amg; blue) in the human brain.

NGF Signaling in CBF Neurons

It is well established that NGF is retrogradely transported from its production sites in the cortex and hippocampus to the CBF consumer neurons. The mechanism by which NGF retrograde transport proceeds and culminates in cell survival remains to be answered, but a series of observations indicate that NGF-bound TrkA complexes are internalized as specialized 'signaling endosomes' by a process involving dynamin and PI3K. These signaling endosomes

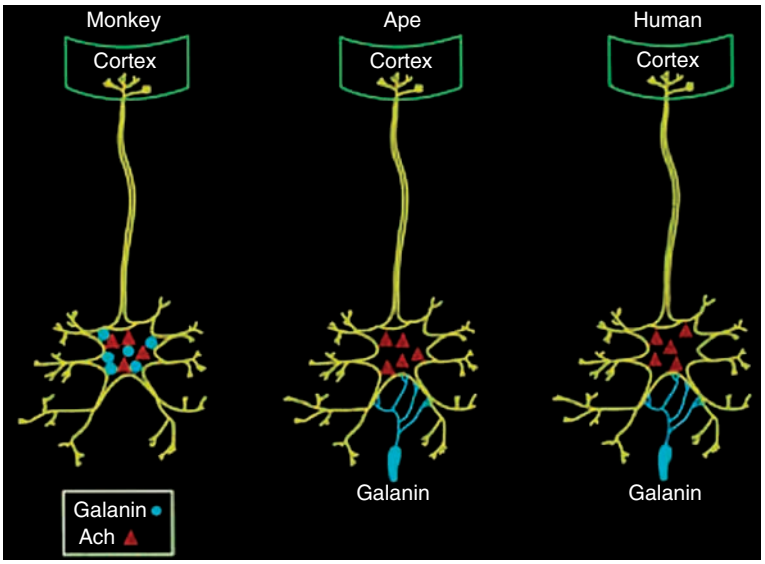

Figure 7 Schematic diagram depicting the relation between acetylcholine (ACh) and galanin within neurons of the primate cholinergic basal forebrain. In the monkey, note that all cholinergic basal forebrain neurons co-localize galanin. By contrast, these neurons are galanin negative in great apes, including humans. In these species, galanin arises from a population of small interneurons (blue) as well as in an extra-forebrain fiber system (not shown), which terminates upon the large cholinergic neurons of the nucleus basalis.

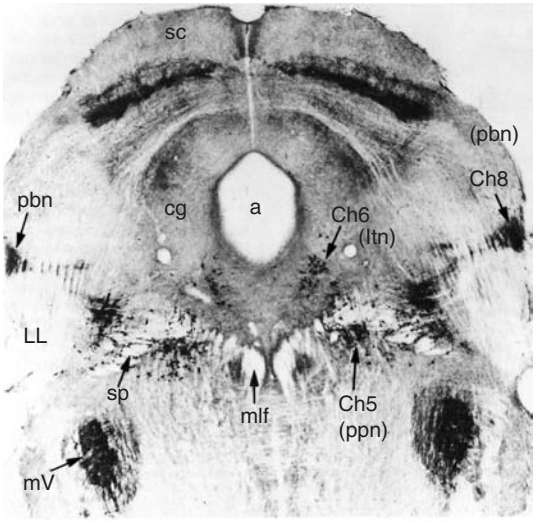

Figure 8 Section stained for the immunohistochemical demonstration of choline acetyltransferase in the rat brain stem. a, cerebral aqueduct; cg, central gray matter; Ch5, cholinergic cell group located within the pedunculopontine nucleus (ppn); Ch6, cholinergic cell group located within the lateral tegmental nucleus (ltn); Ch8, cholinergic cell group located within the parabigeminal nucleus (pbn); LL, lateral lemniscus; mlf, medial longitudinal fasciculus; mV, motor nucleus of fifth cranial nerve; sc, superior colliculus.

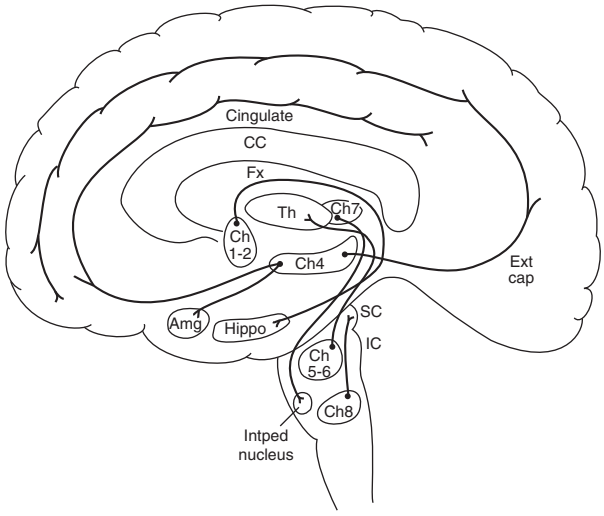

Figure 9 Summary diagram illustrates Ch1–Ch8 neurons and their projection sites. Amg, amygdala; CC, corpus callosum; Ext cap, external capsule; Fx, fornix; Hippo, hippocampus; IC, internal capsule; Intped nucleus, interpeduncular nucleus; Th, thalamus; SC, superior colliculus.

are transported in the minus direction along microtubules, as dynein inhibition or microtubule disassembly blocks the retrograde transport of both NGF and Trk receptors. NGF application to distal axons in compartmentalized sympathetic neurons induces cyclic AMP response element-binding protein (CREB) phosphorylation in an endocytosis-dependent manner. CREB-mediated gene expression is both necessary and sufficient for NGF-dependent survival, indicating that the retrograde transport of signaling endosomes is required for NGF-induced nuclear responses important for cholinergic cell survival. However, there is also evidence that NGF retrograde

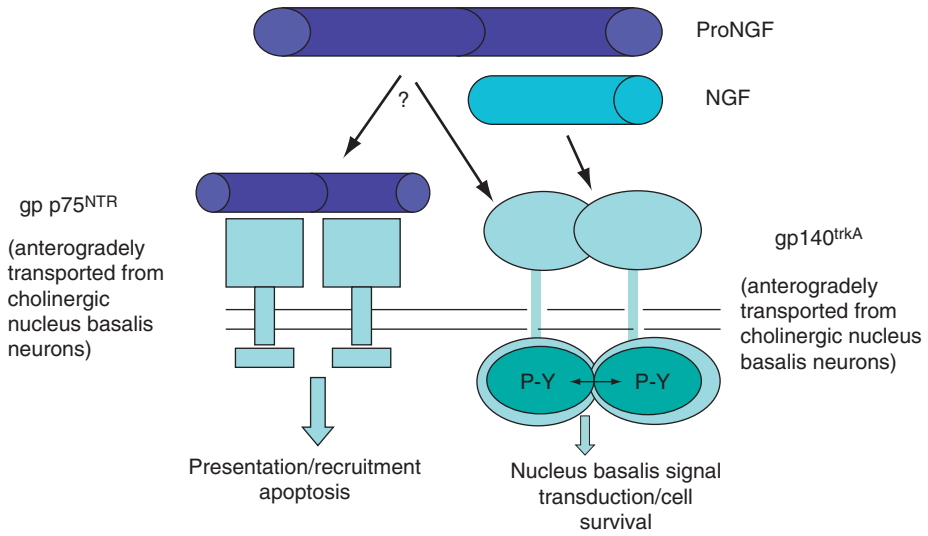

Figure 10 Schematic drawing indicating the interaction between nerve growth factor precursor protein (proNGF), NGF, and its high-affinity (TrkA) and low-affinity ($p75^{NTR}$) glycoprotein (gp) receptors and the putative functional expression in cholinergic basal forebrain neurons. The question mark indicates the ongoing controversy regarding whether proNGF binds with the TrkA receptor as well as with $p75^{NTR}$. P-Y refers to tyrosine phosphorylation.

signaling can occur in the absence of signaling endosomes, suggesting the existence of alternative and possibly parallel mechanisms of retrograde signaling.

CBF Alterations during the Progression of AD

CBF projection neurons have received extensive attention due to their dysfunction in AD. Unbiased stereology has demonstrated that the number of cholinergic neurons within the NB containing either ChAT or VAChT remains unchanged in patients clinically diagnosed with mild cognitive impairment (MCI; a putative prodromal stage of AD) or mild AD, as compared to age-matched controls. In the stereology studies, cortical ChAT enzyme activity was stable in the early stages of AD or increased in the superior frontal cortex as well as the hippocampus of patients with MCI. These observations suggest that basocortical cholinergic tone is preserved in MCI and early AD, supporting the notion that cholinergic NB neurotransmission dysfunction is a late-stage event in AD. By contrast, the number of NB perikarya containing either $p75^{NTR}$ or TrkA was reduced ~50% in MCI and mild AD, compared to controls. Intriguingly, there was a significant positive correlation between the number of TrkA- and $p75^{NTR}$-immunoreactive NB neurons and performance on clinical tests of cognition. In a similar mode, the number of $p75^{NTR}$-immunoreactive NB neurons was significantly correlated with performance on the mini-mental state exam (MMSE) and the global cognitive test. These chemoanatomical clinical pathologic studies suggest that a subpopulation of cholinergic NB neurons undergoes a phenotypic silencing of NGF receptor protein expression in the absence of frank cell loss or cholinergic deficits during the early stages of AD, which correlates with cognitive impairment. Thus, pronounced early defects in NGF receptor expression may be an early pathological event that precedes the eventual loss of cholinergic neurons seen in the later stages of AD.

Quantitative immunoblotting experiments revealed that TrkA and $p75^{NTR}$ protein levels in five cholinergic cortical projection sites (anterior cingulate, superior frontal, superior temporal, inferior parietal, and visual cortex) are decreased during the progression of AD. Interestingly, whereas $p75^{NTR}$ protein cortical levels were stable across the diagnostic groups, cortical TrkA protein was reduced ~50% in mild AD compared to individuals with no cognitive impairment (NCI) and with MCI. Moreover, cortical TrkA levels were positively correlated with cognitive performance. Because TrkA signaling is associated with neuronal survival, this specific reduction of cortical TrkA receptor protein may play a pivotal role in the frank loss of cholinergic NB neurons seen in later stages of AD.

Mechanisms of NGF Receptor Alterations in Cholinergic Neurons in MCI and AD

TrkA Signaling in the Cholinotrophic System

The mechanisms underlying TrkA protein reduction in cholinergic NB perikarya in MCI and the projection sites in early AD are unclear. TrkA gene expression is under positive feedback from NGF signaling,

and this pathway most likely is disrupted in AD by diminished retrograde transport of cortical NGF to CBF consumer neurons. This hypothesis is supported by the observation that NGF mRNA levels are not altered in AD, whereas NGF protein levels are stable or increased in the cortex and decreased in cholinergic NB neurons in late-stage AD. Defective retrograde transport of NGF within cholinergic projection neurons occurs in a segmental trisomic mouse model of Down syndrome (Ts65Dn), a congenital neurodegenerative disorder that mimics some of the neuropathology of AD. In aged rats, CBF neurons exhibit a pronounced reduction in NGF retrograde transport, reduced TrkA protein expression, and severe cellular atrophy. This defect may ultimately reduce TrkA protein in cortical projection sites and further perturbations in NGF signaling within CBF neurons. This putative 'off Trk' cycle of deficient NGF receptor signaling through prosurvival signal transduction molecules such as MAPK and PI3K may contribute to the selective vulnerability of cholinergic NB neurons and to deficits in cortical cholinergic tone observed in late-stage AD.

$p75^{NTR}$ Signaling in the Cholinotrophic System

In contrast, the stability of cortical $p75^{NTR}$ protein expression levels across clinical diagnoses (NCI → MCI → AD) occurs despite a reduction in cholinergic NB perikarya containing $p75^{NTR}$ in both MCI and mild AD relative to NCI. Moreover, cortical $p75^{NTR}$ levels remain stable even in severe AD despite a dramatic reduction of $p75^{NTR}$-immunopositive neurons and fibers in the NB in late-stage AD. Gene expression profiling studies of single cholinergic NB neurons reveal that $p75^{NTR}$ mRNA levels are unchanged in end-stage AD compared to age-matched controls. Therefore, the stability of cortical $p75^{NTR}$ levels is not due to compensatory upregulation of $p75^{NTR}$ by remaining cholinergic NB neurons. On the other hand, the preservation of cortical $p75^{NTR}$ may be related to the *de novo* expression of cortical $p75^{NTR}$-immunoreactive neurons in AD. Alternatively, these data could reflect a redistribution of $p75^{NTR}$ from the CBF soma to the nerve terminals, and receptor accumulation at the terminals due to failed retrograde transport in AD. The preservation of cortical $p75^{NTR}$ protein despite a reduction of cortical TrkA protein in AD could result in a relative increase in $p75^{NTR}$-mediated NGF signaling in cholinergic NB neurons. Traditionally, $p75^{NTR}$ is thought to facilitate NGF signaling by producing high-affinity NGF binding sites and increased TrkA responsiveness to NGF. This concept is being revisited, since $p75^{NTR}$ possesses complex autonomous signaling capabilities in response to NGF that, paradoxically, regulate both cell death and cell survival under different experimental paradigms. NGF activation of $p75^{NTR}$ in the absence of TrkA induces apoptosis in neurons and in mature oligodendrocytes. Retroviral infection of mature $p75^{NTR}$-containing oligodendrocytes with TrkA cDNA reverses NGF-induced apoptosis, suggesting that the coexpression of TrkA suppresses $p75^{NTR}$-mediated proapoptotic signaling. Recent data indicate that activation of the JNK pathway is a primary carrier of the $p75^{NTR}$ proapoptotic signal. NGF binding to $p75^{NTR}$ results in the phosphorylation and activation of JNK, which translocates to the nucleus to phosphorylate the transcription factor c-Jun. Phosphorylated c-Jun-mediated transcription activates a proapoptotic cascade involving the release of cytochrome *c* and caspase activation.

CBF Neurons Display Cell Cycle Proteins in MCI and AD

Accumulating evidence suggests that $p75^{NTR}$-mediated apoptosis involves the activation of cell cycle regulatory molecules. For example, in postmitotic, $p75^{NTR}$-containing retinal neurons, NGF-mediated apoptosis is preceded by an increase in mitotic figures as well as by an upregulation of cyclin B2. Cyclin B2 is typically active at the G_2/M cell cycle transition and apoptosis is prevented in a dose-dependent manner by the delivery of a cyclin B2-dependent kinase inhibitor (CDK1), roscovitine. Intriguingly, CBF neurons in MCI and early AD are characterized by the abnormal expression of cell cycle regulatory proteins (proliferating cell nuclear antigen, cyclin D1, and cyclin B1) and evidence of DNA replication. Perhaps increases in $p75^{NTR}$-mediated activity in the face of reduced TrkA signaling contribute to cholinergic NB neurodegeneration by promoting unscheduled cell cycle reentry and apoptosis in AD. It is likely that shifts in the ratio of TrkA to $p75^{NTR}$ receptors may determine whether neurons survive or die when exposed to NGF. As such, the ~50% reduction in cortical TrkA in the absence of significant changes in $p75^{NTR}$ levels at the onset of AD may signify a relative increase in proapoptotic $p75^{NTR}$ signaling in the cholinobasal projection system.

ProNGF and Cholinergic NB Neuron Survival

The physiological consequences of TrkA and $p75^{NTR}$ signaling upon cholinergic NB cell survival may also depend upon their interactions with the NGF precursor protein, proNGF. ProNGF is the predominant form of NGF present in the cortex of aged, intact humans and is increased ~40–60% in the inferior

parietal cortex of patients diagnosed with MCI or mild AD, compared to NCI. The biological consequences of proNGF accumulation in the cortex during the prodromal stages of AD have yet to be determined. An emerging literature suggests that recombinant proNGF binds TrkA and promotes neuronal survival and neurite outgrowth, similar to mature NGF, but is approximately fivefold less active than mature NGF. Although TrkA-mediated proNGF retrograde transport has not been reported, proNGF accumulation in CBF target sites may be due to reduced cortical TrkA levels and/or retrograde transport to cholinotrophic basal forebrain perikarya. Alternatively, evidence from the rodent cortex indicates that proNGF accumulation is under negative feedback from mature NGF levels; as mature NGF levels rise, proNGF levels are reduced (Figure 11(c)). Moreover, levels of matrix metalloproteinase-9 (MMP-9), an enzyme that degrades mature NGF, are increased in the AD cortex suggestive of increased mature NGF degradation. The resulting decrease in mature NGF may facilitate the proNGF accumulation seen in the prodromal (MCI) and AD cortex (Figure 11(d)). Reductions of mature NGF and poor utilization of proNGF and/or signaling of proNGF via $p75^{NTR}$, in combination with reduced TrkA, may result in cholinergic NB neurodegeneration. Significantly, while TrkA levels in the cortex are positively associated with cognitive scores, cortical proNGF levels are negatively correlated with mental performance. Thus, the concomitant reduction of TrkA and accumulation of proNGF in the cortex may be an early pathobiological marker for the onset of AD.

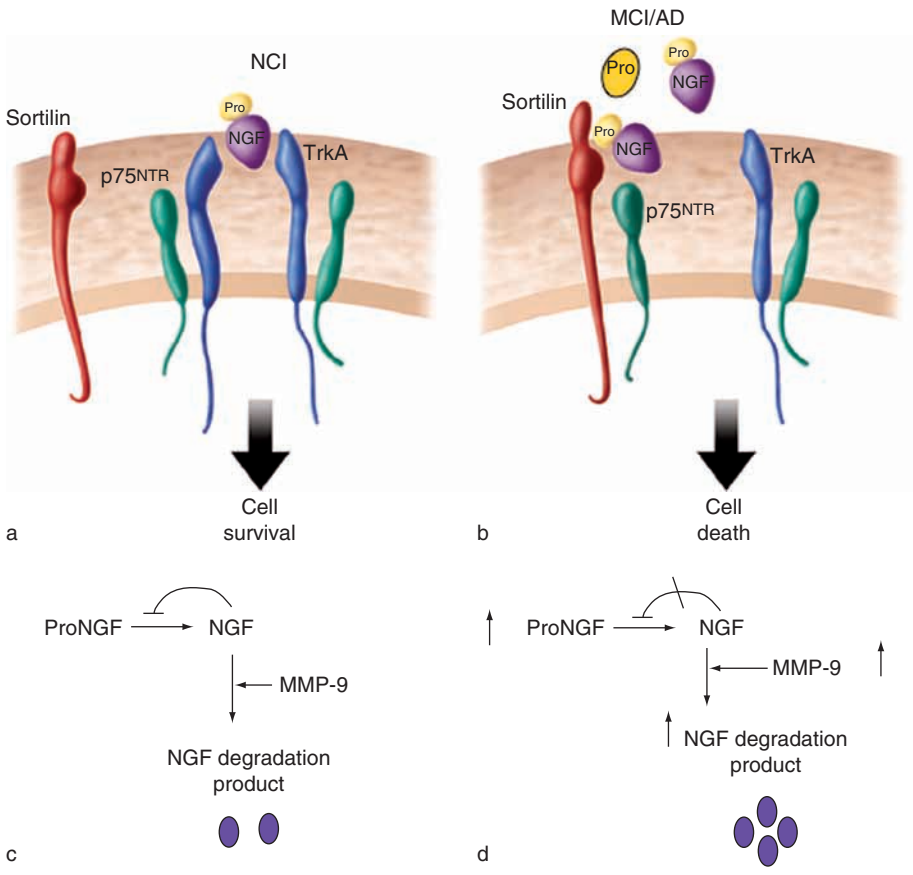

Figure 11 Schematic diagram illustrating putative scenarios for nerve growth factor precursor protein (proNGF)/receptor complexing during the progression of Alzheimer's disease. (a) ProNGF complexes with TrkA to activate cell survival mechanisms in the aged human healthy brain. This binding event is facilitated by the co-expression of $p75^{NTR}$ on the cell surface. (b) Elevated levels of cortical proNGF in the face of reduced cortical TrkA result in increased binding of proNGF to $p75^{NTR}$/sortilin complexes, enhancing the activation of proapoptotic mechanisms during the transition from mild cognitive impairment (MCI) to Alzheimer's disease (AD). (c) Evidence from rodent brain shows that matrix metalloproteinase 9 (MMP-9) degrades mature NGF. Experimental perturbations of MMP-9 activity increase mature NGF levels and reduce proNGF levels, suggesting that proNGF accumulation is under negative feedback from mature NGF levels. (d) MMP-9 levels are increased in the AD cortex, resulting potentially in reduced mature NGF levels, reduced negative feedback on proNGF levels, and proNGF accumulation.

CBF Neurons: Sorting Life from Death

Paradoxically, several investigations indicate that increases in cortical proNGF result in proapoptotic signaling through binding to the $p75^{NTR}$ and other receptor molecules. Data indicate that the proapoptotic effects of $p75^{NTR}$-mediated proNGF signaling are dependent upon interactions between $p75^{NTR}$ and the neurotensin receptor, sortilin. Sortilin is a Vps10p domain trafficking protein that acts as a cell surface coreceptor with $p75^{NTR}$ to mediate proNGF-induced cell death, and blocking this binding event precludes high-affinity binding of proNGF to $p75^{NTR}$. Sortilin binds the 'pro' region of proNGF, and blocking this binding event precludes high-affinity binding of proNGF to $p75^{NTR}$. Moreover, sortilin expression is required for $p75^{NTR}$-mediated apoptosis following proNGF treatments, suggesting that sortilin is a $p75^{NTR}$ binding chaperone associated with its role in the induction of cell death. It is likely that the outcome of $p75^{NTR}$ signaling in response to proneurotrophin depends on the identity of the bound coreceptor. $p75^{NTR}$ may have other receptor chaperone molecules and binding partners, which assist in its role as a 'multitasking' neurotrophin receptor. In this case, if $p75^{NTR}$ and the coreceptor NgR/lingo 1 interact, then an inhibition of axonal growth occurs, as opposed to cell death if sortilin acts as a coreceptor and/or chaperone. It remains to be determined whether sortilin/$p75^{NTR}$/proNGF complexes with motor proteins to form a retrogradely transported ensemble on microtubules, initiating an intracellular cell death response, or whether it is internalized into the somotodendritic compartment of consumer neurons and is anterogradely transported for recycling to the cell surface. Figure 11 diagrams a potential scenario for prosurvival or proapoptotic signaling in cholinergic NB neurons as TrkA levels decline, $p75^{NTR}$ levels remain unchanged, and proNGF levels rise in cortical projection sites during the progression of AD.

Genetic Signature of CBF Neurons

Currently, the molecular and cellular mechanisms underlying CBF dysfunction remain unclear. The advent of single-cell aspiration technology combined with RNA amplification and custom-designed cDNA array technology has provided novel information about the expression levels of several functional classes of mRNAs found in cholinergic NB neurons from normal aged and AD cases. These studies have shown a downregulation of synaptic-related markers (e.g., synaptophysin and synaptotagmin 1) and protein phosphatase 1 and protein phosphatase 2 subunit mRNAs, along with upregulation of cathepsin D acid hydrolase mRNA in AD patients, as compared to normal control cholinergic NB neurons.

Of particular relevance to cholinergic NB neurons was the observation showing a downregulation of trkA, trkB, and trkC expression during the progression of AD (Figure 12). The protein products of trkB and trkC genes mediate the high-affinity binding of

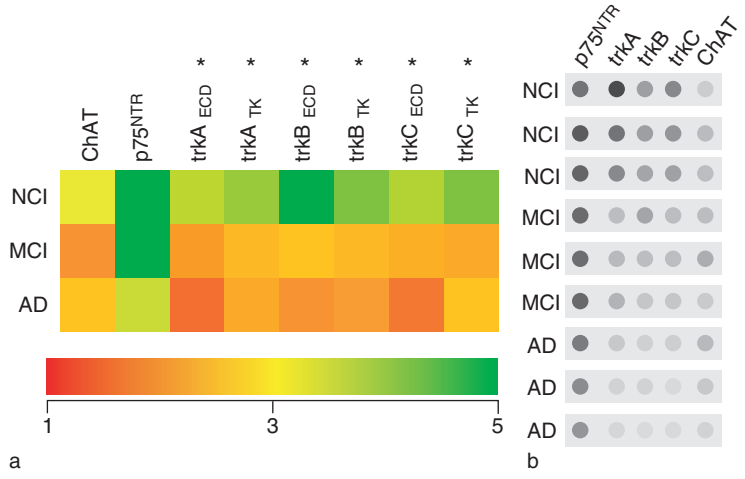

Figure 12 (a) Dendrogram with color-coded scale bar (orange is low expression and green is high expression), illustrating relative expression profiles of $p75^{NTR}$, trkA, trkB, trkC, and choline acetyltransferase (ChAT) derived from individual cholinergic nucleus basalis neurons from persons with no cognitive impairment (NCI), mild cognitive impairment (MCI), and Alzheimer's disease (AD). No significant differences are found for ChAT and $p75^{NTR}$ gene expression. In contrast, statistically significant downregulation (asterisks) of trkA, trkB, and trkC is observed in MCI and AD. Expressed sequence tags identifying extracellular domains (ECD) and tyrosine kinase (TK) domains display downregulation. The decrement of trk gene expression in MCI is intermediate relative to AD, indicating a step-down effect in expression levels from NCI to MCI to AD. (b) Representative custom-designed arrays illustrating expression-level differences between NCI, MCI, and AD. Three individual cases are depicted for each condition.

two neurotrophins, brain-derived neurotrophic factor and neurotrophin-3, respectively. A significant downregulation of trkA was seen in single NB neurons microaspirated from AD and MCI patients compared to normal controls. An intermediate reduction was seen in MCI, with the greatest decrement in mild AD compared to controls. The downregulation of the trk genes was associated with several measures of cognitive decline. By contrast, dysregulation of ChAT and $p75^{NTR}$ mRNA was not found across the three clinical groups examined. Together, these findings suggest a relative selectivity in the alteration of high-affinity neurotrophin receptors within single CBF neurons during the progression of AD, and that neurotrophic dysfunction in these neurons occurs during the earliest stages of cognitive decline.

Cholinergic NB neurons display the classic pathological feature of AD, neurofibrillary tangles (NFTs), which are composed primarily of hyperphosphorylated aggregates of the axonal microtubule-associated protein tau. Using similar single-cell array technology, the expression levels of the six tau isoforms, which are expressed through alternative splicing of a single tau gene on chromosome 17 that contains 16 exons, were assessed within single cholinergic NB neurons. The brain contains three isoforms with three tandem repeats (3Rtau; Unigene annotation MAPT1, MAPT3, and MAPT5) and three isoforms with four tandem repeats (4Rtau; MAPT2, MAPT4, and MAPT6) of 31 or 32 amino acids in the C-terminus of the molecule. Single-cell gene array analysis revealed that the global expression levels of the six tau transcripts and several other cytoskeletal transcripts within individual CBF neurons did not vary among AD and MCI patients and aged controls, consistent with previous reports. Surprisingly, when a 3Rtau/4Rtau ratio was calculated, a significant shift in the ratio was observed in MCI and AD, with a decrement in 3Rtau in relation to 4Rtau levels for all tau transcripts examined. These data suggest a subtle, yet pervasive, shift in the dosage of 3Rtau and 4Rtau within vulnerable CBF neurons in MCI and AD. Interestingly, a similar shift does not occur during normal aging. Shifts in the ratio of tau genes may be a fundamental mechanism whereby normal tau expression is dysregulated, leading to NFT formation. In this vein, hyperphosphorylation of tau in the cortex of transgenic anti-NGF mice (AD11) appears to be linked to NGF deprivation as opposed to a cholinergic deficit, since NGF but not the AChE inhibitor, galantamine, rescues this defect. It is possible that alterations in the NGF/TrkA/$p75^{NTR}$ signaling pathways might disrupt tau isoform stoichiometry, initiating NFT formation and subsequent downstream cytoskeletal neuronal effects within CBF neurons. Taken together, these gene expression studies suggest that a 'shift' in neurotrophic signaling may underlie the differences in cholinergic function in aging and AD.

Therapeutics for Cholinergic NB Dysfunction in AD

Currently there are no treatments to prevent CBF dysfunction in this devastating disease. Several strategies have been employed to influence the neurotransmission or survival of these neurons. Several anticholinesterase inhibitors, including galantamine, have proved only to have modest effects on cognition, and are sometimes associated with unwanted side effects. Esterase inhibitors in combination with an *N*-methyl-D-aspartate (NMDA) receptor agonist (e.g., memantine) have been suggested to provide greater symptomatic improvement in patients with mild AD. Recently, several novel approaches aimed at rescuing CBF neurons based on NGF and its receptors have been under consideration. Various means of delivering NGF or augmenting NGF function in the brain are being developed to deliver neurotrophins to the brain in a fashion that is less invasive, albeit less targeted, than gene therapy. Small peptide analogs of NGF that contain or mimic the putative active region of the NGF β-component should be considered. These molecules possess the potential advantage of crossing the blood–brain barrier due to their low molecular weight, eliminating the need for invasive brain delivery via surgical procedures. Alternatively, the ongoing development of TrkA-specific NGF peptidomimetics has provided the groundwork for another potential approach to improving prosurvival signaling in cholinergic NB neurons in people during the early stages of AD. The importance of defining the neurotrophic molecular mechanisms underlying the selective vulnerability of CBF neurons has gained greater prominence with the findings from a phase I clinical trial demonstrating that *ex vivo* NGF gene therapy improves cognition and induces CBF neuronal sprouting in mild AD. To date, these patients have not shown any adverse affects from these NGF intraparenchymal injections, and early indications are that NGF therapy is efficacious and may even slow the course of the disease.

See also: Acetylcholine Neurotransmission in CNS; Acetylcholinesterase Inhibitors and Alzheimer's Disease; Acetylcholinesterase; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: MRI Studies; Animal Models of Alzheimer's Disease; Axonal Transport and Alzheimer's Disease; Basal Ganglia: Acetylcholine Interactions and Behavior; Cholinergic

Pathways in CNS; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Neuromuscular Junction (NMJ): Acetylcholinesterases; Neurotransmitters and Growth Factors: Overview; Nicotinic Acetylcholine Receptors; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Auld DS, Kornecook TJ, Bastianetto S, et al. (2002) Alzheimer's disease and the basal forebrain cholinergic system: Relations to beta-amyloid peptides, cognition, and treatment strategies. *Progress in Neurobiology* 68: 209–245.

Bruno MA and Cuello C (2006) Activity-dependent release of precursor nerve growth factor, conversion to mature nerve growth factor, and its degradation by a protease cascade. *Proceedings of the National Academy of Sciences of the United States of America* 103: 6735–6740.

Cooper JR, Bloom FE, and Roth RH (2003) *Acetylcholine: The Biochemical Basis of Neuropharmacology*, pp. 151–179. New York: Oxford University Press.

Counts SE and Mufson EJ (2005) The role of nerve growth factor receptors in cholinergic basal forebrain degeneration in prodromal Alzheimer's disease. *Journal of Neuropathology and Experimental Neurology* 64: 263–272.

Counts SE, Perez SE, Kahl U, et al. (2001) Galanin: Neurobiologic mechanisms and therapeutic potential for Alzheimer's disease. *CNS Drug Reviews* 7: 445–470.

DeKosky ST, Ikonomovic MD, Styren SD, et al. (2002) Upregulation of choline acetyltransferase activity in hippocampus and frontal cortex of elderly subjects with mild cognitive impairment. *Annals of Neurology* 51: 145–155.

Ginsberg SD, Che S, Counts SE, et al. (2006) Single cell gene expression profiling in Alzheimer's disease. *NeuroRx* 3: 302–318.

Gotti C, Zoli M, and Clementi F (2006) Brain nicotinic acetylcholine receptors: Native subtypes and their relevance. *Trends in Pharmacological Sciences* 27: 482–491.

Lad SP, Neet KE, and Mufson EJ (2003) Nerve growth factor: Structure, function and therapeutic implications for Alzheimer's disease. *Current Drug Targets – CNS & Neurological Disorders* 2: 315–334.

Longo FM, Xie Y, and Massa SM (2005) Neurotrophin small molecules mimetics: Candidate therapeutics agents for neurological disorders. *Current Medicinal Chemistry – Central Nervous System Agents* 5: 29–41.

Mufson EJ and Kordower JH (2001) Cholinergic basal forebrain systems in the primate central nervous system: Anatomy, connectivity, neurochemistry, aging, dementia, and experimental therapeutics. In: Hoff PR and Mobbs CV (eds.) *Functional Neurobiology of Aging*, pp. 243–281. San Diego, CA: Academic Press.

Roux PP and Baker PA (2002) Neurotrophin signaling through the p75 neurotrophin receptor. *Progress in Neurobiology* 67: 203–233.

Salehi A, Delcroix JD, Mobley WC, et al. (2003) Traffic at the intersection of neurotrophic factor signaling and neurodegeneration. *Trends in Neuroscience* 26: 73–80.

Sofroniew MV, Howe CL, and Mobley WC (2001) Nerve growth factor signaling, neuroprotection, and neural repair. *Annual Review of Neuroscience* 24: 1214–1281.

Tuszynski MT, Thal L, Pay M, et al. (2005) A phase 1 clinical trial of nerve growth factor gene therapy for Alzheimer disease. *Nature Medicine* 11: 551–555.

Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease

S J Teipel, University of Rostock, Rostock, Germany
L T Grinberg, Faculdade de Medicina da Universidade de Säo Paulo and Instituto Israelita de Ensino e Pesquisa Albert Einstein, Sao Paulo, Brazil
H Hampel, Ludwig-Maximilian University, Munich, Germany; and Trinity College Dublin, The Adelaide and Meath Hospital, Incorporating the National Children's Hospital, Dublin, Republic of Ireland
H Heinsen, University Würzburg, Würzburg, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The cholinergic system has received increasing attention from both basic scientists and clinical researchers after the observation that Alzheimer's disease (AD), the most frequent human neurodegenerative disorder, is specifically associated with an early and selective impairment of cholinergic transmission in the brain. Postmortem examinations showed severe neuronal loss in the basal forebrain cholinergic nuclei and a decrease of choline acetyltransferase (ChAT) and cholinesterase activity in the cerebral cortex in AD. Furthermore, anticholinergic drugs induce memory dysfunction in both animal models and humans. Experimental lesions of cholinergic nuclei impair memory function in monkeys. Even in the absence of AD-related neurodegeneration, decline of cholinergic function has been linked to specific cognitive impairment, including working memory, in healthy aging. It has been suggested that the effects of cholinergic lesions on memory are mediated by the primary effect of cholinergic transmission on attention and executive function. Age-associated alterations of the cholinergic system may underlie an increased risk of delirium in aging. Besides its role as a neurotransmitter in the adult brain, several lines of research suggest that acetylcholine (ACh) plays a critical role in cortical development and in establishing synaptic contacts in neuronal networks that will subserve complex cognitive functions in adulthood. However, the cerebral cortex continues to adapt to the environment throughout the entire life span. Cholinergic function plays a critical role in neuronal plasticity in the adult brain. Therefore, cholinergic impairment in AD not only leads to impaired neuronal transmission within the cerebral cortex but also compromises the ability of the brain to compensate for progressive neurodegeneration. Some of the vulnerability of the cholinergic system in aging or AD may be mediated by an amyloid-related decrease of target-specific nerve growth factor. Consistent with the evidence for cholinergic dysfunction, several double-blind placebo-controlled trials have suggested that enhancement of cholinergic function using (acetyl-) cholinesterase inhibitors attenuates cognitive decline in AD. The findings on cholinergic dysfunction, however, have been limited in regard to the time course and the differential involvement of systems and compartments of the cholinergic system. These limitations are partly because evidence was derived from postmortem studies that intrinsically forbid longitudinal observation and are biased toward severe end stages of disease. Therefore, advanced imaging technology was developed to determine the integrity of the cholinergic system *in vivo*. One line of research is based on radiolabeled ligands to specific components of cholinergic transmission together with positron emission tomography (PET). Another approach uses pharmacological intervention with cholinergic and anticholinergic drugs and functional neuroimaging, using PET and functional magnetic resonance imaging (fMRI). Finally, studies have used structural MRI to determine alterations of cholinergic nuclei in the human basal forebrain. Understanding the rationale, implications, and limitations of these approaches requires a basic understanding of the anatomy and biochemistry of cholinergic transmission in the human brain, outlined in the following paragraphs.

The Cholinergic System

Anatomy of the Cholinergic System

According to Mesulam, the cholinergic pathway "is likely to constitute the single most substantial regulatory afferent of the cerebral cortex." Cholinergic projections in the cerebral cortex are involved in arousal, maintenance of attention, and REM sleep. The main source of cholinergic projections to the cerebral cortex are the magnocellular neurons of the nucleus basalis Meynert in the substantia innominata which separates the globus pallidus from the ventral surface of the basal forebrain. Studies in monkeys suggest a corticotopic organization of basal forebrain cholinergic nuclei. The cholinergic nuclei of the basal forebrain can be divided into subpopulations and named by adapting Mesulam's nomenclature. Cell group Ch1 corresponds to the nucleus septi medialis and innervates the hippocampus, cingulate, and retrosplenial cortex. Ch2 corresponds to the vertical limb of the nucleus of the diagonal band of Broca. Both groups are located medially to the nucleus

accumbens septi. Ch3 designates cells associated with the horizontal limb of the diagonal band nucleus and, together with Ch2, innervates the entorhinal and olfactory cortex. The Ch4 group contains the nucleus basalis neurons within the substantia innominata and is the most extensive cholinergic cell group. The Ch4 group can be divided into anterior (Ch4a), intermediate (Ch4i), and posterior (Ch4p) subpopulations. The anterior subpopulation can be divided into medial (Ch4am) and lateral (Ch4al) cell groups. Ch4a extends from the posterior end of the tuberculum olfactorium to the posterior end of the anterior commissure. Ch4i starts at the anterior tip of the thalamus and extends to the posterior end of the ansa peduncularis, Ch4p extends from the posterior end of the ansa peduncularis to the level of the mammillary bodies. In the human brain there are two major pathways from Ch4 neurons to the cortex: the medial pathway supplies the medial frontal cortex and cingulum, and the lateral pathway projects to the dorsal and lateral prefrontal cortex and the temporal, parietal, and occipital neocortex. A cluster of magnocellular cholinergic neurons located in the anterior part of the basal forebrain (ventrolaterally to the anterior commissure) has been named the nucleus subputaminalis (NSP), or Ayla's nucleus. It is interesting that the lateral subdivision of the NSP has been described only in the human brain, not yet in monkey brains, and has been linked to the cortical speech area. Figures 1–6 illustrate the anatomy of the basal forebrain nuclei based on Gallocyanin-stained sections through the basal forebrain of two control cases, a 29-year-old man and a 66-year-old man.

Because all cholinergic nuclei are concentrated in a very small area of the brain, basal forebrain lesions will easily lead to cholinergic deficits. In the same line, anticholinergic effects of medication or other substances will easily lead to cholinergic deficits, particularly in aging and neurodegeneration, in which the cholinergic input into the cortex is already decreased or damaged.

Cholinergic Neurotransmission

Because ACh is among the most biologically potent neurotransmitters of the brain, its synaptic concentration is highly regulated. ACh is synthesized through ChAT from choline and acetyl-Co-A (Figure 7). The cholinesterases (ChEs) inactivate ACh through hydrolyzation (Figure 8).

The primary enzyme is acetylcholinesterase (AChE), but ACh is also a substrate for butyrylcholinesterase. Inferences of the regional distribution of cholinergic projections based on ChE produce too widespread estimates. ChE is widely distributed across the brain, probably as a safeguard against ACh overload within the brain. The localization of cholinergic projections is more precisely inferred from the distribution of the ChAT. ACh acts on two classes of receptors that are characterized by the pharmacological activity of two alkaloids: nicotine and muscarine. The nicotinic ACh receptor is a ligand-gated ion channel eliciting a fast response whereas the muscarinic ACh receptor is G-protein-coupled and elicits a slow response. Both receptor classes exhibit distinct subtypes, are often co-expressed in the same synapse, and occur pre- and postsynaptically.

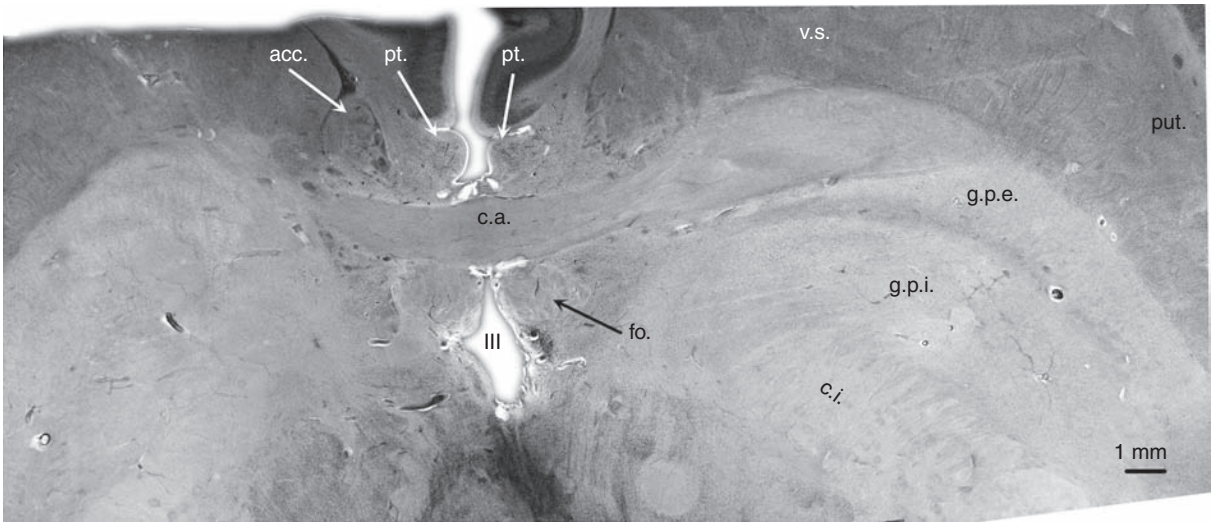

Figure 1 Horizontal section through the basal forebrain. Horizontal gallocyanin-stained 440 μm-thick celloidin-mounted section through the fornix and dorsomedial parts of the anterior commissure of the right hemisphere of a 29-year-old man. Red line in Figure 6 indicates plane of section in this figure. acc., nucleus accumbens.; c.a., anterior commissure; c.i., internal capsule; fo., fornix; g.p.e., external globus pallidus; g.p.i., internal globus pallidus; III, third ventricle; pt., paraterminal (subcallosal) gyrus; put., putamen; v.s., ventral striatum. Reproduced from Grinberg LT and Heinsen H (2007) *Dementia and Neuropsychologia* 1(2), with permission.

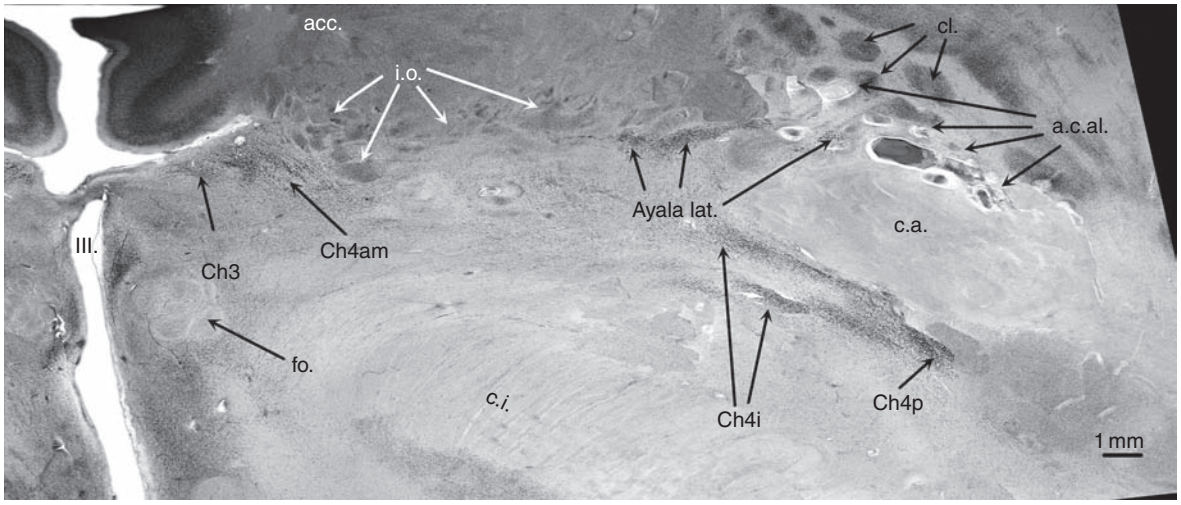

Figure 2 Horizontal section through the basal forebrain. Horizontal gallocyanin-stained 440 μm-thick celloidin-mounted section 4.4 mm ventral to plane of section indicated in **Figure 1**, right hemisphere of a 29-year-old man. Plane of section is indicated in **Figure 6** by the white line. a.c.al., central anterolateral arteries; acc., nucleus accumbens.; Ayala lat., lateral or periputaminal part of Ayala's nucleus; c.a., anterior commissure; Ch4am, basal nucleus of Meynert, anteromedial part; Ch4i, basal nucleus of Meynert, intermediate part; Ch4p, basal nucleus of Meynert, posterior part; c.i., internal capsule; cl., claustrum; fo., fornix; III, third ventricle; i.o., olfactory islands. Reproduced from Grinberg LT and Heinsen H (2007) *Dementia and Neuropsychologia* 1(2), with permission.

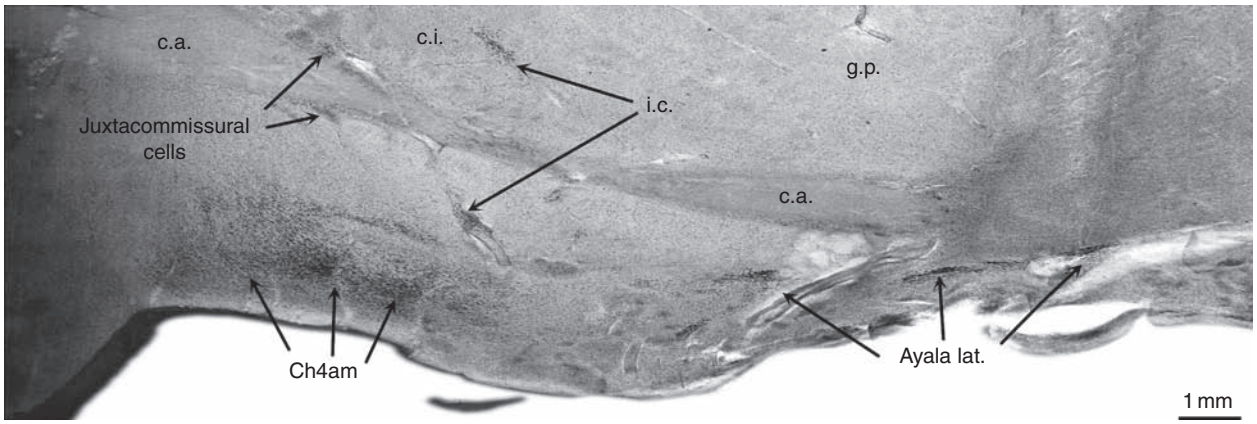

Figure 3 Coronal section through the basal forebrain. Coronal gallocyanin-stained 440 μm-thick celloidin-mounted section through the fornix and dorsomedial parts of the anterior commissure of the right hemisphere of a 66-year-old man. Ayala lat., lateral or periputaminal part of Ayala's nucleus; c.a., anterior commissure; Ch4am, basal nucleus of Meynert, anteromedial part; c.i., internal capsule; g.p., globus pallidus; i.c., interstitial cells. Reproduced from Grinberg LT and Heinsen H (2007) *Dementia and Neuropsychologia* 1(2), with permission.

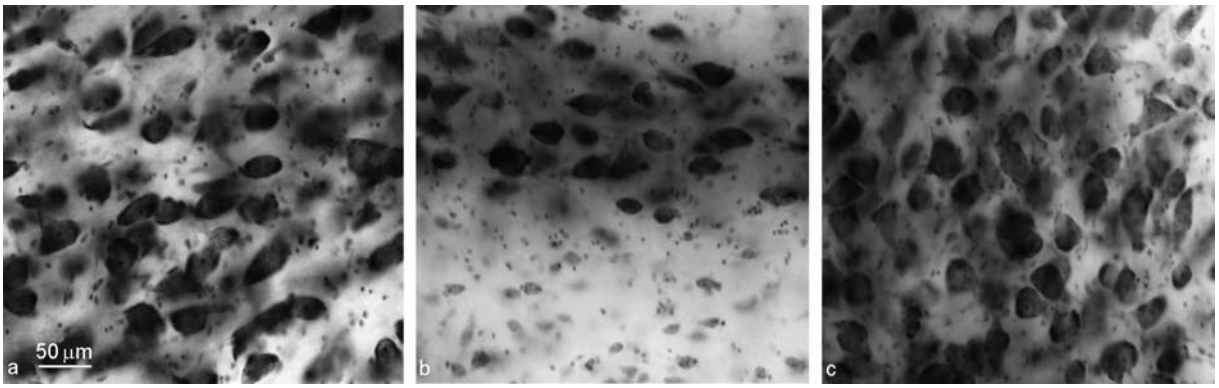

Figure 4 Details from **Figures 1–3**. Microscopic view of (a) large chromophilic neurons of basal nucleus of Meynert, anteromedial part; (b) Ayala's lateral or periputaminal nucleus; and (c) basal nucleus of Meynert, posterior part (66-year-old man). Reproduced from Grinberg LT and Heinsen H (2007) *Dementia and Neuropsychologia* 1(2), with permission.

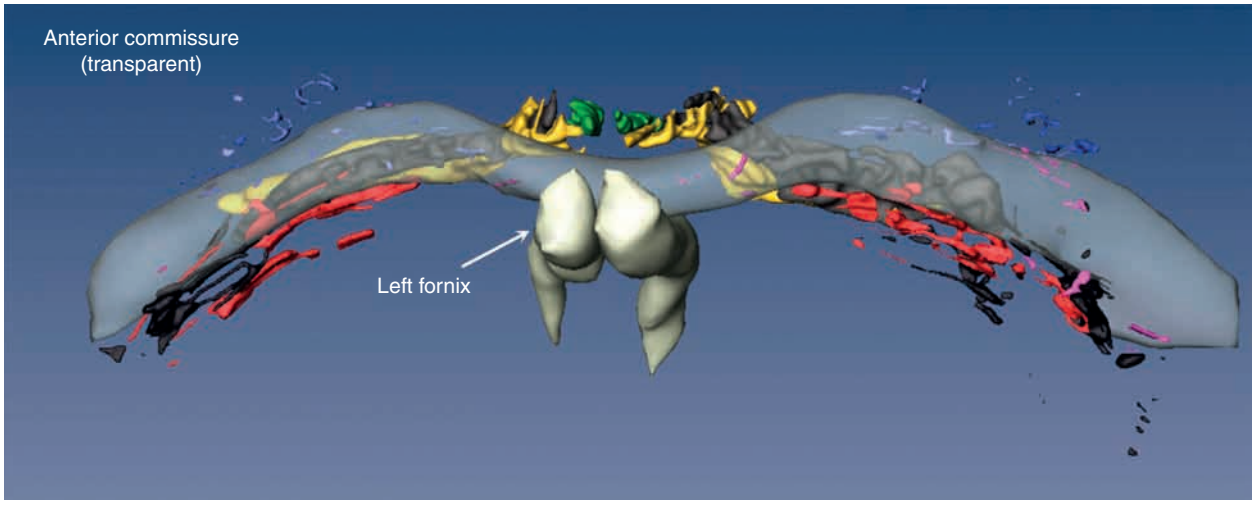

Figure 5 Three-dimensional reconstruction of basal forebrain. Basal forebrain complex seen from dorsal after reconstruction with Amira software, 29-year-old man. Ch2, vertical limb of the nucleus of the diagonal band of Broca; Ch3, cells associated with the horizontal limb of the diagonal band nucleus; Ch4al, basal ncl. of Meynert, anterolateral part; Ch4am, basal ncl. of Meynert, anteromedial part; Ch4i, basal nucleus of Meynert, intermediate part; Ch4p, basal nucleus of Meynert, posterior part. Reproduced from Grinberg LT and Heinsen H (2007) *Dementia and Neuropsychologia* 1(2), with permission.

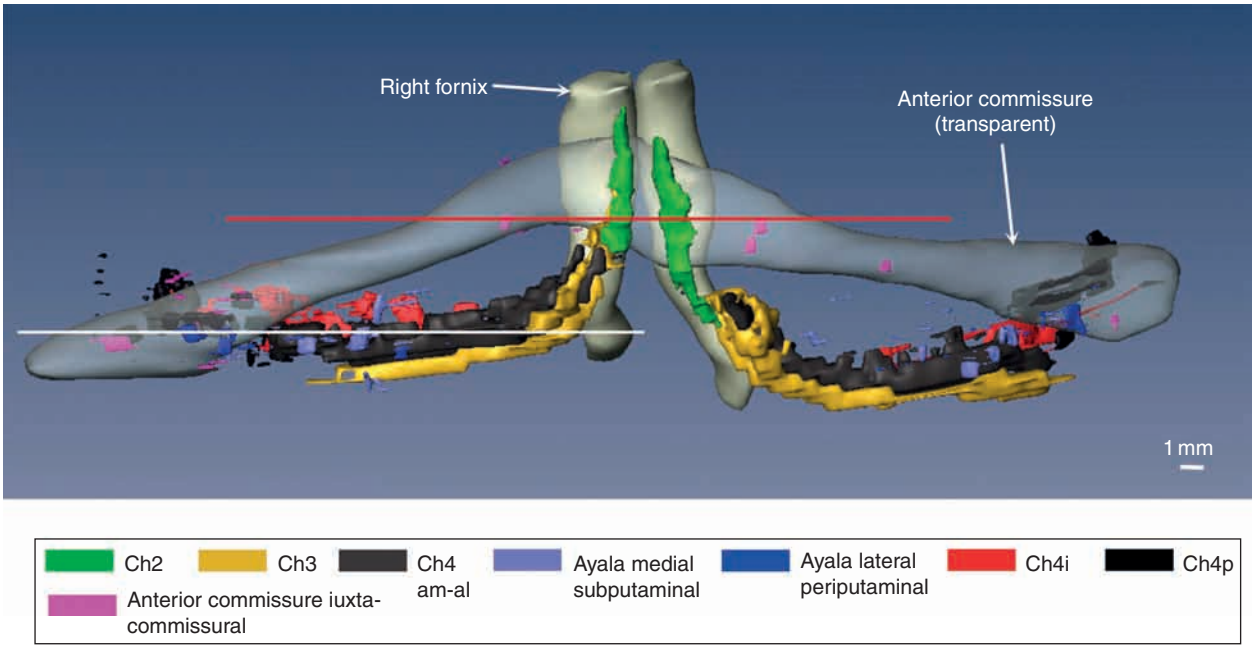

Figure 6 Three-dimensional reconstruction of basal forebrain, view from anterior. Basal forebrain complex seen in a fronto-occipital perspective after reconstruction with Amira software, 29-year-old man. Ch2, vertical limb of the nucleus of the diagonal band of Broca; Ch3, cells associated with the horizontal limb of the diagonal band nucleus; Ch4al, basal ncl. of Meynert, anterolateral part; Ch4am, basal ncl. of Meynert, anteromedial part; Ch4i, basal nucleus of Meynert, intermediate part; Ch4p, basal nucleus of Meynert, posterior part. Reproduced from Grinberg LT and Heinsen H (2007) *Dementia and Neuropsychologia* 1(2), with permission.

Figure 7 Acetylcholine synthesis. The reaction mediated by the choline acetyltransferase activity is shown.

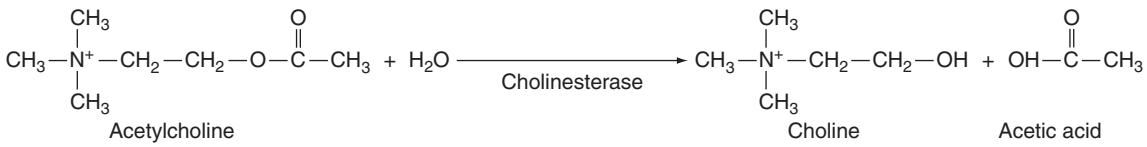

Figure 8 Acetylcholine inactivation. The reaction mediated by acetylcholinesterase activity is shown.

Targeting the Cholinergic System *In Vivo*

The short description of the anatomy and biochemistry of the cholinergic system in the brain reveals several potential targets for *in vivo* imaging of the integrity of the cholinergic system. Radiochemistry has developed ^{11}C-labeled markers for nicotinic and muscarinic ACh receptors and for AChE that can be applied *in vivo* using PET. Cholinergic drugs together with cognitive tests during functional imaging using fMRI or PET have been used as a stress test to map the altered plasticity of the cholinergic system with aging and neurodegeneration. Finally, *in vivo* MRI has been developed to detect anatomical changes in the basal forebrain that may be related to the integrity of cholinergic neurons. The following paragraphs review the technical background of and findings obtained by these approaches.

Pet Imaging of Cholinergic Receptor Binding and ChE Activity

How can PET measure neuroreceptor binding and density *in vivo*? In a neuroreceptor study with PET, the measured activity will always be a mixture of specifically bound, nonspecifically bound, and unbound radiotracer. Besides the binding kinetics and receptor distribution, peripheral clearance, regional cerebral blood flow (CBF), and transport across the blood–brain barrier will influence the profile of the time–activity curve. Therefore, the use of a simple quantitative measure, such as the ratio of a region of interest to a reference region (believed to represent nonspecific binding), will be inadequate to most receptor studies. Model-based approaches have been developed to relate the time–activity curve of a region of interest (the observed response function) to the arterial time–activity curve (the observed input function) through a defined model. Most of these models are based on the notion of compartments. In its most extended form, the compartment model encompasses three compartments, which are derived from a four-compartment model by collapsing two compartments into one (Figure 9).

A radiolabeled ligand is injected intravenously (compartment 1), resulting in an initial plasma concentration (C_{Pl}) that rapidly decreases because of uptake into brain (and other tissues), resulting in an early equilibrium between vascular and extravascular

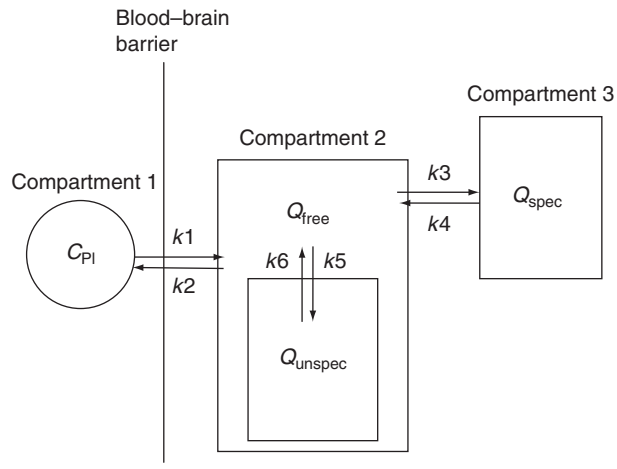

Figure 9 Three-compartment model of ligand–receptor binding. Compartment 1 indicates plasma space; compartment 2, nonspecific binding space in tissue; and compartment 3, specific binding space. C_{Pl} represents ligand concentration in plasma; Q_{free} and Q_{unspec}, quantity of free exchangeable and nonspecifically bound ligand rapidly equilibrating with each other; and Q_{spec}, quantity of receptor-bound ligand. $k1$ to $k6$ are the rate constants for the ligand exchange between compartments in the specified directions. For further details, see text. Modified from Holcomb HH, Links J, Smith C, and Wong D (1989) Positron emission tomography: Measuring the metabolic and neurochemical characteristics of the living human nervous system. In: Andreasen N (ed.) *Brain Imaging: Applications in Psychiatry*, pp. 235–370. Washington, DC: American Psychiatric Press.

brain compartments. The extravascular compartment is divided into three parts. The pool of unbound ligand (quantity of free ligand (Q_{free})) is in rapid equilibrium with the nonspecific bound ligand pool (quantity of ligand reversibly bound to nonspecific proteins (Q_{unspec})). Both compartments are considered a single exchangeable tissue pool (compartment 2). The quantity of free exchangeable ligand in compartment 2 (Q_{free}) tends to equilibrate with the quantity (Q_{spec}) of ligand specifically bound to neuroreceptor in compartment 3. The fractional rate constants k_i govern the exchange of ligand between compartments in the specified direction, assumed to follow a first-order reaction kinetic.

Equilibrium is supposed to occur rapidly between the first (plasma) and second compartments (free and nonspecifically bound ligand). In contrast, the equilibration between the second and third compartments (specifically bound ligand) takes considerable time,

depending on ligand–receptor affinity, and may even not occur within scanning time. There are essentially two models to determine neuroreceptor density and binding characteristics. The first measures receptor saturation with increasing ligand concentration and therefore requires equilibration of ligand between the second and third compartments during scanning time (steady-state model). The second measures the time course of specific brain uptake of ligand over time with different concentrations of an unlabeled receptor-blocking substance (kinetic model). This approach does not require equilibration between the nonspecific and specific compartments. Both models to assess neuroreceptor density and binding kinetics are based on several assumptions, some of which may be reasonable but may not be easily validated. Both approaches gave similar values for receptor density with different dopaminergic ligands (Wong: 17 pmol g^{-1}, Sedvall: 15 pmol g^{-1}). The results with respect to receptor binding affinity, however, are not directly comparable between both models, because different ligands are involved and different parameters are measured. In both approaches the tracer kinetic model requires an input function indicating delivery of tracer to the tissue. In some applications for receptor studies, an arterial input function can be replaced by a reference to tissue devoid of specific binding sites. Single scan acquisition allows for assessment of only one single parameter, whereas dynamic scanning to obtain the time course of tissue activity allows simultaneous estimation of multiple parameters.

Reliable quantification of receptor densities in AD using PET and single-photon emission computed tomography (SPECT) will require an appropriate model to rule out effects of cortical atrophy. Correction of PET data for cerebral atrophy based on corresponding high-resolution MRI scans has already been accomplished in a study on the μ-opioid receptor in the amygdala complex and may be routinely available in the future for other PET studies as well.

Cholinergic Markers in PET Table 1 gives an overview of the most important markers for the cholinergic system in SPECT and PET. Although a decreased nicotinic receptor density in AD has been established from neuropathological studies, studies of the cholinergic system can serve as an example of the pitfalls and controversies associated with *in vivo* receptor imaging using PET. Initial labeling of racemic nicotine with positron emitting ^{11}C in 1982 by Långström and thereafter of the nicotine (+)-(R)- and (−)-(S)-enantiomers was followed by attempts to visualize binding to the brain nicotinic ACh receptors with PET in monkeys and in humans. In AD, uptake of ^{11}C-nicotine into frontal and temporal cortices was less than in healthy age-matched controls. These results are consistent with postmortem evidence of fewer nicotinic

Table 1 PET radioligands for the cholinergic system

Compound	*Label*	*Target*	*Main findings*
2-[18F]fluoro-3-(2(S)-azetidinylmethoxy) pyridine (2-F-A-85380)	^{18}F	$\alpha_4\beta_2$-nicotinic ACh-receptor (responsible for addicitive properties of nicotine?)	Applied to humans; no studies in AD so far reported. 88% occupancy of receptor after smoking of 1 cigarette (dependent smokers)
Nicotine	^{11}C	nAChR	Unspecific binding, kinetic model suggests perfusion as main determinant of distribution
Norchlorofluoroepibatidine	^{18}F	$\alpha_4\beta_2$-nicotinic AChR	Studies in baboons but not in humans because of toxicity of epibatidine
N-methyl-4-piperidyl benzilate (NMPB)	^{11}C	mAChR	Reduced cortical uptake in AD
3-R-quinuclidinyl-4-S-iodobenzilate (RS-IQNB)	^{123}I	mAChR	Reduced uptake temporal and frontal in AD; SPECT marker
Tropanyl benzilate (TRB)	^{11}C	mAChR	Slightly decreased binding in frontal cortex in aging
3-(3-(3-Fluoropropyl)thio)-1,2,5-thiadiazol-4-yl)-1,2,5,6-tetrahydro-1-methylpyridine ([18F] FP-TZTP)FP-TZTP	^{18}F	mAChR (M_2)	Greater receptor avidity in ApoE4 carriers, possibly reflecting lower occupancy by ACh
N-methyl-4-piperidyl acetate (MP4A)	^{11}C	AChE	Reductions in AD
N-[^{11}C]methyl-pi-peridin-4-yl propionate ([^{11}C]PMP)	^{11}C	AChE	Reductions in AD, correlations with cognitive performance

ACh, acetylcholine; AChE, acetylcholinesterase; AD, Alzheimer's disease; ApoE4, apolipoprotein E; mAChR, muscarinic acetylcholine receptor; nAChR, nicotinic acetylcholine receptor; SPECT, single-photon emission computed tomography.

ACh binding sites in AD. However, the validity of nicotinic receptor imaging *in vivo* with ^{11}C-nicotine was first questioned by Nybäck and co-workers. They showed in healthy participants that ^{11}C-nicotine uptake did not decline after potential binding sites were blocked by unlabeled ligand, indicating no measurable impact of receptor availability on tracer uptake. The ratio between tracer uptake in temporal cortex and white matter was constant over time (compared with the previously described linearly ascending ratio between specific binding in the putamen and nonspecific binding in the cerebellum), which is not compatible with specific binding characteristics. Finally, the experimental data could be fitted with a two-compartment model, in which tracer uptake was determined solely by blood flow and blood–brain barrier permeation. A specific binding compartment was not required. Nybäck and colleagues concluded from these results that the regional reduction of ^{11}C-nicotine uptake in AD reflects reduced CBF.

In the following years, new compounds were investigated for their potential to specifically detect nicotinic receptor binding *in vivo*. [^{18}F]norchlorofluoroepibatidine was successfully applied in baboon studies but because of toxicity could not be transferred to human studies. More-selective and less-toxic 3-pyridyl ether derivatives have been used successfully for development of a nicotinic ACh receptor (nAChR) ligand. The recently developed radioligands 2-[^{18}F]fluoro- and 6-[^{18}F]fluoro-3-(2(*S*)-azeti-dinylmethoxy) pyridine allow for visualization of brain $\alpha_4\beta_2$-nAChR with PET in humans. This nAChR subtype is expressed throughout the brain in pre- and postsynaptic locations and has been linked to the addicting properties of nicotine. *In vitro* studies on postmortem brain material using different ligands for autoradiography have obtained conflicting results with reductions of this specific nAChR in some but not all studies. The *in vivo* ligand A-85380 has shown its feasibility and validity in a familial form of frontal lobe epilepsy with mutations in the nAChR α4 or β2 subunit but has not yet been applied to AD.

Several tracers have been applied to detect muscarinic AChRs (mAChRs). In postmortem studies these receptors have been found to be decreased with AD. In aging, autopsy studies have been inconsistent, with series with the greatest age ranges and participant numbers indicating more than 50% reductions in muscarinic receptor numbers in the striatum, hippocampus, and frontal cortex between ages 4 and 93 years. However, declines in cortical or hippocampal muscarinic receptors in series between the ages of 60 and 90 years were modest in some studies or not detectable in other studies. The first ligand to map mAChR *in vivo* in the human brain was 3-quinuclidinyl-4-iodobenzilate labeled with ^{123}I (^{123}IQNB) in SPECT. In older normal adults, this tracer showed high radioligand uptake in the basal ganglia and occipital and insular cortex, and low uptake in thalamus and cerebellum. This regional pattern matches the relative distribution of mAChR in autopsy studies. Studies in AD showed a moderate decrease of tracer uptake in posterior temporal and frontal cortices. Unfortunately, the early uptake of ^{123}IQNB is flow-dependent, so scans had to be performed 21 h after injection, making this probe difficult to handle in a clinical setting. Moreover, the spatial resolution of SPECT cameras is generally much lower than the resolution of PET, lowering the potential use of this approach for clinical application or research.

Binding of ^{11}C-tropanyl benzilate (TRB) using PET was slightly but significantly decreased in the frontal cortex with aging. Binding showed a nonsignificant trend of increase in cerebellum and brain stem with aging. The authors interpreted these effects as indicating small decline of receptor density in frontal cortex in healthy aging. TRB has not yet been applied to studying patients with AD or other neurodegenerative disorders.

In one PET study using ^{11}C-labeled *N*-methyl-4-piperidyl benzilate (NMPB), normal aging was associated with a reduction in muscarinic receptor binding in neocortical regions and thalamus. In AD patients, the three-compartment model appeared capable of dissociating changes in tracer transport from changes in receptor binding but suffered from statistical uncertainty, which required normalization to a reference region, therefore limiting its potential use in the study of neurodegenerative processes. After normalization, no regional changes in muscarinic receptor concentrations were observed in AD. These findings are in contrast to an earlier study combining measurement of mAChR with ^{11}C-NMPB and ^{18}fluorodeoxyglucose (FDG)-PET of cortical glucose consumption and ^{99}Tc-hexamethylpropyleneamine oxime (HMPAO)-SPECT of CBF. In this study, AD patients had reduced ^{11}C-NMPB uptake throughout the entire cortex relative to controls, which was more pronounced than reductions in CBF but less pronounced than changes in cortical glucose metabolism.

Application of M2 muscarinic receptor ligand 3-(3-(3-[^{18}F]Fluoropropyl)thio)-1,2,5-thiadiazol-4-yl)-1,2,5,6-tetrahydro-1-methylpyridine ([^{18}F]FP-TZTP) during PET revealed greater M2 receptor avidity in older adults carrying the apolipoprotein E4 (ApoE4) allele, a risk allele for sporadic AD. This alteration might be due to a decrease in synaptic ACh concentration through lower receptor occupancy. Both age and ApoE4 genotype independently led to a greater decrease in receptor avidity following the infusion of the AChE inhibitor physostigmin, probably reflecting the

increased receptor occupancy by the endogenous ligand ACh. Other markers, such as *N*-^{11}C-methyl-3-piperidyl benzilate, have shown reduced mAChR binding in monkey cortical regions with aging but have not been applied to human studies so far.

In summary, in contrast to the nicotinic markers, a wide range of muscarinic markers are available with PET and SPECT, and these markers have been evaluated using kinetic modeling in aging and in AD. Results, however, are not absolutely consistent to date. Effects of aging on muscarinic receptor density appear to be modest at most, with the effects being concentrated in the frontal lobes. In AD, effects appear to be more widespread if they are detected at all. The most interesting finding to date seems to be the study using [18F]FP-TZTP with PET. This study showed an increased binding volume for the ligand, possibly reflecting decreased availability of endogenous ACh in at-risk groups of AD defined by the presence of the ApoE4 genotype.

A promising approach to quantify cholinergic impairment in AD was presented by Iyo and collaborators, who used an ^{11}C-labeled substrate for AChE (*N*-methyl-4-piperidyl acetate (MP4A)) to measure catalytic activity of brain AChE with PET. This lipophilic tracer is taken up into the brain across the blood–brain barrier, where it is hydrolyzed by brain AChE to a hydrophilic metabolite that remains trapped within the brain. The ratio of metabolized to unmetabolized MP4A gives an estimate of AChE catalytic activity. Using this model in a pilot study with five AD patients and eight healthy age-matched controls, the investigators found AChE activity in AD reduced 20–25% in frontal and occipital cortex and 30–45% in temporal and parietal cortex. Subsequent studies using *N*-[^{11}C]methyl-pi-peridin-4-yl propionate as substrate for AChE confirmed an average decrease of 25–33% of mean cortical AChE activity in patients with moderate AD and 9–15% in mild AD. The extent of reduction in enzyme activity *in vivo* was much less than expected from postmortem findings. Still, the changes were more pronounced than were independently measured reductions of regional CBF. These data suggest that cholinergic denervation may be mild to modest in early stages of disease but appears to be independent of reduced CBF. Loss of cholinergic enzyme activity may precede metabolic and blood flow alterations in early AD, when severe functional impairment and final loss of neurons (both leading to decreased uptake of glucose) may not yet be markedly present. One study investigating cortical AChE activity and cognitive performance suggested that AChE activity is predominantly correlated with measures of working memory and attention, much less with measures of delayed short- and long-term memory. These data agree with experimental evidence that cholinergic transmission is predominantly involved in attention and executive function.

The hope that specific receptor imaging might improve clinical diagnosis and treatment of AD has not yet been fulfilled. The complexity of the methods, their limited availability, and inconsistency of results may be the major reasons why no studies have been reported so far using receptor PET as a diagnostic marker or secondary endpoint in pharmacological trials. Another reason may be the complex regulation of the cholinergic system. Surprisingly, an autopsy study based on the religious order population has shown an increase of ChAT activity in people with mild cognitive impairment whereas patients with AD showed a consistent reduction of cortical ChAT activity. In a subsequent discussion, many researchers argued that this finding may indicate that cholinergic lesions occur at a late stage in the development of the disease. A more differentiated discussion, however, has pointed out that some subregions of the basal forebrain cholinergic nuclei appear to be preserved until late stages of disease and might be the source of new cholinergic synapses into denervated brain areas. The stimulus for this sprouting of cholinergic axons might be the progressive loss of target neurons in the most vulnerable brain areas, such as the entorhinal cortex. These data, together with increased muscarinic receptor avidity reported in some studies of mild stages of AD, suggest a dynamic response of the cholinergic system to progressive neurodegeneration and decrease of ACh concentrations. These mechanisms most likely vary with the clinical stage of the disease and therefore make it difficult to identify patterns of receptor changes that are predictive of AD or respond specifically to treatment.

Perspective of neuroreceptor imaging in Alzheimer's disease Future studies using these techniques will have to address several issues. First, they need to more extensively study alterations in cholinergic transmission and ChE activity in preclinical stages and at-risk stages of AD. The finding of an upregulation of muscarinic receptor binding and ChAT activity in patients with mild cognitive impairment and even cognitively healthy people with ApoE4 genotype suggests preclinical compensatory processes whose breakdown may contribute to the clinical manifestation of dementia. Neuroreceptor imaging may contribute to a better understanding of these mechanisms. The binary discussion (early cholinergic involvement in AD as primary event vs. secondary involvement as consequence of degeneration of cortical input areas) may have to be replaced by a more differentiated view appreciating a stepwise involvement of some cholinergic nuclei and

relative preservation of others that may partly compensate for the loss of cortical cholinergic innervation. Second, studies need to determine alterations in cholinergic binding over time to assess the potential use of these measures to map disease progression and treatment effects. Third, studies will have to consider that AD involves not only the cholinergic system but serotonergic, catecholaminergic, and glutamatergic transmission as well, any of which may contribute to both cognitive and behavioral symptoms of dementia.

Functional Imaging of Cholinergic Treatment Effects

The short-acting AChE inhibitor (AChE-I) physostigmine increases working memory performance in young and older healthy participants and at the same time decreases regional CBF in prefrontal areas that are involved in task performance. These findings may indicate that in healthy adults, enhancement of cholinergic function can improve processing efficiency and thus reduce the effort required to perform a working memory task. A study in monkeys showed that application of an AChE-I could restore the reduced response of regional CBF to vibrotactile stimulation with aging. These data suggest that increased reaction time and reduced working memory performance in aging may be related to a functional decline of cholinergic transmission that can partly be restored by the application of AChE-Is. In agreement with monkey studies, two independent PET studies of the ChE-I donepezil showed a reduction of cortical AChE activity associated with neuropsychological effects of treatment in AD patients. These studies confirmed the notion that application of ChE-I inhibits cortical AChE and that this inhibition is associated with an attenuation of cognitive decline. Interpretation, however, is limited because these studies used an open-label design. Reductions ranged between 27% and 39% at therapeutic doses. The use of higher doses of the presently available ChE-I to reach a higher degree of AChE inhibition is limited because of side effects. Intravenous application of physostigmine decreased cholinesterase activity by 50%. Agents with a longer half-life than physostigmine's and higher inhibition potential than the presently available ChE-I approved for AD are needed to test whether even greater inhibition of ChE might lead to better clinical effects.

The ChE-I rivastigmine increased cortical metabolism in treatment responders compared with nonresponders and placebo-treated AD patients. In open-label studies, the ChE-Is reminyl and metrifonate increased metabolism in specific cortical regions in treatment responders compared with nonresponders. One study found relative preservation of cortical metabolism in temporoparietal and frontal association areas after 12 months of treatment with rivastigmine compared with untreated AD patients. These studies suggest that cholinergic treatment increases cortical metabolism in association with cognitive improvement through treatment. Only one of these studies, however, was placebo-controlled. All studies compared treatment responders with nonresponders. However, this approach has the disadvantage that it is difficult to disentangle treatment effects from other factors influencing cognition. Comparing responders with nonresponders shows the cortical substrate of cognitive change, which might only partly be related to treatment (since by definition both comparison groups are treated in the same way). In a recent study, the effect of donepezil on cortical metabolism was determined during rest and passive audiovisual stimulation in a double-blind crossover design. Effects were compared between donepezil and placebo-treated patients independent of clinical response. Resting state glucose metabolism was increased with donepezil in left prefrontal cortex and decreased in right hippocampus. Cortical response to activation was increased in right hippocampus with donepezil compared with placebo, suggesting that decreased hippocampus metabolism during rest with donepezil treatment was reversible and even led to an overshoot activation. The effects were spatially much more restricted than in earlier studies. This may be related to the use of a double-blind design and the fact that patients were not classified according to clinical response.

The first small-scale studies with functional MRI (fMRI) have shown effects of cholinergic treatment on regional activation during a memory task. During face encoding healthy older patients showed higher activity in the fusiform face area than did AD patients. After treatment with donepezil, AD patients showed fusiform activation similar to that of the controls. Exposure to the ChE-I galantamine produced increases in cortical activation during face recognition in patients with mild cognitive impairment. The available data on the effects of cholinergic treatment on the blood oxygen level-dependent response in fMRI are of limited value because there are no studies using a double-blind, placebo-controlled design.

MRI Studies to Detect Atrophy of Cholinergic Nuclei

Hanyu and colleagues were the first to study atrophy of the substantia innominata as a surrogate measure of atrophy in nucleus basalis Meynert. They measured the thickness of the substantia innominata at its smallest extent in a coronal section through the anterior commissure. They found significant decline of the thickness of the substantia innominata in AD

patients compared with controls. Moreover, they found that the thickness of the substantia innominata was inversely correlated with the clinical response to treatment with the ChE-I donepezil. Studying different types of dementia, they found that the reduction of thickness of the substantia innominata was more pronounced in patients with Lewy body dementia than in patients with AD and more pronounced in AD than in vascular dementia. The conclusiveness of the data, however, was limited because the two-dimensional measure of thickness was not accurately located within the substantia innominata and provided only a very rough estimate of atrophy. Recently an automated technique was proposed to determine signal changes within the substantia innominata on the basis of a proton density weighted scan through the basal forebrain. All MRI scans were normalized to a common standard space. Then a three-dimensional region of interest covering the entire extent of the substantia innominata was extracted. Differences in signal intensity across the entire substantia innominata were determined by means of voxelwise statistics. As the locations of nucleus basalis Meynert have not been mapped in MRI standard space before, postmortem MRI and histological sections of the brain of a nondemented older adult were used to transfer the locations of the nuclei of the nucleus basalis Meynert into standard space. These were used as anatomical reference for the signal changes in the substantia innominata from the *in vivo* data. Significant signal reduction was found in anterior lateral and intermediate nuclei of the nucleus basalis Meynert in AD. Moreover, signal reductions were correlated with regional distribution of cortical gray matter loss in AD, suggesting a somatotopic organization of cholinergic projections from the basal forebrain. Recently, a *z*-score approach was developed to determine signal reductions in single participants compared with a reference group of healthy controls. Significant reductions of signal intensity in the substantia innominata similar to those observed in the independent group comparisons were found.

Diffusion tensor imaging (DTI), another MRI modality, may become interesting in the near future for the detection of cholinergic abnormalities. It can be applied to determine the integrity of subcortical fiber tracts *in vivo*. The trajectories of cholinergic pathways in the human brain have been described. Using the Cholinergic Pathways Hyperintensities Scale with T2-weighted MRI scans, Behl and colleagues showed a better response to cholinergic treatment in those AD patients who had more lesions along cholinergic pathways compared with the patients with low cholinergic lesion load, even though overall load in white matter lesions was comparable between groups. It may be of interest in future studies to apply DTI together with basal forebrain scans to determine the selective loss of cholinergic projections from the basal forebrain to the cerebral cortex in AD. AD may serve as an ideal lesion model to study the neuroanatomy and neuropathology of the human cholinergic system. The advances in mapping and understanding alterations of the cholinergic system in AD can then be applied to other human diseases involving the cholinergic system and serve as target for animal models of cholinergic dysfunction.

Imaging the Cholinergic System *In Vivo*: Future Perspectives

What will be the impact of all these techniques to describe cholinergic changes in AD and aging about 3 years ahead? To address this question three potential fields of application should be delineated: (1) clinical routine diagnostic and treatment, (2) clinical trials to establish new treatments (phase II trials), and (3) clinical neuroscience.

It may well be that none of the techniques described in this article will have a direct impact on clinical routine diagnostic or clinical evaluation of treatment efficacy in AD within the next 3 years. Before it is finally applied in clinical routine, a new diagnostic marker goes through three phases:

1. Establishment of the technical characteristics of the marker: reproducibility, effort of acquisition and analysis, analytical sensitivity, and specificity.
2. Establishment of the validity in selected samples: sensitivity and specificity.
3. Controlled diagnostic trial: application in an intent-to-diagnose design, in which patients are selected according to the indication of the diagnostic measure, not the presence or absence of the disease, so that positive and negative predictive values can be determined.

Most of the techniques to determine the functional or morphological integrity of the cholinergic system *in vivo* are presently in phase I: cholinergic receptor PET, fMRI of cholinergic treatment, and structural MRI of the cholinergic nuclei. Only FDG-PET in combination with cholinergic treatment is in phase II, with first clinical trials having applied PET as secondary outcome parameter, but these appear to be all small-scale monocenter studies in highly selected samples. Up to now, no multicenter phase III trial using PET has been undertaken to map the cholinergic system as diagnostic marker or secondary outcome parameter of treatment. One reason for this may be the high costs and limited availability even of FDG-PET for large-scale multicenter clinical trials.

The already completed European Network for Efficiency and Standardisation of Dementia Diagnosis (NEST-DD) and the recently launched US-based AD Neuroimaging Initiative (ADNI) both involve FDG-PET, although not in combination with cholinergic markers or treatment in a prospective design. Retrospective analysis of a subset of the data of the NEST-DD or ADNI may give some insight into the multicenter feasibility of this approach and the effect of cholinergic treatment on cortical metabolism in FDG-PET following a naturalistic design. There are plans now for multicenter molecular imaging studies that include receptor markers of the cholinergic system. Again, these studies will have to involve highly selected samples and therefore will not provide estimates of positive and negative predictive values of these techniques in a clinical population. Another reason these techniques will not likely be applied in clinical routine diagnostics in the near future is the limited budget of national health systems worldwide. From a socioeconomic point of view, the presently available treatment options for AD do not clearly justify the application of a very expensive methodology to give treatment to patients earlier. However, once these techniques have proven their potential to identify those patients who will have maximum benefit from available treatments or once disease-modifying treatments have become available, this situation will change. The costs of these techniques then will be paid off by the early application of disease-modifying treatments that decrease the costs of dementia care.

For the next years, the development of these techniques will probably have some impact on the development of new drugs. Presently, MRI-based volumetric markers are used as secondary outcome parameters in clinical trials on potential disease-modifying effects in AD. The regulatory authorities have indicated that they might accept biological markers as primary endpoints under certain conditions. It has to be shown that the biological marker is related to a clinically relevant outcome, reflects the supposed mechanism of action of the drug, and has high test–retest reliability. At some point receptor PET and imaging of cholinergic nuclei may be part of clinical trials to identify at-risk people and as a secondary endpoint for treatment. Receptor-PET will additionally play an important role in proof-of-concept studies. Receptor PET has already been applied in testing of new compounds and studying the mechanism of action of drugs, such as determining dopamine release and dopamine receptor antagonism of typical and atypical antipsychotics. Imaging cholinesterase inhibition, dynamic changes of nicotinic and muscarinic binding, and the functional consequence of cholinergic changes on brain reserve capacity will help evaluate the mechanism of action of presently available treatments, such as *N*-methyl-D-aspartate antagonists and ChE inhibitors. They may also help estimate the potential of new compounds to modulate the cholinergic system in the human brain *in vivo* before they enter into phase II and III trials.

The most important impact of these techniques in coming years will be their contribution to understanding of the organization and function of the cholinergic system in the human brain. Aging *per se* is associated with a decline of cholinergic markers in the brain. The total number of nucleus basalis Meynert neurons has been shown in the ninth decade to be 20–30% below that in newborns. If the reduction of neuron numbers is approximately linear with age, these data suggest that reduction would be about 3% per decade. However, within a narrow age range in adults (50–80 years), several studies could not detect significant reductions of neuron numbers, suggesting that the 30% reduction from newborns to centenarians may at least partly reflect developmental changes in childhood and adolescence rather than effects of aging. These findings of a limited reduction of cholinergic projections would fit well with the observation that, in contrast to a widely held belief, aging is not inevitably associated with dementia. It would be of high clinical relevance to confirm the postmortem data by a longitudinal imaging approach applying structural and functional markers of cholinergic projections and receptor availability. This would indicate that normal aging preserves a large enough capacity of the cholinergic system to help the brain adapt to environmental and intrinsic factors under normal conditions. In contrast, AD serves as a paradigm that stresses the cholinergic system and allows inferences about its normal function. The focus of research in the coming years will likely increasingly include the role of the cholinergic system in the maintenance of brain plasticity. Research will focus on one hand on the ability of the brain in healthy aging to respond to external and internal stimuli based on the preservation of a minimum of cholinergic function, and on the other hand on the failure of brain plasticity in the presence of severe cholinergic lesions such as in AD. A better understanding of the molecular basis of brain plasticity will help science not only understand the functional basis of learning in its broadest sense (adaption to a constantly changing environment) but also develop treatment paradigms that do not focus only on lowering the impact of pathological events but on increasing the ability of the brain to adapt to pathological changes.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease:

Neurodegeneration; Alzheimer's Disease: An Overview; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Cholinergic System; Cholinergic Pathways in CNS; Cognition in Aging and Age-Related Disease; Functional Neuroimaging Studies of Aging; Neutrotransmission and Neuromodulation: Acetylcholine.

Further Reading

Arendt T, Bigl V, Arendt A, and Tennstedt A (1983) Loss of neurons in the nucleus basalis of Meynert in Alzheimer's disease, paralysis agitans and Korsakoff's disease. *Acta Neuropathologica* 61: 101–108.

Cullen KM and Halliday GM (1998) Neurofibrillary degeneration and cell loss in the nucleus basalis in comparison to cortical Alzheimer pathology. *Neurobiology of Aging* 19: 297–306.

Eckelman WC (2006) Imaging of muscarinic receptors in the central nervous system. *Current Pharmaceutical Design* 12: 3901–3913.

Gilmor ML, Erickson JD, Varoqui H, et al. (1999) Preservation of nucleus basalis neurons containing choline acetyltransferase and the vesicular acetylcholine transporter in the elderly with mild cognitive impairment and early Alzheimer's disease. *Journal of Comparative Neurology* 411: 693–704.

Heinsen H, Hampel H, and Teipel SJ (2006) Response to Boban et al: Computer-assisted 3D reconstruction of the nucleus basalis complex, including the nucleus subputaminalis (Letter to the editor). *Brain* 129(4): E43.

Holcomb HH, Links J, Smith C, and Wong D (1989) Positron emission tomography: Measuring the metabolic and neurochemical characteristics of the living human nervous system. In: Andreasen N (ed.) *Brain Imaging: Applications in Psychiatry*, pp. 235–370. Washington, DC: American Psychiatric Press.

Iyo M, Namba H, Fukushi K, et al. (1997) Measurement of acetylcholinesterase by positron emission tomography in the brains of healthy controls and patients with Alzheimer's disease. *Lancet* 349: 1805–1809.

Mesulam M (2004) The cholinergic lesion of Alzheimer's disease: Pivotal factor or side show? *Learning and Memory* 11: 43–49.

Mesulam MM, Mufson EJ, Levey AI, and Wainer BH (1983) Cholinergic innervation of cortex by the basal forebrain: Cytochemistry and cortical connections of the septal area, diagonal band nuclei, nucleus basalis (substantia innominata), and hypothalamus in the rhesus monkey. *Journal of Comparative Neurology* 214: 170–197.

Nybäck H, Halldin C, Åhlin A, Curvall M, and Eriksson L (1994) PET studies of the uptake of (S)- and (R)-11C-nicotine in the human brain: Difficulties in visualizing specific receptor binding in vivo. *Psychopharmacology* 115: 31–36.

Potkin SG, Anand R, Fleming K, et al. (2001) Brain metabolic and clinical effects of rivastigmine in Alzheimer's disease. *International Journal of Neuropsychopharmacology* 4: 223–230.

Selden NR, Gitelman DR, Salamon-Murayama N, Parrish TB, and Mesulam MM (1998) Trajectories of cholinergic pathways within the cerebral hemispheres of the human brain. *Brain* 121: 2249–2257.

Teipel SJ, Drzezga A, Bartenstein P, Moller HJ, Schwaiger M, and Hampel H (2006) Effects of donepezil on cortical metabolic response to activation during (18)FDG-PET in Alzheimer's disease: A double-blind cross-over trial. *Psychopharmacology* (*Berl*) 187: 86–94.

Teipel SJ, Flatz WH, Heinsen H, et al. (2005) Measurement of basal forebrain atrophy in Alzheimer's disease using MRI. *Brain* 128: 2626–2644.

Chromaffin Cells: Model Cells for Neuronal Cell Biology

B G Livett, University of Melbourne, Melbourne, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Techniques for Studying Chromaffin Cells

Functional studies on chromaffin cells were initially carried out *in vitro* (with cat and bovine adrenal glands perfused retrogradely through the central lobular vein) or *in vivo* (with anesthetized rats, cats, or dogs cannulated via the adrenal vein, the inferior vena cava, or the dorsal aorta). More recently, studies have been carried out in cell culture on adrenal chromaffin cells isolated in large numbers by collagenase treatment of intact bovine adrenal glands (Figure 1) and maintained as primary monolayer cultures or as suspension cultures. The accessibility of populations and single chromaffin cells in culture, their spherical shape, and the fact that they are nontransformed cells have made them a popular experimental system for addressing a number of important questions in cell biology.

Technical advances now make it possible to study the release of catecholamines from a single chromaffin cell and even from a single chromaffin granule in real time, using either electrophysiological measurements of membrane capacitance changes, electrochemical measurements of the released catecholamine made with carbon fiber microelectrodes, or fluorescence microscopy of chromaffin cells transfected with fluorescent proteins targeted to the secretory granules. Advances in transfection techniques now permit expression of exogenous proteins in a high percentage of chromaffin cells in primary culture. For example, Semliki Forest Virus has been used to transduce cDNAs encoding various proteins fused to enhanced green fluorescent protein (EGFP) into cultured bovine adrenal cells. Pro-ANF–EGFP, the virally transduced fusion protein generated with atrial natriuretic factor (ANF), is specifically targeted to large dense-core vesicles as shown by its co-localization with acidophilic dyes and chromogranin A, making this a useful system for the study of secretory vesicle dynamics. Expression of aequorin Ca^{2+}-sensitive photoproteins with tailored Ca^{2+} affinities and targeted to subcellular compartments (mitochondria, endoplasmic reticulum (ER)) allows Ca^{2+} signaling to be investigated in cellular regions other than the cytosol. Physical separation of noradrenaline and adrenaline chromaffin cell types by differential centrifugation in density gradients and permeabilization of chromaffin cells with digitonin or with *Staphylococcus aureus* alpha toxin to gain direct access to the exocytotic machinery now make it possible to investigate and manipulate the cellular events involved in stimulation–secretion coupling. Together, these sophisticated techniques permit a detailed investigation of the second messengers and signal transduction events that mediate receptor responses in well-defined populations of normal (i.e., nonneoplastic) cells. Consequently, adrenal medullary chromaffin cells occupy an important place in neuroscience research by providing a primary source of normal excitable cells for cell biological studies.

Synthesis and Storage of Catecholamines in Adrenal Chromaffin Cells and Chromaffin Granules

The adrenal medullary catecholamines are low-molecular-weight dihydroxylated phenolic compounds synthesized from the dietary aromatic amino acids phenylalanine and tyrosine by the sequential actions of cytoplasmic and intravesicular enzymes (Figure 2). The physiologically important catecholamines include epinephrine (adrenaline), norepinephrine (noradrenaline), and dopamine. The relative amounts of epinephrine and norepinephrine differ considerably among mammalian species, with the proportion of epinephrine to norepinephrine increasing after birth. The adult human medulla contains large amounts of epinephrine (0.2–0.8 $mg\,g^{-1}$; 80%) and smaller amounts of norepinephrine (0.04–0.16 $mg\,g^{-1}$; 19%) and dopamine ($<0.01\,mg\,g^{-1}$; 1%). The two principal catecholamines, epinephrine and norepinephrine, are stored in separate chromaffin cell types within large vesicles ($\sim$0.5 μm in diameter), which are commonly referred to as chromaffin granules (Figure 3). The concentration of catecholamines within the chromaffin granules is so high ($\sim$0.6 M, or twice the osmolarity of body fluids) that it is unlikely that they occur in free solution. Rather, there is now good evidence for an intragranular storage complex consisting of the catecholamines complexed with adenosine triphosphate (ATP; $\sim$0.2 M), acidic proteins (chromogranins; pI $\sim$ 5.5), and other secretory products such as enkephalins (Figure 4). In addition, the enzyme that synthesizes norepinephrine from dopamine (dopamine β-hydroxylase; DBH) and the proteolytic enzymes that process proenkephalin to metenkephalin and leuenkephalin and congeners are associated with the chromaffin granules. Actin and α-actinin are associated with the outside (cytoplasmic side) of the chromaffin granule membrane and are thought to interact with cytoskeletal elements to facilitate granule movement and fusion. The resulting release of granule contents into the

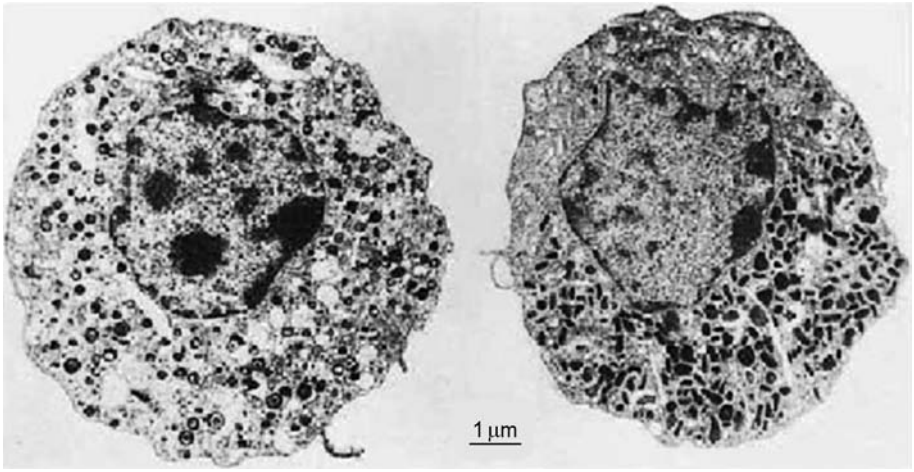

Figure 1 Ultrastructural features of isolated adult bovine adrenal chromaffin cells. Two cell types are present: norepinephrine cells (left) containing dense-core vesicles with electron-dense core contracted away from the vesicle membrane (targetlike appearance), and epinephrine cells (right) with evenly filled pleomorphic vesicles. The epinephrine cell types predominate, in agreement with biochemical analyses of these cell preparations showing an epinephrine/norepinephrine ratio of approximately 3:1. From Livett BG (1984) The secretory process in adrenal chromaffin cells. In: Cantin M (ed.) *Cell Biology of the Secretory Process,* pp. 309–358. Basel: Karger.

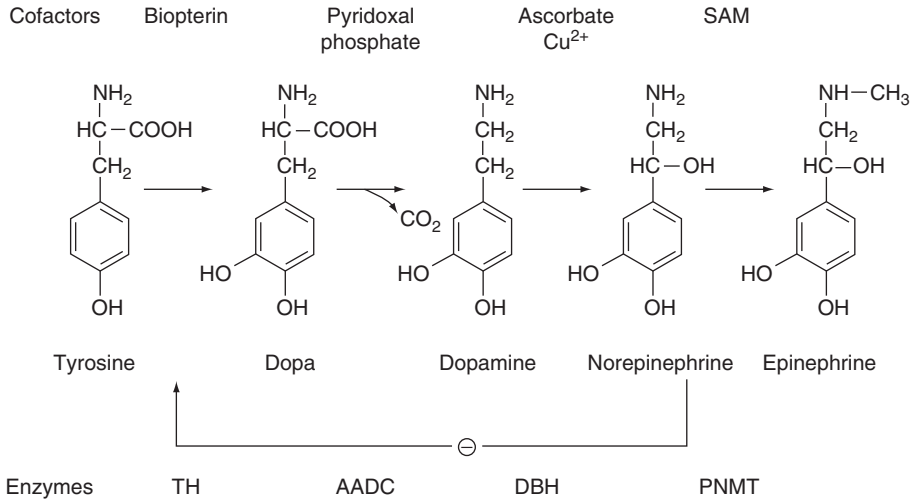

Figure 2 Pathway for the biosynthesis of catecholamines. The pathway from tyrosine to norepinephrine and epinephrine is shown together with the enzymes and cofactors for each step. The feedback inhibition by norepinephrine on tyrosine hydroxylase (TH) is indicated by the arrow. AADC, aromatic amino acid decarboxylase (sometimes called dopa decarboxylase); DBH, dopamine β-hydroxylase; PNMT, phenylethanolamine-*N*-methyltransferase; SAM, *S*-adenosylmethionine.

capillary lumen occurs by a Ca^{2+}-dependent exocytotic mechanism.

The catecholamines are concentrated within the chromaffin granule by a chemiosmotic mechanism (Figure 5) in which a membrane-bound ATPase drives protons into the chromaffin vesicle, thereby creating an electrochemical proton gradient across the chromaffin vesicle membrane. The chemical part (ΔpH) of the gradient supplemented by the electrical potential ($\Delta\psi$) drives the uptake of catecholamines via a specific carrier molecule (the 'catecholamine carrier'). The second component to the electrochemical gradient – the electrical potential – drives the uptake of nucleotides (e.g., ATP) via a separate carrier. The two chromaffin vesicle membrane proteins involved in this process of amine transport, the vacuolar H^+-ATPase (V_o-ATPase) and the catecholamine transporter, have been isolated and their genes cloned and sequenced. The V_o-ATPase is important in providing the energy for amine accumulation in the vesicles, whereas the amine transporter is thought to provide the selectivity for amine uptake. Of interest, the amine transporter in chromaffin cells shows homology to a class of bacterial drug-resistance transporters having 12 transmembrane domains. Like the multidrug-resistance protein from *Bacillus*, the vesicle amine transporter in chromaffin granule membranes binds reserpine and confers cell resistance to the neurotoxin 6-methyl-4-phenylpyridinium (MPP^+)

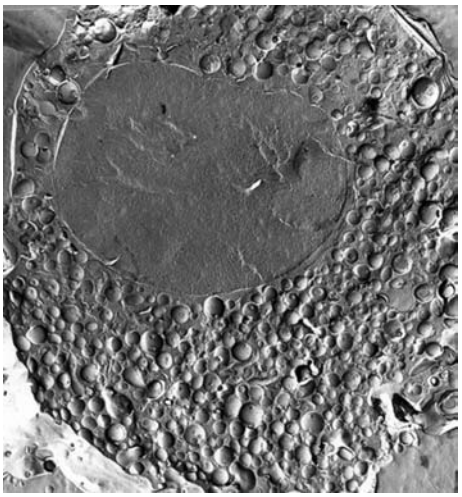

Figure 3 Chromaffin vesicles. The intracellular storage granules containing epinephrine or norepinephrine and a variety of proteins and peptides. The subcellular features were highlighted by the freeze-structure technique. An isolated chromaffin cell was quick-frozen and then fractured. Platinum and carbon, sprayed on the fractured surface, formed a replica of the surface relief; the tissue was then dissolved and a transmission electron micrograph was made of the replica. Micrograph made by Wolfgang Schmidt of the University of Innsbruck; provided courtesy of Professor Stephen Carmichael, Department of Anatomy, Mayo Clinic, Rochester, from the online article (see http://webpages.ull.es/users/isccb12/ChromaffinCell/Primer.html) by Stephen Carmichael and Hans Winkler, 'A Primer on Chromaffin Cells,' based on their article in *Scientific American* (1985) 253: 40–49.

by diverting the uptake of this toxin away from the cellular mitochondria and into the chromaffin granules.

Control of Catecholamine Biosynthesis

In vivo, the synthesis of catecholamines within the adrenal chromaffin cells is subject to neural and hormonal regulation. Tyrosine hydroxylase (TH) is the rate-limiting enzyme, and is known to be controlled acutely (minutes) by phosphorylation and chronically (days) by protein synthesis. In humans and animals exposed to stress (e.g., cold, burns, asphyxia, isolation, immobilization, physical exercise, or childbirth), the rate of synthesis of the biosynthetic enzymes and of the catecholamines increases markedly. Short-term stress initially increases the activity of TH by causing it to be phosphorylated in its N-terminal regulatory domain on several serine residues by a number of different Ser/Thr kinases. These include Ca^{2+}/calmodulin-dependent protein kinase II, cyclic adenosine monophosphate (cAMP)-dependent protein kinase A, protein kinase C, two extracellular signal-regulated kinases (ERK-1 and ERK-2), and a number of related kinases. Nicotine, acting via nicotinic receptors, can sustain the phosphorylation of TH at Ser40 for up to 48 h. Sustained phosphorylation and activation of TH are also observed with histamine and angiotensin II. Sustained phosphorylation of TH at Ser40 provides a novel mechanism for increasing TH activity and leads to increased catecholamine synthesis.

In humans, four splice variants of TH are expressed, with different amino acid insertions immediately prior to one of the key phosphorylated Ser residues that regulate TH activity. TH phosphorylation is increased following stimulation of chromaffin cells with a wide variety of agents, including acetylcholine (ACh), nicotine, K^+, histamine, angiotensin II, pituitary adenylate cyclase-activating peptide-27, caffeine, and nerve growth factor. The phosphorylation state of TH affects its V_{max}, its affinity (K_m) for its substrate tyrosine and for its cofactor tetrahydrobiopterin, its affinity (K_i) for catecholamines responsible for feedback inhibition of TH, and its pH optimum.

Following longer periods of stress lasting hours or days, the amounts of the enzymes TH and DBH increase. The *de novo* synthesis of TH and DBH is under neuronal control whereas the synthesis of phenylethanolamine *N*-methyltransferase (PNMT) is under hormonal control. PNMT is induced by adrenocorticosteroids and this hormonal effect can be mimicked by dibutyryl cAMP and, in hypophysectomized animals, by adrenocorticotropin (ACTH). Synthesis of enzyme cofactors plays a very important role in the regulation of catecholamine biosynthesis. The normal tissue levels of the biopterin cofactor are below the K_m for TH, so the increased synthesis of biopterin (by a guanosine triphosphate cyclohydrolase) is required for long-term activation of TH where the amount of TH enzyme increases. Synthesis of cofactors also plays a crucial role in determining the levels of DBH and PNMT. These two enzymes are subject to rapid intracellular proteolysis and their respective cofactors protect against this proteolysis. Synthesis of the cofactors (ascorbic acid for DBH and *S*-adenosylmethionine for PNMT) is regulated by glucocorticoids under hypothalamopituitary (ACTH) control. Hypophysectomy decreases the number of molecules of DBH and PNMT by increasing their rate of proteolysis. Low concentrations of angiotensin II (10^{-10}–10^{-7} M) evoke long-term effects on TH and PNMT gene induction via a Ca^{2+}-protein kinase C-mediated pathway, whereas higher concentrations (10^{-6}–10^{-5} M) are more effective in the short term to produce catecholamine secretion. This suggests that the major physiological function of this peptide hormone in the adrenal medulla is the long-term regulation of gene expression, rather than the short-term regulation of catecholamine or enkephalin secretion. Control of catecholamine gene expression

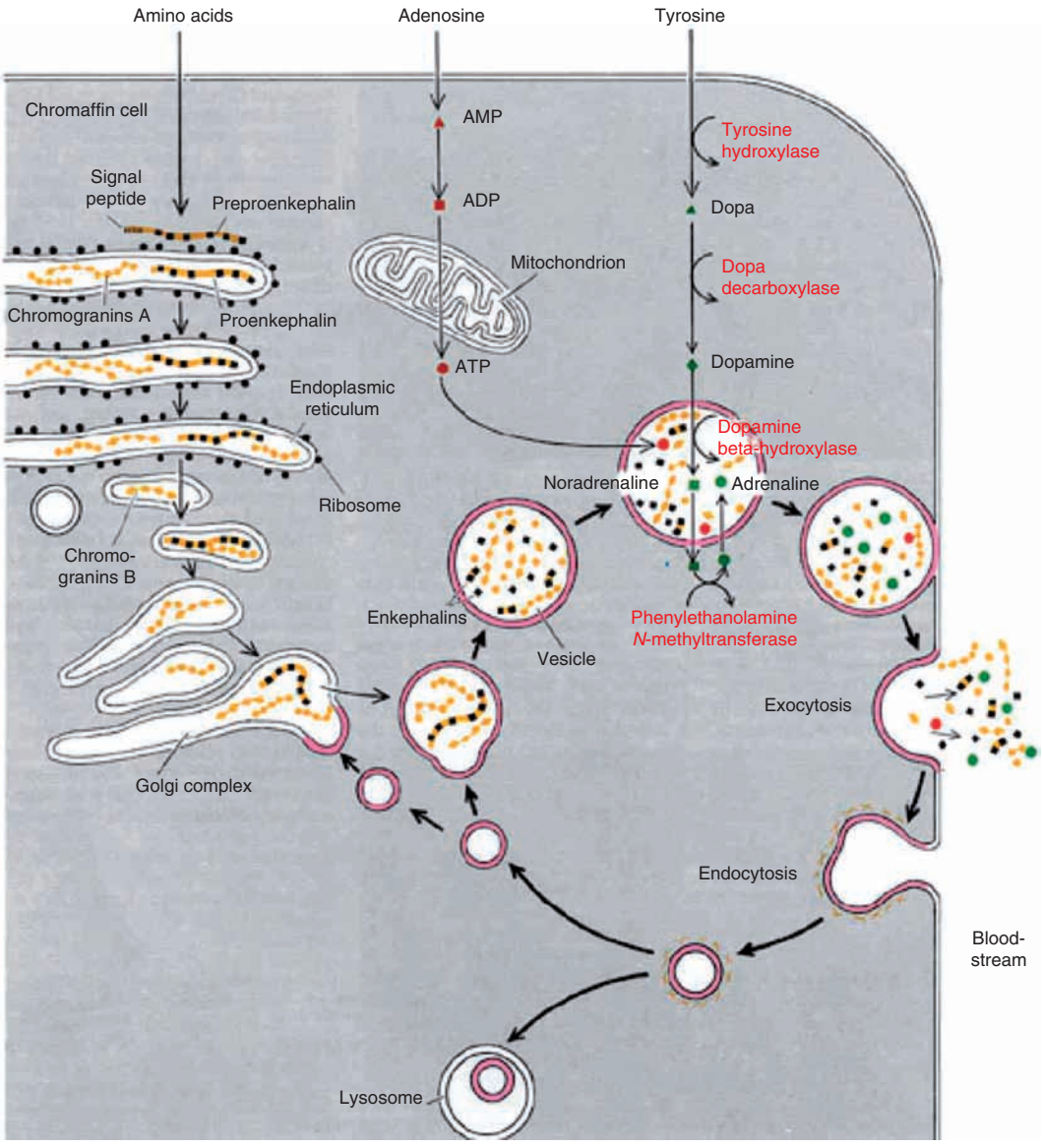

Figure 4 Synthesis and packaging of the contents of chromaffin vesicles. The primary secretary product of most chromaffin cells, adrenaline (epinephrine), is synthesized from tyrosine in four steps, each catalyzed by a different enzyme (see **Figure 2**). Tyrosine is transported into the cell, changed into dopa, and then into dopamine. The dopamine is carried into the chromaffin vesicles and there is transformed into noradrenaline (norepinephrine), which subsequently leaks out of the vesicles into the cytosol and is converted into adrenaline; the adrenaline is pumped back into vesicles for storage. A second substance stored in the vesicles, ATP, is synthesized in mitochondria from ADP, which is derived from adenosine. Amino acids serve as raw material for the manufacture of chromogranins A and B and enkephalin precursors. Proenkephalin, a long polypeptide chain containing enkephalin sequences, is synthesized on ribosomes along the endoplasmic reticulum. As it passes into the reticulum the preproenkephalin loses a short signal peptide; the shortened molecule (now known as proenkephalin) then moves into the Golgi complex and is packaged in newly formed chromaffin vesicles. There it is processed into smaller pieces, including some free enkephalins. The chromogranins are synthesized by a similar process. The various constituents of the chromaffin vesicle (catecholamine, chromogranins, peptides, ATP) are packaged as a multimolecular complex within the vesicle, to reduce their osmotic activity. Upon exocytosis, the vesicle is recycled and reused. Picture courtesy of Professor Stephen Carmichael, Department of Anatomy, Mayo Clinic, Rochester, from the online article (see http://webpages.ull.es/users/isccb12/ChromaffinCell/Primer.html) by Stephen Carmichael and Hans Winkler, 'A Primer on Chromaffin Cells,' based on their article in *Scientific American* (1985) 253: 40–49.

appears to involve several different transcription factors, with the cAMP/cAMP response element (CRE) complex involved in TH expression and the phorbol ester response element (AP-1) being important in mediating PNMT gene expression in response to angiotensin II and nicotine. With their intricate control of TH activity by multiple phosphorylation and of TH gene expression by multiple transcription factors and upstream regulatory domains, chromaffin cells and the related PC12 cells have provided excellent models in which such cell signaling pathways can be investigated and elucidated.

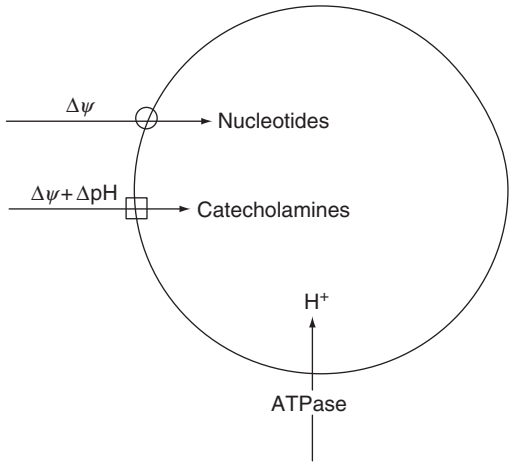

Figure 5 Uptake of amines and ATP into chromaffin granules. The membrane-bound ATPase drives protons into the chromaffin vesicle, creating an electrochemical proton gradient. The chemical part (ΔpH) of the gradient, supplemented by the electrical potential ($\Delta\psi$), drives the uptake of catecholamines by a specific carrier. Only the electrical potential has been shown to drive the uptake of nucleotides via its carrier (O). The specificity of the two carriers has been established by their sensitivity to different inhibitors. From Livett BG (1984) The secretory process in adrenal chromaffin cells. In: Cantin M (ed.) *Cell Biology of the Secretory Process*, pp. 309–358. Basel: Karger.

Control of Catecholamine Secretion

Neuronal

In the adult, control of secretion from the adrenal medulla is mediated through the splanchnic nerve. Classically this innervation has been regarded as purely cholinergic; however, many findings suggest that this may be too restrictive a view of control of chromaffin cell secretion. Studies in the adult rat and dog have provided evidence that stimulation of specific central loci and certain peripheral pathways can be used to achieve independent control over the secretion of epinephrine and norepinephrine. Changes in the proportions of the two amines released can be brought about by merely altering the pattern of electrical stimulation applied to the whole of the efferent innervation to the adrenal in the calf. Different properties of the epinephrine and norepinephrine cells may also contribute to their differential release. Both nicotinic and muscarinic ACh receptors participate in the secretion of catecholamines from adrenal chromaffin cells, but their relative involvement in the release process depends on the concentration of ACh and upon the species being studied. In the rat, cat, dog, and gerbil, nicotinic receptors play a major role in the secretion of adrenal catecholamines, with only a minor contribution from muscarinic receptors. In bovine adrenal chromaffin cells, only nicotinic receptors are linked to secretion, whereas in the chick, muscarinic receptors are principally involved in the secretory response. Significant evidence is now accumulating that noncholinergic cotransmitters in the preganglionic splanchnic nerve fibers also contribute to the control of adrenal medullary function. In particular, the enzyme responsible for nitric oxide synthesis, neuronal nitric oxide synthase, and pituitary adenylate cyclase-activating polypeptide (PACAP) are expressed in preganglionic cholinergic nerves originating in the spinal cord and innervating chromaffin cells in the adrenal medulla. The role of nitric oxide has yet to be fully elucidated, but PACAP is a powerful secretagog and also substantially potentiates catecholamine secretion evoked by cholinergic agonists or by sympathetic nerve stimulation *in vivo*. PACAP also has long-term effects on both secretion and gene expression in chromaffin cells.

Nicotinic Cholinergic Receptors on Chromaffin Cells

Consistent with their neuronal origin, chromaffin cells express neuronal-type nicotinic ACh receptors (nAChRs). Neuronal-type nAChRs are composed of multiple subunits inserted into the plasma membrane and arranged concentrically to form pentameric ACh-gated ion channels. In the central and peripheral nervous system, multiple subunits selected from $\alpha2$–$\alpha9$ and $\beta2$–$\beta4$ form pentameric complexes, either heteromeric (e.g., $\alpha4\beta2$, $\alpha3\beta4$) or homomeric (in the case of $\alpha7$). In the bovine adrenal medulla, the following subsets of neuronal nAChR subunits have been identified: $\alpha3$, $\alpha5$, $\alpha7$, and $\beta4$. It is likely that the functional receptor responsible for catecholamine release from the bovine adrenal medulla contains $\alpha3$ and $\beta4$ subunits, possibly in combination with $\alpha5$, but the precise contribution of the various subunits is not known with certainty. Further, it is not known whether the $\alpha7$ subunit forms a homomeric receptor or heteromeric complex with the other α and β subunits. However, there is currently much interest in $\alpha7$-containing nAChRs because they respond to short pulses of high concentrations of ACh, exhibit a high Ca^{2+} flux, and desensitize rapidly. It is possible they trigger the initial entry of Ca^{2+} into the chromaffin cell in response to the high concentrations of synaptically released ACh, after which further Ca^{2+} entry via $\alpha3\beta4$-containing nAChRs contributes to the observed prolonged release of catecholamines from the adrenal gland.

Peptide Modulation of Secretion

The adrenal medulla has been a rewarding tissue in which to study the functions of neuropeptides. Many neuropeptides, including PACAP, are endogenous components of the splanchnic nerve that innervate the adrenal medulla or of the chromaffin cells, or

both. In some cases (e.g., calcitonin gene-related peptide (CGRP), substance P (SP), galanin, and some of the PACAP), the peptides are present in terminals of sensory nerve collaterals, rather than the preganglionic innervation. Studies with isolated adrenal chromaffin cells in monolayer culture have shown that many of these peptides (notably SP, somatostatin, and certain enkephalin congeners) do not evoke secretion but can modulate the nicotinic secretion of catecholamines. Two binding sites on the β subunit of the neuronal nAChR contribute to the affinity of SP. SP is an inhibitory neuromodulator of the nicotinic response in adrenal chromaffin cells and at the same time protects against nicotinic desensitization. On the basis of these findings *in vitro*, it was proposed that SP may function *in vivo* to maintain a prolonged adrenal catecholamine output during stress. Support for this idea has come from *in vivo* studies on adult rats, pretreated as neonates with capsaicin to kill sensory nerve fibers containing SP in the splanchnic nerve. Thus SP appears to act as a true postsynaptic neuromodulator – inhibiting the release of catecholamines mediated by the release of ACh from the splanchnic nerve under resting conditions while facilitating nicotinic function (through protection against desensitization) under conditions of stress.

The actions of the other neuropeptides are not as well defined, although it is known that, like SP, they are present in the splanchnic nerve or chromaffin cells or both, and many inhibit the nicotinic response. Opioid peptides are expressed in very high concentrations by chromaffin cells, but are also present in their efferent innervation. Their expression in chromaffin cells has led to clinical trials in which encapsulated chromaffin cells are implanted in the subarachnoid space in human cancer patients to alleviate their intractable pain (see later, and see the section 'Further reading'). In addition, opioid peptides (and ATP) released from chromaffin cells have autocrine feedback effects, inhibiting voltage-operated Ca^{2+} channels in the chromaffin cell membrane through actions on G-protein-coupled receptors. Leuenkephalin and metenkephalin are weak inhibitors of nicotinic response, but do enhance the basal release of catecholamines when present in submicromolar concentrations. Many biologically active fragments of the chromaffin granule protein chromogranin have been identified, including catestatin, chromacin, chromostatin, secretoneurin, neuroendocrine secretory protein 55 (NESP55), and vasostatin; they may have actions outside the adrenal, as well as some local modulatory effects. Chromogranin A (CgA), one of the most abundant acidic glycoproteins ubiquitously present in neuroendocrine/endocrine cells, regulates dense-core secretory granule biogenesis. Catestatin (human CHGA352–372) was tenfold more potent than neuropeptide Y at inhibiting the nicotinic secretion of catecholamines from rat PC12 chromaffin cells. Other studies indicate modulatory effects of CGRP, which can potentiate nicotinic responses on rat chromaffin cells. Neurotensin, vasoactive intestinal polypeptide, and vasopressin have all been reported as endogenous peptides within the adrenal medulla, but whether these have a function there or are a developmental consequence and legacy of the common neural crest origin of the adrenal chromaffin cells and other neuronal cell types is not known. Although the detailed mechanisms and physiological importance of these modulatory effects remain to be elucidated, it is clear that the scope for control of secretion extends beyond the original view of splanchnic nerve stimulation and activation of cholinergic receptors on the chromaffin cells.

Nonneuronal

The adrenal medulla of species such as the rat and calf are not functionally innervated by the splanchnic nerve at birth, and nonneurogenic responses have been elicited in response to stresses such as asphyxia. The onset of neural control of adrenal medullary function in the rat first appears at the end of the first postnatal week and matures fully by approximately 10 days of age. A functional role for the endogenous opioids in dampening nonneurogenic release of catecholamines has recently been proposed that may be of survival value to neonates undergoing stress. In addition, glucagon released from the pancreas in response to hypoglycemic stress is a potent secretagog for adrenal medullary catecholamine release. Nonneural influences have been shown to be responsible for the large mobilization and elevation of plasma catecholamines observed 1 h after administration of insulin to adult rats.

Hormonal control of adrenal catecholamine secretion in the adult has been extensively documented. The secretion induced by histamine, angiotensin II, prostaglandins, adenosine, and other factors is dependent on the presence of extracellular Ca^{2+}, as is that evoked by nicotinic agonists; however, unlike nicotinic agonists, these other agents have receptors that are coupled to G-proteins and are not directly depolarizing. The molecular mechanisms by which these noncholinergic receptors induce secretion are hence a major focus of research on chromaffin cells. In many cases, the density of receptors in the adrenal medulla for these noncholinergic agents is so high that this tissue has been used for the isolation and cloning of the receptor (e.g., the histamine H1 receptor). The

G-protein-coupled receptors may act by indirectly depolarizing the chromaffin cells by opening non-selective cation channels or by closing K^+ channels. This depolarization could then recruit voltage-sensitive Ca^{2+} channels. Muscarinic receptors, which are G-protein coupled, contribute substantially to secretion in nonbovine species and act through this mechanism.

Exocytosis and Endocytosis

Much interest has centered around the identification and possible functions of vesicle, cytoplasmic, and plasma membrane proteins thought to be involved in the 'fusion machine' that constitutes exocytosis in the adrenal medulla (Figure 6). The Ca^{2+}-induced exocytosis occurs in two phases: the first phase involves a MgATP-dependent Ca^{2+}-enhanced priming of the secretory apparatus, and the second phase is a Ca^{2+}-dependent and MgATP-independent step that triggers exocytosis. Roles have been proposed for the soluble protein synexin (also termed annexin V[II]) in promoting Ca^{2+}-dependent membrane fusion of chromaffin granules leading to compound exocytosis; however, clearly many other proteins are also involved, including monomeric G-proteins (such as Rab 3A, ARF6, and RhoA), an inhibitory trimeric G-protein (G_o), and cytoplasmic and cytoskeletal proteins such as Exo1, Exo2, synaptobrevin, scinderin, F-actin, munc18, csp, noc2, protein kinase C, myosin, myosin light-chain kinase, and possibly phospholipase D. Trifaro and colleagues have shown that nicotine and depolarizing agents induce disassembly of the cortical actin cytoskeleton in focal areas underneath the plasma membrane, so removing the proposed barrier to chromaffin granule access to the plasma membrane for fusion–exocytosis. Upon stimulation, the uniform cortical distribution of actin and spectrin redistributes into patches. It is believed that voltage-sensitive Ca^{2+} channels at these sites open and permit Ca^{2+}-induced exocytosis, possibly involving phosphorylation of a 20 kDa myosin light chain in the MgATP-dependent priming process preceding exocytosis. Synaptic vesicle proteins such as synaptobrevin and plasma membrane proteins are then proposed to be involved in the so-called fusion engine responsible for vesicle docking to the plasma membrane. The resulting exocytosis brings about hormone secretion, although the details of this mechanism remain to be fully defined.

Membrane capacitance measurements using patch pipettes not only allow exocytosis to be followed in

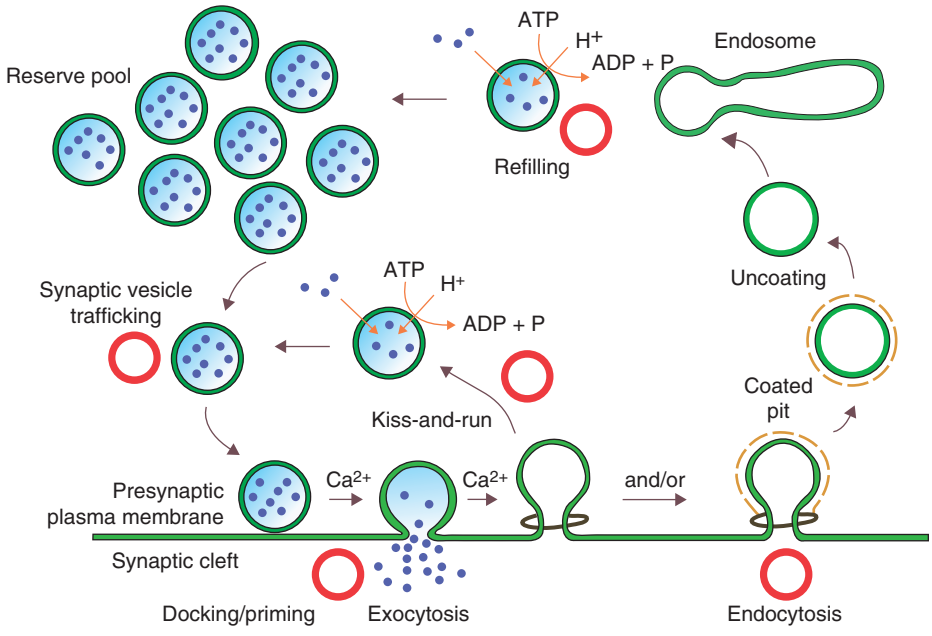

Figure 6 Life cycle of a secretory granule in chromaffin cells and in other secretory cells (neurons, endocrine cells). Step 1: Specific granule membrane and plasma membrane proteins cause the granules to dock with special sites on the plasma membrane. The granules become primed, and then with appropriate stimuli their contents are released ('exocytosis') through a pore created by the fusion of granule and plasma membrane. The fusion pore may only be formed transiently, such that the vesicle is rapidly recovered by endocytosis ('kiss-and-run' exocytosis; step 2). Alternatively, following full fusion, the granule membrane may be recovered by endocytosis through clathrin-coated pits (step 3) and recycled through endosomes (step 4). In both cases, the vesicle is restocked with secretory material and then returned to the secretable pool (step 5). Figure kindly supplied by Dr. Mike Cousin, Membrane Biology Group, Division of Biomedical and Clinical Laboratory Sciences, University of Edinburgh, Scotland.

single cells, but also endocytosis. Such studies on chromaffin cells have demonstrated that the endocytotic retrieval of vesicle membrane following fusion with the cell surface may occur in much less than a minute. In some cases the vesicles do not fully fuse with the membrane but are rapidly retrieved in a process called 'kiss-and-run' exocytosis (Figure 6). Such recycling of vesicles is crucial to permit a constant supply of releasable vesicles.

Ca^{2+} Signaling

The primary signal for exocytosis from chromaffin cells is a substantial increase in cytosolic Ca^{2+} just underneath the plasma membrane. As excitable cells derived from neural crest that are obtainable in large numbers for *in vitro* study, chromaffin cells have been an ideal model system in which to investigate Ca^{2+} signaling. They have been particularly useful for studying calcium channel current autoregulation by materials co-released with catecholamines, such as ATP and opiates. Robust secretion requires concentrations of Ca^{2+} in the subplasmalemmal zone of $>10\,\mu M$. Such concentrations are only effectively achieved by extracellular Ca^{2+} influx through voltage-operated Ca^{2+} channels (VOCCs), the opening of which is synchronized by membrane depolarization. Chromaffin cells posses a variety of types of VOCCs, including in most species L, N, and P/Q channels (5–50%), as well as R and/or T channels in rodents; human adrenal chromaffin cells express mostly P/Q types (60%). Detailed studies have characterized the role of cytosolic Ca^{2+}-buffering proteins on the shape of Ca^{2+} signals in chromaffin cells. By use of pulsed-laser imaging of voltage-clamped cells, it has been shown that depolarization induces Ca^{2+} entry at restricted zones, or 'hot spots,' around the cell surface. Combination of this technique with carbon fiber microelectrode voltammetry has shown these Ca^{2+} hot spots are preferred sites for catecholamine secretion. More recent studies have used chromaffin cells transfected with green fluorescent protein, targeted so as to be expressed as a soluble protein in chromaffin granules. By use of confocal or deconvolution fluorescence microscopy, and total internal reflectance fluorescence microscopy (TIRFM), it is now possible to follow in real time the motion of individual vesicles within the living cell, both deep within the cell and at the specialized release zones around the cell surface, and to observe exocytotic events while manipulating Ca^{2+} levels. These sophisticated methods have contributed significantly to our understanding of the molecular events controlling exocytosis.

Chromaffin cells also possess intracellular Ca^{2+} stores, some of which release their Ca^{2+} in response to the second messenger, inositol (1,4,5)-trisphosphate, and others that are responsive to ryanodine receptor agonists such as caffeine. Different receptors give rise to distinct Ca^{2+} signals in part because they differentially recruit these two stores. Expression of the photoprotein aequorin, engineered so that it is targeted to the ER, has shown directly that Ca^{2+} is released from this organelle by a variety of stimuli and then gets taken up again from the cytosol. Depending on the extent of filling, the ER may act as a sink or source of Ca^{2+} that modulates catecholamine release. Targeted aequorin combined with confocal microscopy has shown that entry of Ca^{2+} through calcium channels can refill the ER to nearly millimolar concentrations. Use of Ca^{2+}-sensitive dyes (e.g., fura-2) has shown substantial interactions between external Ca^{2+} entry and store Ca^{2+} mobilization, such that chromaffin cells can generate complex oscillations in cytosolic $[Ca^{2+}]$. Store Ca^{2+} mobilization is also triggered by Ca^{2+} (a process called Ca^{2+}-induced Ca^{2+} release, or CICR), and this occurs in response to Ca^{2+} release from nearby stores or due to external Ca^{2+} entry through VOCCs. Depletion of store Ca^{2+} causes the influx of external Ca^{2+} entry by a process called capacitative Ca^{2+} entry.

Ca^{2+} removal mechanisms have also been extensively studied in chromaffin cells. These cells possess four principal Ca^{2+} removal mechanisms: plasma membrane Ca^{2+}-ATPases, plasma membrane Na^{+}/Ca^{2+} exchangers (NCXs), smooth ER Ca^{2+}-ATPases, and mitochondrial uniporters. They each play a role in removing cytosolic Ca^{2+} after it has been elevated; however, these mechanisms have different affinities and capacities, as well as different cellular locations. Studies on the mitochondrial uniporter and the NCX indicate that these have significant effects not only on the shape of cellular Ca^{2+} responses, but also on the secretory responses they evoke. Expressing aequorin targeted to mitochondria also confirms that large amounts of Ca^{2+} are taken up into the mitochondria following agonist stimulation and that this limits the extent of exocytosis. Some parameters exhibited by chromaffin cells *in situ*, such as cell polarization and clustering of calcium channels, are not maintained in chromaffin cells in culture, and this is a limitation of isolated cells.

Clinical Uses of Chromaffin Cells in Parkinson's Disease and Chronic Pain

Chromaffin tissue containing catecholamines and enkephalins has been used increasingly over the past decade as a source of cells for transplants in both animal models and human disorders (e.g., Parkinson's disease) in attempts to ameliorate behavioral deficits.

In patients with Parkinson's disease, the use of adrenal medullary tissue transplanted into the brain appears to offer an effective, lasting treatment. Implantation of chromaffin cells which synthesize and secrete a number of pain-reducing neuroactive compounds, including noradrenaline and opioid peptides, has also been proposed for the treatment of intractable cancer pain. Bovine adrenal medullary cells encapsulated into microcapsules of alginate-poly-L-lysine with a liquefied inner core have been introduced into the spinal subarachnoid space in humans to relieve pain. Macro- or microencapsulation of the chromaffin cells within semipermeable membranes is expected to protect the transplant from the immune system of the host. The future of such transplants for the treatment of pain in cancer patients and in patients refractory to morphine remains to be fully evaluated, but is a novel approach with considerable potential.

Bovine chromaffin cells have also provided a convenient functional assay for monitoring the relative efficacy of different synthetic α-conotoxins to inhibit the neuronal-type nicotinic ACh receptor. Two of these, α-conotoxins Vc1.1 and RgIA, recently shown to be potent analgesics in rat models of human neuropathic pain, are selective inhibitors of the α9α10 neuronal nicotinic receptor. Conotoxin Vc1.1 is presently undergoing clinical trials for the treatment of chronic neuropathic pain. It remains to be determined if the α9α10 neuronal nicotinic receptor is present in chromaffin cells and, if so, the extent to which its activation by ACh and nicotine contributes to catecholamine secretion.

See also: Astrocyte: Calcium Signaling; Brain Adrenergic Neurons; Calcium and Signal Transduction; Calcium Channels; Calcium Waves in Glia; Endocytosis and Presynaptic Scaffolds; Exocytosis: Ca^{2+}-Sensitivity; Nicotinic Receptors: Autonomic Neurons; Nicotinic Acetylcholine Receptors; Norepinephrine: CNS Pathways and Neurophysiology.

Further Reading

Aunis D and Langley K (1999) Physiological aspects of exocytosis in chromaffin cells of the adrenal medulla. *Acta Physiology Scandinavia* 167: 89–97.

Borges R and Gandia L (eds.) (2004) *Cell Biology of the Chromaffin Cell. Fundacion Teofilo Hernando, Madrid, © Instituto Teofilo Hernando*, pp. 261–267. LaLaguna and Madrid: Gabinetede publicaciones. http://webpages.ull.es/users/isccb12/TheBook.htm.

Borges R, Machado SK, Alonso C, et al. (2000) Functional role of chromogranins. The intragranular matrix in the last phase of exocytosis. In: Helle KB and Aunis D (eds.) *Chromogranins: Functional and Clinical aspects. advances in Experimental Medicine and Biology,* vol. 482, pp. 3–56. London and New York: Kluwer Academic/Plenum Publishers.

Campos-Caro A, Smillie FI, Dominguez del Toro E, et al. (1997) Neuronal nicotinic acetylcholine receptors on bovine chromaffin cells: Cloning, expression, and genomic organisation of receptor subunits. *Journal of Neurochemistry* 68: 488–497.

Flatmark T (2000) Catecholamine biosynthesis and physiological regulation in neuroendocrine cells. *Acta Physiology Scandinavia* 168: 1–17.

Garcia AG, Garcia-De-Diego AM, Gandia L, et al. (2006) Calcium signaling and exocytosis in adrenal chromaffin cells. *Physiological Reviews* 86: 1093–1131.

Helle KB and Aunis D (eds.) (2000) *Chromogranins: Functional and Clinical Aspects. Advances in Experimental Medicine and Biology,* vol. 482. London and New York: Kluwer Academic/Plenum Publishers.

Hentall ID and Sagen J (2000) The alleviation of pain by cell transplantation. *Progress in Brain Research* 127: 535–550.

Livett BG (1984) Adrenal medullary cells *in vitro*. *Physiological Reviews* 64: 1103–1161.

Livett BG (1984) The secretory process in adrenal chromaffin cells. In: Cantin M (ed.) *Cell Biology of the Secretory Process*, pp. 309–358. Basel: Karger.

Livett BG and Marley PD (1993) Noncholinergic control of adrenal catecholamine secretion. *Journal of Anatomy* 183: 277–289.

Livett BG, Sandall DW, Keays D, et al. (2006) Therapeutic applications of conotoxins that target the neuronal nicotinic acetylcholine receptor. *Toxicon* 48: 810–829.

Marley PD and Livett BG (1985) Neuropeptides in the autonomic nervous system. In: Roses M (ed.) *CRC Critical Reviews in Clinical Neurobiology, Volume 1, No. 3*, pp. 201–283. Cleveland: CRC Press.

Sabban EL (1997) Control of tyrosine hydroxylase gene expression in chromaffin and PC12 cells. *Seminars in Cell and Developmental Biology* 8: 101–111.

Trifaro J-M, Rose SD, Lejen T, et al. (2000) Two pathways control chromaffin cell cortical F-actin dynamics during exocytosis. *Biochemie* 82: 339–352.

Unsicker K, Huber K, Schutz G, et al. (2005) The chromaffin cell and its development. *Neurochemistry Research* 30: 921–925.

Chronic (Repeated) Stress: Consequences, Adaptations

M F Dallman, University of California at San Francisco, San Francisco, CA, USA

© 2009 Published by Elsevier Ltd.

Chronic Stress

What is important for the organism? Evolutionarily, the answers are, first, survival to achieve and sustain reproductive age and, second, reproductive success. The hypothamic–pituitary–adrenocortical (HPA) axis is intimately involved with the former requirement, and its overactivity may inhibit the latter while taking care of living for the moment. Glucocorticoid (GC) secretion is regulated by the HPA axis and is key to survival after severe insults that threaten the organism. GC modulates energy availability, immune responses, learning, and memory, as well as many behaviors. The consequences of GC secretion reveal an organism that is better fit for survival under conditions of challenges. This is clearly so in the face of punctate, acute stressors that, once finished, do not recur on a regular basis. However, with chronic or repeated stressors that cannot be coped with through behavioral changes, the persistently increased secretion of GC may limit the ultimate longevity of the organism. In the case of chronic or persistent inescapable stressors, mechanisms have evolved that may diminish the HPA response to the unrelieved, or repeated stressor input. This is reflected by the process of habituation of the HPA output to a given stressor, well studied in rats by using shock, restraint, or cold.

A careful look at responses to a new stressor in chronically stressed rats that are habituated to a chronic stressor reveals that the new HPA response is most often facilitated compared with that of a stressor-naïve animal. This demonstrates that, although central mechanisms of habituation are engaged, the brain remains aware of the persistent challenge, and the HPA axis overresponds to a new threat. This article first describes the recruitment of limbic networks accompanying chronic or repeated stressors and then the habituation to repeated or chronic stressors, using repeated restraint as a model, and the brain sites and HPA responses important during habituation. It discusses experimental manipulations of these sites and what they may reveal about the plasticity of brain responses in the face of chronic or repeated stressors. Finally, it briefly recaps the upside and discusses the long-term downside of perpetual challenges to the organism – the wear and tear of life.

Chronic Stress and Recruitment of a Limbic Network That Biases Subsequent Autonomic, Behavioral, and Neuroendocrine Responses to Stressors

The HPA axis is not the only motor system that is affected by chronic or repeated inescapable stressors. Chronic stressors bias autonomic responses toward excessive sympathetic outflow; they bias behaviors toward the fearful- and anxiety-like spectrum; they alter neuroendocrine responses so that the HPA responses to novel stressors are facilitated although those to many known stressors are habituated with repetition. In rats, the somatotropic and thyrotropic axes are inhibited by chronic or repeated stressors; the gonadotropic axis may be either inhibited or facilitated, depending on the stressor that is chronically or repeatedly applied. These effects of chronic stressors may be a consequence of the stimulatory actions of GC on corticotropin-releasing factor (CRF) synthesis and secretion in the central nuclei of the amygdala (CeA). From the CeA, CRF axons innervate and may activate all monoaminergic cell groups that have been shown to be important, and facilitated, after the engagement of the chronic stress network. It is the altered activity of the monoaminergic networks, as well as direct effects of CRF from the CeA, that is responsible for many of the altered autonomic, behavioral, and neuroendocrine responses that accompany activation of the emotional limbic brain during persistent, inescapable stressors.

Habituation of HPA Responses with Time after Inescapable Stressors

What are the HPA characteristics of habituation? Because habituation is measured by comparison of the amplitude and duration of the first HPA response to a given stressor with those of subsequent responses, it is far easier to demonstrate that habituation occurs after the use of mild to moderate stressors that do not saturate the response characteristics of the HPA axis. Habituation has typically been studied in responses to repeated restraint, intermittent or chronic cold, footshock, social stressors, and variable chronic stress paradigms. Although some are top-down, processive stressors and others bottom-up, systemic stressors, the HPA responses more often than not habituate with repetition or time. It seems likely that with both top-down and bottom-up cases, the animal becomes adapted, behaviorally and

physiologically, to the repeated or chronic stimuli. Indeed, for restraint, there is evidence, as measured by c-fos expression, that afferent input within the brain is reduced with repetition, whereas with the metabolic challenge of cold, the animal gears up its metabolism and food intake to alter thermogenesis and respond appropriately to the challenge, thus reducing afferent input to the brain. Nonetheless, generally there is evidence that, although habituation may occur with chronicity or repetitions of stressors that are not life-threatening, new central networks have been recruited that allow the animals to hyperrespond to presentation of novel stimuli.

Repeated Restraint

Restraint is probably the most-studied stressor in which the process of habituation has been explored. Restraint, unlike immobilization, simply confines animals in a tube in which they can breathe readily and can move somewhat but in which they cannot turn around. The HPA axis response to an initial bout of restraint is transient. When rats are maintained in a restraining tube for hours, adrenocorticotropin (ACTH) and corticosteroid responses peak at 15–30 min and then return to control concentrations despite continued confinement (Figure 1); red lines show responses to restraint during the first encounter with it). It is likely that a portion of the initial response also reflects the stressor of being picked up and placed into the restraining tube. Nonetheless, the HPA response to the stressor seems clearly damped by the adrenal GC feedback signal since in its absence, ACTH concentrations remain high for the full period of restraint. With chronic, repeated restraint, the amplitude of the HPA response decays over days (Figure 1; blue lines show responses to restraint during the fifth encounter). All efferent components of the response are diminished and curtailed, and this is reflected by the markedly reduced c-fos response in the hypothalamic paraventricular nuclei (PVN) to restraint on the fifth day compared with the first. Moreover, the CRF cells in the PVN that express c-fos after repeated restraint appear to be located more ventromedially, compared to the dorsomedial neuronal CRF response that occurs to the first episode of restraint. All the downstream responses are shortened, and their amplitude is generally diminished. However, it is characteristic of chronic stressors that HPA responses are still observed, but their magnitude and duration decrease.

Why Habituation?

Familiarity with the stimulus, in part through GC-mediated learning, memory, and recall, perhaps reveals that it is not life-threatening, although it may be boring and uncomfortable, and that it can be endured. Habituation to the much more severe stimulus of immobilization, in which rats are tightly restrained and cannot move even their heads, is much slower, although it, too, occurs eventually. Habituation also occurs with other stimuli, such as cold, mild electrical shock, social stress, and subsequent presentations of a previously experienced stimulus in a variable stimulus paradigm (in which different stressors are presented twice daily for up to a month). However, if stimuli are perceived as life-threatening, subsequent repetitions may produce augmented, rather than habituated, responses, as reported for exposures to predator odor and immune challenge. Again, it seems highly appropriate to have mechanisms (that are still unclear) that allow continued, and even increased responses to potentially life-threatening events.

Where in the Brain Does Habituation to Restraint Occur?

Many sites in brain have been examined for their potential role in habituating HPA responses to repeated restraint. With the use of c-fos messenger RNA, it seems clear that there is decreased responsiveness to recurrent restraint throughout the afferent limbic and cortical brain networks involved in transmission of the signal to the CRF motor zone, the hypothalamic PVN, probably as a consequence of cortical inhibition of these sites.

How Does Habituation Occur?

Although cytotoxic lesions of the posterior paraventricular thalamus do not affect the duration or amplitude of an initial ACTH response, they do block habituation of ACTH to repeated bouts of restraint, suggesting that this structure is important for, perhaps, the memory of prior restraint (Figure 2). Moreover, injection of a cocktail of mineralocorticoid receptor and GC receptor (GR) antagonists during each period of repeated restraint also decreases the degree of habituation (Figure 3) although they do not affect the initial HPA response to the stressor. Thus, it appears that cells in the posterior paraventricular thalamus that contain GRs (and presumably mineralocorticoid receptors) and that project to both the basolateral amygdala and the prefrontal cortex strongly affect habituation of HPA responses to restraint, although they do not explain the entire habituation response. Similarly, injection of a GR antagonist into the basolateral amygdala, or into the hippocampus, depending on the type of behavior examined, can block subsequent behavioral responses to a stressor. It seems clear that the GCs aid learning

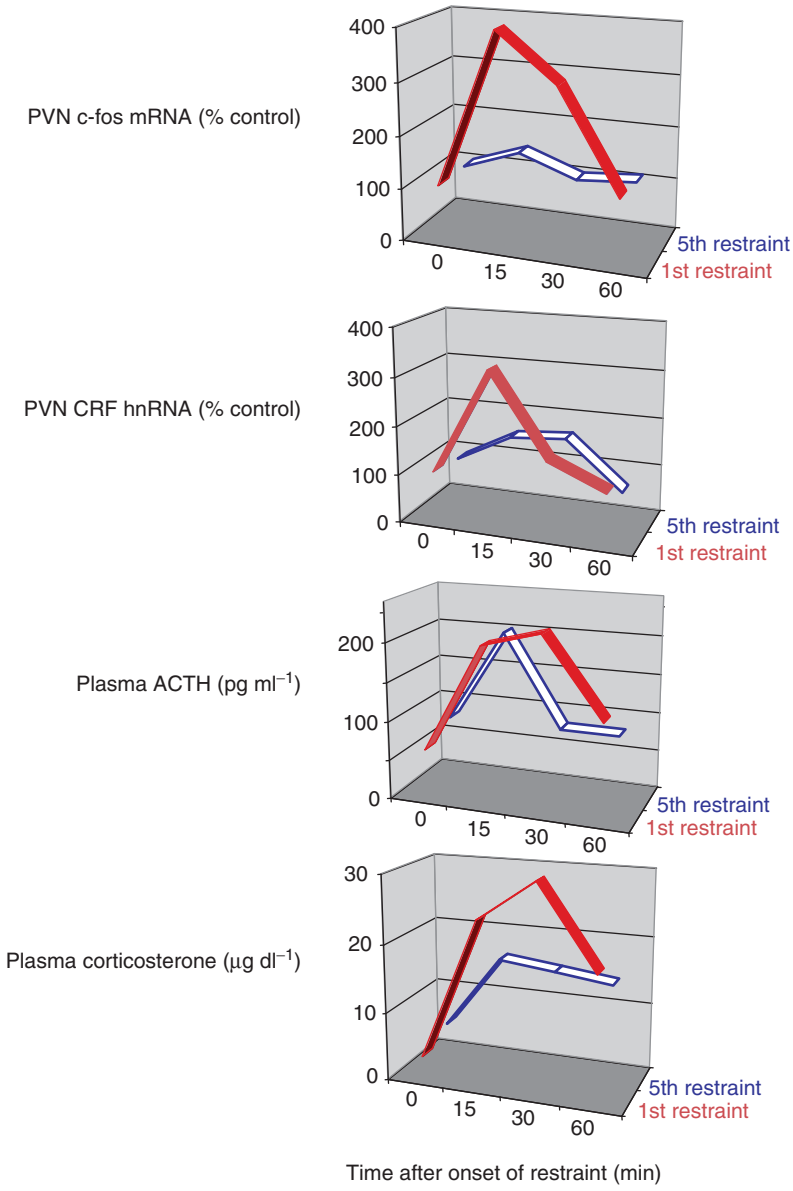

Figure 1 Habituation of HPA responses to restraint: first vs. fifth restraint. ACTH, adrenocorticotropin; CRF, corticotropin-releasing factor; hnRNA, heterogeneous nuclear RNA; mRNA, messenger RNA; PVN, hypothalamic paraventricular nucleus. Redrawn from Girotti M, Pace TWW, Gaylord RI, Rubin BA, Herman JP, and Spencer RL (2006) Habituation to repeated restraint stress is associated with lack of stress-induced c-fos expression in primary sensory processing areas of the rat brain. *Neuroscience* 138: 1067–1081.

of and memory for emotionally laden events and that without the GC action, appropriate responses to the events do not occur.

Facilitation to Novel Stimuli Occurs with Chronic Stress

What Does Facilitation Look Like?

Although HPA responses to repeated stimuli dwindle with repetition, it is characteristic of rats undergoing chronic challenge that they exhibit greater-than-normal HPA responses to new stressors. The example provided in Figure 4 is of rats given intermittent injections of morphine in escalating doses that allow partial withdrawal twice daily, which is an intense stimulus to the HPA axis. However, facilitated responses to novel stressors are characteristic of most chronic challenges and have been observed in rats, mice, sheep, and humans. Facilitated responses are observed at each level (CRF, ACTH, and corticosterone) of the HPA axis in rats. Facilitation of HPA responses to novel challenges, even in an organism undergoing chronic stress with habituated responses,

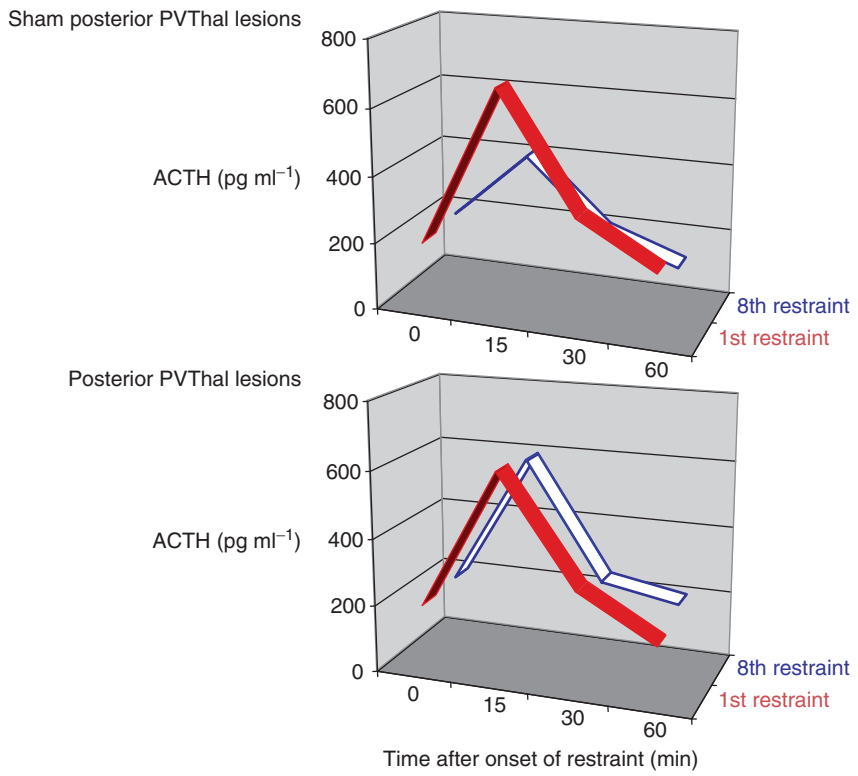

Figure 2 Dishabituation of HPA responses to restraint by lesion of posterior PVThalamus: first vs. eighth restraint. ACTH, adrenocorticotropin; HPA, hypothalamic–pituitary–adrenocortical; PV, paraventricular. Redrawn from Bhatnagar S, Huber R, Nowak N, and Trotter P (2002) Lesions of the paraventricular thalamus block adaptation of hypothalamic–pituitary–adrenal (HPA) responses to repeated restraint. *Journal of Neuroendocrinology* 14: 403–410.

makes a great deal of physiological sense since GC responses of normal magnitude to many stressors are essential for the maintenance of life.

Facilitated Responses Require Sustained or Intermittently Repeated Stressors and Elevated Corticosterone Concentrations

Adrenalectomized rats without GC or with low levels of GC replacement do not exhibit facilitated ACTH responses to novel stimuli under conditions of chronic challenge (Figure 5). Although replacement with GC results in normal, monotonic inhibition of ACTH in chronically stressed, compared with control, rats (Figure 5, left panel), GC concentrations must be maintained at values above the normal daily mean for intact, unstressed rats for facilitated responses to novel stress to be observable (Figure 5, right panel). The corticosterone concentrations that permit facilitation are sufficient to entirely inhibit ACTH responses to novel restraint in control animals that were not chronically challenged. Because GC acts at GRs in the anterior pituitary corticotropes to inhibit both synthesis of proopiomelanocortin, the precursor for ACTH, and secretion of the mature hormone, hypothalamic CRF may increase in response to the acute stimulus, but because of the GC-inhibited corticotrope, ACTH secretion may not occur.

How Does Facilitation Occur?

Facilitation is almost certainly explained by the fact that chronic stressors recruit a limbic network, based on activation of CRF in the central nucleus of the amygdala by GC's acting through GRs. The limbic response network prepares the animal to maintain behavioral, autonomic, and neuroendocrine responsiveness in the face of tonically elevated GC that might otherwise suppress these responses. Facilitated HPA responses to new stimuli are blocked by daily injections with a GR antagonist (RU486) during the period of intermittent challenge. They are also blocked by the injection of a CRF1 receptor antagonist before the novel stimulus is presented. Thus, although the intimate details are still not understood, the stress-induced elevations in GC concentrations promote, not only habituation, but also facilitation, through actions on stressor-recruited limbic structures and on limbic neurons to stimulate CRF synthesis.

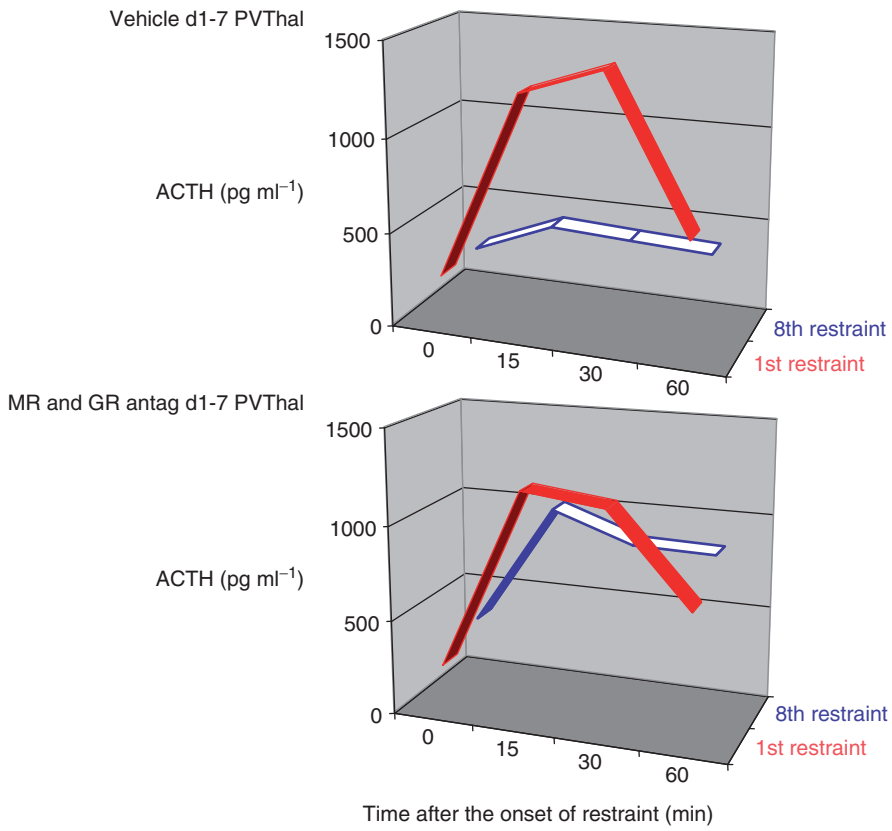

Figure 3 Dishabituation of hypothalamic–pituitary–adrenocortical (HPA) responses to restraint by daily MR and GR antagonists into the posterior PVThalamus: first vs. eighth restraint. ACTH, adrenocorticotropin; GR, glucocorticoid receptor; MR, mineralocorticoid receptor; PV, paraventricular. Redrawn from Jaferi A and Bhatnagar S (2006) Corticosterone can act at the posterior paraventricular thalamus to inhibit hypothalamo–pituitary–adrenal activity in animals that habituate to repeated stress. *Endocrinology* 147: 4917–4930.

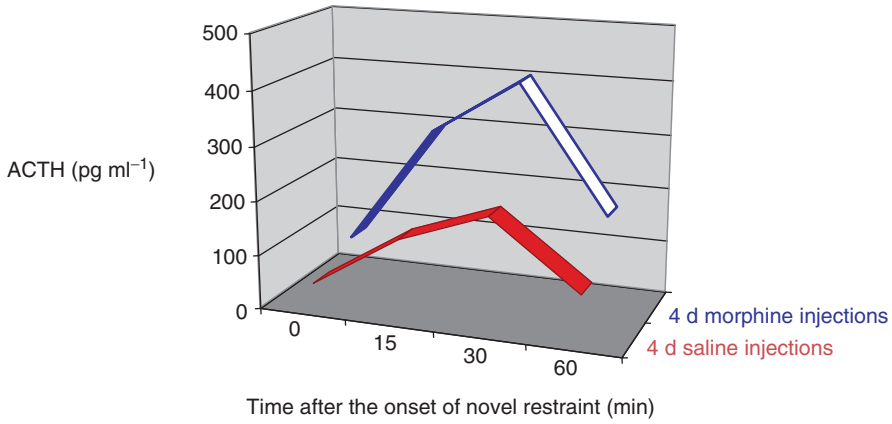

Figure 4 Chronic stress facilitates hypothalamic–pituitary–adrenocortical (HPA) responses to a novel stressor. ACTH, adrenocorticotropin. Redrawn from Houshyar H, Gomez F, Manalo S, Bhargava A, and Dallman MF (2003) Intermittent morphine administration induces dependence and is a chronic stressor in rats. *Neuropsychopharmacology* 28: 1960–1972.

The Up- and Downsides of HPA Responses to Chronic or Repeated Challenge

The Upside

The HPA responses to chronic challenge are essential for maintaining life. Without the GCs, death occurs or metabolism and mentation are not appropriately regeared to meet the challenge. The persistent and overall increases in GC concentrations above basal values that occur during normal physiological responses to chronic stressors sculpt the organism's repertoire to manage the threat flexibly and

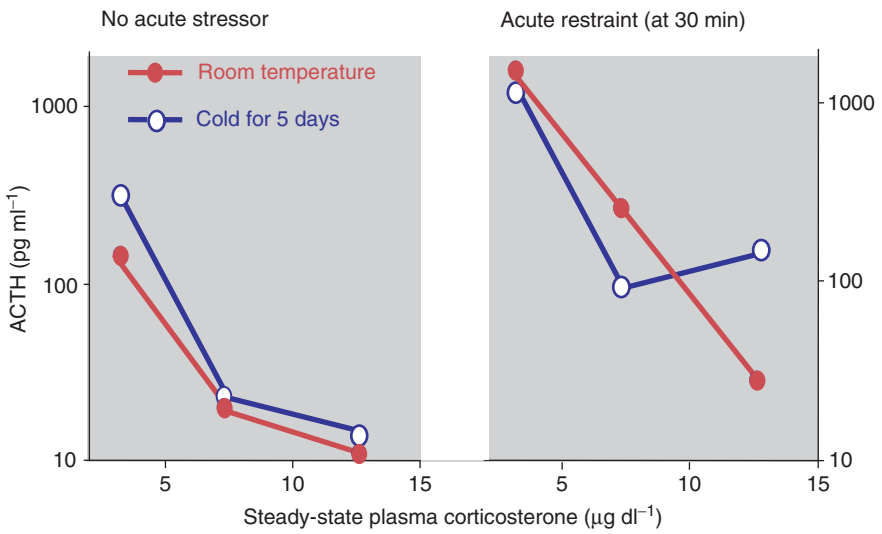

Figure 5 Facilitated ACTH responses to novel restraint require high corticosterone in chronically stressed adrenalectomy (ADX) rats. ACTH, adrenocorticotropin. Redrawn from Akana SF and Dallman MF (1997) Chronic cold in adrenalectomized corticosterone (B)-treated rats: Facilitated corticotropin responses to acute restraint emerge as B increases. *Endocrinology* 138: 3249–3258.

optimally: (1) Metabolic activity is adjusted to produce more gluconeogenesis so that available essential fuel is abundant, drawing on reserves and also rearranging stored metabolic energy toward the center of the animal, where it provides better physical balance for potential escape behaviors. (2) Behaviors are geared more toward fight, flight, anxiety, and ingestive responses, and there is an increased level of arousal that will more accurately pinpoint any new dangers to be assessed. Habituation, mediated at least in part through the GC response, ensures that the HPA response is calibrated and appropriately damped, once the level of the threat is known and evaluated, so that more-than-adequate GC secretion does not occur. Simultaneously, the same challenge-induced GC responses enable full engagement of the stressor-recruited central brain stress-response network, allowing increased responsivity in other central neural systems to ongoing and potential new threats. The HPA response is finely honed to enable and prepare the organism to deal with the ongoing vicissitudes of life.

The Downside

Greatly elevated concentrations of GC, persisting for too long, may exert deleterious effects that will ultimately decrease the life span of the individual. GCs certainly affect learning and memory, and they do so with an inverted U-shaped dose–response relationship. As in the story of Goldilocks and the three bears, the magnitude of the chronic HPA response must be just right, neither too little nor too much. Effects of inappropriately high GC on learning and recall of memory, hippocampal CA3 apical dendritic branching, neurogenesis, and neuronal apoptosis are generally deleterious. As with all things, there is a Janus-like quality to the GCs. They can either augment or inhibit immune responses. During the stressful breeding season, both salmon returning to spawn and the small male Australian marsupial *Antechinus stuartii* die after fertilizing the eggs of the female, and often with evidence of overwhelming infection, probably due to the very high GCs that accompany their efforts. GCs also stimulate not only the learning and memory of high-density palatable calories but also their intake. This is appropriate and plays a large role in seasonal fattening in birds and animals in the wild. However, in industrialized societies, people living under conditions, not only of heightened persistent psychological challenge, but also of inexpensive and readily available, low-quality, high-calorie foods, run into trouble. Stressor-induced increases in palatable food consumption may result in increased abdominal obesity, cardiovascular disease, and type 2 diabetes. All these are known to strongly contribute to morbidity and mortality.

Thus, although they are required for the survival of challenge, the secreted GCs also greatly increase the rate of wear and tear associated with life and may ultimately shorten life span.

See also: Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress and Parasympathetic Control; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress; Sympathoadrenal System: Neural Arm of the Stress Response.

Further Reading

Akana SF and Dallman MF (1997) Chronic cold in adrenalectomized corticosterone (B)-treated rats: Facilitated corticotropin responses to acute restraint emerge as B increases. *Endocrinology* 138: 3249–3258.

Bhatnagar S, Huber R, Nowak N, and Trotter P (2002) Lesions of the paraventricular thalamus block adaptation of hypothalamic–pituitary–adrenal (HPA) responses to repeated restraint. *Journal of Neuroendocrinology* 14: 403–410.

Dallman M, Pecoraro N, and la Fleur S (2005) Chronic stress and comfort foods: Self-medication and abdominal obesity. *Brain, Behavior, and Immunity* 19: 275–282.

Dallman MF, Pecoraro NC, and la Fleur SE (2006) Glucocorticoids, chronic stress, and obesity. *Progress in Brain Research* 153: 75–105.

Dallman MF, Viau V, Bhatnagar S, Gomez F, Laugero K, and Bell ME (2002) Corticotropin-releasing factor (CRF), corticosteroids, stress and sugar: Energy balance the brain and behavior. In: Pfaff DW (ed.) *Hormones, Brain and Behavior,* vol. 1, pp. 571–631. San Diego, CA: Academic Press.

de Kloet ER, Oitzl MS, and Joels M (1999) Stress and cognition: Are corticosteroids good or bad guys? *Trends in Neurosciences* 22: 422–426.

Girotti M, Pace TWW, Gaylord RI, Rubin BA, Herman JP, and Spencer RL (2006) Habituation to repeated restraint stress is associated with lack of stress-induced c-fos expression in primary sensory processing areas of the rat brain. *Neuroscience* 138: 1067–1081.

Herman J and Cullinan W (1997) Neurocircuitry of stress: Central control of the hypothalamo–pituitary–adrenocortical axis. *Trends in Neurosciences* 20: 78–84.

Houshyar H, Gomez F, Manalo S, Bhargava A, and Dallman MF (2003) Intermittent morphine administration induces dependence and is a chronic stressor in rats. *Neuropsychopharmacology* 28: 1960–1972.

Jaferi A and Bhatnagar S (2006) Corticosterone can act at the posterior paraventricular thalamus to inhibit hypothalamo–pituitary–adrenal activity in animals that habituate to repeated stress. *Endocrinology* 147: 4917–4930.

Joels M, Pu Z, Wiegert O, Oitzl MS, and Krugers HJ (2005) Learning under stress: How does it work? *Trends in Cognitive Sciences* 10: 152–158.

Lupien SJ and Lepage M (2001) Stress, memory, and the hippocampus: Can't live with it, can't live without it. *Behavioural Brain Research* 127: 137–158.

McEwen BS (2005) Glucocorticoids, depression, and mood disorders: Structural remodeling in the brain. *Metabolism: Clinical and Experimental* 54(supplement 1): 20–23.

Pecoraro N, Dallman MF, Warne JP, et al. (2006) From Malthus to motive: How the HPA axis engineers the phenotype, yoking needs to wants. *Progress in Neurobiology* 79: 247–340.

Sawchenko PE, Li H, and Ericsson A (2000) Circuits and mechanisms governing hypothalamic responses to stress: A tale of two paradigms. *Progress in Brain Research* 122: 61–78.

Schulkin J, Morgan MA, and Rosen JB (2005) A neuroendocrine mechanism for sustaining fear. *Trends in Neurosciences* 28: 629–635.

Cingulate Cortex

B Y Hayden and M L Platt, Duke University School of Medicine, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Anatomy of Cingulate Cortex

Subdivisions

The mammalian cingulate cortex occupies a vast territory stretching rostrocaudally along the cingulate sulcus on the medial surface of the cerebral hemispheres (Figure 1(a)). In both human and nonhuman primates, the cingulate cortex consists of the cingulate gyrus as well as the cortex lining the superior and inferior banks of the cingulate sulcus. The name 'cingulate,' derived from the Latin word for belt, reflects the fact that the cingulate cortex encircles the corpus callosum. The cingulate cortex can be divided into the more rostral anterior cingulate cortex (ACC) and the more caudal posterior cingulate cortex (PCC). In primates, the ACC consists of Brodmann's areas 24, 25, 32, and 33, while the PCC consists of Brodmann's areas 23, 29, 30, and 31. The clearest distinction between the ACCs and PCCs is that ACC has a thin agranular layer 4, whereas PCC has a pronounced granular layer 4.

The ACC and PCC can be further subdivided. The ACC consists of at least three distinguishable areas – the perigenual anterior cingulate cortex (pACC), the dorsal anterior cingulate cortex (dACC; sometimes referred to as midcingulate cortex, MCC), and the rostral and caudal cingulate motor areas (CMAs). The PCC comprises at least two distinct areas, the more rostral posterior cingulate cortex proper and the more caudal retrosplenial cortex, which constitutes the entire PCC in rodents.

Although there is a great deal of functional overlap between regions of the cingulate cortex, each of the anatomically defined subdivisions serves discrete functions. The pACC and dACC are thought to process emotional and cognitive information, respectively, while the CMAs process motor information. Although response properties of neurons in the CMAs are similar, electrical stimulation in the more caudal areas CMAd/v has a stronger and more consistent effect on movement than stimulation in CMAr. This difference is consistent with the stronger connections to the motor cortex and spinal cord made by neurons in CMAd/v. Response properties of single neurons within the CMAs are similar to those in supplementary motor cortex (SMA), but SMA neurons respond earlier and more phasically than neurons in the CMAs. These findings suggest that SMA initiates movements while the CMAs contribute to their execution. The PCC differs from ACC by virtue of having a closer linkage to processing spatial information and a less direct role in the generation of voluntary action. Within the PCC, the retrosplenial cortex has been implicated in episodic memory formation and consolidation.

Connections

The cingulate cortex has been described as a neural interface between emotion, sensation, and action. This idea is strongly supported by the presence of anatomical connections linking the cingulate cortex with brain areas closely associated with each of these functions. The centrality of motivational and emotional processing to cingulate function is highlighted by its strong reciprocal connections to the reward centers of the brain, including the orbitofrontal cortex (OFC), the basal ganglia, and, in the case of the ACC, the insula (Figure 1(b)). The ACC is a major target of midbrain dopamine neurons, which respond to a variety of reward-related parameters. In addition, the ACC is reciprocally connected with the amygdala, a collection of nuclei that participates in assigning both positive and negative valence to events. Motivational and emotional inputs to ACC may be directly communicated to the PCC via massive reciprocal connections between these two regions.

Both the ACC and PCC are reciprocally connected with the lateral prefrontal cortex (LPFC), a region implicated in executive control, working memory, and rule learning and expression. The strong connections between the cingulate cortex and the LPFC suggest that the cingulate cortex serves as a gateway for incorporating reward-related information into sensorimotor mappings subserved by the LPFC. Furthermore, the ACC sends strong projections to the locus coeruleus (LC), a brain stem nucleus providing widespread noradrenergic inputs to the rest of the brain. Inputs from the LC to the prefrontal cortex may shift the organism between 'exploit' and 'explore' modes of behavior.

The cingulate cortex also shares strong reciprocal connections with motor structures. These include the primary and supplementary motor cortices and the main cortical regions controlling eye movements, the frontal and supplementary eye fields. The PCC is reciprocally connected with several parietal premotor areas, including 7a, 7m (also connected with ACC), and 7/PG. Both PCC and the caudal CMAs send direct projections to the spinal cord, likely pathways

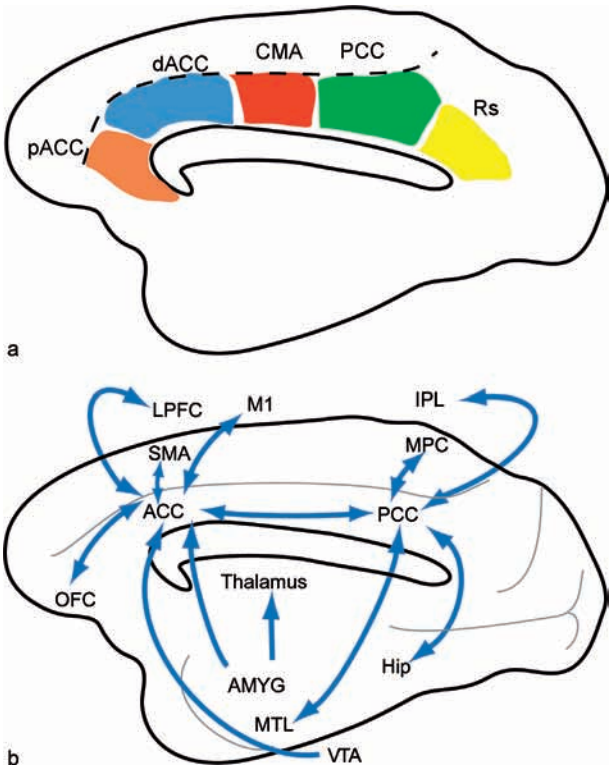

Figure 1 Anatomical subdivisions of the cingulate cortex and adjacent areas of the cortex. (a) The cingulate cortex anterior to the central sulcus is traditionally known as the anterior cingulate cortex (ACC), while the cingulate cortex posterior to the central sulcus is traditionally known as the posterior cingulate cortex (PCC). The ACC is divided into a rostral perigenual ACC (pACC), a central dorsal ACC (dACC), and the caudal motor regions within the sulcus called the cingulate motor areas (CMAs). The posterior cingulate cortex is generally divided into the more rostral posterior cingulate proper (PCC) and more caudal retrosplenial cortex (Rs). (b) Some, but not all, of the major connections between the anterior and posterior cingulate cortices and other areas in the macaque brain. ACC, anterior cingulate cortex; AMYG, amygdala; Hip, hippocampus; IPL, inferior parietal lobule; LPFC, lateral prefrontal cortex; M1, primary motor cortex; MPC, medial parietal cortex; MTL, medial temporal lobe; OFC, orbitofrontal cortex; PCC, posterior cingulate cortex; SMA, supplementary motor area; VTA, ventral tegmental area.

by which the cingulate cortex most directly controls behavior.

The strong connections between PCC and the parietal cortex suggest that it may play a central role in orienting attention. In fact, the parietal, cingulate, and frontal cortices are often thought to comprise a network specialized for attentional control. Indeed, damage to these areas leads to neglect, a selective deficit in orienting attention. Recent neuroimaging work has suggested that the PCC may participate in selecting locations for enhanced attention while ACC plays a more general, nonspatial role in controlling attention. This idea is supported by the finding that blood flow to PCC increases when spatial attention is shifted contralaterally, whereas blood flow to ACC increases in a nonselective fashion. Cingulate cortex is unique among attentional control areas by virtue of its strong connections to limbic regions, which are thought to be important for reward processing. These connections may serve as a gateway by which reward-related information is integrated with executive processes controlling the allocation of attention.

A final set of important connections is that linking cingulate cortex to structures that form long-term memories. The PCC makes strong reciprocal connections with the hippocampus and the medial wall of the temporal lobe, both of which are associated with consolidation of long-term memories. The ACC shares strong connections with the amygdala, which processes information about emotionally relevant stimuli. Both sets of connections provide a potential mechanism for enhancing long-term memories of motivationally significant events.

Each region of the cingulate cortex is most strongly connected, both anatomically and functionally, to adjacent regions of the cerebral cortex. Thus, pACC, involved in emotional aspects of cognition, is located adjacent to the OFC, which is thought to be the primary cortical site for emotional and motivational information processing. The dACC, concerned with more abstract cognitive processes, is located adjacent to the LPFC and pre-SMA, which are thought to participate in rule representations, working memory, and other abstract cognitive processes. The cingulate motor areas, the most directly motor of the cingulate areas, are adjacent to the primary and supplementary motor areas of the frontal lobe. Finally, the posterior cingulate and retrosplenial cortex, which link rewards with locations in space, are adjacent to the parietal and parahippocampal regions, the chief cortical sites for storing and manipulating spatial representations.

Role of Cingulate Cortex in Behavior

Early Ideas of Cingulate Function

As with many other areas of the brain, early hypotheses of cingulate cortex function were based on the anatomy of the region and its connections. Neuroanatomists in the late nineteenth century identified strong connections from the olfactory bulb to the cingulate cortex, suggesting a prominent role for this area in olfaction. The contrast between the large size of the cingulate cortex and the relative unimportance of olfaction in primates, however, weakened this idea. In 1937, the neuroanatomist James Papez

proposed an emotional role for the cingulate cortex, clearly linked to the widespread importance of olfaction to emotional processing in mammals as well as the strong connections to the cingulate cortex from the hypothalamus, which was then thought to be the source of emotions. Initial support for Papez' model came from Wilder Penfield's studies demonstrating that intracranial electrical stimulation of the hippocampal regions adjacent to the cingulate cortex evoked strong emotional reactions in awake human surgical patients. Additional support for the Papez model came from the results of bilateral cingulotomies on humans with severe depression, schizophrenia, and obsessive–compulsive disorder. Although this procedure, in which the entire cingulate cortex was removed, sometimes alleviated unwanted symptoms, it also resulted in blunted emotions and flat affect.

Early electrophysiological studies of the cingulate cortex used methods that had proven successful in elucidating the functions of primary sensory cortical areas, particularly visual cortex. These early studies, however, showed that cingulate neurons in anesthetized monkeys failed to respond to simple stimuli, such as bars and points of light, which strongly activate neurons in visual cortex. Subsequent studies in awake monkeys have shown that cingulate neurons do respond to large textured patches and other behaviorally salient stimuli, probably because such stimuli attract attention. Although most studies have probed cingulate function using visual stimuli, it is not an exclusively visual structure; neuroimaging studies have revealed PCC activation in response to emotionally relevant (threat-related) auditory stimuli, and these responses depend on the locus of auditory attention. Furthermore, when human subjects are hypnotized, pain-related responses in ACC, but not in somatosensory cortex, are reduced. Collectively, these results suggest that the cingulate cortex is not exclusively a sensory or emotional structure, but instead contributes to cognitive processing of motivationally relevant stimuli and for guiding action.

Role of Cingulate Cortex in Reward-Guided Behavior

Early evidence for the role of cingulate cortex in processing motivational and reward-related information came from studies of intracranial self-stimulation (ICSS) in rats. In the ICSS paradigm, a rat can press a lever to receive direct electrical stimulation from an electrode placed in its brain. The extent to which the rat presses the lever serves as an index of the reinforcing effects of activating this region. The strongest ICSS is evoked by stimulation of the medial forebrain bundle (MFB) or the ventral tegmental area (VTA) in the midbrain. ICSS can also be elicited by stimulation in the ACC and the retrosplenial cortex, as well as in areas connected to the cingulate cortex, including the hippocampus, the orbitofrontal cortex, and entorhinal cortex. While MFB and VTA stimulation causes rats to forsake food to the point of starvation, stimulation in cingulate cortex and related structures has much weaker effects on behavior. Notably, the rate of ICSS elicited by cingulate stimulation does not depend on electrical current or pulse frequency, as it does in the MFB or VTA. Also, unlike VTA and MFB, rates of ICSS elicited by cingulate activation gradually increase over time. One possible explanation for the gradually increasing rates of ICSS in cingulate cortex is that the stimulation slowly potentiates the formation of memories linking the animal's behavioral response with the stimulation. These results suggest that cingulate cortex does not simply generate a reward signal, but rather guides the formation of associations between rewards and sensory-guided actions.

Role of Cingulate Cortex in Selecting Action

Electrophysiological and pharmacological studies strongly implicate cingulate cortex in motor planning, possibly by binding reward-related information to action. One important example of the motor functions of cingulate cortex comes from a study of rhesus monkeys performing a task in which the links between sensory cues, rewards, and motor responses were systematically altered. Monkeys were presented with one of two cues that instructed one of two motor responses (either saccade or maintain fixation), and each motor response either was or was not associated with a reward. All combinations of parameters (sensory cue, motor response, mapping rule, and reward) were tested. In this task, ACC neurons represented both the motor plan and the reward contingency but almost never represent the pure sensory attributes of the cues.

One important question about the motor role of the cingulate cortex is the degree to which motor signals within cingulate cortex are abstract versus specific to a particular action. One study of a human patient with an ACC lesion supports the idea that motor coding in cingulate cortex is specific to a particular type of movement. This patient performed poorly on a Stroop task when manual responses were required, but was not affected when the task required verbal responses. This finding is also consistent with reports that electrical stimulation of cingulate cortex during surgical procedures can elicit uncontrolled hand movements. Together, these results demonstrate that motor representations within the cingulate motor areas are effector specific.

One important aspect of motor selection subserved by the cingulate cortex is the correct sequential ordering of responses. When monkeys perform a series of actions to obtain a single reward, many neurons in the

ACC fire selectively before only one of the actions but do not fire when rewards are given randomly. These results imply that one role of the ACC may be to control the order or timing of behavioral sequences.

Role of Cingulate Cortex in Monitoring Behavioral Outcomes

A substantial body of research has investigated the role of the cingulate cortex in monitoring errors and in evaluating and resolving conflict. These studies follow from the observation that negative electrical potentials are generated within the ACC when subjects make errors in many psychological tasks. This error-related negativity, or ERN, appears around 100 ms after the initiation of an incorrect motor response. A second ERN occurs about 250 ms after the subject is told she has made an error. Because the ERN arises after the behavioral response, it must reflect the consequences of an action rather than its planning or initiation. Moreover, since the ERN is generated independently of the modality of the motor response, it likely reflects abstract cognitive processing rather than movement-specific computations.

Subsequent studies have shown that an ERN is also generated on trials in which there is a high probability of error, but no error actually occurs. These trials are characterized as having high conflict, defined as the tension between two or more incompatible competing motor responses. For example, in a Stroop task, subjects must mediate between the habitual tendency to read the word and the instructed task of naming the color in which the word is printed (Figure 2(a)). When the word and the color are incongruent, these two responses conflict with each other. Conflict signals may trigger an executive control function that initiates increased effort to improve performance on subsequent trials. In support of this idea, behavioral studies have shown that trials with larger ERNs are more likely to be subsequently corrected and that subjects performing high-conflict tasks show the greatest increases in reaction time following trials with the greatest ERN.

These observations are supported by several functional magnetic resonance imaging (fMRI) studies showing hemodynamic responses believed to reflect the same processes that generate the ERN. However, two electrophysiological studies have failed to find any evidence for explicit conflict signals in the activity of single ACC neurons. Together, these observations raise the possibility that the cingulate cortex does not explicitly encode response conflict, but instead encodes all possible motor actions and their likelihood of yielding reward. Because high-conflict conditions may be associated with multiple motor plans, which would recruit multiple overlapping groups of neurons, this might lead to what appears to be a conflict-related BOLD signal. Nonetheless, the precise role of cingulate cortex in conflict monitoring remains unclear.

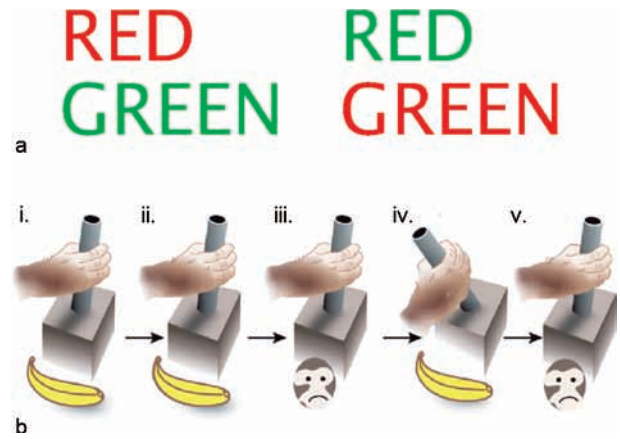

Figure 2 Behavioral monitoring in ACC. (a) Examples showing the low- and high-conflict conditions of the Stroop task. The Stroop task requires subjects to name the color while ignoring the word. The high-conflict condition (right, incongruent word/color) evokes an increased BOLD response in the dorsal ACC. (b) Cartoon indicating the trial structure of the task used by Shima and Tanji. Subjects can perform one of two actions (turn or push a handle). One action is consistently rewarded (i and ii). At some point, the rewarded action is switched (iii). A normal subject will alter the strategy at this point to begin receiving rewards again (iv). Monkeys with reversible ACC lesions fail to reliably adopt the new strategy. Recent work has shown that monkeys with permanent ACC lesions can adopt the new strategy for a few trials, but soon revert to the former strategy (v).

Both electrophysiological and neuroimaging studies consistently report that cingulate cortex seems to be more active when subjects actively search for the optimal strategy to solve a task than when they simply follow arbitrary rules. Enhanced activity is seen in ACC in tasks that require unspecified responses (e.g., word-generation tasks) or in monkeys when movements are triggered internally rather than in response to explicit external cues. Moreover, neurons in ACC are more active when monkeys endeavor to learn the order in which they must press three squares on a computer screen than after they have learned the sequence. Similarly, BOLD activity in ACC is much stronger when human subjects monitor the monetary payoffs associated with pressing buttons and try to maximize their yield than when they merely follow explicit instructions.

Role of Cingulate Cortex in Assigning Motivational Significance to Potential Actions

One model of cingulate cortex function that may account for most of the foregoing observations purports that the general function of cingulate

circuitry is to link reward information to potential sensory-guided actions. This model is supported by the findings (described above) that cingulate cortex participates in both sensory and motor processing and, moreover, appears to integrate reward information into the context of ongoing sensorimotor transformations. According to this view, cingulate cortex provides a functional linkage between limbic areas and cortical sensory and motor areas.

One landmark study supporting this model demonstrated that cingulate cortex participates specifically in reward-based guidance of action. In this study, monkeys grasped a joystick that they could either lift or turn on each trial (Figure 2(b)). One of these actions was rewarded and the other was not. The only way the monkey could identify which action was rewarded was to monitor the consequences of their actions. When reward value associated with one action declined, the monkey could switch to the other action to receive a larger reward. Reward-related switches in behavior elicited high firing from neurons in CMAs. Moreover, temporary lesions in this area abolished the animal's ability to modify its behavior based on changing rewards. Later work has shown that permanent lesions to the ACC do not specifically impair the ability to modify behavior in response to changing outcomes, but instead disrupt the ability to sustain new behavioral strategies following the reward-guided switch. These results suggest that the ACC specifically participates in the reward-based consolidation of new behavioral strategies.

Further support for a role for cingulate cortex in integrating reward outcomes into ongoing sensorimotor transformations comes from the finding that ACC does not represent abstract stimulus–reward mappings when they have no influence on the selection of motor responses. In one study, monkeys were trained to perform two different tasks. One task was a variant of the reward-based switching task, as described above, in which monkeys performed one of two actions (lift or turn a handle) in response to a reward. The other task was a visual discrimination task in which monkeys learned to associate one of two stimuli with a reward and then select the

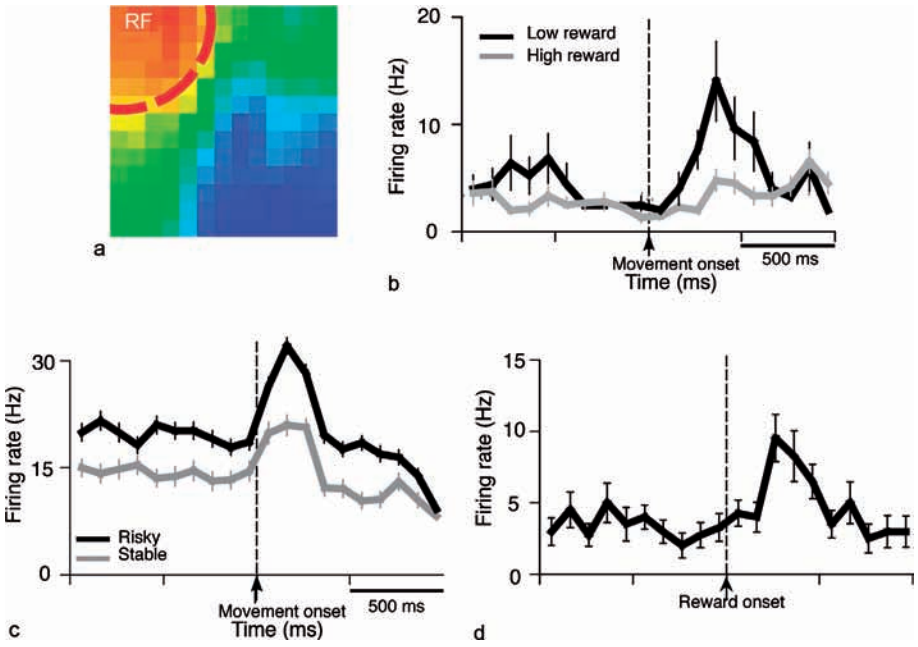

Figure 3 Visuospatial information encoded by neurons in PCC is modulated by reward variables. (a) Neurons in PCC have large contralateral response fields. Color map showing average response of a single neuron following saccades to points in a grid of locations, size 40×40 degrees. Responses are greater when saccades are directed to the upper-left quadrant (response field, RF). (b) PCC neurons encode reward size associated with saccades into their RF. Peri-stimulus time histograms (PSTHs) showing the temporal evolution of the average response of a single PCC neuron in a free choice task. Responses are significantly greater on high reward trials (black line) than on low reward trials (gray line). This modulation affects both the decision period and the postsaccade period of the trial. (c) PCC neurons encode reward uncertainty, or risk, even when average expected payoff is identical. PSTHs showing the average response of a single PCC neuron in a free choice task. Responses are greater when subject chooses to shift gaze to the risky target (black line). (d) PCC neurons respond following unexpected rewards. PSTH showing average response of a single PCC neuron when an unexpected reward is delivered. From McCoy AN and Platt ML (2005) Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature Neuroscience* 8: 1220–1227. Figures reproduced with permission from McCoy et al.

rewarded stimulus. ACC-lesioned monkeys were specifically impaired in their ability to perform the reward-based switching task but their performance was unaffected in the visual discrimination task.

Reward-related computations within cingulate cortex are not limited to the ACC. Neurons in the PCC, for example, represent the subjective value of rewards associated with orienting toward particular locations in space (**Figure 3**). PCC neurons respond after the illumination of targets, saccade onset, and following reward delivery. Because PCC neurons respond in proportion to the size and probability of rewards, as well as subjective preference for uncertain rewards associated with orienting to a visual target, it has been hypothesized that firing rate in PCC is proportional to the subjective utility of a location in space. The responses of PCC neurons to events within a trial may reflect continuous updating of the subjective utility of the preferred location. Collectively, these results suggest that PCC, like ACC, participates in connecting information about rewards to ongoing sensorimotor transformations.

Conclusion

The cingulate cortex has long been viewed as a large and mysterious terra incognita. Nonetheless, a great deal of knowledge has been gained in recent years. This knowledge, derived from single-unit electrophysiology, lesion studies, fMRI, and event-related potential (ERP) studies, implicates the cingulate cortex in the use of reward information to optimize the process of selecting actions based on sensory information.

See also: Attentional Networks; Attentional Networks in the Parietal Cortex; Decision-Making and Vision; Parietal Cortex and Spatial Attention; Representation of Reward; Reward Systems: Human; Reward Decision-Making; Reward Neurophysiology and Primate Cerebral Cortex; Visual System: Multiple Visual Areas in Monkeys.

Further Reading

Botvinick MM, Braver TS, Barch DM, Carter CS, and Cohen JD (2001) Conflict monitoring and cognitive control. *Psychological Review* 108: 624–652.

Bush G, Luu P, and Posner MI (2000) Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences* 4: 215–222.

Gehring WJ, Coles MG, Meyer DE, and Donchin E (1995) A brain potential manifestation of error-related processing. *Electroencephalography and Clinical Neurophysiology* 44(supplement): 261–272.

Hadland KA, Rushworth MF, Gaffan D, and Passingham RE (2003) The anterior cingulate and reward-guided selection of actions. *Journal of Neurophysiology* 89: 1161–1164.

Isomura Y, Ito Y, Akazawa T, Nambu A, and Takada M (2003) Neural coding of "attention for action" and "response selection" in primate anterior cingulate cortex. *The Journal of Neuroscience* 23: 8002–8012.

McCoy AN and Platt ML (2005) Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature Neuroscience* 8: 1220–1227.

Olson CR, Musil SY, and Goldberg ME (1996) Single neurons in posterior cingulate cortex of behaving macaque: Eye movement signals. *Journal of Neurophysiology* 76: 3285–3300.

Paus T (2001) Primate anterior cingulate cortex: Where motor control, drive and cognition interface. *Nature Reviews Neuroscience* 2: 417–424.

Procyk E, Tanaka YL, and Joseph JP (2000) Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nature Neuroscience* 3: 502–508.

Rushworth MF, Walton ME, Kennerley SW, and Bannerman DM (2004) Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences* 8: 410–417.

Shidara M and Richmond BJ (2002) Anterior cingulate: Single neuronal signals related to degree of reward expectancy. *Science* 296: 1709–1711.

Shima K, Aya K, Mushiake H, Inase M, Aizawa H, and Tanji J (1991) Two movement-related foci in the primate cingulate cortex observed in signal-triggered and self-paced forelimb movements. *Journal of Neurophysiology* 65: 188–202.

Shima K and Tanji J (1998) Role for cingulate motor area cells in voluntary movement selection based on reward. *Science* 282: 1335–1338.

Turken AU and Swick D (1999) Response selection in the human anterior cingulate cortex. *Nature Neuroscience* 2: 920–924.

Vogt BA and Gabriel M (1993) *Neurobiology of Cingulate Cortex and Limbic Thalamus*. Boston: Birkhauser.

Circadian Function and Therapeutic Potential of Melatonin in Humans

A J Lewy, J Songer, K Yuhas, and J S Emens, Oregon Health & Science University, Portland, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

The Circadian Function and Therapeutic Potential of Melatonin in Humans

Melatonin is a neurohormone produced by the pineal gland, almost exclusively at night. It was discovered in 1958 by Aaron Lerner, a dermatologist and chemist, who was interested in pigmentation disease. Although melatonin does not have skin-lightening effects in humans, Lerner's work in amphibians eventually led to the discovery that endogenous melatonin production is a central part of the biological timing system. Almost 50 years after its isolation and identification from bovine pineal glands, this neurohormone is now recognized as the best biomarker for endogenous circadian phase. Commercially synthesized from chemical reagents, exogenous melatonin is available for human use, either over-the-counter or by prescription, depending on the country of sale.

The following, complemented by several articles in a 2005 volume of the *Journal of Biological Rhythms* dedicated to human circadian rhythms, provides a comprehensive review of the circadian aspects of melatonin. Work in this area is increasing exponentially. Accordingly, we have included very recent findings, including some as yet unpublished. These findings extend the relevance of this contribution beyond that of a historical review.

Melatonin and 6-Sulfatoxymelatonin

Development of melatonin assays in the 1970s opened up new possibilities for human studies of circadian rhythms. The principal metabolite of melatonin is 6-sulfatoxymelatonin. A convenient way to estimate the total amount of melatonin production in a 24 h period is to collect urine overnight and measure this metabolite, provided that metabolism of melatonin by the liver is within normal limits. However, frequently assessed plasma or salivary levels allow for a more highly resolved use of melatonin as a phase marker.

Melatonin Gas Chromatographic-Mass Spectrometric Assay and Radioimmunoassays

As measured by the 1978 gold-standard gas chromatographic-mass spectrometric (GCMS) assay, melatonin levels in normal humans are less than $10\,\mathrm{pg\,ml^{-1}}$ during the day and increase to several times that amount at night. Melatonin levels are almost always reported in picograms per milliliter, although occasionally picomoles per liter ($43\,\mathrm{pmol\,l^{-1}} = 10\,\mathrm{pg\,ml^{-1}}$) have been substituted. The GCMS assay uses negative chemical ionization and a deuterated internal standard, and it has the greatest precision, accuracy, and sensitivity of any assay. It has directly and indirectly validated the radioimmunoassays (RIAs) that are widely used today because of their convenience and efficiency. One of the first applications of the GCMS assay was to show that virtually all of the circulating levels of melatonin are derived from the pineal gland, thus validating the measures of plasma and salivary melatonin as markers of pineal function. In particular, the timing of endogenous melatonin production can be used as a marker for the phase of the circadian pacemaker (located in the suprachiasmatic nucleus (SCN)).

Bright (High-Intensity) Light

The GCMS assay was also employed in the discovery that light can suppress melatonin production in humans. Humans require brighter light for melatonin suppression than other animals. In addition to melatonin assay methodology, this explains why light suppression of melatonin in humans was not reported until 1980, although it was attempted earlier. The interspecies difference was thought to be adaptive, that is, to be a short-term response to a previous history of light exposure. Several studies have since shown that the intensity of recent ambient light exposure determines the degree of suppression of melatonin to light – the brighter the light, the greater the suppression of melatonin production.

Melatonin Suppression Test

Intensity-dependent suppression of melatonin by light is the basis for the melatonin suppression test (MST). Originally, subjects were exposed to 500 lux light between 2 and 4 a.m. This intensity was used in order to compare the degrees of partial suppression between individuals or groups of individuals (such as manic-depressive patients). Variations of this test have been applied in a number of different ways. Although it is assumed that sensitivity to light as determined by this test reflects the overall sensitivity of the circadian system to light, as presently operationalized this test does not appear to be useful in determining the patency of the retinohypothalamic innervation of the

SCN of the hypothalamus in blind people, due in part to a lack of sensitivity and specificity of this test as a biomarker in the blind.

Suppression of melatonin by light has at least four important implications. First, humans might have biological rhythms that are regulated by exposure to (bright) outdoor light, remaining relatively unperturbed by ordinary-intensity indoor light. Second, bright artificial light might be substituted for sunlight to experimentally, and perhaps therapeutically, manipulate biological rhythms in humans. Third, totally blind people with no light perception might have abnormal circadian rhythms. And fourth, measurement of the melatonin rhythm in sighted people should be done under dim light to avoid masking its true circadian phase.

Seasonal Affective Disorder, or Winter Depression

Based on these implications, the administration of bright light was first used as a treatment in patients with both nonseasonal and seasonal depression. However, more than 25 years later, much still remains to be known about light's efficacy, optimal treatment parameters, and mechanism of action in nonseasonal depression. On the other hand, bright light is now established as the treatment of choice for winter depression, or seasonal affective disorder (SAD). In fact, the most common use of bright light is for the treatment of SAD, although it is also used for treating advanced and delayed sleep phase syndromes, jet lag, and shift-work maladaptation. SAD and winter depression are two terms that are used interchangeably because there do not appear to be other types of mood disorders that recur at the same time each year in patients studied longitudinally – despite the fact that cross-sectional studies show fall and spring peaks in depression and an increased incidence of mania in the summer.

Phase Shift Hypothesis

The leading hypothesis for SAD and the antidepressant mechanism of bright light is the phase shift hypothesis (PSH). According to the PSH, winter depression is at least partly related to a misalignment between circadian rhythms and the sleep/wake cycle that occurs because of the shorter winter days. The prototypical patient with SAD is thought to become phase-delayed in the winter, cueing to the later dawn. Thus, their circadian rhythms, which are tightly coupled to the SCN, shift later with respect to their sleep/wake cycle. Some atypical patients with SAD may be cueing to the earlier dusk and phase advance in the winter. In either case, bright light should be scheduled to provide a corrective realignment. For most (phase-delayed) patients, the optimal timing of light is in the morning, immediately on awakening. The usual intensity is 10 000 lux of ordinary (cool white) fluorescent light, delivered by a light box set at a precise distance from the eyes; 1–2 h of light may be necessary for several days until the antidepressant response is complete, and then the duration of exposure can be reduced. Too much light can also overly phase shift circadian rhythms, causing a misalignment in the opposite direction, which might be the case if patients relapse after a brief response. If morning light is not at all efficacious, a trial of evening light (7–9 p.m.) is in order. Light exposure at both times is not recommended because the phase advance from morning light will be counteracted by the phase delay due to evening light.

Dim Light Melatonin Onset as a Marker for the Phase of the Endogenous Circadian Pacemaker

In sighted people, the best marker for circadian phase is the dim light melatonin onset (DLMO). Suppression of melatonin production by light is to be avoided when using the endogenous melatonin profile as a marker for the underlying phase of the endogenous circadian pacemaker. Removing this masking effect by collecting samples under dim light (light intensities of less than 10–30 lux that begin 1–2 h before the earliest expected rise in melatonin levels) increases the reliability that the melatonin profile is reflecting the true phase of the endogenous circadian pacemaker. The DLMO does not represent the time that melatonin rises under naturalistic lighting conditions, in which the melatonin profile reflects both the timing of the endogenous circadian pacemaker and the gating effect of darkness (suppressant effect of light). The naturalistic profile represents the melatonin signal that is actually presented to the individual living under conditions of ambient light. If a function for endogenous melatonin production in humans is identified, it will be best understood as mediated by the melatonin profile assessed under naturalistic lighting conditions.

Dim Light Melatonin Offset and Melatonin Synthesis Offset

In blind individuals (and in sighted people) studied under naturalistic conditions, the margins of the active phase of melatonin are the melatonin onset (MO) and melatonin offset (MOff). The three features of the melatonin profile that most easily lend

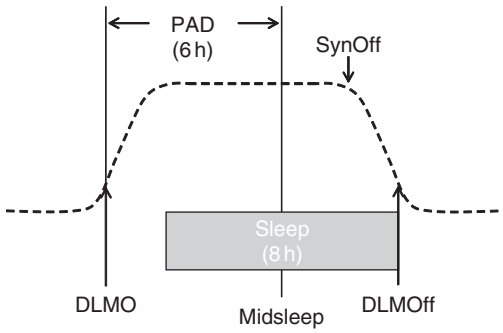

Figure 1 The three features of the melatonin profile that most easily lend themselves for marking circadian phase in sighted people: the dim light melatonin onset (DLMO), dim light melatonin offset (DLMOff), and melatonin synthesis offset (SynOff). PAD, phase angle difference.

themselves for marking the circadian phase in sighted people are the DLMO, dim light melatonin offset (DLMOff), and melatonin synthesis offset (SynOff). These are illustrated in Figure 1. The operational definition of the plasma DLMO (and the DLMOff) is usually the interpolated clock time when levels continuously rise above (or fall below) the threshold of 10 pg ml^{-1}. Saliva concentrations at these times are about one-third those of plasma; the corresponding salivary threshold is therefore 3 pg ml^{-1}. For both saliva and plasma, samples are ideally collected every 30 min.

The SynOff, which is the time when melatonin levels begin to decline, is recommended for assessing the putative second oscillator. For many years, some investigators thought that the human melatonin profile was regulated by two oscillators, which may be true for rodents and other organisms. According to the two-oscillator model, the SynOff is regulated by a separate pacemaker that is cued to dawn. Although the issue has yet to be resolved, there appears to be increasing consensus that if there are two oscillators in humans, they are tightly phase-locked under most circumstances. Therefore, the measurement of any of these markers should provide the same information about circadian phase. In sighted people, the DLMO is the marker of choice because collecting samples usually does not interfere with sleep and can be done at home (using saliva).

Phase Angle Difference, Circadian Misalignment, and Phase Typing

Phase angle difference (PAD) is the time interval between two circadian rhythms, for example, between the DLMO and midsleep. On average, the interval between the 10 pg ml^{-1} plasma DLMO and the midpoint of sleep is 6 h (PAD 6), and the interval between the 2 pg ml^{-1} DLMO ($DLMO_2$) and the midpoint of sleep is 7 h (PAD 7). (Unless otherwise specified, DLMO refers here to the 10 pg ml^{-1} plasma threshold.) PAD 6 appears to represent the optimal mood in SAD patients, before and after treatment with low-dose melatonin (Figure 2). PAD 6 has also been used to operationally define phase-advanced (PAD $6 > 6$ h) and phase delayed (PAD $6 \leq 6$ h) types; the former characterizes people with a DLMO that is phase advanced with respect to sleep, and the latter (typical of patients with SAD) characterizes individuals with a DLMO that is phase delayed with respect to sleep. Phase-delayed SAD patients are optimally treated with bright light in the morning and/or low-dose melatonin in the afternoon/evening; phase-advanced SAD patients are optimally treated with bright light in the evening and/or low-dose melatonin in the morning. Phase typing based on PAD 6 may prove to be useful in the diagnosis and treatment of other disorders as well; however, the circadian component in nonseasonal depression is not likely to be as important as it is in SAD, in which it may explain as much as 65% of the variance in depression ratings (Figure 2).

Zeitgeber Time

The time interval between the DLMO and the entraining zeitgeber (environmental time cue) is the zeitgeber time (ZT) of the DLMO. This can also be used to estimate the phase angle of entrainment (PAE). By tradition, the time interval between waketime (lights on) and the DLMO is the DLMO ZT. On average, the DLMO occurs 14 h after waketime. Thus, the average DLMO ZT is 14.

Phase Angle of Entrainment and Circadian Period

The DLMO ZT can be used to estimate the PAE. However, this assumes that the conveniently measured sleep variable, waketime, faithfully represents the phase of reference for the primary zeitgeber, the light dark cycle, an assumption grounded in the fact that the sleep/wake cycle superimposes structure on the perceived light dark cycle. As more is understood about the circadian response to light (and dark), this reference phase will be calculated with greater accuracy. At present, however, we must deal as best we can with the fact that waking up into bright sunlight results in a shorter PAE (DLMO ZT) than awakening before dawn, a fact that complicates extending findings from studies in which animals with shorter intrinsic periods (taus) entrain to a 24 h zeitgeber with a shorter PAE. (Unless otherwise specified, intrinsic tau is assumed here to be greater than 24 h, which is

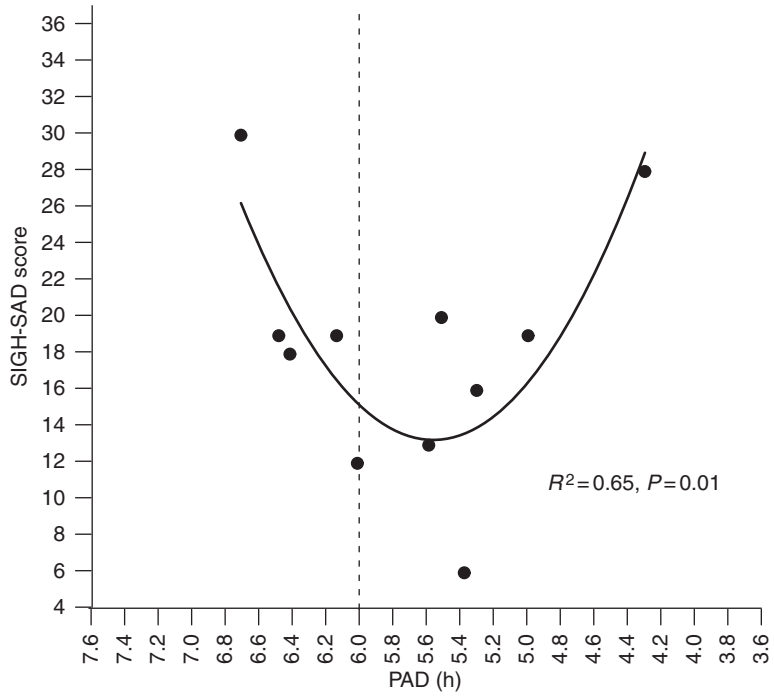

Figure 2 Mood in prototypical SAD patients after treatment with low-dose melatonin given in the afternoon/evening to provide a therapeutic phase advance. PAD 6 appears to represent optimal mood in SAD patients. PAD, phase angle difference; SIGH-SAD, Structured Interview Guide for the Hamilton Depression Rating Scale, for Seasonal Affective Disorder.

clearly the case for the majority of humans.) The animal studies were conducted under conditions in which the light/dark cycle was much more strictly controlled than the measurement of DLMO ZT in humans exposed to their usual light/dark cycles. However, under laboratory conditions, sighted humans with shorter taus have also been shown to have shorter PAEs, a finding consistent with what was found in blind people entraining to a daily dose (10 mg) of melatonin. PAE has also been estimated in humans using the melatonin/sleep onset interval (MSI). Another consideration is that PAD is most meaningful when measured under naturalistic conditions, whereas PAE is most useful when measured under controlled lighting and stabilized sleep/wake experimental conditions.

Circadian Time

The clock time of the DLMO is of importance, and so is its circadian time (CT). By tradition, CT 0 is waketime (lights on). Because the DLMO occurs 14 h on average after waketime, the 10 pg ml^{-1} plasma DLMO is designated CT 14. It is heuristically useful (if not always empirically demonstrable) to refer to the DLMO as CT 14. This is particularly useful for understanding the phase-shifting effects of light and melatonin, that is, at what time these phase-resetting agents cause phase advances versus phase delays.

CT and ZT are sometimes used interchangeably. CT is more appropriately used in free-running blind people, whereas ZT is more appropriately used under conditions of entrainment, the usual circadian status of sighted individuals. However, designating the DLMO as CT 14 in sighted people is also useful for assessing the phase of the light and melatonin phase response curves (PRCs). It has been known for almost 2 decades that scheduling these phase-resetting agents according to CT rather than clock time optimizes their effects because CT takes into account individual differences in circadian phase.

The Light and Melatonin Phase Response Curves

As already mentioned, morning light causes phase advances, and evening light causes phase delays in entrained individuals. These times account for two points on the light PRC, which is usually plotted according to CT. Light and melatonin PRCs are provided in Figure 3. Research in this area is active, and the precise shapes of these curves will probably be refined over the next few years. In essence, however, there is a 12 h zone for phase advances and a 12 h zone for phase delays; that is, a phase-resetting agent will cause phase advances when initiated at any time within its PRC's phase-advance zone and will cause phase delays when initiated at any time within its PRC's phase-delay zone. Until more is known about

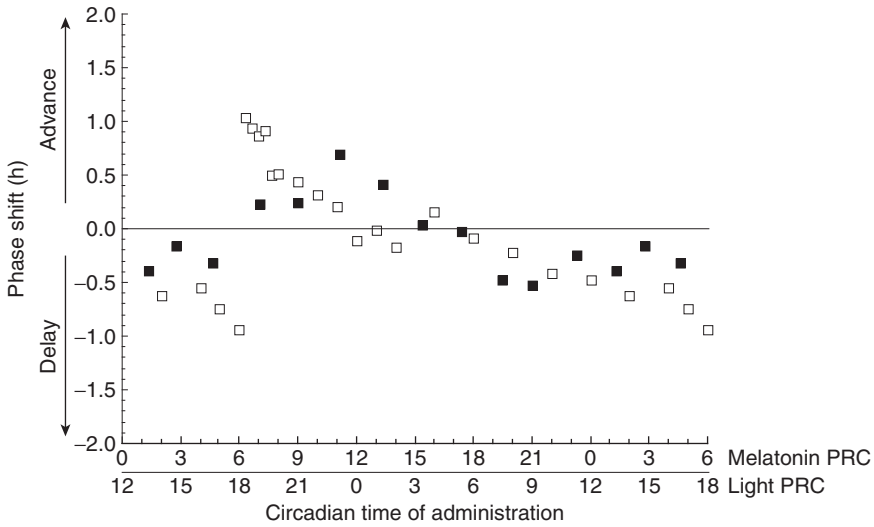

Figure 3 Light and melatonin phase response curves (PRCs). Copyright (1998) From The human phase response curve (PRC) to melatonin is about 12 h out of phase with the PRC to light. *Chronobiology International* by Lewy AJ, Bauer VK, Ahmed S, et al. Reproduced by permission of Taylor & Francis Group, LCC., http://www.taylorandfrancis.com.

these PRCs, it is at least heuristically useful to consider them to be 12 h out of phase with one another and to consider the crossover times to be CT 6 and CT 18, 8 h before and 4 h after the DLMO, respectively. For a person whose DLMO is at 8 p.m., these convert to clock times of noon and midnight, so a useful mnemonic is that the phase advance zone for light (the phase delay zone for melatonin) occurs during the morning hours, whereas the phase delay zone for light (the phase advance zone for melatonin) occurs during the afternoon/evening hours (Figure 4).

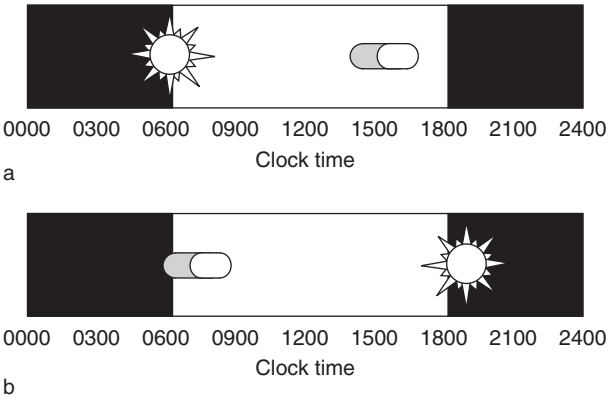

Figure 4 Phase advance and delay zones for a person whose DLMO is at 8 p.m.: (a) the phase advance zone for light occurs during the morning hours and the phase advance zone for melatonin occurs during the afternoon/evening hours; (b) the phase delay zone for light occurs during the afternoon/evening hours and the phase delay zone for melatonin occurs during the morning hours. Reproduced from Lewy AJ and Sack RI (1996) The role of melatonin and light in the human circadian system. In: Buijs RM, Kalsbeck A, Romijn HJ, Pennartz CMA, and Mirmiran M (eds.) *Hypothalamic Integration of Circadian Rhythms*, *Progress in Brain Research*, vol. 111, pp. 205–215. Amsterdam: Elsevier, with permission from Elsevier.

Blind Free-Runners

Approximately 15% of legally blind people are totally blind (unable to detect light). Most totally blind individuals free-run and are called bind free-runners (BFRs). BFRs have observed taus that are usually slightly greater than 24 h (24.1–25.0 h). Almost all BFRs can be entrained to a daily dose of melatonin. A BFR will stop free-running when the exogenous melatonin dose comes into contact with the entrainment point on the melatonin PRC, which for most BFRs is in the advance zone. Tau appears to be related to an inherited trait determined by clock genes. Although not proved, it is convenient to think of people with shorter intrinsic taus as morning types and to think of people with longer intrinsic taus as evening types. Some people have taus slightly shorter than 24 h; they are most assuredly morning types.

Low physiological doses appear to be sufficient for entrainment and may be more effective than some higher doses, presumably because of spillover of these high doses onto the wrong zone of the melatonin PRC. For example, higher doses taken on the advance zone always results in a phase advance, but spillover on to the delay zone reduces the magnitude of the phase advance. Spillover, and other pharmacokinetic parameters, may explain why not all BFRs have been shown to be entrainable to a daily dose of melatonin. Another explanation for failure to entrain has been the finding that initiating treatment on the delay zone will reduce the likelihood of entrainment even after the MO

has transited through a complete 24 h circadian beat cycle (defined later); however, this finding is inconsistent with animal studies and has not been replicated by other research groups (for any doses of melatonin). In any event, low doses (0.5 mg) almost always result in the entrainment of BFRs. They are optimally given around 6 p.m. to entrain most BFRs to the normal phase. BFRs with taus <24 h require melatonin at waketime to be entrained at the normal phase. Low doses are effective at both of these administration times, which has the added benefit of minimizing the soporific side effect that sometimes occurs immediately after taking melatonin. A melatonin dose–response curve has recently been described and appears in Figure 5; the daily phase advance for the lowest entraining dose of melatonin was calculated by subtracting 24 h from tau at entrainment phase (TEP).

As already mentioned, the measurement of MOs is particularly important in doing research in blind people and may have clinical relevance as well. Assessing the MO is essential when determining the CT of exogenous melatonin administration. In research studies, exogenous melatonin is not administered on days when endogenous melatonin levels are assessed. The CT of administration can still be estimated in the following way: if a BFR's MO occurs at 3 a.m. the day before restarting treatment at 6 p.m., melatonin given at 6 p.m. will be on the delay zone (CT 5) of the melatonin PRC (the delay zone begins at CT 6, which is 8 h before the MO). Thus, we can be virtually certain that the BFR has escaped entrainment.

Relative Coordination to Weak Zeitgebers, the Two-Point Tau, the Tau Response Curve, and Tau at Entrainment Phase

A circadian beat cycle in a BFR is completed when the MO transits through 24 h. Every BFR studied to date with high resolution (an MO assessment approximately every 2 weeks through a complete circadian beat cycle) shows relative coordination to as-yet-unidentified weak zeitgebers (Figure 6). The pattern of response in all BFRs is the same; as the MO drifts across the day, the two-point tau between each successive pair of MOs is longer than the BFR's average tau, and as the MO drifts across the night, the two-point taus are shorter than average. These two-point taus can be plotted against the average clock time of the two MOs to produce a tau response curve (tauRC). Figure 7 provides tauRCs from two BFRs. The pattern of response to the weak zeitgebers (which may include social cues) is the same for all BFRs; however, the range of oscillation (the difference between the shortest and longest two-point taus)

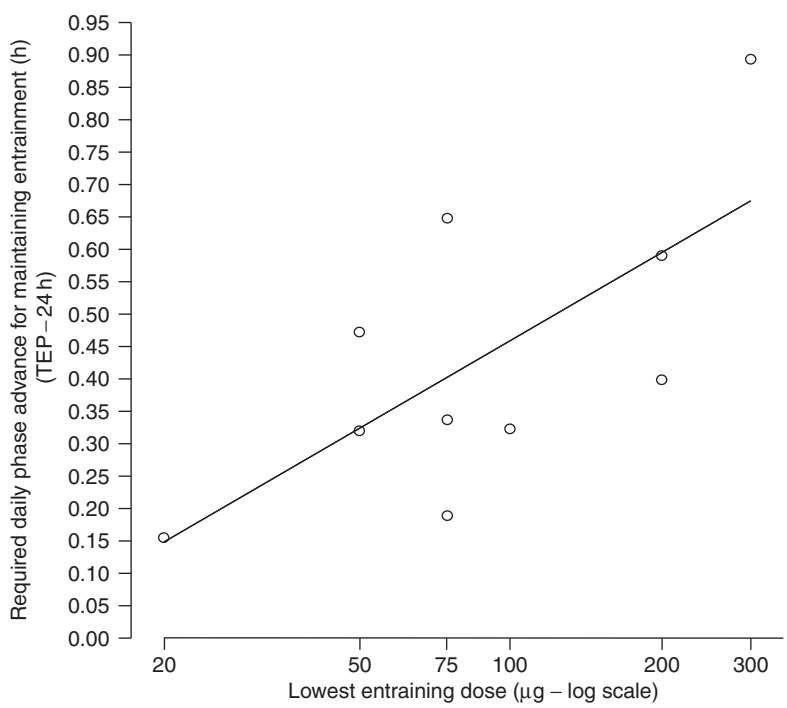

Figure 5 Melatonin dose–response curve. TEP, tau at entrainment phase. Copyright (2005) From Melatonin free-running blind people according to a physiological dose response curve. *Chronobiology International* by Lewy AJ, Emens JS, Lefler BJ, Yuhas K, and Jackman AR. Reproduced by permission of Taylor & Francis Group, LCC., http://www.taylorandfrancis.com.

varies markedly between individuals. Identification of the weak zeitgebers and demonstration that they are regular in strength and timing will validate the use of range of oscillation not just as a bioassay of responsivity but of an individual's sensitivity to the weak zeitgebers. The tauRC of a BFR can also be used to calculate TEP, which is important in calculating the physiological dose–response curve for melatonin. TEP is the two-point tau a BFR would be expected to have while free-running through the clock time of the now-entrained MO.

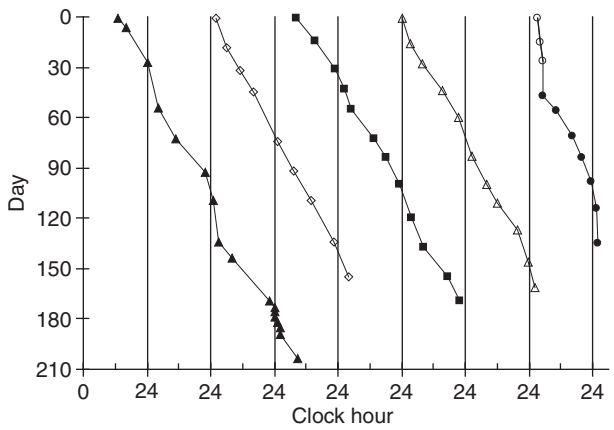

Figure 6 Pattern of response in blind free-runners (BFRs), with melatonin onset (MO) assessment approximately every 2 weeks through a complete circadian beat cycle. A different symbol is used for each BFR. Reproduced from Emens JS, Lewy AJ, Lefler BJ, and Sack RL (2005) Relative coordination to unknown weak zeitgebers in free-running blind individuals. *Journal of Biological Rhythms* 20(2): 159–167.

Natural Entrainment

As already mentioned, most totally blind people are BFRs. Some totally blind people, however, appear to naturally entrain to the unknown weak zeitgebers. A more rigorous operational definition is required for natural entrainment than for entrainment to melatonin. In order to designate a blind person as naturally entrained, a tau very close to 24.00 h must be ruled out, as well as a transient short tau, observed as the MOs move through the short-tau zone during relative coordination. Recent evidence suggests that, among adults, males are less likely to be naturally entrained than females, a potential difference in sensitivity to the weak zeitgebers that is apparent after puberty (in fact, there is a suggestion that prepubertal males may be more likely to be naturally entrained than females). Furthermore, the range of oscillation in adult female BFRs is significantly greater than in adult males. These findings led to the prediction that sighted adult females would have an advanced PAE compared to adult males. This is based on the fact that human intrinsic tau is usually greater than 24 h, thus requiring a daily phase advance for entrainment; more sensitivity to the overall set of zeitgebers would therefore result in an earlier PAE. These findings are consistent with those in *Octodon degus*, a diurnal rodent, in whom social cues and pheromones have been shown to be sex-specific.

The Function of Endogenous Melatonin in Humans

The precise circadian function of endogenous melatonin production remains to be demonstrated. This

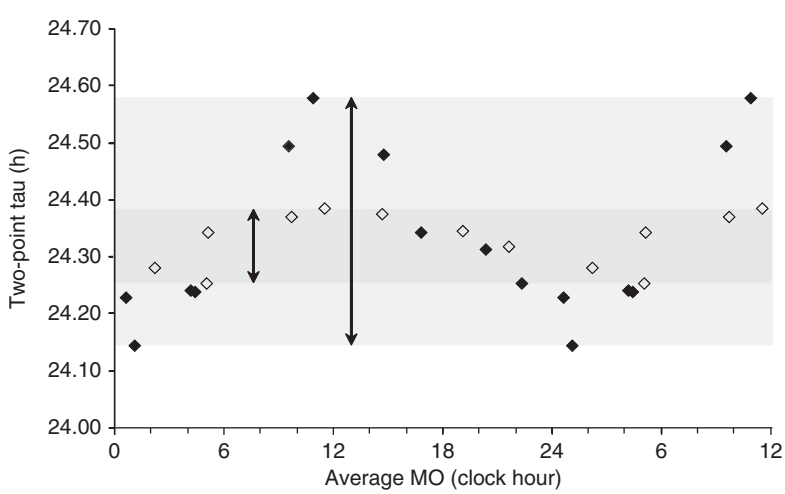

Figure 7 Tau response curves for two blind free-runners (BFRs). The pattern of response to the weak zeitgebers is the same for all BFRs; however, the range of oscillation (the difference between the shortest and longest two-point taus) varies markedly between individuals. MO, melatonin onset.

function, as postulated in the first description of the human melatonin PRC, is dependent on the acute suppressive effect of light, which has been retained in humans who do not seem to have marked seasonal differences in the duration of the naturalistic profile. Taking into account that a truly useful circadian melatonin signal must come from a person who is entrained to the light/dark cycle, the main function of endogenous melatonin in humans may be to synchronize the third trimester fetus and the suckling infant to the mother. At about 3 months of age, the retinohypothalamic tract is fully functional and the infant can independently synchronize to the light/dark cycle. This would be an important area for future research.

See also: Circadian Genes and the Sleep–Wake Cycle; Circadian Rhythms: Influence of Light in Humans; Entrainment of Circadian Rhythms by Light; Genetics of Circadian Disorders in Humans; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Photoreceptors and Circadian Clocks; Pineal Gland and Melatonin; Seasonal Timing: Neural Mechanisms.

Further Reading

Arendt J (1995) *Melatonin and the Mammalian Pineal Gland*. London: Chapman & Hall.

Emens JS, Lewy AJ, Lefler BJ, and Sack RL (2005) Relative coordination to unknown 'weak zeitgebers' in free-running blind individuals. *Journal of Biological Rhythms* 20(2): 159–167.

Lewy AJ, Ahmed S, Jackson JML, and Sack RL (1992) Melatonin shifts circadian rhythms according to a phase-response curve. *Chronobiology International* 9(5): 380–392.

Lewy AJ, Bauer VK, Ahmed S, et al. (1998) The human phase response curve (PRC) to melatonin is about 12 hours out of phase with the PRC to light. *Chronobiology International* 15(1): 71–83.

Lewy AJ, Bauer VK, Cutler NL, et al. (1998) Morning versus evening light treatment of patients with winter depression. *Archives of General Psychiatry* 55: 890–896.

Lewy AJ, Cutler NL, and Sack RL (1999) The endogenous melatonin profile as a marker for circadian phase position. *Journal of Biological Rhythms* 14(3): 227–236.

Lewy AJ, Emens JS, Lefler BJ, Yuhas K, and Jackman AR (2005) Melatonin entrains free-running blind people according to a physiological dose-response curve. *Chronobiology International* 22(6): 1093–1106.

Lewy AJ, Lefler BJ, Emens JS, and Bauer VK (2006) The circadian basis of winter depression. *Proceedings of the National Academy of Sciences of the United States of America* 103(19): 7414–7419.

Lewy AJ and Markey SP (1978) Analysis of melatonin in human plasma by gas chromatography negative chemical ionization mass spectrometry. *Science* 201: 741–743.

Lewy AJ, Sack RL, Miller S, and Hoban TM (1987) Antidepressant and circadian phase-shifting effects of light. *Science* 235: 352–354.

Lewy AJ, Wehr TA, Goodwin FK, Newsome DA, and Markey SP (1980) Light suppresses melatonin secretion in humans. *Science* 210: 1267–1269.

Lockley SW, Skene DJ, James K, et al. (2000) Melatonin administration can entrain the free-running circadian system of blind subjects. *Journal of Endocrinology* 164: R1–R6.

Sack RL, Brandes RW, Kendall AR, and Lewy AJ (2000) Entrainment of free-running circadian rhythms by melatonin in blind people. *New England Journal of Medicine* 343: 1070–1077.

Wright KP Jr., Gronfier C, Duffy JF, and Czeisler CA (2005) Intrinsic period and light intensity determine the phase relationship between melatonin and sleep in humans. *Journal of Biological Rhythms* 20(2): 178–188.

Zatz M (ed.) (2005) *Special Issue: Human Circadian Rhythms: Regulation and Impact*. *Journal of Biological Rhythms* 20(4).

Relevant Websites

http://www.ohsu.edu – Oregon Health and Science University Sleep and Mood Disorders Lab.

http://www.sltbr.org – Society for Light Treatment & Biological Rhythms.

Circadian Gene Expression in the Suprachiasmatic Nucleus

M U Gillette and S-H Tyan, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Regulated expression of a core set of clock genes is the basis of circadian timekeeping in vertebrates, as in other organisms. Clock genes are defined as being necessary to generate normal circadian rhythms, near 24 h oscillations in biological processes. With the enlargement of the brain in vertebrates, control of circadian rhythms became centralized. In lower vertebrates, such as fishes, lizards, and some birds, the pineal gland, near the surface of the brain, is directly photosensitive and exerts significant circadian control. In mammals, the pineal is overlain by enlarged cerebral hemispheres and is no longer photoreceptive. The central circadian clock of mammals resides within the suprachiasmatic nucleus (SCN) of the hypothalamus.

Clock genes are under study in the range of vertebrate taxa, but have been most fully studied in the mammalian SCN. Rhythmic expression of clock genes within the SCN is required for the production of circadian rhythms throughout the brain and body. Circadian clock genes also are expressed rhythmically in the cells of most tissues. Clock gene expression modulates myriad processes that oscillate daily at levels from behavior to tissue-specific cellular function. States of sleep and arousal, release of neuroendocrine hormones, autonomic control of organs, contractile strength of muscle, and production of liver enzymes all oscillate with circadian rhythmicity. Regulated clock gene expression within the SCN enables this central clock to act as the conductor to the orchestra of local body clocks. In concert, these circadian clock-driven oscillations enable the organism to anticipate predictable changes in the environment, to coordinate diverse internal processes, and to align itself adaptively with the major variable in its environment, the cycle of night and day.

Circadian rhythms emerge from cell-autonomous processes, posttranslational modifications (PTMs) of clock proteins that regulate transcriptional-translational feedback loops of clock genes. These cellular processes together form a dynamic timekeeping mechanism with a near-24 h period. Interlocking positive and negative feedback loops wherein clock proteins alternately activate or repress transcription is the apparent basis of the circadian clocks in plants and diverse animals, including mammals. Key to the oscillatory process is the tight regulation of proteins encoded by clock genes; during their lives within the circadian cycle, clock proteins associate dynamically with other clock proteins to form macromolecular assemblies, move between different subcellular compartments, and undergo extensive changes in post-translational state and stability.

The regulated expression of clock genes and proteins underlies the three primary components of the circadian system as it functions within the organism: the central clock, which generates the circadian time base in the SCN; input pathways, by which light entrains the SCN clock to the day–night cycle; and output pathways from the SCN, which transmit timing information beyond the central clock.

1. Within the central clock, clock genes and proteins form the core timekeeping mechanism. Some clock proteins are transcription factors that regulate their own transcription and that of other clock genes. Others are kinases, enzymes that add phosphate groups, reversible PTMs that profoundly alter the functioning and associations of substrate clock proteins.
2. The input pathways, which transmit signals to the SCN, adjust and align the clock's time base. This requires that incoming signaling pathways engage the molecular clockwork. Thus, neurotransmissions to the SCN initiated by light at the retina must activate signal transduction events that target clock proteins and genes.
3. The output pathways, which transmit signals from the clock, are initiated by clock proteins via effects on transcription of genes that are not part of the timekeeping mechanism. These clock-controlled genes enable the central clock to communicate time-of-day information to the brain and body. Outputs include neural and humoral signals from the SCN that are regulated by the clock and bear time-of-day information to peripheral clocks. There, clock proteins drive the expression of tissue-specific output genes throughout the genome. This enables peripheral clocks in the diverse cells of the body to regulate local circadian functions in tissue-specific ways.

This review evaluates present understanding of the expression of circadian clock genes in the mammalian SCN. This is an area of rapid flux; insights are still developing. Extrapolations have been made from the well-studied *Drosophila* model, but must be regarded cautiously, with the recognition that functional differences in some clock elements are emerging. We consider (1) putative circadian clock genes, (2) SCN clockwork mechanisms, (3) the structural and

functional features of clock proteins, (4) patterns of clock gene expression in the SCN, (5) the regulation of clock mRNA/protein expression and function, and (6) the effects of light.

Putative Circadian Clock Genes

Cross-species genomic comparisons of putative clock genes suggest striking similarities in the molecular circuits within timekeeping mechanisms across diverse circadian clockworks. The regulatory circuits are proposed to comprise interacting proteins – transcription factors together with their modifying kinases and phosphatases. These clockwork elements form similar transcriptional-translational feedback loops in phylogenetically distant organisms. Elements of such molecular circuits have been found in unicellular photosynthetic blue-green algae (*Cyanobacteria*), the bread mold (*Neurospora*), the fruit fly (*Drosophila*), the flowering plant (*Arabidopsis*), a number of rodents (rat, hamster, and mouse) and humans (*Homo*). Within animal lineages, the principal clock genes are largely conserved. Where differences appear, they can be traced deep into basal lineages; some clock genes are more similar between mouse (*Mus*) and honeybee (*Apis*) than between bee and fruit fly. This indicates that circadian clocks are ancient and that clockwork elements have been under high selective pressure to be conserved during evolution.

Circadian clock genes are identified by phenotypic profiles as those genes necessary for the generation of circadian rhythms. Key loci are defined by gene mutations, genetic deletions, or molecular manipulations that alter the period by $\geq$3 SD (standard deviations) from mean wild-type rhythms or completely abolish circadian rhythms (Table 1). Although our understanding is growing, the roles of putative clock elements in circadian rhythm generation are less well-established in mammals than in more tractable genetic models, such as *Drosophila*. Orthologs of some fly clock genes exist in mammals and probably perform similar roles in timekeeping. However, mammals may express multiple paralogs, distinct genes with high sequence similarity, probably evolved from a single ancestral gene. For example, *Drosophila Period* (*dPer*), the first animal clock gene identified, exists as a single gene, an ortholog of three paralogous mammalian *Period* genes (*Per1, 2, 3*). Of the three, *Per2* is thought to function as the core clock element, whereas *Per1* probably acts in the input pathway and *Per3* contributes to output functions. Other putative mammalian clock genes are ancestrally distant from those in the fly. New evidence indicates that mouse *Cryptochrome* (*mCry*) *and Timeless* (*mTim*) genes are divergent in sequence and expression pattern from related genes in *Drosophila* but are similar to *Apis*. These differences complicate deciphering timekeeping circuits in mammals.

A working model for the SCN clock has been developed based on overall similarities to the fly model of the structure and regulatory features of putative mammalian clock genes and proteins, their temporal expression patterns, and their functional characteristics. The circadian clock in the SCN is predicted to emerge from interactions among 10 transcription factors encoded by putative clock genes: *CLOCK* (*Clk*), *Bmal1/Mop3* (hereafter, *Bmal1*), *Period* (*Per 1, 2, 3*), *Cryptochrome* (*Cry 1, 2*), *Timeless* (*Tim*),*Rev-erbα* (orphan nuclear receptor), and *Ror* (retinoid orphan receptor). Two putative clock genes encode kinases: *casein kinase I epsilon* (*CKIε*) and *glycogen synthase kinase 3 beta* (*GSK3β*). No phosphatases have yet been identified as mammalian clock elements; however, the strong evidence for clock genes encoding protein kinases in mammals predicts that counterbalancing phosphatases also form the core clock mechanism. An additional feature is that clock elements must be expressed appropriately in temporal and spatial domains. Thus, present models are incomplete and limited by our knowledge, but they are

Table 1 Mutation phenotype and SCN expression patterns of clock genes

Gene	*Mutation phenotype*	*Peak mRNA expression*	*Peak protein expression*
Clock	Longer period (+4 h); arrhythmicity in DD	Constitutive	Constitutive
Bmal1/Mop3	Arrhythmicity in DD	CT 15–18	CT 0–8
Per1	Shorter period (0–1.1 h)	CT 4–6	CT 10–14
Per2	Shorter period (−1.5 h), arrhythmicity in DD	CT 6–12	CT 10–14
Per3	Shorter period (0–0.5 h)	CT 4–9	CT 10
Tim	Homozygous embryonic lethal; antisense ODN abolishes circadian rhythm in SCN	CT 8–12	CT 10–16
Cry1	Shorter period (−1 h)	CT 8–12	CT 12–18
Cry2	Longer period (+1 h)	CT 8–16	CT 12–16

DD, constant darkness; ODN, oligodeoxynucleotide; SCN, suprachiasmatic nucleus.

useful for conceptualizing this complex dynamic process.

Thus far, relatively few clock components have been found in mammals. We might ask if there are new clock components to be discovered: doubtless there are. New candidate clock genes are being reported as potential components of core clock mechanisms. In addition, microarray data have shown that approximately 10% (hundreds of genes) of all detectable transcripts in mice tissue express circadian patterns. Large-scale genomic approaches will likely reveal new candidate genes for testing functionally, which will no doubt cause us to revise core clock mechanisms.

Suprachiasmatic Nucleus Circadian Clockwork Mechanisms

Fundamental to circadian clocks is that different sets of clock proteins act as positive and negative regulators of clock gene expression (Figure 1). This relationship sets up positive and negative feedback loops, dynamic relationships that characterize biological oscillators of many types. Because circadian clock gene expression requires regulating multiple transcriptional, translational, posttranslational, and translocational processes, the time domain of the circadian clock is long relative to other biological oscillators.

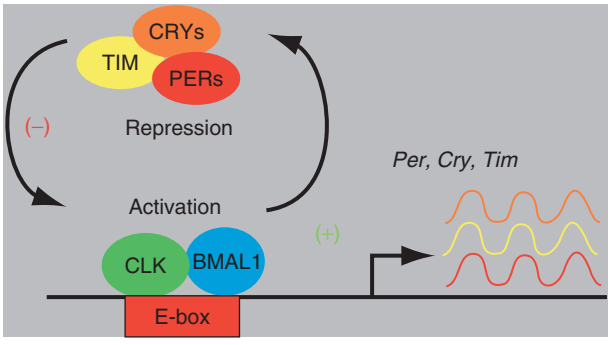

Figure 1 A model of positive and negative transcriptional feedback loops formed by clock proteins within cells of the SCN. When CLK and BMAL1 together bind to E-box sequences in promoters of other genes, they activate transcription. This forms a positive loop, enhancing transcription. In this model, they activate transcription at *Per, Cry,* and *Tim* genes producing mRNAs (small oscillatory lines). These mRNAs are translated in the cytoplasm into PER, CRY, and TIM proteins, which eventually translocate to the cell nucleus. This process is thought to be driven by posttranslational modifications (PTMs) and protein–protein interactions that stabilize the proteins and favor nuclear localization. Within the nucleus, PER, CRY, and TIM associate with the CLK-BMAL1 complex at E-boxes of the *Per, Cry,* and *Tim* gene promoters. This represses transcriptional activity at these genes, closing the negative loop of clock gene regulation. The affinity of the complex changes the association of PER, CRY, and/or TIM, which dissociate from the complex. Upon release, they may be ubiquinated and then degraded. This tips the balance toward the positive feedback loop and the cycle repeats. The period of each cycle is ~24 h.

The proposed mammalian clockwork mechanisms follow. The protein products of the putative clock genes *Clk* and *Bmal1,* CLK and BMAL1, respectively, regulate a positive feedback loop. Under appropriate conditions, these proteins dimerize, translocate to the cell nucleus, bind to E-box (CACGTG)-containing *cis*-regulatory sequences in target genes and initiate transcription. BMAL1 transcriptional activation requires a coactivator, acetylate 3. Recently, CLK itself was found to be a histone acetyl transferase. Among the genes that CLK and BMAL1 activate are the putative clock genes of the *Per* and *Cry* families. In *Drosophila,* target clock genes include *Timeless* (*dTim1*), but it is not known whether they activate mammalian *Timeless* (*mTim*).

Transcription-translation of the *Per* and *Cry* genes initiates a negative feedback loop. PER and CRY accumulate in the cell cytoplasm for part of the circadian cycle. Conformational changes induced by protein–protein interactions and/or PTMs expose nuclear localization sequences (NLS) and target the cell nucleus with these proteins. Nuclear transport/shuttling mechanisms for clock proteins are not understood. Within the nucleus, PER and CRY form multimeric complexes with CLK:BMAL1, still bound to the E-boxes in *Per* and *Cry* promoters. This leads to a reduction in histone acetyltransferase activity, which decreases histone H3 acetylation and transcriptional activation, thereby inhibiting the positive loop. This important control point leads to the cessation of production of mRNA from the *Per* and *Cry* genes, and their proteins are degraded. This closes a negative feedback loop. The loss of CRY and PER proteins poises the system to reenter the positive loop phase. TIM can interact with PER1 to inhibit CLK:BMAL1-activated transcription in cultured cells. Although there is evidence that TIM is necessary for SCN circadian rhythms, it has not yet been shown to act within the negative loop in the SCN clock.

A second bidirectional set of feedback loops is present in mammals, as in flies. These positive and negative loops center around the regulation of and by an orphan nuclear receptor, REV-ERBα. In the positive loop, CLK:BMAL1 binding at an E-box enhancer of the *Rev-erbα* gene activates transcription. The resulting protein, REV-ERBα, initiates the negative loop by binding to retinoic acid-related orphan receptor response elements (ROREs) in the promoter of *Bmal1,* and possibly *Clk* and *Cry1*, thereby repressing transcription. PER:CRY complexes also feedback negatively on the CLK:BMAL1 complex at the E-box site, inhibiting transcription of *Rev-erbα*. As levels of REV-ERBα, decrease, ROR is thought to displace

REV-ERBα on RORE sites, turning on *Bmal1* transcription. Rising levels of PER and CRY again lead to increased BMAL1 expression by repressing *Rev-erbα*, completing an oscillatory loop. These loops are thought to enhance the stability and robustness of the circadian oscillator.

Transcriptional-translational feedback loops do not alone account for an oscillatory cellular clockwork. The critical drive for the dynamics of temporal and spatial oscillation in clock-protein transcription factors resides in their PTM. PTMs can include phosphorylation/dephosphorylation, limited proteolysis, ubiquitination, sumoylation, methylation, and glycosylation. PTMs of clock proteins cause conformational changes in three-dimensional protein states that determine key events in the life of the protein: subcellular localization, intermolecular interaction, translocation, and degradation. Each clock protein is likely to have multiple effectors of PTM. PER2 is phosphorylated both by CKIε and GSK3β. Mutation within the CKIε recognition site of PER2 significantly alters circadian behavioral rhythms, including human sleep–wake cycles. GSK3β is expressed in the SCN and liver of mice, where it exhibits robust circadian oscillation in phosphorylation state, and it phosphorylates PER2 in COS1 cells. Doubtless additional clock genes will be identified that regulate PTM of other clock proteins.

Structural and Functional Features of Clock Proteins

Structural motifs and functional domains of clock proteins are substrates for protein–protein interactions, PTM, and regulation of subcellular distribution (Figure 2). PERs are the most thoroughly studied of the core clock components. In dPER, intermolecular interaction of the two PAS domains (PAS-A and PAS-B) is necessary for PER homodimer formation. All three mammalian PERs contain the PAS-A and PAS-B dimerization domains, nuclear export signal (NES) domains, and CKIε-binding motifs. Studies of a natural period variant in hamster, *tau*, found that it encodes an altered CKIε. CKIε regulates protein stability and subcellular localization of PER. Substituting serine for glycine in the CKIε substrate site of human PER2 shortens the circadian period, possibly by altering degradation or accelerating nuclear entry. CKIε actively regulates the intracellular location of PER by masking or unmasking the nuclear localization signal (NLS) domain, which is adjacent to the CKIε-binding motif in all three PERs. PERs also contain cytoplasmic localization domains (CLD), which are mediators of nucleo-cytoplasmic shuttling. CLDs of PER1 and PER2 are suggested to be nonfunctional. Indeed, mPER2 shows predominantly nuclear staining in the SCN.

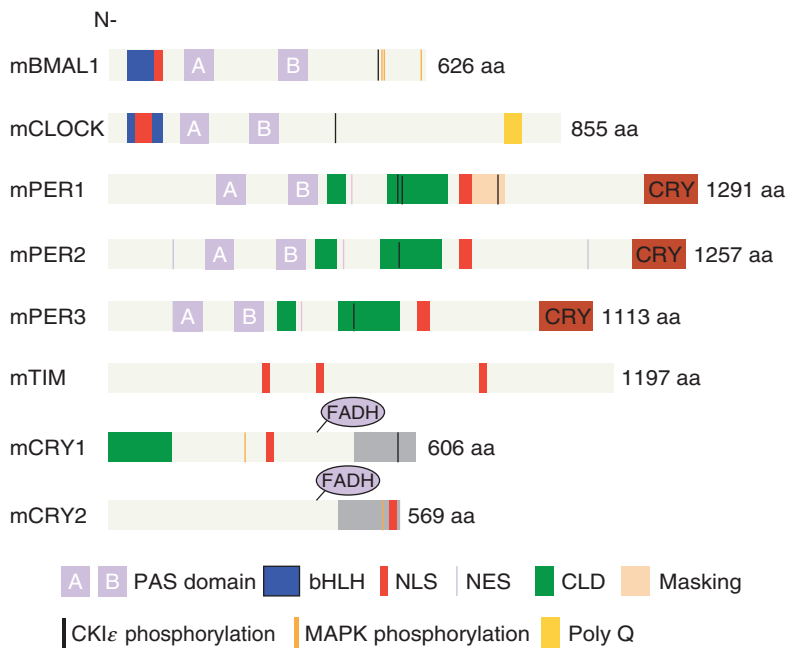

Figure 2 Schematic representation of structural and functional features of mouse clock proteins. A and B represent the PAS-A and PAS-B domains, respectively, sites of protein–protein interactions that enable the formation of dimers and macromolecular assemblies. bHLH, basic helix–loop–helix DNA-binding domains; CKI, casein kinase Iε recognition motif; CLD, cytoplasmic localization domain; CRY, cryptochrome-binding domain; NES, nuclear export signal; NLS, nuclear localization signal; PAS, Period-Arnt-Single-minded transcription factor domain; Poly Q, poly-glutamine-rich sequence.

CRY1 and CRY2 contain N-terminal chromophore-binding domains but exhibit structural divergence in C-terminal extension domains. This suggests that C-terminal domains have unique functions in each CRY. CRY acts as a repressor by directly interacting with the CLK-BMAL1 heterodimer in the nucleus. Thus, the nuclear localization and interaction with this heterodimer are essential for CRY function. An NLS has been found in the N-terminal chromophore-binding sequence in CRY1 and a functional NLS was identified in the C-terminal extension domain in CRY2. How CRY localization is regulated is unknown, but it is likely that PER-binding unmasks the NLS or masks the CLD of CRY.

Intracellular distribution and PTM are also affected by the interactions of PERs with CRYs. The two PAS domains are required for PER–PER and PER–CRY dimerization. PER–CRY dimerization appears important for the nuclear accumulation of both proteins; *Cry1*/*Cry2* double-mutant mice show weak nuclear localization of PER1 and PER2. Moreover, PER–CRY dimerization prevents their degradation through the ubiquitin–proteasome pathway.

CLK and BMAL1 proteins each have one basic helix–loop–helix (bHLH) DNA-binding domain and two PAS dimerization domains. CLK and BMAL1 also contain conserved NES and NLS domains, adjacent to the PAS-A and bHLH domains, respectively. CLK may be regulated by protein kinase G type II (PKG-II) at the end of the night, when they co-immunoprecipitate from SCN extracts. BMAL1 is phosphorylated *in vitro* by two kinases, CKI and mitogen-activated protein kinase (MAPK). MAPK phosphorylates Ser-527, Thr-534, and Ser-599 in BMAL1. A conserved CKI substrate consensus motif exists in the C-terminus of BMAL1, but phosphorylated residues have not been determined. These two kinases may compete to regulate BMAL1 function because their phosphorylation motifs overlap. Transactivation studies support this hypothesis – MAPK-mediated phosphorylation decreases CLK–BMAL1 transcriptional activity, whereas CKI-mediated phosphorylation induces it.

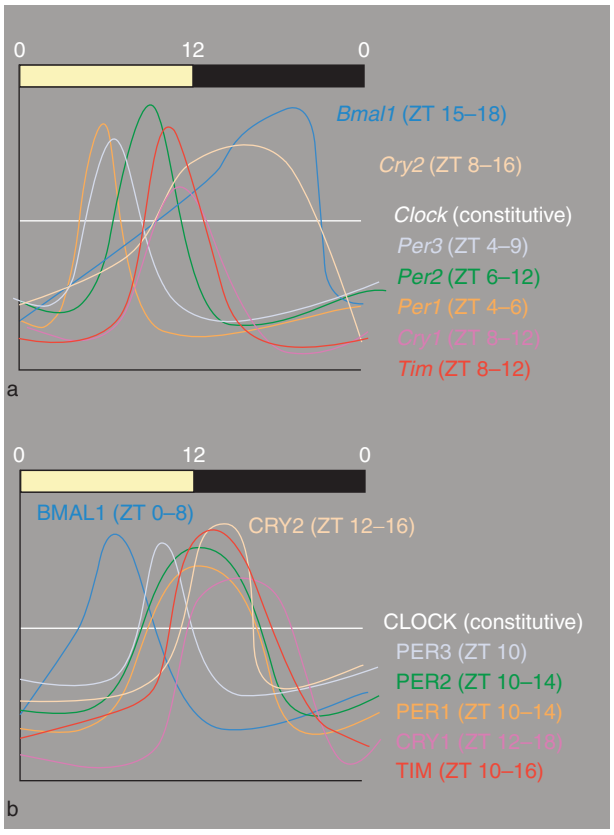

Figure 3 Schematic of circadian expression patterns in the SCN: (a) clock gene mRNAs; (b) clock proteins. Relative mRNA and protein levels of clock gene transcription factors oscillate over the circadian cycle. An exception is those products of the gene *Clock*, which are constitutive. The mRNAs of negative elements are generally high in the daytime; transcripts of *Bmal1*, a positive element, peak in the nighttime. The mRNAs of most clock genes are nearly immediately translated, so the oscillations in clock proteins are offset from mRNA rhythms by ~4 h. BMAL1 peaks midday, whereas proteins of negative elements peak from late day to early night. Horizontal bars: yellow indicates daytime (ZT 0–12 h); black indicates nighttime (ZT 12–24 h).

Patterns of Clock Gene Expression in the Suprachiasmatic Nucleus

Temporal patterns of putative clock mRNA and protein expression have characteristic relationships to phases of the day–night cycle (Figure 3). Most show oscillatory circadian patterns that map to specific phases of the ~24 h cycle. Time of the SCN clock in animals under light–dark conditions is expressed as zeitgeber time (ZT) because the regular daily appearance of light acts as a time-giver. Subjects usually are entrained to a 24 h cycle with 12 h of light (day) and 12 h of darkness (night). Under constant environments, including darkness, phase-relationships persist. The period of these clock-driven oscillations is *circa*, not exactly, 24 h. Reckoning circadian time (CT) includes correcting for the non-24-h period.

Proteins of the positive loop exert primary influence during the daytime. In anticipation, *Bmal1* mRNA increases during the nighttime (CT 15–18), peaks just before dawn (CT 23), and then drops sharply. BMAL1 protein is present in the first half of the day (CT 0–8). *Clk* mRNA and protein are constitutively expressed in the SCN. For CLK and BMAL1 to dimerize, they need to be in the same physical space. Dimerization dynamics are unknown; however, shortly after BMAL1 starts to accumulate, mRNAs of the elements of the negative loop appear (Figure 3). This is

consistent with a model wherein CLK–BMAL1-driven transcription forms the positive loop, activating the transcription of clock genes in the negative loop (Figure 1).

Daytime activation of transcription of elements of the negative loop imposes critical negative feedback that enables the oscillation. The transcription of *Per, Cry,* and *Tim* peaks during the mid- to late circadian day, in antiphase to the *Bmal1* peak (Figure 3). *Per* mRNAs and proteins exhibit robust circadian patterns, with protein lagging mRNA rhythms by ~4 h. Proteins of the negative loop – PER 1, 2, and 3; CRY 1 and 2; and possibly TIM – appear in the late day, with most peaking near the day–night transition. PER-CRY heterodimers are reported to accumulate in the nucleus at night, when they are predicted to interact with CLK–BMAL1 to repress *Per* and *Cry* transcription. However, most *Per* and *Cry* mRNA appears to decrease in the second half of the day. This inconsistency has not been resolved, but it may reflect sampling methods.

Within the SCN, clock gene expression varies regionally. Regions are distinguished as (1) the ventromedial core, receiving inputs from the retina and other afferent regions and expressing vasoactive intestinal peptide (VIP); and (2) the dorsomedial shell, projecting from SCN to the nearby hypothalamus and expressing arginine vasopressin (AVP). In mouse, *Per1* mRNA undergoes circadian oscillation in the shell region, preceding shell *AVP* mRNA rhythms by ~4 h. PER2 rhythms in diurnal rodents peak between ZT 10–18 in the dorsal, ventral, and lateral regions but not centrally. At trough times, PER2 is primarily in the dorsal SCN. PER1 is more widely distributed at its peak, but, as with PER2, it is less dense in the core.

Regulation of Clock mRNA and Protein Expression and Function

Dynamic regulation of clock gene mRNA and protein state is necessary to generate ~24 h oscillations. Regulation of mRNA may include RNA processing (capping, splicing, and polyadenylation), export, decay, and translation, as well as alternative transcriptional start sites. Changes in RNA processing could change clock state by altering the RNA expression pattern or RNA stability. Multiple transcriptional initiation sites and splice variants have been identified for mammalian *Bmal1* and *Bmal2*. Human BMAL2 isoforms exhibit differential transcriptional activity *in vitro*, suggesting that alternative splicing may regulate clock function. In addition, *Tim* may have multiple splice forms because multiple transcript and protein species have been reported.

Protein state is regulated by PTM and protein–protein interactions. Of possible PTMs that can regulate proteins, phosphorylation has been studied most. Changing the equilibrium between phosphorylation–dephosphorylation states may be a key determinant of clock proteins by altering (1) stability in the cytoplasm, (2) transcriptional activator-repressor activity, and (3) nuclear entry. CKIε–mediated phosphorylation of PER2 affects its stability. The phosphorylation state increases PER turnover by altering the rate of ubiquitination, which targets proteins for degradation. In addition to stabilizing clock proteins, phosphorylation is important for regulating transcriptional activator–repressor activity. In cultured mammalian cells, repressing CKI expression decreases BMAL1-dependent transcriptional activity at the *Per1* promoter. MAPK also alters the BMAL1-dependent transcription. Not surprisingly, dephosphorylation has been reported to impact activity of dPER; increased phosphatase PPA2 activity leads to stable, constitutive nuclear expression, but has not been examined in mammals. Nuclear localization of PER2 is affected by CKIε-mediated phosphorylation. In COS1 cells, GSK3β also promotes nuclear translocation of PER2. Thus, these kinases may converge on PER2 in regulating nuclear translocation.

The timing of phosphorylation and oscillating protein–protein interactions shape the period length of circadian clock. PER proteins show similar circadian phosphorylation patterns, with peak phosphorylation levels near CT 21, when most of the PER is nuclear. Phosphorylated CLK is detectable in the nuclear fraction throughout the circadian cycle, with hypophosphorylated forms in the nucleus during the day. A similar pattern of phosphorylation is detected for BMAL1. At the time of negative transcriptional feedback, phosphorylated forms of mammalian clock proteins form complexes in the nucleus, but the significance of nuclear phosphorylated forms of clock proteins is unclear. CKIε is at least partly involved in regulating the nuclear stability of PER proteins and the transcriptional activity of BMAL1.

Protein–protein interactions can regulate receptiveness to PTMs that alter clock protein localization and state. PER appears essential to controlling the subcellular distribution and CKI-dependent phosphorylation of CRYs. CKI and CRYs do not initially bind directly. Rather, both bind to PER simultaneously to form a ternary complex in which CKI can phosphorylate CRY1 or 2. Additional PTMs may be contributed by other kinases. A recent report showed that MAPK directly interacts with and phosphorylates CRY1 and CRY2 *in vitro*, and suggested that MAPK-mediated phosphorylation reduces the repressor activity of CRY.

Regulation of Clock mRNA and Protein Expression by Light

The cell-autonomous SCN clock repeats each cycle in about 1 day. Over successive cycles, the circadian clock drifts away from the 24 h solar cycle on earth. When this desynchronization causes daytime light signals to impinge on the nighttime state of the SCN, the time base of the clock adjusted, or reset, realigning the internal state with the external environment. This process requires that the signaling pathways that transduce neurochemical signals of light from the retina to the SCN engage the molecular clockwork. Light signals in early and late night cause the delay and advance of clock phasing, respectively. In rats, *Per1* mRNA increases in the SCN at 60 min after the light pulse at either time. *Per2* mRNA induction can be observed after light exposure at early, but not late, night. Functionally, light-induced phase delays are attenuated by the co-injection of antisense oligodeoxynucleotide (αODN) to mRNA for *Per1* or *Per2*, whereas only *Per1* αODN blocks phase advance. This indicates that the expression of both these genes is necessary for phase delay of the clock mechanism, but only *Per1* is necessary for phase advance.

Nocturnal light exposure causes changes in PER levels in mouse SCN. PER1 and PER2 increase within 4–6 h after brief light pulse in the early night, when light signals delay clock phase, effectively extending the clock reckoning of daytime. However, during late night, only PER1 expression increases significantly. Light in late night signals an early morning, advancing the clock phase. PER1 peaks ~4 h after this light signal and also early in the next cycle, at 10 h post-stimulus, rather than the usual time, ~CT 12. This interval reflects the amplitude of the advance in clock phase. These data support the hypothesis that both PER1 and PER2 are important for phase delay, whereas only PER1 contributes to phase advances. However, questions remain as to how the light-induced increase in PER affects the negative transcriptional loop and what roles other clock proteins might play.

Light-stimulated changes in regional distributions of PER1 and PER2 within the SCN follow their respective mRNAs and depend on the circadian time of light pulse administration. After a phase-delaying light pulse in early night, the number of PER1-positive cells increases primarily in the ventrolateral core region, whereas PER2-positive cells show greatest increase in the shell. After a phase-advancing light pulse, PER1 expression increases in both the core and shell regions, without significant PER2 increase in either core or shell region. Thus, increased PER2 expression in the shell SCN region correlates with phase delay, whereas the induction of PER1 in the shell correlates with phase advance. Effects of light on other clock elements have not been established.

Summary

Circadian rhythms in mammals are orchestrated by a central clock in the suprachiasmatic nucleus (SCN). The clock emerges from cell-autonomous processes. Clock proteins cause the rhythmic expression of clock genes as well as output genes. Structural features and posttranslational modifications of clock proteins determine their intermolecular associations, subcellular localization, transcriptional activity, and stability. Some clock proteins positively and negatively regulate their own rhythmic expression via feedback on transcription. Patterns of clock gene expression within SCN cells and tissue are complex. The understanding of mammalian clock mechanisms critical to timekeeping is as yet incomplete.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Clock Gene Regulation of Endocrine Function; Clock Genes and Metabolic Regulation; Genetic Regulation of Circadian Rhythms in Drosophila; Sleep and Waking in Drosophila.

Further Reading

Antle MC and Silver R (2005) Orchestrating time: Arrangements of the brain circadian clock. *Trends in Neuroscience* 28: 145–151.

Barnes JW, Tischkau SA, Barnes JA, et al. (2003) Requirement of mammalian *Timeless* for circadian rhythmicity. *Science* 302: 439–442.

Bell-Pedersen D, Cassone VM, Earnest DJ, et al. (2005) Circadian rhythms from multiple oscillators: Lessons from diverse organisms. *Nature Reviews Genetics* 6: 544–556.

Eide EJ, Kang H, Crapo S, Gallego M, and Virshup DM (2005) Casein kinase I in the mammalian circadian clock. *Methods in Enzymology* 393: 408–418.

Hardin PE (2004) Transcription regulation within the circadian clock: The E-box and beyond. *Journal of Biological Rhythms* 19: 348–360.

Hirayama J and Sassone-Corsi P (2005) Structural and functional features of transcription factors controlling the circadian clock. *Current Opinion in Genetics & Development* 15: 548–556.

Lakin-Thomas PL (2006) Transcriptional feedback oscillators: Maybe, maybe not. *Journal of Biological Rhythms* 21: 83–92.

Looby P and Loudon AS (2005) Gene duplication and complex circadian clocks in mammals. *Trends in Genetics* 21: 46–53.

Lowrey PL and Takahashi JS (2004) Mammalian circadian biology: Elucidating genome-wide levels of temporal organization. *Annual Review of Genomics and Human Genetics* 5: 407–441.

Okamura H (2004) Clock genes in cell clocks: Roles, actions, and mysteries. *Journal of Biological Rhythms* 19: 388–399.

Roenneberg T and Merrow M (2005) Circadian clocks: Translation. *Current Biology* 15: R470–R473.

Rubin E, Shemesh Y, Cohen M, et al. (2006) Molecular and phylogenetic analyses reveal mammalian-like clockwork in the honey bee (*Apis mellifera*) and shed new light on the molecular evolution of the circadian clock. *Genome Research* 16(11): 1352–1365.

Saper CB, Lu J, Chou TC, and Gooley J (2005) The hypothalamic integrator for circadian rhythms. *Trends in Neuroscience* 28: 152–157.

Takahashi JS (2004) Finding new clock components: past and future. *Journal of Biological Rhythms* 19: 339–347.

Tischkau SA, Mitchell JW, Pace LA, et al. (2004) Protein kinase G type II is required for night-to-day progression of the mammalian circadian clock. *Neuron* 43: 539–549.

Toh KL, Jones CR, He Y, et al. (2001) An *hPer2* phosphorylation site mutation in familial advanced sleep phase syndrome. *Science* 291: 1040–1043.

Circadian Genes and the Sleep–Wake Cycle

A D Laposky and F W Turek, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

There are two major physiological processes that interact to determine the need for sleep and to direct the temporal organization of sleep–wake states across the day and night. A homeostatic mechanism (process S) governs the propensity for sleep based on the amount of prior wakefulness (i.e., sleep–wake history); the longer an organism is awake, the more likely it will be to fall asleep and to increase sleep intensity. The circadian drive (process C) determines the timing of when sleep and wakefulness occur based on the coordination of an endogenous circadian clock system and entrainment to environmental cues, such as the light–dark cycle. The importance of homeostatic and circadian processes on sleep–wake organization has been demonstrated in organisms as diverse as humans, rodents and flies. Because the functions of the homeostatic and circadian processes are generally conserved, researchers have been able to take advantage of genetic animal models (e.g., mice and flies) to investigate the molecular mechanisms of sleep–wake regulation, as well as the genetic basis of human sleep–wake disorders. One of the interesting advances in this line of research has been the demonstration that circadian clock genes have unique roles in sleep regulation that go beyond just the timing of sleep. In this article, we focus on what is known about the involvement of mammalian circadian clock genes in the control of sleep amount, architecture, and homeostatic drive.

Systems-Level Studies: Suprachiasmatic Nucleus Lesions

The suprachiasmatic nucleus (SCN) consists of a small group of neurons situated just above the optic chiasm in the anterior hypothalamus. This structure represents the major circadian pacemaker in the brain, and its activity underlies the generation and/or coordination of virtually all behavioral and physiological rhythms of the body. It can be argued that the sleep–wake cycle is the master output rhythm of the SCN because most biological functions vary according to whether the organism is asleep or awake. Because of the important role of the SCN and the sleep–wake cycle in daily functioning, there has been an interest in understanding exactly how the SCN is involved in regulating sleep and wakefulness and what aspects of sleep (i.e., timing, quantity, and electroencephalogram (EEG) homeostatic drive) are under circadian control. The complexity of deciphering this role of the circadian system is that the propensity for sleep is also mediated by another factor that is oscillatory in nature – the sleep homeostatic process. Although the circadian and sleep homeostatic processes are essentially linked at the functional level, various experimental paradigms in humans and rodents have been developed to examine their independent roles. Early studies in rats used experimental SCN lesions to remove the influence of the central anatomical circadian pacemaker. These studies demonstrated that SCN lesions eliminated the diurnal sleep–wake rhythm but did not affect overall sleep–wake amount or the homeostatic response (i.e., EEG non-rapid eye movement (NREM) delta [0.5–4.0 Hz] power) to sleep deprivation. These results have contributed to the view that the SCN is involved only in regulating the placement of sleep across the light–dark or circadian cycle. The broader implication of this view is that the circadian system does not play an active role in generating sleep–wake states or sleep homeostatic drive and that there are no common mechanistic links between circadian and sleep–wake regulatory processes.

Following these initial studies in a single species, the laboratory rat, a few subsequent studies involving SCN lesions provided contrasting results to these earlier findings. The first demonstration that SCN lesions could affect sleep time was in the diurnal squirrel monkey, which had a loss of the sleep–wake rhythm but a large increase in NREM sleep time. This finding by Edgar and colleagues established a basis for the opponent-process model, which proposed that the SCN produced a wake-promoting signal across the active phase of the circadian cycle. This major drive for wakefulness could then initiate or contribute to the buildup of sleep homeostatic pressure, eventually leading to sleep initiation near the onset of the major rest period. Another important outcome of the opponent-process model is that it established a foundation for other interesting hypotheses about SCN function, such as that the SCN may also provide a sleep-promoting signal during the rest period that contributes to the dissipation of sleep drive and the cyclical increase in wake propensity. More recent experiments in rats and mice found that SCN lesions not only abolished the sleep–wake rhythm but also increased NREM sleep time, providing further support for the basic tenants of the opponent-process model. Another study in Siberian hamsters showed that following a 5 h phase shift, animals became

arrhythmic accompanied by a large increase in sleep time and an impaired homeostatic response to sleep deprivation. Certainly, it will be important in future studies to determine the extent to which these limited sets of data are replicable in different strains of mice and rats and are generalizable across different nocturnal and diurnal species. It will also be important to recognize that, in addition to the use of EEG NREM delta power as a measure of sleep homeostatic drive, sleep–wake amount and sleep architecture are other variables that are important components of the sleep–wake phenotype. Although the SCN lesion studies do not delineate a specific mechanistic link between circadian and sleep regulatory systems, they do provide important evidence that the circadian system can feed into and cause adjustments in sleep regulatory processes. These studies also established a foundation for interesting hypotheses to be tested regarding what these mechanisms could be, particularly at the genetic level.

Genetics-Level Studies: Mammalian Circadian Clock Genes

The ability of the SCN to entrain to the light–dark cycle and to generate endogenous rhythms in the absence of environmental cues is based on molecular events involving an integrated circuit of circadian clock genes that exists in individual neurons of the SCN. This group of core circadian genes forms a complex transcriptional–translation feedback loop within each cell, which cycles with a near 24 h period. The self-oscillating neurons in the SCN synchronize as a functional unit to create major output signals through both neural transmission and hormonal factors that drive a wide range of output behaviors and physiological processes. One of the most exciting recent discoveries in the circadian field is that the transcriptional–translational feedback loop of circadian clock genes also exists outside of the SCN, in other areas of the brain, as well as in various peripheral tissues (i.e., the lung, heart, muscle, liver, and pancreas). The discovery of circadian clock genes and the development of animal genetic models (see Table 1) have played a crucial role in understanding how the circadian system controls a wide range of output rhythms and have elucidated other potentially important roles of the circadian system that extend well beyond its role as a time keeper.

Clock, Bmal1, and *Npas2*

The first mammalian circadian clock gene was discovered in a forward mutagenesis program at North western University and was given the name *Clock (Circadian locomotor output cycles kaput)*. The *Clock* gene encodes a basic helix–loop–helix (bHLH)-period-arnt-single-minded (PAS) domain-containing transcription factor that is a component of the positive limb of the transcriptional–translational feedback loop of circadian genes. The protein products of *Clock* and another bHLH-PAS circadian gene, *Bmal1* (*Brain and muscle arnt-like 1*, also called *Mop3*) dimerize (CLOCK–BMAL1) to drive the transcription of negative-limb circadian genes, *cryptochrome* and *period*, which in turn feed back to inhibit CLOCK–BMAL1 activity. Mice that harbor a homozygous mutation in the *Clock* gene, the *Clock/Clock* genotype, entrain to a light–dark cycle with respect to locomotor activity, but have an altered circadian period of activity in constant darkness; initially they free-run with a notably long period (~28 h) and then become arrhythmic. Unexpectedly, when sleep–wake patterns were examined, our research group found that *Clock/Clock* mice exhibit a high amount of wake time (+1–2 h day^{-1} compared to wild-type mice) with a reciprocal decrease in NREM sleep time under both 12 : 12 light–dark and constant dark conditions. In the light–dark cycle, the increase in wake time occurs throughout the day, but is more pronounced during the light phase, the time of day when rodents normally have the most consolidated sleep. Furthermore, following short-term sleep deprivation, the homozygous *Clock* mutants recover less REM sleep time than wild-type mice. In combination, these data show that the increase in wake time is not simply a masking effect of the light–dark cycle or an indirect response to a change in circadian period but is a fundamental effect of the mutation. Most importantly, these data represent the first demonstration that an alteration in the function of a core circadian clock gene could have a direct impact on the amount of sleep produced and on the recovery response to sleep loss and opened the door to further inquiries into how circadian clock genes might influence the sleep–wake cycle at many levels.

Because of the remarkable sleep phenotype generated by *Clock* mutant mice, it was of interest to examine sleep in mice with a deletion of *Clock's* dimeric partner, *Bmal1*. Interestingly, we found that *Bmal1*$^{-/-}$ mice produce a high amount of sleep (+1.5 h day^{-1}) compared to their intact controls; a phenotype opposite the *Clock/Clock* mouse. The extra sleep in *Bmal1*$^{-/-}$ mice is achieved entirely during the 12 h dark phase, which is normally the major wake period for rodents. Despite high sleep time, *Bmal1*$^{-/-}$ animals exhibit a severely fragmented sleep pattern throughout the entire light–dark cycle with frequent arousals from sleep and less consolidated bouts of sleep and wakefulness. *Bmal1*$^{-/-}$ mice recover less sleep time in response to sleep

Table 1 Sleep–wake patterns in animal models of circadian clock gene disruption

	Baseline 24 h wake percentage	*Baseline 24 h NREM percentage*	*Baseline 24 h REM percentage*	*Baseline sleep fragmentation*	*Baseline NREM delta activity*	*Recovery response NREM delta activity*[a]	*Source*[b]
Cl/Cl	↑	↓	–	–	–	–	[5]
Npas2	↑	↓	↓				[1]
Npas2	↑	↓	–	–	–	↓	[2]
Bmal1/Mop3	↓	↑	↑	↑	↑	↓	[4]
Cry1,2	↓	↑	↑	↓	↑	↓	[7]
Per1	–	–	–		–	↓	[3]
Per1	–	–	–		–	↑	[6]
Per2	–	–	–		–	↓	[3]
Per2	–	–	–		–	–	[6]
Per3	–	–	–		–	–	[6]
Per1,2	–	–	–		–	↑	[6]

Circadian clock mutant and deficient mice compared to wild-type mice. Empty cells indicate that information was not presented for the respective trait or could not be clearly interpreted. Many different sleep-deprivation paradigms and data analysis formats were used in these studies, making it difficult to directly compare the recovery sleep patterns among many of the animal models. ↑, increase; ↓, decrease; –, no change; EEG, electroencephalogram; NREM, non-rapid eye movement sleep; REM, rapid eye movement sleep.

[a]In the first few hours after sleep deprivation; recovery response in NREM EEG delta power is the standard measurement of sleep homeostatic drive.

[b]Data from:

1. Dudley CA, Erbel-Sieler C, Estill SJ, et al. (2003) Altered patterns of sleep and behavioral adaptability in NPAS2-deficient mice. *Science* 301: 379–383.
2. Franken P, Dudley CA, Estill SJ, et al. (2006) NPAS2 as a transcriptional regulator of non-rapid eye movement sleep: Genotype and sex interactions. *Proceedings of the National Academy of Sciences of the United States of America* 103: 7118–7123.
3. Kopp C, Albrecht U, Zheng B, and Tobler I (2002) Homeostatic sleep regulation is preserved in Per1 and Per2 mutant mice. *European Journal of Neuroscience* 16: 1099–1106.
4. Laposky A, Easton A, Dugovic C, et al. (2005) Deletion of the mammalian circadian clock gene BMAL1/Mop3 alters baseline sleep architecture and the response to sleep deprivation. *Sleep* 28(4): 395–409.
5. Naylor E, Bergmann BM, Krauski K, et al. (2000) The circadian clock mutation alters sleep homeostasis in the mouse. *Journal of Neuroscience* 20: 8138–8143.
6. Shiromani PJ, Xu M, Winston EM, et al. (2004) Sleep rhythmicity and homeostasis in mice with targeted disruption of mPeriod genes. *American Journal of Physiology* 287: R47–R57.
7. Wisor JP, O'Hara BF, Terao A, et al. (2002) A role for cryptochromes in sleep regulation. *BMC Neuroscience* 3: 20 [online].

deprivation and show an attenuated rebound in NREM EEG delta (0.5–4.0 Hz) power, the prototypic measure of sleep homeostasis. Therefore, many components of sleep drive, including sleep time, sleep consolidation, and the compensatory response to sleep loss, are disrupted in *Bmal1*$^{-/-}$ animals. The sleep–wake rhythm is attenuated in *Bmal1*$^{-/-}$ mice across the light–dark cycle and is arrhythmic in constant darkness, but the sleep phenotype remains largely similar in both conditions. It is most interesting to note that although both *Clock* mutant and *Bmal1*$^{-/-}$ mice are arrhythmic in constant darkness, their basic sleep phenotypes – high-wake and high-sleep, respectively – are opposite. This pattern suggests that the changes in their sleep regulatory process are not simply the result of general circadian disruption but are due to a unique function of each of the circadian genes independent of the circadian timing system.

It may seem surprising that sleep regulation in *Clock/Clock* and *Bmal1*$^{-/-}$ mice is quite different, given that these genes are dimeric partners in their function as circadian transcription factors. One possible explanation for the difference in sleep patterns is that the *Clock* mutant mouse harbors a mutation leading to altered CLOCK protein function, whereas *Bmal1*$^{-/-}$ is a gene deletion model. Aside from this basic difference, it is plausible that *Clock* and *Bmal1* exert unique effects on other molecular pathways and multiple physiological systems (e.g., energy metabolism, activity levels, and endocrine function) that may feed back to modulate sleep–wake behavior. For example, a recent study by Turek and colleagues demonstrated that *Clock/Clock* mice develop increased adiposity and symptoms of the metabolic syndrome. *Bmal1*$^{-/-}$ mice may have altered glucose homeostasis and develop joint disease, which leads to abnormalities in locomotor activity. The *Bmal1* gene even plays a role in molecular pathways related to adipogenesis. It will be important to gain an understanding of the multiple molecular pathways in which *Clock* and *Bmal1*, as well as other circadian genes, are involved that may not be directly related to core circadian function. Because clock genes have been

identified in a variety of central and peripheral tissues, there is probably anatomical specificity to their activity at the physiological and molecular levels. Another important pursuit will be to identify the specific brain regions and/or peripheral tissues that are responsible for the effects of *Clock* and *Bmal1*, as well as other circadian genes, on sleep and wakefulness.

Npas2 (*neuronal PAS domain protein 2*) is a paralog of the mammalian *Clock* gene, dimerizes with *Bmal1* to encode a circadian clock transcription factor, and is sensitive to the redox state of the cell, indicating that it may represent a molecular link between energy homeostasis and circadian processes. In contrast to *Clock*, *Npas2* is not expressed in the SCN, but is present in other areas of the brain and shows expression in the cortex, basal ganglia, and limbic structures. The first sleep study in *Npas2*$^{-/-}$ mice provided evidence of an increase in locomotor activity and wake time, particularly during the middle of the dark period. Generally, mice demonstrate a consolidated period of wakefulness just after lights out that lasts for a number of hours. In the middle of the dark phase, mice achieve some sleep, often referred to as a napping period. *Npas2*$^{-/-}$ mice show a delay in the onset of napping, indicating that their nocturnal drive for wakefulness is stronger than animals with an intact *Npas2* gene. A second, more comprehensive study in *Npas2*$^{-/-}$ animals replicated the high-wake pattern and found additional differences in the knockout mice, including alterations in the power density of specific EEG frequency bands and attenuated recovery responses to acute sleep deprivation. The high-wake phenotype in *Npas2*$^{-/-}$ animals is in the same direction as *Clock/Clock* mutant mice, except that *Clock* mutants show a large amount of wakefulness in the light period, whereas extra wake in *Npas2*$^{-/-}$ mice is restricted to the dark phase. Because the anatomical distribution of NPAS2–BMAL1 is different than CLOCK–BMAL1, these data raise the interesting possibility that an SCN-independent role of circadian clock genes is involved in many aspects of sleep regulation. Significantly, a recent study in our laboratory showed that when the SCNs of *Clock/Clock* mice were lesioned, the mice still had more wake time than SCN-lesioned wild-type mice. These data provide evidence for the possibility that the role of *Clock* in mediating wakefulness is occurring in extra-SCN or peripheral sites.

Cryptochrome and *Period*

The CLOCK–BMAL1 and NPAS2–BMAL1 protein complexes drive the transcription of other clock genes, including two *cryptochrome* (*Cry1* and *2*) and three period (*Per1*, *2*, and *3*) genes, which are components of the negative-limb of the circadian transcriptional–translational feedback loop. The *Cry* genes belong to the family of blue-light photoreceptors and are located in retinal ganglion cells, in the SCN, and in various peripheral tissues. The protein products of *Cry1* and *2* form heteromultimeric complexes with PER1, 2, and 3, which translocate to the nucleus to influence the transcriptional properties of CLOCK–BMAL1 and NPAS2–BMAL1. The *Cry* and *Per* genes are thought to play an important role in the ability of the molecular circadian clock to respond to light and in mediating the phase relationship between behavioral rhythms and the light–dark cycle. When the activities of these genes are altered through either mutation or deletion, a variety of circadian abnormalities ensue, including altered entrainment to a light–dark cycle, changes in the free-running period of activity, or even complete loss of circadian rhythmicity. In addition to these varied roles in circadian regulation, a variety of interesting sleep–wake phenotypes have been reported in *Cry1,2*-deficient (*Cry1,2*$^{-/-}$) and *Per* mutant mice.

It is known that *Cry1,2*$^{-/-}$ mice have an attenuated sleep–wake rhythm under the light–dark cycle and have no circadian rhythm of sleep in constant darkness. In addition to the alterations in sleep–wake distribution, there are major changes in sleep time and architecture in these mice. *Cry1,2*$^{-/-}$ mice generate a large amount of NREM and rapid eye movement (REM) sleep time ($\sim$1.5 h more than the wild-type mice amount) throughout the dark phase of the diurnal cycle. In the light phase, NREM sleep is comparable between the genotypes, but REM sleep time is lower in *Cry1,2*$^{-/-}$ mice. Interestingly, *Cry1,2*$^{-/-}$ mice exhibit a notable increase in the duration of individual sleep bouts and have enhanced sleep intensity (i.e., NREM EEG delta power) throughout the light–dark cycle. These data indicate that under baseline conditions, the deletion of both *Cry* genes leads to an enhanced homeostatic drive for sleep. These same measures of increased sleep propensity occur when the mice are arrhythmic in constant darkness – this finding is important because locomotor activity in *Cry1,2*$^{-/-}$ mice is particularly sensitive to the masking effects of light. Therefore, the maintenance of high sleep time in constant darkness suggests that the extra sleep is not simply due to behavioral suppression by light. Following short-term sleep deprivation (i.e., 6 h during the first half of the light phase), *Cry1,2*$^{-/-}$ mice fail to achieve a robust compensatory or homeostatic response in NREM, REM, and NREM EEG delta power. One explanation for this surprisingly attenuated recovery could be that baseline levels of sleep drive were already high and the mice could not generate additional sleep or sleep

intensity, often referred to as a ceiling effect. However, *Clock/Clock* and *Npas2*$^{-/-}$ mice have lower amounts of baseline sleep time and also show smaller responses to sleep loss on various measures of sleep recovery, a pattern that cannot be explained by a ceiling effect. In combination, these results (also see the description of *Per* mutant mice next) indicate that clock genes may exert different effects on sleep under different physiological conditions (i.e., baseline vs. sleep deprivation). It is interesting to note that single *Cry1* and *Cry2* deletion models have been developed that show short and long free-running rhythms of locomotor activity, respectively. Sleep has not been recorded in these mice, and it will be interesting in future studies to determine whether these mice have unique sleep–wake patterns beyond the expected alterations in timing.

Another result in the *Cry1,2*$^{-/-}$ study is that *Per1* and *Per2* mRNA levels are increased in the cortex of sleep-deprived wild-type mice and return to baseline levels following a sleep recovery opportunity. *Per* transcript levels are also high in *Cry1,2*$^{-/-}$ mice under baseline conditions, suggesting that the *Per* gene may have a role in promoting sleep. Other studies have demonstrated that *Per* expression in the cortex is associated with wakefulness and arousal, independent of its expression pattern in the SCN. Two papers were published by separate research teams on sleep–wake patterns in *Per* mutant mice. In the first study, Kopp and colleagues found that PER1-deficient mice have increased wake time throughout the dark period and increased NREM sleep time at the light–dark transition period (i.e., dusk). *Per2* mutant mice exhibited increased wake time during the last 3 h of the light phase (i.e., dawn), consistent with their phase-advanced circadian phenotype (i.e., waking up too early), and had normal sleep–wake times throughout the dark phase. In neither group was the amount of 24 h baseline sleep or wake significantly different from wild-type levels, and no major changes in EEG power spectra were observed. In the second study, by Shiromani and colleagues, sleep in PER1-deficient and *Per2* mutant mice was examined with the addition of PER3-deficient and double *Per1/Per2* mutant mice. Based on the results from this group's genetic models, there was little effect on sleep due to PER1 or PER3 deficiency. On the other hand, *Per2* mutant mice had increased wake time at the end of the light phase, and this pattern was even more pronounced in *Per1/Per2* mutant mice. The altered phase angle of entrainment in *Per2* and *Per1,2* mutant mice is consistent with human studies that have shown an association between alleles in the *Per2* gene and advanced sleep-phase syndrome.

There are a number of changes in the response to sleep deprivation in *Per* mutant mice. In the Kopp study, 6 h of sleep deprivation occurred during the first half of the light phase. Both PER1-deficient and *Per2* mutant mice began to generate a positive rebound in NREM and REM sleep time sooner than wild-type mice, and *Per2* mice had the largest cumulative rebound in REM sleep time of the three groups. PER1-deficient and *Per2* mutant mice produced a significant homeostatic response (i.e., NREM EEG delta power) after sleep deprivation; however, the magnitude of this response was slightly less than wild-type levels in the first 2 h of recovery. Also, NREM delta power remained elevated for 1 h longer in *Per2* mutant mice compared to the other groups. In the Shiromani study, sleep deprivation was placed during the last 6 h of the light period and recovery sleep began at dark onset. At this time of day, *Per3* mutant, PER1-deficient, and *Per1/Per2* mutant mice exhibited longer-lasting recovery responses in EEG NREM delta power (i.e., sleep intensity) than wild-type mice. These results indicate that the rate at which sleep pressure is dissipated may be different among these genetic models. The amplitude (percentage of baseline) of this rebound response was particularly high in PER1-deficient and *Per1/Per2* mutant mice. Furthermore, although all mice produced a significant rebound in REM sleep time, only PER1-deficient and *Per2* mutant mice had elevated NREM sleep time following sleep deprivation. In summary, the results from these two experiments indicate that *Per* genes exert individual effects on the timing of sleep and are also involved in modulating many aspects of the compensatory response to sleep loss.

Summary

The combination of results outlined in this review points to a critical role of circadian clock genes in sleep–wake regulation that extends beyond a strict time-keeping function. It is intriguing that each of these genes appears to influence sleep (baseline patterns and/or compensatory responses to sleep deprivation) in a unique way. To date, a simple model does not explain the different effects of individual clock genes on sleep. For example, the opposite phenotype in *Clock/Clock* (i.e., high-wake) and *Bmal1*$^{-/-}$ (i.e., high-sleep) mice is surprising because the protein products of these genes normally function together in a dimeric capacity. Sleep time is increased in both *Bmal1*$^{-/-}$ and *Cry1,2*$^{-/-}$ mice; however, *Cry* mRNA levels are elevated in *Bmal1*$^{-/-}$ mice, and *Bmal1* expression is high in *Cry1,2*$^{-/-}$ mice, seemingly incompatible patterns with respect to their sleep–wake phenotypes. As another example, *Per1* and

Per2 transcript levels are often elevated during spontaneous or experimentally induced wakefulness and *Per1*, *Per2,* and *Per1,2* mutants show a tendency for high wake time; however, high-sleep *Cry1,2*$^{-/-}$ mice also have elevated *Per1* and *Per2* transcript levels. These types of comparisons reveal that the role of circadian genes in sleep regulation is complex and probably extends to mechanisms beyond the core circadian gene transcriptional–translational feedback loop. One of the difficulties in directly comparing sleep among the different genetic models is the variety of experimental paradigms that have been used to study sleep: various sleep deprivation protocols, different criteria for scoring/analyzing sleep parameters, different genetic backgrounds of the models, mutants versus knockouts, and so forth. Nonetheless, enough evidence has mounted for us to conclude that circadian genes do more than simply control the timing of sleep–wake states. Indeed, individual clock genes may influence sleep, in part, through their activity in circadian regulation, as well as through circadian-independent effects. Because clock genes exist in a wide variety of central and peripheral tissues, there may be an additional component to their role in sleep regulation – anatomical specificity. The exciting finding that *Clock/Clock* mice become obese and demonstrate symptoms of the metabolic syndrome provides a framework to investigate the multiple and common molecular pathways in which the *Clock* gene is involved and that may link sleep with other physiological processes, such as energy metabolism. It has also been demonstrated that a large number of transcripts in central and peripheral tissue show rhythmic expression, presumably due to direct or indirect interactions downstream of the core circadian machinery. For example, the deletion of genes for albumin D-binding protein (DBP), prokineticin-2 (PK2), and transforming growth factor (TGF)-α, all clock-responsive genes, results in alterations in sleep regulation in mice. Another important advance in the field of sleep research has been the convincing demonstration that flies exhibit sleep–wake patterns and that their sleep is under homeostatic regulation. Significantly, mutations to *Drosophila* circadian clock genes (e.g., *Clock*, *Cycle*, *Period*, and *Timeless*) lead to changes in baseline sleep–wake regulation and to some dramatic alterations in the response to sleep deprivation. Undoubtedly, the use of novel model organisms and the development of new genetic techniques will provide many opportunities to continue investigating the apparently ubiquitous roles that circadian clock genes serve in sleep–wake regulation, as well as in the overall health and well-being of the organism.

See also: Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Organization; Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Rhythms in Sleepiness, Alertness, and Performance; Clock Gene Regulation of Endocrine Function; Clock Genes and Metabolic Regulation; Sleep and Circadian Rhythm Disorders in Human Aging and Dementia; Sleep: Development and Circadian Control; Transcription Control and the Circadian Clock.

Further Reading

Antle MC and Silver R (2005) Orchestrating time: Arrangements of the brain circadian clock. *Trends in Neuroscience* 28: 145–151.

Daan S, Beersma DG, and Borbely AA (1984) Timing of human sleep: Recovery process gated by a circadian pacemaker. *American Journal of Physiology* 246: R161–R183.

Dudley CA, Erbel-Sieler C, Estill SJ, et al. (2003) Altered patterns of sleep and behavioral adaptability in NPAS2-deficient mice. *Science* 301: 379–383.

Edgar DM, Dement WC, and Fuller CA (1993) Effect of SCN lesions on sleep in squirrel monkeys: Evidence for opponent processes in sleep wake regulation. *Journal of Neuroscience* 13: 1065–1079.

Franken P, Dudley CA, Estill SJ, et al. (2006) NPAS2 as a transcriptional regulator of non-rapid eye movement sleep: Genotype and sex interactions. *Proceedings of the National Academy of Sciences of the United States of America* 103: 7118–7123.

Hendricks JC, Lu S, Kume K, et al. (2003) Gender dimorphism in the role of cycle (BMAL1) in rest, rest regulation, and longevity in *Drosophila melanogaster. Journal of Biological Rhythms* 18: 12–25.

Ko CH and Takahashi JT (2006) Molecular components of the mammalian circadian clock. *Human Molecular Genetics* 15(2): R271–R277.

Kopp C, Albrecht U, Zheng B, and Tobler I (2002) Homeostatic sleep regulation is preserved in Per1 and Per2 mutant mice. *European Journal of Neuroscience* 16: 1099–1106.

Laposky A, Easton A, Dugovic C, et al. (2005) Deletion of the mammalian circadian clock gene BMAL1/Mop3 alters baseline sleep architecture and the response to sleep deprivation. *Sleep* 28(4): 395–409.

Mistlberger RE (2005) Circadian regulation of sleep in mammals: Role of the suprachiasmatic nucleus. *British Research Reviews* 49: 429–454.

Naylor E, Bergmann BM, Krauski K, et al. (2000) The circadian clock mutation alters sleep homeostasis in the mouse. *Journal of Neuroscience* 20: 8138–8143.

Shaw PJ, Tononi G, Greenspan RJ, and Robinson DF (2002) Stress response genes protect against lethal effects of sleep deprivation in *Drosophila*. *Nature* 417: 287–291.

Shiromani PJ, Xu M, Winston EM, et al. (2004) Sleep rhythmicity and homeostasis in mice with targeted disruption of mPeriod genes. *American Journal of Physiology* 287: R47–R57.

Turek FW, Joshu C, Kohsaka A, et al. (2005) Obesity and metabolic syndrome in circadian Clock mutant mice. *Science* 308: 1043–1045.

Wisor JP, O'Hara BF, Terao A, et al. (2002) A role for cryptochromes in sleep regulation. *BMC Neuroscience* 3: 20 [online].

Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus

R M Buijs, Universidad Nacional Autónoma de México, Mexico City, Mexico
C Escobar, Universidad Nacional Autónoma de México, Mexico City, Mexico
A Kalsbeek, Netherlands Institute for Neuroscience, Amsterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Organisms Exhibit Metabolic Rhythms: Daily and Annual Rhythms

Humans as well as animals exhibit daily rhythms in the sleep–wake cycle and in feeding and drinking behavior. In addition to these clearly visible behavioral rhythms, there are many physiological rhythms that serve to support behavioral rhythms; for example, several hormones and body temperature have a clear daily rhythm, as well as blood glucose and liver enzymes, which control glycogenesis and glycolysis. Diurnal metabolic rhythms are characterized by their dependence on the suprachiasmatic nucleus (SCN) and are independent of food intake; that is, some metabolic rhythms are expressed both during fasting and with a regular feeding schedule. Other metabolic parameters show fluctuations directly related to feeding events. For example, there is a rise in blood concentration of insulin after each meal. Some species exhibit annual fluctuations in metabolic functions, something that is especially observed in organisms that migrate, hibernate, or accumulate fat for winter. In these species, there are certain periods during the year in which they increase food storage, food ingestion, and fat storage.

The SCN Generates Metabolic Rhythms

The SCN is the master circadian clock, strongly entrained by the light–dark (LD) cycle. The daily changes in activity and sleep of the animal are mainly organized by the SCN and are synchronized to the daily change of the light to dark and vice versa via the light information that reaches the SCN through the optic nerve. There is experimental evidence that SCN lesions abolish the expression of daily rhythms involving glucose and corticosterone, body temperature, locomotor activity, and food intake. Because food intake is driven by the SCN, one might conclude that the other metabolic rhythms are derived from SCN-driven food intake rhythms. However, protocols that force animals into arrhythmic food intake and do not alter the SCN do not abolish glucose rhythmicity in the blood or the rhythmicity in liver enzymes. Such is the case with a protocol during which the animal eats every 4 h 6 times a day: under these conditions, the glucose rhythm in the blood is still maintained (Figure 1). In addition, when animals are fasting, the rise in plasma glucose still occurs in anticipation of the coming activity period. The mechanisms behind this daily peak in plasma glucose concentrations were examined by small bolus injections of glucose at different moments of the day. At the moment of the highest glucose concentrations in the blood, the lowest induced peak with injected glucose was also seen, indicating that at that moment not only the 'production' of glucose is the highest but also its 'uptake' (Figure 2). These observations indicate that the organization of the 'simple' glucose peak in the blood that precedes the activity is a complex event that requires an organization at different levels.

How Does the SCN Induce Rhythms in Metabolic Systems?

Hormonal and autonomic pathways are the two main pathways used by the SCN to transmit its rhythmicity to the body (Figure 3).

The Endocrine System

Since a high dose of corticosterone increases glucose output, it has been suggested that the daily corticosterone peak causes the peak in glucose concentrations at the end of the light period. However, in humans, the rise in glucose at the end of the light period is intact after cortisol action has been blocked. Also, no relationship is found between the blood levels of growth hormone (GH) and adrenalin over the 24 h of the day and the daily rhythm of plasma glucose. The 24 h rhythm in GH in rats, for example, is completely different from that in humans, that is, male rats do not display a clear circadian rhythm in GH concentrations, but a strong ultradian rhythm of 3 h. Although there is less and less evidence that links the circadian changes in corticosteroid levels in the blood to the production and release of glucose by the liver, corticosteroids are still assumed to play an important role in the circadian uptake of glucose by the cells of the body. There are several examples that show that glucocorticoids play an important role in preparing the animal for activity. For example, corticosteroids are increased when animals show food anticipatory activity. Moreover, when an animal shows two peaks

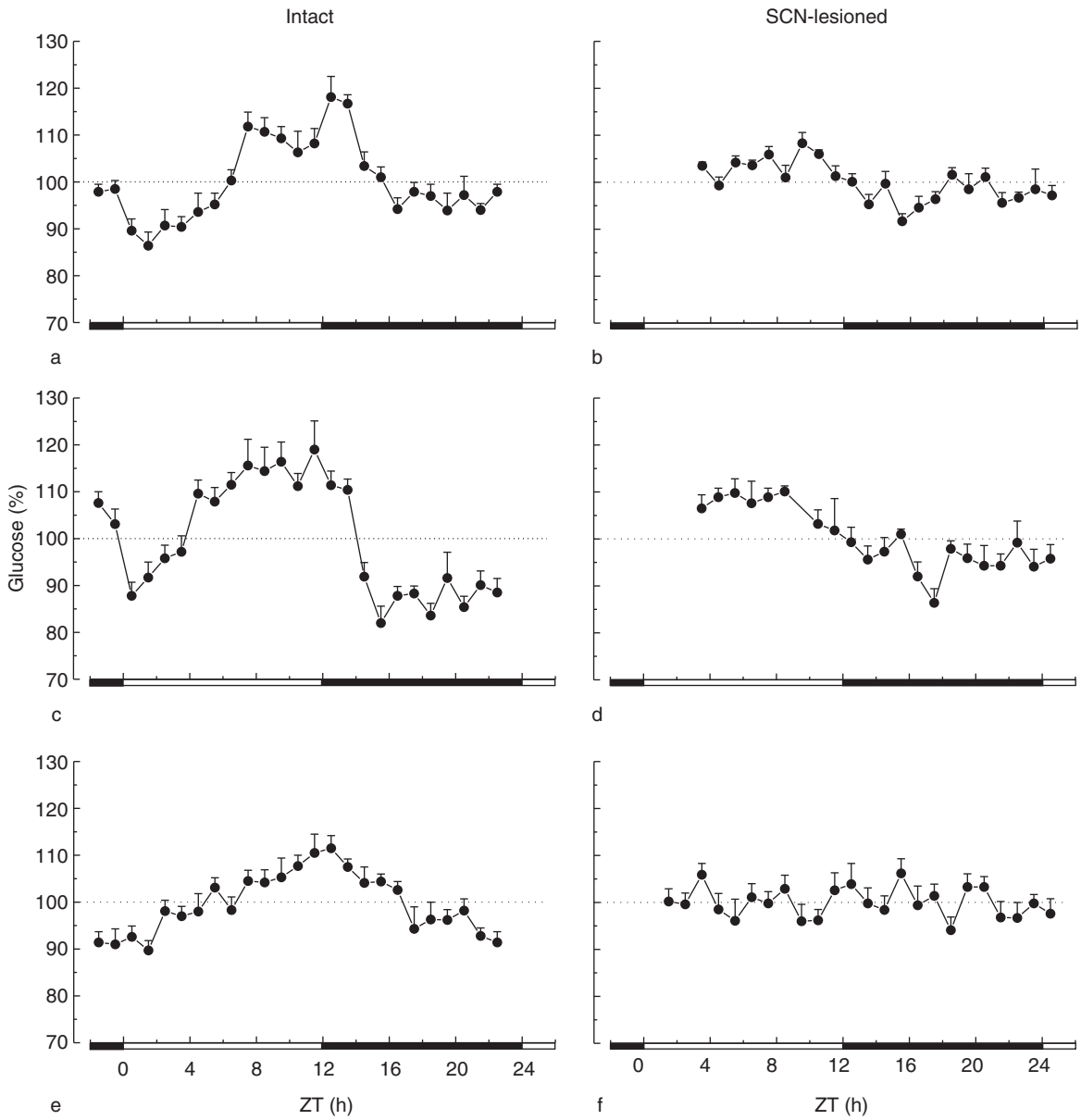

Figure 1 Basal glucose concentrations (values are percentages of the 24 h mean ± SEM) across the LD cycle, in intact rats (left panels; $n = 8$) and in SCN-lesioned rats (right panels; $n = 7$) under *ad libitum* feeding conditions (a, b), during fasting conditions (started at ZT = −12; c, d), and after having been subjected to a scheduled feeding regimen for at least 2 weeks (e, f). The scheduled feeding regimen consisted of a 10 to 12 min feeding opportunity provided every 4 h. Within 2 weeks, the animals learned to use each opportunity, that is, six every day, and consumed a meal of ±3.5 g at every occasion. The black areas indicate the dark period.

of activity in 24 h, two anticipatory peaks in corticosterone are also observed.

The exact mechanism by which the SCN controls plasma glucose concentrations is still being explored. The finding, however, of a simultaneous rise in blood glucose concentration and glucose uptake toward the activity period (Figure 2) suggests that glucose output to the general circulation must be so powerful that it compensates for increased glucose uptake. Of course, it makes perfect sense functionally to prepare an individual for upcoming activity: at the same time that glucose production is increased, the uptake of glucose increases as well. Moreover, it shows that the SCN is able to organize two completely different processes at the same time. The main source for glucose production is the liver. Hepatic glucose production can be stimulated by the sympathetic nervous system either indirectly, via adrenal hormones and glucagon, or directly, by sympathetic innervation of the liver. However, the evidence for the involvement of glucagon and adrenalin for the SCN-regulated anticipatory increase in plasma glucose concentrations toward the

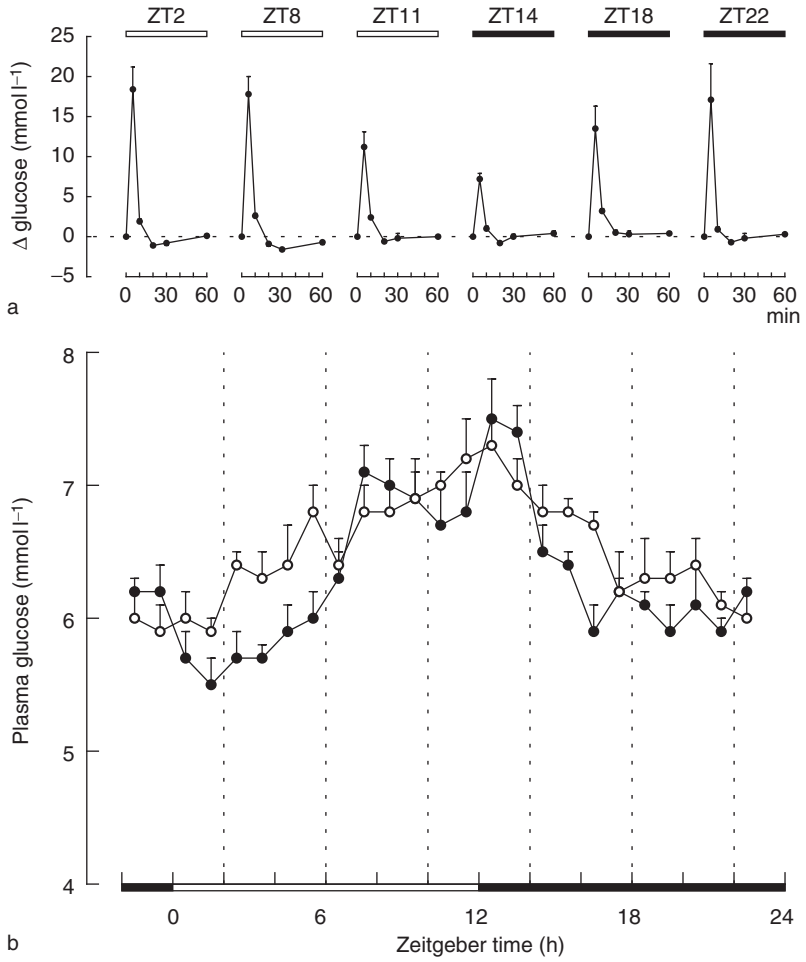

Figure 2 Plasma glucose concentrations after injection of a glucose bolus at different times of the light/dark-cycle (a) and 24 h plasma glucose concentrations during *ad libitum* feeding (closed circles) or during a scheduled-feeding regimen (open circles) (b). Responses in (a) are expressed as the difference with their respective $t = 0$ values. The smaller responses at ZT11 and ZT14 indicate an increased and more rapid uptake of glucose at these times of the LD cycle. The comparable plasma glucose curves of *ad libitum*-fed and scheduled-feeding animals in (b) indicate that the daily rise in plasma glucose concentrations before the dark period is not caused by a day/night rhythm in food intake. In addition, a comparison of the daily plasma glucose curves in (b) with the glucose uptake data in (a) evidences that the daily rise in plasma glucose concentrations is not the consequence of a decrease in peripheral glucose uptake. The black bars in (a) and (b) indicate the dark period.

beginning of the activity period is very weak. There is no daily variation in plasma insulin or glucagon levels during a six-meals-a-day schedule; whereas insulin declines after fasting, glucose, in spite of the fast, still increases in anticipation of the activity period. Moreover, the diurnal increase in adrenalin was also shown to be more related to activity than to the anticipatory peak in glucose.

Interestingly, in addition to the 'spontaneous' circadian changes, also in situations of an acute glucose challenge, the SCN sets the response of the endocrine system and autonomic nervous system (ANS). For example, when confronted with a hypoglycemic event in the early activity period, the low glucose level results in enormous changes in corticosterone secretion, while the same hypoglycemic event in the early sleep period results in hardly any changes in corticosterone.

Recently, the role of the arcuate nucleus has been investigated more closely, especially with respect to the control of metabolism. The arcuate nucleus, which lacks a blood–brain barrier and can be seen as a circumventricular organ (CVO), has intense reciprocal connections with the SCN. It is proposed that the SCN may control the sensitivity of the arcuate for circulating metabolites in a daily manner by means of this interaction. Moreover, several studies have also implicated the arcuate nucleus in the control of seasonal metabolic functions. For example, in hamsters, an increase in VGF protein/mRNA content

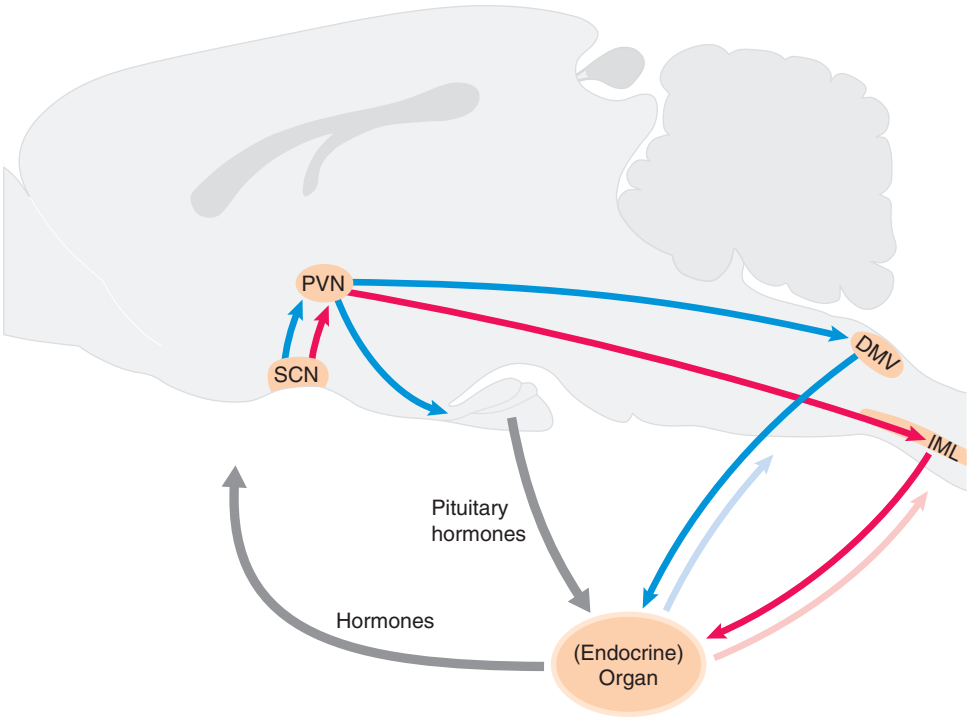

Figure 3 Sagittal scheme of the rat brain illustrates the main pathways by which the SCN communicates its circadian message to the body. Via its connections with the PVN of the hypothalamus (PVN), three different output systems are affected: (1) the neuroendocrine neurons of the PVN responsible for the control of pituitary hormones, (2) parasympathetically projecting neurons in the PVN that specifically target the dorsal motor nucleus of the vagus nerve (DMV), and (3) sympathetically projecting neurons in the PVN that specifically target the spinal cord preganglionic neurons located in the intermediolateral column (IML). The neuronal message prepares and sets the sensitivity of the (endocrine) organs for the arrival of the (pituitary) hormones. In return, the neuronal and hormonal message of the organs will feed back to the CNS (including the SCN), as indicated by the three arrows directed toward the brain.

in the arcuate nucleus precedes the photoperiod-induced increase in food intake and body fat. VGF is a protein involved in the control of food intake and has a widespread distribution in the central nervous system (CNS), including the SCN. Interestingly, in the hamster, the region of the arcuate is an important target for melatonin. This hormone, whose nightly secretion is organized by the SCN, has important seasonal functions. Since the arcuate nucleus is also an important target for SCN neurons, the organization of annual metabolic rhythms can be seen to be tightly linked to daily and seasonal metabolic functions.

Autonomic Output

The output of the SCN to the preautonomic neurons in the hypothalamus (mainly paraventricular nucleus (PVN)) is responsible for sending the circadian signal directly to our organs by the autonomic nerves (Figure 3). The PVN has extensive projections to sympathetic and parasympathetic motor neurons in the spinal cord and in the brain stem, respectively. Early studies using electric stimulation of the hypothalamus combined with blocking the output of the ANS showed that the effects of electrical stimulation on glucose metabolism were mainly reached by the ANS. Interestingly, at that time, it was also shown that exposure to bright light resulted in an increase in sympathetic nerve activity to the adrenal gland, liver, and pancreas, and in a decrease in vagal nerve activity to the pancreas and the liver. This effect disappeared after SCN lesioning, suggesting that light via the retinal terminals in the SCN was able to reach the target organs by the ANS. Recent studies have shown that the dorsomedial hypothalamus (DMH)/PVN area is essential for transmitting the circadian signal to the liver, pancreas, and fat tissue via the ANS. Later anatomical studies showed that after retrograde transneuronal tracer injection in the liver, pancreas, and fat tissue, separate sympathetic and parasympathetic pathways could be made visible that originate in different cells of the SCN. Consequently, the SCN contains neurons that are either preautonomic to the parasympathetic system or preautonomic to the sympathetic system, which projects to the pancreas, liver, or visceral fat. These SCN neurons transmit their message to the PVN, from which separate parasympathetic and sympathetic pathways target

the respective autonomic motor neurons. Recent experiments have shown that the sympathetic innervation to the liver plays a major role in liberating glucose from the liver. Removing GABAergic blockade (γ-aminobutylic acid (GABA) from SCN terminals?) within the PVN by means of Bicucullin (a GABA-A antagonist) infusion showed that the resulting glucose peak was mainly induced via the sympathetic innervation of the liver and not by increases in corticosterone and glucagon.

The control by the SCN of glucose tolerance is more likely to be at the level of the muscle, where the major (insulin-mediated) glucose uptake occurs. The mechanisms by which the SCN affects glucose uptake are still not known.

Integration of Glucose Rhythms and Energy Balance

Interaction of the SCN with Hypothalamic Nuclei Involved in Energy Balance

Sensing energy stores The brain depends entirely on glucose as fuel for its energy, while other organs, such as the heart, kidney, and skeletal muscle, may also rely on free fatty acids (FFAs) or other metabolites. In addition, the brain can actively take up glucose via glucose transporters, independent of insulin, which assures that, when insulin levels are low (fasting), the brain acquires sufficient glucose preferentially, because other tissues are dependent on insulin for glucose uptake. Thus sufficient glucose is always available to the brain.

Of course, in order to be able to balance and direct energy fluxes, the brain needs a precise and clear view of the metabolic state of the body. Circulating humoral factors provide an average signal of the whole body status, while sensory information from specific body regions provides detailed information. Sensory information from tissues important for energy storage, such as the liver (short-term storage) and fat tissue (long-term storage), must be available; recent studies have shown that both liver and fat tissue do provide such sensory information to the brain. In addition, tissues that use up a great deal of energy, such as muscles (skeletal and heart) and kidney, also need to transmit their energy level (and requirements!) to the brain.

Consequently, the brain plays a crucial role in sensing circulating metabolites and hormones; the CVOs, which lack the blood–brain barrier, such as the ventromedial arcuate nucleus (vmARC), area postrema (AP), subfornical organ (SFO), and organum vasculosum of the lamina terminalis (OVLT), likely play an important role in this regard. At present, it seems likely that especially the AP and vmARC play a major role in metabolic sensing and signaling, while the OVLT and SFO functions are more adapted for maintaining the osmolarity and salt levels of the blood. Recently, reciprocal connections between the SCN and vmARC were described, indicating that not only is the SCN kept informed about circulating metabolites, but that it might also be able to gate this information at the vmARC, depending on the time of the day (Figure 4). This hypothalamic integration of peripheral metabolic information results in an immediate modulation of metabolism. For instance, fatty acids and insulin are sensed in the vmARC, which leads to a change in autonomic output, affecting endogenous glucose production by the liver, glycogen synthesis in muscle, and food intake. Also, leptin and glucose are sensed in the vmARC, and their increase leads to inhibition of food intake and activation of anabolic functions, such as promotion of glycogen synthesis and lipid storage.

Interaction with Other Hypothalamic Nuclei

For a better understanding of how the SCN affects metabolism, it is also necessary to have a view on how the different hypothalamic structures influence metabolism. Each hypothalamic structure has its own role and interacts and influences the other structures within the hypothalamus in a specific manner. Clearly, there are main food intake stimulatory centers such as the neuropeptide Y (NPY) and agouti-related peptide (AGRP)-containing neurons of the medial arcuate nucleus, while the food inhibitory centers (the pro-opiomelanocortin (POMC)-containing neurons) are in the lateral arcuate nucleus. In addition to these (according to the current views) main structures for controlling food intake, the ventromedial hypothalamus (VMH), the lateral hypothalamus, and the DMH are essential for integrating and transmitting visceral and hypothalamic information; the PVN seems to play a role mainly as integration and output center (hormonal and autonomic) of the hypothalamus. In addition to these structures that are directly involved in the control of food intake, there are also structures that are involved in related functions, which are also very important for metabolism, such as the medial preoptic area (MPO), which is essential for the control of thermoregulation. The importance of this interaction becomes clear when one evaluates the consequences of fasting for the physiological organization of the body. Fasting results in a diminishment in temperature, but only during the sleep period; it lowers the heart rate, during the day as well as at night; and it increases running wheel activity in rats in spite of the diminishing energy stores. At the same level, induced obesity in rats also affects the

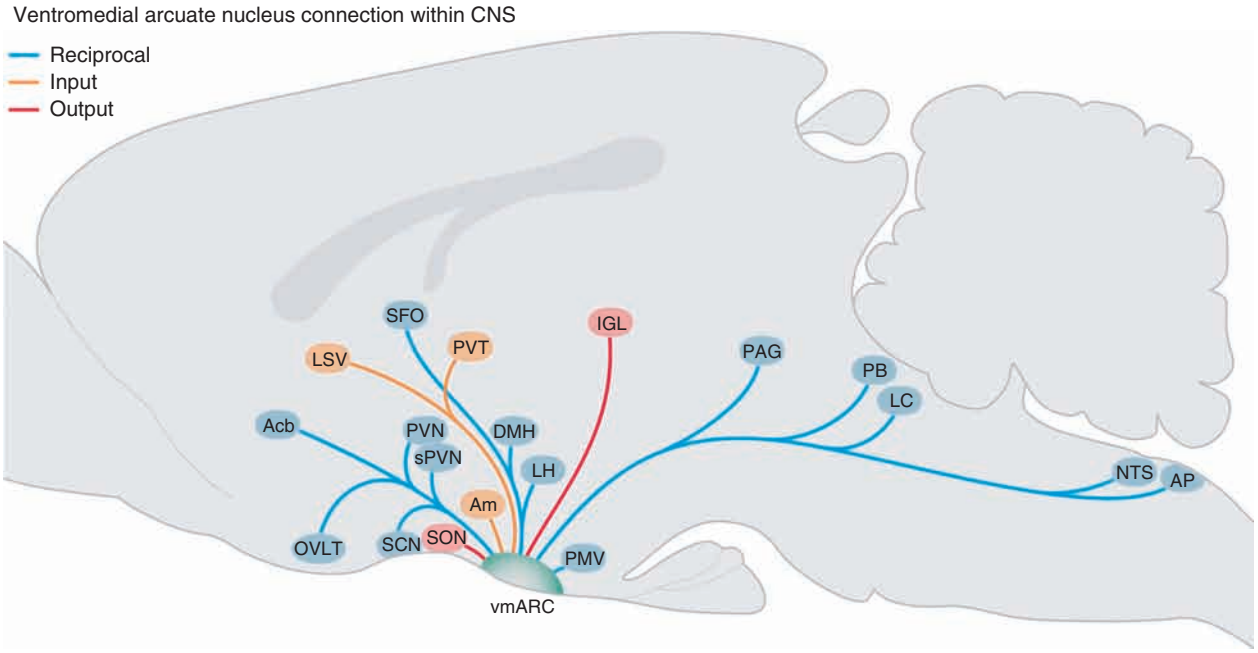

Figure 4 Sagittal scheme of the rat brain to illustrate the main connections of the ventromedial arcuate nucleus (vmARC) within the CNS. Brain areas that have reciprocal connections with the vmARC are indicated in blue. Brain areas with a significant input to the vmARC (i.e., PVT, LSV, and Am) are indicated in orange. Red areas (i.e., IGL and SON) represent those brain areas that receive a significant output from the vmARC. Acb, nucleus accumbens; LSV, lateral septal nucleus ventral part; OVLT, organum vasculosum lamina terminalis; SFO, subfornical organ; LH, lateral hypothalamus; SCN, suprachiasmatic nucleus; SON, supraoptic nucleus; PVN, paraventricular nucleus; sPVN, sub paraventricular nucleus; Am, amygdaloid complex; DMH, dorsomedial hypothalamic nucleus; PVT, paraventricular nucleus of the thalamus; PMV, premammilary nuceus ventral part; IGL, intergeniculate leaflet; PAG, periaquaductal grey; NTS, nucleus tract solitarius, LC, locus coeruleus; AP, area postrema; PB, parabrachial nucleus.

organization of hypothalamic functions; it diminishes the amplitude of the circadian activity and decreases nighttime body temperature. On the other hand, infection also has a large impact on hypothalamic function; it increases corticosterone secretion, diminishes activity, increases temperature, and induces sleep. These observations emphasize the interactions between the different hypothalamic structures and how they affect the functioning of the biological clock. In order to understand these mechanisms, it is important to consider that, in addition to light information that reaches the SCN, mainly in the ventral part, also (metabolic) information from the circulation may reach the SCN via CVOs such as the arcuate nucleus. This site of integration in the SCN (i.e., light and metabolic data target the same area) may allow metabolic information to modify the activity of the SCN similar or contrary to the way light affects the SCN.

What is the Relevance of Metabolic Rhythms?

Metabolic rhythms allow adaptation to daily activities and to annual geographical fluctuations in temperature and food availability, thus allowing animals to adjust their energy balance according to the environment. And even more importantly, because of the regularity of these rhythms, they allow animals to anticipate and prepare for those environmental changes. Consequently, the coherence of endogenous metabolic oscillations with the external cycles of the environment is very important for health and homeostasis. When endogenous rhythms uncouple from the external cycles, individuals cannot exhibit the adequate response at the moment when it is needed and they develop health problems and disease. Such is the case in shift workers who work regularly during the normal sleeping period.

Anticipation of Changes in Energy Requirement

In order to maintain homeostasis, the body needs to anticipate changes in the environment. One of the clearest examples is the anticipation of seasonal changes. Here the SCN–pineal axis is essential for preparing the animal for the upcoming change in food supply, which forces profound changes in reproduction and metabolism. Day length influences the duration of melatonin secretion, and when the period of melatonin secretion is increased or decreased for some time, the reproductive axis, food intake, and fat

storage are simultaneously affected. A number of species (especially humans) are not (notably) influenced by the change in seasons, but in all mammals the biological clock, the SCN, has a special role, in that it takes care of anticipating change in metabolism prior to a period of activity. This anticipatory role of the SCN is characteristic of its involvement in the organization of the physiology in animals, and its influence on metabolism can be seen, for example, in the rise in glucose, corticosterone, and body temperature just before awaking (Figure 1). The appropriate organization of the metabolism is essential for the life of an organism: this may be the reason why, in addition to the circadian system, a number of other physiological systems and/or functions also influence metabolic functions, such as body temperature, fluid balance, the immune system, and reproduction. These functions are all organized within the hypothalamus, where the main tasks of the body are controlled and transmitted to the body by hormones and the autonomic nervous system. Clearly, the organization of metabolic functions takes place within this circuit of brain–body interaction, and a change anywhere in this circuit has repercussions for both brain and body. For example, the metabolic syndrome is a metabolic disease which is characterized by hyperlipidemia, visceral obesity, hypertension, or diabetes. Disturbed circadian rhythms are a prominent feature of the metabolic syndrome.

Metabolic Rhythms without the SCN; Food Entrainment

Under special conditions, when food is restricted to a short daily event, animals show metabolic adjustments that enable them to deal with such a restricted feeding schedule. Also, animals modify their behavioral patterns in order to be active and anticipate meal time. Animals show increased arousal, foraging, and locomotion 2–3 h before expected food access, a preparatory behavior known as 'food anticipatory activity.' This anticipatory behavior appears independent of the SCN-driven nocturnal activity characteristic of most rodents, which remains, although with decreased amplitude. In order to adjust to the feeding schedule, metabolic rhythms show phase shifts and exhibit peaks related to meal time; some of them, like the corticosterone rhythm, exhibit a peak during meal time anticipation, while the largest part of the SCN cellular activity and metabolism remains adjusted to the LD cycle. Metabolic rhythms persist independent of the SCN and continue to oscillate with a 24 h period, probably driven by food-entrained oscillators. This conclusion is based on the observation that food entrainment is also visible in SCN-lesioned animals, indicating that the SCN is not essential for food entrainment. Due to this latter fact, for approximately 30 years, there has been a search for a different biological clock, the food-entrained oscillator, which is thought to drive food-entrained rhythms when food access is uncoupled from the LD cycle. The functional and anatomical identity of this clock remains unknown to date. Recent studies have evidenced that under restricted feeding schedules, food becomes a powerful entraining signal for most of the peripheral organs in which diverse cellular functions exhibit food-entrained rhythms. In both peripheral organs and in brain structures, feeding schedules entrain daily oscillations of clock genes. Thus it has become clear that food, when restricted to a few hours a day, can override the signals of the LD cycle-driven clock and can uncouple the metabolic oscillations from the SCN. However, this does not imply that, under normal circumstances, the SCN is not involved in food entrainment, because in such conditions animals organize their meals according to their activity phase.

In order to approach phenomena of internal desynchronization resulting from jet lag, shift work, or nocturnal activities, it is important to keep in mind that food can exert a powerful influence as an entraining signal to the peripheral organs and metabolic processes. Such circadian disorders can be approached by tightly controlling the feeding schedules in order to shift metabolic rhythms to the expected phases. Also, it remains to be established whether disorganized feeding schedules can be linked to metabolic problems due to uncoupling of peripheral oscillators from the SCN.

The Metabolic Syndrome: A Brain Disease?

Ignoring the Signal of the Clock

The background of the mechanisms causing the metabolic syndrome is incompletely understood. In general, there is agreement that the cause for this outbreak of obesity must lie in the abundance of food in modern societies and the fact that evolution has shaped us for periods of scarcity rather than surplus. In addition, our evolutionary-shaped homeostatic systems have 'learned' to adapt to the everchanging LD cycle, in such a way that the body anticipates the upcoming sleep and activity periods. During the past century, life has changed dramatically in industrialized countries. Food has become abundant and snacking frequency has increased and shifted toward the end of the day, while at the same time the necessity for physical effort has become considerably reduced. Also, physical activity no longer necessarily coincides

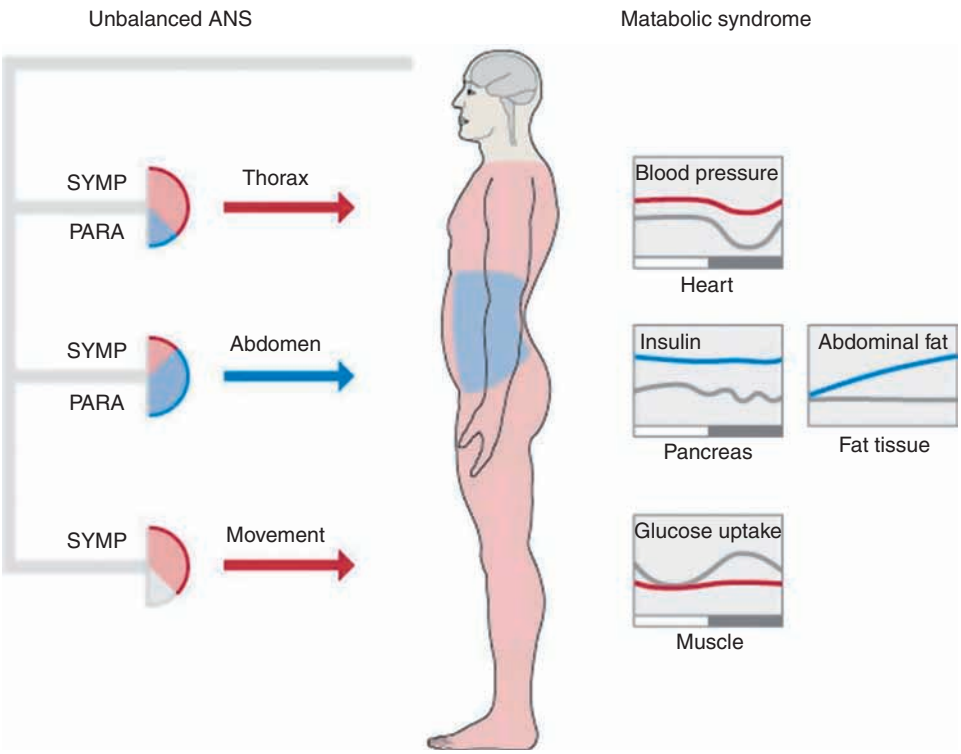

Figure 5 Model of the metabolic syndrome caused by a central nervous deregulation. The disturbed output of the biological clock affects the selective balance of the autonomic nervous system (ANS) in different parts of the body. In the intra-abdominal compartment, the ANS is shifted in favor of the parasympathetic branch, resulting in high insulin secretion, growth of intra-abdominal fat tissue, and fatty liver. Contrarily, in the thorax and movement compartment, the sympathetic branch prevails, leading to high blood pressure and impaired glucose uptake by the muscle. In this model, the symptoms of the metabolic syndrome are result and not cause of the disease.

with the light period. We hypothesize that in such conditions the susceptible brain loses its feeling for internal and external rhythm. As a result, the environment sensed by the brain has become metabolically flattened and arrhythmic. Since the brain uses the ANS to implement the internal rhythmicity, we propose an unbalanced and arrhythmic ANS as a major cause of the metabolic syndrome.

The organization of the output of the SCN via the ANS forms the basis for this proposal. For a long time, fat tissue has been assumed to be targeted by hormones or to be innervated by the sympathetic nervous system only. Recently, it was demonstrated that parasympathetic input from the vagal motor nucleus to fat tissue clearly modulates insulin sensitivity and glucose and fat metabolism in fat tissue in an anabolic way. This allows the brain to stimulate either the burning of fat (sympathetic) or the accumulation of fat (parasympathetic). Moreover, we demonstrated that the brain has an astonishing capacity with respect to the specialization of neurons in the hypothalamus and SCN. For example, preautonomic neurons are present in the SCN and other hypothalamic areas that project either to the intra-abdominal or the subcutaneous body compartment. The significance of this is that the intra-abdominal organs, for example, visceral fat, liver, and pancreas, receive input from the same neurons (Figure 5). Since consuming a meal stimulates the parasympathetic output to the pancreas and results in insulin secretion, the consequence of this hard ANS wiring is that the same stimulus will also increase the insulin sensitivity of the liver and of the visceral fat. Since meals consumed in the evening are known to lead to enhanced insulin secretion, a more efficient uptake of glucose in visceral fat will accordingly result in easier accumulation of intra-abdominal fat. These facts, taken together, indicate that as a consequence of the changed lifestyle in industrialized countries, the metabolic syndrome is on the rise as well, not necessarily because too much food is consumed, but because food is eaten at the wrong time!

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Clock Genes and Metabolic Regulation; Neuropeptides: Food Intake.

Further Reading

Buijs RM, Chun SJ, Niijima A, et al. (2001) Parasympathetic and sympathetic control of the pancreas: A role for the suprachiasmatic nucleus and other hypothalamic centers that are involved in the regulation of food intake. *Journal of Comparative Neurology* 431: 405–423.

Buijs RM and Kalsbeek A (2001) The day–night balance of everyday life: Central and peripheral clocks interacting in the hypothalamus. *Nature Reviews Neuroscience* 2: 521–526.

Buijs RM, La Fleur SE, Van Heijningen C, et al. (2003) The suprachiasmatic nucleus balances sympathetic and parasympathetic output to peripheral organs through separate preautonomic neurons. *Journal of Comparative Neurology* 464: 36–48.

Buijs RM, Wortel J, Van Heerikhuize JJ, et al. (1999) Anatomical and functional demonstration of a multisynaptic suprachiasmatic nucleus adrenal (cortex) pathway. *European Journal of Neuroscience* 11: 1535–1544.

Cailotto C, La Fleur SE, Van Heijningen C, et al. (2005) The suprachiasmatic nucleus controls the daily variation of plasma glucose via its autonomic output to the liver: Are the clock genes involved? *European Journal of Neuroscience* 22: 2531–2540.

Damiola F, Le Minh N, Preitner N, et al. (2000) Restricted feeding uncouples circadian oscillators in peripheral tissues from the central pacemaker in the suprachiasmatic nucleus. *Genes and Development* 14: 2950–2961.

Fleur la SE, Kalsbeek A, Wortel J, et al. (1999) A suprachiasmatic nucleus generated rhythm in basal glucose concentrations. *Journal of Neuroendocrinology* 11: 643–652.

Fleur la SE, Kalsbeek A, Wortel J, et al. (2000) Polysynaptic neural pathways between the hypothalamus, including the suprachiasmatic nucleus, and the liver. *Brain Research* 871: 50–56.

Kalsbeek A and Buijs RM (2002) Output pathways of the mammalian suprachiasmatic nucleus: Coding circadian time by transmitter selection and specific targeting. *Cell Tissue Research* 309: 109–118.

Kalsbeek A, Fliers E, Romijn JA, et al. (2001) The suprachiasmatic nucleus directly generates the diurnal changes in plasma leptin levels. *Endocrinology* 142: 2677–2685.

Kalsbeek A, La Fleur SE, Van Heijningen C, et al. (2004) Suprachiasmatic GABAergic inputs to the PVN control plasma glucose concentrations in the rat via sympathetic innervation of the liver. *Journal of Neuroscience* 24: 7604–7613.

Kalsbeek A, Ruiter M, La Fleur SE, et al. (2003) The diurnal modulation of hormonal responses in the rat varies with stimuli. *Journal of Neuroendocrinology* 15: 1144–1155.

Kalsbeek A and Strubbe JH (1998) Circadian control of insulin secretion and glucose homeostasis is independent of the temporal distribution of feeding. *Physiology and Behavior* 63: 553–560.

Kreier F, Fliers E, Voshol PJ, et al. (2002) Selective parasympathetic innervation of subcutaneous and intra-abdominal fat – functional implications. *Journal of Clinical Investigation* 110: 1243–1250.

Kreier F, Kalsbeek A, Ruiter M, et al. (2003) Central nervous determination of food storage – A daily switch from conservation to expenditure. Implications for the metabolic syndrome. *European Journal of Pharmacology* 480: 51–65.

Kreier F, Kap YS, Mettenleiter TC, et al. (2006) Tracing from fat tissue, liver and pancreas: A neuroanatomical framework for the role of the brain in type 2 diabetes. *Endocrinology* 147: 1140–1147.

Kreier F, Yilmaz A, Kalsbeek A, et al. (2003) Hypothesis: Shifting the equilibrium from activity to food leads to autonomic unbalance and the metabolic syndrome. *Diabetes* 52: 2652–2656.

La Fleur SE, Kalsbeek A, Wortel J, et al. (2001) A daily rhythm in glucose tolerance. A role for the suprachiasmatic nucleus. *Diabetes* 50: 1237–1243.

Niijima A, Nagai K, Nagai N, et al. (1999) Light enhances sympathetic and suppresses vagal outflows and lesions including the suprachiasmatic nucleus eliminate these changes in rats. *Journal of the Autonomic Nervous System* 40: 155–160.

Ruiter M, La Fleur SE, Van Heijningen C, et al. (2003) The daily rhythm in plasma glucagon concentrations in the rat is modulated by the biological clock and by feeding behaviour. *Diabetes* 52: 1709–1715.

Yi C, Van Der Vliet J, Dai J, et al. (2006) Ventromedial arcuate nucleus communicates peripheral metabolic information to the suprachiasmatic nucleus. *Endocrinology* 147: 283–294.

Circadian Organization

F W Turek, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A fundamental characteristic of life on the planet Earth is the 24 h cycle of changes that occur at the cellular, tissue, organ, and whole organism level, to match the period of the rotation of Earth on its axis. Throughout much of the second half of the twentieth century many investigators working on organisms as diverse as the fungus *Neurospora*, the fruit fly *Drosophila*, the hamster *Mesocricetus,* and humans (*Homo sapiens*) conclusively demonstrated that diurnal rhythms were not simply driven by the 24 h changes in the external physical environment, but instead arose from internal biological clocks. Because the endogenous period of these living daily clocks is rarely exactly 24 h, but nevertheless is always close to 24 h, the clocks are called 'circadian' (from the Latin, 'about a day'), and the rhythms that are measured (e.g., sleep–wake cycle, body temperature, feeding, pituitary hormone release) are referred to as circadian rhythms. The nonexact 24 h nature of the internal circadian clock necessitates that it somehow be synchronized to the precise 24 h solar day so as not to become desynchronized from the environmental day–night cycle, and for most organisms the daily light–dark (LD) cycle serves as the major entraining agent of the endogenous circadian clock. (Further details about how circadian clock systems are organized throughout the animal kingdom, and the remarkable progress that has been made in the past decade in unraveling the molecular machinery that underlies the generation of circadian rhythms, can be found elsewhere in this encyclopedia.)

The master circadian clock, located in the bilaterally paired hypothalamic suprachiasmatic nucleus (SCN) in mammals, regulates the timing of literally hundreds of physiological, metabolic, and behavioral parameters. Such timing is thought to be important for the health and well-being of the organism, both in terms of maintaining synchronization of the organism with the external environment (e.g., to search for food or sleep at appropriate time of day or night) and in terms of maintaining synchronization between the myriad of internal circadian rhythms. The SCN has often been referred to as being analogous to the conductor of a symphony. In the presence of the conductor, the individual components of the orchestra play in synchrony with one another. Without the conductor, the individual units will become out of synchrony, with the resulting cacophony of sound. Aging is known to lead to a disruption of the circadian organization of the organism at many levels; the following discussions focus on the mechanisms underlying age-related changes in the circadian clock system, and the possible consequences for the organism as it transitions from a highly synchronized to a less synchronized diurnal state with advancing age.

With advancing age there are changes in several components of the circadian timing system, and each may play a role in the age-related decline in efficacy of the system as a whole. While a decreased strength of input signals to the circadian system may explain some of the age-related decline in sensitivity to both photic and pharmacological stimuli, concurrent changes in the function of the central pacemaking mechanism may account for some changes in sensitivity. Changes in the central pacemaker may occur at the level of the intracellular loops of transcription and translation that have recently been elucidated. In addition, some of the decline in overt circadian rhythmicity may be due to changes in output signals from the circadian clock, or to deteriorating responsiveness of the body's aging effector systems to circadian signals. Thus, aging of the circadian clock system can occur at many levels: from the input to the clock, the clock itself, and its output signals, as well as the end processes affected by circadian signals (Figure 1).

Age-Related Changes in Inputs to the Circadian Clock

A remarkable feature of the mammalian circadian photoreceptive system is that the photoreceptors necessary for vision, the rods and the cones, are not necessary for light to entrain the circadian clock. It has been clearly established in rodents and primates that melanopsin-containing cells exist in the ganglion layer of the retina, and these ganglion cells appear to be the primary photoreceptors relaying light information from the retina to the SCN. While the anatomical and functional importance of these melanopsin-containing ganglion cells has only become apparent in the past few years, two features of the circadian organization of vertebrates/mammals that have been known for decades suggested that light input to the circadian clock was separate from the visual system. First, among the nonmammalian vertebrates, the circadian photoreceptors are not even in the eye, but instead are 'extraretinal' and lie deep in the brain. Second, in mammals, there is a special tract of nerves that projects directly

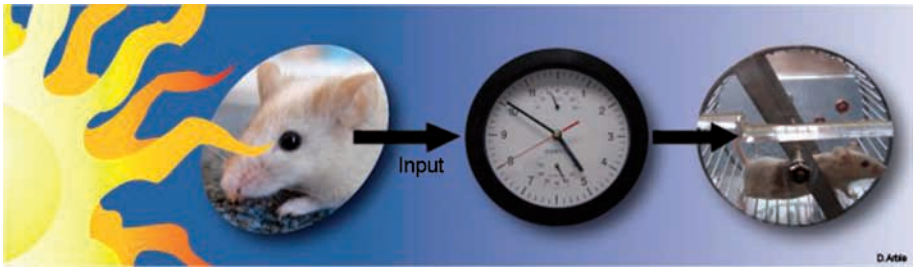

Figure 1 Schematic representation of the circadian clock system in its simplest form. An internal circadian clock must receive input from the external environment, which in the case of most mammals involves photic information from the light–dark environment that is detected by the eye. The internal clock in turn must then communicate temporal information to effector systems regulating such measurable output rhythms as the daily activity/rest cycle of rodents. Age-related changes in the circadian clock system could occur at any level along this pathway, both at the level of the central circadian clock in the suprachiasmatic nucleus and at the level of a multitude of molecular circadian clocks regulating expression of hundreds of clock-controlled genes in cells/tissues outside of the suprachiasmatic nucleus, through the central nervous system and periphery. Courtesy of Deanna Arble.

from the retina into the hypothalamic SCN, referred to as the retinohypothalamic tract (RHT). A second retinal input to the SCN involves an indirect route and goes through the thalamic intergeniculate leaflet. Both traits are separate from the inputs to the visual centers of the brain.

Studies in old rodents indicate that in advanced age (usually >18 months in rats, hamsters, and mice) the circadian photoreceptive system becomes less sensitive to light, based on biochemical, as well as behavioral, measurements. For example, in old rats, mice, and hamsters free-running in constant darkness (DD), the phase shifts which occur in the free-running circadian rhythm of activity in response to a single pulse of light (e.g., 1 h in duration) are less than in a young animal. Light induction of immediate-early genes in the SCN is also reduced in old mice and hamsters. Taken together, the studies of the light sensitivity of the circadian clock system of old rodents has led to the conclusion that in advanced age (1) the sensitivity of the photoreceptive system is decreased, (2) the transition of light information to the SCN is reduced, and/or (3) there is a change in the response at the cellular level of SCN neurons to photic inputs. The use of light therapy to aid in the entrainment of the sleep–wake cycle of elderly humans who are often exposed to very little natural light is partially based on the hypothesis that the photic sensitivity of the circadian system declines with advanced age.

While the light–dark cycle is the major entraining agent of the circadian clock in mammals, other stimuli, including social stimuli, food availability, and the timing of sleep and wake, can entrain and/or influence the temporal profile of rhythms regulated by the circadian clock. As with the response to light, the effects of these nonphotic stimuli on the circadian clock are attenuated in advanced age in rodents. Of particular interest is information reaching the SCN via serotonergic input from the midbrain raphe nuclei, since it appears that this input has a diminished effect on clock function with advanced age; an interesting age-related change in circadian input organization given the association between abnormal circadian rhythms and depression, and the prominent role played by 5-hydroxytryptamine (5-HT; serotonin)-related drugs in the treatment of depression.

Thus, the signaling strengths of both photic and nonphotic inputs into the SCN are diminished in advanced age. Since the proper functioning of input pathways to the central clock is necessary to synchronize the internal timekeeping mechanism with the external environment, age-related changes in the inputs could lead to an age-related pathophysiology of the circadian organization. In addition, as discussed in the next section, age-related changes in the central timing mechanisms can also occur with advanced age.

Age-Related Changes in the Master Central Circadian Clock

A common feature of age-related changes in different cellular, physiological, and behavioral circadian rhythms is a change (usually a decrease) in the amplitude of the rhythms. While a decrease in rhythm amplitude could be due to effects of age on downstream events, there is considerable evidence demonstrating that a decrease in the amplitude of the central core circadian clock is an important component of aging in the organization of the circadian system. This is directly observed in old rats and hamsters, in which the amplitude of the circadian rhythm of the firing rate of SCN neurons (the firing rate is higher during the day than in the night in both nocturnal and diurnal animals) is decreased with old age. This decrease does not appear to be due to an age-related

decrease in the number of SCN neurons, but instead has been attributed to a change in the ability of individual SCN cells to oscillate and/or to a decrease in the coupling strength between individual cells. The diurnal rhythm in glucose uptake within the SCN (also increased during the light period) is also attenuated in advanced age. *In vitro* studies have also found that while individual SCN neurons from young adult mice (2–4 months old) show robust circadian rhythms in spontaneous activity, with periods close to the period of behavioral rhythms in the intact animal, SCN neurons from 'middle-aged' mice (9–11 months old) show a decrease in the amplitude of impulse activity and more variability in the circadian waveform of activity.

A few laboratories have demonstrated that in rodents the expression of some of the core molecular circadian clock genes (e.g., *Clock*, *Bmal 1*, *Per*) is reduced in advanced age. In addition, a number of clock-controlled genes, as well as genes for which expression is induced by light (including core circadian clock genes), show reduced rhythms or induction by light in advanced age. An age-related decrease in gene expression in response to light is consistent with a decrease in the phase-shifting effects of light on measurable output rhythms, such as locomotor activity, with advanced age. Taken together, these data indicate that the age-related changes that are observed in multiple diverse rhythms in mammals may be due to changes within the SCN cells/tissues – changes that alter the ability of the SCN to generate rhythms and to respond appropriately to external entraining stimuli.

Age-Related Changes in Function of the Downstream Effector System

Of course, changes in the amplitude of many circadian rhythms could be due, in part, to age-related changes in the downstream (from SCN) elements for which diurnal timing is only one component of the overall function of the tissue/organ or behavior being measured. For example, a common characteristic observed almost universally in old rodents is a gradual decrease in the amount of locomotor activity (e.g., fewer wheel revolutions, or beam crossings in a cage), which could likely be due to a weakening of muscle strength. Similarly, aging reduces the circulating levels of many hormones, which could be due to diminished activity of the endocrine glands. Thus, when one observes a breakdown or changes in the circadian organization with advanced age, the underlying cause could be at the input, pacemaker, output signaling, and/or one of the many downstream components involved in the actual rhythm being measured.

A recent and far-reaching discovery about the circadian organization of higher organisms has important, but largely unaddressed, implications about the effects of aging on this organization. Just a few years ago it was generally believed that few cells/tissues outside of the SCN contained the machinery capable of generating a self-sustained molecular circadian rhythm in gene and protein synthesis. However, it is now recognized that most, if not all, of the tissues and cells of the body contain not only the core molecular circadian transcriptional and translational elements, but that many of these non-SCN central and peripheral tissues can sustain oscillations *in vitro* for a number of cycles. Furthermore, these cellular clocks, while under the upstream entrainment of the SCN molecular central clock, are regulating at the 'local' level the diurnal rhythms in expression of hundreds of clock-controlled genes. Since we know that aging has an impact on the input pathways to the SCN, as well as the central pacemaker cells in the SCN, it is almost certain that aging will have consequences for the inputs from the SCN to the downstream molecular circadian clocks, as well as on those downstream clocks. The functional consequences for age-related changes at multiple levels in the overall circadian organization await future research. However, the recent finding that aging may have differential effects on central and peripheral oscillators, as well as the resetting properties of these oscillators, raises the intriguing possibility that such age-related changes in timing at the tissue/organ level could contribute to diseases of the brain and periphery.

Attenuating or Reversing Age-Related Changes in Circadian Organization

As with other aging physiological systems, attempts have been made to find the 'fountain of youth' for aging circadian structures. Since the circadian clock becomes less responsive to the entraining effects of photic (as well as nonphotic) stimuli, early attempts to improve circadian rhythmicity involved increasing the strength of the entraining light–dark cycle by increasing light intensity. Modest success in increasing the amplitude of the sleep–wake or activity–rest cycle in rats and hamsters, particularly middle-aged animals (Figure 2), has lent support to the use of bright light therapy in the elderly at the beginning and/or end of the day in an attempt to enhance and increase the robustness of circadian rhythmicity.

It has been suggested that melatonin may be useful in treating a number of human circadian disorders, including age-related insomnia. The melatonin agonist S20242 significantly increases a measure of rhythm stability, the χ^2 value of the periodogram,

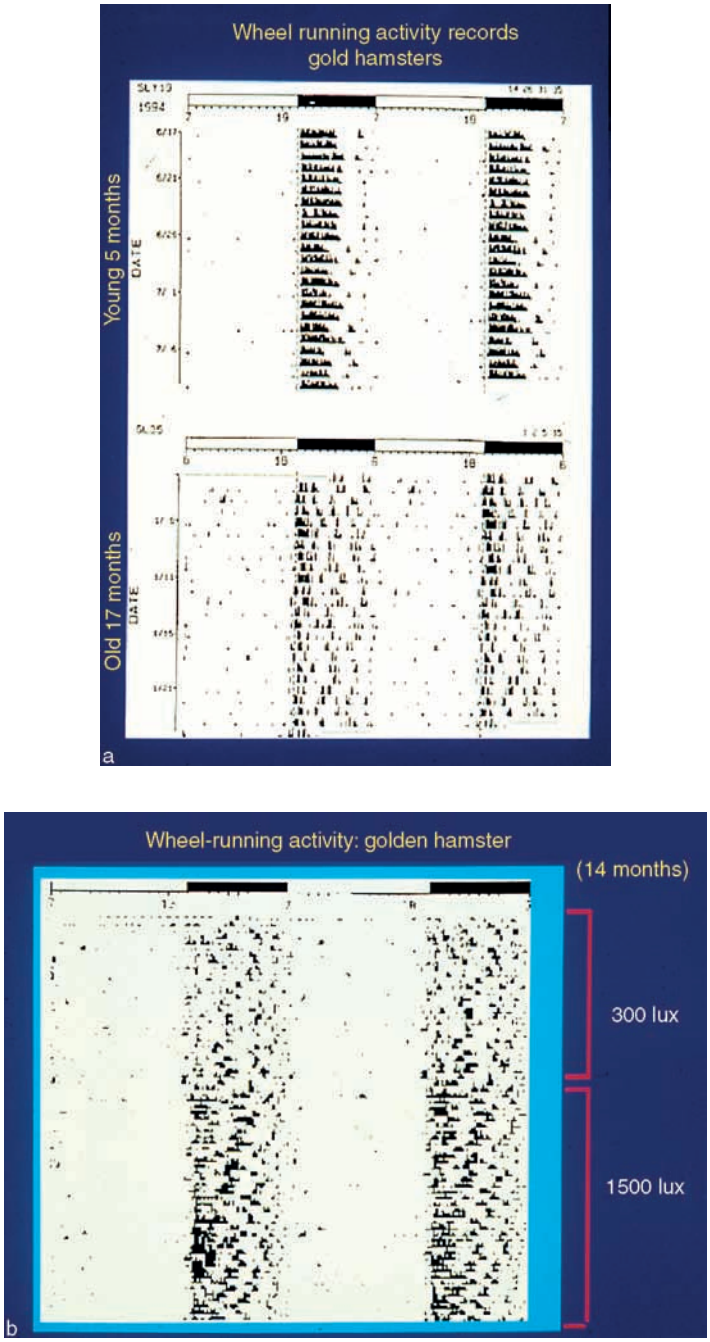

Figure 2 (a) Rhythm of wheel-running activity of male golden hamsters, one young (5 months old) and one old (17 months), showing how with advanced age the activity rhythm becomes less consolidated and more fragmented, with a decrease in total activity. (b) Activity rhythm of a 'middle-aged' male hamster (14 months) exposed to a light–dark cycle in which the light intensity was increased, at the indicated time, from 300 to 1500 lux. The increase in light intensity resulted in a more consolidated activity rhythm. Activity data for all three animals (ages 5, 17, and 14 months) are plotted over 48 h time intervals, and the animals were exposed to a 14:10 light–dark cycle, as indicated by the light and dark bars at the top of each record.

when administered to middle-aged and old rats. This treatment also increases the amplitude of the circadian rhythm of body temperature in middle-aged, but not old, rats. Daily feeding of a different melatonin agonist (S20098) to old hamsters restores their sensitivity to a dark pulse; when the treated food is removed, dark pulses once again fail to phase-shift the old hamsters. Together, these two findings suggest

that exogenous melatonin may be able to restore at least some of the 'youthful' properties to the circadian systems of old animals.

Can Circadian Dysfunction Affect Aging?

A final consideration reverses the question ('what are the effects of aging on the circadian organization?') that has been addressed throughout most of this article. Instead, now a different question is addressed: What are the effects of disrupted circadian rhythms on aging?

A salient characteristic of modern industrial societies is the fact that many jobs require individuals to work at times they would normally be asleep, and to sleep at times they would normally be awake. Such 'shift workers' are constantly changing the times when they are awake or asleep relative to the 24 h day. Thus there is repeated desynchronization or resynchronization between their internal 24 h clock with the external time of day. There is a good deal of epidemiological data to support the hypothesis that there are health consequences of circadian maladaptation and disrupted sleep–wake cycles in shift workers. Shift work has been associated with increased cardiovascular morbidity and mortality, as well as gastrointestinal and metabolic disorders. Surprisingly, there is little experimental evidence from animal models to support the hypothesis that the continuous disruption of normal temporal organization is in fact detrimental to the health of the organism. However, this may well be a classic case whereby the 'absence of evidence' does not mean the hypothesis is incorrect, but instead, is due to the fact that few studies have even addressed the question. While a few early studies indicated that exposure to non-24 h LD cycles, or repeated phase shifts of the external periodic environment to disturb normal circadian organization, decreased the life span of several invertebrate species, remarkably, few similar studies have addressed this question in mammals. In one study of cardiomyopathic hamsters, chronic reversal of the external LD cycle at weekly intervals resulted in a significant decrease in the survival time: the median life span was reduced by 11%. Thus, disrupting normal circadian rhythmicity in an animal susceptible to early mortality due to cardiac disease resulted in a further decrease in longevity. The deleterious effects of repeatedly phase shifting the LD cycle in cardiomyopathic hamsters may be related to reports of increased cardiovascular morbidity and mortality in humans engaged in shift work. Recent studies by Block and colleagues have demonstrated that repeated phase advances in the LD cycle can hasten the death of aged mice, but, interestingly, phase delays did not. Taken together, these findings suggest that experimental or voluntary disruption of the 24 h temporal homeostasis may have adverse effects on the health, well-being, and longevity of the organism.

See also: Circadian Rhythm Models; Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Circadian Organization in Non-Mammalian Vertebrates; Circadian Regulation in Invertebrates; Circadian Function and Therapeutic Potential of Melatonin in Humans; Circadian Rhythms: Influence of Light in Humans; Circadian Rhythms in Sleepiness, Alertness, and Performance; Clock Genes and Metabolic Regulation; Entrainment of Circadian Rhythms by Light; Genetic Regulation of Circadian Rhythms in Drosophila; Genetics of Circadian Disorders in Humans; Mammalian Sleep and Circadian Rhythms: Flies; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Peripheral Circadian Oscillators; Photoreceptors and Circadian Clocks; Psychiatric Disorders Associated with Disturbed Sleep and Circadian Rhythms; Seasonal Hormonal Changes and Behavior; Seasonal Changes in Night-Length and Impact on Human Sleep; Serotonin and the Regulation of Mammalian Circadian Rhythms; Shift Work and Circadian Rhythms; Single Cell Neuronal Circadian Clocks; Sleep and Waking in Drosophila; Sleep and Circadian Rhythm Disorders in Human Aging and Dementia; Sleep: Development and Circadian Control.

Further Reading

Aschoff J, Saint Paul UV, and Wever R (1971) The lifespan of flies under the influence of time shifts. *Naturwissenschaften* 58: 574.

Bliwise DL (1999) Sleep and circadian rhythm disorders in aging and dementia. In: Turek FW and Zee PC (eds.) *Regulation of Sleep and Circadian Rhythms,* vol. 133, ch. 16, pp. 487–526. New York: Marcel Dekker, Inc.

Brock MA (1991) Chronobiology and aging. *Journal of the American Geriatric Society* 39: 74–91.

Davidson AJ, Sellix MT, Daniel J, et al. (1996) Chronic jet-lag increases mortality in aged mice. *Current Biology* 16(21): R914–R916.

Hurd MW, Neher I, and Ralph MR (1996) Circadian rhythms and longevity in hamsters. *Society for Research on Biological Rhythms* Abstract 5: 113.

Kawachi I, Colditz GA, Stampfer MJ, et al. (1995) Prospective study of shift work and risk of coronary heart disease in women. *Circulation* 92: 3178–3182.

Knutsson A, Akerstedt T, and Orth-Gomer K (1986) Increased risk of ischaemic heart disease in shift workers. *Lancet* 2: 89–92.

Kolker DE and Turek FW (2001) Circadian rhythms and sleep in aging rodents. In: Hof PR and Mobbs P (eds.) *Functional Neurobiology of Aging*, *Section V: Homeostasis*, pp. 869–882. San Diego: Academic Press.

Moore-Ede MC and Richardson GS (1985) Medical implications of shift work. *Annual Review of Medicine* 36: 607–617.

Nelson W and Halberg F (1986) Schedule-shifts, circadian rhythms and lifespan of freely-feeding and meal-fed mice. *Physiology & Behavior* 38: 781–788.

Penev PD, Turek FW, Wallen EP, et al. (1997) Aging alters the serotonergic modulation of light-induced phase advances in golden hamsters. *American Journal of Physiology* 272: R509–R513.

Pittendrigh CS and Minis DH (1972) Circadian systems: Longevity as a function of circadian resonance in *Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America* 69: 1537–1539.

Tuchsen F (1993) Working hours and ischaemic heart disease in Danish men: A 4-year cohort study of hospitalization. *International Journal of Epidemiology* 22: 215–221.

Turek FW, Penev P, Zhang Y, et al. (1995) Alterations in circadian system in advanced age. In: Waterhouse J (ed.) *Circadian Clocks and Their Adjustment*, pp. 212–234. London: Ciba Foundation Symposium, Pitman Press.

Turek FW, Scarbrough K, Penev P, et al. (2001) Aging of the mammalian circadian system. In: Takahashi JS, Turek FW, and Moore RY (eds.) *Handbook of Behavioral Neurobiology,* Vol. 12: *Circadian Clocks*, ch. 10, pp. 292–317. New York: Kluwer Academic/Plenum Publishers.

Van Reeth O, Olivares E, Mocauer E, et al. (1997) A melatonin agonist increases the phase-shifting effects of dark pulses in old hamster. *The 27th Annual Meeting of the Society for Neuroscience,* New Orleans, LA, vol. 420.

Circadian Organization in Non-Mammalian Vertebrates

H Underwood, North Carolina State University, Raleigh, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Daily rhythms are a universal feature of eukaryotic organisms; daily oscillations occur at all levels of organization from the molecular to the whole animal. Significantly, most of these daily rhythms persist when the organism is isolated from cyclic environmental cues, showing that they are driven by an internal timekeeper – a circadian clock. When an organism is expressing its endogenous periodicity in constant conditions, it is said to be free-running. The daily light–dark (LD) cycle is the principal cue that synchronizes, or entrains, circadian clocks, although temperature cycles are effective entraining agents for poikilotherms as well. Entrainment of circadian rhythms by environmental cycles ensures that these rhythms occur at the right time of day. Circadian clocks allow organisms to predict, and prepare, for changes in the external environment; for example, a lizard can predict the coming of colder nighttime temperatures and seek shelter while surface temperatures are still conducive for activity or digestive enzyme levels can rise in anticipation of a meal. A popular hypothesis as to the origin of circadian rhythmicity proposes that rhythmicity originally evolved in response to the daily shower of ultraviolet radiation from the sun. Colored molecules are particularly prone to ultraviolet damage; consequently, organisms partitioned their biochemical activities so that light–sensitive reactions occurred during the dark phase of the daily light–dark cycle.

Conceptually, a basic circadian system must comprise at least one photic input pathway to convey entraining light information to the circadian clock, the circadian clock itself, and an output pathway coupling the clock to its overt rhythm. The expectation of early circadian biologists was that specific, presumably neural, sites within the animal played a significant role in generating circadian rhythmicity; that is, one (or several) circadian clocks controlled the myriad rhythms expressed throughout the body.

Much of the work on circadian systems in vertebrates, therefore, has focused on identifying the sites that generate circadian rhythmicity and their photic inputs. Within the last 35 years, remarkable strides have been made in identifying areas that play a major role within the circadian system of vertebrates. This journey has had its share of rewards and surprises as biologists have revealed the existence of multiple photic inputs, multiple circadian pacemakers, and the existence of cellular clocks spread throughout the organism.

Pineal Organ and Melatonin

The first organ that was shown to play a major role within the vertebrate circadian system was the pineal organ; it was demonstrated that removal of the pineal organ caused the activity rhythm of house sparrows to become arrhythmic in constant darkness (DD). Activity rhythms are frequently used by circadian biologists to monitor the behavior of the circadian clock because activity is easily measured, it can be monitored continuously, and it does not require that the animal be restrained or biopsied. Pineal organs are found in virtually all vertebrates, and they develop as evaginations of the roof of the diencephalons. The pineal organ is located on the surface of the brain in many vertebrates, making it an attractive site for investigation because it can be easily removed. Following the initial observation that pinealectomized house sparrows are rendered arrhythmic in constant conditions, investigators showed that pinealectomy had major effects on the activity rhythms of other non-mammalian vertebrates as well, including at least five species of fish, one species of amphibians, six species of lizards, and nine species of birds (Figure 1). Interestingly, pinealectomy did not always produce arrhythmicity; in some cases it caused a change in the period of the free-running rhythm or, more rarely, it caused the activity rhythm to split into two independently free-running circadian components. These three effects of pinealectomy – arrhythmicity, period changes, or splitting – could sometimes be observed between individuals of the same species. The most reasonable model to explain these diverse effects of pinealectomy posits that the pineal organ acts as a circadian pacemaker that controls the period and phase of subordinate oscillators. The subordinate oscillators are mutually coupled, and it is the inter- and intraspecific variations in the strength of coupling between the subordinate oscillators that accounts for the different effects of pinealectomy. If the coupling strength between the subordinate oscillators is strong, the removal of the pacemaker causes a change in the period of the coupled subordinate oscillators, whereas if coupling is weak, the subordinate oscillators do not remain coupled but will split apart and each subordinate clock will begin expressing its own endogenous periodicity.

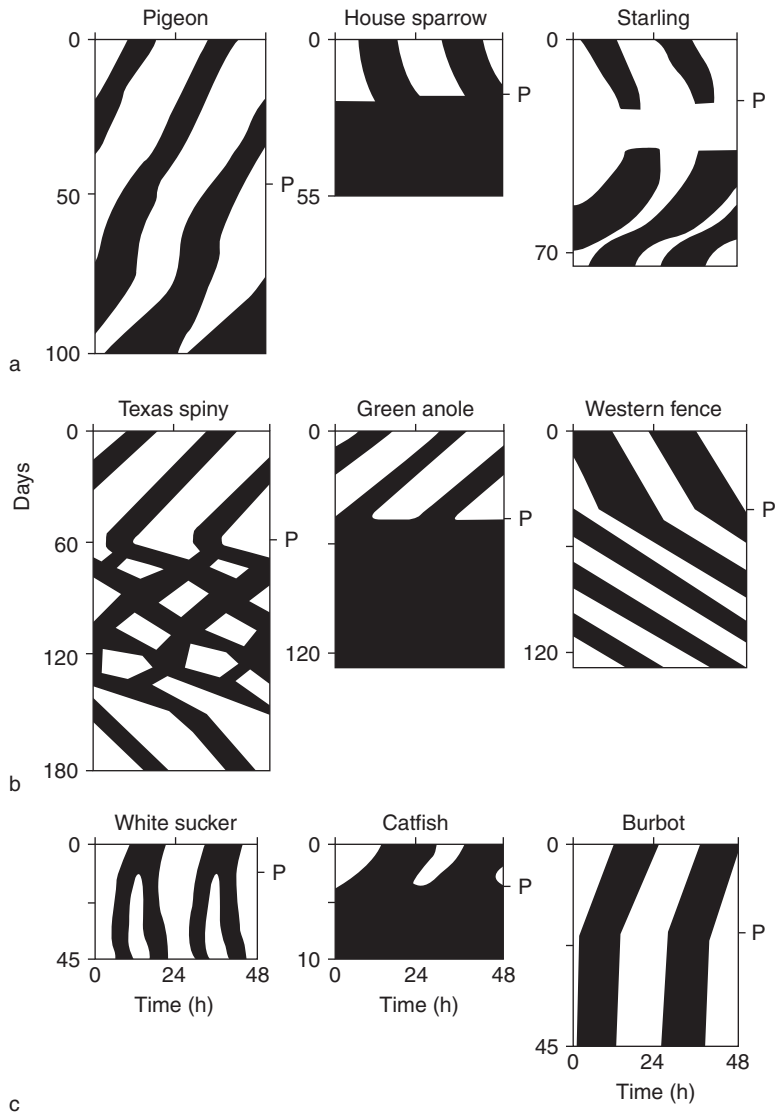

Figure 1 Schema of activity patterns showing the effects of pinealectomy (P) in three classes of vertebrates: (a) birds; (b) lizards; (c) fish. Pinealectomy has one of three kinds of effects: splitting of the circadian activity rhythm into two circadian components (Texas spiny lizard and white sucker), arrhythmicity (house sparrow, green anole, and catfish), or a change in the period of the free-running rhythm (starling, western fence lizard, and burbot). From Underwood H, Wassmer GT, and Page TL (1997) Daily and seasonal rhythms. In: Dantzler WH (ed.) *Handbook of Comparative Physiology, Vol. 13: Comparative Physiology*, pp. 1653–1763. New York: Oxford University Press.

One experiment is particularly informative and bears directly on the question as to whether or not the pineal organ is indeed a pacemaker controlling the organism's entire circadian system. Zimmerman and Menaker transplanted pineal organs from donor birds entrained to light–dark cycles (in which light onsets occurred at either 09.00 or 23.00) into pinealectomized arrhythmic house sparrows in constant darkness (DD) (Figure 2). The recipients began showing a rhythm within a day or two of transplant, and, significantly, the rhythm began from the phase dictated by the donor's pineal, approximately 09.00 or 23.00. The sparrow's pineal, therefore, is indeed the site of a circadian pacemaker because an important property of a pacemaker – its phase – was transplanted along with the pineal.

Pineal organs have an interesting evolutionary history. The principal cell type in the mammalian pineal is a secretory cell, the pinealocyte, which is believed to have evolved from a photosensory cell. Indeed, the main cell type in the pineal organs of fish, amphibians, and reptiles is a cell that resembles in many aspects the cone photoreceptors of the lateral eyes; it is bipolar, it has a stack of photoreceptive discs (the

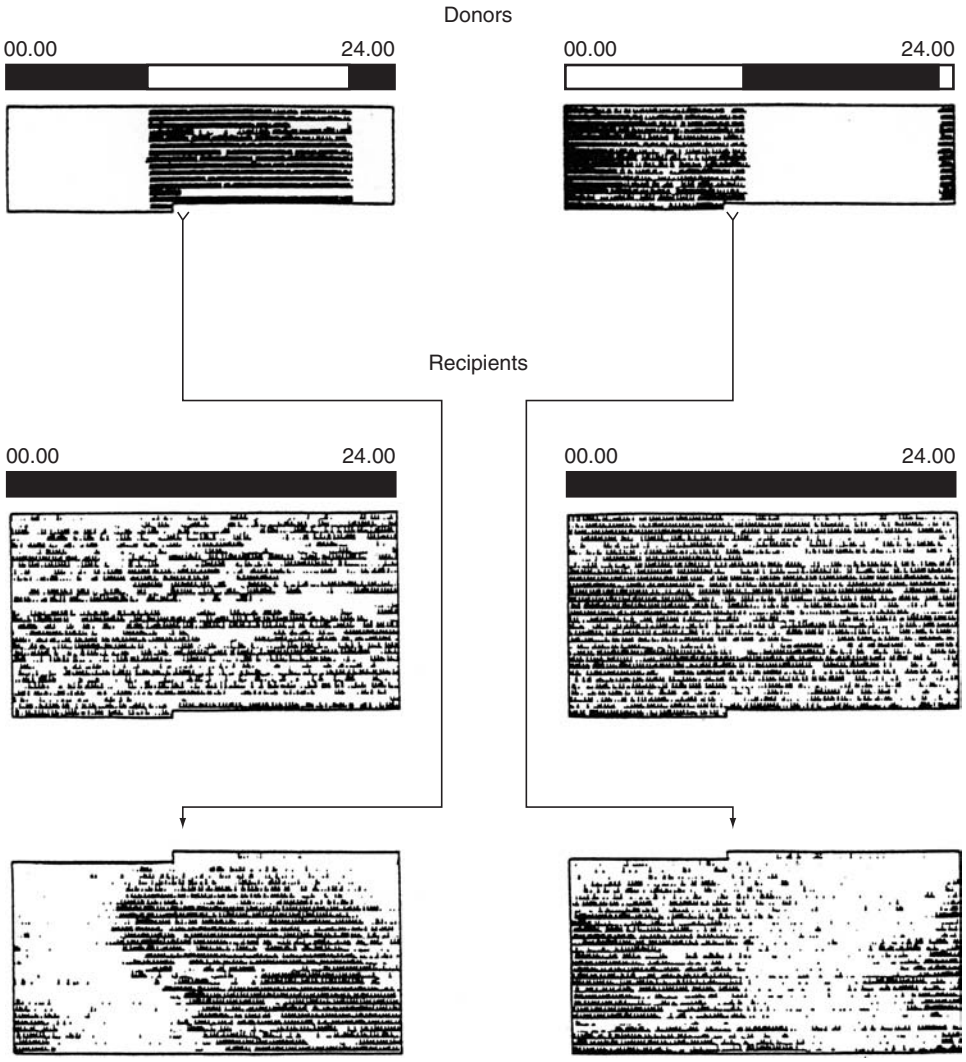

Figure 2 Evidence that the avian pineal is the site of a circadian pacemaker. The perching activity records of the two birds shown in the top records were entrained to light–dark cycles that were 14 h out of phase with one another. Subsequently, their pineal organs were removed and transplanted into two birds (middle records) that had become arrhythmic in constant darkness (DD) following pinealectomy. A free-running rhythm emerged in the transplanted birds (bottom records) and, significantly, the initial activity onsets occurred at the same time of day as the activity onsets in the donor birds. The pineal is the site of a circadian pacemaker because an important property of the pacemaker – its phase – can be transplanted along with the pineal. From Zimmerman NH and Menaker M (1979) The pineal gland: A pacemaker within the circadian system of the house sparrow. *Proceedings of the National Academy of Sciences of the United States of America* 76: 999–1003.

outer segment) containing the photopigments, it has an inner segment containing the nucleus, and it has a synaptic pedicle that makes contact with other neurons. However, in many cases this cell is more degenerate in appearance than a typical cone photoreceptor and, particularly in birds, the outer segment may be reduced to a simple whorl of membranes. A variety of studies have shown that the pineal organs of all the nonmammalian vertebrates are directly light-sensitive. The photosensory cell is also capable of considerable biosynthetic activity; in particular, this cell type can synthesize and secrete the hormone melatonin. Pineal melatonin is always synthesized rhythmically, with higher melatonin synthesis occurring during the night in both diurnal and nocturnal animals. Melatonin is not stored in the pineal, and this lipid-soluble hormone rapidly enters the blood, so blood melatonin levels increase at night as well.

The fact that pinealectomy had major effects on the circadian system immediately focused attention on a role for melatonin in the circadian system because the pineal is one of the few sites that produces melatonin and, in many species, the pineal is the sole source of blood-borne melatonin. To determine if melatonin could affect the circadian system, melatonin was exogenously administered to amphibians, lizards, and

birds using a variety of approaches, including implanting silastic capsules containing melatonin under the skin or injecting or infusing melatonin on a daily basis. Continuous exposure to pharmacological dosages of melatonin administered via silastic capsules caused period changes or arrhythmicity in a variety of birds (e.g., house sparrows, Java sparrows, house finches, Japanese quail, and pigeons) and lizards (Western fence lizard, Texas spiny lizard, ruin lizard, and green anoles). Daily administration of pharmacological doses of melatonin via injections, infusions, or drinking water entrained the activity rhythms of amphibians (Japanese newts), birds (starlings, Japanese quail, and pigeons), and lizards (Western fence lizard). Most convincingly, daily infusions of physiological levels of melatonin that accurately replaced endogenous levels in pinealectomized *Anolis* lizards and pigeons also entrained the activity or feeding rhythms. Taken together, these experiments support the hypothesis that there is a circadian pacemaker in the pineal organ that controls the period and phase of subordinate oscillators via the rhythmic synthesis and release of melatonin.

The pineal organs of a variety of nonmammalian vertebrates have been studied *in vitro*, primarily by measuring the synthesis and secretion of melatonin. In most cases, the pineal exhibited a light-entrainable rhythm of melatonin that persisted in constant conditions, proving that the rhythm is being driven by an intrapineal clock. The pineal pacemaker in comprised of multiple circadian clocks – individual pineal cells of chickens or *Anolis* lizards can exhibit a circadian rhythm of melatonin secretion and are directly light-entrainable *in vitro*. Therefore, the pineal pacemaker is a multioscillator system in which hundreds of individual light-entrainable pineal clocks mutually interact and produce a single coherent output in the form of the daily melatonin pulse.

Studies have shown that the amplitude and duration of the daily melatonin rhythm change in response to seasonal changes in day length in homeotherms and to seasonal changes in day length and temperature in poikilotherms. The pineal organ, therefore, acts as a photo- or photothermoendocrine transducer that converts environmental stimuli into an internal melatonin signal that accurately reflects the ambient environmental conditions. In turn, melatonin controls the phases of subordinate clocks so that the right internal events occur at the proper time of day.

Hypothalamic Clocks

Beginning in the 1970s, investigators used a variety of approaches including lesions of the suprachiasmatic nuclei (SCN) of the hypothalamus, *in vitro* recordings of electrical activity from SCN slices, and SCN transplantation to show that the SCN was the main circadian pacemaker in mammals. Studies in mammals also revealed that lighting information reaches the SCN via a monosynaptic retinohypothalamic tract (RHT) from the retina. Taking note of these experiments, investigators lesioned the suprachiasmatic area of nonmammalian vertebrates, including Japanese quail, Java sparrows, house sparrows, desert iguanas, and ruin lizard, to determine whether the suprachiasmatic area was playing a significant role in the circadian system of these vertebrates as well. As in mammals, these lesions abolished circadian rhythms of activity, rendering the animals arrhythmic in constant conditions. A significant complication in investigations into the role of the SCN, however, is the fact that the anatomy of the SCN in nonmammalian vertebrates is often more complicated that that in mammals.

We must be cautious in concluding that the nonmammalian SCN is the site of a major circadian pacemaker based solely on the effects of SCN lesions. Although the fact that ablation of a specific area renders an animal arrhythmic is compatible with the hypothesis that the site is the locus of a circadian pacemaker, it is not conclusive. It is conceivable, for example, that the pacemaker is located elsewhere and only its output pathway is disrupted by the lesion. Unfortunately, other studies such as *in vitro* studies of the SCN or SCN transplantation studies have yet to be performed in nonmammalian vertebrates. However, the fact that lesions of the suprachiasmatic area disrupt rhythmicity in nonmammalian vertebrates, and the abundant evidence in mammals that the SCN act as circadian pacemakers, strongly supports the hypothesis that the suprachiasmatic area of nonmammalian vertebrates is the site of the subordinate oscillators described previously. The pineal controls the SCN oscillators via the rhythmic secretion of melatonin. Supporting this hypothesis is the fact that melatonin receptors are found within the SCN.

In contrast to mammals, however, the central oscillators in many nonmammalian vertebrates do not maintain rhythmicity in the absence of periodic inputs, either from the daily LD cycle or from other pacemakers such as the pineal organ. For example, activity rhythms dampen into arrhythmicity over a period of days in most birds following pinealectomy. Arrhythmicity is also observed following pinealectomy in some (but not all) fish, amphibians, and lizards. Overall, these data suggest that the hypothalamic clocks often show less autonomously rhythmicity than the mammalian SCN, especially in birds. In these

species, periodic input, such as a daily melatonin pulse from the pineal, is required to maintain rhythmicity; the hypothalamic clocks either damp out or become desynchronized in the absence of this input.

This idea has been expanded to describe a mutual dependence between the pineal and the SCN in some birds. The neuroendocrine loop hypothesis proposes that neither the pineal nor the SCN is autonomously rhythmic in some avian species. Persistent rhythmicity in the pineal requires periodic sympathetic input from the SCN, whereas persistent rhythmicity in the SCN requires a daily pulse of melatonin from the pineal; consequently, neither the pineal nor the SCN is persistently rhythmic on its own.

As previously mentioned, some controversy surrounds the precise location within the nonmammalian hypothalamus of the homolog of the mammalian SCN. In birds, for example, neuroanatomical and neurochemical studies suggest that two anatomically distinct nuclei – the medial hypothalamic nucleus (MHN; also termed the periventricular preoptic nucleus) and the lateral hypothalamic retinorecipient nucleus (LHRN; also termed the visual SCN) – are possible homologs of the mammalian SCN. Lesioning the suprachiasmatic area of birds has been shown to disrupt rhythmicity, but these lesions probably damaged both nuclei. Recent advances in pacemaker localization, however, may clarify the precise location of hypothalamic pacemakers. A small suite of genes have been identified that code for essential components of the molecular clock in all animals from *Drosophila* to mammals. The heart of the molecular clock consists of a transcription–translation negative feedback loop in which cyclic gene transcription causes circadian variations in proteins that feedback and inhibit their own transcription. Because these clock genes have been cloned, it is now possible to determine their expression patterns in different tissues. In birds, the rhythmic expression of clock genes has been observed in the MHN, but not in the LHRN, of Japanese quail, Java sparrows, chickens, and pigeons, but the rhythmic expression of clock genes has been seen in both nuclei in house sparrows. The location of hypothalamic pacemakers in birds, therefore, may depend on species, being located in the MHN of some and in both the MHN and LHRN of others.

The Eyes as Photoreceptors

In mammals, lighting information is transmitted from the eyes to the SCN via a direct, monosynaptic neural pathway, the RHT. The only way light can entrain the mammalian circadian system is via the eyes. Surprisingly, however, early studies showed that entrainment by light–dark cycles can still occur in rodents in which the rod and cone photoreceptors were missing. Subsequent investigations revealed that a photopigment, melanopsin, found in a small subset (1–3%) of ganglion cells, could mediate entrainment in rodless and coneless rodents. These ganglion cells are intrinsically photosensitive, and they have a number of characteristics that are optimized for detecting the presence and intensity of light such as large receptive fields and a higher threshold for light detection than the classical photoreceptors. Significantly, the axons of these ganglion cells comprise the RHT. The melanopsin-containing ganglion cells, therefore, are fully capable of detecting LD information and transmitting this entrainment signal to the SCN. The threshold intensity of light that can entrain the activity rhythm of mice, however, is slightly higher in rodless and coneless animals than in animals with rods and cones, supporting the hypothesis that both the classical photosensory cells and the melanopsin-containing ganglion cells normally mediate entrainment.

Although the eyes play a role in transmitting LD information to the central circadian system of nonmammalian vertebrates, they are not the only photic inputs. As described previously, the pineal organ of nonmammalian vertebrates are directly photosensitive. However, fish, amphibians, lizards and birds in which both the eyes and the pineal organ have been removed are still entrainable by light, showing the existence of extraocular and extrapineal photoreceptors as well (termed extraretinal photoreceptors (ERRs)).

Along with the ERRs, the eyes of nonmammalian vertebrates are clearly involved in the perception of entraining the LD cycle because the light intensity threshold for entrainment is slightly higher in eyeless fish, lizards, and birds than it is in intact animals. Nonmammalian vertebrates possess a RHT, and it is undoubtedly via the RHT that lighting information is being transmitted to the hypothalamic clocks. The photopigment melanopsin is also found in the retinas of nonmammalian vertebrates and may be involved in mediating entrainment to LD cycles, as it does in mammals, although this has yet to be proven.

The Eyes as Circadian Pacemakers

Many events within the vertebrate eye show circadian rhythms, including the shedding of discs from photoreceptor outer segments, photomechanical movements of rods and cones, synaptic morphology, enzyme activities, and melatonin synthesis. Many ocular rhythms are driven by circadian clocks within the

eye; for example, rhythms of melatonin synthesis can persist in retinas held *in vitro*. Interestingly, the eyes of some nonmammalian vertebrates can also act as circadian pacemakers, controlling rhythms outside the eye. Such a role has been established in at least three lizard and two avian species. Removing the eyes of ruin lizards, Texas spiny lizards, and western fence lizards causes marked changes in the period of the free-running activity rhythm in DD, suggesting that the eyes may be the loci of circadian pacemakers that are significantly involved in controlling the period of central circadian clocks. In pigeons, the removal of the eyes abolishes rhythmicity in DD, but only if the pigeons have also been pinealectomized; eye removal alone, or pinealectomy alone, does not abolish rhythmicity, although it does affect the clarity of the activity and body temperature rhythms. Both the eyes and pineal in pigeons synthesize and secrete melatonin rhythmically, and infusions of physiological levels of melatonin that replace endogenous levels in pinealectomized plus blinded pigeons restore rhythmicity. In the pigeon, therefore, both the eyes and pineal are acting as pacemakers that control central oscillators (presumably in the SCN) via the rhythmic synthesis and secretion of melatonin. In the Japanese quail, the ocular pacemaker plays the predominant role because eye removal alone causes the quail to become arrhythmic in DD whereas pinealectomy has little effect. Additional experiments in quail have shown that the ocular pacemaker controls the central hypothalamic clocks both via a neural output and via rhythmic melatonin secretion.

The cells within the retina that contain the circadian pacemaker have not been definitively established, although the photosensory cells are likely sites because the rod/cone photoreceptors are known to possess circadian clocks and they also synthesize melatonin.

Extraretinal Photoreceptors

One of the most intriguing features of the circadian systems of nonmammalian vertebrates is the involvement of extraretinal photoreceptors; fish, amphibians, reptiles, and birds can be entrained by light–dark cycles after removal of their eyes and pineal systems. Blind lizards and birds can be entrained by light cycles in which the light is as dim as full moonlight (0.1 lux). These extraretinal photoreceptors are probably universally present in nonmammalian vertebrates because no cases have been found in which entrainment to a LD cycle failed after blinding, unless the light intensity was very dim.

An interesting question is: Why do animals that possess eyes also need to have extraretinal photoreceptors? We could argue that that primitive organisms possessed direct light sensitivity for entrainment before the development of the eye, but why has extraretinal photoreception persisted in nonmammalian vertebrates after the development of the eyes? A definitive answer to this question has not yet been made, but the fact that extraretinal photoreceptors are a universal feature of present-day nonmammalian vertebrates argues that there must be a significant selective advantage in maintaining them.

The extraretinal photoreceptors are located in the brain; blocking light penetration to the brain eliminates entrainment in blinded animals and localized illumination of the brain can cause entrainment. Extraretinal photoreceptors are presumably well-suited for irradiance detection (rather than a focused representation of the environment) because overlying tissues scatter light. The circadian system is also designed to integrate the total number of photons received and to do so over a long time span. It is not yet known exactly where these extraretinal photoreceptors are located in the brain or whether their nature or location varies between species or between classes of vertebrates. However, various opsins and other members of the phototransduction cascade have been identified in the hypothalamus (e.g., SCN and preoptic areas) and in the periventricular cerebrospinal fluid (CSF)-contacting neurons of the lateral septum of a variety of nonmammalian vertebrates. Most of the functional studies that have attempted to localize extraretinal photoreceptors have used the seasonal reproductive response to changes in day length (photoperiod) as a measure of photoreception, especially in birds, and these studies also implicate the infundibular zone of the hypothalamus and the lateral septum as photosensory sites. It has yet to be definitively established whether these putative photoreceptive areas mediate entrainment of the circadian clock. However, the fact that a major part of the circadian system resides in the hypothalamus supports the idea that the hypothalamus is a likely site for deep brain photoreceptors that mediate entrainment as well as seasonal photoperiodic responses.

Peripheral Oscillators

Until fairly recently, it was assumed that the SCN of mammals, or multiple interacting pacemaking sites (pineal, eyes, and SCN) of nonmammalian vertebrates were solely responsible for driving the myriad overt rhythms seen in individual vertebrates. This model, however, has undergone significant revision with the demonstration that (1) clock genes are expressed not only in neural pacemaker cells such as the SCN but also in a variety of other tissues and

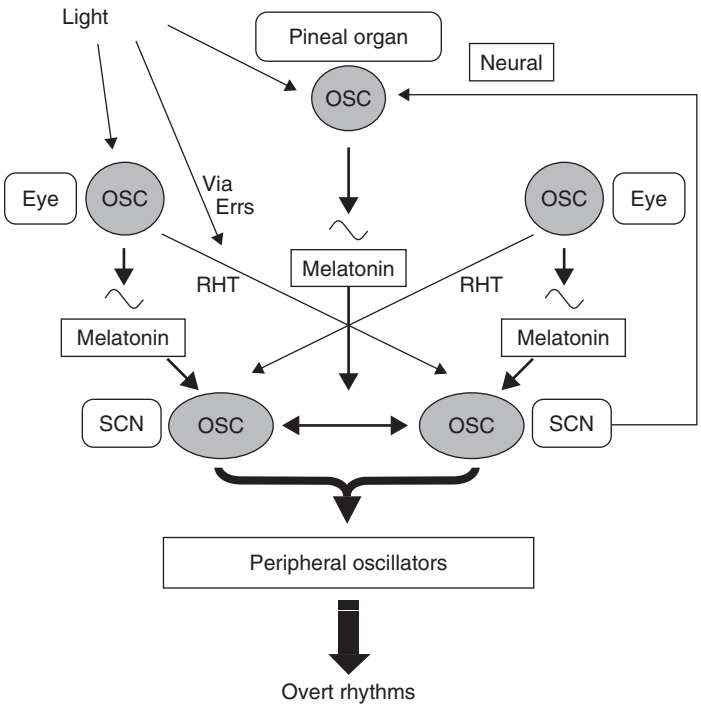

Figure 3 Model of the circadian system of nonmammalian vertebrates. Circadian oscillators (OSC) reside centrally in the suprachiasmatic area of the hypothalamus and more peripherally in the pineal and eyes. The pineal is linked to the rest of the system via melatonin, whereas the eyes may be linked by melatonin or neurally, probably via the retinohypothalamic tract (RHT). Light can entrain oscillators in the pineal and eyes directly, it can entrain the central system via extraretinal receptors (Errs) in the brain, and it can entrain the central system via photoreceptors in the pineal and eyes. In some species, persistent rhythmicity in the pineal and suprachiasmatic nucleus (SCN) may depend on a mutual interaction between them.

(2) rhythmic gene expression can continue in cells, tissues, and organs cultured *in vitro*. In fact, it is likely that every vertebrate cell contains a competent molecular clock. Clocks located outside of the central nervous system have been termed peripheral oscillators.

The existence of peripheral oscillators raises interesting questions about the role of pacemakers located in the brain or eyes (the SCN of mammals or the SCN/pineal/eyes of nonmammalian vertebrates) within the entire circadian system. The crucial parameter that determines whether these pacemakers play a major role may be whether the peripheral oscillators are directly photosensitive. If a peripheral clock can be entrained directly by light, it may not require control by other pacemakers.

Clearly, in mammals the SCN must act as the master pacemaker that controls, via hormonal and/or neural links, the period and phase of peripheral oscillators because peripheral oscillators in mammals cannot be entrained directly by LD cycles; LD cycles entrain the SCN, which, in turn, controls peripheral clocks. However, studies of clock gene expression in zebra fish showed that a variety of organs, such as the kidney and heart, as well as individual cells, exhibit light-entrainable rhythms *in vitro*. In the zebra fish, therefore, the possibility exists that a master clock is not needed to control the period or phases of peripheral oscillators.

Peripheral clocks have also been identified in various tissues in other nonmammalian vertebrates such as birds, but there are no reports that these clocks are directly light-entrainable. More studies are needed to determine which nonmammalian vertebrates have light-entrainable peripheral clocks. Perhaps direct-light entrainment is confined to small fish that are relatively transparent to light.

Investigations into the circadian system of nonmammalian vertebrates over the last few decades showed that, in contrast to mammals, the circadian system is composed of several interacting pacemaker sites, although there are some significant differences between species in the relative roles that these sites play (Figure 3). Also, in contrast to mammals, light can entrain the circadian system via several routes.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Entrainment of Circadian Rhythms by Light; Mammalian

Sleep and Circadian Rhythms: Flies; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Peripheral Circadian Oscillators; Photoreceptors and Circadian Clocks; Photoreceptors: Physiology; Pineal Gland and Melatonin; Sleep: Development and Circadian Control.

Further Reading

Abraham U, Albrecht U, Gwinner E, and Brandstatter R (2002) Spatial and temporal variation of Passer Per2 gene expression in two distinct cell groups of the suprachiasmatic hypothalamus in the house sparrow (*Passer domesticus*). *European Journal of Neuroscience* 16: 429–436.

Dunlap JC, Loros JJ, and DeCoursey PJ (eds.) (2004) *Chronobiology. Biological Timekeeping*. Sunderland, MA: Sinauer Associates.

Roenneberg T and Merrow M (2005) Circadian clocks – the fall and rise of physiology. *Nature Reviews Molecular Cell Biology* 6: 965–971.

Tosini G, Bertolucci C, and Foa A (2001) The circadian system of reptiles: A multioscillatory and multiphotoreceptive system. *Physiology & Behavior* 72: 461–471.

Underwood H (2001) Circadian organization in nonmammalian vertebrates. In: Takahashi J, Turek FW, and Moore RY (eds.) *Handbook of Comparative Physiology, Vol. 12: Circadian Clocks*, pp. 111–140. New York: Kluwer Academic/Plenum Publishers.

Underwood H and Groos GA (1982) Vertebrate circadian rhythms: Retinal and extraretinal photoreception. *Experientia* 38: 1113–1121.

Underwood H, Steele C, and Zivkovic B (2001) Circadian organization and the role of the pineal in birds. *Microscopy Research and Technique* 53: 48–62.

Underwood H, Wassmer GT, and Page TL (1997) Daily and seasonal rhythms. In: Dantzler WH (ed.) *Handbook of Comparative Physiology, Vol. 13: Comparative Physiology*, pp. 1653–1763. New York: Oxford University Press.

Yoshimura T, Yasuo S, Suzuki Y, et al. (2001) Identification of the suprachiasmatic nucleus in birds. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 280: R1185–R1189.

Zimmerman NH and Menaker M (1979) The pineal gland: A pacemaker within the circadian system of the house sparrow. *Proceedings of the National Academy of Sciences of the United States of America* 76: 999–1003.

Circadian Oscillations in the Suprachiasmatic Nucleus

W J Schwartz, University of Massachusetts Medical School, Worcester, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

The suprachiasmatic nucleus (SCN) in the anteroventral hypothalamus is a compact bilaterally paired nucleus straddling the midline, bordering the third ventricle, bounded anteroventrally by the optic chiasm (Figure 1). Best characterized in rodents, it is less than 1 mm long and consists of about 16 000 neurons that are tiny (<10 μm diameter) and very tightly packed. It is the site of an endogenous clock mechanism that governs daily behavioral, physiological, and endocrine rhythms. At the core of the clock is a self-sustaining, light-entrainable pacemaker that oscillates with a circadian (*circa*, about; *dian*, day) period when it is maintained in a constant environment free of external timing cues. Experimental destruction of the SCN by electrolytic lesions leads to a breakdown of the entrainment or generation of a wide range of circadian rhythms, without restoration of function even after prolonged postoperative survival. Transplants of fetal SCN tissue can reestablish overt behavioral rhythmicity in arrhythmic, SCN-lesioned recipients, and the rhythms restored by the transplants display properties that are characteristic of the circadian pacemakers of the donors rather than those of the hosts.

Circadian Rhythmicity of SCN Tissue *In Vivo* and *In Vitro*

In order to begin to understand the clock's mechanism, several assays have been employed as markers for the amplitude, phase, and period of the circadian pacemaker's oscillation. The use of multiple measures helps to distinguish those experimental manipulations that eliminate overt rhythms by actually arresting the pacemaker from those that uncouple a specific output from the still-oscillating pacemaker (loss of the clock's 'hands' rather than damage to its 'gears').

SCN metabolic activity was the first property of the nucleus reported to exhibit circadian rhythmicity. These experiments utilized an autoradiographic method for the *in vivo* determination of the rates of glucose utilization of individual structures within the brain by injecting animals with tracer amounts of 2-deoxy-D-[1–^{14}C]glucose. This assay is useful because regional brain functional activity is closely coupled to regional brain energy utilization, which is dependent on the continuous provision of glucose. SCN glucose utilization is high during the light phase (day) of a light–dark cycle and low during the dark phase (night) in both nocturnally and diurnally active animals. The rhythm persists when animals are maintained in constant darkness (subjective day and night), occurs in fetal animals *in utero*, and continues in acutely prepared SCN explants *in vitro*. A limitation of the deoxyglucose method is that injected animals must be killed in order to prepare brain sections for autoradiography; thus, the metabolic rhythm is not continuously recorded from individual subjects, each monitored longitudinally over several circadian cycles, but instead from groups of subjects sacrificed and sampled at different phases of a single circadian cycle. Methods that rely on such pooled groups, like the deoxyglucose method and those that measure mRNA or protein levels on brain sections, may lead to confusing results if rhythms among individual subjects are not synchronous. Apparent arrhythmicity of such a population might be due either to a loss of rhythmicity of each subject or to a desynchronization of rhythmicity between subjects if each of the individuals expresses a rhythm that differs in phase or period.

One rhythm that can be measured longitudinally in individuals is SCN electrical activity. *In vivo*, the overall firing rate of multiple neurons or the discriminated spikes from single neurons can be recorded using an extracellular electrode; discharge rates are low (around 1 Hz), double in frequency between day and night, and are high during the day or subjective day and low during the night or subjective night in both nocturnal and diurnal rodents. Circadian rhythms of single or multiple unit electrical activities have also been recorded from the SCN in hypothalamic slices *in vitro*. In general, these studies have been performed by sampling the spike activity of single units for short time intervals at different phases across the circadian cycle, pooling the data from a number of slices and phases, and then determining the phase of peak firing rate for the population of units as a whole. Workers using this method must make repeated electrode penetrations of the tissue, and they usually record from individual cells for a few minutes in slices kept for several hours, although there are reports of single neurons held for 14 h and single slices maintained for a few days. Although technically demanding, long-term (>24 h) multiple-unit recordings using single electrodes have also been accomplished in slices. Independent of method, the amplitude of the electrophysiological rhythm *in vitro*

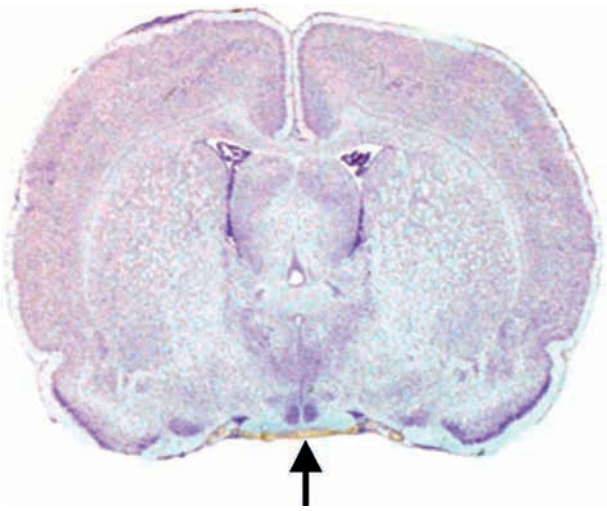

Figure 1 Coronal Nissl-stained section through the rat forebrain, including the bilaterally paired SCN (arrow).

appears to be larger than *in vivo*, with a peak level about 4–5 times higher than baseline values, and without some of the high-frequency components observed *in vivo*. It is not yet understood how neuronal firing rate relates to the coding of efferent signals sent by the SCN to other brain regions.

Most recently, a restricted set of 'clock' genes has been identified and is believed to underlie the actual oscillatory mechanism of the circadian pacemaker; moreover, a significant percentage of the whole genome is modulated by the clock. Rhythms of neuroactive peptides synthesized in the SCN, usually measured in tissue punches and sections at the mRNA and protein levels, provide a valuable window on clock-controlled gene expression (transcription and protein translation and degradation). Within the SCN, mRNA and peptide levels of the neuropeptide transmitter arginine vasopressin (AVP) exhibit circadian rhythmicity, and nuclear run-on analysis has demonstrated that mRNA abundance is regulated by transcription; posttranscriptional mechanisms also contribute to the temporal pattern of SCN AVP expression. This peptide rhythm has been followed longitudinally in the cerebrospinal fluid (CSF) of individual animals and in the culture medium of individual explants. CSF levels are high during the subjective day and low during the subjective night in both nocturnal and diurnal animals, and the rhythm persists *in vitro* in hypothalamic slices and even dissociated cell cultures. The demonstration that two peptide rhythms (AVP and another SCN neuropeptide transmitter, vasoactive intestinal polypeptide (VIP)) can be serially and simultaneously measured *in vitro* is a powerful advantage of this approach.

Multiple Single-Cell Circadian Oscillators within the SCN

New technologies have enabled researchers to establish that SCN rhythmicity is a product of multiple individual cellular circadian oscillators that are coupled together in a tissue network. Fixed multielectrode arrays have provided a method that allows for the simultaneous monitoring of neuronal firing rates from individual dispersed cells in culture. Single cells dissociated from neonatal SCN and cultured at low density express circadian firing rhythms with widely varying phases, in part because different cells express independent circadian periods. This result shows that the SCN is a multioscillator system and suggests that individual SCN cells can act as competent circadian oscillators.

Several laboratories have now taken advantage of rhythmic 'clock' gene transcription to introduce reporter gene constructs for real-time monitoring of gene expression, and high-sensitivity cameras have made it possible to record the oscillation of such bioluminescent (e.g., luciferase) or fluorescent (e.g., enhanced green fluorescent protein (GFP)) reporters from individual cells. The kinetics of such reporters allow for monitoring changes in gene activity that are slower than approximately 1.5 h. GFP fluorescence is bright enough to be visualized in individual cells but requires repeated illumination with potentially damaging consequences. Bioluminescence from luciferase is dimmer but requires delivery of substrate (luciferin) to the region under study.

SCN slices made from neonatal transgenic mice carrying the promoter for the mouse *Period1* (*Per1*) gene linked to the gene for either GFP or luciferase show robust circadian rhythms in light emission. Notably, individual SCN cells in the slices show rather large phase differences in the peaks of their rhythms that persist over repeated cycles (i.e., the individual cellular periods are similar and stable) (**Figure 2**). The phase order does not appear to be a random property of the cellular network in the slice, because it is restored in *Per1::luciferase* slices after a protein synthesis inhibitor is applied to first stop and then reset the cellular oscillations to the same initial phase. These intercellular phase differences are not an artifact of transgenic preparations, because similar differences have been demonstrated by electrophysiological methods in wild-type rat SCN slices. Thus, the duration of high molecular and electrophysiological activities of individual SCN cells coupled together in a slice differs from the composite activities of the tissue as a whole (which generally lasts for most of the subjective day). The circadian waveform of synchronized SCN tissue activity is assembled from subpopulations or individual cells, each expressing

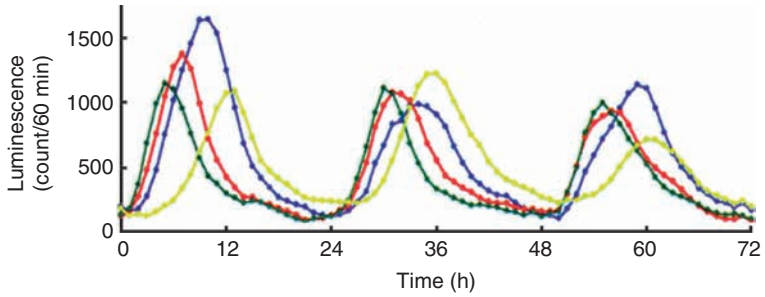

Figure 2 Circadian rhythms of bioluminescence from four individual SCN neurons within a cultured hypothalamic slice made from a neonatal transgenic mouse carrying a mouse *Per1* promoter-driven luciferase reporter. The signal intensity from each cell (in colors for this illustration) is integrated at 1 h intervals and plotted, with the trough time of total SCN tissue bioluminescence set to time = 0. Reprinted with permission from Yamaguchi S, Isejima H, Matsuo T, et al. (2003). Synchronization of cellular clocks in the suprachiasmatic nucleus. *Science* 302: 1408–1412, Copyright (2003) AAAS.

narrower activity peaks sharing a similar period but differing in phase.

The functional significance of heterogeneous cellular phases is not known. It is possible that their distribution and clustering might be configured by afferent input. Perhaps heterogeneous cellular phases might play a role as part of the temporal programming of SCN outputs. Or, induced plasticity of phase differences might permit the encoding of a photoperiodic signal (see below).

Structural and Functional Heterogeneity of SCN Neurons

While the discussion thus far would seem to imply that the SCN contains a homogeneous population of equivalent neurons, it has been recognized since 1980 that this is not the case. Nissl and silver stains, Golgi impregnations, and electron microscopy of the rat SCN were the first techniques to reveal the presence of two predominant subdivisions: cells in the dorsomedial part of the nucleus are smaller and more tightly packed than those in the ventrolateral part. This dorsal/ventral distinction corresponds to regional differences in chemoarchitecture and the topography of afferent and efferent connections. Immunohistochemical identification of peptides in SCN cell bodies marks the dorsomedial (AVP) and ventrolateral (VIP) subdivisions (Figure 3); most SCN inputs (from the retina, raphe, and lateral geniculate) are segregated to the ventral part of the nucleus. The dorsomedial ('shell') and vetrolateral ('core') organization of the rat SCN has been extended to other species as well, but the comparative anatomy appears to be more complex. The golden hamster SCN, for example, also contains a defined retinorecipient caudocentral region; cells expressing the calcium-binding protein calbindin occupy this region but represent a minority of all the cells in the area.

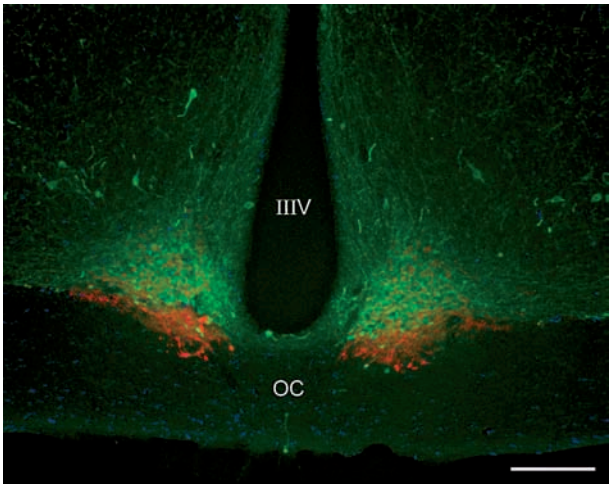

Figure 3 Immunohistochemical AVP and VIP protein expression in a coronal section of the rat SCN, processed for double-label immunofluorescence and viewed using excitation wavelengths of 488 nm (green, for AVP) and 568 nm (red, for VIP). IIIV, third ventricle; OC, optic chiasm. Scale bar = 200 μm. Courtesy of Drs. Matthew Paul and Hoi Pang Low, University of Massachusetts Medical School.

Only a subset of all SCN cells are directly retinorecipient. Estimates of photoresponsiveness range from about one-fifth (by administering a light pulse and then analyzing the immunohistochemical expression of c-Fos, the protein product of a photoinducible gene) to about one-third (by electrophysiological techniques) of SCN cells in the ventral part of the nucleus. No more than about 20% of the cells in this subset can be attributed to any one identified peptidergic phenotype. Moreover, no peptidergic cell type uniformly produces photosensitive responses; for example, VIP neurons positioned ventrally and laterally appear to express photoinducible genes, whereas those located ventrally and medially do not.

That the dorsal and ventral SCN compartments might actually represent two dissociable regional

circadian oscillators was initially suggested by studies of rat SCN slices maintained as long-term organotypic cultures. Hypothalamic slices incubated for weeks (either embedded in plasma clots on coverslips in rotating roller tubes or adhered to filters in stationary cultures) lose at least 70% of their neurons and flatten to a few cell layers thick, but they survive, express immunoreactive AVP and VIP regionally as *in vivo*, and exhibit circadian rhythms of the release of both peptides into the medium. When such organotypic cultures are treated with antimitotics, the two peptide rhythms appear to free-run separately with different circadian periods, indicating that they represent a chemical manifestation of comparable, equipotent oscillators in both the dorsal and ventral SCN subdivisions. More recently, regional dorsal and ventral oscillators have been demonstrated in horizontally bisected coronal slices of SCN from transgenic mice carrying the *Per1*-driven luciferase reporter. In general, studies have suggested that gene expression appears to spread through the SCN like a wave, with dorsal cells peaking earlier in the day than ventral ones.

Regional dorsal and ventral oscillators can also be dissociated *in vivo* by perturbing the external lighting conditions. In the rat, an abrupt phase shift of the light–dark cycle results in a transient desynchronization between the two SCN subdivisions; ventrolateral gene expression and electrical activity shift rapidly, while dorsomedial expression and activity resynchronize only gradually over days to weeks. A stable, 'forced' desynchronization of gene expression between the two subdivisions has also been achieved by exposing rats to an artificially short 22 h light–dark (LD) cycle. In both of these situations, intercellular coupling between SCN subdivisions is lost but coupling within each subdivision is retained, suggesting that inter- and intradivisional synchronizing mechanisms may be different.

Notably, not all SCN cells are rhythmic. The calbindin-containing cells in the caudocentral region of the golden hamster SCN have been shown to be directly retinorecipient by tract-tracing and double-label electron microscopy, and c-*fos* and *Per* mRNA and protein levels are increased in calbindin cells after photic stimulation (as assayed on cut brain sections by *in situ* hybridization and immunohistochemistry). However, circadian rhythms of *Per* gene expression are not detectable in this region, and electrophysiological recordings of identified calbindin cells in hamster SCN slices have demonstrated an absence of endogenous rhythmicity. In rats, a similar regional separation of rhythmic and nonrhythmic (but photoinducible) gene expression has been found; and in electrophysiological studies *in vivo*, photically responsive neurons have been found which do not express the circadian rhythm of discharge rate that can be observed in photically insensitive cells.

Thus, SCN tissue is composed of cellular populations that are structurally, functionally, and spatially heterogeneous. The photoresponsiveness and endogenous rhythmicity of whole SCN tissue emerges through the coupling of individual cells, both rhythmic and nonrhythmic, within and across distinct regional compartments.

Coupling Mechanisms for Coordinating SCN Cellular Oscillators

Intercellular coupling is necessary in order to create an integrated tissue pacemaker that produces a coherent read-out to the rest of the organism. Coupling mechanisms enable photoresponsive cells to reset endogenously rhythmic cells and also synchronize the wide range of individual cellular periods to produce the averaged, genotype-specific circadian period characteristic of behavioral rhythmicity of whole animals. Elucidating SCN intercellular interactions will be critical to understanding circadian properties at the tissue level of organization; several attributes of the clock likely emerge at this level, for example, its regulation of complex behavioral patterns, adaptation to photoperiod, and remarkable day-to-day precision.

SCN cells are coupled together in such a way that environmental lighting can reorganize the tissue into multiple regional oscillating units (**Figure 4**). As described above, light–dark cycles that are either phase-shifted or artificially shortened unmask the presence of dual dorsal and ventral circadian oscillations. Perhaps the most dramatic example of functionally reconfigured SCN tissue is the phenomenon known as 'splitting' in golden hamsters maintained in constant light, in which an animal's single daily bout of locomotor activity dissociates into two components that each free-run with different periods until they become stably coupled 180° (about 12 h) apart. Splitting appears to be the consequence of a reorganized SCN with left and right halves oscillating in antiphase, as mRNAs characteristic of day and night are simultaneously expressed on opposite sides of the paired SCN. In rodents, another effect of exposure to constant light is locomotor arrhythmicity. SCN slices have been made from transgenic mice carrying the *Per1*-driven GFP reporter and rendered behaviorally arrhythmic by constant light; in such preparations, individual cells exhibit robust circadian rhythms of fluorescence but desynchronized in phase, resulting in arrhythmicity of the slice as a whole.

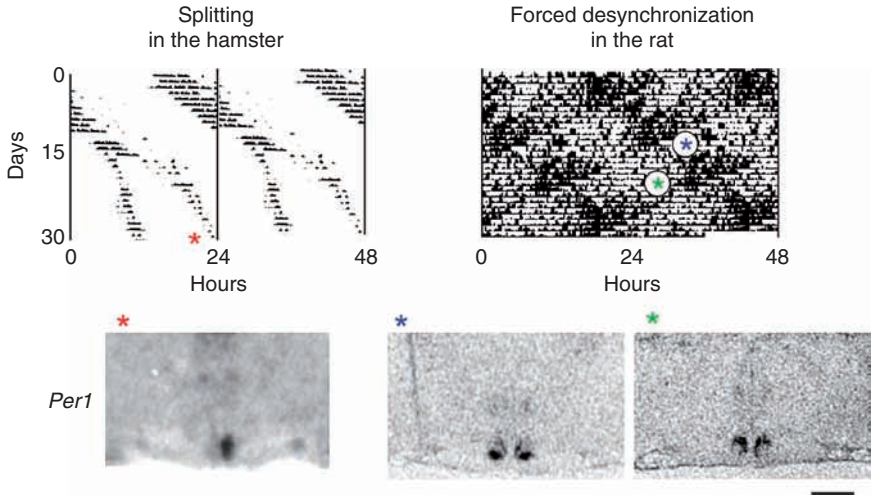

Figure 4 SCN *Per1* gene expression during hamster splitting in constant light (left) and rat forced desynchronization in an 11 h light–11 h dark cycle (right). Locomotor activity is plotted for individual representative animals in double actogram format, with activity over day n followed by day $n+1$ horizontally, succeeded by day $n+1$ and $n+2$ on the next line, and so on. *Per1* mRNA levels are assayed using a specific riboprobe and *in situ* hybridization on coronal brain sections. In splitting, the animal's single daily bout of activity dissociates into two split components that become stably coupled about 12 h apart, at which time *Per1* gene expression is markedly asymmetric; *in situ* hybridization signal is observed unilaterally on either the left or right side of the SCN. In forced desynchronization, two locomotor rhythms with different period lengths are expressed simultaneously, with one entrained to the external 22 h cycle and the other free-running with a period >24 h; at opposite phases for the two rhythms, *Per1* gene expression is restricted to either the ventrolateral or dorsomedial subdivisions. Stars represent times of animal sacrifice, scale bar represents 1 mm. Courtesy of Dr. Horacio O. de la Iglesia, University of Washington. Reprinted with permission from de la Iglesia HO, Meyer J, Carpino A Jr., and Schwartz WJ (2000) Antiphase oscillation of the left and right suprachiasmatic nuclei. *Science* 290: 799–801. Copyright (2000) AAAS. Reprinted with permission from de la Iglesia HO, Cambras T, Schwartz WJ, and Díez-Noguera A (2004) Forced desychronization of dual circadian oscillators within the rat suprachiasmatic nucleus. *Current Biology* 14: 796–800.

Several candidate coupling mechanisms are currently under investigation. Much interest has focused on gamma-aminobutyric acid (GABA). Although still controversial, studies have suggested that GABA's actions on SCN neurons may be inhibitory or excitatory depending on circadian phase. Both effects appear to be mediated by a $GABA_A$ receptor, so the change from inhibition to excitation likely follows a rhythm of the GABA equilibrium potential, perhaps reflecting a circadian oscillation of intracellular chloride concentrations. A $GABA_A$ mechanism may be responsible for coupling oscillations between the ventral and dorsal SCN subdivisions.

A possible role for VIP has been inferred from analysis of 'knockout' mice deficient in the *VIP* or *Vipr2* genes, the latter encoding the $VPAC_2$ receptor. These mice exhibit disrupted behavioral, molecular, and electrophysiological rhythmicity, and long-term SCN neuronal recordings show that many cells are arrhythmic while others are rhythmic but with desynchronized periods. Importantly, these defects appear to be corrected by application of a $VPAC_2$ receptor agonist to $VIP^{-/-}$ cells. How VIP signaling might interact with the GABAergic system (and vice versa) is not known. Gastrin-releasing peptide is another SCN neuropeptide transmitter that possibly functions to synchronize photoresponsive and endogenous rhythmic SCN neurons.

Nonsynaptic mechanisms for intercellular coupling also appear to be involved, including gap junctions, which may couple neurons within subdivisions but not between them. Glial cells are likely to be an important part of any synchronizing mechanism, and intercellular Ca^{2+} waves can travel long distances across cultured SCN astrocytes.

The Search for Morning and Evening Circadian Clocks

A circadian clock mechanism is believed to underlie seasonal time measurement, with daylength (photoperiod) encoded by the circadian rhythm of melatonin production in the pineal gland. The alteration in the duration of the nocturnal melatonin signal, compressed during long summer days and decompressed during short winter days, appears to serve as an endogenous photoperiodic message. Melatonin rhythmicity is controlled by the SCN, and molecular and electrophysiological rhythms of SCN tissue display durations of high activity that expand and contract in concert with changes in the ambient photoperiod.

The photoperiodic clock has been modeled as a complex pacemaker consisting of two mutually coupled circadian oscillators. Variability in the phase relationship between these two oscillators could account for the ability to measure seasonal changes of daylength, with a morning oscillator (M) accelerated by light and synchronized to dawn, and an evening oscillator (E) decelerated by light and synchronized to dusk. Current hypotheses are that M and E might be assembled within SCN tissue from individual cellular oscillators or sets of 'clock' genes that assume different phase relationships as a function of changing photoperiod. Circadian oscillations that behave like M and E have been demonstrated in SCN slices made from the brains of golden hamsters; when the slices are cut in the horizontal (but not in the coronal) plane, the circadian rhythm of multiple-unit neuronal activity exhibits distinct morning and evening peaks. The morning peak follows projected dawn, while the evening peak occurs around projected dusk. The two peaks are differentially affected by changes in the antecedent photoperiod, and they are shifted independently in a phase-dependent manner after stimulation with glutamate. The anatomical basis for these putative M and E oscillators, and whether they might map to sets of dissociable regional SCN oscillators described above, is unknown. However, available data already indicate that left and right 'split' oscillators are not the substrate for M and E.

For the Future: From Intracellular Oscillations to Temporal Programs

While circadian oscillations in the SCN have been identified at multiple levels of biological organization, connecting these levels remains a pre-eminent challenge for researchers. It is not yet understood how cycles of intracellular regulatory molecules lead to rhythms of electrical activity, how individual cellular oscillators are entrained and synchronized within a heterogeneous tissue network, nor how multiple output pathways bring about the entrainment of extra-SCN oscillators elsewhere in the brain and body. The result is a remarkable daily timekeeping system that allows animals to both recognize local time and measure its passage.

See also: Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Peripheral Circadian Oscillators; Seasonal Timing: Neural Mechanisms; Single Cell Neuronal Circadian Clocks.

Further Reading

Antle MC and Silver R (2005) Orchestrating time: Arrangements of the brain circadian clock. *Trends in Neurosciences* 28: 145–151.

Aton SJ and Herzog ED (2005) Come together, right . . . now: Synchronization of rhythms in a mammalian circadian clock. *Neuron* 48: 531–534.

Johnston JD (2005) Measuring seasonal time within the circadian system: Regulation of the suprachiasmatic nuclei by photoperiod. *Journal of Neuroendocrinology* 17: 459–465.

Klein DC, Moore RY, and Reppert SM (eds.) (1991) *Suprachiasmatic Nucleus: The Mind's Clock*. New York: Oxford University Press.

Lee HS, Billings HJ, and Lehman MN (2003) The suprachiasmatic nucleus: A clock of multiple components. *Journal of Biological Rhythms* 18: 435–449.

Michel S and Colwell CS (2001) Cellular communication and coupling within the suprachiasmatic nucleus. *Chronobiology International* 18: 579–600.

Morin LP and Allen CN (2006) The circadian visual system, 2005. *Brain Research Reviews* 51: 1–60.

Schaap J, Albus H, vanderLeest HJ, et al. (2003) Heterogeneity of rhythmic suprachiasmatic nucleus neurons: Implications for circadian waveforms and photoperiodic encoding. *Proceedings of the National Academy of Sciences of the United States of America* 100: 15994–15999.

Silver R and Schwartz WJ (2005) The suprachiasmatic nucleus is a functionally heterogeneous timekeeping organ. In: Young MW (ed.) *Circadian Rhythms, Methods in Enzymology,* vol. 393, pp. 451–465. San Diego: Elsevier Academic Press.

Yamaguchi S, Isejima H, Matsuo T, et al. (2003) Synchronization of cellular clocks in the suprachiasmatic nucleus. *Science* 302: 1408–1412.

Circadian Regulation by the Suprachiasmatic Nucleus

D J Earnest, Texas A&M University Health Science Center, College Station, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Functional Properties of the Mammalian Circadian Clock in the Suprachiasmatic Nucleus

With the prevailing evidence for the endogenous rather than exogenous regulation of circadian or 24 h rhythms in multicellular organisms, identification of the internal biological clock(s) responsible for the generation of these oscillations became the primary focus of many investigations. Through functional studies determining whether the removal of neuroendocrine structures and destruction of nervous tissue alter the expression of overt rhythmicity, the circadian clock in mammals was eventually localized to the hypothalamic suprachiasmatic nucleus (SCN). Ablation of the SCN eliminates circadian rhythms in many behavioral, endocrine, and biochemical processes throughout the body. Transplantation studies have provided unequivocal evidence for SCN function in the central neural control of mammalian circadian rhythms. Experiments using donors and hosts with different circadian genotypes demonstrate that transplantation of fetal SCN tissue restores behavioral rhythmicity in SCN-lesioned hosts and that a specific circadian property, namely period, is conveyed by transplanted pacemaker cells, not the surviving SCN cells in the host, to the reestablished rhythms. SCN-independent clocks play a role, albeit limited, in the generation of circadian rhythms because rhythmic behavioral activity persists in SCN-lesioned rodents treated with methamphetamine or maintained on restricted feeding schedules. Neural substrates for these extra-SCN clocks are unknown, but recent studies implicated the dorsomedial hypothalamic nucleus (DMH) in the circadian regulation of food-anticipatory activity.

In addition to orchestrating circadian rhythms in other brain regions and peripheral systems, the SCN generates self-sustained oscillations in its cellular activities independent of external input. Rhythms in neuropeptide and neurotransmitter levels, cellular metabolism, and neuronal activity are indigenous to the SCN *in vivo* and persist for multiple cycles following the isolation of SCN cells *in vitro*. Circadian oscillations in gene expression are another fundamental property of SCN cells. Rhythmic expression profiles are especially prevalent in the SCN among genes that form the core of the circadian clock mechanism. The molecular clockworks consist of interlocked transcriptional/translational feedback loops in which the expression of core components, *Bmal1*, *Period1* (*Per1*), *Per2*, *Cryptochrome1* (*Cry1*), *Cry2*, and *Rev-erbα*, is rhythmically regulated through mutual interactions between their protein products.

Hierarchy and Cell-Autonomous Nature of Circadian Clocks in Mammals

The molecular dissection of the circadian timekeeping mechanism also brought about the discovery that clock properties are exhibited not only by the SCN but also by other tissues *in vivo* and fibroblast cell lines *in vitro*. The expression and rhythmic regulation of clock genes has been observed in many peripheral cells and tissues, ranging from the liver to endocrine tissues to the heart and skeletal muscles. Clock gene oscillations in peripheral tissues are comparable to those found in the SCN but phase delayed by 4–8 h. Similar to peripheral tissues, Rat-1 and NIH/3T3 fibroblasts exhibit induced oscillations of clock gene expression for several cycles in response to serum shock, forskolin, or corticosteroid treatment. The common molecular organization of the circadian clocks in the SCN, peripheral tissues, and fibroblast cell lines challenged existing dogma on SCN function in the mammalian circadian system by demonstrating that the SCN clock was one among many others throughout the body. These findings also raised many questions about the hierarchical organization of mammalian circadian clocks and functional implications of this canonical clock machinery in the SCN, peripheral tissues, and fibroblast cell lines: Is the SCN a master clock that regulates all or just some of these peripheral clocks? Beyond these molecular oscillations, what other properties are similar in SCN and peripheral clocks? Is clock gene rhythmicity sufficient for peripheral clocks to pace rhythms in downstream cells or tissues, and if not, what distinguishes the SCN from peripheral clocks? What SCN output signals are important in coordinating circadian oscillations in peripheral tissues?

Within the mammalian circadian hierarchy, the SCN appears to function as a master clock that coordinates rhythmicity in other brain regions and many, but not all, peripheral tissues (Figure 1). Consistent with this hierarchical organization, circadian oscillations of gene expression *in vitro* are more robust within the SCN than the peripheral tissues. Circadian rhythms of luciferase-reported *Per1* expression

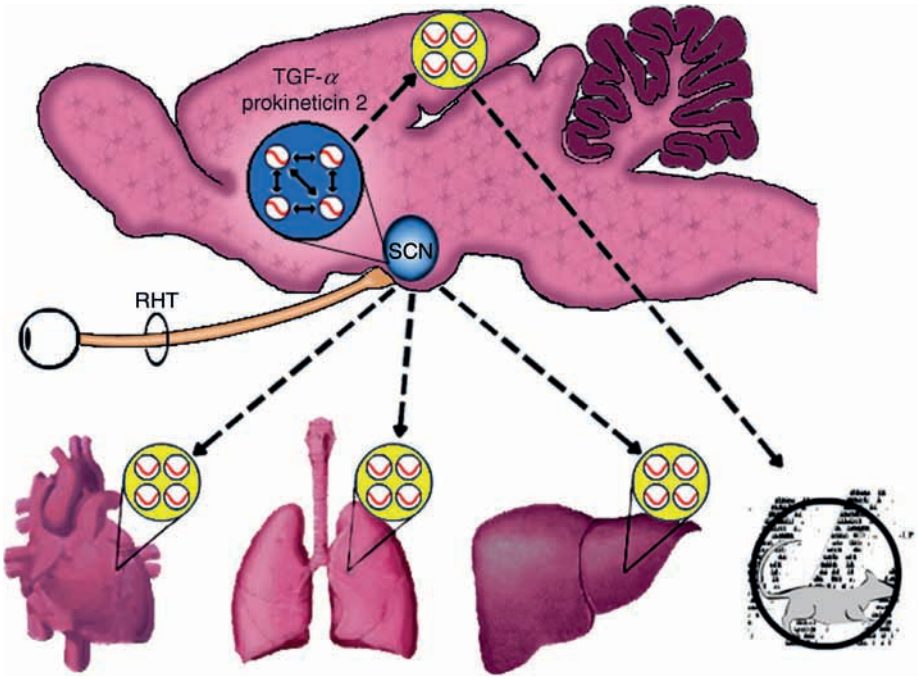

Figure 1 Hierarchy of cell-autonomous clocks in the mammalian circadian system. Entraining light–dark cues are communicated from the retina directly to the master circadian clock located in the suprachiasmatic nucleus (SCN) via the retinohypothalamic tract (RHT). The SCN clock consists of numerous cell-autonomous clocks that, in turn, synchronize the circadian oscillations of clock cells in other brain regions and peripheral tissues such as the heart, lung, and liver. Signals responsible for the local coordination of circadian oscillations among individual clock cells may play a role in the distinctive functional properties of the SCN. SCN control of circadian clocks in other tissues and of behavioral rhythmicity is mediated by diffusible or humoral signals. Many of the signals necessary for SCN pacemaker function are unknown, but transforming growth factor (TGF)-α and prokineticin 2 (PK2) are SCN output signals involved in the circadian regulation of locomotor activity.

persist for many weeks in explant cultures of the SCN but damp after several cycles in cultured lung, liver, kidney, and skeletal muscle tissue. Further evidence for the hierarchical relationship between the SCN and peripheral clocks is based on *in vitro* studies using *Per2-Luc* knockin mice to demonstrate that lesions of the SCN disrupt the normal phase coordination of clock gene rhythmicity among explant cultures of different peripheral tissues. However, some peripheral tissues, such as the liver, are not slaves to the SCN clock but instead can function independently in response to nonphotic cues. Restricted feeding cycles entrain the liver rhythm in *Per1-Luc* expression without influencing the circadian phase of this oscillation in the SCN. The prevalence of this common timekeeping program throughout the body and its functionality even in the absence of the SCN indicate that peripheral clocks provide for the local organization of tissue- or cell-specific processes in time.

Similarities between the SCN and peripheral tissues extend beyond core molecular components to even the basic cellular organization of the circadian clockworks. In the SCN, circadian oscillations are not only an ensemble property of the entire nucleus but are also expressed by individual neurons. Like the composite SCN *in vivo* and in brain-slice preparations, many individual SCN neurons in dissociated cultures maintained on microelectrode arrays express independent oscillations in firing rate. The circadian phase and period of these SCN rhythms in firing rate differ widely among individual neurons despite the presence of synaptic connections. SCN cells in brain-slice preparations also exhibit ensemble rhythmicity in *Per-1*-driven green fluorescent protein (GFP) expression, and the waveform of this rhythm is derived from the multiphasic organization of autonomous oscillations in individual neurons. Recent findings suggest that peripheral tissues may contain cell-autonomous clocks similar to those found in the SCN. Individual NIH/3T3 fibroblasts exhibit circadian rhythms of gene expression that are largely synchronous following serum shock but are marked by broad phase differences in the absence of this treatment.

Functional Differences between Circadian Clocks in the Suprachiasmatic Nucleus, Peripheral Tissues, and Fibroblasts

Despite their apparent parallels at the molecular and cellular levels of organization, the SCN and peripheral

clocks are fundamentally different. The foremost distinction is that the SCN contains the only known clock cells that can be directly entrained by light. Photoentrainment signals from the retina are processed through the retinohypothalamic tract (RHT), which terminates in the SCN (Figure 1). Evidence for the essential role of this retinofugal pathway in circadian photoentrainment is based on the finding that rodent activity rhythms persist but fail to entrain to light–dark cycles following localized transection of the RHT. In contrast, no data are currently available to unequivocally demonstrate that peripheral tissues in mammals also receive direct photic input. In the circadian hierarchy, the SCN clock thus provides a vital link between the external environment and the internal circadian timekeeping system. Further evidence for the preeminent role of the SCN as the gateway for the coordination of local processes with the light–dark environment is provided by the observation that clock gene rhythms are more rapidly entrained by light in the SCN than in peripheral tissues. Moreover, the 4–12 h phase delay of peripheral tissue oscillations relative to SCN rhythms is presumably indicative of the time required for entrainment signals to reach the periphery.

Distinctions between the SCN and peripheral clocks are especially evident in their functional properties. SCN cells function as pacemakers and confer rhythmicity to SCN-lesioned rodents *in vivo* and to other cells *in vitro*, whereas peripheral tissues and cultured fibroblasts do not. Incontrovertible evidence for the unique pacemaker properties of the SCN is founded on transplantation studies. Fetal SCN tissue and immortalized rat SCN cells (SCN2.2), but not other types of tissue (e.g., cerebellum or cerebral cortex) or fibroblasts, reinstate circadian behavioral rhythmicity when transplanted into arrhythmic SCN-lesioned hosts. Similar to transplantation studies, the development of a coculture model has been instrumental in distinguishing the circadian pacemaker properties of SCN cells from the oscillatory behavior of fibroblasts. In this model, SCN2.2 cells are cocultured on companion plates with NIH/3T3 fibroblasts on cell-impermeable inserts. Consistent with SCN function *in vivo*, SCN2.2 cells endogenously generate circadian oscillations of glucose utilization and clock gene expression and confer this metabolic and molecular rhythmicity to cocultured NIH/3T3 fibroblasts (Figure 2). Serum-shock treatment produces rhythms of clock gene expression, but not glucose metabolism, in these fibroblasts. Despite the induction of clock gene oscillations, serum-shocked NIH/3T3 fibroblasts lack circadian pacemaker properties and fail to confer even this molecular rhythmicity to cocultured cells.

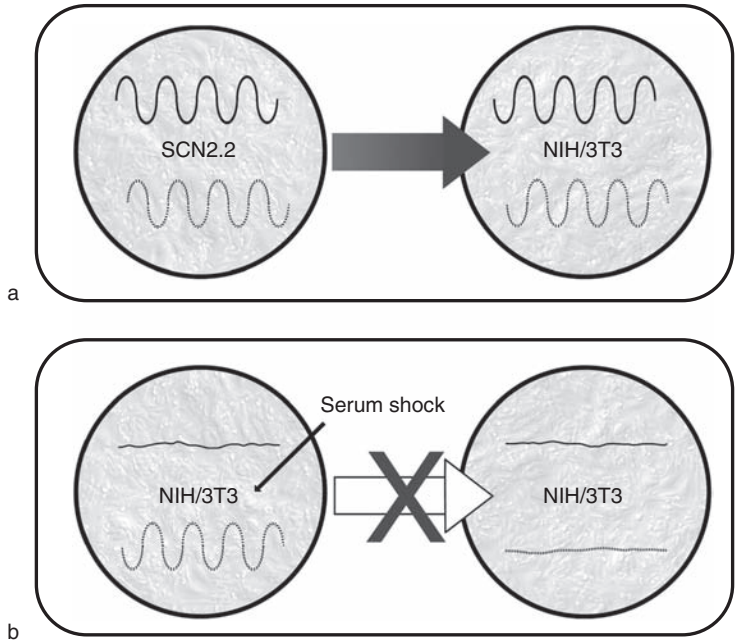

Figure 2 Diagram illustrating the distinctive rhythm-generating and pacemaker properties of SCN2.2 cells in coculture studies: (a) SCN2.2–NIH/3T3 cocultures; (b) serum-shocked (SS) NIH/3T3–untreated NIH/3T3 cocultures. For each cell type the cocultures, the upper (solid) waveform represents the temporal pattern of metabolic activity (2-deoxyglucose uptake) and the lower (dashed) waveform denotes the profiles of clock gene expression (*Per1* and *Per2*). The thick arrows between culture types indicate the presence or absence of signals that coordinate oscillations in cocultured cells. SCN2.2 cells, but not serum-shocked NIH/3T3 fibroblasts, are capable of pacing the metabolic and clock gene rhythms in cocultures of untreated NIH/3T3 cells.

Why peripheral tissues and fibroblasts cannot function as circadian pacemakers is presently unclear. One possible explanation is that all molecular elements of the clock mechanism may not be present or functionally configured. Experiments examining the effects of antisense inhibition of *Clock* reveal important distinctions between SCN2.2 cells and serum-shocked fibroblasts in the regulation of molecular oscillations. Treatment of SCN2.2 cells with *Clock* antisense oligonucleotides disrupted the circadian rhythm of *Per2* expression within these cells and cocultures of untreated NIH/3T3 fibroblasts. However, the circadian oscillations of *Per2, Bmal1*, and *Cry1* expression in serum-shocked fibroblasts showed some alterations in rhythm amplitude but otherwise persisted during antisense inhibition of *Clock*. These differences in *Clock* interactions within the molecular feedback loop could account for distinctions in the circadian pacemaker properties of SCN cells and fibroblasts.

The Search for Critical Output Signals of the Suprachiasmatic Nucleus

An alternative explanation for the distinctive pacemaker function of the SCN is that specific output signals necessary for coordinating rhythmicity among individual clock cells within the SCN and in the periphery may not be expressed or circadian-regulated in other tissues or stimulated fibroblasts. SCN pacemaker regulation of circadian rhythms in downstream cells and tissues is thought to involve diffusible or humoral factors rather than neurotransmission via synaptic connections. Functional implications for diffusible factors in the SCN pacemaker regulation of circadian rhythms were first provided by the observation that transplantation of fetal hypothalamic tissue into the third, or even lateral, ventricle of SCN-lesioned rodents restores circadian rhythms with few, if any, neural connections between the graft and host brain. However, the role of humoral outputs in the circadian pacemaker function of the SCN is directly supported by transplantation experiments determining whether the recovery of circadian function occurs in SCN-lesioned animals when grafted cells are encapsulated so as to prevent the establishment of neural connections with the host brain. Despite their isolation in polymeric capsules that allow graft–host communication only via diffusion of humoral signals, transplants of SCN cells still promote the recovery of circadian behavioral rhythmicity with the donor period in SCN-lesioned hamsters. Coculture studies similarly demonstrate that the rhythmic release of some paracrine signal(s) by the SCN is responsible for its function in pacing circadian oscillations in downstream cells. SCN2.2 cells impose rhythms of metabolic activity and *Per* gene expression on cocultured NIH/3T3 fibroblasts, and these conferred oscillations are mediated by diffusible signals because the two cell types are separated in the coculture environment by a semipermeable membrane preventing physical contact (Figure 2). Furthermore, sustained rhythmicity in NIH/3T3 cells is dependent on continued exposure to this SCN2.2-communicated signal. Following their coculture and subsequent isolation from one another, SCN2.2 cells continue to express circadian rhythms of glucose utilization, whereas NIH/3T3 fibroblasts fail to exhibit self-sustained metabolic oscillations in the absence of SCN2.2-specific outputs. Therefore, identification of specific diffusible or humoral factors that coordinate rhythmicity among individual clock cells in the SCN and among circadian clocks in peripheral tissues is critical in determining which cellular processes are responsible for the distinctive function of the SCN as a circadian pacemaker.

Glucocorticoids were among the first signaling candidates implicated in the coordination of circadian oscillations within peripheral clocks. The SCN regulates the cyclical secretion of glucocorticoid hormones, and this signaling pathway is thought to mediate SCN control of peripheral clocks *in vivo* because adrenalectomy alters the light-induced phase resetting of circadian gene expression in the periphery. Glucocorticoid receptor agonists also induce clock gene oscillations in cultured fibroblasts. Transforming growth factor α (TGF-α) and prokineticin 2 (PK2) were subsequently identified as SCN output signals that play a role in its pacemaker function, at least with regard to the circadian regulation of locomotor activity *in vivo* (Figure 2). TGF-α is rhythmically expressed in the SCN, with peak mRNA levels occurring during the mid-subjective day, and chronic infusion of TGF-α into the third ventricle completely inhibits wheel-running activity and disrupts the circadian organization of the sleep–wake cycle. These effects of TGF-α are mediated through its primary receptor, the epidermal growth factor receptor (EGFR) because mice with a partial loss-of-function mutation for EGFR exhibit increased daytime activity and decreased inhibition of wheel-running activity in response to light during the subjective night. In a similar fashion, PK2 mRNA levels rhythmically fluctuate in the SCN with peak expression during the subjective day, and intraventricular injections of recombinant PK2 during the subjective night, when endogenous levels are low, inhibit nocturnal increases in rat locomotor activity. Significantly, the receptors for TGF-α and PK2 (EGFR and PKR2) are distributed in primary targets of SCN efferent pathways, including the subparaventricular zone and paraventricular nucleus. This observation suggests that TGF-α and PK2 are not involved in

the synchronization of individual cell-autonomous clocks within the SCN but, instead, are critical SCN outputs necessary for pacemaker coordination of downstream clocks in other brain regions controlling locomotor behavior. The neuropeptide vasoactive intestinal polypeptide (VIP) is another output signal that mediates SCN pacemaker function in the circadian regulation of locomotor behavior. Mice with a null mutation of the VIP receptor, $VPAC_2$, show weak or low-amplitude rhythms of wheel-running behavior in conjunction with the disruption of circadian oscillations in SCN expression of core clock genes. $VPAC_2$ receptors are more prevalent within the SCN than in primary targets of its efferent pathways. Coupled with evidence for the essential role of $VPAC_2$ receptor signaling in the generation of the circadian rhythm in SCN neuronal activity, the local distribution of these receptors suggests that VIP may be involved in regulating cell-to-cell communication in the SCN. Thus, $VPAC_2$ receptor signaling may coordinate rhythmicity among populations of cell-autonomous clocks so as to provide for the generation of high-amplitude circadian outputs necessary for SCN pacemaker function.

It is unlikely that glucocorticoid hormones, TGF-α, PK2, and VIP are solely responsible for SCN pacemaker function in coordinating the entire spectrum of circadian processes in downstream cells and tissues. Global analysis of gene expression profiles has provided a valuable screening tool for other output signals that may distinguish the SCN from peripheral and fibroblast clocks. Comparisons of microarray and bioinformatics data indicate that the overall extent of circadian gene expression in the SCN and peripheral tissues *in vivo*, as well as cultured SCN2.2 cells and fibroblasts, was similarly low, ranging from approximately 2% to 10%. However, specific genes with circadian profiles are highly disparate among different tissue and cell types. The canonical clock genes constitute the principal portion of the overlap in circadian gene expression between SCN cells and peripheral tissues or fibroblasts. Other rhythmic genes in common to different mammalian tissues predominantly function as nuclear factors involved in the regulation of DNA-binding and activation of transcription. This intriguing observation suggests that the cyclical regulation of nuclear factors may have important implications in linking the common clockworks to tissue-specific cellular and physiological pathways so as to provide local control over circadian timing. In addition, functional analyses indicate that various mammalian cells and tissues are similar with regard to the diversity of rhythmically expressed transcripts. Within the SCN *in vivo*, SCN2.2 cells, fibroblasts, and peripheral tissues such as the liver, heart, and kidney, transcripts with cyclical profiles control a wide range of biological processes. Yet circadian gene expression is typically more prevalent within certain functional categories, and there is some degree of tissue and cell specificity in this differential distribution of rhythmic genes. In SCN2.2 cells and the SCN *in vivo*, functionally annotated genes with circadian expression profiles are predominantly involved in the regulation of cellular communication, protein dynamics, and metabolic pathways mediating fatty acid recycling, steroid synthesis, glucose metabolism, and mitochondrial energy transduction. The prevalence of rhythmicity among genes mediating intercellular communication may have some significance with regard to SCN-specific signals necessary for the endogenous generation of self-sustained oscillations, local synchronization of clock cell populations, or pacemaker coordination of circadian oscillations in other tissues and cells.

Among rhythmic genes in intercellular communication pathways, inducible nitric oxide synthase (*iNos* or *Nos2*), an isozyme involved in nitric oxide (NO) production, is noteworthy as a circadian signal that is unique to the SCN. The role of *Nos2* in the SCN and the regulation of circadian rhythms has not been examined, presumably because this isozyme is thought to be expressed by macrophages, not SCN cells, and because functional studies suggest that the other isoforms of nitric oxide synthase in the central nervous system are not necessary for the generation or entrainment of circadian rhythms. The activity rhythms of mutant mice lacking neuronal *Nos* (*nNos* or *Nos3*) and endothelial *Nos* (*eNos* or *Nos1*) show no distinct abnormalities with regard to their free-running patterns, entrainment to light–dark cycles, and phase-shifting responses to light. Initial evidence for *Nos2* involvement in the endogenous oscillatory and pacemaker properties of the SCN is derived from the recent observation that antisense inhibition of NOS2 expression abolishes the metabolic rhythmicity of both cell types in cocultures containing antisense-treated SCN2.2 cells and untreated NIH/3T3 fibroblasts. Moreover, *Nos2* function in SCN regulation of circadian rhythms is compatible with the effects of a general NOS inhibitor on rhythmic glucose uptake in SCN2.2–NIH/3T3 cocultures. Similar to *Nos2* antisense treatment, N^G-nitro-L-arginine methyl ester (L-NAME), a reversible inhibitor of all three isoforms of NOS, disrupts the circadian pattern of glucose utilization in SCN2.2 cells. Following treatment, glucose uptake fluctuated rhythmically in both L-NAME-treated SCN2.2 cells and untreated NIH/3T3 fibroblasts, but the peak-to-peak interval for this rhythmicity was increased from 24 to 32 h in both cell types. It is unclear how L-NAME and *Nos2* antisense treatment specifically produce this loss of

rhythmicity in SCN2.2 cultures. One possibility is that the treatment-induced inhibition of NOS2 may disrupt the clock mechanism by altering the biological effects of NO signaling through the intracellular second-messenger cyclic guanosine monophosphate (cGMP) and/or on posttranslational modification of cellular proteins. Another potential explanation for the abolition of metabolic rhythmicity in SCN2.2 cells is that L-NAME and *Nos2* antisense treatment may inhibit NO function as a diffusible messenger in the intercellular coupling or local synchronization of clock cell populations. As coupling strength among a multitude of autonomous clocks decreases from the population mean, changes in rhythm amplitude or period length are expected to occur and eventually result in the loss of rhythmicity from the ensemble. Thus, the disruption of metabolic rhythms in SCN2.2 cultures may reflect treatment-mediated loss of temporal synchrony among populations of clock cells. By affecting the generation of high-amplitude circadian output signals, this desynchrony between clock cells could also account for the effects of *Nos2* antisense treatment on the pacemaker function of SCN2.2 cells in regulating oscillations in cocultured fibroblasts. Together with evidence for the role of NO in resetting the SCN clock, these findings suggest that *Nos2* function in the production of this diffusible messenger may be essential for the endogenous rhythm-generating and pacemaker properties of the SCN.

Future Developments

The localization of cell-autonomous circadian clocks in the SCN, other brain regions, and peripheral tissues throughout the body has yielded a new understanding of the mammalian circadian hierarchy. It is not yet clear whether the canonical clock machinery in peripheral tissues is directly linked to the circadian regulation of any physiological processes. However, this discovery has kindled the search for SCN output signals responsible for the coordination of circadian oscillations in peripheral clocks. Although some candidates for SCN pacemaking signals have been identified, determining their role in the local coupling between individual clock cells and the coordination of circadian oscillations in peripheral clocks remains a major challenge.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Photoreceptors and Circadian Clocks; Serotonin and the Regulation of Mammalian Circadian Rhythms.

Further Reading

Aton SJ and Herzog ED (2005) Come together, right . . . now: Synchronization of rhythms in a mammalian circadian clock. *Neuron* 48: 531–534.

Balsalobre A (2002) Clock genes in mammalian peripheral tissues. *Cell & Tissue Research* 309: 193–199.

Cermakian N and Sassone-Corsi P (2000) Multilevel regulation of the circadian clock. *Nature Reviews Molecular Cell Biology* 1: 59–67.

Herzog ED and Schwartz WJ (2002) A neural clockwork for encoding circadian time. *Journal of Applied Physiology* 92: 401–408.

Kriegsfeld LJ and Silver R (2006) The regulation of neuroendocrine function: Timing is everything. *Hormones & Behavior* 49: 557–574.

Okamura H (2004) Clock genes in cell clocks: Roles, actions, and mysteries. *Journal of Biological Rhythms* 19: 388–399.

Panda S and Hogenesch JB (2004) It's all in the timing: Many clocks, many outputs. *Journal of Biological Rhythms* 19: 374–387.

Reppert SM and Weaver DR (2002) Coordination of circadian timing in mammals. *Nature* 418: 935–941.

Schibler U (2006) Circadian time keeping: The daily ups and downs of genes, cells, and organisms. *Progress in Brain Research* 153: 271–282.

Circadian Regulation in Invertebrates

T L Page, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the invertebrates, circadian clocks control a variety of behavioral, physiological, and developmental processes; and the study of the anatomical and physiological organization of circadian systems of invertebrates has a long and productive history. By far, most of the work has been carried out on arthropods (especially insects), though there have been seminal contributions from the opisthobranch mollusks. Much less is known about other invertebrates, though it is clear that most major groups exhibit circadian rhythms.

Although there are differences in detail, the studies suggest that the general outlines of circadian organization in the invertebrates parallel that in the vertebrates. In invertebrates the circadian system is composed of a population of oscillators that are located both within and outside of the nervous system. Appropriate phase relationships between the circadian system and the external environment are maintained by entrainment to environmental cycles. The most important cycle in the physical environment is the daily light cycle and the transduction of light is (with a few interesting exceptions) carried out largely by extraretinal, nonvisual photoreceptors with a contribution from photoreceptors in the retina. Internal temporal organization (phase relationships among the various oscillators and the rhythms they control) is maintained either by coupling between oscillating components or simply by the subjugation of individual oscillators to direct entrainment by environmental stimuli.

Localization of Circadian Oscillators

Studies on pacemaker localization in invertebrates initially focused on the nervous system. In mollusks, oscillators have been found to exist in both the eye and the central nervous system. The ocular oscillators have been particularly important as model systems for the study of the cellular mechanisms of circadian rhythms. In the arthropods, the brain has consistently been found to be important for generation of circadian timing signals for rhythms in behavior. However, it is clear that circadian oscillations that control physiologically relevant processes frequently arise in structures outside the central nervous system as well.

Pacemakers in the Molluscan Retina

The eyes of several marine opisthobranch mollusks have been shown to exhibit circadian rhythms of spontaneous nerve impulse activity when removed from the animal and maintained in artificial seawater. The rhythm in the isolated eye was first discovered in *Aplysia californica*. Subsequently, similar rhythms were documented in other species of mollusk, including *Navanax inermis*, *Bulla gouldiana*, and *Bursatella leachi plei*. The rhythms are expressed in the frequency of optic nerve large-amplitude, compound action potentials that arise from a population of electrically coupled cells in the eye. The impulses are spontaneously generated and occur even when the eye is maintained in complete darkness. In all species, peak impulse activity occurs during the animal's subjective day, and is likely due to the circadian modulation of potassium conductance. Efforts to further localize the pacemaking cells in the eye through localized lesions have been made in *Bulla* and *Aplysia*, and in both species the results indicate that the rhythm is generated by cells at the base of the eye. This proposition was confirmed in experiments showing periodic changes in membrane conductance in single, isolated cells.

There has also been an effort to assess the functional role of the ocular pacemaker in controlling behavioral rhythms in the mollusks. In both *Aplysia* and *Bulla*, removal of the eyes usually, but not invariably, resulted in disruption of the circadian rhythm in locomotor activity, suggesting these structures may participate as the driving oscillator for the rhythm. These results, combined with the persistence of locomotor rhythmicity in some eyeless animals, indicate that an additional oscillatory structure outside the eyes may play an important role in the temporal organization of behavior. In this context it is notable that in *Aplysia*, the animal's ability to form long-term memories exhibits a circadian rhythm that is not affected by eye removal, indicating that, for this rhythm, both the clock and the photoreceptor for its entrainment are located outside the eyes.

The effects of eye removal on locomotor activity have also been investigated in two pulmonate gastropod species, the terrestrial slug *Limax maximus* and the freshwater snail *Helisoma trivolvis*. Unlike the opisthobranch mollusks, there is no evidence in either of these species that the eye contains a pacemaker, and eye removal appeared to have no significant effect on the ability of animals to generate a circadian rhythm in activity. It seems likely that the pacemaker

for the activity rhythms in these mollusks is located in the brain, though that has yet to be demonstrated.

Pacemakers in the Arthropod Brain

Arthropods exhibit circadian rhythmicity in a variety of behaviors and physiological processes. Three of these – locomotor activity, eclosion, and retinal sensitivity to light – have been investigated extensively and there are several instances in which circadian pacemakers that regulate these rhythms have been localized to the brain (supraesophageal ganglion). In a classic early study on the moths *Hyalophora cecropia* and *Antheraea pernyi,* it was shown that a circadian pacemaker that controls the timing of eclosion (a behavioral sequence that allows the adult moth to emerge from its pupal case after metamorphosis) is located in the brain. The time of day when eclosion occurs is different for the two species. *Hyalophora cecropia* emerges shortly after dawn while *A. pernyi* emerges just before dusk. Removal of the brain did not prevent eclosion, but did disrupt its timing. However, if the brain was reimplanted in the abdomen, normal rhythmicity was restored in both entrained and free-running conditions. The most compelling evidence was provided by experiments in which brains removed from one species were transplanted to the abdomen of the other. Moths that had received these 'switched brain' transplants exhibited normal species-specific eclosion behavior, but the phase of the rhythm was characteristic of the donor and not the host. *Antheraea pernyi* which had received *H. cecropia* brains emerged just after lights-on, while *H. cecropia* that had received *A. pernyi* brains emerged near lights-off. The demonstration that the transplanted brains restored rhythmicity and determined the phase of the rhythm leaves little doubt that circadian pacemaker that regulates the timing of the eclosion rhythm is located in the brains of these moths.

Early studies on cockroaches focused attention on the optic neuropils of the insect brain. The three optic neuropils (from distal to proximal, these are the lamina, medulla, and lobula) form the optic lobe, which is morphologically quite distinct in the cockroach. It was shown that surgical ablation of the cockroach optic lobes or their neural isolation from the midbrain by section of the optic tracts disrupted the locomotor activity rhythm. Although severing the optic tracts invariably disrupted rhythmicity, the rhythm consistently reappeared after several weeks as a consequence of regeneration of the neural connections between the optic lobes and the midbrain. These results led to experiments in which the optic lobes were transplanted between individuals with different free-running periods. Animals that received transplanted optic lobes recovered rhythmicity, and the preoperative period of the donor and the postoperative period of the host were strongly correlated. Thus transplantation of the optic lobes from one animal to another, from which the optic lobes had been removed, both restored the rhythm of locomotor activity and, critically, imposed the period of the donor animal's rhythm on the rhythm of the host. Other experiments have also shown that the optic lobes control the phase of the free-running locomotor activity rhythm and that the tissue could generate a free-running circadian oscillation in multiunit impulse activity when isolated *in vitro*. Studies on crickets and beetles have also indicated pacemakers that are located in the optic lobes. In cockroaches and crickets, the region within the optic lobe that contains the circadian pacemaker for the locomotor activity rhythm was more precisely localized with the use of small lesions. The results suggested that the cells responsible for generating the circadian signal have their somata and/or processes in the ventral half of the optic lobe, between the medulla and the lobula.

The cellular identification of pacemakers in the insect brain came rapidly on the heels of two seminal discoveries. The first was the identification of the sequence of the 'clock gene' *period* and its protein product (Per) in fruit flies. The identification of the gene made it possible to utilize both *in situ* hybridization and immunocytochemistry to identify putative cellular oscillators and to carry out genetic manipulations to evaluate the functional roles of the cells. The second discovery was the finding that antibodies to a crustacean peptide hormone, pigment-dispersing hormone (PDH), specifically stained cockroach optic lobe cells that had a location and projections that made them ideal candidates as the cockroach circadian pacemaker.

In the cockroach, a population of approximately 16 cells in a region that is located at the ventral margin of the medulla was found to be immunopositive for PDH. Subsequently, both lesion and transplantation experiments provided strong evidence that these cells are both necessary and sufficient for driving the circadian rhythm of locomotor activity, suggesting that these cells generate the pacemaking oscillation and that pigment-dispersing factor (PDF; the insect version of crustacean PDH) may be a significant circadian signaling molecule.

In the fruit fly, *Drosophila melanogaster,* several lines of evidence have implicated a small population of neurons, called the lateral neurons, in the regulation of locomotor activity and eclosion rhythms. These neurons are located near the medulla neuropil and express *per* as well as the other clock genes *timeless* and *doubletime*. Genetic manipulations

have shown that these neurons and their rhythmic expression of *per* are essential to rhythmicity. In addition, it is notable that a subset of the lateral neurons produces PDF, and mutations in the *pdf* gene or manipulations of its expression cause major disruptions of the circadian rhythms of locomotor activity, reinforcing the suggestion from the cockroach data that PDF may be a critical signaling molecule in the insect circadian system.

The localization of circadian pacemakers that regulate behavioral rhythms raised the question of whether other rhythms are controlled by the same clocks. A common feature of arthropod visual systems is that the sensitivity of the eye to light exhibits a circadian variation which is reflected in a daily change in the amplitude of the electroretinogram (ERG) recorded in response to a standard light pulse. Depending on the species, several mechanisms may contribute to producing the sensitivity change; however, it does appear, invariably, to be associated with structural changes, including daily migration of screening pigments which surround the rhabdomeres of the ommatidia. In the cricket *Gryllus bimaculatus*, the beetle *Anthia sexguttata*, and the cockroach *Leucophaea maderae*, the pacemaker regulating the ERG amplitude rhythm has been localized to the optic lobe and appears to be the same pacemaker that controls locomotor activity rhythms.

Efforts at localization of circadian pacemakers that regulate retinal sensitivity have also been made in two chelicerate arthropods, the horseshoe crab *Limulus polyphemus* and the scorpion *Androctonus australis*. In both cases the pacemaker that drives the circadian rhythm in ERG amplitude is located in the brain and communicates with the eye via efferent neural projections to the retina (discussed in more detail later).

Lesion studies have also been carried out on decapod crustaceans in an effort to localize pacemakers that regulate both locomotor activity and ERG amplitude rhythms. Much of the work has been on freshwater crayfish, and the results suggest that normal rhythmicity requires both the optic lobes and the cerebral lobes (supraesophageal ganglion), since lesions to either of these structures disrupt, but do not completely abolish, rhythmicity. ERG rhythms persist *in vitro* in preparations that contain only the photoreceptors and the first optic neuropil (lamina ganglionaris), showing that these structures must have the ability to generate a circadian oscillation.

Circadian Pacemakers outside the Nervous System

Although the focus of efforts to find circadian oscillators has been on pacemakers in the nervous system that regulate behavior, there are well-documented cases, all in insects, of circadian pacemakers that reside outside the brain and retina. In the cockroach, studies on circadian deposition of cuticle indicated that this rhythm is driven by a pacemaking system that is independent of the brain. In both *L. maderae* and *Blaberus craniifer*, the new cuticle of the newly molted adult is secreted rhythmically in layers. The rhythm persists in constant conditions with a circadian period that is temperature compensated. Further, in *Blaberus*, the rhythm in cuticle deposition persists in pieces of leg tissue cultured *in vitro*. The results raise the possibility that the epidermal cells that secrete the cuticular material may be autonomously rhythmic.

Several studies have shown that the complex of testis and seminal ducts in moths contains a circadian pacemaker. In the flour moth *Anagasta kueniella*, the release of the sperm from the testis exhibits a daily rhythm that persists in constant darkness in the isolated abdomen. A similar rhythm of release of mature sperm from the testis into the seminal ducts was found in the gypsy moth, *Lymantria dispar*. In this species the rhythm persisted, and could be phase shifted by light, *in vitro* in isolated complexes of seminal ducts and testis. Thus this tissue contains a circadian pacemaker and photoreceptor for entrainment.

In the moth *Samia cynthia*, preparation for pupation and metamorphosis involves a series of morphological and behavioral changes, including an excretion of the contents of the gut which occurs shortly after a cessation in feeding. This 'gut purge' is regulated by the circadian system. Timing of the gut purge appears to be controlled by a circadian clock in the prothoracic gland (PG), an endocrine organ which is responsible for the secretion of the insect steroid hormone ecdysone. A small peak of ecdysteroid occurs in the hemolymph just prior to gut purge. The circadian timing of the ecdysteroid peak persists in decapitated larvae, showing that a brain-centered clock is not necessary. Other studies involving the transplantation of the PGs followed by their local illumination indicated that the glands contain a photosensitive circadian clock that can regulate the timed release of ecdysone. Additional work in the bug *Rhodnius prolixus* and the moth *Galleria mellonella* also provided evidence that insect prothoracic glands contain a photosensitive pacemaker that regulates synthesis of ecdysteroids.

The fact that circadian pacemaking activity is widespread among tissues of the body was particularly clearly demonstrated in experiments on *D. melanogaster* in which the coding sequence of a reporter gene was coupled to the promoter of the clock gene, *period*, allowing tissues in culture to be assessed for the presence of circadian oscillations in transcription. Molecular oscillations were found to be widespread

and to exist even in peripheral cells and tissues where there had never been a hint of endogenous rhythmicity, such as in wings, antennae, proboscises, and legs. The molecular rhythm in the antennae was subsequently shown to drive a circadian rhythm of response in the olfactory receptors of the antennae.

Multioscillator Organization of the Pacemaking System

The data showing that the circadian system in the individual is composed of several oscillators located both within and outside the nervous system raise the question of how temporal order is maintained among the population of oscillators and the circadian rhythms they control. One potential mechanism to ensure appropriate phasing is to have the oscillators directly communicate with one another via internal 'coupling pathways' that allow for hierarchical or mutual entrainment. A second strategy is to give all oscillators direct access to environmental cyclic signals that could entrain the individual pacemakers. The available evidence suggests that both mechanisms are utilized.

The bulk of the information available is on bilaterally paired oscillators, for which the data indicate significant differences among species. In the cockroach, the optic lobe pacemakers are strongly coupled to one another via direct neural connections between the optic lobes, connections probably made by PDH immunopositive cells. In contrast, in the beetle, *Blaps gigas*, the ERG amplitude rhythms from the two eyes of an individual beetle were found to free-run with different periods in constant darkness; and it was shown that the phase and period of the rhythm from one eye could be controlled by local illumination of that eye without any consistent effect on the rhythm from the other eye. The results indicated that each eye was controlled by its own pacemaker, and that the two pacemakers were independent or only weakly coupled. A similar conclusion came from studies on the regulation of locomotor activity by the optic lobe pacemakers in two species of crickets, *Teleogryllus commodus* and *G. bimaculatus*. Thus it appears that within the insect families there are significant species differences in the strength of coupling between bilaterally distributed oscillators, and some appear to rely largely on entrainment by the environmental light cycle to maintain an appropriate phase relationship. Similar species differences have been observed in mollusks. In *Aplysia* maintained in constant darkness, the two ocular pacemakers spontaneously desynchronize and appear to be independent of one another. In contrast, in *Bulla*, the eyes remain closely in phase, suggesting that the pacemakers are coupled. The apparent difference in coupling between the two species was confirmed in studies on the eyes and brain isolated *in vitro*. Following treatments to one eye that induced a phase difference of several hours between the ocular pacemakers, in *Bulla* the eyes tended to come back into phase at a rate of about 1 to 1.5 h day^{-1}, while in *Aplysia* phase separations were maintained. Thus in one case coupling between oscillators is adequate to maintain phase relationships while in the other entrainment by environmental light cycles is crucial.

There is little information available on the coupling relationships between central and peripheral oscillators or among peripheral oscillators. In *D. melanogaster*, data suggest that the central brain pacemaker that regulates circadian rhythms in locomotor activity is neither sufficient nor necessary for a peripheral rhythm in olfactory sensitivity in the antennae. This rhythm is driven by oscillations that are generated in the antennae. In contrast, recent results suggest there is an interaction between pacemakers in the lateral neurons of the brain and in the prothoracic glands that together regulate the circadian rhythm in eclosion.

In cases in which coupling appears to be absent, temporal order among oscillators is likely to be maintained by entrainment to the external environment. In *Drosophila*, for example, peripheral oscillators appear to have access to information about the environmental light cycle through dedicated and distributed photoreceptors (see later).

Localization of Photoreceptors and Mechanisms of Entrainment

Circadian clocks are set to local time by entrainment to daily cycles in the immediate physical environment. While a variety of potential cycles could be utilized, a universal feature of circadian oscillators is that they can be entrained by environmental light cycles, and therefore the circadian system must have access to information from photoreceptors.

Extraretinal Photoreceptors

A striking feature of invertebrate circadian organization is that invertebrate species rarely rely exclusively on photoreceptors of the visual system for entrainment, and with few exceptions, removal of the eyes of mollusks, crustaceans, or insects has little effect on the ability of the circadian oscillators to be entrained by light. This indicates that extraretinal receptors are dominant in detecting light for entrainment in invertebrates.

In view of the wide tissue distribution of circadian oscillators in invertebrates, the question of anatomical localization of the relevant photoreceptors for entrainment is particularly interesting. Data thus far

suggest that the wide distribution of oscillators may also be reflected in a wide distribution of photoreceptors for entrainment. In the fruit fly, for example, it has been shown that the molecular clocks in the brain, legs, antennae, gut, and excretory system can be directly entrained by light *in vitro*, suggesting that the cells that generate the circadian oscillations also contain the molecular machinery necessary for their entrainment. Data from various species of moths paint a similar picture. They suggest that the eclosion rhythm is entrained by a photoreceptor located in the brain; the pacemaker in moth testis can be entrained by light *in vitro*, indicating some cells in the testis–seminal duct complex must be photosensitive, and the photoreceptor for entrainment of the pacemaker in the prothoracic glands appears to be in the gland itself.

Studies on mollusks further reinforce the idea that circadian pacemaker cells may also contain the molecular machinery for photic entrainment. *Aplysia* and *Bulla* eyes isolated *in vitro* can be phase shifted by light, demonstrating that ocular photoreceptors are sufficient for entrainment of the ocular pacemaker. Efforts to identify the specific cells within the molluscan eye that are responsible for entrainment have suggested that the pacemaker cells are the site for phototransduction.

Participation of Retinal Photoreceptors in Entrainment

Although 'organized' photoreceptive structures are frequently bypassed in entrainment in insects, there are at least two insects, the cockroach and the cricket, in which the compound eyes are the exclusive photoreceptors for entrainment. In the cockroaches and crickets, section of the optic nerves between the eyes and the optic lobe, or painting over the compound eyes, eliminated entrainment of the locomotor activity rhythm to light cycles.

In a variety of invertebrate species, even when extraretinal photoreception is sufficient for entrainment, there is evidence that the eyes participate. For example, in *D. melanogaster*, although the eyes are not required, they appear to contribute to the entrainment of behavioral rhythms such as locomotor activity and eclosion. In chelicerate arthropods, both retinal and extraretinal photoreceptors may be involved in the entrainment of ERG amplitude rhythms that are driven by pacemakers in the brain. In *L. polyphemus*, for example, the ERG amplitude rhythm of the lateral eyes can be phase shifted via either retinal photoreceptors in the ventral, median, and lateral eyes or extraretinal photoreceptors in the telson.

In summary, extraretinal photoreception is prevalent, and although retinal photoreceptors may serve as alternate sources of photic information for the circadian system, they are generally not required. With few exceptions (cockroaches, crickets) the results are consistent with the notion that cells that contain the molecular mechanisms for generating circadian oscillations also contain the phototransduction machinery for their entrainment. These observations raise at least two questions of interest. The first concerns the general reliance on extraretinal photoreception for entrainment. It seems likely to be significant that extraretinal pathways are quite prevalent, even in species that have well-developed visual systems, and this raises the possibility that the photoreceptors involved in entrainment may possess unique properties relative to photoreceptors that participate in more 'classical' visual processes. Alternatively, the widespread distribution of extraretinal pathways may simply be a reflection of the hypothetical participation of a photopigment in the early evolution of the molecular machinery of biological clocks. Here the suggestion is that natural selection has maintained the close relationship between phototransduction and the generation of circadian oscillations, as a mechanism for maintaining temporal order among various oscillators and as an alternative to the development of novel coupling pathways and mechanisms between oscillators within the individual.

Cellular Mechanisms of Entrainment by Light

There has been substantial progress in understanding the cellular and molecular mechanisms by which light regulates the phase of the circadian oscillator in the mollusk eye. In *Bulla*, the data suggest that light is absorbed by an opsin-based photopigment. By an unknown mechanism the absorption of photons leads to a depolarization of the putative pacemaker neurons, the basal retinal neurons. The depolarization is both necessary and sufficient for phase shifting the oscillator. Reduction in extracellular calcium concentration or application of the calcium channel blocker nickel chloride blocked the phase shift caused by depolarization, indicating that a voltage-dependent transmembrane calcium flux is a step in the phase-shifting pathway; however, the role of the increase in intracellular calcium is uncertain. In *Aplysia*, an opsin-based pigment is also a likely candidate for photon absorption, and the entrainment pathway involves depolarization. As in *Bulla*, a calcium flux may also be an important element, though the exact role of calcium is uncertain. There is also evidence that protein synthesis is involved in phase shifting by light in *Aplysia*, since protein synthesis inhibition can block light-induced advance phase shifts.

In *Drosophila*, one photopigment that has been shown to participate in entrainment is cryptochrome, and disruptions of the gene coding for this

flavin-containing protein, which has been localized to cells that generate circadian oscillations, cause major disruptions in entrainment. Precisely how cryptochrome is coupled to the molecular machinery of the clock within the cell has not been delineated in detail, but it appears to be different between the central brain clock, where cryptochrome is involved in phototransduction for entrainment but is not required for expression of the rhythm, and the peripheral clocks, where it appears to be part of the molecular clockworks.

Identification of Output Pathways

In addition to discovering the organization of the pacemaking system, another important issue is how circadian oscillators impose periodicity on the various physiological and behavioral processes they control. At this time, there has been no complete description of a pathway for the biochemical and physiological steps that are involved in the representation and transmission of circadian information for the regulation of specific overt rhythms; there has, however, been significant progress in the identification of potential signaling molecules. Two examples will serve to illustrate.

Rhythms in Insect Development

A large number of studies suggest that secretion of a variety of insect hormones, including ecdysone, prothoracicotropic hormone (PTTH), and eclosion hormone, is under the control of the circadian system during development. These hormones are involved in the regulation of various developmental events, including ecdysis – the process by which the insect sheds its exoskeleton. The hormonal control of adult ecdysis (eclosion) has been particularly well studied. Although it appears that there are significant differences in detail among different species, it seems to be consistent that at least two physiologically and anatomically separable circadian oscillators are involved in regulating the timed release of these hormones – one located in the brain and a second located in the prothoracic glands. The clock in the brain appears to interact with the clock in the prothoracic glands to regulate the timing of ecdysis.

Rhythms in Retinal Sensitivity to Light

A particularly nice demonstration of a neural link in circadian regulation comes from studies on the efferent regulation of retinal sensitivity in the horseshoe crab *L. polyphemus*. In *Limulus*, a circadian pacemaker in the brain drives synchronous activity in efferent fibers that project to the lateral, ventral, and median eyes, all of which show a circadian rhythm in sensitivity to light. These efferents drive a variety of functional and structural changes in the eye that alter its sensitivity to light. While it is likely that multiple transmitters are involved, one signaling cascade is known to involve the release of the biogenic amine octopamine. Octopamine in turn stimulates a rise in cAMP, which activates cAMP-dependent protein kinase. The activated kinase mediates several of the known efferent effects on the eye.

In summary, the efforts to identify mechanisms by which circadian clocks regulate the effectors under their control indicate that the pacemakers are able to exploit the full range of signaling mechanisms, from direct neural transmission, to neuromodulatory signals, to circulating hormones.

See also: Circadian Organization; Circadian Rhythm Models; Circadian Organization in Non-Mammalian Vertebrates; Genetic Regulation of Circadian Rhythms in Drosophila; Photoreceptors: Physiology; Sleep and Waking in Drosophila; Visual System Development: Invertebrates.

Further Reading

Arechiga H, Fernandezquiroz F, Demiguel F, et al. (1993) The circadian system of crustaceans. *Chronobiology International* 10: 1–19.

Dunlap J, Loros J, and DeCoursey P (eds.) (2004) *Chronobiology: Biological Timekeeping*. Sunderland, MA: Sinauer Associates, Inc.

Giebultowicz JM (2000) Molecular mechanism and cellular distribution of insect circadian clocks. *Annual Review of Entomology* 45: 769–793.

Helfrich-Forster C (2005) Neurobiology of a fruit fly's circadian clock. *Genes, Brain and Behavior* 4: 65–75.

Homberg U, Reischig T, and Stengl M (2003) Neural organization of the circadian system of the cockroach *Leucophaea maderae*. *Chronobiology International* 20: 577–591.

Page TL (1985) Clocks and circadian rhythms in insects. In: Kerkut G and Gilbert L (eds.) *Comprehensive Insect Biochemistry, Physiology, and Pharmacology VI. Sensory Physiology*, pp. 577–652. Oxford: Pergamon.

Page TL (2001) Circadian systems of invertebrates. In: Takahashi J, Turek F, and Moore RY (eds.) *Handbook of Behavioral Neurobiology 12. Circadian Clocks*, pp. 79–110. New York: Kluwer Academic/Plenum Publishers.

Saunders DS (2002) *Insect Clocks,* 3rd edn., Amsterdam, The Netherlands: Elsevier.

Underwood H, Wassmer GT, and Page TL (1997) Seasonal and diurnal cycles. In: Danzer W (ed.) *Handbook of Comparative Physiology 13: Comparative Physiology*, pp. 1653–1763. New York: Oxford University Press.

Circadian Rhythm Models

P D Smolen and J H Byrne, University of Texas Medical School at Houston, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Many organisms, including animals, plants, and cyanobacteria, undergo 24 h rhythms in physiology and behavior, and these rhythms persist in constant dark conditions. Therefore, self-sustaining internal oscillators with periods of approximately 24 h must have evolved in these organisms. The operation of complex biological systems, such as these circadian oscillators, is generally too complex to understand by intuition. Mathematical models are therefore essential to provide a picture of how the system components work together to determine the dynamics and the response to changes in the organism or the environment. Models are also useful to suggest further experiments, to test whether the system responds in ways that the model predicts, or to verify/disprove hypothetical regulatory mechanisms that are predicted to explain specific system behaviors.

Before experiments characterized the molecular mechanisms of these oscillators, attempts were made to describe their dynamics using phenomenological oscillator models from physics. Such relatively simple models are still important. For example, the Van der Pol equations, originally developed to describe oscillating electrical circuits, have proven useful to model light-induced entrainment and phase-shifting of circadian oscillations in humans and cyanobacteria.

Early Models Illustrated Essential Roles of Negative Feedback and Posttranslational Modifications

Data establishing the molecular nature of the internal circadian oscillators have been obtained for many organisms. Mutations have been found that alter or abolish circadian oscillations in mammals, fruit flies (*Drosophila*), fungi (*Neurospora*), and plants. In all these organisms, circadian oscillations in the rate of transcription of genes and in the levels of gene protein and mRNA products have been characterized. Circadian rhythms of protein phosphorylation and degradation have also been observed. Therefore, the internal oscillators are believed to rely on periodic feedback regulation of gene transcription by gene product proteins, as well as periodic posttranslational modifications. Modeling has proven very useful to characterize more specific schemes of molecular interactions and to suggest experimental tests. Since the 1990s, numerous investigators have developed models that have enhanced the understanding of the ways in which feedback regulation and posttranslational modification are organized to sustain oscillations that are robust to changes in the organism's internal or external environment, and that are entrained by light or temperature cycles.

Sustained oscillations in reaction rates and concentrations can be exhibited by biochemical systems with nonlinear kinetics that operate away from equilibrium due to ongoing energy input (e.g., ATP hydrolysis). Most commonly, to drive oscillations, a particular type of kinetic nonlinearity is required – a negative feedback loop. In a negative feedback loop of chemical reactions, the product of the last reaction inhibits the first reaction. A model proposed by Goodwin in 1965 was the first to simulate circadian oscillations due to negative feedback in which a protein repressed its own gene. This venerable model is still proving useful, for example, to simulate aspects of *Neurospora* oscillations.

Drosophila was the first organism for which a negative feedback loop was found essential for circadian oscillations. Experiments established that the period protein (PER) inhibits the transcription of its own gene (*per*). Early modeling by Goldbeter determined that a negative feedback loop in which PER underwent several posttranslational modifications prior to repressing *per* transcription could sustain circadian oscillations in PER levels and formation. Both a critical number of sequential reactions in the loop (at least six or seven) and a significant, sigmoidal nonlinearity in the repression of *per* by PER (e.g., a Hill function of PER with an exponent of 3 or 4) appear necessary to sustain large-amplitude oscillations. Subsequent experiments have confirmed the necessity of this negative feedback loop. PER is multiply phosphorylated, and specific phosphorylations appear to gate PER nuclear entry. PER nuclear entry may also require the TIM protein, and *tim* mutations block rhythmicity. Nuclear PER interacts with a transcriptional activator protein, CLK, and inhibits CLK's binding to DNA. In this way, PER rhythmically inhibits the transcription of not only *per* but also numerous other genes that are activated by CLK, thereby driving numerous physiological/behavioral aspects of rhythmicity. Transcriptional repression by PER persists for several hours, during which PER is progressively phosphorylated. This process is an important factor in setting the long (~24 h) oscillation period. When PER is maximally phosphorylated, it degrades rather abruptly,

and CLK-dependent transcriptional activation of *per* and other genes then resumes, beginning the next cycle. Figure 1 illustrates this negative feedback loop and simulated PER circadian oscillations.

Experiments have identified analogous negative feedback loops that underlie circadian rhythms in mammals (rhythmic repression of transcription by PER and cryptochrome (CRY) proteins) and in *Neurospora* (rhythmic repression of transcription by FRQ protein) as well as in the flowering plant *Arabidopsis*. Models based on negative feedback have simulated molecular oscillations in these organisms. Indeed, it appears that a generic mechanism of rhythmicity based on a negative feedback loop of transcriptional repression is very widespread. Note, however, that some observations indicate rhythmicity can persist in the absence of these loops, perhaps suggesting that multiple mechanisms of rhythm generation exist. For example, in cyanobacteria, circadian oscillations in the phosphorylation of proteins necessary for circadian rhythms (KaiA–C) persist in the absence of transcription.

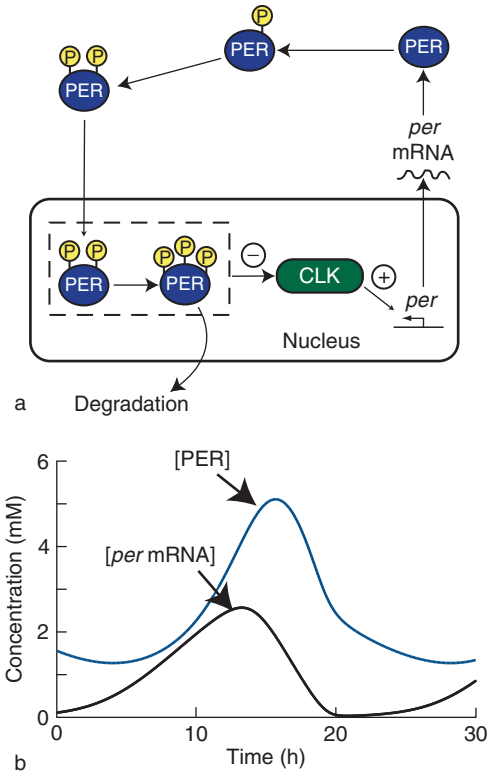

Figure 1 (a) Schematic of the negative feedback loop which appears to be central to circadian rhythm generation in *Drosophila*. Transcription of *per* is activated by the CLK transcription factor. *Per* mRNA is translated, and PER protein then undergoes multiple phosphorylations. Data suggest some phosphorylation must precede PER's nuclear entry. Nuclear PER represses *per* expression by interfering with CLK, closing the negative feedback loop. Subsequently, PER is further phosphorylated over succeeding hours. Fully phosphorylated PER is then rapidly degraded. PER degradation releases *per* repression, allowing a new cycle to begin. When repression is relieved, *per* expression is again induced. (b) Oscillations of *per* mRNA and protein simulated by using the original model of the negative feedback loop from Goldbeter A (1995) A model for circadian oscillations in the *Drosophila* period protein (PER). *Proceedings of the Royal Society of London, Series B* 261: 319–324.

The models with negative feedback loops also rely on nonlinear repression of transcription to sustain large-amplitude oscillations (e.g., sigmoidal Hill functions of FRQ, CRY, or PER levels). However, experiments have failed to demonstrate cooperative binding of these proteins to regulatory regions of their target genes. Therefore, a mechanism for the hypothesized nonlinear repression has not been found. However, a model of circadian rhythm generation in *Drosophila* by Smolen et al. suggests that multiple sequential chromatin modifications that accompany binding of PER or CLK to DNA could account for nonlinear repression. In mammals, binding of CLK and/or PER is known to be accompanied by such rhythmic chromatin modifications (multiple histone acetylation/methylation events).

Recent Models Consider Additional Positive Feedback and Simulate the Effects of Mutations

Experiments have revealed additional molecular interactions that modulate the negative feedback interactions discussed previously. In *Drosophila*, the transcriptional activator CLK indirectly represses its own gene. In mammals, an analogous activator, denoted BMAL1, indirectly represses its own gene. In both organisms, a positive feedback loop is also present. This loop is organized as follows for mammals and *Drosophila* (*Drosophila* in parentheses). First, *bmal1* (*clk*) gene expression increases levels of BMAL1 (CLK) protein. Second, the increased BMAL1 (CLK) activates *cry* and *per* expression. Third, the expressed CRY and PER inhibits the action of BMAL1 (CLK). Finally, this inhibition of BMAL1 (CLK) de-represses *bmal1* (*clk*), increasing *bmal* (*clk*) expression and thereby closing the positive feedback loop. A detail of the mammalian loop is that BMAL1 inhibition reduces expression of the REV-ERBα transcriptional repressor, and this is the way in which *bmal1* expression is de-repressed. In *Neurospora*, a similar positive feedback loop involves FRQ and the activator WCC. Figure 2 illustrates the mammalian positive and negative feedback loops.

Recent models have examined plausible effects of these positive feedback loops. The existence of both positive and negative loops has been suggested to increase the robustness (stability) of the oscillation period and/or amplitude given biochemical parameter

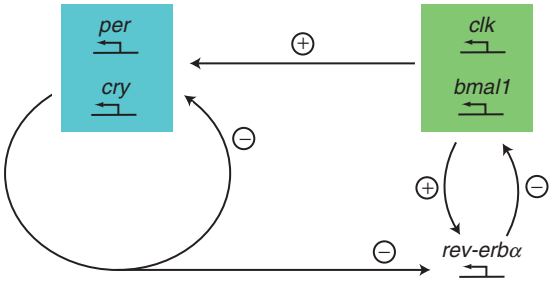

Figure 2 Schematic of the feedback loops that appear to underlie mammalian circadian rhythm generation. *Per* and *cry* expression is induced by the *bmal1* and *clk* gene products. In the first negative feedback loop, PER and CRY proteins repress expression of their own genes via blocking of transcriptional activation by BMAL1/CLK. In a second negative feedback loop, BMAL and CLK proteins indirectly repress their own genes by inducing *rev-erbα*. The REV-ERBα protein is a repressor. In the positive feedback loop, expression of *bmal1* and/or *clk* leads to induction of *per* and *cry*, which inhibits induction of *rev–erbα* repressor by BMAL1/CLK. *Bmal1* expression is therefore de-repressed and can increase further, closing the positive feedback loop. For clarity, proteins are omitted from the figure and multiple *per* and *cry* isoforms are not distinguished.

variations, such as would occur between individuals in a species. However, others have suggested that positive feedback may not increase robustness. Rather, it may primarily drive oscillations in the expression of 'output' circadian genes that are not part of the core oscillatory mechanism but act to govern physiological and behavioral rhythms. Indeed, data suggest that in *Drosophila*, the level of CLK remains relatively constant.

Another perspective on robustness of oscillations has been obtained from studies that examined a range of simplified model network motifs that simulate oscillations of gene expression. These investigations attempted to determine what type of networks exhibit the greatest stability of oscillation period and/or amplitude to parameter changes. In some of these studies, multiplicity of feedback loops was found to increase oscillation robustness.

Since the 1970s, numerous mutations have been found to alter or abolish circadian rhythms. Arrhythmia can be induced by mutation of *Drosophila per* or *clk*, mammalian *clk*, or *frq* in *Neurospora*. Less drastic mutations of these genes, which preserve partial function, can result in lengthened or shortened free-running circadian oscillations (i.e., in constant darkness the period differs from 24 h). Any viable model of circadian rhythm generation should at least qualitatively simulate the effects of such mutations. For mammals, gene duplication has led to several *per* genes that differ slightly in their importance for rhythmicity (*per1–3*). Two other mammalian transcriptional repressors, CRY1 and CRY2, are also core circadian proteins that function similarly to the PERs.

Complex models attempt to capture most of the currently known complexity of the mammalian oscillator (posttranslational modifications and the roles of additional core circadian genes) and therefore consist of as many as 73 differential equations. However, these models only represent regulation of approximately five core genes, and it is likely that additional genes help drive or shape circadian behavioral oscillations. Although there may be value in increasing the complexity of models to incorporate additional genes, it is clearly also necessary to develop 'reduced' models with fewer variables that capture one or a few of the critical mechanisms underlying rhythm generation and modulation, such as feedback loops, or entrainment by light stimuli. For *Drosophila*, the original model with a single negative feedback loop (Figure 1) remains useful. One- or two-variable reduced models with a time delay have also been developed that represent circadian transcriptional feedback in *Drosophila* and in *Neurospora*, as well as effects of light stimuli. However, a weakness of such models is that introducing explicit time delays generally and strongly promotes oscillations, without representing kinetic details of biochemical reactions.

An alternative approach to model reduction begins with a complex model containing many differential equations and projects the dynamics onto a two- or three-dimensional manifold. For example, the 73-variable model of the mammalian oscillator noted previously has been reduced to a two-variable model that retains most of the qualitative dynamics and is much easier to analyze.

Models Have Increased the Understanding of Entrainment and Phase-Shifting by Light Cycles, Light Pulses, and Temperature Changes

Circadian oscillators generically respond to light. Day–night cycles entrain these oscillators. That is, if the constant darkness oscillation period is within a few hours of the light–dark cycle length, then the oscillator will adjust its period to match the light–dark cycle. It is also generically observed that during a background of constant darkness, brief light pulses will shift the phase of the free-running oscillation. A phase–response curve (PRC) is commonly used to present the light pulse data. In plotting this curve, the abscissa represents the time in the free-running cycle at which the pulse is applied, and the ordinate represents the phase shift caused by the pulse. For a variety of organisms, the PRC has comparably sized regions of both positive and negative phase shifts and tends to have a rather abrupt transition from negative to positive phase shifts. Such a curve is a 'type 1' PRC.

For very strong stimuli, type 0 PRCs are occasionally observed. With these PRCs, phases after stimuli cluster near some fixed value and are thus independent of the phase before stimulus.

The biochemical effects of light on core circadian oscillator genes have been examined. In *Drosophila*, light indirectly enhances degradation of the PER protein (via enhanced TIM degradation, which in turn destabilizes PER). In mammals, light enhances expression of *per* genes, and in *Neurospora* light induces *frq* expression. Models of the mammalian and *Drosophila* oscillators have incorporated these effects and have succeeded in simulating light entrainment of oscillations as well as experimental type 1 PRCs. However, given the same model parameter values, these models have usually not succeeded in simulating type 0 PRCs. Figure 3 illustrates simulated light entrainment of *Drosophila* circadian oscillations and a simulated PRC.

Simulations of PRCs and entrainment can help in understanding human circadian rhythm sleep disorders. Advanced and delayed sleep phase syndromes have been identified, and simulations suggest that mutations decreasing or increasing phosphorylation of PER protein could phase-shift the rhythm and account for these syndromes. Experiments have indeed implicated *per* mutations. In other patients, the phase of sleep does not entrain to night and day but, rather, continuously varies. Analogously, in models, modest changes in PER degradation or in other kinetic parameters can alter oscillation period so that entrainment to a 24 h light–dark cycle is no longer possible.

Circadian models incorporating light responses have been seen, for some parameter ranges, to exhibit multiple dynamic behaviors, including coexistence of a steady state and an oscillating solution (hard excitation) or coexistence of two oscillating solutions (birhythmicity). With coexistence, the initial conditions (initial values of the dependent variables) determine which model behavior is observed. For hard excitation, modeling predicts that an appropriate perturbation (e.g., a light pulse of appropriate intensity) could switch the behavior from the oscillating solution to the steady state or vice versa. Indeed, in constant darkness, a single light pulse has been observed to abolish rhythmic oscillations, and oscillations can be restored by a second pulse. This observation is an example of a behavior that can only be understood or predicted with the help of a mathematical model. For certain parameter ranges in some models, chaotic behavior occurs (i.e., seemingly random variable trajectories without periodic behavior). However, chaotic dynamics have not been confirmed experimentally.

In many organisms, circadian rhythms are observed to be temperature compensated. That is, if the environmental temperature is varied within a limited but substantial range (typically 10–15 °C above or below normal), the period of the free-running rhythm in constant darkness is observed to remain almost constant, and near 24 h. Temperature compensation is surprising because the rates of all biochemical reactions increase with temperature, typically by a factor of approximately two for a temperature increase of 10–15 °C. Models have attempted to simulate

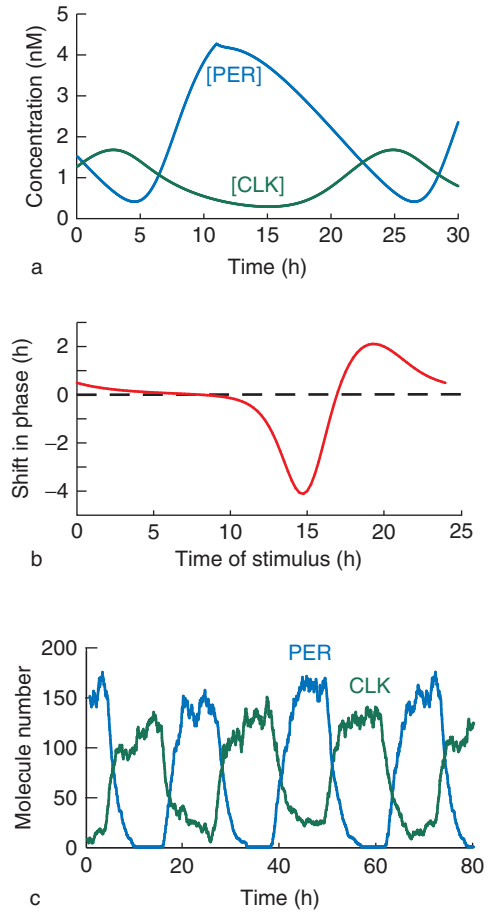

Figure 3 (a) Simulated circadian oscillations of *Drosophila* PER and CLK entrained to a short 22 h light–dark cycle. Note the accelerated degradation of PER at light onset ($t = 11$ h). (b) A simulated *Drosophila* PRC. Constant darkness circadian oscillations were phase-shifted by a brief light pulse, simulated as a 3 h enhancement of the rate of PER degradation. To generate the curve, light pulses were applied at various times during the circadian oscillation. (c) Simulated stochastic oscillations of PER and CLK suggest that relatively reproducible oscillations, with little variation in period, can be sustained when copy numbers of core clock proteins are on the order of 100. (a, b) model used from Smolen P, Hardin PE, Lo BS, Baxter DA, and Byrne JH (2004) Simulation of *Drosophila* circadian oscillations, mutations, and light responses by a model with VRI, PDP-1, and CLK. *Biophysical Journal* 86: 2786–2802. (c) Model used from Smolen P, Baxter DA, and Byrne JH (2002) A reduced model clarifies the role of feedback loops and time delays in the *Drosophila* circadian oscillator. *Biophysical Journal* 83: 2349–2359.

temperature compensation. To date, the only explanation supported by modeling relies on simultaneous acceleration of opposing reactions. For example, sequential phosphorylations of PER protein must be completed during a circadian cycle in *Drosophila* or mammals. At each phosphorylation site, there are opposing phosphorylation and dephosphorylation reactions. If, at each site, the rates of the opposing reactions are increased by similar amounts, then simulations have illustrated that the time required for progression through the sequence of phosphorylations can remain almost constant. This sequence of PER phosphorylations, in turn, appears to be an important determinant of oscillation period length, and as a result, oscillation period does not change appreciably upon a temperature increase.

The above explanation of temperature compensation does rely on 'fine-tuning' of kinetic parameters in the opposing reactions. However, it appears plausible that molecular evolution could lead to such fine-tuning if maintaining a circadian rhythm of proper period in different seasonal temperatures is evolutionarily advantageous to an organism. Indeed, a rhythm of proper period is important for the organism's fitness to ensure that active behaviors (food seeking) and rest/sleep occur at the correct phases of the day–night cycle.

Despite temperature compensation, temperature cycles are observed to entrain circadian rhythms, and temperature pulses can phase-shift rhythms. Models in which temperature affects the rates of opposing reactions can show these effects as well, but the phenomena have not been studied in detail.

Models Illustrate Sensitivity of Rhythms to Stochastic Molecular Noise and Suggest Mechanisms to Enhance Noise Resistance

Models of biochemical systems can be extended to include random variability in the times of individual macromolecular synthesis and degradation events. Observations indicate that the timing of transcription of individual mRNAs is indeed quite random. Also, for some macromolecules, such as specific mRNAs or transcription factors, average molecule numbers per cell can be modest ($\sim$100). In this case, variability in the times of single-molecule synthesis or degradation will lead to significant stochastic fluctuations in macromolecule numbers. These fluctuations can considerably alter the dynamics of biochemical pathways.

It is difficult to determine absolute copy numbers of proteins or mRNAs in neurons, and for circadian gene products, these data are generally not available. Therefore, models including molecule number fluctuations have focused on qualitative predictions of the minimal average copy numbers needed to sustain oscillations of reproducible period and amplitude. Simulations predict that a few hundred copies of PER and other key proteins should be present on average, and average numbers of mRNAs can be somewhat less. Figure 3(c) illustrates *Drosophila* model simulations that sustain reproducible molecular oscillations when the average copy number of key proteins is on the order of 100.

Several theoretical studies have identified mechanisms capable of increasing resistance to either stochastic and environmental fluctuations or noise in genetic oscillators. One conclusion is that positive feedback can enhance noise resistance by generating oscillator dynamics that are characterized by abrupt and large transitions states of high and low gene expression. A related conclusion is that noise resistance is increased in oscillators for which the limit cycle is strongly attracting (i.e., for which perturbations away from the normal periodic time course decay very quickly).

Models Examine the Efficacy of Proposed Mechanisms of Intracellular Synchronization

In the brains of mammals and in *Drosophila*, specific small groups of neurons are responsible for generating the circadian rhythm. This central oscillator, in turn, entrains peripheral oscillators localized in other organs (e.g., the mammalian liver). Two types of synchronization are required in this scheme. Rhythm-generating neurons must synchronize with each other, and the central oscillator must synchronize (entrain) the peripheral oscillators.

In mammals, rhythmogenic neurons are located within the suprachiasmatic nucleus (SCN). Synchronization of these rhythmogenic neurons (and of analogous *Drosophila* neurons) appears to rely on coordinated release of a diffusible neurotransmitter, such as GABA or vasoactive intestinal polypeptide. Electrical coupling of neurons by gap junctions may also play a role. Diffusible neurotransmitters and electrical coupling modulate neuronal electrical activity, and data indicate that variations in electrical activity and Ca^{2+} influx regulate *per* expression. These data therefore suggest a mechanism for rhythm synchronization within the central oscillator. Synchronization of central and peripheral oscillators relies on pulsatile secretion of hormones, such as gonadotropin-releasing hormone, from neurons within or coupled to the SCN.

Models have simulated synchronization of SCN neurons by diffusible neurotransmitters. A plausible

model should also make experimental predictions. One prediction is that efficient synchronization may be achieved even if the average neurotransmitter concentration would dampen individual oscillating neurons. In this case, due to the global neurotransmitter oscillation in the SCN, cells are effectively synchronized. As experimental characterization of SCN neurotransmitter release and possible gap junction coupling proceeds, model refinement will be important to help understand which synchronization mechanisms are most important and to predict how drugs or other environmental changes could alter synchrony.

Directions for Research

Experimental data characterizing the mechanisms of circadian oscillations continue to accumulate rapidly. As a result, no published model of circadian rhythms in mammals, *Drosophila*, or other organisms represents all the known details of rhythm generation. Revision of current models will be necessary as new data become available. However, it is likely that core mechanistic elements represented in current models will remain essential, such as negative feedback loops based on rhythmic repression of clock genes such as *per*, cycle lengths set largely by the time required for multiple phosphorylations of proteins regulating transcription, and relatively simple representations of the known responses to light and temperature.

Conclusion

Models of circadian rhythm generation have helped to determine that circadian oscillators in organisms including animals, fungi, and some prokaryotes share a common element of a negative feedback loop. Due to this loop, core circadian genes are rhythmically repressed by some of their protein products. The 24 h oscillation period is largely required for multiple posttranslational modifications of these core gene products, which are then degraded, allowing the next cycle of transcription. In some organisms, positive feedback loops modulate oscillation phase and amplitude. Multiple feedback loops may increase oscillation robustness to stochastic fluctuations and to mutations. Whether oscillations can be sustained by mechanisms other than transcriptional feedback is still being studied. Models have also helped in understanding some human sleep disorders, predicting mutations in *per* as one causative factor.

Negative feedback loops in which proteins act (directly or indirectly) to repress the transcription of their own genes appear to be a common motif for sustaining biologically important oscillations. Some examples are the cyclical development of somites of the vertebrate embryo, driven by a genetic oscillator in the mesoderm; the oscillatory response of p53 tumor suppressor expression to DNA damage; and oscillations in nuclear levels of the transcription factor NF-κB. Artificial gene circuits with oscillating expression are also being developed and might prove useful for gene therapies that seek to normalize pulsatile release of hormones or other proteins. Models similar to those discussed for circadian rhythms are being applied to characterize these diverse oscillations and predict their responses to environmental and pharmaceutical stimuli in healthy and diseased states.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Genetic Regulation of Circadian Rhythms in Drosophila; Genetics of Circadian Disorders in Humans; Learning and Memory in Invertebrates: *Drosophila*; Sleep and Waking in Drosophila.

Further Reading

Aton SJ and Herzog ED (2005) Come together, right ... now: Synchronization of rhythms in a mammalian circadian clock. *Neuron* 48: 531–534.

Becker-Weimann S, Wolf J, Herzel H, and Kramer A (2004) Modeling feedback loops of the mammalian circadian oscillator. *Biophysical Journal* 87: 3023–3034.

Bernard S, Gonze D, Cajavec B, Herzel H, and Kramer A (2007) Synchronization-induced rhythmicity of circadian oscillators in the suprachiasmatic nucleus. *PLoS Computational Biology* 3: e68.

Forger DB and Peskin CS (2003) A detailed predictive model of the mammalian circadian clock. *Proceedings of the National Academy of Sciences of the United States of America* 100: 14806–14811.

Goldbeter A (1995) A model for circadian oscillations in the *Drosophila* period protein (PER). *Proceedings of the Royal Society of London, Series B* 261: 319–324.

Goldbeter A (2002) Computational approaches to cellular rhythms. *Nature* 420: 238–245.

Indic P, Gurdziel K, Kronauer RE, and Klerman EB (2006) Development of a two-dimensional manifold to represent high dimension mathematical models of the intracellular mammalian circadian clock. *Journal of Biological Rhythms* 21: 222–232.

Kronauer RE, Gunzelmann G, Van Dongen HP, Doyle FJ, and Klerman EB (2007) Uncovering physiologic mechanisms of circadian rhythms and sleep/wake regulation through mathematical modeling. *Journal of Biological Rhythms* 22: 233–245.

Lakin-Thomas PL (2006) Transcriptional feedback oscillators: Maybe, maybe not. *Journal of Biological Rhythms* 21: 83–92.

Shearman LP, Sriram S, Weaver DR, et al. (2000) Interacting molecular loops in the mammalian circadian clock. *Science* 288: 1013–1019.

Smolen P, Baxter DA, and Byrne JH (2002) A reduced model clarifies the role of feedback loops and time delays in the *Drosophila* circadian oscillator. *Biophysical Journal* 83: 2349–2359.

Smolen P, Hardin PE, Lo BS, Baxter DA, and Byrne JH (2004) Simulation of *Drosophila* circadian oscillations, mutations, and light responses by a model with VRI, PDP-1, and CLK. *Biophysical Journal* 86: 2786–2802.

Vilar JM, Kueh HY, Barkai N, and Leibler S (2002) Mechanisms of noise-resistance in genetic oscillators. *Proceedings of the National Academy of Sciences of the United States of America* 99: 5988–5992.

Wagner A (2005) Circuit topology and the evolution of robustness in two-gene circadian oscillators. *Proceedings of the National Academy of Sciences of the United States of America* 102: 11775–11780.

Winfree AT (2001) *The Geometry of Biological Time*, 2nd edn. New York: Springer.

Circadian Rhythms in Sleepiness, Alertness, and Performance

J D Minkel and D F Dinges, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Circadian rhythms (i.e., biological processes that repeat about every 24 h) reflect the incorporation of Earth's daily rotation into living systems. Internal circadian clocks are found in all eukaryotic and even some prokaryotic organisms indicating that natural selection has highly favored circadian rhythms from the beginnings of life for virtually all species.

The suprachiasmatic nucleus (SCN) in the hypothalamus serves as the brain's master clock, in humans and many other animals as well. Neurons in the SCN fire in a 24 h cycle that is driven by a transcriptional–translational feedback loop. Loss of the SCN abolishes the circadian rhythms of a range of behaviors and physiological processes, including sleep, if the animals are not given other external timing cues. Under normal circumstances, the SCN is reset on a daily basis by light inputs from the retina during the day and by melatonin secretion from the pineal gland at night.

The SCN temporally organizes physiological and behavioral tendencies, but the circadian rhythms it promotes are not fixed. This plasticity is believed to reflect adaptation to dynamic demands of the environment such as seasonal changes in daylight. Humans in industrialized societies, however, routinely push the limits of the adaptive capacity of circadian rhythms. Social demands and opportunities (e.g., night shift work, travel across time zones) often lead to abnormal sleep and wake times that are out of phase with biological requirements. The circadian clock has been shaped by natural selection to regulate physiological processes to occur at optimal times relative to each other or relative to the external environment. With the introduction of artificial lighting and 24 h services, however, many humans are awake when their biological clocks are telling them it is time to sleep and trying to sleep when the circadian clock is sending signals throughout the brain that it is time to be awake. The combination of restricting sleep and attaining it at biologically misaligned times is problematic. Approximately 6 million people in the US alone work during the biological night without adapting their internal physiology to these conditions.

In this article, we review what is currently known about circadian modulation of neurobehavioral and affective systems in humans.

Identifying the Circadian Rhythm in Sleep and Wakefulness

Subjective Measures of Fatigue and Alertness

Circadian rhythmicity in neurobehavioral variables has been successfully identified using a wide array of subjective measures of alertness and fatigue. Visual analog scales, Likert-type rating scales, the Karolinska sleepiness scale, the activation–deactivation adjective check list, and the profile of mood states have all demonstrated sensitivity to circadian fluctuations in sleepiness and mood.

Subjective measures may be particularly vulnerable to numerous confounding influences, however. Masking (i.e., the effects of noncircadian factors on the measurement of circadian rhythmicity) frequently prevents underlying circadian signals from being separated from noise. The context in which such measurements are taken (regardless of whether the setting is experimental or environmental) is a major source of masking effects. Demand characteristics, levels of motivation or boredom, distractions, stress, food intake, posture, background noise, ambient temperature, lighting conditions, and pharmacological factors (e.g., caffeine use) are some of the influences that have been identified to mask underlying circadian rhythmicity in subjective measures. These same variables can also create the false appearance of a circadian rhythm, especially in uncontrolled studies where daily schedules may be influenced by sociocultural factors as much as underlying biology.

Circadian rhythms in mood states and sleepiness can be masked by physical, mental, and social activity as well, even in highly controlled experiments. For example, subjects have reported feeling less alert after being challenged to perform. Thus, if not properly controlled, the drop in subjective alertness due to performance effort can mask circadian effects. Similarly, social interactions with staff and/or other subjects in a study can influence subjective variables. In experimental studies, this potential confound can be reduced through careful training of staff and attention to interpersonal interactions among study participants. However, experiments that fail to control social variables risk contaminating true endogenous circadian variance in self-report measures of alertness and fatigue.

These outcomes can also be masked by prior sleep–wake schedules. Sleep and sleep loss have large effects on alertness and performance that can interact with and alter subtler circadian fluctuations. Despite their sensitivity to masking factors, however, subjective scales have been used to index circadian rhythmicity by repeated administration across the day under carefully controlled conditions.

Cognitive Performance

Objective performance measures also demonstrate circadian rhythmicity and are generally less susceptible to masking influences than subjective reports of mood and sleepiness. Tasks that assess psychomotor and/or cognitive speed are particularly sensitive to circadian fluctuations in arousal and somewhat more resistant to masking effects than many other measures. For example, performance on the well-validated Psychomotor Vigilance Test, which shows almost no learning curve and little variance within and between healthy subjects who are fully rested, has been demonstrated to vary with other markers of circadian phase such as melatonin and core body temperature (see Figure 1).

Circadian variation has also been demonstrated in a wide range of cognitive tests including search-and-detection tasks, simple sorting, logical reasoning, memory access, and even school performance and meter-reading accuracy. A number of studies have concluded that different tasks and task parameters may yield different peak phases of circadian rhythmicity, but under strictly controlled laboratory conditions, most of these intertask differences disappear. Thus, it can generally be stated that circadian rhythms of cognitive and psychomotor performance covary. Furthermore, these rhythms are synchronized with physiological markers of circadian phase, such as body temperature.

Electroencephalographic and Ocular Measures

The circadian rhythm in task performance and subjective states reflects functional changes in the brain. Neural activity, as measured by evoked potentials or event-related potentials (ERPs), has been used to measure alertness. ERPs are relatively subtle signals that must be measured over many consecutive stimulus probes in order to separate them from background electroencephalographic (EEG) noise. Therefore, ERPs are usually recorded during repetitive tasks, such as reaction time tasks and search-and-detection tasks. Circadian rhythms in the changes to amplitude and location of ERP waves have been identified and interpreted to reflect fluctuations in alertness. Measurement of EEG during wakefulness has demonstrated circadian rhythmicity, especially in theta and alpha frequencies. The amounts of theta and alpha activity in the resting EEG are associated with levels of alertness.

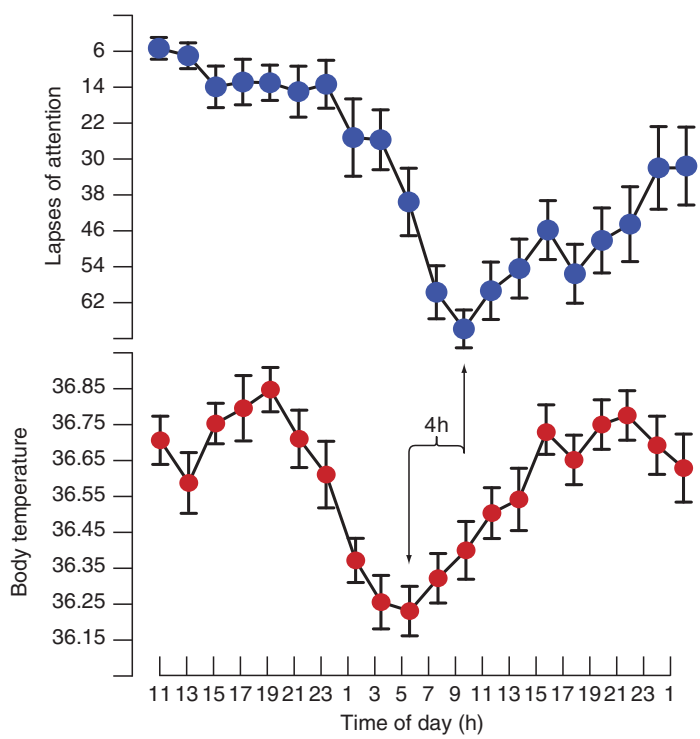

Figure 1 The mean (SEM) circadian minimum (nadir) in body temperature occurred approximately 4 h before the maximum of lapses of attention (note downward deflection on the upper graph indicates worse performance).

Sleep latency measured by EEG and electrooculographic (EOG) activity has been found to exhibit marked circadian variation. Since sleep latency is an indication of propensity of the brain to fall asleep, the circadian system is actively regulating sleepiness and sleep–wake cycles. In addition to shifts toward lower EEG frequencies at night, there is a circadian-mediated increase in slow eyelid closure and slow rolling movements, both of which herald the onset of sleep. Finally, autonomic tone has also been shown to covary with sleep pressure as measured by pupillometry (pupils dilate with sympathetic activation).

Midafternoon Dip

In addition to circadian fluctuations in performance that are tightly linked to physiological markers of phase, some individuals appear to demonstrate a short-term dip in performance in the afternoon (sometimes called the midafternoon dip, siesta, or postprandial dip). This dip has been observed in both field and laboratory studies, but is not consistently demonstrated, suggesting that it may be a relatively weak variable.

The best evidence for the existence of such a midafternoon dip comes from studies of sleep propensity and on the timing of daytime naps. There is much less evidence that the midafternoon dip has deleterious effects on performance measures, which makes the phenomenon's relationship to the biological clock somewhat unclear. Although field studies have reported a decline in performance at this time of day the uncontrolled nature of such experiments prevents strong conclusions about the underlying biology of the phenomenon. As mentioned earlier, external variables can produce the illusion of circadian rhythmicity as well as mask it.

Problems in Detecting Circadian Rhythmicity in Performance

Practice Effects and Other Artifacts

Although repeated administration of the same task is important for identifying circadian fluctuations in performance, increasing familiarity with a task also produces a practice effect (i.e., improved performance with practice on a given task). Practice effects can be difficult to distinguish from the circadian rhythm, but the problem can be circumvented by testing subjects in different orders across times of day, thus balancing out improvements in performance due to learning. This solution assumes however that the practice effect and circadian rhythm are merely additive influences on performance that have the same relationship in every subject. For most tasks, neither assumption can be made with certainty. A better way to deal with the practice effect is simply to train subjects to their full capacity (i.e., until performance levels reach an asymptote) on tasks before attempting to assess circadian rhythms. When possible, the best way to deal with this problem is to use a measure that has no practice effects like simple psychomotor vigilance tasks.

In addition to the practice effect, many of the same variables that serve to mask circadian rhythmicity in subjective estimates of fatigue and alertness can also conceal, accentuate, or otherwise distort such rhythmicity on objective performance measures. The effects of masking can vary from changes in the shape of the circadian curve to fully obscuring the circadian influence. It is therefore difficult to extract meaningful information about the mathematical characteristics of the circadian phase without an understanding of the masking effects that influence the variables under investigation.

Not only is performance affected by learning, aptitude, and other masking factors, but also by neurocognitive processes and strategies that can be conscious or unconscious. For example, subvocalization (a strategy that can improve performance on certain tasks) has been reported to fluctuate with a circadian rhythmicity. Similarly, compensatory effort in the face of drowsiness can maintain objective performance at steady levels despite underlying circadian fluctuations. Finally, neuroimaging studies have begun to demonstrate that a given task can be accomplished using multiple neural pathways. Sleep-deprived subjects have been shown to compensate for their performance impairment by increasing activation in the parietal lobes. Such compensatory plasticity in neural function illustrates that the effects of sleepiness and arousal are not always expressed through differences in performance. For all of these reasons, successful demonstration of circadian rhythmicity in performance measures requires that tasks be structured in ways that minimize or eliminate masking effects.

Interindividual and Intraindividual Variability

A large body of literature supports relatively stable interindividual differences in many parameters related to the human circadian system. An important exception however is the period of the human circadian pacemaker. Although published reports spanning several decades asserted that the period could range from 13 h to 65 h, highly controlled data demonstrated that the period is very close to 24 h with little variation between or within subjects. The free running

period of the human pacemaker is also not heavily influenced by age in healthy subjects. The large variance reported in less-controlled studies therefore probably represents individual differences in the sensitivity of the circadian system's response to masking factors such as physical activity, knowledge of time of day, and exposure to artificial light.

In contrast to the human circadian period, individual differences in circadian amplitude, phase, and mean performance level have been demonstrated. Differences in phase are particularly important because they determine differences in circadian variables relative to time of day. For example, personal preference in timing of work hours is related to individual differences in circadian phase.

The tendency for people to prefer to be active and alert earlier or later (often referred to as 'morningness' or 'eveningness,' respectively) is probably the most substantial source of interindividual variation in circadian rhythmicity. Individuals who prefer different times of day differ endogenously in the circadian phase of their biological clocks. Objective performance measures and physiological variables (e.g., core body temperature) support the morningness/eveningness distinction. This traitlike difference in performance may be seen as a phenotypic aspect of circadian rhythmicity in humans.

Intraindividual differences in circadian rhythms refer to changes within one individual over time, such as differences in morning alertness during the summer compared to the winter. The circadian trough is associated with increased intraindividual variability in performance, reflecting increased wake-state instability similar to what is seen with extended wakefulness. Age is also a source of interindividual variability in the circadian rhythm of alertness and performance, but no consistent relevance of sex has been found.

Circadian Rhythmicity and Sleep–Wake Cycles

Sleep Deprivation

Unmasking the circadian rhythm has been a priority in sleep research for many years. An important method in this domain has been the constant routine procedure (also called the unmasking procedure) which involves keeping subjects awake with a fixed posture in a constant laboratory environment for at least 24 h. By maintaining such a static environment, many of the potential confounding factors (changes in lighting and noise, social interaction, cognitive fatigue, etc.) are eliminated or minimized. Relatively subtle circadian fluctuations can then be readily separated from background noise.

Sleep deprivation, however, can still exert a masking influence on neurobehavioral variables because sleep loss adversely affects performance. Performance may be pushed to extremes by a single night of sleep loss, making circadian fluctuations impossible to detect. This is precisely what happens to sleep latency in the widely used multiple sleep latency test (MSLT). Subjects with extremely high homeostatic sleep pressure fall asleep as quickly as possible regardless of their circadian phase. Nevertheless, when sleep pressure is not extremely elevated and the constant routine procedure is used, both circadian and homeostatic influences on performance can be seen.

Sleep–Wake Regulation

The interaction of circadian and homeostatic influences on performance and alertness has prompted efforts to mathematically model the regulatory processes involved. Current theory and mathematical models posit that the homeostatic process represents the drive for sleep which increases linearly during wakefulness and decreases with sleep. When the 'homeostat' reaches a certain threshold, sleep is triggered; when it is sufficiently dissipated, wakefulness is initiated. The circadian process represents daily oscillatory modulation of these threshold levels, but may have an additional role in promoting wakefulness even in the absence of sleep.

These two processes not only determine the timing of sleep and wakefulness, but also interact to determine waking neurobehavioral performance during sustained wakefulness. Sleep-deprivation experiments clearly show that the circadian process continues to influence performance even as the homeostat increases beyond what would normally initiate sleep.

Forced Desynchrony

Like the constant routine procedure, forced desynchrony is an experimental paradigm that unmasks circadian rhythmicity. In this procedure, study participants stay in an isolated laboratory in which the times for sleep and waking are scheduled to deviate from the normal day. Although humans can entrain to days that are not exactly 24 h, artificial days that are extremely short (e.g., 20 h) or very long (e.g., 28 h) are outside the adaptive capacity of the biological clock. When such a routine is kept for several days, the sleep–wake schedule cannot entrain to endogenous circadian rhythms. Studies using this method have clearly demonstrated that both circadian and homeostatic processes influence sleepiness, performance, and mood. The interaction of the two processes is oppositional during natural waking periods such that

a relatively stable level of alertness and performance can be maintained throughout the day.

Ultraradian Days

The 'ultradian' day (meaning 'very short day') paradigm was designed to sample waking behavior across the circadian cycle without significantly curtailing the total amount of sleep allowed. Studies using this method allow subjects relatively brief periods of sleep and wakefulness (e.g., 30 min of sleep followed by 60 min of wakefulness). Thus, ultraradian schedules maintain the normal ratio of sleep to wakefulness (i.e., 1:2). In one such experiment using 7 min of sleep to 13 min of wakefulness, a clear circadian rhythm emerged for the response time on a choice reaction time task. A movement time component was also recorded. This study showed that even in conditions of extremely short artificial days both circadian and homeostatic influences still interact to determine performance.

Circadian Disorders

The laboratory is not the only place one finds people divorcing their sleep–wake pattern from the tendencies of the biological clocks. Recent labor statistics estimate that about 6 million members of the American workforce engage in work at night. While emergency services have been offered 24 h a day for many years, nonessential services are increasingly being offered throughout the night as well. As expected from the research reviewed above, those who work at the circadian phase for sleep show an increased prevalence of sleep disturbance and related medical and psychiatric conditions relative to those who work during the day.

Field studies have estimated that at best, it takes at least 1 week before the biological clock adjusts to shift work. Complicating the problem is the circadian system's sensitivity to both light and dark. Although in a controlled environment the human circadian system can be shifted a full 12 h over a number of days, in natural settings sunlight will disrupt the reentrainment process. Many shift workers therefore never fully adjust physiologically to their schedules. Melatonin has been investigated as a means of facilitating phase shifts, but its ability to shift circadian phase in the presence of natural light has been found to be fairly weak.

Jetlag is also a common chronobiological problem in modern society. At any given time, approximately 500 000 people are traveling in planes, many of whom will cross many time zones. Most individuals experience unpleasant symptoms for several days after arriving at a destination in a new time zone. Because humans have only very recently achieved high-speed transportation, our circadian systems have not been shaped by evolution to entrain quickly to shifted light–dark cycles. As a result, people experience daytime sleepiness, fatigue, impaired alertness, and difficulty initiating and maintaining sleep for several days after crossing time zones.

Accident Risk

Circadian lows in alertness and performance increase the risk of human-error-related accidents. Motor vehicle accidents caused by drivers falling asleep at

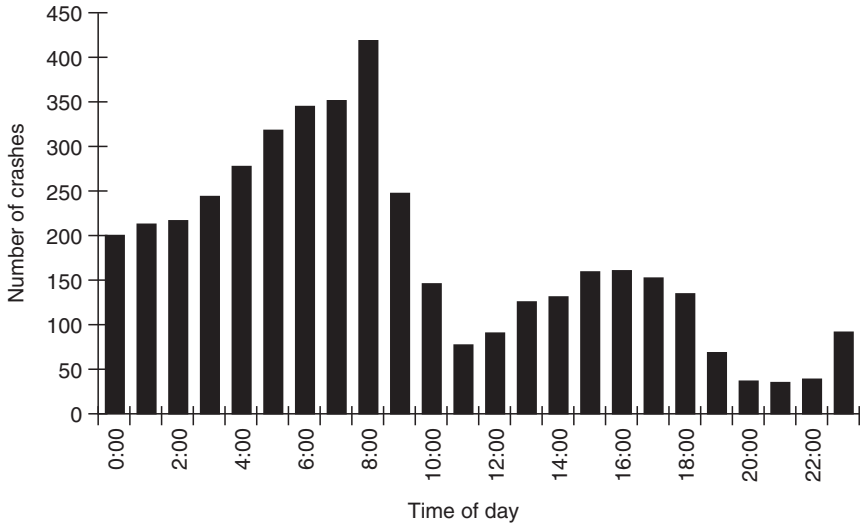

Figure 2 Frequency histogram of time of occurrence during the day of crashes in which the driver was judged to be asleep but not intoxicated. Data for years 1990–92, inclusive. Reprinted from Pack AM, Rodgman E, Cucchiara A, Dinges DF, and Schwab CW (1995) Characteristics of crashes attributed to the driver having fallen asleep. *Accident, Analysis and Prevention* 27: 769–775, with permission from Elsevier.

the wheel show a circadian rhythm quite similar to laboratory-based findings (see **Figure 2**). Fall-asleep accidents increase throughout the night as sustained attention decreases. Such accidents peak around 8:00 a.m., when vigilant performance is at its lowest, then begin to fall as the circadian system increases alertness and arousal.

Conclusion

Despite the difficulties inherent in identifying circadian influences on performance at various levels of alertness, a large body of research employing diverse methodology has clearly dissociated the effects of oscillating circadian rhythms from homeostatic influences on performance and mood. While controlling for factors that mask circadian rhythms is important for scientific and theoretical purposes, it is important to remember that these factors are also an integral part of the regulation of neurobehavioral functions that allow for adaptation to rhythmic changes in the environment.

Understanding circadian rhythmicity in performance and affective processes is important for understanding circadian disorders (such as shift-work- and jet-lag-induced sleep problems) and for correctly anticipating performance lapses and fatigue-related impairments (e.g., drowsy driving). When sleep is chronically restricted, both homeostatic and circadian influences must be understood in order to correctly predict levels of performance impairment that result from the interaction of these two systems.

See also: Circadian Systems: Evolution; Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Electroencephalography (EEG); Metabolic Syndrome and Sleep; Psychophysics of Attention; Shift Work and Circadian Rhythms; Sleep and Waking in Drosophila.

Further Reading

Boivin DB, Czeisler CA, Dijk DJ, et al. (1997) Complex interaction of the sleep–wake cycle and circadian phase modulates mood in healthy subjects. *Archives of General Psychiatry* 54: 145–152.

Cajochen C, Zeitzer JM, Czeisler CA, and Dijk DJ (2000) Dose–response relationship for light intensity and ocular and electroencephalographic correlates of human alertness. *Behavioral Brain Research* 115: 75–83.

Czeisler CA, Buxton OM, and Khalsa SB (2005) The human circadian timing system and sleep–wake regulation. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 375–394. Philadelphia: W.B. Saunders.

Czeisler CA, Duffy JF, Shanahan TL, et al. (1999) Stability, precision, and near-24-hour period of the human circadian pacemaker. *Science* 284: 2177–2181.

Dijk DJ and Von Schantz M (2005) Timing and consolidation of human sleep, wakefulness, and performance by a symphony of oscillators. *Journal of Biological Rhythms* 20: 279–290.

Monk TH (ed.) (1991) *Sleep, Sleepiness, and Performance*. Chichester: Wiley.

Phipps-Nelson J, Redman JR, Dijk DJ, and Rajaratnam SM (2003) Daytime exposure to bright light, as compared to dim light, decreases sleepiness and improves psychomotor vigilance performance. *Sleep* 26: 695–700.

Van Dongen HP and Dinges DF (2005) Sleep, circadian rhythms, and psychomotor vigilance. *Clinics in Sports Medicine* 24: 237–249.

Van Dongen HP and Dinges DF (2005) Circadian rhythms in sleepiness, alertness, and performance. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 435–443. Philadelphia: W.B. Saunders.

Wright KP Jr., Hull JT, and Czeisler CA (2002) Relationship between alertness, performance, and body temperature in humans. *American Journal of Physiology-Regulatory Integrative and Comparative Physiology* 283: 1370–1377.

Circadian Rhythms: Influence of Light in Humans

S W Lockley, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Circadian Organization in Humans

Many aspects of human physiology and behavior are dominated by 24 h rhythms that have a major impact on our health and well-being. For example, sleep–wake cycles, alertness and performance patterns, core body temperature rhythms, and the production of hormones such as melatonin and cortisol are all regulated by an endogenous near-24 h oscillator in the suprachiasmatic nuclei (SCN) of the anterior hypothalamus. The cells in these nuclei spontaneously generate rhythms with a period close to, but not exactly, 24 h, and in order for the circadian pacemaker to ensure that physiology and behavior are appropriately timed to anticipate events in the outside world, environmental time cues must be able to reset this internal clock. The major environmental time cue that resets these rhythms in mammals is the 24 h light–dark cycle generated by the earth's axial rotation. Light information is captured exclusively by the eyes using specialized retinal photoreceptors and transduced directly to the SCN via a dedicated neural pathway, the retinohypothalamic tract (RHT). Each day the light–dark cycle resets the internal clock, which in turn synchronizes the physiology and behavior controlled by the clock.

The major biochemical correlate of the light–dark cycle is provided by the pineal melatonin rhythm. Under normal light–dark conditions, melatonin is produced only during the night and provides an internal representation of the environmental photoperiod, specifically night length. The synthesis and timing of melatonin production requires an afferent signal from the SCN which projects to the pineal gland via the paraventricular nucleus and the superior cervical ganglion. Ablation of this pathway, as occurs in some people due to spinal damage at the upper cervical level, completely abolishes melatonin production, although other circadian rhythms not requiring this projection, such as the cortisol, temperature, and sleep–wake rhythms, persist. Light exposure during the night also inhibits melatonin production acutely and provides an indirect assessment of light input to the SCN via the RHT. Given the close temporal relationship between the SCN and melatonin production, the melatonin rhythm is often used as a marker of circadian phase and the melatonin suppression response as a proxy for RHT–SCN integrity and sensitivity.

The time taken to complete one cycle of a circadian rhythm is defined as the period (τ). A fixed point within a single cycle is termed the phase (ϕ) and is usually measured from the peak (maximum) or trough (minimum) of the rhythm. The circadian period is therefore defined as the duration between one phase and the next. The amplitude of a rhythm is defined as half of the maximum to minimum oscillation of the rhythm.

Circadian Photoreception in Humans

The Effect of Visual Impairment on Circadian Rhythms

The impact of different types of visual impairment on the circadian system has been used as an indirect method to assess the role of light on circadian rhythms and to determine the photoreceptor systems mediating such responses to light. First, these studies have shown that eyes are required for circadian photoreception: individuals without eyes, either through bilateral enucleation or as a result of developmental disorders, are unable to entrain their circadian pacemaker to the 24 h light–dark cycle. Similarly, the majority of totally blind individuals who retain their eyes but cannot consciously perceive light also exhibit circadian rhythms that are not entrained to the 24 h light–dark cycle. As a result of light information failing to reach the SCN to synchronize the clock and its outputs, the pacemaker reverts to its endogenous non-24 h period. Consequently, the timing of physiology and behavior that is controlled by the circadian system, such as the sleep–wake cycle, alertness and performance patterns, the core body temperature rhythm, and melatonin and cortisol production, becomes desynchronized from the 24 h day (Figure 1). These non-24 h rhythms are remarkably stable, ranging from 23.7 to 25.1 h across individuals, and can persist for decades. The resultant clinical disorder is termed 'non-24 h sleep–wake disorder' and is characterized by episodes of good sleep, followed by episodes of poor nighttime sleep and excessive daytime nappitng, followed by good sleep and so on, as the non-24 h circadian pacemaker cycles in and out of phase with the 24 h social day (Figure 1).

Disorders of the visual system do not always attenuate the circadian effects of light, however, demonstrating a functional separation of the visual and circadian photoreception systems. The majority of legally blind individuals who retain some degree of light perception, even with very little usable vision in some cases, have normally entrained circadian rhythms. Primary loss of the central or peripheral visual fields is also not

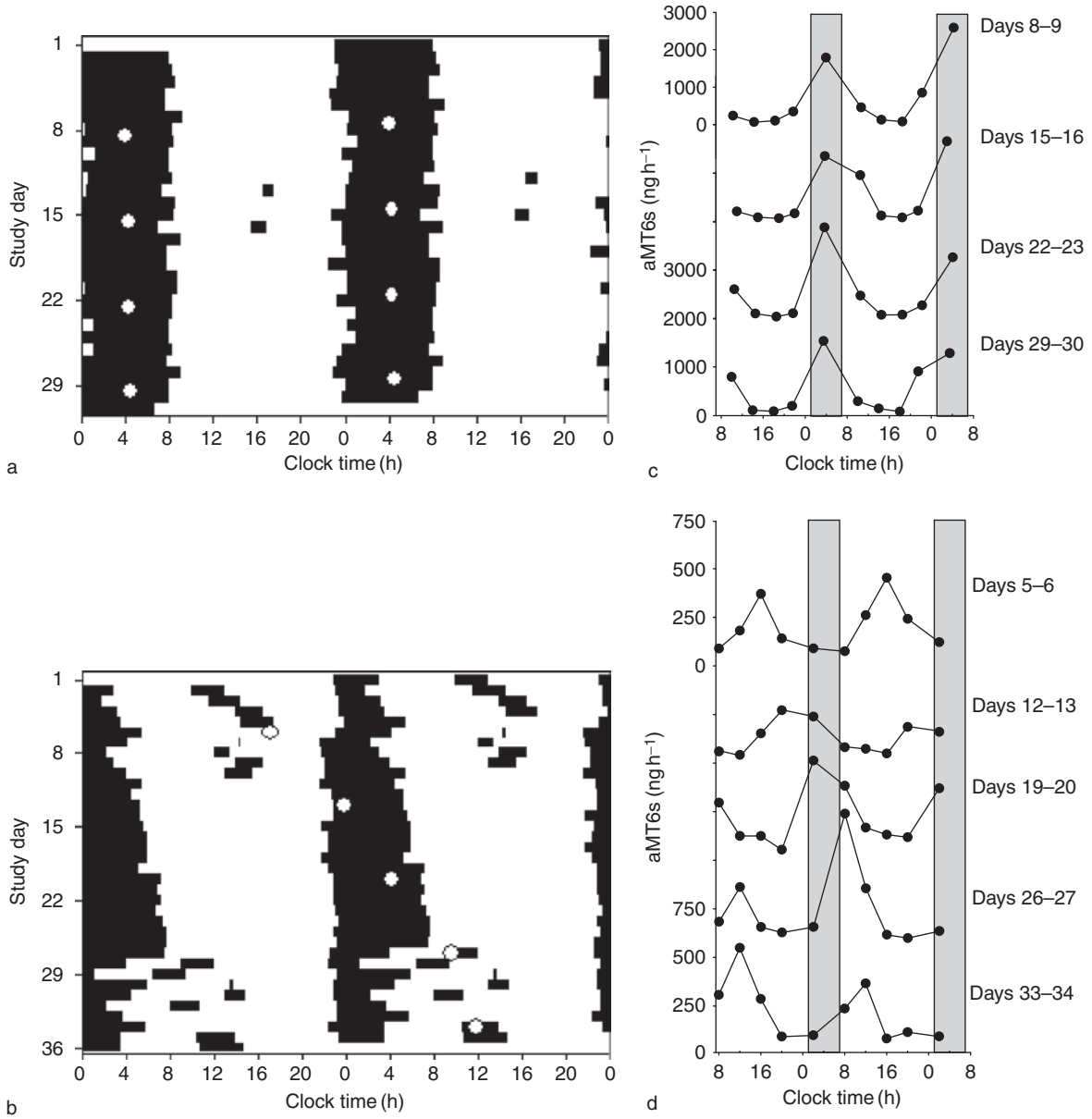

Figure 1 Circadian rhythms of sleep and melatonin in two visually impaired subjects. (a, b) subjects completed daily sleep and nap diaries for 4 or 5 weeks and their sleep times (solid squares) are double-plotted according to time of day (abscissa) and study day (ordinate). Subjects also collected sequential 4 to 8 hourly urine samples for 48 h per week, and the peak time of urinary 6-sulfatoxymelatonin (aMT6s) production each week (open circles) is plotted. (c, d) the weekly rhythms of urinary aMT6s production (solid circles) that correspond to the peak times shown and the gray bars represent the normal range of peak production for aMT6s (01.00–07.00 h). Panels (a) and (c) show data from a visually impaired female aged 35 years with a visual acuity to detect 'hand movements only' (subject 21). Both the sleep–wake cycle and aMT6s rhythm remain entrained to 24 h (a) with normal night-time production of melatonin (c) and such rhythms would be typical of most normally sighted subjects. Note that her visual impairment, while severely disrupting the visual system, does not affect the circadian photoreception system, illustrating the functional separation of the visual and circadian photoreceptor systems. The subject in panels (b) and (d) is a 66-year-old totally blind man with one eye and no conscious light perception (subject 20). He exhibits a sleep–wake cycle and melatonin rhythm that are not entrained to 24 h (b). The aMT6s rhythm exhibits a non-24 h period ($\tau = 24.68$ h) which clearly delays from week to week and is not entrained to the night (d). Closer observation of the sleep patterns shows a cyclic sleep disorder in this subject in which sleep occurs normally at night with no day time naps when the endogenous aMT6s rhythm coincides with the night but becomes very disturbed at night and is accompanied by many daytime naps when the aMT6s rhythm does not occur at night. The timing of the naps follows the timing of the aMT6s rhythm, which represents the internal 'biological' night and the time when the circadian system promotes sleep. The subject tries to remain on a normal social day, this is, sleeping at night and remaining awake in the day, but cannot sleep well when the circadian system in not entrained to the light–dark cycle. This example of non-24 h sleep–wake disorder is typical of most totally blind subjects and highlights the disturbance caused by a failure of light–dark information to reach the hypothalamic circadian pacemaker on our ability to sleep and wake on a 24 h pattern. Adapted from Lockley SW (1997) *Sleep, Melatonin and Other Circadian Rhythms in the Blind*. PhD thesis, University of Surrey, Guildford, UK.

associated with circadian rhythm disorders, suggesting that the circadian photoreceptor system is less sensitive to the spatial distribution and intensity of light than vision. Color blindness, a more specific lesion of the three-cone photopic visual system, also does not attenuate circadian responses to light, as measured by acute suppression of pineal melatonin. Finally, and most definitively, it has been demonstrated that some totally blind people retain normal circadian phase shifting and melatonin suppression responses to light even in the absence of any normal visual response, as assessed by conscious ability to detect light, visually evoked potentials, or electroretinogram (Figure 2). These individuals exhibit normally entrained 24 h rhythms under real-world conditions and do not report sleep disorders. The effects of light in these subjects are mediated exclusively via the eyes because when their eyes are covered, the effects of light are inhibited. Thus, circadian photoreception can be maintained in humans in the absence of a functional visual system, as has also been shown in transgenic and developmental rodent models.

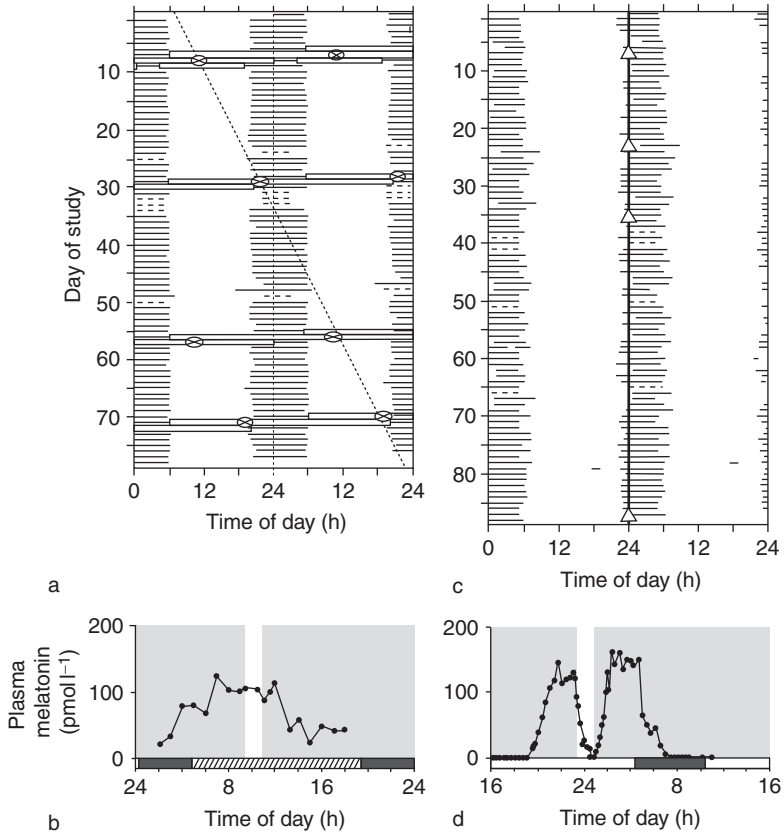

Figure 2 Circadian photoreception in a totally blind subject. (a, c) Subjects completed daily sleep and nap diaries for approximately 11 or 12 weeks and their sleep times (solid lines) are double plotted according to convention in Figure 1. Subjects also had their circadian phase measured at regular intervals from either core body temperature minimum (a, circles) or plasma melatonin onset (c, triangles). (b, d) The results of a melatonin suppression test during which the subjects' eyes were exposed to bright white light (6000–13 700 lux) for 90–100 min during plasma melatonin production. Panels (a) and (b) show the sleep–wake pattern and melatonin suppression response for a 70-year-old blind man with congenital glaucoma. He retained both eyes but reported no conscious light perception. Electroretinogram (ERG) and visually evoked potential (VEP) responses were not detectable. As shown in (a), the subject's circadian system was not entrained to the light–dark cycle and the core body temperature rhythm exhibited a non-24 h period. Consistent with the lack of entrainment by light, light exposure did not have any effect on the melatonin rhythm, confirming that the retina–SCN–pineal pathway was not functional in this patient. Panels (c) and (d) show the sleep–wake pattern and melatonin suppression results for a 21-year-old woman with Leber's congenital amaurosis, a type of retinal dystrophy. The ERG was undetectable, but an abnormal VEP was recorded. As shown in (c), the sleep–wake cycle was normal with stable entrainment to 24 h. The melatonin suppression response was clearly positive, with an immediate suppression of melatonin at lights on (open bars) which ceased once the bright lights were turned off. Despite the subject's lack of conscious light perception, light was still able to stimulate the retina–SCN–pineal pathway; covering the eyes prevented the response. The normal sleep–wake cycle, coupled with the positive melatonin suppression test, indicated that light was able to entrain the circadian system of this totally blind women. Note that a similar sleep–wake pattern accompanied by a negative melatonin suppression test would indicate that (1) the subject was entrained to 24 h by nonphotic time cues or (2) the subject's endogenous circadian period was exactly 24 h. Reproduced from Czeisler CA, Shanahan TL, Klerman EK, et al. (1995) Suppression of melatonin secretion in some blind patients by exposure to bright light. *New England Journal of Medicine* 332: 6–11.

Spectral Sensitivity of Circadian Photoreception

Assessment of the spectral sensitivity of circadian responses has been used as an indirect method to determine the photoreceptor system mediating these effects of light. Using classical photobiology techniques, action spectra have been generated to describe the spectral sensitivity of the melatonin suppression response. These action spectra predict a peak sensitivity in the short-wavelength (blue) range (λ_{max} ~460 nm) that does not match the action spectra for human scotopic (rod) or photopic color (cone) vision or the absorption spectra for the individual red, green, or blue cones (Figure 3). Although full action spectra are not available for circadian phase shifting, comparisons of equal photon density exposures of monochromatic 460 and 555 nm light have shown that the circadian system is also more sensitive to short-wavelength 460 nm light, confirming that the three-cone photopic photoreceptor system (which has a peak sensitivity at 555 nm) is not the primary photoreceptor system mediating the resetting effects of light (Figure 4). These findings are consistent with the concept of a non-rod, non-cone photoreceptor system in the eye. In fact, studies in rodents and primates detailed elsewhere in this encyclopedia have discovered a novel opsin in the eye, melanopsin, which is also present in the human eye. Melanopsin is contained in intrinsically photosensitive retinal ganglion cells which project to the SCN, and these cells have a peak spectral sensitivity in the short-wavelength range (~480 nm), making melanopsin a prime candidate for mediating the circadian effects of light.

Extraocular Circadian Photoreception

Photoreceptors external to the eye that detect light for circadian rhythms are common in nonmammalian vertebrates, such as the pineal gland in birds and the parietal eye and pineal gland in reptiles. There is no evidence, however, for the presence of extraocular circadian photoreceptors in adult mammals despite multiple investigations. The concept of extraocular light sensitivity for circadian response in humans, specifically via exposure to the popliteal region, was widely disseminated following a study published in the late 1990s. These preliminary findings, however, have not stood up to rigorous reevaluation and, coupled with the lack of entrainment observed in enucleated blind subjects as described previously, the current consensus is that human circadian photoreception is mediated exclusively via the eyes.

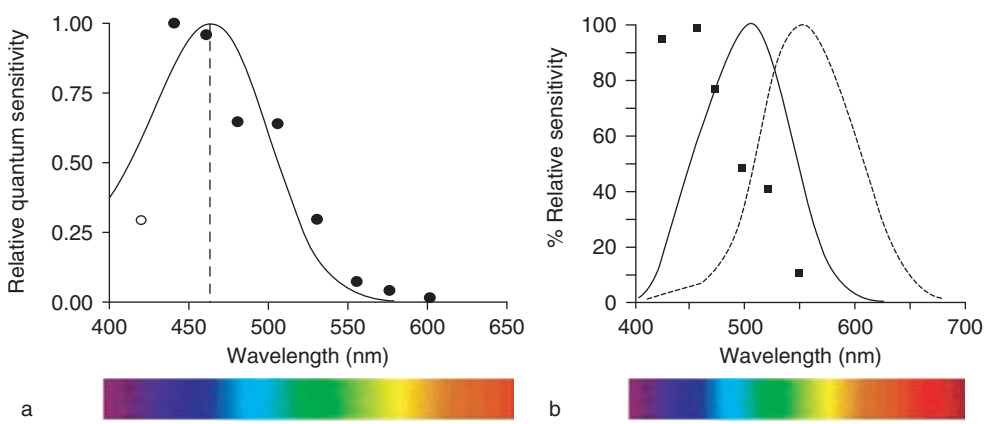

Figure 3 Spectral sensitivity of the light-induced melatonin suppression response. Two action spectra are shown from separate experiments examining the spectral (color) sensitivity of light-induced melatonin suppression. The relative melatonin suppression response is plotted against the wavelength of light exposure. The suppression data were generated from many hundreds of experiments in which subjects were exposed to either 90 min ($n = 72$) (a) or 30 min ($n = 22$) (b) of monochromatic visible light (range, 420–600 nm) of varying intensities during the night after administration of a pupil dilator. Each point in the action spectra represents the half-saturation constant of the dose–response function of one wavelength, normalized to the maximum response. The solid curve in (a) shows the best-fit template for a vitamin A_1 retinaldehyde photopigment and predicts a maximal spectral absorbance (λ_{max}) of 464 nm (range, 446–477). Similar analysis of the data in (b) also predicts a peak sensitivity in the short-wavelength (blue) part of the spectrum with $\lambda_{max} = 459$ nm (range, 457–462). Panel (b) also shows the absorbance spectra for scotopic (rod) vision (solid line) and photopic three-cone color vision (dotted line) and demonstrates that the spectral sensitivity of the melatonin suppression response is blue-shifted relative to conventional visual systems. (a) Adapted from Brainard GC, Hanifin JP, Greeson JM, et al. (2001) Action spectrum for melatonin regulation in humans: Evidence for a novel circadian photoreceptor. *Journal of Neuroscience* 21(16): 6405–6412. (b) Thapan K, Arendt J, and Skene DJ (2001) An action spectrum for melatonin suppression: Evidence for a novel non-rod, non-cone photoreceptor system in humans. *Journal of Physiology* 535(1): 261–267.

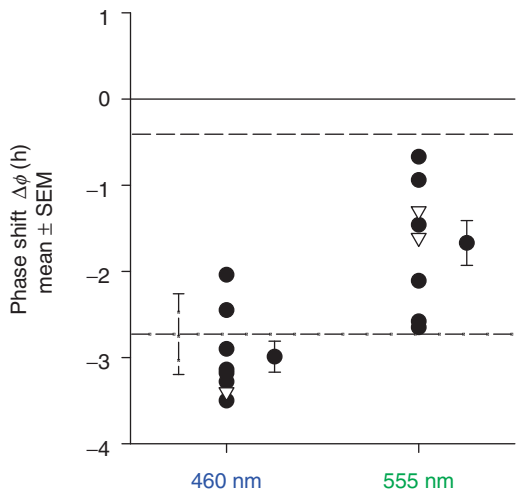

	460 nm	555 nm
Photons cm^{-2} s^{-1}	2.8×10^{13}	2.8×10^{13}
μW cm^{-2}	12.1	10.0
Photopic lux	5.0	68.1
Scotopic lux	116.6	68.7

Figure 4 Spectral sensitivity of the light-induced circadian phase shifting response. Two groups of young healthy subjects ($n = 8$ per group) were exposed to an equal photon density of 460 nm (blue) or 555 nm (green) monochromatic light for 6.5 h during the night following administration of a pupil dilator on a background of <2 lux. The phase shift of the circadian system was assessed using the endogenous melatonin rhythm, measured before and after the experimental light exposure using a constant routine procedure. Both light exposures caused a phase delay (negative sign) of the individuals' plasma (solid circles) or salivary (upside-down triangles) melatonin rhythm. On average, exposure to 460 nm light caused nearly twice the phase shifting response compared to exposure to an equal photon density of 555 nm light (mean ± SEM, -2.98 ± 0.50 vs. -1.67 ± 0.73 h, respectively). The relative photopic lux of the exposures, as would be measured using a standard lux meter, does not correspond to the biologic effects of light, indicating that the spectral sensitivity of the meters ($\lambda_{max} = 555$ nm) is not appropriate for quantifying light for circadian responses. The scotopic lux values, although appearing to correlate with the biological responses, are not applicable because the levels of light used saturate rod responses. Adapted from Lockley SW, Brainard GC, and Czeisler CA (2003) High sensitivity of the human circadian melatonin rhythm to resetting by short wavelength light. *Journal of Clinical Endocrinology and Metabolism* 88(9): 4502–4505.

Measuring Circadian Rhythms in Humans

Circadian Phase and Amplitude

In order to assess the effects of light on the human circadian pacemaker, an accurate measurement of circadian phase and amplitude is required. Common parameters used to assess circadian phase include melatonin and its metabolites or cortisol in plasma, saliva, or urine, and core body temperature. These parameters, however, are susceptible to direct environmental influences that alter their pattern and 'mask' the endogenous circadian component of their rhythms. For example, core body temperature can be directly increased by mild activity or exercise, eating, or taking a shower. Cortisol can be directly affected by stress and melatonin is acutely suppressed by light. A technique is therefore required that removes or minimizes these influences or, if this is not possible, distributes them evenly across all circadian phases so that any effect of the masking stimulus is effectively averaged out. The gold standard method used to achieve these conditions is the constant routine (CR) procedure, developed by Mills and colleagues in the 1970s and later refined by Czeisler. This method requires subjects to be assessed for at least 24 h (usually 30 h or more) in constant conditions; subjects remain awake during continuous semirecumbent bed rest in constant dim light (<15 lux or lower) with no activity and regular (e.g., hourly) identical meals and fluids. Using these techniques, it has been possible to make a systematic review of how the circadian system is affected by a range of properties of light as outlined later.

Circadian Period

Measurement of circadian period requires a longer assessment than can be achieved during a CR procedure. Although a CR can sometimes last two or, rarely, three circadian cycles (48–72 h), the small number of cycles is insufficient to assess circadian period accurately. A range of techniques have been used to measure period in the laboratory, but the forced desynchrony protocol, pioneered by Kleitman in the 1930s and later extended by Czeisler and colleagues, represents the most well-developed method for the accurate and reliable measurement of circadian period. The technique requires that subjects live on a scheduled day length (called a T cycle) that is outside the range of entrainment of the circadian pacemaker; that is, a T cycle so different from 24 h that the circadian pacemaker cannot entrain to it and consequently reverts to its intrinsic period. Examples of forced desynchrony T cycles include 20 min, 90 min, 11 h, 20 h, 28 h, 30 h, and 42.85 h, although the very short T cycles cannot be sustained for more than a few days. Under these conditions, the circadian pacemaker of healthy adults exhibits an average period of ~24.2 h with a range from 23.5 to 24.7 h. Similar non-24 h periods have been observed in subjects living in near-darkness during cave experiments or under dim light in the laboratory.

Resetting the Human Circadian Pacemaker with Light

Circadian Entrainment

In order for the circadian pacemaker to remain entrained to environmental time, a daily phase shift is required that opposes the daily 'drift' in phase as determined by the endogenous period of the pacemaker. For example, if the endogenous period of an individual is 24.5 h, then a phase shift of 0.5 h per day is required to maintain entrainment. As this shift requires a shortening of the period, causing the circadian cycle to be completed in a shorter time, the direction of shift is termed an advance shift. If the period (τ) = 23.8 h, then a phase delay shift of 0.2 h per day is required, lengthening the period and delaying the time taken to complete one circadian cycle. Depending on the circadian phase of exposure, light can both phase advance and phase delay the clock; the magnitude of the phase shift depends on the intensity and number of exposures. The direction and magnitude of phase shifts induced by a stimulus are defined by a phase response curve (PRC), and the PRCs for light are reviewed in detail later. Although light is the most potent environmental stimulus in humans, other nonphotic stimuli, such as scheduled exercise, sleep, food intake, and some drugs (e.g., melatonin), can shift the timing of the clock to a lesser extent and have their own PRCs. In some cases, these nonphotic stimuli are strong enough to entrain the circadian pacemaker, as observed in the minority of totally blind people who are not sensitive to light but remain synchronized with the 24 h day.

As the strength of the light stimulus determines the maximum shift per day that can be achieved, it also determines the 'day' length to which the circadian pacemaker can be entrained – also termed the limits or range of entrainment. Under a dim light–dark cycle (<8 lux), the circadian pacemaker can remain entrained to a 24 h day but the light is not bright enough to achieve the daily phase shift required to entrain to a 23.5 or 24.6 h 'day.' If the intensity of the light during the day is increased, then the limits of entrainment expand. Model predictions suggest that entrainment could occur to day lengths ranging from 22 to 27 h depending on the endogenous circadian period and the intensity of the light.

A practical example of the consequences of the limits of entrainment can be observed in US submariners, who are routinely scheduled to an 18 h 'day.' The combination of the short day length and the relatively dim lighting conditions on board prevents entrainment to the imposed schedule, causing the circadian clock to run at its endogenous non-24 h period, on average ~24.3 h. The resultant desychronization between the circadian system and the sleep–wake schedule is likely to cause decrements in alertness and performance when the submariners are on duty at an inappropriate circadian phase and increase the risk of a fatigue-related accident. Similar problems occur in night-shift workers who also often work during their biological night and have an increased risk of accidents and injuries during this time, as reviewed elsewhere in this encyclopedia.

Properties of Light and Circadian Resetting

The circadian photoreception system is not designed to detect light with the same degree of sensitivity as the visual system but, rather, represents a relatively insensitive irradiance detector. This irradiance detection system predates vision in evolutionary terms and both the anatomical location of the photoreceptors, distributed sparsely in a network across the eye, and the properties of the system, particularly the ability to integrate light information over a long duration of time, reflect this specific role.

The circadian resetting effects of light are determined by a range of factors, including the timing, intensity, duration, pattern, wavelength and history of exposure. These factors have largely been studied separately in order to confirm their individual contribution to the resetting effects of light, although their interactions are beginning to be examined. In real-world conditions, of course, many of these factors will apply simultaneously to the net phase resetting effects of light.

Timing One of the primary factors determining the phase resetting effects of light is the timing of the light exposure. Light can either phase advance or phase delay the circadian system depending on the timing of exposure. Under normal conditions, light exposure in the late day/early night causes a phase delay of the pacemaker, whereas light exposure in the late night/early day will phase advance the clock. The relationship between the timing of a stimulus and the direction and magnitude of the resultant shift is described in a PRC (Figure 5). Two types of light PRCs have been described across many organisms, including humans – a low-amplitude PRC with maximum shifts of several hours and a high-amplitude PRC with maximal shifts of 12 h. Winfree classified these PRCs as type 1 (weak) and type 0 (strong), respectively, according to the shape of their resetting contours (phase transition curves) (Figure 5). In order to achieve type 0 resetting, the light must be of appropriate strength and duration (e.g., 10 000 lux for 5 h) and be timed at a critical phase (approximately equivalent to the phase

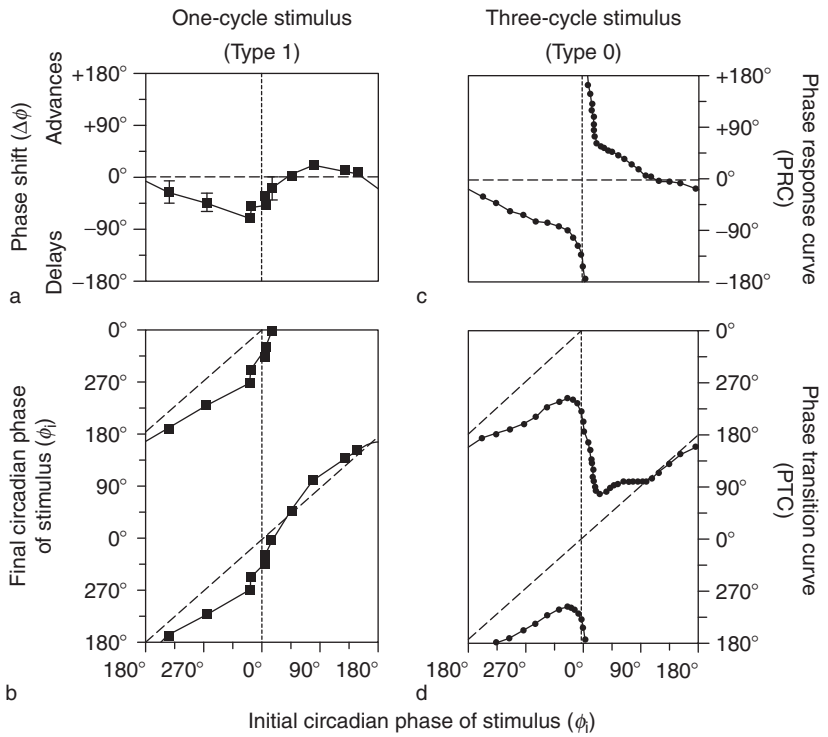

Figure 5 Type 1 and type 0 phase response curves to light. (a, c) Smoothed phase response curve (PRC) following exposure to a single stimulus (a) or three stimuli over consecutive days (c) on a background of <15 lux. The magnitude and direction of the phase shifts are plotted against the initial phase of the stimulus. The magnitude of the shift is expressed in degrees (± SEM); if one circadian cycle is 360°, then a shift of 90° represents one-fourth of the circadian cycle, equivalent to a 6-h phase shift under normally entrained conditions (0.25 × 24 h). Advance shifts are positive on the ordinate and delay shifts are negative. Phase shifts and the timing of the light exposure were assessed from the timing of the core body temperature minimum (0°) under constant routine conditions before and after light exposure. (a) Following a single exposure to bright light (7000–12 000 lux for 5.24 ± 0.22 h; range, 4.0–6.2), a low-amplitude PRC is observed, with maximum delay shifts of −71° and advance shifts of +21°. The asymmetry is caused by the 'drift' between the pre- and postexposure constant routines due to endogenous circadian period which on average causes a net delay ($\tau = 24.2$ h). (c) Following exposure to three light exposures (7000–12 000 lux for 5 h each), the resultant PRC has a much larger amplitude, with maximum advance and delay phase shift of +180° and −180° (equivalent to +12 and −12 h), respectively. These large phase shifts only occur at a particularly sensitive phase, close to the core body temperature minimum (0°). (b, d) Smoothed phase transition curves (PTC) for the corresponding PRC data, with the final circadian phase of the stimulus plotted against the initial phase of the stimulus. The slope of the resultant plot indicates whether the phase resetting response is type 1 (slope ∼1; b) or type 0 (slope ∼0; d). Reproduced from Jewett ME, Kronauer RE, and Czeisler CA (1994) Phase-amplitude resetting of the human circadian pacemaker via bright light: A further analysis. *Journal of Biological Rhythms* 9(3–4): 295–314.

of the core body temperature minimum) such that the amplitude of the circadian oscillation is suppressed. Once the oscillation is suppressed, a phase shift of up to 12 h can be achieved with similar multiple light exposures over the next 2 or 3 days. An analogy for this type of response would be to consider the amplitude of the oscillation with respect to the earth's rotation at the Poles. When the amplitude is nearly zero (i.e., very close to the Pole), a short step in one direction could achieve an immediate phase shift of up to 12 h. When on the equator, however, the amplitude of the earth's rotation with respect to the Poles is very large, and only small phase and amplitude shifts can be achieved by stepping in any particular direction. These weak shifts are analogous to type 1 resetting in which shifts of only a few hours are achieved in response to a single light exposure (Figure 5). In practice, most phase shifts experienced by humans are type 1 shifts since very specialized conditions are required to achieve type 0 resetting.

In animal studies, some species have been shown to be insensitive to light during the middle of the subjective day, creating a so-called 'dead zone' in the PRC. Human PRCs lack a dead zone, as shown in Figure 5, illustrating that the circadian pacemaker is sensitive to the resetting effects of light at all times of day, although the sensitivity to light is lowest in the middle of the day.

The PRC provides an essential tool in calculating when to time light therapy for treatment of circadian rhythm sleep disorders or in anticipating the effects of altered light–dark exposure experienced during transmeridian travel or during shift work.

Intensity The intensity of light is a vital property when considering the resetting strength of the light exposure. There is a nonlinear relationship between the intensity of light and its circadian phase resetting effects such that exposure to relatively dim indoor room light (~100 lux) for 6.5 h at night can stimulate 50% of the maximum effect of an equal-duration exposure with 10- to 100-fold greater illuminance (Figure 6). Thus, the human circadian pacemaker is exquisitely sensitive to light, and light levels commonly experienced from artificial indoor illumination are able to induce considerable resetting effects on the clock.

Duration and pattern The duration of light exposure interacts nonlinearly with the magnitude of suppression such that shorter exposures induce greater effects than would be predicted by a simple linear relationship. When continuous exposures are compared to intermittent patterns, a nonlinear interaction is also observed. Compared to a single continuous 6.5 h bright light exposure (~10 000 lux) at night, an intermittent exposure pattern consisting of six 15 min bright light pulses separated by 60 min of very dim light (<1 lux) caused ~80% of the phase shifting response (3.0 vs. 2.3 h delay, respectively) despite representing only 23% of the total duration (Figure 7). Similar results are observed for multiple exposures. The ability of the circadian system to integrate light information over time presents potential advantages when developing light therapy programs because most of the potential benefit can be realized without requiring prolonged continuous exposure.

Wavelength As outlined previously, the spectral sensitivity of the circadian system is blue-shifted relative to the photopic and scotopic visual systems, indicating the presence of a novel non-rod, non-cone photoreceptor system (Figure 3). Equal photon density comparisons of 460 and 555 nm monochromatic light exposure at night show that 460 nm light is twice as effective at phase delaying the clock (Figure 4). Morning exposure to polychromatic blue light has also been shown to be at least equally as effective at

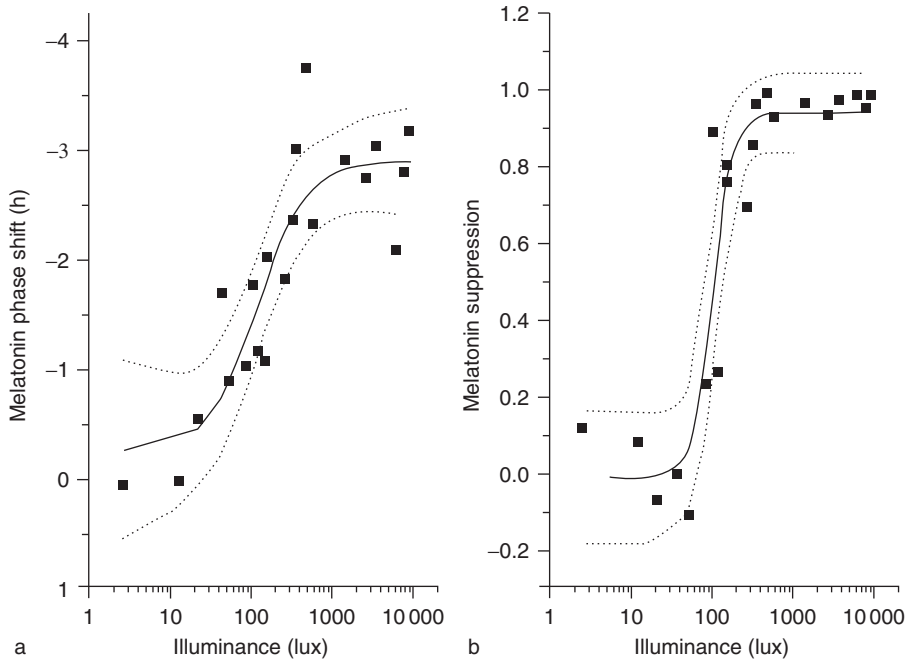

Figure 6 Illuminance–response curves for circadian phase resetting and melatonin suppression. Twenty-three young healthy subjects were exposed to white light (4100 K) for 6.5 h during the night. Each subject was exposed to a different intensity of light, ranging from 3 to 9100 lux (squares) on a background of <10 lux. The phase delay shift in the plasma melatonin rhythm was calculated from the difference in phase from the pre- and postexposure constant routines. Melatonin suppression was calculated from the difference in the area under the curve for the 4 h prior to peak melatonin level on the first constant routine and the corresponding 4 h during the light exposure. The resultant dose–response curves were fitted with a four-parameter logistic model using a nonlinear least squares analysis (solid line) and the 95% confidence intervals are shown (dotted lines). The model for the phase delay response predicts an inflection point of the curve (i.e., the half-saturation constant) at 119 lux and saturation of the response at ~550 lux (a). Similar results were obtained for melatonin suppression, with half-maximal response at 106 lux and saturation at a slightly lower illuminance of ~200 lux (b). These data show that relatively dim light exposure, equivalent to that experienced indoors, is capable of shifting the circadian pacemaker and inducing neuroendocrine responses. Reproduced from Zeitzer JM, Dijk D-J, Kronauer RE, Brown EN, and Czeisler CA (2000) Sensitivity of the human circadian pacemaker to nocturnal light: Melatonin phase resetting and suppression. *Journal of Physiology* 526(3): 695–702.

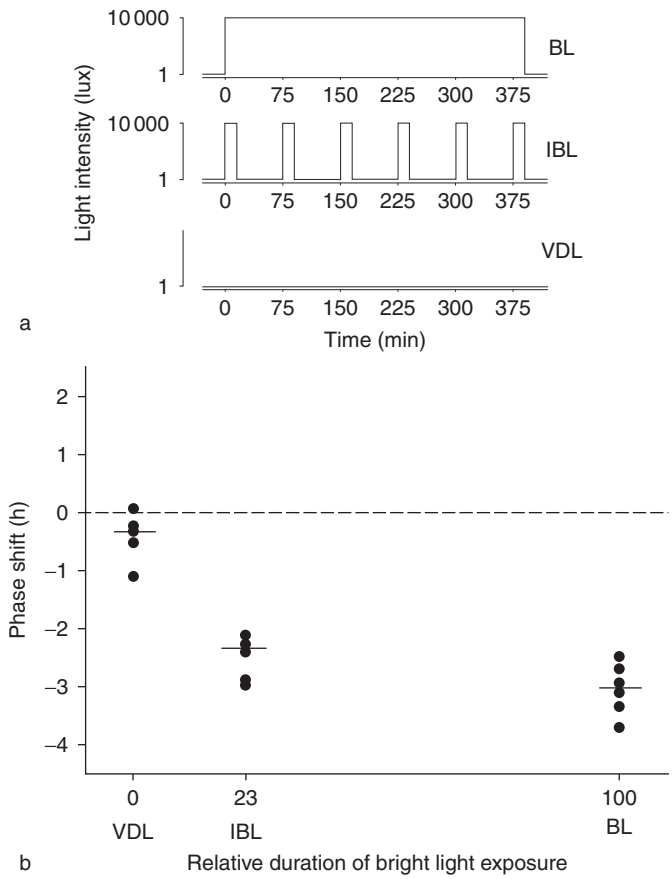

Figure 7 Effect of intermittent light exposure on circadian phase resetting. Twenty young healthy subjects were exposed to white light (4100 K) for 6.5 h during the night. (a) Subjects were randomly assigned to received continuous exposure to bright light (BL; ~9500 lux, $n = 6$), intermittent exposure to bright light (IBL; 6 × 15 min exposures to ~9500 lux each separated by 1 h, $n = 7$), or continuous very dim light (VDL; <1 lux in the angle of gaze, $n = 7$). (b) The phase delay shift was calculated as described for **Figure 6** (solid circles). Despite representing only 23% of the total light exposure duration, intermittent bright light caused ~70% of the phase shift compared to continuous bright light exposure. Reproduced from Gronfier C, Wright KP Jr., Kronauer RE, Jewett ME, and Czeisler CA (2004) Efficacy of a single sequence of intermittent bright light pulses for delaying circadian phase in humans. *American Journal of Physiology Endocrinology and Metabolism* 287: E174–E181.

phase advancing the circadian system as exposure to a 185-fold higher photon density of cool white light.

As a consequence of the finding that the spectral sensitivity of the circadian photoreception system is blue-shifted relative to the three-cone photopic system, a revision of the units used to measure illuminance is required. Currently, lux (1 lux = 1 lumen m^{-2}) is the unit of measurement for illuminance and estimates the amount of light stimulating the visual system. As the three-cone photopic visual system has a peak sensitivity (λ_{max}) of 555 nm, the lux measurements also assume the same spectral sensitivity and therefore so do most commercially available light meters. When measuring the amount of light stimulating a circadian response to light, however, this assumption is incorrect because the peak sensitivity of circadian photoreception is ~460–480 nm. Lux values are therefore inaccurate when quantifying the absolute amount of light stimulating the circadian photoreception system (**Figure 4**), although if the light source remains constant, lux can quantify relative changes in light amount (**Figure 6**). The development of appropriate units and light sensors adapted for circadian photoreception is currently ongoing.

Photic history The effect of a light stimulus on the circadian system is not simply dependent on the intensity of light but also on the relative change in intensity. The circadian photoreception system exhibits adaptation to prior light exposure in determining the response to a current stimulus. As shown in **Figure 8**, under laboratory conditions subjects kept in dim light (<2 lux) for 15 h prior to exposure to 200 lux for 6.5 h showed significantly greater melatonin suppression compared to when they were exposed to 200 lux for 15 h prior to the experimental 6.5 h exposure.

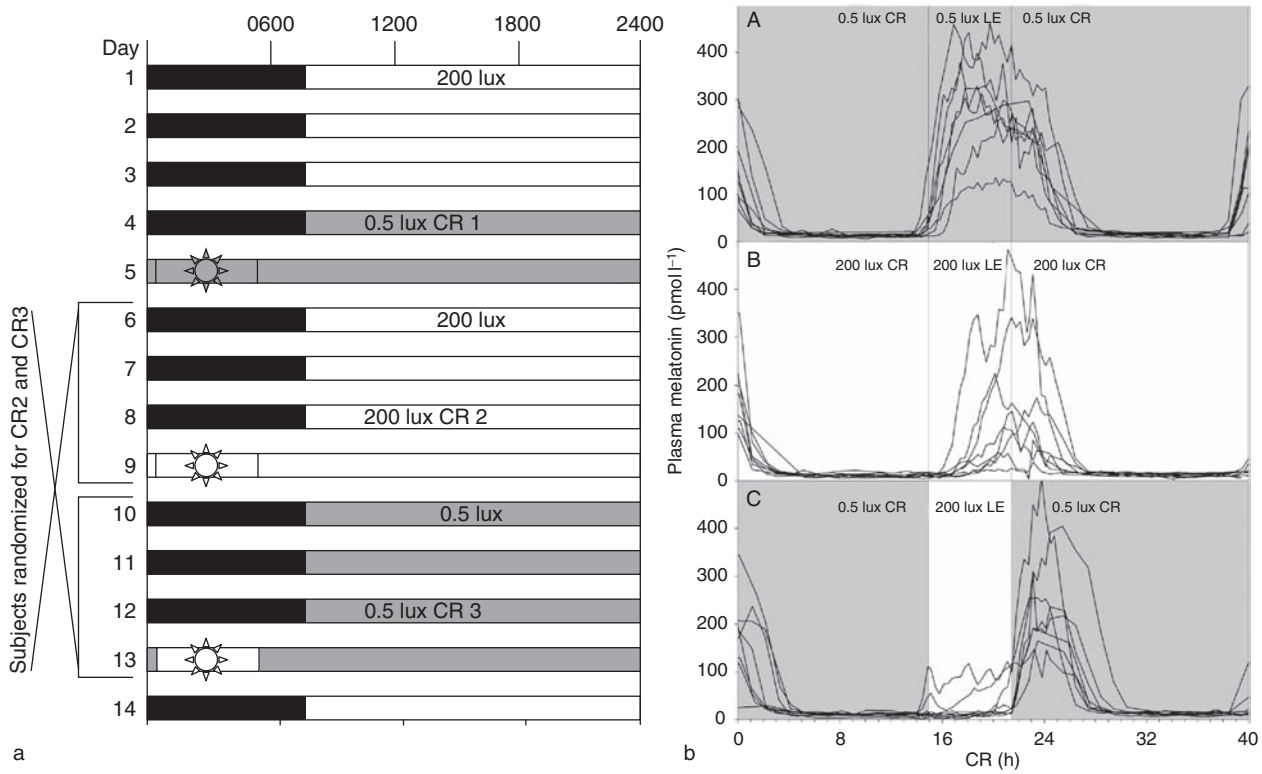

Figure 8 Effect of prior light exposure on the light-induced melatonin suppression response. Eight young healthy subjects were studied in a within-subjects design. (a) Subjects completed three 40 h constant routines (CR), each separated by three baseline days (16 h wake: 8 h sleep in darkness). On the first CR, all subjects were exposed to continuous dim light (<2 lux) after three baselines days at 200 lux. On the second and third CRs, subjects were randomized to receive a 6.5 h exposure to 200 lux (suns) after three baseline days of <2 lux or 200 lux, respectively. (b) The individual plasma melatonin profiles during each 40 h CR are shown for each condition (A–C). Melatonin suppression was significantly greater when 200 lux exposure was preceded by exposure to <2 lux (mean ± SEM, 85.7 ± 6.5%; B) compared to prior exposure to 200 lux (71.2 ± 7.1%; C). Reproduced from Smith KA, Schoen MW, and Czeisler CA (2004) Adaptation of human pineal melatonin suppression by recent photic history. *Journal of Clinical Endocrinology and Metabolism* 89(7): 3610–3614.

Furthermore, subjects routinely exposed to more bright light during the day for 1 week have a reduced melatonin suppression response than following exposure to dimmer light, and it has been suggested that people who routinely work outdoors are less sensitive to the melatonin-suppressing effects of light than those who routinely work indoors. At a cellular level, the melanopsin-containing intrinsically photosensitive retinal ganglion cells also demonstrate an adaptation process which may also be wavelength as well as intensity specific. Further work is required to define the time course of this adaptation response, although these initial studies suggest that the circadian photoreception system is able to integrate light information over hours if not days.

Other Neuroendocrine and Neurobehavioral Effects of Light

In addition to resetting the phase of the circadian pacemaker, light also stimulates other neuroendocrine and neurobehavioral responses. One major response, the inhibitory effects of light on pineal melatonin synthesis, has already been discussed. In addition, light exposure directly stimulates cortisol production in the morning, improves subjective ratings of alertness, improves psychomotor reaction time and lapses of attention, changes the frequency of electroencephalogram (EEG) brain waves that specifically indicate a more alert state, increases heart rate, alters thermoregulatory processes, and induces pupillary constriction. Light exposure has also been shown to stimulate circadian clock gene expression, specifically the expression of PER2 in buccal cells. These effects of light have been collectively termed 'nonvisual' or 'non-image forming' responses and are sometimes grouped under the term 'circadian photoreception' because much of the behavioral and neuroanatomical work that first identified these effects was focused on studies of circadian phase resetting.

Some of the properties of light outlined previously that affect phase shifting responses have also been demonstrated for other effects of light. For example, the acute arousing effects of light at night on alertness, performance, and EEG power are dose dependent

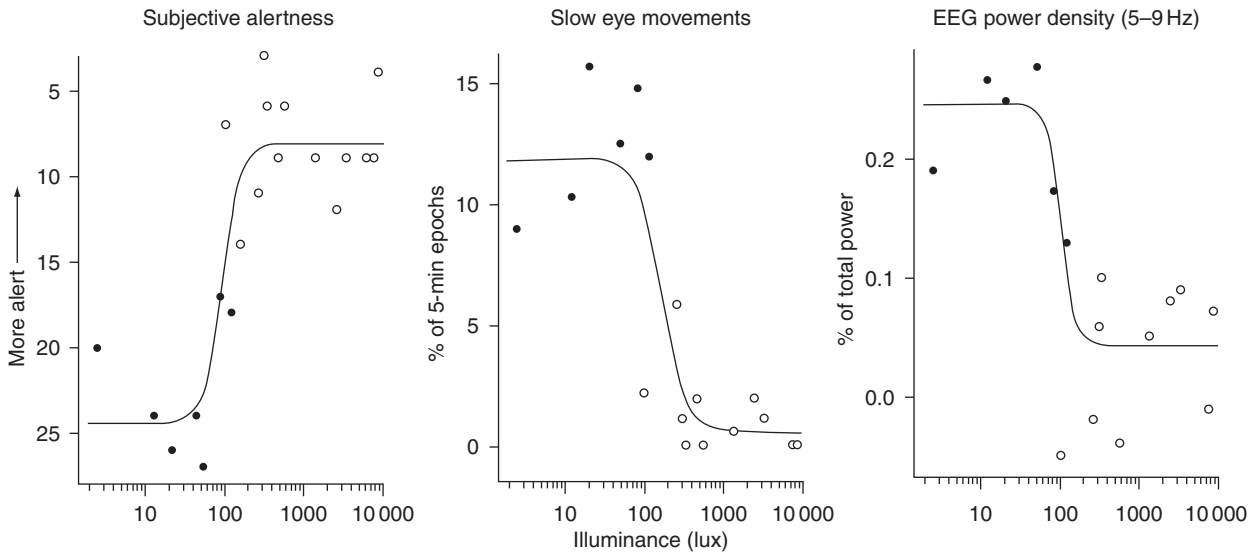

Figure 9 Illuminance–response curves for the acute alerting effects of light. In addition to phase shifting and melatonin suppression (Figure 6), the acute dose-dependent alerting effects of light were also examined in the same subjects. Before, during, and after the 6.5 h light exposure, subjects completed the nine-point Karolinska Sleepiness Scale (1 = very alert, 9 = very sleepy) every 30 min and had continuous electroencephalogram (EEG) and electooculogram (EOG) recordings. The summed alertness ratings during the entire light exposure are shown on the left; the number of 30 s EOG epochs containing at least one slow-rolling eye movement, an objective indicator of fatigue, are shown in the middle; and the percentage EEG power in the theta–alpha frequency range (5–9 Hz) during the last 90 min of light exposure are shown on the right. Open circles represent subjects with >50% suppression of melatonin during the light exposure, and closed circles identify those with <50% melatonin suppression. The logistic regression model fit to the individual data points is also shown for each parameter. Both the subjective and the objective markers of sleepiness exhibit a clear dose-dependent response to light, with increasing intensity causing a reduction in sleepiness, the number of slow eye movements, and theta–alpha EEG power. These reductions in sleepiness were also correlated with the degree of melatonin suppression, although the mechanism of how light induces alertness is not yet known. Reproduced from Cajochen C, Zeitzer JM, Czeisler CA, and Dijk D-J (2000) Dose–response relationship for light intensity and ocular and electroencephalographic correlates of human alertness. *Behavioural Brain Research* 115: 75–83.

(Figure 9) and wavelength dependent (Figure 10). One mechanism by which light could exert its alerting effect is via the suppression of melatonin, the timing of which closely parallels the endogenous rhythm in subjective sleepiness and, when taken exogenously, increases sleepiness. Nighttime light exposure both suppresses melatonin and induces a simultaneous increase in alertness. Furthermore, studies measuring light-induced changes in EEG correlates of arousal have shown that blue light (460 nm) is more effective than green light (555 nm) in activating high-alpha frequencies (9.5–10.5 Hz) which closely parallel the circadian rhythm of melatonin production. Notably, however, light exposure during the day still improves alertness and performance when melatonin is undetectable, suggesting that at least during the biological day, alertness is not mediated via suppression of melatonin. Further separation of day and night responses is also observed in other effects of light; for example, heart rate and temperature are elevated by white light exposure during the night but not the day. These data suggest that it may be possible that multiple light responses have differing spectral sensitivities to light, theoretically mediated through differences in the relative contributions of the short-wavelength sensitive melanopsin-driven system and the longer-wavelength sensitive photopic and scotopic visual photoreceptor systems.

Circadian Rhythm Disorders and Light

Shift work The most common circadian rhythm disorder is shift work sleep disorder, experienced by many of the estimated 15 million shift workers in the United States. Shift workers experience sleep problems, fatigue, poor performance and memory, and gastrointestinal problems and have increased long-term risk of cardiovascular disease, diabetes, and some types of cancer. The underlying source of these problems is the desychronization between the shift work schedule and the light–dark cycle. As illustrated in totally blind subjects, circadian desynchronization is caused by a failure of the internal circadian pacemaker to remain entrained with the environmental light–dark cycle. Shift workers' schedules (and therefore light–dark cycle) often change so rapidly that the circadian system cannot adapt, resulting in wake (work) and sleep occurring at an inappropriate circadian phase. Consequently, sleep duration and quality are reduced,

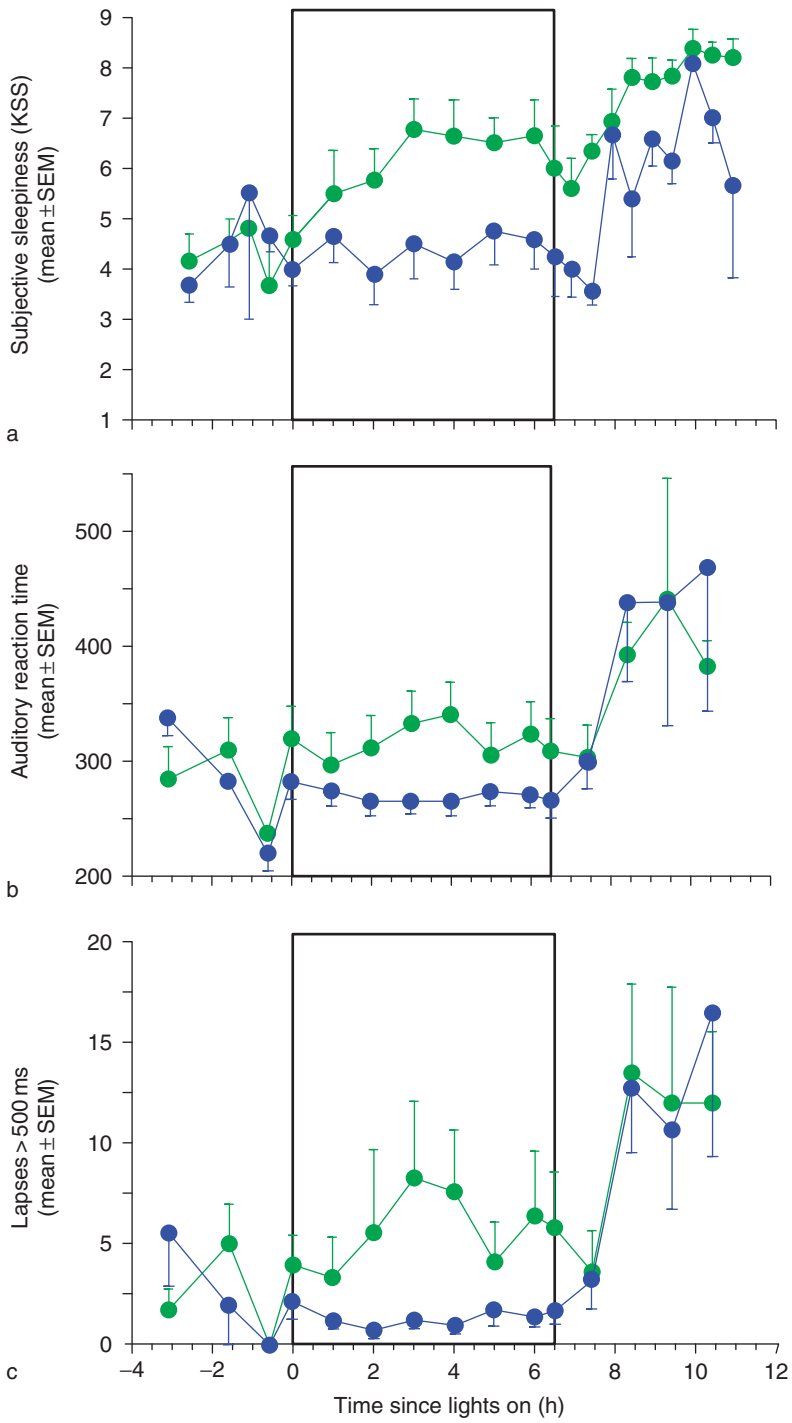

Figure 10 Spectral sensitivity of the acute alerting effects of light. Sixteen subjects were studied to test the spectral sensitivity of the photic phase resetting response (**Figure 4**) and the acute alerting effects of light. Subjects completed the nine-point Karolinska Sleepiness Scale (1 = very alert, 9 = very sleepy) and an auditory psychomotor vigilance test every 60 min before, during, and after exposure to 6.5 h of either 460 or 555 nm monochromatic light (open rectangles) in addition to continuous EEG recordings. A pupil dilator was administered 15 min prior to experimental light exposure. Subjects exposed to 460 nm (blue) light reported significantly lower sleepiness (a), exhibited a shorter auditory reaction time (b), and had fewer lapses (responses >500 ms) (c) compared to subjects exposed to an equal photon density of 555 nm (green) light. Note the rapid deterioration of alertness and performance after cessation of light exposure. The 460 nm light also caused a preferential elevation of high-alpha activity (9.5–10.5 Hz) in the EEG compared to 555 nm light (data not shown), which has been shown previously to correlate highly with subjective sleepiness ratings. Reproduced from Lockley SW, Evans EE, Scheer FAJL, et al. (2006) Short-wavelength sensitivity for the direct effects of light on alertness, vigilance, and the waking electroencephalogram in humans. *Sleep* 29(2): 161–168.

alertness is reduced, performance is impaired, and the risk of on-the-job accidents and injuries increases. Meals are also taken at inappropriate circadian phases, causing postprandial levels of insulin, glucose, and fats to rise, potentially increasing the risk of heart disease and diabetes.

The problems associated with circadian desynchronization due to shift work can be alleviated once the circadian system has fully adapted to the new schedule. A general 'rule of thumb' is that the circadian system takes 1 day to shift by approximately 1 h under uncontrolled conditions; thus, adapting to a 12 h shift from day shift to night shift will require approximately 12 days. This is possible in fixed night or day shifts that persist for several weeks at a time. Unfortunately, if the shift duration lasts only 2 weeks, the shift is changed again only 1 or 2 days after adaptation has been achieved, and desychronization is reestablished. Most shift systems change every few days or are interrupted by days off (when workers tend not to maintain the same schedule as during work days), making it virtually impossible for the circadian system to adapt. A different approach is to have a rotating shift system that is so rapid (e.g., only 1 day on each shift) that it does not induce major shifts in circadian phase. This system, however, requires workers to be on shift at night in a fully unadapted state and therefore on duty during the biological night when the risk of accidents is highest. More detailed information on shift work disorder in humans is provided elsewhere in this encyclopedia.

Jet lag Another common circadian rhythm disorder is jet lag, which is essentially the same problem as experienced by shift workers except that the desynchronization between the circadian system and the environment is caused by the rapid change in light–dark cycle following transmeridian travel. As with shift work, the circadian system cannot adapt to the immediate change in light–dark schedule and the symptoms (insomnia, fatigue, impaired cognition, and gastrointestinal problems) persist until the circadian system has adapted to the new time zone, which will take up to 12 days, depending on the number of time zones crossed. Unlike shift work, however, most episodes of jet lag are isolated, and the magnitude and direction of the phase shift required to adapt are generally predictable. It is possible, therefore, to design light–dark exposure schedules to facilitate circadian adaptation and reduce the number of days required to reentrain to the new time zone. Such schedules require consideration of the PRC for light in order to time light and dark exposure appropriately. Often, travelers will be told to 'get on the new time zone as soon as possible' in order to adapt more quickly and reduce their jet lag symptoms. Although in principle this advice is valid, working through several example itineraries with respect to the PRC illustrates that this approach is oversimplistic and requires a more sophisticated consideration of the timing of light. The following example illustrates how such advice could be counterproductive and demonstrates the general principles in designing jet lag adaptation schedules.

Consider a flight from New York to London leaving at 07.00 h. The flight lasts 7 h and arrives in London at 19.00 h local time, given the 5 h difference in time zones. In order to adapt to UK time, travelers from the United States need to advance their circadian system; eastward travel always requires a phase advance, whereas westward adaptation requires a phase delay. According to the PRC, light exposure at 19.00 h will cause a phase delay, opposite to that required. When the traveler arrives in London, however, minimal circadian adaptation has yet taken place and therefore he or she remains entrained to US time (i.e., 14.00 h). Consultation of the PRC for 14.00 h shows that light at this time and for the next 4 h will cause a phase advance. In fact, light exposure during the entire duration of the flight is also at the correct time to cause a phase advance. The advice to the traveler is therefore to expose him- or herself to as much light as possible for the entire flight and for the first 4 h in London, retiring to bed at 23.00 h and thereby avoiding the phase delaying effects of light that would begin at this phase (**Figure 11**).

Now consider the same flight but this time leaving New York at 19.00 h and arriving in London at 07.00 h. Again, a 5 h phase advance is required to adapt to the new time. Most people would follow the same plan as the daytime flight and expose themselves to light upon landing in London. Closer inspection of the PRC, however, reveals that this plan will in fact prevent adaptation to London time and likely shift their circadian system in the opposite direction than that required, sending their circadian systems toward Hawaii. When landing in London at 07.00 h, the circadian system is still synchronized to US time (i.e., 02.00 h). Light exposure at this time causes a particularly robust phase delay, the opposite direction to that required. It is therefore necessary to avoid light at this time and for the next 4 h until the PRC predicts that light will cause a phase advance (06.00 h) (**Figure 11**). The advice to the traveler in this case is to avoid light during the entire flight and for the first 4 h in London (by sleeping, using an eye mask, or wearing dark sunglasses) and then expose him- or herself to as much light as possible after this time (**Figure 11**). Any schedule can be assessed in this

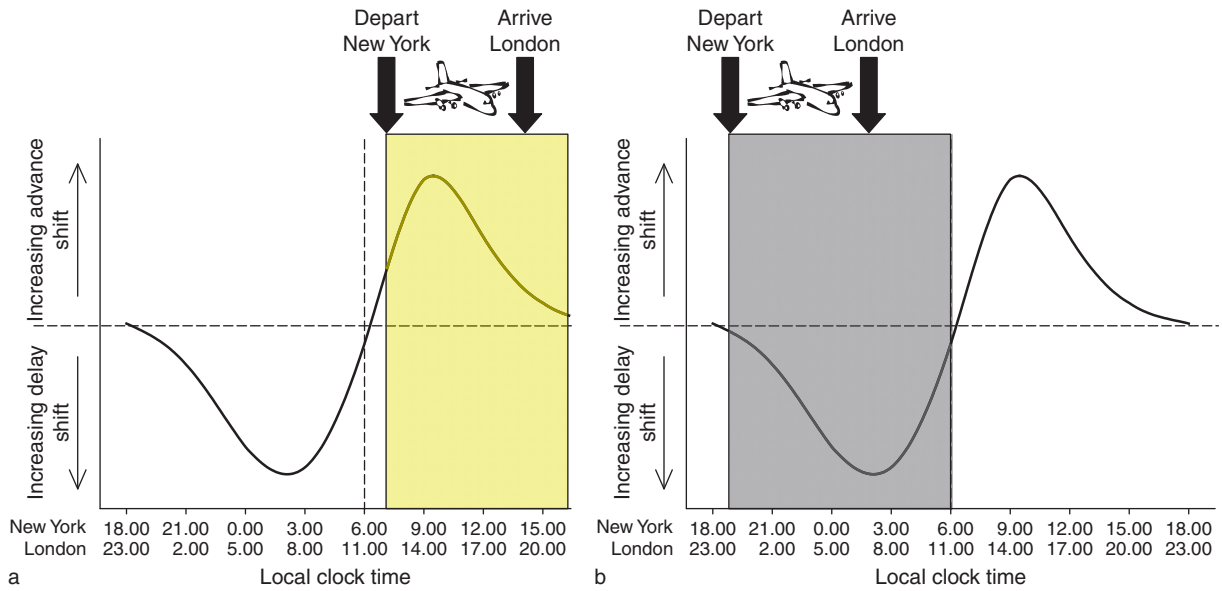

Figure 11 Application of circadian principles in reducing the symptoms of jet lag. Two example flights from New York to London are shown relative to the phase response curve (PRC) for light. Times are shown for both New York and London time zones (5 h apart). The point on the PRC when the change from phase delay to phase advance shifts occurs is shown by a dashed line (06.00 h, New York time). (a) A 7 h flight leaving New York at 07.00 h and arriving in London at 19.00 h local time. Light exposure during the flight and for the first 4 h in London (yellow rectangle) will cause a phase advance shift required for adaptation following an eastward flight. Light exposure after that time should be kept minimal in order to prevent the delay shift which would coincide with 23.00 h London time (18.00 h New York time). (b) Now consider the same flight but leaving New York at 19.00 h and arriving in London at 07.00 h local time. Light exposure during the flight and for the first 4 h in London would cause a phase delay shift, the opposite to that required, and therefore light exposure should be minimal at these times (gray rectangle). After 11.00 h London time, light exposure will cause a phase advance and therefore should be maximized. Adapted from Rajaratnam SMW and Arendt J (2001) Health in a 24 h society. *Lancet* 358: 999–1005.

way, and by judicious use of light exposure (natural light, even on a cloudy day, is very effective and indoor room light is also sufficient if natural light is not accessible) and light avoidance (dark sunglasses will suffice if the eyes cannot be closed), adaptation to the new time zone can be significantly enhanced.

Advanced and delayed sleep phase syndrome Advanced sleep phase syndrome (ASPS) and delayed sleep phase syndrome (DSPS) are circadian rhythm disorders which cause sleep to occur routinely at a particularly early or particularly late time, respectively, relative to normal sleep. These disorders are described in detail elsewhere in this encyclopedia. Treatment of these syndromes requires that the sleep–wake cycle be phase shifted in the appropriate direction. Again, the PRC for light provides the guide to timing; evening light therapy would cause the required phase delay to correct ASPS and morning light therapy would phase advance the circadian system in DSPS. Without knowledge of the PRC, however, light therapy could be timed at the wrong circadian phase, exacerbating the disorder. The exact timing of the light can be determined from a reliable marker of circadian phase which, in an outpatient setting, can be derived simply from salivary or urinary rhythms of cortisol or melatonin and its major metabolite (6-sulfatoxymelatonin).

Unusual environmental photoperiods Some specialized environments present unusual challenges to the circadian system. As outlined previously, submariners scheduled to work an 18 h 'day' in relatively dim light conditions are unable to maintain circadian entrainment to their light–dark cycle and revert to their endogenous circadian period. Similar observations have been made during the Antarctic winter, when only artificial light is available. Interestingly, circadian rhythms can remain entrained at bases where strict schedules are maintained, whereas at bases where researchers are free to keep their own schedules, their circadian rhythms 'free run' on a non-24 h day, highlighting the importance of nonphotic time cues when living under conditions of low light.

Space travel exposes astronauts to challenging light–dark schedules that are the result of both natural and man-made conditions. In the space shuttle, a very high-amplitude, 90 min light–dark cycle is observed as the shuttle orbits the earth. General light levels, however, are remarkably low, averaging

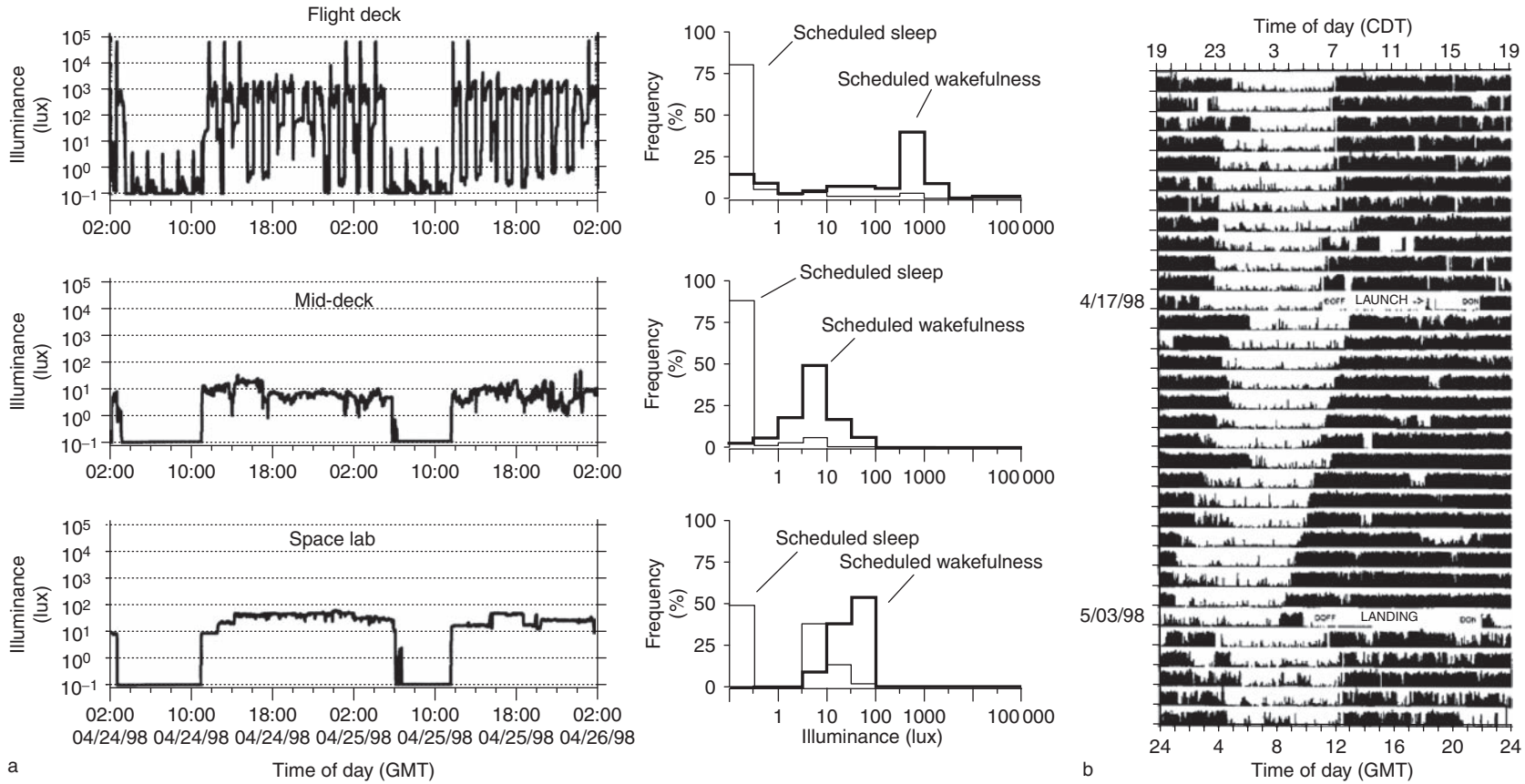

Figure 12 Light exposure and sleep–wake schedules experienced by space shuttle crew. Light levels were measured during the STS-90 shuttle mission using a portable light meter stationed in the habitable quarters of the space shuttle – that is, the flight deck, windowless mid-deck, and the space lab. The time course of illuminance over a representative 48-h episode (a, left) and the corresponding relative frequency distribution of illuminances during wakefulness (thick solid line) and sleep (thin solid line) (a, right) are shown for all three areas. The mid-deck and space lab, where crew members spend most of their time, have stable light–dark patterns with a period of ~23.67 h, although the absolute intensity of light is relatively dim (maximum of 93 and 171 lux, respectively). The flight deck, however, has a much different light–dark pattern. Although the ~23.67-h pattern is apparent, superimposed is an extremely high-intensity light pulse, reaching a maximum of 80 000 lux, caused by the orbital dawn that occurs approximately every 90 min as the shuttle orbits the Earth. This pulse is even observed when the shades are drawn during scheduled sleep. (b) An activity–rest pattern measured using wrist actigraphy from one shuttle crew member before, during, and after the STS-90 mission. Activity levels are shown per minute, with low activity indicating sleep times (except during launch and landing, when monitors were not worn). The 23.67-h schedule imposed on the crew members in order for the shuttle to take off and land on schedule is clearly seen, with sleep start and end times getting progressively earlier day to day. Sleep start time, however, is clearly more variable than wake time, and the duration of sleep obtained varies from night to night. Upon landing, an abrupt 4-h phase delay shift in sleep times is also observed. The low-amplitude light–dark cycles in the mid-deck and space lab and the variable exposures in the flight deck (a), coupled with the requirement to remain on a non-24 h schedule (b), are likely to induce desynchrony between the sleep–wake cycle and the circadian system and expose crew members to the inherent risks caused by the sleeping and waking at an inappropriate circadian phase. Reproduced from Dijk D-J, Neri DF, Wyatt JK, et al. (2001) Sleep, performance, circadian rhythms, and light–dark cycles during two space shuttle flights. *American Journal of Physiology Regulatory, Integrative and Comparative Physiology* 281: R1647–R1664.

10–100 lux in the mid-deck and space lab, respectively, where astronauts spend most of the time (Figure 12(a)). Superimposed on these ambient light levels is a <24 h schedule that is required in order to ensure that the shuttle takes off and lands at the correct times (Figure 12(b)). Although precise measures of circadian phase have not been conducted in space, it is likely that these unusual light–dark cycles cause desynchronization of the circadian system from the astronauts' sleep–wake cycle, impairing sleep and causing fatigue in an extremely dangerous environment. In fact, sleep disorders are highly prevalent in astronauts and hypnotics are the most often used medications in space, suggesting that circadian desynchrony may indeed have a significant negative impact on sleep and alertness. Other space missions are also likely to introduce problems related to circadian entrainment; for example, manned missions to Mars will require both astronauts and Earth-based ground crew to be entrained to a day length of 24.65 h to match that on Mars in order to maintain optimal performance. The proposed missions to set up a lunar base camp also present significant problems because artificially generated light and darkness will be required to allow those living on a lunar base to remain entrained to a 24 h day.

Aging Most adults notice a change in their circadian rhythms as they age such that they choose to sleep and wake at both an earlier time and an earlier circadian time than they did when they were younger. Adolescents tend to sleep and wake at a later circadian phase than younger children and adults. These changes are due in large part to age-related changes in how the circadian system controls sleep–wake timing and are detailed elsewhere in this encyclopedia.

Aging is also associated with a change in light–dark input to the circadian pacemaker which may contribute to the phase advance of sleep experienced by older adults. As the lens in the eye ages, its pigmentation yellows due to a change in the composition of lens proteins resulting in a reduction in the transmission of short-wavelength light that is most effective at resetting the circadian pacemaker (Figure 13).

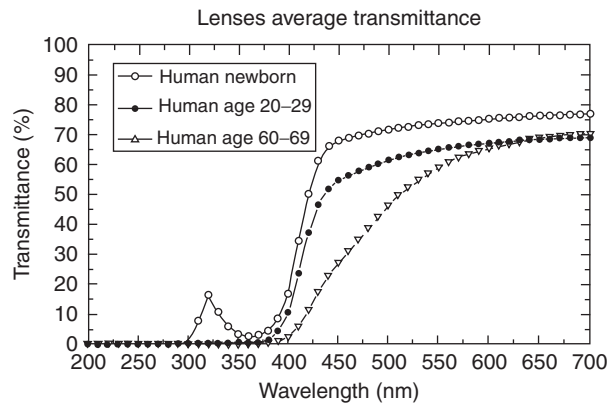

Figure 13 Ocular lens transmission as a function of age. The transmission spectra for postmortem lenses to visible and UV light show a clear age-dependent change. Lenses of older subjects (60–69 years old, $n = 40$) fail to transmit a high proportion of the short-wavelength visible light (400–550 nm) that maximally stimulates the circadian photoreception system compared to the lenses of young adults (20–29 years old, $n = 36$). Newborn lenses ($n = 5$) transmit a higher proportion of all visible and some UV wavelengths. Reproduced from Brainard GC, Rollag MD, and Hanifin JP (1997) Photic regulation of melatonin in humans: Ocular and neural signal transduction. *Journal of Biological Rhythms* 12(6): 537–546.

Psychiatric disorders Circadian rhythm and sleep–wake problems are associated with a range of psychiatric disorders, some of which are related to alterations in light–dark exposure. The most well-characterized light-related disorder is seasonal affective disorder (SAD) or 'winter depression.' This syndrome is characterized by the onset of a depressive episode coincident with the onset of fall/winter and specifically the change in night length during this season. An effective treatment for SAD is ocular light therapy, which aims to ameliorate the lack of light that is thought to initiate the depressive episode. The exact mechanism underlying light therapy is unknown, although light exposure in the morning may be more effective than at other times of day, suggesting that a phase delay of the circadian system may contribute to its onset in some cases. Although clinicians generally prescribe daily bright white light therapy (2500–10 000 lux) for 30–60 min, few systematic studies have been conducted to examine the effect of light intensity, duration, and pattern on clinical outcomes. Given the nonlinear relationship between these properties and the circadian phase resetting effects of light, it may be possible to induce similar therapeutic benefits using dimmer and/or intermittent light exposure, which may improve patient compliance. Similarly, selecting the most potent wavelengths of light may also reduce the intensity or duration of therapy required. Studies have shown that short-wavelength (blue) light exposure for 45 min per day can be an effective therapy for SAD, and further investigations are ongoing.

Another sleep–wake rhythm disorder related to light exposure is observed in patients with Alzheimer's disease (AD). These patients have a highly disturbed rest–activity cycle, with greater activity at night and reduced activity in the day compared to similar-aged patients without AD. These symptoms are most often the ones that prompt institutionalization of AD patients, although this may lead to further exacerbation of rest–activity disorders because robust light–dark

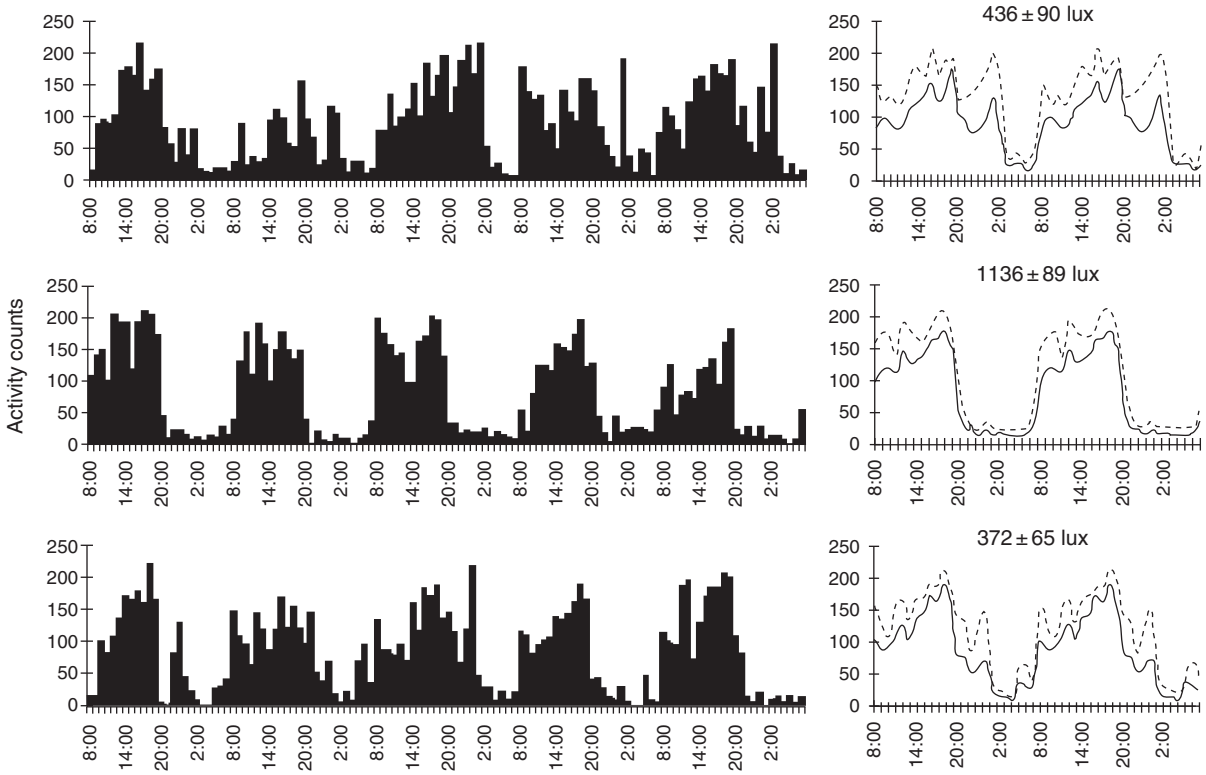

Figure 14 Effect of changes in ambient light exposure on rest–activity patterns in Alzheimer's disease. Patients with Alzheimer's disease ($n = 16$), multi-infarct dementia ($n = 3$), dementia associated with alcoholism ($n = 2$), and normal pressure hydrocephalus ($n = 1$) were studied in a nursing home environment for ~10 weeks each. After a 2-week baseline, bright lights were installed in the ceiling (during the night so that patients would not see the fitting) and remained for 4 weeks, when they were removed again for 4 weeks (baseline 2). Patients' rest–activity and light–dark exposure was assessed using wrist actigraphy during the baseline, during the last 2 weeks of light therapy, and during the last 2 weeks of normal light (baseline 2). During baseline 1, average light levels (± SEM) in the living rooms were 436 ± 90 lux (range, 93–1417). Under the brighter light conditions, light levels increased to 1136 ± 89 lux (range, 790–2190) and returned to 372 ± 65 lux (range, 110–1106). Raw activity data of a patient with Alzheimer's disease are shown on the left for 5 days before (top), during (middle), and after (bottom) light treatment. The right panels show double plots of the average 24 h activity level (solid line) +1 SD (dashed line). On average, the variability of the rest–activity rhythm was reduced and its stability improved during the light exposure compared to the pooled pre- and postbaseline levels. These effects were not observed in patients with severe visual impairment. Reproduced from Van Someren EJW, Kessler A, Mirmiran M, and Swaab DF (1997) Indirect bright light improves circadian rest–activity rhythm disturbances in demented patients. *Biological Psychiatry* 41: 955–963.

cycles are not generally available to those living in nursing homes. Examination of the circadian rhythms of cortisol and melatonin in these patients does not reveal gross abnormalities in phase or arrhythmic patterns that might be expected to be associated with such disturbed rest–activity cycles, suggesting that the rest–activity disorders may be caused by an attenuation of the output of the circadian pacemaker to other efferent targets. Restoration of robust light–dark cycles, however, has a dramatic effect on rest–activity patterns and results in more consolidated bouts of both sleep and wake in institutionalized patients (Figure 14), although the mechanism is unknown. Ongoing research suggests that such light therapy may even improve mortality and morbidity in patients with AD. Investigations of the mechanism underlying the therapeutic benefits and optimization of light therapy (intensity, duration, wavelength, and pattern) are ongoing.

Other psychiatric disorders may also be related to changes in light exposure or sensitivity to light or may respond to light therapy (e.g., unipolar and bipolar depression, schizophrenia, and Parkinson's disease), but further work is required to substantiate these relationships. The association between circadian rhythms and psychiatric disorders is outlined in further detail elsewhere in this encyclopedia.

See also: Circadian Organization; Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Organization in Non-Mammalian Vertebrates; Circadian Function and Therapeutic Potential of Melatonin in

Humans; Circadian Rhythms in Sleepiness, Alertness, and Performance; Entrainment of Circadian Rhythms by Light; Genetics of Circadian Disorders in Humans; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Photoreceptor Adaptation; Photoreceptors and Circadian Clocks; Photoreceptors: Physiology; Psychiatric Disorders Associated with Disturbed Sleep and Circadian Rhythms; Seasonal Changes in Night-Length and Impact on Human Sleep; Shift Work and Circadian Rhythms; Sleep and Circadian Rhythm Disorders in Human Aging and Dementia.

Further Reading

Brainard GC, Rollag MD, and Hanifin JP (1997) Photic regulation of melatonin in humans: Ocular and neural signal transduction. *Journal of Biological Rhythms* 12(6): 537–546.

Brainard GC and Hanifin JP (2005) Photons, clocks, and consciousness. *Journal of Biological Rhythms* 20: 314–325.

Brainard GC, Hanifin JP, Greeson JM, et al. (2001) Action spectrum for melatomin regulation in humans. Evidence for a novel circadian photoreceptor. *Journal of Neuroscience* 21(6): 6405–6412.

Cajochen C, Zeitzer JM, Czeisler CA, and Dijk D-J (2000) Dose–response relationship for light intensity and ocular and electroencephalographic correlates of human alertness. *Behavioural Brain Research* 115: 75–83.

Cajochen C, Jud C, Munch M, et al. (2006) Evening exposure to blue light stimulates the expression of the clock gene PER2 in humans. *European Journal of Neuroscience* 23(4): 1082–1086.

Cajochen C, Munch M, Kobialka S, et al. (2005) High sensitivity of human melatonin, alertness, thermoregulation, and heart rate to short wavelength light. *Journal of Clinical Endocrinology and Metabolism* 90(3): 1311–1316.

Czeisler CA, Shanahan TL, Klerman EK, et al. (1995) Suppression of melatonin secretion in some blind patients by exposure to bright light. *New England Journal of Medicine* 332: 6–11.

Czeisler CA and Wright KP Jr. (1999) Influence of light on circadian rhythmicity. In: Turek FW and Zee PC (eds.) *Neurobiology of Sleep and Circadian Rhythms*, pp. 147–180. New York: Dekker.

Dijk D-J, Boulos Z, Eastman CI, Lewy AJ, Campbell SS, and Terman M (1995) Light treatment for sleep disorders: Consensus reports I–VII. *Journal of Biological Rhythms* 10(2): 105–176.

Dijk D-J, Neri DF, Wyatt JK, et al. (2001) Sleep, performance, circadian rhythms, and light–dark cycles during two space shuttle flights. *American Journal of Physiology Regulatory, Integrative and Comparative Physiology* 281: R1647–R1664.

Dijk D-J and Lockley SW (2002) Integration of human sleep–wake regulation and circadian rhythmicity. *Journal of Applied Physiology* 92(2): 852–862.

Duffy JF and Dijk D-J (2002) Getting through to circadian oscillators: Why use constant routines. *Journal of Biological Rhythms* 17(1): 4–13.

Gronfier C, Wright KP Jr., Kronauer RE, Jewett ME, and Czeisler CA (2004) Efficacy of a single sequence of intermittent bright light pulses for delaying circadian phase in humans. *American Journal of Physiology Endocrinology and Metabolism* 287: E174–E181.

Jewett ME, Kronauer RE, and Czeisler CA (1994) Phase-amplitude resetting of the human circadian pacemaker via bright light: A further analysis. *Journal of Biological Rhythms* 9(3–4): 295–314.

Klerman EB, Shanahan TL, Brotman DJ, et al. (2002) Photic resetting of the human circadian pacemaker in the absence of conscious vision. *Journal of Biological Rhythms* 17(6): 548–555.

Kronauer RE, Forger DB, and Jewett ME (1999) Quantifying human circadian pacemaker response to brief, extended, and repeated light stimuli over the phototopic range. *Journal of Biological Rhythms* 14(6): 500–151. [Erratum in *Journal of Biological Rhythms* 15(2): 184–186; 2000].

Lockley SW, Brainard GC, and Czeisler CA (2003) High sensitivity of the human circadian melatonin rhythm to resetting by short wavelength light. *Journal of Clinical Endocrinology and Metabolism* 88(9): 4502–4505.

Lockley SW, Evans EE, Scheer FAJL, et al. (2006) Short-wavelength sensitivity for the direct effects of light on alertness, vigilance, and the waking electroencephalogram in humans. *Sleep* 29(2): 161–168.

Rajaratnam SMW and Arendt J (2001) Health in a 24-h society. *Lancet* 358: 999–1005.

Ruberg FL, Skene DJ, Hanifin JP, et al. (1996) Melatonin regulation in humans with color vision deficiencies. *Journal of Clinical Endocrinology and Metabolism* 81(8): 2980–2985.

Ruger M, Gordijn MC, Beersma DG, de Vries B, and Daan S (2006) Time-of-day-dependent effects of bright light exposure on human psychophysiology: Comparison of daytime and nighttime exposure. *American Journal of Physiology Regulatory, Integrative and Comparative Physiology* 290(5): R1413–R1420.

Smith KA, Schoen MW, and Czeisler CA (2004) Adaptation of human pineal melatonin suppression by recent photic history. *Journal of clinical Endocrinology and Metabolism* 89(7): 3610–3614.

Thapan K, Arendt J, and Skene DJ (2001) An action spectrum for melatonin suppression: Evidence for a novel-rod, non-cone photoreceptor system in humans. *Journal of physiology* 535(7): 267–267.

Van Someren EJW, Kessler A, Mirmiran M, and Swaab DF (1997) Indirect bright light improves circadian rest–activity rhythm disturbances in demented patients. *Biological Psychiatry* 41: 955–963.

Zeitzer JM, Dijk D-J, Kronauer RE, Brown EN, and Czeisler CA (2000) Sensitivity of the human circadian pacemaker to nocturnal light: Melatonin phase resetting and suppression. *Journal of Physiology* 526(3): 695–702.

Circadian Systems: Evolution

T L Page, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Circadian rhythms are daily oscillations in physiology, metabolism, or behavior that persist in organisms that have been isolated from periodic fluctuations in the environment. The rhythms are thought to have arisen from selection pressure for organisms to organize activities in time as well as in space, and thus to reflect the evolution of an internal timekeeping system – the circadian clock. The fact that circadian rhythms have been described in every major group of eukaryotic organisms, coupled with their discovery in the prokaryotic cyanobacteria, suggests that the origin of circadian clocks occurred very early in evolutionary history, and the molecular data indicate that circadian clocks may have arisen multiple times. While a discussion of the evolution of circadian clocks can be approached from a variety of perspectives, the focus here is on the circadian systems of the metazoa and how the components of the system above the cellular level have been organized through evolutionary history to provide for the temporal organization of the individual.

Fundamental Properties

There are three universal features ('clock-like' properties) of circadian rhythms that provide clues to their adaptive significance and thus to selective pressures that have shaped the evolution of the circadian systems that drive them. First, these rhythms are under the control of regulatory systems that are based on endogenous, self-sustaining oscillators (or pacemakers) whose periods approximate those of the naturally occurring 24-h environmental cycles. The self-sustaining property of the oscillator gives the organism a means to anticipate, rather than simply react to, periodic environmental change. Second, a universal feature of circadian rhythms is that the period of the oscillation is at most only slightly affected by changes in the constant level of ambient temperature. The 'temperature compensation' of period is a functional prerequisite for the underlying oscillator's function in time measurement, since a clock whose motion is temperature sensitive is not particularly useful. Finally, the phase and period of the rhythms are subject to entrainment by a limited number of environmental cycles that are referred to as 'zeitgebers' (literally, time-givers). Zeitgebers are what 'set the clocks to local time' and thus maintain an appropriate phase relationship of the circadian system with the external world. Environmental light cycles are universally effective in the entrainment of circadian rhythms, and while other stimuli such as temperature, sound, or food availability may be effective in some species, it is clear that light intensity, as the most reliable and 'noise-free' environmental cycle, is the dominant zeitgeber. A final generalization can be proposed that is pertinent to the current discussion. Circadian oscillations are generated within individual cells and while intercellular interactions may influence the phase and period of the oscillation, they are not necessary for its generation. This conclusion is based not only on the fact that single-celled organisms exhibit circadian rhythms that possess all of the general properties just described, but also on recent demonstrations that individual cells isolated from multicellular organisms can express circadian rhythms in cell culture.

Function

Circadian systems have two conceptually different functional roles in the temporal organization of the individual. The first is to restrict to a particular time of day aspects of physiology, behavior, and metabolism that are best undertaken at a specific phase of the environmental cycle. A second major function of circadian oscillations is to provide a temporal framework for timing various activities relative to each other. Thus the system provides for an internal 'temporal program' that is an important adjunct to temporal order with respect to the external environment. In both cases, a key factor is the control of phase relationships – both among various rhythms and with respect to the cycles of the immediate external environment.

For this reason, one emerging view is that an important element of selection in the evolution of circadian systems has been the development of entrainment mechanisms that are appropriate to the environmental niche occupied by the individual. Similarly, because multicellular organisms are composed of multiple circadian oscillators (see later), it is plausible to suggest there is strong selection for mechanisms to ensure coordination of phase relationships among oscillators – that is, all oscillators within the individual ultimately must have access to a common phase (time) reference. Conceptually, this could occur either by providing all oscillators with direct access to environmental cycles or by developing coupling mechanisms to allow internal communication between oscillating elements. The discussion that follows is largely based on the proposition that the requirement of mechanisms

to ensure proper phase control (development of a reliable time reference) is a major selective force that provides a framework for interpretation of phylogenetic differences in the circadian organization of major groups of animals.

Physiological Organization and Evolution

Figure 1 illustrates the basic components of a circadian system. The model consists of four essential elements – a pacemaker or oscillator that generates the primary timing signal, photoreceptors for entrainment, and two coupling pathways, one that mediates the flow of entrainment information from the photoreceptor to the pacemaker and another that couples the pacemaker to the effector mechanisms that it controls. A first step in the effort to understand the evolutionary history of the organization of circadian systems in animals is to define the physiological and anatomical organization through the identification of cells or tissues that generate the circadian oscillations, elucidation of the signals by which the oscillations control relevant rhythms, and delineation of the pathways by which the oscillators receive environmental information for entrainment. One can then exploit these data to discuss the evolution of the circadian system in the context of a comparative phylogenetic approach. Here the focal questions are: (1) What do circadian systems of various groups have in common and how are they substantively different? (2) Based on similarities and differences, combined with the natural history and evolutionary relationships among the groups, can one devise a plausible hypothesis about the relevant selective pressures, evolutionary history, and adaptive significance of central features of circadian organization? Addressing the first question would appear to be relatively straightforward, since the answer can be based on hard data derived from physiological and anatomical study of extant organisms. The immediate difficulty here, though, is that our knowledge of the physiological and anatomical organization of circadian systems is both incomplete and rapidly changing – as a result, much of the discussion that follows is as much about the recent evolution of our understanding of how circadian systems are organized as it is about the evolution of circadian organization in natural populations. Attempts to answer the second question are even more difficult and involve much in the way of conjecture, supposition, and frank speculation. This is because much of the relevant information (such as the specifics of the photic environments of ecological niches inhabited by extinct organisms) is lost to history and thus involves a substantial amount of guesswork (hopefully informed).

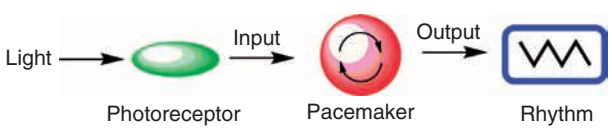

Figure 1 Functionally defined model showing the four major components of the circadian system that drive the circadian rhythm. Environmental light is detected by an entrainment pathway that consists of a photoreceptor and a coupling mechanism. The cycle of light and dark synchronizes a self-sustaining oscillator (pacemaker). The output of the pacemaker regulates the timing of various processes (e.g., locomotor activity) via a second coupling pathway.

Photoreceptors for Entrainment

One especially striking feature of circadian organization is that animals almost never rely exclusively on photoreceptors in the retina for entrainment. With a few notable exceptions (discussed later), removal of the eyes of mollusks, crustaceans, insects, fish, amphibians, reptiles, and birds typically has little effect on the ability of the circadian system to be entrained by light. The results indicate that extraretinal receptors are dominant in detecting light for entrainment in all major groups that have been studied, with the exception of mammals. In mammals, there are no known exceptions to the generalization that they rely exclusively on retinal photoreceptors for entrainment, and this raises the interesting evolutionary question of why mammals have lost all photoreception outside of the retina, or, alternatively, why extraretinal photoreceptors have been retained in all other groups.

Extraretinal photoreceptors that function in entrainment in nonmammalian vertebrates may be widely distributed in the individual. The pineal complex (pineal gland and parietal eye), which is photoreceptive in all vertebrates except mammals, is known to contribute to entrainment. In addition, evidence shows that 'deep brain photoreceptors' (which have been localized to several brain regions) are also involved. Further, as discussed later, recent evidence from zebra fish indicates that photoreceptors for entrainment may be very widespread and reside in several, perhaps even most, tissues.

Interestingly, even when the photoreceptors mediating entrainment are located in the retina, results suggest at least a subset of them represent a subpopulation of cells that are not involved in forming vision. For example, studies in mammals suggest that entrainment is mediated by a unique population of retinal ganglion cells. These cells contain a novel photopigment (melanopsin), are photosensitive, and project directly to a circadian clock in the suprachiasmatic nucleus (SCN) of the hypothalamus. The separation of photoreceptors for image formation (vision)

and for entrainment may be explained by the fact that the two functions have rather different requirements, a suggestion supported by observation that the photoreceptors for entrainment and those for vision have distinctly different physiological properties. For example, photoreceptive retinal ganglion cells involved in entrainment in mammals integrate light information over relatively long timescales, respond only weakly, if at all, to brief stimuli, and show little adaptation.

Localization of Circadian Oscillators

The notion of localization of circadian pacemakers encompasses two separate questions. The first is what cells or tissues are able to generate circadian oscillations, and the second is what roles do those tissues play in driving overt rhythms in physiology, metabolism, and behavior. While it has been known for many decades from studies on protists and fungi that single eukaryotic cells had the capacity to generate circadian oscillations, until recently it had generally been argued that the number of cells or tissues that could function as circadian oscillators in multicellular organisms was limited. The arguments were generally supported by two kinds of observations.

First, in a variety of organisms, vertebrate and invertebrate, lesion and transplantation studies indicated that identifiable neural or endocrine structures (such as the optic lobes of cockroaches, the pineal gland of sparrows, and the SCN of mammals) were responsible for the control of multiple rhythms, suggesting that they were 'master oscillators' whose distributed output determined the temporal organization of the individual. Second, efforts to demonstrate that cells from a wide variety of tissues could generate circadian oscillations *in vitro* met with only limited success, and those successes most frequently involved tissues that were thought to function as 'the clock' in one organism or another. For example, in vertebrates, early studies on pacemaker localization generally proceeded on the assumption that since entrainment by light was a universal feature of circadian clocks, pacemakers would likely be found in close association with known photoreceptive structures. This assumption led to the discovery that three structures that are found in all vertebrate groups, from cyclostomes to mammals – the retina, the pineal gland (which is photoreceptive in all nonmammalian vertebrates), and the suprachiasmatic nuclei of the hypothalamus (which receive direct input from the retina via the retino-hypothalamic tract or from 'deep brain photoreceptors') – played significant roles in vertebrate circadian organization. Precisely what role these structures played varied significantly among species and the variation, at best, only roughly followed phylogenetic lines. However, in one species or another, each of these structures appeared to be essential for expression of overt circadian rhythms, and these structures were also the ones widely shown to generate circadian oscillations *in vitro*.

The notion of a centralized, master oscillator seemed to be particularly true of the most studied group, the mammals. It was known from very early studies that hypothalamic lesions could abolish circadian rhythms in behavior in mammals. This fact, combined with the discovery of a direct connection between the retinas of the eyes and the SCN of the hypothalamus, led to a cottage industry of examining the effects of SCN lesions on various overt rhythms. It was rather quickly shown that lesions to the SCN abolished rhythms in wheel running activity, drinking, plasma corticosterone levels, pineal *N*-acetyltransferase, pineal melatonin, growth hormone secretion, luteinizing hormone secretion, body temperature, and sleep/wake cycles. With one exception (the 'food anticipatory activity' rhythm), there were no convincing reports of rhythms that persisted after complete destruction of the SCN and retained properties that compelled acceptance of the notion that a functionally relevant circadian clock resided outside the SCN. In addition, the disruption of circadian rhythmicity following SCN lesions seemed universal within the mammalian species studied, which ranged from rats to primates, and transplantation and tissue culture studies showed unequivocally that the SCN could generate circadian oscillations that controlled locomotor activity rhythms. These results suggested that mammals were endowed with a restricted and well-defined, centralized locus for the circadian pacemaker. That view was first challenged with the successful demonstration of rhythms in melatonin release from the hamster retina *in vitro*, which showed quite definitively that the retina, in addition to the SCN, contained a perfectly respectable circadian pacemaker.

A more revolutionary change in the landscape came in the past decade, following on the heels of the discovery and characterization of several 'clock genes.' Clock genes are genes that appear to be essential for the generation of circadian oscillations within cells and whose transcription is, in most cases, rhythmic. Interestingly, clock genes appear to be well conserved across phylogenies that range from fruit flies to mammals, indicating strong evolutionary conservation in the molecular mechanism of the clock. By coupling the coding sequences of reporter genes to promoters of clock genes, it was possible to examine tissues in culture rapidly and reliably for the presence of circadian oscillations in transcription in fruit flies, rats, and mice. The results were quite remarkable. Molecular oscillations were found to be pervasive

and widespread in these multicellular organisms, and to exist in peripheral cells and tissues where there had never been a hint of endogenous rhythmicity. In flies, it was shown that isolated wings, antennae, proboscis, and legs were capable of circadian oscillation in culture. In rats and mice, the tissues included (in addition to the SCN) skeletal muscle, liver, lung, fibroblasts, and pineal. Multiple, widely distributed oscillators were also demonstrated in zebra fish, where rhythmic activity in gene transcription has been observed in a variety of isolated tissues. This included not only the 'classical' vertebrate circadian pacemakers – the brain, pineal, and retina – but also the heart, spleen, and kidney. In insects, the generality of the results from fruit flies is buttressed by studies in other insect species that have shown persistent, circadian oscillations in physiology or metabolism *in vitro* in tissues that include endocrine tissue (prothoracic glands), testis, and cuticle. These results, spanning a phylogenetically broad spectrum, lead directly to the conclusion that many, perhaps most, cells in organisms have retained the ability to generate circadian oscillations throughout the evolutionary history of circadian systems.

The data are clear in establishing the fact that animal circadian systems are composed of many cellular oscillators; we presume (with, as yet, little substantiating data) that these oscillators control physiologically relevant rhythms in the cells or tissues in which they reside. This raises the question of what mechanisms have arisen with the evolution of multicellular organization to ensure that correct phase relationships are maintained among the various oscillators and the rhythms they control. One potential mechanism to ensure appropriate phasing is to have the oscillators directly communicate with one another via internal 'coupling pathways' that allow for hierarchical or mutual entrainment. A second strategy is to give all oscillators access to environmental signals that could directly entrain the individual pacemakers. The available evidence suggests that natural selection has capitalized on both options with the dominant strategy in any particular species being correlated with the behavioral ecology of the individual.

Phase Control and the Evolution of Circadian Systems

A perhaps even more surprising discovery in the studies on *in vitro* molecular rhythms in fruit flies and zebra fish was that the circadian oscillations in gene expression in the various tissues could be entrained by light in culture. Thus it appeared that all peripheral oscillators in these two phylogenetically distant organisms also had dedicated photoreceptors. Available data, though limited, suggest that the ability of pacemaker tissues to directly detect light for entrainment extends to all other major groups with the exception of mammals. With one exception (returned to later), in every case in which a tissue has been shown to oscillate *in vitro*, the rhythms were also entrainable by light *in vitro*. In contrast, the ability to detect light for entrainment appeared to be completely absent in any mammalian tissue except the retina. In mammals, the results suggested that coordination of the newly discovered peripheral molecular oscillators with the environment relied on signals distributed from the circadian clock in the SCN, which was, in turn, entrained by photoreceptors in the retina. The data lead to the suggestion that a presumed ancient tie between photoreception and the molecular oscillator was conserved in the evolution of organization of the circadian systems in all major groups of organisms except mammals.

These results suggest that the answer to the question of how temporal organization is maintained is different between the mammals and other major groups. In the invertebrates and nonmammalian vertebrates, the direct coupling to the photic environment could ensure that all of the clocks are appropriately phased, and, therefore (at least initially), there is no reason to presume that internal coupling is critical to the temporal order of the system. In contrast, in mammals, the results indicate that photoreception for entrainment is restricted to the retinas of the eyes and that (with the exception of retinal rhythms) it is necessary for the phase information to be passed from the retina through the SCN. The entrained oscillation in the SCN then distributes timing information directly to effectors or to peripheral oscillators, and internal entrainment pathways assume a newly prominent role in temporal organization. As a specific example of the shift to SCN dominance, one can point to the observation that regulation of the circadian rhythm in pineal melatonin, an important rhythmic signal molecule in the integration and coupling of circadian oscillators in vertebrates, has shifted in mammals from a photosensitive oscillator in the pineal, common to all nonmammalian vertebrates, to regulation in mammals by the SCN (via the sympathetic innervation of the pineal) with the complete loss of pineal photoreception.

This hypothesis results in the two distinct models illustrated in Figure 2. In one case, the common time reference for all rhythms is the external light cycle, while in the other, phase coordination is directly under the control of an internal master clock, and internal coupling pathways for entrainment between oscillators are the critical factor in maintaining internal phase relationships.

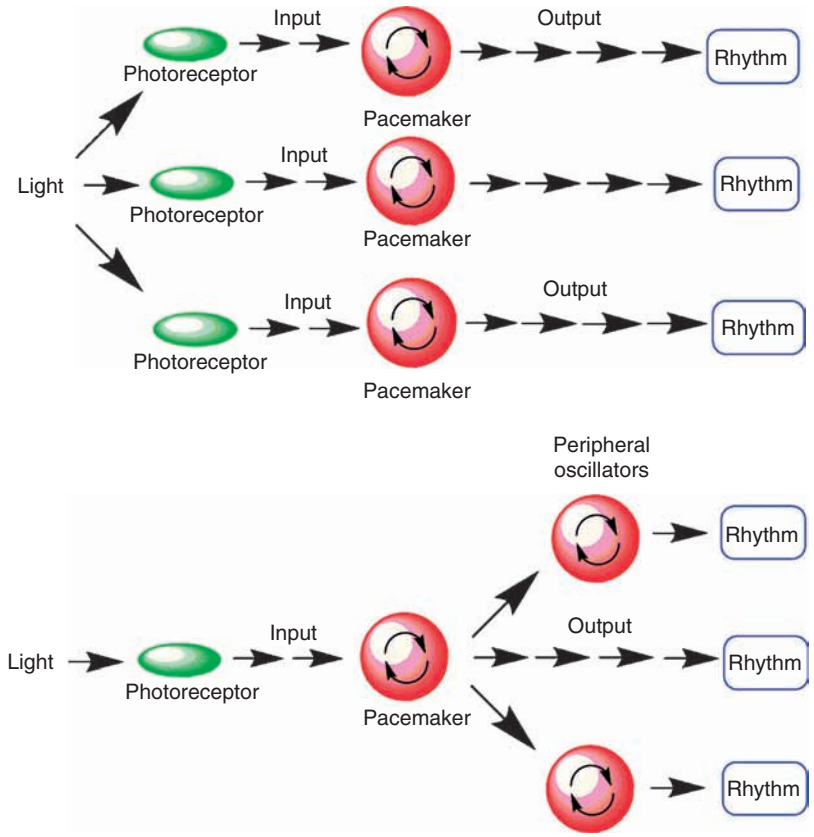

Figure 2 Two models of circadian organization. In the upper model, pacemaking activity arises from a population of distributed oscillators, each with direct access to environmental light information for entrainment, that serve as pacemakers for various rhythms. In the lower model, a centralized pacemaker that is entrained by light either directly drives rhythms in the periphery or regulates the phases of peripheral oscillators that in turn control overt rhythm expression. In this model, the central pacemaker functions as a 'master clock' and serves as the primary source of phase information for maintaining internal temporal organization, while in the upper model, internal temporal organization is dependent on entrainment to the environmental light cycle.

The complete loss of extraretinal photoreception, perhaps driven by the centralization of phase control, appears to be a major step in the evolution of circadian organization and raises the question of why this transition specifically occurred with the mammals. One plausible scenario that has been forwarded derives from the argument that one of the central reasons for the great success of mammals when they initially diverged from their reptilian ancestors approximately 200 Ma was their invasion of the virtually unoccupied nocturnal temporal niche – a time of day when mammals could be active and avoid the dominant, day-active reptilian predators. A switch from a diurnal to nocturnal life style that included hiding in dark burrows during the day would significantly alter the photic environment for entrainment. The nocturnal mammals might only be exposed to the dim light available at transitions in early morning or late evening, or on some days might miss the light signal entirely. For example, laboratory results on flying squirrels given access to a completely dark nest box suggest that these strongly nocturnal animals may only emerge in the light once every several days, and even then only at times near twilight.

Two features of mammalian circadian organization may have arisen as a consequence of the selective pressure of this altered photic environment. First, a strictly nocturnal existence could provide the selective pressures necessary to shift to a more centralized pacemaking system. An internal 'master oscillator' would provide the common phasing signal for maintenance of temporal order among the various peripheral oscillations, shifting the primary source of time information for maintaining internal phase relationships from the now less reliable perception of the daily light cycle to an endogenous clock. Thus even in the absence of a regular exposure to a stable environmental entraining signal, the continued maintenance of appropriate internal phase relationships would be ensured. Second, this shift to a centralized pacemaking system could have led to the centralization of photoreceptor function for entrainment. The restriction of the photoreception to the retina would be advantageous since the photoreceptors are more

directly exposed to light relative to extraretinal photoreceptors buried in tissues of the body. The loss of the extraretinal receptors would obviate any possibility of distributed photoreceptors sending to elements of the circadian system conflicting signals that might result from differential exposure in dim light.

If the arguments presented here are correct, one might look to other organisms with a nocturnal heritage for a similar reorganization of their circadian systems. One group of insects for which we have extensive behavioral, physiological, and anatomical data and that appear both to be strictly nocturnal and to spend daytime in dark hideouts – buried in leaf litter, inside or under decaying logs, or nestled under the bark of trees and thus isolated from the view of light – are the cockroaches (Dictyoptera). Interestingly, cockroaches share several key features of circadian organization with mammals. One remarkable similarity is that all available evidence indicates that photoreception for entrainment in cockroaches, like in mammals, is strictly limited to the retina and there is no evidence for the involvement of extraretinal photoreceptors in entrainment. The retinal photoreceptors entrain a pacemaker that is located in the optic lobes of the brain (adjacent to the eyes). Further, like the mammalian SCN pacemaker, lesions of the optic lobe pacemaker in cockroaches cause rapid loss of most rhythms, including the rhythms in locomotor activity, sensitivity in the photoreceptors of the compound eye, electrical activity in the cervical connectives, and sensitivity of olfactory receptors in the antennae to odor. These results lead to the suggestion that the central pacemakers of both mammals and cockroaches are critical to sustained rhythmicity in nearly all overtly expressed rhythms.

In contrast, in other (day-active) insect models, such as *Drosophila melanogaster*, the available data are more consistent with the view that peripheral oscillations can independently drive sustained, directly entrainable rhythms. For example, in the fruit fly, the data indicate that the central brain pacemaker that regulates circadian rhythms in locomotor activity is neither sufficient nor necessary for the peripheral rhythm in olfactory sensitivity in the antennae. This rhythm is driven by oscillations that are generated and entrained in the antennae. In addition, like mammals, the entrainment information received by the retinal photoreceptors in cockroaches appears to be funneled through the central pacemaker for distribution. In this context, it is notable that a peripheral oscillation in cuticle deposition in cockroach, like the peripheral molecular oscillations in mammals, cannot be entrained by light *in vitro*. The emerging picture is that the two fundamental properties of circadian organization – generation of pacemaking signals for rhythm phase setting and photoreception for entrainment by environmental light cycles – are more centralized in mammals and cockroaches, and the hypothesis posed here is that this centralization traces to the selective pressure of a nocturnal existence.

If a nocturnal existence provides a strong selection pressure for centralization of pacemaker and photoreceptor function, one might also look to moths, another insect group for which we have substantial information on the physiology of the circadian system and which are largely nocturnally active. Based on the physiology, moths appear to have a distributed circadian system with distributed extraretinal photoreceptors much like that of day-active fruit flies, rather than the centralized pacemaker/photoreceptor system of night-active cockroaches. However, rather than escaping light by spending the day in dark burrows, moths often rest in the open during the day, relying on cryptic coloration to protect them from predation. Thus, one could plausibly assume that the peripheral oscillators and their dedicated extraretinal receptors are exposed to, and are readily entrained by, the solar cycle.

There are two other interesting examples where differences in circadian organization between closely related organisms may be traced to differences in temporal behavioral niche. One comes from the nocturnally active, primitive vertebrates, the cyclostomes. In one group of cyclostomes, the lampreys, extraretinal photoreceptors are adequate for entrainment of the activity rhythm. In contrast, in the other group, the hagfish, retinal photoreceptors are required, and the photoreceptive pineal found in other nonmammalian vertebrates is absent. It has been suggested that this difference is related to the fact that lampreys, as parasites on diurnal teleost fish, must be exposed to the light cycle despite a nocturnal activity pattern. In contrast, the hagfish are benthic marine organisms that spend their time feeding inside carcasses of dead fish at depths that are typically over 10 m. Thus it is likely that they are frequently isolated from the daily light cycle.

A second example of a potentially informative dichotomy between closely related species comes from the marine opisthobranch mollusks, *Aplysia californica* and *Bulla gouldiana*. Here the data support the proposition that selection for internal coupling to maintain temporal order between oscillators is related to a nocturnal lifestyle. In both species, circadian oscillators that can be entrained by light *in vitro* reside in cells of the retinas of the two eyes. In *Aplysia*, which are diurnally active and thus can rely on the environmental light cycle as a reliable time reference,

there appear to be no interactions between the ocular pacemakers. When *Aplysia* are placed in constant darkness, isolated from entrainment by the environmental light cycles, the two oscillators rapidly become desynchronized. In contrast, *Bulla* are nocturnally active and bury in mudflats during the day. In *Bulla*, the oscillators of the two eyes are coupled so that even in constant darkness the eyes remain in phase.

Caveats

It is a virtual certainty that the suggestion that evolution has opted for two distinct modes of circadian organization (Figure 2) is an oversimplification and there is good reason to be skeptical that the story is complete. The central arguments are only strongly buttressed by the dichotomies in physiology and behavior represented in the vertebrates by zebra fish and rodents and in the invertebrates by fruit flies and cockroaches and are supported by incomplete data on the physiological organization of the circadian system in other groups. One major omission here is any discussion of the role of internal coupling between oscillators in organisms with distributed circadian systems. A second is the narrow focus on the role of the light cycle in entrainment without recognition of the potential of other environmental factors to drive evolutionary processes in circadian organization. Finally, the hypothesis relies heavily on conjecture about the evolutionary history and natural history of the behavior and photic environment of organisms. Nevertheless, the hypothesis provides a conceptual framework that should be heuristically useful in further investigation on the evolution of circadian organization in the metazoa.

Summary

The view forwarded here is that the photic environment of the organism, largely determined by behavioral strategies, provides a strong selective pressure for dictating the evolution of organization of circadian systems. Most organisms do not escape exposure to daily light cycles, and thus the solar cycle can function as a precise and reliable source of phase information for the myriad of circadian oscillators within the individual. Temporal order can be maintained with a distributed system of oscillators, each with its own photoreceptor, through subjection to entrainment by light. However, for those organisms that have adopted a lifestyle that involves both nocturnal activity and escape from daylight exposure, the external light cycle is a less reliable, less precise source of phase information. As a consequence, it is advantageous to rely more heavily on a centralized, internal clock and on internal coupling for dictating temporal order to the periphery and, concomitantly, to restrict the photoreceptive input to the centralized pacemaker.

See also: Circadian Rhythm Models; Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Circadian Organization in Non-Mammalian Vertebrates; Circadian Regulation in Invertebrates; Genetic Regulation of Circadian Rhythms in Drosophila; Peripheral Circadian Oscillators; Photoreceptor Adaptation; Photoreceptors and Circadian Clocks; Sleep and Waking in Drosophila.

Further Reading

Bell-Pederson D, Cassone VM, Earnest DJ, et al. (2005) Circadian rhythms from multiple oscillators: Lessons from diverse organisms. *Nature Reviews Genetics* 6: 544–556.

Bertolucci C and Foa A (2004) Extraocular photoreception and circadian entrainment in nonmammalian vertebrates. *Chronobiology International* 21: 501–519.

Devlin PF and Kay SA (2001) Circadian photoperception. *Annual Review of Physiology* 63: 677–694.

Dunlap J, Loros J, and DeCoursey P (eds.) (2004) *Chronobiology Biological Timekeeping.* Sunderland, MA: Sinauer Associates, Inc.

Dunlap JC (1999) Molecular basis for circadian clocks. *Cell* 96: 271–290.

Giebultowicz J (2000) Molecular mechanism and cellular distribution of insect circadian clocks. *Annual Review of Entomology* 45: 769–793.

Johnson CH (2005) Testing the adaptive value of circadian systems. *Methods in Enzymology* 393: 818–837.

Menaker M, Moreira LF, and Tosini G (1997) Evolution of circadian organization in vertebrates. *Brazilian Journal of Medical and Biological Research* 30: 305–313.

Page TL (1990) Circadian organization in the cockroach. In: Huber I (ed.) *Cockroaches as Models for Neurobiology Applications in Biomedical Research*, pp. 225–246. Boca Rotan, FL: CRC Press.

Takahashi JS, Turek F, and Moore RY (eds.) (2001) *Handbook of Behavioral Neurobiology XII: Circadian Clocks.* New York: Kluwer Academic/Plenum.

Tamai TK, Vardhanabhuti V, Foulkes NS, et al. (2003) Flies and fish: Birds of a feather. *Journal of Neuroendocrinology* 15: 344–349.

Underwood H, Wassmer GT, and Page TL (1997) Seasonal and diurnal cycles. In: Danzer W (ed.) *Handbook of Comparative Physiology 13: Comparative Physiology*, pp. 1653–1763. New York: Oxford University Press.

Circumventricular Organs

W M Fry and A V Ferguson, Queen's University, Kingston, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Regulation of body fluid homeostasis is complex and critical for survival. It is essential that the body maintain a stable balance whereby there are no net changes in blood volume and osmolality (concentration of dissolved solute) over time. In order to achieve such a balance, complex integration of multiple physiological systems (including behavioral, physiological, and hormonal responses) must take place in a seamless, continuous manner. Although perhaps intuitive, it is important to state at the outset that in order to control such an integrated system regulating blood volume and osmolality, the first essential step is detection of the changes in the interior milieu such that appropriate types and magnitude of responses can be initiated when necessary. This article outlines the roles of the circumventricular organs within the brain in the maintenance of body fluid and mineral balance, with a specific emphasis on the vital roles played by the sensory circumventricular organs (CVOs) in sensing perturbations in these systems and initiating appropriate physiological responses.

The Blood–Brain Barrier

Although the brain is highly vascularized, the tight junctions between vascular endothelial cells which form the blood–brain barrier (BBB) limit free access (by diffusion) of large and ionized solutes from the blood into the neuronal parenchyma. This situation provides protection for neurons from circulating toxins and large shifts in plasma constituents (e.g., Na^+ and glucose), but the BBB also prevents direct monitoring of these variables. Likewise, molecules such as neuropeptides, neurotransmitters, and peptide hormones are prevented from moving from the brain to the circulation, and thus the homeostatic responses that are dependent on such secretion would be severely compromised in the absence of a physiological bridge through the BBB.

Features of CVOs

In most tissue throughout the body, the microvessels are inherently leaky; that is, the movement of solute from the plasma into the surrounding tissue is fairly unrestricted. This is achieved by numerous fenestrations (pores) within the endothelial cells, a lack of tight junctions between endothelial cells, and numerous plasmalemmal vesicles within endothelial cells. Within the brain, however, the BBB is established in cerebral microvessels by tight junctions between endothelial cells and a lack of fenestrations or vesicles. In addition, astrocytic end feet surround the microvessels, probably providing structural and trophic support. Most solute molecules are unable to cross the BBB, and therefore most areas of the brain are not exposed to changes in the composition of these substances in the plasma. Substances can cross the BBB by one of three mechanisms. Small molecules such as oxygen, and lipophilic molecules such as ethanol and steroids, can diffuse across the BBB. The rate of diffusion is proportional to the size and partition coefficient of the molecules. Alternatively, some molecules such as glucose, and some hormones such as leptin and insulin, may be transported across the BBB by specific saturable transporters, the molecular identity and physiological relevance of which have yet to be completely defined. Lastly, the CVOs are physiological 'windows' that contain fenestrated capillaries, through which molecules may pass freely from the circulation into the central nervous system (CNS) (or vice versa). The CVOs play either sensory (permitting molecules in the circulation access to CNS tissue) or secretory (permitting molecules in the CNS access to the circulation) roles. There are three sensory CVOs – the subfornical organ (SFO), the organum vasculosum of the lamina terminalis (OVLT), and the area postrema (AP) – and five secretory CVOs – the median eminence (ME), the neurohypophysis (NH), the intermediate lobe of the adenohypophysis, the subcommissural organ (SCO), and the pineal gland.

Sensory CVOs

The sensory CVOs are primarily composed of neuronal cell bodies, glia, and epithelial cells. They are highly vascularized compared to other areas of the brain, and they exhibit large spaces surrounding the microvasculature (Virchow–Robin spaces). These specializations allow for increased contact between the cells of the CVO and the blood-borne substances. Importantly, they contain a wide variety of receptors to peripheral signals, including regulatory hormones (angiotensin II (Ang II), vasopressin, atrial natriuretic peptide, amylin, and estradiol) and ions (Ca^{2+} and Na^{2+}). In addition, the sensory CVOs have extensive efferent connections with all of the important autonomic control centers in the medulla and hypothalamus.

Subfornical Organ

The SFO is a protrusion of cells on the midline of the wall of the third ventricle, on the dorsal aspect of the lamina terminalis. The SFO sends monosynaptic and polysynaptic connections to the paraventricular nucleus (PVN) and the supraoptic nucleus (SON) of the hypothalamus. Some of these projections are known to excite vasopressin- and oxytocin-producing magnocellular neurons of the PNV and SON. Other projections connect to parvocellular neurons of the PVN that project to the median eminence, medulla, and spinal cord. Minor efferent targets include zona incerta, raphe nuclei, bed nucleus of the stria terminalis, lateral preoptic area, lateral hypothalamus, and arcuate nucleus of the hypothalamus. Afferent neural connections to the SFO come from many of the same areas that receive efferent connections. However, such neural inputs are not extensive, which implies that the majority of sensory input controlling the excitability of SFO neurons is derived from signaling molecules in the systemic circulation.

Organum Vasculosum of the Lamina Terminalis

The OVLT is a midline structure on the anteroventral aspect of the third cerebral ventricle, ventral to the medial preoptic nucleus and dorsal to the optic chiasm. The densest efferent projections from the OVLT are monosynaptic and polysynaptic connections to the magnocellular neurons of the PVN and SON. Additional minor projections target CRH neurons of the PVN, the stria medullaris, and basal ganglia. The strongest afferent connections are received from medial preoptic nucleus, SFO and nucleus tractus solitarius (NTS), whereas minor connections include ventromedial nucleus, arcuate nucleus, and anterior, posterior, and dorsal hypothalamus.

Area Postrema

In rodents and lagomorphs, the AP is a midline structure protruding into the fourth ventricle on the dorsal surface of the medulla, immediately adjacent to the NTS. In most other mammals, the AP is a bilateral structure projecting into the fourth ventricle on either side of the obex. The AP sends dense efferent connections to many of the important autonomic control centers in the hindbrain, including the NTS and lateral parabrachial nucleus (l-PBN). There are also minor connections to the dorsal motor nucleus of the vagus (DMV), nucleus ambiguus, and dorsal tegmental nuclei. The AP receives afferent information from the NTS and l-PBN as well as the PVN and vagus nerve. Importantly, the vagus nerve also provides afferent input to the AP, thus providing sensory information from the viscera and positioning the AP as a unique medullary multimodal sensory integration area.

Properties and Sensory Abilities of CVO Neurons

Much work has described the electrophysiological properties of sensory CVO neurons, with SFO and AP neurons receiving the majority of the attention. In all three sensory CVOs, many neurons fire spontaneous action potentials (firing frequencies usually less than 1 Hz *in vivo*), whereas a subpopulation is typically quiescent. SFO and AP neurons exhibit compact dendritic trees and exhibit high input resistances ($>1\,G\Omega$), observations which are consistent with a prominent role of these cells in detecting humoral inputs as opposed to receiving synaptic input. Voltage-gated ionic currents in SFO and AP neurons are beginning to be thoroughly investigated, a task made easier in recent years with the identification of specific antagonists for genetically identified ion channels. Sensory CVO neurons possess a variety of ion channels that confer electrical excitability, including voltage-gated Na^+ (transient and persistent), Ca^{2+}, K^+ (delayed and transient) currents, and hyperpolarization-activated cation currents (I_h). Of specific interest in sensory CVO neurons, however, are nonspecific cation currents (NSCCs); activation of this type of conductance by circulating regulatory hormones such as Ang II, orexin, and ghrelin, as well as by changes in essential circulating ions such as Ca^{2+}, is often an endpoint leading to changes in membrane potential and thus firing rates. Interestingly, some of these signaling molecules control excitability by modulation of multiple ionic conductances, as illustrated by the actions of Ang II, which modulates transient voltage-gated K^+ currents and Ca^{2+} currents in CVO neurons in addition to its activation of an NSCC. A comprehensive description of all examples of ionic currents that are modulated by hormones and factors related to fluid balance is beyond the scope of this article. However, sensory CVO neurons express a number of receptors for hormones and factors that are directly and indirectly related to fluid balance, including receptors for adrenomedullin, Ang II, amylin, aldosterone, atrial natriuretic peptides, calcium, calcitonin, endothelin, estrogen, ghrelin, glucagon like peptide (GLP)-1, interleukin-1β, orexin, prokineticin 2, relaxin, vasopressin, and vasoactive intestinal peptide. Activation of these receptors by their ligand has in many cases been shown to modulate voltage-gated ion currents (and thus patterns of activity) in sensory CVO neurons.

This modulation of activity represents the mechanism by which information is transduced from the circulation to neural code in the CNS.

Angiotensin II Sensing

The identification in the 1970s of the SFO as the primary CNS site at which circulating Ang II acted to cause drinking first focused attention on the SFO and on the integrative roles of the three 'sensory' CVOs more generally. The initial observations were followed by studies demonstrating dense aggregations of AT_1 receptors in SFO, OVLT, and AP and (by single unit recordings) subpopulations of Ang II-sensitive neurons in all three CVOs. Studies using patch clamp techniques have identified potassium, calcium, and nonselective cationic conductances as potential channels modulated by Ang II, which thus may be responsible for inferring Ang II sensitivity of these neurons.

Sodium Sensation

The SFO and OVLT can detect changes in the concentration of sodium in the extracellular fluid independent of the osmolality. A logical potential signal is the Na^+ ion, whereby high levels of Na^+ can act in the brain as an inhibitory signal for sodium appetite. The identity of the necessary sodium receptor was elucidated in 2000 when tissue distribution of a poorly understood voltage-gated Na^+ channel (called Na_x) was closely examined, revealing a prominent expression in SFO, OVLT, ME, and NH. This restriction of Na_x to areas lacking a BBB implied a possible involvement with regulation of water/salt balance. Electrophysiological studies revealed that the sensor is a Na^+-gated Na^+ channel. Experiments revealed that Na^+ sensitivity was lost in SFO cells of Na_x knockout mice and these mice did not stop ingesting salt when dehydrated. These behavioral defects could be rescued with viral-mediated gene transfer (replacement) of the Na_x channel. Although early work suggested that the Na_x channel was expressed in neurons of the CVOs, subsequent studies have suggested that, at least in mice, Na_x expression is restricted to perineuronal processes of astrocytes in the CVOs, for the first time suggesting that the glial cell is likely to be a very active partner in the mechanisms through which the CVOs detect at least some extracellular signals. How the signals are then transduced from the glia to the neurons has yet to be elucidated.

Osmosensitivity

In addition to Na^+ concentration, the SFO and OVLT can detect changes in the osmolality of plasma. Many SFO and OVLT cells increase or decrease action potential firing rates with increasing and decreasing osmolality, respectively. Current evidence suggests that the osmosensor might be an ion channel (or heteromultimer of channels) of the transient receptor potential (TRP) family, the activation of which results in the NSCCs that have been recorded. These channels are thought to be mechanically gated, such that they open and close in response to changes in pressure on the cell membrane associated with osmotically driven fluid movement. Thus, physical stress on the membranes associated with changing extracellular osmolality modulates activity of these channels, resulting in changes in membrane conductances and changes in action potential frequency. Changes in extracellular osmolality outside the BBB may be transduced to brain centers protected by the BBB.

Calcium Sensation

Calcium is an essential factor in many biological processes, and its concentration in both the extracellular and the intracellular compartments must be maintained within strict limits. For example, calcium entry is required for synaptic release of neurotransmitters and induction of some forms of synaptic plasticity, and changes in $[Ca^{2+}]_i$ may lead to activation of intracellular kinases. Increases or decreases in $[Ca^{2+}]_{ext}$ lead to decreases or increases, respectively, in parathyroid hormone (PTH) release. PTH is a negative feedback hormone released from the parathyroid gland that stimulates Ca^{2+} retention by the kidney, uptake from the gut, and release from bone. The Ca^{2+} sensing receptor (CaR) responsible for detecting $[Ca^{2+}]_{ext}$ has been identified and located in the parathyroid gland, kidney, bone, intestine, and selected sites within the brain. Although the function(s) in brain has not been firmly established, CaR is highly expressed in both neurons and oligodendrocytes of SFO and AP. Activation of this receptor has been shown to increase the excitability of SFO neurons through the modulation of specific ionic conductances (NSCC and I_h). Such observations suggest that SFO may play important roles in the physiological regulation of $[Ca^{2+}]_{ext}$, presumably as a consequence of the ability of these neurons to regulate appropriate autonomic outputs through connections with hypothalamic and medullary autonomic control centers.

Secretory CVOs

The two secretory CVOs that play essential roles in the regulation of fluid and mineral balance are the ME and NH. They represent the 'exit portals' through which neuroendocrine cells of the hypothalamus can secrete their hormones into the systemic circulation for delivery to effector tissues located throughout the body.

Vasopressin and oxytocin are secreted primarily from the NH, which is primarily viewed as a secretory structure composed of axon terminals, glial support cells, and the vasculature supplying and draining the region. There is also limited release of the neurohypophysial peptides from the ME, which is the primary site for other hypophysial hormones of varying importance in the regulation of fluid and mineral balance. These include corticotropin releasing harmone (CRH) (controlling the adrenocortical axis), thyrotropin releasing harmone (TRH) (the thyroid axis), dopamine (prolactin secretion), gonadotropin releasing harmone (GnRH) (reproductive axis), and growth harmone releasing harmone (GHRH)/somatostatin (growth). There likely are important roles for local and circulating factors in regulation at the axon terminal of the processes in the NH and ME through which hypophysial hormones are released.

Roles of CVOs in Osmotic/Mineral Challenges

Water in the body is maintained within two compartments, the extracellular and the intracellular. NaCl is the most important solute, and its concentration in the extracellular compartment primarily determines the distribution of water. Perfusion of the tissues within the body (supplying oxygen and removing metabolites via the circulation) critically depends on the maintenance of fluid balance. Loss of water through evaporation from the skin and lungs, and through excretion in urine and/or feces, usually results in loss from both intracellular and extracellular compartments, whereas vomiting, diarrhea, and hemorrhage result initially in water loss from only the extracellular compartment. The latter conditions also result in significant solute loss. Complex compensatory responses minimize further losses through the skin and urine while modulating cardiovascular tone and output to ensure adequate tissue perfusion. However, these responses do not restore lost water and NaCl. Detection of the fluid/salt imbalance ultimately must lead to activation of neural circuits invoking motivational states of thirst and salt appetite such that lost constituents can be replenished and normal parameters restored. The sensory CVOs have been shown to play essential roles in this process, detecting the perturbations and initiating the integrated physiological response that includes fluid intake, fluid retention, and fluid redistribution until normal balance is restored.

CVOs in the Regulation of Thirst by Intracellular Dehydration

More than 50 years ago, Verney conducted experiments in which he demonstrated release of vasopressin from the posterior pituitary in dogs that had received infusions of hypertonic solutions of NaCl and sucrose into the cerebral arteries. However, a similar infusion of hypertonic urea failed to stimulate vasopressin release. These observations led him to propose that the hypertonic stimulus produced with NaCl or sucrose caused a movement of water out of the cells in the brain, thus activating CNS osmoreceptors and stimulating vasopressin release. He proposed that urea, which rapidly diffuses across cell membranes and does not cause osmotic water movement, did not stimulate the osmoreceptors. Later studies also reported that NaCl and sucrose infusions also caused antidiuresis and drinking, whereas urea was ineffective, again implying that cellular dehydration was an essential prerequisite for the initiation of these responses. Whereas early studies suggested that OVLT lesions blocked drinking and vasopressin secretion in response to hypertonic saline infusion, more recent work has suggested partial roles for both SFO and OVLT in mediating these responses. These studies have also clearly demonstrated that double lesions destroying both SFO and OVLT in the same animal do abolish drinking and vasopressin secretion in response to hypertonic NaCl.

CVOs in the Regulation of Thirst by Hypovolemia

Hypovolemia activates the renin–angiotensin system, resulting in increased circulating concentrations of the hormone Ang II. Ang II has multiple coordinated and convergent actions to regulate water and salt balance. These include direct actions on vascular smooth muscle cells causing arteriolar vasoconstriction, which elevates systemic vascular resistance and therefore increases the systemic blood pressure. Ang II also acts directly on the kidney to stimulate Na^+ reabsorption in the early proximal tubule, and it acts in the adrenal cortex to stimulate release of the mineralocorticoid aldosterone, which in turn causes Na^+ and water reabsorption by the kidney.

Actions of Ang II on CVOs to regulate thirst are direct and indirect. In 1969, Ang II was shown to stimulate thirst, and in 1978 the SFO was identified as the primary CNS site at which circulating Ang II acted to induce such responses (although central administration of Ang II appeared to elicit similar effects through actions at the OVLT). Later studies demonstrated Ang II binding and receptors in the sensory CVOs. Remarkably, Ang II is also found in the brain, where it is apparently used as a neurotransmitter (neuropeptide) in a number of pathways involved with fluid balance. Ang II also activates SFO and OVLT neurons, which send efferent projections to SON and PVN neurosecretory cells, thus stimulating release of vasopressin, oxytocin, and CRH into the circulation through the

posterior pituitary and median eminence (secretory circumventricular organs). These actions have the net effect of reestablishing normal fluid volume.

AP neurons also have the ability to sense changes in circulating Ang II, as well as vasopressin. In addition, they receive direct neural information from the cardiac and arterial baroreceptors. Through efferents to the NTS, DMV, nucleus ambiguus, and l-PBN, AP neurons influence baroreflex sensitivity, heart rate, and sympathetic output, which are essential to an integrated response to hypovolemia.

CVOs in the Regulation of Salt Appetite

Hypovolemia also activates salt appetite (to restore lost solute), albeit on a slower timescale than thirst. Ang II and aldosterone play roles in inducing salt appetite during hypovolemia, perhaps acting synergistically. The possible mechanism of the synergism is quite elegant: aldosterone may act via a nuclear receptor to increase expression of the AT_1 subtype of receptor protein in the SFO, leading to a potentiated response of SFO neurons to Ang II and thus increased salt appetite. Available data also suggest an inhibitory role of AP in salt appetite because AP lesions enhance NaCl solution intake.

CVOs and the Gustatory Aspects of Salt Appetite

The sensory CVOs are thought to play a role in salt appetite associated with gustation. Salt appetite caused by sodium depletion can be sated (temporarily) by the presence of Na^+ on the tongue. Normally this taste elicits an aversive reaction, accompanied by elevated activity of chorda tympani fibers projecting to NTS. However, sodium-deprived rats exhibit decreased responses to NaCl, presumably associated with an inhibition of the aversive reaction. Lesions of the NTS and PBN block the arousal of the salt appetite, indicating the vital role of these areas in the motivation. AP has dense reciprocal connections with NTS and PBN, whereas SFO and OVLT also project to NTS, and it has thus been suggested that modulation of the sensory CVOs by Ang II or other signals may modulate gustatory sensory pathways in the brain stem.

CVOs as Controllers of Integrated Autonomic Function

The previous discussion implies that in addition to their unique position at the blood–brain interface, the sensory CVOs combine broad-ranging sensory abilities with the capability to exert control (through their efferent projections to autonomic control centers) over diverse yet interrelated physiological systems that collectively comprise the integrated control of body fluid balance. The large number of CVO neurons influenced by multiple circulating signals (25–65%) implies that single neurons likely have multiple sensory abilities. Significant proportions of neurons in the three sensory CVOs are also sensitive to circulating signals controlling immune function, energy homeostasis, and reproduction. It is not clear whether single CVO neurons sense multiple signals related to such diverse physiological systems; for example, are amylin- and glucose-sensitive AP neurons also sensitive to Ang II and/or vasopressin? Nonetheless, a picture has begun to emerge of CVO neurons sensing multiple signals and thus potentially controlling in an integrated manner broad physiological outputs. In other words, the 'sensory CVOs' should be considered in a broader context than simply as structures involved in the regulation of body fluid balance.

See also: Angiotensin II; Circumventricular Organs in Neuroendocrine Control; Food and Water Intake: Regulation; Osmoregulation; Salt Appetite; Thirst.

Further Reading

Borison HL (1989) Area postrema: Chemoreceptor circumventricular organ of the medulla oblongata. *Progress in Neurobiology* 32: 351–390.

Brooks VL and Osborn JW (1995) Hormonal–sympathetic interactions in long-term regulation of arterial pressure: An hypothesis. *American Journal of Physiology* 268: R1343–R1358.

Cottrell GT and Ferguson AV (2004) Sensory circumventricular organs: Central roles in integrated autonomic regulation. *Regulatory Peptides* 117: 11–23.

Ermisch A, Ruhle HJ, Landgraf R, and Hess J (1985) Blood–brain barrier and peptides. *Journal of Cerebral Blood Flow & Metabolism* 5: 350–357.

Ferguson AV (1991) The area postrema: A cardiovascular control centre at the blood–brain interface? *Canadian Journal of Physiology and Pharmacology* 69: 1026–1034.

Ferguson AV and Bains JS (1996) Electrophysiology of the circumventricular organs. *Frontiers in Neuroendocrinology* 17: 440–475.

Fitzsimons JT (1998) Angiotensin, thirst, and sodium appetite. *Physiological Reviews* 78: 583–686.

Gross PM (1992) Circumventricular organ capillaries. *Progress in Brain Research* 91: 219–233.

Johnson AK, Zardetto-Smith AM, and Edwards GL (1992) Integrative mechanisms and the maintenance of cardiovascular and body fluid homeostasis: The central processing of sensory input derived from the circumventricular organs of the lamina terminalis. *Progress in Brain Research* 91: 381–393.

McKinley MJ, Allen AM, Burns P, Colvill LM, and Oldfield BJ (1998) Interaction of circulating hormones with the brain: The roles of the subfornical organ and the organum vasculosum of the lamina terminalis. *Clinical and Experimental Pharmacology and Physiology Supplement* 25: S61–S67.

McKinley MJ, Mathai ML, McAllen RM, et al. (2004) Vasopressin secretion: Osmotic and hormonal regulation by the lamina terminalis. *Journal of Neuroendocrinology* 16: 340–347.

McKinley MJ, McAllen RM, Davern P, et al. (2003) The sensory circumventricular organs of the mammalian brain. *Advances in Anatomy, Embryology, and Cell Biology* 172: III–122.

Meisenberg G and Simmons WH (1983) Minireview. Peptides and the blood–brain barrier. *Life Sciences* 32: 2611–2623.

Thunhorst RL (1996) Role of peripheral angiotensin in salt appetite of the sodium-deplete rat. *Neuroscience and Biobehavioral Reviews* 20: 101–106.

Weindl A and Sofroniew MV (1981) Relation of neuropeptides to mammalian circumventricular organs. *Advances in Biochemical Psychopharmacology* 28: 303–320.

Circumventricular Organs in Neuroendocrine Control

A K Johnson, University of Iowa, Iowa City, IA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Certain specialized brain structures have as their major distinguishing feature a dense vascularization which usually lacks the structural components that comprise the blood–brain barrier. In most of the white and gray matter of the central nervous system, capillary endothelial cells form an obstruction that restricts access of many blood-borne ions and molecules to brain parenchyma. In addition, glial cells tightly wrap the outer surface of these vessels to retard access to surrounding neurons. In contrast to typical brain tissue, most of these organs have fenestrated capillaries, and the surrounding glia are relaxed to form large perivascular spaces to allow ready exchange between the blood and interstitial fluid.

Neuroanatomic reports throughout the 1910s and 1920s frequently noted the absence of a blood–brain barrier in several discrete brain locations. After the systemic administration of vital dyes, staining was absent from most of the brain except for several small regions located in areas around the cerebral ventricles and in certain midline structures lying in the subarachnoid space. By the mid-1950s, the body of anatomic findings describing this collection of regions lacking a blood–brain barrier led Hofer to apply the title circumventricular organs (CVOs) to categorize these structures.

The number of CVOs described in amphibians, reptiles, birds, and mammals varies across vertebrate classes and species. For example, in birds, one author has identified 17 CVOs. The general consensus is that mammals have 7–9 CVOs. The areas often referred to as mammalian CVOs are presented in Figure 1. Those most consistently agreed on are the subfornical organ (SFO), organum vasculosum of the lamina terminalis (OVLT), the median eminence (ME), the neural lobe (NL), the pineal gland (PI), and the area postrema (AP). Many, but not all, authors include the intermediate lobe (IL) and the choroid plexus (CP). Although the subcommissural organ (SCO) is usually listed as a CVO, many qualify this structure as unique from the others because it lacks fenestrated capillaries and thereby has an intact blood–brain barrier.

Although there is a 'defect' in the blood–brain barrier associated with most CVOs, they still have properties that retard access from their parenchyma to other brain tissues and the cerebrospinal fluid (CSF). In essence, barrier properties are displaced to the periphery of the organs, and materials gaining access to CVOs do not readily diffuse from the organ into adjacent brain structures. In the region where CVOs and 'normal' brain parenchyma interface, the outward spread from the CVO is impeded. The structural feature giving rise to this limitation of diffusion appears to be a dense array of glia. In addition, tight junctions between the ependymal cells covering the surface facing the ventricles also limit the transfer of agents from CSF to the organ and vice versa. Hence under normal physiologic conditions, CVOs should not be viewed as portals of entry for bulk dissemination of blood-borne agents to the brain or CSF. Rather, CVOs appear as restricted regions allowing for limited localized exchange between blood and brain.

Functional Roles for CVOs

CVOs have been proposed to have a role in many homeostatic and nonhomeostatic functions. Some of these include the maintenance of body fluid balance, blood pressure, temperature, respiration, energy balance, mediation of immune and host-defense responses, pain modulation, protection against toxic substances (e.g., emesis and conditioned taste), biological rhythms, reproduction, parental behavior, lactation, growth, sleep, arousal, and attention. Although some investigators have argued for a global, coordinated action among several or all CVOs for a specific function, others believe it is more likely that each CVO has a unique role and that at most, only small subsets of organs are functionally linked.

CVOs are strategic loci that couple the transfer of information among the CSF, blood, and brain. The possible types of information exchange between these three compartments are summarized in Table 1. The experimental evidence supporting each of these eight classes of interaction varies from substantial to minimal. Evidence for Class I exchanges (e.g., secretion of vasopressin) is the best established, with Classes II and III somewhat less established. Even less functional evidence is currently available for the remaining types of CVO, CSF, and brain interactions.

The function of CVOs for which there is the most general and longest agreement is the role in neurosecretion. The early discoveries of endocrine functions of the NL, ME, PI, and pituitary IL have firmly established this role for these CVOs. A well-described secretory role for the CP and the SCO

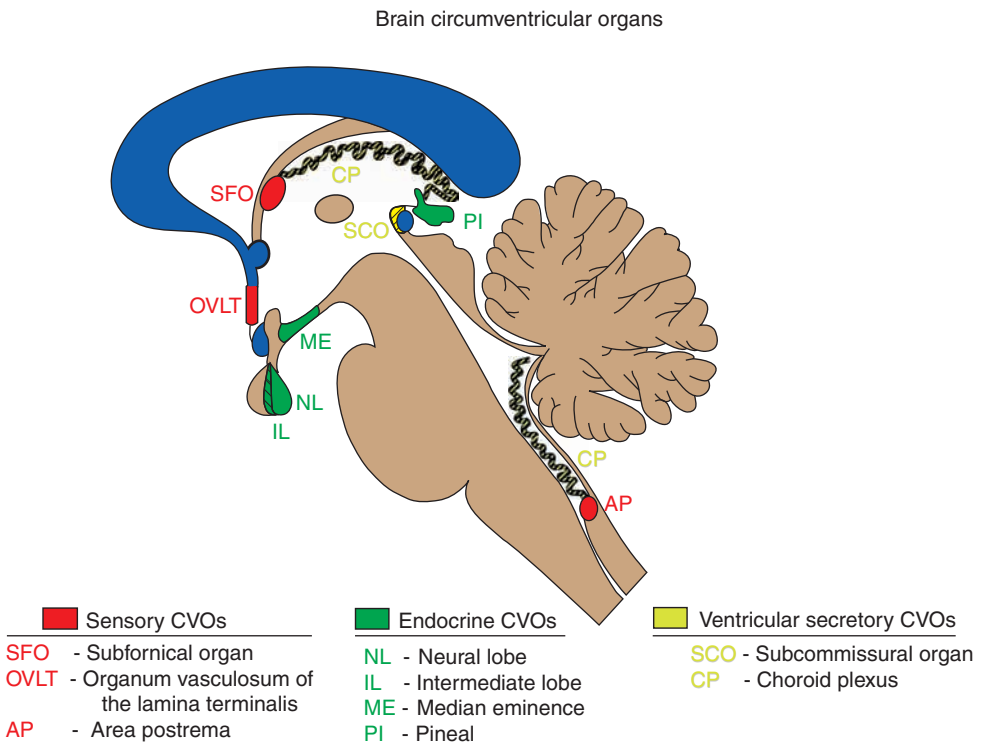

Figure 1 A medial sagittal representation of the human brain depicting nine circumventricular organs (CVOs). The organs have been classified according to their putative primary functions.

Table 1 Classes of hypothesized interactions of circumventricular organs (CVOs) and mode of mediation

Class	*Signal source, target*	*Mode of information transfer*
I	CVO, blood	Humoral
II	Blood, CVO	Humoral
III	CVO, brain parenchyma	Neural
IV	Brain proper, CVO	Neural
V	CVO, CSF[a]	Humoral
VI	CSF, CVO	Humoral
VII	CVO, brain proper	Humoral
VIII	Brain proper, CVO	Humoral

[a]CSF, cerebrospinal fluid.
Adapted from Johnson AK and Gross PM (1993) Sensory circumventricular organs and brain homeostatic pathways. *FASEB Journal* 7: 678–686.

also exists. It is recognized that the CP is the primary source of CSF. The SCO secretes a mucoidlike thread, Reissner's fiber, into the cerebral aqueduct. However, the functional significance, if any, of Reissner's fiber is unknown.

For the remaining three CVOs, the SFO, OVLT, and AP, only a small amount of evidence suggests that they also may secrete materials into either blood or CSF. However, recent research has been more successful in implicating these particular CVOs as targets for blood-borne factors such as hormones or humoral agents (i.e., specific ions or nonspecific osmols). In essence, the SFO, OVLT, and AP constitute a set of sensory CVOs.

The SFO, OVLT, and AP have been especially implicated in the control of body fluid and cardiovascular homeostasis. Findings made in the late 1960s provided evidence that angiotensin II (ANG II) acts on the AP to reduce the gain of baroreceptor-mediated cardiovascular reflexes. These findings were followed by studies implicating the SFO in the control of angiotensin-induced water intake (thirst), vasopressin release, sodium chloride ingestion (salt appetite), and cardiovascular responses. Stimulation and ablation studies of the periventricular region that includes the OVLT indicate that this CVO houses a critical set of the body's osmoreceptors or sodium receptors that influence thirst, vasopressin release, salt appetite, and arterial blood pressure.

Biochemical Properties and Anatomic Connections

Biochemical and immunocytochemical studies have identified many hormonal and neurotransmitter-like factors in CVOs. The CVOs that are primarily neurosecretory contain neurohormones (e.g., vasopressin, adrenocorticotrophic hormone, and melatonin)

that, depending on the specific hormone and CVO, are either synthesized *in situ* or transported to release sites. Sensory CVOs also contain an extensive number of hormonal and neurotransmitter-like factors, including many neuropeptides and biogenic amines. Although some substances may be concentrated in CVOs as a result of extraction from blood or CSF, it is likely that many are transported via axons into the CVOs from other brain regions or that they are synthesized locally; several synthetic enzymes have also been localized in sensory CVOs.

Support for the concept that sensory CVOs are targets for extracellular agents is demonstrated by high concentrations of ligand-binding sites, immunocytochemically identified receptor proteins, and messenger RNAs for the receptors of a number of putative neurotransmitters and hormones. Some of the binding sites or receptors identified in the SFO, OVLT, and AP include those for a number of peptides (e.g., ANG II, atrial natriuretic peptide, calcitonin, endothelin, cholecystokinin, insulin, somatostatin, substance P, vasopressin, various opioids, neuropeptide Y, and vasoactive intestinal peptide), biogenic amines (e.g., catecholamines and serotonin), amino acids (e.g., γ-aminobutyric acid and glycine), and several steroids (e.g., aldosterone, estrogen, glucocorticoids, and progesterone). Evidence from both *in vivo* and *in vitro* electrophysiologic studies indicates that many of the substances with receptor or binding sites in sensory CVOs evoke voltage-related responses in cells within these organs.

Neurons with cell bodies located in brain parenchyma and axons that project and terminate in the NL, ME, IL, or PI are common paths for the release of neurohormones either into the general circulation (e.g., NL or PI) or into specific portal systems (e.g., ME to anterior pituitary). The somata of many of the cells that project into secretory CVOs are points of convergence for multiple types of neural inputs. These inputs typically arise from diverse brain regions, including hypothalamic nuclei, other limbic structures, and brain stem regions. It is also notable that among the CVOs with known neuroendocrine function, the control of melatonin synthesis and secretion in the PI is exerted by sympathetic innervation arising from the superior cervical ganglia.

Recent progress applying refined neuronal track-tracing methods has provided substantial insight into the relationship of sensory CVOs with the rest of the nervous system. The AP, SFO, and OVLT all have a rich distribution of efferents. Notably, many of the targets receiving projections from sensory CVOs are

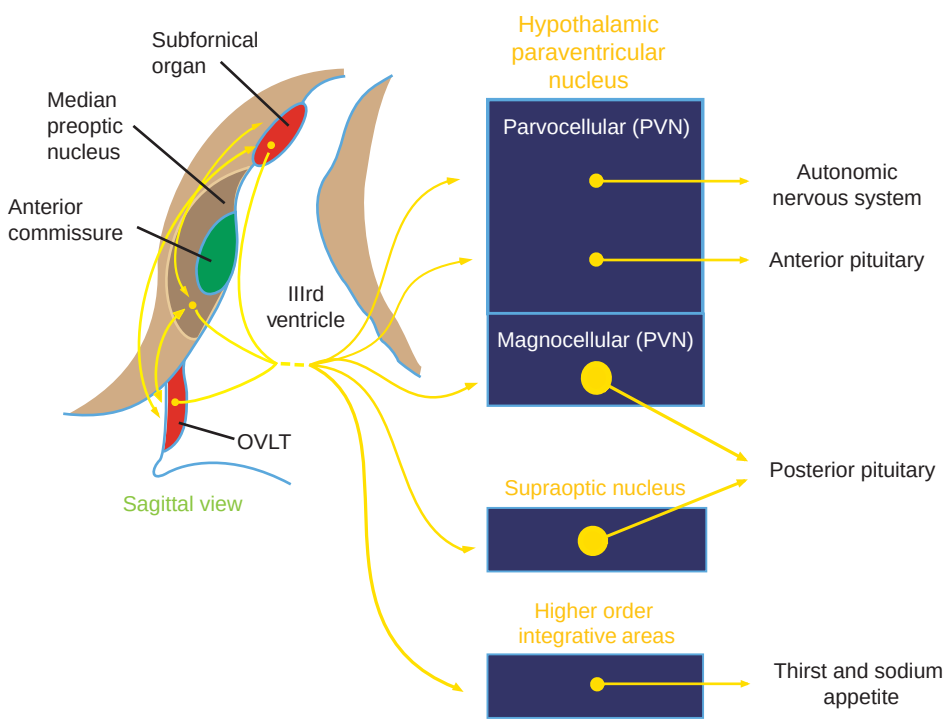

Figure 2 Sensory circumventricular organs (CVOs) associated with the lamina terminalis and descending projections to forebrain regions involved in autonomic, neuroendocrine, and behavioral control of body fluid and cardiovascular homeostasis. OVLT, organum vasculosum of the lamina terminalis. Adapted from Johnson AK and Gross PM (1993) Sensory circumventricular organs and brain homeostatic pathways. *FASEB Journal* 7: 678–686.

brain nuclei that receive input from visceral afferent nerves and from other brain regions implicated in homeostatic processes such as the maintenance of body fluid and energy balance. In essence, sensory CVOs project into an extensive visceral neural network that has been referred to as the visceral neuraxis. This neural network processes (integrates) input related to the control of homeostatic response systems. Outputs from this network control behavioral, autonomic, and neuroendocrine responses to perturbations in homeostatic states.

An example of the relationship between sensory and neurosecretory CVOs and with other effector systems is provided by the relationship of the outflow of the two forebrain sensory CVOs (i.e., SFO and OVLT), along with their related neural circuitry in the lamina terminalis. The SFO and OVLT, along with the adjacent median preoptic nucleus, which is inside the blood–brain barrier, project to several structures implicated in the control of fluid balance and blood pressure. A model for the influence of the outputs from the lamina terminalis is presented in Figure 2. Substantial evidence indicates that in several species, the SFO is sensitive to circulating angiotensin and that the OVLT contains osmoreceptors. Descending pathways from these two CVOs and from the median preoptic nucleus project to multiple regions, influencing physiologic and behavioral mechanisms that maintain cardiovascular and body fluid homeostasis. Among these 'descending' projections are those that terminate in the region of cell bodies of (1) parvocellular paraventricular nucleus (PVN) neurons, which in turn project to the ME and to more-caudal structures involved in the control of sympathetic outflow; (2) magnocellular neurons of the PVN and supraoptic nuclei, which in turn project to the NL and release vasopressin (antidiuretic hormone) and oxytocin; and (3) regions (e.g., periformical hypothalamic region, central nucleus of the amygdala, and bed nucleus of the stria terminalis) implicated in the control of water and sodium intake. Extracellular signals such as ANG II or increased osmolarity reach the SFO- or the OVLT-activating pathways that alter activity of higher-order neurons controlling the secretion of hydration-related hormones, behaviors, and sympathetic outflow. It is interesting to note that a close relationship between the SFO/OVLT and neurosecretory neurons provides a short loop reflex for controlling hormone release. Although the relationship between sensory and secretory CVOs is best illustrated by the mechanisms that maintain body fluid homeostasis, it is likely that future research will fully implicate CVOs in a multitude of critical regulatory processes involved in health and disease.

See also: Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Central Cardiovascular Control; Blood Pressure: Baroreceptors; Cardiovascular Function: Central Nervous System Control; Circumventricular Organs; Immune System–Neuroendocrine Interactions; Neuroendocrine Control of Energy Balance (Central Circuits/Mechanisms); Neuroendocrinology; Sexual Behavior: Neuroendocrine Control.

Further Reading

Carithers JR and Johnson AK (1988) Fine structural studies of the effects of AV3V lesions on the hypothalamo-neurohypophyseal neurosecretory system. In: Cowley AW Jr., Liard JF, and Ausiello DA (eds.) *Vasopressin: Cellular and Integrative Functions*, pp. 301–319. New York: Raven Press.

Ermisch A, Landgraf R, and Rühle HJ (eds.) (1992) *Circumventricular Organs and Brain Fluid Environment: Progress in Brain Research*, vol. 91. Amsterdam: Elsevier.

Gross PM (ed.) (1987) *Circumventricular Organs and Body Fluids*, 3 vols. Boca Raton, FL: CRC Press.

Johnson AK and Loewy AD (1990) Circumventricular organs and their role in visceral functions. In: Loewy AD and Spyer K (eds.) *Central Regulation of Autonomic Regulation*, pp. 247–267. Oxford, UK: Oxford University Press.

Johnson AK and Gross PM (1993) Sensory circumventricular organs and brain homostatic pathways. *EASER Journal* 7: 678–686.

Johnson AK and Thunhorst RL (1997) The neuroendocrinology of thirst and salt appetite: Visceral sensory signals and mechanisms of central integration. *Frontiers in Neuroendocrinology* 18: 292–353.

Johnson AK and Thunhorst RL (2007) The neuroendocrinology, neurochemistry and molecular biology of thirst and salt appetite. In: Laitha A and Blaustein JD (eds.) *Handbook of Neurochemistry and Molecular Neurobiology: Behavioural Neurochemistry, Neuroendocrinology and Molecular Neurobiology*, 3rd edn., pp. 641–687. New York: Springer.

Lind RW and Ganten D (1990) Angiotensin. In: Bjorklund A, Hökfelt T, and Kuhar MJ (eds.) *The Handbook of Chemical Neuroanatomy, Vol. 9: Neuropeptides in the CNS*, pp. 165–286. Amsterdam: Elsevier.

McKinkey MJ, McAllen RM, Davern P, et al. (2003) The sensory circumventricular organs of the mammalian brain: Subfornical organ, OVLT and area postrema. *Advances in Anatomy, Embryology, and Cell Biology* 172: 1–122.

Miller AD and Leslie RA (1994) The area postrema and vomiting. *Frontiers in Neuroendocrinology* 15: 301–320.

Miselis RR, Hyde TM, and Shapiro EE (1986) Disturbances in water balance controls following lesions to the area postrema and adjacent solitary nucleus. In: de Caro G, Epstein AN, and Massi M (eds.) *The Physiology of Thirst and Sodium Appetite*, pp. 279–285. New York: Plenum.

North WG, Moses AM, and Share L (eds.) (1993) *The Neurohypophysis: A Window on Brain Function. Annals of the New York Academy of Sciences*, vol. 689. New York: New York Academy of Sciences.

CIRL/Latrophilins

S Sugita, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Discovery of CL1

CL1 was originally purified as α-latrotoxin binding protein in the absence of Ca^{2+} in efforts to elucidate the mechanisms of α-latrotoxin actions. α-Latrotoxin is a member of latrotoxins that are isolated from the venom of the black widow spider, *Lactrodectus*. Latrotoxins have been used to study the molecular mechanisms of neurotransmitter exocytosis in vertebrates, insects, and crustaceans. α-Latrotoxin is by far the most studied toxin since it is the only one that stimulates neurosecretion from neurons and neuroendocrine cells of vertebrates. α-Latrotoxin has been cloned and sequenced together with three other latrotoxins – α- and γ-latroinsectotoxins and α-latrocrustaceanotoxin. Similar to other latrotoxins, α-latrotoxin is a large protein composed of four domains: a cleaved signal peptide (domain I), a conserved N-terminal domain with three invariant cysteine residues among all the latrotoxins (domain II), a domain composed of 22 ankyrin-like repeats (domain III), and a C-terminal domain (domain IV) that is cleaved during venom maturation. Mature α-latrotoxins are composed of domains II and III, with the molecular weight of 130 kDa (Figure 1).

α-Latrotoxin has a complex mode of action: it stimulates neurotransmitter release from neurons in the presence or absence of extracellular Ca^{2+}, and it also stimulates neurotransmitter release from neuroendocrine cells, but only in the presence of Ca^{2+}. Even in Ca^{2+}-dependent transmitter release, the mechanisms of α-latrotoxin-stimulated exocytosis may differ between neurons and neuroendocrine cells. The Ca^{2+}-independent stimulatory action of α-latrotoxin in neurons has attracted attention because the elucidation of the underlying mechanism will likely lead to the discovery of novel triggering mechanisms of neurotransmitter release.

Regardless of the complex mode of action, α-latrotoxin stimulates exocytosis only after binding to neuronal cell surface receptors. The first receptor to be discovered was neurexin 1α, a neuronal protein with a single transmembrane domain. However, neurexin 1α binds to α-latrotoxin only in the presence of Ca^{2+}. Since α-latrotoxin also acts in the absence of Ca^{2+}, this clearly indicated the presence of another Ca^{2+}-independent toxin receptor. Such a protein was later discovered by two independent research groups and separately termed CIRL1 (Ca^{2+}-independent receptor for latrotoxin 1) and latrophilin 1. As a result, the term CL1, which incorporates both of these names, has been used to represent the protein.

Structure of CLs

CL1, which was purified from the brain by affinity purification with α-latrotoxin column, was originally estimated by SDS-PAGE analysis to be 120 kDa in size. However, cDNA cloning of CL1 revealed the actual size of this protein to be greater than 160 kDa. After careful analysis, it has become clear that CL1 goes through proteolytic processing into two subunits: an N-terminal 120 kDa subunit (p120) and a C-terminal 85 kDa subunit (p85) (Figure 2). The 120 kDa protein purified by α-latrotoxin column is the p120 subunit of CL1.

Immediately after cloning of CL1, two isoforms of CL1 were cloned and termed CL2 and CL3. Although CL2 and CL3 are highly homologous, only CL1 binds tightly to α-latroxin. CL2 may bind weakly to α-latrotoxin, whereas CL3 has no affinity to the toxin. As G-protein-linked receptors, CLs are unusually large (Figure 2). The central part of CLs is occupied by seven transmembrane regions (TMRs) and their connecting loops. The TMRs of CLs are homologous to TMRs of other serpentine receptors, particularly those of the secretin family of peptide hormone receptors. On the N-terminal extracellular side, the TMRs are preceded by at least six domains: (1) a signal peptide; (2) a short cysteine-rich sequence (~90 residues) distantly related to lectins; (3) a longer domain (~290 residues) that is homologous to olfactomedins and the glaucoma gene myocilin; (4) a serine- and threonine-rich linker sequence that may be glycosylated; (5) a long domain (~280 amino acids) that exhibits homology only to BAI1, -2, and -3, which are large serpentine receptors of unknown function characterized as potential inhibitors of brain angiogenesis; and (6) a short (~50 residues) cysteine-rich domain that is located immediately N-terminal of the first TMR and represents a novel type of widely distributed extracellular protein module. The precise site of proteolytic cleavage is identified within this cysteine-rich domain, and this domain is called the G-protein-coupled receptor proteolytic site (GPS) domain (Figure 2).

CLs belong to the group of large N-terminal family B seven-transmembrane (LNB 7TM) receptors, which is composed of at least 33 members, as identified by bioinformatics. All 33 proteins contain the GPS domain, and at least three LNB 7TM receptors, including CL1, appear to be cleaved within this site.

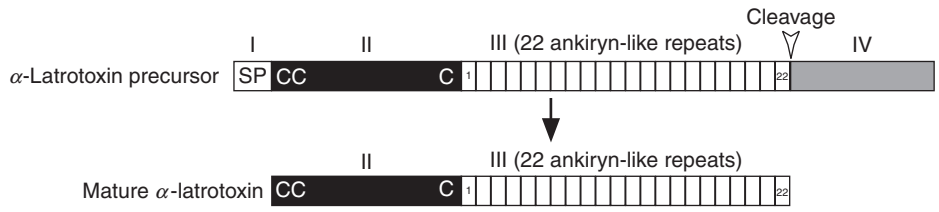

Figure 1 Schematic representation of α-latrotoxin. Composition of α-latrotoxin precursor of four domains: I, signal peptide (SP); II, conserved N-terminal domain. Locations of the only three cysteine residues that are conserved in all latrotoxins are indicated by the letter 'C'; III, central region, composed of 22 ankyrin-like repeats; IV, C-terminal region. α-Latrotoxin is processed by endoproteolytic cleavage after the last ankyrin repeat (open arrowhead). Thus, mature α-latrotoxin is composed of two domains (domains II and III).

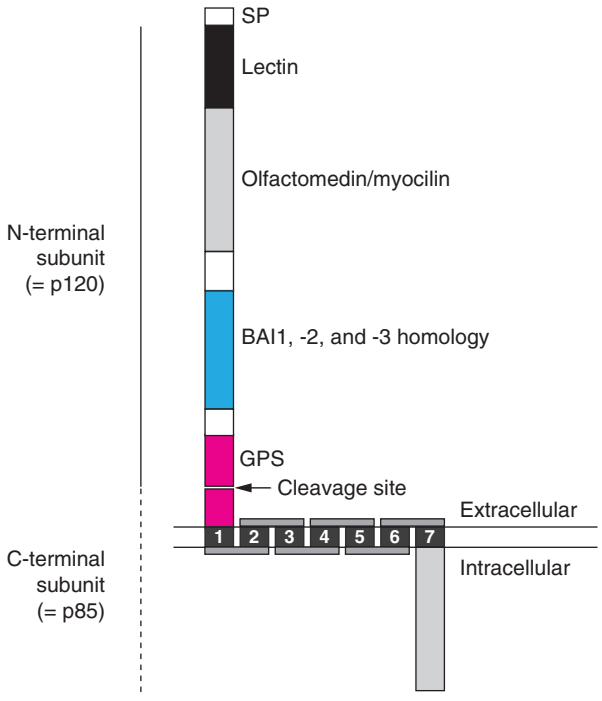

Figure 2 Domain structure of CLs. SP, N-terminal signal peptide; lectin, a lectin-like domain; olfactomedin/myocilin, a domain homologous to olfactomedins and myocilin; BAI1, -2, and -3 homology, a domain homologous to BAI1, -2, and -3; GPS, a cysteine-rich domain that contains a cleaving site. CL1 is cleaved within the GPS domain into two subunits – an N-terminal subunit (p120) and a C-terminal subunit (p85). These two subunits remain associated by noncovalent interactions after cleavage. The C-terminal subunit is a seven transmembrane protein coupled with G-protein.

This observation suggests that the rest of the proteins in the family could also be cleaved at this specific site. Cleavage occurs in the endoplasmic reticulum within the cell, with the N-terminal subunit (p120) remaining associated with the seven transmembrane C-terminal subunit (p85) (Figure 2). The nature and specificity of the cleavage enzymes remain unknown. The biological significance of the cleavage also remains to be fully understood. The mutagenesis in the GPS domain of CL1 that blocks proteolytic cleavage also blocks CL1 trafficking to the cell surface, which indicates that cleavage is critical for delivery of CL1 to the cell surface.

CL1 Is a Major α-Latrotoxin Receptor

The critical role for CL1 in α-latrotoxin action was demonstrated through the phenotypic analysis of CL1 knockout mice. Ca^{2+}-independent binding of α-latrotoxin to brain membranes is impaired in CL1 knockout mice by approximately 75% and Ca^{2+}-dependent binding (approximately two times higher than Ca^{2+}-independent binding) is decreased in CL1 knockout mice by approximately 50%. The binding of α-latrotoxin to the brain membrane is two times greater in the presence of Ca^{2+} than that in the absence of Ca^{2+} since α-latrotoxin can also bind neurexin in the presence of Ca^{2+}. In neurexin 1α knockout mice, Ca^{2+}-dependent binding of α-latrotoxin to brain membranes is decreased by approximately 50%, whereas Ca^{2+}-independent binding of the toxin is not affected, supporting the biochemical finding that binding between neurexin 1α and α-latrotoxin occurs only in the presence of Ca^{2+}.

Synaptosomes lacking CL1 exhibited the same decrease in α-latrotoxin-stimulated glutamate release in the presence and absence of Ca^{2+} (approximately 75%). In contrast, synaptosomes lacking neurexin 1α exhibited only a small decrease in α-latrotoxin-triggered release in the absence of Ca^{2+} (approximately 20%) but a major decrease in the presence of Ca^{2+} (approximately 75%). These findings suggest that CL1 and neurexin 1α account for the majority of α-latrotoxin receptors in the brain and any remaining receptor activity could be accounted for by other CL and neurexin isoforms. They also suggest that CL1 and neurexin 1α act additively in binding α-latrotoxin but not in triggering release. Together, these two receptors act autonomously in binding of α-latrotoxin but cooperatively in transducing the stimulation of neurotransmitter release by α-latrotoxin. However, the underlying mechanisms of the cooperativity between CL1 and neurexin 1α remain to be fully understood.

Role for CL1 in α-Latrotoxin-Stimulated Neurotransmitter Release in the Presence of Extracellular Ca^{2+}

In normal physiological circumstances, 1 or 2 mM of Ca^{2+} is present in the extracellular environment and the binding of α-latrotoxin to neuronal cells is largely mediated by its binding to both CL1 and neurexin 1α. How does α-latrotoxin stimulate neurotransmitter release after binding to CL1 and/or neurexin 1α? The mechanisms underlying α-latrotoxin-induced, extracellular Ca^{2+}-dependent neurotransmitter release in neuroendocrine cells are well established. The key observation is that overexpression of CL1 or neurexin 1α in neuroendocrine cells facilitates α-latrotoxin-induced, extracellular Ca^{2+}-dependent neurotransmitter release that does not rely on intracellular domains of CL1 or neurexin 1α. Based on this finding, it is believed that the α-latrotoxin cell surface receptor is important for recruiting the toxin to the site of action but is not involved in the exocytotic signal, which is transmitted by the toxin. The toxin has long been known to form Ca^{2+}-permeable pores in the plasma membrane of cells sensitive to it. By expressing exogenous toxin receptors (CL1 or neurexin 1α) in cells that are normally insensitive to α-latrotoxin (COS-7, HEK293, or BHK), the toxin easily induced pores. Furthermore, signaling-deficient (intracellularly truncated) mutants of the receptors also facilitated pore formation that correlated with toxin binding irrespective of the receptor structure. Thus, pore formation is the major mechanism by which α-latrotoxin triggers Ca^{2+}-dependent neurotransmitter release in neuroendocrine cells and does not rely on signal transduction by receptors.

How is the hydrophilic α-latrotoxin able to insert itself into the membrane? The three-dimensional structure of α-latrotoxin has been determined and toxin tetramers inserted into the membranes of liposomes have been directly visualized by cryo-electron microscopy. The most relevant feature of the tetramer is the central channel surrounded by toxin monomers; at the lower end, this pear-shaped channel is 25 Å in diameter, and it then widens to 36 Å to be constricted to 10 Å at the top. Thus, the inserted α-latrotoxin tetramer would permeabilize the membrane to all substances that can pass through the central channel, such as small molecules and hydrated cations. Detailed structure–function analysis has been conducted using recombinant α-latrotoxins. It was found that the last two ankyrin-like repeats are not required for binding to neurexin 1α or CL1 but essential for pore formation. This mutant largely lost the ability to stimulate exocytosis from neuroendocrine cells, which supports the fact that pore formation plays a critical role in Ca^{2+}-dependent exocytosis from neuroendocrine cells.

Pore formation may not be the only mechanism underlying α-latrotoxin-stimulated exocytosis from neurons in the presence of extracellular Ca^{2+}. Since CL1 is a G-protein-coupled receptor, it has been hypothesized that the activation of G-protein pathway also contributes to neurotransmitter release. Biochemical analysis suggests that CL1 is linked to $G\alpha_{q/11}$, and the downstream effector of $G\alpha_{q/11}$, phospholipase C (PLC), stimulates extracellular Ca^{2+}-dependent hydrolysis of phosphoinositides. Activated PLC increases the cytosolic concentration of IP_3, which in turn induces release of Ca^{2+} from intracellular stores. This rise in cytosolic Ca^{2+} may increase the probability of neurotransmitter release and, consequently, the rate of spontaneous exocytosis and the amplitude of evoked release. Thus, activation of CL1 leads to extracellular Ca^{2+}-dependent mobilization of intracellular Ca^{2+} and mobilization of Ca^{2+} appears to play a potential role in neurotransmitter release stimulated by α-latrotoxin receptor signaling. The contribution of this receptor signaling pathway to the stimulation of neurotransmitter release is evident from the analysis of a unique α-latrotoxin mutant called $\alpha\text{-LTx}^{N4C}$. This mutant has four extra amino acids that were inserted between domain II and domain III, and it does not seem to form the Ca^{2+}-dependent ionophore, but it stimulates extracellular Ca^{2+}-dependent hydrolysis of phosphoinositides as well as extracellular Ca^{2+}-dependent neurotransmitter release. More important, this mutant can stimulate glutamate release from CA3 neurons in the presence of extracellular Ca^{2+}. Thus, activation of this receptor signaling pathway appears to contribute to the stimulation of neurotransmitter release from neurons in the presence of extracellular Ca^{2+}. In contrast, this mutant is ineffective in stimulating exocytosis from neuroendocrine cells, which supports the belief that the influx of extracellular Ca^{2+} is essential for α-latrotoxin-stimulated exocytosis from neuroendocrine cells. Therefore, this suggests that mechanisms underlying extracellular Ca^{2+}-dependent exocytosis may differ between neuroendocrine cells and neurons.

Role for CL1 in α-Latrotoxin-Stimulated Neurotransmitter Release in the Absence of Extracellular Ca^{2+}

Even in the absence of extracellular Ca^{2+}, α-latrotoxin can stimulate massive neurotransmitter release from neurons. Based on the analysis of CL1 knockout mice, CL1 is the major (~75%) surface receptor for α-latrotoxin, but it remains a mystery how α-latrotoxin

can stimulate Ca^{2+}-independent exocytosis after binding CL1. Unlike wild-type α-latrotoxin, removal of extracellular Ca^{2+} by EGTA abolishes the stimulatory action of α-LTx^{N4C}. Thus, activation of CL1 receptor signaling itself does not explain the stimulation of Ca^{2+}-independent neurotransmitter release by α-latrotoxin. Also, the previously mentioned CL1 signaling pathway ($G\alpha_{q/11}$, PLC) requires the presence of extracellular Ca^{2+} and will not be activated otherwise. It has been hypothesized that secretion triggered by α-latrotoxin in the absence of Ca^{2+} may be attributable to its membrane penetration and/or pore formation. This is further substantiated by the requirement of Mg^{2+} for α-latrotoxin-stimulated release in Ca^{2+}-free conditions, which corresponds to the fact that Mg^{2+} facilitates toxin tetramerization and pore formation. However, it is unclear whether secretion in the absence of extracellular Ca^{2+} can be induced by the pore-mediated influx/efflux of cations or membrane depolarization. The tetrameric wild-type toxin inserted into the neuronal plasma membrane could potentially interact directly with some components of the intracellular transmitter release apparatus, defining a novel hypothetical mechanism proposed recently.

CLs Are Conserved Proteins: Lessons from *Caenorhabditis elegans*

The presence of CL orthologs is reported in *C. elegans* (*lat-1* and *lat-2*) and *Drosophila* (RE25258p/CG8639-PA), which indicates the function of CLs to be conserved evolutionally. Interestingly, null mutation of *lat-1* or RNA interference-mediated knockdown of *lat-1* in *C. elegans* results in a prolonged pharyngeal pump duration and constipation phenotype. Extended pharyngeal pump duration and constipation phenotype are often observed in *C. elegans* mutants with defective neurotransmission, although the defect observed in *lat-1* mutations is not as marked. Furthermore, *lat-1* knockdown worm is more resistant than wild-type worms to lavamisole and imipramine, blockers of acetylcholine receptors, indicating a decrease in the release of acetylcholine. *Lat-1* appears to be expressed in subsets on neurons rather than muscles. These results suggest that *lat-1* may be involved in aspects of neurotransmitter release and that there may be endogenous ligands which exist to activate *lat-1* and/or *lat-2*.

Interestingly, black widow spider venom, from which α-latrotoxin is purified, is toxic to *C. elegans* but purified α-latrotoxin is not. This suggests that venom from the black widow spider contains toxins other than α-latrotoxin. In addition, knockdown of *lat-1*, but not *lat-2* or neurexin 1 homolog, confers resistance to the venom. These results suggest that *lat-1* functions as the receptor for latrotoxin-like proteins in the venom. Furthermore, it has been discovered that emodepside, which belongs to a new class of drugs, the cyclo-octadepsipeptides, binds to nematode *lat-1*. Emodepside causes paralysis in nematodes, including in the pharynx, and facilitates destaining of vesicles stained with FM4-64 in the pharynx, which suggests that emodepside facilitates neurotransmitter release. More important, the pharynxes of *lat-1* mutants are resistant to emodepside, indicating that *lat-1* functions as a receptor for emodepside.

How does emodepside trigger neurotransmitter release after binding to *lat-1*? Unlike α-latrotoxin, emodepside does not have pore-formation abilities. This suggests that it induces exocytosis through the activation of *lat-1*. Similar to CL1, *lat-1* is a G-protein-coupled receptor. In addition, reduction or loss-of-function mutations in Gq, PLC-*β*, and *unc13* in the worm confers resistance to emodepside, suggesting that these pathways are critical for the actions of emodepdide. As discussed previously, the Gq and PLC pathway is activated by mammalian CL1. Thus, *lat-1* and CL1 appear to activate a similar signaling pathway leading to neurotransmitter release. Unc13 is a key priming molecule essential for synaptic vesicle-mediated release of neurotransmitter. Thus, activation of *lat-1* may activate Unc13 and prime neurotransmitter release. Unc13 is also a well-conserved protein and three mammalian homologs (Munc13-1, -2, and -3) are essential for priming of neurotransmitter release in mammalian neurons. When both Munc13-1 and -2 are knocked out, the effects of α-latrotoxin are severely reduced. Thus, α-latrotoxin-dependent exocytosis may involve Munc13 activation via CL1 signaling pathways. However, Munc13s are required for neurotransmitter release stimulated not only by α-latrotoxin but also by other stimulation such as Ca^{2+} influx following depolarization. Thus, the results of Munc13 knockouts do not necessarily demonstrate that α-latrotoxin directly utilizes the Munc13 pathways to trigger neurotransmitter release.

Biological Functions of CLs

CL1 and CL3 are highly enriched in the mammalian brain, whereas CL2 is expressed ubiquitously. This suggests that although the primary function of CLs is in the brain, they function ubiquitously. The conservation of CL1 in lower organisms strongly suggests that CLs are involved in critical physiological functions. Based on the phenotype of *lat-1* in *C. elegans*,

CLs are likely involved in the stimulation of neurotransmitter release. CL1 knockout mice do not show phenotypes, but this lack of phenotype is probably due to the presence of closely related isoforms, CL2 and CL3.

Since CL1 is the major receptor for α-latrotoxin and α-latrotoxin induces neurotransmitter release after binding to presynaptic receptors, CL1 is believed to be localized at the presynapse. To identify the physiological function of CL1, efforts have been made to identify the CL1-interacting proteins by yeast two-hybrid screens using a large intracellular sequence of CL1 as bait. Two groups identified the Shank family of proteins as CL1-interacting proteins, and this interaction was further confirmed by coimmunoprecipitation experiments. This is surprising because the Shank family of proteins is enriched at the postsynapse. The interactions with the Shank family of proteins suggest that CL1 is also expressed at the postsynaptic side, although the role for CL1 at the postsynapse is enigmatic.

The structure of CLs suggests that they are involved in cell-to-cell interaction processes coupled with signal transduction. At least 33 proteins, including CL1, -2, and -3, belong to the LNB 7TM family of proteins. Surprisingly, binding proteins or ligands have been identified for very few of the LNB 7TM proteins. CD97, a member of the LNB 7TM proteins and a homolog to CLs in the GPS and TMR domains, is one of these few examples that has been found to bind CD55. This interaction is mediated by the N-terminal epidermal growth factor-like domain of CD97, which does not exist in CLs. The study to elucidate the physiological significance of this interaction is under way. The extracellular domains of CL1 contain lectin-like and olfatomedin/myocillin-like domains that are considered to be cell adhesion motifs. Identification of the physiological ligand for CL1 will facilitate better understanding of the function of CL1, but no ligand has been identified so far. Further work including identification of ligands is clearly required to elucidate the physiological functions of CLs.

See also: Cell Adhesion Molecules at Synapses; Invertebrate Neurohormone GPCRs; Latrotoxin; Munc13 and Associated Molecules; Neurexins; Neurosecretion (Regulated Exocytosis in Neuroendocrine Cells); Neurotoxins; Neurotoxins and their Neurotoxicology.

Further Reading

Capogna M, Volynski KE, Emptage NJ, and Ushkaryov YA (2003) The alpha-latrotoxin mutant LTX^{N4C} enhances spontaneous and evoked transmitter release in CA3 pyramidal neurons. *Journal of Neuroscience* 23: 4044–4053.

Davletov BA, Shamotienko OG, Lelianova VG, Grishin EV, and Ushkaryov YA (1996) Isolation and biochemical characterization of a Ca^{2+}-independent α-latrotoxin-binding protein. *Journal of Biological Chemistry* 271: 23239–23245.

Foord SM, Jupe S, and Holbrook J (2002) Bioinformatics and type II G-protein-coupled receptors. *Biochemical Society Transactions* 30: 473–479.

Ichtchenko K, Khvotchev M, Kiyatkin N, et al. (1998) α-Latrotoxin action probed with recombinant toxins: Receptors recruit α-latrotoxin but do not transduce an exocytotic signal. *EMBO Journal* 17: 6188–6199.

Khvotchev M and Südhof TC (2000) α-Latrotoxin triggers transmitter release via direct insertion into the presynaptic plasma membrane. *EMBO Journal* 19: 3250–3262.

Krasnoperov V, Bittner MA, Holz RW, Chepurny O, and Petrenko AG (1999) Structural requirements for α-latrotoxin binding and α-latrotoxin-stimulated secretion. A study with calcium-independent receptor of alpha-latrotoxin (CIRL) deletion mutants. *Journal of Biological Chemistry* 274: 3590–3596.

Krasnoperov VG, Bittner MA, Beavis R, et al. (1997) α-Latrotoxin stimulates exocytosis by the interaction with a neuronal G-protein-coupled receptor. *Neuron* 18: 925–937.

Li G, Lee D, Wang L, et al. (2005) N-terminal insertion and C-terminal ankyrin-like repeats of α-latrotoxin are critical for Ca^{2+}-dependent exocytosis. *Journal of Neuroscience* 25: 10188–10197.

Orlova EV, Rahman MA, Gowen B, et al. (2000) Structure of α-latrotoxin oligomers reveals that divalent cation-dependent tetramers form membrane pores. *Nature Structural Biology* 7: 48–53.

Südhof TC (2001) α-Latrotoxin and its receptors: Neurexins and CIRL/latrophilins. *Annual Review of Neuroscience* 24: 933–962.

Sugita S, Ichtchenko K, Khvotchev M, and Südhof TC (1998) α-Latrotoxin receptor CIRL/latrophilin 1 (CL1) defines an unusual family of ubiquitous G-protein-linked receptors: G-protein coupling not required for triggering exocytosis. *Journal of Biological Chemistry* 273: 32715–32724.

Tobaben S, Südhof TC, and Stahl B (2002) Genetic analysis of α-latrotoxin receptors reveals functional interdependence of CIRL/latrophilin 1 and neurexin 1α. *Journal of Biological Chemistry* 277: 6359–6365.

Ushkaryov YA, Volynski KE, and Ashton AC (2000) The multiple actions of black widow spider toxins and their selective use in neurosecretion studies. *Toxicon* 43: 527–542.

Volynski KE, Meunier FA, Lelianova VG, et al. (2000) Latrophilin, neurexin, and their signaling-deficient mutants facilitate α-latrotoxin insertion into membranes but are not involved in pore formation. *Journal of Biological Chemistry* 275: 41175–41183.

Willson J, Amliwala K, Davis A, et al. (2004) Latrotoxin receptor signaling engages the UNC-13-dependent vesicle-priming pathway in *C. elegans*. *Current Biology* 14: 1374–1379.

Clathrin and Clathrin-Adaptors

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

Cells are able to internalize molecules that are outside the plasma membrane by engulfing them in vesicles and drawing those vesicles into the cell interior, a process known as endocytosis. There are several different endocytic pathways, but clathrin-mediated endocytosis is the best understood at the molecular level. This pathway, which is often used for the internalization and recycling of membrane-spanning receptors, involves the formation of special patches of membrane called pits, which are defined by the presence of the cytosolic protein clathrin. Because clathrin is used to internalize many different receptor molecules but does not have affinity for all of them, the formation of a clathrin-coated pit requires the sequential action of numerous factors, beginning with clathrin adaptors whose function is to associate clathrin with particular cargo molecules and recruit accessory factors that help the clathrin to form a polymeric matrix on the inner membrane surface. The polymerization of clathrin is responsible for membrane buckling which allows the vesicle to form. After vesicle scission, the clathrin and adaptor molecules dissociate, leaving them free to contribute to further coated pits and vesicles.

Clathrin

Clathrin is the structural protein which accumulates at the sites that form clathrin-coated pits. The native clathrin molecule is a hexamer comprising three identical copies of a 180 kDa heavy chain and three copies of a 30–40 kDa light chain to form a three-legged structure known as a triskelion. The heavy chain is invariant and contains recognition motifs for clathrin adaptors. In contrast, there are several light-chain isoforms in mammals, resulting from the presence of two genes, the primary transcripts of which can be spliced in different ways. The light chain regulates clathrin self-assembly into polyhedral lattices and is capable of doing so both *in vivo* to form pits and vesicles and *in vitro* to form empty cages that are structurally very similar to clathrin-coated pits. Clathrin has no affinity for membrane lipids or for any known class of transmembrane receptor. Therefore, clathrin assembly at the site of receptor molecules targeted for endocytosis is completely dependent on adaptors and accessory proteins.

Clathrin Adaptors

Clathrin adaptors are required to recruit clathrin to the membrane and assemble clathrin-coated pits at the sites of appropriate cargo molecules. They achieve this through their ability to recognize and bind multiple components of the endocytic pathway simultaneously, including the plasma membrane, clathrin, the cargo targeted for internalization, and various accessory proteins. Two major classes of adaptor are recognized – classic adaptors, which are multimeric, and alternative adaptors known as CLASPs (clathrin-associated sorting proteins), which are either single polypeptides or, occasionally, homodimers.

Classic adaptors are tetrameric with a molecular mass of approximately 300 kDa. The general architecture is $X\beta\mu\sigma$, where X is a variable subunit responsible for membrane binding and the recruitment of accessory proteins, and the $\beta\mu\sigma$ subunits are more structurally conserved. In the well-characterized activator protein (AP)2 adaptor, for example, X is the α subunit and the other three subunits are designated $\beta2$, $\mu2$, and $\sigma2$ (Figure 1). The larger subunits (α and $\beta2$) each comprise an N-terminal trunk domain and a C-terminal appendage joined to the trunk by a flexible hinge region. The α subunit binds to membrane lipids (specifically PtdIns$_{(4,5)}$P$_2$) and to accessory proteins such as epsin-1. The $\beta2$ subunit binds to cargo dileucine motifs, recognizes clathrin via its hinge domain, and recognizes additional accessory proteins via its appendage domain. The $\mu2$ subunit is cargo specific and is absolutely required for cargo recognition, but it also interacts with membrane lipids and with accessory proteins such as stonin-1. The $\sigma2$ subunit is thought to have a purely structural role, facilitating interactions between the other subunits.

A number of different CLASPs have been identified, most of which function as single polypeptides. CLASPs have multiple domains that recognize lipids, the peptide signals in the cytoplasmic tails of their cargo, clathrin, accessory proteins, and in some cases the classic adaptors such as AP2. This modular design allows the adaptors to interact simultaneously with multiple components of the endocytic pathway in a highly synchronized manner. Well-characterized CLASPs include Golgi-associated γ-ear-containing ADP-ribosylation-factor-binding (GGA) proteins, which function in concert with the AP1 classic adaptor

epsin, AP180/clathrin assembly lymphoid myeloid leukemia protein (CALM), and β-arrestin.

Initiation of Endocytic Complex Assembly by Clathrin Adaptors

It is thought that lipid–adaptor interactions initiate the process of clathrin-coated pit formation, followed by interactions with cargo, accessory proteins, and finally clathrin itself. Lipid binding by the α and $\mu 2$ subunits of AP2 (Figure 2) and by epsin is particularly well characterized, and specific domains responsible for interaction with PtdIns$_{(4,5)}$P$_2$ have been identified on many of the alternative adaptor proteins.

In AP2, the N-terminal core domain of the α subunit contains a series of basic side chains, including two clusters of lysine residues which cooperate with lysine residues found at the C-terminus of the $\mu 2$ subunit to bind the phosphate groups of PtdIns$_{(4,5)}$P$_2$. Several different lipid-binding domains have been identified in the CLASP-adaptor proteins, including the epsin NH2-terminal homology (ENTH) domain in epsin, the AP180 N-terminal homology (ANTH) domain in AP180/CALM and the phosphotyrosine-binding (PTB) domain in autosomal-recessive hypercholesterolemia (ARH) and Numb (Figure 3). Unlike the strategy used by AP2, the ENTH domain of epsin binds to both the phosphate head group and the glycerol backbone of PtdIns$_{(4,5)}$P$_2$, which is achieved through a conformational change during binding. Four α-helices in the ENTH domain form a basic pocket to secure the head group, but binding results in a conformation change that forms an additional, amphipathic helix. The hydrophobic face of this helix aligns with the hydrophobic portion of the lipid, whereas the hydrophilic face cooperates to hold the head group in the basic pocket, thus anchoring epsin into the membrane. The ANTH domain of AP180/CALM is also understood in detail. This domain comprises nine α-helices, two of which contact the head group of PtdIns$_{(4,5)}$P$_2$ via three lysine residues and one histidine.

Cargo binding is achieved through the recognition of various peptide motifs and/or specific forms of posttranslational modification on the cytosolic portion of the cargo molecule. It is unclear whether lipid binding or cargo binding occurs first, but it is known that the stability of cargo binding is enhanced by simultaneous interaction with the membrane. For example, the recruitment of AP2 to the plasma membrane may occur through sequential interactions, first between the α subunit and PtdIns$_{(4,5)}$P$_2$, then between the $\mu 2$ subunit and both PtdIns$_{(4,5)}$P$_2$ and cargo carrying the peptide motif Y-X-X-Hy (where Y is tyrosine, X is any amino acid, and Hy is any hydrophobic amino acid), and then between the $\beta 2$ subunit and dileucine motifs on the cargo ([DE]xxxL[LI]). Similarly, studies in yeast suggest that epsin binds to PtdIns$_{(4,5)}$P$_2$ via its ENTH domain and simultaneously to ubiquitylated proteins via two ubiquitin-interacting motifs (UIMs). The removal of either UIM reduces, but does not abolish, the ability of epsin to interact with the membrane, indicating some sort of cooperative-binding behavior.

Once the adaptor has been recruited to the membrane, the scene is set for the recruitment of clathrin

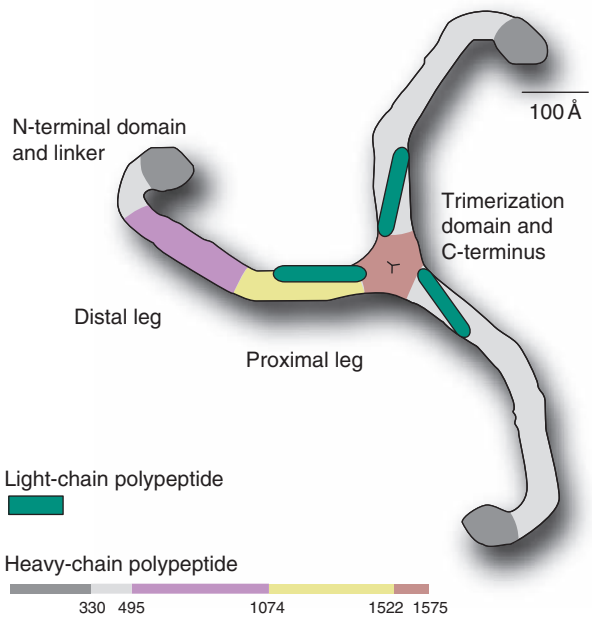

Figure 1 The structure of clathrin.

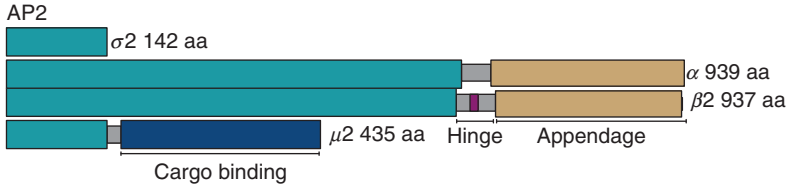

Figure 2 The classical clathrin adaptor activator protein (AP)2. Reproduced from Maldonado-Baez L and Wendland B (2006) Endocytic adaptors: Recruiters, coordinators and regulators. *Trends in Cell Biology* 16: 505–513.

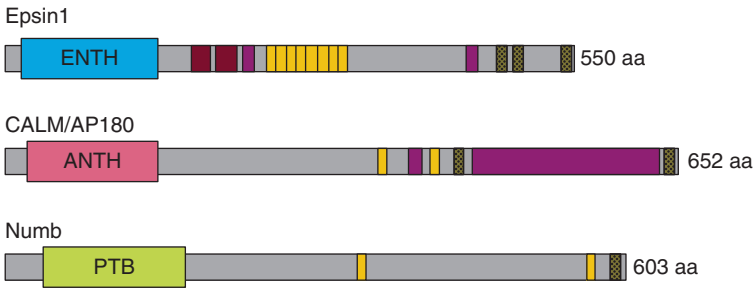

Figure 3 The structures of three alternative adaptor proteins (CLASPs) showing the lipid-binding and cargo-binding domains. AP, activator protein; CALM, clathrin assembly lymphoid myeloid leukemia protein; CLASPs, clathrin-associated sorting proteins. Reproduced from Maldonado-Baez L and Wendland B (2006) Endocytic adaptors: Recruiters, coordinators and regulators. *Trends in Cell Biology* 16: 505–513.

itself. Many studies have confirmed the necessity of adaptor proteins for clathrin localization, and indeed colocalization experiments have shown unambiguously that yeast cells lacking clathrin adaptors fail to generate localized clathrin pits. Clathrin recruitment requires adaptors already bound to the plasma membrane. The presence of lipid- and clathrin-binding motifs in these adaptors allows the adaptors to interact simultaneously with the plasma membrane and clathrin. In the case of AP2, recruitment is achieved by way of the clathrin box – a short motif in the hinge region of the β2 subunit with the sequence LLNLD, which interacts with the β-propeller formed by WD40 repeats at the N-terminal domain of the clathrin heavy chain. There is also evidence for a second clathrin-binding motif in β2 which acts cooperatively with the LLNLD motif. Two distinct classes of clathrin-binding domain have been identified in CLASPs: a type I box with the sequence L-Hy-X-Hy-[DE] (where Hy is any bulky hydrophobic amino acid residue, X is any amino acid, and the negatively charged residue (D or E) can also be represented by the C-terminal carboxyl group) and a type II box with the sequence L-Hy-D-L-L (a shortened version comprising the last three residues is present in AP180). These also interact with the N-terminal domain of the clathrin heavy chain and promote clathrin polymerization with the same efficiency.

The process of cargo recognition, membrane association, and clathrin recruitment would not proceed (or would not proceed with such efficiency) but for the presence of numerous accessory proteins that provide a structural architecture that facilitates the interactions between adaptors and other endocytic components. Many of the clathrin adaptors interact with the same repertoire of accessory proteins, one example being epidermal growth factor receptor pathway substrate (Eps)15, which has been shown to bind AP2, epsin, AP180, Numb, Stonin-2, and membrane cargo proteins. Experiments have shown that Eps15 and other accessory proteins can accelerate the formation of clathrin cages *in vitro*, indicating a similar stimulatory role *in vivo*. Interestingly, Eps15 also interacts with downstream components of the endocytic pathway that are not discussed in this article, such as dynamin. This suggests that Eps15 needs to adopt a position within the forming endocytic structure that will allow it to interact transiently with adaptor proteins that are part of the clathrin-coated bud as well as with endocytic proteins that act later, which are found only at the narrowing bud neck.

Roles of Clathrin and Clathrin Adaptors during Vesicle Formation

During the formation of clathrin-coated vesicles, the plasma membrane encounters increasing degrees of deformation. The regulated assembly of clathrin coats and adaptor complexes at the plasma membrane initiates the membrane's curving to form a vesicle. The driving force seems to come from the combined action of clathrin coats and adaptors recruited to the plasma membrane, providing the mechanical support and energy necessary to deform the membrane.

Biophysical and structural studies demonstrate that clathrin coats assemble with the same degree of curvature as coated vesicles. It has been proposed that localized clathrin-coat assembly can promote membrane deformation because of the intrinsic curvature of the clathrin coat. The constraint of the curved coat on the lipid bilayer might provide some of the force necessary to destabilize and pull the membrane inward. However, a recent thermodynamic model proposes that, instead of driving the curvature of the membrane, clathrin provides not the force but the mechanical support required to maintain an already curved membrane in that state. This thermodynamic model includes parameters that account for clathrin triskelion rigidity, intrinsic triskelion curvature, and free-energy changes resulting from interactions between triskelions. Integrating all these parameters predicts the size distribution of clathrin baskets or coats *in vitro*

and reveals that polymerized clathrin triskelion rigidity is not sufficient to surpass the bending resistance of the lipid bilayer. Thus, to bend the membrane to form a vesicle, other mechanical forces must cooperate with clathrin-coat formation to overcome membrane-bending resistance. In agreement with this idea, the same study demonstrates that the binding of adaptor proteins to the coat stabilizes the interaction between the triskelion subunits and increases their intrinsic curvature and rigidity.

Two models have been proposed to explain how adaptors promote membrane deformation. The first model proposes that endocytic adaptors with amphipathic helices promote the deformation of the plasma membrane by the physical penetration of the amphipathic helix into one leaflet of the membrane, whereas the second one suggests that adaptor proteins also facilitate membrane deformation by clustering at specific areas through interactions with PtdIns$_{(4,5)}$P$_2$ head groups, reaching concentrations high enough to provoke local destabilization and prompting membrane deformation.

Epsin seems to use the mechanism proposed by the amphipathic helix model. The simultaneous recruitment of epsin and polymerization of clathrin coats helps stabilize the deformed membrane. By contrast, the role of the AP180 ANTH domain in membrane deformation is less clear because the domain lacks an amphipathic helix to insert into the membrane. However, although no deformation of the membrane is observed on ANTH-domain binding, AP180 promotes the formation of uniformly sized clathrin-coated vesicles, resulting in a more homogeneous population than vesicles assembled by epsin alone, which might be essential for the regulation of the size of synaptic vesicles. It has been proposed that AP180 might control the size of the clathrin coat, which might be significant for processes dependent on vesicle size, such as synaptic vesicle formation. Thus, adaptors such as AP180 that do not intrinsically deform the membrane might act as proposed in the second model, in which protein clustering leads to membrane deformation. Perhaps both models are correct, and adaptors such as AP180 cooperate with membrane-deforming proteins such as epsin to stabilize the curved state of the membrane. The coordination of these two types of adaptors might guarantee the successful internalization of an endocytic vesicle with a minimal energy cost.

The membrane curvature and membrane bending that are needed at late stages of endocytosis might be mediated by yet another class of endocytic proteins that contain bin-amphiphysin-rvs (BAR) or F-BAR domains. These lipid-binding domains can either bind membranes with a preexisting curvature or assist in membrane bending through the insertion of an amphipathic helix into the hydrophobic layer.

Regulation of Clathrin Adaptors: Interactions and Functions

It is clear that clathrin adaptors play the most significant role in the formation of clathrin-coated pits and vesicles, which suggests they are the prime target for upstream regulation. There is evidence that posttranslational modification is used to regulate the activity and potential interactions of these proteins, allowing them to be controlled by upstream signaling pathways.

One of the most abundant forms of regulatory modification in eukaryotic cells is the phosphorylation of tyrosine, serine, and threonine residues. A relevant example is the phosphorylation of AP2 subunit $\mu 2$ by protein kinase AP2-associated kinase (AAK1). The phosphorylation of a single threonine residue within the lipid-binding domain of subunit $\mu 2$ has a profound effect on its ability to interact with PtdIns$_{(4,5)}$P$_2$ head groups, increasing its affinity over 100-fold. The additional negative charge introduced by phosphorylation is thought to cause a conformation change that alters the relative spacing of the cargo-binding and lipid-binding domains, allowing both sites to bind their respective ligands simultaneously, and leading to a stable complex at the membrane. Other adaptor proteins are negatively regulated by phosphorylation. For example, epsin phosphorylation prevents endocytosis, presumably because the additional negatively charged residue interferes with crucial interactions with the endocytic machinery.

As already stated, the ubiquitylation of cargo molecules can be important for the initiation of vesicle formation by epsin, but an additional complication is that such adaptors can also undergo ubiquitylation at a site distinct from their ubiquitin-binding domain (UBD). *In vitro*, it has been shown that the presence of a single ubiquitin moiety on epsin inhibits its ability to bind to accessory proteins and ubiquitinylated cargo, therefore reducing the efficiency of vesicle formation. It is possible that the ubiquitin moiety interacts with the UBD on the same protein and that this intramolecular binding outcompetes such interactions with other molecules, whereas the UBD on the nonubiquitinylated form of the protein is more likely to partake in intermolecular interactions.

Summary

Clathrin and clathrin adaptors are key molecules in the process of clathrin-mediated endocytosis, which is used for the specific uptake of large extracellular molecules

such as proteins, membrane-localized receptors, and ion channels. Clathrin has no affinity for the plasma membrane or any receptor molecules located therein, so the onus for initiating the formation of clathrin-coated pits lies firmly with the clathrin adaptors, multidomain proteins which simultaneously bind to membrane lipids, specific cargo proteins, clathrin itself, and other endocytosis accessory proteins to recruit all the endocytic machinery into a complex and initiate the process of clathrin polymerization which underlies vesicle formation. Many of the specific domains involved in the assembly of such complexes have been identified and functionally characterized. It is clear that the regulation of endocytosis is mediated in part by the posttranslational modification of clathrin adaptors, a process which directly influences their ability to interact with other components of the endocytic machinery.

See also: Endocytosis and Presynaptic Scaffolds; Synaptic Vesicles; Ubiquitin–Proteasome System and Plasticity.

Further Reading

Conner SD and Schmid SL (2003) Regulated portals of entry into the cell. *Nature* 422: 37–44.

Edeling MA, Smith C, and Owen D (2006) Life of a clathrin coat: Insights from clathrin and AP structures. *Nature Reviews Molecular and Cell Biology* 7: 32–44.

Evans PR and Owen DJ (2002) Endocytosis and vesicle trafficking. *Current Opinion in Structural Biology* 12: 814–821.

Maldonado-Báez L and Wendland B (2006) Endocytic adaptors: Recruiters, coordinators and regulators. *Trends in Cell Biology* 16: 505–513.

Mousavi SA, Malerød L, Berg T, and Kjeken R (2004) Clathrin-dependent endocytosis. *Biochemical Journal* 377: 1–16.

Owen DJ (2004) Linking endocytic cargo to clathrin: Structural and functional insights into coated vesicle formation. *Biochemical Society Transactions* 32: 1–14.

Owen DJ, Collins BM, and Evans PR (2004) Adaptors for clathrin coats: Structure and function. *Annual Review of Cell and Developmental Biology* 20: 153–191.

Slepnev VI and DeCamilli P (2000) Accessory factors in clathrin-dependent synaptic vesicle endocytosis. *Nature Reviews Neuroscience* 1: 161–172.

Sorkin A (2004) Cargo recognition during clathrin-mediated endocytosis: A team effort. *Current Opinion in Cell Biology* 16: 392–399.

Sorkin A and von Zastrow M (2002) Signal transduction and endocytosis: Close encounters of many kinds. *Nature Reviews Molecular and Cell Biology* 3: 600–614.

Szymkiewicz I, Shupliakov O, and Dikic I (2004) Cargo- and compartment-selective endocytic scaffold proteins. *Biochemical Journal* 383: 1–11.

Traub LM (2003) Sorting it out: AP-2 and alternate clathrin adaptors in endocytic cargo selection. *Journal of Cell Biology* 163: 203–208.

Wendland B (2002) Epsins: Adaptors in endocytosis? *Nature Reviews Molecular and Cell Biology* 3: 971–977.

Clock Gene Regulation of Endocrine Function

F W Turek, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A major characteristic of the endocrine system is its high degree of temporal organization. Two of the most prominent temporal variations are those in the circadian range (i.e., about a day, 24 h) and in the ultradian range (e.g., 1–2 h). In addition, some secretory oscillations are also in the 5–20 min range, while others can occur over many days (e.g., estrous and menstrual cycling hormones) or over many months (e.g., seasonal rhythms in reproduction and metabolism). This article focuses on the circadian regulation of the endocrine system with an emphasis on the role of circadian clock genes in this regulation.

Circadian Control of the Endocrine System

It is now recognized that hormones in all the endocrine systems show 24 h fluctuations. For example, the hormones produced by the pituitary gland (e.g., adrenocorticotropic hormone or growth hormone) are synthesized and released to greater or lesser extents on a regular basis throughout the 24 h day. Similarly, hormones involved in energy regulation, such as leptin, insulin, and ghrelin, also show 24 h fluctuations in the bloodstream. These diurnal rhythms are regulated by an internal circadian clock that is entrained by daily environmental cues in the physical environment, primarily the 24 h light–dark cycle imposed on the earth due to the rotation on its axis.

The circadian control of endocrine fluctuations can be direct or indirect. That is, the master circadian clock located in the suprachiasmatic nucleus (SCN) of the hypothalamus in mammals can directly regulate the 24 h pattern of synthesis and release of a particular hormone. This is the case, for example, for the pineal hormone, melatonin, whose release into the bloodstream occurs during the nighttime in mammals and is under direct neural control via a multisynaptic input from the SCN to the pineal gland. Alternatively, the SCN may indirectly control components of, or even the entire, rhythmic 24 h pattern of the circulating levels of a given hormone by regulating a 24 h rhythm in behavior, which in turn may play a major role in the timing of a particular hormonal rhythm. For example, some hormones may be influenced in part by the sleep–wake state of the animal (e.g., growth hormone), or the 24 h rhythm in food intake (e.g., insulin), behavioral rhythms that for most animals are under the tight control of the circadian clock in the SCN. Whether direct or indirect, the SCN control of the 24 h rhythms in endocrine function ensures that the internal endocrine milieu maintains synchrony both within the organism and between the organism and the daily changes in the external physical environment.

Circadian Clock Genes and the Endocrine System

Remarkable progress was made in the last few years of the twentieth century in unraveling the genetic basis of the core molecular transcriptional/translational feedback loop that underlies the generation of circadian rhythms in mammals. While the discovery of the circadian clock genes, and the proteins that are encoded by these genes, has provided great insight into how all living organisms can actually keep track of time on a 24 h basis, their discovery had an interesting side effect. Once the clock genes were discovered and their expression patterns could be followed, a surprising finding was made: these genes were expressed not only in the central master circadian clock, the SCN, but also in most, if not all, the tissues of the body, both in the central nervous system (CNS) and in the periphery. Furthermore, these circadian clock genes were not just expressed in these non-SCN cells, but were also expressed on a circadian basis. Most surprisingly, when tissues were cultured *in vitro*, many (e.g., lung, muscle, liver, pancreas, and ovary) were found to be capable of expressing 24 h rhythms in a dish. In other words, circadian clocks were everywhere, and they were not just keeping time everywhere for themselves, but were actually regulating the timing of the expression of hundreds, if not thousands, of so-called clock-controlled genes (CCGs).

The finding of circadian clock genes in many endocrine tissues, including the pancreas, liver, fat, pituitary, ovary, testis, and adrenal gland, and their temporal regulation of the timing of hundreds of downstream CCGs, has profound implications for endocrine physiology and pathophysiology. The ubiquitous nature of the circadian molecular machinery in endocrine tissues raises the possibility that mistiming at the tissue or organ level could lead to dysfunction of that endocrine tissue/organ at a local level. Just as the molecular circadian clock in the SCN regulates all the endocrine, metabolic, and behavioral circadian rhythms in the whole organism, that same molecular circadian clock machinery is in turn regulating the internal timing within specific cells, tissues,

and organs at a local level. Any disruption of this circadian timing, at the organismal or local level, could in turn lead to dysfunction globally or within a particular endocrine system. Indeed, recent data from mice carrying a mutation of one of the canonical circadian clock genes, named *Clock*, demonstrate how a disruption of circadian clock gene function can have far-ranging effects on diverse endocrine tissues. The *Clock* gene was the first mammalian circadian gene to be identified, and its protein product is known to dimerize with another circadian gene, *Bmal* (also known as MOP3). Together these proteins act as a transcription factor to stimulate the production of two other sets of circadian genes, the period (*Per*) and chrytochrome (*Cry*) genes. The PER and CRY protein products are on the negative feedback loop of the molecular clock as they dimerize to inhibit the transcription of the *Clock* and *Bmal* genes. A mutation of the *Clock* gene leads to an extremely long circadian period (27–28 h) in homozygous *Clock* mutant animals maintained in constant darkness (DD), before leading to a state of arrhythmicity in many animals, that is, no expression of a circadian rhythm in DD. The *Clock* mutation also alters the expression of other circadian genes, particular the *Per* genes, as well as CCGs.

Given the important role played by the circadian clock in regulating rhythmicity in the endocrine system, one might expect that a mutation leading to an alteration in circadian rhythmicity at the molecular level would have important consequences for endocrine function. Indeed, the *Clock* mutant animals are obese, particularly on a high-fat diet, relative to littermate controls, and they show many symptoms in common with the metabolic syndrome in humans. In addition, leptin levels are elevated and hypothalamic peptides involved in energy regulation (e.g., orexin and cocaine-amphetamine-regulated transcript (CART)) show reduced levels of expression. Female *Clock* mutant mice show alterations in the timing of the luteinizing hormone (LH) surge and have irregularities in pregnancy and in the timing of birth, suggesting that the *Clock* mutation causes changes in function across a broad range of endocrine systems.

Circadian Clock Genes and Endocrine Pathophysiology

As noted previously, diurnal rhythmicity is observed in essentially all endocrine systems. For many of these systems, alterations in timing, changes in amplitude, and even the total absence of rhythmicity have been associated with a variety of endocrine, metabolic, and psychiatric disorders. For example, hypercortisolism and alterations in the low point in the cortisol rhythm have been associated with severe depression, while growth hormone levels are markedly suppressed in obese subjects. Aging itself is known to decrease the amplitude of many hormonal rhythms. However, the role that differences or changes in circadian clock gene function *per se* may play in changes in endocrine physiology remains unknown.

Soon after circadian clock genes were discovered in mice, the successful search for human orthologs of genes such as *Clock* and *Per* occurred. Much like the drunk man who looks for his lost keys under the lamppost at night because that is where there is light, initial inquiries as to whether different clock gene alleles were linked to specific circadian phenotypes have focused on looking for genetic linkage of clock gene alleles with two relatively well-characterized unusual circadian phenotypes, delayed sleep phase syndrome (DSPS) and advanced sleep phase syndrome (ASPS). In both DSPS and ASPS, the desired onset of sleep occurs unusually late or early, respectively, independent of environmental factors. Due to the difficulty in being able to accurately monitor endocrine rhythms, as opposed to an easily observable behavioral rhythm such as the sleep–wake cycle, we have much less information about 'circadian disorders' that may exist in endocrine systems that could be due to inherent differences in circadian clock genes. However, the recent demonstration that night eating syndrome (NES), in which a large percentage of daily caloric intake occurs in the evening and during the night, is most likely not a sleep disorder but instead is a disorder of time keeping, has led to the suggestion that NES may just be the tip of the iceberg: a temporal disorder we can easily see or observe that is internal in nature. In contrast, many other metabolic and endocrine temporal disorders that are due to altered circadian clock gene function may go unnoticed, since they are not readily observable and no one is looking for them.

There are two common situations in which humans 'voluntarily' desynchronize their circadian clock from the local light–dark cycle as well as other day–night environmental cues. One condition, commonly referred to as jet lag, occurs when humans move rapidly across time zones. The other occurs when humans work (or play) at a time they normally would be asleep and attempt to sleep when they normally would be awake: the phenomenon of shift work. Both conditions are associated with a variety of physical and performance deficits, as well as changes in metabolic and endocrine function. Intriguingly, some individuals are better able to cope with jet lag and shift work than others. Indeed, the *Diagnostic and Statistical Manual for Mental Disorders* (DSM-IV)

recently classified those individuals who are severely affected by shift work as having shift-work sleep disorder, to distinguish them from the majority of shift workers who, while being affected by circadian disruption and lack of sufficient sleep, are nevertheless able to cope with the shift-work schedule. Such differences between individuals may depend upon differences in circadian clock genes, clearly an area for future research.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Circadian Function and Therapeutic Potential of Melatonin in Humans; Clock Genes and Metabolic Regulation; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Photoperiodic Regulation of Reproductive Cycles; Seasonal Changes in Night-Length and Impact on Human Sleep.

Further Reading

Balsalobre A, Brown SA, Marcacci L, et al. (2000) Resetting of circadian time in peripheral tissues by glucocorticoid signaling. *Science* 289: 2344–2347.

Fukuhara C and Tosini G (2003) Role of peripheral oscillators in mammals. *Frontier in Bioscience* 8: 642–651.

Miller BH, Losee Olson S, Levine JE, Turek FW, Horton TH, and Takahashi JS (2006) Vasopressin regulation of the proestrous luteinizing hormone surge in wild-type and *Clock* mutant mice. *Biology of Reproduction* 75: 778–784.

Rogers NL, Dinges D, Allison KC, et al. (2006) Assessment of sleep in women with night eating syndrome. *Sleep* 29(6): 814–819.

Rosenwasser AM and Turek FW (2005) Physiology of the mammalian circadian system. Section 4: Chronobiology. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practices of Sleep Medicine,* 4th edn., pp. 351–362. New York: WB Saunders.

Takahashi JS, Turek FW, and Moore RY (eds.) (2001) *Handbook of Behavioral Neurobiology, Vol. 12: Circadian Clocks*. New York: Kluwer Academic/Plenum Publishers.

Turek FW, Joshu C, Kohsaka A, et al. (2005) Obesity and metabolic syndrome in circadian *Clock* mutant mice. *Science* 308: 1043–1045.

Van Cauter E and Turek FW (2000) Roles of sleep–wake and dark–light cycles in the control of endocrine, metabolic, cardiovascular and cognitive function. In: McEwen BS (ed.) *Handbook of Physiology, Section 7: The Endocrine System, Vol. 4: Coping with the Environment: Neural and Endocrine Mechanisms*, pp. 313–330. New York: Oxford University Press.

Vitaterna MH, Pinto LH, and Turek FW (2005) Molecular genetic basis for mammalian circadian rhythms. Section 4: Chronobiology. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practices of Sleep Medicine,* 4th edn., pp. 363–374. New York: WB Saunders.

Yoo SH, Yamazaki S, Lowrey PL, et al. (2004) PERIOD2::LUCIFERASE real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues. *Proceedings of the National Academy of Sciences of the United States of America* 101(15): 5339–5346.

Clock Genes and Metabolic Regulation

A Kohsaka and J Bass, Northwestern University and Evanston Northwestern Healthcare Research Institute, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Central Nervous System Organization of Mammalian Circadian Rhythms

The Clock Is Universal

The molecular architecture of the biological clock is universal in diverse organisms, including cyanobacteria, plants, flies, birds, and mammals. Clock genes are expressed in nearly all cells and give rise to cellular circadian rhythms. In mammals, pacemaker neurons in the suprachiasmatic nuclei (SCN) preside over peripheral tissue clocks and maintain temporal synchrony of internal systems to match the external 24 h light–dark (LD) cycle caused by the daily rotation of Earth. The length of the circadian period is a critical, genetically programmed metric of the endogenous clock (circadian comes from *circa diem*, about a day). Period length is also species-dependent; for instance, mice show a shorter and humans exhibit a longer internal period length than 24 h in constant darkness. However, all terrestrial species are entrained to the same external period when exposed to an LD cycle of 24 h.

Light: Zeitgeber and Photoreception

Light is the predominant Zeitgeber signal (time giver) that entrains the biological clock of the SCN. The SCN, in turn, maintains temporal synchrony between internal periodic cycles and the external environment that is thought to enhance overall organismal function and survival. For example, the daily peak in the secretion of glucocorticoid promotes cardiovascular tone during the transition from sleep to wakefulness, and a rise in growth hormone during the rest phase may contribute to metabolic homeostasis, for instance, by promoting hepatic gluconeogenesis.

A major advance in understanding how the SCN receives photic input came from the discovery of the melanopsin photoreceptor pathway by Provencio and colleagues in 2000. In mammals, retinal photoreceptors, which comprise rods, cones, and ganglion cells, receive both visual (sight) and nonvisual (light) information. The nonvisual response, called circadian photoreception, plays an important role in the light entrainment of the circadian system. The discovery of melanopsin, a novel opsin, opened a new window on the detailed mechanism of circadian photoreception. A subset of retinal ganglion cells was subsequently shown to express melanopsin, and genetic ablation of melanopsin reduced the locomotor activity response to light. Melanopsin-containing ganglion cells receive input from the rod–cone system and send neural projections to multiple central nervous system (CNS) regions, including the hypothalamus and thalamus. Retinal photoreceptors project to the SCN via the retinohypothalamic tract (RHT). Recent studies also suggest that melanopsin-containing neurons may also project to broad regions of neocortex and thereby provide direct light input into regions involved in alertness, attention, and arousal independently of the RHT input into the SCN.

Suprachiasmatic Nucleus: Integration of Metabolic and Circadian Signals

Overview In mammals, the SCN comprises a pair of nuclei that undergo 24 h periodic waves of depolarization that are initiated within the dorsal cell mass and followed by depolarization of the ventral cells. SCN lesioning experiments performed in the 1970s showed that the SCN is crucial for 24 h rhythms of locomotor activity, feeding, drinking, and sleep. Conclusive evidence that the SCN contains a master pacemaker came from experiments in the 1990s showing that transplantation of SCN from tau (short-period) mutant hamsters into SCN-lesioned wild-type hamsters resulted in a period length matching that of the donor SCN. However, SCN transplantation into SCN-ablated animals fails to restore 24 h rhythms of glucocorticoid and melatonin, suggesting that direct neural projections are necessary for the full activity of the SCN. Recent studies have also identified several molecules that may play a role in SCN outputs, including transforming growth factor (TGF-)α, prokineticin 2 (PK2), and arginine vasopressin (AVP). Both TGF-α and PK2 are highly expressed in the SCN during the subjective night and reduce locomotor activity when injected into the third cerebroventricle. Interestingly, the receptor for PK2 (PKR2) is widely expressed in CNS regions, including the dorsomedial hypothalamic nucleus (DMH) and lateral hypothalamic area (LHA), areas that are important for activity, feeding, and sleep regulation. Thus, both hard-wired neural networks and secreted signals are involved in the generation of neuroendocrine and behavioral rhythms.

Suprachiasmatic nucleus neural circuits Neural projections of the SCN are quite restricted, largely confined to areas in the hypothalamus and midline thalamus (Figure 1). The vast majority of projections

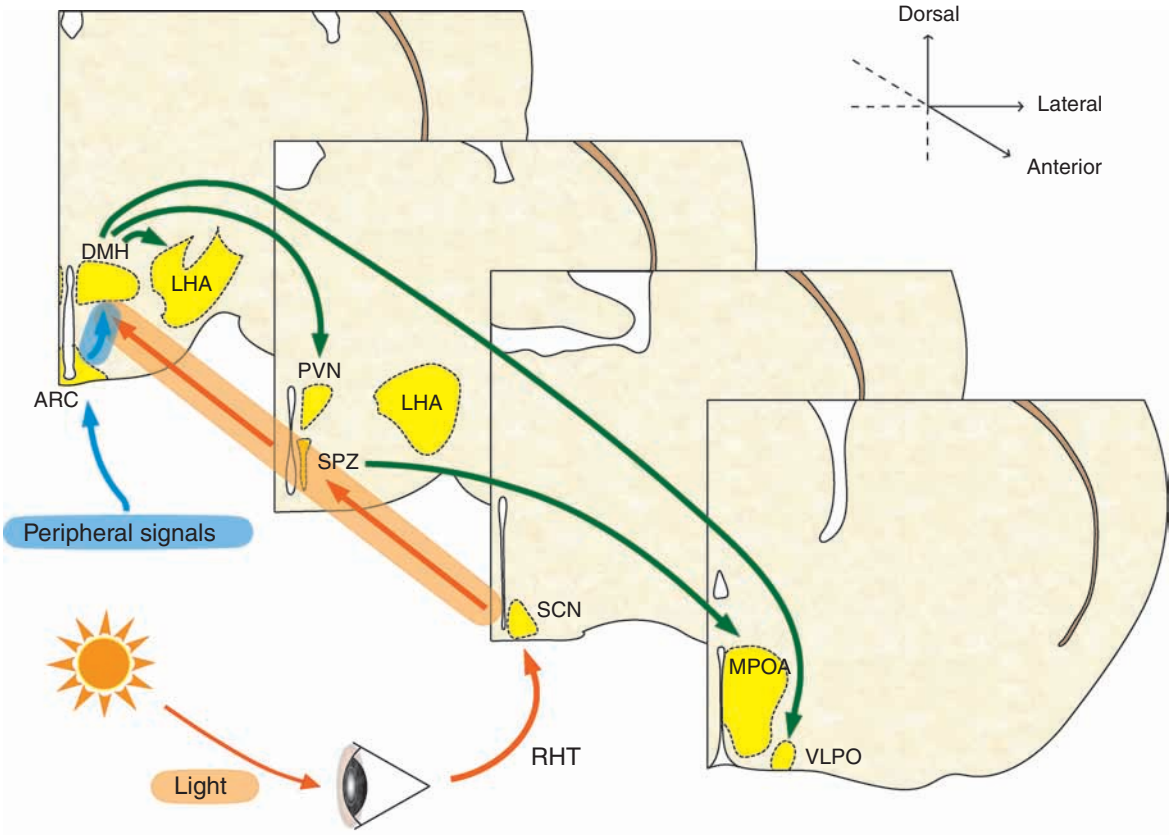

Figure 1 Neuroanatomy of pathways linking circadian and energy centers. Photic information received by retinal ganglion cells is transmitted to the CNS through the RHT, which terminates in retino-recipient cells in the SCN, the master pacemaker cell group. SCN cells then transduce photic information into neural or chemical signals which oscillate with a period of 24 h. The major targets of SCN projections are localized within two cell clusters of the medial hypothalamus, the subparaventricular zone (SPZ) and dorsomedial hypothalamus (DMH). The SPZ in turn innervates the medial preoptic area (MPOA) which plays an important role in both thermoregulation and reproductive function. The DMH projects to wide variety of nuclei involved in energy homeostasis (LHA), sleep (LHA and VLPO), and stress response (PVN). ARC, arcuate nucleus; CNS, central nervous system; DMH, dorsomedial hypothalamus; LHA, lateral hypothalamic area; MPOA, medial preoptic area; PVN, paraventricular nucleus; RHT, retinohypothalamic tract; SCN, suprachiasmatic nucleus; SPZ, subparaventricular zone; VLPO, ventrolateral preoptic nucleus.

from the SCN synapse on cells within the subdomain of two hypothalamic nuclei, localized to the subparaventricular zone (SPZ) and DMH. Axons from SPZ and DMH cell bodies, in turn, project to multiple CNS nuclei important in the regulation of behavior and endocrine function. A current model has proposed that the SPZ constitutes an intermediate relay station that amplifies and distributes signals from the SCN to other regions of brain. Indeed, lesioning studies have supported the hypothesis that the SPZ contributes to the regulation of daily rhythms of sleep, locomotor activity, and body temperature. In addition, DMH projections innervate the ventrolateral preoptic nucleus (VLPO) and LHA, two areas critical in sleep–wake regulation and energy balance. The DMH also projects to the paraventricular nucleus of the hypothalamus (PVN), which plays an important role in the hypothalamic–pituitary–adrenal axis stress response and autonomic tone. Because the DMH receives input from perikarya of the arcuate nucleus (ARC), it is conceivable that both DMH and SPZ may integrate energy and circadian systems.

In addition to preformed neural networks, recent studies indicate that intercellular communication is subject to highly dynamic changes involving synaptogenesis. Such synaptic changes may play an important role in generating daily rhythms of feeding behavior and in coordinating these processes with the 24 h sleep–wake cycles. For example, cells in the hypothalamus that express neuropeptide Y (NPY), agouti-related peptide (AGRP), cocaine- and amphetamine-related transcript (CART), and proopiomelanocortin (POMC) undergo marked changes in synaptic plasticity following an overnight fast and in response to leptin. Horvath and coworkers proposed that orexigenic (i.e., NYP/AGRP) neurons and anorexigenic (i.e., POMC/CART) neurons are reciprocally dominated by either inhibitory or stimulatory inputs that vary across the LD cycle.

Thus, it is also conceivable that daily synaptic rearrangement may impact the diurnal rhythms of neuropeptide gene expression, providing an additional link between circadian and metabolic homeostasis.

Neuroendocrine signals influence the suprachiasmatic nucleus clock Several lines of evidence indicate that the SCN also integrates metabolic signals from both CNS and peripheral tissues. First, neural projections into the SCN originate in several nuclei important in sleep and energy balance such as the raphe and hypothalamus. Second, SCN neurons express receptors for key hormones important in energy balance, including those for leptin. Third, leptin affects the phase of SCN oscillations. Fourth, many genes shown to oscillate in the SCN are involved in intermediary metabolism and energy balance. Thus, the SCN itself may play an important role in integrating environmental and metabolic information.

The Molecular Clock in Neural and Peripheral Metabolic Cells

The Discovery of Mammalian Clock Genes

The discovery of the molecular underpinnings of the circadian system has been a triumph of forward genetics (phenotype to gene) and recently provided further clues concerning interconnections between the circadian and metabolic systems. The molecular clock within the cell was first identified in the fruit fly *Drosophila melanogaster* and in the fungus *Neurospora*. Specifically, chemical mutagens were used to introduce mutations into the genomes of the fly and fungus, and circadian phenotypes were analyzed so researchers could then map and positionally clone the responsible genes. This mutagenesis approach led to the identification of the first clock mutations, *Period* (*Per*) and *Frequency*, in the 1980s. A similar mutagenesis approach was applied to the search for the clock genes in mammals in the early 1990s, resulting in the discovery in 1994 of the first mammalian clock gene, *Clock* (*circadian locomotor output cycles kaput*). The protein encoded by the *Clock* gene belongs to the superfamily of the basic helix–loop–helix (bHLH)-PAS (*Period-Arnt-Single-minded*) transcription factors that are conserved from flies to mammals. Recognition that bHLH-PAS domain proteins function as heterodimers provided the rationale for a two-hybrid screen that identified BMAL1 (or MOP3) as the major partner of CLOCK. CLOCK and BMAL1 heterodimerize in the cytoplasm and translocate to the nucleus to activate transcription from target genes containing an E-box enhancer element (Figure 2).

In addition to these transcriptional activators, the search for mammalian orthologs of the *Drosophila Per* gene led to the discovery of the mammalian *Per* genes (*Per1*, *Per2*, and *Per3*), paralogous members of the PAS protein family. The CLOCK–BMAL1 heterodimer was subsequently shown to activate *Per* gene expression through binding to the M34 E-box element in the *Per* promoter. PER proteins and cryptochromes (CRY1 and CRY2), members of the vitamin B_2-based blue-light photoreceptor/photolyase family, form a heteromultimeric complex that translocates to the nucleus to inhibit binding activity of the CLOCK–BMAL1 heterodimer. Thus, transcriptional activators and repressors form a feedback loop that generates 24 h cycles of gene transcription.

Metabolic Transcription Factors Participate in the Circadian Oscillator

Following the discovery of CLOCK, BMAL1, PER, and CRY proteins, subsequent analyses have revealed that several metabolic transcription factors also participate in generating circadian rhythms. The best characterized of these are REV-ERBα and the retinoic acid receptor-related orphan receptor α (RORα). REV-ERBα inhibits and RORα stimulates the transcription of the *Bmal1* gene by activating ROR elements (RORE) in the *Bmal1* promoter. A peak of *Rev-erbα* transcription occurs early to mid-circadian day and is antiphasic to the *Bmal1* peak, whereas *Rorα* does not oscillate. Hence, the circadian oscillation of *Bmal1* transcription is mediated by the reciprocal activities of REV-ERBα and RORα. In addition, transcription of the *Rev-erbα* and *Rorα* genes is also induced by CLOCK–BMAL1. Interestingly, both *Rev-erbα* and *Rorα* gene transcription increase during adipogenesis. Although interactions among REV-ERBα, RORα, and BMAL1 have been established, how these proteins might coordinate metabolic and circadian processes remains uncertain.

In addition to control at the transcriptional level, posttranslational modification also plays an important role in clock function. For example, several kinases, including casein kinase I epsilon (CKIε) (in mice) and CKII (in flies) have been shown to participate in the regulation of PER phosphorylation. Phosphorylation of PER is thought to trigger the translocation of the PER protein into the nucleus, where it then inhibits the CLOCK–BMAL1 heterodimer. In addition, glycogen synthase kinase 3 (GSK3), a central molecule in carbohydrate metabolism, has been shown to phosphorylate and stabilize REV-ERBα. Together, mounting evidence indicates that transcriptional and posttranslational modification of clock genes involves

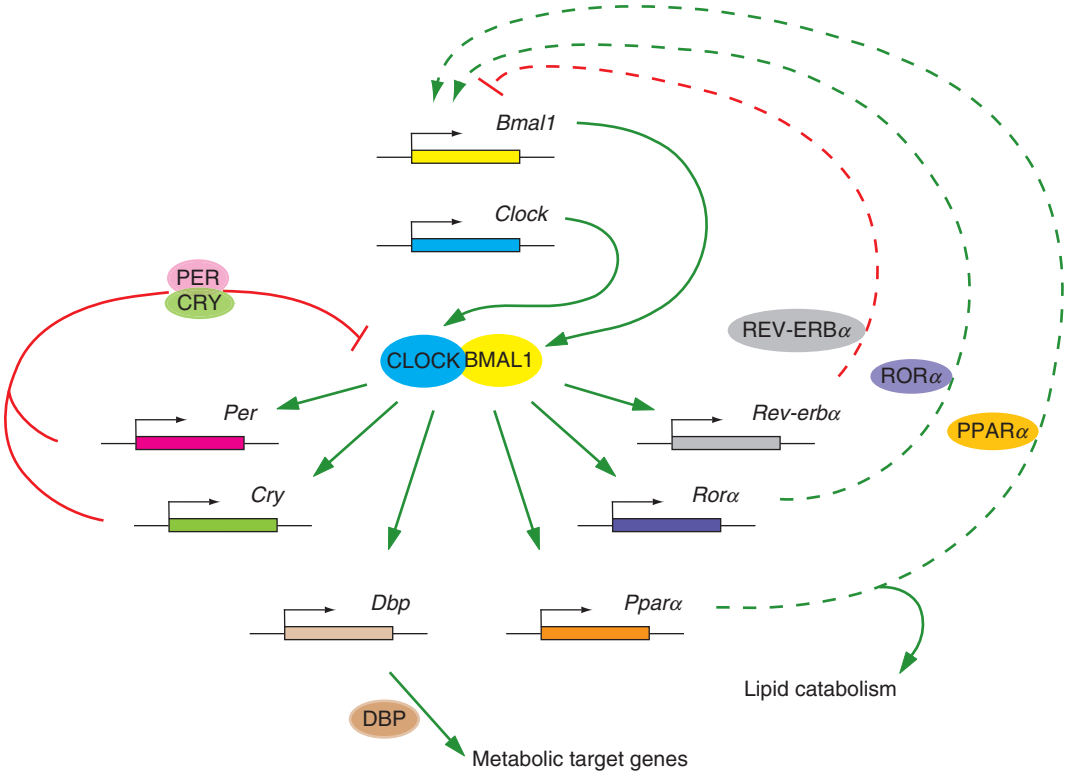

Figure 2 Overlapping gene network regulating the circadian and metabolic systems. The prevailing model illustrating the core autoregulatory feedback loop of the clock consists of transcriptional activators, CLOCK and BMAL1, and repressors, PER and CRY. The CLOCK–BMAL1 heterodimer activates the transcription of *Per* and *Cry* through the E-box enhancer element in their promoters. The PER and CRY proteins multimerize and translocate into the nucleus to inhibit the transcriptional activity of the CLOCK–BMAL1 heterodimer, resulting in reduced expression of the *Per* and *Cry* genes. Several metabolic transcriptional factors that have been shown to interact with clock genes are also illustrated. Particularly important transcription factors regulating both circadian and metabolic systems are REV-ERBα and RORα, which are both induced during adipogenesis. Transcription of these metabolic genes is also activated by CLOCK–BMAL1 through binding to the E-box element in their promoters. REV-ERBα and RORα reciprocally regulate *Bmal1* transcription through the RORE motif in the *Bmal1* promoter; interestingly, REV-ERBα inhibits and RORα stimulates *Bmal1* transcription. PPARα, a nuclear hormone receptor which regulates lipid catabolism, has also been shown to interact with BMAL1. Transcription of *Pparα* is activated by the CLOCK–BMAL1 heterodimer. Conversely, PPARα activates *Bmal1* expression through the PPRE in the *Bmal1* promoter. DBP is illustrated because it is among the best-characterized transcriptional outputs of the core clock pathway in liver cells that is directly linked to both sterol and glucose metabolism.

molecules previously known to mediate glucose and lipid metabolism, suggesting extensive molecular links between the circadian and metabolic pathways.

Clock Genes in Extra-Suprachiasmatic Nucleus Neurons and Peripheral Tissues

A remarkable series of studies by Schibler and colleagues first showed that circadian clock genes are expressed cell-autonomously in peripheral tissues (fibroblasts) and can be reentrained to a 24 h period by serum shock alone. Subsequent *in vivo* studies in transgenic and knockin mice expressing the bioluminescent reporter luciferase as a fusion to the PER protein established the ubiquity of peripheral clock gene expression. Studies using these and other models also showed that clock genes are expressed and function in extra-SCN regions of brain, including the ARC. Collectively, these findings have led to the prevailing model that places the SCN at the top of a neural hierarchy that maintains synchrony between the external environment and multiple behavioral and physiological processes (Figure 3).

Food Entrainment of the Clock

Studies of the anatomy and molecular machinery of the clock also opened a window on the co-regulation of feeding and circadian rhythmicity. In particular, the vexing question of how food and feeding affect the clock has continued to stimulate intensive investigation. A food-entrainable oscillator (FEO) is thought to reside outside of the SCN because SCN-lesioned animals still display an anticipatory burst of activity when food is presented during the subjective night. Yet the anatomic and molecular basis for this

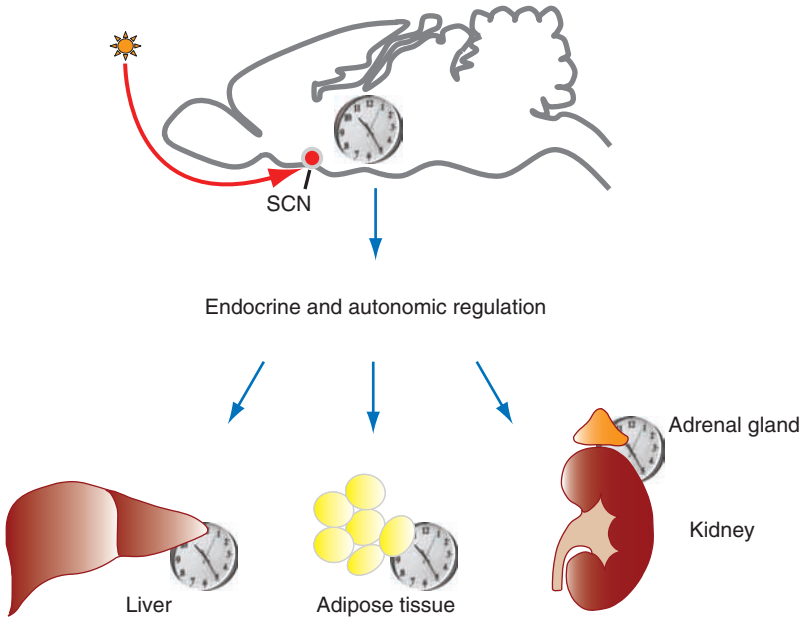

Figure 3 Model of the organization of circadian and metabolic systems. The molecular clock machinery in the suprachiasmatic nucleus (SCN) encodes the master neural pacemaker that is entrained to the daily light–dark cycle. The SCN in turn generates neural and endocrine circadian rhythms and synchronizes the clock machinery in peripheral tissues, as well as in extra-SCN neurons. Processes ranging from glucose and lipid metabolism in liver and adipose tissue to glucocorticoid secretion from the adrenal gland have been shown to be entrained to the SCN; however, each tissue also expresses the full repertoire of clock genes, and in each cell type the outputs of the clock network probably exert distinct physiological effects.

behavioral adaptation to food is still an area of active research. Saper and colleagues reported that DMH-ablated animals show dampened food anticipatory activity (FAA), suggesting that the DMH might contain the FEO. Further focus on the DMH also came from Yanagisawa and colleagues, who showed robust *Per2* oscillation within this location, suggesting a role for the DMH cell cluster in food entrainment. Interestingly, *Per2* mutant mice fail to respond to food entrainment. In contrast, *Clock* mutant mice have persistent FAA, perhaps suggesting differential effects of CLOCK versus PER2 on food-related activity. It is possible that the CLOCK paralog, neuronal PAS domain protein 2 (NPAS2), may compensate for CLOCK and preserve food entrainment in the *Clock* mutant mouse. Additional studies using both genetic models and anatomic approaches promise to increase our understanding of the nature, mechanism, and function of food-entrainable behavior.

Circadian, Sleep, and Metabolic Disorders in Humans

The aforementioned research on the molecular clock and neural clock circuits has also provided an important framework for emerging studies investigating the links between sleep, circadian disruption, diabetes mellitus, and obesity. Here we provide a brief overview of this line of research, highlighting findings that may offer insight into the pathophysiology of metabolic disorders.

Epidemiology of Sleep and Metabolic Disorders

The first series of studies to hint at connections between sleep and obesity came from observational reports on the association between sleep time and body mass index (BMI). Indeed, reduced sleep is a common behavioral perturbation associated with industrialization, the spread of artificial light, increased shift work, and jet lag. Recent studies also showed a correlation between sleep loss and the development of metabolic disorders. Although the precise molecular mechanisms remain unknown, a striking and consistent finding has been the connection between shift work and a propensity for metabolic syndrome in human subjects. In particular, shift work is associated with increased risk of impaired glucose tolerance, hyperinsulinemia, elevated leptin levels, hypertension, and dyslipidemia, all hallmarks of metabolic syndrome. This epidemiological evidence raises the question of whether circadian misalignment may be an etiological factor in abnormalities of glucose and lipid metabolism.

Clinical studies also uncovered a connection among circadian behavioral alterations, sleep, and obesity. One intriguing example is the night-eating syndrome, characterized by the constellation of fragmented sleep

with night-time binge eating. Although the cause of night-eating syndrome remains poorly understood, clinical studies in patients with the disorder have consistently revealed abnormalities in the circadian phase of food intake, often accompanied by obesity. Advanced sleep-phase syndrome (ASPS) is another human circadian disorder that gives rise to changes in sleep. Recently, a missense mutation in the human *Per2* gene was identified in subjects with ASPS. Interestingly, in the delayed sleep-phase syndrome (DSPS), subjects fall asleep and awake at a later time each day. DSPS has been associated with polymorphisms that may alter PER3 phosphorylation. Although still an uncharted area, the discovery of the molecular bases for both ASPS and DSPS now provides a new opportunity for researchers to examine the relationship between circadian gene disorders and metabolism in man.

Narcolepsy and Energy Balance

An additional intriguing link between sleep and energy regulation has emerged from studies of canine narcolepsy and the orexin system in the laboratories of Mignot, Yanagisawa, and Sutcliffe. Orexin is a neuropeptide produced by cells within the LHA and posterior hypothalamus that was originally identified by cDNA cloning of transcripts enriched in this region. The function of orexin was initially thought to involve the regulation of feeding behavior. Surprisingly, positional cloning revealed that mutations in the receptor for orexin give rise to canine narcolepsy. These findings raise the intriguing possibility that orexin may link the regulation of ingestive behavior with control of the sleep–wake state. In support of this hypothesis, Mignot showed that orexin levels in the cerebrospinal fluid of human narcoleptic subjects correlate with BMI. Genetic and pharmacological studies in experimental models have also provided evidence for a role of orexin in alertness, food intake, and energy expenditure. Furthermore, recent neural tracing studies uncovered projections from SCN neurons to orexin-producing cell bodies. Thus, a convergence of clinical, genetic, molecular, and pharmacological analyses points toward a key role for orexin in the neural integration of appetite, body weight, and sleep.

Molecular Interconnection between the Circadian and Metabolic Systems

Role of Cellular Redox State and Chromatin Remodeling in Clock Gene Transcription

Advances from several laboratories have also begun to identify nodal points through which molecular networks involved in both metabolism and circadian rhythms may communicate. Emerging research on nicotinamide adenine dinucleotide cofactors represents an important area of progress. One compelling hypothesis stems from the observations of McKnight and colleagues that demonstrated a link between cellular redox and the transcriptional activity of both the CLOCK–BMAL1 and NPAS2–BMAL1 heterodimers. Specifically, the reduced forms of the nicotinamide adenine dinucleotide cofactors (NAD(H) and NADP(H)) enhance DNA-binding of CLOCK–BMAL1 and NPAS2–BMAL1. Because the ratio of oxidized-to-reduced NAD and NADP cofactors is related to levels of glycolysis and pentose metabolism, these cofactors may act as sensors that couple circadian and metabolic processes.

Another intriguing candidate pathway linking circadian and metabolic processes has emerged from innovative work on chromatin remodeling. Specifically, recent findings suggest that CLOCK possesses histone acetyltransferase (HAT) activity and that chromatin remodeling is important in sustaining oscillations of gene transcription. Because NAD also plays a pivotal role in transcriptional silencing by histone deacetylases, chromatin remodeling may constitute an additional mechanism for metabolic and circadian communication. Together, these observations highlight a potential broad role for NAD cofactor oxidation and chromatin remodeling in the interconnection between behavioral and metabolic outputs of the clock.

Insight from Experimental Genetic Models into Links between Circadian and Metabolic Systems

In the past few years, genetic models in mice have begun to uncover an important role of clock genes in glucose and lipid metabolism *in vivo*. The findings that animals harboring a mutation in the *Clock* gene are hyperphagic and obese first demonstrated that clock genes are necessary for the maintenance of energy balance. As already reviewed, BMAL1 has also been shown to directly affect nuclear hormone receptors important in lipid catabolism and lipogenesis. However, important questions remain concerning whether the genetic disruption of clock genes alters metabolic regulation through direct effects or through mechanisms secondary to the (circadian) behavioral changes in these model animals. Interestingly, within the brain, RNA expression is affected by both circadian time and by the behavioral state of the animal (i.e., sleep and wakefulness). In addition, metabolic genes involved in hepatic lipid metabolism display broad circadian-dependent patterns of expression, and the expression of hepatic RNAs can be shifted by scheduled feeding. In the mediobasal

hypothalami of *Clock* mutant mice, an area outside of the SCN, extensive changes occur in the expression of both core clock genes and genes encoding neuropeptides that regulate organismal energy balance. Thus, an exciting area of investigation involves the elucidation of the role of clock gene expression within extra-SCN neural circuits. Finally, a fascinating yet unanswered question is whether perturbation of metabolic signaling pathways might cause reciprocal changes in clock gene expression.

Future Questions

Studies of the link between circadian and metabolic processes offer an exciting opportunity for us to delve into molecular mechanisms controlling both behavior and energy balance. Significantly, the discovery of clock genes has established the details of the molecular control of circadian rhythms that were previously invisible. Many investigators are now poised to merge analyses at the physiological, molecular, and genetic levels to further dissect the molecular and neural bases for disorders of timing, sleep, diabetes mellitus, and obesity.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Clock Gene Regulation of Endocrine Function; Metabolic Syndrome and Sleep; Narcolepsy; Neuropeptides: Food Intake.

Further Reading

Antle MC and Silver R (2005) Orchestrating time: arrangements of the brain circadian clock. *Trends in Neuroscience* 28: 145–151.

Bass J and Turek FW (2005) Sleepless in America: A pathway to obesity and the metabolic syndrome? *Archives of Internal Medicine* 165: 15–16.

Elmquist JK, Elias CF, and Saper CB (1999) From lesions to leptin: Hypothalamic control of food intake and body weight. *Neuron* 22: 221–232.

Foster RG and Wulff K (2005) The rhythm of rest and excess. *Nature Reviews Neuroscience* 6: 407–414.

Hastings MH, Reddy AB, and Maywood ES (2003) A clockwork web: Circadian timing in brain and periphery, in health and disease. *Nature Reviews Neuroscience* 4: 649–661.

Horvath TL and Diano S (2004) The floating blueprint of hypothalamic feeding circuits. *Nature Reviews Neuroscience* 5: 662–667.

Ko CH and Takahashi JS (2006) Molecular components of the mammalian circadian clock. *Human Molecular Genetics* 15(supplement 2): R271–R277.

Kohsaka A and Bass J (2007) A sense of time: how molecular clocks organize metabolism. *Trends in Endocrinology and Metabolism* 18: 4–11.

Kriegsfeld LJ and Silver R (2006) The regulation of neuroendocrine function: Timing is everything. *Hormones and Behavior* 49: 557–574.

Mistlberger RE (1994) Circadian food-anticipatory activity: Formal models and physiological mechanisms. *Neuroscience and Biobehavior Reviews* 18: 171–195.

Ramsey KM, Marcheva B, Kohsaka A, and Bass J (2007) The clockwork of metabolism. *Annual Review of Nutrition* 27: 219–240.

Roenneberg T and Merrow M (2005) Timeline: Circadian clocks – the fall and rise of physiology. *Nature Reviews Molecular and Cellular Biology* 6: 965–971.

Saper CB, Lu J, Chou TC, and Gooley J (2005) The hypothalamic integrator for circadian rhythms. *Trends in Neuroscience* 28: 152–157.

Schibler U and Naef F (2005) Cellular oscillators: Rhythmic gene expression and metabolism. *Current Opinion in Cell Biology* 17: 223–229.

Shanahan TL and Czeisler CA (2000) Physiological effects of light on the human circadian pacemaker. *Seminars in Perinatology* 24: 299–320.

Staels B (2006) When the Clock stops ticking, metabolic syndrome explodes. *Nature Medicine* 12: 54–55.

Stephan FK (2002) The "other" circadian system: Food as a Zeitgeber. *Journal of Biological Rhythms* 17: 284–292.

Cochlear Development

D M Fekete, Purdue University, West Lafayette, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

The Structure of the Cochlea

The inner ears lie deep in the temporal bones on either side of the hindbrain (Figure 1). In the mammalian inner ear, the ventral portion of the inner ear forms the spiraling cochlea. Everything about the cochlea and the structures within it varies longitudinally from its wide basal turn to its narrow apical turn. Chamber dimensions, epithelial width and thickness, cell numbers, cell dimensions, and innervation patterns all vary systematically along the axis that correlates with frequency selectivity from high frequency (located at the base) to low (located at the apex). For example, the basilar membrane is narrower and stiffer at the base compared to the apex. The cochlea is somewhat analogous to a bicycle wheel comprised of an inner tube surrounded by a tire; only both tube and tire are filled with fluid rather than air, and the structure is spiral rather than truly circular. The 'tire' rests upon the spiral ganglion (the rim of wheel). The axons of the ganglion neurons radiate inward (the spokes of the wheel), gather together (at the hub) to project to the brain as the auditory nerve. The innertube is called scala media and it is filled with a fluid of unusual ionic composition called endolymph. Endolymph is high in potassium and low in sodium. In cross section, scala media looks triangular (Figure 1). One side of the triangle is formed by Reissner's membrane, a thin two-layered epithelial sheet. It is flanked on one side by scala media and on the other side by another fluid-filled chamber, scala vestibuli. The fluid within scala vestibuli is the perilymph. It is typical of normal extracellular fluid, being high in sodium and low in potassium. When viewed in cross section, the basilar membrane forms a suspension bridge across the posterior side of the triangle. This tough elastic matrix supports the organ of Corti, which is a thin, ribbon-shaped tissue that contains mechanosensory hair cells. The apical surfaces of the hair cells are each adorned with a stereociliary bundle that faces scala media. Both the basilar membrane and its passenger, the organ of Corti, are displaced by fluid pressure changes in the cochlea. The opposite side of the basilar membrane contains fibroblasts and blood vessels, and it faces a perilymph-filled chamber called scala tympani. The third side of the triangle is plastered up against the outer wall of the cochlea: it contains a complex tissue called the stria vascularis which is notable for its essential role in endolymph secretion.

The Embryonic Origins of the Cochlea

The epithelial cells lining scala media all share a common embryonic origin in the ectoderm of the otic placode. This includes the hair cell mechanoreceptors, their supporting cells, the apical layer of the stria vascularis, the innermost layer of Reissner's membrane, and the other nonsensory epithelial cells that connect these structures. Together this epithelial sheet tightly seals and ionically isolates the endolymphic fluid by means of junctional connections. The neurons of the auditory (spiral) ganglion also derive from the otic placode, having migrated away from the otic ectoderm very early in embryogenesis. Nearly all of the other soft tissue components housed within the cochlea are of mesenchymal origin, with the exception of neural-crest derivatives that will become melanocytes of the stria vascularis and glial cells of the auditory nerve and ganglion. The entire membranous labyrinth is surrounded by the otic capsule which begins as cartilage but is gradually replaced by bone; both the cartilage and the bone are of mesenchymal origin. The development of both the perilymph chambers and the otic capsule are strongly influenced by factors released from the placode-derived epithelial core, which arises earlier in embryogenesis. Because considerably more is known about the molecular basis of scala media development as compared to the structures around it, these otic-placode-derived cells will be the focus of this article.

Otic Induction

The otic placode is a thickened ectodermal patch located adjacent to the dorsolateral hindbrain and above the branchial arches. Over the course of 1–2 days in the mouse, the otic placode dips into the mesenchyme, forming a cup. The rim of the otic cup constricts, eventually pinching off from the overlying ectoderm to release the otic vesicle into the tissue below. The otic vesicle then begins an extended process of growth and shape changes, called morphogenesis (Figure 2). Not only does it give rise to the cochlea and the neurons of the auditory (spiral) ganglion, but it also forms inner ear vestibular structures including the utricle, the saccule, the semicircular canals, and the vestibular ganglion.

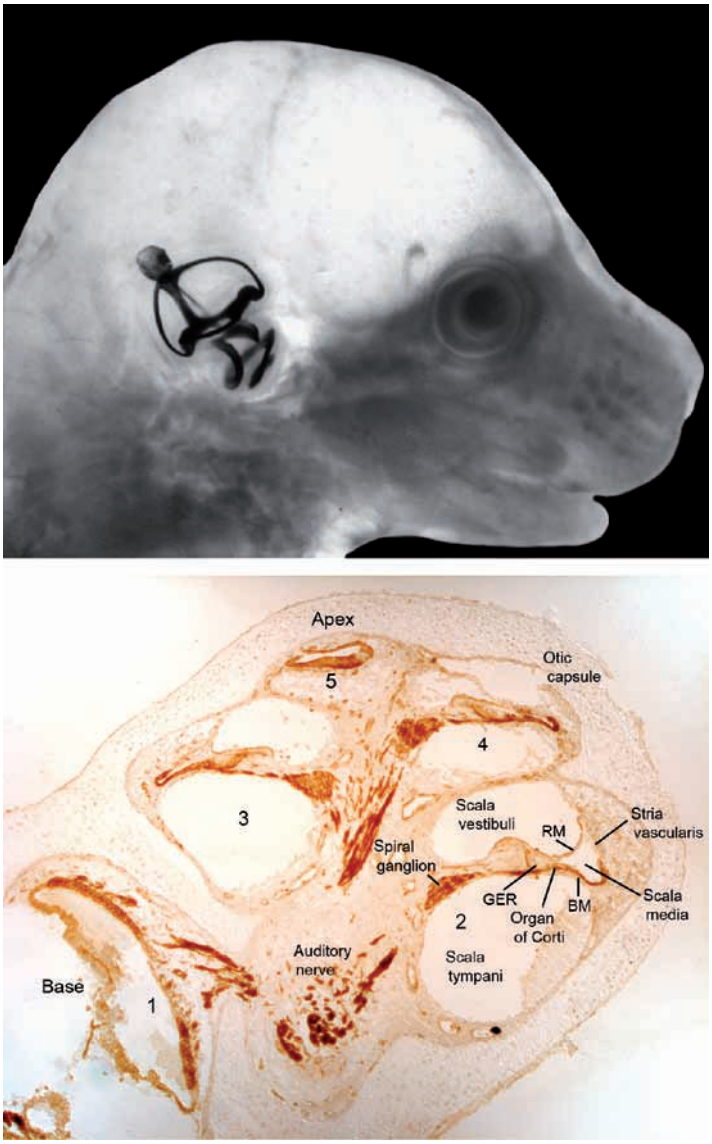

Figure 1 Gross anatomy and histology of the developing cochlea. (Top) The fluid compartment of the mouse inner ear was filled with opaque paint solution at embryonic day 16 and photographed *in situ*. The brain was removed to provide better contrast, and the image was photographically reversed. (Bottom) Histological section through a newborn mouse cochlea immunolabeled for integrin alpha6 (dark brown). This labels axons of the auditory nerve and a subset of epithelial cells lining scala media. Cross sections through the spiraling cochlea are numbered 1–5 from the basal to the apical ends. At this age, the apical turn is less well developed than more basal turns, due to the basal-to-apical gradient of differentiation. Throughout the cochlea, the greater epithelial ridge cells located next to the organ of Corti have not yet undergone programmed cell death, so the epithelium is thicker than in the mature cochlea. Histological artifacts have caused a collapse of Reissner's membrane onto the organ of Corti in some of the cochlear turns. BM, basilar membrane; GER, greater epithelial ridge; RM, Reissner's membrane. (Top) Photograph courtesy of John Brigande (Oregon Health Sciences University). (Bottom) Image generously provided by Donna Whitlon (Northwestern University).

In mammals, the otic placode arises at the level of rhombomeres 5 and 6 and depends upon hindbrain-derived signals for its induction. The endoderm and mesenchyme located beneath the otic placode are also essential for otic induction. Fibroblast growth factors (Fgfs) arising from all three of these tissues convert general preplacode ectoderm to a pre-otic fate. The pre-otic state can be recognized by the expression of the transcription factor, Pax2. Pax2-expressing cells are dispersed within and beyond the otic field, and gradually many of these cells gather together to form the otic placode. The otic state is stabilized by exposure to Wnt ligands. In mice, these essential otic-inducing factors arise sequentially from several different tissue sources. Fgf8 is released from the endoderm, causing the subotic mesenchyme to secrete

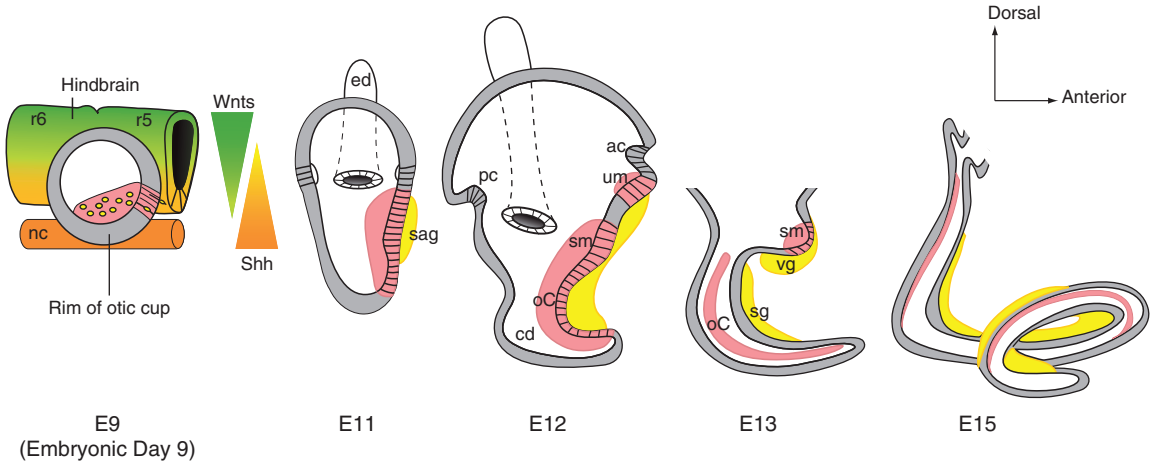

Figure 2 Origin and morphogenesis of the embryonic mouse cochlea. The otic placode originates adjacent to rhombomeres 5 and 6 and deepens to form the otic cup. The hindbrain and notochord generate sources of Wnts and Sonic hedgehog (Shh) that give the inner ear dorsal–ventral polarity. The ventral floor of the otic cup is a neurosensory domain which generates both neuroblasts (yellow) that colonize the statoacoustic ganglion and sensory primordia (pink). The stages of cochlear morphogenesis are shown schematically for the medial half of the otocyst which is portrayed as though it was sectioned in half along the parasagittal plane. cd, cochlear duct; ed, endolymphatic duct; ac, anterior crista; nc, notochord; oC, organ of Corti; pc, posterior crista; r, rhombomere; sag, statoacoustic ganglion; sg, spiral ganglion; sm, saccular macula; um, utricular macula; vg, vestibular ganglion. Staging of the mouse ear is based on the work of Morsli H, Choo D, Ryan A, Johnson R, and Wu DK (1998) Development of the mouse inner ear and origin of its sensory organs. *Journal of Neuroscience* 18: 3327–3335.

Fgf10. Fgf10 then signals to the hindbrain, directing it to secrete both Fgf3 and Wnts. In mice, otic induction is thought to require a combination of mesenchymal Fgf10, neural Fgf3, and neural Wnts, all acting directly on the placodal ectoderm. A direct role for Fgf8 on otic placode is more difficult to ascertain since it is expressed by a number of embryonic tissues, including the otic placode itself.

Axial Polarity and Patterning of the Cochlea

Otic vesicle formation occurs somewhat late relative to the appearance of other embryonic structures, such as the neural tube and the notochord. As a result, the otic ectoderm is poised to use signals arising from these surrounding structures not only for its induction, but also for subsequent patterning. Otic patterning begins at the otic cup stage and is evidenced by asymmetries in gene expression.

The neurogenic portion of the otocyst is located on its ventral–anterior floor (Figure 2). It can be recognized in the mouse by expresson of neurogenin-1 (Ngn1), a gene required for specification of neuroblasts of the otic placode as well as several other neurogenic placodes in the head. Neuroblasts leave the Ngn1 domain from its anterior pole, delaminate from the epithelium into the underlying mesoderm, and migrate medially to colonize both the auditory and vestibular ganglia.

The cochlear duct elongates from the ventral otocyst. The dorsal–ventral (D–V) axis of the duct will come to represent the basal–apical axis of the cochlea. The major tissue types lining the cochlear duct (the organ of Corti, Reissner's membrane, and the stria vascularis) appear to originate from specific positions along its circumference. The primordia for each tissue may be marked by the expression of different combinations of genes as early as E11.5 as discussed below, although this needs to be confirmed by fate mapping. Nonetheless, it is likely that the establishment of anterior–posterior (A–P) and medial–lateral (M–L) axial polarity at the otocyst stage will be related to the specification of cochlear tissue types. Furthermore, D–V axial polarity could underlie the appearance of molecular and structural gradients that mediate frequency selectivity along the basal–apical axis of the mature cochlea.

The regulation of otocyst D–V polarity is best understood in molecular terms. Signals are secreted from the hindbrain and possibly the notochord to influence D–V polarity. Mouse mutants and experimental manipulations of chicken embryos both reveal that Wnts can dorsalize the otocyst while sonic hedgehog (Shh) can ventralize otocyst. Since the cochlea arises ventrally, it has a special need for Shh (Figure 2). The stage of fixation of an otic axis is defined as the earliest time when rotation of the otic vesicle with respect to its surroundings no longer regulates to form an ear of normal orientation.

These experiments have been done in chickens, where there is an elongated cochlear duct that does not spiral. Results show that a cochlea often projects ventrally from a D–V reversed otocyst, indicating that the D–V axis was not fixed at the time of rotation and can be respecified by the surrounding environment if the manipulation is performed before cochlear morphogenesis begins.

There is good evidence that M–L polarity is also under the influence of the hindbrain. Mutations that reduce hindbrain Fgf3 levels are associated with severely underdeveloped otic vesicles. Fgf3 is high in rhombomeres 5 and 6 located adjacent to the otocyst. Retinoic acid, which is enriched in the hindbrain, is also essential for otic development, although it is difficult to distinguish its direct versus indirect effects on the otic ectoderm. For example, retinoic acid can influence rhombomere identity via its effects on Hox genes, and this could indirectly affect the ear by changing the levels of Fgf3 released from the hindbrain. More direct effects of hindbrain-derived retinoic acid are also possible, particularly since retinoic acid receptors are present in the otic ectoderm. Both changes in retinoid signaling and defects in hindbrain rhombomere identity can alter the normal M–L asymmetries in gene expression in the otocyst. Rotations of the otocyst that reverse the M–L axis in chickens generate cochlear ducts in their normal location, but generally the auditory sensory organ fails to differentiate.

The polarity of the third otic axis, anterior–posterior (A–P), is the least understood in terms of molecular signals. This axis is the first of the three to be fixed in the chicken otocyst. Surprisingly, though, the relevant signals do not appear to originate from the hindbrain, as evidenced by failure of hindbrain A–P rotations to reverse the A–P axis of the otocyst even when performed prior to the stage of fixation. Nonetheless, the environment does influence the direction of cochlear outgrowth, since even in A–P reversed ears where the cochlea originates from an abnormally anterior location, the sickle-shaped cochlear duct of the chicken assumes its normal arch in a ventral–medial direction. Presumably, these signals must arise from the mesenchyme surrounding the ventral otocyst.

In summary, signals from surrounding tissues influence otocyst polarity. The requirement for some of these signals, such as Shh, is maintained for an extended period to complete the task of ventralizing the part of the otocyst that generates the cochlea. There are no data yet regarding whether a ventral-to-dorsal gradient in the concentration of Shh protein may underlie the tonotopic organization of the organ of Corti that will later be arrayed along the same axis.

Morphogenesis of the Cochlea

The cochlear duct arises as a tubular outgrowth from the ventral otocyst. It is flattened somewhat along its medial–lateral dimension. As the duct elongates, it makes a sharp anterior turn to run nearly parallel to the A–P axis in mice. Then, the tip of the duct loops upward (dorsally), and laterally to generate the first coil (**Figure 2**). In the mouse, the cochlea stops elongating after 1.75 coils, while other mammals make 1.5–4 coils. The human cochlea coils 2.5 times. Virtually nothing is known about the cellular or molecular basis of cochlear coiling. However, there is some progress on the mechanism of cochlear elongation. At first glance, one might expect such directed outgrowth to be primarily based on cellular proliferation, and certainly this is important. Cell divisions continue throughout the period of duct elongation, at least in nonsensory regions.

In addition to control of cell numbers, it is also theoretically possible to regulate the overall dimensions of an epithelial tube such as the cochlear duct through oriented cell divisions, directed cell movements, and cell shape changes at the level of individual cells. Such mechanisms require that the cells sense the overall polarity of the epithelial sheet in which they reside and orient in response to it through a process known as planar cell polarity (PCP). Indeed, elongation of the cochlea shares a molecular PCP program in common with the elongation of the A–P axis of vertebrate body during gastrulation and neural tube closure. At the cellular level, the morphogenetic processes of gastrulation utilize a choreographed intercalation of cells called convergence-extension that serves to narrow and elongate an epithelial sheet. In zebra fish, oriented cell divisions, whereby the mitotic spindle is aligned with the elongating A–P axis, also drive axis elongation during gastrulation and are regulated by components of a Wnt–PCP pathway.

In the cochlea, data at the resolution of individual cells are lacking, but by analogy to the gastrula, the prediction is that cells will converge along a 'medial–lateral' axis running across the width of the duct and/or they will divide orthogonally to this along the apical–basal axis. Convergence-extension would serve both to narrow a strip of epithelium along the axis of directed cell movement and to lengthen it along the apical–basal dimension (**Figure 3**). Extension of the sensory primordium progresses for several days after the cells have stopped dividing and thus oriented cell divisions cannot explain its lengthening. However, this mechanism could be active in the flanking nonsensory regions, since they remain mitotically active during elongation.

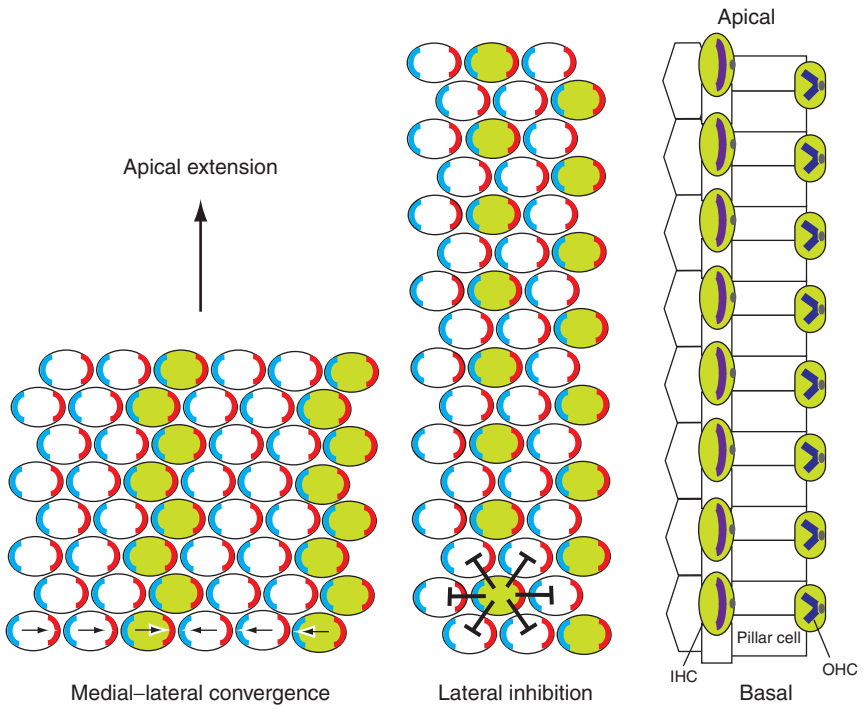

Figure 3 Convergence-extension of the cochlear duct and lateral inhibition of hair cells. A schematic surface view of the development of a strip of the organ of Corti from epithelial cells on the medial side of the inner hair cell row to the first row of outer hair cells. Elongation of the cochlear sensory epithelium in the apical direction could result from directed medial–lateral convergence and intercalation of cells, leading to a longer and thinner shape as shown. Thinning of the cochlear epithelium in the third dimension, within the plane of the paper, involves radial intercalation (not shown) that could also squeeze the cells in the apical dimension; this thinning begins at E14.5, when the duct has already completed approximately one full turn. The polarized localization of hypothetical PCP proteins, shown in blue and red, could provide the cells with directional cues both for medial–lateral convergence underlying cochlear elongation and for polarizing the hair bundle. This may explain why both processes are defective in PCP mutants. Differentiated hair cells are always separated from each other by supporting cells, and this mosaic arrangement probably involves a lateral inhibitory mechanism mediated by Notch signaling. Both the kinocilium and the localized expression of PCP proteins will disappear with maturation of the bundle (see **Figure 5**). IHC, inner hair cell; OHC, outer hair cell.

A second organized cell movement, that of radial intercalation, has been proposed to underlie elongation of the sensory epithelium. During cochlear outgrowth, the sensory primordium reduces from a thickness of several cells to only two cells on average. The cells are presumed to be squeezed apically as a result of these cell rearrangements.

Evidently apical elongation of the cochlea is not simply a question of growth in the direction of least resistance because only apically directed elongation is observed when a section of embryonic mouse cochlea is isolated in culture, even though both ends should be free to elongate under these conditions.

Several mutant strains of mice show shorter and wider cochlear ducts that are found together with defects in neural tube closure. Examples include Dishevelled 1 and 2 double-knockout mice (Dvl DKO) and looptail mutant mice (defective in the Ltap gene). These genes are active in the PCP pathway. In PCP mutants, cochlear outgrowth is stunted even though radial thinning progresses on schedule, leaving medial–lateral intercalation and oriented cell divisions as potential mechanisms that are affected by the mutations.

Some proteins in the PCP pathway have been shown to be asymmetrically organized in individual cochlear cells along the axis of convergence as would be expected if they are endowing the cells with directional sensitivity. However, it is unclear if the timing of this asymmetry is early enough to account for polarized extension of the cochlear duct. Asymmetry in protein localization is usually not detected until E18.5 in the mouse organ of Corti, at which point the cochlea has already completed 1.75 turns. By this age, PCP mutant mice already have shorter and wider cochlear ducts. Interestingly, the direction of medial–lateral cell intercalation is also the primary axis of polarity for hair cell bundles. The same PCP genes that regulate cochlear elongation also affect hair bundle orientation, as discussed later.

What initiates cochlear duct outgrowth? The process fails entirely in knockout mice that lack Eya1,

Six1, Gata3, and Tbx1, genes that are expressed very early in the otic ectoderm and, in the case of Tbx1 and Six1, also in periotic mesenchyme. In Tbx1 and Gata3 mutants, early axial polarity and initial regional specification appear unaffected because generation of the endolymphatic duct, cochleovestibular ganglion, and some sensory specification is observed. Instead, it is the morphogenetic process that fails. In Tbx1 nulls, this may be because the normal boundaries between gene expression compartments break down. In Gata3 nulls, the otocyst separates into dorsal (endolymphatic) and ventral (neurosensory) halves, apparently in response to disruptions in dorsal–ventral gene expression and cell adhesion molecules. The ganglion cells delaminate but fail to express Fgf10 and there is progressive loss of the spiral ganglion. In Eya1 and Six1 nulls, the otocyst itself fails to progress beyond the cystic stage, the otic ganglion is missing, and there is increased cell death. The expansion of dorsal genes into the ventral otocyst in Six1 nulls may explain the failure to develop ventral structures such as the cochlea and the ganglion, although there is also extensive cell death in the early otocyst. Six1 expression depends upon Eya1.

Cases where the cochlear duct is present but abnormal suggest perturbations at later stages. Defects in cochlear length or coiling occur in knockouts of several transcription factors that are expressed in the otic ectoderm, including Pax2, Ngn1, Otx1, Otx2, and a weak allele of Eya1 (Eya1-bor). Of these, Pax2 has the most severe phenotype. It is probable that these genes act cell autonomously in the normal animal. A frameshift mutation in Pax2 underlies a syndromic form of human deafness, coloboma syndrome. Even when the cochlear duct is severely misshapen, considerable specification of different cell and tissue types can occur. For example, neuroblasts, sensory cells, and adjacent nonsensory cell types arise in the Pax2 null mouse but then show later defects in survival (neuroblasts) or spatial patterning. Pax2 requires Shh to maintain its expression in the developing cochlear duct. In contrast to Pax2, the shortened cochlea in Brn4 nulls is probably an indirect effect since this transcription factor is normally expressed in the periotic mesenchyme at the time of duct outgrowth.

Specification of Tissue Types

The major tissues of the cochlear duct derived from the otocyst are the organ of Corti, the stria vascularis, and Reissner's membrane. Figure 4 shows schematically the origins and some gene expression patterns that precede specification of the cochlear tissues. Cochlear prosensory tissues and the greater epithelial ridge cells (which degenerate during the first postnatal week) appear to originate from the medial wall of the duct that expresses lunatic fringe (Lfng) and Pax2 at E10.5. One day later, Lfng and Pax2 expression domains separate, with Lfng shrinking to a small domain that may represent either a sensory-competent state, or perhaps it is the earliest marker of the primordium of the organ of Corti. The cells opposite to this, on the posterior–lateral wall, express Pax2 at E10.5–E11.5 and will eventually become the stria vascularis. In the mature cochlea, the sensory and strial tissues are separated by nonsensory cells, including Hensen's cells and Claudius' cells; they originate from progenitors of the posterior–medial cochlear duct that co-express Bmp4, Pax2, and Lfng at E10.5, with persistent expression of Bmp4 on E15.5. Reissner's membrane originates from the anterior–lateral quadrant of the E10.5–E11.5 cochlear duct, marked by Otx2 expression. The anatomical position of cochlear duct tissues relative to the body axis is significantly altered by the spiraling of the cochlea. So while the organ of Corti is presumed to arise from the Lfng domain on the anterior–medial wall of the ventral otocyst on E11.5, it is eventually located along the posterior wall of the triangular scala media compartment.

Specification of Sensory Cells

Sensory specification of the organ of Corti is an extended process. The transcription factor, Sox2, plays a fundamental role in prosensory specification. A strong mutant allele of this gene, called light coat and circling (Lcc), fails to form any recognizable sensory domain in the cochlea. Sox2 expression is present as early as E9.5 in the mouse. Fgf10 is also expressed at early stages in prosensory domains. Mutants or treatments that block Fgf signaling can significantly reduce the number and arrangement of sensory cells in the organ of Corti.

Cells of the sensory primordium are the first in the cochlear duct to leave the cell cycle. Molecular regulation of cell cycle withdrawal is correlated with expression of the cell-cycle inhibitor, p27kip1. In the absence of this gene, organ of Corti progenitors exit the cell cycle at a later stage than normal, with the result that supernumerary differentiated sensory cells are formed.

Molecules of the Notch signaling system are also expressed within and around the prosensory domain and are essential for its development. The receptor, Notch1, is broadly expressed in the ventral otocyst while several different Notch ligands (Deltas and Jaggeds) are expressed at various time points. Perturbations in the Notch signaling pathway can alter the number of sensory cells in different and seemingly

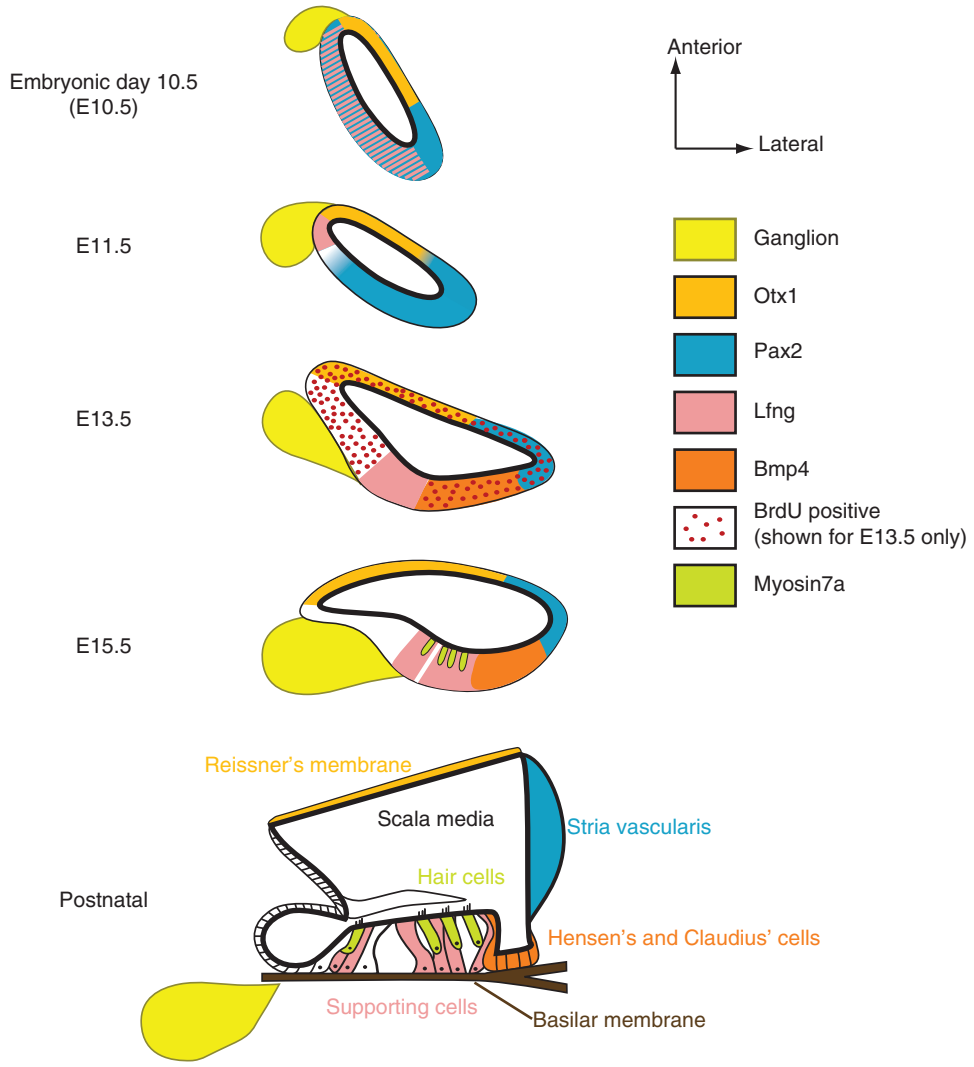

Figure 4 Gene expression domains in the otocyst may correspond to a fate map of cochlear tissue types found in the mature cochlea. The images represent approximate horizontal cross sections through the ventral otocyst (E10.5–E11.5) or the mid-base of the cochlea (E13.5–postnatal). The successive changes in the gene expression domains of the epithelial cells of the cochlear duct are shown. Note that Bmp4 expression is also present at E10.5 and E11.5 in the posterior vesicle but is not shown due to its partial overlap with other genes at these stages. The initial appearance of a zone of nonproliferation reveals the location of the sensory primordium at E13.5, which can be visualized by the absence of DNA synthesis by BrdU (bromo-deoxyuridine) labeling. The thick black line denotes the location of junctional complexes that serve to isolate the endolymphatic fluid. The figure is primarily based on data presented by Burton Q, Cole LK, Mulheisen M, Chang W, and Wu, DK (2004) The role of Pax2 in mouse inner ear development. *Developmental Biology* 272: 161–175; and Chen P, Johnson JE, Zoghbi HY, and Segil N (2002) The role of Math 1 in inner ear development: Uncoupling the establishment of the sensory primordium from hair cell fate determination. *Development* 120: 2495–2505.

paradoxical ways. When this signaling pathway is ectopically activated prior to sensory specification, extra sensory patches are formed in the chicken cochlea. In other words, in a gain-of-function scenario, Notch activation can induce sensory fate. The Notch ligand operative in this inductive event is not known, but based on loss-of-function phenotypes and the timing of expression, it is likely to be Jagged1. However, the Notch receptor is not yet known. Interestingly, ectopic activation of Wnt/β-catenin signaling has similar prosensory effects to ectopic Notch activation in the chicken cochlea.

Notch signaling is active at later stages, when hair cells and supporting cells are acquiring their fates. Here, Notch has been associated not with a prosensory fate but rather the opposite: lateral inhibition of hair cell (i.e., sensory) fate. The process of lateral inhibition is thought to be initiated when some cells begin to acquire a hair cell fate. These nascent hair cells use the Jagged2 and Delta1 ligands to activate Notch1 on neighboring cells that make direct contact with them. The receiving cells will downregulate Jagged2 and Delta1 genes, which reduces their effectiveness at signaling back to the nascent hair cell

(however, in mouse, it seems that only the hair cells ever express these genes so the negative feedback may not be necessary). Upon reception of Notch signaling, the surrounding cells will increase the expression of Hairy/Enhancer of Split (Hes) genes which can block proneural genes of the basic helix–loop–helix family. This includes the proneural gene, Atonal1, that promotes hair cell fate. Once inhibited from assuming the primary hair cell fate, these progenitors will instead take on a supporting cell fate as a secondary option. Overall, Notch signaling from nascent hair cells to surrounding precursors generates a perfect mosaic arrangement where every hair cell is surrounded by supporting cells and avoids contacting any other hair cells. Notch signaling may also be involved in keeping supporting cells in a postmitotic state, since they continue proliferating when Notch signaling is perturbed. However, this effect may be secondary to the increase in hair cells and/or the reduction in supporting cells. In the chicken cochlea, Notch signaling has been proposed to have yet another role in supporting cell development: among the pool of cells that will not become hair cells, reciprocal cell–cell signaling via the Serrate1 (Jagged1) ligand may promote or induce the supporting cell fate.

Hair Bundle Polarity

Once hair cells are specified, they begin a process of differentiation that is most obvious at their lumenal ends, where they elaborate a stereociliary bundle (Figure 5). The process begins with the elongation of numerous microvilli surrounding a central true cilium, called the kinocilium. The kinocilium and the nascent stereocilia are connected by thin fibrous strands. Next, the kinocilium moves to the periphery of the apical surface and for the first time a clear planar polarity is evident. However, the initial polarity assumed by this repositioning of the kinocilium is only partially aligned with adjacent hair cells or with the final polarity of the bundles in the mature ear. Instead, a refinement in orientation occurs later by an unknown process: it is not clear if it is only the bundle that rotates, or the entire cell. Before this reorientation, the stereocilia will lengthen, beginning with those nearest to the kinocilium that will become the tallest row of stereocilia. Lengthening of the next row begins before the taller row reaches its final height. As the stereocilia lengthen, the tethers between them appear to be stretched from the tip of a shorter one to the side of an adjacent taller one. It is possible that these tethers eventually become the tip links that are thought to regulate tension on the mechanically gated ion channels in mature hair cells. In the cochlea, the kinocilium degenerates as a final differentiation step, revealing that it is not required for transduction and may primarily be needed to establish a polarized bundle. Once development is completed, all hair cells in the organ of Corti assume the same bundle polarity, with the tallest row of stereocilia located at the outer (abneural or lateral) edge of the spiral.

It has been suggested that there are three levels of planar cell polarity in a hair cell-bearing sensory patch. The first is at the level of the individual cell, where the kinocilium becomes asymmetrically localized to the rim of the apical surface. The second is a local level, where small groups of hair cells assume similar orientations, but these are not perfectly aligned across the entire organ. The final level is global, which in the cochlea is manifest by a near-perfect orientation of the bundles with their axis of maximum sensitivity oriented orthogonally to the long axis of the cochlea. In the search for genes that control bundle polarity, it is at the local level where defects have been reported in various mouse mutants. Several of the responsible genes belong to the so-called PCP pathway because their orthologs in flies have been implicated in similar planar orientation defects. In the cochlea, some of the proteins encoded by PCP genes are known to be asymmetrically localized to either the neural or abneural side of sensory cells near their lumenal surface, shown schematically in Figure 5. Thus the molecules are in the right location to influence cytoskeletal rearrangments directionally across the cell surface along the axis of bundle polarity. In flies, the Frizzled receptors, usually affiliated with signaling via wingless (Wnt) ligands, have been implicated in PCP, even though wingless itself has not been shown to play a role. Instead, the directional polarity within the fly epithelium may be transmitted from cell to cell through direct cell contacts. In contrast, Wnts appear to play a more active role in hair cell polarity in the mammalian cochlea. Not only is there evidence for Frizzleds being involved in hair cell polarity in the mouse cochlea, but changes in Wnt concentrations can also alter bundle polarity of hair cells grown in culture.

Spiral Ganglion Neurons

The first cells that can be recognized during inner ear development are the neuroblasts that arise from the Ngn1-expressing floor of the otic cup and otocyst. Neuroblasts fail to develop in the absence of the neurogenic gene, Ngn1. Neuroblasts are still in a proliferative state at the time they delaminate from otic ectoderm. The neuroblasts migrate around the anterior pole and initially gather into a single ganglion just medial to the otocyst. They will soon split into a dorsal vestibular ganglion and a ventral spiral ganglion

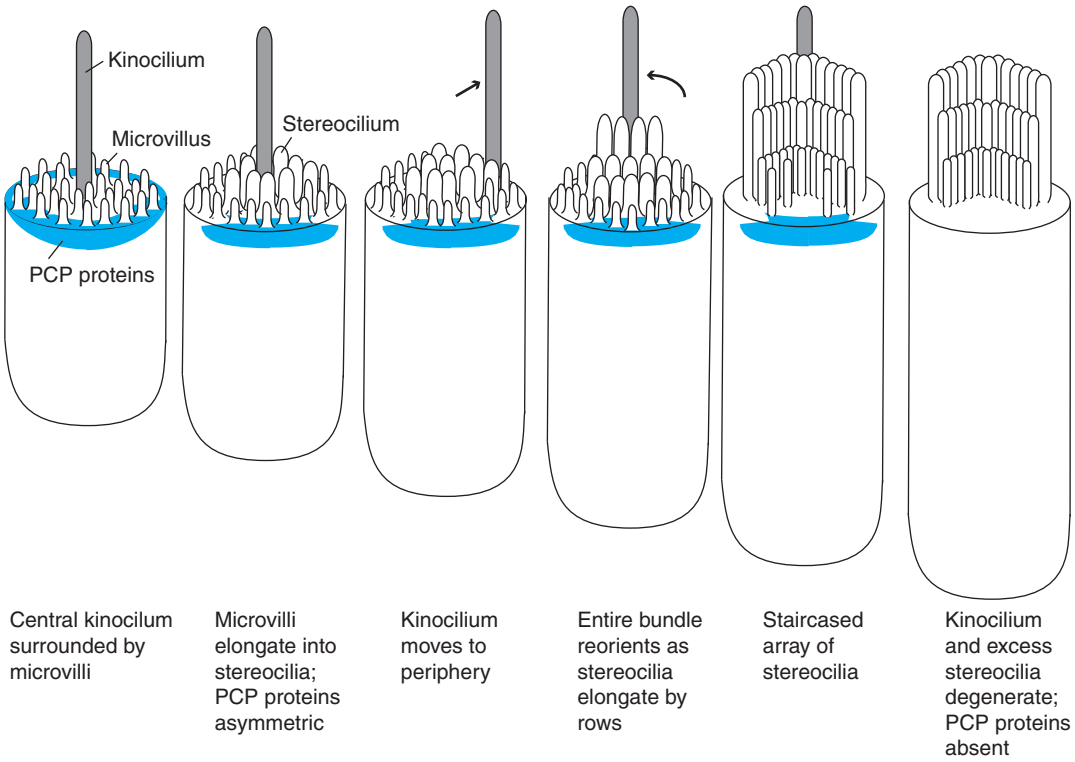

Figure 5 Hair bundle development and PCP proteins. The sequence of events leading to the polarized stereociliary bundle of an outer hair cell is shown as viewed from the neural (medial), or inner hair cell side. PCP proteins (blue) are initially located around the apical circumference of the cell, but then become localized to either the medial rim (as shown) or to the lateral rim, or both. When the kinocilium (gray) moves circumferentially, it tends to localize within 45° of its final location. The directionality of the initial movement and final localization of the kinocilium could be facilitated by the action of asymmetrically distributed PCP proteins interacting with the cytoskeleton. Figure based in part on schematics by Goodyear R, Kros CJ, and Richardson GP (2006) The development of hair cells in the inner ear. In: Eatock RA, Fay RR, and Popper AN (eds.) *Springer Handbook of Auditory Research, Vol. 27: Vertebrate Hair Cells*, Fay RR and Popper AN (series eds.), pp. 20–94. New York: Springer; and Kelley MW (2006) Hair cell development: Commitment through differentiation. *Brain Research* 1091: 172–185.

(Figure 2). It is not yet known whether the two populations of neuroblasts are distinct at the time of delamination. Lineage studies in the chicken otocyst indicate that clonally related cells can disperse across both ganglia, suggesting that common progenitors are used for both neuronal populations. On the other hand, most clones are restricted to one or the other ganglion, so it is possible that most neuroblasts have already acquired auditory or vestibular identity before they leave the otocyst. One likely candidate for providing auditory identity is the transcription factor Gata3. Gata3 is expressed in the neurogenic region of the otocyst and is retained in spiral ganglion neurons but not vestibular neurons. Furthermore, when the gene is deleted, spiral ganglion neurons fail to form.

Neuronal differentiation requires the basic helix–loop–helix transcription factor, NeuroD. NeuroD null mice generate neuroblasts, but the cells never differentiate and soon die.

Spiral ganglion neurons are bipolar: one process projects out to the organ of Corti and the other projects through the auditory nerve into the cochlear nucleus. The neurons require neurotrophic support from the periphery for their survival. During normal development, a large number of neurons undergo programmed cell death, perhaps responding to the need to balance neuron numbers with hair cell targets. Either NT3 or BDNF can support the survival of spiral ganglion neurons, which express the Trk receptors for both neurotrophins. During normal development, NT3 supports axons innervating the majority of the cochlea, while BDNF supports the most apical fibers. The peripheral processes of the spiral ganglion neurons reach the organ of Corti in advance of the efferent fibers originating from the brain stem. Thus, it is not surprising that the efferents appear to use the afferents to find their way from the nerve to the hair cells. Further discoveries may reveal exactly how growing auditory 'dendrites' find their way to the sensory periphery and how the centrally directed axons navigate into the cochlear nucleus of the brain stem.

See also: Auditory System: Efferent Systems to the Auditory Periphery; Cochlear Mechanics; Deafness; Forebrain: Early Development; Hair Cells: Sensory Transduction; Hair Cell Differentiation; Helix–Loop–Helix (bHLH) Proteins: Proneural; Helix–Loop–Helix (bHLH) Proteins: Hes Family; Notch Signal Transduction: Molecular and Cellular Mechanisms; Notch Pathway: Lateral Inhibition; Retinoic Acid Signaling and Neural Patterning; Sonic Hedgehog and Neural Patterning; Sox Gene Expression; Wnt Pathway and Neural Patterning.

Further Reading

Burton Q, Cole LK, Mulheisen M, Chang W, and Wu DK (2004) The role of Pax2 in mouse inner ear development. *Developmental Biology* 272: 161–175.

Goodyear R, Kros CJ, and Richardson GP (2006) The development of hair cells in the inner ear. In: Eatock RA, Fay RR, and Popper AN (eds.) *Springer Handbook of Auditory Research, Vol. 27: Vertebrate Hair Cells*. Fay RR and Popper AN (series eds.), pp. 20–94. New York: Springer.

Holme RH, Bussoli TJ, and Steel KP (2002) Table of gene expression in the developing ear. http://www.ihr.mrc.ac.uk/hereditary/genetable (accessed December 2006).

Kelley MW (2006) Hair cell development: Commitment through differentiation. *Brain Research* 1091: 172–185.

Kelley MW, Wu DK, Popper AN, and Fay RR (eds.) (2005) *Springer Handbook of Auditory Research, Vol. 26: Development of the Inner Ear.* Fay RR and Popper AN (series eds.), New York: Springer.

Lewis J and Davies A (2002) Planar cell polarity in the inner ear: How do hair cells acquire their oriented structure? *Journal of Neurobiology* 53: 190–201.

Morsli H, Choo D, Ryan A, Johnson R, and Wu DK (1998) Development of the mouse inner ear and origin of its sensory organs. *Journal of Neuroscience* 18: 3327–3335.

Rubel EW and Fritzsch B (2002) Auditory system development: Primary auditory neurons and their targets. *Annual Reviews of Neuroscience* 25: 51–101.

Van Camp G and Smith RJH (2006) Hereditary hearing loss homepage. http://webhost.ua.ac.be/hhh/ (accessed December 2006).

Relevant Website

http://www.jax.org/hmr – Hereditary Hearing Impairments in Mice. The Jackson Laboratories.

Cochlear Mechanics

R B Patuzzi, University of Western Australia, Perth, WA, Australia

© 2009 Published by Elsevier Ltd.

Introduction

The performance of the normal mammalian cochlea is quite magnificent. It can detect air pressure fluctuation (sound) as small as 1/2 000 000 000th of atmospheric pressure, which produces vibrations within the cochlea of only atomic dimensions. It can discriminate frequencies separated by only a small fraction of a musical semitone (a 1% frequency change can change vibration tenfold), can cope briefly with a 1 000 000-to-1 range of sound pressure without permanent damage, fits neatly in a bony capsule the size of a pea, provides its own battery, and consumes only about 50 μW, or 1% of the power of a solar-powered calculator. Of course, it also has problems. It distorts so badly (about 3% distortion, comparable to a $5 radio) that some sounds are not heard in the presence of others (called masking or suppression), and when two pure tones are presented together, it produces four or more of its own. Unfortunately, this distortion is no excuse for a cheap stereo: Different frequencies vibrate different areas in the cochlea, and the mechanical separation of frequencies occurs before the distortion takes place. As a result, the ear's distortion is of a special type, and we can discriminate external distortion from our own. This distortion not only colors our perception of music (harmony is unlikely to exist without it), it is used clinically in detecting deafness of mechanical origin. The cochlea also has a tendency to whistle like a poorly fitted hearing aid as a result of internal electromechanical feedback, and in some people these sounds can be heard by others nearby (known as 'objective tinnitus,' different from 'subjective tinnitus,' which is not heard by others). It can also be damaged by excessive sound and various prescription drugs in high doses, including antibiotics, loop diuretics, and even aspirin. It is interesting that it achieves its high performance by combining electrical and mechanical phenomena to cancel internal friction in a way that engineers have only recently used in the surface acoustic wave 'chirp filters' of modern RADAR. Overall, the function of the cochlea is so dependent on the vibration of the organ of Corti within it that a full understanding of normal sound perception and deafness is impossible without an understanding of cochlear mechanics. Here, the processes producing cochlear vibration are summarized, including the frequency-dependent vibration, the active amplification of vibration by hair cells, the nonproportional growth of vibration with sound level, the nonlinear two-tone interactions such as distortion generation and two-tone suppression (masking), and finally how these phenomena are altered to produce sensorineural deafness of cochlear origin.

Cochlear Vibration and the Active Process

In a normal ear, incoming sound vibrates the ear drum and middle ear bones, causing pressure fluctuations in the cochlear fluids and a vibration of the ribbonlike basilar membrane that divides the cochlea roughly in two along its length. The cochlea's coiling is largely irrelevant to its mechanical behavior because the speed of longitudinal sound waves in cochlear fluid is so fast that pressure fluctuations distribute in the cochlea within microseconds, so their wavelengths are far longer than any of the cochlea's structural features (including its coiling). Moreover, the sound vibration detected by the stretch-sensitive hair cells and conveyed to the auditory neurons is not the longitudinal fluid waves but the transverse interface wave along the basilar membrane and organ of Corti (OC). Just as the longitudinal pressure waves within bulk water of a pond travel at about 1000 m s^{-1} while the transverse surface ripples travel at a leisurely few meters per second, the transverse cochlear waves travel at a very slow pace when compared with the longitudinal waves in bulk fluid (with delays of a few milliseconds from base to apex, depending on species). As shown in Figure 1(a), this vibration takes the form of Bekesy's traveling waves or ripples that mostly travel away from the stapes and toward the cochlear apex. For a pure-tone stimulus of a particular frequency, the wave grows as it travels (because the basilar membrane becomes less stiff along its length), before reaching a maximum at its resonant characteristic place, beyond which it collapses abruptly so that no vibration exists beyond its cutoff region. This characteristic place is positioned along the cochlear length according to a species-specific, logarithmic, place–frequency map, with high characteristic frequencies (CFs) toward the cochlear base and lower frequencies closer to the cochlear apex. The mapping is largely determined by the grading in mass and (mainly) stiffness of structures along the cochlear partition (the basilar membrane plus the OC sitting on it). Fortunately, this grading and the cochlear shape have evolved to avoid reflections within. These would

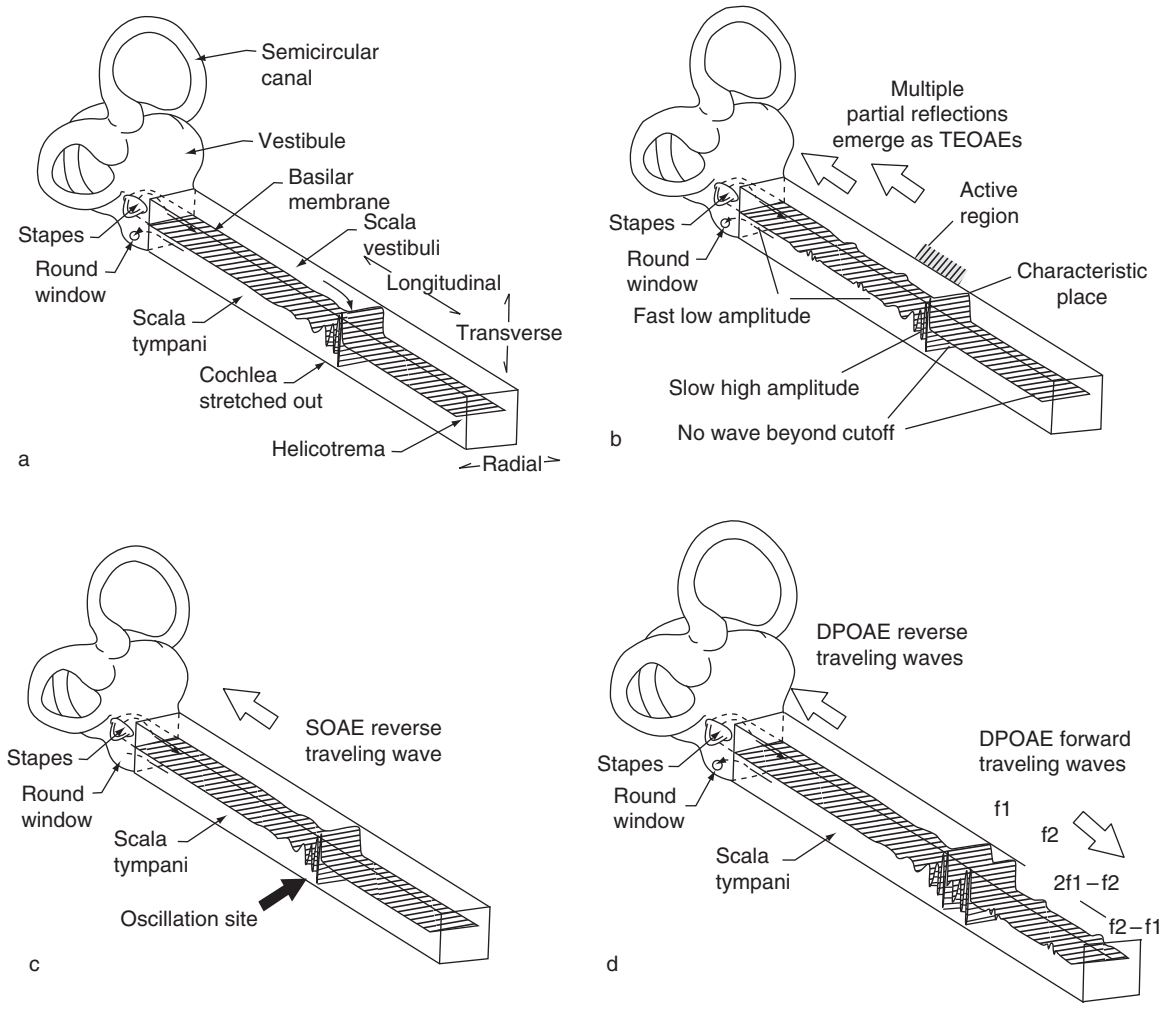

Figure 1 (a) Schematic of the cochlea, uncoiled to show the traveling wave, which normally travels rapidly away from the stapes before slowing and growing as it encounters less stiff regions of the cochlear partition. It is the transverse (up-down) vibration of the partition that stimulates hair cells and neurons. (b) While reflections are minimized, slight reflections occur from multiple sites in any imperfect cochlea, with echoes emerging as transient evoked otoacoustic emissions (TEOAEs). Only outer hair cells (OHCs) in a small active region before each characteristic place assist vibration, which is the reason masking or suppression is most effective for suppressor tones vibrating this region. (c) Some OHCs can become overzealous, oscillating spontaneously, producing spontaneous otoacoustic emissions (SOAEs) that can be detected in the ear canal by quiet sensitive microphones. (d) Two tones with similar frequencies (f1 and f2) interact because of mechanical distortion to produce other traveling waves that propagate to their own characteristic places and also back out through the middle ear to the ear canal, where they can be measured as distortion product otoacoustic emissions (DPOAEs).

be confusing and would limit timing resolution and therefore our directional discrimination, partly derived from timing differences between our two ears. Having said that, a normal (but always imperfect) cochlea does produce slight reflections from small imperfections along its length (Figure 1(b)). Originally known as Kemp echoes, they are now called transient evoked otoacoustic emissions (TEOAEs), are idiosyncratic of any individual, are delayed according to their frequency and travel time from the stapes to the point of reflection and back, and can be measured in the ear canal by a sensitive but quiet microphone. They are evoked by clicks and less often by tone bursts, provide some frequency-specific information, and are used as a rapid screening test for normal cochlear vibration.

Detecting this vibration and communicating it to the brain are the function of the OC, which sits on the basilar membrane and contains two types of vibration-sensitive hair cells: the inner hair cells (IHCs), which lie in a single row and synapse with about 95% of the primary afferent neurons (type I), and the outer hair cells (OHCs), which lie in three or four parallel rows, outnumber the IHCs almost three to one, and synapse with only 5% of the afferent neurons (type II). Experimentally, the molecular vibration has been detected in many ways (see Figure 2, showing the

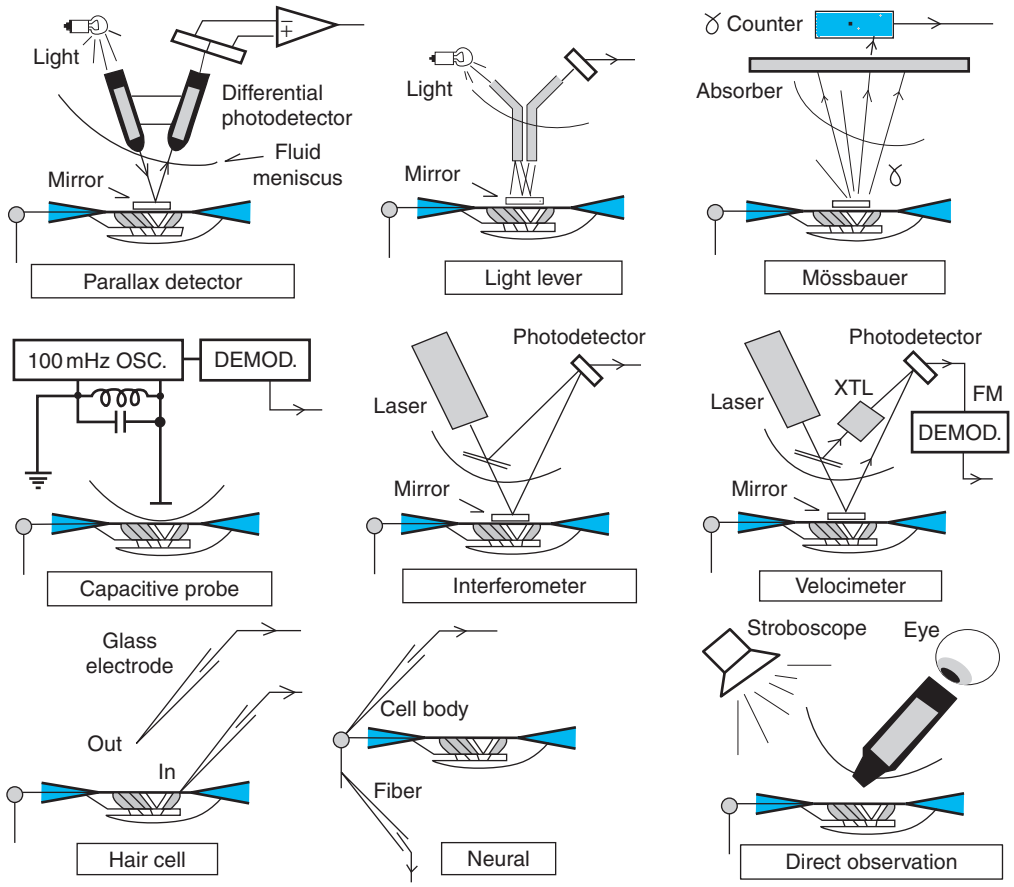

Figure 2 Most of what we know of cochlear vibration has come from experiments in which the vibration has been observed directly in living animals using many techniques. Optical techniques include direct observation, the parallax detector, the light-lever, the interferometer, and the velocimeter. Doppler shift of gamma ray emission has also been used (the Mössbauer effect), as has the capacitive probe. A great deal can be determined by monitoring cochlear vibration indirectly through responses of hair cells and neurons, detected using electrophysiological techniques. As shown here, the organ of Corti hangs from beneath the basilar membrane. This is most often the case experimentally, when the approach to the vibrating structures is often via the fluid-filled scala tympani. DEMOD., demodulator; OSC., oscillator.

OC upside down, as it usually is in experimentation in which vibration is detected from its scala tympani side). The transverse (up-down) vibration of the OC (point 1 of Figure 3) produces a radial shearing between its tectorial membrane and reticular lamina and ultimately a deflection of the hair bundles at the hair cell apex, leading to the opening and closing of the mechanoelectrical transduction (MET) channels on each stereocilium. This produces a fluctuation in the electrical resistance of the hair cell apex (points 2 and 2′ of Figure 3). Because there is a voltage across the apical membrane of the hair cells which drives K^+ ions into the cells from the K^+-rich endolymph of the fluid compartment of scala media, the opening and closing of the MET channels produces a fluctuating hair cell receptor current. The driving voltage (110 mV for IHCs and 170 mV for OHCs) is made up of two components: the positive potential above the hair cells in scala media (the endocochlear potential (EP) of about +95 mV) and an additional component due to each hair cell's membrane potential (about −40 mV in IHCs and −70 mV in OHCs). The EP is maintained by the cochlea's epithelial 'battery charger' (stria vascularis on the cochlear wall of scala media), while the membrane potentials are maintained by each cell's metabolic resources. The receptor currents (points 3 and 3′ of Figure 3) produce receptor potentials within the hair cells, which also depend on the electrical properties of each 'cell's basolateral wall (points 4 and 4′ of Figure 3). While the receptor potentials within IHCs and OHCs cause the release of neurotransmitters (point 6 of Figure 3) and spiking in the afferent neurons (point 7 of Figure 3), the receptor potential of OHCs also changes the shape of a set of 'motor proteins' (prestin) arranged in a systematic network on the inside surface of the OHC basolateral wall. There appears to be so much prestin that its concerted movement changes the OHC length (electromotility), and possibly similar forces arise from small movements of the hair bundles themselves

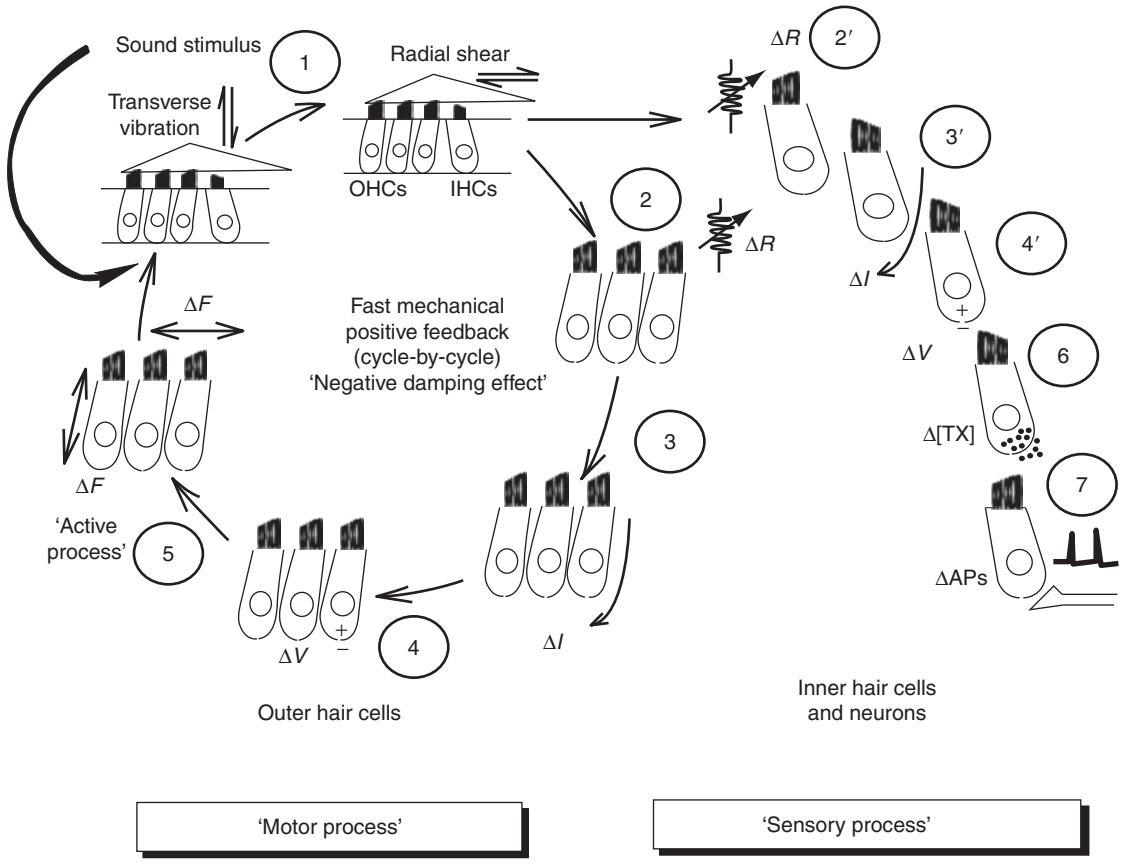

Figure 3 The outer hair cells (OHCs) assist cochlear vibration by a process of electromechanical feedback. Transverse vibration of the organ of Corti produces radial shear (R) of the hair bundles (1), opening and closing mechanoelectrical transduction channels at the hair cell apex (2), thereby modulating hair cell receptor current (I; 3) and producing intracellular receptor potentials (V; 4). In outer hair cells, this leads to feedback forces (F) due to either electromotile prestin in the basolateral wall or possibly forces (F) generated by the hair bundles themselves (5). Overall, these OHC forces act to cancel friction within the organ of Corti, producing nearly a 1000-fold increase in vibration at low sound levels. This enhanced vibration is detected by adjacent inner hair cells (IHCs) by an analogous process. Overall, the processes associated with OHCs can be referred to as 'motor processes' and those associated with IHCs can be referred to as 'sensory processes.' Cochlear hearing loss can therefore be subcategorized as motor losses, sensory losses, or mixed cochlear losses when both are disrupted.

(point 5 of Figure 3). Because the opening of MET channels is synchronous with the sound vibration, the receptor current and receptor potential also increase and decrease through a cycle of a pure tone, so neurotransmitter release, neural firing, and OHC length all fluctuate through a cycle of a tonal stimulus. This cycle-by-cycle fluctuation produces feedback forces crucial to normal vibration. Just as children on a swing move their legs back and forth at the swing's natural frequency to increase vibration by canceling friction in the hinge, the OHCs act to cancel friction and increase vibration of the OC. In a normal cochlea, the vibration is increased about 1000-fold (60 dB) relative to the vibration without OHC assistance. The mechanism enhancing vibration is known as the active process, the partial cancellation of friction is known as negative damping, and the combination of all these steps to produce large vibration is often referred to as the cochlear amplifier. As indicated in Figure 3, the processes associated with IHCs and the 95% of afferent neurons synapsing with them (type I) can usefully be called sensory processes because their ultimate role is to 'sense' whatever vibration occurs, while the processes associated with OHCs can be called motor processes because they ultimately allow OHCs to assist vibration, like in-built hearing aids. This distinction is useful because it is possible to subcategorize various types of cochlear hearing loss in this way.

Nonlinear Growth of Macromechanical and Micromechanical Vibration with Sound Level

The overall transverse vibration of the OC in mammals is often referred to as macromechanical vibration, indicating that the OC and basilar membrane are moving as a whole. The vibration of the component

parts of the OC (basilar membrane, hair cells, tectorial membrane, etc.) relative to each other is known as micromechanical vibration. For some time, these two vibration modes were thought to be separable and simply cascaded ('the first and second filters'), but it has become increasingly clear that they must be considered together because their resonances are tightly coupled. This coupling occurs partly because their respective mass and stiffness parameters are not greatly different and partly because micromechanical motion determines hair bundle deflection, production of the electrical signals from vibration (forward transduction), and ultimately the electromotile forces feeding back to the OC (reverse transduction). This action of electromotility (which occurs in a limited active region just before the characteristic place; Figure 1(b)) is crucial to understanding cochlear vibration because it partly (and sometimes wholly) cancels friction, thereby determining cochlear sensitivity, frequency selectivity, and dynamic range (because the electric signals and therefore feedback forces distort or clip at higher sound levels). If OHCs totally cancel friction in a local region, it oscillates, generating sound that propagates back to exit into the ear canal as a spontaneous otoacoustic emission (SOAE; Figure 1(c)). About 50% of people have one, and in some it is heard as an 'objective tinnitus.' In any case, the progressive clipping of the electrical signals and feedback forces at higher sound levels (saturation feedback) not only allows the cochlea to compress a 1 000 000-to-1 range of sound level (its 120 dB dynamic range) into a mere 100-to-1 range of vibration amplitude; it creates dramatic two-tone interactions: low-frequency intermodulation (in which hearing sensitivity fluctuates dramatically through a cycle of an intense low-frequency tone), two-tone suppression at higher frequencies (in which the cycle-by-cycle modulation is not obvious, but vibration due to one tone is reduced by the presence of a second tone of appropriate frequency and intensity), and distortion product generation (in which the nonlinear interaction between two tones produces other frequency components). If one tone has a frequency f1 and the second tone has a slightly higher frequency f2, then the most prominent distortion frequencies in most mammals are f2−f1 (the simple or quadratic distortion tone, which indicates asymmetric distortion) and 2f1−f2 (the cubic distortion tone, which indicates symmetric distortion). These distortion products (Figure 1(d)) are readily detected in the external ear canal of normal ears as distortion product otoacoustic emissions (DPOAEs), which are easily shown to be produced within the cochlea (because the f1 and f2 frequencies must be close, with $f1f2^{-1} \sim 1.3$, indicating sharp filtering, and because cochlear disruption abolishes them). Because the DPOAEs are disrupted by any pathology that disrupts OHC function (the motor processes), they can also be used for objective hearing screening like the TEOAEs, especially in infants. As for which elements in the cochlea cause the nonlinear phenomena, there are potentially five: nonlinear hair bundle gating stiffness, nonlinear mechanoelectrical transduction at the apex of hair cells, further distortion of receptor potentials by nonlinear conductances at the OHC basolateral wall, nonlinear electro-mechanical transduction from OHC membrane potential to prestin forces, and hair bundle adaptation at the apex of the hair cells. In some cases it is clearly the forward transduction of the MET channels that determines the clipping and distortion, but in others the cause is not so clear.

Why Does the Wave Travel from Base to Apex? It's Cochlear Hydrodynamics, Not Stapes Drive

One important point to appreciate about vibration within the cochlea is that it is pressure fluctuations in the fluid surrounding the OC that drives vibration, not forces between adjacent sections along the cochlear length. That is, the traveling wave is not like a ripple traveling along a shaken rug but is more like a row of boats moored next to each other, coupled only by the water movement beneath them. If we push one boat downward, pressure changes in the water are transmitted between adjacent boats, and a wave of boat motion is observed. The consequences of this point are important.

First, the sound energy can be fed to the cochlear fluids at almost any point around the perimeter of the cochlea, and it will travel almost instantaneously (at almost $1000\,\mathrm{m\,s^{-1}}$) to all other regions of the cochlear fluid, producing the usual transverse traveling wave on the basilar membrane, traveling in the same direction, at the same speed, and with the same properties as usual. This point is important because it means that sounds can be fed to the cochlea from almost any point around its perimeter, and by almost any means, and the resultant traveling wave and nerve stimulation will be near normal. For example, although sound is normally fed to the fluid through vibration of the stapes, it can be fed through vibration at the apex, from spontaneous vibration of the OHCs themselves, or through the surrounding bone, and the sound perception is maintained. In fact, this mechanism of 'bone conduction' is the dominant mode of cochlear stimulation when the ear canals or middle ear are blocked (e.g., with wax, exudate, growths, or deformity) and has recently been used in a special implanted hearing aid that vibrates the bone directly (the bone anchored hearing aid).

Second, the transverse traveling wave that travels from base to apex is not stopped by immobilizing the

partition in the base (removing or cementing one boat along a line of boats does not destroy all wave motion beyond the problem boat, whereas cutting a rug along its length does). Rather the hydrostatic fluid pressure simply passes the problem, and the wave recommences on the other side. The implications of this are that moderate damage and scarring of the cochlear partition in the basal regions produce a drop in vibration only in the base, but vibration in more apical low-frequency regions is relatively unaffected. That is, it is possible to have a high-frequency mechanical loss without a consequent low-frequency one. This is precisely what happens with most cases of acoustic trauma, which occurs most often at high frequencies, and is also true in the case of other frequency-specific lesions of many kinds.

Third, because the transverse vibration is driven by pressure fluctuations in the fluid rather than through the partition itself, if the pressure were short-circuited by a small perforation through the cochlear partition, all vibration in surrounding and more apical regions would be reduced or absent. Fortunately, it is unusual to suffer such a punctate hole, because the cochlea is encased in the hard temporal bone. On the other hand, the fluid pressure can be effectively short-circuited if one region of the cochlear partition is resonant, because it has a very low acoustic impedance. As the transverse wave for a pure tone travels along the cochlear partition from base to apex, it will encounter a region where the partition is at or near resonance at that stimulus frequency, and the fluid pressure will be short-circuited (or at least greatly reduced), effectively 'stealing' the fluid pressure required to drive the next section of partition toward the apex. The net result is that a traveling wave travels out from the basal end, initially traveling rapidly and at low amplitude (like a tsunami in the deep ocean), but when it reaches its resonance region, the wave slows dramatically and increases in amplitude (because resonance means low acoustic impedance to vibration) but then suffers a 'cutoff' phenomenon, abruptly falling to a very low amplitude and ultimately extinguishing entirely. This high-frequency cutoff is not evident in ripples on water because it requires that the surface be resonant, requiring the combination of a restoring stiffness and an inertial mass. Although simple water surfaces possess a restoring force (gravity mostly, but partly surface tension), there is very little mass at the surface itself (rather than bulk fluid), and a cutoff phenomenon does not occur, so we are not familiar with it. A cutoff would be observed, however, on a water surface loaded with mass (e.g., boats or water lilies). In any case, the combination of mass, stiffness, and surrounding fluid pressure coupling means that no vibration can travel (forward or backward) along a surface if the stimulus frequency is above the resonance frequency of the surface at that point. Such a wave is said to be 'evanescent.' This has the practical significance that any sound produced by the cochlear partition itself cannot travel outward as a transverse traveling wave if its frequency is above the resonance frequency of its generation site. For example, distortion tones with frequencies higher than that of their generation site (e.g., f2 + f1, 2f1 + f2) are 'stillborn' and unable to propagate. (This does not mean that distortion products higher in frequency than the resonance frequency of their generation site cannot influence distant sites, but this must occur via a longitudinal sound wave and not as a transverse traveling wave.) Conversely, lower-frequency distortion products (f2−f1 and 2f1−f2) can propagate in the forward direction to their appropriate site, to be perceived as Tartini's phantom bass, and backward to the stapes, middle ear, and ear canal, to emerge as DPOAEs.

Sensorineural Hearing Loss and the Frequency Threshold Curve

Hearing loss in humans is normally categorized as conductive, sensorineural, or mixed (conductive plus sensorineural), and the sensorineural component can itself be broken down into its cochlear and retrocochlear components. It is now possible to categorize cochlear lesions into two basic types, as mentioned earlier: motor losses, in which motor processes are disrupted (left half of Figure 3), and a sensory loss, in which sensory processes are disrupted (right half of Figure 3). In some pathologies, both motor and sensory processes are impaired, and we can define a mixed cochlear loss. As an aid to memory, the many types of cochlear hearing loss described below can also be categorized simply as one of the four Bs of cochlear deafness: (1) a 'block' of MET channels or the primary afferent synapse by drugs or overstimulation, (2) a mechanical 'bias' of the hair bundles by osmotic or hydrostatic pressures, (3) a drop in the cochlear 'battery' driving the receptor currents through the outer and inner hair cells, and (4) activation of the crossed olivo-cochlear 'bundle' (drugs such as salicylate may mimic its action). To discuss these cochlear pathologies in detail, we must consider the upside-down 'audiogram' of a single afferent neuron, called its frequency tuning curve (FTC). As shown in Figure 4, the normal FTC of each cochlear afferent neuron has a very sensitive 'tip region' centered around the neuron's CF (where vibration is most sensitive), and a less sensitive 'tail region' at lower frequencies (where the low-amplitude vibration passes the neuron on its way to its own more apical characteristic place). Above the cutoff frequency of a neuron, there is no response at all because

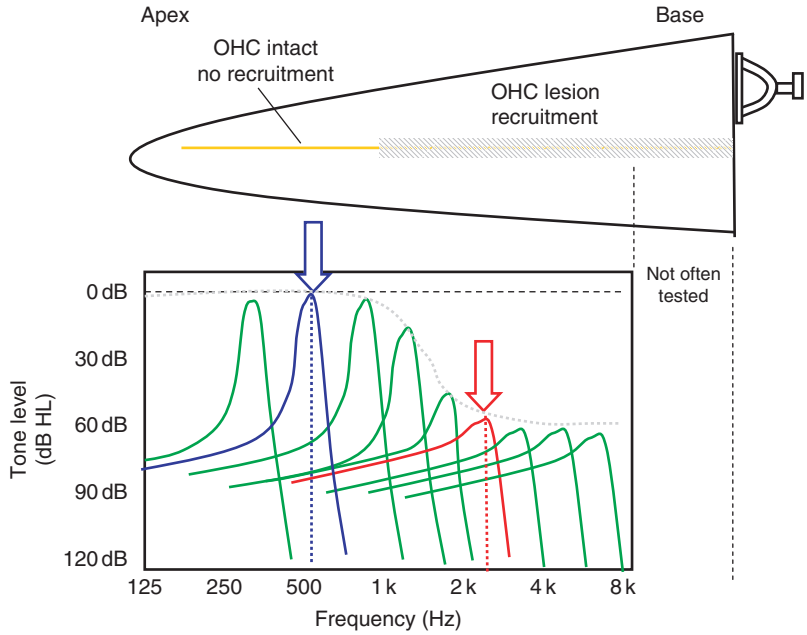

Figure 4 Schematic showing neural frequency tuning curves (FTCs) for neurons spread across the cochlea, and in this case with damaged hearing at higher frequencies. The normal localized vibration of the cochlear partition with pure tone stimulation results in each primary afferent neuron possessing a sharp frequency threshold curve (FTC), which plots the sound pressure required to produce neural firing just above spontaneous rate (vertical scale) versus the frequency of the pure tone stimulus (horizontal scale). Each FTC for each neuron possesses a sharp tip region and a wider low-frequency tail, with no response above the cutoff frequency (because higher-frequency traveling waves never reach that neuron). Taken together, all contributing neurons result in an overall sensitivity contour known as the 'audiogram,' shown here as a dashed curve skimming the tops of all FTCs. Here an example is shown of the change in FTC shape as we move from a normal low-frequency region of the cochlea (the apical regions) into a region with outer hair cell (OHC) damage (e.g., damaged by acoustic trauma). The FTCs in the normal cochlear apex are sharp with a narrow and sensitive tip region (red), but the FTCs in the damaged cochlear base are less sensitive and blunt (blue). As a result, in the normal apical region, sound must be increased substantially above threshold (red arrow) to stimulate adjacent higher-frequency neurons, but in the base only small increases in sound level above threshold (blue arrow) are required to produce an upward spread of recruitment of neural firing. This influence of broader FTCs in regions with damaged OHCs is exacerbated by the loss of nonlinear compression in these regions. The icon of the cochlea above the audiogram and FTCs highlights the normal apex and abnormal base of the example, and that the normal clinical audiogram (125 Hz–8 kHz) does not test the extreme basal region of the human cochlea.

the traveling wave never propagated that far. It is important to note that the mechanical sensitivity around the CF and the FTC tip region relies heavily on the action of the active process of OHCs to cancel friction alone (see Figure 3), and any damage to OHCs alone reduces sensitivity near the tip region without changing sensitivity on the stiffness-dominated low-frequency tail. This is important in understanding the changes in perception with OHC damage, as discussed next.

Motor Losses: OHC Problems with No IHC/Neural Disruption (Noise Trauma, Hydrops, and Fistula)

As illustrated in Figure 4, motor losses that are due to disruption of the active process only are associated with elevation of the single fiber FTC for frequencies near CF, with little change on the low-frequency tail. Because the active assistance of OHCs enhances cochlear vibration about 1000-fold, the motor loss associated with OHC disruption alone is usually limited to this amount (60 dB). A motor loss can be due to blockage of the MET channels of OHCs by moderate noise trauma, ototoxic drugs, or mechanical biasing of the OC by low-frequency sound or hydrostatic/osmotic imbalances, as in perilymph fistulas and endolymphatic hydrops (which in some cases probably produces Ménière's syndrome). Motor losses can also be caused by disruption of OHC motility directly, due to some drugs (e.g., salicylate) or excessive efferent neuron activity. The changes in OC vibration reduce sensitivity, the sharpness of FTC tuning, nonlinear compression, two-tone suppression, and distortion tone generation. These changes dramatically change the perception of speech and music and cause the phenomenon of (hyper) recruitment, which is the abnormally rapid increase in perceived sound loudness as sound level is increased. (Recruitment is best called hyperrecruitment because there is some neural recruitment in any normal ear.) Recruitment clearly has its origin in

cochlear mechanics, and certainly has always been associated with OHC lesions. It has two major aspects. First, although the active assistance of OHCs enhances cochlear vibration by about 60 dB for low sound levels, as sound level increases and the nonlinear processes associated with OHCs clip or saturate, the OHC assistance becomes less important (the so-called saturation feedback process that provides the ear's 120-dB dynamic range). In other words, OHC assistance is important for low-level sounds but largely irrelevant for very high sound levels. As a result of this progressive saturation of the OHC assistance, normal vibration in a healthy ear grows very slowly with sound level ($\sim$0.2 $dBdB^{-1}$). However, once OHC assistance is abolished, there is not only a 1000-fold elevation of hearing threshold (a 60 dB sensorineural hearing loss), but compressive growth of vibration is abolished, and things revert to a simple proportional growth, or linearity results (1 $dBdB^{-1}$). With no compression, the increase in neural firing rate with increase in sound level is perceived as abnormally rapid by a central nervous system, previously programmed to expect a much slower growth. The second factor contributing to recruitment is also mechanical in origin. In a normal cochlea, the resonance of cochlear vibration is so good that the tip of the FTCs of single auditory neurons is extraordinarily sharp. They have low-frequency slopes of a few hundred decibels per octave and high-frequency slopes of almost 1000 dB per octave, with a very narrow bandwidth. As a result, at auditory threshold very few neurons are stimulated, and as sound level is increased, there is little recruitment of adjacent neurons until 40–50 dB above threshold (because all adjacent FTCs are so sharp). Above this sound level, however, neurons innervating more basal (high- frequency) neurons are stimulated, but neurons in more apical regions are not. The rate of this upward spread of recruitment in a normal cochlea with nonlinear compression is very slow. In a cochlea with no OHC assistance to vibration (a motor loss), not only is the nonlinear compression absent but the FTC is much broader, without a well-defined tip region (see **Figure 4**). As a result, although only a few neurons may be stimulated near auditory threshold, just a few decibels above threshold there is already a rapid upward recruitment of neural firing. This abnormally rapid upward spread of recruitment adds to the abnormally rapid increase in firing rate of individual neurons because of the lack of compression, and the increase in sound loudness is perceived as far too rapid and troublesome. Although the absence of nonlinear compression in cochlear vibration can be replaced to some extent by similar nonlinear compression in more-sophisticated hearing aids, the sharp tuning of the individual FTCs cannot be restored. The more-sophisticated hearing aids attempt (with only partial success) to overcome this problem using multiband filtering, with nonlinear compression added independently to each frequency band.

Sensory Losses: IHC/Neural Problems with Normal OHCS

Sensory losses are due to disruptions of IHCs or synaptic function (e.g., Mg^{2+} perfusion into the cochlea) and are associated with elevation of the single fiber FTC across the whole frequency range. Without a mechanical involvement, tuning, nonlinear compression, distortion tone generation, and two-tone interactions are not greatly affected, and there appear to be little or no changes in the OAEs in the ear canal. Such sensory losses do not appear to be common, although one recently recognized form (in perhaps 10% of congenitally deaf infants) is auditory neuropathy, in which OHC receptor currents and OAEs seem normal but cochlear nerve and auditory brain stem responses are abnormal or absent. Some of these cases are likely to benefit from cochlear implantation.

Mixed Cochlear Losses: Drug Blockade, Cochlear Hypoxia, and Morphological Disruption

One of the simplest cochlear manipulations that impairs the function of IHCs and OHCs simultaneously is the introduction of drugs that block MET in both inner and outer hair cells. Such drugs include aminoglycoside antibiotics (e.g., kanamycin, streptomycin, gentamycin), anticancer drugs (e.g., *cis*-platinum), and some antimalarial drugs (e.g., quinine). Another way to disrupt motor and sensory processes and produce a mixed cochlear loss is to reduce the driving voltage for the hair cell receptor currents, which is most easily accomplished by dropping the endocochlear potential above the hair cells by disrupting the cochlea's 'battery charger' (stria vascularis). This can be done with hypoxia (disruption of strial blood supply from stroke or local embolism) or with pharmacological blockade of strial ion transport by pharmacological blockers such as ouabain or the diuretics furosemide, bumetanide, or ethachrynic acid. Changes in the cochlea that are associated with morphological changes to the OC, and to the hair cells in particular, also alter OC vibration. Such changes can occur with intense or prolonged acoustic overstimulation, and high doses for long periods of some ototoxic drugs (e.g., aminoglycosides, anticancer drugs). Such pathology with morphological changes has the added complication that the passive mechanical properties of the

OC (the stiffness, mass, and geometry of its constituent parts) are also altered.

Concluding Remarks

While the description provided here is true of most mammals, the role of cochlear vibration is different in different classes of vertebrates. For example, in lower vertebrates (e.g., Amphibia, turtles), cochlear vibration is relatively simple and not particularly resonant, and neural firing is determined largely by electrical resonance of the basolateral wall of the main sensory hair cells. This is also partly true in birds, but in many reptiles, micromechanical vibration becomes more resonant, more independent of macromechanical vibration, and more important in determining sensitivity and frequency tuning of the neural response. It seems likely that electrical tuning is useful up to a relatively low frequency of a few kilohertz, above which basic constraints of cell surface area, ion channel density, and animal temperature limit its usefulness. Beyond that, micromechanical resonances become useful in determining sensitivity and frequency selectivity, but only if internal friction can be overcome. In some animals with hair cells without electromotile prestin, the negative damping is performed by the hair bundles and MET channels themselves (as appears true in insect sensilla), but in mammals, the evolution of prestin and its electromotility appears to have at least assisted hair bundle motility, if not replaced it. Finally, the extraordinary sensitivity of mammalian hair cells to molecular displacements also requires the cochlea to be protected against excessive displacements. There are many lines of this defense, including the Eustachian tube, the slip joint of the human malleoincudal joint, the hard temporal bone, the pressure bleed of the helicotrema, and the viscous coupling of IHCs to the overlying tectorial membrane. While the hair bundles of OHCs are embedded into the tectorial membrane and are therefore displacement coupled, the OHCs possess a variety of homeostatic mechanisms to dynamically maintain the angle of their hair bundles in the face of constant buffeting, including a slow Ca^{2+}-controlled contractile process, similar to that in smooth muscle.

See also: Auditory Scene Analysis; Cochlear Development; Cochlear Prosthesis; Deafness; Hair Cells: Sensory Transduction; Sensory Aging: Hearing; Tinnitus.

Further Reading

Manley GA (2000) Cochlear mechanisms from a phylogenetic viewpoint. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11736–11743.

Patuzzi RB (1996) Cochlear micromechanics and macromechanics. In: Dallos P, Popper AN, Fay RR, and Popper A (eds.) *The Cochlea*, pp. 186–257. New York: Springer.

Patuzzi R and Robertson D (1988) Tuning in the mammalian cochlea. *Physiological Reviews* 68: 1009–1082.

Robles L and Ruggero M (2001) Mechanics of the mammalian cochlea. *Physiological Reviews* 81: 1305–1352.

Cochlear Prosthesis

G E Loeb, University of Southern California, Los Angeles, CA, USA
B S Wilson, Research Triangle Institute, Research Triangle Park, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The cochlear implant is the most successful and widely applied neural prosthesis developed to date. According to the National Institutes of Health (NIH) Consensus Statement on Cochlear Implants (1995), "A majority of those individuals with the latest speech processors for their implants will score above 80 percent correct on high-context sentences, even without visual cues." There are now more than 110 000 recipients of cochlear implants, far more than for any other type of neural prosthesis.

Cochlear implants provide an information-carrying interface between electronic systems and neuronal systems. They provide sensations of sound in patients suffering from profound or severe deafness as a result of trauma (either from head trauma or overexposure to intense sounds), administration of ototoxic drugs (e.g., kanamycin, streptomycin, and cisplatin) at damaging levels, or various infectious and hereditary diseases. Each of these insults causes loss of the sensory hair cells in the cochlea, as illustrated schematically in Figure 1 (not to scale). These cells normally transduce the mechanical vibrations of sound waves (entering the fluid-filled cochlear chambers via the middle ear ossicular chain) into neuroelectric impulses to the brain. The axons of the spiral ganglion cells that make up the auditory nerve convey these impulses from the organ of Corti on the basilar membrane (where the hair cells are located) to the cochlear nucleus in the brain stem. In the absence of the hair cells, the remaining spiral ganglion cells cannot respond to sound no matter how loud it is amplified. A cochlear implant activates these cells directly by electrical currents from stimulating electrodes implanted surgically into the scala tympani, one of the cochlear chambers. The auditory nerve activity so elicited is interpreted as sound by the central nervous system, with loudness and pitch generally dependent on the amplitude and location, respectively, of the stimulating current.

During the past 30+ years of active research in this field, over 25 different designs of varying complexity have been developed, three of which are now in widespread use. Each of these current designs share the general components shown in Figure 2, including external parts similar to a hearing aid (battery-powered, wearable control box with a microphone), plus a radio telemetry link that conveys information and power to implanted electronics. The implanted electronic circuitry generates the stimulation pulses and conveys them to the cochlea via the array of electrodes. The designs differ in the number and arrangement of electrode contacts in the cochlea and in the manner in which the acoustic signal is processed to derive the electrical stimulation patterns. In recent revisions of this basic arrangement, the body-worn external processor has been replaced by a behind-the-ear (BTE) processor, similar or identical in shape and appearance to a BTE processor for a hearing aid, as illustrated in Figure 2. Fully implantable systems – with the microphone and speech processor moved to the inside – are under development by each of the three major implant manufacturers and by a research group at the University of Michigan. One or more of these systems may be available within the next several years.

Electrode Arrays

All present-day cochlear prostheses attempt to make use of the tonotopic organization of the normal cochlea, in which high-frequency sounds are transduced at the basal end of the mechanically tuned basilar membrane, and lower frequencies are represented progressively toward the apex of the cochlear spiral. A multielectrode array is implanted in the scala tympani so as to permit local stimulation of subsets of the spiral ganglion cells. The selectivity of this stimulation depends on the condition of the spiral ganglion cells (which may themselves be damaged), the physical proximity of the electrodes to the cells, and the temporospatial distribution of stimulation pulses (see below). The proximity depends on the shape, size, and other design features of the electrode array. Electrode arrays have been built largely by hand from corrosion-resistant metals, such as platinum and iridium, and with flexible carriers, such as silicone elastomer. Their designs have been somewhat constrained in the past by fabrication complexity and desirable surgical handling properties. In general, the spatial specificity of stimulation with implants is much lower than that demonstrated in neural tuning curves for normal hearing, with acoustic stimuli. Such a broad activation of the nerve with electrical stimuli most likely limits the simultaneous number of perceptually separable channels to four to eight, even if more than eight electrodes are used. Substantial improvements in implant performance might be

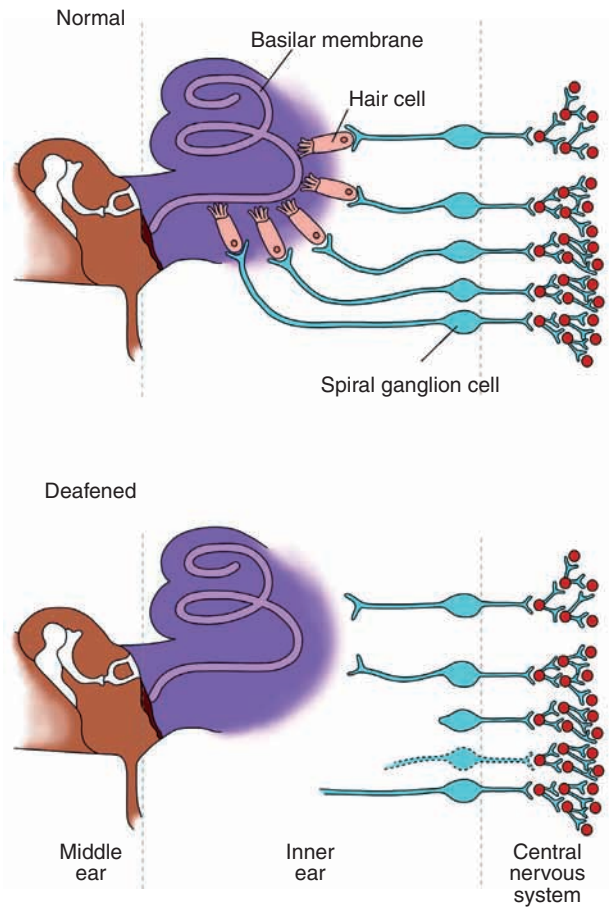

Figure 1 Illustrations of anatomical structures in the normal and deafened ears. Note the absence of sensory hair cells in the deafened ear. Also note the incomplete survival of spiral ganglion cells and changes in neural processes peripheral to the cells that are still viable. The illustrations do not reflect details of the structures and are not to scale. From Dorman MF and Wilson BS (2004) The design and function of cochlear implants. *American Scientist* 92: 436–445.

realized with an increase in the number of perceptual channels, for example, with new types or placements of electrodes. It is also possible to create 'virtual channels' by distributing the stimulation current between two adjacent, simultaneously active electrode contacts. While this does not permit more perceptual channels to be presented at any one time, it may improve the tonal representation of sound and, hence, the quality of musical perception.

Speech Processors

Most multichannel prostheses use stimulation schemes based on the principle of a frequency-channel vocoder (voice coder) synthesizer. Intelligible speech can be synthetically reproduced acoustically by four or more single-frequency sound sources, each temporally modulated in proportion to the instantaneous power present in the band of frequencies immediately surrounding the center frequency of the source oscillator. Cochlear prostheses filter the incoming acoustic information similarly into bands corresponding to the place-pitch sensations evoked by each available stimulating electrode site. They use the amplitude of each band to modulate the electrical stimulation of the corresponding electrode, compressing the range of modulation greatly to deal with the very limited dynamic range of electrical stimulation from perceptual threshold to uncomfortable loudness (typically 6–20 dB, compared with 100–120 dB for normal hearing).

A block diagram of the most common stimulation strategy, continuous interleaved sampling (CIS), is presented in Figure 3. This strategy filters a microphone input into bands and represents relatively slow envelope variations in those bands as modulated trains of biphasic pulses at the electrodes. The output of each band-pass channel is directed to a single electrode, with low-to-high channels assigned to apical-to-basal electrodes, thereby recreating the tonotopic order of frequency mapping in the normal cochlea. The pulses are delivered in a nonoverlapping sequence across the electrodes in order to eliminate undesirable interactions that otherwise would be produced by summation of electric fields from simultaneously stimulated electrodes.

Parameters in the CIS and other strategies include number of channels and associated electrodes, cutoff frequency of the envelope detectors for each channel, and the rate of pulse presentations at each electrode. Results from recent studies have indicated improvements in speech reception performance with increases in the number of channels up to four to eight, and with increases in rate up to 1500 pulses/second/electrode and beyond, at least for some patients. Results for some patients show quite large improvements with increases to very high rates, for example, 5000 pulses/second/electrode. Rates at and above about 2500 pulses/second are likely to elicit a stochastic pattern of responses within and among auditory neurons, rather than the artefactual synchronization of the neurons at lower stimulus pulse rates. This may underlie the improvements observed with very high rates. We note that benefits of high rates come at a cost, however, because the high-rate stimulation consumes more electrical power and requires much higher data transmission rates between the external speech processor and the implanted stimulator.

Clinical Results

The vast majority of cochlear implant patients have experienced no lasting side effects from their

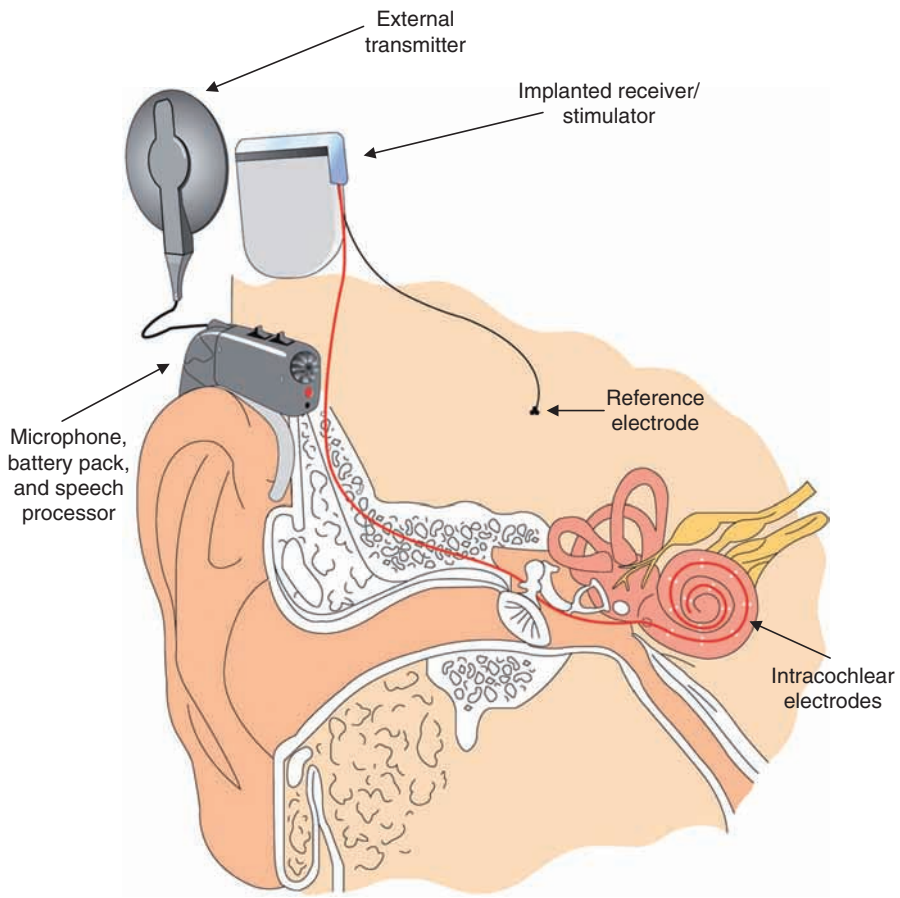

Figure 2 Components of the Med-El TEMPO+ cochlear prosthesis. Other systems share the same basic components. However, the electrode designs and details of processing and transmission of stimulus information across the skin differ widely among systems. In the TEMPO+ system, the microphone, battery pack, and speech processor are incorporated in a behind-the-ear (BTE) housing, much like the BTEs of hearing aids. A thin cable connects the output of the speech processor (a radiofrequency signal with encoded stimulus information) to an external transmitting coil that is positioned opposite to an implanted receiver/stimulator. The transmitting coil is held in place with a pair of magnets, one in the center of the coil and the other within the housing of the receiver/stimulator. The receiver/stimulator is implanted in a flattened or recessed portion of the skull, behind and slightly above the pinna. The reference or 'ground' electrode is implanted at a location remote from the cochlea, usually in the temporalis muscle. For some implant systems, a metallic band around the outside of the receiver/stimulator package serves as the reference electrode. An array of active electrodes is inserted into the scala tympani of the cochlea through the round window membrane or through a larger drilled opening (cochleostomy) near the round window. In present practice, a cochleostomy is used far more frequently than an insertion through the round window. Drawing courtesy of the manufacturer, Med El GmbH, of Innsbruck, Austria.

surgery or subsequent stimulation, and they have been enthusiastic and regular users of their devices for everyday communications. They differ greatly, however, in their performance on objective tests of speech perception. Some patients can understand speech without visual cues (e.g., over the telephone) almost immediately upon first use of the device, but many others have difficulties, particularly in noisy environments. A minority of patients learn only slowly to cope with distorted sounds that are far from normal hearing. Identification of the factor or factors underlying this variability in outcomes is a focus of current research.

Cochlear implants generally provide only modest benefit for adults who lost their hearing prior to acquiring language, probably because their auditory nervous systems have missed a critical period for development. However, congenitally or perinatally deafened children appear to derive substantial benefits if implanted early (usually before 3 and increasingly before 1 or 2 years of age). They usually acquire sufficient speech perception and production skills to permit their education in 'mainstream' classrooms rather than in special schools for the deaf that teach sign language. The use of cochlear prostheses and educational mainstreaming in young, deaf children originally stirred some controversy in the deaf community, some of whom view deafness as an alternative culture rather than a disability. There are also some patients who are unsuitable for cochlear

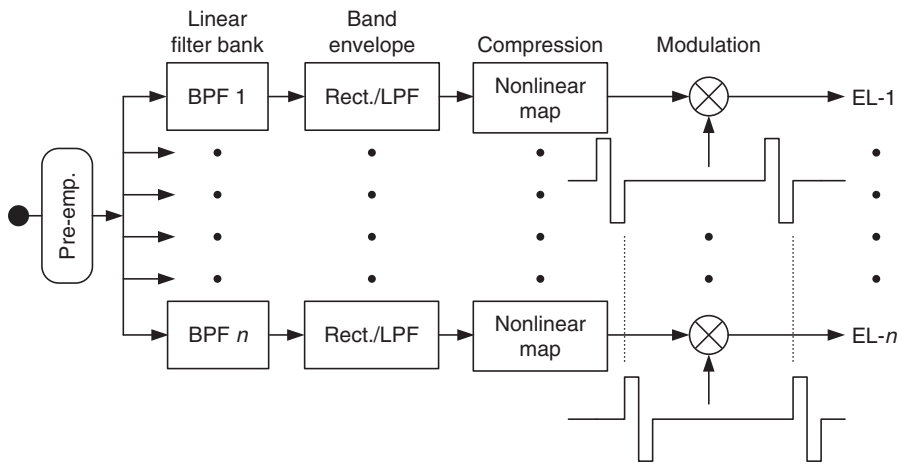

Figure 3 Signal processing steps for a CIS strategy for converting an acoustic input into a pattern of stimulation at the cochlear electrode array. The acoustic signal is amplified with pre-emphasis of the most important frequencies, then filtered into bands (BPF) whose center frequencies correspond to the tonotopic place of the electrode contacts (EL-1 to EL-*n*). The energy in each band is extracted by rectification and low-pass filtering (LPF) and its dynamic range is compressed by mapping onto the highly limited range of electrical pulse amplitudes from threshold for hearing to maximal comfortable loudness. This continuously modulated information about the loudness required in each band is converted into biphasic pulses with variable current and fixed duration that are delivered in a sequential, nonoverlapping, and fixed order to the various electrodes in the array.

implants because of mechanical deformation or damage to the cochlea itself. Research is underway to bypass the cochlea and stimulate more central auditory structures directly. Notwithstanding those few exceptions, cochlear implants have emerged as the standard of rehabilitative treatment for most cases of sensorineural deafness.

See also: Auditory System: Central Pathways; Cochlear Mechanics; Cochlear Development; Deafness; Sensory Aging: Hearing; Speech Perception: Cortical Processing.

Further Reading

Dorman MF and Wilson BS (2004) The design and function of cochlear implants. *American Scientist* 92: 436–445.

Fishman KE, Shannon RV, and Slattery WH (1997) Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. *Journal of Speech Language and Hearing Research* 40: 1201–1215.

Lee DS, Lee JS, Oh SH, et al. (2001) Cross-modal plasticity and cochlear implants. *Nature* 409: 149–150.

Loeb GE (1985) The functional replacement of the ear. *Scientific American* 252: 104–111.

Loeb GE (1990) Cochlear prosthetics. *Annual Review of Neuroscience* 13: 357–371.

Loeb GE (2005) Are cochlear implant patients suffering from perceptual dissonance? *Ear and Hearing* 26: 435–450.

NIH Consensus, Statement (1995) Cochlear Implants in Adults and Children. Bethesda, MD: Office of Medical Applications of Research, National Institutes of Health. (Also available in *JAMA* 274: 1955–1961.)

Rubinstein JT, Wilson BS, Finley CC, and Abbas PJ (1999) Pseudospontaneous activity: Stochastic independence of auditory nerve fibers with electrical stimulation. *Hearing Research* 127: 108–118.

Sharma A, Dorman MF, and Spahr AJ (2002) Rapid development of cortical auditory evoked potentials after early cochlear implantation. *Neuroreport* 13: 1365–1368.

Spelman FA (2006) Cochlear electrode arrays: Past, present and future. *Audiology and Neurotology* 11: 77–85.

Svirsky MA, Silveira A, Suarez H, Neuburger H, Lai TT, and Simmons PM (2001) Auditory learning and adaptation after cochlear implantation: A preliminary study of discrimination and labeling of vowel sounds by cochlear implant users. *Acta Oto-Laryngologica* 121: 262–265.

Wilson BS (2006) Speech processing strategies. In: Cooper HR and Craddock LC (eds.) *Cochlear Implants: A Practical Guide*, 2nd edn., pp. 21–69. London: Whurr.

Wilson BS, Finley CC, Lawson DT, Wolford RD, Eddington DK, and Rabinowitz WM (1991) Better speech recognition with cochlear implants. *Nature* 352: 236–238.

Wilson BS, Schatzer R, Lopez-Poveda EA, Sun X, Lawson DT, and Wolford RD (2005) Two new directions in speech processor design for cochlear implants. *Ear and Hearing* 26: 73S–81S.

Cognition in Aging and Age-Related Disease

E A Kensinger, Boston College, Chestnut Hill, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

As the average age of the population increases, there is growing interest in understanding the cognitive and neural changes that accompany aging. It is now clear that significant cognitive decline is not an inevitable consequence of advancing age. This realization has spurred researchers to examine what separates high-performing older adults from lower-performing older adults and to investigate how the changes with successful aging differ from those that result from age-related disease. Data acquired using behavioral testing, functional neuroimaging, and structural neuroimaging are beginning to inform these issues, although a number of questions remain.

Cognitive Declines with Healthy Aging

Not all cognitive domains are affected equally by age, and not all cognitive processes show age-related decline. If an older adult were asked to list the cognitive declines that have been most notable to him or her, it is likely that at least one of the following would make the list: problems paying attention to relevant information and ignoring irrelevant information in his or her environment, word-finding difficulties, or problems remembering the context in which information was learned.

Much of the research within the field of cognitive aging is focused on understanding whether the pattern of age-related cognitive decline can be explained by domain-general (or 'core') cognitive changes or whether changes in domain-specific processes are required to describe the data.

Domain-General Theories of Cognitive Aging

Domain-general theories of aging are based on the hypothesis that there is a shared ability that cross-cuts all of the tasks on which older adults are impaired: these theories suggest that although aging affects a range of cognitive functions, there is one central or core deficit underlying the myriad changes. This article focuses on three core deficits that have been proposed to explain the pattern of age-related declines: changes in sensory perception, changes in inhibitory ability, and changes in speed of processing.

Sensory deficits The sensory deficit hypothesis of aging proposes that the cognitive changes with aging may be attributed to changes in sensation (i.e., deficits in vision and hearing). Indeed, aging is associated with dramatic sensory declines: by the eighth decade of life, the majority of older adults have significant hearing loss and a reduced ability to discriminate colors and luminance. Support for the hypothesis that these sensory deficits may underlie cognitive changes has come from two lines of research. First, across a wide range of cognitive tasks, older adults' performance correlates strongly with their sensory abilities. Second, in young adults, cognitive impairments can arise when the to-be-processed stimuli are degraded. For example, when asked to remember words pronounced against a noisy background, or when asked to match digits and symbols written in low-contrast font, young adults perform comparably to older adults. Thus, it is plausible that age-related deficits on many cognitive tasks stem, at least in part, from reductions in sensory processing. For example, changes in audition may result in older adults' slowed performance on tasks requiring auditory processing and could explain older adults' poorer memory for auditory information. It is also possible, however, that in at least some instances the correlation derives from a common influence underlying both the sensory and cognitive changes. For example, individuals who have greater brain atrophy or dysfunction may be more likely to have both sensory deficits and cognitive impairments. Regardless of the precise mechanism through which sensory deficits relate to cognitive ones, evidence of significant sensory changes with age underscores the importance of modifying testing procedures to minimize the influence of sensory deficits on older adults' cognitive test performance (e.g., using louder or higher-contrast stimuli).

Inhibition Hasher, Zacks, and colleagues have proposed that older adults' cognitive deficits may relate to their inability to ignore irrelevant information in the environment while focusing attention on goal-relevant information. If an older adult is seated at a restaurant that has many tables in close proximity to one another, he or she may have difficulty paying attention to the conversation at his or her table while ignoring the conversations at nearby tables. In the laboratory, inhibitory deficits can result in responses to previous (but not current) targets. For example, after reading the sentence "Before going to bed, please turn off the STOVE," older adults will be more likely than young adults to believe that the

target word was LIGHT rather than STOVE. The age-related increase in this type of error is thought to occur because older adults have a hard time inhibiting the strong association present in the 'garden path sentence' (e.g., between the idea of going to bed and the idea of turning off the light). Inhibitory deficits also frequently emerge when older adults are required to task-switch or to set-shift. On these tasks, older adults must first pay attention to one aspect of a stimulus (e.g., match items based on their shape) and then another (e.g., match items based on their color). Older adults often have a harder time than young adults when they must ignore the previously relevant dimension (e.g., the shape of the items) and instead focus their attention on another stimulus attribute (e.g., the color of the items).

These data clearly indicate that inhibitory deficits can occur on a range of tasks requiring the ability to selectively attend to information in the environment or to inhibit a strong association or response. However, inhibitory deficits may impair performance not only on tasks that directly assess inhibitory ability, but also on assessments of working memory capacity: if older adults have a hard time distinguishing relevant from irrelevant information, this likely means that they store task-irrelevant information, reducing the storage capacity available for task-relevant information.

Speed of processing Older adults have a slower speed of processing than young adults. This slowed processing is noted at the motor level, but it also is apparent at a cognitive level. For example, older adults will tend to be slightly slower than young adults when they must slam on the brakes at a red light; this slowing may primarily be due to motor changes, because the association of red = stop remains very strong with aging. The reaction time differences between young and older adults will be exaggerated, however, if older adults must decide whether to slam on the brakes or to hit the gas as they approach a light that has just turned yellow. This additional slowing likely results because of the increased cognitive processing that must occur before the appropriate action can be selected.

Salthouse and colleagues have suggested that this decline in processing speed may underlie the age-related changes in cognitive function. It is apparent how slowed speed of processing could be detrimental to performance on any type of timed task. Importantly, however, a slower speed of processing could also manifest itself on nontimed tasks. For example, imagine that I read aloud the following arithmetic problem, and ask you to solve it in your head, with no time limit: "Jimmy walks up to a store counter with three packs of gum, each costing 50 cents. He gives the sales clerk $5. Because the clerk is out of dollar bills, she gives Jimmy his change in quarters. How many quarters does Jimmy receive from the sales clerk?" On the face of it, this is a task of working memory ability (the ability to store various pieces of information and to update the information as you work through the problem) that might be thought of as being independent from a measure of speed of processing, because there is no time limit for solving the task. However, if it takes someone a little longer to process the phrase "Jimmy walks up to a store counter with three packs of gum," it is possible that they will have a harder time attending to the phrase "costing 50 cents." Similarly, if it takes someone longer to multiply 3 by 50, it is possible that by the time that calculation is completed they will have forgotten the amount of money that Jimmy gave to the clerk. In other words, cognitive performance can suffer because the slowed mental operations cannot be carried out within the necessary time frame, and because the increased time between mental operations can make it more difficult to access previously processed information. Thus, a slower speed of processing may lead to a poorer encoding of information and a reduced ability to store information. In support of the hypothesis that processing speed changes may underlie much of the cognitive decline with aging, controlling for speed of processing often eliminates age differences on cognitive tasks, and longitudinal studies have shown a strong relation between changes in speed of processing and changes in performance on a large number of cognitive tasks.

Domain-Specific Theories of Cognitive Aging

In contrast to the domain-general theories of cognitive aging, domain-specific theories propose that some age-related declines may not be explained by core deficits that affect all aspects of cognition, but rather by changes that have a larger impact on one area of cognition than on another.

Word-finding difficulties and transmission deficits Older adults often have difficulties retrieving the appropriate name for a person, place, or thing. These word-finding problems can be manifest in various ways: excessive use of pronouns (due to difficulty generating the proper nouns), decreased accuracy and increased reaction time when asked to name items, and increased tip-of-the-tongue experiences. The tip-of-the-tongue state occurs when a person has access to a word's meaning, but not the phonological features of the word. The word seems just out of reach. Older adults tend to have more tip-of-the-tongue

experiences than young adults, particularly for proper names, and the accuracy of the phonological information available during a tip-of-the-tongue state (e.g., the first letter of the word, the number of syllables) tends to be more accurate for young adults than for older adults.

Burke, MacKay, and colleagues have suggested that these word-finding difficulties result from the fact that, with age, the links connecting one unit to another within the memory system become weaker. Thus, more links must be active in order for older adults to generate the correct name for an object or a person. This transmission deficit will mean that older adults will be relatively good at generating words when there are lots of links converging onto the word but will show larger impairments when trying to generate words that have fewer associated links. Most everyday objects have many semantic associations (e.g., individuals know that apples are red, of waxy texture, round, etc.). This convergence onto the word 'apple' (referred to by Burke and colleagues as summation of priming) makes it relatively easy to generate the name of the object. In contrast, proper names (with the exception of nicknames) are arbitrary and do not benefit from the same summation of priming (e.g., there is nothing that is required for someone to be a Jane vs. a Linda). Consistent with a transmission deficit, older adults remain relatively good at generating words of everyday objects for which they know a lot of semantic information but show larger impairments when asked to generate proper nouns.

Contextual memory and associative binding deficits Episodic memory can be thought of as including memory for two different types of information: memory for the item previously encountered (which typically can be based on familiarity) and memory for the broader context in which that item was encountered (which requires recollection). For example, when I pass by someone on the street, I might recognize that I have seen the person before (item memory). I also may remember that I met the person at a recent conference (an item–context association) or that the person is my colleague's husband (an item–item association).

Older adults remain quite good at using familiarity to recognize previously encountered people or items. However, older adults are particularly impaired at using recollection to remember the contextual details of an event. They seem to have difficulties binding together multiple event details into one cohesive memory. These deficits arise when trying to remember both item–item associations and item–context associations ('source' memory). Naveh-Benjamin and colleagues have proposed that this associative memory deficit underlies older adults' episodic memory difficulties.

Two broad types of memorial deficits may underlie this decreased ability to remember item–item or item–context associations. First, older adults have difficulties initiating effective encoding 'strategies' that would promote memory for the associative details of an experience (as proposed by Craik, Jennings, and colleagues). When they are given a strategy to use as they learn information (e.g., if they are asked to tell a story that binds the item to its context), older adults often perform as well as young adults on tasks requiring associative or contextual memory. Thus, at least some of the age-related deficits in remembering contextual details seem to result from deficits at encoding. Second, older adults seem to have difficulties either in forming a long-lasting 'bond' between an item and its context, or in retrieving that bound representation. Thus, given retrieval support (e.g., an untimed recognition task), older adults will tend to perform better than when given little retrieval support (e.g., a recall task).

Preserved Cognitive Function with Healthy Aging

Although the previous sections have focused on age-related declines, aging is not associated with across-the-board deficits in cognitive. In fact, some aspects of cognition remain markedly stable, or even improve, as individuals age.

Crystallized Intelligence

Crystallized intelligence refers to the ability to retrieve and use information that has been acquired throughout a lifetime. It often is contrasted with fluid intelligence, the ability to store and manipulate new information. As discussed earlier, fluid intelligence processes tend to be disrupted by healthy aging. Crystallized intelligence, in contrast, remains stable across the life span. Thus, older adults are very good (often better than young adults) at defining words, answering questions that rely on general world knowledge (e.g., "Who wrote the 'Star Spangled Banner'?"), detecting spelling errors, or carrying out skills related to jobs that they have held for many years.

Emotion Regulation

Another important area of preservation (or enhancement) is within the realm of emotion regulation. After about the age of 60, the ability to regulate emotion seems to start to improve. Thus, older adults show lower rates of depression than young adults. Compared to young adults, their good moods last longer,

and they are able to rebound more quickly from negative mood states. Older adults seem to focus more on positive information in their environment and to choose activities (e.g., spending time with close family or friends) based on their potential for emotional fulfillment.

At least some aspects of memory for emotional information also seem to be preserved with aging. While older adults tend to show overall poorer memory than young adults on a variety of tasks requiring retrieval of contextual information, studies of 'flashbulb memories' (memories for a public event that was highly surprising and emotional) have suggested that older adults, like young adults, are more likely to remember contextual details if the event contains emotional relevance than if it does not. The emotional memory benefit for older adults may be particularly pronounced for positive information, although a number of studies have suggested that older adults receive a memory boost when asked to remember negative experiences as well.

Neural Changes with Healthy Aging

Although healthy aging is associated with brain changes, not all regions are affected equally. To date, the vast majority of studies investigating the neural changes with healthy aging have used structural magnetic resonance imaging (MRI) (allowing examination of the volume of various brain regions) and functional MRI (fMRI)or positron emission tomography (PET) (allowing indirect measurements of neural activity as individuals are performing a task). These studies have suggested that the largest changes in structure and function occur in the prefrontal cortex and in the medial temporal lobe, while other cortical and subcortical regions remain relatively preserved across the life span. Current research also is focused on examining how the connections between regions are affected by aging. Diffusion tensor imaging studies are being conducted to examine age-related changes in white matter tracks, and structural equation modeling of fMRI and PET data is being used to investigate age-related changes in the functional connections between brain regions. These methods may provide new insights into the neural changes that mediate older adults' cognitive decline.

Changes in Prefrontal Cortex

The prefrontal cortex shows notable changes with aging. At a structural level, there is evidence of atrophy, both in the gray matter and in the white matter. The gray matter declines may reflect reductions in the number of cells (due to cell death) or may be a sign of neuronal shrinkage. The white matter changes reflect axonal abnormalities and may result in slowed neurotransmission. It is plausible that these white matter changes may mediate the cognitive slowing that accompanies healthy aging.

Prefrontal function also is altered with aging. Across a range of working memory and episodic memory tasks, older adults seem to show a different pattern of prefrontal activity than young adults, with reduced activity in some prefrontal regions and increased activity in other regions. As noted by Cabeza and colleagues, particularly in the prefrontal cortex, older adults often show a hemispheric asymmetry reduction. In other words, on tasks leading to unilateral prefrontal activity in young adults, older adults will tend to show bilateral recruitment. It is not yet clear whether this bilateral recruitment reflects compensatory activation, or whether it is a result of pathological changes (e.g., hemispheric release from inhibition). Some of the strongest evidence in favor of the compensatory view has come from studies comparing the pattern of neural recruitment in older adults who perform a task as well as young adults (high performers) and in older adults who perform the task more poorly (low performers). These studies have found that the high performers tend to recruit the prefrontal cortex bilaterally, whereas the low performers show unilateral prefrontal activity. To the extent that the prefrontal activity underlies the initiation of goal-relevant task strategies, it would make sense that the older adults who recruit additional prefrontal regions would be those who would be best able to perform the tasks. Nevertheless, future studies are required to clarify whether this compensatory hypothesis can account for all of the findings of bilateral prefrontal recruitment with age.

Medial Temporal Lobe Changes

The hippocampus proper is the other region that shows large age-related change. While there are structural changes in the hippocampus, it is not clear whether there is significant cell loss in the hippocampus with aging, or whether the structural changes are related more to neuronal atrophy (shrinkage). There also is ambiguity regarding the regions of the hippocampus that are most affected by aging. High-resolution MRI, allowing the distinction of the various hippocampal subfields (CA1, CA3, dentate gyrus, subiculum) may allow better assessment of the structure (and function) of these regions.

Functionally, the hippocampus tends to be under-recruited by older adults during both the encoding and the retrieval phases of recollective or associative memory tasks, and these functional changes often

correlate with the older adults' reduced performance on the tasks. Given the critical role of the hippocampus in forming vivid and detailed memories, it makes sense that the functional and structural changes in this region would correspond with older adults' difficulties in remembering the context in which information was learned.

Changes in Emotion Processing Regions

The fact that older adults show improved emotion regulation, and a memory enhancement for many types of emotional information, is consistent with the neural evidence indicating that emotion processing regions (particularly the amygdala and orbitofrontal cortex) are relatively spared in healthy aging. The amygdala shows minimal atrophy with healthy aging; its atrophy is on par with the decline in whole-brain volume (with a 1–3% reduction in volume every decade). Similarly, the orbitofrontal cortex seems to undergo little volumetric decline with age, particularly as compared to other regions of the prefrontal cortex.

Mild Cognitive Impairment

As discussed previously, some cognitive impairment is a natural part of the aging process. For some adults, however, advancing age is associated with fairly severe impairments in recent memory. Although the deficits do not impair their ability to function in daily life (and thus they do not meet the criteria for dementia), these individuals have cognitive impairments that exceed those that typically accompany healthy or successful aging. 'Benign senescent forgetfulness' was the first term used (by Kral and colleagues) to describe these individuals' impairments, although there were no strict diagnostic criteria associated with the concept. 'Age-associated memory impairment' (proposed by Crook and colleagues) was the first attempt at a standardized definition, requiring an individual to have subjective memory complaints and to perform at least 1 standard deviation below the mean for young adults on a standardized memory task. This concept was criticized by a number of researchers, who believed that the concept was too restrictive. Mild cognitive impairment (MCI) (defined by the Mayo Clinic Alzheimer's Disease Research Center) is the most recent in a series of attempts to characterize these individuals who straddle the boundary between healthy aging and dementia. A diagnosis of MCI requires subjective memory complaints and impairment in one area of cognition (scores must be more than 1.5 standard deviations below age-scaled norms), but with deficits not severe enough to interfere with activities of daily living or to result in a diagnosis of dementia.

There has been tremendous interest in defining this group of individuals, because as treatments that slow or reverse the development of Alzheimer's disease (AD) become available, it will be critical to have a method for diagnosing individuals at risk for, or in the prodromal stages of, the disease. Individuals with MCI seem to be an excellent population to be the targets of such treatments, because they are at increased risk for development of dementia, and of AD in particular. In fact, by some estimates, the vast majority of patients with MCI will eventually meet the diagnostic criteria for dementia.

The link between MCI and AD is supported not only by the high conversion rate, but also by the overlapping neuropathological and genetic features. Like AD patients, those with MCI have significant structural and functional changes in the medial temporal lobe. They also have alterations in the concentration of amyloid beta protein, the protein associated with neuritic plaque formation in AD. Moreover, the $\varepsilon 4$ allele of the apolipoprotein E, associated with an increased risk of developing AD, is overrepresented among individuals with MCI.

Alzheimer's Disease

AD was first described by Alois Alzheimer in 1907. In the landmark publication, Alzheimer reported a case study of a woman with severe psychiatric symptoms and memory deficits. An autopsy conducted upon her death revealed a large quantity of intracellular neuritic plaques and extracellular neurofibrillary tangles, now recognized as the pathologic hallmarks of AD.

Cognitive Changes in AD

Although dementia (a loss of intellectual function severe enough to interfere with daily activities) can result from a variety of etiologies, AD is by far the most common cause, accounting for an estimated two-thirds of all cases of dementia. Because AD can only be confirmed at autopsy, its clinical diagnosis must be an exclusionary one. Thus, the clinical profile of AD requires memory impairment plus decline in one other area of cognition (language, motor function, attention, executive function, personality, or object recognition). The deficits must have a gradual onset, and they must progress continually and irreversibly. When these criteria are met, a diagnosis of 'probable' AD is given. When made by a trained clinician, this diagnosis will be accurate in the vast majority (80–90%) of cases.

Episodic memory In contrast to healthy older adults, who remain able to successfully remember

previously encountered information (though perhaps not the context in which it was encountered), the most notable deficit for patients with mild AD is an inability to remember information encountered in the recent past. This deficit extends across different types of encoding tasks (e.g., incidental or intentional; deep or shallow processing) and exists regardless of the stimulus materials (e.g., pictures, words, faces, autobiographical events, emotional stimuli) or the task's retrieval demands (e.g., recall, forced-choice recognition, yes–no recognition). In fact, deficits in episodic memory tend to be the best way of distinguishing people with AD from healthy older adults.

Semantic memory In contrast to episodic memory, semantic memory (general world knowledge) is relatively spared with mild AD. As the disease progresses, however, significant semantic deficits arise. The deficits are particularly pronounced on word-finding tasks, with the extent of such deficits being useful for tracking the severity of AD. Current research is examining the extent to which the breakdown in semantic memory is due to changes in the structure of the memory networks (i.e., to a problem with the storage of such knowledge) or to difficulties retrieving the stored information (i.e., a problem with access).

Working memory and executive function In addition to deficits in long-term memory, Alzheimer's patients also show deficits in the online processing of information (working memory). Deficits are particularly pronounced on tasks requiring dual-task performance, suggesting that a primary deficit in AD may be in executive functions, the ability to flexibly shift attention and to attend to goal-relevant information.

Neural Changes in AD

Although plaques and tangles often are apparent throughout the brain in the later stages of AD, early in the course of the disease, the medial temporal lobe regions are those most affected. Even early in the disease, the hippocampal formation shows marked atrophy, and a volumetric reduction in the entorhinal cortex (which serves as the site of input to the hippocampus) is one of the best indicators that an individual has early AD. Given the essential role of the hippocampus for memory formation and retrieval (as demonstrated by the link between hippocampal damage and amnesia), it makes sense that patients with mild AD would be best identified by their difficulties remembering recently learned information (as discussed previously).

The amygdala is another region of the medial temporal lobe that is affected early in the disease process. This region, through its interactions with other medial temporal lobe structures, is thought to be essential for enhancing individuals' memories for highly emotional events. It is likely because of the amygdalar damage that patients with AD do not show a memory benefit for emotional information.

In addition to the medial temporal lobe changes, the nucleus basalis also shows significant cell loss in mild AD. This region of the ventral forebrain contains many of the brain's cholinergic neurons; thus, the damage to this region impedes cholinergic neurotransmission. Many of the first approved therapies for AD have been aimed at increasing the amount of acetylcholine in the brain, by blocking the function of acetylcholine esterase (the enzyme that breaks down the neurotransmitter acetylcholine). The minimal effectiveness of these acetylcholine esterase inhibitors in affecting cognitive function in AD patients suggests that the acetylcholine deficiency is not the only cause of the cognitive dysfunction in AD (and, indeed, at least by late-stage AD, there is marked depletion of other neurotransmitters, including norepinephrine, dopamine, and serotonin). The underwhelming effect of the cholinesterase inhibitors has led to continued research on alternate therapeutic options for AD, which may prove to be more effective in altering the progression of the disease.

Neural Changes in Later-Stage AD

In later stages of the disease, there is increased atrophy throughout the medial temporal lobe, and the cellular abnormalities also become apparent in the frontal lobe and throughout the temporal lobe. Because the frontal lobe is essential for higher-level executive functions, and the temporal neocortex is critical for the ability to retrieve semantic information, the advancing neuropathology in these regions clearly is linked to the cognitive deficits that arise in moderate AD. By the very late stages of AD, neuropathology is abundant throughout the cortical and subcortical structures, even within primary sensory regions (e.g., auditory, visual, and motor cortex).

Individual Differences in Aging

Unlike many diseases or illnesses that can be linked to a specific cause, MCI and AD appear to arise due to a combination of many factors, both environmental and biological. Research has begun to elucidate some of the traits that can increase the likelihood of disease development ('risk factors' such as prior head injury, mutations in the apolipoprotein, presenilin-1, and presenilin-2 genes, or having the $\varepsilon 4$ allele of apolipoprotein E) or that seem to reduce the probability of disease ('protective factors' such as high education

level, intake of antioxidants, or having the ε2 allele of apolipoprotein E). Research continues to investigate the reliability of these factors in predicting disease development and to examine whether any of the protective factors can alter disease progression once the pathological hallmarks of disease are present.

More broadly, researchers are beginning to recognize the necessity of taking an individual-differences approach to understanding the cognitive and neural changes that accompany aging. Research is now exploring the potential differences between high-performing and low-performing older adults. It is plausible that a better understanding of what differentiates the most successful agers from those who age less gracefully will lead to ideas for behavioral or neurobiological interventions that can boost the performance of individuals who experience significant age-related cognitive decline.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Aging and Memory in Humans; Alzheimer's Disease: MRI Studies; Cognition: An Overview of Neuroimaging Techniques; Dementia; Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Episodic Memory; Executive Function and Higher-Order Cognition: Neuroimaging; Hippocampus; Memory Consolidation: Systems; Prefrontal Cortex: Structure and Anatomy; Priming.

Further Reading

Burke DM and Mackay DG (1997) Memory, language, and ageing. *Philosophical Transactions of the Royal Society of London, B Series (Biological Sciences)* 352: 1845–1856.

Cabeza R (2002) Hemispheric asymmetry reduction in older adults: The HAROLD model. *Psychology and Aging* 17: 85–100.

De Toledo-Morrell L, Goncharova I, Dickerson B, Wilson RS, and Bennett DA (2000) From healthy aging to early Alzheimer's disease: *In vivo* detection of entorhinal cortex atrophy. *Annals of the New York Academy of Sciences* 911: 240–253.

Grady CL and Craik FIM (2000) Changes in memory processing with age. *Current Opinion in Neurobiology* 10: 224–231.

Kensinger EA (2006) Remembering emotional information: Effects of aging and Alzheimer's disease. In: Welsh EM (ed.) *Frontiers in Alzheimer's Disease Research*. Hauppauge, NY: Nova Science Publishers.

Li KZH and Lindenberger U (2002) Relations between aging sensory/sensorimotor and cognitive functions. *Neuroscience and Biobehavioral Reviews* 26: 777–783.

Light L (1992) The organization of memory in old age. In: Craik FIM and Salthouse TA (eds.) *The Handbook of Aging and Cognition*, pp. 111–165. Hillsdale, NJ: Erlbaum.

Mather M and Carstensen LL (2005) Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences* 9: 296–502.

Park DC and Gutchess AH (2002) Aging, cognition, and culture: A neuroscientific perspective. *Neuroscience and Biobehavioral Reviews* 26: 859–867.

Salthouse TA (1996) The processing-speed theory of adult age differences in cognition. *Psychological Review* 103: 403–428.

Cognition: An Overview of Neuroimaging Techniques

S A Bunge, University of California at Berkeley, Berkeley, CA, USA
I Kahn, Howard Hughes Medical Institute at Harvard University, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Anatomical Techniques

Anatomical techniques are used in a variety of ways in the service of the study of cognition, for example, to localize neuropathies in patients with cognitive disabilities or to compare the size of specific brain structures between groups of subjects through volumetric analysis. Additionally, anatomical techniques are used in conjunction with functional techniques in order to localize brain activity.

The earliest technique for imaging brain structure was computed tomography (CT). Today, it has been largely replaced by the much more powerful technique of magnetic resonance imaging (MRI). MRI provides excellent detailed structural information and enables the naked eye to distinguish gray matter (neuronal cell bodies) from white matter (myelinated tracts). New anatomical techniques, such as diffusion tensor imaging, have been developed to specifically visualize myelinated tracts. These methods can be used to track the normal and abnormal development of neural pathways in childhood.

Functional Techniques

Functional techniques have been the dominant force in cognitive neuroscience because they enable us to determine when and where neural activity in the brain is associated with the ability to perform a particular cognitive task. These techniques allow us to examine the workings of the human brain throughout the life span, in sickness and in health.

It is common to design several task variants that differ slightly in terms of task requirements. On the basis of this approach, differences in brain activation between the task variants enable us to isolate brain structures that are implicated in hypothesized cognitive processes. For example, if one wanted to identify brain regions involved in actively maintaining information in mind (referred to as working memory), one might parametrically vary the number of items (or the load) that subjects needed to maintain and then identify brain regions in which the level of activation varied according to working memory load. Although parametric variation is arguably the most powerful functional brain imaging manipulation, a more standard approach is to compare two task variants that differ only in that task 1 but not task 2 is hypothesized to engage a specific cognitive process. This approach relies on the assumption of 'pure insertion'; that is, the assumption that the insertion of the new task requirement affects only the cognitive process of interest.

Classes of Functional Neuroimaging Techniques

A variety of noninvasive functional neuroimaging techniques are available for use in humans (Figure 1); these techniques are categorized into two main classes. The first consists of methods for directly measuring electrical activity associated with neuronal firing, such as electroencephalography (EEG) and magnetoencephalography (MEG). The second main class consists of methods for indirectly measuring neuronal activity, which operate under the principle that neural activity is supported by increased local blood flow and metabolic activity. These methods include positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and near-infrared spectroscopy (NIRS). More information on each of these methods is provided in the following sections.

Direct measures of neural activity: EEG and MEG
EEG is the oldest functional brain imaging technique, dating back to Berger's discovery in 1929 that brain electrical activity could be recorded from electrodes placed on the scalp. This technique is still widely used today because of its ability to provide real-time measurements of brain activity. EEG records transient electrical dipoles generated by the net flow of electrical current across the cellular membrane during neuronal depolarization associated with postsynaptic potentials. Global EEG is used to measure neural activity during different brain states, such as sleeping and waking. A more powerful tool for cognitive neuroscience than global EEG measurements consists of event-related potentials, which refer to EEG activity averaged over a series of instances (or trials) triggered by the same event (e.g., the presentation of a visual stimulus).

Whereas EEG records the electrical activity associated with neuronal depolarization, the newer technique of MEG records the magnetic field produced by this electrical activity. The signal measured by both techniques results primarily from the activity of pyramidal neurons, which constitute roughly 70% of cells in the neocortex and are oriented perpendicular to the cortical sheath. EEG records electrical activity oriented perpendicular to the surface of the brain,

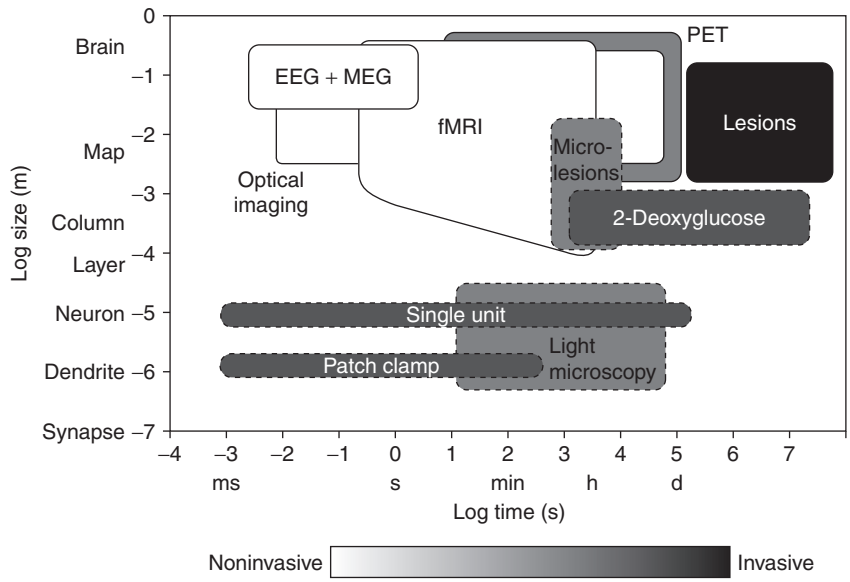

Figure 1 Relative spatial and temporal sensitivities of different functional brain imaging techniques. The level of invasiveness of each technique is indicated by a grayscale; more highly invasive techniques are shown in darker gray. Techniques used in animals only are outlined with a dashed line. Adapted from Cohen MS and Bookheimer SY (1994) Localization of brain function using magnetic resonance imaging. *Trends in Neuroscience* 17(7): 268–277.

whereas MEG records activity oriented parallel to the surface of the brain. Thus, EEG measures the activity of pyramidal cells in cortical gyri and the depths of the sulci, whereas MEG is sensitive primarily to the activity of pyramidal cells in the superficial parts of the sulci, and it is therefore more limited in its scope.

The major challenge with both EEG and MEG is referred to as the 'inverse problem,' which is the challenge of identifying the source of the underlying signal. This source can be a great distance from the point on the scalp at which it is measured, and it is affected by factors such as head shape and dipole location and orientation. Thus, it is necessary to build source localization algorithms to determine the likely source of a signal. One advantage of MEG over EEG is that the signal is less sensitive to factors such as head shape. Thus, the two techniques are complementary in that EEG samples from more neurons, but MEG is less sensitive to signal distortion. As such, some researchers take the approach of measuring EEG and MEG simultaneously to take advantage of what each technique has to offer.

Indirect measures of neural activity: PET Roy and Sherrington first showed in 1890 that brain stimulation led to a local increase in blood flow to active populations of neurons. Landau and others subsequently used radioactive tracers to measure regional cerebral blood flow (rCBF) in animals (1955), and in 1963 this technique was first applied to humans. The most widely used radioactive tracer is $^{15}O_2$, an oxygen molecule in which one electron has been removed from the atom to give an unstable form that will emit one positron to revert to the stable form $^{16}O_2$. In a PET scan, a small amount of radioactive tracer is injected into a vein. The tracer enters the brain after approximately 30 s, and in the following 30 s radiation in the brain rises to its maximal value; a picture of the rCBF is taken during this time frame.

The greatest advantage of PET over the more recent method of fMRI is the choice of radioactive tracer. Researchers can synthesize radiopharmaceutical compounds that bind to dopamine or serotonin receptors (C-11 or F-18 *N*-methylspiperone), opiate receptors (C-11 carfentanil), etc. PET is likely to continue to be important for understanding the role of various neurotransmitters in cognition. However, there are several disadvantages of PET that have led to its being surpassed by fMRI as the most widely used indirect measure of brain activity. The first of these is cost; PET facilities require not only a PET camera but also a cyclotron, which is used to produce the radioactive tracers. The second disadvantage is the poor temporal resolution – on the order of 1 min compared to 6–8 s for fMRI. Finally, PET is more invasive than fMRI, requiring the injection of a radioactive tracer (albeit one with a half-life of a few minutes), and thus it is not suitable for use in children or other special populations.

Indirect measures of neural activity: fMRI fMRI is currently the most widely used brain imaging technique for a number of reasons, including (1) widespread availability of MRI scanners and technology,

(2) relatively low cost per scan, (3) the lack of recognized risks for properly screened subjects, (4) good spatial resolution, and (5) better temporal resolution than other indirect neuroimaging methods.

In addition to observing that brain activity was associated with local changes in blood flow, Roy and Sherrington noted that the increase in total oxygenated blood delivered far exceeded the demand. It is this surplus of oxygen that is detected with fMRI, with what is known as the blood oxygen-level dependent (BOLD) contrast. Oxygenated hemoglobin, or oxyhemoglobin, has magnetic properties different from those of deoxyhemoglobin. Thus, the deployment of a strong magnetic field (typically 1.5–4 tesla when used in humans) enables one to measure changes in the ratio of oxy- to deoxyhemoglobin in local draining venules and veins. An increase in blood flow in response to a specific stimulus (be it brain stimulation or the presentation of a visual or auditory stimulus) is referred to as a hemodynamic response. This response is sluggish with respect to the underlying neural activity; for example, the hemodynamic response in visual cortex to a brief (e.g., 30 ms) visual stimulus peaks roughly 4–6 s after the onset of the stimulus.

Related Techniques

Perfusion fMRI enables direct measurement of the hemodynamic response, unlike BOLD fMRI, which relies on the comparison of hemodynamic responses under different conditions. This technique is less widely used than BOLD fMRI because of its much lower sensitivity. Magnetic resonance spectroscopy is used to measure the relative concentration and distribution of specific ions or compounds, such as hydrogen, phosphorus, or carbon.

Optical Brain Imaging

NIRS operates according to the same principle as optical brain imaging techniques that have been used in animals. Both techniques capitalize on the fact that changes in hemoglobin concentration in cortical tissue affect the absorption of infrared light by the tissue. However, NIRS is noninvasive, whereas optical imaging in animals is accomplished by exposing the surface of the brain. Because the cortical surface is not exposed in NIRS, light scattering by the skull leads to reduced spatial resolution.

During the past few years, Gratton, Fabiani, and colleagues have developed a new analytic approach to the analysis of optical data, known as the event-related optical signal (EROS). The EROS signal is based on measures of the optical properties of cortical tissue, which change when the tissue is active. The changes in these optical properties are likely to be due to changes in light scattering as a function of neuronal activity. Because this type of optical signal is directly linked to neural activity, rather than blood flow, it provides a much higher temporal resolution than the signal typically analyzed in NIRS studies. Additionally, the EROS signal has a high degree of spatial resolution, in that it can be localized to an area of less than 1 cm^3.

A drawback of all three optical imaging techniques noted previously is that they can only be used to measure activity on or near the surface of the brain (within 3–5 cm of the surface). However, these techniques are advantageous in that they are relatively inexpensive and completely noninvasive. Furthermore, because these techniques are not as sensitive to head movement as fMRI, they hold much promise for the study of clinical populations and children.

Trade-Offs between Temporal and Spatial Resolution

Direct measures of neural activity, such as EEG and MEG, have exquisite temporal resolution. They are sensitive to changes in neural activity on a millisecond resolution. However, these methods have poor spatial resolution in that it is difficult to pinpoint the precise origin of the signal. In contrast, indirect measures of neural activity have better spatial resolution than EEG or MEG but poor temporal resolution. The spatial and temporal characteristics of these various methods are shown in Figure 1. Thus, there is currently a trade-off between high temporal precision and high spatial precision. One solution to this problem is to use fMRI data to limit the possible sources of neural activity in EEG and MEG data; another – and even more challenging – solution is multimodal imaging, for example, the concurrent acquisition of fMRI and EEG data.

Advantages and Limitations of Neuroimaging Techniques

To date, much of what we know about the functions of different brain regions comes from neuropsychological studies in humans and lesion studies in animals. A distinct drawback of neuroimaging techniques is that, unlike neuropsychological and lesion techniques, they cannot demonstrate the necessity of a brain region for a specific cognitive process. However, neuroimaging techniques can demonstrate that humans without neurologic damage routinely recruit this region to perform the task. Moreover, it is possible to compare levels of activation across individuals and to state that greater activation of this region is

associated with better performance on a task. Better yet, it is possible to compare activation between trials within an individual and to state that on trials in which the subject made an error, activation was lower in this region than on trials in which the subject performed correctly. Thus, functional brain imaging techniques can be used to characterize a region's contribution to specific cognitive processes. Moreover, brain imaging techniques can be paired with techniques for temporarily disrupting neural activity in a temporally and spatially precise manner (transcranial magnetic stimulation).

There are several important advantages of neuroimaging techniques over neuropsychological ones. First, neuropsychological studies necessarily rely on the output of behavior as the critical dependent measure, whereas neuroimaging studies can focus on cognitive processes that take place prior to – or are not associated with – a behavioral response. For example, it is impossible to determine whether lesions that result in a long-term memory impairment are associated with a deficit at encoding and/or retrieval. With brain imaging, however, one can identify brain regions associated with the effective encoding of memories separately from brain regions associated with effective retrieval.

Second, neuroimaging techniques enable us to identify the entire neural circuit underlying a cognitive process. Lesion studies in animals can accomplish this only piecemeal by lesioning each area in turn. Neuropsychological studies in humans would be hard-pressed to accomplish this at all, given the limited availability of patients with specific brain lesions. This point is an important one because with lesion studies it is possible to completely overlook – or instead to overestimate – the contributions of a specific region to cognitive function. For example, the hippocampus has long been thought to be the primary structure contributing to the encoding and retrieval of memories. However, brain imaging has demonstrated that in the healthy brain, prefrontal cortex routinely contributes to memory encoding and retrieval.

The third point is also related to the limited availability of neuropsychological patients: neuroimaging techniques enable one to examine the role of a given human brain region in cognition even if one does not have access to specific patient groups. Moreover, it can be very difficult (or next to impossible, depending on the brain region) to find a patient with a lesion limited to the region in which one is interested because most brain lesions are rather coarse. Furthermore, cortical reorganization can lead to recovery of function over time, which could lead one to assume mistakenly that a lesioned brain region is not normally involved in a specific cognitive process.

Finally, neuroimaging techniques with high temporal resolution, such as EEG and MEG, provide us with important clues about brain mechanisms. It is helpful to say that brain regions X and Y are involved in memory retrieval, but if we can say that brain region X is active a short time after region Y, we can begin to gain insight into the cascade of events leading to memory retrieval.

A common criticism of neuroimaging studies is that they can be highly unconstrained and atheoretical. Like any scientific tool, however, brain imaging can be used wisely or foolishly. Certainly, many exploratory brain imaging studies have been conducted, particularly in the initial phase of brain imaging, when it was important to validate the new techniques. However, the second generation of imaging studies, starting in the late 1990s, has been much more focused. On the whole, cognitive neuroscientists carefully devise experimental manipulations designed to test psychological theories or specific predictions about brain function. Brain imaging techniques are likely to be an important force in cognitive neuroscience for the foreseeable future.

Contributions to the Study of Cognition

Neuroimaging research has been used to enrich our understanding of the neural basis of a wide variety of cognitive abilities, including attention, language, and memory. In addition, neuroimaging techniques have been used to gain insight into the etiology of neurobehavioral disorders, for surgical planning, and to assess functional recovery after brain damage.

One area in which neuroimaging studies are making important contributions is that of memory. For example, studies of long-term semantic memory have begun to uncover the ways in which stored information is organized in the brain. These studies show that different attributes of an object are stored in a distributed manner across several brain regions, with visual form information stored in a region that processes form, and functional information stored near a region that processes motion. Furthermore, studies of memory encoding have been able to pinpoint the precise regions for which level of activation during stimulus processing predicts subsequent memory for that stimulus. Additionally, studies of memory retrieval are being used to adjudicate between models of episodic memory positing that recollection (i.e., distinct remembrance of an item and the context in which it was previously encountered) and familiarity (a vague sense that the item has previously been encountered) are either distinct processes or merely

on a continuum of memory retrieval. These examples show how brain imaging studies can be used to test or adjudicate between psychological models.

Perhaps the most important contribution of neuroimaging to the field of cognition will be in the study of higher cognitive functions, which are not highly developed in nonhuman species and are therefore best studied in humans. Carefully designed brain imaging experiments have begun to fractionate the cognitive processes that underlie language, reasoning, problem solving, and other high-level mental functions, but further investigation is necessary.

See also: Cognitive Neuroscience: An Overview; Cognitive Control and Development; Electroencephalography (EEG); Event-Related Potentials (ERPs) and Cognitive Processing; Executive Function and Higher-Order Cognition: Neuroimaging; fMRI: BOLD Contrast; Magnetic Resonance Spectroscopy; Neuroimaging; Perfusion MRI; Single Photon Emission Computed Tomography (SPECT): Technique; Voxel Based Morphometry.

Further Reading

Andreassi JL (1995) *Psychophysiology: Human Behavior & Physiological Response,* 3rd edn. Hillsdale, NJ: Lawrence Erlbaum.

Cohen MS and Bookheimer SY (1994) Localization of brain function using magnetic resonance imaging. *Trends in Neuroscience* 17(7): 268–277.

Frith CD and Friston KJ (1997) Studying brain function with neuroimaging. In: Rugg M (ed.) *Cognitive Neuroscience*, pp. 169–195. Cambridge, MA: MIT Press.

Kutas M and Dale A (1997) Electrical and magnetic readings of mental functions. In: Rugg M (ed.) *Cognitive Neuroscience*, pp. 197–242. Cambridge, MA: MIT Press.

Levy I, Hasson U, Avidan G, Hendler T, and Malach R (2001) Center-periphery organization of human object areas. *Nature Neuroscience* 4(5): 533–539.

Raichle ME (1998) Behind the scenes of functional brain imaging: A historical and physiological perspective. *Proceedings of the National Academy of Sciences of the United States of America* 95(3): 765–772.

Rorden C and Karnath HO (2004) Using human brain lesions to infer function: A relic from a past era in the fMRI age? *Nature Reviews Neuroscience* 5(10): 813–819.

Wilkinson D and Halligan P (2004) The relevance of behavioural measures for functional-imaging studies of cognition. *Nature Reviews Neuroscience* 5(1): 67–73.

Cognition: Basal Ganglia Role

A R Aron, University of California, at San Diego, La Jolla, CA, USA
R A Poldrack, University of California at Los Angeles, CA, USA
S P Wise, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

The basal ganglia receive widespread input from the cerebral cortex and thalamus and return that projection through output to thalamocortical pathways. The basal ganglia also receive important neurochemical inputs, in particular a prominent dopaminergic projection from the midbrain. Because damage to the basal ganglia – in a broad range of neurological disorders – produces impairments in movement, expert opinion long held that the basal ganglia had an exclusively motor role. It is now clear, however, that these structures also contribute to a wide range of cognitive functions. These functions include learning, memory, skill, planning, switching, sequencing, timing, and the processing of rewards and other feedback. This article reviews evidence relating some of these cognitive functions to the basal ganglia.

Anatomical and Neurochemical Considerations

To understand how the basal ganglia contribute to cognitive functions, it helps to consider the pathways linking their component structures: the striatum, globus pallidus, substantia nigra, and subthalamic nucleus (STN) (Figure 1). Some of these pathways facilitate basal ganglia output, others inhibit it. In the next section, we discuss how different pathways engage in initiating, inhibiting, and switching behaviors, and also briefly mention a role in sequencing a series of behaviors.

In addition to the pathways illustrated in Figure 1, larger circuits – often called loops or modules – link different parts of cortex with distinct regions within the basal ganglia (Figure 2). Each of these circuits, which involve cortex, striatum, pallidum, and thalamus, carry different kinds of information. The terms often applied to the loops, such as associative (or complex), motor, oculomotor, and limbic, or, as illustrated in Figure 2, motor, spatial, visual, and affective, are not particularly significant. More important is the concept, which contrasts with earlier ideas about the anatomy of cortex–basal ganglia interactions, that many inputs and outputs of the basal ganglia involve cortical areas that have little direct sensory or motor function. The outputs of the basal ganglia do not, as previously believed, preferentially target motor areas of the cerebral cortex; instead, they target multiple sectors of cortex, including prefrontal cortical regions implicated in cognitive processing. This fact explains how damage to the striatum, such as occurs in Huntington's disease, can produce a range of cognitive and affective symptoms, even before motor performance deteriorates. The modules illustrated in Figure 2 take the form of open rather than closed loops. That is, different loops share information, through corticocortical connections among others, so that diverse influences can contribute to the basal ganglia's final output. Most behaviors probably require cycling through multiple loops in an interactive fashion.

The neurochemical inputs to the basal ganglia also play a vital role in their function (Figure 3). In particular, prominent dopaminergic pathways originating in the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) of the midbrain primarily innervate the dorsal and ventral striatum, respectively. As elaborated in the following discussions, several current models posit that the dopaminergic input to the dorsal striatum provides a reinforcing signal that binds the representations of stimuli and responses into stable associations, whereas the dopaminergic input to the ventral striatum plays a key role in rewarding or affective aspects of motivated behavior.

Overall, the cognitive function of the basal ganglia could be viewed as one of assigning behaviorally or biologically relevant feedback (Figure 3) to the computations performed by the various parallel loops involving the cerebral cortex, thalamus, and basal ganglia (Figure 2), with the basal ganglia's output controlled by the several different pathways converging on the pallidum (Figure 1). In the next section, these pathways are discussed in the context of preparing, initiating, inhibiting, and switching behaviors, while the subsequent two sections address the role of the basal ganglia in the processing of reward and feedback, and in the learning of new actions.

Cognitive Control: Preparing, Initiating, Inhibiting, and Switching Responses

Many lines of evidence in humans, nonhuman primates, and rodents implicate frontal cortex–basal ganglia loops in the control of responses, including both lower-order stimulus–response behaviors and the higher-order rules guiding such behaviors.

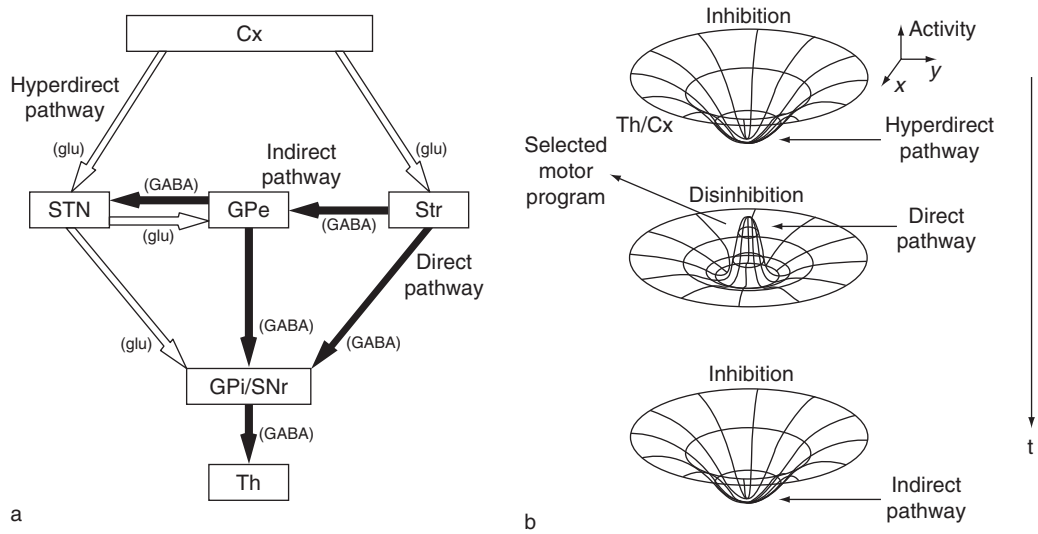

Figure 1 (a) Schematic diagram of hyperdirect, direct, and indirect pathways through the basal ganglia. Open arrows represent excitatory glutamatergic connections; filled arrows represent inhibitory γ-aminobutyric acid (GABA)ergic connections. Cx, cerebral cortex; STN, subthalamic nucleus; GPe, external segment of the globus pallidus; Str, striatum; GPi, internal segment of the globus pallidus; SNr, substantia nigra pars reticulata; Th, dorsal thalamus; glu, glutamatergic projections. (b) Schematic diagram showing the center-surround model of dynamic basal ganglia function, depicting activity changes in thalamus or cortical targets across time (vertical) caused by sequential inputs through the hyperdirect pathway (top), which inhibits all thalamocortical programs, the direct pathway (middle), which disinhibits some thalamocortical programs with concurrent suppression of alternatives, and the indirect pathway (bottom), which provides further inhibition of currently active programs to facilitate sequencing, among other functions. Reproduced from Nambu A, Tokuno H, and Takada M (2002) Functional significance of the cortico-subthalamo-pallidal 'hyperdirect' pathway. *Neuroscience Research* 43: 111–117, with permission from Elsevier.

Stimulus–Response Behavior

In some studies of stimulus–response behaviors, participants receive a cue that signals which task – of two or more tasks – they should perform once a stimulus appears. According to current thinking, this task cue evokes a top-down process by which the prefrontal cortex biases response channels within the basal ganglia. This bias could be understood to result from frontal inputs to the striatum, which excite a subpopulation of striatal cells and therefore prime the striatum for a particular type or class of sensory input (Figure 4(a)). If the expected stimulus appears, then the participants' reaction times will be quicker. Consistent with this model, individuals respond more quickly with adequately long intervals between the task cue and the subsequent stimulus, whereas they respond more slowly with shorter intervals. Functional imaging and neurophysiological data support these behavioral observations by showing that predictive cues lead to increased activation of striatum, compared to nonpredictive cues.

Response initiation begins once the stimulus appears. As illustrated in Figure 5(a), this event engages the direct pathway of the basal ganglia, eventually leading to sufficient activation of motor cortex to generate the motor response or support an ongoing motor command initiated through other mechanisms. Inhibiting an already initiated response could occur via frontal activation of the STN, either via the hyperdirect or indirect pathways (Figure 5(b); also see Figure 1). The STN, in turn, activates the internal segment of the globus pallidus (GPi), which inhibits thalamocortical output. Functional imagining has shown activation of the STN by response inhibition in humans. Further, patients with implanted STN stimulators (for treatment of Parkinson's disease) have improved response inhibition when the stimulator is turned on compared to when it is turned off, and rodents with lesions of the STN have impaired response inhibition.

In addition to their role in inhibiting already initiated responses, the basal ganglia may also be important for inhibiting incongruent or irrelevant responses or those recently performed as a movement sequence unfolds. For example, in performing a sequence of movements in which an action by the index finger is followed by a programmed movement of the thumb, the neural representation of the index-finger movement should be inhibited after the completion of the first part of the sequence. This kind of inhibition could also depend on the indirect pathway (Figure 1(a)).

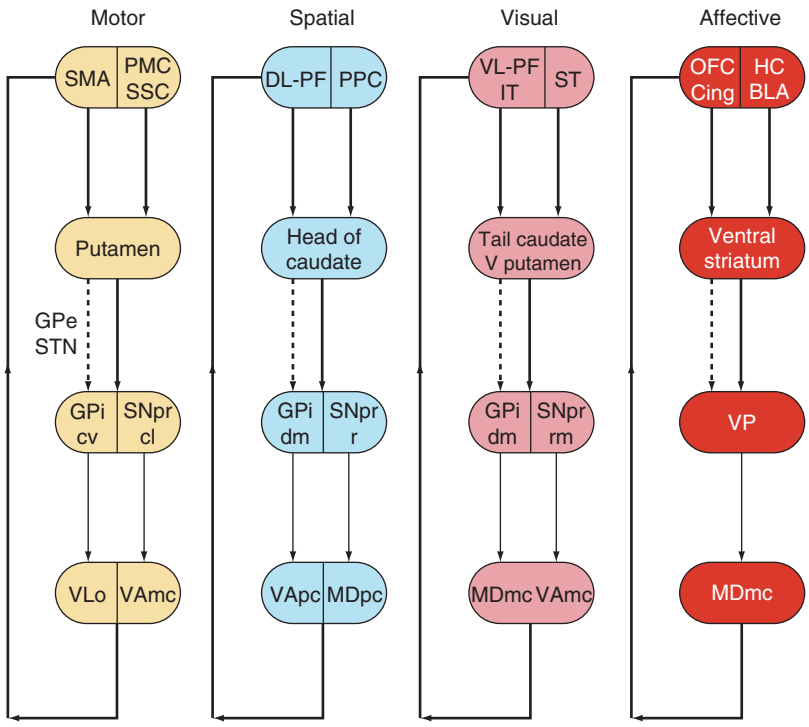

Figure 2 Cortical–basal ganglia loops, modified from the scheme of Alexander et al. Four parallel loops are shown, with possible functions labeled at the top. In the top row of rounded rectangles, the cortical areas in the left parts of bisected boxes are those that receive the thalamic projections indicated. The second row represents the striatum, the third row the pallidum, and the bottom row the dorsal thalamus. Dashed lines indicate net inhibitory influences of the indirect striatal output pathway. SMA, supplementary motor area; PMC, premotor cortex; SSC, somatosensory cortex; DL-PF, dorsolateral prefrontal cortex; PPC, posterior parietal cortex; VL-PF, ventrolateral prefrontal cortex; IT, inferior temporal cortex; ST, superior temporal gyrus; OFC, orbitofrontal cortex; Cing, anterior cingulate cortex; HC, hippocampal cortex; BLA, basolateral amygdala; V putamen, ventral putamen; GPe, external segment of globus pallidus; STN, subthalamic nucleus; GPi, internal segment of the globus pallidus; SNpr, substantia nigra pars reticulata; VP, ventral pallidum; VLo, oral (anterior) part of the ventrolateral thalamic nucleus; VA, ventral anterior nucleus of the thalamus, including the magnocellular (VAmc) and the parvocellular (VApc) parts; MDmc, magnocellular part of the mediodorsal nucleus of the thalamus. Subdivisions of the pallidum are labeled as follows: cv, caudoventral; cl, caudolateral; dm, dorsomedial; r, rostral; rm, rostromedial. Reproduced from Lawrence A, Sahakian BJ, and Robbins TW (1998) Cognitive functions and corticostriatal circuits: Insights from Huntington's disease. *Trends in Cognitive Sciences* 2: 379–387, with permission from Elsevier.

Task Switching

The same ideas that apply to stimulus–response behavior could contribute to understanding task switching, as well. A cardinal function of the basal ganglia involves switching between different tasks (or responses). Switching between tasks can be tested with a paradigm such as that shown in Figure 4(b). On some trials, the experimenter cues the subject to respond according to the 'letter' task (say aloud the letter in the upcoming stimulus, and ignore the digit), whereas on other trials the cue instructs a response according to the 'digit' task (say aloud the digit, and ignore the letter). Individuals have a slower reaction time on trials in which the they switch to a different task compared to trials on which they repeat the same task, a phenomenon known as the switch cost (Figure 4(c)). Switching likely requires a combination of both facilitating (or preparing) some responses and inhibiting others. Many studies have shown that patients with Parkinson's and Huntington's disease are impaired at switching tasks, and neuroimaging studies in healthy persons have shown increased activation of the striatum for trials with switches, compared to trials lacking them.

In terms of the model shown in Figure 4(a), frontostriatal inputs assist switching. As described earlier, the prefrontal cortex may bias the relevant response channels within the striatum in advance of the stimulus. If the expected stimulus later appears, then reaction time will be quick. If the respondent repeats the same task or response, reaction time will stay quick, because the relevant response channels remain primed, both by the top-down signal and by the just-performed response. However, if the respondent needs to switch tasks, the prefrontal cortex needs to activate or bias new response channels, and possibly inhibit previously

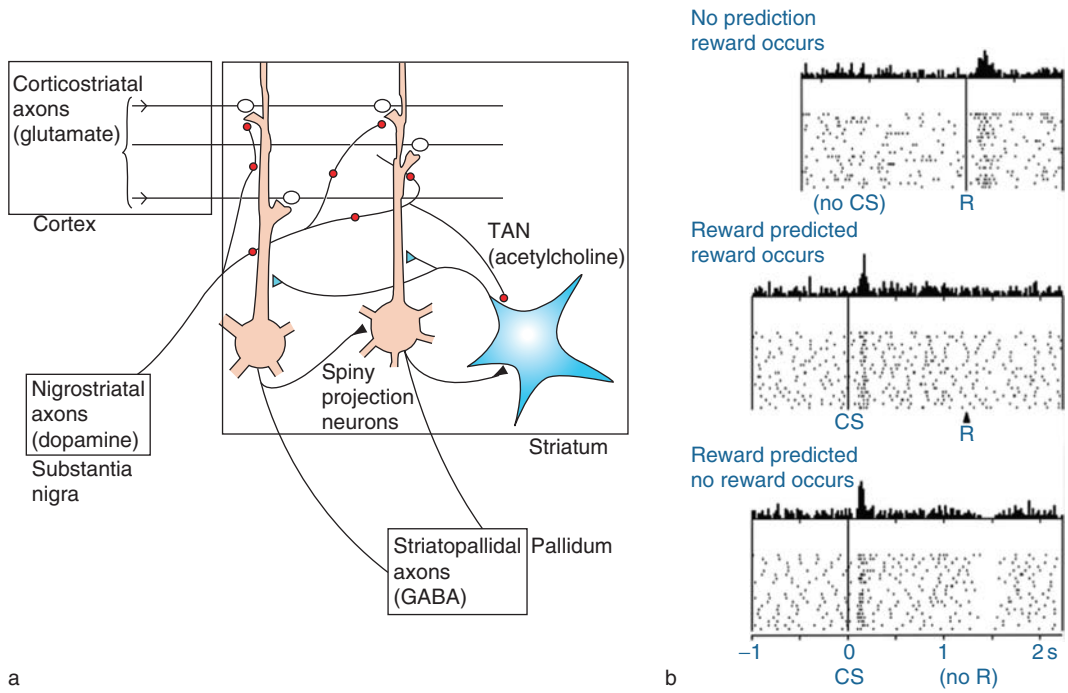

Figure 3 (a) Schematic diagram of striatal microcircuitry involved in reward-related motor learning. Corticostriatal axons make glutamatergic synaptic contacts (open circles) with the medium spiny projection neurons and cholinergic interneurons, thought to be the tonically active neurons (TAN). Nigrostriatal axons make dopaminergic synapses (filled circles) in the vicinity of corticostriatal synapses (as illustrated) and thalamostriatal synapses (not shown). Spiny projection neurons make γ-aminobutyric acid (GABA)ergic synapses (black triangles) on other spiny projection neurons and TANs, which, in turn, make cholinergic synapses (gray triangles) on spiny projection neurons. Phasic activity in the dopamine neurons produces pulses of dopamine able to act on corticostriatal and thalamostriatal synapses. This leads to strengthening of some synapses and weakening of others, which in turn facilitates selection of particular pathways. (b) Dopamine neurons convey an error in the prediction of reward. Each line of dots indicates a trial and each dot signifies the time that a single dopamine neuron fired an action potential in relation to the alignment event (middle vertical line). Top: No conditioned stimulus (CS). A drop of liquid reward (R) occurs unexpectedly, which constitutes a positive error in the prediction of reward, and the neuron increases its discharge rate. Middle: A learned, conditioned stimulus predicts a reward, and the reward occurs according to the prediction; hence, no error in the prediction of reward. The dopamine neuron does not respond to the predicted reward (at the time indicated by R), but instead increases its discharge rate following the reward-predicting CS (left). Bottom: Omission of predicted reward. The activity of the dopamine neuron decreases at the approximate time of an expected reward. (a) Adapted from Wickens JR, Reynolds JN, and Hyland BI (2003) Neural mechanisms of reward-related motor learning. *Current Opinion in Neurobiology* 13: 685–690, with permission from Elsevier. (b) Reproduced from Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

active ones. The interference between a currently relevant response and the recently primed one could partly explain the switch cost.

Rewards: Modulation of Striatal Signal Processing by Biological Feedback

Initiating, switching, inhibiting, and sequencing responses applies most to human behaviors in which task instructions are well understood or performance is relatively automatic. However, in many real-world situations the task to perform is not obvious and must therefore be learned or selected on the basis of biologically relevant feedback.

In many research settings with human participants and in nearly all studies with nonhuman ones, feedback is given in the form of primary reinforcers, such as water or food. Neurobiologically, feedback is coded by the firing rate of dopaminergic cells of the midbrain (Figure 3). In monkeys, this rate does not change significantly for highly overlearned behaviors, but, during the process of learning, these cells show significant modulation, which conveys to the basal ganglia either a reward signal or a signal that reflects an error in the prediction of reward. Therefore, as illustrated in Figure 3(b), when an unpredicted reward occurs, the dopaminergic cells increase their discharge rates, but for highly predicted rewards they do not. When a predicted reward fails to materialize, these cells decrease their discharge rate. With repeated experience, the cells become unresponsive to rewards *per se*, but discharge in response to signals, such as

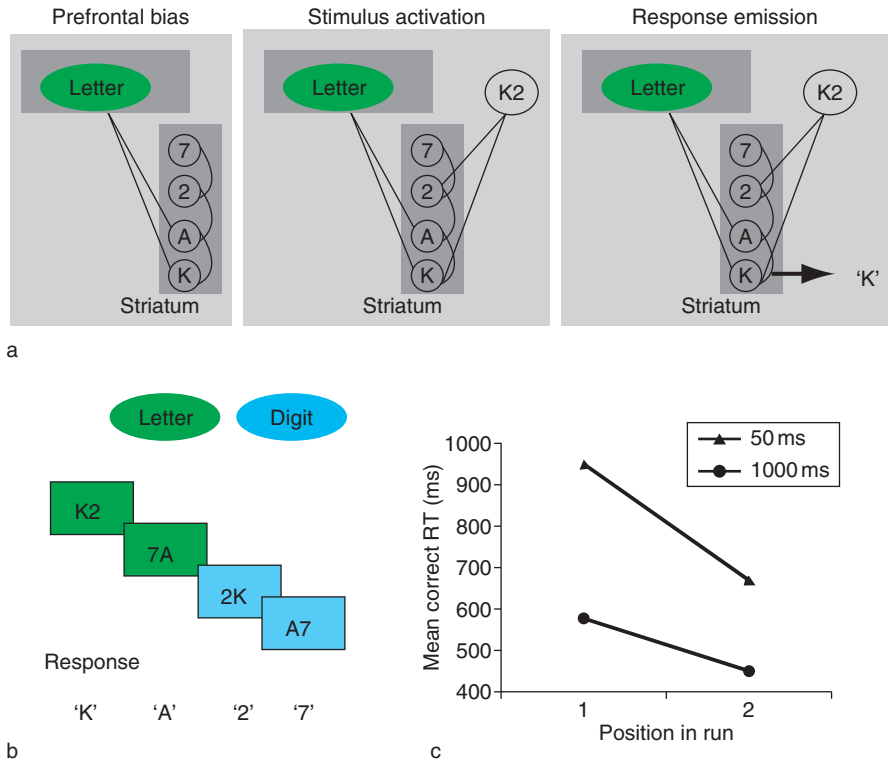

Figure 4 (a) A schematic model of how the prefrontal cortex could create top-down bias within the striatum to guide responding. The model is shown in three panels, representing different stages of task processing. Left panel: The person's knowledge of the upcoming task (letter task) creates bias over a subpopulation of striatal response channels by biasing 'A' and 'K.' Middle panel: The visual stimulus is shown ('K2'), which triggers striatal response channels for both a letter and a digit. Right panel: The striatal response channels are in mutual competition (e.g., via lateral inhibition), and the top-down prefrontal bias creates an advantage in favor of the letter channel, so that response is 'K.' (b) Illustration of a task-switching paradigm. On each trial the person responds as fast as possible to a stimulus such as 'K2' or '7A' by attending to the relevant task (letter or digit), which is cued (either 50 or 1000 ms) prior to the stimulus by means of a green- or yellow-colored rectangle. The third trial in the series is a switch trial because the subject must now attend to the digit, whereas previously the letter was relevant. In this paradigm, there are runs of two: the first position is always a switch (of task), the second position is always a repeat (of task). (c) Switching the task is associated with a longer reaction time (RT) than is repeating the task, and this effect is especially evident when the cue–stimulus interval is short (50 ms). This phenomenon is called the preparation effect.

conditioned stimuli, that predict reward. In accord with these findings in monkeys, studies in rats show that dopamine release increases as they approach a lever that they press to produce a reward. In rats trained to associate an initially neutral cue with the reward, the presentation of those cues also leads to an increase in dopamine release.

These data thus show that the dopaminergic signal to the basal ganglia provides feedback about response outcome in biologically significant terms. This feedback has the proposed function of strengthening some striatal synapses and weakening others. This synaptic plasticity, in turn, facilitates specific striatal output pathways, thus increasing the probability that a given input will trigger a given response. Empirical evidence provides support for this model. Neurons in the basal ganglia are highly sensitive to locations or movements associated with different levels of predicted reward. When monkeys choose to turn a handle leftward or rightward, and each choice leads to a different likelihood of reward, the activity of many striatal neurons correlates best with the predicted value of one of the two movements. Fewer neurons have activity that reflects the relative value of the two choices or the movements selected. Similar properties have been observed in the eye-movement part of the striatum, a part of the caudate nucleus. When monkeys make saccadic eye movement to either a left or a right target, and, for example, only saccades to the left target produce a reward for a given block of trials, neuronal activity is greater for leftward saccades than it is during a subsequent block of trials, when the same saccade fails to produce a reward. This higher level of activity correlates with quicker reaction times for the saccade target that will produce a reward. In terms of the model discussed earlier (Figure 4(a)), the higher

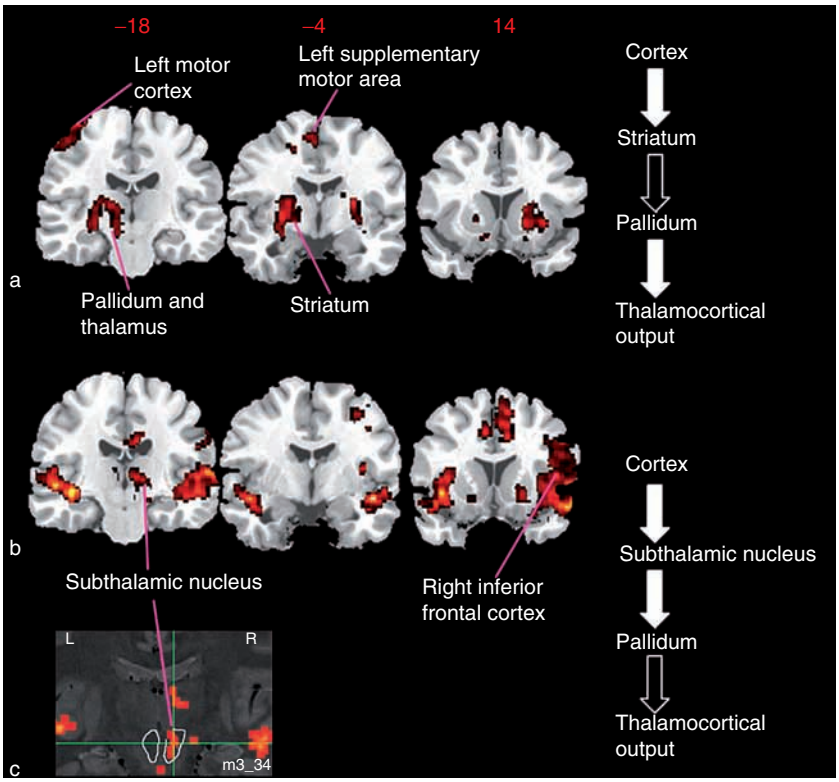

Figure 5 Functional imaging of executing a motor response and inhibiting an already initiated motor response. (a) Executing a motor response with the right hand. Activation is seen in the left (contralateral) supplementary motor area, striatum, pallidum, thalamus, and left motor cortex, consistent with the direct pathway through basal ganglia (right panel). (b) Inhibiting an initiated right-hand response activates the right subthalamic nucleus (confirmed by a high-resolution scan; see (c), in which the subthalamic nucleus is outlined by the white line, one in each hemisphere). This activation could result from inputs from the hyperdirect pathway (right panel), by which motor areas of the cortex excite the subthalamic nucleus, which, in turn, excites the pallidum and so suppresses thalamocortical output. Activation also occurs in the right inferior frontal cortex, consistent with the long-standing idea that the prefrontal cortex plays a key role in behavioral inhibition. At the top, the numbers in red denote Montreal Neurological Institute positions (in millimeters, posterior or anterior to the midpoint of the anterior commissure–posterior commissure line) of coronal slices through the brain. In the far right, unfilled arrows signify inhibitory projections, filled (white) arrows indicate excitatory ones. Adapted from Aron AR and Poldrack RA (2006) Cortical and subcortical contributions to Stop signal response inhibition: Role of the subthalamic nucleus. *Journal of Neuroscience* 26: 2424–2433.

level of activity reflects the current priming or bias for movements to that location, based on the prior probability of rewards.

Many functional imaging studies confirm a reward signal in the basal ganglia, especially in the ventral striatum. In some studies, activation in the striatum correlates with predicted reward; in others, it correlates with errors in the predicted reward or with negative reinforcement. Overall, it is thought that the response-channel biases described in the previous section occur in the striatum as a result of biologically relevant feedback from the midbrain dopaminergic system. In the aforementioned studies of eye movements, the striatum receives both cortical inputs bearing spatial and motor signals and dopaminergic inputs signaling reward or errors in reward prediction. This integrated signal, then, influences the motor system through its pallidal output. A key component of the basal ganglia's role in cognition thus involves the assignment of biologically relevant feedback to the computational operations of each of the loops illustrated in Figure 2, especially those involving the prefrontal cortex. In modulating the information processing of all cortical–basal ganglia loops at the striatal level, the dopaminergic inputs can influence a wide range of cognitive behaviors, including learning, memory, skills, and habits, the topics of the next section.

Learning, Memory, Skills, and Habits

Given its role in the adjustment of behavior based on feedback, it is not surprising that the basal ganglia play an important role in the learning of new skills and habits. For example, disorders of the basal ganglia impair learning of new motor skills, such as

rotary pursuit, in which the individual attempts to keep a stylus positioned on a moving turntable. Because patients with basal ganglia disorders have impaired motor control, it can be difficult to interpret any deficits strictly in terms of an impairment in motor skill learning. However, it has been shown that these patients are impaired at motor learning even when the task is made easier, assuring that the patients and the controls start out at the same level of performance. Interestingly, some forms of motor skill learning, such as mirror tracing, are not impaired in patients with basal ganglia disorders. This can be explained based on the model that the basal ganglia may play a role in learning those tasks that require predictions generated internally (known as open-loop tasks, such as rotary pursuit), but not in tasks in which behavior is directly guided by external stimuli (known as closed-loop tasks, such as mirror tracing). This is consistent with the observation that patients with basal ganglia disorders are disproportionately impaired on tasks that require self-initiated movements, compared to visually guided ones.

The involvement of the basal ganglia in learning extends to perceptual and cognitive skills, as well. One paradigm that has been extensively investigated is probabilistic classification learning, in which an individual learns to classify stimuli into categories based on imperfect feedback (see Figure 6(a)). Patients with Parkinson's or Huntington's disease are impaired at this form of learning (Figure 6(b)), and neuroimaging studies have demonstrated activity in the striatum and dopaminergic midbrain areas during classification learning in normal persons (Figure 6(c)). This form of learning is often referred to as habit learning

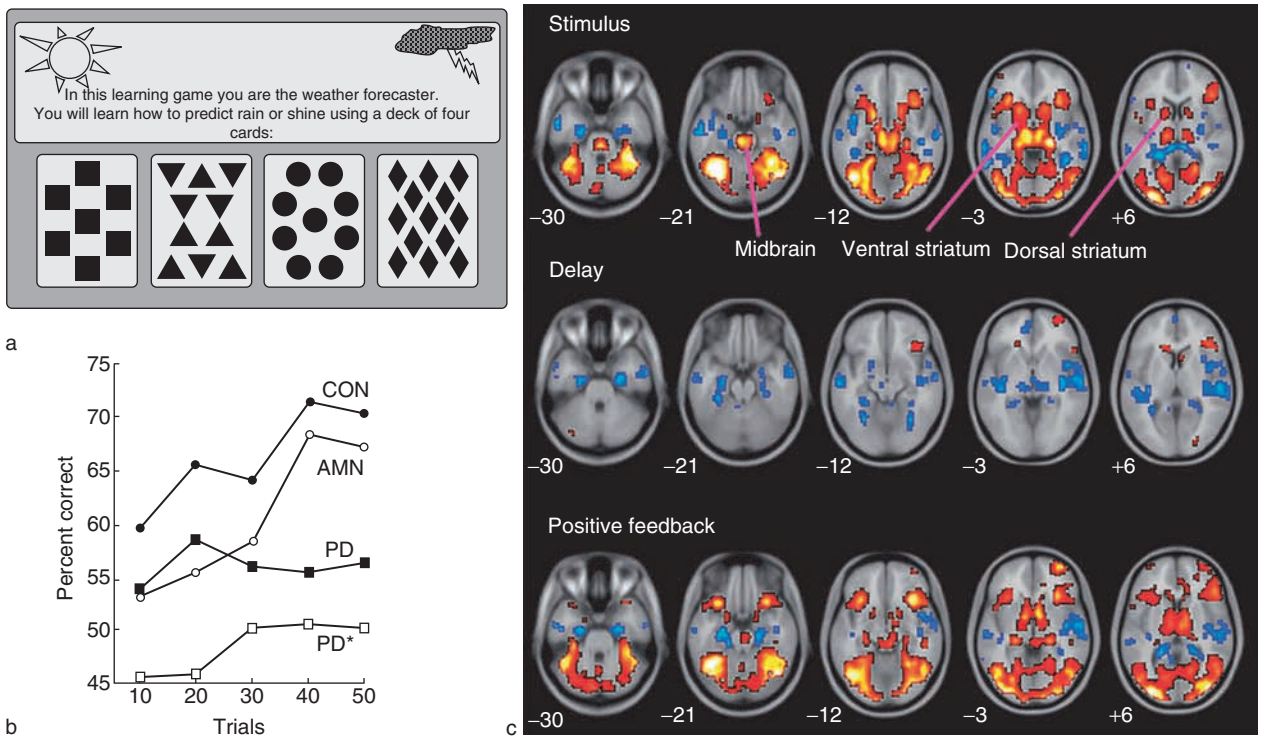

Figure 6 Probabilistic classification task. (a) The task involves the presentation of cards, which the person classifies as either predicting rain or sunshine. The stimulus is similar to a set of playing cards, and the feedback indicates a rain or sunshine outcome. (b) Patients with Parkinson's disease (PD) are impaired at learning the probabilistic classification task relative to controls (CON), particularly for advanced PD (PD^{+}), whereas patients with amnesia (AMN) due to medial temporal lobe damage have little or no impairment. It has been argued that the impairment in PD results from dysfunction of the basal ganglia. (c) Activation of basal ganglia and midbrain structures (putatively the substantia nigra and ventral tegmental area) during different phases of probabilistic classification in normal controls. Functional images are shown in axial format, starting at –30 mm beneath the anterior commissure–posterior commissure line, and rising dorsally to +6 mm above it. The stimulus (top row) and the feedback (bottom row) activate midbrain, as well as ventral and dorsal regions of the striatum, compared to a no-task baseline. This finding is consistent with the proposed role of the nigrostriatal system in mediating stimulus–response learning. The delay period (middle row) between the person's response and the outcome is associated with much less activation. Blue denotes 'deactivations' (i.e., regions of the brain where the functional magnetic resonance imaging signal is less for stimulus, delay, or feedback, compared to baseline); red/orange regions indicate the reverse. (b) Reproduced from Knowlton BJ, Mangels JA, and Squire LR (1996) A neostriatal habit learning system in humans. *Science* 273: 1399–1402, with permission. (c) Adapted from Aron AR, Shohamy D, Clark J, et al. (2004) Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *Journal of Neurophysiology* 92: 1144–1152, with permission.

because it involves the gradual acquisition of stimulus–response associations without the necessity of conscious memory: this kind of probabilistic classification learning occurs normally in amnesic patients, who cannot consciously remember the details of the task.

As discussed in the previous section for stimulus–response learning, one important role of the basal ganglia in classification learning involves the processing of feedback. Neuroimaging studies have shown that learning with feedback more strongly engages the striatum than does learning without feedback. Patients with Parkinson's disease, who are impaired at learning the task with feedback, can learn the same information normally if they do not need to process feedback to do so. Thus, the role of the basal ganglia in some forms of learning may relate to a more general role in the integration of behaviorally relevant feedback with response selection and adjustments to ongoing behavior.

Learning is associated with changes in activity in medium spiny projection neurons (MSNs) and tonically active cholinergic interneurons (TANs) in the striatum. MSNs show changes as rats learn to run a maze. MSNs shift from firing throughout the behavior to firing selectively at the beginning and end of the behavior sequence. These patterns are also sensitive to extinction and reacquisition of the maze-running skill, and their firing is highly predictive of the animal's behavior. In primates, the observed learning-related changes in the putamen, as well as in the prefrontal cortex, closely mirror behavioral learning curves. In the caudate nucleus, similar changes appear earlier during the learning process, when cells appear to code for stimulus–response contingencies before they are fully reflected in behavior. These results are consistent with lesion studies which suggest that the caudate is responsible for learning of new stimulus–response relations (perhaps by quickly coding the reward associations of individual places or responses to those places), whereas the putamen is responsible for the learned expression of these relations. TANs in the striatum appear to be sensitive to the motivational relevance of stimuli regardless of their valence (appetitive or aversive).

The range of findings suggests that the striatum may in some cases compete with the medial temporal lobe to determine behavior. For example, in a cross-maze task, rats must learn to run from a particular starting point to a particular arm of the maze in order to receive food. The rat's response from a different starting point in the maze can be used to probe how it learns the task; early in training, rats will go to the same place in the maze (the place-learning strategy), which is dependent upon the hippocampus, whereas later in training they will make the same movements (a response-learning strategy), which is dependent upon the striatum. However, inactivation of the hippocampus or stimulation of the striatum early in training leads to the use of a response-learning strategy, whereas inactivation of the striatum or stimulation of the hippocampus late in training causes the animal to revert to a place-learning strategy. Thus, both systems appear to be able to drive behavior, and performance is determined by the relative balance of activity between the two systems. An alternative view is that early in training rats use external cues, such as those from vision, to determine the distance and direction to their next goal, whereas later in learning they use internal signals. These internal signals consist of efference copy, which matches the motor command that produces the response, and proprioceptive feedback that tells the motor system whether the current response is proceeding as planned. Thus, the place-learning strategy might be better construed as an exteroceptive strategy, and the response-learning strategy as an interoceptive one. This formulation agrees with the idea, expressed previously, that the basal ganglia's core function may involve a general role in the integration of behaviorally relevant feedback with response selection. Along these lines, functional imaging studies in human study participants show that activity in the striatum and medial temporal lobe correlates negatively across participants and across learning trials. Further, dysfunction of the basal ganglia is associated with a general increase in the engagement of the medial temporal lobe during the learning of tasks that are normally dependent upon the striatum. The mechanisms by which this competition occurs are not yet understood, but they could rely upon biasing signals from the prefrontal cortex (**Figure 4(a)**).

Synthesis

The basal ganglia play a nearly ubiquitous role in cognition. Their contribution can perhaps best be appreciated by considering them as a device that resolves the selection demands that confront a behaving organism. These demands require prioritizing, scheduling, planning, sequencing, and generally controlling the way in which context elicits behavior in accordance with external and internal constraints. Which behavior occurs depends on many factors: top-down biases from the prefrontal cortex that prime particular response channels, the salience and prior rewarding properties of stimuli, the prepotency of various alternatives, and the degree of competition from other cognitive systems. By integrating different information through the interconnections among parallel cortical–basal ganglia loops, moderating this flow

according to the various pathways that influence basal ganglia output, and incorporating signals from neurotransmitter systems that monitor outcomes in biologically relevant terms, the basal ganglia can control systems of thought as well as action.

See also: Appetitive Systems: Amygdala and Striatum; Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia and Oculomotor Control; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopamine: Cellular Actions; Executive Function and Higher-Order Cognition: Assessment in Animals; Motor Skill Learning; Procedural Learning: Striatum; Reward Systems: Human; Reward Processing: Human Imaging; Task Switching.

Further Reading

Aron AR and Poldrack RA (2006) Cortical and subcortical contributions to Stop signal response inhibition: Role of the subthalamic nucleus. *Journal of Neuroscience* 26: 2424–2433.

Aron AR, Shohamy D, Clark J, et al. (2004) Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *Journal of Neurophysiology* 92: 1144–1152.

Barnes TD, Kubota Y, Hu D, et al. (2005) Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature* 437: 1158–1161.

Brasted PJ and Wise SP (2004) Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum. *European Journal of Neuroscience* 19: 721–740.

Hikosaka O, Nakamura K, and Nakahara H (2006) Basal ganglia orient eyes to reward. *Journal of Neurophysiology* 95: 567–584.

Jog MS, Kubota Y, Connolly CI, et al. (1999) Building neural representations of habits. *Science* 286: 1745–1749.

Knowlton BJ, Mangels JA, and Squire LR (1996) A neostriatal habit learning system in humans. *Science* 273: 1399–1402.

Lawrence A, Sahakian BJ, and Robbins TW (1998) Cognitive functions and corticostriatal circuits: Insights from Huntington's disease. *Trends in Cognitive Sciences* 2: 379–387.

Mink JW (1996) The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology* 50: 381–425.

Nambu A, Tokuno H, and Takada M (2002) Functional significance of the cortico-subthalamo-pallidal 'hyperdirect' pathway. *Neuroscience Research* 43: 111–117.

Packard MG (1999) Glutamate infused posttraining into the hippocampus or caudate-putamen differentially strengthens place and response learning. *Proceedings of the National Academy of Sciences of the United States of America* 96: 12881–12886.

Pasupathy A and Miller EK (2005) Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature* 433: 873–876.

Poldrack RA, Sabb FW, Foerde K, et al. (2005) The neural correlates of motor skill automaticity. *Journal of Neuroscience* 25: 5356–5364.

Redgrave P, Prescott TJ, and Gurney K (1999) The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience* 89: 10009–10023.

Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

Shohamy D, Myers CE, Grossman S, et al. (2004) Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology. *Brain* 127: 851–859.

Wickens JR, Reynolds JN, and Hyland BI (2003) Neural mechanisms of reward-related motor learning. *Current Opinion in Neurobiology* 13: 685–690.

Relevant Websites

http://web.mit.edu – Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research.

http://en.wikipedia.org – Wikipedia: The Free Encyclopedia.

Cognition: Cerebellum Role

J E Desmond and C L Marvel, The Johns Hopkins Medical Institutions, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The cerebellum has been historically linked to the control of posture, gait, and skilled voluntary movement. Yet, numerous cases have been reported – dating back to the early 1800s – of nonmotor deficits, such as mental impairment, in patients with cerebellar degeneration and atrophy. Despite these reports, a distinct bias persists among many researchers and physicians that cerebellar-related cognitive deficits are attributed to some type of motor response or to lesion encroachment onto other brain regions. Indeed, some of the earliest functional neuroimaging studies used cerebellar metabolism as a baseline from which neocortical cognitive activation was compared because it was assumed that the cerebellum was silent during cognition.

A new wave of interest in nonmotor functions of the cerebellum began to emerge in the early 1990s. With functional neuroimaging in the hands of a rapidly increasing number of neuroscientists, increasingly more cerebellar activations that were difficult to explain in terms of simple motor processes were reported, and these results were seen to converge with results reported in the patient literature. This article reviews some of the relevant findings supporting a role for the cerebellum in executive function. Executive function refers to some of the highest levels of cognitive control, including planning, organization, decision making, problem solving, and logical analysis. Executive function is typically required when information needs to be coordinated or manipulated, or when prepotent or habitual responses need to be inhibited. Transynaptic neuroanatomical tracing studies have provided convincing evidence that lateral and posterior regions of the cerebellar cortex project through ventral portions of the dentate nucleus and the thalamus to connect with dorsolateral prefrontal regions implicated in executive control. Thus, an anatomical foundation exists by which cerebro-cerebellar networks can influence higher level cognition.

Functional Neuroimaging Studies

Functional neuroimaging methods such as positron emission tomography and functional magnetic resonance imaging (fMRI) are sensitive to regional metabolic changes in blood flow that occur as a result of increased neuronal activity. As summarized here, neuroimaging has revealed cerebellar activation in many investigations that focus on executive function.

Verbal Working Memory

Cerebellar activations have been consistently observed by a number of researchers in verbal working memory tasks. In their simplest form, these tasks require the online maintenance of verbal information via a 'phonological loop' – that is, the refreshment of phonological information through repeated articulatory rehearsal. For example, keeping a phone number in mind long enough to dial a telephone would typically require verbal working memory. Manipulation of the stored information is hypothesized to engage 'central executive' processes of working memory. For example, the *n*-back task entails sequentially presenting letters to the subject, whose task is to decide if the present letter matches the one that occurred *n* letters ago. Comparisons of activation in the *n*-back task relative to those observed in more maintenance-based tasks have shown that although the cerebellum responds to both kinds of working memory tasks, more lateral cerebellar activations are observed for the manipulation task, in conjunction with greater dorsolateral prefrontal activation.

Verbal Fluency

Verbal fluency is a frequently used measure of executive function. Common neuropsychological testing involves asking a subject to say as many words as possible in 60 s that begin with a specific letter. Usually, this procedure involves organizing output into meaningful clusters in order to speed responses. For example, if prompted with the letter S, one may first try to think of S words that begin with the letters S–E (set, settee, sell, self, seep, etc.). When neural activations during letter fluency are compared to a control condition (e.g., counting forward from 1), subjects increase activation in the left prefrontal cortex and right cerebellum.

Some of the earliest demonstrations of cognition-related cerebellar activations were reported for a particular type of verbal fluency task referred to as verb generation. This involves the presentation of a noun, such as 'ball,' and the subject responds with an appropriate action verb, such as 'throw.' As a control condition, the noun is simply read and repeated. This and other fluency tasks, such as word stem completion, are associated with robust right posterior cerebellar and left prefrontal activations. Interesting practice-related effects have been associated with

these activations. If subjects are given a second run of the verb-generation task using the same set of nouns that they saw on the first run, left prefrontal and right cerebellar activation diminish concurrently and increase again when novel nouns are presented. These results demonstrate the close functional coupling of the cerebro-cerebellar network during cognitive operations. However, this coupling may reflect complementary rather than redundant functionality. For example, in the stem completion task, word stems with few possible completions (e.g., PSA__) elicit greater cerebellar activation compared to completion of stems with many possible completions (e.g., STA__), whereas the reverse is true for activation in left prefrontal regions. One possible explanation for this dissociation is that cerebellar activation during word stem completion is related to word search operations, whereas prefrontal activation is related to word selection.

Cognitive Flexibility

A popular neuropsychological test for assessing the ability to shift cognitive sets according to changes in stimulus contingencies is the Wisconsin Card Sorting Test (WCST). In the WCST, subjects are required to sort cards according to one of three possible dimensions: number, color, or shape. The subject must determine the correct sorting strategy on the basis of right/wrong feedback that is presented after the subject's response to each card. The experimenter can change the correct sorting strategy suddenly without warning, and the subject must then determine the new dimension for sorting. Impaired performance on the task is observed as a tendency to perseverate on the incorrect strategy in light of new information that contingencies have changed. As a neuropsychological test, the WCST appears to be sensitive to frontal lobe dysfunction, but the specificity of the test to the frontal lobes has been questioned. Indeed, neuroimaging studies have revealed a large network of cortical and subcortical activations, including bilateral cerebellum, when contrasting WCST performance with sensorimotor control tasks. Such control tasks effectively rule out simple visual- or response-elicited explanations for cerebellar activation, but they do not address whether the cerebellum may be responding to the nonexecutive working memory requirements of the task, such as the need to keep the current relevant dimension of the stimuli in mind. Recent investigations, however, have tried to separate higher order from lower order cognitive operations in the task. For example, by contrasting the WCST task with a task that cues the subject when a dimension change occurs, simple working memory maintenance operations are presumably subtracted out, and what remains are the most executive components of the task. Cerebellar activations have been shown to persist when such subtractions are performed.

Reasoning/Planning

A number of different studies have examined the neural substrates of higher order skills used in hypothesis testing and logical inference. These investigations have applied one of three main approaches. The first requires the subject to solve a puzzle, with the control condition attempting to equate the sensorimotor aspects of manipulating puzzle pieces in the absence of planning. The second requires the subject to form a hypothesis about features presented within a group of multidimensional stimuli (e.g., containing letters, numbers, shapes, and colors or composed of a large grid of squares, a subset of which are filled, whereas others are unfilled). On each trial, the subject determines which features of the stimuli are relevant or correct. Control conditions typically involve presenting identical stimuli but asking the subject to attend and respond to only one dimension. The third approach requires subjects to decide whether three-sentence syllogisms, presented with either semantic or nonsemantic content, are logically valid. An effective control condition for this type of study replicates much or all of the working memory requirements of the reasoning task. It entails reading three sentences, the first two of which are plausible premises, whereas the third contains unrelated content and therefore requires no reasoning. For each of these approaches, prominent cerebellar activation has been observed, suggesting a role for the cerebellum in reasoning and planning.

Dual Tasks

Patients with executive dysfunction often exhibit profound difficulties performing multiple tasks at once. In healthy subjects, performance typically declines for both tasks relative to performance in either task alone, but subjects are nevertheless able to perform both tasks to some degree. Dual-task performance has thus been used by neuroimagers to investigate whether there are specific neural correlates of cognitive coordination. Of particular interest are tasks that show unique regions of activation during dual-task performance relative to single-task performance. Such studies have shown that in cases in which neither prefrontal nor cerebellar activation are evident for either task alone, dual-task performance recruits new prefrontal, parietal, and cerebellar regions.

Random Number Generation

Random number generation (RNG) requires a subject to randomly generate numbers, usually 1–10, while suppressing the natural tendency to count in sequence or to repeat the same number habitually. Thus, RNG involves both fluency and inhibition. When subjects are forced to generate random numbers at increasing rates, numbers tend to become less random and more sequential. Neuroimaging investigations have thus used pacing stimuli to control the rate of RNG and have typically contrasted activation for this task with that obtained from counting in sequence. Medial portions of the cerebellum have been shown to increase activation with increasing rate of performance, consistent with either motoric or timing functions. However, more lateral portions of the cerebellar hemispheres have exhibited diminution of activation at faster rates of RNG compared to serial counting. These occur in parallel with similar diminutions seen in dorsolateral prefrontal and parietal regions. These results suggest that a cortical–subcortical executive network is involved in suppressing habitual responses and that at increasing rates, network activity and behavioral performance begin to break down.

Theory of Mind

The ability to infer intention or mental states in others is an important aspect of human social interaction and has been referred to as possessing a 'theory of mind.' Both verbal and nonverbal approaches have been used in neuroimaging investigations of theory of mind. An example of a nonverbal approach would be to present a subject with frames of a cartoon and ask the subject to choose an ending frame from two or three choices using a button press. For the theory of mind condition, the choice of the ending requires inferring the intention of the character depicted in the cartoon, whereas the control condition requires understanding physical causality but does not involve making inferences of motives. An example of a verbally based approach would be to ask subjects in the theory of mind condition to read a brief story describing the actions of a character and then to make up a story explaining the conditions that led to the character's behavior. As a control condition, subjects read aloud a story that requires no attribution of intentions, thereby equating the same amounts of reading and word production produced by the theory of mind task. In both approaches, strong cerebellar activation has been observed in the theory of mind condition relative to the control condition. For the nonverbal tasks, left cerebellar/right prefrontal activation has been observed, whereas the laterality of frontocerebellar activation is reversed for the verbally based approach.

Patient Studies

Evidence for cerebellar contribution to executive function can be drawn from neuroimaging and neuropsychological studies of cerebellar-affected clinical populations. In these cases, the cerebellum may be disrupted early on as a result of maldevelopment or during adulthood following years of normal function.

Neurological Disorders

Compelling evidence for the cerebellum's participation in executive function comes from studies of patients with neurological disorders confined to the cerebellum. These data have been derived from neuropsychological studies of patients with cerebellar stroke, tumor resection, cerebellar degeneration, and cerebellitis. In general, executive dysfunction seems to be most pronounced in cerebellar patients with large bilateral infarcts, followed by those with unilateral infarcts, and least in those with pancerebellar involvement. However, deficits are not always evident or as severe as might be expected given the size of the lesion. Heterogeneity of lesion type and time since lesion onset are two key factors that affect a patient's degree of impairment. Improved function has been noted in those with nonprogressive lesion types (e.g., stroke) after several months of recovery.

Relatively few studies have directly addressed the cognitive contributions of the cerebellum in clinical populations. This is due, in part, to the difficulty in acquiring patients with diseases confined to the cerebellum with an absence of disease elsewhere in the brain. However, one seminal study, conducted by Schmahmann and Sherman, tested 20 patients with cerebellar diseases on a variety of neuropsychological tests that involved executive skills. Patients were most impaired in working memory manipulation, verbal fluency, and complex figure copying. During complex figure copying, one is presented with a drawn design and asked to copy it while the figure is in plain view. Although this may appear to be a straightforward task, patients have difficulty copying the figure if they cannot develop a drawing strategy, such as where to begin. They may also be unable to organize and maintain a long-term plan for drawing so that the design is redrawn in sections. There may be a focus on details without consideration for the overall configuration. This results in a drawing of several geometric

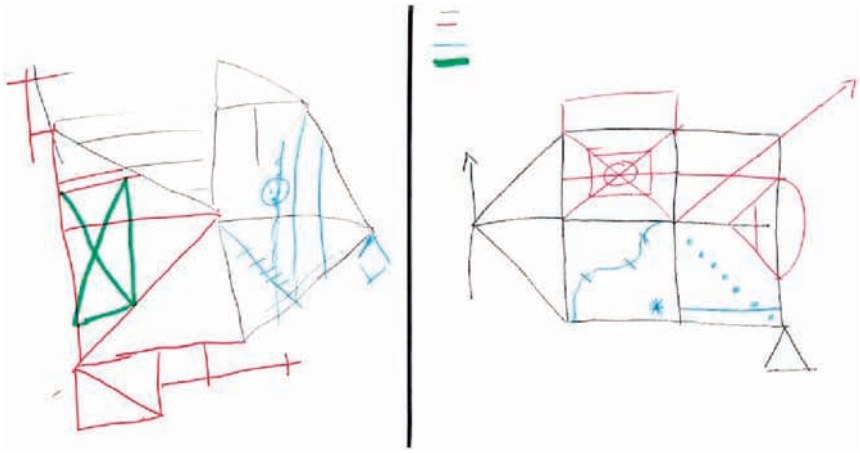

Figure 1 Copying complex figures relies on executive skills, such as planning, organizing, and maintaining a strategy for sequential actions. In these drawings, a copy of the Rey–Osterreith complex figure (ROCF; left) was drawn by a patient with a symptomatic postcerebellitis infection. The same patient copied the Taylor complex figure (TCF; right) 3 months later. Using the colored pencil lines as a key to the sequence of the patient's actions, one can see that the patient applied a part-based approach to the ROCF. She drew the figure piecemeal, completing the major sections of the right side first (black). Next, she drew the general composition of the left side (red) and then went back to fill in details on the right side (blue). Finally, she returned to the left side to add a few more details to complete the drawing (green). The final construction is awkward in design but includes all relevant pieces of the original figure. In sharp contrast, 3 months later the patient applied a highly organized copying strategy that included a configurational approach to the TCF. The patient began by drawing the general outline of the figure (black). She worked her way clockwise by filling in the upper left and right quadrants next (red). Finally, she filled in the bottom right and left quadrants (blue). This efficient strategy allowed her to complete the figure quickly before being asked by the experimenter to draw with the green pencil. In addition to improving her copying speed, the recovered patient drew neater, stronger lines that rendered a highly accurate depiction of the complex figure. Reproduced from Schmahmann JD and Sherman JC (1998) The cerebellar cognitive affective syndrome. *Brain* 121(Part 4): 561–579, by permission of Oxford University Press.

shapes that are not well integrated into a whole. Often, figure parts are misplaced or missing altogether. Without intact skills for creating and implementing a copying strategy, a patient can eventually produce a rough version of the figure but inefficiently and usually with errors. To illustrate such deficits, Figure 1 shows two copies of complex figures drawn by a patient during the symptomatic phase of a postcerebellitis infection (left, Rey–Osterreith complex figure) and by the same patient 3 months later during recovery (right, Taylor complex figure). It is clear that during the symptomatic phase of the illness the patient was unable to appreciate the overall configuration, which resulted in an inefficient part-based approach to figure copying. After recovery, the patient was able to rely on her improved executive skills to use a configurational approach, which yielded a highly organized and accurate replication.

Although patients with cerebellar disease may exhibit executive dysfunction, the current opinion posits that disrupted communication between the cerebellum and the frontal lobe contributes to these deficits. A breakdown in the frontocerebellar network may result in an inability of the cerebellum to contribute to the frontal lobe's temporal organization of mental processes. Consequently, an otherwise coordinated sequence of cognitive events unravels, leading to a phenotype of cognitive impairment. The term dysmetria of thought has been given to describe this phenomenon in patients with cerebellar disease.

Developmental Disorders

There are a number of developmental disorders that involve the cerebellum. However, maldevelopment is not typically confined to the cerebellum and often influences additional areas of the brain. With increasing popularity in the use of neuroimaging tools, however, researchers have been able to associate structural and functional abnormalities of the cerebellum with executive dysfunction in several developmental disorders, as summarized here.

Attention-deficit/hyperactivity disorder Attention-deficit/hyperactivity disorder (ADHD) is characterized by inattention and hyperactive or impulsive behavior. Executive dysfunction is a prominent feature of ADHD and may be related to an underlying core disturbance in response inhibition. A fairly consistent finding in ADHD brain research is reduction in total cerebellar size, most markedly in the posterior vermis (lobes VIII–X). Although nearly all brain imaging research in ADHD has been conducted

in children and adolescents, evidence suggests that cerebellar abnormalities are not progressive and are not related to medication treatment. Studies using fMRI in ADHD have reported decreased cerebellar activation during verbal working memory and response inhibition. However, the severity and consistency of cerebellar-related cognitive impairments in ADHD has not been firmly established. For example, several fMRI studies of executive-type tasks in ADHD children have implicated frontal–striatal rather than frontocerebellar pathways. It is possible that cerebellar disturbances in ADHD disrupt executive function through miscommunications with a larger neural network that includes the frontal lobe and basal ganglia.

Autism spectrum disorder Autism spectrum disorder (ASD) is marked by impaired interpersonal relations, poor communication skills, and reduced or repetitive interests and behaviors. It is believed that poor cognitive planning, flexibility, and inhibition underlie the social and behavioral abnormalities associated with ASD. The most consistent postmortem finding in ASD is decreased cerebellar Purkinje neurons in the absence of neurodegeneration. Genetic disturbances in cerebellar development may lead to a disruption in normal cerebellar functions and contribute to the onset of autistic symptoms. fMRI studies that target cerebellar functions in autism are beginning to emerge in the literature. For example, decreases in cerebellar activation have been reported in ASD subjects while observing (but not while imitating) emotional expressions. Attention-related decreases in cerebellar activation also have been reported. Cognitive and social deficits that are inherent to ASD may be attributed to maldevelopment of the cerebellum and/or its communication with other brain structures. However, there is an obvious need for future studies on this topic in order to better characterize the role of the cerebellum in autism.

Dyslexia Dyslexia is a developmental disorder identified in children who fail to attain reading and writing skills commensurate with their level of education and intelligence. Poor motor coordination and delayed motor milestones have also been noted in dyslexic children. Although there are several hypotheses regarding the fundamental disturbance in dyslexia, one view is that dyslexics have cerebellar-related difficulties with automatization of skills. Automaticity – the ability to perform a skill fluently and without conscious effort in doing so – is essential for multitasking. Impaired automaticity in dyslexia may contribute to difficulties in phonological processing, word recognition, verbal working memory, and, ultimately, fluent reading. It is generally accepted that the cerebellum plays a key role in automatization, and that cerebellar damage can force one to pay extra attention to skill sequences that never become automatic. Accordingly, dyslexics have difficulties performing tasks that require the blending of two or more skills. For example, balancing ability (e.g., standing on one foot) becomes disproportionately affected when dyslexics perform a counting task at the same time. Cerebellar structural abnormalities, such as decreased anterior lobe volume, have been reported in this population. Moreover, these abnormalities have been correlated with cognitive tests of verbal ability, such as spelling, word identification, and rapid letter naming. Although there is persuasive evidence to support cerebellar-related cognitive deficits in dyslexia, there has not been full agreement on this in the literature. Notably, some researchers maintain that the cerebellum is only secondarily related to dyslexic cognitive impairments, with primary deficits attributed to cortical areas. However, the cerebellar hypothesis provides an interesting and parsimonious explanation for both the motor and the cognitive deficits observed in this disorder.

Schizophrenia Schizophrenia is a psychiatric disorder that involves an array of perceptual, emotional, motor, and cognitive impairments. Cognitive deficits are widespread but are found in several important components of executive function, such as attention, working memory, sequencing, and problem solving. Data supporting cerebellar-related cognitive disturbances in schizophrenia have come from functional neuroimaging studies of working memory, verbal recall, and social cognition. In addition, structural imaging studies have found decreased cerebellar volume, especially of the vermis, although some reports have also noted increased cerebellar volume or normal cerebellar volume. Perhaps the most compelling study to date on this topic, conducted by Ho and colleagues, examined 155 neuroleptic-naive schizophrenia patients on a comprehensive cognitive battery. In addition, patients were examined for cerebellar soft signs (motor impairments referable to the cerebellum), including unsteady tandem gait, intention tremor, poor balance, and flaccid muscle tone. Thirty-two patients were identified as having cerebellar soft signs. Relative to the 123 patients without cerebellar soft signs, these 32 patients exhibited specific deficits of attention, working memory, problem solving, and figure copying in addition to having overall smaller cerebellar volume. These results suggest that decreased cerebellar volume in

schizophrenia is related to impairments of motor coordination and higher order cognitive skills. Given that cerebellar abnormalities have been reported in many, but not all, regional brain volume studies of schizophrenia, this phenomenon may apply to only a subset of patients.

Individuals born very preterm Children born very preterm (i.e., <33 weeks of gestation) often show signs of putative cerebellar dysfunction, such as clumsiness and poor motor coordination. Although neuroimaging studies are only beginning to emerge on this topic, data suggest that total cerebellar volume is decreased in these individuals. Moreover, volumetric decreases in lateral cerebellar hemispheres have correlated with performance on tests of abstract reasoning and visuospatial construction. Functional neuroimaging of adolescents born preterm has revealed bilateral reductions in cerebellar activity during response inhibition, despite normal task performance. Because the pre-, peri-, and neonatal periods represent a vulnerable time for brain formation, individuals who are unable to complete the full gestation (approximately 38 weeks) run the risk of maldevelopment. This risk is not specific to the cerebellum, and indeed, abnormalities have been reported in multiple cortical and subcortical brain areas in this population. Therefore, cerebellar contributions to cognitive and motor deficits observed in preterm individuals may be further attributed to underlying factors, such as hypoxia, sepsis, or malnutrition around the time of birth.

Substance Abuse

A recent view on substance abuse holds that executive dysfunction is responsible for the loss of self-control that leads to drug and alcohol addiction. Evidence implicating the cerebellum in substance abuse suggests that the cerebellum may contribute to cognitive deficits (e.g., poor inhibition) that segue into drug-seeking behavior. Structural brain imaging of alcoholics has revealed volume reductions throughout the cerebellum, particularly in the anterior vermis. Similar cerebellar abnormalities have been reported in children exposed to high prenatal doses of alcohol. Functional neuroimaging has shown that activity in the left anterior cerebellum and cerebellar vermis increases during alcohol craving. Moreover, alcoholics have demonstrated excessive right anterior cerebellar activation while performing a task of verbal working memory. Studies have also shown that abnormal cerebellar blood flow is associated with the use of cocaine, marijuana, methylphenidate (Ritalyn), 3,4-methylenedioxymethampetamine (MDMA, or ecstasy), and amphetamine. For example, blood flow to the cerebellar vermis increases in cocaine users when viewing cocaine-related cues. Over-responsive activity in the left cerebellum of cocaine users has been associated with high demands on working memory and self-inhibition. A similar pattern of excessive left cerebellar activity has been observed in heavy marijuana users while performing a complex decision-making task. In nearly all cases, findings from the structural and functional imaging studies described previously have implicated the prefrontal cortex in conjunction with the cerebellum. This pattern suggests that a disrupted frontocerebellar circuit contributes to drug and alcohol addiction. However, specifying the contributions of the cerebellum to substance dependence first requires a deeper understanding of the cerebellum's role in cognition.

Cerebro-Cerebellar Interactions

Most complex behaviors require a response of some kind, and this inevitably invokes motor systems. Given the involvement of the cerebellum in motor control, it may be difficult to completely separate cognitive- and motor-related cerebellar functions. However, the results of Hulsmann et al. illustrate how cognitive and motor involvement of the cerebellum falls along a continuum, reflecting progressive activation of cerebro-cerebellar closed loops, rather than a dichotic absolute. Figure 2 shows that for a self-paced finger response (button press), activations occurring close in time to the finger response, and soon thereafter, are observed in primary sensorimotor cortex as well as in medial spinocerebellar regions. In contrast, looking 3 or 4 s prior to the response, activations are observed in neocortical areas associated with motor planning, such as supplementary and premotor cortex, as well as in more lateral regions of the neocerebellum. This example is derived from a relatively simple behavior. However, it is not difficult to extrapolate from these results and hypothesize that as higher order planning is required – occurring even more remotely in time from the final motor response – neocortical regions, such as parietal and dorsolateral prefrontal regions, are in communication with the most lateral and posterior regions of the cerebellum during executive planning operations. As described in the introduction, anatomical studies have demonstrated the existence of these closed-looped cerebro-cerebellar circuits. The successive activation of these circuits during the time course of simple and complex behaviors gives us a beginning framework for understanding why the cerebellum is active in so many functions, from motor control to cognition.

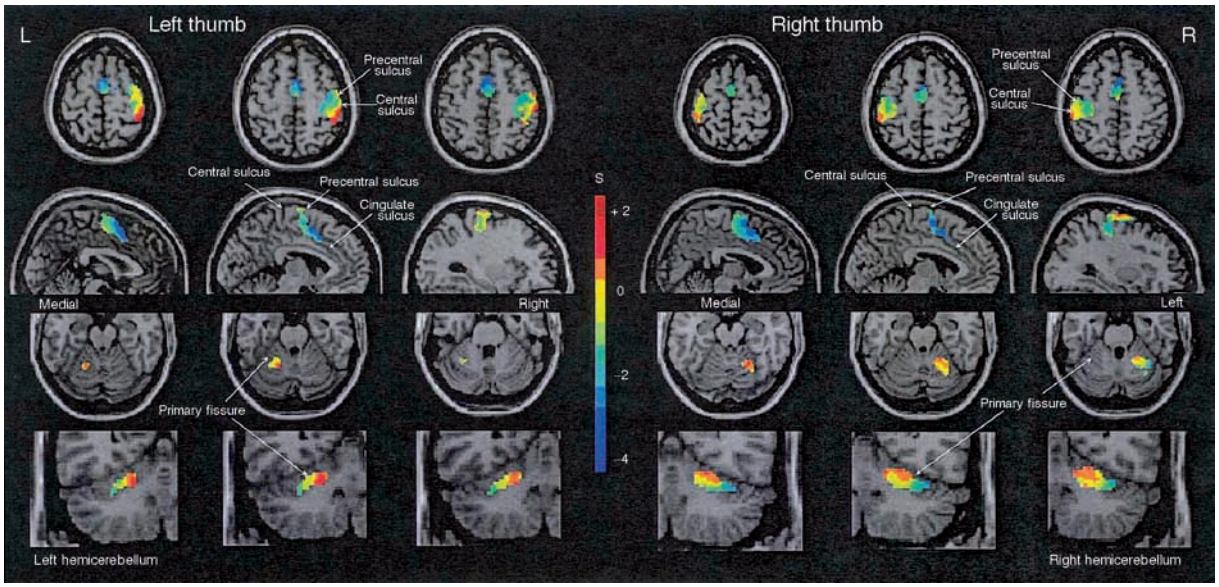

Figure 2 Color-coded map depicting the time course of maximum activation relative to a self-paced thumb movement. Earliest activations prior to movement are depicted in blue and green and occur in cingulate motor areas, supplementary motor, and premotor areas, as well in the lateral neocerebellar regions. Activations occurring at the time of movement and slightly thereafter are observed in primary sensory and motor cortex as well as in medial spinocerebellar regions. The sequential progression of regional activations during this simple motor behavior demonstrates how cerebro–cerebellar interactions could progress in similar closed-loop circuits during executive function. Reproduced from Hulsmann E, Erb M, and Grodd W (2003) From will to action: Sequential cerebellar contributions to voluntary movement. *NeuroImage* 20(3): 1485–1492, with permission from Elsevier.

See also: Cerebellum and Oculomotor Control; Cerebellum: Clinical Pathology; Cerebellum: Evolution and Comparative Anatomy; Cognitive Deficits in Schizophrenia; Executive Function and Higher-Order Cognition: Neuroimaging; Executive Function and Higher-Order Cognition: Definition and Neural Substrates; Frontal Lobe Syndrome; Procedural Learning: Cerebellum Models; Reasoning and Problem Solving: Models; Short Term and Working Memory; Substance Abuse and Dependence; Synaptic Plasticity: Cerebellum; Task Switching.

Further Reading

Calarge C, Andreasen NC, and O'Leary DS (2003) Visualizing how one brain understands another: A PET study of theory of mind. *American Journal of Psychiatry* 160(11): 1954–1964.

Chen SH and Desmond JE (2005) Temporal dynamics of cerebro-cerebellar network recruitment during a cognitive task. *Neuropsychologia* 43(9): 1227–1237.

Collette F, Olivier L, Van der Linden M, et al. (2005) Involvement of both prefrontal and inferior parietal cortex in dual-task performance. *Brain Research. Cognitive Brain Research* 24(2): 237–251.

Daniels C, Witt K, Wolff S, Jansen O, and Deuschl G (2003) Rate dependency of the human cortical network subserving executive functions during generation of random number series – A functional magnetic resonance imaging study. *Neuroscience Letters* 345(1): 25–28.

Dapretto M, Davies MS, Pfeifer JH, et al. (2006) Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* 9(1): 28–30.

Desmond JE, Gabrieli JDE, and Glover GH (1998) Dissociation of frontal and cerebellar activity in a cognitive task: Evidence for a distinction between selection and search. *NeuroImage* 7: 368–376.

Goel V, Buchel C, Frith C, and Dolan RJ (2000) Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage* 12 (5): 504–514.

Ho BC, Mola C, and Andreasen NC (2004) Cerebellar dysfunction in neuroleptic naive schizophrenia patients: Clinical, cognitive, and neuroanatomic correlates of cerebellar neurologic signs. *Biological Psychiatry* 55(12): 1146–1153.

Hulsmann E, Erb M, and Grodd W (2003) From will to action: Sequential cerebellar contributions to voluntary movement. *NeuroImage* 20(3): 1485–1492.

Lie CH, Specht K, Marshall JC, and Fink GR (2006) Using fMRI to decompose the neural processes underlying the Wisconsin Card Sorting Test. *NeuroImage* 30(3): 1038–1049.

Nicolson RI and Fawcett AJ (2005) Developmental dyslexia, learning and the cerebellum. *Journal of Neural Transmission Supplementum* 69: 19–36.

Nosarti C, Rubia K, Smith AB, et al. (2006) Altered functional neuroanatomy of response inhibition in adolescent males who were born very preterm. *Developmental Medicine and Child Neurology* 48(4): 265–271.

Schmahmann JD and Sherman JC (1998) The cerebellar cognitive affective syndrome. *Brain* 121(Part 4): 561–579.

Sullivan EV, Harding AJ, Pentney R, et al. (2003) Disruption of frontocerebellar circuitry and function in alcoholism. *Alcoholism, Clinical and Experimental Research* 27(2): 301–309.

Valera EM, Faraone SV, Biederman J, Poldrack RA, and Seidman LJ (2005) Functional neuroanatomy of working memory in adults with attention-deficit/hyperactivity disorder. *Biological Psychiatry* 57(5): 439–447.

Cognition: Neuropharmacology

A B Hains and A F T Arnsten, Yale Medical School, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

The prefrontal cortex (PFC) is instrumental in coordinating, controlling, and executing cognitive and emotional processes using appropriate judgment, flexibility, and attention. The ability to organize appropriate responses to a changing environment requires maintenance and updating of relevant information as well as control over the processing of new incoming information. This careful choreography allows the regulation of impulses, language, attention, decision making, and error correction and is commonly referred to as the executive functions. These abilities require the representation of information not currently in the environment (sometimes referred to as working memory) and are mediated by the PFC. The PFC receives ascending inputs from diverse areas of the brain. The dorsal and lateral surfaces of the PFC interconnect with the sensory and motor association cortices and are key for the regulation of behavioral responses and attention; the ventral and medial regions of PFC interconnect with brain regions involved with emotion, such as the amygdala, hypothalamus, and nucleus accumbens. The ventral surface is often referred to as the orbital frontal cortex (OFC) because it sits above the orbits of the eyes. In rodents, the PFC is much smaller; there is a medial portion (the prelimbic and infralimbic mPFC) that is needed for cognitive control and an orbital area more ventral-laterally that has functions similar to the OFC in primates. The homology of the mPFC in rodents to the dorsolateral PFC in primates is controversial. It is likely that it is more homologous to the mPFC in primates and involves a combination of affective and cognitive functions.

Single-unit recording studies in monkeys have shown that PFC neurons are able to hold modality-specific information online over a delay and to use this represented information to guide behavior in the absence of environmental cues. PFC neurons can also fire in relationship to an abstract rule that is used to govern action. A unique feature of PFC neurons is their ability to maintain information in the presence of distracting stimuli. Delay-related firing also can serve as the basis for behavioral inhibition (e.g., having to look away from a remembered visual stimulus or reverse reward contingencies). Neuromodulators can have powerful effects on the patterns of PFC neuronal response and on PFC cognitive functioning.

The PFC is highly sensitive to its neurochemical state. Evidence to date suggests that the dorsolateral PFC is particularly sensitive to catecholamines (dopamine (DA) and norepinephrine (NE)), whereas the ventromedial PFC is more influenced by serotonin (5-HT). This view has emerged from both lesion studies in nonhuman primates that selectively deplete catecholamines (vs. 5-HT) and from pharmacological studies in humans that preferentially block catecholamine (vs. 5-HT) transporters. However, there are critical exceptions to this generalization; for example, NE α2A-adrenoceptor stimulation can improve ventromedial PFC functions, and 5-HT2A receptor blockade can impair dorsolateral PFC function. Acetylcholine (ACh) also has powerful influences on PFC functions, although there has been little research on cholinergic mechanisms in the monkey PFC. Most of the research examining direct cholinergic mechanisms in PFC have been performed in rats using attentional tasks, and thus it is difficult to compare results across paradigms.

It is noteworthy that the PFC also projects back down to the monoamine and ACh cells, and thus it is positioned to regulate its own modulatory input. Interestingly, the NE and DA cells respond to informative cues and receive projections from the dorsolateral PFC, whereas the ACh cells respond to rewards and response choices, and these cells receive inputs from the OFC. Recent studies indicate that NE cells also receive projections from the OFC and anterior cingulate and that these projections may regulate their tonic firing rate rather than the response to specific stimuli. Thus, there is a coordinated interplay between the subregions of the PFC and their modulatory inputs.

Dopamine

The pioneering work of Goldman-Rakic first revealed the powerful influences of DA on the working memory functions of the dorsolateral PFC. This initial study found that depleting catecholamines in PFC was as destructive as removing the cortex itself. Although this first study focused on DA, it is now known that the depletion of both DA and NE may be especially deleterious.

The DA inputs to the PFC arise from the dorsal cells of the substantia nigra and from the ventral tegmental area in the midbrain. DA neurons project throughout the primate cortex, with highest levels in primary and secondary motor cortices. Within the PFC, the medial PFC actually has a denser innervation than the

dorsolateral PFC, despite the established importance of DA to dorsolateral PFC function. Thus, the quantity of innervation cannot be equated with the importance of an input because it is possible that some delicate inputs may be relatively sparse because they are so powerful.

There are two families of DA receptors: the D1 family (D1 and D5) and the D2 family (D2, D3, and D4). There are currently no drugs available that distinguish between D1 and D5 receptors; thus, reference to D1 usually means D1 or D5. The highest levels of DA receptors in the PFC are the D1 family, and these receptors have been the focus of most research on DA mechanisms in the PFC. D1/D5 receptors can be found in both superficial and deep layers of the PFC, with D1 receptors concentrated on spines and D5 receptors on shafts of pyramidal cells. In contrast, the D2 receptor family is more concentrated in layer V, the cells that project to the basal ganglia. Thus, the D2 family may be especially involved with modulating the response output. D4 receptors are especially numerous on γ-aminobutyric acid (GABA)ergic interneurons in the PFC.

Most research on DA influences on PFC cognitive function has focused on spatial working memory. In monkeys, spatial working memory generally has been assessed using an oculomotor or classical delayed response task. This paradigm requires the monkey to remember a spatial location over a brief delay (seconds). The location changes on each trial so that memory must be continuously updated. Electrophysiological studies of PFC neurons record single units from monkeys performing the oculomotor version of this task. Many neurons in PFC exhibit spatially tuned firing during the delay period; that is, the cells fire more for a preferred than for a nonpreferred direction. This is considered the electrophysiological signature of spatial working memory. In rats, spatial working memory is often assessed in a spatial delayed alternation task, within a T-maze, or using automated operant procedures. Delayed nonmatch-to-position is often substituted for delayed alternation, but it is substantially easier due to diminished proactive interference. In humans, spatial working memory is tested in a variety of paradigms, including a search task on the CANTAB battery and N-back tasks developed from the monkey delayed-response task. In addition, human studies often use the Wisconsin Card Sorting task, Stroop Interference task, and Stop-Signal tasks, which emphasize the inhibitory functions of the PFC.

The D1 family of receptors has powerful effects on spatial working memory function. Either too little or too much D1/D5 receptor stimulation impairs working memory; that is, there is an inverted-U dose–response curve that occurs within normal physiological parameters (see Figure 1). Insufficient levels of DA D1 stimulation probably contribute to cognitive deficits in Parkinson's disease and may also contribute to normal age-related cognitive decline. Excessive levels of D1 receptor stimulation occur during stress exposure and contribute to stress-induced PFC dysfunction. This inverted U in working memory abilities has been observed in monkeys, rats, mice, and human subjects. In humans, an inverted U has been observed in relation to the COMT genotype, whereby the substitution of methionine for valine weakens enzymatic degradation of DA and shifts the U curve rightward. An inverted U is also observed in single-unit recordings of monkeys performing a spatial working memory task. D1 receptor stimulation is especially important for suppressing cell responses to nonpreferred spatial directions. These suppressive effects are mediated by the activation of cyclic adenosine monophosphate (cAMP) intracellular signaling. The effects of D1 receptor stimulation probably depend on the endogenous state of the neuron (e.g., whether it is broadly or narrowly tuned) and on the cognitive demands on the subject. This may explain why D1 receptor blockade or DA depletion has little effect on some kinds of executive function such as self-ordered tasks.

The influence of the D2 family is more complex and less well studied. There has been almost no research on D3 receptors mechanisms in the PFC, and drugs which distinguish between D2 and D3 receptors are scarce. Early studies showed that D2/D3 receptor blockade did not impair working memory, but rat studies suggest that D2/D3 receptor stimulation (e.g., during stress exposure) may impair

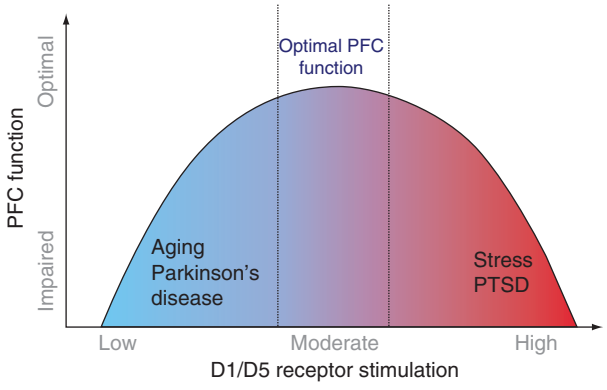

Figure 1 A schematic representation of the inverted-U relationship between D1/D5 receptor stimulation and PFC function; moderate levels of receptor stimulation produce enhanced delay-related firing in PFC neurons as well as optimal performance on working memory tasks, whereas insufficient or excessive D1/D5 receptor stimulation erodes spatial mnemonic tuning and impairs working memory. PFC, prefrontal cortex; PTSD, posttraumatic stress disorder.

working memory performance. Electrophysiological studies show that D2 receptor stimulation increases response-related firing, but not delay-related firing, of PFC neurons in monkeys performing a working memory task. Because some response-related firing occurs after the motor response is initiated, it may act as corollary discharge, informing the brain that a motor command has been initiated. This may have particular relevance to schizophrenia, in which hallucinations may involve weakened corollary discharge.

Recordings from rat PFC slices and from awake monkeys indicate that D4 receptor stimulation inhibits GABAergic interneurons via G_i inhibition of cAMP signaling. Thus, D4 receptor stimulation may have an overall excitatory effect on PFC circuits. However, D4 also appears to inhibit some pyramidal cells, and so the picture may be more complex. Behavioral studies of D4 antagonists may find an inverted-U dose–response curve on working memory, but these agents have not been studied extensively. It should be noted that NE has a higher affinity for D4 receptors than for adrenoceptors and thus that the D4 receptor should properly be considered a catecholamine receptor rather than a DA receptor.

Norepinephrine

It is now appreciated that NE has just as powerful an effect on NE function as DA. In contrast to DA D1, NE appears to dissociate its beneficial and detrimental effects at different types of adrenoceptors. Thus, these may be more practical targets for pharmacological treatment of PFC dysfunction.

The NE innervation of the PFC arises from the locus coeruleus (LC) in the pons. As with DA, there is a delicate innervation of both the supragranular and granular layers, with particular density in layer I. The superficial layers are the focus of cortical–cortical connections and have dense α1- and α2-adrenoceptor binding. There are three α2 subtypes; of these, the α2A subtype is particularly prominent in the monkey PFC. Electron microscopy studies have documented α2A receptors postsynaptically in the spines and dendrites of PFC pyramidal cells, as well as in presynaptic locations on NE axons and terminals. β-Receptor binding is densest in the middle layers (i.e., layer IV), which receives projections from the thalamus.

As with DA, NE has dual effects on PFC function depending on the amount of NE release. Moderate levels of NE release during normal waking have vital, beneficial effects on PFC function via higher-affinity postsynaptic α2A-adrenoceptors, whereas high concentrations of NE released during stress impair PFC function via lower-affinity α1-adrenoceptors and possibly β1-adrenoceptors (see Figure 2). In most parts of brain, NE has potent effects via β-adrenoceptors. In contrast, the PFC seems more dramatically altered by α-adrenoceptors. Figures 3(a) and 3(b) illustrate the

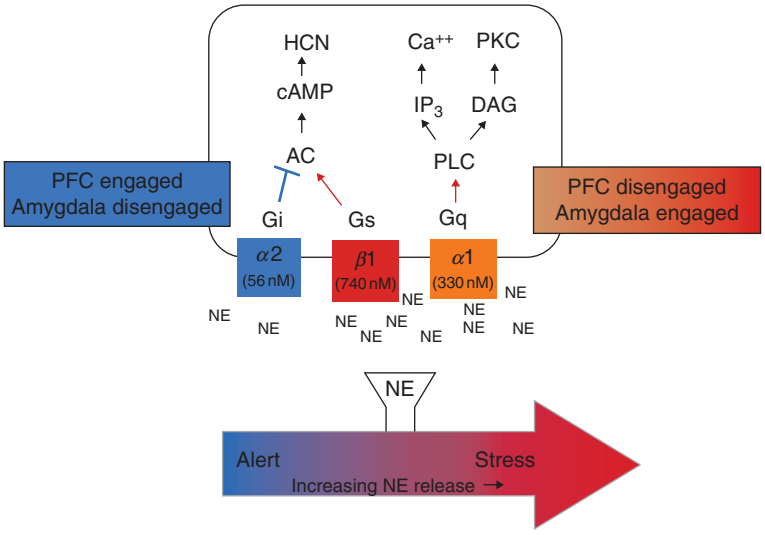

Figure 2 Levels of NE release determine the functional status of the PFC based on varying affinities for adrenergic receptors. Moderate concentrations of NE during the nonstressed, alert state lead to improved PFC function via binding to higher-affinity α2A-adrenoceptors (shown in blue). Higher concentrations of NE release in the PFC, which are observed in conditions of stress, lead to PFC impairment due to the binding to lower-affinity α1- and possibly β1-adrenoceptors. Recent studies have started to uncover downstream, intracellular signaling mechanisms that mediate these NE actions. In contrast, the amygdala is modulated in a manner opposite to the PFC, whereby α2-adrenoceptors impairs, and β- and α1-adrenoceptors strengthen, amygdala regulation of behavior. Thus, NE may act as a chemical switch to determine whether behavior is regulated by higher-order PFC operations or more primitive mechanisms mediated by the amygdala. AC, adenylyl cyclase; cAMP, cyclic adenosine monophosphate; DAG, diacylglycerol; IP_3, inositol triphosphate; NE, norepinephrine; PFC, prefrontal cortex; HCN, hyperpolarization-activated cyclic nucleotide-gated channels; PKC, protein kinase C; PLC, phospholipase C.

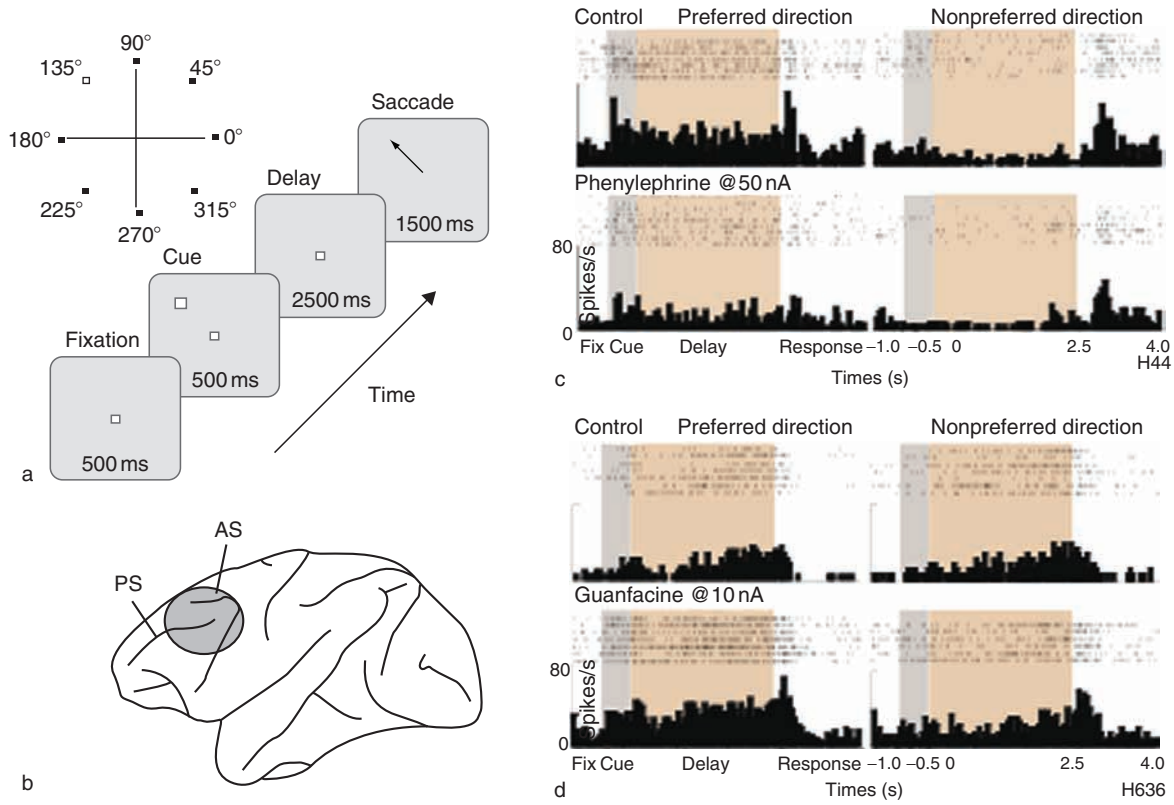

Figure 3 The delay-related activities of neurons of the monkey PFC provide a cellular representation of working memory: (a) the oculomotor delayed-response task often used for *in vivo* electrophysiological recordings; (b) an electrode is lowered into the monkey's dorsolateral PFC to record responses while the monkey is performing the task; (c) the response of delay-related neurons altered by the iontophoretic application of phenylephrine, an $\alpha 1$ agonist, acting on adrenoceptors; (d) the response of delay-related neurons altered by the iontophoretic application of guanfacine, an $\alpha 2A$ agonist, acting on adrenoceptors. As shown in (a), following a cue at a specific orientation, there is a delay period during which the monkey must keep the previously presented cue in mind in order to guide the appropriate saccade (response). As shown in (c) and (d), cells show preferences for particular orientations (preferred vs. nonpreferred), and specific neurons fire selectively during either the cue, delay, or response periods. The iontophoretic application phenylephrine suppresses delay-related firing (c), whereas the application of guanfacine enhances delay-related firing (D). PS, principle sulcus; As, arcuate PFC, prefrontal cortex. This figure was generously supplied by Dr. Min Wang, Yale University.

influence of $\alpha 1$- versus $\alpha 2$-adrenoceptor stimulation on the delay-related firing of PFC neurons in monkeys performing an oculomotor delayed response task. Stimulation of $\alpha 1$-adrenoceptors with phenylephrine reduces delay-related cell firing for the preferred direction (Figure 3(c)), whereas stimulation of $\alpha 2A$-adrenoceptors with guanfacine enhances delay-related firing for the preferred direction (Figure 3(d)). However, β mechanisms are just beginning to be understood and may involve opposing actions at $\beta 1$ versus $\beta 2$ receptors that are obscured with nonselective compounds. The administration of $\alpha 2A$ agonists improves spatial working memory in mice, rats, monkeys, and humans, especially in subjects with PFC dysfunction. In addition to working memory, the $\alpha 2A$ agonist, guanfacine, has been shown to improve a number of other PFC functions, such as lessening distractibility, improving reversal performance and other measures of behavioral and cognitive inhibition (e.g., in humans, improving performance on the Stroop interference task), and strengthening conditional motor learning. It has little effect on posterior cortical functions and, indeed, can impair amygdala function. Conversely, the blockade of endogenous $\alpha 2$ receptors in the PFC with yohimbine impairs working memory, weakens no-go responding, and induces locomotor hyperactivity. Thus, this system is particularly relevant to attention-deficit hyperactivity disorder (ADHD). Recordings from PFC neurons in monkeys performing spatial working memory tasks show similar effects; $\alpha 2A$-adrenoceptor stimulation increases delay-related firing for the preferred direction (see Figure 3(d)), whereas yohimbine suppresses delay-related firing. These enhancing effects on working memory result from the stimulation of postsynaptic $\alpha 2A$ receptors in the PFC, inhibiting cAMP production, closing HCN channels on dendritic spines, and strengthening the functional connectivity of PFC circuits. Based on

research in animals, guanfacine is now in use for the treatment of ADHD, Tourettes syndrome, posttraumatic stress disorder (PTSD), and mild traumatic brain injury. Recent studies indicate that mutations on the Disordered In Schizophrenia 1 (DISC1) gene may lead to excessive cAMP signaling in patients with schizophrenia-like illnesses and that agents such as guanfacine may be useful in treating PFC deficits in these disorders as well.

In contrast to α2A-adrenoceptors, stimulation of α1-adrenoceptors impairs working memory. α1-Adrenoceptor antagonists have no effect on working memory when infused into the PFC during nonstress conditions, but they protect PFC cognitive function under conditions of uncontrollable stress. Conversely, infusions of α1-adrenoceptor agonists such as phenylephrine mimic the stress response and impair working memory. These detrimental actions are mediated by phosphotidyl inositol–protein kinase C intracellular signaling. Detrimental α1-adrenoceptor actions have been observed at the cellular levels as well, at which iontophoresis of phenylephrine onto PFC neurons decreased the delay-related cell firing necessary for working memory function (see **Figure 3(c)**). As with cognitive performance, this collapse in delay-related firing was reversed by a protein kinase C inhibitor. These mechanisms are especially relevant to PTSD, bipolar disorder, and schizophrenia, all of which are caused by, or worsened by, stress exposure. The α1-adrenoceptor antagonist prazosin is now used to treat patients with PTSD, and all atypical antipsychotic medications have α1 and 5-HT2 receptor-blocking properties. 5-HT2 receptors, like α1 receptors, are coupled to phosphotidyl inositol–protein kinase C intracellular signaling. Genetic and biochemical studies suggest that phosphotidyl inositol–protein kinase C intracellular signaling is overactive in bipolar disorder and schizophrenia (e.g., due to loss of function mutations in the genes encoding for DAG kinase or RGS4). Most antimanic agents reduce the activity of this signaling pathway. Thus, these agents may restore PFC regulation of behavior, thought, and affect by normalizing phosphotidyl inositol–protein kinase C intracellular signaling.

Serotonin

The raphe nuclei send widespread and dense 5-HT projections throughout the cortex, including the PFC. Electron microscopy studies indicate that these 5-HT projections make connections with both pyramidal cells and interneurons in the PFC. PFC neurons project back to the raphe nuclei, where they may have an inhibitory influence. There are a multitude of 5-HT receptor subtypes (at least 13); however, the localization and function of these receptors within the PFC are poorly understood. The 5-HT1 family appears to be more densely localized in superficial layers, whereas 5-HT2 receptors are more concentrated in middle layers.

It is well established that 5-HT plays a critical role in affective regulation by the OFC, whereas its effects on dorsolateral PFC function appear to be more complex. The most unambiguous evidence for the role of 5-HT in executive processes comes from studies of reversal learning in marmoset monkeys. In reversal learning, subjects are presented with two stimuli and initially learn to associate reward with one stimulus. After this pairing has been established, a subsequent reversal between stimulus and reward takes place, such that the animals must now make the previously unrewarded response in order to receive a reward. Studies in humans as well as lesion studies in nonhuman primates indicate that the performance of this task depends on the OFC as well as the ventromedial PFC. Selective reductions in 5-HT via injections of 5,7-dihydro-testosterone (DHT) directly into the OFC of marmoset monkeys impairs performance on serial discrimination reversal tasks. In contrast, performance is preserved on an attentional set-shifting task that requires subjects to reverse rules and learn new rules regarding a perceptual dimension of a stimulus (i.e., shape or color) or a specific feature of a percept (see **Figure 4**). Monkey studies have shown that set-shifting performance depends on the dorsolateral PFC. Similar results have been observed in humans, in whom rapid tryptophan depletion impairs reversal learning while preserving attentional set-shifting performance. Interestingly, studies in monkeys as well as rats have shown that OFC depletion of DA had no effect on performance of a serial reversal learning task.

In addition to regulating attentional set shifting, the lateral PFC inhibits inappropriate motor responses, which can be evaluated by performance on a stop-signal task. A study in humans found that increases in NE, via NE reuptake inhibition, led to faster reactions on a stop-signal task. Performance on a probabilistic learning task, an analog to reversal learning, was not affected by NE reuptake inhibition. Conversely, increased basal 5-HT levels produced impairments on a probabilistic learning task but did not influence stop-signal performance. These findings support the findings from depletion studies in humans and monkeys. Together, these data have led to the hypothesis that 5-HT and catecholamines have dissociable and specific roles in the modulation of PFC operations.

One significant barrier to interpreting studies using global depletions or enhancements of 5-HT action is

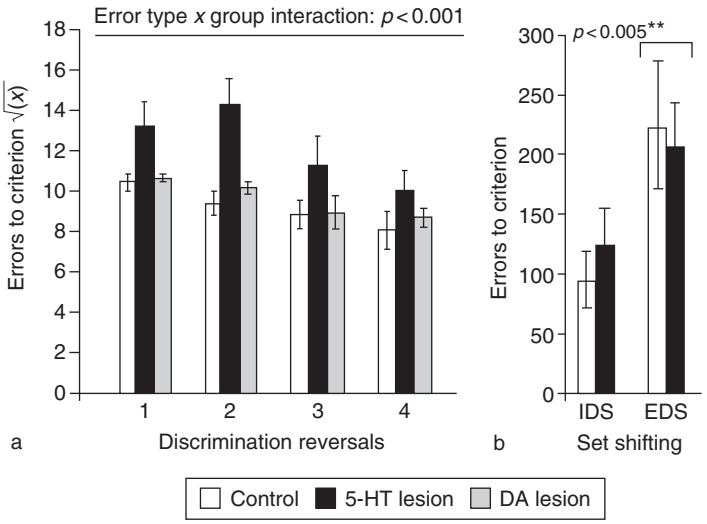

Figure 4 Effects of prefrontal serotonin depletions on learning and attention: (a) depletions of prefrontal serotonin (but not prefrontal dopamine) disrupt discrimination reversal learning; (b) prefrontal serotonin depletions have no effect on shifting an attentional set. In (b) note the greater number of errors made in performing a discrimination requiring a shift of attentional set (EDS) compared to one involving the maintenance of a previously acquired attentional set (IDS), in both groups. 5-HT, serotonin; DA, dopamine; EDS, extra-dimensional shift; IDS, intradimensional shift. Courtesy of Dr. Angela Roberts, Cambridge University, UK.

unraveling the heterogeneous actions of different receptor subtypes. Future studies need to examine the role of specific 5-HT receptor subtypes and determine whether dorsolateral PFC operations are altered when the 5-HT system is probed more selectively. For example, electrophysiological studies indicate that 5-HT2 receptors may influence spatial working memory operations, and these findings need to be pursued at the behavioral level.

In summary, 5-HT plays a distinct, if complicated role in the affective flexibility and the execution of primed behaviors and many of the activities of the mPFC and the OFC. Through its connections with limbic areas, the medial OFC is critical in connecting associative information regarding outcomes with the representational memory processes of the PFC. Ultimately, this positions the OFC to regulate the generation and use of outcome expectancies, which underlie the ability to make adaptive decisions. Thus, it is not surprising that studies of major depressive disorder in humans implicate dysfunction of the OFC as well as the dorsal raphe nucleus. Future studies using receptor-selective agents and behavioral paradigms probing a wide array of PFC-mediated cognitive functions will bring an improved understanding of the complex role of 5-HT in higher-order cognitive and affective processes.

Acetylcholine

Cholinergic neurons project from the basal forebrain to virtually all layers and all regions of the cortex, with approximately 75% of cholinergic cortical innervation originating from the nucleus basalis. Cholinergic axons make synapses in all layers of the monkey PFC, targeting pyramidal, and nonpyramidal cells. Most synapses on pyramidal cells target dendritic shafts, although there are some on spines and very few on the cell bodies. There are two general families of cholinergic receptors: muscarinic and nicotinic receptors. There are five muscarinic subtypes. Of these, the M2 receptor serves as both a presynaptic autoreceptor and a postsynaptic receptor on PFC pyramidal and nonpyramidal cells. In PFC pyramidal cells, M2 receptors are observed on dendritic spines receiving excitatory (asymmetric) inputs. The M1 receptor is also found postsynaptically on dendritic shafts and spines. Nicotinic receptors are often found presynaptically on the terminals of monoamine or ACh axons, where they can regulate transmitter release. They are most dense in layer I of the monkey PFC, which contains the highest levels of catecholamine fibers.

Although there have been detailed anatomical studies of the ACh inputs to the PFC, there has been little or no research on the function of these inputs in monkeys. Many studies have found alterations in working memory with systemic administration of muscarinic and nicotinic compounds, but these agents have powerful actions throughout the brain (e.g., in the thalamus), and thus alterations in cognitive performance cannot be necessarily attributed to PFC actions. In contrast, there has been a great deal of progress outlining ACh actions in the rodent mPFC, in which

infusions of muscarinic blockers such as scopolamine directly into the mPFC induce working memory deficits. Much of this research has focused on the important role of ACh in attention, specifically vigilance. These effects are evident both in the early stages of sensory processing and in subsequent top-down processing, which is regulated by the PFC.

It is widely accepted that ACh-mediated enhanced processing of sensory information underlies the cognitive process of attention. Attention encompasses a variety of operations that together contribute to the detection and discrimination of stimuli; and the integrity of attention processes contributes to the efficacy of higher-order cognitive functions such as learning and memory. Numerous behavioral studies have implicated basal forebrain cortical cholinergic inputs in sustained attention functioning. Tasks measuring sustained attention, or vigilance, require an animal to distinguish signal trials, which typically use a panel of lights illuminated for a short duration of time, from nonsignal trials, in which no light is illuminated. Accuracy is measured by training the animals to make a designated response for both signal and nonsignal trials. A range of stimulus intensities, presentation intervals, and the introduction of distracters are often used to manipulate the difficulty of this task.

Performance on sustained attention as well as distracter tasks stimulates ACh release. Behavioral and electrophysiological studies have implicated the association areas of the cortex, most notably the mPFC, in the ACh facilitation of sustained attention. First, the PFC appears particularly sensitive to ACh. In rats, local administration of ACh stimulated mPFC neuronal activity. In addition, several studies have indicated that the mPFC requires ACh to effectively suppress the processing of irrelevant or distracting information. One study observed increases in ACh release in the frontal cortex of rats as they recovered from a distracter introduced in a sustained attention task. A separate study found rats with ACh depletion to be unimpaired on tasks of attention under normal conditions; however. the introduction of distracters into the task resulted in a significant decrease in performance as compared to their sham counterparts. These findings further support the importance of ACh to the maintenance of information and attention under challenging or unpredictable conditions, such as the presentation of distracters or the erosion of signal.

The modification of information processing and more complex operations by knowledge-driven or goal-oriented information is referred to as top-down regulation. Basal forebrain cortical cholinergic input is mediated by efferent projections from the PFC; thus, the PFC is well positioned to orchestrate cholinergic innervation throughout the brain. PFC-driven top-down control of signal processing is thought to be one mechanism through which the PFC filters distracting stimuli and directs thought and behavior appropriately. In primates, ACh cells probably receive information from the OFC regarding reward or other affective information that may motivate attentional state. It was recently shown in rodents that PFC cholinergic stimulation resulted in increased ACh release in the posterior parietal cortex, a region that has been implicated with attention. Thus, following stimulation, the PFC is positioned to deploy ACh release in other cortical regions.

Overactivity of ACh has been linked to attention performance deficits, specifically to an increased response to nonsignal stimuli. This impaired signal-to-noise processing results in overprocessing, leading to generically amplified and unfiltered information and ultimately to false-positive responses. However, underactive cholinergic input to the cortex may likewise contribute to cognitive and attention deficits. In Alzheimer's disease, cholinergic inputs from the basal forebrain deteriorate, and this loss of input is may contribute to impaired signal detection as well as subsequent impairments in the PFC-mediated or top-down regulation of attention. Although this ACh dysregulation appears to contribute to the symptoms of disorganized thought and behavior, at this time, no conclusive studies have distinguished the ramifications of ACh underactivity from gross impairments in executive centers that would otherwise direct this input. In sum, studies of dysregulated ACh processing in disease states reiterate that moderate levels of ACh are necessary for the appropriate signal-driven activation of cortical processing centers and for the subsequent regulation of attention via PFC-dictated ACh release.

Summary

The PFC subserves executive functions by organizing information required for future thought and action. A vast body of literature across species indicates that the cognitive tasks of the PFC are sensitive to a variety of neurochemicals and that different neurotransmitter systems have distinct roles in cognitive functions of the PFC. However, there are several formative obstacles to fully delineating the neurochemical modulation of executive functions, most notably, the myriad actions of specific receptors and the broad range of tasks regulated by the various subregions of the PFC. Understanding the intricacies of these powerful neurochemical influences on PFC function is key to our understanding of the etiology and treatment of many neuropsychiatric illnesses, including schizophrenia, ADHD, and PTSD, as well as the decline in PFC cognitive functions with advancing age.

See also: Acetylcholine Neurotransmission in CNS; Dopamine; Executive Function and Higher-Order Cognition: Definition and Neural Substrates; Norepinephrine: CNS Pathways and Neurophysiology; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Strategic Control of Memory; Working Memory: Capacity Limitations.

Further Reading

Arnsten AFT (2007) Catecholamine and second messenger influences on prefontal cortical networks of representational knowledge: A rational bridge between genetics and the symptoms of mental illness. *Cerebral Cortex* 17(Supplement 1): 16–17.

Arnsten AF and Li BM (2005) Neurobiology of executive functions: influences on prefrontal cortical functions. *Biological Psychiatry* 57: 1377–1384.

Arnsten AF and Robbins TW (2002) Neurochemical modulation of prefrontal cortical function in humans and animals. In: Stuss DT, and Knight RT (eds.) *Principles of Frontal Lobe Function*, pp. 51–84. Oxford: Oxford University Press.

Brozoski TJ, Brown RM, Rosvold HE, and Goldman PS (1979) Cognitive deficit caused by regional depletion of dopamine in prefrontal cortex of rhesus monkeys. *Science* 205: 929–932.

Chamberlain SR, Muller U, Blackwell AD, Clark L, Robbins TW, and Sahakian BJ (2006) Neurochemical modulation of response inhibition and probabilistic learning in humans. *Science* 311: 861–863.

Clarke HF, Walker SC, Dalley JW, Robbins TW, and Roberts AC (2007) Cognitive inflexibility after prefrontal serotonin depletion is behaviorally and neurochemically specific. *Cerebral Cortex* 17: 18–27.

Goldman-Rakic PS (1995) Cellular basis of working memory. *Neuron* 14: 477–485.

Robbins TW (2005) Chemistry of the mind: Neurochemical modulation of prefrontal cortical function. *Journal of Comparative Neurology* 483: 140–146.

Sarter M and Bruno JP (1997) Cognitive functions of cortical acetylcholine: Toward a unifying hypothesis. *Brain Research Reviews* 23: 28–46.

Sarter M, Hassemo M, Bruno JP, and Givens B (2005) Unraveling the attentional functions of cortical cholinergic inputs: Interactions between signal-driven and cognitive modulation of signal detection. *Brain Research Reviews* 48: 98–111.

Vijayraghavan S, Wang M, Williams GV, and Arnsten AFT (2007) Dopamine D1 receptor stimulation alters tuning of prefrontal cortical neurons during spatial working memory. *Nature Neuroscience* 10: 376–384.

Wang M, Ramos BP, Paspalas CD, et al. (2007) α2A-Adrenoceptors strengthen working memory networks by inhibiting cAMP-HCN channel signaling in prefrontal cortex. *Cell* 129: 397–410.

Wang M, Vijayraghavan S, and Goldman-Rakic PS (2004) Selective D2 receptor actions on the functional circuitry of working memory. *Science* 303: 853–856.

Cognitive Control and Development

D Amso and B J Casey, Sackler Institute for Developmental Psychobiology, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Immature cognition is characterized by a greater susceptibly to interference from competing inputs, representations, or responses. Cognitive control is the ability to reduce this interference. This ability has been included in a number of theories of attention and memory and referred to in a number of ways, including executive function, attentional bias, and supervisory attention. Each of these coined terms is suggestive of a mechanism that is required to direct or guide our attention or actions. The following question is addressed in this article: How is it that we gain this ability with development?

Measures of cognitive control across development are necessarily constrained by the changing repertoire of behaviors available at different points in development and by age-appropriate techniques. At the core of these tasks is the demand for selection of one target, action, location, or dimension in the presence of competing alternatives. While holding this defining feature constant, researchers have been able to manipulate aspects of cognitive control paradigms to be appropriate for use at different ages while still allowing valid examination of this one construct. We review findings from paradigms used across development.

Evidence of Cognitive Control in Infancy

In the context of the infant, the environment is full of sensory and perceptual information competing for attention. For learning to take place, there must be a developing control mechanism that allows for suppression of irrelevant information so that selection of information relevant for current learning or behavior is possible. Some of the earliest research targeting control over behavior in infancy made use of reaching paradigms and originally posited that cognitive control was not apparent until the end of the first postnatal year. However, research into the development of the oculomotor system has shed new light on control over behavior in infancy using an age-appropriate measure. Although it is clear that cortical influences on behavior are in place at birth, robust endogenous or voluntary control over eye movements is not apparent until approximately 3–4 months of age, with earlier behavior being largely reflexive and subcortically mediated. In this context, the developing ability to voluntarily suppress responses to one stimulus in order to select another reflects emerging components of cognitive control.

Infant eye-tracking studies have used negative priming paradigms to index cognitive control. In a negative priming paradigm, a distractor object or location that has previously been ignored on one trial becomes the target to be selected on the next trial. Typically, responses to it are slowed relative to a control stimulus. The simultaneous presentation of the target and distractor generates competition that needs to be resolved by inhibition of the distractor in favor of the target. Infants have been shown to have the capacity to suppress an irrelevant stimulus in order to select a competing salient alternative.

An attempt has been made to link observable changes in oculomotor control in infancy to the development of certain visual pathways. Endogenous control over eye movements indicates the development of the frontal eye fields and the strengthening of its projections to the oculomotor basal ganglia thalamocortical circuit. This circuit, one of several closed loops that project from cortex to the basal ganglia and subsequently back to cortex, receives projections from both the frontal eye fields and the parietal cortex and appears to be critical for the production of voluntary saccades. Infants as young as 4 months have been shown to be capable of inhibiting an automatic saccade to a peripheral flashing target, suggesting rudimentary functioning of this circuit early in life. Overriding an orienting response takes effort because the visual system is designed to orient to salient or novel stimuli in the environment.

By the end of the first postnatal year, manual dexterity (e.g., reaching for objects) provides another measure beyond simple looking behavior to document the development of cognitive control. The classic Piagetian A-not-B task that relies on reaching behavior has been used to document changes in cognitive control between 8 and 12 months. Infants watch a toy hidden in location A and are prompted to reach and retrieve it. After several trials of the toy being hidden at location A, the toy is hidden in location B in full view of the infant. Infants younger than 12 months show the A-not-B error (i.e., they perseverate and continue to search in location A). This error appears to be a failure in the ability to suppress a response to the salient competing location in favor of the correct one (i.e., cognitive control).

Nonhuman primate lesion studies suggest that the A-not-B task is dependent on development of the prefrontal cortex. Neural processes underlying this behavior in humans have been examined using a looking

version of the task in combination with scalp electrode recordings. These recordings show an increase in frontal as well as posterior power values for only those infants who do not commit the A-not-B error, implicating the prefrontal cortex and related posterior circuitry in accurate performance of the A-not-B task.

Evidence of Cognitive Control in Early Childhood

Cognitive control paradigms in preschool children tend to focus on emerging proficiency in rule use and representational thought. In general, young children have difficulty if they are required to override a competing rule or dimension. Errors of cognitive control have been shown using tasks originally adapted from the adult neuropsychology literature, including the Stroop task and Wisconsin Card Sorting Task. In the Stroop task, subjects are required to name the color of the ink in which a word is written. The classic finding is a slowing of response times when the color of the ink is incongruent with the word (e.g., the word 'blue' written in red ink) relative to a color–word congruent example. Variations of this task have been developed for use with children, especially preschoolers who are nonreaders. For example, in the day–night Stroop-like task, subjects are asked to respond "day" to a card depicting a night scene and "night" to a card depicting a day scene. In both the adult and the child versions of the Stroop task, subjects must suppress the interfering salient information (color or image) in order to respond according to the appropriate rule. Children 3–5 years of age tend to have difficulty overriding the prepotent response, with performance not becoming more adultlike until later childhood. Manipulations that reduce the competing demands of the task or provide young children with more time to compute the answer result in even 4-year-olds performing well.

The Dimensional Change Card Sort is a simplified version of the Wisconsin Card Sorting Task and a more complicated version of the A-not-B task. Three- and 4-year-old children are presented with two cards that can be sorted in either of two ways (e.g., a red truck can be sorted by either color or shape). Children are generally able to sort the cards accurately by the preswitch dimension (e.g., color), but 3-year-olds perseverate on this dimension during the postswitch phase. That is, when the rules change such that they are required to respond based on the other dimension (e.g., shape), these children continue to sort by color, even though they can articulate the new rule. This behavior is consistent with the child not being able to suppress aspects of the situation that were previously relevant in the service of making the goal-appropriate action.

Evidence of Cognitive Control in Late Childhood and Adolescence

As children's fine motor skills improve, more precise measures of cognitive control, which index improvements in reaction time as well as accuracy across a wide age range, can be applied. Tasks used to measure cognitive control during this period of development include negative priming, Stroop tasks, go/no-go tasks, and Simon tasks. In all cases, children have a more difficult time ignoring or inhibiting irrelevant salient information or prepotent responses in favor of the relevant items or responses. In the Simon task, for example, a stimulus attribute is relevant for a response (e.g., left button press), whereas its location (e.g., right side of display) is interfering and irrelevant (e.g., a blue square appears on the left of the screen but requires a response with the right hand). The Simon effect is the difference in accuracy or reaction time between trials in which stimulus and response are on the same side and trials in which they are on opposite sides, with responses being generally slower and less accurate when the stimulus and response are on opposite sides. The size of the Simon effect decreases with age, suggesting that susceptibility to interference decreases with age.

Behavioral evidence from developmental studies of antisaccade tasks and go/no-go tasks in this age range provide similar results. In a go/no-go task, subjects are asked to respond to one stimulus (target) and withhold responding to a different stimulus (nontarget) as the stimuli are presented in succession. Both measures provide evidence of continued improvement in cognitive control into adolescence. Similarly, attention tasks that include distracting peripheral information, as in the case of the flanker task and stop signal tasks, show comparable developmental changes. Performance across these tasks continues to develop over childhood and does not reach full maturity until approximately 12 years (**Figure 1**). Importantly, these age-related differences are not observed on these tasks in the absence of interfering information.

Cortical Organization Underlying the Development of Cognitive Control

Advances in neuroimaging techniques have provided developmental scientists the opportunity to safely track cognitive and neural processes underlying

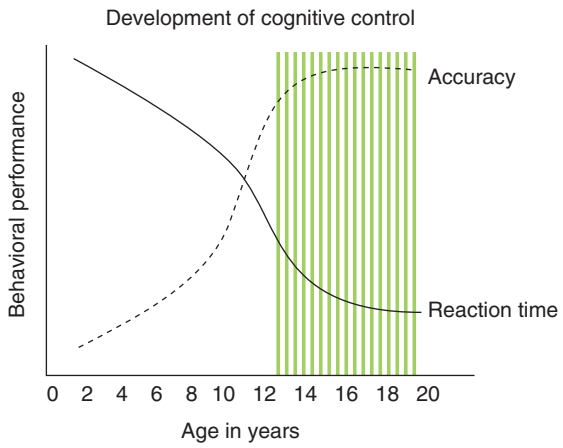

Figure 1 Model of developmental change in cognitive control with age as a function of accuracy and reaction time. From Casey BJ, Frontostriatal and frontocerebellar circuitry underlying cognitive control. In: Mayr U, Awh E, and Keele SW (eds.) *Developing Individuality in the Human Brain*. American Psychological Association, Washington, DC, 2005, reprinted with permission.

human development. These methods have advanced the field of developmental neuroscience by providing evidence of changes in cortical organization in the developing brain *in vivo*. Although these measures provide only an indirect index of a very dynamic process of regressive and progressive changes in brain structure and function across development (e.g., changes in myelination, dendritic pruning, and vascular, neuronal, and glial density), they provide information about regional changes in brain development.

One of the most influential magnetic resonance (MR) methods is functional magnetic resonance imaging (fMRI). This method provides *in vivo* measures of brain activity by indexing changes in blood oxygen levels in the brain. Studies using this method have begun to track developmental changes in patterns of brain activity that coincide with changes in cognitive control across childhood.

Prefrontal involvement in the development of cognitive control in children has been demonstrated most convincingly in fMRI studies that parametrically manipulate the salience of the competing information that must be suppressed. By definition, cognitive control is the ability to attenuate interference from competing sources and thus should be needed more when the interfering information becomes increasingly salient. For example, in the case of the go/no-go task, as the number of go trials preceding a no-go trial increases, so too do the number of errors (i.e., incorrectly responding on a no-go trial). The imaging data show that correctly withholding a response on a no-go trial is associated with activity in the ventral prefrontal cortex and basal ganglia, and activity in these regions correlates with successful task performance. Adults show a monotonic increase in prefrontal and parietal cortices as the number of go trials preceding a no-go trial increase. Children, on the other hand, maximally activate these regions regardless of whether they have to withhold a response following one, three, or five go trials (responses) and make significantly more errors. These findings provide evidence that susceptibility to interference is greater in children than adults but is attenuated with age. This increase in cognitive control with age is paralleled by maturational differences in the recruitment of underlying prefrontal and parietal cortices.

Many paradigms besides the go/no-go task have been used to study the neural development of cognitive control (e.g., antisaccade, Simon, flanker, and Stroop tasks). Children with effective cognitive control on these tasks typically recruit additional or larger portions of prefrontal cortex than those activated by adults, suggesting age-related differences in recruitment of this region. Children also recruit posterior parietal regions consistently activated in adults on these tasks, unless they have poor cognitive control, in which case they do not recruit these regions. These findings suggest that the improved ability to suppress irrelevant information may first require mature activation of posterior parietal regions prior to maturation of prefrontal regions. Tasks of lesser cognitive demand (e.g., selective attention tasks without response competition) do not appear to show age-related differences. This developmental pattern of brain activity parallels that of behavioral performance on cognitive tasks that do not require suppression of competing information, for which children also show an adultlike pattern.

Collectively, these imaging studies of cognitive control show that children recruit distinct but often more diffuse regions of prefrontal cortex than do adults. Based on cross-sectional and longitudinal studies, the pattern of activity within brain regions central to task performance becomes more focal or fine-tuned with increased activity, whereas regions not correlated with task performance show attenuation in activity with age. This pattern, observed across a variety of paradigms, has been suggested to reflect development within, and refinement of, projections to and from these regions.

Development of Brain Connectivity and Cognitive Control

Another MR method that has begun to be used in developmental research is diffusion tensor imaging (DTI). This method detects changes in white matter microstructure based on properties of diffusion.

Diffusion of water molecules in white matter tracts is affected by myelin and the orientation and regularity of fibers. As such, this method provides an index of regional brain connectivity and strengthening of connections among regions. Together with fMRI, this methodology has been used to track the refinement of cortical circuits involved in the fine-tuning of systems underlying the development of cognitive control.

Only a small number of studies have linked brain connectivity and cognitive development measures, although indirect measures of white matter suggest regional development in prefrontal cortex and presumably function. Across all these studies, strengthening of connections throughout the brain is shown to occur with development; however, connections between those regions implicated in cognitive control appear to be protracted and are also shown to correlate with behavioral performance. One such study showed that development of working memory was positively correlated with prefrontal–parietal connectivity consistent with imaging studies showing differential recruitment of these regions in children relative to adults. A second study examining the degree of connectivity between the prefrontal cortex and the basal ganglia and performance on the go/no-go task showed a tight association between connectivity and performance that was not present in a comparison corticospinal fiber tract that also showed changes with development.

Findings from the imaging literature suggest that development of cognitive control is paralleled by a fine-tuning of cortical systems, with protracted refinement of prefrontal regions relative to posterior cortical and subcortical regions. The reported shift in cortical architecture and function is presumably an experience-driven maturational process that reflects fine-tuning of neural systems with experience and development, and strengthening of cortico-cortical and cortico-subcortical connections. These imaging studies map onto findings from animal and human postmortem studies indicating that pruning and elimination of connections, in combination with strengthening of relevant ones, occur during this time period, and illustrate the subtle interplay between neuroanatomic and physiologic changes in neural circuitry and cognitive maturation.

A Developmental Model of Cognitive Control

The imaging literature on the development of cognitive control suggests that maturity of prefrontal regions is critical for the improvement of this ability with age. However, two other regions, the parietal cortex and basal ganglia, have been implicated repeatedly in directing or guiding attention and actions. Specifically, the basal ganglia have been implicated in signaling prefrontal cortex when competing actions require top-down adjustment of behavior, whereas the posterior parietal cortex has been implicated in signaling prefrontal cortex when competing stimuli require top-down biasing of attention in favor of one stimulus attribute or location over another. Both of these regions (basal ganglia and parietal cortex) are part of unique circuits that project both to and from prefrontal cortex, thus providing a means for signaling prefrontal regions to help impose top-down control of behavior (Figure 2).

Although these structures are neuroanatomically intact at birth, refinement of function and connectivity within and between these regions continues gradually across development. In this regard, each region has been suggested to have different developmental trajectories, whereby subcortical systems such as the basal ganglia develop before cortical systems. Within cortical systems, posterior regions (parietal cortex) have been suggested to develop before anterior ones such as the prefrontal cortex. The regional differences in functional development of brain regions may explain the protracted development of cognitive control. For example, intact signaling by subcortical regions, but immature top-down control due to

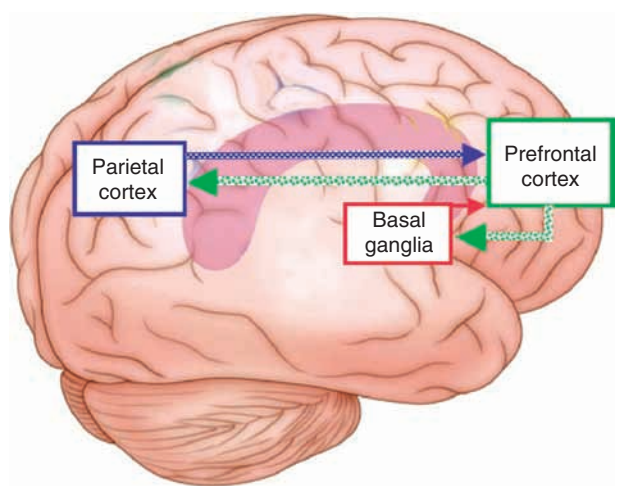

Figure 2 Simplified drawing of prefrontal circuits implicated in cognitive control. Each circuit has projections both to and from the prefrontal cortex. The basal ganglia and posterior parietal systems signal the prefrontal cortex when new or competing response or sensory information is encountered so that the prefrontal cortex can immediately adjust behavior to accommodate or suppress new or salient information during goal-oriented actions. Poor cognitive control early in development results from immature top-down control of prefrontal cortico-cortical and cortico-subcortical projections.

immature cortico-subcortical or cortico-cortical connections, could result in poor regulation of attention and behavior as seen in the review of studies of cognitive development. This developmental model of cognitive control suggests more of a refinement in the functional representations of intact brain regions as connections between brain regions are strengthened through experience over the course of development, as opposed to reorganization of brain circuitry with development or the development of new functional regions.

Conclusions

A hallmark of cognitive development is the ability to select information and actions from competing sources. Collectively, this article tracks this ability across development and shows how cognitive control emerges early in development and continues to be refined into adulthood. A recurring theme across all ages discussed in this article is the propensity to commit perseverative errors early in the development of cognitive control. For example, after frequent presentations of a toy at location A in the A-not-B task, infants cannot shift their response to a toy hidden at location B and continue to perseverate and search for the toy in location A. In the card sorting task, preschoolers have difficulty switching to a new rule after using another rule to sort cards. In the go/no-go task, increasing the number of go trials preceding a no-go trial increases the interference from a competing prepotent response and results in a decrease in the ability to withhold responding to the nontarget. Across the different tasks, the pattern of behavior suggests immature maintenance of relevant information (e.g., current hiding location and rule or task demands) from interference from the competing representation of a prior action.

Cognitive development is characterized as the growing ability to filter and suppress irrelevant information and actions in favor of more relevant ones as these neural systems develop. The prefrontal cortex has been shown to be involved in guiding actions by supporting representations of relevant information against interference from competing sources. Thus, developmental changes in prefrontal circuitry observed throughout childhood presumably contribute to the emergence of this ability.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Cognition: An Overview of Neuroimaging Techniques; Cognitive Neuroscience: An Overview; Diffusion Tensor Imaging (DTI); Executive Function and Higher-Order Cognition: Neuroimaging; Executive Function and Higher-Order Cognition: Definition and Neural Substrates; Functional Connectivity; Prefrontal Cortex: Structure and Anatomy.

Further Reading

Bunge SA, Dudukovic NM, Thomason ME, Vaidya CJ, and Gabrieli JD (2002) Immature frontal lobe contributions to cognitive control in children: Evidence from fMRI. *Neuron* 33: 301–311.

Casey BJ (2005) Frontostriatal and frontocerebellar circuitry underlying cognitive control. In: Mayr U, Awh E, and Keela SW (eds.) *Developing Individuality in the Human Brain*. Washington, DC: American Psychological Association.

Diamond A (1991) Neuropsychological insights into the meaning of object concept development. In: Carey S and Gelman R (eds.) *The Epigenesis of Mind: Essays on Biology and Knowledge*, pp. 67–110. Hillsdale, NJ: Lawrence Erlbaum.

Diamond A, Kirkham NZ, and Amso D (2002) Conditions under which young children can hold two rules in mind and inhibit a prepotent response. *Developmental Psychology* 38: 352–362.

Durston S, Davidson MC, Tottenham N, et al. (2006) A shift from diffuse to focal cortical activity with development. *Developmental Science* 9(1): 1–8.

Gogtay N, Giedd JN, Lusk L, et al. (2004) Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences of the United States of America* 101: 8174–8179.

Hasher L and Zacks RT (1998) Working memory, comprehension, and aging: A review and a new view. *Advances in Research and Theory* 22: 193–225.

Johnson MH (2005) *Developmental Cognitive Neuroscience: An Introduction,* 2nd edn. Oxford: Blackwell.

Klingberg T, Forssberg H, and Westerberg H (2002) Increased brain activity in frontal and parietal cortex underlies the development of visuospatial working memory capacity during childhood. *Journal of Cognitive Neuroscience* 14: 1–10.

Liston C, Watts R, Tottenham N, et al. (2006) Frontostriatal microstructure predicts individual differences in cognitive control. *Cerebral Cortex* 16(4): 553–560.

Luna B and Sweeney JA (2004) The emergence of collaborative brain function: fMRI studies of the development of response inhibition. *Annals of the New York Academy of Sciences* 1021: 296–309.

Munakata Y (1998) Infant perseveration and implications for object permanence theories: A PDP mode of the AnotB task. *Developmental Science* 1: 161–184.

Ridderinkhof KR, van der Molen MW, and Band GPH (1997) Sources of interference from irrelevant information: A developmental study. *Journal of Experimental Child Psychology* 65: 315–341.

Schlaggar BL, Brown TT, Lugar HM, Visscher KM, Miezin FM, and Petersen SE (2002) Functional neuroanatomical differences between adults and school-age children. *Science* 296: 1476–1479.

Sowell ER, Thompson PM, Holmes CJ, Jernigan TL, and Toga AW (1999) *In vivo* evidence for post-adolescent brain maturation in frontal and striatal regions. *Nature Neuroscience* 2(10): 859–861.

Zelazo PD, Muller U, Frye D, and Marcovitch S (2003) The development of executive function. *Monographs of the Society for Research in Child Development* 68(3)Serial No. 274.

Cognitive Deficits in Schizophrenia

D S Manoach, Massachusetts General Hospital, Charlestown, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Why Study Cognitive Deficits in Schizophrenia?

For many years, schizophrenia was considered to be a profound mystery, beyond the reach of science. Over the past several decades, major advances in the neurosciences have drastically changed the way schizophrenia is viewed, have deepened our understanding of it, and have led to better treatments. Most of all, it has become abundantly clear that schizophrenia is a brain disease. It is characterized by 'positive' psychotic symptoms such as hallucinations and delusions and by 'negative' symptoms such as amotivation, social withdrawal, and apathy. Positive symptoms are the most dramatic manifestations of schizophrenia and have been the primary focus of drug development. They can be effectively managed by medications in the majority of individuals, but a full return to premorbid levels of function is rare. The focus on symptoms has contributed to a relative neglect of the subtler but severely disabling cognitive deficits until recently.

Since the initial conceptualization of the disease as 'dementia praecox,' cognitive dysfunction has been recognized as integral to schizophrenia. Rather than being an epiphenomenon of psychosis, cognitive deficits reflect the pathology that gives rise to schizophrenia. They are probably more closely related to predisposing genes and are better predictors of social function and outcome than are the symptoms that define schizophrenia. Increasingly, the goal of schizophrenia research is to understand the nature and neural basis of cognitive deficits and to effectively treat them. Aside from the obvious benefits of treatment, there are other important reasons to understand cognitive deficits. First, because cognitive abnormalities predate the onset of illness, they may identify individuals at high risk for developing schizophrenia and enable early intervention. Second, their presence in healthy relatives of individuals with schizophrenia suggests that cognitive deficits are markers of genetic vulnerability to illness that can help to identify genes that confer increased risk of schizophrenia. Third, schizophrenia is a heterogeneous illness, and identifying core cognitive deficits and their genetic bases may lead to more valid subtyping, which has important implications for treatment and prognosis. In summary, understanding cognitive deficits, along with relatively preserved abilities in schizophrenia, will implicate underlying neural circuitry, guide the search for neuropathology, lead us to susceptibility genes, and thereby provide clues to etiology (or etiologies) and contribute to a theoretical understanding of this enigmatic disorder.

Article Overview

The literature documents impairments in a wide range of neurocognitive functions in schizophrenia. In fact, most meta-analytic studies of cognition in schizophrenia demonstrate that no domain tested is spared. The aim of this article is not to provide an exhaustive list of cognitive deficits; rather it is to highlight a subset that seems most central to schizophrenia and that holds the most promise for shedding light on its pathophysiology. One criterion for identifying core deficits is whether they are present in increased rates in healthy relatives of patients with schizophrenia. This would suggest that they are markers of genetic vulnerability to illness, or 'cognitive endophenotypes.' The endophenotype concept in psychiatry is discussed in the section titled 'Cognitive endophenotypes.' Meta-analytic studies of relatives show the strongest effect sizes for impairments on tasks with high executive function demands, such as set-switching, inhibition, and working emory (see 'Executive function deficits'). Disturbances of memory, thought, and language are described in the sections titled 'Declarative memory' and 'Thought, language, and semantic memory disturbances.' The section titled 'Challenges and potential confounds in the study of cognition' outlines some of the difficulties in identifying specific cognitive deficits. And the section that asks 'Are there fundamental deficits that give rise to widespread cognitive disturbance?' considers whether there are more elementary deficits that can parsimoniously account for the range of cognitive dysfunction. Finally, the advent of sophisticated neuroimaging techniques and the revolution in genetics hold tremendous promise for illuminating the neural basis of cognitive deficits in schizophrenia (discussed in the section titled 'Promising advances in the study of cognition').

Cognitive Endophenotypes

Schizophrenia arises from a poorly understood and complex interaction of multiple genetic, environmental, and epigenetic risk factors. Identifying genes that predispose to schizophrenia will lead to early

identification of individuals at risk, foster a better understanding of schizophrenia at a molecular level, improve pharmacotherapy, and guide the development of preventive interventions. Since only a small portion of individuals who have genes that predispose to schizophrenia will become ill, we cannot rely on the presence of symptoms alone to identify individuals who carry susceptibility genes. Because cognitive deficits are associated with illness within families and are present at a higher rate among family members than in the general population, they are can be used to identify individuals who lack any overt symptoms of illness but who carry susceptibility genes. These cognitive deficits can be considered endophenotypes, or markers of genetic vulnerability to illness. (The endophenotype concept was first presented in relation to psychiatric disorders by Gottesman and Shields in 1972.) Endophenotypes are manifestations of pathology that may not be obvious without specialized testing and that are thought to be more closely related to the genes that predispose to illness than are symptoms. Because endophenotypes fill gaps in the causal chain between genes and more distal, complex disease manifestations such as symptoms, they may provide clues to etiology. The characteristics of useful endophenotypes are listed in Figure 1. Candidate cognitive endophenotypes include deficits in executive functions such as working memory and inhibition. Abnormalities in several psychophysiological measures have also been proposed as candidate endophenotypes. These include smooth pursuit eye tracking, the P300 event-related potential, and indices of sensorimotor gating such as backward masking, prepulse inhibition of the startle blink, and gating of the auditory P50 event-related potential component. These abnormalities in early information processing are mentioned here because they might have ramifications for cognitive processing down the line, although such links are still more theoretical than established.

Executive Function Deficits

Executive functions are a diverse set of cognitive abilities that allow flexible rather than reflexive responses to events. They are necessary for coordinating thought and action in the pursuit of goals. These goals may be immediate, such as finding food, but they are often remote, higher order goals such as earning an academic degree. Executive functions include switching flexibly from one task to another; inhibiting strongly called for, but inappropriate, response tendencies; problem-solving; planning; and using context to select an appropriate response. The prefrontal cortex is thought to play a pivotal role in mediating executive function on the basis of studies of monkey neurophysiology, frontal lesion patients, and human neuroimaging. Neuroimaging studies consistently implicate prefrontal neural circuitry in the executive function deficits of schizophrenia.

In schizophrenia, executive function deficits are reflected in behavior that is stimulus-bound rather than guided by context, perseverative, and stereotyped. The seminal studies that first established a direct link between deficient executive function and abnormal prefrontal activity employed the Wisconsin Card Sort Test (WCST) (Figure 2). Although clinically rich and sensitive to prefrontal network dysfunction, the WCST is a complex, multidimensional task that requires sustained attention, concept formation, performance monitoring, task switching, and working memory for successful performance. Therefore, poor performance and differential brain activation in schizophrenia compared to healthy individuals cannot be attributed to a specific deficit. More-recent studies employ tasks that constrain demands to identify specific executive processes that are intact or impaired.

Inhibition

Inhibition is the ability to suppress automatic or prepotent responses (e.g., looking toward a suddenly appearing object). Individuals with schizophrenia consistently show increased errors and frequently also show increased latencies for correct responses on tasks requiring inhibition, such as the Stroop interference paradigm, the go, no-go task, and the antisaccade task (Figure 3). In real life, deficits in inhibition might translate into behavior that is impulsive and stimulus-bound.

Task Switching

Task switching refers to moving flexibly from one set of rules to another in response to changing environmental contingencies. Task switching requires processes that are not necessary when simply repeating the same task and generally incurs costs in the form of increased response latency and errors. Clinical observations of perseveration – the contextually inappropriate and unintentional repetition of responses – and performance of neuropsychological tests such as the WCST have led to the presumption of task-switching deficits in schizophrenia. Using tasks that isolate cognitive processes, however, several groups have reported that task switching is normal. For example, while performing random sequences of prosaccades and antisaccades (Figure 3), patients showed normal task-switching costs. But unlike healthy individuals, their errors tend to be in the same direction of the prior saccade. This suggests an abnormally strong influence of the prior response. Thus, instead of reflecting

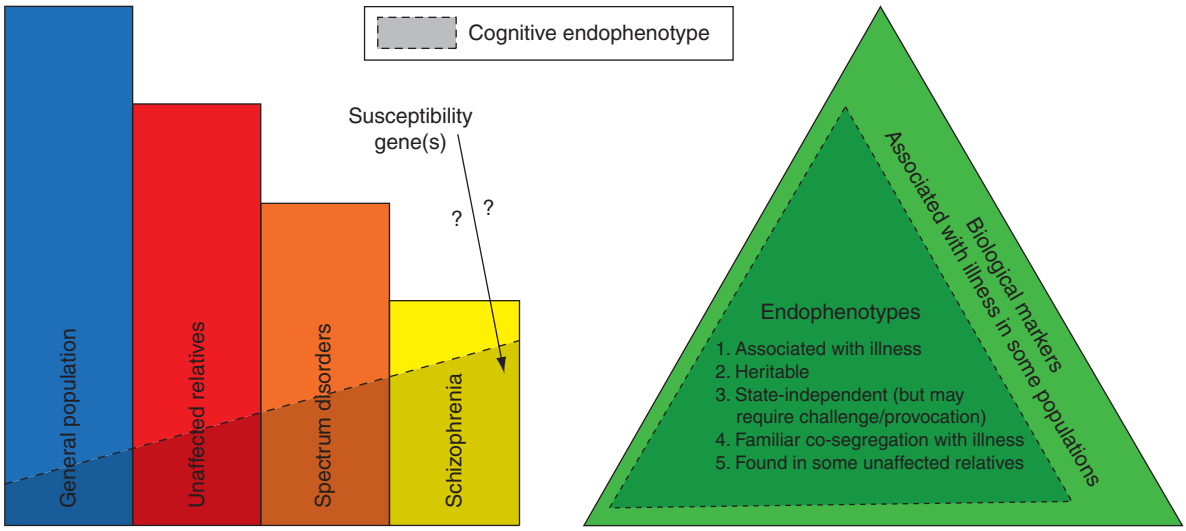

Figure 1 A schematic model of the frequency (left) and characteristics (right) of a cognitive endophenotype (e.g., inefficient working memory, formal thought disorder, antisaccade errors). It is most frequent in schizophrenia, then in schizophrenia spectrum disorders, then in unaffected family members, and then in the general population. It is presumed to reflect the influence of a susceptibility gene or genes. Enodophenotypes are a subset of biological markers that can be distinguished by the criteria listed, which suggests that they are genetic in origin (as opposed to primarily environmental, epigenetic, or multifactorial). The right side of the figure is adapted from Gould TD and Gottesman II (2006) Psychiatric endophenotypes and the development of valid animal models. *Genes, Brain, and Behavior* 5: 113–119 with permission from I. Gottesman, who owns the copyright.

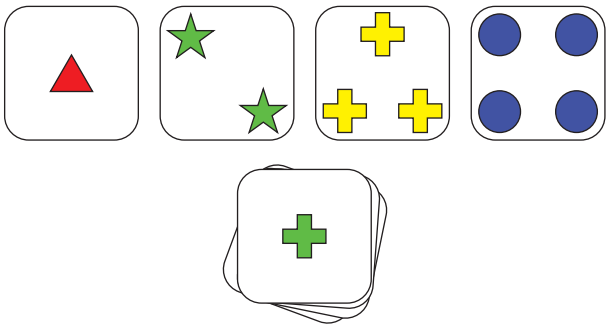

Figure 2 WCST: the participant is presented with four prototype cards and given a stack of cards to match to one of the prototypes by placing it beneath the prototype in a separate pile. The participant is not told the sorting rule (color, shape, or number) but has to figure it out on the basis of feedback from the examiner (i.e., a statement indicating that the match was correct or incorrect). If participants make a mistake, they must try again with the next card. Following ten correct sorts, unbeknownst to the participant, the rule changes, and the participant must again figure out the operative sorting rule based on feedback.

a problem with switching rules, perseveration may reflect an unwanted repetition of responses. Such dissociations between intact and impaired ability suggest that there are selective impairments of executive function in schizophrenia that can help pinpoint dysfunction in specific neural circuitry.

Working Memory

Working memory refers to actively holding information 'online' in the mind's eye and manipulating it in the service of guiding behavior. It is hypothesized to be a temporary store whose contents are continually updated and scanned in response to immediate information processing demands. Daily activities, from mentally rehearsing a phone number to considering alternative perspectives and outcomes, depend on it. In schizophrenia, there is abundant evidence of both behavioral impairment and anomalous patterns of brain activity during working memory performance. What is less clear is whether particular subcomponents of working memory are differentially affected. Working memory is comprised of distinct stages (encoding, maintaining and manipulating information, and then selecting the appropriate response) and operates on different domains of information (e.g., spatial, verbal, and object features). Neuroimaging and neurophysiological evidence suggests that different stages and domains of working memory require different strategies and cognitive processes for success and are therefore associated with distinct patterns of neural activity, although there is also a great deal of overlap in these patterns. Regardless of which subcomponents of working memory are disrupted in schizophrenia, impoverished internalized representations may lead to behavior that is unduly driven by external stimuli.

Response Monitoring

To perform well on any challenging task, it is necessary for one to evaluate one's performance so as to learn

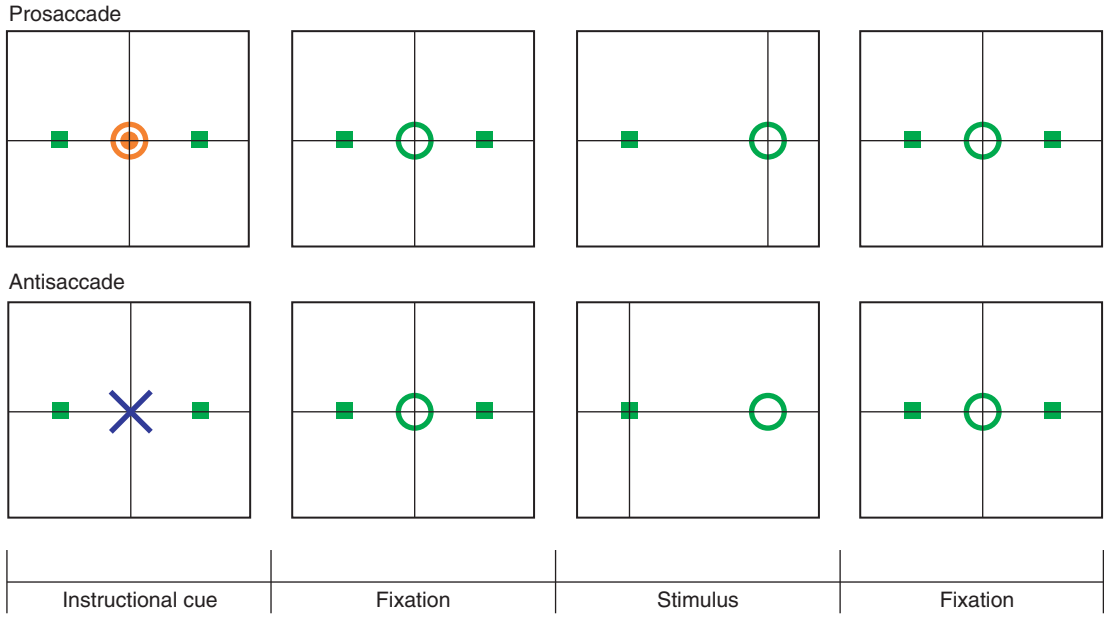

Figure 3 An example antisaccade paradigm. The cross hair represents the point of regard. A trial begins with a visual cue instructing either a prosaccade or an antisaccade (in this case an orange circle for prosaccades and a blue X for antisaccades). Following the disappearence of the cue, participants maintain fixation on the center until a stimulus appears over one of the two peripheral dots, the side randomly determined. Participants are instructed to look toward the suddenly appearing stimulus on prosaccade trials. For antisaccade trials, they are instructed to look in the opposite direction. After stimulus offset, fixation returns to the center as participants await the beginning of the next trial. While prosaccades are a relatively automatic response, antisaccades require executive control. To perform an antisaccade correctly one must suppress the prepotent response of looking toward a suddenly appearing stimulus (i.e., prosaccade) and generate the novel behavior of looking in the opposite direction. Out come measurements include directional accuracy of the saccadic eye movement and latency to initiate the saccade. Individuals with schizophrenia generally perform normally on prosaccade trials but show increased errors (failures of inhibition) and, depending on task parameters, may also show increased latency for correct responses on antisaccade trials.

from errors and, if an error was made, slow down on the next try. These processes – performance evaluation, reinforcement learning, and remedial performance adjustment on the subsequent trial (i.e., post-error slowing) – are all thought to rely on the anterior cingulate cortex. Diminished performance evaluation and diminished reinforcement learning in schizophrenia are presumed on the basis of relative reductions in anterior cingulate cortex activity and error-related negativity (ERN) following error commission on a range of cognitive tasks. ERN is an event-related potential that is thought to reflect error-based reinforcement learning, or the modification of stimulus–response mappings in order to optimize behavioral outcomes. In spite of these abnormalities, post-error slowing and error positivity (PE), an event-related potential associated with both error awareness and post-terror slowing, are often found to be intact. Such dissociations have been observed within single studies and suggest that performance evaluation and remediation are implemented by separate systems, only one of which may be impaired in schizophrenia. Because response monitoring is essential to optimal performance, a deficit could lead to generalized performance impairments and contribute to rigid, perseverative behavior.

Declarative Memory

Clinical observation suggests that individuals with schizophrenia have problems with memory. But memory is a broad term that encompasses the myriad ways that experience leaves durable traces in the brain. Declarative memory, particularly for verbal materials, has been a focus of research and is found to be deficient. Declarative memory includes both episodic memory (memory for events) and semantic memory (memory for facts) and can be divided into stages that engage distinct neural circuits: acquiring and organizing new information (encoding), filing and retaining it (storage), and retrieving it when necessary. Encoding and retrieval are both active processes that are vulnerable to disruptions of attention and depend on prefrontal networks. The storage of new information, in contrast, is not under voluntary control and is critically dependent on hippocampal mediation. There is

ample evidence of deficient encoding and retrieval in schizophrenia, but in spite of reports of structural and functional hippocampal abnormalities, storage deficits, as distinct from encoding and retrieval deficits, are mild or nonexistent.

Thought, Language, and Semantic Memory Disturbances

In addition to abnormalities in the content of thought (i.e., delusions and hallucinations), individuals with schizophrenia also have abnormalities in the form of thought. Formal thought disorder refers to impairments in the organization and control of thoughts as expressed through language. It ranges in severity and type, from disturbances of logical concept formation to peculiar word usage. Although viewed as a symptom, formal thought disorder has cognitive origins, and milder manifestations have been documented in healthy relatives, suggesting that it may be an endophenotype. Whether formal thought disorder reflects a primary thought or language disturbance has been debated since it was first described by Bleuler as "loosening of associations." Although not specific to schizophrenia – it is seen in other psychotic disorders, and even healthy individuals show instances of thought slippage – qualitatively, the flavor of formal thought disorder in schizophrenia is unique. It is frequently characterized by confusion, instability, and unusual word and phrase usage, in contrast to, for example, the loosely combined ideas and jocularity often seen in mania. At its most extreme, language is incoherent. Milder manifestations include vagueness (language that lacks specific information) or its opposite, overly specific and excessively qualified speech (e.g., a four-legged pig).

Even patients without overtly thought disordered language may show abnormalities of language production and comprehension at the levels of individual words, sentences, and discourse. Studies examining the organization and retrieval of individual words from semantic memory suggest that there is no loss of lexico-semantic knowledge in schizophrenia, but that the storage and/or access to words is less organized than in healthy individuals. At the level of whole sentences, patients have difficulty in the 'online' use of semantic context to constrain their expectations of what will follow, and this impairs comprehension. This may arise because patients have difficulty in building up context, due to deficits in combining semantic with syntactic information within sentences and in generating bridging inferences across sentences. This can lead sentence and discourse processing to be inappropriately dominated by semantic associative relationships between individual words at the expense of constructing a more complete meaning.

Challenges and Potential Confounds in the Study of Cognition

Characterizing deficient versus preserved cognitive abilities has proved an elusive goal in schizophrenia, and the literature is replete with ambiguous or contradictory results. Several factors contribute to this difficulty.

Heterogeneity

Schizophrenia is a heterogeneous illness with no known etiology or pathognomomic signs. In most cognitive studies, only a subset of patients shows deficits on any given measure, and the particular subset showing a deficit varies from measure to measure. Thus, compared with demographically matched healthy individuals, patients with schizophrenia show a greater range of values on outcome measurements ranging from simple reaction time to indices derived from neuroimaging studies. The effects of course variables (e.g., first episode vs. chronic, acute exacerbation vs. partially remitted state), medication, severity of illness, and the interaction of illness with factors that normally contribute to individual differences (e.g., gender, age, intelligence) also contribute to measurement heterogeneity. Thus, depending on the composition of the sample, a particular cognitive deficit may or may not be detectable.

Paradoxically, in spite of this heterogeneity, schizophrenia has defied meaningful subdivision. Most subtyping schemes are based on phenomenology rather than biology, lack temporal stability, and fail to provide an adequate account of the variability in neuroanatomic abnormalities, cognitive dysfunction, outcome, response to medications, and other signs of schizophrenia. The failure of phenomenology to map onto specific biological substrates suggests that it may not be the most valid organizing principle for illness heterogeneity. Identifying core cognitive deficits that can be mapped onto specific neural substrates may lead to more meaningful subtyping.

Variability

In addition to showing a wide range of performance across individuals, there is more variable cognitive performance within individuals, both within a single session (i.e., from trial to trial) and over separate sessions (i.e., test–retest) compared with demographically matched healthy individuals. For example, while performing a working memory task, patients with

schizophrenia show greater variability of response time within a single session and reduced test–retest reliability of brain activation across sessions. Even the demonstration of hand preference on simple tasks is less stable over two sessions. Increased variability and unreliability of performance are among the most consistent findings in schizophrenia research. Although frequently viewed as a measurement confound, it may be more productive to regard variability as intrinsic to schizophrenia and as having a neurological basis that requires explanation. This is supported by findings that frontal lobe damage in humans leads to increased variability of cognitive performance.

Amotivation

Amotivation is a core negative symptom of schizophrenia that makes it difficult to determine whether poor performance reflects a true deficit or that the patient was unwilling or unable to exert the effort necessary for optimal performance. Because tasks differ in the amount of effort required, amotivation may be more detrimental to the performance of some tasks than to others.

Medications

Antipsychotic medications have thus far targeted symptoms rather than cognition, and there is a vast and often contradictory literature on whether conventional antipsychotics impair cognition and whether atypical agents might improve it. Adjunctive anticholinergic drugs may also be detrimental to memory and other cognitive abilities.

Generalized Deficit

Meta-analytic studies demonstrate widespread cognitive dysfunction, suggesting that schizophrenia is characterized by a global blunting of performance. In addition to potential impairments attributable to amotivation and medication, patients may appear to have a generalized cognitive deficit because schizophrenia interferes with education and occupational attainment and many cognitive measurements are sensitive to these factors. In addition, acute psychosis and associated disruptions of attention may prevent optimal engagement in cognitive testing. For these reasons, simply demonstrating that patients perform worse on a particular task is not terribly revealing.

To some extent, the concept of a 'generalized deficit' is a straw man. Existing data do not support such a simplistic explanation. And a generalized deficit cannot account for the findings of highly selective cognitive deficits in unaffected relatives. Sorting out cognitive deficits that reflect the pathology of schizophrenia from those that are epiphenomena of having and being treated for a severe, chronic mental illness is an ongoing challenge. Strategies for identifying specific deficits in schizophrenia include demonstrating a pattern of both proficient and deficient performance across tasks that are matched for discriminating power and finding tasks on which a cognitive abnormality is actually advantageous to performance. Successful employment of this superiority strategy is illustrated by a series of numerosity studies in which the intact perceptual organizational ability of controls interfered with the rapid counting of elements (i e., perceptual gestalts had to be broken for elements to be counted). Because patients failed to form gestalts, they were faster at counting.

Are There Fundamental Deficits That Give Rise to Widespread Cognitive Disturbance?

The seeming generality of neurocognitive deficits in schizophrenia leaves one searching for a parsimonious explanation – are there more-basic problems that can organize and account for the diverse deficits seen? Some investigators have hypothesized that many cognitive deficits stem from deficient working memory, which leads to a failure to guide behavior on the basis of internalized representations such as schemata and ideas. Deficits earlier in the information processing pipeline, such as in sensorimotor gating, have been proposed to account for higher order difficulties in inhibiting thoughts, speech, and action. While a theoretical case can be made for a number of integrative explanations, empirical evidence of links between elementary and higher order deficits is lacking.

Attention

A wide range of cognitive deficits in schizophrenia could be viewed as failures of the effective deployment and control of attention. The term attention refers to a broad array of cognitive functions that enable the individual to select relevant aspects of either the internal or external milieu (i.e., thoughts or environmental stimuli) for further processing while keeping others at bay. Particularly during acute phases of schizophrenia, patients describe difficulties filtering the continuous barrage of stimuli present at every moment and focusing on what is relevant. For example, they may have difficulty selectively attending to the voice of the person addressing them rather than to noise from the refrigerator. Inattention, however, cannot account for findings of deficits when attention is controlled or for information processing deficits on measurements

that minimize its contribution, such as the auditory mismatch negativity, an event-related brain potential that is sensitive to stimulus deviation from a repetitive pattern and that is reduced in chronic schizophrenia.

Deficient Automation

In healthy individuals, practice can increase the speed of performance and reduce both the variability of responses and the error rate, reflecting a shift from controlled to more automatic processing and a corresponding shift in the brain networks that support performance. Tasks, or components of tasks, that have become automated proceed efficiently with reduced demands on attention. If schizophrenia were characterized by deficient automation, increased attention would be required for task components that should have been automated, and consequently fewer resources would be available for other, higher-order task demands. Although there are only a few studies of automation *per se* in schizophrenia, deficient automation provides a plausible account for increased variability and unreliability of performance across a range of tasks.

Context Processing

As illustrated in the section titled 'Executive function deficits,' the balance between past and present influences on behavior is upset in schizophrenia. Patients have trouble learning from error feedback, suggesting that the influence of previous experience is too weak. On the other hand, they also show perseveration of responses from trial to trial, indicating that the influence of the recent past is too strong and/or the influence of the current contingency is too weak. These difficulties in appropriately using information from both the past and the present to optimize behavior can be said to reflect deficient context processing. A number of executive functions are subsumed under the term 'context processing' (e.g., working memory since a representation of context has to be held online). Computational models have shown that a deficit in a single context processing module can mimic patient performance on a variety of cognitive tasks.

Promising Advances in the Study of Cognition

Imaging Genetics

An extremely promising avenue of research, made possible by advances in genetics, links variability in cognitive function and associated brain activation to genetic variation. This work is strengthening the case for suspected susceptibility genes and identifying potential mechanisms for the development of schizophrenia. For example, in groundbreaking work, Egan and colleagues in 2001 reported that a common functional polymorphism in the catechol-*O*-methyltransferase (COMT) gene, which reduces the amount of dopamine available in the synapses of the prefrontal cortex, and is slightly more common in individuals with schizophrenia than their unaffected family members, was associated with both poorer performance of the WCST and increased prefrontal cortical activation during performance of a working memory task in both patients with schizophrenia and healthy control participants. This suggest that prefrontal cortex function is influenced by genes that increase susceptibility to schizhophrenia. Increased prefrontal activation is thought to reflect inefficient function (i.e., increased resources are necessary to support a given level of performance). Several groups have shown that prefrontal inefficiency is also present in unaffected siblings of schizophrenia patients. The siblings showed greater activation than controls even though their task performance did not diffuse (Figure 4). These findings suggest that physiological indices derived from neuroimaging, because they more directly reflect the effects of genes (e.g., increased metabolism of dopamine), may prove more sensitive endophenotypes than behavioral ones.

Failures to find increased frequencies of the high risk COMT polymorphism in schizophrenia in some studies (e.g., of Asian populations) are not surprising given the number of factors that contribute to variability in association studies and point to the need for caution in generalizing these findings. In addition, particular polymorphisms account for only a small portion of the variance in cognitive performance and regional activation, as would be expected in a complex, multigenetic illness. A gene such as COMT likely interacts with other genetic and environmental factors to either diminish or exacerbate its effect on brain function, and investigations of the effects of gene–gene and gene–environment interactions on cognition are just getting under way.

Neuroimaging Advances

Neuroimaging findings continue to form the crux of many theoretical conceptualizations of schizophrenia, and continued technical advances hold tremendous promise for understanding its pathophsiology. The search for the neural basis of cognitive dysfunction in schizophrenia has evolved from key brain regions (e.g., prefrontal cortex or hippocampus) to also considering the integrity of neural circuits and

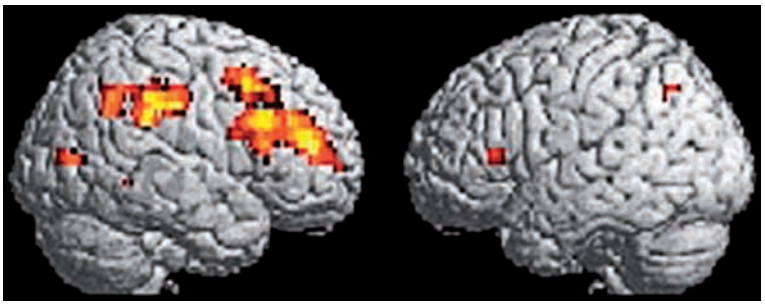

Figure 4 Regions of increased, inefficient fMRI activation in the unaffected siblings of schizophrenia patients relative to demographically matched control participants during performance of a working memory paradigm. Behavioral performance did not discriminate between the groups, but fMRI activation did, suggesting that imaging indices may be more sensitive to genetic risk for schizophrenia. Regions of increased activation include right dorsolateral and ventrolateral prefrontal cortices. Statistical group difference maps are rendered onto canonical single-participant lateral brain surfaces. Adapted from figure 2 in Callicott JH, Egan MF, Mattay VS, et al. (2003) Abnormal fMRI response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. *American Journal of Psychiatry* 160: 709–719.

the timing of neuronal processes across different regions. Advances in functional magnetic resonance imaging (fMRI) are providing ever higher spatial resolution. In combination with methods that provide high temporal resolution such as magnetoencephalography and electroencephalography, it is possible to examine the neural correlates of cognitive processes at each stage of performance and pinpoint exactly where and when they go awry. Complementary structural MRI can reveal morphological correlates of abnormal activation and cognitive performance. Since performance of even simple cognitive tasks is the product of coordinated activity in a distributed network, it is also of interest to assess the integrity of white matter, which can be done *in vivo* using a relatively new MRI technique, diffusion tensor imaging. Recent evidence suggests that white matter physiology contributes to individual differences in cognitive processing in health, and an emerging literature in schizophrenia links regionally reduced white matter integrity to cognitive dysfunction. By impairing communication in local circuitry and between connected brain regions, white matter pathology may contribute to cognitive dysfunction in schizophrenia. Together, these techniques can identify the functional and structural correlates of cognitive dysfunction, guide the search for neuropathology, and provide targets for intervention aimed at improving cognition in schizophrenia.

See also: Cognitive Dysfunction in Psychiatric Disorders; Executive Functions: Eye Movements and Neuropsychiatric Disorders; Inhibitory Control over Action and Memory; Psychiatric Disorders: Functional Genomics; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Schizophrenia: Genetics.

Further Reading

Callicott JH, Egan MF, Mattay VS, et al. (2003) Abnormal fMRI response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. *American Journal of Psychiatry* 160: 709–719.

Cirillo MA and Seidman LJ (2003) Verbal declarative memory dysfunction in schizophrenia: From clinical assessment to genetics and brain mechanisms. *Neuropsychology Review* 13: 43–77.

Egan MF, Goldberg TE, Kolachana BS, et al. (2001) Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 98: 6917–6922.

Gottesman II and Gould TD (2003) The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry* 160: 636–645.

Gould TD and Gottesman II (2006) Psychiatric endophenotypes and the development of valid animal models. *Genes, Brain, and Behaviour* 5: 113–119.

Green MF (1998) *Schizophrenia from a Neurocognitive Perspective*. Boston: Allyn and Bacon.

Green MF (2001) *Schizophrenia Revealed: From Neurons to Social Interactions*. New York: W. W. Norton.

Heinrichs RW (2005) The primacy of cognition in schizophrenia. *American Psychologist* 60: 229–242.

Lee J and Park S (2005) Working memory impairments in schizophrenia: A meta-analysis. *Journal of Abnormal Psychology* 114: 599–611.

Knight RA and Silverstein SM (2001) A process-oriented approach for averting confounds resulting from general performance deficiencies in schizophrenia. *Journal of Abnormal Psychology* 110: 15–30.

Kuperberg G and Caplan D (2003) Language dysfunction in schizophrenia. In: Schiffer RB, Rao SM, and Fogel BS (eds.) *Neuropsychiatry*, 2nd edn., pp. 444–466. Philadelphia: Lippincott Williams and Wilkins.

Manoach DS (2003) Prefrontal cortex dysfunction during working memory performance in schizophrenia: Reconciling discrepant findings. *Schizophrenia Research* 60: 285–298.

Mathalon DH, Fedor M, Faustman WO, et al. (2002) Response-monitoring dysfunction in schizophrenia: An event-related brain potential study. *Journal of Abnormal Psychology* 111: 22–41.

Montague PR, Hyman SE, and Cohen JD (2004) Computational roles for dopamine in behavioural control. *Nature* 431: 760–767.

Nuechterlein KH, Barch DM, Gold JM, et al. (2004) Identification of separable cognitive factors in schizophrenia. *Schizophrenia Research* 72: 29–39.

Snitz BE, Macdonald AW III, and Carter CS (2006) Cognitive deficits in unaffected first-degree relatives of schizophrenia patients: A meta-analytic review of putative endophenotypes. *Schizophrenia Bulletin* 32: 179–194.

Stuss DT, Murphy KJ, Binns MA, et al. (2003) Staying on the job: The frontal lobes control individual performance variability. *Brain: A Journal of Neurology* 126: 2363–2380.

Relevant Website

http://www.schizophreniaforum.org – Schizophrenia Research Forum.

Cognitive Dysfunction in Psychiatric Disorders

D M Barch, Washington University in St. Louis, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Schizophrenia is an extremely disabling neuropsychiatric disorder that affects approximately 1% of the general population. Researchers and theorists have long recognized that abnormalities in cognitive function are a key component of schizophrenia, and the past two decades have witnessed a relative explosion of research on cognition in schizophrenia, much of it couched within the framework of understanding the cognitive neuroscience of schizophrenia. This emphasis is partly due to the research suggesting that cognitive function in schizophrenia is one of the most critical determinants of quality of life in schizophrenia and because of the hope that understanding the neurobiology of cognitive dysfunction in schizophrenia will provide clues to the etiology of this disorder. Research on cognitive function in individuals with schizophrenia has identified deficits in many different cognitive domains, and it is beyond the scope of this article to review the literature relevant to all possible cognitive deficits in schizophrenia. Instead, this article focuses on the three cognitive domains that received the majority of research attention and consistent evidence suggesting impairment in schizophrenia: (1) working memory, (2) executive control, and (3) episodic memory.

Working Memory

Working memory is typically defined as the ability to maintain and manipulate information over short periods of time. It is widely agreed that working memory involves several different component processes. Baddeley's influential theory of working memory distinguishes among four major subcomponents: (1) a short-term storage buffer for visual information that is often referred to as the visual–spatial scratch pad, (2) a short-term storage buffer for verbal information referred to as the phonological loop that includes articulatory rehearsal and phonological storage, (3) a central executive component that guides the manipulation and transformation of information held within the storage buffers, and (4) the recently described episodic buffer. A number of studies suggest that articulatory rehearsal is particularly dependent on regions of left ventrolateral prefrontal cortex (VLPFC), including Brodmann's areas 44 and 45. Figure 1 shows the location of these regions. In contrast, the processing or storage of phonological representation is thought to dependent on regions of left posterior parietal cortex (PPC) (see Figure 1), again based on data from both lesions studies and imaging studies in healthy humans.

The specific component processes of the visual–spatial scratch pad are less clear, although it has been hypothesized that we use covert shifts of attention to the spatial locations to be remembered, a process that has been referred to as attention-based rehearsal, thought to be supported by regions such as right PPC. Studies of spatial working memory also consistently activate regions such as the frontal eye fields and the supplementary eye fields, regions in which information about the direction of visual saccades may be processed and/or maintained. There are a host of different processes that may fall into the central executive, including those involved in the manipulation of information being stored in the buffers, protection from interference due to competing information or decay across time, temporal coding or sequence, and updating of the contents of working memory. Many of these have been assumed to be supported by the dorsolateral regions of prefrontal cortex (DLPFC; see Figure 1), typically Brodmann's areas 46 and 9 bilaterally. In addition, the dopamine system is thought to be critical for a range of working memory functions.

Working Memory in Schizophrenia

Phonological Loop

Research on working memory in schizophrenia has suggested that the operation of the phonological loop is relatively intact in this disorder. For example, a number of studies have shown that individuals with schizophrenia demonstrate (1) relatively intact performance on digit span forward tasks, particularly when the number of items is at or below working memory span (7 ± 2) and when there is no verbal interference; (2) no disproportionate impairment for recall of lists with phonologically similar versus dissimilar items, suggesting an intact ability to represent phonological information; (3) intact serial position curves among individuals with schizophrenia; and (4) equal impairments on verbal and nonverbal working memory tasks. However, some studies have suggested greater impairments in verbal than tone working memory and a small but significant impairment in digit span on meta-analysis.

The vast majority of the working memory functional activation studies report intact activation of VLPFC regions during working memory performance in individuals with schizophrenia, one of the regions

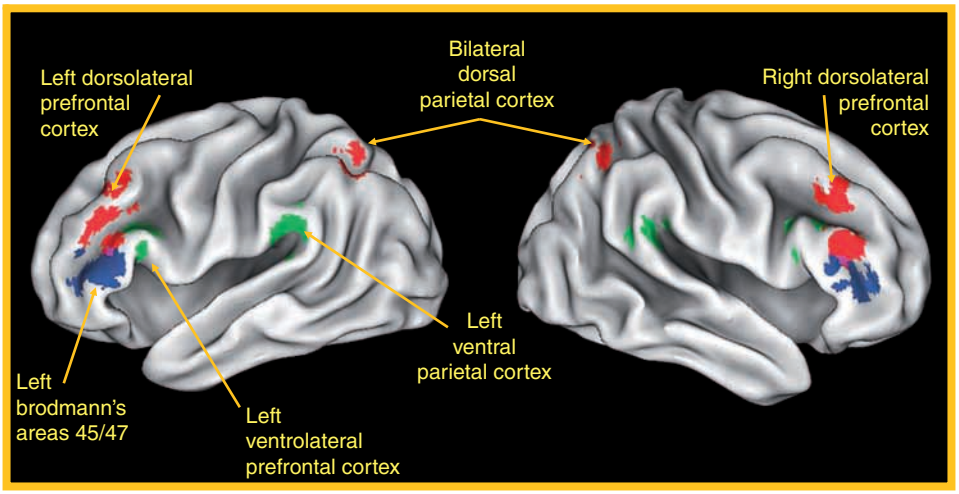

Figure 1 Frontal and parietal regions involved in the central executive and phonological loop function of working memory, and frontal regions involved in semantic processing. Regions plotted in red are the dorsolateral prefrontal and dorsal parietal regions thought to support a range of central executive functions. The left hemisphere regions plotted in green are the ventrolateral prefrontal and parietal regions thought to support phonological loop function. The left hemisphere inferior prefrontal regions plotted in blue are those thought to be involved in semantic processing.

thought to support the phonological loop. These findings are consistent with results of postmortem work which did not find cell density changes in Brodmann's area 44, although such changes are found in Brodmann's area 9.

A number of these same studies report abnormal activation in PPC in schizophrenia, in terms of either the degree of activation or connectivity with other brain regions. However, there are two dissociable regions of PPC activated by working memory tasks, with a left ventral PPC region (see Figure 1) sensitive to the type of information (verbal vs. nonverbal) that may support phonological storage. More dorsal and bilateral regions of PPC (see Figure 1) are sensitive to load irrespective of material type, and they may play a role in the maintenance or updating of information in working memory across stimulus domains. To date, it is not clear whether the PPC regions showing abnormal activation in schizophrenia should be characterized as ventral or dorsal. However, work specifically focused on this question does suggest impaired dorsal PPC activation in schizophrenia but intact ventral PPC activity.

Visual–Spatial Scratch Pad

In terms of the visual-spatial scratch pad in schizophrenia, there is certainly a large body of evidence for impairments on visual–spatial working memory tasks and on classic measures of the visual–spatial scratch pad, such as memory-guided saccade performance. There are relatively few functional neuroimaging studies in schizophrenia that have used visual–spatial working memory tasks that could not be accomplished by verbal recoding (small numbers of spatial locations are typically verbally labeled by participants, turning the task into a verbal working memory task). However, work using an occulomotor delayed response task demonstrated impaired activity in the frontal eye fields, supplementary eye fields, and inferior parietal lobule among individuals with schizophrenia – regions that may support visual scratch pad function.

Central Executive Function

There is very consistent evidence that individuals with schizophrenia have difficulty with processes that are part of the central executive component of working memory. For example, individuals with schizophrenia show deficits on (1) working memory tasks with all different material types, with little evidence for differential deficits across material types; and (2) measures of a range of central executive functions, including manipulation, interference control and/or dual-task coordination, and information updating and temporal indexing. In addition, the first-degree relatives of individuals with schizophrenia also show evidence of consistent impairments in central executive components of working memory.

As described previously, the cognitive neuroscience literature has linked many of the central executive components of working memory to DLPFC function. Many, although not all, functional neuroimaging studies using tasks that engage central executive components of working memory have found disturbed DLPFC activation in schizophrenia. These studies

have used tasks such as the Wisconsin Card Sorting Task, mental arithmetic, self-ordered pointing, the Sternberg Item Recognition Task, and various versions of the Nback and delayed response tasks. The most common finding on these tasks is that individuals with schizophrenia show reduced DLPFC activation. However, several studies have reported increased activation in DLPFC among individuals with schizophrenia that sometimes co-occurred with decreased DLPFC activation in different regions. One interesting hypothesis put forth to explain these differential findings of hypo-versus hyper-DLPFC activity is the idea that the memory load–DLPFC response curve may be different for individuals with schizophrenia compared to healthy controls. Specifically, it may be that at low load levels, individuals with schizophrenia need to devote more resources to working memory, leading to hyperactivation, but that higher load levels exceed their capacity, leading to reduced DLPFC activity. However, relatively few studies have reported such hyper-DLPFC activation in schizophrenia, and the regions that show hyperactivation tend not to overlap for the most part with regions showing hypoactivation because they are located either anterior (particularly on the right) or more inferior (particularly on the left).

Although many studies of working memory in schizophrenia report abnormalities in the activation of specific regions during working memory tasks, a growing number of studies suggest that individuals with schizophrenia demonstrate altered patterns of connectivity between DLPFC and other working memory-related regions such as PPC and temporal cortex. Furthermore, there is evidence from methods other than functional imaging for altered DLPFC function in schizophrenia, including (1) reduced gray matter volume; (2) cellular and molecular abnormalities; (3) reduced *N*-acetylaspartate concentrations, particularly in dorsal regions; and (4) a relationship between poor working memory performance (as measured by the Nback task) and altered D1 receptor availability in DLPFC. Although there have been negative findings in each of these areas as well, the weight of the evidence suggests structural, cellular, and molecular abnormalities as well as functional abnormalities in the DLPFC of individuals with schizophrenia.

Executive Control in Schizophrenia

Context Processing

Individuals with schizophrenia also show impairments in other aspects of executive control thought to be dependent on DLPFC and potentially dopamine function. For examine, a growing body of research suggests that individuals with schizophrenia show deficits in the ability to represent and maintain context information, a process thought to be dependent on intact function of dopamine in DLPFC. A number of studies have demonstrated impaired context processing in schizophrenia, in individuals at risk for schizophrenia, and in individuals with schizotypal personality disorder. In addition, a number of studies have demonstrated impaired DLPFC activity during performance of context processing tasks in schizophrenia among both medication naïve first-episode patients and chronic patients.

Conflict Detection/Error Monitoring

An alternative (or complimentary) theory suggests that disturbances in the ability to detect conflict or errors in ongoing processing, due at least in part to the function of anterior cingulate cortex (ACC), may lead to deficits in the ability to recruit and regulate cognitive control functions that may be subserved at least in part by DLPFC regions. There is mixed behavioral evidence in schizophrenia regarding impairments on indices thought to reflect the ability to detect or respond to conflict or errors, with some studies finding evidence for impairments and others not. However, event-related potential (ERP), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET) studies have provided more consistent evidence for altered conflict/error processing and abnormal ACC function in schizophrenia. There is also evidence for ACC abnormalities in schizophrenia from structural and postmortem studies. However, only a handful of the functional neuroimaging studies on working memory in schizophrenia have reported abnormal ACC activity, and most of these have been in the direction of greater ACC activity. It is possible that such results reflect the increased errors and conflict experienced by individuals during ongoing cognitive processing, which would be expected to lead to enhanced ACC activity. Further research is needed to clarify the nature of functional abnormalities in ACC among individuals with schizophrenia and to clarify their relationship to DLPFC deficits.

Episodic Memory

As with trying to understand the nature of working memory deficits in schizophrenia, it is useful to briefly discuss the cognitive neuroscience literature on the processing and brain regions involved in episodic memory. Research has long indicated the importance of the hippocampal formation in the formation

of long-term memories, with most theories of hippocampal function focusing on its role in the rapid binding of novel configurations of information. These theories have been supported by both human functional imaging studies and work in amnestic patients. However, it is also very clear that PFC structures also make important contributions to episodic memory. Damage to the PFC can also lead to episodic memory deficits, although these deficits are not selective to episodic memory. It has been hypothesized that PFC regions contribute to episodic memory via their strategic contributions to memory formation and retrieval. Consistent with this hypothesis, studies have shown activation of prefrontal regions such as Brodmann's areas 45 and 47 when participants are asked to process verbal information using semantic elaboration strategies and increased activation during encoding in frontal regions such as Brodmann's areas 45 and 47 are very strongly predictive of subsequent memory performance. Work on episodic memory has also distinguished between recollective versus familiarity-based processing in episodic memory. Recollection is the ability to retrieve information associated with specific learning episodes, whereas familiarity refers to feelings of knowing that may be related to prior exposure to information. Imaging and lesion data suggest that the DLPFC is important for recollection compared to familiarity, although there is opposing evidence. In contrast, imaging and lesion data suggest that the hippocampus is important for both recollection and familiarity, although there are conflicting views on this.

Episodic Memory in Schizophrenia

Binding Functions

A meta-analysis comparing performance on associative and item memory tests in individuals with schizophrenia concluded that there was evidence for a 20% greater impairment in associative compared to item memory in individuals with schizophrenia. However, few of the studies that have compared item and associative memory have dealt with the ubiquitous problems of discriminating power. Recent studies have used tasks derived from the animal literature on hippocampal function, such as the transitive interference test. Other research has shown that individuals with schizophrenia are impaired on the critical conditions of this task that require relational processing but not on conditions that require the learning of simpler associative reinforcement mappings. Taken together, these results suggest impairments in relational processing or binding in schizophrenia of the type thought to be specifically supported by the hippocampal formation.

Many recent functional imaging studies of episodic memory have provided relatively consistent evidence for abnormal hippocampal activation in schizophrenia, at both encoding and retrieval. These failures to show task-related hippocampal activity at retrieval have been interpreted as reflecting a failure in explicit recollection among individuals with schizophrenia. There are a number of other sources of evidence that individuals with schizophrenia have deficits in hippocampal volume and shape. The evidence for hippocampal abnormalities in schizophrenia extends to the cellular level as well, with much theoretical and empirical work focusing on *N*-methyl-D-aspartate (NMDA) receptor hypofunction. NMDA receptors are dense in the hippocampus, and NMDA receptor antagonists disrupt hippocampal long-term potentiation, one of the mechanisms thought to support memory encoding and/or binding. In humans, NMDA receptor antagonists such as ketamine or phenycyclidine (PCP) can impair episodic memory and elicit many clinical phenomena analogous to the symptoms of schizophrenia, including hallucinations, delusions, and disorganized speech. Thus, many theorists have argued that such NMDA receptor hypofunction contributes to hippocampal abnormalities and episodic memory deficit in schizophrenia.

Impairments in Strategic Contributions to Episodic Memory

A number of researchers have hypothesized that individuals with schizophrenia experience deficits in the use of beneficial strategies that may be supported by PFC. A number of studies suggest that individuals with schizophrenia are impaired in their ability to generate effective mnemonic strategies and encode information less elaborately than controls. However, when provided with strategies that promote successful episodic encoding, individuals with schizophrenia are typically able to benefit as much as controls from these strategies. Importantly, individuals with schizophrenia show evidence of impaired recollection but relatively intact familiarity, which may relate to both binding deficits and strategic impairments. For example, when individuals with schizophrenia are asked to make judgments about the source of their responses, they consistently provide fewer responses indicative of recollective experiences.

Prefrontal Abnormalities in Schizophrenia during Episodic Memory

Many of the functional imaging studies of episodic memory in schizophrenia have demonstrated abnormalities in a range of prefrontal regions, including both decreased and increased activity. These studies have revealed consistent abnormalities in PFC function, but

in different regions than those found to be impaired in studies of working memory in schizophrenia. For example, episodic memory studies of schizophrenia consistently find reduced activity in Brodmann's areas 47 and 45 (Figure 1), regions associated with semantic elaboration or encoding of information in episodic memory. Abnormal activity in these regions may reflect deficits in the ability to generate or apply effective encoding or retrieval strategies that utilize semantic processing. However, people with schizophrenia can appropriately engage these regions when given external support in the use of effective encoding strategies.

Specificity of Cognitive Deficits to Schizophrenia

A number of studies have examined whether the pattern or profile of cognitive impairments found in schizophrenia are specific to this disorder or found in other psychiatric disorders as well. There are several levels at which one could address the question of specificity. One could ask whether the pattern of cognitive deficits and associated neurobiological abnormalities is specific to schizophrenia as opposed to any psychiatric disorder, such as nonpsychotic major depression. At the behavioral level, a number of studies have shown that individuals with nonpsychotic major depression do not show the same impairments (either no deficit or less of a deficit) in working memory and executive function as shown by individuals with schizophrenia, although individuals with major depression do show impairments in episodic memory function. Furthermore, individuals with major depression do not show the same impairments in DLPFC function during working memory and/or executive control tasks as found in schizophrenia. Individuals with bipolar disorder can show deficits in many of the same domains as individuals with schizophrenia, particularly in terms of working memory and episodic memory, with impairments more severe during manic than euthymic states. However, there is evidence that these impairments tend to be less severe than those found in schizophrenia. There has been little work directly comparing individuals with schizophrenia and those with bipolar disorder on functional brain activation during cognitive performance. However, there is evidence for reduced anterior middle frontal gyrus activity among euthymic bipolar patients performing a Stroop interference control task.

It is more difficult to distinguish between individuals with schizophrenia and those with other psychotic disorders (e.g., schizoaffective, brief psychotic disorder, delusional disorder) in terms of behavioral impairments, although there are data suggesting that cognitive deficits are more stable in individuals with schizophrenia compared to those with other psychotic disorders. There is preliminary evidence that impairments in DLPFC function are more severe among individuals with schizophrenia. Further work comparing both behavioral and functional brain activation in a range of cognitive domains between individuals with schizophrenia and individuals with other types of psychosis (both affective and nonaffective) will help to clarify the specificity of cognitive deficits and functional brain activity disturbances found in schizophrenia.

Summary

Individuals with schizophrenia show deficits in a range of cognitive domains; disturbances are particularly apparent in working memory, executive function, and episodic memory. Within the domain of working memory, individuals with schizophrenia show relatively little evidence of impairments in the phonological loop and relatively little impairment in activation of ventral prefrontal cortex regions (e.g., Brodmann's area 44) or dorsal PPC regions associated with this subcomponent of working memory. Individuals with schizophrenia do not show evidence of differential impairments in spatial working memory, but they do show evidence of impaired visual–spatial scratch pad function and altered dorsal PPC activity, a region that may support the visual–spatial buffer. There is consistent evidence for impairments in the central executive components of working memory, associated with impaired DLPFC and dorsal PPC functions. Individuals with schizophrenia also show clear evidence of impairments in other types of executive controls functions associated with DLPFC, such as context processing. There is mixed behavioral and imaging evidence for impairments in error detection and conflict processing associated with ACC function in this disorder. In the domain of episodic memory, individuals with schizophrenia show deficits suggestive of impairments in both relational processing and the generation and application of effective encoding strategies. Furthermore, individuals with schizophrenia show evidence of alterations in hippocampal activity, as well as impairments in the activation of regions associated with semantic encoding and successful subsequent memory retrieval. The types of cognitive impairments and changes in functional brain activation found in schizophrenia are dissociable from those found in depression and affective psychoses (e.g., bipolar disorder). However, further research is needed to determine what cognitive and functional brain abnormalities are specific to schizophrenia compared to other nonaffective psychoses.

See also: Cognition: Cerebellum Role; Cognitive Deficits in Schizophrenia; Episodic Memory; Executive Function and Higher-Order Cognition: Neuroimaging; Prefrontal Cortex: Structure and Anatomy; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Short Term and Working Memory.

Further Reading

Achim AM and Lepage M (2005) Episodic memory-related activation in schizophrenia: Meta-analysis. *British Journal of Psychiatry* 187: 500–509.

Barch DM (2005) The cognitive neuroscience of schizophrenia. *Annual Review of Clinical Psychology* 1: 321–353.

Braver TS, Barch DM, and Cohen JD (1999) Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry* 46: 312–328.

Danion JM, Rizzo L, and Bruant A (1999) Functional mechanisms underlying impaired recognition memory and conscious awareness in patients with schizophrenia. *Archives of General Psychiatry* 56(7): 639–644.

Davidson LL and Heinrichs RW (2003) Quantification of frontal and temporal lobe brain-imaging findings in schizophrenia: A meta-analysis. *Psychiatry Research* 122(2): 69–87.

Green MF, Kern RS, Braff DL, and Mintz J (2000) Neurocognitive deficits and functional outcome in schizophrenia: Are we measuring the 'right stuff'? *Schizophrenia Bulletin Special Issue: Psychosocial Treatment for Schizophrenia* 26(1): 119–136.

Manoach DS (2003) Prefrontal cortex dysfunction during working memory performance in schizophrenia: Reconciling discrepant findings. *Schizophrenia Research* 60(2–3): 285–298.

Meyer-Lindenberg AS, Olsen RK, Kohn PD, et al. (2005) Regionally specific disturbance of dorsolateral prefrontal–hippocampal functional connectivity in schizophrenia. *Archives of General Psychiatry* 62(4): 379–386.

Olney JW, Newcomer JW, and Farber NB (1999) NMDA receptor hypofunction model of schizophrenia. *Journal of Psychiatric Research* 33(6): 523–533.

Snitz BE, Macdonald AW III, and Carter CS (2006) Cognitive deficits in unaffected first-degree relatives of schizophrenia patients: A meta-analytic review of putative endophenotypes. *Schizophrenia Bulletin* 32(1): 179–194.

Weiss AP, Schacter DL, Goff DC, et al. (2003) Impaired hippocampal recruitment during normal modulation of memory performance in schizophrenia. *Biological Psychiatry* 53: 48–55.

Yonelinas AP, Otten LJ, Shaw KN, and Rugg MD (2005) Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience* 25(11): 3002–3008.

Cognitive Neuroscience: An Overview

C M Wessinger and E Clapham, University of Nevada, Reno, NV, USA

© 2009 Elsevier Ltd. All rights reserved.

Understanding how brain enables mind is the goal of cognitive neuroscience. More broadly stated, cognitive neuroscience is concerned with understanding the neurobiological basis of mental processes – in other words, thinking about how the brain is involved in thinking. Thinking about thinking. There is little doubt that human brains can engage in the task of thinking about thinking and likely have been able to do so for some time. Unfortunately, throughout most of human history, more practical concerns necessarily occupied human thought. Humans were more concerned with thinking about and acting on behaviors that were related to their survival: storing enough food to enable them to survive the winter, killing the bear before the bear killed them, and finding shelter. They had to concentrate on understanding and working within their environment in order to survive to struggle another day. Anything that did not specifically contribute to their survival was a luxury. Thus, thinking about thinking would have been a luxury – a luxury that humans could not afford. However, as soon as civilization developed to a point that day-to-day survival did not occupy every waking thought, humans could afford to think about thinking. They began to think about how others thought and why they did what they did. Oedipus Rex, the ancient Greek play that deals with parent–child conflicts, as well as Mesopotamian and Egyptian theories designed to explain religion and the universe are excellent early attempts to understand why people act the way they do. Unfortunately, there are significant pieces of the equation missing. These thinkers and philosophers did not systematically explore the relationship between brain, behavior, and mind. This is where modern-day science and scientific methodology come into play. The notion that we must observe, manipulate, and measure the environment in order to understand the world and our place within it is central to scientific methodology. When we apply such techniques to understanding how the brain enables mind, we call that cognitive neuroscience.

Certainly, one could make a case that cognitive neuroscience has been around ever since scientists in different disciplines began exploring the common goal of understanding how brain enables mind. Such early explorations were quite illuminating and many have stood the test of time. Some early theories that specific areas of the brain gave rise to specific functions were put forth by the phrenologists Franz Joseph Gall and J G Spurzheim, who declared that the brain was organized around 35 distinct functions ranging from such basic cognitive functions as language and color perception to more philosophical concerns such as hope and self-esteem. Each of these areas was supported by a specific brain region, and with increased proficiency in one of these areas, the corresponding brain representation would grow in size, in turn causing a corresponding bump on the skull. However, once subjected to strict scientific methodology, phrenology was shown to be untrue. It took the great neurologist John Hughlings Jackson to again substantiate that a structure–function relationship between the brain and behavior existed. Rather than being associated with bumps on the skull, his hypotheses were based on careful clinical observations of patients. While seizing, some of his epileptic patients would move in characteristic patterns, almost with a systematic rhythm. This led him to propose a somatotopic organization – that is, a map of the body in the brain. Also at approximately this time, the French neurologist, Paul Broca, reported perhaps the most famous neurological case – one of his patients had suffered a stroke and now was unable to speak but was still able to understand speech. Postmortem examination showed damage in the left frontal area, which is now known as Broca's area and still believed to play a major role in language output. In contrast, German neurologist Carl Wernicke reported a patient who babbled incessantly following a stroke, but that is all it was – babbling – the words kept coming but made no sense. Wernicke's patient had a lesion in a more posterior region in the left hemisphere, which has become known as Wernicke's area, and it is still believed to play a major role in understanding language. Neuroanatomists were realizing similar divisions of the brain, but on a more cellular level. Perhaps the most famous is Korbinian Brodmann, who analyzed cellular organization of the cerebral cortex using a Franz Nissl developed stain to visualize the different layers of cortical neurons, dividing the cerebral cortex into more than 50 distinct regions. This method of looking at cellular differences in cortical layers is known as cytoarchitectonics. Many others contributed to this work, including Constantin Von Economo, Gerhardt von Bonin, and Percival Bailey. More recently, similar techniques have been used to demonstrate more than 30 different cortical areas in the visual system alone. Such early

investigations still guide cognitive neuroscience today, particularly with regard to helping the field understand structure–function relations within the brain and how these combine to enable mind.

The formal birth of the field known as cognitive neuroscience was relatively recent. One story as to how the field developed involved a taxi ride shared by two eminent brain scientists – biopsychologist Michael S Gazzaniga and behavioral-turned-cognitive psychologist George Miller – on the way to a dinner meeting in the early 1970s. The dinner was being held so scientists from many different disciplines could discuss how to pursue a common goal: understanding how the brain enables the mind. Such a joining of forces deserved an appropriate name – not just a renaming of biopsychology or cognitive neuropsychology, well-respected disciplines already striving to understand the structure–function relations of the brain–mind problem, but a name that incorporated the investigations of information processing akin to cognitive science: one that reflected a merging of all of these many disciplines and more. After much discussion, the term 'cognitive neuroscience' seemed ideal. Another story concerning the birth of the field involves a similar meeting with other eminent brain scientists such as Larry Squire, a neuroscientist renowned for his research exploring the biological basis of memory, and Stephen Kosslyn, a cognitive psychologist well-known for his mental imagery research. However, the exact events that led to the adoption of the term cognitive neuroscience are not important. What is important is the fact that many scientists using quite varied methodologies are working together, and making headway, toward the common goal of cracking the code of the brain and understanding how this approximately 6-pound lump of soft, squishy, neuronal tissue allows us to think and feel and emote and cry and to dream – how the brain enables mind. Many of the varied methodologies that contribute to this exciting and rapidly expanding field are described here.

In order to help us understand the inherently multidisciplinary approach of cognitive neuroscience, we need to understand some of the key methodologies. Techniques ranging from exploring the molecular basis of the brain to exploring consciousness are utilized, combined, dissected, and manipulated by researchers in the quest to understand the neurobiological underpinnings of mental operations. However, cognitive neuroscience is more than a simple recombination of various methodologies and techniques. The field is extremely innovative in that it borrows the best techniques from the best disciplines in order to provide the best innovative insights into how brain enables mind. Such innovation can be more fully appreciated by exploring some of the key methodologies and techniques.

One key avenue in understanding how brain enables mind is grounded in understanding the relationship between brain anatomy and cognition. Many studies have tried to relate differences in intelligence to differences in brain size and structure. Differences in neuronal type, size, density, and connectivity have also been explored in trying to understand structure–function relations between brain anatomy and cognition. These explorations of brain structure can be loosely grouped under the term neuroanatomy. Formally, neuroanatomy is the study of structure of the nervous system. Primary questions of interest to neuroanatomy are identifying the various parts and components of the nervous system, as well as describing how these parts and components are interconnected. To further complicate matters, it is necessary to use specialized techniques and methods to ask these questions at multiple levels. Not only is it important to know the size and shape of a neuron but also it is important to know where that neuron connects. Where does the input come from and where does the output go? This very debate was truly a key development in the birth of neuroscience, and neuroscience is key to cognitive neuroscience. Two brilliant neuroanomists were debating the interconnectivity of the nervous system. Camillo Golgi invented a silver stain that permitted visualization of single neurons. Using Golgi's stain, Santiago Ramón y Cajal demonstrated that the neurons were discretely bound, individual entities that did not touch. This view was contrary to that of Golgi, who believed the whole brain was a continuous mass of tissue sharing a common cytoplasm, or syncytium. Cajal extended the field by being the first to identify the unidirectional (from axon to dendrite) nature of electrical transmission within neurons. These findings were elucidated with early microanatomical techniques. Microanatomy or fine anatomy has now developed to the point of exploring the organization of neurons and the connections. This is opposed to gross anatomy, which is concerned with the general structures and connections visible to the naked eye. Gross and microanatomical techniques continue to inform the field of cognitive neuroscience by providing essential knowledge concerning brain and neuronal structure and connectivity.

Another key technique employed by cognitive neuroscientists that is grounded in neuroscience is physiology. This is a branch of neuroscience that deals with the functions and processes of life, as well as the physical actions and reactions associated with these processes. Simply stated, physiology measures bodily responses such as blood pressure, heart rate, and

neuronal electrical activity to the environment. One physiological measure of great import is neuronal electrical activity. Despite the fact that the electrical potential is extremely small, when large numbers of neurons are active simultaneously, they produce enough electrical signals that when summed can be measured by electrodes placed on the scalp. Electroencephalography (EEG) involves recording a continuous stream of overall brain activity while the person performs a variety of tasks, sometimes in an experimental setting and sometimes as part of everyday life. These continuous recordings of continuous activity have proven extremely useful clinically and have helped with epilepsy, sleep disorders, and anxiety.

EEG can only provide limited information when exploring cognitive processes because the recordings tend to reflect global brain activity, as opposed to specific cognitive or perceptual processes and functions. However, by synchronizing EEG collection with specific cognitive and perceptual events, it is possible to disentangle the brain waves and see changes in electrical activity associated with changes in cognitive and perceptual processes. This technique is called event-related potential (ERP) because the neuronal potentials are related to specific events. Given that neurons produce unidirectional electrical transmission and constraints of Farady's law of induction (which states that a changing electric field produces a magnetic field orthogonal to the direction of flow of electricity), they produce miniscule magnetic fields. Magnetoencephalography (MEG) records these summed magnetic potentials at the scalp in a manner similar to ERPs.

One great advantage of these techniques is that millisecond accuracy is possible in terms of linking summed electrical signals specific to an event; thus, cognitive neuroscientists can understand exactly when and how the brain is responding to a specific event. One major drawback of these techniques is the very notion of summed electrical signals. Such summing results in limited spatial resolution. Furthermore, because the summed signals must be collected at the scalp, localization of depth is also problematic. It is difficult to determine where in the brain the signals occur.

This problem can be overcome by taking advantage of the inherently multidisciplinary nature of cognitive neuroscience by bringing other techniques to bear. One technique that can help inform EEG, MEG, and a variety of other cognitive neuroscience techniques is magnetic resonance imaging (MRI). Formally, MRI is a branch of neuroimaging that is a clinical specialty that uses noninvasive techniques to produce images of the brain. Interestingly, MRI also takes advantage of Faraday's law in that it uses a supercooled, superconducting electromagnet to produce a strong electrical field. This magnetic field is then systematically manipulated to produce images of the body. These images currently provide great detail concerning the gross anatomy, and even microanatomy, of the nervous system. Such information, when coupled with ERP and/or MEG techniques, can provide essential structure–function relations within the human brain. Such techniques can help surgeons understand not only tumor location but also functions compromised by the tumor and weigh them against functions that may be compromised by the surgery.

A recent interesting use of MRI in conjunction with other techniques involves trying to understand the biological basis of autism, an increasingly more problematic developmental disorder. Hutsler and colleagues showed that increases in dendritic spine density could be related to increases in autistic symptoms. Certainly, these data do not indicate that dendritic spine density is responsible for autism, just that there is some connection, and this connection needs to be investigated further. This is an excellent example of the utility of cognitive neuroscience: by combining neurology, neuropsychology, radiology, and microneuroanatomy, researchers are making tremendous strides in understanding the biological basis of autism.

Within the past 15 years, MRI has been extended to include a functional component. By taking advantage of what from a radiologist's viewpoint is noise, cognitive neuroscientists are able to see the brain in action – almost. It is well-known that changes in cognitive and perceptual activity result in localized changes in neuronal activity, which result in localized changes in metabolic needs, which result in localized changes in blood flow, which result in localized changes in blood oxygenation. The end of this cascade, the localized changes in blood oxygenation, can be visualized with functional magnetic resonance imaging (fMRI) in order to infer localization of function. Also, when MRI signals are time locked to cognitive and perceptual changes, localization of function, as well as localization of time, can be inferred. This is similar to positron emission tomography (PET), another neuroimaging technique that can image the blood flow change portion of the cascade to infer localization of function, except that PET uses ionizing radiation and fMRI does not. Another advantage that fMRI has over PET is greater spatial and temporal resolution. PET provides resolution in terms of minutes and centimeters for temporal and spatial properties, respectively. However, fMRI, particularly event-related fMRI, can provide temporal accuracy in seconds and spatial accuracy in millimeters.

So far, we have reviewed several biologically based techniques and have a greater understanding of the neuroscience part of cognitive neuroscience. But what about the cognitive part? Is there really a need for more when trying to understand how brain enables mind? Certainly. Without the 'cognitive,' there would be no true understanding of the mind portion of the equation.

Cognition can be defined as any form of information processing, mental operation, or intellectual activity such as thinking, reasoning, remembering, imagining, or learning. Thus, cognitive psychology, a major contributor to the field of cognitive neuroscience, is the study of behavior related to cognition. The principles and goals of cognitive psychology rest on two key concepts. First, mental operations depend on internal representations. Second, mental operations undergo transformations.

The notion that mental operations depend on internal representations is not that far-fetched; after all, by definition, if the operations are mental and taking place in the mind, then they are necessarily internal. It is also easy to accept and understand the notion that mental operations have to undergo transformations. After all, we do not directly perceive and act in the world; rather, our perceptions, thoughts, and actions depend on what our sensory systems are designed to detect within our environment, and our senses are not capable of translating everything in our environment into useable information. For example, we do not readily see infrared light or directly sense changing electrical and magnetic fields, but these physical properties do exist in the environment – they are just not detectable by our sensory systems. We need to use other physical-detection tools specifically designed to detect these aspects of our environment. Nonetheless, we still perceive, think about, and act within our environment reasonably well. This is because our brains take the information provided by our sensory systems and transform it into signals that can be understood, interpreted, and acted upon. Michael Posner and Richard Mitchell first conducted a classic study illustrating both of these key principles in the mid-1970s. Using a series of letters, they were able to clearly demonstrate that mental processes were internal and that these processes were necessarily manipulated and transformed when acting on them. The experiment is quite simple: it requires only that one present letter pairs to participants and ask them if the letters are the same or different. Why is it a classic experiment? Because of how the letters were paired and how the question of same or different was asked. In the simplest form, four different letters are used – two consonants and two vowels – and they are displayed in either upper or lower case. Using these stimuli, different levels of sameness can be asked. The lowest level is based on the physicality of the letters: are they physically the same? AA and aa are two examples of stimuli that would be judges as physically the same, Aa and aA are examples that would be correctly judged not the same. The next level of question involves identity, or the name of the stimuli. The previous four examples would all be judged the same based on name identity. AB and AE are examples of stimuli that would be judged as different. The next higher question involves judging if the letters are the same or different based on more detailed information concerning the letters – whether the letters are consonants or vowels. Reaction time data were collected during this experiment and clearly show that the mental processes cannot necessarily be overtly witnessed (they are internal) and various mental processes take various amounts of time, depending on the level of complexity involved in the process (or number of transformations necessary to perform the task). That is, the physical judgment was the easiest and fastest, the type judgment task was the most difficult and slowest, and the identity task (naming) was somewhere in between.

Another key technique that is often considered under the cognitive domain of cognitive neuroscience is neuropsychology. This is a science that seeks to integrate psychological observations of behavior and mind with neurological observations of the nervous system and the brain. This field is closely related to neurology, a branch of medicine concerned with understanding and treating disorders of the nervous system. One difference is that neuropsychology is specifically concerned with understanding and treating cognitive, perceptual, and other psychological deficits related to nervous system compromise. Such deficits can come in the form of vascular disorders such as strokes or bleeds, tumors, trauma, and disease.

These deficits of structure can also cause deficits of function, which is where cognitive neuroscience can benefit. The real strength of cognitive neuroscience is the ability to take a variety of seemingly disparate technologies and bits of data and integrate them into a coherent explanation of how brain enables mind. One of the most fruitful paradigms in cognitive neuroscience that depends on converging data involves the effects of brain lesions on behavior. Fundamental concepts, such as the notion that language resides primarily in the left hemisphere and the fact that the left hemisphere controls the right side of the body (and vice versa), are grounded in such a combination of techniques. The logic is pretty straightforward. If mental processes and behavior depend on processing within a certain brain region, then damage

to this region should impact the processes and behavior. Unfortunately, it is not that straightforward. Just because behaviors are disrupted following brain damage does not necessarily mean that the region damaged is responsible for the disrupted process, only that it is involved in the processes.

This is why cognitive neuroscientists often design research paradigms that involve at least two tasks, and two different types of brain compromise. Such a combination of techniques provides the most information because it is useful in dissociating structure–function issues. This notion of separating or dissociating various mental processes into relatively independently functioning units is quite attractive. A classic example of this approach can be illustrated by considering the findings of Broca and Wernicke. Broca's patient had trouble with speech output, but his language comprehension was intact. Wernicke's patient has problems with language comprehension, but his speech output was intact. These were two different language disorders in two different people. Behaviorally, these two language functions are dissociable. By adding the component of neuroanatomy – what part of the brain is compromised – the dissociation is strengthened. A modern example of a similar double dissociation can be illustrated with the phenomenon of blindsight. Blindsight is extremely interesting because it allows cognitive neuroscientists to pursue what is perhaps the highest level of cognition possible by the human brain – consciousness – in an established patient model. Blindsight comes about following damage to cortical regions of the primary visual pathway. Such damage results in corresponding regions of blindness in the visual field. Under certain testing conditions, some patients demonstrate residual vision within this area of blindness – residual vision without awareness. That is, they can respond correctly to some stimuli while denying actually having 'seen' the stimuli. Hence the term blindsight.

Interestingly, few patients with such damage demonstrate blindsight, and those who do so demonstrate the phenomenon differently. A study by Wessinger and colleagues showed this in terms of a dissociation: one participant (FN) can detect direction of motion but cannot detect differences in shape, whereas another participant (CLT) can detect differences in shape but not differences in motion. These data suggest a dissociation of motion and shape processing in blindsight. Interestingly, MRIs demonstrate differences in the location and extent of their damage. FN has greater damage in his inferior occipital lobe and CLT has greater damage in his parietal lobe – both have limited sparing around the calcarine fissure. These data suggest that it is not the damage to primary visual cortex that results in differences in blindsight; rather, it is compromise of higher order visual processing (i.e., extrastriate) modules. CLT's motion module is damaged, and thus does not have motion blindsight, but it does have shape blindsight because his form regions are less damaged. The opposite is true for FN: his form processing capabilities are too damaged to provide any useful form processing data, but his motion processing capabilities still provide some limited information – information that is sufficient to drive motion blindsight. Overall, these data indicate that blindsight may depend on surviving remnants of primary visual cortex and vestiges of connections to surviving, albeit severely compromised, extrastriate visual modules.

Blindsight is not the only interesting exploration of consciousness undertaken by the multidisciplinary approach of cognitive neuroscience. In an extremely interesting combination of philosophy and neuroscience, Benjamin Libet has made a career of investigating consciousness.

In a groundbreaking and often controversial series of experiments, Libet investigated the neural time factors in conscious and unconscious processing. These experiments are the basis for his 'backward referral hypothesis.' Libet and colleagues conclude that awareness of a neural event is delayed ~500 ms following the onset of the stimulating event, and more important, this awareness is referred back in time to the onset of the stimulating event. Thus, we are not aware of the event until after it occurs, although we often think we are aware of the event from the onset. Fortunately, backward referral of our consciousness is not so delayed that we act without thinking. The actual beginning of the act occurs sufficiently after the awareness of the intent to act, giving us time to override inappropriately triggered behavior. This ability to detect and correct errors is what Libet believes is the basis for free will. Cognitive neuroscience is investigating consciousness and free will and making headway.

Certainly, this is not an exhaustive review and explanation of the methods that converge to form the field of cognitive neuroscience. In fact, there are a whole host of animal techniques not covered. By definition, the field of cognitive neuroscience – a multidisciplinary science that borrows from many seemingly disparate fields in order to develop coherent explorations of the neurobiological underpinnings of cognition – is dynamic. The field is in constant flux, always expanding and growing and exploring new questions related to furthering our understanding of how brain enables mind. To some, such constant change may seem disconcerting, but to others it is what makes cognitive neuroscience exciting. Such constant change is in fact why cognitive

neuroscience came about in the first place. By combining the best of many disciplines, cognitive neuroscience is asking and answering interesting questions about how the brain enables mind, putting the field at the forefront of helping humans understand their place within the world.

Given that scientists from many seemingly disparate fields were already pursuing the common goal of understanding how brain enables mind, why was it necessary to formally define a multidisciplinary approach that has now become known as cognitive neuroscience? It is hoped that by raising exciting issues ranging from the development of the neuron doctrine to the use of cytoarchitectonics to relate brain morphology to autism, from the collection of behavioral data in cognitive paradigms to the integration of converging data in dissociation paradigms, and the combination of philosophy and neuroscience to understand consciousness, this article helps you understand why cognitive neuroscience emerged as a new multidisciplinary field of study. In sum, it was to accurately represent the merging of many different brain sciences, ranging from a neuroscientific and structural approach to a cognitive and functional approach. A neuroscientific approach is concerned with understanding the basic mechanisms of neural action that result in processing and communication of information within neuronal assemblies. It is also necessary to consider the anatomical substrates or structures that are composed of these neuronal assemblies. A cognitive approach is concerned with characterizing the cognitive processes that lead to specific functions or behavior. Cognitive neuropsychology is concerned with similar characterizations with the added dimension of trying to understand what structures are necessarily involved in various cognitive processes. Similarly, cognitive science is concerned with characterizing the computational algorithms necessary for such cognitive processes. By incorporating the concerns of these wide-ranging disciplines, cognitive neuroscience is more than just a sum of its parts; rather, it is a stronger, more comprehensive approach to understanding how the brain enables the mind – one that begins at the molecular level and continues to the problem of understanding human conscious experience.

See also: Cerebral Cortex; Cognition: An Overview of Neuroimaging Techniques; Cognitive Control and Development; Contextual Interactions in Visual Perception; Electroencephalography (EEG); Executive Function and Higher-Order Cognition: Neuroimaging; fMRI: BOLD Contrast; History of Neuroscience: Early Neuroscience; Neuroanatomy Methods in Humans and Animals; Neuroimaging; Phrenology; Positron Emission Tomography (PET).

Further Reading

Banich MT (2004) *Cognitive Neuroscience and Neuropsychology,* 2nd edn. Boston: Houghton Mifflin.

Farah MJ and Feinberg TE (eds.) (2000) *Patient Based Approaches to Cognitive Neuroscience.* Cambridge: MIT Press.

Finger S (1994) *Origins of Neuroscience: A History of Explorations into Brain Functions.* New York: Oxford University Press.

Gazzaniga MS (2000) (editor-in-chief) *The New Cognitive Neurosciences.* Cambridge: MIT Press.

Gazzaniga MS, Ivry RB, and Mangun GR (2002) *Cognitive Neuroscience: The Biology of the Mind,* 2nd edn. New York: Norton.

Hutsler JJ, Love T, and Zhang H (2007) Histological and magnetic resonance imaging of cortical layering and thickness in autism spectrum disorders. *Biological Psychiatry* 61: 449–457.

Kandel ER, Schwartz JH, and Jessell TM (1991) *Principles of Neural Science,* 3rd edn. New York: Elsevier.

Kosslyn SM and Koenig O (1992) *Wet Mind: The New Cognitive Neurosciences.* New York: Free Press.

Libet B (2005) *Mind Time: The Temporal Factor in Consciousness.* Cambridge, MA: Harvard University Press.

Posner MI and Mitchell RF (1967) Chronometric analysis of classification. *Psychological Review* 74(5): 392–409.

Rains GD (2002) *Principles of Human Neuropsychology.* Boston: McGraw-Hill.

Wessinger CM, Fendrich R, and Gazzaniga MS (1999) Variability of residual vision in hemianopic subjects. *Restorative Neurology and Neuroscience* 15: 243–253.

Wessinger CM, Fendrich R, and Gazzaniga MS (2005) Cognitive neuroscience: What is it and why? In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience,* 3rd edn. (on CD-ROM). Oxford: Elsevier.

Color Vision

P Gouras, Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Color vision is best known by humans' perception of it. It creates an aspect of vision that cannot be appreciated by any other means. But it is an illusion of reality that comes from the interaction of millions of neurons within the darkness of our skull projecting us into a multicolored universe, which would otherwise remain unknown. In addition to an esthetic impact, color vision facilitates detecting objects in the external world (Figure 1).

Color vision uses two properties of light to create visual contrasts. One property is effective energy, that is, the number of quanta in the light. The other is the frequency of vibration, or wavelength, of the absorbed quanta. How our brain separates these two properties of light and then recombines them to create color is the mystery of color vision.

All colors perceived by humans can be quantitatively defined by the mixtures of three international standards. Colors can also be described by three psychological concepts: their hue, best realized by each of the four most basic colors, red, green, yellow, and blue; their saturation, which is determined by the whiteness or blackness mixed with hue; and last, their luminance, or brightness. Describable by these variables, about 2 million different colors are perceivable by humans.

Relating color perception to physical and psychological dimensions, however, does not explain the physiology that creates these impressions in our brain. The purpose of this chapter is to try to explain the physiological basis of color. But explaining color vision by the responses and circuitry of neurons has pitfalls. Most of the information available at present comes from peripheral levels of the nervous system, the retina and thalamus, and even here, questions remain. The perception of color involves the entire brain, especially cerebral cortex. Color vision is also associated with the perception of form and the state of consciousness, about which our understanding is still vague. Nevertheless, progress is being made in exploring these higher levels of visual perception and color vision, especially with new techniques that can now monitor many neurons simultaneously in multiple parts of the brain in conscious humans or trained nonhuman primates.

In this description of color vision mechanisms, the human visual system is emphasized because we have a unique form of color and have much less of an idea of color perception in other animals. In addition, human color vision has been intensively studied over centuries, providing valuable insight into this process and intriguing us to find physiological explanations. But there are many diurnal animals, such as birds and fish, that have an even more elaborate form of color vision than humans and all other mammals.

The Photodetectors, Opsins

Unique proteins coupled to an isomer of vitamin A, 11-*cis* retinaldehyde, are the means by which we detect light. These proteins are located in the lipid membrane of a specialized organelle, the outer segment of photoreceptor cells, where they absorb light transmitted through the eye and focused on a monolayer of these receptors. Absorption of a quantum of light alters a double bond in 11-*cis* retinaldehyde, which activates the opsin molecule, allowing it to catalyze a complex biochemical reaction leading to a change in the electrical properties of the photoreceptor that can either stop or start the release of a synaptic transmitter molecule onto second-order retinal cells. This process, called 'phototransduction,' amplifies the energy transfer of a single quantum of absorbed light a million times and more.

Each opsin molecule is an antenna tuned to a particular band of the visible spectrum. By altering the amino acid composition of the opsin molecule, nature can tune these opsins to absorb light from the infrared to the ultraviolet region of the spectrum. The absorption spectra of the opsins an organism uses reflect the particular ecology of the organism, and this is under strong evolutionary pressure. The genes for the three human cone opsins have been cloned by Jeremy Nathans – a major contribution to understanding the initial stages of color vision.

The Evolution of Vision

Vision must have begun with diurnal organisms that could distinguish light from dark. The earliest photoreceptors, undoubtedly developing in broad daylight, were conelike, as first surmised by the comparative biologist Gordon Walls and confirmed by recent molecular genetics. Rods, which are extremely sensitive nocturnal photoreceptors, evolved later. Rods play no significant role in color vision; their response saturates at light levels at which cone function and

Figure 1 Difference between color vision using both wavelength and energy contrasts (top) and achromatic vision using only energy contrasts (bottom).

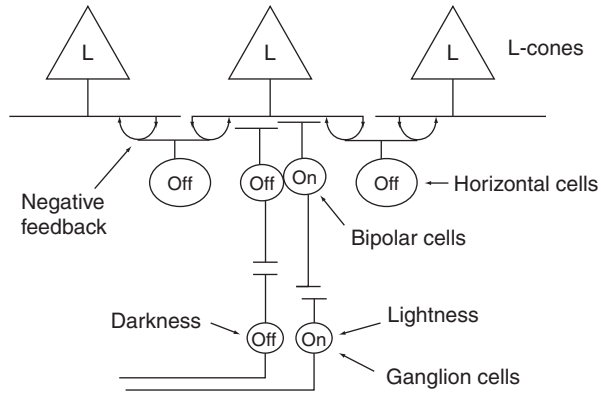

Figure 2 Key points in cone circuitry of human retina. Cones synapse on two types of bipolar cells, called on- and off-bipolars, which synapse on two types of ganglion cells, called on- and off-ganglion cells. In the fovea, these bipolars synapse with only one cone; elsewhere such bipolars synapse with several cones, which reduces their spatial resolution. Horizontal cells mediate antagonistic interactions between neighboring cones. Amacrine cells also mediate antagonistic interactions in the inner plexiform layer but have been omitted in this figure for simplicity.

color vision are optimal. In the evolution of vision, shadows must have been triggers for the detection of movement. A shadow depolarizes a cone, causing the release of a synaptic transmitter that depolarizes (excites) a second-order neuron, called an 'off-bipolar.' The reappearance of light with movement hyperpolarizes the cone, stopping transmitter release. This disinhibits (excites) a parallel second-order bipolar cell, an 'on-bipolar.' These two bipolar cell systems form a 'push–pull' circuit signaling dark and light, respectively. Cone bipolar cells excite third-order retinal ganglion cells, which transmit their signals to the central nervous system (Figure 2).

Cones respond optimally over enormous ranges of ambient illumination, owing to their ability to adapt to light. As ambient light increases, a cone's sensitivity decreases, but its response speed and temporal accuracy increase. This adaptation results in large part from the biochemistry of the cone photo-transduction process converting light into a neural signal.

Cones receive negative feedback from second-order neurons called horizontal cells. Horizontal cells are depolarized by the transmitter released by cones, being depolarized when cones are depolarized and hyperpolarized when cones are hyperpolarized. Horizontal cells send an antagonistic (inhibitory) signal back to cones, which counteracts the cone's response to light or dark. Whenever cones are hyperpolarized by light, horizontal cells are also hyperpolarized, stopping the release of their inhibitory transmitter, which after a slight synaptic delay depolarizes groups of neighboring cones, that is, opposes their response to light. This antagonism also occurs when cones are depolarized by darkness. Such negative feedback increases spatial and temporal resolution. With two cone types, as needed for color vision, this feedback can also narrow the spectral response of cones.

Color Vision

In a world seen by cones, all with the same opsin, objects appear brighter or darker but colorless, as in a black and white photograph (Figure 1). Whether an object is bright or dark depends on a comparison between the light energy it reflects and that reflected by its background. The lighter the background, the darker an object appears. The darker the background, the brighter an object appears. Absolute energy is discarded for relative energy, which is all that is needed for survival.

Inhabiting a world with only energy contrasts is a handicap because most objects differ in spectral reflectance, which is often a stronger clue to their presence than is their energy reflectance. Energy contrasts are also more misleading because of shadowing. Therefore, early in the evolution of vision, a mechanism to detect spectral as well as energy contrast must have evolved to create color vision. For this, at least two different cone types are needed, one absorbing best at one part and the other absorbing best at another part of the visible solar spectrum (Figure 3). This allows the brain to compare the responses of two cone systems to assess spectral contrast and create color vision.

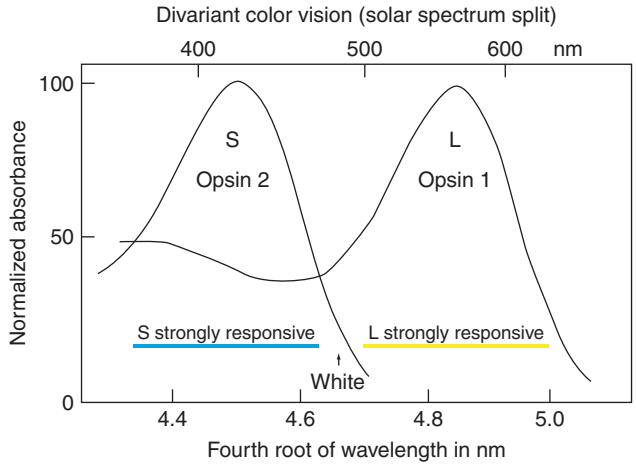

Figure 3 Normalized absorption spectra of the short (S-cone) and long (L-cone) wavelength-sensitive opsins that provide a divariant color vision system to many mammals and 2% of human males. When plotted against the fourth root of wavelength, opsin absorption curves have similar shapes. The horizontal lines under the absorption spectra indicate how the solar (white) spectrum is split into blue and yellow regions.

The primordial cones must have started with an opsin sensitive to longer wavelengths in the spectrum, an 'L-cone' opsin. Here, chromatic aberration is less than at short wavelengths, and sufficient quantal energy exists to activate opsins. Subsequently, a second cone system must have evolved with a short wavelength-sensitive ('S-cone') opsin (Figure 3). Short wavelengths are poorly transmitted by the ocular media and also exhibit considerable chromatic aberration. For this reason, the human S-cone system has been relegated to detecting only spectral and not energy contrast. Energy contrast provides a higher spatial resolution than does spectral contrast and therefore is more disturbed by chromatic aberration. Some small animals with tiny eyes and relatively large lenses, which provide greater depth of focus that reduces chromatic aberration and allows more ultraviolet light to reach the retina, have ultraviolet rather than blue-sensitive opsins expanding their range of vision. This has allowed many highly diurnal organisms, birds, fish, and reptiles, to expand the range and richness of their color vision.

Adding an S-cone system to the mammalian retina led to the evolution of unique S-cone-responsive ganglion cells (Figure 4).

On-bipolars from S-cones excite one dendritic arbor of this ganglion cell, and off-bipolars from L-cones excite a more superficial arbor of dendrites from the same ganglion cell. The inner plexiform layer of the retina is layered into off- and on-bipolar laminas, undoubtedly to prevent off-bipolars from exciting on-ganglion cells and on-bipolars from exciting off-ganglion cells, and this is reflected in the dendritic inputs to this ganglion cell.

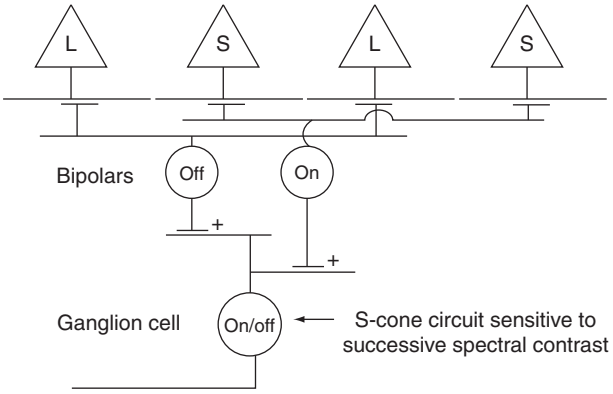

Figure 4 A circuit that transmits signals of S-cones from the retina to the brain. Each of such retinal ganglion cells receives excitatory inputs from S-cone-specific on-bipolars and L-cone-specific off-bipolars. L, long; S, short.

Such S-cone-activated ganglion cells are optimized to detect the appearance of short wavelengths as long wavelengths leave their receptive field. The L-cones excite the cell as long wavelengths leave the receptive field and as short wavelengths (blue light) enter the receptive field. This is an example of 'successive color contrast.' It illustrates a principle in the processing of neural signals for spectral contrast. The signals from the two different cone mechanisms are either antagonistic or out of phase with each other. The S-cones produce on-responses and the L-cones produce off-responses. This facilitates the response to maximum spectral contrast, that is, when short wavelengths enter as long wavelengths leave the receptive field of the cell.

The evolution of an S-cone system also added a new horizontal cell type to the retina. This so-called H2 horizontal cell allows L-cones to antagonize S-cones but may not allow S-cones to antagonize L-cones, although this latter point is controversial. H1 horizontal cells mediate antagonistic interactions among L-cones, being both pre- and postsynaptic to all L-cones they contact. H2 cells are presynaptic to both L- and S-cones but are likely only postsynaptic to S-cones and therefore do not antagonize L-cones. This arrangement appears likely because many ganglion cells in primate retina receive inputs from L-but not S-cones. If S-cones were presynaptic to L-cones, S-cone signals would be detected in all L-cone-responsive cells, which is not the case.

Color vision with these two cone systems splits the 'white' solar spectrum in half, with one cone mechanism more sensitive to one end and the other more sensitive to the opposite end of the visible spectrum. This arrangement allows the brain to determine whether an object reflects more or less light from these two different parts of the spectrum, a key to determining its color.

Energy versus Spectral Contrast

Energy contrast is determined by comparing the outputs of neighboring cones, but this cannot work for spectral contrast. To detect spectral contrast, the brain must compare the output of groups of cones containing both cone types in the same area of visual space. Spectral contrast requires a comparison different from that required for energy contrast. Small spots or narrow slits provide insufficient clues for determining spectral contrast because they do not cover a sufficient population of cones, but they are sufficient for detecting energy contrast. That is why the spatial resolution of chromatic vision is much less than that of achromatic vision (Figure 5).

Figure 5 shows two output cells, both of which can be retinal ganglion cells. The circuitry of the cell hypothetically coding for blue is well established, as illustrated in Figure 4. There is, however, another way to interpret the blue cell in Figure 5. It may be transmitting a signal for the presence of short-wave light, which could be used at a higher level to code for both blue and white. This hypothesis envisions color-specific cells to be formed in cerebral cortex. The existence of 'yellow on–blue off' cells is moot at present because such cells are rare in the retina.

This touches on an important point in the formulation of human color vision. The German physiologist Ewald Hering proposed that when blue and yellow colors are mixed, the color perceived does not appear as a mixture of yellow and blue but as a color that appears white or gray without any trace of blue or yellow. This concept led to the idea of 'opponent' colors. Certain color pairs such as blue and yellow or red and green (see below) are antagonistic to each other. Certainly at the level of retina one finds that cells excited by S-cones are being inhibited by L-cones, mediated by horizontal cells and probably amacrine cell antagonism. But there is a difference between such cone antagonism and color antagonism. The antagonistic interactions between different cone mechanisms narrow the relatively broad action spectrum of cone opsins, very much as oil drops do in the cones of certain diurnal birds, reptiles, and fishes. Antagonism

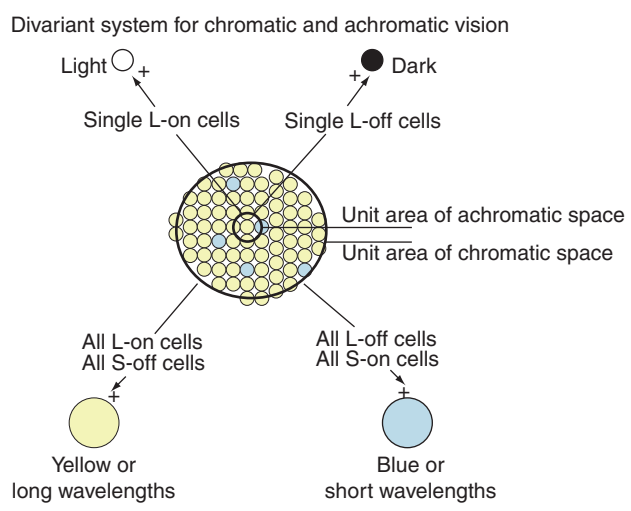

Figure 5 A mosaic of S- and L-cones needed for divariant color vision. S-cones are less common than L-cones because only the latter are involved in energy contrast while both cone types are involved in spectral contrast. To compare the responses of L-with those of S-cones, necessary for color vision, a relatively large group of cones must be sampled. For energy contrasts, single L-cones can be compared with their neighboring L-cones. L, long; S, short.

between colors, however, may occur only after the neural processing of cone signals is sufficient to create color-selective neurons, which seems to occur first in visual cortex.

But regardless of where in the brain the cells uniquely selective for color are located, the logic by which S- and L-cone signals can be used to code for color can be understood, as shown in Figure 6. Here both energy and spectral factors are considered to illustrate what arrangements must be used to code for colors in a divariant color vision system. The lightness or darkness of objects is determined by the achromatic L-cone system. The chromatic quality is determined by the activity of both cone types. If the S-cone system is absorbing light strongly, that is, its on-system is active, the colors perceived will be white, gray, or blue. Whether it is white or gray will depend on how strongly the L-cone-on system is active to detect an object that is being compared to its background. If the L-cone-on system is inactive, the L-cone-off system will be active, and the color perceived will be blue and not too bright. If the S-cone system is not absorbing light, the S-cone-off system is active, and the colors perceived will be yellow or brown. If the L-cone-off system is also active, the color will be black and dark.

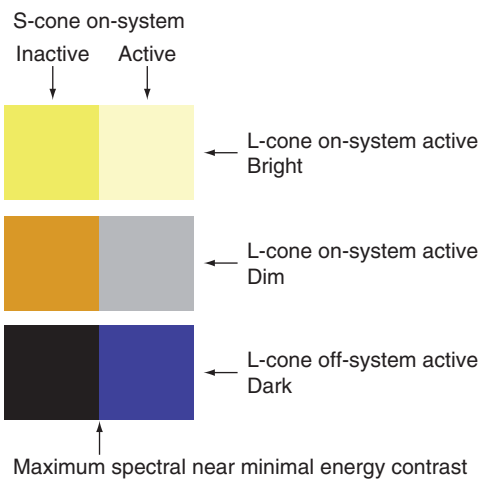

Figure 6 Divariant Color Vision.

Trivariant Color Vision

A major change occurred in primate vision when the gene for the long wavelength-sensitive opsin on the X chromosome reduplicated into two tandem genes with slightly altered sequences, probably in parallel with the evolution of a fovea in primates. This created two different longer wavelength opsins, each expressed in separate cones by gene programming. This split the long wavelength or yellow region of the spectrum into a red and green region (Figure 7).

These are called middle (M) and long (L) wavelength-sensitive opsins, the former absorbing more in the middle and the latter more in the long wavelength part of the spectrum. Because they straddle the yellow region of the divariant spectrum, they are closer together spectrally than the S- and L-cones of the primordial system. This may be why there is a separate H1 horizontal system interacting exclusively with L- and M-cones to physiologically narrow the action spectrum of these cones.

The Midget System's Double-Duty Role

This second system for obtaining chromatic contrast differs from the earlier divariant system because the retinal ganglion cells involved in this new system

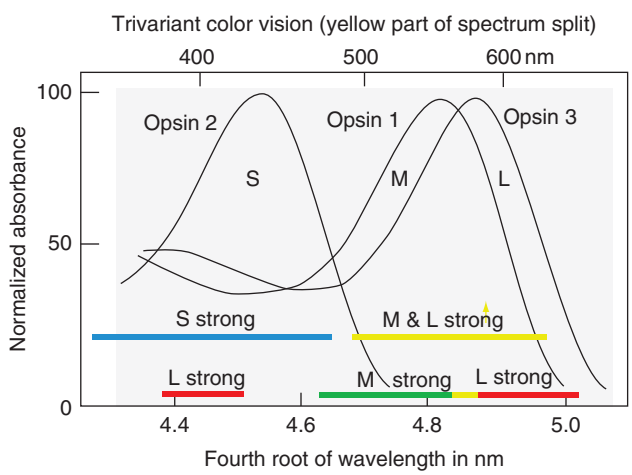

Figure 7 Splitting the long wavelength sensitive cone opsin into two slightly different opsins enables a new comparison of spectral reflectance operating over the longer wavelength part of the spectrum to occur. The L strong bar on the left is due to the greater absorption of the L-cone compared to M-come opsin in this short wave region of the spectrum, the so-called beta band.

transmit 'both' chromatic and achromatic contrast information. The new system is intimately involved with the midget cells of the fovea, which mediate high-resolution vision (Figure 2). Each L- and M-cone-mediating midget ganglion cell has a 'double-duty' role transmitting both high-resolution achromatic and low-resolution chromatic contrast. Models (Figure 8) exist that can explain how visual cortex separates these two forms of contrast by specific and parallel circuits.

For achromatic contrast, the small receptive field centers of single L- or M-cones detect local 'lightness' or 'darkness' by comparing the outputs of neighboring on- and off-center cells respectively, achieving the spatial resolution of single foveal cones. For chromatic contrast, on the other hand, a larger number of single cone signals are pooled in a highly selective way. L-cone-on and M-cone-off signals excite the same cell, and M-cone-on and L-cone-off signals excite another cell. These two cells sensing the same region of chromatic space could also inhibit one another to increase their selectivity for color. To minimize the center-surround antagonism present in each midget cell input, many adjacent cells are pooled, which increases the ratio of center to surround to make their receptive fields more coextensive and therefore more ideal for spectral contrast detection.

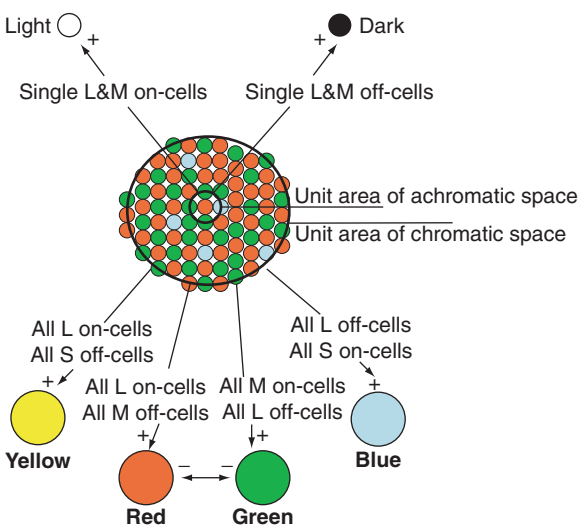

Figure 8 Trivariant system for chromatic and achromatic vision. Both chromatic and achromatic contrast detection can be derived from the same population of midget cells. For achromatic contrast, single long (L) and middle (M) on-cells excite a cortical cell and off-cells excite another cortical cell in linear arrays producing orientation-selective cells for light or dark with high spatial resolution. For chromatic contrast, L on-cells and M off-cells or L off-cells and M on-cells excite the same cortical cell in a coextensive, nonoriented manner to develop spectrally selective cells. A cell responsive to red may inhibit one responsive to green in the same area of chromatic space. Such cells reflect single color opponency.

The color gamut discriminable by a trivariant system is increased by the addition of a second comparison of L- and M-cones which occurs in parallel with the earlier divariant system. These two systems are generally described as red/green and blue/yellow opponent systems. Now the latter system compares the mixed signal of L- and M-cones with that of the S-cones (Figure 9).

The Retinex Theory

An interesting problem arises when the responses of two different cone systems sensing an area of visual space are compared in order to determine color. If the illuminant contains more energy from one end of the spectrum than from the other, as occurs with tungsten versus sunlight, colors will change in the direction of the illuminant. Human vision is only minimally affected by such a change, a phenomenon called 'color constancy.' Edwin Land examined this phenomenon over many years and suggested a mechanism that could be responsible for color constancy.

Land's 'retinex' (retina + cortex) model involves a mechanism that normalizes the responses of each cone mechanism before they are compared to determine color. This model can accurately predict the appearance of colors. Therefore, it is reasonable to expect that such a normalizing occurs at a point in the nervous system before the comparisons of cone mechanisms are made to determine color.

Double Opponent Cells

An insight into how normalizing could occur physiologically has come from the detection of so-called double opponent neurons in the visual cortex of

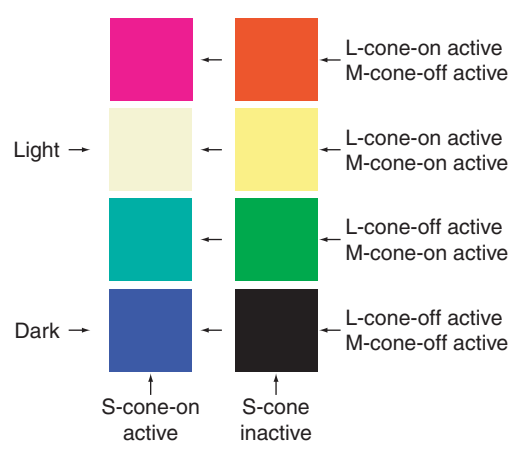

Figure 9 Trivariant color vision. Figure shows the variety of colors that can be discriminated by a trivariant color vision system combining chromatic and achromatic contrasts and their relationship to the three cone responses.

monkeys. Double opponent cells are organized to maximize spectral contrast by constructing concentrically organized cells which have one type of color-sensing cell exciting it in the center and the opposite color-sensing cell exciting and the same type of color-sensing cell inhibiting it in its surround (Figure 10). This arrangement illustrates the extraordinary selectivity of neural circuitry.

Double opponency will facilitate 'color constancy,' as Land's model requires, because it produces widespread antagonism of responses affecting each cone mechanism separately. This will tend to normalize the response of each cone mechanism before it is compared with another cone mechanism for color perception. Double opponency also provides an explanation of 'simultaneous color contrast,' in which colors are enhanced by surrounding them by an opposite, 'opponent,' color. A red color, for example, will appear 'redder' when surrounded by green.

It is important to realize that the creation of color-selective cells from the input of midget ganglion and geniculate cells must occur in visual cortex because such red and green color-selective cells, illustrated in Figures 8 and 10, have coextensive receptive fields, a phenomenon not seen in precortical midget cell responses. This supports the hypothesis that color-selective cells are formed ultimately in visual cortex.

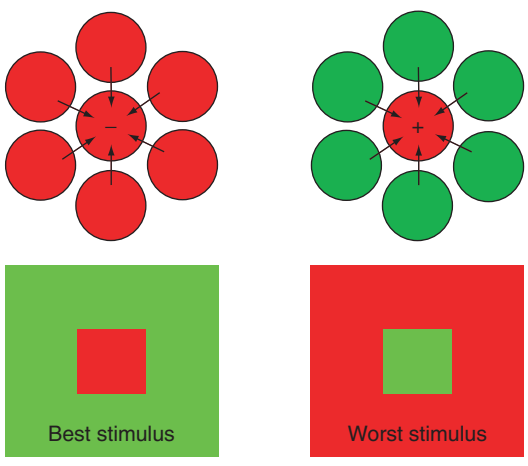

Figure 10 Double opponency; pure spectral contrast detection. A single opponent cell that is excited by L-cone-on and M-cone-off inputs can be transformed into a double opponent cell that is specific for red/green spatial contrast, pure spectral contrast. All the neighboring color-selective cells affect the central cell in the following ways. The neighboring 'red-selective' single opponent cells serving adjacent areas of space inhibit the central cell; all the neighboring 'green-selective' cells excite the central cell. Similar interactions among this population of cells will transform all such single opponent cells into double opponent cells. Similar double opponent cells can be formed for green, blue, or yellow selective cells. L, long; M, medium.

The Organization of Color Vision in Visual Cortex

Information on this topic is still rudimentary, but some important concepts are apparent (Figure 11). There is anatomical segregation of the inputs from the parvo and magno systems of the lateral geniculate nucleus. Only the parvo system is involved in high spatial resolution vision and color vision, the subject of this chapter. The magno system uses a different and parallel system of retinal ganglion cells that receive synergistic inputs from L- and M-cones and is thought to be involved in the detection of motion. The parvo system has often been referred to as the 'what is it' system and the magno to the 'where is it' system.

There is evidence that the signals of the parvo system used for chromatic contrast targets a different part of visual cortex than those for high spatial resolution achromatic vision. The former targets cytochrome oxidase rich 'blobs' whereas the latter targets larger 'interblob' regions of the cortex. Double opponent cells have been found in blobs. Blobs take less space than those devoted to achromatic processing, probably reflecting the higher spatial resolution of the latter. The parvo system targets the lower part of layer 4C and from there projects to layers 2 and 3 in striate cortex (V1). The parvo system projects from V1 to V2, V3, and V4, where separation from the magno system becomes more complete. The magno system targets upper layer 4C, where cells project to layer 4B

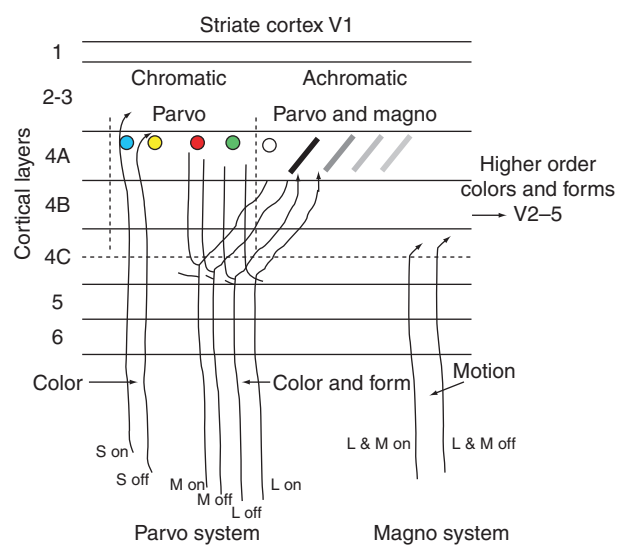

Figure 11 Synaptic targets of the different functional types of cells projecting to striate cortex from the lateral geniculate nucleus. There is a separation of parvo from magno cells, and within the former system, there is segregation of chromatic contrast-detecting cells into chromatic areas (blobs) and achromatic areas (interblobs). The S-cone system targets only the chromatic areas. S, short; V, visual cortex.

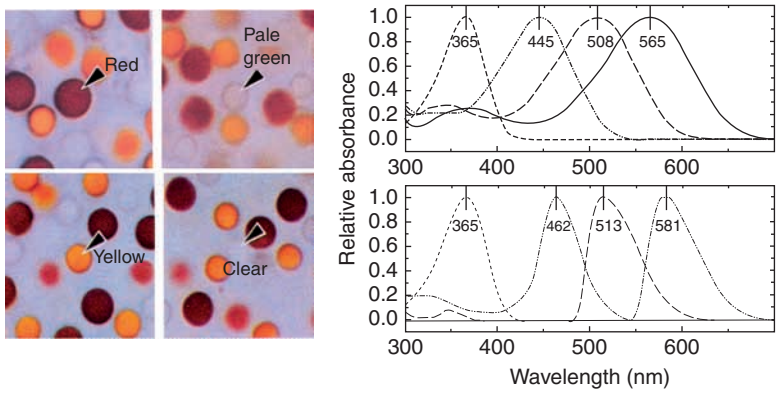

Figure 12 Turtle cones filtered by oil droplets. Panels on the left show the microscopic appearance of avian oil droplets: red, yellow, pale green, and transparent (Teruya Ohtsuka). On the right (above) are the absorption spectra of the four varieties of cones involved in the color vision of the bird Budgerigar (*Melopsittacus undulatus*); the graph below illustrates how oil droplets narrow these spectra (Timothy H. Goldsmith).

and from there to a higher area of visual cortex, V5, after passing through V2.

Color and Form Vision

A key point in understanding color vision is to determine how clues from chromatic contrast are used together with those from achromatic contrast to detect and distinguish objects. In this, both contours and surface textures must be important. The texture of objects is perhaps more determined by chromatic clues and associated with their colors. The achromatic clues provide higher spatial resolution, and the chromatic clues may provide greater reliability. Some mechanism must sort through these clues from contrast to assemble the most 'logical' representation of an object. Some of this logic must be influenced by the memory of known objects, and here the interactions of higher visual areas may be important.

Is There a Separate Color Area in Visual Cortex?

This question is closely associated with the discovery that visual cortex is subdivided into multiple areas where there is a separate re-representation of visual space. Evidence has been presented that area V4 contains cells processing only colors, but this hypothesis has not received strong support. The strongest support for such a hypothesis comes from extremely rare cases of people who lose color vision because of a disease in adult life but retain high spatial resolution vision. Such clinical material is often difficult to interpret. There are no examples of a genetic abnormality that dissociates these two processes from birth. Further studies will ultimately clarify this controversial topic. Color represents a combination of clues from both chromatic and achromatic contrast detectors. This computation must be tagged to objects which are detected by clues from these same detectors. It would seem reasonable that all visual areas processing form cannot easily separate, in time or space, chromatic from achromatic contrast detectors.

Tetravariant Color Vision

Primates with trivariant color vision and all other mammals with divariant color vision are deprived of the full universe of colors that is available to diurnal birds, fish, reptiles, and lizards that have tetravariant color vision, having added an ultraviolet-sensitive cone mechanism to increase their ability to detect chromatic contrast and colors. Figure 12 shows that birds do at both ends of the solar spectrum what primates do only at the yellow end. They have evolved an ultraviolet-sensitive cone which expands spectral contrasts at the short wavelength end of the spectrum. Such diurnal animals also use a unique method to narrow the action spectrum of each cone mechanism by having a carotenoid-containing oil droplet within the light path of each cone (Figure 12, left). Red, yellow, pale green, and transparent oil droplets are present in their L-, M-, S-, and ultraviolet-sensitive cones, respectively. This strategy narrows the action spectrum of these cones, as shown in Figure 12 (below, right), and this increases their chromatic contrast detection. The oil droplets also reduce the cones' sensitivity to light, but these organisms are active only in the presence of abundant sunlight, making sensitivity to light trivial.

Higher primates have not used either of these strategies to expand their color vision, probably because their eyes are so large that much less ultraviolet light

reaches their photoreceptors and oil droplets decrease the ability of mammals to operate at dim light levels.

See also: Cone Photopigment Evolution; Photoreceptor Adaptation; Photoreceptors: Physiology; Phototransduction; Representation of Color; Retinal Bipolar Cells; Retinal Color Mechanisms; Vision: Light and Dark Adaptation; Visual Development.

Further Reading

Conway BR, Hubel DH, and Livingston M (2002) Color contrast in macaque V1. *Cerebral Cortex* 12: 915–925.

Dacey DM and Lee BB (1994) The blue on-opponent pathway in the primate retina originates from a distinct bistratified ganglion cell. *Nature* 367: 731–735.

Dowling JE (1998) *Creating Mind: How the Brain Works*. New York: Norton.

Goldsmith TH (2006) What birds see. *Scientific American* July: 51.

Hofer H, Singer B, and Williams DR (2005) Different sensations from cones with the same photopigment. *Journal of Vision* 5: 444–454.

Kingdom FAA and Mullen KT (1995) Separating colour and luminance information in the visual system. *Spatial Vision* 9: 191–219.

Land E (1977) The retinex theory of color vision. *Scientific American* 237: 108–128.

Lennie P and Movshon JA (2005) Coding of color and form in the geniculate striate visual pathway (invited review). *Journal of the Optical Society of America. A, Optics and Image Science* 22: 2013–2053.

Michael C (1978) Color vision mechanisms in monkey striate cortex: Dual opponent cells with concentric receptive fields. *Journal of Neurophysiology* 41: 572–588.

Nathans J (1989) The genes for color vision. *Scientific American* 260: 44–49.

Ohtsaka T (1985) Relation of spectral types to oil droplets in cones of turtle retina. *Science* 229: 874–877.

Coma

S Laureys, M Boly, G Moonen, and P Maquet, University of Liège, Liège, Belgium

© 2009 Elsevier Ltd. All rights reserved.

Two Dimensions of Consciousness: Arousal and Awareness

Consciousness is a multifaceted concept that has two dimensions: arousal, or wakefulness (i.e., level of consciousness), and awareness (i.e., content of consciousness). One needs to be awake in order to be aware (rapid eye movement (REM) sleep and lucid dreaming being notorious exceptions). Figure 1 shows that in normal physiological states (green), level and content are positively correlated (with the exception of the oneiric activity during REM sleep). Patients in pathological or pharmacological coma (i.e., general anesthesia) are unconscious because they cannot be awakened (pink). The vegetative state (VS; blue) is a unique dissociated state of consciousness (i.e., patients are seemingly awake but lack any behavioral evidence of 'voluntary' or 'willed' behavior). There is, of course, an irreducible philosophical limitation in knowing for certain whether any other being possesses a conscious life. Awareness is a subjective first-person experience, and its clinical evaluation is limited to evaluating patients' motor responsiveness. In addition to its clinical and ethical importance, studying these disorders offers a still largely underestimated means of studying human consciousness. In contrast to unconscious states such as general anesthesia and deep sleep (in which impairment in arousal cannot be disentangled from impairment in awareness), VS is characterized by a unique dissociation of arousal and awareness, offering a lesional approach to identifying the neural correlates of awareness. A major challenge is to unequivocally disentangle patients residing in the clinical 'gray zone' between the vegetative and the minimally conscious state (MCS). The locked-in syndrome (LIS) is a rare but horrifying situation in which patients awaken from their coma fully aware but remain mute and immobile; it is called a 'pseudodisorder of consciousness' because patients superficially look unconscious but in reality are fully aware only unable to show it due to severe paralysis.

Arousal is supported by several brain stem neuronal populations (i.e., the reticular activating system) that directly or via nonspecific thalamic nuclei project to cortical neurons. Hence, focal damage of the brain stem or diffuse damage of the cerebral hemispheres may cause reduced arousal. The evaluation of eye-opening and brain stem reflexes is a key to the clinical assessment of the functional integrity of the arousal systems. Awareness depends on the functional integrity of the cerebral cortex and its reciprocal subcortical connections (each of its aspects partly residing in spatially defined brain areas), but its underlying neural code remains to be elucidated. Therefore there is, at present, no validated objective 'consciousness meter.' The estimation of the multiple dimensions of consciousness requires the interpretation of several clinical signs, mainly based on the observation of 'voluntary' interaction with the examiner or the environment. Brain death, coma, VS, MCS, and LIS (see Figure 2) are all solely defined by clinical criteria. Many scoring systems have been developed for a standardized assessment of consciousness in severely brain-damaged patients.

Nosology of Disorders of Consciousness

Brain Death

Brain death means human death determined by neurological criteria. The current definition of death is the permanent cessation of the critical functions of the organism as a whole (i.e., neuroendocrine and homeostatic regulation, circulation, respiration, and consciousness). Most countries, including the US, require death of the whole brain including the brain stem, but some (e.g., UK and India) rely on the death of the brain stem only, arguing that the brain stem is at once the through-station for nearly all hemispheric input and output, the center generating arousal (an essential condition for conscious awareness), and the center of respiration. Clinical assessments for brain death, however, are uniform and are based on the loss of all brain stem reflexes and the demonstration of continuing cessation of respiration (by performing a standardized apnea test) in a persistently comatose patient (Table 1). There should be an evident cause of coma, and confounding factors such as hypothermia, drugs, and electrolyte and endocrine disturbances should be ruled out.

Brain death is classically caused by a massive brain lesion (e.g., trauma, intracranial hemorrhage, anoxia) which increases intracranial pressure to values superior to mean arterial blood pressure and hence causes intracranial circulation to cease and damages the brain stem due to herniation. Using the brain stem formulation of death, however, unusual but existing cases of catastrophic brain stem lesion (often of hemorrhagic origin) sparing the thalami and cerebral cortex can be declared brain death in the absence of clinical brain stem function, despite intact intracranial circulation. Hence a patient with a primary brain

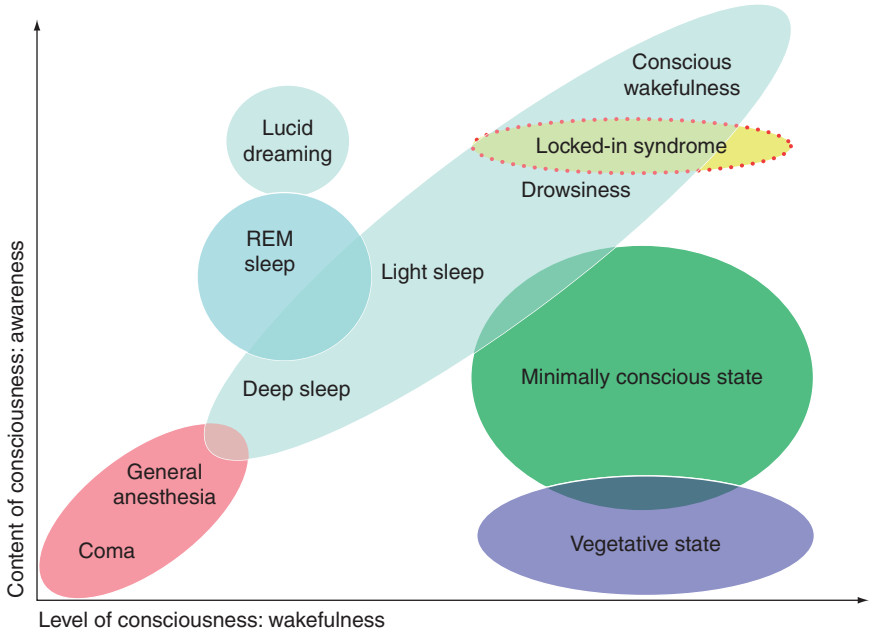

Figure 1 Oversimplified illustration of the two major dimensions of consciousness: the level of consciousness (i.e., arousal or wakefulness) and the content of consciousness (i.e., awareness or experience). Adapted from Laureys S (2005) The neural correlate of (un)awareness: Lessons from the vegetative state. *Trends in Cognitive Sciences* 9: 556–559.

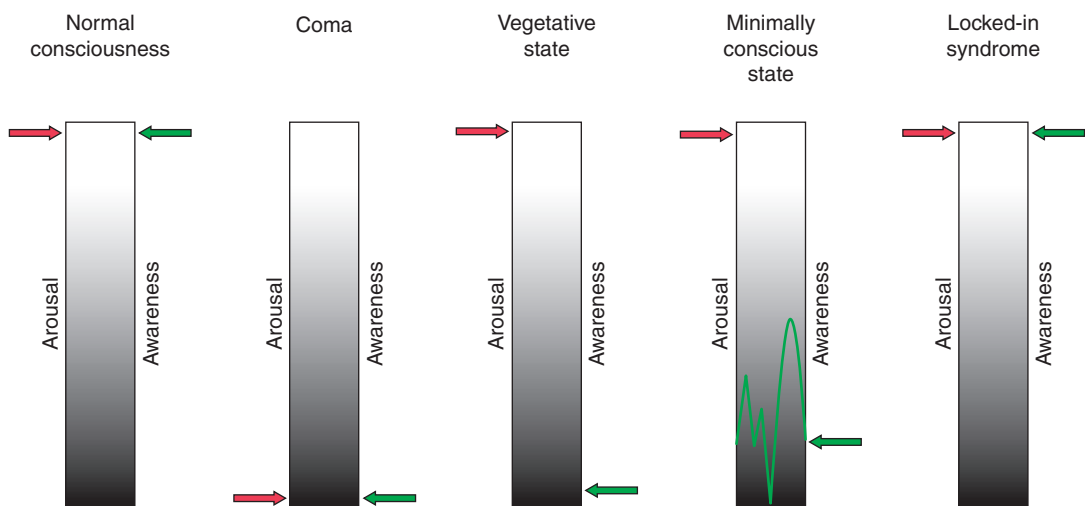

Figure 2 Graphical representation of the two dimensions of consciousness: arousal (red arrow) and awareness (green arrow) and their alterations in coma, the vegetative state, the minimally conscious state, and the locked-in syndrome. Adapted from Laureys S, Owen AM, and Schiff ND (2004) Brain function in coma, vegetative state, and related disorders. *Lancet Neurology* 3: 537–546.

stem lesion (who did not develop raised intracranial pressure) might theoretically be declared dead by the UK doctrine but not by the US doctrine.

Coma

Coma is a state of unarousable unresponsiveness characterized by the deficiency of the arousal systems (clinically assessed as the absence of stimulation-induced eye opening after having ruled out bilateral ptosis) and thus also by absence of awareness. The comatose patient lacks the sleep–wake cycles that can be observed in the VS. To be clearly distinguished from syncope, concussion, or other states of transient unconsciousness, coma must persist for at least 1 h. In general, comatose patients who survive begin to awaken and recover gradually within 2–4 weeks. This recovery may go no further than VS or MCS, or these may be stages (brief or prolonged) on the way to more complete recovery of consciousness. There are two main causes of coma: (1) bihemispheric

diffuse cortical or white matter damage secondary to neuronal or axonal injury and (2) brain stem lesions bilaterally affecting the subcortical reticular arousing systems (i.e., pontomesencephalic tegmentum and/or paramedian thalami).

VS

VS was defined in 1972 by Bryan Jennett and Fred Plum to describe those patients who 'awaken' from their coma (meaning they open their eyes spontaneously or on stimulation) but remain unaware of self or environment (meaning they show only reflex motor responses; Table 2). According to the *Oxford English Dictionary*, the term 'vegetative' describes "an organic body capable of growth and development but devoid of sensation and thought." It is very important to stress the difference between persistent and permanent VS, which are, unfortunately, too often abbreviated identically as PVS, causing unnecessary confusion. 'Persistent VS' has been arbitrarily defined as a VS still present 1 month after the acute brain damage, but the term does not imply irreversibility. In 1994, the US Multi-Society Task Force on PVS concluded that 3 months following nontraumatic brain damage and 12 months after traumatic injury, the condition of VS patients may be regarded as 'permanent.' Only in cases of 'permanent VS' do the ethical and legal issues surrounding withdrawal of treatment arise. It is essential that experienced examiners employing adapted standardized clinical assessment scales repeatedly establish the behavioral absence of any sign of conscious perception or deliberate action before making the diagnosis of VS.

MCS

In 2002, the Aspen Neurobehavioral Conference Workgroup published the diagnostic criteria for MCS to subcategorize patients above VS but unable to functionally communicate their thoughts and feelings. On a reproducible or sustained basis, MCS patients show limited but clearly discernible evidence of awareness of self or environment (Table 3). The emergence of MCS is characterized by the recovery of interactive communication or functional use of objects. Further improvement is more likely than in VS patients. However, some patients may remain permanently in MCS. At present, no time intervals for 'permanent MCS' have been agreed on.

'Akinetic mutism' (a condition characterized by severe poverty of movement, speech, and thought without associated arousal disorder or descending

Table 1 Criteria for brain death (American Academy of Neurology guidelines)

- Demonstration of coma
- Evidence for the cause of coma
- Absence of confounding factors, including hypothermia, drugs, electrolyte, and endorcrine disturbances
- Absence of brain stem reflexes
- Absent motor responses
- Apnea
- A repeat evaluation in 6 h is advised, but the amount of time is considered arbitrary
- Confirmatory laboratory tests are required only when specific components of the clinical testing cannot be reliably evaluated

Table 2 Criteria for the vegetative state (US Multi-Society Task Force on Persistent Vegetative State guidelines)

- No evidence of awareness of self or environment and an inability to interact with others
- No evidence of sustained, reproducible, purposeful, or voluntary behavioral responses to visual, auditory, tactile, or noxious stimuli
- No evidence of language comprehension or expression
- Intermittent wakefulness manifested by the presence of sleep–wake cycles
- Sufficiently preserved hypothalamic and brain stem autonomic functions to permit survival with medical and nursing care
- Bowel and bladder incontinence
- Variably preserved cranial nerve and spinal reflexes

Table 3 Criteria for the minimally conscious state (Aspen Neurobehavioral Conference Workgroup)

Clearly discernible evidence of awareness of self or environment, on a reproducible or sustained basis, by at least one of the following behaviors:

- Purposeful behavior (including movements or affective behavior that occurs in contingent relation to relevant environment stimuli and are not due to reflexive activity), such as:
  - Pursuit eye movement or sustained fixation occurring in direct response to moving or salient stimuli
  - Smiling or crying in response to verbal or visual emotional (but not neutral) stimuli
  - Reaching for objects demonstrating a relationship between object location and direction of reach
  - Touching or holding objects in a manner that accommodates the size and shape of the object
  - Vocalizations or gestures occurring in direct response to the linguistic content of questions
- Following simple commands
- Gestural or verbal yes/no response (regardless of accuracy)
- Intelligible verbalization

Emergence from the minimally conscious state requires reliable and consistent demonstration of at least one of the following behaviors:

- Functional interactive communication: accurate yes/no responses to six of six basic situational orientation questions (including items as, "Are you sitting down?" and "Am I pointing to the ceiling?") on two consecutive evaluations
- Functional use of two different objects (such as bringing a comb to the head or a pencil to a sheet of paper) on two consecutive evaluations

motor tract impairment) is an outdated term that should be avoided and is now considered to be a subcategory of the minimally conscious syndrome.

LIS

The category of LIS (pseudocoma) was introduced by Fred Plum and Jerome Posner in 1966 to refer to the constellation of quadriplegia and anarthria brought about by the disruption of, respectively, the brain stem's corticospinal and corticobulbar pathways. The syndrome describes patients who are aroused and awake but selectively deefferented (i.e., have no means of producing speech or limb or face movements). Usually the anatomy of the responsible lesion in the brain stem is such that locked-in patients are left with the capacity to use vertical eye movements and blinking to communicate their awareness of internal and external stimuli. Acute vascular bilateral ventral pontine lesions are its most common cause. Patients with such lesions often remain comatose for some days or weeks, needing artificial respiration, and then gradually wake up, remaining paralyzed and voiceless, superficially resembling someone in a coma or a VS. Table 4 lists the clinical criteria of LIS. The syndrome is subdivided, on the basis of the extent of motor impairment, into (1) classical LIS (characterized by total immobility except for vertical eye movements or blinking); (2) incomplete LIS (permitting remnants of voluntary motion); and (3) total LIS (consisting of complete immobility including all eye movements, combined with preserved consciousness).

Clinimetric Evaluation, Diagnosis, and Prognosis

Consciousness Scales

In 1974, Graham Teasdale and Bryan Jennett published the Glasgow Coma Scale (GCS). This standardized bedside tool to quantify arousal and awareness now is the gold standard in coma research. The GCS has tree components assessing eye opening and verbal and motor responses. Some authors have disagreed that evaluation of spontaneous or stimulation-induced opening of the eyes is sufficiently indicative of brain stem arousal system activity and have proposed coma scales that include brain stem reflexes. However, only the GCS has known a widespread use because the others generally are more complex. A simpler system, the Glasgow-Liège Scale (GLS), combines the GCS with five brain stem reflexes (i.e., fronto-orbicular, oculocephalic, pupillary, and oculocardiac reflexes). The increasing use of intubation and mechanical ventilation has rendered the verbal component of the GCS unassessable in many coma patients. Recently, the Full Outline of Un-Responsiveness (FOUR; also the number of components tested: eye, motor, brain stem reflexes, and respiratory function) scale proposed a hand-position test (i.e., asking the patient to make a fist or a 'thumbs-up', or 'V-for-victory' sign) as an alternative to the verbal component of the GCS.

The GCS, GLS, and FOUR scales lack reliability in assessing chronic disorders of consciousness. For these patients, more-sensitive scales are the Coma Recovery Scale-Revised (CRS-R), the Sensory Modality Assessment and Rehabilitation Technique (SMART), and the Wessex Head Injury Matrix (WHIM). The CRS-R is a recent scale specifically developed to differentiate VS from MCS and explicitly incorporating current diagnostic criteria. The basic structure of the CRS-R is similar to the GCS, but its subscales are much more detailed, targeting more-subtle signs of recovery of awareness.

Table 4 Criteria for the locked-in syndrome (American Congress of Rehabilitation Medicine)

- Sustained eye opening (bilateral ptosis should be ruled out as a complicating factor)
- Quadriplegia or quadriparesis
- Aphonia or hypophonia
- A primary mode of communication that uses vertical or lateral eye movement or blinking of the upper eyelid to signal yes/no responses
- Preserved awareness of the environment

Brain Death

Because many of the areas of the supratentorial brain (including the neocortex, thalami, and basal ganglia) cannot be tested for clinical functions accurately in a comatose patient, most of the bedside tests for brain death (such as cranial nerve reflexes and apnea testing) directly measure functions of only the brain stem. Since the first clinical definition of brain death nearly 50 years ago, no patient in apneic coma properly declared dead on the basis of neurological criteria has ever regained consciousness. Complementary neurophysiological tests such as electroencephalography (EEG), event-related potentials (ERPs), angiography, Doppler sonography, and scintigraphy reliably and objectively confirm the clinical diagnosis.

Coma

The management and prognosis of coma depend on many factors, such as etiology, the patient's general medical condition, age, clinical signs, and complementary examinations. After 3 days of observation,

absence of pupillary or corneal reflexes, stereotyped or absent motor response to noxious stimulation, isoelectrical or burst suppression pattern EEG, bilateral absent cortical responses on somatosensory evoked potentials, and (for anoxic coma) biochemical markers such as high levels of serum neuron-specific enolase are known to herald bad outcome. Visual and brain stem auditory evoked potentials are of limited prognostic use. Prognosis in traumatic coma survivors is better than in anoxic cases. Predicting outcome in toxic, metabolic, and infectious comatose states is challenging, and many unexpected cases of recovery have been reported.

VS

Whereas coma and brain death characteristically are acute conditions (lasting no more than days to weeks), the VS and MCS may become chronic entities. Vegetative patients, unlike brain dead or comatose patients, can move extensively, and studies have shown how difficult it is to differentiate 'automatic' from 'willed' movements in these patients. This results in an underestimation of behavioral signs of awareness and hence a misdiagnosis. It is known that when the diagnosis is made with insufficient care, up to one in three 'vegetative' patients actually are at least minimally conscious. Physicians also frequently tend to erroneously diagnose VS in older demented nursing home residents. Clinical testing for absence of awareness is much more problematic and slippery than testing for absence of arousal, brain stem reflexes, and apnea in irreversible coma. Given appropriate medical treatment, meaning artificial nutrition and hydration, VS patients may survive for many years.

Over the past decade, investigators have struggled to find an objective test that could reliably predict outcome for vegetative individuals. In contrast to coma and brain death, there are no validated diagnostic or prognostic markers for VS patients. The chances of recovery depend on the patient's age, etiology (worse for anoxic causes), and time spent in the VS. Recent data suggest that damage to the corpus callosum and brain stem indicate bad outcome in traumatic VS.

MCS

Because criteria for MCS have only recently been introduced, there are few clinical studies of patients in this condition. As stated above, it remains very challenging to behaviorally differentiate minimally conscious from vegetative patients because both are, by definition, noncommunicative. VS is one end of a spectrum of awareness, and the subtle differential diagnosis with MCS necessitates repeated evaluations by experienced examiners. As with VS, traumatic etiology has a better prognosis than does nontraumatic (anoxic) MCS. Preliminary data show that overall outcome for MCS is better than for VS.

LIS

In acute LIS, eye-coded communication may be difficult because of fluctuating arousal and limited control of voluntary eye movements. More than half the time, it is family members, not the physician, who first realize that the patient is aware. Distressingly, the diagnosis takes on average more than 2.5 months. In some cases it has taken 4–6 years before aware and sensitive patients, locked in an immobile body, have been recognized as being conscious. Some memoirs written by locked-in patients well illustrate the clinical challenge of recognizing the syndrome. Striking examples are *Look Up for Yes*, by Julia Tavalaro, and *Only the Eyes Say Yes*, by Phillippe Vigand.

While motor recovery remains very limited in LIS, life expectancy (with adequate medical care) may be several decades. Sensory evoked potentials are not reliable predictors of prognosis, but motor evoked potentials might evaluate the potential motor recovery. Eye-controlled computer-based communication technology now allows patients to control their environment, use a word processor coupled to a speech synthesizer, and access the Internet. Outsiders often assume that the quality of life with LIS is so poor that it is not worth living. Recent surveys, however, have revealed that chronic LIS patients self-report meaningful quality of life and that the demand for euthanasia is infrequent.

Residual Cerebral Function

Brain Death

The EEG in brain death shows absent electrocortical activity (i.e., isoelectric recording) with a sensitivity and specificity of around 90%. This makes the EEG the most validated and, because of its wide availability, preferred confirmatory test for brain death. Somatosensory evoked potentials typically indicate arrest of conduction at the cervicomedullary level, and brain stem auditory evoked potentials usually show only a delayed wave I (originating in the cochlear nerve). Cerebral angiography and transcranial Doppler sonography document with very high sensitivity and 100% specificity the absence of cerebral blood flow in brain death. Similarly, radionuclide cerebral imaging such as single-photon emission computed tomography (CT) and positron-emission tomography (PET) demonstrate the hollow-skull sign confirming

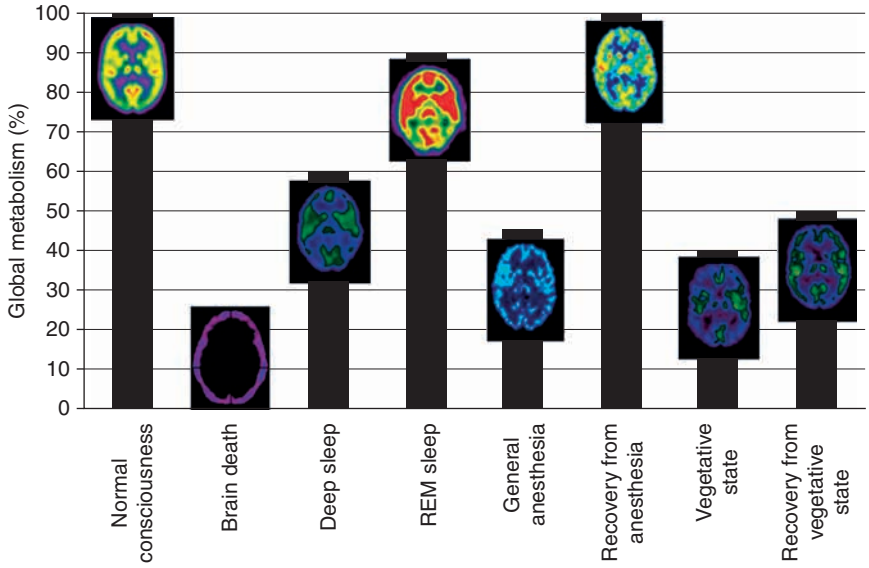

Figure 3 Cerebral metabolism in conscious wakefulness, in brain death (hollow skull sign confirming the absence of neuronal function in the whole brain in irreversible coma), physiological (slow wave sleep) and pharmacological (general anesthesia) modulation of arousal reflecting massive global decreases in cortical metabolism (in REM sleep, metabolic activity is paradoxically prominent); and in wakefulness without awareness (i.e., the vegetative state). Note that recovery from the vegetative state may occur without substantial increase in overall cortical metabolism, emphasizing that some areas in the brain are more important than others for the emergence of awareness. Images of halothane-induced loss of consciousness from Alkire M, Pomfrett CJ, Haier RJ, et al. (1999) Functional brain imaging during anesthesia in humans: Effects of halothane on global and regional cerebral glucose metabolism. *Anesthesiology* 90: 701–709.

the absence of neuronal function in the whole brain (Figure 3).

Anatomopathology in brain death patients receiving maximal artificial means of support will inevitably end up showing the so-called respirator brain, and after about a week, an autolyzed liquefied brain will pour from the opened skull.

Coma

The electrical activity of the brain as measured by the EEG tends to become nonreactive and slower as the depth of coma increases, regardless of the underlying cause. As stated above, the bilateral absence of cortical potentials (called N20 waves because they occur after about 20 ms) on somatosensory evoked potentials herald bad outcome. In cases in which cortical potentials are present, so-called endogenous ERPs might be useful. The presence of 'mismatch negativity,' that is, a negative component elicited after 100–200 ms by any change or 'mismatch' in a sequence of monotonous auditory stimuli (i.e., an 'oddball paradigm') indexes some persistent automatic information processing and is correlated with recovery of at least minimal consciousness.

Cortical metabolism in coma survivors is on average 50–70% of normal reference values. Cerebral metabolism correlates poorly with the level of consciousness, as measured by the GCS, in severely head-injured patients. A global depression of cerebral metabolism is not unique to coma. When anesthetic drugs are titrated to the point of unresponsiveness, the resulting reduction in brain metabolism is similar to that observed in pathological coma. Another example of transient metabolic depression can be observed during slow wave sleep. In this daily physiological condition, cortical cerebral metabolism can drop to nearly 40% of normal values, although in REM sleep, metabolism returns to normal waking values (Figure 3).

VS

In VS the EEG most often shows a diffuse slowing (i.e., generalized polymorphic delta or theta rhythm); only sporadically is it of very low voltage or isoelectric. Somatosensory evoked potentials may show preserved primary somatosensory cortical potentials, and brain stem auditory evoked potentials often show preserved brain stem potentials in vegetative patients. Endogenous evoked potentials measuring, for example, the brain's response to complex auditory stimuli such as the patient's own name (compared with other names) record a so-called P300 response (i.e., a positive wave elicited around 300 ms poststimulus when patients detect an unpredictable target in the regular train of stimuli). In brain-damaged patients, the use of emotionally meaningful stimuli such as the patient's own name increases the chances of obtaining

a P300 response. Recent data show that the P300 is not a reliable marker of awareness but rather indicates automatic processing, because it could be recorded in well-documented VS patients who never recovered (Figure 4).

In contrast to the functional decapitation observed in irreversible coma or brain death, vegetative patients show substantially reduced (40–50% of normal values) but not absent overall cortical metabolism. In some vegetative patients who subsequently recovered, global metabolic rates for glucose metabolism did not show substantial changes (Figure 3). Hence, the relationship between global levels of brain function and the presence or absence of awareness is not absolute. It rather seems that some areas in the brain are more important than others for its emergence.

Anatomopathologic features in anoxic VS encompass multifocal laminar cortical necrosis, diffuse leucoencephalopathy, and bilateral thalamic necrosis; VS following blunt head injury classically shows diffuse white matter damage with neuronal loss in thalami and hippocampi. However, these postmortem studies have not permitted a detailed regional topography of cerebral damage characteristic of VS. Voxel-based statistical analyses of metabolic PET data have identified a metabolic dysfunction in a wide frontoparietal network encompassing the polymodal associative cortices: bilateral lateral frontal regions; parietotemporal and posterior parietal areas; and mesiofrontal, posterior cingulate, and precuneal cortices (Figure 5), known to be the most active 'by default' in resting nonstimulated conditions.

In some other conditions, patients also show merely reflex or automatic motor activity while behaviorally seeming 'awake.' In absence seizures (characterized by brief episodes of unresponsiveness and staring, frequently accompanied by purposeless eye blinking and smacking), functional magnetic resonance imaging (MRI) studies have shown decreases in blood oxygen level-dependent signals in a wide

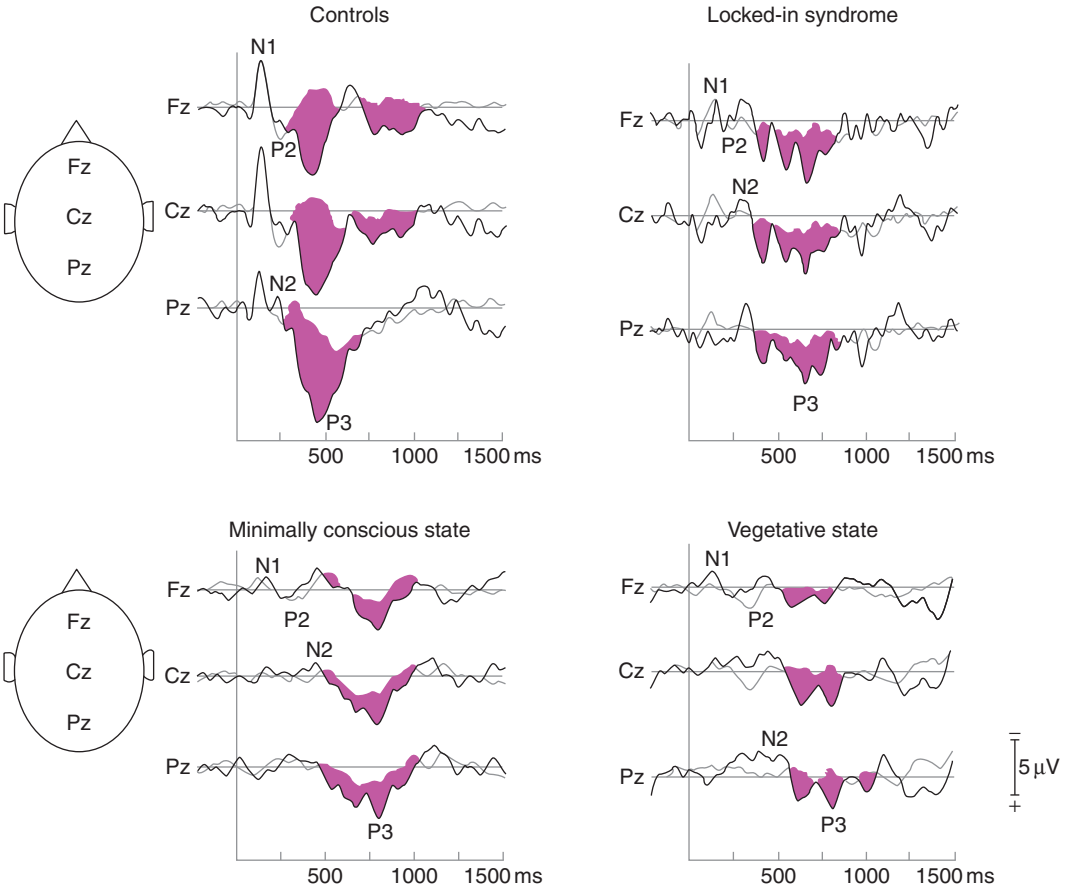

Figure 4 Endogenous event-related potentials to the patient's own name (thick traces) and to other first names (thin traces) in five healthy controls, four patients with locked-in syndrome, six minimally conscious state patients, and five vegetative state patients. Note that the differential P300 response (in pink) can also be observed in some well-documented VS patients (who never recovered) and hence is not a reliable indicator of awareness. Adapted from Perrin F, Schnakers C, Schabus M, et al. (2006) Brain response to one's own name in vegetative state, minimally conscious state, and locked-in syndrome. *Archives of Neurology* 63: 562–569.

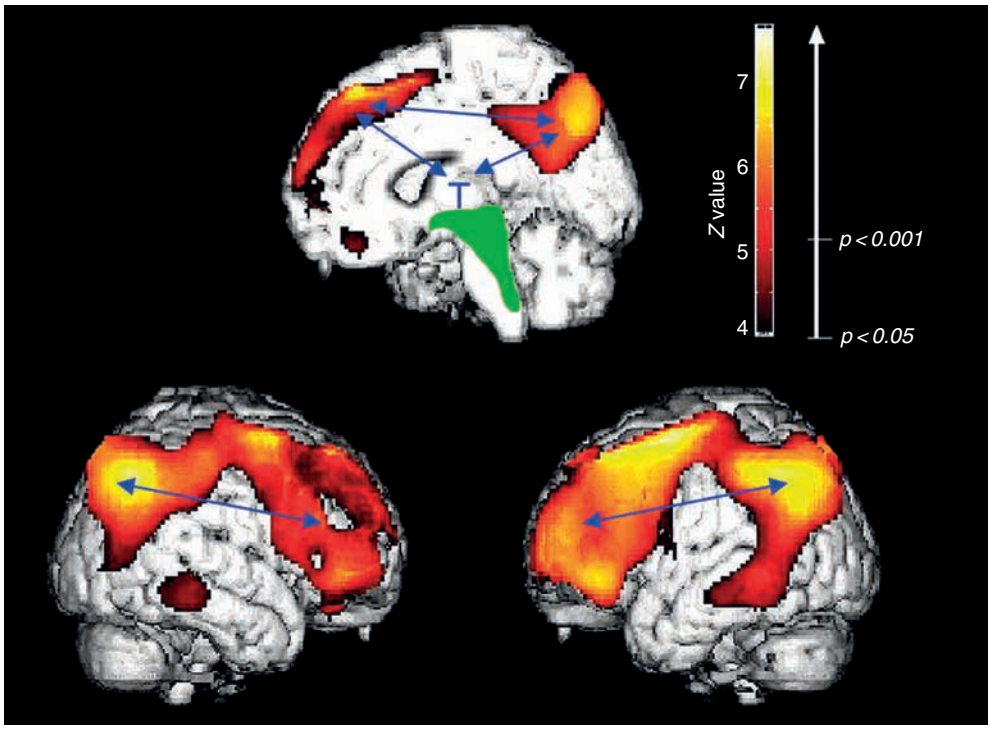

Figure 5 The common hallmark of the vegetative state is a metabolic dysfunctioning of a widespread cortical network encompassing medial and lateral prefrontal and parietal multimodal associative areas. This might be due to either direct cortical damage or corticocortical or thalamocortical disconnections (schematized by blue arrows). Also characteristic of the vegetative state is the relative sparing of metabolism in the brain stem (encompassing the pedunculopontine reticular formation, hypothalamus, and basal forebrain, represented in green), allowing for the patients' preserved arousal and autonomic functions. Adapted from Laureys S (2005) The neural correlate of (un)awareness: Lessons from the vegetative state. *Trends in Cognitive Sciences* 9: 556–559.

frontoparietal network, similar to that identified in VS patients. In some temporal lobe seizures that impair consciousness (coined 'complex partial') while preserving automatic activities such as picking, fumbling, or even cycling, single-photon emission CT has shown similar widespread deactivation in frontal and parietal association cortices. In contrast, temporal lobe epilepsy with sparing of consciousness (termed 'simple partial') is not accompanied by such frontoparietal dysfunction. Another example of transient nonresponsiveness with preserved automatic behavior such as ambulation can be observed in sleepwalkers. Again, deactivations in large areas of frontal and parietal association cortices were demonstrated by means of SPECT imaging during sleepwalking. Taken together, these observations point to the critical role of the frontoparietal associative cortices in the generation of conscious awareness.

However, awareness seems not exclusively related to the activity in this 'global workspace' cortical network but also, it is important to note, to the functional connectivity within this system and with the thalami. Long-range frontoparietal and thalamocortical (with nonspecific intralaminar thalamic nuclei) 'functional disconnections' have been identified in VS. Moreover, recovery is paralleled by a functional restoration of this frontoparietal network and part of its thalamocortical connections.

The most relevant question regards possible residual perception and cognition in vegetative patients. In well-documented VS patients, noxious high-intensity electrical stimulation (experienced as painful in controls) resulted in activation in brain stem, thalamus, and primary somatosensory cortex, but hierarchically higher-order areas of the pain matrix encompassing the anterior cingulate cortex failed to activate (Figure 6). It is important to note that the activated cortex was isolated and functionally disconnected from the frontoparietal network, considered critical for conscious perception. Similarly, auditory stimulation in VS was found to activate primary auditory cortices but not higher-order multimodal areas from which they were disconnected. The activation in primary cortices in awake but unaware patients confirms Crick and Koch's early hypothesis (based on visual perception and monkey histological connectivity) that neural activity in primary cortices is necessary but not sufficient for awareness.

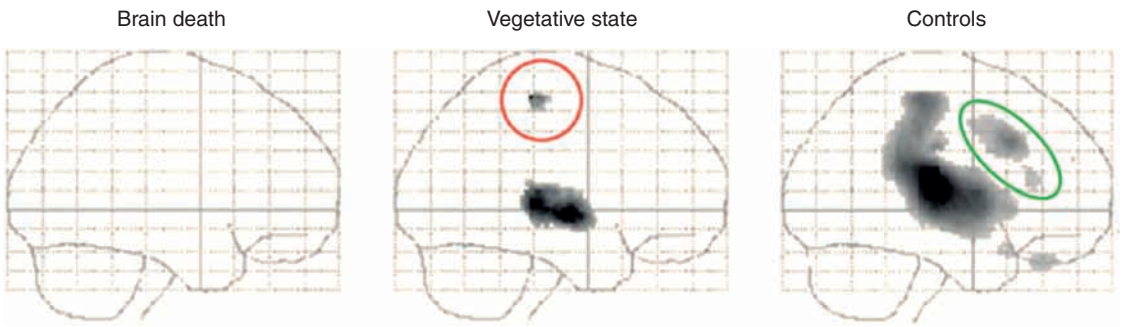

Figure 6 High-intensity somatosensory stimuli fail to induce any subcortical or cortical neural activation in irreversible coma with clinical absence of brain stem reflexes (i.e., brain death). In the vegetative state, subcortical (upper brain stem and thalami) and cortical (primary somatosensory cortex; red circle) activation can be observed. However, this preserved cortical activation is limited to the primary cortex and fails to reach higher-order associative cortices from which it is functionally disconnected. In healthy volunteers, reporting the stimuli as painful, stimulation resulted in a wide neural network activation (the so-called pain matrix) including the anterior cingulate cortex (green ellipse). Data from Laureys S, Faymonville ME, Peigneux P, et al. (2002) Cortical processing of noxious somatosensory stimuli in the persistent vegetative state. *Neuroimage* 17: 732–741 and shown on 'glass brains.'

MCS

PET studies measuring cerebral metabolism at rest are unable to reliably differentiate vegetative from minimally conscious individuals. Functional imaging can here be of utmost value in objectively differentiating activation patterns measured during external stimulation characteristic of either clinical entity. Complex auditory stimuli with emotional valence, such as personalized narratives or the patient's own name, activate language-related cortical areas not observed during presentation of meaningless stimuli in MCS. Such context-dependent higher-order auditory processing in MCS patients (often is not assessable at the patient's bedside) indicate that content does matter when talking to MCS patients.

However, given the absence of a thorough understanding of the neural correlates of consciousness, functional neuroimaging results must be used with caution as proof or disproof of awareness in severely brain-damaged patients. Recently, Adrian Owen, from Cambridge, proposed a more powerful approach for identifying 'volition without action' in noncommunicative brain-damaged patients (rather than the use of passive external stimulation paradigms): scanning patients when they are asked to perform a mental imagery task. Reproducible and anatomically specific activation in individual patients during tasks that unequivocally require 'willed action' or intentionality for their completion unambiguously reflect awareness. Of course, negative findings in the same circumstances cannot be used as evidence for lack of awareness. In one exceptional VS patient, task-specific activation was observed, unequivocally demonstrating conscious awareness in the absence of reliable behavioral motor signs of voluntary interaction with the environment. It is interesting that the patient subsequently recovered. Other studies also showed that VS patients with atypical brain activation patterns on functional neuroimaging afterwards showed clinical signs of recovery of consciousness, albeit sometimes many months later.

Investigations using MRI diffusion tensor imaging, which gauges the integrity of the white matter, are increasing scientific understanding of the brain mechanisms underlying recovery from VS. A team led by Nicholas Schiff of Cornell University, for instance, recently used diffusion tensor imaging to show the regrowth of axons in the brain of Terry Wallis, an Arkansas man in posttraumatic MCS who started talking in 2003 after 19 years of silence.

LIS

Classically, the EEG is relatively normal (or minimally slow) and reactive to external stimuli in LIS, but unreactive alpha rhythms (i.e., 'alpha coma' patterns) may also be observed. Cognitive ERPs and brain–computer interfaces may document awareness and permit communication in the extremely challenging cases of complete LIS. Residual cognition in LIS has long remained *terra incognita*. Recently, standard neuropsychological testing batteries have been adapted and validated for eye response communication. In classical LIS caused by a brain stem lesion, these studies have shown preserved attention, memory, executive functioning, and phonological and lexicosemantic performance.

PET scanning has shown significantly higher metabolic levels in the brains of patients in LIS compared with patients in VS. Voxel-based comparisons with healthy controls showed that no supratentorial cortical area has a significantly impaired metabolism in classical LIS. Conversely, hyperactivity was observed in bilateral amygdala of acute, but not chronic, LIS patients. The amygdala is known to be involved in emotions, especially negative emotions such as fear and anxiety. The absence of metabolic signs of

reduced function in any area of the gray matter reemphasizes the fact that LIS patients suffer from a pure motor de-efferentation and recover an entirely intact intellectual capacity. The increased activity in the amygdala might relate to the terrifying situation of an intact awareness in a mute but sensitive being. Healthcare workers should be aware of this condition, adapt their bedside behavior, and consider pharmacological anxiolytic therapy.

Conclusion

Brain death, coma, VS, and MCS represent different pathological alterations of both dimensions of consciousness (involving arousal and awareness) or, for LIS, of the motor signs of consciousness. At the bedside, the evaluation of conscious perception and cognition in these conditions is difficult and sometimes erroneous. Electrophysiological and functional neuroimaging studies can objectively describe the regional distribution of cerebral activity at rest, under various conditions of passive stimulation, and during 'active' mental imagery tasks. These studies are increasing our understanding of the neural correlates of arousal and awareness and will improve the diagnosis, prognosis, and management of disorders of consciousness. But at present, much more data and methodological validation are awaited before functional neuroimaging studies can be proposed to the medical community as a tool for disentangling the clinical gray zone that separates conscious from unconscious survivors of acute brain damage.

See also: Awareness: Functional Imaging; Brain Trauma; Coma and Other Pathological Disorders of Consciousness; Consciousness: Theories and Models; Consciousness: Philosophy; Consciousness: Neural Basis of Conscious Experience; Reticular Activating System; Stroke: Injury Mechanisms; Vegetative State.

Further Reading

Alkire M, Pomfrett CJ, Haier RJ, et al. (1999) Functional brain imaging during anesthesia in humans: Effects of halothane on global and regional cerebral glucose metabolism. *Anesthesiology* 90: 701–709.

American Congress of Rehabilitation Medicine(1995) Recommendations for use of uniform nomenclature pertinent to patients with severe alterations of consciousness. *Archives of Physical Medicine and Rehabilitation* 76: 205–209.

Bernat JL (2006) Chronic disorders of consciousness. *Lancet* 367: 1181–1192.

Giacino JT, Ashwal S, Childs N, et al. (2002) The minimally conscious state: Definition and diagnostic criteria. *Neurology* 58: 349–353.

Giacino JT, Hirsch J, Schiff N, and Laureys S (2006) Functional neuroimaging applications for assessment and rehabilitation planning in patients with disorders of consciousness. *Archives of Physical Medicine and Rehabilitation* 87: 67–76.

Laureys S, Faymonville ME, Peigneux P, et al. (2002) Cortical processing of noxious somatosensory stimuli in the persistent vegetative state. *Neuroimage* 17: 732–741.

Laureys S (2005) The neural correlate of (un)awareness: Lessons from the vegetative state. *Trends in Cognitive Sciences* 9: 556–559.

Laureys S (2005) Science and society: Death, unconsciousness and the brain. *Nature Reviews Neuroscience* 6: 899–909.

Laureys S, Owen AM, and Schiff ND (2004) Brain function in coma, vegetative state, and related disorders. *Lancet Neurology* 3: 537–546.

Laureys S, Pellas F, VanEeckhout P, et al. (2005) The locked-in syndrome: What is it like to be conscious but paralyzed and voiceless? *Progress in Brain Research* 150: 495–511.

Majerus S, Gill-Thwaites H, Andrews K, and Laureys S (2005) Behavioral evaluation of consciousness in severe brain damage. *Progress in Brain Research* 150: 397–413.

Multi-Society Task Force on PVS (1994) Medical aspects of the persistent vegetative state (1). *New England Journal of Medicine* 330: 1499–1508.

Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, and Pickard JD (2006) Detecting awareness in the vegetative state. *Science* 313: 1402.

Perrin F, Schnakers C, Schabus M, et al. (2006) Brain response to one's own name in vegetative state, minimally conscious state, and locked-in syndrome. *Archives of Neurology* 63: 562–569.

Quality Standards Subcommittee of the American Academy of Neurology (1995) Practice parameters for determining brain death in adults (summary statement). *Neurology* 45: 1012–1014.

Schiff ND, Rodriguez-Moreno D, Kamal A, et al. (2005) fMRI reveals large-scale network activation in minimally conscious patients. *Neurology* 64: 514–523.

Tavalaro J and Tayson R (1997) *Look Up for Yes*. New York: Kodansha America.

Vigand P and Vigand S (2000) *Only the Eyes Say Yes*. New York: Arcade.

Voss HU, Uluc AM, Dyke JP, et al. (2006) Possible axonal regrowth in late recovery from the minimally conscious state. *Journal of Clinical Investigation* 116: 2005–2011.

Coma and Other Pathological Disorders of Consciousness

N D Schiff, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Published by Elsevier Ltd.

Introduction

Coma is an unconscious brain state that may be produced by a variety of acute or chronic brain insults. Coma is uniquely characterized by the total absence of patterned behavioral arousal; comatose patients are completely unresponsive and lay motionless in an eyes-closed state. There are no spontaneous periods of wakefulness, nor induced eye opening with vigorous stimulation that may be observed in other unresponsive brain states. By definition, and to distinguish coma from syncope, concussion, or other forms of transient unconsciousness, the term coma implies that the state has endured for at least 1 h. Coma is a transient condition and, if uncomplicated by concurrent systemic illness, comatose patients who survive begin to awaken within 10–14 days. Some patients recover rapidly, whereas others recover more gradually, first entering into a vegetative state or minimally conscious state (as defined later) or other global disorders of consciousness that may be confused with coma. Time frames and probabilities of recovery from coma, vegetative state, or minimally conscious state depend greatly on the nature of the underlying brain injury producing the initial coma. In addition to pathologic causes of coma, general anesthesia reflects a pharmacologically induced coma and studies suggest some overlap in mechanisms at the neuronal circuit level between regional effects of anesthetics and focal brain injuries that may produce acute coma. This article briefly reviews the common mechanisms producing coma and related disorders of consciousness. Emphasis is placed on the neuronal circuits supporting forebrain arousal and integrative cerebral function disrupted in these conditions. Finally, some suggestions regarding the possible overlap of selective brain injuries producing coma and the mechanisms of action of general anesthetics are proposed.

Underlying Mechanisms of Pathologic Coma

Although coma may resemble deep, sleep-like unconsciousness, it is differentiated by a total lack of purposeful response to internal or external stimuli. The motionless resting state of a comatose patient may give way with effortful stimulation to a grimace in response to painful stimuli or stereotyped withdrawal movements of the limbs generated by spinal reflexes. Such movements, however, do not localize the source of external stimulation nor demonstrate organized sequential components characteristic of purposeful avoidance. As coma progresses to deeper stages, such primitive reflex responses may not be evident.

Coma results from a complete disabling of the brain's arousal mechanisms that maintain the alert waking state, including responsiveness to external stimuli, intentional behaviors, and postural control. Arousal regulation includes control of the overall level of neuronal activity (firing rates and metabolic demand) in the brain and organization of spatiotemporal dynamics (patterns of cerebral activity reflected in the electroencephalogram (EEG) and in the spinal cord associated with wakeful behavior). Many different neurological conditions can produce coma, but statistically most coma arises as a result of very diffuse global damage to the brain resulting either from a loss blood flow during cardiac arrest or from severe head trauma typically due to road traffic accidents. As a general rule, if a patient is comatose, one of three possible mechanisms (or combinations of these three) accounts for this level of brain dysfunction: (1) the patient has a diffuse metabolic disturbance (loss of blood flow, oxygen deprivation (e.g., smoke inhalation), poisoning, etc.), (2) the patient has sustained bilateral injuries to the cerebral hemispheres, or (3) the patient has sustained focal injuries to a relatively small set of regions of the upper brain stem or thalamus.

There are several different neuronal populations concentrated in the brain stem and basal forebrain that contribute to maintaining the state of wakefulness, collectively grouped as the 'arousal systems.' The concept of an ascending arousal system located in the dorsal regions of the upper brain stem and central thalamus arose in the 1940s. Initially, the roles of the midbrain reticular formation (MRF) and the central thalamus (primarily intralaminar thalamic nuclei (ILN) and closely associated regions) were emphasized as the origin of this system. Experimental studies demonstrated that electrical stimulation of the midbrain and central thalamus produced EEG desynchronization and behavioral arousal as evidenced by eye opening and eliciting of purposeful movements in lightly anesthetized animals. A modern view incorporates these early observations into a conceptual framework that identifies the control of arousal *per se* as interdependent on the output of glutamatergic and cholinergic neurons located in both the brain stem and the basal forebrain, and several other brain stem neuronal populations including serotoninergic

and noradrenergic neurons, as well as histaminergic nuclei and orexin-producing neurons located in the hypothalamus. Many experimental studies have sought to identify a specific subset of necessary or sufficient neuronal groups uniquely responsible for arousal reflected in the EEG and behavior without clear evidence that any single group is indispensable. Consistent with the results of these experimental studies, it has been seen that only relatively large bilateral injuries to the dorsal regions of the upper pons, midbrain, or central regions of the thalamus produce coma in humans on the basis of a small structural brain injury. Moreover, such comas arising from focal injuries are usually quite short lived, lasting only hours or a few days.

Circumscribed injuries to the brain stem that produce acute coma are concentrated bilaterally in the rostral pons and dorsal midbrain – regions containing cholinergic neurons and other afferents from the brain stem arousal systems that have strong anatomical projections to the basal forebrain, central thalamus, and wide regions of the cerebral cortex. Bilateral focal injury to the central thalamus may also produce acute coma when the injuries involve both sides of the brain and affect either the anterior or the posterior thalamic intralaminar nuclei. This type of injury often includes bilateral brain stem lesions centered on the upper midbrain (MRF). Both isolated injuries of the dorsal brain stem or the central thalamus may produce initial coma with an invariably short duration. Restricted bilateral brain stem lesions often give way to faster and more complete recoveries than bilateral ILN injuries (particularly if the midbrain is involved on one or both sides along with the central thalamus). Patients with brain stem injuries producing coma may recover consciousness within 1 week if they survive. In contrast, patients with isolated central thalamic injuries producing acute coma may show recovery of eye opening and patterned arousal within 1 or 2 days but remain behaviorally unconscious for many weeks or months, if recovering at all. Of note, some patients with bilateral injuries involving the central thalamus and midbrain do recover from enduring global disorders of consciousness over very long intervals. The mechanisms underlying this protracted recovery process are not well understood.

Although coma may arise in the setting of such focal injuries, in most cases coma is associated with widespread neuronal dysfunction or death due to loss of oxygen (anoxia) following a cardiac arrest or head trauma that produces disconnection of neurons by shearing injury of their axonal projections. Nonetheless, in the setting of coma produced by anoxia or brain trauma, it is the severity of damage to the thalamus that best indexes the likely outcome.

Related Global Disorders of Consciousness

Following a coma produced by a severe brain injury (traumatic or nontraumatic), patients either worsen with progressive neuronal damage leading to brain death (total cessation of all brain function) or improve with the initial signs of recovery heralded by brief periods of eye opening without behavioral responsiveness. Comas that do not resolve with more complete recovery of consciousness give way to the vegetative state, minimally conscious state, and other related clinical syndromes. These conditions are often confused with coma, particularly in the popular press, but each is distinct in clinical presentation and natural history following coma.

Vegetative State

In 1972, Jennett and Plum defined the clinical syndrome of the 'persistent vegetative state' noting the dissociation of a limited recovery of arousal as judged by the recovery of an 'eyes-open' state that alternates with an 'eyes-closed' state in a patient with otherwise no evidence of awareness of self or their environment. Understanding of the mechanisms underlying the clinical syndrome of vegetative (VS) has evolved since that time, particularly with respect to the basis of enduring VS in patients who survive cardiac arrest. Current usage arbitrarily identifies VS lasting more than 30 days as the persistent vegetative state. The modifier 'persistent' is often misinterpreted to indicate the likely permanence of VS. An expectation for the permanence of VS, however, depends on both the mechanism of injury and the elapsed time from the injury.

Because VS typically arises following coma, it is not surprising that patterns of structural injuries producing VS overlap those producing coma. Autopsy studies of patients remaining in VS show a common pattern of widespread disconnection of the corticothalamic system leading to extensive loss of thalamic neurons, particularly the central thalamic neurons of the ILN and closely adjacent components of thalamic association nuclei that can alone when damaged produce coma. This pattern is seen following both diffuse axonal injury (DAI) from brain trauma and oxygen deprivation resulting from cardiac arrest, with more cortical damage seen with the latter type of injury. Significant brain stem damage, however, is not commonly found at autopsy for patients remaining VS after structural injuries, emphasizing that VS is primarily a disorder of cerebral integration at the level of the corticothalamic systems. The damage to the thalamus following DAI is indirect as a result of

transneuronal degeneration, whereas oxygen deprivation causes rapid neuronal death after approximately 6 min. This underlying difference in mechanism is likely to play an important role in the very different expected time course of recovery and point beyond which permanence is expected in VS resulting from traumatic brain injury compared to cardiac arrest and similar injuries.

Neuroimaging studies of VS patients have focused on measurements of brain energy usage and functional activation of cerebral networks in response to sensory stimuli. Studies of cerebral metabolism in VS patients using positron emission tomography scans show that the brain in VS is severely reduced in resting metabolism, usually remaining at 30–50% of normal levels. These observed metabolic levels are lower than those found in pharmacologic coma produced by surgical anesthesia. Functional neuroimaging studies of VS patients show a widespread failure of functional connections with loss of responses to elementary sensory stimuli in cortical regions remote from the primary sensory cortices. Specific deficits of integrative responses in the medial parietal region have been consistently observed in VS patients. Occasionally, behavioral and physiological variations of VS have been identified, including evidence of isolated islands of metabolic activity. In one patient who had remained clinically in VS for 20 years, rare single understandable words were emitted, without contingent relationship to environmental stimulation. In this VS patient, left frontotemporal cerebral structures that operated at an abnormally low metabolic rate, but at nearly twice the rates of remaining brain, correlated with remaining elements of the human language system. Similar isolated behavioral fragments have been identified in other VS patients. Typically, patients express easily identifiable, stereotypical, emotional–limbic responses such as grimaces. These emotional expressions likely reflect isolated limbic mechanisms activated by brain stem and basal forebrain structures that lie outside of the thalamocortical systems that typically undergo overwhelming injury in VS.

Minimally Conscious State

VS is importantly differentiated from minimally conscious state (MCS) by evidence of unequivocal but inconsistent evidence of awareness of self or the environment. Behavioral evidence of responsiveness in MCS varies widely from isolated tracking of objects in the visual field to high-level responses such as intermittent verbalization and communication. Although the findings distinguishing VS and MCS patients on the clinical exam may be subtle, pathological studies suggest significant differences in the underlying structural pathology associated with each condition. Some patients with clinical histories consistent with MCS demonstrate no evidence of thalamic injury or severe DAI – findings never observed in VS patients at autopsy. The differences indicate a potential anatomical substrate for residual cerebral function in some MCS patients. Consistent with this picture from autopsy studies, a small number of neuroimaging studies show preservation of distributed network activation in response to auditory stimuli involving cortical association regions in the superior temporal gyrus not seen in VS patients and evidence in some MCS patients of large-scale networks responsive to human language. The presence of recruitable large-scale cerebral networks in some MCS patients suggests a potential substrate for further recovery in these patients. The observations likely underlie the rare, but verified, cases of late recovery of spoken language in some severely brain-injured patients with clinical histories consistent with long-standing MCS.

Akinetic Mutism

The syndrome of akinetic mutism spans a range of related behavioral disturbances in which patients may outwardly appear attentive and vigilant (as often evidenced by deliberate visual tracking of the examiner around the room) but exhibit a paucity of behaviors even when prompted. The pattern of brain injury most commonly associated with this syndrome is bilateral damage of the anterior medial regions of the cerebral cortex, typically arising as a result of a rupture of an arterial aneurysm located at the connection of the anterior cerebral arteries between the two cerebral hemispheres. These cortical regions are closely associated with the central thalamus and basal ganglia, and a similar syndrome can arise from bilateral injuries to these subcortical structures (as well as the MRF). This behavioral syndrome is also sometimes referred to as akinetic mutism and is characterized by severe memory loss, markedly slowed behavioral responses, and a listless appearance. The pattern of injuries associated with this syndrome overlaps with those producing acute coma and VS.

Prognosis in Coma, Vegetative State, and Minimally Conscious State

Coma is a very serious illness and studies indicate that 40–50% of patients in a coma after brain trauma, and 54–88% of patients comatose after a cardiac arrest, will die. Although few studies have evaluated outcomes other than survival or death, patients suffering traumatic brain injuries have a significantly higher likelihood than cardiac arrest patients to make

a good recovery. This is possibly because of the younger age of patients with traumatic brain injuries and the delayed mechanisms of neuronal death following brain trauma and axonal disconnection. Recovery after prolonged traumatic coma is well described and, unlike nontraumatic causes, unconsciousness for 1 month does not necessarily preclude significant recovery. Specific prognostic factors in coma are too numerous for review herein.

Prognosis in VS depends similarly on the underlying mechanism of brain injury. Patients who remain in VS for 3 months after a cardiac arrest or other nontraumatic brain injury producing a loss of blood flow or brain oxygen are considered to be in a permanent vegetative state. Time frames for recovery following posttraumatic VS are considerably longer, and 1 year is required for the expectation of permanence. Very few studies have continued to follow patients remaining in VS at 1 year after brain trauma, and some reports suggest that a small percentage of such patients may show some recovery of conscious awareness past the 1 year time frame.

Prognosis in MCS is the least well characterized because the diagnostic category is relatively new and available studies indicate that significant further recovery after 1 year in MCS can occur. Importantly, MCS patients show faster changes in rate of recovery within the first year after injury from either traumatic or nontraumatic injuries compared to VS patients. Thus, it is important that the sometimes small clinical distinctions be made between these two conditions within the first year after injury. All verified examples of late recovery of communication in severely brain-injured patients have come from the MCS patient population.

Underlying Mechanisms of Pharmacologic Coma

The previous discussion identified an important role for specific brain stem, basal forebrain, and central thalamic neuronal populations in mechanisms of coma and other global disorders of consciousness. Functional neuroimaging studies suggest that pharmacologic coma induced by anesthetics is associated with regionally selective reductions in cerebral metabolic activity or blood flow in the same structures that produce coma with structural injury. In most studies, general anesthetics lower average resting cerebral metabolism to approximately 60–65% of normal level when unconsciousness is achieved. As noted previously, this is a somewhat higher level of metabolism than identified in VS patients. Similar profiles of overall reduction in metabolic activity and regional effects on cerebral substructures have been seen for volatile anesthetics such as halothane and intravenous agents (e.g., propofol). Frontal and parietal cortices (higher level association areas of cerebral cortex) appear to be more strongly affected by anesthetic agents, and a marked reduction of cerebral metabolism in the thalamus and midbrain has been noted for most anesthetics. Decreased co-activation of these cortical and subcortical neurons is interpreted to account for the initial reductions in attention, working memory, and arousal associated with general anesthetics. These functional studies hint at an intersection of common action of anesthetics in MRF and thalamus. The selective suppressive effects of anesthetics on the midbrain and thalamus overlap with the clinical evidence provided previously that functional disturbances of these structures play a key role in mechanisms underlying coma and vegetative state. It is possible that, in general, anesthetics act first to disable selective integrative cerebral functions followed by more global impairment of the corticothalamic system as overall neuronal firing is suppressed through disabling of the midbrain and thalamus systems. It is problematic, however, to assess the functional contribution of specific structures to loss of consciousness seen with anesthetic agents because all generate widespread reductions of cerebral blood flow and metabolism. It is not clear whether anesthesia primarily overwhelms and suppresses neuronal firing rates or whether it might alter specific patterns of neuronal activity within the distributed networks of the frontal and parietal cortices (particularly the medial parietal regions). The further elucidation of the relative importance of the alteration of global cerebral activation versus specific patterns of forebrain activity associated with anesthetic agents will further elaborate the precise relationships between pathological and pharmacological forms of unconsciousness.

See also: Awareness: Functional Imaging; Coma; Consciousness: Theories and Models; Consciousness: Theoretical and Computational Neuroscience; Consciousness: Neural Basis of Conscious Experience; Vegetative State.

Further Reading

Adams JH, Graham DI, and Jennett B (2000) The neuropathology of the vegetative state after acute insult. *Brain* 123: 1327–1338.

Giacino JT, Ashwal S, Childs N, et al. (2002) The minimally conscious state: Definition and diagnostic criteria. *Neurology* 58: 349–353.

Gusnard DA and Raichle ME (2001) Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience* 2: 685–694.

Jennett B (2002) *The Vegetative State*. Cambridge, UK: Cambridge University Press.

Jennett B, Adams JH, Murray LS, et al. (2001) Neuropathology in vegetative and severely disabled patients after head injury. *Neurology* 56: 486–490.

Laureys S, Faymonville ME, Peigneux P, et al. (2002) Cortical processing of noxious somatosensory stimuli in the persistent vegetative state. *Neuroimage* 17: 732–741.

Laureys SL, Owen AM, and Schiff ND (2004) Brain function in coma, vegetative state and related disorders. *Lancet Neurology* 3: 537–546.

Maxwell WL, MacKinnon MA, Smith DH, McIntosh TK, and Graham DI (2006) Thalamic nuclei after human blunt head injury. *Journal of Neuropathology & Experimental Neurology* 65: 478–488.

Parvizi J and Damasio AR (2003) Neuroanatomical correlates of brainstem coma. *Brain* 126: 1524–1536.

Posner JB, Saper CB, Schiff ND, and Plum F (2007) *The Diagnosis of Stupor and Coma,* 4th edn., London: Oxford University Press.

Schiff ND, Ribary U, Moreno D, et al. (2002) Residual cerebral activity and behavioral fragments in the persistent vegetative state. *Brain* 125: 1210–1234.

Schiff ND, Rodriguez-Moreno D, Kamal A, et al. (2005) fMRI reveals large-scale network activation in minimally conscious patients. *Neurology* 64: 514–523.

Wijdicks EF (2001) The diagnosis of brain death. *New England Journal of Medicine* 344: 1215–1221.

Command Systems

J Jing, Mount Sinai School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Animals are characterized by their ability to adaptively distribute their activity among various behaviors, such as foraging, predator avoidance, and reproduction, and do so by integrating the sensation, motivation, and learning experiences. In animals with a well-developed central nervous system (CNS), most, if not all, of these behavioral activities are organized by the CNS. Once a behavior is selected, it has to be executed in an appropriate spatial and temporal patterning. Accordingly, one of the most basic question facing neuroscientists is to elucidate the underlying neural mechanisms that select, initiate, and generate coordinated behavioral activity. One central issue is whether there are specific neural substrates that are responsible for the initiation and generation of behaviors (localist representation), or whether these brain functions are distributed among neuronal elements (distributed representation). The concept of command neurons/systems was initially used to describe some specific neuronal elements/systems that are important for initiation of behavior, and would be consistent with an encoding scheme more closely related to a localist representation. This article provides a historical and current account of the 'command' concept and illustrate how this concept may be essential to our understanding of brain functions, such as sensorimotor transformation, neural coding, and decision making.

Command Systems: Command versus Decision Making

Historical Issues

The debate between proponents of localist and distributed representation or processing has been going on since the beginning of modern studies of the brain in the 1800s and has undergone a variety of manifestations. Currently, most neuroscientists agree that the CNS utilizes both strategies of information processing. Historically, however, it is important to demonstrate that there are specific elements, such as command neurons, that can initiate or release a distinct behavior. Because of apparent utilities of invertebrate preparations where some neurons are large enough to be easily accessible to electrophysiological recordings, many early studies were carried out using these relatively simple preparations. Indeed, the 'command neurons' concept came into usage in the 1960s after Wiesma and Ikeda found that stimulation of a small bundle of neuronal fibers in the ventral nerve cord of crayfish drove the metachronal beating of swimmrets. Subsequently, similar neurons/fibers that can evoke/release a specific pattern of behavior have been identified in both invertebrates (e.g., tail flip escape in crayfish, swimming in leech, escape in the mollusk *Tritonia*, and feeding in the mollusk *Pleurobranchaea*) and vertebrates (e.g., startle escape in teleost fish).

Demonstration of command neurons shows that in the CNS there are specific neuronal elements that occupy a privileged position by releasing a specific behavior; therefore, these neurons may be critical decision-making sites where expression and coordination of different behaviors may be determined. However, in the late 1970s, Kupfermann and Weiss were concerned with the free usage of the term 'command neuron' by some investigators, and they wrote an article to clarify its definition. One of their main suggestions was that command neurons, if useful for understanding the neural basis of behavior, should be used to refer to specific neurons or a neuronal population (as command systems) that are both necessary and sufficient for a behavior, in addition to their substantial activity when the behavior naturally occurs. This strict definition was intended to help investigators characterize actual roles of command neurons/systems in a behavior and exclude pretenders, such as neurons that are not essential for expression of the overall behavior but may initiate a component of the motor pattern when activated (e.g., motor neurons). However, experts' reactions to the criteria at the time were diverse and controversial, and currently only a few examples that meet all these criteria have been found. Here, we adopt a view that, rather than limiting ourselves to a strict definition, a more fruitful approach is to examine the neural basis of behavior from the perspective of network architecture and determine, in this context, what may constitute a command system.

Network Architecture: Command and Central Pattern Generators

Consider a prototypical hierarchical network organization of a motor system (Figure 1(a)). From an evolutionary standpoint, sensory neurons and motor neurons are perhaps the earliest developed neuron types controlling movements. As animals and their behaviors become more complex, interneurons emerge to situate between sensory and motor elements, thus providing additional computational capabilities.

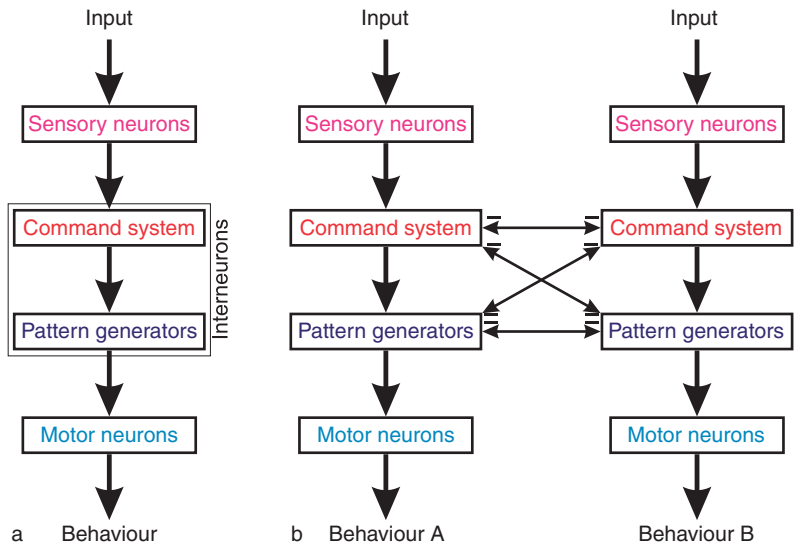

Figure 1 Network architecture underlying behavior. (a) A prototypical motor network organized in a hierarchical manner. The four most basic layers of neural elements organized hierarchically are shown: sensory neurons, interneurons including command systems and pattern generators (or central pattern generators), and motor neurons. In the scheme shown here, sensory inputs act on the sensory neurons, which then act on the command systems, which in turn act on the pattern generators. Finally, pattern-generating elements act on the motor neurons to produce a behavior. (b) Schematic diagram of putative inhibitory interactions between two behavioral networks that may mediate suppression of one incompatible behavior by another.

In addition, in some motor networks, interneurons differentiate further into two layers: a command system and a pattern-generating circuit or a central pattern generator (CPG). The command system integrates sensory information from sensory neurons and activates the CPG, which in turn drives motor neurons to produce a motor pattern. From the perspective of sensorimotor transformation, it is apparent that command neurons are closer to the sensory processing, whereas CPG elements represent movement features. The mechanisms for generating specific motor patterns are intrinsic to the local circuits formed by CPG elements, whereas command system provides excitatory, inhibitory, or modulatory inputs to the CPG.

There are two caveats in considering the prototypical network as a representation of a motor system. First, not all motor networks can be strictly stratified into these layers of elements; actual networks may be variants of this organization. For example, although we can clearly identify command neurons in crayfish lateral giant and fish Mathner cell-mediated escape circuits, there are no pattern generators that intervene between the command neurons and the motor neurons, perhaps because escape is a reflex-like behavior which does not require a dedicated CPG. Another example of a variation is the crustacean stomatogastric network, in which many of the pattern-generating elements are motor neurons rather than interneurons. In contrast, in vertebrate motor systems, there is a proliferation of more layers of neurons at sensory and cognitive processing stages. Despite these differences in network structure, some fundamental principles are commonly applicable in different networks.

Hierarchical and Parallel Organizations

A second notable point is that many neuroscientists now recognize that realistic motor networks do not follow a strict hierarchical organization as depicted in Figure 1(a). Rather, there are multiple parallel, feedforward and feedback interactions among network elements. Historically, hierarchical organization was recognized first. Then, in the 1980s, there was an emphasis on parallel organization (and de-emphasis on hierarchical organization) in neural networks. The current consensus is that both types of organizations are present in the CNS.

Some aspects of the parallel pathways within a hierarchical network are illustrated in Figure 2, which also includes the peripheral structures that are involved in motor behaviors, such as head and other sensory organs that may gather salient information regarding the environment or the goal object, and the plant or motor apparatus that implements the motor behavior. Notably, sensory neurons may provide inputs not only to command systems but also to pattern generators or even motor neurons. How much sensory information is channeled to each of the components may depend on the type of

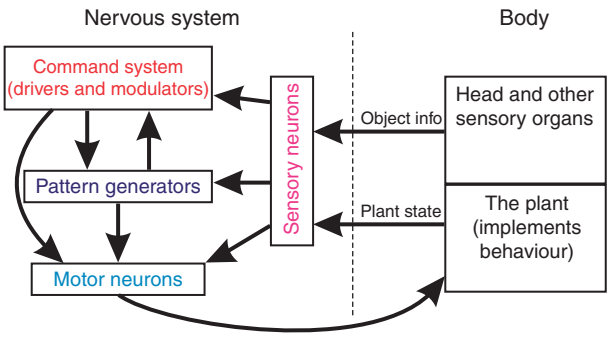

Figure 2 A schematic diagram illustrating a motor network with both parallel and hierarchical organizations. Also depicted are the peripherals: head and other sensory organs that detect salient information about the environment and goal objects, and the plant or the motor apparatus that implements the behavior. Sensory information may be conveyed to command systems, pattern generators, and motor neurons. There are also bidirectional interactions between command systems and pattern generators. Although sensory information is channeled in parallel to different layers, different types of sensory information may preferentially influence different network elements. In one possible scheme illustrated here, information about the goal object that is detected through vision, or tactile/chemical sensory processors, may preferentially act on the command system to select an appropriate behavior, whereas plant state information through proprioceptive receptors may preferentially act on the pattern-generating elements and motor neurons to adjust the motor behavior online to execute behavior adaptively.

behavior. For example, in a reflex behavior, sensory neurons may predominantly act on motor neurons; whereas in a rhythmic behavior sensory neurons may act more via interneurons (command systems/CPG elements). In some cases, this parallel organization of sensory pathways may also be correlated with the types of sensory inputs involved. Thus, object information from the head sensory organs may preferentially act on the command systems to select an appropriate behavior, whereas the proprioceptive information from the plant may preferentially act on the CPG or motor neurons to adjust motor behavior online to execute behavior adaptively. Another feature of the depicted network is that some interactions between layers of elements are not one way but, rather, both ways; for example, feedback from CPG elements to command systems has been well documented. In fact, some command neurons have been shown to serve pattern-generating functions as well. Because structure is clearly related to function, it is critical to determine the network architecture for the motor system under study and also the implications of the architecture for its network functions.

Drivers and Modulators

From the network architecture perspective, it appears that command neurons that drive a CPG circuit are present in many behaviors, especially stimulus-dependent behaviors, that require external stimuli to activate the CPG. Even for autoactive behaviors, higher order neurons may modulate CPG, thereby producing multiple versions or types of motor outputs. Furthermore, with the exception of a few specialized networks in which single neurons may act as a command, it is common for an ensemble of neurons to be involved in command function; therefore, use of command systems or higher order neurons is preferable. In such a command system, at least two major types of interneurons can be distinguished: one type is the drivers, which are the so-called command neurons, and the other is the modulators, which may not initiate a full-blown behavior but, rather, modulate motor outputs. Thus, command systems, as broadly defined here, include both command elements and modulatory elements. Because these elements in command systems are important for expression of specific behaviors, contemporary research focuses on the decision-making role of these neurons – that is, how they act on pattern-generating circuits and how they interact with each other to produce adaptive behaviors. To illustrate these and other relevant issues, we use model systems to discuss the command functions.

Model Systems and Command Functions

As with all branches of biology, many problems are studied using model organisms. Vertebrate models are necessary for ultimate understanding of ourselves. Invertebrate models are particularly informative in uncovering fundamental principles, many of which have been verified in vertebrate models. In terms of command functions, the most extensively studied invertebrates include mollusks (e.g., *Aplysia*, *Clione*, *Lymnaea*, *Pleurobranchaea*, and *Tritonia*), leech, and crustacean (crayfish lateral giant, stomatogastric nervous systems (STNS)). Well-known vertebrate model systems include teleost fish escape, frog prey capture, turtle scratching, locomotory networks (e.g., lamprey), and primate saccade. In the following sections, selected examples are used to discuss how command function contributes to expression of behaviors. In particular, it is useful to distinguish command functions for related behaviors implemented by a single motor apparatus and those for different behaviors implemented by different motor apparatus. Note that such a distinction is operational and only for the convenience of discussion.

Related Behaviors: Reconfiguration of Multifunctional Central Pattern Generators

Our ability to generate a variety of behaviors using a single or related motor apparatus contributes to the

flexibility and diversity of behavioral production. For instance, using our legs, we walk, run, and jump, and we can walk forward and backward. Indeed, one of the major principles derived from work with model systems is that the CPGs that control these motor apparatus are multifunctional; that is, the CPGs can generate multiple motor patterns. Specifically, the operations of CPG depend on the multiple nonlinear processes at the cellular, synaptic, and network levels. Neuronal properties have a set of building blocks at

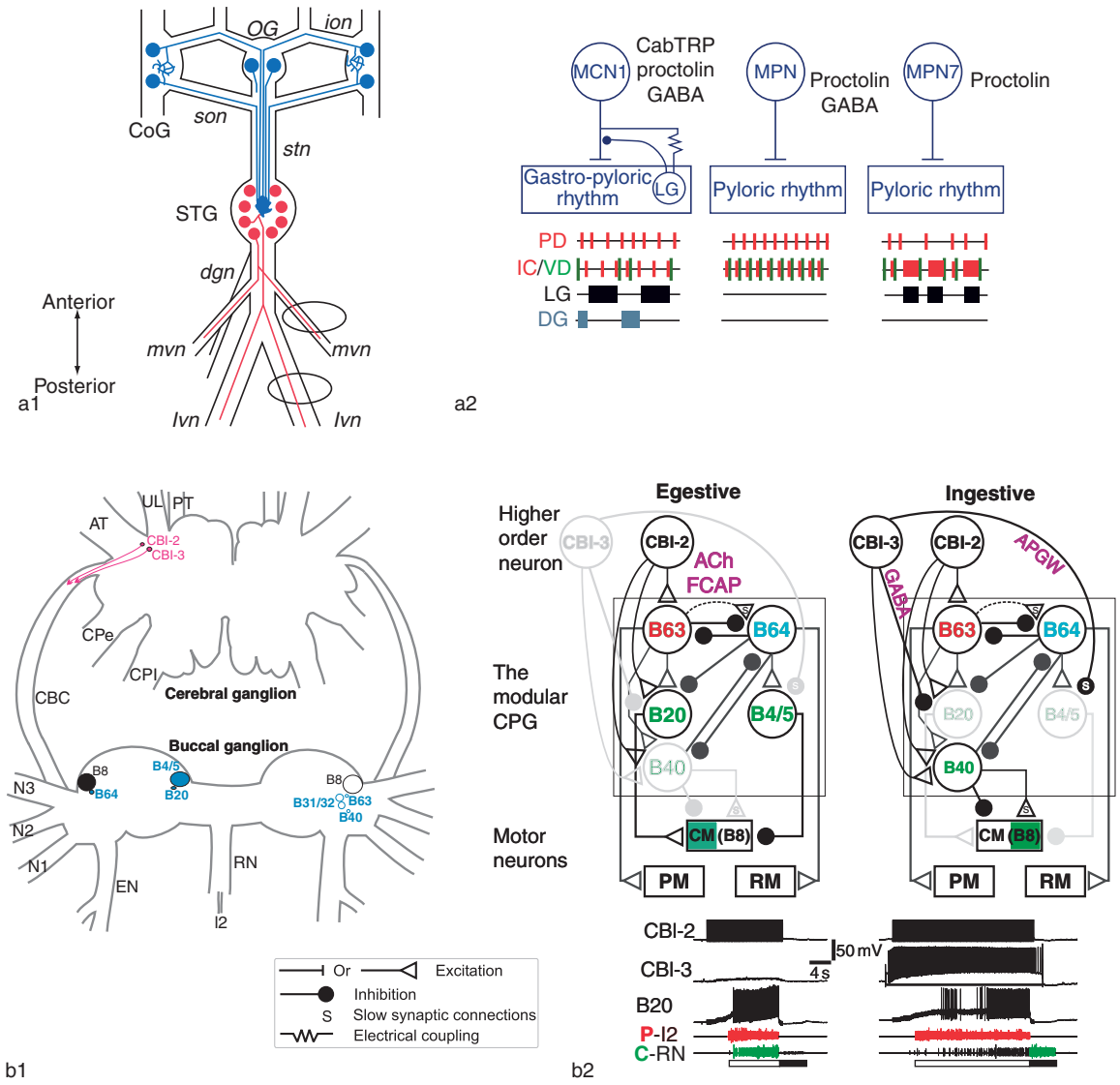

Figure 3 Generation of a variety of related behaviors. (a) Crustacean stomatogastric nervous system and actions of modulatory projection neurons on motor patterns. (a1) Schematic diagram of stomatogastric nervous system showing neuronal distributions in the three ganglia: commissural (CoG) or esophageal (OG) ganglia and the stomatogastric ganglion (STG), involved in the control of foregut movement including pyloric and gastric mill rhythm. Modulatory projection neurons are located in the CoG or OG, which are connected to STG through stomatogastric nerve (stn). CPG elements, many of which also innervate muscles through peripheral nerves (dgn, mvn, and lvn), are located in the STG. (a2) Three projection neurons – MCN1, MPN, and MCN7 – all contain peptide protoclin, but they promote distinct motor patterns as shown in the activity of CPG elements PD, IC, VD, LG, and DG in part because they contain different complements of cotransmitters (e.g., GABA and peptide CabTRP). Reprinted by permission from Macmillan Publishers Ltd: *Nature,* Nusbaum MP and Beenhakker MP (2002) A small-systems approach to motor pattern generation. *Nature* 417: 343–350, copyright 2002. (b) *Aplysia* CNS involved in feeding and combinatorial actions of cerebral–buccal interneurons (CBIs) on the CPG to effect motor program switching. (b1) Schematic diagram of *Aplysia* CNS showing neuronal distributions in cerebral ganglion and buccal ganglion involved in the control of radula movements. Higher order neurons, or CBIs, are located in cerebral ganglion, whereas CPG interneurons and motor neurons are located in buccal ganglion (the left side shows cells on the rostral surface and the right side shows cells on the caudal surface). Food information enters the brain, in part, through the cerebral nerves (AT, UL, and PT), whereas the buccal mass (including the

each level – for example, cellular-level properties (ionic channels and active and passive electric properties of the cell) and synaptic-level properties (nature of synapses, short-term dynamics, and long-term plasticity). All these nonlinear processes can be modulated by neurotransmitters and hormones, and therefore the CPG can assume a variety of functional configurations that correspond to different motor outputs. One of the most well-studied systems is the crustacean stomatogastric ganglion (STG), which contains all CPG elements necessary for control of movements of the foregut, including pyloric and gastric mill rhythms. A large body of work has shown that a variety of neuromodulators (e.g., dopamine, serotonin, and neuropeptides) can alter cellular and synaptic properties of stomatogastric CPG neurons to produce different motor patterns.

Importantly, higher-order neurons/command systems are also prominent sources for many of the neuromodulators, in addition to fast-acting neurotransmitters. Two examples illustrate how elements of a command system are composed of a heterogeneous population of interneurons, which allows the command system to reconfigure or modulate the CPG, thereby producing a variety of motor patterns. In STNS, these higher-order neurons are termed projection neurons. Their somata are located in commissural or esophageal ganglia, but their axons innervate the STG, where they release neurotransmitters and neuromodulators to act on the CPG (**Figure 3(a1)**). An important insight gained from this work is that projection neurons with the same neuropeptide may elicit different versions of motor patterns because these neurons contain different complements of cotransmitters. As shown in **Figure 3(a2)**, all three projection neurons contain proctolin. However, the motor patterns that these projection neurons promote are distinct, in part due to the fact that they have different complements of cotransmitters.

In *Aplysia* feeding network, two major ganglia are involved (**Figure 3(b1)**). The cerebral ganglion contains a number of higher-order interneurons, termed cerebral–buccal interneurons (CBIs), that receive sensory information from head sensory organs and project their axons to the buccal ganglion, where they activate and/or modulate the feeding CPG. The buccal ganglion contains the feeding CPG elements and also motor neurons that innervate the feeding apparatus, the radula. Notably, the *Aplysia* feeding system may entail an additional level of complexity, not present in some simpler networks, in that most of the CPG elements are interneurons, which form a local circuit that underlies motor patterns and provides timing information to motor neurons. Furthermore, studies suggest that the CPG interneurons did not evolve as a duplicate of single motor neurons to simply provide excitatory drive to them. Rather, CPG interneurons provide both excitation and inhibition to motor neurons and represent a higher-level abstraction in that they appear to be organized as functional modules that encode behaviorally relevant movement parameters. In particular, the radula produces two major types of motor programs, ingestive and egestive, that involve different coordination patterns of two sets of radula movements: protraction–retraction versus opening–closing. Protraction initiates first, and it always precedes retraction in each cycle of a motor program. In ingestion, closing occurs during retraction; in egestion, closing occurs during protraction. Different CPG interneurons are responsible for protraction or retraction. Furthermore, because closing occurs at different times in the two programs, two sets of CPG interneurons control radula closing for these two programs, and therefore the activity levels of these interneurons will determine what types of programs are generated.

Such a modular organization engenders an elegant mechanism with which higher-order neurons can act on the CPG to produce different motor patterns.

feeding apparatus, the radula) and the gut are innervated by buccal nerves (N1, N2, N3, RN, and EN). I2 is a branch of RN and contains the axons of protraction motor neurons (PM), such as B31/32. Closing motor neuron (CM), B8, exits the buccal ganglion through RN. (b2) Motor program switching. The pattern-generating network appears to be organized in a modular manner. Protraction interneuron, B63, and retraction interneuron, B64, mutually inhibit each other and mediate alternative protraction (open bar) and retraction (filled bar) movements. Two sets of CPG elements, B20–B4/5 and B40, mediate closing B8 activity for egestive and ingestive programs, respectively. In egestive programs, B20 excites B8 during protraction, whereas B4/5 inhibits it during retraction; thus, B8 is mostly active during protraction (green). In ingestive programs, B40, active during protraction, provides fast inhibition and slow excitation on B8 so that B8 is mostly active during retraction (green). The ability of CBI-2 to induce a motor program is mediated in part by its excitation (using ACh and possibly peptide FCAP) on B63, which is a critical initiating element of the CPG. (Left) CBI-2 promotes an egestive program by exciting B20, which in turn excites B4/5 (the latter connection not shown for clarity). (Right) Coactivation of CBI-3 with CBI-2 converts the program to ingestive because CBI-3 inhibits B20 through a fast GABAergic synapse, whereas it inhibits B4/5 through a peptidergic (APGW) mechanism. CBI-3 also promotes B40 activity, reinforcing the ingestive program. Data are summarized from Jing J and Weiss KR (2001) Neural mechanisms of motor program switching in *Aplysia*. *Journal of Neuroscience* 21: 7349–7362; Jing J, Vilim FS, Wu JS, Park JH, and Weiss KR (2003) Concerted GABAergic actions of *Aplysia* feeding interneurons in motor program specification. *Journal of Neuroscience* 23: 5283–5294; and Hurwitz I, Kupfermann I, and Weiss KR (2003) Fast synaptic connections from CBIs to pattern-generating neurons in *Aplysia*: Initiation and modification of motor programs. *Journal of Neurophysiology* 89: 2120–2136, and references cited within.

Figure 3(b2) illustrates how two CBIs use their different sets of cotransmitter complements to initiate and modulate programs to effect motor program switching. Specifically, CBI-2 contains the fast-acting excitatory neurotransmitter ACh and peptide cotransmitter FCAP, both of which promote activation of the CPG (in this case, an egestive program). By contrast, when coactivated with CBI-2, CBI-3 exerts fast GABAergic and slow peptidergic (APGW) inhibitions separately on the two egestive CPG elements active at two different phases to convert the program to ingestive. Similarly, combinations of other sets of CBIs (e.g., CBI-2 and CBI-12) can act on the modular CPG to produce changes in other movement parameters, thus constituting a flexible and adaptive mechanism for generation of a variety of related behaviors. In summary, the CPG contains basic mechanisms that underlie motor patterns, whereas various combinations of higher order interneurons act on the CPGs and sculpt a functional configuration appropriate for any given condition.

Different Behaviors: From Behavioral Hierarchy to Adaptive Decision Making

We constantly make behavioral decisions. In addition to selecting a behavior from related behaviors that are generated by a single or related motor apparatus, we also choose among behaviors that are implemented by different motor apparatus. In an early series of studies, Davis and others asked why animals generally perform one type of behavior at any given time. Specifically, animal behaviors appear to be organized into a priority sequence and some behaviors take precedence over others, which has been termed a 'behavioral hierarchy.' In terms of the neural mechanisms, it is presumed that some type of inhibitory interactions might exist between the networks that mediate incompatible behaviors. Although interactions may involve elements in both the command systems and the CPG, elements of command systems appear to be one of the major sites of actions (Figure 1(b)). Indeed, inhibitory interactions between command elements in one network and neural elements in the other network have been identified in several model systems, such as *Pleurobranchaea* (feeding inhibition of withdraw and swimming over feeding), *Clione* (withdraw/hunting over swimming), and leech (feeding over tactile responses).

Recent studies of behavioral selection, however, have revealed a more complex picture, which can be illustrated by work on *Pleurobranchaea* and leech. First, even suppression of incompatible behaviors does not have to be mediated by pure inhibition, as is the case in the suppression of avoidance turning by escape swimming in *Pleurobranchaea* (Figure 4(a)). Mild noxious stimuli cause prolonged activity in the serotonergic interneurons that support avoidance turning by providing excitation to turning interneurons. Strong noxious stimuli, however, set the network in a state in which the serotonergic interneurons as well as turning interneurons are co-opted into a swimming rhythm and prevented from driving avoidance turning. Second, interactions between network elements may involve both excitation and inhibition. In *Pleurobranchaea*, although suppression of feeding by escape swimming involves inhibition of feeding command neurons by command and CPG elements in the escape swimming network, the serotonergic neurons involved in swimming do not inhibit feeding command neurons. Rather, they provide monosynaptic fast and slow excitation to feeding modulatory neurons that enhance feeding activity, and they also provide slow, albeit weak, excitation to feeding command neurons. Similarly, in the interactions between networks that mediate shortening and swimming in leech (Figure 4(b1)), there are prominent excitatory inputs from interneurons for shortening to the triggering and gating swimming interneurons that are part of swimming command systems, with only one of the swimming gating neurons being inhibited by the shortening interneurons. Based on these and other findings in the leech, Kristan and colleagues proposed a combinatorial code scheme that could account for how leech makes a behavioral decision to shorten, swim, crawl, or bend (Figure 4(b2)). It is suggested that a different but partially overlapping population of interneurons is active at any given time, and this specific combination of interneurons determines which behavior is to be expressed (see also Figure 3(b2)).

Two major insights can be gained from the previously described studies. First, they support the postulate that the neurons that comprise behavioral networks constitute a heterogeneous population. Although some neurons are more dedicated to a single behavior, most neurons are multifunctional. For instance, although the serotonergic interneurons in *Pleurobranchaea* and *Tritonia* were initially identified as part of escape swimming CPG, studies performed to date indicate that these neurons may also function as gating mechanisms because they are activated by mild to strong noxious stimuli, and the activity level in these neurons then determines which behavior is expressed. At lower levels of activation, these neurons enhance the overall arousal level of the animal and promote withdrawal, avoidance, locomotion, or even feeding by providing both fast and slow modulatory excitations to networks mediating these behaviors. At higher levels of activation, these neurons promote escape swimming, resulting in more dedicated swimming

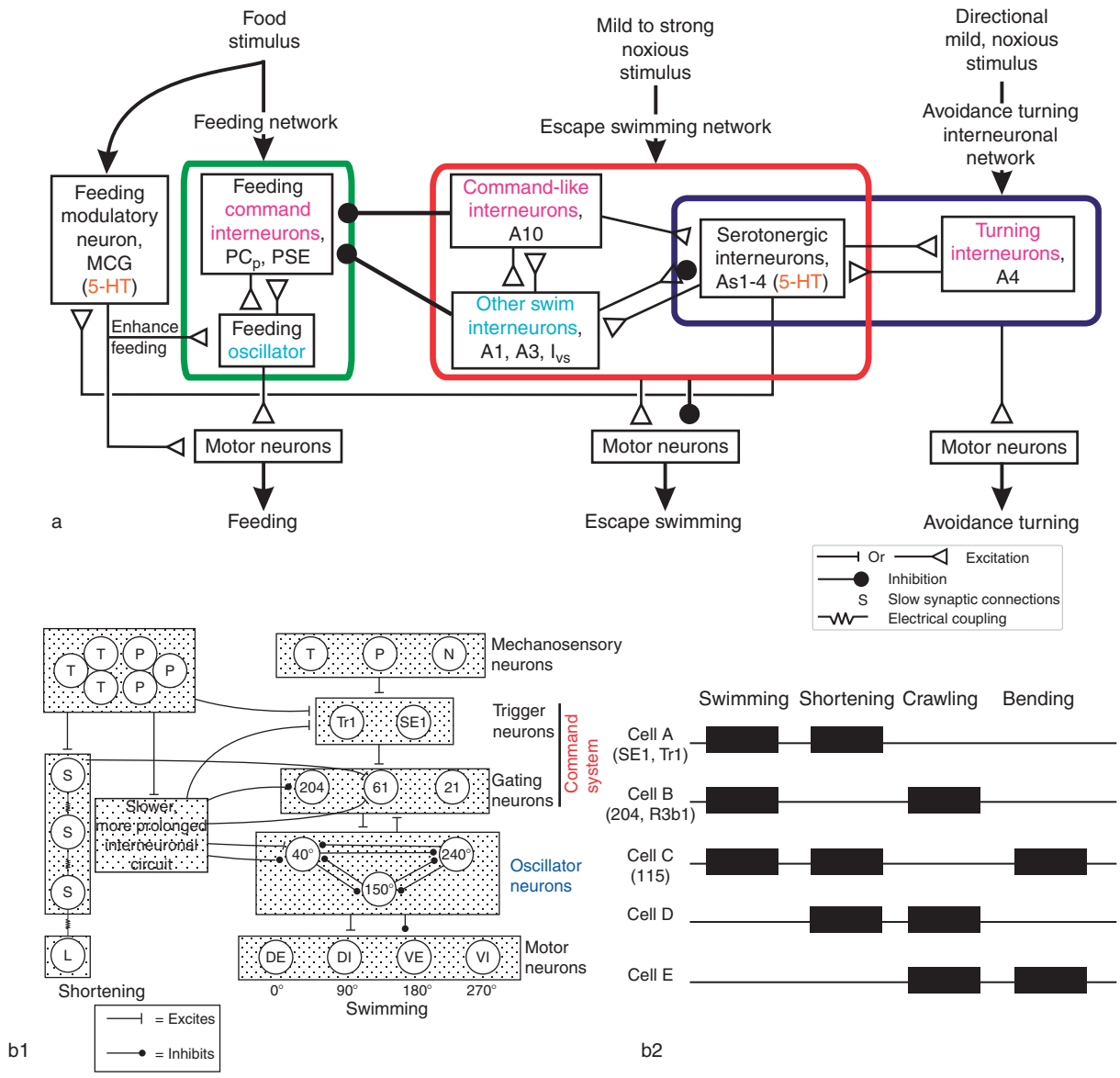

Figure 4 Interacting neural networks underlying behavioral decisions. (a) Highly interactive networks that mediate feeding, escape swimming, and avoidance turning in *Pleurobranchaea*. Suppression of feeding by escape swimming is mediated by inhibition of feeding command neurons by swimming command-like neurons and other CPG neurons. However, serotonergic swimming interneurons As1–4 do not inhibit feeding command neurons; rather, they provide slow, albeit weak, excitation to them (not shown). In addition, As1–4 provide fast and modulatory excitations to feeding modulatory neurons that enhance the feeding. As1–4 are also integrative members of the turning network by providing strong excitation to turning interneuron A4 to promote A4 unilateral activity that underlies avoidance turning. Suppression of avoidance turning by escape swimming is mediated in part by co-option of As1–4 and A4 into swimming rhythm when sufficiently strong noxious stimuli set the network in a state to produce escape swimming. Finally, As1–4 may also promote locomotion by driving activity in locomotor neurons (not shown). 5-HT, serotonin. Data are summarized from Jing J and Gillette R (2000) Escape swim network interneurons have diverse roles in behavioral switching and putative arousal in *Pleurobranchaea*. *Journal of Neurophysiology* 83: 1346–1355; and Jing J and Gillette R (2003) Directional avoidance turns encoded by single interneurons and sustained by multifunctional serotonergic cells. *Journal of Neuroscience* 23: 3039–3051, and references cited within. (b) Network interactions underlying decision making in leech. (b1) Interactions between shortening and swimming networks. Although shortening suppresses swimming, shortening interneurons only inhibit one swimming interneuron (204) while providing excitations to many other swimming interneurons. (b2) A combinatorial code scheme for decision making. It is proposed that at any given moment, an overlapping but distinct set of interneurons is activated (black rectangle) that determines which behavior – swimming, shortening, crawling, or bending – is expressed. Names of interneurons of the three neuron types, A–C, that have already been documented are indicated in parentheses at left, whereas cell types D and E are postulated. Reproduced with permission from The Society for Integrative and Comparative Biology: Esch T and Kristan WB Jr., (2002) Decision-making in the leech nervous system. *Integrative and Comparative Biology* 42: 716–724, by permission of Oxford University press.

interneurons phasically inhibiting other behavioral networks to prevent incompatible behaviors from occurring. From this perspective, these neurons are functionally analogous to vertebrate arousal networks such as the raphe nucleus and locus coeruleus. Second, these highly interactive networks may well contribute to the flexibility in behavioral selection and expression in animals. Thus, even simple animals such as mollusk and annelid are not rigid robots and do not act as servants to a multitude of environmental stimuli. Rather, animals have an intrinsic repertoire of motor behaviors that can be gracefully recruited into action by these decision-making neurons/systems which form elaborate networks and choose to respond appropriately to salient stimuli at any given time. Indeed, vertebrates are characterized by further development of decision-making systems into more complex forebrain structures (e.g., cortex, basal ganglia, and hypothalamus).

More Elaborate Neural Networks Involved in Motor Commands in Vertebrates

The work in invertebrate model systems is informative for the understanding of neural organization of vertebrate motor systems. For example, motor system organization for a number of vertebrate behaviors, such as locomotion, scratching, respiration, occulomotor control, and feeding, shares a grossly similar architecture with some behaviors in invertebrates (Figure 5). This similarity is first reflected in the fact that there are clearly CPG networks, either in the spinal cord or in the brain stem, that are responsible for execution of these motor behaviors. These CPG networks are composed mostly of interneurons, and unlike some invertebrates, each functional group is composed of a population of a large number of interneurons rather than one or a few. However, the essential building blocks for the CPG networks (or, as termed by Grillner and others, motor infrastructure) are similar to those in invertebrates. Interestingly, based in part on experimental work and theoretical/computational analysis in turtle scratching, lamprey locomotion, and other limb behaviors, Bizzi, Grillner, Stein, and others proposed a modular organization for limb control (compare the modular organization proposed for an invertebrate CPG in *Aplysia* feeding). One version of the proposal stipulates that extensor and flexor of the hip and knee are controlled by separate functional modules or unit CPGs such that a variety of motor behaviors (e.g., different forms of scratching and locomotion) can be constructed through flexible combinations of these modules.

Whereas CPG networks are mostly located in spinal cord or brain stem, command signals largely originate from descending fibers of the higher brain centers, such as brain stem (midbrain and hindbrain), diencephalon (thalamus and hypothalamus), and telencephalon (basal ganglia and cortex) (Figure 5(a)). These command signals can activate/modulate the specific CPG network appropriate for a given condition. Indeed, vertebrate higher level networks that fulfill command and other higher-order functions are much more elaborate. First, there may be multiple command sites for a behavior, such as the two command sites demonstrated for locomotion – that is, the mesencephalic (MLR) and the diencephalic locomotor (DLR) regions that may provide more flexible means to access the locomotor CPG. Second, there may be multiple layers of commands that exert more sophisticated controls of a behavior. For example, Swanson proposed that motivated behaviors such as ingestive/reproductive/defensive and exploratory/foraging behaviors are controlled by command sites in the brain stem (e.g., MLR) called 'motor pattern initiators,' which in turn are controlled by 'pattern controllers' in the diencephalon (e.g., DLR).

Increased complexity of vertebrate motor systems results from additional, higher levels of control centers (Figure 5(b)). These control centers include basal ganglia, whose output nuclei, pallidium, are spontaneously active and exert a tonic GABAergic inhibitory control over lower command sites and the CPGs. On the other hand, input nuclei, the striatum, are normally quiescent, and when active they inhibit palladium also using GABA, thereby initiating behaviors through a disinhibitory mechanism. An additional level of control is the excitatory influence from the cortical networks that are responsible for aspects of sensory processing, motor control, and cognition. There are multiple parallel and hierarchical interactions in subcortical and cortical regions, such as those involved in primate saccade: superior colliculus, visual cortex; lateral intraparietal area (parietal cortex); frontal eye fields (prefrontal cortex); and medial eye fields (premotor cortex). Finally, vertebrates also evolved specialized neural systems that can significantly influence behavioral decisions: for example, the cerebellum for motor learning and precision motor control, prefrontal cortex for working memory, and hippocampus for declarative memory.

Importantly, similar to invertebrate motor networks, variations in network architectures are also present in vertebrates. For instance, teleost fish developed a specialized Mauthner cell-associated command system in the brain stem for startle, and the Mauthner cell's excitatory outputs to spinal motor neurons are direct, which is adaptive considering the importance of response speed in escape. Some primate behaviors, such as reaching and, in particular, fine motor skills (e.g., speech and hand/finger

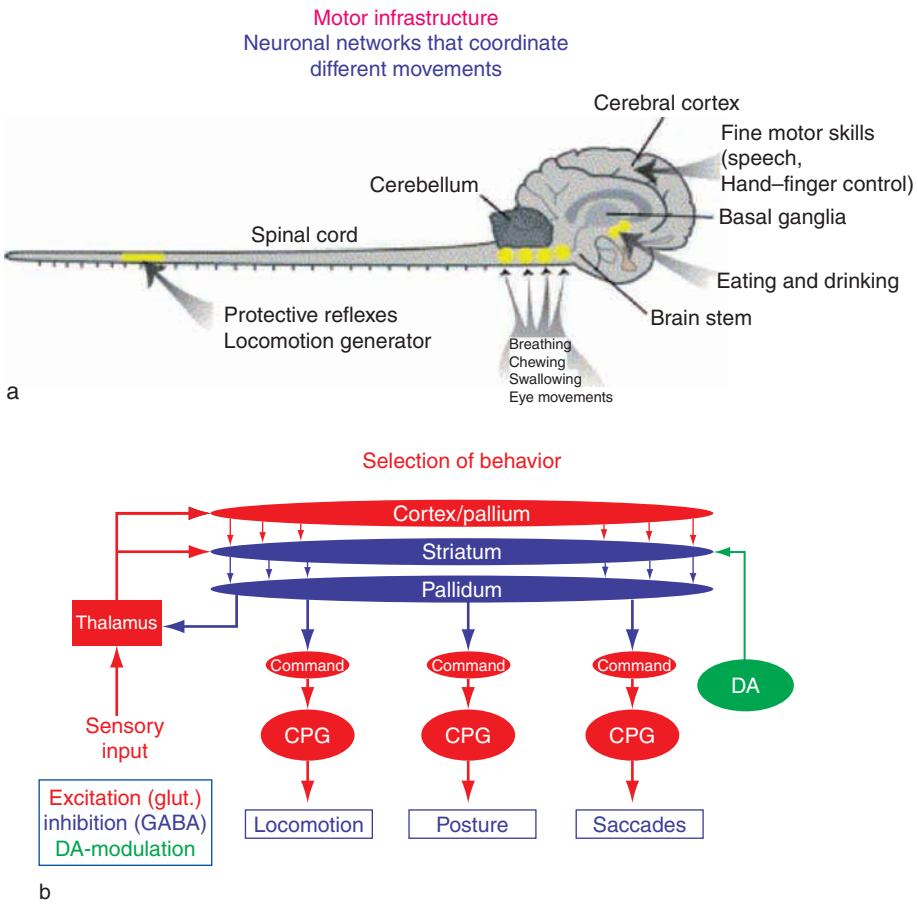

Figure 5 Neural organization of vertebrate motor networks. (a) Illustration of vertebrate CNS including spinal cord, brain stem, cerebellum, basal ganglia and cerebral cortex. In the spinal cord and brain stem are the CPG networks or motor infrastructure that mediate a variety of behaviors, such as protective reflexes, locomotion, breathing, chewing, swallowing, and eye movements. Eating and drinking are also heavily influenced by neural networks located in the hypothalamus. Fine motor skills (e.g., speech and hand–finger control) are under increasingly direct control of cortical neurons, particularly in primates. (b) Simplified diagram of vertebrate motor systems involved in behavioral selection. Motor networks for locomotion, posture, and saccade, as illustrated here, are organized in at least two major layers – the CPG networks located in the spinal cord or brain stem and the command sites (likely gultaminergic (glut.)) in midbrain. The command sites are further under the control of cerebral cortex/pallium and basal ganglia (input nuclei, striatum; output nuclei, pallidum). In general, command sites are under tonic GABAergic inhibition of pallidum which is spontaneously active. Specific striatum neurons which have high activation threshold and are normally inactive may be activated, in part, from appropriate sensory input, channeled through thalamus. Upon activation, these striatum neurons, also GABAergic, inhibit pallidum, thus initiating a behavior through a disinhibitory mechanism. Cortex also exerts excitatory influence on striatum to initiate an appropriate behavior. Also illustrated are midbrain dopaminergic (DA) neurons that exert a modulatory influence to striatum to promote expression of specific behaviors. Reproduced from Grillner S (2006) Biological pattern generation: The cellular and computational logic of networks in motion. *Neuron* 52: 751–766, with permission from Elsevier.

control), represent another type of specialization in which the cortical neurons play a larger role in direct control of these behaviors, although spinal/brain stem interneurons may still play a role.

Summary

This article presented an overview of the command concept in the context of neural architecture underlying behavior, which may be a useful vantage point for understanding network functions in both invertebrate and vertebrate motor systems. With regard to sensorimotor transformations, the command systems lie between neural elements responsible for sensory processing and motor execution, and they function as decision makers and serve other higher order functions by transforming sensory information into signals that activate and modulate CPG networks for an appropriate behavior. In particular, command systems that comprise only dedicated command neurons are rare. More commonly, command systems are composed of a population of neurons that can be differentiated based on their morphology, synaptic inputs and outputs, transmitter phenotype, etc. Some may play a role in driving the CPG, whereas others may play a role in modulating the CPG.

These interneurons function based on the principle of 'division of labor' commonly observed in evolved systems, such as biological and social systems. A major theme of current and future research, therefore, is to delineate the functional roles of these heterogeneous neural components in the command systems, in particular, and any neural system in general. In this regard, the advantage to directly link a neuron or a neuronal population to their functional roles in a behavior in invertebrate model systems is and will remain essential. Furthermore, developments and improvements in techniques in both invertebrates and vertebrates (e.g., use of genetic organisms, computational approaches, and recording of neural populations) promise to provide a more integrative understanding of how animals make behavioral decisions and how the behavioral sequences are mediated.

See also: Behavioral Hierarchies; Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Computational Neuroethology; Motor Primitives; Neuromodulation; Pattern Generation; Population Coding; Startle Response.

Further Reading

Briggman KL, Abarbanel HD, and Kristan WB Jr. (2005) Optical imaging of neuronal populations during decision-making. *Science* 307: 896–901.

d'Avella A and Bizzi E (2005) Shared and specific muscle synergies in natural motor behaviors. *Proceedings of the National Academy of Sciences of the United States of America* 102: 3076–3081.

Edwards DH, Heitler WJ, and Krasne FB (1999) Fifty years of a command neuron: The neurobiology of escape behavior in the crayfish. *Trends in Neurosciences* 22: 153–161.

Esch T and Kristan WB Jr. (2002) Decision-making in the leech nervous system. *Integrative and Comparative Biology* 42: 716–724.

Ewert J-P (2003) Command neurons and command systems. In: Arbib MA (ed.) *The Handbook of Brain Theory and Neural Networks*, pp. 233–238. Cambridge, MA: MIT Press.

Frost WN and Katz PS (1996) Single neuron control over a complex motor program. *Proceedings of the National Academy of Sciences of the United States of America* 93: 422–426.

Getting PA (1989) Emerging principles governing the operation of neural networks. *Annual Review of Neuroscience* 12: 185–204.

Grillner S (2006) Biological pattern generation: The cellular and computational logic of networks in motion. *Neuron* 52: 751–766.

Hurwitz I, Kupfermann I, and Weiss KR (2003) Fast synaptic connections from CBIs to pattern-generating neurons in *Aplysia*: Initiation and modification of motor programs. *Journal of Neurophysiology* 89: 2120–2136.

Jing J and Gillette R (2000) Escape swim network interneurons have diverse roles in behavioral switching and putative arousal in *Pleurobranchaea*. *Journal of Neurophysiology* 83: 1346–1355.

Jing J and Gillette R (2003) Directional avoidance turns encoded by single interneurons and sustained by multifunctional serotonergic cells. *Journal of Neuroscience* 23: 3039–3051.

Jing J and Weiss KR (2001) Neural mechanisms of motor program switching in *Aplysia*. *Journal of Neuroscience* 21: 7349–7362.

Jing J and Weiss KR (2005) Generation of variants of a motor act in a modular and hierarchical motor network. *Current Biology* 15: 1712–1721.

Jing J, Vilim FS, Wu JS, Park JH, and Weiss KR (2003) Concerted GABAergic actions of *Aplysia* feeding interneurons in motor program specification. *Journal of Neuroscience* 23: 5283–5294.

Korn H and Faber DS (2005) The Mauthner cell half a century later: A neurobiological model for decision-making? *Neuron* 47: 13–28.

Kristan WB Jr., Calabrese RL, and Friesen WO (2005) Neuronal control of leech behavior. *Progress in Neurobiology* 76: 279–327.

Kupfermann I and Weiss KR (1978) The command neuron concept. *Behavioral and Brain Sciences* 1: 3–39.

Marder E, Bucher D, Schulz DJ, and Taylor AL (2005) Invertebrate central pattern generation moves along. *Current Biology* 15: R685–R699.

Nusbaum MP and Beenhakker MP (2002) A small-systems approach to motor pattern generation. *Nature* 417: 343–350.

Page M (2000) Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences* 23: 443–512.

Poggio T and Bizzi E (2004) Generalization in vision and motor control. *Nature* 431: 768–774.

Shadmehr R and Wise SP (2005) *Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge, MA: MIT Press.

Swanson LW (2003) *Brain Architecture: Understanding the Basic Plan*. New York: Oxford University Press.

Sherman SM and Guillery RW (1998) On the actions that one nerve cell can have on another distinguishing 'drives' from 'modulators'. *Proceedings of the National Academy of Sciences of the United States of America* 95: 7121–7126.

Stein PS, Grillner S, Selverston AI, and Stuart DG (eds.) (1997) *Neurons, Networks, and Motor Behavior*. Cambridge, MA: MIT Press.

Communication in Frogs and Toads

M J Ryan, University of Texas, Austin, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Anurans, commonly known as frogs and toads, are among the most vociferous of animals. Their nocturnal cacophony characterizes many landscapes and seasons, such as vernal ponds in the temperate-zone spring and rainforests in the neotropical wet season. This serenade is an annual mating ritual. Males broadcast long-distance calls to attract females and repel other males, and females use these calls to choose mates.

Anuran sexual communication has emerged as an especially tractable system for integrative studies of brain, behavior, and evolution. There are about 5000 species of frogs, and the long-distance mating or advertisement calls is their most obvious signal. A caveat, however, should be implicit throughout this account. Anurans encompass incredible diversity in communication, and this article merely describes some of the more common themes in the manner in which these animals acoustically communicate with each other.

Frogs have a variety of calls that communicate different types of information. These include long-distance mating or advertisement calls, territorial calls, aggressive calls, encounter calls, distress calls, release calls, short-distance courtship calls, and rain calls. Not all species produce all these call types, and most have a rather modest repertoire. One glaring exception, however, is *Boophis madagascariensis* (Rhacophoridae), a frog found in the forests of eastern Madagascar. Call notes produced by these males have been classified into 28 types (Figure 1), the largest known call repertoire of any anuran, although little is known of the function of these call types.

Most studies of anuran communication concentrate on one call in the frog's repertoire, the mating call, which mediates social interactions during mate attraction. Because the mating call figures so strongly in female mate choice, this dyadic interaction is critical in ensuring successful reproduction; furthermore, it can drive speciation between populations by promoting mate recognition of only local populations, and female preferences among appropriate mates can cause the evolution of increasingly elaborate signals through the action of sexual selection.

Origin and Mechanisms of Sound Production

Modern frogs produce sound by expelling air from the lungs through the larynx. The larynx sits on a ring of cricoid cartilages on the hyoid plate. It consists of two arytenoid cartilages that close on the top of the larynx and which, inside, support a pair of vocal folds, one extending across the interior of each arytenoid cartilage (Figure 2). As air passes through the lungs, it vibrates the vocal folds, and it is this vibration that brings about the pressure fluctuations that are perceived as sound. The mass and tension of the vocal folds are primarily responsible for the spectral properties of the emitted sound. In some species such as the túngara frog, *Physalaemus pustulosus*, masses can be added to the vocal cord to decrease its fundamental frequency of vibration (Figure 2). The opening and closing of the arytenoid cartilages can shape the waveform of the sound.

The frog's larynx opens into the floor of the mouth. In many frogs, there are vocal slits in the mouth that open into a vocal sac that expands as the call is produced. The vocal sac is usually quite thin and enhances the coupling of the sound to the environment. The vocal sac seems not to be a cavity resonator; in *Pseudacris crucifer* and *Physalaemus pustulosus*, the dominant frequency of the call does not change when males call in a helium–oxygen mixture (as it would if it were a resonator, since frequency is the speed of sound divided by wavelength, and the speed of sound varies with medium density while the emphasized wavelength of a resonator remains constant). In an unusual case of sound coupling, the large external tympanic membrane of the male bullfrog couples sound to the environment. Thus these frogs call as well as hear through their ears. Frogs can also make use of the external environment to enhance signal coupling. A treefrog in Malaysia, *Metaphrynella sundana*, for example, calls in tree holes and varies the call's dominant frequency to match the hole's resonance.

Vocal sacs can be quite varied in appearance (Figure 3) and have other functions besides acoustic coupling. Air is recycled between the vocal sac and the lungs; thus buccal pumping is not required to inflate the lungs after each call. This saves energy, in *P. pustulosus* at least, and allows males to call at a faster rate, which is more attractive to females. A male frog's large pulsating vocal sac can also serve as a visual cue. In both *Hyla squirella* and *P. pustulosus*, the vocal sac is a visual cue that further increases the attractiveness of the acoustic cue to the female.

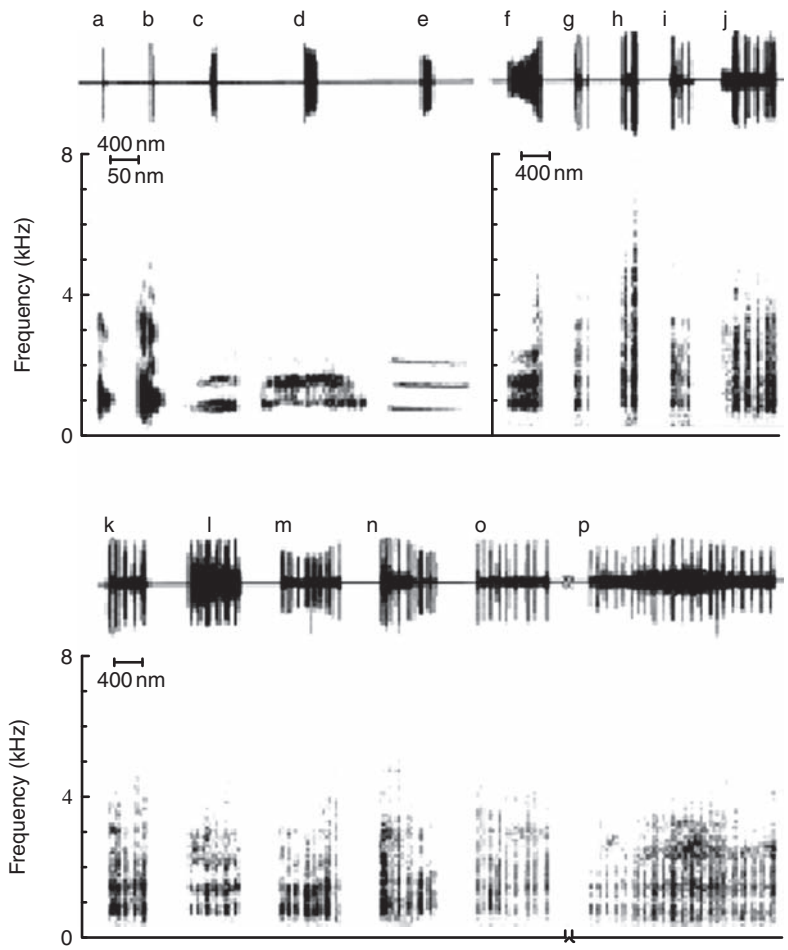

Figure 1 Waveforms and sound spectrograms of representative call notes produced by males of *Boophis madagascariensis*: a, toc note; b, short click note; c, short rip note; d, loud click note; e, tone-like note with a fundamental frequency of 670 Hz; f, long rip note; g, creak note; h ± p, iambic (I) notes with increasing numbers of pulses: (h) I2, (i) I3, (j) I4, (k) I5, (l) I6, (m) I7, (n) I8, (o) I9, (p) I23. Except for I21, all notes from I10 to I22 were also produced, but are not illustrated. Analysis filter bandwidth: a and b: 300 Hz; c and d: 150 Hz; all other spectrograms: 59 Hz. Ambient recording temperatures were in the 18–22 °C range. Timescale for sound spectrograms a ± e has been magnified to provide increased temporal resolution. Reproduced from Narins PM, Lewis ER, and McClelland BE (2000) Hyper-extended call note repertoire of the endemic Madagascar treefrog *Boophis madagascariensis* (Rhacophoridae). *Journal of Zoology (London)* 250: 283–298, with permission from Blackwell Publishing.

Receivers

In frogs, more so than in most other animals, the neural properties of the receiver are inextricably linked to sexual communication. This is most clearly illustrated by Capranica's matched-filter hypothesis, which suggests that among species, the tuning of two auditory end organs matches the distribution of spectral energy in the mating call.

The Auditory System

Most research has concentrated on how the frog's auditory system decodes the mating call. But frogs do attend to other sounds. Reed frogs (*Hyperolius nitidulus*) in Africa can hear the roar of fire rolling across savanna, and they respond by jumping into the water. But there is little doubt that the hearing of frogs has primarily evolved to enhance sexual communication.

All anurans have two inner ear organs, the amphibian papilla (AP) and the basilar papilla (BP), that are sensitive to airborne sound. Furthermore, the saccule is sensitive to surface vibrations. In some frogs, such as the white-lipped frog, *Leptodactylus albilabris*, the surface vibrations that result from the vocal sac's striking the ground are perceived by the receiver as a salient part of the vocal display.

Tympanic membrane and middle ear Frogs lack pinnae. Some frogs, such as some members of the genus *Atelopus*, lack both external ears and middle ears, even though these frogs clearly communicate acoustically.

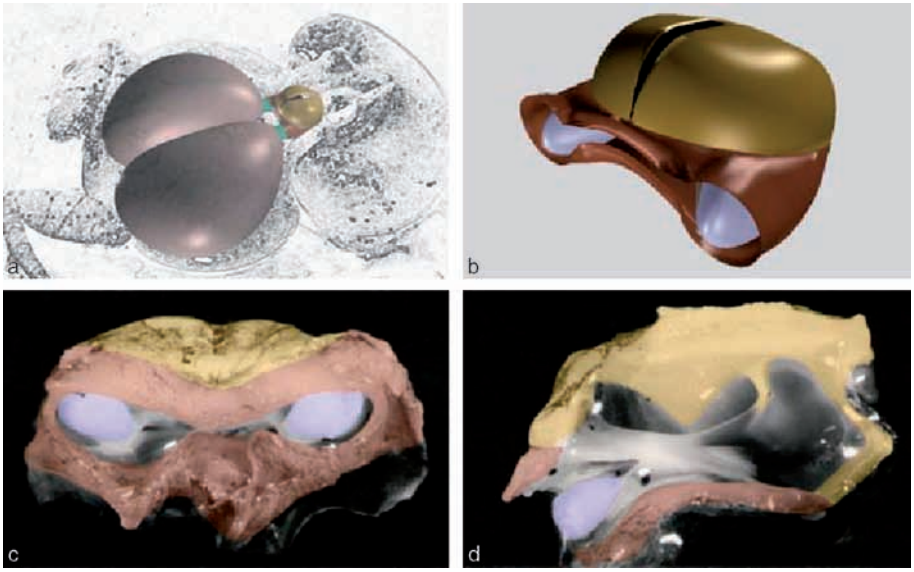

Figure 2 The laryngeal morphology of the túngara frog. The arytenoid cartilages are in yellow, the cricoid cartilage in red, the fibrous masses in blue, the vocal folds in white, the bronchi in green, and the lungs in pink. (a) Approximate position of the larynx and lungs in the calling frog. (b) Simplified illustration of the larynx without bronchi or lungs. (c) A view of the larynx from the lungs showing the intrusion of the fibrous masses into the bronchi. (d) A medial section of the larynx showing the attachment of the fibrous mass to the vocal fold. Reproduced from Gridi-Papp M, Rand AS, and Ryan MJ (2006) Complex call production in túngara frogs. *Nature* 441: 38, with permission.

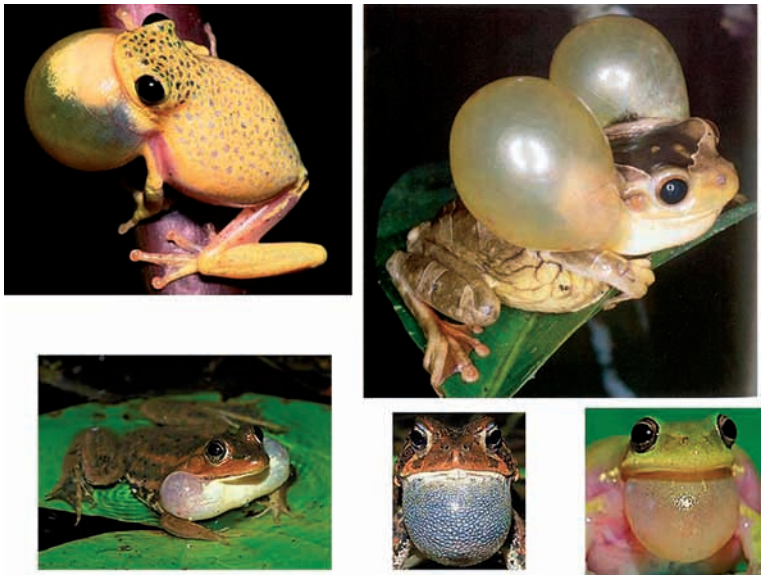

Figure 3 A sample of the diversity of vocal sacs of frogs. Clockwise from top left: African reed frog (*Hyperolius bayoni*), from www.exploratorium.edu/frogs/mainstory/hbayoni.html; Amazonian arboreal frog (*Phrynohyas coriacea*), from Hofrichter R (ed.) (2000) *Encyclopedia of Amphibians* (p.162). Toronto, Canada: Key Porter Books; squirrel treefrog (*Hyla squirella*), American toad (*Bufo americanus*), and carpenter frog (*Rana virgatipes*), all from www.naturesound.com.

In most frogs, however, there is a tympanic membrane on the outside of the head and a collumella in the middle ear, which transduces outer ear vibrations to the oval window of the inner ear. This vibration causes disturbance of the inner ear fluid, which in turn causes movements of membranes in both the AP and BP.

Inner ear Both the AP and BP are involved in sound detection, but the details of these structures are quite different. The AP has about 1000–1500 hair cells in the bullfrog, *Rana catesbeiana*, and its membrane is tonotopically organized. The AP is most sensitive to lower frequencies, from a few hundred Hz

to about 1500 Hz. The AP usually has two peaks of frequency sensitivity. In the bullfrog, for example, there is a lower peak of about 100–200 Hz and a mid-frequency peak of about 500–600 Hz. Neurons most sensitive to the lower-frequency peak exhibit two-tone suppression, but this is not true of the mid-frequency peak. The BP, by comparison, has fewer hair cells, 60–90 in the bullfrog, and is not tonotopically organized. The hair cells in the BP tend to be most sensitive to the same frequency, about 1400–1500 Hz in bullfrogs.

The tuning of the AP and BP of a species tends to coincide with the frequency peaks in its mating call. In frogs that call in a narrow frequency range, there is a match between the tuning of either the AP or the BP and call frequency, whereas in frogs that call over a wider frequency range, the tuning of both end organs match frequency peaks in the call. A recent review of these data showed that species variation in the carrier frequency of calls explains 82% of the variation in the tuning of the inner ears organs (Figure 4).

At least one species of frog, the concave-eared torrent frog (*Amolops tormotus*), uses ultrasonics in communication. This Chinese frog calls near rushing streams, and the ultrasonics are above most of the spectral energy in the background noise. There is

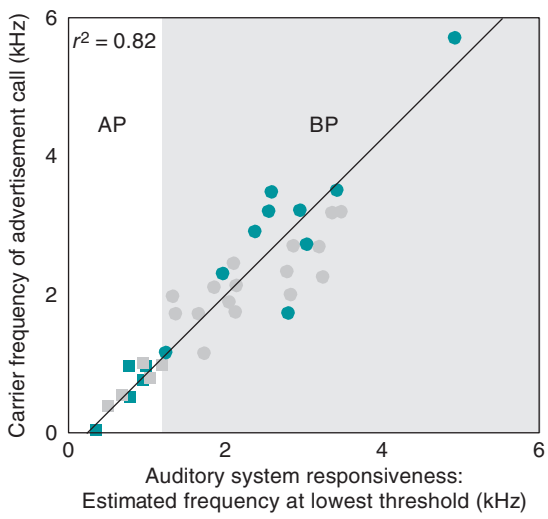

Figure 4 Scatter diagram of the spectral peaks of the advertisement calls of 24 anurans (plus three populations of one species) versus estimates of the best excitatory frequency of the peripheral auditory system. Green circles represent data in which closed-field stimulation was used in the neurophysiological recordings, and gray symbols show data in which free-field stimulation was used. Squares represent low frequencies attributed to the amphibian papilla (AP), and circles represent sensitivities attributed to the basilar papilla (BP). The white and gray backgrounds estimate the range of frequency sensitivities of the AP and BP, respectively. Modified from figure 7.1 of Gerhardt HC and Schwartz JJ (2001) Auditory tuning and frequency preferences in anurans. In: Ryan MJ (ed.) *Anuran Communication*, pp. 73–85. Washington, DC: Smithsonian Institution Press.

some behavioral evidence to suggest that the ultrasonics influence the male's calling behavior, and neurophysiological studies of evoked potentials show clearly that the frogs can hear these sounds.

Central Processing

Much of the integration of acoustic information from the two inner ear organs takes place in the torus semicircularus, a large auditory nucleus in the midbrain homologous to the mammalian inferior colliculus. Electrophysiological studies show more complex processing in the torus than in the inferior colliculus: Some cells respond to simple stimuli such as tones, whereas others respond only to specific combinations of sounds such as those found in the species' mating call. Studies using electrophysiology and immediate early gene (IEG) expression have mapped spatial variation in auditory activity in the torus in response to call variation and found evidence for parallel processing of different aspects of acoustic stimuli. In túngara frogs, combining IEG expression measures throughout the torus provides sufficient information to discriminate between conspecific and heterospecific calls as well as between simple and complex conspecific calls.

Further acoustic processing takes place in thalamic auditory nuclei. As with the torus, parallel processing occurs here. Lesioning studies in gray treefrogs show that female phonotaxis requires a functioning torus semicircularis but not dorsal thalamus. Dorsal thalamic nuclei might function in motivation or attention relative to phonotactic readiness.

Thalamic and midbrain auditory nuclei have anatomical projections to various forebrain regions that receive inputs from multiple sensory modalities. Electrophysiological and metabolic measures have found acoustic responsiveness in auditory system targets in the hypothalamus and telencephalon. In the túngara frog, IEG analyses have also shown that correlated patterns of neural activity within the hypothalamus vary with signal salience. Lesioning dopaminergic cells in the green treefrog, *Hyla cinerea*, impaired phonotaxis behavior, and the residual phonotaxis behavior correlated with the number of surviving dopamine cells in one region of the hypothalamus. These IEG and lesioning studies suggest the existence of social neural networks that integrate auditory and motivation information and influence the motor outputs that constitute phonotaxis.

Behavior and Evolution

Other Modes and Multimodal Communication

Frogs communicate in modalities besides the auditory one. Tadpole social behavior, for example, is strongly

Figure 5 Visual signaling in *Phrynobatrachus kreftii*. Female (above left) and male without vocal sac expansion (above right), with partly expanded vocal sac (below right), and with fully expanded vocal sac (below left). Note the bright yellow subgular region of the male. Photograph by Walter Hödl at Zigi River (Amani Nature Reserve), Tanzania. From figure 1 of Hirshmann W and Hödl W (2006) Visual signaling in *Phyrnobatrachus krefftii* Boulenger, 1909 (Anura: Ranidae). *Herpetologica* 62: 18–27.

influenced by olfactory cues. In aggregations of tadpoles of the American toad, *Bufo americanus*, siblings are more likely to be found in proximity to one another than to nonsiblings. These and many species of tadpoles can discriminate sib from nonsib and paternal half-sib from sib using odor cues that become incorporated into the jelly matrix of the nest.

Visual cues are important in sexual communication in many diurnal frogs, such as the poison frogs of the family Dendrobatidae. *Dendrobates pumilio* on different islands in the Bocas del Toro region of Panama show striking differences in color pattern, and females' preference for local males is strongly influenced by visual cues. Other diurnal frogs, such as *Hylodes* and *Atelopus zeteki*, have dynamic visual displays, waving their limbs at other frogs in a stereotyped pattern. Both frogs live near rushing streams, where the background noise might not be conducive to acoustic communication when rain is abundant.

The vocal sac can also serve as a visual cue. Male poison frogs, *Epipedobates femoralis*, attack models with pulsating vocal sacs. In a diurnal frog from Tanzania, *Phrynobatrachus krefftii*, males silently inflate their bright yellow vocal sacs as a visual signal (Figure 5). The vocal sac can also contribute to multimodal communication.

Male Calling

Most frogs form choruses to vocally advertise for females. As choruses are associated with reproduction, there are seasonal patterns; choruses form in the breeding season but not out of it. There are also diurnal patterns. Most frogs call at night, but some, perhaps most notably poison frogs, call during the day. There are also diel patterns as most frogs do not call from dusk to dawn. In the El Verde forest of Puerto Rico, which is rich in frogs of the genus *Eleutherodactylus*, there is acoustic partitioning of frequencies within choruses, but when call frequencies are similar between sympatric species of *Eleutherodactylus*, they call at different times at night.

Temporal patterning also occurs at a finer scale. Only a few species of frogs, such as the neotropical *Smilisca sila*, synchronize their calling. In this case, synchronization makes the frogs less susceptible to predation by the frog-eating bat *Trachops cirrhosus*. Most frogs, however, alternate calls. Call alternation usually occurs within a small neighborhood, and thus, the summed acoustic output of an entire chorus can lack any easily discernible structure. Some of the best studies of these types of interactions have been conducted with the neotropical treefrogs *Hyla microcephala*, *H. phlebodes*, and *H. ebraccata*. Males of these species adjust their patterns of calling to avoid overlap of calls with conspecifics as well as similar-sounding congenerics. Males of both *Eleutherodactylus coqui* and *Hyperolius broadleyi* have especially impressive abilities to produce calls in gaps between the calls of other males.

In many species of chorusing frogs, nonrandom spacing of calling males occurs whether or not males defend territories such as oviposition sites. In some species, such as the neotropical *Eleutherodactylus diastema* and the North American spring peeper,

Pseudacris crucifer, the amplitude of the nearest neighbor's call, rather than the neighbor's physical distance, appears to be more important in determining nearest-neighbor distances.

Mate Recognition

One of the most important decisions an animal can make is with whom to mate. This single act can ensure reproduction between individuals and the persistence of a genetic lineage. The mating call is the dominant cue used by males to advertise their presence and identity to females, and females attend carefully to call attributes when choosing a mate. There is variation in mating calls among species, populations, and males within populations. Much of this signal variation is salient to a female and influences her choice of mates.

Species recognition Species recognition occurs when individuals recognize and prefer members of their own species as mates over members of other species. As noted above, the frog's auditory system is biased toward perceiving and responding to calls of their own species in preference to calls of other species. This functional correlation between sender and receiver predicts that females should be able to discriminate easily among calls in favor of conspecifics. Phonotaxis experiments with numerous species have shown this to be the case. In addition, a number of studies have used synthetic calls and systematically deconstructed them to demonstrate the features salient to females. Two main points have emerged. Not all the acoustic detail is salient, and different species rely on different constellations of call parameters for call recognition. An especially interesting example is mate discrimination between two treefrogs *Hyla cinerea* and *H. gratiosa*. These species co-occur throughout much of their range, and when they do so, their calls are more similar to one another than they are to any of the other local species. Females from each species are able to discriminate between the two calls, but they do so differently. *H. cinerea* uses spectral cues to make the discrimination whereas *H. gratiosa* attends mostly to temporal features.

If there is strong selection on females to mate with conspecific males rather than heterospecific ones, then the constellation of species at the breeding site should influence aspects of one another's communication system. This phenomenon is known as reproductive character displacement, and its occurrence has been perhaps best demonstrated by studies of frogs. These studies usually compare signal variation of two species in areas in which they do not co-occur (allopatry) and in areas where they do co-occur (sympatry). Reproductive character displacement predicts that calls of the two species evolve to be more different in sympatry. This is the case in two species of chorus frogs in Florida, *Pseudacris nigrita* and *P. ferriarum*. Their calls are a series of pulses, and the pulse rate is known to be important in mate recognition. In the area of sympatry, *P. ferriarum* has a faster pulse rate compared with allopatry, about 30 pulses s^{-1} compared with 20 pulses s^{-1}. This faster rate enhances the difference between *P. ferriarum* and *P. nigrita*, whose rate is about 20 pulses s^{-1} whether these frogs are sympatric or allopatric with *P. nigrita*. A similar situation occurs in the Australian frogs *Litoria ewingii* and *L. verreauxii* (Figure 6). Their pulse rates are similar in allopatry. But the calls of each species differ between sympatry and allopatry. Specifically, in sympatry the *L. ewingi* call is slower (68 versus 73 pulses s^{-1}) and the call of *L. verreauxii* is much faster (138 vs. 84 pulses s^{-1}). In phonotaxis experiments, the species can distinguish each other's calls in sympatry, but *L. ewingii* in sympatry cannot distinguish its call from the similar calls of *L. verauxii* in allopatry. Thus not only have the calls of the two species diverged in sympatry, but these differences are salient to the receiver, and if the calls had not diverged in sympatry, there would be breakdown of species recognition.

Most studies of reproductive character displacement have concentrated on evolutionary changes in the signal rather than the receiver, but the receiver can also respond to selection to better avoid calls of sympatric heterospecifics. There are two species of gray treefrogs, *Hyla chrysoscelis* and *H. versicolor*, that are indistinguishable morphologically but differ in chromosome number (diploid and tetraploid) and the pulse pattern of their call (short pulses and long pulses). *H. chrysoscelis* is usually allopatric with

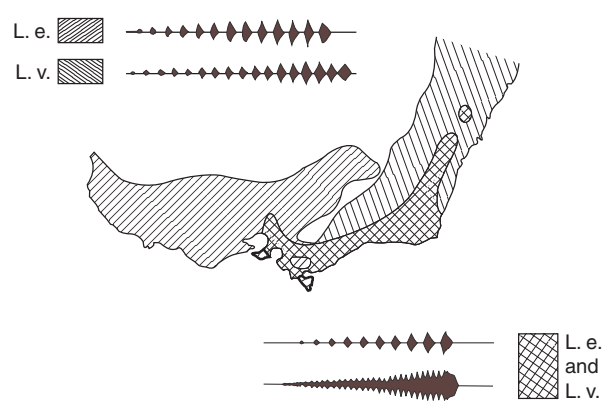

Figure 6 The geographic distributions of *Litoria ewingii* (L. e.) and *L. verreauxi* (L. v.) and the waveforms of their advertisement calls. The two species are sympatric in the crosshatched area. Modified from Halliday T (1980) *Sexual Strategy* (p. 74). Chicago: University of Chicago Press.

H. versicolor, and females prefer the longer-pulsed calls produced by their males. When they are sympatric with *H. versicolor*, however, such a preference could increase their chances of mating with the wrong species. In sympatry, this preference for longer-pulsed calls among conspecific males by *H. chrysoscelis* is significantly reduced.

Sexual selection Sexual selection by female choice occurs when female mating preference results in some males in the population having greater reproductive success than others. The most thorough studies in frogs show that (1) there is female choice of mates, (2) there is substantial variation in male mating success as a function of some aspects of the male's phenotype, (3) there is a correlation between this aspect of the male's phenotype and some aspect of his call, and, (4) those aspects of the call make the male more attractive to females in phonotaxis experiments. As an example, larger male túngara frogs in the wild have greater mating success than smaller ones do, larger males have lower-frequency calls because they have a larger larynx, and in phonotaxis experiments, females are more attracted to synthetic calls with lower frequencies than to those with higher frequencies.

Females of species in addition to túngara frogs prefer lower-frequency calls. Even more species prefer calls with faster pulse rates. A general theme is that females prefer more acoustic energy. They often tend to prefer faster calls, longer calls, higher-amplitude calls, and calls that stimulate one rather than two peripheral end organs. A recurrent question ever since Darwin is why females prefer some males over others. One hypothesis is that females should choose males who provide an immediate and direct benefit in reproductive success. There is good evidence for this in frogs. In some cases, females prefer males that increase the number of offspring sired. In bullfrogs, larger males have territories that afford the eggs more protection from leech predation. In other cases, females prefer larger males because they fertilize more eggs. This is not because of differences in sperm supply, however, but because there is mechanically a better match of the male's and female's gononpores during external fertilization. The better the match, the more the sperm come into contact with the eggs. Females of some species, such as túngara frogs, accrue this advantage in fertilization success through a simple preference for larger males, while others, such as the Australian frog *Uperoleia rugosa*, choose males who are not only larger than average but of a specific size relative to the female's own size.

Another hypothesis for the evolution of female choice is based on indirect benefits: Females should choose males whose genotypes are superior for survival. In one study, male gray treefrogs that made energetically more expensive calls were preferred by females, and half-sib breeding experiments have shown that tadpoles of these males metamorphose faster and at larger size. This effect, however, is exhibited only under certain ecological conditions. The call, in this case, seems to advertise genetic variation for at least some components of survival. Another type of genetically based choice is to shun matings with genetically similar males to avoid inbreeding. In American toads, information about male relatedness is encoded in the calls, and females prefer calls of less-related males. An analogous study in túngara frogs, however, showed no such an effect.

A more recent hypothesis to explain female mate choice is sensory exploitation. As mentioned above, the response properties of the female's auditory system are biased toward acoustic parameters of the species' mating call. Males are under selection to make their calls more attractive to females, and one way to accomplish this is to evolve calls with acoustic parameters that exploit the females' sensory, neural, and cognitive biases. In túngara frogs, males always produce a whine whose dominant frequency matches the best excitatory frequency of the AP, and males can add up to seven secondary components or chucks that stimulate the best excitatory frequency of the BP. Males add chucks to their calls when they vocally compete with one another. Simple calls, whines only, are both necessary and sufficient to attract females and elicit calling from other males. The chuck by itself is not salient to either sex, but the addition of chucks to a whine makes the whine more attractive to females and elicits more calling from males. There seems to be nothing special about the chuck except that it stimulates the female's auditory system. Other sounds, such as noise, tones, bells, and whistles, increase the whine's attractiveness to the same degree as a chuck.

Closely related species that do not produce a chuck have the same BP tuning, so it seems clear that the chuck and its dominant frequency evolved in túngara frogs to stimulate the tuning of the inner ear organ that already existed. In addition, in some of these closely related species without chucks, females prefer the calls of their own species to which the chuck of a túngara frog is added. As this short review of female choice shows, there are a number of different scenarios that can favor the evolution of female choice among conspecifics, and these forces could act in concert.

Eavesdroppers on Frog Communication by Predators

When animals communicate, there is, at minimum, a sender and an intended receiver. But other animals

Figure 7 The frog-eating bat *Trachops cirrhosus* capturing a calling male túngara frog. Photo by Merlin Tuttle, Bat Conservation International.

might be listening to exploit the communication system. These other animals are called eavesdroppers. They are especially common in sexual communication systems in which long-distance signals are loud and conspicuous.

The best known eavesdropper on frog calls is the frog-eating bat *Trachops cirrhosus* (Figure 7). These bats feed on frogs, and in Panama they are especially fond of túngara frogs. These bats are especially sensitive to the relatively low-frequency sounds in frog calls, and they use the calls to localize the male. Like female túngara frogs, these bats are attracted to simple, whine-only calls, and also like female túngara frogs, they are more attracted to whines with chucks than to whines alone. A group of flies in the genus *Corethrella* shows analogous reactions to túngara frog calls. These flies take a blood meal from frogs; they use the frog's call for localization; and although the flies can locate males producing simple calls, they are preferentially attracted to complex calls. These cases of eavesdropping illustrate nicely the conflict often seen in communication systems between sexual and natural selection. The former favors males that make conspicuous signals that are more attractive to females, but these more attractive calls are simultaneously selected against because of their low survival probabilities.

See also: Communication Networks and Eavesdropping in Animals; Communication in Terrestrial Animals; Game Theory and the Economics of Animal Communication; Multimodal Signaling in Animals; Seismic and Vibrational Signals in Animals; Sexual Selection and the Evolution of Animal Signals; Signal Transmission in Natural Environments.

Further Reading

Feng AS, Narins PN, Xu C-H, et al. (2006) Ultrasonic communication in frogs. *Nature* 440: 333–336.

Fritzsch B, Ryan MJ, Wilczynski W, Hetherington TE, and Walkowiak W (eds.) (1988) *The Evolution of the Amphibian Auditory System*. New York: Wiley.

Gerhardt HC and Huber F (2002) *Acoustic Communication in Insects and Anurans*. Chicago: University of Chicago Press.

Gridi-Papp M, Rand AS, and Ryan MJ (2006) Complex call production in túngara frogs. *Nature* 441: 38.

Hoke KL, Burmeister SS, Fernald RD, et al. (2004) Functional mapping of the auditory midbrain during mate call reception. *Journal of Neuroscience* 24: 11264–11272.

Hoskin CJ, Higgie M, McDonald KR, and Moritz C (2005) Reinforcement drives rapid allopatric speciation. *Nature* 437: 1353–1356.

Narins PM, Grabul DS, Soma KK, Gaucher P, and Hödl W (2005) Cross-modal integration in a dart-poison frog. *Proceedings of the National Academy of Sciences of the United States of America* 102: 2425–2429.

Ryan MJ (1985) *The Túngara Frog: A Study in Sexual Selection and Communication*. Chicago: University of Chicago Press.

Ryan MJ (2001) *Anuran Communication*. Washington, DC: Smithsonian Institution Press.

Waldman B and Bishop PJ (2004) Chemical communication in an archaic anuran. *Behavioral Ecology* 15: 83–93.

Welch AM, Semlitsch RD, and Gerhardt HC (1998) Call duration as an indicator of genetic quality in male gray tree frogs. *Science* 280: 1928–1930.

Wilczynski W and Capranica RR (1984) The auditory system of anuran amphibians. *Progress in Neurobiology* 22: 1–38.

Communication in Terrestrial Animals

K McComb and D Reby, University of Sussex, Brighton, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Research on vocal communication in terrestrial mammals had until recently lagged behind work on birds and anurans (frogs and toads), perhaps because of the greater difficulty of recording, analyzing, and playing back acoustically more complex signals to animals that are usually free-ranging and difficult to confine for experimental purposes. One exception is the nonhuman primates, which have received intense scrutiny because of their shared ancestry with humans. Although primates can be classified as terrestrial mammals, we do not attempt to comprehensively cover research on this group here but, rather, focus on other terrestrial mammals, where advances in research techniques (enabled by the revolution in digital technology) are beginning to yield interesting results. Research on these mammals integrates study at several levels of investigation, including the functional anatomy of the vocal apparatus, the acoustic structure of the calls, the context of their emission, and their relationship and coevolution with the social structures and reproductive strategies of the animals that give them. As a consequence, research in this area uses a range of methods and tools, including anatomical investigation, sound analyses, and resynthesis and playback experiments. Although mammalian vocal communication is qualitatively different from human speech, humans are also mammals and our mechanisms of vocal production are largely shared. This means that a common framework for understanding the structure and evolution of mammal vocalizations already exists – the source–filter theory of voice production. The realization that this is the case has led to major advances in our scientific understanding of terrestrial mammal calls.

How Mammal Calls Are Produced

The Source and the Filter

In most terrestrial mammals, a vocalization is the product of a source signal, generated by vibration of the vocal folds in the larynx, that is subsequently filtered in the cavities of the vocal tract. This source signal is typically a periodical wave with a fundamental frequency and harmonic overtones that are integer multiples of it (Figure 1). The fundamental frequency (F_0), sometimes called the glottal pulse rate, is determined by the rate of vibration of the vocal folds and is responsible for the pitch of the vocalization: the higher the rate of vibration, the higher pitched the vocalization is perceived to be. However, after being generated at the larynx, the source signal must pass through the cavities of the vocal tract before radiating out through the mouth and nostrils into the environment. During this process, the vocal tract acts as a tube of air with natural resonances and selectively amplifies certain frequencies in the source spectrum. This filtering process shapes the spectral envelope of the signal, producing peaks called 'formants.' The perceptual correlate of formant frequencies is more difficult to characterize, but it is often referred to as the timbre of the vocalization.

One of the important consequences of the source–filter theory of vocal production is that characteristics of vocalizations that arise from inherent properties of the filter can vary independently from those that arise from the source, and either or both may provide receivers with important information. In particular, two key sender characteristics are likely to be coded in mammal calls: information on individual identity, which in large terrestrial mammals usually results from random variation in the shape of the vocal apparatus or the way in which it is operated by individuals, and information on fitness (including size, condition, and age), usually a consequence of physical or physiological constraints that affect the range of variation of acoustic features or the rate of calling.

In principle, pitch should be inversely related to the length and mass of the vocal folds: therefore, smaller animals generally produce higher calls. However, although this is broadly the case across species (mice have higher pitched calls than elephants), empirical data show that within species, pitch is a relatively poor indicator of sender physical attributes, mostly reflecting large-scale, age-related variation (in most species the calls of juvenile individuals are higher pitched than those of adult females, which are in turn higher than those of adult males). However, pitch does not accurately reflect finer scale differences between individuals of the same age/sex classes. In contrast, pitch varies within individuals, and its range, variation, and quality are likely to convey information on motivational state. For example, it is predicted that pitch should rise with increasing levels of fear and stress since vocal folds are likely to become stiffer in more tense animals. Another characteristic of the source that may convey information on motivational state is its periodicity. Many mammalian calls are characterized by a lack of periodicity, caused by irregular vibration of the vocal folds, which confers a harsh quality on the

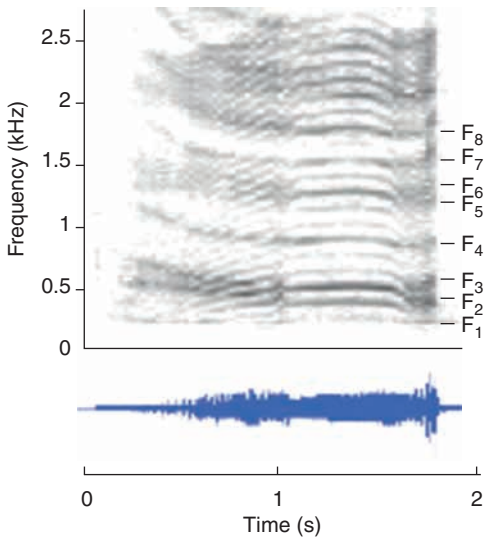

Figure 1 Waveforms and spectrogram of a common roar from a red deer stag showing the key acoustic features. Common roars are complex calls with modulated source (fundamental frequency) and filter (formants). The narrow, evenly spaced frequency bands rising in the first half of the roar and decreasing toward the end represent the fundamental frequency and harmonics. The vocal tract resonances, or formants, are represented by the darker, less regularly spaced energy bands (F_1–F_8) that decrease during calling as the stag lengthens his vocal tract. Harmonics and formants move independently due to source–filter independence.

resulting vocalizations. This quality tends to be associated with cries indicative of pain (e.g., distress calls) or intensely aggressive vocalizations (e.g., harsh roars in red deer).

Variation in the shape of the vocal tract affects the position of the formants relative to one another and contributes to the individual quality of different voices, as studies of deer (e.g., reindeer, fallow deer, and red deer) and some primate species have shown. More important, there is a direct inverse relationship between the length of the vocal tract and the spacing of formants that results in the presence of size-related information in the calls of many mammalian species (e.g., primates, canids, and deer). Both between and within species, the larger the animal, the lower its resonance frequencies, and consequently the closer the spacing of these resonances (formants) in the frequency domain.

Finally, physiological costs associated with the production of sequences of calls are likely to limit the rate at which individuals can call, and this characteristic potentially provides receivers with an honest indication of the physical condition of the caller, as has been shown in red deer.

The ability of individuals to determine each other's identity, physical status, or internal state from vocalizations dramatically increases the level of functionally relevant information that can be exchanged. Whereas in some cases identity may be inferred from the location of the caller or by using visual or olfactory signals, acoustic cues are likely to be of primary importance when individuals range widely. Such cues can provide receivers with instantaneous information on the location and attributes of the caller and may represent the only effective signaling modality in nocturnal or forest-dwelling species. There is a considerable body of evidence indicating that the vocalizations of terrestrial mammals contain information on the identity and physical attributes of the caller.

Contexts in Which Terrestrial Mammals Call

Intraspecific

Social communication Functionally, vocal communication underlies and facilitates social behavior in terrestrial mammals, allowing individuals to stay in contact with others in their own group (including offsprings) even when widely separated, to defend territories, to interact with others of differing rank without constantly getting into costly fights, and to communicate about certain external events. Individual differences in vocalizations, arising from variation in source and filter characteristics, encode individual identity in a range of large mammals, enabling the social recognition that is at the basis of more complex social behavior. Differences in the fundamental frequency contour have been identified as important in broadcasting information on identity in wolves and elephants, whereas individuality in formant frequencies has been demonstrated in fallow deer, roe deer, red deer, elephants, and rhesus macaques.

In order to study vocal recognition, researchers use playback experiments in which vocalizations from specific individuals are recorded and broadcast back to receivers through a loudspeaker. For example, playback experiments have been used to assess whether animals can discriminate the calls of particular close associates (e.g., mother or offspring) from those of other individuals. Such studies have shown that mother–young recognition can be at least partially mediated by voice cues in a range of species (pinnipeds, sheep, deer, bats, and primates) and that vocal recognition occurs between adult individuals in, for example, red deer, red squirrels, African elephants, and several nonhuman primates. In certain species, the number of individuals that listeners can discriminate by voice can be quite large: playback experiments have demonstrated that female elephants are familiar with the calls of approximately 100 others

in the population and social recognition is possible over distances of up to 2.5 km. Memory for the calls of social companions can also persist for very long periods. For example, there is evidence that fur seals and African elephants can still recognize the calls of particular individuals months and sometimes years after these individuals have separated from them.

Calling and territoriality There is abundant evidence that terrestrial mammals use loud calls to advertise ownership of a territory and defend it against rivals in the vicinity. Territorial defense may be carried out by single individuals calling alone (e.g., a male roe deer barking to delineate a breeding territory) or by groups of individuals (often relatives) cooperating to defend a joint territory or range (e.g., howler monkeys). In the latter case, individuals call in chorus, engaging in vocal battles with neighboring groups that involve howling back and forth. Howling in wolves is used in a similar manner, but empirical evidence suggests that calling serves an intergroup spacing function rather than explicitly representing a territorial advertisement. Much of the work on wolves has been conducted using human simulations of wolf howling (rather than playback) to elicit replies from the study animals. Where playbacks were conducted, wolves called back at the loudspeaker and gave lower pitched howls as they got closer. This could suggest that wolves vocalize with a lower pitch as they get closer to ritualistically advertise a stronger motivation to attack, although the lower pitch could also simply be a by-product of softer calling when the wolves perceive their attacker to be close.

Vocalizations used in territorial defense can be quite complex. Gibbons deliver their territorial vocalizations in songs that are highly structured, and these appear to function in intrasexual competition, with the complexity of songs increasing as the interaction proceeds and more complex songs eliciting faster approaches to playback. This is supported by comparative evidence indicating that singing performance in males is also reduced when energy reserves are limited.

Finally, when groups of individuals cooperate to defend shared resources such as territories and young (e.g., lions and many nonhuman primates), the number of individuals per group can often be more important than interindividual differences in determining the outcome of a contest. Assessment of resource-holding potential would be expected to vary accordingly, and playback experiments on lions, chimpanzees, and black howler monkeys have demonstrated that groups assess one another on the basis of relative group size, with vocalizations providing an important cue to this.

Communication within the social group We have focused on examining communication between groups; however, communication also plays a key role in coordinating activities within groups, with the ability to discriminate the calls of group members from those of other individuals obviously providing the crucial basis for this. Among large terrestrial mammals, female lions have been shown to use roaring to recruit members of their own pride if some individuals are absent when simulated intruders' roaring is played to them. In some species, there is evidence that during within-group communication, callers may direct their calls at specific individuals. Thus, male chimpanzees have been observed to pant-hoot more when particular allies are nearby than when they are absent, and in African elephants affiliated females preferentially respond to each other's rumbles and often call in sequence when approached by dominant individuals. Specific vocal signals between dominants and subordinates within stable social groups also play a key role in allowing individuals of different ranks to interact with minimal aggression and in reconciling opponents after fights. Moreover, within the social group, mammals also use vocalizations to signal to one another about certain external events (e.g., predators, food, and social events). These calls can be functionally referential (denoting objects or events external to the caller) as well as emotional (reflecting the internal state or subsequent behavior of the caller).

Sexual Communication

In addition to their social function, mammal vocalizations are also crucial to sexual behavior, influencing the outcome of contests between males as they compete for females, affecting mate attraction and patterns of mate choice, and sometimes directly influencing reproductive physiology.

Loud calls and sexual behavior Many large mammals have evolved loud calls specific to the reproductive season that function to attract individuals of the opposite sex and advertise resource-holding potential to competitors in the vicinity. These calls are often very conspicuous and seem specifically adapted for attracting the attention of a wide audience. The loud reproductive calls of polygynous deer, which typically serve several functions, provide some of the best examples.

Calling rate and male–male contests A long-term study of a free-ranging population of red deer on the Island of Rum identified roaring rate as an assessment cue that determines the outcome of roaring contests

between red deer stags. When one red deer stag is challenging another for the possession of a harem, he typically approaches his opponent and the two start to exchange roars directly. After this roaring conversation, the stags may or may not escalate to either a parallel walk, in which they parade tensely up and down in parallel (usually roaring as they do so), or a full-blown fight. Fights are more often preceded by roaring contests in which the challenger roars more frequently than his rival. Moreover, roaring rates in contests are well correlated with fighting ability. When roars are played back from a loudspeaker to simulate the presence of a rival stag, stags increase their roaring rate to match that of the challenger unless the challenger is roaring so fast that they are apparently unable to do so. This provides experimental evidence that stags only continue with a challenge if they can outroar their opponent. As a result of roaring contests, and the further visual and vocal assessment possible in parallel walks, relatively few challenges end in costly fights. In the closely related fallow deer, harem-holding bucks groan at higher rates in the presence of nearby vocal males, suggesting that the signal conveyed by groaning rates in this species is primarily a threat aimed at rival males. Bucks achieve their highest call rates immediately after copulating, thereby transmitting information to nearby males on their condition and motivational state. In line with the hypothesis that groaning rate acts as a signal between males, immature males decrease their rates of groaning in response to playbacks of groans from mature males, whereas mature males increase their groaning rates in this situation. Overall, these results suggest that fallow deer bucks vary their groaning rate in relation to the presence and quality of competitors, including the indirect costs of inviting contests with surrounding and potentially stronger males.

Effects of calling on females In both fallow and red deer, there is evidence that rutting calls do not function solely to mediate competition between males. In playbacks conducted under controlled conditions on deer farms, red deer females exposed to male roars during the pre-rut period conceived earlier than females who were not, providing evidence that calling can advance ovulation. Moreover, red deer roaring appears to influence mate choice decisions. Hinds are more likely to look at and move toward a speaker simulating a stag roaring at a higher rate, even at points in the roaring sequence when both are delivering the same number of roars. They even appear to prefer callers that initiate bouts of roaring. High levels of roaring activity may thus confer advantages in intersexual choice as well as intrasexual competition.

Red deer females also appear to attend to the particular roaring characteristics associated with the stag whose roars they hear most often. Hinds are able to discriminate between the roars of their current harem holder and those of other neighboring stags, in accordance with previous findings that red deer roars contain sufficient acoustic cues to identify the caller. Estrous females could benefit by choosing to mate with males that are most familiar to them (i.e., those that are able to spend more time and effort in retaining them within a harem and in roaring at them). It is thought that over the course of the breeding period, red deer hinds progressively become familiar with the vocalizations of the stag in whose harem they spend the most time. The degree of familiarity that a given female has with a particular stag's roars should result from the cumulative extent of her exposure to his roars. A combination of factors is likely to affect this, including the duration of exposure, the rate of roaring during exposure, and the loudness of the roars (harem holders' roars will be received at a higher amplitude simply because of proximity). Studies of long-term groaning rates in fallow deer also indicate that the amount of calling is correlated with reproductive success, therefore providing support for the familiarization hypothesis.

Alternative routes to honest signaling in mammal calls As discussed previously, males and females use rate of calling as an indicator of the quality or resource-holding potential of particular males. In red deer, roaring rate is believed to be an honest indicator of fighting ability (and an evolutionarily stable assessment cue) because of the direct and indirect energetic costs associated with maintaining high roaring rates. Beyond the usual costs associated with phonation, roaring involves strenuous muscular activity related to tilting the head back and retracting the larynx toward the sternum each time a roar is given. However, although roaring rate provides a reasonable indication of a stag's condition, it does not give direct information about a stag's body size – a crucial parameter in predicting the outcome of fights and a good indicator of reproductive success. Studies of the detailed acoustic structure of roars indicate that formant frequencies given when the vocal tract is fully extended are inversely related to body weight and that stags attend to these acoustic features when assessing opponents. Research on the responses of hinds to playbacks of roaring also shows that females prefer roars in which lower formants indicate larger individuals. As discussed at the beginning of the chapter, formant frequencies are inversely related to the length of the vocal tract, which in red deer is constrained by the length of the neck. This provides an example of how physical

constraints can generate honesty in mammal calls independently of any costs linked to their production.

Loud mating calls in other mammals Males from polygynous pinniped species characterized by extreme sexual dimorphism also have loud sexual advertisement calls (e.g., northern elephant seals and common seals) and calling interactions between males on land or under water may be attended to by rival males, potential mates, or both. It remains to be seen whether receivers alter their subsequent behavior on the basis of which male dominates in a calling interaction. In elephant seals, females are unusual in also producing loud calls during matings (copulation calls) that attract other males in the vicinity and indirectly result in the caller being mated by the most dominant male.

Interspecific Communication

We previously discussed the use of terrestrial mammal calls that are directed primarily at other members of the same species. There are also cases of terrestrial mammals producing calls that function in the context of interspecific interactions. One such example is that of pursuit-deterrence calls, typically given by ungulates. A study of roe deer revealed that barks, previously identified as 'alarm calls,' in fact function to elucidate the cause of disturbance. In this communication system, calls inform any predator that might be present that it has been detected. The likelihood of barking in response to a predator-like disturbance is independent of the presence of (related or unrelated) conspecifics in close vicinity, demonstrating that it is not an alarm call. In other cases, animals appear to have developed the ability to listen to the calls of predators and extract sophisticated information that enables them to adjust their defensive behavior according to the specific threat. Work on harbor seals has shown that this species has learned to actively avoid dangerous 'mammal–eating' killer whales in comparison with the local fish-eating ones that pose no threat, and that this discrimination is based on acoustic recognition.

Finally, there is a growing body of research on the impact of domestication and artificial selection on vocal communication between companion animals (particularly domestic cats and dogs) and their human keepers. Within this literature, there is plentiful evidence that vocal signals may be used by pets to specifically alter the behavior of their owners.

See also: Agonistic and Affiliative Signals: Resolutions of Conflict; Animal Communication: Honesty and Deception; Communication in Frogs and Toads; Facial Expression in Primate Communication; Kinship Signals in Animals; Referentiality and Concepts in Animal Cognition; Sexual Selection and the Evolution of Animal Signals; Signal Transmission in Natural Environments; Visual Signaling in Animals; Vocal Communication in Birds.

Further Reading

Clutton-Brock TH and Albon SD (1979) The roaring of red deer and the evolution of honest advertisement. *Behaviour* 69: 145–169.

Cowlishaw G (1996) Sexual selection and information content in gibbon song bouts. *Ethology* 102: 272–284.

Cox CR and Le Boeuf BJ (1977) Female incitation of male competition: A mechanism in sexual selection. *American Naturalist* 111: 317–335.

Deecke VB, Slater PJB, and Ford JKB (2002) Selective habituation shapes acoustic predator recognition in harbour seals. *Nature* 420: 171–173.

Fitch WT (2000) The evolution of speech: A comparative review. *Trends in Cognitive Sciences* 4: 258–267.

Harrington FH (1989) Aggressive howling in wolves. *Animal Behaviour* 35: 7–12.

Insley SJ (2000) Long-term vocal recognition in the northern fur seal. *Nature* 406: 404–405.

Kitchen DM (2004) Alpha male black howler monkey responses to loud calls: Effect of numeric odds, male companion behaviour and reproductive investment. *Animal Behaviour* 67: 125–139.

McComb K, Moss C, Durant S, Baker L, and Sayialel S (2001) Matriarchs act as repositories of social knowledge in African elephants. *Science* 292: 491–494.

McComb K, Packer C, and Pusey A (1994) Roaring and numerical assessment in contests between groups of female lions *Panthera leo*. *Animal Behaviour* 47: 379–387.

McComb K, Reby D, Baker L, Moss C, and Sayialel S (2003) Long-distance communication of cues to social identity in African elephants. *Animal Behaviour* 65: 317–329.

McElligott AG and Hayden TJ (1999) Context-related vocalization rates of fallow bucks, *Dama dama*. *Animal Behaviour* 58: 1095–1104.

Reby D, Cargnelutti B, and Hewison AJM (1999) Contexts and possible functions of barking in roe deer. *Animal Behaviour* 57: 1121–1128.

Reby D and McComb K (2003) Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of red deer stags. *Animal Behaviour* 65: 519–530.

Reby D and McComb K (2003) Vocal communication and reproduction in deer. *Advances in the Study of Behaviour* 33: 231–264.

Reby D, McComb K, Cargnelutti B, Darwin CJ, Fitch WT, and Clutton-Brock TH (2005) Red deer stags use formants as assessment cues during intra-sexual agonistic interactions. *Proceedings of the Royal Society of London, Series B* 272: 941–947.

Rendall D, Rodman PS, and Emond RE (1996) Vocal recognition of individuals and kin in free-ranging rhesus monkeys. *Animal Behaviour* 51: 1007–1015.

Soltis J, Leong K, and Savage A (2005) African elephant vocal communication I: Antiphonal calling behaviour among affiliated females. *Animal Behaviour* 70: 579–587.

Communication in the Honeybee

A Michelsen, University of Southern Denmark, Odense, Denmark

© 2009 Elsevier Ltd. All rights reserved.

The Life of Honeybees

Honeybees are social insects and live in colonies of up to about 70 000 individuals. There are three types of bees: workers, drones, and queens. The social cooperation in insect colonies, which is unparalleled in the animal kingdom, ensures that a large number of jobs are done. The jobs include activities as diverse as collecting food, defending the colony, building the nest, feeding larvae, cleaning cells, keeping a constant temperature in the nest, and so on. The probability that an individual does a particular job depends on its age and genetic disposition (which varies in honeybee colonies because the queen has mated with many drones), but each individual decides which job to do and for how long. Obviously, much communication among the individuals is needed in order to coordinate the work that needs to be done.

The in-nest jobs are performed by young individuals, whereas the outside tasks are performed by older bees. Visual communication is not possible inside the dark nest, where the bees have to use chemical and mechanical signals. Vision plays a role in outdoor orientation, but the spatial resolution of the eyes limits the use of visual signals. For example, the swarming bees are kept together by odor from the queen and not by vision when a swarm is flying to a new nest.

Chemical Communication

Modern instrumentation has made it possible to sample and identify small amounts of chemical substances, as well as to measure the responses of olfactory sense organs. Nevertheless, chemical communication in honeybees remains largely unexplored. About 20 pheromones have been identified, but a similar number probably remain to be discovered. The effects of the pheromones on the receiving bees vary from strong physiological responses to minor modifications of behavior.

Most pheromones are volatile, but the main queen pheromones are probably transported from bee to bee during antennal contacts. Examples of volatile pheromones are the orientation odors from the Nasonov gland (at the tip of the abdomen), which workers disperse by fanning when they attract bees to the entrance of a new nest, during swarming, and when marking a source of water (which itself is relatively odorless). Another example is the alarm odors released from the sting, which remains in the skin when mammals are stung. These substances alert the workers to danger, release the attack reaction, and guide the workers to the previously stung spot (when stung, it is therefore wise for mammals, including humans, to remove the sting as soon as possible).

The pheromones produced by the mandibular glands of queens inhibit ovary development in workers (who are females), inhibit the rearing of new queens (and thus swarming), and also seem to calm the workers in the colony and improve their cooperation. Some of these effects are also caused by the footprint pheromone from the queen's tarsal glands. A release of inhibition occurs after the death of the queen, and the workers begin to treat some young larvae as queen larvae. After some weeks without the queen, some of the workers begin to lay eggs (which develop into drones because the workers have not mated). These 'drone mothers' may produce so much queen pheromone that the doomed colony returns to almost normal behavior before the last workers die.

In passing, it should be mentioned that the sense of olfaction is located in the pore plates on the antennae. The antennae are mobile and often far apart, and bees are therefore able to detect the direction to sources of odor by comparing the responses from the two antennae.

Communication by Tactile and Vibration Signals

By observing the behavior of bees inside the hive in dark-red light, which is invisible to the bees, one finds that most bees pass each other without touching. Somehow the bees must be able to sense the presence of other bees and keep a short distance from them. Some cases of touching are known to involve communication. A very common activity is mouth-to-mouth exchange of food (trophallaxis), in which the bees touch each other with their antennae and tongues (Figure 1). The extent of food exchange in a colony can be studied by feeding a few bees radioactively marked food. Within a couple of hours the radioactivity spreads to a large fraction of the bees in the colony. The purpose of trophallaxis is not simple, because food does not always flow from bees with plenty of food to bees with very little food. The opposite also occurs. In the dorsoventral abdominal vibration (DVAV) signal (also known as the shaking signal), vibrations are often combined with touching: a worker vibrates its abdomen up and down,

Figure 1 Two bees (upper left) begging food from another bee (center). Magnification 1.6×. Photo copyright Axel Michelsen.

frequently while grasping another worker or a queen. These signals are very common, and their occurrence is temporally correlated with swarming and foraging activities.

During swarming, the colony is divided into two, and repeated swarming may result in several colonies. The process begins with the construction of queen egg cups, in which the queen deposits eggs. The queen receives an increasing number of DVAVs until the cups are sealed and then a diminishing number until she leaves the nest with about half the bees. The swarm settles not far from the nest, and the shaking workers now perform DVAVs on both queen and workers while other workers engage in house hunting. When the house hunting has been finished, the number of DVAVs received by the queen again diminishes before the swarm takes off for the new nest. The message of the DVAVs to the queen probably is "Reduce your activity." Despite a very convincing correlation between DVAVs and foraging, the message of the DVAVs to the workers is not so obvious. A peak of DVAV activity in the morning may mean "Wake up and go to work," whereas at other times, the message may be "Change your course and do something else." More research is obviously needed.

Several other signals involve vigorous body motion and in some cases also vibrations. The exit of a swarm (both from the nest and from the swarm's temporary position outdoors) is signaled by workers performing buzzing runs. The workers run very fast through the colony or swarm and increase the activity of the other workers, and just before takeoff, all the bees are on the move. During shaking dances (not to be confused with shaking signals), the workers shake their bodies from side to side and invite nearby workers to groom them. On their return to the nest, the nectar foragers transfer the nectar to food-storing bees. If the foragers are not able to find a food-storing bee, the foragers may perform tremble dances, which seem to increase the recruitment of the food-storing bees. Still other dances (jostling, spasmodic) have been described. All these dances look different to human observers, but it is not obvious how the bees distinguish them in the darkness of the nest.

The bees performing tremble dances also emit substrate-borne vibration signals, that is, body vibrations that cause bending wave vibrations in the wax comb. This requires close contact between the vibrating body and the substrate, for example, between the vibrating thorax of the bee and the comb. The most powerful bending waves are made by queens during the preparation for swarming, and these signals may travel up to 20 cm before decaying below the threshold of the receiving bees' perception. Tooting signals last some seconds and are made by the old queen during the days before she leaves with a swarm. During this time, the young queens are still confined to their queen cup cells by workers, who resist the young queens' attempts to leave the cell. The confined queens respond to the toots with quacking vibrations, which are emitted as a long series of short (about 0.1–0.2 s) vibrations separated by pauses of a similar duration. Although these signals travel as bending waves and are perceived by the bees as substrate vibrations, they can be heard by humans as airborne sound because the vibrating comb acts as a loudspeaker. Tooting may be heard meters away from a hive and tell the beekeeper that the colony is about to swarm.

The concert of tooting and quacking goes on for days. Because of the limited range of the signals, the queen is constantly walking from comb to comb through the nest. About every 10 cm, she stops and toots. It is obvious that the confined queens respond to the toots by quacking, but the primary receivers of the toots may be the workers guarding the confined queens. They perform DVAV signals on the queen cells. A large rate of DVAVs on the queen cell is correlated with a delay of emergence of the confined queen and also with the later behavior of the confined queen when she leaves the queen cell. She may either begin to toot and later leave with a new swarm (many DVAVs), or she may kill the other (still confined) queens and not swarm (few DVAVs).

Dance Language

The zoologist Karl von Frisch (Nobel laureate, 1973) is mainly known for discovering that honeybees can recruit other bees to visit a food source by performing a dance in which information on the distance and direction from the hive to the food is codified. The dance language communication about remote events is an exception to the general rule that animals tend to communicate about immediate events connected with

the actor and its surrounds. Furthermore, in the dance language, a symbolic code is used to transmit an impressive amount of information.

The dances are normally performed on the vertical wax combs in the hive. In the waggle dance, the dancer moves in a straight line (the waggle run) and circles back, alternating between a left and a right return path so that the entire dance path takes on a figure-eight shape. During the waggle run, the dancer wags her body from side to side 13–15 times per second and emits 280 Hz sounds by vibrating her wings. The angle between the direction of the waggle run and the vertical indicates the angle which the recruits should later maintain between their flight path to the food and the direction to the projection of the sun on the horizon (known as the sun's azimuth). The distance to the food is indicated by the duration of the waggle run.

A few bees (called follower bees) keep close contact with the dancer during most of the dance, and these bees may be recruited to search for the food at the position indicated in the dance. The dancer carries the odor of the flowers visited, and this information makes it easier for the followers to find the food. This is understandable since the dance does not contain any information about the shape or color of the flowers. The dance language is also used by the bees for indicating the position of resin, water, and suitable new nests (during swarming). At short distances, the presence of the food is signaled by round dances. These dances have a circular dance path but contain the same elements as the waggle dances. The average direction of the waggle run points to the direction to food, but the correlation is weak. The information is therefore not available to the follower bees in limited number of dances followed.

Karl von Frisch and his students made several other discoveries when investigating the dance language. For example, their experiments showed that bees have an internal clock, which allows the dancer and followers to compensate for the rapid changes in the sun's azimuth as the sun is moving over the sky. It was also shown that the vision of honeybees is sensitive to the plane of polarization of light. The polarization of the sky depends on the time of the day and the position of the sun. The bees can therefore use this sensory capacity for determining the position of the sun on cloudy days, provided that they can see a small patch of blue sky. It has also been found that the preferred orientation for polarized light in about 150 single eyes in the bee's composite eye forms a map of the average polarization of the sky.

For many years after the discovery of the dance language, most biologists intuitively believed that it made the bees' collection of food more efficient. Very recently, this question has been investigated in experiments in which the effects of dances with and without indication of direction were compared. One can cause the dancers to become disoriented by forcing the bees to dance on a horizontal surface in diffuse light. The results show that the direction information increases food collection when food sources are scarce, whereas it does not matter when there is plenty of food.

Several parameters of the dance vary with the distance and direction of the food, so this information is redundant. Therefore, one cannot be sure which part of the dance is perceived by the follower bees. This problem has been studied by means of manipulated dances performed by a mechanical robot deployed on the dance floor of an observation hive (Figure 2). The robot was made of brass, and the wings were simulated by a single piece of razor blade, which was vibrated by an electromagnet. A step motor rotated the model and caused it to waggle during the waggle runs. Other motors moved the model in a figure-eight path. All motors were controlled by a computer, which made it possible for the investigators to create dances that differed from the normal waggle dance.

The experiments showed that although the follower bees can obtain the specific information from the movements of the robot, live dancers are more efficient (which was not surprising!). With the computer, the investigators could manipulate dances to provide the bees with conflicting information about distance or direction. For example, the position of the waggle run in the figure eight could be shifted so that opposite directions were indicated by the waggle run and the '8.' The bees now followed the instructions given by the waggle run and ignored the information given by the dance figure. The waggle run appears

Figure 2 A mechanical robot model of a dancing honeybee during its performance of a waggle dance among the bees on the dance floor of an opened observation hive. A number of bees are facing the robot and running along with it, and some of these bees later appear at the location indicated by the robot's dance. Magnification 3×. Photo copyright Axel Michelsen.

to be the 'master parameter' for the transfer of information about both direction and distance.

The follower bees thus obtain information about distance and direction to the food mainly by 'observing' the waggle run. The follower bees can obtain the information in total darkness, so the observation does not have to include vision. Karl von Frisch suggested that the dance sounds might travel as vibrations from the dancer, through the wax comb, to the follower bees, and that the followers might touch the dancer. Some evidence supports these suggestions, but the vibrations seem to be to attract followers rather than to transmit specific information about the direction and distance to food. Studies using high-speed video have shown that follower bees have a chance of touching the dancer with one or both antennae during each waggle. The most obvious information to be obtained from touching is whether the follower bee is facing the dancer's abdomen laterally or from behind, but this information is not sufficient for perceiving the direction of the waggle run.

The problem of how the bees determine the direction of the waggle run is complicated by the waggle motion: the body of the dancer moves from side to side like a pendulum (fixed 1 mm in front of the head) 13–15 times per second. The direction of the waggle run is defined from the position of the head and the average position of the tip of the abdomen. It is difficult to imagine how this average position can be determined when the body makes such rapid lateral movements. One hypothesis focused on the finding that the complicated motion of the dancer during the waggle run gives rise to two three-dimensional fields of intense airflows, oscillating with the frequencies of waggle (13–15 Hz) and wing vibration (280 Hz), respectively. This discovery suggested a third strategy: the follower bees might keep a small distance from the dancer and estimate her dance path by means of receptor organs sensitive to such air flows. However, studies of air flows around models showed that the patterns of oscillatory air flows were much too complicated, especially lateral to the dancer, to provide the follower bees with reliable information about the direction of the waggle run.

Recent studies, in which the positions of the follower bees around the dancers were correlated with their later success in finding the food, have shown that the specific information about the location of the food is available only to the follower bees located behind the dancer, within the angle of the waggle motion. Measurements with tiny anemometers (instruments for detecting air flows) have shown that the passing of the tip of the abdomen of the waggle dancer is associated with a distinct maximum of air flow. The temporal pattern of the maxima of air flows experienced by the follower bees behind the dancer depends on the position of the follower bee relative to the direction of the waggle run. One may therefore speculate that the temporal pattern is used as a cue for calculating the direction to the food. Two mechanisms contribute to the maximum of air flow associated with the passing of the tip of the dancer's abdomen. The oscillatory waggling air flow makes a minor contribution. Most of the air flow hitting follower bees behind the dancer is generated by the vibrating wings and travels away from the dancer as a jet air flow (Figure 3). In contrast to the oscillatory air flows, in which the masses of air flow to and fro, the air in the jet flows in only one direction (away from the dancer).

The honeybee waggle dances have, until recently, been considered a very stereotyped behavior, in which the main variation was the duration and direction of the waggle run, although the duration of dancing and the tendency to vibrate the wings were known to depend on the motivation of the dancer. It is now known that the duration of the return run (between consecutive waggle runs) varies with the profitability of the food source. Furthermore, the dancers may generate two kinds of jet air flows, spatially narrow jets and broad jets, and both may be switched on and off by an adjustment of the position of the wings. A similar flexibility is found in the dance following. The patterns of following dances are related to the foraging experiences of the follower bees and to the presence, on the dance floor, of dancers announcing other targets. Finally, the patterns in which the followers move

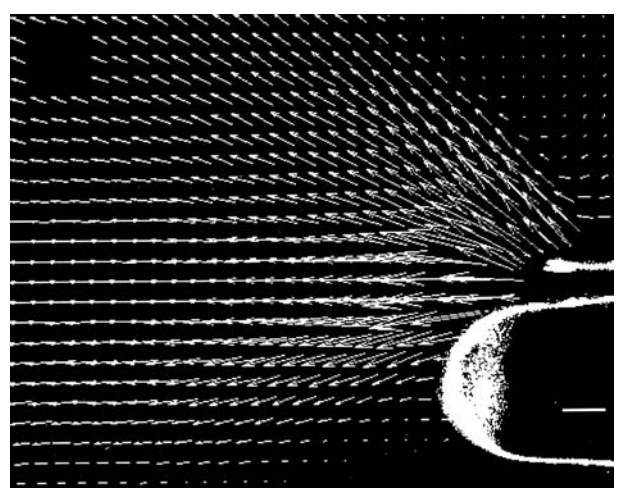

Figure 3 A jet air flow caused by wing vibration in a model. The tip of the abdomen and of a metal wing were photographed when the wing was moving up. The air flow is visualized by means of a laser technique, and the air velocities are indicated as vectors (arrows). The largest vectors correspond to approximately 300 mm s^{-1}. Such jet air flows can also be measured (by means of hot wire probes) behind live dancers. Scale bar = 100 mm s^{-1}. Photo copyright Axel Michelsen.

with the dancer also show much flexibility. It seems fair to conclude that the dance language will keep investigators busy for many years to come.

See also: Conditioning: Simple Neural Circuits in the Honey Bee; Learning and Memory in Invertebrates: Honey Bee; Olfaction in Invertebrates: Honeybee; Pheromones and other Chemical Communication in Animals; Seismic and Vibrational Signals in Animals.

Further Reading

Biesmeijer JC and Seeley TD (2005) The use of waggle dance information by honey bees throughout their foraging careers. *Behavioral Ecology and Sociobiology* 59: 133–142.

Frisch K von (1967) *The Dance Language and Orientation of Bees*. Cambridge, MA: Harvard University Press. [Paperback edition published 1993 with a foreword by Tom Seeley, in which the main progress from 1967 to1993 is reviewed; Cambridge, MA: Harvard University Press].

Judd TM (1995) The waggle dance of the honey bee: Which bees following a dancer successfully acquire the information? *Journal of Insect Behavior* 8: 343–354.

Michelsen A (1999) The dance language of honeybees: Recent findings and problems. In: Hauser MD and Konishi M (eds.) *The Design of Animal Communication*, pp. 111–131. Cambridge, MA: MIT Press.

Michelsen A (2003) Signals and flexibility in the dance communication of honeybees. *Journal of Comparative Physiology A* 189: 165–174.

Michelsen A, Andersen BB, Storm J, Kirchner WH, and Lindauer M (1992) How honeybees perceive communication dances, studied by means of a mechanical model. *Behavioral Ecology and Sociobiology* 30: 143–150.

Michelsen A, Kirchner WH, and Lindauer M (1986) Sound and vibrational signals in the dance language of the honeybee, *Apis mellifera*. *Behavioral Ecology and Sociobiology* 18: 207–218.

Michelsen A, Towne WF, Kirchner WH, and Kryger P (1987) The acoustic near field of a dancing honeybee. *Journal of Comparative Physiology A* 161: 633–643.

Rohrseitz K and Tautz J (1999) Honey bee dance communication: Waggle run direction coded in antennal contacts? *Journal of Comparative Physiology A* 184: 463–470.

Seeley TD (1995) *The Wisdom of the Hive: The Social Physiology of Honey Bee Colonies*. Cambridge, MA: Harvard University Press.

Seeley TD, Mikheyev AS, and Pagano GJ (2000) Dancing bees tune both duration and rate of waggle-run production in relation to nectar-source profitability. *Journal of Comparative Physiology A* 186: 813–819.

Seeley TD and Visscher PK (1988) Assessing the benefits of cooperation in honeybee foraging: Search costs, forage quality, and competitive ability. *Behavioral Ecology and Sociobiology* 22: 229–237.

Sherman G and Visscher PK (2002) Honeybee colonies achieve fitness through dancing. *Nature* 419: 920–922.

Tautz J (1996) Honeybee waggle dance: Recruitment success depends on the dance floor. *Journal of Experimental Biology* 199: 1375–1381.

Tautz J, Rohrseitz K, and Sandeman DC (1996) One-strided waggle dance in bees. *Nature* 382: 32.

Towne WF and Gould JL (1988) The spatial precision of the honey bees' dance communication. *Journal of Insect Behavior* 1: 129–155.

Winston ML (1987) *The Biology of the Honey Bee*. Cambridge, MA: Harvard University Press.

Communication Networks and Eavesdropping in Animals

P K McGregor, Duchy College, Cornwall, UK

© 2009 Elsevier Ltd. All rights reserved.

Communication Networks

A communication network is a group of several animals within signaling and receiving range of each other. Such networks are likely to be ubiquitous because most signals travel farther than the average spacing between individuals. Indeed, a network will exist unless a signal can never be received by more than one receiver or unless a receiver can never receive more than one signal simultaneously. Communication networks can be based on signals of any modality and over any distance, from the ocean-spanning songs of the great whales to the crowded hives of social insects. Therefore, the social environment in which communication has evolved and in which signals are now produced and received is best thought of as a network.

Given such arguments, it is surprising that a network perspective has not been more widely adopted in studies of animal communication. Communication was, and in many instances still is, treated as occurring between two individuals – the signaler–receiver dyad. Notable exceptions are studies of chorusing animals, in particular insects and anuran amphibians, which have long adopted a network perspective and demonstrated the value of doing so.

While a dyad is the simplest relationship possible between the three basic components that characterize communication (the signaler, the signal, and the receiver), there are two important reasons to adopt a network, rather than a dyadic, perspective. First, there are communication behaviors that occur in a network but cannot occur in a dyad. Second, a network perspective better explains some striking signaling behaviors that are not adequately explained by a dyadic view.

Communication Behaviors Specific to Networks

A range of communication behaviors involve more than a signaler–receiver dyad. A good example is eavesdropping, and for this reason it is considered at length later in this article. Eavesdropping involves a least three individuals (signaler, intended or target receiver, and eavesdropper). The term 'bystander' is often used for individuals that are present during an interaction (including signaling interactions) but which do not take part in the interaction – therefore, eavesdroppers are an example of bystanders. The presence of bystanders can affect the nature of signaling interactions, and the effects of such an audience are another example of an aspect of communication behavior that is revealed by adopting a network perspective (and is therefore discussed more fully below).

Victory displays are signals produced by the winner (but not the loser) after an aggressive interaction and are a form of signaling that seems likely to function in a network context. Such displays were first described in waterfowl but also occur in other bird groups (parrots, penguins, divers, shrike, and a songbird), other vertebrates (including an amphibian, a reptile, and mammals), and some invertebrates (such as spiders and crickets). There is little evidence for or against a network function of victory displays, but even if such signals are directed within the winner–loser dyad, their conspicuous nature will ensure that information can travel considerably more widely.

Signaling Behavior Best Explained by Networks

There are two signaling behaviors that are commonly cited as being best explained by a network perspective: signal timing in choruses and the high level of signals used at close range. The widely broadcast nature of many signals also raises the possibility that some aspects of signaling behavior are associated with attempts to direct signals in a network.

Signal synchrony in choruses Choruses are collections of signaling individuals, from groups of several vocalizing mammals (e.g., howling wolves), through thousands of frogs in breeding aggregations, to many tens of thousands of flashing fireflies. Often there is a high degree of signal synchrony in the chorus, but such synchrony is not universal. In anuran amphibians, patterns such as call alternation, following, and overlapping have been described, as having a degree of flexibility in the pattern. Similar variation may exist in the visual waving signals of fiddler crabs. Regardless of the particular pattern of signaling in a chorus, it seems likely to result from the effect of the signals of neighboring conspecifics on the timing of signal production. The number of such influential neighbors varies and is usually only a part of the chorus. However, usually more than one neighbor influences signal timing; therefore, chorus synchrony cannot be explained adequately by considering communication as a dyad.

While several advantages of synchrony have been suggested, such as enhancing group detection or confusing predators, recent studies have focused on competition for receiver attention as the driving force behind the exact timing of signal production. Synchrony could result from individuals each trying

to ensure that their signal is the first to be detected, as females of many species prefer the leading call in a group of calls.

High level of close-range signals From a dyadic perspective, a puzzling aspect of signaling in close-range aggressive encounters is the high amplitude or level employed – human antagonists nose to nose, yet shouting at each other, for example. Such high signal levels seem unnecessary when the individuals are so close to one another. Considered from a network perspective, such high levels are more explicable – there may be more-distant intended receivers in addition to the opponent.

Most studies of parent–offspring communication have adopted a dyadic perspective. In order to do so, these studies treat groups of young being provisioned by adults (e.g., a brood of nestlings) as a unit. However, recognizing that such a system is in fact a network allows explicit consideration of features of begging such as the high amplitude of calls and the role of competition between nest mates.

Directing signals in a network The observation that many signals are widely broadcast leads to the idea of communication networks (see above). However, it is easy to imagine advantages to directing signals to particular individuals in a network in both aggressive and mate attraction contexts. Several features of signaling could achieve such targeting of one particular receiver among many. One suggestion is that songbirds match song types (also termed 'matched countersinging'), that is, they select songs from their repertoire that are most similar to those being sung by the individual to which they seek to direct signals. In many taxonomic groups, the relative timing of signals could be used in the same way, with close temporal association indicating signals directed between particular individuals.

Eavesdropping

Eavesdropping is best defined as the use of information in signals by individuals other than the primary target. It is therefore a receiving behavior that is found only in networks as it requires a minimum of three individuals: a signaler, a target receiver, and the eavesdropper. This definition encompasses most of the ways the term 'eavesdropping' has been used in the animal communication literature without obscuring interesting aspects of the behavior. Eavesdropping is also a word in everyday use in the English language, and dictionary definitions emphasize that eavesdropping refers to acoustic signals and that the behavior is carried out secretly. These two features unnecessarily restrict the use of eavesdropping as a technical term in animal communication. However, whether information gathered without the source's knowledge has particular value or salience (as it seems to have in humans) is an interesting subsidiary question.

Networks and Eavesdropping

The value of the communication network perspective does not depend solely on a demonstration of eavesdropping. While eavesdropping is a communication behavior that can occur only in a network context, there are other communication behaviors and aspects of communication that are best explained from a network perspective. The current degree of prominence of eavesdropping in the literature simply reflects the fact that it was one of the first communication network behaviors to be clearly identified as such and for which experimental paradigms were developed. There is considerable scope to develop experimental investigations of other network behaviors as well as key influences of the network environment on communication behavior.

Two Categories of Eavesdropping

There are two ways in which individuals who are not the primary target of signals can use information in those signals, that is, there are two categories of eavesdropper: interceptive and social. Interceptive eavesdroppers intercept signals intended (in an evolutionary sense) for another individual, whereas social eavesdroppers gather information by attending to the signaling interactions of other individuals. The reason for making this distinction is that the nature of information transfer differs between the two, with resulting differences in the selection pressures they impose.

The two categories of eavesdropping often differ in several other aspects. Interceptive eavesdroppers are usually different species from the signaler; there is commonly a cost of eavesdropping to the signaler; and most often, absolute information is gathered from broadcast signals. Social eavesdroppers are usually the same species as the signaler; the payoff to the signaler can be zero, negative, or positive; and both absolute and relative information can be gathered from directed signaling interactions. While such differences may be common, they are not universal, and they are not necessary or sufficient conditions for distinguishing between interceptive and social eavesdropping. Therefore, such differences do not constitute a sound basis for a working definition of eavesdropping in general or for the two categories of eavesdropping, despite the fact that they have been so used. The definitions used in this article are considerably less likely to result in exceptions or to

generate confusion and will not obscure important characteristics of this behavior.

Interceptive Eavesdropping

Some of the clearest examples of interceptive eavesdropping involve widely broadcast signals, such as the signals designed (in an evolutionary sense) to attract potential mates. Predatory interceptive eavesdroppers can locate potential prey from the prey's mate-attraction signals; an example is the bats that prey on calling male frogs. Parasites can locate potential hosts in a similar way, with parasitic flies attracted to the chemical or acoustic mate attraction signals of their hosts.

An elaboration of this basic pattern is shown by the geckos that prey on female crickets attracted by the calls of male crickets rather than preying directly on the calling males. The effect of the presence of predatory interceptive eavesdroppers is shown by the stream by stream variation in male guppy *Poecilia reticulata* courtship coloration, variation that is explained by differences in the predator guild between streams and the differing abilities of these predators to detect, and therefore intercept, guppy courtship colors.

It is also possible for prey to intercept predators' signals and thereby become aware of the presence of predators. An example is the seals that can discriminate between the calls of transient, mammal-eating killer whale *Orcinus orca* pods and those of resident, fish-eating pods.

Examples involving predator–prey and host–parasite relationships illustrate the costs interceptive eavesdroppers impose on signalers. However, there are examples in which the presence of interceptive eavesdroppers has no cost or possibly even a benefit. One instance is the tendency of individuals to respond to the alarm calls warning of an approaching predator, even if the call is given by a different species. For example, primates respond to the alarm calls of birds and ground foragers in mixed-species bird flocks in forests when these birds are alerted by 'sentinel' species in the canopy. The possible benefit of interceptive eavesdroppers reacting to the alarm call is greater confusion in which to escape. Such examples demonstrate that interceptive eavesdropping should not be defined on the basis of costs and benefits, although clearly costs and benefits are crucial to understanding evolution of the behavior.

Within a species, interceptive eavesdropping has been noted in the contexts of mate attraction and detecting the presence of predators. Signals designed to attract mates from afar have to travel widely and therefore are particularly likely to be intercepted by potential competitors for mates. For example, male fiddler crabs *Uca tangeri* wave in a distinctive fashion (high-intensity waving) when they have detected an approaching female. Experimental studies have shown that other males respond to high-intensity waving by males near them by increasing their own high-intensity waving even when they cannot see the female – such males are interceptive eavesdroppers.

We saw above that interceptive eavesdropping on signals designed to warn others of the presence of predators occurs between species. The same behavior occurs within a species. Convict cichlids *Cichlasoma nigrofasciatum* guarding their young produce a visual signal – fin flicking – to alert young to the approach of a potential predator. Parents were shown to fin-flick in response to fin-flicking models in the absence of predators. Such interceptive eavesdropping is likely to result in their young being alerted to a predator more quickly than if they relied solely on their own vigilance to detect predators. The primary targets of the signal are likely to be relatives of the signaler (i.e., the signaler's young); however, there is no obvious cost to the signaler of interceptive eavesdropping by unrelated adults.

The term 'eavesdropping' has been used in the context of autocommunication, that is, when an animal broadcasts a signal that it also receives in order to investigate the environment. Well-known examples of autocommunication are the biosonar of cetaceans, the bioradar of bats, and the use of electric impulses by electric fish. The use of information in the autocommunication of others could be considered a variant of interceptive eavesdropping: the eavesdropper intercepts a message whose intended receiver is the same individual as the signaler. For example, dolphins can identify objects from the echolocation clicks directed at the object by another dolphin. Similarly, bats can increase their foraging efficiency by paying attention to the echolocation calls of other foraging bats and using information from conspecifics to identify areas of high prey density – and in some instances, to identify suitable individuals of other bat species from which to steal captured prey.

Social Eavesdropping

The defining feature of social eavesdropping is that information is gathered from signaling interactions between others. Therefore, attempts to demonstrate social eavesdropping experimentally have concentrated on ensuring that information likely to elicit a difference in response is contained only in the interaction. A realistic simulation of the presence of a signaler can be achieved by playback of signals alone in acoustic and electric signaling modalities. Playback

can then present subjects with an eavesdropping opportunity – a simulated signaling interaction – in which potentially confounding features (e.g., the amount of song) can be held constant. Achieving a comparable degree of control with visual signals is theoretically possible with video playback but requires a degree of coherence between the perceptual abilities of the subjects and the signal output of the playback system that is yet to be achieved in nonhuman animals. There is now experimental evidence for social eavesdropping in two taxa employing different signal modalities: territorial birds interacting acoustically and fish interacting with visual signals.

Acoustic Interactions in Songbirds

Experimental investigations of social eavesdropping in songbirds have used playback to simulate the presence of rival territorial males. Often playback is interactive in the sense that details of the playback stimulus are determined by the response of the rival male song by song, achieving a realistic interaction.

The earliest experiments involved interactions between playback and a territorial neighbor of the subject. Neighbors were presented either with a more aggressive simulated intruder (signaled by overlapping songs of increasing length) or a less aggressive intruder (alternating songs of decreasing length). When subsequently presented with noninteractive playback of the simulated intruder, male great tits *Parus major* responded to previously aggressive intruders by keeping their distance and overlapping song, while less aggressive intruders were approached quickly. A similar experiment investigated the extra-pair behavior of females in relation to their mate's ability to respond to a simulated intruder. Female great tits were more likely to visit neighboring territories if their mates had responded less vigorously to the simulated intruder than a neighbor had, and females tended to visit the territory of the neighbor who had performed well during playback. In this study, female visits were not translated into extra-pair young. However, a comparable study in black-capped chickadees *Poecile atricapilla* did find an effect of playback treatment on brood paternity.

Interpretation of interactions between playback and a real male could be complicated by changes in behavior of the live male. In the experiments with black-capped chickadees, the behavior of males after each playback treatment was noted, and no effect on subsequent behavior was detected.

This potential problem can be avoided by simulating dyadic interactions with two loudspeakers. Such experiments with nightingales *Luscinia megarhynchos* showed that subjects could associate roles during an interaction with the location of the interactant.

The simulation can be further refined by placing the interacting loudspeakers outside the territory boundary of the subject (ensuring the subject does not approach the loudspeakers during playback), then simulating an intrusion by one of the interactants by playback from within the boundary. This design presented playback treatments that differed only in relative song timing during the interaction, thereby demonstrating that the male great tit subjects had eavesdropped on the interaction as a whole and associated the role of each interactant with features of his song. This approach has been extended to show that great tits combine information from an interaction they took part in with information from an interaction they could eavesdrop on. They have also been shown to combine information on relative features of an interaction with absolute information on singing behavior of each interactant.

Visual Interactions in Fish

Two approaches have been used with visually signaling fish to demonstrate social eavesdropping. In both, the subjects were encouraged to behave as eavesdroppers because they could see the interaction without being seen themselves (achieved by using one-way glass and different light levels between compartments) and could not join the interaction because they were in an adjacent aquarium.

The first approach controlled changes that occur in interacting fish (winner and loser effects) and that might give information on the outcome of interactions rather than (or as well as) information from the dynamics of the interaction. In experiments with male Siamese fighting fish *Betta splendens*, two interactions were staged between similarly sized males displaying across a transparent barrier. However, the subjects could see only one interaction, and the other acted as a control. The staged interactions were allowed to proceed for 15 min, a time that was long enough for a difference in display activity to be detected by subjects but too short for losing males to adopt submissive coloration. Each of the four interacting males (two seen and two unseen) was then introduced into the subject's compartment in turn and the response measured. The male subjects responded to individuals they had seen lose by approaching and displaying sooner than they did to individuals they had seen win. No such differences were seen in response to the unseen males, even though they too had won or lost displays, suggesting that the information used by subjects when responding

was gathered by social eavesdropping. Similar experiments with green swordtails *Xiphophorus helleri* have found comparable evidence for social eavesdropping in aggressive contests between males. Experiments with female fighting fish as subjects followed a protocol similar to that of the experiments with males. The main differences were that the seen and unseen interactions were separated by some days rather than being concurrent and that the females were allowed to move freely so as to exhibit a proximity preference for either male. Females visited seen winners first and more often and spent more time near and displaying to seen winners than to seen losers. Unseen losers were visited first more often than unseen winners, with no difference in the time spent near or displaying to either male. These results are consistent with females' eavesdropping on male–male interactions.

The second approach used to experimentally establish the importance of observed interactions in the future aggressive behavior of male fighting fish has been to decouple the experience of interacting fish from the experience of the subject. This was achieved by placing a small compartment between interacting fish into which the subject could not see. This compartment was filled with water, allowing the males to interact across the gap between them and allowing the subject to see both males involved in a real interaction. However, the compartment could also contain two fish separated by an opaque partition, allowing each of the males seen by the subject to interact with a fish that could not be seen. In this 'apparent' interaction, the subject's interpretation of the aggressive signals given by each visible male is decoupled from that male's actual experience with a hidden fish. Another experiment replaced the two males in the separating compartment with a mirror; therefore, in the apparent interaction, the males were interacting with their own reflection, again decoupling the subject's interpretation from the interactants' experiences. The results showed that the subjects' subsequent response was based on information from interactions, providing support for social eavesdropping.

Effects of an Audience

In communication networks, signaling interactions are likely to be within receiving range of other receivers that do not take part in the interaction. The effect of the presence of such an audience is likely to influence the features of signaling interactions, either because it is an apparent audience (i.e., it is currently present and detected by the interactants) or because it has been present in the evolutionary past, despite being absent during any particular interaction.

Both types of audience impose a selection pressure on signalers, and an audience of predatory or parasitic interceptive eavesdroppers can impose severe selection pressures. Evolutionary audiences result in selection to make signals less easy for interceptive eavesdroppers to locate. Apparent audiences may make the cost of signaling too high to be sustained, and a common response to the presence of a predatory or parasitic eavesdropper is for signaling animals to cease signaling.

The selection pressures imposed on signaling interactions by social eavesdroppers are often less clear cut. It has been argued that the low amplitude of song used by birds in particularly aggressive contexts or when mating is about to occur is an adaptation to prevent social eavesdropping. In some cases, similar considerations seem to have produced the opposite outcome, that is, signaling interactions seem more conspicuous than necessary to transmit information between participants (as discussed above in the context of the high level of close-range signals). Here it may be reasonable to assume that high-amplitude signals are, at least partly, 'intended' for eavesdroppers, that is, that signalers are attempting to communicate more widely in the network than their opponent can. An interesting development of such ideas is that the opponents in an interaction may be subject to different selection pressures, with individuals that are winning wishing to advertise the fact widely and those losing perhaps wishing to withhold that information.

See also: Agonistic and Affiliative Signals: Resolutions of Conflict; Localizing Signal Sources; Signal Production and Amplification in Birds; Signal Identification: Peripheral and Central Mechanisms; Signal Transmission in Natural Environments.

Further Reading

Bradbury JW and Vehrencamp SL (1998) *The Principles of Animal Communication.* Sunderland, MA: Sinauer.

Butlin RK, Guilford T and Krebs JR (eds.) (1993) The evolution and design of animal signalling systems. *Philosophical Transactions of the Royal Society of London, Series B* 340: 161–225.

Dawkins MS (1995) *Unravelling Animal Behaviour,* 2nd edn. Harlow: Longman.

Espmark Y, Amundsen T, and Rosenqvist G (eds.) (2001) *Animal Signals: Signalling and Signal Design in Animal Communication.* Trondheim, Norway: Tapir Academic Press.

Gerhardt HC and Huber F (2002) *Acoustic Communication in Insects and Anurans: Common Problems and Diverse Solutions.* Chicago, IL: Chicago University Press.

Greenfield MD (2002) *Signalers and Receivers: Mechanisms and Evolution of Arthropod Communication*. Oxford, UK: Oxford University Press.

Hauser MD (1996) *The Evolution of Communication*. Cambridge, MA: MIT Press.

McGregor PK (ed.) (2005) *Animal Communication Networks*. Cambridge, UK: Cambridge University Press.

McGregor PK and Peake TM (2000) Communication networks: social environments for receiving and signalling behaviour. *Acta Ethologica* 2: 71–81.

Maynard-Smith J and Harper D (2003) *Animal Signals*. Oxford, UK: Oxford University Press.

Searcy WA and Nowicki S (2005) *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton, NJ: Princeton University Press.

Wyatt TD (2003) *Pheromones and Animal Behaviour: Communication by Smell and Taste*. Cambridge, UK: Cambridge University Press.

Comparative Biology of Invertebrate Neuromuscular Junctions

H L Atwood and M K Klose, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Invertebrates Investigated

Analyses of synaptic transmission and its alteration have been carried out in many different preparations, and neuromuscular junctions (NMJs) have provided a great deal of our understanding of basic principles. Studies of both vertebrate and invertebrate NMJs have generated important findings, yet relatively few of the many invertebrate phyla have actually been investigated in detail, and our knowledge is effectively confined to a few species of arthropods, mollusks, and nematodes, with occasional cursory observations from other phyla. Major contributions to studies of synaptic transmission have been derived from giant neurons of a few mollusks (the squid giant synapse, and synapses of central neurons of the marine gastropod *Aplysia*). However, at the level of NMJs, arthropod species have proved to be more amenable to physiological analysis. By far the bulk of current information on invertebrate neuromuscular synaptic structure and function derives from selected arthropods – decapod crustaceans, and members of two orders of insects, Orthoptera (grasshoppers, locusts, and crickets) and Diptera (flies; in particular, *Drosophila melanogaster*). However, recent advances have also been made in the nematode *Caenorhabditis elegans*. Thus, we base the present account mainly on arthropods and nematodes, groups which represent more species than in any other invertebrate phyla, and which promise to be of continuing interest due to the recent emphasis on genetic factors affecting synaptic structure and function.

During the past century and a half, several basic features of synaptic physiology applicable to invertebrates and often to vertebrate synapses, both central and peripheral, were worked out primarily in decapod crustaceans, with more recent contributions from insects (which share many features with crustaceans). Decapod crustaceans provided the best access for physiologists due to the advantages of their large motor neurons and the accessibility of their NMJs for physiological experiments. It is worth noting the results of general significance which have emerged. In brief, these include: (1) establishment of the amino acids γ-aminobutyric acid (GABA) and glutamate (Glu) as neurotransmitters, (2) presynaptic inhibition mediated by GABA, (3) short-term facilitation due to increase in the number of quanta released per nerve impulse, (4) long-term facilitation, a form of long-term potentiation, which involves semipermanent recruitment of synapses, (5) differential effectiveness of individual terminals of a single neuron, now known to occur commonly in the mammalian central nervous system (CNS), (6) modification of physiological and structural properties of individual neurons through chronic alteration of their activity levels, referred to as long-term adaptation, and (6) distinctive properties of synaptic transmission in uniquely identified 'phasic' and 'tonic' motor neurons, with consequent emergence of the 'identified neuron' concept for invertebrate nervous systems. All of these features have proved to be widely utilized in other nervous systems. Striking short-term changes in transmission (frequency facilitation, short-term depression, and responses to neuromodulating amines and peptides) are especially prominent at crustacean NMJs. Thus, crustacean motor neurons continue to have advantages for combined physiological, optical, and ultrastructural studies of glutamate-mediated neurotransmission.

Despite the manifest advantages of crustacean neuromuscular junctions, experimental studies during the past two decades have increasingly moved to insects, especially *Drosophila*. There is a good reason for this: the *Drosophila* genome has been sequenced, the many genes and their products are known, and excellent techniques for genetic manipulation have been developed. These advantages are currently missing for the crustaceans and for most other invertebrates. Insect and crustacean NMJs have much in common, both in structure and function, so findings are often applicable to both groups. The *Drosophila* larval neuromuscular junction has proved to be amenable to electrophysiological and optophysiological investigations, and currently is the most heavily studied invertebrate neuromuscular junction.

Advantages of genetics and molecular information also attach to the nematode *C. elegans*, and thus NMJs of that organism are also of increasing interest to neurobiologists. However, the small size and difficulty of obtaining electrophysiological recordings have so far largely impeded the types of physiological experimentation that have been successful in arthropods. A much larger nematode, *Ascaris*, has been investigated electrophysiologically, but lacks other features that would make it a favorite of neurobiologists. Limited descriptive studies are also available for a few mollusks and annelids, but their NMJs are less easily accessed for experimentation and also are generally smaller and less visible than those of arthropods. Thus, our account of invertebrate NMJs will of necessity emphasize those of arthropods.

Confusion about synaptic terminology currently exists in the scientific literature, especially that concerned with the *Drosophila* larval neuromuscular junction; therefore, we present our preferred definitions here. 'Neuromuscular junction' is the innervation supplied to a muscle fiber by a motor neuron, and may include several ramifications on the muscle fiber's surface (Figure 1). In the CNS, the equivalent is often referred to as a 'synaptic connection' (the collection of synapses supplied by one neuron to another). 'Synapse' is the specialized contact between pre- and postsynaptic membranes, characterized in electron micrographs by uniform separation and often by conspicuous electron density of the adhering membranes. The contact region is rich in adhesion molecules and is physically very robust; it can be isolated from preparations of the CNS by methods which destroy most cell-to-cell contacts.

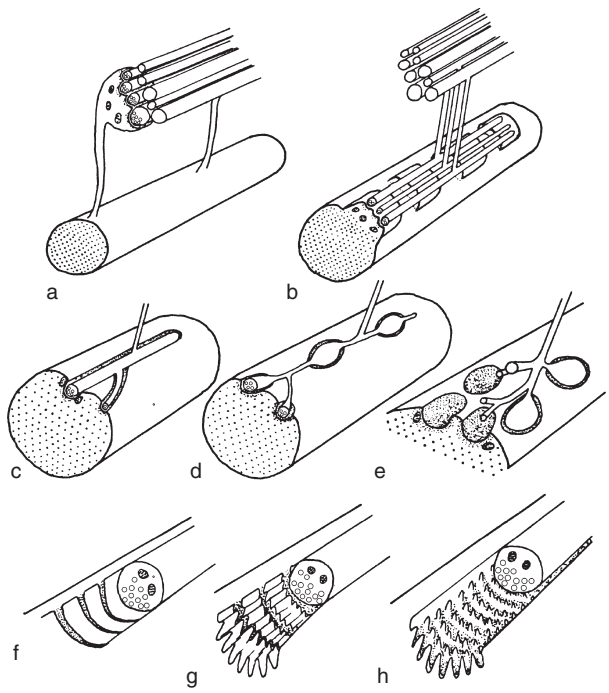

Figure 1 Morphological features of neuromuscular junctions in different phyla. (a) Muscle arms from each muscle cell travel to a motor nerve: each arm expands into a head section that synapses with a few motor axons. Widespread: found in platyhelminths, nematodes, echinoderms, onychophorans, and protochordates. (b) Short pillars extend outward from the muscle cells to receive branches from motor axons. Present in some annelids, mollusks, and insects. (c) Motor axon terminals fit into grooves in the surface of the muscle cells. Synapses occur along the entire length of the axon terminal. Common in arthropods, mollusks, annelids, and fish. (d) Motor axon terminals make varicosities that contact the surface of the cell either with or without shallow indentations at each varicosity. Common in crustaceans, insects, some mollusks, and fish. (e) Compact cluster of varicosities present in some slow vertebrate skeletal muscles, especially those of reptiles, birds, and mammals. (f) Groove with transverse infoldings in which a single motor axon terminal makes functional juxtapositions with each ridge. Characteristic of amphibia. (g) Groove with both longitudinal and transverse infoldings. Seen in some reptilian fast muscle cells. (h) Groove highly infolded both longitudinally and transversely. Seen in fast mammalian muscles. Adapted from Hoyle G (1983) *Muscles and Their Neural Control*. New York: Wiley.

Within the synapse, various specializations for transmission occur. Electrical synapses of the CNS possess gap junctions in which numerous ion-conducting channels (connexons) link pre- and postsynaptic membranes; chemical synapses of the CNS and NMJs are endowed with arrays of postsynaptic transmitter receptors (typically distributed over most of the postsynaptic membrane) and with presynaptic calcium channels and additional structural features that provide for rapid release of neurotransmitter from synaptic vesicles.

An 'active zone' is a specialized region of the presynaptic membrane of chemical synapses functionally adapted for preferential vesicle exocytosis during neurotransmission. The active zone of many synapses of the CNS typically occupies the entire presynaptic membrane, whereas large central synapses and NMJs often have a more localized active zone, and sometimes more than one for an individual synapse. The essential feature of the active zone is the juxtaposition of voltage-gated calcium channels and synaptic vesicles 'docked' at the presynaptic membrane. This association must be spatially tight in order for neurotransmission to occur in the time observed for synaptic delay.

Muscle Innervation

Diversity across Phyla

Invertebrate NMJs are generally quite different from the well-studied amphibian and mammalian counterparts. In different phyla, morphological arrangements vary, as illustrated in Figure 1. In some invertebrates, muscle extensions innervate motor neurons, while in the more 'advanced' phyla (arthropods, mollusks, and most chordates), motor axons travel to muscles in peripheral nerves to form NMJs. Although NMJs of skeletal muscles in mammals usually exhibit compact 'end-plate' structures and complexly contoured postsynaptic folds not found among invertebrates, simpler varicosities are often found in some vertebrate classes. Invertebrate NMJs have a variety of forms, commonly varicose or filiform, with numerous small synapses.

In the arthropod groups best known physiologically – insects and crustaceans – major muscles for locomotion are innervated by peripheral nerves which convey a small complement of motor axons to target muscles. Typically, an entire muscle is innervated by only one to six motor axons, which branch

within the muscle to supply each muscle fiber with many branches and synaptic contacts (multiterminal innervation). Individual muscle fibers often receive more than one motor axon, each physiologically distinct (polyneuronal innervation). The presynaptic nerve terminals are often varicose, with many small synaptic contacts concentrated on the varicosities (sometimes loosely termed 'boutons'). To illustrate general features of innervation in arthropods, we have selected two examples: crustacean limb muscles and insect (*Drosophila*) larval abdominal muscles. In addition, brief comparative overviews of motor innervation in nematodes, and in selected muscles of mollusks, are included.

Crustacean Limb Muscles

An illustrative example (Figure 2) is the well-studied musculature of crayfish claws and walking legs. Individual muscles are innervated by one to four excitatory motor neurons (all glutamatergic), and by one or two inhibitory neurons (GABAergic). Each neuron typically provides synapses to hundreds or thousands of muscle fibers in one or more muscles. In decapods generally, several muscles receive two motor axons, traditionally termed 'fast' and 'slow,' after the types of contraction they evoke when separately stimulated. More recently, they have often been termed 'phasic' and 'tonic,' based on the nature of their normal nerve impulse discharges. The types of contractions evoked by these axons depend in part on their selective innervation of fast and slow muscle fibers, respectively, but also on their radically different modes of synaptic transmission (to be considered later).

Peripheral inhibitory innervation is a common (but not invariant) feature of arthropod muscles. For decapod leg muscles, there are generally three inhibitory neurons, two of which innervate primarily the 'opener' and 'stretcher' muscles which share an excitatory neuron. These muscles can contract independently only when one of the two specific inhibitory neurons is also active. The third inhibitory neuron, termed the 'common inhibitor,' innervates all of the leg muscles (with the possible exception of the claw opener muscle). Its likely function is to speed up the locomotor response through selective inhibition of slowly contracting muscle fibers in all leg muscles. In many decapod limb muscles, inhibitory axons act directly on their target muscle fibers, generating inhibitory postsynaptic potentials (Figure 2). In some instances, the inhibitory axons also act on presynaptic terminals of motor axons, producing presynaptic inhibition.

Variations in innervation occur in other decapod muscles. Abdominal muscles of crayfish and lobsters receive excitatory and inhibitory innervation, but the

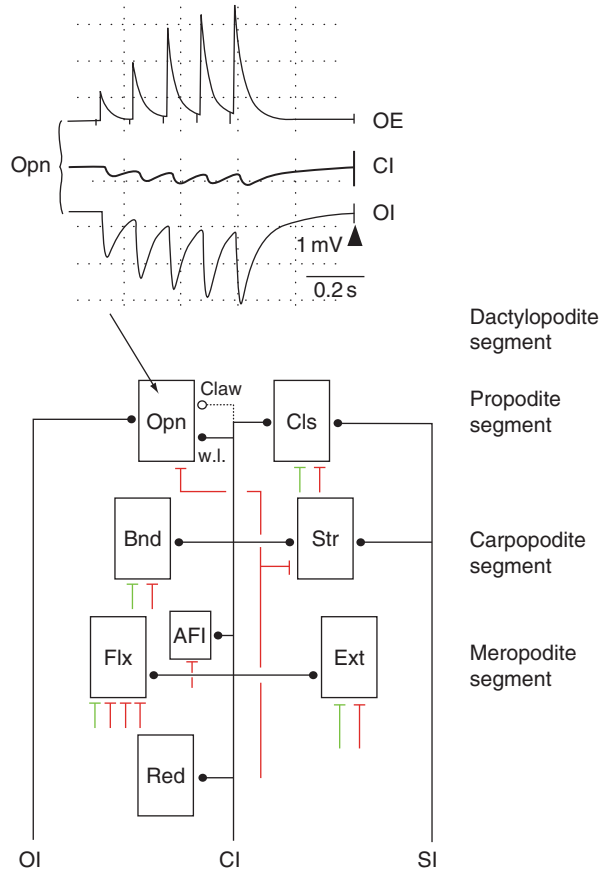

Figure 2 Innervation of crayfish limb muscles. Limb segments containing the muscles are indicated. The paired muscles of the claws and walking legs (w.l.) are each supplied by one to four excitatory axons (red and green 'T' connections) and by one or two inhibitory axons. Paired excitatory axons of the closer (Cls), bender (Bnd), and main extensor (Ext) muscles are differentiated as phasic (green) and tonic (red) subtypes. The opener (Opn) and stretcher (Str) muscles share a single excitatory motor axon. Representative synaptic potentials recorded from the opener muscle with an intracellular microelectrode in response to stimulation of isolated axons illustrate facilitating depolarizing excitatory junction potentials for the opener excitor (OE) axon, and facilitating hyperpolarizing inhibitory junction potentials for the OI and CI axons. The OI axon generates a larger potential than does the CI axon. In the claw, CI innervation is vestigial. Other muscles of the limb are the main flexor muscle (Flx), with quadruple excitatory innervation (one axon exhibits phasic properties); the accessory flexor muscle (AFI), with a single excitatory axon; and the reductor of the meropodite (Red), with undefined excitatory innervation. OI, opener muscle inhibitor; CI, common inhibitory; SI, stretcher muscle inhibitor. Adapted from Wiens TJ (1993) The closer muscle is a second target for the stretcher inhibitor motoneuron of the crayfish's thoracic limbs. *Journal of Comparative Physiology A* 173: 435–444.

inhibitory effects are entirely postsynaptic. In addition, fast and slow contractions are generated by separate sets of muscles arranged in parallel and separately innervated by phasic and tonic motor neurons. In stomach muscles, there is no peripheral inhibitory innervation. At least one of these muscles receives cholinergic

excitatory innervation, while the majority receive glutamatergic innervation. It would seem that the complex physiological scenario of the limb muscles reflects the more refined uses of these appendages, although the means of achieving the required control would probably not have been arrived at by a human engineer.

Drosophila Larval Abdominal Muscles

The second example is an important one because of the large number of recent studies pertaining to it. The *Drosophila* larval abdomen, which accounts for most of the animal's mass at that stage of development, is segmentally arranged; each hemisegment is supplied by its own nerve (Figure 3). A simple dissection exposes all of the musculature and its innervation. In the well-characterized abdominal segments 2–7, each hemisegment has 30 individual muscles (each actually a single muscle fiber) innervated by 31 well-characterized motor neurons, all uniquely identified. Muscle identification is based on position, orientation, and shape, in addition to molecular expression patterns, while motor neuron nomenclature is based on either the origin of the axon or its target, determined by way of retrograde or anterograde single-cell dye labeling. Each muscle receives innervation from one or two glutamatergic excitatory motor neurons supplying prominent boutons, on each of which a number of small neuromuscular synapses occur (Figure 4). Two types of these terminals can be distinguished: type 1b ('big') and type 1s ('small'). Each muscle receives innervation from type 1b terminals, and some receive in addition type 1s terminals. The neurons supplying the latter innervate several different muscle fibers (a type of common excitatory innervation). The type 1 neurons provide the primary locomotor drive for this system. Type 1b terminals produce on average smaller synaptic potential amplitudes while type 1s (smaller bouton size) terminals generally produce larger amplitudes. Small clear synaptic vesicles containing glutamate are found in these boutons along with a small number of dense-core vesicles containing, in many cases, pituitary adenylyl cyclase-activating peptide (PACAP)-like neuropeptide. Other modulatory peptides are found; for example, in muscle 8, dense-core vesicles reveal leukokinin immunoreactivity. Thus, peptidergic cotransmission likely occurs at the glutamate-releasing synapses; its physiological role is not well defined.

No peripheral inhibitory neurons have been found, but two additional types of innervation occur on some muscle fibers. Type 2 innervation consists of long strings of small boutons, each forming one or two neuromuscular synapses (Figure 4). These neurons contain both glutamate and octopamine. Additional innervation, type 3, occurs on only a few muscle fibers and is neurosecretory: type 3 terminals of muscle 12 show insulin-like immunoreactivity and contain other peptides as well. Electrical excitation of type 2 and

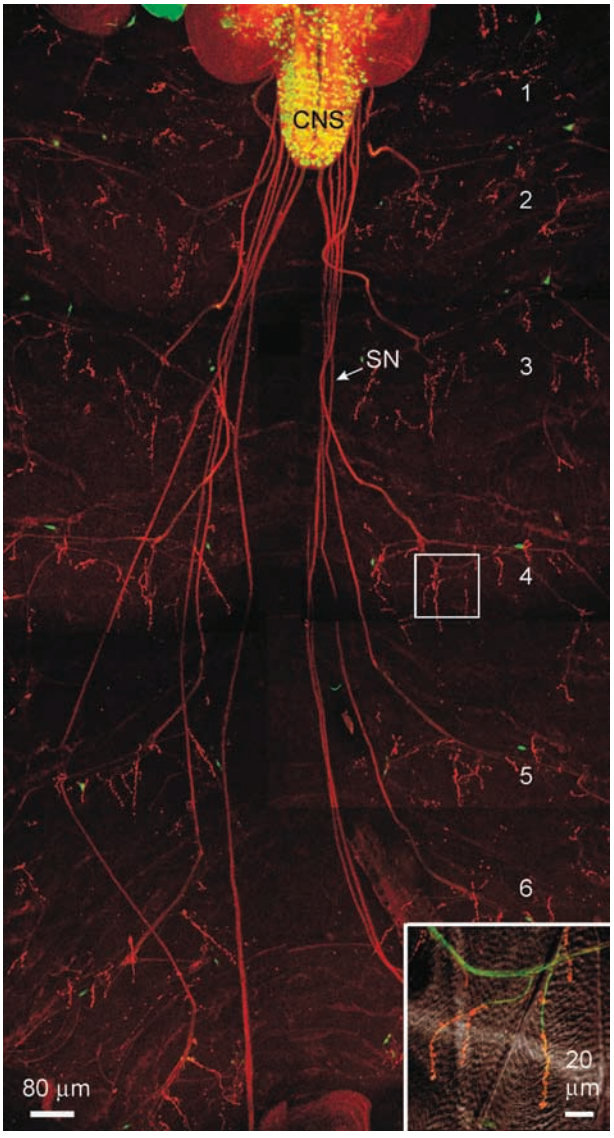

Figure 3 *Drosophila* larval muscle innervation. Neuromuscular junctions are shown in an immunostained specimen expressing green fluorescent protein and treated with a fluorescent antibody for the endocytotic enabler protein dynamin (produced by the *shibire* gene). Segmental nerves (SN), appearing red, descend from the central nervous system (CNS), to supply each abdominal segment (numbered) with motor and sensory innervation. Each abdominal hemisegment has 30 muscle fibers ('muscles'), each supplied by 1–4 efferent axons (see Table 1). Inset shows an enlargement of exposed neuromuscular junctions in part of one segment; here, brighter staining indicates a relatively high level of dynamin in synapse-bearing boutons. The specimen is from a D42-Gal4 strain containing one copy of the transgene *UAS-shibire*^{ts1} and one copy of transgene *UAS-gfp* (green fluorescent protein). Specimen prepared and photographed by Dr. Balaji Iyengar, University of Toronto.

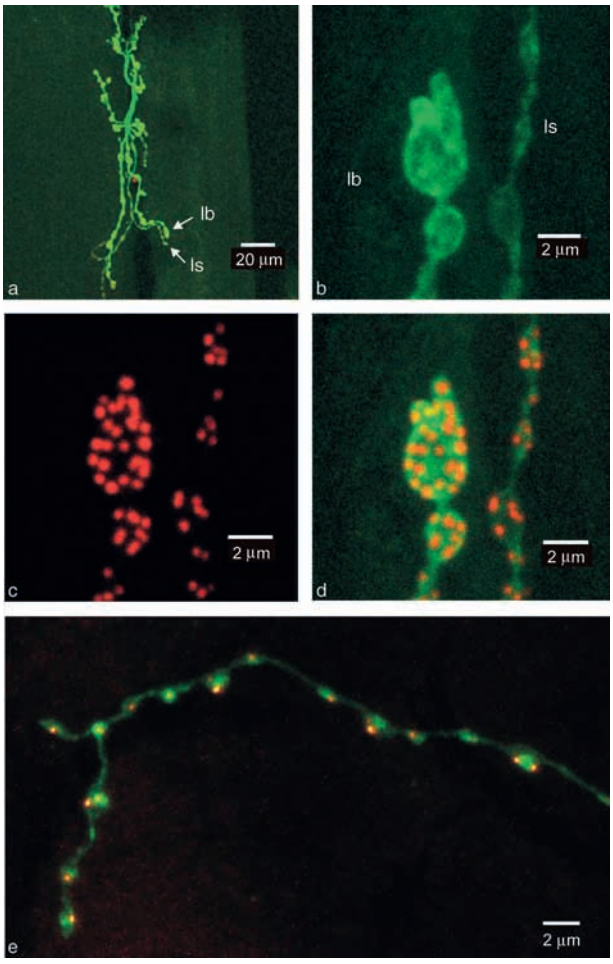

Figure 4 Neuromuscular junctions and synapses of larval *Drosophila*. Three types of innervation are illustrated, together with staining of individual synapses. (a) Low-power view of the neuromuscular junctions on abdominal muscles 6 and 7 formed by two parallel motor axons which supply type 1 (1b and 1s) innervation. Three other panels (b–d) show enlargements of two types of varicosities, type 1b (large varicosity) and type 1s (smaller varicosity), belonging to the two different motor axons. The varicosities are stained with anti-horseradish peroxidase antibody (b, green) and antibody nc82, against the active zone structural protein bruchpilot (c, red), to illustrate multiple synapses on each bouton; the merged image is shown in panel (d). Type 1s boutons have higher emission rates per synapse compared to type 1b boutons. Note the variable number of synapses per bouton for both axon types. (e) A third type of innervation (type 2) observed on a few of the abdominal muscle fibers, and characterized by long strings of numerous small varicosities. The merged image of anti-horseradish peroxidase and nc82 staining demonstrates only one or two synapses on each varicosity. (a–d) Prepared and photographed by J Dason, University of Toronto. (e) Prepared and photographed by Dr. G Macleod, University of Texas at San Antonio.

type 3 terminals reveals no prominent fast postsynaptic electrical responses, but small postsynaptic currents (probably caused by glutamate release) have been detected at type 2 boutons. Thus, type 1 terminals are responsible for fast neuromuscular transmission, while type 2 and type 3 terminals play mainly modulatory roles, as yet not well defined.

Significant innervation features are summarized in Table 1. Although up to four different motor neurons can innervate a single muscle cell, only one neuron supplying each type of bouton (1b, 1s, 2, or 3) innervates a given muscle (bouton-type specific exclusion). The most common scenario is triple innervation (1b, 1s, and 2) as seen on muscles 1–4, 8–10, 13, 14, 20–24, and 30.

Other Arthropod Muscles

Insect muscles show a variety of innervation patterns, some simpler than that of the *Drosophila* larva, and others more complex. Simplicity of innervation is exemplified by longitudinal flight muscles of adult *Drosophila*; the six individual muscle fibers ('muscles') of this group each receive unineuronal excitatory innervation. Each motor neuron impulse leads to a large excitatory junction potential (EJP) and muscle spike, eliciting oscillatory myogenic contractions at a high frequency, necessary for flight. Examples of more complex innervation are found in the much-studied locust leg tibial extensor and flexor muscles. The extensor muscle has 'fast' and 'slow' excitatory axons (analogous to those of crustacean limb muscles), and receives a branch of a common inhibitory axon as well as modulatory octopaminergic innervation. The antagonistic flexor muscle is innervated by nine excitatory motor neurons which are selectively, but not exclusively, distributed to subsections of the muscle, and by two inhibitory neurons. Many fibers are innervated by four or more axons. This innervation pattern is the most complex discovered to date for an insect muscle.

Other arthropod groups have been investigated sparsely. Limb muscles of the primitive arachnid horseshoe crab *Limulus* and a tarantula spider show less differentiation of muscle fibers and their attendant NMJs than observed in insects and crustaceans. Muscles are polyneuronally innervated, often by two or three motor axons which generate EJPs of different amplitudes. The main flexor muscle of the *Limulus* leg has as many as 12 excitatory axons. Peripheral inhibition may also occur, since the muscle responds to GABA. In general, the neuromuscular properties found so far in arachnids are functionally less specialized than in crustaceans and insects.

Nematodes

The general features of multiterminal and polyneuronal innervation occur in most other invertebrate groups, including annelids, mollusks, and nematodes. In the latter group, efferent muscle nerves are replaced by

Table 1 Innervation of *Drosophila* larval abdominal muscles

Property	*Type 1b innervation*	*Type 1s innervation*	*Type 2 innervation*	*Type 3 innervation*
Muscles innervated	1–24, 28, 30[1]	1–4, 6–10, 12–16, 18–24, 30[1]	1–5, 8–24, 30[1]	12 (15, 16)[1]
Terminal length	Short[2–4]	Longer[2–4]	Very long[2–4]	
Terminal shape	Minimal terminal branching[2–4]	More elaborate branching[2–4]	Most elaborate branching[2–4]	
Bouton size	Big, 2–6 μm[5]	Small, 1–4 μm[5]	Smallest, 1–2 μm	Small, 1–3 μm
Synapses/bouton	Many (20–40)[2]	Fewer (3–15)[2]	Least (1 or 2)	
T-bars/synapse	0–4[2]	0–3[2]		
Neuroactive substances	Glu[6]/PACAP-like peptide[7]/leukokinin[8]	Glu[6]/PACAP-like peptide[7]/leukokinin[8]	Glu[6]/octopamine[9,10]/proctolin[11]	Insulin-like peptide, other peptides[12]
Dense-core vesicles	Few[4]	Intermediate[4]	Many, elliptical[13]	Many, spherical[13]
Clear vesicles	Numerous, smaller (39-nm diameter)[14]	Fewer, larger (45 nm diameter)[14]		
Readily releasable pool fraction	Low proportion of vesicle pool is readily releasable[15]	High proportion of vesicle pool is readily releasable[15]		
Ca^{2+} sensitivity of release	Low[15]	High[15]		
Initial EJP size	Intermediate to small[2,16]	Large to intermediate[2,16]	Small	No EJP
Response pattern	Tonic-like (more pronounced facilitation)[16]	Phasic-like (depression)[16]		
Mitochondria/terminal	Prominent[2,4]	Small[2,4]	Very small[17]	
K^+ channels	Shaker/Shab[18]	Shaker/Shab[18]	Shal	Shal[19]
Subsynaptic reticulum	Profuse folding[2,17]	Some folding[2,17]	Little folding[17]	Minimal, bouton on surface[17]
GluRIIA and GluRIIB levels	↑ IIA, ↑ IIB[20]	↓ IIA, ↑ IIB[20]	↓ IIA, no IIB[20]	

Notes and references: muscles 15 and 16 may have type 3 innervation (indicated in parentheses). Excitatory junction potential (EJP) properties were observed in HL3 solution and may be different in the intact organism. Superscript numbers refer to the following references: (1) Hoang B and Chiba A (2001) *Developmental Biology* 229: 55–70; (2) Atwood HL, Govind CK, and Wu C-F (1993) *Journal of Neurobiology* 24: 1008–1024; (3) Shayan AJ and Atwood HL (2000) *Journal of Neurobiology* 43: 89–97; (4) Renger JJ, Ueda A, Atwood HL, et al. (2000) *Journal of Neuroscience* 20: 3980–3992; (5) Lnenicka GA and Keshishian H (2000) *Journal of Neurobiology* 43: 186–197; (6) Johansen J, Halpern ME, Johansen KM, et al. (1989) *Journal of Neuroscience* 9: 710–725; ((7) Zhong Y and Pena LA (1995) *Neuron* 14: 527–536; (8) Cantera R and Nässel DR (1992) *Cell and Tissue Research* 269: 459–471; (9) Keshishian H, Chiba A, Chang TN, et al. (1993) *Journal of Neurobiology* 24: 757–787; (10) Rivlin PK, St Clair RM, Vilinsky I, et al. (2004) *Journal of Comparative Neurology* 468: 596–613; (11) Anderson MS, Halpern ME, and Keshishian H (1988) *Journal of Neuroscience* 8: 242–255; (12) Gorczyca M, Augart C, and Budnik V (1993) *Journal of Neuroscience* 13: 3692–3704; (13) Budnik V, Zhong Y, and Wu C-F (1990) *Journal of Neuroscience* 10: 3754–3768; (14) Karunanithi S, Marin L, Wong K, et al. (2002) *Journal of Neuroscience* 22: 10267–10276; (15) Atwood HL, Karunanithi S, Georgiou J, et al. (1997) *Invertebrate Neuroscience* 3: 81–87; (16) Kurdyak P, Atwood HL, Stewart BA, et al. (1994) *Journal of Comparative Neurology* 350: 463–472; (17) Jia X-X, Gorczyca M, and Budnik V (1993) *Journal of Neurobiology* 24: 1025–1044; (18) Ueda A and Wu C-F (2006) *Journal of Neuroscience* 26: 6238–6248; (19) Covarrubias M, Wei AA, and Salkoff L (1991). *Neuron* 7: 763–773; (20) Marrus SB, Portman SL, Allen MJ, et al. (2004) *Journal of Neuroscience* 24: 1406–1415.

arms of muscle fibers which extend to the major nerve cords to receive synaptic contacts (Figure 1). This mode of innervation is also found in echinoderms, platyhelminth worms, and the onychophoran *Peripatus*.

The success of genetic approaches in *C. elegans* has led to use of its neuromuscular system for discovery of synaptically important molecular pathways. Both excitatory (cholinergic) and inhibitory (GABAergic) motor neurons supply all of the longitudinal locomotory muscle cells. The complete wiring diagrams for the nervous system of both *C. elegans* and the larger parasitic nematode *Ascaris* have been worked out through painstaking serial sectioning. The structural similarities between the two species are remarkable. In *C. elegans*, which has about 20 000 genes, there are 302 neurons, 95 muscle-cells, about 7000 synapses, and 2000 neuromuscular junctions altogether. All cells are specified and identified. The neurons and their distributions are much the same in *Ascaris*. Physiological research is a much easier task in the latter, and it is likely that physiological features in the two species (and in other nematodes) are basically the same.

Motor neurons are confined to the longitudinal nerve cords and connecting commissures, and have relatively few processes. In fact, complex neuropile

structures like those of arthropods and mollusks do not occur in nematodes. Physiological investigations in *Ascaris* have shown that the electrical length constants are very long (centimeters) for nematode motor neurons, and that these neurons do not produce action potentials. Instead, electrical activity and resultant synaptic transmission are graded. Probably tonic release of neurotransmitter at the NMJs accounts for maintained muscular contraction. Stimuli applied to neuronal processes in nerve cords and commissures do evoke synaptic potentials, either depolarizing or hyperpolarizing, in the muscle fibers, sometimes with graded muscle spikes superimposed. The connections of the excitatory and inhibitory motor neurons to different longitudinally arranged muscle groups produce the typical oscillatory waves responsible for locomotion.

Mollusks

Muscle innervation in the mollusks is very diverse, but generally follows the organizational scheme of polyneuronal, multiterminal innervation seen in other major invertebrate groups. Motor axons may be sent to muscles from central ganglia, but often arise instead from local small ganglia within the muscle, a feature which creates semiautonomous muscular activity. Unlike arthropods, the molluscan groups employ acetylcholine as the major excitatory neurotransmitter. Peripheral inhibition has been found in a few molluscan muscles, but is apparently less utilized than in arthropods and nematodes. On the other hand, numerous modulatory neuropeptides occur in molluscan muscles; many of them are likely cotransmitters. They may exert potentiating effects on the muscle, probably through activation of membrane calcium channels, or inhibitory effects through activation of potassium channels; such dual, concentration-dependent effects are outlined for one class of peptides, the myomodulins. Another class (buccalins) acts presynaptically, modulating neurotransmitter release.

Intracellular recordings from muscle cells in a variety of different muscles show EJPs of different amplitudes (polyneuronal innervation). Accompanying contractions in the muscle are usually graded in accordance with the depolarization produced by EJPs and are stronger when muscle action potentials are initiated. However, an unusual form of regulation is present in certain muscles of bivalve mollusks. The anterior byssal retractor muscle (ABRM) of the mussel *Mytilus* develops an extremely prolonged, powerful contraction ('catch' contraction) in response to application of acetylcholine, the major excitatory transmitter substance. Application of serotonin, contained in a second set of nerves, relaxes the catch. Thus, the serotonergic nerves are referred to as 'relaxing nerves.' Serotonin may produce its effect by promoting sequestration of Ca^{2+} within the muscle cells.

Fast and slow contractions are often elicited in molluscan muscles participating in swimming or escape responses. A particularly good example has been described in the pteropod mollusk *Clione limacine*, which is specialized for swimming at variable speeds. Swimming muscles contain slow-twitch and fast-twitch muscle fibers, which are recruited for slow and rapid swimming, respectively. The former are innervated by small motor neurons which have small innervation fields. A second type of input is produced by large motor neurons which have widespread innervation fields, including both fast- and slow-twitch muscle fibers. These 'general excitor' neurons become active when fast swimming is required, and their activity generates spikelike electrical events and strong contractions in muscle cells throughout the innervation field. This system represents a counterpart to the specialized phasic and tonic neuromuscular systems of crustaceans, in which differences in speed of response are accomplished through activation of motor neurons with appropriate synaptic properties and distributions.

Structural Features of Invertebrate Neuromuscular Junctions

Synapses

NMJs typically are formed by varicose terminals or linear filiform branches of motor axons, although in nematodes muscle processes reach the motor neurons in the nerve cords to form the junctions. Invariably, small synapses, often numerous, are found at invertebrate NMJs (Figure 4). A single varicosity on an arthropod muscle fiber typically bears 5–40 individual synapses. At higher resolution in electron micrographs, synapses are characterized by rigidly aligned, electron-dense pre- and postsynaptic membranes uniformly separated by 10–20 nm (Figures 5 and 6). On the postsynaptic (muscle) side, a complex reticular structure (subsynaptic reticulum) extending from the muscle fiber receives the synapses; the postsynaptic membrane exhibits regular arrays of particles in freeze-fracture replicas, and these are thought to be the receptors for the major neurotransmitter. On the presynaptic side, prominent presynaptic densities (typically 0–4 per synapse) are associated with clusters of clear synaptic vesicles, some of which appear to be attached to the presynaptic density by microfilaments. The general arrangement of synapses differs from that generally seen in skeletal muscles of mammals, in which a single localized large consolidated end-plate is often found on each muscle fiber.

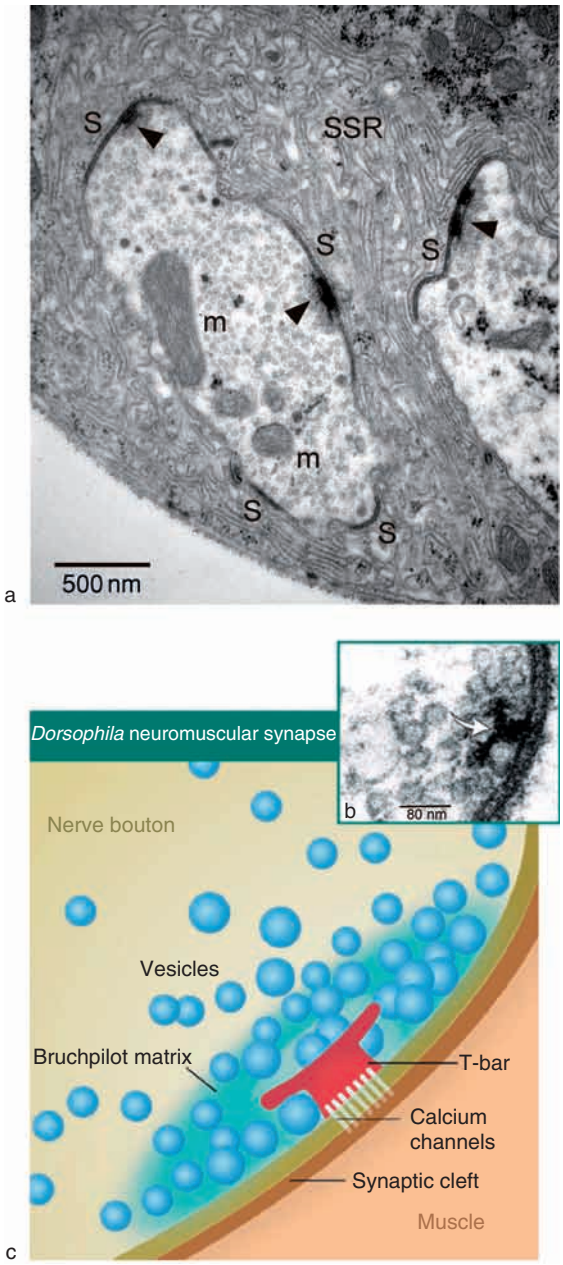

Figure 5 Specializations of *Drosophila* neuromuscular junctions. (a) Overview of two presynaptic boutons illustrating variable T-bar (arrowheads) morphology. In the right-hand example, two dense pillars share a single T-projection. In the middle example, very fine filaments emanate from the T-bar. In the left-hand example, a clear separation appears between the dense pillar and the presynaptic membrane. (b) Enlarged view of a single T-bar (arrow) with clustered synaptic vesicles, and a docked vesicle under one arm of the T-bar. (c) Diagram of the active zone, illustrating the relationship between the T-bar, synaptic vesicles, calcium channels, and a structural protein Bruchpilot which surrounds and is closely associated with the active zone. Antibody (nc82) labeling of this protein reveals the presence of a synapse endowed with one or more active zones. SSR, subsynaptic reticulum; S, synaptic; m, mitochondrion. Electron micrographs taken by Dr. L Marin, University of Toronto. From Atwood HL (2006) Gatekeeper at the synapse. *Science* 312: 1008–1009.

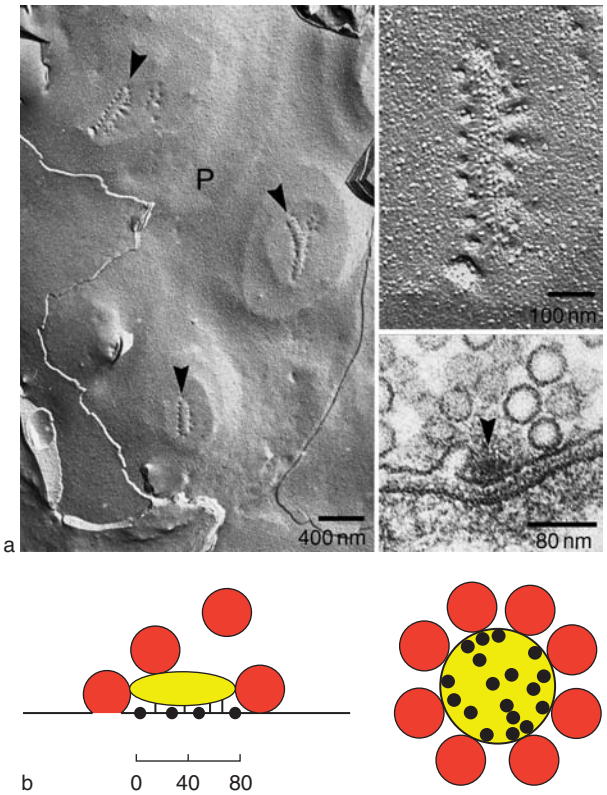

Figure 6 Active zone of the American lobster, *Homarus americanus*. (a) Freeze-fracture replicas (left image) of an excitatory synaptic bouton's P-face, showing three synaptic contacts (raised plateaus), with active zones (arrowheads) consisting of variably shaped collections of large intramembranous particles surrounded by images of vesicular fusion (shown at higher magnification in right top image). Transmission electron micrograph (lower right) shows a presynaptic dense body (arrowhead) with an adjacent image of vesicular fusion, and additional nearby synaptic vesicles. (b) Diagrams based on the transmission electron micrographs and freeze fracture replicas. Face view (left) and orthogonal (transverse) view (right) are drawn to scale. Presynaptic Ca^{2+} channels (black circles) appear in apparently random arrangement throughout the presynaptic membrane region occupied by the dense bar (yellow). The latter appears to be tethered just above the presynaptic membrane, as seen in transmission electron micrographs. Synaptic vesicles (red) are typically 'docked' around the edges of the dense bar, and exocytosis occurs preferentially at that location, as illustrated (scale in nanometers). (a) Reproduced from Atwood HL and Govind CK (1990) Activity-dependent and age-dependent recruitment and regulation of synapses in identified crustacean neurones. *Journal of Experimental Biology* 153: 105–107, with permission from the company of Biologists. (b) Reprinted from Pearce J, Govind CK, and Shivers RR (1986) Intramembranous organization of lobster excitatory neuromuscular synapses. *Journal of Neurocytology* 15: 241–252, with kind permission of Springer Science and Business Media.

Active Zones

The synaptic specialization termed the 'presynaptic active zone,' denoting a small region specialized for intense secretory activity, was originally recognized at amphibian NMJs, but similar features have been

characterized at invertebrate NMJs and at central synapses. In transmission electron micrographs of arthropods (Figures 5 and 6) and nematodes (Figure 7), the 'signpost' of the active zone is an electron-dense adhesion on the presynaptic membrane, variable in appearance. In crustaceans it is often hemispherical (Figure 6), while in *Drosophila* and some other insects it has a typical T-shaped appearance, and is referred to as a 'T-bar' (Figure 5). In all cases, the electron-dense structure appears to be

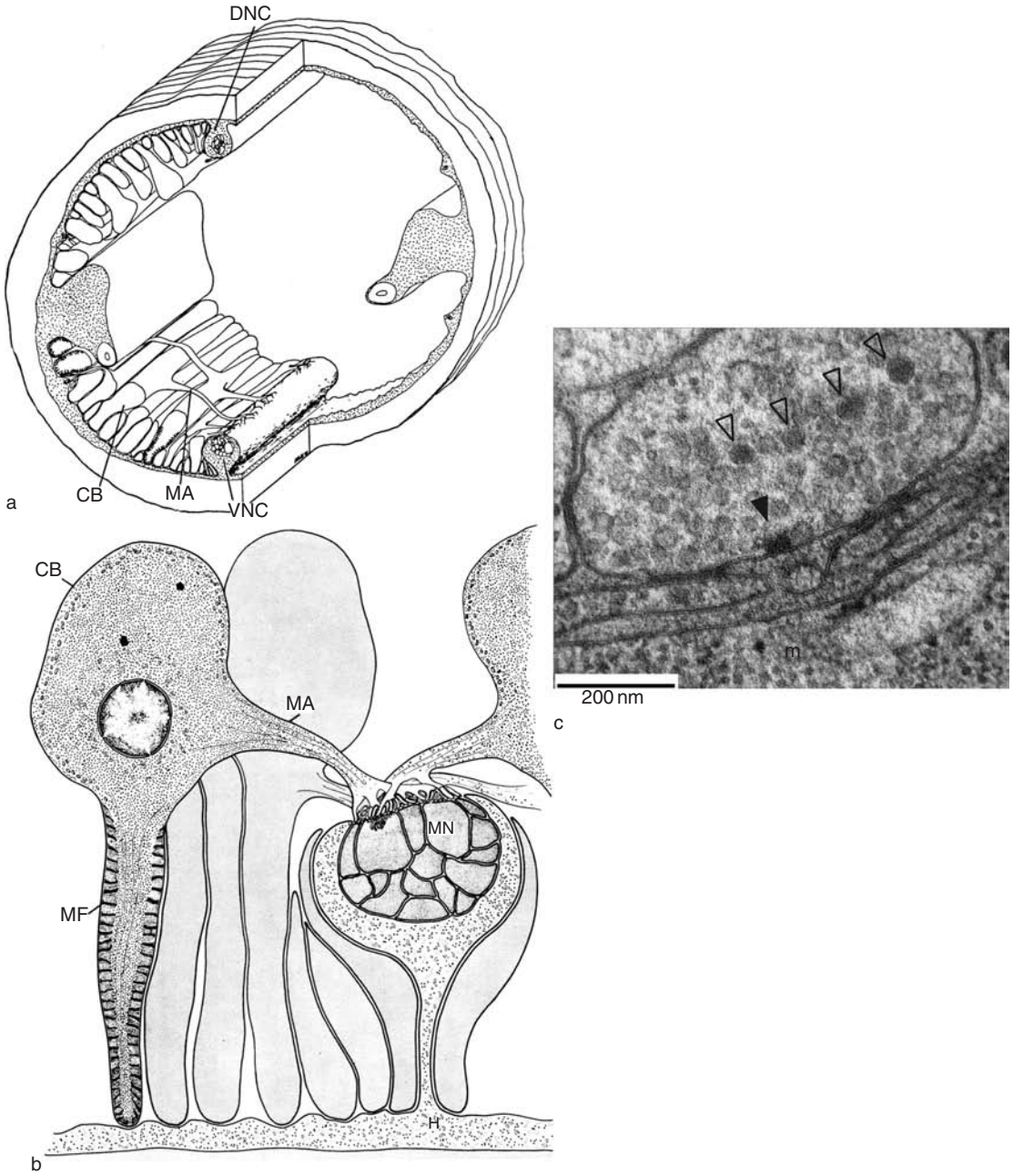

Figure 7 Neuromuscular junctions of nematodes. (a, b) Drawings of the neuromuscular arrangements in *Ascaris*. Muscle cells with contractile regions (MF) and large noncontractile cell bodies (CB) extend 'muscle arms' (MA) to dorsal and ventral nerve cords (DNC and VNC, respectively). Synapses are made between muscle arms and motor neurons (MN) in the nerve cords. (c) Neuromuscular junction in *Caenorhabditis elegans*. Cross section of a neuromuscular junction obtained after high-pressure freezing to preserve synaptic components. The nerve terminal contains numerous clear synaptic vesicles and four dense-cored vesicles (open arrowheads); a dense projection (closed arrowhead) has a clear synaptic vesicle docked beside it. The granular process of the postsynaptic muscle fiber (m) receives the synapse and appears to possess closely associated laminar subdivisions. (a, b) From Hoyle G (1983) *Muscles and Their Neural Control*. New York: Wiley. (c) From Weimer RM, Gracheva EO, Meyrignac O, Miller KG, Richmond JE, and Bessereau JL (2006) UNC-13 and UNC-10/rim localize synaptic vesicles to specific membrane domains. *Journal of Neuroscience* 26: 8040–8047; image courtesy of Drs. J-L Bessereau and R Weimer.

slightly detached from the presynaptic membrane, but tethered to it (Figure 6). In this respect, it bears similarity to some vertebrate sensory synapses which have large electron-dense structures, either round or ribbon-shaped, attached at the active zone. The presynaptic density appears to be the focal point within the synapse for synaptic vesicle accumulation at the presynaptic membrane; most of the vesicles that adhere to the presynaptic membrane ('docked' vesicles) are found around the presynaptic density and some are attached to it by fine filaments (Figures 5 and 6). Thus, cytoskeletal proteins are thought to promote the arrangement and movement of synaptic vesicles.

Evidence from freeze-fracture replicas and immunostaining strongly supports the view that presynaptic voltage-gated calcium channels are closely associated with the presynaptic density, and that the clear synaptic vesicles are preferentially released along its sides, close to the presynaptic calcium channels (Figure 6). Prominent intramembranous particles visible in freeze-fracture replicas are thought to be mainly synaptic calcium channels. In many arthropods, these are not regularly arrayed, as at vertebrate NMJs. Many occur on the membrane surface under the dense projection, but the small gap between the latter and the presynaptic membrane may facilitate diffusion of Ca^{2+} to the edges of the active zone, where vesicle fusion occurs. The active zone provides the necessary spatial arrangement for fast neurosecretion; vesicles to be released must be very close to presynaptic calcium channels in order to be activated in 200 μs or less by the calcium trigger that is known to precipitate vesicle fusion with the presynaptic membrane.

The presynaptic voltage-gated Ca^{2+} channels (VGCCs) responsible for exocytosis have been partially characterized in *Drosophila* and can be compared with known vertebrate counterparts. Combined studies using selective pharmacological disruption of calcium channels and immunocytochemical characterization of calcium channel distribution suggest that N-type and P/Q-type voltage-gated calcium channels are the main players of vertebrate synaptic transmission. Both are high-voltage activated and they are sensitive to ω-conotoxin GVIA and ω-agatoxin VIA, respectively. At least 10 vertebrate α1 subunits (the pore-forming components) and at least 4 β subunits (supplementary modulatory components) have been cloned. In contrast, *Drosophila* has only three genes encoding pore-forming α1 subunits. *DmCa1D* encodes an α1 subunit homologous to that of vertebrate Cav1 VGCCs, *DmCa1A* (*cacophony*) encodes an α1 subunit homologous to that of Cav2 VGCCs, and cDNA has been cloned for a homolog of the α1 subunit of vertebrate Cav3 VGCCs. Only one gene coding for a β subunit has been identified in *Drosophila*. The *Drosophila* presynaptic calcium channels, while homologous to N- and P/Q-type channels in location and function, are not highly sensitive to either ω-conotoxin GVIA or ω-agatoxin VIA.

The *Drosophila* neuromuscular junction contains several distinct calcium channels; three have been described for the presynaptic terminal. A *cacophony*-based calcium channel, sensitive to the spider toxin *Plectreurys* toxin II (PLTXII), is confined to the active zone, and is solely responsible for fast synchronized transmitter release. A La^{3+} and flunarizine-sensitive Ca^{2+} channel, likely more diffusely distributed, is involved in endocytosis. A third type, a non-*cacophony* PLTX-sensitive calcium channel thought to be involved in endocytosis, is likely also diffusely distributed.

Other differences between invertebrate and vertebrate calcium channels involve the synaptic protein-interacting ('synprint') site. In vertebrates, key docking proteins (syntaxin 1, SNAP25, and synaptotagmin 1) bind to an intracellular domain of the II–III linker region of the α1 subunit of mammalian N-type and P/Q-type calcium channels. Injection of rat synprint peptide to competitively dissociate interaction of the synprint site with synaptic proteins diminishes synaptic transmission, supporting the hypothesis that a physical association is necessary for fast synaptic transmission. However, the invertebrate homologs neither appear to require such a region nor a physical interaction with synaptic proteins for fast transmission. Work done on *Lymnaea* has revealed that despite lack of direct interaction between synaptic proteins and calcium channels, synprint peptide injection disrupts synaptic transmission as it does in vertebrate preparations. This supports the hypothesis that neurotransmitter release is blocked downstream of calcium entry, and suggests that earlier results may need to be reexamined.

Several additional synaptic active zone molecules have recently come to light in studies of *Drosophila* and *C. elegans*. As yet, the coordinating roles of these molecules are not well understood, but many are important for synapse formation and optimal release of neurotransmitter. As an example, the *Drosophila* anchoring protein Bruchpilot, a structural protein with coiled-coil domains and homologies to a mammalian active zone protein, is essential for localizing the T-bar to the presynaptic membrane. *Drosophila* mutants lacking this protein also do not exhibit the normal T-shaped presynaptic density or normal presynaptic calcium channel distribution. Although they can maintain evoked neurotransmission, its effectiveness is reduced to about one-fourth of normal, and the timing of exocytosis is less precise. An antibody to Bruchpilot (nc82) is now used extensively to reveal the number and distribution of active zones

(or synapses with active zones) at the *Drosophila* neuromuscular junction (Figures 4 and 5).

Several examples of active zone-associated molecules have been discovered in *C. elegans*, using screens for mutants with abnormal locomotion and defective presynaptic morphology. Among these are RPM-1 (regulator of presynaptic morphology), which regulates formation and spacing of active zones in motor neurons. This protein possesses guanine nucleotide exchange factor and zinc finger domains, suggesting a role in regulating formation of protein complexes. Similar proteins are found in *Drosophila* presynaptic terminals (Highwire) and at mammalian synapses (Pam), indicating conservation of function. As yet, a general model of invertebrate active zones that would account for the roles of all of the structural and regulatory proteins has not emerged.

Neuroactive Substances and Receptors

General Features

The majority of invertebrate neurons communicate via chemical synapses, which employ several major neurotransmitters. Complete definition of these neuroactive substances involves the characterization of their biosynthetic pathways, their receptors, and their second-messenger pathways. Glutamate is the primary excitatory neurotransmitter of arthropod NMJs, while acetycholine is a major excitatory transmitter of NMJs in nematodes, annelids, mollusks, and a few crustacean stomach muscles. The primary transmitter of invertebrate inhibitory NMJs is GABA, though glutamate can also activate inhibitory receptors in certain cases.

In the presynaptic terminal, glutamate is synthesized from α-ketoglutarate by way of glutamate dehydrogenase and is taken up into vesicles to be mobilized for release. In inhibitory terminals, glutamate is further metabolized by glutamate decarboxylase into GABA. Both GABA and glutamate are actively removed from the synaptic cleft by transporters in adjacent cell membranes.

Primary neurotransmitters acting directly at invertebrate NMJs are often co-released with agents that can exert effects elsewhere. Proctolin, a pentapeptide unique to arthropods, is released as a cotransmitter at some NMJs, but exerts its effects on the extrasynaptic muscle membrane. Some neuroactive substances are released at locations remote from the neuromuscular junction upon which they act, and access their targets either by diffusing to nearby NMJs (neuromodulator) or by entering into the circulation and targeting distant sites (neurohormone). Neuromodulatory neurons of insects often contain octopamine (OCT), while peripheral endings of some neurosecretory cells contain several neuropeptides. In fact, all of the glutamatergic and GABA-containing neurons contain variable numbers of dense-core vesicles, most of which are probably loaded with neuropeptides. A few cases of peptidergic cotransmission with glutamate have been investigated. Representative neurotransmitters released at invertebrate NMJs as well as some of the better-known neuroactive substances that can modulate transmission are presented in Table 2.

Numerous receptors, both pre- and postsynaptic, are present in and around invertebrate NMJs. The sum of activity of excitatory and inhibitory ionotropic synaptic receptors determines the ultimate effect of transmission on electrical responses of the postsynaptic target. Additional receptors, including extrasynaptic, metabotropic, neuromodulatory, and neurohormonal receptors, modulate the effects of primary neurotransmission.

Several excitatory and inhibitory receptors were initially identified in crustacean and insect neuromuscular synapses through dose–response experiments using several analogs of the endogenous transmitters. Receptors have been historically categorized by their preferred agonist, a substance that potently activates the receptor in a way similar to the endogenous transmitter. Further classification is based on antagonist pharmacology, location (pre/post and synaptic/extrasynaptic), and biophysical characteristics (desensitization, conductance, current and ion selectivity).

Postsynaptic Ionotropic Glutamate Receptors

Ionotropic glutamate receptors are the main transducers of excitatory neuromuscular activity in arthropods and they are also essential for rapid neuronal signaling in the vertebrate CNS. Vertebrate glutamate receptors, classified by the agonist of highest potency, include α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors, which have fast kinetics and are typically impermeable to calcium, and *N*-methyl-D-aspartate (NMDA) receptors, which are slower and do conduct calcium. Invertebrate ionotropic glutamate receptors show a mixture of the vertebrate characteristics, having fast kinetics and calcium permeability. They are not very responsive to the major vertebrate agonists. Early studies in *Drosophila*, locust, and crayfish neuromuscular preparations found quisqualate to be the most potent agonist of excitatory ionotropic receptors at the neuromuscular junction. The quisqualate-type invertebrate receptors are inhibited by antagonists such as the polyamine-containing toxins from the venoms of spiders (argiotoxins) and wasps (philanthotoxins). Interestingly, NMDA has been found to induce presynaptic effects through a metabotropic receptor in the crayfish NMJ. Other glutamate

Table 2 Invertebrate endogenous neuroactive substances

Transmitter	*Glu*	*GABA*	*Ach*	*5-HT*	*Dopa*	*Oct*	*Hist*	*Adren*	*Noradr*	*NO*	*FMRF*	*SCP/MM/ BUC*	*Proct*
Insect													
Drosophila	±	–	na	na	Δ	Δ	na	na	na	Δ	Δ	na	Δ
Locust/ cockroach	±	–	(+)	Δ	Δ	Δ	na	na	na	?	Δ	na	Δ
Crustacean Crayfish/ lobster/crab	±	–	(+)	Δ	Δ	Δ	Δ	na	na	Δ	Δ	na	Δ
Nematode *Caenorhabditis elegans*		–	+	Δ/(+)	Δ	Δ	–	na	na	Δ	Δ/±	na	na
Annelid Earthworm/leech		–	+	Δ	Δ	Δ	na	Δ	Δ	Δ	Δ	na	na
Mollusk *Aplysia/Mytilus*		–	+	Δ/–	Δ/–	Δ	na	Δ	Δ	Δ	Δ/±	Δ	na

Ach, acetylcholine; Adren, adrenaline; BUC, buccalins; Glu, glutamate; FMRF, FMRFamide-related peptide family; Dopa, dopamine; Hist, histamine; NO, nitric oxide; Oct, octopamine; MM, myomodulins; Noradr, noradrenaline; Proct, proctolin; 5-HT, serotonin; SCP, small cardioactive peptides. +, –, and ± indicate that the effects of a primary neurotransmitter on neuromuscular transmission can be excitatory, inhibitory, or both, respectively, depending on target cell (na, not applicable). Other neuroactive subsatances can act as neuromodulators or neurohormes to modulate the effects of a primary transmitter (denoted by Δ); parentheses indicate cursory observations from a few muscles. A role for NO in modulating neuromuscular transmission has been suggested but not fully substantiated in some systems (denoted by ?), including the locust slow extensor tibiae preparation. Data based on the following sources: Brezina V, Bank B, Cropper EC, et al. (1995) *Journal of Neurophysiology* 74: 54–72; Osborne RH (1996) *Pharmacology and Therapeutics* 69: 117–142; Nistri A and Constanti A (1979) *Progress in Neurobiology* 13: 117–235; Muneoka Y, Fujisawa Y, Matsuura M, et al. (1991) *Comparative Biochemistry and Physiology C* 98: 105–114.

receptors found postsynaptically on the muscle membranes of locusts and crayfish pass chloride ions and are therefore inhibitory. It is uncertain whether these receptors occur directly at the synapse.

Glutamate released into the synaptic cleft binds to receptors located primarily, if not entirely, at the postsynaptic density (PSD), causing an increase in the channel open probability. Binding of several glutamate molecules is required for a single channel to open. Once open, channels are permeable to Na^+, K^+, and Ca^{2+}. Biophysical characterization of ionotropic quisqualate-sensitive receptors reveals a high degree of similarity between several species of arthropods in most categories (Table 3). Application of a high concentration of agonist (10 mM) activates these nonselective excitatory receptors, revealing a high conductance level (100–150 pS) and a current of 7–9 pA (in single-channel muscle patches held at –70 mV). This steplike change in channel conductance contributes to the rapid change in muscle membrane potential typically preceding muscle contraction. The probability of a channel being open is dependent upon the concentration of glutamate in the cleft. Bursts of channel openings quickly reach a peak and then decay over the duration of glutamate's presence. Prolonged exposure to glutamate causes channel desensitization. Although the time constant of desensitization is concentration dependent, recovery is not.

Unlike the interspecies similarities that exist in conductance levels and pharmacology, desensitization characteristics are quite different among the arthropods. Postsynaptic receptor populations in *Drosophila* desensitize more slowly than in crayfish and recover from desensitization much more slowly than in crayfish or locust. Desensitization kinetics also differentiate subtypes of the quisqualate receptors within species. Outside-out patches from crayfish muscle revealed several types of ionotropic glutamate receptors, which differ mainly in desensitization kinetics. One type of glutamate receptor completely desensitizes, another partially desensitizes, while a third does not desensitize. The time constant of desensitization of both the completely and partially desensitizing crayfish channels are very similar, while they differ in the different locust (25 and 5 ms) and *Drosophila* (19 and 2 ms) channels.

Quantal currents (produced by release of glutamate from single synaptic vesicles) can be explained from the observed channel kinetics under certain assumptions. The rapid rise time of quantal currents at the crayfish NMJ indicates that glutamate concentration in the cleft should be approximately 1 mM, to open channels rapidly. At this concentration, the time constant of desensitization is similar to the decay constant of the quantal current. Assuming that responding postsynaptic receptors are not the nondesensitizing type, desensitization may terminate the quantal current. However, most of these data come from recordings

Table 3 Arthropod postsynaptic ionotropic glutamate receptors

Postsynaptic ionotropic receptor	*Agonist*	*Antagonist*	*i(pA)*	*g (pS)*	P_o	T_{rise} *(ms)*	α *(ms)*	β *(ms)*	τ_d *(ms)*	τ_{recov} *(ms)*	K_d	*Ion*
Drosophila DGluRIIA	Quis > glu	Arg^1_{636}	~9[1]	120[1]	0.6[2]	0.2[3]	0.1[2]	0.2[2]	19[1]	150[2] (mixed pop)	2 mM[4]	Na^+; Ca^{2+}, $K^{+(5)}$
Drosophila DGluRIIB	Quis > glu	Arg^1_{636}	~7[1]	120[1]	0.6[2]	0.2[3]	0.1[2]	0.2[2]	2.0[1]	–	2 mM	Na^+; Ca^{2+}, $K^{+(5)}$
Locust	Quis > glu[5]	Arg^6_{636},PhTx	~8[7]	125–150[7,8]	0.7[9]	0.6[3]	6[8] 0.2[8]	–	25[8,10] 3.0[10]	1–20 4[8]		Na^+; Ca^{2+}, $K^{+(5)}$
Crayfish	Quis > glu[5]	NSTX[11]	~8[12]	100[12]	0.6[10]	0.4[13]	3[3]	0.2[3]	1–2[14]* Nondesensitizing	2–3[10]	1 mM[15]	Na^+; Ca^{2+}, $K^{+(5)}$

This table illustrates a variety of receptor types found at invertebrate neuromuscular junctions. The well-characterized ionotropic glutamate receptor reveals great similarity across invertebrate species except for desensitization properties (splits). Arg, argiotoxin636 (phenolic polyamine); *g*, conductance (single channel); glu, glutamate; *i*, current; Ion, indicates ion selectivity; KA, kainate; K_d, agonist concentration at one-half maximal activation; NSTX, nephila spider toxin (phenolic polyamine); PhTx, philanthotoxin; P_o, open probability (with 10 mM glutamate); Quis, quisqualate; T_{rise}, rise time of single-channel currents (10–90%); α, opening rate; β, closing rate; τ_d, time constant of desensitization; τ_{recov}, time constant of recovery from desensitization; *, up to four different excitatory ionotropic glutamate receptors have been identified at the crayfish neuromuscular junction based on desensitization characteristics – an incompletely desensitizing channel, a completely desensitizing channel, a nondesensitizing channel, and a channel which can make transitions from one state to another. Superscript numbers refer to the following references: (1) DiAntonio A, Petersen SA, Heckmann M, et al. (1999) *Journal of Neuroscience* 19: 3023–3032; (2) Heckmann M and Dudel J (1997) *Biophysical Journal* 72: 2160–2169; (3) Dudel J, Adelsberger H, and Heckmann M (1997) *Invertebrate Neuroscience* 3: 89–92; (4) Heckmann M, Parzefall F, and Dudel J (1996) *Pflügers Archives* 432: 1023–1029; (5) Nistri A and Constanti A (1979) *Progress in Neurobiology* 13: 117–235; (6) Kerry CJ, Ramsey RL, Sansom MS, et al. (1988) *Brain Research* 459: 312–327; (7) Patlak JB, Gration KA, and Usherwood PN (1979) *Nature* 278, 643–645; (8) Dudel J, Franke C, Hatt H, et al. (1990) *Neuroscience Letters* 114: 207–212; (9) Bates SE, Sansom MS, Ball FG, et al. (1990) *Biophysical Journal* 58: 219–229; (10) Dudel J, Franke C, and Hatt H (1990) *Biophysical Journal* 57: 533–545; (11) Klose MK, Atkinson JK, and Mercier, AJ (2002) *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 187: 945–952; (12) Hatt H, Franke C, and Dudel J (1988) *Pflügers Archives* 411: 8–16; (13) Tour O, Parnas H, and Parnas I (1998) *Biophysical Journal* 74: 1767–1778; (14) Franke C, Hatt H, and Dudel J (1987) *Neuroscience Letters* 77, 199–204; (15) Dudel J, Adelsberger H, and Heckmann M (2000) In: Wiese K (ed.) *Crustacean Nervous Systems*. Berlin: Springer; (16) Dudel J, Franke C, and Hatt H (1990) *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 166: 757–768.

of individual or small groups of receptors isolated from the muscle, without any indication of the full composition or relative contribution of receptor types located at the PSD. How individual receptors contribute to the synaptic currents evoked during neuromuscular transmission is best documented at the *Drosophila* larval neuromuscular junction. Genetic studies, supported by immunocytochemistry, have been employed to assess the relative contributions of individual receptor subtypes to synaptic responses.

Determining the full complement of receptors of an organism requires the integration of molecular and genetic studies, which in *Drosophila* and *C. elegans* have allowed a more complete elucidation of these receptors. The *Drosophila* model is presently ideal since it allows for combined pharmacological, kinetic, and genetic studies to examine the physiological mechanisms of synaptic transmission. Over 30 glutamate receptor genes have been identified in the *Drosophila* genome (15 in *C. elegans*), several of which are ionotropic, including *DGluRI* (kainite), *DNMDAR* (NMDA), and *DGluRIIA*, *B*, *C*, *D*, and *E*. Five muscle-specific subunits have been identified: DGluRIIA, DGluRIIB, DGluRIIC, DGluRIID, and DGluRIIE. DGluRIIA and DGluRIIB subunits are found in all somatic muscles but not in the CNS. The two genes are adjacent in the genome, have similar genomic organization, and are closely related, though their products share only 44% amino acid identity. In the putative pore region, critical for Ca^{2+} permeability, the sequence of the DGluRIIA is identical to that for vertebrate channels, while the sequence of DGluRIIB is divergent.

The postsynaptic excitatory ionotropic glutamate receptor in *Drosophila* muscles consists of four subunits. The GluRIIC, GluRIID, and GluRIIE subunits are required, in addition to either the GluRIIA or GluRIIB subunit. Immunocytochemistry reveals GluRIIB-containing receptors are abundant at both type 1b and type 1s boutons, while GluRIIA-containing receptors are most abundant at type 1b boutons. This suggests that postsynaptic sensitivity is determined by a mixture of both types. Genetic deletion of one or the other receptor subunits allowed characterization of the biophysical characteristics of each receptor type (see Table 3). In GluRIIA-containing mutants, postsynaptic sensitivity was greater than in GluRIIB-containing mutants. Increased levels of DGluRIIA increase quantal size, while decreased levels decrease quantal size. Conversely, overexpressing DGluRIIB decreases quantal size. These data also suggest that postsynaptic sensitivity is determined by a combination of both GluRIIA and GluRIIIB receptors. However, evidence from single-channel recordings of receptors on the muscle membrane and relative sensitivity to argiotoxin suggests that DGluRIIA receptors contribute more to normal (wild-type) EJP amplitude and time course than do DGluRIIB receptors. The slow recovery from desensitization of *Drosophila* receptors may in part explain decreased amplitude of EJPs during fictive locomotion, when a high rate of nerve impulse activity (50–100 Hz) is often attained. Functionally, this may serve as a negative feedback control of muscle contraction.

Other Invertebrate Neuromuscular Receptors

Features of other receptors found at invertebrate NMJs are summarized in Table 4. Biophysical studies have been confined mainly to those of arthropods. Ionotropic GABA receptors (GABA-A type) are found in postsynaptic muscle membranes of many arthropods and in presynaptic terminals of crustacean NMJs. Both in the muscle and in the presynaptic nerve terminals, GABA opens chloride channels, effectively counteracting the depolarizing effects of concurrent inward current flow. In muscle membranes, this is sufficient to attenuate or diminish action potential amplitude in muscle fibers that produce electrogenic responses (Figure 8). In nerve terminals, the presynaptic action potential is reduced in amplitude, and consequently the nerve terminal's transmitter output decreases (presynaptic inhibition).

Three different ionotropic GABA channels have been characterized in crayfish. In the opener muscle, a GABA channel exhibits a conductance of 29 pS, with a rise time of about 1 ms (50 mM GABA) and a 0.12 mM K_d (agonist concentration at half maximal binding). In the deep abdominal extensor muscle the conductance is greater, the rise time is slower, and the K_d is higher. When the deep abdominal extensor muscle is denervated, conductance drops from 35 to 31 pS, rise time increases from 5 to 10 ms, and the sensitivity to GABA increases slightly from 1 to 0.8 mM. These changes following denervation are attributed to an alteration in subunit composition, analogous to that seen in vertebrate nicotinic receptors following denervation.

In *Drosophila*, 10 genes appear to code for ionotropic GABA receptor subunits, while *C. elegans* possesses 37 genes. Single-channel recordings from *Drosophila* muscle reveal GABA receptors with conductance substates of 40, 80, and 110 pS. The 40 and 80 pS conductances were also elicited by application of glutamate. Nothing is known about the role of GABA receptors in *Drosophila* neuromuscular transmission, but in *C. elegans* GABAergic neuromuscular transmission is of major importance for locomotion.

Several other presynaptic receptors have the potential to modulate transmitter release at invertebrate NMJs. At the lobster stretcher muscle in the

Table 4 Representative invertebrate neuromuscular receptors

Species	*Natural ligand*	*Ionotropic/ metabotropic*	*Pre-/ postsynaptic*	*Agonists*	*Antagonists*	*Excitatory/ inhibitory/ modulatory*	*Identifying characteristics*
Insects							
Drosophila body wall DmGluRA	Glutamate[1,2]	Metabotropic	Pre	(S)-4C3HPG/DCG-IV/ (1S,3S)-ACPD	MCCG-I	Modulatory	Attenuates facilitation; potentiates transmission
Drosophila body wall	Octopamine[3]	Metabotropic	Pre		Yohimbine	Inhibitory	
Drosophila body wall	Insulin-like peptide[4]		Post				
Drosophila body wall	Glutamate and GABA[5]	Ionotropic	Post			Inhibitory	Cl^-
Schistocerca extensor tibeae	Acetylcholine[6]	Ionotropic	Pre	Nicotine	Tubocurarine	Excitatory	
Schistocerca extensor tibeae	Acetylcholine[6]	Ionotropic	Pre	Muscarine/mecholine		Excitatory	
Schistocerca foregut	Acetylcholine[7]	Ionotropic	Post	Nicotine	Tubocurarine	Excitatory	
Schistocerca foregut	Acetylcholine[7]	Ionotropic	Post	Muscarine/acetyl-β-methylcholine/ oxotremorine	Atropine	Excitatory	
Locust extensor tibeae	GABA[8,9]	Ionotropic	Post	Muscimol	Picrotoxin; bicuculline[9]	Inhibitory	Cl^-
Locust extensor tibeae	Glutamate[10,11] (GABA insensitive)	Ionotropic	Post	Ibotentate/aspartate	Picrotoxin	Inhibitory	Cl^-
Schistocerca slow extensor tibiae	Proctolin[12]		Post			Modulatory	Sustained catch contraction
Locust oviduct	Proctolin[7]	Metabotropic	Post			Modulatory	
Locust oviduct	FMRF/ FLRFamide[13]	Metabotropic	Post	Bztc		Modulatory	↑ OR, ↓ muscle activity*
Locust oviduct	Tyramine[14]	Metabotropic				Modulatory	↓ Muscle tension, ↓ proctolin-induced contraction
Locust slow extensor tibeae[7;15]	Octopamine[7,15–17]	Metabotropic	Pre	Naphazoline > Clondine	Phentolamine = metoclopramide[17]	Excitatory	(2A) ↑ spontaneous release, ↑ evoked release
Locust fast and slow extensor tibiae	Octopamine[7,15,16]	Metabotropic	Pre		Phentolamine > metoclopramide	Excitatory	(2B) ↑ relaxation rate
Locust myogenic bundle	Octopamine[7,15,16]		Post	Indolalkamine/5-hydroxytryptamine	Chlorpromazine/yohimbine	Inhibitory	(1) Slows myogenic bundle rhythm
Locust myogenic bundle	Serotonin[15]		Post			Excitatory	Accelerates myogenic bundle rhythm
Locust foregut[15]	Dopamine[7]		Post			Excitatory	↑ [DA] induces contractions
Locust foregut[15]	Dopamine[7]		Post			Modulatory	↓ [DA] reduces contractions

Continued

Table 4 Continued

Species	*Natural ligand*	*Ionotropic/ metabotropic*	*Pre-/ postsynaptic*	*Agonists*	*Antagonists*	*Excitatory/ inhibitory/ modulatory*	*Identifying characteristics*
Periplaneta americana slow coxal depressor	Proctolin[18]		Post			Excitatory	
Crustaceans							
Crayfish (3 types)	GABA[9,19]	Ionotropic	Pre	β-GP			** DEAM ** Denervated ** Opener
Crayfish deep abdominal extensor	Glutamate[20]	Metabotropic	Pre	Quisqualate > NMDA	APV/CNQX	Inhibitory	
Crayfish opener	GABA[19,21,22]	Ionotropic	Post	Ibotenate; muscimol[9]	Picrotoxin; bicuculline; isoflurane	Inhibitory	
Cancer borealis gm6 muscle	Acetylcholine[23,24]	Ionotropic	Post	Nicotine[25]	Picrotoxin	Excitatory	
Cancer borealis gm6 muscle	Glutamate[23]	Ionotropic	Post			Excitatory	
Lobster stretcher	Glutamate[26]	Metabotropic	Pre		Pertussis	Inhibitory	Opens K^{+} $^{+}$channel
Lobster stretcher	GABA[26]	Metabotropic	Pre	Baclofen		Inhibitory	Opens K^{+} $^{+}$channel
Lobster stretcher	GABA[26]	Ionotropic	Pre			Inhibitory	Cl^{-}
Lobster stretcher	GABA[26]	Ionotropic	Post			Inhibitory	Cl^{-}
Crayfish	Acetylcholine	Ionotropic	Post			Excitatory	
Crayfish	Octopamine[27]		Pre			Modulatory	↑ Quantal release
Lobster	Octopamine[27]		Post			Excitatory/ modulatory	↑ Muscle tone
Annelids							
Annelids	Acetylcholine[28]	Ionotropic	Post	Levamisole		Excitatory	
Earthworm body wall	GABA[28]	Ionotropic	Post		Bicuculline methiodide/ picrotoxin	Inhibitory modulatory	Cl^{-}
Leech body wall	GABA[28]	Ionotropic	Post		Bicuculline methiodide	Inhibitory/ modulatory	Cl^{-}
Leech body wall	FMRFamide[28]		Post			Modulatory	Muscle tone ↑ by KNEFIRFamide and ↓ by SDPNFLRFamide
Leech body wall	Serotonin[29]	Metabotropic		TET/*N*-PIPER/1-PIPER; IAMP/2M-5-HT/αM-5-HT/5MT***		Modulatory	
Mollusks							
Aplysia intrinsic buccal muscle 3	Acetylcholine[30]	Ionotropic	Post		Hexamethonium/tubocurarine	Excitatory	
Aplysia intrinsic buccal muscle 3	Glutamate [30]	Ionotropic	Post			Excitatory	

Aplysia intrinsic buccal muscle 3	FMRF[30]					Modulatory	↑ B3 EJP amplitude; ↓ B9 EJP amplitude
Aplysia intrinsic buccal muscle 3	Small cardioactive peptide[30]	Metabotropic				Modulatory	↑ EJP amplitude and ↑ rate of relaxation
Aplysia radular opener	Myomodulin[31]	Metabotropic				Modulatory	
Lymnaea head retractor muscle	Ach[32]	Ionotropic					
Lymnaea head retractor muscle	Glutamate[32]	Ionotropic					
Lymnaea head retractor muscle	GABA[33]	Ionotropic			Picrotoxin	Inhibitory	
Lymnaea head retractor muscle	FMRFamide[32]					Modulatory	
Mytilus edulis	Acetylcholine[34,35]	Ionotropic	Post			Excitatory	Eserine potentiates Ach-induced catch contraction
Mytilus edulis anterior byssus retractor	Serotonin[34]	Metabotropic	Post		Mersalyl	Inhibitory/ modulatory	↓ Ach-induced contraction through AC
Mytilus edulis anterior byssus retractor	Dopamine[34,36,37]	Metabotropic	Post		Bromo-LSD/methysergide	Modulatory/ inhibitory	↓ Ach-induced contraction through AC
Mytilus edulis anterior byssus retractor	FMRFamide[35,37]		Post			Excitatory/ inhibitory	↓ [FMRF]relaxes Ach-induced catch contraction; ↑ [FMRF] causes contraction
Mytilus edulis anterior byssus retractor	Octopamine[38]		Post			Modulatory	↑ Ach-induced catch contraction
Nematodes							
C. elegans body wall	Acetylcholine[28,39]	Ionotropic	Post	Leviamisole	Bungarotoxin/tubocurare	Excitatory	Na^+/K^+
C. elegans body wall	Acetylcholine[39]	Ionotropic	Post	Nicotine	Dihydro-β-erythroidine	Excitatory	
C. elegans body wall	GABA[28]	Ionotropic	Post	Muscimol		Inhibitory	Cl^-
Ascaris summ body wall	FMRFamide[40]		Post			Modulatory	
Ascaris summ pharynx/body wall	Serotonin[41, ****]					Excitatory/ modulatory	

Continued

Table 4 Continued

Species	*Natural ligand*	*Ionotropic/ metabotropic*	*Pre-/ postsynaptic*	*Agonists*	*Antagonists*	*Excitatory/ inhibitory/ modulatory*	*Identifying characteristics*
Nematode body wall	Serotonin[42]	Metabotropic	Post			Modulatory	↓ Ach-induced contraction through AC
Nematode pharynx	Serotonin[42]	Metabotropic	Post			Modulatory	↓ Ach-induced contraction through IP_3
Nematode pharynx	Glutamate[43,44]	Ionotropic	Post	Ibotenate/kainate/ quisqualate	Picrotoxin (weak)		Cl^-
Nematode body wall	FMRFamide[42]	Metabotropic	Post			Excitatory/ inhibitory	↓ [FMRF] relaxes Ach-indcued contraction; ↑ [FMRF] causes contraction

Notes: *, Evidence supports one receptor mediating both excitatory and inhibitory activity, and not separate receptors. Peptides activate different second-messenger systems through one receptor. **, Pharmacologically the same but kinetically different receptors. ***, Four distinct serotonin responses observed in leech pharynx, suggesting several receptors. ****, Reverse transcription polymerase chain reaction reveals alternative splicing results in four isoforms of serotonin receptors. TET, (6)-8-hydroxy-2-(di-*n*-propylamino)tetralin; N-PIPER, *N*-(3-trifluoromethylphenyl) piperazine; 1-PIPER, 1-(*m*-chlorophenyl)-piperazine; IAMP, (6)-2,5-dimethoxy-4-iodoamphetamine; 2M-5-HT, 2-methyl-5-hydroxytryptamine; αM-5-HT, α-methyl-5-hydroxytryptamine; 5MT, 5-methoxytryptamine; (*S*)-4C3HPG, (*S*)-4-carboxyl-3-hydroxyphenylglycine; DCG-IV, 2-(2,3-dicarboxycyclopropyl)glycine; (1*S*,3*S*)-ACPD, 1-aminocyclopentane-1*S*,3*S*-dicarboxylate; AC, adenylyl cyclase; Ach, acetylcholine; APV, (±)-2-amino-5-phosphonovaleric acid; Bztc, benzethonium chloride; β-GP, β-guanidinopropionic acid; CNQX, 6-cyano-7-nitroquinoxaline-2,3-dione; MCCG-I, (2*S*,3*S*,4*S*)-2-methyl-2-(carboxycyclopropyl) glycine ; ↓, decreases/relaxes. Superscript numbers refer to the following references: (1) Bogdanik L, Mohrmann R, Ramaekers A, et al. (2004) *Journal of Neuroscience* 24: 9105–9116; (2) Zhang D, Kuromi H, and Kidokoro Y (1999) *Neuropharmacology* 38: 645–657; (3) Nishikawa K and Kidokoro Y (1999) *Brain Research* 837: 67–74; (4) Gorczyca M, Augart C, and Budnik V (1993) *Journal of Neuroscience* 13: 3692–3704; (5) Delgado R, Barla R, Latorre R, et al. (1989) *FEBS Letters* 243: 337–342; (6) Fulton BP (1982) *Neuroscience* 7: 2117–2124; (7) Osborne RH (1996) *Pharmacology and Therapeutics* 69: 117–142; (8) Cull-Candy SG (1986) *Journal of Physiology* 374: 179–200; (9) Nistri A and Constanti A (1979) *Progress in Neurobiology* 13: 117–235; (10) Cull-Candy SG (1978) *Journal of Physiology* 276: 165–181; (11) Dudel J, Franke C, Hatt H, et al. (1989) *Brain Research* 481: 215–220; (12) Baines RA, Lange AB, and Downer RG (1990) *Journal of Comparative Neurology* 297: 479–486; (13) Lange AB, Orchard I, Wang Z, et al. (1995) *Proceedings of the National Academy of Sciences of the United States of America* 92: 9250–9253; (14) Donini A and Lange AB (2004) *Journal of Insect Physiology* 50: 351–361; (15) Roeder T (2002) *Microscopy Research Techniques* 56: 237–247; (16) Evans PD (1981) *Journal of Physiology* 318: 99–122; (17) Howell KM and Evans PD (1998) *Journal of Experimental Biology* 20(part 13): 2053–2060; (18) Adams M and O'Shea E (1983) *Science* 221: 286–289; (19) Dudel J, Adelsberger H, and Heckmann M (1997) *Invertebrate Neuroscience* 3: 89–92; (20) Parnas I, Dudel J, Parnas H, et al. (1996) *European Journal of Neuroscience* 8: 116–126; (21) Fiszer DP and De Robertis E (1973) *FEBS Letters* 33: 45–48; (22) Dudel J, Adelsberger H, and Heckmann M (2000) In: Wiese K (ed) *Crustacean Nervous Systems*. Berlin: Springer; (23) Lingle C and Auerbach A (1983) *Journal of General Physiology* 81: 571–588; (24) Marder E and Paupardin-Tritsch E (1980) *Brain Research* 181: 223–227; (25) Marder E and Paupardin-Tritsch E (1978) *Journal of Physiology* 280: 213–236; (26) Miwa A, Ui M, and Kawai N (1990) *Journal of Neurophysiology* 63: 173; (27) Breen C and Atwood HL (1983) *Nature (London)* 303: 716–718; (28) Walker RJ, Holden-Dye L, and Franks CJ (1993) *Comparative Biochemistry and Physiology C* 106: 49–58; (29) O'Gara BA, Illuzzi FA, Chung M, et al. (1999) *General Pharmacology* 32: 669–681; (30) Keating C and Lloyd PE (1999) *Journal of Neurophysiology* 82: 1759–1767; (31) Evans CG, Vilim FS, Harish O, et al. (1999) *Journal of Neurophysiology* 82: 1339–1351; (32) Kononenko NL, and Zhukov VV (2005) *Zoology (Jena)* 108: 217–237; (33) Kononenko NL and Zhukov VV (2005) *Rossiskii Fiziologicheski Zhurnal Imeni I M Sechenova* 91: 300–308; (34) Twarog BM, Muneoka Y, and Ledgere M (1977) *Journal of Pharmacology and Experimental Therapeutics* 201: 350–356; (35) Muneoka Y and Saitoh H (1986) *Comparative Biochemistry and Physiology C* 85: 207–214; (36) Marchand-Dumont G and Baguet F (1975) *Pflügers Archives* 354: 87–100; (37) Gies A (1986) *Comparative Biochemistry and Physiology C* 84: 61–66; (38) Matsuura M, Kamura M, and Muneoka Y (1983) *Comparative Biochemistry and Physiology C* 76: 305–311; (39) Richmond JE and Jorgensen EM (1999) *Nature Neuroscience* 2: 791–797; (40) Trailovic SM, Clark CL, Robertson AP, and Martin RJ (2005) *Molecular Biochemistry and Parasitology* 139: 51–64; (41) Huang X, Duran E, Diaz F, et al. (1999) *Molecular Biochemistry and Parasitology* 101: 95–106; (42) Walker RJ, Franks CJ, Pemberton D, et al. (2000) *Acta Biologica Hungarica* 51: 379–394; (43) Wolstenholme AJ and Rogers AT (2005) *Parasitology* 131(supplement): S85–S95; (44) Holden-Dye L and Walker RJ (2006) *International Journal of Parasitology* 36: 395–402.

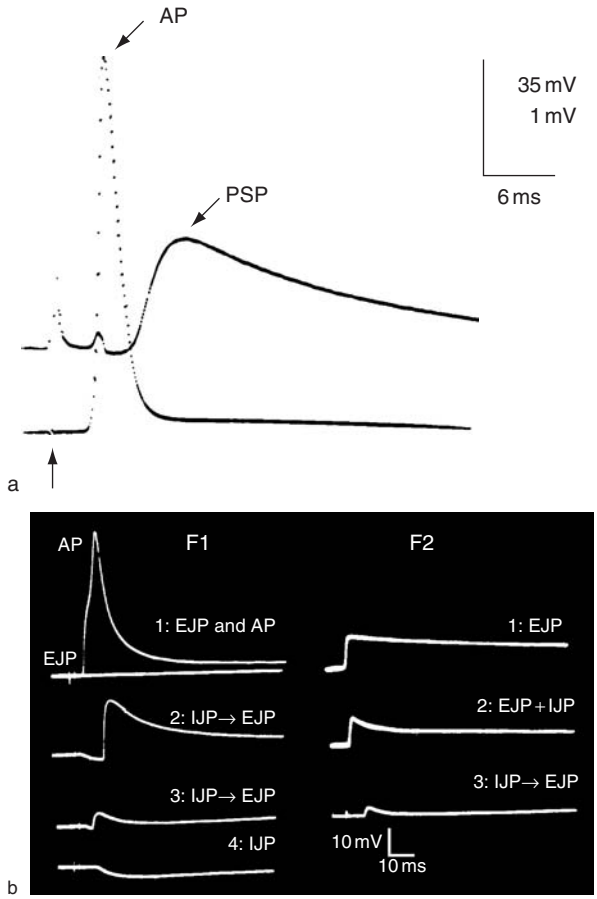

Figure 8 Excitatory and inhibitory electrical responses at crustacean neuromuscular junctions. (a) Simultaneous intracellular records of the presynaptic action potential (AP) and the excitatory junction potential (PSP, postsynaptic potential) from a muscle fiber in the leg opener muscle of the crayfish, *Procambarus clarkii*. Stimulus artifact is indicated by an arrow before the AP's occurrence. (b) Examples of excitatory and inhibitory potentials from two different muscle fibers of the leg stretcher muscle of a crab, *Grapsus grapsus*. In fiber 1 (F1), a single stimulus to the excitor axon (1) produces an excitatory junction potential (EJP), which triggers a muscle action potential (AP). A single stimulus to the stretcher inhibitor axon (4) generates a hyperpolarizing inhibitory junction potential (IJP). Stimulating the inhibitor 8 ms before the excitor leads to disappearance of the AP (2). When the inhibitory stimulus leads the excitatory stimulus by 2 ms, the EJP is drastically reduced, indicating strong presynaptic inhibition (3). In fiber 2 (F2), a large EJP (1) is shortened when the IJP occurs synchronously (2), and is drastically reduced (indicating presynaptic inhibition) when the inhibitory stimulus arrives 2 ms in advance. (a) Recording by Dr. M Wojtowicz, University of Toronto; (b) recordings by Dr. H L Atwood, University of Toronto.

carpopodite joint, three presynaptic receptors may contribute, in varying degrees under differing conditions, to the modulation of transmitter release: an ionotropic GABA receptor (GABA-A type) that activates a chloride channel, a metabotropic GABA receptor (GABA-B type) that activates potassium channels, and a metabotropic glutamate receptor that also activates potassium channels. Metabotropic receptors coupled to G-proteins can modulate release of transmitter directly or indirectly. Both Ca^{2+} and K^{+} channels can be modulated by metabotropic receptors through phosphorylation mediated by second-messenger pathways. Opening of K^{+} channels by metabotropic receptors hyperpolarizes the nerve terminal, altering membrane excitability over a prolonged period of time (several hundred milliseconds or more). Duration of presynaptic inhibition can be influenced by activity of metabotropic receptors, although its major component is attributed to ionotropic GABA-A receptors.

The *Drosophila* genome encodes three putative metabotropic glutamate receptors and three putative metabotropic GABA receptors. One of the former is located in the periactive zones of neuromuscular synapses and is sensitive to glutamate. It regulates activity-dependent synaptic facilitation in addition to sculpting or fine-tuning synaptic morphology.

Presynaptic ryanodine receptors are also involved in regulating transmitter release. In *C. elegans*, loss-of-function mutations of the ryanodine receptor (UNC-68) cause a large reduction in amplitude of excitatory postsynaptic currents and a loss of large-amplitude mini postsynaptic currents. At these synapses, ryanodine receptors likely affect quantal size through presynaptic mechanisms involving regulation of intraterminal calcium stores. The generality of these effects has yet to be determined.

Retrograde Modulation of Synaptic Strength

Although the predominant operation of neuromuscular junctions and other synapses is to convey signals anterogradely from the presynaptic neuron to its postsynaptic targets, there is evidence for retrograde signaling that determines the morphology and response properties of the presynaptic nerve terminals. Two types of evidence have recently been gathered from *Drosophila* NMJs. First, retrograde signaling pathways involving receptors and ligands of the transforming growth factor-β (TGF-β) superfamily of secreted trophic molecules have been implicated as determinants of presynaptic morphology and performance. Second, an acute secretory mechanism involving an isoform of the calcium receptor synaptotagmin, known to be a key partner in calcium-mediated vesicle fusion, has been demonstrated in the postsynaptic region.

Evidence for synaptic retrograde signaling in *Drosophila* was initially derived from observations on effects of manipulating postsynaptic glutamate receptors. Genetic deletion of the GluRIIA subunit, which chronically reduces quantal size (due to reduced postsynaptic receptor sensitivity) leads to an increase in quantal content (number of transmitter-filled vesicles

presynaptically released per nerve impulse), resulting in a postsynaptic recording equal in amplitude to wild type. This indicates a compensatory adjustment of transmitter release in the presynaptic terminal to maintain the functional effectiveness of transmission. Thus, synaptic strength (amplitude of postsynaptic potentials) is influenced by retrograde signaling, which modulates transmitter release to compensate for changes in postsynaptic sensitivity.

Further searches for the mechanism of this effect demonstrated that postsynaptic calcium/calmodulin-dependent protein kinase II (CaMKII) can control retrograde modulation of synaptic strength in *Drosophila*. Chronic increase of postsynaptic CaMKII activity decreases transmitter release, while chronic decrease of postsynaptic CaMKII activity has the opposite effect. Increased quantal content in CaMKII mutants occurs without a decrease in quantal size and is accompanied by an increase in presynaptic T-bars at the active zone.

An additional component of the retrograde signaling pathway came to light with the finding that the complex presynaptic protein Highwire (Hiw), a putative E3 ubiquitin ligase and structural regulator, controls synaptic growth of the neuromuscular junction through negative regulation of effects of bone morphogenetic protein (BMP), a member of the TGF-β superfamily. The presynaptic BMP receptor gene *wishful thinking* (*wit*) is required for the retrograde signal-induced alteration in presynaptic form and function. Thus, a *Drosophila* version of BMP is thought to be a ligand secreted postsynaptically and acting on a presynaptic receptor (which is in turn regulated by Hiw). A presynaptic signal is then received and transduced by the motor neuron's nucleus, leading to modifications of the presynaptic terminal. There is an overlap in mechanisms regulating developmental, structural, and activity-dependent plasticity.

Evidence for a postsynaptic secretory mechanism comes from additional studies of acute alterations in transmitter release, which apparently can occur through retrograde signaling. Postsynaptic calcium influx through glutamate receptors is sensed by an isoform of the calcium-binding protein synaptotagmin (Syt4), triggering vesicle fusion. The contents of the postsynaptic vesicles initiate a large increase of quantal release from the presynaptic terminal (spontaneous calcium-independent vesicle release). In this case, the unknown retrograde signal stimulates the presynaptic cAMP-dependent protein kinase pathway, leading to increased transmitter release. Chronic decreases in many of the proteins involved in this pathway also produce structural alterations to the synaptic areas, suggesting that this pathway may also be involved in longer term changes in synaptic function.

Physiological Features

The numerous distributed synapses provided to a muscle fiber each typically release on average small amounts of neurotransmitter in response to a single presynaptic action potential (sometimes less than 1 quantum per 100–1000 trials in some crustacean tonic axons). In this, they resemble the majority of vertebrate and invertebrate central synapses more closely than NMJs of skeletal muscles in higher vertebrates. The resulting electrical response (postsynaptic potential) of the muscle fiber is often called a junctional potential, either excitatory (EJP) or inhibitory (IJP), according to the type of postsynaptic receptor and ionophore engaged (**Figure 8**). EJPs are capable of eliciting contraction in the responding muscle fiber without provoking a muscle action potential, since the muscle fiber's resting membrane potential is often close to the threshold potential for excitation–contraction coupling. In fact, many invertebrate muscle fibers lack the appropriate mix of ion channels for excitability, and thus do not produce action potentials even when the EJP is large. When a muscle action potential does occur, its amplitude is often variable, depending on the rate of rise and amplitude of the EJP. In contrast, IJPs are usually of small amplitude and act to reduce the depolarization of the EJP, since the IJP's reversal potential is usually close to the muscle fiber's resting potential. The speed and force of the contraction are governed by the rate of rise and amplitude of the electrical events produced by the combined effects of EJPs and IJPs. In addition, the contractile properties of the muscle fiber (which range widely in arthropods and mollusks) play an important role in the ultimate performance of the neuromuscular system.

Representative electrical responses recorded at crustacean (**Figures 8–10**), *Drosophila* (**Figure 11**), and nematode (**Figure 12**) NMJs illustrate typical features and range of invertebrate signaling. In crustaceans, the presynaptic action potential can be recorded intracellularly together with the evoked EJP or IJP (**Figure 8(a)**). The synaptic delay is brief (0.5–1 ms, depending on temperature), and at many junctions the EJP is small at low frequencies of activation. Inhibitory transmission may decrease the amplitude and speed up the decay of the EJP, and eliminate graded muscle action potentials (**Figure 8(b)**). In many limb muscles, drastic attenuation of EJP amplitude indicates presynaptic inhibition mediated by axo-axonal synapses. The *Drosophila* larval preparation shows no sign of peripheral inhibition; two EJPs mediated by different axons can be readily detected (**Figure 11**). In nematodes, EJPs (sometimes accompanied by muscle action potentials) and IJPs (sometimes

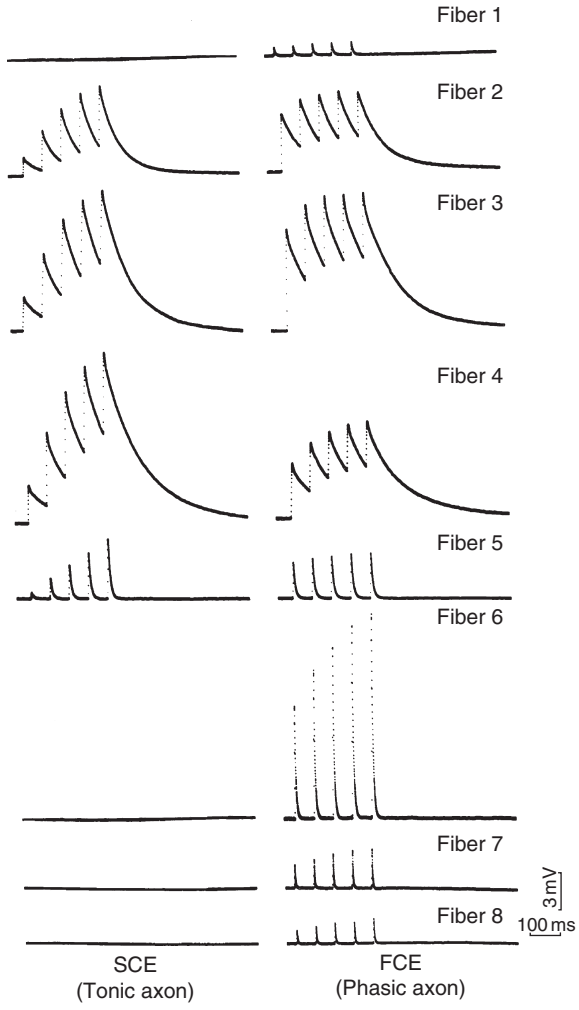

Figure 9 Variation in electrical responses at crustacean neuromuscular junctions. Intracellular records of excitatory junction potentials were taken from eight identified muscle fibers in the closer muscle of the walking leg of the crab *Eriphia spinifrons*. Five stimuli were delivered at a frequency of 16 Hz to either the slow (tonic) excitor (SCE) or the fast (phasic) excitor (FCE). A wide range of excitatory junction potentials, differing in amplitude, time course, and facilitation, was found. Some fibers in this sample (fibers 1, 6, 7, and 8) appeared to be innervated by only one of the two excitors (FCE). From Rathmayer W and Erxleben C (1983) Identified muscle fibers in a crab. I. Characteristics of excitatory and inhibitory neuromuscular transmission. *Journal of Comparative Physiology* 152: 411–420.

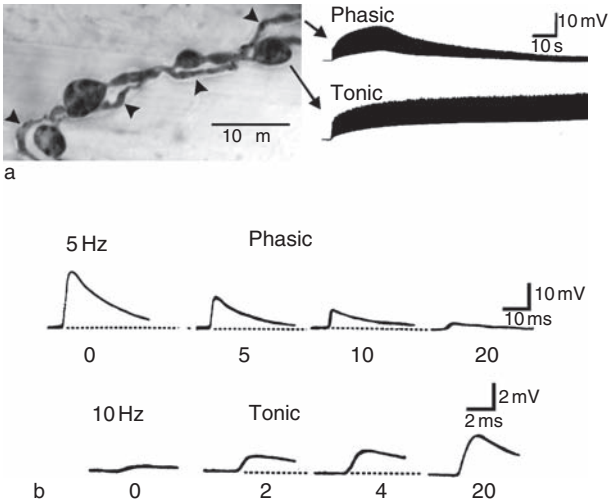

Figure 10 Physiological features of crustacean phasic and tonic neuromuscular transmission. (a) Nerve terminals of phasic and tonic axons supplying the claw closer muscle of the crayfish *Procambarus clarkii* differ in morphology and physiology. Boutons of tonic terminals are considerably larger and generate an excitatory junction potential which is initially small, but facilitates dramatically with 20 Hz stimulation, and does not show depression. Phasic terminals are filiform, and produce an excitatory junction potential which is relatively large, but which shows only moderate transient facilitation at 20 Hz, and soon depresses. (b) Demonstration of inverse relationship between facilitation and depression in phasic axon 3 and tonic axon 6 supplying abdominal extensor muscles of the crayfish, *Procambarus clarkii*. Stimulation at the indicated frequencies was sustained for 20 min, leading to depression and facilitation, respectively, of the phasic and tonic excitatory junction potentials. (a) From Lnenicka GA and Morley EJ (2002) Activity-dependent development and plasticity of crustacean motor terminals. In: Wiese K (ed.) *The Crustacean Nervous System*, pp. 266–282. Berlin: Springer. (b) From Nguyen PV, Marin L, and Atwood HL (1997) Synaptic physiology and mitochondrial function in crayfish tonic and phasic motor neurons. *Journal of Neurophysiology* 78: 281–294.

in combination with EJPs) are elicited by appropriate nerve stimulation (Figure 12). These events are often superimposed on a background of depolarization attributed to tonic release of neurotransmitter.

Synaptic Diversity of Single Neurons

Characteristically, in response to trains of action potentials, both EJPs and IJPs exhibit short-term facilitation and depression, but the manifestations of these processes vary enormously at different terminals of individual neurons. Diversity of terminal physiology is best known in crustaceans (Figure 9). Individual muscle fibers, which differ in physiological and biochemical features, are thought to specify the physiological properties and also the structural features of the innervating terminals. In particular, the probability of transmitter release per synapse and the extent of short-term facilitation and depression are variable features which are characteristically different for terminals innervating specific muscle fibers. Experimental manipulation of muscle fibers to alter their electrical properties leads to compensatory adjustments in physiological responses evoked by the innervating terminals on the affected muscle fiber. This line of evidence indicates a retrograde influence from the muscle fiber to the presynaptic nerve terminal.

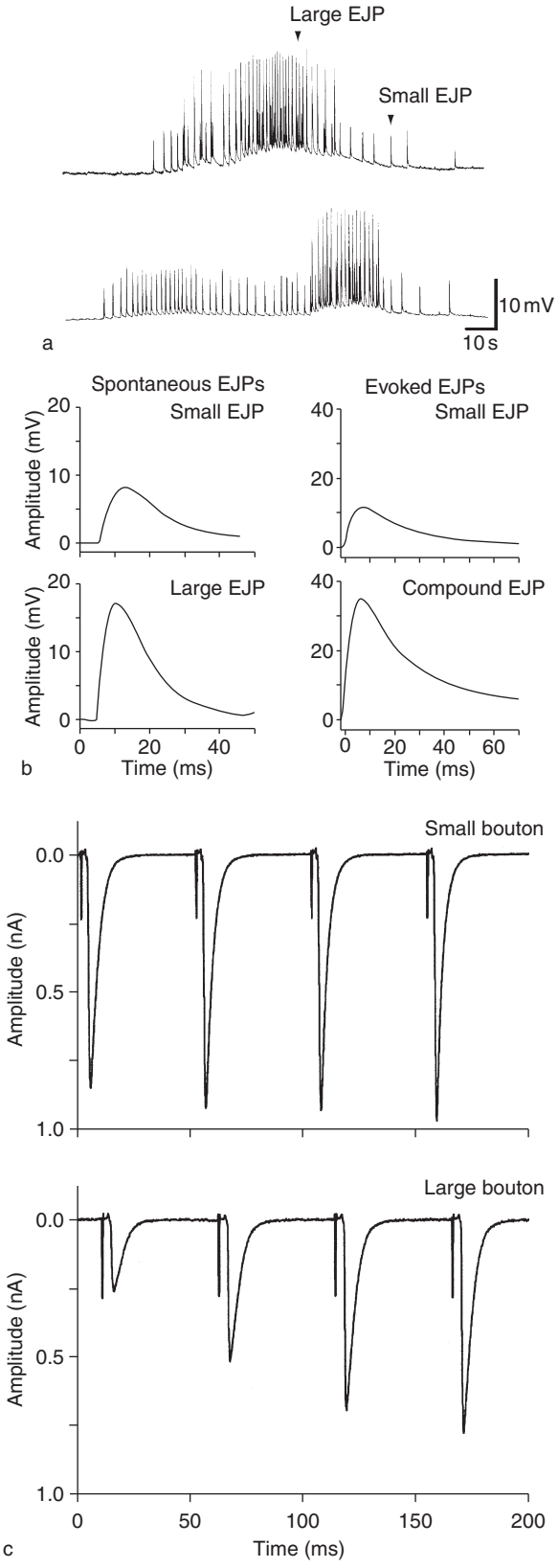

Figure 11 Electrical responses at *Drosophila* neuromuscular junctions. (a) Spontaneously occurring excitatory junction potentials (EJPs) of muscle 6, recorded with an intracellular electrode during fictive locomotion. Two motor neurons are active, producing different patterns of excitatory potentials. Large and small EJPs occur in bursts; bursts of small EJPs are more prolonged. (b) Spontaneously occurring EJPs averaged to illustrate their relative amplitudes. In a different preparation, direct stimulation of the motor axons produces small EJPs and compound EJPs (at higher stimulus strength and recruitment of the second motor axon). (c) Extracellular recording from small and large boutons (type 1s and 1b, respectively) with a 'macropatch' electrode. Recordings were made in HL3 solution (20 mM Mg^{2+}, 1.5 mM Ca^{2+}). Responses of type 1s boutons typically showed little or no facilitation during repeated stimulation; those of type 1b boutons were initially smaller, but often showed short-term facilitation in this physiological solution. From Kurdyak P, Atwood HL, Stewart BA, et al. (1994) Differential physiology and morphology of motor axons to ventral longitudinal muscles in larval *Drosophila*. *Journal of Comparative Neurology* 350: 463–472.

Retrograde signals like those described in *Drosophila* could be partly responsible for muscle cell specification of presynaptic terminal properties. But the general effects on growth and transmitter release ascribed to the BMP pathway may not be enough to account for the wide range of response properties observed in crustaceans (Figure 9). Additional refinements, which may include other retrograde signals, quantitative regulation of signaling effectiveness, and regulation of local protein synthesis in or near the presynaptic terminal, are very likely involved and have yet to be elucidated.

Proximate factors responsible for variation in release properties are in part structural – terminals with higher output have on average more active zones per synapse, and the individual active zones are on average larger – and in part physiological, possibly related to higher impulse-linked calcium entry per synapse in high-output boutons. Biochemical or immunohistochemical observations have not been made to determine molecular differences among boutons.

Synaptic Specialization of Different Motor Neurons

Synaptic differentiation among motor neurons is best known in crustaceans, where large differences between phasic and tonic motor neuron subtypes are strikingly apparent. Very similar neurons are found also in insects. Phasic neurons, normally silent but active for brief episodes of intense activity, produce EJPs of large amplitude which fatigue rapidly on repeated stimulation, whereas tonic neurons, active continuously during normal activity, produce smaller EJPs which often facilitate dramatically and are very resistant to depression during sustained activity (Figure 10). The physiological differences are linked to morphological ones: tonic axons have larger terminal varicosities which contain more mitochondria (Figure 10(a)). These very different types of motor neurons often innervate the same muscle fibers in crustacean limb muscles. Muscles innervated exclusively by either phasic or tonic motor neurons

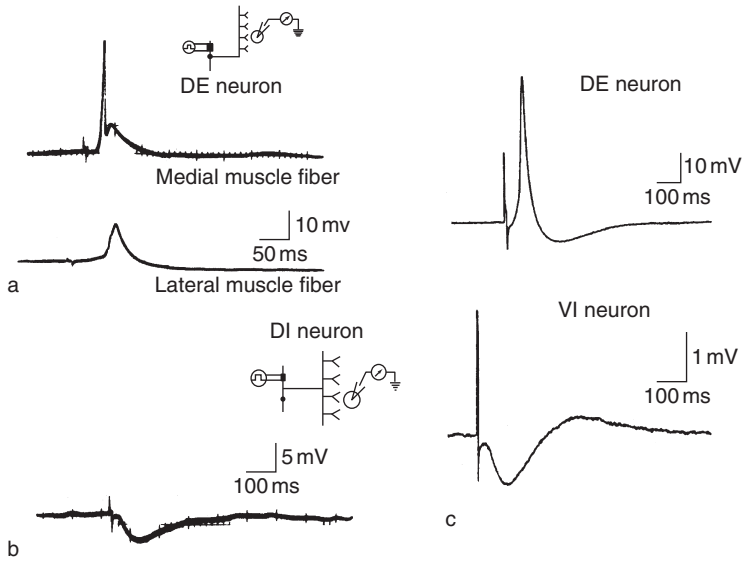

Figure 12 Electrical responses at nematode neuromuscular junctions. Intracellular recordings from muscle fibers of *Ascaris lumbricoides* with stimulation of identified neurons in transverse commissures. (a) Stimulation of dorsal excitatory (DE) neurons produces slow depolarizations, which evoke action potentials in medial muscle fibers. (b) Stimulation of a dorsal inhibitory (DI) neuron generates a hyperpolarizing response in dorsal muscle fibers. (c) Biphasic responses (combined depolarization and hyperpolarization) produced by stimulation of DE and ventral inhibitory (VI) neurons. From Walrond JP, Kass IS, Stretton AO, et al. (1985) Identification of excitatory and inhibitory motoneurons in the nematode *Ascaris* by electrophysiological techniques. *Journal of Neuroscience* 5: 1–8.

are also prevalent; abdominal muscles of crayfish and lobsters provide the best known examples.

At the physiological level, explanations for the phasic–tonic differences have become available. The large differences in synaptic fatigue are linked to a smaller store of docked and releasable synaptic vesicles, and a larger fractional release of available vesicles, in phasic neurons. In addition, smaller and less active mitochondria in these terminals are less effective in maintaining intracellular ion homeostasis, vesicle recycling, and transmitter supply. The larger fractional release in phasic terminals, which ensures a much higher probability of release per synapse, is due to higher Ca^{2+} sensitivity and more rapid kinetics of intracellular Ca^{2+} receptors governing synaptic vesicle exocytosis, rather than to larger Ca^{2+} entry at each synapse during an action potential. This feature was demonstrated in experiments with 'caged' Ca^{2+} injected into the terminals. Photolysis of the caging compound produced about tenfold more release in phasic than in tonic terminals at the same final intracellular Ca^{2+} concentration. In combination with slower Ca^{2+}-binding kinetics of the intracellular Ca^{2+} receptors at the active zone in tonic nerve terminals, this differentiating feature accounts for most of the phasic–tonic phenotypic peculiarities.

Molecular differences that could account for these remarkable features of performance remain sparse. A regulatory Ca^{2+}-binding protein, frequenin, that affects transmitter release in *Drosophila*, is more strongly expressed in crustacean phasic neurons, but it is not clear that this can account for the observed physiological differences. Very likely, many molecular differences are responsible.

In other invertebrates, phasic–tonic differentiation also occurs, though usually less strikingly than in crustaceans. For example, in the *Drosophila* larva, type 1s boutons fire less often during locomotion, have a higher output per bouton, and facilitate less than do type 1b boutons (Figure 11). Compared to type 1b terminals, they are more 'phasic' in their response properties and roles in locomotion. The quantal output per synapse is larger in type 1s boutons, but only by a factor of 2–3, rather than 100–1000 as in crayfish phasic and tonic terminals. Thus, the *Drosophila* larval preparation does not provide a clear-cut model for investigating this aspect of synaptic differentiation.

Conclusion

This brief survey of invertebrate NMJs brings out general features of our current knowledge. Physiological and biophysical findings on details of synaptic transmission, relevant receptors, and ion channels are based largely on experimentally amenable NMJs of a few species of arthropods. On the other hand, modern genetic and molecular information relevant to synaptic structure and function has been built up using small organisms (*Drosophila* and *C. elegans*)

which lend themselves less well to physiological analysis. The *Drosophila* larval preparation, although small, can be used for physiological analyses and there are many recent studies on the effects of various mutations and genetic manipulations on synaptic performance. In *C. elegans*, physiological work is an order of magnitude more difficult, and most of what we know by direct measurement about nematode neurophysiology has been obtained from the much larger *Ascaris*.

Considering known invertebrate systems in total, it is clear that the major invertebrate phyla share a few common features, but differ in detail greatly from each other and from vertebrates. Conserved features in the invertebrates which distinguish their NMJs from those of higher vertebrates are the prevalence of numerous small neuromuscular synapses and active zones, generally widely distributed on the muscle fiber's surface; the common occurrence of polyneuronal innervation, often including peripheral inhibitory neurons; and, frequently, striking exploitation of short-term synaptic plasticity (facilitation and depression). But there are great differences among the major invertebrate phyla. Arthropods use glutamate and GABA, major central neurotransmitters of vertebrates, as their neuromuscular transmitters, and acetylcholine as their major central transmitter; mollusks and nematodes use acetylcholine as their major peripheral excitatory transmitter. Arthropods and mollusks innervate their muscles with branches of motor axons that conduct action potentials to the neuromuscular junction; nematodes have motor neurons that do not conduct action potentials, release transmitter tonically, and synapse with muscle arms that travel directly to the longitudinal nerve cords. A combination of these styles of innervation is found in onychophorans.

Several solutions to the problem of obtaining maximum range of motor activity from a limited set of neurons and muscle fibers have been developed in arthropods and mollusks. Synaptic differentiation at different terminals of a single motor axon permits variable recruitment of muscle fibers in the innervation field according to functional demand. In contrast, synaptic differentiation of specialized phasic (fast) and tonic (slow) neurons, often innervating the same muscle fibers, has evolved as a method of generating different modes of locomotion. These neurons are fundamentally different in their physiology. While retrograde signaling from the muscle fiber may account for differentiation of terminals associated with a single neuron, it cannot explain the differences between phasic and tonic neurons innervating the same muscle fiber. Further analysis will be needed to uncover the root causes of phasic–tonic presynaptic neuromuscular differentiation.

See also: Calcium Channels; Development of Drosophila Neuromuscular Junctions; Neuromodulation of Calcium Channels; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): Aging; Presynaptic Events in Neuromuscular Transmission; Voltage-Gated Calcium Channels.

Further Reading

Atwood HL (2006) Gatekeeper at the synapse. *Science* 312: 1008–1009.

Atwood HL and Govind CK (1990) Activity-dependent and age-dependent recruitment and regulation of synapses in identified crustacean neurones. *Journal of Experimental Biology* 153: 105–107.

Atwood HL and Karunanithi S (2002) Diversification of synaptic strength: Presynaptic elements. *Nature Reviews Neuroscience* 3: 497–516.

Atwood HL and Wojtowicz JM (1986) Short-term and long-term plasticity and physiological differentiation of crustacean motor synapses. *International Review of Neurobiology* 28: 275–362.

Budnik V and Gramates LS (eds.) (1999) *Neuromuscular Junctions in Drosophila*. San Diego: Academic Press.

Budnik V and Ruiz-Canada C (eds.) (2006) *The Fly Neuromuscular Junction: Structure and Function*. San Diego: Academic Press.

Davis RE and Stretton AOW (1989) Signaling properties of *Ascaris* motorneurons: Graded active responses, graded synaptic transmission, and tonic transmitter release. *Journal of Neuroscience* 9: 415–425.

DiAntonio A, Petersen SA, Heckmann M, et al. (1999) Glutamate receptor expression regulates quantal size and quantal content at the *Drosophila* neuromuscular junction. *Journal of Neuroscience* 19: 3023–3032.

Dudel J, Franke C, and Hatt H (1990) A family of glutamatergic, excitatory channel types at the crayfish neuromuscular junction. *Journal of Comparative Physiology A* 166: 757–768.

Hoang B and Chiba A (2001) Single-cell analysis of *Drosophila* larval neuromuscular synapses. *Developmental Biology* 229: 55–70.

Hoyle G (1983) *Muscles and Their Neural Control*. New York: Wiley.

Kurdyak P, Atwood HL, Atwood, and Stewart BA (1994) Differential physiology and morphology of motor axons to ventral longitudinal muscles in larv *Drosophila*. *Journal of Comparative Neurology* 350: 463–472.

Nguyen PV, Marin L, and Atwood HL (1997) Synaptic physiology and mitochondrial function in crayfish tonic and phasic motor neurons. *Journal of Neurophysiology* 78: 281–294.

Osborne RH (1996) Insect neurotransmission: Neurotransmitters and their receptors. *Pharmacology and Therapeutics* 69: 117–142.

Pearce J, Govind CK, and Shivers RR (1986) Intramembranous organization of lobster excitatory neuromuscular synapses. *Journal of Neurocytology* 15: 241–252.

Rathmayer W and Erxleben C (1983) Identified muscle fibers in a crab. I. Characteristics of excitatory and inhibitory neuromuscular transmission. *Journal of Comparative Physiology* 152: 411–420.

Satterlie RA (1993) Neuromuscular organization in the swimming system of the pteropod mollusk *Clione limacine*. *Journal of Experimental Biology* 181: 119–140.

Usherwood PNR (1977) Neuromuscular transmission in insects. In: Hoyle G (ed.) *Identified Neurons and Behavior of Arthropods*, pp. 31–48. New York and London: Plenum.

Walrond JP, Kass IS, Atwood, Stretton AO, et al. (1985) Identification of excitatory and inhibitory motoneurons in the nematode *Ascaris* by electrophysiological techniques. *Journal of Neuroscience* 5: 1–8.

Weimer RM, Gracheva EO, Meyrignac O, Miller KG, Richmond JE, and Bessereau JL (2006) UNC-13 and UNC-10/rim localize synaptic vesicles to specific membrane domains. *Journal of Neuroscience* 26: 8040–8047.

White JG, Southgate E, Thomson JN, et al. (1976) The structure of the ventral nerve cord of *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 275: 327–348.

White JG, Southgate E, Thomson JN, et al. (1976) The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 314: 1–340.

Wiens TJ (1993) The closer muscle is a second target for the stretcher inhibitor motoneuron of the crayfish's thoracic limbs. *Journal of Comparative Physiology A* 173: 435–444.

Wiese K (ed.) (2002) *The Crustacean Nervous System*. Berlin: Springer.

Zhai RG and Bellen HJ (2004) The architecture of the active zone in the presynaptic nerve terminal. *Physiology* 19: 262–270.

Comparative Neurobiology: History

R L Reep, University of Florida, Gainesville, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Three interrelated conceptual threads run through the history of comparative neurobiology: understanding the relationship between nervous system activity and behavior, reconstructing the evolution of the nervous system, and uncovering the mechanisms that generate the patterns and transformations seen during development of the nervous system. These three themes were regular points of departure during the nineteenth and early twentieth centuries, but technical limitations often resulted in long periods of little or no progress beyond certain limits. Since the 1930s, the development of ever more refined techniques for electrophysiological study of the nervous system has led to detailed analyses of the patterns of neuronal activity that underlie sensory and motor functions and behavior, and has provided maps of the localization of these processes in the brains of vertebrates. The development of sensitive axonal tracing techniques in the 1960s, together with the formulation of systematic methods by which to compare phenotypic characters, led to combination of the two approaches in efforts to understand homologies among vertebrate nervous systems. Since the 1980s, the explosion of molecular data on gene expression has begun to provide a framework by which to understand themes and variations in development and evolution, including that of the nervous system. In invertebrates, the combination of electrophysiology, characterization of the architecture of single neurons, and analysis of gene expression patterns is enabling dynamic descriptions of the operation of small neural networks having defined functions.

Origins

Recognition of the brain as the organ which is responsible for organizing sensation, movement, cognition, emotions, and behavior has many origins in antiquity, but it became firmly established in the nineteenth century. Along the way came the recognition that brains in animals perform many of these same functions. In the third century AD, Galen performed dissections on the nervous systems of many mammals, described the sympathetic chain, distinguished between sensory and motor nerves, and concluded that cerebral convolutions could not be the seat of intelligence, since they were well developed in donkeys. He also focused on the ventricles as a storage place for animal spirits, and thus the seat of the soul. Based upon wounds suffered by soldiers, Galen proposed that the brain substance was the site of intellect. Fascination with the role of the ventricles in higher functions persisted into the Renaissance and beyond. Leonardo da Vinci performed wax casts of the ventricles in cattle. Vesalius commented that animals had ventricles that were similar to those in humans. Descartes chose the pineal as the seat of the soul, partly because it was surrounded by cerebrospinal fluid that he also believed to contain the animal spirits. In the seventeenth century, Thomas Willis developed the first modern theory of functional localization when he ascribed different functions to the cerebral cortex and corpus striatum. In the eighteenth century, it was discovered that the medulla contained centers for cardiovascular and respiratory functions, through experiments on dogs and rabbits. Studies on dogs also showed that dorsal spinal roots conveyed sensory signals whereas ventral roots were motor. One of the most important findings of the nineteenth century was the localization of the dog motor cortex by Fritsch and Hitzig in 1870.

The history of comparative neurobiology is inextricably linked to several major trends in biological thought, revolving around comparative anatomy, embryology, and evolution. In the early nineteenth century, Cuvier and others who embraced Naturphilosophie conceived of species as being immutable, each one being the perfect design product of a creator who had particular functional needs in mind. Interestingly, Cuvier gave priority to the nervous system in establishing his system of animal taxonomy, arguing that all other systems were subordinate to it. In contrast, his contemporary and former friend, Étienne Geoffroy Saint-Hilaire, emphasized that function follows form, and that final causes cannot serve as explanatory principles. Geoffroy suggested that philosophically there is only a single animal. He emphasized the existence of a unity of plan for vertebrates, and thought that archetypal structures could be modified in many ways and that each of these allows a certain range of functional capacity (this is an early example of the concept of developmental constraint). However, these modifications were not conceived to occur within the framework of descent with modification later articulated by Darwin. Homology (Geoffroy's analogy) was properly recognized by the principle of connections – relative position and spatial relationships rather than simply shape. This presages the use of relative position of nuclei and their patterns of neuronal connections as

modern criteria for homology. Geoffroy Saint-Hilaire's proposal to extend the vertebrate archetype to arthropods was generally viewed skeptically and contains many fanciful, erroneous elements. However, since the 1980s, molecular biological studies have revealed deeply conserved genes and gene expression patterns across phyla, including segmental and dorsoventral patterning. Now it is commonplace to expect to find ever more intricate similarities in developmental mechanisms, because the acceptance of a common tree of life is firmly embedded in biological thinking.

Biologists of Geoffroy Saint-Hilaire's time knew nothing about genes, yet they realized that some as yet undiscovered portion of the cell was directing transformations such as differentiation. It then became possible to imagine realistic (but then unknown) scenarios by which transformations had created the extensive diversity of biological forms coming to light through nineteenth century study of fossils and extant organisms. Darwin's revolutionary synthesis firmly established descent from common ancestry as the guiding principle in studies of relatedness of form, rather than Cuvier's functional morphology or Geoffroy's unity of plan. Darwin also emphasized that a true theory of inheritance must include a mode of transmission and a mode of development. He emphasized the modular nature of individual body parts, which evolved with a degree of independence, thus glimpsing one of the current preoccupations of evolutionary developmental biology.

Neuroanatomy in an Evolutionary Context

Comparative neurobiology began to flourish in the latter half of the nineteenth century, at a time when naturalists traveled the world collecting specimens of previously undescribed life forms, fossil discoveries were providing an ever expanding record of diversity in the past, geologists were finding evidence that the Earth was much older than had been thought, and Darwin's theory of evolution was being expounded and digested. The practice of sectioning brains and staining sections began in the mid-nineteenth century, shortly after the exposition of the cell theory by Schleiden and Schwann in 1837. Degeneration methods began in 1852 with Waller. Weigert developed myelin staining in 1881, around the same time that Nissl developed a method for staining cell bodies and Cajal began his intense study of neurons using the Golgi stain. Cajal championed the view that the nervous system is composed of separate cells (neurons) rather than a continuous reticulum as envisioned by Golgi. Toward the close of the century, Sherrington formalized the concept of the synapse. Around this time, several centers devoted to comparative neuroanatomy were being established in Europe and North America. Even Sigmund Freud joined the fray, investigating crayfish neurons.

Comparative neurobiology naturally embraced an evolutionary context for its discoveries. However, in spite of Darwin's emphasis on the use of actual genealogies to define taxonomic relatedness, another dominant tendency of the day was adopted, the ancient belief in scala naturae – a phylogenetic scale of nature that viewed evolution as progressive and linear, with increases in complexity and intelligence from fish to amphibians, reptiles, birds, and mammals, culminating with humans. This viewpoint was to dominate comparative neurobiology until the 1980s. For Cajal there was a discernible increasing complexity in behavior and elaboration of nervous system elements as one progressed from unicellular organisms, which lack a nervous system and exhibit rudimentary responses to stimuli, through coelenterates, with sensory and motor neurons, to worms, in which the primitive scheme of sensory and motor neurons only is embellished by the appearance of the interneuron or association neuron. Cajal even stated that with the advent of the association neuron, multicellular organisms became true animals. This qualitative leap was, of course, due to the integrative richness made possible by the interposition of interneurons between sensory and motor neurons. Although we now know that the situation is not so simple (e.g., coelenterates have integrative capacities), these first glimpses of the neural machinery, which undoubtedly was at the heart of generating complex sensory impressions and coordinated movement expressed as behavior, sparked a new era in comparative brain studies that continues to the present. As important as association neurons were, Cajal felt that an even greater step was taken when the psychomotor neuron appeared. These, too, were association neurons but were farther removed from sensory and motor elements, and were organized into concentrated masses, such as the cerebral ganglia of arthropods and the cerebral hemispheres of vertebrates.

Cajal enumerated several ways in which the general organizational structure of vertebrate nervous systems resembled the ganglionic nervous systems of invertebrates: axons of spinal cord motor neurons innervate striated muscles, primary sensory neurons originate outside the central nervous system (CNS) and terminate in the spinal cord, second-order association neurons lie within the cord, higher order neurons transmit neural activity to the cerebral cortex, and corticospinal neurons run the length of the CNS. Cajal envisioned the transformation from the invertebrate to the vertebrate plan to be driven by proliferation of neurons and their processes, elaboration of

neural morphology to increase capacity for information processing, and progressive unification of neural tissue into a continuous mass to increase conduction speed and conserve neuronal cytoplasm.

In 1906, JB Johnston viewed gross features of the vertebrate brain and patterns of neuronal connections as having evolved through adaptation to particular modes of life, and by embryonic development, whereby important roles are played by special structures adapted to the embryonic mode of life and inherited structures adapted to ancestral modes of life. Similarly, Cajal emphasized the limited range of stimuli provided to the CNS by the sense organs, and this limitation led him to speak of a substance of which we would remain forever ignorant, whose essence would inevitably escape our understanding. Johnston and Cajal both viewed cerebral cortex as responsible for organizing the most complex functions.

One aspect of scala naturae thinking was the concept that some brain traits were older than others. For Edinger, Herrick, Ariëns Kappers, and others, olfaction seemed to be the most primitive sense; therefore, the olfactory areas of the brain were viewed as the most ancient. The telencephalon in Agnathans was believed to receive only olfactory projections, hence it was said to consist of a paleopallium and paleostriatum. Fish were believed to have evolved higher olfactory centers, represented by the hippocampus or archipallium. Thalamic projections to the striatum were thought to have originated in amphibians, whereas those to the neopallium were believed to have appeared first in reptiles. With mammals the neopallium and neostriatum became predominant. Thus in mammals olfactory cortex was named paleocortex, hippocampus was archicortex, and the six-layered isocortex became neocortex. (Ironically, according to ancient usage, the prefix 'archi-' should have been used for the most ancient region.)

The application of purely architectonic methods, together with names derived from the scala naturae mindset, led to erroneous designations that only became clarified much later, with the advent of histochemistry, axon tracing techniques, and methods to trace developmental lineages. Thus, histochemical staining for acetylcholinesterase revealed that the avian striatum comprised a small ventral region, compared to the voluminous territory that had been designated paleostriatum, archistriatum, neostriatum, and hyperstriatum. An overhaul of this terminology has been made recently for avian brains, and it also includes corrections for telencephalic pallial regions erroneously designated as belonging to the striatum. Another case involves ray-finned fishes, in which the telencephalon exhibits eversion during development. This produces reversed locations of the medial and lateral pallium, compared to other fishes. Proper interpretation of axonal labeling patterns was required to correctly identify these regions.

The architectural parcellations of human brains by Brodmann and the Vogts were mirrored by similarly detailed studies of a wide variety of other animals. These descriptive anatomical studies culminated in the 1930s with the publication of the Kappers, Huber, and Crosby volumes. Anatomical comparative neurobiology had temporarily run its course until revived in the 1960s by the advent of more refined techniques for tracing axonal pathways. However, these early descriptive anatomical studies set the stage for a profusion of electrophysiological studies that were soon to come.

Functional Mapping

The tradition of functional localization began with the discovery of excitable motor cortex in dogs by Fritsch and Hitzig in 1870. In the twentieth century, study of functional localization was ignited by the development of microelectrode recordings in anesthetized animals. Woolsey's group set out to map sensory and motor areas of the cerebral cortex in a wide variety of mammals, as a way of gaining insight into the neural events that underlay known variations in behavior and anatomy. This tradition has resulted in many fundamental findings, including detailed studies of topographic maps within the visual, auditory, somatosensory, and motor systems, including subcortical areas of the brain stem and thalamus. A key insight has been the presence of expanded representations in the cortex that reflect specializations of the sensory periphery. Functional mapping has also revealed cortical fields that are common to all mammalian species. These include primary and secondary sensory areas for vision, audition, and somatic sensation, and primary motor cortex. This work provides a basis for hypotheses concerning the elaboration of new cortical areas (e.g., multiple visual areas in primates), and for defining the developmental processes by which they originate.

Specializations are important because they represent the boundaries of known evolutionary potential, and because they represent the neural underpinnings of specialized behaviors. The sensory periphery exerts a powerful influence on the development and organization of the central regions that process this information. This is vividly illustrated in the range of variation in sensory specializations exhibited among vertebrates. These include electroreceptors, which appear to have evolved several times; eyes, including four-eyed fish and reductions in blind cave fish; hearing, including echolocation in bats and odontocete cetaceans; tactile

specializations, including body-wide distribution of sensory hairs in naked mole rats, sirenians, and hyraxes; lateral line detectors in fishes and amphibians; and dense mechanoreceptors on catfish barbels, alligator domes, the platypus bill, and the nasal appendages of the star-nosed mole. Among the many striking central elaborations associated with sensory systems are the vagal lobes of certain teleosts. These are voluminous spiral outgrowths of the visceral sensory column in the medulla which process gustatory information from similarly spiraled epibranchial organs located in the oral cavity. Catfish have enlarged medullary facial lobes which process gustatory information originating from the perioral barbels. Electroreceptive fish exhibit an enlarged region of the medulla which receives inputs from an electrosensory lateral line. Mormyrid electric fish possess a hugely enlarged region of the rostral cerebellum. Mammalian examples of central specializations associated with sensory input include neuronal aggregations in the brain stem, thalamus, and cerebral cortex for vibrissal input; multiple visual areas in primate cerebral cortex; expanded central representations in the platypus for tactile receptors and electroreceptors that reside on the bill; expanded cerebral cortical representations in star-nosed moles to accommodate tactile input from elaborate nasal appendages; expanded representations of the forepaw digits in raccoon cerebral cortex; and expanded auditory processing regions in echolocating bats and cetaceans. The cortical neuron aggregates known as barrels, which are associated with vibrissal input, are curious because some species exhibit solid aggregations whereas others have hollow barrels. The developmental determinants of these two types of morphology are not well understood, but appear to involve factors including the distribution of thalamocortical afferents. Understanding how such factors influence the development of particular adult morphologies will illuminate the basis of species-typical neural variation.

Simpler Systems

At the mid twentieth century, Herrick noted that much of the early research in neuroanatomy was motivated by the idea that the simpler brains of nonmammalian vertebrates present fundamental features of the human brain, without the complexities seen in humans and other mammals. Herrick's approach was to study the brain of the tiger salamander, chosen due to its simple behaviors, rapid development, and small brain. His primary interest was the study of the brain as the organ which regulates behavior.

Herrick's approach – to use a simpler vertebrate to understand the human brain – was taken a step further by Wiersma and others in the 1950s, who argued that general principles of the neural control of behavior could be derived from studying and modeling invertebrate systems. This view, that particular systems could reveal general principles of neural function, had precedent in the studies of axon conduction by Hodgkin and Huxley in the 1930s, using the squid giant axon, and those of neurotransmission by Katz and Miledi using the frog neuromuscular junction. Wiersma discovered giant interneurons in the crayfish that are constant in location across individuals, have characteristic firing properties, and produce characteristic movements upon stimulation (thus presaging the later discovery of many other identified neurons). Command neurons have been found to initiate, gate, or terminate patterned neural activity generated by pacemaker networks intrinsic to specific ganglia, in a variety of invertebrate systems, including crayfish swimmerets, insect walking, *Aplysia* feeding, and cricket stridulation. The same command neuron may trigger different movement sequences, depending upon environmental variables. Another striking example of behavioral organization in an invertebrate was Roeder's discovery that mating behavior in the male praying mantis was programmed in the thoracic and abdominal ganglia. This patterned behavior was released upon disconnection from the brain (either experimentally or by having the head bitten off by a female), apparently through cessation of tonic descending inhibition.

The development of microelectrode recording techniques had a major impact on the long-standing goal of understanding the relationship between brain activity and behavior. On the one hand, it was now possible to examine the firing patterns of single neurons. On the other hand, there were vast numbers of neurons in vertebrate brains. The complexity of studying vertebrate systems presented three approaches to the problem, outlined by Don Maynard: (1) shift the focus from single neurons to single brain regions, each consisting of many neurons; (2) examine a few single neurons in great detail, in the hope that general principles will emerge; and (3) explore simpler systems in great detail. In the 1970s, the third strategy led to an intensive focus on invertebrates, which presented the following advantages: fewer nerve cells arranged in discrete aggregations (ganglia), large cell bodies superficially located and thus accessible for recordings, and interesting behaviors. Furthermore, significant aspects of behavior and plasticity could be studied in isolated preparations, such as one thoracic ganglion and its associated legs or wings, the lobster stomatogastric ganglion, or *Aplysia* gill withdrawal. Those in this new field of neuroethology sought to understand the relationship between patterns of activity in single neurons and behavior. A major

advance in invertebrate neuroethology was the discovery that invertebrates possess identified neurons – neurons that have the same cell body location, neurite branching pattern, and basic physiological properties across individuals. Similarly, within an individual there are serially homologous neurons, wherein identified neurons in different ganglia have similar locations and morphologies (but may have different functions). In 1965, Bullock and Horridge summarized the voluminous literature on invertebrate nervous systems. Interestingly, this was the same year that Cohen and Jacklet reported the existence of identified neurons in insects. With the increasing use of microelectrodes to perform intracellular recording and staining, it became possible to correlate cell body location, neurite branching pattern, and firing properties. As a harbinger of the complexity present in even simpler systems, it became apparent that single invertebrate neurons may have extensive neuritic fields, often with several functional compartments that are electrotonically isolated from each other. This raised important issues, still unanswered, concerning the degree to which single invertebrate neurons may sometimes function as several neurons in one. Thus, although invertebrates have many fewer neurons than do vertebrates, the functional properties of their individual neurons may sometimes be greater in complexity.

The discovery of identified neurons and the development of intracellular recording and staining techniques led to a profusion of fruitful studies of the neural networks mediating behavior, mostly in arthropods and mollusks. Recording from several identified neurons simultaneously allows one to investigate network properties in great detail. For specific behavioral sequences, especially those exhibiting repetitive movements, it was found that the pattern of motor neuron output is closely associated with the pattern of movement (e.g., repeating sequences of flexion and extension in limb-based locomotion). Fundamental features of these patterns were often found to persist in isolated preparations lacking sensory input or connections of motor neurons to the muscles they normally innervate. Thus, centrally generated patterns could be studied independent of the actual behavior, and free from the modifying influence of sensory input. By 1966, such isolated systems included cicada song, crustacean heart rhythm, insect flight, crayfish swimmeret beating, insect respiration, insect copulation, and cricket song. There was a general feeling of optimism that the relatively simple nature of the neural components, together with the freedom to study them independent of sensory or motor output, held the promise of understanding the underlying basis for central pattern generation. This view was held not only with respect to the neuroethology of particular systems of interest, but with the aim of deriving general principles of information flow in neural networks. Throughout the 1970s–90s, this resulted in studies of the neural networks mediating behaviors such as walking, flying, escape, stridulation, and digestion. It became clear that the membrane properties of neurons play a key role in pattern generation. Pacemakers that drive the intrinsic rhythm can consist of one or a few neurons with intrinsic bursting oscillatory firing patterns that are either endogenous or are triggered by extraneous factors such as neuromodulators or descending input, as in the crustacean stomatogastric ganglion. More frequently, the rhythm is generated synaptically by a distributed group of interconnected neurons that collectively exhibit oscillatory firing, even though none of the units alone does so, as in locust flight. Reciprocal inhibition, plateau potentials, adaptation, and postinhibitory rebound are other features of these neuron networks that shape the output of central pattern generators (CPGs) in complex ways. However, the details of central integration that produce and modulate such output have proved daunting to unravel, even in the case of rhythmic behaviors. For example, it has proved difficult to specify the cellular basis for observed properties of the frequency and phase of motor output patterns. Intersegmental coordination of movement patterns is mediated by sensory feedback in some cases, and by central interneurons in others. It has also become apparent that groups of neurons may participate in more than one CPG, switching from one to the other under the influence of factors such as neuromodulators. Thus, one outcome of these findings is the striking complexity present in what were initially thought to be simple systems.

The organization of behavior into coherent modules (e.g., feeding, mating, escape) was noted by early neurobiologists, including Charles Sherrington. Later neuroethological studies, particularly those carried out in mollusks, established the neural bases for producing the singleness of action noted by Sherrington. In some cases hormones act as neuromodulatory signals that inhibit certain behaviors during the performance of others. Suppression of one behavior can also occur via synaptically mediated signaling that occurs during another behavior. These mechanisms produce a network of behavioral hierarchies, demonstrating plasticity under certain conditions and thus representing a neural substrate for interpreting concepts such as drive and motivation. In addition to their usefulness for neuroethology, invertebrates have also provided insights into the cellular mechanisms of nerve conduction, synaptic transmission, and synaptic plasticity. In 2000, Eric Kandel was awarded the

Nobel Prize in Physiology or Medicine for his group's work on synaptic plasticity involved in learning and memory in *Aplysia*.

Known vertebrate central pattern generators include those that generate characteristic output for respiration, locomotion, chewing, swallowing, coughing, crying, laughing, sneezing, fear responses, saccadic eye movements, and the hippocampal gamma and theta rhythms. As with invertebrate CPGs, these networks are viewed as providing a basic output pattern that is then modified by a variety of local, ascending, and descending inputs. Swimming in fish, lampreys, tadpoles, and salamanders involves segmental CPGs, on the right and left sides of the spinal cord, that exhibit reciprocal inhibition. Spinal CPGs are controlled by descending input from the brain stem. The basal ganglia play a major role in selecting which CPGs are activated by release from tonic inhibition, and dopaminergic influences are an important factor in all vertebrates. Mammalian limb-based locomotion appears to involve spinal CPGs for each portion of the limb (hip, knee, ankle, foot), and more proximal CPGs entrain those located more distally.

Among vertebrates, one of the best studied examples of neuroethology is birdsong. The modular nature of birdsong has been fundamental to success in analyzing the influence of environmental manipulations, brain lesions, and hormones on song learning, song production, and song perception. This rich subject includes detailed studies of the brain circuitry involved in these functions, comparisons among many species, and investigations of sexual dimorphism, development, and seasonal changes. Another well-studied system is echolocation in bats. The neural representations of specific components of ultrasonic-frequency modulated signals and constant-frequency signals in the inferior colliculus and auditory cortex have been precisely correlated with their behavioral significance. These findings have revealed species-typical neural specializations associated with different signaling patterns and environments, and have also demonstrated specializations at the level of individuals in some species. The coevolution of some bats and their ultrasound-detecting moth prey is a fascinating example of ongoing interspecific neurobehavioral competition, and this has a counterpart in the similar coevolution of echolocating odontocete cetaceans and their clupeid fish prey.

Brain Size

One of the simplest comparisons to make among vertebrates is that of brain size. Comparisons among humans have a long tradition including divining (often in a *post hoc* manner) the moral and intellectual merits of famous individuals based upon the size of their brain. In the mid-nineteenth century this was added to by the burgeoning popularity of phrenology, examining the external bumps of the skull as a window into the soul, due to their close proximity to the brain. Although these speculations may seem laughable now, they illustrate the magnitude of importance that had come to be placed upon the brain as the organ most centrally responsible for human 'identity.' Close behind considerations of brain size have been speculations concerning the significance of gyral patterns in the cerebral cortex. These have been with us since Galen, and until recently have centered on an imagined correspondence between gyral complexity and intelligence. Theories of the past 30 years have focused instead on mechanistic models that take into account the following observations: (1) gyral patterns are lineage-specific; (2) within a lineage, larger brains tend to be more gyrated; (3) gyral crowns exhibit thicker gray matter and greater axon density; and (4) sulcal boundaries between adjacent gyri are often also the site of functional boundaries. Thus, physical features involving relative thickness of gray matter versus underlying white matter, mechanical tension inherent in axon bundles, and sequences of developmental events are given primacy.

Harry Jerison championed the use of relative rather than absolute brain size, reasoning that because larger animals require larger brains, one should compare the excess brain volume present in order to get at issues of intelligence. Sophisticated quantitative procedures have been brought to bear, and there has been an enduring desire to find the basis of intelligence in its many manifestations, represented in brain anatomy. Thus, in a fundamental sense the twentieth century school of thought championed most vigorously by Jerison has its roots in the nineteenth century craniometry of Broca and others. This has raised countless questions within humans alone, not to mention the question of comparing measures of intelligence across taxa. In the late 1970s, the focus shifted from intelligence to more readily defined biological variables. Thus, relative brain size in bats was found to correlate with feeding strategy. Manatees appear to have relatively small brains because dietary and energetic factors drive selection for large body size, perhaps independently of large brain size. Large dinosaurs appear to have small brains simply because they follow the general negative allometric trend for reptiles. The brain case of *Archaeopteryx* suggests that its brain exhibited some traits associated with

flight, characteristic of modern birds. Primates may have relatively large brains due in part to energetic trade-offs between gut and brain.

Total brain size is likely to be a crude measure of intelligence or any other functional variable. Increased understanding of functional localization has resulted in attempts to use allometric analysis to examine the relative enlargement or diminishment of brain regions rather than whole brains. Thus, the ratio of superior to inferior colliculus volume provides a rough index of visual versus auditory dominance, and correlates surprisingly well with known behavioral specializations (e.g., visually dominant frugivorous bats vs. echolocating aerial insectivorous bats). One of the problems with this approach is that most volumetric data exist for major brain divisions rather than for the specific nuclei that comprise functional systems. Nevertheless, recent findings by Barbara Finlay and colleagues have shown significant developmental linkages between the sizes of brain regions in mammals. Because regions differ with respect to important factors like duration of neurogenesis, they scale with each other in a nonlinear fashion. Thus, for example, larger brains will exhibit disproportionately large cerebral hemispheres. Models of brain scaling based upon schedules of neurogenesis explain much of the observed variation in adult brain composition across taxa. An important conclusion from this work is that developmental plasticity allows for specific regions to increase in size, but this is balanced by constraints that produce compensatory changes in other regions. These constraints may be related to rhombomeric and prosomeric compartmentalization of the brain during development.

Evolutionary Developmental Neurobiology

Theories of brain evolution that were proposed prior to the flood of new information on neural connections tended to be based on a scala naturae view of evolution. Phrases such as the 'phylogenetic scale' and terms like 'higher' and 'lower' were regularly used. Paul McLean's triune brain theory posited that the mammalian brain consists of a core that is dominant in reptiles, an intermediate limbic system, and a neopallial cerebral cortex. The popularized notion that there exists a reptilian brain inside each human brain has proved to be a powerful metaphor, but this cultural cooptation has distorted the theory, so that its biological essence has been overshadowed. Thus, in discussions of the triune brain theory, humans often stand in for mammals generally, ignoring the astounding range of perceptual and behavioral diversity among mammals. McLean's theory is in essence a combination of scala naturae thinking with functional localization that focuses on three collections of regions within the brain, ascribes functional properties to each of these three collections, and then uses this as a framework by which to compare taxa. However, we now know that brain regions and collections of regions do not evolve independently. Rather, functional circuits involve multiple brain regions and portions of each of the three systems recognized in the triune brain theory. The scala naturae viewpoint persisted until the 1980s, as neurobiologists were slow to realize the significance of the modern evolutionary synthesis, with its implications that (1) there is no chain of being – rather, evolutionary changes occur independently in each radiations and (2) there are no primitive extant animals – rather, living forms display a mixture of primitive and derived phenotypic traits.

Like other soft tissue characters, most brain traits do not fossilize. Therefore, they can be compared only among extant taxa. Thus, reconstructing brain evolution has been analogized to finding fruit on the ground beneath a tree and trying to reconstruct which fruit fell from which branch. This reality necessitates use of the comparative method, whereby one gathers information about the status of a suite of characters (in this case, neural) in a number of living taxa, then infers the status of the traits in a common ancestor by reference to known evolutionary trees derived from the fossil record or molecular data, together with outgroup analysis and the principle of parsimony (Occam's razor). The result is a pattern of the degree and direction of transformation from the ancestral state among the taxa examined. This leads to working hypotheses concerning which traits are likely to be homologous (similar due to derivation from a common ancestor) or homoplastic (similar due to parallel or convergent evolution). Glenn Northcutt (see **Figure 1**) has led the way in these kinds of studies, and has been a major influence in helping comparative neurobiologists make the transition from scala naturae thinking to an outlook that is congruent with the modern evolutionary synthesis. (He was also instrumental in forming the JB Johnston Club in 1980. This organization's annual meeting provides an important venue for presentations on comparative neurobiology, and the club's official journal, *Brain, Behavior and Evolution*, is an important outlet for publications.)

Of fundamental importance for comparative neurobiologists is the notion of a common plan for the vertebrate nervous system, first suggested by nineteenth century embryologists and anatomists, and

Figure 1 A fanciful chinoiserie scroll, showing Glenn Northcutt and Ted Bullock exploring the wonders of animal nervous system. Presented to Ted on his 80th birthday.

elaborated upon in the past 50 years through studies of gene expression, neurochemistry, and neural connections. Features common to all vertebrates include (1) a dorsally located ectodermal neural tube that exhibits segmental structure and contains the following major divisions: telencephalon, diencephalon, mesencephalon, pons, cerebellum, medulla, and spinal cord; (2) longitudinally oriented functional zones for somatic, visceral, and special sensory and motor regions; (3) cranial nerves associated with specific functional systems, brain regions, and central nuclei; (4) sensory ganglia of cranial nerves derived from ectodermal placodes and neural crest; (5) autonomic and spinal ganglia derived from neural crest; (6) major divisions of the brain with characteristic nuclear groups having particular patterns of connections with other brain regions; and (7) common location of particular groups of neurons identifiable based upon neurotransmitter or other biochemical markers (e.g., dopaminergic projections from the midbrain).

Through the work of von Baer in the early nineteenth century, it was recognized that during vertebrate embryogenesis (1) features common to a group of animals (e.g., medial and lateral nasal processes in mammals) appear earlier than do specialized features present in particular species (e.g., multiple nasal appendages in star-nosed moles), (2) specialized features develop from more general ones (e.g., expanded vagal lobes present in some teleosts develop from the visceral sensory column of the medulla), and (3) embryos of one species do not pass through the adult stage of any other species; rather, divergent adult morphologies result from divergent development. Thus, Darwin proposed that commonality of embryonic features was a key indicator of common descent. Experimental embryology of the late nineteenth century and early twentieth century defined the germ layers and explored the processes of cell lineages, induction, restriction, determination, and differentiation. Diffusible morphogens were suspected to be important factors in inductive interactions and embryonic patterning.

The rise of molecular biology and methods for analyzing complex networks has made possible the long-awaited marriage of evolutionary theory and developmental biology. Garstang, among other biologists of the nineteenth and early twentieth centuries, emphasized that evolutionary changes (new heritable adult traits) can only arise as a result of alterations in development. However, it was not until the 1980s that evolutionary developmental biology came of age. Some of the most stunning findings have involved the deeply conserved nature of genes and gene expression patterns, exemplified by the Hox genes and other genes involved in segmentation and patterning of the body (including the nervous system) in invertebrates and vertebrates. Thus, the notion of a common plan in the vertebrate nervous system must be extended to encompass a common plan for early developmental patterning of the nervous system in invertebrates and vertebrates. This naturally puts a primacy on variation in developmental characters and processes when attempting to reconstruct phylogenetic changes in the brain. Modern thinking about the relationship between ontogeny and phylogeny was reinvigorated in 1977 upon the publication of Stephen Jay Gould's book of the same name. Coming as it did at the beginning of the explosion of studies of the molecular genetics that underlies developing morphologies, this work not only clarified much of the older thinking but also set the stage for understanding phenomena such as heterochrony and homoplasy. Comparative

neurobiologists came to realize that the key to understanding homologies was to analyze developing nervous systems in addition to their adult endpoints.

The development of phylogenetic systematics (cladistics) provided comparative neurobiologists with a method by which to make strong inferences about the evolution of neural characters. These schemes often relied on existing phylogenies derived from paleontological data, together with the principles of parsimony and outgroup comparisons applied to the distribution of the neural character in question. Another approach was to develop the phylogenetic tree *de novo*, based upon the status of multiple neural characters. One of the most scientifically charged applications of this approach was the contention by John Pettigrew that megachiropteran bats were more closely related to primates than to the microchiropteran bats, based primarily upon their patterns of retinotectal connections. This finding provoked a long and lively debate that involved bat biologists as well as neuroscientists. Argument pivoted upon issues familiar from interpretations of character data in many other systems: the relative weight which should be given to various neural and nonneural characters, the independence of characters, and ability to infer the developmental trajectory of a character by observing its organization in the adult. Similar considerations apply to the question of the role of the reptilian dorsal ventricular ridge in mammalian brain evolution. Current debate related to that issue centers on the relative primacy of neural connections versus developmental data.

The analysis of the developmental processes rather than the adult characters that represent their endpoints is providing a rich basis upon which to make comparisons. Stuart Kauffman and others have developed methods for modeling complex networks consisting of many interacting units. This has facilitated understanding the dynamics of adaptive evolution on fitness landscapes, building on the earlier work of Sewall Wright, RA Fisher, and JBS Haldane. (This viewpoint, that evolution explores the functional possibilities inherent in various architectural combinations, would undoubtedly please Geoffroy Saint-Hilaire to no end.) Eric Davidson and others have emphasized that alterations in gene expression in developing systems often occur because of changes in the *cis*-regulatory elements that modulate spatiotemporal patterns of gene expression. These perspectives, together with the deeply conserved aspect of many developmental genes, suggest that under certain conditions genetic regulatory networks exhibit self-organization into ordered states, and that changes in the structure and expression patterns of genetic regulatory networks become the means by which evolution explores fitness landscapes through phenotypic change, subject to general constraints such as the laws of physics and metabolic possibility, and subject to local constraints such as the current habitat and climate. One of the most interesting constraints is the current state of the genetic regulatory network. This biases the direction of change and the basins of attraction (stable states) which the network is likely to find. Thus, as noted by Gould in his magnum opus, *The Structure of Evolutionary Theory*, the random mutation scheme inherited from Darwin is being supplemented with the view that complex evolving networks are biased and generate some of their own order. A central issue is the extent to which this occurs. If developmental constraints in complex gene networks make some transformations more likely to occur, then much of what has been considered parallel evolution (one kind of homoplasy) may be explained as basins of attraction that represent recurrent stable states for complex gene networks. This may have occurred in megachiropteran bat retinotectal connections and vertebrate electroreception, among other cases.

Because brains play a major role in the generation of behavior, and some behavior is adaptive, it follows that if environmental influences alter the brain they can alter behavior, and thus fitness (defined as reproductive success). Therefore, phenotypic plasticity becomes another significant variable to consider in theories of brain evolution. This is particularly true for brains because of their role as a dynamic interface between the organism and the environment, through sensory inputs and motor outputs organized as behavior. Furthermore, nervous systems exhibit a wide range of physiological plasticity, as occurs during learning. Cichlid fish are proving very useful in this endeavor because of their wide range of adaptive phenotypes and rapidity of change. Songbirds have also been an important example due to hormonally driven seasonal changes in brain structure related to song production and learning and song perception. Future research will explore the range of such plasticity, which is found in phenotypes involving factors such as the regulation of receptors and action of hormones and neuromodulators, but also in more conserved traits such as patterns of neuronal connectivity.

See also: Allometric Analysis of Brain Size; Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brain Scaling Laws; Brains of Primitive Chordates; Cajal's Place in the History of Neuroscience; Cetacean Brains; Evolution of Vertebrate Brains; History of Neuroscience: Early Neuroscience; Neuroanatomy Methods in Humans and Animals.

Further Reading

Ariëns Kappers CU, Huber CG, and Crosby EC (1936) *Comparative Anatomy of the Nervous System of Vertebrates, Including Man*. New York: Hafner.

Bullock TH and Horridge GA (1965) *Structure and Function in the Nervous Systems of Invertebrates*. San Francisco: W.H. Freeman.

Burrows M (1996) *The Neurobiology of an Insect Brain*. Oxford: Oxford University Press.

Butler AB and Hodos W (2005) *Comparative Vertebrate Neuroanatomy; Evolution and Adaptation*, 2nd edn., Hoboken, NJ: Wiley.

Carew TJ (2001) *Behavioral Neurobiology*. Sunderland, MA: Sinauer.

Davidson EH (2006) *The Regulatory Genome. Genetic Regulatory Networks in Development and Evolution*. Amsterdam: Elsevier.

Fentress JC (ed.) (1976) *Simpler Networks and Behavior*. Sunderland, MA: Sinauer.

Finger S (1994) *Origins of Neuroscience*. New York: Oxford University Press.

Gould SJ (1977) *Ontogeny and Phylogeny*. Cambridge, MA: Harvard University Press.

Gould SJ (2002) *The Structure of Evolutionary Theory*. Cambridge, MA: Harvard University Press.

Kaas JH (ed.) (2006) *Evolution of Nervous Systems, a Comprehensive Reference* (4 volumes). Amsterdam: Elsevier.

Kauffman SJ (1993) *The Origins of Order*. Oxford: Oxford University Press.

Neiuwenhuys R, Ten Donkelaar HJ and Nicholson C (eds.) (1998) *The Central Nervous System of Vertebrates* (3 volumes). Berlin: Springer.

Northcutt RG (2001) Changing views of brain evolution. *Brain Research Bulletin* 55: 663–674.

Striedter GF (2005) *Principles of Brain Evolution*. Sunderland, MA: Sinauer.

Complexins

J Tang, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Complexins (also called synaphins) are small synaptic proteins. They form a small protein family with four isoforms, I–IV. Complexins I (synaphin II) and II (synaphin I) were first identified as co-binding proteins with SNARE complex in the immunoprecipitates by syntaxin antibody. Their name derives from the feature that they preferentially bind to SNARE complex, but not SNARE monomer or heterodimer. Composed of 134 amino acids, both complexin I and complexin II are highly negative charged proteins. The amino acid sequence identity of complexins I and II can reach 84%. The homology is also high for complexin I and II orthologs in different species, especially for complexin II, which is 100% identical among mouse, rat, and human. Complexins III and IV were discovered by protein profile search in a genomic database (CELERA). They have a higher molecular weight, containing 158 and 160 amino acids, respectively. Protein sequence alignment of complexins III and IV revealed that these two isoforms are homologous (58% identity) to each other, and their orthologs in human, mouse, and rat are also very similar. However, there is only limited homology between complexins III/IV and complexins I/II (24–28% identity), indicating that these four isoforms can be further divided into two subfamilies.

All complexins have an evolutionary conserved central region, which is responsible for SNARE complex binding. Outside this region, the similarity is only limited to certain species or one of the two subfamilies. The most striking difference between vertebrate subfamilies is that complexins III/IV contain a CAAX consensus sequence at their C-terminus for posttranslational prenylation that is not present in complexins I/II but is found in some invertebrate complexins. With the exception of this CAAX motif, most invertebrate complexins (with only one or two known isoforms) are more similar to complexins I/II. These data suggest that complexins I/II and some other complexins may lose this CAAX sequence after separation from complexins III/IV.

Expression and Localization

Complexins I–III appear late in brain development on postnatal day 6 (P6), and their levels increase and reach a plateau at approximately P20, a period of synaptogenesis. Complexins I and II are highly enriched in the brain, with much less in retina. In addition, low levels of complexin II were found ubiquitously expressed in peripheral tissues, such as kidney, muscle, and lung. Immunostaining revealed that complexins I/II were neuron specific, with no immunoreactivity detected in glia. The distribution of complexins I and II in the brain and retina, as well as their subcellular localization, has been extensively studied using immunohistochemistry and immunocytochemistry. Both complexins are widely expressed in the brain but with distinct distribution patterns. Complexin I is present in a subpopulation of neurons in cortical layers III–VI, with characteristic punctuate labeling around neuron soma. Cortical layers II, III, V, and VI are the main regions for complexin II, but there are additional granular or stringlike signals scattered over these layers. In hippocampus, complexin I staining is mainly concentrated in the cell soma in dentate gyrus, the CA1–3 region, and some scattered neurons, whereas complexin II covers both cell bodies and synaptic layers of these neurons. Complexin II also highlights the granule cells of cerebellum (complexin I has weaker staining), with ubiquitous granular staining in the molecular layer. Basket and stellate cells are only immunoreactive to complexin I. In Purkinje cell soma, neither complexin I nor complexin II is present. Many other regions, including the spinal cord, contain both complexin I and complexin II, but the pattern is quite different.

Within neurons, complexin immunoreactivity is observed primarily in somas, axons, and presynaptic terminals. Electron microscopy results of complexin labeling have shown signals in both axosomatic and axodendritic synapses, usually associated with synaptic vesicles and plasma membrane. Postsynaptic labeling has also been detected in some dendrites and dendritic spines with much lower intensity. Interestingly, complexin I is remarkably localized in the axosomatic synapses, and complexin II is preferentially present in terminals of axodendritic synapses. The small degree of overlap indicates that they might work mainly in different synapses. The accumulation of complexins in synapses is time dependent. In the early stage (P14), complexin staining of pyramidal neurons is limited to cell somas on hippocampus slice. In comparison, synaptic vesicle proteins already exhibit a punctate distribution. After maturation, the intensity of labeling in the hippocampus is remarkably increased in synaptic layers. These *in vivo* findings of the differential distribution of complexins I and II, as well as the enrichment in synapses during maturation,

are consistent with observations of *in vitro* cultured hippocampal neurons.

In retina, complexin I and II antibodies stain GABAergic amacrine cells. Whether glycinergic amacrine cells contain complexins I/II is controversial since there is supportive evidence for both. Small amounts of complexins I and II have been detected in horizontal cells and the soma of amacrine cells in the inner nuclear layer (INL), whereas they are much more widely distributed in the inner plexiform layer (IPL), as well as ganglion cells in the ganglion cell layer (GCL). Both complexins are absent in photoreceptors and bipolar cells since their staining does not overlap with vesicular glutamate transporter 1 (VGLUT1)-labeled synaptic terminals of these cells.

Unlike complexins I and II, complexin IV is exclusively expressed in retina, whereas complexin III is preferentially located in retina but also may be found in certain areas of the brain, mainly cortex, hippocampus, inferior colliculus, and striatum. Only trace amounts of complexins are detected in other tissues. In retina, complexin III immunoreactivity is present in both the outer plexiform layer and the IPL. Double labeling with VGLUT1 indicated that complexin III is expressed in photoreceptors and bipolar cells. In the IPL, complexin III antibody also labels glycinergic amacrine cells. Similar to the distribution of complexin III, both photoreceptor and bipolar cell terminals contain complexin IV. However, complexin IV-stained terminals in IPL are located in the middle of this layer, which is different from complexin III-containing terminals (in IPL), which are next to the INL and GCL. Moreover, within photoreceptors, complexin III is present in both cone and rod photoreceptor terminals, whereas complexin IV exists only in rod photoreceptor terminals.

Similar to the subcellular localization of complexins I/II, complexins III and IV exhibit a characteristic presynaptic distribution in primary hippocampal neurons. The distribution is dependent on the prenylation modification of the CAAX motif at the C-terminus since cysteine-to-serine mutation in both complexin III and complexin IV altered their localization (like soluble protein). *In vivo* assay with HEK293 cells has confirmed that they can be farnesylated at cysteine within the CAAX motif. Farnesylation enabled complexins III/IV to associate with membrane, which may account for the accelerated accumulation of complexins III/IV in presynaptic terminals as early as D14 in *in vitro* cultured neurons.

Structure and Biochemical Features

Nuclear magnetic resonance and circular dichroism studies have revealed that complexin has no tertiary structure except a stable α helix at the N-terminus of the central region (residues 29–64). Binding of the second half of this α helix to the SNARE complex might induce the extension of the helix to the C-terminal half of the central region and result in the complete interaction with the SNARE complex. Complexin binding did not cause major conformational changes to SNARE complex. Based on fluorescence resonance energy transfer measurement and the crystal structure of the complexin/SNARE complex, it is clear that complexin binds to the four-helix bundle of SNARE complex in an antiparallel manner with a 1:1 stoichiometry (Figure 1). Within this complex, complexin is located in the groove between syntaxin and synaptobrevin and has no direct contact with SNAP-25. This is in agreement with the biochemical result that the substitution of SNAP-25 has the least effect on complexin binding.

The binding of complexin to SNARE complex is mediated by the combination of hydrogen bonding and hydrophobic and ionic interactions in the central region from residues 48 to 70: three arginines (R48/R59/R63) of complexin form salt bridges with three aspartate residues (D57/D65/D68) from synaptobrevin, tyrosine 52 has extensive contact with synaptobrevin by hydrogen bond and lysine 69 and tyrosine 70 respectively contribute to the ionic interaction and hydrogen bonding with syntaxin aspartate D218. These residues are conserved in all complexin isoforms from various species. The substitution of them by alanine completely abolishes binding to the SNARE complex.

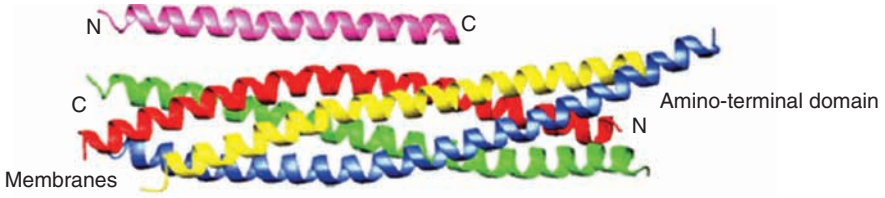

Figure 1 Three-dimensional structure of the complexin/SNARE complex. Complexin is shown in pink and binds antiparallel to the mini SNARE complex, which is composed of syntaxin C-terminal SNARE motif (yellow), and corresponding SNARE motifs from synaptobrevin (red) and SNAP-25 (green and blue). The approximate locations of the membranes and the N-terminal domain of syntaxin are indicated. From Rizo J and Südhof TC (2002) Snares and Munc18 in synaptic vesicle fusion. *Nature Reviews Neuroscience* 3: 641–653.

The high conservation of the primary protein sequence within this binding region among all complexins indicates that SNARE complex binding represents the most fundamental function of complexin. As expected, even complexins III/IV maintain the SNARE complex binding ability, although the affinity of complexin IV is much lower compared to that of the other three isoforms. The similarity of the general structure between rat and squid SNARE/complexin complex further confirms this conservation at the tertiary structure level, suggesting an essential physiological function for this interaction.

Although complexin binding does not induce major changes in the four-helix bundles of the SNARE complex, it promotes an extremely tight contact between synaptobrevin and syntaxin/SNAP-25 at the C-terminus of the SNARE complex. The stabilization of the SNARE complex at the C-terminus may help overcome the repulsive forces from the two opposing membrane structures and hold them in close association.

Biochemically, the characteristic feature of complexins is that they can selectively bind to the SNARE complex. The interaction is rapid ($k_{on} \sim 5 \times 10^7$ m s^{-1}) and with high affinity (K_D of approximately 10–36 nM by independent methods). However, this binding does not affect the rate and kinetics of the SNARE complex assembly. In contrast, the extent of SNARE complex formation might determine the affinity of complexin binding. In addition, calcium has little effect on this interaction.

The relationship between complexin and other SNARE complex binding proteins, such as synaptotagmin 1 and α-SNAP, has also been investigated. Experiments using both recombinant and native proteins have confirmed some of the previously discussed results and uncovered a striking finding that there is competition between complexin I, synaptotagmin 1, and α-SNAP for SNARE complex binding. In the hydrophilic environment, α-SNAP has the highest affinity to the SNARE complex. It can easily replace complexin from the immobilized SNARE complex, but the reverse competition fails under the same conditions. Similarly, complexin I is more efficient than synaptotagmin 1 for SNARE complex binding. Mutation (R48/R59/K69/Y70A) of complexin I completely abolishes its SNARE complex binding ability, as well as the inhibitory effect on synaptotagmin 1 for SNARE complex binding, indicating that the competition is specific. However, when SNARE complex is embedded in the phospholipid bilayer to mimic its physiological status, the binding affinity for complexin I decreases dramatically (~70 times lower) compared to that in solution using stopped-flow fluorescence. The major difference results from slower k_{on} and faster k_{off} values produced in the membraneous environment. This change directly leads to the result that synaptotagmin 1 can replace complexin from the SNARE complex more efficiently than the reverse competition in a strictly calcium-dependent manner. The EC_{50} of free calcium concentration for this competition is 50 μM, which is within the physiological calcium concentration range, supporting the physiological relevance of this replacement. Structural studies indicate that complexin binding is able to stabilize the SNARE complex. However, this effect does not block SNARE complex disassembly because NSF can still easily take the SNARE complex apart in the presence of complexin with the help of α-SNAP and ATP, which is consistent with the idea that α-SNAP binds tighter than complexin to the SNARE complex.

As an important presynaptic protein involved in synaptic transmission, complexin can be regulated. Serine 93 of complexin II has been found to be phosphorylated in okadaic acid (OA)-treated primary cortical neurons by mass spectrometry. Since OA is thought to be a selective inhibitor of protein phosphatase 2A (PP2A), complexin II is a potential target of PP2A. It is unknown why complexin I is not detected for the same phosphorylation since residues 91–102 are identical for complexins I and II. Either residues outside this region determine the specificity of the modification or the phosphorylation on complexin I is very transient and weak. In addition, sequence ^{115}SILD^{118} matches the consensus motif of S/T-x-x-D/E for protein kinase CK2 phosphorylation, and *in vitro* assay has confirmed that serine 115 can be phosphorylated. Interestingly, the phosphorylated complexins I/II exhibit higher affinity for SNARE complex binding, suggesting that this modification could potentially regulate SNARE function. However, phosphorylation-specific antibody can only detect complexin II in the immunoprecipitates by complexin antibody, but not in whole brain lysate in the presence of protein phosphatase inhibitors, indicating that only a small fraction of complexin II is phosphorylated *in vivo*. The absence of complexin I signal suggests that either it is not phosphorylated by CK2 *in vivo* or the phosphorylation is very transient or very sensitive to protein phosphatase.

Mechanism of Complexin Function

Microinjection of complexin II-specific antibody into *Aplysia* buccal ganglia induces a transient increase in IPSC amplitude. Application of complexin II recombinant protein causes a sustained depression of IPSC, which can be reversed by further addition of recombinant α-SNAP. Overexpression of complexins I/II in PC12 cells suppresses the exocytosis of both large dense core vesicles and small synaptic vesicles.

These results suggest that complexins have a negative regulatory role in synaptic transmission.

Injection of competitive complexin fragment for SNARE complex binding into squid axons inhibits synaptic transmission. A similar reduction is also observed by knocking down complexin expression in mast cells and pancreatic β cells by RNAi. These results indicate that complexin is required for vesicle fusion. Complexin has been found to be able to promote the oligomerization of the SNARE complex, which is medicated by the incorporation of two individual SNARE motifs of SNAP-25 into different SNARE complexes. However, neither biochemical nor biophysical (CD and fluorescence spectroscopy) data from another research group support this function of complexin. Regardless of the controversial explanation, oligomerization of the SNARE complex has been detected and confirmed. The formation of the ring structures of the SNARE complex might bring the two membrane structures in close association and provide a scaffold for other protein binding, and it may be essential for efficient transmitter release.

Mouse genetic studies may provide the evidence for the physiological function of complexin *in vivo*. Complexin I knockout mice have strong ataxia and die within 2–4 weeks, although with an enhanced feeding regime they have a normal life span (>2 years). Complexin II null mice show no obvious phenotype, indicating that complexins I and II are partially redundant. Deletion of both complexin I and II causes the mice to die within a few hours after birth. Complexins have no effect on brain development and synaptogenesis since no abnormality is found in brain structure, cytoarchitecture, or synapse numbers and density. At the molecular level, the expression of critical synaptic proteins, including SNAREs and synaptotagmin 1, is unaffected in complexin single and double knockout mice. Excitatory postsynaptic currents (EPSCs) of neurons in hippocampal autapic culture from either complexin I or complexin II single knockout mice are indistinguishable from those of control littermates, whereas double knockout neurons exhibit a 66% reduction, which further confirms the redundancy of complexin and also indicates that complexins are necessary for exocytosis. Normal sucrose response and unchanged miniature amplitude and frequency rule out the possibilities of presynaptic readily releasable pool size reduction and postsynaptic dysfunction. Together with indistinguishable vesicle refilling rate and synaptic ultrastructures (active zone length, vesicle number, and docked vesicles) between double mutant and control, results strongly indicate that neurons without complexins I/II are capable of translocation, docking, and priming of synaptic vesicles. The defect in synaptic transmission must occur following vesicle priming and before final fusion. In agreement with these results, both vesicular and synaptic release probabilities are reduced. As expected, facilitation of transmitter release is observed in the early phase of train stimulation in double mutant neurons, whereas consistent depression of EPSC amplitude is detected in control neurons. Titration of external calcium concentration has revealed that the reduction of release probability is due to the impaired calcium sensitivity of evoked vesicle release. Insufficient calcium entry caused by changes of calcium channels can be ruled out since bypassing calcium channels with calcimycin does not eliminate the difference between mutant and control neurons. The dysfunction at the level of the SNARE complex and its associated proteins (including calcium sensor) during the last calcium-triggering step is more likely a reasonable explanation.

Calcium-dependent synaptic transmission can be divided into two components: fast synchronous and slower asynchronous release. In complexin double knockout mice, only the synchronous part is impaired, with the asynchronous part unchanged. This selective decrease of the fast component of vesicle release is very similar to the phenotype of synaptotagmin 1 knockout neurons, which has a more complete abolishment of synchronous release. However, the specific disturbance of the last step of calcium-dependent exocytosis in complexin deletion mice remains unsolved since all known biochemical characteristics of complexins are calcium independent. The similarity between synaptotagmin 1 null and complexin double knockout mice, together with the fact that synaptotagmin 1 has calcium-dependent SNARE complex binding, raises the possibility to explain the phenotype in association with the effects from other SNARE binding proteins.

Discoveries from three groups raise new hypotheses for complexin functions in exocytosis. Using *in vitro* liposome-based fusion assays or a cell fusion system with flipped SNARE proteins on plasma membrane, two groups have shown that synaptotagmin 1 can reverse the inhibitory effect of complexin on liposome and cell fusion, and the release of this clamp is strictly calcium dependent. The function of complexin here was contradictory to the results using complexin double knockout mice. One of the groups provided data on the calcium-dependent co-binding of synaptotagmin 1 and complexin with SNARE complex in liposomes as molecular evidence to explain the relief effect of synaptotagmin 1. However, their results did not indicate whether the co-binding was due to synaptotagmin 1 and complexin binding to the SNARE complex together or only by synaptotagmin 1 binding to liposomes (synaptotagmin 1 has calcium-dependent phospholipids binding). An unanswered question is how the formation of the larger complex (complexin,

synaptotagmin 1, and SNARE complex) resulted in clamp relief.

Another hypothesis is based on the finding that complexin competes with synaptotagmin 1 for SNARE complex binding. As mentioned previously, with the SNARE complex embedded in membrane, synaptotagmin 1 can efficiently replace complexin from the SNARE complex in a strict calcium-dependent manner within the physiological concentration range. This hypothesis emphasizes the function of complexin as an activator to stabilize the SNARE complex after its formation in the absence of calcium. Energetically, the primary effect of complexin binding to the SNARE complex is to overcome repulsive forces and catalyze the SNARE complex to a 'metastable' status for further fusion reactions; with a secondary effect to prevent fusion before calcium entry. Calcium-induced simultaneous binding of synaptotagmin 1 to the SNARE complex and phospholipids removes complexin clamp and triggers vesicle fusion. This hypothesis suggests that primed vesicles containing a partly assembled SNARE complex are either substrates for asynchronous Ca^{2+}-triggered release or activated by complexin binding to SNARE complexes, which transforms them into a substrate for Ca^{2+}-triggered synchronous release. This hypothesis includes a 'clamp' function for complexins that is secondary to their role in activating SNARE complexes and only clamping of activated SNARE complexes produced by complexin binding (Figure 2). In this model, complexin and synaptotagmin 1 function in the same pathway for synchronous release but are not involved in the other asynchronous process, which explains the similarity in phenotype between knockout mice.

To test the hypothesis *in vivo*, a complexin 1/synaptobrevin 2/Venus fusion protein was introduced

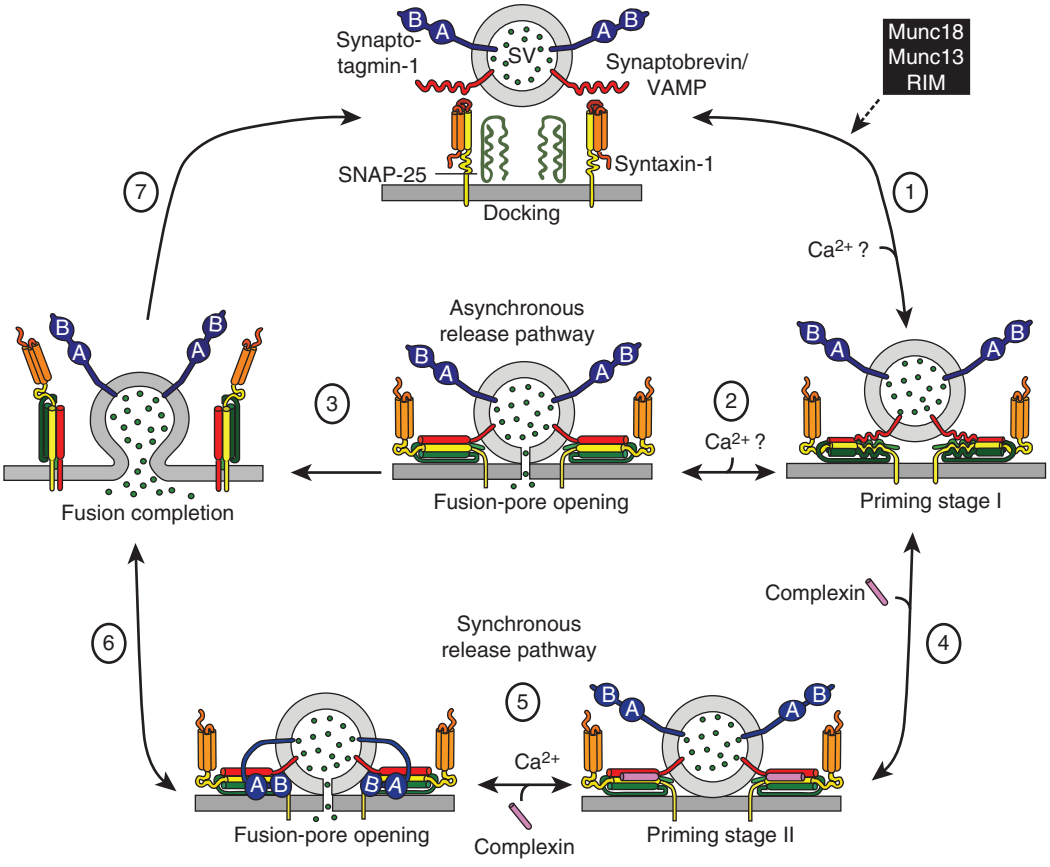

Figure 2 Model for complexin and synaptotagmin 1 function in Ca^{2+}-triggered release. Docked vesicles containing unassembled SNARE complexes (top) are primed for release by partial SNARE complex assembly catalyzed by Munc18, Munc13, and RIM (step 1). The resulting primed vesicles form the substrates for two release pathways: asynchronous release, in which full assembly of the SNARE complexes leads to fusion-pore opening followed by complete fusion (steps 2 and 3), or synchronous release, in which 'superpriming' by binding of complexins to assembled SNARE complexes (step 4) activates and freezes SNARE complexes in a metastable state (referred to as priming stage 2). This stage is substrate for fast Ca^{2+}-triggered release when Ca^{2+} binding to synaptotagmin 1 induces its binding to phospholipids and SNARE complexes, with the latter reaction displacing complexin and resulting in fusion pore opening (step 5). Again, opened fusion pores can then dilate to complete fusion (step 6). Fused vesicles undergo endocytosis and the *cis*-SNARE complex is dissembled by α-SNAP and NSF and ready for the next cycle (step 7). From Tang J, Maximov A, Shin OH, Dai H, Rizo J, and Südhof TC (2006) A complexin/synaptotagmin 1 switch controls fast synaptic vesicle endocystosis. *Cell* 126: 1175–1187.

into cortical neurons by lentivirus infection. With a flexible linker at the C-terminus, exogenous complexin folded back and bound to the assembled SNARE complex into which the fused synaptobrevin 2 was incorporated. The high affinity of intramolecular complexin binding to the SNARE complex did counteract the effect of synaptotagmin 1 and inhibit its binding to the SNARE complex, but not α-SNAP binding. Thus, the expression of this fusion protein in wild-type neurons produced a phenotype similar to synaptotagmin 1 knockout mice without a significant change in vesicle pool size. Mutations (R48/R59/K69/Y70A) on complexin could eliminate this effect, indicating that the effect was specific.

Although complexin II knockout mice appeared normal and their hippocampal neurons had no significant difference in EPSC amplitude and paired-pulse facilitation (PPF), titanic stimulation-induced long-term potentiation (LTP) in both CA1 and CA3 regions was impaired. The decrease of LTP in CA1 was frequency dependent, and the difference was significant only at 100 Hz and not at 10 and 30 Hz. Mossy fiber-CA3 LTP was expressed presynaptically, so the reduction of LTP could be directly caused by presynaptic depletion of complexin II. However, Schaffer collateral fiber-CA1 LTP was mediated mainly by the postsynaptic activation and insertion of AMPA receptors. Direct application of NMDA induced the same potentiation of EPSC, suggesting that the postsynaptic machinery was presumably intact. One possibility was that complexin acted downstream of the retrograde signals from the postsynaptic machinery at a later stage of LTP. Another was that complexin may be involved in postsynaptic function, although its expression level was very low. In agreement with the electrophysiological finding, additional behavior studies of complexin II-deficient mice showed a severe impairment in Morris water maze, suggesting that complexin II plays an essential role in synaptic plasticity, even learning and memory.

The conservation of all complexin isoforms for SNARE complex binding suggested that complexins III/IV may have similar functions. Expression of complexins III and IV fully rescued the reduction of EPSC amplitude in complexin I/II double knockout neurons. Although the farnesylation of complexin III had little effect on rescue, it did regulate complexin IV function, with the mutant (C158S) having much less rescue activity.

Pathological Function of Complexin

Huntington's disease (HD) is a neuronal degenerative disorder characterized by motor, emotion, and cognitive dysfunction. It is caused by an expanded CAG repeat in the huntingtin gene, which results in a polyglutamine repeat expansion in protein. A selective and progressive depletion of complexin II has been detected in brains of HD model mice (R6/2), which expresses the first axon of the huntingtin gene with an expanded CAG repeat. Measurement of protein samples from postmortem brains of HD cases has confirmed the decrease of complexin II in striatum, which is accompanied by a significant reduction of synaptobrevin 2 and Rab3A. Unexpectedly, in contrast to the results from striatum, the protein levels of all three proteins are remarkably elevated in samples from hippocampus. The selective depletion or increase of limited synaptic proteins indicates that this is not due to progressive neuronal degeneration since most synaptic proteins are unaffected. Moreover, complexin II knockout mice and R6/2 mice have a common impairment of LTP in the CA3 region, suggesting that the depletion of complexin II may contribute to the LTP reduction in R6/2 mice. Cross of R6/2 mice with complexin II knockout mice does not cause additional deteriorative effects on the phenotype of R6/2 mice, indicating that loss of complexin II shares the same mechanism underlying R6/2 phenotype. However, whether complexin II has a causal effect and how complexin II is involved in the pathological process of HD are unknown.

Complexins have also been widely employed as an indicator in many mental disorders, such as schizophrenia, bipolar disorder, and major depression. Studies in this field have provided very controversial results: both an increase and a reduction of complexin II:complexin I ratios have been reported for schizophrenia from different groups. Furthermore, genetic association analysis in Japanese schizophrenia patients found a very weak linkage between complexin and schizophrenia. All these inconsistencies indicate that the issues of complexin functions in these disorders and the responses to different drugs are quite open.

Future Directions

Much progress has been achieved in the past few years, but more questions are still open and need to be addressed. Complexins are in a pool with various proteins within the presynaptic terminals; the relationship between them needs to be reevaluated and explored. One example is Munc18. Studies have demonstrated that Munc18 is incompatible with SNAP-25 for syntaxin binding, implying that Munc18 functions as a binding partner for the closed form of syntaxin but not for the open form of syntaxin within the SNARE complex. However, data have demonstrated the existence of a SNARE/Munc18 complex. In addition, unpublished results from GST-complexin pull-down

assay indicate an efficient co-binding of Munc18. Thus, whether complexin and Munc18 bind together with the same SNARE complex or not is a very interesting question. This can be further expanded for other proteins. As another example, in complexin double knockout mice, there is approximately 34% EPSC left. It is known that complexin III is also expressed in hippocampus. By making triple knockout mice or performing complexin III knockdown on the background of complexin I/II double knockout neurons, we can answer the question of whether the remaining response is due to the existence of complexin III.

See also: Calcium Channels and SNARE Proteins; Huntington's Disease; NSF and SNAPs; Retinal Pharmacology: Inner Retinal Layers; SNAREs; Synaptotagmins; Triplicate Repeats: Huntington's disease.

Further Reading

Bowen ME, Weninger K, Ernst J, Chu S, and Brunger AT (2005) Single-molecule studies of synaptotagmin-1 and complexin binding to the SNARE complex. *Biophysical Journal* 89: 690–702.

Chen X, Tomchick DR, Kovrigin E, et al. (2002) Three-dimensional structure of the complexin/SNARE complex. *Neuron* 33: 397–409.

Giraudo CG, Eng WS, Melia TJ, and Rothman JE (2006) A clamping mechanism involved in SNARE-dependent exocytosis. *Science* 313: 676–680.

McMahon HT, Missler M, Li C, and Südhof TC (1995) Complexins: Cytosolic proteins that regulate SNAP-receptor function. *Cell* 83: 111–119.

Pabst S, Hazzard JW, Antonin W, et al. (2000) Selective interaction of complexin with the neuronal SNARE complex. Determination of the binding regions. *Journal of Biological Chemistry* 275: 19808–19818.

Pabst S, Margittai M, Vainius D, Langen R, Jahn R, and Fasshauer D (2002) Rapid and selective binding to the synaptic SNARE complex suggests a modulatory role of complexins in neuroexocytosis. *Journal of Biological Chemistry* 277: 7838–7848.

Reim K, Mansour M, Varoqueaux F, et al. (2001) Complexins regulate a late step in Ca^{2+}-dependent neurotransmitter release. *Cell* 104: 71–81.

Reim K, Wegmeyer H, Brandstatter JH, et al. (2005) Structurally and functionally unique complexins at retinal ribbon synapses. *Journal of Cell Biology* 169: 669–680.

Schaub JR, Lu X, Doneske B, Shin YK, and McNew JA (2006) Hemifusion arrest by complexin is relieved by Ca^{2+}-synaptotagmin I. *Nature Structural & Molecular Biology* 13: 748–750.

Tang J, Maximov A, Shin OH, Dai H, Rizo J, and Südhof TC (2006) A complexin/synaptotagmin 1 switch controls fast synaptic vesicle exocytosis. *Cell* 126: 1175–1187.

Yamada M, Saisu H, Ishizuka T, Takahashi H, and Abe T (1999) Immunohistochemical distribution of the two isoforms of synaphin/complexin involved in neurotransmitter release: Localization at the distinct central nervous system regions and synaptic types. *Neuroscience* 93: 7–18.

Computational Approaches to Motor Control

R Shadmehr, Johns Hopkins School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

There are an infinite number of movements that we can do at any given moment. How do we select what kind of movement to perform? We perform actions to acquire rewarding states. For example, we may reach in order to pick up a cup of tea and bring it to our mouth. The drinking of the tea is the rewarding state that we have chosen as our goal, and the reach is the action that allows us to experience that rewarding state. The problem of what to do begins with a catalog of all potential rewarding states and proceeds with an evaluation of these possibilities to decide on one that is most valuable.

Our ability to choose from among the infinite number of potentially rewarding states depends on our brain's ability to assign value to those states. This value includes a measure of cost and a measure of reward. The basal ganglia play a fundamental role in this ability to assign value to potentially rewarding states and help in choosing the goal. However, once the goal has been chosen, the problem of how to perform it remains. The cerebellum plays a fundamental role in this ability to control the process of performing actions.

The problem of motor control is summarized in Figure 1. Through our belief about costs and rewards of various states, we choose a goal state. We use the cost and rewards associated with our goal to generate motor commands that we believe will bring us closer to the goal state. Our motor commands cause states of our body to change. We sense these state changes through our sensory system. However, we also predict what the sensory changes should be through internal models (called forward models). We combine what we predicted with what we actually sensed to form beliefs about the state of our body. We compare our belief about our current state with our goal state and continue to produce further motor commands until either the goal state is achieved or we abandon that goal and consider another one.

The Problem of What to Do

We select actions based on cues that are available to us. Some of these cues tell us about the position of our body and objects in the environment with respect to a spatial map. For example, suppose that a rat is released into a pool of water from some random starting point. A platform is positioned in a specific location just below the waterline and cannot be seen. The platform is always at the same location in the pool. Rats dislike being wet and will try to find a way to elevate themselves. The normal rat can learn to locate the platform position by paying attention to the visual cues that surround the pool. For example, the rat might learn that the platform is near the wall with a particular feature. This requires learning a spatial map of where the platform is located with respect to the surrounding visual cues. With repeated swims, the animal learns a spatial map. This spatial map is analogous to a reward function that associates places in the pool with the likelihood of the platform (and therefore the likelihood of not having to be wet). Once they learn the map, rats can find the platform regardless of where they are released into the water because the map is related to the cues on the walls. If the platform is removed, the normal animal will spend most of its time searching in the region where the platform used to be.

Learning of a spatial map depends on the hippocampus. If a genetically altered rat with a malfunctioning hippocampus is given the same training, he will not learn the spatial map and will spend equal time in each quadrant. Therefore, selecting an action based on a spatial map likely relies on the hippocampus. Sometimes, certain cues are rewarding no matter where they are located. Consider a pool where there are two hidden platforms: one that is large enough for the rat to mount and one that is too small. Both have a distinct visual cue associated with them: a little flag attached to each platform, each of a different color, sticking out of the water. Suppose that the flag attached to the large platform is red, and the flag attached to the small platform is green. The platforms may be positioned in any part of the pool, and their locations will change from trial to trial. Therefore, in this experiment, the animal needs to learn that the red flag indicates the location of the suitable platform and is a rewarding object.

In another version of the experiment, the large platform will always be located in a particular spatial location, but the flag atop it will be a random color. In this version of the experiment, the animal needs to learn that it is not the color of the flag that is important but the spatial location. The two versions of the experiment are depicted in Figures 2(a) and 2(b). There is a natural competition between the learning systems that might be involved in these two conditions: Is the platform the same 'place' as before (where place refers to a location in the spatial map), or is the platform always where the red flag is located? After a few trials in which the flags

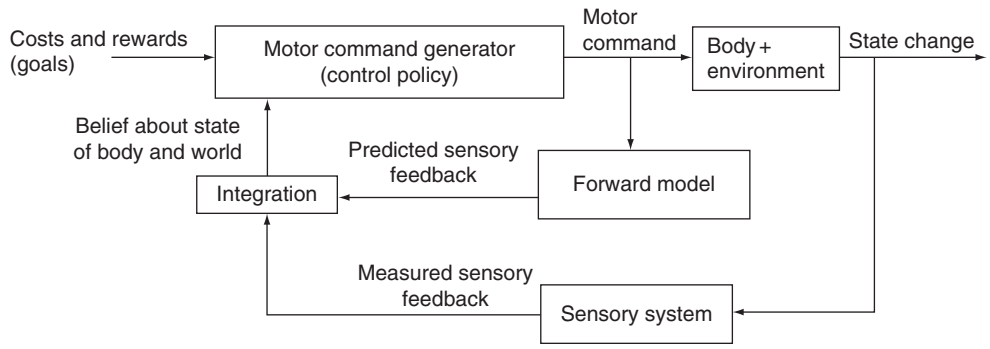

Figure 1 Schematic of motor control for voluntary movements.

move around, in the first experiment the animal should learn that the spatial map is not a good indicator of the platform and that therefore the values associated with places in the spatial map should be near zero. However, the value associated with the red flag should rise. In the second experiment, the animal should learn that it is the spatial location that is of value and that the flag colors are irrelevant.

Packard and McGaugh performed both experiments by having their animals swim eight times per day for a number of days. They recorded the number of times the animals mounted the small platform and labeled these as errors. In the first experiment, where reward was associated with the red flag, healthy animals gradually learned to swim to the red flag (Figure 2(a)). It is interesting that animals with damage to the medial temporal lobe (which includes the hippocampus and the fornix, a major afferent–efferent pathway of the hippocampus) learned the task just as well as the healthy controls. However, animals with damage to the caudate nucleus (a part of the basal ganglia) were much slower in learning the association. After days of training, they continued to attempt to mount the platform under the green flag. Therefore, it appears that the ability to associate reward to stimuli regardless of spatial location depends on the basal ganglia.

In the second experiment, where reward was associated with a spatial location, healthy animals gradually learned to swim to that location and ignore the color of the flag (Figure 2(b)). Animals with damage to the caudate nucleus performed similarly to the healthy controls. However, animals with damage to the medial temporal lobe were much slower in learning the association. Therefore, the ability to associate reward to a spatial location depends on the medial temporal lobe.

The Major Reward System in the Brain Is the Neurotransmitter Dopamine

How does reward get associated with a stimulus? That is, how do we learn the value of a location on a spatial map or the color of a flag? This issue is still poorly understood, but a major piece of the puzzle appears to be the neurotransmitter dopamine. Neurons that release dopamine have their cell bodies in the midbrain area (an area called the substantia nigra) of the brain stem. They project to three main areas: the striatum (encompassing the caudate and putamen nuclei of the basal ganglia), the hippocampus, and the prefrontal cortex. About 80% of the dopamine in the brain is in the basal ganglia.

The earliest studies of neuronal activity in the dopaminergic cells of the midbrain led to few published reports. The investigators conducting those experiments often reported frustration because the neurons never seemed to 'do anything.' For example, the dopaminergic cells discharged at about 2–4 Hz, but this rate never seemed to change. Although they did not recognize it at the time, the people who made these observations had actually stumbled on a crucial and important property of dopaminergic cells. The earliest recordings were made in monkeys that performed highly overlearned movements – a simple flexion and extension of the arm – in a highly stereotyped and repetitive manner. The monkeys performed the task perfectly and learned nothing during the recording sessions.

When Wolfram Schultz and his colleagues reexamined the activity of midbrain dopaminergic cells years later, they placed their monkeys in a situation involving learning, and the cells showed significant modulation in their activity level. They found that these cells carried a signal that conveyed a measure of the predicted reward: when an unexpected reward occurred, the dopaminergic cells increased their discharge rates. After repeated experience, the cells became unresponsive to rewards *per se* but discharged after signals that predicted reward. The omission of expected reward led to a decrease in the firing rate of dopaminergic neurons. Thus, dopaminergic cells seemed to encode both errors in predicted reward and stimuli that predicted reward.

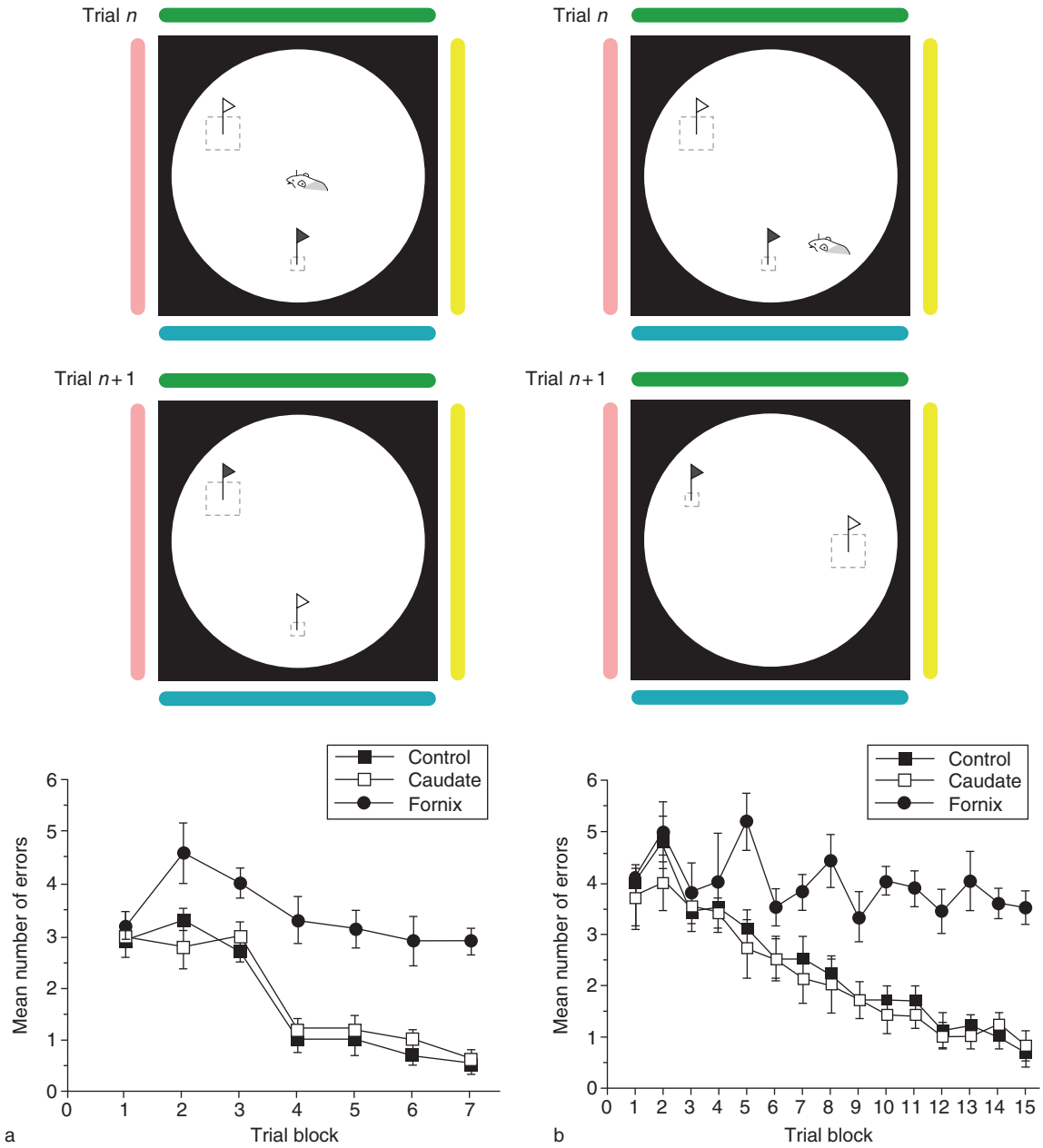

Figure 2 Learning to associate reward to locations on a spatial map depends on the hippocampus, while learning to associate reward to arbitrary visual cues depends on the striatum. (a) Two hidden platforms, one small and one large, are positioned in a constant position with respect to the colored walls. A flag is attached to each platform. On each trial, the flag is randomly changed, but the position of the platforms remains constant with respect to the walls. Each trial block is eight trials. Errors refer to the number of times the animal attempted to mount the small platform. (b) The position of the platforms randomly changes on each trial, but the colors of the flags atop each platform remain constant. Error bars are standard errors of the mean (SEM). Data from Packard MG and McGaugh JL (1992) Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: Further evidence for multiple memory systems. *Behavioral Neuroscience* 106: 439–446.

Okihide Hikosaka and his colleagues have studied the prediction of reward by neurons in the caudate nucleus. In their experiment, the monkey had to make either a left or a right saccadic eye movement, cued by a light spot, in order to keep the trials coming. However, for a block of consecutive trials, only movements to the left target would produce a reward. Then the experimenters reversed that contingency, and after the reversal, only movements to the right target produced reward. The activity of striatal neurons related to leftward saccades increased after just one trial when a leftward target became associated

with reward. In conjunction with that increase in activity, the monkeys responded faster to that cue. Then, after the experimenters switched the reward contingency so that rightward, not leftward, movements produced rewards, the cell slowly returned to its original, lower level of activity as the monkey responded more slowly to the leftward target. Therefore, rewarding behaviors are performed faster, which is certainly quite familiar to parents who request that their child do something. It is important to note that this reduced reaction time is associated with increased activity in the caudate nucleus of the basal ganglia. The basal ganglia seem to be crucial for associating reward to stimuli.

Balancing Expected Rewards with Motor Costs

Associating reward to sensory states is important, but there is another crucial piece of the puzzle: costs of motor commands. Although drinking a cup of tea may provide us with a pleasant sensory state, if that were the only constraint, then we should reach as fast as possible. But people do not reach as fast as possible. Why do we reach with a specific trajectory and speed? What is the cost that prevents us from moving at maximum speed?

The obvious cost is a metabolic cost of the motor commands. Therefore, in choosing how we should move, we need to consider how much reward we might get when we achieve our goal state and the costs associated with the motor commands that we would produce to get to that state. However, for something as simple as an eye movement, it is hard to imagine that the metabolic costs are very high. Is there some cost other than metabolic cost that constrains our movements?

It turns out that a more compelling motor cost is noise. Large motor commands produce noisy, variable movements. That is, if our objective is to reliably achieve a rewarding state, then large motor commands are a bad idea because they will increase the variance of our movements. These ideas were first described by Harris and Wolpert in a theoretical paper that showed that the cost of minimizing variance of our movements nicely explains the trajectories that animals choose for their eye and arm movements. At the heart of the theory is the assumption that large motor commands make our movements more variable and therefore difficult to control. An experiment by Jones and colleagues has demonstrated this idea (Figure 3). In this experiment, participants produced a given force with their thumb flexors (Figure 3(a)). To guide force production, the participants viewed a cursor that displayed thumb force, but the experimenters analyzed the data during a 4-s period in which this feedback had disappeared. They observed that variance of force was larger when the force magnitude was larger: standard deviation of the force grew linearly with the size of the force

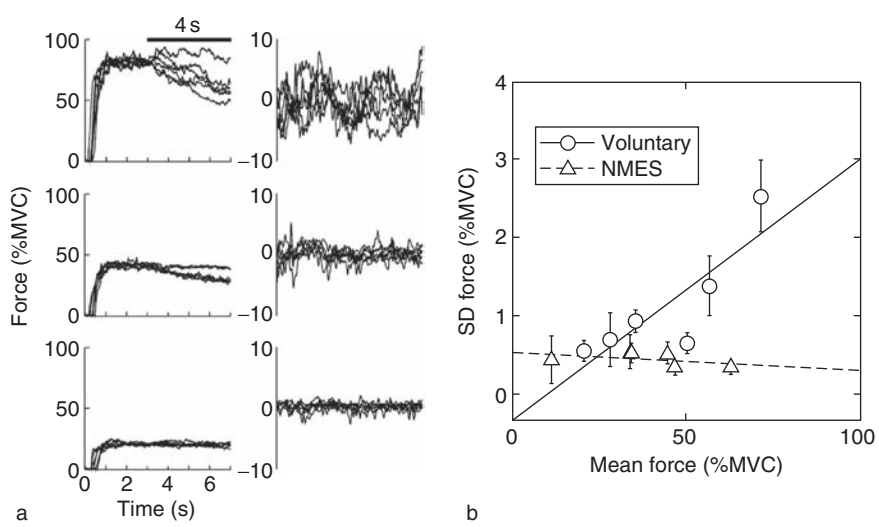

Figure 3 The standard deviation of noise grows with mean force in an isometric task. Participants produced a given force with their thumb flexors. In one condition (labeled 'voluntary'), the participants generated the force, whereas in another condition (labeled neuromuscular electrical stimulation (NMES)), the experimenters stimulated their muscles artificially to produce force. To guide force production, the participants viewed a cursor that displayed thumb force, but the experimenters analyzed the data during a 4 s period in which this feedback had disappeared. (a) Force produced by a typical subject. The period without visual feedback is marked by the horizontal bar on top (4 s), and is expanded in the second column. (b) When participants generated force, noise (measured as the standard deviation (SD)) increased linearly with force magnitude. MVC, maximum voluntary contraction. From Jones KE, Hamilton AF, and Wolpert DM (2002) Sources of signal-dependent noise during isometric force production. *Journal of Neurophysiology* 88: 1533–1544.

(Figure 3(b)). Therefore, large motor commands produce inaccurate movements because the standard deviation of the noise in the signal grows with the size of the signal.

The Problem of How to Do It: Minimizing Costs to Find a Control Policy

Todorov and Jordan have suggested that the movements we make are a reflection of a cost that includes both a measure of expected rewards and a measure of motor costs. They suggested that our movements are optimal with respect to a cost function that includes both motor costs and expected rewards. For example, if we use the vector **y** to represent our belief about our current state, vector **r** to represent our goal state, and vector **u** to represent the motor commands, the objective is to find the motor commands **u** that minimize the cost:

$$\sum_{t=1}^{N}(\mathrm{y}^{(t)}-\mathrm{r})^{T}T^{(t)}(\mathrm{y}^{(t)}-\mathrm{r})+\mathrm{u}^{(t)T}L^{(t)}\mathrm{u}^{(t)} \qquad [1]$$

where matrices T and L are measures of the weighted costs associated with achieving the rewarding state and producing the motor commands.

To get to the rewarding state, we need to find the motor commands that will minimize this cost (eqn [1]). To find the motor commands, we need to understand how our motor commands produce changes in our states. For example, if we intend to make a reaching movement, we need to know how the motor commands to our arm muscles are expected to produce sensory feedback. The term 'internal model' (or forward model) refers to this ability of the brain to predict the sensory consequences of motor commands. In a gross simplification of the problem, these predicted consequences may be written as a linear function of our motor commands:

$$\begin{aligned}\hat{\mathbf{x}}^{(t+1|t)} &= \hat{A}\hat{\mathbf{x}}^{(t)}+\hat{B}\mathbf{u}^{(t)}\\ \hat{\mathbf{y}}^{(t)} &= \hat{C}\hat{\mathbf{x}}^{(t)}\end{aligned} \qquad [2]$$

where $\hat{\mathbf{x}}^{(t)}$ represents the state of the body and the world that we interact with, $\mathbf{u}^{(t)}$ is our motor command, $\hat{\mathbf{y}}^{(t)}$ is our expected sensory feedback, and $\hat{\mathbf{x}}^{(n+1|n)}$ is our predicted state at time $t+1$, given our previous predictions and observations at time t.

Now our body and the world that we interact with have their own dynamics, represented as follows:

$$\begin{aligned}\mathbf{x}^{(t+1)} &= A\mathbf{x}^{(t)}+B\mathbf{u}^{(t)}+\varepsilon_{u}^{(t)}\\ \mathbf{y}^{(t)} &= C\mathbf{x}^{(t)}+\varepsilon_{y}^{(t)}\end{aligned} \qquad [3]$$

These dynamics are affected by noise: $\varepsilon_{u}^{(n)}$ is a stochastic variable representing motor noise, and $\varepsilon_{y}^{(n)}$ is a stochastic variable representing sensory noise. As we send motor commands, we receive a continuous stream of sensory feedback **y**. We combine what we predicted with what we observed to update our estimate of the state of our body and the world:

$$\hat{\mathbf{x}}^{(t+1|t+1)}=\hat{\mathbf{x}}^{(t+1|t)}+K^{(t)}(\mathbf{y}^{(t)}-\hat{\mathbf{y}}^{(t)}) \qquad [4]$$

In this equation, the term $K^{(t)}$ is a mixing gain (or a Kalman gain) that determines how much we should change our belief based on the errors between what we predicted and what we observed. Therefore, eqn [2] describes how we make predictions about sensory feedback, and eqn [4] describes how we combine the actual sensory observations with our predictions to update our beliefs.

Our problem is to try to get as much reward as possible for as little cost as possible. How do we generate motor commands that produce the greatest amount of reward? If eqn [2] is an accurate model of how motor commands produce changes in the state of the body and the state of the cursor that we are observing on the screen, then we can use it as a set of constraints with which to minimize eqn [1]. This is the classic linear quadratic problem in optimal control. Solving this problem yields a linear feedback control law that specifies the motor commands as a function of our belief about our state with respect to our goal:

$$\mathbf{u}^{(t)}=G^{(t)}\hat{\mathbf{x}}^{(t|t-1)} \qquad [5]$$

This feedback control law is called a control policy. The new variable G in a time-dependent matrix that represents the gain of the sensorimotor pathway. This way of formulating the problem of motor control is extremely powerful because, in principle, it can account for behavior whether we look at a single act, like a reaching movement, or our actions over the course of a day.

The theory states that our aim is specified as a long-term goal, and our actions try to achieve that goal by implementing a control policy that relies on expected costs and rewards and on our belief about where we are at any given time with respect to that goal. Therefore, the computational problem of motor control may be described as having four components:

1. We live in a world where, at any given time, many actions are possible. For any possible action, we need to know the costs associated with it as well as the sensory states that are rewarding (eqn [1]). We choose to perform an action because we believe it will be more rewarding than other potential actions. The way that we perform that action will depend on our internal estimates of costs and rewards.

2. Once we choose to acquire a rewarding state, we need to know how our motor commands produce changes in things that we can observe (eqn [2]). This is a system identification problem associated with our body or tool that we may be trying to control. Learning of this map is called forming a forward model.

3. We must learn how to actually produce the motor commands that are needed so that we minimize the costs and maximize the reward. That is, we need to figure out the 'best' motor commands that bring the cursor to the target and get it to explode. This is the constrained minimization problem (minimize eqn [1] under the constraints of eqn [3]). The result of the minimization is a feedback control law, or a control policy, that specifies our motor response to the sensory states that we observe in our body and the environment (eqn [5]). Learning of this feedback control law is sometimes called forming an inverse model.

4. As we generate motor commands, we make predictions about the sensory consequences (eqn [2]). When actual sensory feedback arrives, we integrate our observations with our predictions and form a belief about how our motor commands have affected the state of our body and the world around us (eqn [4]). Combining predictions with observations is sometimes referred to as Bayesian integration.

At the heart of the approach is the idea that we make movements to acquire rewards. The way that we make a movement, that is, how quickly we move, what trajectory we choose to execute, and how we respond to sensory feedback, depends on our control policy, which itself depends on our belief about the costs and rewards associated with that movement.

Effects of Damage to the Basal Ganglia on Representation of the Cost Function

Writing instruments are one of the most common tools that we use in our daily lives. One of the striking features of damage to the human striatum is micrographia, an impairment of writing characteristics in which letter sizes become very small and writing speed becomes slow. This condition is most common in degenerative diseases of the basal ganglia such as Parkinson's disease. However, it can also occur in the case of focal lesions. For example, consider patient FF, an individual who suffered an ischemic stroke in the left basal ganglia, in the head of the caudate nucleus and the anterior part of the putamen. When FF was asked to copy a four- or an eight-letter string of letters, the writing with the right hand was much smaller than with the left hand.

Micrographia can be a complex disorder, manifesting itself sometimes in self-generated writing and not in copying of sentences or present in copying letters but not copying geometric shapes. However, the basic observation of smallness of writing size and slowness of writing speed has been puzzling. Current theories suggest that the disorder may be related to a function of the striatum in forming complex repetitive and sequential motor actions. But if we assume that a fundamental role of the striatum is to associate sensory states with their expected reward properties, then the optimal control formulation of motor control provides a different perspective.

In the optimal control framework, there are no desired trajectories for our movements. That is, we do not move our hand along a particular path because that path is somehow hardwired in our brain. Rather, the path is a result of a control policy (eqn [5]), which itself is a result of minimization of a cost function (eqn [1]). The cost function depends on two quantities: an expected reward and a motor cost. What matters is the ratio between these two costs. In theory, movements become small and slow if the cost associated with the motor commands becomes large compared with expected rewards. Therefore, micrographia may be a reflection of a relative loss of expected reward, or increased motor costs, associated with the use of the writing instrument.

If we assume that the control policies that govern our movements (eqn [5]) require an internal measure of costs and rewards (eqn [1]), and if we further assume that the striatum is a crucial node where the brain forms associations between visual cues and expected rewards, then perhaps we can understand bradykinesia (slowness of movements) in Parkinson's disease as a reflection of a disorder in this implicit measure of costs and rewards. Mazzoni and colleagues recently reported that although patients with Parkinson's disease reach more slowly than age-matched controls, they are nevertheless able to make fast movements without a loss in accuracy if forced to move fast by the experimental task. However, the patients take longer (require more trials) to accumulate a set number of movements at the required speed. Thus bradykinesia (slowness of movements), one of the hallmarks of Parkinson's disease, may represent decreased intrinsic value of a given motor task because of an imbalance between an estimate of an implicitly determined cost for making a fast movement and the expected rewards.

Cerebellar Damage and the Ability to Predict Sensory Consequences of Motor Commands

In the optimal control framework, control policies generate motor commands on the basis of beliefs

about the state of the body and the environment (eqn [5]). In eqn [4], we see that the state estimate $\hat{\mathbf{x}}^{(t)}$ depends on two quantities: a prediction $\hat{\mathbf{y}}^{(t)}$ and an observation $\mathbf{y}^{(t)}$. The prediction comes from an internal model that uses a copy of the motor commands to estimate the state change that is expected to occur (eqn [2]). The observation comes from the sensory system. The predicted value acts as a prior belief that 'colors' the observed value. That is, our beliefs are not based on our observations alone. Rather, our beliefs are a combination of what we predicted and what we observed.

An important part of the optimal control framework is that motor commands continuously depend on predictions about state, which in turn continuously depend on predictions about sensory consequences of motor commands (the internal feedback pathway that relies on the forward model in **Figure 2**). Some movements are so fast that there is no time for the sensory system to play a role. A prominent example is control of saccades, rapid eye movements that move the eyes to a new location, typically within 50–80 ms. Such movements are too brief for visual feedback to influence saccade trajectory. In fact, the brain actively suppresses visual processing during saccades to reduce the perception of motion. Furthermore, proprioceptive signals from the eyes do not play any significant role in controlling saccade trajectories. Thus, the brain must guide saccade trajectories in the absence of sensory feedback. How is this accomplished? A plausible solution is for the brain to use an internal estimate of the state of the eye, derived from a copy of ongoing motor commands. This idea is supported by the observation that both natural and drug-induced variability of saccade velocity and duration have little influence on saccade amplitude. That is, variability of the saccade's trajectory is partially corrected as the saccade progresses.

What are the neural substrates of this internal feedback? The cerebellum is known to be critical for many aspects of saccade control and adaptation. The projections from the superior colliculus to the cerebellum may provide the efference copy. Indeed, Takeichi and colleagues have shown that adaptive changes in saccade amplitude are already reflected in the nucleus reticularis tegmenti pontis, a major source of input to the cerebellum. Together, the superior colliculus–cerebellar–brain stem side loop seems important for steering saccade trajectories midflight. This side loop is a likely candidate for the role of forward model of the plant to produce midflight corrective feedback.

A simple experiment can test whether the brain has the capacity to predict consequences of self-generated motor commands. In an experiment described by Nowakand colleagues, the participant holds a force transducer that measures grip forces (participant's right hand in **Figure 4**). Attached to the transducer is a basket. The experimenter drops a ball into the basket. When the ball drops, it exerts a downward

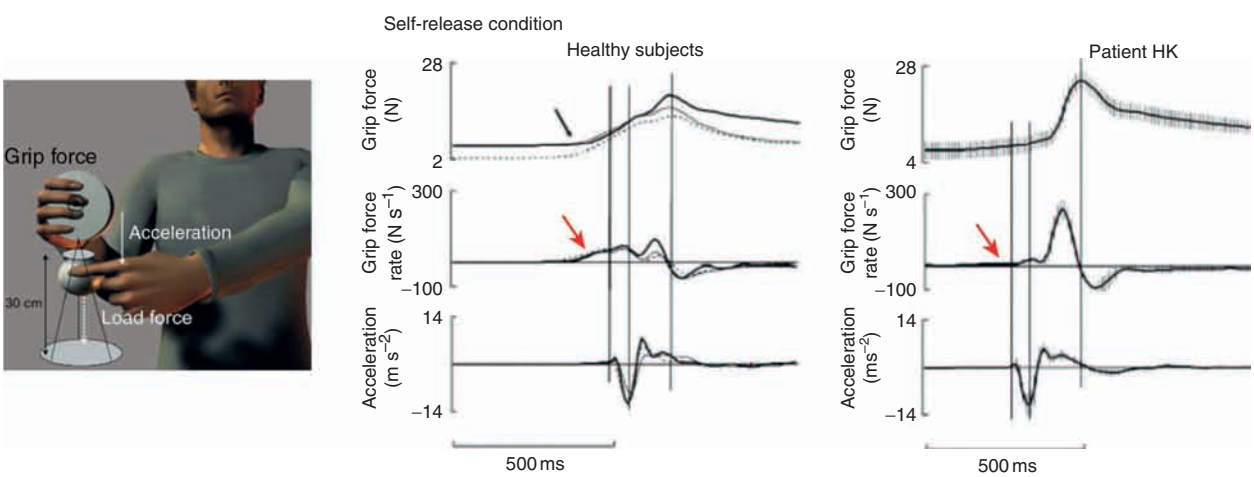

Figure 4 The cerebellum participates in predicting the sensory consequences of motor commands. In an experiment described by Nowak and colleagues, the subject holds a force transducer that measures grip forces. Attached to the transducer is a basket. When a healthy individual drops the ball, the brain predicts the consequences: on impact, the ball will exert a downward force. The traces indicate the grip force of the hand holding the transducer (right hand in the picture), rate of change in the grip force, and acceleration of that hand along the vertical dimension. The first vertical line indicates time of ball impact on the basket. The second vertical line is time of maximum hand acceleration. The third vertical line is time of maximum grip force. To maintain grip, in the healthy subject the right hand exerts a greater force right before the ball hits the basket (blue and red arrows, 50 ms before ball impact). Patient HD, who does not have a cerebellum, does not show this predictive response. Rather, she responds to the sensory feedback from the ball impact by increasing the grip force. From Nowak DA, Timmann D, and Hermsdorfer J (2007) Dexterity in cerebellar agenesis. *Neuropsychologia* 45: 696–703.

load force on the hand. The participant responds by squeezing the transducer so that it will not slip out of the hand. Because there are delays in sensing the drop of the ball, the grip response by the participant is delayed with respect to the drop of the ball (about 100 ms after ball hits the basket, the grip force increases).

Nowak and colleagues have described patient HK, who did not have a cerebellum because of a very rare developmental condition. When the experimenter dropped the ball into the basket, both the healthy individual and HK showed the delayed response. Therefore, the sensory feedback pathways appeared intact in HK. In a subsequent trial, the participant (rather than the experimenter) dropped the ball (**Figure 4**). In a healthy individual, the cerebellum predicts that release of the ball will result in an increased downward load when the ball hits the basket. In anticipation of this event, the healthy individual squeezes harder (with the right hand holding the force transducer) around the time that the ball is released. HK, however, could not make this prediction. Rather, she responded to the perturbation the same way that she responded when the ball was dropped by the experimenter. Therefore, the cerebellum appears to be required for the ability to predict the sensory consequences of motor commands; that is, it participates in representing a forward model.

Individuals with cerebellar damage are often profoundly impaired in learning to control tools. If the cerebellum is the crucial site for learning forward models (eqn [2]), then it probably makes its contribution to control of reaching via its outputs to the thalamus, which in turn projects to the cerebral cortex. Indeed, artificial stimulation of the ventrolateral thalamus impairs the ability of essential tremor patients to learn the reach adaptation task, even though the stimulation improves the patient's tremor. These results suggest that a crucial function of the cerebellum is to learn forward models that predict sensory consequences of motor commands.

See also: Basal Ganglia and Oculomotor Control; Basal Ganglia: Motor Functions; Bayesian Models of Motor Control; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Cerebellum and Oculomotor Control; Cerebellum: Clinical Pathology; Cerebellum: Models; Finger Movements: Control.

Further Reading

Barbarulo AM, Grossi D, Merola S, Conson M, and Trojano L (2007) On the genesis of unilateral micrographia of the progressive type. *Neuropsychologia* 45: 1685–1696.

Chen H, Hua SE, Smith MA, Lenz FA, and Shadmehr R (2006) Effects of human cerebellar thalamus disruption on adaptive control of reaching. *Cerebral Cortex* 16: 1462–1473.

Derkinderen P, Dupont S, Vidal JS, Chedru F, and Vidailhet M (2002) Micrographia secondary to lenticular lesions. *Movement Disorders* 17: 835–837.

Fiorillo CD, Tobler PN, and Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299: 1898–1902.

Girard B and Berthoz A (2005) From brainstem to cortex: Computational models of saccade generation circuitry. *Progress in Neurobiology* 77: 215–251.

Guthrie BL, Porter JD, and Sparks DL (1983) Corollary discharge provides accurate eye position information to the oculomotor system. *Science* 221: 1193–1195.

Harris CM and Wolpert DM (1998) Signal-dependent noise determines motor planning. *Nature* 394: 780–784.

Hopp JJ and Fuchs AF (2004) The characteristics and neuronal substrate of saccadic eye movement plasticity. *Progress in Neurobiology* 72: 27–53.

Jones KE, Hamilton AF, and Wolpert DM (2002) Sources of signal-dependent noise during isometric force production. *Journal of Neurophysiology* 88: 1533–1544.

Jurgens R, Becker W, and Kornhuber HH (1981) Natural and drug-induced variations of velocity and duration of human saccadic eye movements: Evidence for a control of the neural pulse generator by local feedback. *Biological Cybernetics* 39: 87–96.

Keller EL and Robinson DA (1971) Absence of a stretch reflex in extraocular muscles of the monkey. *Journal of Neurophysiology* 34: 908–919.

Kording KP and Wolpert DM (2004) Bayesian integration in sensorimotor learning. *Nature* 427: 244–247.

Lauwereyns J, Watanabe K, Coe B, and Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. *Nature* 418: 413–417.

Maschke M, Gomez CM, Ebner TJ, and Konczak J (2004) Hereditary cerebellar ataxia progressively impairs force adaptation during goal-directed arm movements. *Journal of Neurophysiology* 91: 230–238.

Nowak DA, Timmann D, and Hermsdorfer J (2007) Dexterity in cerebellar agenesis. *Neuropsychologia* 45: 696–703.

Optican LM (2005) Sensorimotor transformation for visually guided saccades. *Annals of the New York Academy of Sciences* 1039: 132–148.

Optican LM and Quaia C (2002) Distributed model of collicular and cerebellar function during saccades. *Annals of the New York Academy of Sciences* 956: 164–177.

Packard MG and McGaugh JL (1992) Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: Further evidence for multiple memory systems. *Behavioral Neuroscience* 106: 439–446.

Robinson DA (1975) Oculomotor control signals. In: BachyRita P and Lennerstrand G (eds.) *Basic Mechanisms of Ocular Motility and Their Clinical Implications*, pp. 337–374. Oxford, UK: Pergamon.

Smith MA and Shadmehr R (2005) Intact ability to learn internal models of arm dynamics in Huntington's disease but not cerebellar degeneration. *Journal of Neurophysiology* 93: 2809–2821.

Takeichi N, Kaneko CR, and Fuchs AF (2005) Discharge of monkey nucleus reticularis tegmenti pontis neurons changes during saccade adaptation. *Journal of Neurophysiology* 94: 1938–1951.

Thiele A, Henning P, Kubischik M, and Hoffmann KP (2002) Neural mechanisms of saccadic suppression. *Science* 295: 2460–2462.

Todorov E (2005) Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system. *Neural Computation* 17: 1084–1108.

Todorov E and Jordan MI (2002) Optimal feedback control as a theory of motor coordination. *Nature Neuroscience* 5: 1226–1235.

Troyer AK, Black SE, Armilio ML, and Moscovitch M (2004) Cognitive and motor functioning in a patient with selective infarction of the left basal ganglia: Evidence for decreased non-routine response selection and performance. *Neuropsychologia* 42: 902–911.

Tsien JZ, Huerta PT, and Tonegawa S (1996) The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory. *Cell* 87: 1327–1338.

VanGemmert AW, Teulings HL, and Stelmach GE (2001) Parkinsonian patients reduce their stroke size with increased processing demands. *Brain and Cognition* 47: 504–512.

Vaziri S, Diedrichsen J, and Shadmehr R (2006) Why does the brain predict sensory consequences of oculomotor commands? Optimal integration of the predicted and the actual sensory feedback. *Journal of Neuroscience* 26: 4188–4197.

Wolpert DM, Ghahramani Z, and Jordan MI (1995) An internal model for sensorimotor integration. *Science* 269: 1880–1882.

Wolpert DM, Miall RC, and Kawato M (1998) Internal models in the cerebellum. *Trends in Cognitive Sciences* 2: 338–347.

Computational Methods

T J Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Computational neuroscience uses techniques from computer science and applied mathematics to simulate and analyze computational models of neurons and neural systems at many levels of investigation. Digital computers have continued to increase in speed, making it possible to approach more complex neural systems. The number of investigators using computational tools is expanding, and a variety of new journals, summer schools, and scientific conferences that focus on computational neuroscience have proliferated (Table 1). The second edition of a comprehensive handbook on brain theory is available, and there are 33 specialized articles covering many aspects of computational neuroscience in this encyclopedia. Comprehensive introductions to the methods used to develop computational models are also available. This article provides an overview and a discussion of the major methods and techniques in computational and theoretical neuroscience.

The brain regulates behavior by gathering, storing, and accessing information. The brain solves computational problems with specialized circuits, which evolved to process information rapidly and efficiently, in contrast with digital computers, which can be programmed to solve many different types of algorithms inefficiently. For example, the retina is dedicated to visual transduction and image processing and cannot be reprogrammed to process sounds. Because of the close coupling between structure and function in a dedicated system, the anatomy and physiology of a brain region provide important clues to its function. The connectivity between neurons and their properties is shaped by the environment during development and remain plastic even in adulthood. Thus, as the brain processes information, it changes its own structure in response to the information. This plasticity is important in allowing brains to respond flexibly to a changing world through adaptation and learning.

One of the major constraints on the evolution of brains has been their high energy cost. In humans, 20% of the energy budget is consumed by the brain, which constitutes 2% of the body weight. The energy constraint has consequences for coding strategies and communications protocols.

Brain Models

Brain models used as an adjunct to experimental techniques have several advantages: (1) models provide intuition about the possible behaviors of complex, dynamical brain systems, especially when they are nonlinear and have feedback loops; (2) the predictions of a model make explicit the consequences of the underlying assumptions, and comparison with experimental results can lead to new insights and discoveries; and (3) the results of difficult experiments can be simulated with a model, such as reversible lesions of selected channels or neurons, to optimize the design of the experiment for distinguishing between competing explanations.

There are three different types of brain models. The first type, called an interpretive model, is used to analyze experimental data in order to determine whether they are consistent with a particular computational assumption. For example, Apostolos Georgopoulos has used a vector-averaging technique to compute the direction of arm motion from the responses of a population of cortical neurons, and William Newsome and his colleagues have used signal detection theory to analyze the information from single cortical neurons responding to visual motion stimuli. In these examples, the computational model was used to explore the information in the data but was not meant to be a model for the actual cortical mechanisms. Nonetheless, these models were highly influential and have provided new ideas for how the cortex may represent sensory information and motor commands. These models in turn have affected experimental design, which has then led to improved models.

A second type of model, called a confirmatory model, has been used extensively to test whether a set of data can account for the phenomena being studied. In many biophysical experiments, such as the classic Hodgkin–Huxley studies of the squid action potential, sets of data are collected under a variety of conditions, and a model is later constructed to integrate the data into a unified framework. A highly realistic Monte Carlo model of synaptic transmission in the chick ciliary ganglion recently predicted that the majority of vesicles in the nerve terminal were not released at active zones, identified by presynaptic and postsynaptic specializations, but ectopically where they stimulated extrasynaptic receptors; this surprising prediction was subsequently confirmed experimentally. This type of model is most effective when most of the variables in the model have been measured experimentally and

Table 1 Resources for computational neuroscience

Summer schools and conferences
Computation and Systems Neuroscience (COSYNE) (February)
Okinawa Computational Neuroscience Summer Course (June)
Canadian Summer School in Computational Neuroscience (June)
Cold Spring Harbor Laboratories Summer Course on Computational Neuroscience: Vision (July)
Computational and Neural Systems (CNS) Conference (July)
Telluride Neuromorphic Engineering Workshop (July)
Advanced European Course in Computational Neuroscience (August)
Summer Course on Methods in Computational Neuroscience at the Marine Biological Laboratory in Woods Hole (August)
Frankfurt Institute for Advanced Studies Summer School in Theoretical Neuroscience and Complex Systems (August)
Goettingen Course on Computational Neuroscience (September)
Neural Information Processing Systems Conference and Workshop (December)
Selected journals
Journal of Computational Neuroscience (Springer Netherlands)
Neural Computation (MIT Press)
Neural Networks (Elsevier)
Network: Computation and Neural Systems (Taylor and Francis)
Biological Cybernetics (Springer Verlag)
Computer programs
MCell: Monte Carlo models of subcellular chemical signaling
GENESIS: realistic compartmental models of neurons and networks
NEURON: realistic compartmental models of neurons and networks
NSL: neural simulation language for large-scale models of neural systems
PDP++: parallel distributed processing models based on abstract neural networks

only a few unknown parameters need to be fit to the experimental data. One danger with this approach is that even if the model fits the data, the resulting model may not be unique. However, automated techniques have been developed for systematically exploring large parameter spaces to determine all combinations of parameters that fit the data. As the number of experiments increases, the number of possible solutions that fit all the data should converge to a unique set.

Finally, a third type of model starts with a general principle and produces an abstract model that implements the principle within known biological constraints. These models can be quite fruitful in helping to motivate experiments that might not have been otherwise undertaken. An example of this approach is the model of coupled nonlinear oscillators analyzed by Nancy Kopell, Bard Ermentrout, and others, which has led to new experiments on fictive swimming in the lamprey spinal cord. One of the strengths of this type of model is that it can be used to identify the critical variables that determine the qualitative behavior of a system. One corresponding weakness is that because much of the fine detail is often absent, it may not be possible to make detailed comparisons with data.

Dynamical systems theory provides a global analysis of neural systems, but is mainly applicable when the numbers of parameters and variables are small. Most models of neural networks involve a large number of variables, such as membrane potentials, firing rates, and concentrations of ions, with an even greater number of unknown parameters such as synaptic strengths, time constants, and conductances. In the limit that the number of neurons and parameters is very large, techniques from condensed matter physics become applicable in predicting the average behavior of large networks. There is a midrange of systems where neither type of limiting analysis is possible, but where simulations can be performed. One danger of relying solely on computer simulations is that they may be as complex and difficult to interpret as the biological system itself.

Compartmental Models

At the cellular and molecular levels, significant advances have taken place in modeling specific neurons and synapses based on biophysical measurements of ionic mechanisms. Markov models that are used to model ionic channels can be applied to every aspect of synaptic signaling, including transmitter release and the intracellular second messenger systems that modulate synaptic transmission. The original Hodgkin–Huxley models for the fast sodium and delayed rectifier potassium channels are special cases of a Markov model, as are the detailed biophysical models of receptor kinetics.

The pioneering work of Wilfrid Rall on the electrical properties of dendrites was based on the analysis of simplified dendritic geometries. It is now possible to simulate multicompartment models of dendrites from the geometries of reconstructed neurons. Voltage-dependent sodium and calcium channels have been observed in the dendrites of cortical neurons, which greatly increases the complexity of synaptic integration. The experimental confirmation first predicted by compartmental models that active currents can carry information in a retrograde direction from the cell body up to the distal synapses also has theoretical significance for Hebbian forms of synaptic plasticity. Another intriguing observation made with modeling techniques is that the wide variety of spiking patterns in cortical neurons can be reproduced from the same distribution of ionic channels by varying only the geometry of the dendritic tree.

Network Models

Realistic models with several thousand cortical neurons can be explored on the current generation of workstation, which allows the dynamics of cortical columns to be explored in detail. The first model for the orientation specificity of neurons in the visual cortex was the feedforward model proposed by Hubel and Wiesel, which assumed that the orientation preference of cortical cells was determined primarily by converging inputs from thalamic relay neurons. Experimental evidence now favors this model over other models in which the orientation specificity was determined by local cortical circuits. Simulations of orientation columns have shown that the intrinsic circuits in the cortex could be used to increase feature selectivity, amplify weak signals, and suppress noise as well as to perform gain control to extend its dynamic range. This is an important step forward in understanding the function of visual cortex.

A good example of an abstract model that provided a conceptual framework for understanding experimental results is the strongly nonlinear Hopfield network, which exhibits attractor dynamics. This model can be analyzed and convergence proven from any initial state to one of several possible attractor states. Realistic network models of spiking neurons with similar self-sustaining dynamics are now used to study delay-period activity in prefrontal cortical areas and head direction cells in the hippocampus.

Although thalamic neurons that project to the cortex are called relay cells, they almost surely have additional functions since the visual cortex makes massive feedback projections back to them. As an example of a speculative model that has led to a new computational hypothesis for the thalamus, Crick has proposed that the relay cells in the thalamus may be involved in visual attention, and has provided an explanation for how this could be accomplished based on the anatomy of the thalamus. This searchlight model of attention and other hypotheses for the function of the thalamus are being explored with computational models, and new experimental techniques are being used to test these models. Detailed models of thalamic networks can reproduce the low-frequency oscillations observed during sleep states, when feedback connections to the thalamus affect the spatial organization of the rhythms.

Learning Models

The precise timing of single spikes carries information about the locations of objects for echolocating bats and electric fish and for visual motion in invertebrates. Pyramidal neurons are capable of initiating spikes with high precision, and spike-timing-dependent synaptic plasticity (STDP) has been found in the hippocampus and cerebral cortex to be sensitive to timing difference between pre- and postsynaptic events in the millisecond range. Although spike timing has also been suggested as an information code for the cerebral cortex, experimental evidence instead points toward a role for temporal patterns of spikes in regulating the flow and storage of information in the cortex, such as top-down control of attention and expectation.

Classical conditioning occurs on a timescale of seconds, and recent recordings from dopamine neurons in the ventral tegmental area of monkeys has confirmed the predictions of temporal difference (TD) reinforcement learning, which hypothesizes that the dopamine neurons signal the prediction of future reward. The dopamine neurons project diffusely throughout the cortical mantle and the basal ganglia, where they influence actions and regulate synaptic plasticity. TD learning takes advantage of temporal order to encode causal relationships. It has also been used to model honeybee foraging and is consistent with STDP, which suggests that it may be a general computational principle in nature.

Technology for Brain Modeling

Simulations of large-scale neural systems currently are performed on fast workstations and multiprocessor supercomputers. Computer programs for performing these simulations are freely available (Table 1). New technology, however, is needed to scale up simulations from thousands to millions of neurons. A new approach to massively parallel models has been introduced by Carver Mead, who uses subthreshold complementary metal-oxide semiconductor (cMOS) very large-scale integrated (VLSI) circuits with components that mimic the analog computational operations performed by the brain. For example, silicon chips that model the visual processing found in retinas have been built. Analog VLSI cochleas have also been built that can analyze sound in real time. These chips use analog voltages and currents to represent the signals and are extremely efficient in their use of power compared to digital VLSI chips. A new branch of engineering called neuromorphic engineering has arisen to exploit this technology.

Protocols have been designed for long-distance communication between analog VLSI chips that use the equivalent of all-or-none spikes in the same way that long-distance communication between neurons is accomplished. Many of the design issues that govern the evolution of biological systems also arise in these neuromorphic systems, such as the trade-off in cost between short-range connections and expensive long-range communication. Computational models

that quantify this trade-off and apply a minimization procedure can predict the overall organization of topographical maps and columnar organization of the cortex.

The increase in computer memory capacity and advances in machine learning algorithms have greatly increased our ability to automatically segment serial thin sections from electron microscopy and to reconstruct neural circuits. The goal of connectomics is to produce complete connectivity diagrams. Although connectivity cannot give a complete picture including the temporal dynamics of neural circuits, there are many questions that can be answered, such as whether the neurons in the brain are organized to minimize wire length.

Concluding Remarks

Although brain models are becoming increasingly accepted into neuroscience as tools for interpreting data and generating hypotheses, we are still a long way from having explanatory theories of brain function. For example, despite the relatively stereotyped anatomical structure of the cerebellum, we still do not understand its computational functions. Although the cerebellum is thought to be involved in motor coordination, functional imaging suggests that the cerebellum is also involved in higher cognitive functions such as attention and language. Modeling studies may help in exploring competing hypotheses and suggesting a new type of explanation, such as a common computational operation used in motor coordination, attention, and language. This has already occurred in the oculomotor system, which has a long tradition of using control theory models to guide experimental studies.

At this stage in our understanding of the brain, a model should be considered only a provisional framework for organizing thinking. Many partial models need to be explored at many different levels of investigation, each model focusing on a different scientific question. As computers become faster, and as software tools become more flexible, computational models should become increasingly important in analyzing and interpreting experimental data as well as helping to design critical experiments to test alternative hypotheses.

See also: Axonal Pathfinding; Consciousness: Theoretical and Computational Neuroscience; Executive Function and Higher-Order Cognition: Computational Models; Hippocampus: Computational Models; Hodgkin–Huxley Models; Memory: Computational Models; Spike-Timing-Dependent Plasticity Models.

Further Reading

Anderson CH and Eliasmith C (2003) *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press.

Arbib AM (2003) The Handbook of Brain theory and Neural Networks. 2nd edn. Cambridge, MA: MIT Press.

Chen BL, Hall DH, and Chklovskii DB (2006) Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences of the United States of America* 103: 4723–4728.

Churchland PS and Sejnowski TJ (1992) *The Computational Brain*. Cambridge, MA: MIT Press.

Coggan JS, Bartol TM Jr., et al. (2005) Evidence for ectopic neurotransmission at a neuronal synapse. *Science* 39: 446–451.

Crick F (1984) Function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 81: 4586–4590.

Dayan P, Abbott LF, and Abbott (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.

Kistler WM and Gerstner W (2002) *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge, UK: Cambridge University Press.

Koch C and Segev I (1996) *Methods in Neuronal Modeling: From Synapses to Networks*, 2nd edn. Cambridge, MA: MIT Press.

Laughlin SB and Sejnowski TJ (2003) Communication in neuronal networks. *Science* 301: 1870–1874.

Mainen ZF and Sejnowski TJ (1996) Influence of dendritic structure on firing pattern in model neocortical neurons. *Nature* 382: 363–366.

Mead C (1989) Analog VLSI and neural systems. Reading, MA: Addison Wesley Publishing Company.

Prinz AA, Billimoria CP, and Marder E (2003) An alternative to hand-tuning conductance-based models: Construction and analysis of data bases of model neurons. *Journal of Neurophysiology* 90: 3998–4015.

Rieke F, Warland D, de Ruyter van Steveninck R, and Bialek W (1996) *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Segev I, Rinzel J, and Shepherd GM (1995) *The Theoretical Foundation of Dendritic Function: Selected Papers of Wilfrid Rall with Commentaries*. Cambridge, MA: MIT Press.

Steriade M, McCormick DA, and Sejnowski TJ (1993) Thalamocortical oscillations in the sleeping and aroused brain. *Science* 262: 679–685.

Computational Neuroethology

H J Chiel, Case Western Reserve University, Cleveland, OH, USA
R D Beer, Indiana University, Bloomington, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

Definitions

Ethology is the study of behavior that allows animals to survive and reproduce in natural environments (i.e., adaptive behavior). Neuroethology is the study of the neural basis of adaptive behavior. Computational neuroethology is the use of modeling and simulation to study the neural control of adaptive behavior. In contrast, the broad area of computational neuroscience generally focuses on modeling microscopic properties of neurons (e.g., channel biophysics, dendritic spines), single neurons (e.g., multiple conductance models of complex neurons and their morphology), neural circuits, and the dynamics of neural circuitry, without reference to biomechanics or environmental phenomena.

History

In 1990, both Randall D Beer and Dave Cliff independently called for the creation of a new field, computational neuroethology. Beer succinctly summarized the rationale for the new field:

> The working assumptions of computational neuroethology can be summarized as follows: (1) the ability to flexibly cope with the real world is a defining characteristic of intelligent behavior, and more fundamental than conscious deliberation; (2) adaptive behavior derives from a structural congruency between the dynamics of an intelligent agent's internal mechanisms and the dynamics of its external environment; (3) modeling the neural control of behavior in simpler whole animals will provide insights into the nature of the dynamics required for adaptive behavior, and eventually lead to an understanding of the successive elaborations of these mechanisms which are observed in higher animals.

Recently, Cliff has published a review of this area that provides an excellent overview and a summary of key examples of computational neuroethology, with an emphasis on research that has focused on artificial agents.

Conceptual Framework

The conceptual framework of computational neuroethology draws on classical ethology, ecological psychology, biomechanics, and dynamical systems theory. The first key concept of the framework (see Figure 1) is that behavior is embodied, that is, it is not purely due to neural commands but emerges from the ongoing interaction of the central nervous system with a complex sensory and mechanical system. For example, in response to motion in the periphery of its visual system, a monkey will move both its eyes and head, thus using its motor system to reposition its sensory structures. As another example, the optimal angle for throwing a soccer ball from over the head is determined not simply by Newtonian mechanics but by the mechanical advantage that human arms can generate. More generally, embodiment implies that the central nervous system does not 'stand above' the sensory and motor system in the process of responding to the environment; rather, preprocessing by sensory structures, active perception (movement of sensory structures as part of the perceptual process), and the filtering of motor commands by muscles and body structures are crucial for generating adaptive behavior.

The second key concept of the framework is that behavior is situated within a particular environment. For example, changes in optic flow as a fly approaches a surface allow the fly to trigger a rapid landing response. Recently, investigators used optic flow to control a flying robot. Another example of the situated nature of behavior is provided by studies showing that animals change leg stiffness during running to match the elasticity of the surface. In humans, activation of muscles changes depending on whether pointing movements occur with or against gravity. More generally, the situated nature of behavior implies that it is not solely due to the actions of the agent but emerges from the ongoing interactions of the agent with its environment.

Thus, the third key concept of the framework is that adaptive behavior emerges from the coupling of the nervous system, body, and environment. For example, in a virtual reality environment, it is possible to predict human behavior in response to goals and to obstacles by modeling goals as attractors and obstacles as repellers with a relatively simple damped mass spring model for the heading error relative to each object. Sophisticated routes that avoided obstacles and reached goals emerged from this analysis without explicit path planning. As another example, the behavior of a hexapod robot with insectlike reflexes emerged from its ongoing interaction with soft, irregular surfaces.

The centrality of coupled dynamics for adaptive behavior is not surprising. Evolution cannot select for brains or bodies in isolation. Rather, the likelihood

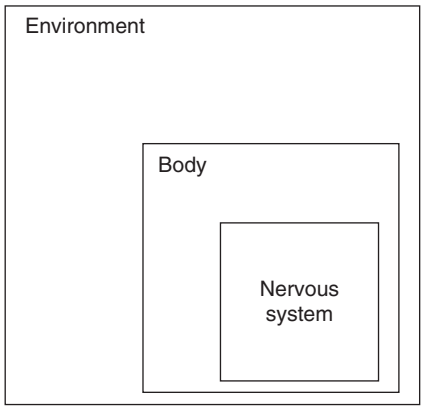

Figure 1 A schematic illustration of the framework of computational neuroethology. Nervous systems are embodied within the biomechanics of a body, and animals are situated within an environment. Adaptive behavior emerges from the coupling of the dynamics of the nervous system, the body, and the environment.

that an animal will survive or leave more offspring than other competitors depends on that animal's ability to coordinate its nervous system and body effectively within a particular environment. Thus, evolution selects an overall coupling of the dynamics of the components that effectively generates adaptive behavior.

Finally, an implication of the three concepts is that understanding adaptive behavior requires the analysis of the dynamical properties of the coupled system. Within each component – the nervous system, the body, and the environment – are complex processes that unfold over time according to rules, and thus each component is a dynamical system. Since each of the systems is coupled to one another, they are nonautonomous, but the entire coupled system may be viewed as an autonomous dynamical system.

The conceptual framework has several important implications. First, it suggests that stable behavioral patterns can be viewed as attractors within the dynamical system defined by the entire coupled system. Second, given that environments change over time, the conceptual framework suggests that understanding the transient structure of the coupled system may be as important as understanding its possible steady state behaviors. Third, any slow dynamical process that allows an organism to improve its interaction with its environment may be adaptive; thus, learning may not be a distinct process but may simply be these slower dynamical processes.

Goals and Methodology

Computational neuroethology has been pursued with two distinct, albeit complementary, goals. From the perspective of a biologist, computational neuroethology is another approach to understanding the mechanisms of adaptive behavior in animals. Thus, the models and analyses are in the service of understanding general principles that may illuminate the adaptive behavior of a single animal or a particular group of animals. The value of the results will depend on how well they explain behavior in animals, generate testable hypotheses, and suggest general principles for the function of similar organisms. More abstract models may be of great utility if they provide general principles that could apply to many organisms. For example, a theoretical approach for all forms of locomotion – constructal theory – has recently been described that provides very general principles for running, flying, and swimming.

Investigators whose primary focus is in the area of computer science or engineering study computational neuroethology with different goals. For these investigators, computational neuroethology provides a new methodology for developing principles for intelligent behavior, as well as a means of designing adaptive artificial devices that have the flexibility of animals. From this perspective, the models and analyses are in the service of understanding adaptive behavior of any agent, whether it corresponds to a biological organism or not. As a consequence, the value of the results will be the insight they provide into a general understanding of adaptive behavior and intelligence or their utility for artificial devices such as robots to function in complex, open-ended environments.

Distinguishing these two different goals is important. A robot that can avoid obstacles using an algorithm based on studies of the escape response in the cockroach may be of great utility, and thus a success from the point of view of engineering. From a scientific viewpoint, however, the success of the robot demonstrates that the proposed algorithm could actually work in the physical world but does not prove that it does work that way in the biological system. Thus, an engineering success may not be the same as a scientific proof. Conversely, it is important for investigators interested in animals to be willing to develop models that may omit or greatly simplify biological details in the search for general principles. A highly detailed model of insect flight may be so computationally intensive and so complex that the general principles of operation, such as the importance of nonstationary aerodynamics or the relative importance of inertial and viscous forces, may be obscured.

These observations have led investigators to pursue computational neuroethology using a variety of models, ranging from highly abstract to highly realistic. Abstract agents provide the opportunity to confront the most general questions about the mechanisms of adaptive behavior. The study of 'frictionless brains'

can provide insights similar to the study of frictionless planes in physics. All features of the agent are known, unlike a biological organism. Careful choices of models of the environment, the body, and the neural network controlling the agent can generate systems that are tractable to detailed mathematical analysis, providing deep insights into the underlying dynamics of adaptive behavior. For example, a study of evolved neural networks for locomotion elucidated the importance of the mechanics of the body for many features of the evolved pattern generators. However, it may be difficult to make quantitative predictions from abstract agents to any particular animal.

Realistic models provide the ability to directly connect model predictions with experimental results. Quantitative predictions can be made about the effects of specific perturbations. Realistic models have drawbacks, however. In addition to the large number of parameters that must be fit by experimental data, the 'brittleness' of models in response to small errors in parameters, and the high computational cost associated with simulating realistic models, it is often difficult to understand or derive general principles from models if they rival the original biological system in complexity.

Thus, some of the most effective work in computational neuroethology has been done by the use of a bracketing methodology, in which both abstract and highly detailed models are studied, with the goal of developing models of intermediate complexity that retain some analytical tractability but whose predictions can be tested experimentally. For example, studies of the mechanisms of insect locomotion have made use of simplified 'template' models of running or climbing movements, which have then been compared to highly detailed 'anchor' models based on the biology of particular animals (e.g., cockroach or gecko).

Investigators of computational neuroethology also debate the value of simulations versus robotic models. On one hand, simulations can be reconfigured rapidly. Combinations of different parameter values can be systematically explored, and if the model as a whole needs to be restructured, this can also be done relatively quickly. On the other hand, if an investigator wishes to ensure that the agent is interacting with the environment, a robot provides the actual physics of the world for free and ensures that any algorithm that works well in simulation can also work well in the presence of fallible sensors, unreliable actuators, and a complex, changing environment. But robots can be expensive to build and difficult to quickly reconfigure, and if they fail to function properly, the investigator needs to ensure that the failure is relevant to the original hypothesis. For example, if a robot is built to examine a perceptual problem but uses wheels rather than legs to negotiate its environment, the investigator needs to be sure that the question of interest (whether biological or theoretical) is unaffected by this change in the periphery. Once again, combined studies of simulations and robotics are likely to be more effective than a sole reliance on one or the other.

Examples

This section provides an overview of some recent work in computational neuroethology, but it is not exhaustive. Rather, the examples illustrate some of the key aspects of the conceptual framework, as well as the models that have been generated, from highly abstract to highly specific; the animals and behaviors that have been studied, ranging from invertebrates to humans; and the fruitful interplay of simulation and robotics.

Embodiment: Biomechanics and Neural Control

A combined mechanical and neural model of walking in the cat revealed that the mechanics of the body played a crucial role in stabilizing or destabilizing gait, depending on the coordination rule used. To study the role of sensory reflexes in locomotion, investigators created a realistic model of the mechanics of the hindlimbs of the cat and a simplified finite state model of the neural control that was driven by sensory inputs from the model body. One mechanism for triggering the transition of a leg from supporting weight on the ground (the stance phase) to moving forward before initiating the next step on the ground (the swing phase) is the removal of load from the limb as other limbs begin supporting the body's weight. The force in the ankle extensor muscle monitors the unloading of the limb. When this rule was used alone to drive walking gaits, the resulting movements were stable even in response to perturbation. The reason was that if the other limbs were not yet taking up load, the limb that was about to switch from stance to swing was delayed by this mechanical input, and this stabilized the gait. In contrast, when the extension of the hip was used alone as the signal to terminate stance, the resulting gaits were unstable because the mechanical system tended to increase the hip extension independent of whether the body could afford to lift that limb. This model provided strong evidence for mechanical coupling through the body as a stabilizing mechanism for locomotion. Further support for this view is provided by the recent work of investigators who have used a dynamic periphery and a simple neural network to create a fast bipedal

walking robot (RunBot) that takes advantage of peripheral biomechanics to stabilize its gait and can improve its walking with experience.

A neuromechanical model of feeding in the marine mollusk *Aplysia californica* made it possible to predict the roles of mechanical reconfiguration during feeding. Investigators created a kinetic model of the feeding apparatus and showed that the change in shape of the central grasper from spherical to ellipsoidal changed the forces exerted by a separate muscle, a protractor muscle that wraps around the grasper and was stretched by the grasper's change in shape. As a consequence, the protractor muscle became stronger, both because of its position on its length–tension curve, and because of its enhanced mechanical advantage, allowing it to generate stronger protractions. Subsequently, *in vitro* and *in vivo* studies showed that the model predictions were correct. These results have been used to define the concept of neuromechanical modulation – the ability of one set of motor neurons to affect the actions of a different set of motor neurons through the changes in the mechanics of the periphery.

These results emphasize the importance of understanding motor neuronal activity within the context of biomechanics. More generally, it is clear that motor behavior is best understood within an embodied context.

Situatedness: Coupling among the Environment, the Periphery, and the Nervous System

A striking illustration of the importance of the coupling among the environment, the body, and the nervous system was provided by a simulation of lamprey swimming. The neuromechanical model performed well without sensory feedback in water flowing at a constant velocity. In contrast, in the absence of sensory feedback, when the model encountered a barrier consisting of water flowing at a higher velocity, the model lamprey failed to swim across the barrier. When the model lamprey was provided with appropriate sensory feedback but no other changes were made to the neural circuit, the model could swim effectively through the barrier. The faster flowing current induced rapid sensory inputs that induced a faster oscillation of the body, which in turn allowed it to move through the faster water. Thus, a new and effective behavioral pattern emerged from the interaction of all the components.

Another example of the coupling of the environment and the body to generate adaptive behavior is provided by the studies of insect flight. A robotic fly that operated in mineral oil, whose wing dimensions and viscous interactions with the fluid surrounding it were scaled to have similar Reynolds number to those of the fruit fly, was used to demonstrate that delayed stall due to translational motion of the wings was insufficient to account for the ability of *Drosophila* to fly. Rotation of the wings during stroke reversals was also critical, allowing the insect to take advantage of rotational circulation and wake capture. These additional forces not only enhanced lift but also provided the insect with the ability to change direction and steer. Recent studies have focused on the steering mechanisms underlying rapid right-angle turns (body saccades) in fruit flies, using three-dimensional, high-speed infrared video to capture the actual motions of the wings and body of the flies and using the dynamically scaled robot to examine the underlying forces responsible for the turns. Surprisingly, they found that inertial forces dominated turning behavior due to subtle tilts of the stroke plane angle and small changes in stroke amplitude.

These results emphasize the importance of understanding behavior through detailed studies of the interactions of biomechanics and environmental forces, which in turn may both simplify and limit what the nervous system can do at every instant.

Active Perception

Perception within the context of ongoing behavior, or active perception, has been one of the most heavily investigated areas of computational neuroethology. Physical constraints on active sensing systems, as well as the need for appropriate sensorimotor coordination, determine whether systems can simply react to events or can anticipate them. Combining models of electrosensory and mechanosensory stimuli with prey strike trajectories performed by electric fish (recorded using infrared video), investigators demonstrated the relative contributions of different frequency components of the electric sense and mechanosensory inputs on prey capture. Recently, investigators have developed a simple robot capable of underwater localization of targets using electrolocation.

A variety of other active perceptual processes have been studied. Cricket phonotaxis, the ability of a female cricket to locate a male cricket by its calling song, has been studied using a series of increasingly complex and realistic robot and neural network models. Investigators recently demonstrated that a model neural circuit driving a legged robot could successfully detect and locomote to sound sources outdoors. Using a slightly modified model of human navigation, investigators controlled a differential drive robot and showed that it negotiated obstacles

more smoothly than did previous methods. A robotic whisker array demonstrated that moments induced by whisker contact with objects or flows could be used for accurate reconstruction of the inputs.

These results demonstrate that when sensing is both embodied and focused on environmentally relevant features, complex perceptual feats can be accomplished efficiently.

Cognition and Intelligence

Computational neuroethology provides a novel approach for creating intelligent behavior. Rather than relying primarily on an abstract analyzer and manipulator of symbolic representations of the world, it emphasizes that highly complex and adaptive behavior, including behavior that is cognitive, emerges from the interaction of brain, body, and environment. For example, investigators have created a series of models of specific behavioral competencies – frog prey capture, rat spatial navigation, monkey saccades and grasping, and human language – which are inspired by known anatomy and neurophysiology. Investigators have argued that the successive elaboration of functional schemas provides insights into prey capture, predator avoidance, negotiation of obstacles, recognition and categorization of one's location in egocentric space relative to external space, the ability to imitate motor actions, and ultimately the ability to recognize, imitate, and generate 'combinatorially open' abstract motor sequences that can become the basis for language.

A complementary approach is creating agents that have 'minimally cognitive' capabilities, such as the ability to categorize objects, to focus attention, to manipulate objects, or to signal to other agents. By making the agents as simple as possible, but capable of an arguably cognitive behavior, investigators can analyze in detail the mechanisms underlying their behavior. Investigators have pursued this line of inquiry using genetic algorithms as a means of searching for agents that produce cognitive behaviors (e.g., categorical perception, selective attention, associative learning) so that the neural circuits are not hand designed, which could introduce *a priori* biases into their function, and then analyzing how they work. Recent work analyzed an agent evolved to perform categorical perception: the agent could catch circular objects and would avoid diamond-shaped objects. The agent's behavior could be explained in terms of the dynamics of its interaction with the environment and the dynamics of its nervous system, but no 'representations' of either circles or diamonds were found. These results suggest that a dynamical systems approach to embodied and situated agents may well provide insights into the nature of cognition and intelligence in biological organisms, and in artificial devices, without recourse to abstract symbol manipulation.

Summary and Conclusions

Although the field of computational neuroscience is still relatively young, it has become increasingly important for understanding adaptive behavior. As computers become faster and cheaper, it is more feasible to create sophisticated models that incorporate features of the environment, of biomechanics, and of neural control. Successful analyses of complex adaptive behavior using computer simulation and robotics have encouraged investigators to adopt the methodology of computational neuroethology. At the same time, simulated organisms and robots capable of solving challenging real-world problems have emerged from this research, suggesting that it is a fruitful approach for generating artificial devices that have the flexibility of biological organisms.

What is most exciting about recent work in computational neuroethology is both the growing empirical support for the larger conceptual framework and the progress in clarifying general principles for adaptive behavior. Thus, it is clear that many aspects of motor behavior can be understood only by examining neural activity within its appropriate biomechanical and environmental context. Similarly, the process of perception involves active and ongoing engagement of the organism with its environment, so perception must be studied within both a biomechanical and an environmental context. Finally, cognition and intelligence must be understood within a behavioral context and emerge from an embodied and situated agent, not a disembodied pure reasoner.

See also: Animal Intelligence: The Search for Animal Intelligence; Artificial Intelligence; Auditory Localization; Consciousness: Theories and Models; Consciousness: Theoretical and Computational Neuroscience; Executive Function and Higher-Order Cognition: Computational Models; Executive Function and Higher-Order Cognition: Definition and Neural Substrates; Pattern Generation; Referentiality and Concepts in Animal Cognition.

Further Reading

Arbib MA (2003) *Rana computatrix* to human language: Towards a computational neuroethology of language evolution. *Philosophical Transactions of the Royal Society of London A – Mathematical Physical and Engineering Sciences* 361: 2345–2379.

Beer RD (in press) Beyond control: The dynamics of brain-body-environment interaction in motor systems. In Sternad D (ed.) *Progress in Motor Control V: A Multidisciplinary Perspective*. Berlin: Springer.

Beer RD (2003) The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior* 11: 209–243.

Chiel HJ and Beer RD (1997) The brain has a body: Adaptive behavior emerges from interactions of nervous system, body, and environment. *Trends in Neurosciences* 20: 553–557.

Cliff D (2002) Neuroethology, computational. In: Arbib MA (ed.) *The Handbook of Brain Theory and Neural Networks,* 2nd edn., pp. 737–741. Cambridge, MA: MIT Press.

Full RJ and Koditschek DE (1999) Templates and anchors: Neuromechanical hypotheses of legged locomotion on land. *Journal of Experimental Biology* 202: 3325–3332.

Holmes P, Full RJ, Koditschek D, and Guckenheimer J (2006) The dynamics of legged locomotion: Models, analyses, and challenges. *Society for Industrial and Applied Mathematics Review* 48: 207–304.

Nelson ME and MacIver MA (2006) Sensory acquisition in active sensing systems. *Journal of Comparative Physiology A* 192: 573–586.

Pearson K, Ekeberg O, and Buschges A (2006) Assessing sensory function in locomotor systems using neuro-mechanical simulations. *Trends in Neurosciences* 29: 625–631.

Webb B (2006) Validating biorobotic models. *Journal of Neural Engineering* 3: R25–R35.

Conditioned Reflex

D S Woodruff-Pak, Temple University, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

The Discovery of the Conditioned Reflex

The Nobel laureate physiologist Ivan Petrovich Pavlov discovered the conditioned reflex in the first decade of the twentieth century. The initial work involved the autonomic nervous system with Pavlov's use of the salivating reflex in dogs. He discovered that presenting a neutral stimulus such as a bell that never normally elicited salivation could come to elicit salivation if presented repeatedly before giving the dog meat powder. The learned response (called the 'conditioned response') was the dog's salivation to the bell. Presentation of the bell followed by presentation of the meat powder led to the formation of an association. Throughout the century following the discovery of classical (Pavlovian) conditioning, it has become one of the most widely employed paradigms for the investigation of associative learning and memory and its neurobiological substrates. Pavlov was the first scientist to observe that old dogs condition more slowly than young dogs. Thus, research on classical conditioning in normal aging has a long and illustrious history.

Classical Conditioning

The standard format for the presentation of stimuli in classical conditioning is called the delay procedure. A neutral stimulus such as a tone, light, or tactile vibration is called the conditioned stimulus (CS). It is presented for a fixed duration (optimally around half a second) before the presentation of a reflex-eliciting stimulus, called the unconditioned stimulus (US). In the delay procedure, the CS and US overlap and typically end together. Learning occurs when the organism responds with the reflex to the CS before the onset of the US. The learned response is called the conditioned response.

A second format for the presentation of stimuli in classical conditioning is called the trace procedure. In this case, the neutral stimulus that is the CS is presented and then turned off before the US is presented. There is a blank period of time, called the 'trace,' between the CS offset and US onset. Pavlov thought that in the trace conditioning procedure, a 'memory trace' had to be formed in the blank period between CS offset and US onset. The two major classical conditioning procedures, delay and trace, are illustrated in Figure 1. Although there are a number of additional formats for classical conditioning, virtually all the studies of conditioned responses in normal aging have used the delay or trace classical conditioning procedures.

Classical Conditioning of the Eyeblink Response

Whereas Ivan Pavlov discovered the phenomenon of classical conditioning using a conditioned reflex in the autonomic nervous system, the conditioned reflexes used in contemporary studies of classical conditioning, such as the conditioned eyeblink response, are controlled in the central nervous system. Skeletal muscle innervated by the cranial nerves controls the eyeblink. Eyeblink classical conditioning is a common paradigm for investigations of classical conditioning in general and classical conditioning in normal aging in particular. Fear conditioning, in which a neutral CS is paired with a moderately painful US, is another promising model system that has been elaborated extensively on a behavioral and neurobiological level. Studies of aging effects on fear conditioning have been reported, but the results are not consistent. The magnitude of age-related fear responses appears to be small, especially in comparison with age-related changes in eyeblink classical conditioning.

The neural circuitry that supports classical conditioning of the eyeblink response is almost completely mapped, and the behavioral and neurobiological parallels in this form of associative learning extend to all mammals that have been studied, including humans. Processes of normal aging affect eyeblink classical conditioning similarly in all species in which older organisms have been tested – mice, rats, rabbits, cats, and humans.

For several decades, the promise of eyeblink classical conditioning as a model system for the study of learning and memory in aging has been evident. Among the significant advantages of this model system for investigating processes of normal aging are that (1) age differences in the classically conditioned eyeblink responses are large and (2) striking parallels exist between the age differences in eyeblink conditioning in nonhuman mammals and humans. The initial parallels in normal aging and eyeblink conditioning were demonstrated between rabbits and humans. On no other species is there such a large body of parametric data on classical conditioning as on the classically conditioned eyeblink response in the rabbit. Indeed, much of the general literature on classical conditioning is based on data collected in the rabbit with the nictitating membrane (third eyelid of the rabbit) response and in the human with the

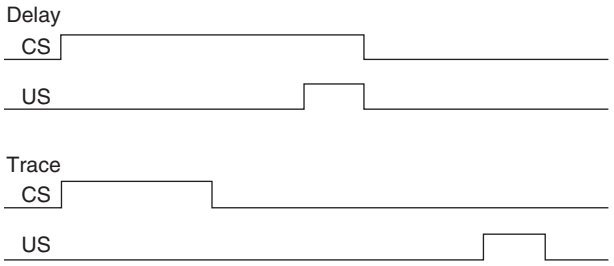

Figure 1 (Top) Delay classical conditioning paradigm in which the conditioned stimulus (CS) and unconditioned stimulus (US) overlap. (Bottom) Trace classical conditioning paradigm in which the CS is turned off before the US is turned on. The blank period between CS offset and US onset is called the 'trace' period.

eyeblink. Now techniques for behavioral and electrophysiological testing in rats, including neonatal rats, have been developed, along with techniques to test eyeblink conditioning in normal, mutant, and transgenic mice and rodent slice preparations.

Ernest Hilgard was the first to study eyeblink classical conditioning in animals and did classic studies on human eyeblink conditioning as well. His work established the close correspondence in properties of the conditioned eyeblink response in humans and others animals, suggesting that the underlying neuronal mechanisms of memory storage and retrieval are the same in all mammals, including humans. Here, perhaps more than in any other form of learning, neuronal mechanisms of memory elucidated in infrahuman mammals apply directly to the human condition. Isadore Gormezano was the first to publish eyeblink conditioning studies in the rabbit and to introduce measurement of the nictitating membrane extension response.

Initial research on aging and classical conditioning studied rabbits or humans. However, there are advantages of other species, such as mice and rats, for research on aging. In particular, mice and rats, with their short gestation period, large litter size, relatively small space requirements, and short life spans, are highly desirable for research on development and aging. Neither the genetic controls nor the breeding environments presently available for rabbits equal those in place for the mouse and rat. Consequently, many of the recently published studies on eyeblink classical conditioning, including those on aging, involve rodent subjects.

Eyeblink Conditioning in Aging Mammals

Eyeblink classical conditioning displays natural age-related deficits in all species in which it has been tested, including humans, cats, rabbits, rats, and mice. The generalization first observed in humans, that age-related deficits appear in middle rather than old age, holds for the species that have been tested over the adult life span: mice, rabbits, and humans.

Rodents Mice and rats are a useful species for research on normal aging. They have the advantage of small space requirements and short life spans. Breeding colonies maintained by the National Institute on Aging facilitate research on aging by making a number of strains of rodents, at all points of the life span, available to researchers.

Techniques for rodent eyeblink conditioning were initially developed in rats, and age differences in eyeblink classical conditioning in rats were first document in John Disterhoft's laboratory. Jeansok Kim, Lu Chen, and Shaowen Bao, in Richard F Thompson's laboratory, developed eyeblink conditioning initially in the C57BL/6 mouse. This group demonstrated that the mouse preparation exhibited normal associative learning and memory. The C57BL/6J strain was used to start the C57BL/6NIA barrier-reared strain used extensively for research on aging and also commonly used to build transgenic mouse strains. Research on normal aging and eyeblink conditioning in mice has been carried out exclusively with the C57BL/6 strain. The first published study of eyeblink conditioning in aging C57BL/6 mice was carried out in Kirino's laboratory, and age-related impairment was observed in 20-month-old mice in the delay procedure. The CS was a 1 kHz, 85 dB sound pressure level tone. Interpretation of those results was problematic because C57BL/6 mice have age-related hearing impairment beginning about mid-life, with thresholds at all frequencies greatly elevated in the second year of life. C57BL/6 mice begin to exhibit progressive, high-frequency, sensorineural hearing loss at 3 months of age that extends to middle and lower frequencies by 16 months. Some studies have used a white noise CS so that the frequency spectrum would be covered up to around 15 months of age. This strategy makes the CS clearly audible to C57BL/6 mice in the younger age groups, but severe hearing losses in C57BL/6 mice have been reported between 16 and 23 months of age. CS intensity affects the rate of acquisition, and perceived CS intensity is undoubtedly lower in C57BL/6 mice older than 16 months. Age-related hearing loss accounts for some portion of the observed age-related impairment in eyeblink classical conditioning in the older groups.

In other species, such as rabbits and humans, hearing deficits have been ruled out as the cause of age-related impairment of eyeblink conditioning, but those species do not suffer the magnitude of hearing loss that affects C57BL/6 mice. In future studies of aging in mice, the use of CSs in other sensory

modalities and the use of different mouse strains that do not demonstrate significant age-related hearing loss will clarify normal aging effects on eyeblink conditioning.

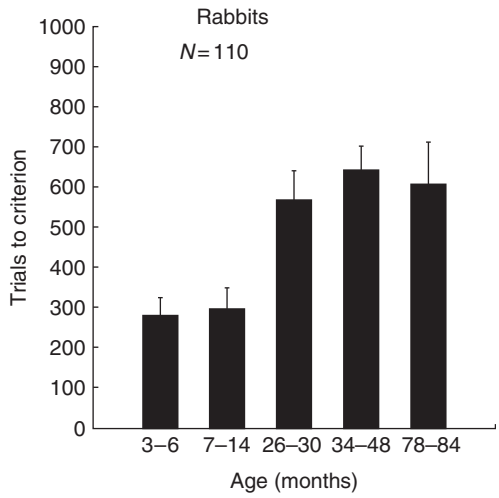

Figure 2 Age differences in the 750 ms-delay eyeblink classical conditioning procedure in 110 rabbits ranging in age from 3 to 84 months. There was statistically significant age-related impairment in trials to a learning criterion of eight conditioned responses in nine consecutive trials between the ages of 21 and 30 months with no significant additional decline in the older groups.

Rabbits The rabbit has some major advantages as an animal model in classical conditioning research because of its size and, especially, its docile nature. Rabbits tolerate well the restraint required in classical conditioning studies. In some laboratories, research on eyeblink conditioning and aging has occurred long enough so that hundreds of rabbits of various ages have been run. Using a delay paradigm in which the interval between the CS and US was 750 ms, 110 rabbits ranging in age from 3 to 54 months in Diana Woodruff-Pak's laboratory showed significant age-related impairment in trials to learning criterion between the ages of 21 and 30 months, with no significant additional decline in the older groups (see Figure 2). These results were paralleled closely in John Disterhoft's laboratory using the 600 ms short CS trace paradigm. In 180 rabbits older than the age range of 2–36 months, age-associated learning impairments were seen by 24 months of age. Comparisons of young and older rabbits in 600 ms and 750 ms delay and trace eyeblink classical conditioning paradigms in Woodruff-Pak's laboratory indicated that older rabbits performed significantly worse in the 750 ms than in the 600 ms delay paradigm, but they performed as well as young rabbits in the 600 ms delay paradigm. Both age groups of rabbits were more impaired in the trace paradigm, with older rabbits equally impaired in the 600 ms and 750 ms trace paradigm (Figure 3).

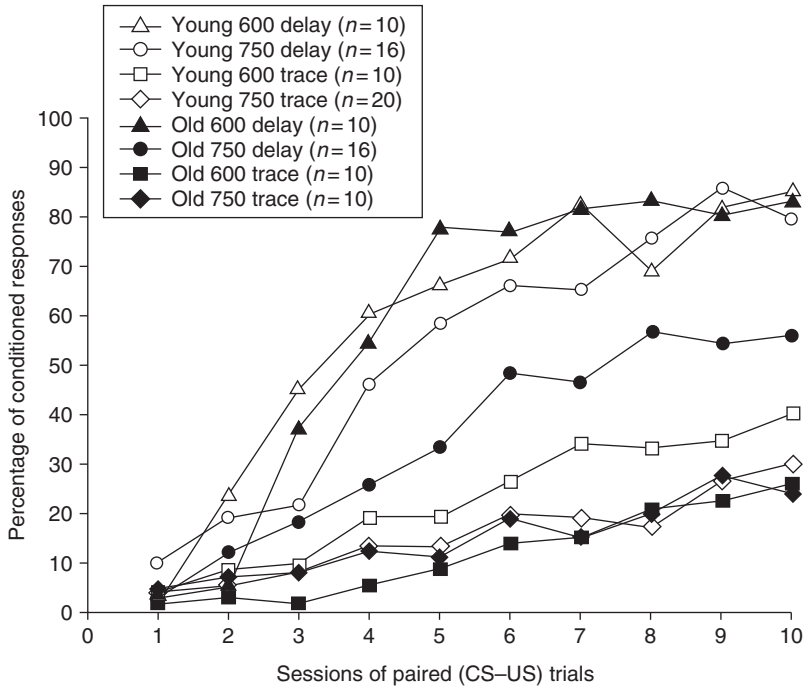

Figure 3 Age differences between 4-month-old and 26-month-old rabbits in delay and trace eyeblink classical conditioning paradigms with a 600 ms or 750 ms interval between the CS and US (called the interstimulus interval, ISI). There were statistically significant effects of age, ISI, paradigm, and training session.

Since rabbits have a life expectancy that has been estimated to be about 8 years, these animals may be useful as animal models for studies of brain and behavior relationships in aging. Assuming that human life expectancy is almost 80 years, it is tempting to make a linear extrapolation from the human to the rabbit life expectancy and equate a human decade with a rabbit year. In this scheme, a 60-year-old human would be equivalent to a 6-year-old rabbit. However, there are a number of reasons that a linear interpolation of human to rabbit life expectancy probably is inappropriate. For example, at birth the rabbit is much more mature than the human, and rabbits reach puberty by 6 months. According to the linear extrapolation between human and rabbit age, a 6-month-old rabbit would be the equivalent of a 5-year-old human. However, since 5-year-old children cannot reproduce, the linear interpolation between human and rabbit age is inaccurate in early life. Using reproductive capacity as a biomarker, a 6-month-old rabbit is roughly comparable to an adolescent aged 12–14 years. The age of 3 years in rabbits might compare to the human age decade of the 50s, and humans in their 60s might correspond in age to 4-year-old rabbits. Age changes in reproductive capacity appear to be a useful means to parallel human and rabbit age, and age differences in learning and memory occur at similar times in the life course of these two species when reproductive capacity is used as a biomarker to assess age.

Humans The first research on aging and human eyeblink conditioning was carried out in the early 1950s by the Russian scientists Gakkel and Zinina and reported in the United States by Jerome in the first *Handbook of the Psychology of Aging* in 1959. Differences between eyeblink classical conditioning in young and older nursing home residents in a single 70–90-trial session were reported. These results were replicated and extended to normal, community-residing older adults. The main and striking result was the relative inability of the older participants to acquire the conditioned eyeblink response.

The early studies of eyeblink conditioning in normal aging in humans involved comparisons of college students and older volunteers. The first study to investigate eyeblink conditioning over the adult life span was carried out in Woodruff-Pak's laboratory in the 400 ms delay eyeblink classical conditioning procedure. The 103 participants represented the adult life span between the ages of 18 and 83 years. Participants received a series of 108 trials; 96 trials were paired presentations of the tone CS and corneal air puff US, and 12 trials were presentations of the tone CS alone. Before the session, neutral instructions were given that were designed neither to facilitate nor to inhibit eyeblinks. Results replicated earlier reports that age differences in acquisition of conditioned eyeblink responses were large. A novel finding was that age differences in eyeblink conditioning occurred early in adulthood, in the age decade of the 40s. These results were replicated almost identically in Paul Solomon's laboratory. Donald Powell's laboratory reported age differences in both eyeblink and heart rate classical conditioning.

This result was extended in the Woodruff-Pak laboratory to demonstrate that the effects of normal aging in eyeblink conditioning appear in humans in the age decade of the 40s and increase in the 50s but that there is no further statistically significant impairment in the age decades of the 60s through the 80s.

Brain Substrates of Eyeblink Classical Conditioning

The essential site of the plasticity for learning resides in the ipsilateral cerebellum, and the hippocampus, while not essential, can affect the rate of conditioning. Significant changes occur in the cerebellum and in the hippocampus during normal aging. It may now be possible to identify and characterize the critical age-related changes in the brain underlying the age-associated decline in eyeblink classical conditioning.

Involvement of the Cerebellum in Associative Learning

A variety of techniques, including electrophysiological recording of multiple and single units, electrolytic and chemical lesions, physical and chemical reversible lesions, neural stimulation, genetic mutations, and pharmacological manipulation, have been used to demonstrate that the cerebellum is essential for all forms of eyeblink classical conditioning. More specifically, the dorsolateral cerebellar interpositus nucleus ipsilateral to the conditioned eye is the essential site for acquisition and retention of conditioned eyeblink responses. The cerebellar cortex is also of demonstrated involvement during normal acquisition. Rabbits with aspirations of cerebellar cortex or mutant mice that lose all Purkinje neurons acquire conditioned responses at a dramatically slowed rate and to a lower maximal performance level than do intact organisms. Tests of patients with lesions in various parts of the cerebellum, along with functional imaging studies of the brain during eyeblink conditioning, have demonstrated that the cerebellum's role in eyeblink conditioning is essential in humans as well as in rabbits and other nonhuman mammals.

Involvement of the Hippocampus in Eyeblink Classical Conditioning

The hippocampus and septohippocampal acetylcholine system are involved in basic associative learning of the sort represented by eyeblink conditioning. In a series of studies using delay eyeblink classical conditioning in rabbits, it was observed that activity recorded in the CA1 pyramidal cell region of the hippocampus forms a predictive 'model' of the amplitude–time course of the learned behavioral response, but only under conditions in which behavioral learning occurs. This response is generated largely by pyramidal neurons. The role of the hippocampus in classical conditioning of the eyeblink response is called 'modulatory' because manipulations of the hippocampus can impair or enhance the rate of acquisition. The memory trace itself is not in the hippocampus, but the hippocampus can markedly influence the storage process.

Although the hippocampus is not necessary for normal acquisition in the delay procedure, it is necessary for trace conditioning. A long-lasting neuronal plasticity is formed in the hippocampus following eyeblink conditioning. This change is essential for learning to occur in the trace eyeblink conditioning procedure, at least until the learning is consolidated. Age-related electrophysiological changes in the hippocampus have been associated with age-related impairment in trace eyeblink conditioning.

Normal Aging in the Hippocampus and Cerebellum

Neuron numbers vary widely among individuals, in part as a consequence of chance developmental variations. The neuron cell loss required for functional impact appears to differ widely among brain regions, and the magnitude of neuron loss in normal aging also varies among brain regions. Until recently, it was widely accepted that neuron death in most brain regions was an inevitable result of normal aging. With the development of stereological methods for counting neurons and the application of these techniques to research on aging, it has become apparent that in the neocortex and hippocampus of humans and many other species, there is little age-related decline in neuron number.

Although there is stability of neuron number in the older hippocampus, electrophysiological investigations of aging in the mammalian hippocampus have revealed that some aspects are compromised. In learning-impaired older rabbits, single-unit records of clusters of pyramidal neurons in CA1 showed diminished responding. Age-related impairments in trace eyeblink conditioning involve an increased after-hyperpolarization in hippocampal pyramidal neurons due in turn to an increased calcium-activated potassium current.

Whereas neuron number is maintained in neocortex and hippocampus, there is significant Purkinje neuron loss associated with aging in the cerebellar cortex. For example, we used unbiased stereology to assess the total number of Purkinje neurons in cerebellar cortex over the life span of C57BL/6 mice. Statistically significant age differences in Purkinje neuron number were observed beginning at 18 months. Similar Purkinje neuron loss in late middle age has been reported in other mammalian species. In addition to significant neuronal loss, functional capacity of older Purkinje neurons is also impaired. Purkinje neuron number has been associated with the rate and magnitude of delay eyeblink classical conditioning in young as well as older organisms. The normal age-related impairment and loss of Purkinje neurons is associated with age-related deficits in delay eyeblink classical conditioning.

Resolution in humans using techniques such as magnetic resonance imaging (MRI) is not fine enough to reveal presence or absence of individual neurons – not even the largest neurons in the brain – Purkinje neurons. Anatomical MRI studies of cerebellar volume over the adult life span indicated significantly smaller cerebellar volume by the age of 50 years. Assessment of cerebellar volume with MRI in humans tested with eyeblink classical conditioning demonstrated that age-related changes in cerebellar volume are highly correlated with age-related deficits in eyeblink classical conditioning.

Amelioration of Eyeblink Classical Conditioning in Older Organisms

A successful means of ameliorating impaired eyeblink conditioning in older rabbits is the use of cognition-enhancing drugs. Several types of drugs are of documented efficacy in this model system. Drugs that promote the availability of acetylcholine, such as acetylcholinesterase inhibitors, are effective in facilitating eyeblink conditioning. In particular, drugs that target nicotinic acetylcholine receptors have shown great efficacy in eyeblink conditioning in young as well as older rabbits. Another drug mechanism that has ameliorated impaired eyeblink conditioning in older rabbits is the blockade of calcium channels. Older organisms have increases in calcium influx in hippocampal pyramidal neurons that apparently impair learning, and drugs that block calcium channels can improve the function of these neurons.

See also: Aging of the Brain; Aging: Brain Potential Measures and Reaction Time Studies; Conditioned Taste Aversion; Eyeblink Conditioning; Procedural Learning: Classical Conditioning; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Chen L, Bao S, Lockard JM, Kim JJ, and Thompson RF (1996) Impaired classical eyeblink conditioning in cerebellar lesioned and Purkinje cell degeneration (*pcd*) mutant mice. *Journal of Neuroscience* 16: 2829–2838.

Christian KM and Thompson RF (2004) Neural substrates of eyeblink conditioning: Acquisition and retention. *Learning & Memory* 11: 427–455.

Disterhoft JF and Oh MM (2006) Pharmacological and molecular enhancement of learning in aging and Alzheimer's disease. *Journal of Physiology, Paris* 99: 180–192.

Kishimoto Y, Suzuki M, Kawahara S, and Kirino Y (2001) Age-dependent impairment of delay and trace eyeblink conditioning in mice. *Neuroreport* 12: 3349–3352.

Luft AR, Skalej M, Schulz JB, et al. (1999) Patterns of age-related shrinkage in cerebellum and brainstem observed *in vivo* using three-dimensional MRI volumetry. *Cerebral Cortex* 9: 712–721.

Oswald B, Knuckley B, Mahan K, Sanders C, and Powell DA (2006) Prefrontal control of trace versus delay eyeblink conditioning: Role of the unconditioned stimulus in rabbits (*Oryctolagus cuniculus*). *Behavioral Neuroscience* 120: 1033–1042.

Powell DA, Buchanan SL, and Hernandez LL (1981) Age-related changes in classical (Pavlovian) conditioning in the New Zealand albino rabbits. *Experimental Aging Research* 7: 453–465.

Power JM, Wu WW, Sametsky E, Oh MM, and Disterhoft JF (2002) Age-related enhancement of the slow outward calcium-activated potassium current in hippocampal CA1 pyramidal neurons *in vitro*. *Journal of Neuroscience* 22: 7234–7243.

Solomon PR and Groccia-Ellison M (1996) Classic conditioning in aged rabbits: Delay, trace, and long-delay conditioning. *Behavioral Neuroscience* 110: 427–435.

Thompson LT, Moyer J, James R, and Disterhoft JF (1996) Trace eyeblink conditioning in rabbits demonstrates heterogeneity of learning ability both between and within age groups. *Neurobiology of Aging* 17: 619–629.

Woodruff-Pak DS (2006) Stereological estimation of Purkinje neuron number in C57BL/6 mice and its relation to associative learning. *Neuroscience* 141: 233–243.

Woodruff-Pak DS and Steinmetz JE (eds.) (2000) *Eyeblink Classical Conditioning Vol I: Applications in Humans*. Boston: Kluwer Academic Publishers.

Woodruff-Pak DS and Steinmetz JE (eds.) (2000) *Eyeblink Classical Conditioning Vol II: Animal Models*. Boston: Kluwer Academic Publishers.

Woodruff-Pak DS, Vogel RW III, Ewers M, Coffey J, Boyko OB, and Lemieux SK (2001) MRI-assessed volume of cerebellum correlates with associative learning. *Neurobiology of Learning and Memory* 76: 342–357.

Wu WW, Oh MM, and Disterhoft JF (2002) Age-related biophysical alterations of hippocampal pyramidal neurons: Implications for learning and memory. *Ageing Research Review* 1: 181–207.

Conditioned Taste Aversion

M R Foy and J G Foy, Loyola Marymount University, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Conditioned taste aversion (CTA) is a learned association of taste and visceral distress. CTA occurs when an animal learns to avoid a newly encountered taste after suffering adverse postingestive effects from a noxious substance. From an evolutionary perspective, CTA is thought to prevent animals from eating poisonous substances. From an experimental perspective, CTA has long been considered an important learning paradigm in which neural events have been correlated with different phases of learning, including acquisition, retention, and extinction. Recent advances have been made in understanding the neural pathways involved in CTA learning, and how this learned behavior is affected by the normal aging process. The aim of this article is to provide a brief review of the major components of the CTA paradigm, the neural pathways, the biochemical and genetic factors implicated in CTA, and the effects of aging related to CTA.

The Conditioned Taste Aversion Paradigm

CTA has been known by zoologists and others since the late 1800s and was introduced into the laboratory for intensive investigation in the mid-1950s. The CTA protocol is highly reproducible, and the memory of associations made by the pairing of relevant stimuli in CTA conditioning is long-lasting, which permits analysis of multiple memory phases and systems. In a typical CTA experiment, animals are originally exposed to tap water for drinking, and are adapted to a shortened fluid session (e.g., 30 min per day). During the conditioning phase, a novel taste (saccharin- or sucrose-flavored water) becomes the conditioned stimulus (CS), the substance that produces adverse effects (lithium chloride, LiCl) is the unconditioned stimulus (US), and the behavior following exposure to the US (nausea or behavioral malaise) is the unconditioned response (UR). Thirty minutes of sodium saccharin (0.1% solution) drinking, followed by an intraperitoneal injection of LiCl (0.15 M, 2% body weight) from minutes to hours after exposure to the CS, elicits in rats severe gastrointestinal distress that is followed by strong aversion to saccharin on the next day. In a two-bottle preference test, the animal is given the opportunity to choose between the familiar (water) and novel (saccharin) fluid over an extended period of time (20-day test period) following the single CS–US conditioning procedure. The benefit of using a two-bottle preference test is that the animal is not forced to drink the novel fluid (as opposed to a single-bottle technique, wherein the thirsty animal is offered only the novel fluid, and thus its consumption is dependent upon its motivation and how palatable it finds the novel fluid). Furthermore, the neophobic avoidance of novel stimuli (calculated as the ratio of the volume of the novel fluid to total fluid intake) can be identified more readily in a two-bottle test. In Figure 1(a), a schematic diagram of a typical CTA learning paradigm is shown.

Studies using CTA have found that some animals show less preference for the novel substance prior to poisoning (taste neophobia) than do others. The length of the interval between exposure to the novel CS substance and the noxious US agent (delay) can be manipulated, with longer delays associated with stronger CTA in some animals than in others. Also, the number of trials required to extinguish the aversive behavior shows variability: extinction rates are faster under conditions of deprivation and previous exposure to the preferred substance (familiarity). Furthermore, researchers have identified significant sex, strain, and species differences in CTA conditioning, with several of these aspects of CTA learning having relevance to aging.

Although often considered a form of classical conditioning, CTA does not fit neatly into this classification, nor does it share all of the characteristics of instrumental learning, suggesting that CTA may actually be a unique type of learning. Several critical temporal properties that distinguish CTA from other forms of associative learning include the fact that CTA learning can be acquired following a single pairing of the CS and US stimuli. In other forms of associative learning, multiple CS–US presentations are necessary for conditioning to occur. Some have argued this to be a 'biological preparedness' component of CTA learning, since illness is most easily associated with taste and the association can be made in one trial.

A second parameter unique to CTA is the wide range of interstimulus intervals (ISIs) used in CTA experiments. During CTA training, a CS–US interval of several hours can be tolerated, whereas in other forms of associative learning (e.g., classical and instrumental conditioning), ISIs longer than several seconds become ineffective in establishing conditioned responses. However, like other forms of learning, CTA shows a delay of reinforcement gradient, and conditioning strength can be decreased when the US is presented much longer (e.g., >12 h) after the CS.

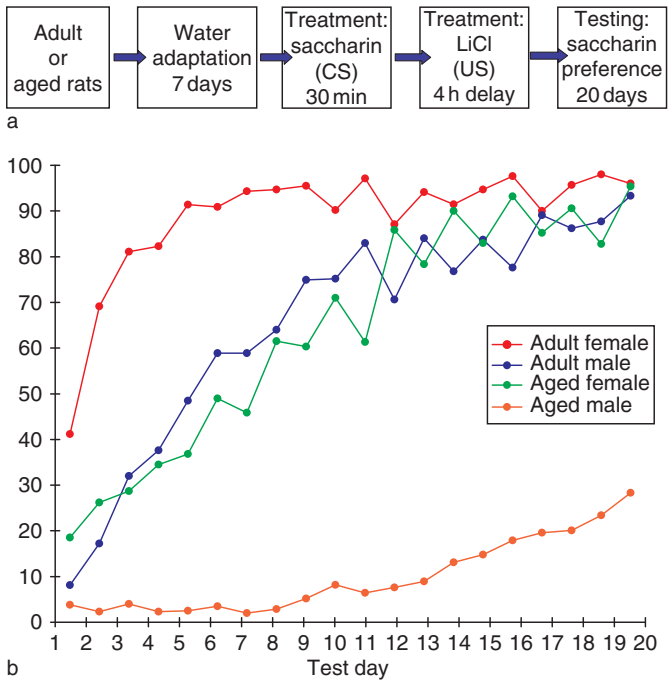

Figure 1 (a) Schematic diagram of a typical conditioned taste aversion learning paradigm. In this design, rats are adapted to a 30 min tap water drinking schedule over 7 days. On the conditioning day, all rats are administered the conditioning stimulus (CS; 0.1% sodium saccharin solution) during the normal 30 min water session, followed 4 h later by a single unconditioned stimulus (US; 0.15 M LiCl, 20 ml kg^{-1}, intraperitoneal injection). For 20 days, the rats are then given daily two-bottle preference tests: familiar tap water and novel sodium saccharin solution during the 30 min fluid sessions. (b) Saccharin preference scores for adult (3 months old) male and female rat groups, along with aged (24 months) male and female rat groups. Two-bottle saccharin preference scores for adult male and female rats treated with LiCl after a 4 h delay show a significant sex difference across the 20-day test period between the two groups. Adult males had a slower extinction rate (stronger conditioned taste aversion as measured by lower saccharin preference) compared to adult females (faster extinction rate and higher saccharin preference). This sex difference is very pronounced in the aged group; aged male rats show a much slower extinction rate compared to aged females. Adapted from Foy MR and Foy JG (2003) Reversal of long-delay conditioned taste aversion learning in rats by sex hormone manipulation. *Integrative Physiological and Behavioral Science* 38: 203–213, and Foy MR, Foy JG, and Thompson RF (2005) Aging and delayed conditioned taste aversion learning in rats. *Pavlovian Society Meeting.* Anaheim, CA, 29 September 2005.

Several theories have been proposed to suggest that associative learning mechanisms are responsible for the formation of immediate taste aversions, in which animals are made ill soon (e.g., within minutes) after consuming a novel substance. However, for long-delay learning in which the CS–US interval is quite long (e.g., hours), a nonassociative theory has been proposed to employ the nonassociative learning mechanisms of habituation and sensitization (dishabituation) in combination with neophobic avoidance of novel stimuli.

The use of the CTA paradigm has provided significant advantages for investigating the phenomena and mechanisms of simple associative learning. Single-trial acquisition facilitates the correlation of neural events with learning, and targeted interventions using anatomical and pharmacological procedures can be readily examined in CTA. For example, one promising line of CTA research in rats has identified a sexual dimorphism found in long-delay (4 h) CTA learning. Figure 1(b) illustrates the sex differences in long-delay CTA learning in adult and aged male and female rats that have been correlated with the presence of circulating sex hormones.

Neural Circuitry Implicated in CTA

Research on neural circuitry and related structures involved in CTA learning originates largely from studies of the effects of reversible and irreversible brain lesions, and disruptions of protein synthesis within specific structures during the acquisition and retention of CTA. Generally, these studies reveal that as many as ten sites are relevant for CTA learning and memory to occur. Brief synopses of the major circuitry and structures involved in CTA are delineated in the following sections.

Area Postrema

The small region known as the area postrema (AP), which lacks a blood–brain barrier, lies on the floor of the fourth ventricle. It is presumed to be the site for

the detection of toxins, and the reflexive action to produce illness. Lesions of the AP before conditioning obliterate avoidance responses in CTA. Lesions of the AP after acquisition of LiCl-induced CTA have no effect on retention of the sucrose aversion, suggesting that the AP is not necessary for long-term aspects of CTA. Evidence thus far supports the view that the area postrema is necessary for the acquisition of CTA induced by LiCl, scopolamine, gamma radiation, and histamine, but not for other noxious agents such as amphetamine, nicotine, ethanol, and rotation-induced illness, for which the sites of action are not yet established. CTAs induced by different unconditioned stimuli may be mediated by different neural mechanisms.

Nucleus Tractus Solitarius

Gustatory information from the facial, glossopharyngeal, and vagus nerves (cranial nerves VII, IX, and X, respectively) converge in the brain stem, within the nucleus tractus solitarius (NTS). Neurons in the NTS project to the pontine parabrachial nucleus (PBN). Lesions within the NTS do not always markedly impair CTA, and rats can learn to avoid noxious substances with lesions to this area. The primary deficit observed with NTS lesions appears to be sensory rather than associative; animals with NTS lesions do not show the typical preference for sweet taste and aversion to unpleasant taste seen in intact animals.

Pontine Parabrachial Nucleus

The pontine parabrachial nucleus (PBN) is involved in the initial integration of gustatory with visceral information. Lesions to the PBN are more disruptive to CTA acquisition and retention than are lesions to any other site. PBN-lesioned rats become incapable of acquiring CTA under conditions of delay between CS and US stimuli. However, if conditioning occurs prior to PBN lesioning, CTA expression is unimpaired. It is strongly suspected that the PBN is necessary for CTA acquisition, but not CTA expression, and that the PBN is involved in associating the taste with illness.

Thalamus

Lesion studies involving the thalamus have produced equivocal effects on CTA. Lesions that spare structures providing olfactory cues may confound the results, suggesting that further research is needed. The general finding is that thalamic lesions, especially those of the medial ventrobasal thalamic complex that include the thalamic gustatory nuclei, are sufficient to prevent CTA learning.

Hypothalamus

Hypothalamic neurons are closely related to the influence of motivational state on behavioral responsiveness to gustatory stimuli. Lesions of hypothalamic nuclei responsible for regulating feeding feedback (feeding center or lateral hypothalamic area, satiety center, or ventromedial hypothalamic nucleus) have facilitatory effects on CTA (slower extinction rates) when rats are maintained at normal weights. Obese rats with ventromedial hypothalamic (VMH) lesions and water-deprived rats also have slower extinction rates.

Globus Pallidus

The globus pallidus (GP) plays an important role in the regulation of feeding behavior. Lesions of the ventromedial region of the GP impair acquisition and retention of CTA and are responsible for stronger and more persistent neophobia. These findings suggest that GP neurons are critical in CTA acquisition and retention and also play a role in the attenuation of neophobia.

Amygdala

In the limbic system, parts of the amygdala appear to be involved in the assessment of taste, as well as in the acquisition of CTA. The effects of amygdala lesions on CTA may be determined by the extent of damage to fibers from the insular cortex to amygdala and some key nuclei. Lesions of the basolateral nucleus prevent CTA acquisition, whereas lesions of the central nucleus appear to have no effect on CTA.

Hippocampus

Hippocampal lesions attenuate, but do not abolish, CTA acquisition. Hippocampectomized animals also have slower extinction rates than do intact animals. It has been proposed that the hippocampus is involved in the evaluation of the novelty or safety of the stimulus.

Prefrontal Cortex

The prefrontal cortex (PFC) is involved, generally, in the organization of goal-directed, adaptive behaviors, and more specifically in neural control of feeding. Studies mostly show that deficits in CTA acquisition and retention are associated with PFC lesions, especially marked when lesions are located in the dorsomedial PFC. The number of cells showing c-Fos protein expression is increased in the prefrontal cortex during extinction. These results suggest that the medial PFC may play a substantial role in the formation, retrieval, and extinction of CTA, although further study is necessary.

Table 1 Summary of neural structures and putative relations to conditioned taste aversion learning and memory

Learning/memory feature	*Neural structures*
Taste detection and assessment	Area postrema; nucleus of the solitary tract; amygdala via facial, glossopharyngeal, and vagal nerves
Neophobia	Hippocampus, globus pallidus
Acquisition of CTA	Pontine parabrachial nucleus, thalamus, globus pallidus, amygdala, prefrontal cortex
Strength of CTA	Insular cortex, amygdala, hippocampus, prefrontal cortex
Retention of CTA	Pontine parabrachial nucleus, globus pallidus, amygdala, prefrontal cortex
Extinction of CTA	Hypothalamus, hippocampus, insular cortex, prefrontal cortex

Insular Cortex

The insular cortex (IC) may play a role in the integration of visceral reactions and stress. Whereas lesions of both the IC and the amygdala abolish CTA acquisition, lesions of the IC alone tend to attenuate acquisition and retention of CTA, not because the animals are incapable of aversive responses, but because they retain this ability. The difficulty caused by IC lesions appears to stem from their inability to associate aversive qualities to previously pleasant tastes. Lesions in the anterior IC are more likely to correspond with CTA deficits than are regions of the posterior IC. Studies have confirmed that the IC is involved in taste processing and memory formation and, particularly, mediation of associative aspects of taste responses. Long-term extinction of CTA is eliminated when new protein synthesis and β-adrenergic transmission in the IC is blocked in the late post-retrieval period. The IC may be involved in the consolidation of memory for CTA and require intact noradrenergic activity into the basolateral amygdala.

The gustatory cortex (GC), a small region of the IC, appears to be involved in the processing, encoding, and storage of motivation-independent taste familiarity information. Lesion studies indicate that the GC is involved in CTA with delay between the taste exposure and the illness, but not under no-delay conditions. Other studies indicate that while the gustatory cortex is involved in the acquisition of CTA, it is not essential for acquisition, but rather for the retention, of the aversion.

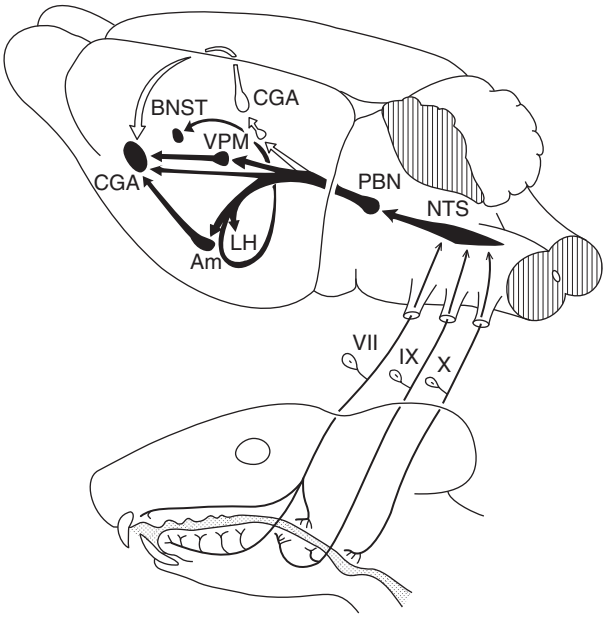

Figure 2 Gustatory afferent pathways in the rat. VII, facial nerve, including the chorda tympani and greater superficial petrosal nerve; IX, glossopharyngeal nerve; X, superior laryngeal nerve belonging to the vagal nerve; NTS, nucleus tractus solitarius; PBN, parabrachial nucleus; VPM, ventral posteromedial nucleus of the thalamus; CGA, insular cortical gustatory area; BNST, bed nucleus of the stria terminalis; LH, lateral hypothalamus; Am, amygdala. Ipsilateral projections, black arrows; contralateral projections, white arrows. Copyright 1989 from *Neural Mechanisms* by Yamamoto T, edited by Cagen RH. Reproduced by permission of Routledge/Taylor &Francis group, LLC.

Anatomical Summary

A cautionary approach to eating or drinking novel substances, even when they taste good, seems prudent and evolutionarily advantageous. Likewise, avoiding substances that have caused illness but which would normally be considered pleasant to taste is adaptive, but only if the substance continues to cause illness. Thus some degree of neophobia, aversion, and extinction constitutes critical features of CTA in the normal individual, beyond, of course, being able to detect tastes and to assess these tastes along pleasant–unpleasant dimensions. Delineating the neural pathways implicated in CTA thus must account for these elements (Table 1), which clearly involve gustatory afferent pathways (Figure 2), as well as pathways involving visceromotor integration with higher-order learning.

In order for CTA to occur, the individual must be able to taste and to assess that taste. CTA thus requires sensory coding and taste discrimination, subserved by neural pathways involving the facial, glossopharyngeal, and vagal nerves to the area postrema, NTS, and amygdala, three areas where the taste is detected and evaluated. Both novelty of taste and how to respond to it require accessing memories

corresponding to the taste and regulating subsequent feeding behavior accordingly. The hippocampus and globus pallidus appear to be involved when the individual evaluates a novel taste and, although finding it pleasant, exhibits cautionary feeding behavior. When a substance makes an individual ill, the substance must be associated with the illness and remembered as being associated, and feeding behavior must be regulated on the basis of these cognitive processes in order for there to be acquisition of a conditioned taste aversion. Neural structures involved in acquisition of CTA include the PBN, thalamus, globus pallidus, amygdala, and prefrontal cortex.

Once a CTA has been established, how much avoidance of the substance the individual engages in appears to be subserved by the insular cortex, amygdala, hippocampus, and prefrontal cortex. There is strong consensus that the processes of retention of a learned response and extinction are distinct, both cognitively and behaviorally. They also appear to involve distinct neural substrates. Retention of CTA appears to involve the PBN, globus pallidus, amygdala, and prefrontal cortex. The hypothalamus, hippocampus, IC, and PFC have been linked with extinction in CTA.

Effects of Aging on CTA Learning

An extensive literature associates normal aging with mild, but significant, changes in learning and memory function. Many of these studies employ techniques that examine molecular, cellular, and behavioral changes associated with the aging process. It remains unclear, however, whether these age-dependent declines are related to an increasing deficit in learning, to reduced influences of contextual cues, or to whether the performance on learned tasks deteriorates with age and thereby obscures a relatively unimpaired ability to learn.

The CTA paradigm has been employed in developmental studies of learning because it minimizes performance factors that may confound age comparisons on learning ability, while providing an opportunity for demonstrating an ability to associate different types and qualities of stimuli. Parameters within the CTA protocol, such as the intensity of the CS, intensity of the US, interval between the CS and US, and effects of preexposure can all be systematically modified in order to study the effects of aging. The behavioral measures that can be studied include acquisition of CTA (indicated by a reduced preference for sweet taste compared to preference scores in unconditioned subjects), strength of conditioning (relative preference for the sweet taste over the nonsweet taste), and extinction rates (magnitude or frequency of preference scores relative to a time unit, usually measured in days).

The effects of aging as revealed in the CTA paradigm can be summarized as follows: parameter changes that weaken conditioning tend to highlight age differences. For example, a long CS–US delay (several hours) typically weakens the strength of the conditioned aversion. Aged rats show stronger conditioned aversion differences than do young rats when the CS–US delay is long (e.g., 3 and 6 h); age differences are not as apparent in short-delay paradigms. This effect also appears to be mediated by the strength of the US, which is typically correlated with the strength of the conditioned response. With long delays, age differences are eliminated when the US is relatively intense (e.g., 0.4 M LiCl). Aged rats develop stronger CTAs than do young rats at both short and long delays when the US is relatively weak (e.g., 0.15 M LiCl).

Aged individuals show markedly slower CTA extinction rates compared to younger individuals. The ability to acquire a CTA is not diminished in rats as old as 24 months. In fact, compared to younger rats, aged rats show an early and strong aversion to the novel CS following CS–US conditioning, and a much stronger resistance to extinction to the CS during extinction trials. This superiority is independent of learned safety ability, exposure to relevant contextual cues, physical characteristics associated with aging (such as body weight), and ability to metabolize the US. Referring back to Figure 1(b), typical extinction rates over 20 days in a 4 h delay CTA paradigm are illustrated comparing adult and aged rats. The finding of an age-related increase in resistance to extinction is consistent with those of other aging studies that used instrumental learning tasks, such as shuttle avoidance, maze learning, or operant bar pressing.

Another parameter known to affect CTA learning is preexposure to the CS. Daily preconditioning reduces CTA acquisition in young rats, but not aged rats, whereas a single US preexposure 1 h prior to conditioning reduces CTA acquisition in both young and aged animals. In summary, aged individuals appear to form a strong initial trace in the CTA paradigm that lasts considerably longer, compared to younger individuals.

Mechanisms underlying age differences in CTA learning are not yet well defined or understood. Some authors have speculated that aged rats may be less sensitive to contextual cues, rendering them less vulnerable to interference effects. Another line of research suggests that slowed metabolic processes known to be associated with aging may slow a hypothetical biochemical clock, making stimuli disparate in time subjectively more contiguous, and more

strongly associated in aged rats, compared to young rats in whom long CS–US delays are typically associated with weak CTA traces.

Biochemical and Metabolic Factors Related to Aging and CTA

Aging is associated with decreased adrenal androgen sensitivity to adrenocorticotropin hormone (ACTH), and with decreased concentrations of adrenal androgens and steroids in both men and women. In adult rats, water deprivation and preexposure to the preferred substance (familiarity) typically decrease the rate of extinction independent of acquisition of the CTA, whereas ACTH and testosterone are associated with prolonged extinction rates. Male rats are more neophobic and require more extinction trials than females. These sex differences can be reversed by gonadectomy and restored in gonadectomized animals by hormone replacement. Animals implanted with a testosterone metabolite (e.g., 5α-dihydrotestostoerone) show slower CTA extinction rates, regardless of sex, than do animals implanted with estrogen (17β-estradiol). Elevated testosterone levels in water-deprived male rats (e.g., following injection with corticotropin) are also linked with slower extinction rates.

Genetic Factors Related to Aging and CTA

Long-term memory is dependent on the production of new proteins in the brain via the regulation of gene expression, transcription, and translation mechanisms. By blocking protein synthesis with specific drugs at different time points, it is possible to determine when new proteins are required for long-term taste aversion learning. Blockade of protein synthesis impairs CTA learning and is associated with at least three specific proteins: c-Fos, inducible cAMP element repressor (ICER), and spinophilin. In the CTA paradigm, the antimetabolites 1β-D-arabinofuranosylcystosine (ara-CTP), cytosine arabinoside (ara-C), and the protein synthesis inhibitor anisomycin impaired CTA at long-delay time intervals but not short-delay time intervals (e.g., 1 h).

Substances that interfere with the formation of cAMP response element-binding protein (CREB) transiently reduce the level of CREB protein during training and impaired CTA memory when tested 3–5 days later when injected into the amygdala several hours before CTA training. These substances had no effect on CTA retrieval when administered after CTA memory formation, and did not affect short-term CTA memory.

The central nucleus, which is connected to the IC, has been proposed to be critical for the establishment of long-term CTA memory in the behaving rat. This process involves long-term changes subserved by CRE-regulated gene expression in amygdala neurons, and the amygdala may retain some CTA-relevant information over time, rather than merely modulating the gustatory trace during acquisition of CTA.

Clinical Features of Aging Relevant to the CTA Paradigm

Aging is associated with structural, physiological, and genetic changes that could be involved in CTA learning and memory function. Aging typically reduces sensitivity to taste, with greater quantities of the taste substance required; however, aged rats are more neophobic than are young rats, suggesting that the typical quantities of the US used in CTA paradigms are sufficient for taste detection. Conflicting findings have been reported as to whether aging significantly affects the vagus nerve. Aging is also associated with changes (usually cell loss) to many of the pathways involved in CTA: gastric mucosa and the NST, volume of the prefrontal cortex, insula, anterior cingulate gyrus, superior temporal gyrus, inferior parietal lobule, medial nucleus of the stria terminalis, hippocampus, and precuneus. Aging appears to have little effect on the volume of the amygdala, although impaired amygdala–hippocampus interactions have been found in aged rats with cognitive impairments. Shrinkage of the globus pallidus only occurs in men. The number of neurons in some regions of the hypothalamus (e.g., corticotropin-releasing neurons of the paraventricular nucleus) increases whereas oxytocin neurons in the paraventricular nucleus remain intact during aging. The hypothalamic–pituitary–adrenal (HPA) axis shows reduced responsiveness to glucocorticoid feedback inhibition in human aging.

Conclusions

The neural processes and behaviors associated with CTA learning have been delineated. Aging is linked with strong aversion responses under conditions in which these aversions are weakened in younger individuals. Aging is also associated with slower extinction rates, compared to younger individuals. Aging-associated biochemical and metabolic factors that affect key areas of neural pathways in CTA might be responsible for these results. Systematic investigations of the role that neural structures, hormones, and genes play in behavioral changes associated with CTA

hold promise for elucidating the neural mechanisms associated with behavior, understanding specific processes associated with aging, and developing interventions for age-related cognitive decline.

See also: Conditioned Reflex; Conditioning: Theories; Memory Representation; Procedural Learning: Classical Conditioning; Reward and Learning; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Bernstein IL (1999) Taste aversion learning: A contemporary perspective. *Nutrition* 15: 229–234.

Bures J, Bermudez-Rattoni F, and Yamamoto T (1998) *Conditioned Taste Aversion: Memory of a Special Kind*. Oxford: Oxford University Press.

Chambers KC and Bernstein IL (2003) Conditioned taste aversions. In: Doty RL (ed.) *Handbook of Olfaction and Gustation*, pp. 905–934. New York: Dekker.

DeCola JP and Fanselow MS (1995) Differential inflation with short and long CS–US intervals: Evidence of a nonassociative process in long-delay avoidance. *Animal Learning & Behavior* 23: 154–163.

Desmedt A, Hasvi S, and Dudai Y (2003) Differential pattern of camp response element-binding protein activation in the rat brain after conditioned aversion as a function of the associative process engaged: Taste versus context association. *Journal of Neuroscience* 23: 6102–6110.

Foy MR and Foy JG (2003) Reversal of long-delay conditioned taste aversion learning in rats by sex hormone manipulation. *Integrative Physiological and Behavioral Science* 38: 203–213.

Garcia J, Ervin FR, and Koeling RA (1966) Learning with prolonged delay of reinforcement. *Psychonomic Society* 5: 121–122.

Grigson PS, Shimura T, and Norgren R (1997) Brainstem lesions and gustatory function: II. The role of the nucleus of the solitary tract in Na^+ appetite, conditioned taste aversion and conditioned odor aversion in rats. *Behavioral Neuroscience* 111: 169–179.

Kiefer SW (1985) Neural mediation of conditioned taste aversion. *Annals of the New York Academy of Science* 443: 100–109.

Misanin JR, Collins M, Rushanan S, et al. (2002) Aging facilitates long-trace taste-aversion conditioning in rats. *Physiology and Behavior* 75: 759–764.

Mitchell D, Scott DW, and Mitchell LK (1977) Attenuated and enhanced neophobia in the taste-aversion 'delay of reinforcement' effect. *Animal Learning & Behavior* 5: 99–102.

Revusky SH and Garcia J (1970) Learned associations over long delays. In: Bower GH and Spence JT (eds.) *Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 4, pp. 1–84. New York: Academic Press.

Yamamoto T (1989) Role of the cortical gustatory area in taste discrimination. In: Cagan RH (ed.) *Neural Mechanisms in Taste*, pp. 197–219. Boca Raton: CRC Press.

Yamamoto T, Shimura T, Sako N, et al. (1994) Neural substrates for conditioned taste aversion in the rat. *Behavioral Brain Research* 65: 123–137.

Relevant Website

http://www.ctalearning.com – Conditioned Taste Aversion: An Annotated Bibliography.

Conditioning: Simple Neural Circuits in the Honeybee

R Menzel, Freie Universität Berlin, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

The Antennal Lobe and the Mushroom Body Are Sequentially Involved in the Olfactory Memory Trace

In honeybee olfactory learning, the pathways for the conditioned stimulus (CS, odor) and the unconditioned stimulus (US, sucrose) are well defined. The US pathway is implemented in a single identified neuron, the VUMmx1 neuron. The olfactory pathway (the CS pathway) consists of the olfactory receptor neurons projecting through the antennal nerve to the antennal lobe (AL); the first-order integration neuropil; the olfactory projection neurons (PNs), which connect the AL with the mushroom body (MB); the lateral horn (also called lateral protocerebrum); and other higher-order brain regions. CS and US pathways converge anatomically at three sites: the AL, the lip region of the MB, and the lateral horn. For two of these regions, the AL and the MB, it has been shown with local injections of the putative transmitter of the first ventral unpaired median neuron of the maxillary segment (VUMmx1) as a substitute for the reward stimulus that olfactory memory induces learning by forward pairing (but not by backward pairing). These findings corroborate earlier observations using local cooling as a retrograde amnestic treatment, which documented a cooling-sensitive memory phase for the AL during the first 1–2 min after a single learning trial and such a phase for the MB for the first 5–6 min. The early and late short-term-memory phases after single-trial olfactory conditioning were related to the AL and MB, respectively, on the basis of such findings.

Single Neuron Activity Correlate with Olfactory Learning

The reward neuron VUMmx1 responds to sucrose with long-lasting spike activity and to various visual, olfactory, and mechanosensory stimuli with low-frequency spike activity. Depolarizing VUMmx1 immediately after CS presentation (forward-pairing, CS^+) leads to learning, but backward-pairing does not (CS^-). Thus VUMmx1 constitutes the neural correlate of the US in associative olfactory learning. After differential conditioning in which an odor CS^+ is forward paired with sucrose reward (US) and another odor CS is backward paired with US, presentation of the CS^+ alone activates and CS^- inhibits VUMmx1, indicating that the reward pathway learns about the stimuli, a property that may be the neural substrate of second-order conditioning. In addition, VUMmx1's response to the US after presentation of the CS^+ is greatly reduced, but the response to CS^- remains normal, indicating that the response of VUMmx1 to US after conditioning depends on US expectation. This property of VUMmx1 is sufficient to explain the behavioral phenomenon of blocking and may thus reflect its neural substrate. These results demonstrate that the single identified neuron VUMmx1 is a sufficient neural substrate for the reinforcing function of the US sucrose in olfactory conditioning and has properties that allow explaining second-order conditioning and blocking.

Many other neurons belonging to different tracts leading from the AL toward the MB, or which are extrinsic to the MB, have been tested for associative plasticity by means of the differential conditioning paradigm. PNs connecting the AL with the MB calyx were less clear in CS^+-specific plasticity when intracellular recordings were applied, but recent extracellular multiunit recordings provided compelling evidence for associative plasticity in the minute range after conditioning. Most interesting, particular PNs specifically enhance or reduce their responses to CS^+ in about equal proportions, and a CS^+-specific shift in the power spectrum of local field potentials indicates an increase of precise spike timing for the CS^+ but not for the CS^-. This property may explain findings from optical imaging experiments in which Ca^{2+} activity of the postsynaptic sites of lateral PN neurons in the AL glomeruli was monitored during differential olfactory conditioning. No specific CS^+ and CS^- effects were seen. It appears that the glomerulus as a whole reliably codes odors in a combinatorial pattern together with other glomeruli, whereas single PNs undergo specific associative plasticity. In the case of a balance between associative rise and fall of activity in PNs originating in the same glomerulus, no learning effect will be seen when the whole glomerulus is imaged. It will have to be elucidated in the future whether the associative rise and fall in the rate responses of particular PNs or the precise timing of spikes in particular subsets of PNs represents the learning-related signature in the AL-MB network.

Two other observations support the conclusion that the AL network is indeed involved in an early and short-lasting memory trace. First, Ca^{2+} imaging of glomerular activity after bath application of the Ca^{2+} indicator (Ca-green FM ester) gave consistently higher responses to CS^+ than to CS^-. The fluorescence signal in such a situation may come predominantly from the presynaptic terminals of olfactory receptor

neurons (and possibly also from local interneurons and glia cells), indicating that synaptic processing inside the glomerulus converts CS^+-specific plasticity into both enhancement and reduction of odor responses. Second, the postsynaptic sites of lateral PN neurons within the glomeruli show spontaneous Ca^{2+} fluctuations. After stimulation with an odor, these spontaneous activity fluctuations are more strongly correlated between the glomeruli that had been activated simultaneously by the odor, and the correlations between the spontaneous activity fluctuations suffice to reconstruct the stimulus. It is concluded that these modifiable fluctuations could provide an ideal substrate for Hebbian reverberations and sensory memory in a neural system. It is important to note that all these associative and nonassociative changes seen in the AL glomeruli and PNs last only for a few minutes, corroborating the conclusion that the AL may be involved in the establishment of an olfactory memory trace only shortly after the learning trial.

Two kinds of MB-extrinsic neurons have been studied with respect to associative plasticity, a single identified neuron, the pedunculus-extrinsic neuron 1 (PE1), and neurons in the protocerebral-calycal tract (PCT). The PE1 neuron leaves the alpha lobe of the MB and receives its input across the peduncle of the MB at two bands of putative postsynaptic specializations. PE1 responds to a large range of odors. Differential conditioning leads to a CS^+-specific reduction of odor responses, whereas US-only presentations (sensitization) cause an increase of odor responses. Since the CS^+-specific responses – as recorded intracellularly from isolated bee heads – were lost after a few minutes, it was initially concluded that PE1 may be related to short-term memory. However, recent extracellular recordings from the intact animal lasting many hours have shown that the CS^+-specific response reduction is stable over the lifetime of the recording (several hours). Furthermore, the synapses between the MB-intrinsic neurons (Kenyon cells) and the PE1 were found to undergo long-term potentiation if PE1 was depolarized during the tetanic electric stimulation of the Kenyon cells. It is not yet clear how learning-related response reduction and associative long-term potentiation in PE1 can be reconciled.

CS^+-specific effects – as compared with CS^- or US-only effects – were seen in neurons of the PCT. Many of these neurons are γ-aminobutyric acid immunoreactive and project back from the alpha lobe of the MB to its input site, the calyx. They terminate both at the presynaptic sites of the PNs and the postsynaptic sites of the MB intrinsic neurons, the Kenyon cells. It is conceivable that PCT neurons provide an adapted inhibitory feedback from the output site of the MB to its input. Since PCT neurons also cross between the sensory domains of the MB, they could also be involved in context-specific forms of learning.

Neural Circuits Underlying Olfactory Conditioning: Signaling Cascades Mediating CS and US Stimuli

The different contributions of AL and MB in associative learning are also evident at the level of the molecular machinery implicated in learning and memory formation. Processing CS and US stimuli – even as early as during training – activates different signaling cascades *in vivo*. Stimulation of the antenna with sucrose, the US, induces a short transient activation of the cyclic adenosine monophosphate (cAMP)-dependent protein kinase A (PKA) in the AL, whereas odor stimulation (CS) or mechanical stimulation of the antennae does not affect PKA activity in the ALs. This US-induced PKA activation in the AL is mediated by octopamine, the putative transmitter of the reward neuron VUMmx1. The dense innervation of the VUMmx1 in the AL, the main localization of the PKA in AL interneurons, and the biochemical measurements support a general modulatory function of US-induced processes mediated via the cAMP/PKA cascades in the AL. The specific reinforcing function of octopamine during learning in the AL was supported by silencing the expression of the octopamine receptor which impairs olfactory acquisition but not odor discrimination.

Although the calyces, the olfactory input area of the MB, are also densely innervated by the octopaminergic VUMmx1 neuron, stimulation with sucrose (US) does not lead to PKA activation in the MB calyces *in vivo*. This, and the fact that octopamine can, in principle, stimulate PKA in cultured Kenyon cells of the MBs, supports the idea that octopamine receptors that receive input from VUMmx1 in the MB are most likely coupled to Ca^{2+}-regulated pathways. The signaling cascade mediating the US function in olfactory conditioning in the MB has not yet been identified.

Molecular Signaling Cascades in the ALs Are Critical for the Induction of Olfactory Memory Traces

Direct *in vivo* measurements demonstrate that the temporal characteristics of the US-induced PKA activation in the AL is critically influenced by its pairing with CS stimuli during olfactory conditioning. A single CS/US forward pairing, which induces a weak olfactory memory, leads to a transient increase in PKA activity that returns to basal levels 60 s after

the conditioning trial. Three CS/US forward-pairings in succession, which induce a long-term memory (LTM), prolong PKA activation in the AL up to more than 3 min. In contrast, US stimulation alone and, independent of the number, US/CS backward pairings induce a significantly shorter PKA activation. This close correlation between training parameters on the one hand and the temporal relation between US and CS stimulation, the number of conditioning trials, and the dynamic properties of PKA activation in the AL on the other hand leads to the hypothesis that prolonged PKA activation in the AL is involved in LTM formation. The hypothesis was tested by photolytic release of caged cAMP in the AL to artificially prolong PKA activation during olfactory conditioning. A local replay of the prolonged PKA activation in the AL *in vivo*, combined with a single conditioning trial, is sufficient to induce long-lasting memory. This provides direct functional evidence for a link between conditioning parameters, PKA activation in the AL, and its contribution to LTM formation in intact animals.

The use of the uncaging technique allowed further identification of the molecular processes underlying the prolonged PKA activation in the AL and thus the mechanisms contributing to induction of LTM in the AL. Nitric oxide, which is required for LTM formation in the honeybee, mediates the prolongation of the PKA activity by activation of the soluble guanylate cyclase and increase of cyclic guanosine monophosphate (cGMP). Inhibition of the soluble guanylate cyclase impairs both the prolonged PKA activation in the AL and LTM formation. Similar to the uncaging of cAMP, photorelease of caged cGMP in the AL – in combination with single-trial conditioning – induces long-lasting memory. Although the target of cGMP has not yet been identified *in vivo*, the synergistic activation of honeybee PKAII by cAMP and cGMP points to a direct function of cGMP in the prolonged increase in PKA activity ($\approx$25%) in the AL.

These findings demonstrate that molecular processes localized in the neural circuit of the AL critically contribute to LTM formation. Prolongation of PKA activation in the AL in conjunction with single-trial conditioning, however, does not reach the level of conditioned responses after multiple-trial conditioning, and thus it seems feasible that additional molecular processes or neural circuits may contribute to the prolongation. The molecular and neural targets of the early events in the AL that finally lead to LTM are yet unknown.

The signaling cascade that involves the Ca^{2+}-phospholipid-dependent protein kinase C (PKC) is implicated in a parallel-acting system contributing to olfactory memory at the AL level. In contrast to US-specific activation of the PKA in the AL, PKC is activated by both the US and the CS. The temporal pattern of PKC activation is independent of the sequence of CS and US stimulation or the number of conditioning trials. Since inhibition of PKC activation during the conditioning phase affects neither acquisition nor memory formation, conditioning-induced PKC activation in the AL seems not to be required for memory induction. The analysis of the learning-induced changes in PKC activity in the AL hours and days after training revealed a function of PKC in memory maintenance. In contrast to a single conditioning trial, repeated conditioning trials that induce LTM cause an increase in PKC activity beginning 1 h after conditioning and lasting up to 3 days. This long-lasting PKC activation can be dissected into two mechanistically independent phases. In the early phase (1–16 h), a constitutively active PKC, the PKM, contributes to the elevated activity. PKM is formed by cleavage of the activated PKC by the Ca^{2+}-dependent protease calpain. Blocking calpain activity during conditioning prevents PKM formation in the AL and impairs memory in a time window between 1 and 16 h. This treatment does not affect acquisition, the early memory phase up to 30 min, and memory after 1 day. The late phase of the training-induced increase in PKC activity (1–3 days) in the AL is unaffected by calpain blockers but requires protein and RNA synthesis. Thus, PKM formation in the AL is an independent parallel process required for maintaining a midterm memory phase. The function of the late phase (1–3 days) is still unclear but is probably one of several parallel mechanisms occurring in different brain circuits required for the formation of the late phase of LTM.

In summary, currently evidence exists for a function of the AL circuitry only in early processes of learning and memory formation, like the nitric oxide-mediated PKA activation involved in LTM induction and the early PKM formation required for maintaining a memory phase in the hours range.

Glutamate-Mediated Signaling Cascades in the MBs Contribute to LTM Formation

Glutamate is the major excitatory transmitter in mammals and plays a central role in neuronal plasticity in vertebrates. Its function in the insect brain, however, is only poorly understood. Although glutamate receptors have been identified in honeybees and manipulation of glutamate function using pharmacological tools supports a function in learning processes, the results are a matter of controversial discussion.

Photolytic uncaging of glutamate in the honeybee MB revealed direct evidence for a defined spatial and temporal contribution of glutamate in LTM formation. Uncaging glutamate immediately after a weak training protocol (single-trial training) in the MB improves the formation of a late memory phase (2 days) and thus mimics the effect of a strong training protocol. Uncaging shortly before the weak training has no effect, pointing to a defined function of glutamate with respect to early memory processing rather than learning. This function of glutamate is restricted to the MB circuitry, since glutamate release in the AL does not improve learning or retention. Taken together, glutamate action is restricted to the MBs, where it contributes to induction of a late memory phase in a time window shortly after training. Although it is still unclear whether this late memory phase is mechanistically identical with the LTM induced by a strong training (three-trial conditioning), the findings provide an example for parallel molecular processes underlying learning and memory formation *in vivo* occurring in different neural circuits.

Parallel Molecular Processes Contribute to LTM Memory Formation

As in other systems, LTM in honeybees can be divided into an early phase (eLTM, 1–2 days), which requires protein synthesis, and a transcription-dependent late phase (lLTM, $\geq$3 days). In all model systems investigated so far, inhibition of PKA activity during the training period results in a loss of both eLTM and lLTM. Although this suggests a single PKA-triggered process, experiments considering the impact of the satiation level in appetitive learning and memory formation in the honeybee provided evidence of a more complex function of the cAMP/PKA cascade in LTM induction. Three-trial conditioning in hungry animals leads to perfect acquisition and induction of midterm memory and both LTM phases. However, feeding bees 4 h before olfactory conditioning impairs acquisition and memory formation. This, and the different basal PKA-activity in honeybee brains at different satiation levels, points to a contribution of the cAMP/PKA cascade. Elevating the low basal PKA-activity levels in animals fed 4 h before conditioning specifically rescues the transcription-dependent lLTM. Acquisition, midterm memory, and eLTM are still impaired. The fact that both eLTM and lLTM require PKA for their induction, but PKA rescue in fed animals improves only lLTM but not eLTM, points to the existence of two parallel cAMP/PKA pathways implicated in memory formation: one process appears to trigger events leading to translation-dependent eLTM; the other is involved in inducing cascades required for transcription-dependent lLTM. Localization of the neural circuits related to these independent cAMP/PKA pathways will provide important information on the organization of the neural network involved in LTM formation.

See also: Communication in the Honeybee; Conditioning: Theories; Learning and Memory in Invertebrates: Honey Bee; Olfaction in Invertebrates: Honeybee; Procedural Learning: Classical Conditioning.

Further Reading

Abel R, Rybak J, and Menzel R (2001) Structure and response patterns of olfactory interneurons in the honeybee, *Apis mellifera*. *Journal of Comparative Neurology* 437: 363–383.

Balfanz S, Strunker T, Frings S, and Baumann A (2005) A family of octopamine receptors that specifically induce cyclic AMP production or Ca^{2+} release in *Drosophila melanogaster*. *Journal of Neurochemistry* 93: 440–451.

Brandt R, Rohlfing T, Rybak J, et al. (2005) A three-dimensional average-shape atlas of the honeybee brain and its applications. *Journal of Comparative Neurology* 492: 1–19.

Faber T, Joerges J, and Menzel R (1999) Associative learning modifies neural representations of odours in the insect brain. *Nature Neuroscience* 2: 74–78.

Farooqui T, Robinson K, Vaessin H, and Smith BH (2003) Modulation of early olfactory processing by an octopaminergic reinforcement pathway in the honeybee. *Journal of Neuroscience* 23: 5370–5380.

Fiala A, Müller U, and Menzel R (1999) Reversible downregulation of protein kinase A during olfactory learning using antisense technique impairs long-term memory formation in the honeybee, *Apis mellifera*. *Journal of Neuroscience* 19: 10125–10134.

Galán RF, Weidert M, Menzel R, Herz AV, and Galizia CG (2006) Sensory memory for odors is encoded in spontaneous correlated activity between olfactory glomeruli. *Neural Computation* 18: 10–25.

Ganeshina O and Menzel R (2001) GABA-immunoreactive neurons in the mushroom bodies of the honeybee: An electron microscopic study. *Journal of Comparative Neurology* 437: 335–349.

Grohmann L, Blenau W, Erber J, Ebert PR, Strunker T, and Baumann A (2003) Molecular and functional characterization of an octopamine receptor from honeybee (*Apis mellifera*) brain. *Journal of Neurochemistry* 86: 725–735.

Grünbaum L and Müller U (1998) Induction of a specific olfactory memory leads to a long-lasting activation of protein kinase C in the antennal lobe of the honeybee. *Journal of Neuroscience* 18: 4384–4392.

Grünewald B (1999) Physiological properties and response modulations of mushroom body feedback neurons during olfactory learning in the honeybee *Apis mellifera*. *Journal of Comparative Physiology A* 185: 565–576.

Hammer M (1993) An identified neuron mediates the unconditioned stimulus in associative olfactory learning in honeybees. *Nature* 366: 59–63.

Hammer M (1997) The neural basis of associative reward learning in honeybees. *Trends in Neurosciences* 20: 245–252.

Hammer M and Menzel R (1998) Multiple sites of associative odour learning as revealed by local brain microinjections of octopamine in honeybees. *Learning & Memory* 5: 146–156.

Hildebrandt H and Müller U (1995) Octopamine mediates rapid stimulation of PKA in the antennal lobe of honeybees. *Journal of Neurobiology* 27: 44–50.

Hildebrandt H and Müller U (1995) PKA activity in the antennal lobe of honeybees is regulated by chemosensory stimulation *in vivo*. *Brain Research* 679: 281–288.

Leboulle G and Müller U (2004) Synergistic activation of insect cAMP-dependent protein kinase A (type II) by cyclicAMP and cyclicGMP. *FEBS Letters* 576: 216–220.

Locatelli F, Bundrock G, and Müller U (2005) Focal and temporal release of glutamate in the mushroom bodies improves olfactory memory in *Apis mellifera*. *Journal of Neuroscience* 25: 11614–11618.

Lopatina N, Ryzhova I, and Chesnokova E (2002) The role of non-NMDA-receptors in the process of associative learning in the honeybee *Apis mellifera*. *Journal of Evolutionary Biochemistry and Physiology* 38: 211–217.

Maleszka R, Helliwell P, and Kucharski R (2000) Pharmacological interference with glutamate re-uptake impairs long-term memory in the honeybee *Apis mellifera*. *Behavioural Brain Research* 115: 49–53.

Mauelshagen J (1993) Neural correlates of olfactory learning in an identified neuron in the honey bee brain. *Journal of Neurophysiology* 69: 609–625.

Menzel R (1999) Memory dynamics in the honeybee. *Journal of Comparative Physiology A* 185: 323–340.

Menzel R, Erber J, and Masuhr T (1974) Learning and memory in the honeybee. In: Barton-Browne L (ed.) *Experimental Analysis of Insect Behaviour*, pp. 195–217. Berlin: Springer.

Menzel R and Manz G (2005) Neural plasticity of mushroom body-extrinsic neurons in the honeybee brain. *Journal of Experimental Biology* 208: 4317–4332.

Müller U (1996) Inhibition of nitric oxide synthase impairs a distinct form of long-term memory in the honeybee, *Apis mellifera*. *Neuron* 16: 541–549.

Müller U (1997) Neuronal cAMP-dependent protein kinase type II is concentrated in mushroom bodies of *Drosophila melanogaster* and the honeybee *Apis mellifera*. *Journal of Neurobiology* 33: 33–44.

Müller U (2000) Prolonged activation of cAMP-dependent protein kinase during conditioning induces long-term memory in honeybees. *Neuron* 27: 159–168.

Peele P, Ditzen M, Menzel R, and Galizia CG (2006) Appetitive odour learning does not change olfactory coding in a subpopulation of honeybee antennal lobe neurons. *Journal of Comparative Physiology A* 192(10): 1083–1103.

Riedel G, Platt B, and Micheau J (2003) Glutamate receptor function in learning and memory. *Behavioural Brain Research* 140: 1–47.

Rybak J and Menzel R (1998) Integrative properties of the Pe1-neuron, a unique mushroom body output neuron. *Learning & Memory* 5: 133–145.

Si A, Helliwell P, and Maleszka R (2004) Effects of NMDA receptor antagonists on olfactory learning and memory in the honeybee (*Apis mellifera*). *Pharmacology, Biochemistry, and Behavior* 77: 191–197.

Zannat MT, Locatelli F, Rybak J, Menzel R, and Leboulle G (2006) Identification and localisation of the NR1 sub-unit homologue of the NMDA glutamate receptor in the honeybee brain. *Neuroscience Letters* 398: 274–279.

Conditioning: Theories

A R Wagner, Yale University, New Haven, CT, USA
E H Vogel, Universidad de Talca, Talca, Chile

© 2009 Elsevier Ltd. All rights reserved.

Background

Notions Received from Pavlov

Much of the progress in understanding the basic mechanisms of learning is due to research conducted with Pavlovian (i.e., classical) conditioning, a paradigm in which animals change their response to a target stimulus, referred to as a conditioned stimulus (CS), as a function of the pairings of a CS with another stimulus, designated as an unconditioned stimulus (US). In the simplest Pavlovian conditioning procedure, one CS is paired with one US. As the result of one or more of these pairings or trials, the animal progressively changes its response to the CS, acquiring a conditioned response (CR). Subsequently, presentation of the CS without the US leads to a progressive diminution in the amplitude and frequency of the CR, in a process known as extinction.

The foundations of acquisition and extinction of the CR, as well as two related phenomena of generalization and discrimination, were intensively investigated by Pavlov and his followers more than a century ago. In generalization, it is seen that the strength of the CR is an increasing function of the similarity between the tested stimulus and the originally trained CS. In discrimination, acquisition trials with a given CS (i.e., A+) are intermixed with nonreinforced trials with another CS (B−). At first, this A+/B− training may result in responding to B as well as A (presumably due to generalization from A to B), but as training continues the animals progressively lose their response to B. In a special type of discrimination training in which A is followed by the US when alone, but not when presented in compound with B (i.e., A+/AB−), B acquires the capacity to inhibit the response elicited by other CSs, which is referred to as an instance of 'conditioned inhibition.'

The predominant explanation of all of these phenomena has been associative. That is, it is assumed that acquisition of the CR results from the strengthening of CS–US excitatory associations, that conditioned inhibition results from the development of inhibitory CS−US associations, and that extinction results from the weakening of excitatory CS−US associations or the suppressing effects of inhibitory associations.

Figure 1 depicts a prototypical network that can be used to describe the associative approach. It is assumed that environmental events, CSs and USs, activate their respective sensory processing units, represented by circles, which, by means of a net of interconnections (of strengths V and λ), influence the activity of a special 'adaptive' unit, which is responsible for the generation of the response. Pavlovian conditioning takes place because the adaptive unit, which prior to training is activated only by the US, can also, with training, come to be activated by the CS unit via its excitatory connection. Generalization is understood as the capability of similar CSs to partially activate the sensory unit of a trained CS. Likewise, conditioned inhibition is understood as the development of a negative connection between a CS and the adaptive unit. Extinction can be understood as either the weakening of the excitatory connection or as the superimposing of a new inhibitory connection (see the section titled 'Inhibitory learning during extinction').

The Linear Rule

The same computational learning rule was proposed within two major theoretical traditions, the Hull–Spence approach and the so-called stochastic models. This rule, known as the 'linear rule,' consists of the computation of the change in the strength of a habit (in the case of Hull–Spence theory) or the probability of a response (in the case of stochastic models) controlled by a CS, as a function of the discrepancy between a maximal or asymptotic associative level and the current associative level of the CS. This can be expressed formally by the equation $\Delta V_i = \theta(\lambda - V_i)$, where ΔV_i is the change in the excitatory associative strength of CS_i, θ is a learning rate parameter, and λ is the maximal associative strength supported by the US that occurred. According to this equation, changes in associative strength accrue to all stimuli for which internal representation is active on the occasions of reinforcement or nonreinforcement, regardless of any other stimuli present in the trial.

Contingency Effects and the Delta *P* Rule

In the late 1960s several lines of converging evidence suggested that the linear rule was not sufficient to explain conditioning. The first set of observations, concerning what are known as contingency effects, demonstrated that conditioned responding depends not only on the probability of the US given a CS, but also on the probability of the US in the absence

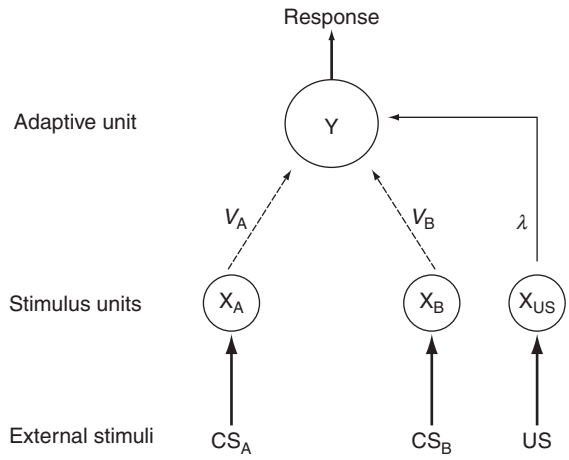

Figure 1 A network representation of Pavlovian conditioning (CS, conditioned stimulus; US, unconditioned stimulus). The circles represent processing units and the lines ending in arrows represent links among units. Continuous and discontinuous lines represent modifiable (V_i) and unmodifiable (λ) connections, respectively. The letters X_i and Y, represent the activity of the stimuli and adaptive units, respectively.

of the CS. In the typical contingency experiment several groups of animals receive the same number of CS–US pairings, but receive a different number of occurrences of the US in the absence of the CS. The outcome is that CR frequency or amplitude is negatively correlated with the probability of the US in absence of the CS.

The earliest account of contingency effects posited that animals must compute the correlation between the CS and US. Consider the contingency table shown in Table 1, containing the frequency of occurrence of the CS and the US in the presence and absence of each other. Excitatory conditioning takes place when the probability of the US in the presence of the CS (P(US/CS) = a/a + b) is greater than the probability of the US in the absence of the CS (P(US/no CS) = c/c + d). Inhibitory learning occurs when the probability of the US in the presence of the CS is smaller than the probability of the US in the absence of the CS, and no learning occurs when the two probabilities are equal. This idea has been summarized in what is known as the ΔP rule, which posits that the CR is a function of P(US/CS) − P(US/no CS), which, like a correlation coefficient, can vary between +1.0 (maximal excitation) and –1.0 (maximal inhibition).

Table 1 Contingency table

	US present	US absent
CS present	a	b
CS absent	c	d

The Rescorla–Wagner Rule

Another challenge to the linear rule came from a series of experiments that demonstrated that in conditioning involving a compound of cues, what is learned from one of the cues on a trial appears to depend not only upon its own current associative value, but also upon the associative value of the other cues present. The most influential observation is known as 'blocking.' If some CS, X, is reinforced in compound with another CS, A, the prior reinforcement of A alone reduces (blocks) the acquisition of conditioned responding to X. The phenomenon of conditioned inhibition, when subjected to careful analysis, revealed a similar dependence. That is, it was observed that training with AX, nonreinforced, causes X to become inhibitory only to the degree that A has been reinforced to be excitatory.

The Rescorla–Wagner model was designed to account for such effects. Formally, the model posits that the amount of change in the strength of the CS–US association for each CS present on a trial is a function of the error of the prediction of the US by all the CSs that are present in the trial – that is, that $\Delta V_i = \theta\ (\lambda - \sum V_i)$. This departs from the linear operator rule by introducing the aggregate associative strength in the computation of the learning to each CS. That is, when a CS is paired with a US or is nonreinforced, the change in the associative value of that CS does not depend only upon its own associative value, but also upon the aggregate value of all CSs present on that trial. In terms of Figure 1, the learning rule states that the more the adaptive unit is activated by all the CSs present on a trial, the less the associative strength will be increased if the US is presented, but the more the associative strength will be decreased if the US is not presented.

A major appeal of the Rescorla–Wagner rule was that it could account for contingency effects. It could do so by assuming that the experimental context can function like other stimuli in isolation and in compound with the training CS. Thus, US-alone trials can be seen as context-US trials, and CS–US trials can be seen as CS plus context-US trials. Application of the model with different frequencies of such assumed trials in the manner of Table 1 leads to quantities that co-vary with ΔP. In this regard, the Rescorla–Wagner rule, can be regarded as an approximate

heuristic for judging correlational relationships, as seen in contingency effects, without appealing to any complex computational capacity in the animals. Of course, the model has several limitations and these have stimulated theorizing during the past 30 years. Indeed, the history of associative learning theory after 1972 can be traced, in part, as the attempt of subsequent theorists to overcome presumed shortcomings of this model. The following sections introduce several current issues and discuss how they relate to the Rescorla–Wagner model.

Current Issues

Inhibitory Learning during Extinction

The term 'extinction' is commonly used to refer to a procedure in which a well-trained CS is presented repeatedly without reinforcement. The term also refers to the typical outcome of this procedure, which is a progressive decrease in the strength and frequency of the CR as a function of the number of nonreinforced presentations of the CS. The Rescorla–Wagner model construes the strength of a CS–US association as falling on a simple continuum from excitation to inhibition. During extinction the association becomes less excitatory and closer to being inhibitory. By this account there should be no difference between a CS that is mildly excitatory, because of modest CS–US training or because of extensive CS–US training, and some CS-alone extinction.

This unicontinuum account has been challenged as not easily accounting for the fact that an extinguished CR tends to reappear after extinction – that is, to show an increase not seen with a simple, modestly trained CS. This postextinction recovery of the CR happens under a number of circumstances including (a) when unsignaled USs are introduced after extinction (reinstatement), (b) when extinction is carried out in a context different from the one used in training and a subsequent test is carried out in the original context (renewal), and (c) after the mere passage of time (spontaneous recovery).

The demonstrations of reinstatement, renewal, and spontaneous recovery have encouraged the idea that extinction involves the development of a separate inhibitory CS–US association, which counteracts the intact excitatory tendency. In order to account for the postextinction recovery effects, it is further assumed that the inhibitory association construed during extinction is more susceptible to disruption by events like the passage of time or contextual changes than is the excitatory association. Although several descendants of the Rescorla–Wagner model have included separate excitatory and inhibitory CS–US connections during extinction, there is not yet a satisfactory theoretical rationalization of the presumed greater 'fragility' of the inhibitory learning.

Elementistic versus Configural Stimulus Representation

One of the major issues that separates current theories of associative learning is whether the associated stimuli are basically coded in elemental or configural fashion. Elementistic theories emphasize separate processing and representation of stimuli, no matter whether they are presented independently or in compound with other cues. By contrast, 'configural' accounts propose that compound stimuli are processed as unique exemplars and form associations independent of those formed by component elements. Thus, the cognitive representation of a compound is qualitatively different from the representation of its elements, although some of the associative strength acquired by a configuration can generalize to similar cues.

Although elemental and configural theories are often pitted against each other, it is important to appreciate that all existing theories recognize both elemental and configural processes. This can be seen in Figure 2, which depicts the general features of the Rescorla–Wagner model and Pearce's configural model. The Rescorla–Wagner model supposes that association develops for each element of a compound AB, but these include a context-dependent configural element X_{AB} as well as the context-independent elements X_A and X_B. The usual assertion is that the configural element X_{AB} is less salient than are the context-independent elements in learning situations involving compounds. On the other hand, configural models suppose that associations develop to experienced compounds, rather than to the elements of the compounds, but allow that association generalizes from one compound to another according to the distance that separates them along their elemental dimensions (denoted by the similarities links ${}_iS_j$ that connect the stimulus units in Figure 2). The usual assertion is that the generalization leaves much that is not shared from one configuration to others, even in cases of separable compounds.

Some phenomena are more intuitive from an elemental perspective. For example, when individuals are trained with two stimuli, A and B, each paired separately with the US, it can be anticipated that more responding will be seen in response to the novel compound, AB, than to either of the training stimuli. Other phenomena are more intuitive from a configural perspective. For example, it can be anticipated that the individuals can learn nonlinear

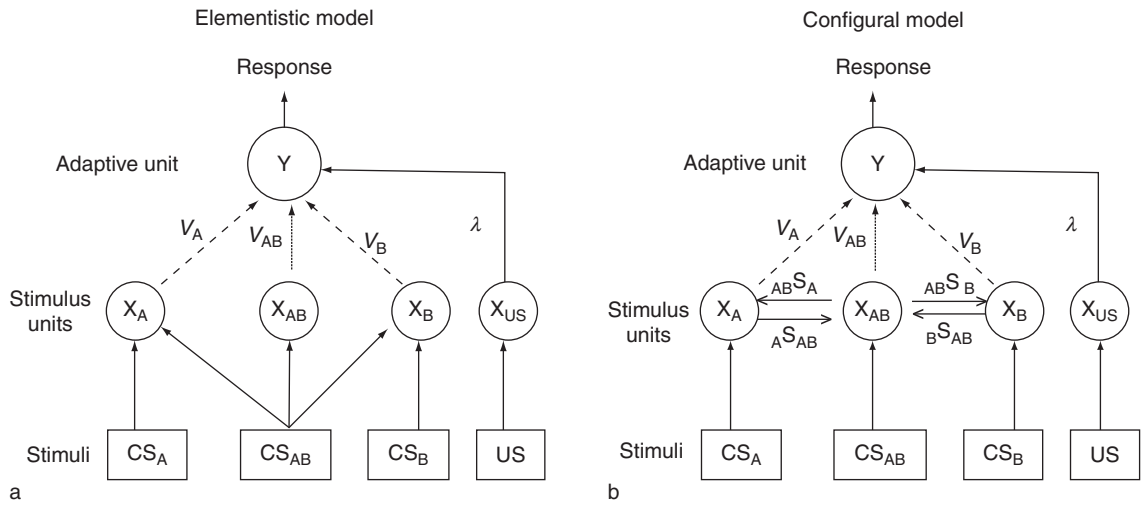

Figure 2 A network representation of elementistic (left panel) and configural (right panel) models of Pavlovian conditioning (CS, conditioned stimulus; US, unconditioned stimulus). In the elementistic model, the representational unit of A is equally activated by the A and AB stimuli, and the representational unit of B is equally activated by the B and AB stimuli. In the configural model, the AB stimulus activates directly its own representational unit AB and exerts an indirect influence over the A and B units via generalization links ($_{AB}S_A$ and $_{AB}S_B$). In the configural model, the A and B stimuli also influence the activity of the AB unit via their respective generalization links ($_AS_{AB}$ and $_BS_{AB}$).

discriminations, such as 'negative patterning,' in which A and B are separately paired with the US, but AB is not. But, all existing theories recognize both elemental and configural influences in order to provide some reasonable approach to these well-demonstrated phenomena. Advocates of the different theory types have nominated numerous additional phenomena as being more or less congenial to the different theories, or have come to different conclusions about their integrative usefulness. It is as possible, however, that the essential need is to understand better the variable influence of elemental versus configural processing in different circumstances.

Representation of Time in Conditioning

In the models that have been mentioned, the basic unit of analysis is the trial. For example, in the Rescorla–Wagner model the parameters, θ and λ, for the occurring stimuli, are specified constants on each trial, and ΔV_i is computed trial by trial. Such models are not equipped, without modification, to address important within-trial temporal phenomena. The most notable phenomenon is the tendency for the CR to be initiated and to peak at a time after CS onset that increases with the delay between the CS and US during training. The timing of the CR appears to become discriminated so as to occur at about the time at which the US has been presented. A striking instance of this can be seen when conditioning involves training with two different CS–US intervals on intermixed occasions. With sufficient training the result is a bimodal CR, with peaks at the two intervals at which the US was experienced.

The general approach to such phenomena has been to assume that the CS initiates a stimulus trace that is composed of a number of temporally differentiated components, and that ΔV is computed for each of the components on a moment-to-moment basis. In this case, just those components that are relatively active at the moment of the US are strengthened and become capable of eliciting the CR, whereas those that are relatively more active at moments of non-reinforcement are extinguished.

The notion of a componential CS representation was anticipated by Pavlov, who hypothesized that the timing of the CR was the result of discrimination learning that takes place among the many components that comprise the CS during its duration, some of which could be reinforced, nonreinforced, or partially reinforced. Modern models frequently employ a moment-by-moment application of the Rescorla–Wagner rule in computing how the temporal discrimination is formed, but differ in the pattern of activation of the components that the CS is assumed to initiate. Figure 3 presents three examples of the pattern of CS-component activity used in different componential models. The top plot depicts a so-called tap delay line, in which the CS onset initiates a sequential activation of elements in series. Each element activates the next element in the line but may have a period of activity that overlaps a number of the subsequent elements. Another possibility shown in

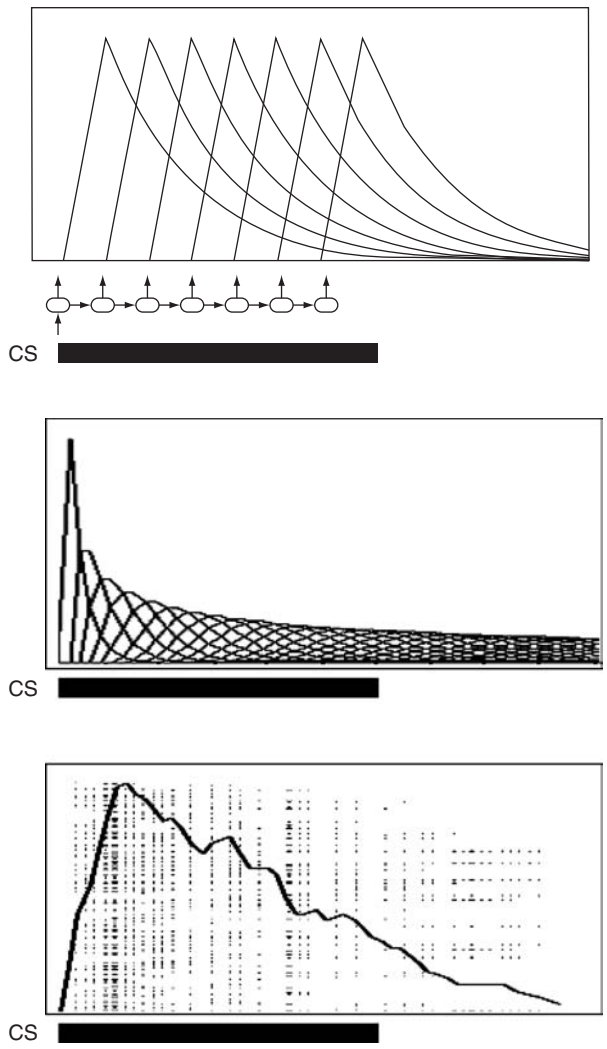

Figure 3 Time courses of three examples of componential conditioned stimulus (CS) representations. The black rectangle at the bottom of each plot represents the duration of the CS. The top plot represents a tap delay line, in which the CS activates a sequence of elements (the circles connected by arrows), which in turn give rise to identical nonmonotonic courses of activity. The middle plot represents the gamma functions, in which the CS gives rise to successive patterns of activity of varied shapes. The bottom plot presents a raster plot, with the activation of binary elements, in which each of the elements is represented on the ordinate and its state of activity varies over the duration of the CS. Superimposed on the raster plot is a summary function describing the aggregate number of elements that are active at each moment.

the middle plot is to assume that the CS generates a population of signals that vary in rise and decay parameters. These functions, known as gamma functions, are different from the delay line in that they are initiated at the same time, but nonetheless have different distinctive courses of activity. Finally, the bottom plot shows another example in which the CS influences the pattern of activity of a population of binary elements, each of which has 'on' and 'off' states that vary according to some characteristic stochastic process. One alternative or another may be defended on some formal grounds, as consistent with some sense of neurobiological plausibility, or as shadowing some particular data pattern in some conditioning preparation. Beyond such differences, the need for a componential representation of the CS has become relatively uncontroversial in the past 20 years.

Stimuli Trained in Compound Do Not Have a Fully Common Fate

One of the most straightforward implications of the Rescorla–Wagner rule is that there is a common fate for all stimuli that are trained in compound. Because a common error term is used to compute the learning that accrues to all of the stimuli present in a trial, equally salient stimuli should experience equal associative change. In fact, in many situations one can be most impressed with the way in which cues of different excitatory and inhibitory value, when trained in compound, appear to be commonly moved by reinforcement or nonreinforcement, as the model assumes. However, the results of a very clever experimental design recently employed by Robert Rescorla suggest that the change in associative strength of CSs trained in a compound do not have a fully common fate, as each is constrained to a degree by its own starting associative level.

The experiment proceeded as follows. Two stimuli, A and B, were originally trained to be equally excitatory by simple reinforced trials (A+ and B+). Two additional stimuli, C and D, were trained to be equally inhibitory by being nonreinforced in compound with an otherwise reinforced stimulus X (X+, CX−, and DX−). It could be assumed (without testing) that the compounds AC, AD, BC, and BD would all have equal total strength, since each contained one excitatory and one inhibitory element. Then one of these compounds, AC, was selected and further reinforced, so that it could be expected that both A and C would be incremented. The question posed was whether the increments to A and C were equal. If learning was determined by the error term in the Rescorla–Wagner rule ($\lambda - \sum V_i$), A and C should have gained exactly the same amount of associative strength at the end of the second phase. Whether they did, or not, was tested by comparing responding to the AD and BC compounds, each of which contained one of the originally reinforced CSs, one of the originally nonreinforced CSs, one of the additionally incremented CSs, and one nonincremented CS. The data clearly showed more responding to BC than to AD, indicating that the originally inhibitory C must

have been increased more than the originally excitatory A by the AC+ treatment.

This differential learning about elements of compounds can be handled by various modifications of the Rescorla–Wagner rule. One approach begins with a componential representation of the CSs, as described in the previous section. Assume that the associative strength of each of the nominal stimuli is determined by the summed strength of a large number of components of varying strengths. Assume further that whereas the total strengths are relatively unbounded, there is a limited excitatory or inhibitory strength that can be acquired by each individual component. Following this reasoning, it may be that when the excitatory A and the inhibitory C were reinforced in compound, although acted upon by a common error term, a greater proportion of the A components than the C components reached asymptote and were prevented from further increment. It remains to be seen whether this or some other modification in the learning rule will prove more useful, but the Rescorla data clearly imply that some constrained rule is necessary.

Priming Phenomena

The Rescorla–Wagner rule embodies the assumption that the more predicted is a US by the aggregation of CSs that precede it (i.e., the greater is $\sum V_i$), the less effective that US will be in producing excitatory learning. A substantial body of data indicate (a) that diminished learning is but one of the behavioral measures by which a predicted stimulus may be seen to be less effective than it otherwise would be, and (b) that prediction by a preceding CS is but one of the prestimulus manipulations by which diminished processing can be brought about. The larger pattern of data has encouraged a theoretical approach that has more general scope than does the Rescorla–Wagner model.

Among the other ways by which a predicted stimulus may be seen to be less effective than it otherwise would be, one that has been central to theoretical development, is the 'conditioned diminution of the UR.' This refers to the fact that a US that is preceded by an excitatory CS with which it has previously been paired elicits a UR that is of diminished amplitude, in comparison to when it is unannounced or is preceded by the same CS after it is extinguished. This phenomenon of conditioned diminution is not peculiar to the stimuli customarily employed as USs in Pavlovian conditioning, but has been demonstrated as well, using a measure of the orienting response, with stimuli that would more commonly be used as CSs. For example, a light that has been repeatedly preceded by a distinctive tone produces less of an orienting response when it is preceded by the tone than when not, and this diminution depends on the integrity of the tone–light association.

Another prestimulus manipulation by which diminished processing of a stimulus can be brought about is by a shortly preceding exposure to the same stimulus. The phenomenon of short-term habituation is well known, whereby the UR to a stimulus is decremented on the second of two exposures, depending on the interval of time between the two. Studies in eyeblink conditioning of the rabbit with a paraorbital US have demonstrated that application of the US shortly before a CS–US pairing diminishes the UR observed to the US in the pairing and decreases the amount of excitatory learning that it produces. Studies have similarly shown that application of the CS shortly before a CS–US pairing decreases the amount of conditioning that accrues to the CS.

The general pattern that emerges from these observations is that a training stimulus, whether CS or US, is rendered less effective than it otherwise would be, if on the occasion of its occurrence if it is preactivated or 'primed' in the memory system. This can be accomplished by presentation of the stimulus itself, or by presentation of a stimulus that can retrieve the training stimulus as a result of prior associative learning. A formal model that incorporates these notions was formulated by Allan Wagner and his colleagues, and given the acronym, SOP (for 'sometimes opponent process').

Detailed exposition of the quantitative features of SOP is beyond the scope of the present treatment, but the essence of the formulation can be appreciated through the network depicted in Figure 4. Stimuli are assumed to activate a sequence of nodes, Al followed by A2. The central assumption by which the priming effects are addressed is that activity of the secondary node recurrently inhibits the primary node, making it transiently less susceptible to activation by its initiating stimulus. The assumption is further made that, of the activity sequence that is occasioned by any stimulus, it is the A2 node that is the adaptive node that can come to be activated by a formally neutral stimulus. By this reasoning, if the A2 node of a US is activated either by a recent instance of that stimulus or by the precedence of an associated CS, the A1 node of that US will be less susceptible to activation, so as to produce a diminished UR and less associative learning. The same reasoning would apply if the A2 node of a CS is activated by a recent instance of that stimulus or the precedence of an associated stimulus. In this case the CS should show diminished associability.

Applications of the SOP model, like the Rescorla–Wagner model, have often imputed associative

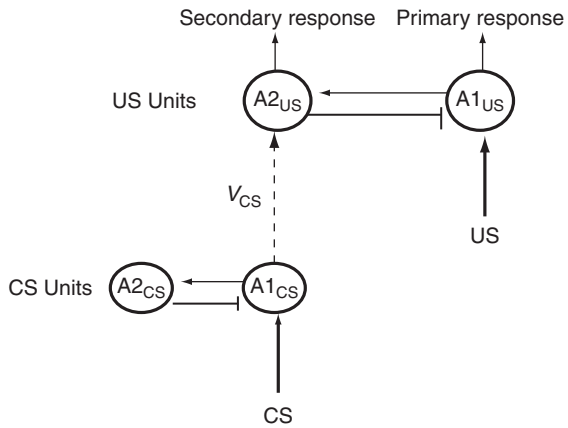

Figure 4 A network representation of the SOP model (see text for discussion). Each stimulus is represented by a primary ($A1_i$) and a secondary unit ($A2_i$) unit. The conditioned stimulus (CS) and the unconditioned stimulus (US) activate their respective primary units, which in turn activate the corresponding secondary units. The CS influences the activity of the US-secondary unit by means of its associative value, V_{CS}. The links ending in a vertical line represent inhibitory feedback from each secondary unit to its corresponding primary unit. The activity in the $A1_{US}$ units is responsible for the generation of the primary unconditioned response, whereas the activity in $A2_{US}$ is responsible for the generation of the secondary unconditioned response and the conditioned response.

influence to the experimental context in which Pavlovian conditioning occurs. US preexposure in one context diminishes the associative learning produced by the US in that context. Similarly, CS preexposure in one context diminishes the associability of the CS in that context. Both these effects can be rationalized as being due to the preexposure context acquiring, through associative learning, the ability to activate the representation of the preexposed stimulus to the A2 state. Such a context-specific associative effect may be implicated in many observations of long-term habituation, but at the moment there is not clear understanding of why the habituation decrement observed in some URs (e.g., freezing) is more context specific than is that of others (e.g., startle) elicited by the same stimulus.

One feature of SOP pointed to in Figure 4 is that the CR and the UR supported by a US are not assumed to be identical. The UR is assumed to involve a primary response to the A1 activity, followed by a secondary response to the A2 activity, whereas the CR is assumed to involve only the latter response. In some conditioning preparations, such as the rabbit eyeblink reinforced by paraorbital shock, the CR and UR are separable eyelid closures are similar. In other preparations, such as the conditioned emotional response reinforced by a shock US, the two responses may appear antagonistic, as the CR (and secondary UR) is freezing, whereas the prominent initial UR is increased activity.

Divergence of Response Measures

Much of the diversity among current theories of Pavlovian conditioning stems from differences regarding the manner in which the CS is represented. By contrast, there has been relatively less attention paid to the manner in which the US is assumed to be represented. There is, however, reason to suppose that the US is represented by dissociable elements, just as is the CS. The most obvious call for such an assumption comes from the substantial data showing that different Pavlovian conditioning responses reinforced by the same nominal US may fail to be well correlated.

The term 'divergence of response measures' was first coined in reference to the finding with rabbits, conditioned with a paraorbital shock US, that over a portion of the CS–US interval function, evidence of conditioning with the eyeblink CR decreased while evidence of conditioning with a heart rate CR increased. In a more striking example in the same preparation, it has been shown that the same backward pairings, of a paraorbital shock US followed by a CS, produced evidence of net inhibitory associative learning to the CS when assessed through conditioned eyeblink responding, but produced evidence of net excitatory associative learning when assessed through conditioned suppression of drinking. One possible interpretation of this phenomenon is that the conditioned eyeblink and the conditioned emotional response may be reinforced by different aspects of the US, and that the different aspects engage differential temporal processing.

Figure 5 is intended to hint at this frequently assumed notion of separable processing of two US components, one representing the sensory–perceptual aspect of the US and the other representing its emotional attributes. Association of the CS with the sensory–perceptual aspect of the US is assumed to be responsible for the CR, whereas the association of the CS with the emotional attribute is thought to be responsible for the conditioned emotional response (CER). This is consistent with findings suggesting that the rabbit eyeblink is controlled by a US processing center in the cerebellum, whereas the CER is controlled by a center in the amygdala. Different centers in the amygdala activated by the US appear to similarly produce variation in common measures of the CER, such as freezing, startle, heart rate changes, and defecation. Included in Figure 5 is the supposition

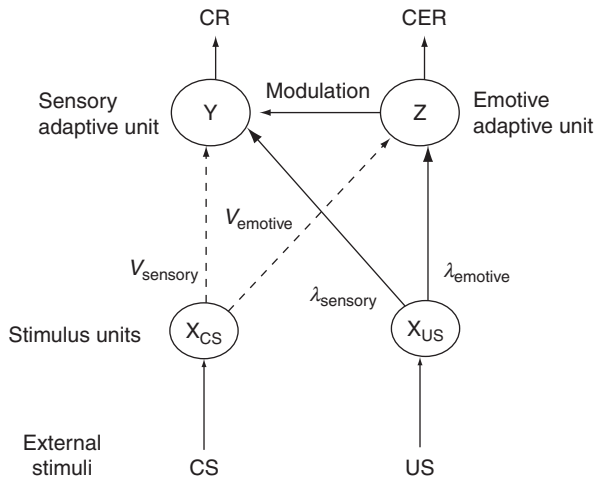

Figure 5 A network representation of a modulatory model of Pavlovian conditioning. It is assumed that the presentation of the unconditioned stimulus (US) produces activation of a sensory unit (Y) and an emotive unit (Z) via its $\lambda_{sensory}$ and $\lambda_{emotive}$ links, respectively. The sensory unit controls the conditioned response (CR) and can be associatively activated by the conditioned stimulus (CS) via the link $V_{sensory}$. The emotive unit controls the conditioned emotional response (CER), modulates the sensory unit, and can be associatively activated by the CS via $V_{emotive}$.

that a consequence of the CER is also a modulation of the motor CR occasioned by the same CS.

Learning Occurring with Absent Stimuli

The Rescorla–Wagner model and the elaborations and alternatives that have been mentioned are alike in assuming that CS–US association and CS extinction occur only via episodes in which the designated stimuli occur. There are a number of phenomena that challenge this restriction, some of which have been given ready interpretation, but others which remain problematic and in search of explanation. 'Second-order conditioning' occurs when CSl–US pairing leads CS1 to elicit a CR, and subsequent CS2–CSl pairings cause CS2 to acquire similar responding. This well-documented phenomenon might be entertained as suggesting that, during the CS2–CS1 training, CS2 was paired with a representation of the absent US retrieved by CS1 to establish an effective CS2–US association. However, analysis of the learning that occurs favors the interpretation that the CS2–CS1 pairing establishes an association between CS2 and the CR produced by CS1 during this training. 'Sensory preconditioning' occurs when CS2–CS1 pairing precedes CS1–US pairing and subsequent CS2 presentation evokes the CR previously acquired to CS1. Again, this might be interpreted as evidence that CS1 retrieved representation of CS2 on the occasions of the CS1–US pairings, so that the absent CS2, as well as CS1, formed associations with the US. However, analysis of the learning that occurs favors the interpretation that only CS2–CS1 and CS1–US associations are involved – that is, that upon testing, CS2 retrieves the representation of CS1, which then evokes the CR. This supposes that an absent stimulus may mediate responding but does not necessarily mediate associative learning.

There are two conditioning phenomena that are related to the previous examples but about which interpretation is less secure. In one case, a CS is paired with a food US in phase 1 of training and with a malaise-inducing agent in phase 2. The test in this instance is not with the responding to the CS, but rather with the response to the food US that had been employed in phase 1. The finding was that, as a result of the phase 2 training, the food was less accepted. The 'absent cue' interpretation is that during phase 2 the CS retrieved representation of the food US during the pairings with malaise, thereby strengthening a food–malaise association that made the food less attractive.

A second interesting case that has been reported is an instance of apparent retrospective revaluation termed 'unovershadowing.' In initial training, a compound CS, AB, was paired with a US, so that A acquired a tendency to elicit a CR, but less than if it had been trained in isolation (i.e., without the overshadowing influence of B). In phase 2, cue B was extensively extinguished, after which cue A was tested. The challenging observation has been that cue A is more responded to as a result of the extinction of B. The 'absent cue' interpretation is that on trials in which B is extinguished, it retrieves representation of A, the association of which with the US is not extinguished but, rather is enhanced.

The two challenging phenomena have sometimes been interpreted via modification to the SOP model. The model explicitly assumes that $A1_{CS}$–$A1_{US}$ conjunctions are necessary for excitatory learning and that $A1_{CS}$–$A2_{US}$ conjunctions are necessary for inhibitory learning. The data from the first example might be accounted for if it were assumed that $A2_{food}$–$A1_{malaise}$ conjunctions strengthened the food–malaise association. The data from the unovershadowing example might be accounted for by assuming that $A2_{A}$–$A2_{US}$ conjunctions strengthened the A–US association. Whatever the virtues of the separate modifications of SOP might be, it should be apparent that the two appear incompatible: pairing of a retrieved A2 state of a stimulus with the A1 state or the A2 state of another stimulus cannot be similarly strengthening of an association, at least within the usual understandings of the model.

There is yet another theoretical approach that may account for several of the reported effects of learning with absent cues. This is the 'comparator theory' proposed by Ralph Miller and his colleagues. This approach assumes that learning occurs by mere CS–US contiguity (like the linear operator rule) and that responding to a given CS is a function of the comparison of the CS's associative strength with the associative strengths of any other cues with which the CS has become associated. Thus, for instance, the low level of responding initially seen in response to the overshadowed cue A would not be due to the low associative strength of the cue, but to the comparison with the similar or greater associative strength of the overshadowing cue B, with which A was associated. After extinction of B, performance in response to A can be assumed to increase thanks to a more advantageous comparison with B.

The comparator theory deals well with phenomena like unovershadowing, in which the target cue (e.g., the overshadowed cue A) is transparently affected by a comparator cue (e.g., the overshadowing cue B). However, the theory at this time is necessarily complicated by extensions which appeal to mechanisms that regulate when target and comparator cues sustain a competitive or facilitative relationship. The issue of learning with absent cues clearly begs further empirical and theoretical clarifications.

See also: Aversive Emotions: Molecular Basis of Unconditioned Fear; Conditioned Reflex; Conditioned Taste Aversion; Conditioning: Simple Neural Circuits in the Honey Bee; Executive Function and Higher-Order Cognition: Computational Models; Eyeblink Conditioning; Fear Conditioning and Synaptic Plasticity; Prediction Errors in Neural Processing: Imaging in Humans; Procedural Learning: Classical Conditioning; Reinforcement Models.

Further Reading

Bouton ME (2004) Context and behavioral processes in extinction. *Learning & Memory* 11: 485–494.

Brandon SE, Vogel EH, and Wagner AR (2002) Computational theories of classical conditioning. In: Moore JW (ed.) *A Neuroscientist's Guide to Classical Conditioning*, pp. 270–349. New York: Springer.

Bower GH (1994) A turning-point in mathematical learning-theory. *Psychological Review* 101: 290–300.

Church RM and Kirkpatrick K (2001) Theories of conditioning and timing. In: Mowrer RR, and Klein SB (eds.) *Contemporary Learning: Theory and Application*, pp. 211–263. Hillsdale, NJ: Erlbaum.

Delamater AR (2004) Experimental extinction in Pavlovian conditioning: Behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology Section B – Comparative and Physiological Psychology* 57: 97–132.

Denniston JC, Savastano HI, Blaisdell AP, et al. (2003) Cue competition as a retrieval deficit. *Learning and Motivation* 34: 1–31.

Holland PC (2005) Amount of training effects in representation-mediated food aversion learning: No evidence of a role for associability changes. *Learning & Behavior* 33: 464–478.

Hall G (1996) Learning about associatively activated stimulus representations: Implications for acquired equivalence and perceptual learning. *Animal Learning & Behavior* 24: 233–255.

Kehoe EJ (1990) Classical conditioning: Fundamental issues for adaptive network models. In: Gabriel M and Moore JW (eds.) *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pp. 389–420. Cambridge, MA: MIT Press.

Pearce JM and Bouton ME (2001) Theories of associative learning in animals. *Annual Review of Psychology* 52: 111–139.

Pearce JM (2002) Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior* 30: 73–95.

Pickens CL and Holland PC (2004) Conditioning and cognition. *Neuroscience and Biobehavioral Reviews* 28: 651–661.

Rescorla RA and Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non reinforcement. In: Black AH and Proasky WF (eds.) *Classical Conditioning II: Current Theory and Research*, pp. 64–99. New York: Appleton-Century-Crofts.

Wagner AR (1981) SOP: A model of automatic memory processing in animal behavior. In: Spear NE and Miller RR (eds.) *Information Processing in Animals: Memory Mechanisms*, pp. 5–47. Hillsdale, NJ: Lawrence Erlbaum.

Wagner AR (2003) Context-sensitive elemental theory. *Quarterly Journal of Experimental Psychology* 56B: 7–29.

Wagner AR and Brandon SE (1989) Evolution of a structured connectionist model of Pavlovian conditioning (AESOP). In: Klein SB and Mowrer RR (eds.) *Contemporary Learning Theories: Pavlovian Conditioning and the Status of Traditional Learning Theory*, pp. 149–189. Hillsdale, NJ: Lawrence Erlbaum.

Cone Photopigment Evolution

G H Jacobs, University of California, Santa Barbara, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Essential Features of Cone Photopigments

Cones are delicate, elongated cells located at the outermost portion of the neural retina (Figure 1(a)). Along with rods, they constitute the two types of vertebrate photoreceptors. These receptors contain photopigments that convert photon energy into neural signals, thus serving as the essential first stage in the complex process that culminates in vision. Rods and cones in all vertebrates have different distribution patterns across the retina, different patterns of connectivity to downstream neurons, and inherent differences in sensitivity and temporal responsivity. These differences condition the manner in which rod and cone operation is linked to vision. Cones are traditionally conceived as supporting daylight vision by providing the information required for high spatial acuity and good temporal resolution. Subpopulations of cones in the retina of any given species typically contain different types of photopigments. This latter feature allows cones to also provide the signals that can be used to support color vision. In general, there are predictive relationships between the number of cone pigment types and the dimensionality of color vision: those retinas that contain only a single pigment are linked to an absence of color vision (a condition referred to as 'monochromatic'), those with two pigment types yield a single dimension of color vision (dichromatic), those with three types support two dimensions (trichromatic), etc.

Cone pigment is packed in a stack of parallel membranes that make up the distal portion of cone receptors, a region called the outer segment (Figure 1(a)). The physical arrangement is such as to increase the total pigment density in each receptor, thus maximizing the likelihood that incoming light will be absorbed by the pigment. Cone pigments are members of a family of G-protein-linked receptor molecules in which a chain of amino acids is embedded in a lipid membrane. The folding of the amino acid chain of the visual pigment is such that it traverses the membrane in seven helical palisades (Figure 1(b)). This protein, called opsin, has an accessory group – a chromophore – that is covalently linked to the opsin in the seventh of the transmembrane segments of the molecule (at position 312 in Figure 1(b)). The chromophore is a derivative of vitamin A and all vertebrate cone pigments are made from one of two chromophores: retinal or 3-dehydroretinal. The chromophore used in the manufacture of the cone pigments of birds and mammals is retinal; cone pigments of fishes, reptiles, and amphibians are constructed from either retinal or 3-dehydroretinal. Absorption of a photon of light by the pigment causes a conformational change (an isomerization) in the chromophore. The isomerization of the chromophore from an 11-*cis* to an all-*trans* configuration initiates a cascade of molecular events that produces a change in the flow of ionic current across the membrane of the cone outer segment. Characteristic features of this cascade include considerable speed and a large amplification of signal strength. Once initiated, the signal can spread along the extent of the receptor membrane and so trigger a change in the rate of release of neurotransmitter from the synaptic terminal of the cone, in turn informing second-order cells about receptor activation.

As for all other pigments, an absorbance spectrum provides an essential characterization of a visual pigment. Such spectra plot the probability that absorption will occur as a function of the wavelength of the light (Figure 1(c)). Opsin itself absorbs maximally in the ultraviolet (UV) wavelengths, below 300 nm. When it is bound to the chromophore, the absorbance spectrum shifts toward the longer wavelengths. The exact spectral positioning of the resultant pigment depends on the structure of the particular opsin. Although there are well-documented exceptions, generally each cone contains only a single type of photopigment. When they are properly scaled, the shapes of the absorbance spectra of all photopigments are very similar. Consequently, it is convenient and conventional to specify photopigments by the wavelength value to which they are most sensitive (λ_{max}).

Linking Opsin Genes and Cone Opsins

The genes specifying the opsins of the three types of human cone photopigment were first isolated and sequenced about 20 years ago. This achievement served as a catalytic event in the quest to understand the evolution of photopigments. As had long been inferred from early work on the inheritance of the common color vision defects, the genes for two of these opsins are located on the q-arm of the X-chromosome, where they lay in a head-to-tail tandem array. Because of the relative positioning of their absorption spectra, the two products are conventionally termed M (middle-wavelength-sensitive)

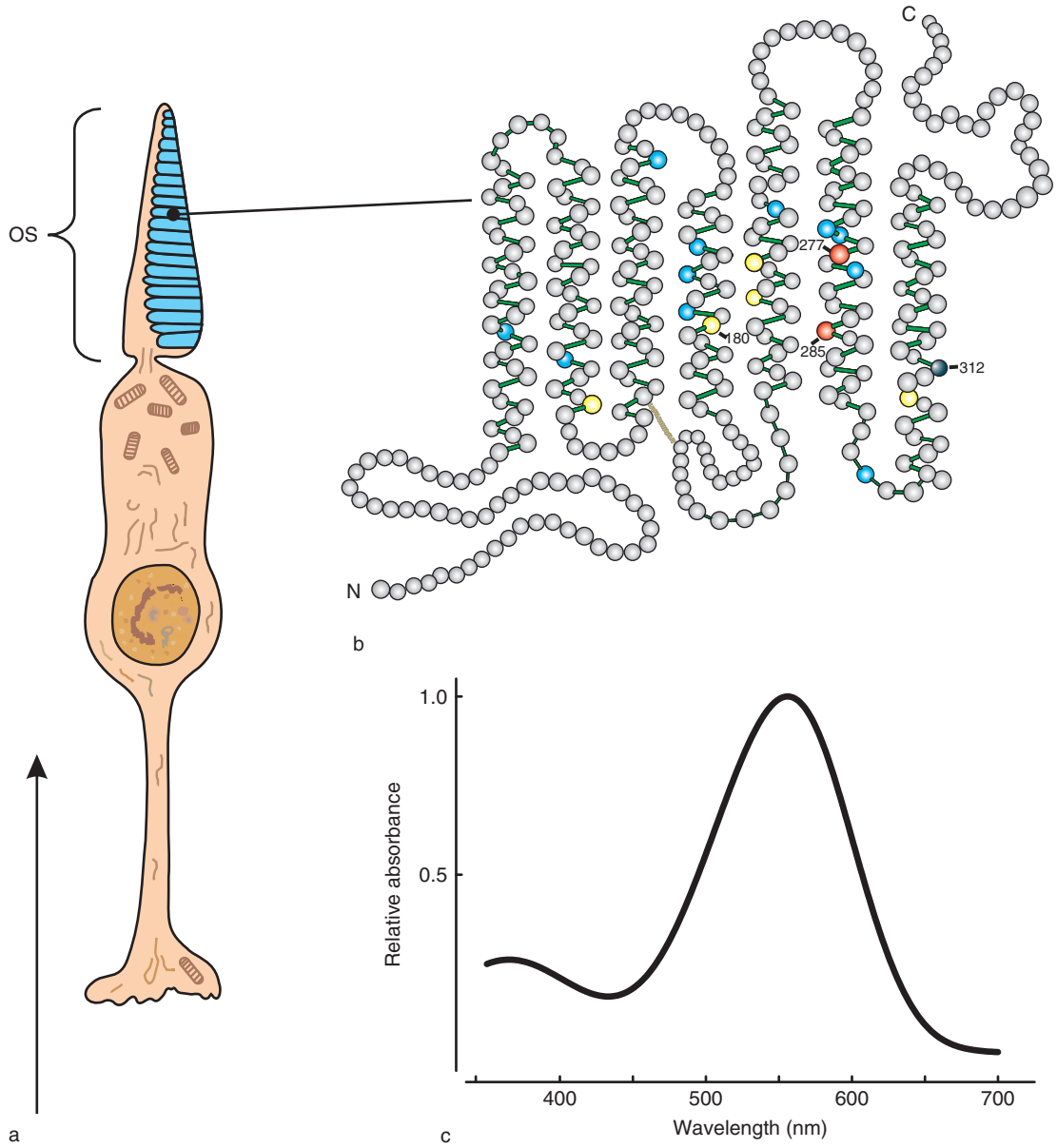

Figure 1 (a) Sketch of a vertebrate cone photoreceptor. The normal pathway followed by incoming light is indicated by the arrow. The outer segment (OS) at the top consists of a stack of photopigment-rich disks. The synaptic terminal of the photoreceptor where signal transmission to other retinal cells occurs is located at the bottom. (b) Schematic representation of the M/L primate cone opsin molecule. The 364 amino acids in this molecule are shown as small circles. The three residues that are responsible for the spectral tuning of these pigments in New World monkeys are located at numbered locations (see text). The chromophore attachment site is at residue 312. (c) Absorption curve for a typical cone photopigment.

and L (long-wavelength-sensitive) pigments. As shown in the schematic drawing (Figure 1(b)), each of these genes specifies an opsin consisting of 364 amino acids. Reflective of their common evolutionary history, the structures of these two genes are very similar, yielding proteins that have amino acid sequences that are ~96% identical. The gene for the third type of human cone pigment, the S (short-wavelength-sensitive) pigment, is located on chromosome 7. It specifies an opsin consisting of approximately 350 amino acids. The sequence of this gene is considerably divergent from the other two, producing an amino acid sequence for the S opsin that is only about 40–45% identical to that of the M and L opsins.

This early work on the structure of opsin genes inspired two lines of research. First, the availability of opsin clones led to the isolation of other opsin genes, and over the past two decades hundreds of

opsin sequences derived for many different species have been resolved. These sequences are deposited at the website of the National Center for Biotechnology Information (NCBI). Second, it quickly became possible to correlate opsin sequences with photopigment spectra, thus providing a critical link between gene structure and visual function. This was accomplished in two different ways. The first was to compare opsin sequence variations in animals whose cone pigments were known from other types of measurements to have different absorption spectra. Another was to directly produce opsins in artificial expression systems and then make *in vitro* measurements of the pigment product. The latter proved to be a particularly powerful approach because it could be pursued in opsins with engineered sequence variations, including both those having targeted mutations and those that were constructed as opsin gene chimeras. These approaches eventually made possible an understanding of the linkages between gene structure variations and variations in cone absorption spectra.

Changes at only a few critical locations in opsin genes are associated with shifts in the spectral absorption properties of the photopigment. For example, the M/L photopigments of New World (platyrrhine) monkeys occupy any one of five discrete spectral locations having λ_{max} values in the range from about 530 to 560 nm. This variation is controlled by changes at only three locations in opsin molecule (these are indicated in Figure 1(b)). Dimorphic amino acid substitutions at these three locations – alanine/serine at position 180, tyrosine/phenylalanine at 277, and alanine/threonine at 285 – yield pigments at the five locations. In other primates, substitutions at these same locations plus an additional two sites are involved in the spectral tuning of M/L cone pigments. This very conservative mechanism for spectral tuning allows inferences to be drawn about M/L pigments in animals where direct pigment measurements are not available. Cone opsin gene sequence information in turn can then be used to directly infer some features of visual experience, particularly about the nature of color vision.

Opsin Gene Families and Their Evolutionary Histories

Comparisons of the amino acid sequences of photopigment opsins from a variety of species has led to the consensus conclusion that all vertebrate visual pigments can be classified into five groups. These groups are considered to be paralogous, that is, they share a common evolutionary ancestry. An opsin family tree for these five groups appears in Figure 2,

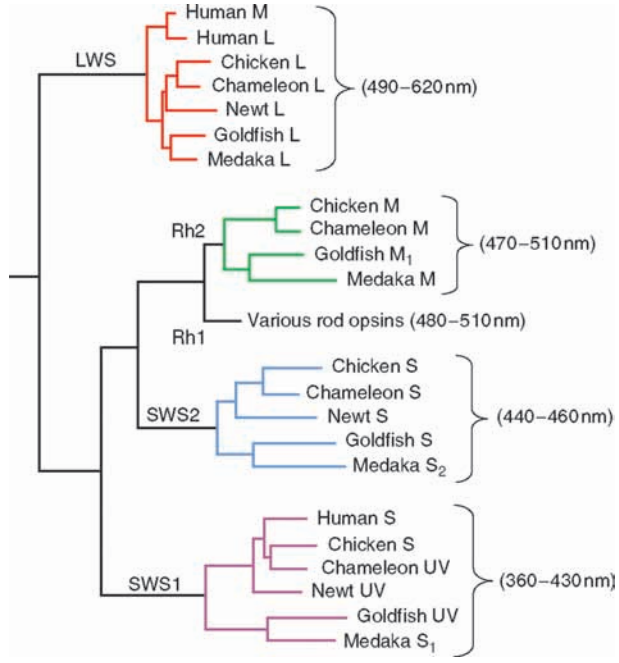

Figure 2 Opsin gene family tree. Shown are representative contemporary species that draw their pigments from the five families. The ranges of λ_{max} values characteristic of the pigments from each of the families are indicated to the right of the brackets. The tree construction is derived from Hisatomi O and Tokunaga F (2002) Molecular evolution of proteins involved in vertebrate phototransduction. *Comparative and Biochemical Physiology B* 133: 509–522.

where results from a number of species representing different vertebrate groups are illustrated and the spectral locations of pigments characteristic of each group are noted. In total, vertebrate opsin genes produce photopigments whose λ_{max} values cover the spectrum from about 360 to 620 nm. The spectral diversity of pigments specified by some of these gene families is restricted – pigments from the SWS2 (short-wavelength-sensitive) group have peak values extending across a range of only about 20 nm – whereas pigments from other groups spread over a much more expansive range; for example, the pigments from the LWS (long-wavelength-sensitive) family cover a peak range of some 130 nm. Of the five groups, pigments from the Rh1 family are typically expressed in rods while the other four are usually cone pigments.

Although the general relationships among the opsin genes seem to be reasonably settled, it is much less clear as to the timing of important evolutionary changes. One suggestion is that photopigments arose very early in the evolution of life, perhaps in motile microorganisms similar to green algae. The divergence of a single type of opsin gene into two

separate families was apparently also an ancient event, variously estimated as having occurred 800–1100 Ma. The products of this divergence would have permitted animals to have two types of photopigments with respective maximum sensitivities in the short and long wavelengths. In understanding subsequent changes that may have occurred, recent work on the opsin genes of the lamprey, an agnathan (jawless) fish, has been particularly informative. This animal has predominantly cone-based vision that originates from the activation of four types of cone photopigments. Each type of pigment is drawn from one of the four classes of cone opsin genes. That fact provides strong evidence that four of the five classes of vertebrate opsin genes arose prior to the evolution of jawed vertebrates, that is, by the early Cambrian period (~540 Ma). Rod pigments then appeared only after all the cone opsin gene families were in place. These facts imply that, at least potentially, a relatively elaborate capacity for daylight vision, including color vision, has been characteristic of the entire history of vertebrates, with rod-mediated (dim light) vision a relatively more recent invention. That scenario also receives support from examinations of the structure of mammalian retinas, which show that there is much more elaborate neural circuitry for the analysis of cone as opposed to rod signals. All of these observations suggest that rods and their associated neural circuitry were superimposed on an organization originally evolved to support cone-based vision.

From Figure 2, it can be seen that all four classes of cone opsin genes are present in some contemporary birds, fish, and reptiles. Note however that the only mammal represented in Figure 2, *Homo sapiens*, has pigments drawn from only two of the four gene families, LWS and SWS1. That example, and many others, indicates that although four opsin gene families have been maintained throughout the history of vertebrates, that arrangement is not uniformly true for all contemporary vertebrates. The particular exceptions offered by mammals are the focus of the following sections.

Mammalian Cone Pigment Evolution

Paleontological evidence suggests that the early mammals were mostly strongly nocturnal, and that is believed to have had a strong impact on mammalian visual adaptations. This photic history may have led to three types of changes in the visual system: average increases in rod/cone ratios, loss of colored oil droplets from mammalian photoreceptors that have been retained in many other vertebrates, and the disappearance of any representation of cone pigments from two gene families, namely Rh2 and SWS2 (see Figure 2). The latter fact implies that the retinas of ancestral mammals were limited to two types of cone pigments, one absorbing in the short wavelengths and one in the middle to long wavelengths. Based on evidence that a total of only five amino acid sites in LWS opsins are involved in the spectral tuning of all mammalian LWS cone pigments, it is possible to infer the evolution of spectral change in these pigments. Figure 3 shows a tree typology for mammalian LWS pigments for a number of representative mammals. By this accounting, the original mammalian LWS pigment had a spectral peak close to 530 nm. Divergences since then have yielded cone pigments that span a spectral range from about 510 to 560 nm. Note that among the contemporary mammals shown in Figure 3, only the human has acquired a second type of LWS pigment. That addition is also characteristic of some other primates, as described in the next section.

An interesting exception to the rule that mammals have retained only two of the four cone opsin gene families occurs in some Australian marsupials. At least four species from this group have recently been found to have a total of three types of cone photopigments. In addition to pigments drawn from the SWS1 and LWS gene families, these animals have a pigment from the Rh2 family. Recent behavioral experiments indicate that this third cone pigment does support trichromatic color vision in these marsupials. Why some marsupials have retained pigments derived from the Rh2 opsin gene family, while all other mammals seem to have lost them, remains at present a mystery.

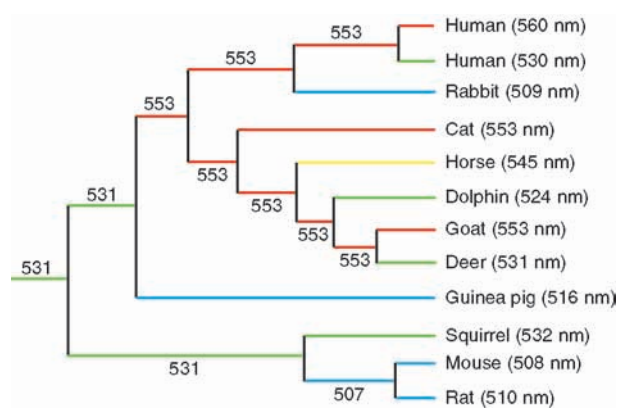

Figure 3 A proposed tree of mammalian LWS pigments. The numbers on the branches are λ_{max} values for the pigments as obtained from a model of spectral tuning; those in parentheses represent direct measurements. These inferences were derived from Yokoyama S and Radlwimmer FB (1999) The molecular genetics of red and green color vision in mammals. *Genetics* 153: 919–932.

Primate Cone Pigment Evolution

The number of cone pigments and the dimensionality in color vision this allows are summarized for the various groupings of extant primates in Table 1. A major point is that, among eutherian mammals, only primates have been able to move beyond two classes of cone photopigment and the visual confines this creates. But not all of them have done this, and the ones who have done it have accomplished it in two quite distinct ways. In one instance, the single LWS X-chromosome opsin gene duplicated at a point close to the base of the radiation leading to contemporary catarrhine primates (Old World monkeys, apes, and humans), perhaps 30–35 Ma. This led, through subsequent changes that are still matters of speculation, to the presence of two genes coding for spectrally distinct M and L pigments (Figure 3). This arrangement provides routine trichromatic color vision to all members of these species. There are some exceptions to this, notably among humans (see later).

Some New World monkeys have been able to achieve the same trichromatic end by following a quite separate pathway. In these primates, there are alleles of the X-chromosome opsin gene, usually three, each coding for a spectrally distinct M/L cone pigment. This allows heterozygous females, those getting different versions of the gene on their two X-chromosomes, to produce two spectrally distinct M and L pigments. Because they lack a second X-chromosome, male New World monkeys do not have a second M/L pigment and neither do homozygous females. The consequence is that among these New World species there is significant polymorphism of cone photopigments and color vision. There are two lineages of New World monkeys that do not follow this plan – one has routinely trichromatic color vision like the catarrhines, the other has lost color vision entirely due to gene mutation.

Among the species in the third major branch of the primate order, the more primitive strepsirrhines (lemurs, bushbabies, etc.), there is considerable variation in M/L pigment arrangements. Some species are like nonprimate mammals in routinely having only two classes of cone photopigments; others have only a single cone pigment and must therefore be monochromatic, while still others resemble the majority of New World monkeys, that is, they are polymorphic, with some females having a total of three types of cone photopigment while all other individuals have only two.

There are a number of clear instances of evolutionary change having occurred in primate cone opsin genes. As the preceding summary documents, these would include (1) loss of S-opsin gene function in some platyrrhine and strepsirrhine lineages, (2) the appearance of opsin gene polymorphisms in most platyrrhines and in a few of the strepsirrhines, and (3) the addition of a third opsin gene in all catarrhines and in some female platyrrhine and strepsirrhines. Finally, there are the intriguing polymorphisms of X-chromosome opsin genes found in humans. These include the much-studied variations classically associated with color vision defects, found mostly in males, and more subtle variations detected among both males and females. Evolutionary change of this kind is believed to be virtually absent in other catarrhines and so must have arisen in human history. There are clear variations in the incidence of these polymorphisms among human populations. The meaning of that fact has led to much speculation, but relatively little real insight.

Evolution of SWS1 Cone Pigments

Cone pigments derived from genes in the SWS1 family fall into two spectral classes having respective peaks in the UV, with λ_{max} of ~360–390 nm, and at

Table 1 Cone photopigments and color vision in primates

Suborder[a]	*Infraorder*[a]	*Exemplars*	*Cone pigments and color vision*[b]
Strepsirrhini	Lemuriformes Chiromyiformes Loriformes	Lemur, loris, galago	Monochromatic; dichromatic; polymorphic
Haplorhini	Tarisiformes	Tarsier	Dichromatic
Haplorhini	Eusimiiformes Platyrrhini	New World monkeys	Majority polymorphic; one genus is monochromatic; one genus is routinely trichromatic
Haplorhini	Eusimiiformes Catarrhini	Old World monkeys, apes, humans	Routinely trichromatic (with notable polymorphism in humans that is not seen in other catarrhines)

[a]Classification follows that of Groves C (2001) *Primate Taxonomy*. Washington: Smithsonian Institution Press.
[b]Dimensionality of color vision inferred from cone pigment complement and direct tests of color vision. Further description is given in the text.

longer wavelengths (λ_{max} of ~395–435 nm). The latter are identified as S pigments in Figure 2, although elsewhere they have alternatively been referred to as VS (violet-sensitive) pigments. The picture relating gene sequence changes and spectral tuning for the pigments in this group is less clear-cut than for the LWS pigments. Comparative examinations of opsin gene sequences have led to the conclusion that the ancestral pigments in this group were UV pigments. In different lineages, these have been converted to S pigments and, sometimes, then reconverted back to UV pigments. The amino acid substitutions associated with tuning changes may involve as many as eight residues, with perhaps three that are key to spectral shifting.

Until about 15 years ago, it was believed that, although the locations of their λ_{max} values may vary from species to species, all mammals had S pigments. It is now known that this is not true, that in fact some mammals possess the ancestral UV sensitivity. The evolutionary events that might account for this essential difference among the mammals are unknown. Mammals with UV pigments are mostly rodents (common examples: house mouse and rat), but some rodents have acquired S pigments (e.g., most members of the squirrel family), so the shift from UV to S pigments must have occurred in the more distal branches of rodent phylogeny. The former are nocturnal, the latter diurnal, so it might be thought that a shift to a diurnal lifestyle was somehow linked to a change from a UV to an S pigments. However, there are many mammals possessing S pigments that are classified as nocturnal, and at least one rodent classified as diurnal (the Chilean *Octodon degus*) is known to have a UV photopigment. So there seems to be no completely compulsive relationship between lifestyle and S pigment type. Further clouding this picture is the fact that it is not at all obvious how heightened sensitivity to the UV might particularly benefit a nocturnal animal.

An interesting twist in the story of mammalian SWS1 pigments is that there has been a loss of function of these opsin genes in some lineages. This was first discovered in a search for S pigments in the retinas of two species of nocturnal primates (the owl monkey and the bush baby). No S pigment could be found and, subsequently, it was discovered this was due to the fact that the SWS1 genes in these animals contained fatal mutational changes that obviated the expression of any pigment. This loss seems to be specieswide. That fact carries the implication that the loss of S pigment must be in some way adaptive. Subsequently, a similar loss of S cone function has been documented for some rodents and for some carnivores. Remarkably, this loss of function of the S-cone opsin gene is apparently universal among many species of marine mammal, including all of the pinnipeds and all of the cetaceans. At present, there is no satisfactory explanation of what caused some mammalian lineages to abandon their S cones and, with them, the potential for a dimension of color vision and some additional sensitivity to the short wavelengths.

Cone Pigment and Color Vision Evolution

As the preceding discussion makes clear, considerable progress has been made toward understanding the evolution of cone opsin genes and the cone pigments they specify. Since, as already noted, inferential linkages can be drawn between cone pigments and color vision, it is tempting to suppose that cone pigment evolution is also the evolution of color vision. That equation is too simplistic – there is much more to color vision than just cone pigments. In particular, in order to support color vision, the nervous system must have a means of comparing signals generated in cones containing different types of photopigments. Consequently, adding a new type of photopigment to a retina will broaden the window for spectral absorption (i.e., a larger range of wavelengths will have a probability of being absorbed by the pigment), but it will not automatically add a new dimension of color vision. Since simply capturing more photons may by itself prove advantageous, a new pigment type conceivably could evolve in the absence of any new color vision.

Even assuming that the evolution of a new cone pigment adds new color vision to the perceptual repertoire, the quality of that new capacity will depend on factors quite beyond the mere appearance of a new pigment type. In particular, the total number of cones containing the various pigment types and their spatial distribution across the retina can have a major impact on how acute the color vision capacity will be. For instance, both the domestic cats and humans with deuteranopic color vision (a dichromatic defect) have two photopigment types with respective spectral peaks that are not too greatly different for the two species. However, the acuteness of color discrimination in the color-defective human system is much greater than that in the cat, simply because of vastly greater number of cones in the human retina. Finally, to add color vision, one needs to add a new cone pigment type and assure that this new pigment type gets selectively expressed in cones. Coexpression of cone pigments (i.e., two pigments being present in one photoreceptor) does in fact occur in some rodent retinas, and this arrangement allows, at best, for only a very weak color vision capacity.

Summary

From the current understanding of the relationships between opsin genes and photopigments, it is now possible to discern the stages through which contemporary vertebrates have acquired their cone pigments. Because the linkages between photopigments and visual behavior have been much studied in these species, one can, with appropriate due caution, use ideas about cone pigment evolution to provide predictive insights into the visual capacities of ancestral species and this may, in turn, permit an enhanced understanding of their photic worlds and their visual behaviors.

See also: Color Vision; Non-Photoreceptor Photoreception; Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Ganglion Cells: Anatomy; Retinal Color Mechanisms; Retinal Development: An Overview.

Further Reading

Bowmaker JK (1998) Evolution of colour vision in vertebrates. *Eye* 12: 541–547.

Collin SP and Trezise AEO (2004) The origins of colour vision in vertebrates. *Clinical and Experimental Optometry* 87: 217–233.

Dominy NJ (2004) Color as an indicator of food quality to anthropoid primates: Ecological evidence and an evolutionary scenario. In: Ross CF and Kay RF (eds.) *Anthropoid Origins: New Visions*, pp. 615–644. New York: Kluwer Academic/Plenum.

Grooves C (2001) *Primate Taxonomy*. Washington, DC: Smithsonian Institution Press.

Hisatomi O and Tokunaga F (2002) Molecular evolution of proteins involved in vertebrate phototransduction. *Comparative and Biochemical Physiology B* 133: 509–522.

Hunt DM, Cowing JA, Wilkie SE, Parry JWL, Poopalsundaram S, and Boamaker JK (2004) Divergent mechanisms for the tuning of shortwave sensitive visual pigments in vertebrates. *Photochemical and Photobiological Sciences* 3: 713–720.

Hunt DM, Jacobs GH, and Bowmaker JK (2005) The genetics and evolution of primate visual pigments. In: Kremers J (ed.) *The Primate Visual System: A Comparative Approach*, pp. 73–126. Chichester: Wiley.

Jacobs GH (1993) The distribution and nature of colour vision among the mammals. *Biological Reviews* 68: 413–471.

Jacobs GH (2004) Photopigment variations and the evolution of anthropoid vision. In: Ross CF and Kay RF (eds.) *Anthropoid Origins: New Visions*, pp. 645–664. New York: Kluwer Academic/Plenum.

Jacobs GH and Rowe MP (2004) Evolution of vertebrate colour vision. *Clinical and Experimental Optometry* 87: 206–216.

Nathans J (1999) The evolution and physiology of human color vision: Insights from molecular genetic studies of visual pigments. *Neuron* 24: 299–312.

Neitz J, Carroll J, and Neitz M (2001) Color vision: Almost reason enough for having eyes. *Optics and Photonics News* 12: 26–33.

Yokoyama S (2002) Molecular evolution of color vision in vertebrates. *Gene* 300: 69–78.

Yokoyama S and Radlwimmer FB (1999) The molecular genetics of red and green color vision in mammals. *Genetics* 153: 919–932.

Relevant Website

http://www.ncbi.nlm.nih.gov – Opsin sequence information is catalogued at the NCBI website.

Congenital Muscular Dystrophy

C G Bonnemann, Children's Hospital of Philadelphia; and University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Published by Elsevier Ltd.

Introduction and History of the Concept of Congenital Muscular Dystrophy

Congenital muscular dystrophy (CMD) refers to neuromuscular conditions with clinically apparent muscle weakness at birth or in early infancy, and in which the muscle biopsy reveals a dystrophic or myopathic picture without other diagnostic features suggestive of one of the structural congenital myopathies or of a mitochondrial or denervating disorder. The congenital muscular dystrophies are genetically and clinically quite heterogeneous. However, with the ongoing clarification of the genetic and molecular basis of an increasing number of conditions, new delineations of molecular subgroups and their associated clinical phenotypes have been introduced. The following is a brief historical review of the evolution of the concept of CMD up to the molecular era in CMD research.

The concept of congenital muscular dystrophies was first introduced by Frederick Eustace Batten in 1903, who carefully described patients with a

> disease [that] is congenital or starts in early infancy, and is characterized by smallness, lack of power, and loss of tone in all the muscles of the body without localized atrophy or hypotrophy of individual muscles or groups of muscles.(...) The disease is but slowly progressive for the child may, as development takes place, learn to sit up, and possibly to stand with support. The child usually learns to talk at the normal age, and intellectually is often in advance of his years. As a rule these children never learn to walk but adopt some strange method of getting about.

This phenotype corresponds closely to what today is known as laminin 2 (merosin)-negative CMD. In 1930, Otto Ullrich gave an equally detailed description of a phenotype which he called "scleratonic muscular dystrophy" because of the striking coexistence of low muscle tone and joint laxity as well as contractures. This condition is now is known to be caused by deficiency of collagen VI and is referred to as CMD type Ullrich. In 1960 in Japan, Yukijo Fukuyama published a series of papers describing in great detail a form of CMD with abnormalities of brain formation as well as with mental retardation that was prevalent in Japanese patients. In 1973, Victor Dubowitz recognized the clinical phenotype of rigid spine CMD, a proportion of which is now known to be caused by SEPN1 mutations. Muscle–eye–brain disease was described in 1977 by Pirkko Santavouri in the Finnish population. Arthur Earl Walker in 1942 and Mette Warburg in 1978 described a syndrome of severe eye and brain malformations that was designated as Walker–Warburg syndrome by Roger Williams and Verne Caviness in 1984 who also recognized its relationship to Fukuyama CMD and thereby muscular dystrophy. This was confirmed by William Dobyns in 1989 who proposed diagnostic criteria for Walker–Warburg syndrome and also referred to the cortical malformation as lissencephaly type II. The molecular age of CMD begins with the description of merosin/laminin 2 deficiency by Fernando Tome in 1994 and molecular confirmation of LAMA2 gene mutations by Pascale Guicheney's group in 1995 as the first identification of a gene mutated in CMD. Subsequently, there has been a considerable expansion of genetically defined conditions with the identification of mutations in various genes in the group of α-dystroglycanopathies, in the collagen VI genes, as well as in selenoprotein N. This molecular expansion is likely to continue.

The identification of this expanding number of genes involved in CMD has influenced the classification of the CMDs, which is now based largely on pathophysiolgical principles. However, since some of the defining clinical aspects are in good agreement with the molecular classification, the best approach to this group of disorders will take both the clinical syndrome and its range of underlying possible genetic etiologies into account. It is of interest to note that many disorders that are now classified as CMD affect components of the extracellular matrix as well as the interface between matrix and muscle. These CMDs of the extracellular matrix include those affecting α-dystroglycan glycosylation and α-dystroglycan's ligand merosin/laminin 2 as well as collagen VI. Intracellular molecules involved in causing CMD include selenoprotein N (SEPN1) as well as the inner nuclear matrix component lamin A/C. Thus, there is no single pathophysiological concept that would be applicable for all disorders now subsumed under CMD.

Disorders of α-Dystroglycan Glycosylation

Dystroglycan is a central component of the dystrophin-associated glycoprotein complex. This complex is composed of proteins that co-purify with dystrophin on wheat germ agglutinin columns from muscle membrane preparations. The complex includes an intracellular subcomplex (including syntrophin and dystrobrevin) as well as membrane spanning

molecular components. The membrane spanning part of the complex is composed of α- and β-dystroglycan as well as the sarcoglycan complex. β-Dystroglycan binds to dystrophin's cysteine-rich C-terminal domain and to α-dystroglycan on the extracellular side. α-Dystroglycan in turn binds to laminin 2 (merosin), one of the major components of basement membranes around muscle. Other α-dystroglycan ligands are known, including neurexin in the central nervous system and agrin at the neuromuscular junction. Both α- and β-dystroglycan are transcribed from the same gene DAG1 on chromosome 3p21 and are subsequently cleaved by posttranslational processing. α-Dystroglycan in particular undergoes significant further posttranslational glycosylation, including N-linked glycosylation at one site, but more important, O-linked glycosylation in the serine and threonine-rich region of the peptide, giving it mucin-like characteristics. Thus, the molecular weight of the processed α-dystroglycan is 156 kDa in muscle, whereas the weight of the core peptide is only 74 or 75 kDa. The molecular weight is lower in brain (120 kDa) and in heart (140 kDa), suggesting that there are tissue-specific modifications of this process. The O-linked glycosylation of α-dystroglycan that is of particular relevance for the pathogenesis of the CMDs is of the O-mannose linked type, a rather unusual form of O-linked glycosylation for which only very few targets are known. The enzymatic machinery involved is directly affected in the pathogenesis of an expanding class of CMDs now also referred to as secondary dystroglycanopathies. Mutations in the α-dystroglycan gene are not known to occur and are probably not compatible with life because inactivation of the *dag1* gene in the mouse is lethal in very early embryonic development (at the formation of Reichert's membrane). Some of the genes mutated in the group of disorders of O-linked glycosylation of α-dystroglycan are actual glycosyltransferases with proven enzymatic activities (protein *O*-mannosyl transferase 1 and 2 (POMT1 and POMT2, respectively) and protein *O*-mannose β1,2-*N*-acetylglucosaminyltransferase 1 (POMGnT1)), whereas the roles of others are less clear (Fukutin, Fukutin-related protein/FKRP, and LARGE), although putative functions have been suggested. The end result of mutations in all of these genes is an incomplete *O*-mannose-linked glycosylation of α-dystroglycan with direct effects on its function as a receptor for extracellular matrix components such as laminin 2 (merosin), agrin, neurexin, and presumably others. This impaired binding capacity of α-dystroglycan appears to lie at the basis of both the muscular dystrophy and the associated brain and eye malformations. The resulting muscular dystrophy is associated with membrane fragility, as evidenced by the elevated creatine kinase (CK) levels in affected patients, whereas the typical cortical malformation is a disorder of neuronal migration referred to as lissencephaly type II. This type of lissencephaly is associated with overmigration of neurons beyond the limits of the regular cortical plate, likely because of breaches in the glia limitans, the basement membrane covering the cortical surface of which α-dystroglycan is a crucial component. The result is a very disorganized layer of overmigrated neurons covering the original brain surface with resulting loss of normal gyration and increased thickness of the cortex. Observations in mouse models carrying tissue-specific conditional inactivations of the *dag1* gene to circumvent the early embryonic lethality suggest that most, if not all, of this pathology is directly related to loss of normal α-dystroglycan function. The biochemical and immunohistochemical hallmark of all of these disorders is a reduction of the glycoslyated epitope of α-dystroglycan (evident on Western blotting as well as on immunohistochemistry) while the core protein for α-dystroglycan is unaffected.

Several clinical syndromes are known in this group of disorders; their classification depends on the involvement of muscle, brain, and eyes. Ongoing genotype/phenotype analysis suggests that most of these phenotypes can be caused by mutations in more than one of the genes related to the pathway and that each of the genes in this pathway has a broadening range of phenotypes associated with it.

Walker–Warburg Syndrome

The most severe of the phenotypes is Walker–Warburg syndrome (WWS). Affected infants have extensive lissencephaly type II, hydrocephalus, absence of the corpus callosum, and fusion of the hemispheres due to overmigrating neurons. Rarely, there may be an occipital encephalocele associated with this malformation. There is significant pontocerebellar hypoplasia. Eye abnormalities may be severe and include microphthalmia, malformations of the anterior chamber, cataract, congenital glaucoma, and retinal dysplasias. Mental retardation is profound and epilepsy develops frequently. Life expectancy is limited so that only 5–10% of patients survive beyond 5 years. Genes known to cause WWS include POMT1, mutated in approximately 20% of patients; other genes known to be involved include POMT2, Fukutin, as well as FKRP, but together these do not account for all patients with WWS.

Muscle–Eye–Brain Disease

Muscle–eye–brain disease (MEB) is somewhat milder in that the brain malformation may be less complete, there may be lissencephaly as well as pachygyria, or sometimes there may also be polymicrogyric-appearing cortex. Abnormalities of the white matter

on magnetic resonance imaging (MRI) may be present, even transiently. Most consistently, there is a small pons as well as cerebellar hypoplasia with conspicuous cerebellar cysts on imaging. Eye involvement includes congenital myopia, glaucoma, as well as retinal dysplasias and optic disc abnormalities. The myopia and retinal abnormalities can be progressive. Mental retardation is invariant with a high possibility of the development of epilepsy. Patients may gain the ability to walk; however, there may be loss of ambulation because of increasing spasticity, ataxia, and weakness. This condition was originally described in Finland but is now known to occur worldwide. Mutations in MEB were first found in POMGnT1, located on chromosome 1p32–34, but mutations have also been found in FKRP as well as in POMT1 and -2.

Fukuyama Congenital Muscular Dystrophy

Fukuyama congenital muscular dystrophy (FCMD) is related to MEB but has its highest prevalence in Japan based on a founder mutation in the gene Fukutin (retrotransposon insertion into the gene locus). The spectrum of possible brain malformations in FCMD is very similar to MEB, whereas the eye findings tend to be less severe, mostly consisting of myopia. Profound mental retardation and the development of epilepsy are common. Patients in other countries with a FCMD-like clinical disease have had mutations in other genes along the pathway, whereas mutations in Fukutin outside of Japan have been associated with a more severe WWS-like disease and also with a severe (initially steroid-responsive) LGMD form.

Other Phenotypes

There are also patients with CMD or milder muscular phenotypes associated with mental retardation and brain malformations that may range from substantial to minor brain malformations, such as cerebellar cysts only, or with a normal appearing brain on imaging. These clinical syndromes may be associated with mutations in the same genes that are also responsible for the more severe phenotypes described. Mutations in the gene LARGE are associated with abnormal white matter and other subtle brain abnormalities, mental retardation, and muscular dystrophy but no eye involvement in the initial case, thus not formally qualifying as MEB. This condition is referred to as MDC1D. There is evidence that LARGE interacts with α-dystroglycan in the Golgi apparatus to facilitate subsequent *O*-mannosyl-linked α-dystroglycan glycosylation and in addition functions as glycosyltransferase in its own right. There are also patients with a milder LGMD-like muscular dystrophy, with mental retardation and microcephaly that are caused by mutations in POMT1, the gene originally associated with WWS. This disorder is now also classified among the limb–girdle muscular dystrophies as LGMD2K. Finally, at the mildest end of the spectrum, at least from the perspective of brain involvement, there are patients with muscular dystrophy only, without brain involvement or mental retardation. Mutations in the gene FKRP cover the broadest range of possible severity, ranging from severe WWS to a late adulthood onset form of pure LGMD. The severity of FKRP-associated muscular dystrophy can vary from severe CMD but normal brain imaging and intelligence (referred to as MDC1C) via Duchenne- and Becker-like severities to a milder adult-onset form of LGMD (referred to as LGMD2I). This supports the notion that these disorders should be viewed as a spectrum of a broadness that currently is unique in neuromuscular diseases. From a clinical perspective, it is important to realize that there may be cardiac involvement in addition to the skeletal muscle involvement in FKRP-related disease and also in the other dystroglycanopathies.

It is clear that there must be additional undiscovered genes causing hypoglycosylation of α-dystroglycan because not all patients with biochemical evidence for an α-dystroglycanopathy can be accounted for by mutations in the currently known genes (see Table 1 for additional possible phenotypes associated with biochemical evidence for α-dystroglycan deficiency).

Laminin 2 (Merosin)-Deficient Congenital Muscular Dystrophy

Mutations in a component of the ligand for α-dystroglycan, laminin 2 (also referred to as merosin), cause a CMD that occurs with an estimated frequency of 30–40% of all CMDs in Europe and has been recognized in most countries. It may be less common in North America. This condition is also referred to as MDC1A. Laminin 2 (merosin) is a heterotrimer composed of laminin α_2, laminin β_1, and laminin γ_1. The mutations occur in the LAMA2 gene on chromosome 6q22–23 coding for the heavy chain of laminin 2, laminin α_2. Laminin α_2 is also a component of laminin 4 occurring at the neuromuscular junction, which is therefore also affected by LAMA2 mutations, although the clinical consequences of this are not clear. Laminin 2 is a crucial component of basement membranes around muscle fibers, Schwann cells in the peripheral nervous system, whereas in the brain it is found around capillary blood vessels between the astrocyte foot process and the vessel and as a component of the lamina limitans on the surface of the brain. Laminin 2 also interacts with integrin β_1/α_7 as another receptor on muscle and also with components of the basement membrane, including nidogen, perlecan, and collagen type IV. Mutations in integrin α_7 have been described

Table 1 The congenital muscular dystrophies

Disease entity	*Locus* *protein product* *gene symbol*	*Helpful clinical features*	*CNS involvement*	*Laboratory testing*
Primary merosin/laminin 2 deficiency				
CMD with primary laminin 2 (merosin) deficiency (MDC1A)	6q22–q23 Laminin α_2 LAMA2	Sitting and standing with support as maximal motor ability if complete deficiency; neuropathy; epilepsy in approximately 30%; possible subclinical cardiomyopathy; generally normal mental development	Abnormal white matter signal (T2 MRI); 5% occipital pachy- or agyria; pontocerebellar atrophy (rare)	Mostly complete laminin α_2 deficiency on IH/WB; secondary reduction of integrin α_7 possible; mutation analysis[b]
α-Dystroglycanopathies – secondary merosin/laminin 2 deficiency				
CMD with partial merosin deficiency (MDC1B)	1q42 Not known	Rare; variety of severity; delayed onset possible; limb girdle weakness; generalized muscle hypertrophy; early respiratory failure possible	Abnormal white matter and structural changes possible	Partial deficiency of laminin α_2 on IH/WB; α-DG significantly reduced on IH; linkage analysis
Fukutin-related proteinopathy (MDC1C)	19q13.3 Fukutin-related protein FKRP	Often reminiscent of MDC 1A, but severity more variable, from severe CMD to LGMD; generally normal mental development; cases with structural brain involvement and mental retardation increasingly recognized, including MEB and WWS	Range from normal to significant structural abnormalities, ranging from cerebellar cysts to typical MEB and WWS	α-DG with diminished MW on WB, or reduction of IH using antibodies against glycosylated isotopes; secondary reductions in laminin α_2 on IH/WB; mutation analysis[b]
LARGE-related CMD (MDC1D)	22q12.3 Acetylglucosaminyltransferase-like protein LARGE	Only one patient described; congenital muscular dystrophy with profound mental retardation may eventually blend with the MEB/WWS spectrum	White matter changes; hypoplastic brain stem; mild pachygyria (similar to MEB)	IH/WB comparable to MDC1C; mutation analysis[b]
Fukuyama CMD (FCMD)	9q31 Fukutin FCMD	Frequent in Japanese population; never walk; mental retardation; epilepsy common – clinical overlap with MEB	Lissencephaly type II/ pachygyria; hypoplastic brain stem cerebellar abnormalities	IH/WB comparable to MDC1C; mutation analysis
Muscle–eye–brain disease (MEB)	1q32–q34 Protein-O-linked mannose β1,2-*N*-acetylglucosaminyltranferase 1 POMGnT1, also FKRP, Fukutin	Severe weakness and mental retardation; large head; prominent forehead; flat midface; walking rarely achieved; ocular involvement (e.g., severe myopia, retinal hypoplasia); deterioration because of spasticity	Lissencephaly type II/ pachygyria; eye malformations; brain stem and cerebellar abnormalities	IH/WB comparable to MDC1C; mutation analysis (genetic heterogeneity)
Walker–Warburg syndrome (WWS)	9q34.1 O-Mannosyltranferase 1 POMT1, also POMT2, FKRP, Fukutin	Severe; lethal within first years of life because of severe CNS involvement	Lissencephaly type II/ pachygyria; hydrocephalus; encephalocele; hypoplastic brain stem; cerebellar abnormalities; eye malformations	IH/WB comparable to MDC1C; mutation analysis (genetic heterogeneity)

Other matrix disorders (merosin/laminin 2 positive)				
Ullrich CMD (UCMD)	21q22.3 and 2q37 $\alpha_{1/2}$ and α_3 collagen VI COL6A1, COL6A2, COL6A3	Distal joint hyperextensibility; proximal contractures; motor abilities variable; precludes independent ambulation in severe cases; soft palmar skin	No	IH for collagen VI with severe to mild deficiency; mutation analysis[a]
Integrin α_7	12q13 Integrin α_7 ITGA7	Very rare; delayed motor milestones; walking within 2–3 years of life	No	Absence of integrin α_7 on IH (secondary reduction possible); mutation analysis[b]
Other CMDs				
Rigid spine muscular dystrophy (RSMD)	1p36–p35 Selenoprotein N SEPN1	Delayed walking; predominantly axial weakness with early development of rigidity of the spine; restrictive respiratory syndrome	No	Normal expression of laminin α_2; mutation analysis
CMD merosin-positive	4p16.3	Severe muscle weakness of trunk and shoulder girdle muscles; mild to moderate involvement of facial, neck, and proximal limb muscles; normal intelligence	No	Normal expression of laminins; dystrophin; sarcoglycans and β-dystroglycan
Lamin A/C-related CMD	1 Lamin A/C LMNA	Early motor deterioration; prominent axial weakness with dropped head syndrome; early development of spinal rigidity	No	Lamin A/C IH present
CMD with microcephaly/calf hypertrophy	Not known	Joint contractures associated; severe psychomotor retardation; no walking; striking enlargement of the calf and quadriceps muscles; CK grossly elevated	Megacisterna magna; cerebellar hypoplasia; white matter changes	Mild to moderate partial deficiency of laminin α_2 on IH
CMD with adducted thumbs	Not known	Rare; adducted thumbs; toe contractures; generalized weakness; delayed walking; ptosis; external ophthalmoplegia; mild mental retardation	Mild cerebellar hypoplasia	Normal expression of laminin α_2 and α-DG on IH
CMD with mental retardation and microcephaly	Not known, FKRP not yet excluded	Microcephaly; delayed psychomotor development; generalized muscular wasting and weakness with mild facial involvement; calf pseudohypertrophy; joint contractures; severe mental retardation	Pontocerebellar hypoplasia; focal cortical dysplasia; white matter changes; cerebellar cysts	Normal expression of laminin α_2
CMD with cerebellar atrophy	Not known	Delayed motor milestones; mild intellectual impairment	Moderate to severe cerebellar hypoplasia; no white matter abnormalities	Normal expression of laminin α_2

[a]Currently not available as diagnostic test; only performed on a research basis.

CK, creatine kinase; CMD, congenital muscular dystrophy; CNS, central nervous system; DG, dystroglycan; IH, immunohistochemistry; MEB, muscle–eye–brain disease; MW, molecular weight; WB, Western blot.

Adapted from Schessl J, Zou Y, and Bönnemann CG (2006) Congenital muscular dystrophies and the extracellular matrix. *Seminars in Pediatric Neurology* 13: 80–89.

in three patients with a congenital muscle disease, but this appears to be extremely rare. Of the six domains that are recognized in laminin 2, the C-terminal G-domain is the most important domain for interaction with α-dystroglycan; specifically, the globular subdomains G4 and G5 have been shown to interact with α-dystroglycan. Mutations abolishing the expression of this domain as well as that of domains I and II are usually associated with severe disease. In the typical case, there is a complete deficiency of laminin 2 in the muscle biopsy on the basis of truncating mutations. Partial deficiencies may be caused by compound heterozygosity of a truncating mutation and an in-frame deletion or a missense mutation. The absence of laminin 2 disrupts the lamina dense of the basement membrane, causing dysfunction of the sarcolemma, as evidenced by elevated CK levels in the plasma of patients and in mouse models of laminin 2 deficiency. There is evidence that there may be both membrane instability and an increased rate of apoptosis in mice with laminin 2 deficiency.

Clinically, primary deficiency of laminin 2 is a severe disease with profound hypotonia at birth. Some patients may have distal contractures; however, these are not prominent. Standing with support is frequently the maximal motor achievement, whereas free ambulation typically is not achieved, although rare exceptions exists. Dysphagia and later respiratory compromise are common. Milder clinical courses can be seen with partial deficiency of laminin 2, resembling a milder CMD or a limb–girdle muscular dystrophy, with sometimes prominent contractures. A typical finding in patients with laminin 2 deficiency is abnormal white matter signal on T2-weighted MRI. This abnormal signal spares the more compacted fiber tracks such as the corpus callosum or the internal capsule. The nature of this abnormality is not completely understood but likely does not reflect de- or dysmyelination; however, it may reflect abnormal water content of the white matter. A small but significant percentage of patients with primary laminin 2 deficiency have structural abnormalities of the brain as well, including occipital pachygyria and cerebellar hypoplasia. There is also a higher incidence of epilepsy in patients with primary laminin 2 deficiency (approximately 30%). Thus, there are some clinicopathological features that are common to both the α-dystroglycanopathies and primary laminin 2 deficiency.

Disorders of Collagen VI (CMD Types Ullrich and Bethlem)

Mutations in the extracellular matrix component collagen type VI underlie a CMD known as type Ullrich as well as its milder allelic counterpart Bethlem myopathy. Collagen VI is composed of three α chains encoded by three genes: COL6A1 and COL6A2 on chromosome 21q22 or COL6A3 on chromosome 2q37. The individual chains have relatively short collagenous domains of 335–336 amino acids as well as N-terminals and C-terminal domains, frequently including von Willebrandt factor A domains as the most prominent component. The C-terminal globular domain in the α_3 chain is the largest. Collagen VI undergoes a complex assembly process in which all three α chains are required to form a heterotrimeric monomer, presumably by C-terminal association followed by zipper-like formation of the triple-helical collagenous domain. This monomer then undergoes further intracellular assembly that is stabilized by the formation of disulfide bridges involving cysteine residues that are located at the N-terminal part of the triple-helical domains of all three chains. Tetramers are then secreted and assemble end-to-end in the extracellular matrix to form beaded microfibrils. Collagen VI is a prominent component of most extracellular matrices, in which it forms beaded microfibrils that are found in close connection with cells and basement membranes as well as in the interstitial space in many tissues, including muscle, tendon, skin, cartilage, and intervertebral discs. Thus, collagen VI mutations present as a combination of disorders of both muscle and connective tissue in that joint hyperlaxity and contractures as well as abnormal skin are important aspects of the clinical phenotype in addition to the weakness. It is still unclear how collagen VI links or relates to the DAP complex, if at all. However, the molecular basis and consequences of this disorder appear to be different from those of the α-dystroglycanopathies in that the disease is less overtly 'dystrophic' but still leads to significant weakness and disability. There is evidence from work in a mouse model of collagen deficiency suggesting that apoptosis mediated by mitochondrial dysfunction may play a major role in the pathogenesis of the disease. This process appears to be mediated by a breakdown of the mitochondrial potential because of dysfunction of the mitochondrial permeability transition pore, triggering apoptosis. The transition permeability pore can be blocked by cyclosporine A and some of its derivates, thus providing the potential for treatment for this group of disorders.

CMD Type Ullrich

Clinically, there are two recognized disorders associated with mutations in collagen type VI – the more severe type Ullrich and the milder type Bethlem. In the Ullrich type of CMD (UCMD), there is hypotonia and weakness present at birth in conjunction with striking hyperlaxity of the distal joints. The hyperlaxity in particular is present in the vast majority of

affected infants, it is often associated with a characteristic prominent calcaneus. Typically, there are also contractures present that coexist with the distal laxity, either present at birth, including kyphoscoliosis or torticollis, or developing in early childhood, affecting more proximal joints and the spine. In severe cases, independent ambulation is not achieved or is lost after a variable and delayed period of ambulation. This deterioration is also caused by the progressive nature of the contractures. A prominent but treatable complication is respiratory insufficiency. There are also abnormalities of skin that include abnormal scar formation, peculiar softness of the skin in the hands and soles, and keratosis pilaris over the extensor surfaces of the arms and legs. In contrast to the CMD types associated with abnormalities of α-dystroglycan and laminin 2, there is no excessive elevation of CK, which is frequently normal or only mildly elevated. The immunohistochemical hallmark of this disorder is the absence of collagen VI from the muscle or just from the basement membrane.

Bethlem Myopathy

In the milder phenotype of Bethlem myopathy, there is also early evidence of hypotonia, although the weakness is much less pronounced and walking is achieved at a normal or slightly delayed age. Later contractures set in at the Achilles tendons, elbows, pectoralis, and the long finger flexors. However, the degree of contractures is variable and can be pronounced even with only mild weakness, or there can be minimal contractures in conjunction with much more pronounced weakness. Ambulation is usually preserved until adulthood, although approximately 50% of patients will require assistive devices with ambulation by age 60 years. Restrictive respiratory impairment can be a complication in Bethlem myopathy, as can abnormal scar formation and the formation of keloids.

Mutations causing UCMD occur in all three chains forming collagen VI and are of two types. Autosomal recessive mutations were found initially, often as null mutations, causing complete absence of collagen VI. *De novo* dominant negative mutations have been added as a mutational mechanism in approximately half of patients. Mutations in the latter mechanism are deletions or splice mutations causing in-frame skipping of exons located in the N-terminal part of the triple-helical domains of all three collagen VI chains, allowing for full assembly of the mutant chain into the collagen VI heterotrimer and subsequently into the secreted tetramer so that at the end of the assembly process only 1 of 16 tetrameres will be composed of wild-type chains only. In Bethlem myopathy, mutations have been dominant, causing less severe interference with collagen VI assembly compared to the UCMD mutations. It is becoming increasingly clear that there are phenotypes intermittent between Ullrich and Bethlem, suggesting that there may be a spectrum of severity from severe Ullrich to milder Bethlem disease.

Congenital Muscular Dystrophy with Spinal Rigidity

There are two forms of CMD caused by mutations in intracellular genes that are associated with prominent rigidity of the spine but which are also associated with other phenotypes.

Selenoprotein N-Related CMD

Mutations in selenoprotein N (SEPN1) are associated with CMD with spinal rigidity (RSMD1) and also with other morphological phenotypes, including multi-minicore disease, Mallory body myopathy, and congenital fiber type disproportion, whereas the clinical phenotype is very similar in all of these histologically different conditions. SEPN1 is localized to the endoplasmic reticulum, but its functions there are unknown. Selenoproteins in general incorporate a selenocysteine residue at the position of an in-frame UGA stop codon. For this to occur, the presence of a hairpin structured selenocysteine insertion sequence (SECIS) in the 3′ UTR of the mRNA as well as other protein factors are required. Some selenoproteins of known function are involved in redox systems and in thyroid hormone synthesis, such as in members of the glutathione peroxidase or iodothyronine deiodinase families. Severe nutritional selenium deficiency can cause a form of cardiomyopathy in children referred to as Keshan disease (endemic in certain regions of China where the selenium content in soil is extremely low).

Clinically, patients present with early onset of hypotonia but only mildly delayed motor milestones. There is atrophy of muscles, particularly affecting the inner thigh. Early spinal rigidity can be noted as early as 4 or 5 years of age and is clearly evident by 10 years of age. There may also be scoliosis. Of clinical importance is the early onset of nocturnal respiratory failure, which requires regular careful monitoring of oxygenation in sleep and the timely introduction of assisted ventilation at night. There is no known cardiac involvement.

Lamin A/C-Related CMD

Lamin A/C is an intermediate filament of the inner nuclear matrix with direct interactions with emerin, the protein mutated in X-linked Emery–Dreifuss

muscular dystrophy. Dominant mutations in the gene coding for lamin A/C (LMNA) more frequently cause autosomal dominant Emery–Dreifuss muscular dystrophy, but a number of other and disparate phenotypes are also caused by mutations in this gene, including LGMD1B, arrhythmogenic familial cardiomyopathy, lipodystrophy Dunnigan type, a form of CMT2B, mandibuloacral dysplasia, Hutchinson–Guilford progeria, and variants of Werner syndrome. It has become apparent that there is also an early onset presentation of the disease presenting with significant axial weakness including head drop because of weakness of the neck extensors, as well as early rigidity of the spine, thus making this an important differential diagnosis for patients with CMD and early rigidity of the spine.

Other CMD phenotypes for which genes have not yet been identified are included in Table 1.

See also: Axonal Pathfinding: Extracellular Matrix Role; Dystrophin, Associated Proteins, and Muscular Dystrophy; Glycosylation: General Overview; Mitochondrial Dysfunction in Nervous System Injury.

Further Reading

Barresi R and Campbell KP (2006) Dystroglycan: From biosynthesis to pathogenesis of human disease. *Journal of Cell Science* 119(part 2): 199–207.

Bertini E and Pepe G (2002) Collagen type VI and related disorders: Bethlem myopathy and Ullrich scleroatonic muscular dystrophy. *European Journal of Paediatric Neurology* 6: 193–198.

Brockington M and Muntoni F (2005) The modulation of skeletal muscle glycosylation as a potential therapeutic intervention in muscular dystrophies. *Acta Myologica* 24(3): 217–221.

Colognato H and Yurchenco PD (2000) Form and function: The laminin family of heterotrimers. *Developmental Dynamics* 218: 213–234.

Davies KE and Nowak KJ (2006) Molecular mechanisms of muscular dystrophies: Old and new players. *Nature Reviews Molecular Cell Biology* 7(10): 762–773.

Jimenez-Mallebrera C, Brown SC, Sewry CA, and Muntoni F (2005) Congenital muscular dystrophy: Molecular and cellular aspects. *Cellular and Molecular Life Sciences* 62: 809–823.

Kanagawa M and Toda T (2006) The genetic and molecular basis of muscular dystrophy: Roles of cell-matrix linkage in the pathogenesis. *Journal of Human Genetics* 51: 915–926.

Lampe AK and Bushby KM (2005) Collagen VI related muscle disorders. *Journal of Medical Genetics* 42: 673–685.

Lisi MT and Cohn RD (2007) Congenital muscular dystrophies: New aspects of an expanding group of disorders. *Biochimica et Biophysica Acta* 1772(2): 159–172.

Muntoni F (2004) Journey into muscular dystrophies caused by abnormal glycosylation. *Acta Myologica* 23: 79–84.

Muntoni F and Voit T (2004) The congenital muscular dystrophies in 2004: A century of exciting progress. *Neuromuscular Disorders* 14: 635–649.

Rankin J and Ellard S (2006) The laminopathies: A clinical review. *Clinical Genetics* 70: 261–274.

Schachter H, Vajsar J, and Zhang W (2004) The role of defective glycosylation in congenital muscular dystrophy. *Glycoconjugate Journal* 20(5): 291–300.

Schessl J, Zou Y, and Bönnemann CG (2006) Congenital muscular dystrophies and the extracellular matrix. *Seminars in Pediatric Neurology* 13: 80–89.

van Reeuwijk J, Brunner HG, and van Bokhoven H (2005) Glyc-O-genetics of Walker–Warburg syndrome. *Clinical Genetics* 67: 281–289.

Connectionist Models

T T Rogers, University of Wisconsin–Madison, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Connectionism is a theoretical framework for cognition whose principal tenets are (1) that all cognitive phenomena arise from the propagation of activation among simple neuronlike processing units and (2) that such propagation is mediated by weighted synapselike connections between units. Connectionist theories are typically instantiated as computer models, that is, computer programs that simulate how activation propagates through the system of interconnected units specified by the theory. Such models have profoundly influenced virtually every subdomain of cognitive science: perception and attention; word recognition, reading, derivational morphology, and other aspects of language; episodic, semantic, and short-term memory; action and other forms of sequential processing; executive function and cognitive control; many aspects of cognitive development; and even emotion. Many researchers, however, consider the connectionist framework to be incommensurable with the mainstream view that the human mind is a symbol-processing system and, for this reason, consider connectionist theories to be fundamentally flawed, or at least insufficient to serve as a general framework for cognition. This article considers the historical roots of connectionism, reviews key aspects of the approach, and addresses some of the criticisms it faces.

History

The roots of connectionism lie in the mid-1940s, when Warren McCullough and Walter Pitts showed how individual neurons could be viewed as simple computing devices, assuming that:

1. Neurons can assume two states (on or off, that is, 1 or 0).
2. Each neuron has a fixed threshold above which it assumes the on state.
3. A given neuron's activation state is influenced by the other neurons from which it receives connections.
4. The influence of a sending neuron on the receiving neuron depends on the value of the synapse or weight connecting them.

McCullough and Pitts demonstrated that individual neurons could compute the basic two-bit logic functions (AND and OR) and that networks of such neurons could compute more complex logic functions (e.g., exclusive-OR (XOR)).

At around the same time, Donald Hebb had become interested in the cellular basis of learning and memory, and he suggested a mechanism for changing the efficacy of a synapse to support a simple form of associative learning. Specifically, he suggested that, "When an axon of cell A is near enough to excite B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (Hebb, 1949). This insight – that correlated activity in a system of neurons can produce synaptic change – provided a basis for understanding the neural mechanisms that support learning and memory.

These contributions provided the foundations for the connectionist enterprise. Like other approaches, the connectionist framework assumes that cognitive systems are information processing systems that take in information via sensory organs, transform the information to form internal representations of the environment, and from these representations generate outputs in the form of overt behaviors. In contrast to some other approaches, the connectionist framework posits that the functional elements that support cognitive information processing are something like the abstract neurons described by McCullough and Pitts – units that take on activation states depending on their inputs and transmit these states to other units via weighted connections. It also posits that the principal mechanism for learning and memory is something like that proposed by Hebb – learning is effected through changes to the weights that connect processing units; such weight changes are partly (although not necessarily completely) driven by patterns of correlated activity across units in the system; and such correlated patterns in turn depend on experience.

These foundational ideas were developed in the 1950s and 1960s, largely through the work of Frank Rosenblatt, who introduced the perceptron, a form of neural network similar in some respects to those described by McCullough and Pitts. Rosenblatt derived a learning algorithm (a rule for changing the weights between units) that allowed the perceptron to solve a much wider range of learning problems than the Hebb learning rule. Rather than changing weights in proportion to the correlation between connected units, the perceptron learning rule minimizes the error generated by an input. That is, for each input, the algorithm compares what the network actually produced to the pattern it should have produced (the target pattern) and adjusts the values of each

weight so as to reduce the difference between these (the error, sometimes referred to as delta because it describes a difference between two terms). This delta-rule learning algorithm thus requires a teaching signal (the network must be told what pattern it is to generate in response to each input), but given this signal it can learn input–output mappings that are well beyond the Hebb rule.

Rosenblatt died tragically in a boating accident in 1969 at the age of 41. The same year, in a book called *Perceptrons*, Marvin Minsky and Seymour Papert published a forceful critique of perceptron-based theories of cognitive function. The book showed that networks of simple perceptrons were incapable of learning to compute nonlinearly separable mappings between input and output; for instance, they could not learn second-order logical functions such as XOR. Largely on this basis, the authors argued that perceptrons could not, in principle, provide a basis for understanding human cognition. This book, coupled with the intensive lobbying (by Minsky) of various funding agencies and other scientists, seriously dampened general interest in and funding for neural network modeling for the next 10 years, although important advances were made during this period by a number of scientists working independently.

The Modern Period

The 1980s witnessed a renaissance of interest in connectionism, for three principal reasons. The first was the publication of some landmark papers that set aside the difficult questions about learning raised by Minsky and Papert but that used connectionist principles and implemented computer models to understand a range of cognitive phenomena. Perhaps most influential was the interactive activation and competition model of word recognition, which accounted for a very wide range of behavioral phenomena in the domain and resolved several important puzzles pertaining to lexical processing (see Figure 1). Among other things, this model introduced the notions that the component processes involved in word recognition proceed in parallel and not in serial; that processing is cascaded, so that a later process can begin to operate on the outputs of earlier processes before the earlier processes have themselves completed; and that bottom-up and top-down processes continually interact throughout processing. These notions were antithetical to the then-dominant view that cognitive processes were carried out by informationally encapsulated modules whose operation proceeded largely in serial.

The second major development was the discovery of a solution to the perceptron-learning problems raised by Minsky and Papert. The discovery was made independently by various investigators, but its importance to cognitive science was made clear in an influential article by Rumelhart, Hinton, and Williams. The key insight concerned the nature of the transfer function that governed how a single unit sets its activation state in response to inputs. Previous investigators had, for simplicity's sake, assumed either a linear transfer function or a step function; and the limited the scope of the perceptron-learning rule was a direct consequence of these choices. Rumelhart and colleagues investigated a logistic transfer function which (1) is nonlinear like a step function but (2) varies smoothly with net input so that it is differentiable (and, in fact, has a comparatively simple derivative; see Figure 2(b)). Because the transfer function is nonlinear, groups of units arranged in series are capable of computing the nonlinearly separable mappings that cannot be learned in a multilayered perceptron with linear units. Because

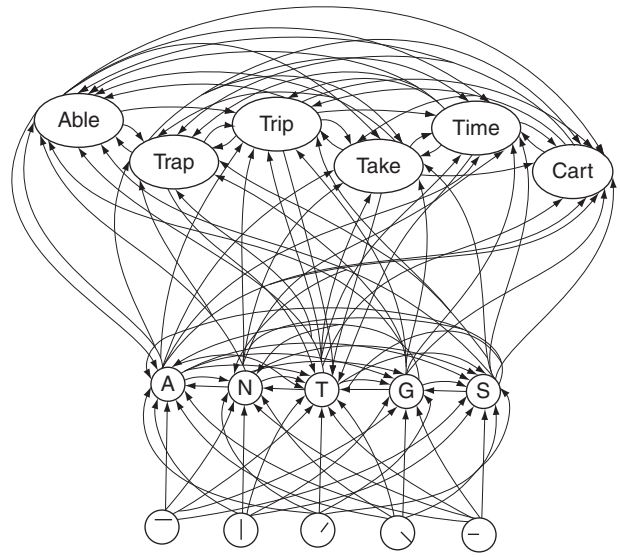

Figure 1 The interactive activation model of word recognition. The model consists of feature detectors organized hierarchically so that those toward the bottom represent simple visual features, those in the middle represent individual letters, and those at the top represent individual words. The visual appearance of an individual word activates basic features at the bottom, which pass their activation up through the hierarchy via weighted connections. Features excite the letters with which they are compatible and inhibit the letters with which they are incompatible. Individual letters inhibit one another, excite word units with which they are compatible, and inhibit word units with which they are incompatible. The activation of word-level units can feed back down to influence letter-level units, inhibiting incompatible letters and further exciting compatible units. Processing of a given input thus depends on the interactive activation of representations at all levels of the hierarchy. From McClelland JL and Rumelhart DE (1981) An interactive activation model of context effects in letter perception, Part 1: An account of basic findings. *Psychological Review* 88: 375–407.

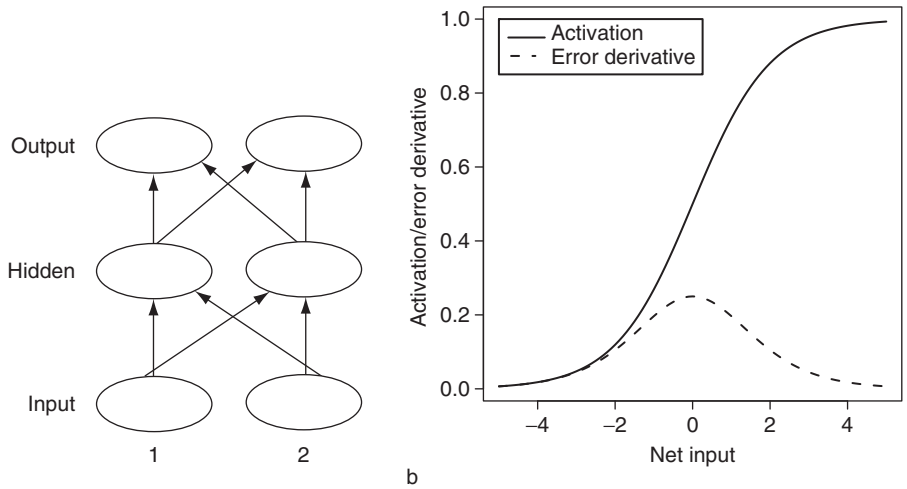

c

Patterns	Input 1	Input 2	Output 1	Output 2
1	0	0	1	0
2	0	1	0	1
3	1	0	0	1
4	1	1	1	0

d

From: \ To:		Input **1**	Input **2**	Hidden **1**	Hidden **2**	Output **1**	Output **2**
Input	**1**	0.0	0.0	–8.7	–7.6	0.0	0.0
Input	**2**	0.0	0.0	8.9	8.6	0.0	0.0
Hidden	**1**	0.0	0.0	0.0	0.0	13.0	–13.0
Hidden	**2**	0.0	0.0	0.0	0.0	–13.0	13.0
Output	**1**	0.0	0.0	0.0	0.0	0.0	0.0
Output	**2**	0.0	0.0	0.0	0.0	0.0	0.0
Bias		0.0	0.0	–4.57	3.8	6.44	–7

Figure 2 Exclusive- or (XOR) mapping: (a) architecture of a six-unit network for computing the exclusive- or (XOR) mapping; (b) logistic transfer function indicating how a unit's activation state changes with its net input; (c) a model environment containing patterns for the XOR learning problem; (d) one-weight matrix learned by the backpropagation learning rule. In (a), each oval indicates a neuronlike processing unit, and each arrow indicates a synapselike connection between units. In (b), a given receiving unit computes its net input by taking the dot product of the activations of sending units from which it receive projections times the values of the interconnecting weights. The dotted line indicates the derivative of the activation function, which can be used by the backpropagation learning rule to adjust the magnitudes of weights in the network so as to reduce error at the output layer. In (c), to process each pattern, the input units in (a) are set to the values indicated in the table (1 or 0), and this activation spreads forward in the network as determined by the values of the interconnecting weights and the transfer function. The activations of the two output units are then compared to the desired outputs shown in the table; the difference is converted to a measure of error, and the derivative of this error with respect to each weight is computed in a backward pass from outputs toward inputs. Each weight is adjusted by a small amount so as to reduce error at the output. In (d), the gray cells indicate logically possible connections between neurons that are not present in the network architecture, so their values are fixed to zero. The line labeled bias indicates each unit's intrinsic bias, which is a constant added to the unit's net input. When each of the input patterns in (c) is presented to a network with these weight values, outputs closely approximate the desired outputs also listed in (c).

the function is differentiable, it is possible to compute how each weight in such a network should be adjusted to improve performance. It is therefore possible to train a network of such units by (1) presenting an input and computing the forward propagation of activation from inputs toward outputs; (2) computing a measure of error, that is, a measure of the discrepancy between the observed and desired (target) output; and then (3) computing the derivative of the error with respect to each weight in the network in a backward pass from outputs toward inputs, and using these to adjust each weight in the network so as to reduce the error at the output. Networks trained with this backpropagation learning rule can, in principle, learn any input-output mapping, so long as there is an intervening layer of at least two units between inputs and outputs.

The third watershed event was the publication in 1986 of *Parallel Distributed Processing*, a two-volume work that laid out a comprehensive rationale for basing cognitive theories on principles derived from neural computation, reviewed the methodological state of the art, reported models of cognitive function across a remarkably broad variety of cognitive domains, and linked these models to core principles derived from theories of neural information processing. The two volumes highlighted some of the key differences between the then-dominant mind-as-computer approach to cognition and a neurally inspired approach. For instance, whereas computers

employ blindingly fast serial processors to achieve their computational power, individual neurons operate on an orders-of-magnitude slower time scale but are massively parallel. Despite the relatively slow computations achieved by a single neuron, neural systems achieve remarkably rapid and accurate performance in tasks that seriously challenge even the fastest serial computers (e.g., speech recognition). This suggests some fundamental difference in the way that parallel and serial systems process information; and if the mind is a product of a massively parallel processing system, it follows that the endeavor to understand cognition as the product of a serial symbol-processing system may not be optimal.

Basic Properties of Connectionist Networks

Connectionist models have seven general properties: (1) a set of units, (2) an activation state, (3) a weight matrix, (4) an input function, (5) a transfer function, (6) a learning rule, and (7) a model environment.

A Set of Units

Units are abstract neurons, simple processors that set their activation level by combining incoming signals and then transmit their activation state to other units through weighted connections. Some models construe individual units as representing particular theoretical constructs – for instance, explicit stimulus properties (such as colors, shapes, individual letters, or phonemes) or even individual words, objects, or concepts. Other models suggest that such theoretical constructs are best represented as patterns of activity across collections of units, so the activation of any individual unit has no directly interpretable significance. Figure 2 (a) shows the architecture of a feed-forward model for computing the XOR function.

An Activation State

At any given point in time, every unit in a network has a certain activation state. The state of the entire network, then, may be conceived as a one-dimensional vector, with each element corresponding to a single unit and containing that unit's current activation. When processing any given input pattern, the state of the XOR network is represented with a vector of six elements.

A Weight Matrix

Units can influence one another's activation via synapselike connections, or weights. Weights have direction; they project from a sending unit i to a receiving unit j. The weights in a network with n units can thus be conceived as an $n \times n$ matrix, with each element indicating the value of the weight projecting from one neuron to another, elements below the diagonal coding weights projecting in one direction (i.e., from i to j) and elements above the diagonal coding weights projecting in the reverse direction (i.e., from j to i). Most networks also include a vector of bias weights, that is, a vector then includes one constant for each unit in the network that reflects that unit's base activation in the absence of input. The architecture of the network (the arrangement of units into groups or layers) constrains which units project to which other units; where no connection exists between two units, the value in the weight matrix is fixed to zero. Figure 2(d) shows the matrix for a configuration of weights that solve the XOR mapping; gray cells in the matrix indicate values fixed to 0 because no connection exists between the corresponding units (A in Figure 1).

An Input Function

At any given time, a receiving unit j receives input from some set of sending units via the intermediating weight. The input function associated with a network determines how these inputs are combined to provide a net input to the receiving unit. The most frequent input function is simply the dot product of the vector of sending activations and the weight vector; that is, for each sending unit, one computes the product of the unit's activation and the value of the weight projecting to the receiving unit. These products are then summed across all sending units, and the result is the receiving unit's net input.

A Transfer Function

The transfer function determines how a unit's current net input influences its activation state. As previously mentioned, the logistic transfer function (a sigmoid bounded at 0 and 1) is very commonly used due to its desirable computational properties; but linear functions, step functions, radial basis functions, and other choices all appear in the literature. Models can also vary in how the transfer function integrates time. In many models, units update their activation instantaneously, but it is possible to gradually update the unit's state so as to simulate temporally extended network behaviors. Figure 2(b) shows the logistic transfer function and its derivative.

A Learning Rule

Although the weights in a connectionist model can sometimes be set by hand (i.e., specified by the

modeler), a key aspect of their appeal is their ability to discover useful weights as the consequence of applying some learning rule, that is, some rule for determining how weights should change as the consequence of processing an input pattern. There now exist a variety of learning rules, including many variants of the Hebb rule, in which weights are adjusted on the basis of correlated activity between pairs of units; various forms of error-correcting learning, such as the delta rule and backpropagation; and unsupervised learning rules that come to categorize, or cluster, their inputs without any explicit indication of what the desired output should be. There remains considerable controversy regarding the significance or utility of particular learning rules, the principal issue being an apparent trade-off between computational power and biological plausibility. Specifically, backpropagation remains the most powerful learning rule yet discovered, but it makes use of error signals that, at least at first blush, appear to have no analog in real brains. In contrast, variants of the Hebb rule closely reflect changes to synaptic potentiation in real brains; consequently, such rules are sometimes viewed as more biologically plausible, but they are also known to be fairly limited in the kinds of computations they can learn to perform.

A Model Environment

Because, in the connectionist view, all information processing involves the propagation of activation among units, it follows that, in any given task, both the stimulus and the behavior must be represented as patterns of activation across subsets of input and output units. The theorist must explicitly specify how such inputs and outputs are coded; in models of reading, for instance, the modeler must specify how the visual appearance of a word form and the motor act of articulating a word are to be coded as patterns of activity across units. The model environment consists of a collection of input–output pairs that constitute the mappings to be learned. Because a model's behavior is strongly influenced by the structure of its inputs and outputs, decisions about how these are best represented are often key to the theory that a given model is intended to exemplify. The training environment for the XOR problem is shown in Figure 2(c).

Appeal of Connectionist Models

Models with the seven properties just described provide an appealing basis for understanding aspects of cognition that seem especially challenging from traditional mind-as-computer approaches to cognition. For instance, such models exhibit:

1. Content-addressable memory. Connectionist models are good at pattern completion. Given some partial information about a previously learned pattern, the propagation of activation through a connectionist net can reinstantiate the full pattern; and moreover, any part of the full pattern can later be used as the cue for retrieval. This is also clearly true of human memory, but it is not naturally achieved in computer systems, in which a given data record can be retrieved only if one knows the correct address (i.e., the physical location on the storage medium where the record begins).
2. Generalization. In most connectionist networks, all patterns are processed through the same network of units and weights. Consequently, new inputs that share structure with previously learned items tend to generate similar outputs to the previously learned items. This ability to generalize past learning to new situations is critical to connectionist accounts of memory and language.
3. Graceful degradation. As any software engineer knows, computer programs are extremely brittle – tiny flaws in a program can completely destroy its ability to function. In contrast, connectionist networks exhibit graceful degradation. When weights or units are removed from the system, such networks can often generate outputs that are completely or at least partially correct. The human cognitive system also appears to exhibit graceful degradation, in that brain damage rarely if ever produces an all-or-nothing pattern of cognitive impairment. Furthermore, because the elements of a connectionist model can correspond, at least in principle, to particular brain regions, it is possible to investigate how brain damage in a particular locus should influence behavior under a given theory. Thus, connectionist models offer a valuable tool for the development of neuropsychological theories.
4. A mechanism for cognitive change. More traditional information-processing approaches are often used to characterize a child's mental processes at different stages of development, but under such approaches it is often unclear what mechanism subserves the transition from one stage to another. Connectionist models provide explicit hypotheses about mechanisms of change, in the form of the learning rule used to train the model, explicit assumptions about the nature of the developing child's experience, and explicit assumptions about how maturational change affects the architecture of the neural network

supporting the behavior. So, such models also offer a means of generating explicit hypotheses about mechanisms underlying cognitive development.

Symbolic versus Connectionist Approaches

Connectionist theories are sometimes viewed as providing a fundamentally different approach to cognition than the more traditional symbol-processing approach. Symbol-processing theories assume that the mind is a computational system consisting of two basic elements: (1) symbols, that is, structured representations that stand for things in the world, and (2) rules or operations, that is, algorithms for manipulating, combining, and responding to symbols. For instance, to generate the past-tense form of a verb such as 'look' might require, under a symbol-processing view, two symbols – one standing for the word 'look' and one standing for the English past-tense suffix '-ed' – and a rule that says, "'IF the goal is to generate the past tense from the present tense, THEN attach the past-tense suffix to the present-tense form." Applying the rule to the symbols then generates the correct past-tense form, 'looked.' Moreover, such a symbolic approach has a certain generative power; after learning a new present-tense verb (e.g., 'mave'), it is possible to infer the past-tense form ('maved') simply by applying the past-tense rule to the newly learned symbol.

Because symbol-processing systems are generative (i.e., they can produce an infinite variety of new representations and behaviors from a finite set of starting elements), many scientists believe that they must underlie much of human cognition, which is also generative. Connectionist models, in contrast, are thought by some to be capable of showing, at best, only a limited range of generativity. The current debate concerning the merits of connectionist approaches often hinge on these issues, which are best illustrated by an example.

Consider that in English, approximately 80% of verbs have a past tense formed by simply adding '-ed' to the present-tense form. The remaining 20% have past-tense forms that may be somewhat related to the present-tense form (e.g., 'give' → 'gave') or may even be completely unrelated to it (e.g., 'go' → 'went'). How do people come to learn which past-tense forms to produce from a given present-tense verb?

One commonsense hypothesis is that the present- and past-tense forms of familiar verbs are linked together in memory, by virtue of prior learning. Several empirical facts about language appear, however, to challenge this view. For instance, (1) people can reliably generate past-tense forms for novel verbs, and because these have never before been encountered, the behavior cannot depend on simply remembering the correct form; (2) children often generate incorrect forms that they are unlikely ever to have heard (e.g., producing 'goed' as the past tense of 'go'); and (3) a similar kind of behavior is observed following some kinds of brain damage, and it is absurd to imagine that brain damage causes someone to suddenly remember some past-tense form they have never before experienced or produced.

Such phenomena suggest that past-tense formation is subserved by something other than an association in memory between present- and past-tense forms. Symbol-processing theories propose that this something other is a rule like the one articulated earlier. In this view, the rule is used to generate past-tense forms for the majority of verbs, including novel verbs. Exceptions to the rule are stored in a separate lookup system or lexicon; and there may be additional rules for handling such exceptions (e.g., IF the present tense is listed in the lexicon, THEN look up the past-tense form in the lexicon; OTHERWISE use the rule to generate the past-tense form). Such a proposal nicely explains the empirical facts: because novel verbs such as 'mave' are not stored in the lexicon, the rule is used to generate a lawful past-tense form; children, on initially acquiring the rule, may mistakenly overapply it to exception words such as 'go' (thus accounting for their production of forms such as 'goed' that they have never encountered); patients who, after brain damage, generate forms such as 'goed' or 'drived' may have an intact rule system but a degraded exception system and so apply the rule to previously known exceptional forms.

It turns out, however, that the associative-learning account is not as limited, nor is the symbolic approach as powerful, as we might initially suspect. Connectionist networks trained to generate past-tense from present-tense forms (with real English words) actually exhibit all of the behaviors listed. The reason has to do with the remarkable ability of such models to generalize from previous experience. For instance, although such a network is never trained with a nonword such as 'mave', it does, during training, encounter real verbs with a similar structure: 'save', 'pave', 'rave', and so on. Most of these words form their past tense by adding '-ed' to their present-tense forms; and because this is so, new items such as 'mave' tend to produce a similar output. Such systematic tendencies in the language can also produce patterns of overregularization early in network training, as when children produce 'goed' instead of 'went', or when the knowledge in the network is perturbed by removing connection weights or units (as a model analog of brain damage).

Moreover, connectionist models so trained explain other observations that challenge the symbolic approach. For example, although exception words do not, like most verbs, simply take -'ed' in their past-tense form, there are some very systematic tendencies among the exceptions – consider 'weep' → 'wept', 'keep' → 'kept', 'sleep' → 'slept', and so on. These formations are not just consistent with one another; they are also quite similar to the past-tense form that would be generated if the rule were applied – the only difference is that, for these words, the long vowel is shortened in the past tense. Such observations are difficult to explain under the rules-and-symbols view. If exceptions are generated by a lookup table, why are past-tense forms so systematic across different subgroups of exceptions? And, if the rule and exception systems are functionally independent of one another, why are so many exceptional forms so similar – often differing in just one or two phonemes – to the regular forms that would be generated under the rule system?

Connectionist theories of past-tense formation explain such observations by suggesting that past-tense forms for both regular and exception words are generated through the same set of units and weights. The weights are shaped by systematic structure in the language, via the operation of a simple learning rule such as backpropagation. Such learning mechanisms excel at discovering weights that simultaneously capitalize on systematic input–output relationships, where they exist, and tolerate mild or even strong deviations from such regularities where necessary. Thus, although mappings such as 'keep' → 'kept' are in some sense exceptions, they still share considerable structure with the regular verbs and with one another; so these exceptional forms actually benefit substantially by being processed through the same system that handles the regular verbs. And, the connectionist account predicts empirical findings that are difficult to reconcile with the rules-and-symbols view. For example, when presented with a new nonsense verb such as 'neep', people are actually more likely to generate an irregular past-tense form (e.g., 'nept') than the regular form (i.e., 'neeped'), even though, according to the rule-based view, they should always apply the past-tense rule to all novel forms.

Past-tense formation thus exemplifies a domain in which (1) there exist very intuitive and empirically fleshed-out reasons for believing that behavior must be supported by rules and symbols but for which (2) simulations with connectionist models demonstrate that there exist alternative explanations for the same behaviors, within which there are no analogs to rules and symbols. Similar issues and controversies arise in the study of reading, language comprehension, word and object recognition, and categorization and semantics. In all these domains, connectionist theories usually explain behavior without reference to rules and symbols, and consequently such theories are sometimes viewed as challenging the view that human minds are symbol-processing systems. Other domains of human behavior – including, arguably, language production, reasoning, problem solving, and mathematics – seem to many to require a symbolic approach; consequently, challenges to symbolic explanations of even relatively simple phenomena (such as past-tense formation or single-word reading) may be viewed either as flawed in principle (because it is difficult to see how such accounts can be extended to higher-level cognition) or as undermining the foundational assumptions of cognitive science.

There are, however, other less dialectical ways of conceiving the relationship between symbolic and connectionist approaches. One view is that human cognition is subserved by some hybrid of connectionist and symbol-processing systems. For instance, recent symbolic accounts of past-tense formation suggest that the exception system is supported by a connectionist-like associative memory (thus explaining the systematicity among exceptions) but that there also exists a separate and independent rule system for handling regular items that bear little resemblance to previously learned forms.

Another view is that symbolic and connectionist approaches account for human behavior at different levels of description. Symbolic approaches may, for instance, provide valid descriptions of human behavior much as a flowchart captures the operation of a computer program. The flowchart is a useful description of the program, even though it yields no insight into the causal mechanisms by which the underlying hardware carries out the resulting computations. Connectionist theories, in contrast, may capture a level of description that explains how cognitive behaviors that, roughly speaking, look symbolic, can arise from mechanisms similar to those that govern neural information processing.

Biological Plausibility

Connectionist models are sometimes criticized for making implausible claims to biological faithfulness. For instance, it is not clear to what the units in many connectionist models are intended to correspond: single neurons, cortical columns, or whole populations of neurons. Similarly, single units can be used to represent whole words, objects, or other entities that, in the brain, are probably represented by widely distributed patterns of activity across thousands of neurons; unit activations are represented with single

real-valued numbers, whereas actual neurons transmit information via discrete all-or-nothing action potentials; weights in connectionist networks often can take any real positive or negative value, whereas in the brain, some synapses are exclusively inhibitory and others are exclusively excitatory; in connectionist models, different units are often treated as functionally and structurally equivalent to one another, whereas in the brain there exist many different kinds of neurons, which may in turn participate in various different cytoarchitectonic structures; and so on. For such critics, claims that connectionist models capture important aspects of neural information processing may seem far-fetched.

One response to this criticism is to point out that the utility of any model system lies in abstracting away from the many complex details of the natural system, so as to investigate those factors that, under a given theory, are necessary to explain the relevant phenomena. Thus, it is certainly possible to construct neurally faithful computer simulations in which there exist multiple different kinds of neurons, connected via different kinds of synapses, organized into cortical columns, and projecting to one another via neuroanatomically faithful tracts. Should the theorist demonstrate that such a simulation replicates the behavior of interest, however, it would be comparatively difficult to understand why this should be so because the model is nearly as complex as the system itself. Connectionist models may be viewed as making simplifying assumptions about the details of the underlying neural systems so as to allow the theorist to focus on a relatively circumscribed set of factors of explicit theoretical interest.

It is also worth noting that the investigation of the properties of abstract computational systems can lead to important neuroscientific discoveries, even when the computational systems are not intended as faithful models of neural systems. For instance, in 1988 Sutton described an algorithm for reward-based learning of sequential structure (temporal-difference learning) in work that was purely computational (i.e., not guided by neuroscience). Nevertheless, it was subsequently discovered that dopamine signals in cortex, striatum, and midbrain closely track the reward signal generated by the algorithm, and it now appears as though some forms of learning may be supported by neural mechanisms that effectively implement temporal-difference learning. So, insights gained through computer simulations, even when they may seem divorced from what we currently know about neural processing mechanisms, can promote new insights into how neural systems function.

See also: Cognition: An Overview of Neuroimaging Techniques; Cognitive Neuroscience: An Overview; Connectionist Models of Language Processing; Executive Function and Higher-Order Cognition: Computational Models; History of Neuroscience: Early Neuroscience; Recognition Memory; Word Recognition; Word Learning.

Further Reading

Churchland PS and Sejnowski TJ (1990) Neural representation and neural computation. *Special Issue: Action Theory and Philosophy of Mind. Philosophical Perspectives* 4: 343–382.

Cohen JD, Braver TS, and O'Reilly R (1996) A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. *Philosophical Transactions: Biological Sciences* 351(1346): 1515–1527.

Elman JL (1990) Finding structure in time. *Cognitive Science* 14: 179–211.

Elman JL, Bates EA, Karmiloff-Smith A, Johnson M, Parisi D, and Plunkett K (1996) *Rethinking Innateness: A Connectionist Perspective on Development.* Cambridge, MA: MIT Press.

Fodor JA and Pylyshyn ZW (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28: 3–71.

Hebb DO (1949) *The Organization of Behavior.* New York: Wiley.

Joanisse MF and Seidenberg MS (1999) Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7592–7597.

Marcus GF (2001) *The Algebraic Mind.* Cambridge, MA: MIT Press.

McClelland JL and Patterson K (2002) "Words or rules" cannot exploit the regularity in exceptions. *Trends in Cognitive Sciences* 6: 464–465.

McClelland JL and Rumelhart DE (1981) An interactive activation model of context effects in letter perception, Part 1: An account of basic findings. *Psychological Review* 88: 375–407.

Pinker S and Ullman MT (2002) The past and future of the past tense. *Trends in Cognitive Sciences* 6: 456–463.

Plaut DC, McClelland JL, Seidenberg MS, and Patterson K (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103: 56–115.

Rogers TT and McClelland JL (2004) *Semantic Cognition: A Parallel Distributed Processing Approach.* Cambridge, MA: MIT Press.

Rumelhart DE, McClelland JL, and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations.* Cambridge, MA: MIT Press.

Rumelhart DE, McClelland JL, and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models.* Cambridge, MA: MIT Press.

Connectionist Models of Language Processing

L Onnis, University of Hawai'i, Honolulu, HI, USA
H Christiansen, Cornell University, Ithaca, NY, USA
N Chater, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Connectionist psycholinguistics is an emerging approach to modeling empirical data on human language processing using connectionist computational architectures. Over the last 20 years, a wide range of psycholinguistic phenomena have been modeled, such as speech processing, impaired and normal reading, aphasic word production, and structural priming in sentence production. This article focuses on syntax, and especially on the theoretical relationship between traditional notions in syntax (constituency, structure dependency, and recursion) and connectionist networks, which do not appear to make reference to these notions intrinsically.

Connectionist models, or neural networks, are neurally inspired devices based on numerical computation rather than symbolic manipulation. They are formed by several simple processing units, called nodes, which mimic the activity of single neurons. Each node is excited or inhibited by information coming from other units to which it is connected. A set of homogeneous nodes that collectively represents some information is often called a layer. For instance, there could be a phoneme layer, a word layer, and so on. A popular architecture in connectionist psycholinguistics is the feed-forward network (Figure 1), in which an input layer receives signals (e.g., raw words) from outside the network, an output layer sends signals to outside the network, and an intermediate layer of so-called hidden units does not directly connect outside of the network. Hidden unit layers are often considered the internal states that represent the knowledge acquired by the network at the end of training. Information flows from input to output, and training proceeds by adjusting the weights between single units until the network produces a desired output (for instance, the production of the correct word syntactic category in a given context).

Two aspects of neural networks are particularly interesting for syntactic processing. First, networks learn to perform some specific task, such as process an English sentence, from being exposed to relatively unanalyzed raw input, such as a sequence of words. This feature makes them suitable for explaining a great deal of language learning and language processing in a psychologically plausible manner. Second, the knowledge of the networks is not represented in any explicit way but rather emerges as a pattern of neuronal activity between the interconnected nodes in a parallel distributed processing fashion. This is particularly interesting because it contrasts with more-traditional symbolic views in cognitive science in suggesting that humans process sentences by transforming representations according to sets of rules. For instance, a representation of the sentence *the girl liked a boy* requires the constituents *the*, *a*, *girl*, *boy*, *liked*, and an explicit representation of the syntactic relationships between these constituents. The rules of syntax govern how constituents can be combined together, allowing, for instance, *the boy liked a girl* as a legal sentence but not *boy girl a liked the*. The following phrase structure rules describe the relationship between constituents in the above sentence (S stands for sentence; NP, noun phrase; VP, verb phrase; det, determiner; N, noun; and V, verb):

S → NP VP
NP → (det) N
VP → V (NP)

Besides such constituent structure, to capture the full generativity of human language, recursion needs to be introduced, for instance by adding a new rule that adds a potential prepositional phrase (PP) to the NP (prep stands for preposition):

NP → (det) N (PP)
PP → prep NP

These rules are recursive because the expansion of the right-hand sides of each can call the other. For example, the complex NP *the flowers in the vase* has the simple NP *the vase* recursively embedded within it. Because this process can be applied arbitrarily often, constructions of arbitrary complexity can in principle be generated. Constituency and recursion are some of the most fundamental concepts in linguistics. Because both are defined in terms of relations between symbols, symbolic models of language processing therefore incorporate these properties by fiat. This article discusses how constituency, structure dependency, and recursion may fit into a connectionist framework and the possible implications this work may have for linguistics and psycholinguistics.

Constituency

Connectionist models can address constituency in three increasingly radical ways. First, some models – especially early ones – are implementations of

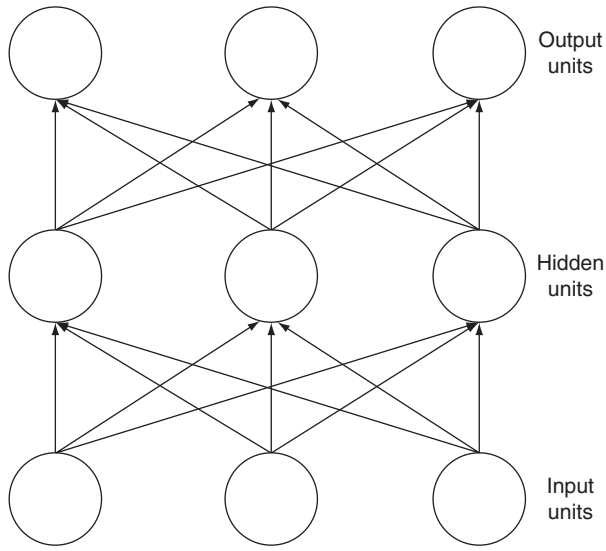

Figure 1 A standard feed-forward network containing three layers of units. Information flows bottom up from input to output units. These types of networks are normally trained using the back-propagation algorithm, which minimizes the discrepancy between the network's actual and desired output. Redrawn from Onnis L, Christiansen H, and Chater N (2006) Human language processing: Connectionist models. In: Brown K (ed.) *Encyclopedia of Language and Linguistics,* pp. 401–409. Oxford: Elsevier, with permission from Elsevier.

symbolic language-processing models in neural hardware. For example, some models have implemented context-free grammars containing explicit representations of the constituent structure of a sentence in just the same way as a nonconnectionist implementation of the same model would have. Connectionist implementations of this kind provide feasibility proofs that traditional symbolic models of language processing are compatible with a brain style computational architecture, although they add little new to the treatment of constituency. The remaining two classes of connectionist models actually learn to process constituent structure rather than having this ability hardwired. One approach is to have a network learn from input tagged with information about constituent structure. One study trained a network to map a combination of orthographic and co-occurrence-based semantic information about a word onto a structured representation encoding the minimal syntactic environment for that word. With an input vocabulary consisting of 20 000 words, this model has an impressive coverage and can account for certain results from the psycholinguistic literature concerning ambiguity resolution in sentence processing. Another model, known as a simple synchrony network, was trained to parse sentences pre-encoded as parts of speech. The network takes the part of speech tags for the sentence constituents as input and is trained to output the parse tree fragment of a given constituent when that constituent is queried. The network learns to parse a corpus of written English reasonably well. However, because in these models constituent structure is compiled either into the input or the output representations, this style of model does not offer any fresh insight into how linguistic constituency might operate based on connectionist principles.

The third class of connectionist models addresses the more ambitious problem of learning the constituent structure of a language from untagged linguistic input. Such models have the potential to develop a new or unexpected notion of constituency and hence may have substantial implications for theories of constituency in linguistics and psycholinguistics.

To understand how the more radical connectionist models address constituency, one can divide the problem of finding constituent structure into two interrelated parts: segmenting the sentence into chunks which correspond, to some extent, to linguistic constituents and categorizing these units appropriately. The first problem is an aspect of the general problem of segmenting speech into appropriate units (phonemes, words, etc.) and more generally is an aspect of perceptual grouping. The second problem regards the general problem of classifying linguistic units – for instance, recognizing different classes of phonemes or establishing the parts of speech of individual lexical items. The segmentation and classification problems need not be solved sequentially. Indeed, there may be mutual influence between the decision to segment a particular chunk of language and the decision that it can be classified in a particular way. Nonetheless, it is useful to keep the two aspects of the analysis of constituency conceptually separate.

It is also important to stress the difference between the problem of assigning constituent structure to novel sentences when the language is known and the problem of acquiring the constituent structure of an unknown language. Statistical symbolic parsers are able to make some inroads into the first problem. For highly stylized language input, and given a prestored grammar, they can apply grammatical knowledge to establish one or more possible constituent structures for novel sentences. But symbolic methods are much less advanced in acquiring the constituent structure of language because this requires solving the hard problem of learning a grammar from a set of sentences generated by that grammar. It is therefore in relation to the acquisition of constituency that connectionist methods, with their well-developed learning methods, have attracted the most interest.

One may begin by focusing on the problem of classifying, rather than segmenting, the linguistic input. One connectionist model learns parts of speech

of individual words by clustering words together on the basis of the immediate linguistic contexts in which they occur. The rationale is based on the replacement test: if two words are observed to occur in highly similar immediate contexts in a corpus, they probably belong to the same syntactic category. Using a single layer network with Hebbian learning to store co-occurrences between target words and their immediate neighbors, each target word was associated with a vector representing the contexts in which it typically occurs. A competitive learning network classified these vectors grouping together words with similar syntactic categories. This method is able to operate over unrestricted natural language, in contrast to most symbolic and connectionist models. From a linguistic perspective, the model slices lexical categories too finely, producing, for example, many word classes that correspond to nouns or verbs. On the other hand, the words within a class tend to be semantically related, which is useful from a cognitive perspective. The same method can be extended to classify sequences of words as NPs, VPs, and so forth. An initial classification of words is used to recode the input as a sequence of lexical constituents. Then short sequences of lexical constituents are classified by their context, as before. The resulting groups of phrase (e.g., determiner–adjective–noun) are readily interpretable as NPs, VPs, PPs, and so on, but again, these groupings are too linguistically restrictive (i.e., only a small number of NPs are included in any particular cluster). Moreover, this phrasal-level classification has not yet been implemented in a connectionist network.

A different attack on the problem of constituency involves training simple recurrent networks (SRNs) on linguistic input. An SRN involves a crucial modification to a feed-forward network: the current set of hidden unit values is copied back to a set of additional input units and paired with the next input to the network (Figure 2). The current hidden unit values can thus directly affect the next hidden unit values, providing the network with a memory for past inputs. This enables it to tackle online sentence processing, in which the input is revealed sequentially over time.

Segmentation into constituents can be achieved in two ways by an SRN trained to predict the next input. One is based on the assumption that predictability is higher within constituents than across constituent boundaries, and hence that high prediction error indicates a boundary. This method has been advocated as potentially applicable at a range of linguistic levels, but in practice it has been successfully applied only on corpora of unrestricted natural language input in finding word boundaries. Even here, the prediction strategy is a very partial cue to segmentation. If the network is provided with information about

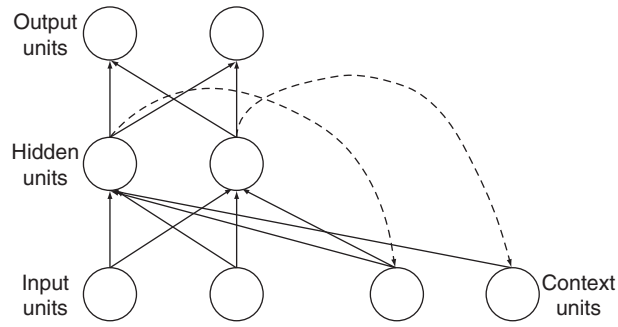

Figure 2 A simple recurrent network (SRN). An SRN is essentially a standard feed-forward network equipped with an extra layer of so-called context units. At each time step, an input propagates through the hidden units to the outputs (solid arrows). The hidden unit activation at the previous time step is copied back to the context layer (dashed arrows) and paired with the current input (solid arrows). Thus the hidden units influence the processing of subsequent inputs, providing a limited ability to deal with sequential inputs. Redrawn from Onnis L, Christiansen H, and Chater N (2006) Human language processing: Connectionist models. In: Brown K (ed.) *Encyclopedia of Language and Linguistics,* pp. 401–409. Oxford: Elsevier, with permission from Elsevier.

naturally occurring pauses between utterances (or parts of utterances), an alternative method is to assume that constituent boundaries occur where the network has an unusually high expectation of an utterance boundary. The rationale is that pauses tend to occur at constituent boundaries and hence the prediction of a possible utterance boundary suggests a constituent boundary may have occurred. This approach seems highly applicable to segmenting sentences into phrases but has also primarily been used for finding word boundaries in real corpora of language, when combined with other cues.

So far, this article has considered how SRNs might find constituents. But how well do they classify constituents? At the word level, cluster analysis of hidden unit activations shows that, to some extent, the hidden unit patterns associated with different word classes group naturally into syntactic categories, for SRNs trained on simple artificial grammars. These results are important because they show that even though the SRN may not learn to classify constituents explicitly, it is nevertheless able to use this information to process constituents appropriately. Another way of assessing constituency learning is to see whether networks generalize to predict novel sentences of a language. The logic is that to predict successfully, the SRN must exploit linguistic regularities which are defined across constituents and hence develop a notion of constituency to do so. However,

it has been pointed out that this type of evidence is not compelling if the novel sentences are extremely similar to the network's training sentences. That is, to show substantial evidence for generalization across constituents, the network should be able to handle novel sentences in which words appear in sentence locations where they have not previously occurred. For example, a novel sentence might involve a particular noun in object position that has previously occurred only in subject position. This property of the human brain, called systematicity, captures the idea that if one can understand the sentence *Bob loves Mary*, one can also understand *Mary loves Bob*; the underlying knowledge that supports understanding of the first sentence enables understanding the second. Hence, to generalize effectively, the network must presumably develop some abstract category of nouns.

The challenge of systematicity can be addressed by presenting evidence that connectionist models are in fact able to attain strong generalization. In the training corpus of one study, the noun *boy* was prevented from ever occurring in a noun phrase conjunction (i.e., noun phrases such as *John and boy* and *boy and John* did not occur). During training, the SRN had therefore been presented only with singular verbs following *boy*. Nonetheless, the network was able to correctly predict that a plural verb must follow *John and boy* as prescribed by the grammar. In addition, the network was still able to correctly predict a plural verb when a prepositional phrase was attached to *boy* as in *John and boy from town*, providing even stronger evidence for strong generalization. In contrast, when the SRN was presented with ungrammatical lexical items in the second noun position, as in *John and near*, it did not activate the plural nouns. Instead, it activated lexical items that were not grammatical given the previous context. This suggests that the SRN is able to make nonlocal generalizations based on the structural regularities in the training corpus. If the SRN relied solely on local information, it would not have been able to make correct predictions in either case. Thus, the network demonstrated sophisticated generalization abilities, ignoring local word co-occurrence constraints while appearing to comply with structural information at the constituent level.

Another strong test of generalizations that connectionist models seem to fail concerns equivalence relations: one SRN was trained on sentences like *A rose is a rose* or *A tulip is a tulip*, but when given a novel sentence fragment like *A blicket is a*..., the SRN could not predict that *blicket* was going to be the next word because it activated all the words that it had seen in this sentence position (e.g., *rose*, *tulip*, etc.). This suggests that it did not develop abstract variable-based frames like *a X is a X*, where *X* is a variable that can be bound to any word. In reply to this negative evidence, another study presented a model of sentence production that could generalize well if the input message was encoded in several separate representations that were linked together. The model had a dual-pathway architecture, with one pathway representing the mapping of object semantics to word forms and another representing and mapping objects (and the words that describe them) into appropriate sentence positions. The model contained a bank of units representing event roles (agent, patient, etc.) that fed onto a layer of lexical units (e.g., DOG, CAT, etc.). This way the model could represent the different roles of a word, say *dog*, in various events while activating a common semantics that all dogs have. Semantic units connected to word label units that represented the phonetics of each word, allowing the learning of a word label for each meaning. Because there was only one set of semantic units for different event role units, learning the mapping of the semantic feature DOG to the word label *dog* allowed the model to generalize this word to other event roles. Hence, even if the model had experienced only the sentence *the dog chases the cat*, it was able to generalize and produce a novel sentence *the cat chases the dog*. A variation of this model was further able to capture an important psycholinguistic phenomenon called structural priming, namely, a tendency for speakers to reuse previously experienced sentence structures. Because structural priming occurs in the absence of lexical and conceptual repetition and in the face of thematic role differences, it has been taken as convincing evidence that people make use of abstract syntactic representations and constructions, such as a the double-object dative construction (e.g., *John gave Mary a book*). The connectionist model could reproduce structural priming without having inbuilt syntactic constructions.

Another model, called CSCP, was developed to perform both comprehension and production of complex, multiclausal sentences using a larger corpus than previous models had. In this respect, neural networks promise to scale up to real samples of natural language. The model deals with several sentence comprehension and production phenomena that involve discovery of constituent structures such as multiple verb tenses and voices, adverbs and adjectives, prepositional phrases, relative and subordinate clauses, and sentential complements. The CSCP network has a semantics component, which is used to help the net predict the next word. The semantic component is also used to produce a word. When a word is selected for production, it is fed back into the comprehension input, and the model proceeds to produce the next word, and so on. Therefore, the model also claims that language production is learned primarily by

formulating implicit predictions while attempting to comprehend one's language.

This discussion must note that connectionist models do not mirror classical constituency precisely. That is, they do not derive rigid classes of words and phrases that are interchangeable across contexts. Rather, they divide words and phrases into clusters without precisely defined boundaries and treat them differently depending on the linguistic contexts in which they occur. This context-sensitive constituency can be viewed as either the undoing of connectionist approaches to language or their radical contribution. The potential problem with context-sensitive constituency is the productivity of language: to take Noam Chomsky's famous example, how do we know that the sentence *colorless green ideas sleep furiously* is syntactically correct, except by reference to a context-insensitive representation of the relevant word classes? This seems necessary because each word occurs in a context where it has rarely been encountered before. But this problem may not be fatal for context-sensitive notions of constituency. A neural network was trained to mutually associate two input sequences – a sequence of word forms and a corresponding sequence of word meanings. The network was able to learn a small artificial language successfully; it was able to regenerate the word forms from the meanings and vice versa. The network was tested to see whether, by passing information from form to meaning and back, it could recreate a sequence of word forms presented to it. Ungrammatical sentences were recreated much less accurately than grammatical sentences, and the network was thus able to distinguish grammatical from ungrammatical sentences. It is important to note that this was true for sentences in which words appeared in novel grammatical combinations, as specified by the systematicity criterion and as exampled in Chomsky's famous sentence. Thus, the context sensitivity of connectionist constituency may not rule out the possibility of highly creative and novel use of language, because abstract relations may be encoded at a semantic level, as well as at the level of word forms.

If the apparent linguistic limitations of context-sensitive constituency can be overcome, then the potential psychological contribution of this notion is enormous. First, context-sensitivity seems to be the norm throughout human classification. Second, much data on sentence processing can be naturally explained by assuming that constituents are represented in a fuzzy and context-bound manner.

Structure Dependency

A linguistic concept intimately tied to constituency is structure dependency, the notion that grammatical knowledge relies on the structural relationships between constituents rather than on the linear sequence of items. One of the fiercest arguments leveled at distributional models of learning (of which neural networks can be seen as an instantiation) concerns the uninformativeness of such mechanisms for detecting linguistically relevant properties. The most relevant properties of language appear to be abstract, such as phrase structure configurations, grammatical relations, and syntactic categories, whereas information contained in raw input pertains to serial position, adjacency, and co-occurrence relations among words. To the extent that connectionist models, and in particular SRNs, process information relying simply on the linear sequencing of linguistic elements (at any level of analysis – phonemes, morphemes, words, etc.), they would seem doomed to fail to capture structure dependency. It turns out that SRNs can master a classic example of structure dependency in English, auxiliary fronting. In the generative linguistics framework, declaratives are turned into questions by fronting the correct auxiliary. For example, to turn the declarative form *The man who is hungry is ordering dinner*, it is correct to front the main clause auxiliary, as in 1(a), but fronting the subordinate clause auxiliary produces an ungrammatical sentence, as in 1(b):

1(a) *Is the man who is hungry ordering dinner?*
1(b)* *Is the man who hungry is ordering dinner?*

A structure-independent rule would move the first *is*, but the correct structure-dependent rule involves only the movement of the *is* from the main clause. Because children do not erroneously move the first *is* to the front of the sentence, and because they receive too little direct evidence for inferring correct auxiliary fronting, it has been claimed that structure dependency could not be learned from positive evidence by an associative mechanism. To address this challenge, SRNs were trained to predict the next lexical category from a corpus of child-directed speech that contained no relevant examples of correct auxiliary fronting. The networks, tested on completely novel sentences, produced more accurate predictions for grammatical test sentences such as *Is the boy who is hungry nearby?* than for ungrammatical sentences, such as **Is the boy who hungry is nearby?* The prediction of the well-formed relative clause continuation, V (i.e., *is*), was highly preferred over the ill-formed version, ADJ (i.e., *hungry*). This pattern of predictions reflected the networks' sensitivity to the statistical properties of the corpus. The networks distinguished chunks of lexical categories that were more frequent in the training input from less-frequent ones (i.e., the lexical categories corresponding to PRON V ADJ (*who is hungry*) vs. PRON ADJ V (*who hungry is*)). This work suggests that the networks were sensitive to the indirect statistical evidence present in the

corpus, developing an appropriate bias toward the correct forms of AUX questions without having a built-in rule for question formation.

Recursion

As with constituency, connectionist models have dealt with recursion in three increasingly radical ways. The least radical approach is to hardwire recursion into the network (e.g., as in the model that implemented phrase structure rules) or to add an external symbolic (first-in-last-out) stack to the model. In both cases, recursive generativity is achieved entirely through standard symbolic means, and although this is a perfectly reasonable approach to recursion, it adds nothing new to symbolic accounts of natural language recursion. The more radical connectionist approaches to recursion aim for networks to learn to deal with recursive structure. One approach is to construct a modular system of networks, each of which is trained to acquire different aspects of syntactic processing: one trained to map words onto case-role assignments, another trained to function as a stack, and a third trained to segment the input into constituent-like units. Although such a model displays complex recursive abilities, the basis for these abilities and their generalization to novel sentence structures derive from the configuration of the stack network, combined with the modular architecture of the system, rather than being discovered by the model.

The most radical connectionist approaches to recursion attempt to learn recursive abilities with minimal prior knowledge built into the system. In this type of model, the network is most often required to discover both the constituent structure of the input and how these constituents can be recursively assembled into sentences. As with the similar approach to constituency described above, such models may provide new insights into the notion of recursion in human language processing.

Before these modeling efforts are discussed, the extent to which recursion is observed in human language behavior needs to be assessed. It is useful to distinguish simple and complex recursion. Simple recursion consists of recursively adding new material to the left (e.g., the adjective phrases (AP) in *the gray cat* → *the fat gray cat* → *the ugly fat gray cat*) or right (the PPs in *the flowers in the vase* → *the flowers in the vase on the table* → *the flowers in the vase on the table by the window*) of existing phrase material. In complex recursion, new material is added in more complicated ways, for example, through center embedding of sentences (*the chef admired the musicians* → *the chef who the waiter appreciated admired the musicians*). Psycholinguistic evidence shows that people find simple recursion relatively easy to process, whereas complex recursion is almost impossible to process with more than one level of recursion. For instance, the following sentence, with two levels of simple (right-branching) recursion, *The busboy offended the waiter who appreciated the chef who admired the musicians*, is much easier to comprehend than the comparable sentence with two levels of complex recursion, *The chef who the waiter who the busboy offended appreciated admired the musicians*. Because recursion is built into symbolic models, there are no intrinsic limitations on how many levels of recursion can be processed. Instead, such models must invoke extrinsic constraints, such as the competence–performance distinction, to accommodate the human performance asymmetry on simple and complex constructions. The radical connectionist approach models human performance directly without the need for extrinsic performance constraints.

The SRN model developed by Jeff Elman was perhaps the first connectionist attempt to simulate human behavior on recursive constructions. This network was trained on sentences generated by a small context-free grammar incorporating center embedding and a single kind of right-branching recursive structure. In related work, SRNs were trained on a recursive artificial language incorporating four kinds of right-branching structures, a left-branching structure, and center embedding. The behavior of these networks was qualitatively comparable to human performance in that the SRN predictions for right-branching structures were more accurate than for sentences of the same length involving center embedding, and performance degraded appropriately when depth of center embedding increased. Further simulations corroborated these results, suggesting that semantic bias (incorporated via co-occurrence restrictions on the verbs) can facilitate network performance in center-embedded constructions in a way similar to the semantic facilitation effects found in human processing. Using abstract artificial languages, it has been demonstrated that the SRN's general pattern of performance is relatively invariant across network size and training corpus, and the humanlike pattern of performance likely derives from intrinsic constraints inherent in the SRN architecture.

Connectionist models of recursive syntax typically use toy fragments of grammar and small vocabularies. Aside from raising concerns over scaling up, these factors make it difficult to provide detailed fits with empirical data. Nonetheless, some attempts have recently been made toward fitting existing data and deriving new empirical predictions from the models. One SRN model fits grammaticality ratings data from several behavioral experiments, including an account of the

relative processing difficulty associated with the processing of center embeddings (with the following relationship between nouns and verbs: $N_1_N_2_N_3_V_3_V_2_V_1$) versus cross-dependencies (with the following relationship between nouns and verbs: $N_1_N_2_N_3_V_1_V_2_V_3$). Human data have shown that sentences with two center embeddings (in German) were significantly harder to process than comparable sentences with two cross-dependencies (in Dutch). A measure of grammatical prediction error (GPE) was developed that allowed network output to be mapped onto human performance data. GPE is computed for each word in a sentence and reflects the processing difficulties that a network experiences at a given point in a sentence. Averaging GPE across a whole sentence, the SRN fit human data concerning the greater perceived difficulty associated with center embedding in German compared with cross-serial dependencies in Dutch. The simulation results demonstrated that the SRNs exhibited the same kind of qualitative processing difficulties as humans do on these two types of complex recursive constructions. Further simulations derived novel predictions concerning other types of recursive constructions, and these predictions were later confirmed experimentally. Single-word GPE scores were mapped directly onto reading times, providing an experience-based account for human data concerning the differential processing of singly center-embedded subject and object relative clauses in human participants with different levels of reading comprehension ability.

Just as the radical connectionist approach to constituency deviates from classical constituency, the above approach to recursion deviates from the classical notion of recursion. The radical models of recursion do not acquire true recursion because they are unable to process infinitely complex recursive constructions. However, the classic notion of recursion may be ill-suited for capturing human recursive abilities. Indeed, the psycholinguistic data suggest that people's performance may be better construed as being only quasi-recursive. The earlier mentioned semantic facilitation of recursive processing further suggests that human recursive performance may be partially context-sensitive; for example, the semantically biased *The bees that the hive that the farmer built housed stung the children* is easier to comprehend than neutral *The chef that the waiter that the busboy offended appreciated admired the musicians* even though both sentences contain two center embeddings. This dovetails with the context-sensitive notion of constituency and suggests that context sensitivity may be a more pervasive feature of language processing than is typically assumed by symbolic approaches.

Connectionist Models of Language Processing

What is the significance of connectionist models of language processing? Will connectionism ultimately replace, complement, or simply implement symbolic approaches to language? Early connectionists addressed this issue by attempting to show that connectionism could, in principle, capture aspects of language and language processing. These models showed that connectionist networks could in principle acquire parts of linguistic structure without extensive innate knowledge. Recent work has moved toward a connectionist psycholinguistics that captures detailed psychological data. This article has outlined several ways in which constituency and recursion – two fundamental properties of linguistic knowledge – may be accommodated within a connectionist framework, ranging from direct implementation of symbolic systems to the acquisition of constituency and recursion from untagged input. It has focused on the radical approach because this approach has the greatest potential impact on psycholinguistics and linguistic theory.

However, much of this research is still preliminary. Work is required to decide whether promising, but limited, initial results can eventually be scaled up to deal with the complexities of real language input or whether a radical connectionist approach is beset by fundamental limitations. Another challenge is to find ways – theoretically and practically – to interface models, which have been proposed at different levels of linguistic analyses, with one another (e.g., interfacing models of morphology with models of sentence processing).

Nevertheless, the connectionist models described in this article have already influenced the study of language processing. First, connectionism has helped promote a general change toward replacing box-and-arrow diagrams with explicit computational models. Second, connectionism has reinvigorated the interest in computational models of learning, including learning properties, such as recursion and constituent structure, which were previously assumed to be present *a priori* in humans and therefore taken for granted. Finally, connectionism tends to discard the separation between competence and performance as artificially construed and possibly misleading. From this perspective, linguistic recursion is a conceptual artifact of the competence–performance distinction instead of a necessary characteristic of the underlying computational mechanism. In this light, the problem facing connectionist models of language processing is not whether they can implement some kind of recursive mechanism but whether they will be able to account for the (limited) recursive structure found in

natural language behavior purely in terms of nonsymbolic computation.

Connectionism has already had a considerable impact on the psychology of language. However, the final extent of this influence depends on the degree to which practical connectionist models can be developed and extended to deal with complex aspects of language processing in a psychologically realistic way. Recent models have already started to scale up to more realistic corpora and sentence processing behaviors. If realistic connectionist models of language processing can be provided, then the possibility of a radical rethinking, not just of the nature of language processing, but of the structure of language itself, may be required.

See also: Language Development; Psycholinguistics; Sentence Production; Sentence Comprehension; Statistical Learning of Language; Word Production; Word Recognition; Word Learning.

Further Reading

Allen J and Seidenberg MS (1999) The emergence of grammaticality in connectionist networks. In: MacWhinney B (ed.) *The Emergence of Language*, pp. 115–151. Mahwah, NJ: Lawrence Erlbaum.

Cairns P, Shillcock RC, Chater N, and Levy J (1997) Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology* 33: 111–153.

Chang F, Dell GS, and Bock K (2006) Becoming syntactic. *Psychological Review* 113(2): 234–272.

Christiansen MH, Allen J, and Seidenberg MS (1998) Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes* 13: 221–268.

Christiansen MH and Chater N (1994) Generalization and connectionist language learning. *Mind and Language* 9: 273–287.

Christiansen MH and Chater N (1999) Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23: 157–205.

Christiansen MH and Chater N (2001) Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences* 5: 82–88.

Christiansen MH and Chater N (eds.) (2001) *Connectionist Psycholinguistics*. Westport, CT: Ablex.

Elman JL (1990) Finding structure in time. *Cognitive Science* 14: 179–211.

Finch S and Chater N (1993) Learning syntactic categories: A statistical approach. In: Oaksford M and Brown GDA (eds.) *Neurodynamics and Psychology*, pp. 295–321. New York: Academic Press.

Fodor J and Pylyshyn Z (1988) Connectionism and cognitive architecture. *Cognition* 28: 3–71.

Hadley RF (1994) Systematicity in connectionist language learning. *Mind and Language* 9: 247–272.

Henderson J and Lane P (1998) A connectionist architecture for learning to parse. In: *Proceedings of 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, pp. 531–537. Canada: University of Montreal.

Kwasny SC and Faisal KA (1990) Connectionism and determinism in a syntactic parser. *Connection Science* 2: 63–82.

MacDonald MC and Christiansen MH (2002) Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review* 109: 35–54.

Marcus GF (1998) Rethinking eliminative connectionism. *Cognitive Psychology* 37: 243–282.

Miikkulainen R (1996) Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science* 20: 47–73.

Onnis L, Christiasen M, and Chater N (2006) Human langauge processing: Connectionist models. In Brown K (ed.) *Encyclopedia of Language and Linguistics*, pp. 401–409. Oxford: Elsevier.

Reali F and Christiansen MH (2005) Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science* 29(6): 1007–1028.

Rohde DLT and Plaut DC (2003) Connectionist models of language processing. *Cognitive studies* 10(1): 10–28.

Weckerly J and Elman JL (1992) A PDP approach to processing center-embedded sentences. In: *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, pp. 414–419. Hillsdale, NJ: Lawrence Erlbaum.

Connectivity of Primate Reward Centers

S N Haber, University of Rochester School of Medicine, Rochester, NY, USA

© 2009 Published by Elsevier Ltd.

Introduction

At the center of the neural network that mediates reward-based learning is the cortical–basal ganglia circuit. This network is critical for the development of appropriate goal-directed behaviors and good decision making. Consistent with these normal behaviors, this circuit is also associated with pathological risk-taking and addictive behaviors. The frontal–basal ganglia network, in general, mediates all aspects of action planning, including reward and motivation, cognition, and motor control. However, specific regions within this network play a unique role in different aspects of reward processing and the evaluation of outcomes, including reward value, anticipation, predictability, and risk. Other brain areas also contain cells that respond to reward, including the hypothalamus that specifically process information about basic drives to maintain homeostasis of the animal. However, this article focuses on those areas that use information about reward to predict, evaluate, and plan future actions, based on perceived positive outcomes.

The prefrontal cortex, the ventral striatum (VS), the amygdala, and the dopamine neurons of the ventral tegmental area (VTA) are at the heart of the reward circuit (Figure 1). The anterior cingulate cortex (ACC) and orbital prefrontal cortex (OFC) areas mediate different aspects of reward-based behaviors, error prediction, value, and the choice between short- and long-term gains. In addition to these prefrontal areas, the dorsolateral prefrontal cortex (DLPFC) is engaged when working memory is required for monitoring incentive-based behavioral responses. Cells in the striatum respond to the anticipation of reward and reward detection. Reward prediction and error detection signals are generated, in part, from the midbrain dopamine cells. In addition, the medial temporal cortex also plays a central role in the motivation and memory for the development of incentive- and goal-directed behaviors. Moreover, although the VS and the VTA dopamine neurons are the basal ganglia areas most commonly associated with reward, reward-responsive activation is not restricted to these areas but is found throughout the striatum and substantia nigra pars compacta (SNc).

Prefrontal Cortex

The main areas of prefrontal cortex involved in the reward circuit are the ACC and OFC. Each of these regions comprises several specific cortical areas. However, based on specific roles for mediating different aspects of reward and error prediction, they can be functionally grouped into the dorsal ACC (dACC), which includes parts of areas 24 and 32; the ventromedial prefrontal cortex (vmPFC), which includes parts of areas 32, 25, and 14; and the OFC, which includes areas 11, 13, and 12. The vmPFC plays a role in monitoring correct responses based on previous experience and the internal milieu. It is selectively connected to the hippocampus, parahippocampal, and amygdala, as well as the hypothalamus. The OFC is engaged when a previously learned response is no longer appropriate and the response needs to be suppressed. This more lateral area is connected to multisensory regions, parts of the DLPFC, and the amygdala. The dACC is a unique part of prefrontal cortex, in that it contains within it a representation of many diverse frontal lobe functions, including motivation, cognition, and motor control. However, its overall functional role is to monitor these functions in conflict situations, which is reflected in its widespread connections with limbic, cognitive, and motor cortical areas.

The Ventral Striatum

The striatum is the main input structure of the basal ganglia. It receives a massive and topographic input from cerebral cortex (and thalamus). These afferent projections to the striatum terminate in a general topographic manner, such that the ventromedial striatum receives input from reward-related areas, the central striatum receives input from associative cortical areas, and the dorsolateral striatum receives cortical input from sensorimotor areas. The VS is the striatal region most associated with reward, supported not only by its cortical inputs but also by physiological, pharmacological, molecular, and imaging studies. Collectively, these studies demonstrate its key role in the acquisition and development of reward-based behaviors and its involvement in drug addiction and drug-seeking behaviors.

Although the VS is similar to the dorsal striatum in most respects, there are also some unique features. The VS contains a subterritory, called the shell, that plays a particularly important role in the circuitry

underlying goal-directed behaviors, behavioral sensitization, and changes in affective states. Because there is no histochemical or cytoarchitectonic boundary between the VS and dorsal striatum, the best way to define it is by its afferent projections from cortical areas that mediate different aspects of reward processing, the vmPFC, OFC, dACC, and medial temporal lobe. Finally, and of particular importance, although both the dorsal striatum and VS receive input from the cortex, thalamus, and brain stem, the VS alone also receives a dense projection from the amygdala and from the hippocampus.

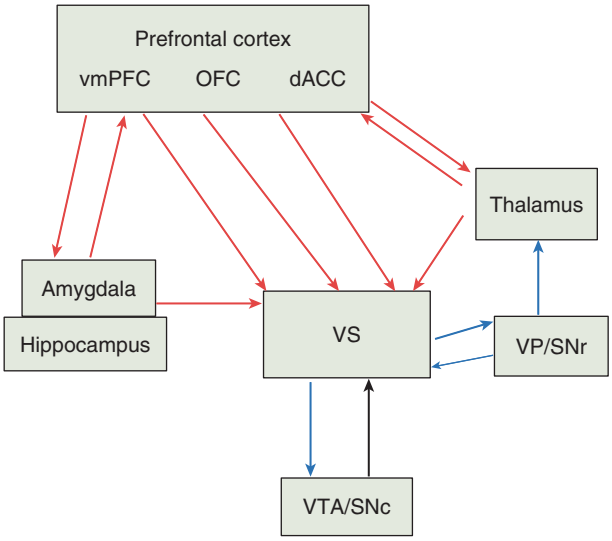

Figure 1 Diagram demonstrating the key elements of the reward circuit. Red arrows indicate excitatory glutamatergic projections; blue arrows indicate inhibitory GABAergic projections. dACC, dorsal anterior cingulate cortex; GABA, γ-aminobutyric acid; OFC, orbital prefrontal cortex; SNc, substantia nigra pars compacta; SNr, substantia nigra pars reticulata; vmPFC, ventromedial prefrontal cortex; VP, ventral pallidum; VS, ventral striatum; VTA, ventral tegmental area.

The vmPFC, OFC, and dACC project primarily to the rostral striatum, including the nucleus accumbens, rostral caudate nucleus, and the medial and ventral rostral putamen. Terminals from these cortical areas show a complex interweaving and convergence, providing an anatomical substrate for modulation between circuits to affect incentive learning and habit formation. The shell receives the most limited projection from cortex and is primarily derived from the vmPFC and agranular insular cortex, areas involved in monitoring the internal milieu. In addition, inputs from both the hippocampus and amygdala terminate in overlapping regions of the shell. The terminal field that receives input from the vmPFC and amygdala continues from the shell dorsally to lie adjacent to the lateral ventricles rostral to the anterior commissure. The OFC and dACC projection fields lie, for the most part, lateral to these fibers from the vmPFC. They occupy the ventral caudate and putamen and the centromedial part of the caudate nucleus. The entire reward-related striatum, as defined by vmPFC–OFC–dACC inputs, occupies a large rostral region and encompasses at least 22% of the entire striatum. The large extent of this region is consistent with the findings that diverse striatal areas are activated following reward-related behavioral paradigms (Figure 2).

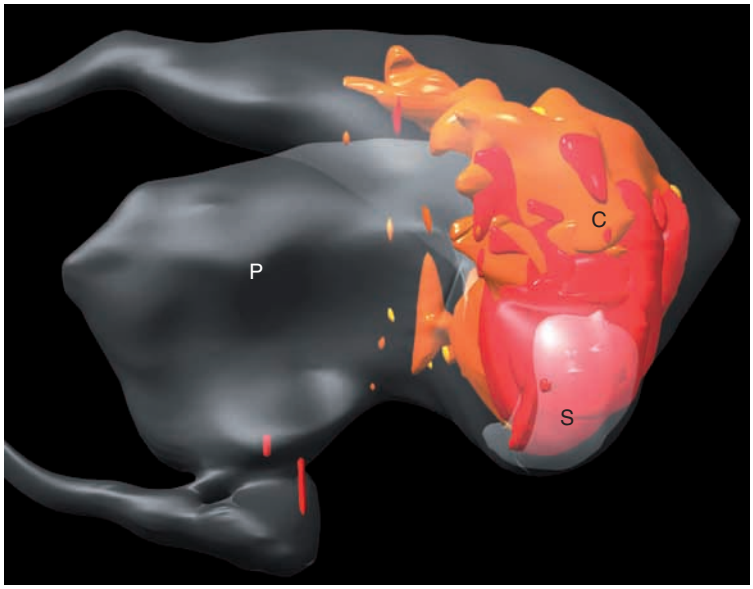

Figure 2 Lateral view of a 3-D rendering of the striatum. Projections from the dorsal anterior cingulate are illustrated in orange and from the vmPFC and orbital prefrontal cortex in red. Note the large rostral striatal region associated with the reward processing areas of cortex. C, caudate nucleus; P, putamen; S, shell.

The Amygdala

The amygdala is a prominent limbic structure that plays a key role in the emotional coding of environmental stimuli and provides contextual information used for adjusting the motivational level that is based on immediate goals. It comprises the basolateral nuclear group (BLNG), the corticomedial region, and the central amygdaloid complex. Overall, the basal nucleus and the magnocellular division of the accessory basal nucleus are the main sources of inputs to the VS. The lateral nucleus has a relatively minor input to the VS. The amygdala innervates the entire ventromedial striatum and has few to no inputs to the dorsal striatum in primates. The basal and accessory basal nuclei innervate both the shell and ventromedial striatum outside the shell. Moreover, there is a topographical organization to that innervation, such that inputs from basal nucleus subdivisions target different regions of the VS. The shell is set apart from the rest of the VS by a specific set of connections derived from the medial part of the central nucleus (CeM) and periamygdaloid cortex and the medial nucleus of the amygdala. The BLNG, the source of all amygdaloid input to the all the VS outside the shell, processes higher-order sensory inputs in all modalities except olfactory and responds to highly palatable foods and multimodal cues. The CeM and periamygdaloid cortex inputs directed to the shell add important information on whether a stimulus matches the animal's internal drive state. Thus, much like the topography of prefrontal inputs to the VS, the nuclei of the amygdala that mediate various functions, target subterritories of the ventromedial striatum.

The Midbrain Dopamine Neurons

Dopamine neurons play a key role in the incentive learning and acquisition of newly acquired behaviors. Behavioral and pharmacological studies of dopamine pathways have led to the association of the mesolimbic pathway and nigrostriatal pathway with reward and motor activity, respectively; however, more recently both of these cell groups have been associated with reward. The midbrain dopamine neurons project widely throughout the brain. However, studies of the rapid signaling that is associated with incentive learning and habit formation focus on the dopamine striatal pathways. Before turning to this projection to the striatum, it is important to understand the organization of the midbrain dopamine cells.

The midbrain dopamine neurons are divided into two fundamental groups: the VTA and the SNc (Figure 3). These cells are also referred to as the dorsal and ventral tier neurons, based on projection and chemical signatures. The dorsal tier includes the VTA and the dorsal part of the SNc (also referred to as the retrorubral cell group). The cells of the dorsal tier are calbindin positive and contain relatively low levels of mRNA for the dopamine transporter and D2 receptor subtype. They project to the VS, cortex, hypothalamus, and amygdala. The ventral tier of dopamine cells is calbindin-negative, has relatively high levels of mRNA for the dopamine transporter and D2 receptor, and projects primarily to the dorsal striatum. The relatively high levels of the dopamine transporter indicate that the ventral tier cells have tight control over regulating extracellular dopamine. Moreover, high levels of the transporter molecule result in these cells being more vulnerable to degeneration in Parkinson's disease and to *N*-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-induced toxicity, whereas the dorsal tier cells are selectively spared. As already mentioned, despite these distinctions both cell groups respond to unexpected rewards but do not respond to movement-related activity.

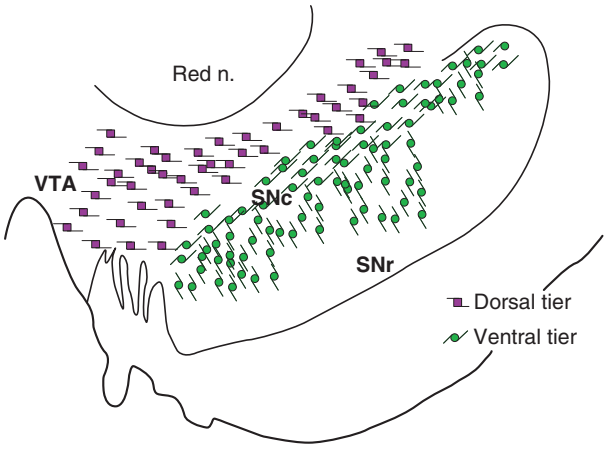

Figure 3 Schematic demonstrating the arrangement of the dopamine cell groups in the midbrain. The dorsal tier includes the ventral tegmental area and retrorubral cell groups. The ventral tier includes the main part of the pars compacta. n., nucleus; SNc, substantia nigra, pars compacta; SNr, substantia nigra, pars reticulata; VTA, ventral tegmental area.

Afferent Projections

Input to the midbrain dopamine neurons is primarily from the striatum, from both the external segment of the globus pallidus and the ventral pallidum, and from the brain stem. In addition, there are projections to the dorsal tier from the bed nucleus of the stria terminalis, the sublenticular substantia innominata, and the extended amygdala (the bed nucleus of the stria terminalis) (Figure 4(a)).

The striatal–nigral projection is the most massive projection to the SN and terminates in both the

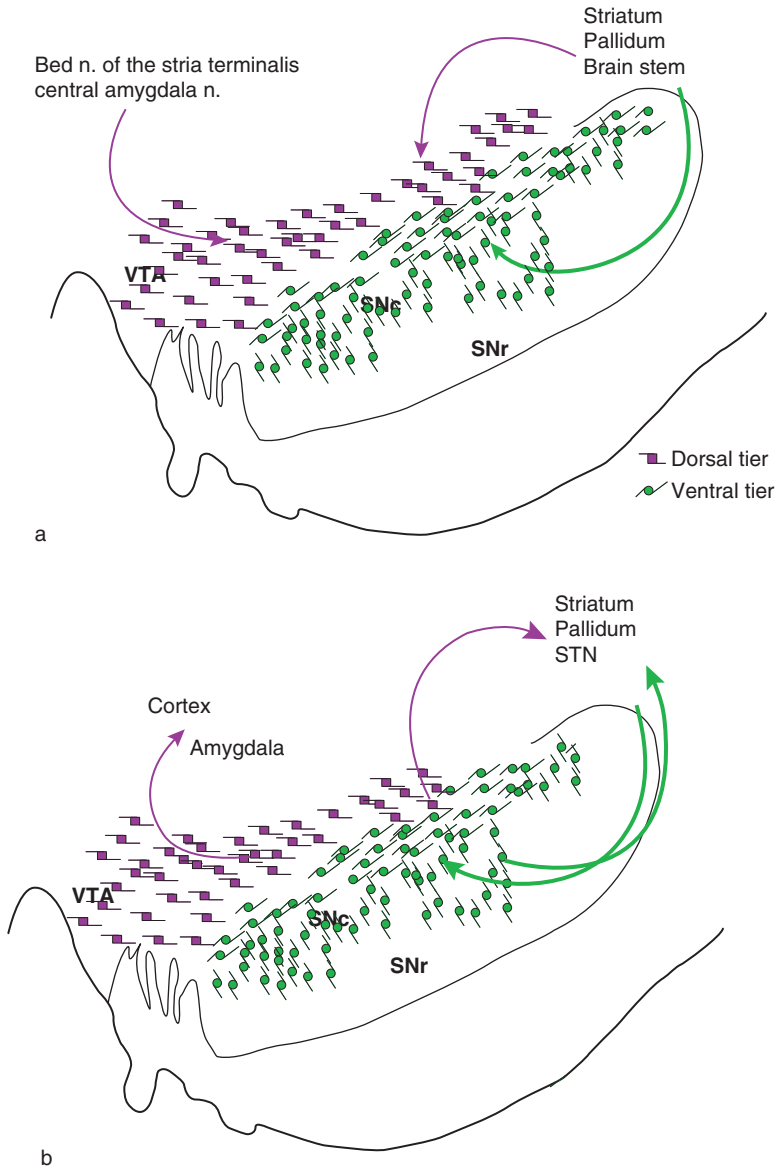

Figure 4 Diagram of the projections of the midbrain dopamine neurons: (a) main afferent projections; (b) main efferent projections. n., nucleus; SNc, substantia nigra, pars compacta; SNr, substantia nigra, pars reticulata; STN, subthalamic nucleus; VTA, ventral tegmental area.

VTA-SNc and the substantia nigra pars reticulata (SNr). The ventral (like the dorsal) striatal–nigral connection terminates throughout the rostrocaudal extent of the SN. There is an inverse ventral–dorsal topography to the striatal–nigral projections. The dorsolateral striatal–nigral inputs are concentrated in the ventrolateral SN. These striatal cells project primarily to the SNr, but also terminate on the cell columns of dopamine neurons that penetrate deep into the SNr. In contrast, the ventral striatal–nigral inputs terminate in the VTA, medial and dorsal SNr, and much of the dorsal part of the ventral tier. Thus, the VS projects not only throughout the rostrocaudal extent of the SN, but also covers a wide mediolateral range.

Descending projections from the central nucleus of the amygdala also terminate in a wide mediolateral region, but they are limited primarily to the dorsal tier cells. In addition, there are projections to the dorsal tier from the bed nucleus of the stria terminalis and from the sublenticular substantia innominata that travel together with those from the amygdala. Both the external segment of the globus pallidus and the ventral pallidum project to the SN. The internal segment of the globus pallidus does not. The pallidal

projection follows a similar inverse dorsal–ventral organization as the striatal–nigral projection. Thus, the ventral pallidum projects dorsally, primarily to the densocellular region of the pars compacta. The pedunculopontine nucleus sends a glutamatergic input to the dopaminergic cells bodies. In addition, there is a serotonergic innervation from the dorsal raphe nucleus, although there is disagreement regarding whether fibers terminate primarily in the pars compacta or pars reticulata. Other brain stem inputs to the dopamine neurons include those from the superior colliculus, the parabrachial nucleus, and the locus coeruleus. Finally, in primates, there is a small and limited projection from the prefrontal cortex to the midbrain dopamine neurons, to both the VTA and SNc. However, relative to the density of its other inputs, this projection is weak.

Efferent Projections

The striatum receives the largest efferent projection from the midbrain dopamine neurons (Figure 4(b)). As with the descending striatal–nigral pathway, there is a mediolateral and an inverse dorsoventral topography arrangement to the projection. The ventral pars compacta neurons project to the dorsal striatum and the dorsally located dopamine neurons project to the VS. The shell region receives the most limited input, primarily derived from the VTA. The rest of the VS receives input from the entire dorsal tier. In addition, there are some afferent projections from the medial region and dorsal part of the SNc. In contrast to the VS, the central striatal area (the region innervated by the DLPFC) receives input from a wide region of the SNc. The dorsolateral (motor-related) striatum receives the largest midbrain projection from cells throughout the ventral tier. In contrast to the dorsolateral region of the striatum, the VS receives the most limited dopamine cell input. Thus, in addition to an inverse topography, there is also a differential ratio of dopamine projections to the different striatal areas. The ascending midbrain dopaminergic striatal afferents terminate on the spines of the medium spiny projection neurons, in close apposition to cortical synapses, and are in a position to influence cortical input. Thus, a relatively smaller midbrain projection to the VS has a large impact on the modulation of the larger cortical projection.

The dorsal tier cells also project widely throughout the primate cortex and are found not only in the granular areas but also in the agranular frontal regions, parietal cortex, temporal cortex, and, albeit sparsely, occipital cortex. The majority of dopamine cortical projections are from the parabrachial pigmented nucleus of the VTA and the dorsal part of the SNc. The VTA also projects to the hippocampus, albeit to a lesser extent than found in neocortex. The dopamine cells that project to functionally different cortical regions are intermingled with one another, in that individual neurons send collateral axons to different cortical regions. Thus, nigral–cortical projection is a more diffuse system compared to the nigral–striatal system and can modulate cortical activity at several levels. Dopamine terminals are located throughout layer I, providing a general modulation of cortical cells at the distal apical dendrites and in layers V–VI, in specific cortical areas. The basolateral complex and the central amygdala nuclei have reciprocal connections to the midbrain. Projections to the amygdala arise from the dorsal tier. These terminals form symmetric synapses primarily with the spiny cells of the amygdala. Dopamine fibers also project directly to the subthalamic nucleus (STN) and the ventral pallidum.

Completing the Cortical–Basal Ganglial Reward Circuit

In addition to the cortical and amygdaloid projections and the midbrain, inputs to the striatum also include projections from the ventral pallidum, the dorsal and median raphe nuclei, and the midline thalamic nuclei. Finally, in addition to the extrinsic inputs, the projection neurons also receive input from interneurons and from local collaterals of other medium spiny cells. The output from the VS targets the ventral pallidum and the VTA of the midbrain, and the medial and dorsal SNc and SNr.

The ventral pallidum is an especially important component of the reward circuit in that cells in this forebrain region respond specifically during the learning and performance of reward-incentive behaviors and it is an area of focus in the study of addictive behaviors. This ventral extension of the globus pallidus was identified based on histological criteria and found to receive its primary input from the VS. The ventral pallidum extends not only ventrally, below the anterior commissure, but also rostrally to invade the rostral and ventral portions of the VS, sending fingerlike extensions into the anterior perforated space. The ventral pallidum has common features of both the external and internal segments. The external pallidal component of the ventral pallidum lies ventrally and rostrally to the external segment and projects to the striatum, the STN, and the SN. The internal segment component of the ventral pallidum lies primarily in the medial and rostral portions of the internal segment proper and projects to the mediodorsal nucleus of the thalamus, and to the midline thalamic nuclei, which in turn project to the prefrontal

cortex, thus completing the basic reward circuit through the basal ganglia (see Figure 1).

Complex Network Features of the Reward Circuit

To develop an appropriate behavioral response to external environmental stimuli, information about motivation and reward needs to be combined with a strategy and an action plan for obtaining the goal. In other words, it is not sufficient to desire or want, for example, to win at a card game. We must understand the rules of the game, remember the cards played, and so forth before executing the play. In addition, there is a complex interaction between the desire to put cards in play and the inhibition of impulse to play them too early. The cortico-cortical interaction plays a central role in strategic planning, memory, and impulse control that allows the ability to inhibit responses that will result in long-term loss. The cortical–basal ganglia circuit serves a different function. This system is particularly important in procedural and incentive-based learning leading to the development of goal-directed behaviors. The resulting action plans for obtaining a goal require a combination of reward processing, cognition, and motor control. Emphasis has generally been placed on the separation between these functions within the striatum, with no substantial overlap among these different cortical–basal ganglia circuits (referred to as parallel processing). However, recent work clearly shows that there are dual corticalbasal ganglia systems permitting both parallel and integrative networks. The integrative role for the basal ganglia is critical in learning new behavior-guiding rules, associations, and habits.

The Cortical–Striatal Network

Focal terminal projections from the different reward-processing areas of the prefrontal cortex (vmPFC, OFC, and dACC) converge in specific regions of the VS (Figure 5; see also Figure 2). The areas of convergence also receive input from the amygdala. These convergent points provide an anatomical substrate for the integration of different reward-processing circuits within specific striatal areas. Areas of convergence among the different reward-related cortical areas are likely to be relatively large and may represent interactions (hot spots) of plasticity to integrate reward value, predictability, and salience. These regions may especially be critical for the development of appropriate action plans and habits, and they are also especially vulnerable to the development of addictive disorders. Moreover, there are also specific points of convergence from the vmPFC, OFC, dACC, and those

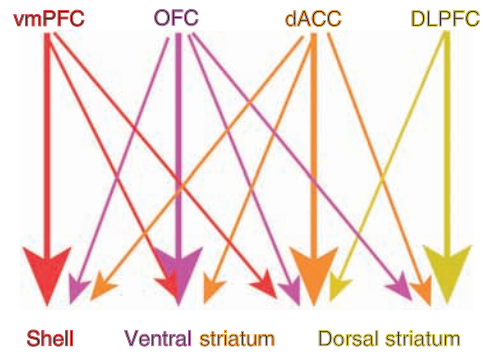

Figure 5 Diagram showing convergence of cortical projections from different reward-related regions of the prefrontal cortex. Note that there is also some convergence between reward areas and the dorsolateral prefrontal cortex. dACC, dorsal anterior cingulate cortex; DLPFC, dorsal lateral prefrontal cortex; OFC, orbital prefrontal cortex; vmPFC, ventral medial prefrontal cortex.

from the DLPFC. The DLPFC plays an important role in executive functions and projects primarily to a large rostral region of the caudate nucleus, extending into the putamen. The hot spots between reward-related cortical areas and the DLPFC form a potential zone for dynamic restructuring of neural ensembles fundamental to learning and habit formation. Finally, the premotor terminals in the striatum converge with those from the DLPFC, placing these projections in a pivotal position in the striatum to integrate reward processing with planned action outcomes. In addition, whereas the projections from the cortex terminate in a general topography through the basal ganglia structures, the dendrites and axons within each structure extend outside of the functional boundaries, increasing the cross-talk between functional regions. This is particularly prominent in the globus pallidus, where the dendrites are extensive, stretching among limbic, cognitive, and motor control areas. In this way, the distal dendrites from one region receive information from other functional areas of the striatum.

The Striatal–Nigral–Striatal Network

The role of dopamine in reward is well established. However, the latency between the presentation of the reward stimuli and the activity of the dopamine cells is too short to reflect higher cortical processing necessary for linking a stimulus with its rewarding properties. The fast burst-firing activity is likely, therefore, to be generated from other input such as in the brain stem glutamatergic nuclei. An interesting issue is, then: How do the dopamine cells receive information concerning reward value? The largest forebrain input to the dopamine neurons is from the striatum. However, this is a relatively slow γ-aminobutyric acid

(GABAergic) inhibitory projection and unlikely to result in the immediate, fast burst-firing activity. Nonetheless, the collective complex network of prefrontal cortex, amygdala, and hippocampal inputs integrates information related to reward processing and memory to modulate striatal activity. These striatal cells then impact directly a subset of medial dopamine neurons, which, through a series of connections, influences the dorsal striatum. Thus, although the short latency burst firing activity of dopamine that signals the immediate reinforcement is likely to be triggered from brain stem nuclei, the cortical–striatal midbrain pathway is in the position to influence dopamine cells to distinguish rewards and modify responses to incoming salient stimuli over time. This pathway is further reinforced via the nigral–striatal pathway, placing the striatal–nigroalstriatal pathway in a pivotal position for incentive-based learning.

As already mentioned, projections from the striatum to the midbrain are arranged in an inverse dorsal–ventral topography and there is also an inverse dorsal–ventral topographical organization to the midbrain striatal projection. When considered separately, each limb of the system creates a loose topographical organization such that the VTA and medial SN are associated with the limbic system, and the lateral and ventral SN are related to the associative and motor-striatal regions. However, each functional region also differs in its proportional projections that significantly alter its relationship with other functional regions. The VS receives a limited midbrain input, but projects to a large region. In contrast, the dorsolateral striatum receives a wide input, but projects to a limited region. In other words, the VS influences a wide range of dopamine neurons, but is itself influenced by a relatively limited group of dopamine cells. On the other hand, the dorsolateral striatum influences a limited midbrain region, but is affected by a relatively large midbrain region.

With this arrangement, while the VS receives input from the vmPFC, OFC, dACC, and amygdala, its efferent projection to the midbrain extends beyond the tight VS–dorsal tier dopamine–VS circuit. It terminates also in the ventral tier to influence the dorsal striatum. Moreover, this part of the ventral tier is reciprocally connected to the central (or associative) striatum. The central striatum also projects to a more ventral region than it receives input from, which, in turn, projects to the dorsolateral (or motor) striatum (Figure 6). Taken together, the interface between different striatal regions via the midbrain dopamine cells is organized in an ascending spiral interconnecting different functional

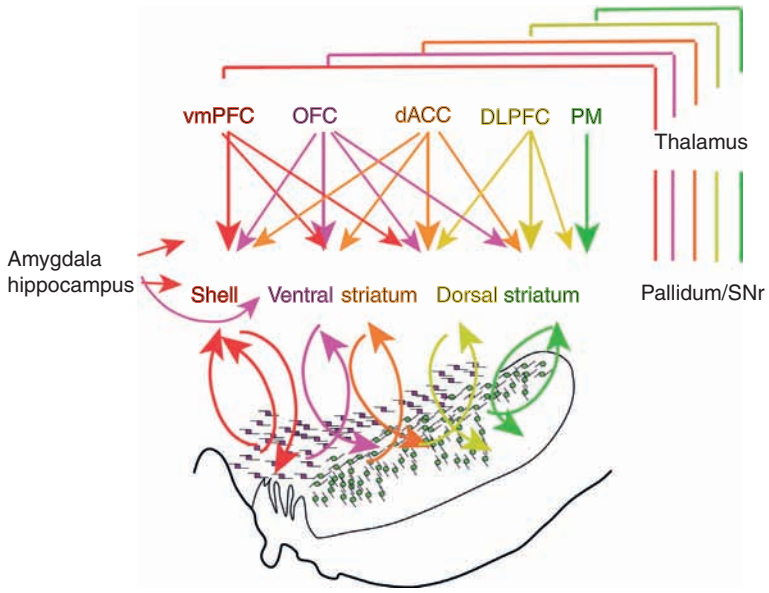

Figure 6 Diagram showing the complex interactions between the prefrontal cortex and the basal ganglia. Illustrated are areas of interaction through which the reward circuit interfaces with associative circuits that mediate action plans. The colors represent different cortical functional regions. Note that the parallel system is carried from the cortex to the striatum, to the output structures of the pallidum and pars reticulata, to the thalamus, and back to the cortex. There is some convergence between terminals from different functional areas of cortex in the striatum. Moreover, the dopamine neurons are in a position to channel information from limbic to cognitive to motor circuits. Also illustrated are the parallel functional circuits from the cortex, through the basal ganglia, and back to cortex via the thalamus. dACC, dorsal anterior cingulate cortex; DLPFC, dorsal lateral prefrontal cortex; OFC, orbital prefrontal cortex; PM, premotor area; SNr, substantia nigra pars reticulata; vmPFC, ventral medial prefrontal cortex.

regions of the striatum and creating a feed-forward organization, from reward-related regions of the striatum to the cognitive and motor areas.

Summary

The reward circuit comprises several cortical and subcortical regions that form a complex network to mediate different aspects of incentive learning. The cortical–basal ganglial network is a fundamental component of this circuit, leading to adaptive behaviors and good decision making. This complex network involves a system in which the prefrontal cortex exploits the basal ganglia for additional processing of rewards to effectively modulate learning, which leads to the development of goal-directed behaviors and action plans. However, the reward circuit does not work in isolation. The development and modification of goal-directed behaviors require continual processing and updating of outcomes relative to response. As such, this complex circuitry interfaces with pathways that mediate cognitive function and motor planning. Within each of the cortical–basal ganglial structures, there are regions of convergence linking up areas that are associated with different functional domains (**Figure 6**). In this way the reward circuit impacts cognition and motor control through several different interactive networks, including cortico-cortical connections, convergence between cortical terminals in the striatum, and a feed-forward system via the midbrain. This complex set of systems allows information about reward channeled in different ways through cognitive and motor-control circuits to mediate the development of appropriate action plans.

See also: Addiction: Neurobiological Mechanism; Appetitive Systems: Amygdala and Striatum; Cortical Processing of the Reward Value of Food; Dopamine: Cellular Actions; Neuropsychology of Primate Reward Processes; Prefrontal Contributions to Reward Encoding; Reward and Learning; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Berns GS (2004) Something funny happened to reward. *Trends in Cognitive Science* 8: 193–194.

Elliott R, Dolan RJ, and Frith CD (2000) Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. *Cerebral Cortex* 10: 308–317.

Everitt BJ and Robbins TW (2005) Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience* 8: 1481–1489.

Haber SN (2003) The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy* 26: 317–330.

Heimer L, DeOlmos JS, Alheid GF, et al. (1999) The human basal forebrain, part II. In: Bloom FE, Bjorkland A, and Hokfelt T (eds.) *Handbook of Chemical Neuroanatomy*, pp. 57–226. Amsterdam: Elsevier.

Koob MD (1999) Drug reward and addiction. In: Zigmond MJ, Bloom FE, Landis SC, Roberts JL, and Squire LR (eds.) *Fundamental Neuroscience*, pp. 1261–1279. San Diego, CA: Academic Press.

Rolls ET (2000) The orbitofrontal cortex and reward. *Cerebral Cortex* 10: 284–294.

Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36: 241–263.

Volkow N and Li TK (2005) The neuroscience of addiction. *Nature Neuroscience* 8: 1429–1430.

Wise RA (2004) Dopamine, learning and motivation. *Nature Reviews Neuroscience* 5: 483–494.

Consciousness: Neural Basis of Conscious Experience

B Libet[†], Davis, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Basic Definitions and Concepts

Conscious or subjective experience or awareness is clearly accessible only introspectively to the individual having the experience. Consequently, the valid operational definition of it in scientific investigations is the introspective report by the individual, under conditions of reliability and credibility. That requires the availability of human study participants, awake and forthcoming, in whom simultaneous observations and manipulations of neural functions can be made. This combination of requirements imposes severe limitations on potential studies.

The observer must ask the individual the appropriate question about the latter's conscious experience and be confident that the individual understands the question. The report is often most conveniently verbal, but it may be nonverbal (e.g., a sign made by a finger) if the latter clearly represents the individual's introspective experience. Any report should be made only after sufficient time for introspection and not as part of any speedy reaction-time procedure. The issue of the accuracy and reliability of the individual's communication of introspective experience must be dealt with in each study individually. Suitable control tests can in fact be devised, and satisfactory reliability of reports is obtainable.

Physical Nature; System Property

Conscious subjective experience is a primary phenomenon whose nature cannot be defined by any other externally observable event, whether molecular or behavioral. Subjective experience arises as a system property of appropriate neural activities in the brain. It is a 'nonphysical' event, in that it cannot be detected or described by direct physical measurements, although a correlation with neural patterns may be discernible in relation to an individual's report of an experience. It is well known that the properties of a system are not predictable from those of the elements in the system. A 'nonphysical' property could, in principle, appear as a system property of the 'physical' neuronal activities. Even a complete knowledge of the observable neural processes in the brain of another individual would not in itself tell us what that individual is experiencing or feeling.

There are no *a priori* rules governing the relationship between conscious mental events and brain events; the rules must be discovered. The phenomenon of subjective referral of a sensory experience in the spatial dimension illustrates this principle. The spatial configuration of a subjectively experienced visual image, for example, is considerably different from the configuration of the brain neuronal activities that accompany and represent the sensory input and give rise to the visual images; the image that is seen, subjectively, is quite different from the neuronal pattern for that image. The cerebral neuronal representation is actually a spatial distortion of the original image coming into the eye, but subjectively the image is 'referred' to the original source of the input in a way that corrects the neuronal distortion. Direct electrical stimulation of primary sensory cortex provides an even more obvious demonstration of this; a stimulus applied to the lateral portion of the postcentral gyrus produces a sensation that is felt to be located not at the stimulated site of the brain, but is rather subjectively referred to, and felt in, the hand or arm and in a direction reversed from the postcentral representation. Clearly, a complete knowledge of the neuronal representation would not, without validation by the individual's report, tell us where or what sensation is being subjectively experienced. Similarly, the experience of color cannot be described by finding that certain neural units may respond to specific frequency bands of the light spectrum.

Although each individual has access to and can be certain about only his own conscious experiences and feelings, we do commonly concede and accept the premise that other human beings have their subjective experiences (except for those of us who want to adopt Bishop Berkeley's solipsism as a serious view of life). We are also confident that one person can communicate something about his subjective experiences to another. But the validity of what is communicated depends upon the degree to which both individuals have had similar or related experiences. A congenitally completely blind person can never share the conscious experience of a visual image, regardless of how detailed a verbal description he is given by a sighted individual. The same limitation applies to all experiences in less dramatic, more subtle ways. For example, electrical stimulation of somatosensory cortex can produce sensations related to but sufficiently different from those generated by normal sensory input, so that the individuals could only relate some roughly understandable approximation

[†]Deceased.

of these experiences to the experimental observer, in whom similar modes of sensory generation had never been employed.

Reports of Conscious Experience versus Behavioral Detection

There is an imperative corollary of the foregoing operational definition of conscious experience: any measured indicator that can be dissociated from or independent of subjective awareness would be invalid, not acceptable as a primary indicator of subjective experience. This would even include behavioral responses that depend on cognitive and decision-making processes, unless these were validated by the primary evidence of the individual's introspective reports. For example, we have all had the experience of driving an automobile and becoming subjectively engrossed in thoughts not related to the mechanics of driving; nevertheless, the driving proceeds while all sorts of sensory signals are being properly recognized, evaluated, and acted upon (usually successfully), without any consciously introspective awareness or later recallability. Among experimental paradigms, signal detection studies provide sophisticated examples of a generally unacceptable approach to conscious experiences. The forced-choice responses in such studies could be made independent of introspective awareness of the signal, although they may be excellent indicators of whether some type of detection has occurred. Even when individuals are giving confidence ratings of their responses in a signal detection study, they are rating their forced choices and are not necessarily directly reporting their subjective experiences. There is, in fact, evidence that signal detection can occur with signals that are distinctly below the threshold required for any conscious awareness of the signal. Indeed, most sensory signals probably do not reach conscious awareness; however, many of them lead to modified responses and behaviors, as in simple everyday postural and walking activities, and have therefore clearly been detected and utilized in complex brain functions.

Do Animals Have Conscious Experience?

There is no way to answer this in any absolute sense. But it should be clear that it is difficult, if not impossible, to carry out valid studies of conscious experience in nonhuman animals. Obviously, we cannot meaningfully ask an animal to report about a shared introspective experience, the validity of which we have confidence. Second, as just noted, complicated cognitive and purposeful behaviors can proceed even in human beings without introspective awareness of them, and so we cannot safely assume that such behaviors in animals are expressing subjective experience. Some investigators and writers have proposed that adaptive, problem-solving behaviors in animals indicate conscious thought and experience. However, even in humans, the most complex and even mathematical problem solving can and often does proceed at unconscious levels, as has been repeatedly described by many creative thinkers, artists, and others. Again, one must be careful to hold to the primary criterion of conscious experience *per se*, as a phenomenon that cannot automatically be described by any nonvalidated behavioral expression.

Are Computers Conscious?

Some enthusiasts of machines with artificial intelligence have tended to equate complexity of abstract processing with the existence of conscious experience. As we have noted, there is no such necessary connection even in the human brain. Again, there is the fundamental error in assuming that conscious experience can be identified based on purely behavioral criteria, whether for humans or machines. It has been argued that a machine might eventually be constructed with such sophistication that an external observer could not distinguish its behavior and responses from those of a human being. Even if we grant such a possibility, for the sake of argument, such an apparent identity of behavioral responses does not require that the two systems, machine and human brain, are alike in every respect. In fact, they are constituted differently and operate with different mechanisms. The unique makeup of the brain, and indeed certain special neuronal actions in it (see later), may be essential correlates of what we recognize as conscious subjective experience.

Studies of Neural Basis

Effects of Cerebral Lesions

It was, of course, long known that lesions of the cerebral cortex that destroyed a primary sensory area resulted in severe sensory impairment. For example, with loss of striate cortex area 17 patients are blind; that is, they report being unaware of signals in the affected visual field. But further analyses are showing that these patients can detect signals without any reportable awareness, by pointing correctly to them in forced-choice tests; this is the 'blindsight' phenomenon. In a related manner, patients with lesions that abolish their reportable awareness of differences in color can correctly discriminate among different colors in

forced-choice tests, and with lesions that obliterate conscious identifiability of the shapes and locations of objects, patients can retain correct visuomotor guidance (of their hands) to these objects. A similar discrepancy, between conscious and unconscious (or nonconscious) processes, can be produced by lesions in the frontal medial cortex: patients exhibit the 'alien-hand sign,' in which the affected arm and hand perform spontaneous purposeful-looking movements that are not consciously willed or amenable to direct conscious control. A variety of other cerebral lesions can produce other interesting distortions of, or losses in, various kinds of awareness.

The foregoing kinds of findings in neuropsychology indicate cortical areas that are necessary for developing conscious experiences of sensory and volitional events. But they do not demonstrate that these areas are sufficient for this function; that is, they do not show where the conscious experience actually forms or arises. This distinction between necessity and sufficiency has at times been neglected, leading to potentially erroneous conclusions about the 'seat of consciousness'; it presents a problem that is difficult to address experimentally. Also, the neuropsychological findings do not themselves provide evidence on the physiological activities and interactions of the neuronal groups involved in the conscious process.

Regional Cerebral Blood Flow and Metabolism

The ability to measure regional cerebral blood flow (rCBF) and relating this to psychological functions in awake humans was introduced over 25 years ago by N Lassen and D Ingvar. The analysis has since been technically improved, as to localizability and quantitativeness, by the introduction of positron emission topography (PET) scan and by the coupling of this with the structural mapping by magnetic resonance imaging (MRI). These techniques are essentially non-invasive, except for the administration of presumably harmless tagged agents, and have thus opened a highly productive avenue for studying brain activity in the awake and responsive human. However, while these methods tell us which areas or nuclei in the brain are changing in their rate of metabolism, they do not give information on the physiological dynamics involved; that is, the information is essentially topographical. The timing of the change is also still only possible to within some seconds, for PET scans, but MRIs are faster. The rCBF studies in which individuals are asked to perform conscious mental operations are of special interest to the issue of brain and conscious experience. In one such study, individuals were asked to imagine moving their fingers without actually moving them; a selective increase in blood flow appeared in the supplementary motor area (SMA). Such findings might bear on the question of whether and where a conscious mental function can initiate or influence neural processes. The time resolution of the rCBF method, however, as well as the absence of indicators of the precise relative timing of the onset of the mental imagining, vitiate any answers to the question. It has been shown experimentally that conscious intention to move, for example, actually follows onset of electrophysiological activity associated specifically with the volitional process.

It should be noted that, although the machine for the functional MRI (fMRI) responses can show changes within a few milliseconds, the relevant fMRI responses cannot appear for perhaps 1000 ms or more after the relevant neuronal event. This is because fMRI measures a local change in blood flow or in some neural metabolite. Such changes require a substantial time to become evident after the relevant neural event has initiated the processes that produce them. That is, the time of a relevant fMRI change is delayed by the slower circulatory and metabolic changes that follow the neural activity, even though the machine itself can be very fast. Timing of the relevant neuronal activities can only be obtained by direct recordings of the electrical or magnetic changes that immediately are manifested by the relevant neural activities.

Electrophysiological Stimulation and Recording

The numerous studies with extracranial (scalp) recordings of event-related potentials (ERPs) over recent decades have not generally been sufficiently relevant to the issue of conscious experience. This is partly because the associated psychological tasks for the study participants were mostly those of detection of, recognition of, and responsiveness to signals, functions that could all potentially be performed unconsciously unless direct introspective reports of awareness are demanded. Scalp recordings also suffer from uncertainties of localization of the neuronal sources of the ERPs, although this is improving with recent technological advances. Whether the recording of magnetic rather than electrical fields will improve localizability significantly remains to be seen. Also, even when conscious reports are appropriately included, it is extraordinarily difficult to design experiments that may indicate causal interrelationships, as there is little possibility to manipulate one of the independent variables, the cerebral neural activity.

Intracranial electrodes can greatly improve both the localizability of electrical manifestations and the ability to manipulate neuronal functions in some

quantitative fashions by electrical stimulation. On the other hand, the availability of such electrodes in loci suitable for study in awake, responsive patients is not common, and the kinds of experimental procedures permitted by the therapeutic and risk factors are greatly restricted. Fruitful experimental avenues have, however, been found. Stimulation can elicit conscious experiences, especially when applied in the cerebral sensory systems. It can also reversibly interfere with or 'disrupt' normal cerebral functions. This latter ability can be used to investigate functions of cerebral areas where stimulation does not itself elicit any conscious responses; it has been employed fruitfully in recent years by GA Ojemann and colleagues to analyze the language functions of temporal (and other) cortical areas.

Many neural responses to stimulation can develop without leading to any conscious experience. For example, stimulating the cortical surface can produce large electrophysiological direct cortical responses (DCRs) with no awareness by the individual. Similarly, the primary evoked potential elicited by a peripheral sensory stimulus leads to no conscious sensation if the appropriate later event-related potentials do not accompany the primary one. Unique kinds and durations of neuronal activities appear to be required to elicit a conscious sensory experience, but not unconscious forms of stimulus detection.

In recent years the ability to deliver brief magnetic pulses repetitively has been achieved, thus permitting effective stimulation of a cortical area with a magnetic device placed on the overlying scalp. This makes it possible to stimulate the cortex without the invasive surgery required to place an electrical stimulating wire in a local region, expanding the pool of potential individuals greatly, to include normal individuals. However, magnetic stimulation may be deficient in some important ways, relative to an intracranial electrical stimulator. Stimulation with an electrode inserted intracranially can be more localizable and affect more identifiable groups of stimulated neural elements. Also, one can be more certain of a threshold level for eliciting a neural or a subjective response; with a magnetic stimulus, it is possible that each single stimulus pulse at an apparently threshold level is actually producing an undetected series of repetitive neural responses.

Time Factors in Conscious and Unconscious Mental Functions

Neural Delay for Sensory Experience

Substantial duration of specific cerebral activations, of up to about 500 ms, have turned up experimentally as one of the most interesting neuronal requirements. This implies there is a substantial delay before neuronal adequacy for a sensory experience is achieved. However, further evidence indicates there is a subjective referral of the experience back to the time of the initial sensory signal that normally arrives at the cerebral cortex within 15–25 ms after a sensory stimulus; the primary evoked potential represents the response to this early signal, which comes up by way of the fast, specific projection system. Subjectively, then, the skin sensation would appear to have no significant delay, even though the experience is not neuronally elicitable until some hundreds of milliseconds later. Such a subjective referral in the temporal dimension is analogous to that in the spatial dimension. Both referrals serve to project the subjective image closer to the spatial and temporal features of the real stimuli, even though the adequate neuronal representation distorts both the spatial pattern and the timing of the process.

Neural Delay for Intention to Act

When individuals perform a self-initiated voluntary motor act, a slow surface-negative potential is recordable at the vertex (the 'readiness potential,' or RP); this begins 550 ms before the muscle is activated, in a fully spontaneous act. However, conscious intention to act appears only 350 ms after onset of the RP, although 200 ms before the act. This indicates that initiation of a voluntary act is made unconsciously by the brain before there is any awareness of the wish to act. But, also, there is time after conscious intention appears during which the volitional process could be consciously stopped or vetoed.

'Time-On' Theory for Conscious versus Unconscious Functions

Based on their earlier lines of evidence indicating a requirement for a substantial duration of neural activities in order to elicit a conscious experience, whether a sensation or intention (to move now), in 1991 Libet and colleagues devised and tested a theory to explain how the brain may distinguish between conscious and unconscious mental functions. The theory proposes the duration, or 'time-on,' of appropriate neural activities as one controlling factor in achieving this distinction: durations must exceed minimum values of up to about 500 ms to achieve conscious awareness of a mental process, but durations distinctly below those minimum values can mediate unconscious mental functions (without awareness). A specific experimental test of the proposal was carried out in relation to detection of a sensory signal with and without awareness.

Unity of Conscious Experience

One of the most mysterious and seemingly intractable problems in the mind–brain relationship is that of the unitary and integrated nature of conscious experience. This phenomenon is somehow a product of a brain with an estimated 100 billion neurons, each of which may have thousands of interconnections with other neurons. There is increasing evidence that many functions of cerebral cortex are localized, apparently organized into specialized columns of neurons. In spite of the enormously complex array of structures and functions, whatever does reach awareness is experienced as unified and integrated.

A recent neurophysiological observation may offer one possible route to investigating the issue of experiential integration, or, as it is often referred to, the 'binding problem.' This is the discovery by Charles M Gray and Wolf Singer of widespread synchronization of oscillatory neuronal electrical potentials in response to a visual image. It has led to some speculation that a 'correlation' model (based on the widely occurring synchronization of some activities) might represent the neural coding for a unified mental image in an otherwise chaotic background. This speculation is still to be tested.

A testable field theory We may view subjective experience as if it were a field, produced by appropriate though multifarious neuronal activities of the brain. This conscious mental field (CMF) would have the attribute of a unitary, integrated experience or awareness, and it would also have a causal ability to affect or alter neuronal function. The theory of a CMF makes some crucial experimental predictions. If local areas of cerebral cortex could independently contribute to or alter the larger, unitary CMF, it should be possible to demonstrate such contributions when (1) that cortical area is completely isolated or cut off from neuronal communication with the rest of the brain, but (2) the area remains *in situ*, alive, and kept functioning in some suitable manner that sufficiently resembles its normal behavior. A fuller discussion of this theory and the experimental design to test it have been previously elaborated. The point here is to show that it is possible to formulate a theory and design an experimental test of it, even when the theory deals with the profound issues of conscious unity and with the question of whether the conscious function can actually influence brain function.

See also: Awareness: Functional Imaging; Consciousness: Theories and Models; Consciousness: Philosophy; Consciousness: Neurophysiology and Visual Awareness in; Consciousness: Theoretical and Computational Neuroscience; Event-Related Potentials (ERPs) and Cognitive Processing; fMRI: BOLD Contrast; Perfusion MRI; Positron Emission Tomography (PET).

Further Reading

Amassian VE, Somasundaram M, Rothswell JC, et al. (1991) Parasthesias elicited by single pulse magnetic coil stimulation of motor cortex in susceptible humans. *Brain* 114: 2505–2520.

Ciba Foundation Symposium 174 (1993) *Experimental and Theoretical Studies of Consciousness* (Nagel T, chair). Chichester: Wiley.

Eccles JC (1994) *How the Self Controls Its Brain*. Berlin: Springer.

Goldberg G, Mayer NH, and Toglia JU (1981) Medial frontal cortex infarction and the alien hand sign. *Archives of Neurology* 38: 683–686.

Ingvar D and Phillipson L (1977) Distribution of cerebral blood flow in the dominant hemisphere during motor ideation and motor performance. *Annals of Neurology* 2: 230–237.

Libet B (1973) Electrical stimulation of cortex in human subjects, and conscious sensory aspects. In: Iggo A (ed.) *Handbook of Sensory Physiology, Volume 2, Somatosensory System*, pp. 743–790. Berlin: Springer.

Libet B (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8: 529–566.

Libet B (1993) *Neurophysiology of Consciousness*. Boston, MA: Birkhauser.

Libet B (1994) A testable field theory of mind–brain interaction. *Journal of Consciousness Studies* 1: 119–126.

Libet B (2006) Reflections on the interaction of the mind and brain. *Progress in Neurobiology* 78(3–5): 322–326.

Libet B, Alberts WW, Wright EW, et al. (1975) Cortical representation of evoked potentials relative to conscious sensory responses and of somatosensory qualities in man. In: Kornhuber HH (ed.) *The Somatosensory System*, pp. 291–308. Stuttgart: G. Thiene.

Libet B, Wright EW Jr., Feinstein B, et al. (1979) Subjective referral of the timing for a conscious sensory experience, a functional role for the somatosensory specific projection system in man. *Brain* 102: 191–222.

Libet B, Pearl DK, Morledge DE, et al. (1991) Control of the transition from sensory detection to sensory awareness in man by the duration of a thalamic stimulus. The cerebral time-on factor. *Brain* 114: 1731–1759.

Nagel T (1979) *Mortal Questions*. Cambridge: Cambridge University Press.

Ojemann GA, Fedio P, and van Buren JM (1968) Anomia from pulvinar and subcortical parietal stimulation. *Brain* 91: 99–116.

Penfield W (1958) *The Excitable Cortex in Conscious Man*. Liverpool: Liverpool University Press.

Weiskrantz L (1986) *Blindsight: A Case Study and Implications*. Oxford: Clarendon.

Consciousness: Neurophysiology and Visual Awareness in

S L Macknik and S Martinez-Conde, Barrow Neurological Institute, Phoenix, AZ, USA

Published by Elsevier Ltd.

Consciousness as a Neurophysiological Problem

Consciousness is the feeling of life's experience. It is a difficult concept to define precisely, however, because the word 'consciousness' means different things in different contexts. For instance, one may say that we are unconscious during sleep, despite the fact that, while asleep, we sometimes experience powerful and highly salient dreams (these dreams fit the definition of consciousness as stated in the first sentence above). Does this mean that it is possible to be conscious of our unconscious experiences (an obvious semantic contradiction)? Or does this mean that dreams during sleep are not really unconscious? Because of these and many other semantic difficulties, the theorist Francis Crick (who also described the structure of DNA with James Watson), and his collaborator Christof Koch, suggested that we set aside the semantics and avoid defining the term consciousness linguistically. Instead, we should strive to establish the neural correlates of experience, or the neural correlates of consciousness (NCC). By determining the NCC, we will eventually arrive at a neurophysiological definition of consciousness. Following from this proposal, the field is currently at work to define consciousness by its neurophysiological underpinnings. Today, thanks in large extent to Crick and Koch's efforts to validate the neurobiological study of consciousness, most neuroscientists and cognitive scientists are receptive to, and familiar with, the concept that consciousness, as with any other brain function (sensory, motor, emotional, and/or cognitive), must have a neurobiological substrate.

In studying the NCC, one of the main strategies has been to make use of the vast library of knowledge that neurobiologists have accumulated, for many decades, on the primate visual system. Therefore, our discussion of the neurophysiological underpinnings of consciousness focuses on those studies conducted in the visual system. The visual system of primates and humans offers the most extensive library we have available on the anatomy, physiology, and psychophysics of perceptual experience. The authors of this article have begun to research the nature of the neurobiological substrate of consciousness by isolating the neural correlates of a very small facet of our conscious experience: the experience of visibility, which is arguably the most basic function of the visual system. By 'visibility,' we do not mean the entire process of vision, but simply whether a stimulus is visible, or not. In this sense, visibility is the beginning of (or the necessary condition for) visual perception, not the conclusion. Without visibility, the stimulus cannot have significance or meaning: we perceive stimulus attributes such as color, motion, and depth only if the stimulus is also visible. Thus, when the underpinnings of visibility are discovered, they will include at least some of the circuits that cause awareness. This does not imply that visibility circuits must come before other types of circuits. For instance, it may not be necessary that a stimulus be visible before other types of processing take place (i.e., the corneal reflex). Whereas an unconscious visual reflex such as the accommodation of the eye may be a type of visual experience, stimuli that invoke visual reflexes are not necessarily consciously visible. That is, we are not aware of them despite the fact that they produce an appropriate and automatic response. As far as we know, visibility may take place subsequently to the processing of other visual attributes, or all visual attributes may come together in parallel, or there may be a disunity of processing for different attributes (as suggested by Semir Zeki), or there could be a combination of all these possibilities. Whatever the sequence of events that eventually leads to visual awareness, we cannot be aware of a visual stimulus, by our definition, unless it is visible.

Requirements for Establishing the NCC

Let us assume that visual awareness is correlated to brain activity within specialized neural circuits, and that not all brain circuits maintain awareness. It follows that the neural activity that leads to reflexive or involuntary motor action may not correlate with awareness because it does not reside within awareness-causing neural circuits.

By our proposal, a stimulus that has become just-noticeable, or just-visible, has already activated awareness-producing circuits. Thus, awareness is neither a final process in vision, nor a very low-level process, but it is instead an intermediate-level process. A less than just-noticeable stimulus may activate neurons within early visual areas, but within circuits that do not maintain awareness; therefore, the stimulus remains invisible. Varying the level of salience of a percept (for instance, by varying its brightness above and beyond the just-noticeable threshold) may equate to varying the level of visibility and awareness.

The idea of consciousness as an intermediate-level process, limited to intermediate brain areas, was first

proposed by Jackendoff. According to Jackendoff's proposal, the 'homunculus' in the frontal lobe is probably not conscious: we do not have direct conscious access to our thoughts, but only to images, sounds, speech, and other bodily feelings associated with intermediate brain representations. Jackendoff's theory provides a solution for the classical problem of the infinite regress of consciousness: given that the top homunculus is not conscious, there is no need for infinite homunculi at higher and higher levels in the brain in order to read the output of the previous, conscious homunculus.

Let us begin with the hypothesis that there is a minimal set of conditions necessary to achieve visibility, in the form of a specific type (or types) of neural activity within a subset of brain circuits. This minimal set of conditions will not be met if the correct circuits have the wrong type of activity (too much activity, too little activity, sustained activity when transient activity is required, etc). Moreover, if the correct type of activity occurs, but solely within circuits that do not maintain awareness, visibility will also fail. Finding the conditions in which visibility fails is critical to the research described here: although we do not yet know what the minimal set of conditions is, we can nevertheless systematically modify potentially important conditions to see if they result in stimulus invisibility. If so, the modified condition is potentially part of the minimal set.

To establish the minimal set of conditions for visibility we need to answer at least four questions:

1. What stimulus parameters are important to visibility?
2. What types of neural activity best maintain visibility (transient vs. sustained firing, rate codes, bursts of spikes, etc. – that is, what is the neural code for visibility)?
3. What brain areas must be active to maintain visibility?
4. What specific neural circuits within the relevant brain areas maintain visibility?

We must also determine the set of standards that will allow us to conclude that any given brain area, or neural circuit within an area, is responsible for generating a conscious experience. Parker and Newsome developed a "list of idealized criteria that should be fulfilled if we are to claim that some neuron or set of neurons plays a critical role in the generation of a perceptual event." If one replaces the words "perceptual event" with "conscious experience," Parker and Newsome's list can be used as an initial foundation for the neurophysiological requirements needed to establish whether any given neuron or brain circuit may be the neural substrate of awareness. The list is as follows:

1. The responses of the neurons and of the perceiving subject should be measured and analyzed in directly comparable ways.
2. The neurons in question should signal relevant information when the organism is carrying out the chosen perceptual task. Thus, the neurons should have discernable features in their firing patterns in response to the different external stimuli that are presented to the observer during the task.
3. Differences in the firing patterns of some set of the candidate neurons to different external stimuli should be sufficiently reliable in a statistical sense to account for, and be reconciled with, the precision of the organism's responses.
4. Fluctuations in the firing of some set of the candidate neurons to the repeated presentation of identical external stimuli should be predictive of the observer's judgment on individual stimulus presentations.
5. Direct interference with the firing patterns of some set of the candidate neurons (e.g., by electrical or chemical stimulation) should lead to some form of measurable change in the perceptual responses of the subject at the moment that the relevant external stimulus is delivered.
6. The firing patterns of the neurons in question should not be affected by the particular form of the motor response that the observer uses to indicate his or her percept.
7. Temporary or permanent removal of all or part of the candidate set of neurons should lead to a measurable perceptual deficit, however slight or transient in nature.

However, visual circuits that may pass muster with Parker and Newsome's guidelines may nevertheless fail to maintain awareness, as explained below. To isolate awareness-maintaining processes, some additional strategies and principles must be added to guide the search for the NCC.

The first strategy concerns the use of illusions as the tool of choice to test whether a neural tissue maintains awareness. Visual illusions, by definition, dissociate the subject's perception of a stimulus from its physical reality. Thus, visual illusions are powerful devices in the search for the NCC, as they allow the differentiation of neural responses to the physical stimulus from neural responses that correlate to perception. Our brains ultimately construct our perceptual experience, rather than reconstruct the physical world. Therefore, an awareness-maintaining circuit should express activity that matches the conscious

percept, irrespective of whether it matches the physical stimulus. Neurons (circuits, brain areas) that produce neural responses that fail to match the percept provide the most useful information because they can be ruled out, unambiguously, as part of the NCC. The result is that the search for the NCC can be focused to the remaining neural tissue. Conversely, neurons that do correlate with perception have not necessarily been shown to be critical to awareness, as they could potentially be maintaining awareness or they may simply play a support role (among other possibilities) without causing awareness themselves; they would correlate with perception either way.

The second strategy derives from a major contribution of Crick and Koch's framework: the distinction between explicit and implicit representations. In an explicit representation of a stimulus feature, there is a set of neurons that represent that feature without substantial further processing. In an implicit representation, the neuronal responses may account for certain elements of a given feature; however, the feature itself is not detected at that level. For instance, all visual information is implicitly encoded in the photoreceptors of the retina. The orientation of a stimulus, however, is not explicitly encoded until the level of area V1, where orientation-selective neurons and functional orientation columns are first found. Crick and Koch propose that there is an explicit representation of every conscious percept.

We propose the following corollary to Crick and Koch's idea of explicit representation. Before one can test a neural tissue for its role in the NCC, such tissue must be shown to explicitly process the test stimulus. This corollary constrains the design of neurophysiological experiments aimed to test the participation of specific neurons, circuits, and brain areas in the NCC.

For instance, if one found that retinal responses do not correlate with auditory awareness, such a discovery would not carry great weight. The neurons in the eye do not process auditory information, and so it is not appropriate to test their correlation to auditory perception. However, this caveat also applies to more nuanced stimuli. What if area V1 was tested for its correlation to the perception of faces versus houses? Faces and houses are visual stimuli, but area V1 has never been shown to process faces or houses explicitly, despite the fact that visual information about faces and houses must implicitly be represented in area V1. Therefore, one cannot test area V1's correlation to awareness using houses versus faces and expect to come to any meaningful conclusion about V1's role in the NCC. Because that form of information is not explicitly processed in V1, it would not be meaningful to the NCC if neurons in V1 failed to modulate their response when the subject is presented with faces versus houses.

It follows that some stimuli are inadequate for localizing awareness within neural tissues, because no appropriate control exists to test for their explicit representation. For example, binocular rivalry stimuli pose a special problem in the study of visual awareness. Binocular rivalry (discovered by Charles Wheatstone in 1838) is a dynamic percept that occurs when two disparate images that cannot be fused stereoscopically are presented dichoptically to the subject (i.e., each image is presented independently to each of the subject's eyes). The two images (or perhaps the two eyes) appear to compete with each other, and the observer perceives repetitive undulations of the two images, so that only one of them dominates perceptually at any given time (if the images are large enough then binocular rivalry can occur in a piecemeal fashion, so that parts of each image are contemporaneously visible).

This bistable stimulus has been used as a tool to assess the neuronal correlates of visibility but has generated controversy because of conflicting results. Some human functional magnetic resonance imaging (fMRI) studies report that blood oxygenation level-dependent (BOLD) activity in V1 correlates with visual awareness of binocular rivalry percepts. In contrast, other human fMRI studies, and also single-unit recording studies in primates, suggest that activity in area V1 does not correlate with visual awareness. One possible reason for this discrepancy is that none of these studies determined that the visual areas tested contain the interocular suppression circuits necessary to mediate binocular rivalry. That is, since binocular rivalry is a process of interocular suppression, the neural tissue underlying the perception of binocular rivalry must be shown to produce interocular suppression – explicitly. Otherwise, it cannot be demonstrated that binocular rivalry is a valid stimulus for testing the NCC in such tissue. Thus, awareness studies using binocular rivalry are valid only in those areas that have been shown to maintain interocular suppression. If binocular rivalry fails to modulate activity within a visual area, one cannot know, by using binocular rivalry alone, whether the perceptual modulation failed because awareness is not maintained in that area or because the area does not have circuits that drive interocular suppression. This is more than just a theoretical possibility: we have shown, as we describe further below, that the initial binocular neurons of the early visual system (in areas V1 and V2) are binocular for excitation, but they nevertheless fail to process interocular suppression explicitly.

Since there is no monoptic form of binocular rivalry, one cannot use binocular rivalry by itself to test the strength of interocular suppression. One could use binocular rivalry in tandem with a different stimulus, such as visual masking stimuli (described below), to test for the explicit representation and strength of interocular suppression. But in such a case, the role of the tissue in maintaining visibility and awareness would have been probed by the visual masking stimuli, thus obviating the need for binocular rivalry stimuli. Because one must rely on non-binocular rivalry stimuli to determine the explicit representation and strength of interocular suppression in a given area, it is not possible to unambiguously interpret the neural correlates of perceptual state using binocular rivalry alone.

Visual masking is a type of visual illusion that has been used as a primary tool to isolate the NCC. Visual masking illusions come in different flavors, but in all of them a visible target (a visual stimulus, such as a rectangle), or some specific aspect of a visible target (for instance, the semantic content of a word displayed visually), is rendered invisible by changing the context in which the target is presented, without actually modifying the physical properties of the target itself (Figure 1). That is, the target becomes less visible due solely to its spatial and/or temporal context. Visual masking illusions allow us to examine the brain's response to the same physical target under varying levels of visibility. All we need to do is measure the perceptual and physiological effects of the target when it is visible versus invisible and we will determine many, if not all, of the conditions that cause visibility.

Our visual masking studies have shown that binocular neurons in areas V1 (the first stage in the visual hierarchy where information from the two eyes is combined) and V2 of humans and monkeys can integrate excitatory responses between the eyes. However, these same neurons do not express interocular suppression between the eyes. That is, binocular neurons in V1 are largely binocular for excitation while nevertheless being monocular for suppression. In summary, most early binocular cells do not explicitly process interocular suppression, and so these neurons cannot process binocular rivalry explicitly. Thus, binocular rivalry is an inappropriate stimulus to probe these visual areas for the NCC. This finding renders the results from binocular rivalry studies in the early visual system uinterpretable with respect to localizing the NCC: the fact that early visual areas are not correlated to awareness of binocular rivalry is equivalent in significance to concluding that these areas are not correlated to auditory awareness.

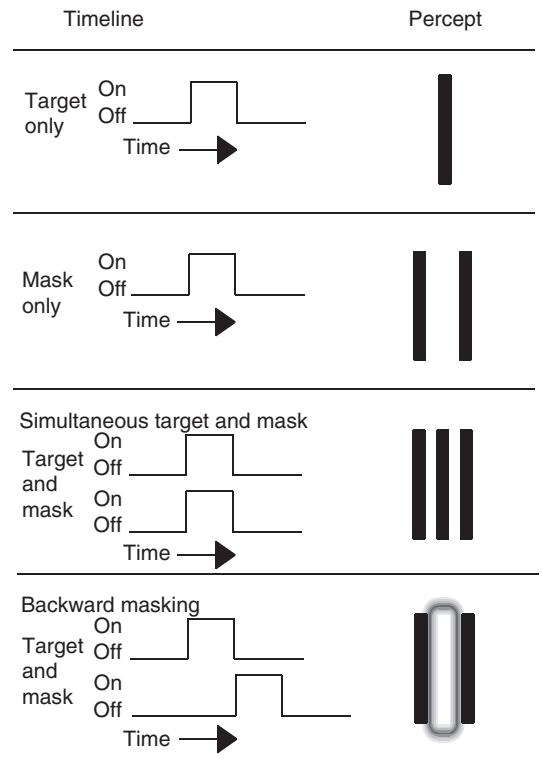

Figure 1 Temporal sequencing of stimuli in backward visual masking. A target and/or mask presented alone, or in simultaneous combination, is visible. When the mask is presented after the target with the appropriate timing (backward masking), the target is rendered invisible. Visual masking illusions provide a powerful tool for investigating the neural underpinnings of visibility and invisibility.

We propose that, to test for explicit processing in neural tissue, one should use a visual illusion with two modes of operation: one mode to ensure that the tissue processes the stimulus explicitly, and one mode to test the correlation to awareness. In visual masking, the monoptic mode establishes that the neural tissue processes masking stimuli explicitly, and then the dichoptic mode can be used to probe the NCC.

The third strategy involves controlling for the effects of attention when designing experiments to isolate the NCC. Attention is a process in which the magnitude of neural activity is either enhanced or suppressed by high-level cognitive mechanisms. Attention may increase or decrease the likelihood of awareness of a given visual stimulus. But attention is a process distinct from awareness itself, as shown by Phillip Merikle and others. For instance, low-level bottom-up highly salient stimuli (such as flickering lights) can lead to awareness and draw attention, even when the subject is actively attending to some other task, or not attending to anything (i.e., when the subject is asleep). Thus, attention can manipulate awareness, but the opposite is also true. This double-dissociation suggests that the two

processes are mediated by separate brain circuits. It follows that in experiments to isolate the NCC, if the subject is conducting a task that requires attention to the stimulus of interest, then attentional and awareness processes may be confounded. Therefore, experiments to isolate the NCC should control for the effects of attention. If experimental manipulation of attentional state affects the magnitude of neural response, then the neural mechanism of interest may not be related to awareness, but instead to attention.

Therefore, we add the following three criteria to Parker and Newsome's list:

8. The candidate neurons should be tested with an illusion that allows dissociation between the physical stimulus and its perception. If the candidate set of neurons is capable of maintaining awareness, the neural responses should match the subjective percept, rather than the objective physical reality of the stimulus.
9. The candidate neurons must explicitly process the type of information or stimulus used to test them.
10. The responses of the neurons, and of the perceiving subject, should be measured with experimental controls for the effect of attention.

Techniques to Study the Neurophysiology of Consciousness

The search for the NCC requires the localization of circuits in the brain that are sufficient to maintain awareness. To this end, brain areas have been sought within the ascending visual hierarchy that correlate, or more importantly, fail to correlate, with visual perception. For instance, the circuits of the brain that are critical to the visibility of targets must be circuits whose activity is suppressed during target invisibility due to visual masking. A corollary: if we identify circuits in which the target response is not suppressed during masking, we can rule out those circuits as sufficient to maintaining visual awareness. This section discusses the techniques used to identify parts of the brain whose activity correlates with visibility and visual awareness.

Table 1 compiles some of the neurophysiological techniques used in awareness research, as well as their strengths and weaknesses. All of these techniques may be paired with psychophysical measures of awareness. No single technique is perfect, but the combination of single-unit electrophysiology with anatomical analysis, fMRI, macroscopic optical recording, and two-photon laser scanning microscopy (2PLSM) may be most powerful to discover the cortical circuits that maintain visual awareness *in vivo*.

Single-Unit Recordings Accompanied by Anatomical Analysis

David Hubel and Torsten Wiesel championed the combined use of single-unit recordings and anatomical staining techniques to achieve high-resolution cortical maps. The tungsten microelectrode, invented by Hubel, was similar to the electrodes used in humans by neurosurgeons such as Wilder Penfield and Benjamin Libet, but it was about 100–1000 times smaller at the tip, so it could be used to discover meaningful, if limited (on the scale of tens of microns), functional patterns within brain areas. Hubel and Wiesel combined electrode penetrations that were angled perpendicularly, obliquely, and tangentially to the cortical surface, with histological staining techniques to reconstruct the electrode tracks in postmortem tissue. Clever application of these techniques led to the first model (on the scale of tens to hundreds of microns) of the two-dimensional pattern of visual function across the surface of the primary visual cortex. They discerned functional domains: repeating patterns of physiological circuits across the surface of the cortex, such as ocular dominance and stimulus orientation columns. They also determined general principles of structure and function of the cortical layers: this added a third dimension to their functional anatomical circuit model.

Ultimately, Hubel and Wiesel found that mapping the three-dimensional cortex with a one-dimensional device was too limiting. By combining their recordings with anatomical staining techniques, such as the Nauta method, radioactive 2-deoxyglucose uptake, horseradish peroxidase staining, and radioactive amino acid pathway tracing, they were able to better relate their electrode tracks to the direct postmortem analysis of cortical and subcortical maps.

Modern versions of these methods are used today to determine the physiological underpinnings of visibility and awareness in anesthetized and awake monkeys (Figure 2) and humans.

Functional Magnetic Resonance Imaging

fMRI is the magnetic imaging of blood flow across differential behavioral conditions. Magnetic resonance imaging (MRI) scanners allow the radiologist to view the density of protons in the various tissues of the body in three-dimensional space, which essentially results in a very high resolution X-ray like image of the internal structures of the body. Blood is one of the tissues that can be imaged in the MR scanner. The magnetic properties of blood change as a function of neural activity: neural activity causes neurons to use more oxygen from the blood, and oxygenated blood has different magnetic properties

Table 1 Summary of neurophysiological techniques to study the NCC

Techniques	*Noninvasive*	*Awake*	*High temporal resolution*	*In vivo application*	*Directly visualizes macroscopic maps*	*High spatial resolution*	*Provides 3-D anatomical reconstruction*	*Provides subcellular anatomical reconstruction*	*Samples from deep structures*	*Provides laminar mapping*
Anatomical staining					X	X	X	X	X	X
Radiography					X	X	X		X	X
Extracellular metal electrodes		X	X	X					X	X
Intracellular glass microelectrodes		X	X	X				X	X	
EEG/evoked potentials	X	X	X	X					X	
fMRI	X	X		X			X		X	
Macroscopic optical recording		X		X	X	X				
2PLSM		X	X	X		X	X	X		X

Courtesy of the Barrow Neurological Institute.

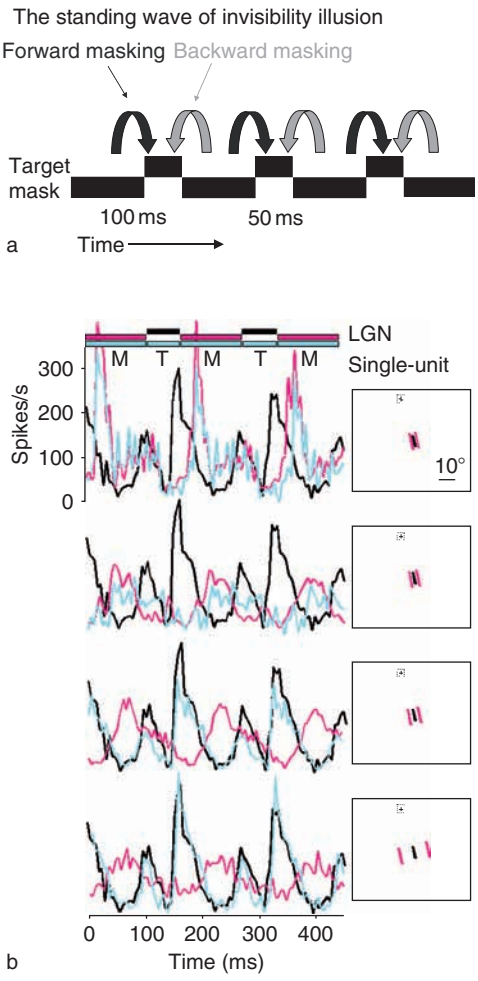

Figure 2 The electrophysiology of visual masking. (a) The standing wave of invisibility (SWI): a temporal sequence of target and mask presentations that results in very powerful masking of the target. (b) A single-neuron recording from the lateral geniculate nucleus (LGN) of a monkey. The stimulus presentation layout on the monkey's computer monitor is illustrated in the inset of each row. The cross is the monkey's fixation point. The target (T) is centered over the receptive field of the LGN neuron and color-coded in black (actual presentation color was white on a black background). The mask (M) is color-coded in purple (actual color same as target) and is presented at four different distances from the target (one distance per row). The physiological traces are represented in black (when the target was presented alone), purple (when the mask was presented alone), and blue (when the target and mask were presented together in the SWI condition). Notice that when the mask is near the target, the target response during the SWI condition (blue) is lower than the target response when presented alone; this is the neural correlate of visual masking. As the mask is moved away from the target, the target response during the SWI condition recovers its magnitude, just as the target becomes more visible perceptually. Reproduced from Macknik SL, Martinez-Conde S, and Haglund MM (2000) The role of spatiotemporal edges invisibility and visual masking. *Proceedings of the National Academy of Sciences of the United States of America* 97: 7556–7560, with permission from the National Academy of Sciences of the United States of America.

than deoxygenated blood. Therefore, functional MR imaging of blood flow (fMRI) is particularly amenable to magnetic imaging analysis. The basis is the following: iron (deoxygenated hemoglobin) is magnetic, whereas rust (oxygenated hemoglobin) is not magnetic. Thus, by isolating the parts of the brain in which blood changes its magnetic properties in correlation to a task, we can infer where the neurons of the brain have responded functionally. It was this line of reasoning that led Ogawa and colleagues to discover that BOLD signal can be used to map brain function noninvasively.

fMRI has sparked a major revolution in the discovery and mapping of human cortical brain areas. However, the method has relatively low spatial resolution (on the order of millimeters, rather than the micron resolution we enjoy with anatomical staining techniques). It also has low temporal resolution (the hemodynamic neural response function occurs over seconds, rather than the microsecond resolution we enjoy with electrophysiology).

Figure 3 shows a map of many of the retinotopic visual areas that may be relevant to maintaining visual awareness. These visual areas can be probed with illusory stimuli, such as visual masking stimuli, to determine whether BOLD activity correlates with the subject's perception (Figure 4).

Macroscopic Optical Imaging

Lack of oxygen not only changes blood from nonmagnetic to magnetic, but also turns its color from red (rusty) to bluish (iron). The regions of the brain's surface that are active during a given task remove more oxygen from the blood, and so these regions appear bluer than the regions where neurons are less active. Gary Blasdel and Guy Salama discovered that when red light was shone onto the exposed surface of the cortex, more red photons were absorbed by the active regions, as blue blood absorbs more red light than red blood. This decrease in reflectance can be measured with a video camera to map cortical function.

fMRI and macroscopic optical recording of intrinsic signals have the same physical substrate (blood oxygenation level), and therefore the same slow temporal resolution. However, macroscopic optical recording enjoys a much higher spatial resolution than fMRI: the resolution of the capillary network just below the brain's surface is tens to hundreds of microns, depending on the tissue.

Figure 5 illustrates how macroscopic optical maps may be used to examine area V1 responses to standard visual stimuli. Figure 6 shows an optical image

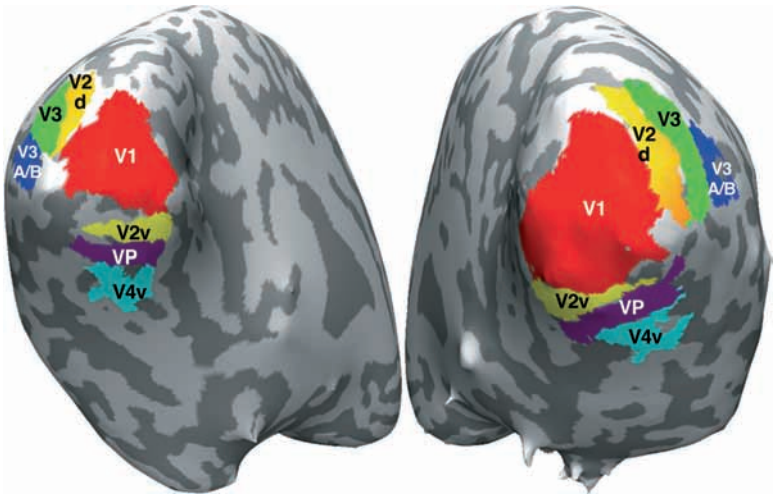

Figure 3 Example of retinotopy mapping from a human subject. Visual areas delineated by retinotopic mapping analysis are indicated in different colors. Reproduced from Tse Pu, Martinez-Conde S, Schlegel AA, and Macknik SL (2005) Visibility, visual awareness, and visual masking of simple unattended targets are confined to areas in the occipital cortex beyond human V1/V2. *Proceedings of the National Academy of Sciences of the United States of America* 102: 17178–17183, with permission from the National Academy of Sciences of the United States of America.

of V1 responses to visual masking illusions, such as the standing wave of invisibility.

Two-Photon Laser Scanning Microscopy

One of the great problems with standard light microscopy is that the light source is inevitably brighter than the tissue of interest, which drastically decreases the quality of the imaging. To make matters worse, the images are also contaminated by light coming from sources outside of the plane of focus. The invention of fluorescence microscopy was a major leap forward for microscopic analysis: it allowed for the illumination of tissue with one color of light (the excitation wavelength) while imaging only the fluoresced photons of a different color (the emission wavelength). Thus, fluorescence microscopes filter out the photons outside the emission wavelength, which improves the quality of the imaging greatly. The main problem with fluorescent imaging in biological tissues is that fluorescence, by its quantum mechanical nature, turns a high-energy (bluish) photon into a lower energy (reddish) photon. Because brain tissue tends to absorb blue photons, one must use extremely high-powered light sources to image tissues more than ~100 microns deep; this carries the risk of burning the brain.

In 1990, Denk et al. used two-photon fluorescence excitation to work around these problems. The idea behind the two-photon quantum effect is that a fluorescent molecule, which has an excitation wavelength of, say, 400 nm, will fluoresce a photon when excited by two photons of 800 nm (which each have exactly half the energy of a single 400 nm photon). So the same fluorescent dye will fluoresce a greenish emission photon after receiving a single bluish (400 nm) excitation photon, or two simultaneous infrared (800 nm) photons. The advantage is that the brain is relatively transparent to infrared light (as compared with blue light) and so one can image much deeper and with a much lower power light source (i.e., infrared) than when using a blue light source. The illumination method is moreover inherently focused on a given focal plane because the two infrared photons must land on the same exact fluorescent dye molecule within femptoseconds of each other to result in the emission of a green photon. Any excitation (infrared) light that is scattered by irrelevant tissue outside the focal point does not cause noise in the imaging (it is simply lost), and so only the single point of focus (in three dimensions) of the laser will generate fluorescence. Because of these features, 2PLSM solves all of the problems above: it combines fluorescence (which allows one to filter out noise from the light source) with the ability to use an infrared light source to image deeply within the cortex (on the order of 1 mm), and furthermore to maintain an optical resolution that is roughly the size of a single photon.

2PLSM has not yet been used to examine circuits that are critical to visibility or consciousness. However, it is only a matter of time before this method is employed to determine the microscopic neural interactions that are the fundamental basis of awareness.

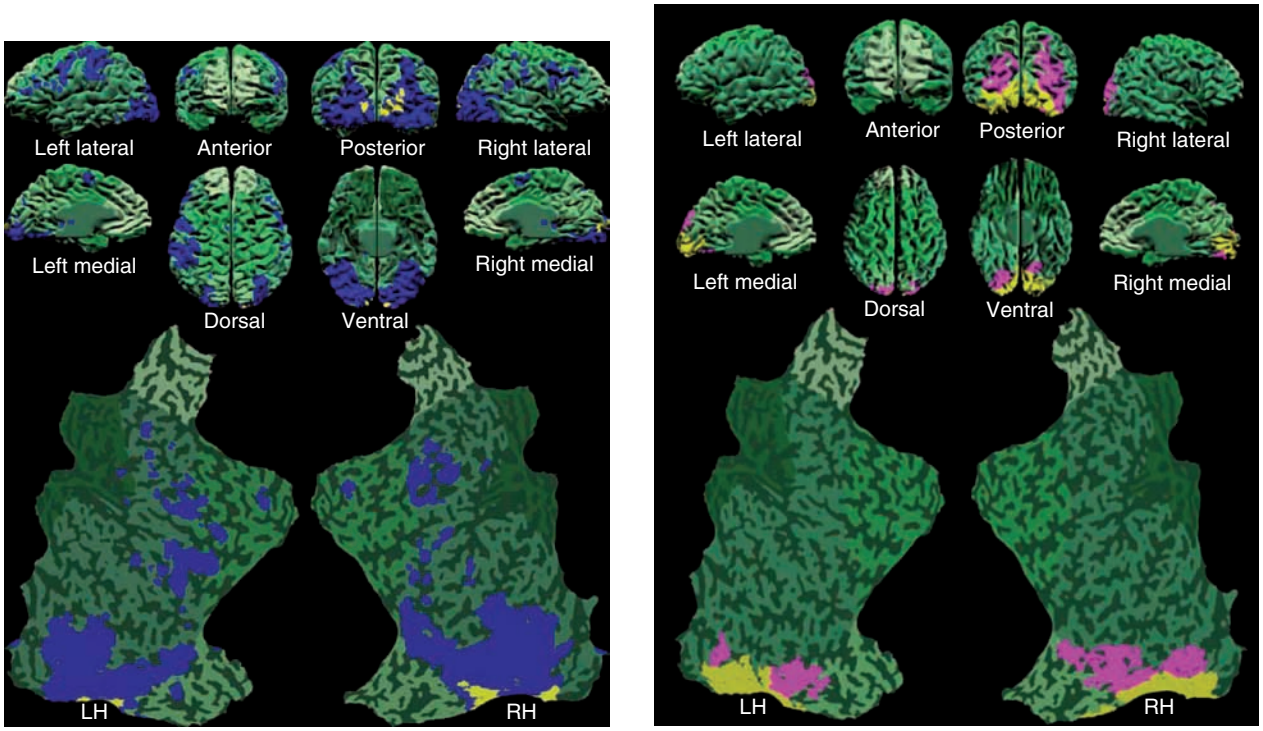

Figure 4 Human cortical responses to visible and invisible stimuli. Green shading represents the various cortical brain regions. (a) The union of responses from 17 normal human subjects viewing target stimuli alone (blue). These blue areas represent the areas of the brain that encode simple unattended target stimuli. The yellow areas represent the cortical modulation during visual masking of targets. Therefore, the yellow areas represent the portions of the occipital lobe that encode target visibility. These results suggest that awareness of simple attended targets is processed within the occipital cortex. (b) Further analysis of the occipital lobe's role in awareness of visibility, in an individual subject. Here the yellow regions represent the occipital areas that correlate with target perception during monoptic visual masking, but not during dichoptic visual masking. Since these areas explicitly process visual masking but fail to correlate with dichoptic masking, they can be ruled out as sufficient for maintaining the awareness of targets. The pink areas correlate with target perception during both monoptic and dichoptic masking and thus cannot be ruled out. The results from (a, b) combined suggest that awareness of simple unattended targets is maintained by circuits downstream of area V2 but nevertheless confined to the occipital lobe. Reproduced from Tse Pu, Martinez-Conde S, Schlegel AA, and Macknik SL (2005) Visibility, visual awareness, and visual masking of simple unattended targets are confined to areas in the occipital cortex beyond human V1/V2. *Proceedings of the National Academy of Sciences of the United States of America* 102: 17178–17183, with permission from the National Academy of Sciences of the United States of America.

Current Findings on Visibility and Visual Awareness

We have begun to establish the minimal set of conditions necessary to maintain the awareness of simple unattended stimuli. We have determined that the target's spatiotemporal edges are crucial to the target's perceptual visibility. Visibility at the spatiotemporal edges is encoded by transient bursts of spikes in the early visual system. If these bursts are inhibited, visibility fails. The specific circuits that maintain visibility are not yet known, but we propose that lateral inhibition plays a critical role in sculpting our perception of visibility, both by causing interactions between stimuli positioned across space and by shaping the responses to stimuli across time.

We have tested several retinotopic visual areas of the early visual cortex for their correlation to visibility, using visual masking illusions. We first confirmed that the perceptual strength of visual masking is equivalent in monoptic and dichoptic conditions. This established that neurons that may potentially maintain visual awareness should correlate equally well to both monoptic and dichoptic masking. Next, we established that all retinotopic areas tested (lateral geniculate nucleus (LGN), V1, V2, V3, and V4) explicitly processed visual masking with monoptic visual masking illusions. We retested with dichoptic masking and found that activity in the LGN, V1, and V2 did not match perception. Therefore, we were able to rule out these areas as part of the NCC. Only V4 responses (and, weakly, area V3

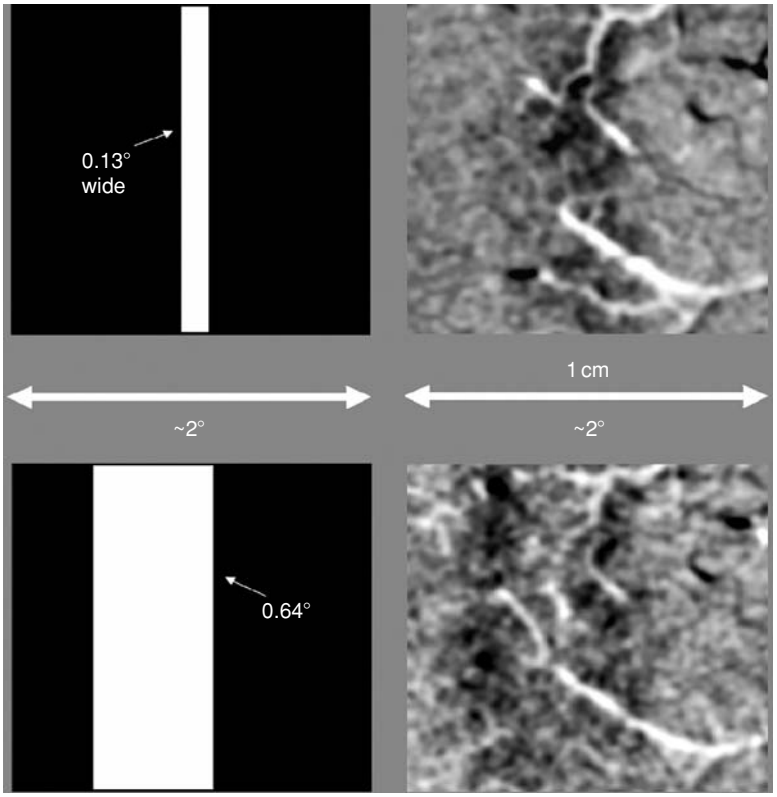

Figure 5 The optical image of a flickering bar. The left column represents the layout of the stimuli used to stimulate the cortical window imaged in the right column. Top: Image of the area V1 cortical intrinsic signal generated by a flickering bar 50 ms on and 100 ms off, 0.13° wide with an orientation of 132°. The imaged patch is 1 cm^2 and was approximately 10–12° below and to the left of the foveal representation and subtended about 4° of visual angle (as measured with microelectrode penetrations at each edge of the image) at the anterior-medial border of the operculum. The vertical meridian is parallel to the lower edge of the image; the fovea is to the right. Bottom: Image of the intrinsic signal generated by a flickering bar 0.64° wide in the same piece of cortex (the center of the bar was also shifted approximately 0.29° away from the fovea). Notice that the widened optical image of the bar has shifted in position and split into two edges, showing that edges generate much stronger signals that the interiors of objects at the level of V1. Reproduced from Macknik SL, Martinez-Conde S, and Haglund MM (2000) The role of spatiotemporal edges invisibility and visual masking. *Proceedings of the National Academy of Sciences of the United States of America* 97: 7556–7560, with permission from the National Academy of Sciences of the United States of America.

responses) matched perception, making area V3 the lower boundary in the visual hierarchy for the processing of visual awareness. Moreover, we found that only areas within the occipital lobe correlate with awareness of simple unattended targets. This finding provided us with the first experimental evidence of an upper boundary (the edge of the occipital lobe) for the processing and maintenance of visual awareness. These results localized visual awareness of simple unattended targets to a relatively small part of the brain, between area V3 and the edge of the occipital lobe. Future research will more precisely localize the specific circuits within this region that maintain awareness.

The results support Jackendoff's theory that consciousness is an intermediate-level process, and not a low- or high-level process. However, it remains possible (contrary to Jackendoff's proposal) that higher brain areas, such as the frontal lobes, maintain awareness of higher cognitive functions and that only sensory awareness is maintained at intermediate levels.

Future Challenges

The field of the neurobiology of consciousness is at its very incipient stage. We and others have begun to narrow down the potential brain regions or circuits that may maintain various types of sensory awareness. However, we are far from having a solid understanding of the NCC. Once the NCC are localized, a bigger challenge will be to determine the specific nature of the underlying circuits that maintain awareness. This is further complicated by the possibility that there may be multiple different circuits that maintain awareness of different types of events.

On the bright side, the field is converging on a set of techniques and methods that can be used to test neural tissue at multiple levels of resolution, both

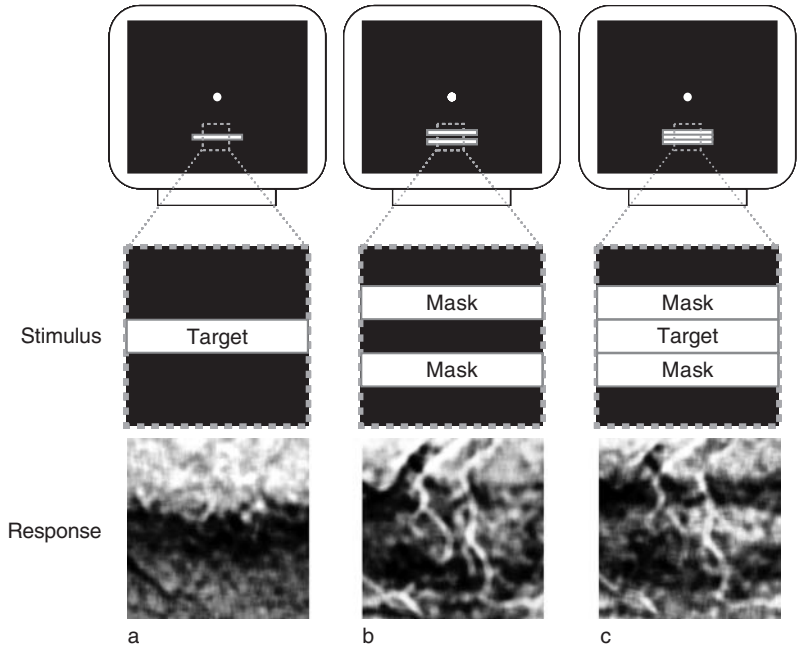

Figure 6 Optical imaging of the standing wave of invisibility (SWI). (a) A visual target with a width of 0.12° and the correlated optical image. (b) Two masking stimuli presented without a target and the correlated response. (c) The SWI stimulus (target and masks were never presented on the display at the same time) and the correlated response. Notice that the image of the target in (c) is now missing, compared to the response seen in (a), despite the fact that both targets were physically identical. Reproduced from Macknik SL and Haglund MM (1999) Optical images of visible and invisible percepts in the primary visual cortex of primates. *Proceedings of the National Academy of Sciences of the United States of America* 96: 15208–15210, with permission from the National Academy of Sciences of the United States of America.

invasively and noninvasively. Once these techniques are fully developed, long hours in the lab and clever experimental design will accomplish much of the remaining work.

See also: Attentional Networks; Awareness: Functional Imaging; Blindsight: Residual Vision; Coma and Other Pathological Disorders of Consciousness; Consciousness: Theories and Models; Consciousness: Theoretical and Computational Neuroscience; Consciousness: Neural Basis of Conscious Experience; Contextual Interactions in Visual Perception; Contextual Interactions in Visual Processing.

Further Reading

Blasdel GG and Salama G (1986) Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature* 321: 579–585.

Crick FHC (1994) *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Touchstone.

Denk W, Strickler JH, and Webb WW (1990) Two-photon laser scanning fluorescence microscopy. *Science* 248: 73–76.

Hubel DH and Wiesel TN (2005) *Brain and Visual Perception: The Story of a 25-Year Collaboration*. New York: Oxford University Press.

Jackendoff R (1987) *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Koch C (2004) *The Quest for Consciousness: A Neurobiological Approach*. CO: Roberts and Publishers.

Libet B (2006) Reflections on the interaction of the mind and brain. *Progress in Neurobiology* 78: 322–326.

Macknik SL (2006) Visual masking approaches to visual awareness. *Progress in Brain Research* 155: 179–217.

Macknik SL and Haglund MM (1999) Optical images of visible and invisible percepts in the primary visual cortex of primates. *Proceedings of the National Academy of Sciences of the United States of America* 96: 15208–15210.

Macknik SL and Livingstone MS (1998) Neuronal correlates of visibility and invisibility in the primate visual system. *Nature Neuroscience* 1: 144–149.

Macknik SL and Martinez-Conde S (2004) Dichoptic visual masking reveals that early binocular neurons exhibit weak interocular suppression: Implications for binocular vision and visual awareness. *Journal of Cognitive Neuroscience* 16: 1–11.

Macknik SL and Martinez-Conde S (2004) The spatial and temporal effects of lateral inhibitory networks and their relevance to the visibility of spatiotemporal edges. *Neurocomputing* 58–60C: 775–782.

Macknik SL, Martinez-Conde S, and Haglund MM (2000) The role of spatiotemporal edges in visibility and visual masking. *Proceedings of the National Academy of Sciences of the United States of America* 97: 7556–7560.

Malonek D and Grinvald A (1996) Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy – Implications for functional brain mapping. *Science* 272: 551–554.

Martinez-Conde S (2006) Fixational eye movements in normal and pathological vision. *Progress in Brain Research* 154: 153–178.

Martinez-Conde S, Macknik SL, and Hubel DH (2000) Microsaccadic eye movements and firing of single cells in

the striate cortex of macaque monkeys. *Nature Neuroscience* 3: 251–258.

Martinez-Conde S, Macknik SL, and Hubel DH (2002) The function of bursts of spikes during visual fixation in the awake primate lateral geniculate nucleus and primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 99: 13920–13925.

Martinez-Conde S, Macknik SL, and Hubel DH (2004) Moving our eyes to keep the world visible: The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience* 5: 229–240.

Martinez-Conde S, Macknik SL, Troncoso XG, and Dyar TA (2006) Microsaccades counteract visual fading during fixation. *Neuron* 49: 297–305.

Merikle PM, Smile KD, and Eastwood JD (2001) Perception without awareness: Perspectives from cognitive psychology. *Cognition* 79(1–2): 115–134.

Ogawa S, Lee TM, Kay AR, and Tank DW (1990) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America* 87: 9868–9872.

Parker AJ and Newsome WT (1998) Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience* 21: 227–277.

Penfield W and Jasper H (1954) *Epilepsy and the Functional Anatomy of the Human Brain*. London: Churchill.

Tse PU, Martinez-Conde S, Schlegel AA, and Macknik SL (2005) Visibility, visual awareness, and visual masking of simple unattended targets are confined to areas in the occipital cortex beyond human v1/v2. *Proceedings of the National Academy of Sciences of the United States of America* 102: 17178–17183.

Wheatstone C (1838) On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions* 128: 371–394.

Zeki S (2003) The disunity of consciousness. *Trends in Cognitive Science* 7: 214–218.

Relevant Website

http://neuralcorrelate.com – Online Demonstration of the Standing Wave of Invisibility.

Conciousness: Philosophy

G Tononi, University of Wisconsin, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

Consciousness is everything we experience, from a city scene teeming with shapes, colors, and noises to darkness and silence. It has been defined as what abandons us every night when we fall into dreamless sleep and returns the next morning when we wake up. We know that consciousness depends on the functioning of the brain; yet many philosophers have claimed that understanding how the brain produces subjective experience may lie forever beyond the realm of scientific explanation. Scientists often find it safer to dispense with explanations and to explore instead which parts of the brain are important for consciousness and which aspects of neual activity accompany it. For example, clinical observations and classic experiments have shown that the level of consciousness is associated with an activated electroencephalogram. Brain lesions have demonstrated that different brain areas are important for specific contents of consciousness. New experimental tools are refining our appreciation of when and where changes in neural activity are correlated with changes in conscious experience.

Important as it is, however, gathering data on the neural correlates of consciousness is no substitute for a principled explanation. Otherwise, how can we establish whether newborn babies or animals are conscious and to what extent, or whether a conscious artifact can be constructed with nonneural ingredients? And how can we determine, in the absence of verbal reports, whether a person with akinetic mutism – awake with eyes open, but mute, immobile, and unresponsive – is conscious or not, or how much consciousness there is during sleepwalking or psychomotor seizures? A genuine understanding of consciousness requires, in addition to empirical studies, a theoretical analysis.

The Two Problems of Consciousness

Two main problems need to be addressed. The first problem has to do with the necessary and sufficient conditions that determine whether consciousness is present or not. For example, why are changes of neural activity in thalamocortical regions so important for conscious experience, whereas changes in activity in cerebellar circuits are not, given that the number of neurons in the two structures is comparable? Why is consciousness strikingly reduced during deep slow-wave sleep, given that the level of neuronal firing is similar to that of wakefulness?

The second problem has to do with the necessary and sufficient conditions that determine the specific nature of consciousness. For example, what makes the activity of specific cortical areas contribute specific dimensions of conscious experience – auditory cortex to sound, visual cortex to shapes and colors? What aspect of neural organization is responsible for the fact that shapes look the way they do, and different from the way colors appear or pains feel? Solving the first problem means that we would know whether a physical system can generate consciousness and to what extent – the *level* of consciousness. Solving the second problem means that we would know what kind of consciousness is generated – the *content* of consciousness.

Consciousness and the Brain: Facts and Enigmas

Let us briefly consider the first problem by examining some facts about consciousness and the brain and some of the baffling questions they pose:

1. Consciousness is produced by certain parts of the brain and not, or much less, by others. For example, we know that certain portions of the thalamocortical system are essential for consciousness. Other regions of the brain, such as the cerebellum, are not and can be stimulated or lesioned without giving rise to changes in conscious experience. Yet the cerebellum has as many neurons and is every bit as complicated as the thalamocortical system. Why is consciousness associated with some but not with other neural structures?
2. Consciousness is associated with a distributed neural system: there is no single area where it all comes together. For example, different areas of the thalamocortical system each contribute different dimensions to conscious experience, but no single area is solely responsible for consciousness.
3. Consciousness can be split if the thalamocortical system is split. Studies of split brain patients, whose corpus callosum was sectioned for therapeutic reasons, show that each hemisphere has its own, private conscious experience.
4. Within the thalamocortical system, the activity of certain neurons correlates well with conscious experience; that of others does not. For example, during binocular rivalry, the activity of certain visual cortical neurons follows what the subject consciously

perceives, while that of other neurons follows the stimulus, whether the subject is perceiving it or not. What determines whether the firing of neurons within the thalamocortical system contributes directly to consciousness or not?

5. Many neural processes in the thalamocortical system can influence conscious experience yet do not seem to contribute directly to it. For example, what we see and hear depends on elaborate computational processes in the cerebral cortex that are responsible for object recognition, depth perception, and language parsing; yet such processes remain largely unconscious.

6. Neural activity in sensory afferents to the thalamocortical system usually determines what we experience at any given time. However, such neural activity does not appear to contribute directly to conscious experience. For example, although retinal cells can tell light from dark and convey that information to visual cortex, their rapidly shifting firing patterns do not correspond well with what we perceive. Moreover, a person who becomes retinally blind as an adult continues to have vivid visual images and dreams. Why is it that the activity of retinal cells *per se* does not contribute directly to conscious experience, but only indirectly through its action on thalamocortical circuits?

7. Neural processes occurring in brain regions whose inputs and outputs are closely linked to the thalamocortical system, such as the basal ganglia, are important in the production and sequencing of action, thought, and language. Yet such processes do not seem to contribute directly to conscious experience. Why is this so?

8. When first learning a new task, we are conscious of every detail of the task; we are slow, make mistakes, and must make an effort. When we have learned the task well, we become much better, faster, and effortless, but also less aware of it. During learning, many cortical areas are involved, whereas well-learned tasks may employ different neural circuits. Why does automatic performance become less conscious?

9. The mode of neuronal firing in the thalamocortical system is important for consciousness. For example, cortical neurons fire almost as much during deep slow-wave sleep as during wakefulness, but the level of consciousness is much reduced in the former condition. Similarly, in absence seizures, neural firing is high and synchronous; yet consciousness is seemingly lost. Why is this the case?

10. The firing of the same cortical neurons may correlate with consciousness at certain times, but not at others. For example, multiunit recordings in the primary visual cortex of monkeys show that, after a stimulus is presented, the firing rate of many neurons increases regardless of whether the animal reports seeing a figure or not. After 80–100 ms, however, their discharge accurately predicts the conscious detection of the figure. What determines when the firing of the same cortical neurons contributes to conscious experience and when it does not?

Many more facts and puzzles could be added to this list. This state of affairs is not unlike the one biologists faced when, knowing a great deal about similarities and differences between species, fossil remains, and breeding practices, they still lacked a theory of how evolution might occur. What was needed, then as now, were not just more facts, but a theoretical framework that could make sense of them. Unfortunately, theoretical approaches that try to provide a coherent explanation for some of the basic facts about consciousness and the brain are few and far between. Here, in order to offer a tentative but at least unified perspective on the issues that need to be addressed, we discuss a theoretical approach, according to which consciousness is related to the brain's ability to integrate information. This theory comprises (a) an examination of phenomenology indicating that consciousness has to do with information integration, (b) a definition of what information integration is and how it can be measured, (c) an attempt at accounting for basic facts about consciousness and the brain, and (d) some corollaries and predictions.

Phenomenology: Consciousness as Information Integration

In this view, perhaps the most important thing to realize about consciousness is that, when one experiences a particular conscious state – for example, the one experienced when reading *this particular phrase* here and now – each of us is gaining access to an extraordinarily large amount of information. This information has nothing to do with how many letters or words we can take in at any one time, which is a very small number. Instead, the occurrence of a particular conscious state is extraordinarily informative because of the very large number of alternative conscious states that it rules out. Just think of all possible written phrases you could read, multiply them by the number of possible fonts, ink colors, sizes in which you could read them; then think of the same phrases spoken aloud, or read and spoken, or think further of all other possible visual scenes you might experience, multiplied by all possible sounds you might hear at the same time, by all possible moods you might be in, and so on *ad libitum*.

The point is simply this: every time we experience a particular conscious state out of such a huge repertoire of possible conscious states, we gain access to a correspondingly large amount of information. This conclusion is in line with the classical definition of information as reduction of uncertainty among a number of alternatives. For example, tossing a fair coin and obtaining heads corresponds to $\log_2(2) = 1$ bit of information, because there are just two alternatives; throwing a fair die yields $\log_2(6) = 2.59$ bits of information, because there are six equally likely possibilities. Similarly, the information generated by the occurrence of a particular conscious state lies in the large number of different conscious states that *could potentially* have been experienced but were not. Although no attempt has been made to estimate the size of the repertoire of conscious states available to a human being, it is clear that such repertoire must be extraordinarily large, and so is the information yielded by entering a particular conscious state out of this repertoire. This point is so simple that its importance has been overlooked.

Another key aspect of this theory is that the information associated with the occurrence of a conscious state is information from the perspective of an *integrated* system. When each of us experiences a particular conscious state, that conscious state is experienced as an integrated whole – it cannot be subdivided into independent components, that is, components that are experienced independently. For example, the conscious experience of *this particular phrase* cannot be experienced as subdivided into, say, the conscious experience of how the words look independently of the conscious experience of how they sound in our mind. Similarly, we cannot experience visual shapes independently of their color, or perceive the left half of the visual field of view independently of the right half. Furthermore, we cannot simultaneously experience both aspects of an ambiguous figure, nor can we simultaneously think two different thoughts. If we could, this would be tantamount to having two separate 'centers' of consciousness. Separate centers of consciousness exist, of course, but each is a different person with a different brain (or a different hemisphere of a split brain). Indeed, if two people each experience a different conscious state in two separate rooms, the information to which each person gains access depends only on the repertoire of conscious states available to that person. A superordinate consciousness associated with the joint states of two different people considered together is an absurd notion, because the two people are causally independent and do not constitute an integrated system.

Finally, it is important to appreciate that every conscious experience takes some time. Studies of how a percept is progressively specified and stabilized – a process called *microgenesis* – indicate that it takes up to 100–200 ms to develop a fully formed sensory experience and that the surfacing of a conscious thought may take even longer. Other evidence indicates that a single conscious moment cannot extend beyond 2–3 s. Although it is arguable whether conscious experience unfolds more like a series of discrete snapshots or a continuous flow, its time scale is certainly comprised between these lower and upper limits. Thus a phenomenological analysis indicates that consciousness has to do with the ability to integrate a large amount of information over a short period of time.

Measuring Information Integration

If consciousness is integrated information, then a physical system should be able to generate consciousness to the extent that it can rapidly enter any of a large number of available states (information); yet it cannot be decomposed into a collection of causally independent subsystems (integration). How can one identify such an integrated system, and how can one measure its repertoire of available states?

At first sight, it might seem that all one needs to do is choose a system, for example, the brain, and measure the repertoire of states that are available to it with their probability. One could then calculate the information associated with the occurrence of each brain state, just as one can measure the information associated with tossing a coin or a die, by using the entropy function, that is, the weighted sum of the logarithm of the probability of system states: $H = -\Sigma p(s)\log_2 p(s)$. Measuring the available repertoire would easily account for why a seemingly similar task can be performed unconsciously (or nearly so) by a simple device and consciously by a human being. For example, when a retinal cell, or even a photodiode – a simple semiconductor device that changes its electrical resistance depending on the illumination – detects complete darkness, it generates a minimal amount of information, since it can only discriminate between darkness and light. When we consciously detect complete darkness, however, we perform a discrimination that is immensely more informative: we are not just ruling out light, but an extraordinary number of other possible states of affairs, including every possible frame of every possible movie, every possible sound, and every possible combination of them.

Measuring information this way, however, is insufficient, because it is completely insensitive to whether the information is available to an integrated system. To give a simple example, consider a collection of 1 million photodiodes constituting the sensor chip of a digital camera. From the perspective of an external

observer, such a chip can certainly enter a very large number of different states, as could easily be demonstrated by presenting it with all possible input signals. However, because of the absence of any physical interaction among the photodiodes, the chip as such cannot integrate any information: the state of each element is causally independent of that of other elements. In other words, what we have is 1 million photodiodes with a repertoire of two states each, rather than a single integrated system with a repertoire of $2^{1\,000\,000}$ states. Thus to measure integrated information, it is essential to know whether a set of elements constitutes a causally integrated system or can be broken down into a number of independent or quasi-independent subsets among which no information can be integrated.

To see how one can achieve this goal, consider an extremely simplified neural system composed of a set of elements. Each element could represent, for instance, a group of locally interconnected neurons that share inputs and outputs, such as a cortical minicolumn. We could further assume that each element can go through discrete activity states, corresponding to different firing levels, each of which lasts for a few hundred milliseconds. Finally, for the present purposes, let us imagine that the system is disconnected from external inputs, just as the brain is disconnected from the environment when it is dreaming.

Effective Information

Consider now a subset S of elements taken from the system. We want to measure the information generated when S enters a particular state out of its repertoire, but only to the extent that such information can be integrated (i.e., it can be the result of causal interactions within the system). To do so, we start by dividing S into two complementary parts A and B, give maximum entropy to the outputs from A (i.e., substitute its elements with independent noise sources) and determine the entropy of the resulting responses of B. Specifically, we define the *effective information* between A and B as $EI(A \rightarrow B) = MI(A^{Hmax} \rightarrow B)$. Here $MI(A;B) = H(A) + H(B) - H(AB)$ stands for mutual information, the standard measure of the entropy or information shared between a source (A) and a target (B). Since A is substituted by independent noise sources, the entropy that B shares with A is due to causal effects of A on B. In neural terms, we try out all possible combinations of firing patterns as outputs from A and establish how differentiated is the repertoire of firing patterns they produce in B. Thus if the connections between A and B are strong and specialized, different outputs from A will produce different firing patterns in B, and $EI(A \rightarrow B)$ will be high. On the other hand, if the connections between A and B are such that different outputs from A produce scarce effects, or if the effect is always the same, then $EI(A \rightarrow B)$ will be low or zero. Note that, unlike measures of statistical dependence, effective information measures causal interactions and requires perturbing the outputs from A. Moreover, by enforcing independent noise sources in A, effective information measures all possible effects of A on B, not just those that are observed if the system were left to itself. Also, $EI(A \rightarrow B)$ and $EI(B \rightarrow A)$ will generally differ; that is, effective information is not symmetric. For a given bipartition of a subset, the sum of the effective information for both directions is indicated as $EI(A \rightleftharpoons B) = EI(A \rightarrow B) + EI(B \rightarrow A)$.

Information Integration

Based on the notion of effective information for a bipartition, we can assess how much information can be integrated within a system of elements. To this end, it is important to realize that a subset S of elements cannot integrate any information if there is a way to partition S in two complementary parts A and B such that $EI(A \rightleftharpoons B) = 0$. In such a case we would obviously be dealing with two (or more) causally independent subsets, rather than with a single, integrated subset. More generally, to measure how much information can be integrated within a subset S, we should search for the bipartition(s) for which $EI(A \rightleftharpoons B)$ reaches a minimum. Since $EI(A \rightleftharpoons B)$ is necessarily bounded by the maximum entropy available to A or B, to be comparable over bipartitions, $\min\{EI(A \rightleftharpoons B)\}$ should be normalized by $\min\{H^{max}(A), H^{max}(B)\}$. Thus the *minimum information bipartition* of subset S, or MIB(S), is its bipartition for which the normalized effective information reaches a minimum, corresponding to $\min\{EI(A \rightleftharpoons B)/(\min\{H^{max}(A), H^{max}(B)\})\}$ for all A. The capacity for *information integration* of subset S, or $\Phi(S)$, is simply the value of $EI(A \rightleftharpoons B)$ for the minimum information bipartition: $\Phi(S) = EI(MIB(S))$. The symbol Φ is meant to indicate that the information (the 'I') can be integrated within a single entity (the 'O'). This quantity is also called ^{MIB}*complexity*, for minimum information bipartition complexity. In most cases, the bipartitions for which the normalized value of EI will be at a minimum, everything else being equal, will be bipartitions that cut the system in two halves, that is, midpartitions.

Complexes

If $\Phi(S)$ is calculated for every possible subset S of a system, we can establish which subsets are actually capable of integrating information, and how much of it. Consider thus every possible subset S of *m* elements out of the *n* elements of a system, starting

with subsets of two elements ($m = 2$) and ending with a subset corresponding to the entire system ($m = n$). A subset S with $\Phi > 0$ is called a *complex* if it is not included within a larger subset having higher Φ. For a given system, the complex with the maximum value of $\Phi(S)$ is called the *maximum* or *main complex*, where the maximum is taken over all combinations of $m > 1$ out of n elements of the system. Note that a complex can be causally connected to elements that are not part of it. The elements of a complex that receive inputs from or provide outputs to other elements not part of that complex are called *ports in* and *ports out*, respectively. In summary, a system can be analyzed to identify its complexes – those subsets of elements that integrate information – and each complex will have an associated value of Φ – the amount of information it integrates.

Consciousness, Information Integration, and the Brain

If consciousness has to do with the ability to integrate information, and if such ability can be measured as suggested previously, it follows that a physical system will have consciousness to the extent that it constitutes a complex having high values of Φ. How do these concepts apply to the brain, and can they account, at least in principle, for some of the facts and puzzles listed previously? Can they shed any light, for instance, on why the thalamocortical system is essential for consciousness whereas the cerebellum is not, or on why consciousness is reduced during slow-wave sleep?

Although performing an exhaustive analysis of complexes on the brain itself is not feasible at present, it is nevertheless possible to obtain some insights by considering computer simulations that capture basic aspects of brain organization. To be sure, the effective information between subsets of neurons is heavily constrained by biological factors affecting maximum firing rates, firing duration, and postsynaptic excitability. Everything else being equal, however, the single most important determinant of a network ability to integrate information is its anatomic organization. Connection patterns can determine whether a system is highly modular (i.e., splits into small, nearly independent complexes) or whether it forms a single, large complex. Moreover, complexes having the same number of elements and connections may have widely different values of Φ depending on how the connections are arranged. It is worth considering a few representative instances.

Thalamocortical System

Using small networks of linear elements, it is possible to show that Φ is maximized by having cach element develop a different connection pattern with the rest of the complex (functional specialization) while ensuring that, no matter how the system is divided in two, the two parts are well connected (functional integration). The thalamocortical system – the brain structure most directly associated with consciousness – appears to be an excellent candidate for the rapid integration of information. It is composed of a large number of widely distributed elements that are functionally specialized. At the same time, these elements are intricately interconnected and exhibit short- and long-range synchronization, a hallmark of functional integration. Indeed, computer simulations suggest that the connectivity of the visual thalamocortical system, characterized by many local groups of heavily connected neurons, forward connections that increase functional specialization, back connections that can distribute information widely, and a large number of entangled short loops, is well suited to generating high values of Φ. It is not yet clear to what extent the connectivity may differ in other cortical regions, or whether certain cortical layers and cell types may play a privileged role in integrating information among distant areas. However, it is clear that the organization of the thalamocortical system is well suited to forming a complex capable of integrating a large amount of information.

Cerebellum

This brain region contains probably more neurons and as many connections as the cerebral cortex, receives mapped inputs from the environment, and controls several outputs. However, in striking contrast to the thalamocortical system, lesions or ablations indicate that the direct contribution of the cerebellum to conscious experience is minimal. As suggested by computer simulations, the reason may be that the organization of synaptic connections within the cerebellum is radically different from that of the thalamocortical system and is not well suited to the integration of information. Specifically, the organization of the connections is such that individual patches of cerebellar cortex tend to be activated independently of one another, with little interaction between distant patches. This suggests that cerebellar connections may not be organized so as to generate a large complex of high Φ, but rather very many small complexes each with a low value of Φ.

Split Brain

When the corpus callosum is sectioned, consciousness is split. The level of consciousness of the dominant hemisphere and most of its contents are not altered severely after the operation. The nondominant hemisphere also appears to be conscious, although it

loses some important abilities. Some information, for example, emotional arousal, seems to be shared across the hemispheres, probably due to subcortical common inputs. As illustrated by simple computer models, a 'callosal' cut produces, out of a large complex corresponding to the connected thalamocortical system, two separate complexes. However, because there is great redundancy between the two hemispheres, their Φ value is not greatly reduced compared with when they formed a single complex. The analysis of complexes also identifies a complex corresponding to both hemispheres and their subcortical common inputs, although with much lower Φ values. That is, there is a sense in which the two hemispheres still form an integrated entity, but the information they share is minimal.

Cortical Input and Output Systems

Why is it that retinal cells do not appear to contribute directly to visual consciousness, given that they evidently influence it by providing powerful signals to the thalamocortical system? According to computer simulations, input circuits providing powerful inputs to a complex add nothing to the information it can integrate if their effects are entirely accounted for by ports in, that is, by the elements of the complex to which they connect. Similar considerations apply to output circuits triggered by ports out.

Basal Ganglia and Cortico-Subcortical Loops

The basal ganglia are large neural structures that contain many circuits arranged in parallel, some implicated in motor and oculomotor control, others, such as the dorsolateral prefrontal circuit, in cognitive functions, and others, such as the lateral orbitofrontal and anterior cingulate circuits, in social behavior, motivation, and emotion. Further specialization appears to occur within each of these broad domains. Each basal ganglia circuit originates in layer V of the cortex, and through a last step in the thalamus, returns to the cortex, not far from where the circuit started. Similarly arranged cortico-cerebello-thalamo-cortical loops also exist. Are such subcortical loops that are connected at both ends to the thalamocortical system necessarily part of the main thalamocortical complex? According to computer models, adding a loop to a large complex does not necessarily produce a larger complex with higher Φ. Instead, the elements of the main complex and those of the loop form a joint complex that can only integrate the limited amount of information exchanged within the loop. Thus subcortical loops could implement specialized subroutines that influence the states of the main thalamocortical complex, even if the neural interactions within the loop remain informationally insulated from those of the main complex. Informationally insulated cortico-subcortical loops may thus constitute substrates for unconscious processes that influence conscious experience.

Slow-Wave Sleep

Even if the anatomic connectivity is the same, a change in a few key parameters may radically alter the Φ value of a complex, with corresponding consequences on consciousness. When the brain enters the deepest stages of slow-wave sleep, consciousness is substantially reduced; yet neurons fire almost as intensely as in wakefulness. However, because of certain conductances modified by changes in neuromodulatory levels, neurons undergo periods of hyperpolarization, during which they are much further from their firing threshold. Computer simulations indicate that, if during such periods a part of the thalamocortical complex is perturbed in many different ways, the resulting states of the rest of the complex would be much less differentiated than during normal wakefulness. Although this conjecture has not yet been explicitly tested *in vivo*, if it were confirmed, it would indicate that the ability to integrate information is significantly reduced during slow-wave sleep.

Spatial and Temporal Scales

Values of Φ are dependent on both spatial and temporal scales that determine what counts as a state of the system. In general, the right 'grain size' in both space and time should be the one at which Φ/t reaches a maximum. Biological considerations suggest that synchronous firing of heavily interconnected groups of neurons sharing inputs and outputs (e.g., cortical minicolumns) may produce significant effects in the rest of the brain, whereas asynchronous firing of various combinations of individual neurons should not.

Thus, Φ values may be higher when considering as elements cortical minicolumns rather than individual neurons, even if their number is lower. In regard to time, Φ values in the brain are likely to show a maximum between tens and hundreds of milliseconds. It is clear, for example, that if one were to perturb one half of the brain and examine what effects this produces on the other half, no perturbation would produce any effect whatsoever after just a tenth of a millisecond ($\Phi = 0$). However, after 100 ms, for example, there is enough time for differential effects to be manifested. On the other hand, given the duration of conduction delays and of postsynaptic currents, longer intervals will not increase Φ values. Thus the time scale of

neurophysiologic interactions needed to integrate information among distant cortical regions is consistent with that required by the microgenesis of conscious experience.

A Few Corollaries and Predictions

Several corollaries and predictions derive from viewing and measuring consciousness as the amount of information that can be integrated within a complex of elements. Some of them are straightforward. For instance, stimulation and recording studies should indicate that groups of neurons in the thalamocortical system form a large complex of high Φ, whereas those in the cerebellum do not. Moreover, they should indicate that effective information among various cortical areas in high during waking and REM sleep and low in slow-wave sleep. Other predictions are less intuitive. For example, as long as neurons belong to a complex with high Φ, they should contribute to conscious experience whether they are firing or not, just as the sound of an orchestra is specified both by the instruments that are playing and by those that are silent. This is because the level of consciousness is a property of a system – a complex with high Φ – not of a state. Another prediction is that the same neurons can contribute or not to consciousness depending on whether they are part of the main thalamocortical complex. Yet another prediction states that when a task becomes automated, its execution fades from consciousness because the neural circuits that carry it become informationally insulated.

At a more general level, this approach makes it easy to see that consciousness provides an adaptive advantage and may have evolved precisely because it corresponds to the ability to rapidly integrate a lot of information. If such information is about the environment, the implication is that the more an animal is conscious, the larger the number of variables it can take into account jointly to guide its behavior. Also, in this view, consciousness would not be an all-or-none property, but rather would increase in proportion to a system's ability to integrate information rapidly. Indeed, this approach provides a principled way of establishing, through the analysis of complexes, whether consciousness increases during development, whether and to what extent it is present in animals, and how much consciousness is left in neurologic conditions such as psychomotor seizures and akinetic mutism. Finally, an intriguing implication is that it should be possible to construct a conscious artifact, provided its physical basis is a complex having high Φ/t.

In summary, according to this theoretical framework, consciousness comes about when the parameters of a physical system are such that it can integrate a large amount of information over a short period of time. The level of consciousness can be determined, at least in principle, by evaluating $\Phi(S)/t$ for all subsets of a system at the spatial and temporal scale at which this quantity reaches a maximum. Such an evaluation identifies complexes – subsets of elements that can genuinely integrate information – and determines how much information they can integrate. At present, the validity of this theoretical framework rests on its ability to account, in a coherent manner, for some elementary but puzzling facts about consciousness and the brain. However, several experimental and theoretical developments are needed to establish whether it can provide a useful approach to the first problem of consciousness. Moreover, it remains to be seen whether the second problem of consciousness – what determines the way consciousness is – can be usefully addressed within the same framework. An interesting possibility is that, just as the amount of effective information that can be integrated within a complex would determine the level of consciousness, the space defined by the values of effective information among the elements of a complex would determine its content. Undoubtedly, a full understanding of how the brain generates consciousness remains a formidable task. However, if experimental investigations can usefully be complemented by principled theoretical approaches, it may not lay necessarily beyond the reach of science.

See also: Consciousness: Theories and Models; Consciousness: Neurophysiology and Visual Awareness in; Consciousness: Theoretical and Computational Neuroscience; Consciousness: Neural Basis of Conscious Experience; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Baars BJ (1997) *Inside the Theater of Consciousness*. New York: Oxford University Press.

Bachmann T (2000) *Microgenetic Approach to the Conscious Mind*. Amsterdam: John Benjamins.

Crick F and Koch C (2003) A framework for consciousness. *Nature Neuroscience* 2: 119–126.

Dehaene S and Naccache L (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79: 1–37.

Edelman GM and Tononi G (2000) *Consciousness: How Matter Becomes Imagination*. New York: Basic Books.

Rees G, Kreiman G, and Koch C (2001) Neural correlates of consciousness in humans. *Nature Reviews Neuroscience* 3: 261–270.

Tononi G (2003) *Galileo and the Photodiode*. Bari, Italy: Laterza.

Zeman A (2003) *Consciousness: A User's Guide*. New Haven: Yale University Press.

Consciousness: Theoretical and Computational Neuroscience

C Koch, California Institute of Technology, Pasadena, CA, USA
G Tononi, University of Wisconsin–Madison, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

What Phenomena Does a Science of Consciousness Need to Explain?

There are many definitions of consciousness, with some scholars arguing that it cannot be defined at all and others stating that it can only be defined by example. A common philosophical definition is "Consciousness is what it is like to be something." This would include such experiences as what it feels like to see red, or to be angry. This what-it-feels-like-from-within definition expresses what many believe is the most unique and irreducible characteristic of consciousness: being conscious means to experience something. This characteristic is called the phenomenal aspect of consciousness (from the Greek 'phenomenon,' i.e., appearance or experience). "What it feels like to be me, to see red, or to be angry" also emphasizes the subjective or first-person perspective of consciousness: it is a subject, an I, who is having the experiences, and experiences are inevitably private. What it feels like to have a particular experience is also called the quale of that experience: the quale of red is what is common to such disparate conscious states as seeing a red sunset, the red flag of China, arterial blood, or a ruby gemstone. All four distinct subjective states share 'redness.' There are countless qualia (the plural of quale): the ways things look, sound, and smell, the way it feels to have a pain, the way it feels to have thoughts and desires, and so on. To have an experience means to have qualia, and the quale of an experience is what specifies it and makes it different from other experiences.

The science of consciousness must explain the exact relationship between brain states and phenomenal, mental states. This is the heart of the classical mind–body problem. It can be dissected into several subproblems:

1. Why is there any experience at all? Or, put differently, why does a brain state feel like anything? In philosophy, this is referred to by some as the Hard Problem (note the capitalization), or as the explanatory gap between the material, objective world and the subjective, phenomenal world. Throughout the ages, many scholars have argued that the exact nature of this relationship will remain a central puzzle of human existence, without an adequate reductionistic, scientific explanation. However, as similar sentiments have been expressed in the past for the problem of seeking to understand life or to understand of what material the stars are made out of, it is best to put this question aside for the moment and not be taken in by defeatist arguments.
2. Why is the relationship among different experiences the way it is? For instance, red, yellow, green, cyan, blue, and magenta are all colors that can be mapped onto the topology of a circle. Why? Furthermore, as a group, these color percepts share certain communalities that make them different from other percepts, such as seeing motion or smelling a rose.
3. Why are feelings private? As expressed by poets and novelists, we cannot communicate an experience to somebody else except by way of example.
4. How do feelings acquire meaning? Subjective states are not abstract states but have an immense amount of associated explicit and implicit feelings. Think of the unmistakable smell of dogs coming in from the rain or the crunchy texture of potato chips.
5. Why are only some behaviors associated with conscious states? Much brain activity and associated behavior occurs without any conscious sensation.

Another aspect of the mind–body problem is the question of freedom of choice or free will. This is orthogonal to the problem of consciousness; whether volition is illusory or truly 'free' in the libertarian sense does not answer the question of how subjective states relate to brain states. On the other hand, the perception of free will, what psychologists call the feeling of authorship or agency (e.g., "I raised my left hand"), is a subjective state with an associated quale no different in kind from the quale of a toothache or hearing a high C.

Unless and until science can explain how subjective experiences arise from physical states, its claim to provide an adequate account of all phenomena in the universe will ring hollow. Sources of data about subjective states come not only from persons that can directly report (i.e., talk) about their subjective experiences, but also from nonlinguistic competent individuals – newborn babies or locked-in syndrome patients – and from appropriately trained animals. There are three reasons to assume that many species, in particular those with complex behaviors, such as mammals, share at least some aspects of consciousness with humans:

1. Similar neuronal architectures. Except for size, there are no large-scale, dramatic differences between the cerebral cortex and thalamus of mice, monkey, humans, and whales. In particular, the macaque

monkey is taken to be a powerful model organism to study visual perception, since it shares with the human visual system three distinct cone photopigments, binocular stereo, a foveated retina, and similar eye movements.

2. Similar behavior. An appropriately trained macaque monkey behaves in a manner similar to that of a trained human observer. Almost all human behaviors have precursors in the animal literature. Take the case of pain. The behaviors seen in humans when they experience pain – facial contortions, moaning, yelping, or other forms of auditory vocalization; motor activity such as writhing; expressing fear at the prospect of a repetition of the painful stimulus – can be observed in mammals and most other animals. Likewise for the physiological signals that attend pain – change in blood pressure, dilated pupils, sweating, increased heart rate, release of stress hormones, and so on. Indeed, the debate concerning cortical pain responses in premature babies shows the fallacy of relying on language as sole criteria for consciousness.

3. Last, there is a great evolutionary continuity between *Homo sapien*s and other mammals. Thus, while it is likely that certain aspects of consciousness – in particular those relating to the notion of self and to abstract, culturally transmitted knowledge – are not widespread in nonhuman animals, there is little reason to doubt that other mammals share conscious feelings – sentience – with humans. The extent to which nonmammalian species, such as crows, squids, or bees, with very complex, nonstereotyped behaviors, including delayed matching and non-matching-to-sample and other forms of learning, are conscious is difficult to answer at this point in time.

Distinctions between States and Levels of Consciousness

As long as you are awake, you experience a specific conscious state with a specific content. You see a face, you hear the radio, or you think of your mother. This usage of conscious states to refer to individual experiences is quite distinct from the expression of states of consciousness, which is used in a different context. Clinicians talk about impaired states of consciousness as in "the comatose state," "the persistent vegetative state (PVS)," or "the minimal conscious state (MCS)." Here, state refers to a different level of consciousness, from a total absence in coma and in PVS to a minimal amount of consciousness in MCS, as compared to full-blown waking consciousness. The repertoire of distinct conscious states (in the sense of distinct experiences) that are accessible to a patient in MCS is presumably minimal (mainly pain and discomfort, at times limited visual or other sensory experiences), immeasurably smaller than the possible conscious states that can be experienced by a healthy brain. A related use of differentiation is when psychologists refer to altered states of consciousness – for instance, state after taking psychedelic drugs or during dreaming, when insight and introspection are usually missing and events often have a stronger emotional connotation than in normal life. In both cases, the basic physiology of the brain is changed, affecting the space of possible conscious experiences. (This dual use of the word 'state,' referring both to a specific state of consciousness as well as to different levels or modes of consciousness, is analogous to the use of 'state' in physics. There, the state of a system can both refer to its position in an underlying phase space and to different states of matter – solid, liquid, gas, and plasma.)

Neuronal Correlates of Consciousness

It is often assumed in the debate concerning the nature of consciousness that there are few relevant facts and therefore everybody is entitled to their own theory of consciousness. Nothing could be further from the truth. There is an immense amount of psychological, clinical, and neuroscientific data and observations that need to be accounted for. In particular, the contemporary focus on the neuronal basis of consciousness in the brain – rather than on eristic philosophical debates concerning exact definitions, whether true zombies can exist, or whether or not the Hard Problem can or cannot be solved – has given neuroscientists the tools to effectively attack this problem. As defined by FC Crick and C Koch, the neuronal correlates of consciousness (NCC) are the minimal neuronal mechanisms jointly sufficient for any one specific conscious percept (minimal, as the entire brain is sufficient to give rise to consciousness). For instance, it is likely that neural activity in the cerebellum does not underlie any conscious perception, and thus it is not part of the NCC. That is, trains of spikes in Purkinje cells (or their absence) will not induce a sensory percept, although they may ultimately affect some behaviors (such as eye movements).

Every phenomenal state will have associated NCC: one for seeing a red patch, another one for seeing grandmother, yet a third one for hearing a tone, and so on. Perturbing or inactivating the NCC for any one specific conscious experience will affect the percept or cause it to disappear. If the NCC could be induced artificially – for instance, by drugs or with the aid of extracellular electrical stimulation – the subject will experience the associated percept.

What characterizes the NCC? What are the communalities between the NCC for seeing and for

hearing? Will the NCC involve all pyramidal neurons in cortex at any given point in time? Or only a subset of long-range projection cells in frontal lobes that project to the sensory cortices in the back? Only layer 5 cortical cells? Neurons that fire in a synchronous manner? These are some of the proposals that have been advanced over the past years.

The NCC for the current content of consciousness has to be distinguished from the general, enabling factors in the brain that are needed for any form of consciousness to occur and that also control the level of consciousness. One very general enabling factor is a proper blood supply. Inactivate the heart and consciousness ceases within a fraction of a minute. This does not imply that the neural correlates of consciousness is in the heart (as Aristotle thought). A more specific neuronal enabling factor for consciousness is the reticular activating system. This heterogeneous set of cells in the midbrain and posterior hypothalamus release neuromodulators such as acetylcholine, noradrenaline (norepinephrine), serotonin, and histamine. These neuromodulators control the excitability of thalamus and forebrain and mediate the alternation between wakefulness and sleep. Acute lesions of the reticular activating system can result in loss of consciousness and coma. However, eventually the excitability of thalamus and forebrain recovers and consciousness can return. Another enabling factor for consciousness are the five or more intralaminar nuclei (ILN) of the thalamus. These project strongly to the basal ganglia and, in a more distributed manner, to much of neocortex. Comparatively small (1 cm^3 or less), bilateral lesions in the thalamic ILN completely knock out all awareness. Of course, it is possible that future research will demonstrate hidden aspects of neuronal specificity, such that particular neuronal subpopulations in the ILN are responsible for different conscious modalities (e.g., vision vs. audition).

In functional magnetic resonance imaging (fMRI) studies, much of the contemporary work aimed at establishing the NCC has concentrated on visual consciousness. Visual psychologists have perfected a number of techniques – masking, binocular rivalry, continuous flash suppression, motion-induced blindness, change blindness, inattentional blindness – in which the seemingly simple and unambiguous relationship between a physical stimulus in the world and its associated percept in the privacy of the subject's mind is disrupted. In particular, a stimulus can be perceptually suppressed for seconds or even minutes at a time: the image is projected into one of the observer's eyes but invisible, not seen. This permits the footprints of visual consciousness to be tracked in the brain. In binocular rivalry, a small image – say, of a horizontal grating – is presented to the left eye and another image – say, a vertical grating – is shown to the corresponding location in the right eye. In spite of the constant visual stimulus, observers 'see' the horizontal grating alternate every few seconds with the vertical one. The brain does not allow for the simultaneous perception of both images. It is possible, though difficult, to train a macaque monkey to report whether it is currently seeing the left or the right image. The distribution of the switching times and the way in which changing the contrast in one eye affects these leaves little doubt that monkeys and humans experience the same basic phenomenon.

In a series of elegant experiments, N Logothetis and colleagues recorded from a variety of visual cortical areas in the awake macaque monkey while the animal performed under binocular rivalry condition. In primary visual cortex, only a small fraction of cells modulated their response as a function of the percept of the monkey. The majority of cells increased their firing rate in response to one or the other retinal stimulus with little regard to what the animal perceived at the time. In contrast, in a high-level cortical area such as the inferior temporal (IT) cortex, almost all neurons responded only to the perceptual dominant stimulus (in other words, a 'face' cell only fired when the animal indicated by its performance that it saw the face and not the pattern presented to the other eye). This makes it likely that the NCC involves activity in neurons in the IT lobe. Lesions in the homologous area in the human brain are known to cause very specific deficits in conscious face or object recognition. However, it is possible that specific interactions between IT cells and neurons in parts of the prefrontal cortex are necessary in order for the NCC to be generated. However, sophisticated functional brain imaging experiments in human volunteers demonstrate quite conclusively that hemodynamic activity in V1 does follow the percept rather than the retinal stimulus, unlike the majority of neurons in monkey primary visual cortex. This discrepancy between single cell and fMRI data, operating at quite different spatiotemporal scales, has not yet been resolved.

There is a lively debate of the extent to which neurons in primary visual cortex simply encode the visual stimulus or are directly responsible for expressing the observer's conscious percept. That is, is V1 part of the NCC? It is clear that retinal neurons are not part of the NCC for visual experiences. While retinal neurons often correlate with visual experience, the spiking activity of retinal ganglion cells does not accord with our visual experience (e.g., there are no photoreceptors at the blind spot, yet we do not see a hole in our field of view; there is a dramatic drop in cone photoreceptor density away from the fovea, yet

the visual field appears everywhere colorful; the eyes moves three to four times each second, yet the world is seen as stable). Furthermore, in dreams, the eyes are closed while we experience vivid visual scenes. Other empirical problems of contemporary interest include the question of whether each conscious attribute is all-or-none, whether perception is not continuous but is actually quantized in batches of variable duration, and whether different percepts (i.e., color, motion, position) are simultaneously experienced or not.

In order to advance these and similar questions, it will be imperative to record from large numbers of neurons simultaneously at many locations throughout the cortico-thalamo system. Our growing ability to manipulate identified populations of cortical or thalamic cells, in a reversible, transient, deliberate, and delicate manner, using tools from molecular biology, combined with optical stimulation techniques, opens the possibility of moving from correlation (observing that a particular conscious state is associated with some neural or hemodynamic activity) to causation. Of course, this requires the development of appropriate behavioral assays and model organisms amenable to large-scale genomic analysis and manipulation.

Zombie Behaviors

Many actions in response to sensory inputs are rapid, transient, stereotyped, and unconscious. They could be thought of as cortical reflexes and are characterized by rapid and somewhat stereotyped responses, in addition to a slower, all-purpose conscious mode. Consciousness deals more slowly with broader, less stereotyped aspects of the sensory inputs (or a reflection of these, as in imagery) and takes time to decide on appropriate thoughts and responses. A general purpose, consciousness mode is needed because otherwise a vast number of different zombie modes would be required to react to unusual events. The conscious system may interfere somewhat with the concurrent zombie system. It would be a great evolutionary advantage to have both zombie modes that respond rapidly in a stereotyped manner, and a slightly slower system that allows time for thinking and planning more complex behavior. This latter aspect would be one of the functions of consciousness. It seems possible that visual zombie modes in the cortex mainly use the dorsal stream in the parietal region. However, parietal activity can affect consciousness by producing attentional effects on the ventral stream, at least under some circumstances. The conscious mode for vision depends largely on the early visual areas (beyond V1) and especially on the ventral stream. It is well known that seemingly complex visual processing (such as recognizing animals in natural scenes) can be accomplished by cortex within 130–150 ms, far too slow for conscious perception to occur. Furthermore, reflexes such as the vestibuloocular reflex (VOR) take place at even more rapid timescales. It is plausible that such behaviors are mediated by a purely feed-forward moving wave of spiking activity while conscious activity requires more sustained, reverberatory neural activity that in turn requires excitatory feedback to maintain its activity above a critical threshold. This would be true of any cortical region: fast, transient, and feed-forward activity can trigger behavior without conscious sensation.

Consciousness and Information Integration

Progress in the study of the NCC on one hand, and of the neural correlates of zombie behaviors on the other, may lead to a better understanding of what distinguishes neural structures or processes that are associated with consciousness from those that are not. But even if we come closer to this goal, we still need to understand why certain structures and processes have a privileged relationship with subjective experience. For example, why is it that neurons in corticothalamic circuits are essential for conscious experience, whereas cerebellar neurons, despite their huge numbers, are not? And what is wrong with cortical zombie systems that makes them unsuitable to yield subjective experience? Or why is it that consciousness wanes during slow-wave sleep early in the night, despite levels of neural firing in the corticothalamic system that are comparable to wakefulness? Other questions are even more difficult to address in the absence of a theory. For example, is consciousness present in animals that have a nervous system considerably different from ours? And what about com puterized robots or other artifacts that behave intelligently but are organized in a way radically different from that of human brains?

To address these questions we need a theoretical approach that establishes at the fundamental level what consciousness is, how it can be measured, and what requisites a physical system must satisfy in order to generate it. The information integration theory of consciousness represents such an approach. According to this theory, the most important property of consciousness is that it is extraordinarily informative. Whenever a particular conscious state is experienced, it rules out a huge number of alternative experiences, and, classically, the reduction of uncertainty among a number of alternatives constitutes information. For example, with the experience of reading a particular

phrase in a book, a huge number of other possible experiences are ruled out. Just think of all possible written phrases that could have been read, in all possible fonts, ink colors, and sizes, then think of the same phrases spoken aloud, or read and spoken, or think further of all other possible visual scenes that might have been experienced, and then multiplied by all possible sounds that might have been heard at the same time, and by all possible moods that the reader might have been in, and so on. Thus, every time we experience a particular conscious state, it is out of a huge repertoire of possible conscious states, and we gain access to a correspondingly large amount of information. This point is so simple that its importance has been overlooked. It is essential to realize, however, that the information associated with the occurrence of a conscious state is integrated information. When a particular conscious state is experienced, that conscious state is an integrated whole – it cannot be subdivided into components that are experienced independently. For example, the conscious experience of reading a particular phrase cannot be experienced as subdivided into, say, the conscious experience of how the words look independently of the conscious experience of how they sound in the mind of the reader. Similarly, visual shapes cannot be experienced independently of their color, or the left half of the visual field of view perceived independently of the right half. Based on these and other considerations, the theory claims that a physical system can generate consciousness to the extent that it can integrate information. This requires that the system has a large repertoire of available states (information) yet it cannot be decomposed into a collection of causally independent subsystems (integration).

Importantly, the theory introduces a measure of a system's capacity to integrate information. This measure, called Φ, is obtained by determining the minimum repertoire of different states that can be produced in one part of the system by perturbations of its other parts. Although this measure is not easy to calculate exactly for realistic systems, in principle it can be applied to any system of elements. For example, by using simple computer simulations, it is possible to show that Φ is high for neural architectures that conjoin functional specialization with functional integration, like the mammalian thalamocortical system. Conversely, Φ is low for systems that are made up of small, quasi-independent modules, like the cerebellum, or for networks of randomly or uniformly connected units. Thanks to the theoretical notion of consciousness as information integration, it becomes possible to provide a principled account for key neurobiological observations concerning consciousness, including (1) the association of consciousness with certain neural systems rather than with others, (2) the fact that neural processes underlying consciousness can influence or be influenced by neural processes that remain unconscious, (3) the reduction of consciousness during slow-wave sleep and generalized seizures, and (4) the time requirements on neural interactions that support consciousness. Recently, the notion that consciousness has to do with the brain's ability to integrate information has been tested directly by measuring electroencephalogram responses evoked by transcranial magnetic stimulation (TMS) when study participants were awake and conscious, compared to when they were deeply asleep early in the night – a time during which consciousness is much reduced. During quiet wakefulness, an initial response at the stimulation site was followed by a sequence of waves that moved to connected cortical areas several centimeters away. During slow-wave sleep, by contrast, the initial response was stronger but was rapidly extinguished and did not propagate beyond the stimulation site. Thus, the fading of consciousness during certain stages of sleep may be related, as predicted by the theory, to the breakdown of information integration among specialized thalamocortical modules.

See also: Blindsight: Residual Vision; Cerebral Cortex; Coma and Other Pathological Disorders of Consciousness; Consciousness: Neural Basis of Conscious Experience; Consciousness: Neurophysiology and Visual Awareness in; Consciousness: Theories and Models; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging.

Further Reading

Blake R and Logothetis NK (2002) Visual competition. *Nature Reviews Neuroscience* 3: 13–21.

Bogen JE (1995) On the neurophysiology of consciousness. I: An overview. *Consciousness & Cognition* 4: 52–62.

Chalmers D (1995) *The Conscious Mind: In Search of a Fundamental Theory.* Oxford, UK: Oxford University Press.

Crick FC and Koch C (1995) Are we aware of neural activity in primary visual cortex? *Nature* 375: 121–123.

Crick FC and Koch C (2003) A framework for consciousness. *Nature Neuroscience* 6: 119–127.

Dawkins MS (1993) *Through Our Eyes Only? The Search for Animal Consciousness.* Oxford, UK: Oxford University Press.

Edelman GM and Tononi G (2000) *Consciousness: How Matter becomes Imagination.* New York: Basic Books.

Goodale MA and Milner AD (2004) *Sight Unseen.* Oxford, UK: Oxford University Press.

Han CJ, O'Tuathaigh CM, van Trigt L, et al. (2003) Trace but not delay fear conditioning requires attention and the anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the United States of America* 100: 13087–13092.

Kim C-Y and Blake R (2004) Psychophysical magic: Rendering the visible 'invisible.' *Trends in Cognitive Science* 9: 381–388.

Koch C (2004) *The Quest for Consciousness: A Neurobiological Approach.* Denver, CO: Roberts.

Logothetis N (1998) Single units and conscious vision. *Philosophical Transactions of the Royal Society of London B* 353: 1801–1818.

Massimini M, Ferrarelli F, Huber R, et al. (2005) Breakdown of cortical effective connectivity during sleep. *Science* 309: 2228–2232.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford, UK: Oxford University Press.

Slater R, Cantarella A, Gallella S, et al. (2006) Cortical pain responses in human infants. *Journal of Neuroscience* 26: 3662–3666.

Thorpe S, Fize D, and Marlot C (1996) Speed of processing in the human visual system. *Nature* 381: 520–522.

Tononi G (2004) An information integration theory of consciousness. *BMC Neuroscience* 5: 42–72.

Tononi G and Edelman GM (1998) Consciousness and complexity. *Science* 282: 1846–1851.

Villablanca JR (2004) Counterpointing the functional role of the forebrain and of the brainstem in the control of the sleep-waking system. *Journal of Sleep Research* 13: 179–208.

Consciousness: Theories and Models

B Baars, Neurosciences Institute, San Diego, CA, USA
A K Seth, University of Sussex, Sussex, UK

© 2009 Published by Elsevier Ltd.

Introduction

There are only a few basic ways to study conscious brain events. One is to compare conscious and unconscious states, and here we can look to classical experiments on the brain stem and basal nuclei such as the reticular formation, zona incerta, and parts of the thalamus. There is a large literature on neuromodulation and cortical activation in waking, sleep, rapid eye movement (REM) dreams, comalike states, general anesthesia, and so on.

Conscious contents are commonly indexed by voluntary reports in humans and by accurate match-to-sample responses in related species. In humans, the simplest approach is to ask about easily described events, such as a visual coffee cup or a mental image of a rotating kitchen chair. Conscious contents such as these can be reported with good accuracy under the proper conditions, such as immediate report, freedom from distraction, and independent verification. These conditions apply to numerous standard experiments.

Traditional sensory science compares conscious stimuli to one another, a method that dates back to Newton's prism experiments. However, in a recent wave of research, conscious events are compared to unconscious events which appear to be physiologically similar, although they cannot be reported as conscious even under optimal conditions. Many experimental paradigms allow for such comparisons. Such experiments may be said to isolate the issue of conscious content as such in that they treat it as an experimental variable. A classic example is binocular rivalry, in which two different images are shown simultaneously to the two eyes and conscious perception alternates between them. As long as the input to the two eyes cannot be integrated into a single, coherent visual event, they compete against one another for conscious dominance. The brain differences between conscious perception and unconscious input registration provides evidence about visual consciousness as such.

With the advent of this burgeoning new literature, the idea that conscious cognition cannot be studied scientifically has largely faded. Instead, there is now a growing need for theoretical frameworks and neuroscience-based models able to integrate the rapidly growing body of experimental data. Theories and models of consciousness are also needed to guide future experiments and to make testable predictions.

Major Features

Much of the evidence to be accounted for is summarized in Table 1 and in the following discussion. Needless to say, this is an open set that will expand as we learn more.

Conscious Events Recruit Widespread Brain Activity

Many studies suggest that conscious contents mobilize the frontal and parietal brain regions, even if they originate posteriorly, as in the case of sensory events. For example, Dehaene and colleagues used visual backward masking to compare conscious and unconscious words, using a pattern mask presented immediately afterward. Masked words are unconscious, but they are not physically blocked from entering the eyes. Brain imaging analysis showed that both conscious and unconscious words activated the cortical vision and word recognition areas, which are known to analyze stimulus location, shape, color, and word identity. However, conscious words triggered 12 times more activity in these regions than identical unconscious ones. In addition, conscious words evoked a great deal of additional activity in the parietal and frontal cortex. It seems as if the stimulus-related activity in the conscious case may be widely distributed from visual regions to other areas in the brain. This is, in fact, predicted by several current theories.

An alternative hypothesis is that, because any conscious event is presumed to be reportable, the prefrontal-parietal component may be related to readiness to report rather than conscious experience as such. This is a difficult alternative to test conclusively. As we gain a more detailed understanding of the specific functional regions of the cortex, it may become possible to do so.

Many sources of evidence suggest that conscious events show wider cortical activity than matched comparison cases. A major theoretical question, therefore, concerns the contribution of frontoparietal activity in conscious perception.

Integration and Dissemination of Focal Information

The visual cortex is clearly involved in the integration of relatively simple visual features into objects

Table 1 Some basic features of conscious brain events

Feature	*Description*
1	Consciousness is characterized by irregular, low-amplitude, fast (12 to 70 Hz) electrical field activity
2	Consciousness is associated with activity in the thalamocortical complex, modulated by subcortical nuclei
3	Consciousness involves distributed cortical activity related to conscious contents
4	Conscious events are integrated into unitary consistent scenes
5	Integrated conscious scenes occur serially; only one is experienced at a time
6	Conscious scenes are metastable and reflect rapidly adapting discrimination in perception and memory
7	Conscious scenes comprise a wide multimodal range of contents and involve multimodal sensory binding
8	Conscious scenes have a focus/fringe structure; focal conscious contents are modulated by attention
9	Consciousness is subjective and private and is often attributed to an experiencing self, which may involve distinctive brain regions
10	Conscious experience is reportable by humans, verbally and nonverbally
11	Consciousness facilitates various forms of learning; even implicit learning initially requires consciousness of the stimuli from which regularities are unconsciously inferred
12	Conscious scenes have an allocentric (world-centered) character, yet are shaped by egocentric (subject-centered) frameworks; they generally show intentionality (aboutness)
13	Consciousness is needed for much decision making, adaptive planning, and voluntary control

Adapted from Seth AK, Izhikevich E, Reeke GN, and Edelman GM (2006) Theories and measures of consciousness: An extended framework. *Proceedings of the National Academy of Sciences of the United States of America* 103(28): 10799–10804.

and scenes. Because humans in a natural environment are conscious of objects and scenes rather than isolated features (such as contrast or spatial frequency), it is plausible to think that there is a close link between these end products of visual integration and visual consciousness. N Logothetis and colleagues recorded single-cell activity in response to different kinds of binocular rivalry from the V1/V2 areas, middle temporal (**MT**)/middle superior temporal (MST) cortical areas (motion), V4 area (color), and finally inferotemporal cortex (IT)/superior temporal sulcus (STS) cortical areas, the object perception regions of the macaque temporal lobe. Only 20–40% of neurons sampled in early visual regions responded to the reported (conscious) percept, with a roughly equal number responding to the unreported percept. However, in the temporal lobe, 90% of the neurons responded only to the conscious percept. Because object recognition combines many basic features such as color, motion, and location into a single percept, this region is a plausible place for conscious objects to arise. One hypothesis is that visual integration settles into a stable equilibrium when object and scenes are identified and is then disseminated to other parts of the brain. In this view conscious contents represent the result of integration, but the integrated percept then interacts with numerous regions elsewhere in the brain.

V Lamme and others have raised the question of whether visual consciousness can be said to arise from the visual cortex alone or whether it requires selective attention and other top-down influences from other regions. For example, dorsal visual stream activity is widely considered to be unconscious in itself, but it may provide allocentric and egocentric spatial maps for objects defined in the ventral visual stream. Thus, dorsal parietal regions may be contextual with respect to conscious object and scene perception. Parietal maps may be necessary for conscious visual perception, but they may not enter into experience directly. Lamme has proposed that reentrant activity in the visual cortex may be the key to visual consciousness. Cortical reentry, like other forms of reentry, will have the effect of stabilizing and strengthening activity in particular neural pathways.

F Crick and C Koch have suggested that consciousness may require competition among coalitions of neurons, in which winning coalitions determine the contents of consciousness at a given time. Such neuronal coalitions bear similarities to Hebbian cell assemblies on a very large and dynamic scale. They also suggest that unconscious processing may consist largely of feed-forward cortical waves, whereas consciousness may involve standing waves created by bidirectional signal propagation. Direct brain imaging and electrophysiological evidence for such phenomena has been found by W Freeman, R Srinivasan, P Nunez, and others. A distinctive aspect of Crick and Koch's view is the reintroduction of the controversial notion of an inner homunculus; they suggest that conscious perception may involve anterior brain regions looking at posterior activity.

Fast Cortical Interactions (Gamma–Theta)

A conscious task, such as pointing to a coffee cup, involves a great many cortical and subcortical regions. A range of evidence and theory suggests that synchronized or coherent firing of large numbers

of neurons plays a part. Gamma and alpha-range synchrony have been implicated in the conscious interpretation of ambiguous stimuli. However, coherent oscillatory activity of large populations of cells has been observed in other tasks as well, including theta activity in episodic retrieval (involving the medial temporal lobe (MTL) and higher cortex), in μ oscillations in motor regions, during preparation of motor acts (?), and possibly in 10 Hz coherence during endogenous mentation in posterior cortex. These circa 10 Hz oscillations may serve to group faster rhythms such as beta and gamma. Only delta oscillations (<2.5 Hz) are generally associated with unconscious states.

Several studies suggest that synchronous activity during conscious perception is widely distributed in the brain. However, not all brain areas are firing in synchrony – if they did, the result might resemble an epileptic seizure, in which one single rhythm entrains large regions of cortex. Rather, there seem to be rapid moments of synchrony between different brain regions that need to work together to accomplish the task of conscious perception. In deep sleep and other unconscious conditions, the individual firing rate of cortical neurons may not change, but rapid interactive linking between regions seems to be disrupted.

Current Theories and Models

Consciousness is often viewed as an architectural aspect of the nervous system because it is involved with a variety of specific contents and functions. Such architectures can be expressed neurocomputationally, or in terms of symbolic computations. Neural network mathematics has broad flexibility and explanatory potential. Formally, even symbolic information processing models are expressible in neurocomputational terms. In practice, however, it is a challenge to scale neural networks up to the massive complexity of real brains.

Qualitative and Hybrid Architectures

B Baars has summarized a body of cognitive and brain evidence for a globalist view of conscious cognition, interpreted in a theoretical framework called global workspace theory (GWT). GWT assumes the existence of multiple knowledge sources that work to identify vague, ambiguous, or unpredictable stimuli in a cooperative and competitive fashion, by posting hypotheses about the identity of the input in a common memory space, a blackboard or global workspace. This approach was surprisingly successful when first applied in the 1970s to the difficult problem of computer-based speech recognition in a noisy acoustical environment. Today, global workspace architectures are widely used in robotics and artificial intelligence. They can be implemented in both symbolic and neural network formats.

S Dehaene, J-P Changeaux, and colleagues proposed a neuronal implementation of a global workspace architecture, called neuronal global workspace. In this model, sensory stimuli mobilize excitatory neurons with long-range cortico-cortical axons, leading to the genesis of a global activity pattern among workspace neurons. Any such global pattern can inhibit alternative activity patterns among workspace neurons, thus preventing the conscious processing of alternative stimuli. The global neuronal workspace model predicts that conscious presence is a nonlinear function of stimulus salience (i.e., a gradual increase in stimulus visibility should be accompanied by a sudden transition of the neuronal workspace into a corresponding activity pattern). Such ignition of the workspace is consistent with the notion that consciousness is an all-or-nothing, as opposed to a graded, process.

The neuronal global workspace has been used to model a number of experimental phenomena in consciousness. In one study, spontaneous workspace activity blocked external sensory processing, in a manner analogous to inattentional blindness, in which normal subjects engaged in effortful mental activity fail to notice salient but irrelevant stimuli. A similar phenomenon is the attentional blink, in which subjects are unable to report a stimulus S2 presented within a short time interval (100–500 ms) following the presentation of a first stimulus S1. In a neuronal global workspace model of this paradigm, the presentation of S1 evoked both bottom-up and top-down signal flow, leading to the sort of sustained activity that might underlie conscious reportability in human subjects. However, the presentation of S2 immediately after S1 evoked only bottom-up activity because of the global activity constraints imposed by S1. Plausibly, this absence of sustained activity relating to S2 corresponds to the lack of conscious access apparent in the attentional blink.

An advantage of the GWT approach is that it is clearly functional; that is, it describes a known architecture for integrating and distributing focal information in a large-scale, massively parallel, brain-like organ. However, scaling neural net architectures up to the level of real brains present major challenges. Current neuronal GWT models make use of simplified network architectures where entire brain areas are represented by single neurons.

Several other groups are currently applying GW-based ideas to the question of conscious cognition. In another approach to circumventing the difficulty

of capturing high-level and low-level features in a single model, S Franklin and colleagues made use of a hybrid computational approach, combining in their learning IDA (LIDA) model several types of existing computational methods using a quasi-neuronal activation-passing design. High-level conceptual models such as LIDA can provide insights into the processes implemented by the neural mechanisms underlying consciousness, without necessarily specifying the mechanisms themselves Although it is difficult to derive experimentally testable predictions from large-scale architectures, this hybrid architecture approach is broadly consistent with the major empirical features discussed in this article. It predicts, for example, that consciousness may play a central role in the classic notion of cognitive working memory, selective attention, learning, and retrieval.

A variety of neurocomputational models have been proposed to deal with aspects of consciousness. Some of these make global or qualitative claims, whereas others make quite specific experimental predictions.

Dynamical and Field Theories

Several theorists have advanced nonlinear dynamical theories that are thought to be able to model larger-scale neuronal architectures and time-locked nonlocal effects among brain regions. Some have used formalisms borrowed from physics, such as electromagnetic fields, wave mechanisms (W Freeman, P Nunez, R Srinivasan, and R Wallace), and even quantum mechanics (K Pribram and S Hameroff). The notion that quantum effects play a critical role in consciousness has been widely criticized on the grounds that significant quantum effects seem unlikely in neurophysiological environments unless body-temperature versions can be found.

ER John proposed that consciousness emerges from global negative entropy, which is a property of an electrical field resonating in a critical mass of brain regions. John suggested that sensory stimuli can evoke localized synchronous neuronal activation, causing deviations from a baseline brain state of maximum entropy. These deviations establish spatially distributed islands of local negative entropy, which correspond to fragments of sensation. These fragments can be bound into a unified global percept via integrative interactions within thalamocortical (TG) loops. In similar vein, P Nunez and R Srinivisan described large-scale cortical dynamics in terms of synaptic action fields, which are continuous mathematical functions representing the density of active synapses. According to these authors, consciousness depends on resonance in these fields within characteristic frequency ranges (e.g., 4–15 Hz).

Neural field theories may account well for the unified nature of conscious experience, as well as for associated properties including seriality and metastability. Although they have inspired empirical work using quantitative electroencepahlography (EEG), detailed modeling of the generation of conscious neural fields is still needed.

Neural Darwinism

Because the thalamocortical (TC) complex has extraordinarily interactive connectivity compared to other brain structures, such as the cerebellum, it is attractive to suppose that the different modalities of consciousness make use of the TC complex as a common anatomical substrate. This is not to deny the need for content-specific cortical regions.

An extensively developed TC theory of consciousness is provided by G Edelman's neural Darwinism (ND). ND suggests that brain development and dynamics are selectionist in nature, not instructionist, in contrast to computers, which carry out explicit symbolic instructions. Selectionist processes have four features:

1. A set of elements are characterized by diversity, such as populations of neuronal groups in the brain.
2. These elements can reproduce or amplify.
3. A process of selection operates on the products of diversity; for example, differential reproductive success serves to select some offspring and not others.
4. Degeneracy, the ability of different combinations of elements to perform the same function, is inherent in such systems.

In the brain, selectionism applies both to neural development and to moment-to-moment functioning. Selection in the brain is directed by value, which reflects the salience of an event and can be positive or negative. Value is analogous to selective pressure in evolution; the magnitude of a value signal corresponds to fitness. In the brain, it involves pleasure, pain, and emotional salience networks. Finally, according to ND, spatiotemporal coordination in the brain is enabled by reentry, the recursive exchange of signals among neural areas across massively parallel reciprocal connections.

According to ND, the neural systems underlying consciousness arose in evolution as a result of their ability to integrate a large number of sensory inputs and motor outputs occurring in parallel. This integration allows the discrimination of sensorimotor signals in a high-dimensional space, yielding adaptive behavior. Primary (sensory) consciousness is associated with an interaction between perceptual categorization and memory, where primary consciousness

refers to the presence of a multimodal scene of perceptual and motor events (William James's specious present or Edelman's remembered present). Primary consciousness, by this view, provides an animal with increased discriminatory selectivity, flexibility, and planning capacity when responding to complex environments, compared to its preconscious ancestors. Edelman, along with many other theorists, distinguished between primary consciousness and higher-order consciousness which in humans is associated with verbal reportability, the attribution of sensory scenes to a temporally stable self, and the ability to project conscious experiences both forward and backward in time.

Recent extensions of ND focused on informational aspects of consciousness. In the dynamic core hypothesis (DCH), Edelman and colleagues suggested that the occurrence of any particular conscious scene constitutes a highly informative discrimination for the reason that conscious scenes are at once integrated (every conscious scene is experienced all of a piece) and differentiated (every conscious scene is distinct from others). The DCH claims that conscious qualia are these discriminations and that the neural mechanisms underlying consciousness consist of a functional cluster in the TC system, within which reentrant neuronal interactions yield a succession of differentiated yet unitary metastable states. They suggested that the boundaries of the dynamic core shift over time, with some neuronal groups leaving and others being incorporated, these transitions occurring under the influence of internal and external signals.

According to the DCH, high values of a measure called neural complexity accompany consciousness. Neural complexity measures the extent to which the dynamics of a neural system are both integrated and differentiated. The component parts of a neurally complex system are differentiated; however, as larger and larger subsets of elements are considered, they become increasingly integrated. Modeling studies suggest that the distinctive reentrant anatomy of the TC system is ideally suited to producing dynamics of high neural complexity.

Similar to the DCH, and proposed by one of its originators (G Tononi), the information integration theory of consciousness (IITC) claims that consciousness corresponds to the capacity of a system to integrate information. A system is deemed capable of information integration to the extent that it has available a large repertoire of states and that the states of each element are causally dependent on the states of other elements. Like the DCH, the IITC is based on the notion that the occurrence of any conscious scene simultaneously rules out the occurrence of a vast number of alternatives and therefore constitutes a highly informative discrimination. Also like the DCH, the IITC proposes that the TC system provides the neuroanatomical substrate for the neural processes that underlie consciousness.

The IITC proposes a novel measure of the quantity of consciousness generated by a system. This measure, Φ, is defined as the amount of causally effective information that can be integrated across the weakest link of a system. Unlike neural complexity, Φ measures directed causal interactions within a system. According to the IITC, consciousness as measured by Φ is characterized as a disposition or potentiality. The contents of any given conscious scene are specified by the value, at any given time, of the variables mediating informational interactions within the system. A distinguishing feature of the IITC is that Φ is proposed to be a sufficient condition for consciousness, so that any system with sufficiently high Φ is conscious.

Both neural complexity and Φ can be considered to measure the spatial aspect of the balance between differentiation and integration in the dynamics of a neural system. A Seth and colleagues suggested useful additional measures of neural dynamics to measure other aspects of this balance, including complexity across time (temporal complexity) and across different levels of description (recursive complexity). Temporal complexity reflects the fact that consciousness extends over time in several ways. Many experiments suggest that it takes $\sim$100 ms for sensor stimuli to be incorporated into a conscious scene and that neuronal readiness potentials can appear several hundred milliseconds before the awareness of an intention to act. Recursive complexity reflects the fact that brains show rich organization at multiple levels, from molecular interactions within individual synapses to reentrant interactions among distinct brain regions. The phenomenal structure of consciousness also appears to be recursive; individual features of conscious scenes are gestalts and share organizational features with the conscious scene as a whole. Therefore, a combination of distinct measures may be needed to adequately characterize the neural dynamics relevant to consciousness.

Other Approaches

Because theoretical approaches to consciousness remain in their infancy, there is a large and growing body of alternative approaches which we are unable to describe fully here. These include the interplay of conscious experience with attention, immediate memory, selective attention, and input to an executive self.

Other approaches have focused on neuroanatomical constraints. S Zeki proposed that consciousness arises at micro features from local cortical circuits.

This micro-consciousness view is at variance with much of the evidence summarized here and is a minority position. B Merker challenged the dominant view that conscious contents depend on cortex and proposed an alternative model in which basic consciousness arises from the activity of subcortical areas including the thalamus, superior colliculus, and zona incerta.

Summary

What should we expect from a neural theory of consciousness? One thing we should not ask for is that the theory produce consciousness. After all, a theory of hurricanes is not itself windy. Rather, a useful theory should help move from establishing correlations between brain activity and conscious experience toward developing explanations that link features of brain activity with features of conscious phenomenology, as well as accounting for the relevant experimental evidence. Because consciousness is a rich biological phenomenon, the theories surveyed in this article vary in emphasis, level of abstraction, and the extent to which they provide satisfying explanations of conscious phenomena. At present, no single effort accounts for all the evidence, but we have seen marked progress in the last few decades.

See also: Awareness: Functional Imaging; Consciousness: Philosophy; Consciousness: Neurophysiology and Visual Awareness in; Consciousness: Theoretical and Computational Neuroscience; Consciousness: Neural Basis of Conscious Experience; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Baars BJ (1988) *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

Baars BJ (2002) The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences* 6(1): 47–52.

Cosmelli D, David O, Lachaux J-P, et al. (2004) Waves of consciousness: Ongoing cortical patterns during binocular rivalry. *Neuroimage* 23(1): 128–140.

Crick F and Koch C (2003) A framework for consciousness. *Nature Neuroscience* 6(2): 119–126.

Dehaene S, Naccache L, Cohen L, et al. (2001) Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience* 4(7): 752–758.

Dehaene S, Sergent C, and Changeux J-P (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences of the United States of America* 100(14): 8520–8525.

Edelman GM (1989) *The Remembered Present*. New York: Basic Books.

Edelman GM and Tononi G (2000) *A Universe of Consciousness: How Matter Becomes Imagination*. New York: Basic Books.

Franklin S and Graesser A (1999) A software agent model of consciousness. *Consciousness and Cognition* 8(3): 285–301.

John ER (2001) A field theory of consciousness. *Consciousness and Cognition* 10(2): 184–213.

Lamme V and Roelfsema P (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience* 23(11): 571–579.

Llinas RR, Ribary U, Contreras D, and Pedroarena C (1998) The neuronal basis for consciousness. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 353: 1841–1849.

Merker B (2007) Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences* 30(1): 63–81.

Nunez PL and Srinivasan R (2006) A theoretical basis for standing and travelling brain waves measured with human EEG with implications for an integrated consciousness. *Clinical Neurophysiology* 117: 2424–2435.

Seth AK and Baars BJ (2005) Neural Darwinism and consciousness. *Consciousness and Cognition* 14(1): 140–168.

Seth AK, Izhikevich E, Reeke GN, and Edelman GM (2006) Theories and measures of consciousness: An extended framework. *Proceedings of the National Academy of Sciences of the United States of America* 103(28): 10799–10804.

Sheinberg DL and Logothetis NK (1997) The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences of the United States of America* 94(7): 3408–3413.

Srinivasan R, Russell DP, Edelman GM, and Tononi G (1999) Increased synchronization of magnetic responses during conscious perception. *Journal of Neuroscience* 19: 5435–5448.

Tononi G (2004) An information integration theory of consciousness. *BMC Neuroscience* 5(1): 42.

Relevant Websites

http://www.assc.caltech.edu – Association for the Scientific Study of Consciousness.

http://www.sci-con.org – Science and Consciousness.

Contextual Interactions in Visual Perception

D Sagi, Weizmann Institute of Science, Rehovot, Israel

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The visual world is sampled by the photoreceptors in the eyes' retinas, which produce an array of local light measurements. However, perception is definitely nonlocal. Both the subjective measures of stimulus qualities and the objective measures of related discrimination thresholds strongly depend on the spatial and temporal contexts within which the stimulus is embedded. This is true for points of light as well as for higher order visual features such as color, motion, depth, texture, and shape. These effects are thought to be of major importance in assigning meaning to the visual input and in the formation of behaviorally relevant perceptual objects by the brain. The underlying neuronal mechanisms are termed 'contextual interactions.'

Contextual interactions affect all aspects of visual processing, including low-level processes at the sensory end, which are related to basic image qualities (Figure 1) and higher level processes that are concerned with the identification of visual objects. The low-level processes at the sensory end are especially intriguing because they cannot be abolished by familiarity with the stimulus, and thus introduce a profound gap between the objectively measured physical world and our intimate knowledge of it as obtained with our perceptual system. This gap is explained by the nontrivial relationship between measurements and their causes: the latter is believed to constitute the content of perception. Typical examples include a variety of perceptual constancies, such as size and color constancies, which demonstrate that perception corresponds to actual object properties in the world (distal), such as actual size and surface reflectance, rather than to the corresponding retinal (proximal) measurements. Since perception involves reasoning about sensory effects by attributing measurements to causes, it is based on a large variety of assumptions that are made by the brain while the sensory measurements are being transformed into perceivable objects. These assumptions are found to underlie many perceptual processes and are probably the result of long-term adaptation to the environment, partly through a slow evolutionary process and partly through brain plasticity operating on a time scale of human life. As such, they can be viewed as constraints on the perceived world, capturing regular events occurring in the environment, the consequences of which are implemented in the brain to allow for improved processing and information transmission in a perceptual system with inherently limited measuring and processing devices. These constraints are crucial for resolving ambiguities when different scenes produce equivalent measurements, by choosing the more reasonable perception, the one that makes sense. Sensory ambiguities are resolved within the perceptual system without reference to, and often in conflict with, real-world knowledge that can be obtained by nonsensory means, as shown in Figures 1 and 2. By exploring the choices made by our visual system in generating perceptions, it is possible to theorize about the content of perceptual objects. Current research supports the notion that perception concerns the description and segmentation of visual surfaces. The biological mechanisms required for recovering such properties from the visual stimulus can be found in early stages of visual processing and include processes that detect boundaries between surfaces as well as surface characteristics such as color, texture, and layout in space and time. All of these processes are context sensitive, meaning that these processes are affected by remote image parts, or that different image attributes affect each other.

Phenomenology

First-Order Spatial Interactions

First-order spatial interactions are contextual effects that are related to the spatial distribution of light and to influences on the perception of light intensity (brightness).

Subjective estimates of light intensity The most basic contextual effect in visual perception is seen when attempting to evaluate the amount of light falling on a small retinal area (definitions: 'luminance' is the physical quantity of light; 'brightness' is the perceived quantity of light). The light will be evaluated as being less intense when the target is surrounded by light of a higher intensity, whereas it will be evaluated as more intense when the target is surrounded by light of a lower intensity. This phenomenon, illustrated in Figure 1, is termed 'simultaneous contrast'; Figure 1 demonstrates that the perceptual qualities of dark and bright (or black and white) do not correspond to absolute light intensities but rather to gradients of intensity. Similarly, the perceived color of an object depends on the colors surrounding the object ('color contrast,' 'chromatic induction'). In a related phenomenon, called Mach bands, sharp intensity transients are

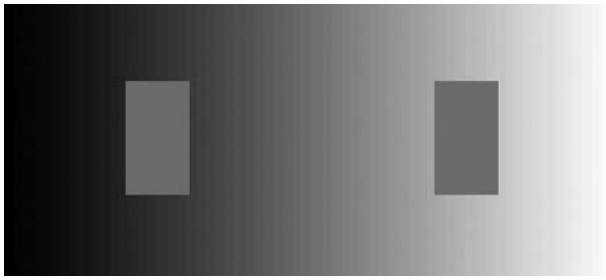

Figure 1 Context effects in brightness perception – first-order spatial interactions. The two bars have the same intensity; however, their perceptual effects differ significantly because of the different background. What we perceive are 'white' surfaces that are brighter than their immediate surround and 'black' surfaces that are darker than their immediate surround.

enhanced so that the dark side of an edge looks darker whereas the bright side of an edge looks brighter. The occurrence of Mach bands can be explained by lateral inhibition between neighboring light-detection arrays; high response is predicted when the neighboring arrays receive different levels of stimulation, and no response is predicted when the neighboring arrays receive the same level of stimulation. This type of transformation takes part in sensory tasks involving different sense organs and affects different features within the same sensory organ (see later, discussions of spatial interactions in the perception of motion and depth). A hierarchy of lateral-inhibition operations within the visual system was proposed to account for spatial interactions involved in the perception of light contrast

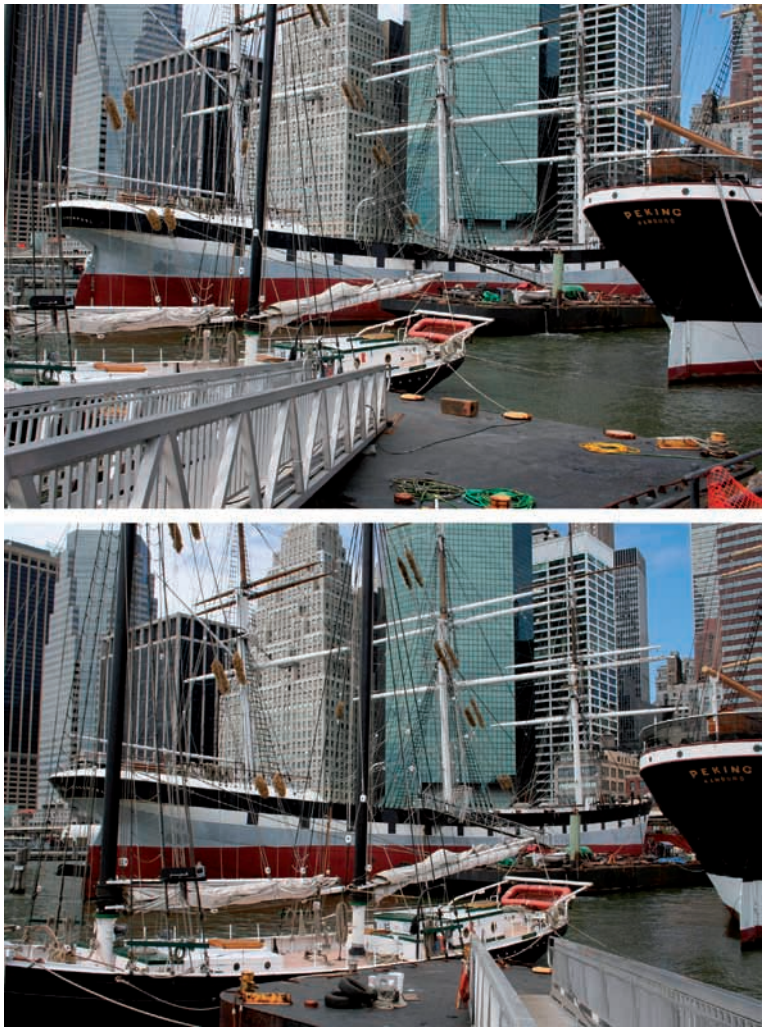

Figure 2 The Wavertree (South Street Seaport Museum, New York, NY). These images, rich in details and contours, do not pose a particular challenge to the visual system, and objects are effortlessly segmented. Such an achievement is impossible without contextual interactions within the perceptual system. However, details of shape are not reliable and the ship in the background (top image) is seen as curved. The distortion is even stronger under natural viewing conditions when the ship covers a larger part of the field of view. A slightly different point of view produces a more reasonable percept (bottom image). It is difficult to estimate the frequency of such misperceptions in our daily encounters with the visual world.

(see the section titled 'Second-order spatial interactions'). Lateral inhibition can be used to explain, at least in part, 'simultaneous contrast,' because sensory systems equipped with lateral inhibition respond to relative intensity rather than to the absolute intensity level. However, such an explanation cannot be complete because the perceived simultaneous contrast also depends on long-range interactions and on the spatial arrangement of the edges in the stimulus.

Sensitivity to changes in light intensity Discrimination thresholds depend on the light distribution in the area around the target. For a small, fixed-size target of light on a background of variable size, incremental threshold increases when the background size is initially increased – a larger increment is required to induce an appreciable difference in light intensity (desensitization) – but then decreases (sensitization) with a further increase of size.

Neurophysiological correlates Light-discrimination thresholds and the corresponding spatial interactions are successfully modeled with receptive fields of a shape found in the mammalian retina and the lateral geniculate nucleus (LGN). These receptive fields integrate light over some spatial areas, with regions in the center of the integration zone and its periphery having opposite signs (ON–OFF receptive field). This can be viewed as an implementation of the lateral-inhibition theory suggested by Mach. Thus, a neuron with such a receptive field responds to contrast – that is, to spatial variations in light intensity. Given the stimulus presented in Figure 1, such neurons, with the appropriate receptive-field size and retinal location, will respond vigorously to the left bar if their center is positively weighted (ON) and their surround is negatively weighted (OFF), whereas the right bar activates neurons with a receptive field having an opposite weight. Since this effect is possible with bars having a wide range of sizes (it can be experienced by changing the viewing distance) while receptive sizes are limited, most neuronal activity occurs around the edges of the bars, at locations of maximum contrast, and not within the uniform region within the bars. Thus, the brightness of the bars is thought to be perceived through a process of interpolation or filling-in (see later).

Second-Order Spatial Interactions

Second-order spatial interactions consist of contextual effects that are related to the distribution of light contrast in the visual field and to the perception of light contrast.

Subjective estimates of light contrast When attempting to evaluate the amount of light contrast within a textured stimulus region, the contrast will be evaluated as being weaker when the target is surrounded by a texture of higher contrast, whereas it will be evaluated as having more contrast when surrounded by a texture of lower contrast. This phenomenon, presented in Figure 3, suggests the existence of lateral inhibition within arrays of contrast detectors.

Sensitivity to changes in light contrast The detection threshold for light contrast is affected by the contrast around the target. For a localized and oriented contrast target (Gabor signal), detection thresholds increase (desensitization) in the presence of nearby high-contrast stimuli (flankers), but when the distance between the target and the flankers is increased, then there is a range in which the detection threshold decreases (sensitization). These effects are strongest when the target and flankers are co-oriented and co-aligned, as if lying on the same contour.

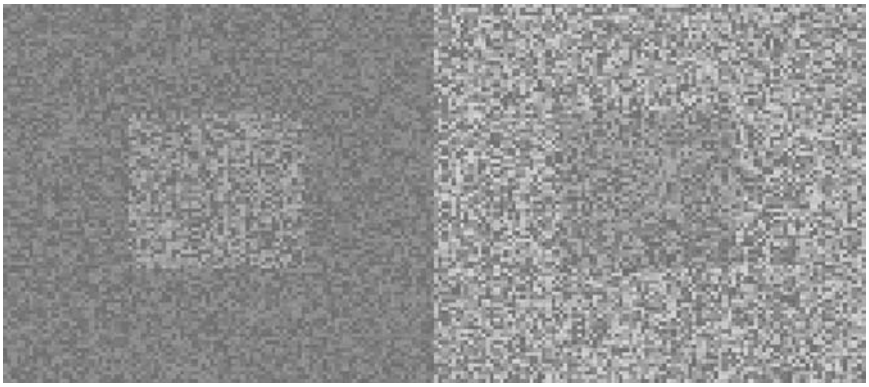

Figure 3 Second-order spatial interactions. The two smaller squares have the same contrast but their perceptual effects differ significantly because of the different background. Adapted from Chubb C, Sperling G, and Solomon JA (1989) Texture interactions determine perceived contrast. *Proceedings of the National Academy of Sciences of the United States of America* 86(23): 9631–9635.

Neurophysiological correlates Contrast-discrimination thresholds and the corresponding contextual effects can be modeled by receptive fields of neurons found in the primates' primary visual cortex and by their lateral interactions. These neurons respond to oriented contrast – that is, to nonisotropic first-order changes taking place along one direction on the retinal surface. In addition, these neurons are also affected by stimuli placed outside their classically defined receptive field (meaning that neuronal activity cannot be elicited by these stimuli when presented in isolation) if these are properly aligned with the stimuli that are placed within the receptive field. Both inhibitory and excitatory contextual interactions have been found in the visual cortex.

Spatial Interactions with respect to the Perception of Motion

The detection of motion is carried out by special mechanisms that are sensitive both to translation of luminance (first-order motion) and to contrast (second-order motion), with their resulting motion vectors combined. The perceived motion within a retinal area depends on the surrounding motion ('motion induction'). In the case of ambiguous random-dot cinematograms, when the moving target region is perceived with equal probability as moving to the left or to the right, an unambiguous surround biases the target motion in the surround direction. This result indicates the existence of lateral interactions between motion detectors that integrate motion signals to reduce local ambiguities. On a larger spatial scale, inhibitory interactions are found, with antagonistic interactions between motion signals of the same direction. The neurophysiological correlates of these effects are believed to reside in the middle temporal (MT) visual area, the specialty of which is the analysis of motion, whereby neuronal response to movement is suppressed in the presence of surround movement in the same direction.

Spatial Interactions with respect to the Perception of Depth

The detection of a three-dimensional structure is based on multiple cues (see later), but at the most basic level it is based on the detection of small displacements between features of the two retinal images captured by the two retinas (stereopsis). Stereoscopic information is often highly ambiguous and thus the existence of unique perception implies a strong dependence on other depth cues, if available, or on internal brain structure. For example, the lines seen in Figure 4 can be matched between both eyes in different ways, because each green line is a potential match for every red line (the 'correspondence problem'); however, one perceives flat surfaces. Such a product of the perceptual system cannot be achieved without spatial integration. The range of spatial interactions between isolated stereoscopic stimuli is limited to a few minutes of arc when lines are used and to three times the object's size when Gaussian blobs are used as matched features. However, with textured stimuli, disparity information is found to propagate across much larger distances of a few degrees with both real and illusory disparity edges. Neurophysiological correlates of these effects are thought to reside in visual area V2.

Plasticity of Contextual Interactions

There are many indications that perceptual processes are affected by experience. Context effects are found

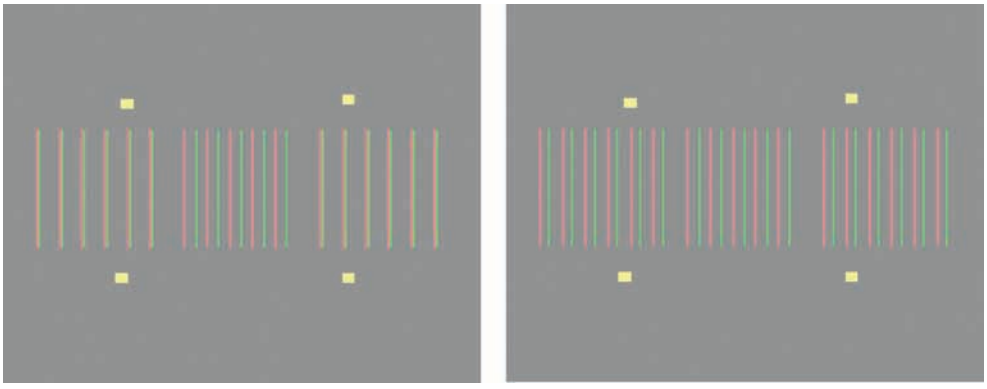

Figure 4 Ambiguous stereo images (require red–green glasses). The center part of each of the two displays is ambiguous with respect to disparity and can be seen either in front of the page or at the back when viewed alone. In the left-hand image, the flanking lines, which form depth plans different from each of the ambiguous center states, do not resolve the ambiguity, but in the right-hand image, the flanking lines attract the central bars to form a continuous plan (the depth of unambiguous regions should reverse with reversal of the glasses, viewing distance >40 cm is recommended).

to change with practice. Studies of texture segmentation show a large decrease in processing time (threefold) due to practice, and this is found to be highly specific to visual features known to be encoded in the primary visual cortex. In particular, second-order spatial interactions are found to propagate over larger distances after practice, implying increased range of excitatory interactions. First-order context effects are generally more stable and do not change with practice. However, a recent report by Adini et al. show changes in contrast discrimination thresholds after practicing with stimuli that activate second-order interactions, indicating that first-order contextual interactions are affected by practice. Electrophysiological recordings from the primary visual cortex of primates show experience-dependent activity, with receptive-field shape modified with practice. Of particular interest here is the finding by Crist et al. that changes in receptive-field size depend on the specific configuration practiced. This latter finding also shows that contextual interactions in the visual cortex are task dependent – that is, they can be specifically modulated by the current behavioral context. Psychophysical experiments show that excitatory spatial interactions are task dependent. Overall, current research provides strong evidence for different time scales involved in the adaptability of contextual interactions to the environment.

Functional Significance

Surface Properties

The brain infers (or, according to some views, reconstructs) the properties of objects in the visual field from the available measurements made by the eye; however, these measurements do not always have a unique interpretation. To reduce the number of possible interpretations, the brain relies on contextual information that may be found in the retinal image or is made available through other senses. The usefulness of such 'cues' depends on the validity of the assumptions implemented in the brain as constraints on the structure of objects in the world and their behavior.

This problem arises, for example, when estimating the light-reflectance properties (i.e., lightness) of a surface, when querying whether it is dark or bright and regarding its color. The recovery of surface reflectance requires, in addition to measuring the light reflected from the object (brightness, perceived luminance), knowledge of the light illuminating the object as well as of the geometric properties of the surface. However, the illumination cannot be measured directly and is inferred from the distribution of light reflected from regions surrounding the target object. In Figure 1, our visual system seems to assume that the illumination is gradually increasing from left to right, so that bars on the left and on the right that reflect the same light intensity (as measured by the eye), which is a product of the incident light and the surface reflectance, are expected to have different reflectance properties: higher on the left and lower on the right. Thus, it is assumed that our perceptions report the reflectance properties of objects rather than the light reflected. There is also an uncertainty regarding the shape of the surface presented in Figure 1. It is possible that the light source is uniform, and so is the surface reflectance, but the surface is curved so that the darker region is from a surface section that is more tilted relative to the direction of light and thus reflects less light. Often variations of surface shades are perceived as depth variations; however, this is not the case in Figure 1. In general, the geometric properties of the surface are also not directly available for the brain to use and therefore the recovery of a three-dimensional structure is required. Furthermore, other effects, such as cast-shadow, transparency, and air-transfer function, need to be taken into account. The reflected light reaches the eyes after traveling in air with unknown properties that need to be taken into account when the distance between the object and the eyes is large (atmospheric effects). As will be shown later, the three-dimensional structure is not available to the brain unambiguously and its recovery makes use of shading cues. Given this situation, one may doubt whether reliable perception is possible at all. The answer is that perception is not always reliable (as shown in Figure 2 and in many illusions that push our perceptual abilities to the limit) and is heavily dependent on intrinsic assumptions that are based on regularities in the world in which we live. Contextual interactions can be viewed, at least in part, as an implementation of these environmental regularities in the brain.

The recovery of geometrical properties of objects from the retinal image is ambiguous due to a reduction in dimensionality. The retinal sampling array is two-dimensional (2-D) whereas the world around us is three-dimensional (3-D); thus the lost dimension needs to be determined from contextual cues. In the case of recovering the 3-D structure of objects from their 2-D retinal projections, the brain makes use of geometric information related to perspective, shadows (cast and attached), occlusion, transparency, image blur, atmospheric perspective, optic-flow and motion parallax, texture gradients, retinal size, accommodation, eye vergence, and eye disparity. The recovery of depth information from these cues requires further assumptions about the relationship between the depth measures available from the cues and the actual depth. In addition, contextual information related to viewing

conditions needs to be integrated with the visual measurements to yield a reliable perception of the three-dimensional structure. Furthermore, ambiguity can often be resolved by combining information from different cues.

Segmentation

Surface properties are used by the visual system to locate boundaries of objects, since boundaries are often marked by a change in brightness and texture. The process of boundary detection is thought to involve contextual interactions within low-level vision, and recent studies have suggested neurophysiological correlates in the visual cortex. Object boundaries are marked by differences in luminance across the boundary when objects differ from their background in their reflectance properties. These luminance differences can be detected by edge detectors that are based on first-stage receptive fields of the type proposed to account for first-order spatial interactions that operate on different spatial scales. However, such a mechanism is not sufficient for successful image segmentation because object surfaces often include intensity edges that do not delineate the boundaries of objects but instead are produced by other causes such as surface texture. In Figure 1 the boundaries of objects are denoted by intensity differences (edges), but in Figure 3 the intensity edges are present everywhere, not only at the boundaries of objects. When looking at Figure 3, first-stage receptive fields are activated by the textured surface, though to different degrees: the higher contrast parts produce larger responses and the lower contrast parts produce lower responses. These responses do not mark the locations of boundaries but rather the changes in these responses correspond to the relevant boundaries. Thus, the task of boundary detection here involves the detection of spatial variations in the activity of first-stage receptive fields. Early texture models attempted to define global texture properties that would enable discrimination between two given textures based on their image statistics. Later models introduced local feature detection into the texture process, and assumed that segmentation is done by comparing statistics of such features. Current theories assume a local 'textural gradient' detection mechanism, thus avoiding global processes, or transferring it to the final decision stage with a simple feed-forward design. Recent theories of texture segmentation contain two processing stages that correspond to the two stages discussed earlier: first- and second-order spatial interactions.

In addition to boundary-sensitive processes, image segmentation is supported by grouping processes that bind different image parts to form objects. The underlying mechanism is presently not clear, but recent results suggest a mechanism that is based on linking local orientation signals across space. Such a mechanism can be useful in enhancing long contours, as in Figure 5, in particular closed contours, and in texture images that correspond to an object's boundaries. It was suggested that the neuronal implementation of these contour interactions includes non-isotropic excitatory connections between orientation-specific detectors within the same network that produce second-order spatial interactions. Global effects that were found, such as higher sensitivity to long and closed contours, are believed to be achieved via propagation of activity within the network.

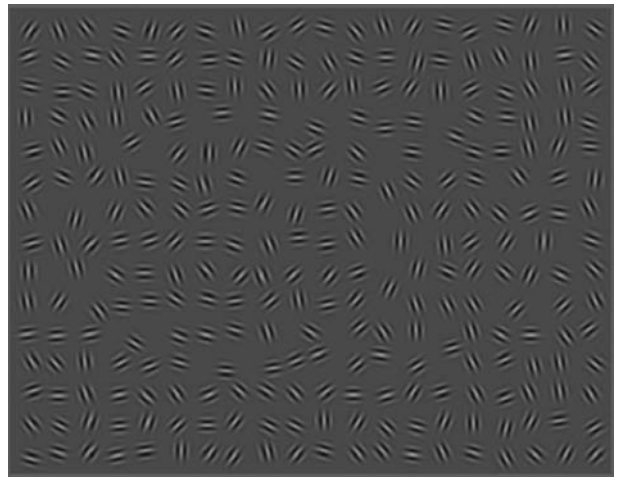

Figure 5 Evidence for efficient contour integration within the visual system. The elipse created by the closely aligned (Gabor) patches clearly stands out from the noisy background. Without the 'built-in' connectivity that allows for orientation signals to propagate along contours, we would see all patches as having equal saliency. Note that shorter contours are not as salient as the closed one is. Adapted from Kovacs I and Julesz B (1983) A closed curve is much more than an incomplete one: Effect of closure in figure–ground segmentation. *Proceedings of the National Academy of Sciences of the United States of America* 90(16): 7495–7497.

See also: Color Vision; Contextual Interactions in Visual Processing; Perception of Surfaces and Forms; Perceptual Learning: Neural Mechanisms; Perceptual Learning and Sensory Plasticity; Stereoscopic Vision; Vision for Action and Perception; Vision: Light and Dark Adaptation; Visual Cortex in Humans.

Further Reading

Adelson EH (1993) Perceptual organization and the judgment of brightness. *Science* 262(5142): 2042–2044.

Adini Y, Sagi D, and Tsodyks M (2002) Context-enabled learning in the human visual system. *Nature* 415(6873): 790–793.

Allman J, Miezin F, and McGuinness E (1985) Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local–global comparisons in visual neurons. *Annual Review of Neuroscience* 8: 407–430.

Bakin JS, Nakayama K, and Gilbert CD (2000) Visual responses in monkey areas V1 and V2 to three-dimensional surface configurations. *Journal of Neuroscience* 20(21): 8188–8198.

Beck J (1983) Textural segmentation, second-order statistics, and textural elements. *Biological Cybernetics* 48(2): 125–130.

Caelli T and Julesz B (1978) On perceptual analyzers underlying visual texture discrimination: Part I. *Biological Cybernetics* 28(3): 167–175.

Cannon MW and Fullenkamp SC (1991) Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vision Research* 31(11): 1985–1998.

Cataliotti J and Gilchrist A (1995) Local and global processes in surface lightness perception. *Perception & Psychophysics* 57(2): 125–135.

Chang JJ and Julesz B (1984) Cooperative phenomena in apparent movement perception of random-dot cinematograms. *Vision Research* 24(12): 1781–1788.

Chubb C, Sperling G, and Solomon JA (1989) Texture interactions determine perceived contrast. *Proceedings of the National Academy of Sciences of the United States of America* 86(23): 9631–9635.

Crist RE, Li W, and Gilbert CD (2001) Learning to see: experience and attention in primary visual cortex. *Nature Neuroscience* 4(5): 519–525.

Field DJ, Hayes A, and Hess RF (1993) Contour integration by the human visual system: Evidence for a local 'association field' *Vision Research* 33(2): 173–193.

Freeman E, Sagi D, and Driver J (2001) Lateral interactions between targets and flankers in low-level vision depend on attention to the flankers. *Nature Neuroscience* 4(10): 1032–1036.

Gilbert CD, Sigman M, and Crist RE (2001) The neural basis of perceptual learning. *Neuron* 31(5): 681–697.

Julesz B (1965) Texture and visual perception. *Scientific American* 212: 38–48.

Julesz B (1980) Spatial nonlinearities in the instantaneous perception of textures with identical power spectra. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 290(1038): 83–94.

Julesz B (1981) A theory of preattentive texture discrimination based on first-order statistics of textons. *Biological Cybernetics* 41(2): 131–138.

Julesz B (1986) Stereoscopic vision. *Vision Research* 26(9): 1601–1612.

Kapadia MK, Westheimer G, and Gilbert CD (2000) Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *Journal of Neurophysiology* 84(4): 2048–2062.

Karni A and Sagi D (1991) Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the United States of America* 88(11): 4966–4970.

Kovacs I and Julesz B (1993) A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences of the United States of America* 90(16): 7495–7497.

Kuffler SW (1953) Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology* 16(1): 37–68.

Lamme VA (1995) The neurophysiology of figure–ground segregation in primary visual cortex. *Journal of Neuroscience* 15(2): 1605–1615.

Landy MS and Bergen JR (1991) Texture segregation and orientation gradient. *Vision Research* 31(4): 679–691.

Li W and Gilbert CD (2002) Global contour saliency and local colinear interactions. *Journal of Neurophysiology* 88(5): 2846–2856.

Li Z (2000) Pre-attentive segmentation in the primary visual cortex. *Spatial Vision* 13(1): 25–50.

Lu ZL and Sperling G (2001) Three-systems theory of human visual motion perception: Review and update. *Journal of the Optical Society of America A: Optics, Image Science, and Vision* 18(9): 2331–2370.

Malik J and Perona P (1990) Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A* 7(5): 923–932.

Marr D and Hildreth E (1980) Theory of edge detection. *Proceedings of the Royal Society of London: Series B, Biological Sciences* 207(1167): 187–217.

McKee SP and Mitchison GJ (1988) The role of retinal correspondence in stereoscopic matching. *Vision Research* 28(9): 1001–1012.

Murakami I and Shimojo S (1996) Assimilation-type and contrast-type bias of motion induced by the surround in a random-dot display: Evidence for center-surround antagonism. *Vision Research* 36(22): 3629–3639.

Nawrot M and Sekuler R (1990) Assimilation and contrast in motion perception: Explorations in cooperativity. *Vision Research* 30(10): 1439–1451.

Nothdurft HC (1985) Sensitivity for structure gradient in texture discrimination tasks. *Vision Research* 25(12): 1957–1968.

Petrov Y (2002) Disparity capture by flanking stimuli: A measure for the cooperative mechanism of stereopsis. *Vision Research* 42(7): 809–813.

Polat U, Mizobe K, Pettet MW, et al. (1998) Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature* 391(6667): 580–584.

Polat U and Sagi D (1993) Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments. *Vision Research* 33(7): 993–999.

Polat U and Sagi D (1994) Spatial interactions in human vision: from near to far via experience-dependent cascades of connections. *Proceedings of the National Academy of Sciences of the United States of America* 91(4): 1206–1209.

Pollen DA (1999) On the neural correlates of visual perception. *Cerebral Cortex* 9(1): 4–19.

Ramachandran VS and Cavanagh P (1985) Subjective contours capture stereopsis. *Nature* 317: 527–530.

Rubenstein BS and Sagi D (1990) Spatial variability as a limiting factor in texture-discrimination tasks: Implications for performance asymmetries. *Journal of the Optical Society of America A* 7(9): 1632–1643.

Sagi D and Julesz B (1985) 'Where' and 'what' in vision. *Science* 228(4704): 1217–1219.

Sagi D and Hochstein S (1985) Lateral inhibition between spatially adjacent spatial-frequency channels? *Perception & Psychophysics* 37(4): 315–322.

Sagi D and Julesz B (1987) Short-range limitation on detection of feature differences. *Spatial Vision* 2(1): 39–49.

Stettler DD, Das A, Bennett J, et al. (2002) Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron* 36(4): 739–750.

VanEssen DC, DeYoe EA, Olavarria JF, et al. (1989) Neural responses to static and moving texture patterns in visual cortex of the macaque monkey. In: Lam DMK and Gilbert C (eds.) *From Single Cells to Visual Perception*, pp. 137–154. The Woodlands, TX: Portfolio Publishing.

von Bekesy G (1967) Mach band type lateral inhibition in different sense organs. *Journal of General Physiology* 50(3): 519–532.

Westheimer G (1965) Spatial interaction in the human retina during scotopic vision. *Journal of Physiology* 181(4): 881–894.

Westheimer G (1967) Spatial interaction in human cone vision. *Journal of Physiology* 190(1): 139–154.

Westheimer G and Levi DM (1987) Depth attraction and repulsion of disparate foveal stimuli. *Vision Research* 27(8): 1361–1368.

Contextual Interactions in Visual Processing

A Das, Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

History

Perception: Back to Gestalt from a New Reductionism

Gestalt psychology – an influential theory of perception early in the twentieth century – proposed that perception was determined not by the elemental sensations of light and dark but by laws of similarity, good continuation (analogous to smoothness), closure, symmetry, etc. that grouped such elements within a larger visual context (Figure 1). The founder of this school, Max Wertheimer, first came to prominence with his studies of 'phi motion,' the illusory percept of smooth motion that appears when a spot of light flashes in sequence over a set of discrete points in space; this illusion of motion – which underlies our perception of smooth motion in movies – cannot be reduced to its underlying elemental light flashes.

Two historical exigencies shortened the life of Gestalt psychology in its initial incarnation. First, the rise of Hitler forced Wertheimer and many of the other proponents of Gestalt ideas to flee Germany as refugees. More important, Gestalt ideas could not point to any underlying neurophysiological mechanisms. By contrast, the discovery of apparently elemental neuronal receptive fields in the retina and later in visual cortex starting building a powerful body of physiological and anatomical evidence for a reductionist model of visual perception. Namely, that the visual input was decomposed first into the smallest independent elements of light and dark. These were then built up into independent short lines and so on through a set of progressively more complex but independent features higher along the visual pathway, toward the final percept.

Recent work in the field has revived more holistic ideas of visual perception reminiscent of Gestalt concepts, this time with neurophysiological and anatomical backing. Using perceptual tests and a quantitative measure of curve smoothness, Field and colleagues showed that our visual systems find it progressively easier to pick out 'smoother' curves. This gave a rigorous basis for the Gestalt concept of 'good continuation' and was further backed up with physiological and anatomical results showing selectivity for smoothness in V1 neurons. More generally, quantitative perceptual tests suggest that we appear to parse visual scenes into surfaces, contours, and objects rather than into elemental edges. Our interpretation of objects in the scene then influences our perception of their color, the presence (or absence) of edges, and other nominally elemental aspects of the visual scene. Many of these results are now finding corroborative evidence in physiological results.

Neurophysiology: The Changing Concept of the Receptive Field

Starting with the recordings by Hartline in the limulus retina in the 1930s, visual neurophysiologists described neurons in the visual pathway as having small, independent, spatially delimited receptive fields (RFs). This was exemplified by Kuffler's characterization of small, independent circularly symmetric center-surround RFs in the retina and Hubel and Wiesel's description of small, independent elongated orientation-tuned simple cell and complex cell RFs in primary visual cortex. Notably, the response of each neuron was independent of the stimulation of other neurons with different RFs. These discoveries laid the basis for the influential, reductionist view of visual perception.

This classical description of elemental, independent RFs was a consequence of the types of simple stimuli used by early neurophysiologists – small isolated single spots or bars of light which stimulated neurons vigorously (or suppressed responses sharply) but only when shown within crisp, small regions of space.

There were several early indications, however, that the idea of the spatially limited RF was oversimplified and that responses of neurons to stimuli within their receptive fields could be influenced by distant visual inputs. As early as 1953, Barlow showed that the responses of a particular class of ganglion cells in the frog retina – the 'on–off' cells – were suppressed by spots of light outside their receptive fields. This was distinct from the inhibitory (or excitatory) subfields that Kuffler and others had described for the receptive field, where a small spot of light could, on its own, cause a reduction (or increase) over background activity. Barlow showed that a spot of light that was incapable of changing an on–off neuron's firing pattern on its own – thus, by definition, lying outside the classically defined receptive field – could entirely suppress the response to a second spot of light lying inside the receptive field. In 1964, McIlwain showed that the responses of (cat) retinal ganglion cells to stimuli inside their RFs increased when moving a stimulus (e.g., a small black disk) up to 90° away in the periphery. In 1965, Hubel and Wiesel

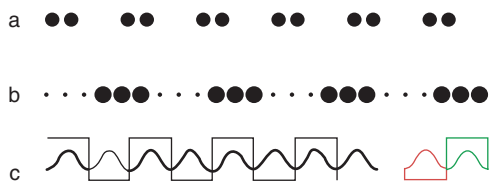

Figure 1 Some Gestalt principles used by our visual system to group simple elements into visual objects: (a) proximity (grouping the dots into nearby pairs); (b) similarity (grouping the dots into alternating sets of small and large dots); (c) good continuation (we see this figure as a sinusoid and a square wave, not as alternating pairs of the more complicated objects shown in red and green to the side).

described the phenomenon of 'hypercomplexity' (now known as end-stopping) in cat area 18, where the response of a neuron to its preferred bar stimulus could be suppressed entirely by a second, collinear bar outside the neuron's receptive field.

Evidence accumulated through the 1970s and 1980s established this basic parsing of the input receptive field of neurons in all visual areas – a small classically defined RF where stimuli were capable of driving neuronal responses directly, and a surround where stimuli could suppress or facilitate the neuron's response to a stimulus in its classical RF. During this period, Blakemore and Tobin, Nelson and Frost, and later Allman and colleagues showed that the modulatory surrounds of RFs could be characterized by their tuning for properties similar to the properties of the 'classically' defined RFs ('CRFs'). Their studies using a particular class of surround stimuli – bars or gratings or moving dots in an annulus outside the RF – showed that the influence of the surround was typically strongest (most inhibitory or most excitatory) when its features (orientation, spatial frequency, and direction of motion) matched the features of the optimal stimulus inside the neuron's classical RF.

This parsing into a classical RF and modulatory surround is not rigid and some authors prefer to define a neuron's RF as encompassing the entire complex of responses, both direct and modulatory. Furthermore, the bounds separating the core of an RF from its surround are not fixed but depend on the stimulus being used. Bright, high-contrast stimuli induce neuronal responses with smaller, crisper cores, whereas low-contrast stimuli or stimuli surrounded by busy textured surrounds can drive the same neurons from core RF regions that extend out much further. The distinction between core and surround in a receptive field, although not precise, is still useful and is likely to reflect the neuroanatomical distinction between the feed-forward (or local) inputs to a neuron versus the inputs from longer range connections and feedback from higher visual areas.

Neuroanatomy: The Substrate underlying Contextual Influences

Since the 1980s, new, sensitive anatomical tracing techniques have made it possible to see precise, long-range anatomical interconnections both within and between cortical regions, giving a plausible anatomical basis for contextual interactions in vision. In 1979, Gilbert and Wiesel showed that pyramidal neurons in V1 not only make local connections within their cortical columns (over a few hundred micrometers) but also send out axonal collaterals that extend for many millimeters and connect with distant neurons in discrete and well-separated clumps. Later work has shown that these long-range connections are precise and selective for neuronal response properties (Figure 2). Thus, neurons selective for vertical lines, for example, at one point of visual space, send out axonal connections that skip over other orientations to connect selectively with distant cortical columns of other vertical-preferring neurons with receptive fields at a different point in visual space, up to 10 RF diameters away. This anatomical network provides a plausible substrate for orientation-tuned contextual interactions. Thus, a neuron responding to a vertical edge in its RF will receive long-range axonal input from other neurons responding to more distant vertical edges (but not other orientations) in their RFs. This axonal input could be facilitatory, from direct excitatory pyramidal-to-pyramidal connections, or it could be inhibitory since the long-range connections also target local inhibitory interneurons.

V1 also receives extensive and widely divergent feedback from higher visual areas, with single locations in the higher areas sending connections over V1 regions extending many millimeters (i.e., the same lateral extent as the intrinsic axonal collaterals). This feedback can be very rapid. The fibers interconnecting different visual areas are myelinated and fast, unlike the slower unmyelinated fibers within a cortical region. In 1996, Movshon and Newsome showed that the transit time between V1 and area MT, a cortical area specializing in visual motion processing one level beyond V1, is only 1.4 ms. Feedback from higher areas contributes a significant component to the contextual modulation of neuronal responses in V1. Bullier and colleagues showed that a number of contextual modulatory responses, even ones with short latencies, are considerably reduced or even eliminated on silencing higher visual areas either by cooling or with the inhibitory neurotransmitter GABA. Most of these feedback connections lack the precise orientation-specific patterns characteristic of intrinsic connections. They are thus believed to mediate overall attentional or task-dependent

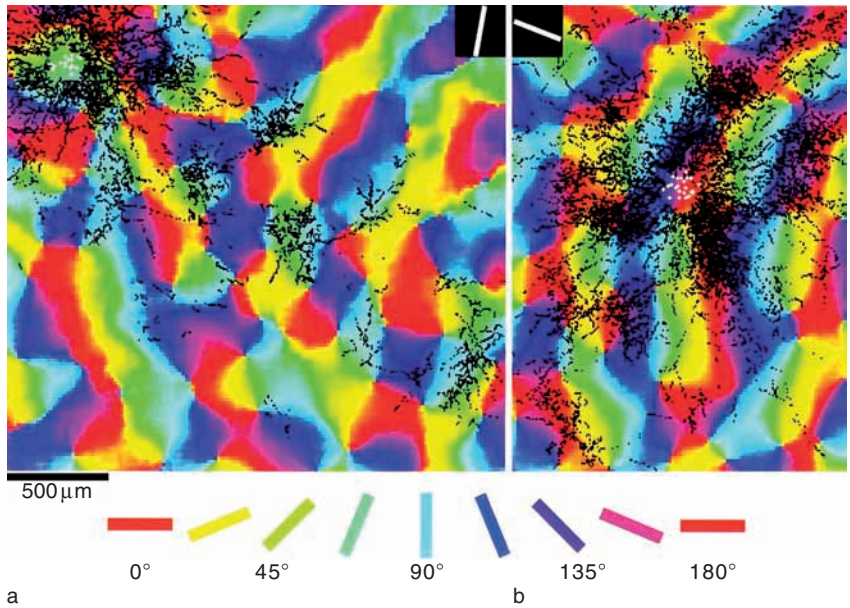

Figure 2 The anatomy of axonal arbors radiating out from V1 pyramidal neurons, shown superimposed on an optically imaged orientation map from the same region of V1 (in tree shrew). The optical image is color-coded to identify the orientation preference of each column (key below). The anatomical trace was obtained by first optically imaging the animal's V1, injecting a neural tracer into a selected orientation column and processing the tissue after allowing the tracer to spread along axonal arbors. In the immediate vicinity of the injection site (white dots: cell bodies that have taken up the tracer) the axonal arbors spread uniformly, independent of local orientation. At more distant sites, however, the axonal arbor forms clumps of connections with other neurons in columns of the same orientation as the injection site, skipping over intervening cortical columns of different orientations. Reprinted from *Journal of Neuroscience* 17(6): 2112–2127. Bosking WH, Zhang Y, Schofield B, and Fitzpatrick D: Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. (1997).

modulations rather than orientation-tuned contextual modulations in V1.

The extent and sources of these long-range inputs stand in contrast with the mostly local connections and feed-forward inputs that are believed to determine RF cores. Thus, the input whose geometric alignment determines the core elongated shapes and orientation tuning of V1 neurons is mostly feed forward from lower areas (e.g., lateral geniculate nucleus (LGN)) combined with connections within the cortical column (i.e., extending a few hundred micrometers).

The Role of Contextual Modulation

The patterns of surround-induced modulation discovered in physiological studies during the 1980s were interpreted as mechanisms of adaptation or of sharpening the discrimination among features in the visual input. The typical stimulus consisted of a center at the preferred orientation or direction of motion of the neuron and a surround consisting of a uniform background pattern similar to the center but with variable orientation or direction. Such surrounds typically induced suppression or facilitation tuned to the angular difference between center and surround. The overall response was weakest when the surround blended with the center, and it was strongest when the surround was antagonistic. These center–surround interactions were interpreted as a way of accentuating regions of visual contrast while suppressing regions of uniform visual texture. They were also believed to be mechanisms for perceptual constancy; for example, a means of integrating (and thus discounting) the color over the background, allowing us to see objects in their true color independent of the illumination.

Work since the 1980s has uncovered a much larger and more complex repertory of neurophysiological center–surround interactions. Neuronal responses to contextual stimuli seem well matched to parsing visual scenes into perceptually relevant midlevel features – into smooth contours, the inside versus the outside of objects, foreground versus background, identifying occlusion of one object by another, etc. These broad classes of contextual interactions are laid out in the following sections.

Grouping Lines into Smooth Contours

The Gestalt idea that the visual system is sensitive to smooth contours (the principle of 'good continuation') has been borne out by quantitative perceptual

as well as physiological experiments. Using contours consisting of disjointed line segments, Field, Hayes, and Hess showed that the ease with which a contour could be detected in a noisy background was related, quantitatively, to its smoothness. Subjects' detection performance decreased monotonically as the contours were made more sharply curved or had increased angular or positional jitter between the line elements. Field and colleagues quantified this interaction by introducing the concept of an 'association field' tuned for smoothness around each visual line element (Figure 3).

Our sensitivity to smoothness in contours may be mediated through the mutual facilitation between V1 neurons responding to the individual elements of the smooth curve. Kapadia showed, for example, that a (monkey) V1 neuron responding to a line inside its RF was strongly facilitated by flanks lying outside the classical RF. The degree of facilitation was strongest for flanks that were collinear and nearby, and it diminished as the flanks were drawn further apart, to a side, or at an increased angle. In a parallel perceptual study in which subjects were asked to detect a barely visible low contrast line, the ease of detection improved almost twofold with a collinear flank; the improvement decreased monotonically as the flanking line was moved away, as in the physiological experiments (Figure 4).

The close match between the physiological pattern of facilitation, the perceptual association field, and the anatomical layout of axonal collaterals in V1 (Figure 2) suggests that V1 intrinsic connections may mediate contour detection; findings by Grinvald of orientation-tuned feedback from V2 open the possibility, however, that feedback from higher areas also plays a role.

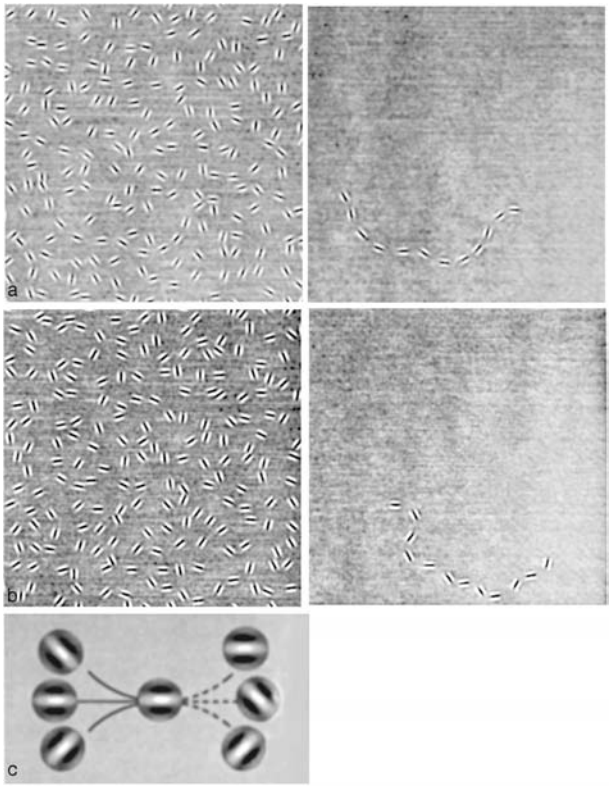

Figure 3 (a) A smooth contour with 30° jumps between adjacent elements (right) is easy to pick out of a complex background (left). (b) A contour with 60° jumps between adjacent elements (right) is perceptually less 'smooth' and is much harder to see against a background. (c) The proposed association field mutually facilitates the perception of line elements that lie on smooth contours relative to each other (left) while suppressing the perception of elements that are orthogonal or displaced (right). Reproduced from Field DJ, Hayer A, and Hess RF (1993) Contour integration by the human visual system: Evidence for a local "Association Feild". *Vision Research* 33(2): 174–193, with permission from Elsevier.

Completing Occluded Surfaces

Smooth curves broken up into short line segments are a special case of a feature that is common to everyday perception. Namely, that in natural environments we rarely see objects in their entirety. They are generally partially occluded by other objects in front of them. Yet, we (and other animals with sophisticated visual systems) have no problem perceptually interpolating and identifying object boundaries behind occluders – reaching, for example, for a fruit obscured by a branch. This process of identifying occluded objects is believed to involve the intermediate step of assigning 'border ownership': when one object is seen as lying in front of another, the visual boundary between the two is perceived as 'belonging' to and defining the outlines of the object in front. The object in the background is seen as continuing behind the occluder.

The process of identifying occlusion and border ownership was believed to be largely bottom–up and hierarchical. Our visual systems were believed to first identify local cues suggesting the presence of one surface lying across another: 'T' junctions, corners, and disparity (i.e., depth) ordering between surfaces. These cues were then used to assign border ownership locally (T junction means that the straight edge lies on top), building up a sequential process of identifying surfaces and their boundaries (Figure 5).

Quantitative perceptual work, for example, by Nakayama and Shimojo, suggests that the process is not entirely bottom-up. Rather, we form intermediate hypotheses about surfaces and objects early in the perceptual process, which then influence our perception of local features. By presenting visual stimuli dichoptically (i.e., with one image to the left eye and a slightly shifted image to the right eye so as to give

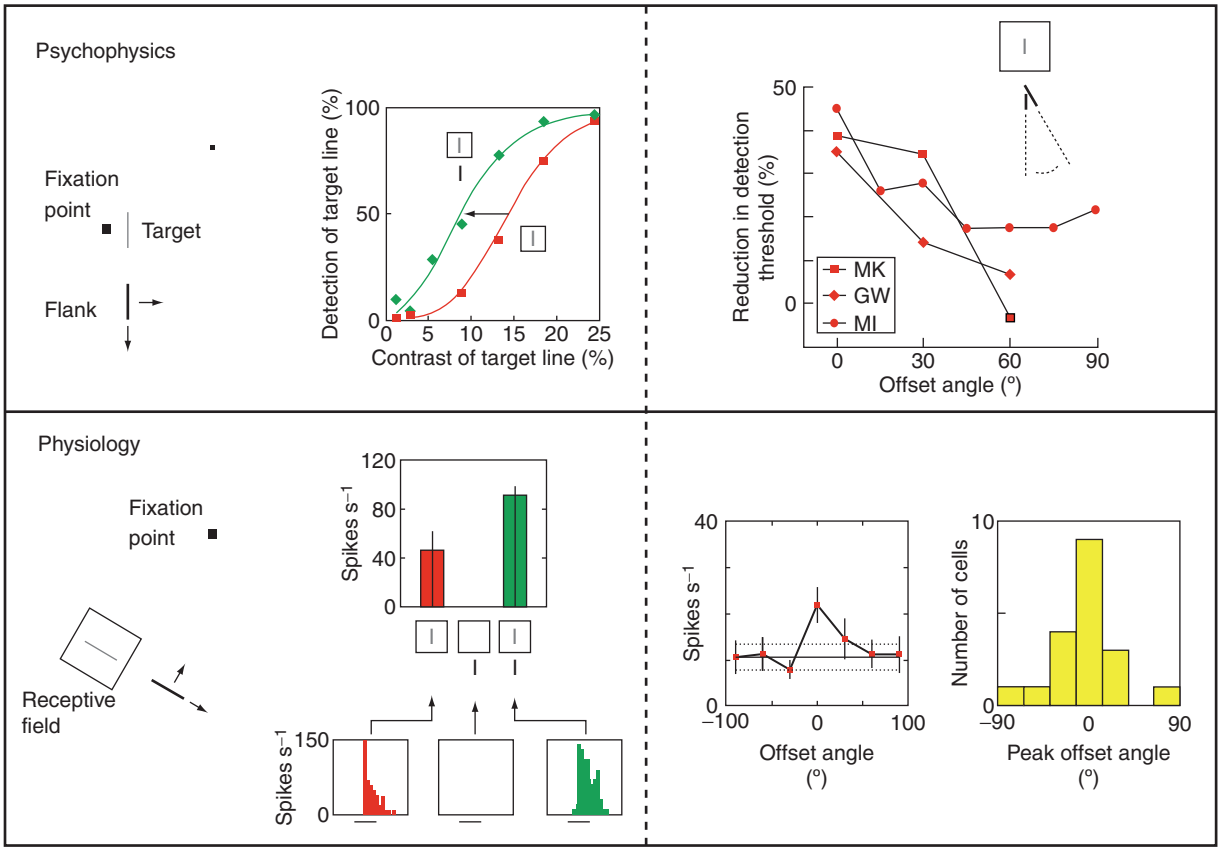

Figure 4 Comparing perceptual results for detecting a low-contrast line (in human subjects) with the physiological responses of V1 neurons to a similar line (in monkeys). Top left: Human subjects were asked to detect a low-contrast line ('target') that was flashed on while the subject maintained steady fixation. The subjects' detection performance as a function of target contrast (red curve) improved when the single low-contrast line was flanked by a collinear flank (green curve). For the closest flank position, the threshold for detection is reduced to roughly 50%. Bottom left: Responses of a monkey V1 neuron to a low-contrast line flashed in its RF (square) while the animal maintained steady fixation. The response to the line alone (red) was facilitated by a second line outside the RF (green) even though the second bar did not stimulate neuronal responses by itself (empty box showing line outside the RF). Top right: The threshold for human subjects' ability to detect a dim line improves 50% with a collinear flank; this improvement reduces steadily as the flank is rotated away from collinear. Bottom right: The facilitation by a collinear flank of a monkey V1 neuron's response drops as the flank is rotated away from collinear. Reproduced from Kapadia MK, Ito M, Gilbert CD, and Westheimer G (1995) Improvements in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. *Neuron* 15(4): 843–856, with permission from Elsevier.

the percept of objects in three dimensions), they showed that simply switching images between the left and right eye dramatically changed our interpretation of a scene. With an object seen as lying behind an occluder, we automatically 'complete' the occluded object with perceptually salient continuous boundaries. Just switching the order of images so that the image on the left now gets presented to the right eye and vice versa turns occluders lying above the plane of the object into holes extending below. Despite the presence of the same local information (T junctions, corners, etc.), switching images breaks the foreground object into discontinuous strips that are much more difficult to identify, even eradicating the percept of continuous edges, a nominally 'elementary' percept.

These perceptual results are supported by physiological results showing that the grouping of lines into smooth contours – the facilitation of neuronal responses by collinear flanks – is not automatic. Rather, it depends on being able to interpret the gaps between lines as occluders. In V1, a small fraction of neurons respond thus to illusory lines lying behind occluders (Figure 6). In V2, on the other hand, a large number of neurons (possibly 50%) have such a disparity-selective response to the presence of collinear flanks (Figure 7).

Explicitly Identifying Border Ownership

Several studies by Rudiger Von der Heydt and colleagues have identified neurons at various stages along the visual pathway – in V4 and V2 and, to a limited

extent in V1 – that signal border ownership *per se*. That is, the response of a neuron to an edge at its preferred orientation, running through the neuron's receptive field, depends on whether the edge 'belongs' to an 'object' or 'foreground' lying on one side or the other of the neuron's RF (Figure 8). The neuron's response does not depend on lower order features of the edge such as contrast polarity (i.e., whether the edge is white against black or black against white). These neurons appear capable of generalizing across objects cues; that is, whether objects are defined through solid blocks of color or through disparity shifts in random dot patterns. Note that in all cases the distant boundaries marking whether the object lies on one side or the other of the neuron's RF lie far outside the neuron's classical RF; these neurons are capable of spatially integrating contextual cues that extend at least 20° (i.e., more than 10 times the diameter of the classical RF). Moreover, the neuronal response difference signaling 'object to one side' versus 'other side' emerges at a latency no greater than the initial response of the neuron, suggesting a rapid processing of the contextual information regarding border ownership (Figure 8), thus indicating that the processing occurs intracortically within the same cortical region rather than involving feedback from higher areas.

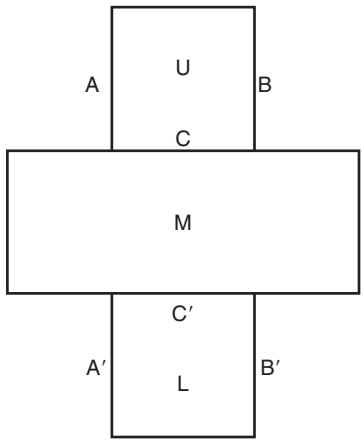

Figure 5 A simple scheme for extracting depth order amongst occluded surfaces using T junctions: the junctions, for example between lines A and C or B and C, help identify this as a horizontal bar (M) lying atop a vertical bar (U + L).

Figure versus Background

The responses of neurons even in V1 are modulated depending on whether their RFs lie inside 'figures' or

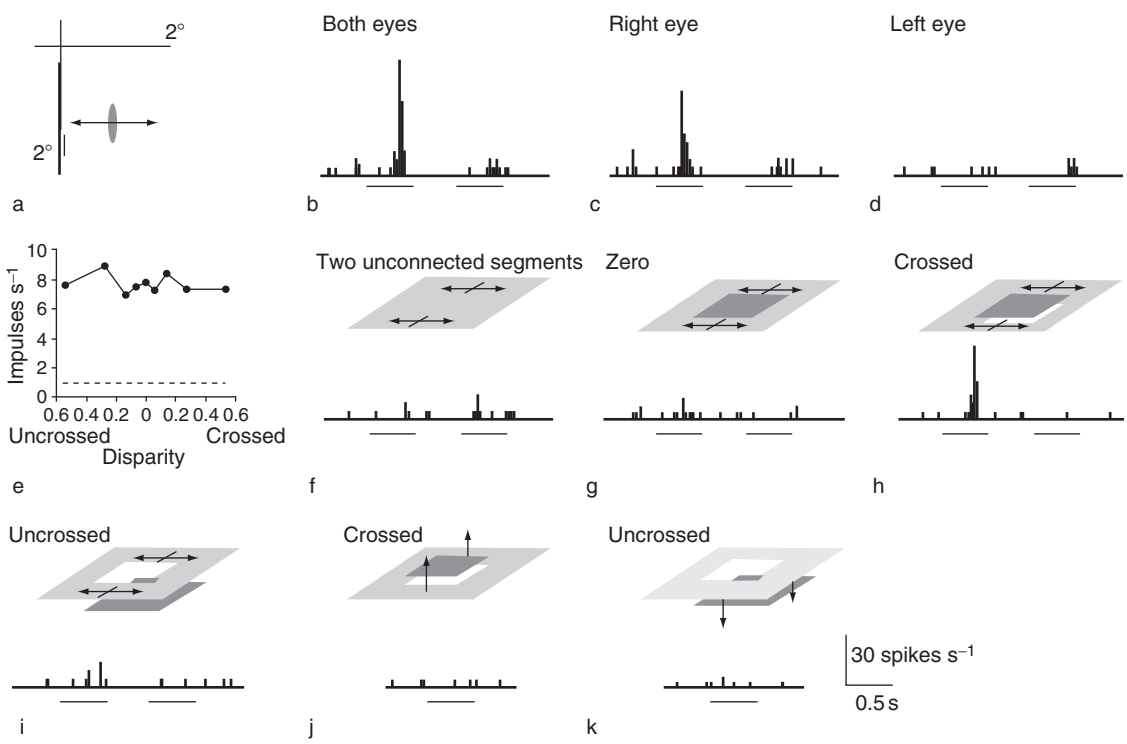

Figure 6 Neurons in monkey V1 respond to a line that appears to be completed behind an occluder. The neuron (RF shown in (a) responses to a line in its RF in (b–d)) shows no response to an unconnected pair of lines outside its RF (f) but does respond when simultaneously shown a patch that could be interpreted as a foreground surface occluding a continuous line underneath (h). Controls (g, i, j, k) to show that the cell respond only to a stimulus geometry that is consistent with an occluder lying above a line. Reprinted from *Nature* 401: 269–272. Sugita Y Grouping of image fragments in primary visual cortex (1999).

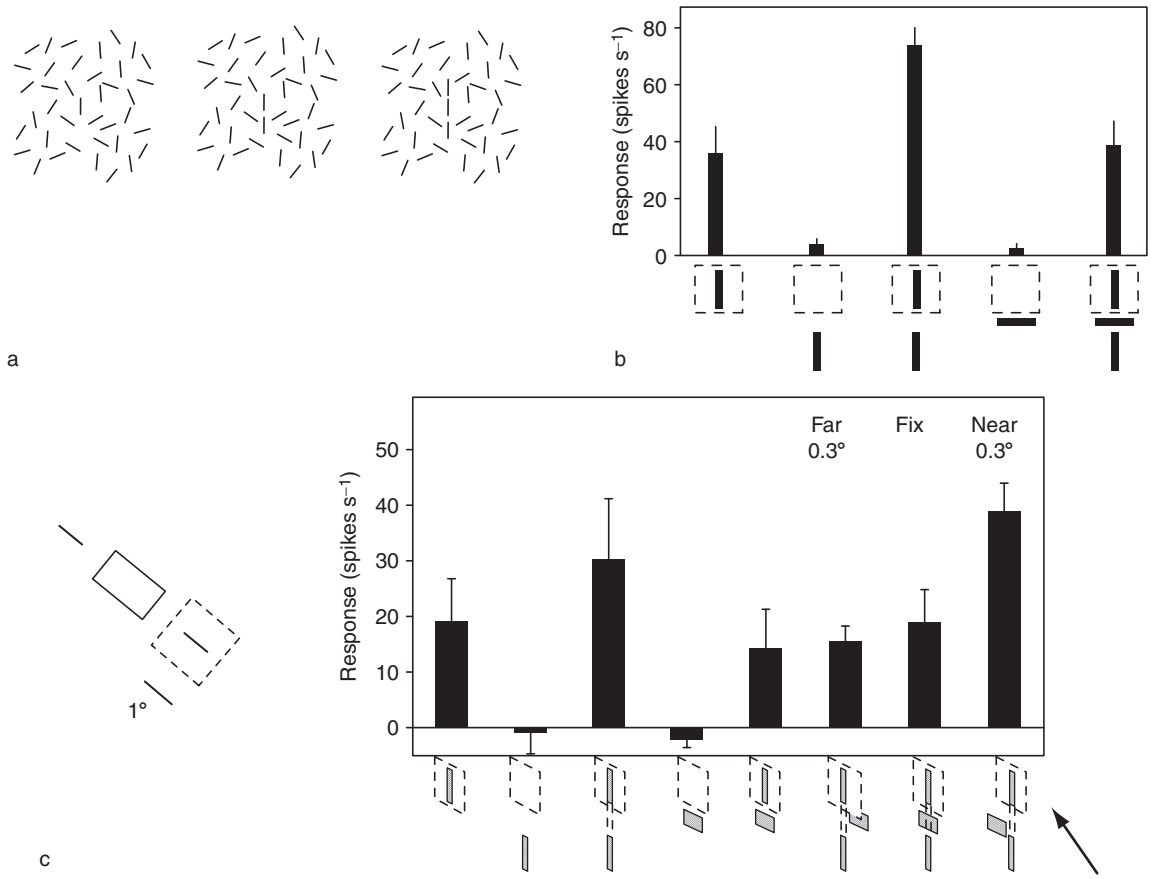

Figure 7 Collinear lines provide facilitation only when they lie on the same plane and are seen as potentially continuous. A V2 neuron's response to a single line (vertical center line) in a complex background (a) improves when flanked with collinear lines on the same depth plane (b). This improvement is 'vetoed' by an orthogonal line on the same depth plane (b, extreme right) but returns when the perpendicular line is placed on a different depth plane consistent with being an occluder lying above a continuous line (arrow, last item in c). Reproduced from Bakin JS, Nakayama K, and Gilbert CD (2000) Visual responses in monkey areas V1 and V2 to three-dimensional surface configurations. *Journal of Neuroscience* 20(21): 8188–8198, with permission.

on the 'background.' In some of the earliest studies of neurophysiological correlates of figure–ground separation, Zipser and Lamme created (flat, two-dimensional) rectangular figures distinguished from their backgrounds by some simple visual contrast – a pattern of dots moving in one direction in the figure versus the opposite direction in the background, or a pattern of lines tilted in one direction in the figure versus the orthogonal tilt in the background. With V1 neuronal RFs lying on the figure boundary, or at various distances away either into the figure or into the background, the neurons showed a consistently higher delayed response inside the figure versus outside on the background (Figure 9). Unlike the 'border ownership' response modulation, the neuronal response signaling 'figure versus Background' arrives with a considerable latency of 50–100 ms after the initial onset of the neurons' responses to visual stimulation. Furthermore, this neuronal response difference is seen only when recording from alert behaving animals and disappears under anesthesia. These observations suggest that the figure versus background contextual modulation is mediated through feedback from higher areas.

Border Ownership and Visual Motion Perception

The assignment of boundary ownership can radically alter our perception of motion. A line seen moving behind an aperture appears to move perpendicular to its edge even though its actual direction of motion may be quite different, a perceptual confound that gives rise to the classical 'barber pole' illusion. The critical feature of such a visual scene, as pointed out by Nakayama, is that the ends of the line be interpreted as being occluded by the aperture as opposed to 'belonging' to the line. On the other hand, a moving line seen in the foreground has ends that clearly belong to the line and there is no ambiguity about the direction of motion of the line.

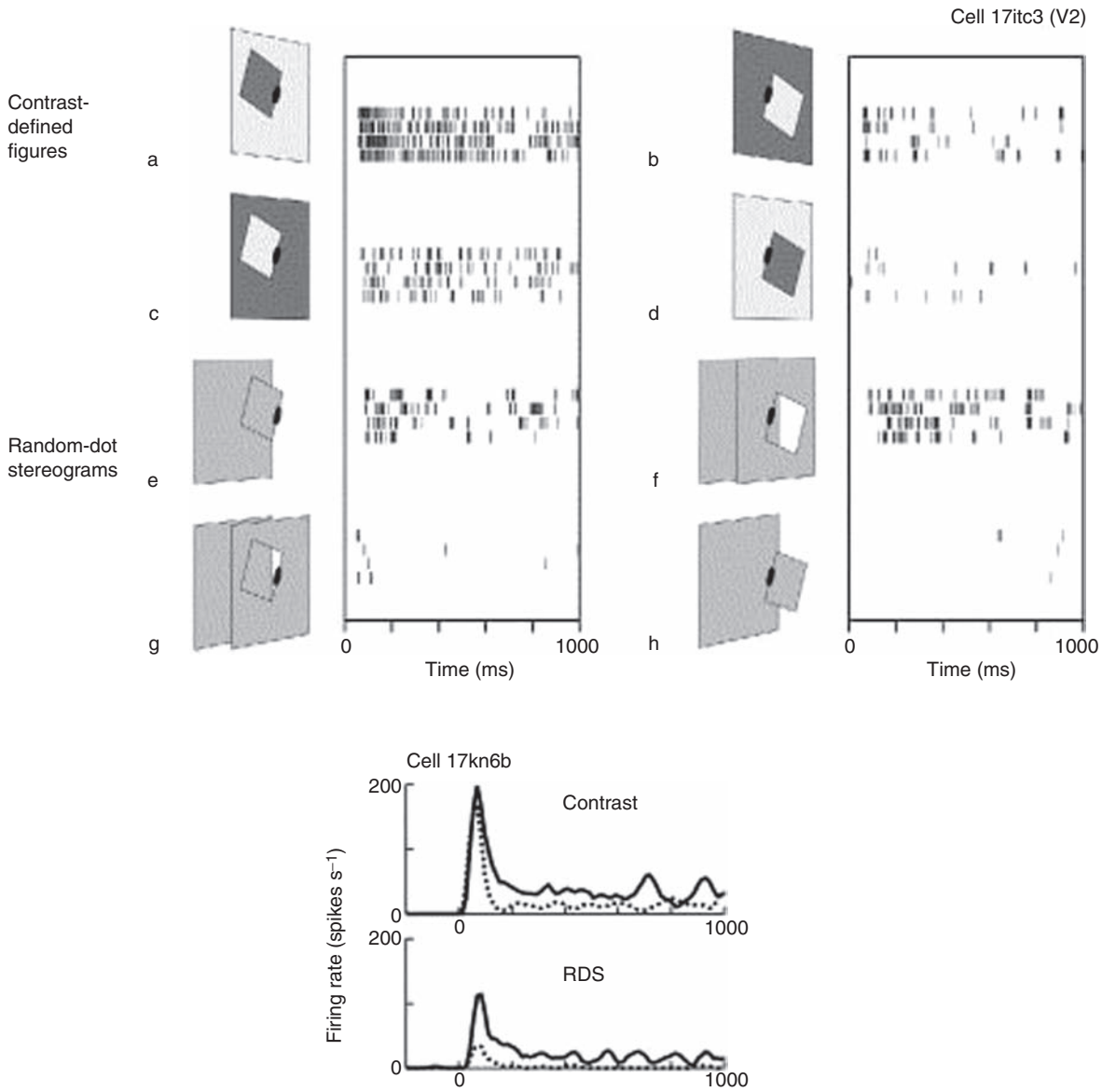

Figure 8 Cell (monkey V2) responds to 'foreground' versus 'background' independent of whether the 'foreground' is defined by contrast or by disparity. Top (a, c): Cell responds vigorously when a 'foreground' object defined as a square of contrast different from the background lies to the left of the RF. Note that the boundaries of the object are well beyond the boundaries of the cell's RF. (b, d) There is no response when the foreground object lies to the right of the RF even though the local visual stimulus running through the cell's RF is the same (compare a, b and c, d). (e–h) The visual stimulus here consists of a random dot stereogram, presented dichoptically to the two eyes with different disparities over the image such that a portion of the image appears to float above a background. The cell responds vigorously when the floating foreground lies to the left of the border running through the RF (e, f), but not when it lies to the right of the RF (g, h). Bottom: Comparing the response of another V2 neuron to foreground (continuous line) vs. background (dotted line) lying on one side of the RF, whether defined by contrast or by disparity (RDS: random dot stereogram). Reprinted from *Neuron* 47: 155–166. Qiu FT and von der Heydt R Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules (2005).

Albright and colleagues used this insight to construct a 'barber diamond' in which they could vary the perceived direction of motion by changing the depth order of the surfaces surrounding the diamond. The pair of bounding surfaces (e.g., top right/bottom left) positioned in front of the moving lines were seen as an aperture occluding the lines, whereas the other pair, behind the moving lines, became the background. The motion of the line ends seen against the background boundaries then determines the perceived direction of motion of the lines (Figure 10). Switching the depth order of the surfaces switches

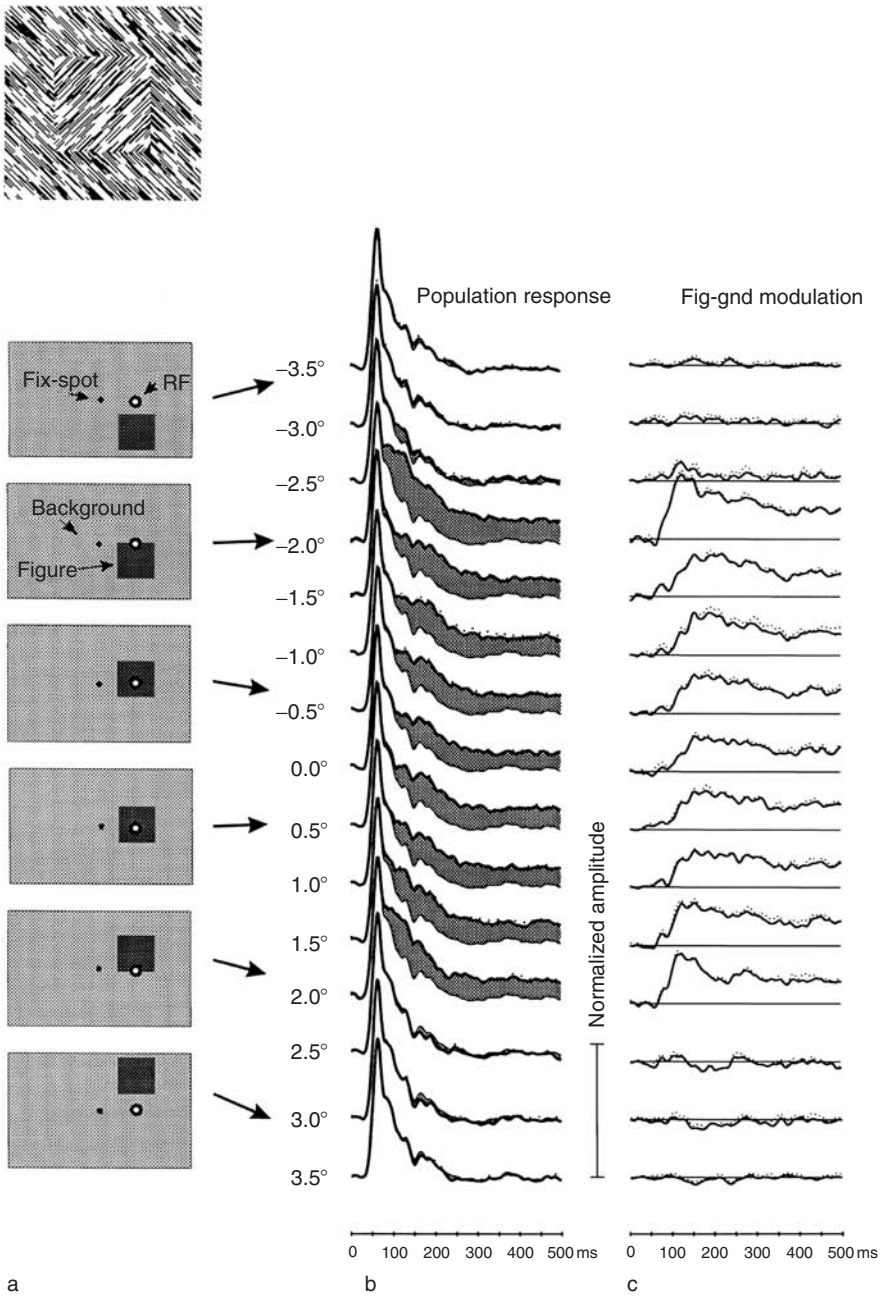

Figure 9 V1 neurons respond with different temporal dynamics depending on whether their RFs lie on the foreground, background, or border of a figure. Top: Figure versus background defined by a texture of oriented lines. Bottom: (a) The figure was flashed on at different locations relative to the neuron's RF while the monkey maintained fixation. (b), (c) At RF locations lying within the figure, the neurons showed a delayed sustained activity that was higher than when the RF lay on the background. When the RF lay on the figure boundary, the delayed response was stronger and had a shorter latency. Reprinted from *Cerebral Cortex* 9(4): 406–413. Lamme VAF, Rodriguez-Rodriguez V, and Spekreijse H Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. (1999).

'aperture boundary' and 'foreground' and changes the perceived direction of line motion. Neurons in macaque monkey area MT, a visual cortical area that is believed to be critical to the perception of visual motion, switched their response in accordance with the perceived shift of line motion (Figure 11). Not only does this show that neurons in MT can integrate contextual information lying well outside their classical RFs but also it is an example of an 'intermodal' interaction in which information about

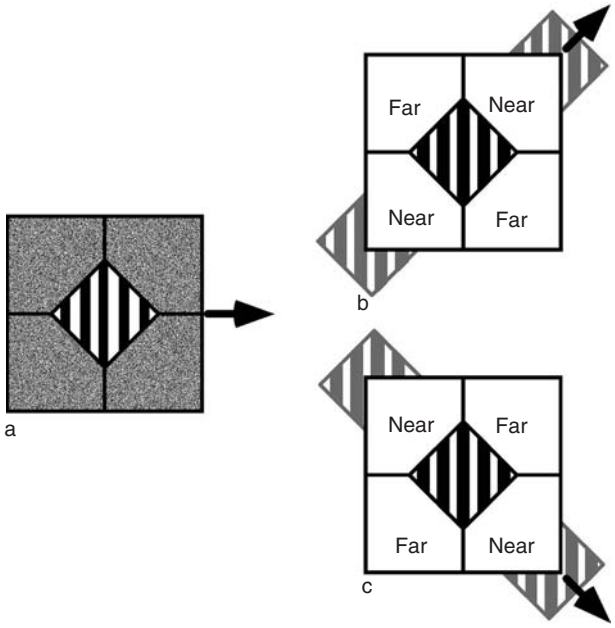

Figure 10 The 'barber diamond' illusion. With vertical stripes moving horizontally in the diamond-shaped aperture, (a) the perception is also, typically, of an object moving horizontally seen through an aperture in an occluding screen. When the visual stimulus is presented dichoptically, however, with diagonal pairs of quadrants floating above and behind the vertical stripes, (b,c) the percept changes dramatically. With the disparity arranged so that the top-left/bottom-right are far while the top-right/bottom-left are near, the percept is clearly of a rectangular strip moving diagonally to top right as shown over the far surface; the angle of the diagonal motion switches on switching the order of the far/near quadrants (c). Reproduced from Duncan RO, Albright TD, and Stoner GR (2000) Occlusion and the interpretation of visual motion: Perceptual are neuronal effects of context. *Journal of Neuroscience* 20(15): 5885–5897, with permission.

depth (disparity) influences response tuning to motion through an intermediate interpretation of the scene into a set of surfaces.

Filling 'Holes' in the Visual Stimulus

Perceptually, the visual system is able to bridge holes or scotomas in the visual field. This is the case for the pair of holes present naturally in all individuals – the blind spot, or exit point of the optic nerve in each eye, a region approximately $5 \times 8°$ in size ('large enough to conceal 76 full moons') lacking photoreceptors and hence any visual input in that eye. Not only do we not notice the hole – even monocularly, shutting one eye at a time – but also visual objects straddling the blind spot are seen as continuous even if they have gaps that fall within the blind spot. Mechanisms of visual fill-in are present not only at the blind spot – which could be special, being part of the visual system from birth – but also appear to be a dynamic property of neurons in adult visual cortex. Ramachandran and Gregory simulated the blind spot with 'artificial scotomas' – visual stimuli that are busy everywhere except for a small patch of neutral gray or black. Maintaining fixation steadily on a point in the stimulus bleeds in the surrounding color or texture, perceptually filling the 'scotoma.'

There are likely physiological correlates of visual fill-in at multiple levels of the visual cortex. In (anesthetized cat or monkey) V1, RFs expanded, reversibly, when presented with artificial scotomas. This expansion appears to be mediated by increased coupling among neurons with nonoverlapping RFs but similar orientation preferences – thus likely mediated by the orientation-tuned long-range axonal collaterals in V1. In V2 and V3 (but not V1) in alert rhesus monkeys, neurons with RFs within similar artificial scotomas also start responding as though expanding to include the initially distant texture.

Signaling Surface Luminance

The perceived color and luminance of a surface can often be very different from their physical correlates. Thus, if the spectral distribution of light reflected off a surface (as measured using spectrophotometers) remains constant while changing the hue or brightness of a distant surround, we can get a dramatic percept of the surface color or brightness changing in the 'opposite' direction. In real life, this leads to 'perceptual constancy,' in which we get a largely veridical sense of the intrinsic color and reflectance of surfaces independent of the illumination.

In the visual pathway, neurons that could signal perceived brightness are first encountered in V1 but not earlier. In the retina and LGN, neuronal RFs largely transmit local information on brightness or color differences; that is, signaling the presence of an edge and which side of the edge is darker. Responses are not modulated by the global modulation of the brightness of distant surrounds. Paradiso and colleagues showed that in V1, on the other hand, neurons with RFs on a surface of uniform and constant brightness modulate their responses with the modulation of distant surrounds, integrating luminance information from regions of visual space more than ten RF dimensions away.

Perceptual Learning and Task-Dependent Changes in Contextual Influences

Perceptual training on new visual discrimination tasks can lead to distinct improvements in our

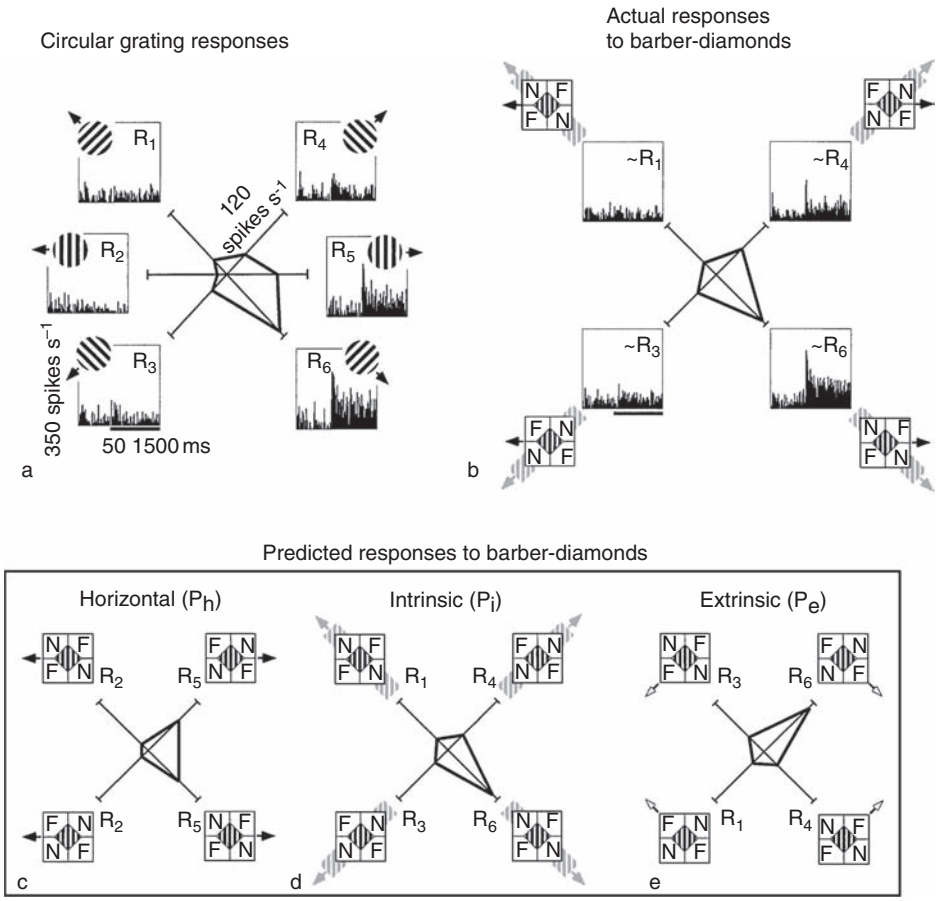

Figure 11 Neurons in monkeys, area MT respond to perceived motion, as determined by distant contextual cues, rather than local motion. (a) Response of the neuron to circular gratings: the neuron is tuned to motion to the bottom right. (b) Response of the neuron to 'barber diamond' motion. The neuron responds when the perceived motion is to the bottom right. (c–e) Predictions of what the neuronal response would have been under different conditions. (c) If the neuron were responding only to local motion of objects in its classical receptive field it would see a 'horizontal' motion for all possible arrangements of occluder quadrants, giving a sub-optimal response to rightward moving stimuli vs. left. (d) If responding to the 'barber diamond' motion (as determined by a human observer) the neuron would respond when the occluder quadrants are arranged to give perceived motion to the bottom right. (e) If the neuron were responding to the line ends moving behind the near-occluder boundary it would respond to top right ('extrinsic'). Clearly the neuron responds to the perceived motion as in (d). Reproduced from Duncan RO, Albright TD, and Stoner GR (2000) Occlusion and the interpretation of visual motion: Perceptual are neuronal effects of context. *Journal of Neuroscience* 20(15): 5885–5897, with permission.

visual abilities. This is true not only for complex stimuli (e.g., learning to remember and discriminate between new faces) but also for simple visual tasks (e.g., learning to discriminate the tilts of different lines). Perceptually, improvements in low-level tasks do not transfer well; learning to discriminate between near-vertical lines lying in one quadrant of visual space will not transfer, for example, to lines at other orientations or other locations in visual space. This suggests that perceptual improvement in such low-level tasks leads to changes in early visual areas (V1, V2, etc.) with relatively small and orientation-selective RFs. By contrast, learning to recognize faces probably involves changes in high-level areas such as inferior temporal cortex with face-selective neurons that have large RFs.

Training on tasks of intermediate levels of complexity – learning to discriminate some simple geometric shapes – leads to dramatic, task-dependent changes in the contextual modulation responses of V1 neurons. Gilbert and colleagues showed that training on simple geometric patterns leads to very specific pattern-dependent improvements; training on one geometric pattern (a 'bisection task' – whether the middle line in a triplet of parallel

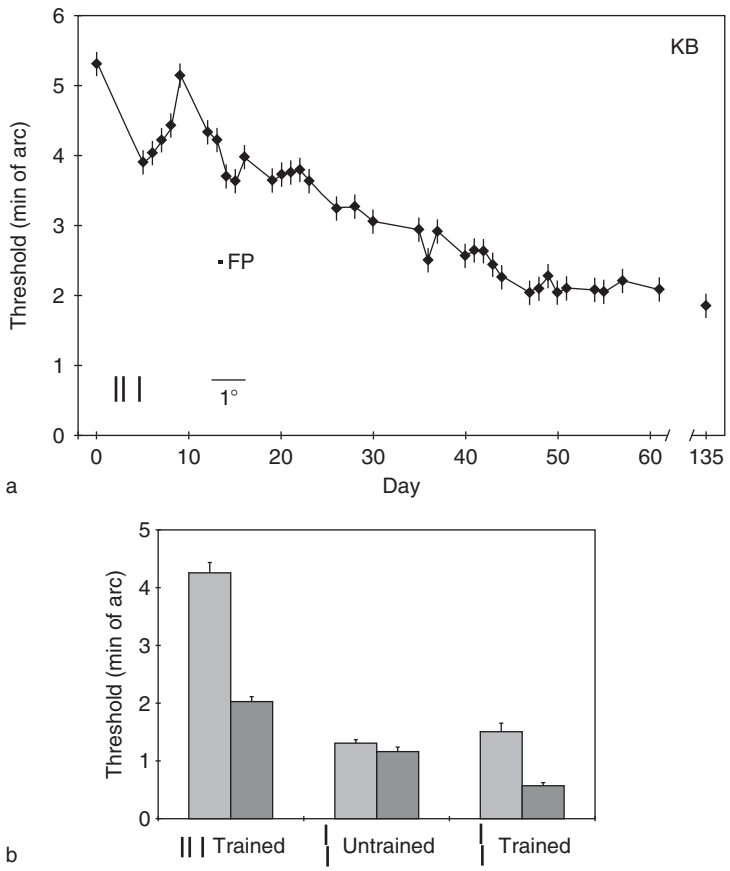

Figure 12 The visual system shows long-term task-selective improvements in simple perceptual discriminations. (a) Human observers trained on a bisection task — determining whether the center bar of a vertical triplet was offset to one side or the other of the center position — improve over 2 months of training by more than a factor of 2. (b) This improvement is task-specific and occurs only for the vertical bisection task (left, as indicated by icon below figure) but not for a vernier task on which the subject had not been trained (middle; determining whether two lines were offset to the right or left relative to each other), even though the two discrimination tasks are very similar and are presented at that same position in visual space. This task selectivity is not because subjects are already at their best for the vernier task: training specifically on the vernier task leads to specific improvement in the task (right). Reprinted from *Journal of Neurophysiology* 78: 2889–2894. Crist RE, Kapadia MK, Westheimer G, and Gilbert CD Perceptual learning of spatial localization: Specificity for orientation, position and context (1997).

lines is closer to one side or the other) does not transfer to other very similar patterns (a 'vernier' task – detecting small lateral shifts between two almost collinear lines; Figure 12). They then trained monkeys on the same pair of tasks, showing the animals the same stimulus for both tasks but cueing the animals to perform either the bisection or the vernier match. After the animals had learned to perform the geometric tasks, neurons in V1 showed a task dependence in their contextual modulation. With the same set of lines surrounding the neuron's RF, the response of the neuron depended on the specific geometric match in which the animal was engaged, implying that intracortical interactions between the different line elements were being further modulated by higher order feedback about the task (Figure 13).

Conclusion

An earlier reductionist view of visual processing that involved breaking scenes into independent elements (tiny independent edge segments) and building the percept hierarchically along parallel and independent channels (depth, color, and motion) is giving way to a view in which visual elements interact to give information about surfaces and objects early in visual processing. The strong perceptual and physiological evidence for such contextual interactions is backed up, in some cases, with plausible underlying anatomy. It is likely that this visual processing has evolved to match the statistical properties of natural scenes, and, in particular, to detect the types of visual object properties (object boundaries, textures, and occlusion of one object by another) that are important in interpreting natural scenes.

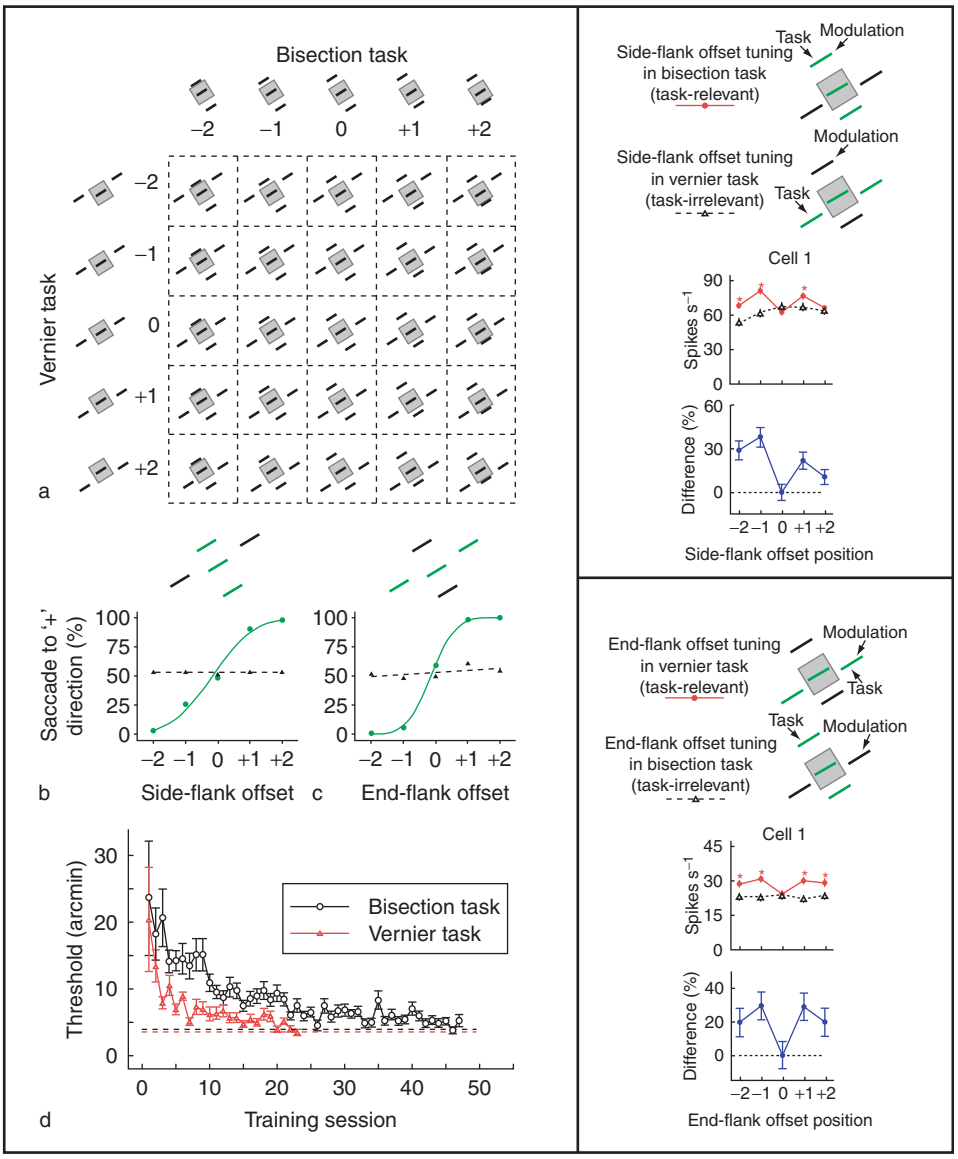

Figure 13 Neurons in monkey V1 respond differentially to the same visual stimulus depending on task context. Left: (a) Monkeys were trained on both a bisection task and a vernier task using the same visual stimulus (a set of five bars with the center bar at various offsets relative to the surrounding four bars). The color of the bars cued the animal as to whether to look at bisection offsets (b) or vernier offset (c). The animals improved on both tasks over training sessions (d). Right, top: The pair of side flanks gave a strong, separation-tuned modulation of a neurons's response when the animal engaged in a bisection task (where the flanks were relevant to the task) vs. when engaging in a vernier task (where the same side flanks at the same positions were irrelevant to the task). Right, bottom (Different neuron): the end flanks gave a strong, position-tuned modulation when the animal was engaged in a (relevant) vernier task but not during the (irrelevant) bisection task. Reprinted from *Nature Neuroscience* 7(6): 651–657. Li W, Peich V, and Gilbert CD Perceptual learning and top-down influences in primary visual cortex (2004).

See also: Contextual Interactions in Visual Perception; Decision-Making and Vision; Perception of Surfaces and Forms; Retinal Ganglion Cells: Receptive Fields; Vision for Action and Perception; Visual Attention.

Further Reading

Albright TD and Stoner GR (2002) Contextual influences on visual processing. *Annual Review of Neuroscience* 25: 339–379.

Allman J, Miezin F, and McGuinness E (1985) Stimulus specific responses from beyond the classical receptive field. *Annual Review of Neuroscience* 8: 407–430.

Bakin JS, Nakayama K, and Gilbert CD (2000) Visual responses in monkey areas V1 & V2 to three-dimensional surface configurations. *Journal of Neuroscience* 20(21): 8188–8198.

Bosking WH, Zhang Y, Schofield B, and Fitzpatrick D (1997) Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of Neuroscience* 17(6): 2112–2127.

Crist RE, Kapadia MK, Westheimer G, and Gilbert CD (1997) Perceptual learning of spatial localization: Specificity for orientation, position and context. *Journal of Neurophysiology* 78: 2889–2894.

Duncan RO, Albright TD, and Stoner GR (2000) Occlusion and the interpretation of visual motion: Perceptual and neuronal effects of context. *Journal of Neuroscience* 20(15): 5885–5897.

Field DJ, Hayes A, and Hess RF (1993) Contour integration by the human visual system: Evidence for a local 'association field.' *Vision Research* 33(2): 173–193.

Huang Z, MacEvoy SP, and Paradiso MA (2002) Perception of brightness and brightness illusions in the macaque monkey. *Journal of Neuroscience* 22: 9618–9625.

Kanizsa G (1979) *Organization in Vision: Essays on Gestalt Perception*. New York: Praeger.

Kapadia MK, Ito M, and Gilbert CD (1995) Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. *Neuron* 15: 843–856.

Lamme VAF, Rodriguez-Rodriguez V, and Spekreijse H (1999) Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cerebral Cortex* 9(4): 406–413.

Li W, Piech V, and Gilbert CD (2004) Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience* 7(6): 651–657.

Nakayama K and Shimojo S (1990) Toward a neural understanding of visual surface representation. *Cold Spring Harbor Symposia on Quantitative Biology* LV: 911–924.

Qiu FT and von der Heydt R (2005) Figure and ground in the visual cortex: V2 combines stereoscopic cues with Gestalt rules. *Neuron* 47: 155–166.

Ramachandran VS and Gregory RL (1991) Perceptual filling in of artificially induced scotomas in human vision. *Nature* 350: 699–702.

Sugita Y (1999) Grouping of image fragments in primary visual cortex. *Nature* 401: 269–272.

Zeki S (1980) The representation of colours in the cerebral cortex. *Nature* 284: 412–418.

Cornelia De Lange Syndrome

M A Deardorff and I D Krantz, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The first description of a patient with Cornelia de Lange syndrome (CdLS) has been attributed to Vrolik (1849) who described the case as an extreme example of oligodactyly. In 1916 Brachman provided a detailed account of another CdLS patient in a paper whose title translates as "a case of symmetrical monodactyly representing ulnar deficiency, with symmetrical antecubital webbing and other abnormalities (dwarfism, cervical ribs, hirsutism)." However, it was not until the 1930s that the condition began to be classified as a distinct entity. Cornelia de Lange, a pediatrician in Amsterdam, described two unrelated young girls with strikingly similar features and in honor of the city in which she worked she termed the condition "degeneration typus amstelodamensis." In recognition of her major contribution, the disorder has subsequently been widely described as CdLS, although some publications refer to Brachmann-de Lange syndrome. The prevalence of this syndrome has been estimated to be approximately 1 in 10 000. Since these early reports there have been over 100 papers describing various aspects of this disorder and it was not until recently that the underlying etiology was elucidated as being caused by disruption of the normal functioning of the cohesin complex. This implicated the multiprotein cohesin complex, well characterized for its role in chromosomal cohesion and segregation during mitosis and meiosis, for the first time with a human developmental disorder.

Clinical Features

Structural Features

The CdLS phenotype consists of neurocognitive and somatic growth delays in addition to numerous structural differences. The typical structural anomalies seen in CdLS include: characteristic facial features (including synophrys, long eyelashes, ptosis, depressed nasal bridge with an uptilted tip of the nose and anteverted nares, small widely spaced teeth, small head, and low-set ears), hirsutism, various ophthalmologic abnormalities, abnormalities of the upper extremities (ranging from subtle changes in the phalanges and metacarpal bones to oligodactyly and severe reduction defects primarily affecting the ulnar structures), and gastroesophageal abnormalities (gastroesophageal reflux, malrotation, pyloric stenosis) (Figure 1). Other frequently seen findings include ptosis, myopia, cryptorchidism, hypospadias, congenital diaphragmatic hernia, cardiac defects (most commonly atrial septal defects), seizures, and hearing loss (sensorineural, conductive and mixed have been reported).

Neurodevelopmental Features

The mental retardation seen in CdLS is variable with IQs ranging from less than 30 to 102 with an average of 53. Many affected individuals also demonstrate autistic behavior, including self-destructive tendencies, and they may avoid or reject social interactions and physical contact. Historically, most individuals with classical CdLS have been reported to have severe to profound mental retardation. More recently, reports of CdLS children and adults with milder cognitive retardation have suggested a broader range of intellectual potentials. Achievement of all developmental milestones are generally found to be significantly delayed. On average, children with CdLS sit at 12 months, have their first words at 18 months, walk alone at 24 months, speak (put words together) by 4.5 years, and are toilet trained by 3 years. In severely affected individuals many or all of these milestones are never met. There is no evidence to indicate that individuals with CdLS experience regression (loss of skills) and new skills are acquired even into adulthood. Areas of relative strength were noted to be visual–spatial memory and perceptual organization, while some of the weakest areas involved language skills, especially expressive language. Less than 5% of children with CdLS have language skills within normal to low-normal ranges. In general, those children with essentially no speech and more severe cognitive delays were found to also have more severe physical manifestations of CdLS. CdLS individuals with a greater degree of mental retardation had more pronounced autistic features, hyperactivity, self-injury, aggression, and sleep disturbances.

Clinical Variability

While the facial features seen in individuals with classical CdLS are striking and easily recognizable (Figure 1), and may be one of the most useful diagnostic signs, there is marked variability and a milder phenotype has been consistently described. With increasing recognition of a milder CdLS phenotype, both isolated and familial cases have been diagnosed and reported. The mild phenotype, which has been estimated to account for approximately 20–30% of

Figure 1 Phenotypic features of CdLS. Top row: characteristic facial features including arched eyebrows connected in the midline (synophrys), ptosis, long eyelashes, short upturned nose, long philtrum with thin upper lip, and small chin (micrognathia) with low set and posteriorly rotated ears. Bottom row: phenotypic variability in the involvement of the forearms ranging from severe olygodactyly on the left, through variable types of reduction defects to small hands as seen on the lower right.

the CdLS population (although likely an underestimate due to decreased ascertainment), is distinguished by less significant psychomotor and growth retardation, a lower incidence of major malformations, and milder limb anomalies.

Genetics

Although CdLS is one of the most distinctive dysmorphic syndromes known, having been formally described over 70 years ago, identification of the molecular etiology had remained elusive until recently. Due to a lack of informative families and conserved chromosome abnormalities, standard positional cloning approaches had been difficult to undertake. The CdLS gene was initially presumed to map to chromosome 3q26 based on phenotypic similarities between individuals with duplication of this region and individuals with CdLS. Recently mutations in three genes, *NIPBL* and *SMC1A* (also known as *SMC1L1*) and *SMC3*, have been found to cause CdLS in approximately half of studied probands.

NIPBL, located on chromosome 5p, is a regulator of the cohesin complex and is the human homolog of the *Drosophila Nipped-B* gene. While the role of NIPBL in humans or mammals has not been elucidated, much of what is known about its function comes from work in *Drosophila*. *Nipped-B* was identified in *Drosophila* using a screen to map genes involved in long-range enhancer promoter interactions. Once isolated Nipped-B was found to be homologous to *Saccharomyces cerevisiae* sister chromatid cohesion protein 2 (SCC-2) that played a critical role in sister chromatid cohesion during mitosis and meiosis. It was subsequently shown that through interactions with the cohesin complex, Nipped-B has a dual, but related, role in sister chromatid cohesion, as well as in long-range enhancer–promoter interactions. Preliminary studies evaluating genotype–phenotype correlations in CdLS suggest that some of the phenotypic variability seen in CdLS (severity of cognitive and growth retardation and severity of limb defects) is dependant on the type of mutation identified (with missense mutations for the most part resulting in a milder phenotype when compared to protein truncating mutations).

SMC1A, located on the X chromosome, is a subunit of the cohesin complex and has also recently been demonstrated to cause CdLS when mutated. Even though localized to the X chromosome, this gene is not silenced by X inactivation and both males and females have been reported to be equally affected by mutations, although in familial cases males appear to be more severely affected. SMC3 another subunit that functions in concert with SMC1, was subsequently found to be mutated in a single patient with CdLS. Overall individuals with mutations in *SMC1A* and *SMC3* appear to be more mildly affected, both cognitively and structurally, than individuals with CdLS caused by mutations in *NIPBL*. None of the reported individuals with *SMC1A* or *SMC3* mutations had overt limb abnormalities, several individuals had normal head circumferences and a few appeared to have very subtle features of CdLS and would appear for all intents and purposes as having apparent isolated mental retardation.

CdLS was the first developmental disorder shown to be caused by disruption of cohesin function.

Subsequently, the Roberts/SC phocomelia syndrome (an autosomal recessive condition of mental retardation, growth delay, craniofacial anomalies including cleft lip and palate, upper and lower limb defects, and chromosomal mitotic abnormalities including cohesion defects) was found to be caused by mutations in *ESCO2*, which is required for the establishment of cohesion during S-phase. The implication of the cohesin complex and its regulation in causing the highly conserved phenotypes of specific human developmental disorders was somewhat surprising giving its fundamental role in chromosomal cohesion and is likely due in part to disruption of its function as a mediator of long-range enhancer–promoter interactions.

The Cohesin Complex

In eukaryotic cells, replicated DNA remains physically connected from its synthesis in S-phase until separated during anaphase. This phenomenon called sister chromatid cohesion is essential for the equal separation of the duplicated genome. The cohesin complex is a multimer that consists of at least four proteins: SMC1 (SMC1A in humans), SMC3, SCC3 (STAG 1 and 2 in humans), and SCC1 (RAD21 in humans). Like other structural maintenance of chromosome proteins, SMC1 and SMC3 fold back upon themselves to form antiparallel coiled coils with their N- and C-termini forming ATPase head domains (Figure 2). Their hinge regions interact to form the SMC heterodimer, which acts as a ring-like structure to encircle chromatids. The head domains of SMC3 and SMC1 interact with SCC1 that along with SCC3 serves to form a clasp on the ring-like structure. NIPBL orthologs are known to be a subunit of the adherin protein, which is required for sister chromatid cohesion. It is believed that adherins are required for the attachment of cohesin to chromosomes.

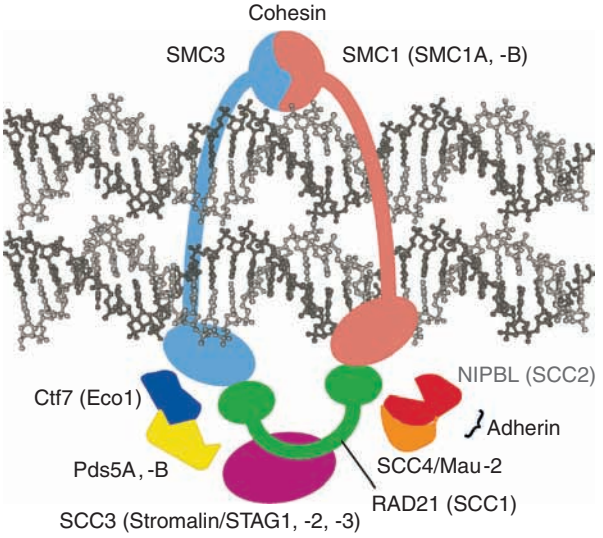

Figure 2 The cohesin complex and its key regulators. The four main proteins composing the cohesin complex (SMC1, SMC3, SCC1, and SCC3) are indicated. The adherin complex of which NIPBL is a component is in the lower left and is involved in loading the complex on and off of the chromosomes.

Long-Range Enhancer–Promoter Interaction and Control of Gene Expression

While chromosomal cohesion may be mildly perturbed in CdLS (there is some evidence for mild precocious sister chromatid separation in CdLS probands), this mechanism is not felt to be the major contributor resulting in the multiple clinical findings seen in CdLS. Rather the critical means by which disruption of NIPBL, SMC1A and SMC3 appears to result in the CdLS phenotype is through the effects on long-range enhancer–promoter interactions and resultant disruption of transcriptional regulation. For most genes the region immediately upstream of the minimal promoter contains the requisite transcription factor binding sites to regulate expression of the gene; however, for many genes multiple *cis*-acting distal elements (most commonly enhancers, but also repressors and insulators) are required for correct spatiotemporal expression. These elements may be located upstream, downstream, or within introns and can reside greater than 1 Mb from the target gene. Disruption of these long-range regulatory interactions can result in human disease phenotypes through either global or partial tissue-specific loss or gain of expression. The majority of genes responsible for these disorders when disrupted are transcription factors that in turn regulate the expression of tissue-specific targets critical for normal morphogenesis of the organ systems involved. In the study that first identified *Nipped-b* in Drosophila, alteration of this gene and the resulting disruption of long-range enhancer–promoter interaction was noted to affect the regulation of the critical transcription factors called homeobox genes. Homeobox genes are critical in normal body plan formation and in the development of the brain and other organ systems.

See also: Autism; Developmental Disability and Fragile X Syndrome: Clinical Overview.

Further Reading

Allanson JE, Hennekam RCM, and Ireland M (1997) De Lange syndrome: Subjective and objective comparison of the classical and mild phenotypes. *Journal of Medical Genetics* 34: 645–650.

Berney TP, Ireland M, and Burn J (1999) Behavioural phenotype of Cornelia de Lange syndrome. *Archives of Disease in Childhood* 81: 333–336.

Deardorff MA, Kaur M, Yaeger D, et al. (2007) Mutations in cohesin complex members *SMC3* and *SMC1A* cause a mild variant of Cornelia de Lange syndrome with predominant mental retardation. *American Journal of Human Genetics* 80(3): 485–494.

de Knecht-van Eekelen A and Hennekam RCM (1994) Historical study: Cornelia C. de Lange (1871–1950) – a pioneer in clinical genetics. *American Journal of Medical Genetics* 52: 257–266.

DeScipio CA, Kaur M, Yaeger D, et al. (2005) Chromosome rearrangements in cornelia de lange syndrome (cdls): report of a der (3)t(3;12) (p25.3;p13.3) in two half sibs with features of cdls and review of reported CdLS cases with chromosome rearrangements. *American Journal of Medical Genetics A* 137: 276–282.

Gillis LA, McCallum J, Kaur M, et al. (2004) *NIPBL* Mutational analysis in 120 individuals with Cornelia de Lange syndrome and evaluation of genotype-phenotype correlations. *American Journal of Human Genetics* 75: 610–623.

Jackson L, Kline AD, Barr MA, and Koch S (1993) de Lange syndrome: A clinical review of 310 individuals. *American Journal of Medical Genetics* 47: 940–946.

Kleinjan DJ and van Heyningen V (2005) Long-range control of gene expression: Emerging mechanisms and disruption in disease. *American Journal of Human Genetics* 76: 8–32.

Krantz ID, McCallum J, DeScipio C, et al. (2004) Cornelia de Lange syndrome is caused by mutations in *NIPBL*, the human homolog of the *Drosophila Nipped-B* gene. *Nature Genetics* 36: 631–635.

Musio A, Selicorni A, Focarelli ML, et al. (2006) X-linked Cornelia de Langesyndrome owing to SMC1L1 mutations. *Nature Genetics* 38: 528–530.

Nasmyth K, Peters JM, and Uhlmann F (2000) Splitting the chromosome: Cutting the ties that bind sister chromatids. *Science* 288: 1379–1385.

Opitz JM (1985) Editorial comment: The Brachmann-de Lange syndrome. *American Journal of Medical Genetics* 22: 89–102.

Rollins RA, Morcillo P, and Dorsett D (1999) *Nipped-B*, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. *Genetics* 152: 577–593.

Russell KL, Ming JE, Patel K, et al. (2001) Dominant paternal transmission of Cornelia de Lange syndrome: A new case and review of 25 previously reported familial recurrences. *American Journal of Medical Genetics* 104: 267–276.

Tonkin ET, Wang TJ, Lisgo S, et al. (2004) NIPBL, encoding a homolog of fungal Scc2 type sister chromatid cohesion proteins and fly *Nipped-B*, is mutated in Cornelia de Lange syndrome. *Nature Genetics* 36: 636–641.

Vega H, Waisfisz Q, Gordillo M, et al. (2005) Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion. *Nature Genetics* 37: 468–470.

Relevant Websites

http://www.cdlsusa.org – Cornelia de Lange Syndrome Foundation, Inc.

http://www.cdlsworld.org – CdLS-World is an international 'hub' for worldwide organizations and communities united by Cornelia de Lange Syndrome.

Corpus Callosum: Agenesis

D Kamnasaran, The Hospital for Sick Children, Toronto, ON, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The corpus callosum was first described in 1812 by Johann Christian Reil and has since been considered an intriguing structure with respect to its development, function, and aftermath of malformation. It represents a unique commissure of placental mammals (eutherians), with a composition of 200–350 million fibers in the human and about 7 million fibers in the mouse, to connect the two cerebral hemispheres, and to perform neocortical sensory functions across the central nervous system (CNS) midline, in addition to providing an advantage for bimanual coordination with accelerated interhemispheric transmission periods. Numerous factors such as language dominance, handedness, and disorders such as schizophrenia and epilepsy are postulated to influence the morphology of the corpus callosum. Pathological assessments of the corpus callosum have indeed captivated the medical community over the centuries. The absence or malformation of this structure correlates with dispensability to some extent among patients and therefore is not mandatory for a normal life. Studies of human subjects with malformations of the corpus callosum are very limited to understanding the structural and molecular processes involved in the development of the corpus callosum. The use of genetically engineered model (GEM) organisms, especially the mouse, has come to a forefront to accelerate our understanding of these processes. This article describes the development of the corpus callosum in humans and mice and the underlying factors (including genetic), which are implicated in acallosal defects.

Embryology of Human Corpus Callosum Development

The development of the corpus callosum occurs in a spatiotemporally defined manner, that is, in anteroposterior and ventrodorsal polarities. Events that initiate the development of the corpus callosum commence between the 6th and 8th week of gestation. During this period, the lamina reuniens forms antecedently near the paraphysis by the agglomeration of the rostral wall of the telencephalon (Figure 1(a)). It is the lamina reuniens that is the origin of commissure fibers for the corpus callosum and anterior commissure. Between the 8th and 10th week, the bundle of commissure fibers augment medially from the ventrolateral wall of the right and left cerebral hemispheres and extend over the midline within the lamina reuniens in order to develop the anterior commissure. It is during the 8th week of gestation that the sulcus medianus telencephai medii (SMTM) develops by a process that involves the medial infolding of the lamina reuniens (Figure 1(b)). Cells from the dorsal lamina reuniens subsequently develop into the hippocampal commissure. These events are a prerequisite for the superior migration of cells from the lamina reuniens by the 9th week of gestation. At this stage, a mass of cells undergo programmed cell death (apoptosis) leaving a fraction of cells that develop into the massa commissularis (Figure 1(b)). It is the massa commissularis (part of the midline glial structures such as glial sling and indusium griseum glia) that serves as a bridge for the migration of the callosal axons in order to establish the higher-order structure of the corpus callosum (Figures 1(b)–1(d)). Different mechanisms may exist for the migration of axons to the corpus callosum between the 11th and 20th week of gestation. However, based on recent magnetic resonance imaging (MRI) findings and studies in mouse models, it is most likely that the initial regions of the corpus callosum to develop are the lamina rostralis (part of the rostrum) and the anterior region of the body (Figure 1(e)). Callosal axons of the rostral regions are speculated to use pioneering axons derived from the cingulated cortex and midline glial structures, while those from the caudal region (future regions of the body) may follow the axons of the fornix and hippocampal commissure to cross the midline, in order to establish the primary architecture of the corpus callosum (Figure 1(e)). Additional axons are then progressively appended to develop the genu and finally the splenium of the corpus callosum (Figure 1(e)). Augmentation of the cortical plates in the hemispheres is accompanied by progressive crossing of callosal axons through the corpus callosum, with distal extensions into the contralateral hemisphere, a process required in the production of a thicker corpus callosum. The principal architecture of the corpus callosum is defined prenatally, with completion of neuronal cell division and migratory events. Size increases are noted among subjects up the age of 25 years, initially in the splenium, and then accompanied by size increases in the body and rostrum. Such a size increase is postulated as a requirement to mediate higher-order neurological functions.

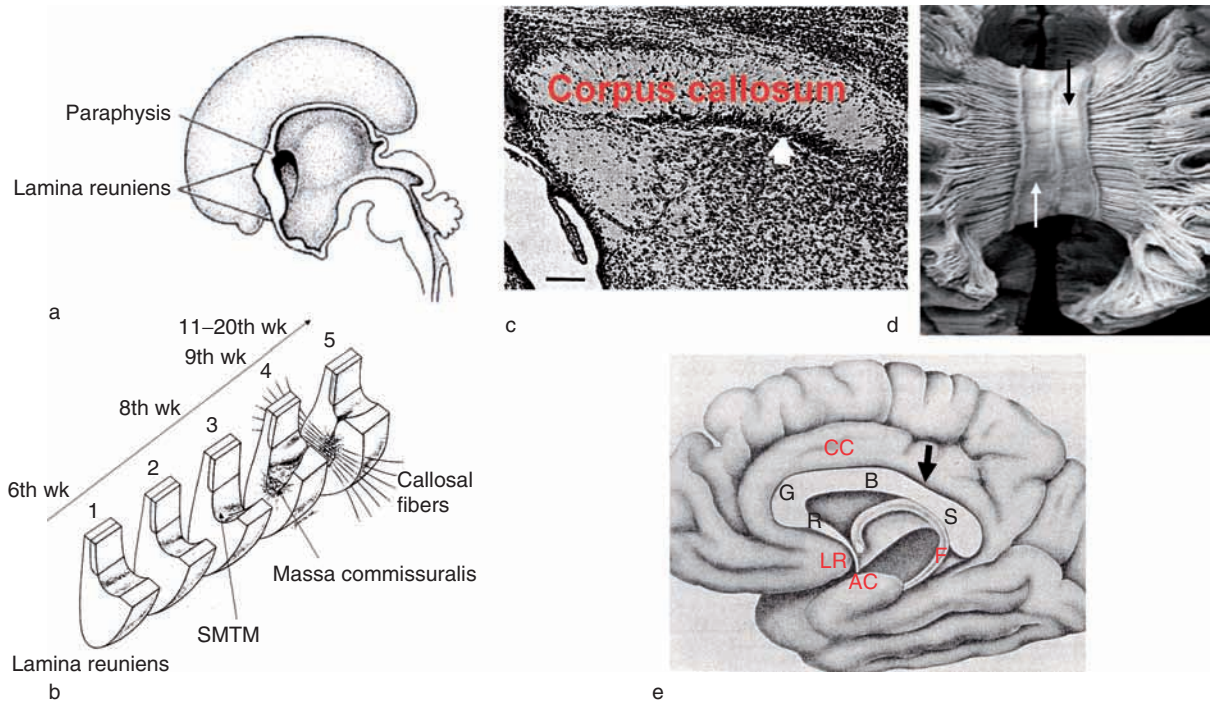

Figure 1 The development of the corpus callosum. (a) The human rostral telencephalon at about 7 weeks of gestation. Thickening of the lamina reuniens is the initial stage in the development of the corpus callosum. (b) Schematic of the progressive events involved in the development of the corpus callosum between the 6th and 20th week of gestation. See text for details. (c) A sagittal section of a normal BALB/cCF mouse strain showing an example of the glial sling structure required for callosal axon guidance. The glial sling is a mass of glial cells flanking the corpus callosum (see white arrow). Scale = 0.1 mm. (d) A ventral view of the interhemispheric connective structures showing the indusium griseum required for callosal axon guidance. The indusium griseum glia (see arrow) is made up of primarily glial cells involved in the guidance of axonal projections. (e) A sagittal section of the brain demonstrating the regions of the corpus callosum (black arrow). SMTM, sulcus medianus telencephaii medii; R, rostrum; G, genu; B, body; S, splenium; LR, lamina rostralis; F, fornix; AC, anterior commissure; CC, cingulated cortex. Adapted from Kamnasaran D (2005) Agenesis of the corpus callosum: Lessons from humans and mice. *Clinical and Investigative Medicine* 28: 267–282; Gentry LR, Thompson B, and Godersky JC (1988) Trauma to the corpus callosum: MR features. *American Journal of Neuroradiology* 9: 1129–1138.

Embryology of Mouse Corpus Callosum Development

The development of the murine corpus callosum is astonishingly similar to humans. Initial events are documented as early as embryonic day 15.5, a stage where callosal axons originate from the cingulated cortex, and by embryonic day 16, from the frontal cortex. Richards and colleagues recently describe possible callosal projection pathways; however, these are yet to be precisely defined. In general, with the use of axonal mapping experiments, some evidence is suggestive of callosal axons originating from differentiating neurons of the cortical plate (specifically layers 2/3 and 5). These axons initially extend radially and downward into the cortical white matter, then change direction and migrate toward the midline. At the midline, the fibers migrate between the hippocampal commissure and pia membrane, and over a mass of glial cells including those from the midline glial structures such as the glial sling and indusium griseum glia (Figures 1(c) and 1(d)). The glial sling and indusium griseum glia act as a bridge between the two cerebral hemispheres for the crossing of fibers over to the contralateral side.

It is postulated that the size of the cerebral hemispheres dictates the precision of these events, a phenomenon not documented in humans. Interestingly, an alternative migratory pathway may be mediated by the hippocampal commissure. At E15, the first hippocampal axons cross the midline and continue onto the contralateral fornix and fimbria to targeted regions during the formation of hippocampus. At this stage, some callosal axons bundle just rostral to the hippocampal commissure. At E16, the callosal axons cross the midline by associating with the hippocampal axons, suggestive of a role of the hippocampal commissure as pioneering axons for the midline crossing of callosal axons. In support of this hypothesis, in acallosal murine strains, the hippocampal axons cross the midline at least 2 days later than normal due to a cleft extending deep between the hemispheres. This delay impairs the proper midline crossing of the callosal axons resulting in acallosal mice. Furthermore, patients

with semilobar holoprosencephaly have an intact hippocampal commissure and with a partial corpus callosum (splenium), suggesting that perhaps the presence of the hippocampal commissure may serve to direct the partial midline migration of callosal axons among these patients.

As soon as the callosal axons gain entry into the contralateral cortex, reentrance into the gray matter is noted, however, at specific positions depending on the site of their origin. By 2 days prior to birth, the callosal axons form synapses with relevant targets. After birth, an increase of 2 mm is noted in the size of the corpus callosum. Between 11 and 50 days after birth, myelination of the callosal axons is documented leading to a progressive augmentation of the corpus callosum.

Factors Shaping the Development of the Corpus Callosum

There are specific physical building blocks that are critical to remain functionally intact as prerequisites for the normal development of the corpus callosum. First, normal development of prosencephalic (embryonic forebrain) structures are absolutely crucial to prevent malformation of the corpus callosum. An early defect in the development of the overall prosencephalon, including the midline, is anticipated to result in acallosal defects accompanied with other CNS anomalies (described below). A classic example of the latter is observed among patients with holoprosencephaly (the improper patterning of the prosencephalon into the bilateral cerebral hemispheres resulting in an absent midline to varying extents).

Second, since callosal axons are derived from the cortex, the proper development of the cortex including processes of proliferation within the ventricular zones, migration to relevant regions, and differentiation into neuronal and glial cell lineages are critical. Furthermore, within the cortex, pioneering axons are necessary for the guidance or direct fasciculation of callosal axons, and reside in regions such as in the cingulated cortex. In this instance, malformed callosal axons, parent cell bodies, or callosal guidance pioneer axons can serve as sources for callosal agenesis. An excellent example of a callosal axon migration defect leading to callosal agenesis is demonstrated among patients with lissencephaly.

Third, intermediate structures or molecules involved in callosal axon guidance are critical to remain functionally intact to prevent malformation of the corpus callosum. Current evidence supportive of this postulate includes the following. Numerous axonal guidance molecules have been identified and include secreted factors, and molecules that are prevalent within the cell membrane or extracellular matrix. Among these groups are included those with chemoattractant, chemo-repellent, growth-permissive, and growth-suppressive properties. Proteins that serve as transcription factors to regulate the expression of these molecules, as well as intracellular signal transduction molecules, have also been identified (see below). The growth cones (tips) of axons have the ability to sense these molecules that will inevitably influence the migration, growth, or death during callosal development. It must also be noted that these axonal guidance molecules are present in variable amounts at intermediate structures of the prosencephalon to fine-tune axonal migratory or projection pathways. Examples of axonal guidance intermediate structures include the midline glial structures such as the glial wedge (part of radial glial scaffold), glial sling (Figure 1(e)), and indusium griseum glia (Figure 1(e)) and glial tunnel (associated with the anterior commissure). Evidence in support of the necessity of these intermediate structures is from several sources. For instance, it is noted that glial wedge explants grown in collagen gels have demonstrated the secretion of short-range diffusible growth regulatory molecules, such as Slit2, that act as factors implicated in the guidance of callosal axons. Most remarkable are findings from grafting and lesioning experiments of the glial sling in the murine developing brain that induces acallosal defects. Furthermore, the reorientation of the glial wedge and glia within the indusium griseum in organotypic slices can induce misguidance of the callosal axons away from the midline.

In general, the absence of these functional building blocks of the corpus callosum coupled with the timing of the absence influences to a large extent the severity of the malformation. For instance, in this manner, it is hypothesized that between the 12th and 20th week of gestation, an insult in the migration of the callosal axons will spare the normal development of the genu, but partial callosal agenesis of varying extents will be noted among other regions of the corpus callosum.

Agenesis of the Corpus Callosum in Humans

Congenital Malformations of the Corpus Callosum

The incidence of agenesis of the corpus callosum varies worldwide based on autopsy reports and other factors such the population studied and diagnostic method utilized. However, an incidence of about 1 in 19 000 (0.005%) subjects in the general population is estimated, and among the pediatric population, the incidence is about 0.0005–0.7%.

The incidence is also noted as being 2–3% in the mentally impaired population and in about 47% of patients with an existing CNS anomaly. Examples of subjects with a normal morphology of the corpus callosum are demonstrated in Figures 2(a) and 2(b), using T1-weighted MRI scans.

Agenesis of the corpus callosum is prevalent as two types. The first type manifests as partial or complete agenesis of the corpus callosum, and is associated with malformations of contiguous and phylogenetically related structures. The most cardinal feature is the lack of commissural fibers to cross the midline. These fibers instead migrate ipsilaterally and reside along the superomedial region of the lateral ventricles in the form of a structure known as Probst bundles (Figures 2(c) and 2(d)). The Probst bundles indent and partition the anterior horns of the lateral ventricles. Frequently, the third ventricle may be displaced upward, and the septum pellucidum is absent. Since the corpus callosum forms in an anteroposterior polarity, consequently, partial agenesis is manifest by a structurally intact genu and body (i.e., anterior regions), and the absence of the splenium and rostrum (i.e., posterior regions) (Figure 2(e)). The second type of callosal agenesis is associated with perturbations in the formation of the forebrain analage, resulting in an increased incidence of callosal agenesis associated with other CNS anomalies. In this type of callosal agenesis, Probst bundles are not documented. Examples of the second type of callosal agenesis include disorders with polymicrogyria and heterotopias of the gray matter.

Symptoms of Patients with Congenital Callosal Agenesis

Subjects with isolated agenesis of the corpus callosum are generally asymptomatic with an incidence estimated as low as 10–17% of all callosal agenesis cases. Isolated callosal agenesis cases are most frequently detected during fetal ultrasound screenings, or during follow-up after birth, with subjects not presenting any clinical symptoms in due course. However, symptomatic subjects are evident when patients have callosal agenesis associated with other

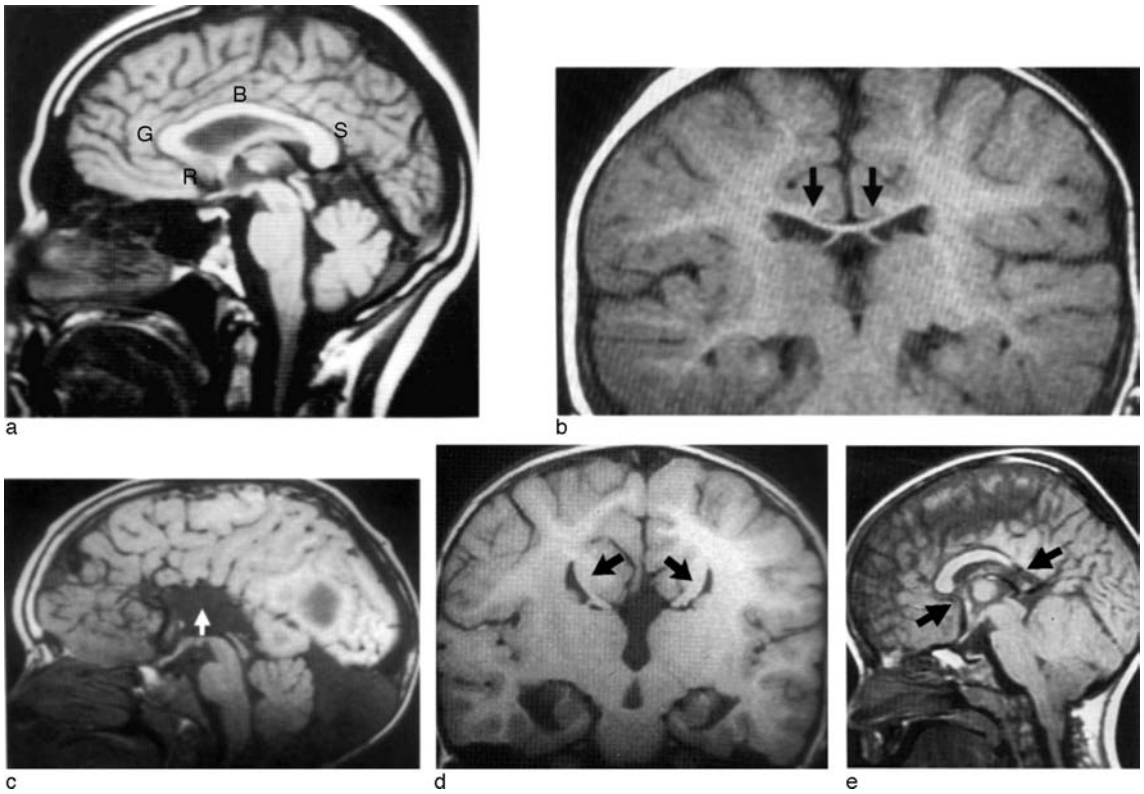

Figure 2 Magnetic resonance imaging (MRI) images from normal and acallosal-defect subjects. (a) Sagittal image of a normal mature corpus callosum from a 15-year-old subject. R, rostrum; G, genu; B, body; S, the splenium. (b) Coronal image of a normal corpus callosum (black arrows). (c) Sagittal image showing agenesis of the corpus callosum with the medial hemispheric sulci radiating into the third ventricle (white arrow). (d) Coronal image of true agenesis of the corpus callosum with Probst bundle formation (black arrows). (e) Sagittal image showing a patient with Chiari II malformation and partial agenesis of the corpus callosum. The rostrum and splenium are absent (black arrows). Adapted from Kamnasaran D (2005) Agenesis of the corpus callosum: Lessons from humans and mice. *Clinical and Investigative Medicine* 28: 267–282.

CNS anomalies and with the tendency to demonstrate neurological symptoms such as sleep disorders, feeding disorders, neuropsychiatric problems, language and social communication disorders, motor impairments, mental delay, and epileptic seizures. Complete agenesis of the corpus callosum is estimated between 53% and 74% of all diagnosed callosal agenesis cases, while partial agenesis of the corpus callosum is estimated among 26–47% of these cases. Complete agenesis of the corpus callosum is more frequently associated with malformations of cortical development, while partial agenesis of the corpus callosum is more correlated with malformations of the posterior CNS fossa. As high as 85% of all callosal agenesis cases are associated with other primary and secondary CNS anomalies, while 62–65% with other non-CNS anomalies. Examples of the spectrum include macrocephaly, microcephaly, hypertelorism, broad and depressed nasal bridge and cleft lip/palate, and sometimes costovertebral defects, gastrointestinal, cardiovascular, and genitourinary anomalies. Patients usually have normal cognitive abilities, but most function at lower than the average range resulting in learning disablities. These patients also demonstrate delayed response periods to interhemisperhic tasks, and impairments in asynchronous distal movements, tactuomotor learning, and cross-localization of finger stimulation. In general, there is no unique prognosis of patients with callosal agenesis.

Differences between Congenital Callosal Agenesis Patients and Those with Complete or Partial Callosotomy

Patients with congenital acallosal defects were originally utilized as natural models of the split brain to understand the role of the corpus callosum and other topologically relevant regions of the CNS. The split brain refers to a completely surgically sectioned corpus callosum/interhemispheric callosal commissure, as in patients treated with a surgical procedure known as callosotomy. A preponderance of patients with congenital acallosal defects are however noted to lack most symptoms of the disconnection syndrome (neurological symptoms as a consequence of a loss of communication between the cerebral hemispheres), as noted in patients treated by callosotomy. To elaborate, congenital callosal agenesis subjects, for instance, do not typically experience difficulties for tachistoscopic tasks requiring interfield comparisons, or tasks involving recognition and verbal identification of an object placed unilaterally to either hemisphere. Callosotomized patients, however, are dysfunctional with respect to these tests. One possible explanation for this phenomenon is that patients with a congenital acallosal brain may utilize neural plasticity and reorganization of the subcallosal migratory tracts to compensate for the lack of the principal commissure during early stages of gestational development and possibly before birth. In this instance, the anterior commissure and subcortical commissures can therefore act to compensate and to prevent impaired visual interhemispheric tasks. In support of this hypothesis, an enlarged anterior commissure is reported among some congenital acallosal patients.

Acquired Malformations or Lesions of the Corpus Callosum

Apart from congenital callosal agenesis, acquired malformations or lesions of the corpus callosum also exist, but are relatively rare. These lesions are most often underestimated as a consequence of absent neurological symptoms and the type of cranial imaging detection method utilized. The prevalence is estimated at 3% in the general population. Several examples are described below.

Vascular or arteriovenous malformations These represent about 9–11% of all cerebral vascular anomalies among patients. In a large fraction of these patients, intraventricular hemorrhaging occurs resulting in ischemic-induced callosal atrophy (Figure 3(a)). Infarcts are uncommon (8%) in the corpus callosum since ischemic injury is rare in the dense white matter typically residing within this commissure. This may be a consequence of the resilient angioarchitecture of the corpus callosum that can supply blood from a combination of both anterior and posterior cerebral arteries to prevent atrophy.

Repair of hydrocephalus Dilated ventricles as a result of increases in cerebrospinal fluids most often indents the middle and posterior regions of the corpus callosum causing ischemia. Ischemia occurs because of a reduction in the number of capillaries leading to atrophy of the posterior corpus callosum (Figure 3(b)). Repair of the hydrocephalus after ventriculoperitoneal shunting can induce both diffuse and focal lesions of the corpus callosum. This occurs because overdrainage can distend pericallosal artery branches resulting in ischemia and demyelination.

Trauma Among subjects who have died of severe head injuries, about 16–100% are noted to have traumatic lesions of the corpus callosum, which are typically hemorrhagic. These lesions are also prevalent in some nonfatal head traumas. Larger lesions extend throughout the corpus callosum, and smaller lesions are most prevalent in the splenium of the corpus callosum (Figure 3(c)). Focal atrophy of the corpus callosum may be a consequence of Wallerian

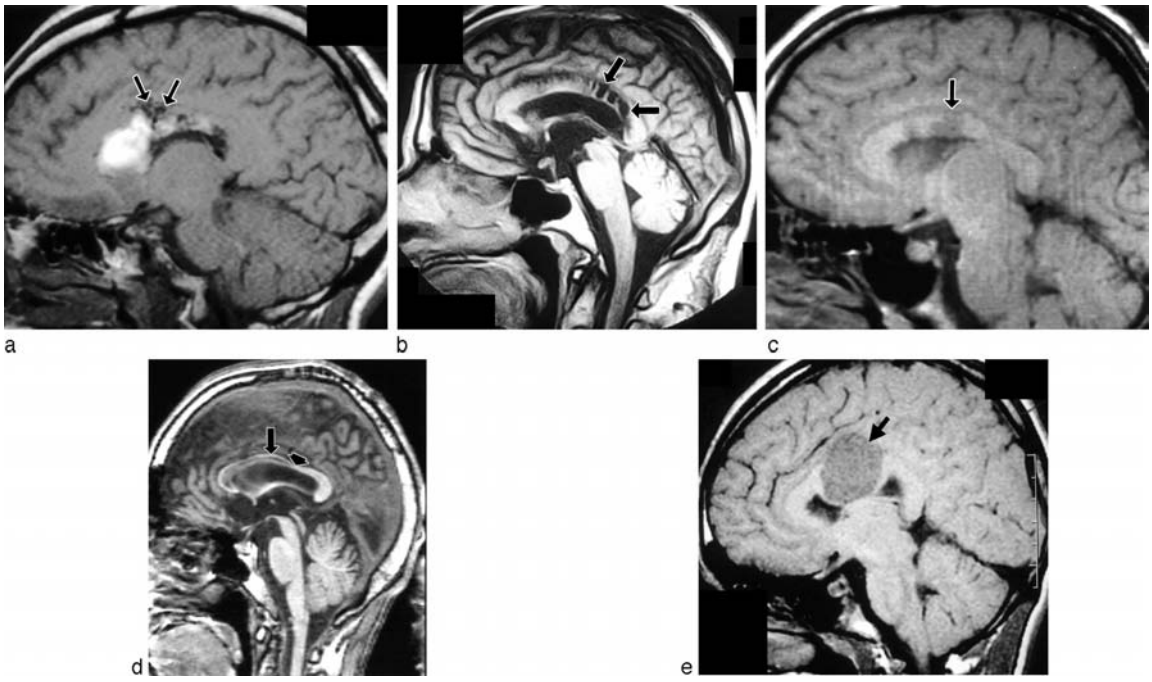

Figure 3 Magnetic resonance imaging (MRI) examples of acquired malformations/lesions of the corpus callosum. (a) An early forties male with arteriovenous malformation and intraventricular hemorrhage (arrows). (b) A mid-forties male with cystic lesions (arrows) associated with long-term hydrocephalus and multiple prior shunt revisions. (c) An early 20-year-old male with a tramautic nonhemorrhagic hypointense lesion (arrow) of the corpus callosum, after a motor vehicle accident. (d) A mid-fifties male with the demyelinating disease, Marchiafava–Bignami disease, after 30 years of alcohol abuse. Atrophy of the corpus callosum (arrows) is noted. (e) A 4-year-old girl with gliomas in the body of the corpus callosum (arrow). Adapted from Bourekas EC, Varakis K, Bruns D, et al. (2002) Lesions of the corpus callosum: MRI and differential considerations in adults and children. *American Journal of Roentgenology* 179: 251–257.

degeneration or direct callosal injury. Wallerian degeneration refers to antegrade degeneration of axons in conjunction with their myelin sheaths secondary to proximal axonal injury or cell death.

Demyelinating diseases Chronic alcoholics eventually present at late stages atrophy of the corpus callosum as observed in Marchiafava–Bignami disease (Figure 3(d)). In these subjects, the central regions of the corpus callosum are most commonly affected, and atrophy is postulated as a result of demyelination induced by necrosis. Lesions can also be multifocal within the corpus callosum. Other types of demyelinating diseases leading to callosal atrophy include multiple sclerosis with patients usually showing symptoms within the disconnection syndrome, and progressive multifocal leukoencephalopathy.

Tumors High-grade infiltrative tumors along the white matter tracts across the corpus callosum are common in glioblastoma multiforme (GBM), a high-grade glioma (Figure 3(e)). GBMs are the most common type of adult primary brain tumors. Moreover, as high as 65% of lipoma cases are documented with lesions in the corpus callosum and in the pericallosal cistern, but these patients are usually asymptomatic. Other types of tumors that induce lesions within the corpus callosum include lymphomas, carvenous angiomas, and juvenile pilocytic astrocytoma.

Generalized etiology The etiology of callosal agenesis is very much heterogeneous. Apart from the lack of functional building blocks during callosal development, callosal agenesis is also observed among conceptuses of mothers with alcohol abuse, maternal rubella infections, congenital toxoplasmosis, and diabetes. Although these cases clearly demonstrate a most likely induced teratogenic gestational environment, and mothers demonstrating increased associations/risk with callosal agenesis conceptuses, the underlying pathological mechanisms are yet to be determined.

Understanding the Genetic Etiology of Congenital Agenesis of the Corpus Callosum

The genetic etiology of agenesis of the corpus callosum is enthralling and extremely genetically heterogeneous for both isolated callosal agenesis and callosal agenesis associated with other complex CNS malformation syndromes. This extreme genetic heterogeneity may be a consequence of the complex embryological processes underlying the development

of the corpus callosum. Although several mechanisms may be implicated in callosal agenesis, the same genetic defects responsible for callosal agenesis can be used to comprehend the causes of other associated CNS malformations – in those instances where callosal anomalies are associated with other complex CNS anomalies. Interestingly, the prevalence is skewed with higher detection of acallosal defects among males compared to females (1.6:1 ratio), suggestive of possible X-linked modifier genes regulating sex-specific incidences of callosal agenesis.

Genetic syndromes/disorders/malformations The familial recurrence of callosal agenesis is relatively rare. This is suggestive that callosal agenesis may be primarily caused by *de novo* or sporadic genetic mutations (either by germline or early postzygotic mechanisms) rather than a familial mutation. Autosomal recessive, dominant, and X-linked forms are reported in at least 22 malformation syndromes with autosomal inheritance, and at least 13 malformation syndromes with X-linked inheritance (Table 1). A total of 16 malformation syndromes are yet to be associated with a specific genetic mode of inheritance among affected families. A few families with isolated callosal agenesis have been reported with autosomal dominant, autosomal recessive, and X-linked inheritance. Moreover, at least 56 malformation syndromes and metabolic disorders associated with agenesis of the corpus callosum have been documented to date (Table 1). Currently, at least 33 genes were identified with mutations among some of these disorders. Syndromes, such as Joubert, septo-optic dysplasia, and Kabuki, are rarely associated in patients with partial or complete agenesis of the corpus callosum. Furthermore, patients with inborn errors of metabolism such as neonatal adrenoleukodystrophy, kinky hair disease, mucopolysaccharidosis, nonketotic hyperglycinemia, pyruvate dehydrogenase deficiency, Zellweger syndrome, and glutaric aciduria type II demonstrate associations with accallosal defects. For both the callosal agenesis-associated syndromes that are rare, and inborn errors of metabolism disorders, the genetic perturbations underlying the CNS malformations may not necessarily be directly related to the cause of callosal agenesis, a hypothesis yet to be tested.

Chromosome aberrations Apart from the use of syndromes to discern the genetic basis of the development of the corpus callosum, the use of chromosomal anomalies has been informative to define possible minimal critical regions (MCRs) for agenesis of the corpus callosum in the human genome. More than 11% of all callosal agenesis patients tested have been documented with a chromosome anomaly. This may however be an underestimate since most of these cases were analyzed using classic cytogenetic karyotyping methods that are sensitive to detecting only large chromosomal rearrangements. Recent usage of bacterial artificial chromosome (BAC) genome microarrays has improved the sensitivity and detection of submicroscopic chromosome rearrangements among patients that were previously undetected by utilizing classic cytogenetic karyotyping methods. In general, human chromosomes 8, 11, 13–15, and 18 are most commonly associated with agenesis of the corpus callosum. A wider spectrum of frequent chromosomal aneuploidies is anticipated with the use of higher-resolution karyotyping methods such as with BAC genomic microarrays. Numerical and structural aberrations involving chromosomes 1, 2, 3, 4, 6, 11, 13, 15, 16, 21, 22, and X suggest that a reduction in gene dosage of one or more genes on these chromosomes contributes to agenesis of the corpus callosum. An increase in gene dosage as a possible cause of agenesis of the corpus callosum is suggested by reports of partial trisomies (i.e., two normal chromosomes and a partial chromosome), or duplications of chromosomes 4, 5, 6, 8, 11, 12, 14, and 19, triploidy (i.e., three copies of every chromosome) and trisomies (i.e., three copies of a chromosome) of chromosomes 13, 18, 21, and 22. Figure 4 illustrates a genetic map of candidate regions for callosal agenesis on human chromosomes defined by the use of affected patients with chromosome aberrations. Many of these chromosomes harbor genes that are already identified with mutations among patients with acallosal defects (Table 1).

Agenesis of the Corpus Callosum in Mouse

Comprehending the cause of callosal agenesis in humans is very much limited with respect to unraveling perturbations in structural and molecular pathways. To circumvent this limitation, model organisms have been utilized, not only to understand the structural basis of the development of the corpus callosum, but also to validate the roles of genes discovered. The mouse has served as an excellent model system among other models such as the cat, rat, dog, or even chimpanzee to provide effectively outstanding structural and molecular insights into the development of the corpus callosum.

Murine Strains with Spontaneous Callosal Agenesis

It must be first recognized that some murine strains demonstrate spontaneous acallosal defects at different frequencies (as shown in Table 2). These strains are hypothesized to have mutations in yet-to-be-identified genes that collectively contribute toward

Table 1 Syndromes/disorders/malformations associated with agenesis of the corpus callosum

Syndrome/disorder	*Gene(s)*	*Location on chromosome*	*Mouse model with accallosal defects*
Acrocallosal syndrome		7p13	
Aicardi syndrome		Xp22	
Alpha-thalassemia/mental retardation syndrome	*XH2*	Xq13	tbd[a]
Andermann syndrome	*SLC12A6*	15q13-q15	No
Apert syndrome	*FGFR2*	10q26	No
Baller–Gerold syndrome			
Coffin–Siris syndrome			
Crouzon syndrome	*FGFR2*	10q26	No
Dincsoy syndrome			
Facio-auriculo-vertebral syndrome			
FG syndrome		Xq12-q21.31	
Frontonasal dysplasia sequence	*EFNB1*	Xq12	No
Fryns syndrome			
Glutaricaciduria type II	*ETFDH, ETFA, ETFB*	19q13.3, 15q23-q25, 4q32-qter	No
Gorlin syndrome	*PTC*	9q31, 9q22.3	No
Greig cephalopolysyndactyly syndrome	*GLI3*	7p13	
HSAS syndrome	*L1CAM*	Xq28	Yes
Hydrolethalus syndrome	*HLYS1*	11q23-q25	tbd
Joubert syndrome	*NPHP1*	9q34.3, 2q13	tbd
Kabuki syndrome		8p23.1-p22	
Kinky hair disease			
Lowry–Wood syndrome			
Malpuech syndrome			
Marden–Walker syndrome			
Marshall–Smith syndrome		2q3	
MASA syndrome	*L1CAM*	Xq28	Yes
Meckel–Gruber syndrome		17q22-q23 8q24	
Microphthalmia-linear skin defects syndrome		Xp22.31	
Microgastria-limb reduction defects association			
Microcephalic osteodysplastic primordial dwarfism type I/III			
Miller–Dieker syndrome		17p13.3	
Mowat–Wilson syndrome	ZFHX1B	2q22	No
Neonatal adrenoleukodystrophy	*PEX1, PEX10, PEX13, PEX26*	1, 2p15, 22q11.21, 12p13.3, 7q21-q22	No
Neu–Laxova syndrome			
Oral-facial-digital syndrome	*CXORF5*	Xp22.3-p22.2	tbd
Oculu-Auriculo-vertebral spectrum			
Optiz G syndrome	*MID1*	Xp22	tbd
Peters-plus syndrome			
Proud syndrome	*ARX*	Xp22.13	tbd
Pyruvate decarboxylase deficiency	*PDHA1*	Xp22.2-p22.1	No
Radial aplasia-thrombocytopenia syndrome			
Rubinstein–Taybi syndrome	*CREBBP*	16p13.3	No
Seckel syndrome		3q22-q24, 18p11.31-q11.2, 14q21-q22	
Septo-optic dysplasia	*HESX1*	3p21.2-p21.1	Yes
Shapiro syndrome			
Simpson–Golabi–Behmel syndrome	*GPC3*	Xq26	No
Sotos syndrome	NSD1	5q35	tbd
Toriello–Carey syndrome			

Continued

Table 1 Continued

Syndrome/disorder	*Gene(s)*	*Location on chromosome*	*Mouse model with accallosal defects*
Vici syndrome			
Walker–Warburg syndrome	*POMT1, FCMD*	9q34.1, 9q31	tbd
Warburg Micro syndrome	*RAB3GAP*	2q21.3	tbd
Winter–Baraitser syndrome			
X-linked hydrocephalus	*L1CAM*	Xq28	Yes
Yunis–Varon syndrome			
Zellweger syndrome	*PEX1, PEX2, PEX3, PEX5, PEX6, PEX12, PEX14, PEX26*	1, 2p15, 22q11.21, 1q22, 12p13.3, 1p36.2, 7q21-q22, 6q23-q24	No

[a]tbd, to be determined.

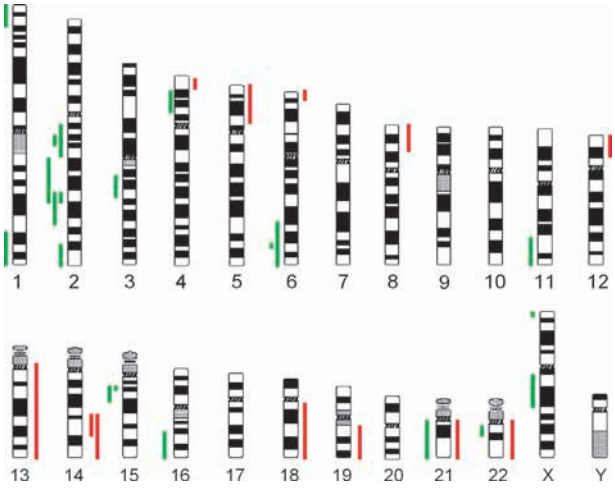

Figure 4 Genetic map of agenesis of the corpus callosum in the human genome, based on patients with chromosome aberrations. Green bars on the left of chromosome ideograms are examples of intervals reported in patients with a decrease in gene dosage. Red bars on the right of chromosome ideograms are examples of intervals reported in patients with an increase in gene dosage.

Table 2 Spontaneous callosal agenesis murine strains

Strain	*Complete callosal agenesis (%)*	*Callosal hypoplasia (%)*
BALB/cWah2	2	17
BALB/cWah1	20	55
129/J	30	71
129/ReJ	2	77
I/LnJ	100	100
Wah1	8	57
Wah1/Wah2	4–7	43–50
Wah1/I/LnJ	0	69
Wah2	4	33
Wah2/129	0	9
Wah2/I/LnJ	15	70
129	2	77
129/I/Ln	9	73
C129F2	24	33
RI-1	100	100
BALB/c	2–21	17–55
129/SuEv	6	50

callosal agenesis. Breedings undertaken among these strains show variations in callosal agenesis. For instance, breedings between the BALB/cWah1 and 129/ReJ strains have a normal corpus callosum, but crosses between strains BALB/c and I/LnJ or I/LnJ and 129/ReJ, have agenesis of the corpus callosum. With this knowledge, modeling a mutant candidate callosal agenesis gene with one of the strains listed in Table 2 will increase the possibility of observing a callosal agenesis phenotype. This phenomenon has been well documented with the modeling of the *Emx1* (a transcription factor of unknown function), *bApp* (a β-amyloid precursor transmembrane glycoprotein), and *L1Cam* (transmembrane glycoprotein with roles in cell adhesion, neurite outgrowth, and axon bundling) genes that result in true agenesis of the corpus callosum with Probst bundle formation. For instance, when both copies of the *Emx1* gene are defective in a 129Sv strain, all mice have partial or complete callosal agenesis. However, in a hybrid 129Sv-C57BL/6 strain, mice develop only a slight atrophy or reduction in size after the eighth generation.

Genetically Engineered Models of Callosal Agenesis

Callosal agenesis models were developed in the past by utilizing nongenetic manipulation methods such as with the use of prenatal irradiation. With these methods, the perturbed molecular (genetic) parameters are unknown. GEMs have come to a forefront to model and understand the role of candidate callosal agenesis genes in mice. GEMs, however, do have advantages and disadvantages to comprehend the pathogenesis of callosal agenesis (Table 3), which to some extent impose caution in understanding this anomaly in humans. For instance, some of these differences are explicit when callosal agenesis is

observed among patients affected with Anderman syndrome and with mutations in the *SLC12A6* (Solute carrier family 12, member 6) gene; however, no acallosal defects are observed in the murine GEM when one or both copies of the *Slc12A6* gene are defective. Other examples are listed in Table 1.

Table 3 Pros and cons of using murine genetically engineered models to study acallosal defects

Advantages	*Disadvantages*
Thorough knowledge on neuroanatomy and neuroembryology	Strains with spontaneous acallosal defects
Large panel of genetic markers available for genetic mapping experiments	Differences in the spatiotemporal expression of genes
Numerous mutant strains available from public consortiums	Differences in gene function
Efficient and improved technologies for the creation of genetically engineered mice	Differences in gene redundancy

Complete agenesis of the corpus callosum Numerous GEMs were created to model true agenesis of the corpus callosum, that is, mice with Probst bundle formation (as listed in Table 4) or agenesis of the corpus callosum associated with other CNS malformations (as listed in Table 1). GEMs with true agenesis of the corpus callosum are more informative to directly correlate gene function with the molecular–structural parameters required for callosal axon guidance and migratory events leading toward the development of the corpus callosum. Several genes have been identified to result in complete true agenesis of the corpus callosum and have diverse functions. These include *Vax1* (a transcription factor implicated in the formation of the midline commissure tracts and guidance of axons mediated by molecules such as Netrin-1, EphB3, and EphB2), *Nfia* (a transcription factor that regulates the development of the midline glia and axonal chemorepellent molecules such as

Table 4 Mouse models for callosal agenesis (true complete agenesis and hypoplasia)

Gene	*Location on mouse chromosome*	*Function*	*Phenotype in mouse model*	*Location on human chromosome*	*Mutation in human disease with acallosal agenesis*
bApp1	16 C3	Transmembrane glycoprotein	ACC[a], *HCC*[b]	21q21.2	
Creb1	1 C1/C5	Transcription factor	HCC	2q32.3-q34	Deleted among acallosal agenesis patients
Emx1	6 C3	Transcription factor	ACC, HCC	2p14-p13	
EphA5	5 E1	Transmembrane receptor	ACC, HCC	4q13.1	
Frizzled-3	14 D1	Transmembrane receptor	HCC	2q32.3-q34	
L1Cam	X A6-B	Transmembrane glycoprotein	ACC, HCC	Xq28	CRASH/HSAS syndrome, MASA syndrome, X-linked hydrocephalus
Map1B	13 D1	Regulating dynamics of laminin cytoskeletal elements	ACC, HCC	5q13	
Marcks	10 B1	Intracellular substrate for protein kinase C	ACC	6q22.2	
Mena	1 H5	Regulating dynamics of actin cytoskeletal elements	ACC	1q42	
Netrin-1	11 B3	Soluble protein with chemotaxic role	ACC	17p12-q13	
Nfia	4 C4-C6	Transcription factor	ACC	1p31	
Pac1	6 B3	G-coupled receptor	ACC, HCC	7p14	
p35	11 B5	Activator of CDK5	HCC	17q11.2	
Slit2	5 B3	Soluble protein with chemotaxic role	ACC	4p15	
Tlx	10 B2	Transcription factor	HCC	6q21	
Vax1	19 D3	Transcription factor	ACC	10q26	

[a]ACC, complete agenesis of the corpus callosum.
[b]HCC, hypoplasia (dysgenesis) of the corpus callosum.

Slit2), *Map1B* (a structural protein with role in microtubule assembly and neuronal morphogenesis, by mediating axonal guidance perhaps through laminin-enhanced axonal growth/migration), *Netrin-1* (a soluble protein involved in axonal guidance), *Marcks* (an intracellular signal transduction substrate for protein kinase C), *Mena* (a suppressor of the Abl nonreceptor kinase that regulates actin cytoskeletal elements), *Slit2* (a soluble protein involved in axonal guidance), and *EphA5* (a tyrosine kinase receptor, regulating axonal guidance).

Hypoplasia/dysgenesis of the corpus callosum Atrophy of the corpus callosum maybe misdiagnosed as true agenesis of the corpus callosum, and may be a consequence of an aberrant event during cortical development (such as in preterm infants or perinatal asphyxia). Genes implicated with this phenotype, however, have crucial roles such as in the growth of the corpus callosum. The GEMs demonstrate hypoplasia (dysgenesis) of the corpus callosum (see also Table 4), and include mutations in genes such as *Frizzled-3* (a transmembrane receptor with possible roles in axon growth by coordinating cytoskeletal polarity), *Tlx* (a transcription factor with unknown function), *p35* (an activator of cyclin-dependent kinase 5), and *Creb1* (a cAMP response element-binding protein involved in cellular growth and differentiation).

Concluding Remarks

Numerous autopsy cases and patients have been diagnosed with agenesis of the corpus callosum for over two centuries and have perplexed both the medical community and developmental biologists as to what are the underlying causes. An understanding of both the structural and molecular (genetic) mechanisms implicated in accallosal defects have been minimal thus far, most likely as a consequence of the immense complexity apparent as a consequence of genetic heterogeneity and/or possible phenotypic modifications by the gestation environment. Both structural and molecular parameters have been determined in a complementary human–mouse system using standard genetic and developmental biology approaches. With the availability of the human and mouse genomic sequences plus improved mouse modeling resources and mutation detection technologies, our understanding of these processes is anticipated to accelerate exponentially.

See also: Forebrain Development: Holoprosencephaly (HPE); Hemispheric Specialization and Cognition; Holoprosencephaly; Lissencephaly; Split-Brain Patients.

Further Reading

Barkovich AJ (2005) *Pediatric Neuroimaging*, 4th edn. New York: Raven.

Bourekas EC, Varakis K, Bruns D, et al. (2002) Lesions of the corpus callosum: MR imaging and differential considerations in adults and children. *American Journal of Roentgenology* 179: 251–257.

Dobyns WB (1996) Absence makes the search grow longer. *American Journal of Human Genetics* 58: 7–16.

Gentry LR, Thompson B, and Godersky JC (1988) Trauma to the corpus callosum: MR features. *American Journal of Neuroradiology* 9: 1129–1138.

Kamnasaran D (2005) Agenesis of the corpus callosum: Lessons from humans and mice. *Clinical and Investigative Medicine* 28: 267–282.

Magara F, Muller U, Li ZW, et al. (1999) Genetic background changes the pattern of forebrain commissure defects in transgenic mice underexpressing the beta-amyloid-precursor protein. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4656–4661.

Nebes RD (1990) The commissurotomized brain. In: Boiler F and Grafman J (eds.) *Handbook of Neuropsychology*, vol. 4, section 7, pp. 3–168. Amsterdam: Elsevier.

Rakic P and Yakovlev PI (1968) Development of the corpus callosum and cavum septi in man. *The Journal of Comparative Neurology* 132: 45–72.

Richards LJ, Plachez C, and Ren T (2004) Mechanisms regulating the development of the corpus callosum and its agenesis in mouse and human. *Clinical Genetics* 66: 276–289.

Schmidt SL and Lent R (1987) Effects of prenatal irradiation on the development of cerebral cortex and corpus callosum of the mouse. *The Journal of Comparative Neurology* 264: 193–204.

Sherr EH, Owen R, Albertson DG, et al. (2005) Genomic microarray analysis identifies candidate loci in patients with corpus callosum anomalies. *Neurology* 65: 1496–1498.

Silver J, Lorenz SE, Wahlsten D, and Coughlin J (1982) Axonal guidance during development of the great cerebral commissures: Descriptive and experimental studies, *in vivo*, on the role of preformed glial pathways. *The Journal of Comparative Neurology* 210: 10–29.

Takahashi T, Kinsman S, Makris N, et al. (2003) Semilobar holoprosencephaly with midline 'seam': A topologic and morphogenetic model based upon MRI analysis. *Cerebral Cortex* 13: 1299–1312.

Wahlsten D (1984) Growth of the mouse corpus callosum. *Brain Research* 317: 59–67.

Cortical Control of Eye Movements

P H Schiller and E J Tehovnik, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Several cortical areas contribute to the generation of visually guided saccadic eye movements. The major areas we shall consider here area striate cortex (area V1), the lateral intraparietal sulcus (LIP), the frontal eye fields (FEFs), and the medial eye fields (MEFs). Figure 1 shows a view of the monkey cortex with these areas indicated by the arrows; also named are some of the major sulci.

Electrical stimulation of these cortical areas elicits saccadic eye movements. In areas V1, LIP, and the FEFs constant vector saccades are elicited. The amplitude and direction of such saccades depends on where within each structure the electrode is placed. The vector is constant in the sense that it is unaffected by the initial position of the eye in orbit. Also shown in the upper portion of Figure 2(a) are the effects of electrical stimulation in the superior colliculus (SC), which are similar to those obtained in V1, LIP, and the FEFs. Not shown is that similar saccades also can be elicited by the electrical stimulation of area V2. It may be said that these areas use predominantly a retinocentric code. By contrast, a different kind of coding operation is evident in the MEFs. In this structure, electrical stimulation moves the eyes to a particular orbital position. Different subregions of the MEF code different orbital positions, which are arranged in topographic order.

Electrical stimulation in the SC, FEFs, and MEFs can elicit saccadic eye movements at low current levels and can produce saccades with latencies as short at 15–25 ms. By contrast, whereas electrical stimulation of areas V1 and LIP can elicit eye movements at correspondingly low thresholds, the minimal saccadic latencies are notably higher, at 35–50 ms.

How do the signals from areas V1, LIP, the FEFs and the MEFs reach the brain stem oculomotor centers? These regions have numerous connections. All project to the SC. However, the FEFs and the MEFs also have direct connections with the brain stem oculomotor centers. To determine whether the commands from all of these areas pass through the SC, we can ablate the SC and then stimulate the various cortical areas to determine whether saccades can still be elicited. The results of such an experiment are shown in Figure 2(b). After SC ablation, saccades can no longer be elicited from V1, V2, and LIP. However, stimulation of the FEFs and MEFs still elicits the same kinds of saccades as before the SC ablation. These findings have led to the suggestion that there are two major cortical systems for saccadic eye-movement generation: the anterior and the posterior. The posterior system passes through the SC, whereas the anterior system gains direct access to the brain stem oculomotor centers. These observations then raise the question of why we have two such systems. In what follows, we examine these areas in more detail using several different methods that include single-cell recordings, lesions, microstimulation, and pharmacological manipulations.

The Neural Circuits of Visually Guided Saccadic Eye Movements as Revealed by Anatomical and Physiological Studies

In the retina, there are several different classes of retinal ganglion cells. Two of these classes, the midget and the parasol cells, are believed to play a central role in the generation of visually guided saccadic eye movements. The parasol cells project to the magnocellular layers of the lateral geniculate nucleus (LGN), whereas the midget cells project to the parvocellular layers. The major projections of these two classes of cells from the LGN of the primate are to layers 4Cα and β in the V1.

Area V1 makes extensive reciprocal connections with numerous extrastriate areas and also projects to the SC from layer 5. These cortico-tectal cells are comprised of complex cells and are driven predominantly by the parasol system. Although the superficial layers of the SC receive direct connections from a class of small, relatively slowly conducting retinal ganglion cells, this input does not have the capacity to activate the deeper layers of the SC; when the visual cortex is cooled or ablated, neurons in stratum opticum and below can no longer be driven visually in the SC. Thus it appears that in primates the deeper layers of the SC are under cortical control.

In V1, many cells receive selective input from either the midget or parasol cells, but there is a significant portion of cells that receive a convergent input from these two systems. Such convergent input is also evident in V4. However, the middle temporal area (MT) and other regions of the parietal lobe receive an input that appears to be dominated by the parasol system.

In area LIP of the parietal lobe, several different classes of cells involved in saccadic eye movements generation have been identified. In some regions of this structure neurons discharge prior to the initiation of saccadic eye movements, whereas in other regions neurons are found that discharge vigorously while the

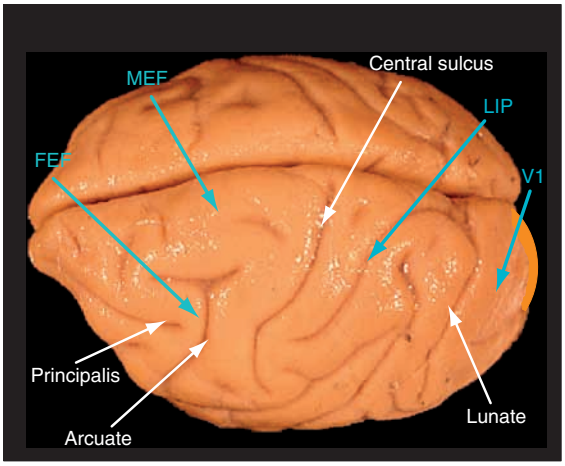

Figure 1 View of the monkey cortex. FEF, frontal eye field; LIP, lateral intraparietal area; MEF, medial eye field.

animal maintains fixation. The firing of these neurons is also affected by different reward contingencies.

As in LIP, in the FEFs several different classes of neurons have been identified. Neurons discharge vigorously in association with a particular size and direction of saccades irrespective of the starting eye position. These neurons, as already noted, appear to carry a vector code. Some of these neurons discharge prior to the initiation of eye movements, and some discharge during or after saccade initiation. Neurons coding for different sizes of saccades are organized topographically within the FEFs such that lateral portions of FEFs code for small amplitude saccades while medial portions code for large amplitude saccades. Neurons in the superficial layers discharge to both the onset of visual targets and saccade initiation, whereas neurons in the deeper layers discharge mainly during saccade initiation. Neurons in

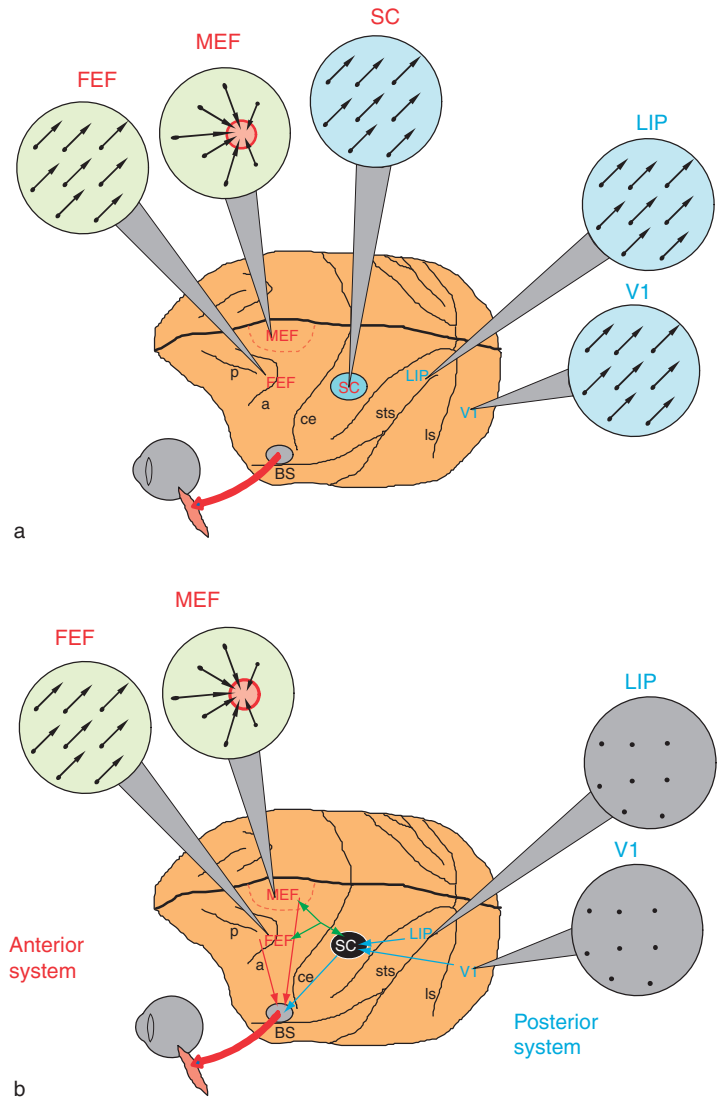

Figure 2 Effects of electrical stimulation in the superior colliculus (SC) which are similar to those obtained in V1, the lateral intraparietal area (LIP), and the frontal eye field (FEF): (a) intact animal; (b) SC-ablated animal. a, arcuate; BS, brain stem; ce, central sulcus; ls, lunate; MEF, medial eye field; p, principalis; sts, superior temporal sulcus.

the FEFs are believed to be involved in target selection as well as in the generation of sequences of eye movements.

In the MEFs, most neurons code for the position of the eyes in orbit. A group of neurons has been identified in the MEFs that discharge vigorously for fixations within different parts of the visual field such that neurons in anterior MEFs discharge for fixations of targets in the contralateral field, neurons in posterior MEFs discharge for fixations of targets in the central field, neurons in medial MEFs discharge for fixations of targets in the lower field, and neurons in lateral MEFs discharge for fixation of targets in the upper field. Vigorous activity of these cells occurs when the eyes shift to the orbital position coded by them. This coding operation, as already noted, has also been confirmed by electrical microstimulation studies. Finally, much like the neurons in LIP, the neurons in the MEFs are modulated by different reward contingencies.

The Effects of Selected Lesions on Saccadic Eye-Movement Generation

When monkeys are examined informally following lesions of the SC, there are only subtle deficits. The animals make fewer eye movements per unit time, and, following, unilateral lesions they tend to orient toward the side of the lesion. Deficits are even less evident after FEF and MEF lesions. However, when the SC and the FEFs are both removed bilaterally, a dramatic deficit surfaces – monkeys can no longer generate saccadic eye movements at all. This suggests that the anterior and posterior systems are essential for the generation of saccades.

When more careful measurements are taken, several specific deficits are revealed by studies in which the SC, the FEFs, or the MEFs have been removed. One set of studies has examined the distribution of saccadic latencies to single visual targets. These studies have established an interesting fact – the distribution of saccadic latencies made to single targets is bimodal. The first mode peaks around 100 ms. These short-latency saccades have been termed express saccades. The second mode peaks around 135–145 ms, and the eye movements have been called regular saccades. Following SC lesions that disrupt the input from posterior cortex to the brain stem, express saccades are entirely eliminated. Furthermore, the latency of the regular saccades also becomes longer. By contrast, when the FEFs or the MEFs are ablated, the bimodal distribution of saccades is unaffected. Thus, it appears that the posterior system is essential for the generation of express saccades.

Using several other tests, we have established that the FEFs play a significant role in the selection of targets for the generation of saccadic eye movements as well as in the planning of sequences of saccadic eye movements. Following lesions of the FEFs, monkeys showed long-term deficits in both target selection and in the execution of sequences of saccadic eye movements.

Eye Movements Elicited by Two Visual Targets and by Electrical Stimulation at Two Brain Sites

Experiments have assessed eye movements elicited by the simultaneous presentation of two visual targets and compared such performance with what is obtained when two brain sites are electrically stimulated simultaneously. The way such experiments have been carried out is to first map the receptive fields of neurons located at the tips of two electrodes and to then place a visual target into each receptive field. The eye movements elicited this way can then be compared with the eye movements elicited by electrically stimulating the neurons at the tip of each electrode. Results obtained at two sites – area V1 and the FEFs –-are compared here. When in area V1, the location of the two receptive fields at the tip of the two electrodes is close to 90 angular deg relative to the center of fixation. Saccades are made to visual targets presented at either one or the other location; vector-averaged saccades that land the eye in between these two points are rare. The same result is obtained with electrical stimulation of V1. By contrast, at two sites in the FEFs or in the SC, electrical stimulation produces vector-averaged saccades as depicted in Figure 3. These observations suggest that electrical stimulation in V1 produces predominantly a visual percept, whereas in the FEFs and the SC it produces a motor command.

Eye Movements Elicited by Combined Visual and Electrical Stimulation

Using similar procedures, we can determine what happens when a visual target placed into the receptive field of a set of neurons is coupled with the delivery of electrical current. When this is done in the FEFs or the MEFs, combining electrical stimulation with the visual target always leads to facilitation; the likelihood of producing a saccade is significantly increased. This occurs even when the current levels used are below the threshold of eliciting a saccadic eye movement in the absence of a visual stimulus.

This effect can also be obtained in some subregions of the LIP. LIP is a rather complex structure, however, and in some subregions the electrical stimulation results, instead, in a decreased disposition to make a saccade to the target placed into the receptive or motor field of the stimulated cells. Furthermore, in

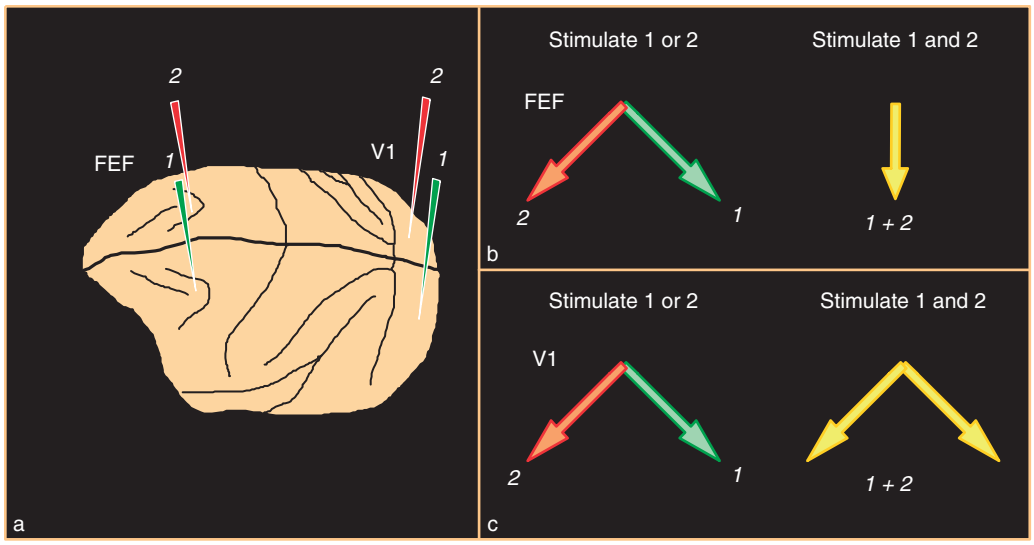

Figure 3 Effects of stimulating two sites in the frontal eye fields (FEFs) and in area V1: (a) top view; (b) FEF; (c) V1.

yet some other subregions of LIP, electrical stimulation delays the initiation of a saccade. This happens in those regions of LIP, as previously noted, in which neurons discharge vigorously while fixation is maintained. The characteristics of these cells and the delay produced by electrical stimulation suggest that these subregions in LIP play a significant role in determining the time when the eye movement should be initiated. Presumably this occurs subsequent to when the object under view has been satisfactorily analyzed.

When a similar procedure is used in V1, an interesting finding emerges. Stimulation in the lowers layers, especially in layer 5, which contains neurons that project to the SC, produces facilitation; the monkey is more likely to make a saccade to the visual target in the receptive field when it is paired with electrical stimulation. By contrast, in the upper layers electrical stimulation produces interference. Systematic variation of the location of the visual target relative to the center of the receptive field shows that the interference is effective only when the visual target is within the receptive field. Thus, the area affected by the electrical stimulation is confined to the receptive field.

The Visual Percept Elicited by Electrical Stimulation of Area V1

These considerations suggest that electrical stimulation of the SC, the FEFs, and the MEFs produces a command to generate a saccadic eye movement, but that in V1 it creates predominantly a visual percept that under some circumstances, as defined by other brain areas, can lead to an eye movement. This, then, raises the question as the nature of the visual percept created by electrical stimulation of V1. It is well known that electrical stimulation of this area in humans produces a starlike image called a phosphene. To assess what percepts are produced in monkeys, procedures have been used in which various stimulus parameters are varied for a visual stimulus which is paired with electrical stimulation of neurons whose receptive fields are at another location. The monkeys have been extensively trained to make saccadic eye movements to the larger and brighter of two simultaneously presented visual stimuli. The size and contrast of the visual stimulus is varied while the current delivered through the microelectrode to the neurons whose receptive fields are at another location is kept constant. Under these conditions, when the visual target is very small and dim, the monkey makes saccades to the receptive field site of the electrically stimulated neurons, whereas when the visual target is large and bright, the monkey makes a saccadic eye movement to the visual stimulus. When the percepts elicited by the visual stimulus and the electrical stimulus are similar, the monkey is expected to select them with equal probability, just as the animal does when two identical visual stimuli are presented. Thus, by systematically varying the size and contrast of the visual target we can establish the point of equivalence. These studies show that when 20–120 μamp current pulses are delivered at 200 Hz within 80 ms trains, the visual percept created has 6–12% contrast and 15–20 min of visual angle in size at locations 2.5–3.5° from the fovea.

The Role of Cortical GABAergic Inhibitory Circuits in the Generation of Saccadic Eye Movements

In the natural environment subsequent to each saccadic eye movement, dozens of objects impinge on the

retina. Two central tasks with each fixation are the analysis of the scene and a decision as to where to look next. To be able to generate an accurate saccade to a desired location, a decision has to be made not only where to look but also where not to look; otherwise, all the visual signals would probably create a vector-averaged saccadic eye movement that would not be helpful in analyzing the visual scene. To maintain high accuracy for the direction of gaze, the nervous system relies on a complicated interplay between excitatory and inhibitory circuits. The substantia nigra makes extensive GABAergic inhibitory connections with the SC. The planning of a saccade requires that a specific region of the SC be disinhibited while inhibition is maintained in other regions. Furthermore, the disinhibited region must also receive an excitatory signal. When muscimol, a γ-aminobutyric acid (GABA) agonist, is injected locally in the SC, an animal has difficulties generating saccades with vectors represented by the injected neurons. By contrast, the administration of bicuculline, a GABA antagonist, produces irrepressible saccades with vectors coded by the affected neurons.

The administration of these two agents, muscimol and bicuculline, in the cortex has clarified the role of GABAergic circuits in vision and eye movement. Such studies have examined how both visual processing and saccade generation are affected. The results show that in the FEF muscimol and bicuculline produce effects similar to those seen in the SC. Muscimol interferes with saccade generation, and bicuculline produces irrepressible saccades. The processing of visual information is only mildly affected. The administration of similar dosages in LIP has little effect on either saccade generation or visual processing. In area V1, both agents produce a major disruption in visual analysis; when monkeys are asked to detect or discriminate visual stimuli they fail when either muscimol or bicuculline is administered. These findings suggest that the GABAergic circuits in the FEFs and SC play a central role in eye-movement generation, whereas in V1 they play a major role in analyzing the visual scene.

Outline of the Basic Tasks and the Neural Circuitry Involved in the Generation of Visually Guided Saccadic Eye Movements

Figure 4 outlines the basic steps and brain areas involved in the generation of visually guided saccadic eye movements. Five steps are identified and formulated in terms of task questions:

1. What are the objects in the visual scene? During maintained fixation, the visual scene is analyzed to determine what the objects are in the scene. This is accomplished by numerous brain areas that include V1, V2, V4, LIP, and inferotemporal area (IT).
2. Which object to look at? A decision process has to be generated that picks out the object which is to be the target of the saccadic eye movement. The prime areas involved are V1, V2, LIP, FEF, and MEF.
3. Which object not to look at? To be able to generate an accurate saccade to a desired location, it is necessary to inhibit regions activated by visual stimuli toward which saccades should not be generated. The areas that play a role in this decision process include V1, V2, and LIP.
4. Where are the objects in space? The position of the object in space to be looked at must be computed. Prime candidates for this computation include V1, V2, FEF, and SC, in which retinocentric coordinates are arranged in topographic order.
5. When to initiate the saccadic eye movement? The decision to initiate a saccade presumably arises when the central scene at the fixation site has been satisfactorily analyzed. Areas LIP and MEF have been shown to contain neurons whose activity must cease before an eye movement can be initiated.

Figure 5 outlines the basic circuitry involved in these tasks. Three major classes of retinal ganglion cells that project to area V1 via the LGN are shown. The midget and parasol system remains separate only in part in V1; there are many neurons that receive a convergent input from these to channels. The layer 5 complex cells in V1, which are driven predominantly by the parasol system, project to the stratum opticum of the SC. Cortico-cortical projections show that area MT and the parietal lobe receive inputs that are dominated by the parasol system, whereas area V4 and the temporal lobe receive a mixed input from the midget and parasol systems. The question marks in the figure indicate pathways and areas whose composition and destination are not yet known.

The posterior system involves cortical areas in posterior cortex, which sends its signal to the brain stem through the SC. Furthermore, the SC is under inhibitory control by the substantial nigra, a control that is modulated by cortex. The projection from the retina to the superficial layers of the SC does not appear to have direct access to the deeper layers of this structure. The anterior system from the FEF and MEF of the frontal lobe connects with the SC, but also has direct projections to the brain stem oculomotor centers. The FEF contains its own inhibitory circuits for keeping saccadic eye-movement generation in check. The coordinate execution of various types of saccadic eye movements is accomplished in part by the extensive interconnections, both feed-forward and feedback, among the brain structures depicted in Figure 5.

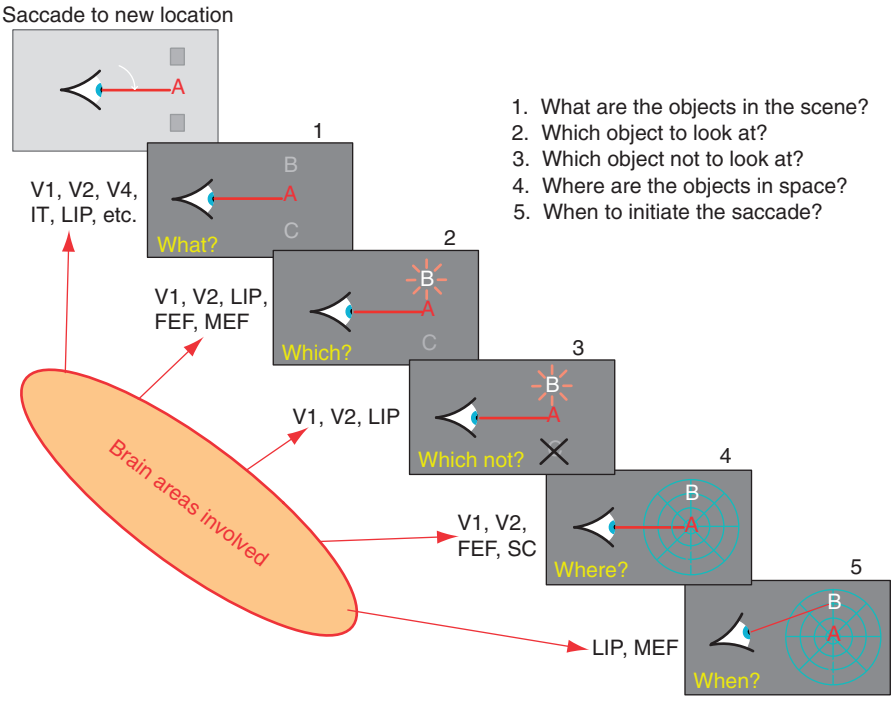

Figure 4 Five basic steps and the brain areas involved in the generation of a visually guided saccadic eye movements, formulated in terms of task questions. FEF, frontal eye field; IT, inferotemporal visual areas; LIP, lateral intraparietal area; MEF, medial eye field; SC, superior colliculus.

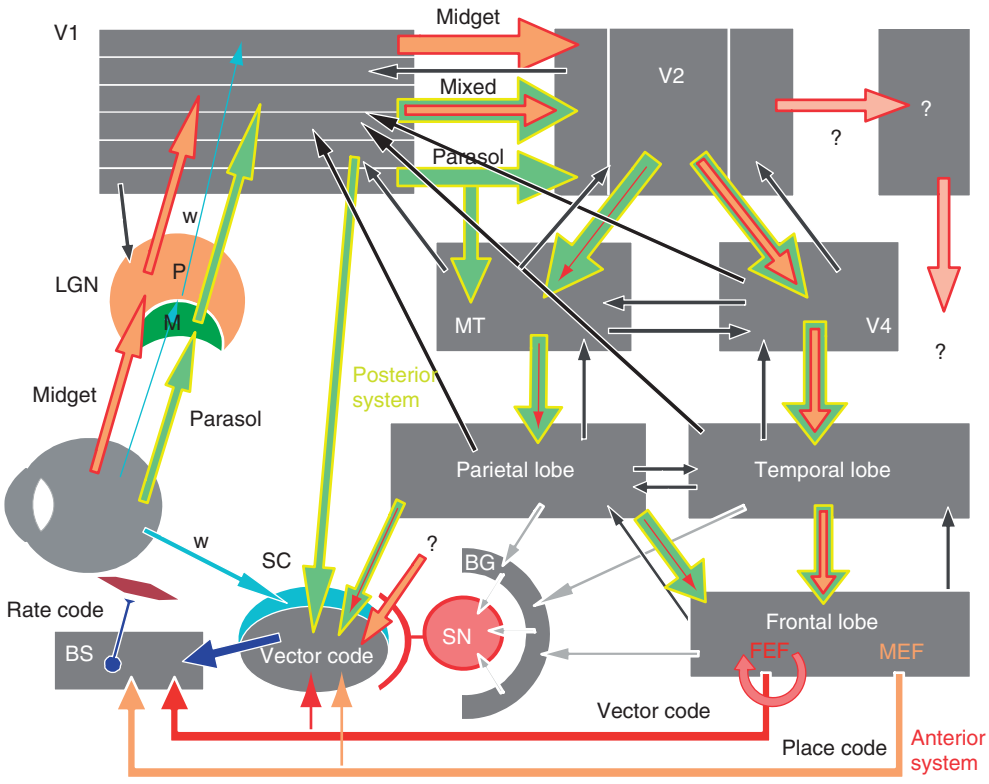

Figure 5 Basic circuitry involved in the tasks listed in **Figure 4**. BS, brain stem; FEF, frontal eye field; LGN, lateral geniculate nucleus; M, magnocellular; MEF, medial eye field; MT, middle temporal visual area; P, parvocellular; SC, superior colliculus; SN, substantia nigra; w, W-like cells.

See also: Basal Ganglia and Oculomotor Control; Brainstem Control of Eye Movements; Cerebellum and Oculomotor Control; Executive Functions: Eye Movements and Neuropsychiatric Disorders; Eye and Head Movements; Eye Movement Disorders; Oculomotor Control: Anatomical Pathways; Oculomotor System: Models; Thalamus and Oculomotor Control; Vergence Eye Movements; Vestibulo-Ocular Reflex.

Further Reading

Andersen RA, Essick GK, and Siegel RM (1985) Encoding of spatial location by posterior parietal neurons. *Science* 230: 456–458.

Glimcher PW (2003) The neurobiology of visual-saccadic decision making. *Annual Review of Neuroscience* 26: 133–179.

Kurylo DD and Skavenski AA (1991) Eye movements elicited by electrical stimulation of area PG in the monkey. *Journal of Neurophysiology* 65: 1243–1253.

Mann SE, Thau R, and Schiller PH (1988) Conditional task-related responses in monkey dorsomedial frontal cortex. *Experimental Brain Research* 69: 460–468.

Schall JD and Thompson KG (1999) Neural selection and control of visually guided eye movements. *Annual Review of Neuroscience* 22: 241–259.

Schiller PH (1997) Past and present ideas about how the visual scene is analyzed by the brain. In: Rockland KS, Kaas JH, and Peters A (eds.) *Cerebral Cortex, Vol. 12: Extrastriate Cortex in Primates*, pp. 59–90. New York: Plenum Press.

Schiller PH (1998) The neural control of visually guided eye movements. In: Richards JE (ed.) *Cognitive Neuroscience of Attention: A Developmental Perspective*, pp. 3–50. Mahwah, NJ: Lawrence Erlbaum.

Schiller PH and Tehovnik EJ (2001) Look and see: How the brain moves your eyes about. *Progress in Brain Research* 134: 127–142.

Tehovnik EJ, Slocum WM, Carvey CE, and Schiller PH (2005) Phosphene induction and the generation of saccadic eye movements by striate cortex. *Journal of Neurophysiology* 93: 1–19.

Tehovnik EJ, Slocum WM, and Schiller PH (2003) Saccadic eye movements evoked by microstimulation of striate cortex. *European Journal of Neuroscience* 17: 870–878.

Tehovnik EJ, Sommer MA, Chou IH, Slocum WM, and Schiller PH (2000) Eye fields in the frontal lobes of primates. *Brain Research: Brain Research Reviews* 32: 413–448.

Cortical Plasticity and Learning: Mechanisms and Models

D V Buonomano and H A Johnson, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sensory and motor processing, as well as cognition in general, is critically dependent on cortical computations. A fundamental aspect of cortical function is the ability of the cortex to adapt and remodel itself in an experience-dependent fashion. Thus, to a certain extent, cortical circuits are optimized to perform specific computations. Experience-dependent cortical plasticity has been demonstrated in sensory and motor areas, although most studies have been performed in primary sensory areas, in large part because of the relative ease in characterizing cortical responses and manipulating sensory experience. Two general features of primary sensory cortex are generally examined: topography and stimulus selectivity. Topography refers to the fact that neighboring sites in the cortex represent neighboring areas of the sensory surface. Stimulus selectivity refers to the fact that neurons respond to a specific subset of sensory features, such as tone frequency, skin location, or orientation of visual edges. Stimulus sensitivity is generally defined by the receptive field characteristics of a neuron.

Experience-Dependent Cortical Plasticity

Experience-dependent changes in the receptive field and cortical map topography have been reported throughout various developmental stages and sensory modalities. The common principle in these studies is that plasticity seems to be driven by the ability of cortical networks to detect patterns and changes in the spatiotemporal structure of the input to which they are exposed. More specifically, the correlational structure of the inputs to the cortex, or lack thereof, seems to be a primary determinant in cortical organization.

It is well established that the removal of patterned input during development has profound consequences on cortical function. In kittens, for example, eyelid suture leads to a dramatic decrease in the number of neurons in the visual cortex responsive to the deprived eye. Changes in the structure of the input patterns also alter cortical function; rearing kittens in environments in which they are predominately exposed to lines of a certain orientation will result in an overrepresentation of neurons selective to this orientation. Similar experience-dependent alterations in cortical function have been reported in other modalities. Presentation of auditory noise during early ontogenesis alters the normal development of cortical tonotopic maps, and removing whiskers early in development produces changes in the functional somatotopy of the barrel cortex.

In adults, sensory experience also elicits cortical plasticity. These experiential changes can consist of complete sensory deprivation or more-subtle changes in the input structure, such as those that would occur during perceptual learning. Merzenich and colleagues demonstrated in monkeys that after amputation of a finger or transection of a peripheral nerve, the cortical areas previously responsive to that finger would develop responses to neighboring fingers. The same mechanisms underlying reorganization induced by broad sensory deprivation are also assumed to causally contribute to certain forms of learning. This has been directly examined in the context of perceptual learning. For example, studies in auditory cortex suggest that frequency discrimination learning results in increased representation of the trained frequencies, and orientation discrimination learning is associated with shifts in the orientation tuning of neurons in area V4, consistent with improved discrimination of the trained orientations.

One shared principle in the above studies is the relevance of the correlational structure of the inputs. The correlational structure refers to whether different afferents are generally activated in a synchronous or near-synchronous manner. For example, two close points on the skin are more likely to be coactivated than are two distant points. Theoretical work has demonstrated that the information embedded in the correlational structure of the inputs driving cortical activity can be used to guide the formation or reorganization of topographic maps and their receptive fields. The importance of temporal correlations in the plasticity of cortical receptive fields was directly demonstrated by experiments in which the proximal or distal portions of the fingers of monkeys were consistently co-stimulated – that is, there was an artificially high likelihood that the portions of the adjacent fingers would be co-stimulated. Cortical maps of the primary sensory cortex from these animals revealed a high incidence of neurons that had multiple-digit receptive fields (Figure 1).

Together, a large body of experimental data indicates that the amount of stimulation and the structure of the input patterns to the cortex play instructive roles in the formation of topographic maps and receptive fields. Furthermore, it is hypothesized that what the spatiotemporal patterns of action potentials arriving in the cortex represent – whether they are generated

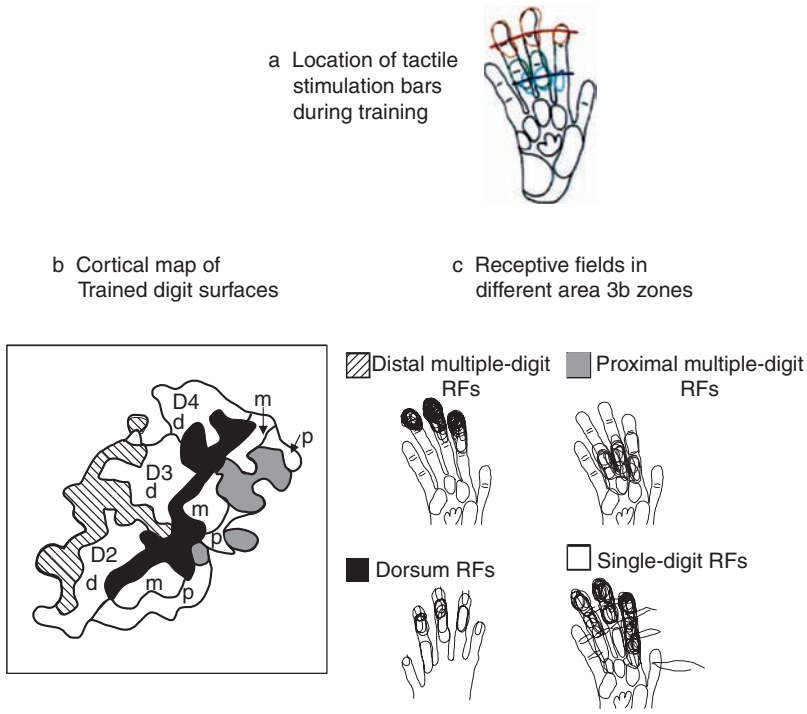

Figure 1 Cortical map reorganization. After monkeys received behavioral training where digits 2 through 4 were consistently co-stimulated, the receptive fields and map of hand representation in area 3b of somatosensory cortex was reorganized to reflect the experienced training pattern. (a) Two tactile stimulation bars alternately made simultaneous contact with the distal (red line) or proximal (blue line) portions of digits 2 through 4. (b) In normal untrained monkeys, cortical maps rarely show multiple-digit receptive fields. In contrast, this cortical map of a trained monkey shows areas of multiple-digit receptive fields which are specific to either the distal (striped area) or proximal (grey area) portions of the digits. D2, D3, and D4 are areas representing digits 2, 3, and 4, respectively. Areas responding to distal, medial, and proximal digit receptive fields are designated d, m, and p, respectively. (c) Multiple-digit receptive fields (RFs) in a trained monkey. Adapted from Wang X, Merzenich MM, Sameshima K, and Jenkins WM (1995) Remodeling of hand representation in adult cortex determined by timing of tactile stimulation. *Nature* 378: 71–75.

by auditory, somatosensory, or visual stimuli – may not be particularly relevant. This notion is supported by experiments that show that rerouting visual information to the auditory cortex produces 'auditory' neurons that respond selectively to orientation of bars of light. This suggests that there may be a set of universal synaptic and cellular learning rules that underlie cortical organization and plasticity.

Models of Cortical Plasticity

Given the inherent complexity of cortical networks and the computations performed, theoretical approaches have proven critical in establishing a framework to study cortical plasticity by generating predictions and demonstrating, as a proof of principle, that a given set of learning rules is sufficient to account for experimental observations. Computational models of cortical plasticity have focused primarily on two interrelated issues: (1) describing the rules that govern changes in synaptic strength and (2) characterizing how cortical features, such as topography, receptive fields, and stimulus selectivity, emerge in an experience-dependent manner.

Structure from Correlations

To understand the models of cortical plasticity, it is useful to consider the problem of stimulus selectivity from the perspective of a single neuron, for example, a neuron that receives four inputs (Figure 2). If all these inputs are always coactivated (100% correlated), they are functionally indistinguishable; that is, from the perspective of the postsynaptic neuron, it is not possible to discriminate between them. On the other hand, if one pair of neurons is always activated in a correlated manner and the other pair is also highly correlated, yet both pairs are rarely coactive, they may represent two distinct and unrelated sensory stimuli (e.g., stimulus A and stimulus B). In this case, how would the postsynaptic neuron develop a selective response to one but not the other? From the perspective of an individual neuron, two related learning rules must be in place for stimulus selectivity to emerge in a general and robust manner:

1. Associative synaptic plasticity: If a neuron is to develop selective responses to a subset of its synaptic inputs (representing a subset of external

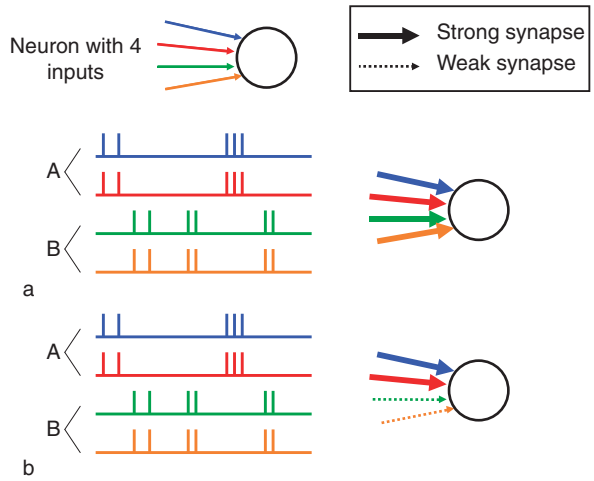

Figure 2 Cortical learning rules. A neuron receives four sources of input, whose synapses onto the neuron are initially weak. (a) Associative synaptic plasticity. The activation pattern of the four inputs is organized into two subsets (A and B), which are active independent of each other and can be thought of as representing two separate stimuli, each consisting of two inputs with correlated activity. Because of this correlation, the inputs from group A will cooperate to drive the postsynaptic neuron, and both inputs onto the neuron will be strengthened in accordance with associative plasticity rules. Since the two inputs representing stimulus B are also correlated, a similar processes will occur. The result is potentiation of all inputs and a postsynaptic neuron that will respond robustly to both stimuli but not exhibit any stimulus selectivity. (b) Associative synaptic plasticity with competition. However, if there is an additional mechanism, such as postsynaptic normalization or a sliding long-term potentiation/long-term depression threshold, competition occurs between the inputs representing the different stimuli, resulting in potentiation of one set of inputs, depression of the other, and the formation of a stimulus-selective neuron.

stimuli), some synapses have to become stronger. In computational models this is achieved by assuming that coincident presynaptic inputs cooperate to drive postsynaptic depolarization, which results in associative synaptic potentiation.

2. Competition: The strengthening of synapses alone will not necessarily lead to selectivity. Thus, the second condition for cortical plasticity is that uncorrelated groups of inputs compete for control of the postsynaptic cell. Functionally, this competition takes the form of decreases in the synaptic strength of a subset of inputs.

Associative synaptic plasticity is sufficient to account for the strengthening of the inputs that will comprise the preferred stimulus. Specifically, if all inputs onto a postsynaptic neuron are initially weak, when a subset of these synapses is consistently activated together, they may cooperate to drive postsynaptic activity, leading to the selective potentiation of this subset. However, if a second independent subset is also consistently coactivated, they will also be potentiated. Without competition, the cell will be driven by more than one input pattern and thus will not be stimulus-selective (Figure 2(a)). Hence, any set of synaptic learning rules capable of accounting for the emergence of stimulus-selective responses must also provide a mechanism by which different inputs 'compete' for control of the postsynaptic cell (Figure 2(b)). Two general classes of models have been proposed to account for this competition: postsynaptic normalization and a sliding long-term potentiation/long-term depression (LTP/LTD) threshold.

Postsynaptic normalization If two stimuli, A and B, are capable of engaging associative plasticity mechanisms, the synaptic weights from both inputs will grow until some saturation point is achieved. Postsynaptic normalization assumes that, through a homeostatic mechanism, neurons actively maintain a fixed level of total synaptic input. Although there is no consensus regarding the mechanisms by which such a process would be expressed, for didactic purposes one could imagine that the total number of postsynaptic glutamate receptors was constant. In this manner, if one synapse is potentiated by a factor X, all synapses would become weaker by a specific fraction (a function of X and the total number of synapses). According to this model, all synapses would be depressed by the same amount after each stimulus presentation, but the synapse that started off a bit stronger would eventually dominate because it would undergo proportionally more LTP after each stimulus presentation.

Sliding threshold for LTP and LTD A second model of cortical plasticity is referred to as the sliding-threshold learning rule, or BCM (representing the initials of the authors of the model, Bienenstock, Cooper, and Munro). In this model, associative plasticity and competition are tightly linked. Here, presynaptic activity can result in either LTP or LTD, depending on the voltage of the postsynaptic neuron. Specifically, if the postsynaptic cell is depolarized above a critical value (the LTP/LTD threshold, θ) during presynaptic activity, LTP ensues; if the postsynaptic voltage is below this threshold, LTD ensues. Mathematically this can be expressed as follows:

$$\Delta W_{\text{pre, post}} = \alpha V_{\text{pre}}(V_{\text{post}} - \theta)$$

where V reflects the degree of activation of the pre- or postsynaptic cell, ΔW is the change in the weight of the synapse, and α is a constant which controls the rate of synaptic change. If θ is equal to zero, we are left with the standard Hebbian term of synaptic plasticity.

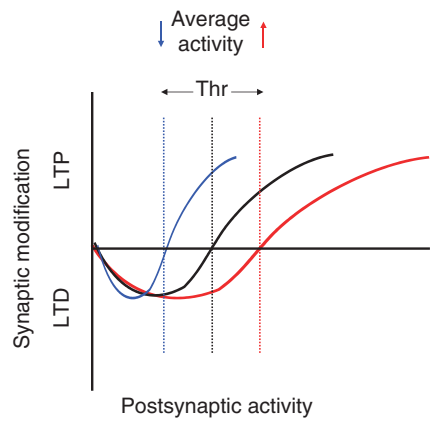

Figure 3 Sliding long-term potentiation–long-term depression (LTP–LTD) threshold. The BCM model (named for its authors, Bienenstock, Cooper, and Munro) postulates a sliding threshold for the induction of LTP–LTD. Synaptic input which drives postsynaptic activity above this threshold level undergoes LTP, and one which drives postsynaptic activity below this level undergoes LTD. If the average activity rate for the neuron increases over time (red trace), the threshold slides to the right, thus requiring more postsynaptic activity for the induction of LTP and facilitating the induction of LTD. The reverse is also true: if the average activity rate decreases (blue trace), the threshold slides to the left, and less activity is needed for inputs to drive enough postsynaptic activity to induce potentiation.

When θ is positive, the level of postsynaptic activity determines whether LTP or LTD will occur. Competition arises from the fact that θ is not a constant but varies as a running average of activity in the postsynaptic cell (Figure 3). Specifically, θ changes in a manner proportional to the average postsynaptic activity, albeit at a slow rate (hypothesized to occur on the scale of hours). Assuming a state in which all the inputs onto the postsynaptic neuron are initially strong, the cell will be nonselective, responding to any and all stimuli. However, because the average cell activity rate will also be high, precisely because of the absence of selectivity, θ will increase. As this happens, there will be a point when the weaker of the two inputs will produce a depolarization that falls below θ and will undergo LTD. This will create a divergence in the synaptic weights representing different stimuli, resulting in a stimulus-selective neuron.

These models differ significantly in relation to the competitive mechanisms but share a number of common features, particularly in regard to the associative potentiation of active synapses. In addition, both models rely on the calculation of some global (cellwide) parameter capable of influencing all synapses of the postsynaptic cell. In the normalization model, a global mechanism must estimate the sum of all synaptic weights onto a cell and adjust these weights in a cellwide manner. While there is no direct experimental evidence for this exact mechanism, related forms of plasticity have been observed experimentally. Specifically, a form of homeostatic plasticity referred to as synaptic scaling appears to adjust the strength of all synapses in a manner proportional to overall activity levels. Alternatively, the sliding-threshold model requires neurons to estimate their activity level, and the result of this estimate must be locally available at each synapse to control for the induction of LTP or LTD. Experimental evidence also provides some support for such a sliding-threshold; however, to date, these studies are indirect.

In addition to postsynaptic normalization and the sliding-threshold model, it has been proposed that spike timing-dependent plasticity (STDP) may provide a mechanism for competition between stimuli and thus the formation of stimulus-selective responses. STDP refers to an experimentally well-characterized form of associative synaptic plasticity in which presynaptic spikes arriving before a postsynaptic spike produce LTP, while those arriving after result in LTD. This state of affairs can result in active competition between groups of presynaptic cells representing different inputs. However, in contrast to postsynaptic normalization and the sliding-threshold rules, competition is limited to stimuli that occur in close temporal proximity (within the time window for STDP). Thus STDP by itself does not represent a general learning rule for cortical organization and plasticity.

At the theoretical level, both the postsynaptic normalization and sliding-threshold models provide a general mechanism that is sufficient to account for the formation of stimulus-selective neurons in a self-organizing manner. Note, however, that the mechanisms discussed would not explain why neighboring neurons would have similar receptive field characteristics, because each postsynaptic neuron is independent of the other. Topography is postulated to arise from the lateral interaction within cortical networks. Specifically, it has been shown that if neurons excite their near neighbors and inhibit their far ones, 'macroscopic' features of cortical maps can be simulated, including the topography and the presence of ocular dominance stripes observed in primary visual cortex.

Mechanisms and Cellular Correlates of Cortical Plasticity

Normal cortical development and cortical plasticity rely on a complex set of activity-dependent and molecular guidance mechanisms. Although these mechanisms are incompletely understood, there is significant agreement between the experimental findings and the

theoretical models, particularly in regard to the role of synaptic plasticity in experience-dependent cortical plasticity. The majority of excitatory neurons in the cortex that have been examined exhibit both LTP and LTD. The LTP is typically associative in nature and dependent on the *N*-methyl-D-aspartate (NMDA) subtype of glutamate receptors, which act as coincidence detectors. Experiments in both the visual system and the barrel cortex (the area of the rodent somatosensory cortex that represents the whiskers) reveal alterations in sensory-evoked responses after deprivation consistent with LTP or LTD. In the barrel cortex model of cortical plasticity, removal of all but one or two whiskers produces enhanced responses to the spared whiskers and decreased responses to the cut whiskers. Evidence for the role of LTD in decreasing the responsiveness comes from the observation that the induction LTD in this pathway is occluded after sensory deprivation. Pharmacological and genetic manipulations that alter normal NMDA receptor function (or downstream biochemical pathways) and impair the induction of associative LTP also impair cortical plasticity. For example, pharmacological blockade of NMDA receptors prevents the normal reorganization of the barrel cortex observed after removing whiskers. Also, in the visual cortex, the pattern of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor phosphorylation produced by LTP/LTD *in vitro* resembles those patterns produced *in vivo* in response to protocols that induce cortical plasticity. These and other experiments demonstrate a role of associative, NMDA receptor-dependent LTP and LTD in cortical plasticity.

Theoretical studies have focused primarily on the role of synaptic plasticity and competition as the mechanisms underlying cortical plasticity. However, it is clear that changes in excitatory synaptic strength alone provide an incomplete picture. For example, deprivation of sensory inputs is known to result in changes in dendritic and axonal morphology, intrinsic membrane excitability, as well as changes in the levels of γ-aminobutyric acid-mediated inhibition. Thus, cortical plasticity is believed to rely on complex and highly orchestrated changes at many cellular and synaptic loci in parallel. The mechanisms may be dependent on the forms of plasticity being examined, the cortical areas involved, as well as the developmental stage. Indeed, evidence suggests that cortical plasticity induced by sensory deprivation relies on different loci in young and adult animals. Specifically, early in development, depression and potentiation of the deprived and spared inputs, respectively, is observed. In adults, however, the potentiation of the spared inputs appears to be the dominant response to deprivation. In addition, there is evidence suggesting that in adults plasticity relies more heavily on synaptic plasticity occurring in supra- and infragranular layers of the cortex, and less on thalamocortical plasticity.

Cortical Plasticity: Spatiotemporal Stimuli

Most studies of cortical plasticity have focused primarily on the neural response to simple spatial stimuli – such as somatic receptive fields or ocular dominance preference. However, a fundamental aspect of cortical function is the processing and categorization of complex spatiotemporal stimuli, such as animal vocalizations. It is well established that perceptual learning of spatiotemporal stimuli results in robust improvement in the ability to discriminate between complex stimuli, and it is widely accepted that perceptual learning is dependent on cortical plasticity.

In contrast to changes in spatial receptive field properties, which can be explained in terms of simple potentiation and depression of the inputs to a neuron, the mechanisms underlying the response properties to complex stimuli are not well understood. One general class of models proposes that the cortical circuits are intrinsically able to process spatial and temporal stimuli as a result of the inherent complexity of cortical networks and the presence of time-dependent neuronal properties. These models, referred to as state-dependent networks, consist of recurrently connected excitatory and inhibitory neurons and the presence of time-dependent properties, particularly short-term synaptic plasticity. In the same manner that long-term plasticity provides a trace or memory determined by a set of events that occurred sometime in the past, short-term synaptic plasticity provides a 'trace' that a given synapse was recently (on the scale of hundreds of milliseconds) active. Consequently, the state of a cortical network is dependent on the previous stimuli to which it was exposed, and the response to subsequent inputs will inherently encode both the stimulus and the context in which it was presented. Relatively few experiments have examined the neural mechanisms underlying experience-dependent plasticity to complex stimuli. However, it has been shown in primates and ferrets that the neural responses to animal vocalizations are indeed dependent on experience. Furthermore, exposure to specific auditory patterns paired with stimulation of the nucleus basalis results in reorganization of primary auditory cortex and changes in the neuronal sensitivity to temporal sequences. Additional studies are needed to elucidate the mechanisms by which neurons develop selective responses to spatiotemporal stimuli.

Conclusion

Cortical plasticity allows circuits to flexibly adjust and optimize processing in a manner that reflects the environmental and behavioral demands. This ability encompasses both relatively simple features such as the formation of topographic maps and the spatial receptive field characteristics of neurons, as well as the processing of complex stimuli, such as speech and vocalization recognition. Theoretical and experimental studies have established that associative forms of synaptic plasticity and synaptic competition are critical to cortical plasticity. While the molecular and cellular mechanisms underlying associative synaptic plasticity are partially understood, the mechanisms underlying competition remain to be elucidated. Similarly, much remains to be learned regarding the contribution of other cellular and synaptic loci, including intrinsic neuronal excitability and inhibition, in cortical plasticity.

See also: Adult Cortical Plasticity; Hebbian Plasticity; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Neuronal Plasticity after Cortical Damage; Plasticity and Activity-Dependent Regulation of Gene Expression; Spontaneous Patterned Activity in Developing Neural Circuits; Synaptic Mechanisms of Learning; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Bienenstock EL, Cooper LN, and Munro PW (1982) Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience* 2: 32–48.

Blakemore C and Cooper GF (1970) Development of the brain depends on the visual environment. *Nature* 228: 477–478.

Buonomano DV and Merzenich MM (1995) Temporal information transformed into a spatial code by a neural network with realistic properties. *Science* 267: 1028–1030.

Buonomano DV and Merzenich MM (1998) Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* 21: 149–186.

Feldman DE and Brecht M (2005) Map plasticity in somatosensory cortex. *Science* 310: 810–815.

Fox K and Wong RO (2005) A comparison of experience-dependent plasticity in the visual and somatosensory systems. *Neuron* 48: 465–477.

Heynen AJ, Yoon BJ, Liu CH, et al. (2003) Molecular mechanism for loss of visual cortical responsiveness following brief monocular deprivation. *Nature Neuroscience* 6: 854–862.

Kilgard MP and Merzenich MM (2002) Order-sensitive plasticity in adult primary auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America* 99: 3205–3209.

Miller KD, Keller JB, and Stryker MP (1989) Ocular dominance column development: Analysis and simulation. *Science* 245: 605–615.

Monfils MH, Plautz EJ, and Kleim JA (2005) In search of the motor engram: Motor map plasticity as a mechanism for encoding motor experience. *Neuroscientist* 11: 471–483.

Sharma J, Angelucci A, and Sur M (2000) Induction of visual orientation modules in auditory cortex. *Nature* 404: 841–847.

Song S and Abbott LF (2001) Cortical development and remapping through spike timing-dependent plasticity. *Neuron* 32: 339–350.

Turrigiano GG and Nelson SB (2004) Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience* 5: 97–107.

von der Malsburg C (1973) Self-organization of orientation sensitive cells in the striata cortex. *Kybernetik* 14: 84–100.

Wang X, Merzenich MM, Sameshima K, and Jenkins WM (1995) Remodeling of hand representation in adult cortex determined by timing of tactile stimulation. *Nature* 378: 71–75.

Cortical Processing of the Reward Value of Food

E T Rolls, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

The Pleasantness of the Taste of Food

The modulation of the reward value of a sensory stimulus, such as the taste of food, by motivational state – for example, hunger – is one important way in which motivational behavior is controlled. The subjective correlate of this modulation is that food tastes pleasant when a person is hungry, and tastes hedonically neutral when a person has eaten to satiety. To make the results relevant to understanding these processes and to the control of food intake in humans, complementary evidence is provided by neurophysiological studies in nonhuman primates, and by functional neuroimaging studies in humans. The macaque taste and olfactory system may be particularly relevant to understanding these processes in humans, for the primate taste system does not have a pontine taste area and direct subcortical connections, but instead relies on thalamic-to-cortical processing.

Rolls and colleagues have found that the modulation of taste-evoked signals by motivation is not a property found in early stages of the primate gustatory system. The responsiveness of taste neurons in the nucleus of the solitary tract and in the primary taste cortex (in the frontal operculum and anterior insula) is not attenuated by feeding to satiety. (Figure 1 shows a schematic diagram of a lateral view of some of the primate brain taste, olfactory, visual, and somatosensory pathways.) In contrast, in the secondary taste cortex, in the caudolateral part of the orbitofrontal cortex, it has been shown that the responses of the neurons to the taste of glucose decreased to zero while the monkey ate it to satiety, during the course of which the behavior turned from avid acceptance to active rejection. This modulation of responsiveness of the gustatory responses of the orbitofrontal cortex neurons by satiety could not have been due to peripheral adaptation in the gustatory system or to altered efficacy of gustatory stimulation after satiety was reached, because modulation of neuronal responsiveness by satiety was not seen at the earlier stages of the gustatory system, including the nucleus of the solitary tract, the frontal opercular taste cortex, and the insular taste cortex.

Sensory-Specific Satiety

In the secondary taste cortex, it was also found that the decreases in the responsiveness of the neurons were relatively specific to the food with which the monkey had been fed to satiety. For example, in seven experiments in which the monkey was fed glucose solution, neuronal responsiveness decreased to the taste of the glucose but not to the taste of blackcurrant juice (see example in Figure 2). Conversely, in two experiments in which the monkey was fed to satiety with fruit juice, the responses of the neurons decreased to fruit juice but not to glucose. An example of this sensory-specific decrease in the responsiveness of a neuron after feeding to satiety is shown in Figure 2. This neuron had taste, olfactory, and visual responses to food, and the satiety was produced by a food, blackcurrant juice, with of all these components.

This evidence shows that the reduced acceptance of food which occurs when food is eaten to satiety, and the reduction in the pleasantness of its taste, are not produced by a reduction in the responses of neurons in the nucleus of the solitary tract or frontal opercular or insular gustatory cortices to gustatory stimuli. Indeed, after feeding to satiety, humans reported that the taste of the food on which they had been satiated tasted almost as intense as when they were hungry, though much less pleasant. This comparison is consistent with the possibility that activity in the frontal opercular and insular taste cortices, as well as in the nucleus of the solitary tract, does not reflect the pleasantness of the taste of a food, but rather its sensory qualities, independently of motivational state. On the other hand, the responses of the neurons in the caudolateral orbitofrontal cortex taste area and in the lateral hypothalamus are modulated by satiety, and it is presumably in areas such as these that neuronal activity may be related to whether a food tastes pleasant, and to whether the food should be eaten. In addition to providing an implementation of sensory-specific satiety (probably by habituation of the synaptic afferents to orbitofrontal neurons with a time course of the order of the length of a course of a meal), it is likely that visceral and other satiety-related signals reach the orbitofrontal cortex (as indicated in Figure 1) (from the nucleus of the solitary tract, via thalamic areas), and there modulate the representation of food, resulting in an output that reflects the reward (or appetitive) value of each food.

It is an important principle that the identity of a taste, and its intensity, are represented separately from its pleasantness. Thus it is possible to represent what a taste is, and to learn about it, even when we are not hungry.

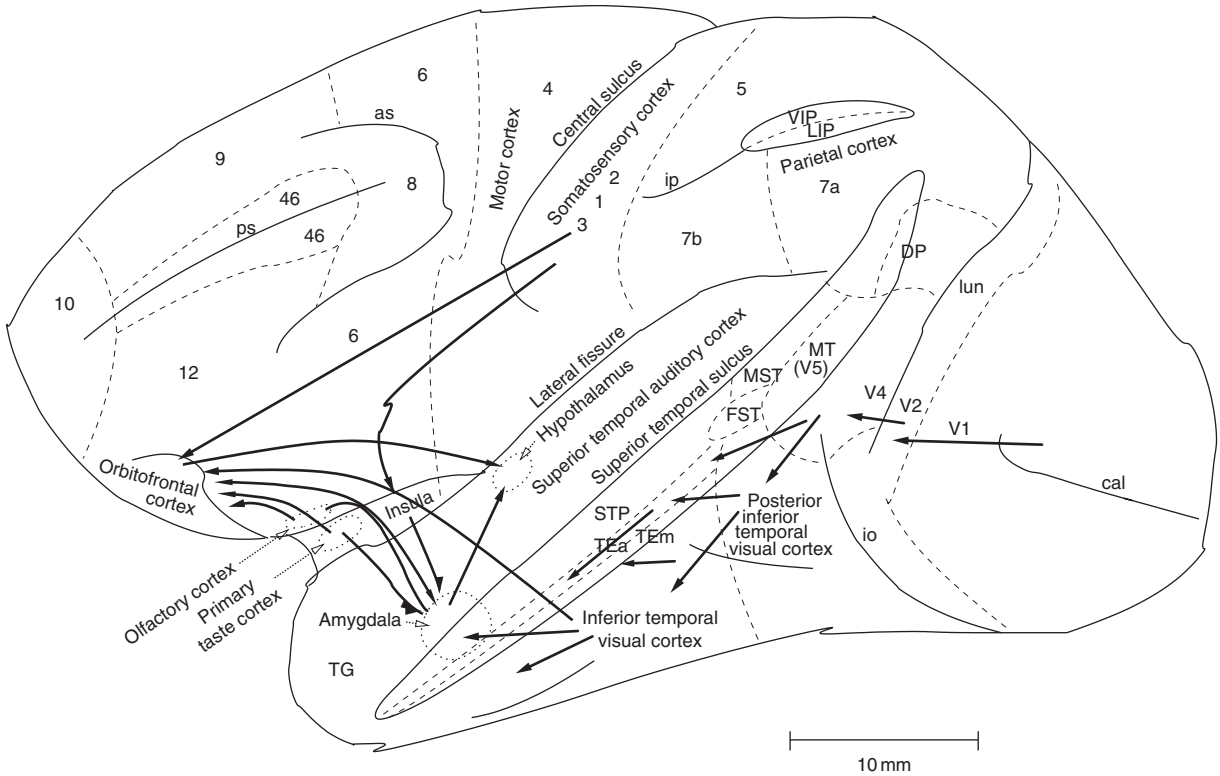

Figure 1 Some of the pathways involved in sensory processing involved in the control of food intake and emotion are shown on this lateral view of the brain of the macaque monkey. Connections from the primary taste and olfactory cortices to the orbitofrontal cortex and amygdala are shown. Connections are also shown in the 'ventral visual system' from V1 to V2, V4, the inferior temporal visual cortex, etc., with some connections reaching the amygdala and orbitofrontal cortex. In addition, connections from the somatosensory cortical areas 1, 2, and 3 that reach the orbitofrontal cortex directly and via the insular cortex, and that reach the amygdala via the insular cortex, are shown. as, arcuate sulcus; cal, calcarine sulcus; cs, central sulcus; lf, lateral (or Sylvian) fissure; lun, lunate sulcus; ps, principal sulcus; io, inferior occipital sulcus; ip, intraparietal sulcus (which has been opened to reveal some of the areas it contains); sts, superior temporal sulcus (which has been opened to reveal some of the areas it contains); AIT, anterior inferior temporal cortex; FST, visual motion processing area; LIP, lateral intraparietal area; MST, visual motion processing area; MT, visual motion processing area (also called V5); PIT, posterior inferior temporal cortex; STP, superior temporal plane; TA, architectonic area including auditory association cortex; TE, architectonic area including high-order visual association cortex, and some of its subareas TEa and TEm; TG, architectonic area in the temporal pole; V1–V4, visual areas V1–V4; VIP, ventral intraparietal area; TEO, architectonic area including posterior visual association cortex. The numerals refer to architectonic areas, and have the following approximate functional equivalence: 1, 2, 3, somatosensory cortex (posterior to the central sulcus); 4, motor cortex; 5, superior parietal lobule; 7a, inferior parietal lobule, visual part; 7b, inferior parietal lobule, somatosensory part; 6, lateral premotor cortex; 8, frontal eye field; 12, part of orbitofrontal cortex; 46, dorsolateral prefrontal cortex.

Representation of the Pleasantness of the Smell and Sight of Food in the Brain

It has also been possible to investigate whether the olfactory representation in the orbitofrontal cortex is affected by hunger, and thus whether the pleasantness of odor is represented in the orbitofrontal cortex. In satiety experiments, Critchley and Rolls showed that the responses of some olfactory neurons to a food odor are decreased during feeding to satiety with a food (e.g., fruit juice) containing that odor. In particular, seven of nine olfactory neurons that were responsive to the odors of foods, such as blackcurrant juice, were found to decrease their responses to the odor of the satiating food. The decrease was typically at least partly specific to the odor of the food that had been eaten to satiety, potentially providing part of the basis for sensory-specific satiety. It was also found for eight of nine orbitofrontal cortex neurons that had selective responses to the sight of food that they demonstrated a sensory-specific reduction in their visual responses to foods following satiation. These findings show that the olfactory and visual representations of food, as well as the taste representation of food, in the primate orbitofrontal cortex are modulated by hunger. Usually a component related to sensory-specific satiety can be demonstrated. Hunger does not influence the responses of neurons in the inferior temporal visual cortex to visual stimuli, so that the representation of the pleasantness or reward

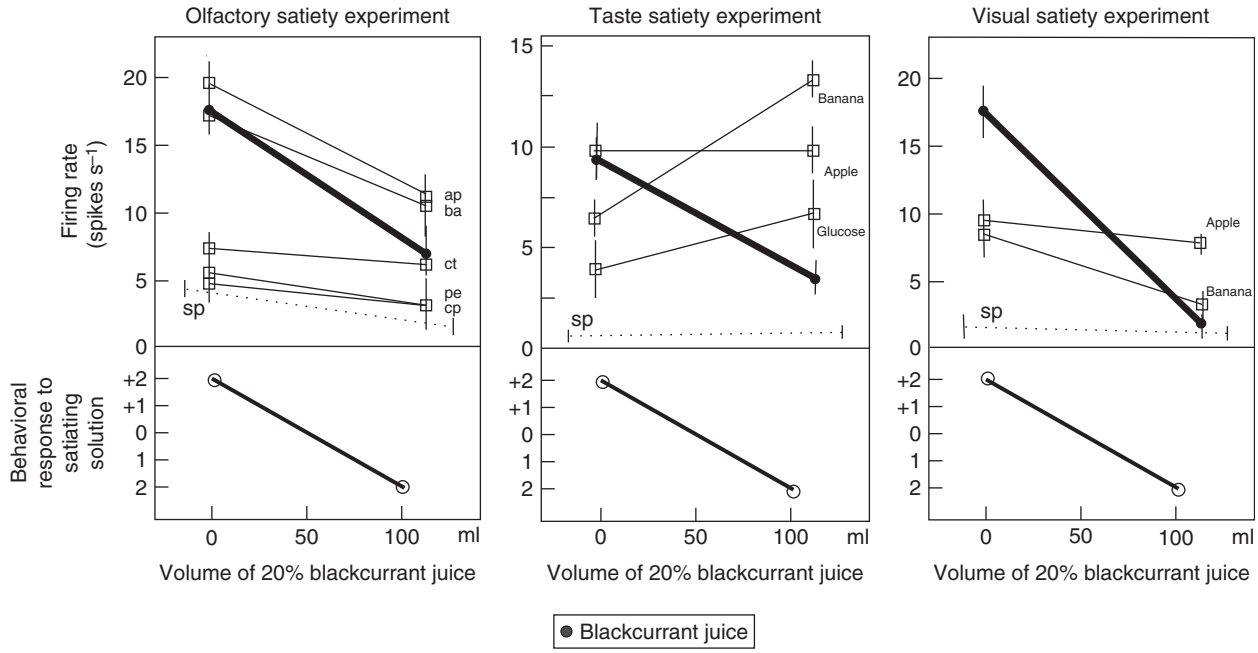

Figure 2 Orbitofrontal cortex neuron with visual, olfactory, and taste responses, showing the responses before and after feeding to satiety with blackcurrant juice. The solid circles show the responses to blackcurrant juice. The olfactory stimuli included apple (ap), banana (ba), citral (ct), phenylethanol (pe), and caprylic acid (cp). The spontaneous firing rate (sp) of the neuron is shown. Adapted from Critchley HD and Rolls ET (1996) Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. *Journal of Neurophysiology* 75: 1673–1686.

value of the sight of food is a special property of the orbitofrontal cortex, which receives visual inputs from the inferior temporal visual cortex.

These findings link at least part of the processing of olfactory and visual information in this brain region to the control of feeding-related behavior. This is further evidence that part of the olfactory representation in this region is related to the hedonic value of the olfactory stimulus, and in particular that at this level of the olfactory system in primates, the pleasure elicited by the food odor is at least part of what is represented.

As a result of the neurophysiological and behavioral observations showing the specificity of satiety in the monkey, experiments were performed to determine whether satiety was specific to foods eaten in humans. It was found that the pleasantness of the taste of food eaten to satiety decreased more than for foods that had not been eaten. One consequence of this is that if one food is eaten to satiety, appetite reduction for other foods is often incomplete, and this will lead to enhanced eating when a variety of foods is offered. Because sensory factors such as similarity of color, shape, flavor, and texture are usually more important than are metabolic equivalences in terms of protein, carbohydrate, and fat content, in influencing how foods interact in this type of satiety, this has been termed 'sensory-specific satiety.' It should be noted that this effect is distinct from alliesthesia, in that alliesthesia is a change in the pleasantness of sensory inputs produced by internal signals (such as glucose in the gut), whereas sensory-specific satiety is a change in the pleasantness of sensory inputs and is accounted for at least partly by the external sensory stimulation received (such as the taste of a particular food), in that it is at least partly specific to the external sensory stimulation received.

To investigate whether the sensory-specific reduction in the responsiveness of the orbitofrontal olfactory neurons might be related to a sensory-specific reduction in the pleasure produced by the odor of a food when it is eaten to satiety, Rolls and Rolls measured humans' responses to the smell of a food which was eaten to satiety. It was found that the pleasantness of the odor of a food (but much less significantly its intensity) was decreased when persons ate the food to satiety. It was also found that the pleasantness of the smell of other foods (i.e., foods not eaten in the meal) showed much less decrease. This finding has clear implications for the control of food intake, for ways to keep foods presented in a meal appetitive, and for effects on odor pleasantness ratings that could occur following meals. In an investigation of the mechanisms of this odor-specific, sensory-specific satiety, Rolls and Rolls

allowed humans to chew a food without swallowing, for approximately as long as the food is normally in the mouth during eating. They demonstrated sensory-specific satiety with this procedure, showing that the sensory-specific satiety does not depend on food reaching the stomach. Thus at least part of the mechanism is likely to be produced by a change in processing in the olfactory pathways. It is not yet known which stage of olfactory processing is the earliest at which this modulation occurs. It is unlikely to be in the receptors, because the change in pleasantness found was much more significant than was the change in the intensity.

The increase of food intake that can occur when a variety of foods is available, as a result of the operation of sensory-specific satiety, may have been advantageous in evolution in ensuring that different foods with important different nutrients were consumed. However, today, humans have a wide variety of foods readily available, and this may be a factor that can lead to overeating and obesity. In a test of this in the rat, it has been found that variety itself can lead to obesity.

Mouth Feel of Fat: Orbitofrontal Cortex, Primary Taste Cortex, and Amygdala

Texture in the mouth is an important indicator of whether fat is present in a food, which is important not only as a high-value energy source, but also as a potential source of essential fatty acids. In the orbitofrontal cortex, Rolls, Critchley, and colleagues have found a population of neurons that responds when fat is in the mouth. An example of such a neuron is shown in Figure 3. The fat-related responses of these neurons are produced at least in part by the texture of the food rather than by chemical receptors sensitive to certain chemicals, in that such neurons typically respond not only to foods, such as cream and milk, containing fat, but also to paraffin oil (which is a pure hydrocarbon) and to silicone oil ($Si(CH_3)_2O)_n$). Moreover, the texture channels through which these fat-sensitive neurons are activated are separate from viscosity-sensitive channels, in that the responses of these neurons cannot be predicted by the viscosity of the oral stimuli, as illustrated in Figure 3. Some of the fat-related neurons do, though, have convergent inputs from the chemical senses, in that in addition to taste inputs, some of these neurons respond to the odor associated with a fat, such as the odor of cream. Feeding to satiety with fat (e.g., cream) decreases the responses of these neurons to zero on the food eaten to satiety, but if the neuron receives a taste input from, for example, glucose taste, that is not decreased by feeding to satiety with cream. Thus there is a representation of the macronutrient fat in this brain area, and the activation produced by fat is reduced by eating fat to satiety.

For some neurons, a combination of taste, fat texture, oral viscosity, and temperature is represented in the macaque primary taste cortex in the rostral insula and adjoining frontal operculum. These oral sensory properties of food, and also the sight and smell of food, are also represented in the primate amygdala. Interestingly, the responses of these amygdala neurons do not correlate well with the preferences of the macaques for the oral stimuli, and feeding to

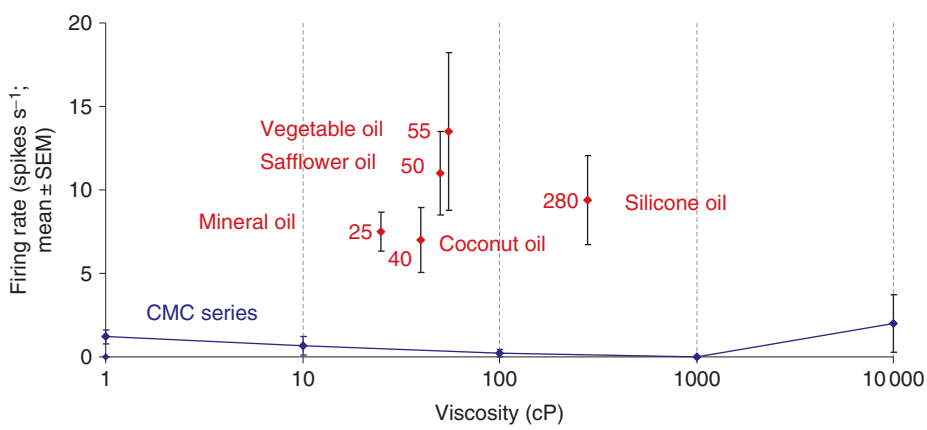

Figure 3 A neuron in the primate orbitofrontal cortex responding to the texture of fat in the mouth independently of viscosity. The cell (bk265) increased its firing rate to a range of fats and oils (the viscosity of which is shown in centipoises, cP). The information that reaches this type of neuron is independent of a viscosity-sensing channel, in that the neuron does not respond to the carboxymethyl cellulose (CMC) viscosity series. The neuron responded to the texture rather than to the chemical structure of the fat in that it also responded to silicone oil ($Si(CH_3)_2O)_n$) and paraffin (mineral) oil (hydrocarbon). Some of these neurons have taste inputs. Adapted from Verhagen JV, Rolls ET, and Kadohisa M (2003) Neurons in the primate orbitofrontal cortex respond to fat texture independently of viscosity. *Journal of Neurophysiology* 90: 1514–1525.

satiety does not produce the large reduction in the responses of primate amygdala neurons to food that is typical of orbitofrontal cortex neurons.

Imaging Studies in Humans

Taste

In humans it has been shown in neuroimaging studies using functional magnetic resonance imaging (fMRI) that taste activates an area of the anterior insula/frontal operculum, which is probably the primary taste cortex, and part of the orbitofrontal cortex, which is probably the secondary taste cortex. O'Doherty et al. showed that the human amygdala was as much activated by the affectively pleasant taste of glucose as by the affectively negative taste of NaCl, and thus provided evidence that the human amygdala is not especially involved in processing aversive as compared to rewarding stimuli.

Another study has shown that umami taste stimuli, of which an exemplar is monosodium glutamate (MSG) and which captures what is described as the taste of protein, activate cortical regions of the human taste system similar to those activated by a prototypical taste stimulus, glucose. A part of the rostral anterior cingulate cortex (ACC) was also activated. When the nucleotide 0.005 M inosine 5′-monophosphate (IMP) was added to MSG (0.05 M), the blood oxygenation-level-dependent (BOLD) signal in an anterior part of the orbitofrontal cortex showed supralinear additivity, and this may reflect the subjective enhancement of umami taste that has been described when IMP is added to MSG. Overall, these results illustrate that the responses of the brain can reflect inputs produced by particular combinations of sensory stimuli with supralinear activations, and that the combination of sensory stimuli may be especially represented in particular brain regions.

Odor

In an investigation of where the pleasantness of olfactory stimuli might be represented in humans, O'Doherty et al. showed that the activation of an area of the orbitofrontal cortex to banana odor was decreased (relative to a control vanilla odor) after bananas were eaten to satiety. Thus activity in a part of the human orbitofrontal cortex olfactory area is related to sensory-specific satiety, and this is one brain region where the pleasantness of odor is represented.

An important issue is whether there are separate regions of the brain discriminable with fMRI that represent pleasant and unpleasant odors. To investigate this, we measured the brain activations produced by three pleasant and three unpleasant odors. The pleasant odors chosen were linalyl acetate (floral, sweet), geranyl acetate (floral), and α-ionone (woody, slightly food related). The unpleasant odors chosen were hexanoic acid, octanol, and isovaleric acid. We found that these unpleasant odors activated dissociable parts of the human brain. Pleasant, but not unpleasant, odors were found to activate a medial region of the orbitofrontal cortex. Further, there was a correlation between the subjective pleasantness ratings of the six odors given during the investigation with activation of a medial region of the rostral orbitofrontal cortex. In contrast, a correlation between the subjective unpleasantness ratings of the six odors was found in regions of the left and more lateral orbitofrontal cortex. Activation was also found in the anterior cingulate cortex, with a middle part of the anterior cingulate activated by both pleasant and unpleasant odors, and a more anterior part of the anterior cingulate cortex showing a correlation with the subjective pleasantness ratings of the odors. These results provide evidence that there is a hedonic map of the sense of smell in brain regions such as the orbitofrontal cortex and cingulate cortex.

Olfactory–Taste Convergence to Represent Flavor, and the Influence of Satiety

To assess how satiety influences the brain activations to a whole food which produces taste, olfactory, and texture stimulation, we measured brain activation by whole foods before and after the food is eaten to satiety. The aim was to determine using a food that has olfactory, taste, and texture components the extent of the region that shows decreases when the food becomes less pleasant, in order to identify the different brain areas where the pleasantness of the odor, taste, and texture of food are represented. The foods eaten to satiety were either chocolate milk or tomato juice. A decrease in activation by the food eaten to satiety relative to the other food was found in the orbitofrontal cortex, but not in the primary taste cortex (see Figure 4). This study provided evidence that the pleasantness of the flavor of food, and sensory-specific satiety, are represented in the orbitofrontal cortex.

Oral Viscosity and Fat Texture

The viscosity of food in the mouth is represented in the human primary taste cortex (in the anterior insula), and also in a midinsular area that is not taste cortex, but which represents oral somatosensory stimuli. In these regions, the fMRI BOLD activations are proportional to the log of the viscosity of carboxymethyl cellulose in the mouth. Oral viscosity is also represented in the human orbitofrontal and perigenual cingulate cortices, and it is notable that the perigenual

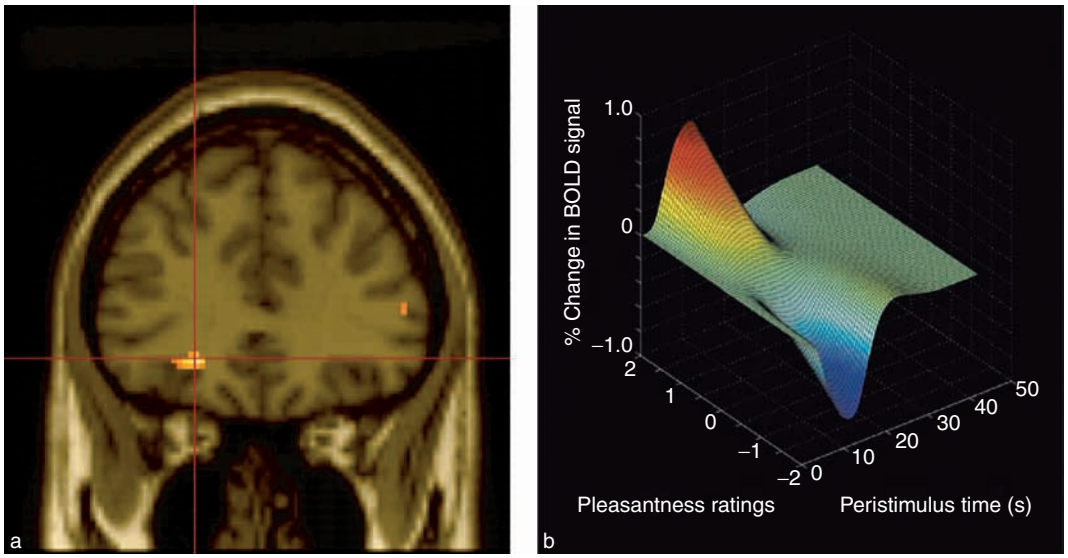

Figure 4 Areas of the human orbitofrontal cortex with activations correlating with pleasantness ratings for food in the mouth. (a) Coronal section through the region of the orbitofrontal cortex from the random-effects group analysis, showing the peak in the left orbitofrontal cortex (Talairach coordinates *X*, *Y*, *Z* = −22, 34, −8; z-score = 4.06), in which the blood oxygenation-level-dependent signal in the voxels, shown in yellow, was significantly correlated with the individual's subjective pleasantness ratings of the foods throughout an experiment in which the individuals were hungry and found the food pleasant, and were then fed to satiety with the food, after which the pleasantness of the food decreased to neutral or slightly unpleasant. The design was a sensory-specific satiety design, and the pleasantness of the food not eaten in the meal, and the blood oxygenation-level-dependent activation in the orbitofrontal cortex, were not altered by eating the other food to satiety. The two foods were tomato juice and chocolate milk. (b) Plot of the magnitude of the fitted hemodynamic response from a representative single individual against the subjective pleasantness ratings (on a scale from −2 to +2), and peristimulus time in seconds. Adapted from Kringelbach ML, O'Doherty J, Rolls ET, et al. (2003) Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. *Cerebral Cortex* 13: 1064–1071.

cingulate cortex, an area in which many pleasant stimuli are represented, is strongly activated by the texture of fat in the mouth and also by oral sucrose.

The Sight of Food

Visual stimuli associated with the taste of glucose activate the orbitofrontal cortex and some connected areas, consistent with the primate neurophysiology. Consistent with these findings, Pelchat and colleagues found that after consuming a monotonous diet, study participants who were instructed to imagine foods that they craved showed more activation in some brain areas, including part of the insula, than did participants who had consumed a normal diet.

Satiety Signals

Satiety signals can be thought of as occurring in a sequence. In addition to sensory-specific satiety, which involves reduced activity in cortical areas that represent the pleasantness of food, further temporally overlapping signals include gastric distension, duodenal chemosensory signals, glucose utilization, and hormonal (including leptin) signals. Many of these signals are likely to have influences, in some cases via the lateral hypothalamus, on the orbitofrontal cortex, and by modulating sensory activity there, to set the current pleasantness and reward value of food. Investigation of hunger and satiety signals has been a major and difficult issue for many years, but whatever the details of these signals, they must influence processing in brain areas such as the orbitofrontal cortex. The outputs of the orbitofrontal cortex then reach brain regions such as the striatum, cingulate cortex, and dorsolateral prefrontal cortex, where behavioral responses to food may be elicited, because these structures produce behavior which makes the orbitofrontal cortex reward neurons fire, as they represent a goal for behavior. At the same time, outputs from the orbitofrontal cortex and amygdala, in part via the hypothalamus, may provide for appropriate autonomic and endocrine responses to food to be produced, including the release of hormones such as insulin.

See also: Central Gustatory System and Ingestive Behavior; Conditioned Taste Aversion; Flavor Physiology; Food and Water Intake: Regulation; Gastrointestinal Signals: Stimulation; Gastrointestinal Signals: Satiety;

Neuropeptides: Food Intake; Olfactory Cortex Physiology; Retronasal Olfaction; Reward Neurophysiology and Orbitofrontal Cortex; Sensory Aging: Chemical Senses; Taste: Vertebrate Taste Bud Physiology; Taste: Vertebrate Central Pathways.

Further Reading

Cabanac M (1971) Physiological role of pleasure. *Science* 173: 1103–1107.

Critchley HD and Rolls ET (1996) Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. *Journal of Neurophysiology* 75: 1673–1686.

Kadohisa M, Rolls ET, and Verhagen JV (2005) Neuronal representations of stimuli in the mouth: The primate insular taste cortex, orbitofrontal cortex, and amygdala. *Chemical Senses* 30: 401–419.

Kringelbach ML, O'Doherty J, Rolls ET, et al. (2003) Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. *Cerebral Cortex* 13: 1064–1071.

Norgren R (1984) Central neural mechanisms of taste. In: Darien-Smith I (ed.) *Handbook of Physiology –The Nervous System III. Sensory Processes 1*, pp. 1087–1128. Washington, DC: American Physiological Society.

O'Doherty J, Rolls ET, Francis S, et al. (2000) Sensory-specific satiety related olfactory activation of the human orbitofrontal cortex. *NeuroReport* 11: 893–897.

Pelchat ML, Johnson A, Chan R, et al. (2004) Images of desire: Food-craving activation during fMRI. *NeuroImage* 23: 1486–1493.

Rolls ET (1981) Central nervous mechanisms related to feeding and appetite. *British Medical Bulletin* 37: 131–134.

Rolls ET (2005) *Emotion Explained*. Oxford: Oxford University Press.

Rolls ET (2006) Brain mechanisms underlying flavour and appetite. *Philosophical Transactions of the Royal Society London Series B* 361: 1123–1136.

Rolls ET (2007) Sensory processing in the brain related to the control of food intake. *Proceedings of the Nutrition Society* 66: 96–112.

Rolls ET and Rolls JH (1997) Olfactory sensory-specific satiety in humans. *Physiology and Behavior* 61: 461–473.

Rolls BJ, Rolls ET, Rowe EA, et al. (1981) Sensory specific satiety in man. *Physiology and Behavior* 27: 137–142.

Rolls ET, Critchley HD, Browning AS, et al. (1999) Responses to the sensory properties of fat of neurons in the primate orbitofrontal cortex. *Journal of Neuroscience* 19: 1532–1540.

Rolls ET, Kringelbach ML, and de Araujo IET (2003) Different representations of pleasant and unpleasant odors in the human brain. *European Journal of Neuroscience* 18: 695–703.

Verhagen JV, Rolls ET, and Kadohisa M (2003) Neurons in the primate orbitofrontal cortex respond to fat texture independently of viscosity. *Journal of Neurophysiology* 90: 1514–1525.

Verhagen JV, Kadohisa M, and Rolls ET (2004) The primate insular/opercular taste cortex: Neuronal representations of the viscosity, fat texture, grittiness, temperature and taste of foods. *Journal of Neurophysiology* 92: 1685–1699.

Corticomotoneuronal System

R N Lemon, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Definition

The term 'corticomotoneuronal (CM)' is used exclusively to denote a direct monosynaptic linkage between the axon of a corticospinal neuron and spinal motor neurons. The term was first introduced by Bernhard and Bohm in their classic papers in the early 1950s. They demonstrated that electrical stimulation of the motor cortex evoked responses in limb nerves of the monkey which had such short latencies that they could only have resulted from a direct, monosynaptic response of motor neurons to the descending volleys in the corticospinal tract (CST) excited by cortical stimulation. This suggestion was subsequently confirmed physiologically by Phillips, who made intracellular recording in alpha motor neurons of monosynaptic CM excitatory postsynaptic potentials (EPSPs) in response to stimulation of the motor cortex in the baboon. Anatomical confirmation was provided by Kuypers and others, who demonstrated labeled boutons from the CST terminating on dendrites of spinal motor neurons. The term CM is quite specific to the monosynaptic connection and should not be confused with more general terminology that refers to the influence of the motor cortex over motor neurons and muscles through less direct (oligosynaptic) or otherwise undefined pathways. The term CM is not usually applied to corticobulbar and other cortical projections which innervate cranial motor nuclei.

Anatomy of the CM System

The CM system is made up of a subset of CST fibers that terminate upon spinal motor neurons. There are at least two aspects to the anatomy of the CM system. One relates to the cortical origin of the CM system and its course between cortex and spinal cord, and the second to the spinal termination of CM fibers upon spinal alpha motor neurons. Because the CM system appears to be uniquely present in primates, most of the experimental evidence referred to below has been obtained in humans or in nonhuman primates (mainly Old World macaque monkeys).

Pattern of Termination

Little is known about the number, size, or density of CM terminals on individual motor neurons, but a single CST axon probably makes only one or only a few contacts with a given target motor neuron, and those contacts are located on medium to distal dendrites. There is as yet no direct evidence for CM effects on gamma motor neurons.

Distribution of CM Projections to Different Motor Nuclei

Early studies by Kuypers and others using silver degeneration provided widespread evidence in primates, including humans, that CST fibers terminated widely among the motor nuclei that supply limb muscles, and particularly those supplying muscles of the distal extremity (hand, foot, and, in some species, tail). This was confirmed by later studies using anterograde neuroanatomical tracers. In the macaque, the density of CS terminals among the lateral motor nuclei shows an increasing rostral to caudal gradient (from low cervical (C7) to high thoracic (T1) segments); this gradient corresponds to the increasing density of the motor neurons innervating the intrinsic hand muscles.

Origins of CM Projections

An important issue is to discover which components of the CST give rise to CM projections. The CST itself is derived from widespread regions of the frontal and parietal lobes. The major cortical area that terminates directly among motor neurons is the classical primary motor cortex (termed M1, Brodmann area 4). There is accumulating anatomical evidence that most of the CST that gives rise to CM projections arises from M1 and rather less from the secondary motor areas (such as the supplementary motor area (SMA) and premotor areas). Recent studies using retrograde transneuronal labeling techniques have suggested that cells with CM projections are mainly located in the rostral bank of the central sulcus and that CM neurons are not found on the convexity of the precentral gyrus. These studies have also indicated that CM projections arose from CST neurons with both large and small cell bodies.

The postcentral (somatosensory) cortex also gives rise to large numbers of corticospinal projections, and those originating from cortical area 3a have been shown to give rise to CM connections. It has been speculated that these connections might involve gamma motor

neurons, which provide motor innervation to muscle spindles, whose feedback is thought to converge heavily on cortical area 3a. There is no evidence to date for a specific molecular marker distinguishing CM from other corticospinal neurons.

Physiology of the CM System

Electrical Stimulation Methods

The CM system has classically been identified by monosynaptic responses in target motor neurons or muscles evoked by direct electrical stimulation of the CST at the level of the medullary pyramid (the pyramidal tract, PT) or of the motor cortex. Such stimulation gives rise to one or more descending volleys of CS activity. These methods have been complemented in humans by the use of transcranial magnetic and electrical stimulation.

Identification of Monosynaptic CM Effects in Motor Neurons

The monosynaptic nature of CM responses to the earliest corticospinal volley has been inferred from the delay (~1.0 ms) between arrival of the volley and the onset of the response, which delay is shorter than the minimum value normally associated with disynaptic transmission (1.3 ms or greater). All CM synapses are excitatory, and responses take the form of a compound EPSP. Any inhibitory effects are thought to be mediated by spinal inhibitory interneurons activated by CST inputs. The monosynaptic nature of the CM connection is also shown by its ability to follow high-frequency stimulation of the PT (up to 500 Hz or higher).

If a CM-EPSP is large enough, it will discharge the motor neuron. In awake or lightly sedated animals, stimulation of the PT can be shown to evoke short-latency responses in motor units. These responses are probably monosynaptic because of their short and invariant latency.

Interpretation of effects resulting from electrical stimulation of the motor cortex is more difficult because this stimulation usually gives rise to repetitive high-frequency (~600 Hz) activation in the CST. The first wave of excitation is the D-wave, due to direct activation of CST neurons, and this is followed by I-waves, thought to be due to indirect, transsynaptic activation of CST neurons. It is possible to demonstrate that each successive wave of CS activity gives rise to a CM-EPSP.

Physiological Properties of CM Synapses

CM synapses are excitatory and involve glutamate as the neurotransmitter. The size of the unitary CM-EPSP synapse (i.e., the amplitude of the EPSP evoked in a target motor neuron by activation of a single CM input) is not known, but it is probably quite small (50–100 μV). High-frequency stimulation causes temporal facilitation of CM-EPSPs and may explain in part the powerful motor response to stimulation of the cortex, such as transcranial magnetic stimulation (TMS; see the section titled 'Transcranial stimulation in humans'). CM synapses are not thought to be under control of presynaptic inhibition. Excitation of motor neurons by CM inputs is severely depressed or abolished by general anesthetics.

Distribution of CM Effects in Different Motor Nuclei

The size of compound CM-EPSPs (i.e., the amplitude of the EPSP evoked in a target motor neuron by activation of all its CM inputs) varies considerably, both within a motor nucleus innervating a single muscle and across motor nuclei. In general, the largest effects are seen in the motor nuclei supplying the most distal (hand, foot) muscles and in extensor muscles of the upper limb (such as the common extensor of the digits) and dorsiflexors of the foot (such as the tibialis anterior). In monkeys, CM effects are weaker or absent in motor nuclei supplying more proximal (shoulder, hip) muscles. In humans tested with trancranial stimulation of the cortex, these same general rules apply, and there is evidence that the compound EPSP is probably larger in humans than in most nonhuman primates that have been investigated. In humans, putative CM effects have also been reported in many other muscle groups, including those of respiratory muscles, girdle, trunk, and sphincter muscles.

Origins of CM Effects

Because CST fibers arising from different regions of the cortex and terminating at different spinal levels are intermixed at the level of the pyramidal tract, PT stimulation activates the CST in an indiscriminate manner. To investigate which cortical regions give rise to CM effects, local, single-pulse electrical stimulation of different cortical areas can be used. Such studies indicate that M1 gives rise to large CM effects in hand motor neurons, consistent with the anatomical findings reviewed already. Much weaker effects are obtained from secondary motor areas such as the SMA, and CM effects are generally absent after stimulation of the premotor cortex.

Transcranial Stimulation in Humans

TMS and transcranial electrical stimulation (TES) have allowed detailed investigation of CM effects in awake humans with a wide variety of neurological diseases.

Both forms of stimulation result in high-frequency, repetitive activation of the CST, which complicates interpretation of the effects, normally recorded as a motor evoked potential (MEP) in the electromyogram (EMG) or from single motor units. The earliest muscle responses are probably due to the arrival of waves of activity in the CST induced by TMS (dominated by I-waves) or TES (dominated by D-waves), and these responses have properties which confirm their CM origin. However, it is probable that MEPs in whole EMG recordings undoubtedly include motor neuron responses to CM inputs as well as to indirect, oligosynaptic inputs. Motor responses, probably including CM effects, have also been evoked by transmastoid electrical stimulation, which is thought to activate the CST at brain stem level.

Identification of CM Effects by Spike-Triggered Averaging

Fetz and Cheney first showed that CM cells can be identified in the awake animal by use of correlational methods to demonstrate that a motor cortex neuron exerts a facilitatory effect on a motor neuron pool that is consistent with CM action. This method involves spike-triggered averaging (STA) of EMG activity. In this approach, EMG activity recorded from limb muscles is full wave rectified and averaged with respect to the spontaneous, natural discharge of a single neuron, recorded from the contralateral motor cortex. If the neuron makes direct CM connections with the motor neurons of the recorded muscle, there should be a postspike facilitation (PSF) of EMG activity. This PSF reflects the CM-EPSPs generated by the cortical neuron causing motor neurons to fire, resulting in the discharge of motor unit(s) in the EMG that is a time locked to the spikes of the cortical neuron. Because of the small size of unitary EPSPs from CM cells, the probability of such CM cell-triggered discharge of motor neurons is low, and therefore large numbers of spikes (typically at least 2000–4000) must be averaged before any PSF can be detected in the average. The amplitude, form, and latency of the PSF are used to judge whether the PSF is indeed the result of CM action. It is now generally accepted that the STA method reveals a spectrum of postspike effects in EMG, which may include some effects due to synchronization among populations of cortical neurons.

'Pure' PSF effects in STA are characterized by a sharp onset and offset at latencies consistent with CM action. This means that the onset latency should be equal to the central conduction time for action potentials in a particular CM neuron to reach the motor neuron pool from the cortex, plus one synaptic delay plus the peripheral conduction time from motor neuron pool to muscle.

In contrast, 'synchrony' PSF usually consists of slower, broader features in the STA with onset latencies that are much shorter than the predicted total conduction delay and which in some cases precede the trigger spike. This implies that the spikes used to generate the STA emanate from a neuron that is not itself a genuine CM cell but whose discharges are synchronized with one or more CM cells. Although such effects lack any direct evidence of CM action, some investigators take them as indicative of a more general involvement of the recorded neuron in influencing motor output. Both pure and synchrony PSF can occur in the same STA.

STA has demonstrated that PSF is most commonly derived from large CST neurons with fast conducting axons. However, in agreement with anatomical studies, PSF is also produced by small, slower conducting CST neurons. Recording in awake animals is generally biased toward the larger neurons, and it may be that CM contribution from the much more numerous smaller, but undersampled, CST neurons has been seriously underestimated.

Stimulus-triggered averaging of EMG has also been used to look at the organization of corticospinal outputs from cortical motor areas. This method is useful for detecting weaker outputs and for mapping the whole motor output. It undoubtedly excites multiple corticospinal neurons, including CM cells. Even single stimuli can excite quite complex effects within the cortex, and identification of motor effects as unequivocally CM is more problematic, but in general there is agreement between stimulus-evoked effects from a cortical site and spike-triggered effects from single neurons recorded at the same site.

Finally, it should be emphasized that most CM cells, as well as producing PSF in one muscle or group of muscles, have also been shown to produce postspike suppression of EMG activity in other muscles. This is most clearly seen in CM cells that facilitate a group of agonist muscles (e.g., wrist flexors) but suppression of activity in antagonists (e.g., wrist extensors). Suppression is considered to be mediated via spinal inhibitory interneurons.

Distribution of CM Effects in Different Motor Neuron Pools

PSF has been observed to occur in many upper limb and particularly hand muscles. It is generally larger in more distal muscles, especially intrinsic hand muscles. It has been demonstrated to occur during a wide range of different types of upper limb and hand movements, including different types of hand

grasp, precision and power grip, fractionated finger movements, wrist extension and flexion movements, elbow movements, and whole arm movements such as are used in reaching and pointing. For a given CM cell and target muscle combination, PSF is generally found to be present across a range of tested tasks and movements, although there may be significant changes in its form and amplitude.

The 'muscle field' of a CM cell defines the group of muscles in which it exerts PSF. When a reasonable sample of muscles is tested (e.g., at least five active muscles, recorded simultaneously with the CM cell), it is relatively rare to find that the CM cell produces PSF in a single target muscle. To date all studies have shown that PSF is found only in a subset of the muscles that are coactivated with the cortical neuron, and this has been interpreted as a key characteristic of the CM system: it is a 'one to not-so-many' divergent control system, with the capacity to facilitate activity in a relatively small, focused group of muscles. This capacity is further underlined by the finding that in some cases the muscle field represents a group of muscles that act as functional synergists: that is, they act together to produce a particular type of movement. For example, CM cells can be found which facilitate both long extrinsic digit muscles and short intrinsic hand muscles, whose combined activity is needed to produce digit movements involving action over a multiarticulate system of carpal, metacarpal, and phalangeal joints.

Development of the CM System

It is generally recognized that the CM system, like the CST more generally, has a protracted timetable of postnatal development. Studies in both humans and nonhuman primates have indicated that the CST extends its fiber projection along the full length of the spinal cord long before birth, but both the myelination of the tract and its growth, arborization, and synaptic termination within the spinal gray matter occur postnatally. In the macaque monkey, there is direct anatomical and electrophysiological evidence for a rapid development of the CM projection to the cervical enlargement during the first 6 postnatal months. This time course matches exactly the monkey's ability to grip and manipulate fine objects.

Kuypers first showed that there are striking differences in the degree of development of the CM system across different species. Although in some species, such as the rat and the raccoon, there are sparse CST projections into the vicinity of the cervical motor neuron pools, electron microscopy and electrophysiological studies have since demonstrated that there are no functional CM connections in these species, and it would now appear that the CM system is a uniquely primate feature. Both physiology and anatomy demonstrate that even within the primates, this system is developed to very different extents. For example, the CM system is absent in tree shrews, lemurs, and the marmoset. It is well developed in some New World monkeys such as the capuchin monkey and in Old World monkeys, including the macaque. CM connections are generally most numerous in the great apes and in humans.

There has been considerable interest in the possible significance of these differences. There is a clear relationship across species between the extent of CST projections among the motor neuron pools of the spinal gray matter and the 'index of dexterity,' a linear score of hand junction used to assess hand/paw use in a wide variety of mammalian species. This index reflects functional attributes of the skilled forelimb used, and in particular the use of digits for prehensile purposes and for manipulation. It also assesses the extent to which independent use of the digits is developed in a given species. All the primates known to have high levels of dexterity (index 6 or above) possess a well-developed CM system.

Lesions Affecting CM Function

Lesions of the CST at a cortical, capsular, brain stem, or spinal level will of course also involve the CM system. However, there is currently no means of investigating a lesion selective for CM connections alone. Classical work has shown that complete, bilateral pyramidotomy in macaques (cutting both pyramidal tracts at the brain stem medullary level) has relatively small effects on many motor functions, including walking, running, and climbing, although causing a permanent deficit in fine motor skill and independent finger movements. A wealth of studies have demonstrated that CM neurons are particularly active during such movements. It has therefore been speculated that deficits in fine motor control are largely the result of interrupting CM fibers terminating on motor neurons supplying hand and digit muscles. However, the observed deficits indicate a breakdown in fine sensorimotor control rather than a pure motor deficit. This implies not only a deterioration in motor performance, but also in the ability to correctly interrogate sensory feedback from the hand, a function likely to reflect CST projections to dorsal rather than ventral spinal gray matter.

With subtotal lesions of the CST, a somewhat different picture emerges. Such lesions include unilateral lesions interrupting all or most of the fibers in the lateral CST. The effects are initially similar to a complete lesion, a striking deterioration in the speed,

force, and accuracy of more distal hand and/or foot movements. Interestingly, these deficits often show a substantial degree of recovery, the extent of which seems to be related to the amount of CST fiber spared. These results both confirm the importance of the CM system and the CST for distal control and underline the capacity for plastic mechanisms to compensate for subtotal damage.

Pathological Diseases Affecting CM Function

Pathological damage and disease that affect the CST include stroke, cerebral palsy, and spinal cord injury. These disorders disrupt CST function in general and will, of course, also affect the CM system. There is some evidence that large fibers in the CST, including those giving rise to CM projection, may be especially vulnerable to damage in stroke and spinal cord injury. In the early phase of some forms of ALS, there is evidence of lowered cortical threshold to TMS for evoking MEPs in hand muscles, and this may be due to an excitotoxic glutaminergic mechanism operating within the motor cortex. Coupled with this have been observations that the primary peak responses of single motor units to TMS may be enhanced during onset of ALS, suggesting that the CM system may be a particular target of the disease process; later in the disease, there is evidence for a selective loss of the fastest corticospinal neurons, many of which make CM connections. Finally, monitoring of descending CST volleys and/or the response of upper and lower limb motor neurons to them has been shown to be a very effective form of interoperative monitoring during neurosurgical procedures in the brain and spinal cord, preventing potential damage to the CST.

Functional Significance of the CM System

The first point is that the CM system is only a component of the CST. The CST is known to influence spinal circuitry involved in a wide variety of sensory, motor, and autonomic functions. In addition to the CST, there are a number of descending pathways operating in parallel to it. Thus, the CM system does not operate in isolation, and it is important to consider how the CM system works alongside these parallel systems. Cortical control of limb movements is of course highly developed in many species without CM connections, and these species have provided good models for control via indirect or oligosynaptic pathways.

The electrophysiology, neuroanatomy, development, and comparative biology of the CM system have all demonstrated its role in the control of distal movements, movements which are under high levels of voluntary control. Although most research has focused on the role of CM control of digit and hand movements, there is plentiful evidence that many upper- and lower-limb muscles receive CM projections. A general function for the CM system is in providing the capacity for fractionation of movements and the control of small groups of muscles in a highly selective manner. Thus, the CM system may contribute to the 'analytical' basis of motor control. The breaking down of relatively hardwired motor patterns can be regarded as an important feature of voluntary motor actions and one that is probably essential for the acquisition of new motor skills. CM action could be superimposed on background drive to motor neurons, provided by other descending motor pathways, so as to sculpture selective patterns of muscle activity.

The available data suggest that the CM system is a recently evolved feature and that it subserves evolutionarily newer aspects of motor behavior, including voluntary control of relatively independent finger movements. Although such movements are not exclusive to species with CM connections, they are far and away better developed in primates than in nonprimates.

Skilled hand movements are thought to be driven in a feed-forward manner by an internal model, a model that is richly supplied with somatosensory and other sensory inputs so that it is constantly updated in terms of the outcome of each movement. The internal model system 'learns' from these updates to generate predictable motor outcomes. This might explain why CM commands apparently bypass the spinal mechanism: Their control signals are faithfully transmitted to the motor neurons because the outcome is already predictable on the basis of past programs stored in the internal model that are regularly updated after each new action. Many of the species with well-developed CM systems possess the capacity to acquire motor skill using fine hand and finger movements. Thus, it may be that the CM system provides an efferent pathway that is accessed by the extensive motor network of the primate brain for the development of adaptive motor programs. These programs exhibit massive flexibility in the combinatorial patterns of selective muscle recruitment, and the CM system is well placed to provide this.

See also: Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Corticospinal Development; Finger Movements: Control; Map Plasticity and Recovery from Stroke; Spinal Motor Neurons: Properties; Synaptic Mechanisms of Learning.

Further Reading

Armand J, Olivier E, Edgley SA, and Lemon RN (1997) The postnatal development of corticospinal projections from motor cortex to the cervical enlargement in the macaque monkey. *Journal of Neuroscience* 17: 251–266.

Baker SN and Lemon RN (1998) Computer simulation of post-spike facilitation in spike-triggered averages of rectified EMG. *Journal of Neurophysiology* 80: 1391–1406.

Cheney PD, Belhaj-Saif A, and Boudrias M-H (2004) Principles of corticospinal system organisation and function. In: Eisen A (ed.) *Clinical Neurophysiology of Motor Neuron Diseases*, pp. 59–96. Amsterdam: Elsevier.

Day BL, Dressler D, Maertens de Noordhout A, et al. (1989) Electric and magnetic stimulation of human motor cortex: Surface EMG and single motor unit responses. *Journal of Neurophysiology* 412: 449–473.

Dum RP and Strick PL (2002) Motor areas in the frontal lobe of the primate. *Physiology & Behavior* 77: 677–682.

Eisen A and Krieger C (2004) Classic Charcot amyotrophic lateral sclerosis. In: Eisen A (ed.) *Clinical Neurophysiology of Motor Neuron Diseases*, pp. 469–486. Amsterdam: Elsevier.

Kuypers HGJM (1981) Anatomy of the descending pathways. In: Brookhart JM and Mountcastle VB (eds.) *Handbook of Physiology: The Nervous System II*, pp. 597–666. Bethesda, MD: American Physiological Society.

Lemon RN and Griffiths J (2005) Comparing the function of the corticospinal system in different species: Organizational differences for motor specialization? *Muscle & Nerve* 32: 261–279.

Lemon RN, Kirkwood PA, Maier MA, Nakajima K, and Nathan P (2004) Direct and indirect pathways for corticospinal control of upper limb motoneurons in the primate. *Progress in Brain Research* 143: 263–279.

Maier MA, Armand J, Kirkwood PA, Yang H-W, Davis JN, and Lemon RN (2002) Differences in the corticospinal projection from primary motor cortex and supplementary motor area to macaque upper limb motoneurons: An anatomical and electrophysiological study. *Cerebral Cortex* 12: 281–296.

Porter R and Lemon RN (1993) *Corticospinal Function and Voluntary Movement*. Oxford: Clarendon.

Rathelot JA and Strick PL (2006) Muscle representation in the macaque motor cortex: An anatomical perspective. *Proceedings of the National Academy of Sciences of the United States of America* 103: 8257–8262.

Schieber MH and Rivlis G (2005) A spectrum from pure post-spike effects to synchrony effects in spike-triggered averages of electromyographic activity during skilled finger movements. *Journal of Neurophysiology* 94: 3325–3341.

Corticospinal Development

J H Martin, K M Friel, I Salimi, and S Chakrabarty, Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The corticospinal (CS) system is the principal motor system for controlling skilled movements. In humans, this system is so important that damage to it during development or later in life invariably produces significant motor impairment. The primary motor cortex and adjoining premotor and somatic sensory cortices are the principal origins of the CS system. Neurons in these cortical areas project to the spinal cord through the CS tract. This is the last motor pathway to develop, growing into the spinal cord during the late prenatal and early postnatal periods. Even with this early start, development of the CS system is protracted, in humans especially, in whom it takes 10–15 years to gain some of its mature characteristics. It is remarkable that most CS system development occurs while the individual is acquiring motor skills during early postnatal life.

Like other neural systems, development of the CS system depends on a complex interplay between factors intrinsic to the developing central nervous system (CNS) – including the CS neurons themselves, the regions through which CS axons grow to reach their spinal gray matter targets, and the spinal gray matter – and on neural activity and behavioral experience. This article first considers milestones of CS system development in several species, including humans, and then examines, in a bottom-up sequence, the critical steps in development of the CS system's role in skilled movement control.

Phases and Time Course of CS System Development

While there are a bewildering number of steps leading to development of this motor system, four major steps can be identified, and each is governed by different mechanisms. First, precursor cells differentiate into CS neurons; this occurs prenatally. Second, CS axons grow through the brain and into the spinal cord to form the CS tract. Once CS axons are in the cord, there is outgrowth from the tract into the gray matter. This is the pathfinding stage and leads to the initial selection of postsynaptic targets. Pathfinding occurs primarily prenatally in humans, monkeys, and cats but primarily postnatally in rodents. Both the commitment to become a CS neuron and axon outgrowth and guidance to postsynaptic target neurons are determined by cell-specific and regional-intrinsic factors. Through genetic approaches, many key molecules have been identified recently. The third step is to establish connectional specificity at the circuit level, which involves refinement of the gray matter terminations. This typically includes both elimination of axon branches and growth of new branches to nearby targets. For spinal circuits, physiological studies in humans indicate that this happens during the first 2 years, while anatomical studies in animals show that refinement occurs within the first postnatal weeks and months, depending on the species. The rapid phase of myelination of CS tract axons occurs during this period. Fourth, the cortical motor representation forms, and the CS system's role in limb control is expressed. During this period, many other neural systems are developing and contributing to movement control. An individual's motor experience and other activity-dependent processes during the third and fourth stages have important long-term developmental functions.

CS Neuron Differentiation

CS neurons are layer-5 pyramidal neurons in a variety of cortical motor and somatic sensory areas. Early in development, CS neurons are distributed from the frontal to the occipital poles but later are restricted primarily to the posterior frontal and anterior parietal cortices. The reduction in the number of CS neurons is principally due to the elimination of spinal axon branches, not widespread cell death. The homeodomain transcription factor Otx1 is required for axon branch elimination. Mice in which the gene for Otx1 has been deleted fail to develop the limited distribution of CS neurons restricted to the somatic sensory and motor cortices.

Significant progress has been made recently in determining factors important for differentiation of CS neurons from precursor cells. It has been proposed that the combinatorial expression of many genes delineates the population of CS neurons. Experiments show that some of these genes are needed for development of the CS tract. For example, in mice without the forebrain embryonic zinc fingerlike transcription factor (Fezl), cortical neurons fail to grow an axon into the caudal brain stem and spinal cord (Figure 1). Moreover, the loss of Fezl results in the loss of other molecular markers of layer 5 and 6 neurons, including another transcription factor required for CS tract formation, Ctip2. This suggests that Fezl acts upstream to regulate neuronal differentiation.

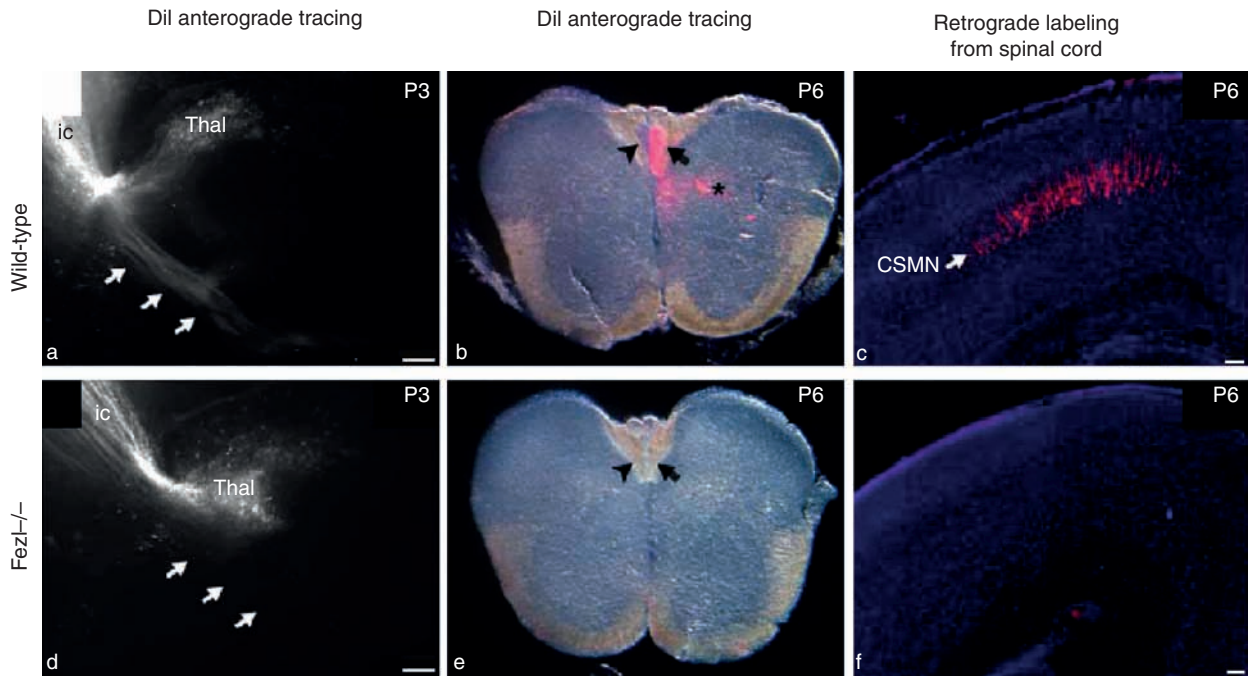

Figure 1 Forebrain embryonic zinc fingerlike transcription factor (Fezl) is important for specification of CS neurons. Images along the top row are from a wild-type mouse and along the bottom row, from a Fezl −/− mouse. (a, d) Anterograde labeling (bright white) of cortical efferent axons. Note that the wild-type shows projections to the brain stem and spinal cord (arrows), but the knockout mouse does not; both show projections to the thalamus (thal). (b, e) Anterograde labeling (red) of CS axons in the spinal cord of the wild-type but not the knockout mouse. (c, f) Retrograde labeling (red) of CS neurons in the wild-type mouse only. Scale bars = 100 μm. DiI, 1,1′-dioctadecyl-3,3,3′,3′-tetramethylindocarbocyamine perchlorate; Ic, internal capsule; CSMN, corticospinal motor neuron; P3, postnatal day 3; P6, postnatal day 6. From Molyneaux BJ, Arlotta P, Hirata T, Hibi M, and Macklis JD (2005) Fezl is required for the birth and specification of corticospinal motor neurons. *Neuron* 47: 817–831.

Development of the Projection to the Brain Stem and Spinal Cord

Pathfinding from Cortex and Subcortical White Matter Tracts

Traversing the CNS to reach its targets is a complex problem for any projection neuron, but particularly for a CS neuron because its axon has the longest distance to travel. Pathfinding is organized by complex tissue molecular cues that are detected by the primary growth cone of the axon. Studies suggest that cortical pyramidal cells simplify the task of reaching their final targets, for example, spinal cord neurons, by achieving a sequence of intermediate targets, such as first projecting into the internal capsule. Many chemorepellant and chemoattractant molecules have been identified as necessary for keeping CS neuron axons 'on tract,' else the axons stray into aberrant locations.

Decussation and Midline Crossing

Functionally, the CS system exerts its effects on the contralateral spinal cord and contralateral muscle through a descending projection that is primarily crossed. Most axons decussate in the medullary pyramid. In the rodent, the axons course from the ventral to dorsal surfaces as they cross from one side to the other, forming an X shape (Figures 2(a) and 2(b)). Three sets of molecules and receptor families have been identified that play key roles in the crossing of CS axons: (1) the L1 cell adhesion molecule (L1-CAM), which is a member of the immunoglobulin superfamily of cell adhesion molecules; (2) the Eph A4 receptor and Ephrin B3, which is a protein that repels axon growth and is the membrane bound ligand for EphA4; and (3) members of the roundabout (Robo) family of transmembrane proteins and Slit, another protein with axon growth-repulsive properties.

L1-CAM is expressed along the CS tract; CD24, which is a ligand for L1-CAM, is localized at the pyramidal decussation. L1 knockout mice display profound errors in CS axon decussation in the medulla and guidance into the appropriate spinal white matter regions (Figure 2(b)). Human L1 mutations produce hypoplasia of the CS tract and spasticity, as well as a variety of other brain structural and functional impairments. Patients with L1 mutations show elevations in transcranial magnetic stimulation (TMS)

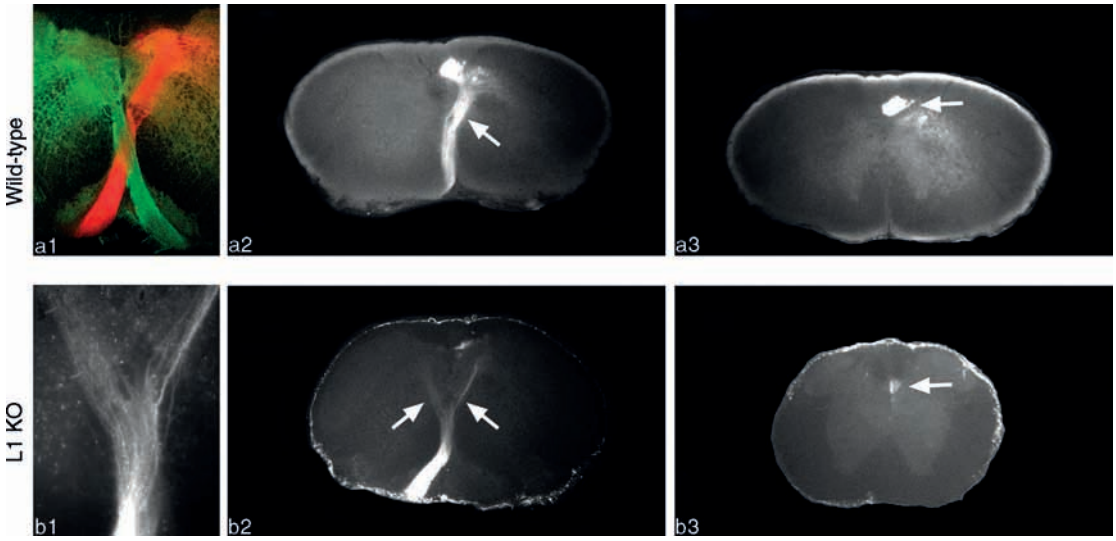

Figure 2 The pyramidal tract decussation and the role of L1-CAM in forming the contralateral descending CS projection. (a1) Image from a normal mouse showing the pyramidal decussation. Axons from the two sides are labeled red and green. (Image courtesy of Drs. Kyoko Itoh and Vance Lemmon.) (a2) Normal decussation in a wild-type mouse (arrow; similar to the red axons in (a1)). (a3) Descending axons in the CS tract (arrow). (b) The decussating axons from a L1-CAM mutant. (b1), (b2) Aberrant decussation from an L1-CAM mutant ((b1) is a high-magnification view). (b3) There are fewer contralateral descending axons (arrow) and no ipsilateral descending axons. Arrows in (a2) and (b2) mark decussating axons; arrows in (a3) and (b3) mark the presence of the dorsal corticospinal tract in the dorsal column. Magnification ×23 (a2, a3, b2, b3), ×44 (b1). From Itoh K, Cheng L, Kamei Y, et al. (2004) Brain development in mice lacking L1-L1 homophilic adhesion. *Journal of Cell Biology* 165: 145–154.

thresholds and increased latencies for evoking motor responses. These CS functional defects correlate with impairments in performance on several tests of hand/digit dexterity.

CS axons in EphA4 knockout mice decussate in the pyramid and descend into the appropriate white matter regions of the cord, but they terminate bilaterally in the spinal gray matter (Figure 3). This is because ephrin B3 is a midline growth barrier. Normally, it is present extensively along the spinal midline (Figure 3), thereby limiting recrossing. In the medulla, ephrin B3 is located only dorsal to the pyramidal decussation and at this level does not block CS axon crossing. Two Robo proteins, Robo1 and 2, are receptors for Slit. Robo is expressed on CS axons. CS axons may be prevented from aberrant recrossing through interactions between these Robo proteins on developing axons and Slit at the floor plate of the hindbrain and spinal cord. A third member of this family, Robo3, is functionally different from the other two members because it is required for decussation. In patients with a rare genetic disorder, horizontal gaze palsy with progressive scoliosis, mutation in the Robo3 gene prevents decussation of the CS tract. While most CS axons decussate in the medulla, approximately 10% do not and project to the ipsilateral cord. Moreover, once in the spinal cord, many CS axons – whether or not they decussate in the medulla – cross the midline before terminating on spinal neurons. The logic of the pattern of decussation of CS axons is not yet known.

Collateral Branching into the Gray Matter

Once CS axons reach the caudal brain stem, a small contingent pioneer the path into the spinal cord, followed later by waves of axons that further populate the CS tract. Long-distance growth of the primary descending CS axon is followed by formation of side, or collateral, branches that extend into the surrounding gray matter after a variable delay period. Spinal gray matter innervation is mediated by target-specific chemotropic factors that induce branching. In tissue explant experiments, for example, neurites from a portion of the prospective forelimb area of the sensorimotor cortex grow toward a cervical spinal explant but not to a lumbar explant.

Development of Connectional Specificity between CS Axon Terminals and SC Neurons

Topographic Refinement of CS Terminations in the Spinal Gray Matter

When CS axons initially grow into the spinal gray matter, their termination pattern is not what it is in maturity. The most common pattern is that CS axons have a widespread spinal termination pattern found early in development that is subsequently refined, or 'pruned,' to a more restricted distribution in maturity (Figure 4(a)). Terminations that are normally present

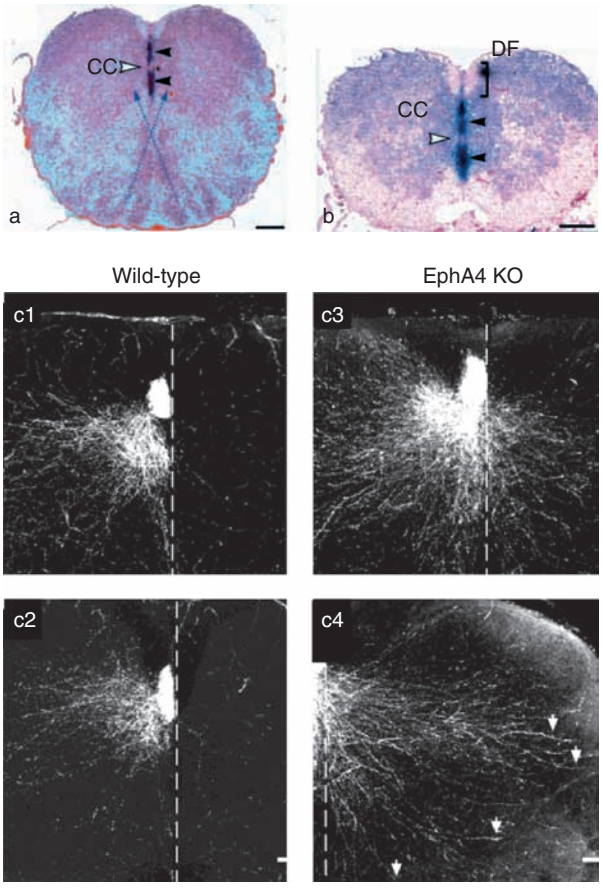

Figure 3 Ephrin B3 and its receptor EphA4 prevent recrossing of CS axons. (a, b) Ephrin B3 messenger RNA expression (dark blue staining; arrow heads) in the caudal medulla of a normal mouse, at the level of the pyramidal decussation ((a), dashed arrows) and in the spinal cord (b). (c1)–(c4) Region of the central canal. (c) Effects of EphA4 knockout on the development of the laterality of CS axon terminations in a 7-day-old mouse. In the normal mouse, terminations remain predominantly contralateral to the CS neurons of origin (c1) and (c2), whereas in the knockout mouse (c3) and (c4) there are abundant ipsilateral terminations. The arrows in (c4) point to aberrant terminations in the ventral and lateral gray matter. The distribution of terminal axons is also more extensive in the knockout mouse, suggesting a role for ephrin-Eph receptor interactions in determining regional distributions as well as laterality. Scale bars = 200 μm. (a, b) From Kullander K, Croll SD, Zimmer M, et al. (2001) Ephrin-B3 is the midline barrier that prevents corticospinal tract axons from recrossing, allowing for unilateral motor control. *Genes & Development* 15: 877–888. (c) From Coonan JR, Greferath U, Messenger J, et al. (2001) Development and reorganization of corticospinal projections in EphA4 deficient mice. *Journal of Comparative Neurology* 436: 248–262.

early in development and are eliminated later on are often termed 'transient terminations'. The rhesus monkey shows a pattern of progressive increase in CS axon terminal growth during early postnatal development rather than a broader distribution of terminations that is subsequently refined. Figure 4(b) shows that terminations within the central region of the gray matter become progressively denser (more red), especially within the lateral motor nucleus (yellow arrow). Irrespective of the particular pattern, the refined distribution of connections ultimately is achieved late in development: by 6–7 weeks in the cat, 8 months in the monkey, and after several years in the human. An important question for future research is why the CS system in some animals prunes back exuberant connections while others use a delayed outgrowth strategy.

CS terminations early in development are able to excite their spinal targets. This has been shown in the cat and rat by electrically stimulating CS axons in the CS tract and recording postsynaptic field potentials (i.e., focal synaptic potentials) in the spinal gray matter. In immature cats, for example, stimulation of the CS system evokes postsynaptic responses throughout the dorsoventral extent of the gray matter (Figure 5(a)). This broad distribution is similar to the one of CS terminations. By contrast, in mature animals, when CS axon terminals are restricted to the middle region of the cord, responses are largely limited to the same middle layers (Figure 5(b)). The refinement process restricts the effects of CS activation to particular spinal motor circuits. *In vitro* studies in rats show that synapse elimination is *N*-methyl-D-aspartate receptor-dependent. Although the anatomy and physiology of this process have been studied, the functional logic of the final termination patterns of CS axons has yet to be elucidated.

During postnatal development, local CS axon branch growth and presynaptic bouton formation are abundant (Figure 6; see also Figure 4(b) and increasing termination density). When axon branch elimination is present (Figure 4(a)), local growth occurs concurrently and complements the refinement process. Local axonal branch growth leads to a remarkable increase in the strength of CS evoked responses during early development (Figure 5; note increase from 30 to 700 μV range). As CS axon terminals mature, they also are better able to recruit spinal motor circuits into action (Figure 7). This has been tested in the cat by stimulating CS axons in the pyramid and recording motor responses in peripheral motor nerves. Immature CS connections require stronger electrical stimulation of CS axons, and a greater number of electrical stimuli, to evoke a motor response (Figure 7(a)) compared with more mature animals (Figure 7(b)). While comparable spinal recordings in the developing monkey have not been done, TMS has been shown to be effective in evoking motor responses only after about 3 months of age. This correlates with the progressive late growth of CS axon terminals and is consistent with the finding in the cat (Figure 7) that CS connections with spinal circuits become more effective in evoking motor responses later in development. In the monkey,

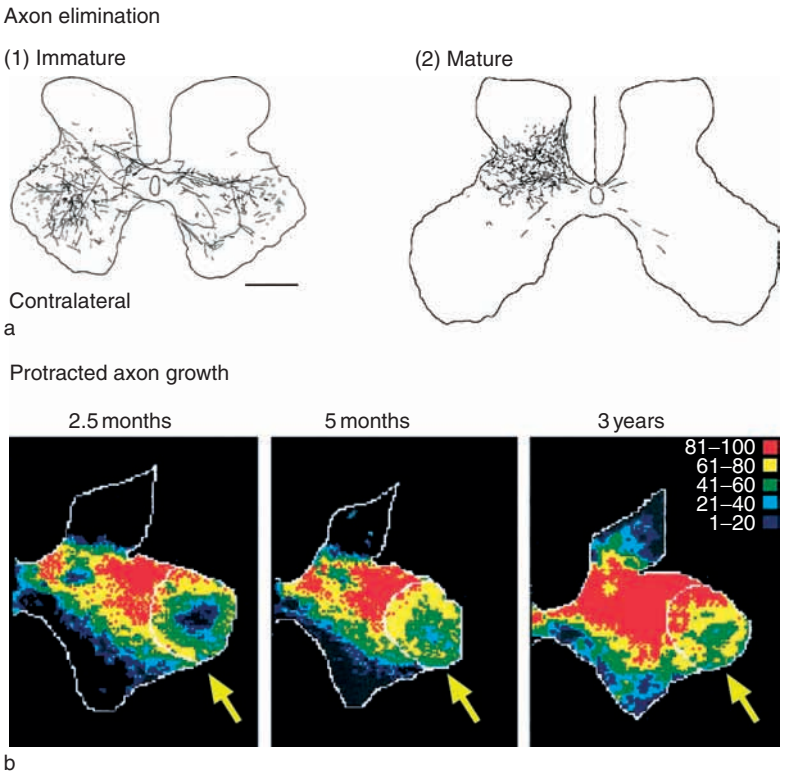

Figure 4 The distribution of CS axon terminations is refined postnatally. (a)(1) Extensive dorsoventral and bilateral CS terminations in a 5-week-old cat. (a)(2) An adult cat, where terminations are predominantly contralateral and restricted dorsoventrally. Scale bar: 1 mm. (b) Increase in CS axon terminal density in monkeys at three different ages. These sections were taken from the rostral portion of the first thoracic spinal cord segment. Color codes the density of terminations, on a scale from 1 to a maximum of 100. The arrow marks the region of the lateral motor nucleus, where motor neurons innervating distal arm muscles are located. Note that at 2.5 months, very little of the nucleus is densely labeled (red), and there are areas without any label, whereas at 3 years, the entire nucleus is labeled, and labeling in the outer rim is dense. (a1) From Li Q and Martin J (2000) Postnatal development of differential projections from the caudal and rostral motor cortex subregions. *Experimental Brain Research* 2000: 134–198. (b) From Armand J, Olivier E, Edgley SA, and Lemon RN (1997) Postnatal development of corticospinal projections from motor cortex to the cervical enlargement in the macaque monkey. *Journal of Neuroscience* 17: 251–266.

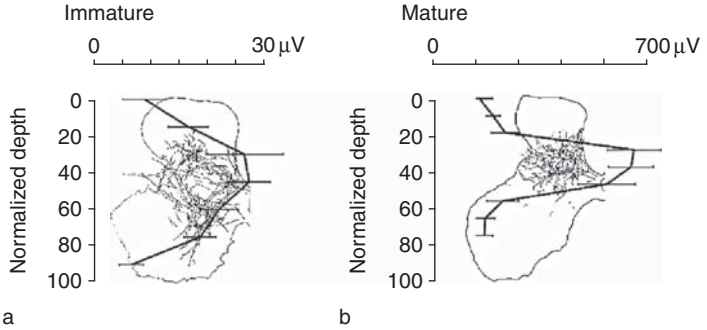

Figure 5 CS synaptic refinement. The amplitude of the monosynaptic CS evoked response is plotted at different depths (normalized; dorsal horn surface is zero and the ventral horn ventral surface, 100). Each graph overlies a schematic view of the distribution of terminations at that age. (a) The immature cat (5-week-old) shows small postsynaptic responses from the dorsal to the ventral surface. (b) By contrast, responses in the mature cat are restricted to the middle region, where the densest terminations are located. From Meng Z and Martin JH (2003) Postnatal development of corticospinal synaptic actions. *Journal of Neurophysiology* 90: 683–692.

TMS studies have also shown that there is an exponential increase in CV and decrease in TMS threshold within the first 1.5 years and a slower increase in CV and decrease in threshold later in development. Stronger connections between the motor cortex and spinal cord motor circuits are critical for the CS system's control over motor behavior.

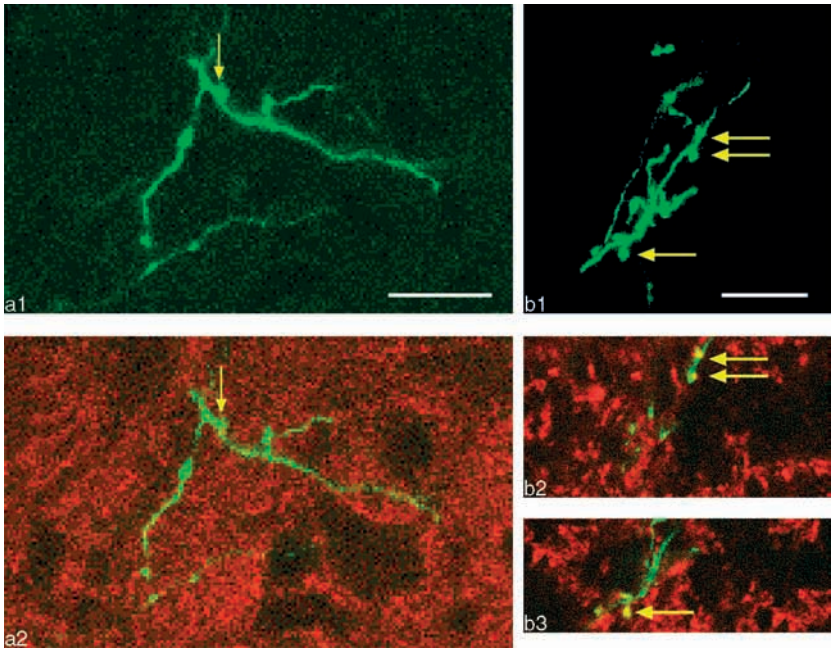

Figure 6 Confocal images from (a) an immature cat (4-week-old) and (b) a mature cat ((a1), (a2), (b2), and (b3): 1 μm optical slices; (b1): 7 μm projection image). The top two images show CS axons labeled with the anterograde tracer BDA (green). The lower row of images shows staining for the synaptic vesicle protein synaptophysin (red), together with axon staining (green). Double labeled boutons are yellow and highlighted with the yellow arrow. (a1) CS axon terminal with sparse branches. (b1) Highly branched mature axon terminal. Scale bar = 25 μm (a1, a2), 50 μm (b1–b3). From Meng Z, Li Q, and Martin JH (2004) The transition from development to motor control function in the corticospinal system. *Journal of Neuroscience* 24: 605–614.

Electrophysiological studies in humans using TMS also point to an early period of transient ipsilateral and ventral CS terminations, as in the cat. TMS up to about 1 year evokes bilateral motor responses (Figure 8(a)), consistent with bilateral terminations. The amplitude of the ipsilateral response becomes reduced relative to the contralateral response during the first year (Figure 8(a)). This is similar to the early postnatal elimination of ipsilateral CS branches in the cat. The TMS threshold for evoking muscle contraction is lower in preterm and very young infants, only to become higher during the first 3 postnatal months. This could be due to the elimination of some ventral CS axon terminal branches, thereby reducing synaptic effects. Between about 1 and 15 years, TMS threshold systematically decreases (Figure 8(b)). This threshold reduction is due to several factors, including increased CS axon terminal branching in the spinal cord, stronger connections in the spinal gray matter, and increased myelination of CS axons, which results in more-synchronous activation of spinal motor circuits.

Activity-Dependent Refinement of CS Axon Terminal Connections

The basis for the CS system's motor control functions depends on a refined pattern of connectivity between zones in the motor cortex that integrate motor control signals from various sources and spinal motor circuits. What drives postnatal refinement of CS axon terminations in the spinal gray matter? Recent studies show an important role for the level and pattern of CS neuronal activity in shaping development of the regional distribution of their spinal terminations, consistent with activity-dependent synaptic competition. Activity of the CS system can be selectively reduced by intracortical infusion of the γ-aminobutyric acid agonist muscimol and augmented by electrical stimulation of CS axons. Unilateral activity reduction results in an aberrant pattern of contralateral spinal terminations. The axons populate more dorsal laminae (Figure 9(b); yellow). This defect reflects a failure of silenced axons to maintain connections established earlier during development as well as a failure to add new branches and presynaptic sites (Figures 9(c) and 9(d)). Animals with this aberrant pattern of CS terminations have significant arm control impairments (see section on motor skill development).

The contralateral active CS system not only develops the normal contralateral projection but also maintains significant ipsilateral terminations. Thus, the reduction in termination space of the silenced side is balanced on that side by maintenance of ipsilateral

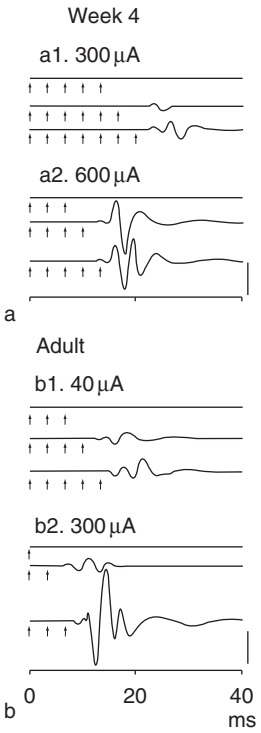

Figure 7 Ability of immature cat CS tract to recruit threshold motor responses is much less than in the adult. Each trace is an average recording from the median nerve contralateral to pyramidal tract electrical stimulation. The arrows indicate the time of occurrence of pyramidal stimuli. Note that more pulses and higher currents are needed to evoke responses in the 4-week-old animal than in the adult cat. Calibrations: (a) 500 μV; (b) 300 μV. From Meng Z, Li Q, and Martin JH (2004) The transition from development to motor control function in the corticospinal system. *Journal of Neuroscience* 24: 605–614.

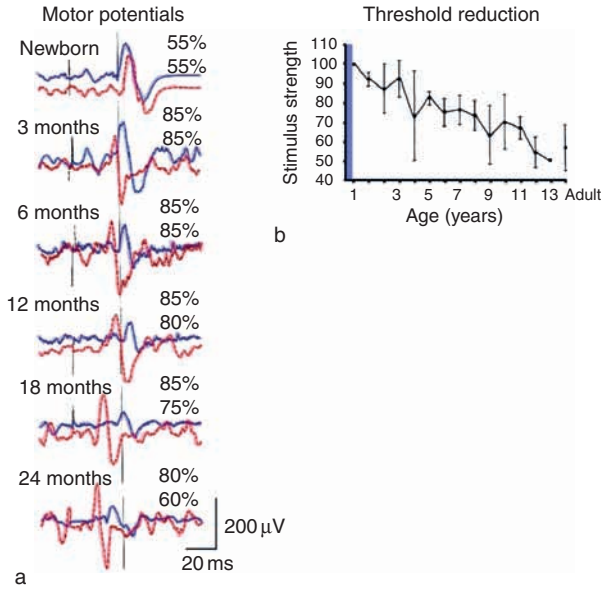

Figure 8 Transcranial magnetic stimulation (TMS) is used to activate the human CS system noninvasively. (a) Contralateral (red) and ipsilateral (blue) motor responses evoked by TMS over the motor cortex; 100% is maximal stimulus current. The amplitude of the ipsilateral response (blue) decreases with age. (b) The current for TMS to evoke a threshold motor response decreases from about the first to the 14th year. The *y*-scale of the graph plots current, as mean maximal current ±SD. Prior to 1 year of age, responses (blue) can be evoked at a lower threshold than at 1 year. (a) Modified from Eyre JA, Taylor JP, Villagra F, Smith M, and Miller S (2001) Evidence of activity-dependent withdrawal of corticospinal projections during human development. *Neurology* 57: 1543–1554. (b) From Nezu A, Kimura S, Uehara S, Kobayashi T, Tanaka M, and Saito K (1997) Magnetic stimulation of motor cortex in children: Maturity of corticospinal pathway and problem of clinical application. *Brain & Development* 19: 176–180.

terminations of the active system, which are the terminations that are normally eliminated. These topographic changes persist into maturity. The presence of ipsilateral terminations of the active side developing in parallel with the silenced CS terminations implies competition between the two sides during early development. Consistent with this idea is the observation that bilateral inactivation of the motor cortices produces a relatively normal pattern. This suggests that the changes in the laterality of CS terminals occurring after activity blockade are due to activity-dependent competition between developing CS terminals. The overall density of terminations after bilateral inactivation, however, is less than in controls, suggesting that the silenced CS terminations are less effective in competing with other spinal neural systems in securing terminations.

Another way to examine the role of activity of the two sides in shaping the mature pattern of terminations is to augment activity unilaterally. Electrical stimulation of CS axons in the medullary pyramid for several weeks during early postnatal life results in maintenance of substantial ipsilateral (and contralateral) terminations at 8 weeks (Figure 10(a)). Normally, there are predominantly contralateral terminations at this age. Projections from the nonstimulated side were displaced dorsally and laterally as a consequence of this CS system stimulation (Figure 10(c)). Both the maintenance of ipsilateral terminations and the displacement of nonstimulated axons are consistent with the activity-dependent competition model: The stimulated axons are more competitive at securing spinal synaptic space at the expense of the nonstimulated axons.

In hemiplegic cerebral palsy, the pattern of functional CS connectivity tested with TMS is similar to the anatomical distribution of CS axons in cats after unilateral CS system inactivation. TMS of the less impaired side evokes bilateral responses (Figure 11). TMS of the impaired side fails to evoke significant responses (not shown in figure). The aberrant ipsilateral effects are observed only after CS damage early in development, possibly before ipsilateral CS terminations are eliminated; in adults, stroke that produces hemiparesis does

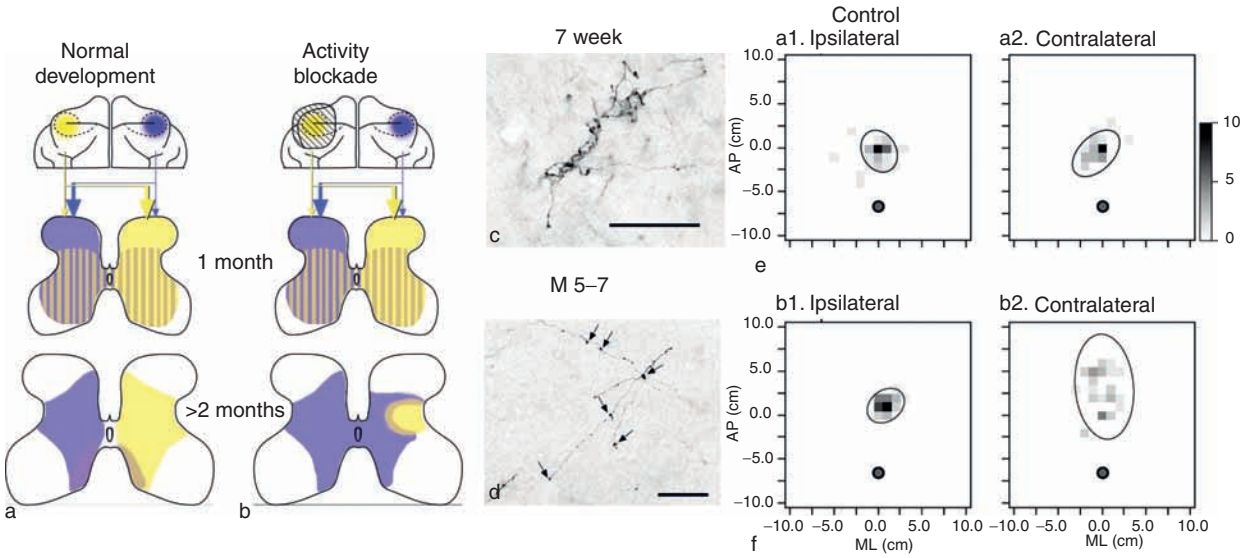

Figure 9 (a) Normally there are overlapping CS terminations in the spinal cord at 1 month in the cat. These terminations are refined into predominantly contralateral terminations after 2 months. (b) After motor cortex inactivation between weeks 5 and 7, the silenced axons (yellow) failed to populate as much of the contralateral gray matter. In contrast, the active side (blue) has a bilateral termination pattern. The ipsilateral axons are largely maintained transient terminations. (c) Morphology of a normal CS axon terminal at 7 weeks, and (d) morphology after inactivation. Arrows in (d) point to presynaptic boutons. Presynaptic boutons were too many to label in (c). Scale bars: 100 μm. (e, f) Results of behavioral experiments in which animals were trained to reach and grasp a food reward located at the origin (0,0). The start of the reach is indicated by the filled circle. (e) Data from an animal that received unilateral saline infusion into the primary motor cortex. (f) Animal that received unilateral muscimol infusion to block activity. Each square indicates the location of the end point of a reach. The more end points at that spot, the darker the square. The ellipses enclose the end points of 95% of the reaches. The area and the location of the ellipses are measures of accuracy. Reaching accuracy for the control animal is normal for both limbs. By contrast, accuracy for the contralateral limb in the animal in which the motor cortex was inactivated was significantly worse than for the ipsilateral side, which was normal. The end points for the contralateral side were both overshot (i.e., farther from the origin) and more dispersed. AP, anteroposterior; M5–7, muscimol inactivation between postnatal weeks 5–7; ML, mediolateral. (a–d) From Friel K and Martin JH (2005) Role of sensory–motor cortex activity in postnatal development of corticospinal axon terminals in the cat. *Journal of Comparative Neurology* 485: 43–56. (e, f) From Martin JH, Hacking A, and Donarummo L (2000) Impairments in prehension produced by early postnatal sensorimotor cortex activity blockade. *Journal of Neurophysiology* 83: 895–906.

not augment the ipsilateral response. These findings are consistent with the hypothesis that during early development, the impaired side is rendered much less competitive in securing and maintaining spinal synaptic space than the normal or less impaired side is.

Use-Dependent Development of the CS Tract

Motor experience comprises both directed control of muscle by motor centers in the brain and spinal cord and sensory information received by the nervous system as a consequence of movement. To determine the role of motor experience on CS development, limb use can be prevented or curtailed specifically during key periods when the CS system is developing. Preventing limb use during early development has a profound effect on CS axon terminal development. One way to prevent limb use during development is to block contraction of selected forelimb muscles using botulinum toxin A. Like motor cortex activity blockade, preventing limb use prevents the growth of CS axon terminals and presynaptic sites (Figures 12(a) and 12(b)). When limb use is regained, axon branching and presynaptic site density remain significantly reduced compared with controls, which indicates that CS axon terminals do not recoup lost connections (Figures 12(c) and 12(d)); moreover, the regional distribution of CS axon terminations remains aberrant (Figures 12(e) and 12(f)). These findings are consistent with the synaptic competition model, similar to what is observed after CS neural activity blockade. However, there is an important distinction: Preventing motor experience changes the patterns of neural activity in the developing motor and somatic sensory system but does not eliminate activity, as in the blockade experiments. This suggests that the developing CS system uses forms of temporally specific activity-dependent plasticity, such as long-term potentiation and long-term depression, to shape connections and function.

Development of CS Control of Skilled Motor Behavior

What is the relationship between development of the CS system and development of skilled limb control?

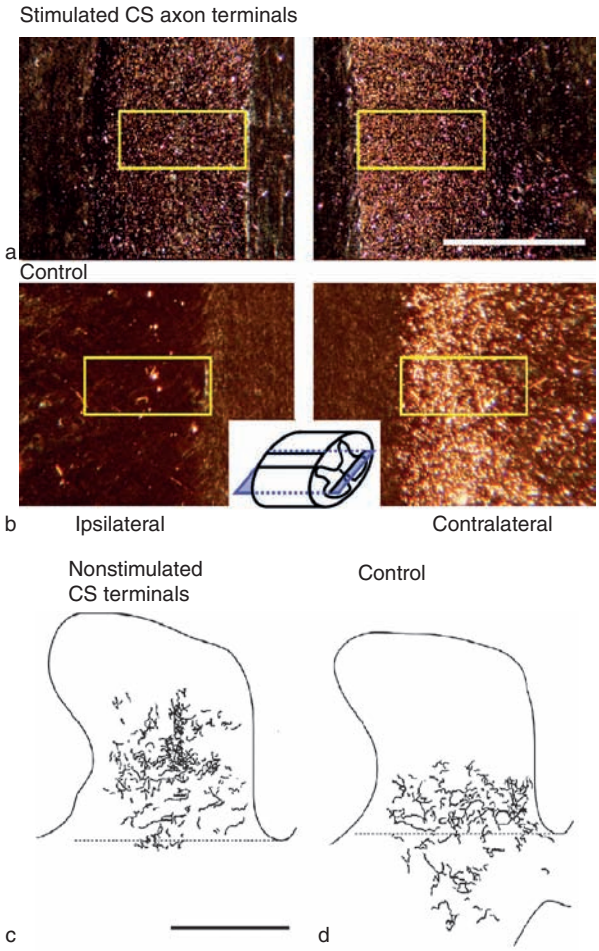

Figure 10 (a) CS tract electrical stimulation during early development promotes maintenance of ipsilateral terminations (left). Axons were labeled with wheat germ agglutinin conjugated with horseradish peroxidase, which appears as bright bronze particles. Inset shows plane of section in (a) and (b). Distribution of CS axon terminals from (c) the nonstimulated side (i.e., ipsilateral to the side of stimulation) and (d) a control stimulation (reticular formation). The intra-axonal distribution of tracer used in parts (c) and (d) is shown as the solid black lines. The dotted lines mark the ventral boundary of lamina 6. Note that stimulation displaced developing CS terminals dorsally in (c). Scale bar = 1 mm (a–d). From Salimi I and Martin JH (2004) Rescuing transient corticospinal terminations and promoting growth with corticospinal stimulation in kittens. *Journal of Neuroscience* 24: 4952–4961.

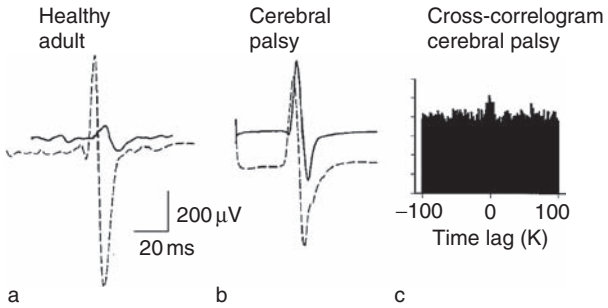

Figure 11 Ipsilateral (solid line) and contralateral (dotted) motor responses (flexor digitorum indicis) evoked by TMS in (a) a control adult and in (b) a patient with spastic hemiplegia acquired as a consequence of a stroke during infancy (i.e., cerebral palsy). Note that there is an abnormally large-amplitude ipsilateral response in the patient. (c) Correlogram for simultaneous electromyographic recordings from the two sides. The *y*-scale of the graph plots probably of occurrence $\times 10^{-2}$, from 0 to 15, and the *x*-scale plots time lag in milliseconds (K). The peak at zero indicates likely activation of electromyogram on both sides by a common signal. This suggests bilateral CS axon terminations. From Eyre JA, Taylor JP, Villagra F, Smith M, and Miller S (2001) Evidence of activity-dependent withdrawal of corticospinal projections during human development. *Neurology* 57: 1543–1554.

In many species there is clear evidence of co-development of skills, the spinal terminations of the CS tract, and the cortical motor representation. Monkeys begin to develop relatively independent finger movements during the period when CS axon terminals become dense within the intermediate zone and when CS terminals establish synapses within the motor nuclei. Humans develop relatively independent finger movements after about 1–2 years, as the thresholds for evoking motor responses begin to decline. However, as the CS system matures, other motor systems develop, as do sensory and cognitive systems. All these changes are likely to contribute to the expanding motor repertoire. Thus, because many CNS systems are developing during these early periods, establishing causal links between CS system developmental milestones and development of skilled control is difficult. Another way to obtain insight into how the CS system contributes to developing motor skills is to perturb normal CS tract development by changing CS activity levels or motor experience and determine the effect of these manipulations on skilled limb control.

Early postnatal blockade of motor cortex activity, which alters the termination pattern of CS tract axons, has a profound effect on visually guided skilled limb control. After the blockade is removed, there is a permanent aiming error in reaching, in which the end point of the movement overshoots the target (Figure 9(e)). During visually guided locomotion, there is also a movement end point defect, in which the foot is placed too far forward on the substrate. The common hypermetric defect (i.e., overreach and overstep) suggests a common underlying mechanism, such as a deficit in recruiting a muscle synergy for stabilizing limb position at the end of the movement. There also is a grasping impairment. After animals' motor experience is prevented and they are then allowed to regain limb use, their aim of reaching movements is normal, but their digit coordination during grasping is impaired (Figure 12(g)), similar to the grasping defect after postnatal inactivation.

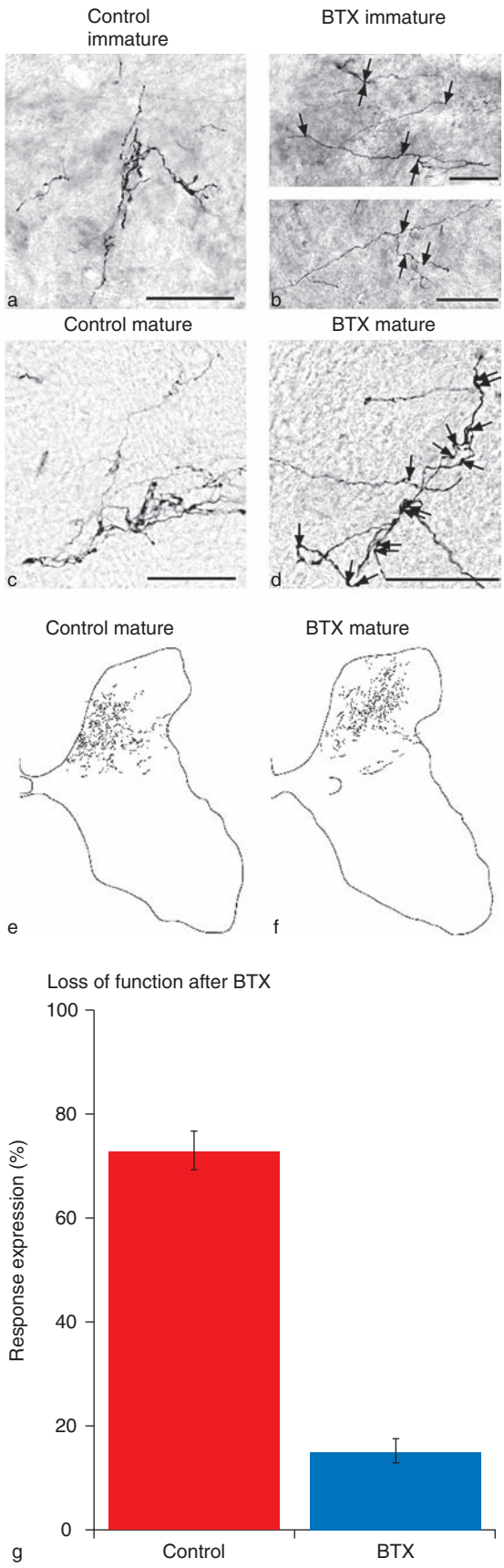

Development of the Cortical Motor Map

The CS tract carries the output from the primary motor cortex, as well as adjoining regions. In maturity, several characteristics of the motor representation in the primary motor cortex are thought to be critical for the normal expression of skilled movements, including a requisite size of the representation and the need for particular joints to be represented. In animals, the motor map can be assessed using microstimulation, whereby electrical stimulation through a microelectrode is used to excite a small population of cortical neurons. In development, the motor representation is first detected at 2 months in the cat; prior to this, motor cortex microstimulation does not evoke motor responses. The absence of a motor map before week 8 could be due to several factors, including weak intracortical synapses and weak CS synapses on spinal motor circuits (see **Figure 7**).

During the following month, there is an increase in the percentage of sites from which microstimulation evokes a motor response and a concomitant decrease in the current threshold (**Figure 13**). The threshold reduction suggests more-efficient activation of spinal motor control circuits, reflecting synaptic strengthening at both the cortical and the spinal levels (see **Figure 6**). The joints and muscles represented increases, from only proximal forelimb movements and muscles initially to both proximal and distal. In addition, as animals grow older, effects at multiple joints are produced at a higher percentage of sites. These multijoint sites are thought to play a role in encoding interjoint motor synergies. The role of experience and CS system activity in shaping motor map development is being studied in the cat: Training reduces the threshold for evoking responses and

Figure 12 Effects of preventing limb use in a cat between weeks 3 and 7. Morphological changes after preventing limb use persist into maturity. (a, c) Control micrographs of normal corticospinal axon terminals at 8 weeks and in maturity, respectively. (b) Micrograph of terminal at 2 months, which is immediately after the period of disuse. (d) Micrograph of terminal in a mature animal in which BTX was injected into forelimb muscles between weeks 4 and 7. Arrows in (b) and (d) mark axon varicosites (boutons). Calibrations in (a)–(d): 100 μm. (e) Regional distribution of contralateral CS terminals in a mature control and (f) in a mature animal in which BTX was injected into forelimb muscles between weeks 4 and 7. Calibration for (e) and (f) corresponds to the length of the bar in (c), but representing 500 μm. (g) Once the effects of BTX wear off, there is a permanent coordination impairment. The occurrence of coordinated digit flexion and forearm supination is plotted (mean ± SD) for the control side (red) and the side in which BTX was injected (blue). BTX, botulinum toxin, From Martin JH, Choy M, Pullman S, and Meng Z (2004) Corticospinal development depends on experience. *Journal of Neuroscience* 24: 2122–2132.

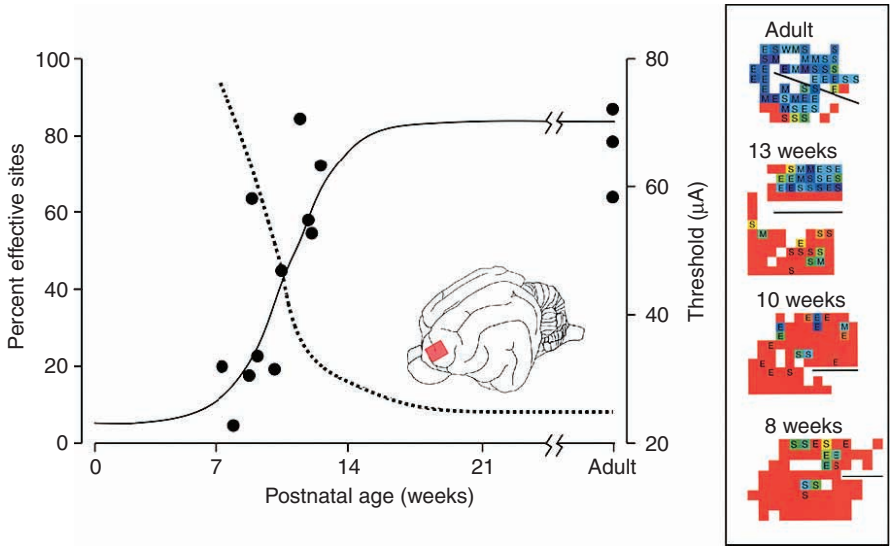

Figure 13 Development of the cortical motor map. The graph plots the percentage of sites from which microstimulation evoked a motor response at the various ages tested (solid line and dots; left scale) and the mean threshold for evoking the responses (dashed line; right scale). Each dot shows percentage data from a separate animal. Maps of the forelimb area at four selected ages are shown on the right. Color represents the amount of current needed to evoke a response. This threshold reduction can be seen on maps of the forelimb area of the cat motor cortex as a change from a preponderance of high-threshold (red) to low-threshold sites (blue). Red corresponds to a stimulation current of 100 μA, indigo to 1 μA. The general location of this area is indicated as the red box on the inset brain drawing. For the various maps, the lines indicate the cruciate sulcus, which roughly divides the rostral and caudal forelimb areas. E, elbow; M, multijoint responses; S, shoulder; W, wrist. From Chakrabarty S and Martin JH (2000) Postnatal development of the motor representation in primary motor cortex. *Journal of Neurophysiology* 84: 2582–2594.

increases the representation of coordinated muscle synergies, whereas preventing motor experience augments thresholds and reduces muscle synergy representation. These representational changes return to control levels several months after normal experience returns, reflecting plasticity that persists throughout the animal's life.

Conclusion and Implications for Rehabilitation

The CS system develops over a protracted pre- and postnatal period before achieving the mature distribution of spinal terminations. Activity-dependent synaptic competition between developing CS axon terminals, and possibly with other spinal synapses, shapes the formation of a complex set of stable connections between the motor cortex and particular spinal motor circuits. Both the level of CS neural activity and the pattern of activity, as reflected in the patterns of limb use, are critical to this system's development. The outcome is incorporation of the CS system into the developing motor systems and the expression of this system's contributions to skilled motor behavior.

Activity- and use-dependent development of the CS system assures that neural events and experience, as connections are being made, play an important role in forming the circuits for controlling movements that an individual makes throughout life. However, activity and use dependence also creates a vulnerability to deviations from an optimal functional state of the motor systems. It has been suggested that the vulnerability produced by activity-dependent competition between the developing CS tracts on the two sides of the brain could lead to a progressive worsening of the effects of a perinatal CS system lesion.

Can activity-dependent competition be harnessed to promote CS system function after early brain damage? Lessons can be learned from visual system development. It is well known that periods of monocular visual deprivation produced by cataracts or strabismus can lead to a worsening of sight as a child grows up. After correction of the ocular defect, vision is promoted in the impaired eye by periods of monocular deprivation of the normal-functioning eye. Visual acuity in the unimpaired eye is maintained at a normal level by permitting daily controlled visual experience. Work in the developing cat CS system points directly to overcoming functional impairment by rebalancing CS terminations later in development. There also is evidence in humans that rehabilitation strategies can be used to rebalance function in hemiplegic patients. A motor behavioral analog of monocular deprivation

is constraint-induced therapy, in which children with spastic hemiplegia have their unimpaired (or less impaired) arm physically restrained. Constraint-induced therapy forces them to use their affected arm exclusively during daily activities, which can lead to improvement in motor function. Convergence of animal research and human clinical studies leads to the optimistic view that some forms of developmental motor disorders might be significantly improved by harnessing activity-dependent processes shaping development of the CS system.

See also: Anterior-Posterior Spinal Cord Patterning of the Motor Pool; Corticomotoneuronal System; Games in Monkeys: Neurophysiology and Motor Decision-Making; Motor Skill Learning; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Spinal Motor Neurons: Properties.

Further Reading

Arlotta P, Molyneaux BJ, Chen J, Inoue J, Kominami R, and Macklis JD (2005) Neuronal subtype-specific genes that control corticospinal motor neuron development *in vivo*. *Neuron* 45: 207–221.

Armand J, Olivier E, Edgley SA, and Lemon RN (1997) Postnatal development of corticospinal projections from motor cortex to the cervical enlargement in the macaque monkey. *Journal of Neuroscience* 17: 251–266.

Chakrabarty S and Martin JH (2000) Postnatal development of the motor representation in primary motor cortex. *Journal of Neurophysiology* 84: 2582–2594.

Eyre JA, Miller S, Clowry GJ, Conway EA, and Watts C (2000) Functional corticospinal projections are established prenatally in the human foetus permitting involvement in the development of spinal motor centres. *Brain* 123: 51–64.

Eyre JA, Taylor JP, Villagra F, Smith M, and Miller S (2001) Evidence of activity-dependent withdrawal of corticospinal projections during human development. *Neurology* 57: 1543–1554.

Friel K and Martin JH (2005) Role of sensory–motor cortex activity in postnatal development of corticospinal axon terminals in the cat. *Journal of Comparative Neurology* 485: 43–56.

Gordon AM, Charles J, and Wolf SL (2005) Methods of constraint-induced movement therapy for children with hemiplegic cerebral palsy: Development of a child-friendly intervention for improving upper-extremity function. *Archives of Physical Medicine and Rehabilitation* 86: 837–844.

Kamiguchi H, Hlavin ML, and Lemmon V (1998) Role of L1 in neural development: What the knockouts tell us. *Molecular and Cellular Neurosciences* 12: 48–55.

Kamiyama T, Yoshioka N, and Sakurai M (2006) Synapse elimination in the corticospinal projection during the early postnatal period. *Journal of Neurophysiology* 95: 2304–2313.

Kullander K, Croll SD, Zimmer M, et al. (2001) Ephrin-B3 is the midline barrier that prevents corticospinal tract axons from recrossing, allowing for unilateral motor control. *Genes & Development* 15: 877–888.

Li Q and Martin JH (2001) Postnatal development of corticospinal axon terminal morphology in the cat. *Journal of Comparative Neurology* 435: 127–141.

Martin JH (2005) The corticospinal system: From development to motor control. *Neuroscientist* 11: 161–173.

Martin JH, Choy M, Pullman S, and Meng Z (2004) Corticospinal development depends on experience. *Journal of Neuroscience* 24: 2122–2132.

Martin JH, Kably B, and Hacking A (1999) Activity-dependent development of cortical axon terminations in the spinal cord and brain stem. *Experimental Brain Research* 125: 184–199.

Meng Z, Li Q, and Martin JH (2004) The transition from development to motor control function in the corticospinal system. *Journal of Neuroscience* 24: 605–614.

Molyneaux BJ, Arlotta P, Hirata T, Hibi M, and Macklis JD (2005) Fezl is required for the birth and specification of corticospinal motor neurons. *Neuron* 47: 817–831.

Nezu A, Kimura S, Uehara S, Kobayashi T, Tanaka M, and Saito K (1997) Magnetic stimulation of motor cortex in children: Maturity of corticospinal pathway and problem of clinical application. *Brain & Development* 19: 176–180.

Ohno T, Maeda H, and Sakurai M (2004) Regionally specific distribution of corticospinal synapses because of activity-dependent synapse elimination *in vitro*. *Journal of Neuroscience* 24: 1377–1384.

Ohno T and Sakurai M (2005) Critical period for activity-dependent elimination of corticospinal synapses *in vitro*. *Neuroscience* 132: 917–922.

Porter R and Lemon R (1993) *Corticospinal Function and Voluntary Movement*. Oxford: Oxford Science.

Salimi I and Martin JH (2004) Rescuing transient corticospinal terminations and promoting growth with corticospinal stimulation in kittens. *Journal of Neuroscience* 24: 4952–4961.

Corticothalamic Connections: Structure and Function

F Briggs and W M Usrey, University of California at Davis, Davis, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

For sensory systems, neurons in the thalamus provide the cortex with information about the external environment. The cortex, in turn, gives rise to feedback pathways terminating in the thalamus. As a result of this plan, neurons in the cortex not only respond to thalamic input but also help shape the nature of this input. This article summarizes the state of knowledge regarding corticothalamic circuits and their function on the basis of results primarily from studies of the visual system. Given the ubiquitous nature of corticothalamic connections, it seems likely that many of the fundamental properties of corticothalamic circuits are shared across sensory systems.

From Cortex to Thalamus: A Tale of Two Pathways

The corticothalamic pathway begins with neurons in the deepest layers of the cerebral cortex and ends in the thalamic nuclei serving that sensory modality. The two main types of corticothalamic projections are defined in part by their cortical layer of origin and in part by the thalamic targets they innervate. Corticothalamic projections from neurons in layer 6 typically target primary thalamic nuclei – subdivisions of the thalamus that receive ascending information from the periphery and relay that information to cortex. Axons of these corticothalamic cells are called type 1 axons. In contrast, corticothalamic projections from layer 5 neurons target secondary thalamic nuclei – subdivisions of the thalamus that receive mainly efferent cortical input and whose proposed function is to sort and relay information to various cortical targets, including higher-level areas. Axons of this type of corticothalamic neuron are called type 2 axons. One important feature distinguishing the two types of corticothalamic projections is the presence or absence of axonal branching within the reticular nucleus of the thalamus. Layer 6 corticothalamic axons (type 1) send axonal collaterals to the reticular nucleus whereas layer 5 corticothalamic axons (type 2) project few if any collaterals to the reticular nucleus. Instead, type 2 axons frequently send collateral projections that target a variety of nonthalamic structures. For the visual system, type 1 axons from neurons in layer 6 of primary visual cortex primarily terminate in the lateral geniculate nucleus (LGN) and the reticular nucleus, and type 2 axons from neurons in layer 5 terminate in the pulvinar nucleus (Figure 1).

Cortical Projections to the LGN

Layer 6 of primary visual cortex contains a myriad of morphologically and physiologically distinct cell types. In the macaque monkey, only about 14% of layer 6 neurons give rise to axons that target the LGN, although higher percentages have been found in carnivores and rodents. Layer 6 corticogeniculate neurons are generally short pyramids with basal dendrites restricted to layer 6 and apical dendrites that arborize in layer 4C – the primary target of ascending LGN input to cortex and analogous to layer 4 in carnivores and rodents. Such dendritic arborization allows corticogeniculate neurons to sample both intracortical activity and ascending LGN activity. Along these lines, physiological studies have shown that a subset of corticothalamic neurons receives direct thalamocortical input. Corticogeniculate neurons also give rise to local axons that synapse in layer 4C. As a result of this organization, corticogeniculate neurons are in a strategic position to influence the transmission of information from thalamus to cortex.

The primate LGN contains magnocellular, parvocellular, and koniocellular neurons that are distinct in their laminar location, visual physiology, and cortical projection patterns. Corticogeniculate axons provide synaptic input to all three classes of LGN neurons. In the macaque monkey, neurons in the upper tier of layer 6 have axons that target the parvocellular layers of the LGN, and neurons in the bottom tier have axons that project primarily to the magnocellular layers. In addition, a small percentage of corticogeniculate neurons in the lower tier of layer 6 provide input to either the parvocellular or the koniocellular layers. Although layer 6 neurons receive mixed intracortical input from both the magnocellular and parvocellular streams, there is evidence supporting a bias of magnocellular or parvocellular stream input to individual layer 6 neurons. Finally, although corticogeniculate projections to the LGN are topographically organized, the precision of these projections may not be as high as that of thalamocortical projections.

Corticogeniculate neurons share an intimate relationship with cells in the reticular nucleus. The reticular nucleus is a unique structure compared with its neighboring thalamic nuclei as it contains only γ-aminobutyricacidergic (GABAergic) neurons that provide inhibitory input to other thalamic nuclei, including the LGN. Because neurons in the reticular

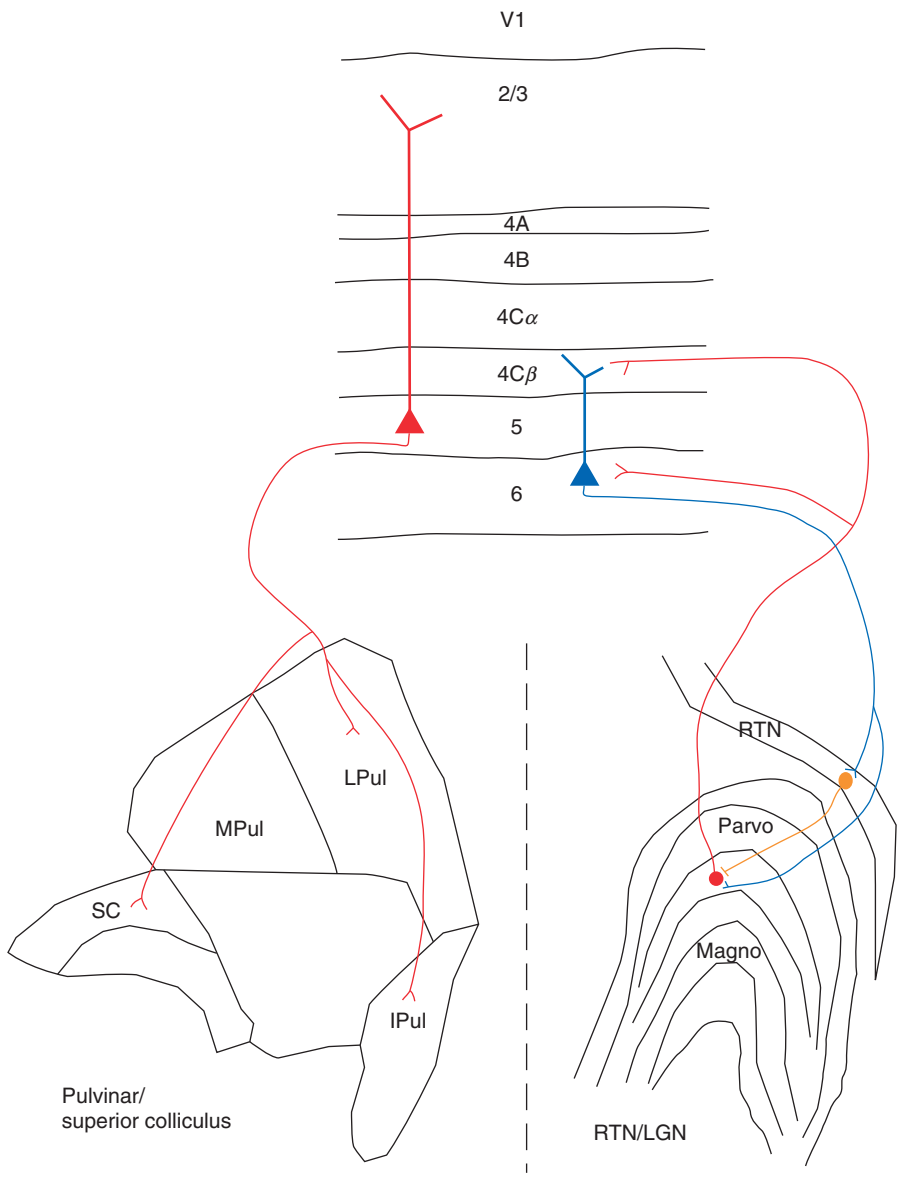

Figure 1 Type 1 and type 2 corticothalamic axons. Type 1 corticogeniculate axons (blue, right side; dendrites are thick and axons are thin) originate from neurons in layer 6 of V1 and project to the lateral geniculate nucleus (LGN) and the reticular nucleus (RTN). RTN neurons (orange) make inhibitory synapses onto thalamocortical neurons. Thalamocortical axons (red, right side) display type 2 characteristics and project to specific compartments within V1. Type 2 corticothalamic axons (red, left side) originate from neurons in layer 5 of V1 and target the pulvinar and superior colliculus. IPul, inferior pulvinar; LPul, lateral pulvinar; MPul, medial pulvinar; SC, superior colliculus.

nucleus receive direct input from thalamocortical axons as well as corticothalamic axons, these neurons are well situated to play an important role in the transfer of information between thalamus and cortex, and vice versa.

Cortical Projections to the Pulvinar

Layer 5 of primary visual cortex contains a separate and distinct population of corticothalamic neurons. These neurons are tall pyramids with dendritic and local axonal arbors that extend into the superficial layers of cortex. As a result, these neurons are better positioned to sample local circuit input rather than direct thalamocortical input. Indeed, local circuit inputs impinging on layer 5 corticothalamic neurons originate from all the cortical layers, with particularly strong input coming from neurons in layers associated with the magnocellular stream.

Corticothalamic neurons in layer 5 have type 2 axons that project to the pulvinar nucleus rather than the LGN. Notably, these axons frequently give rise to collateral branches that target the superior

colliculus and other regions of the brain stem associated with motor function. Within the pulvinar nucleus, the specific subdivisions innervated by layer 5 axons include the lateral margin of the lateral pulvinar, the central inferior pulvinar, and the medial inferior pulvinar. Although corticopulvinar axons generally avoid the reticular nucleus, there is some evidence for weak input to this structure. Additional evidence indicating that type 1 and type 2 corticothalamic axons may not be entirely separable comes from findings that some neurons in deep layer 6 with type 1 axons provide weak input to the pulvinar.

Corticothalamic Physiology and Function

Conduction Latency

Some of the earliest physiological experiments performed on the corticothalamic pathway used electrical stimulation to identify corticothalamic neurons and to determine the time required for action potentials to travel from cortex to thalamus. These experiments were conducted in multiple species and systems and included different classes of corticothalamic projections. By providing electrical stimulation to the thalamic nucleus of interest, investigators measured the time required for action potentials to back-propagate to simultaneously recorded neurons in the cortex. These experiments frequently utilized a collision test to ensure that action potentials following stimulation were indeed caused by antidromic rather than orthodromic mechanisms. Results from these studies demonstrate diversity in the corticothalamic population as conduction latency among neurons can range from just a few milliseconds to more than 50 ms. Accordingly, across species and systems, corticothalamic cells are often categorized as fast-, medium-, or slow-conducting cells. It has further been suggested that these categories of neurons might also be distinct in their sensory response properties and the specificity of their input to specific regions of their target structures.

Synapses and Receptors

Corticothalamic neurons in layers 5 and 6 use glutamate as their neurotransmitter. Beyond sharing a common neurotransmitter, however, the anatomical and physiological properties of their synapses are quite different. In fact, the synaptic distinctions between type 1 and type 2 axons have led some to hypothesize that the two pathways serve distinct functions: those of modulators and drivers, respectively.

Type 1 axons from neurons in layer 6 are strikingly similar in appearance across modalities and species. These axons tend to be thin and sparsely branching with many short terminal sprouts containing small synaptic swellings that contain round glutamate-filled vesicles (termed 'RS profiles'). In the LGN, synapses are made onto the distal dendrites of excitatory relay neurons at extraglomerular sites. In addition to providing input to excitatory neurons, type 1 axons also provide input to inhibitory neurons, both within the relay nucleus (on dendritic shafts, outside glomeruli) and within the reticular nucleus. In the reticular nucleus, type 1 axons activate both ionotropic and metabotropic receptors on their postsynaptic targets. Relay neurons in the LGN also receive type 1 input via ionotropic and metabotropic receptors; however, local circuit neurons may receive input only via ionotropic receptors. Activation of non-NMDA ionotropic receptors causes a stereotypical fast Na^+-based excitatory postsynaptic potential (EPSP) in the target neuron. NMDA receptor activation is somewhat more complicated in that it requires a prior depolarization to remove the Mg^{2+} block before the ion channel can be opened. Once the channel opens, an EPSP results; however, the delay is longer than that of non-NMDA EPSPs. As for the metabotropic receptors, glutamate activation starts a series of second messenger cascades leading to release of intracellular Ca^{2+} and a reduced K^+ leak conductance, which depolarizes the cell and causes a slow and long-lasting EPSP. While ionotropic receptor activation should permit faster, perhaps higher-fidelity transfer of action potentials from pre- to postsynaptic cells, the long-term effects of metabotropic receptor activation likely impose a control over certain activity patterns through a sustained depolarization of postsynaptic cells.

In addition to corticogeniculate effects exerted directly onto LGN relay neurons through distal synaptic connections, relay cells are also affected indirectly by corticogeniculate activity through connections with local circuit neurons and neurons in the reticular nucleus. Local circuit neurons release GABA as their neurotransmitter, activating postsynaptic zones of relay neurons and other local circuit cells containing ionotropic $GABA_A$ receptors. These interactions cause fast inhibitory postsynaptic potentials (IPSPs) in target neurons. Neurons in the reticular nucleus exert their inhibition onto relay cells via $GABA_B$ receptors. These receptors are located on the distal dendrites of relay cells. As a result of this organization, corticogeniculate activity might influence the firing mode of LGN relay neurons through polysynaptic interactions with the reticular nucleus.

Thalamic relay cells tend to operate in one of two modes: tonic and burst. In tonic mode, the resting membrane voltage is around -60 mV with Na^+ and K^+ voltage-dependent conductances primarily active. Faster, short-term changes in membrane voltage associated with ionotropic receptor activation are

consistent with tonic mode operation. Additionally, the prolonged depolarization brought about by direct corticogeniculate activation of metabotropic glutamate receptors may serve to keep a relay neuron in tonic mode for a longer duration. A switch to burst mode requires a certain chain of events initiated by a prolonged period of hyperpolarization. Such hyperpolarization could be achieved by $GABA_B$ receptor activation, especially at more distal portions of thalamic relay cell dendrites. Once in the prolonged hyperpolarized state, T-type Ca^{2+} channels, which are also located at distal dendritic sites, are de-inactivated and can be activated by a depolarizing input. This input causes a low-threshold Ca^{2+} action potential, on top of which rides a burst of high-frequency Na^+ action potentials. Following the burst of action potentials and subsequent inactivation of the calcium conductance and other associated conductances, either the pattern can be repeated by another prolonged hyperpolarization, or the cell can depolarize toward −60 mV and return to tonic mode. In summary, the variety of receptor types and channels located at or near corticogeniculate synapses onto thalamic relay neurons allows this pathway to play a role in determining the firing mode of thalamic relay neurons. Based on these findings, the corticothalamic pathway has been suggested to modulate the mode of LGN activity and information transfer from thalamus to cortex (described further in the next section).

In contrast to type 1 axons, type 2 axons from neurons in cortical layer 5 tend to be thick and quite richly branched. Their terminal swellings tend to be larger and heterogeneously shaped. The synapses at these swellings tend to contact the more proximal dendrites of postsynaptic neurons in the pulvinar nucleus – often within glomeruli and often forming triads with other thalamic afferents and local circuit neurons, but also at extraglomerular regions. The presynaptic terminals of these neurons are large and contain round glutamate-filled vesicles ('RL profiles'). The postsynaptic receptors are ionotropic (likely both NMDA and non-NMDA). The resultant EPSP is similar to that described above for postsynaptic ionotropic receptors. However, the synapses of type 2 axons are more proximal to the cell body and are often combined with processes from other neurons in glomeruli, likely leading to different current outcomes at the postsynaptic cell's axon hillock compared with the current outcomes of type 1 axon activation. Based on these findings, type 2 axons have been suggested to serve as drivers, rather than modulators, of postsynaptic activity during sensory processing. It is not known whether synapses from type 2 axons can activate postsynaptic metabotropic glutamate receptors (or whether such receptors even exist). However, type 2 afferents from the retina to the LGN, another driver pathway, do not activate metabotropic glutamate receptors in their target neurons. When encapsulated within a glomerulus or as a portion of a triad, synapses from type 2 axons contact local circuit neurons in combination with thalamocortical neurons. The inhibitory contributions of the local circuit cells, as described above, could likewise exert a form of feedforward inhibition. However, the influence of inhibition at the microlevel within a glomerulus might have a profoundly different outcome at the postsynaptic cell body.

The Influence of Corticothalamic Feedback on Sensory Processing

Despite the anatomical strength of the corticothalamic pathway (e.g., individual LGN neurons receive two to five times more synapses from the cortex than from the retina), we know remarkably little about the function of this pathway during sensory processing. Nevertheless, a number of proposed roles have been assigned to the corticothalamic pathway, based mainly on studies of the pathway from cortex to the LGN. These proposed roles can be sorted into two broad categories: (1) the corticothalamic pathway serves to sharpen the receptive field properties of thalamic neurons, and (2) the corticothalamic pathway serves to enhance the transmission of sensory information from periphery to cortex. Although less is known about the functional properties of the corticopulvinar pathway, recent theories have been introduced that suggest a role for this pathway in relaying information from one cortical area to the next.

Past studies in the visual, auditory, and somatosensory systems have led researchers to suggest that corticothalamic feedback from neurons in layer 6 serves to sharpen the receptive fields and/or shift the tuning of thalamic responses to sensory input. In the visual system, the strength of the extraclassical or suppressive surround that extends beyond the classical receptive field has been argued to depend on corticogeniculate feedback. Because of the suppressive nature of this surround, stimuli restricted within the classical receptive field would elicit greater responses than would stimuli that extend into the surround. As a result, the extraclassical surround serves to sharpen the borders of the classical receptive field. The notion that corticothalamic feedback serves to sharpen receptive fields is also supported by work in the somatosensory and auditory systems. In most cases, these studies support the view that feedback enhances responses to stimuli within the center of the receptive field and diminishes responses to stimuli that extend into the surround. A variation on this theme is found in the auditory thalamus of echolocating bats. Here, a process referred to as

egocentric selection occurs in which cortical feedback appears to shift the tuning of thalamic neurons toward that of activated corticothalamic neurons. In the process, the transfer of stimulus-matched information is enhanced, and nonmatched information is weakened and ignored.

Corticothalamic feedback has been proposed to influence information transfer in a number of ways. As described above, corticothalamic feedback might serve to shift thalamic neurons between burst and tonic modes of firing. While no direct evidence for this hypothesis exists, the necessary machinery is present in the receptors and channels located at corticothalamic synapses. Determining the mode of LGN activity has significant ramifications for the transfer of information from LGN to cortex, as burst spikes are not only more effective than tonic spikes at driving cortical responses but also differ in subtle ways from tonic spikes in their visual response properties.

Distinct from burst and tonic activity, corticogeniculate feedback has also been proposed to synchronize the firing of LGN neurons. Although the source of this synchrony has been the subject of debate, separate studies have shown that near-synchronous spikes from two LGN neurons that provide convergent input to a common cortical cell are more likely to drive cortical responses than spikes displaced by little more than 10 ms. Thus, there appears to be a mechanism in place to preferentially detect synchronous inputs.

Finally, corticothalamic feedback has been suggested to gate the flow of information from periphery to cortex. In the most extreme case, corticothalamic connections are believed to contribute to the neuronal circuits that underlie sleep, an obvious case in which thalamocortical transmission is diminished or disrupted. At the other extreme, spatial attention has been argued to increase the activity of LGN neurons. If LGN activity is indeed influenced by attention, then it seems likely that the corticothalamic pathway plays a key role in mediating these effects. Last, somewhere in between these two extremes is the notion that corticothalamic feedback serves to increase the gain of LGN responses to otherwise weak stimuli, such as low-contrast visual stimuli.

Conclusion

Although we have a good understanding of the components of the corticothalamic pathway, our understanding of the function of this pathway during sensory processing is limited. This could be because the definitive experiment to determine corticothalamic function has yet to be conceived or the tools to study such an anatomically complex pathway have yet to be developed. Alternatively, we may not be thinking about the corticothalamic pathway in the right context. The brain does not function in separable pieces but as a flowing, dynamic whole. The key to understanding the corticothalamic pathway may thus lie in examining it within the context of a larger functional circuit including sensory afferents, thalamocortical afferents, corticothalamic efferents, and intrinsic circuits. A daunting task indeed, but perhaps the visual pathway of the future.

See also: Cerebral Cortex; Cortical Plasticity and Learning: Mechanisms and Models; Corticothalamic Connections: Ultrastructure; Neural Patterning: Arealization of the Cortex; Visual Cortical Models of Orientation Tuning.

Further Reading

Alitto HJ and Usrey WM (2003) Corticothalamic feedback and sensory processing. *Current Opinion in Neurobiology* 13: 440–445.

Briggs F and Callaway EM (2001) Layer-specific input to distinct cell types in layer 6 of monkey primary visual cortex. *Journal of Neuroscience* 21: 3600–3608.

Fitzpatrick D, Usrey WM, Schofield BR, and Einstein G (1994) The sublaminar organization of neurons in layer 6 of macaque striate cortex. *Visual Neuroscience* 11: 307–315.

Jones EG (1985) *The Thalamus*. New York: Plenum Press.

Llinas RR and Steriade M (2006) Bursting of thalamic neurons and states of vigilance. *Journal of Neurophysiology* 95(6): 3297–3308.

Sherman SM (2001) Tonic and burst firing: Dual modes of thalamocortical relay. *Trends in Neurosciences* 24: 122–126.

Sherman SM and Guillery RW (2002) The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1695–1708.

Sherman SM and Guillery RW (2005) *Exploring the Thalamus and Its Role in Cortical Function*, 2nd edn. Boston: MIT Press.

Sillito AM and Jones HE (2002) Corticothalamic interactions in the transfer of visual information. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1739–1752.

Steriade M, Jones EG, and McCormick DA (eds.) (1997) *Thalamus*. New York: Elsevier.

Suga N and Ma X (2003) Multiparametric corticofugal modulation and plasticity in the auditory system. *Nature Reviews Neuroscience* 4: 783–794.

Usrey WM (2002) Spike timing and visual processing in the retinogeniculocortical pathway. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1729–1737.

Usrey WM, Alonso J-M, and Reid RC (2000) Synaptic interactions between thalamic inputs to simple cells in cat visual cortex. *Journal of Neuroscience* 20: 5461–5467.

Wiser AK and Callaway EM (1996) Contributions of individual layer 6 pyramidal neurons to local circuitry in macaque primary visual cortex. *Journal of Neuroscience* 16: 2724–2739.

Worgotter F, Eyding D, Macklis JD, and Funke K (2002) The influence of the corticothalamic projection on responses in thalamus and cortex. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1823–1834.

Corticothalamic Connections: Ultrastructure

K S Rockland, RIKEN Brain Science Institute, Saitama, Japan

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Corticothalamic (CT) connections are one of several major output systems to subcortical centers. In general, they are considered to reciprocate thalamocortical (TC) connections, and are numerically and functionally important. In the primary sensory thalamic nuclei, CT terminations have been thought to have a modulatory role, in contrast with inputs from the sensory periphery (retina, cochlea, or somatosensory systems), which are essential ('driving') for the receptive field response properties of neurons in the recipient nuclei. In associational thalamic nuclei, the anatomical organization is typically more complex, in that CT terminations originate from multiple cortical areas, and are composed of multiple subtypes. This has led to the suggestion that different subpopulations operate in modulatory or driving modes. This and other proposals about the role of CT connections are briefly discussed in the last section of this article.

CT connections were identified as fiber tracts as early as the end of the nineteenth century and were identified in finer detail at about the same time, in Golgi-stained preparations. During the twentieth century, major advances were made in defining neurons of origin in various cortical areas, terminal fields in the target thalamic nuclei, and basic synaptic microcircuitry. These three topics are developed later in this article. Overall, emphasis is on associational thalamic nuclei, which receive connections from associational cortical areas and are concerned with more integrative or multimodal processing. The lateral geniculate nucleus and other primary sensory nuclei might be considered something of a special case as regards CT terminations, and, moreover, are extensively discussed in textbooks and other source material. Finally, species-specific and nucleus-specific features have been largely neglected here (but see the section 'Further reading' for literature by Rouiller and Welker).

Cells of Origin

Layers

By injecting retrograde tracers into the thalamus, projection neurons can be visualized histologically in cortical areas. CT projecting neurons are found in the deeper cortical layers, in layers 5 and 6. These layers also contain other projection neurons, including some corticocortical neurons. Sublaminar location (i.e., layer 6A vs. layer 6B) and the proportion of neurons in layers 5 and 6 will vary, depending on the species of animal and the specific corticothalamic connection. Some nuclei, such as the lateral geniculate nucleus, receive connections only from layer 6, but most of the association thalamic nuclei receive CT connections from both layers 5 and 6. In that case, the population in layer 6 is much larger. When several cortical areas project to the same thalamic nucleus, terminations may originate from neurons only in layer 5 of one area, only from neurons in layer 6 of another area, or from a mixed population in yet another area (see 'Further reading,' Rouiller and Welker). In rats and cats at least, some nuclei receive bilateral connections.

Subtype Characteristics

CT neurons in layers 5 and 6 are distinguished by several distinct properties. (1) They have distinctive terminal morphology (see later). (2) They have distinctive extrinsic collateralization. At least some neurons in layer 6 send collaterals to the reticular nucleus of the thalamus (RTN). Layer 5 neurons, in contrast, frequently send collaterals to the superior colliculus or other brain stem regions, but not to the RTN. (3) They have distinctive intrinsic cortical collateralization. Layer 5 cells have extensive horizontal collaterals within layers 3 and 5; but intracellular fills of layer 6 corticothalamic neurons in cat auditory cortex have shown more confined recurrent collaterals, vertically oriented in the vicinity of the cell body. Further differences have been documented in the intrinsic collateralization of layer 6 neurons in monkey visual cortex, suggestive of additional subdivisions within the layer 6 population. (4) A physiological study in cat motor cortex demonstrated distinct spike trains for neurons in layer 5 and provided evidence of at least two subdivisions within the CT neurons in layer 6. CT neurons in layer 5 are fast conducting, show step-related activity, have sensory receptive fields, and are spontaneously active; one population of neurons in layer 6 is slow conducting, and shows none of the properties of the neurons in layer 5; and a second population of layer 6 neurons, distinguished as moderately conducting, has precisely timed step-related behavior.

Terminal Arbors

Light Microscopy

Background Terminal arbors can be visualized by Golgi stains of fixed tissue, by intracellular injection

of various tracers, or by extracellular injections of anterograde tracers such as biocytin, *Phaseolus vulgaris* leukoagglutinin, or biotinylated dextran amines. The most precise approach is intracellular injection of identified parent neurons in the cortex (*in vivo*), with subsequent complete visualization (in histology preparations) of the axon trajectory and terminations. As intracellular injections are technically demanding, most data have been derived from relatively large extracellular injections (diameter ~1.0 mm). These allow confident identification of thalamic terminations as having originated from neurons within the cortical injection site, but the exact cells of origin are obscured within the tracer injection.

Morphological subtypes Starting in the 1990s, several studies using high-resolution anterograde tracers reported multiple types of CT arbors at the light microscopic level. These have been grouped in two broad categories: type 1 and type 2 (Figures 1–4). Type 1 terminations consist of small, stalked terminations ('drumstick-like,' or *boutons terminaux*), which look very similar to dendritic spines. The axon is usually thin, and the terminal portion is very long (>1.0 mm). Type 2 terminations consist of predominantly large, beaded boutons (*boutons en passant*). The axon is slightly thicker (diameter ~ 1.0 μm) and

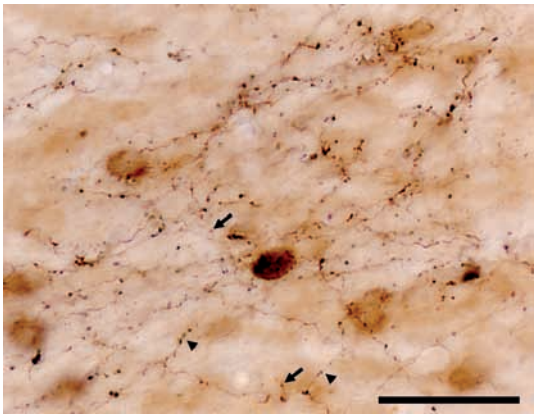

Figure 1 Light micrograph of lateral geniculate nucleus (parvocellular layer 4; macaque monkey). Corticothalamic terminations anterogradely labeled by an injection of biotinylated dextran amine in area V1. Note small terminations that are beaded (arrowheads) or stalked (arrows). Some thalamocortical neurons as well have been retrogradely labeled by the injection. Scale bar = 50 μm.

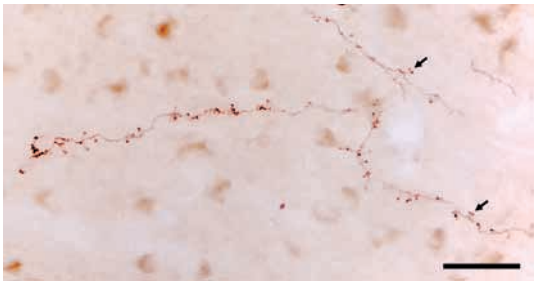

Figure 2 Light micrograph of medial pulvinar (macaque monkey). A segment of a corticothalamic axon, anterogradely labeled by an injection of biotinylated dextran amine in parietal cortex. This is a type 1 axon, characterized by an abundance of stalked ('drumstick') terminations. These occur in various patterns: singly, in clusters, in doublets (arrows), and intermixed with some beaded terminations. The overall configuration is more divergent than is seen in V1 axons in the lateral geniculate nucleus, and the stalked terminations are more pronounced. Scale bar = 50 μm.

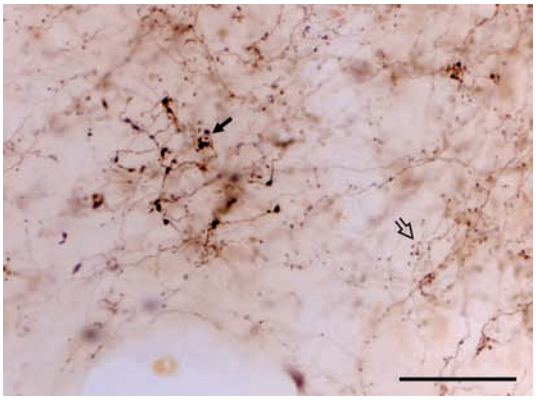

Figure 3 Light micrograph of medial pulvinar (macaque monkey). A field of corticothalamic terminations anterogradely labeled by an injection of biotinylated dextran amine in parietal cortex. A small arbor with type 2, larger boutons (solid arrow) is visible in a field of type 1 terminations (hollow arrow). Scale bar = 50 μm.

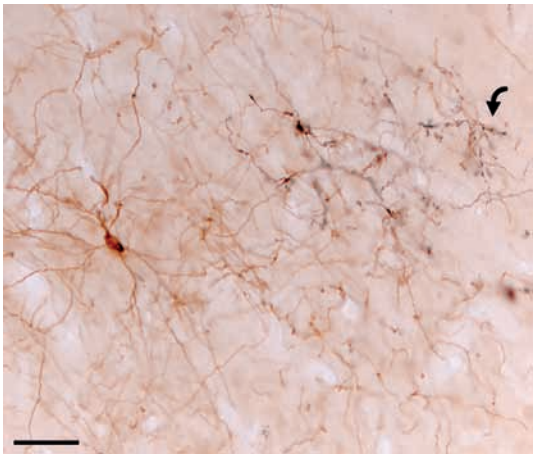

Figure 4 Light micrograph of lateral pulvinar (macaque monkey). Cortical injections in temporal areas were made of adenovirus and biotinylated dextran amine. A thalamocortical neuron (brown) has been labeled with green fluorescent protein (reacted with anti-green fluorescent protein antibody and 3,3′-diaminobenzidine), after infection by adenovirus. Anterogradely labeled corticothalamic terminations (in black) are also visible. Curved arrow indicates a concentration of type 2, large boutons, targeting proximal dendrites of a thalamocortical neuron (soma not visible). Scale bar = 50 μm.

the terminal portion forms a well-defined arbor, frequently about 0.25 mm in diameter. Because of the distinctive shapes of the terminal field, an alternate terminology was proposed, of 'elongated' for type 1 and 'round' for type 2, but the classical terminology has so far been more popular.

The two broad types of connections have been associated with neurons in different cortical layers. Type 1 terminations are thought to arise from neurons in layer 6, while type 2 terminations arise from neurons in layer 5. This conclusion is based (1) on direct evidence from laminar-specific injections in auditory cortex of cat, and from a small number of single cell injections in rats and (2) on indirect numerical evidence. That is, there are large numbers of type 1 arbors and large numbers of retrogradely filled CT neurons in layer 6, and small numbers of type 2 arbors and small numbers of retrogradely filled CT neurons in layer 5.

In the primate association thalamus, further morphological subdivisions have been noted in both type 1 and type 2 CT terminations. On the basis of arbor shape, arbor size, and subtle differences in bouton size and shape, two further subdivisions have been proposed for type 1 terminations, and four subdivisions for type 2. This is consistent with ongoing characterization of subpopulations of layer 5 and layer 6 pyramidal cells, although the functional implications are still unclear.

In the visual system, CT connections to the lateral geniculate nucleus (LGN) are only type 1 (Figure 1). Type 2 large-bouton terminations do occur, but these belong to retinogeniculate terminations. In the LGN, this strong morphological dichotomy correlates with functional differences, where the type 2 retinogeniculate terminations have a driving influence (e.g., they are essential for normal visual response properties of LGN neurons), while the type 1 CT terminations are more modulatory. The driving type 2 versus modulatory type 1 distinction has been hypothesized to apply more generally. Thus, when the two types of CT terminations occur together, as in association thalamic nuclei, type 1 terminations are proposed to be modulatory and type 2 driving. Consistent with this distinction, type 1 terminations have been shown to use metabotropic receptors, associated with slow depolarization, while type 2 terminations are associated with the faster ionotropic receptors.

Electron Microscopy

Type 1 and Type 2 Boutons

Electron microscopic (EM) studies in the 1960s and 1970s reported the occurrence of large (type 2) and small (type 1) excitatory boutons. The positive identification of these boutons as CT required a prelabeling step, whereby anterograde tracer injections were made in cortical areas *in vivo*, followed by perfusion and histological processing. As early as 1979, this was achieved by intracortical injections of tritiated amino acids. Also used for this purpose were horseradish peroxidase conjugated to wheat germ agglutinin, or biotinylated dextran amine (Figure 5).

At EM magnifications, both small and large boutons were seen to contain round synaptic vesicles, characteristic of excitatory terminations, and were accordingly called LR ('large round,' up to 9.0 μm^2) and SR ('small round,' 0.2–2.0 μm^2). Both are associated with a postsynaptic density, but this is more pronounced for SR boutons. LR boutons contain mitochondria, but SR boutons often do not. LR boutons in addition have filamentous contacts (Figure 6). These are established with somata or dendrites of projection neurons. Their function is not known, but possibilities include an adhesive role (i.e., desmosomoid attachments), a role in the establishment and maintenance of precise connections, or a role in some form of energy-dependent intercellular communication. The extensive networks of agranular endoplasmic reticulum localized along both sides of filamentous contacts have been suggested to be sites for the storage and release of calcium ions.

Another particular feature identified by EM is that type 2, but not type 1, boutons frequently occur within glomeruli (also called 'synaptic islands'). These are large aggregations of excitatory and inhibitory synaptic terminals associated with a large central dendrite and encapsulated by astrocytic cytoplasm (Figures 6–8). The central element can sometimes be an LR terminal. The inhibitory terminations arise from external sources, such as the RTN or pretectal nucleus, or from local circuit neurons, and are identifiable by their flat synaptic vesicles Thalamic local circuit neurons have specialized dendrites which are both pre- and postsynaptic. The configuration and density of glomeruli vary with species and nuclei. In rats, except for the LGN, the glomeruli do not contain presynaptic dendrites of inhibitory neurons. Glomeruli, defined as aggregations of synapses wrapped by glia, occur in other areas besides the thalamus, such as the cerebellar nuclei.

Postsynaptic Target

EM verification of synaptic contacts is essential for specification of microcircuitry. The main postsynaptic targets in the thalamus are TC relay neurons and interneurons. Immunolabeling for γ-aminobutyric acid (GABA) is often used in discriminating between GABA-positive interneurons and GABA-negative relay neurons.

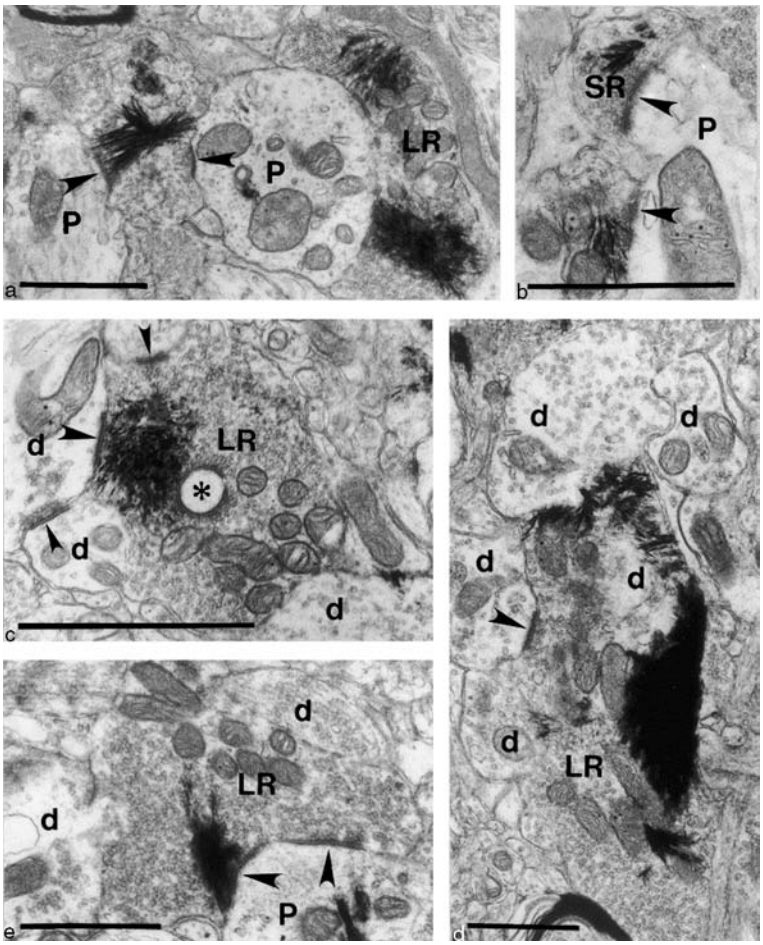

Figure 5 Electron micrographs of wheat germ agglutinin-conjugated horseradish peroxidase-labeled corticothalamic boutons in the lateral posterior nucleus, retrogradely labeled by an injection in parietal cortex (macaque monkey). (a) Examples of two labeled boutons. At the right, a large bouton with round vesicles (LR) is in close apposition with the proximal projection neuron dendrite (P). The medium-size bouton on the left forms two synapses (arrowheads): one with the proximal projection neuron dendrite (right-hand P) and another with a smaller diameter branch (left-hand P). (b) Two small, round vesicle-containing boutons (SR) with asymmetric synapses (arrowheads) on a distal projection neuron dendrite (P). (c) Example of a complex synaptic arrangement with a labeled LR bouton. It forms two asymmetric synapses (two arrowheads in the top left corner) on two vesicle-containing dendrites (labeled d), one of which forms a symmetric synapse (bottom left) with another vesicle-containing dendrite; the asterisk denotes a spinelike process, most likely belonging to a projection neuron dendrite, that protrudes into the LR bouton and has one continuous synapse encircling it. (d) An example of an LR bouton surrounded by vesicle-containing dendrites (labeled d) in a glomerulus. Arrowhead points to a synapse between the bouton and one of the vesicle-containing dendrites. (e) An example of an LR bouton forming long synapses (arrowheads) on a labeled projection neuron dendrite (P) and surrounded by two vesicle-containing dendrites (labeled d). All scale bars = 1 μm. Reprinted from Taktakishvili O, Sivan-Loukianova E, Kultas-Ilinsky K, et al. (2002) Posterior parietal cortex projections to the ventral lateral and some association thalamic nuclei in *Macaca mulatta*. *Brain Research Bulletin* 59: 135–150, with permission from Elsevier.

In primates and cats, but not rodents, thalamic nuclei contain 15–25% inhibitory interneurons. Both SR and LR boutons can synapse on inhibitory neurons, in variable proportions, depending on specific nucleus and species. (The LGN in all species, including rodents, contain a proportion of inhibitory interneurons.)

TC relay cells can be identified on the grounds of being GABA negative. They can also be identified by lack of the synaptic vesicles that occur in presynaptic dendrites of local circuit neurons. Relay cells can be labeled by intracellular tracers or, more commonly, by retrograde tracer injections in cortex, although these often fill only cell somata and very proximal dendrites. Proximal–distal location along a dendrite can be inferred on the basis of dendritic caliber.

Type 1 synapses tend to target distal dendrites, while type 2 synapses are usually located more proximally (Figure 4). In some nuclei, however, the distribution of LR and SR boutons is more even. Neurons in the centromedian interlaminar nucleus, for example, exhibit more intermingling of LR and

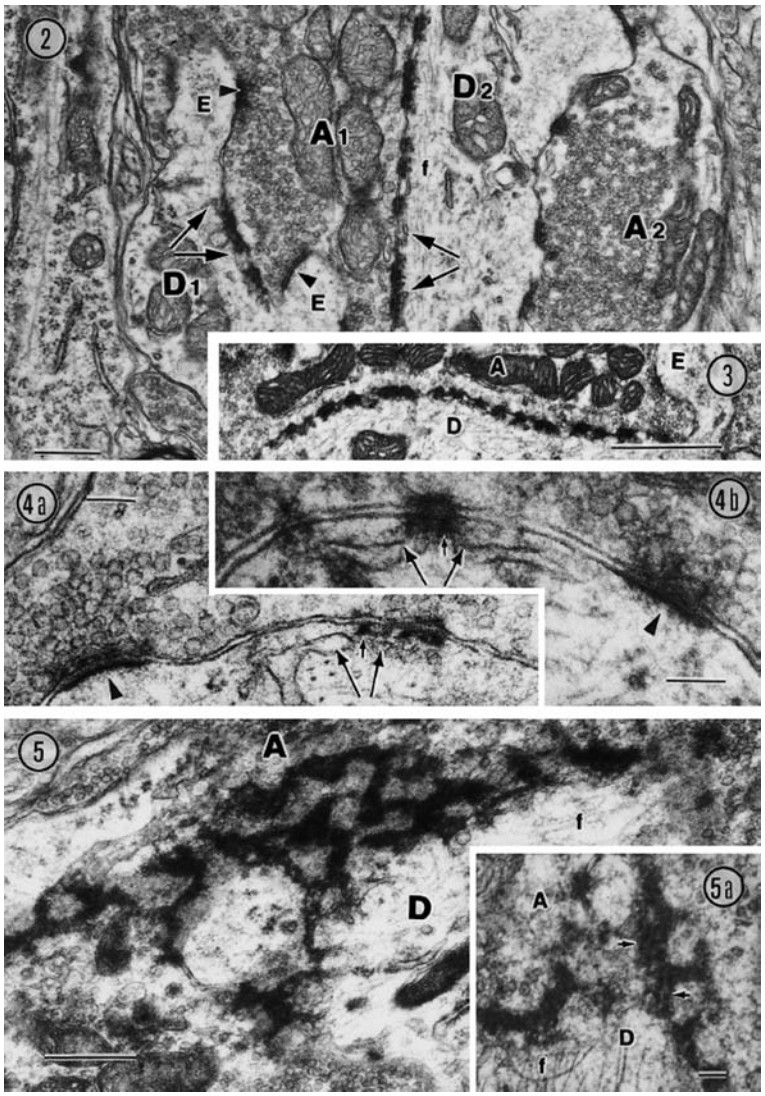

Figure 6 Micrograph 2: Part of a synaptic glomerulus in rat dorsal lateral geniculate. Two dendrites (D_1, D_2) are associated with two large axon terminal profiles (A_1, which is almost certainly a retinal afferent terminal, and A_2), and both dendrites establish asymmetrical filamentous contacts with A_1. Note that the excrescences (E) invaginating A_1 from D_1 make synaptic contacts (arrowheads) rather than filamentous contacts, and that neurofilaments (f) and agranular reticulum (arrows) are concentrated below the contact zones in both dendrites. Scale bar = 0.5 μm, ×24 000. Micrograph 3: extensive asymmetrical filamentous contact between a large presumptive lemniscal afferent axon terminal (A) and the primary dendrite (D) of a projection neuron in ventrobasal thalamus. A postsynaptic excrescence (E) protrudes from the dendrite. Scale bar = 1.0 μm, × 20 000. Micrographs 4a and 4b: differences between axodendritic filamentous and synaptic contacts. Note the clustering of vesicles and the continuous postsynaptic densities at the synaptic contacts (arrowheads) and the dendritic agranular reticulum (arrows) and neurofilaments associated with the filamentous contacts; 4a: ×80 000, 4b: Scale bar = 0.1 μm, × 120 000. Micrographs 5 and 5a: Tangentially sectioned asymmetrical filamentous contacts between lemniscal axon terminals (A) and projection cell dendrites (D) in ventrobasal thalamus. 5: Scale bar = 0.5 μm, ×35 000; 5a: Scale bar = 0.1 μm, × 50 000. Reprinted from Lieberman AR and Spacek J (1997) Filamentous contacts: The ultrastructure and three-dimensional organization of specialized non-synaptic interneuronal appositions in thalamic relay nuclei. *Cell and Tissue Research* 288: 43–57, with permission from Springer-Verlag GmbH.

SR boutons. Not known is whether there is any systematic difference in connectivity, depending on the type of relay neuron. TC relay neurons are all excitatory and glutamatergic, but they are further divisible, on the basis of dendritic morphology and depending on whether they contain parvalbumin or calbindin.

Examples of postsynaptic targets in five nuclei In the LGN, CT terminations are only type 1 (small), and type 2 (large) boutons are from the retina, but the synaptic topology is more complex in the auditory and somatosensory systems, each system consisting of several subnuclei. In association nuclei,

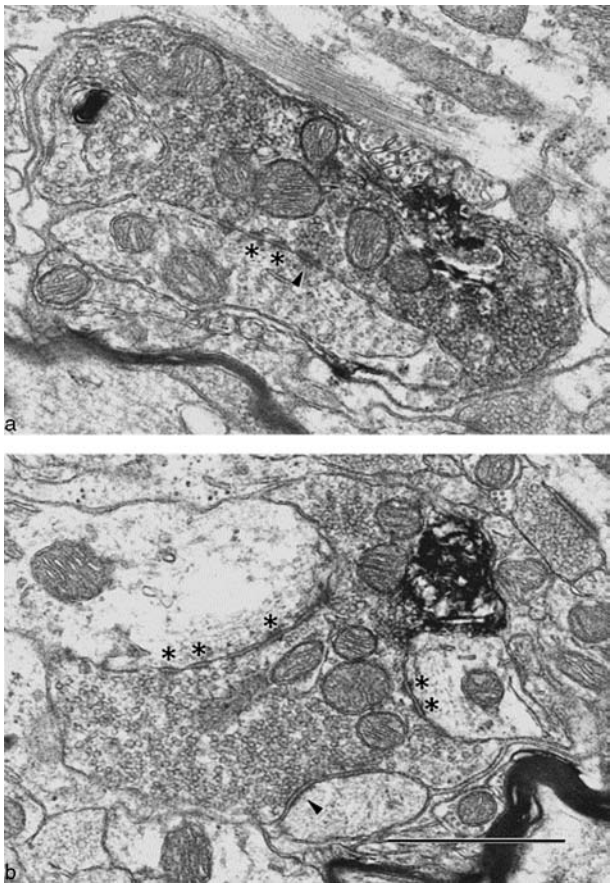

Figure 7 Electron micrographs from the rat mediodorsal thalamus. Large round axon terminals with reaction product after injections of wheat germ agglutinin-conjugated horseradish peroxidase into the basolateral amygdaloid nucleus. (a) A labeled large round terminal with round synaptic vesicles, an asymmetrical synaptic contact (arrowhead), and puncta adherentia (asterisks) on a dendritic shaft. Note the presence of glial enwrapment. (b) A labeled bouton that establishes an asymmetrical synaptic contact (arrowhead) on a dendritic spine. Puncta adherentia can also be seen (asterisks). Scale bar = 1 μm. Reprinted from Kuroda K, Yokofujita J, and Murakami K (1998) An ultrastructural study of the neural circuit between the prefrontal cortex and the mediodorsal nucleus of the thalamus. *Progress in Neurobiology* 54: 417–458, with permission from Elsevier.

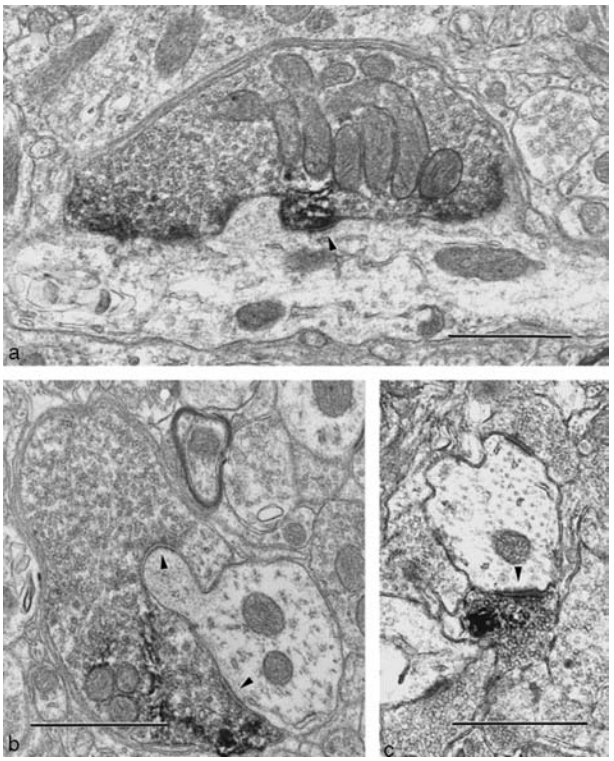

Figure 8 Electron micrographs from the rat mediodorsal thalamus. Examples of labeled boutons following injections of wheat germ agglutinin-conjugated horseradish peroxidase into the piriform cortex. (a) A labeled large round axon terminal with numerous mitochondria and asymmetric synaptic contacts (arrowhead) on a longitudinally sectioned dendrite. (b) A labeled large round bouton synapsing with a dendritic shaft and spine. Note the round synaptic vesicles, the asymmetrical synaptic membrane specialization (arrowheads), and the glial ensheathment. The postsynaptic thickening is less pronounced than that seen in small round terminals. (c) A wheat germ agglutinin-conjugated horseradish peroxidase-labeled bouton identified as a small round terminal on the basis of the size and the absence of glial processes. All scale bars = 1 μm. Reprinted from Kuroda K, Yokofujita J, and Murakami K (1998) An ultrastructural study of the neural circuit between the prefrontal cortex and the mediodorsal nucleus of the thalamus. *Progress in Neurobiology* 54: 417–458, with permission from Elsevier.

CT terminations are both type 1 and type 2, but these can originate from different areas in different combinations. Type 2 boutons are also used by amygdala neurons, which project to the mediodorsal (MD) thalamus (Figure 7), and type 1 boutons can be from the brain stem, as is the case in the LGN. Thus, positive identification of CT terminations requires prelabeling with cortical injections of anterograde tracers. Cortical injections are also necessary to define the exact source area. This following five examples describe synaptic topology (different nuclear and species patterns are more extensively discussed elsewhere; see section 'Further Reading'):

Example 1: Cat ventral posterior (primary somatosensory) thalamic nucleus. The distribution of four types of synapses has been reported in detail on four physiologically identified relay neurons. In particular, this work provided an estimate of the total number of synapses per cell, calculated as 5584–8797. Of the total, LR synapses constituted approximately 15%; SR, approximately 50%; F ('flat vesicle,' inhibitory synapses), approximately 35%; and presynaptic dendrites, 35%. For all four cells, LR (from principal afferents) and F synapses were concentrated on proximal dendrites, and SR (from corticothalamic) were on distal dendrites, with variable proportions on intermediate dendrites. Some LR synapses occurred

distally as well, but sharply fewer than SR. Comparable studies have been carried out in the LGN and medial geniculate body (MGB).

Example 2: Cat pulvinar nucleus. Several EM studies have demonstrated that the cat pulvinar nucleus, as in the monkey, receives a mix of type 1 and type 2 CT terminations. The synaptic distribution in the neuropil was determined as 3.5% LR profiles, 84.2% SR profiles, 7.3% F profiles, and 5% presynaptic dendrites. For TC cells (identified on the basis of dendritic features, and being GABA negative), LR profiles were calculated to comprise 2.1% of the synapses and SR boutons were calculated to comprise 85.6%. For interneurons (identified as GABA positive), LR boutons represented 11.8% of the synaptic contacts, and SR boutons represented 75.6%. As compared with the cat LGN, the pulvinar nucleus received fewer LR boutons and significantly more SR boutons; the ratio of LR to SR synapses was given as 1:5 in the LGN (where LR are retinogeniculate and SR are at least partly corticogeniculate), and 1:24 for the pulvinar (where both LR and SR are CT). It is important to keep in mind that these are numbers in aggregate, and it is not known whether proportions differ significantly in relation to specific relay cells. When CT projections to the cat pulvinar were prelabeled by anterograde tracers in areas 5 and 7, LR terminations were found to target approximately equal numbers of GABA-positive and GABA-negative profiles, while SR synapses were mainly on small-caliber, GABA-negative dendrites, presumed to be distal dendrites of TC projection neurons.

Example 3: Cat lateral posterior (LP) nucleus. In contrast with the mixed type 1 and type 2 terminations in the pulvinar, terminations from area 17 (or area V1) are mainly or exclusively of type 2, from layer 5 neurons. These target large dendrites, presumably proximal dendrites of TC projection neurons, and, as well, presynaptic dendrites of interneurons within glomerular arrangements. SR terminations in the LP nucleus are not labeled from injections in area 17, but originate from other cortical areas.

Example 4: Inferior pulvinar nucleus in monkey. This subdivisions receives mainly type 2 terminations from area 17 (area V1), but receives a mix of type 1 and type 2 terminations from extrastriate areas.

Example 5: Mediodorsal nucleus. In the rat, CT terminations from prefrontal cortex are reported to be only type 1. LR terminations occur, but originate from neurons in the amygdala or piriform cortex (Figures 7 and 8). The organization of the MD nucleus in monkey is very different, and in this species, injections in frontal cortex label both type 1 and type 2 boutons. Type 2 terminations occur in glomeruli, where they target both dendrites of TC projection neurons and presynaptic dendrites of inhibitory interneurons. Projections from the amygdala are also large, and similar to type 2 CT projections.

RTN output The reticular nucleus of the thalamus is a key structure that participates in a three-way connectivity, receiving collaterals from TC neurons and from type 1 CT terminations. The RTN sends inhibitory connections to the thalamus. In the cat ventroposterior nucleus (Figure 9), tracer-labeled terminations from RTN are reported to target predominantly dendrites of relay neurons (82%), but also dendrites of interneurons (8.5%), and somata (9.3%). In the monkey, tracer injections in the anterior RTN produced terminations in four subdivisions of the motor thalamus. Synaptic contacts were mainly on distal dendrites of both projection and local circuit

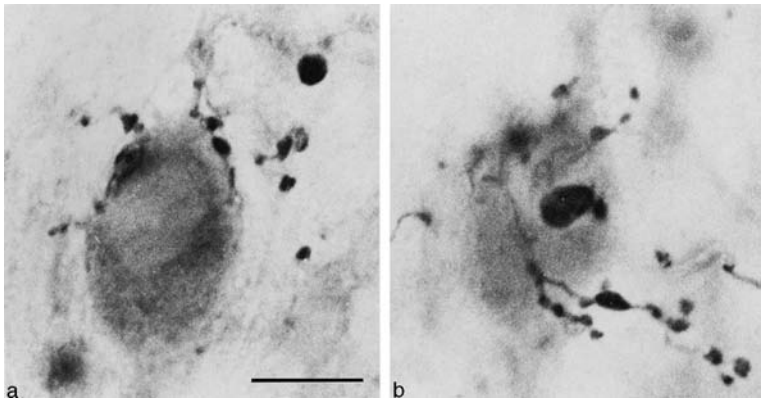

Figure 9 (a, b) Views in two planes of focus of phytohemagglutinin-L-labeled terminal axons from reticular thalamic nuclei and their boutons, ending in the vicinity of a cell body in the ventroposterior lateral nucleus of cat. Cell body is identified by nonspecific staining. Scale bar = 12.5 μm. Reprinted from Liu XB, Warren RA, and Jones EG (1995) Synaptic distribution of afferents from reticular nucleus in ventroposterior nucleus of cat thalamus. *Journal of Comparative Neurology* 352: 187–202, Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

neurons. The proportion of synapses on local circuit neurons was noted to be higher (48%, 62%, and 65%, respectively) than what has been reported for sensory thalamic nuclei, although in range with previous reports of RTN terminations in the MD and the anterior ventral nuclei.

RTN cortical inputs For reticular neurons, the CT terminations are more numerous than are TC terminations (70% vs. 20–25% for thalamic synapses), and are evenly distributed over both proximal and distal dendritic locations. The TC synapses are larger, more proximally located, and have multiple vesicle release sites. CT terminations are associated with a higher number of glutamate receptors(GluR4) in reticular neurons than in the thalamus, and this has been proposed as a basis for differences in synaptic strength (Figure 10).

Convergence of Multiple Afferents

The quantification of specific inputs onto identified postsynaptic neurons has not yet been achieved, and it is not known how many thalamic neurons are contacted by how many different type 1 and type 2 CT axons. A representative type 2 axon might have 200 boutons, which are likely to contact portions of at least 10–20 neurons, lying relatively close together. At the light microscopic level, instances can sometimes be found of single neurons that are heavily entwined by numerous type 2 boutons, possibly from one arbor (Figures 4 and 11), and a serial reconstruction of large CT terminals in the mouse posterior nuclear group demonstrated that these encapsulate the postsynaptic dendrite, often near proximal branch points. Type 1 axons typically have more boutons (up to 2000) and extend over much wider territories. Thus, they clearly contact more postsynaptic neurons, possibly with only a few boutons (1–5?) per cell. Given the low ratio of LR to SR boutons, there are likely to be at least two distinct input patterns: cells with a mix of type 1 and type 2 inputs, and those with only type 1 inputs. These latter would derive from multiple type 1 axons, possibly even from different cortical areas. These questions need to be further addressed by continued EM analysis at the level of identified cells.

Finally, it is important to note that thalamic nuclei receive a variety of excitatory inputs. The LGN receives primary sensory input from the retina, CT

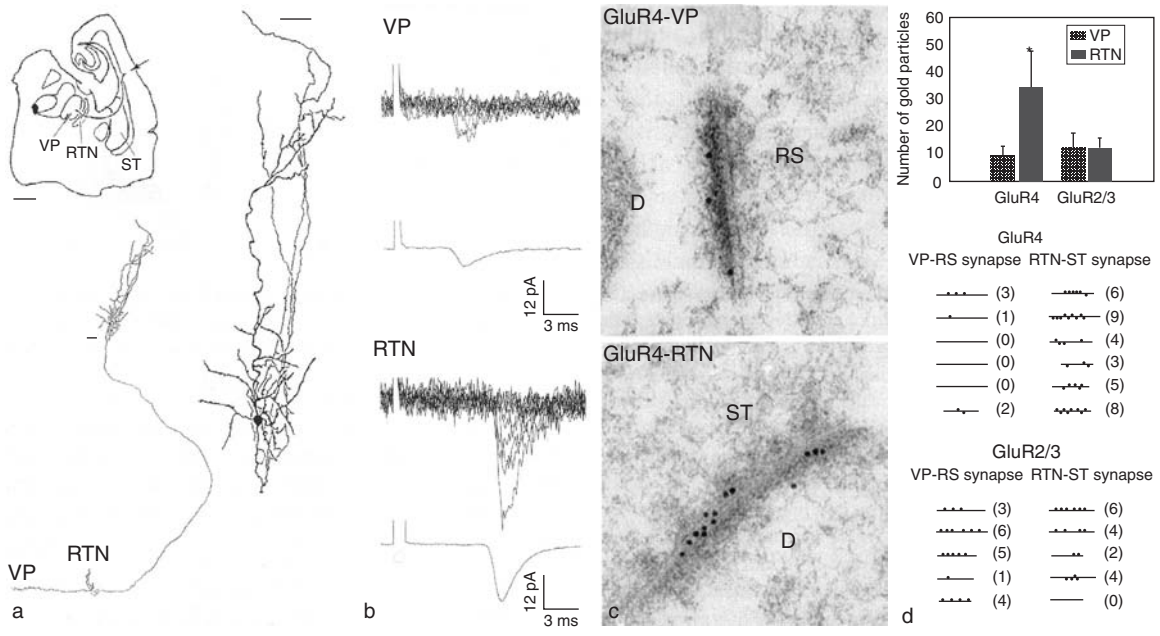

Figure 10 (a) Location and camera lucida drawing of a typical layer 6 corticothalamic cell from a mouse thalamocortical slice preparation. The axon branches to innervate both the reticular nucleus (RTN) and the ventral posterior (VP) nucleus (ST, corticothalamic synapses in the RTN). (b) Ten superimposed consecutive traces and their mean, showing excitatory postsynaptic currents recorded from a VP cell and an RTN cell in response to minimal stimulation of corticothalamic fibers. Minimal excitatory postsynaptic currents in RTN cells have approximately three times greater amplitudes than do those in VP cells. (c) Immunogold labeling for glutamate receptor (GluR4) subunits at corticothalamic synapses in the VP (RS synapse) and RTN (ST synapse), showing a larger number of particles in the RTN (D, dendrite). (d) Quantification of immunogold particles representing GluR4 and GluR2/3 receptor subunits at postsynaptic densities of four serially sectioned corticothalamic synapses in the VP and RTN. The number of GluR4 particles at corticothalamic synapses on RTN cells is approximately three times greater than at corticothalamic synapses on VP cells. Reprinted from Jones EG (2002) Thalamic organization and function after Cajal. *Progress in Brain Research* 136: 333–357, with permission from Elsevier.

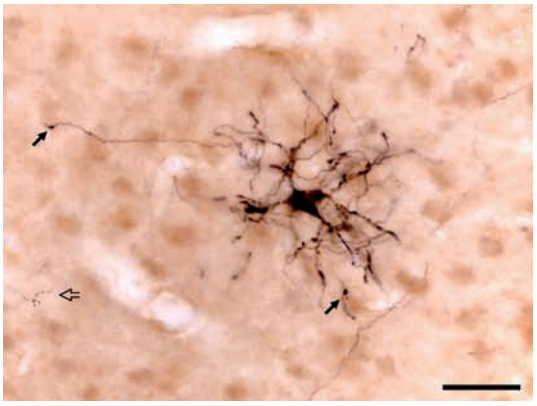

Figure 11 Light micrograph of medial pulvinar (macaque monkey). An injection of biotinylated dextran amine in parietal cortex shows type 2 corticothalamic terminations wrapping proximal dendrites of a retrogradely labeled thalamocortical neuron. Although the arbor is spatially delimited, individual boutons appear to target multiple neurons (solid arrows), and one divergent branch is visible at left. Hollow arrow indicates type 1 terminations in the same field. Scale bar = 50 μm.

type 1 input from primary visual cortex and some other visual cortical areas, and cholinergic input from the brain stem. The retina input is numerically smallest, although it is known to have an essential (or 'driving') role. The medial pulvinar nucleus, a representative associational nucleus in the monkey, receives inputs from several cortical areas and, in selective patches, from the amygdala and superior colliculus. The convergence and interaction of these inputs, especially at the level of single cells, is only partially understood.

Reciprocity of CT and TC Connections

The reciprocal nature of the connections between thalamus and cortex has been well documented and is often considered a major characteristic of this network. The concept of reciprocity, however, is actually complicated. First, other connections play important roles. In the thalamus, these include the amygdala, brain stem nuclei, and, importantly, RTN. The bidirectional circuitry involving cortex, thalamus, and RTN may be better described as a larger, three-way connectivity loop. Second, connections between cortex and thalamus are not two-way, but occur as a fanout, where projections are among multiple areas and nuclei. Not all of these are directly reciprocal. Third, very little precise information is available at the cell-to-cell level. Do layer 5 neurons, which synapse on a small number of TC neurons, receive connections back from these same neurons, either directly or through an interposed synapse in layer 4? Similarly, concerning layer 6 neurons. But for these it seems likely that only some of the widely separated TC neurons contacted by a layer 6 arbor would actually project back to that same layer 6 neuron.

TC terminations from MD to prefrontal cortex in rats have been demonstrated to target multiple populations. These include presumptive CT pyramidal neurons in the deeper layers, but also neurons in layer 3, which do not project to the thalamus. This and other results clearly show that while some monosynaptic reciprocal loops may exist, these are only one component of rich connectivity networks

CT Terminations: Functional Roles

There has been much discussion of the function of CT terminations. One of the better understood roles has been the excitatory influence of cortical terminations (type 1) on RTN neurons, which in turn produce prolonged inhibitory postsynaptic potentials in TC neurons. This circuitry is crucial in the generation of sleep rhythms. More generally, large-scale interactions between cortex and thalamus have been proposed to recruit large constellations of neurons in synchronous activity appropriate to cognitive phenomena such as perceptual binding and directed attention.

As mentioned earlier in this article, the strong dichotomies of type 1 and type 2 CT terminations have been interpreted in the general framework of modulatory and driving influences. In this view, type 2 CT terminations are compared with primary sensory inputs and are proposed to have a similar essential role (driving, or feedforward) in receptive field properties, whereas the small bouton type 1 terminations would be more modulatory (feedback). When multiple areas are known to use different combinations of type 1 and type 2 boutons, this is proposed to reflect differential feedback or feedforward effects. This hypothesis has been very influential in emphasizing the importance and richness of CT terminations, but it has tended to overlook issues such as subtypes of CT terminations and of their postsynaptic TC neurons, convergence of multiple inputs, and the marked spatial difference in arbors, where type 1 CT arbors are widely divergent within a nucleus, in contrast with more spatially delimited type 2 arbors.

Another proposal is based on the differential collateral pattern of type 1 and type 2 CT terminations, and in particular, the fact that type 2 terminations typically branch to output centers of the brain stem or spinal cord. The thalamic terminations can thus be seen as transmitting information about both the environment and the body and concurrent instructions to motor centers.

Understanding the role of CT terminations, the anatomical substrates, and the network connectivity

is still at an early stage. As noted by Guillery, further work at the level of single cells will be important to resolve finer aspects of microcircuitry.

See also: Corticothalamic Connections: Structure and Function; Presynaptic Receptor Signaling; Presynaptic Development: Functional and Morphological Organization; Thalamus and Oculomotor Control; Thalamus: Evolution in Vertebrates; Ultrastructural Organization of Release Sites in the Calyx of Held.

Further Reading

Crick F and Koch C (1998) Constraints on cortical and thalamic projections: The no-strong-loops hypothesis. *Nature* 391: 245–250.

Deschenes M, Veinante P, and Zhang Z-W (1998) The organization of corticothalamic projections: Reciprocity versus parity. *Brain Research Reviews* 28: 286–308.

Feig S and Harting JK (1998) Corticocortical communication via the thalamus: Ultrastructural studies of corticothalamic projections from area 17 to the lateral posterior nucleus of the cat and inferior pulvinar nucleus of the owl monkey. *The Journal of Comparative Neurology* 395: 281–295.

Guillery RW (2005) Anatomical pathways that link perception and action. *Progress in Brain Research* 149: 235–256.

Jones EG (2007) *The Thalamus*. Cambridge, UK: Cambridge University Press.

Jones EG (2002) Thalamic organization and function after Cajal. *Progress in Brain Research* 136: 333–357.

Jones EG (2002) Thalamic circuitry and thalamocortical synchrony. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1659–1673.

Kultas-Ilinsky K, Sivan-Loukianova E, and Ilinsky IA (2003) Reevaluation of the primary motor cortex connections with the thalamus in primates. *The Journal of Comparative Neurology* 457: 133–158.

Kuroda M, Yokofujita J, and Murakami K (1998) An ultrastructural study of the neural circuit between the prefrontal cortex and the mediodorsal nucleus of the thalamus. *Progress in Neurobiology* 54: 417–458.

Lieberman AR and Spacek J (1997) Filamentous contacts: The ultrastructure and three-dimensional organization of specialized non-synaptic interneuronal appositions in thalamic relay nuclei. *Cell & Tissue Research* 288: 43–57.

Liu XB, Warren RA, and Jones EG (1995) Synaptic distribution of afferents from reticular nucleus in ventroposterior nucleus of cat thalamus. *Journal of Comparative Neurology* 352: 187–202.

Rockland KS (1996) Two types of corticopulvinar terminations: Round (type 2) and elongate (type 1). *The Journal of Comparative Neurology* 368: 57–87.

Rouiller EM and Welker E (2000) A comparative analysis of the morphology of corticothalamic projections in mammals. *Brain Research Bulletin* 53: 727–741.

Sherman SM and Guillery RW (2006) *Exploring the Thalamus and Its Role in Cortical Function*. Cambridge: MIT Press.

Steriade M, Jones EG, and McCormick DA (1997) *Thalamus* vol. 1. Oxford: Elsevier.

Taktakishvili O, Sivan-Loukianova E, Kultas-Ilinsky K, et al. (2002) Posterior parietal cortex projections to the ventral lateral and some association thalamic nuclei in *Macaca mulatta*. *Brain Research Bulletin* 59: 135–150, with permission from Elsevier.

Wang S, Eisenback MA, and Bickford ME (2002) Relative distribution of synapses in the pulvinar nucleus of the cat: Implications regarding the "driver/modulator" theory of thalamic function. *The Journal of Comparative Neurology* 454: 482–494.

Corticotropin-Releasing Hormone and Urocortins: Binding Proteins and Receptors

M O Huising, The Salk Institute for Biological Studies, La Jolla, CA, USA and Department of Animal Physiology, Radboud University Nijmegen, Nijmegen, The Netherlands
W W Vale, The Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Published by Elsevier Ltd.

Discovery of CRF and Its Role in the Stress Response

Corticotropin-releasing factor (CRF) was identified in 1981 as the key hypothalamic factor to induce the release of adrenocorticotropic hormone (ACTH) from the anterior pituitary gland. This feat was accomplished by the purification of the peptide from sheep hypothalami. CRF is the most potent known stimulator of ACTH release, although other hypothalamic factors, most notably arginine vasopressin (AVP) and oxytocin, can synergize with CRF to augment the release of ACTH. Nevertheless, the observation that the release of ACTH from the anterior pituitary gland is almost completely blocked by administration of CRF antisera established that CRF is the predominant regulator of pituitary ACTH release. Mature bioactive CRF is a 41-amino-acid neuropeptide that is cleaved from a larger precursor protein lacking intrinsic biological activity (Figure 1). CRF is produced in the parvocellular neurons of the paraventricular nucleus (PVN) of the hypothalamus. These neurons project to the median eminence and release their neuropeptide content into the capillary bed of the hypophysial portal system. This portal system carries CRF to the anterior pituitary corticotropes, which respond by releasing ACTH into the general circulation to induce the release of glucocorticoids from the adrenal cortex. This endocrine axis, referred to as the hypothalamic–pituitary–adrenal (HPA) axis, controls reactions to stress. The stress response is an adaptive response aimed at maintaining or restoring homeostasis in the face of a plethora of potentially adverse stimuli from diverse origins. The glucocorticoids – cortisol in humans and corticosterone in rodents – that are released as a result of HPA axis activation are responsible for many of the hallmarks of the stress response, such as increased energy mobilization via enhanced liver glycogenolysis and the general inhibition of vegetative functions, including digestion, growth, reproduction, and immunity. Importantly, glucocorticoids engage in negative feedback at the level of both the pituitary gland and the hypothalamus to inhibit the synthesis and release of ACTH and CRF, aimed at a timely termination of the stress response. Many of the behavioral effects that are associated with stress, such as increased anxiety, improved cognition, increased depression, and reduced feeding behavior, are often the direct effect of the activation of central CRF circuits.

CRF: Founding Member of a Small Family of Neuropeptides

Shortly after the discovery of CRF in sheep, the CRF genes of humans, mice, rats, and other mammals were identified. Urotensin-I (UI), a peptide resembling CRF, was isolated from a fish species (*Catastomus commersonii*) following the discovery of CRF in mammals. This peptide was initially regarded as the fish ortholog (i.e., the same gene in a different species) of mammalian CRF, until a second peptide also resembling CRF was identified from the same fish species. As this second CRF-like peptide shared much higher similarity with mammalian CRF, it was clear that it represented the actual ortholog of mammalian CRF. The mammalian ortholog of fish UI was identified in 1995 with the isolation and characterization of urocortin 1 (Ucn1). Another peptide with resemblance to the Ucn1 of mammals and the UI of fishes was isolated from the skin of the waxy monkey tree frog (*Phylomedusa sauvagei*), a species from the Amazonian rain forest. This peptide was named sauvagine and although it is only known from the skin of a single amphibian species, sauvagine has been frequently used in the characterization of many aspects of CRF signaling. In recent years, two additional members of the CRF family of neuropeptides were independently discovered by two groups employing a bioinformatics approach. The current nomenclature for these peptides is urocortin 2 (Ucn2) and urocortin 3 (Ucn3), although one of the identifying groups initially named them stresscopin-related peptide and stresscopin, respectively. All members of the vertebrate CRF family have several characteristics in common with CRF. All are preprohormones that contain at their C-termini mature neuropeptides that are presumed to be liberated by cleavage at a mono- or dibasic cleavage site (Figure 1). In keeping with their discovery via bioinformatics, the existence of mature neuropeptides for Ucn2 and Ucn3 has been predicted, and their bioactivity following *de novo* synthesis has been established, but neither mature peptide has been directly isolated from an endogenous source.

CRF Family Neuropeptides Signal via One of Two CRF Receptors

Two receptors for CRF have been identified and are termed CRFR1 and CRFR2. Both are class B

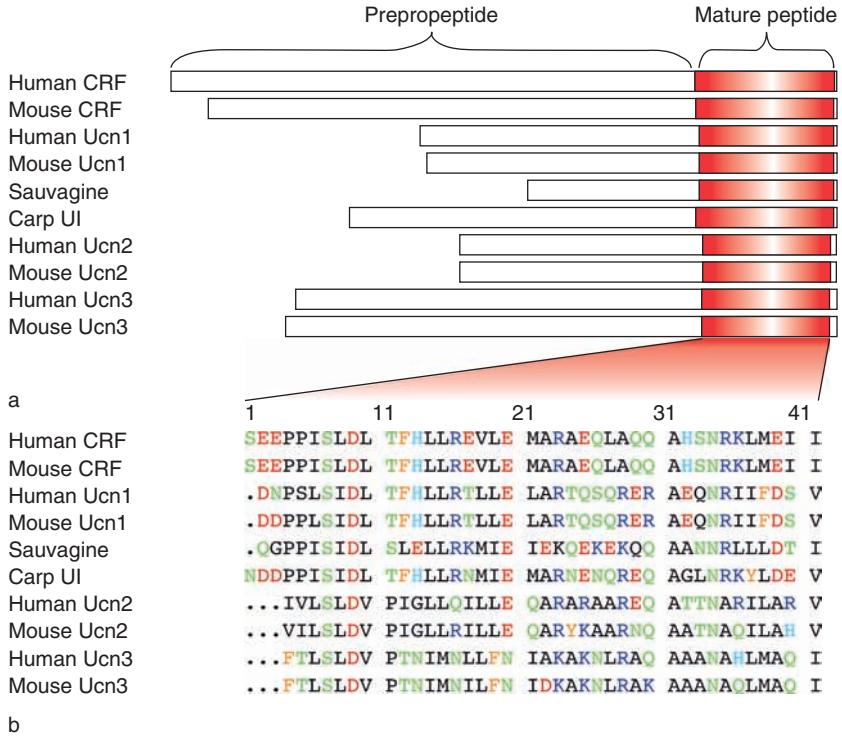

Figure 1 Corticotropin-releasing factor (CRF) family ligands (Ucn1, Ucn2, and Ucn3; urocortins) are encoded as preprohormones. (a) The C-terminal mature neuropeptides are liberated from the prepropeptide sequence by cleavage at mono- or dibasic cleavage sites. At the amino acid level CRF and urocortins share moderate amino acid similarity. (b) Amino acids with similar biochemical properties are color-coded for clarity.

(secretin-like) seven-helix transmembrane G-protein-coupled receptors and consist of 395 to 440 amino acids. CRFR1 and CRFR2 share approximately 70% amino acid identity, with most of their amino acid differences concentrated in the N-terminal region. As the N-terminal domain (together with the extracellular loops) is responsible for ligand binding, this accounts for the substantial differences in the pharmacology of the receptors. CRFR1 is the principal receptor activating pituitary corticotropes during the stress response and has nanomolar affinity for CRF. CRFR1 binds Ucn1 with a similarly high affinity, but has no appreciable affinity for physiological doses of Ucn2 and Ucn3. The CRFR2 receptor has highest affinity for Ucn1 and also binds Ucn2 and Ucn3 with high affinity. The affinity of CRFR2 for CRF is about 10- to 30-fold lower than its affinity for the urocortins, but is still below 100 nM. This illustrates the redundancy of the CRF system, wherein multiple ligands share, and under certain circumstances may even compete for, two different receptors (Figure 2).

A further layer of complexity is added to CRF signaling by the observation that multiple transcripts can be generated from both CRFR genes. The CRFR1 gene (*CRHR1*, humans; *Crhr1*, rodents) consists of 14 exons in humans and 13 exons in rodents. More than one dozen alternative transcripts have been found that result from the skipping of one or several exons during gene transcription. The predicted proteins from some of these splice variants partially or completely lack the extracellular domain (ECD), or lack one or several transmembrane domains. Nevertheless, the mechanism responsible for these CRFR1 splice isoforms is unknown and the physiological roles played by CRFR1 splice variants remain presently unclear; direct evidence for the presence of corresponding CRFR1 proteins is lacking for most of these splice isoforms.

A clear example of how alternative splicing can generate functionally different receptor proteins is found in the mammalian gene encoding CRFR2. Two different receptor isoforms, CRFR2α and CRFR2β, are transcribed from the *CRFR2* gene by splicing alternative start exons onto a common set of downstream exons. A third isoform, CRFR2γ, is produced in a similar fashion in primates only. The resulting receptor proteins differ only at the extreme N-terminus of their ECD. As the majority of the ECD is encoded by exons four, five, and six, which are shared by all CRFR2 isoforms, their pharmacological properties do not differ substantially. Nevertheless, the different CRFR2 isoforms vary markedly in their central and

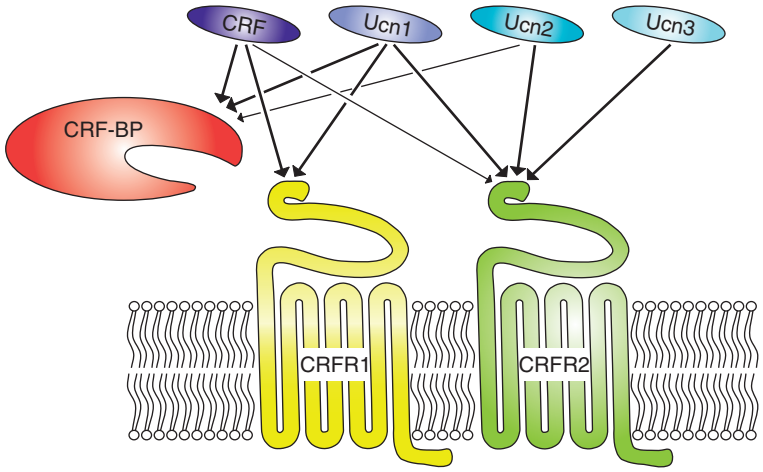

Figure 2 Corticotropin-releasing factor (CRF) and the urocortins (Ucn1, Ucn2, and Ucn3) signal through one of two CRF receptors (CRFR1 and CRFR2). Both CRF receptors differ in the repertoire of ligands that they bind. Furthermore, CRF and Ucn1, and to a lesser extent Ucn2, bind to CRF-binding protein (CRF-BP). The arrows pointing away from each ligand indicate the ligand potential targets. The thickness of the arrows indicates the approximate affinity for each of the ligands.

peripheral tissue distribution, as the transcription of each isoform is driven by its own unique promoter sequence. In rodents, the expression of CRFR2α is largely restricted to the central nervous system (CNS), with expression in the olfactory bulb, ventromedial hypothalamus (VMH), the lateral septum (LS), and the bed nucleus of the stria terminalis (BNST). In contrast, the expression of CRFR2β within the CNS is restricted to the choroid plexus. In the periphery of rodents, the situation is completely reversed. Expression of CRFR2β is widespread, with highest expression levels in skeletal muscle, heart, and skin. Interestingly, CRFR2α is the predominantly expressed isoform in the human periphery. By comparison, CRFR1 is expressed in brain and pituitary as well as in peripheral sites, including adrenal gland, spleen, testis, and skin.

Recently, several other splice variants were discovered that are generated from the CRFR2 gene. One particularly interesting splice variant stems from the deletion of exon six, resulting in a premature stop codon caused by a shift in the amino acid reading frame. The resulting 143-amino-acid protein consists of the complete ECD of CRFR2α followed by a unique hydrophilic 38-amino-acid C-terminal tail. The resulting soluble CRFR2α can abrogate the response to CRF and Ucn1 *in vitro* and may serve as a natural modulator of the actions of CRF and urocortins.

A two-step model has been proposed for hormone binding and receptor activation of CRF receptors. First, the C-terminus of the ligand binds the extracellular domain of the receptor, followed by the partial insertion of its N-terminus into the transmembrane region. This results in a conformational change that leads to exchange of guanosine diphosphate (GDP) for guanosine triphosphate (GTP) on the Gα subunit of the G-protein that is coupled to the third intracellular loop of the receptor. The activated Gα subunit dissociates from the G$\beta\gamma$ dimer, and both species can activate various downstream signaling pathways. In many cells, activation of the G-protein activates the adenylyl cyclase–cyclic adenosine monophosphate (cAMP)/protein kinase A (PKA) pathway. This pathway results in downstream modulation of target proteins by PKA as well as transcriptional regulation via cAMP response element-binding protein (CREB). Depending on the cell type that expresses CRF receptors, other signaling pathways can be activated downstream of CRF receptors via different G-proteins. These pathways include extracellular signal-regulated kinase/mitogen-activated protein kinase (ERK/MAPK), protein kinase C (PKC), nuclear factor-kappa B (NF-κB), and several others.

CRF-Binding Protein

A unique player in the network of CRF signaling exists in the form of a soluble 37 kDa protein that binds CRF with high affinity. This CRF-binding protein (CRF-BP) was initially discovered in late gestational maternal plasma, where it prevents activation of the stress response via the substantial amounts of CRF from placental origin that circulate prior to parturition. The CRF-BP gene consists of seven exons that encode a protein of 322-amino-acids, unrelated to the CRF receptors. In fact, its primary amino acid sequence bears no resemblance to that of any other known protein. CRF-BP is characterized by ten conserved cysteine residues that form five consecutive disulfide bridges, and a single N-linked glycosylation site. CRF-BP orthologs have been demonstrated for a number of vertebrate species and were recently

discovered in species of insects. In mammals, CRF-BP binds CRF and Ucn1 with an affinity that is several fold higher than the affinity of CRF receptors for either peptide. Human CRF-BP also binds mouse, but not human, Ucn2 with nanomolar affinity and does not bind Ucn3 (Figure 2). Interestingly, several peptide fragments, such as CRF(6–33) and CRF(9–33), which consist only of the core of the mature peptide hormone, retain the capacity to bind to CRF-BP with high affinity, but have lost the ability to activate CRF receptors. This indicates that different areas of the peptide hormone are involved in binding to the receptor and to the binding protein.

The high affinity of CRF-BP for CRF and Ucn1 suggests an antagonistic role by preventing these hormones from activating their receptors. Indeed, CRF-BP is expressed in anterior pituitary corticotropes and is capable of abrogating the CRF-induced release of ACTH from pituitary cells *in vitro*. CRF-BP is widely expressed within the CNS, including the cerebral cortex, the amygdala, BNST, olfactory bulb, and various nuclei within the brain stem and hypothalamus. Peripherally, complexes of CRF-BP and CRF are rapidly cleared from the circulation, supporting an antagonistic role for CRF-BP.

Central CRF Pathways Are Involved in Anxiety and Depression

Although initially discovered as an activator of the HPA axis, CRF is produced at many other sites in the brain outside of the parvocellular section of the PVN. Within the hypothalamus CRF is produced in several other nuclei, including the preoptic, supraoptic, and suprachiasmatic nuclei. Prominent sites of CRF localization outside of the hypothalamus include the cortex, the central nucleus of the amygdala (CeA), the BNST, the locus coeruleus (LC), and the parabrachial nucleus (PB). These nuclei are generally associated with autonomic function. Consensus is that within the CNS, CRF acts as a neurotransmitter by activating downstream neurons.

Central CRF pathways are involved in behavioral patterns that are commonly associated with stress, such as changes in arousal and anxiety. Several rodent models have been developed for the measurement and quantification of the behavioral manifestations of arousal and anxiety. Transfer of a rodent from a familiar nonstressful environment, such as its home cage, to a completely new environment is considered an acutely stressful event for the animal. This transition is immediately reflected in a series of behavioral changes that can be measured as a reduction in locomotor activity and reduced exploration of the novel environment, a reduction in rearing, and an increase in grooming and freezing. Injections of CRF (intracerebroventricularly; directly into the ventricles of the brain) can reproduce this panel of behaviors associated with novelty stress. Mice that lack CRFR1 display a marked reduction in anxiety-related behaviors in addition to blunted HPA activation. This suggests that the majority of the anxiogenic effects of CRF are mediated via CRFR1. Data regarding the role of CRFR2 in anxiety are less clear. Both anxiogenic (anxiety-promoting) and anxiolytic (anxiety-reducing) effects have been attributed to CRFR2, and the direction and magnitude of the effects mediated via CRFR2 may be dependent on the precise anatomical location of CRFR2 expression as well as on confounding factors such as gender.

Dysregulation of CRF pathways is implicated in the etiology of depression. CRF concentrations are significantly elevated in the cerebrospinal fluid of depressed patients, and reduction of these levels by electroconvulsive therapy correlates with alleviation of depression symptoms. Several rodent models have been developed that display behavior reflecting some traits associated with depression in human patients. In one of these models a mouse is suspended by the tail and the severity of depression is quantified as the fraction of the time the animal is immobile. In another test mice are subjected to involuntary swimming and depression is measured as a reduction in swimming and climbing behavior and increased immobility. In these animal models for depression, selective CRFR1 antagonists reduce immobility, implicating central CRF pathways that employ CRFR1 in the pathophysiology of depression. Recently, mice were generated that are deficient for the selective CRFR2 agonist Ucn2. One of the major findings in these Ucn2-deficient mice is their antidepression-like phenotype, as measured by reduced immobility in the forced swim test and the tail suspension tests. These experiments demonstrate a prominent role for CRFR1 and CRFR2 in the development of depressive disorders.

Feeding, Digestion, and Metabolism

Several CRF pathways are involved in the regulation of feeding and digestion. Central circuits employing CRF regulate key aspects of appetite control and feeding behavior, while peripheral CRF receptors relay signals that affect gastrointestinal motor functions. Much of the integration of central metabolic cues with peripheral indicators of nutrient status, such as circulating levels of leptin and insulin, takes place in the hypothalamus. It has long been known that injection of CRF and Ucn1 directly into the brain ventricle produces anorexic effects. Activation of either

CRF receptor produces anorexic effects, but with different kinetics. Intracerebroventricular injection of CRFR1-specific agonists results in a rapid inhibition of food intake, while the specific activation of CRFR2 results in prolonged anorexia that requires several hours to establish. These findings are supported by studies on CRF-deficient mice: mice lacking CRFR1 are insensitive to short-onset anorexia induced by a general CRF receptor agonist such as Ucn1, while CRFR2-deficient mice fail to mount prolonged hypophagia following Ucn1 administration. The arcuate nucleus has strong CRFR1 expression while the ventromedial hypothalamus features prominent CRFR2 expression. Both nuclei play key roles in the regulation of food intake and metabolism. The VMH receives abundant projections of Ucn3, and, to a lesser extent, Ucn1, while it is only sparsely innervated by CRF. Ucn2 and Ucn3 are also detected in the arcuate nucleus. These observations suggest that many of the anorectic effects of CRF family peptides are mediated via CRFR2 and that the urocortins constitute its endogenous ligands, modulating appetite control and feeding behavior. CRFR2 expression levels in the VMH drop during fasting and are correlated with circulating leptin levels, suggesting that plasma leptin levels directly or indirectly affect the level of CRFR2 expression in this nucleus. Direct stimulation of the VMH by CRFR2 agonist injection or electrical stimulation not only results in reduced food intake, but also activates the sympathetic nervous system. Via this route energy expenditure is increased through the stimulation of brown adipose tissue (BAT) thermogenesis. Thus, feeding behavior is enhanced and BAT thermogenesis is decreased during starvation, when plasma leptin and consequently VMH CRFR2 expression are low. During times of surplus body energy stores, increased leptin concentrations and CRFR2 levels drive increased BAT thermogenesis and inhibit food intake.

CRF signaling is involved in the regulation of gastrointestinal motor function. Although the inhibition of gastromotor function that is observed during stress is partially mediated via central CRFR2, CRF and urocortins have direct effects on the gastrointestinal tract. Peripheral administration of urocortins delays gastric emptying and inhibits gut motility. These inhibitory effects on gastromotor functions are blocked by selective CRFR2, but not CRFR1, antagonists. In contrast to the inhibition of gastric motility by central and peripheral CRFR2 activation, colonic motility can be stimulated during stress via CRFR1-mediated activation of the parasympathetic nervous system as well as via CRFR1 present in the colonic myenteric nervous system. The physiological relevance of the dichotomy in the activation of different stretches of the gastrointestinal tract is explained by adaptive value in times of stress. Inhibitions of gastric and gut activity and colon evacuation are different ways of inhibiting digestion, thus contributing to the maximal allocation of resources toward responding to stress.

Cardiovascular Effects of Urocortins via CRFR2

It has long been known that CRF influences cardiovascular function. During acute stress, heart rate, blood pressure, and cardiac output increase as the result of activation of the sympathetic nervous system. This overall stimulation of the cardiovascular system is an important adaptive response, assisting an animal to actively cope with a stressful event. Central injection of CRF mimics the cardiovascular effects that are normally observed following stress, provided that the sympathetic nervous system is intact. Since coadministration of a CRFR1-specific antagonist attenuates CRF-induced increases in heart rate, the CRF-induced activation of the sympathetic nervous system is, at least in part, mediated via CRFR1.

In contrast to the generally stimulatory effects of central CRF injections on the cardiovascular system, CRF delivered peripherally causes vasodilatation and a drop in mean arterial blood pressure. Systemic CRF and Ucn1 also increase cardiac contractility. These direct effects of systemically administered CRF and Ucn1 on the cardiovascular system suggest the presence of local CRF receptors. Indeed, CRFR2 is expressed in significant amounts in cardiac tissue, including myocardium, epicardium, arterioles of atrium and ventricle, and aorta. Its role in the cardiovascular system was elegantly demonstrated by comparing the hemodynamic properties of normal mice with mice lacking CRFR2. The latter mice have an elevated mean arterial blood pressure that is insensitive to systemic Ucn1 administration, whereas normal mice respond to Ucn1 with significant and prolonged (>1 h) hypotension. Ucn1 and Ucn2 are potential endogenous ligand(s) for CRFR2 in the cardiovascular system, as both urocortins are expressed in isolated cardiac myocytes. Alternatively, it is possible that CRF family peptides reach the heart via nerve terminals or through the general circulation to activate cardiac CRFR2.

In addition to the effects on blood pressure and cardiac contractility mediated via CRFR2 *in vivo*, urocortins display cardioprotective effects. The expression of Ucn1 and Ucn2 in cultured cardiomyocytes is increased after hypoxia followed by reoxygenation,

a model for cardiac ischemia. The presence of urocortin peptides in media reduces the percentage of cultured cardiomyocytes that can enter apoptosis and necrosis following hypoxia. CRFR2 is required for these cardioprotective effects, since Ucn2 and Ucn3 fail to rescue cardiomyocytes isolated from CRFR2-deficient mice from apoptosis. Moreover, the administration of Ucn2 and Ucn3 immediately following experimentally induced ischemia of the intact rat heart results in a substantial reduction of the infarct zone. These cardioprotective effects of urocortins acting through CRFR2 are of particular interest in the context of cardiac ischemia. In particular, the fact that Ucn2 and Ucn3 lack the affinity for CRFR1 required for HPA activation or induction of anxiety makes them promising targets in the quest for therapeutical interventions aimed at alleviation of the adverse outcome of cardiac ischemia in humans.

CRF and Pregnancy

CRF and Ucn1 as well as their cognate receptors and CRF-BP are expressed in different tissues of the female reproductive system. CRF plays a remarkably versatile role affecting pregnancy, from blastocyst implantation to parturition. Mammalian reproduction, in which the fetus is carried inside the womb, is an awkward phenomenon from an immunological point of view. The developing fetus is only 50% 'self': the paternal contribution to the fetus qualifies it as a foreign body and would as such subject it to attack by the maternal immune system. For obvious reasons, this would be incompatible with successful reproduction. CRF protects the fetus from attack by the maternal immune system and simultaneously prevents the developing fetal immune system from attacking maternal tissues. Around the time of blastocyst implantation into the uterus, CRF is produced in the embryonic trophoblast and the maternal decidua, the contact interface of maternal and fetal tissues. Locally, CRF increases the expression of an important proapoptotic factor, Fas ligand (FasL), via CRFR1. FasL induces apoptosis in cells carrying its cognate binding partner, Fas, which is expressed by many immune cells. As Fas only appears at the cell surface following leukocyte activation, CRF-induced upregulation of FasL at the site of implantation allows the specific deletion of only those immune cells that are activated at the maternal–fetal interface.

CRF is expressed in the placenta of primates, but not rodents, and CRF from placental origin is detectable in the maternal circulation beginning approximately halfway through gestation. Circulating CRF levels rise sharply toward the end of pregnancy and, close to term, reach levels equivalent to peak CRF levels measured in the hypothalamic–pituitary capillary bed during acute HPA activation. Yet, these high levels of CRF do not activate the HPA axis, because they are bound to a surplus of CRF-BP, which is also produced by the placenta. In the fetal circulation, increasing levels of CRF will activate the fetal HPA axis. This results in glucocorticoid release from the fetal adrenal gland, a process that drives the maturation of a number of fetal organs, including the lungs in preparation for delivery.

Throughout much of gestation, CRF in the placenta evokes quiescence of uterine smooth muscle cells via the CRFR1-dependent activation of the PKA pathway. In a series of intricate molecular switches that involve the oxytocin-dependent phosphorylation of CRFR1, CRFR1 is coupled to a different G-protein toward parturition. This causes a shift in the intracellular cascade that is activated upon ligand binding to CRFR1 from PKA to PKC. Instead of inducing myometrial relaxation, CRF now can assist in the induction of myometrial contractility and labor.

Summary

CRF was initially noted for its potent capacity to activate the stress response. In the decades that followed its discovery, three additional ligands related to CRF, as well as two receptors and a soluble binding protein, have been discovered. While CRF still has strong ties to 'stress,' we now know that CRF signaling is involved in a multitude of other central and peripheral processes. These include the regulation of anxiety, arousal, and depression, the modulation of food intake and sympathetic nervous system activation, control of cardiovascular output as well as cardioprotection, and a combination of fascinating roles in the regulation of human reproduction. Many of these processes are at the heart of human well-being, which has sparked widespread interest in CRF. The challenge that lies ahead will be the development of strategies to intervene in isolated CRF pathways without causing widespread activation of CRF receptors throughout the body.

See also: Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Mammalian Neuropeptide Families; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Discovery; Neuropeptides: Enteric Nervous System; Neuropeptides: Food Intake; Neuropeptides: Mental Disease; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress, the HPA Axis and Depressive Illness.

Further Reading

Charmandari E, Tsigos C, and Chrousos GP (2005) Endocrinology of the stress response. *Annual Review of Physiology* 67: 259–284.

Chen A, Perrin M, Brar B, et al. (2005) Mouse corticotropin-releasing factor receptor type 2α gene: Isolation, distribution, pharmacological characterization and regulation by stress and glucocorticoids. *Molecular Endocrinology* 19: 441–458.

Coste SC, Quintos RF, and Stenzel-Poore MP (2002) Corticotropin-releasing hormone-related peptides and receptors: Emergent regulators of cardiovascular adaptations to stress. *Trends in Cardiovascular Medicine* 12: 176–182.

Heinrichs SC and Koob GF (2004) Corticotropin-releasing factor in brain: A role in activation, arousal, and affect regulation. *The Journal of Pharmacology and Experimental Therapeutics* 311: 427–440.

Hillhouse EW and Grammatopoulos DK (2002) Role of stress peptides during human pregnancy and labour. *Reproduction* 124: 323–329.

Hillhouse EW and Grammatopoulos DK (2006) The molecular mechanisms underlying the regulation of the biological activity of corticotropin-releasing hormone receptors: Implications for physiology and pathophysiology. *Endocrine Reviews* 27: 260–286.

King BM (2006) The rise, fall, and resurrection of the ventromedial hypothalamus in the regulation of feeding behavior and body weight. *Physiology & Behavior* 87: 221–244.

Lewis K, Perrin MH, Blount A, et al. (2001) Identification of urocortin III, an additional member of the corticotropin-releasing factor (CRF) family with high affinity for the CRF2 receptor. *Proceedings of the National Academy of Sciences of the United States of America* 98: 7570–7575.

McLean M and Smith R (2001) Corticotrophin-releasing hormone and human parturition. *Reproduction* 121: 493–501.

Owens MJ and Nemeroff CB (1991) Physiology and pharmacology of corticotropin-releasing factor. *Pharmacological Reviews* 43: 425–473.

Reyes TM, Lewis K, Perrin MH, et al. (2001) Urocortin II: A member of the corticotropin-releasing factor (CRF) neuropeptide family that is selectively bound by type 2 CRF receptors. *Proceedings of the National Academy of Sciences of the United States of America* 98: 2843–2848.

Westphal NJ and Seasholtz AF (2006) CRH-BP: The regulation and function of a phylogenetically conserved binding protein. *Frontiers in Bioscience* 11: 1878–1891.

Winsky-Sommerer R, Boutrel B, and de Lecea L (2005) Stress and arousal: The corticotrophin-releasing factor/hypocretin circuitry. *Molecular Neurobiology* 32: 285–294.

Zorilla EP, Taché Y, and Koob GF (2003) Nibbling at CRF receptor control of feeding and gastrocolonic motility. *Trends in Pharmacological Sciences* 24: 421–427.

Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress

P E Sawchenko, Z F Yuan, F Laplante, R A Rissman, and J C Bittencourt, Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Corticotropin-releasing factor (CRF) is a neuropeptide that was first isolated from the hypothalamus on the basis of its capacity to stimulate the release of adrenocorticotropic hormone (ACTH) from the anterior pituitary gland. Its localization to parvocellular neurosecretory neurons of the paraventricular hypothalamic nucleus (PVH) defined the central limb of the hypothalamic–pituitary–adrenal (HPA) axis. It was recognized early on, however, that the central distribution of CRF is quite broad and includes aspects of the limbic forebrain associated with affective behaviors and brain stem cell groups involved directly in central autonomic control. The subsequent demonstration that the peptide could act within the brain to elicit stresslike behavioral (increased anxiety, arousal, decreased appetite) and autonomic (sympathomimetic) responses gave rise to the hypothesis that the central CRF system may act at multiple levels of the central nervous system to activate, and even integrate, complementary modes of stress adaptation. This notion has provided a dominant unifying framework guiding much contemporary work in the field of stress neurobiology.

The idea that individual signaling molecules may serve similar or complementary functions at diverse sites of expression has been entertained for some time. While this is not a universal generalization, certain catecholamines (epinephrine, norepinephrine) that serve as principal effectors of the sympathoadrenal system in the periphery are also prominently expressed, and situationally active, in aspects of central stress-related circuitry. In addition, there are a number of 'brain–gut peptides' (including CRF) whose expression in the peripheral autonomic and/or enteric nervous systems on one hand, and in the brain on the other, is associated with defined and complementary roles in digestive functions and appetitive control, respectively. Here we consider the functional and anatomical organization of the central CRF system and evaluate evidence bearing on its hypothesized role as an arbiter of multiple modes of stress adaptation.

The CRF Family of Signaling Molecules

CRF exerts its biological activities by binding G-protein-coupled receptors that activate adenylate cyclase. The type 1 receptor (CRFR1) is expressed by anterior pituitary corticotropes and can account for the neuroendocrine actions of the peptide. CRFR1 is expressed widely in brain (Figure 1), though its distribution is more suggestive of dominant roles in cortical, cerebellar, and sensory information processing than in the stress-related behavioral and autonomic functions for which CRF is better known. A structurally related type 2 receptor (CRFR2) exists in two principal forms; one variant (CRFR2β) is expressed in the periphery and the choroid plexus, while a second (CRFR2α) is expressed in brain, where its distribution is distinct from, and more limited than, that of CRFR1. The binding affinity of CRF for CRFR2 is relatively low. This information, coupled with the recognition that certain nonmammalian vertebrates express not only obvious CRF homologs but also other structurally related peptides such as sauvagine (from frog) and urotensin I (from fish) prompted an effort to identify new mammalian CRF family members, culminating in the cloning of three novel peptides, termed urocortins (Ucns 1–3). Ucn 1 binds both CRFRs with high affinity, while Ucn 2 and 3 are CRFR2-selective ligands. Each of the Ucns exhibits a distinctive and discrete expression pattern in brain (Figure 1), with variable potentials for interacting with cell groups expressing CRFR2 (e.g., Ucn 3) or both receptors (Ucn 1). There is a strong basis for considering these components of the central CRF system as functioning in a unified manner in stress adaptations. Shortly after the identification of Ucn 1, evidence was provided that this peptide can act centrally to mediate some stress-related functions originally attributed to CRF. Moreover, a substantial body of evidence supports an interaction of CRFR1- and CRFR2-dependent mechanisms in sculpting a variety of neuroendocrine, autonomic, and behavioral responses to stress.

In keeping with their binding affinities, a strong general correspondence has been found between the distribution of neurons that are activated in response to central administration of CRF and sites of CRFR1 expression, while central Ucn 1 activates both CRFR1- and CRFR2-bearing targets. The principal exceptions are robust activational responses provoked by either peptide in a set of interconnected cell groups that

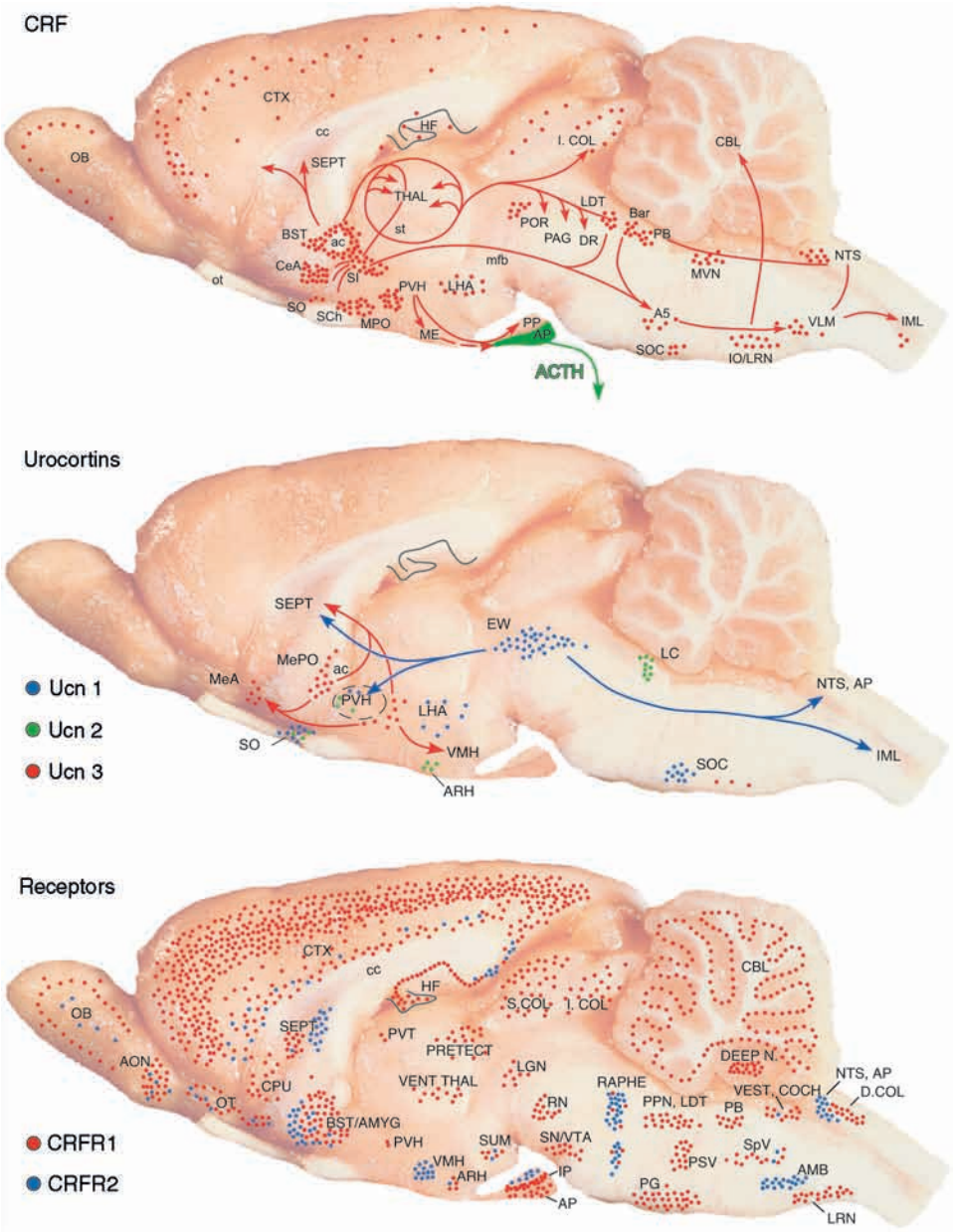

Figure 1 Central distribution of corticotropin-releasing factor (CRF) signaling molecules. Schematic representation on a sagittal view of the rat brain of major sites of cellular expression of CRF-related peptides and receptors; major axonal projections of peptidergic neurons are indicated by arrows. In addition to being expressed by neurosecretory neurons of the paraventricular hypothalamic nucleus (PVH) that comprise the final common path for effecting stress-induced activation of adrenocorticotrophic hormone (ACTH) secretion, CRF (top) is broadly expressed in brain, including in limbic forebrain (BST, CeA) and central autonomic (NTS, VLM, PB, IML) cell groups. The expression of urocortins (Ucn 1–3; middle) is more restricted. The dominant seat of Ucn 1 expression is the Edinger–Westphal nucleus (EW), whose outputs include projections to central autonomic cell groups (PVH and NTS). Ucn 2 is weakly expressed in brain, and axonal projections immunoreactive for this peptide have not been demonstrated. Ucn 3 is expressed in a few discrete forebrain regions that project to some major sites of type 2 CRF receptor expression (medial amygdala, lateral septum and ventromedial hypothalamus). The two known CRF receptor subtypes (CRFR1 and CRFR2, bottom) are differentially distributed, with CRFR1 expression being more widespread. Receptor expression in demonstrated sites of peptide action in eliciting stresslike physiological and behavioral responses is limited mainly to the parabrachial nucleus (CRFR1) and NTS (CRFR2). A5, A5 catecholamine cell group; ac, anterior commissure; AMB, nucleus ambiguus; AMYG, amygdala; AON, anterior olfactory nucleus; AP, area postrema; ARH, arcuate nucleus hypothalamus; Bar, Barrington's nucleus; BST, bed nucleus stria terminalis; CBL, cerebellum; cc, corpus callosum; CeA, central nucleus amygdala; COCH, cochlear nucleus; CPU, caudoputamen; CTX, cerebral cortex; D.COL., dorsal column nuclei; DEEP N., deep nucleus cerebellum; DR, dorsal raphe nucleus; EW, Edinger–Westphal nucleus; HF, hippocampal formation; I. COL, inferior colliculus; IML, intermediolateral column spinal cord; IO, inferior olivary complex; LC, locus coeruleus; LDT, laterodorsal tegmental nucleus; LGN, lateral geniculate nucleus; LHA, lateral hypothalamic area; LRN, lateral reticular nucleus; ME, median eminence; MeA, medial nucleus amygdala; MePO,

comprise a core circuitry for central autonomic and neuroendocrine regulation – responses not readily explicable on the basis of CRFR distributions. These cell groups, collectively termed the central autonomic system (CAS), are of critical importance as most are implicated as principal sites of CRF action in eliciting stress-related responses but have been found to display limited capacity for CRFR expression or CRF binding. These include the central nucleus of the amygdala (CeA), which has been identified repeatedly as a site of CRF action relevant to behavioral (anxiety-like) and autonomic stress responses; the PVH, whose hypophysiotropic CRF neurons receive a CRF-immunoreactive input, defining a substrate for posited feedback and feed-forward effects on the central limb of the HPA axis, and which is otherwise a sensitive site for peptide effects on ingestive behavior and cardiovascular responses; and the nucleus of the solitary tract (NTS) and ventrolateral medulla (VLM), the linchpins of interoceptive information processing in brain and where CRF acts at low doses to elicit sympathomimetic effects on cardiovascular parameters. Many would include among this grouping the locus coeruleus, whose widespread noradrenergic output is thought to set levels of arousal and 'behavioral vigilance' and where unit activity is enhanced by local administration of CRF. Overall, the lack of an apparent basis in CRFR expression for the generalized CAS activation elicited by central peptide injection poses an unresolved challenge to the widely held view of CRF as a common arbiter of multiple modalities of stress responses.

The CAS and Its Role in Adaptive Responses to Stress

The basic organization of circuitry considered to compose the CAS is shown schematically in Figure 2. Integral components include aspects of the bed nucleus of the stria terminalis (BST), the PVH, the CeA, the lateral parabrachial nucleus, the A1 and C1 catecholamine cell groups of the VLM, and the medial division of the NTS. The NTS is the pivotal sensory structure in this system as it is the principal central nervous system recipient of primary visceral sensory input carried by the vagus and glossopharyngeal nerves and projects directly to all other components, which, in turn, are extensively interconnected with one another. Both the NTS and the VLM also receive noxious inputs by way of the marginal zone of the spinal dorsal horn, indicating a broader range of sensory control over this circuitry. Major effector arms of the CAS include the main cell groups governing the neural arm (sympathetic preganglionic neurons) and endocrine arm (CRF-expressing neurosecretory neurons) of the stress response. Equally direct access to the limbic forebrain (BST and CeA) provides a basis for invoking mechanisms involved in evaluating the emotional significance of environmental events and in effecting appropriate changes in behavior. Though less elaborately interwoven into this system, the locus coeruleus is considered here along with it as it is interconnected with other CAS components, mediates relevant arousal-like functions, is among the sites at which cellular effects of CRF are best characterized, and, like *bona fide* CAS components, receives a CRF and/or Ucn input that arises from other CAS structures.

The validation of inducible immediate-early genes such as *c-fos* and its protein product, Fos, to serve as markers with which to identify neurons that are activated in response to salient environmental events has enabled substantial progress to be made in unraveling central stress-related circuitry. This work has identified the CAS as being recruited to action in a nearly invariant manner in response to a wide variety of acute stressors, though the extent to which it may mediate adaptive responses may vary as a function of the nature of the challenge. Thus, in so-called physiological stress models (e.g., hemorrhage or immune challenges), core CAS structures are dominant seats of cellular activation, and evidence implicating projections originating from the sensory arm of these pathways in mediating HPA and other adaptive responses in these two models has been provided. Models of the other major, 'emotional' stress category (e.g., restraint, footshock) are also accompanied by widespread CAS activation, though the extent of this activation is variable and less pervasive than that seen in physiological models. The fact that the CAS is so generally called into play in response to diverse challenges is consistent with the view that it may constitute a system for reflexive mediation of generalized alarm responses, in a manner coarsely analogous to the sympathoadrenal system in the periphery. It must be noted, however, that for no emotional stress model

median preoptic nucleus; mfb, medial forebrain bundle; MPO, medial preoptic area; NTS, nucleus solitary tract; OB, olfactory bulb; ot, optic tract; OT, olfactory tubercle; PAG, periaqueductal gray; PB, parabrachial nucleus; PG, pontine gray; POR, perioculomotor raphe; PP, posterior pituitary; PPN, pedunculopontine nucleus; PSV, principal sensory nucleus trigeminal nerve; PVT, paraventricular nucleus thalamus; RAPHE, midbrain raphe nucleus; RN, red nucleus; SCh, suprachiasmatic nucleus; S. COL, superior colliculus; SEPT, septal region; SI, substantia innominata; SN, substantia nigra; SO, supraoptic nucleus; SOC, superior olivary complex; SpV, spinal trigeminal nucleus; st, stria terminalis; SUM, supramamillary nucleus; THAL, thalamus; VENT THAL, ventral thalamus; VEST., vestibular nucleus; VLM, ventrolateral medulla; VMH, ventromedial n, hypothalamus; VTA, ventral tegmental area.

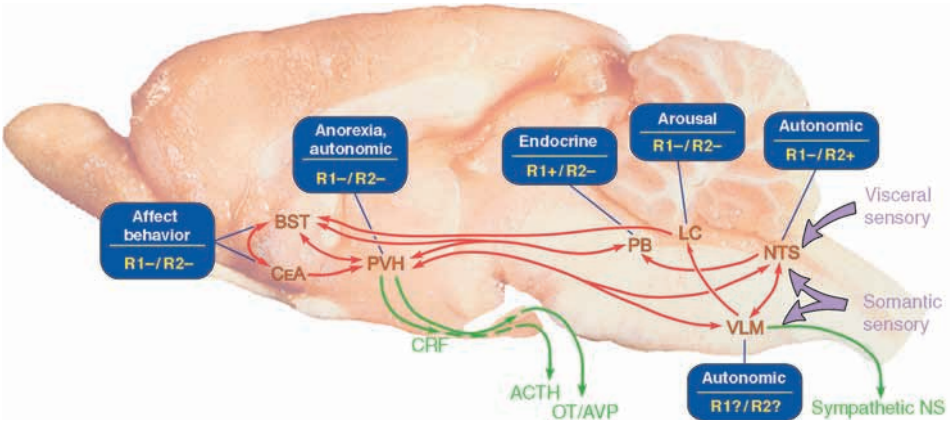

Figure 2 Organization of the central autonomic system in rat brain. Note extensive, often bidirectional interconnections (arrows) among component cell groups. Major sources of sensory input (purple) and motor output (green) are highlighted. The status of each cell group as regards its capacity for CRFR1 (R1) and CRFR2 (R2) messenger RNA expression is indicated, as are some stress-related functions modulated by CRFR agonist/antagonist injections in each. ACTH, adrenocorticotrophic hormone; AVP, arginine vasopressin; BST, bed nucleus stria terminalis; CeA, central nucleus amygdala; CRF, corticotropin-releasing factor; LC, locus coeruleus; NS, nervous system; NTS, nucleus solitary tract; OT, olfactory tubercle; PB, parabrachial nucleus; PVH, paraventricular nucleus; VLM, ventrolateral medulla.

Table 1 Cellular activation profiles elicited by central injection of CRF-related peptides

Peptide	*CRFR1-expressing cell groups*	*CRFR2-expressing cell groups*	*Central autonomic system cell groups*
CRF	+	−	+
Ucn 1	+	+	+
Ucn 2	−	−	+
Ucn 3	−	−	+

Table 2 Effects of CRFR disruption on CRF-induced CNS responses

Disruption	*CRFR1-expressing cell groups*	*CRFR2-expressing cell groups*	*Central autonomic system cell groups*
CRFR antagonist	↓	−	↓
CRFR1 knockout	↓	−	−
CRFR2 knockout	−	−	↑
Double CRFR KO	↓	−	−

has the circuitry underlying CAS activation been identified, and for some such paradigms, the evidence supports a modulatory, rather than a mediating, role.

To what extent does the CRF signaling system participate in the brain's processing of stressful information? Central injections of CRF or Ucn 1 elicit patterns of cellular activation (e.g., Fos induction) that share substantial similarities with those elicited by various stressors, including recruitment of all cell groups of the CAS, as defined here. These responses can be blocked in a dose-related manner by coadministration of nonselective CRFR antagonists. But the evidence supporting an involvement of CRF signaling mechanisms in shaping the brain's response profile in stressful situations is less compelling. Studies involving central administration of CRFR antagonists have reported variable, and generally limited, effectiveness in mitigating stress-induced activation of CAS components. These issues are raised by recent studies of ligand- or stress-induced central nervous system activational profiles in mice genetically engineered to lack functional expression of either or both of the known CRFRs. Thus, central administration of CRF to mice lacking in CRFR1 expression gives rise to markedly reduced activation of cell groups in which CRFR1 is prominently expressed, including cerebral cortex and cerebellum, but CAS responses are unaffected relative to those of genetically intact mice (Table 1 and 2). Moreover, CRFR1-deficient animals display no alteration of CAS responses to acute physiological (intravenous endotoxin injection) or emotional (restraint) stressors relative to wild-type controls, while CRFR2-deleted animals showed exaggerated recruitment of some CAS cell groups in

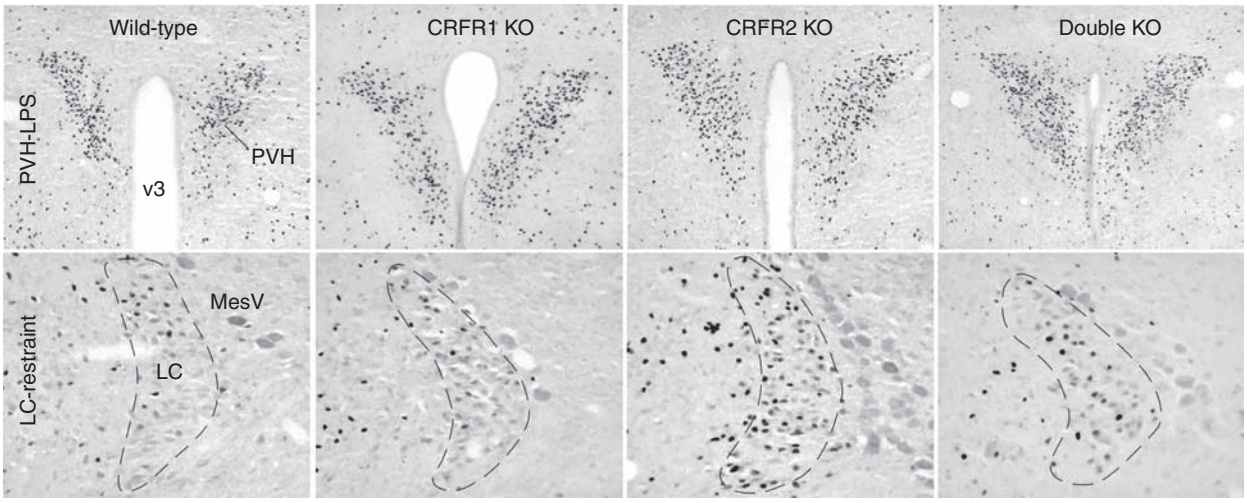

Figure 3 Stress-induced cellular activation of central autonomic cell groups in wild-type and corticotropin-releasing factor receptor (CRFR)-deficient mice. Immunohistochemical localization of Fos protein (a generic marker of neuronal activation) induced in the paraventricular nucleus (PVH; top) in response to a physiological stressor (intravenous lipopolysaccharide (LPS)), and in the locus coeruleus (LC) in response to an emotional stressor (30 min restraint), as a function of CRFR status. None of the CRFR knockouts (KOs) disrupt stress-induced responses in central autonomic system, which are enhanced in the LC of CRFR2 mutants. MesV, mesencephalic trigeminal nucleus; V3, third ventricle.

response to the same challenges (Figure 3). Mice lacking both receptors display stress-induced response profiles like those of CRFR1 mutant animals.

These findings define a basic paradox. On one hand, studies employing functional assays strongly support an involvement of the central CRF system in modulating endocrine, autonomic, and behavioral responses to stress and point toward CAS structures as key sites of action in this regard. Yet on the other hand, these cell groups display limited capacity for CRFR expression, and their engagement in response to ligand administration or stress is not critically dependent on the integrity of either or both of the known CRFRs.

Issues in Understanding CAS Activation by Central Peptide Administration and Stress

Questions remain as to the basis for signaling by CRF-related peptides at demonstrated sites of action that are known to exhibit limited expression of cognate ligands and/or receptors and as to how these elements participate in the response of CAS components to central administration of ligand or stress. A number of possible means for reconciling the seemingly pervasive ligand–receptor misalignment in this system may be considered.

CRFR Protein Expression

The lack of a firm understanding of CRFR distributions at the protein level remains a major impediment for work aimed at pinpointing central sites and mechanisms of peptide action. Immunohistochemical studies using a variety of CRFR antisera have been published, but these often describe results at odds with the consensus of analyses at the messenger RNA level, and not one has been conducted to an unimpeachable standard of histochemical rigor. There is abundant evidence to document axonal transport of receptors, which may undergo differential trafficking to assume distinctive intracellular localizations in different cell types. Thus, 'presynaptic' receptor localization may define substrates for reconciling instances of apparent mismatch, such as recent electrophysiological evidence supporting presynaptic CRFR1-mediated effects in the CeA. Moreover, stress-inducible expression of CRFR1 has been described in the PVH, as has a capacity for modulation of CRFR2 expression. The range of sites and conditions under which such capabilities may be manifested as receptor protein expression remains to be determined.

Novel Receptors?

As noted above, CAS activation induced by central peptide injection can be blocked in a dose-related manner by coadministration of CRFR antagonists but is unabated in mice lacking either or both of the known CRFRs. These findings are consistent with the existence of a novel CRFR expressed in one or more CAS components. It may be noted that three distinct CRFRs have been cloned in a catfish species, with all three subtypes exhibiting distinctive distributions, including in brain. Arguing against the existence of a third receptor is the fact that recently isolated receptor isoforms in other G-protein-coupled receptor families have all displayed structural

similarities to existing receptors, and searches of human and mouse genome databases for genes related to sequences of various domains of the known CRFRs have failed to identify a gene with significant relatedness to CRFRs or other members of the secretin receptor family.

Receptor Cross-Talk

CAS activation observed in intracerebroventricular injection studies could be a result of interaction of CRF peptides with heterologous receptors. There are precedents for a degree of promiscuity among some subgroups of G-protein-coupled receptors, which may bind more than one structurally related ligand. While there appears to be no precedent for such cross-talk in the CRF system, several members of the secretin receptor family, including receptors for glucagon-like peptide 1, pituitary adenylate cyclase-activating peptide, and vasoactive intestinal polypeptide, are distributed in at least some CAS components and could comprise pharmacologic targets for CRF-related ligands. Further structural characterization of the ligand-binding domains of secretin receptor family members should help define potential participants in such interactions, and their involvement can then be tested using pharmacologic and genetic tools.

Interaction of CRFR- and CRFR2-Dependent Mechanisms

Whether or not the central CRF system proves to be indispensably involved in any given mode of stress adaptation, a growing body of evidence supports an interaction of CRFR1- and CRFR2-dependent mechanisms in shaping such responses. The generally hypo- and hyper-responsive phenotypes of CRFR1 and CRFR2-deficient mice, respectively, may be viewed as indicative of antagonistic interactions. However, there is evidence from both genetic and pharmacologic studies that the two receptors' influences on feeding suppression, anxiety, and HPA axis activation may be cooperative and exerted over distinct time domains. Studies of ligand and stress effects in CRFR-mutant mice provide ample evidence of receptor interplay in CAS sites of action for the above-mentioned responses. The anatomical organization that provides for convergent CRFR influences on a wide range of stress-related endpoints is completely obscure.

Other Ligand-Binding Proteins

A CRF-binding protein (CRF-BP) distinct from CRFRs was isolated as a circulating factor capable of binding and reversibly neutralizing CRF actions, thus preventing inappropriate pituitary–adrenal activation during pregnancy in primates. The CRF-BP is expressed prominently in brain, where its distribution suggests multiple potential sites and modes of interaction with CRF. The question of whether the CRF-BP may serve roles beyond its capacity to passively sequester peptide remains controversial. From synthetic CRF fragments that are bound with high affinity by the CRF-BP, but not either CRFR, evidence has been obtained to support an active role for the BP in signaling. This may be significant as the CRF-BP is more broadly expressed in the CAS than either known CRFR.

In a similar vein, a novel CRFR2α variant has been recently cloned and expressed, and its truncated form predicts that it exists as a soluble receptor (CRFR2α-S). Its abundance in brain rivals that of the membrane-bound form, raising the possibility that CRFR2α-S may serve a role in modulating signaling by CRFR2 agonists. The fact that the reported central distribution of the soluble receptor more closely mimics that of CRFR1 than CRFR2 is ostensibly at odds with this view.

Role of Ucns

Like other CRFR agonists tested to date, central administration of Ucn 2 or 3 strongly activates CAS cell groups; unexpectedly, though, sites of CRFR2 expression are unresponsive. Resolution of this paradox is needed to understand the basis for CRFR2-mediated signaling and its interaction with CRFR1-dependent mechanisms. At least three alternate explanations of this phenomenology may be considered: that CRFR2 activation may be CRFR1-dependent, that the naturally occurring form(s) of Ucn 2 and 3 may differ in activity from those predicted by complementary DNA sequences, and that access of centrally administered CRFR2 agonists may be limited by the actions of CRFR2α-S. Interest in the central functions of Ucns centers about their capacities to elicit stresslike anorexic and anxiolytic behavior, actions originally attributed to CRF, and, as putative endogenous CRFR2 ligands, to modulate a range of CRFR1-mediated activities. Each of the Ucns displays a restricted and unique central distribution (Figure 2). Only Ucn 1 pathways have been found to innervate CAS structures (mainly NTS) with any prominence. Ucn 3-containing projections are best aligned with major sites of CRFR2 expression in limbic forebrain (lateral septum, medial amygdala, ventromedial hypothalamus), which have the capacity to influence CAS mechanisms, albeit indirectly.

Significance

A wealth of evidence has implicated the central CRF system in the modulation of neuroendocrine, autonomic, and behavioral responses to stress, and CAS cell groups have been repeatedly implicated as

important sites of action in this regard. Studies involving repeated or chronic central administration of CRF have implicated elevated peptide levels in the genesis of stress-related physiological and behavioral disorders. Nevertheless, the relative paucity of evidence for cellular expression of CRFRs within the CAS and the general failure of pharmacologic or genetic disruption of CRFR signaling to mitigate stress-induced CAS activational responses pose stern questions as to mechanisms that provide for CRF system involvement in stress adaptations. Small-molecule CRFR1 antagonists are currently being evaluated in clinical trials for the treatment of major depression, other anxiety-related disorders, and at least one systemic condition, irritable bowel syndrome. Additional applications of drugs that target this signaling system in a range of stress-related illnesses are being explored avidly. Deeper understanding of the sites and mechanisms through which such agents act is essential to exploiting them in an informed and effective manner.

See also: Chronic (Repeated) Stress: Consequences, Adaptations; Corticotropin-Releasing Hormone and Urocortins: Binding Proteins and Receptors; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress; Sympathoadrenal System: Neural Arm of the Stress Response.

Further Reading

Bakshi VP, Smith-Roe S, Newman SM, Grigoriadis DE, and Kalin NH (2002) Reduction of stress-induced behavior by antagonism of corticotropin-releasing hormone 2 (CRH2) receptors in lateral septum or CRH1 receptors in amygdala. *Journal of Neuroscience* 22: 2926–2935.

Bale TL and Vale WW (2004) CRF and CRF receptors: Role in stress responsivity and other behaviors. *Annual Review of Pharmacology and Toxicology* 44: 525–572.

Bittencourt JC and Sawchenko PE (2000) Do centrally administered neuropeptides access cognate receptors?: An analysis in the central corticotropin-releasing factor system. *Journal of Neuroscience* 20: 1142–1156.

Chadwick DJ, Marsh J, and Ackrill K (eds.) (1993) *Corticotropin-Releasing Factor, Ciba Foundation Symposium 172*. London: Wiley.

Dautzenberg FM, Kilpatrick GJ, Hauger RL, and Moreau J-L (2001) Molecular biology of the CRH receptors – in the mood. *Peptides* 22: 753–760.

De Souza EB, and Nemeroff CB (eds.) (1990) *Corticotropin-Releasing Factor: Basic and Clinical Studies of a Neuropeptide*. Boca Raton, FL: CRC Press.

Grammatopoulos DK and Chrousos GP (2002) Functional characteristics of CRH receptors and potential clinical applications of CRH-receptor antagonists. *Trends in Endocrinology and Metabolism* 13: 436–444.

Herkenham M (1987) Mismatches between neurotransmitter and receptor localizations in brain: Observations and implications. *Neuroscience* 23: 1–38.

Radulovic J, Ruhmann A, Liepold T, and Spiess J (1999) Modulation of learning and anxiety by corticotropin-releasing factor (CRF) and stress: Differential roles of CRF receptors 1 and 2. *Journal of Neuroscience* 19: 5016–5025.

Saper CB (1995) Central autonomic system. In: Paxinos G (ed.) *The Rat Nervous System*, 2nd edn., pp. 107–128. San Diego, CA: Academic Press.

Spina M, Merlo-Pich E, Chan RKW, et al. (1996) Appetite-suppressing effects of urocortin, a CRF-related neuropeptide. *Science* 273: 1561–1564.

Van Pett K, Viau V, Bittencourt JC, et al. (2000) Distribution of mRNAs encoding CRF receptors in brain and pituitary of rat and mouse. *Journal of Comparative Neurology* 428: 191–212.

Vaughan J, Donaldson C, Bittencourt J, et al. (1995) Urocortin, a mammalian neuropeptide related to fish urotensin I and to corticotropin-releasing factor. *Nature* 378: 287–292.

Cotransmission

G Burnstock, Royal Free and University College Medical School, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Early Background

For many years, the understanding of neurotransmission has been dominated by the concept that one neuron releases only a single transmitter, known as 'Dale's Principle.' This idea arose from a widely adopted misinterpretation of Dale's suggestion in 1935 that the same neurotransmitter was stored in and released from all terminals of a single neuron, a suggestion which did not specifically preclude the possibility that more than one transmitter may be associated with the same neuron. Several lines of evidence emerged which were inconsistent with the single transmitter concept, and it is now known that individual neurons contain and can release a large number and variety of substances which are capable of influencing target cells. This phenomenon of 'cotransmission' is widespread, involving virtually all known transmitter systems.

Early hints of cotransmission came in the 1950s with evidence for the involvement of both noradrenaline (NA) and acetylcholine (ACh) in sympathetic transmission. Koelle identified acetylcholinesterase in some adrenergic neurons in 1955, while Burn and Rand introduced the concept of a 'cholinergic' link in adrenergic transmission in 1959. Another line of evidence provided by Hillarp concerned the coexistence of adenosine 5′-triphosphate (ATP) with catecholamines, first in adrenal chromaffin cells and later in sympathetic nerves. Inconsistencies in the single transmitter hypothesis provided by these and other studies from the early literature were rationalized in an article by Burnstock in 1976 with the provocative title,"Do some nerve cells release more than one transmitter?" Today it is widely accepted that cotransmission is an integral feature of neurotransmission.

A role for ATP as a cotransmitter in sympathetic, parasympathetic, sensory–motor, and enteric nonadrenergic, noncholinergic (NANC) inhibitory nerves was supported by research from Burnstock and colleagues, while Hökfelt and colleagues focused on the co-localization, vesicular storage, and release of peptides from both peripheral and central nerves.

Furness and Costa introduced the concept of 'chemical coding' to describe the combination of potential neurotransmitters found in enteric nerves, and this concept has since been applied to other nerve types, in both peripheral and central nervous systems CNSs (see Figure 1). Co-localized substances are not necessarily cotransmitters; they can (especially peptides) act as pre-and/or postjunctional neuromodulators of the release and actions of the principal cotransmitters. The proportions of cotransmitters vary considerably between species and organs and show plasticity of expression during development and in pathological conditions. In general, classical transmitters are contained in small synaptic vesicles, whereas peptides are stored in large granular (dense-cored) vesicles (LGVs), although small molecule transmitters are sometimes stored together with peptides in LGVs.

Pharmacological studies of pre- and postjunctional neuromodulation provide evidence which is complementary to the concept of cotransmission. For example, parallel presynaptic modulation of transmitter overflow supports the concept of closely associated co-release, while postjunctional synergism between co-localized transmitters provides justification of cotransmission in terms of transmitter economy.

Sympathetic Nerve Cotransmission

There is substantial evidence to show that NA, ATP, and neuropeptide Y (NPY) are cotransmitters in sympathetic nerves, having differentially important roles as transmitters and neuromodulators, depending on the tissue, the species, and the parameters of stimulation. Most of the early studies establishing the model of cotransmission of NA and ATP were made on the vas deferens, a tissue with a high density of sympathetic nerves. Subsequently, numerous studies demonstrated that cotransmission of NA and ATP also occurs in other visceral organs and many different blood vessels in a variety of species.

The first indication that ATP might be released from sympathetic nerves was the demonstration that stimulation of sympathetic nerves by Su, Bevan, and Burnstock in 1971 led to release of tritium from taenia coli preincubated in [^{3}H]adenosine (which is taken up and converted to [^{3}H]ATP). Later Sol Langer and colleagues suggested that the substantial residual NANC response of the cat nictitating membrane observed after depletion of NA by reserpine was due to the release of ATP remaining in the sympathetic nerves. In retrospect, there was a good indication that the excitatory junction potentials (EJPs) recorded in the guinea pig vas deferens when the electrophysiology of sympathetic nerve smooth muscle transmission was first described by Burnstock and Holman in 1960 were due to ATP released as a cotransmitter from sympathetic nerves rather than to NA. It was puzzling at the time that adrenoceptor antagonists failed to block

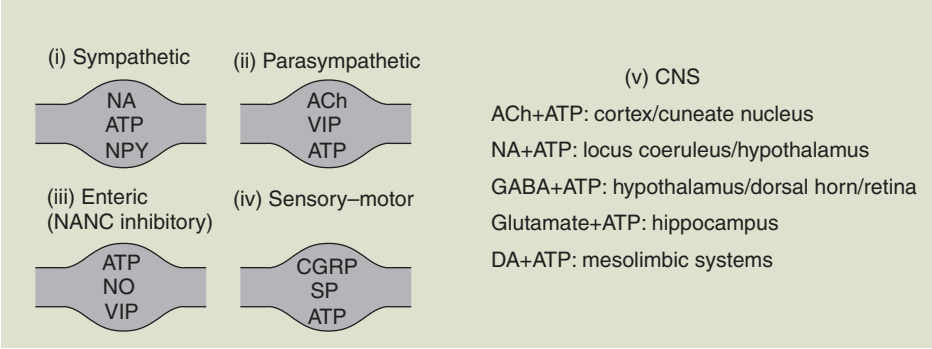

Figure 1 Schematic diagram of the principal cotransmitters with ATP in the nervous system. Nerve terminal varicosities of (i) sympathetic, (ii) parasympathetic, (iii) enteric (NANC inhibitory), (iv) sensory–motor neurons, and (v) central nervous system (CNS). Reproduced from Burnstock G (2007) Physiology and pathophysiology of purinergic neurotransmission. *Physiological Reviews* 87: 659–797, with permission from The American Physiological Society.

the EJPs, although guanethidine, a drug that prevents the release of sympathetic transmitters, was effective. It was more than 20 years later that NANC EJPs were shown to be blocked by desensitization of the ATP (P2) receptors by α,β-methylene ATP and mimicked by ATP. After destruction of sympathetic nerves with 6-hydroxydopamine, purinergic nerve-mediated responses were abolished. ATP is co-stored with NA in small and large vesicles. Differential prejunctional modulation of the release of NA and ATP by various agents has been shown in the vas deferens, perhaps indicating that NA and ATP are stored in different vesicles.

Cotransmission of NA and ATP in perivascular sympathetic nerves supplying the rabbit aorta, portal vein, and saphenous, pulmonary and mesenteric arteries and the dog basilar artery has been demonstrated. Electrophysiological studies have shown that in a number of vessels, the electrical response to stimulation of perivascular sympathetic nerves is biphasic: an initial fast, transient depolarization or EJP of the vascular smooth muscle is followed by a slow, prolonged depolarization. The EJP and slow depolarization are mimicked by the effects of ATP and NA, respectively. Considerable variation exists in the proportions of NA and ATP utilized by sympathetic nerves. For example, in guinea pig submucosal arterioles, both vasoconstriction and EJPs, evoked in response to electrical stimulation of sympathetic nerves, are mediated exclusively by ATP, with NA assuming the role of a neuromodulator, acting through prejunctional α_2-adrenoceptors to depress transmitter release. At the other extreme, in rat mesenteric arteries, the purinergic component is relatively small. In addition, it has been noted that the purinergic component is optimal with short bursts of low-frequency stimulation, whereas longer durations of higher frequency favor adrenergic transmission.

NPY has been found to be present in LGV in most sympathetic nerves. The release of NPY, as well as

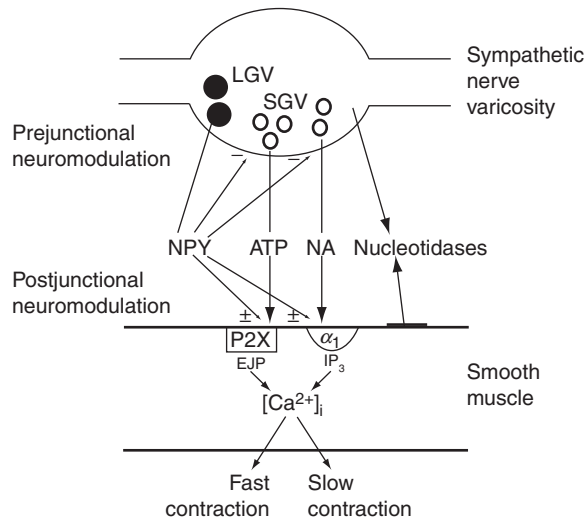

Figure 2 Schematic of sympathetic cotransmission. Adenosine 5′-triphosphate (ATP) and noradrenaline (NA) released from small granular vesicles (SGV) act on P2X and α_1 receptors on smooth muscle, respectively. ATP acting on inotropic P2X receptors evokes excitatory junction potentials (EJPs), increase in intracellular calcium ($[Ca^{2+}]_i$), and fast contraction while occupation of metabotropic α_1 adrenoceptors leads to production of inositol triphosphate (IP_3), increase in $[Ca^{2+}]_i$, and slow contraction. Neuropeptide Y (NPY) stored in large granular vesicles (LGV) acts on release both as a prejunctional inhibitory modulator of release of ATP and NA and as a postjunctional modulatory potentiator of the actions of ATP and NA. Nucleotidases are released from nerve varicosities and are also present as ectonucleotidases.

NA and ATP, in response to electrical stimulation of sympathetic nerve terminals is prevented by guanethidine. The major role of NPY in the vasculature, and in the vas deferens, appears to be that of a pre- and/or postjunctional modulator of sympathetic transmission since it has little direct postjunctional action or causes contraction only at high concentrations (see Figure 2). Direct vasoconstrictor actions of NPY have, however, been demonstrated in some vessels. At the prejunctional level, NPY has potent inhibitory

effects, reducing the release of NA and ATP from sympathetic nerves. Postjunctionally, NPY generally acts to enhance the actions of sympathetic nerve stimulation, NA, and ATP.

Although 5-hydroxytryptamine (5-HT) immunofluorescent nerves have been localized in a number of vessels, for the most part 5-HT is not synthesized and stored in separate nerves but is taken up, stored in, and released as a 'false transmitter' from sympathetic nerves. Enkephalins have been shown to coexist with NA in cell bodies and fibers of some postganglionic sympathetic neurons. Both NA and met-enkephalin are co-released by transmural stimulation in a guanethidine-sensitive manner. The functional significance of sympathetic coexistence of opioids is likely to be related to their prejunctional inhibitory effects on sympathetic transmission.

Parasympathetic Nerve Cotransmission

The classical evidence for cotransmission of ACh and vasoactive intestinal polypeptide (VIP) in certain postganglionic parasympathetic neurons comes from pharmacological studies performed by Lundberg in 1981 on cat salivary glands. ACh and VIP are released from the same parasympathetic nerve terminals in response to transmural nerve stimulation. During low-frequency stimulation, ACh is released to cause an increase in salivary secretion from acinar cells and also to elicit some minor dilatation of blood vessels in the gland. VIP is preferentially released at high frequencies to cause marked vasodilatation of blood vessels, and while it has no direct effect on acinar cells, it acts as a neuromodulator to enhance both the postjunctional effect of ACh on acinar cell secretion and the release of ACh from nerve varicosities via prejunctional receptors. Vasodilator nerves to the uterine arteries in the guinea pig contain immunoreactivity to VIP, which coexists with dynorphin, NPY, and somatostatin. NPY-like immunoreactivity has been reported in some of the choline acetyltransferase-/VIP-containing neurons of the parasympathetic ciliary, sphenopalatine, otic, and pterygopalatine ganglia with targets including the iris and cerebral vessels.

Autonomic control of penile erection, involving relaxation of the smooth muscle of the corpus cavernosum as well as dilatation of other penile vascular beds, has traditionally been attributed to the vasodilator effects of ACh and VIP released from parasympathetic nerves. More recently, nitric oxide (NO) released from nerves arising from nerves in the pelvic ganglia have been claimed to play a role in smooth muscle relaxation leading to penile erection. Fibers containing NO synthase, shown by lesion studies to arise from parasympathetic cell bodies in the sphenopalatine ganglia, have been localized in the adventitia of cerebral arteries, and many of these also contain VIP. A functional role for perivascular neuronal NO in cerebral arteries has been identified in studies showing that stimulation of adventitial nerve fibers causes vascular relaxation, which is attenuated by inhibitors of NO synthase.

Parasympathetic nerves supplying the urinary bladder utilize ACh and ATP as cotransmitters, in variable proportions in different species, and by analogy with sympathetic nerves, ATP again acts through P2X ionotropic receptors to produce EJPs and fast contraction, while the slow component of the response is mediated by a metabotropic receptor, in this case muscarinic. There is also evidence for parasympathetic, purinergic cotransmission to resistance vessels in the heart and airways.

Sensory–Motor Nerve Cotransmission

The neuropeptides substance P (SP) and calcitonin gene-related peptide (CGRP) are the principal transmitters of primary afferent nerves and have been shown to coexist in the same terminals. Furthermore, with the use of colloidal gold particles of different sizes, they have been shown to coexist in the same large granular vesicles. The motor (efferent) function of sensory nerves has been demonstrated in rat mesenteric arteries, where evidence exists for a role for CGRP as the mediator of vasodilatation following release from sensory–motor nerves. In contrast, SP is not co-released with CGRP by electrical stimulation, and SP has little or no vasodilator action on rat mesenteric arteries. While it is possible that SP released from nerves supplying the microvasculature could produce vasodilatation via SP receptors on endothelial cells, it is most unlikely to reach the endothelium without degradation in larger blood vessels. It may be that the role of the coexisting SP is either trophic or sensory (and not motor).

Other peptide and nonpeptide substances, including neurokinin A, somatostatin, VIP, and ATP, have been described in capsaicin-sensitive sensory neurons. Unmyelinated sensory neurons containing cholecystokinin (CCK)/CGRP/dynorphin/SP have been shown to project to cutaneous arterioles in guinea pig skin. Neurons from the same ganglia which contain CCK/CGRP/SP innovate arterioles of skeletal muscle, CGRP/dynorphin/SP nerve fibers mostly supply the pelvic viscera, and CGRP/SP fibers run mainly to the heart, large arteries, and veins. There is also evidence for a sensory role for ATP, and it has been proposed that ATP may coexist in sensory nerve terminals with SP and CGRP.

Intramural Nerve Cotransmission

Intrinsic neurons exist in most of the major organs of the body. Many of these are part of the parasympathetic nervous system, but certainly in the gut and perhaps also in the heart and airways, some of these intrinsic neurons are derived from neural crest tissue, which differs from that which forms the sympathetic and parasympathetic systems and appears to participate in local reflex actions independent of the CNS.

The enteric nervous system contains several hundred million neurons located in the myenteric plexuses between muscle coats and the submucous plexus. The chemical coding of these nerves has been examined in detail. A subpopulation of these intramural enteric nerves provides NANC inhibitory innervation of the gastrointestinal smooth muscle. It seems likely that three major cotransmitters are released from these nerves. ATP produces fast inhibitory junction potentials, NO also produces inhibitory potentials but with a slower time course, while VIP produces slow tonic relaxations. The proportions of these three transmitters vary considerably in different regions of the gut and in different species; for example, in some sphincters, the NANC inhibitory nerves utilize largely VIP, in others largely NO, while in nonsphincteric regions of the intestine, ATP is more prominent. In recent papers, evidence has suggested that ACh and ATP are fast excitatory cotransmitters to myenteric neurons and that there may be co-localization of ACh, ATP, and serotonin in enteric S neurons. Detailed studies have allowed a very complete mapping of the complex neuronal markers and projections of enteric neurons. Several peptidergic substances, including NPY, VIP, peptide histidine isoleucine, SP, and CGRP, have been identified in enteric neurons, often coexisting (up to five peptides in the same neuron) with the neurotransmitters NA, ACh, 5-HT, NO, and ATP. The precise roles of the coexisting substances, however, have not for the most part been established, except for the proposed combinations of ACh and SP in the excitatory nerves and ATP, NO, and VIP in the NANC inhibitory neurons involved in peristaltic reflexes.

Studies of intrinsic cardiac neurons in culture have shown that subpopulations of intrinsic nerves in the atrial and intra-atrial septum contain and/or release cotransmitters, including ATP, NO, NPY, Ach, and 5-HT. Many of these nerves project to the coronary microvasculature and produce potent vasomotor actions. NO and ATP have been shown to be the mediators of NANC vasodilatation of the rabbit portal vein.

CNS

Evidence for ATP's being a cotransmitter with established neurotransmitters in the CNS has been reported. In preparations of affinity-purified cholinergic nerve terminals from the rat cuneate nucleus, ATP and ACh are co-released. Co-release of ATP with catecholamines from neurons in the hypothalamus and locus ceruleus has been reported, and there is recent evidence for co-release of ATP with γ-aminobutyric acid (GABA) in dorsal horn and lateral hypothalamic neurons, and for ATP with glutamate in the hippocampus. Co-localization of functional nicotinic and ionotropic nucleotide receptors has been identified in isolated cholinergic synaptic terminals in the midbrain; ATP and dopamine (DA) are probably co-released from the terminals of ventral tegmental neurons in the nucleus accumbens.

Co-release and interaction of two fast inhibitory cotransmitters, GABA and glycine, in synaptic bouton preparations of the sacral dorsal commissural nucleus of the sacral spinal cord have been described. Earlier papers showed that glycine/GABA cotransmission occurred in brain stem motor neurons and spinal interneurons. Co-release of NA and DA from neurons in the cerebral cortex has also been reported. Neurons in the tuberomammillary nucleus in the posterior hypothalamus contain histamine, GABA, galanin, enkephalin, and SP as cotransmitters. GABA/somatostatin cotransmission has been reported at synapses in a subpopulation of amygdala projection neurons to the nucleus tractus solitarius, which might inhibit cardiovascular reflex responses to fear or emotion-related stimuli.

CCK is co-localized with DA in rat mesencephalic neurons, and with glutamate in corticostriatal neurons; released CCK appears to be involved in locomotor behavior. Synthesis and storage of glutamate, ACh, and GABA in basal forebrain neurons projecting to the entorhinal cortex have been reported. Galanin is co-stored with enkephalin, and often NPY, in some neurons of the substantia gelatinosa, whereas tachykinins and enkephalin, galanin, and SP are co-stored in neurons in deeper layers of the dorsal horn. Co-storage of galanin and neurotensin, as well as CGRP and SP, has been shown by postembedding immunocytochemistry to be present in LGVs. Reverse transcription polymerase chain reaction (RT-PCR) and *in situ* hybridization studies have shown co-storage of oxytocin and vasopressin mRNA in LGVs in hypothalamic neurosecretory neurons.

Physiological Significance of Cotransmission

Several major themes have emerged about the physiology of cotransmission (see Figure 3).

Fast and Slow Cotransmitters: Different Firing Patterns

Although single presynaptic action potentials release small molecule neurotransmitters, trains of impulses are needed to release neuropeptides. For sympathetic and parasympathetic cotransmission, release of ATP is favored at low-frequency stimulation, whereas NA and ACh are released at higher frequencies. There are instances where more than one fast cotransmitter is released (e.g., glutamate and ATP) together with one or more peptides. See Figures 3(a) and 3(b).

Different Cotransmitters Act on Different Postjunctional Cells

Neurons using multiple transmitters may project to two or more targets (Figure 3(c)). For example, ACh released at low-frequency stimulation from parasympathetic nerves supplying salivary glands acts on acinus cells to produce secretion and a minor dilatation of vessels, whereas at higher frequency stimulation, its cotransmitter VIP causes powerful vasodilatation of vessels in the glands and postjunctional enhancement of ACh-induced saliva secretion.

Presynaptic Neuromodulation of Cotransmitter Release

A cotransmitter can feed back on presynaptic receptors that increase or decrease its own release or that of its cotransmitter(s) (Figure 3(d)). For example, ATP released as a cotransmitter with glutamate from primary afferent fibers in lamina II of the spinal cord can act on prejunctional $P2X_3$ receptors to facilitate the release of its cotransmitter, glutamate, whereas adenosine resulting from ectoenzymatic breakdown of ATP acts on presynaptic P1 receptors to inhibit glutamate release. Both NA and ATP can prejunctionally modulate sympathetic transmission, NA via prejunctional α_2-adrenoceptors and ATP via P1 receptors following breakdown to adenosine. Modulation of cotransmitter release and presynaptic action by other agents also occurs and might provide a new level of synaptic flexibility, in which individual neurons utilize more than one transmitter but retain independent control over their synaptic activity.

Synergism

There are an increasing number of reports of the synergistic actions of cotransmitters (Figure 3(e)). ATP and NA released from sympathetic nerves have synergistic actions on smooth muscle of vas deferens and blood vessels, and ATP released with ACh from motor neurons facilitates the nicotinic actions of ACh at the skeletal neuromuscular junction. Cooperativity between receptors for ATP and *N*-methyl-D-aspartate (NMDA) in induction of long-term potentiation in hippocampal CA1 neurons has also been demonstrated. Thyrotropin-releasing hormone and serotonin have been reported to have synergistic actions in spinal cord neurons.

In view of the evidence for cotransmitter synergy, the reports that known nucleotide P2 receptor antagonists, such as suramin, have actions on nonpurinergic receptors need to be questioned. For example, the claims that suramin and reactive blue 2 have antagonistic actions on NMDA and GABA receptor channels in hippocampal neurons are probably explained by blockade of the P2 receptor-mediated responses of the cotransmitter ATP, thereby removing its synergistic potentiating effect.

The mechanisms underlying cotransmitter synergism are not well understood. However, it has been suggested that postjunctional synergism between the responses of vas deferens to NA and ATP is caused by the ability of NA to potentiate the contractile responses to ATP by sensitizing smooth muscle cells to Ca^{2+} via an inhibitory action on myosin light chain phosphatase, an action mediated by protein kinase C.

Negative Cross-Talk

Coapplication of nicotinic and P2X receptor agonists produces less than the additive responses predicted by independent receptor activation (Figure 3(f)). Inhibitory interactions between $5\text{-}HT_3$ and P2X receptors have been described in submucosal and myenteric neurons. Cross-inhibition of $GABA_A$ and glycine receptors has been demonstrated in rat sacral dorsal commissural neurons.

Cotransmitters and Trophic Factors

Some co-stored and co-released substances can act as long-term (trophic) factors, as well as neurotransmitters (Figure 3(g)). For example, ATP can act on P2 receptors, or P1 (adenosine) receptors after ectoenzymatic breakdown, to promote vascular cell proliferation, motility, differentiation, or death. NPY released from sympathetic nerves has cardiovascular trophic effects in end-stage renal disease. There is growing evidence that neurotrophic factors might be synthesized, stored, and released from nerve terminals together with fast neurotransmitters.

Excitatory and Inhibitory Cotransmitters

Although cotransmitters generally have similar actions on postjunctional cells, there are some examples of cotransmitters having opposite actions. For example, in the mammalian uterus, one or other cotransmitter dominates depending on the hormonal and/or tonic status of the postjunctional muscle cells

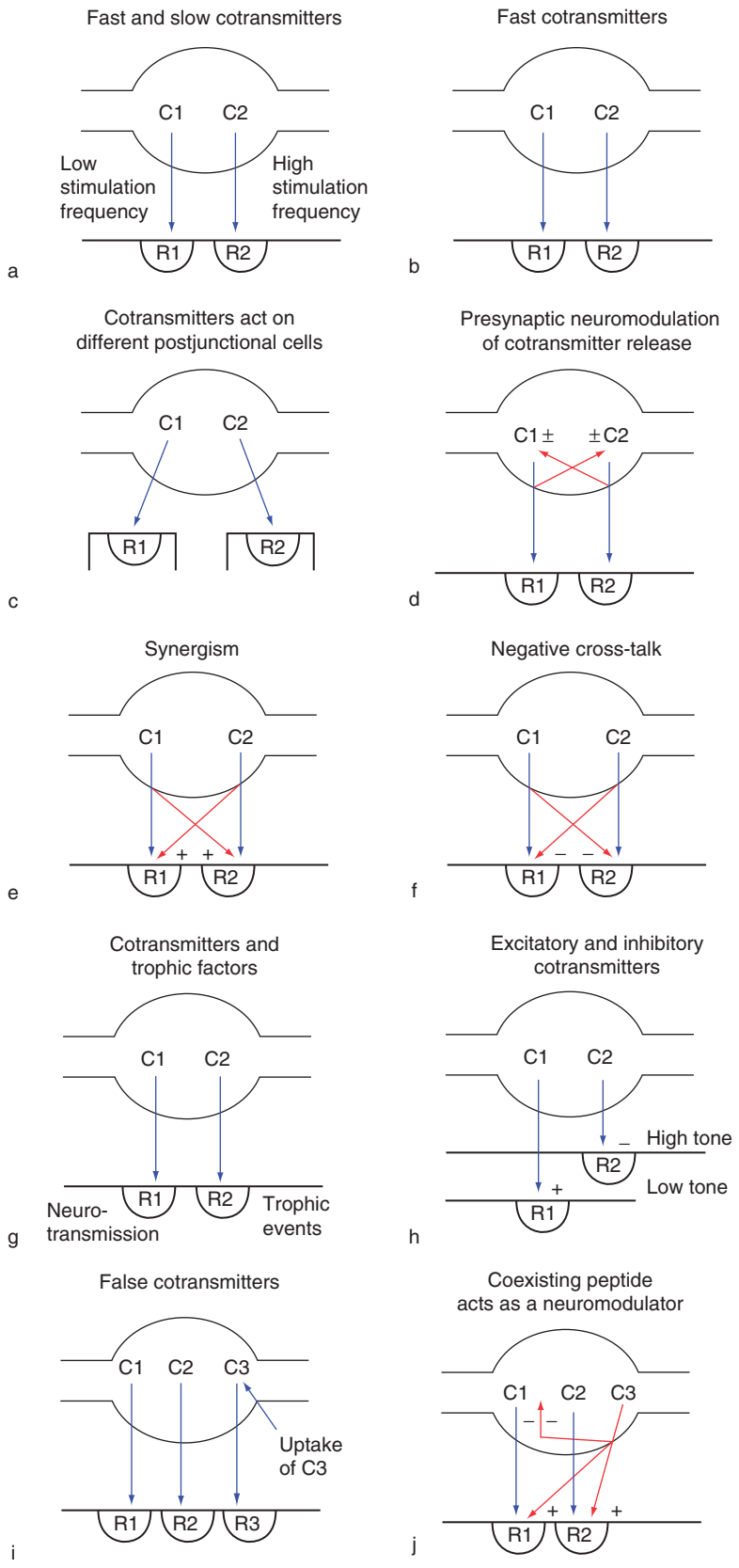

Figure 3 Spectrum of signaling variations offered by cotransmission (blue arrows = neurotransmission; red arrows = pre- or postjunctional neuromodulation). (a) Fast transmission is usually produced by small molecules (C1) released at low-frequency nerve stimulation acting on ionotropic receptors (R1), whereas slow transmission is usually produced by release of peptides (C2) or other

(Figure 3(h)). Brain-derived neurotrophic factor (BDNF) increases the release of ACh and reduces NA release from sympathetic nerves to cause a rapid shift from excitatory to inhibitory transmission.

False Cotransmitters

For example, it has been known for some time that sympathetic nerves take up 5-HT, which can then be released as a 'false transmitter' rather than a genuine 'cotransmitter.' A false transmitter is a substance actively taken up and subsequently released by a neuron that does not synthesize it (Figure 3(i)).

Coexisting Peptide Acting as a Neuromodulator

For example, NPY released from sympathetic nerves acts as a pre- and postjunctional modulator of ATP and NA release and postjunctional actions (Figure 3(j)).

Cotransmitter Plasticity: Control of Transmitter Expression

Cotransmitter plasticity occurs during development and aging, following trauma or surgery, and after chronic exposure to drugs, as well as in disease. There were some outstanding early studies of the factors influencing cotransmitter expression in sympathetic nerves, and a physiological role for neuropoietic cytokines in the control of VIP expression during the development of cholinergic sympathetic neurons was proposed. A more recent study presented evidence that cholinergic differentiation in sympathetic neurons is promoted by neurotrophic factors from three different protein families (glial cell line-derived neurotrophic factor, neurotrophin 3, and ciliary neurotrophic factor), whereas noradrenergic differentiation is promoted by nerve growth factor. In another study, BDNF was claimed to switch sympathetic neurotransmission to the heart from an adrenergic excitation to cholinergic inhibition; it was also shown that the action of BDNF was mediated by the P75 neurotrophic receptor. Histamine, galanin, and GABA acting as cotransmitters in neurons of the tuberomammillary nucleus (hypothalamus) have independent control mechanisms. Changes in chemical coding of myenteric neurons in ulcerative colitis have been reported, with a shift from cholinergic to more SP-positive cotransmission. In a study using primary cultures of neonatal rat spinal neurons, evidence was presented for the regulation of SP (NK1) receptor expression by CGRP.

CGRP-like immunoreactivity was found earlier than SP-like immunoreactivity in cerebrovascular nerves during development, and increased in old age, while the density of SP-like immunoreactive fibers did not change. NA and NPY also show different expression in cerebrovascular nerves during development. Direct evidence for changes in transmitter ratio in disease comes from a study of hypertension in which the purinergic component of sympathetic cotransmission has been claimed to be enhanced to the extent that it is the dominant component of the sympathetic response. In many isolated blood vessels, contractions produced by sympathetic nerve stimulation or due to vasoconstrictors including catecholamines and 5-HT are greater after the endothelium is removed or during antagonism of endothelium-derived relaxing factor. While part of this effect is likely to involve postjunctional mechanisms, evidence has been presented that substances released from the endothelium may act prejunctionally to influence neurotransmitter release from nerves. This may or may not involve endothelial-derived NO. ATP is also released from endothelial cells in response to physiological stimuli such as hypoxia or shear stress and may thus modulate the activity of perivascular nerves via prejunctional P1 receptors following breakdown to adenosine and diffusion through the vessel wall. Conversely, in the microvasculature, where neural-endothelial separation is small, cotransmitters released from nerves could act directly on endothelial cells to influence the release of endothelium-derived factors.

molecules at high frequency stimulation acting on G protein-coupled receptors (R2). (b) Cotransmitters C1 and C2 can both be fast messengers acting via ionotropic receptors (R1 and R2). (c) Cotransmitters C1 and C2 act on receptors (R1 and R2) localized on different postjunctional cells. (d) Cotransmitters C1 and C2 not only act postjunctionally via R1 and R2 receptors but can also act as prejunctional modulators to either inhibit (−) or enhance (+) the release of C1 and/or C2. (e) Cotransmitters C1 and C2 act synergistically to enhance the combined responses produced via R1 and R2 receptors. (f) Cotransmitters C1 and C2 act to inhibit the responses evoked via R1 and/or R2 receptors. (g) Cotransmitter C1 evokes neurotransmission via R1 receptors, while C2 evokes long-term (trophic) responses of postjunctional cells via R2 receptors. (h) Cotransmitter C1 produces excitation via R1 receptors when the postjunctional smooth muscle target has low tone, with C2 having little influence; however, when the smooth muscle tone is high, the dominant response might be relaxation produced by C2 via R2 receptors. (i) Substance C3 is taken up by nerve terminals rather than being synthesized and stored, as is true for the cotransmitters C1 and C2. C3 can then be released on nerve stimulation to act on postjunctional R3 receptors. In these circumstances, C3 would be known as a 'false transmitter.' (j) A coexisting substance C3 (often a peptide) can be synthesized and stored in a nerve but not act directly via a postjunctional receptor to produce changes in postjunctional cell activity. It could, however, act as a prejunctional inhibitor (−) of the release of the cotransmitters C1 and C2 or as a postjunctional enhancer (+) of the responses mediated by R1 and R2. Reproduced from Burnstock G (2004) Cotransmission. *Current Opinion in Pharmacology* 4: 47–52, with permission from Elsevier.

Concluding Comments

To establish that compounds shown to be localized in nerves are actually cotransmitters, several criteria need to be satisfied:

1. The substance is synthesized and stored in the nerve.
2. The substance is released on nerve stimulation.
3. Specific receptors for the substance need to be identified on postjunctional sites that, when occupied, lead to changes in postjunctional activity.
4. A transport system needs to be present for the substance itself or its breakdown products, uptake of which leads to renewal of messenger storage in nerve terminals.

It has been particularly difficult to establish cotransmitter roles for the many peptides found in nerves, partly because specific receptors and physiological roles have not been established for some of these and partly because of the lack of selective antagonists. In some enteric neurons, up to five neuropeptides have been identified. It is important to distinguish between neuromodulator, neurotransmitter, and neurotrophic roles for released peptides.

It is becoming clear that ATP is a primitive signaling molecule that has been retained as a cotransmitter in every nerve type in both the peripheral and central nervous systems, although the relative role of ATP varies considerably in different species and pathophysiological conditions. ATP appears to become a more prominent cotransmitter in stress and inflammatory conditions. Most nerves contain and release ATP as a fast cotransmitter together with classical transmitters such as ACh, NA, glutamate, GABA, and one or more peptides. In view of this, it is recommended that the terms 'adrenergic,' 'cholinergic,' 'peptidergic,' 'purinergic,' 'aminergic,' and 'nitrergic' not be used when nerves are described, although adrenergic, cholinergic, peptidergic, purinergic, aminergic, or nitrergic transmission is still meaningful.

See also: Acetylcholine Neurotransmission in CNS; Adenosine Triphosphate (ATP); Adenosine Triphosphate (ATP) as a Neurotransmitter; Gamma-Aminobutyric Acid (GABA); Glutamate; Neutrotransmission and Neuromodulation: Acetylcholine; Noradrenaline; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation.

Further Reading

Burnstock G (1976) Do some nerve cells release more than one transmitter? *Neuroscience* 1: 239–248.

Burnstock G (1990) Co-transmission. *Archives Internationales de Pharmacodynamie et de Therapie* 304: 7–33.

Burnstock G (1996) Cotransmission with particular emphasis on the involvement of ATP. In: Fuxe K, Hökfelt T, Olson L, Ottoson D, Dahlström A, and Björklund A (eds.) *Molecular Mechanisms of Neuronal Communication. A Tribute to Nils-Ake Hillarp*, pp 67–87. Oxford: Pergamon Press.

Burnstock G (1999) Purinergic cotransmission. *Brain Research Bulletin* 50: 355–357.

Burnstock G (2003) Purinergic receptors in the nervous system. In: Schwiebert EM (ed.) *Current Topics in Membranes, vol. 54: Purinergic Receptors and Signalling*, pp. 307–368. San Diego, CA: Academic Press.

Burnstock G (2004) Cotransmission. *Current Opinion in Pharmacology* 4: 47–52.

Burnstock G (2007) Physiology and pathophysiology of purinergic neurotransmission. *Physiological Reviews* 87: 659–797.

Fujii S, Kato H, and Kuroda Y (2002) Cooperativity between extracellular adenosine 5′-triphosphate and activation of *N*-methyl-D-aspartate receptors in long-term potentiation induction in hippocampal CA1 neurons. *Neuroscience* 113: 617–628.

Furness JB, Morris JL, Gibbins IL, et al. (1989) Chemical coding of neurons and plurichemical transmission. *Annual Review of Pharmacology and Toxicology* 29: 289–306.

Hökfelt T, Elfvin LG, Elde R, et al. (1977) Occurrence of somatostatin-like immunoreactivity in some peripheral sympathetic noradrenergic neurons. *Proceedings of the National Academy of Sciences of the United States of America* 74: 3587–3591.

Jo YH and Role LW (2002) Coordinate release of ATP and GABA at *in vitro* synapses of lateral hypothalamic neurons. *Journal of Neuroscience* 22: 4794–4804.

Kupfermann I (1991) Functional studies of cotransmission. *Physiological Reviews* 71: 683–732.

Lundberg JM (1996) Pharmacology of cotransmission in the autonomic nervous system: Integrative aspects on amines, neuropeptides, adenosine triphosphate, amino acids and nitric oxide. *Pharmacological Reviews* 48: 113–178.

Merighi A (2002) Costorage and coexistence of neuropeptides in the mammalian CNS. *Progress in Neurobiology* 66: 161–190.

Nusbaum MP, Blitz DM, Swensen AM, et al. (2001) The roles of co-transmission in neural network modulation. *Trends in Neurosciences* 24: 146–154.

Sneddon P and Burnstock G (1984) Inhibition of excitatory junction potentials in guinea-pig vas deferens by α,β-methylene-ATP: Further evidence for ATP and noradrenaline as cotransmitters. *European Journal of Pharmacology* 100: 85–90.

Wu LJ, Li Y, and Xu TL (2002) Co-release and interaction of two inhibitory co-transmitters in rat sacral dorsal commissural neurons. *Neuroreport* 13: 977–981.

Cranial–Cervical Dystonia

C Kenney and J Jankovic, Baylor College of Medicine, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Definitions

Dystonia is a neurologic syndrome characterized by involuntary, sustained (tonic) or spasmodic (rapid or clonic), patterned, repetitive muscle contractions, which frequently cause twisting (e.g., torticollis), flexing or extending (e.g., writer's cramp, retrocollis), and squeezing (e.g., blepharospasm) movements or abnormal postures. Dystonia can involve skeletal muscles in any part of the body; distribution of symptoms may be focal, segmental, multifocal, unilateral, or generalized. Characteristic features include co-contraction of agonist/antagonist muscles and involuntary overflow contraction into adjacent muscle groups. Dystonic movements are typically action induced, although they may also occur at rest, particularly when severe or after peripheral injury.

Blepharospasm, an involuntary closure of the eye produced by spasmodic contractions of the eyelids and eyebrows (affecting the orbicularis oculi and corrugator muscles), is an example of focal dystonia, occurring particularly in adult women. An increased frequency of blinking, associated with a feeling of irritation or dryness in the eyes, usually precedes the sustained closure of the eyes. Only about 20% of patients with blepharospasm continue to have isolated (focal) blepharospasm, called essential blepharospasm. In the majority of patients, blepharospasm is associated with dystonic movements in the adjacent facial, oromandibular, laryngeal, and neck muscles. The disorder may vary from a slightly annoying condition to a disabling one that interferes with daily activities, such as reading, watching television, and driving. Bright light usually exacerbates blepharospasm. As a result, many patients wear sunglasses both outside and inside. The spasms may be transiently alleviated by pulling on an upper eyelid, pinching the neck, talking, humming, yawning, singing, sleeping, relaxing, reading, concentrating, looking down, and other maneuvers or sensory tricks ('geste antagonistique'). Blepharospasm may be confused with tic disorders or ptosis.

Oromandibular dystonia consists of involuntary spasms of jaw, mouth, and tongue muscles that produce trismus (jaw clenching); bruxism (tooth grinding); involuntary tongue movements; and opening, closure, or deviation of the jaw. These symptoms may interfere with chewing, speaking, and swallowing. Secondary complications of oromandibular dystonia include dental wear and temporomandibular joint (TMJ) syndrome. Involuntary contractions of the mouth and jaw muscles can cause disabling dysarthria and chewing difficulties, as well as severe discomfort and social embarrassment. Some patients with oromandibular dystonia have involvement of the pharyngeal and laryngeal musculature, resulting in dysphagia, strained voice (spasmodic dysphonia), respiratory difficulties, and involuntary vocalizations, such as humming, grunting, belching, and gasping. As in other forms of dystonia, alleviation of oromandibular dystonia may be attained by various sensory tricks (geste antagonistique), including touching the lips or chin, chewing gum, biting on a toothpick, talking, and applying pressure to the submental area. With this phenomenon unique to dystonia, touching a nearby body part ameliorates the abnormal posture temporarily.

Spasmodic dysphonia presents typically as focal dystonia induced by speech. Adductor dysphonia leads to overcontraction of the thyroarytenoid muscles causing a strained pattern of speech. Less commonly, abductor dysphonia causes dystonic contraction of the posterior cricoarytenoid muscles, resulting in a whispering or breathy voice. Spasmodic dysphonia may present in isolation or as part of a more diffuse dystonic disorder.

Cervical dystonia is a focal dystonia characterized by sustained, involuntary contractions of neck muscles that result in abnormal movements and postures of the head. Onset occurs typically between the third and fifth decade of life. The term spasmodic torticollis is commonly used to describe this form of dystonia. However, since the movement is not always spasmodic and may consist of abnormal postures other than torticollis (e.g., turning of the head), such as laterocollis (head tilt), antercollis (neck flexion), and retrocollis (neck extension), the term cervical dystonia is preferred as a generic descriptor of dystonic movements or postures involving the neck. Categorized as a focal dystonia, cervical dystonia is often seen in association with other movement disorders, such as dystonia of adjacent (and other) body parts and tremor. Head oscillation caused by cervical dystonia can be divided into two broad categories: essential tremor and dystonic tremor. Essential tremor usually has a higher frequency (>7 Hz) than the dystonic tremor (3–7 Hz). Dystonic tremor, a component of the dystonia, is most obvious when the patient attempts to maintain primary position of the head by resisting the dystonic 'pulling.' About

25% of patients with cervical dystonia have postural tremor in their hands that is phenomenologically identical to essential tremor, in addition to head tremor. A majority of patients (75%) note neck pain as a symptom of cervical dystonia.

The combination of blepharospasm, oromandibular dystonia, and cervical dystonia is sometimes referred to as Meige syndrome, but the term-cranial–cervical dystonia is more appropriate to describe this form of segmental dystonia. In most series of patients with cranial–cervical dystonia, women outnumber men by a ratio of 2:1, and the onset is usually in the sixth decade of life. The disorder is a lifelong condition, and spontaneous remissions are rare.

Causes

Since the muscle spasms of cranial–cervical dystonia are typically exacerbated by stress and relieved by relaxation, as well as by various sensory 'tricks,' the disorder is often incorrectly presumed to be mental or psychological. This misconception is fostered by the lack of diagnostic imaging or other laboratory tests for abnormalities, as the vast majority of these patients have 'idiopathic' dystonia.

Probably the second most common cause of cranial–cervical dystonia is tardive dystonia, which is similar to the more common tardive dyskinesia because it is caused by drugs that block dopamine receptors, such as the major tranquilizers and certain antiemetic drugs. In contrast to orofacial and other movements that are typically present in tardive dyskinesia, the movements in tardive dystonia are more sustained and resemble those of idiopathic dystonia. Patients treated for Parkinson's disease with levodopa may also experience cranial–cervical dystonia. Structural lesions in the upper brain stem, diencephalon, and basal ganglia, resulting from stroke, multiple sclerosis, thalamotomy, hydrocephalus, and other causes, have been reported to cause cranial–cervical dystonia. In addition, cranial–cervical dystonia can be seen in patients with Wilson disease and other metabolic and neurodegenerative disorders. Finally, trauma to the eyelids, jaw, or neck can trigger dystonic movements in the injured area, so-called peripherally induced dystonia. One example of peripherally induced movement disorder is hemifacial spasm. This form of segmental myoclonus must be differentiated from cranial–cervical dystonia. Hemifacial spasm consists of involuntary, recurrent twitches of the eyelids and perinasal, perioral, zygomaticus, platysma, and other muscles of one side of the face. The condition may be mild, causing only social embarrassment or psychological distress, but in some patients, it causes unilateral blepharospasm that can interfere with vision. Though rarely familial, hemifacial spasm is an example of a peripherally induced movement disorder in which the muscular contractions result from a compressive or irritative lesion of the ipsilateral facial nerve. Facial and jaw spasms, called oculomasticatory myorhythmia, can be a complication of central nervous system (CNS) involvement in Whipple disease. The rhythmic contractions of the eyelids, face, and mouth may be synchronous with convergent eye oscillations. Hemimasticatory spasm, consisting of unilateral jaw closure spasms and ipsilateral facial hemiatrophy, is another facial movement disorder that is clinically similar to cranial–cervical dystonia and hemifacial spasm.

Pathogenesis and Pathophysiology

Dystonia is characterized by impaired inhibition at multiple levels of the CNS. Overflow movements in adjacent muscles are thought to derive from a failure to suppress neuronal excitability in surrounding neural circuits (surround inhibition). No specific abnormality or lesion has been identified on brain imaging studies or at autopsy of the brain of patients with cranial–cervical dystonia. Electrophysiologic studies of patients with cranial–cervical dystonia have demonstrated abnormalities in the R2 component of the blink reflex, indicating increased excitatory of the brain stem interneuronal pathway. Presumably, impaired descending control of blink circuits is related to dysfunction of the cortico-striato-thalamo-cortical circuits. Functional magnetic resonance imaging (fMRI) studies show deficient activation of the primary motor and ventral premotor cortex along with overactivity in the somatosensory cortex and caudal supplementary motor area. Interestingly, botulinum toxin (BTX) partially reverses the overactivity of the somatosensory cortex and caudal supplementary motor area after treatment, exerting its clinical effects peripherally and centrally. Dystonia is generally regarded as a motor disorder, but the aforementioned overactivity of the somatosensory cortex implicates the sensory system as well. Furthermore, patients with blepharospasm often complain of sensory symptoms such as dry or itchy eyes which improve after BTX injections. The importance of tactile and proprioceptive input can be demonstrated at the bedside in dystonia using a sensory trick. Existing clinical and experimental lends support to the idea that sensorimotor integration is impaired in dystonia.

Clinical, pharmacologic, and biochemical studies have suggested dopaminergic and noradrenergic preponderance in some cases. This biochemical abnormality may be genetically determined, since family history of a movement disorder is present in at least one-third of patients with cranial dystonia. A familial form of young-onset cervical dystonia with a dramatic response to levodopa has been reported,

but the genetic mutation remains undiscovered. Adult-onset cervical dystonia may occur in families with the DYT7 locus localized to the short arm of chromosome 18 (18p) by linkage analysis; however, the molecular basis is poorly understood. Dystonia may also occur in individuals with deletions of the entire short arm of chromosome 18, thereby implicating a host of novel genes with unknown function and many others involved in important cellular pathways: GNAL, CIDEA, NDUFV2, TGIF, DAP-1, and AFG3L2.

Treatment

While the cause of cranial–cervical dystonia is still poorly understood, the treatment of dystonias has markedly improved in the past decade. Investigation should focus on potentially treatable causes of dystonia. If none can be identified, then strategies designed to provide symptomatic relief without side effects should be discussed with the patient. About one-third of patients benefit from some pharmacologic therapy with agents such as anticholinergics, benzodiazepines, or muscle relaxants (baclofen). Dystonic patients usually require high doses of anticholinergic treatment (trihexyphenidyl 6–80 mg/day) for noticeable improvement. Trihexyphenidyl should be administered at 1 mg/day and slowly titrated by 2 mg/week using a BID dosing schedule. The pediatric population tolerates these doses more readily than adults. Side effects may limit aggressive titration: dry mouth, blurred vision, and confusion. Baclofen reduces spinal cord interneuron firing by activation of presynaptic γ-aminobutyric acid (GABA) receptors. As with trihexyphenidyl, slowly titrated large doses may be required for amelioration of dystonia. Side effects include sedation, dizziness, and dry mouth. Tizanidine, a centrally acting muscle relaxant, improves dystonia and spasticity by stimulating noradrenergic α-2 receptors. Longer acting benzodiazepines such as clonazepam provide symptomatic relief with less concern of addiction. In some patients, tardive dystonia responds well to tetrabenazine, a dopamine-depleting drug. Dose-limiting side effects include sedation, akathisia, depression, and parkinsonism. Finally, reports exist of familial cervical dystonia that responds favorably to levodopa.

The most important advance in the symptomatic treatment of blepharospasm and other focal dystonias has been the introduction of BTX. Seven serotypes (A–G) exist in nature, two of which are available as commercial preparations (A and B). An injection of this complex protein into the contracting muscles inhibits acetylcholine release at the neuromuscular junction, resulting in a satisfactory, albeit temporary, relief of the dystonic spasms in a majority of patients. Specifically, BTX type A cleaves synaptosome-associated protein (SNAP-25), a protein required for synaptic fusion with the presynaptic membrane and eventual release of acetylcholine into the neuromuscular junction. A placebo-controlled trial, reported in 1987, provided support for and validation of encouraging reports from various open trials. In subsequent studies, BTX injections were reported to produce moderate or marked improvement in spasms and function in over 90% of patients with blepharospasm and cervical dystonia. The average latency from the time of the injection to the onset of improvement is 2–7 days, and the average duration of benefit is 3–4 months. A meaningful reduction in the mouth, jaw, and tongue spasms and an improvement in chewing and speech were noted in more than 70% of patients with oromandibular dystonia treated with BTX. Patients with dystonic jaw closure responded better than those with jaw-opening dystonia. Temporary swallowing problems, noted in about 10–15% of patients, were the most common complication of injections into the masseter, submental, and other oromandibular muscles. Other common side effects include excessive weakness or pain at the site of injection. BTX injections provide the most effective relief of oromandibular dystonia, and early treatment may prevent dental and other complications, including TMJ syndrome and oral and dental problems associated with oromandibular dystonia. Patients with jaw closure dystonia generally benefit from injections into the masseter and, occasionally, the temporalis muscles, and those with jaw-opening dystonia benefit from submental or lateral pterygoid injections. Patients with blepharospasm should be injected into the pretarsal portion of orbicularis oculi rather than the preseptal region to prevent ptosis. Adjacent muscles may also be involved in blepharospasm such as corrugator supercilii or procerus. Injections for blepharospasm can lead to dry eyes, tearing, or diplopia. A small subset of patients treated with BTX develop antibodies against BTX and no longer benefit from subsequent injections. Factors that predispose patients to antibody formation include large doses injected in intervals shorter than 3 months.

Treatment of cranial–cervical dystonia requires a detailed knowledge of the local anatomy and the individual's condition. When skilled clinicians use BTX therapy, side effects should occur less frequently, that is, in less than 15% of all treatment sessions. The potential side effects are related to the site of injection: an injection in the eyelids may cause ptosis, blurred vision, diplopia, tearing, or eyelid hematoma; an injection into the oromandibular muscles may cause chewing, speaking, and swallowing difficulties; and an injection into the cervical muscles can produce neck weakness and swallowing problems. Function is

impaired in less than 2% of the patients who experience these complications. Furthermore, the complications usually improve spontaneously within 3 weeks. The functional improvement, experienced by nearly all patients after a BTX injection, is difficult to express numerically. Many patients could not work, drive, watch TV, or read before the injections; after the injections, most can function normally as a result of reduced eyelid and eyebrow spasms. Similarly, after BTX therapy, patients with oromandibular dystonia can speak, chew, and swallow more easily, and those with cervical dystonia usually experience improved control of their head position and marked reduction in neck pain. BTX injection is now considered by many physicians as the treatment of choice for blepharospasm, oromandibular dystonia, cervical dystonia, and spasmodic dysphonia. While surgical microvascular decompression of the facial nerve has a high rate of success in the treatment of hemifacial spasm, this treatment is associated with potentially serious risks, such as permanent facial paralysis, deafness, stroke, and death. Therefore, a local injection of BTX into involved facial muscles offers a useful alternative to surgical therapy. Nearly all patients improve; the complications, consisting of mild facial asymmetry, are minimal and transient. The approach can be individualized by injecting only those muscles in which contractions are most disturbing to the patient. In the authors' experience with treating a large number of patients with hemifacial spasm, the average duration of improvement was 5 months, longer than in any of the other spasmodic disorders.

For patients unresponsive to conventional pharmacologic intervention, options include intrathecal baclofen, selective peripheral denervation, and deep brain stimulation (DBS). DBS is most appropriate for patients with disabling generalized dystonia rather than focal or segmental dystonia. To date, most investigators advocate targeting the globus pallidus interna (GPi) for the treatment of medically refractory dystonia. Patients with DYT1 dystonia appear to have the most robust response. Case reports and case series of cervical dystonia and cranial–cervical dystonia look promising, though reports of long-term follow-up are lacking. In contrast to essential tremor, the postoperative improvement of GPi-DBS may be delayed several months. Phasic dystonic movements tend to improve more rapidly than tonic movements. DBS parameters differ in dystonia when compared to essential tremor or Parkinson's disease in that the pulse width and voltage requirements are typically higher. DBS-related complications when targeting the GPi are infrequent but include intracerebral hemorrhage, infection, dysarthria, weakness, and numbness.

See also: Botulinum and Tetanus Toxins; Deep Brain Stimulation; Deep Brain Stimulation and Movement Disorder Treatment; Dopaminergic Agonists and L-DOPA; Dystonia: Myoclonus–Dystonia; Functional Imaging of the Motor System; Segawa Dopa Responsive Dystonia; Torsion Dystonia.

Further Reading

Bhidayasiri R and Tarsy D (2006) Treatment of dystonia. *Expert Review of Neurotherapy* 6: 863–886.

Bittar RG, Yianni J, Wang S, et al. (2005) Deep brain stimulation for generalized and spasmodic torticollis. *Journal of Clinical Neuroscience* 12: 12–16.

Defazio G, Brancati F, Valente EM, et al. (2003) Familial blepharospasm is inherited as an autosomal dominant trait and relates to a novel unassigned gene. *Movement Disorders* 18: 207–212.

Defazio G and Livrea P (2002) Epidemiology of primary blepharospasm. *Movement Disorders* 17: 7–12.

Forget R, Tozlovanu V, Iancu A, and Boghen D (2002) Botulinum toxin improves lid opening in blepharospasm-associated apraxia of lid opening. *Neurology* 58: 1843–1846.

Geyer HL and Bressman SB (2006) The diagnosis of dystonia. *Lancet Neurology* 5: 780–90.

Hallett M (2002) Blepharospasm: Recent advances. *Neurology* 59: 1306–1312.

Horwath-Winter J, Bergloeff J, Floegel I, Haller-Schober EM, and Schmut O (2003) Botulinum toxin: A treatment in patients suffering from blepharospasm and dry eye. *British Journal of Ophthalmology* 87: 54–56.

Houser M and Waltz T (2006) Meige syndrome and pallidal deep brain stimulation. *Movement Disorders* 20: 1203–1205.

Jankovic J (2001) Can peripheral trauma induce dystonia and other movement disorders? Yes! *Movement Disorders* 16: 7–12.

Jankovic J (2006) Botulinum toxin therapy for cervical dystonia. *Neurotoxicity Research* 9: 145–148.

Jankovic J (2006) Treatment of dystonia. *Lancet Neurology* 5: 864–872.

Kenney C and Jankovic J (2006) Tetrabenazine in the treatment of hyperkinetic movement disorders. *Expert Review of Neurotherapy* 6: 7–17.

Kenney C, Simpson R, Hunter C, et al. (2007) Short-term and long-term safety of deep brain stimulation for the treatment of movement disorders. *Journal of Neurosurgery* 106: 621–625.

Kupsch A, Benecke R, Muller J, et al. (2006) Pallidal deep-brain stimulation in primary generalized or segmental dystonia. *New England Journal of Medicine* 355: 1978–1990.

Misbahuddin A, Placzek MR, Chaudhuri KR, Wood NW, Bhatia KP, and Warner TT (2002) A polymorphism in the dopamine receptor DRD5 is associated with blepharospasm. *Neurology* 58: 124–126.

Németh AH (2002) The genetics of primary dystonias and related disorders. *Brain* 125: 695–721.

Tarsy D and Simon DK (2006) Dystonia. *New England Journal of Medicine* 345: 1184–1192.

Relevant Websites

http://www.dystonia-foundation.org – Dystonia Medical Research Foundation.

http://www.wemove.org – Worldwide Education and Awareness for Movement Disorders.

Cross-Modal Interactions Between Vision and Touch

K Sathian and S Lacey, Emory University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Although this might seem heretical, it is now generally accepted that many extrastriate visual cortical areas are active during tactile perception, in both the sighted and the blind. Here we review the literature on such cross-modal recruitment of visual cortex during tactile perception. We consider the attempts that have been made to address a variety of issues that remain largely unresolved, including why such cross-modal recruitment of visual cortex occurs, whether the underlying reasons differ in the sighted and the blind, the nature of the underlying perceptual representations, and how they relate to performance differences between the blind and sighted.

Involvement of Visual Cortical Areas in Tactile Perception in the Sighted

Extrastriate visual cortical areas were first shown to be active during tactile perception in a positron emission tomography (PET) study conducted in our laboratory on normally sighted human volunteers. In this study, the experimental task was tactile discrimination of the orientation of gratings presented to the immobilized fingerpad. When this task was contrasted with a control task requiring tactile discrimination of grating groove width, activation was found at a focus in extrastriate visual cortex, close to the parieto-occipital fissure. This focus is in the vicinity of the human V6 complex of visual areas and had previously been found to be active during visual discrimination of grating orientation. The possibility that this cross-modal parieto-occipital activation might be merely epiphenomenal was ruled out in a subsequent study from our laboratory, using transcranial magnetic stimulation (TMS) to test whether disrupting the function of this cortical region was detrimental to performance. TMS applied directly over the parieto-occipital focus, and at sites close to it, significantly impaired discrimination of grating orientation when the magnetic pulses were appropriately timed with respect to the stimuli. However, TMS at or near the parieto-occipital focus did not affect discrimination of grating groove width, and when applied at more distant sites did not affect performance on either task. This pattern of results permitted the inference that the activation found in the earlier PET study was functionally meaningful.

Analogous findings were reported in a study using repetitive TMS (rTMS) at 1 Hz to decrease local cortical excitability. This study took advantage of a dissociation between subjective magnitude estimates of perceived interdot distance, which increase with physical interdot distance (up to 8 mm), and those of perceived roughness, which peak at intermediate values (around 3 mm). The main finding was that rTMS over the primary somatosensory cortex (S1) affected judgments of roughness but not interdot distance, whereas rTMS over the medial occipital cortex affected distance but not roughness judgments. Further, a congenitally blind patient who suffered a stroke that damaged the occipital cortex bilaterally was impaired on judgments of interdot spacing but not roughness. These observations, together with the studies of the grating orientation task reviewed in the previous paragraph, indicate that visual cortical involvement occurs in macrospatial (large-scale) tactile tasks but not in microspatial tasks in which the parameters of interest vary in the range of $\leq$3 mm. This conclusion is consistent with psychophysical studies suggesting that macrospatial tactile tasks are preferentially associated with visual processing and that vision seems to be better than touch for perceiving macrospatial features, the reverse being true for microspatial features.

Another example of visual cortical recruitment during tactile perception is offered by work showing that the MT complex, an area located in the dorsal (visuospatial) pathway and specifically involved in perception of visual motion, is also active during presentation of tactile motion stimuli, even in the absence of any task. Psychophysical studies suggest that both vision and touch may engage a common representation of motion; tactile motion perception can be used to disambiguate the direction of motion in an ambivalent visual motion display. Within the ventral visual pathway, which is specialized for form processing, the lateral occipital complex (LOC) is a region that is object-selective for visual stimuli. The LOC is also recruited cross-modally during haptic object recognition, as revealed by numerous functional magnetic resonance imaging (fMRI) studies. Further, activation has been reported in the LOC during tactile perception of two-dimensional forms presented to the immobilized right index fingerpad, and studies of patients with lesions involving the LOC suggest that it is necessary for haptic as well as visual shape perception. Multisensory object-selective activity in the LOC is stronger for graspable visual objects than for other visual stimuli and does not appear to

extend to the auditory modality. These studies, along with cross-modal priming effects observed in psychophysical and fMRI studies, cross-modal interference effects in memory tasks, and category-specific representations that overlap between vision and touch, imply that both visual and haptic shape perception may engage a unitary neural representation, as in the case of motion perception.

Visual Imagery and Multisensory Representations

The research reviewed so far has established that visual cortical processing is routinely involved in normal tactile perception in the sighted, especially during macrospatial tasks. Such processing is quite task-specific, so the extrastriate visual cortical areas known to be specialized for particular visual tasks are involved during performance of the same tasks tactually. An obvious question is whether visual imagery underlies this cross-modal recruitment of visual cortex into nonvisual tasks. Subjects in our laboratory consistently report mental visualization of tactile stimuli during macrospatial tasks such as discrimination of grating orientation or tactile form, which are associated with visual cortical recruitment, but not during microspatial tasks such as discrimination of grating groove width or gap detection, which do not tend to involve visual cortical activity. Visual imagery might be triggered by lack of familiarity with the tactile stimuli or tasks used; indeed, such cross-modal translation may be a general phenomenon, especially during complex information processing. Left-lateralized LOC activity has been found during mental imagery of shape generated from previous visual encounters in the sighted or from haptic encounters in the blind. An fMRI study from our laboratory found that inter-individual variations in the magnitude of haptic shape-selective activity in the right LOC (ipsilateral to the hand used for haptic perception) were strongly predicted by ratings of the vividness of visual imagery. In contrast, activation strengths in the left LOC were uncorrelated with these visual imagery scores, pointing to a role for factors other than visual imagery in cross-modal visual cortical recruitment. Despite the differences between these studies in the lateralization of imagery effects, they suggest that a role for visual imagery in cross-modal visual cortical recruitment cannot be ruled out.

However, some have argued that LOC activity during haptic perception is not due to visual imagery. Vision and touch can each encode object shape and motion, and visual cortical activation during tactile perception might reflect a multisensory representation rather than a specifically visual image, especially given the evidence reviewed earlier for common representations of motion and shape across vision and touch. That similar illusions can be experienced with both visual and tactile inputs support this idea. In the Müller–Lyer illusion, two lines of identical length and endcapped with fins that point either outward or inward are seen as shorter or longer, respectively, than the true length. This visual illusion has also been demonstrated in a haptic format and appears to be independent of visual experience because blindfolded, normally sighted individuals, those with low vision, late blind subjects, and congenitally blind individuals are all equally susceptible to the illusion. In the same vein, length distortion is an illusion that was first described haptically and subsequently found to occur in vision too. Length distortion refers to the overestimation of the distance between two points as a function of the increasing length of an indirect pathway between them. The same effect can be observed, albeit of a smaller magnitude, in the visual version of the task. These studies suggest a common visuo-haptic representation of space that may be based on body-centered reference points (e.g., the body midline or the positions of the hands relative to one another) because instructions to use these in judging Müller–Lyer line length virtually eliminated the illusion in both modalities whereas external reference points (a frame surrounding the illusion figure) had no effect. The use of these body-centered reference points may be a common mechanism for integrating spatial information from vision and touch.

The evidence for multisensory representations suggests the existence of direct, bottom-up somatosensory projections into the visual cortical areas that are implicated in tactile perception, as opposed to the top-down projections from the prefrontal into visual cortical areas that would be required to support a process such as visual imagery. These competing possibilities can be tested empirically by studying connectivity in experimental animals, by electrophysiological studies in animals or humans to pinpoint the timing of activity in different regions, or by analyzing fMRI data to reveal effective connectivity. A study from our laboratory using structural equation modeling of effective connectivity, based on the correlation matrix between the time courses of fMRI activity in various regions, revealed both bottom-up and top-down paths in a network of haptically shape-selective areas, including foci in the postcentral sulcus (corresponding to Brodmann's area 2 of S1), the intraparietal sulcus (IPS), and the LOC. This suggests that potential neural substrates exist for both visual imagery and multisensory representations, in relation to visual cortical involvement during haptic shape perception.

Neurophysiological and neuroanatomic studies in monkeys have also illuminated our understanding of the mechanisms involved in cross-modal visual cortical recruitment. Some neurons in area V4 (a nonprimary area in the ventral visual pathway) were found to be selective for the orientation of a tactile grating when it served as a cue for matching to a subsequently presented visual stimulus but not when it was task-irrelevant. The requirement for task-relevance implicates top-down, rather than bottom-up, inputs. Multisensory inputs have been reported in early sensory cortical areas that are traditionally considered unisensory, including the primary visual cortex (V1) and auditory association cortex. Analysis of the laminar profile of these projections suggests that both top-down and bottom-up inputs are present. Thus, the weight of evidence indicates the existence of multisensory representations that are flexibly accessible via both vision and touch, potentially involving interactions between bottom-up sensory inputs and top-down processes such as visual imagery. However, empirical studies addressing the format of such representations are in their infancy, and currently we cannot rule out as an alternative to unified multisensory representations that there might be independently derived spatial representations for vision and touch that can interact at later stages of processing (e.g., in cross-modal comparison).

If vision and touch produce separate representations, how do these affect one another? The classic work of Rock and Victor suggested that vision dominates touch; when the visual width of an object was distorted by a lens, individuals who both saw and felt the object were unaware of the cross-modal conflict and reported the distorted width perceived visually rather than the true width perceived tactually. Thus, this sensory conflict was resolved in favor of the visual percept. More recent research by Klatzky, Lederman, and their colleagues showed the situation to be more complex and to depend on the specific object property. For example, under neutral instructions the haptic salience of object properties is hardness > texture > shape; under haptically biased instructions this salience changes to texture > shape > hardness. But in combined visual and haptic perception, as well as in haptic perception with concurrent visual imagery, this saliency is reversed and changes to shape > texture > hardness/size. Visual imagery can therefore influence tactile perception in the absence of visual perception, and moreover, different modalities may be optimal for different object properties. Thus, there is not a simple dominance of visual over tactile perception.

Other approaches have also identified interactions between vision and somatosensory representations of the body. One line of work has demonstrated that noninformative vision of the body can improve tactile spatial acuity as measured using grating orientation discrimination, and that this cross-modal enhancement is abolished by TMS over the primary somatosensory cortex. Another line of research pertains to the rubber hand illusion – when a rubber hand is positioned in alignment with the subject's own concealed hand, if the subject sees the rubber hand being touched, he or she experiences illusory touch, indicating a feeling of ownership of the synthetic hand. Such a feeling of ownership is associated with activity in the ventral premotor cortex, a region where, in monkeys, neurons exhibit visual receptive fields that are anchored to the arm and move with it.

Tactile Perception in the Blind and Its Involvement of Visual Cortex

Despite the common belief that blindness is associated with superior nonvisual perception, research on this has not yielded uniform results. Some studies of haptic perception of three-dimensional shape have observed better performance in the blind, others have shown equal performance in sighted and blind individuals, and still others, even better performance in the sighted. It has been argued that superior performance of the blind on haptic tasks may stem not from heightened sensitivity but, rather, from using more efficient sensorimotor strategies or from practice in attending to cues that the sighted typically ignore. However, blind Braille readers out-perform the sighted on many tests involving tactile perception of two-dimensional patterns with the fingerpad. These include tactile recognition of Braille-like dot patterns, judging the orientation of bars or detecting gaps in them, discriminating grating orientation, and determining whether the central dot in a linear three-dot array is offset laterally by a small (<1 mm) distance. In contrast, blind individuals do not differ significantly from the sighted in their ability to discriminate bar length or laboratory textures such as gratings or real-world textures such as sandpapers or abrasive stones. There is some evidence that the superior tactile spatial performance of the blind (when it does occur) is due to practice. Thus, practiced sighted subjects do as well as the blind on the offset task and at discriminating patterns generated with the Optacon (a vibrotactile reading aid for the blind). They can also match the deaf-blind in decoding speech by the Tadoma method (feeling the speaker's face and neck). The role of specific practice with reading Braille is unclear; one study reported that the grating orientation discrimination threshold was lowest on the Braille-reading finger, whereas another study found no significant difference on this test between blind subjects who read Braille and

those who did not and that offset detection was no better on the hand used for reading Braille than the other hand.

Visual deprivation studies have investigated whether superior performance in the blind may be due to cortical reorganization. Following neonatal visual deprivation, rodents navigate a maze for a food reward faster than normal and demonstrate altered somatosensory cortical representations of the whiskers as well as novel somatosensory responsiveness in the anterior occipital cortex. In monkeys whose eyelids were sutured shut for the first year of life and opened later, the dorsal extrastriate visual cortex had lower visual but greater somatosensory responsiveness than in the controls, although this procedure did not affect haptic discrimination ability. The human somatosensory cortex also reorganizes after visual deprivation, probably as a consequence of increased hand use. Expansion of the representation of the Braille-reading finger has been reported in a study using evoked potentials and TMS. Others have found disordered cortical somatotopy as revealed by magneto-encephalography (MEG), and mislocalizations of touch on the reading fingers, in subjects who read Braille using multiple fingers. In an early report of cross-modal plasticity in blind humans, PET scanning was used to show that visual cortical areas in the occipital lobe were more active metabolically in the early blind than in the late blind or sighted. This was interpreted as an index of greater synaptic activity in the early blind, perhaps reflecting incomplete developmental pruning of synapses. Subsequently, many PET and fMRI studies found that the medial occipital cortical regions of blind subjects were recruited during tactile reading of Braille or embossed Roman letters. The medial occipital cortical activations were specific to early blind subjects, in contrast to late blind and sighted subjects, who deactivated these regions. Complementary rTMS studies demonstrated functional involvement of the medial occipital cortex in tactile identification of Braille or Roman characters by blind subjects but not in sighted subjects for Roman letters. As in the activation studies, the rTMS effects over the medial occipital cortex were found in the early blind but not the late blind, implicating cross-modal plasticity during a critical period of visual development. The functional role of the visual cortex in Braille reading is supported by the report of an early blind person who became unable to read Braille after a bilateral infarct of the occipital cortex.

Was the visual cortical activation observed during Braille reading in the blind due to sensory or higher-order cognitive processes such as attention and language? Most of the studies cited in the preceding paragraph could not resolve this because the tactile reading tasks were compared to low-level conditions. It turns out that visual cortical activity occurs in the blind during a variety of language tasks, including generating verbs in response to tactile or auditory input, and verbal recall in the absence of any sensory input. Such activity is more extensive in the early blind than the late blind. Its functional relevance is shown by its greater strength during semantic than phonological processing and by the correlation of its magnitude with semantic and syntactic complexity and with verbal recall ability.

The conclusion from these studies is that the visual cortex is clearly active during language processing in the blind. However, the extent of its involvement in tactile processing is still unclear, as is its potential relationship to performance differences between blind and sighted individuals. Compared to activity during Braille-reading, there was less activation in the medial occipital cortex of the blind during discrimination of angles or the width of grooves cut in homogeneous Braille fields. Vibrotactile frequency discrimination evoked extensive responses in the visual cortex in congenitally blind subjects, with less extensive responses in the late blind. Category-selectivity during haptic perception of three-dimensional form, similar to that noted in sighted subjects, was found in the inferotemporal cortex of blind subjects, although the category-selective voxels were located more ventrally in the blind than in sighted subjects. Unfortunately, most of these studies used rest controls, making it difficult to disentangle sensory from linguistic processing. A study of Braille-reading that did attempt to control linguistic processes, using an auditory word control, actually found medial occipital cortical activity in the late blind but not the early blind. Thus, more research is needed to address the extent to which tactile processing engages the visual cortex in the blind and its links with cross-modal recruitment in the sighted.

See also: Multisensory Convergence and Integration; Neural Circuitry in the Somatosensory System; Sensorimotor Integration: Models; Somatosensory Perception; Somatosensory Cortex: Functional Architecture; Somatosensory Receptive Fields; Somatosensory Cortex; Tactile Coding in Peripheral Neural Populations; Tactile Texture; Vision for Action and Perception; Visual Deprivation; Visual Cortical Models of Orientation Tuning.

Further Reading

Amedi A, von Kriegstein K, van Atteveldt NM, Beauchamp MS, and Naumer MJ (2005) Functional imaging of human cross-modal identification and object recognition. *Experimental Brain Research* 166: 559–571.

Balakrishnan JD, Klatzky RL, Loomis JM, and Lederman SJ (1989) Length distortion of temporally extended visual displays: Similarity to haptic spatial perception. *Perception and Psychophysics* 46: 387–394.

Botvinick M and Cohen J (1998) Rubber hands 'feel' touch that eyes see. *Nature* 391: 756.

Ehrsson HH, Spence C, and Passingham RE (2004) That's my hand!: Activity in premotor cortex reflects feeling of ownership of a limb. *Science* 305: 875–877.

Fiorio M and Haggard P (2005) Viewing the body prepares the brain for touch: Effects of TMS over somatosensory cortex. *European Journal of Neuroscience* 22: 773–777.

Heller MA, Brackett DD, Wilson K, Yoneyama K, and Boyer A (2002) The haptic Müller-Lyer illusion in sighted and blind people. *Perception* 31: 1263–1274.

Hollins M (1989) *Understanding Blindness*. Hillsdale, NJ: Lawrence Erlbaum.

Klatzky RL, Lederman SJ, and Reed C (1987) There's more to touch than meets the eye: The salience of object attributes for haptics with and without vision. *Journal of Experimental Psychology: General* 116: 356–369.

Merabet L, Thut G, Murray B, et al. (2004) Feeling by sight or seeing by touch? *Neuron* 42: 173–179.

Peltier S, Stilla R, Mariola E, LaConte S, Hu X, and Sathian K (2007) Activity and effective connectivity of parietal and occipital cortical regions during haptic shape perception. *Neuropsychologia* 45: 476–483.

Rock I and Victor J (1964) Vision and touch: An experimentally created conflict between the two senses. *Science* 143: 594–596.

Sathian K (2005) Cross-modal plasticity in sensory systems. In: Selzer ME, Clarke S, Cohen L, Duncan P, and Gage F (eds.) *Textbook of Neural Repair and Rehabilitation,* pp. 180–193. Cambridge, UK: Cambridge University Press.

Sathian K (2005) Visual cortical activity during tactile perception in the sighted and visually deprived. *Developmental Psychobiology* 46: 279–286.

Sathian K, Zangaladze A, Hoffman JM, and Grafton ST (1997) Feeling with the mind's eye. *NeuroReport* 8: 3877–3881.

Zangaladze A, Epstein CM, Grafton ST, and Sathian K (1999) Involvement of visual cortex in tactile discrimination of orientation. *Nature* 401: 587–590.

Cyclic AMP (cAMP) Role in Learning and Memory

T W Abrams, University of Maryland School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Historic Background

From the time of the anatomical discovery of synapses in the later nineteenth century, microscopists such as Tanzi speculated that synapses would be modified as a result of experience. It was further proposed by Donald Hebb that synaptic modification during learning would be initiated by appropriately timed neuronal activity. However, prior to the mid-twentieth century, there was insufficient information about cellular signaling mechanisms to permit the development of hypothetical proposals for how activity might modulate synaptic strength. This changed in the late 1950s with the identification of cyclic adenosine monophosphate (cAMP) by Earl Sutherland and Theodore Rall as the intracellular signal responsible for the modulatory effects of hormones such as epinephrine and glucagon. According to the classical model of cAMP signaling, receptors for hormones or neurotransmitters stimulate the enzyme adenylyl cyclase (AC), which synthesizes cAMP from adenosine triphosphate (ATP). cAMP exerts its modulatory effects primarily by activating cAMP-dependent protein kinase, known as protein kinase A (PKA), which transfers the tertiary phosphate group from ATP to specific serines or threonines in consensus sequences within target proteins. PKA was also identified in the 1950s by Edmond Fischer and Edwin Krebs. By stimulating the phosphorylation of target proteins, which vary depending on the cell-type, cAMP initiates modulatory effects that are cell-type specific. The understanding that cAMP regulates the activity of enzymes and other effector proteins by stimulating their phosphorylation gave rise to a general model of intracellular signaling. Although these early discoveries initiated a period of rapid progress in analyzing cAMP signaling, some of the other proteins that participate in the cAMP pathway would not be characterized for several decades.

The cAMP signaling model was highly influential. Neurobiologists interested in modulation of synaptic connections began to explore the possibility that cAMP might be involved in the experience-dependent alteration of synaptic strength. Strikingly, in the mid-1960s, Sutherland, Eduardo De Robertis, and their colleagues found that AC activity is enriched in brain homogenates that contain synaptic terminals, as is phosphodiesterase, the enzyme responsible for hydrolyzing cAMP and terminating its actions. Within a few years, Paul Greengard and colleagues had found that both PKA and phosphoproteins regulated by cAMP are also enriched in fractions containing presynaptic terminals. Greengard began to characterize a few of the most abundant cAMP-dependent phosphoproteins in synaptic terminals, including a presynaptic protein that was later named synapsin. During the same period, initial biochemical evidence suggested that in neurons, cAMP might mediate some slow modulatory effects of neurotransmitters.

In 1972, Seymour Kety published an essay discussing the possible involvement of noradrenergic signaling in associative learning, which he proposed is mediated by the modification of synaptic connections. He speculated that this noradrenergic modulation of synapses during learning might involve cAMP. However, it was not yet possible to conduct cellular studies on synaptic modulation in the mammalian brain. The first physiological evidence for the importance of cAMP at synapses came at the end of the 1960s from recordings at the mammalian neuromuscular junction. Analogs of cAMP or inhibitors of phosphodiesterase caused an increase in the amplitude of the synaptic response. In the 1970s, Eric Kandel, James Schwartz, and Vincent Castellucci studied short-term synaptic modulation during sensitization in the marine mollusk *Aplysia*; sensitization is a nonassociative form of behavioral plasticity in which a noxious stimulus increases the magnitude of a defensive withdrawal reflex. They showed that the noxious stimuli that caused behavioral sensitization increased cAMP levels in the central nervous system (CNS) and that this effect was mimicked by the modulatory neurotransmitter serotonin. Both serotonin and cAMP increased the strength of the direct synaptic connections between sensory neurons and motor neurons in the circuit for the defensive withdrawal reflex. Collaboratively with Paul Greengard, they demonstrated that this strengthening of the synaptic connections of sensory neurons to motor neurons was mediated by PKA.

It is noteworthy that the groundbreaking studies of the cAMP signaling pathway were recognized by three independent Nobel Prizes: for the discoveries of cAMP and the mechanism of hormone action (to Sutherland in 1971), for discoveries concerning protein phosphorylation (to Fischer and Krebs in 1992), and for the discovery of G-proteins, which were first identified in studies of the regulation of AC (to Gilman and Rodbell in 1994). In 2000, Greengard and Kandel received the Nobel Prize for physiology or medicine (which was shared with Carlsson) for research on signal

transduction in the nervous system; a substantial portion of the seminal work by these two investigators involved their studies of cAMP-dependent signaling at synapses.

This article summarizes the present understanding of the molecular mechanisms by which AC, cAMP, and PKA act to modulate synapses during learning. Before synaptic plasticity is discussed, the central elements of the cAMP signaling system are briefly reviewed.

The cAMP Signaling Pathway

In animals, AC enzymes contain two transmembrane domains, each with six membrane-spanning α helices. Between the two sets of six α helices is a large C1 cytoplasmic loop. A second cytoplasmic domain (C2) forms the C-terminus. Enzymatic activity of AC, catalyzing the synthesis of cAMP from ATP, requires both of these cytoplasmic domains. In mammals there are nine isoforms of membrane-bound AC. In addition, there is a cytoplasmic AC isoform that may be important in neuronal development. Metabotropic receptors that are positively coupled to AC interact with the enzyme via stimulatory, guanine nucleotide-binding G-proteins (G_s; known generally as G-protein-coupled receptors). These heterotrimeric G regulatory proteins consist of an α subunit and tightly associated $\beta\gamma$ subunits. In the inactive state, guanosine diphosphate (GDP) is bound to $G_{s\alpha}$, and $G_{s\alpha}$ is associated with $G_{\beta\gamma}$. When the AC-coupled receptor binds to a transmitter, guanosine triphosphate (GTP) is exchanged for GDP, and $G_{s\alpha}$ dissociates from $G_{\beta\gamma}$. The free GTP-bound G_s now stimulates AC. AC isoforms are also regulated by specific isoforms of $G_{\beta\gamma}$; whether the regulation is stimulatory or inhibitory depends on the specific type of AC. Moreover, several AC isoforms are inhibited by the α subunit of the inhibitory G-protein, G_i.

cAMP directly activates certain ion channels and also interacts directly with some other signaling proteins, such as specific GTP exchange factors, which facilitate the activation of G-proteins by GTP. However, most cAMP effects are mediated via PKA. PKA is a tetramer complex with two catalytic (C) subunits and two regulatory (R) subunits. In the tetramer, the R subunits inhibit the catalytic site of the C subunits. When cAMP concentrations are elevated, two cAMP molecules bind to each R subunit, which causes the dissociation of the C subunits. There are two classes of regulatory subunit, each of which has two isoforms: RIα, RIβ, RIIα, and RIIβ. There are two isoforms of the catalytic subunit, Cα and Cβ. The specificity of action of PKA depends on the substrate proteins expressed in different cells and the localization of PKA (discussed in the section titled 'Localization of PKA via A kinase anchoring proteins is necessary for cAMP-dependent synaptic plasticity').

Interactions among Intracellular Signaling Systems

cAMP is traditionally considered the prototypical intracellular messenger, in part because cAMP was the first intracellular messenger to be recognized. However, by the 1950s, another cytoplasmic signal, calcium, was known to be important for initiating key physiological processes in excitable cells. The earliest evidence for the possible role of calcium in triggering muscle contraction was obtained years before the identification of cAMP. Moreover, it had also begun to be appreciated that calcium plays a central role in both triggering and regulating the release of neurotransmitter at the synapse. By the 1960s, it was generally believed that cAMP was a universal intracellular messenger, whereas calcium was thought to be involved in a few specific roles, and only in excitable cells. It is now understood that calcium is also a ubiquitous intracellular signal. It is interesting that some of the earliest evidence suggesting that calcium is an important intracellular signal that participates in diverse cellular mechanisms was obtained in studies of the cAMP pathway. Particular forms of phosphodiesterase (the enzyme that degrades cAMP) were found to be activated by calcium. In the early 1970s, it was discovered that this activation by calcium occurred via a small, 17 kDa, heat-stable calcium regulator protein, which had been purified from brain. This calcium regulator protein, later named calmodulin, is highly conserved throughout evolution. Calmodulin binds four calcium ions; in the calcium-bound state, calmodulin activates a wide variety of calmodulin-dependent proteins. By the mid-1970s, calcium–calmodulin was found to stimulate some forms of AC. It is now widely appreciated that in general, multiple signaling pathways interact within single cells, as occurs with the calcium–calmodulin and the cAMP pathways. (Yet another example of interactions between signaling cascades, involving PKA and mitogen-activated protein (MAP) kinase, is discussed in the section titled 'How does the cAMP pathway initiate synaptic modifications during learning?'.) Although the calcium–calmodulin-sensitive AC was first identified in mammalian brain, its central role in synaptic plasticity and learning in mammals was not appreciated until after this AC isoform had been demonstrated to be involved in learning in invertebrates.

Genetic Evidence for the Role of cAMP in Learning

During the 1970s, a number of learning mutants were isolated in *Drosophila* by screening for deficits in an olfactory conditioning task. The first mutant that was successfully analyzed at the biochemical level was *dunce*. The *dunce* gene codes for an isoform of phosphodiesterase, and the learning deficit in the mutant appears to result from the tonic elevation of cAMP. When cAMP is tonically increased, cAMP modulatory effects may be chronically activated, occluding any modulatory changes initiated during learning. In addition, the tonically activated cAMP pathway in *dunce* has been shown to result in altered synaptic physiology in these flies and altered synaptic morphologies. Thus in *dunce*, developmental changes and changes in synaptic physiology may also contribute to the learning deficit.

The Role of Calmodulin-Sensitive AC

In the early 1980s, a second *Drosophila* learning mutant, named *rutabaga*, was found to have a lesion in calmodulin-sensitive AC. In *Aplysia*, calmodulin-sensitive AC, which is also regulated by the serotonin receptor and G_s, was proposed to act as an associative integrator during classical conditioning of the defensive withdrawal reflex. Mammals express nine types of membrane-bound AC; two of these expressed in the brain, AC1 and AC8, are stimulated by calcium–calmodulin. The cloning of calmodulin-sensitive ACs from mammals has allowed for exploration of the roles of these specific AC isoforms in learning. Daniel Storm and colleagues have examined the learning deficits resulting from knockout of AC1 and AC8. When placed in a bright chamber and given the choice, mice normally move into a darkened chamber. Pairing of shock with entry into the darkened chamber results in inhibition of the normal behavioral preference; that is, the animals learn to avoid the shock, which is known as inhibitory avoidance learning. In a mutant mouse in which the genes for both AC1 and AC8 were knocked out, inhibitory avoidance learning was normal when tested at 5 min after training; however, in the double-knockout mouse, memory at 30 min, 1 day, and 8 days was greatly impaired. These double-knockout mice were also tested in a fear-conditioning paradigm, in which animals learn to fear either a distinctive context or a conditioned stimulus (such as a tone) that is paired with a shock. In context-dependent conditioning, mice that are exposed to a shock delivered in a distinctive conditioning chamber display freezing behavior when they return to that chamber. This type of learning about the context involves the hippocampus, as well as the amygdala. The double-knockout mice showed normal context-dependent conditioning at 24 h but displayed reduced freezing compared with wild-type mice when tested at 8 days. In cue-dependent fear conditioning, when a shock is paired with a tone, animals learn to freeze on presentation of the tone. Cue-dependent conditioning involves the amygdala but not the hippocampus. Cue-dependent fear conditioning was normal in the double-knockout mice at both 24 h and 8 days. As in the fly mutants, it is difficult to know whether the impairment in learning is not at least partially due to developmental abnormalities when synapses form in the absence of normal cAMP signaling. Hippocampal synapses appeared normal in terms of synaptic strength and very shortterm synaptic plasticity, but certainly more subtle changes could have contributed to a learning deficit. Storm and colleagues also evaluated transgenic mice that overexpress AC1 in a memory task in which mice pay particular attention to novel objects. Compared with wild-type mice, mice overexpressing AC1 as adults showed improved memory, that is, enhanced preference for the novel object, at both 1 day and 5 days. In contrast, the double-knockout mice lacking AC1 and AC8 showed no memory for the novel object when tested at 1 h. Based on the deficits in the double-knockout mice, calmodulin-sensitive AC appears to be important for memory as early as 1 h after the initial exposure to objects, but not in the first 5 min.

The evidence from *Drosophila* and mouse strongly implicates calmodulin-sensitive AC in learning. However, the importance of convergent activation of calmodulin-sensitive AC is still uncertain in these two systems, and in *Aplysia*. In particular, in mammalian brain, much of the activation of AC1 appears to be mediated via calcium–calmodulin, and the role of any dual activation involving neurotransmitter remains unclear.

cAMP-Dependent Signaling Plays a Central Role in Long-Term Learning

Memory mechanisms must function over dramatically different time periods. Short-term memory lasts for minutes to tens of minutes. In contrast, long-term memory lasts hours to days or even months or years. These two forms of memory are mechanistically distinct in that long-term memory requires gene transcription and protein synthesis, whereas short-term memory does not. Undoubtedly, there are multiple molecular mechanisms with different durations that overlap to provide a continuous, relatively stable memory.

One approach to addressing the question of the role of cAMP in learning is to study transgenic mice

that express dominant negative R subunits of PKA in hippocampus and forebrain. Mice were tested with the Morris water maze spatial learning task. In this paradigm, mice are placed in a circular pool, and they learn to escape the water by finding a hidden platform located just below the surface. With repeated trials, the mice learn the location of the hidden platform using spatial cues. Training was conducted over 14 days, and the animals' searching behavior was tested on the following day, when the submerged platform was removed. Wild-type mice spent more than twice as much time searching in the quadrant where the platform was previously located as they spent searching in the other three quadrants; in contrast, mice expressing dominant negative R subunits showed little preference for the previous platform location. These dominant negative-PKA mice were also tested with fear conditioning. With cue-dependent fear conditioning, the amygdala-dependent task, no deficits were observed either 1 h or 24 h posttraining. With context-dependent fear conditioning, the dominant negative-PKA mice showed normal learning when tested 1 h after training but displayed a dramatic memory deficit when tested 24 h after training. Although these studies with transgenic mice did not implicate either calmodulin-sensitive AC or PKA in cue-dependent fear conditioning, studies by Joseph LeDoux and colleagues demonstrated that a PKA inhibitor delivered into the lateral amygdala blocked memory at 24 h but not at 1 h. Thus, cAMP-dependent signaling may also be involved in memory for amygdala-dependent fear conditioning. One general conclusion from these behavioral experiments on transgenic mice with deficits in the cAMP pathway is that AC and PKA are particularly important in longer-term memory. The precise time frame in which cAMP-dependent signaling is critical seems to vary depending on the learning task and also on the how the cAMP cascade is altered. Mechanisms by which the cAMP pathway modifies synaptic connections during activity-dependent plasticity are considered below.

How Does the cAMP Pathway Initiate Synaptic Modifications during Learning?

By the 1970s, it had been demonstrated that cAMP can cause strengthening of synapses, and one PKA-dependent, synapse-associated phosphoprotein, synapsin, had been identified. Nevertheless, understanding of the molecular mechanisms by which cAMP exerts its effects is still modest. The incompleteness of understanding of cAMP-dependent regulation of synaptic strength is at least partly because multiple mechanisms are involved, depending on the type of synapse. This heterogeneity of mechanisms of synaptic plasticity that are initiated by cAMP will be evident below.

In any analysis of an alteration in the strength of a synaptic connection, the fundamental initial question is whether the underlying change takes place presynaptically or postsynaptically. The former type of mechanism involves changes in calcium handling or the vesicle release apparatus, whereas the latter type involves changes in properties or numbers of receptors. The issue of pre- versus postsynaptic contributions has been analyzed in a number of examples of synaptic plasticity that are believed to occur during learning. (Synaptic plasticity is defined simply as a change in the strength of a synaptic connection.) Two of the most extensively studied systems are long-term potentiation of the synaptic connections between CA3 and CA1 pyramidal cells in hippocampus and serotonin-induced facilitation of synaptic connections of sensory to motor neuron in *Aplysia*. Both 'potentiation' and 'facilitation' refer to an enhancement of synaptic strength; in contrast, 'synaptic depression' refers to a decrease in synaptic strength. At the CA3-to-CA1 pyramidal cell synapse, most recent attention has focused on postsynaptic mechanisms of plasticity and primarily on pathways other than cAMP, particularly calmodulin-dependent kinase II, protein kinase C, and MAP kinase. At the *Aplysia* synapse of sensory to motor neuron, most analysis since the 1970s has focused on presynaptic mechanisms of facilitation, and the PKA pathway has received the greatest emphasis. This has led to the widespread impression that learning in mollusks utilizes presynaptic mechanisms of plasticity involving cAMP and that learning in mammals utilities postsynaptic mechanisms for altering synaptic strength that are largely independent of cAMP. This dichotomy reflects the history of the research on plasticity at the *Aplysia* synapse of sensory to motor neuron and the hippocampal CA3-to-CA1 synapse rather than fundamental differences between the phyla. Indeed, both within *Aplysia* and within mammalian CNS, learning involves a diverse array of pre- and postsynaptic mechanisms. (A substantial number of recent studies of *Aplysia* synapses have focused on the contribution of increases in postsynaptic calcium to the strengthening of these synaptic connections. These mechanisms are discussed here, as cAMP has not yet been implicated in these postsynaptic processes.) The diverse examples discussed in the next section illustrate some of the ways that cAMP contributes to plasticity during learning. It is also important to emphasize that this discussion of cAMP-initiated synaptic plasticity should not be interpreted as suggesting that other signaling pathways do not also play important roles in these same forms of plasticity.

Presynaptic Modulation by cAMP

Evidence for Effects of cAMP That Enhance Transmitter Release during Long-Term Potentiation at Mammalian CNS Synapses

Long-term potentiation (LTP) is a persistent strengthening of synaptic connections induced by a brief period of high-frequency presynaptic activity. Many central synapses display LTP, though the required patterns of stimulation for induction vary depending on the synapse. In the cerebellum, synapses between the granule cells and the Purkinje neurons undergo LTP when the granule cell axons, called parallel fibers, are stimulated at 2–8 Hz for $\geq$10 s. Modulation of the strength of these parallel fiber synapses is believed to contribute to at least some types of motor learning. Parallel fiber LTP is initiated by presynaptic calcium influx through voltage-activated calcium channels. cAMP is known to be involved because bath-applied inhibitors of PKA block LTP induction. Linden and Storm and their colleagues have analyzed the contribution of AC and cAMP to parallel fiber LTP. They found that knockout of calmodulin-sensitive AC1, which is expressed in cerebellum, completely eliminated parallel fiber LTP. To assess the relative contributions of pre- and postsynaptic PKA, inhibitors of the kinase were delivered pre- and postsynaptically. Unexpectedly, neither pre- nor postsynaptic delivery of PKA inhibitors blocked LTP. This suggested that inhibitors might not adequately diffuse from the soma where they were injected to the synaptic terminals. Subsequent experiments attempted to improve delivery by transiently expressing an endogenous PKA inhibitor protein, using a viral vector. Expression of the PKA inhibitor in the presynaptic granule cell, but not in the postsynaptic Purkinje neuron, blocked LTP induction. By the use of extracellular application of antibodies against the intravesicular domain of the synaptic vesicle protein synaptotagmin, it was possible to directly demonstrate that cAMP increases the rate of vesicle fusion. Together, this evidence suggests that activity in the presynaptic parallel fiber terminals, and the accompanying increase in intracellular calcium, results in calmodulin-dependent activation of AC1. The increase in cAMP somehow initiates facilitation of transmitter release.

A similar form of cAMP-dependent LTP occurs in hippocampus at the synapse between dentate granule cells and CA3 pyramidal cells. LTP, elicited by 1 s of 100 Hz stimulation of the dentate granule cells' axons, called mossy fibers, may also involve a postsynaptic contribution by PKA. However, there is substantial evidence that mossy fiber LTP can occur with a similar or identical mechanism as parallel fiber LTP in cerebellum. Mossy fiber LTP is largely eliminated in mutant mice in which AC1 has been knocked out. Presynaptic injection of a PKA inhibitor results in comparable elimination of most of the mossy fiber LTP. It should be pointed out that as of yet, no deficits in hippocampal-dependent learning have been identified that result from mutations that impair mossy fiber LTP. This lack of a deficit may be a consequence of which specific learning tasks are routinely evaluated rather than a reflection of a lack of relevance of plasticity at the mossy fiber synapses for behavioral plasticity in general. Other central synapses, including corticothalamic synapses, undergo LTP that appears to involve a similar presynaptic, PKA-dependent mechanism.

Presynaptic Targets of PKA That Contribute to LTP or Synaptic Facilitation

It has been difficult to identify the active-zone proteins that are the critical targets for PKA during facilitation of transmitter release. This is in part because the active zone includes more than two dozen proteins that participate in regulating synaptic vesicle trafficking, including docking (initial interaction of vesicles with proteins associated with the plasma membrane at the active zone), priming (the prefusion step that makes the vesicle competent for exocytosis), and fusion (the final step in transmitter release). Two proteins that have been implicated in PKA-dependent enhancement of transmitter release are RIM1α and synapsin.

Phosphorylation of RIM1α Plays a Necessary Role in PKA-Dependent LTP at Mossy Fiber and Parallel Fiber Synapses

Rab3A is a small G-protein that binds GTP. When in the GTP-bound state, Rab3A attaches to synaptic vesicles and influences the probability of vesicle release. Knockout of Rab3A had only subtle effects on transmitter release, perhaps because there are three Rab3 isoforms. The most striking phenotype that was observed in the Rab3A knockout mouse was that LTP is lacking at mossy fiber synapses. The importance of Rab3A in PKA-dependent mossy fiber LTP was initially puzzling because Rab3A is not a substrate for PKA. However RIM1α, with which Rab3A interacts, is phosphorylated by PKA. Like Rab3A, RIM1α is required for mossy fiber LTP. RIM1α is a scaffolding protein that interacts with a number of active-zone proteins, including munc13 and synaptotagmin, as well as Rab3A. RIM1α may participate in vesicle priming, and it also influences release probability. Knockout of RIM1α eliminates LTP at both mossy fiber and parallel fiber synapses. This has been explored in the most detail by Sudhof and Linden and their colleagues at

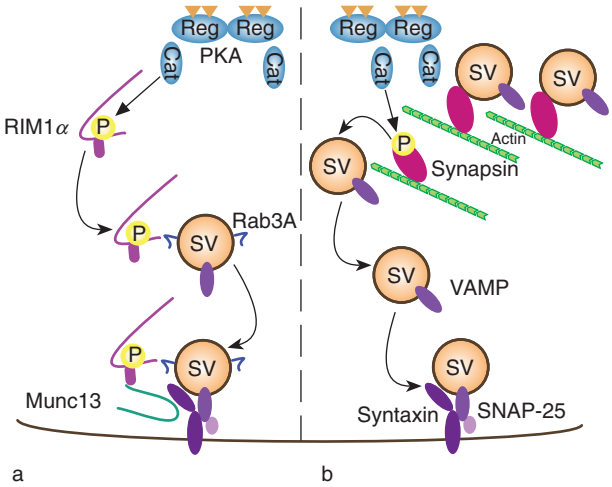

Figure 1 Presynaptic mechanisms of short-term, protein kinase A (PKA)-dependent synaptic enhancement: two alternative mechanisms by which PKA acts to enhance transmitter release. (a) Phosphorylation (P) of RIM1α enhances vesicle priming and increases the probability of vesicle release. PKA, containing two regulatory and two catalytic subunits, is shown above, with cyclic adenosine monophosphate (cAMP) bound to the regulatory subunits. The dissociated catalytic subunit phosphorylates RIM1α at serine 413 near the N-terminus. RIM1α interacts both with Rab3A, which is coupled to synaptic vesicles (SVs), and with munc13. The RIM1α–Munc13–Rab3A complex participates in vesicle priming. Munc13 is shown interacting with syntaxin. (b) Phosphorylation of synapsin (red) recruits vesicles from the reserve pool. Synapsin is shown attached to actin and synaptic vesicles, tethering the vesicles in the reserve pool to the cytoskeleton. Phosphorylation by activated PKA catalytic subunit results in dissociation of synapsin from the vesicle, enabling the vesicle to be mobilized into the readily releasable pool of vesicles. In both (a) and (b), the synaptosome-associated protein (SNAP) receptors (SNAREs; VAMP, syntaxin, and SNAP-25) are shown participating in vesicle priming. (Numerous other presynaptic proteins that do not participate in these specific modulatory mechanisms are omitted from the figure; moreover, proteins participating in one modulatory mechanism are omitted from the other panel. Nevertheless, these omitted proteins play critical roles in vesicle trafficking at these synapses.)

the cerebellar synapses of granule cells to Purkinje neurons (i.e., in parallel fiber LTP). PKA phosphorylates RIM1α at two serine residues, at the 413 and 1548 positions. Mutation of serine 413 to the nonphosphorylatable amino acid alanine blocked parallel fiber LTP; a similar mutation of serine 1548 had no effect. This suggests that phosphorylation of RIM1α at a single residue may mediate the increase in the probability of vesicle release during LTP at both parallel fiber synapses and mossy fiber synapses (Figure 1(a)).

Phosphorylation of Synapsin by PKA Regulates the Readily Releasable Pool of Vesicles

Despite the identification of synapsin a decade or more before other active-zone proteins, it has been difficult to delineate unambiguously its role either in normal exocytosis or in synaptic plasticity. This is, in part, because synapsin does not directly contribute to the exocytosis processes or influence release probability. Synapsin knockout animals do not show striking deficits in synaptic transmitter release. The primary role for synapsin may be the regulation of vesicle availability from the reserve pool of synaptic vesicles at the active zone. Synapsin binds to synaptic vesicles and also to actin filaments, as well as to other cytoskeletal proteins. Through these interactions, synapsin has been proposed to tether the reserve pool of synaptic vesicles to the cytoskeleton. In the absence of synapsins, localization of the reserve pool of synaptic vesicles at the active zone is impaired. Synapsin-dependent regulation of vesicles in the reserve pool is particularly important during high-frequency activity; with long periods of activity, excitatory synapses from synapsin knockout mice display more-severe synaptic depression and impaired recovery from synaptic depression. The interaction of synapsin both with vesicles and with actin depends on phosphorylation, though at different amino acids. Phosphorylation of synapsin at serine 9 either by PKA or by calmodulin-dependent kinase I or IV results in dissociation from synaptic vesicles; phosphorylation at another site by calmodulin-dependent kinase II or MAP kinase triggers dissociation from actin.

Analysis of the role of synapsin is also complicated because there are three synapsin isoforms in mammals. In contrast, *Aplysia* expresses only a single synapsin isoform, and it is in this marine snail that the clearest role for synapsin in learning has been identified. In addition to behavioral sensitization and classical conditioning, the defensive withdrawal of *Aplysia* reflex undergoes habituation. During repeated stimuli, the synaptic connections from the sensory neurons to postsynaptic motor neurons of the reflex circuit become progressively weaker. In a habituated animal, noxious stimuli cause a recovery of the behavioral response called dishabituation. Like sensitization, dishabituation involves the release of serotonin, which acts to enhance transmission at the previously depressed synapses of sensory neurons to motor neurons. John Byrne and colleagues have studied the regulation of synapsin during this serotonin-initiated reversal of synaptic depression. The PKA and calmodulin-dependent kinase I/IV phosphorylation site at the N-terminus of synapsin is highly conserved throughout evolution, suggesting it is central to this synaptic protein's function; the PKA consensus sequence, including serine 9, is also present in *Aplysia*. Exposure to serotonin resulted in the dissociation of synapsin from the main pool of synaptic vesicles at the active zone, as evidenced by immunocytochemistry. The dispersal of synapsin from

the active zone depended on PKA. It is believed that by phosphorylating synapsin and thereby releasing vesicles in the reserve pool that are tethered to the cytoskeleton, PKA increases the availability of vesicles for release in the active zone (Figure 1(b). This increase in vesicle availability triggered by PKA phosphorylation of synapsin may be an important step in the serotonin-induced facilitation of previously depressed synapses of sensory neurons to motor neurons. The full mobilization of vesicles into the readily releasable pool requires docking and priming, which involve other active zone proteins; these steps are likely to be regulated by other protein kinases, including protein kinase C. (Synapsin was recently found to also interact with Rab3A and may participate in docking, but regulation of this interaction by PKA has not been demonstrated.)

The importance of phosphorylation of synapsin at serine 9 was explored using hippocampal pyramidal cells that expressed a mutant protein in which this serine was changed to alanine. Vesicle release was monitored by preloading vesicles with the fluorescent dye FM4-64, applied extracellularly, and then measurement of the disappearance of the fluorescence, which is indicative of the rate of release. Extended periods of release were elicited with prolonged depolarization stimulated by high potassium saline. The decline in fluorescence at presynaptic terminals stimulated by depolarization was dramatically reduced in neurons expressing synapsin with the serine-to-alanine substitution. On the other hand, the mutant synapsin did not affect release triggered by single action potentials. Phosphorylation of serine 9 that was induced by depolarization depended primarily on PKA, as opposed to calmodulin-dependent kinase. This suggests that calcium influx during presynaptic activity activates cAMP-dependent kinase, probably via calmodulin-sensitive AC; phosphorylation by PKA of serine 9 in synapsin mobilizes vesicles from the reserve pool, contributing to the maintenance of transmitter release during high-frequency firing.

Postsynaptic Modulation by cAMP: Regulation of Glutamate Receptors during LTP

In contrast to modulation of vesicle release by PKA, which has been difficult to analyze due to the large numbers of presynaptic proteins that participate in vesicle trafficking, modulation of postsynaptic responses to neurotransmitter have been relatively easy to characterize. At the synapse, postsynaptic receptors are directly modulated by phosphorylation. PKA-dependent modulation of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors occurs during LTP at the CA3-to-CA1 pyramidal cell synapse in hippocampus. High-frequency stimulation of the CA3 axons (e.g., with trains of 100 Hz for $\leq$1 s) results in synaptic potentiation that persists for tens of minutes or longer, depending on the details of the induction protocol. As is the case for most fast excitatory synapses in the mammalian CNS, these synapses utilize glutamate as their neurotransmitter. At these CA3-to-CA1 pyramidal cell synapses, changes in localization of AMPA-type glutamate receptors (GluRs) have been demonstrated to play a central role in the early phase of LTP. AMPA receptors are composed of combinations of GluR subunits. The GluR1 subunit of the AMPA receptor is phosphorylated both at serine 845 by PKA and at serine 831 by calcium–calmodulin-dependent kinase II. Typically, at naive synapses, serine 845 of GluR1 is already phosphorylated. Long-term depression (LTD) is initiated by low-frequency synaptic activation and involves the dephosphorylation of serine 845 by a phosphatase (depending on the protocol, one of several serine/threonine phosphatases may be involved). The loss of the phosphate at serine 845 triggers the endocytic removal of the GluR1-containing receptor from the membrane. Phosphorylation by PKA is therefore required for stabilizing GluR1-containing AMPA receptors at the synapse (Figure 2). Phosphorylation of serine 845 in GluR1 by PKA also increases the open probability of the AMPA receptor, which increases the postsynaptic current elicited by the neurotransmitter glutamate. The dephosphorylation of GluR1 that occurs during LTD has the opposite effect: reducing the current through the AMPA receptor, as well as destabilizing the AMPA receptors. At CA3-to-CA1 synapses that have undergone LTD, high-frequency stimulation initiates potentiation that reverses the effect of LTD (de-depression). In addition to maintaining normal AMPA receptor localization and function, PKA plays a key role in restoring synaptic strength during de-depression. Thus, when GluR1 subunits have been dephosphorylated, reversal of depression is mediated by PKA in the postsynaptic dendritic spines. The activation of PKA during high-frequency presynaptic activity is triggered by calcium influx through *N*-methyl-D-aspartate (NMDA) receptors, probably via activation of calmodulin-sensitive AC (AC1 and AC8) in the dendritic spines.

Localization of PKA via A Kinase Anchoring Proteins Is Necessary for cAMP-Dependent Synaptic Plasticity

For coupling among proteins in signaling cascades, including the cAMP cascade, the co-localization of interacting proteins is critical. Anchoring and scaffolding proteins mediate the assembly of protein

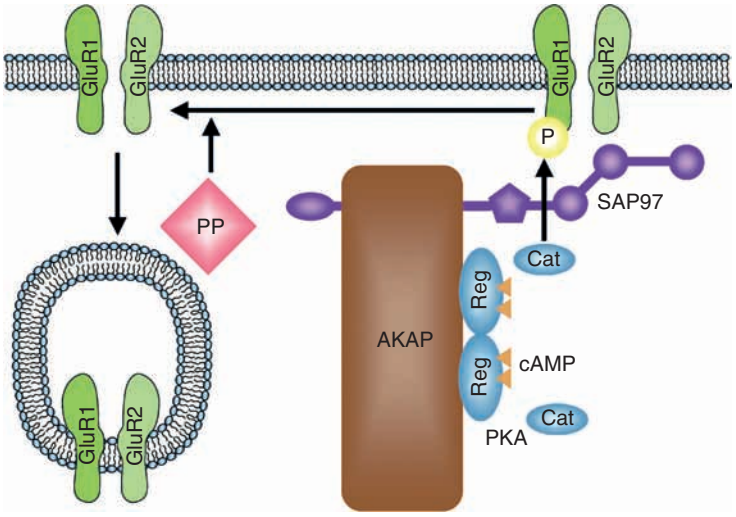

Figure 2 Protein kinase A (PKA) acts postsynaptically to increase α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor response to glutamate. PKA is co-localized with AMPA receptors by binding to A kinase anchoring protein (AKAP) 79/150. AKAP79/150 and AMPA receptors are brought together through their mutual interactions with the scaffolding protein SAP97. Under basal conditions at CA3-to-CA1 pyramidal cell synapses, PKA phosphorylation of glutamate receptor (GluR) 1 at serine 845 (indicated by P) increases the AMPA receptor open probability and stabilizes the receptors in the postsynaptic membrane. During long-term depression (LTD), dephosphorylation of serine 845 by protein phosphatase (PP) decreases the open probability and promotes endocytic removal of the GluR1-containing AMPA receptors from the membrane. When moderate levels of calcium influx through the *N*-methyl-D-aspartate receptor (not shown) initiate LTD, PKA is displaced from the postsynaptic density, shifting the kinase–phosphatase balance in favor of dephosphorylation of serine 845. The protein phosphatase responsible for regulating serine 845 is not specified, though calcineurin is activated by calcium influx during initiation of LTD. When higher-frequency presynaptic stimulation initiates potentiation of synapses that have undergone LTD, de-depression involves phosphorylation of serine 845 to increase AMPA receptor current and to promote AMPA receptor stabilization. Cat, catalytic subunit of PKA; Reg, regulatory unit of PKA.

complexes, including the enzymes that produce intracellular messengers, the channels that mediate calcium elevation (e.g., NMDA receptors), protein kinases, phosphatases, and the target proteins that are regulated by phosphorylation. This localization of signaling proteins within microdomains achieves both specificity and efficiency of the modulatory cascades. Because the co-localization of signaling molecules within a microdomain varies among cell types, co-localization also contributes to the cell specificity of the effects of each signaling cascade. Localization of PKA is achieved through A kinase anchoring proteins (AKAPs), which have been extensively characterized by John Scott. AKAPs anchor the regulatory subunits of PKA, localizing the protein kinase via the targeting domains of each particular AKAP. One such AKAP, AKAP79/150, targets both PKA and the phosphatase calcineurin to the postsynaptic density by interacting with two scaffolding proteins, PSD-95 and SAP97. SAP97 binds to AMPA receptors, whereas PSD-95 binds to NMDA receptors. The RII subunit of PKA binds to AKAP79/150, positioning the kinase proximal to both NMDA and AMPA GluRs. AKAP localization of PKA is required for the normal PKA phosphorylation of GluR1 at serine 845 and for the stabilization of GluR1-containing AMPA receptors in the postsynaptic membrane (Figure 2). During LTD and de-depression, the initiating signal for the plasticity is calcium influx through NMDA receptors. Therefore the AKAP complex localizes PKA close to both the source of the initiating signal, the NMDA receptor, and the target protein, the AMPA receptor. There must be similar co-localization of calmodulin-sensitive AC with the channels that mediate calcium influx, the NMDA receptors, to enable PKA activation during de-depression. With LTD, there is loss of GluR1-containing AMPA receptors from the surface of the postsynaptic membrane, as described above (Figure 2). Activation of calcium influx through NMDA receptors by low levels of activity, which induce LTD, displaces the PKA that contains RIIβ from the postsynaptic density. The importance of co-localization of PKA can be tested by means of the Ht31 peptide, which corresponds to the PKA-binding domain of an AKAP; this peptide can be used experimentally to disrupt the PKA–AKAP association. When the Ht31 peptide was injected postsynaptically, both GluR1 and GluR2 were lost from the membrane, presumably through endocytosis, as during LTD. Thus, it seems that part of the mechanism of LTD is

to displace PKA from the postsynaptic density, thereby eliminating the PKA-dependent stabilization of AMPA receptors.

AKAP-dependent localization of PKA is also critical for presynaptic mechanisms of synaptic plasticity. Serotonin-dependent initiation of long-term synaptic facilitation of the connection of sensory neurons to motor neurons in *Aplysia* requires the RII subunit of PKA. Injection of the Ht31 peptide into the presynaptic *Aplysia* sensory neuron to disrupt PKA–AKAP coupling completely blocked long-term synaptic facilitation tested at 24 h after exposure to serotonin.

A clear link between AKAP-mediated localization of PKA and learning has been demonstrated in amygdala-dependent fear conditioning by LeDoux and colleagues. PKA in the lateral amygdala is required for both context-dependent and cue-dependent fear conditioning, when learning is evaluated at 24 h after training. Injection of a membrane-permeant form of the Ht31 peptide into the amygdala disrupts formation of stable memory for fear conditioning evaluated at the 24 h test point. Thus AKAP-mediated localization of PKA is essential for cAMP-dependent presynaptic and postsynaptic mechanisms of plasticity, and also for formation of long-term memory.

Role of cAMP and PKA in Initiating Long-Term Synaptic Plasticity

Synaptic plasticity that is initiated by transient increases in an intracellular messenger typically lasts only minutes or tens of minutes, ending when the intracellular signal disappears. In the case of cAMP, when AC is no longer activated, cAMP levels drop as a result of hydrolysis by phosphodiesterase; protein phosphatases then rapidly reverse the phosphorylation of proteins that had been targets of PKA. The initiation of more-persistent alteration of synaptic strength requires gene transcription and protein synthesis, much as is true for long-term learning (see the section titled 'cAMP-dependent signaling plays a central role in long-term learning'). Thus, long-term synaptic plasticity involves mechanisms distinct from the mechanisms that mediate short-term synaptic plasticity, which typically depends on covalent modification of existing proteins. One of the early steps in the initiation of stable long-term synaptic changes and long-term learning is the transient expression of immediate early genes. These immediate early genes code for transcription factors and other proteins important for the transition from short-term to long-term plasticity. These proteins then initiate a later wave of transcription and protein synthesis which contributes more directly to stable synaptic changes.

cAMP-Dependent Phosphorylation Plays an Important Role in Initiating Stable Transcription-Dependent Long-Term Synaptic Plasticity in Both *Aplysia* and Mammals

In *Aplysia*, repeated brief exposures to serotonin result in long-term facilitation of the synaptic connections of sensory neurons to motor neurons, lasting 24 h or more. Initiation of long-term facilitation at these synapses depends on cAMP and PKA. In hippocampus, LTP at CA3-to-CA1 pyramidal cell synapses lasts either tens of minutes or hours depending on the pattern of high-frequency stimulation. Stable late-phase LTP that lasts several hours is initiated when the presynaptic Schaeffer collateral axons from the CA3 pyramidal cells are stimulated with repeated trains of high-frequency stimuli, whereas single trains elicit only short-lasting LTP. Initiation of late-phase LTP involves cAMP, similar to long-term facilitation in *Aplysia*. Storm and colleagues examined mice in which both of the calmodulin-sensitive ACs, AC1 and AC8, had been knocked out; the CA3-to-CA1 pyramidal cell synapses in these animals displayed no late-phase LTP, measured 80 min after the end of high-frequency presynaptic stimulation (Figure 3). Injection of an inhibitor of PKA into the postsynaptic CA1 pyramidal neurons similarly

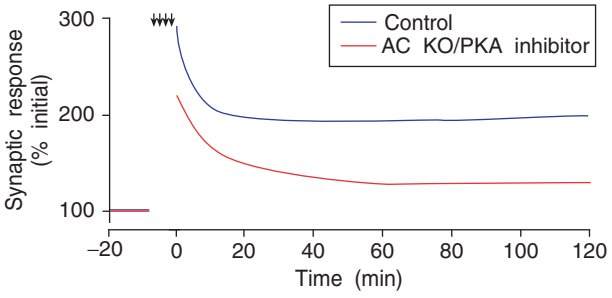

Figure 3 Contribution of adenylyl cyclase (AC) and protein kinase A (PKA) to late-phase long-term potentiation (LTP) at CA3-to-CA1 pyramidal cell synapses. Blue curve illustrates typical stable, late-phase LTP. Late-phase LTP was initiated by a series of four high-frequency 100 Hz stimulus trains (arrows), which activated the presynaptic CA3 pyramidal cell axons. Depending on the study, the train duration was 200 ms or 1 s and the interval between trains was 6 s or 5 min. The late phase of LTP was almost completely eliminated by postsynaptic injection of a PKA inhibitor or by knockout (KO) of the calmodulin-dependent ACs (AC1 and AC8; red curve). Graph is illustrative of the effects of blockade of the cyclic adenosine monophosphate pathway observed by Wong et al. and Duffy and Nguyen. Note the partial decrease in LTP with blockade of the PKA pathway even at early time points, which is before the transcription-dependent phase of LTP first appears. This suggests that PKA may also contribute to early-phase LTP, at least with some protocols; a PKA-dependent contribution to early-phase LTP could be mediated by modulation of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor open probability or trafficking (see Figure 2).

blocked the late phase of LTP. Furthermore, Kandel and colleagues found that a dominant negative form of the regulatory subunit, which reduced PKA activity by as much as 50%, completely eliminated the late phase of LTP. Thus, even in regions of the mammalian CNS where short-term synaptic plasticity is independent of cAMP, long-term synaptic potentiation requires phosphorylation by PKA.

Presynaptic Contributions to Late-Phase LTP at CA3-to-CA1 Pyramidal Cell Synapses

Induction of late-phase LTP at CA3-to-CA1 pyramidal cell synapses is believed to be initiated by postsynaptic calcium influx through NMDA-type GluRs and calmodulin-dependent activation of AC. Siegelbaum and Kandel and colleagues examined the effects of treatment with cell-permeant cAMP analogs on presynaptic release sites at these same synapses. To monitor exocytosis at active release sites, they applied the lipophilic dye FM1-43, which labels vesicles during fusion with the plasma membrane. They found, 2 h after exposure to the cAMP analog, a sixfold increase in the number of active release sites without any change in total number of synaptic sites. This dramatic increase in functional release sites reflects conversion of preexisting silent release sites to active release sites (or conversion of release sites of very low probability to release sites of high probability). Because this increase in functional release sites depended on protein synthesis and because this change developed gradually, appearing at 2 h but not at 30 min, this enhancement of release must correspond to a component of long-term synaptic potentiation. The induction of the increase in active release sites initiated by the cell-permeant cAMP analog required functional postsynaptic GluRs that responded to the neurotransmitter released from the CA3 pyramidal cell terminals. This suggests that the mechanism for initiating the change in release involves the postsynaptic CA1 neuron and must reflect transsynaptic signaling. As discussed above, late-phase LTP at these same synapses involves postsynaptic activation of calcium–calmodulin-dependent AC and postsynaptic protein synthesis. Although these experiments in which increases in active release sites were imaged involved pharmacologically induced plasticity initiated by a cAMP analog, it likely that this coordinate interaction of pre- and postsynaptic cells in producing stable synaptic changes is a general phenomenon.

Persistently Activated Protein Kinase Contributes to Initiation of Long-Term Synaptic Facilitation

In order to understand how transient changes in synaptic connections initiated by intracellular messengers might be converted to more stable changes, James Schwartz and his colleagues explored the possibility that a specific protein kinase might be modified so as to upregulate phosphorylation. During the mid-1980s, they obtained evidence for persistent activation of PKA during induction of long-term synaptic facilitation in *Aplysia*; this persistent activation of PKA results from proteolysis of the regulatory subunit, removing the inhibitory regulation of the catalytic (C) subunit. Schwartz and his colleagues also observed that calmodulin-dependent kinase II can be persistently activated; since then, the persistent activation of calmodulin-dependent kinase II in mammalian hippocampus has been analyzed in detail. More recently, during LTP in hippocampus, protein kinase C was found to be expressed in a truncated form that lacks its inhibitory domain and is therefore autoactive. Thus, the persistent activation of serine/threonine protein kinases appears to be a general mechanism contributing to persistent synaptic plasticity.

How do serotonin and cAMP initiate degradation of the PKA regulatory subunits? Analogous to the requirements for induction of the late phase of LTP, initiation of stable long-term facilitation of *Aplysia* synapses of sensory neurons to motor neurons requires repeated exposures to serotonin spanning tens of minutes. PKA then triggers long-term facilitation of synapses of sensory neurons to motor neurons, a process that depends on both transcription and protein synthesis. One of the early steps in initiation of long-term facilitation is the proteolysis of the RI isoform of the regulatory subunit of PKA. This degradation of RI is mediated by the ubiquitin–proteasome pathway. Ubiquitin is a small, 76-amino-acid protein; proteins that are targeted for degradation are covalently labeled via the attachment of a chain of individual ubiquitin monomers. The proteasome is a large multiprotein proteolytic complex that cleaves ubiquitinated proteins into polypeptides. The ubiquitin-mediated degradation of RI subunits of PKA requires two steps. First, cAMP binding to both sites on RI initiates a conformational change that makes the subunit vulnerable to ubiquitination. Second, one of the immediate early genes that is induced in the presynaptic sensory neuron by repeated exposures to serotonin is a ubiquitin hydrolase; this hydrolase disassembles ubiquitin chains, accelerating the processing of proteins through the ubiquitin–proteasome pathway and the recycling of ubiquitin. Normally, RI subunits persist for 2 weeks, but with serotonin exposure, a percentage of RI subunits is degraded in less than 2 h. This degradation of the R subunits results in a population of PKA C subunits that are not bound to R subunits and therefore are tonically active. In a series of elegant experiments, Schwartz and Kandel and their colleagues found that blocking

the proteasome pathway in the presynaptic sensory neurons prevents induction of long-term facilitation. Injecting an excess of C subunits overcomes this effect, initiating substantial long-term facilitation. Thus, the ubiquitin-mediated degradation of the R subunits and the resulting reduction in the inhibitory regulation of C subunits are critical for the induction of stable long-term synaptic facilitation by serotonin. The RI subunits are primarily localized to the cell body, which suggests that the primary role of the autoactive PKA catalytic units is to initiate the transcription cascade that results in long-term facilitation. This process represents a positive feedback loop in which serotonin and cAMP initiate transcription of immediate early genes; the transcription of ubiquitin hydrolase then results in degradation of RI, which enhances PKA activity, potentiating further activation of the transcription cascade. The necessity for a subsequent increase in cAMP, which binds to RI and enables its degradation after ubiquitin hydrolase has been expressed, may be one reason that long-term learning requires repeated training and that long-term synaptic plasticity requires repeated exposure to the initiating stimulus, serotonin.

Initiation of Long-Term Plasticity by cAMP Involves Phosphorylation of Transcription Factors, Including CREB: Interactions between PKA and MAP Kinase

Many of the genes induced by cAMP have an upstream control region that contains a specific DNA consensus sequence, known as the cAMP responsive element (CRE). Activation of transcription of CRE-regulated genes occurs via phosphorylation of transcription factors, including the CRE binding protein (CREB), which then bind to the CRE. The role of the CRE in initiating long-term plasticity in *Aplysia* was tested by injecting into the nucleus of the presynaptic sensory neuron short DNA dimers that include the CRE sequence. These oligonucleotides should bind any activated transcription factors that interact with the CRE, including CREB, and should therefore block activation of the native CRE. Indeed, this 'transcription factor decoy' completely blocked long-term facilitation, measured 24 h after treatment with serotonin. In *Aplysia* and *Drosophila*, CREB has been demonstrated to play a central role in long-term synaptic facilitation and long-term learning, respectively. For example, in *Drosophila*, long-term memory was blocked by overexpression of an endogenous inhibitory isoform of CREB, which inhibits CREB activation of CRE-mediated transcription. In mammals, the situation is more complex because there are two other CREB-related genes, CREM and ATF1, which code for transcription factors that also bind to the CRE. Manipulations of the CREB gene in mammals have produced limited effects on long-term plasticity, perhaps because of compensatory increases in the expression of these other CREB family members.

The analysis of the role of CREB in long-term plasticity is also complicated because CREB can actually be phosphorylated by more than half a dozen protein kinases. In *Aplysia* CNS and in mammalian brain, it has been found that MAP kinase is required for CREB activation by cAMP. The interaction between the two signaling cascades has been most thoroughly analyzed by Storm and colleagues using CRE-regulated transcription of a reporter gene. Although PKA does not directly phosphorylate CREB in hippocampal neurons, PKA is required for the activation of CRE-mediated transcription triggered by calcium influx and cAMP synthesis. MAP kinase was also required for the initiation of CRE-mediated transcription. During exposure to an activator of AC, paired with calcium influx produced by membrane depolarization, MAP kinase is translocated to the nucleus. Both the activation and the translocation of MAP kinase depend on PKA (Figure 4). A second kinase, RSK2, is translocated to the nucleus along with MAP kinase. RSK2 is activated by MAP kinase, as is MSK1. Both RSK2 and MSK1 are CREB kinases that directly phosphorylate CREB. Thus, PKA participates in translocating and activating other kinases, which in turn are responsible for activation of CRE-mediated transcription.

Summary

cAMP and PKA play central roles in synaptic plasticity and learning. Because of interactions among signaling cascades, it is often difficult to analyze the contributions of one pathway in isolation. At some specific synapses, cAMP-dependent protein phosphorylation contributes to short-term synaptic plasticity and short-term learning. Postsynaptically, short-term modulatory mechanisms involve the regulation of neurotransmitter receptors. Receptor open probability or conductance may be modulated, as well as the stability of receptors in the plasma membrane. Presynaptically, because the vesicle trafficking apparatus involves more than 30 participating proteins, there are many possible targets for modulation by PKA; it is likely that these mechanisms have not yet been fully characterized. PKA-dependent phosphorylation of two presynaptic proteins, synapsin and RIM1α, is known to contribute to synaptic plasticity. Synapsin primarily regulates trafficking between the reserve pool and the readily releasable pool of vesicles, whereas RIM1α influences vesicle priming and the probability of release. One important interaction between cascades is the activation of calmodulin-sensitive AC by calcium influx. The AC–cAMP–PKA pathway depends on co-localization

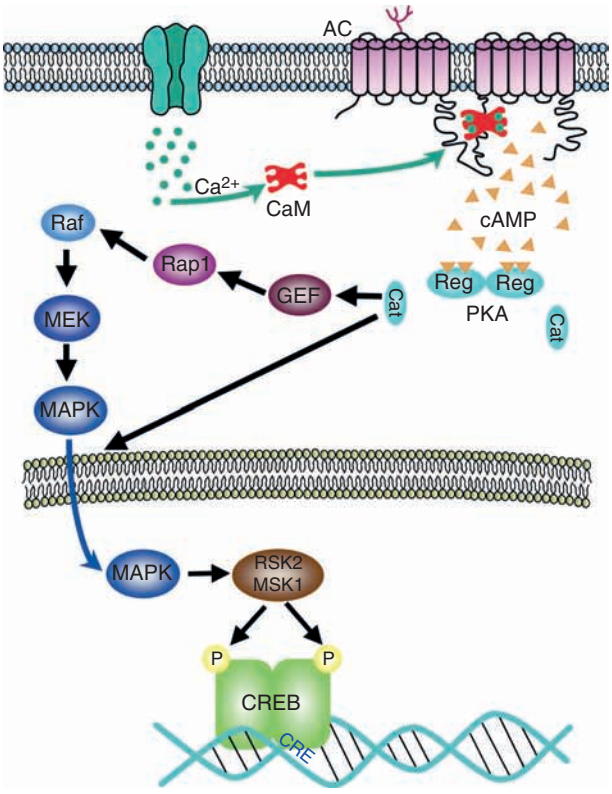

Figure 4 Interactions between protein kinase A (PKA) and mitogen-activated protein (MAP) kinase in activation of cyclic adenosine monophosphate (cAMP) responsive element (CRE)-mediated transcription. Calcium influx, via either voltage-sensitive calcium channels (shown here) or *N*-methyl-D-aspartate-type glutamate receptors, acts via calmodulin (CaM) to stimulate adenylyl cyclase (AC). cAMP (yellow triangles) binds to the regulatory subunit of PKA, releasing the catalytic subunits. PKA activates the mitogen-activated protein kinase (MAPK) cascade via the small G-protein Rap1. Small G-proteins are regulated by guanosine triphosphate (GTP) exchange factors (GEFs), which increase the rate of exchange of GTP for guanosine diphosphate (GDP). Phosphorylation of a Rap1-specific GEF activates Rap1, which in turn stimulates the protein kinase Raf. Raf phosphorylates the kinase MEK, which phosphorylates MAPK. The translocation of MAPK from the cytoplasm to the nucleus depends on activity of PKA via an unknown mechanism. In the nucleus, MAPK activates one of two CRE binding protein (CREB) kinases, RSK2 or MSK1, which phosphorylates CREB. Phosphorylated CREB binds to CRE, activating transcription of CRE-regulated genes. A specific activation pathway based on results from several studies is illustrated here, but depending on the specific cell type, the co-localization of these signaling proteins varies. Thus, there can be additional pathways for the activation of MAPK, including Ras. cAMP also binds to a subclass of GEFs called exchange proteins directly activated by cAMP (EPACs), not shown; however, because of a lack of appropriate co-localization, EPACs do not appear to mediate the activation of the MAPK cascade.

of these signaling proteins with upstream regulators, such as calcium channels, and with downstream target proteins. Co-localization of PKA with target proteins is mediated by AKAPs. The binding of the regulatory subunit of PKA to specific AKAPs and the resulting co-localization of PKA with target proteins is necessary for induction of PKA-dependent plasticity. Initiation of long-term synaptic plasticity and long-term learning involve activation of transcription cascades. cAMP and PKA play a central role in this process by activating transcription of genes regulated by the CRE, a DNA regulatory element. Activation of CRE-mediated transcription by PKA is indirect. The signaling between PKA and the CRE involves a cascade of interacting proteins, including MAP kinase and downstream CREB kinases that directly activate the transcription factor CREB in the nucleus.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Glutamate Receptor Organization: Ultrastructural Insights; Long-Term Depression: Cerebellum; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Mechanisms of Learning.

Further Reading

Abel T, Nguyen PV, Barad M, et al. (1997) Genetic demonstration of a role for PKA in the late phase of LTP and in hippocampus-based long-term memory. *Cell* 88: 615–626.

Angers A, Fioravante D, Chin J, et al. (2002) Serotonin stimulates phosphorylation of *Aplysia* synapsin and alters its subcellular distribution in sensory neurons. *Journal of Neuroscience* 22: 5412–5422.

Banke TG, Bowie D, Lee H, et al. (2000) Control of GluR1 AMPA receptor function by cAMP-dependent protein kinase. *Journal of Neuroscience* 20: 89–102.

Chain DG, Casadio A, Schacher S, et al. (1999) Mechanisms for generating the autonomous cAMP-dependent protein kinase required for long-term facilitation in *Aplysia*. *Neuron* 22: 147–156.

Dash PK, Hochner B, and Kandel ER (1990) Injection of the cAMP-responsive element into the nucleus of *Aplysia* sensory neurons blocks long-term facilitation. *Nature* 345: 718–721.

Dell'Acqua ML, Smith KE, Gorski JA, et al. (2006) Regulation of neuronal PKA signaling through AKAP targeting dynamics. *European Journal of Cell Biology* 85: 627–633.

Greenberg SM, Castellucci VF, Bayley H, and Schwartz JH (1987) A molecular mechanism for long-term sensitization in *Aplysia*. *Nature* 329: 62–65.

Impey S, Obrietan K, Wong ST, et al. (1998) Cross talk between ERK and PKA is required for Ca2+ stimulation of CREB-dependent transcription and ERK nuclear translocation. *Neuron* 21: 869–883.

Liu J, Hu JY, Schacher S, and Schwartz JH (2004) The two regulatory subunits of *Aplysia* cAMP-dependent protein kinase mediate distinct functions in producing synaptic plasticity. *Journal of Neuroscience* 24: 2465–2474.

Lonart G, Schoch S, Kaeser PS, et al. (2003) Phosphorylation of RIM1alpha by PKA triggers presynaptic long-term potentiation at cerebellar parallel fiber synapses. *Cell* 115: 49–60.

Ma L, Zablow L, Kandel ER, and Siegelbaum SA (1999) Cyclic AMP induces functional presynaptic boutons in hippocampal CA3-CA1 neuronal cultures. *Nature Neuroscience* 2: 24–30.

Menegon A, Bonanomi D, Albertinazzi C, et al. (2006) Protein kinase A-mediated synapsin I phosphorylation is a central modulator of Ca^{2+}-dependent synaptic activity. *Journal of Neuroscience* 26: 11670–11681.

Moita MA, Lamprecht R, Nader K, and LeDoux JE (2002) A-kinase anchoring proteins in amygdala are involved in auditory fear memory. *Nature Neuroscience* 5: 837–838.

Snyder EM, Colledge M, Crozier RA, et al. (2005) Role for A kinase-anchoring proteins (AKAPS) in glutamate receptor trafficking and long term synaptic depression. *Journal of Biological Chemistry* 280: 16962–16968.

Storm DR, Hansel C, Hacker B, et al. (1998) Impaired cerebellar long-term potentiation in type I adenylyl cyclase mutant mice. *Neuron* 20: 1199–1210.

Wang H and Storm DR (2003) Calmodulin-regulated adenylyl cyclases: Cross-talk and plasticity in the central nervous system. *Molecular Pharmacology* 63: 463–468.

Yovell Y and Abrams TW (1992) Temporal asymmetry in activation of *Aplysia* adenylyl cyclase by calcium and transmitter may explain temporal requirements of conditioning. *Proceedings of the National Academy of Sciences of the United States of America* 89: 6526–6530.

Cyclic Nucleotide-Gated and Hyperpolarization-Activated Channels

M Biel and S Michalakis, Ludwig-Maximilians Universität München, Munich, Germany
A Ludwig, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
F Hofmann, Technische Universität München, Munich, Germany

© 2009 Elsevier Ltd. All rights reserved.

Structural Features of Cyclic Nucleotide-Gated and Hyperpolarization-Activated Cyclic Nucleotide-Gated Channels

Cyclic nucleotide-gated (CNG) and hyperpolarization-activated cyclic nucleotide-gated (HCN) channels are members of an extended superfamily of cation channels that is characterized by a principal building unit containing six transmembrane helices (S1–S6) and an reentrant pore-loop (ion selectivity filter) between S5 and S6 (Figure 1). Four channel subunits are assembled around the central ion conducting pore to form a tetrameric complex. CNG channels pass monovalent cations, such as Na^+ and K^+, but do not discriminate between them. Ca^{2+} is also permeable but at the same time acts as a voltage-dependent blocker of monovalent cation permeability. By providing an entry pathway for Ca^{2+}, CNG channels control a variety of cellular processes that are triggered by this cation. HCN channels conduct Na^+ and K^+ with permeability ratios of approximately 1:4 and are blocked by millimolar concentrations of Cs^+. Despite this preference for K^+ conductance, HCN channels carry an inward Na^+ current under physiological conditions. HCN channels can also conduct Ca^{2+}, but not CNG channels. At 2.5 mM external Ca^{2+}, the fractional Ca^{2+} current of HCN2 and HCN4 is approximately 0.5%, whereas for native CNG channels it is 10–80%.

Native CNG channels are heterotetramers composed of homologous A subunits (CNGA1–CNGA4) and B subunits (CNGB1 and CNGB3). The subunit stoichiometries have been determined for the channels expressed in rod photoreceptors (3 CNGA1:1 CNGB1a), cone photoreceptors (2 CNGA3:2 CNGB3), and olfactory neurons (2 CNGA2:1 CNGA4:1 CNGB1b). The HCN channel family comprises four members (HCN1–4). In contrast to CNG channels, the exact subunit composition and stoichiometry of native HCN channels has not been determined experimentally. However, there is evidence for the presence of both homo- and heterotetrameric channels *in vivo*.

CNG and HCN channels differ from each other with regard to their mode of activation. CNG channels are typical ligand-gated channels. Their activation obligatorily requires binding of cAMP or cGMP to a cyclic nucleotide-binding domain (CNBD) in the C-terminus. This domain is homologous to CNBDs of other cyclic nucleotide-binding proteins, such as cAMP- and cGMP-dependent protein kinases or cAMP-regulated guanine nucleotide exchange factors. In contrast to CNG channels, HCN channels are principally operated by membrane hyperpolarization. Sensitivity to voltage is conferred by a positively charged voltage sensor motif (S4) in the channel core. HCN channels also contain a CNBD in the C-terminus that shows strong homology to the CNBD of CNG channels. Unlike in CNG channels, binding of cAMP to the CNBD is not sufficient for HCN channel activation. However, it affects the voltage dependence of channel activation and, hence, modulates the voltage range where HCN channels can operate. The extent of regulation by cAMP varies between different HCN channels. It is profound in HCN2 and HCN4 (shift of $V_{0.5}$ by +10 to +20 mV) but weak in HCN1 and HCN3 (shift of $V_{0.5}$ by −5 to +7 mV). The crystal structure of the proximal C-terminus of HCN2 containing the CNBD and the short linker sequence that connects the CNBD with the S6 helix (the C-linker) has been determined. Together with results from functional studies, the structural data indicate that in the absence of cAMP the proximal C-terminus assembles into a twofold symmetry (dimer of dimer) that inhibits HCN channel gating, probably via an interaction between the C-linker and the channel core. Binding of cAMP to the CNBD promotes a conformational change to a fourfold symmetry and relieves this inhibition. Differences in the magnitude of the response to cAMP among the four HCN channel isoforms are largely due to differences in the extent to which the CNBD inhibits basal gating.

Functional Roles of CNG Channels in Visual and Olfactory Transduction

CNG channels are molecular switches converting alterations of cellular cAMP and cGMP levels into changes in the membrane potential. In outer segments of rod and cone photoreceptors, the CNG channel is opened in the dark by high cGMP concentrations. The resulting inward current depolarizes the cell and triggers synaptic glutamate release. Light-induced hydrolysis of cGMP leads to the closure of the CNG channel and, hence, to cessation of synaptic transmission. Olfactory transduction takes place in the cilia of olfactory receptor neurons. It relies on odorant-induced synthesis of cAMP and subsequent activation

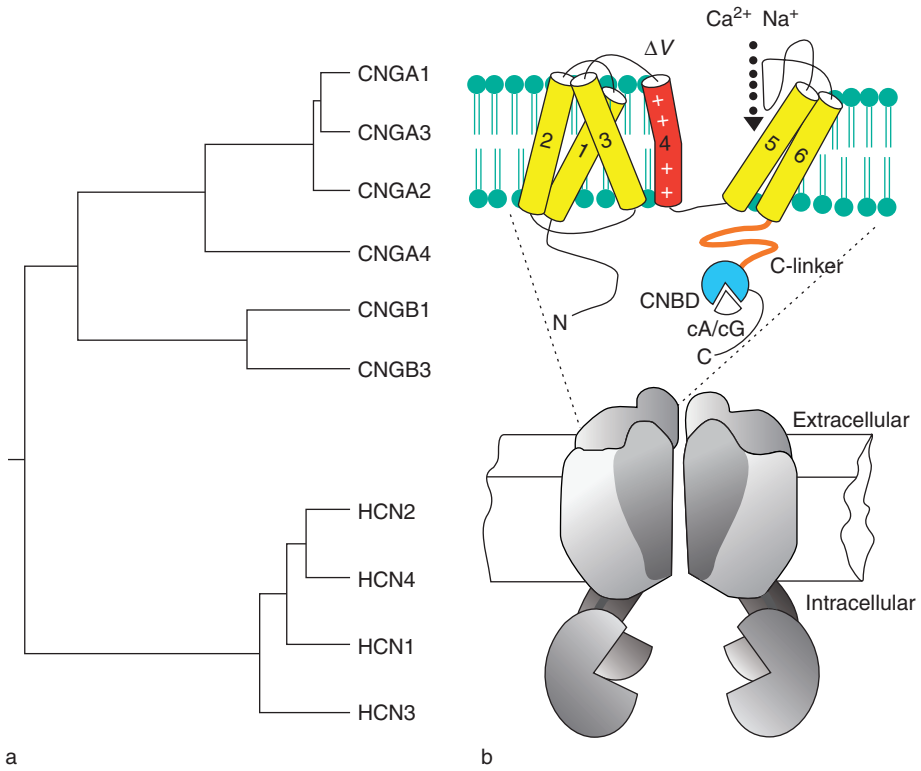

Figure 1 (a) Phylogenetic tree of human CNG and HCN channel subunits. (b) The membrane topology of CNG and HCN channel subunits (top). The native channels are homo- or heterotetramers (bottom). 1–6, transmembrane segments 1–6; C, C-terminus; cA, cyclic AMP; cG, cyclic GMP; CNBD, cyclic nucleotide binding domain; N, N-terminus.

of a CNG channel. Influx of Na^+ and Ca^{2+} depolarizes the membrane and triggers action potentials. Depolarization is amplified by an outward Cl^- current that is activated by Ca^{2+} entering the cell through the CNG channels.

Channelopathies and Knockout Models of the Rod Photoreceptor CNG Channel

CNGA1 Subunit

Mutations in the CNGA1 gene (Table 1) have been found in patients suffering from retinitis pigmentosa (RP). RP comprises a genetically diverse group of progressive degenerative diseases affecting the photoreceptors of the retina. The most common symptoms of RP include night blindness, progressive concentric reduction of the visual field, an abnormal accumulation of pigmentation in the retina, and blindness. At the cellular level, the disease is characterized by a primary impairment or total loss of rod function and structure followed by a secondary degeneration of the cones. RP has been mapped to more than 30 genes encoding proteins involved in the visual transduction pathway (e.g., rhodopsin) or required for the maintenance of rod architecture (e.g., peripherin). Mutations in CNGA1 lead to an autosomal recessive form of RP that accounts for approximately 1% of cases of RP. Three of the five mutant alleles identified are null mutants (E76Stop, K139Stop, and deletion of most or all of the CNGA1 transcription unit) because they would encode proteins that are lacking essential channel domains. The remaining two alleles (S316F and R654DStop) encode channels that, in heterologous expression systems, fail to reach the plasma membrane. To date, mice lacking the CNGA1 subunit are not available. However, transgenic mice revealing an approximately 50% reduction of CNGA1 transcript levels due to the overexpression of a CNGA1 antisense mRNA have been produced. These mice show some histological features reminiscent of RP (e.g., reduced number of photoreceptors and apoptotic death of retinal cells).

CNGB1a Subunit

The CNGB1 gene locus consists of at least 33 exons encoding several splice variants. Rods express a 240-kDa isoform (CNGB1a) containing a long cytosolic N-terminus that is also translated as a separate cytosolic protein (glutamic acid rich protein (GARP)). Shorter variants of the subunit lacking the GARP part are present in the native olfactory CNG channel

Table 1 Overview of diseases associated with CNG channels

Type	*Subunit*	*Animal model*	*Human disease*
Rod channel	CNGA1	CNGA1 antisense expressing mice: retinal degeneration	Retinitis pigmentosa
	CNGB1a	CNGB1-deficient mice: impaired rod function and retinal degeneration	Retinitis pigmentosa
Cone channel	CNGA3	CNGA3-deficient mice: loss of cone function, cone degeneration, enhanced hippocampal LTP, impaired cued fear conditioning	Achromatopsia, progressive cone dystrophy
	CNGB3	Canine models (null deletion or missense mutation): cone degeneration	Achromatopsia, progressive cone dystrophy
Olfactory channel	CNGA2	CNGA2-deficient mice: impaired olfaction, impaired hippocampal LTP	
	CNGA4	CNGA4-deficient mice: impaired odor adaptation, impaired olfaction	
	CNGB1b	CNGB1-deficient mice: impaired odor adaptation and impaired olfaction	

LTP, long-term potentiation.

(CNGB1b) and have been identified in sperm cells and other tissues. A CNGB1-deficient mouse model covering all potential CNGB1 splice variants was generated by deleting exon 26, which encodes the pore-forming region and the S6 segment. These mice show an impairment of both vision and olfaction. In this section, we describe the retinal phenotype; the olfactory phenotype is described later.

CNGB1-deficient mice (Table 1) show a slow-progressing degeneration of the retina caused by apoptotic death and concurred by retinal gliosis. The degeneration primarily affects rod photoreceptors and is detectable in these cells as early as postnatal day 15. Cones are quite unaffected up to 6 months of age but start to degenerate subsequently. At the age of approximately 1 year, the retina of CNGB1-deficient animals is essentially devoid of both rods and cones. The principal structure of the inner retina is unaffected by the deletion of CNGB1. However, rod bipolar cells and horizontal cells reveal some morphological alterations (e.g., sprouting extensions, retraction of processes, and misplacement of cell bodies). CNGB1-deficient mice show no detectable rod-mediated responses in electroretinograms (ERGs), whereas at up to 6 months of age, cone-mediated responses are normal. Later, the cone responses also decrease, and at the age of 12 months knockout mice are legally blind. In agreement with the ERG measurements, the vast majority of isolated rod photoreceptors of young mice (postnatal day 17) fail to respond to light. Only in a few cells can a tiny dark current be detected. The severe retinal phenotype of CNGB1-deficient mice was unexpected given that CNGB1 was considered to be a modulatory subunit not needed for principal channel formation. This conundrum could be solved by Western blots and immunohistochemistry showing that in CNGB1-deficient mice, only trace amounts of the CNGA1 subunit are present on the rod outer segment. Thus, in contrast to cell lines used for heterologous expression (e.g., *Xenopus* oocytes or HEK293 cells), rod photoreceptors are not capable to produce substantial amounts of the homomeric CNGA1 channel and/or cannot target homomeric channels to the rod outer segment. It is not known why photoreceptors, unlike *Xenopus* oocytes or HEK293 cells, cannot efficiently target homomeric CNGA1 channels to the plasma membrane. It is tempting to speculate that the requirement of CNGB1 for targeting CNGA1 to the outer segment may have evolved to ensure that only heteromeric channels reach the outer segment membrane.

The total loss of CNG channel expression explains why the retinal phenotype of CNGB1-deficent mice is very similar to that found in RP patients with null mutations in the CNGA1 gene. Interestingly, RP patients have been identified who carry a point mutation in the CNBD of CNGB1. The functional consequences of this mutation, G993V, which involves a highly conserved residue in the CNG channel family, have not been determined in expression systems. Importantly, it is not known whether the mutant protein is targeted to the plasma membrane of rod outer segments. If the mutant protein is not targeted, the disease mechanism in CNGB1-deficient mice would be equivalent to that leading to human CNGB1-related RP. Alternatively, if the mutant CNGB1 is produced and assembled with CNGA1, it might impair the cGMP-dependent activation of the channel.

Taken together, the genetic data imply that photoreceptors degenerate in response to the absence (or downregulation) of functional CNG channels in outer

segments. The molecular pathways leading to the initiation of cell death in CNGA1/CNGB1-deficient rods are unknown. One possible mechanism refers to the 'equivalent light' hypothesis. According to this hypothesis, the absence of functional CNG channels is equivalent to the permanent closure of channels occurring under continuous bright light conditions. Continuous exposure of experimental animals to light has been shown to result in photoreceptor degeneration. In support of the equivalent light hypothesis, mutations of several other genes leading to constitutive activation of the visual cascade (e.g., null mutants of the guanylyl cyclase and activating mutants of rhodopsin) have been identified in RP patients. It is unclear, however, how constitutive activation of phototransduction leads to retinal degeneration. Low Ca^{2+} levels resulting from the loss of Ca^{2+} influx through CNG channels as well as toxicity induced by reactive oxygen species may play roles in this process.

Channelopathies and Knockout Models of the Cone Photoreceptor CNG Channel

CNGA3 Subunit

Mutations in the CNGA3 subunit cause achromatopsia (total colorblindness or rod monochromacy) (Table 1). CNGB3 and GNAT2 (which encodes cone-specific transducin) have been identified as additional achromatopsia genes. Achromatopsia is a rare autosomal recessive disorder (incidence of approximately 1 in 30 000), in which functional cones are absent from the retina. Affected individuals show a total loss of color discrimination, photophobia, nystagmus, and poor visual acuity. Electroretinography reveals absent photopic (cone) responses and normal scotopic (rod) responses. Individuals with incomplete achromatopsia retain some color vision. Approximately 50 mutations have been identified in the CNGA3 gene of patients suffering from achromatopsia. Most of the mutations are amino acid substitutions compared with a few stop codon mutations and deletions or insertions. The effect of several mutations on channel function was examined in heterologous expression systems. CNGA3 subunits carrying mutations in the S1 region (Y181C, N182Y, L186F, and C191Y) failed to be targeted to the plasma membrane, indicating that the folding and/or subunit assembly was disturbed in these mutants. Impaired channel targeting may also explain the loss of function induced by several other mutations. Alternatively, mutations may alter the biophysical properties of the channels. For example, the N471S and R563H mutants were found to have an increased cGMP affinity which may prevent complete closure of the channels in response to light stimulation. Consequently, these 'gain-of-function' mutations may lead to a sustained and potentially toxic elevation of intracellular Ca^{2+} levels in cone outer segments. A T369S mutation in the pore region of CNGA3 has been linked to incomplete achromatopsia. When co-expressed with the CNGB3 subunit, this mutant induced robust currents. However, the channel revealed a lower efficacy of Ca^{2+} blockage at physiological membrane potentials which tentatively would lead to a lower signal-to-noise ratio of the light response and, hence, explain the lower light sensitivity observed in the patients.

CNGA3-deficient mice display the principal hallmarks of achromatopsia described previously. As the only exception, deletion of CNGA3 has no obvious effect on visual acuity. This finding may be explained by the fact that mice lack the cone-enriched macula that confers high-acuity vision in humans. CNGA3-deficient mice reveal a complete loss of cone function. By contrast, the rod photoreceptor system remains fully intact (structurally and functionally) over the life span of the animals. Cones of CNGA3-deficient mice show a delayed postnatal migration behavior, form irregular synapses, and reveal a progressive degeneration. Interestingly, dorsal cones expressing medium-wavelength sensitive opsin survive significantly longer (up to 22 months) than ventral cones expressing mostly short-wavelength sensitive opsin (<3 months). The reason for this difference remains to be determined. CNGA3-deficient cones fail to transport opsins into outer segments, downregulate various proteins of the phototransduction cascade, and induce apoptotic death. As discussed previously for CNGA1/B1-deficient rods, cones of CNGA3-deficent mice may degenerate as a result of the constitutive activation of the visual cascade.

The CNGA3 subunit has been identified in a number of tissues different from retina, including a subpopulation of oligoribonucleotides (ORNs), sperm, kidney, and brain. CNGA3-deficient mice reveal normal fertility, indicating that CNGA3 is not essential for sperm function. There is evidence that CNGA3 may be involved in the modulation of long-term potentiation (LTP) in some brain regions and the formation of amygdala-dependent memory.

CNGB3 Subunit

A missense mutation in the CNGB3 gene (S435F) was identified in colorblind individuals originating from the Pingelap Atoll of Micronesia. On this small island, achromatopsia is very frequent and it affects nearly 10% of the native population. Several other mutations of the CNGB3 gene have been identified, most

of which lead to truncated proteins. Overall, mutations in the CNGB3 gene account for approximately 50% of all cases of achromatopsia. A missense mutation in exon 6 of the CNGB3 gene (D262N, corresponding to human D267N) and a CNGB3 null deletion were also identified in two canine inbred strains (Table 1) with cone degeneration.

The S435F mutation has been functionally characterized in *Xenopus* oocytes. CNGA3/CNGB3$_{S435F}$ heteromeric channels exhibited a more than fourfold increase in cAMP affinity and a modest increase in apparent affinity for cGMP. In addition, the mutant channel had a slightly higher open probability in the presence of saturating cyclic nucleotide concentration than wild-type channels. It is unclear whether these subtle effects can account for the severe clinical phenotype of the patients. Alternatively, it is possible that the mutation prevents correct protein folding and/or assembly with the CNGA3 subunit *in vivo*. The finding that null mutations of the CNGB3 gene induce achromatopsia strongly suggests that the CNGA3 subunit (like the CNGA1 subunit) is not able to form homomeric channels *in vivo*.

Knockout Models of the Olfactory CNG Channel

CNGA2 Subunit

Unlike all other CNG channel genes, the CNGA2 gene is localized on the X chromosome. Mutations of this gene have not been identified in humans. However, four different CNGA2-deficient mouse lines have been generated to define the physiological role of this subunit. CNGA2-deficient mice display elevated neonatal lethality due to their inefficiency to locate the mother and suckle milk. Indeed, initial electroolfactogram (EOG) recordings showed that the mice exhibit no detectable responses to odorants. Moreover, in perforated patch experiments, ORNs of CNGA2-deficient mice did not respond to odors or activators of the cAMP system. Together, these findings indicate that olfaction is conferred by a cAMP/CNGA2-mediated signaling pathway. In addition, the results strongly argue against a major contribution of other signaling cascades which have been suggested in the past (e.g., IP3-mediated olfaction). Studies have confirmed this principal conclusion and also uncovered that CNGA2-deficient mice are not totally anosmic. Very small but reproducible EOG responses to a variety of odorants were found in CNGA2-deficient mice. In behavioral tests, CNGA2-deficient mice could detect and discriminate several odorants. The molecular components conferring this residual olfaction remain to be determined. Pharmacological blockers suggest that the respective signaling pathway(s) is cAMP independent. There is evidence that a small subgroup of ORNs projecting to the so-called necklace glomeruli could be involved in non-cAMP-mediated olfaction. These neurons do not express CNGA2 but, rather, express the CNGA3 channel and several other components that support a cGMP signaling pathway.

In addition to its key role in olfactory transduction, CNGA2 is required for normal development of the olfactory epithelium and olfactory bulb. In comparison to wild-type mice, CNGA2 knockouts display a thinner olfactory epithelium and show lower expression of an olfactory marker protein. Moreover, the olfactory bulb is smaller, and tyrosine hydroxylase expression is reduced in the majority of periglomerular neurons but is retained in necklace glomeruli.

The olfactory CNG channel was identified in murine and rat hippocampus, suggesting a role of this protein in memory formation. In support of this hypothesis, hippocampal LTP evoked by a theta burst stimulation protocol was reduced (Table 1) in CNGA2-deficient mice. By contrast, several measures of basal synaptic activity as well as LTP induced by high-frequency stimulation seemed to be similar in wild-type and knockout mice. Behavioral tests to determine the memory performance of CNGA2-deficient mice have not been reported, leaving the question of whether or not the observed alterations of the hippocampal LTP are physiologically relevant unanswered.

CNGA4 and CNGB1b Subunits

Mice deficient for the CNGA4 subunit are fertile and show no obvious morphological differences with respect to wild-type mice. Moreover, the olfactory epithelium and the olfactory bulb are normally developed in these mice. Maximal EOG responses induced by 3-isobutyl-1-methylxanthine (IBMX), an inhibitor of the phosphodiesterase, are comparable in both genotypes. However, the dose–response curves of the response are shifted to approximately threefold higher IBMX concentrations in the CNGA4-deficient mice, suggesting that the mutant CNG channel is less sensitive to cAMP. Indeed, the apparent cAMP affinity of the CNG channel as determined from excised patch recordings is approximately tenfold lower in the knockouts than in wild-type mice. Thus, CNGA4 is clearly required to confer high cAMP affinity to the olfactory channel. CNGA4-deficient mice also show a defect in odor adaptation (Table 1). The decay time of the EOG in response to an odorant stimulus is profoundly slower in knockouts than in wild-type mice. Electrophysiological recordings in excised patches from ORNs clarified why CNGA4 is needed

for odor adaptation. The recordings revealed that in the absence of CNGA4, the kinetics of Ca-calmodulin-dependent inhibition are profoundly slower than in its presence. Thus, CNGA4 is required for fast feedback inhibition of channel activity by Ca-calmodulin. The results also suggested that the Ca-calmodulin binding site identified in the CNGA2 subunit is probably not involved in Ca-calmodulin-mediated inhibition *in vivo*. Another study confirmed this notion by demonstrating that apocalmodulin is preassociated with an IQ motif present in the C-linker of CNGA4. The authors also showed that another IQ motif present in the N-terminus of the CNGB1b subunit is necessary for inhibition. It was speculated that one apocalmodulin molecule binds to both sides, thereby cross-linking the CNGA4 and the CNGB1b subunit. If Ca^{2+} flows through the channel pore, it interacts with the preassociated apocalmodulin and induces a conformational change in the channel that decreases its cAMP sensitivity. As a consequence, the CNG channel closes and the olfactory transduction pathway is switched off.

RP patients carrying the G993V mutation were reported to show no olfactory impairment. In contrast, CNGB1-deficient mice that lack both the retinal CNGB1a and the olfactory CNGB1b subunit reveal a profound olfactory phenotype (**Table 1**). As reported for CNGA2-deficient mice, CNGB1 knockouts show an increased postnatal mortality that is probably due to a deficit in locating their mother's nipples to suckle milk. The knockout mice have a decreased olfactory performance and their EOG responses are significantly smaller than those of wild-type littermates. Moreover, fast odor adaptation is impaired, confirming that CNGB1 is absolutely necessary for Ca^{2+}-calmodulin-dependent inhibition. Interestingly, protein levels of the CNGA2 and the CNGA4 subunits are dramatically reduced in the cilia of CNGB1 knockout mice. Similarly, ciliary expression levels of CNGB1b and CNGA2 are extremely low in CNGA4-deficient mice. These findings show that ORNs have evolved a mechanism to control the composition of the CNG channel in the cilia to ensure that only channels containing all three subunits are available for olfactory signal transduction.

See also: Neuronal Pacemaking; Photoreceptors: Physiology; Phototransduction; Retinitis Pigmentosa.

Further Reading

Biel M, Seeliger M, Pfeifer A, et al. (1999) Selective loss of cone function in mice lacking the cyclic nucleotide-gated channel CNG3. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7553–7557.

Brunet LJ, Gold GH, and Ngai J (1996) General anosmia caused by a targeted disruption of the mouse olfactory cyclic nucleotide-gated cation channel. *Neuron* 17: 681–693.

Craven KB and Zagotta WN (2006) CNG and HCN channels: Two peas, one pod. *Annual Review of Physiology* 68: 375–401.

Hofmann F, Biel M, and Kaupp UB (2005) International Union of Pharmacology: LI. Nomenclature and structure–function relationships of cyclic nucleotide-regulated channels. *Pharmacological Reviews* 57: 455–462.

Hüttl S, Michalakis S, Seeliger M, et al. (2005) Impaired channel targeting and retinal degeneration in mice lacking the cyclic nucleotide-gated channel subunit CNGB1. *Journal of Neuroscience* 25: 130–138.

Kaupp UB and Seifert R (2002) Cyclic nucleotide-gated ion channels. *Physiological Reviews* 82: 769–824.

Ludwig A, Budde T, Stieber J, et al. (2003) Absence epilepsy and sinus dysrhythmia in mice lacking the pacemaker channel HCN2. *EMBO Journal* 22: 216–224.

Michaelides M, Hunt DM, and Moore AT (2004) The cone dysfunction syndromes. *British Journal of Ophthalmology* 88: 291–297.

Munger SD, Lane AP, Zhong H, et al. (2001) Central role of the CNGA4 channel subunit in Ca^{2+}-calmodulin-dependent odor adaptation. *Science* 294: 2172–2175.

Nolan MF, Malleret G, Dudman JT, et al. (2004) A behavioral role for dendritic integration: HCN1 channels constrain spatial memory and plasticity at inputs to distal dendrites of CA1 pyramidal neurons. *Cell* 119: 719–732.

Nolan MF, Malleret G, Lee KH, et al. (2003) The hyperpolarization-activated HCN1 channel is important for motor learning and neuronal integration by cerebellar Purkinje cells. *Cell* 115: 551–564.

Robinson RB and Siegelbaum SA (2003) Hyperpolarization-activated cation currents: From molecules to physiological function. *Annual Review of Physiology* 65: 453–480.

Stieber J, Herrmann S, Feil S, et al. (2003) The hyperpolarization-activated channel HCN4 is required for the generation of pacemaker action potentials in the embryonic heart. *Proceedings of the National Academy of Sciences of the United States of America* 100: 15235–15240.

Stieber J, Hofmann F, and Ludwig A (2004) Pacemaker channels and sinus node arrhythmia. *Trends in Cardiovascular Medicine* 14: 23–28.

Relevant Websites

http://iuphar-db.org – International Union of Pharmacology Receptor Database.

http://www.sph.uth.tmc.edu – The University of Texas School of Public Health Retinal Information Network.

Cysteine-String Proteins (CSPs)

F Schmitz, Saarland University Medical School, Homburg/Saar, Germany
R Fernández-Chacón, Universidad de Sevilla, Seville, Spain

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Intercellular communication at chemical synapses occurs via exocytosis of neurotransmitter-containing synaptic vesicles in presynaptic nerve terminals. In the synaptic vesicle cycle, neurotransmitter-filled synaptic vesicles translocate (dock) to the presynaptic plasma membrane, where they are transformed into a fusion-competent primed state, ready for exocytosis. After a suprathreshold depolarization (usually an action potential), voltage-gated Ca^{2+}-channels open and the resulting Ca^{2+} influx triggers the completion of the final fusion event and neurotransmitter release. After exocytosis, synaptic vesicle membrane is retrieved and recycled for a new round of the synaptic vesicle cycle. The synaptic vesicle cycle is the result of extremely fast, precisely timed, and highly regulated cascades of protein–protein interactions. The assembly and disassembly of protein–protein complexes relies on the proper folding and native conformation of the participating proteins. Presynaptic nerve terminals work as largely autonomic subcellular units that are often separated by large distances from the biosynthetic machinery of the cell body. Consequently, dysfunctional proteins cannot be immediately replenished. This suggests that nerve terminals are equipped with first-aid mechanisms that protect them from the consequences of malfolded or denatured proteins (aged in the course of duty during continuous activity in the synaptic vesicle cycle).

Chaperone proteins are important components of the cellular rescue machinery. They ensure the proper folding of proteins, participate in the refolding of damaged or aged proteins and the untangling of protein aggregates, and aid in the stabilization and restructuring of protein conformations. Chaperone proteins typically bind to short stretches of hydrophobic amino acids that are exposed on partially or completely unfolded proteins and stabilize them until they are translocated into the hydrophobic interior of the protein or covered by other protein–protein interactions. Many protein–protein interactions involve hydrophobic stretches of amino acids which eventually need to be shielded from the aqueous surrounding if the protein–protein complex is disassembled. Obviously, the maintenance of correct protein folding is particularly important in presynaptic nerve terminals, where many protein–protein interactions are assembled and disassembled.

CSPα is an essential part of the chaperone system in presynaptic terminals. The importance of this chaperone protein was initially appreciated in *Drosophila* and *Drosophila* mutants and was recently highlighted by several studies analyzing CSPα-deficient knockout mice. Important properties of the (CSP) based chaperone system, including insights that emerged from the analyses of these CSPα-deficient mice, are summarized in this article.

Molecular Anatomy of CSPs: J-Domain Proteins Highly Conserved in Evolution

CSP genes, cDNAs, and proteins have been characterized in a variety of species and are highly conserved in evolution (Figure 1(a)). Invertebrates (e.g., *Drosophila*) have a single CSP gene, whereas mammals have three CSP genes (CSPα, CSPβ, and CSPγ). CSPα is the main isoform in the brain; CSPβ and CSPγ transcripts are absent from the brain as indicated by northern blot analyses. Hair cells from the inner ear express both CSPα and CSPβ; the testis expresses all CSP isoforms. CSPβ and CSPγ have same domain structure as CSPα, but their function is entirely unknown.

CSPs are characterized by distinct domains (Figure 1):

- An N-terminal Dna-J domain, the signature domain of Hsp40-type co-chaperones. J domains comprise a stretch of approximately 70-amino-acids which is highly conserved from bacteria (Dna-J protein) to humans.
- A central string of multiple cysteine residues. In mouse CSPα, 13 cysteine residues are found within a 22-amino-acid stretch present in the central portion of the protein.
- The linker region (L region) between the J domain and the cysteine-string region. The linker domain of CSP is highly conserved in evolution and has been implicated in regulating exocytosis.
- A variable C-terminus. The C-terminus is subject to differential splicing and generates C-terminally truncated proteins in *Drosophila* and mammals (named CSP1, CSP2, and CSP3; Figure1(a)).

The J domain of CSP mediates binding to Hsp70-type chaperone proteins. Hsp70-type chaperone proteins are a large family of evolutionary highly conserved proteins that includes the constitutively expressed

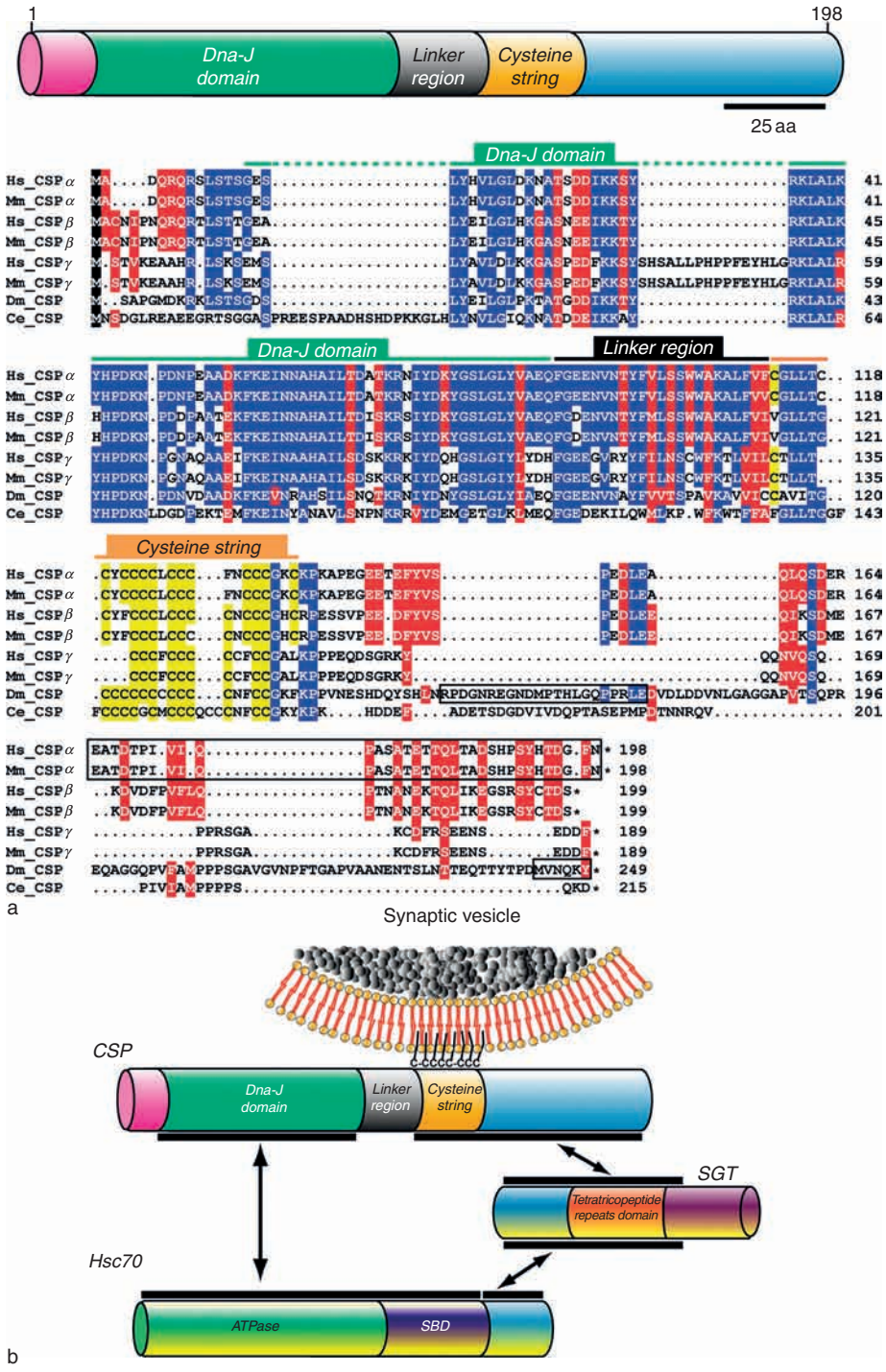

Figure 1 Primary structure of vertebrate and invertebrate CSPs: (a) general domain structure (top) and sequence alignment of CSPs (bottom); (b) the CSP–Hsc70–SGT complex. In (a), the amino acid sequences of human (Hs), mouse (Mm), *Drosophila melanogaster* (Dm), and *Caenorhabditis elegans* (Ce) are aligned for maximal homology using the single-letter amino acid code. The three mammalian CSPs are identified on the left as CSPα, CSPβ, and CSPγ; the single CSP genes in *Drosophila* and *C. elegans* are Dm CSP1 and Ce CSP, respectively. Residues that are identical in > 50% of the sequences are highlighted in blue, and residues that are similar or identical in 50% of the sequences are in red. The initiator methionines are shown in white on a black background, and the cysteine repeats are highlighted in yellow. Sequence in empty boxes at the C-terminus indicate alternatively spliced regions. The protein topology is illustrated by a colored horizontal bar above the sequence, with the Dna-J domain indicated by green and the linker region in black. Asterisks at the C-termini of the sequences indicate stop codons. Genbank accession number for the sequences analyzed are NP_079495 (Hs CSPα), NP_149096 (Hs CSPβ), NP_775921 (Hs CSPγ), AB87080 (Mm CSPα), Q9UF47 (Mm CSPβ), Q8N7S2 (Mm CSPγ), Q03751 (Dm CSP1), and NAAK39254 (Ce CSP). As shown in (b), CSP binds via its J domain to the ATPase domain of Hsc70. The TPR repeats of SGT bind to the C-terminus of Hsc70 as well as to the cysteine-string region and the C-terminus of CSP. Thus, binding of Hsc70 to CSP is secured by two different molecular connections. ATP, adenosine triphosphate; CSP, cysteine-string protein; SBD, substrate-binding domain; SGT, small glutamine-rich tetratricopeptide; TPR, tetratricopeptide repeat.

heat-shock cognate protein Hsc70, the stress-induced Hsp70, and many other Hsp70-type members. Hsp70 proteins bind to short, strongly hydrophobic stretches on proteins. These hydrophobic stretches are often found on denatured proteins or on disassembled protein complexes which are mediated by hydrophobic interactions.

Hsp70 proteins contain a N-terminal adenosine triphosphatase (ATPase) domain and a C-terminal substrate-binding domain (SBD). The ATPase domain binds ATP/adenosine diphosphate (ADP) with distinct affinities. The intrinsic ATPase activity of Hsp70 is low, but it can be strongly activated by the binding of the J domain to the ATPase domain of Hsp70 and probably, although to a lesser extent, also by the binding of client proteins to the SBD.

In the ATP-bound form, the affinity for client proteins is low. Nevertheless, ATP-Hsp70 does bind client proteins; however, binding is transient and release occurs rapidly. Binding of the J domain and subsequent ATP hydrolysis locks Hsp70 into its high-affinity conformation in which the refolding of malfolded proteins probably takes place. Subsequent binding of a novel ATP promotes the release of renatured client proteins, making Hsp70 ready for a new round of the chaperone cycle (Figure 2). In addition to the activation of ATPase activity by CSP binding, the ATPase activity of Hsp70 is further enhanced by the simultaneous binding of the small, glutamine-rich tetratricopeptide repeat (TPR) protein small glutamine-rich tetratricopeptide (SGT) (Figures 1(b) and 2). SGT binds to the cysteine-rich region of CSP (in cooperation with the CSP C-terminus). Previously, SGT had also been shown to bind to the C-terminus of Hsc70, and thus it represents a bridge molecule between CSP and Hsp70. SGT is ubiquitously expressed. Recently, a brain-specific isoform of SGT, called SGTβ, has also been described.

The cysteine residues in the cysteine-rich central portion of CSP are heavily palmitoylated and

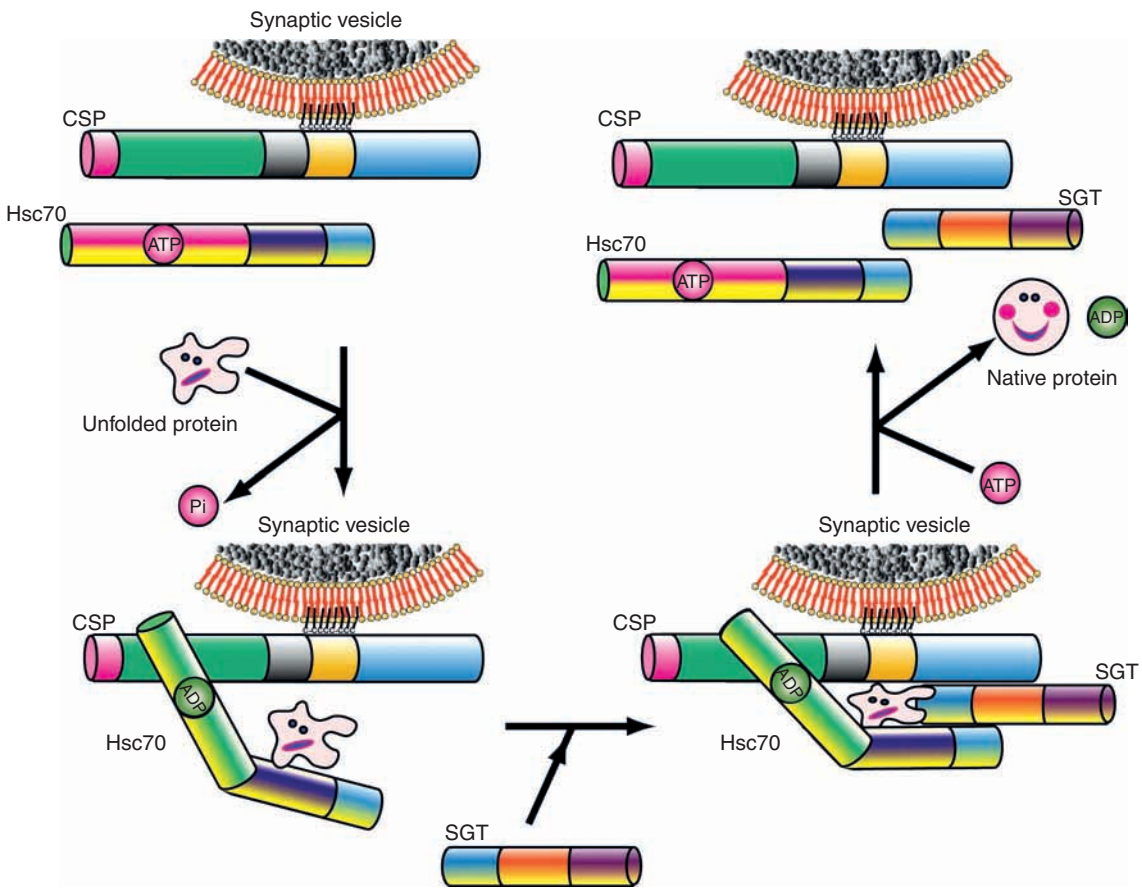

Figure 2 Mechanistic model of the CSP–Hsp70–SGT chaperone complex at the synapse. CSP is anchored via its acylated cysteine residues in the cysteine-string region to the synaptic vesicle membrane. Then, a complex formed by Hsc70 and denatured client protein recruits the TPR protein SGT. The binding of CSP to Hsc70 induces ATP hydrolysis and a conformational change of Hsc70, which provides the energy for the refolding of the denatured client protein. The subsequent binding of ATP releases the renatured client protein and induces the disassembly of the trimeric complex. ADP, adenosine diphosphate; ATP, adenosine triphosphate; CSP, cysteine-string protein; Pi, inorganic phosphate; SGT, small glutamine-rich tetratricopeptide.

responsible for membrane targeting. Vesicle-bound CSP recruits Hsp70 to synaptic vesicles (Figure 1(b)). The trimeric complex consisting of CSP, Hsp70, and SGT is formed *in vivo* on the surface of synaptic vesicles and is able to refold denatured luciferase *in vitro*.

CSP Is Localized to Synaptic Vesicles and Secretory Granules

CSP proteins are associated with the membrane of synaptic vesicles. Originally, CSP was assumed to be brain-specific. Further studies clarified that CSP is also expressed outside the brain, in the endocrine, neuroendocrine, and exocrine tissues. In exocrine and endocrine glands, CSP has been localized to various secretory granules: the chromaffin granules in chromaffin cells, zymogen granules in exocrine pancreatic cells, insuline-containing secretory granules of neuroendocrine β-cells, hormone-containing secretory granules of the rat neurohypophysis, and secretory granules in secretory epithelia (e.g., the mammary gland). The localization in synaptic vesicles and secretory granules indicates an importance of CSP for secretion. CSP was also found in some nonsecretory tissues (e.g., the heart, kidney, spleen, testis, and adipose tissue), although in lower amounts. The significance of CSP at the latter localizations is not well understood.

Insights from CSP-Deficient Flies

The deletion of the CSP gene in *Drosophila* using transposon techniques is semi-lethal. The majority of flies without a functional CSP gene die. Only a minor fraction (4%) survives. These suffer from uncoordinated motor behavior and die prematurely. Interestingly, they become totally paralyzed and die within hours when the temperature is increased. At the cellular level, motor deficits result from a presynaptic defect in neurotransmission. Initially, it was believed that the recycling of vesicles in the presynaptic terminals would be the prime defect in CSP-deficient flies because Hsc70 (in cooperation with the J domain protein auxilin) mediates the uncoating of clathrin-coated vesicles. Further analyses, however, showed that the endocytotic pathway is not impaired in CSP-deficient flies. Instead, CSP is necessary for normal, evoked neurotransmitter release in flies. Characteristically, the presynaptic secretion defect in mutant flies was temperature dependent. Neurotransmitter release and evoked postsynaptic responses were already strongly reduced at 22 °C; at higher temperatures (29 °C or 32 °C), evoked responses were nearly completely lost. The temperature-dependent block of synaptic transmission is consistent with an increase in protein instability or unfolding in the absence of the CSP chaperone function. The temperature-sensitivity of the defects and the existence of escapers are untypical for a distinct and uniform loss-of-function effect, but are more compatible with a main role of CSP as a chaperone that works on denatured synaptic client proteins. Although a very useful and important model system was available with the CSP-deficient flies, pinpointing the precise molecular function of CSP and explaining the molecular defect in secretion in CSP mutant flies remained difficult.

Studies of CSP Interactions with Voltage-Dependent Ca^{2+} Channels and Other Presynaptic Proteins

Several studies suggested that voltage-dependent presynaptic Ca^{2+} channels (VDCCs) and their modulation are main targets of CSP. Early studies on CSP function described that the expression of Torpedo N-type VDCCs in oocytes was inhibited on injection of CSP antisense RNA. Later, the putative connection between CSP and VDCCs was supported by imaging studies of exocytosis in nerve terminals of *Drosophila* CSP mutants. Those studies showed that neurotransmitter release of CSP mutants was normalized on the application of Ca^{2+} ionophores that bypassed VDCC activation but not on the application of depolarizing solutions. Biochemical studies with vertebrate CSP led researchers to propose that CSP directly binds VDCCs by a direct interaction of CSP with the synaptic protein interaction (synprint) site of P/Q-type and N-type VDCCs and indirectly via G-protein subunits that regulate N-type VDCCs. Nevertheless, other observations challenged the CSP–Ca^{2+} channel hypothesis. Patch-clamp measurement of L-type Ca^{2+} currents at peptidergic nerve terminals of *Drosophila* CSP mutants showed no differences from wild-type flies. In addition, direct measurements of cytosolic Ca^{2+} levels in synapses of *Drosophila* CSP mutants revealed that Ca^{2+} influx elicited by depolarization was even higher than normal, ruling out that a deficit in Ca^{2+} influx through VDCC caused the impairment in neurotransmitter release. Instead, CSP appeared to more directly influence the synaptic vesicle exocytotic machinery. This notion was supported by experiments of CSP overexpression in neuroendocrine cells that showed altered exocytosis but had no effect on Ca^{2+} channel activity, arguing against a major importance of CSP in regulating VDCCs.

In conclusion, the reported heterogeneous abnormalities of VDCC function did not prove there is a major role for CSP as a physiological regulator of VDCC function. The observed multiple, and partly controversial, observations can be well explained by context-dependent activation of the CSP chaperone system induced by partially unfolded and/or aged proteins. Such an argument is particular relevant in overexpression experiments in which transfected cells are forced to overexpress proteins that might be difficult to fold (e.g., membrane proteins with multiple membrane-spanning regions) and which – if unproperly folded – will activate the CSP chaperone system. Also overexpression of CSP constructs could generate unphysiological side effects due to the sequestration of the physiological interaction partner of CSP and/or due to an overload of the cellular palmitoylation system.

Further studies proposed the interaction of CSP with key proteins implicated in synaptic vesicle fusion. Using CSP fusion protein as a bait protein, synaptotagmin 1 was identified as a potential interacting protein for CSP in rat brain homogenates. The interaction has been described to be phosphorylation dependent. A genetic and biochemical interaction between CSP and syntaxin in *Drosophila* has been controversially reported. Syntaxin was suggested as a possible target of CSP because the overexpression of syntaxin 1A compensated for the loss of CSP in *Drosophila*. A direct interaction between CSP and syntaxin was observed in some studies but not in others. Synaptobrevin 2 and CSP were found to co-immunoprecipitate from rat brain homogenates, although a direct interaction between these proteins was not found *in situ*. Only in cell-culture experiments that employed an overexpression of proteins as CSP reported to directly interact with synaptobrevin 2.

Recently, the vesicular γ-aminobutyric acid (GABA) transporter (VGAT) was found in a complex with CSP, Hsc70, and glutamate decarboxylase (GAD), suggesting an involvement of CSP in neuratransmitter synthesis and filling of synaptic vesicles with the neurotransmitter GABA. Also, Rab-α-guanosine diphosphate (GDP)-dissociation inhibitor (GDI), which recognizes GDP-complexed rab proteins and solubilizes rab-proteins from synaptic vesicle membranes after vesicle fusion, has been found in a complex that contained Hsc70 and CSP. This finding might indicate that either particular proteins of the complexes or the stability of the complex itself need to be chaperoned to maintain the required enzymatic activity. Yet these observations are still based on biochemical experiments, and the physiological relevance has not yet been determined.

Elucidating the Biological Function of Cysteine-String Proteins in CSPα-Deficient Mice

Recently, CSPα knockout mice (CSPα KO mice) were generated in which the CSPα gene was inactivated by homologous recombination, allowing a detailed molecular and functional analysis of CSPα in mouse. CSPα is the only CSP isoform in brain.

Immediately after birth, CSPα-deficient animals were normal and did not show any obvious defects. The synapse structure and function of CSPα-deficient mice was analyzed at different postnatal stages. Although the defects were initially inconspicuous, CSPα-deficient mice started to suffer from progressive muscle weakness and a severe deficiency in locomotive behavior that began between the second and third postnatal weeks of life. Locomotor defects in CSPα-deficient mice (e.g., delay in righting time (the time it took to get into the upright position after mice have been laid on their back)) dramatically increased with age. Parallel to the developing muscle weakness and deficits in locomotor performance, CSPα-deficient mice showed abnormal, decreased mean compound muscle action potentials (CMAPs) at postnatal day (P)45 but – most interestingly – not at P13 or P23. Thus, defects in CSPα-deficient mice were prominent in older animals but not yet present in young mice. These physiological defects in muscle strength and CMAPs correlated with a morphological degeneration in the presynaptic nerve terminal of the neuromuscular junction (NMJ), indicating a presynaptic defect.

Certain fast synapses in the mouse, which are absent in *Drosophila*, are enormously useful for the quantitative analysis of Ca^{2+} currents and neurotransmitter release. In particular, giant presynaptic terminals of the calyx of Held synapse in the auditory pathway of the brain stem are ideal for measuring VDCCs because the presynaptic terminals are directly accessible for patch-clamp recordings. Interestingly, detailed measurements of P/Q- and N-type Ca^{2+} currents performed at the calyx of Held did not reveal any abnormalities in CSPα KO mice compared to control littermates. The absence of any defects in VDCC function cannot be explained by compensation of the CSPα-deficiency by other CSP isoforms because CSPα is the only CSP isoform in the brain. Similarly, also L- and T-type VDCCs in ribbon synapses of the inner ear were found to be normal in CSPα KO mice. These findings indicate that in the physiological context CSPα is not essential for regulating voltage-gated Ca^{2+} channels.

Synaptic transmission was measured in the calyx synapse by using afferent fiber stimulation and by

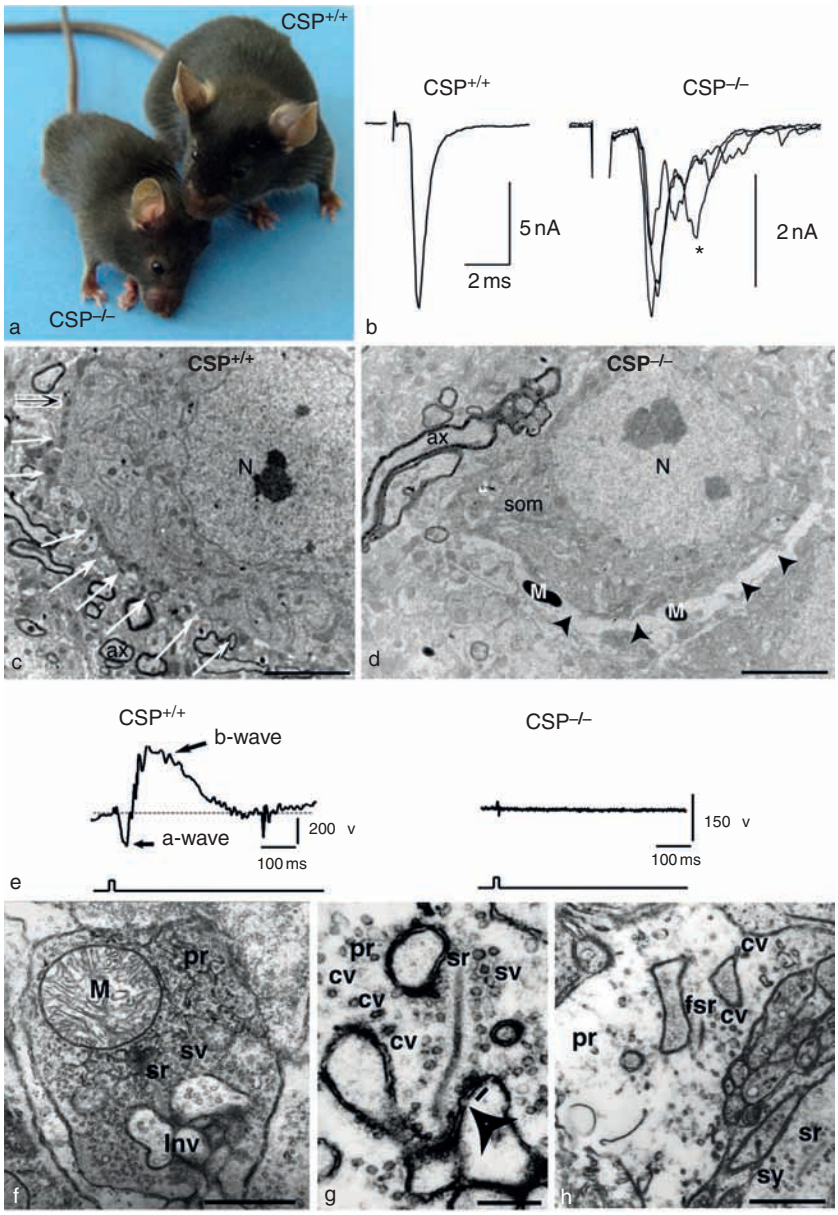

Figure 3 Use-dependent loss of function and presynaptic degeneration in CSPα-deficient mice, revealing the consequences of the loss of the chaperone activity *in vivo*: (a) at 4 weeks of age, CSPα-deficient mice (CSP$^{-/-}$) are significantly smaller and show severe defects in the motor system compared to control littermates (CSP$^{+/+}$); (b) asynchronous, delayed, and amplitude-decreased EPSCs in the calyx of Held of CSP mutants (CSP$^{-/-}$) compared to control littermates (CSP$^{+/+}$) at postnatal day (P)20–P23, indicating alterations in neurotransmitter release (note the different scales in the wild-type and knockout graphs); (c) presynaptic calyx terminals in wild-type animals (CSP$^{+/+}$); (d) presynaptic calyx terminals in CSPα-deficient knockout mice (CSP$^{-/-}$); (e) electroretinogram (ERG) activity in wild-type and CSPα-deficient knockout mice; (f) wild-type photoreceptor ribbon synapses; (g, h) degenerated presynaptic photoreceptor ribbon terminals of CSPα-deficient mice. The asterisk in (b) indicates postsynaptic currents of extra-calycal origin. As shown in (c, d), the functional defects observed by electrophysiology are accompanied by a morphological degeneration of the presynaptic calyx terminals. In wild-type animals (c), the typical large, cup-shaped presynaptic terminal can be observed, which is filled with synaptic vesicles and mitochondria. In CSPα-deficient mice (d), the presynaptic terminal shows a pathological translucent appearance with few synaptic vesicles and condensed mitochondria (M). As shown in (e), at P45, CSP$^{-/-}$animals are completely blind and show no electrical activity in the electroretinogram (ERG); both the b-wave and the a-wave of the ERG (present in the wild-type animals) have disappeared. The loss of vision is accompanied by a morphological degeneration of the presynaptic ribbon photoreceptor terminals; the wild-type photoreceptor ribbon synapses (f) are characterized by large presynaptic terminals which are filled with synaptic vesicles and which contain a synaptic ribbon (sr) anchored to the active zone. As shown in (g and h), the degenerated presynaptic photoreceptor ribbon terminals of CSPα-deficient mice have a strongly decreased density of synaptic vesicles, an increased number of coated vesicles (cv), and an abnormal architecture of the presynaptic terminal. Remarkably, in the active zone of CSPα-deficient mice omega-shaped exocytotic figures could be occasionally observed. These exocytotic figures are hard to capture via conventional electron microscopy due to the high speed of exocytosis and the slow velocity of chemical fixation. The visibility of exocytotic fusion events and, in addition, rows of synaptic vesicles docked to the active zone waiting for exocytosis ((g) indicated by the solid line next to the arrowhead) could

parallel, paired recordings of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor-mediated postsynaptic responses. At early stages (P9–P11), measurements of evoked excitatory postsynaptic currents as response to presynaptic fiber stimulation did not reveal any alterations in the amplitude and kinetics of postsynaptic responses. This indicates normal coupling of Ca^{2+} channels and the exocytotic machinery, as well as a normal performance of exocytotic machinery, ruling out significant defects in exocytosis at that time. In addition to evoked postsynaptic responses, spontaneous miniature excitatory postsynaptic currents (mEPSCs) were also found to be normal in terms of kinetics and amplitude in CSPα-deficient mice.

In contrast to the normal synaptic function at P9–P11, synaptic function at P20 showed severe defects in the calyx synapses of the brain stem. Evoked EPSCs were asynchronous and lower in amplitude than in the wild-type animals (Figure 3(b)). Interestingly, in wild-type animals, the calyx synapses showed synaptic depression after repeated stimulation; in contrast, the calyx synapses of CSPα-deficient mice at P20 showed a synaptic facilitation instead of synaptic depression. Such an increase in synaptic facilitation is consistent with a decrease in release probability. In contrast to the decreased evoked postsynaptic responses, amplitudes of the spontaneous mEPSCs were increased, which might be due to a compensatory upregulation of postsynaptic receptors. These physiological changes in CSPα-deficient mice were accompanied by a structural degeneration of the presynaptic terminals.

Progressive synaptic defects and neurodegeneration were also observed in the retina and particularly affected the tonically active photoreceptor ribbon synapses. Initially, after eye opening at P14, electroretinogram (ERG) analyses demonstrated that vision was only slightly affected in CSPα KO mice but vision faded completely away by P45 (Figure 3(e)). This loss of vision correlated with a morphological degeneration of the photoreceptor ribbon synapse (Figures 3(g) and 3(h)). Neurodegeneration started in the presynaptic ribbon synapses and initially spared other synapses, indicating that neurodegeneration starts in synapses which are particularly active. These data indicate that CSPα is necessary to keep synapses running for extended periods of time and to prevent neurodegeneration. Interestingly, tonically active ribbon synapses from inner-ear hair cells were spared from structural and functional defects because of their expression of an additional CSP isoform, CSP*β*.

These findings of CSPα support a major role of CSP as a synaptic chaperone protein because defects were initially small and then dramatically increased with age. Consistently, synapses that are heavily used are the first ones that degenerate in a principally ubiquitous degeneration of presynaptic nerve terminals. CSPα-deficient mice finally died after 2–3 months.

Recent studies by the Südhof group demonstrated that synaptosomal-associated protein of 25 kDa (SNAP-25) is an important substrate of CSPα. In CSPα-deficient mice, SNAP-25 levels and the formation of the soluble *N*-ethylmaleimide-sensitive factor attached protein receptor (SNARE) complex of exocytosis are decreased. The assembly and dissociation of the SNARE complex require hydrophobic interactions, which probably need protective shielding in the nonassembled form. It is conceivable that CSPα is a chaperone that maintains the SNARE complex during multiple rounds of exocytosis. This hypothesis is further supported by recent findings demonstrating that the protein α-synuclein, a protein whose gene is mutated in cases of familial Parkinson's disease, enhances SNARE complex formation and, by this means, is able to compensate for the defects generated by CSPα-deficiency. The neuronal overexpression of α-synuclein in transgenic mice prevented the neurological and lethal phenotype of CSPα KO mice. Furthermore, the analysis of KO mice lacking both CSPα and α-synuclein accelerated the neurological phenotype of CSPα KO mice. These experiments convincingly showed that α-synuclein functionally cooperates with CSPα to prevent synaptic degeneration. Remarkably, the rescue occurs only if α-synuclein is able to bind phospholipids because transgenic expression of the A30P mutant of α-synuclein, a Parkinson-associated point mutant not able to bind phospholipids, did not rescue early lethality in CSPα KO mice. Synuclein itself does not interact with Hsc70 or SGT but, rather, acts downstream of CSPα. However, the cooperative action between α-synuclein and CSPα in

indicate delays in the exocytotic fusion process. Such a delay has been documented by electrophysiology (b). Scale bar = 5 μm (c), 2.5 μm (d), 1 μm (f), and 500 nm (g, h). ax, axon; CSP, cysteine-string protein; cv, coated vesicle; EPSC, excitatory postsynaptic current; fsr, floating [detached] synaptic ribbon; Inv, invagination of postsynaptic dendrites into the presynaptic terminal (typical normal morphology); M, mitochondria; N, nucleus; pr, presynaptic terminal; som, soma of the principal neuron in the trapezoid body (superior olivary complex of the brain stem); sr, synaptic ribbon; sv, synaptic vesicle; sy, synapse. (b, d) reprinted from *Neuron*, 42, Fernandez-Chacon R, Wölfel M, Nishimune H, et al., The synaptic vesicle protein CSPα prevents presynaptic degeneration, 237–251, Copyright (2004), with permission from Elsevier. (e, h) Reproduced from Schmitz F, Tabares L, Khimich D, et al. (2006) CSPα-deficiency causes rapid and massive photoreceptor degeneration. *Proceedings of the National Academy of Sciences of the United States of America* 103: 2926–2931, with permission from National Academy of Sciences, USA.

neuroprotection involves chaperoning of the SNARE complex in a reaction that requires phospholipids binding.

Interestingly, SNAP-25 levels were already present in decreased amounts at early stages (P5), before signs of neurodegeneration were obvious. This further indicates that SNARE proteins are particularly important synaptic substrates of CSP-containing chaperone complexes. Initially, the remaining SNAP-25 might be sufficient to allow core complex formation. But prolonged activation of the synaptic vesicle cycle probably provokes irreversible and cumulative ageing that can no longer be compensated for and which leads to the observed nerve terminal degeneration.

Concluding Remarks and Perspectives

The analysis of the CSP-dependent chaperone system in CSPα-deficient mice has shown that this system is essential to prevent use- and age-dependent degeneration of presynaptic terminals. These findings further put chaperones at center stage for neurodegenerative diseases of the human brain. The hallmark of many neurodegenerative diseases is malfolded proteins that subsequently generate harmful aggregates, for example, in Alzheimer's disease, Huntington's disease, spinocerebellar ataxia, and Parkinson's disease. Transgenic approaches demonstrated that the activation of the Hsp70-chaperone system is able to improve to some extent the neurological symptoms of neurodegenerative diseases in mouse model systems. Therefore, the knowledge and further analysis of chaperone systems in the brain are of crucial importance to understanding neurodegeneration in humans. Elucidating the cross-talk between the Hsp70-based chaperone system and the ubiquitin proteasome system (UPS) will be highly relevant to understanding the mechanisms of protein quality control in neurons. Hsp70-binding proteins such as the TPR-protein CHIP can probably shuttle (triage) proteins from the chaperone system to the proteasome degradation system if malfolded proteins cannot be recovered. Candidate proteins to be chaperoned by CSPα and Hsp70 are just beginning to be elucidated. Future studies aiming to complete the list of CSPα substrates will be very important in identifying sensitive target molecules that lead to synaptic failure in neurodegenerative disorders.

See also: Development of Drosophila Neuromuscular Junctions; Exocytosis: Ca^{2+}-Sensitivity; Neuromuscular Junction (NMJ): Mammalian Development; Neurosecretion (Regulated Exocytosis in Neuroendocrine Cells); Protein Folding and the Role of Chaperone Proteins in Neurodegenerative Disease; SNAREs; Ultrastructural Organization of Release Sites in the Calyx of Held.

Further Reading

Borst JG and Sakmann B (1996) Calcium influx and transmitter release in a fast CNS synapse. *Nature* 383: 431–434.

Braun JE, Wilbanks SM, and Scheller RH (1996) The cysteine string secretory vesicle protein activates Hsc70 ATPase. *Journal of Biological Chemistry* 271: 25989–25993.

Chamberlain L and Burgoyne RD (2000) Cysteine-string protein: The chaperone at the synapse. *Journal of Neurochemistry* 74: 1781–1789.

Chamberlain L, Henry J, and Burgoyne RD (1996) Cysteine-string proteins are associated with chromaffin granules. *Journal of Biological Chemistry* 271: 19514–19517.

Chandra S, Gallardo G, Fernandez-Chacon R, Schlüter OM, and Südhof TC (2005) Alpha-synuclein cooperates with CSPalpha in preventing neurodegeneration. *Cell* 123: 383–396.

Evans GJ, Morgan A, and Burgoyne RD (2003) Tying everything together: The multiple roles of cysteine string protein (CSP) in regulated exocytosis. *Traffic* 4: 653–659.

Fernandez-Chacon R, Wölfel M, Nishimune H, et al. (2004) The synaptic vesicle protein CSPα prevents presynaptic degeneration. *Neuron* 42: 237–251.

Fink AL (1999) Chaperone-mediated protein folding. *Physiological Reviews* 79: 425–449.

Gundersen CB and Umbach JA (1992) Suppression cloning of the cDNAs for a candidate subunit of a presynaptic calcium channel. *Neuron* 9: 527–537.

Hartl FU and Hayer-Hartl M (2002) Molecular chaperones in the cytosol: From nascent chain to folded protein. *Science* 295: 1852–1858.

Leveque C, Pupier B, Marqueze B, et al. (1998) Interaction of the cysteine string proteins with the alpha1A subunit of P/Q-type calcium channel. *Journal of Biological Chemistry* 273: 13488–13492.

Magga JM, Jarvis SE, Arnot MI, Zamponi GW, and Braun JE (2000) Cysteine string protein regulates G protein modulation of N-type calcium channels. *Neuron* 28: 195–204.

McDonough H and Patterson C (2003) CHIP: A link between the chaperone and proteasome systems. *Cell Stress & Chaperones* 4: 303–308.

Morales M, Ferrus A, and Martienz PM (1999) Presynaptic calcium-channel currents in normal and csp mutant drosophila peptidergic terminals. *European Journal of Neuroscience* 11: 1818–1825.

Ranjan R, Bronk P, and Zinsmaier KE (1998) Cysteine string protein is required for calcium secretion coupling of evoked neurotransmission in Drosophila but not for vesicle recycling. *Journal of Neuroscience* 18: 956–964.

Schmitz F, Tabares L, Khimich D, et al. (2006) CSPα-deficiency causes rapid and massive photoreceptor degeneration. *Proceedings of the National Academy of Sciences of the United States of America* 103: 2926–2931.

Südhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Tobaben S, Thakur P, Fernandez-Chacon R, Südhof TC, Rettig J, and Stahl B (2001) A trimeric protein complex functions as a synaptic chaperone machine. *Neuron* 21: 987–999.

Zinsmaier KE, Eberle KK, Buchner E, Walter N, and Benzer S (1994) Paralysis and early death in cysteine-string protein mutants of drosophila. *Science* 263: 977–980.

Zinsmaier KE, Hofbauer A, Heimbeck G, Pflugfelder GO, Buchner S, and Buchner E (1990) A cysteine-string protein is expressed in retina and brain of drosophila. *Journal of Neurogenetics* 7: 15–29.

Cysticercosis: Cerebral

M A Goldberg, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

The invasion of the central nervous system by the larva of the pork tapeworm, *Taenia solium*, results in a neurological disorder known as 'cerebral cysticercosis.' Until recently, this condition was primarily encountered in Third World countries, but improved diagnostic methods and changing migration patterns have resulted in recognition of this disease in developed countries. Cerebral cysticercosis is common in Mexico, Central and South America, Southeast Asia, South Africa, and in parts of Eastern Europe. In some undeveloped countries it is the most common cause of epilepsy. The parasite spreads through unsanitary porcine agricultural conditions, and in many countries its eradication follows the adoption of simple sanitary measures. The primary infection in humans is an intestinal infestation by the adult tapeworm resulting from ingestion of contaminated pork. This condition is relatively benign, producing only mild gastrointestinal symptoms. Mature segments from the adult worms contain eggs which pass out with human feces into the soil. Pigs, usually living in close association with humans, become infected by consuming food contaminated with these eggs. Human infestation with larva occurs through autoinfection by anal–oral transfer or fecal contamination of food. Ingestion of the eggs results in the development of the larval form which then invades tissues. Skeletal muscle and the central nervous system are the primary sites for this organism. In tissue, the larva develop into a cystic structure typically 1–2 cm in diameter or larger and with a characteristic scolex and three-layered membrane. They may survive in the CNS for up to 15 years.

Five clinical syndromes of cerebral cysticercosis can be defined: (1) intraventricular, (2) meningitic, (3) parenchymal, (4) spinal, and (5) mixed. The clinical manifestations depend largely on the anatomical site of encystment.

Cysts in the 3rd or 4th ventricle obstruct the flow of cerebral spinal fluid (CSF) and can result in obstructive hydrocephalus. The cyst may move within the ventricular cavity and produce a ball-valve effect resulting in intermittent symptoms. Sudden death may occur in these patients, occasionally without prior symptoms.

In the parenchymal form, symptoms are closely related to the site of encystment. Cortical cysts frequently give rise to partial seizures, and epilepsy due to this disease is quite common. Persistent motor or sensory loss or ataxia from cerebellar cysts occurs less frequently.

The meningitic form is common, and may be the result of the rupture or death of an arachnoidal cyst or its transformation into the racemose form of the organism which lacks a scolex and can enlarge within the basal cisterns. Symptoms of meningeal involvement vary from mild headache to a syndrome of chronic meningitis with communicating hydrocephalus. Spinal involvement may occur and evolve into complete spinal subarachnoid block.

The diagnosis should be considered in individuals who have resided in endemic areas and have epilepsy, meningitis, or increased intracranial pressure. Computerized tomography is useful in evaluating such patients but MRI is more sensitive. Hydrocephalus may be observed and the exact site of obstruction determined; parenchymal cysts are often calcified, and the presence of small calcifications which may be surrounded by a lucent area, either singly or scattered diffusely, is helpful in suggesting the diagnosis. Specific clinical diagnostic criteria have recently been proposed.

Viable cysts appear as small, lucent areas that do not enhance after the intravenous infusion of contrast material. CSF examination is usually normal at this stage. The longtime survival of these cysts suggests a lack of immunological or host responses, supporting the view that the brain is an immunologically privileged site. When the larva dies, proteins leak out of the cyst and an immunological response takes place. A more severe meningitis with many white blood cells, elevated protein, and low glucose in the CSF may occur. The histological appearance of the lesions at this time shows a chronic inflammatory response with lymphocytes and plasma cells, and the surrounding brain will enhance when contrast material is given prior to CT or MRI scan (Figures 1 and 2). In addition to the cellular response there may be a significant antibody response. Specific immunoglobulins to cysticercosis have been identified in serum and in CSF using enzyme-linked techniques and immunoblots and may be of diagnostic assistance. The apparent discrepancy

Figure 1 Axial T1-weighted, postcontrast MRI of the brain demonstrating a cysticercus (arrows) within the trigone of the right lateral ventricle.

Figure 2 Coronal T1-weighted, postcontrast MRI of the brain demonstrating a cysticercus (arrows) within the trigone of the right lateral ventricle.

between the immunological response seen in symptomatic patients and the long survival of cysts has led investigators to suggest that it is usually the death of the organism that produces most of the clinical features of this disorder, as well as the greatest inflammatory and immunological response.

Treatment of cerebral cysticercosis is usually symptomatic with anticonvulsants and anti-inflammatory agents. Two antihelminthics, praziquantel and albendazole, have been reported to be effective in the treatment of cysticercosis. The role of these agents in treatment and the long-term outcome remains controversial as they are associated with serious toxicity, and there is considerable evidence that many cases of CNS cysticercosis undergo spontaneous resolution and healing. Intraventricular cysts can be removed surgically, and newer endoscopic techniques are quite safe.

See also: Epilepsy; Epilepsy: Channelopathies; Parasitic Diseases: Nervous System Effects.

Further Reading

Bergsneider M (1999) Endoscopic removal of cysticercal cysts within the fourth ventricle. *Journal of Neurosurgery* 91: 340–345.

Carpio A, Escobar A, and Hauser WA (1998) Progress in epilepsy research: Cysticercosis and epilepsy: A critical review. *Epilepsia* 39: 1025–1040.

Carpio A, Santillan F, Leon P, et al. (1995) Is the course of neurocysticercosis modified by treatment with antihelminthics agents? *Archives of Internal Medicine* 155: 1982–1988.

Del Brutto OH, Rajshekhar V, White AC Jr., et al. (2001) Proposed diagnostic criteria for neurocysticercosis. *Neurology* 57: 177–183.

Tsang VCW, Brand JA, and Boyer AE (1989) An immunoelectrotransfer blot assay and glycoprotein antigens for diagnosing human cysticercosis. *Journal of Infectious Disease* 159: 50–59.

Cytokine Receptors in Glia

H W G M Boddeke, U Eisel, and K Biber, University of Groningen, Groningen, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Cytokines and chemokines are secreted, soluble proteins that are involved in cell survival, proliferation, differentiation, and a range of specific cellular activities. They act through transmembrane receptors and thus initiate a variety of complex intracellular signaling cascades. In the central nervous system, a number of cytokines are essentially involved in neurodevelopment and homeostasis and also play a role in neuroinflammatory processes and neurodegenerative pathology. Particularly glial cells, including astrocytes and microglia, are the primary source of cytokines in the nervous system. Astrocytes, microglia, and oligodendroctyes express a specific repertoire of cytokine receptors, allowing them to react properly on cytokines released from surrounding glia and infiltrating immune cells. In the following sections the various groups of cytokine receptors, their signaling, and their possible functions in glia are addressed.

The Interferon/Interleukin-10 Receptor Family

Interferon (IFN) and interleukin-10 (IL-10) receptors are related structurally. IFN receptors have been categorized as type I and type II receptors. Type I receptors bind IFN-α and IFN-β, whereas type II receptors bind interferon-γ. Both type I and type II receptors are heterodimers consisting of the transmembrane protein chains IFN-αRI and IFN-αRII, and IFN-γRI and IFN-γRII, respectively.

Signaling of type I receptors is initiated by phosphorylation of the receptor-associated proteins tyrosine kinase-2 (TYK-2) and Janus kinase (JAK), which subsequently phosphorylate IFN-αRI. IFN-αRI then binds to signal transducer and activator of transcription 2 (STAT2), which is then tyrosine phosphorylated and associates with STAT1α. The STAT1α–STAT complex dissociates from the receptor and binds the interferon-regulating factor IRF-9. The resulting trimer translocates to the nucleus and binds to IFN-stimulated regulatory element responsive elements (ISREs). ISREs are found in the promoters of all IFN-regulated genes and activate a specific pattern of gene expression.

In the brain, IFN-α has direct effects on neurosecretory hypothalamic neurons and regulates the hypothalamic–pituitary–adrenocortical axis. Generally, IFN-β induces anti-inflammatory activity and thus is used for the treatment of multiple sclerosis. Type I IFN receptors have been localized in astrocytes, microglia, and oligodendrocytes, which respond to IFN-β by suppression of the production and release of proinflammatory cytokines, induction of anti-inflammatory cytokines such as IL-10, downregulation of class II major histocompatibility complex (MHCII) expression, and suppression of matrix metalloprotease activity.

The IFN-γ receptor subunits RI and RII are constantly associated with JAK-1 and JAK-2, respectively. After dimerization and activation of JAK-1 and JAK-2, IFN-γRI is phosphorylated, which leads to phosphorylation of STAT1α. Subsequently, homodimerized STAT1 translocates to the nucleus and binds gamma activation sites (GASs) in the promoter of IFN-γ-inducible genes, thus regulating their transcription. In general, IFN-γ is produced by T helper (Th1) cells and natural killer cells and produces a proinflammatory (Th1-like) response. Type II IFN-γ receptors are expressed by astrocytes, microglia, and oligodendrocytes. Thus, IFN-γ is a potent activator of microglia and astrocytes, resulting in expression of MHCII, CD40, and B7. Furthermore, INF-γ induces upregulation of a variety of proinflammatory cytokines. In oligodendrocytes, IFN-γ induces apoptosis.

The IL-10 receptor has structural similarity to the IFN-γ receptor and is composed of an IL-10R1 and -R2 subunit. After binding IL-10, the IL-10R1 subunit associates with IL-10R2. The receptor complex then activates the JAK–STAT pathway, including phosphorylation of the receptor-associated proteins TYK and JAK-1, ultimately leading to activation of STAT3, STAT1, and STAT5, which regulate the expression of IL-10-responsive genes, including c-*myc*, *bcl-2*, and *bcl-xL*.

IL-10 is generally considered an anti-inflammatory cytokine that inhibits the production and secretion of a variety of inflammatory cytokines. Furthermore, IL-10 suppresses the expression of MHCII, CD40, and B7 molecules. Astrocytes and microglia produce IL-10 and express high-affinity IL-10 receptors under pathological conditions. Activation of IL-10 receptors in astrocytes leads to decreased expression of intercellular cellular adhesion molecule-1 (ICAM-1) and enhances the release of nerve growth factor (NGF). IL-10 suppresses cytokine and chemokine expression in both astrocytes and microglia. Furthermore, IL-10 has a protective effect on microglia and oligodendrocytes.

Interleukin-6-Type Receptors

IL-6 belongs to the family of cytokines (IL-6-type cytokines) which also includes ciliary neurotrophic factor (CNTF), leukemia inhibitory factor (LIF), oncostatin M (OSM), IL-11, cardiotrophin-1 (CT-1), and the recently discovered IL-27 and cardiotrophin-like cytokine (CLC); CLC is also known as novel neurotrophin 1 (NNT1) or B cell-stimulating factor-3 (BSF-3). IL-6-type cytokines bind to specific receptors that share as a common feature the use of glycoprotein 130 (gp130) in their signaling cascade. Various variants of signaling complexes have been described for different cytokines from this family. IL-6 and IL-11 bind to α receptor subunits (IL-6R or IL-11R) that recruit two gp130 subunits (gp130/gp130 homodimer) for signaling. NNT1/BSF-3/CLC/LIF and CNTF also bind to their corresponding α receptor subunits, which recruit a dimer of a LIF receptor (LIFR) and a gp130 (gp130/LIFR heterodimer). The same signaling complex can also be activated by OSM that signals through either gp130/LIFR or gp130/OSMR heterodimers.

Two forms of IL-6R, a membrane bound and a soluble extracellular form, are known. The soluble IL-6R either results from proteolysis (shedding) of membrane-bound IL-6R or from alternative splicing. Since the membrane-bound IL-6R does not contain any signaling domains in its short cytoplasmic part, both membrane-bound IL-6R and soluble IL-6R have very similar effects. Thus, soluble IL-6-like receptors bound to an agonist ligand act as agonist in any gp130-expressing cell, independent of the presence of IL-6R, a process that is called trans-signaling. Indeed, this has been described in astrocytes and neurons. Cytokine-activated gp130 complexes do not contain intrinsic kinase domains, but are known to associate with JAKs. Upon cytokine binding and dimer formation, JAKs phosphorylate several tyrosine residues in the cytoplasmic domain of gp130. These tyrosine residues are essentially involved in the activation of various intracellular signaling cascades, such as the mitogen-activated protein kinase (MAPK) pathway, or in the activation of STAT.

The co-receptor gp130 is expressed ubiquitously and is thus also found in the major glia types. There is, however, limited information concerning the receptor expression for IL-6-type cytokines in glia cells. Evidence for the expression of IL-6-type cytokine receptors in glia cells is mostly based on functional studies showing effects of these cytokines in glia cells. Among the members of the IL-6 family, IL-6 has been intensively studied in glia cells. There is convincing evidence that glia cells are main producers of IL-6 in brain pathology; however, contradictory findings have been published concerning the expression of IL-6 receptor in astrocytes. Trans-signaling most likely enables astrocytes to respond to IL-6, although several reports suggest astrocytic expression of IL-6R. In unchallenged astrocytes, expression levels of CNTF receptors are rather low. CNTF and its receptor are clearly upregulated in the central nervous system (CNS) following lesions, when CNTF is suggested to be involved in the regulation of astrogliosis. Little is yet known about effects of CNTF in oligodendrocytes and microglia. LIF is an important regulator of astrocyte differentiation and is crucial for the expression of glial fibrillary acidic protein (GFAP). Accordingly, LIF-deficient mice display less astrogliosis in response to neuronal injury. Therefore, a regulation of astrocytic activity (astrogliosis) after CNS lesion is also suggested for LIF. Recent results, moreover, indicate a similar function for LIF in the regulation of microglia activity. Exciting findings have been published concerning the effects of LIF in oligodendrocytes: LIF has been found to enhance the survival of oligodendrocytes and stimulated myelination of axons. Accordingly, LIF is now discussed as a potential therapeutic agent in multiple sclerosis. OSM is currently being discussed as a means to control the growth of glioblastoma cells, but little is yet known about OSM-dependent effects in other glia cell types.

The Interleukin-2 Receptor Family/γC

Of the IL-2 receptor family, the cytokine receptors for IL-2, IL-4, and IL-15 have been found to be expressed in glia. These cytokine receptors contain a ligand-binding subunit (also designated as the α chain) and a commonly shared γC chain. Particularly, IL-2 and IL-15 cytokine receptors contain a third chain. IL-2 and IL-15 cytokine receptors mediate signal transduction by activation of STAT5. IL-4 cytokine receptors use STAT6 for signal transduction.

The IL-2 receptor consists of two to three receptor subunits: the α chain, the β chain, and the common γ chain. The high-affinity receptor is composed of all three subunits; the intermediate affinity receptor only contains the IL-2β and IL-2γ subunit. The IL-2α subunit is not required for signal transduction but confers high affinity to the receptor complex. Furthermore, it has been suggested that the oligomeric receptor complex consisting of the receptor subunits IL-2β, IL-2γ, and IL-15α serves as a receptor for the cytokine IL-15 in microglia.

After IL-2-dependent oligomerization of the IL-2 subunits, leading to either high- or intermediate-affinity receptors, JAK-3, which is associated with the IL-2γ subunit, activates the IL-2β-associated

JAK-1. Both JAKs phosphorylate tyrosine residues in β chain-associated lymphocyte-specific protein tyrosine kinase ($p56^{lck}$), as well as tyrosine residues in both the β and γ chains. This leads to activation of secondary signaling molecules, including STAT5 and STAT3 and SH2 domain-containing protein tyrosine phosphatase (SHP-2), ultimately leading to induction of cell-specific genes and cell proliferation.

IL-2 is a potent T cell growth factor and has a variety of neuromodulatory activities. Thus, IL-2 may be involved in memory function, regulation of sleep, and modulation of the neuroendocrine axis. In the brain, IL-2 synthesis has been described for astrocytes, neurons, and possibly micoglia. Expression of IL-2R has been found in microglia, astrocytes, and oligodendrocytes. In microglia, IL-2 activity has been associated with the production of nitric oxide. Furthermore, IL-15 signaling in microglia has been associated with activation of JAK-1, cell proliferation, and inhibition of nitric oxide production. In astroctyes, IL-2 increases astrogliosis. IL-2 has been shown to promote oligodendrocyte proliferation, and maturation inhibition of IL-1-induced oligodendrocyte progenitor proliferation has been reported as well.

Phosphorylation of JAK-1 and JAK-3, which are stably associated with the α chain and with the γC chain, respectively, occurs after binding of IL-4 to its receptor. Subsequently, phosphorylation of STAT6 takes place; STAT6 then binds to N4 STAT binding elements and interacts with a variety of transcription factors, including nuclear factor-kappa B (NF-κB), CCAAT/enhancer binding protein (C/EPBα), and cAMP response element-binding protein (CBP). IL-4 receptors are found in astrocytes and microglia and oligodendrocytes. IL-4 is an anti-inflammatory cytokine that also effectively suppresses production of tumor necrosis factor-α (TNF-α), IL-1β, and nitric oxide (NO) in glial cells, as well as the expression of MHCII and CD40.

Tumor Necrosis Factor Receptor Family

TNF receptors are activated by cytokines and growth factor molecules of the TNF ligand family. Based on the presence or absence of a death domain, members of the TNF receptor family have been divided into two groups. Death domain-containing receptors (e.g., Fas) are solely potent inducers of cell death. Despite the presence of a death domain, TNFR1 activation mostly activates a multitude of signaling responses, whereas apoptosis is induced only under certain conditions. Receptors lacking the death domain, such as TNFR2 and CD40, play a role in gene transcription, differentiation, or cell survival.

The most abundant and best-described TNF ligand in the brain is TNF-α, which acts upon TNFR1 and TNFR2, with TNFR1 mediating most of the TNF-α effects. After TNF-α-induced trimerization of TNFR1 subunits, a TNFR1-associated death domain (TRADD) protein is recruited. The TNFR1–TRADD complex can bind TNFR-associated factor 2 (TRAF2) and receptor-interacting protein (RIP) in order to activate c-Jun N-terminal kinase (JNK) and NF-κB. Additionally, the TNFR1–TRADD complex can recruit Fas-associated death domain protein (FADD), which ultimately initiates caspase activation and apoptosis. TNFR2 was for a long time considered to be a modulator of TNFR1 or even only a 'ligand-passing molecule' because of its low affinity for soluble TNF-α. However, after it was shown that the TNF-α membrane precursor acts as the prime activator for TNFR2, independent signaling pathways were described for also TNFR2. TNFR2 is known to be capable of binding TRAF2 and to induce NF-κB. Co-signaling of both receptors in some cell types leads to a modulation of TNFR1 signals.

TNF-α is primarily produced not only by microglia, but also, to a certain extent, by astrocytes and neurons. TNF-α is a pleiotropic and proinflammatory cytokine and induces a variety of cellular responses, including proliferation, differentiation, gene transcription, and apoptosis. TNF-α is a major activator of microglia and astrocytes and both TNFR1 and TNFR2 have been found in astrocytes, microglia, and oligodendrocytes. Both receptors become strongly upregulated in glia upon cellular stress. Although TNF-α is directly toxic to oligodendrocytes, and transgenic mice overexpressing TNF-α clearly exhibit marked white-matter lesions, other studies show that TNFR2 signaling is important for remyelination. In ischemia, TNFR2 antagonizes TNFR1 effects and induces cell survival signals mediated via protein kinase B (PKB)/Akt phosphorylation and NF-κB activation. Upon ischemia, TNF-α and TNFRs become upregulated for hours in both neurons and glial cells. Mice without TNFR1 (TNFR1 knockout mice) show a strong resistance to ischemia and have high levels of phoshorylated PKB/Akt, whereas TNFR2 knockout mice are highly sensitive. Interestingly, NF-κB activation by TNFR1 does not induce neuroprotection and is not activated by the same signaling pathway as seen in TNFR2-mediated NF-κB activation. TNFR2-mediated and NF-κB-dependent protection may be a result of the long-lasting activation kinetics. TNF-α-induced activation of the transcription factor NF-κB is crucial for production of various cytokines and chemokines, adhesion molecules, and acute-phase proteins. NF-κB also activates regulators of apoptosis and cell proliferation.

Interleukin-1 Receptors

IL-1 receptors belong to the Toll/IL-1 receptor superfamily, as defined by the presence of a so-called Toll/IL-1 receptor (TIR) domain. Besides IL-R1, this family includes IL-1 receptor accessory protein (IL-1AcP; which is essential for IL-1 signaling), a number of Toll-like receptors, and the putative IL-18/IL-1-γ receptor IL-1Rrp (IL-1 receptor-related protein). After binding IL-1α/β, IL-R1 heterodimerizes with IL-1AcP. The TIR domains of the IL-R1–IL-1AcP complex interact with a number of adapter molecules, including the IL-1 receptor-associated kinases IRAK1 and IRAK2. Activation of NF-κB occurs by association of the death domain-containing cytosolic protein MyD88 to both IL-1RAcP and IRAK2, which leads to binding of tumor necrosis-associated factor 6 (TRAF6) to the signaling complex and subsequent activation of NF-κB-inducing kinase (NIK).

The regulators of IL-1 signaling, IL-1 receptor antagonist (IL-1Ra) and type 2 IL-1R (IL-1R2), are present in the brain. IL-1Ra is an IL-R1 antagonist; IL-1R2 is a membrane-bound IL-1 binding protein, incapable of signal transduction, which is also situated in glial cells. Astrocytes, microglia, and oligodendrocytes express IL-1R1, which can be activated by IL-1α and IL-1β. Glia-derived IL-1 is a potent inducer of proinflammatory effects in chronic degenerative diseases and traumatic or toxic lesions. Resting microglia constitutively express IL-1Ra, which may contribute to the 'immune-privileged' state of the central nervous system. The relative levels of IL-1 and IL-1Ra most likely contribute to the overall state of reactivity toward immunological challenges. IL-1β induces astrogliosis and the expression of cell adhesion molecules (ICAM-1 and VCAM-1), as well as astrocytic expression of various cytokines and chemokines, including TNF, IL-6, CCL2, CXCL8, and CXCL10.

IL-1 expression is induced rapidly in response to CNS injury and is a key mediator of experimentally induced neuronal damage. However, IL-1-mediated neuronal death requires co-culturing of neurons and glia. IL-1-induced neuronal cell death can be prevented by pretreatment with IL-1 receptor antagonist and caspase inhibitors. It is assumed that IL-1 activates IL-1R1 receptor in astrocytes to induce caspase-dependent neuronal death, which is dependent on the release of free radicals and may contribute to neuronal cell death in CNS diseases.

Transforming Growth Factor-β Receptors

Transforming growth factor-β (TGF-β) belongs to the TGF-β superfamily, a large family of structurally related cytokines and growth factors. In mammals, three TGF-β isoforms (TGF-β1, TGF-β2, and TGF-β3) are known. The effects of all TGF-β isoforms are mediated by a transmembrane serine/threonine kinase receptor complex consisting of TGF-β type I and TGF-β type II receptors. Since these receptors function as heterodimers, both receptor types must be expressed in order to allow TGF-β-dependent effects. Upon TGF-β binding and heterodimer formation, type II receptor phosphorylates the type I receptor kinase domain, which subsequently activates intracellular Smad proteins (vertebrate homolog of the *Drosophila* gene 'mothers against decapentaplegic' (MAD)). Two different receptor-regulated Smads (R-Smads), Smad2 and Smad3, are known to interact with type I TGF-β receptors. Upon activation, these R-Smads associate with the common-partner Smad (Co-Smad), Smad4, to form heterooligomers, which enter the nucleus and induce gene transcription.

TGF-βs regulate a variety of cellular functions, including cell survival, apoptosis, proliferation, and cellular differentiation. It is known today that TGF-βs act in a highly contextual manner, meaning that the outcome of TGF-β signaling greatly depends on the cell type and environment. A large body of evidence shows the presence of all three TGF-β isoforms in the CNS. Whereas TGF-β2 and TGF-β3 are found under physiological conditions, TGF-β1 expression is rapidly induced in damaged nervous tissue. The cellular source of TGF-β isotypes is widespread; thus, all major cell types in brain, including astrocytes, microglia, and oligodendrocytes, show expression of TGF-β isotypes.

The expression of TGF-β receptors in brain cells is less clear. Few reports show TGF-β receptor (mRNA or protein) expression in brain or cultured neurons and astrocytes. However, despite a less clear characterization of the TGF-β receptor expression pattern in glia cells, it has been convincingly shown that all major glia cell types (astrocytes, oligodendrocytes, Schwann cells, and microglia) functionally respond to TGF-β stimulation, strongly indicating that they functionally express TGF-β receptors. In astrocytes, TGF-β1 treatment downregulates the expression and release of various proinflammatory cytokines (IL-1β or TNF-α) and induces the expression of various extracellular matrix molecules (ICAM-1 or VCAM-1). Similarly, an anti-inflammatory role of TGF-β1 in microglia has been described; TGF-β treatment disrupts endotoxin signaling, controls MHCII expression, and induces and changes the expression pattern of microglia inflammatory genes toward an anti-inflammatory phenotype. The importance of TGF-β in downregulating microglia activity has been corroborated in TGF-β1-deficient mice, which

display increased microgliosis and neuronal death in response to excitotoxic injury.

Chemokine Receptors

Chemokines are small-sized proteins that are able to induce migration of cells; currently, approximately 50 chemokines and 20 receptors are known. Based on the position of four conserved cysteine residues in the N-terminal region of the protein, the chemokine family has been divided into four groups. The two largest groups are CXC (or α) and CC (or β), in which the first two cysteines are either separated by one amino acid residue or are adjacent to each other respectively. The two small groups are the C chemokines (or γ), with only one cysteine in the N-terminal region, and the CX3C chemokines (or δ), in which the first two cysteines are separated by three amino acid residues. Chemokine receptors are named after the types of chemokines they preferentially bind. Thus, CC receptors bind CC chemokines, and so on. Functionally, only one exception has been reported in the literature, the CC chemokine CCL21, which, in addition to its primary receptor CCR7, also activates the CXC receptor CXCR3. Chemokine receptors all belong to the family of G-coupled protein coupled receptors (GPCRs). A majority of chemokine-induced responses are inhibited by pertussis toxin (PTX), indicating that G_i proteins mediate these effects. Chemokine receptors activate several intracellular targets, such as adenylcyclase, phospholipases, and molecules like Rho, Rac, and Cdc42. They are also able to activate signaling cascades of major kinases, such as MAPK and phosphatidylinositol 3-kinase (PI3K) pathways. The actions of chemokines are not restricted to chemoattraction but are very diverse. Chemokines are important in early development, formation of secondary lymphoid organs, wound healing, angiogenesis and angiostasis, regulation of expression of adhesion molecules, and development of Th1/Th2 profiles, and they have also been implicated in certain aspects of tumor growth and metastasis.

Glial cells (astrocytes and microglia) are the primary cellular source of chemokines in the CNS under various conditions and these cells also express a large variety of chemokine receptors. In Table 1 a summary of the chemokine receptors expressed in glia is given. An induction of chemokine receptor expression in glia cells has been described in various brain diseases such as multiple sclerosis, HIV infection, infarcts/stroke, and neoplastic disease.

Various chemokines have been shown to induce astrocyte proliferation and/or migration *in vitro*. Similarly, microglia express a large number of chemokine receptors and respond to chemokine stimulation with chemotaxis. Since endangered neurons express and release chemokines that recruit microglia and CXCL10, a role of chemokines in neuron–microglia signaling has been proposed. This assumption is mainly based on *in vitro* findings, though recent studies using chemokine receptor-deficient mice indeed show significant changes in microglia behavior in response to neuronal injury.

Table 1 Summary of the chemokine receptors expressed in glia and their chemokine ligands expressed in brain

Chemokine receptor	*Ligand*	*Glial cell type*
CCR1	CCL3	Microglia
	CCL5	Astrocytes
	CCL6	Oligodendrocytes
	CCL7	
	CCL8	
CCR2	CCL2	Microglia
	CCL7	Astrocytes
	CCL8	
CCR3	CCL5	Microglia
	CCL7	Astrocytes
	CCL8	Oligodendrocytes
	CCL11	
CCR4	CCL22	Astrocytes
CCR5	CCL3	Microglia
	CCL4	Astrocytes
	CCL5	
	CCL8	
CCR6	CCL20	Microglia
		Astrocytes
CCR7	CCL19	Microglia
	CCL21	
CCR8	CCL1	Astrocytes
CCR10		Astrocytes
CXCR1	CXCL1	Microglia
	CXCL6	Astrocytes
	CXCL8	Oligodendrocytes
CXCR2	CXCL1	Microglia
	CXCL2	Astrocytes
	CXCL6	Oligodendrocytes
	CXCL8	
CXCR3	CXCL9	Microglia
	CXCL10	Astrocytes
	CXCL11	Oligodendrocytes
	CCL21	
CXCR4	CXCL12	Microglia
		Astrocytes
		Oligodendrocytes
CXCR5	CXCL13	Astrocytes
CXCR6	CXCL16	Microglia
		Astrocytes
CX3CR1	CX3CL1	Microglia
		Astrocytes

See also: Glial Growth Factors; Growth Factors: Neuronal Atrophy; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neurotransmitters and Growth Factors: Overview; Retrograde Neurotrophic

Signaling; Sleep and Sleep States: Cytokines and Neuromodulation; Stress, Cytokines and Depressive Illness.

Further Reading

Ambrosini E and Aloisi F (2004) Chemokines and glial cells: A complex network in the central nervous system. *Neurochemical Research* 29: 1017–1038.

Balkwill F (1998) The molecular and cellular biology of the chemokines. *Journal of Viral Hepatitis* 5: 1–14.

Bottner M, Krieglstein K, and Unsicker K (2000) The transforming growth factor-βs: Structure, signaling, and roles in nervous system development and functions. *Journal of Neurochemistry* 75: 2227–2240.

Fontaine V, Mohand-Said S, Hanoteau N, et al. (2002) Neurodegenerative and neuroprotective effects of tumor necrosis factor in retinal ischemia: Opposite roles of TNF receptor 1 and TNF receptor 2. *Journal of Neuroscience* 22: RC216.

Hanisch UK and Quirion R (1996) Interleukin-2 as neuroregulatory cytokine. *Brain Research Reviews* 21: 246–284.

John GR, Lee SC, Song X, et al. (2005) IL-1-regulated responses in astrocytes: Relevance to injury and recovery. *Glia* 2: 161–176.

Ledeboer A, Breve JJ, Poole S, et al. (2000) Interleukin-10, interleukin-4 and transforming growth factor-beta differentially regulate lipopolysaccharide-induced production of pro-inflammatory cytokines and nitric oxide in co-cultures of rat astroglial and microglial cells. *Glia* 30: 134–142.

Murakami M, Kamimura D, and Hirano T (2004) New IL-6 (gp130) family cytokine members, CLC/NNT1/BSF3 and IL-27. *Growth Factors* 2: 75–77.

Nguyen VT and Benveniste EN (2000) IL-4-activated STAT-6 inhibits IFN-gamma-induced CD40 gene expression in macrophages/microglia. *Journal of Immunology* 165: 6235–6243.

Paglinawan R, Malipiero U, Schlapbach R, et al. (2003) TGFβ directs gene expression of activated microglia to an anti-inflammatory phenotype strongly focusing on chemokine genes and cell migratory genes. *Glia* 44: 219–231.

Pinteaux E, Parker LC, Rothwell NJ, et al. (2002) Expression of interleukin-1 receptors and their role in interleukin-1 actions in murine microglial cells. *Journal of Neurochemistry* 83: 754–763.

Sawada M, Suzumura A, Hosoya H, et al. (1999) Interleukin-10 inhibits both production of cytokines and expression of cytokine receptors in microglia. *Journal of Neurochemistry* 72: 466–477.

Strle K, Zhou JH, Shen WH, et al. (2001) Interleukin-10 in the brain. *Critical Reviews in Immunology* 21: 247–449.

Unsicker K and Strelau J (2000) Functions of transforming growth factor-β isoforms in the nervous system. Cues based on localization and experimental *in vitro* and *in vivo* evidence. *European Journal of Biochemistry* 267: 6972–6975.

Wajant H, Grell M, and Scheurich P (1999) TNF receptor associated factors in cytokine signaling. *Cytokine & Growth Factor Reviews* 1: 15–26.

Wesemann DR and Benveniste EN (1995) Cytokine and chemokine receptors and signalling. In: Kettenmann H and Ransom BR (eds.) *Neuroglia*, 2nd edn., pp. 142–149. New York: Oxford University Press.

Cytoskeletal Interactions in the Neuron

A Adebola and R K H Liem, Columbia University College of Physicians and Surgeons, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

The cytoskeleton is a dynamic arrangement of interconnected actin filaments, microtubules (MTs), and intermediate filaments (IFs) that extends throughout the cytoplasm. This structure plays a vital role in the development and maintenance of cell and tissue structure. Due to the complex nature of the cytoskeleton, early studies focused on identifying mechanisms that governed the organization and function of individual filament networks. Although these studies ascribed distinct functions to each, it has become increasingly clear that many dynamic cellular processes, such as cell migration, cell polarization, cell division, growth cone guidance, and intracellular transport, require the cooperative activity or cross-talk of multiple filament networks. In recent years, there has been tremendous focus on identifying and characterizing the structural and regulatory components required for facilitating such cytoskeletal cross-talk. Studies have revealed that proteins once thought to associate specifically with actin, MTs, or IFs are indeed capable of associating with more than one cytoskeletal network and thus may be involved in mediating cytoskeletal cross-talk. One such class of proteins is the plakins. Plakins were originally identified as IF-binding proteins that functioned to cross-link IFs to adhesion complexes of epithelia. Advancements in molecular cloning, whole-genome sequencing, and gene-targeting technologies have revealed that plakins are not limited to their capacity to associate with IFs. Instead, newly identified plakin genes and alternatively spliced isoforms have been shown to interact with other cytoskeletal elements, namely, actin and MTs. Interestingly, many of the newly identified proteins are highly expressed in the nervous system. This article provides an overview of plakin cross-linking properties from both nonneuronal and neuronal model systems.

Plakins

Plakins are giant-sized (200–800 kDa) modular proteins that link cytoskeletal elements to each other and to specialized sites at the plasma membrane. Maintaining such linkages lends mechanical strength to cells and tissue structures and ensures proper cellular development. The importance of plakins as cytolinkers is evident by the fact that mutations in several plakins are linked to a diverse array of diseases in humans and mice, in which cell and tissue integrity is compromised.

The plakin family is composed of the mammalian plakins desmoplakin, plectin, bullous pemphigoid antigen-1 (BPAG1), microtubule cross-linking factor 1 (MACF1), envoplakin, periplakin, and epiplakin and the invertebrate plakins *shortstop* (shot) and *vab-10* (Figure 1). Analysis of the genomic organization of plakins reveals that various protein modules are encoded by the gene loci of plakins. These modular interaction units, which allow for interaction with different cellular elements, are alternatively combined to produce distinct proteins with unique functions that are expressed in a tissue-specific manner (Figure 1). The modules include a calponin-type actin binding domain (ABD), a plakin domain, a coiled-coil rod domain, a plakin (plectin) repeat domain (PRD), a PRD associated linker (L) subdomain, a spectrin repeat-containing rod domain, two EF-hand motifs, a Gas2-related homology domain (GAR), and a domain containing a series of glycine–serine–arginine (GSR) repeats. The function of these modules is discussed in further detail later.

Because plakins were first identified as components of adhesion complexes in epithelia, most of our knowledge concerning plakin function stems from epithelial cell biology. Autoimmune and inherited disease states, such as plakin-associated skin-blistering disorders, provided insight into the function of these molecules. As additional cDNAs of plakins were cloned, they were found to be more complex than epithelial plakins, and their expression profiles were not limited to epithelia. Many of the new plakins were highly expressed in the nervous system. Their functions as linkers in the nervous system are just beginning to become clear. We therefore first describe the known functions of plakins in epithelial cells and proceed to their roles in the nervous system.

Connections between the Cytoskeleton and Cell Junctional Complexes

Cell-anchoring junctions are cytoskeleton-associated sites of cell–cell and cell–extracellular matrix (ECM) adhesions. In epithelial cells, three types of cell–cell junctions are identified: tight junctions, adherens junctions, and desmosomes (Figure 2). Cell–ECM adhesions include focal adhesions and hemidesmosomes (Figure 2). The formation and stabilization of these adhesion complexes are essential for molecular events that govern how tissues acquire and maintain proper architecture. In general, anchoring junctions are composed of transmembrane adhesion molecules of the cadherin (cell–cell adhesions) or integrin (cell–ECM adhesions) families and an electron-dense submembrane

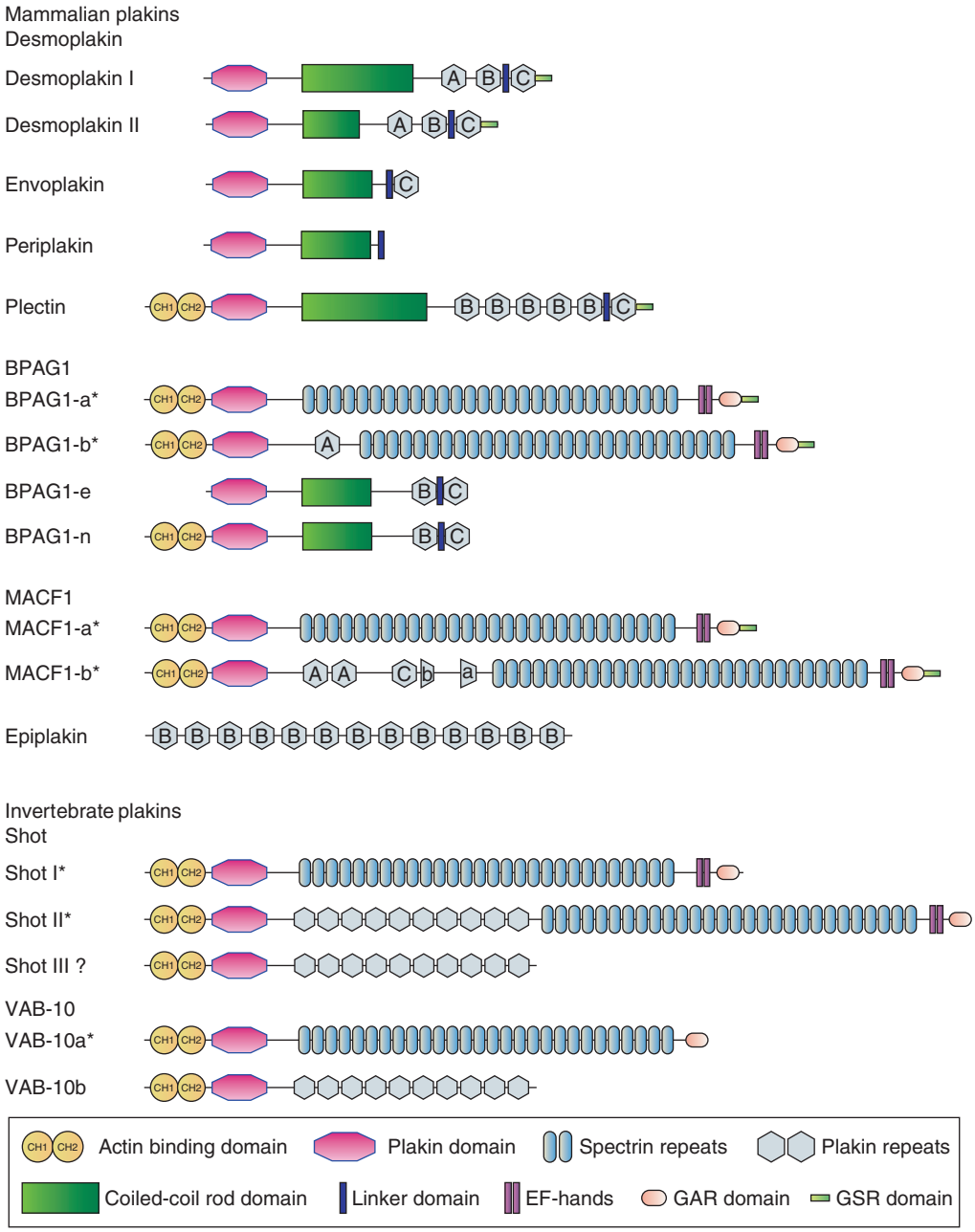

Figure 1 Schematic diagram of mammalian and invertebrate plakins. The plakin family is composed of several interaction modules that include calponin-type acting binding domain (ABD), plakin domain, coiled-coil rod domain, spectrin repeat-containing rod domain, plakin repeat domain (PRD), Gas2-related (GAR) domain, and glycine–serine–arginine (GSR) repeat-containing domain. The calponin-type ABD is composed of two calponin homology domains, CH1 and CH2, arranged in tandem. Although not shown here, there is extensive alternative splicing of the ABD. Such splicing alters the actin binding capacity of these proteins. The plakin domain is common to all but one of the plakin proteins. The coiled-coil rod domain is present in several of the epithelial plakins and is thought to mediate homo- and hetero-dimerization of plakins. The spectrin repeat-containing rod domain is present in MACF1, BPAG1, Shot, and VAB-10 isoforms. These proteins (asterisks) are sometimes referred to as spectraplakins because they share similarities with members of both the spectrin and the plakin families. The PRD is composed of plakin repeat motifs. It is important for binding to intermediate filaments. For some plakins, the presence of a linker (L) subdomain aids in interactions with intermediate filaments. PRDs based on sequence similarity are divided into subgroups called A, B, and C. For MACF1-b, the PRDs labeled 'a' and 'b' are equivalent to half of the A and B PRDs. The number of spectrin repeats and PRs drawn here are descriptive and may not reflect the actual number that exists in reality. EF-hand motifs appear in several plakins; however, it is not known if this domain can bind Ca^{2+}. The microtubule binding domain of BPAG1 and MACF1 isoforms is composed of a GAR domain, which binds microtubules, and a GSR domain that bundles microtubules. The GSR domain of plectin is also able to bind microtubules; however, its presence in desmoplakin functions to regulate desmoplakin's intermediate filament binding activity. In *Drosophila* Shot and *C. elegans* VAB-10 the GAR domain serves as the microtubule binding region. The existence of Shot III is still controversial. This figure is not drawn to scale. Adapted from Jefferson JJ, Leung C, and Liem RKH (2004) Plakins: Goliaths that link cell junctions and the cytoskeleton. *Nature Reviews Molecular Cell Biology* 5: 542–553, with permission.

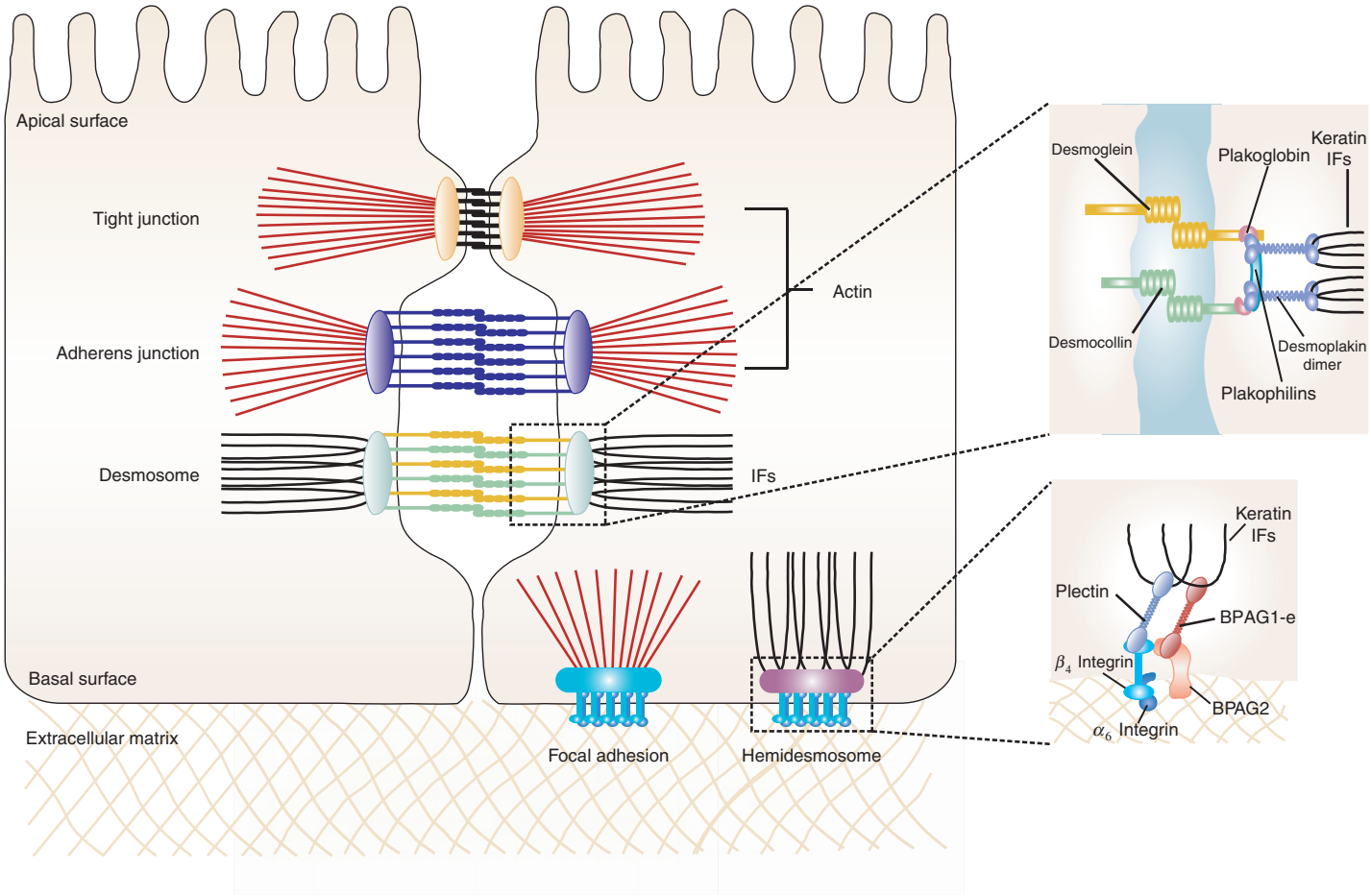

Figure 2 Mammalian cell adhesion complexes. The schematic illustrates adhesion complexes present in mammalian epithelia. Epithelial cells are polarized cells with an apical surface that is exposed to the environment and a basal surface that contacts the underlying extracellular matrix (ECM). Cell junctions, sites of cell–cell and cell–ECM contact, are generally composed of transmembrane adhesion proteins and an electron-dense submembrane plaque made up of signaling proteins that contribute to the transduction of intracellular signals and adaptor proteins that tether the junctional complex to the cytoskeleton. There are five different types of cell junctions illustrated in this figure. First, tight junctions or zona occludens (ZO) are selective permeability barriers that seal adjacent cells together. At these junctions, claudins and occludins are adhesive transmembrane proteins that are linked to the actin cytoskeleton through ZO proteins. Second, adherens junctions are cell–cell junctions composed of classical cadherins linked to the actin cytoskeleton through adaptor proteins such as vinculin and catenins. Third, desmosomes are cell–cell junctions that are formed by homotypic interactions between desmosomal cadherins, termed desmocollins and desmogleins, and are linked to intermediate filaments (IFs) through desmoplakin dimers. The armadillo proteins plakoglobin and plakophilin mediate interactions between desmoplakin and the desmosomal cadherins. Desmoplakin can also bind directly to the juxtamembrane region of desmocollin. Fourth, focal adhesions are cell–ECM junctions that are formed by cell adhesion proteins of the integrin family. These junctions are linked to actin filaments by adaptor proteins such as talin, filamin, and vinculin. Fifth, hemidesmosomes are cell–matrix junctions composed of the transmembrane proteins $\alpha_6\beta_4$ integrin and BPAG2 (bullous pemphigoid antigen-2). These transmembrane proteins, which interact with components of the ECM, are linked to cytoplasmic IFs through interactions with the plakin proteins, plectin and BPAG1-e (the epithelial form of bullous pemphigoid antigen-1). Adapted from Jefferson JJ, Leung C, and Liem RKH (2004) Plakins: Goliaths that link cell junctions and the cytoskeleton. *Nature Reviews Molecular Cell Biology* 5: 542–553, with permission.

plaque made up of signaling proteins that contribute to the transduction of intracellular signals and adaptor proteins that tether the junctional complex to the cytoskeleton. Several members of the plakin family serve as essential adaptor proteins for desmosomes and hemidesmosomes (**Figure 2**). Desmosomes are specialized intercellular junctions present in epithelia and cardiac muscle, where IFs are linked to desmosomal cadherins through interactions with desmoplakin. At hemidesmosomes, specialized cell–ECM junctions present in epithelia, BPAG1-e (the epidermal isoform of BPAG1), and plectin provide links between IFs and the transmembrane hemidesmosomal proteins $\alpha_6\beta_4$ integrin and BPAG2, a type XVII collagen. In the absence of these plakins, links between adhesion complexes and IFs are compromised and tissues under mechanical stress are subject to shearing.

The ability of plakins to connect cell adhesion targets at the plasma membrane with cytoskeletal elements is mediated through interactions with the various protein modules that make up these cytolinkers. Generally, epithelial plakins exhibit a tripartite architecture defined by a central rod domain flanked by globular N- and C-terminal domains, whereby each globular domain is composed of protein modules that can interact with distinct cytoskeletal or cell adhesion targets. At the N-terminus of all plakins (except for an unusual plakin called epiplakin) is a protein module called the plakin domain (**Figure 1**). This domain has been shown to interact with transmembrane components of desmosomal and hemidesmosomal junctions. The crystal structure of an N-terminal segment of the plakin domain of BPAG1 has been solved. It consists of a pair of three α-helical bundles aligned in an antiparallel array that are connected by a helical linker region. This structure is analogous to the secondary and tertiary structure that defines spectrin repeats found in spectrin family members, although plakin domains are more similar to each other than to spectrin-repeat regions found in spectrin proteins. Although the crystal structure of the entire plakin domain has not been solved, protein domain prediction programs suggest that the complete domain may consist of an SH3 domain flanked by two or more pairs of spectrin repeats. The presence of a putative SH3 domain suggests that plakins may serve as molecular scaffolds that recruit signaling proteins to sites of cell–cell contact, cell–ECM contact, or cytoskeletal activity. In addition to the plakin domain, some plakins also have a calponin-type ABD preceding the plakin domain at its N-terminus. This domain is described in more detail later.

Whereas the plakin domain mediates connections with cell–junctional complexes, it is the C-terminal end of several epithelial plakins that mediates connections with IFs. Connections to the IF network are mediated through a region at the C-terminus composed of plakin repeat motifs. Plakin repeats are described as repeats of 38 residues that form a β-hairpin followed by two antiparallel α-helices. Four and a half of these repeats form a globular PRD. PRDs based on sequence similarity are divided into subgroups termed A, B, and C. The PRDs have been shown to function as IF binding domains. For some plakins, IF binding also requires the presence of a linker (L) subdomain. The number of PRDs varies among different plakins. For example, the desmosomal protein desmoplakin contains three PRDs, whereas hemidesmosomal proteins plectin and BPAG1-e contain six and two, respectively. Although it is thought that the primary function of plakin repeats is to mediate binding to cytoplasmic IFs, there is evidence that they may have additional functions. First, there are plakin isoforms in which the plakin repeat motifs are not arranged as the 4.5-repeat PRD that is seen in mammalian epithelial plakins. Second, plakin homologs present in *Drosophila* contain plakin repeats, yet the *Drosophila* genome lacks genes encoding cytoplasmic IFs. The plakin repeats in both *Drosophila* and *Caenorhabditis elegans* are not arranged as PRDs. In fact, invertebrate plakins are quite different in structure from the epithelial plakins described so far. Although their N-terminal region contains a plakin domain, the remainder of the invertebrate plakins diverges from the epithelial plakins (**Figure 1**). In addition to the plakin repeats, some *Drosophila* and *C. elegans* plakin isoforms contain a series of spectrin repeats which is thought to confer some flexibility to the molecule as well as serve as a site for protein interaction. These isoforms also contain a C-terminal region that consists of EF-hand motifs and an MT binding region in turn composed of a GAR domain.

The *Drosophila* plakin homolog is called *Shortstop* or Shot (also referred to as *kakapo*). As mentioned previously, the domain architecture of Shot diverges from mammalian epithelial plakins. However, it is quite similar to mammalian plakins that are expressed highly in the nervous system. These plakins include BPAG1-a (neuronal BPAG1 isoform) and MACF1-a (neuronal MACF1 isoform). Insights into the function of Shot came from studies concerning position-specific (PS) integrins. In *Drosophila*, PS integrins, which are similar to mammalian β_1 integrins, have been shown to be important for integrin-mediated adhesions between dorsal and ventral wing surfaces and between muscle and epidermis (hemiadherens junctions). Mutations of PS integrins result in defects in epidermal integrity and epidermal muscle attachment. A genetic screen for mutants that produce

similar phenotypes led to the identification of Shot. Shot behaves similarly to mammalian plakins in that it anchors integrin-mediated adhesion complexes to cytoskeletal elements. However, since *Drosophila* lack cytosplasmic IFs, Shot anchors the junctional complex to MTs, through interactions with its GAR domain. In *Drosophila*, MTs seem to have taken on the role of IFs as the filaments that stabilize cells subject to mechanical stress.

In *C. elegans*, muscle cells are mechanically linked to the cuticle through electron-dense structures similar to mammalian hemidesmosomes called fibrous organelles (FOs). FOs consist of an IF array anchored at the basal surface to the underlying basement membrane and associated muscle and at the apical surface to the cuticle. VAB-10, the plakin homolog in *C. elegans*, functions similarly to hemidesmosomal plakins. However, unlike hemidesmosomes, FOs do not contain integrins or a BPAG2 homolog; instead, other transmembrane proteins (myotactin, MUA-3, and MUP-4) that might function analogously are present. Two VAB-10 isoforms have been identified – VAB-10A, which localizes to FOs, and VAB-10B, which localizes to bands between FOs. Loss of both VAB-10 isoforms results in epidermal detachment from the cuticle and muscle and perturbation of IFs and circumferential actin bundles. The role of VAB-10 in the nervous system has not been elucidated.

Thus far, our discussion on the role of plakins as key integrators of plasma membrane components and the cytoskeleton has been restricted to epithelia and muscle, tissues that are subject to substantial mechanical stress. Are plakins involved in the formation and/or maintenance of adhesive connections between neurons or between neurons and the underlying ECM? Such connections would have important implications for development and maintenance of neural structures. Several reports hint that indeed plakins may be involved in events that require neural adhesion. For example, mutations in *Drosophila* Shot affect sensory and motor axon growth, terminal arborization at the neuromuscular junction (NMJ), the assembly of presynaptic specializations at the NMJ, and the formation of neuronal dendrites in the central nervous system (CNS) – all cellular events that require cell adhesion.

In mammals, the plakin domain of BPAG1 interacts with the LAP protein erbin. LAP proteins, which include erbin, scribble, and densin-180, are PDZ binding proteins that are involved in the clustering of receptors at synaptic junctions. Erbin interacts with ErbB2, a mammalian epidermal growth factor receptor that in turn interacts with neuregulin. Neuregulin is a signaling protein that is essential for the formation and maintenance of the NMJ. Erbin has been found to localize to postsynaptic sites at the NMJ and sites of ErbB2 localization in the CNS. Thus, BPAG1 in conjunction with erbin may play an important role in the regulation of neuregulin signaling and subcellular organization of ErbB proteins, which in turn could play an important role in the formation and maintenance of NMJs. Despite these connections to NMJ integrity, there is no direct evidence that shows that plakins link the cytoskeleton to junctional complexes in the nervous system.

Cytoskeletal Cross-Linking

Intermediate Filament and Actin Connections

Disruption of IF organization due to pharmacological perturbations of actin function illustrates the interdependence of these two cytoskeletal filament networks. In epithelia, initiation of keratin IF formation is thought to be dependent on a functional actin-rich cortex. *In vitro* studies with *Xenopus laevis* egg extracts suggest that actin cables act as templates upon which cytoplasmic keratin IFs assemble after transition into the first mitotic interphase. A role for actin in IF motility has been identified from studies that compare the difference in motile properties of keratin and vimentin IFs in epithelial cells. These studies show that in the absence of MTs, keratin squiggles (short filaments) continue to move, whereas vimentin squiggles do not, suggesting that actin might be the alternative mode of transport for keratin. Additional evidence supporting actin-based IF motility includes reports that more than half of the total myosin Va (actin-associated molecular motor protein) in the CNS associates with neuronal IFs. Deletion of myosin Va, which is abundant in nerve fiber tracts of brain and spinal cord, results in altered distribution of IFs within axons.

Members of the plakin family may be mediators of IF–actin interactions involved in IF organization and IF motility. Several plakins are able to interact with both actin and IFs. These interactions are mediated by the IF-interacting PRD described previously and an N-terminal ABD composed of two calponin homology (CH) domains arranged in tandem (Figure 1). CH domains are protein modules of approximately 100 residues that are present in a number of structural and regulatory actin binding proteins, such as α-actinin, calponin, filamins, dystrophin, and spectrins. Although similar in secondary structure, CH domains are divided into several subtypes based on their primary sequence. Studies show that although the CH1 subtype has the ability to bind actin and the CH2 subtype does not, it is the combination of both CH1 and CH2 arranged in tandem that constitutes a functional ABD. Transfection and *in vitro* binding

assays demonstrate that the CH1–CH2 motifs of mouse BPAG1, MACF1, plectin, and *Drosophila* Shot serve as functional ABDs that bind directly to actin and co-localize with actin filaments.

The possible importance of plakin-mediated IF–actin cross-linking is seen in disease states in which plakin function has been perturbed. For instance, deletions within the BPAG1 gene are linked to a mild skin blistering phenotype and severe sensory neuron loss in the dystonia musculorum (*dt*) mouse mutant. *Dt* mice display dystonic movements and usually die before weaning. Those animals that survive past the weaning stage exhibit lower motor neuron degeneration due to sensory neuron loss. The degenerating sensory neurons display abnormal accumulation of neuronal IFs and disorganization of the MT network. The observed skin blistering phenotype seen in *dt* mice is due to the absence of BPAG1-e. This results in disruption of connections between hemidesmosomes and IFs, leaving epithelial cells vulnerable to shearing due to mechanical stress. Disorganization of the neuronal cytoskeleton is linked to the absence of several neuronal BPAG1 isoforms. One such isoform is BPAG1-n, which has the capacity to connect actin and IFs and is thought to be important in anchoring neuronal IFs to the axonal actin cytoskeleton. Loss of this connection could contribute to the abnormal accumulation of neuronal IFs seen in *dt* mice. However, it is important to note that BPAG1-n is a minor neuronal isoform and that cytoskeletal defects seen in *dt* mice likely involve other neuronal BPAG1 isoforms, especially BPAG1-a.

As described previously, plectin, another member of the plakin family, also has a functional ABD and a C-terminal PRD-containing region that can interact with IFs. Plectin has many alternate promoters that allow for its tissue-specific expression. One particular isoform (plectin 1C) has been reported to be expressed in neurons and can link actin filaments and IFs. However, elucidation of the importance of plectin in the nervous system may require the isoform-specific knockout of this plakin. Plectin may play a role in regulating the dynamics of actin because cultured plectin null fibroblasts display perturbations in cell motility and Rho/Rac/Cdc42-mediated actin rearrangements.

Intermediate Filament and Microtubule Connections

It has long been known that IFs also associate with MTs. Evidence for this association comes from electron microscopy studies that reveal IF–MT are linked at the ultrastructural level. Pharmacological perturbation studies also show that inhibition of MT function results in reorganization of IFs as well as impairment of intracellular transport. For instance, cultured sympathetic neurons treated with nocodazole exhibit profound inhibition of neurofilament (NF) transport.

NFs were traditionally thought to undergo slow axonal transport. However, observations of living sympathetic neurons expressing GFP-tagged NFs reveal that transported NF structures move at rates similar to those seen for fast axonal transport, although the movements are frequently interrupted by prolonged pauses. It has been shown that during anterograde transport in mature sensory neurons, NFs pause 99% of the time. Molecules that are responsible for the transport of NFs include motor proteins of the kinesin and dynein families that regulate anterograde and retrograde transport of NFs, respectively.

Plakins may play a role in mediating NF–MT interactions given that mutations in the plakin gene BPAG1 lead to disorganization of MT and neural IF networks. Several reports suggest that the primary function of BPAG1 is to stabilize the MT network. Once the integrity of this network is perturbed, many MT-based functions such as intracellular transport may be impaired. BPAG1-a and BPAG1-b isoforms associate with MTs via their microtubule binding region, which is composed of a GAR domain that binds MTs and a C-terminal tail that contains GSR repeats, which is thought to be important for MT bundling. BPAG1-b also has a PRD, which could interact with IFs, but there are no published reports of this interaction. Furthermore, there is little BPAG1-b present in the brain.

In addition to associating with and stabilizing MTs, one report has revealed a potential interaction of a component of the minus-end-directed dynein motor with a region of BPAG1 that encompasses a portion of the spectrin-repeat rod domain. Although this would implicate BPAG1 as a participant in retrograde transport, it is important to note that there have been conflicting reports concerning the role of BPAG1 in axonal transport. The discrepancies may be due to the fact that several plakins are expressed in the nervous system and so may compensate for one another.

Some BPAG1 isoforms may have the ability to cross-link IFs with MTs. This cross-linking could serve as a molecular brake against rapid transport of IFs and thus contribute to the frequent IF pausing that is observed during NF axonal transport. The plakins may also contribute to the stabilization of cytoskeletal networks in specific regions of the cell experiencing mechanical stress.

The other major plakin expressed in the nervous system is MACF1. MACF1-a has a similar domain structure as BPAG1-a and is the primary isoform of MACF1 in the nervous system. However, it has no

obvious IF binding domains. Another isoform, MACF1-b, has a series of PRDs that could bind to IFs. However, MACF1-b is expressed primarily in muscle tissues, and there is no evidence that it functions in the nervous system to link MTs and IFs.

Actin and Microtubule Connections

As mentioned previously, the cytoskeleton plays a vital role in dynamic cellular processes such as neuronal pathfinding. Proper development of the nervous system as well as functional recovery after insults to the brain due to disease or injury are dependent on mechanisms that guide the processes of developing neurons to their correct target. Mediating this guidance behavior is a highly dynamic sensorimotor structure at the distal ends of axons and dendrites called the growth cone. While surveying the extracellular environment, growth cones sense guidance cue information and integrate it into cytoskeletal changes that result in directed movement toward a specific target. It is well documented that the dynamic behavior of the growth cone is dependent primarily on the organization of the actin and MT cytoskeletons. Advances in high-resolution time-lapse imaging have made it possible to directly visualize cytoskeletal changes that underlie growth cone behaviors and in recent years have sparked an explosion of research examining the role of cooperative cytoskeletal activity in growth cone guidance.

The growth cone cytoskeleton is divided into three major domains: an actin-rich peripheral (P) domain, a central (C) domain predominated by MTs, and a transition (T) domain at the interface of the P and C domains (Figure 3). The P domain is characterized by filopodia, which are thin finger-like structures present at the leading edge of the growth cone, and lamellipodia, which are veil-like structures situated between filopodial protrusions. Filopodia are composed of actin bundles and extend forward to survey the environment, whereas lamellipodia make contact with the ECM and are composed of a dense actin meshwork. The actin cytoskeleton is arranged in a polarized manner such that the plus ends of actin filaments, sites of polymer growth, are directed toward the leading edge. A dense array of stable, polarized MT bundles in the axon shaft extend into the C domain, where they defasciculate and exhibit complex dynamic behaviors that include looping, bending, splaying out, and extension

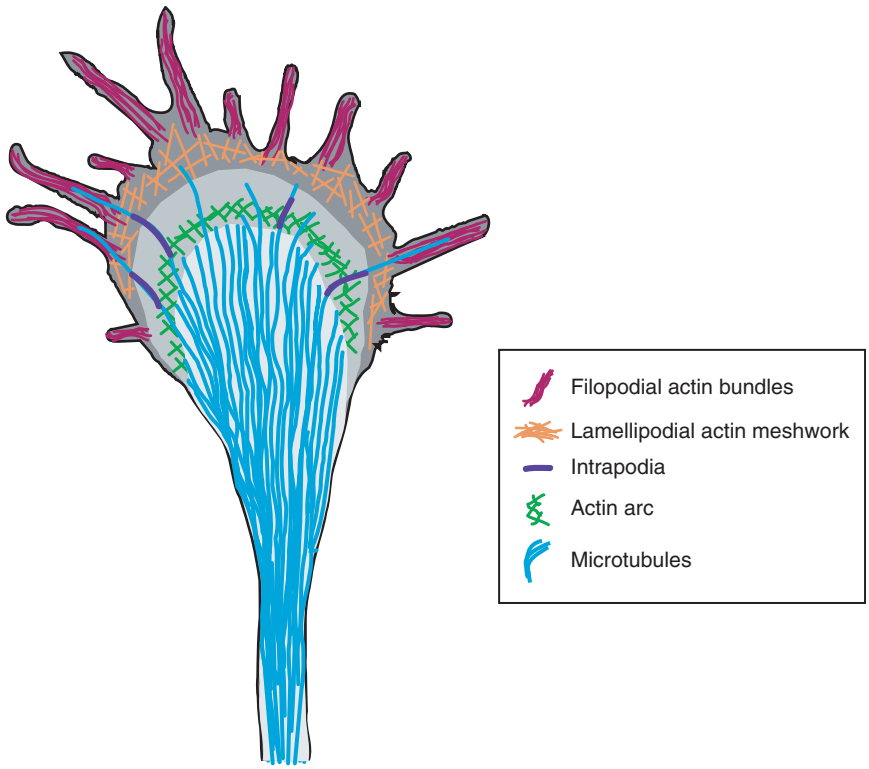

Figure 3 Neuronal growth cone. The schematic illustrates the organization of the growth cone cytoskeleton. The peripheral region (dark gray) consists of two different actin filament populations: actin bundles (pink) present in filopodial structures and an actin meshwork (orange) present in lamellipodial structures. Microtubules (blue) which are bundled in the axon shaft enter the central domain (light gray), where they defasciculate. Microtubules can also be seen extending into peripheral regions of the growth cone. The transition domain (gray) is at the interface of the peripheral and central domains and is composed of actin arcs (green) and intrapodia (purple) that co-localize with microtubules extending into the periphery.

into more peripheral regions of the growth cone (through probing with their plus ends). This domain is also packed with intracellular membrane structures such as vesicles and organelles. The T domain, a site of active membrane ruffling and actin–MT interaction, contains MT-associated actin arcs (contractile actin bundles) that are positioned perpendicular to filopodia and intrapodia, motile actin and β_1 integrin-rich protrusive structures.

Early studies of the growth cone suggested that the actin and MT filament systems are interdependent. For example, application of actin filament depolymerizing drugs results in robust MT advancement into the P domain, whereas inhibiting MT activity results in the formation of lamellipodial and filopodial structures along the length of the axon shaft. The suggestion that the two filament networks are physically linked is supported by imaging studies that show that myosin-driven actin retrograde flow in filopodial structures aids in the rearward movement of MTs. This retrograde transport contributes to MT clearance from the P domain. Retrograde flow also contributes to the formation of MT loops, which signifies growth cone pausing and MT buckling and breakage. The breakage creates new minus ends that quickly depolymerize and new plus ends that can elongate into peripheral regions. Thus, this transport mechanism, which contributes to the cycle of MT clearance and MT advancement in the P domain, reflects the interdependence of actin and MT networks of the growth cone.

Whereas retrograde transport serves to regulate the spatial distribution of cytoskeletal elements in the growth cone, it is clear that the molecular machinery required for eliciting growth cone turning responses is dependent on factors that control stabilization of actin and MTs. Studies show that focal application of MT-stabilizing drugs on one side of the growth cone results in directed growth toward the side in which MTs have been stabilized. Interestingly, this also enhances lamellipodial protrusion on that side before turning occurs. Conversely, when MTs are pharmacologically destabilized on one side of the growth cone, the growth cone turns away from this region and instead grows in the direction containing stabilized MTs. Focal disruption of stable actin bundles with actin depolymerizing agents inhibits actin–MT interaction in that region and consequently results in directed growth toward the side of stabilized filopodia, where MTs are still able to extend. Factors involved in this process include cytoskeletal-associated proteins that function to stabilize the actin and MT cytoskeletons, as well as proteins that serve to physically link the two cytoskeletal systems. Such factors include members of the plakin family.

As mentioned previously, plakins are able to cross-link different cytoskeletal elements to each other. Genetic studies in *Drosophila* suggest that the cytoskeletal cross-linking activity of plakins is involved in several processes that underlie growth cone activity. Reduction in the activity of the *Drosophila* plakin homolog Shot, which localizes throughout the growth cone, results in defects in axon extension, dendrite formation, and axonal terminal arborization at the NMJ. The activity of Shot is dependent on its ability to interact with both actin and MTs. Interestingly, Shot and its mammalian counterparts, MACF1-a and BPAG1-a, interact with the microtubule plus-end protein EB1, which has been localized to the growth cone. EB1 is important for MT capture events at the cell cortex. However, its role in growth cone function has not been clearly defined. Defects in BPAG1 result in severe sensory neuron degeneration, in which affected neurons exhibit abnormal accumulation of neuronal IFs in axon swellings and disorganization of MT network. BPAG1 is expressed in axonal and dendritic processes and has been shown to stabilize MTs; it may therefore be involved in stabilization of MTs required for growth cone activity. However, a role for BPAG1 in the growth cone has not been elucidated. Cultured MACF null endothelial cells show defects in MT activity. In these cells, MTs no longer stabilize at the actin-rich cortex and they no longer grow along actin bundles. This suggests that in the context of the neuron, MACF may function in stabilizing actin–MT interactions. Further studies are needed to establish definitive conclusions.

Concluding Remarks

Cross-linking cytoskeletal elements to each other and to specialized sites at the plasma membrane is important for many dynamic cellular processes. This article focused primarily on members of the plakin family, key facilitators for such linkages. Because plakins were initially identified as key structural components in epithelia, most of our knowledge centered on their role in ensuring proper development and maintenance of epithelial structures. Further analysis of the gene loci of plakins indicated that plakins are complex molecules that contain numerous interaction modules that are alternatively combined in a tissue-specific manner. This analysis led to the identification of several plakin proteins that are highly expressed in the nervous system. Current studies suggest that the presence of these complex molecular

goliaths is critical for proper neuronal function. This is seen with mutations in the BPAG1 gene in mice, which result in severe sensory neurodegeneration, where degenerating neurons exhibit disorganization of NF and MT networks. It is also seen in invertebrate systems, in which reduction in the activity of plakins leads to defects in axonal growth, axonal and dendritic arborization, and defects at the NMJ. Although it is likely that plakins are important for neuronal function, definitive evidence to support this claim is lacking. Because there are multiple plakins that are expressed in the nervous system, conditional knockout of tissue-specific isoforms will be useful in elucidating the roles of plakins in neurons.

In addition to plakins, there are a host of other proteins that play essential roles in mediating the dynamics and physical connections between different cytoskeletal elements. Proteins once thought to exclusively associate with one filament system have now been shown to interact with multiple networks. For example, microtubule-associated proteins have been shown to interact with the actin cytoskeleton and with NFs. Motor proteins of the kinesin, dynein, and myosin families are involved in multiprotein motor complexes that operate along multiple cytoskeletal tracks to transport common cellular cargo. In addition to proteins that function to physically link cytoskeletal networks, there is a plethora of signaling proteins that regulate the dynamics of cytoskeletal elements. The best examples of such regulatory elements are members of the Rho family of GTPases. Once thought to play a role in actin organization and dynamics, these molecules are also known to regulate MT dynamics. It is an exciting time in the field of neuronal cell biology, in which it is hoped that insights into the structural and regulatory components of cytoskeletal connections will provide a better understanding of the molecular events that govern the development and maintenance of neuronal structures.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Cytoskeleton in Plasticity; Growth Cones; Intermediate Filaments; Microtubules: Organization and Function in Neurons; Neurofilaments: Organization and Function in Neurons.

Further Reading

Akhmanova A and Hoogenraad CC (2005) Microtubule plus-end tracking proteins: Mechanisms and functions. *Current Opinion in Cell Biology* 17: 47–54.

Chang L and Goldman RD (2004) Intermediate filaments mediate cytoskeletal crosstalk. *Nature Reviews Molecular Cell Biology* 5: 610–613.

Dehmelt L and Halpain S (2003) Actin and microtubules in neurite initiation: Are MAPS the missing link? *Journal of Neurobiology* 58: 18–33.

Dent EW and Gertler FB (2003) Cytoskeletal dynamics and transport in growth cone motility and axon guidance. *Neuron* 40: 209–227.

Getsios S, Huen AC, and Green KJ (2004) Working out the strength and flexibility of desmosomes. *Nature Reviews Molecular Cell Biology* 5: 271–281.

Goldstein LS and Yang Z (2000) Microtubule-based transport systems in neurons: The roles of kinesins and dyneins. *Annual Review of Neuroscience* 23: 39–71.

Goode BL, Drubin DG, and Barnes G (2000) Functional cooperation between the microtubule and actin cytoskeletons. *Current Opinion in Cell Biology* 12: 63–71.

Grunwald IC and Klein R (2002) Axon guidance: Receptor complexes and signaling mechanisms. *Current Opinion in Neurobiology* 12: 250–259.

Jefferson JJ, Leung C, and Liem RKH (2004) Plakins: Goliaths that link cell junctions and the cytoskeleton. *Nature Reviews Molecular Cell Biology* 5: 542–553.

Luo L (2000) Rho GTPases in neuronal morphogenesis. *Nature Reviews Neuroscience* 1: 173–180.

Nievers MJ, Schaapveld RQJ, and Sonnenberg A (1999) Biology and function of hemidesmosomes. *Matrix Biology* 18: 5–17.

Ozaki M (2002) Neuregulins and the shaping of synapses. *The Neuroscientist* 17: 146–154.

Rodriguez C, Schaefer AW, Mandato CA, Forscher P, Bement WM, and Waterman-Storer CM (2003) Conserved microtubule–actin interactions in cell movement and morphogenesis. *Nature Reviews Molecular Cell Biology* 5: 599–607.

Cytoskeleton in Plasticity

A Ferrari and Y Goda, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Overview

Neurons are connected to each other and to their target cells via highly specialized structures called synapses. Morphologically, the synapse resembles an asymmetrical intercellular junction, with a presynaptic compartment specialized for neurotransmitter release and a postsynaptic side attuned to the reception of released neurotransmitters and signal propagation. Functionally, the synapse serves one in 10^{14} internodes in a very complex network, and modulating the efficacy of synaptic transmission – the synaptic strength – is central for the coordination of multiple activities in our body. Despite the plasticity of synaptic strength, historically, the synapse was regarded as a stable, static structure. In recent years, particularly with advances in imaging technology, experimental evidence has revealed a dynamic nature of the synapse that encompasses its morphology.

This article focuses on the relationship between the neuronal cytoskeleton and its changes during synaptic plasticity, particularly in dendritic spines. We concentrate mostly on the actin cytoskeleton because it is the major cytoskeletal component present at the synapse and the best studied. We describe the morphological characteristics of a synapse at pre- and postsynaptic terminals, and will also discuss the different forms of synaptic plasticity and the possible intracellular signaling pathways involved in coordinating functional and structural changes. Finally, the importance of actin regulation is highlighted through some examples of human diseases. All the data and information reported in this article comes from studies on excitatory synapses, where the synapse structure–function relationship has been actively investigated.

The Cytoskeleton at the Synapse

Structurally, synapses in the central nervous system (CNS) represent strongly adhesive junctions that are similar to adherens junctions in nonneuronal cells. Conserved properties between these two types of intercellular contacts include (at synapses) (1) Ca^{2+}-dependent homophilic cell adhesion molecules (the cadherin–catenin complex), (2) electron-dense material at both sides of the cellular membrane (the postsynaptic density (PSD) and the active zone), and (3) the intercellular distance between the opposite membranes (~25 nm; the synaptic cleft). In a speculative view, the physiological role of the synapse could have been refined during evolution, over a simple junction by the addition of the molecular machinery specific to conveying information across the intercellular gap. On the presynaptic side, the machinery is involved in the organization of synaptic vesicles (SVs) – such as their clustering, docking, fusion, and retrieval, which are needed for regulated neurotransmitter release. On the postsynaptic side, the assembly of neurotransmitter receptors modulates intracellular signaling and controls the excitability of the cell.

The Presynaptic Terminal

In the late 1980s the first electron microscopy (EM) structural reports on the presynaptic organization were published from various preparations. Strikingly, the presynaptic cytoarchitecture shared a high degree of homology, from insects to fish and rodents. The presynaptic cytomatrix consists mainly of fibrillar structures that extend from the specialized presynaptic plasma membrane, called the active zone, deep into the presynaptic bouton. Among the high molecular weight proteins, piccolo and bassoon and other cytomatrix proteins are thought to provide the framework along which SVs are guided toward the active zone, where SVs fuse to release their contents, neurotransmitters. The major components of the presynaptic cytoskeleton are filamentous actin (F-actin), synapsin, fodrin, and, at some synapses, microtubules. F-actin is enriched throughout the synaptic bouton and colocalizes partially with synapsin. Synapsin has been suggested to regulate the availability of SVs for neurotransmitter release by sequestering the reserve SV pool to the F-actin scaffold.

The Postsynaptic Terminal

The postsynaptic compartment represents a cellular specialization where neurotransmitter receptors are enriched by being anchored to the PSD. In electron micrographs, the PSD is highly electron dense, which reflects the fact that it is extremely protein rich, and is roughly 20 nm thick and 200–400 nm long. More than 200 different proteins have been reported to be associated with the PSD, including receptors, scaffolding proteins, and other proteins involved in signal transduction and receptor trafficking. The PSD has a crucial role in mediating signals from the glutamate receptors to intracellular pathways, and, therefore, in modulating activity-dependent responses. For instance, scaffold proteins such as postsynaptic density-95 (PSD-95) protein physically link the *N*-methyl-D-aspartate (NMDA) subtype of glutamate receptors

Table 1 Key actin regulatory proteins and actin effectors

Protein	*Function*	*Phenotype*
ADF/cofilin	Actin-binding protein	Inhibition leads to defective LTP late phase
Cdc42	Small Rho-GTPase	Inhibition causes increased filopodia and decreased spine formation
Cortactin	Actin-binding protein	Overexpression increases spine length and inhibition decreases spine number
Eph4	Receptor tyrosine kinase	Knockout shows increased spine density and elongation
Gelsolin	Actin-binding protein	Knockout shows impaired LTD and increased NMDAR current
Intersectin	Cdc42 guanosine exchange factor (GEF)	Inhibition leads to less spine formation and increased filopodia
Kalirin	Rac guanosine exchange factor (GEF)	Positively regulates spine density
LIMK	ADF/cofilin regulator	Knockout shows spines with wider heads and thinner necks, enhanced LTP, synaptic depression, and miniature excitatory postsynaptic currents (mEPSCs) frequency
N-WASP	Interacts with ARP2/3	Downstream effector of Cdc42; its first blockade leads to increased filopodia
Oligophrenin	Rho-activating protein (GAP)	Inhibition leads to decrease in spine length
PAK	Rac effector	Upstream of LIMK, inhibition causes increased spine head size with reduced spine length. Increased LTP, decreased LTD
PIX	Cdc42 and Rac guanosine exchange factor (GEF)	Inhibition results in reduced spine and synapse number
Profilin	Actin-binding protein	Inhibition causes spine elongation and spine head loss
Rac	Small Rho-GTPase	Activation leads to increased filopodia and lamellipodia; inhibition leads to decreased spine number
Rho	Small Rho-GTPase	Its activation leads to reduced spine length and density, decreased branching

to regulators, which in turn couple to the different signaling cascades that impact the postsynaptic cytoskeleton. Such pathways include Rac/Cdc42, Ras/MAPK, or Rho/Rap (see Table 1 for a summary of representative proteins involved).

Dendritic Spines

Dendritic spines are distinct protrusive structures that decorate the main shaft of the dendrite and harbor the postsynaptic terminals. A majority of excitatory synapses are formed on spines, whereas inhibitory synapses have a smooth appearance. The dendritic spines were first described by Ramón y Cajal toward the end of the nineteenth century. Cajal insightfully postulated that morphological plasticity of spines might have an important role in brain function. It was only in the middle of the twentieth century that EM studies performed by Gray established spines as the reception sites for neurotransmitters, thereby mediating neurotransmission at most excitatory synapses. Spines are formed of two distinct compartments: one is the neck, which connects the spine to the dendritic shaft, and the second, the body of the spine, called the spine head, is an enlarged structure attached at the end of the spine neck housing the main PSD and structural proteins of the spine. The spine head diameter can range from ~0.6 to ~1.1 μm, and comes in various shapes, such as thin, stubby, mushroom, cuplike, or branched, as identified by Cajal. The heterogeneity of spine shapes might represent their functional diversity.

An important function of spines is Ca^{2+} regulation and signaling. Accordingly, the common feature of all spines is the presence of endoplasmic reticulum (ER) compartments that are interconnected to the ER in the dendrite, and the presence of inositol triphosphate (IP3) receptors and Ca^{2+}-ATPase. Upon neurotransmitter release from the presynaptic terminal and activation of postsynaptic receptors, Ca^{2+} influx into the dendritic spine can occur through the NMDA receptors and voltage-gated Ca^{2+} channels. The extent to which Ca^{2+} level rises in the dendritic spine and the time window in which it occurs are critical parameters that regulate the extent and directionality of synaptic plasticity. The physical structure of the spine has been suggested to play a key role in Ca^{2+} compartmentalization where the neck restricts Ca^{2+} movement between the spine head and the dendritic shaft. In general, small and thin spines show higher relative increases in Ca^{2+} and a higher degree of plasticity and motility. On the contrary, larger mushroom-shaped spines might experience smaller changes in Ca^{2+} levels and are more stable. The contribution of different spine types to particular forms of synaptic plasticity remains to be clarified.

Dendritic Spines: Cytoskeleton

The actin filaments and actin regulatory proteins have an essential role in conferring the unique spine shape and driving spine motility. Microtubules, which are present in the dendritic shaft, are usually absent from dendritic spines. As a major component of the

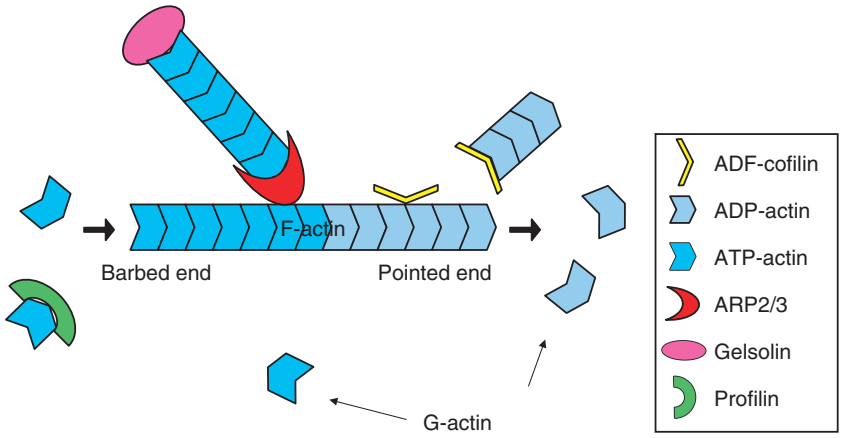

Figure 1 Dynamic interactions with actin modulators regulate the balance of F- and G-actins. Adapted from Carlisle HJ and Kennedy MB (2005) Spine architecture and synaptic plasticity. *Trends in Neuroscience* 28: 182–187, with permission from Elsevier.

scaffold structure of the spine, actin forms a dense lattice around the ER in the spine neck and head. Activity modulates actin dynamics and the balance between monomeric G-actin and filamentous F-actin. Moreover, interfering with the functional activity of actin modulators, in turn, impairs synaptic plasticity.

F-actin is a double-stranded filament composed of G-actin, and is a dynamic structure that undergoes continuous turnover, where monomers are added at the 'barbed' end and removed from the 'pointed' end (Figure 1). Growth of filaments depends on the availability of ATP-bound G-actin versus ADP-bound G-actin, because only the ATP-bound G-actin is added onto a growing filament. As the filament 'ages,' ATP is exchanged for ADP, whereupon G-actin is lost from the filament. Various proteins regulate actin turnover. Actin-depolymerization factor (ADF)/cofilin binds preferentially to ADP-bound actin to promote actin depolymerization, and it can also sever filaments. Profilin promotes actin growth by exchanging ADP with ATP on G-actin monomers and, therefore, maintains the supply of ATP-bound G-actin pool needed for F-actin growth. Other actin-interacting proteins include ARP2/3, which initiates the nucleation of new actin filaments and/or branching of preexisting actin filaments, and gelsolin, which is an actin-capping protein that prevents addition of new G-actin to the barbed growing end of F-actin filaments. Calmodulin also interacts with actin and modulates the activities of fodrin, myosin, and the microtubule-associated protein (MAP-2) and Ca^{2+} signaling. These properties of calmodulin serve to couple the structure of the spine to functional synaptic plasticity involving Ca^{2+}-regulated events.

Forms of Synaptic Plasticity

Two major and best-studied forms of synaptic plasticity are long-term potentiation (LTP) and long-term depression (LTD), which represent the adaptability of synapse to variable activity regimes. The balance between the two phenomena is thought to regulate experience-dependent processes such as learning and memory storage in the brain and control the dynamic range of neuronal network activity. Synaptic strength changes during LTP and LTD are mirrored by structural plasticity of synapses.

LTP and Activity-Dependent Actin Remodeling

LTP was first described in the early 1970s as an increase in synaptic strength mediated by coincident activation of pre- and postsynaptic neurons. By increasing synaptic strength of coactive inputs, LTP is thought to play a role in shaping and consolidating neuronal circuit activity for the encoding of specific experiences. LTP induction involves presynaptic activation, for example, by brief tetanic bursts (50–100 Hz). There are several forms of LTP that are either dependent or independent of NMDA receptor activation. In both cases, its induction requires a local rise in postsynaptic Ca^{2+} that must reach a certain threshold to activate specific kinases, such as calmodulin-dependent protein kinase II (CaMKII), protein kinase C (PKC), and protein kinase A (PKA). The signaling culminates in the increase in the number of functional α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) subtype of glutamate receptors, either by rapid insertion of receptors into the PSD and/or by phosphorylation of the receptors already present at the PSD, thereby enhancing the postsynaptic response to presynaptic glutamate release.

It has been long postulated that such physiological alterations in synaptic strength might be reflected by structural changes in the synapse. Early ultrastructural studies showed correlation between high-frequency stimulation, PSD remodeling, and spine duplication, which suggested that morphological changes could

contribute to the increase in synaptic efficacy associated with LTP. In one study, active synapses were marked based on their increased Ca^{2+} level. Subsequent three-dimensional reconstruction of EM serial sections demonstrated that LTP induces an increase in the number of spines adjacent to a mature spine that shared the same presynaptic terminal, with an overall effect of increasing the number of activated synapses.

In recent years it has been possible to visualize in real time the morphology of dendritic spines undergoing LTP induction by using fluorescently labeled dendrites. Following LTP induction, the spine head is enlarged and the spine neck appears to thicken. After repetitive stimulation, spines can split to give rise to two new spines. Time-lapse studies of green fluorescent protein (GFP)-tagged actin have permitted following the live behavior of the pre- and postsynaptic actin cytoskeleton in response to different stimulation protocols aimed at inducing synaptic plasticity. LTP induction by high-frequency stimulation caused an increase of F-actin content. Fluorescence resonance energy transfer (FRET) experiments also showed in real time, upon tetanic stimulation, a shift in the balance between G-actin and F-actin toward F-actin in the dendritic spines, with a net elongation of actin filaments and an increase in their stability. In addition, an *in vivo* study elegantly showed, after LTP induction in the dentate gyrus, an increase in F-actin in the spines, as identified by phalloidin staining. Such an increase was dependent on NMDA receptor activation, actin polymerization, and ADF/cofilin inactivation, but not on *de novo* protein synthesis, which suggested that F-actin increase occurred in the early phase of LTP. Indeed, the increase was apparent within 20 min after LTP induction and lasted for at least 5 weeks.

Long-Term Depression

LTD is the phenomenon by which synapses become less efficient in transmitting neuronal signals. As a counterpart to LTP, it is also important for adapting neural networks to physiological activity requirements. LTD has been characterized in various systems and brain areas, including the cerebellum, the hippocampus, the cerebral cortex, and the striatum. Experimentally, LTD induction requires low-frequency stimulation (e.g., 1 Hz) over a prolonged time period (i.e., 5–15 min). In the cerebellum, LTD plays an important role in the trial-and-error motor learning processes, and it has been most thoroughly characterized at the synapses formed between parallel fibers and Purkinje cells. In the hippocampus, LTD has been primarily investigated at the CA3 to CA1 connections, in association with spatial learning in rodents.

Similar to the differences in the polarity of synaptic strength changes between LTP and LTD, actin cytoskeletal dynamics behave in the opposite direction during LTD, compared to LTP. In the FRET experiments using fluorescently labeled actin, low-frequency stimulation to induce LTD shifted the balance between G-actin and F-actin toward G-actin in the dendritic spines, leading to a net loss of postsynaptic F-actin. These changes were accompanied by the shrinkage of the spine and, sometimes, by spine disappearance.

Postsynaptic Signaling Pathways

What are the possible biochemical signaling pathways that mediate actin remodeling and structural plasticity at the synapse? Intracellular signaling pathways involve different protein families that act in a coordinate manner, interact with each other, and form specific complexes in order to exert their final effect on actin dynamics, and, therefore, on spine morphology and turnover. At the spine, the following classes of proteins participate in signaling: (1) neurotransmitter receptors, which initialize the signal at the cell membrane, (2) scaffold proteins, which physically link interacting proteins in specific complexes, (3) the Rho-GTPases, which execute changes in actin dynamics, (4) GEFs (guanine exchange factors), which activate Rho-GTPases, (5) GAPs (GTPase-activating proteins), which inactivate Rho-GTPases, and, finally, (6) various adaptor proteins, which ensure signal specificity by bringing specific scaffold proteins into contact with its mediators (Table 1). As the effector of actin dynamics, major players in spine remodeling are the Rho family GTPases, such as Ras, Rap, Cdc42, and Rac. Although these actin effectors represent independent pathways, there is a high degree of cross-talk between them. Signal specificity is therefore facilitated by spatial localization of effectors and coeffectors via scaffolding proteins.

Upon LTP or LTD induction, NMDA receptor activation triggers a number of signaling cascades. These include activation of GEFs for Ras (RasGRF) and therefore Ras activation, CaMKII activity modulation, interaction with SynGAP, and inhibition of Ras. In addition, indirect pathways modulate activities of gelsolin and profilin, which are proteins that regulate actin dynamics by F-actin capping and ADP–ATP exchange on G-actin.

Some activity-dependent changes involve spine formation, in which EphB receptors (EphBRs) might have a role. EphBRs are activated through ligand binding, and lead to kalirin-7 tyrosine phosphorylation. Kalirin-7 is a Rho-GEF with a Rac1 GEF domain, and upon activation, it translocates to the spine head. Overexpression of a kalirin-inactive mutant or of a Rac1-dominant negative mutant blocks the effect of EphB ligand on spine formation. Signaling downstream of EphBR–kalirin is mediated by Rac,

p21-activated kinases (PAK), and LIM motif-containing kinase (LIMK), which modulate ADF/cofilin activity, to exert changes in actin dynamics. An alternative pathway engaged following EphBR activation involves intersectin. Intersectin is a scaffold protein that interacts with several signaling proteins, among which is Cdc42. A large complex composed of intersectin, Cdc42, and a Wiskott-Aldrich syndrome protein, N-WASP, nucleates new F-actin branches via activation of ARP2/3 by N-WASP.

Other examples of receptor-mediated postsynaptic intracellular signaling are represented by glutamate receptor interacting protein (GRIP), which binds to AMPA receptors, and homer, which binds to IP3 receptors and to metabotropic glutamate receptors (mGluRs). Both GRIP and Homer link to the actin cytoskeleton via Shank, and, either via cortactin or via fodrin. The ability of high molecular weight proteins such as GKAP and Shank to interact with three postsynaptic scaffold proteins, PSD-95, GRIP, and Homer, completes the link between the different classes of glutamate receptors (NMDA, AMPA, and mGluR). Scaffold proteins, therefore, can modulate interactions between signals originating from different sources. Collectively, the postsynaptic scaffold has a dual function: shaping the postsynaptic organization and the signaling responses elicited by it. Despite the extensive interconnectivity, the ability of each receptor subtype to mediate and/or respond to distinct signaling suggests that the PSD might be segregated into substructural domains.

Mental Retardation and Dendritic Spines

The importance of a proper synapse assembly and its morphological regulation is exemplified by a class of human cognitive disorders involving mental retardation (MR). MR, which is reflected by an abnormally low IQ (usually <70), affects 2–3% of the population, and has its onset early during postnatal development. Golgi staining studies of brains of MR patients performed in the 1970s revealed abnormalities in dendritic structures. In particular, dendritic spines were predominantly long and thin, with few mushroom and stubby spines. Since elongated spines resemble filopodia-like structures typical of immature synapses, the etiology of MR could be attributed to the failure of the nervous system to wire correctly. This is due to either genetic or environmental factors, and consequently leads to a suboptimal network performance. The link between disease and spine structure is substantiated by the identification of specific genetic loci, in particular, of genes coding for modulators of actin cytoskeleton. An example is X-linked mental retardation (XLMR). At least 13 genes have been implicated in the pathogenesis of XLMR, and many of the proteins encoded by them are involved in the regulation of Rho family GTPases. The best characterized of such proteins thus far are: oligophrenin 1, PAK3, and ARHGEF6 (also called αPix).

Oligophrenin 1 negatively regulates Rho-GTPases such as RhoA, Rac, and Cdc42 via its GAP domain. The gene encoding oligophrenin 1, *OPHN1*, was first identified as a translocation on chromosome X in a female patient affected by XLMR. Subsequently, other familial forms of the disease associated with a mutation in the *OPHN1* gene have been described. A more direct link between oligophrenin 1 and aberrant spine morphology has been established by recent interference RNA (RNAi) studies. Knocking down of oligophrenin 1 expression in hippocampal slice cultures significantly reduces spine length in the CA1 region. Oligophrenin 1 also interacts with Homer, a postsynaptic scaffold protein for glutamate receptors. This indicates that oligophrenin 1 might be important not only for spine formation and maintenance, but might also have a more direct role in modulating synaptic transmission to regulate cognitive processing.

PAK3 belongs to the family of PAK and acts downstream of Rac and Cdc42. So far, three point mutations of PAK3 have been described in XLMR families: R419X, R67C, and A365E. R419X abolishes the kinase activity of the enzyme, R67C disrupts the interaction with Rho-GTPases, and A365E lies within a highly conserved motif in the catalytic domain that probably affects its kinase activity.

Another MRX gene is *ARHGEF6*, which encodes a GTP exchange factor. In its monomeric form, ARHGEF6 can bind to both Cdc42 and Rac with the same affinity. As a homodimer, however, its affinity for Rac is higher, and it therefore functions as a specific Rac GEF. ARHGEF6 also interacts with PAK, thereby regulating its activity. Mutations in *ARHGEF6* comprise translocations and frame shifts, resulting in a protein that lacks either exon 2 or a portion of exon 10 and exon 11. This could affect the integrin-mediated signaling that leads to activation of Rho-GTPases. ARHGEF6 has been shown to interact with the complex that localizes to the focal adhesions to regulate integrin clustering, including β-parvin (PARVB) and the integrin-linked kinase (ILK).

Concluding Remarks

Advances in molecular and imaging techniques have facilitated a better understanding of the importance of dendritic spine heterogeneity. Current evidence indicates that spines are highly plastic, and they undergo complex changes in both shape and number over time in an activity-dependent manner. A bewildering array of postsynaptic signaling pathways can

be activated by synaptic activity, and this in turn is reflected in the complexity of protein interactions in the PSD. A key future goal will be to understand in quantitative terms how these various postsynaptic mechanisms are coupled to synaptic activity, including neurotransmitter release from the presynaptic terminal, how different signaling systems cross-talk with each other, and how specific patterns of synaptic activity manifest in particular forms of cytoskeletal remodeling.

See also: Cytoskeletal Interactions in the Neuron; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Postsynaptic Development: Neuronal Molecular Scaffolds; Spine Plasticity; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Bonhoeffer T and Yuste R (2002) Spine motility. Phenomenology, mechanisms and function. *Neuron* 35: 1019–1027.

Carlisle HJ and Kennedy MB (2005) Spine architecture and synaptic plasticity. *Trends in Neuroscience* 28: 182–187.

Dillon C and Goda Y (2005) The actin cytoskeleton: Integrating form and function at the synapse. *Annual Review of Neuroscience* 28: 25–55.

Engert F and Bonhoeffer T (1999) Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature* 399: 66–70.

Fischer M, Kaech S, and Knutti D (1998) Rapid actin-based plasticity in dendritic spines. *Neuron* 20: 847–854.

Fukazawa Y, Saitoh Y, Ozawa F, et al. (2003) Hippocampal LTP is accompanied by enhanced F-actin content within the dendritic spine that is essential for late LTP maintenance *in vivo*. *Neuron* 38: 447–460.

Kennedy MB, Beale HC, Carlisle HJ, et al. (2005) Integration of biochemical signaling in spines. *Nature Reviews Neuroscience* 6: 423–434.

Luo L (2002) Actin cytoskeleton regulation in neuronal morphogenesis and structural plasticity. *Annual Review of Cell and Developmental Biology* 18: 601–635.

Matus A (2000) Actin-based plasticity in dendritic spines. *Science* 290: 754–758.

Okamoto KI, Nagai T, Miyawaki A, et al. (2004) Rapid and persistent modulation of actin dynamics regulates postsynaptic reorganization underlying bi-directional plasticity. *Nature Neuroscience* 7: 1104–1112.

Ramakers GJ (2002) Rho proteins, mental retardation and the cellular basis of cognition. *Trends in Neuroscience* 25: 191–199.

Sheng M and Kim MJ (2002) Postsynaptic signaling and plasticity mechanisms. *Science* 298: 776–780.

Deafness

J E Hawkins, Ann Arbor, MI, USA

© 2009 Published by Elsevier Ltd.

This article is a revision of the previous edition article by Joseph Hawkins, © 2004 Elsevier B.V. All rights reserved.

Deafness (anacusis) is the complete inability to hear speech and other sounds, however much they may be amplified. Persons with partial deafness (hypoacusis, dysacusis) are described colloquially as being hard of hearing, or clinically as having a hearing loss. Such deficiencies are all too common, especially in developed countries, with their overcrowded cities, excessive environmental noise and other forms of pollution, and their 'diseases of civilization.' In the United States alone, more than 20 million persons over 3 years of age have trouble hearing. Thus, some 8% of us are, like Chaucer's Wyf of Bathe, 'som-deel deef.' As our population ages and noise levels continue to rise, these figures can only increase. Moreover, some 4.8 million persons are deaf in the proper, absolute sense of that word. Comparable estimates are available for the UK, where the total number of deaf persons has been calculated to be 1 774 000. This figure includes 15 000 deaf mutes; 30 000 totally deaf; 70 000 'deaf to all natural speech' (who could hear speech amplified by hearing aids); and 1 659 000 persons who are hard of hearing.

Assessment of Hearing

The degree of auditory impairment, or hearing loss, is usually estimated by subjective audiometry, using pure tones and speech at precisely measured intensities. Thresholds, that is, sound pressure levels at which tones are just audible, or speech barely intelligible, are compared with established average normal values. The difference in decibels (dB) is the patient's hearing level. The audiogram (Figure 1) shows hearing levels for tones over a range of about six octaves, including the frequencies between 125 and 4000 Hz, which are most important for speech and music. Hearing levels between 0 and 20 dB ISO (i.e., based on the audiometric calibration of the International Standards Organization) are generally considered normal; those between 20 and 40 dB represent a mild hearing loss. Between 40 and 55 dB the loss is regarded as moderate, between 55 and 70 dB as moderately severe. Hearing levels between 70 and 90 dB indicate that the loss is severe and, above 90 dB, that the loss is profound. In these terms, deafness can be defined as a hearing level of 92 dB or more.

Other types of screening tests are used for identifying neonates who are at risk for hearing problems. One important test is based on otoacoustic emissons (OAEs) measured from the external ear canal with a microphone. Normal OAEs indicate normal middle-ear function and normal function of the inner ear at least for the outer hair cells, which are the likely generators of OAEs. Brain stem response measures are also used for hearing screening. The neuronal potentials of the brain stem, as elicited by brief sounds, are recorded with one electrode on the skin over the mastoid process and another on the scalp at the vertex. Normal brain stem responses indicate normal function of the middle ear, inner ear, auditory nerve, and certain important pathways within the brain stem. In electrocochleography (EcoG), the responses of the inner ear (cochlear microphonics and nerve action potentials) are recorded from the cochlea with a fine needle electrode piercing the tympanic membrane to make contact with its bony wall at the promontory. Pain is minimal for the adult, but for children sedation or anesthesia is advisable. On occasion, measurement of the threshold for the acoustic reflex, the brief contraction of the stapedius muscle of the middle ear in response to sudden sounds of high intensity, can be informative.

Causes of Impaired Hearing

When deafness is congenital, as it is in 1 or 2 in 1000 births, it is generally of genetic origin. The majority of such cases of hereditary hearing loss are nonsyndromic, that is, not accompanied by other structural or functional anomalies. Others may be the result of one or another of a large number of genetically determined syndromes involving malformation of the external or middle ear, or degenerative changes in the cochlea. The loss may be associated with visual defects, as in retinitis pigmentosa (Usher syndrome); musculoskeletal malformations, as in mandibulofacial dysostosis (Treacher–Collins syndrome) and osteogenesis imperfecta; pigmentary abnormalities, as in the Waardenburg syndrome; renal disease, as in the Alport syndrome; metabolic disorders such as goiter, as in Pendred's syndrome; various neurological conditions; or trisomy 13, 18, and 21. Also, maternal disease during pregnancy (e.g., rubella and syphilis) or perinatal conditions (e.g., jaundice, anoxia, and trauma) may be responsible. When hearing impairment is acquired, it may be attributable to one or more well-recognized causes, including viral or bacterial disease (e.g., mumps, measles, herpes zoster oticus, meningitis, and

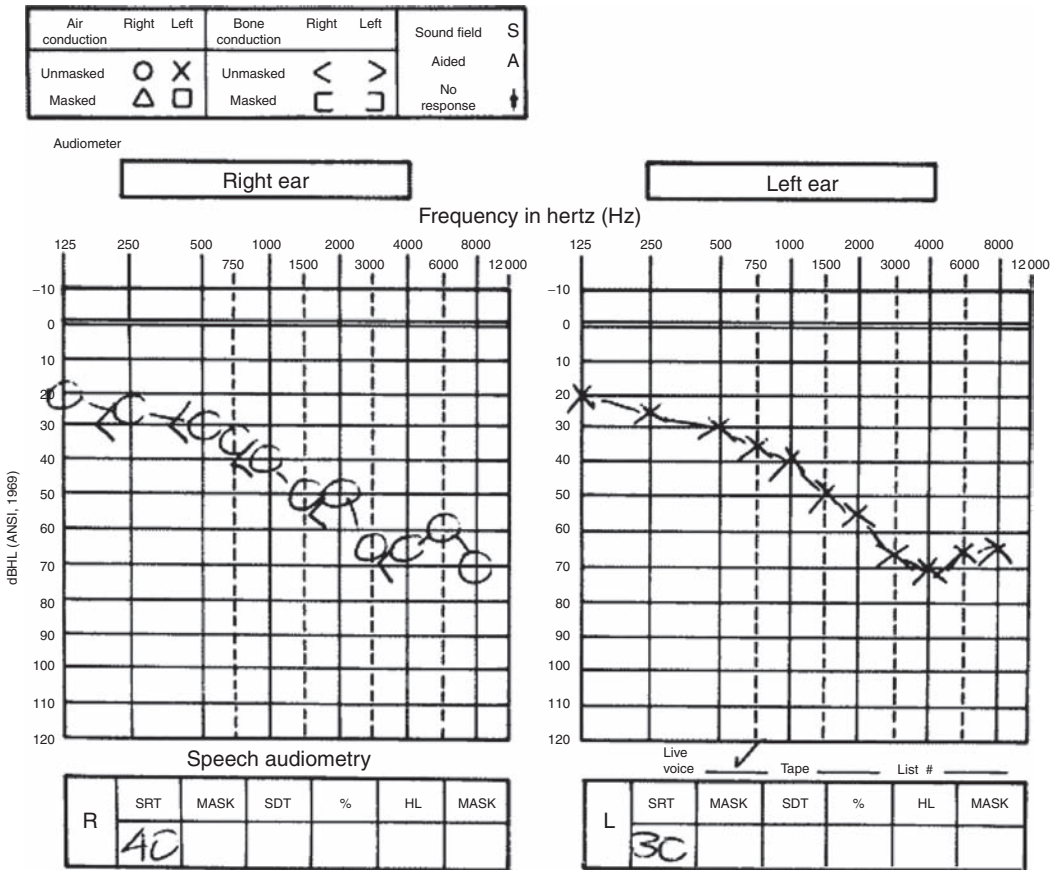

Figure 1 Audiogram showing moderate, symmetrical, sensorineural hearing loss in an octogenarian subject of the NIDCD-sponsored 'Old Time Ears Project' (a functional and histopathological study of presbycusis by CI Berlin and JB Nadol). Speech reception thresholds (SRT) for spondees are 40 and 30 dB. In addition to presbycusis, the hearing loss suggests the lasting effects of numerous experimental exposures some 50 years ago to intense tones and noise and possibly those of still earlier quinine treatment for malaria. Nowadays the subject finds binaural hearing aids of help in the theatre and lecture room, but not essential for ordinary conversation. Published with permission.

syphilis), or head injury with temporal bone fracture. Other causes include ototoxic medications, in particular the aminoglycosidic antibiotics such as neomycin or gentamicin (especially when used with a loop diuretic), the antitumor agent cisplatin, the antimalarial quinine, or even aspirin in large doses. Occupational or recreational exposure to high levels of noise sooner or later produces hearing loss, not least among today's youth. In adults, otosclerosis and Meniere's disease can be causes of impaired hearing, with the aging process itself the most common cause of all. In presbyacusis, which may have a genetic, an environmental, or a temporal basis, both ears tend to be equally affected (Figure 1). In Meniere's disease and sudden deafness, the loss is often confined to one ear. In Meniere's there may also be evidence of an autoimmune condition. In sudden deafness a viral infection may be responsible, or possibly occlusion of a small artery supplying the membranous labyrinth. Infection with cytomegalovirus can also be a cause of adult deafness, especially in AIDS.

Sites and Types of Abnormality

The site of hearing impairment is usually peripheral rather than central, that is, in the outer, middle, or inner ear rather than in the brain stem or higher centers. If the cochlea is normal but sounds are prevented from reaching it by malformation or blockage of the external canal or injury to the drum membrane or the chain of ossicles, the hearing loss is said to be of the conductive type. Thresholds for airborne sounds tend to be reduced more or less equally at all frequencies, but the loss is not complete. Hearing by bone conduction, that is, with a vibrator held against the mastoid process, the forehead, or the teeth, may be normal. For such a patient, a hearing aid to amplify sounds is often of great help. Surgical repairs, as in Wullstein's and Zöllner's types of tympanoplasty, can restore sound conduction by the ossicular chain and thus give significant hearing improvement. In otosclerosis (more appropriately called otospongiosis, as

in French usage), the impairment is most often of the conductive type. Sounds are prevented from reaching the cochlea because the stapes footplate is unable to vibrate, being sealed in the oval window by deposits of newly formed, abnormal bone. Fortunately, an effective surgical treatment is available. In this stapedectomy, the fixed and immobile stapes are removed from the oval window and the opening into the vestibule is covered with a membrane consisting of vein, fascia, or other suitable material. As a substitute for the stapes, a short length of wire or plastic is attached to the long process of the incus to make contact with the new membrane. Like tympanoplasty, stapedectomy has proved to be a dramatic advance in surgery for the betterment of hearing in patients with conductive hearing loss.

Sensorineural Impairment

In sensorineural hearing impairment, there are usually pathological changes in both the cochlea and its nerve, but the middle ear may be intact. Most often, extensive loss of the sensory hair cells and supporting cells of the organ of Corti has occurred, with secondary degeneration of at least some of the cochlear nerve fibers and spiral ganglion cells. This most prevalent type of hearing loss occurs with aging and after injury by intense noise or ototoxic medications. The changes tend to occur first in the basal turn of the cochlea, which responds to the higher frequencies of sound, but they may spread to involve the upper turns and thus affect hearing for lower frequencies as well. A slow progression of this type of change is commonly observed in presbyacusis, the hearing loss associated with aging and earlier in life with certain genetic disorders. During treatment with an ototoxic antibiotic, especially in combination with a loop diuretic, a similar but rapidly progressive loss of hearing can occur. The sensory epithelium of the cochlea may also degenerate in the wake of viral diseases such as mumps; in the late stage of Meniere's disease; in sudden deafness; and in the destructive capsular form of otosclerosis, which is not limited to the oval window and stapes but may extend so far as to violate the periosteal lining of the bony labyrinth and injure Corti's organ as well (Figures 2 and 3). Because the resulting impairment is both conductive and sensorineural in character, it is called mixed. In man and other mammals, sensorineural loss is permanent because the hair cells do not regenerate. Ongoing studies in several laboratories have shown that in birds the sensory cells can reappear after noise or ototoxic injury. There is a hope that gene transfer or stem cell treatment may eventually be devised to encourage human hair cells to regenerate and replace their losses after injury. Because sensorineural change tends to diminish hearing for the higher frequencies, presbyacusic and other patients with this form of impairment have difficulty in understanding sibilant and fricative speech sounds and in distinguishing among unvoiced stop consonants. They are at a particular disadvantage when trying to converse in noisy surroundings and in groups. A hearing aid may be helpful in a church, theater, or lecture hall, but the results are often disappointing because of the user's inability to resolve the spectral details of speech sounds, especially in the presence of noise, despite amplification.

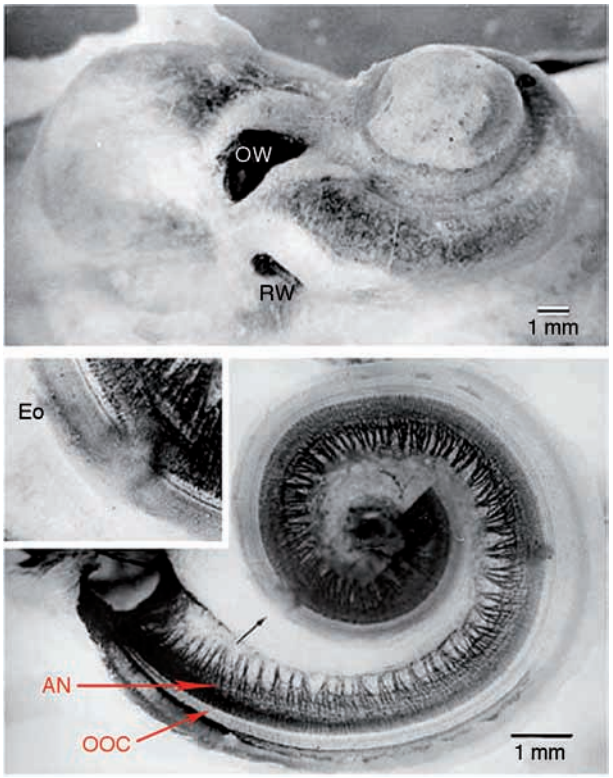

Figure 2 A case of deafness in a 73-year-old woman with capsular otosclerosis who had also received ototoxic aminoglycosides for sepsis following surgery. Top, Otic capsule, right ear, thinned by drilling and with stapes removed, showing otosclerotic focus surrounding both oval window (OW) and round window (RW). Bottom, Basal turn of cochlea. The organ of Corti (OOC) is clearly absent from most of the basal half-turn. The auditory-nerve fibers (AN) coursing across the osseous spiral lamina.were stained with osmium; their density is less than normal in most of the basal half-turn. At black arrow, Endosteum (Eo) is thickened and connective tissue has grown onto the basilar membrane as a result of the otosclerosis. From Johnsson LG, Hawkins JE Jr. et al. (1982) Cochlear and otoconial abnormalities in capsular otosclerosis with hydrops. *Annals of Otology, Rhinology and Laryngology* 91(supplement 97): 3–15.

Sensorineural impairment of hearing may be accompanied by the sometimes distressing symptom called tinnitus, a subjective ringing in the ears that may also take the form of a buzzing or roaring noise.

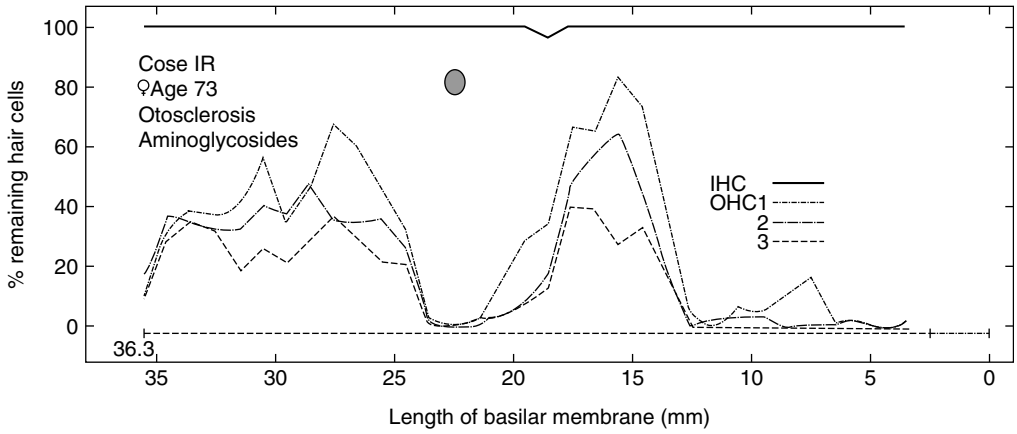

Figure 3 Corresponding cytocochleogram of the same ear as **Figure 2**. The *x*-axis plots the length of the basilar membrane from the cochlea's basal end (0 mm) to the apex (36.3 mm). There is a loss of outer hair cells from the basal turn (below 13 mm), attributable to ototoxic agents, and from the second turn (18–25 mm) in the region of the thickened endosteum, attributable to the capsular otosclerosis. Inner hair cells are present except for the most basal 3 mm of Corti's organ, where nerve fibers have also degenerated. Reprinted from Johnsson LG, Hawkins JE Jr., et al. (1982) Cochlear and otoconial abnormalities in capsular otosclerosis with hydrops. *Annals of Otology, Rhinology and Laryngology* 91(supplement 97): 3–15, with permission.

Tinnitus often accompanies attacks of Meniere's disease, in which the acuity of hearing tends to vary (fluctuant hearing loss). It also occurs during quinine treatment for malaria and salicylate treatment for arthritis. Furthermore, tinnitus may also be present after injury by noise, and unless the impairment is symmetrical, sounds may produce an unpleasant sensation of dissimilar pitch in the two ears (diplacusis). Unilateral tinnitus when hearing thresholds are near normal may indicate a problem with the auditory nerve. Such a retrocochlear lesion may be produced by a tumor such as an acoustic neuroma, which can interrupt conduction in the auditory nerve. A final characteristic symptom of sensorineural hearing loss is an unusually rapid increase in the perceived loudness of sound as its intensity level is increased: a phenomenon known as the recruitment of loudness.

Animal Models for Deafness

Animal models are very important tools for the study of normal hearing and disease processes. Genetic forms of deafness and their underlying otopathology have historically been studied in certain types of domestic and laboratory animals that are either born deaf or become deaf at an early age. Among these are blue-eyed white cats; waltzer, shaker, and other strains of mice; Dalmatian dogs; and Hedlund mink. The changes seen in the inner ear can resemble closely those occurring in certain types of congenitally deaf patients. Molecular biological techniques can be used to knock out (or overexpress) a protein in mouse models. The resulting loss of function can lead to important knowledge of the role of that protein in the normal ear. Noise-induced and ototoxic hearing losses have been studied experimentally in a variety of species to understand why humans who are exposed to such conditions lose their hearing. Presbyacusic changes in the cochlea vary with the species examined. In the dog they more nearly resemble those found in aging human ears, whereas those seen in the largely vegetarian rhesus monkey tend to be milder and less extensive.

The Cochlear Prosthesis

In deafened patients, even with extensive degeneration or absence of cochlear hair cells, it is often possible to restore a significant degree of useful hearing by surgically implanting a cochlear prosthesis. This device consists of a microphone, a speech processor, and an array of up to 24 tiny metal electrodes inserted through the round window of the cochlea into scala tympani. These electrodes stimulate the remaining nerve fibers, bypassing the missing hair cells. Most patients who have received an implant have emerged from a world of silence. Many are able to carry on a normal conversation. Lip-reading ability is usually enhanced. Over 60 000 implants have been placed over the past 20 years. In the past, the prosthesis scored its greatest successes in patients who had become deaf after learning to speak. Now, striking results are being obtained even in congenitally deaf infants and children, provided they receive the implant during the optimal period for language learning. Many such operations are now performed around the age of 12 months. In some patients, a cochlear implant will not work because there is no functioning auditory nerve. In these patients, an auditory brain stem implant (ABI) is placed in the

cochlear nucleus in the brain stem. ABI users have not generally had speech perception as good as that possible with a cochlear implant One challenge of the future will be to improve such devices to allow speech perception for all forms of deafness.

See also: Auditory System: Central Pathways; Cochlear Prosthesis; Hair Cell Regeneration; Hair Cells: Sensory Transduction; Otoacoustic Emissions; Tinnitus.

Further Reading

Ballantyne J, Martin MC, and Martin A (eds.) (1993) *Deafness,* 5th edn. London: Whurr.

Chen CY, Halpin C, and Rauch SD (2003) Oral steroid treatment of sudden sensorineural hearing loss: A ten year retrospective analysis. *Otology and Neurotology* 24: 728–733.

Cheng AG, Cunningham LL, and Rubel EW (2005) Mechanisms of hair cell death and protection. *Current Opinion in Otolaryngology and Head and Neck Surgery* 13: 343–348.

Colletti V and Shannon RV (2005) Open set speech perception with auditory brainstem implant? *Laryngoscope* 115: 1974–1978.

Dennis JM and Neely JG (1991) Basic hearing tests. *Otolaryngology Clinics of North America* 24: 253–276.

Forge A and Schacht J (2000) Aminoglycoside antibiotics. *Audiology and Neurotology* 5: 3–22.

Goldfarb A and Avraham KB (2002) Genetics of deafness: Recent advances and clinical implications. *Journal of Basic and Clinical Physiology and Pharmacology* 13: 75–88.

Hawkins JE and Johnsson LG (1985) Otopathological changes associated with presbyacusis. *Semin Hearing* 6: 115–133.

Johnsson LG, Hawkins JE Jr., Rouse RC, and Linthicum FH Jr. (1982) Cochlear and otoconial abnormalities in capsular otosclerosis with hydrops. *Annals of Otology, Rhinology and Laryngology* 91(supplement 97): 3–15.

Kawamoto K, Ishimoto S, Minoda R, Brough DE, and Raphael Y (2003) Math1 gene transfer generates new cochlear hair cells in mature guinea pigs *in vivo*. *Journal of Neuroscience* 23: 4395–4400.

Linthicum FH (1993) Histopathology of otosclerosis. *Otolaryngology Clinics of North America* 26: 335–352.

Morton CC and Nance WE (2006) Newborn hearing screening – A silent revolution. *New England Journal of Medicine* 354: 2151–2164.

National Institutes of Health (2000) *Healthy People 2010, Vol. II, 28: Vision and Hearing*. http://www.healthypeople.gov/document/html/volume2/28vision.htm (accessed Jul. 2008).

Niparko JK, Kirk KI, Mellon NK, Robbins AM, Tucci DL, and Wilson DL (eds.) (2000) *Cochlear Implants. Principles and Practices*. Philadelphia: Lippincott Williams & Wilkins.

Ohlemiller KK (2006) Contributions of mouse models to understanding of age- and noise-related hearing loss. *Brain Research* 1091: 89–102.

Pickles JO (2004) Mutation in mitochondrial DNA as a cause of presbyacusis. *Audiology and Neurotology* 9: 23–33.

Schuknecht HF (1993) *Pathology of the Ear,* 2nd edn. Philadelphia: Lea and Febiger.

Talaska AE and Schacht J (2007) Mechanisms of noise damage to the cochlea. *Journal of Audiological Medicine* 5: 3–9.

Whitehead ML, Lonsbury-Martin BL, Martin GK, and McCoy MJ (1996) Otoacoustic emissions: Animal models and clinical observations. In: Van De Water T, Popper AN, and Fay RR (eds.) *Clinical Aspects of Hearing*, pp. 199–257. New York: Springer.

Willott JF (2007) Factors affecting hearing in mice, rats, and other laboratory animals. *Journal of American Association for Laboratory Animal Science* 46: 23–27.

World Health Organization (2006) Deafness and hearing impairment. Fact sheet N°300. http://www.who.int/mediacentre/factsheets/fs300/ (accessed Jul. 2008).

Decision-Making and Neuroeconomics

A Rustichini, University of Minnesota, Minneapolis, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Decision Theory in Economics

Decision theory as developed in economics is an axiomatic analysis. The theory proceeds by first defining a set of choices that a subject (the decision maker, DM) faces. A choice is a finite set of options that are offered to the DM; a decision is the selection of one of these options. These are the observed data: it is in principle possible to collect them by explicitly asking a real DM to choose among the options, under the condition that the object selected is actually delivered to him. No speech acts (that is, statements like "I prefer this option to this other option") are necessary or allowed among the potential observations.

The method and the main results of the theory are best illustrated in a simple and concrete example of choice environment, choice under risk. In this environment the options are lotteries. A common lottery ticket provides an example: a winning number is drawn at random, and with such a ticket, a person is entitled to a payment if the winning number is the one he has, and he receives no payment otherwise. In general, a lottery is a contract specifying a set of outcomes (the payments made to the subject in our example) and a probability for each of these outcomes. The probability is typically specified in advance and precisely known to the subject.

For example, a lottery with two outcomes can be formally described with a pair $(x, p, y, 1-p)$, to be interpreted as this lottery gives the outcome x with probability p, and the outcome y with probability $1-p$. With this notation, a lottery with monetary payments (\$10, 1/2, \$0, 1/2) gives a 50/50 chance of a payment of 10 dollars and nothing otherwise. Lotteries do not need to award a monetary amount: the theory applies also to situations where x is an expensive jewel and y is a night in jail (if, for example, you are contemplating robbing a jewelry store and considering the risk of getting caught in the act). For simplicity of exposition we confine ourselves mostly to lotteries with monetary amounts in the rest of the article.

Revealed Preferences

The only feature that we can observe of our subject is the decisions he makes. We do not, of course, observe his preferences directly. However, we may interpret his choices as a 'revelation' he makes of his preferences: if when he is presented with a choice between lottery L_1 and L_2 he chooses L_1, we may say that he reveals he prefers L_1 to L_2. It is in this sense and in this sense only that we can say that the DM prefers something. The two descriptions of his behavior, one with the language of decisions and the other with that of preferences, are by the definition we adopt perfectly equivalent. Since the language of preferences seems more intuitive, it is the one used typically by decision theory, and is the one used here. But how do we describe the behavior or preferences of our subject?

Axioms

Even with simple lotteries with two monetary outcomes, by varying the amounts and the probabilities we can obtain an infinite set of possible lotteries, and by taking pairs of these lotteries we can obtain infinitely many choices. So to completely describe the behavior of a subject we should in principle list the infinite set of decisions he makes. The theory considers instead subjects whose decisions can be described by a short list of simple principles or axioms.

The first axiom requires that the preferences are complete: for every choice between the two lotteries L_1 and L_2, either L_1 is preferred to L_2 or L_2 is preferred to L_1. The occurrence of both possibilities is not excluded: in this case, the subject is indifferent to the two lotteries. When the subject prefers L_1 to L_2 but does not prefer L_2 to L_1, then we say that he strictly prefers L_1 to L_2. The second axiom requires the preferences to be transitive: if the DM prefers L_1 to L_2 and L_2 to L_3, then he prefers L_1 to L_3.

The next two axioms are also simple, but have more of a technical nature. Suppose we have two lotteries, $L_1 = (x, p, y, 1-p)$ and $L_2 = (z, q, w, 1-q)$. Take any probability r, a number between 0 and 1. Imagine the following contract. We will run a random device, with two outcomes, *Black* and *White*, the first with probability r. If *Black* is drawn, then you will play according to the lottery L_1; if *White* comes out, you will play according to the lottery L_2. This new contract is a lottery as well. If you do not care about how you get the amounts of money, then this is the lottery with four outcomes described as $(x, rp, y, r(1-p), z, (1-r)q, w, (1-r)(1-q))$. We write this new lottery as $rL_1 + (1-r)L_2$. The two axioms that we mentioned are expressed using this new concept.

The next axiom, third in our list, requires that if you strictly prefer L_1 to L_2, then for some number r, you strictly prefer $rL_1 + (1-r)L_2$ to L_2. This seems

reasonable: when r is close to 1, the composite lottery is very close to L_1, so you should strictly prefer it to L_2 just like you strictly prefer L_1.

Finally, suppose that you strictly prefer L_1 to L_2. Then for any lottery L_3, you also strictly prefer $rL_1 + (1-r)L_3$ to $rL_2 + (1-r)L_3$. Again, this seems reasonable. When, in the description we gave above, *White* is drawn, then you get in both cases L_3; when *Black* is drawn, you get in the first case L_1 and in the second L_2. Overall, you should prefer the first lottery, $rL_1 + (1-r)L_3$.

Representation of Preferences

A fundamental result in decision theory (due to von Neumann and Morgenstern (vNM)) is that subjects having preferences that satisfy these four axioms behave as if they had a simple numerical representation of their preferences. A numerical representation of a preference is a function that associates to a lottery a single number, called the utility of the lottery, that we can write as $U(L)$. This function is called a representation of the preferences if whenever L_1 is preferred to L_2, then the utility of L_1 is larger than the utility of L_2, that is, $U(L_1) > U(L_2)$.

It should be clear that the existence of such a function restricts the preferences of the subject. For example, suppose that L_1 is preferred to L_2 and L_2 to L_3. If such a function exists, then $U(L_1) > U(L_2)$ and $U(L_2) > U(L_3)$. It follows that $U(L_1) > U(L_3)$, and since the utility is a representation of preferences, L_1 is preferred to L_3. That is, preferences with a utility representation must be transitive.

The vNM theorem also states that the preference order satisfies the four preceding axioms if and only if there exists a utility function with a very simple form that represents the preference. The form is simple because it is the expectation of the utility of each outcome, according to some function u of outcomes. For example, the expected utility of the lottery $L = (x, p, y, 1-p)$ is

$$U(L) = pu(x) + (1-p)u(y). \qquad [1]$$

Cardinal and Ordinal Utilities

For neuroeconomics, and any research program that tries to determine how decisions are implemented, the utility function is probably the most interesting object. This function ties observed behavior with a simple one-dimensional quantity, the utility of the option, and predicts that the decision between two options is taken by selecting the option with the highest utility. If one is interested in determining the neural correspondents of the objects we have introduced, he must first know whether these objects are unique. For example, we may formulate the hypothesis that the decision is taken depending on some statistics of the firing rate of a group of neurons associated with each of the options. We may also consider that this firing rate is proportional to the utility we determine from observed choice behavior. Then we need to know whether this utility is uniquely determined. This introduces us to a fundamental distinction in decision theory, that between cardinal and ordinal representation.

An ordinal representation of a preference is any utility function such that $U(L_1) > U(L_2)$ if and only if L_1 is strictly preferred to L_2. There are clearly many such functions. For example, if M is any increasing function, then $M(U(L_1)) > M(U(L_2))$ if and only if L_1 is strictly preferred to L_2. So we say that an ordinal representation is unique only up to increasing (or monotonic) transformation, like the one we have used from U to $M(U)$.

Consider now the function u is eqn [1], and take two numbers $a > 0$ and b. Replace the u function in (1) with the new function v defined for any value z by $v(z) = au(z) + b$. If we replace the u in eqn [1], we obtain a new function on lotteries which also represents preferences and has the form of expected utility. Since these transformations leave the observed choices and preferences unchanged, the u in [1] is not unique. However, these are the only possible transformations we can apply. A second remarkable part of the vNM theorem is that if two functions u and v represent the preferences of a subject as expected utility (that is, as in [1]), then it must be that $v(z) = au(z) + b$ for some positive number a and some number b. In this case, the two functions are said to be linear transformations of each other, and representations like these are said to be cardinal representations. A different but equivalent way of saying this is that if we consider functions on a range of monetary prizes between a minimum of 0, say, and a maximum value M, and we agree to normalize the utility function u to $u(0) = 0$ and $u(M) = 1$, then there is a unique such function that, once substituted in eqn [1], represents the preferences of the DM.

However, the observed decision between two choices is determined by the function U, and this is unique only up to monotonic transformations. So even if we agree to normalize $U(0) = 0$ and $U(M) = 1$, still there are infinitely many such Us. Can we do better than this? We can, if we agree to extend the set of observed data to include errors and time in the decision process.

Stochastic Choice

The next step in our survey brings us closer to the models and results currently applied in neuroscience: the theory of stochastic choice.

To illustrate and motivate this new point of view, we begin with an interesting finding, discovered in the 1940s by an Iowa researcher, D Cartwright. He asked subjects to pick one of two alternatives. In keeping with our previous exposition, one may think of these alternatives as lotteries, although the original experiments were based on cognitive tasks, such as discriminating among geometric shapes or angles of different width. By changing the parameter appropriately, the experimenter could make the choice more or less difficult. For example, setting the width of two angles to be more similar would make choosing the wider angle a more difficult task. Also, by asking the subject to make the same choice repeatedly, at some time distance, he could test the frequency of the choice of one or the other of the alternatives in different decision problems. He could then construct what we can call the empirical random choice: the frequency of choice of each of the options from a given set of alternatives.

He had, however, also measured the response time for each choice (that is, the time that passed between the moment the choice was presented and the time the choice was made by the subject), and he then plotted the average response time for each decision problem against the minimum frequency of any of the two choices in that same problem. The key finding was that the longest response time was observed when the minimum frequency was approaching 50%. In other words, those problems in which the subject was more likely to select, over time, both options were also those in which he was taking more time to decide.

The result is robust and has been replicated since in many different environments. A related test is the measure of the response time measured as function of the distance between the two options, when a measure of distance is natural, for example, the distance between two numbers when the subject has to pick the largest. A robust finding is that the response time decreases with the distance (typically the log of the time is linear in the log of the distance). This is the 'symbolic distance' effect.

The finding of Cartwright suggests a model of decision in which two opposing forces push in the direction of each of the options. When the difference between these two forces is large, the decision is frequently in favor of the favored option, and the decision is taken quickly. When they are the same, the frequency of choice of the two options becomes closer, and the response time becomes longer.

For our purposes of outlining a theory of the decision process when the decision considers economic choices, it is important to note that for economic choices the same result holds. Suppose we determine the utility of a subject from the observed choices, that is, the quantity $U(L)$ for every lottery L. We can now measure the distance between the utility of any two lotteries in a choice and conjecture that the analog of the Cartwright results hold in this situation: the closer the two options in utility, the longer the time to decide, and the higher the minimum probability of choosing any of the two. This conjecture has been confirmed in several studies. There is one last hurdle: we have just seen that the U in [1] is not unique; in fact, even after normalization we have infinitely many such functions. So how can we measure in a meaningful way the distance in utility between two options?

Economic Theories of Stochastic Choice

The experimental evidence reviewed in the previous section suggests that when faced with a choice between two options many times, the subject may not always choose the same option in each instance. This is the opposite of what the utility theory we have reviewed so far predicts: in that theory, if the utility of one of the two is larger, it should always be chosen. Still, a large body of experimental and empirical evidence suggests that at best we can determine the frequency according to which one option is chosen, and that this frequency is not necessarily (in fact, is typically not) concentrated on one option. The key idea of the stochastic theory of choice is that the relative frequency of the choice of one option over the other gives a way to get a better measure of the utility of the two options.

The method of analysis is again axiomatic. Let $p(L_1, L_2)$ denote the frequency according to which lottery L_1 is chosen out of the set $\{L_1, L_2\}$, that we called the empirical random rule. Thus, it gives us evidence on whether L_1 is preferred to L_2: for example, this is plausible if that frequency is larger than 1/2. But it tells us more. If we compare the same frequency of choices out of two other lotteries $\{L_3, L_4\}$, and we observe that $p(L_1, L_2) > p(L_3, L_4)$, then we may say that the difference in utility between L_1 and L_2 is larger than the difference between L_3 and L_4. An appropriate set of axioms on the new evidence (the frequency of choice) guarantees that there is a utility

function (that we denote V to distinguish it from the very different object U) such that

$$p(L_1, L_2) > p(L_3, L_4)$$

if and only if

$$V(L_1) - V(L_2) > V(L_3) - V(L_4) \qquad [2]$$

This function is a cardinal object: that is, if a different function W also satisfies [2], for the same p, then W is a linear monotonic transformation of V. That is, there are two numbers $a > 0$ and b such that $W = aV + b$. Another way of stating this delivers us the unique object that we were looking for: if we normalize V to $V(0) = 0$ and $V(M) = 1$, then this function is unique. This is the object we were looking for: a unique summary, entirely constructed from behavioral observations, of the value that the DM attaches to different options.

In the plan of determining the neural basis of decision, we have now two final steps. First, we have to produce a reasonable model of the decision process that produces a stochastic choice with the required coherence expressed formally by the axioms underlying the representation in eqn [2]. Second, we have to specify and test the neural basis implementing this process. Let us begin with the first.

The Random Walk Model of Decision

The random walk model is a model of the decision process, originally formulated by Ratcliff. The process was originally thought of as an explanation of memory retrieval. In this task, a subject is presented with an object and is asked to decide whether it is the same he saw some period earlier. The model describes the two objects (one in front of the subject, the other in his memory) as a vector of characteristics. The decision is taken by marking, for each characteristic, whether it is the same in the two objects (in this case mark a score of 1) or not (and then mark -1). When the sum of the evidence in favor and against the conclusion that the two objects are the same reaches a critical threshold, then the decision is taken.

The process provides the basis for a general model of decision. Let us consider for simplicity the case in which the DM has to choose between two alternatives. The essential elements of the model are a space of evidence, a stochastic process on this space, and two barriers. In the memory retrieval example, the space of evidence is the set of integers, where a number is the number of agreements minus the number of disagreements. The process on the space is a random walk: as the DM proceeds to consider the next characteristic in the list, he can add or subtract one unit to the current evaluation state, depending on whether he finds an agreement or a disagreement. Finally, the process stops and the decision is taken when a barrier (high or low) is reached. The decision process has been reduced to a random walk on a space and a stopping time described by two barriers. At the stopping time, the decision corresponding to the barrier is taken.

The Predictions of the Model

The model has several parameters: first, those describing the process, for example, the mean and the variance of the random walk. The other parameters are the barriers. The model yields two observed variables: the probability that one of the two decisions is taken, and the time needed to reach the decision. Of course, an important qualitative aspect of this model of the decision process is that choice is not deterministic: when the subject is presented with the same choice, he may sometimes choose one option, and sometimes the other. The connection between the theory of stochastic choice in economics and the random walk model is clear: the probability of choice of each option studied in the model of stochastic choice is the same object induced by the random walk model when the latter hits the barrier.

The model gives sharp predictions on the two variables. For example, if the drift in favor of one of the two options is stronger, then the probability that that option is chosen increases. Also, when the difference in drift between the two choice is small, then the time to take a decision increases. The model provides a good fit with the observed data on choice and on response time in a variety of situations. The question is: do we have any evidence that suggests a neural basis for this?

Decision Theory in Neuroscience

Intense research on the neural foundation of the random walk model of decision has been developed in the past years. To illustrate the method and the findings, we begin again with a classical experiment.

In the experiment, the subject (for example, a rhesus monkey) observes a random movement of dots. A fraction of the dots is moving to the left and a fraction to the right. The monkey has to decide whether a larger fraction of dots is moving to the right or to the left, being instructed to do this after intensive training. If it makes the right guess, it is compensated with a squirt of juice. Single neuron recordings of neurons show that the process of deciding on the direction is the outcome of the following process. Some neurons are associated with the movement to the left, and others to the right. The overall firing rate of the 'left' and 'right' neurons is of course roughly proportional to the number of dots moving in the two directions.

The decision is taken when the difference between the cumulative firing in favor of one of the two alternatives is larger than a critical threshold.

We suggest that the mental operation that is performed when the subject has to choose between two economically valuable options consists of two steps. First, the individual has to associate a utility to each of the two options. Second, he has to decide which of these two computed quantities is larger. This second step is a simple comparison of quantities. The first is completely new and is specific to economic analysis. Note two important features of this model: first, even if the DM assigns (somewhere in his brain) a strictly larger utility to one of the two options, he still does not choose for sure that option; he has only a larger probability of doing so. Second, the DM has a single utility or preference order over outcomes. The choice outcome is not deterministic, because the process from utility evaluation to choice is random.

What is the evidence supporting this view? Let us begin from the step involving comparison of quantities. There are experiments involving comparison of numbers, run with human subjects, that confirm the basic finding that the response time is decreasing with the distance between the two quantities that are being compared. For example, if subjects have to decide whether a number is larger or smaller than a reference number, then the response time is approximately decreasing exponentially with the distance between the two numbers. So there is experimental evidence that suggests that the operation of comparing quantities follows a process that is close to that described by the random walk model.

The last missing element is this: do we have evidence that there are areas of the brain where neurons fire in proportion to the utility of the two options?

The Computation of the Utility

Building on previous results, we examine the neural basis of the computation of the utility. In this experiment, a monkey is offered the choice between two quantities of different food or juices, for example, three units of apple or one unit of raisin.

By varying the quantities offered of juice of each type, the experimenter can reconstruct, from 'revealed preferences', the utility function of the monkey. This function can be taken to be, for the time being, an artificial construct of the theorist observing the behavior. The choices made by the subjects have the typical property of the random choice. For example, between any amount less than or equal to 2 units of apple-tea and 1 unit of raisin, the monkey always chose the apple juice. With 3 units of tea and 1 of apple, the frequency of choice was 50/50 between the two. With 4 units of tea, the monkey always went for the tea. This is the revealed preference evidence.

At the same time, experimenters can collect single neuron recordings from areas that are known to be active in evaluation of rewards (for example, area 13 of the orbito-frontal cortex). They could then plot the average firing rate over several trials (on the y-axis) against the estimated utility of the option that was eventually chosen on the x-axis. One then obtains a clear, monotonic relationship between the two quantities.

A Synthesis

We have now the necessary elements for an attempt to provide a synthesis of the two approaches, one based on economic theory and the other on neuroscience.

Consider a subject who has to choose between two lotteries. When a subject considers each of them, he can assign to it an estimate of the expected utility of each option. This estimate is likely to be noisy. When he has to choose between the two lotteries, he can simply compare the (possibly noisy) estimate of the two utilities, so the choice between the two lotteries is now determined by the comparison of these two values. At this stage, the choice is reduced to the task of comparing of two numerical values, just as the task that the random walk model analyzes.

In summary, this model views the decision process as the result of two components. The first reduces the complex information describing two economic options to a numerical value, the utility of each option; the second performs the comparison between these two quantities, and determines, possibly with an error, the larger of the two. The comparison in this second step is well described by a random walk model of decision.

Risky and Ambiguous Choice

In this final section we illustrate how this theory can provide a simple explanation of the choice behavior of subjects facing a special type of uncertainty: ambiguity. We begin with a precise definition of this idea. Our simple examples of lotteries were meant to illustrate the uncertainty that individuals face in real economic life: any choice is typically between options that promise not certain, but rather random, outcomes. In our examples, the probability over the outcomes was clearly defined: the description of a simple random device (for example, a toss of a coin) would not leave any doubt. But in common real choices, the outcomes do not come with clearly defined probabilities attached. For example, a car insurance is a contract stipulating payments depending on events

(car accidents) for which we do not know the exact probability of their occurrence, although we can formulate a subjective estimate of it, or perhaps gather information from actuarial data. When the situation defining the probability of an option is not familiar, the ambiguity of the option may be particularly troubling. For example, an investor considering two different assests may look very different at those of a US company compared to those of a company in Malaysia: the confidence that he has on his forecasts may be fundamentally different in the two cases.

Ambiguity can be produced in experiments by reducing the information available to the subject on the likelihood of the outcomes. For example, he may be told that a lottery L will pay an amount x if a white ball is drawn from an urn containing 90 balls and y if a black one is drawn. He is not told, however, the proportion of black and white balls.

The barrier in the random walk model of choice that we have considered can be interpreted in the case of economic choices as expressing the level of satisfactory evidence which is considered necessary to make a decision. If this level is independent of the quality of the information, then the prediction on response time is clear: when the information is inferior, it would require a longer time to reach the same level of satisfactory evidence. We argue that the information-processing device (the brain) should also consider that inferior information gives a smaller incentive to wait and see. In the random walk model of decision, the trade-off between waiting and the accuracy of the information translates into different distances of the barrier from the initial point of the decision process: a barrier which is farther will require, all else being equal, more time and will allow more accurate evaluation. How should the barrier be set? This depends on the quality of information that the *DM* has. Consider, for example, the case in which the signal observed provides no information at all. Then clearly it is optimal to decide immediately, since observing the signal is not going to improve the quality of the decision. In this case the barrier is very close to the initial starting point of the process. Let us apply this to choices involving risky or ambiguous lotteries. For example, let us compare the situation in which a subject has to choose between a risky lottery and a certain amount to be paid for sure with a similar choice in which the lottery is ambiguous.

In these two types of choice, the quality of the information is different. When the quality is inferior (as in the case of ambiguous choices), the barrier should be closer to the initial point than in the risky choice because the evidence gathered is less useful. So a less informative signal has the effect of reducing the level of evidence considered necessary before the information processing stops and a decision is reached.

The main hypothesis we derive is that the choices made with a less informative signal (an ambiguous lottery compared to a risky one) may induce a shorter response time. Correspondingly, we should observe a weaker activation than those found during choices made with a more informative one. This is exactly what we find in several experimental analyses of choice.

See also: Decision-Making in Financial Markets; Delayed Reinforcement: Economics; Game Theory and the Economics of Animal Communication; Games in Monkeys: Neurophysiology and Motor Decision-Making; Neuroeconomics: History; Reward Decision-Making; Social Cognition.

Further Reading

Block HD and Marschak J (1960) Random orderings and stochastic theories of responses. In: Olkin J, Ghurye S, Hoeffding W, Madow W, and Mann H (eds.) *Contributions to Probability and Statistics,* Stanford: Stanford University Press.

Cartwright D and Festinger L (1943) A quantitative theory of decision. *Psychological Review* 50: 595–621.

Cartwright D (1941) The relation of the decision time to the categories of response. *American Journal of Psychology* 54: 174–196.

Davidson D and Marschak J (1959) Experimental tests of stochastic decision theory. In: West Churchman C (ed.) *Measurament Definitions and Theories,* New York: Wiley.

Debreu G (1958) Stochastic choice and cardinal utility. *Econometrica* 26(3): 440–444.

McFadden D and Richter M (1991) Revealed stochastic preferences. In: Chipman JS, McFadden D, and Richter MK (eds.) *Preferences, Uncertainty and Optimality.* Boulder: Westview Press.

Padoa-Schioppa C and Assad J (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441: 223.

Ratcliff R (1978) A theory of memory retrieval. *Psychological Review* 85: 59–108.

Schall JD (2001) Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience* 2: 33–42.

Shadlen MN and Newsome WT (1996) Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences of the United States of America* 93: 628–633.

Shadlen MN and Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology* 86: 1916–1936.

Sigman M and Dehaene S (2005) Parsing a cognitive task: A characterization of mind's bottleneck. *PLoS Biology* 3(2): e37.

Tremblay L and Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. *Nature* 398: 704–708.

Decision-Making and Vision

A K Churchland and M N Shadlen, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Vision involves more than processing the light information received through the eyes. It involves interpreting information, classifying it, drawing inferences, and making decisions about its meaning in the context of goals, plans, and attitudes. The study of visual processing dovetails with the study of cognitive function. When we approach the study of vision from the 'processing' perspective, we tend to ask questions about the mathematical transformations of low-level features, such as contrast contours, to more complex entities, such as depth, motion energy, curvature, texture, objects, and faces. When we approach the topic from the cognitive perspective, we tend to ask questions about how information leads to a behavioral response, inference, or decision. Thus, a central question in the study of both vision and cognition is how the brain processes information to select among different behavioral alternatives. In this article we exploit knowledge about the neurobiology of vision to gain insight into the neurobiology of decision making.

A decision is a commitment to a particular choice or proposition at the expense of competing alternatives. It is usually based on evidence bearing on the likelihood of the alternatives and on the costs and benefits associated with the choices. This blanket formulation applies to complex decisions, like for whom to vote or what car to buy, and to simple ones, like whether to stop or go or whether a note is sharp or flat. The latter type of decision can be studied using the tools of neuroscience, especially those developed to study perception. We think understanding these simple decisions opens a window onto higher brain function in general.

Many decisions involve contingency and deliberation. By contingency, we mean that the process can occur without necessitating a particular outcome. Contingencies defy an 'if A then B' recipe. Instead we are inclined to use the language of probability to describe the outcome, or at least to predict it: if A then probably B, but maybe C or D. Of course, sometimes decisions are based on such overwhelming evidence that one outcome is guaranteed. We might view such decisions as diminutive examples, especially if they are part of a continuum. How we classify such easy decisions is not so important, but as objects of study, they are of limited utility to the neurobiologist. This is because the decision process is hard to separate from the cause and consequence – stimulus and response in the laboratory setting.

By deliberation, we mean that process is guided by more than one piece of information that is acquired from the environment or retrieved from memory over time. Again, not all decisions contain this element, but from a purely practical point of view, if the process of deciding is drawn out over time, it is easier to study. In short, decisions that do not necessitate immediate action can be peeled away from the evidence and outcome of the process. Decisions that incorporate deliberation also necessitate a stratagem for terminating the process. Thus many decisions are really double – they are a decision about the alternatives and a decision to commit based on the current state of knowledge.

The combination of these elements – contingency, deliberation, and a rule for termination – has led to the idea that many decisions are accomplished by a mechanism that accumulates evidence until some criterion level has been achieved. The neurobiology of decision making exploits this condition because it permits exposure of neural responses associated with the decision process and dissociated at least partially from the immediate representation of evidence or the outcome of the decision, typically a motor act. This separation of motor response from sensation – both temporally and logically (via contingency) – is a basic element of higher cognition.

The Basic Accumulator Model for Decisions

In the laboratory, stimuli can be configured so that the evidence for each alternative is present but sometimes weak. These kinds of stimuli are ideally suited to accumulation of evidence. We will describe work from a number of laboratories that have created stimuli of this kind to encourage animals to accumulate evidence over time. However, we are confident that this kind of decision making is not constrained to artificial laboratory situations. As we will detail at the end, a number of ecological situations are well suited to this kind of strategy also.

The premise behind the accumulator model for decision making is that evidence accumulates to better inform the decision. This is not always the case. Which decisions benefit from accumulation of evidence? A reasonable benchmark is the improvement

in accuracy expected from an accrual of independent samples of evidence. The simplest example of when accumulation of evidence is beneficial is a process that adds a series of independent values in time. Suppose that a series of independent samples with mean μ and standard deviation σ arrive at rate kt. The accumulated signal at $t = T$ would have a mean μkT and standard deviation $\sigma\sqrt{kT}$. (Recall that the sum of two independent random numbers has mean and variance equal to the sum of the their means and variances, respectively.) Thus the signal-to-noise ratio should improve as $\sqrt{t}$.

If accumulation is truly benefiting the subject in a psychophysical experiment, we might expect the threshold stimulus intensity that is required to achieve some level of accuracy to fall by the square root of the stimulus duration. The graph in **Figure 1** shows results, which conform reasonably well to this prediction. Note that slope of threshold versus viewing duration is very close to -0.5. It is often the case that performance improves at a rate substantially less than what is predicted from the $\sqrt{t}$ relationship. That could be a sign that the samples of evidence (in time) are not independent, or that not all are used, either because there is some forgetfulness (e.g., a leak of an integrator) or that some samples are ignored (e.g., because the process has terminated before all the samples were counted). In other words, a departure from $\sqrt{t}$ does not rule out accumulation.

Formal models of the process just described are termed accumulator models. In these models, a decision variable reflects the evidence accumulated over time (**Figure 2(a)**). Models of this kind are also known

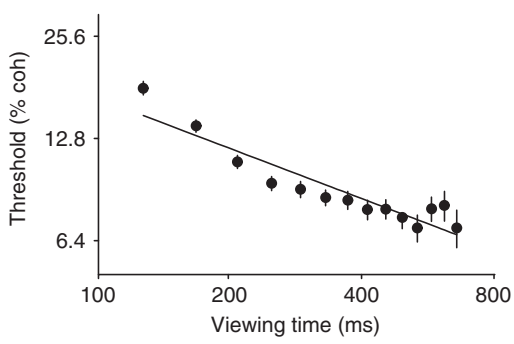

Figure 1 Accuracy improves as a function of viewing time in the motion discrimination task. Each point in the graph was obtained by measuring the monkey's accuracy as a function of motion strength for one viewing duration. From these data, we estimate a threshold: the motion strength required to reach 81% correct. The graph shows that when given extra viewing time, the monkey achieves this level of accuracy with weaker motion strength (coh, coherence). The best fitting line has a slope of –0.46 (95% confidence interval is –0.59 to –0.33). Adpted from Gold JI and Shadlen MN (2003) The influnce of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *Journal of Neuroscience* 23(2): 632–651.

as sequential sampling models because new evidence is added at each moment in time. The rate of accumulation (k) might depend on a number of factors, including the salience of the sensory stimulus being evaluated to make the decision. When a sufficient amount of evidence has been accumulated, the decision variable reaches a threshold value, known as a bound, and the decision process is terminated. The bound provides a reliable way to determine when to cease accumulating evidence and commit to a course of action. When decisions are among two alternative courses of action, the process might be modeled by two accumulators, each accumulating evidence for one alternative (**Figure 2(b)**). And if the evidence for one alternative is evidence against the other, then the race can be conceived as a single process/accumulation terminating in an upper or lower bound, for the two alternatives, respectively (**Figure 2(c)**).

This last scenario is termed a diffusion or random walk to bound model. It has special status in the field of statistical decision making because, under appropriate assumption, it offers the most efficient algorithm for deciding between two alternatives. It also played a key role in Alan Turing's scheme to break the German enigma cipher during World War II. In all cases, a higher bound requires more evidence before the decision process is terminated, while a lower bound requires less evidence. As we will describe, bound height may be somewhat flexible and slowly change depending on the demands of the decisions being made. A variety of accumulator models have been proposed on theoretical grounds or to explain different data sets. Some restrict accumulation to positive quantities, some to positive or negative values drawn from normal distributions; some contain leak and some involve strong interactions between accumulators. Though the details of the models differ, they share a common theme, which is that they attempt to describe both the content of a decision and the time it takes to reach it with a single mechanism. Further, by using a bound to terminate the decision process, accumulator models collectively describe a process that is controlled by internal signals in the brain, rather than by external sensory cues. Here, we highlight differences between kinds of accumulator models only as needed.

We focus on a number of different kinds of decisions that appear to rely on the accumulation of evidence. We begin with perceptual decisions whereby individuals are required to make judgments based on visual stimuli. Next, we describe a simpler kind of decision to act or not. Last, we outline decisions that may be well-described by the accumulation of evidence, but require some key changes to the models described elsewhere in this article.

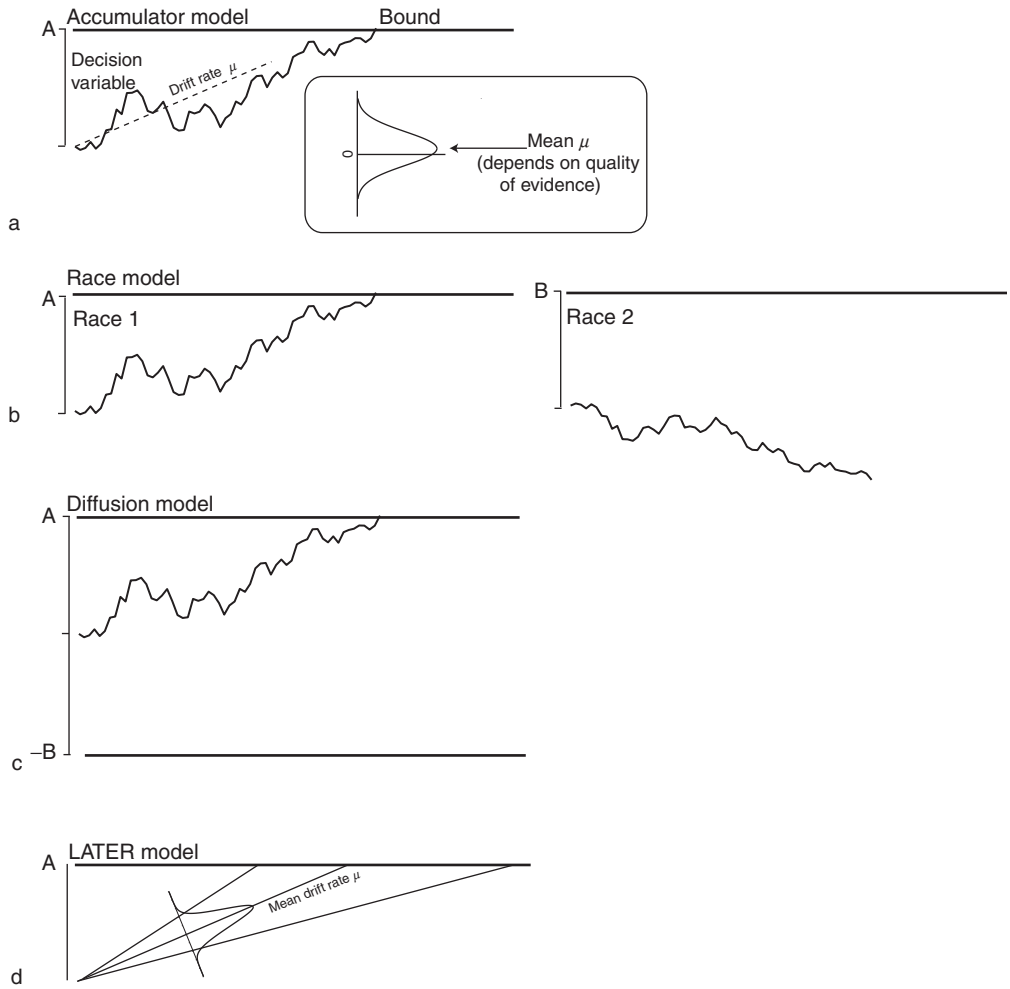

Figure 2 Accumulator models. (a) Basic accumulator model. Samples of momentary 'evidence' are accumulated into a decision variable. The process terminates when the decision variable reaches a bound. The momentary evidence is modeled as a probability distribution with mean μ and variance σ^2 (insert). The decision variable has the appearance of a random walk. The jagged trace shows one example. The average of many such traces would have a slope equal to μ and a variance equal to $n\sigma^2$, where n is the number of samples drawn in time. (b) Race model for decisions between two alternatives. Conventions are the same as in a except that there are two races, one for each possible outcome. Each decision variable is the accumulation of momentary evidence for one of the outcomes. (c) Diffusion model. In a two-choice decision, if the evidence for one choice is evidence against the other, then the two decision variables are identical except for a sign change. The accumulation can be drawn on a single graph with an upper and a lower bound. Horizontal line labeled 'A' reflects the bound for one choice; horizontal line labeled 'B' reflects the bound for the other choice. (d) The LATER (linear accumulation to threshold with ergodic rate) model for decisions to 'go' or not. The decision signal rises linearly from a starting point until the bound is reached. The rate of rise during this time varies from trial to trial and can rise slowly (thin, shallow line), at an intermediate rate (medium line), or quickly (thin, steep line). The variability in slope explains reaction time variability. The linear rise approximates the accumulation of a fixed value, which is drawn from a normal (Gaussian) distribution on each trial. The slopes are therefore normally distributed and so are the reciprocals of the reaction times.

Simple Perceptual Decisions

Behavioral Observations

Considerable behavioral evidence suggests that both humans and monkeys suffer more errors when they make decisions quickly. In a recent study by our group, humans performed a random-dot motion-discrimination task. Individuals viewed a dynamic random-dot kinematogram and decided the net direction of motion of the dots. Some of the dots moved together, in the same direction, while the remaining dots moved randomly. On each trial, the percentage of dots moving in the same direction, termed the coherence, was set to make the trial easy or more difficult. In our version of the task, individuals terminated the dot motion when they were ready with a decision by making an eye movement to one of two choice targets (Figure 3(a)). Accuracy (percentage of trials whereby the individual correctly reported the direction of motion) and speed (reaction time) were measured

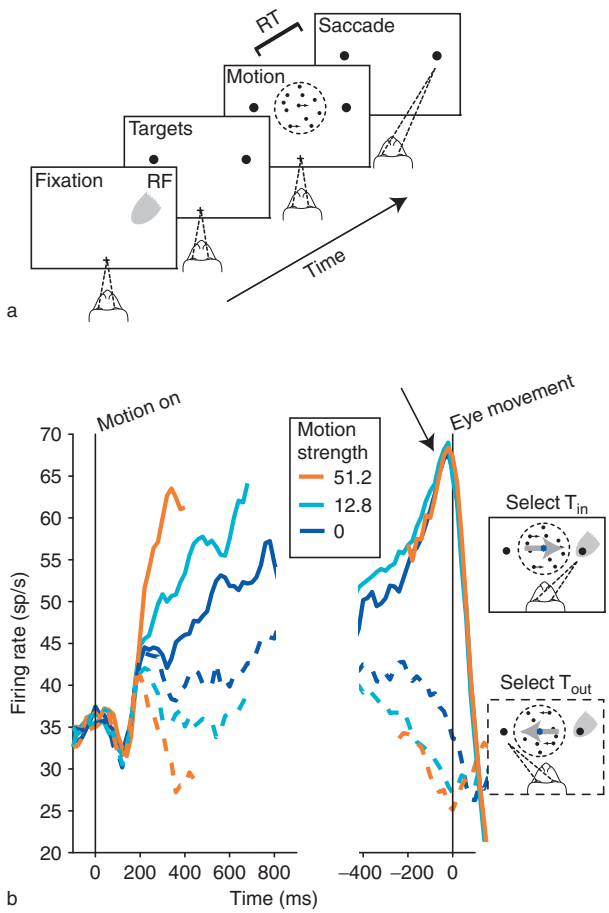

Figure 3 A neural correlate of a decision variable. (a) Reaction-time (RT) direction-discrimination task. A random-dot motion task is used to study decision making in humans and monkeys. The task begins with the appearance of a fixation point and two choice targets. After a random interval, a patch of moving dots appears. Some of the dots in the stimulus move coherently in the same direction while the remaining dots move randomly. When the individual is ready with a decision, he terminates the trial by making a saccade to one of two choice targets. When monkeys are performing the task, they are rewarded for correct responses with a drop of juice or water. (b) Average response from 54 lateral intraparietal area neurons on the direction-discrimination task. The firing rates (sp/s, spikes/second) of neurons in the lateral intraparietal area were measured while the monkey performed a RT version of the random-dot direction-discrimination task shown in (a). In each experiment, one of the choice targets was in the neuron's response field. Responses are grouped by motion strength (line color) and the choice the monkey made at the end of the trial in favor of the direction associated with the target in (T_{in}; solid) or out (T_{out}; dashed) of the neuron's response field. The responses are aligned to two events in the trial. On the left, responses are aligned to the onset of stimulus motion. Response averages in this portion of the graph are drawn to the median RT for each motion strength and exclude any activity within 100 ms of eye movement initiation. On the right, responses are aligned to initiation of the eye movement response. Response averages in this portion of the graph show the buildup and decline in activity at the end of the decision process. Only correct choices are included in these graphs for motion coherences >0%. Arrow indicates the approximate time at which responses to different coherences coalesce for T_{in} choices. From Roitman JD and Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience* 22: 9475–9489.

both when individuals were given no instructions about reaction time and when they were instructed to aim for a specified reaction time on the most difficult trials. Both individuals were able to comply with the instructions, and their reaction times changed accordingly. Changes in accuracy accompanied the changes in speed: when a fast reaction time was specified, both responded quickly, but both made more errors and hence had reduced accuracy.

An accumulator model described the individuals' behavior well, and also helped explain how both speed and accuracy changed with instruction. Because the model was required to account for both speed and accuracy with a single mechanism, it provided a stringent test of accumulator models. The model successfully described the behavior; further, changes in each of the model's three parameters were examined to see whether the model could produce the behavioral effects that were observed: rate of accumulation and bound height (both described previously) as well as residual time (an additional contributor to response latencies meant to encompass sensory and motor delays that are not part of the decision process). There was little change to drift rate and residual time when the speed–accuracy trade-off changed. Instead, changes in the speed–accuracy trade-off were best explained by changes in bound height. Specifically, when the individuals had fast reaction times and low accuracy, they had a low bound height, and when they had slow reaction times and high accuracy, they had a high bound height. Intuitively, this makes sense: increasing the bound height can be thought of as raising one's standards for how certain one has to be in order to make a decision. Raising one's standards improves performance, but it also takes more time. Similar findings have been obtained by BA Reddi and colleagues.

Notice that the rate of evidence accumulation – that is, the average drift rate of the decision variable – is governed by the intensity of stimulus motion. In contrast, the level of the bounds is controlled by the brain. J Palmer and co-workers showed that this level governs the trade-off between speed and accuracy. That may yield important insights into why performance may be different among individuals or even for the same individual under different circumstances. Unless decision time is measured, it is possible that the reason

for poor performance is a low level for the bound. Put simply, the decision maker might ignore useful information because the brain has already come to the end of the decision process. This also accounts for improvements in performance that fall below the expected $\sqrt{t}$ improvement.

The brain might employ deeper principles to set the bound. For example, it has been suggested that the bound could be set to maximize the overall rate of harvesting reward. RY Cho and colleagues tested this using a variant of the choice-reaction-time (RT) motion task. Reaction time as a function of error rate for a number of different task difficulties (coherences) was measured and was observed to be very close to that of an optimal estimator for some individuals. The estimator took into account the overall reward rate, which depended on the individual's reaction time, error rate, and the intertrial interval. Although not all of the individuals tested appeared to have adopted an optimal strategy, the close match between the optimal estimator and the best 30% of individuals suggests that neural systems may invest in trying to find the speed–accuracy trade-off that maximizes reward.

Taken together, these behavioral studies provide evidence for accumulators by demonstrating that the time it takes to reach a decision is linked to the accuracy of that decision. The model is not without its drawbacks. For example, in its simplest form, it fails to account for the RT on error trials, and it tends to overestimate the degree of skew seen in the distribution of RTs measured experimentally. A variety of extensions have been proposed to remedy these failures. In our view, they should not allow us to lose sight of the essential computational principle. The appeal of the accumulator model is that it provides an explanation for both choice and decision time, using a single mechanism – hence their success in a wide variety of decision tasks, including brightness discrimination, letter discrimination, and lexical decisions.

Physiological Observations

The key features of the accumulator model have been observed in neural recordings obtained from awake monkeys while they performed the choice-RT direction-discrimination task. Monkeys perform this task almost identically to the way humans do. Just as in the Palmer experiments with humans, monkeys indicated their decisions about direction by making a saccadic eye movement to a 'choice target.' By recording from neurons that encode one of the choice targets – its salience, importance, or the fact that it will be a target of an upcoming eye movement – it is possible to observe a neural correlate of an evolving decision during the period in which it is formed. Recordings from several brain areas have shed light on the process, but only one area has been studied during the choice-RT version of the random-dot motion task.

Neurons in the lateral intraparietal (LIP) area are candidates for providing evidence for the neural basis of accumulator models. Area LIP receives input from sensory areas (middle temporal (MT) and medial superior (MST) areas) that are sensitive to the motion stimulus used in the task. Further, it projects to areas known to play a role in the generation of saccadic eye movements. The position of area LIP in a sensory-motor middle ground suggests it is suited to participate in decisions that use incoming sensory information to make choices about outgoing movements. LIP neurons are selective for the direction of a saccade, so they respond more vigorously on decisions when the choice target is in the neuron's response field relative to when the choice target is in the opposite hemifield. In other words, their firing rates indicate the outcome of the monkey's decision on the motion task. However, the dynamics of the LIP response in the period leading up to the saccade suggest that these neurons play a role in the formation of the decision.

There is a gradual change in firing rates of LIP neurons in the period between onset of random-dot motion and the end of the decision process signaled by an eye movement. These changes are depicted by the traces in Figure 3(b). These traces are averages from many neurons over many trials. All of the solid traces correspond to trials that ended with an eye movement to the target in the neuron's response field (T_{in}). All of the dashed traces correspond to trials that ended with an eye movement to the opposite target (T_{out}). Not surprisingly, at the end of the decision, when the monkey is about to make an eye movement, the responses are clearly larger for T_{in} choices than for T_{out} choices (compare solid to dashed curves). In fact, the responses look like they reach the same firing rate, regardless of motion strength for all of the T_{in} choices (arrow). At the beginning of the trial, when the motion is first turned on, the responses are nearly identical for all choices and for all motion strengths.

What is happening in area LIP between the onset of random-dot motion and the eye movement response? In the first 100 ms after the dots appear there is a dip in the firing rate. This dip does not depend on the direction and strength of the motion, and it is the same regardless of the choice. It seems to signal the beginning of the decision process, but there is no sign yet of an accumulation of evidence for the T_{in} or T_{out} choice. The first sign that the neurons are affected by the direction of motion occurs ~200 ms after onset of the motion, when the solid and dashed

curves begin to separate. From this point forward, the trajectories represent the accumulation of evidence for and against the T_{in} choice. When the motion is strong and toward T_{in} (solid orange), the evidence from the visual cortex is overwhelmingly positive: the accumulation has a large positive slope. When the motion is strong and toward T_{out} (dashed orange), the evidence is overwhelmingly negative: the accumulation has a large negative slope. When the motion is weak, the evidence is less definitive and the accumulation has a smaller positive (for T_{in} choices) or negative (for T_{out} choices) slope.

Consider the case when the motion is entirely ambiguous (0% coherence motion). From moment to moment, the direction-selective neurons in the visual cortex are providing evidence that is sometimes positive and sometimes negative, but whatever the sign, the evidence is weak. The accumulation of such evidence is a random walk that meanders away from its starting point in either the positive or the negative direction. The average of all such random walks would be a flat line with zero slope. But we have sorted the responses in accordance with the monkey's choices. According to the accumulator model, the trials that end in a T_{in} choice (**Figure 3(b)**; solid dark blue) are those random walks that happen to wander to high firing rates – high enough to reach the bound (threshold) that terminates the decision process. The average of this half of the data should have a positive slope, and that is exactly what we see in **Figure 3(b)**. The other half of the trials ended in a T_{out} choice. According to the accumulator model, these are the random walks that never wandered far enough in the positive direction to cause a T_{in} choice. Instead, the decision terminated when the LIP neurons with the other target in their response field (our neurons' T_{out}) reached a high firing rate. Accordingly, this half of the data should appear to meander in a somewhat negative direction. That is exactly what we see in the dashed dark blue curve. The intermediate motion strengths (light blue) can be understood in much the same way.

The hallmarks of evidence accumulation and bound crossing just described are present in LIP neurons. First, as the monkey watches the motion stimulus, LIP neurons undergo ramplike changes in firing rate (**Figure 3(b)**, left panel). Importantly, responses to motion that is in the direction of the monkey's response field ramp up progressively at a buildup rate that is dependent on the quality of the evidence – the motion coherence. Around the time of the saccade, neural responses that have diverged, as a function of motion strength, coalesce to a common firing rate. About 70 ms after the responses come together, the saccade is executed (**Figure 3(b)**, right panel). This is a signature of the bound: when the firing rate of the LIP neurons reaches a critical value, the decision process ends with a decision in favor of the direction associated with the choice target in the neuron's response field. On the other hand, when the monkey chooses the opposite direction, the responses do not achieve a stereotyped level (**Figure 3(b)**, right panel, dashed curves). This observation suggests that these choices do not occur when the spike rate reaches a lower bound. Rather, the absence of a stereotyped termination value suggests that another process terminates the decision, presumably neurons with the chosen target in their response field. The observation thus supports a race between accumulators, like LIP neurons with each of the choice targets in their response fields, rather than the diffusion model in **Figure 2(c)**.

An accumulation to threshold model is able to reproduce the firing rate dynamics of LIP neurons well. Further studies have suggested that the momentary evidence that is accumulated in area LIP comes from a difference signal in area MT, that the accumulation represented in area LIP appears to be through a process resembling integration, and that LIP activity appears to be causally linked to the monkey's choice and response time.

Decision-related activity resembling the patterns in **Figure 3** have been observed in a number of brain areas for a variety of tasks. A gradual rise and decline in firing rates during decision making has been reported in the superior colliculus (SC), frontal eye field (FEF), and dorsolateral prefrontal cortex in studies using the random-dots task, although these studies did not measure reaction times. There is some evidence from human functional magnetic resonance imaging (fMRI) studies that a similar process underlies decision making in the human brain. In the SC, where neurons are associated with saccade-related and preparatory activity, responses of monkeys performing a two-choice position-identification task are related to reaction times and perhaps accuracy. And, in premotor cortex, responses appear to reflect different stages of the decision process when monkeys are engaged in a somatosensory discrimination task. Whether an accumulation to threshold model can explain the neural responses and behavior in all of these systems is not yet known.

Deciding to 'Go'

The behavioral and physiological observations thus far have been about perceptual decisions. However, accumulator models have been successful at describing even simpler decisions: whether to make a particular movement or not. These decisions arise when the

evidence is overwhelming for one choice, or in which there is only one possible choice. In that case, the only real decision is whether to go or not. Such decisions take more time than one would expect based on simple sensory and motor latencies. Even a simple eye movement to a single target takes ~200 ms to initiate, and there is a lot of variability in this response time from trial to trial. RHS Carpenter has championed the idea that this variable latency furnishes contingency – the possibility of doing something else – thereby imbuing even a simple motor act with the seeds of higher brain function.

A simple accumulator model has been proposed to describe these kinds of decisions. In this model, known as LATER (linear accumulation to threshold with ergodic rate), a decision signal rises linearly from a starting point until it reaches a threshold. Because (1) these decisions take very little time and (2) the accumulation is mainly positive, the process can be described very simply as a ramp with variable slope (Figure 2(d)). It is very easy to see from the geometry in Figure 2(d) that the distribution of decision times will have its median where the mean accumulation ramp intersects the bound. The distribution of decision times will be skewed to the right. Indeed, if the distribution of ramp slopes is Gaussian, then the reciprocals of decision times will also obey a Gaussian distribution.

Because the LATER model primarily describes reaction times (i.e., how long it takes to reach the bound) rather than accuracy, it is well suited to decisions about when and whether to 'go.' The simple model has been exploited to study the role of prior probability (i.e., bias), urgency, and stimulus strength on the decision process. Bias might arise when there is some uncertainty about which action is about to be instructed. Urgency refers to a time pressure on initiation after an instruction to go. Stimulus strength might be something like the intensity of the instruction or the ease of its interpretation, much like the coherence of random-dot motion. These studies exploit the change in the median reaction time and the shape of the distribution to gain insight into the process. It seems that bias and urgency affect the height of the bound in Figure 2(d), whereas stimulus intensity affects the rate or rise – what we would term the evidence accumulation.

This simple idea has been elaborated slightly to incorporate a race between two processes. As we have already noted, when there are two or more options, it may be sensible to conceive of two or more accumulators that reach a decision point, thus committing the organism to acting and selecting. In our view, the LATER type of accumulation makes sense when a single random number can effectively characterize the rate of accumulation. This will tend to be true when decision times are short and/or when the momentary evidence is the same at each moment in time. These conditions do not hold in the motion experiments reviewed earlier, but they do hold in some circumstances.

The countermanding task is a case in point. In countermanding, a second 'stop' signal instructs an individual to cancel a movement in the period between the onset of any eye movement target and movement initiation. When the stop signal is presented a long time before the individual is to make a saccade, the movement is easily withheld. However, when the stop signal occurs close to the time when the individual makes his saccade, the individual is sometimes unable to withhold the movement. The decision to go or cancel is well described by a race between a go and a stop signal.

There is physiological support of the kind of simple rise of a decision variable to a threshold, as postulated by LATER, to explain decisions to go. DP Hanes and JD Schall showed that movement neurons in the FEF exhibit a rapid rise in firing rate just before eye movement begins. These neurons, which are known to project to brain stem oculomotor structures, undergo a linear rise in firing rate after a target appears to instruct an eye movement. Hanes and Schall showed that the saccadic reaction times measured in monkeys could be explained by the rate of this rise in firing rate (Figure 4). Moreover, the firing rate reached a common value ~10–20 ms before the eye movement began, consistent with the presence of a threshold for initiation (Figure 4).

In countermanding, Hanes and Schall showed that these movement neurons also fail to achieve a critical level of discharge rate on trials in which the monkey successfully cancels a saccade. When the 'stop' signal comes too late and the saccade ensues, the FEF movement cells discharge as they do for saccades. This provides partial evidence for the type of race model hypothesized to underlie the decision process in countermanding. Indeed, a putative neural correlate of the other race – the stop signal – has been observed in the FEF and SC. However, the evidence for an accumulator-race model is still preliminary. It rests on the observation that fixation cells fire more and movement cells fire less when an eye movement is canceled. This is hardly surprising: by definition, the fixation cell responds when the animal does not make an eye movement. Thus, the argument that these cells participate in a race rests on subtle comparisons of firing rates in epochs predicted by the computational theory.

It seems secure that a rise in firing rate to a threshold level will account for the variable interval between

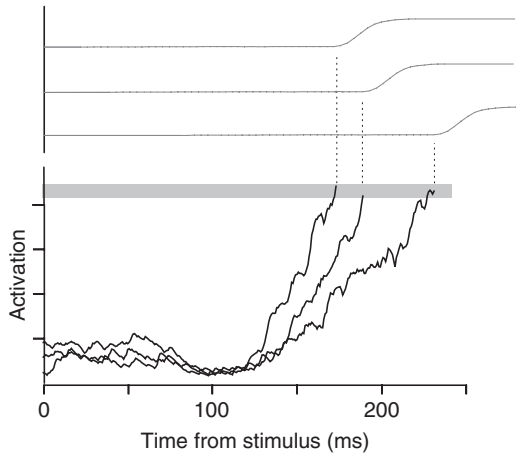

Figure 4 A neural correlate of a decision to go. Responses from a neuron in the FEF were recorded while a monkey made saccadic eye movements to a peripheral target. The three responses accompanied saccades to the same visual target in the response field of the neuron. Eye position traces (gray traces, top) show that the eye movements were nearly identical in amplitude, duration, and velocity but occurred at short, intermediate, and long latencies after the instruction to go. The firing rate began to increase at roughly the same latency from onset of the target (go signal), but the rate of increase was faster for the shorter latency saccades. Notice that the firing rate reaches a common value (horizontal gray bar) just preceding saccade initiation, consistent with the basic accumulator and LATER models in **Figures 2(a)** and **2(d)**. Adapted from Schall JD and Bichot NP (1998) Neural correlates of visual and motor decision processes. *Current Opinion in Neurobiology* 8: 211–217.

the onset of a visual target and movement initiation. It remains to be seen what role neurons in the FEF and other candidate structures play. Responses in the LIP, FEF, and SC can be altered by prior probability, reward expectation, uncertainty, and urgency, but it remains to be seen whether these responses interact with accumulation in the manner prescribed by LATER.

Limitations and Extensions of Accumulator Models

Certain kinds of tasks are less well suited to the simplest form of accumulator models. Some might necessitate modifications to the models in Figure 2, whereas others may simply require a different mechanism altogether. Simply put, accumulating evidence as it arrives in time is not always an optimal strategy for decision making. For example, when attempting to detect a relevant stimulus that occurs at an unpredictable time, accumulating evidence may be a poor strategy. If the brain were to accumulate evidence at the start of the trial, it might accumulate nothing but noise for the lengthy period of time while no stimulus is present. The accumulation of noise would cause a decision variable like the one depicted in Figure 1 to drift away from the starting point (in a random walk) toward one of the bounds, before the relevant information is even present. Not surprisingly, visual detection tasks typically do not appear to be mediated by the kind of accumulation strategies discussed here. When attempting to detect a low-contrast visual stimulus, humans perform only slightly better when the stimulus is presented for a longer durations. Up to ~80 ms, there is very clear improvement, as if stimulus intensity were accumulating in time (a phenomenon known as Bloch's law). But longer durations support only modest improvement in sensitivity. The improvement is explained by assuming that there is no accumulation of signal at all, but that longer viewing durations afford more independent opportunities to detect the stimulus. More opportunities decrease the probability that the stimulus will be missed.

Even without accumulation, it is possible to preserve the architecture of decision mechanism in Figure 2 for detection tasks. A time-varying decision variable is compared to a threshold in a race against other competing interpretations. Suppose the accumulation of momentary evidence is not perfect, but 'leaks' away. This way, even if noise causes the decision variable to drift toward one of the bounds, it will return to a neutral position fairly quickly. In principle, such a leaky accumulator could obey a time constant appropriately matched to the time course of the signals it is trying to detect – in other words, those governed by physiological responses to the visual stimulus. This approach has been championed with great success in psychophysics and cognitive decision making.

There is some evidence for this type of decision mechanism in neurophysiology. EP Cook and JH Maunsell trained monkeys to detect the onset of coherent motion. The monkey stared at a fixation point while random dots appeared in a patch of a video display. At the beginning of each trial, the dot display was completely random, like the 0% coherence condition in the discrimination task described earlier. On half of the trials, the motion became coherent after a random time interval. The monkey released a lever to indicate that he detected the coherent motion. Cook and Maunsell recorded from direction-selective neurons in area MT and the ventral intraparietal (VIP) area while the monkey performed the detection tasks. Neurons in area MT are strongly selective for the direction of stimulus motion when it is presented in a part of the visual field known as a neuron's receptive field. Neurons in area VIP receive a strong input from MT neurons and many exhibit responses that are selective for the direction of motion as well. The patch of random dots was aligned to the neuron's response field and the motion was in the

direction that would cause an increase in firing rate. Indeed, on trials in which the patch changed to coherent motion, the neurons increased their firing rates. The monkey detected the change when the instantaneous firing rate exceeded a criterion level. Importantly, the detection decision does not appear to be based on an accumulation of the MT response but rather on an instantaneous measure of firing rate. This makes sense because the brain cannot know when to begin accumulating evidence. (The actual estimate of instantaneous firing rate, whether performed by the brain or in data analysis, is achieved by smoothing the time series of spikes using a filter. The process can be described as leaky integration, but this should not be confused with accumulation.)

Accumulation of evidence in time may be misguided for other kinds of decisions as well. For example, some decisions involve comparing information across time. It would be detrimental to accumulate information in the time gap between two stimulus presentations. There is a need for some flexible control of integration in neural circuits. Indeed, as evidenced by R Romeo and colleagues in a two-interval vibrotactile comparison, neurons in the prefrontal cortex appear to encode both the vibration of the first stimulus, its memory in the time gap before a second comparison stimulus, and then the accumulation of evidence as the second stimulus is compared to the first. Clearly this requires an elaboration of the accumulation models depicted in Figure 1. On the other hand, storage (i.e., working memory) and accumulation are close cousins. Both rest on integration in the calculus sense, for both involve a stable value at a particular moment based on the history of inputs up until that moment. In this sense, working memory is just an accumulation of an impulse. In principle, a common mechanism might underlie both working memory and accumulation of evidence. Thus investigating not only the circuits involved in decisions, but also the cellular mechanisms of the neurons in those circuits, will be illuminating.

Real-World Decisions

The simple accumulator model will need to be elaborated to handle more complex decisions. There is some evidence that it can explain choice and decision times in decisions that are more complicated than a direction discrimination. However, a critical limitation is that samples of momentary evidence are accumulated in a single number, what we term a decision variable. But some decisions involve combining apples and oranges: perhaps evidence arrives from two different sensory modalities. One possible solution is that the brain does not accumulate raw evidence but rather evidence as it bears on a decision. For example, evidence from different sources may be more or less reliable. Studies of human perception and action suggest that the brain can combine such evidence by giving appropriate weights to the more and less reliable sources. The simple accumulators in Figure 2 need to be elaborated to explain such observations.

Certainly much work is still needed to bridge the gap between simple decision making in the laboratory and complex decision making in the real world. However, it would be a mistake to think that the kinds of mechanisms that have successfully described decision making so far would be ill suited to decisions in a more natural context. Some interesting examples from ecology suggest that animals accumulate evidence for and against particular choices. In prey selection, for example, a predator may observe a number of potential prey before making a decision about which to choose. Wolves are known to do this, and it is a strategy which greatly assists their hunting: a wolf might survey a herd of moose, for example, for quite some time, taking into account each animal's age, size, physical health, and even dental health before launching an attack. Making a choice too quickly could be detrimental to a wolf because a young healthy moose can retaliate quite effectively, injuring or even killing a wolf. This kind of decision making raises the question of how accumulators might work on different timescales. The difference between a LIP neuron responding to dot motion for about 800 ms and a wolf surveying his prey for several hours or even days is considerable. Nevertheless, these ecological observations suggest that animals in the wild might accumulate evidence and fine-tune speed–accuracy trade-offs the same way that monkeys and human subjects do in a laboratory setting.

See also: Contextual Interactions in Visual Perception; Contextual Interactions in Visual Processing; Retina: An Overview; Vision for Action and Perception; Vision: Light and Dark Adaptation; Visual–Vestibular Interactions.

Further Reading

Carpenter R and Williams M (1995) Neural computation of log likelihood in control of saccadic eye movements. *Nature* 377: 59–62.

Colby CL, Duhamel JR, and Goldberg ME (1993) Ventral intraparietal area of the macaque: Anatomic location and visual response properties. *Journal of Neurophysiology* 69: 902–914.

Cook EP and Maunsell JH (2002) Dynamics of neuronal responses in macaque MT and VIP during motion detection. *Nature Neuroscience* 5: 985–994.

Ernst MO and Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415: 429–433.

Gold IJ and Shalden MN (2003) The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *Journal of Neuroscience* 23(2): 632–651.

Gold JI and Shalden MN The neural basis of decision making. *Annual Review Neuroscience* 30 (in press).

Good IJ (1979) Studies in the history of probability and statistics. XXXVII A.M. Turing's statistical work in World War II. *Biometrika* 66: 393–396.

Hanes DP and Schall JD (1996) Neural control of voluntary movement initiation. *Science* 274: 427–430.

Hanks TD, Ditterich J, and Shadlen MN (2006) Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience* 9: 682–689.

Heekeren HR, Marrett S, Bandettini PA, et al. (2004) A general mechanism for perceptual decision-making in the human brain. *Nature* 431: 859–862.

Link SE (1992) *Wave Theory of Differnce and Similarity.* Hillsdade, NJ: Earlbaum Associates.

Link SW and Heath RA (1975) A sequential theory of psychological discrimination. *Psychometrika* 40: 77–105.

Machens CK, Romo R, and Brody CD (2005) Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science* 307: 1121–1124.

Palmer J, Huk AC, and Shadlen MN (2005) The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision* 5: 376–404.

Platt ML and Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400: 233–238.

Ratcliff R and Rouder JN (1998) Modeling response times for two-choice decisions. *Psychological Science* 9: 347–356.

Reddi BA, Asrress KN, and Carpenter RH (2003) Accuracy, information, and response time in a saccadic decision task. *Journal of Neurophysiology* 90: 3538–3546.

Roitman JD and Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience* 22: 9475–9489.

Romo R, Hernandez A, and Zainos A (2004) Neuronal correlates of a perceptual decision in ventral premotor cortex. *Neuron* 41: 165–173.

Schall JD and Bichot NP (1998) Neurol correlates of visual and motor decision processes. *Current Opinion in Neurobiology* 8: 211–217.

Shadlen MN and Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology* 86: 1916–1936.

Sugrue LP, Corrado GS, and Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. *Science* 304: 1782–1787.

Watson AB (1986) Temporal sensitivity. In: Boff KR, Kaufman L, and Thomas JP (eds.) *Handbook of Perception and Human Performance*, pp. 6.1–6.43. New York: Wiley.

Decision-Making in Financial Markets

P Bossaerts, California Institute of Technology, Pasadena, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Finance

Financial markets are a relatively recent phenomenon. They are markets where agents trade claims to future monetary payoffs, commonly referred to as securities or assets. Almost invariably, the payoffs are uncertain. That is, they entail financial risk. The first modern organized financial market was set up in Antwerp, in the lower countries (present-day Belgium), in 1460. At the time, the securities were claims on future payments by established buyers of goods and services, as well as sovereigns. Only much later (around the middle of the nineteenth century) was the type of asset – namely, equity – that most people would be familiar with in modern-day financial markets introduced. Equity, or common stock, is a claim on the residual earnings of firms, which means that the holder is entitled to the cash flow of the firm after the firm pays all debtors (which includes suppliers, employees, and banks). Much more recently, derivative securities were introduced (although many of them were already occasionally traded by the Dutch in the sixteenth century). These are claims on the payoffs on other securities. Options constitute the classical example of derivatives. Nowadays, a bewildering variety of securities are traded on and off organized exchanges.

Finance studies the behavior of agents when faced with financial risk, as well as the pricing of such risk in markets. Finance is both descriptive (why are prices the way they are?) and normative (how to compute optimal portfolios?). In many respects, finance is a subfield of economics, but there are a couple of features that really set it apart. The differences will be discussed shortly, because they may lead to fresh insights for neuroscience. Far more than economics, finance relies on probability theory. This is not surprising: early developments of probability theory derived from attempts to understand monetary games, as exemplified by Pascal's proposal for computing winning odds. That is, probability theory originated in many respects from efforts to formalize financial risk.

Many Facets of Risk

One dimension in which finance and economics differ substantially is that the former explicitly deals with the many facets of risk. Rather than summarizing the attractiveness of monetary gambles in terms of a single-dimensional index, ever since the work of Harry Markowitz, in the early 1950s, finance scholars have represented risk in terms of expected payoff, payoff variance, skewness, and even kurtosis. These constitute mathematical characterizations of payoff distributions. The expected payoff is just the mean payoff that one anticipates. The variance is the (square of the) typical ('standard') deviation from the mean. The skewness measures whether payoffs below the mean are more likely (negative skewness) or not (positive skewness). Kurtosis measures to what extent extreme payoffs (both above and below the mean) are more likely, compared to a benchmark payoff distribution – namely, the well-known bell-shaped gaussian distribution.

Analyzing Risky Payoffs

Let p_t denote the payoff on a monetary gamble in trial period t. Let x_t denote the anticipated, or expected, payoff. Mathematically:

$$x_t = \mathrm{E}[p_t]$$

where E denotes the expectations operator (whereby all possible outcomes are multiplied by their respective probabilities and summed). The payoff variance is defined as follows:

$$\upsilon_t = \mathrm{E}[(p_t - x_t)^2]$$

So, the payoff variance is the anticipated deviation from the expected payoff.

The skewness is defined as the expectation of the third-order power of the difference from the expected payoff:

$$\phi_t = \mathrm{E}[(p_t - x_t)^3]$$

The kurtosis is proportional to the fourth-order power:

$$\kappa_t \sim \mathrm{E}[(p_t - x_t)^4]$$

(The constant of proportionality is such that the kurtosis of a gaussian random variable equals 3, independent of its expected outcome and its variance.) Like variance, skewness and kurtosis are anticipated deviations from the expected payoff, but negative deviations are counted separately in computing skewness (because the third power of a negative number is negative, unlike the second and fourth powers), and the fourth power with which the kurtosis is computed effectively puts a heavy weight on outliers.

While there may be many other facets of risk, finance in general limits its attention to these four relationships. In the case of common stock, often only the first two (expected payoff and payoff variance) are considered (see Figure 1). It is known, however, that expected payoff and payoff variance provide an incomplete description of the risk of other securities, such as options. Humans are sensitive to skewness and kurtosis as well, and the most convincing way to confirm this is to consider the prices of options, which include huge premia for the skewness and the kurtosis of their payoffs.

Economists take a different approach. Ever since Bernoulli proposed logarithmic expected utility, economists have generally insisted on representing the attractiveness of risky gambles in terms of a single-dimensional index. There are very few exceptions. Even the celebrated Prospect Theory remains a representation of preferences in terms of a single-dimensional expected utility index. In Prospect Theory, the index is basically a multiplication of (subjective) probabilities of each of the possible payoffs with their respective utilities (nonlinear transformations of the payoffs). Of course, one can justify the finance approach in terms of a mathematical operation on expected utility – namely, a Taylor series expansion in terms of expectation, variance, skewness, and kurtosis. As such, an agent with logarithmic expected utility (i.e., Bernoulli's agent) prefers higher expected return, is averse to variance, prefers positive skewness (while averse to negative skewness), and is averse to kurtosis.

But the Taylor series approach is by no means necessary. Moreover, by not insisting that monetary gambles are ranked in terms of a single-dimensional index, one avoids properties that many would find counterintuitive. For instance, in expected utility theory and its variations (such as Prospect Theory), there is a tight relationship between the curvature of the utility function and risk aversion. In particular, the degree to which the utility function becomes less curved as wealth increases and, hence, the degree to which utility of an extra dollar decreases with wealth determine one's risk aversion. The faster the utility of an extra dollar decreases, the more risk averse one is. As such, fear of uncertainty is tied to decreasing utility for (certain!) extra monetary units, a rather counterintuitive property (see Figure 2).

After it was introduced in the 1950s, the finance 'mean-variance' approach proved extremely useful for determining optimal portfolios of securities. Indeed, it is rather complicated mathematically to

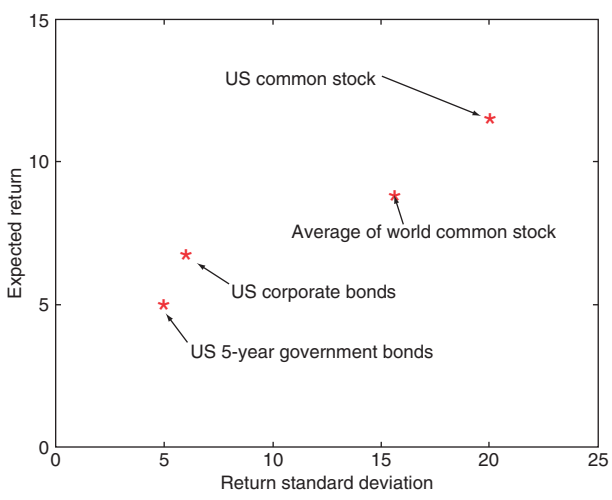

Figure 1 Trade-off between expected payoff and payoff standard deviation (square root of variance) from investing in several asset classes, based on historical data from 1926 to 2004. Payoffs are expressed in percentage returns (i.e., in percent per dollar invested). Returns are computed from investing at the end of the year and selling at the end of the subsequent year. Notice the trade-off: higher expected returns (e.g., on US common stock) imply higher risk (higher payoff variance). The degree to which one requires higher expected payoffs in order to be willing to accept higher risk measures one's risk aversion. Based on data from Center for Research in Security Prices of the University of Chicago.

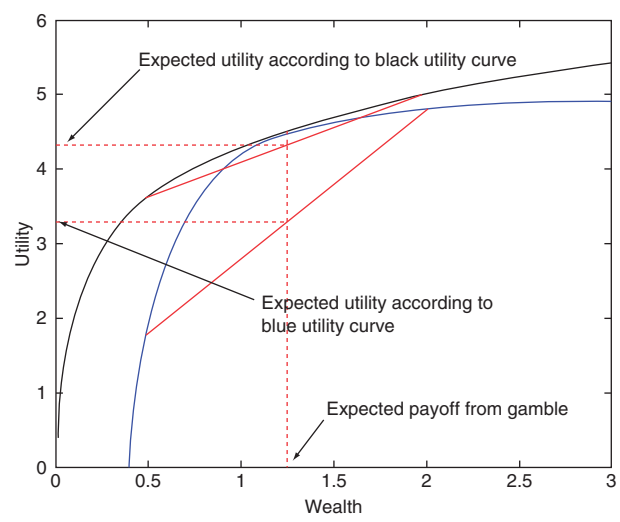

Figure 2 Effect of curvature of the utility function on risk aversion in the context of expected utility preferences. Shown are two utility functions. The black one is less curved than the blue one, in the sense that the degree to which additional utility of a dollar decreases with wealth is smaller for the former. This translates into lower risk aversion: the expected utility of a gamble that pays 0.5 or 2 with equal chance is higher for the black utility curve than for the blue one, implying that a person with the black utility function is willing to pay more for the gamble compared to someone with the blue utility curve. Using a simple geometric trick, the expected utility can readily be located by intersecting the vertical line through the expected payoff (vertical dotted red line) and the chord that connects the utility of the low payoff (0.5) with that of the high payoff (2) (solid red lines).

analyze the attractiveness of simultaneous risky gambles with correlated risks if one insists on, say, logarithmic expected utility, let alone Prospect Theory. In contrast, the properties of optimal portfolios in a simple mean-variance framework are very intuitive, and, after their discovery, led to profound insights in the 1960s about the pricing of risk in financial markets. In particular, it was realized that financial markets would not compensate investors for the entire risk as measured by a security's payoff variance, but only to the extent the payoff covaries (correlates) with the payoff on all other available securities. The covariation risk has become known as the 'beta' risk of a security. Laboratory experiments have demonstrated that the principle is correct. In such experiments, groups of human participants trade securities over anonymous, electronic markets. The securities are short-lived and their payoff patterns are deliberately kept simple and intuitive.

But there are other reasons why the finance approach is convenient. Among other things, it allows one to describe risk differences in very intuitive, statistical terms, and it facilitates learning. Both properties are relevant to neuroscience.

Describing Differences between Environmental and Financial Risks

It is likely that, through evolutionary pressures, the human brain is optimized to deal with environmental risks. This would inspire one to believe that the human brain is well adapted to financial risks as well. Yet financial risks and environmental risks differ substantially.

Two differences stand out. First, many environmental risks are adequately described by the gaussian distribution. This is not so for financial risks, which often display substantial excess kurtosis (extreme events happen far more often than predicted under the gaussian distribution). See Figure 3, for instance, which compares the risks of investing in Microsoft common stock and weather-related risks. Second, environmental risks often emerge in clusters, which means that the chance of their occurrence depends on whether they have just occurred. In contrast, the expected payoffs from investing in securities are rather independent over time, as witnessed by the ubiquitous warning in investment prospectuses that "past performance is no indication of future performance."

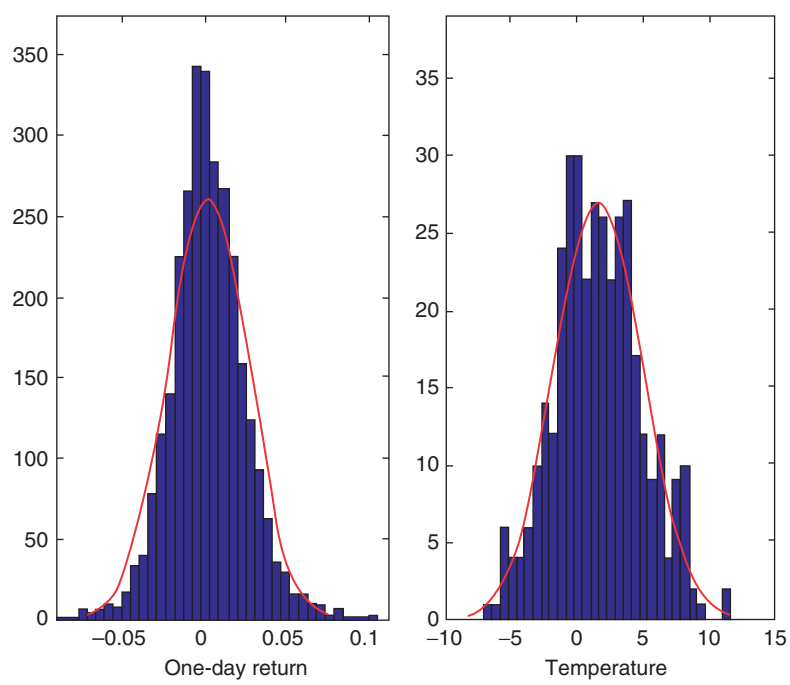

Figure 3 Histograms of daily returns (1-day payoffs per dollar invested) on Microsoft common stock (14 March 1986 till 31 December 1997; left panel) and of daily average temperature in Geneva, Switzerland in January (1995–2006). Red lines indicate best gaussian fit. The financial data (left panel) exhibit far more kurtosis than does the gaussian distribution, in the sense that extremely high and low returns are far more frequent. The histogram of the financial data is cut at a distance of about 4 standard deviations from the mean: seven data points are lost as a result; in particular, the maximum (17.97%) and the minimum (−30.12%) are not shown. The histogram of the temperature data (right panel) is not trimmed. The kurtosis of the financial data is estimated to be 15; the kurtosis of the temperature data is about 3, which is equal to the theoretical kurtosis of a gaussian random variable. Source: Center for Research in Security Prices of the University of Chicago (financial data) and the University of Dayton (Ohio) average daily temperature archive (temperature data).

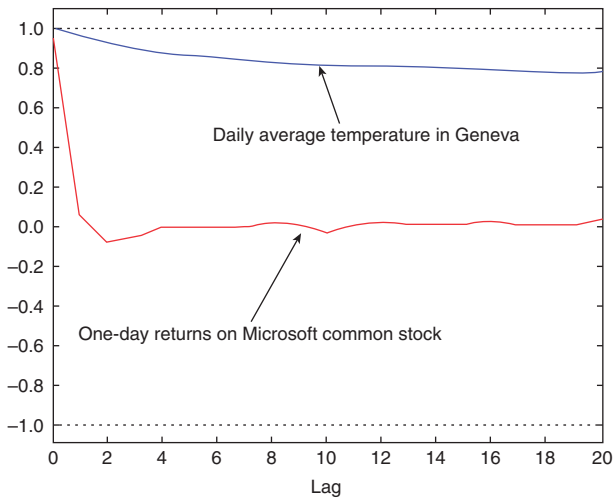

Figure 4 Time series correlation of daily returns on Microsoft common stock and of the average daily temperature in Geneva, Switzerland. Shown are the estimated correlations between the outcome of a series and its outcome *k* days in the past, where *k* is the lag indicated on the horizontal axis. A time series correlation close to zero indicates absence of predictability (past and future are uncorrelated); correlations close to 1 or −1 imply substantial predictability (the past is a very good indication of the future). Time series correlations of lag 0 are equal to 1, by definition (see **Figure 3** caption for data details).

Figure 4 illustrates the difference by comparing the time series correlation of payoffs on Microsoft common stock and that of the average daily temperature in Geneva, Switzerland.

One could claim that the differences between environmental and financial risks may explain why humans' attitudes toward financial risks are often 'irrational.' Before jumping to this conclusion, however, it is important to consider that, as we shall discuss soon, the brain seems to be using the very statistical language in terms of which the disparities between financial and environmental risks are most readily detected.

Learning in a Finance Framework

Learning is easy within the language of finance. This is because monetary gambles are described in terms of statistical moments of their payoffs (expectation, variance, skewness, kurtosis), and these moments can be estimated in an unbiased and efficient way using simple averages of past outcomes. This contrasts with learning in the context of single-dimensional utility indices. With the exception of classical expected utility indices, it is not obvious how updating should be done. Controversy remains even today, for instance, on how individuals with Prospect Theory preferences update their (biased) beliefs.

Predicting Changes in Prediction Risk

Financial payoffs are almost never equal to their predictions; actual and expected payoffs generally differ. That is, there is prediction risk. A second respect in which finance differs significantly from economics, and in fact from almost all other fields, is in its concern for tracking prediction risk. This is because financial prediction risk at times increases suddenly, generally after unusually large or small payoffs, and it decreases only gradually afterwards (see **Figure 5**). A simple statistical model known as generalized autoregressive conditional heteroskedasticity (GARCH) describes this process. As proposed by Robert Engle in the mid-1980s, in the GARCH model, risk increases after a large error in assessing prediction risk. If no further errors occur, prediction risk gradually decreases.

In computational neuroscience, one often distinguishes between estimation risk and (pure) risk. The former is the expectation of the size of the difference between estimates and true values of parameters that describe the distribution of outcomes; it reflects 'ignorance.' The latter is the expectation of the size of the difference between outcomes and the theoretical value of these parameters; 'output noise' is sometimes used as a synonym. When one is only concerned about prediction, this distinction may not be very relevant. All that matters is the prediction risk – that is, the expected size of the difference between outcomes and predictions. For other purposes, the distinction is crucial (e.g., when one is interested in testing whether the theoretical value of underlying parameters is different from zero). From a practical point of view, finance is about prediction, and, hence, the terms risk and prediction risk are used interchangeably. So, the payoff standard deviation plotted in **Figure 1** is really the prediction risk, measured as the historical standard deviation of the prediction error. (Likewise, the expected payoffs plotted in **Figure 1** are really predictions based on historical data.)

Prediction risk is of utmost importance in finance, among other factors, because the value of derivative securities such as options depends crucially on the risk in predicting the payoff of the underlying asset. Therefore, correct valuation of financial securities requires close monitoring of changes in prediction risk. This implies, in particular, tracking of prediction risk errors, to be used to update one's estimate of future prediction risk. The importance of this endeavor is further underscored by the recent introduction in many organized financial markets of securities with payoffs that depend on realized prediction risk (realized volatility).

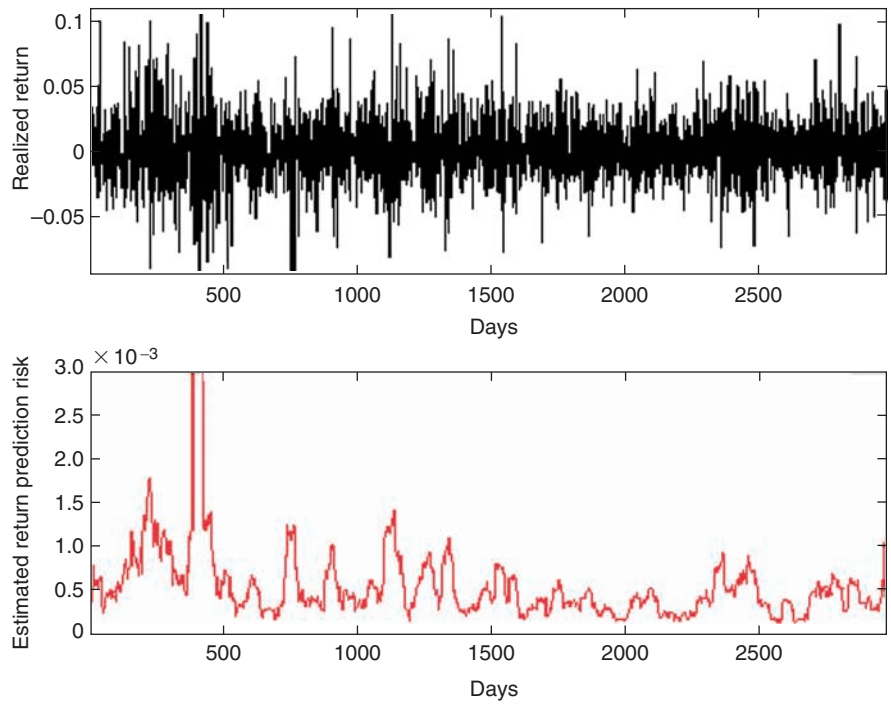

Figure 5 Historical daily returns on Microsoft common stock (top panel) and estimate of return prediction risk (bottom panel), over 2985 trading days (14 March 1986 till 31 December 1997). Return prediction risk is estimated as a 40-day moving average of the squared return. The sudden increases in prediction risk are visible despite the smoothness that the moving-average filter attempts to impose. Once prediction risk has increased, it decreases only gradually over time. These are the two main features of a generalized autoregressive conditional heteroskedasticity (GARCH) process. The GARCH model not only describes prediction risk in financial markets at the daily level (displayed here), but also at the minute or even subminute level.

Finance is one of the few fields in which prediction risk is considered to be stochastic. In many other fields, prediction risk is assumed to change deterministically. For instance, in the celebrated Kalman filter model, prediction risk changes according to a nonstochastic set of difference equations. The Kalman filter model applies to cases in which the underlying causes behind outcomes change slowly over time. The model is inadequate as a description of financial risk, however, wherein prediction risk often increases suddenly with past prediction risk errors, which means that prediction risk is stochastic. This is exactly what the GARCH model does capture.

Prediction Risk: Comparing The Kalman Filter and GARCH Models

Let p_t denote the payoff on a monetary gamble in trial period t and let x_t denote the predicted (expected) payoff. Let e_t denote the prediction error. That is,

$$e_t = p_t - x_t$$

Let υ_t denote the prediction risk, measured as the expectation of the square of the prediction error. In the Kalman filter model, υ_t changes over time deterministically, as follows:

$$\upsilon_{t+1} = a^2\upsilon_t + q$$

for some scalars a and q. In the simplest GARCH model,

$$\upsilon_{t+1} = a^2\upsilon_t + \kappa(e_t)^2$$

There are more complicated GARCH models, wherein the evolution of the prediction risk depends on lags of the prediction risk and of the squared prediction error beyond one period. But the essential difference with the Kalman filter model is that prediction risk changes stochastically because of the presence of a random term (the square of the prediction error e_t).

It may seem that the GARCH model is rather complicated. Yet, there is a relationship between the GARCH model and the Rescorla–Wagner reinforcement learning (RL) model from neuroscience. Specifically, a simple application to prediction risk of the Rescorla–Wagner rule produces a particular case of the GARCH model. This means that a simple RL model on prediction risk errors generates a GARCH model.

Reinforcement Learning Models for Risk Prediction as Special Cases of GARCH

The prediction risk error is defined as the difference between the realized prediction risk, measured as the squared prediction error $(e_t)^2$, and the anticipated prediction risk (υ_t):

$$\delta_t = (e_t)^2 - \upsilon_t$$

Applying the temporal difference (TD) model to update prediction risk, one obtains:

$$\begin{aligned}\upsilon_{t+1} &= \upsilon_t + \kappa\delta_t \\ &= \upsilon_t + \kappa[(e_t)^2 - \upsilon_t] \\ &= (1-\kappa)\upsilon_t + \kappa(e_t)^2\end{aligned}$$

which is a GARCH model with the restriction that the coefficient a^2 equals $(1-\kappa)$.

Neuroscience: The Finance Perspective

Evidence has accumulated recently that one of the primary roles of the dopaminergic system in the primate brain is to facilitate reward learning. Electrophysiology studies have confirmed that phasic firing of midbrain dopamine neurons correlates with errors in prediction (of cumulative anticipated rewards), and that these prediction errors affect future predictions through a simple but powerful Rescorla–Wagner rule. Functional magnetic resonance imaging (fMRI) scanning studies have found corresponding blood-oxygen-level-dependent (BOLD) signals in key subcortical dopaminoceptive areas such as ventral striatum and putamen. From a finance point of view, however, the Rescorla–Wagner rule is not fully satisfying as a model to predict rewards. This is because it does not take into account prediction risk.

Learning and Prediction Risk

Prediction risk is a key factor in determining how prediction errors should affect one's estimate of future payoffs. If the prediction risk is large, then a large prediction error is to be expected, and, hence, the prediction error should not have a large impact on future predictions. Conversely, if the prediction risk is small, then a large prediction error should cause one to substantially update one's prediction.

In terms of the Rescorla–Wagner rule of learning, this means that prediction errors should be scaled (normalized by) prediction risks in order to obtain a scale-free prediction error with which to update predictions. In general, the scaling is not simply a matter of dividing the prediction error by the (square root of the) prediction risk, but in certain cases it is. This suggests the hypothesis that the dopaminergic system encodes not just the prediction error, but also a scale-free version.

Accommodating Prediction Risk in the Rescorla–Wagner Learning Rule

When prediction risk is high, prediction errors should not change one's future predictions much because the prediction errors were expected to be sizeable anyway. Therefore, one should not update one's prediction with a constant κ, as in the traditional model. Using the notation introduced in previously, the traditional model reads as follows:

$$x_{t+1} = x_t + \kappa e_t$$

Instead, the update should be based on a scale-free prediction error η_t. In prediction of rewards in the model underlying the standard Kalman filter, the scaling takes on a very simple form:

$$\begin{aligned}x_{t+1} &= x_t + (\tilde{\kappa}/\upsilon_t^{1/2})e_t \\ &= x_t + \tilde{\kappa}(e_t/\upsilon_t^{1/2}) \\ &= x_t + \tilde{\kappa}\eta_t\end{aligned}$$

where, as before, υ_t denotes the prediction risk for trial period t ($\upsilon_t = \mathrm{E}[(e_t)^2]$). (Optimal prediction can be shown to imply that $\tilde{\kappa}$ should be proportional to the covariance between the outcomes and the scaled prediction error η_t.) Recent electrophysiological evidence from the nonhuman primate brain obtained in Wolfram Schultz' lab supports this hypothesis: at the time of reward delivery, the encoding of prediction errors in midbrain dopamine neurons appears to be scale-free (see Figure 6).

If dopaminergic neurons are able to attune firing to prediction risk, the brain must not only keep track of predictions, but also prediction risk. It seems that the brain does so, and the relevant signal may even be in the dopaminergic system itself. First, Schultz' lab has discovered that (other) midbrain dopamine neurons fire with delay during the anticipation period and throughout the period of reward delivery and that this delayed firing increases with prediction risk. This may provide the input signal with which raw prediction errors are scaled at the time of reward delivery. Second, fMRI analysis of certain subcortical dopaminoceptive regions of the human brain has also revealed a delayed signal that increases in prediction risk (see Figure 7).

One could go a step further. Because prediction risk generally changes over time, there is a need for the brain to encode prediction risk errors, with which to update prediction risk. To date, it is not known whether and where this is accomplished in the brain.

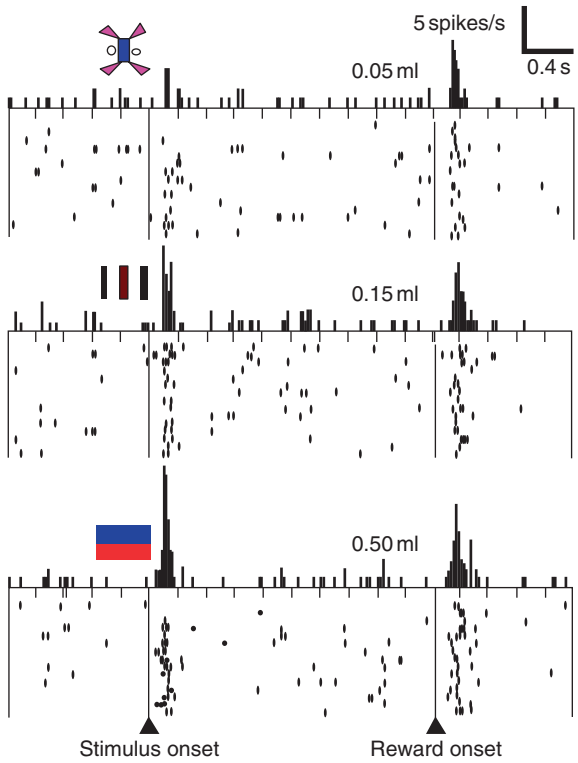

Figure 6 Spiking of primate midbrain dopamine neurons in three different types of trials involving a stochastic juice reward. In the first type, shown on top, the predicted reward (expected reward) and prediction risk (reward variance) are low. Predicted reward and prediction risk increase in the second trial type (middle) and again in the third trial type (bottom). The increase in predicted reward can be seen in the spiking of the neurons at stimulus onset, which is known to reflect prediction errors relative to baseline. The spiking at the time of reward is known to reflect prediction errors relative to expectation at stimulus onset. Only rewarded trials are shown, which is why the spiking at reward onset is always positive. Notice that the size of the spiking at reward onset does not reflect the differences of the sizes in prediction risk across trial types. In other words, the prediction error encoded in spiking at reward onset is scaled in a way that accommodates prediction risk. Reprinted from Tobler PN, Fiorillo CD, and Schultz W (2005) Adaptive coding of reward value by dopamine neurons. *Science* 307: 1642–1645, with permission from AAAS.

There is some recent suggestive evidence that insula is crucial to encoding uncertainty-related phenomena, including prediction risk, but there is no account of prediction risk errors. Some of the evidence on activation of the midbrain locus coeruleus seems to suggest that norepinephrine may play the role for prediction risk errors that dopamine plays for prediction errors. Peter Dayan's group have recently started to distinguish between predicted and unpredicted uncertainty and have pointed to the role of norepinephrine in capturing the latter. Unpredicted uncertainty has an obvious relationship with prediction risk errors.

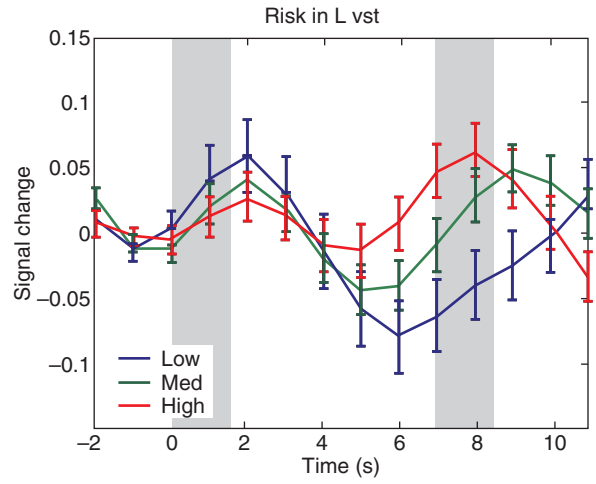

Figure 7 Functional magnetic resonance imaging evidence of activation of the subcortical dopamine system in the human brain modulated by prediction risk (measured as reward variance). Shown are time courses of the blood-oxygen-level-dependent signal in the left ventral striatum (L vst) in trials stratified by prediction risk. Left-hand gray area indicates period of stimulus onset (when reward prediction and prediction risk for the trial are conveyed); right-hand gray area indicates period of reward realization. Note that the prediction risk signal emerges with a delay: activation differentiation is maximal only after 7 s, which is approximately 2 s after peak activation in the canonical hemodynamic response. Reproduced from Preuschoff K, Bossaerts P, and Quartz S (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51: 381–390, with permission from Elsevier.

A Dual Role for Encoding Prediction Risk

The encoding of prediction risk not only facilitates learning: it can simultaneously play a crucial role in the determination of preferences. In finance, prediction risk is a crucial determinant of the desirability of a gamble. In Figure 1, the prediction risk of various securities is plotted on the horizontal axis (prediction risk is measured in terms of variance). As such, the encoding of prediction risk not only allows the brain to correctly update predictions, but at the same time, it also provides one of the parameters with which the brain can rank gambles. It is significant in this regard to note that the evidence for a prediction risk signal in the left ventral striatum of the human brain (see Figure 7) emerges even when there is no role for learning, and when individuals' behaviors (choice, reaction time) and brain activation (across trials) indeed do not show any evidence of updating. That is, the prediction risk signal plays a purely perceptual role.

The dual role of the encoding of prediction risk implies economy. The resulting efficiency is a powerful argument why the brain would not want to characterize the desirability of gambles in terms of a single-dimensional index, in contrast to economic theory. Of course, there is still the possibility that

the brain does build a single-dimensional index on the basis of signals of expected payoff and payoff variance, but the point is that the raw inputs with which the index is computed come from a decomposition of the features of a random payoff in terms of statistical moments, as is done in finance. Moreover, the statistical inputs need not always be weighted in the same way, thus potentially explaining why attitudes toward risk appear to be unstable.

See also: Decision-Making and Neuroeconomics; Delayed Reinforcement: Neuroscience; Delayed Reinforcement: Economics; Games in Monkeys: Neurophysiology and Motor Decision-Making; Neuroeconomics: History; Reward Decision-Making.

Further Reading

Bossaerts P and Plott C (2004) Basic principles of asset pricing theory: Evidence from large-scale experimental financial markets. *Review of Finance* 8: 135–169.

Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987–1008.

Engle RF (2002) New frontiers for ARCH models. *Journal of Applied Econometrics* 17: 425–446.

Fiorillo CD, Tobler PN, and Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299: 1898–1902.

Huang CF and Litzenberger RC (1998) *Foundation of Financial Economics*. New York: Prentice Hall.

Markowitz H (1952) Portfolio selection. *Journal of Finance* 7: 77–91.

Preuschoff K, Bossaerts P, and Quartz S (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51: 381–390.

Schultz W (1998) Predictive reward signal of dopamine neurons. *Journal of Neurophysiology* 80: 1–27.

Sharpe WF (1964) Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Tobler PN, Fiorillo CD, and Schultz W (2005) Adaptive coding of reward value by dopamine neurons. *Science* 307: 1642–1645.

Yu AJ and Dayan P (2002) Expected and unexpected uncertainty. Ach and NE in the neocortex. In: Becker S, Thrun S, and Obermayer K (eds.) *Advances in Neural Information Processing Systems,* 15, pp. 157–164. Cambridge, MA: MIT Press.

Relevant Website

http://nobelprize.org – R Engle's Nobel lecture, on GARCH models; H Markowitz' Nobel lecture, on portfolio theory; W Sharpe's Nobel lecture, on 'beta' pricing models.

Declarative Memory System: Anatomy

W A Suzuki, New York University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Think for a moment about all the memories you conjure up in a typical day. Perhaps something in today's newspaper reminds you of a recent debate with a friend, a TV program brings to mind a memorable vacation, or a perfume sampled at the department store recalls a favorite aunt. Our ability to retrieve the factual and event-based information from our past is called declarative memory. More formally, declarative memory is defined as the facts and events from our own personal histories that can be consciously brought to mind. We form and retrieve declarative memories every hour of the day. Without a functioning declarative memory system, as is the case in amnesia, we would be unable to hold down a job, keep track of our daily appointments, or find our way around. Declarative memory not only plays a critical part of our daily lives, but the facts and events in our declarative memory help shape our personalities and define who we are.

Surprisingly, the first systematic studies of human memory did not begin until the late 1800s. In 1881, the French psychologist Theodule Ribot published *Les Maladies de la Memoire* (Diseases of Memory). In it, Ribot provided the first systematic description of human amnesic patients, noting that when brain injury causes memory impairment, recent memories are more susceptible to disruption than remote memories. This principle has become known as Ribot's Law. In 1885, Hermann Ebbinghaus published *Uber das Gedachtnis* (On Memory), in which he was the first to apply systematic experimental approaches to the study of normal human memory. For example, he showed that some memories are only short-lived while others are long lasting, and the amount of repetition can influence the duration of the memory. In 1890, the famous American psychologist William James published *Principles of Psychology* in which he described distinctions between what he called primary memory, defined as information that forms the focus of current attention, and secondary memory, defined as memory that persists even after it has dropped out of consciousness.

By the beginning of the twentieth century a major debate in the field concerned whether memory function was localized to a particular brain area or was distributed throughout the brain. The results from early experimental studies were mixed. However, in the 1920s Karl Lashley, a well-known psychologist at Harvard University, performed a series of influential experiments that addressed this debate. He trained rats to run in a simple spatial maze and then examined the effect of lesions to different parts of their cortex. He found that the larger the cortical lesion, the more severe the memory impairment on the rat's maze performance. He referred to these findings as the Law of Mass Action and concluded that memory was distributed over widespread cortical areas. This view dominated the field for more than 25 years, until findings from a single landmark study transformed our understanding of the localization of memory in the brain.

In 1957, Scoville and Milner described the effects of experimental bilateral lesions of the medial temporal lobe in 10 patients. These surgeries were done in an attempt to relieve a variety of psychiatric or neurological conditions, including schizophrenia, manic depressive psychosis, and, in one case, epilepsy. One striking and completely unexpected result of these experimental procedures was a profound memory loss that appeared to be correlated to the amount of bilateral hippocampal damage sustained by the patients. This memory loss was most obvious and easiest to study in patient H.M., who underwent the surgery in an effort to relieve severe and intractable epilepsy. Other patients also exhibited memory loss, but were more difficult to study because of their persistent psychiatric conditions. H.M., who received one of the largest medial temporal lobe resections, was described as being able to recall nothing of recent events of his life, though his childhood memories remained intact. Moreover, while H.M.'s memory impairment was profound, affecting memory for all sensory modalities tested (i.e., visual, auditory, somatosensory, and even olfactory memory), his general cognitive and intellectual abilities were spared.

This landmark study demonstrated several fundamental principles of memory organization in the brain. First, it showed definitively that memory could be localized to a particular brain area, namely the medial temporal lobe. Second, it demonstrated that memory could be studied independently of other general cognitive functions. Third, it led the way to more recent demonstrations that the medial temporal lobe has a critical and selective role in establishing declarative memory for facts and events. Fourth, this report was the catalyst for the ensuing experimental studies focused on defining more precisely the neuroanatomical basis of declarative memory. The following sections describe findings from the

series of experimental lesion studies and neuroanatomical studies in animals that, cumulatively, led to the identification of the subset of medial temporal lobe structures important for declarative memory.

The Amygdala, Hippocampus, and Development of an Animal Model of Human Amnesia

The amygdala and the hippocampus are two prominent structures that sit just deep to the cortex of the medial temporal lobe. The amygdala is a rounded almond-shaped structure that is situated in the anterior portion of the medial temporal lobe behind the cortex of the temporal pole. The hippocampus is an irregular tube-shaped structure that is located just posterior to the amygdala. The noted neurosurgeon Dr. William Scoville, who performed the surgery on patient H.M., described the medial temporal lobe resection as removing 8 cm of the medial temporal lobe tissue, including the temporal pole, amygdala, and approximately two-thirds of the anterior–posterior length of the hippocampus (Figure 1). Scoville and Milner initially hypothesized that bilateral damage to the hippocampus was responsible for the severe memory impairment, but because H.M.'s resection also included bilateral damage to the amygdala, they could not rule out the possibility that combined damage to both the hippocampus and the amygdala was collectively responsible for the deficit.

To address the question of which medial temporal lobe structure or structures underlie declarative memory, experimentalists attempted to develop an animal 'model' of human amnesia. Nonhuman primates have been particularly valuable animal model systems because their general brain organization, intelligence, and dominant visual system mirror our own. The strategy was to determine if bilateral medial temporal lobe lesions in an animal model could reproduce the profound memory impairment seen in patient H.M. If the deficit could indeed be reproduced, this animal model then becomes a very useful tool with which to examine the effect of damage to specific structures or combinations of structures on memory. In 1978, Mishkin was the first to successfully develop an animal model of human amnesia in the monkey. This study focused on recognition memory, defined as the ability to determine that a given stimulus has been seen before. Recognition memory is a form of declarative memory known to be impaired in human amnesic patients. In his seminal study, Mishkin examined the effect on recognition memory of combined damage to the hippocampus and the amygdala, made to resemble the damage sustained by patient H.M. Animals with separate lesions of either

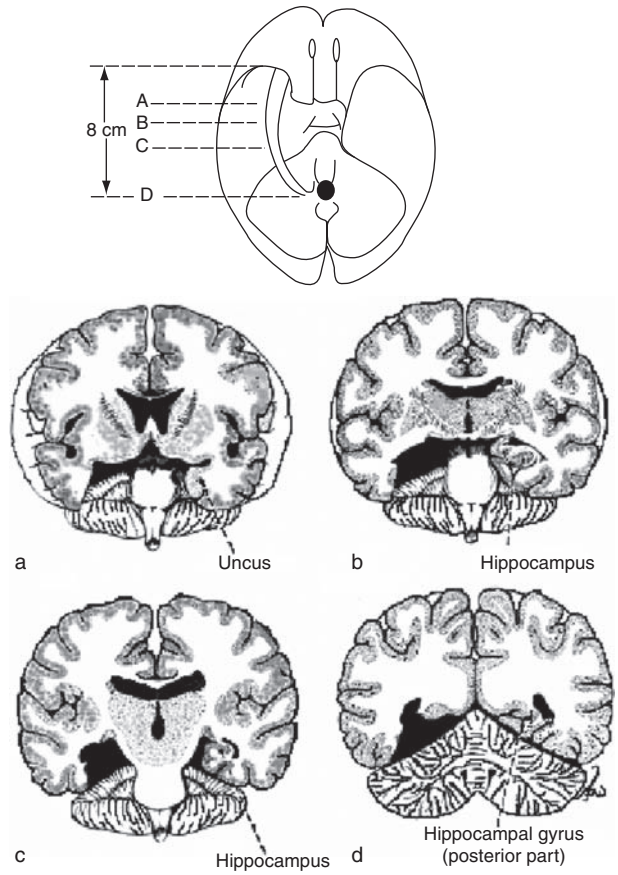

Figure 1 This is a reproduction of the original from the 1957 work of Scoville and Milner, illustrating Scoville's estimate of H.M.'s medial temporal lobe resection. The drawing at the top shows a ventral view of the human brain, indicating the estimated boundaries of the lesion. (a through d) Coronal sections arranged from anterior (a) to posterior (d), showing the estimated extent of the lesion. In these coronal sections, the left side shows the estimated amount of tissue removed while the right side is shown intact in order to label the different target brain structures. Reproduced from Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry* 20: 11–21, with permission from the BMJ Publishing Group.

the hippocampus or the amygdala were also tested. All of the lesions included some inadvertent damage to the surrounding cortical areas. Recognition memory was tested using a delayed nonmatching to sample task (Figure 2(a)). In this now widely used task, animals are first shown either a single sample stimulus or a series of to-be-remembered sample stimuli. After a variable delay interval, animals are presented with a sample stimulus and a novel stimulus. They must learn to choose the novel, or 'nonmatching,' stimulus to receive a reward. Mishkin's study showed convincingly that combined, but not separate, lesions of the hippocampus and amygdala produced a severe recognition memory deficit (Figure 2(b)). These results suggested that it was combined damage to

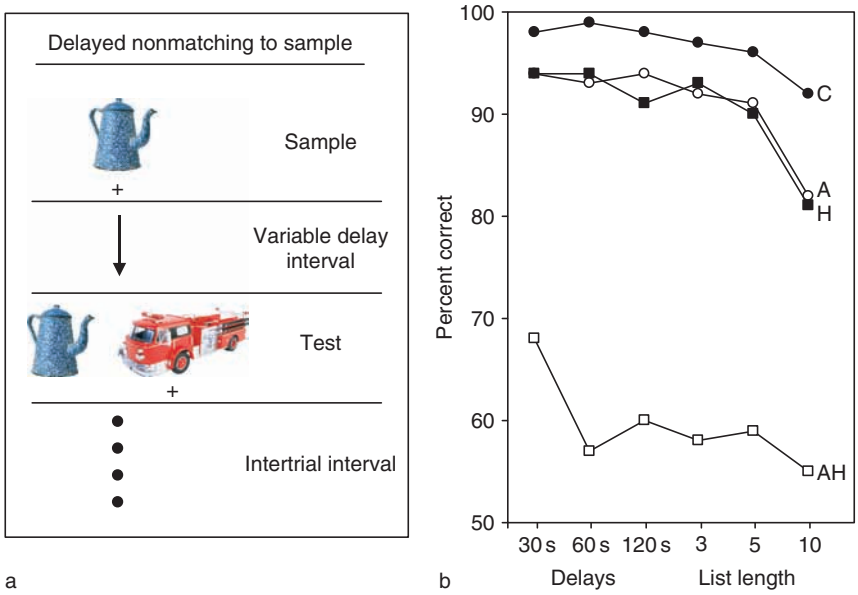

Figure 2 (a) This panel shows a schematic diagram of the delayed nonmatching to sample task, a benchmark task used for studying recognition memory in monkeys. Animals are first presented with a rewarded (+) 'sample' stimulus over a central food well. After displacing the sample object and retrieving the reward, there is a variable delay interval (typically between 8 s and 10 min), after which time the sample stimulus and a novel stimulus are presented over two lateral food wells. The animals learn to choose the novel, or 'nonmatching,' test stimulus to receive a reward (+). After the intertrial interval, the sequence is repeated using a novel pair of objects. The objects used in this task are typically 'junk' three-dimensional objects, including small toys, containers, or other similarly sized objects. (b) Illustration of the performance of groups of monkeys with combined damage to the amygdala and hippocampus (AH lesion group), or separate amygdala (A group) or hippocampal (H group) lesions on the delayed nonmatching to sample task. Delay interval indicates the time delay between the sample and the test phase of the task. List length refers to a condition where lists of 3, 5, or 10 sample stimuli were presented before the test phase. This manipulation increases memory demand by increasing interference between the sample stimuli. Note the profound memory deficit of the AH group even on the shortest delay interval used. Adapted from Mishkin M (1978) Memory in monkeys severely impaired by combined but not by separate removal of amygdala and hippocampus. *Nature* 273: 297–298.

the hippocampus and amygdala that was responsible for H.M.'s severe memory deficit.

Other scientists replicated the severe memory impairment observed in monkeys following large medial temporal lobe lesions (Figure 4(a)). This lesion has been referred to by Zola-Morgan and his colleagues as the H+A+ lesion, where H refers to hippocampal damage and A to amygdala damage, with the + symbols indicating inadvertent damage to the surrounding cortex made to access the underlying hippocampus and amygdala (Figure 3). Additional studies showed that the memory impairment exhibited by the H+A+ group paralleled key features of human amnesia. Specifically, it was shown that the impairment extended to both the visual and the somatosensory modalities, that it was long lasting, and that the memory impairment was selective for only certain forms of memory while other forms of memory known to be spared in human amnesia were intact.

Findings from human neuropsychological studies reported at about the same time that the effects of H+A+ lesions being characterized showed that bilateral damage limited to the hippocampus resulted in significant memory impairment. However, the memory impairment was not as severe as that of patient H.M. To confirm these observations in human patients experimentally, Zola-Morgan and his colleagues tested a group of monkeys with bilateral lesions limited to the hippocampus and the adjacent cortex (H+ group) on the delayed nonmatching to sample task. The H+ lesion group exhibited significant memory impairment on this task, but similar to the observations in human patients, the deficit was not as severe as the deficit seen following large H+A+ lesions (Figure 4(a)). Note that this finding differed from Mishkin's report of no memory impairment following hippocampal lesions (H lesion in Figure 2(b), which also included damage to the surrounding cortex), most likely because that study used the relatively less sensitive prelesion training strategy while the H+ group was tested using a more sensitive postlesion training strategy.

These findings suggested that additional damage sustained by the H+A+ group, but not the H+ group, must be underlying this more severe deficit. Based on the original report by Mishkin, the first

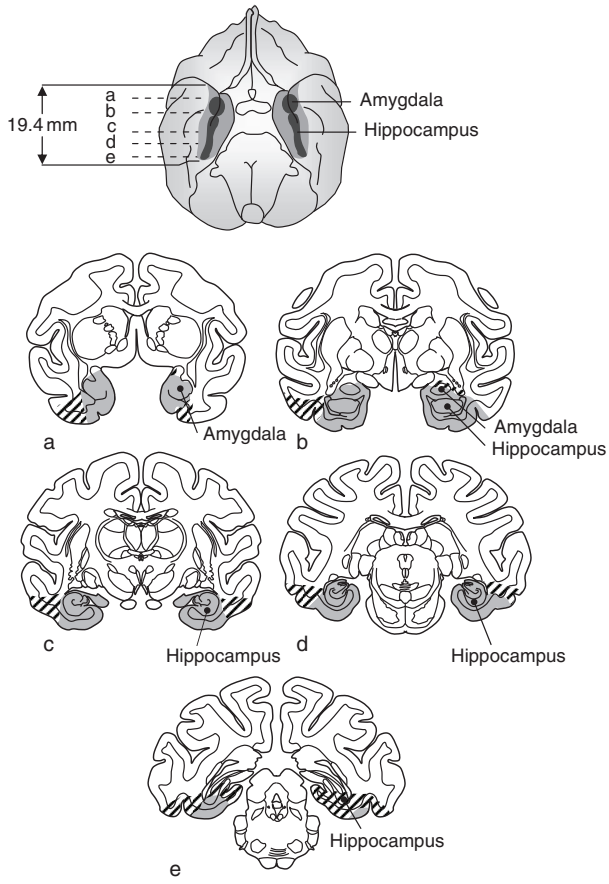

Figure 3 Schematic illustration of the extent of damage sustained in Zola-Morgan's H+A+ lesion group. Shown on the top is a ventral view of a macaque monkey brain. Shaded areas indicate the approximate location of the intended lesion. Also shown are the relative locations of the amygdala and hippocampus, which are situated just below the cortex of the ventral medial temporal lobe. Drawings (a)–(e) show coronal sections through the lesion site, illustrating the largest (stippled) and smallest (gray area) extent of damage. Note the similarities between the lesion illustrated here and the illustration of H.M.'s presumed lesion in **Figure 1**. Also note the similarity between the intended lesion of the AH group of Mishkin (**Figure 2(b)**) and the H+A+ group of Zola-Morgan and colleagues. The major difference was simply the nomenclature used by the two groups. Adapted from Zola-Morgan S and Squire LR (1985) Medial temporal lesions in monkeys impair memory on a variety of tasks sensitive to human amnesia. *Behavioral Neuroscience* 99: 22–34.

obvious candidate was the additional damage to the amygdala. To test this idea, Zola-Morgan and colleagues developed techniques that would allow selective damage of the amygdala without involvement of the surrounding cortex. In contrast to the prediction from Mishkin's study, Zola-Morgan showed that animals with bilateral lesions limited to the amygdala performed completely normally on the delayed nonmatching to sample task (**Figure 4(b)**). Even more convincingly, it was shown that the addition of a selective amygdala lesion to a H+ lesion (H+A lesion group) did not exacerbate the memory deficit observed following H+ lesions alone. These findings showed definitively that the additional damage of the amygdala was, in fact, not contributing to the severe memory impairment of the H+A+ group. This left the possibility that the inadvertent damage to the surrounding cortex was underlying the more severe memory deficit of the H+A+ group. This possibility was particularly surprising because ever since the original description of H.M., attention had focused almost exclusively on the likely role of the hippocampus and amygdala in memory. Cortical damage was acknowledged (**Figures 1 and 3**), but generally considered unrelated to the memory impairment. Two additional pieces of evidence emerging at the time further strengthened the idea that the surrounding cortical areas may be important for declarative memory.

The Entorhinal Cortex

The first piece of evidence came from anatomical studies examining the connections of the monkey entorhinal cortex, a brain area strongly associated with the hippocampus. In the early 1900s the famous Spanish neuroanatomist Ramón y Cajal first noted the massive projections from layers II and III of the entorhinal cortex to the dentate gyrus, part of the hippocampal region. Based on this strong anatomical projection, Cajal speculated that the functions of the entorhinal cortex and hippocampus would also be strongly linked. To emphasize this strong anatomical link, the term 'hippocampal formation' is often used to refer to the hippocampal region, including the dentate gyrus, hippocampus proper, and subicular complex, together with the entorhinal cortex.

Boundaries and connections In the monkey, the entorhinal cortex is situated medial to the rhinal sulcus on the ventromedial aspect of the anterior temporal lobe (**Figures 5 and 7**). Important early anatomical studies by Van Hoesen and colleagues described some of the cortical inputs to the monkey entorhinal cortex. However, the full complement of anatomical projections to this area was not known. To address this question, Insausti, Amaral, and colleagues analyzed a series of retrograde tracer injections made throughout the monkey entorhinal cortex. Retrograde tracers are transported from the axons that terminate in the injection site to the cell bodies of origin, where the tracer can be visualized. Plotting the locations of retrogradely labeled cell bodies resulting from a given injection site identifies all the brain areas that project to that site, while the number of labeled cells in each brain area provides an indication of the relative strength of that projection.

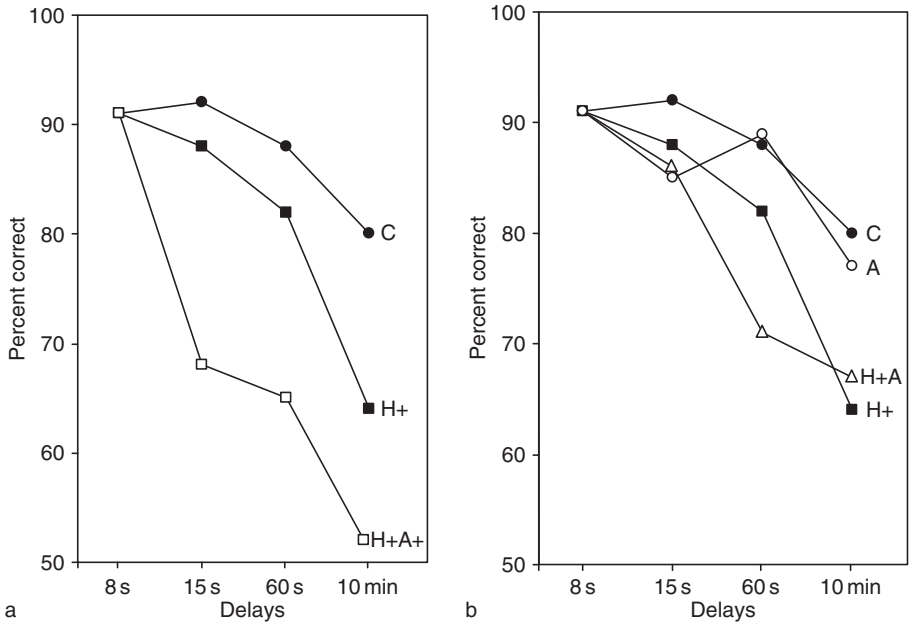

Figure 4 (a) This graph illustrates the memory deficit exhibited by the H+ and H+A+ group on the delayed nonmatching to sample task. Note that the H+ group exhibited a significant but less severe memory impairment compared to the H+A+ group. For both groups, the most severe impairment is seen at the longest delay intervals. (b) This graph illustrates the effect of selective amygdala damage on the delayed nonmatching to sample task. This graph makes two key points: first, that bilateral lesions limited to the amygdala that do not involve the surrounding cortex (A group) do not cause memory impairment, and second, that the addition of an A lesion to an H+ group (H+A) does not exacerbate the memory deficit of the original H+ group. This study showed definitively that the amygdala was not contributing to the severe visual recognition memory deficit exhibited by the original H+A+ group. C, control group. Adapted from Zola-Morgan S, Squire LR, and Amaral DG (1989) Lesions of the amygdala that spare adjacent cortical regions do not impair memory or exacerbate the impairment following lesions of the hippocampal formation. *Journal of Neuroscience* 9: 1922–1936; and Zola-Morgan S, Squire LR, Amaral DG, and Suzuki WA (1989) Lesions of perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. *Journal of Neuroscience* 9: 4355–4370.

Figure 5 illustrates the cortical connections of the monkey entorhinal cortex. A major finding was that the entorhinal cortex receives almost two-thirds of its direct cortical inputs from the surrounding perirhinal and parahippocampal cortices. Earlier anatomical studies suggested that these areas are higher-order association areas that receive convergent inputs from multiple sensory modalities. Other direct inputs to the entorhinal cortex include olfactory and somatosensory inputs from the piriform cortex and insula, respectively, spatial inputs from the retrosplenial cortex, auditory inputs from the superior temporal gyrus, and multimodal sensory inputs from the orbitofrontal cortex. The discovery that the major cortical inputs to the entorhinal cortex originate in the surrounding perirhinal and parahippocampal cortices suggested that these adjacent cortical areas were more intimately linked to the hippocampal formation than previously appreciated.

The second critical piece of evidence highlighting the possible role of the perirhinal and parahippocampal cortices in memory was a reexamination of the extent of the lesion in Zola-Morgan's original H+A+ group. This reexamination revealed that not only was the entorhinal cortex removed in this group, but also a much larger extent of both the perirhinal and the parahippocampal cortices was damaged than had originally been appreciated. These findings, together with the newly described connections of the entorhinal cortex, strongly suggested that the extra damage to the surrounding perirhinal and parahippocampal cortices was underlying the more severe memory impairment in the H+A+ group relative to the H+ group.

The Perirhinal and Parahippocampal Cortices

Lesion studies To test the hypothesis that the perirhinal and parahippocampal cortices contribute to memory, Zola-Morgan and colleagues examined the effects of bilateral lesions limited to these areas that spared the adjacent and underlying entorhinal cortex, hippocampus, and amygdala (PRPH lesion group) on a battery of memory tasks, including the delayed nonmatching to sample task. Consistent with this hypothesis, PRPH lesions produced a severe recognition memory impairment that was similar in severity to the deficit exhibited by the original H+A+ group (Figure 6). Like human amnesia, the impairment of

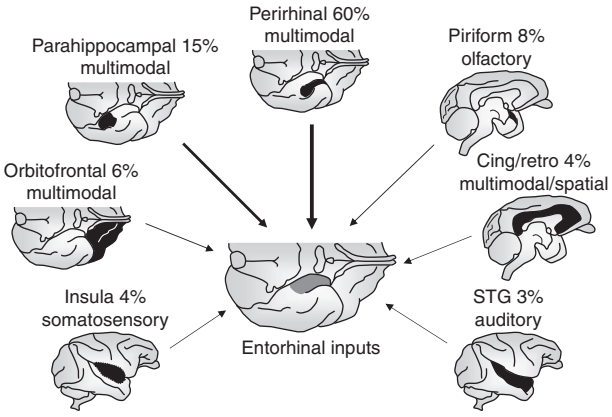

Figure 5 Illustration of the relative strength of the cortical inputs to the monkey entorhinal cortex. These data were taken from one representative case of retrograde tracer injection in the entorhinal cortex (Case IM-6 of Insausti and colleagues). The entorhinal cortex receives nearly two-thirds of its direct cortical input from the adjacent perirhinal and parahippocampal cortices. Other inputs include olfactory inputs, somatosensory inputs as well as strong multimodal input from the orbitofrontal cortex, and spatial input from the retrosplenial cortex as well as auditory input from the superior temporal gyrus (STG). Reproduced from Suzuki WA (1996) Neuroanatomy of the monkey entorhinal, perirhinal and parahippocampal cortices: Organization of cortical inputs and interconnections with amygdala and striatum. *Seminars in the Neurosciences* 8: 3–12.

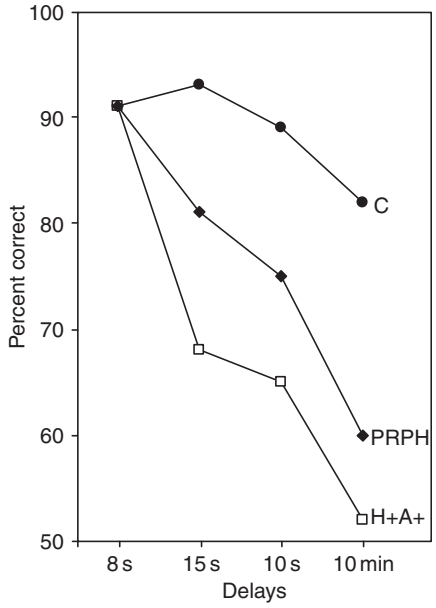

Figure 6 Illustration of the severe deficit exhibited by the perirhinal and parahippocampal (PRPH) group on the delayed nonmatching to sample task. The performance of the H+A+ group is shown for comparison. Adapted from Zola-Morgan S, Squire LR, Amaral DG, and Suzuki WA (1989) Lesions of the perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. *Journal of Neuroscience* 9: 4355–4370.

the PRPH lesion group encompassed multiple sensory modalities and was long lasting. Findings from numerous other studies confirmed that damage including the perirhinal and parahippocampal cortices resulted in severe memory impairment. These lesion findings provided definitive evidence that the perirhinal and parahippocampal cortices contributed to memory. Subsequent neuroanatomical studies focused on these areas helped clarify how these areas contribute to memory.

Boundaries and connections A major factor that slowed progress in identifying the memory functions of the perirhinal and parahippocampal cortices was the fact that at the time, very little was known about these cortical medial temporal lobe areas. For example, both the boundaries and the nomenclature applied to these areas had been highly inconsistent, making it difficult to know where one area ended and the next one began. Amaral and colleagues performed a preliminary reevaluation of the boundaries of these areas as part of their description of the cortical inputs to the entorhinal cortex. The current boundaries of the perirhinal and parahippocampal cortex are informed by data from tract tracing studies, cytoarchitectonic analyses of the cell staining and density patterns, as well as chemoarchitectonic analyses of neuropeptide staining patterns. Figure 7 illustrates the boundaries of the perirhinal and parahippocampal cortices on a ventral view of the monkey brain as well as in coronal sections. By comparing the boundaries of the perirhinal and parahippocampal cortices shown in the coronal sections in Figure 7 to the boundaries of the H+A+ lesion shown in the coronal sections of Figure 3, one can gain an appreciation for how much of the perirhinal and parahippocampal cortex was damaged in the original H+A+ group.

What are the connections of the perirhinal and parahippocampal cortices? Early anatomical studies suggested that both the perirhinal and parahippocampal cortices received convergent input from areas involved in processing multiple sensory modalities. But like the entorhinal cortex, it was unclear if all of the inputs to the perirhinal and parahippocampal cortices had been identified. To address this question, a series of neuroanatomical studies were undertaken to define the full complement of cortical inputs to these areas. The results from a series of retrograde tracer injections made throughout the perirhinal and parahippocampal cortices revealed two major findings. First, these areas receive a massive convergence of input from widespread areas, including visual, somatosensory, and auditory association areas, as well as higher order multimodal association areas, including the frontal lobe and the dorsal bank of the

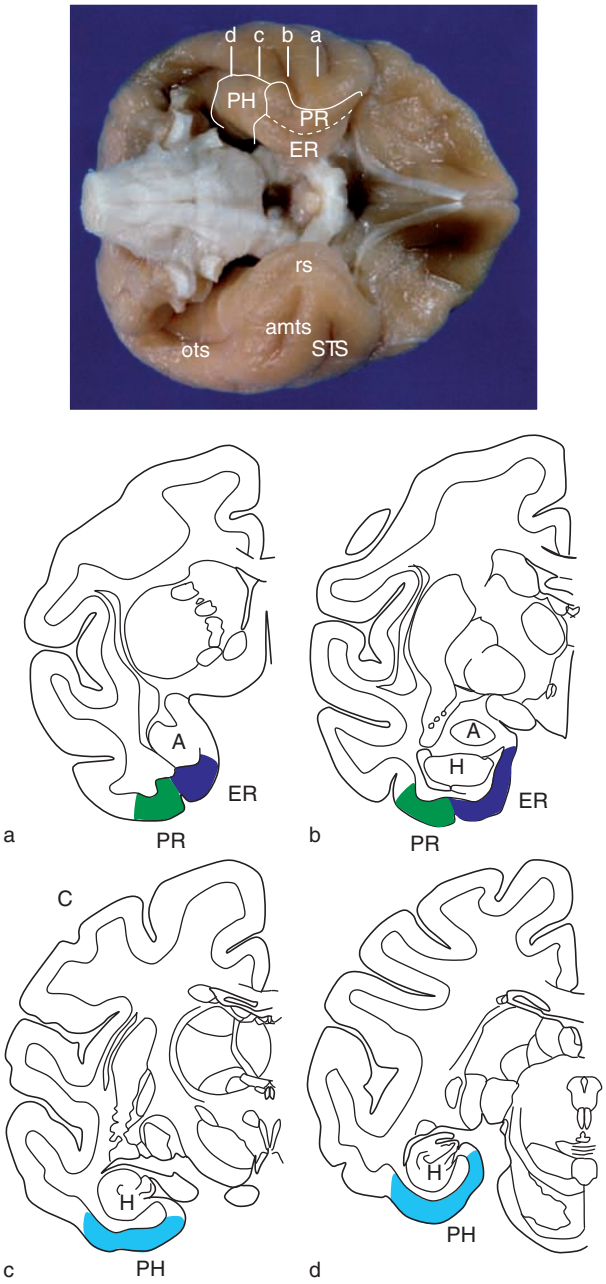

Figure 7 (Top) Photograph of the ventral view of a macaque monkey brain, where the boundaries of the entorhinal (ER), perirhinal (PR), and parahippocampal (PH) cortices are shown on the right hemisphere and the major sulci are labeled on the left hemisphere. The dashed line on the right hemisphere shows the location of the rhinal sulcus, which forms the boundary between the entorhinal cortex medially and the perirhinal cortex more laterally. Drawings (a)–(d) illustrate individual coronal sections through the perirhinal and parahippocampal cortices. The locations of these coronal sections are indicated by vertical lines in the ventral view. Note that by comparing this figure with **Figure 3**, one can gain an appreciation of the extent of the perirhinal and parahippocampal cortices damaged in the H+A+ group. A, amygdala; amts, anterior middle temporal sulcus; H, hippocampus; ots, occiptotemporal sulcus; rs, rhinal sulcus; STS, superior temporal sulcus.

superior temporal sulcus. Second, the perirhinal and parahippocampal cortices receive distinctly different sets of cortical inputs, suggesting potentially different roles in memory.

The major input to the perirhinal cortex (**Figure 8**) arises from the adjacent visual area TE, a major component of the 'ventral visual pathway' important for processing visual object information. The second strongest input to the perirhinal cortex originates in the parahippocampal cortex, suggesting strong cross-talk between the two areas. Other inputs-include multimodal sensory input from the dorsal bank of the superior temporal sulcus and orbitofrontal cortex as well as somatosensory input from the insular cortex. These projections are largely reciprocal. The powerful inputs from visual area TE, taken together with the devastating effects of lesions, including the perirhinal cortex on the delayed nonmatching to sample task, suggest that perirhinal cortex is particularly important for visual object memory.

Like the perirhinal cortex, the parahippocampal cortex also receives strong visual input, but this input comes from more posterior visual areas TEO and V4 (**Figure 9**). In contrast to the perirhinal cortex, with its prominent inputs from the highest order visual area TE, the parahippocampal cortex receives more prominent input from the so-called dorsal visual processing pathway, important for processing visual–spatial information. These spatial inputs arise from the retrosplenial cortex, posterior parietal cortex, dorsal bank of the superior temporal gyrus, and dorsolateral regions of the prefrontal cortex. These findings suggest that the parahippocampal cortex may be particularly important for spatial memory.

Neuroanatomy of Declarative Memory

Figure 10 illustrates the connections of the key medial temporal lobe structures important for declarative memory. These areas include the hippocampus, together with the surrounding and strongly interconnected entorhinal, perirhinal, and parahippocampal cortices. The connections of these areas exhibit a hierarchical organization, with the hippocampus acting as the ultimate site of information convergence. The hippocampus receives its major input and provides it major output to the entorhinal cortex, which receives its major cortical inputs and outputs from the surrounding perirhinal and parahippocampal cortices. The latter two areas act as a gateway of sensory information flow between the hippocampal formation, on the one hand, and widespread areas of neocortex, on the other.

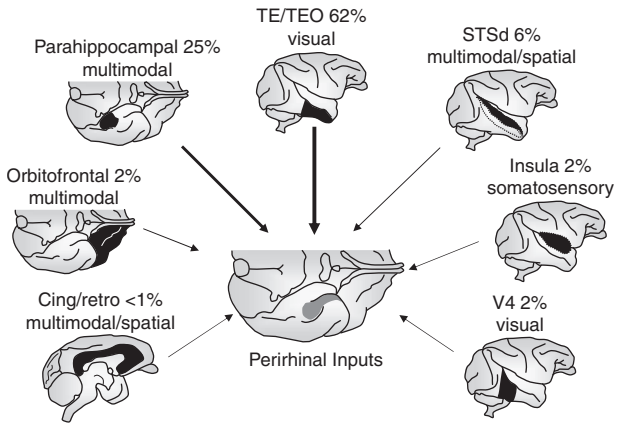

Figure 8 Schematic illustration of the relative strength of the cortical inputs to the perirhinal cortex. These data were taken from one representative case of retrograde tracer injection in the perirhinal cortex (Case M3–90 of Suzuki and Amaral). The strongest input to the perirhinal cortex arises from visual area TE and TEO, important for processing visual object information. The parahippocampal cortex also provides a very strong input to the perirhinal cortex. Other inputs include somatosensory input from the insula as well as multimodal input from the dorsal bank of the superior temporal sulcus (STSd) and orbitofrontal cortex. Thus, while the perirhinal cortex is a site of multimodal sensory convergence, its major input originates in visual areas TE and TEO. Reproduced from Suzuki WA (1996) Neuroanatomy of the monkey entorhinal, perirhinal and parahippocampal cortices: Organization of cortical inputs and interconnections with amygdala and striatum. *Seminars in the Neurosciences* 8: 3–12.

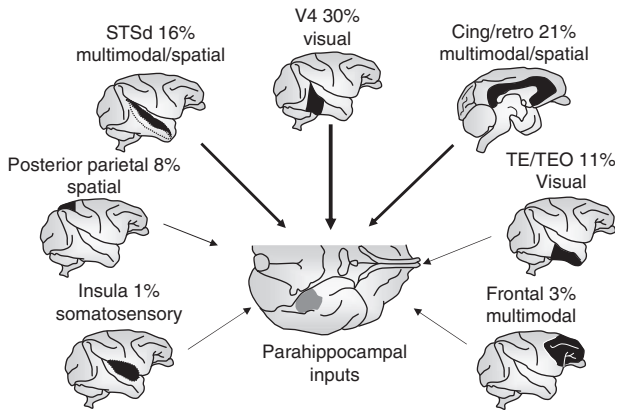

Figure 9 Schematic illustration of the relative strength of the cortical inputs to the parahippocampal cortex. The percentages shown here are taken from one prototypical case of retrograde trace injection in the parahippocampal cortex (Case M2–90 of Suzuki and Amaral). Like the perirhinal cortex, the parahippocampal cortex also receives strong visual input, but this input arises from more posterior visual area V4. Relative to the perirhinal cortex, the parahippocampal cortex receives disproportionately more direct input from the so-called dorsal visual processing stream, important for processing spatial information. These inputs arise from the retrosplenial cortex, the dorsal bank of the superior temporal sulcus, the posterior parietal lobe, and the dorsolateral prefrontal cortex. Reproduced from Suzuki WA (1996) Neuroanatomy of the monkey entorhinal, perirhinal and parahippocampal cortices: Organization of cortical inputs and interconnections with amygdala and striatum. *Seminars in the Neurosciences* 8: 3–12.

Understanding the anatomical connections of these medial temporal lobe areas provides the foundation for elucidating the role of these structures in declarative memory. For example, these anatomical connections suggest that the severe memory impairment exhibited by the PRPH group is not due exclusively to damage of the perirhinal and parahippocampal cortices, but also to the fact that the PRPH lesion cuts off virtually all of the cortical interconnections of both the entorhinal cortex and the hippocampus (Figure 10). Importantly, numerous studies have confirmed that damage limited to the hippocampal formation both in humans and in monkeys also produce significant, though less severe, memory deficits than the deficits seen following damage to the surrounding cortex. The powerful and reciprocal interconnections between these medial temporal lobe structures suggest that functional interactions between these areas may also be important for memory. This idea is supported by findings showing that the extent of medial temporal lobe damage correlates with the severity of impairment on the delayed nonmatching to sample task. However, the differences in the connections of these areas, particularly between the perirhinal and parahippocampal cortex, suggest that these cortical areas may also participate in different forms of memory, with perirhinal cortex important for object memory and the parahippocampal cortex important for spatial memory. This idea is also supported by experimental findings in monkeys.

The lesion and anatomical studies in the monkey model system also make strong prediction with respect to the actual brain damage sustained by patient H.M. Specifically, it is clear that the impairment in patient H.M. is substantially more severe than the memory impairment in patients with damage limited to the hippocampal formation bilaterally. Findings from animal studies predict that that patient H.M.'s lesion involves the hippocampus together with the surrounding cortex. Unfortunately, for many years the precise extent of H.M.'s lesion could not be confirmed with structural magnetic resonance imaging (MRI) scans because it was thought that the surgical clips left in place after his original surgery were not compatible with imaging. However, it was later determined that the clips would not pose a danger to H.M., and in 1997 Corkin and colleagues brought the study of the neuroanatomy of declarative memory full circle when they published a historic MRI analysis of H.M.'s brain lesion. This study confirmed that H.M. sustained bilateral damage to the amygdala as well as approximately the anterior

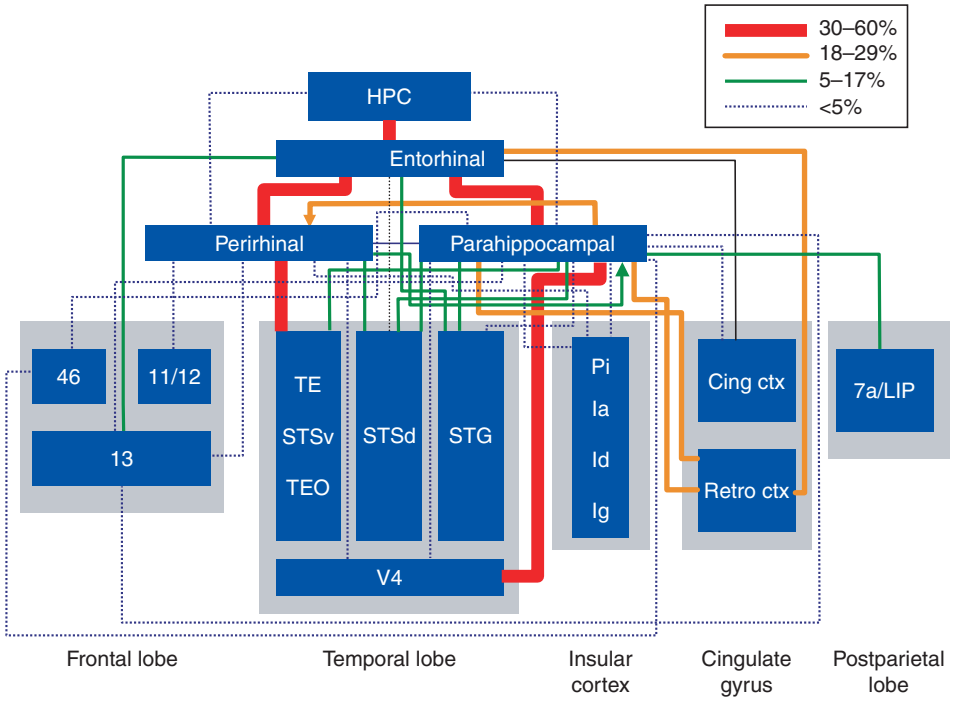

Figure 10 Schematic illustration of the connections of the hippocampus and the entorhinal, perirhinal, and parahippocampal cortices. The relative strength of the connections estimated from retrograde tracer experiments is illustrated by the color and thickness of the interconnecting lines (see key in upper right). Note the hierarchical organization, with the hippocampus acting as the pinnacle of the hierarchy. The hippocampus receives its major inputs and outputs from the entorhinal cortex, which receives its major inputs from the perirhinal and parahippocampal cortices. These latter two cortical areas act as the major gateway of sensory information flow between widespread areas of the neocortical mantel, on the one hand, and the hippocampal formation, on the other. Note the differences in the patterns of inputs between the perirhinal and parahippocampal cortices. These anatomical connections provide a framework for understanding the mnemonic functions of the medial temporal lobe. HPC, hippocampus; Ia, agranular portion of the insula; Id, dysgranular portion of the insula; Ig, granular portion of the insula; LIP, lateral intraparietal area; STG, superior temporal gyrus; STSd, dorsal bank of the superior temporal sulcus. Reproduced from Suzuki WA and Amaral DG (1994) Perirhinal and parahippocampal cortices of the macaque monkey: Cortical afferents. *Journal of Comparative Neurology* 350: 497–533.

half of the hippocampus. Moreover, consistent with the predictions from the animal studies, the MRI scans confirmed that the resection also involved all of the entorhinal cortex and parts of the perirhinal cortex bilaterally, though the parahippocampal cortex was spared. These findings, taken together with studies of other well-characterized amnesic patients with MRI or histological confirmation of the medial temporal lobe damage, support the idea that results from experimental animal studies are directly applicable to humans. These parallel findings reinforce the value of animal models in elucidating the principles of memory organization in the brain.

The discussion here has emphasized the lesion and anatomy studies in monkeys. However, important studies in the rodent model system have also provided fundamental contributions to our understanding of the medial temporal lobe areas important for declarative memory. Taken together, these experimental studies have not only identified the structures important for declarative memory, but have also laid the groundwork for the next generation of questions concerning how each of these medial temporal lobe areas contribute and/or interact to acquire new declarative memories. These studies also form the foundation for clinical strategies to treat various neurological conditions wherein declarative memory is impaired, including both normal human aging and Alzheimer's disease. Progress in the basic research of the brain basis of declarative memory will be critical in addressing these pressing clinical issues.

See also: Amnesia: Declarative and Nondeclarative Memory; Amygdala: Structure and Circuitry in Primates; Amygdala: Structure and Circuitry in Rodents and Felines; Animal Models of Amnesia; Entorhinal Cortex; Episodic Memory; Functional Amnesia; Hippocampus; Neuroanatomy Methods in Humans and Animals; Neuropsychological Testing; Neuropsychology: Theoretical Basis; Perirhinal Cortex.

Further Reading

Amaral DG, Insausti R, and Cowan WM (1987) The entorhinal cortex of the monkey: I. Cytoarchitectonic organization. *Journal of Comparative Neurology* 264: 326–355.

Corkin S, Amaral DG, GilbertoGonzalez R, et al. (1997) H.M.'s medial temporal lobe lesion: Findings from magnetic resonance imaging. *Journal of Neuroscience* 17: 3964–3979.

Insausti R, Amaral DG, and Cowan WM (1987) The entorhinal cortex of the monkey: II. Cortical afferents. *Journal of Comparative Neurology* 264: 356–395.

Lavenex P and Amaral DG (2000) Hippocampal-neocortical interaction: A hierarchy of associativity. *Hippocampus* 10: 420–430.

Mishkin M (1978) Memory in monkeys severely impaired by combined but not by separate removal of amygdala and hippocampus. *Nature* 273: 297–298.

Mishkin M and Murray EA (1994) Stimulus recognition. *Current Opinion in Neurobiology* 4: 200–206.

Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry* 20: 11–21.

Squire LR and Zola-Morgan S (1991) The medial temporal lobe memory system. *Science* 253: 1380–1386.

Stefanacci L, Buffalo EA, Schmolck H, et al. (2000) Profound amnesia after damage to the medial temporal lobe: A neuroanatomical and neuropsychological profile of patient E.P. *Journal of Neuroscience* 20: 7024–7036.

Suzuki WA (1996) Neuroanatomy of the monkey entorhinal, perirhinal and parahippocampal cortices: Organization of cortical inputs and interconnections with amygdala and striatum. *Seminars in the Neurosciences* 8: 3–12.

Suzuki WA and Amaral DG (1994) Perirhinal and parahippocampal cortices of the macaque monkey: Cortical afferents. *Journal of Comparative Neurology* 350: 497–533.

Suzuki WA and Amaral DG (2003) Where are the perirhinal and parahippocampal cortices? A historical overview of the nomenclature and boundaries applied to the primate medial temporal lobe. *Neuroscience* 120: 893–906.

VanHoesen GW and Pandya DN (1975) Some connections of the entorhinal (area 28) and perirhinal (area 35) cortices of the rhesus monkey. III. Efferent connections. *Brain Research* 95: 48–67.

Zola-Morgan S and Squire LR (1985) Medial temporal lesions in monkeys impair memory on a variety of tasks sensitive to human amnesia. *Behavioral Neuroscience* 99: 22–34.

Zola-Morgan S and Squire LR (1993) Neuroanatomy of memory. *Annual Review of Neuroscience* 16: 547–563.

Zola-Morgan S, Squire LR, Amaral DG, and Suzuki WA (1989) Lesions of the amygdala that spare adjacent cortical regions do not impair memory or exacerbate the impairment following lesions of the hippocampal formation. *Journal of Neuroscience* 9: 1922–1936.

Zola-Morgan S, Squire LR, Amaral DG, and Suzuki WA (1989) Lesions of perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. *Journal of Neuroscience* 9: 4355–4370.

Decoding Neuron Transcriptome by SAGE

Y C Kim and S M Wang, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Each somatic cell in the body of an individual contains the same genome with the same genes. However, the genes expressed in different cells can be very different in order to fulfill specific physiological activities. The brain is one of the most complex tissues, having different cell types, and the genes in the brain are expressed most actively in all human tissues. To fully understand brain function, the activity of each gene in brain tissue needs to be studied. One approach to doing this is to identify the direct gene products, the transcripts, expressed in the brain. Serial analysis of gene expression (SAGE) has been applied for this purpose.

Basics of SAGE

SAGE is a technique widely used for transcript identification. SAGE collects a short sequence, named a SAGE tag, from a long transcript to represent the whole transcript. Information from the SAGE tag can be used to detect and quantify the gene activities in the analyzed cells. The SAGE tags derived from transcripts of known genes represent the activity of known genes; the SAGE tags derived from unknown transcripts represent the activity of unknown genes, and such SAGE tags provide clue to discovering novel genes.

A standard procedure for SAGE tag generation is shown in Figure 1. Starting from an mRNA sample isolated from the target cells, cDNAs are synthesized and are digested by restriction enzyme NlaIII (CATG). The 3′-most cDNAs after the last CATG site are recovered, and SAGE tags are released after the last CATG site. SAGE tags are concatemerized, cloned, and sequenced. SAGE tags are extracted from the sequences with a copy number for each SAGE tag representing its quantitative information.

The standard SAGE is the most widely used; it is composed of 14 bp and is extracted as just described. Several variations of SAGE have been evolved for studying specific issues. LongSAGE elongates the standard SAGE tag to a 21 bp tag. The extended length improves the specificity of the transcripts and increases the mapping frequency of the tags to the reference genomic tags. SuperSAGE extends the length further to 26 bp, which allows superSAGE tags to be used directly as microarray probes and makes it useful for studying gene expression in species in which there is no genome or expressed sequence tag (EST) information. Cap analysis of gene expression (CAGE) collects 21 bp tags from the 5′ end of cDNAs to determine the 5′ end of transcripts and the location of the promoter. Gene identification signature (GIS) collects longSAGE tags from both the 5′ and the 3′ ends of the same transcript to determine the 5′ and 3′ ends of the detected transcripts.

The full-length cDNA and ESTs provide high specificity for the transcripts detected, but they have limited sensitivity for detecting low-abundance transcripts. SAGE has a high sensitivity for detecting both known and unknown transcripts present from high- to low-abundance levels. Comparing the full-length cDNA and ESTs, however, SAGE tags have a lower specificity for the detected original transcripts due to its short length. This could be improved by increasing the tag length into 3′ EST using the generation of longer cDNA fragments from SAGE tags for gene identification (GLGI) or rapid amplification of cDNA 3′ ends (3′-RACE) methods.

Three types of SAGE data analysis can be used to extract information from SAGE data.

1. *Quantitative analysis*. By using statistical analysis, the alternative gene activities among different samples represented by SAGE tag copies can be determined. Several standard statistical methods are used for this purpose, such as the Audic and Claverie pairwise test, Fisher's exact test, Greller and Tobin test, the *R* test, and general chi-squared tests. The Audic and Claverie test is efficient when detecting the differential expression of genes in pairwise comparisons. The chi-squared test is more efficient when there are variations with weakly expressed genes. An online program, IDEG6, provides several statistical programs specifically formatted for SAGE data analysis to identify differential gene expression.

2. *Determining the original gene for the detected SAGE tags*. This is achieved by comparing the experimental SAGE tags with a reference tag database based on the known transcript sequences. A reference SAGE database contains virtual SAGE tags extracted from known transcripts, including well-characterized full-length cDNA, mRNA, and ESTs. The widely used databases include the SAGEmap database at the National Center for Biotechnology Information (NCBI; the largest official collection of experimental SAGE tags exists on NCBI GEO for many species; the standard SAGE tag collection from human can be found as a platform data GPL4 and the longSAGE

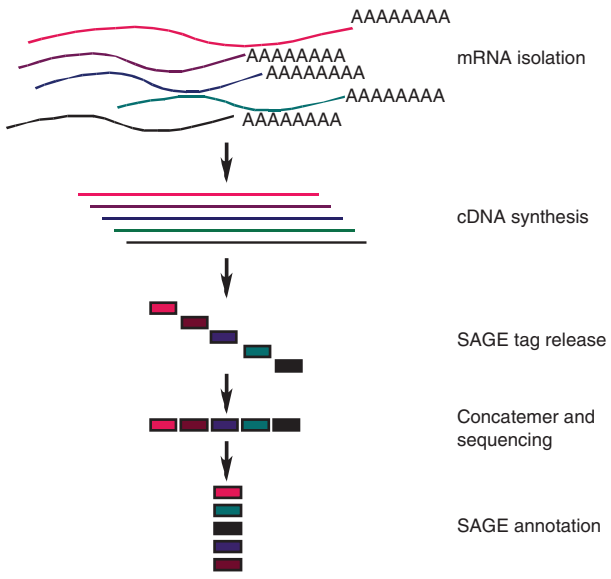

Figure 1 SAGE procedure. mRNA is isolated from the targeted sample. After being converted into cDNA, SAGE tags are released from the 3′ part after the last restriction site CATG and concatemerized for sequencing. SAGE tags are extracted from the sequences. Their gene origins are identified through matching to known transcript sequences.

tags from human samples are in GPL1485) and the SAGE Genie database at the National Cancer Institute (NCI). The SAGEmap tags are extracted from known transcripts or genes. SAGE Genie has a large collection of deposited tags from human and mouse, and it provides various tools for tag analysis. The experimental SAGE tags are compared with the virtual SAGE tags in the database. When an experimental SAGE tag matches a virtual SAGE tag, this implies that the experimental SAGE tag originated from the same mRNA or gene represented by the virtual SAGE tag. When the experimental SAGE tag does not match any virtual SAGE tag, the experimental SAGE tag is considered an unknown transcript.

3. *Identifying novel transcribed regions in the genome*. By mapping SAGE tags to the genome sequences, the origin of the transcripts detected by the SAGE tag can be identified. Those mapped to the intergenic regions or antisense regions, or intronic regions, represent potential novel transcribed regions in the genome. This is mainly achieved by using the longSAGE tags; the standard 14 bp SAGE tags are too restrictive for this type of study due to their short length.

When analyzing SAGE data, we need to pay attention to several factors that affect the quality of SAGE tag interpretation. These factors include experimental artifacts, sequencing errors, multiple loci in the genome containing the same SAGE tag sequence, incomplete annotated reference databases, transcript isoforms, alternate splicing, and single nucleotide polymorphisms (SNPs). When planning downstream functional studies, it is better to use different methods to confirm the results, for example, to use Northern blotting or real-time polymerase chain reaction (PCR) to confirm that the candidate genes are differentially expressed in different samples and to use reverse transcription (RT)-PCR to confirm the presence of transcripts from the novel gene candidates.

Applying SAGE to Brain Transcriptome Studies

Since SAGE was introduced, it has been increasingly used for transcriptome study. One of the major tissue types analyzed by SAGE is the human brain. This has been largely achieved by Gregory Riggins's laboratory, supported by the NCI's Cancer Genome Anatomy Project.

The Total SAGE Tags Collected from the Human Brain

A total of 95 standard SAGE and 8 longSAGE libraries were constructed from RNA samples of human brain sources. A total of 8 million SAGE tags have been collected from those libraries, which account for 25% of the 30 million SAGE tags collected so far that are of human origin. Table 1 summarizes the brain SAGE data collected from each sample. The accumulation curve (Figure 2) shows that even using this scale of SAGE tag collection, a saturated stage of transcript detection in brain has not been reached. This implies that more unknown transcripts expressed in the brain remain to be identified.

The SAGE Tags from Normal Brains and Brains with Tumors

Of the 103 SAGE libraries, 12 are from normal human tissues and 91 are from different types of brain tumors (Table 2). Most of the standard SAGE samples are from brain cancers, such as astrocytoma, ependymoma, glioblastoma, medulloblastoma, meningioma, and oligodendroglioma. The information from the normal tissue provides the gene expression information in normal conditions, and the information from the tumors is valuable for studying tumorigenesis and for identifying tumor-specific markers for diagnosis and classification.

Comparing Brain SAGE Tags to Known Transcript Sequences

Comparing the SAGE tags with the known human transcript sequences shows that, in the individual SAGE

Table 1 SAGE tags collected from brain tissues

SAGE library	*Total SAGE tags*	*Unique SAGE tags*
LSAGE_Brain_fetal_normal_B_S1	305 546	113 723
LSAGE_Brain_glioblastoma_CL_28d	29 873	14 725
LSAGE_Brain_glioblastoma_CL_4d	28 166	14 529
LSAGE_Brain_glioblastoma_CL_stem_cells	26 381	13 780
LSAGE_Brain_Medulloblastoma_CL_425C2	62 642	27 165
LSAGE_Brain_Medulloblastoma_CL_UW228-N	94 937	39 295
LSAGE_Brain_Medulloblastoma_CL_UW228-V2	80 827	34 482
LSAGE_Brain_normal_substantia_nigra_B_1	41 773	20 528
SAGE_Brain_astrocyte_normal_CL_NHA5	51 481	20 955
SAGE_Brain_astrocytoma_grade_I_B_H1043	75 922	30 880
SAGE_Brain_astrocytoma_grade_II_B_H127	114 489	31 165
SAGE_Brain_astrocytoma_grade_II_B_H359	105 764	35 555
SAGE_Brain_astrocytoma_grade_II_B_H388	106 285	34 081
SAGE_Brain_astrocytoma_grade_II_B_H501	128 309	38 047
SAGE_Brain_astrocytoma_grade_II_B_H516	108 116	42 250
SAGE_Brain_astrocytoma_grade_II_B_H518	116 022	36 310
SAGE_Brain_astrocytoma_grade_II_B_H530	102 439	34 715
SAGE_Brain_astrocytoma_grade_II_B_H563	88 568	30 054
SAGE_Brain_astrocytoma_grade_III_B_343	100 158	29 620
SAGE_Brain_astrocytoma_grade_III_B_407	108 312	32 819
SAGE_Brain_astrocytoma_grade_III_B_439	107 824	34 620
SAGE_Brain_astrocytoma_grade_III_B_584	103 008	30 772
SAGE_Brain_astrocytoma_grade_III_B_828	99 939	32 109
SAGE_Brain_astrocytoma_grade_III_B_H1020	51 573	30 738
SAGE_Brain_astrocytoma_grade_III_B_H1055	109 886	37 250
SAGE_Brain_astrocytoma_grade_III_B_H272	96 059	34 171
SAGE_Brain_astrocytoma_grade_III_B_H970	106 982	35 207
SAGE_Brain_astrocytoma_grade_III_B_R140	118 733	36 465
SAGE_Brain_astrocytoma_grade_III_B_R927	107 344	33 747
SAGE_Brain_ependymoblastoma_B_819	33 975	14 845
SAGE_Brain_ependymoma_B_1150	62 373	23 324
SAGE_Brain_ependymoma_B_1394	56 314	22 956
SAGE_Brain_ependymoma_B_239	46 653	20 459
SAGE_Brain_ependymoma_B_H580	68 614	25 883
SAGE_Brain_ependymoma_B_R1023	122 690	40 027
SAGE_Brain_ependymoma_B_R353	73 822	27 211
SAGE_Brain_ependymoma_B_R455	51 825	19 611
SAGE_Brain_ependymoma_B_R510	84 073	30 595
SAGE_Brain_ependymoma_B_R512	75 379	27 374
SAGE_Brain_ependymoma_B_R582	52 189	17 282
SAGE_Brain_ependymoma_B_R628	120 431	39 836
SAGE_Brain_fetal_normal_B_S1	305 546	79 405
SAGE_Brain_glioblastoma_B_GBM1062	59 762	15 504
SAGE_Brain_glioblastoma_B_H1110	68 986	25 812
SAGE_Brain_glioblastoma_B_H1353	124 805	37 773
SAGE_Brain_glioblastoma_B_H1371	49 338	22 909
SAGE_Brain_glioblastoma_B_H1425C	88 990	28 204
SAGE_Brain_glioblastoma_B_H833	100 600	34 143
SAGE_Brain_glioblastoma_B_pooled	56 428	24 020
SAGE_Brain_glioblastoma_B_R20	101 053	35 474
SAGE_Brain_glioblastoma_B_R336	102 322	34 762
SAGE_Brain_glioblastoma_B_R70	99 099	32 108
SAGE_Brain_glioblastoma_CL_28d	110 624	33 890
SAGE_Brain_glioblastoma_CL_4d	115 212	37 454
SAGE_Brain_glioblastoma_CL_H392	55 990	19 417
SAGE_Brain_glioblastoma_CL_H54+EGFRvIII	56 982	19 079
SAGE_Brain_glioblastoma_CL_H54+LacZ	66 908	20 539
SAGE_Brain_glioblastoma_CL_stem_cells	108 414	33 840
SAGE_Brain_glioblastoma_control_CL_H247	60 428	19 237
SAGE_Brain_glioblastoma_CS_TA2	46 678	21 568

Continued

Table 1 Continued

SAGE library	*Total SAGE tags*	*Unique SAGE tags*
SAGE_Brain_glioblastoma_hypoxia_CL_H247	71 765	24 770
SAGE_Brain_medulloblastoma_B_1273	38 614	17 236
SAGE_Brain_medulloblastoma_B_96–04-P019	52 645	22 017
SAGE_Brain_medulloblastoma_B_97–05-P015	69 971	24 631
SAGE_Brain_medulloblastoma_B_97–05-P312	74 295	28 029
SAGE_Brain_medulloblastoma_B_98–04-P117	32 570	14 766
SAGE_Brain_medulloblastoma_B_98–04-P494	43 068	16 247
SAGE_Brain_medulloblastoma_B_98–05-P040	89 258	34 995
SAGE_Brain_medulloblastoma_B_98–05-P608	48 451	22 675
SAGE_Brain_medulloblastoma_B_98–09-P558	85 984	33 598
SAGE_Brain_medulloblastoma_B_98–13-P301	45 342	17 313
SAGE_Brain_medulloblastoma_B_C609	74 612	32 456
SAGE_Brain_medulloblastoma_B_DL5	83 671	28 721
SAGE_Brain_medulloblastoma_B_DL7	68 392	25 370
SAGE_Brain_medulloblastoma_B_H1322	59 498	28 969
SAGE_Brain_medulloblastoma_B_H1413	61 853	22 833
SAGE_Brain_medulloblastoma_B_H275	72 318	26 151
SAGE_Brain_medulloblastoma_B_H306	60 454	24 714
SAGE_Brain_medulloblastoma_B_H484	57 469	25 712
SAGE_Brain_medulloblastoma_B_H876	67 404	25 648
SAGE_Brain_medulloblastoma_B_H972	85 376	30 675
SAGE_Brain_Medulloblastoma_CL_425C2	62 642	21 950
SAGE_Brain_medulloblastoma_CL_4-HCR	66 599	26 573
SAGE_Brain_medulloblastoma_CL_D283	66 316	25 279
SAGE_Brain_medulloblastoma_CL_H341	43 920	15 220
SAGE_Brain_medulloblastoma_CL_mhh-1	47 858	18 542
SAGE_Brain_Medulloblastoma_CL_UW228-N	94 937	29 409
SAGE_Brain_Medulloblastoma_CL_UW228-V2	80 827	26 713
SAGE_Brain_meningioma_grade_I_B_SF4271	69 253	24 434
SAGE_Brain_meningioma_grade_I_B_SF4465	54 647	19 464
SAGE_Brain_meningioma_grade_II_B_SF2176	48 711	16 920
SAGE_Brain_meningioma_grade_II_B_SF3622	44 857	16 264
SAGE_Brain_meningioma_grade_III_B_SF2366	59 786	20 161
SAGE_Brain_normal_cerebellum_B_1	50 385	24 357
SAGE_Brain_normal_cerebellum_B_BB542	40 500	19 163
SAGE_Brain_normal_cortex_B_BB542	94 233	31 683
SAGE_Brain_normal_cortex_B_pool6	62 451	25 053
SAGE_Brain_normal_leptomeninges_B_AL2	72 473	26 220
SAGE_Brain_normal_peds_cortex_B_H1571	77 554	24 161
SAGE_Brain_normal_substantia_nigra_B_1	41 773	17 106
SAGE_Brain_normal_thalamus_B_1	24 015	10 115
SAGE_Brain_oligodendroglioma_B_1001	32 442	14 369
SAGE_Brain_oligodendroglioma_B_H988	27 864	13 982
Total	8 010 914	692 735

SAGE data are from the SAGE Genie database. Total unique SAGE tags count the SAGE tags that overlap among different libraries only once.

tag set, between 22% and 74% of SAGE tags match a known transcript sequences and that the overall matching rate is 36% (Table 3). The matched sequences are included in a total 32 968 UniGene clusters that include most of the genes annotated in the human genome, indicating that most of the known human genes are expressed in brain. Significantly, between 24% and 78% of SAGE tags in individual SAGE tag sets do not match known transcript sequences, with an overall unmatched rate of 64%. Although a certain portion of these might be due to experimental artifacts, they contain a substantial number of unknown transcripts or the genes in the genome that remain to be identified. The quantitative distribution of the mapped SAGE tags shows that most of the detected transcripts are present at a low-abundance level (Figure 3).

Mapping SAGE Tags to Human Genome Sequences

The longSAGE tags can be mapped specifically to the genome; therefore, we can use the longSAGE

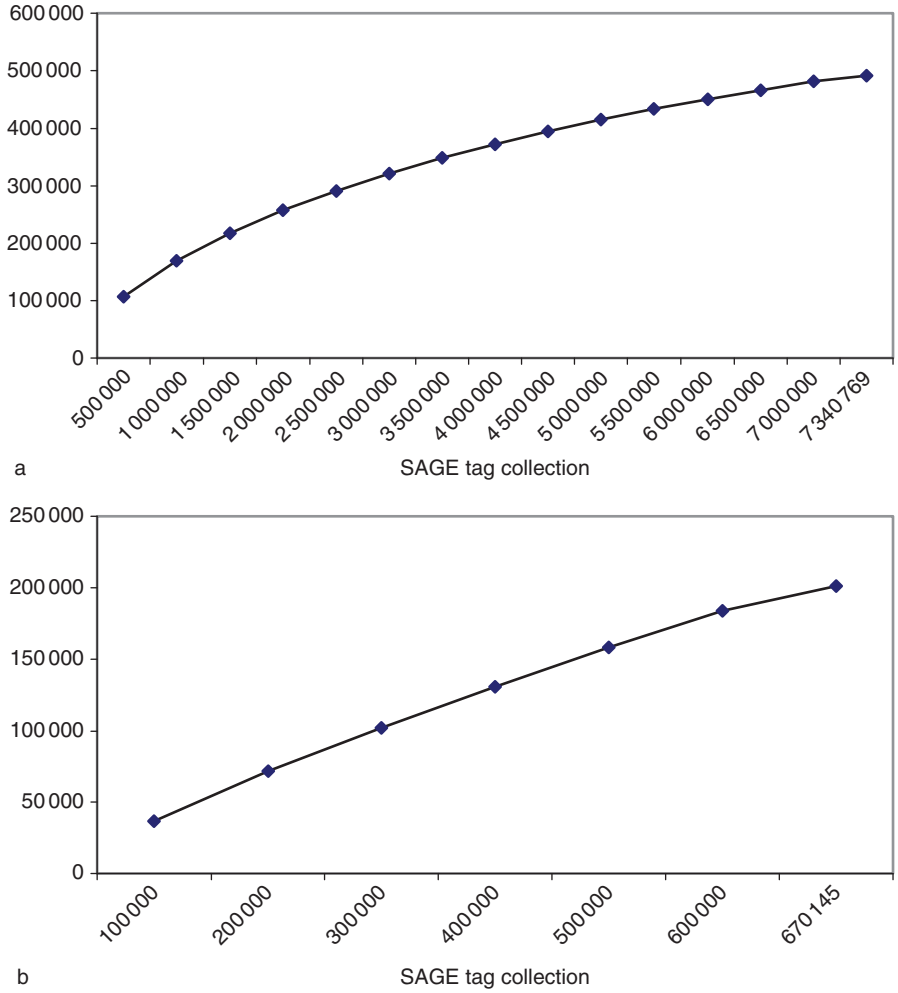

Figure 2 Accumulation of brain SAGE tags: (a) regular 14 bp SAGE tags; (b) longSAGE tags. Total brain SAGE tags were used for the plotting. SAGE tags were divided into 100 000 per set and the unique tags from each set were compared with the previous ones. Only the new unique tags identified in later SAGE tag set were kept to indicate the detection of new transcripts. The curves show that the saturated stage of transcript detection in the brain has not been reached.

Table 2 Summary of SAGE tags from different brain tissues

Tissue type	*Libraries*	*Total tags*	*Unique tags*
SAGE tags			
Normal	10	820 411	155 895
Astrocytoma	20	2 055 732	248 029
Ependymoma	12	848 338	139 266
Glioblastoma	19	1 544 384	208 623
Medulloblastoma	27	1 734 344	252 366
Meningioma	5	277 254	63 530
Oligodendroglioma	2	60 306	24 555
Subtotal	95	7 340 769	491 287
LongSAGE tags			
Normal	2	347 319	124 835
Glioblastoma	3	84 420	33 679
Medulloblastoma	3	238 406	81 295
Subtotal	8	670 145	201 448
Total	103	8 010 914	692 735

to identify the genome origin of the detected transcripts. There are 201 448 unique longSAGE tags from brain tissues. Mapping these brain longSAGE tags to human genome sequences shows that, of the longSAGE tags mapped to the genome, 15% map to the loci that have no annotated gene information and that represent potential novel transcribed regions in the genome (Table 4).

In conclusion, brain SAGE data provide one of the richest sources of transcriptome information. The brain SAGE data will be a valuable resource for studying brain biology and pathology and for studying human genome annotation.

See also: Gene Expression in Normal Aging Brain; Gene Expression Dysregulation in CNS Pathophysiology; Tumors of the Brain and the Spinal Cord.

Table 3 Matching brain SAGE tags to known transcript sequences

Type of SAGE tags	*Classification*			
	Unique tags	*Matched tags*	*Matched UniGene clusters*	*Novel tags*
SAGE tags				
Normal	155 895	90 192 (58)	24 656	65 703 (42)
Astrocytoma	248 029	134 352 (54)	27 697	113 677 (46)
Ependymoma	139 266	85 501 (61)	24 158	53 765 (39)
Glioblastoma	208 623	118 207 (57)	26 222	90 416 (43)
Medulloblastoma	252 366	133 806 (53)	27 474	118 560 (47)
Meningioma	63 529	41 547 (65)	18 440	21 982 (35)
Oligodendroglioma	24 555	18 209 (74)	13 275	6 346 (26)
Subtotal	491 287	220 770 (45)	31 833	270 517 (55)
LongSAGE tags				
Normal	124 835	36 836 (30)	16 378	87 999 (70)
Glioblastoma	33 679	15 202 (45)	16 276	18 477 (55)
Medulloblastoma	81 295	27 882 (34)	12 877	53 413 (66)
Subtotal	201 448	51 404 (22)	18 167	185 803 (78)
Total	692 735	272 174 (39)	32 974	420 561 (61)

SAGE tags were matched to the SAGEmap database, http://www.ncbi.nlm.nih.gov. Values in parentheses are percentages.

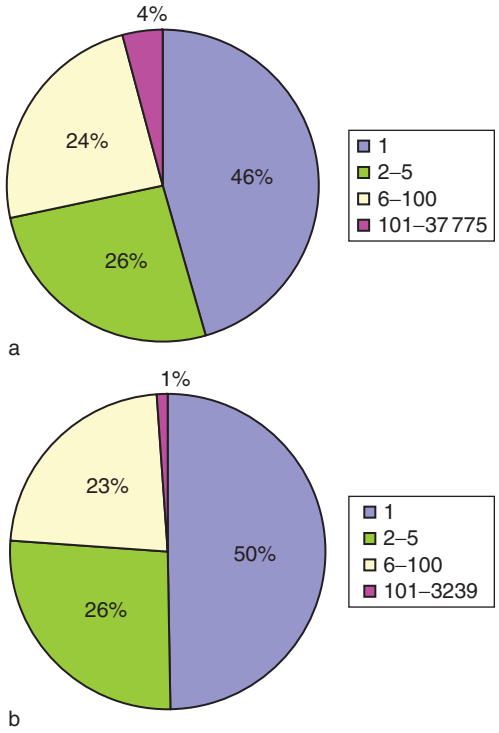

Figure 3 Quantitative distribution of mapped SAGE tags: (a) matched regular SAGE tags; (b) matched longSAGE tags. The SAGE tags and longSAGE tags that matched known transcript sequences were used for the analysis; each SAGE tag was grouped based on its copy number. The results show that the majority of the detected transcripts were expressed at the low-abundance levels, whereas only smaller numbers of transcripts were present at the higher-abundance levels.

Table 4 Novel transcribed region in the human genome mapped by novel longSAGE tags

Classification	*Number*[a]
Total brain unique longSAGE tags	201 448
Not mapped to genome	113 822
Mapped to genome	87 626 (100)
Matched to known gene	74 916 (85)
Novel	12 710 (15)

[a]Values in parentheses are percentages.

Further Reading

Boon K, Osorio EC, Greenhut SF, et al. (2002) An anatomy of normal and malignant gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 99: 11287–1192.

Matsumura H, Ito A, Saitoh H, et al. (2005) SuperSAGE. *Cell Microbiology* 7: 11–18.

Saha S, Sparks AB, Rago C, et al. (2002) Using the transcriptome to annotate the genome. *Nature Biotechnology* 20: 508–512.

Shiraki T, Kondo S, Katayama S, et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America* 100: 15776–15781.

Velculescu VE, Zhang L, Vogelstein B, and Kinzler KW (1995) Serial analysis of gene expression. *Science* 270: 484–487.

Wang SM (ed.) (2005) *SAGE: Current Technologies and Applications*. Norfolk: Horizon Scientific Press.

Wang SM (2007) Understanding SAGE. *Trends in Genetics* 23: 42–50.

Wei CL, Ng P, Chiu KP, Wong CH, et al. (2004) 5′ Long serial analysis of gene expression (longSAGE) and 3′ longSAGE for transcriptome characterization and genome annotation. *Proceedings of the National Academy of Sciences of the United States of America* 101: 11701–11706.

Relevant Websites

http://telethon.bio.unipd.it – IDEG6.
http://www.ncbi.nlm.nih.gov – NCBI GEO and SAGEmap.
http://cgap.nci.nih.gov – SAGE Genie.

Deep Brain Stimulation

C C McIntyre, Cleveland Clinic Foundation, Cleveland, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Clinical Applications of Deep Brain Stimulation

During the past two decades, deep brain stimulation (DBS) has evolved from an experimental technique to a well-established therapy for a range of medically refractory neurological disorders. To date, the most effective application of DBS technology has been in the treatment of movement disorders. DBS can alleviate symptoms and improve the quality of life for disabling diseases such as Parkinson's disease (PD), essential tremor (ET), and dystonia in appropriately selected patients. Surgical interventions for movement disorders have a long history, beginning with early studies that used lesions to eliminate activity in localized brain regions. In the mid-1980s, Benabid and colleagues found that stimulation of the brain had effects similar to lesioning, specifically that high-frequency stimulation (~100 Hz) of the motor thalamus produced tremor arrest. These early studies led to the advent of DBS as an experimental therapy that eventually overtook lesioning as the functional neurosurgery treatment of choice for movement disorders. Thalamic DBS has virtually replaced ablative lesions of the thalamus for the treatment of ET. DBS of the subthalamic nucleus or globus pallidus internus (GPi) has largely replaced pallidotomy in the treatment of the cardinal motor features of PD (tremor, rigidity, and bradykinesia). GPi DBS has emerged as an effective therapy for dystonia. In addition, multiple pilot studies are examining the utility of DBS for epilepsy, obsessive–compulsive disorder, Tourette's syndrome, and depression.

DBS Components and Implantation

Currently, Medtronic (Minneapolis, MN, USA) produces the only commercially available Food and Drug Administration-approved DBS medical device systems. However, as a result of the growing clinical demand for DBS technology, additional commercial entities are beginning to enter the marketplace. The fundamental components of the Medtronic DBS system consist of a quadripolar stimulation lead and an implanted pulse generator (IPG). The stimulation lead is a cylindrical column 1.27 mm in diameter with four narrowly spaced platinum–iridium electrode contacts near the tip of the lead. Each electrode contact is a cylinder 1.5 mm in length and 1.27 mm in diameter, resulting in a surface area of ~6 mm^2. The IPG is battery powered, uses constant voltage stimulation, and the stimulation parameters can be adjusted via telemetry. Typically, a single electrode contact is used to deliver the stimulation and the clinician has the opportunity to adjust the amplitude (voltage), width, and frequency of the stimulus pulses.

The implantation procedures for a DBS system are different at different centers, but in general the procedure combines noninvasive imaging, stereotactic neurosurgery, and microelectrode neural recording and stimulation. The patient's head is fitted with a stereotactic head frame and then undergoes computed tomography (CT) and/or magnetic resonance imaging (MRI) before entering the operating room (**Figure 1(a)**). The results of the CT/MRI scan are merged with a brain atlas, and the neurosurgeon identifies the x, y, and z location of the target brain structure within the context of the stereotactic frame (**Figure 1(b)**). An electrode trajectory is then identified to avoid major blood vessels and ventricles, a burr hole is drilled in the skull, and a microelectrode drive system is attached to the stereotactic frame. The microelectrode is inserted into the brain along the planned trajectory, and detailed neural recordings are made as the probe is advanced in micrometer increments (**Figure 1(c)**). The microelectrode recordings from the different deep brain structures exhibit characteristic firing patterns that the neurosurgeon and/or neurophysiologist use to further define the brain region where the DBS electrode will be implanted. Integration of the neuroanatomical (imaging and brain atlas) and neurophysiological (microelectrode) data allows definition of a target location in the stereotactic reference frame for placement of the permanent DBS electrode (**Figure 1(d)**). The DBS electrode is implanted, the individual electrode contacts are tested, and their location is verified by intraoperative X-ray. The burr hole is capped with the DBS lead left externalized, traditionally concluding the first surgical day, which lasts approximately 4 h for each lead implanted. In a subsequent surgical procedure, the electrode lead is internalized and connected to the IPG, which is implanted in the chest just below the collarbone.

Extracellular Electrical Stimulation

The fundamental purpose of DBS is to modulate pathological neural activity with extracellular electric fields. When stimulating within the central nervous system, the electrode is placed within a complex electrical environment in which there exist three general

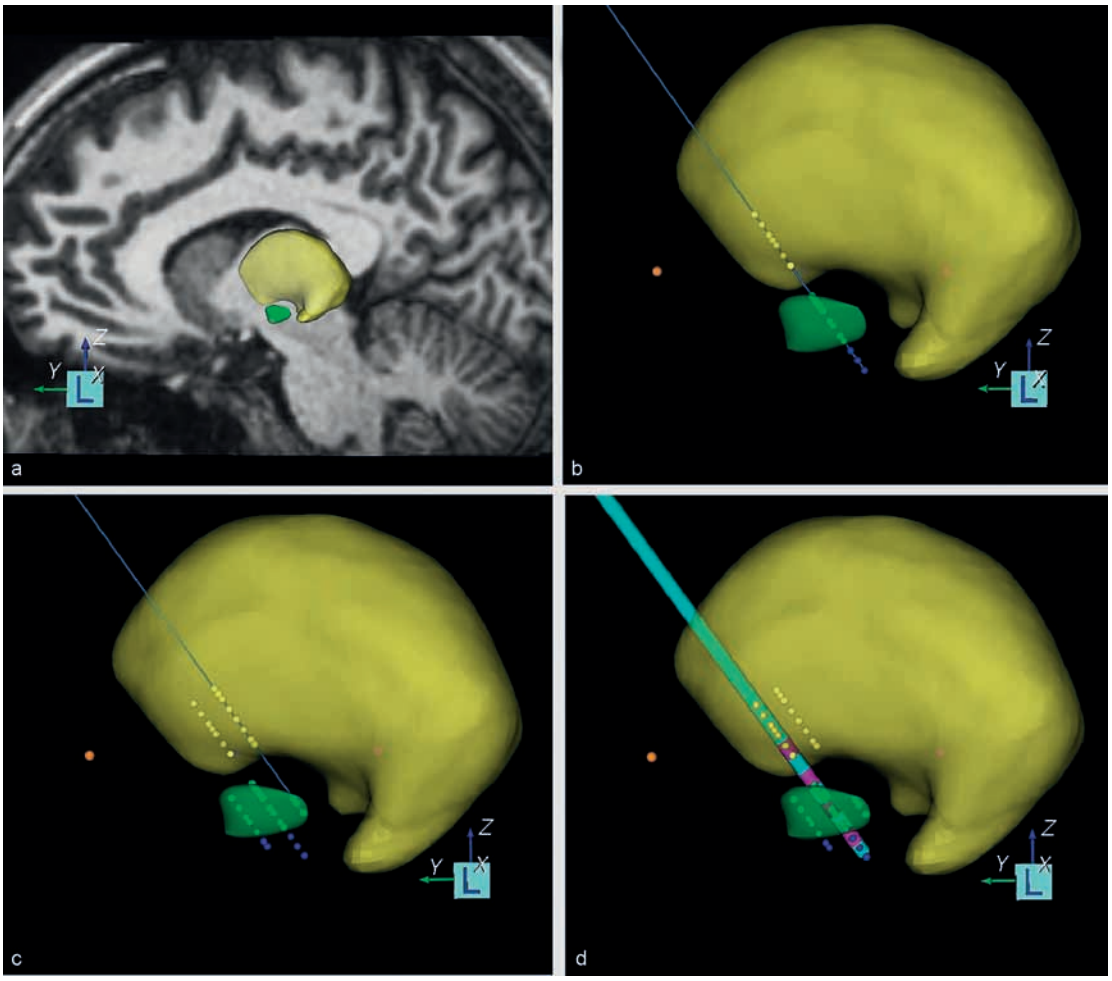

Figure 1 DBS electrode implantation. (a) DBS electrodes are commonly implanted into the thalamus (yellow volume) or subthalamic nucleus (green volume). (b) Following identification of a target in reference to the stereotactic frame and MRI, microelectrode recording of neural activity is acquired along a trajectory that goes through the target (blue line). Single-unit extracellular recordings of neurons representative of thalamus (yellow dots), subthalamic nucleus (green dots), and substantial nigra (blue dots) are recorded to define the electrophysiological borders of the deep brain structures. (c) Multiple trajectories may be evaluated to map the area around the target. (d) The DBS electrode (pink electrode contacts) is placed within the anatomical and electrophysiological environment of the subthalamic nucleus.

classes of neurons that can be affected: local cells, afferent inputs, and fibers of passage. Local cells represent neurons that have their cell body in close proximity to the electrode. Afferent inputs represent neurons that project to the region near the electrode and whose axon terminals make synaptic connections with local cells. Fibers of passage represent neurons where both the cell body and axon terminals are far from the electrode, but the axonal process of the neuron traces a path that comes in close proximity to the electrode. The electric field generated by DBS is nondiscriminately applied to each class of neurons, thereby affecting action potential signaling across a wide range of neurons.

The electric field generated by DBS is a three-dimensionally complex phenomenon that is distributed throughout the brain. This field is applied to the three-dimensional geometry of the surrounding neural processes (i.e., axons and dendrites). The applied field generates an extracellular voltage distribution along each neural process, which generates transmembrane current flow. Because of the laws of conservation of current, if the applied field generates depolarization in one part of a neuron, hyperpolarization will be generated in another part to compensate. The distribution of depolarization/hyperpolarization for a given neuron can be qualitatively predicted by the second spatial derivative of the potential distribution along each neural process. The generation of an action potential in response to the applied stimulation will depend on the response of the voltage-gated sodium channels to the membrane polarization. As a result, a neuron can either be activated or suppressed in response to extracellular

stimulation. This activation or suppression can happen in many different ways and in different sections of the neuron depending on its positioning with respect to the electrode and the stimulation parameters used.

Activation or inhibition of the different classes of neurons surrounding the electrode can result in dramatically different physiologic and/or behavioral outputs. Given the stimulation parameters and electrode geometries currently used, it is likely that DBS directly affects every class of neurons surrounding the electrode: local cells, afferent inputs, and fibers of passage. This complicates our ability to specifically address the role of activation or inhibition of each class of neurons in the physiological effects and therapeutic mechanisms of DBS. However, a number of generalized concepts can be used to bias stimulation to a specific class of neurons. For example, fibers of passage are preferentially activated by cathodic (negatively pulsed) stimulation with short ($<100\,\mu s$) pulse widths, and local cells are preferentially activated by anodic (positively pulsed) stimulation with long ($>100\,\mu s$) pulse widths.

DBS Parameter Selection

DBS is an attractive alternative to surgical ablation of brain tissue for a variety of reasons. DBS allows for bilateral procedures without the high incidence of side effects associated with ablative procedures. In addition, side effects associated with the stimulation are reversible, and DBS allows for customization of the therapy to the individual patient's needs over time via alteration of the stimulation parameters. Furthermore, DBS does not destroy tissue, allowing the opportunity for patients to benefit from emerging restorative therapies. However, programming DBS devices for maximal clinical benefit can be a difficult and time-consuming process that typically requires a highly trained and experienced individual to achieve optimal results. Although guidelines exist on stimulation parameter settings that are typically effective for movement disorders, it is not feasible to clinically evaluate each of the thousands of stimulation parameter combinations that are possible. In addition, application of DBS technology to epilepsy, dystonia, and neuropsychiatric disorders is especially problematic because the beneficial effects of stimulation can take weeks to months to manifest and it is currently unclear what stimulation paradigms are optimal for these different disorders.

The physician has the opportunity to adjust a range of stimulation parameters to customize the DBS to each patient. In addition to selecting an electrode contact to deliver the stimulation, the clinician must set electrical parameters of polarity, amplitude, pulse width, and frequency. Typical therapeutic stimulation parameter settings for PD use a square-wave cathodic stimulus pulse $\sim$3 V in amplitude and $\sim$90 μs in duration, delivered at a frequency of $\sim$130 Hz. The specific parameters for a given patient are determined by testing various stimulation parameter combinations and addressing the clinical outcome. This process is guided by clinical intuition and requires periodic adjustments over a period of months to address various therapeutic and side effects of the stimulation.

Mechanisms of Action

Substantial clinical evidence supports the use of DBS technology to treat a wide range of neurological disorders. However, debate continues in the scientific community on the therapeutic mechanisms of DBS. Currently, there exist four general hypotheses to explain the therapeutic mechanism(s) of DBS: (1) stimulation-induced alterations in the activation of voltage-gated currents that block neural output near the stimulating electrode (depolarization blockade), (2) indirect inhibition of neuronal output via activation of axon terminals that make synaptic connections with neurons near the stimulating electrode (synaptic inhibition), (3) synaptic transmission failure of the efferent output of stimulated neurons as a result of transmitter depletion (synaptic depression), and (4) stimulation-induced disruption of pathological network activity (jamming).

Depolarization blockade and synaptic inhibition represent two of the earliest hypotheses to explain the similarity between the therapeutic benefit of ablation and high-frequency stimulation. Single-unit recordings of local cells in the stimulated nucleus support both of these hypotheses. However, the limitation of these hypotheses is that they do not take into account the possible independent activation of the efferent axon of projection neurons. Theoretical and experimental results show that the axon plays the most important role in the activation of neurons surrounding the electrode during extracellular stimulation. Additionally, the response of the cell body does not necessarily represent the output of the axon during DBS. Therefore, although synaptic inhibition and/or depolarization blockade may occur in the cell body, the functional effect of these phenomena may have limited significance in the therapeutic mechanism(s) of DBS.

If DBS does generate efferent outputs in neurons surrounding the electrode, why do ablation and DBS generate similar therapeutic effects? One possibility is that neurons activated by the stimulus train are unable to sustain high-frequency synaptic action on

efferent targets due to depletion of neurotransmitter. In turn, synaptic depression represents an attractive connection between the functional effects of DBS and ablation. However, several *in vivo* experimental studies have shown increases in transmitter release and changes in firing of efferent nuclei consistent with activation of neurons around the electrode and subsequent synaptic action on their targets during high-frequency stimulation. Therefore, although some level of synaptic depression is undoubtedly occurring throughout the stimulated network of neurons, this phenomenon does not appear to be sufficient to block signal transmission between nuclei.

The abnormal motor activity effectively controlled by DBS is most likely generated by altered patterns, increased neuronal synchronization, and/or low-frequency rhythmic oscillation of neurons within the basal ganglia and thalamus. It is possible that DBS overrides these altered patterns and low-frequency oscillatory activity and replaces it with tonic high-frequency output. Analysis of theoretical and neurophysiological data suggests that this more regularized pattern of activity may be more easily compensated for by the remaining elements of the basal ganglia–thalamocortical network. However, limitations in understanding of the dynamics of the basal ganglia network hinder scientific definition of a mechanistic connection between network activity induced by DBS and therapeutic behavioral outcomes.

In summary, a number of questions remain to be answered before the true mechanisms of DBS are explained. Advances in knowledge on the electrophysiological effects of DBS are necessary for a more detailed understanding of this medical technology. Continued scientific advancement in this area will lead to novel stimulation paradigms, surgical techniques, and medical device designs that will improve the clinical outcomes for patients receiving DBS therapies in the future.

See also: Basal Ganglia: Functional Models of Normal and Disease States; Cognition: Basal Ganglia Role; Connectionist Models; Deep Brain Stimulation and Movement Disorder Treatment; Deep Brain Stimulation and Parkinson's Disease; Diffusion Tensor Imaging (DTI); Epilepsy; Obsessive–Compulsive Disorder; Statistical Tests and Inferences.

Further Reading

Alexander GE, Crutcher MD, and DeLong MR (1990) Basal ganglia–thalamocortical circuits: Parallel substrates for motor, oculomotor, 'prefrontal' and 'limbic' functions. *Progress in Brain Research* 85: 119–146.

Benabid AL, Pollak P, Gao D, et al. (1996) Chronic electrical stimulation of the ventralis intermedius nucleus of the thalamus as a treatment of movement disorders. *Journal of Neurosurgery* 84: 203–214.

Bergman H, Feingold A, Nini A, et al. (1998) Physiological aspects of information processing in the basal ganglia of normal and parkinsonian primates. *Trends in Neuroscience* 21: 32–38.

Lee DC, McIntyre CC, and Grill WM (2003) Extracellular electrical stimulation of central neurons: Quantitative studies. In: Finn WE and Lopresti PG (eds.) *Handbook of Neuroprosthetic Research Methods*. Boca Raton, FL: CRC Press.

Lozano AM, Dostrovsky J, Chen R, and Ashby P (2002) Deep brain stimulation for Parkinson's disease: Disrupting the disruption. *Lancet Neurology* 1: 225–231.

McIntyre CC, Savasta M, Kerkerian-LeGoff L, and Vitek JL (2004) Uncovering the mechanism(s) of action of deep brain stimulation: Activation, inhibition, or both. *Clinical Neurophysiology* 115: 1239–1248.

Montgomery EB and Baker KB (2000) Mechanisms of deep brain stimulation and future technical developments. *Neurology Research* 22: 259–266.

Obeso JA, Olanow CW, Rodriguez-Oroz MC, Krack P, Kumar R, and Lang AE (2001) Deep-brain stimulation of the subthalamic nucleus or the pars interna of the globus pallidus in Parkinson's disease. *New England Journal of Medicine* 345: 956–963.

Vidailhet M, Vercueil L, Houeto JL, et al. (2005) Bilateral deep-brain stimulation of the globus pallidus in primary generalized dystonia. *New England Journal of Medicine* 352: 459–467.

Volkmann J, Herzog J, Kopper F, and Deuschl G (2002) Introduction to the programming of deep brain stimulators. *Movement Disorders* 17(supplement 3): 181–187.

Walter BL and Vitek JL (2004) Surgical treatment for Parkinson's disease. *Lancet Neurology* 3: 719–728.

Deep Brain Stimulation and Movement Disorder Treatment

A W Laxton, C Hamani, E Moro, and A M Lozano, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

History and Development

The first attempts at therapeutic deep brain stimulation (DBS) in humans began in the 1950s and 1960s. At this time, reliable surgical stereotactic frames were being developed to enable the precise localization of subcortical targets for lesioning procedures used to treat movement disorders. Electrical stimulation was used in these early procedures to identify desirable target locations for lesion placement. Over time, electrical stimulation itself was used to treat movement disorders, psychiatric disorders, epilepsy, cerebral palsy, and pain. Long-term application of these therapies was limited by the unavailability of suitable implantable electrodes and batteries, however. Furthermore, the success of levodopa to treat Parkinson's disease (PD) in the early 1970s led to a general decline in functional neurosurgery until the late 1980s. In the 1990s, DBS became an accepted therapeutic option for a variety of movement disorders, including PD, levodopa-induced dyskinesias, essential tremor (ET), and dystonia. More than 30 000 DBS operations for movement disorders have been performed worldwide.

Description

The goal of DBS is to improve neurological function through the targeted application of electrical current to specific brain regions or structures. It involves the placement of electrode leads into desired brain regions (see Figure 1). These electrode leads are then connected to a programmable internal pulse generator (IPG). The IPG is placed subcutaneously in the subclavicular region. Each currently available DBS lead has four electrode contacts at its tip, typically 1.5 mm wide and 0.5 or 1.5 mm apart. The DBS leads are made out of polyurethane tubing, and the electrodes are made out of platinum and iridium. The IPG consists of a silver vanadium oxide or lithium ion cell and integrated circuits encased in titanium with an insulating coating.

The typical wave shape of the delivered current of stimulation is square. Stimulation parameters are commonly set to a frequency in the range of 130–185 Hz, a pulse width in the range of 60–120 μs, and an amplitude in the range of 2–4 V, according to the type of movement disorder treated and the type of IPG implanted. Although stimulation can be delivered continuously or in cycles, the continuous mode is used most commonly. The longevity of the IPG battery depends on the stimulation parameters set for each patient. With typical stimulation parameters, the IPG needs to be replaced every 3–5 years.

Indications and Targets

DBS is appropriate for PD patients who have shown a positive response to dopamine agonist medications, such as levodopa, in the past, whose response to dopaminergic agents is of limited or decreasing duration, and who experience significant motor fluctuations or dyskinesias in response to levodopa. The primary goals of DBS are to reestablish the maximal benefit patients have previously achieved with levodopa or to allow a levodopa dose reduction which will diminish drug-induced dyskinesias while maintaining the patient in a good 'on' state throughout the day. DBS may also be used for patients with severe, disabling tremor such as ET or primary and to a lesser extent secondary dystonia that is unresponsive to medical treatments.

The subthalamic nucleus (STN) is the most commonly used target in patients with PD. For patients with ET or tremor secondary to multiple sclerosis, head injury, cerebellar disease or stroke, the most commonly used DBS target is the ventral intermediate nucleus of the thalamus (Vim). For segmental or generalized dystonia, the internal segment of the globus pallidus (GPi) is currently used most commonly although the pallidal-receiving thalamus and the STN may also be useful (see Figure 2).

More recently, pedunculopontine nucleus DBS has been investigated in PD patients, with promising preliminary results. There have been a few reports of the efficacious application of GPi DBS in patients with Huntington's disease and other types of chorea. Patients with Tourette's syndrome have also recently been treated with thalamic and GPi DBS, with variable results.

Anatomy and Pathophysiology

Structural lesions, neural degeneration, problems in connectivity, neurotransmitter deficiencies, and intrinsic disruptions in neural activity are likely contributors to the disturbances in normal motor physiology that underlie movement disorders. These disturbances are funneled to the output structures of the basal ganglia and cerebellum and conveyed to the thalamus and then to the corresponding cortical terminal fields to cause

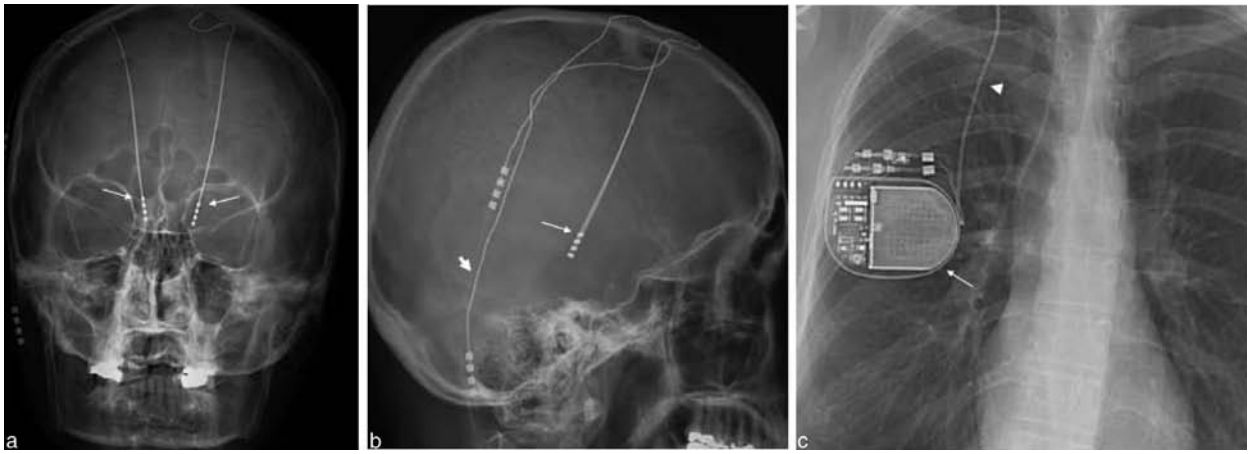

Figure 1 X-ray images of a patient who had subthalamic nucleus deep brain stimulation electrodes implanted. Anteroposterior (a) and lateral (b) skull X-rays showing the bilaterally implanted electrodes with four contacts each (small arrows are pointing at the most dorsal contacts). The large arrow in (b) is pointing to one of the extension cables that connect the electrodes to the internal pulse generator (IPG). (c) Chest X-ray showing the extension cables (arrowhead) that connect the IPG (small arrow) to the electrodes.

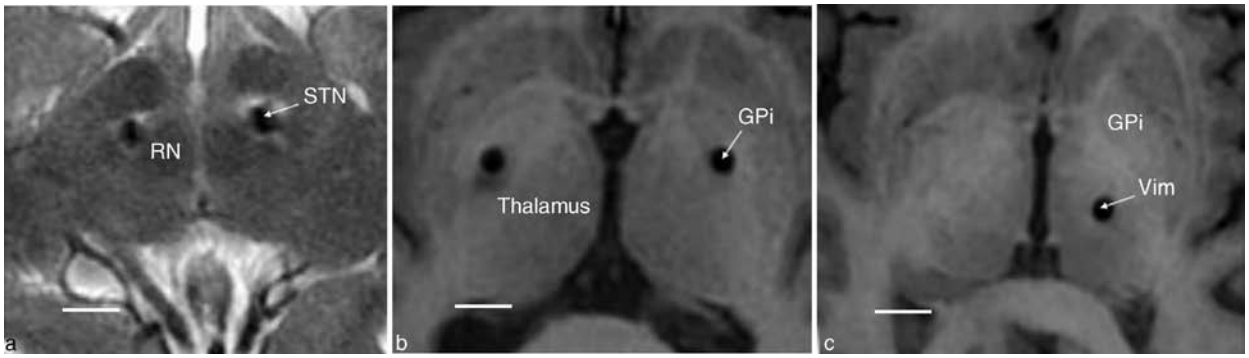

Figure 2 Magnetic resonance images of electrodes (arrows) (a) implanted in the subthalamic nucleus (STN), (b) globus pallidus internus (GPi), and (c) ventral intermediate nucleus of the thalamus (Vim). RN, red nucleus.

the widespread disturbance in neural activity and function that is characteristic of movement disorders. Several theories have been proposed for how the activity within this circuit is regulated. There is evidence for alterations in neuronal firing rates, changes in patterns of discharges and neuronal bursting behavior, the emergence of oscillatory behavior among specific groups of neurons, enhanced correlated activity across neurons, and the expansion of sensorimotor receptive fields as potential contributors to the pathophysiology of movement disorders.

Movement disorders arise when the normal activity within the motor circuit is disturbed. In PD, the activity is disrupted to a large extent because dopaminergic output from the degenerating substantia nigra pars compacta (SNc) is decreased. In primary generalized dystonia, the pattern of GPi neuronal discharge is altered for unknown reasons, leading, it is thought, to dysregulated thalamocortical activity. In ET, groups of thalamic neurons fire synchronously with the peripheral limb tremor. It is believed that DBS interferes with this dysfunctional activity, leading to more-normal motor function.

Mechanism of Action

Electrical brain stimulation affects a broad range of neural elements, including neuronal cell bodies, afferent and efferent axons, and glia. The electrophysiological properties associated with each of these elements vary depending on anatomical location. The effects of electrical stimulation on these elements are incompletely understood but are thought to vary with the intrinsic properties of the stimulated elements and with the stimulation parameters used (i.e., frequency, amplitude, and pulse width).

Because high-frequency stimulation (HFS) of certain targets mimics the effects of lesions to those targets,

it had been believed that stimulation was inhibitory. However, this simple view ignores the possibility of stimulation's activating inhibitory circuits, the potential complex polysynaptic downstream effects of local activation, the possible differential effect of stimulation on cell bodies versus axons, and the possibility that HFS may lead to neurotransmitter depletion or synaptic failure. To better account for the mechanisms of DBS, several explanations have been proposed.

It has been observed that single pulses in the vicinity of GPi neurons produce a cessation of spontaneous activity for 15–25 ms. Because of the latency and duration of this effect, it has been suggested that γ-aminobutyric acidergic (GABAergic) synaptic transmission may be involved. That is, stimulation within GPi causes GABA release from either pallidal (pars externa) or striatal axons projecting onto GPi neurons, or local dendritic release of GABA, resulting in GPi neuronal inhibition. Although the GPi also receives glutamatergic afferents from the STN, the larger GABAergic input is believed to trump these excitatory signals. Further supporting the synaptic transmission proposal, studies have demonstrated increased firing rates in the GPi following STN HFS, and decreased firing rates in the thalamus following GPi HFS. Similarly, glutamate release in the downstream GPi has also been found to be increased following STN HFS. Computer modeling studies have suggested that DBS may inhibit neuronal cell bodies and excite axons. Functional imaging also corroborates the role of HFS in activating axonal elements as increased upstream (antidromic) and downstream (orthodromic) activity has been demonstrated. A variation on the synaptic transmission proposal states that although HFS is excitatory, the stores of neurotransmitter are soon depleted, resulting in functional inhibition.

The ability of HFS to alter intrinsic membrane properties has also been posited as a potential mechanism of the DBS effect. Some *in vitro* studies suggest that HFS transiently depresses voltage-gated Ca^{2+} channels, producing a 'depolarization block.' Alternatively, HFS may also directly activate Na^{+} channels, leading to depolarization. The relevance of these *in vitro* findings awaits further investigation.

The seemingly contradictory explanations for DBS effects may actually be complementary. That is, the effect of DBS may depend on several mechanisms, and the effects may vary depending on what structures are being stimulated and how far a structure is from the stimulating electrode. In addition, since the pathophysiology of most of the movement disorders is still not entirely known, the stimulation of the same target in different diseases may require different parameters of stimulation (e.g., low-frequency GPi stimulation in Huntington's disease and high-frequency in PD). The net effect of DBS, however, appears to be to disrupt the dysfunctional neuronal activity associated with movement disorders.

Operative Procedure

DBS surgery involves brain imaging, target identification, and the insertion of DBS electrodes and an IPG. The STN and GPi can be visualized with magnetic resonance imaging. The Vim target is estimated indirectly, relative to the position of the anterior and posterior commissures. Intraoperative electrophysiological mapping is used to characterize the location of the recording electrode in relation to the desired target, to assess the therapeutic effect of an electrode in a particular location, and to identify the potential unwanted side effects associated with stimulation in a given location. Approaches may involve one or more microelectrodes (single unit recording; local field potentials; microstimulation) or macroelectrodes (usually, the DBS electrode itself; local field potentials; macrostimulation). The majority of DBS centers use a combination of these techniques.

Standard mapping of the STN begins with microelectrode recordings starting in the range of 10–15 mm above the chosen target. The single unit activity recorded depends on the specific trajectory chosen but may include the following structures, in order (see Figure 3): the thalamus (with its characteristic bursting cells), the zona incerta (exhibiting a low firing rate), the STN (with its characteristic dense cellular activity in the range of 30–50 Hz, and occasional 4–6 Hz bursting activity synchronized to the patient's tremor), and then the substantia nigra pars reticulata (characterized by a more regular firing rate in the range of 50–70 Hz, without bursting or tremor cells). When the STN has been reached, testing begins to determine whether the cellular activity changes in response to the patient's active or passive limb movements. The presence of movement-related activity suggests that the electrode may be in a therapeutically desirable location. At this point, microstimulation may also be performed within the STN to check for tremor arrest or benefit to the other clinical signs or to detect the presence of paresthesias, muscle contractions, double vision, or other unwanted side effects (related to current diffusion into the areas surrounding the STN, which aids in the identification of the correct electrode locations). The microelectrodes may be repositioned to sample other STN areas and to identify the portion of the STN populated by movement-responsive neurons – the 'motor territory.' The microelectrodes are then removed and replaced by the DBS electrode. This insertion is done using X-ray guidance to ensure that the electrode is placed in the desired location. Once in place, each contact of the DBS electrode

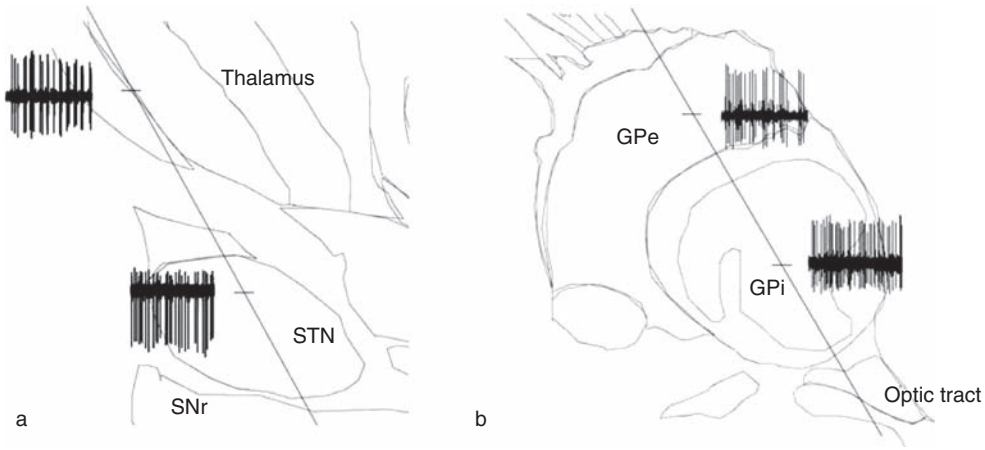

Figure 3 Schematic representation of sagittal sections of the brain that are 12 mm (a) and 20 mm (b) from the midline. The oblique line in the center of the figures represents a typical microelectrode trajectory during the physiological exploration of the subthalamic nucleus (STN) (a) and globus pallidus (b) in patients with Parkinson's disease. (a) During the exploration of the subthalamic region, neuronal activity may be recorded from the thalamus, STN, and substantia nigra pars reticulata (SNr). Examples of a thalamic bursting cell and an irregularly firing STN cell are shown on the left side of the trajectory line. (b) Neuronal activity recorded from the globus pallidus externus (GPe) and globus pallidus internus (GPi). Stimulation of the optic tract often induces phosphenes and maps the bottom of the trajectory. For details on the neuronal firing characteristics in each of these regions, reader is referred to the text.

is stimulated with incrementally increasing voltage to check the threshold for stimulation-induced clinical benefit and unwanted side effects. The DBS electrode position can be adjusted to minimize adverse effects.

The general procedure and rationale for intraoperative electrophysiological mapping of the GPi and Vim are similar to those described for the STN, except for differences in regional anatomy and associated cellular characteristics. The microelectrode trajectory to the GPi traverses the striatum, external segment of the globus pallidus (firing activity around 50 Hz), the GPi (high-frequency activity around 80 Hz with occasional 4–6 Hz tremor-associated bursts), and then either the optic tract (inferiorly) or the internal capsule (medially and posteriorly). As with the STN, movement-related cells are sought within the GPi, and microstimulation is used to identify side effects such as muscle contractions (posterior limb of the internal capsule) or visual phosphenes (optic tract). To reach the Vim, the microelectrode may traverse the following thalamic nuclei, depending on trajectory: ventral oralis anterior/posterior (Voa/Vop; tonic firing rate around 20 Hz; active movement-responsive activity), Vim (around 25 Hz; passive movement-responsive activity), ventral caudalis (Vc; cellular activity similar to Vim but displays tactile/sensory-responsive activity), and finally beyond the thalamus into the medial lemniscus. Stimulation of the Voa/Vop and Vim should produce tremor suppression; stimulation of the Vc and medial lemniscus elicits paresthesias.

Advantages, Disadvantages, and Alternatives

Prior to the development of DBS, lesioning procedures were the standard surgical approach to treat movement disorders. Lesions are made with high-radiofrequency stimulation to the same targets used in DBS therapy. While lesion surgery is still practiced, bilateral lesions can be associated with a high incidence of adverse effects (e.g., cognitive impairment, dysarthria, and dysphagia) and is only rarely used. The adjustability and reversibility of DBS are two of its main advantages over lesioning procedures. With DBS therapy, stimulation parameters can be adjusted to enhance or maximize a positive effect or to decrease unwanted side effects (e.g., paresthesias, muscle contractions, and dysarthria). Postoperative programming of the DBS system involves adjustments to the following parameters: choice of contact, type of stimulation (monopolar or bipolar), mode of stimulation (continuous or cycled), amplitude (voltage), frequency, and pulse width. Another interesting advantage of DBS over lesioning is the possibility of performing a bilateral procedure with a lower incidence of the severe complications that have been found with bilateral ablative surgery.

The major disadvantages of DBS include its expense; the need to replace IPGs (batteries) every 3–5 years, which involves minor surgery; and because it entails implanted foreign bodies, an increased infection risk relative to lesions. In addition, DBS surgery requires the

involvement and coordination of a large multidisciplinary team (neurologist, psychiatrist, neuropsychologist, neurophysiologist, and neurosurgeon) in order to select the appropriate candidates before surgery and to manage medication reductions, DBS programming, and possible DBS complications after surgery.

Outcomes

For patients with PD, STN DBS has been shown to produce improvements in the range of 60–70% in overall motor functioning (based on the motor component of the Unified Parkinson's Disease Rating Scale). Stimulation-associated improvements in the off-medication state vary across the various motor domains of PD and are approximately as follows: tremor, 90%; rigidity, 75%; gait, 60%; bradykinesia, 50%; postural instability, 40%; and dystonia, 30–50%. Some of these improvements, particularly in tremor and rigidity, can persist for 5 or more years. The benefits in gait and postural disability, however, wane by 2 years. STN DBS improves dyskinesias by allowing dopaminergic medication dosages to be reduced (on average by approximately 50%) and through a direct antidyskinetic effect. STN DBS has also been found to improve patients' activities of daily living and quality of life. The benefits of GPi DBS in PD may approach those seen with STN DBS, but the outcomes are less well characterized.

Vim stimulation is associated with up to 80–90% tremor reduction in PD and ET patients. In comparison with STN and GPI stimulation, however, Vim DBS does not improve the bradykinesia and rigidity associated with PD. For patients with ET, Vim stimulation may be more effective at treating hand tremor than voice and head tremor.

For primary generalized dystonia, GPi DBS is associated with an approximately 50% improvement in motor functioning and reduction in disability, on the basis of the Burke–Fahn–Marsden Dystonia Scale, and an improved quality of life related to general health and physical functioning. These beneficial effects can be expected to last for 1 or more years. GPi DBS may be more effective in improving motor function in the neck, trunk, and limbs than in improving facial movements and speech. The emergence of DBS over the last decade is a major advance in the surgical treatment of movement disorders.

See also: Animal Models of Parkinson's Disease; Basal Ganglia: Internal Organization; Basal Ganglia: Physiological Circuits; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Deep Brain Stimulation; Deep Brain Stimulation and Parkinson's Disease; Dystonia: Classification, Genetics and Therapeutics; Parkinsonian Syndromes; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Abosch A and Lozano A (2003) Stereotactic neurosurgery for movement disorders. *Canadian Journal of Neurological Sciences* 30: S72–S82.

Benabid AL, Pollak P, Gervason C, et al. (1991) Long-term suppression of tremor by chronic stimulation of the ventral intermediate thalamic nucleus. *Lancet* 337: 403–406.

Deuschl G, Schade-Brittinger C, Krack P, et al. (2006) A randomized trial of deep-brain stimulation for Parkinson's disease. *New England Journal of Medicine* 355: 896–908.

Dostrovsky JO and Lozano AM (2002) Mechanisms of deep brain stimulation. *Movement Disorders* 17: S63–S68.

Gildenberg PL (2005) Evolution of neuromodulation. *Stereotactic and Functional Neurosurgery* 83: 71–79.

Haber SN and Gdowski MJ (2004) The basal ganglia. In: Paxinos G and Mai JK (eds.) *The Human Nervous System*, 2nd edn., pp. 676–738. London: Elsevier.

Hamani C, Richter E, Schwalb JM, and Lozano AM (2005) Bilateral subthalamic nucleus stimulation for Parkinson's disease: A systematic review of the clinical literature. *Neurosurgery* 56: 1313–1321; discussion 1321–1324.

Hassler R, Riechert T, and Mundinger F (1960) Physiological observations in stereotaxic operations in extrapyramidal motor disturbances. *Brain* 83: 337–350.

Krack P, Batir A, Van Blercom N, et al. (2003) Five-year follow-up of bilateral stimulation of the subthalamic nucleus in advanced Parkinson's disease. *New England Journal of Medicine* 349: 1925–1934.

Kupsdch A, Beneke R, Muller J, et al. (2006) Pallidal deep-brain stimulation in primary generalized or segmental dystonia. *New England Journal of Medicine* 355: 1978–1990.

Lang AE, Deuschl G, and Rezai AR (eds.) (2006) Deep brain stimulation for Parkinson's disease. *Movement Disorders* 21 (supplement 14): S167–S327.

Lozano AM, Lang AE, Galvez-Jiminez N, et al. (1995) Effect of GPi pallidotomy on motor function in Parkinson's disease. *Lancet* 346: 1383–1387.

Perlmutter JS and Mink JW (2006) Deep brain stimulation. *Annual Review of Neuroscience* 29: 229–257.

Rodriguez-Oroz MC, Obeso JA, Lang AE, et al. (2005) Bilateral deep brain stimulation in Parkinson's disease: A multicentre study with 4 years follow-up. *Brain* 128: 2240–2249.

Vidailhet M, Vercueil L, Houeto J-L, et al. (2005) Bilateral deep-brain stimulation of the globus pallidus in primary generalized dystonia. *New England Journal of Medicine* 352: 459–467.

Deep Brain Stimulation and Parkinson's Disease

M C Rodriguez-Oroz, J M Matsubara, P Clavero, J Guridi, and J A Obeso, University Clinic and Medical School, University of Navarra, Pamplona, Spain

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Deep brain stimulation (DBS) of the ventral intermediate nucleus of the thalamus (Vim) in Parkinson's disease (PD) has been performed since the late 1980s, mimicking the effect of the lesion against tremor. Bilateral approaches with DBS and bilateral thalamotomy were similarly effective, but the incidence of adverse events was much lower with DBS. Thus, DBS of the Vim replaced thalamotomy in patients with tremor. The realization that DBS delivered similar benefits with fewer adverse events than did lesions, the need for bilateral surgery in the majority of severe PD patients, the reversibility of this approach in contrast to the lesion, and a study in 1-methyl, 4-phenyl-1,2,3,6 tetrahydropyridine monkeys showing high antiparkinsonian efficacy using high-frequency stimulation of the subthalamic nucleus (STN) led to the introduction of DBS of the STN in the therapeutic armamentarium for advanced PD patients. Concomitantly, this technique was also applied to the globus pallidus pars interna (GPi) in an attempt to achieve the benefit of a bilateral approach without the undesirable adverse effects induced by the lesions. Initial results were positive, leading to a multicenter trial begun in 1996 to assess the safety, viability, and short-term efficacy of DBS of both STN and GPi. The positive findings of this trial allowed a prompt spreading of this therapy, which continues to the present, with stimulation of the STN predominating. This article reviews the most important aspects of bilateral DBS of the STN and GPi.

Methodological Aspects

The success of surgical treatment with DBS in PD depends on several factors. The selection of the appropriate candidate for DBS is the first critical point. Thus, the clinician has to evaluate in every patient the known negative and positive predictive factors for the outcome of this therapy. The second key factor is the accuracy in targeting the STN or GPi during the surgical procedure and prior to the implant of the therapeutic electrode. The precise location of the electrode in the nucleus determines a better antiparkinsonian benefit without stimulation-induced adverse events. Last, the careful exploration of the benefit obtained with the different programming of the electrodes, the adequate selection of the parameters for continuous stimulation, and the concomitant adjustment of medication guarantee an optimization of the therapy.

Selection of PD Patients for Surgical Therapy with DBS

The experience gained in the last decade with PD patients submitted to DBS has allowed us to define a profile of patients with a higher probability of obtaining better results with this therapy. Numerous factors need to be taken into account in the evaluation of whether a patient with PD is a good candidate to be treated with DBS.

Advanced age, concomitant severe medical illness, and neuropsychological and psychiatric disturbances can be a drawback in the outcome of this therapy. Although age-related comorbidities seem to be more important than age *per se*, and so no specific age cutoff has been defined, in practice, the vast majority of centers exclude patients older than 75. Serious systemic or neurologic disease other than PD should be regarded as a contraindication for DBS, as they would significantly compromise DBS benefits or amplify surgical risks. Previous functional surgery should not exclude additional DBS procedures. Neurobehavioral contraindications may include primary psychotic disorders, primary and compensated bipolar disorders, severe treatment refractory depression, severe substance abuse, and severe personality disorders.

It is important to note that factors related to the PD itself and levodopa-induced side effects have to be carefully analyzed. Dementia, frequently seen in advanced PD patients, should be considered a contraindication for surgery despite the lack of outcome studies. Levodopa-resistant symptoms (i.e., freezing of gate, instability, speech disturbances) are not significantly improved by DBS and, although not a formal contraindication, should be considered in every particular case. As a rule, DBS improves PD signs that are alleviated with levodopa therapy. Preoperative levodopa response is a predictive factor for the positive response to surgery, with the possible exception of drug-refractory but otherwise classical rest tremor. Currently, all studies that have evaluated predictors of benefit from DBS have found that levodopa response is the best one. Thus, a test evaluating the motor improvement in response to a high dose of levodopa is crucial to estimate the achievable antiparkinsonian benefit with DBS. A minimum 25–50% reduction in the Unified Parkinson's Disease Rating Scale (UPDRS) motor part (III) score following the

administration of an effective dose of levodopa in the practically defined off-medication state (a minimum of 12 h without antiparkinsonian medication) is considered a significant response. There is some controversy about the dose of the levodopa challenge used for the test: the first dose of levodopa in the morning or a suprathreshold dose (50% more of this dose) is the most frequently employed.

Another important issue is the outcome of DBS surgery in patients with dopaminergic-induced psychiatric disturbances. Although some publications indicate patients with gambling and dopamine dysregulation syndrome can improve both the parkinsonism and the dopaminergic side effects with DBS of the STN, there is not enough data in the literature at present to draw conclusions. Thus, no recommendation can be made on this particular and relevant aspect of PD until more experience is accumulated.

In addition to the previous criteria, it is recommended to operate on patients with a score higher than 30 in the motor UPDRS scale in the off medication state and with a disease duration of at least 5 years (to allow for atypical forms of parkinsonism to become evident).

A question recently raised is the benefit of this therapy in patients with genetic forms of PD. The available data indicate these patients can be successfully operated on with bilateral DBS of the STN. The inclusion and exclusion criteria are the same as for idiopathic PD.

Surgical Issues

Accuracy in targeting the STN or GPi is critical for the outcome of DBS in PD, as previously explained. Stereotactic functional surgery is the procedure universally employed. Thus, the first step of the intervention is the stereotactic targeting of the nucleus by using computed tomography, magnetic resonance imaging, venticulography, or neuronavigational systems. They are all safe and reliable techniques that have been used effectively by experienced groups. Image fusion may improve the accuracy but is not mandatory. Imaging techniques should be coupled with physiological and/or clinical intraoperative testing for target refinement. The physiological definition of the nuclei is usually undertaken with microelectrode or semimicro electrode recording of the neuronal activity of the nucleus (STN or GPi) (Figure 1). Macroelectrodes can also be used for recording but provide very poor spatial resolution. When physiological methods are employed, once the target is precisely identified, microstimulation with the electrode used for micro or semimicrorecording and/or macrostimulation must beused for refining the final target location. In those centers that do not use physiological

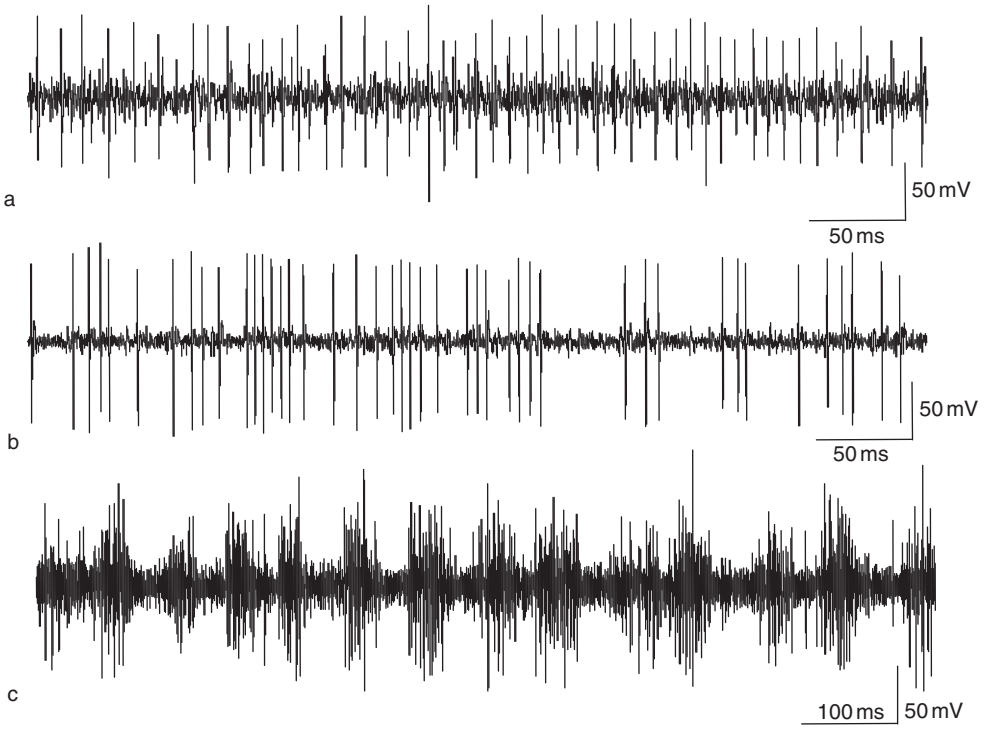

Figure 1 Microrecording of the neuronal activity of the STN in a patient with Parkinson's disease. Example of the three types of neuronal discharges: (a) tonic, high-frequency pattern; (b) irregular activity with long pauses; (c) oscillatory, slow activity.

Figure 2 Coronal magnetic resonance image of a patient which leads for DBS bilaterally implanted in the STN.

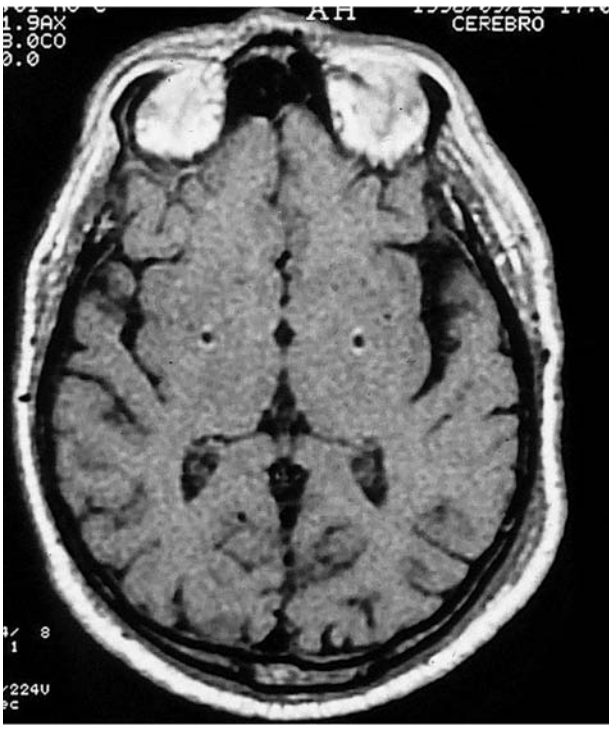

Figure 3 Axial magnetic resonance image of a patient which leads for DBS bilaterally implanted in the Gpi.

parameters to better define the intended target, clinical intraoperative testing with a macroelectrode or with the DBS electrode can also be performed directly. In any case, and independent of the method used, once the therapeutic electrode has been positioned (Figures 2 and 3), intraoperative DBS stimulation is conducted to assess the benefit of the therapy against parkinsonian signs and to test the possible induction of stimulation-related adverse events that could thereafter preclude the utilization of this treatment. It has to be known that in some instances, there is an 'impact' or 'lesionlike' effect that consists of an amelioration of the parkinsonian signs after the mechanical introduction of the therapeutic electrode into the nucleus. In these cases, the intraoperative stimulation with the DBS electrode does not provide any relevant information.

Patients remain awake during the whole intervention in off-medical condition. Local anesthesia is used in the area of the burn hole. In some particular cases (i.e., severe tremor or dystonia), antiparkinsonian medication, sedation, or general anesthesia is administered. This situation is far from ideal because it interferes with the neuronal activity of the nucleus and its recognition by the physician while recording and because it does not allow testing of the antiparkinsonian efficacy and side effects of the stimulation.

Once the therapeutic electrode has been placed in the nucleus and the effect of the stimulation has been evaluated, the next step is to confirm its location by intraoperative fluoroscopy before securing it to the burn hole with a specific device.

Apart from the therapeutic electrode(s), the rest of the hardware (external cable and battery) has to be placed under the subcutaneous cellular tissue and connected to the lead already implanted (Figure 4). The battery or internal pulse generator (IPG) is placed in a tissue pocket located in the upper chest or abdomen and connected to the therapeutic electrode through the external cable. This procedure is undertaken with general anesthesia and can be done in the same surgery or later.

Another important decision when bilateral procedures are planned is whether they should be done in the same surgery. Available data show no significant differences between surgical time and patient clinical outcome, and every case should be individually evaluated.

Postoperative Management

Postoperative antiparkinsonian treatment should be restarted as soon as possible to relieve discomfort and limit the risk of acute dopaminergic withdrawal and malignant hyperthermia. The postoperative levodopa equivalent dose (LED) should be the same as the presurgical dose and composed primarily of levodopa to facilitate simple titration during the process of programming the stimulation. A confusional state may be present in some patients, but it generally lasts a few days and can be managed easily with atypical antipsychotics such as clozapine or quetiapine.

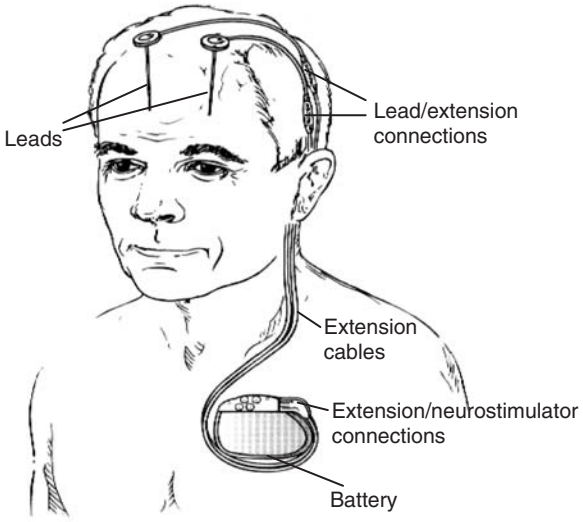

Figure 4 Schematic representation of the implanted device for DBS.

Neuroimaging to confirm the electrode location and to rule out intracranial adverse events is recommended.

Programming can be initiated at any time after the whole device has been implanted depending on the center and the situation of the patient. When a lesion-like effect is present, the programming can be delayed until the parkinsonian symptoms reappear. Programming should be started in the off-medication state to select and set up the parameters with the highest antiparkinsonian efficacy. Parameters inducing permanent adverse events should be ruled out. Stimulation with the greatest efficacy and fewest side effects is selected for therapy. Once programming has been selected, gradual reduction of drug doses has to be initiated to adjust the medication to the new motor state. Levodopa and STN stimulation have similar and additive effects in the antiparkinsonian and dyskinetic states. Thus, in PD patients treated with DBS of the STN, medication titration is needed for fine adjustment of the motor situation in the majority of patients. In cases with DBS of the GPi, this additive effect of stimulation and medication is not as strong, and LED cannot be reduced.

Treatment should be maintained with levodopa to allow the programming to be adjusted adequately to both off- and on-medication motor conditions. Once programming is stable, dopaminergic agonists can help to achieve a more stable motor benefit than levodopa.

After this first adjustment of stimulation and medication, the frequency of visits to the neurologist depends on the patient's needs and on the available resources of each center. In any case, follow-up visits are recommended.

Results of Bilateral DBS in PD

The multicenter study conducted in 1996 to assess the safety, viability, and efficacy of DBS of the STN and GPi was not randomized, and no specific guidelines were issued for allocating patients to STN or GPi surgery. Both groups showed significant improvement induced by stimulation. Over the years, many groups all over the world have applied DBS of the STN and GPi to treat PD, but studies properly comparing the therapeutic efficacy of the two targets are limited.

Nevertheless, STN DBS has become the preferred approach in most centers. There appear to be three main reasons. First, the impression that the clinical relief of the cardinal features of PD obtained by STN DBS was greater than that obtained by GPi DBS spread quickly after the initial trial despite absence of validated data. Second, from a surgical point of view, it may be easier to physiologically define the sensorimotor region and surrounding structures of the STN than those of the GPi. Finally, and as an established fact, STN DBS allows a significant reduction in the daily dose of dopaminergic drugs that is not achieved with GPi DBS. This is a very attractive clinical possibility considering the profusion of motor problems (i.e., dyskinesia) and nonmotor problems (e.g., psychiatric complications, sleep disturbance) associated with chronic levodopa treatment. Overall, the issue of advantages and disadvantages of either target remains unresolved.

Bilateral DBS of the STN

Bilateral DBS of the STN has become the most frequent surgical therapy for PD patients.

Since the first patients operated on by the Grenoble group early in the 1990s, a huge number of publications have described the effect of this therapy in PD patients at different points in the postoperative follow-up. Some of them, because of the number of patients included or long-term follow-up, are more indicative of the benefit–risk ratio of this treatment. A meta-analysis of the outcome of DBS of the STN based on a review of the literature from 1993 until June 2004, while acknowledging between-study heterogeneity, reached the conclusion that the treatment is highly effective in PD patients. It improves UPDRS-II or activities of daily living (ADL) scores by 50%, UPDRS-III scores by 52%, dyskinesia scores by 69.1%, daily off-medication time by 68.2%, and quality of life by 34.5%. LED was reduced in 55.9%. After June 2004, and therefore not included in this meta-analysis, many more articles about DBS of the STN in PD have been published with similar findings. Because of the profusion of data, only those studies with long-term follow-up (4–5 years) published after the meta-analysis are reviewed in this article (Table 1).

Table 1 Summary of the most relevant long-term follow-up studies of DBS of the STN

Study	*Follow-up (months)*	*Age at surgery (years)*	*Mean H&Y (off)*	*Disease duration (years)*	*Preoperative levodopa response (UPDRS III off/on: % change)*	*UPDRS II score (postop stim on and med off vs. preop med off; % change)*	*UPDRS III score (postop stim on and med off vs. preop med off; % change)*	*LED (% change)*	*Dyskinesia (% change)*	*Duration of daily 'off' (% change)*
Rodriguez-Oroz ($N=49$)	36–48	59.8 ± 9.8	4.3 ± 0.8	15.4 ± 6.3	59.78	43	50	35	59	56
Visser-Vandewalle ($N=20$)	53.6 ± 2.6	60.9 ± 8.1	NR	15 ± 4.4	NR	59	42.8	47.2	79	78
Østergaard ($N=18$)	48 ± 2	63 ± 8	3.7 ± 0.8	19 ± 5	54.19	42	55	29	90	67
Shupbach ($N=30$)	60	54.9 ± 9.1	5	15.2 ± 5.3	69	NR	49.51	54.56	79	67

[a]Data reported as mean ± SD/range.
H & Y, Hoehn and Yahr scale; LED, levodopa equivalent dose; *N*, effective sample size; NR, not reported; UPDRS, Unified Parkinson's Disease Rating Scale; II, UPDRS activities of daily living scale; III, UPDRS motor scale.

Rodriguez-Oroz et al. reported the results of a multicenter study that is the follow-up of the first multicenter trial aimed at assessing the safety, viability, and efficacy of DBS of either the STN or the GPi. Overall, 49 patients treated with DBS of the STN were available for study after 4 years. Stimulation of the STN in off-medication condition induced a reduction of 50% in the motor UPDRS score and of 43% in the ADL score compared with the preoperative medication off state. All cardinal features were also improved except speech. In the on-medication situation and in comparison with the preoperative on-medication state, stimulation did not improve either the motor or the ADL-UPDRS scores. Motor fluctuations were reduced with an increment of the time spent in the on-medication condition without dyskinesias. Dyskinesia severity was reduced by 59% with respect to baseline and LED in 35%. Comparison of the improvement induced by STN stimulation at 1 year and at 4 years showed a worsening in axial signs (speech, postural stability, and gait) in both off- and on-medication conditions. This result is similar to the one obtained in a cohort of 49 patients after 5 years of follow-up by Krack et al. that was included in the meta-analysis. Adverse events were frequent (53% of patients) and will be detailed below.

In addition to the multicenter study, there are other studies with very similar results (Table 1). Visser-Vandewalle reported that the benefit of stimulation of the STN in the off-medication condition was maintained after 4 years in the UPDRS motor score and in all cardinal signs of PD (tremor, rigidity, bradykinesia and gait) when compared with the off-medication state at baseline. Antiparkinsonian benefit was lower at 4 years than at 3 months (i.e., total motor UPDRS score reduction of 54% at 3 months and 43% at 4 years). In the on-medication state, stimulation-linked improvement in total UPDRS motor score, rigidity, and akinesia at 3 months vanished at 4 years, and gait worsened significantly. Dyskinesias, motor fluctuations, and ADL were similarly improved by 79%, 78%, and 59%, during the 4 years of follow-up, and LED was reduced by 42.8%.

Another cohort of 18 patients has shown similar results after 4 years of follow-up. Stimulation induced significant benefit in the off-medication situation compared with the off-medication preoperative state in ADL and motor UPDRS scores except for the axial signs (postural stability, gait, speech), which also were significantly worsened compared with the 1-year and basal condition. Stimulation in the on-medication state did not improve the total motor score, and there was even a worsening of speech and postural stability compared with the on-medication condition at baseline. The reduction in LED was 29%. It is interesting that the benefit of levodopa in axial signs and total motor UPDRS was significantly lower at 4 years.

Shüpbach et al. showed a 49.5% UPDRS motor score benefit of stimulation in the off-medication state, a less-striking improvement in the axial signs, and a 40% benefit in the ADL. Motor complications and dyskinesias were reduced by 67% and 79%, respectively, although motor complications were significantly higher at the 5-year evaluation than at the 6-month follow-up. The reduction in LED was 54%.

Recently the effect of DBS of the STN has been compared with the best medical management in two trials. Deuschl et al. studied 156 patients in a multicenter, randomized-pairs trial with 6 months of follow-up. Stimulation produced greater improvement in the quality of life and the various motor evaluations (UPDRS, motor fluctuations, etc.) than the best medical treatment. In the other trial, only young patients (younger than age 55) with no severe PD were included in a study aimed at assessing the benefit of this surgery in comparison with the best medical treatment at an early stage of the disease. Similar to the previous trial, the patients who underwent surgery early after the onset of PD had an outcome superior to that of patients treated with the best medical therapy.

In summary, bilateral stimulation of the STN produces a significant benefit on the off-medication motor and ADL UPDRS scores and in cardinal features of PD patients after long-term follow-up, a benefit that is less striking in axial signs. In the on-medication condition, stimulation tends to lose efficacy in the motor and ADL scores. Remarkably, a significant decline in the effect of stimulation in the off-medication states between the short- and long-term evaluations was observed in all the studies, mainly in the axial signs. However, this worsening in axial features was also seen in the on-medication condition, indicating that it may reflect the progression of the disease and extension of the pathological process beyond the nigrostriatal dopaminergic system. Moreover, in terms of quality of life, parkinsonian signs, and motor complications, this therapy seems to be more beneficial than the best medical therapy.

Bilateral DBS of the GPi

Bilateral stimulation of the STN became the preferred approach for DBS in PD. Thus, the data about the efficacy of DBS of the GPi are less abundant. With the exception of the multicenter study initially undertaken in 1996 and the follow-up of this trial at 4 years after surgery, the majority of studies conducted to assess the outcome of this intervention are open trials, the numbers of patients are small, and the follow-up

Table 2 Summary of the most relevant studies evaluating the antiparkinsonian benefit of deep brain stimulation of the GPi

Study	*Follow-up (months)*	*Age at surgery (years)*	*Mean H&Y (off)*	*Disease duration (years)*	*Pre-operative levodopa response (UPDRS III off/on; % change)*	*UPDRS II score (postop stim on/med off vs. preop med off; % change)*	*UPDRS III score (postop stim on/med off vs. preop med off; % change)*	*LED (% change)*	*Dyskinesia (% change)*	*Duration of daily 'off' (% change)*
Rodrigues ($N=11$)	24.8	61.45	2.7	14.54	50	NR	46	UCH	76	NR
Volkmann ($N=11$)[b]	12 ($n=10$)	NR	NR	11.2 ± 2.7	44	50	56	UCH	58	NR
	36 ($n=9$)					30	43		63	
	60 ($n=5$)					10	24		64	
Durif ($N=5$)	36	64	4.3 ± 0.2	15 ± 2	72	9	32	UCH	50	10
Ghika ($N=6$)[c]	24	55	4.2	15.5	33 minimum	>50	>50	UCH	75	>50
Obeso ($N=36$)	6	55.7 ± 9.8	NR	14.5	NR	36	33.3	UCH	67	35
Rodriguez-Oroz ($N=20$)	46.8	55.8 ± 9.4	NR	15.4	NR	28	39	UCH	76	45
Loher ($N=10$)	12	65	4.2 ± 0.8		NR	34	41	UCH	71	NR
Kumar ($N=17$)	2	52.7 ± 8.5	NR	14.8 ± 6.1	NR	39	31	UCH	68	NR
Anderson ($N=10$)	10	54 ± 12	4	10.3 ± 2	57	23	39	3	89	NR

[a]Data reported as mean ± SD/range.
[b]Four patients who experienced no efficacy were operated on in the STN after 2 years ($n=2$) and 3 years ($n=2$).
[c]Three patients experienced worsening after 1 year with stimulation-resistant gait ignition failure.
H&Y, Hoehn and Yahr scale; LED, levodopa equivalent dose; NR, not reported; *N*, effective sample size; UCH, unchanged; UPDRS, Unified Parkinson's Disease Rating Scale; II, UPDRS activities of daily living scale; III, UPDRS motor scale.

after surgery is variable. The most important studies of DBS of the GPi are summarized in Table 2.

In the first multicenter study, the final evaluations were undertaken after 6 months of surgery in 36 patients bilaterally treated with DBS of the GPi. Stimulation in the off-medication condition compared with the same state at baseline induced a significant improvement of 33% and 35% in motor UPDRS and ADL scores, respectively. All cardinal features were also improved. In the on-medication state compared with baseline on-medication, stimulation provided 26.6% and 30% reduction in the motor UPDRS and the ADL scores, respectively. Not all cardinal features were improved in this medication condition. Dyskinesias were reduced by 66.6%, and the time spent in the on-medication condition without dyskinesias was increased by 56%. The total dose of antiparkinsonian medication was unchanged.

An extended follow-up of this multicenter study that evaluated the effect of this treatment in 20 patients after 4 years showed the benefit of stimulation was maintained during this period of time in the off-medication state. There was a reduction of 39% and 28% in motor UPDRS and ADL scores, respectively, with respect to the preoperative off-medication state. The improvement on postural stability and speech initially observed was lost at 4 years in the off-medication state. Furthermore, the on-medication assessment at 4 years showed a progressive decline in motor UPDRS, gait, and ADL scores with respect to the 1-year situation in the on-medication state, with no significant changes in the off-medication scores. Dyskinesias were reduced by 76% with respect to the baseline situation. These data indicate DBS of the GPi maintains its efficacy in the long-term follow-up in the off-medication state but not in the on-medication situation, with a worsening in the motor state, predominantly in the axial features. The antidiskinetic effect was maintained, and the dose of medication was unchanged with respect to baseline (a slight increment was not significant).

In contrast to this study, Volkmann et al. reported that in 11 patients treated with bilateral DBS of the GPi, there was a significant improvement in the parkinsonian condition in the off-medication situation in the first year after the surgery, but that it started to decline after 3 years and was not maintained at 5 years. Cardinal features were not equally improved, and the axial signs (speech, swallowing, gait, and posture) were the least improved. Actually, there were four patients with lack of efficacy mainly due to gait deterioration after 1 year who required a new-intervention with bilateral stimulation of the STN, with significant improvement. In on-medication condition, stimulation induced a significant benefit in the motor and ADL UPDRS scores but no benefit in individual cardinal features with respect to the preoperative on-medication status in the short term. After 5 years, the benefit had vanished. On the other hand, the reduction in on-condition dyskinesias was maintained during the 5 years. The dose of antiparkinsonian medication was unchanged.

Some other studies with a smaller number of patients and variable follow-up report similar results. Ghika et al. reported six patients in whom bilateral stimulation of the GPi improved ADL and motor UPDRS scores in the off-medication state, the time spent in off condition during the day, and the on dyskinesias after 2 years. During this follow-up period, four patients developed motor fluctuations and needed additional medication, and three patients developed gait initiation failure resistant to stimulation of the GPi. Stimulation did not improve the on-medication state at any interval after surgery. The improvement in the severity of on-condition dyskinesias was significant during the whole period. Another study, also conducted with a small number of patients, found similar results in the off-medication state in the short term but found that after 3 years, the motor scores showed a progressive decline. There was no improvement in the on-medication condition at any point. Benefit was maintained in the on-medication dyskinesias during the entire period.

Recently another study of 11 patients treated with bilateral DBS of the GPi for 2 years reported similar results. There was a benefit in the off-medication state in the motor UPDRS score, with all cardinal signs improved. There was a slight benefit of 18% in the on-medication status in comparison with the on-medication state preoperatively. Dyskinesias were improved by 76%, and no change in the medication dosage was found.

In summary, all studies with short-term follow-up have shown that bilateral GPi stimulation induces a dramatic decrease of on-medication dyskinesia, with or without change in the on-medication motor and ADL UPDRS scores, and a significant improvement in the off-medication motor and ADL UPDRS scores, with different degrees of benefit in the cardinal features. It is important to note that in spite of the antiparkinsonian benefit, the dopaminergic treatment cannot be reduced. In the long-term follow-up, the antidyskinetic effect is maintained, but there is a tendency to lose the antiparkinsonian benefit in the on-medication state and, in many reports, also in the off-medication state, mainly in the axial signs. Actually, for a significant number of patients, the antiparkinsonian benefit is no longer

sustained. It is interesting that some of them have had successful surgery on the STN. Although nowadays the STN is the preferred target for most of the surgical teams, until the results of a randomized-to-either-target, prospective, double-blind, multicenter study are available, the GPi could be considered in some patients in whom dyskinesias are the most relevant problem.

Adverse Events

Complications or adverse events (AEs) associated with DBS can be categorized as related to the surgery itself, to the implanted device, or to the stimulation and/or disease progression.

Surgical and Device-Related AEs

The most severe complication associated with the surgical procedure is the intraoperative bleeding which occurs in 1–5% of operated patients. Hemorrhages should be handled according to accepted neurosurgical practice. Misplacement of leads is an infrequent complication of the surgery. It is associated with lack of efficacy and significant stimulation-induced side effects. Another intervention for re-placement of the electrode may be needed.

Superficial infections of the surgical wounds can occur immediately after surgery. Erosions or infections in the skin covering the device (extension cable, IPG, cap of the burn hole) are not infrequent (from 1% to 15%), even months or years following surgery. They must be treated with intravenous or oral antibiotics with good tissue penetration. Deep infections involving purulent collections around hardware must be surgically treated, often requiring device removal. They may be related to systemic infections (sepsis), cellulitis, or skin erosion exposing the implant.

Device-related or hardware complications such as lead or extension wire fracture and consumption or malfunction of the IPG are frequent long-term complications. In these circumstances, loss of effect, intermittent stimulation, or tolerance can be the signs indicating the abnormal functioning of the device. Once the broken or malfunctioning component is identified, it must be replaced.

Stimulation and/or Disease-Related AEs

Stimulation AEs are problems that clearly arise because of the stimulation of the structures surrounding the nucleus or the nucleus itself such as dyskinesias in the case of STN and sometimes GPi. It has been reported that GPi DBS may induce dyskinetic movements when stimulating with more dorsal contacts. On the contrary, stimulation-induced akinesia was observed in the same patients when activating the most ventral contacts of the electrode. Persistent AEs, such as muscle contraction due to stimulation of the corticospinal fibers, can also be seen in both STN and GPi DBS. This effect is unbearable, requiring this particular programming to be ruled out.

The majority of stimulation-induced AEs has been observed in STN DBS. Among the most frequent ones are paresthesias, speech dysfunction (dysarthria and hypophonia), eyelid apraxia, and ocular and visual disturbances. In these cases, modulation of the parameters of stimulation can ameliorate the AEs, but if the result is not satisfactory, the parameters should be changed.

Weight gain is a common AE in STN-treated patients. Candidates for a surgical procedure may be given preoperative nutritional counseling to prevent excessive weight gain. Uncontrolled long-term observations suggest that the weight of most patients normalizes after 1–2 years.

Some other AEs are present with and without stimulation and are also frequent in PD patients not treated with DBS. Thus, the relation to stimulation or to the disease is not clear at this moment. Psychiatric, cognitive, and behavioral disturbances have been associated with STN DBS. Depression, apathy, abulia, desinhibition, mania and hypomania, hallucination, and other disturbances have been described. One important point to consider is the possibility that a large number of patients might already have psychiatric complications or cognitive deficits which could be worsened after surgery. Some studies indicate that patients with depression after STN DBS had depressive episodes before surgery. Nevertheless, cognitive and mood changes have been reported more frequently in STN DBS than in GPi DBS and may be related to the spreading of the current to limbic or associative portions of the nucleus or areas surrounding the STN.

Memory decline, psychiatric disturbances, and axial problems (gait and speech difficulties, disequilibrium, and falls) are also frequently described as AEs in patients treated with DBS but seem to be more frequent among those patients treated with DBS STN, as reported in the multicenter study at 4-year follow-up. It is interesting that all these complications are also frequently seen in advanced PD patients treated chronically only with pharmacological means, so they may be related to the disease. In this context, the multicenter study shows that for the STN DBS-treated patients, depression, thought disorders, rising from a chair, and gait were more affected at baseline in those patients who were to exhibit AEs at 4 years' follow-up. These data support the idea that the AEs exhibited

by STN-treated patients could be mainly due to the disease progression rather than to stimulation. At present no conclusion can be reached, and well-designed studies will be needed to clarify these crucial aspects in relation to the disease and to DBS therapy.

In summary, DBS for PD is a successful therapy with a good long-term evolution result. Adverse events associated with this therapy are frequent, but in the majority of cases, they do not preclude the continuation of the therapy. The greatest motor benefits associated with DBS of the STN have to be prospectively evaluated in studies with patients randomly operated in the STN or in the GPi. The recent impression that DBS of the STN produces more AEs than DBS of the GPi also needs to be confirmed. It is necessary to rule out the possibility that the disease evolution is the major contributor to AE.

See also: Animal Models of Parkinson's Disease; Deep Brain Stimulation; Deep Brain Stimulation and Movement Disorder Treatment; Dopamine; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopaminergic Agonists and L-DOPA; Parkinsonian Syndromes; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Anderson VC, Burchiel KJ, Hogarth P, Favre J, and Hammerstad JP (2005) Pallidal vs subthalamic nucleus deep brain stimulation in Parkinson disease. *Archives of Neurology* 62: 554–560.

Deuschl G, Herzog J, Kleiner-Fisman G, et al. (2006) Deep brain stimulation: Postoperative issues. *Movement Disorders* 21(supplement 14): S219–S237.

Deuschl G, Schade-Brittinger C, Krack P, et al. (2006) A randomized trial of deep-brain stimulation for Parkinson's disease. *New England Journal of Medicine* 355: 896–908.

Durif F, Lemaire JJ, Debilly B, and Dordain G (2002) Long-term follow-up of globus pallidus chronic stimulation in advanced Parkinson's disease. *Movement Disorders* 17: 803–807.

Ghika J, Villemure JG, Fankhauser H, Favre J, Assal G, and Ghika-Schmid F (1998) Efficiency and safety of bilateral contemporaneous pallidal stimulation (deep brain stimulation) in levodopa-responsive patients with Parkinson's disease with severe motor fluctuations: A 2-year follow-up review. *Journal of Neurosurgery* 89: 713–718.

Gross RE, Krack P, Rodriguez-Oroz MC, Rezai AR, and Benabid AL (2006) Electrophysiological mapping for the implantation of deep brain stimulators for Parkinson's disease and tremor. *Movement Disorders* 21(supplement 14): S259–S283.

Hariz MI (2002) Complications of deep brain stimulation surgery. *Movement Disorders* 17(supplement 3): S162–S166.

Joint C, Nandi D, Parkin S, Gregory R, and Aziz T (2002) Hardware-related problems of deep brain stimulation. *Movement Disorders* 17(supplement 3): S175–S180.

Kleiner-Fisman G, Herzog J, Fisman DN, et al. (2006) Subthalamic nucleus deep brain stimulation: Summary and meta-analysis of outcomes. *Movement Disorders* 21(supplement 14): S290–S304.

Kumar R, Lang AE, Rodriguez-Oroz MC, et al. (2000) Deep brain stimulation of the globus pallidus pars interna in advanced Parkinson's disease. *Neurology* 55(12 supplement 6): S34–S39.

Lang AE, Houeto JL, Krack P, et al. (2006) Deep brain stimulation: Preoperative issues. *Movement Disorders* 21(supplement 14): S171–S196.

Loher TJ, Burgunder JM, Pohle T, Weber S, Sommerhalder R, and Krauss JK (2002) Long-term pallidal deep brain stimulation in patients with advanced Parkinson disease: 1-year follow-up study. *Journal of Neurosurgery* 96: 844–853.

Østergaard K and AaSunde N (2006) Evolution of Parkinson's disease during 4 years of bilateral deep brain stimulation of the subthalamic nucleus. *Movement Disorders* 21: 624–631.

Rezai AR, Kopell BH, Gross RE, et al. (2006) Deep brain stimulation for Parkinson's disease: Surgical issues. *Movement Disorders* 21(supplement 14): S197–S218.

Rodriguez-Oroz MC, Obeso JA, Lang AE, et al. (2005) Bilateral deep brain stimulation in Parkinson's disease: A multicentre study with 4 years follow-up. *Brain* 128: 2240–2249.

Schupbach WM, Maltete D, Houeto JL, et al. (2007) Neurosurgery at an earlier stage of Parkinson disease: A randomized, controlled trial. *Neurology* 68: 267–271.

Visser-Vandewalle V, van der Linden C, Temel Y, et al. (2005) Long-term effects of bilateral subthalamic nucleus stimulation in advanced Parkinson disease: A 4 year follow-up study. *Parkinsonism & Related Disorders* 11: 157–165.

Volkmann J, Allert N, Voges J, Sturm V, Schnitzler A, and Freund HJ (2004) Long-term results of bilateral pallidal stimulation in Parkinson's disease. *Annals of Neurology* 55: 871–875.

Voon V, Kubu C, Krack P, Houeto JL, and Troster AI (2006) Deep brain stimulation: Neuropsychological and neuropsychiatric issues. *Movement Disorders* 21(supplement 14): S305–S327.

Welter ML, Houeto JL, Tezenas du Montcel S, et al. (2002) Clinical predictive factors of subthalamic stimulation in Parkinson's disease. *Brain* 125: 575–583.

Delayed Reinforcement: Economics

C Harris, University of Cambridge, Cambridge, UK

© 2009 Published by Elsevier Ltd.

Introduction

In general, an economic subject faces many dimensions of choice. How much should she consume, and when should she consume it? What sort of career should she choose, and what sort of education and training should she undertake in order to prepare best for that career? What sort of spouse should she choose? Should she have children, and if so how many and when? How much should she work, and when should she retire? In which, among the bewildering array of assets on offer, should she place her savings?

This article focuses on just one of these dimensions: when should a subject consume? In other words, the focus is on the essence of intertemporal choice: time. The author does not consider the question of how much she should consume. The author does, however, allow for uncertainty. This is essential, for three reasons. First, in practice, most problems of intertemporal choice involve uncertainty. Second, even in the laboratory, there may be some uncertainty in a subject's mind as to whether she will actually receive the rewards being promised to her. Third, it is not possible to make sense of even the most basic concepts of the theory of intertemporal choice such as discount factors and discount rates unless the subject makes choices involving uncertainty.

Both economists and psychologists frequently analyze problems of intertemporal choice in terms of a subject's discount function. It is therefore important to begin by understanding what the use of a discount function does and does not entail. One thing that it does not entail is an assumption that the subject actually has a discount function. The point is that she behaves as if she had a discount function. The discount function itself is simply a convenient device that quantifies her choices. What the use of a discount function does entail is three assumptions about the nature of her choices: they must be complete, transitive, and substitutable. This is because a discount function is simply a particular kind of utility function: it assigns a utility to each date at which the subject could consume.

The three requirements of completeness, transitivity, and substitutability are controversial from both a theoretical and an empirical point of view. It is therefore somewhat surprising that psychologists are willing to work with discount functions. Nonetheless, there is one respect in which they are rather permissive: they are satisfied by essentially any discount function. Narrowing the range of possible shapes for the discount function therefore seems to be a higher priority than widening the enquiry to include behavior that is not consistent with having a discount function. Psychologists have taken an experimental approach to this problem, and they are inclined to interpret the data as suggesting that discount functions are hyperbolic. Following Samuelson, economists were inclined to assume for conceptual and mathematical convenience that discount functions were exponential. However, with the arrival of more powerful computers and more powerful analytical techniques, they have been increasingly willing to respond to the challenge to this paradigm contained in Strotz.

Much of the attraction of exponential discounting derives from the fact that it satisfies both dynamic consistency and stationarity. These requirements can be expressed as follows. Suppose that we are given two dates, $t > s > 0$. Then dynamic consistency means that, if the subject decides at date 0 that she would like to consume at date t, then she would not want to revise her decision at date s if given the opportunity to do so. Stationarity means that, if the subject decides at date 0 that she would like to consume at date t, then she would decide a date s to consume at date $s + t$.

Consider the following example. At 2 p.m., a subject is asked when she would like to eat supper. She chooses 7 p.m. Two hours later (i.e., at 4 p.m.) she is unexpectedly given the option of revising her choice. If she still chooses 7 p.m., then she is acting in a dynamically consistent manner. If, however, she chooses a different time (say 6 p.m.), then she is acting in a dynamically inconsistent manner. This change could be explained by the arrival of new information. For example, she may have learned at 3 p.m. that her 5 p.m. appointment had been cancelled. In this case, the dynamic inconsistency arises because she was asked to make an unconditional choice. Had she instead chosen a pair of times for supper, one to apply in the event that the appointment was on and the other to apply in the event that it was off, then she would have chosen the pair {7 p.m. if appointment, 6 p.m. if no appointment} and no revision in her choice would have been necessary. The dynamic consistency requirement embodied in an exponential discount function does not rule out changes that arise from the arrival of new information. It simply rules out changes that occur in the complete absence of any new information. As such, it is more reasonable than it might appear at first sight.

Now consider the same example, but from the point of view of stationarity. Stationarity requires that, when the subject is given the option of revising

her choice at 4 p.m., she chooses 9 p.m.! The point is that, at 2 p.m., she wanted to have supper 5 h later. So, at 4 p.m., she should again want to have supper 5 h later. Stated in this way, the requirement of stationarity appears ridiculous: if the subject wanted to delay supper by 5 h at 2 p.m., then, at 4 p.m. when she is presumably hungrier, she should want (at the very minimum) to delay supper by less than 5 h. The problem does not lie with stationarity, however. It lies with an incompletely specified model. If hunger is an important factor in the subject's decision making, then it should be included in the model. Once it is included, the subject's choices become stationary again: for any given level of hunger, she will choose the same delay until supper. So stationarity too is more reasonable than it may appear at first sight.

Unfortunately, the exponential discounting model appears to be rejected by the experimental data. Dynamic consistency, stationarity, or both must therefore be wrong. From an experimental point of view, it would be hard to abandon stationarity: experiments are ultimately based on the idea that subjects will keep on doing the same thing if they are placed repeatedly in the same situation. It is therefore more sensible to abandon dynamic consistency, at least in its present form.

The experimental data also suggest that subjects are more patient with respect to any given delay, the further in the future that delay occurs. The present article proposes a method for deriving discount functions with this property and uses the method to derive a unified discount function that incorporates all three of the most popular discount functions (namely, exponential, hyperbolic, and generalized hyperbolic) as special cases.

Finally, if dynamic consistency is abandoned, then a subject's decision problem becomes an intrapersonal game. In this game, the interests of the future self are not necessarily aligned with those of the current self, and the current self has to try to anticipate the behavior of the future self. Anticipation is made easier by stationarity, which ensures that the outlook of the future self will be identical to that of the current self, but made harder by the fact that the future self may face circumstances different from those faced by the current self. Two extremes can be identified: a naive subject fails completely to anticipate the behavior of her future self (and fails to notice that she is failing to anticipate), and a sophisticated subject anticipates the behavior of her future self perfectly.

An Idealized Example

Consider the following idealized example of choice over delayed outcomes. Suppose that there is an infinite sequence of possible dates at which the subject can eat an apple: $t = 0$ (today), $t = 1$ (tomorrow), $t = 2$ (the day after tomorrow), and so on. Denote the set of all possible dates by $\mathbb{T} = \{0, 1, 2, \ldots\}$; define a probability distribution p over $\mathbb{T}$ to be simple if the number of dates t for which $p(t) > 0$ is finite; and denote by $\mathbb{P}$ the set of simple probability distributions (SPDs) over $\mathbb{T}$. The idea is then that:

1. The experimenter will offer the subject a choice of two SPDs.
2. The subject will choose between them.
3. The experimenter will use a suitable randomizing device to generate a date t according to the SPD chosen.
4. The subject will then wait until that date, at which point the experimenter will give her the apple and she will eat it straight away.

Faced with this problem, it can be argued that the perfect subject should proceed as follows. First, before she is actually asked to make any choice, she should formulate her preferences as between any pair of SPDs she could conceivably be offered. In other words, for any pair of SPDs p_1 and p_2, she must decide that (i) p_1 is at least as good as p_2; or (ii) p_2 is at least as good as p_1; or (iii) p_1 and p_2 are equally good. Second, her preferences should be consistent in the sense that, if she thinks that p_1 is at least as good as p_2 and that p_2 is at least as good as p_3, then she must also think that p_1 is at least as good as p_3. Third, if we start with three SPDs p_1, p_2 and p_3, and if we create two compound probability distributions $\tilde{p}_1$ (in which p_1 is chosen with probability γ and p_3 is chosen with probability $1-\gamma$) and $\tilde{p}_2$ (in which p_2 is chosen with probability γ and p_3 is chosen with probability $1-\gamma$), then the subject should again be consistent, this time in the sense that if she thinks that p_1 is at least as good as p_2 then she should also think that $\tilde{p}_1$ is at least as good as $\tilde{p}_2$. The logic behind this requirement is that, in moving from $\tilde{p}_2$ to $\tilde{p}_1$, we simply replace the intermediate outcome p_2 with the better intermediate outcome p_1. Substitutability, as defined later, incorporates a second requirement: provided that the probability distribution over final dates is the same, the subject should not care whether the probabilistic procedure used to arrive at this probability distribution takes place in two stages (as in a compound probability distribution) or in one (as in a simple probability distribution). This requirement is often called the reduction of compound lotteries.

More precisely, if we express the statement that the subject thinks that p_1 is at least as good as p_2 by writing $p_1 \text{ B } p_2$, then it can be argued that the subject's preferences should satisfy the following three requirements:

Completeness: For all $p_1, p_2 \in \mathbb{P}$, we have either p_1 B p_2 or p_2 B p_1 or both.
Transitivity: For all p_1, p_2, $p_3 \in \mathbb{P}$, if p_1 B p_2 and p_2 B p_3, then p_1 B p_3.
Substitutability: For all p_1, p_2, $p_3 \in \mathbb{P}$ and all $0 < \gamma \leq 1$, if we put $\tilde{p}_1 = \gamma p_1 + (1-\gamma)p_3$ and $\tilde{p}_2 = \gamma p_2 + (1-\gamma)\, p_3$, then p_1 B p_2 if and only if $\tilde{p}_1$ B $\tilde{p}_2$.

These requirements can be challenged on both theoretical and empirical grounds. For example, why would the subject want to evaluate the vast number of hypothetical alternatives envisaged by the requirement of completeness? Why not simply wait until she is actually asked to make a choice, and only then evaluate the pair of dates that she is actually offered? Moreover, if she evaluates only a single pair of dates, then the consistency issue addressed by the requirement of transitivity – which in general involves three pairs of dates – never arises. Finally, the requirement of substitutability (for all its theoretical appeal) has not fared well when subjected to experimental testing. A full discussion of these challenges would, however, take us too far away from our main theme.

If the requirements of completeness, transitivity, substitutability, and a fourth technical requirement are all satisfied, then the following result holds:

Expected-Utility Theorem: There exists a utility function $u : \mathbb{T} \to \mathbb{P}$ that represents the subject's preferences in the sense that p_1 B p_2 if and only if

$$\sum_{t \in \mathbb{T}} p_1(t)\, u(t) \geq \sum_{t \in \mathbb{T}} p_2(t)\, u(t)$$

The fourth requirement says roughly that, if the subject strictly prefers p_1 to p_2, and if p_3 is a third simple probability distribution, then there is a small positive probability γ such that: (1) the subject strictly prefers the compound probability distribution $\widehat{p}_1$ (in which p_1 is chosen with probability $1-\gamma$ and p_3 is chosen with probability γ) to p_2; and (2) the subject strictly prefers p_1 to the compound probability distribution $\widehat{p}_2$ (in which p_2 is chosen with probability $1-\gamma$ and p_3 is chosen with probability γ). More precisely:

Archimedean axiom for all p_1, p_2, $p_3 \in \mathbb{P}$ such that p_1 B p_2 but not p_2 B p_1 (i.e., p_1 is strictly preferred to p_2), there exists $\gamma \in (0, 1)$ such that, if we put $\widehat{p}_1 = (1-\gamma)\, p_1 + \gamma\, p_3$ and $\widehat{p}_2 = (1-\gamma)\, p_2 + \gamma\, p_3$, then: (1) $\widehat{p}_1$ B p_2. but not p_2 B $\widehat{p}_1$ (i.e., $\widehat{p}_1$ is strictly preferred to p_2); and (2) p_1 B $\widehat{p}_2$ but not $\widehat{p}_2$ B p_1 (i.e., p_1 is strictly preferred to $\widehat{p}_2$).

We shall not discuss this requirement further.

In other words, there exists a utility function that characterizes the subject's choices in the sense that she chooses p_1 over p_2 if and only if the expected utility associated with p_1 exceeds the expected utility associated with p_2. (The expected utility of an SPD p is calculated by multiplying the probability $p(t)$ with which the date t is generated by the utility $u(t)$ associated with that date, and then adding up over all dates that occur with positive probability.) This result does not say that the subject has a utility function, only that she behaves as if she had one.

The utility function provided by the Expected-Utility Theorem is not unique: for all scalars $\alpha \in \mathbb{R}$ and $\beta > 0$, the utility function $\upsilon : \mathbb{T} \to \mathbb{R}$ defined by the formula $\upsilon(t) = \alpha + \beta\, u(t)$ does just as well. However, this is also the only sense in which u is not unique. More precisely, the following result holds:

Cardinal-Information Theorem: If there is a second utility function $\upsilon : \mathbb{T} \to \mathbb{R}$ that also represents the subject's preferences, then there exist scalars $\alpha \in \mathbb{R}$ and $\beta > 0$ such that $\upsilon(t) = \alpha + \beta\, u(t)$ for all $t \in \mathbb{T}$.

This result is essential for what follows. Indeed, important concepts in the analysis of intertemporal choice such as the discount factor and the discount rate are defined directly in terms of the utility function. Yet many different utility functions can be used to represent the subject's choices. We therefore need to make sure that any such concept is independent of which utility function is used to represent the subject's choices. In this way we can guarantee that it has a meaning directly in terms of what we observe, namely, the subject's behavior.

Consider the following analogy. We can measure temperature on the Celsius scale (corresponding to u) or on the Fahrenheit scale (corresponding to υ). Water freezes at 0° and boils at 100° on the Celsius scale, and it freezes at 32° and boils at 212° on the Fahrenheit scale. So statements like "the temperature is 0°" or "the temperature increased by 100°" do not make sense until one specifies whether the temperature scale is Celsius or Fahrenheit. However, statements like "when he heated the pan with a Bunsen burner the temperature rose by twice as much as when he heated it with a candle" make sense even if we do not know which temperature scale is being used. It is to statements like these that we must therefore confine ourselves. (The definitions of discount factor and discount rate given below both conform to this pattern.) Finally, in order to move between the Celsius scale and the Fahrenheit scale, we need to set $\alpha = 32$ and $\beta = \frac{9}{5}$.

Dynamic Consistency and Stationarity

Up to this point, we have assumed that, at date 0, the subject chooses a date at which to consume. She is then committed to that date. This is rather artificial: in an experimental setting, the experimenter can provide the external discipline that ensures that the subject really is committed to wait until the chosen

date, but in everyday settings, a subject is likely to have opportunities to reconsider her choices.

Suppose by way of an example that $u(1) > u(2) > u(0) > u(3) > u(4) > \ldots$. In other words, while the subject knows that apples are good for her, given the choice of eating the apple today ($t = 0$) or tomorrow ($t = 1$), she would rather eat it tomorrow. Similarly, given the choice of eating the apple today ($t = 0$) or the day after tomorrow ($t = 2$), she would rather eat it the day after tomorrow. However, given the choice of eating the apple today ($t = 0$) or 3 days from now ($t = 3$), she would rather eat it today: even though she does not particularly like apples, she recognizes that she really must eat an apple from time to time.

Suppose further that the experimenter offers the subject a choice of eating the apple today ($t = 0$) or tomorrow ($t = 1$). Then the subject will certainly choose tomorrow. Suppose, however, that, when tomorrow actually arrives, the experimenter offers the subject a new choice: she can either have the apple right away (i.e., at $t = 1$, as originally planned), or she can have it on the following day (i.e., at $t = 2$). On the face of it, the subject is better off than she was before: she still has the option that she originally chose, but in addition she has the option of waiting another day. What will she do?

If we accept the logic of the preceding two sections, then the answer is that she should formulate a new utility function $\tilde{u}$ over the set $\tilde{\mathbb{T}} = \{1, 2, 3, \ldots\}$ of remaining dates. The question, then, is what relationship should this new utility function bear to the old utility function u? There does not seem to be a compelling answer to this question, but there are three possibilities.

Dynamic Consistency

One possible answer is that the subject will rank the dates $1, 2, 3, \ldots \in \tilde{\mathbb{T}}$ in exactly the same way that she ranked the dates $1, 2, 3, \ldots \in \tilde{\mathbb{T}}$. More explicitly, for all $t_1, t_2 \in \tilde{\mathbb{T}}$, $\tilde{u}(t_1) \geq \tilde{u}(t_2)$ if and only if $u(t_1) \geq u(t_2)$. In other words, the subject has preferences over absolute (or calendar) dates, and these preferences do not change with the passage of time. In the context of the example given above, that means that if (at date 0) the subject preferred to eat the apple at date 1, then (at date 1) the subject will still prefer to eat the apple at date 1.

Stationarity

Another possible answer is that the subject's preferences are really a kind of perspective on the future. The subject will therefore rank the dates $1, 2, 3, \ldots \in \tilde{\mathbb{T}}$ in exactly the same way that she ranked the dates $0, 1, 2, \ldots \in \mathbb{T}$. More explicitly, for all $t_1, t_2 \in \tilde{\mathbb{T}}$, $\tilde{u}(t_1) \geq \tilde{u}(t_2)$ if and only if $u(t_1 - 1) \geq u(t_2 - 1)$. In other words, the subject has preferences over relative dates, and these preferences do not change with the passage of time. In the context of the example given previously, that means that if (at date 0) the subject preferred to eat the apple 1 day later (i.e., at date 1), then (at date 1) the subject still prefers to eat the apple 1 day later (i.e., at date 2).

Exponential Discounting

Dynamic consistency ensures that a subject never wants to revise plans formulated at date 0, even if she is given the opportunity to do so. Stationarity means that a subject will approach a new planning problem in exactly the same way that she approached earlier planning problems. Can both of these advantages be incorporated within a single model? The answer is that they can, but only if the discount function takes a very particular form: it must be exponential.

In order to explain why, note that dynamic consistency implies that u and $\tilde{u}$ are simply two different representations for the same preferences over the dates in $\tilde{\mathbb{T}} = \{1, 2, 3, \ldots\}$. The Cardinal-Information Theorem therefore implies that we can find scalars $\alpha_{DC} \in \mathbb{R}$ and $\beta_{DC} > 0$ such that

$$\tilde{u}(t) = \alpha_{DC} + \beta_{DC}\, u(t) \qquad [1]$$

for all $t \in \tilde{\mathbb{T}}$. Similarly, stationarity implies that there are scalars $\alpha_S \in \mathbb{R}$ and $\beta_S > 0$ such that

$$\tilde{u}(t) = \alpha_S + \beta_S\, u(t-1) \qquad [2]$$

for all $t \in \tilde{\mathbb{T}}$. Inverting eqn [1] to express $u(t)$ in terms of $\tilde{u}(t)$, and then using eqn [2] to substitute for $\tilde{u}(t)$, we obtain

$$u(t) = \alpha_E + \beta_E\, u(t-1) \qquad [3]$$

for suitable scalars $\alpha_E \in \mathbb{R}$ and $\beta_E > 0$. (The scalars in question are $\alpha_E = \beta_{DC}^{-1}(\alpha_S - \alpha_{DC})$ and $\beta_E = \beta_{DC}^{-1}\beta_S$.)

Now, eqn [3] is a simple difference equation. If $\beta_E \neq 1$, then it has the unique solution

$$u(t) = \frac{\alpha_E}{1-\beta_E} + \beta_E^t\left(u(0) - \frac{\alpha_E}{1-\beta_E}\right) \qquad [4]$$

and, if $\beta_E = 1$, then it has the unique solution

$$u(t) = u(0) + t\,\alpha_E \qquad [5]$$

At first sight, these equations may appear to be rather complicated. Nothing could be further from the truth: they can be simplified dramatically by choosing a more convenient utility scale. More precisely, we can replace u with an equivalent utility function $\upsilon = \alpha + \beta\, u$ for suitably chosen α and β. If $\beta_E \neq 1$ and $u(0) = \frac{\alpha_E}{(1-\beta_E)} \neq 0$ then one uses $\alpha = -\frac{\alpha_E}{(1-\beta_E)}$ and $\beta = \left|u(0) - \frac{\alpha_E}{(1-\beta_E)}\right|^{-1}$; if

$\beta_E \neq 1$ and $u(0) - \frac{\alpha_E}{(1-\beta_E)} = 0$ then one uses $\alpha = -\frac{\alpha_E}{(1-\beta_E)}$ and $\beta = 1$; if $\beta_E = 1$ and $\alpha_E \neq 0$ then one uses $\alpha = -u(0)$ and $\beta = |\alpha_E|^{-1}$; and if $\beta_E = 1$ and $\alpha_E = 0$ then one uses $\alpha = -u(0)$ and $\beta = 1$. In this way, we find that there are essentially only five possible choices for the subject's utility function:

1. $\upsilon(t) = \beta_E^t$ for some $\beta_E \neq 1$;
2. $\upsilon(t) = -\beta_E^t$ for some $\beta_E \neq 1$;
3. $\upsilon(t) = 0$;
4. $\upsilon(t) = t$; and
5. $\upsilon(t) = -t$.

These cases occur, respectively, if (1) $\beta_E \neq 1$ and $u(0) - \frac{\alpha_E}{(1-\beta_E)} > 0$; (2) $\beta_E \neq 1$ and $u(0) - \frac{\alpha_E}{(1-\beta_E)} < 0$; (3) either $\beta_E \neq 1$ and $u(0) - \frac{\alpha_E}{(1-\beta_E)} = 0$, or $\beta_E = 1$ and $\alpha_E = 0$; (4) $\beta_E = 1$ and $\alpha_E > 0$; and (5) $\beta_E = 1$ and $\alpha_E < 0$. Cases 1, 2, and 3 could be summarized by saying that, viewed from the perspective of date 0, the utility of eating the apple at date t is the product of two factors: (1) an instantaneous utility (i.e., a utility that is obtained at the actual moment of consumption) that takes one of the three values 1, −1, and 0; and (2) a discount factor $\beta_E \neq 1$ raised to the power of t. There is nothing in the theory that we have sketched that requires that $\beta_E < 1$, so β_E can take any value in $(0,1) \cup (1, \infty)$. Cases 4 and 5 are limiting cases of cases 1 and 2, obtained when β_E tends to 1. This is easier to see when preferences are represented by the utility function u. Indeed, rewriting eqn [4] in the form $u(t) = \frac{(1-\beta_E^t)}{(1-\beta_E)}\ \alpha_E + \beta_E^t\ u(0)$, we find that $u(t) \to t\ \ \alpha_E + u(0)$ as $\beta_E \to 1$. In other words, we obtain eqn [5]. There are other ways of taking limits. For example, if we let $\beta_E \to 1$ in case 1, then $\widehat{u}(t) \to 1$. This is, of course, case 3, the case in which the subject is indifferent among all dates. Similarly, if we let $\beta_E \to 1$ in case 2, then $\widehat{u}(t) \to -1$. This is again case 3.

The Discount Factor

Intuitively speaking, the discount factor at date t tells us what fraction of the utility gain from an increase in consumption at date t the subject would still obtain if that increase in consumption were deferred until date $t+1$. This is a measure of the subject's patience at date t: the bigger the fraction of the utility gain that she still gets, the more patient she can be said to be.

Now, in our model, there are only two possible levels of consumption at date t: either the subject eats one apple or she eats no apples. However, to say that she eats no apples is ambiguous: it leaves open the question of when she does eat the apple. The best way to resolve this ambiguity is to introduce a new possibility into our model, namely, that the subject never eats the apple. This possibility can be captured by adding the new date ∞ (i.e., infinity) to the existing list of dates $\mathbb{T} = \{0, 1, 2, \ldots\}$. Choosing $t \in \mathbb{T}$ then means choosing to eat the apple at date t, and choosing $t = \infty$ means choosing never to eat the apple. It is also helpful to impose the requirement that $u(t)$ converges to $u(\infty)$ as t gets large. This ensures that deferring the consumption of the apple indefinitely and never consuming it yield essentially the same utility. Having to introduce the date ∞ into the model is the one small price that we have to pay for eliminating the level of consumption from the model.

We can now make precise the definition of the discount factor outlined previously. Indeed, the utility gain from eating the apple at date $t+1$ instead of never eating it is $u(t+1) - u(\infty)$, the utility gain from eating the apple at date t instead of never eating it is $u(t) - u(\infty)$, and the discount factor at date t is

$$\delta(t, 1) = \frac{u(t+1) - u(\infty)}{u(t) - u(\infty)}$$

For example, if $u(t) = \beta_E^t$ for some $\beta_E < 1$, then $u(\infty) = 0$ and $\delta(t, 1) = \beta_E$ for all t. Notice that $\delta(t, 1)$ is unchanged if we replace u by the equivalent utility function $\upsilon = \alpha + \beta u$. The discount factor is therefore a genuine behavioral measure.

The Discount Rate

In order to derive some alternatives to the exponential-discounting model, it will be helpful to treat t as a continuous variable. In the theory that we have sketched, a single indivisible good is consumed once. It is therefore relatively easy to adapt the theory to the continuous-time case. We can then ask what fraction of the utility gain from an increase in consumption at date t the subject would still obtain if that increase in consumption were deferred until date $t + \tau$. More precisely, we can introduce a generalization of the discount factor, namely,

$$\delta(t, \tau) = \frac{u(t+\tau) - u(\infty)}{u(t) - u(\infty)}$$

This is again a measure of the subject's patience at date t. Like $\delta(t, 1)$, $\delta(t, \tau)$ is unchanged if we replace u by the equivalent utility function $\upsilon = \alpha + \beta u$. It too is therefore a genuine behavioral measure.

Now, we would ideally like to make τ as small as possible, to ensure that we measure the subject's patience at date t as precisely as possible. However, $\delta(t, \tau) \to 1$ as $\tau \to 0$. We therefore look instead at the discount rate

$$\rho(t) = \lim_{\tau \to 0} \frac{1 - \delta(t, \tau)}{\tau}$$

This tells us how quickly $\delta(t, \tau)$ falls below 1 as the delay τ grows from 0. It is therefore a measure of the subject's impatience at date t. Elementary calculus shows that

$$\rho(t) = \frac{-u'(t)}{u(t) - u(\infty)} \qquad [6]$$

For example, if $u(t) = \beta_E^t$ for some $\beta_E < 1$, then $\rho(t) = -\ln(\beta_E) > 0$.

In some respects, the terminology of discount factor and discount rate is unfortunate: it might have been better to use the terms patience factor and impatience rate instead.

Beyond Exponential Discounting

According to the exponential-discounting model, the subject's discount rate ρ is constant. However, if there is one thing on which the experimentalists all agree, it is that ρ decreases with time: a given delay is regarded as less serious, the further in the future it occurs.

One way of modeling this stylized fact is to make $\rho(t)$ a weighted average of a short-run discount rate λ and a long-run discount rate μ in which the weight placed on μ increases with time. For example, we could put

$$\rho(t) = \frac{1}{1+\nu t}\lambda + \frac{\nu t}{1+\nu t}\mu \qquad [7]$$

Here ν is a strictly positive parameter which measures the speed with which the discount rate moves from its short-run value λ to its long-run value μ, and the weight placed on μ increases from 0 when $t = 0$ to 1 when $t = \infty$.

If we substitute this formula for $\rho(t)$ in eqn [6], then we obtain a simple differential equation for the utility function u, $(1+\nu t)\,u'(t) + (\lambda + \mu\nu t)\ (u(t) - u(\infty)) = 0$. It is easy to verify that this equation has the general solution

$$u(t) = C\ \exp(-\mu t)(t+\nu t)^{-\frac{(\lambda-\mu)}{\nu}} + u(\infty) \qquad [8]$$

where C is an arbitrary constant. This solution is consistent with our requirement that $u(t) \to u(\infty)$ provided that either $\mu > 0$ and $\lambda \in \mathbb{R}$, or $\mu = 0$ and $\lambda > 0$. Assuming this to be the case, we can choose the utility scale in such a way that $u(\infty) = 0$. There are then three basic possibilities: $C = 1$, $C = -1$, and $C = 0$. Focusing on the case $C = 1$, we obtain

$$u(t) = \exp(-\mu t)\ (t+\nu t)^{-\frac{(\lambda-\mu)}{\nu}} \qquad [9]$$

This is an attractively simple functional form with both exponential and hyperbolic features: $u(t)$ is the product of the exponential discount factor $\exp(-\mu t)$ and the hyperbolic discount factor $(1+\nu t)^{-\frac{\lambda-\mu}{\nu}}$. Moreover, the parameters λ, μ, and ν are all readily interpretable.

Equation [9] is sufficiently general to encompass both exponential and hyperbolic discounting. Indeed, if we set $\mu = 0$, then we obtain

$$u(t) = (1+\nu t)^{-\frac{\lambda}{\mu}} \qquad [10]$$

This is the generalized hyperbola of Harvey and Loewenstein and Prelec. And, if we specialize further by putting $\lambda = \nu$, then we obtain

$$u(t) = (1+\nu t)^{-1} \qquad [11]$$

This is the original hyperbola of Herrnstein. Third, if we put $\lambda = \mu$, then we obtain

$$u(t) = \exp(-\mu t) \qquad [12]$$

This is the exponential used by Samuelson.

Equation [9] also offers a flexible functional form for statistical testing. For example, while it is widely thought that the discount rate $\rho(t)$ is decreasing in t, it would be desirable to be able to know the statistical significance of this finding. Because eqn [9] is valid for both $\lambda - \mu > 0$ and $\lambda - \mu < 0$, it can be used as the basis of a one-tailed test of whether $\lambda - \mu > 0$ (as expected) or not. Similarly, while it is widely thought that discount functions are not exponential, it would be desirable to know the statistical significance of this finding. Equation [9] can be used as the basis of a two-tailed test of whether $\lambda - \mu = 0$ or not (as expected). Third, it is sometimes claimed that discount rates are close to 0 for large t. Because eqn [9] is valid for all $\mu \geq 0$ (provided at least that $\lambda > 0$), it can be used as the basis of a test of whether the restriction $\mu = 0$ is rejected by the data or not.

Equation [9] does, however, have one fault. Except in the special case where $\lambda = \mu$ (i.e., the case of exponential discounting), it is difficult to incorporate into theoretical models in a mathematically tractable way. This difficulty can be overcome. Notice that there are essentially four elements in the construction of eqn [9]: (1) the short-run discount rate λ, (2) the long-run discount rate μ, (3) the pattern of adjustment $t \to \frac{t}{1+t}$, and (iv) the speed of adjustment ν. The trick is to change the (seemingly simple) pattern of adjustment

$$t \to \frac{t}{1+t}$$

into the (seemingly complex) pattern of adjustment

$$t \to \frac{\theta(1-\exp(-t))}{\exp(-t) + \theta(1-\exp(-t))}$$

where θ is a nonnegative parameter. (Notice that θ is not the rate of adjustment.) Making this change leads to the (thankfully simple!) discount function

$$u(t) = \exp(-\mu t)\ (\exp(-\nu t) + (1-\exp(-\nu t))\theta) \qquad [13]$$

introduced by Harris and Laibson. This discount function is the product of two factors: the exponential discount function $\exp(-\mu t)$ and the supplementary discount function $\exp(-\nu t) + (1 - \exp(-\nu t))\,\theta$. The supplementary discount function attaches weight 1 to date 0 and weight θ to date ∞, and the difference between these two weights decays exponentially at rate ν. If $\theta < 1$, there is therefore a transparent sense in which the present is overweighted relative to the future.

The idea behind eqn [13] is to divide time into two phases. During both of these phases, discounting takes place at rate μ. However, at the point of transition between them, an additional discount factor θ is applied. Finally, the transition between them takes place at rate ν. The perceptive reader may wonder why we do not assume that discounting takes place at different rates in the two phases. For example, why not assume that discounting takes place at rate $\widehat{\lambda}$ during the first phase, that discounting takes place at rate $\widehat{\mu}$ during the second phase, that an additional discount factor $\widehat{\theta}$ is applied at the transition between the two phases, and that the transition between the two phases takes place at rate $\widehat{\nu}$? The answer is that this leads to no increase in generality: under these assumptions the discount function still takes the form given in eqn [13] with $\theta = \frac{\widehat{\beta}\widehat{\nu}}{(\widehat{\nu}+\widehat{\lambda}-\widehat{\mu})}$, $\mu = \widehat{\mu}$, and $\nu = \widehat{\nu} + \widehat{\lambda} - \widehat{\mu}$. In other words, one of the parameters $\widehat{\theta}$, $\widehat{\lambda}$, $\widehat{\mu}$ and $\widehat{\nu}$ is redundant. This construction is a compromise between the homogeneity of exponential discounting (which involves only a single phase) and the complete inhomogeneity of discount functions like that given in eqn [9] (which involve a continuum of phases).

Intrapersonal Games

If exponential discounting is rejected by the data, then dynamic consistency, stationarity, or both must be wrong. The author concludes this article by exploring the theoretical implications of abandoning dynamic consistency, but retaining stationarity, doing this in the context of an example.

Suppose that $u(1) > u(2) > u(0) > u(3) > u(4) > \cdots > u(\infty)$. This is the same as the example in the section titled 'Dynamic consistency and stationarity,' except that the new date ∞ has been added. Suppose, however, that the subject is now free to choose any date she wishes.

One possible approach to this problem is to assume that the subject is naive. At date 0, the subject's preferences are represented by the utility function $u_0 = u$. More explicitly, we have $u_0(1) > u_0(2) > u_0(0) > u_0(3) > u_0(4) > \cdots > u_0(\infty)$. She will therefore plan to consume at her most preferred date, namely, $t = 1$. However, at date 1, stationarity implies that her preferences over future dates can be represented by the utility function u_1 given by the formula $u_1(t) = u_0(t-1)$. More explicitly, we have $u_1(2) > u_1(3) > u_1(1) > u_1(4) > u_1(5) > \ldots > u_1(\infty)$. She will therefore plan to consume at her most preferred date, which is now $t = 2$. And so on, indefinitely.

This reads like a textbook example of procrastination: every day the subject plans to eat the apple the following day, with the result that she never eats it. The weakness of the analysis is that the subject never learns from her repeated failure to eat the apple that her plan to eat the apple the following day is unrealistic.

A second possible approach is to assume that the subject is sophisticated. From the very outset, she recognizes that she cannot bind her future self to take the actions that she deems desirable today. So, instead, she tries to anticipate the choice c_t that she will make at date t in case she has not already consumed the apple. In this case, there are three stable patterns of behavior. They are summarized in Table 1. Each column of this table contains the choices made at a particular date. The first three rows contain the choices made in the three possible behaviors of the sophisticated subject. The fourth row contains the choices made by a naive subject, for the purposes of comparison. A value of 1 denotes that the subject decides to eat the apple, and a value of 0 denotes that she decides not to eat it.

If the behavior of the sophisticated subject follows the first pattern, then she ends up consuming the apple straight away. This is in marked contrast to the naive subject, who never consumes it. This pattern of behavior is supported by the following logic.

At date 0, the subject recognizes that if she does not eat the apple immediately, then (because she will not

Table 1

Date	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	...
Behavior 1 of sophisticated subject	1	0	0	1	0	0	1	0	0	1	0	...
Behavior 2 of sophisticated subject	0	1	0	0	1	0	0	1	0	0	1	...
Behavior 3 of sophisticated subject	0	0	1	0	0	1	0	0	1	0	0	...
Behavior of naive subject	0	0	0	0	0	0	0	0	0	0	0	

eat it at dates 1 or 2) she will end up eating it at date 3. Since $u(0) > u(3)$ (and therefore $u_0(0) > u_0(3)$), she decides to eat the apple straight away. Similarly, at date 1, she recognizes that if she does not eat the apple immediately, then (because she will not eat it at date 2) she will end up eating it at date 3. Since $u(2) > u(0)$ (and therefore $u_1(3) > u_1(1)$, by stationarity), she decides not to eat the apple. Similarly, at date 2, she recognizes that if she does not eat the apple immediately, then she will eat it at date 3. Since $u(1) > u(0)$ (and therefore $u_2(3) > u_2(2)$, by stationarity), she decides not to eat the apple. And so on.

Notice that the naive subject ends up with the lowest possible utility, namely, $u(\infty)$, and that (under behavior 1) the sophisticated subject ends up with the better utility $u(0)$. Sophistication, therefore, does pay in this example. However, the sophisticated subject does better still under behavior 3 (which results in a utility of $u(2)$) and best of all under behavior 2 (which results in a utility of $u(1)$).

Five other points are worth noting. First, we do not obtain a unique prediction for the behavior of a sophisticated subject. This is typical of all but the simplest games. But we do still get clear predictions. For example, a naive subject will never eat the apple, whereas a sophisticated subject will eat it within the first three periods. Second, although the model is stationary, none of the three predictions for the behavior of the sophisticated subject is stationary. While it may not be reasonable to rule out behaviors that are nonstationary, it is reasonable to look for behaviors that are. One way of obtaining stationary behavior is to allow the subject to randomize. Third, randomization can be given an attractive interpretation in this model. It need not mean that the subject deliberately randomizes between eating and not eating the apple (e.g., by tossing a coin). It could mean instead that she is unsure whether she will eat the apple tomorrow or not: all she knows is that she will eat it with some probability. Fourth, the game-theoretic approach we have adopted in the analysis of the sophisticated subject effectively reinstates dynamic consistency in a more subtle form. Previously, a subject was deemed to be dynamically consistent if the future self would always find it to be in her own interests to go along with any plan formulated by the current self. Now the current self recognizes that the interests of the future self may not coincide with her own and chooses the best plan that the future self will be willing to go along with. Finally, the behavior of a naive subject is absurdly naive: even pigeons can do better than that. Similarly, the behavior of the sophisticated subject is absurdly sophisticated: even the most intelligent human being would struggle to foresee as much as she does. The truth probably lies somewhere in between.

See also: Decision-Making in Financial Markets; Decision-Making and Neuroeconomics; Delayed Reinforcement: Neuroscience; Game Theory and the Economics of Animal Communication; Games in Monkeys: Neurophysiology and Motor Decision-Making; Neuroeconomics: History; Prediction Errors in Neural Processing: Imaging in Humans; Reward Decision-Making; Social Cognition.

Further Reading

Frederick S, Loewenstein G, and O'Donoghue T (2002) Time discounting and time preference: A critical review. *Journal of Economic Literature* 40: 351–401.

Harvey CM (1986) Value functions for infinite-period planning. *Management Science* 32: 1123–1139.

Herrnstein RJ (1981) Self-control as response strength. In: Bradshaw CM, Szabadi E, and Lowe CF (eds.) *Quantification of Steady-State Operant Behavior*, pp. 3–20. Amsterdam: Elsevier/North-Holland.

Koopmans TC (1960) Stationary ordinal utility and impatience. *Econometrica* 28: 287–309.

Koopmans TC, Diamond PA, and Williamson RE (1964) Stationary utility and time perspective. *Econometrica* 32: 82–100.

Kreps D (1990) *A Course in Microeconomic Theory.* New York: Harvester Wheatsheaf.

Laibson D (1997) Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112: 443–477.

Loewenstein G and Prelec D (1992) Anomalies in intertemporal choice: Evidence and an interpretation. *Quarterly Journal of Economics* 107: 573–597.

Mazur JE (1987) An adjustment procedure for studying delayed reinforcement. In: Commons ML, Mazur JE, Nevins JA, and Rachlin H (eds.) *Quantitative Analysis of Behavior: The Effect of Delay and of Intervening Events on Reinforcement Value*, vol. 5, pp. 55–73. Hillsdale NJ: Erlbaum.

O'Donoghue T and Rabin M (1999) Doing it now or later. *American Economic Review* 89: 103–124.

Rachlin H and Green L (1972) Commitment, choice and self-control. *Journal of the Experimental Analysis of Behavior* 17: 15–22.

Samuelson PA (1937) A note on measurement of utility. *Review of Economic Studies* 4: 155–161.

Starmer C (2000) Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38: 332–382.

Strotz RH (1956) Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23: 165–180.

Delayed Reinforcement: Neuroscience

C M Bradshaw, S Body, and E Szabadi, University of Nottingham, Nottingham, UK

© 2009 Elsevier Ltd. All rights reserved.

Theoretical Bases

A basic principle of human and animal behavior is that organisms tend to carry out voluntary actions which lead to rewarding outcomes. Couched in such simple terms, this principle conveys little information, because the terms 'voluntary action' and 'rewarding outcome' are defined with reference to one another, leading to a proposition that is essentially circular. The problem is not resolved merely by replacing the everyday notions of voluntary action and reward by technical terms such as 'operant behavior' and 'reinforcement.' What is needed is first a formal statement of the reinforcing relation which allows the key terms to be defined rigorously, and second a means of anchoring the statement to some underlying biological mechanism. The burden of fulfilling the first requirement has been shouldered variously by psychologists, behavioral ecologists, and economists, while the task of addressing the second requirement has been taken on by behavioral neuroscientists. While considerable advances have been made with each of these endeavors, a complete understanding of reward-motivated behavior awaits a successful integration of the two approaches. This article argues that one area where such an integration is nearly within reach is the analysis of the effects of delayed reinforcement.

It is generally agreed that the efficacy or 'value' of a reinforcer is an inverse function of its delay. Specifying the mathematical form of this function is an important step in the direction of defining the reinforcement process. Choice schedules have played a crucial part in establishing the nature of this function. Faced with a choice between two reinforcers of different sizes, animals will, other things being equal, prefer the larger one. Similarly, faced with a choice between two reinforcers associated with different delays, animals will, other things being equal, prefer the one that can be obtained sooner. In real life, of course, other things are seldom equal, and choice between rewarding outcomes inevitably involves consideration of the relative merits of the two rewards along more than one dimension. For example, in the situation that has been most extensively studied in the operant behavior laboratory, the study participant must choose between a smaller reinforcer that may be obtained after a brief delay and a larger reinforcer whose delivery is considerably postponed; in such an 'intertemporal choice' situation, the participant's choice behavior necessarily reflects the influences of both size and delay of reinforcement.

In a pioneering experiment by George Ainslie, pigeons repeatedly chose between a smaller reward, A, delivered after a delay d_A, and a larger reward, B, delivered after a longer delay, d_B. With d_A set at zero, the birds showed a strong preference for A. However, when both delays were progressively increased by the same amount, D, preference for A weakened, and eventually reversed, so that at high values of D the birds actually preferred B to A. Ainslie's demonstration of preference reversal delivered, in the eyes of many psychologists, the coup de grâce to models of choice based on classical microeconomic theory, in which the value of a commodity is presumed to decline exponentially as a function of delay. Exponential delay discounting is epitomized by eqn [1]

$$V = q\mathrm{e}^{-kd} \qquad [1]$$

where V is the hypothetical value of a reward, q is its size or quantity, d is delay, and k is an exponential decay parameter. In keeping with economic notions of 'rational choice,' exponential delay discounting functions decline in parallel (provided that k is held constant), with the result that if A is preferred to B when their respective delays are d_A and d_B, A will still be preferred to B when their delays are $d_A + D$ and $d_B + D$. However, as pointed out by Ainslie, preference reversal is entirely compatible with hyperbolic delay discounting, as shown in **Figure 1**.

An elegantly simple statement of hyperbolic delay discounting, championed by James Mazur, is given in eqn [2]

$$V = \frac{q}{1 + Kd} \qquad [2]$$

where K is a parameter expressing the rate of (hyperbolic) delay discounting.

In fact, exponential discounting (eqn [1]) can be made to accommodate preference reversal, provided that k is allowed to vary inversely with q, in other words, if the rate of discounting is allowed to be higher for small reinforcers than for large ones. Such may indeed be the case when humans make choices between hypothetical monetary reinforcers. However, there is good evidence that this so-called 'magnitude effect' does not apply when animals make choices between food reinforcers, suggesting that eqn [2] is to be preferred to eqn [1].

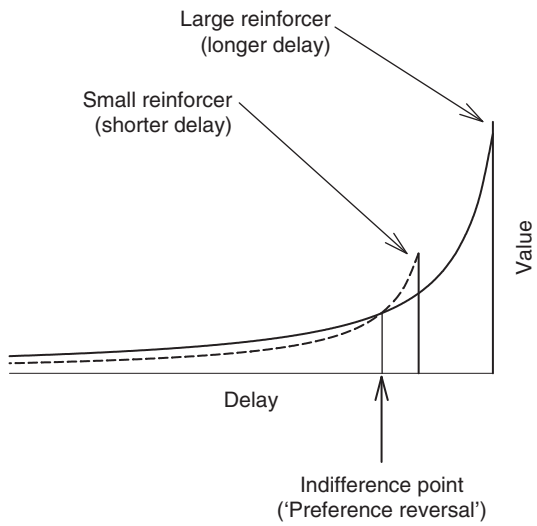

Figure 1 Ainslie's theoretical account of preference reversal, based on the principle of hyperbolic delay discounting. The vertical lines represent the instantaneous values of a large (continuous line) and a small (broken line) reinforcer. The curves show the decline in value that results from imposition of a delay between the response and the delivery of the reinforcer. At the indifference point ('preference reversal'), the values of the two reinforcers are equal. If the organism makes its choice when the delay to the small reinforcer is relatively short (i.e., at a delay value to the right of the indifference point), the smaller reinforcer is preferred. If the delays to both reinforcers are increased, so that the organism is required to make its choice at some time point to the left of the indifference point, the larger reinforcer will be preferred. Note that the intersecting value curves are a consequence of hyperbolic discounting (exponential discounting gives rise to curves that decline in a parallel fashion: see text).

Recent work in our laboratory suggests that eqn [2] somewhat oversimplifies the discounting function. The problem arises from the fact that eqn [2] implies that, if d is held constant, V is directly proportional to q, the physical size of the reinforcer. There are, however, good reasons to doubt this supposition. For instance, when rats make choices between reinforcers consisting of different numbers of food pellets, preference for the larger reinforcer is stronger when the two reinforcers consist of one and two pellets than when they consist of three and six pellets. This suggests a nonlinear relation between V and q. According to one model, the 'multiplicative hyperbolic model' (MHM), the relation is hyperbolic, as defined by eqn [3]

$$V = \frac{q}{q+Q}\,\frac{1}{1+Kd} \qquad [3]$$

where Q is a parameter which modulates the effect of reinforcer size upon value. It should be emphasized at this point that the role of Q is not merely to explain away embarrassing departures from the straightforward predictions of eqn [2]. As will be seen later in this article, neurobiological interventions can alter animals' choices between reinforcers by altering their sensitivity to delay and/or their sensitivity to size. Q provides a vehicle for formalizing and quantifying the effects of interventions on sensitivity to reinforcer size.

According to MHM, eqn [3] may be expanded in order to accommodate the effects of other variables on reinforcer value. For example, probability of reinforcement, p, can be accommodated thus:

$$V = \frac{q}{q+Q}\,\frac{1}{1+Kd}\,\frac{1}{1+H\theta} \qquad [4]$$

where θ is the odds against the occurrence of the reinforcer ($[1/p] - 1$) and H is a an odds-discounting parameter. Equation [4] allows risky choice, a matter of great interest to students of foraging behavior, to be considered within the framework of MHM.

It will not have escaped the reader's notice that V is not a directly measurable quantity. As an intervening variable, it is useful only to the extent that it can be linked to some explicit feature of behavior. Two ways of dealing with V can be found in the literature. The first approach is to link it to notions of 'response strength' and to infer changes in V from changes in the rate or vigor of responding. The second approach is to write equations which allow V to be canceled out; this is the indifference or null equation approach. The fundamental assumption underlying the derivation of indifference equations is that when the animal displays indifference between two reinforcers, A and B, their values are equal: $V_A = V_B$. This equality may be expanded using the definition of V specified by eqn [4], and then reduced to simple relations among the quantitative features of the two reinforcers. For example, in an experiment in which an animal makes choices between a small reinforcer, A, delivered after a short delay and a larger reinforcer, B, delivered after a longer delay, the delay to B that results in indifference between A and B ($d_{B(50)}$) is a linear function of the delay to A (d_A):

$$d_{B(50)} = \frac{1}{K}\left[\frac{q_B(Q+q_B) - q_A/(Q+q_A)}{q_A/(Q+q_A)}\right] + d_A\left[\frac{1+Q/q_A}{1+Q/q_B}\right] \qquad [5]$$

Equation [5] is one member of a family of null equations that may be derived to describe indifference between two reinforcers; it is particularly suited to experiments in which the sizes of the two reinforcers are held constant and delays are systematically varied (see **Figure 2** for an example). Other equations can be derived to describe situations in which size, rather than

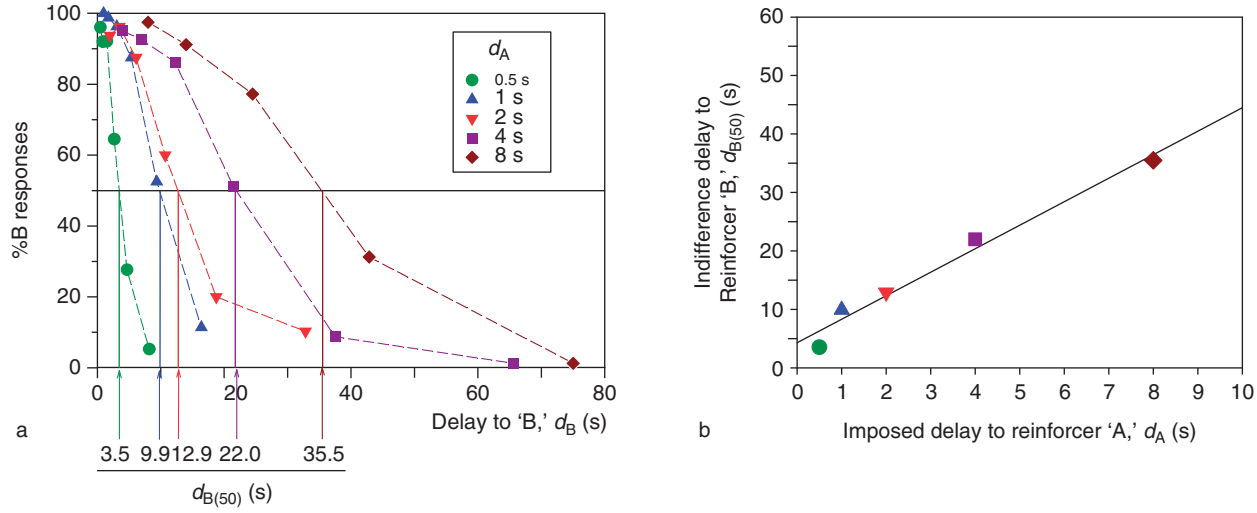

Figure 2 Derivation of a linear indifference function: example of data obtained from one rat. (a) Preference functions: percent responding on lever B (%B) is plotted against the delay to the larger of two reinforcers (d_B). Each plot shows the data obtained from one phase of the experiment, in which the delay to the smaller reinforcer (d_A) was fixed at the value shown in inset panel. The horizontal line signifies indifference (%B = 50), and the vertical lines indicate the indifference delays ($d_{B(50)}$) obtained in each phase (values shown under the abscissa). (b) Linear indifference function ($d_{B(50)}$ vs d_A) derived using the indifference delays shown in (a). The fitted line is the linear function specified by eqn [5]. Unpublished data from Kheramin S, Body S, Bradshaw CM, Szabadi E, Deakin JFW, and Anderson IM.

delay, is systematically manipulated, or in which reinforcer probability is the variable of interest.

As well as specifying the form of the indifference relation (in the case of eqn [5], a linear relation), null equations also offer a means of characterizing the effects of interventions on the discounting parameters. A more ambitious goal, which, unfortunately, is often thwarted by the intrinsic variability of the behavioral data, is to derive quantitative estimates of the discounting parameters pertaining to individual animals. For example, the slope of the linear function specified by eqn [5] is determined entirely by the sizes of the two reinforcers and the size-discounting parameter Q. The delay-discounting parameter K may be obtained using the formula $K = (\text{slope} - 1)/\text{intercept}$.

Indifference equations highlight a feature of choice behavior which has important implications for neurobiological analyses of reinforcement processes: choice usually entails a compromise. In the situation characterized by eqn [5], the compromise is one between size and delay. Indifference implies that a balance has been struck between the competing attractions of the larger size of reinforcer B and the shorter delay associated with reinforcer A. This balance can be upset, and hence the indifference point displaced, by altering the animal's sensitivity to either size or delay of reinforcement, or both. This is illustrated in **Figure 3**, which shows the effects of hypothetical interventions on the linear indifference function defined by eqn [5]. The left-hand graph shows the effect of an intervention that alters sensitivity to delay (via an increase in K), the middle graph shows the effect of an intervention that alters sensitivity to size (via a reduction of Q), and the right-hand graph shows the effect of an intervention that alters sensitivity to both of these features of reinforcers (via increases in both K and Q).

If the foregoing analysis is correct, it follows that the reluctance of an individual animal to select the larger and more delayed of two reinforcers does not necessarily indicate that it is especially sensitive to delay of reinforcement; it is equally possible that it differs from its peers only in its sensitivity to the relative sizes of the two reinforcers. Therefore, interpretation of induced shifts of preference favoring smaller, less delayed reinforcers in terms of 'accelerated delay discounting' or 'delay aversion' should be made with due caution. Moreover, the time-honored anthropomorphisms 'impulsive' and 'self-controlled' choice have little to recommend them. While they have served a useful purpose in highlighting the distinction between different uses of the term 'impulsiveness' in clinical and lay discourse, in particular the distinction between impulsive choice (selection of smaller earlier reinforcers) and impulsive action (acting rapidly in situations where delay is advantageous), they can be seriously misleading if they are identified with hypothetical behavioral processes. As we have seen, models such as MHM imply that altered sensitivity to delay of reinforcement is neither a necessary nor a sufficient condition for a shift in preference in favor of smaller earlier reinforcers.

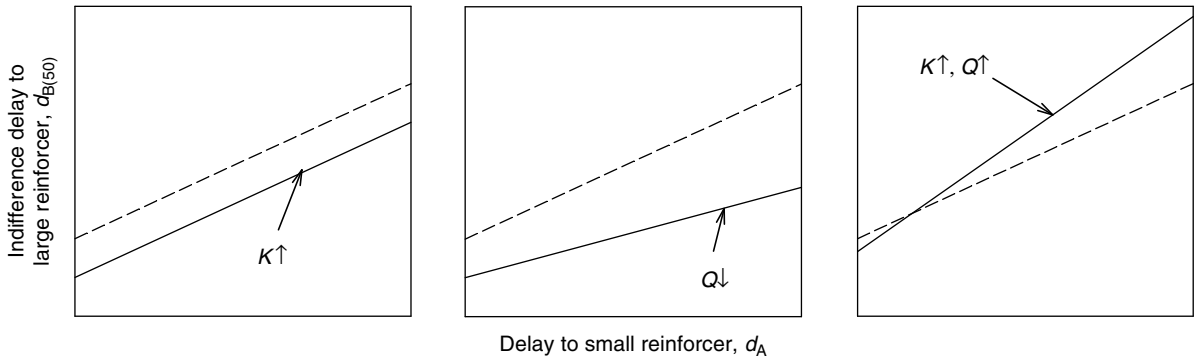

Figure 3 Effects of three hypothetical interventions on the linear indifference function defined by MHM (eqn [5]). Ordinates: indifference delay to the larger of two reinforcers, $d_{B(50)}$; abscissas: delay to the smaller reinforcer, d_A. In each graph, the broken line represents the 'baseline' function and the continuous line represents the function derived in the presence of an intervention that alters the delay-discounting parameter, K (left-hand graph), the size-discounting parameter, Q (middle graph), or both parameters (right-hand graph). Note that preference for the smaller, earlier reinforcer (A) can be promoted by either an increase in K or a reduction of Q, but the slope of the function is altered only in the latter case. If K and Q are both increased, the baseline and altered functions may intersect, resulting in stronger preference for A at low values of d_A, and weaker preference for A at high values of d_A, in the presence of the hypothetical intervention.

Methodology

Numerous methods have been devised to investigate intertemporal choice behavior. A full description of these methods is beyond the scope of this article, and the interested reader is referred to the articles listed in the 'Further reading' section. Some of the most commonly used approaches are outlined in this section.

Free-Operant Concurrent Schedules

In these schedules, the study participant has continuous access to two operanda that are associated with different outcomes. One type of concurrent schedule that lends itself well to studies of delayed or probabilistic reinforcement is the concurrent-chains schedule, in which equal initial-link schedules (e.g., a pair of concurrent variable-interval schedules) provide access to mutually exclusive terminal-link schedules (e.g., response-independent reinforcer delivery following different delays). Relative response rate in the initial links provides the measure of preference, and equality of response rates is taken to indicate equality of the reinforcing values of the terminal links. Of course, precise indifference may not be obtained in any one pair of schedules; however, the indifference point may be estimated from a series of concurrent schedules by application of the 'generalized matching law.'

The 'time-left' procedure is a concurrent chain schedule which yields continuous, graded measure of preference. Equal initial-link schedules provide access, at unpredictable times after the start of a trial, to two mutually exclusive terminal links which deliver reinforcement either after a fixed delay or, after a variable delay, at the end of the trial. Early in the trial, the fixed delay is shorter than the variable delay, whereas toward the end of the trial, the variable delay is shorter. Preference for the variable delay accordingly increases from approximately zero at the start of the trial, to near 100% at the end of the trial. The locus of the indifference point (50% choice of the variable delay) depends on the length of the fixed delay and the sizes of the reinforcers offered in the two terminal links. It has been shown that the indifference delay is a linear function of the length of the fixed delay, in accord with eqn [5].

Discrete-Trials Concurrent Schedules

These schedules comprise a series of trials in each of which the participant chooses between two mutually exclusive alternatives. Preference is measured as proportional choice of one alternative across a number of trials. Discrete-trials concurrent schedules tend to engender exclusive preference; this is the expected outcome of repeated exposure to choice between mutually exclusive alternatives, A and B, since alternative A will always be chosen whenever $V_A > V_B$, and vice versa (see earlier). However, indifference points may be determined by systematic manipulation of a relevant independent variable (e.g., delay to one reinforcer), and estimation of the value of that variable corresponding to 50% choice of each alternative.

Conventional T-maze tasks are a type of discrete-trials concurrent schedule; they can be adapted to the study of intertemporal choice by delaying reinforcer delivery in one of the goal boxes by some specified time. Most studies employing T-maze tasks have used proportional choice of the delayed reinforcer as a simple measure of preference, a measure which can

only yield an indifference point if preference is tested using a range of delays.

J. L. Evenden and C. N. Ryan introduced a protocol for examining the effect of delay in a discrete-trials lever-pressing task. In each free-choice trial, lever A delivers a small reinforcer immediately, whereas lever B delivers a larger reinforcer after a delay, d_B; d_B is progressively increased in successive blocks of trials. The value of d_B at which preference switches from B to A may be taken as the indifference delay, $d_{B(50)}$. In our laboratory, the method has been extended, using a range of delays to the smaller reinforcer (d_A), to enable linear indifference functions to be obtained between $d_{B(50)}$ and d_A, as prescribed by MHM.

An important technical development in the estimation of indifference points was Mazur's introduction of the adjusting delay schedule. In this schedule, the participant undergoes a series of trials in which the sizes of the two reinforcers and the delay to the smaller reinforcer are fixed, while the delay to the larger reinforcer is allowed to vary in accordance with the participant's choices. When the participant shows a preference for the larger reinforcer, the delay to that reinforcer is increased; when preference shifts to the smaller reinforcer, the delay to the larger reinforcer is reduced. Extended training on this schedule results in the establishment of a quasi-stable delay to the larger reinforcer, which is taken as the indifference delay.

J. B. Richards and colleagues have devised a variant of Mazur's method, the adjusting-amount schedule, in which rats make repeated choices between a delayed liquid reinforcer of fixed volume and an immediate reinforcer, the volume of which is adjusted in accordance with the animals' choices. The volume of the variable reinforcer at which indifference is achieved is often taken as an index of the 'value' of the fixed-volume delayed reinforcer. However, it should be noted that the logic of this method is based on the assumption that value is linearly related to reinforcer size, in accordance with eqn [2], an assumption that is contradicted by models such as MHM, which assume that value is constrained by an asymptotic function such as that defined by eqn [3]. An important practical advantage of the adjusting-amount schedule is the rapid (within-session) attainment of stable indifference that has been reported in many studies.

Effects of Neurobiological Interventions on Choice of Delayed Reinforcers

The 5-Hydroxytryptaminergic Pathways

Twenty years ago, Phillipe Soubrié proposed that the ascending 5-hydroxytryptaminergic (5-HTergic, or serotonergic) pathways may play an important role in enabling organisms to tolerate delay of reinforcement, and that dysfunction of these pathways might therefore promote selection of small immediate reinforcers in preference to larger delayed reinforcers. Several studies have since investigated the effects of lesions of the 5-HT pathways, and acute treatment with drugs whose mode of action is believed to involve these pathways, on choice between small immediate and large delayed rewards. Unfortunately, these experiments have produced varied, and sometimes conflicting, results. The results of experiments investigating the effects of central 5-HT depletion are summarized in Table 1.

Several early experiments employed intraraphe injections of the selective neurotoxin 5,7-dihydroxytryptamine (5,7-DHT) to destroy the ascending 5-HTergic pathways. These studies generally indicated that loss of central 5-HT was associated with a shift of preference toward smaller, less delayed reinforcers.

Table 1 Effects of central 5-HT depletion on intertemporal choice behavior

5-HT depletion method	*Behavioral method*	*Number of preference functions/ indifference points*	*Effect on preference for smaller, earlier reinforcer*	*Reference*
Intra-raphe 5,7-DHT	Adjusting delay	1	↑	Wogar et al. (1993)
Intra-raphe 5,7-DHT	Time-left procedure	1	↑	Al-Ruwaitea et al. (1999)
Intra-raphe 5,7-DHT	Adjusting delay	5	↑	Mobini et al. (2000a)
Intra-raphe 5,7-DHT	Discrete-trials choice	1	↑	Mobini et al. (2000b)
i.c.v.5,7-DHT	Discrete-trials choice	1	0	Winstanley et al. (2003)
i.c.v.5,7-DHT	Discrete-trials choice	1	0	Winstanley et al. (2004)
i.c.v. 5,7-DHT	Discrete-trials choice	1	0	Denk et al. (2005)
*p*CPA	T-maze choice	NA	↑	Bizot et al. (1999)
i.c.v. 5,7-DHT	Discrete-trials choice	1	0	Winstanley et al. (2005)

5,7-DHT, 5,7-dihydroxytryptamine; i.c.v., Intracerebroventricular injection; *p*cpa, *p*-chlorophenylalanine.

More recent experiments, generally employing intracerebroventricular injections of the same neurotoxin, have not replicated this result. The reason for this discrepancy is unclear; obvious candidates are different techniques for inducing 5-HT depletion and different behavioral paradigms used to measure choice behavior. A weakness of most of these studies is the use of a single preference function or indifference point, which, as discussed previously, can cloud the interpretation of any changes observed. However, this is unlikely to be the sole reason for the discrepant results, because the only study to have used a multiple-indifference-point approach found that 5-HT-depleted rats showed a parallel displacement of the linear indifference function defined by eqn [5], consistent with a selective increase of the delay-discounting parameter K (see **Figure 3**, left-hand graph). This pattern of effect implies that the shift of preference favoring smaller earlier reinforcers should occur at all values of delay, and therefore the outcome of the experiments should not have been affected by the use of different delays in different experiments. The possible role of the 5-HTergic pathways in risky choice is a largely unexplored area. Two recent studies found no evidence that 5-HT depletion promotes selection of large low-probability rewards (a putative model of gambling); however, this is clearly a topic that requires further research.

There have been some attempts to explore the effects of drugs that interact with 5-HTergic mechanisms on intertemporal choice. Earlier studies, which mainly used T-maze-based choice paradigms, reported that 5-HT uptake-inhibiting drugs generally promoted choice of the larger, more delayed of two reinforcers, in keeping with Soubrié's proposal that endogenous 5-HT contributes a restraining influence on the selection of small immediate gains. More recent experiments, using discrete-trials operant procedures, have indicated that 5-HT_{1A} receptor stimulation tends to promote the selection of smaller earlier reinforcers, possibly via an interaction between 5-HTergic and dopaminergic mechanisms in the ventral striatum. Unfortunately, all these studies have examined the effects of drugs on a single preference function. Confirmation that the reported effects reflect changes in the rate of delay discounting must await the results of parametric studies that allow clear separation of delay discounting from sensitivity to reinforcer size.

The Dopaminergic Pathways

The possibility that the central dopaminergic pathways may contribute to the control of behavior by delayed reinforcement has attracted considerable interest because of the putative involvement of dopaminergic dysfunction in attention-deficit/hyperactivity disorder (ADHD), a condition in which 'impulsiveness' is a prominent feature. Unfortunately, a clear picture of dopamine's role in intertemporal choice has yet to emerge.

Selective dopamine depletion from the nucleus accumbens and local injection of dopamine receptor antagonists into this region have generally been found to have little effect on preference for the smaller and earlier of two reinforcers. This is in contrast to the effect of excitotoxic lesions of this structure, which have been found to promote the selection of small immediate reinforcers (see later). In one study, injection of the selective catecholamine-depleting neurotoxin 6-hydroxydopamine into the orbital prefrontal cortex (OPFC) had a complex effect on intertemporal choice, promoting the selection of the larger, more delayed reinforcer to an increasing degree as the delay to the smaller reinforcer was progressively increased (see **Figure 3**, right-hand graph). This finding can be interpreted on the basis of MHM (eqn [5]) in terms of a dual effect on delay discounting (K was increased) and sensitivity to relative reinforcer size (Q was increased).

A number of studies have investigated the effects of systemic treatment with psychostimulant drugs, which facilitate dopaminergic function. Earlier studies yielded disconcertingly disparate results; however, it now seems that these apparently inconsistent findings may be reconcilable. As shown in **Table 2**, there are reports that the catecholamine-releasing agent amphetamine can promote the selection of either the smaller immediate reinforcer or the larger delayed reinforcer in intertemporal choice schedules. A methodological factor whose importance has come to light in more recent experiments is the presence or absence of a cue during the delay interval. There are theoretical and empirical reasons for believing that an intradelay cue functions as a conditioned reinforcer whose reinforcing efficacy is inversely related to the duration for which it is presented. Consistent with this notion, it has been found that rats generally tolerate much longer delays to reinforcement when a cue is presented during the delay than when no cue is provided. Since amphetamine and related drugs are believed to enhance the effect of conditioned reinforcers, it might be expected that these compounds would promote the selection of the more delayed of two reinforcers when the delay is signaled by an explicit cue, and this indeed turns out to be the case. In the absence of such a cue, the predominant effect of amphetamine is to promote selection of the smaller earlier reinforcer. Unfortunately there are no published studies that have

Table 2 Effects of amphetamine and related drugs on intertemporal choice behavior

Drug	*Behavioral method*	*Number of preference functions/ indifference points*	*Effect on preference for smaller, earlier reinforcer*	*Reference*
d-Amphetamine 0.5 mg kg^{-1}	Discrete-trials choice	NA	↑	Charrier and Thiébot (1996)
d-Amphetamine 0.1 mg kg^{-1}	Discrete-trials choice	1	↑	Evendon and Ryan (1996)
Methamphetamine 0.5, 1, 2 mg kg^{-1}	Adjusting amount	1	↓	Richards et al. (1999)
d-Amphetamine 0.5, 1 mg kg^{-1}	Adjusting amount	1	↓	Wade et al. (2000)
d-Amphetamine 0.3, 1, 1.6 mg kg^{-1}	Discrete-trials choice	1	↑↓[a]	Cardinal et al. (2000)
d-Amphetamine 0.3, 1, 1.5, 2.3 mg kg^{-1}	Discrete-trials choice	1	↑	Winstanley et al. (2003)
d-Amphetamine 0.4, 0.6, 0.8 mg kg^{-1}	Discrete-trials choice	1	↑↓[b]	Isles et al. (2003)
Methamphetamine 0.3, 1, 1.7, 3 mg kg^{-1}	Concurrent-chains schedule	1	↓	Pitts and Febbo (2004)

[a]↑ when intradelay cue was not used; ↓ when intradelay cue was used.
[b]Dose-dependent effect: lower doses promoted choice of the larger reinforcer, higher dose promoted choice of the smaller reinforcer.

incorporated a range of indifference points, and therefore it remains unknown whether the effects of amphetamine on intertemporal choice reflect a selective influence on delay discounting.

The Nucleus Accumbens Core

Excitotoxic bilateral destruction of the nucleus accumbens core (NAC) has been found to displace preference functions in favor of smaller immediate reinforcers. However, it has yet to be established, on the basis of parametric studies, whether this is due to facilitation of delay discounting.

The Prefrontal Cortex

Excitotoxic lesions of the ventral–medial and dorsal–medial (anterior cingulate) prefrontal cortex have been found to have no effect on intertemporal choice behavior. However, destruction of the OPFC produced a complex pattern of effect (Figure 4); preference was minimally affected when short delays to the smaller reinforcer (d_A) were employed, whereas the lesion promoted preference for the larger reinforcer to an increasing degree as d_A was progressively increased. Analysis of the linear indifference functions using MHM (eqn [5]) indicated that the OPFC-lesioned subjects had a higher rate of delay discounting (K was increased) and a greater sensitivity to relative reinforcer size (Q was increased) compared to sham-lesioned control subjects.

Future Directions

There is, as yet, no unanimous agreement on the most appropriate model of delay discounting; however, there is a good level of agreement about the range of phenomena that a successful model should accommodate – preference reversal and indifference functions are prominent among these phenomena. More controversial is the issue of how the influences of different features of reinforcers, for example, delay and magnitude, may be disentangled. The model advocated in this article (MHM) is an attempt to address this problem; while we make no claim that MHM is unique in this respect, we are convinced that formal models, of which MHM is one example, can provide a framework for quantifying the effects of various features of reinforcers, which may be of considerable help in the quest to identify the neurobiological substrate of choice behavior.

Emerging evidence for the involvement of the NAC and OPFC in intertemporal choice, discussed earlier, is of particular interest in the light of recent neurophysiological evidence that these structures are active during prereinforcer delays. It has been suggested that neural activity in these areas may help the organism to bridge the gap between an operant response and a rewarding outcome, and may therefore play an important role in 'anticipation' of rewarding events. An important question that should be addressed in future research is whether the putative involvement of these structures in intertemporal choice reflects their

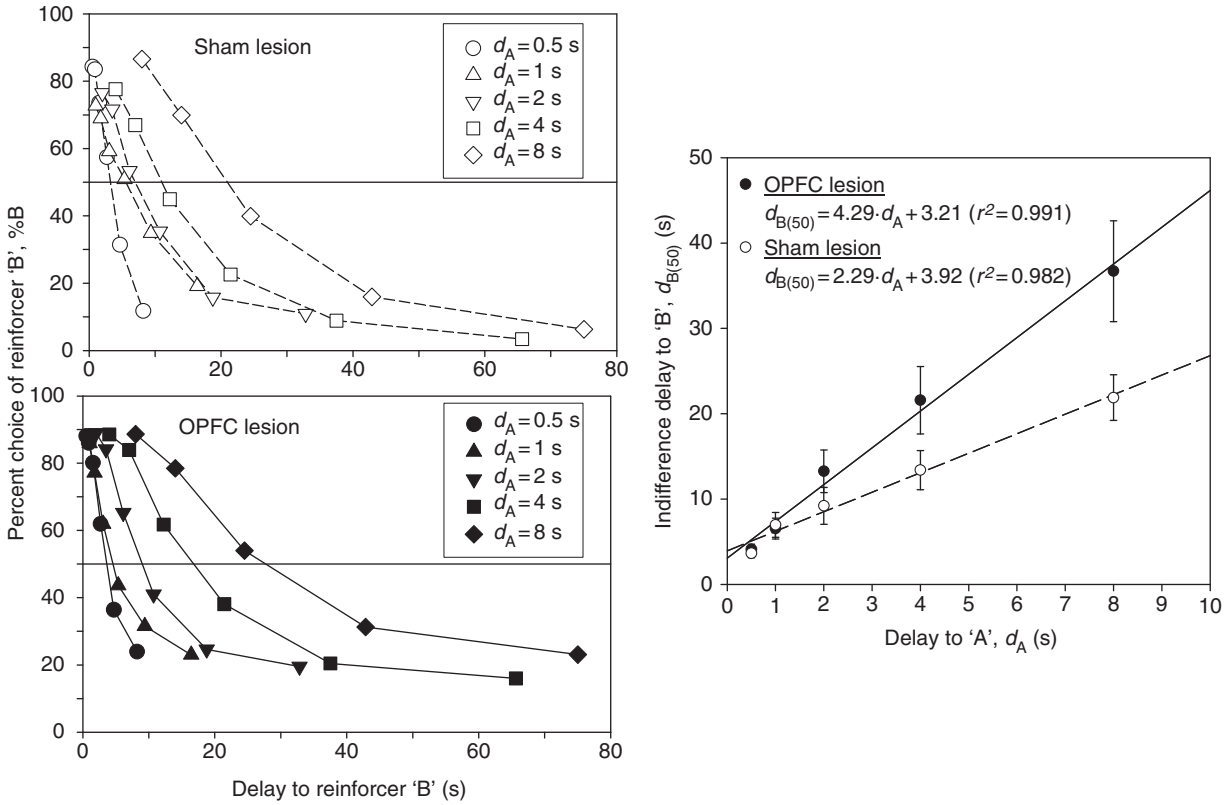

Figure 4 The effect of excitotoxin-induced lesions of the OPFC on intertemporal choice. (Left-hand graphs) Preference functions: percent responding on lever B (%B) is plotted against the delay to the larger of two reinforcers (d_B). Each plot shows the data obtained from one phase of the experiment, in which the delay to the smaller reinforcer (d_A) was fixed at the value shown in the inset panel. The horizontal line signifies indifference (%B = 50). (Right-hand graph) Linear indifference functions derived for the control group of sham-lesioned rats (open circles) and the OPFC-lesioned rats (filled circles). Ordinate: indifference delay to the larger reinforcer ($d_{B(50)}$); abscissa: imposed delay to the smaller reinforcer (d_A). Points are group mean data (± SEM); lines are best-fit linear functions. Data from Kheramin S, Body S, Mobini S, et al. (2002) *Psycyhopharmacology* 165: 9–17.

participation in integrated cortico-striato-thalamo-cortical circuits.

See also: Amphetamines; Delayed Reinforcement: Economics; Psychopharmacology of Reward and Appetite in Rats; Reinforcement Models; Reward and Learning; Reward Decision-Making; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Al-Ruwaitea ASA, Chiang T-J, Al-Zahrani SSA, Ho M-Y, Bradshaw CM, and Szabadi E (1999) Effect of central 5-hydroxytryptamine depletion on tolerance of delay of reinforcement: Evidence from performance in a discrete-trials 'time-left' procedure. *Psychopharmacology* 141: 22–29.

Bizot J-C, Le Bihan C, Puech AJ, Hamon M, and Thiébot MH (1999) Serotonin and tolerance to delay of reward in rats. *Psychopharmacology* 146: 400–412.

Cardinal RN, Robbins TW, and Everitt BJ (2000) The effects of d-amphetamine, chlordiazepoxide, alpha- flupenthixol and behavioural manipulations on choice of signalled and unsignalled delayed reinforcement in rats. *Psychopharmacology* 152: 362–375.

Cardinal RN, Robbins TW, and Everitt BJ (2004) Choosing delayed rewards: Perspectives from learning theory, neurochemistry, and neuroanatomy. In: Vuchinich RE and Heather N (eds.) *Choice Behavioural Economics and Addiction*, pp. 181–213. Oxford: Elsevier.

Charrier D and Thiébot MH (1996) Effect of phychotrophic drugs on rat responding in an operant paradigm involving choice between delayed reinforcers. *Pharmacology, Biochemistry, and Behavior* 54: 149–157.

Denk F, Walton ME, Jennings KA, Sharp T, Rushworth MFS, and Bannerman DM (2005) Differential involvement of serotonin and dopamine systems in cost–benefit decisions about delay or effort. *Psychopharmacology* 179: 587–596.

Evenden JL and Ryan CN (1996) The pharmacology of impulsive behaviour in rats: The effects of drugs on response choice with varying delays of reinforcements. *Psychopharmacology* 128: 161–170.

Ho M-Y, Mobini S, Chiang T-J, et al. (1999) Theory and method in the quantitative analysis of 'impulsive choice' behaviour: Implications for psychopharmacology. *Psychopharmacology* 146: 362–372.

Isles AR, Humby T, and Wilkinson LS (2003) Measuring impulsivity in mice using a novel operant delayed reinforcement task:

Effects of behavioural manipulations and d-amphetamine. *Psychopharmacology* 170: 376–382.

Kacelnik A (2003) The evolution of patience. In: Loewenstein G, Read D, and Baumeister R (eds.) *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, pp. 115–138. New York: Russel Sage.

Kheramin S, Body S, Mobini S, et al. (2002) Effects of quinolinic acid-induced lesions of the orbital prefrontal cortex on inter-temporal choice: A quantitative analysis. *Psychopharmacology* 165: 9–17.

Mazur JE (1997) Choice, delay, probability, and conditioned reinforcement. *Animal Learning & Behavior* 25: 131–147.

Mazur JE (2001) Hyperbolic value addition and general models of animal choice. *Psychological Review* 108: 96–112.

Mazur JE (2006) Mathematical models and the experimental analysis of behavior. *Journal of the Experimental Analysis of Behavior* 85: 285–291.

Mobini S, Chiang T-J, Al-Ruwaitea ASA, Ho M-Y, Bradshaw CM, and Szabadi E (2000a) Effect of central 5-hydroxytryptamine depletion on inter-temporal choice: A quantitative analysis. *Psychopharmacology* 149: 313–318.

Mobini S, Chiang T-J, Ho M-Y, Bradshaw CM, and Szabadi E (2000b) Effects of central 5-hydroxytryptamine depletion on sensitivity to delayed and probabilistic reinforcement. *Psychopharmacology* 152: 390–397.

Pitts RC and Febbo SM (2004) Quantitative analyses of methamphetamine's effects on self-control choices: Implications for elucidating behavioral mechanisms of drug action. *Behavioural Processes* 66: 213–233.

Richards JB, Sabol KE, and de Wit H (1999) Effects of methamphetamine on the adjusting amount procedure, a model of impulsive behavior in rats. *Psychopharmacology* 146: 432–439.

Sagvolden T, Johansen EB, Aase H, et al. (2005) A dynamic developmental theory of attention-deficit/hyperactivity disorder (ADHD) predominantly hyperactive/ impulsive and combined subtypes. *Behavioral and Brain Sciences* 28: 397–468.

Schultz W (2004) Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Current Opinion in Neurobiology* 14: 139–147.

Soubrié P (1986) Reconciling the role of central serotonin neurons in human and animal behavior. *Behavioral and Brain Sciences* 9: 319–364.

Sozou PD (1998) On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society, London, Series B* 265: 2015–2020.

Wade TR, deWit H, and Richards JB (2000) Effects of dopaminergic drugs on delayed reward as a measure of impulsive behavior in rats. *Psychopharmacology* 150: 90–101.

Winstanley CA, Dalley JW, Theobald DEH, and Robbins TW (2003) Global 5-HT depletion attenuates the ability of amphetamine to decrease impulsive choice in rats. *Psychopharmacology* 170: 320–331.

Winstanley CA, Theobald DEH, Cardinal RN, et al. (2004) Contrasting roles for basolateral amygdala and orbitofrontal cortex in impulsive choice. *Journal of Neuroscience* 24: 4718–4722.

Winstanley CA, Baunez C, Theobald DEH, et al. (2005) Lesions to the subthalamic nucleus decrease impulsive choice but impair autoshaping in rats: The importance of the basal ganglia in Pavlovian conditioning and impulse control. *European Journal of Neuroscience* 21: 3107–3116.

Wogar MA, Bradshaw CM, and Szabadi E (1993) Effect of lesions of the ascending 5-hydroxytryptaminergic pathways on choice between delayed reinforcers. *Psychopharmacology* 111: 239–243.

Dementia

G K Gouras, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Dementia is derived from the Latin words for 'away/from' (*de*) and 'mind' (*mens*), and is defined as an abnormal decline in cognitive function, with relative sparing of alertness and attention, that results in a significant change from the normal premorbid functioning of an individual.

Dementia generally is defined by decline in two or more domains of cognitive function, usually including memory. Neuropsychological testing can be used to test different cognitive domains, and can also be used to compare an individual to normal controls of the same age and education level. Mild cognitive impairment (MCI) has emerged as a term that describes an intermediate stage between 'normal' aging-related cognitive decline, also called benign forgetfulness of aging, and dementia. MCI indicates that cognition has declined from previous levels, or is abnormal compared to age- and education-matched controls, but that deficits are not severe enough to impair social or occupational functioning. It is when deficits impair daily functioning that the severity criterion for dementia is reached. It should be noted that this definition can be remarkably imprecise; given the same performance on neuropsychological testing, an individual in an intellectually demanding profession could meet the criterion for dementia, whereas another individual with few daily intellectual demands might not. Thus, the clinical distinction between MCI and dementia may not be sharp. The pathological distinctions between MCI and dementia also appear imprecise, as studies suggest that the pathology underlying MCI is often that of early Alzheimer's disease (AD), the most common cause of dementia.

An important point in educating patients is that dementia is not a specific diagnosis, but rather a symptom that can result from many different disorders. The most common causes of dementia include AD and other neurodegenerative diseases, vascular dementia, and the (pseudo) dementia of depression (Table 1).

Decline from previous functioning distinguishes dementia from mental retardation. However, individuals with mental retardation can also develop a superimposed dementia, as generally occurs in individuals with Down syndrome, who develop AD-like neuropathology and dementia as they age. Another key distinction is the difference between dementia and delirium; the latter state is also referred to as encephalopathy or acute confusional state. In delirium, the predominant abnormality is a disturbance of alertness or attention. It is important to note that delirium is not a specific diagnosis, but, like dementia, is a symptom that can result from many different causes. However, whereas the usual causes of dementia are neurodegenerative diseases, the typical causes of delirium are toxic–metabolic derangements. Thus, delirium is often characterized as acute and reversible, and dementia as chronically progressive and irreversible, though these are really characteristics of the typical diseases causing delirium and dementia, and not inherently of delirium and dementia. It is also important to note that dementia and delirium frequently co-occur. Some diseases, such as dementia with Lewy bodies, may cause both delirium and dementia. In addition, patients with dementing illnesses are more prone to develop delirium in response to urinary tract infections, medication adjustments, anesthesia, or even simple changes in environment. 'Sundowning' refers to the disorientation and agitation that can develop as night falls in cognitively impaired elderly persons removed from the familiar and comforting surroundings of their own homes. In fact, delirium often provides the first indication of an incipient dementia that had previously not been recognized.

Dementia and delirium should be viewed as symptoms, each term designed to trigger a search for specific causative diseases. As with the rest of neurology, the key features in the diagnosis of dementing diseases are the time course and pattern of anatomical involvement. A subacute, rapidly progressive course suggests Creutzfeldt–Jakob disease (CJD) or paraneoplastic limbic encephalitis, whereas insidious onset and slow progression suggest chronic neurodegenerative diseases. Within the neurodegenerative diseases, the key differentiating feature is the pattern of anatomical involvement, particularly early in the course of the illness. Prominent early involvement of memory (hippocampus) and naming or visuospatial function (temporoparietal polymodal association cortices) is strongly suggestive of AD. In contrast, when nonfluent aphasia or bizarre inappropriate behavior is prominent early in the course, the diagnosis is likely to be frontotemporal degeneration (FTD). In addition to the history of the present illness and the neurobehavioral examination, use of techniques such as brain imaging, including magnetic resonance imaging (MRI), single-photon emission tomography (SPECT), and positron emission tomography (PET) may assist in defining the pattern of anatomical involvement. A detailed workup also includes obtaining a family history, social history

Table 1 Major categories of dementia

Neurodegenerative diseases
Alzheimer's disease
Parkinson's disease and diffuse Lewy body disease
Frontotemporal dementia (Pick's disease; corticobasal degeneration; progressive supranuclear palsy)
Trinucleotide repeat neurodegenerative diseases
(Huntington's disease; spinocerebellar ataxias, multiple system atrophies)
Creutzfeldt–Jakob disease (prion diseases)
Storage diseases (i.e., leukodystrophies)
Dementia pugilistica
Vascular dementias
Psychiatric diseases (dementia of depression, schizophrenia, among others)
Toxic–metabolic encephalopathies (toxins, infectious, endocrine, vitamin deficiency, nutritional, systemic illness, inborn errors of metabolism)
Structural brain lesions and other neurological diseases
Brain trauma
Chronic subdural hematoma
Brain tumors (i.e., frontal meningioma)
Paraneoplastic limbic encephalitis
Normal pressure hydrocephalus
Obstructive hydrocephalus
Granulomatous diseases (i.e., sarcoid)
Multiple sclerosis
Epilepsy

(including alcohol intake, history of smoking, and possible exposures to toxins), list of medications, and prior and current medical conditions.

The pathophysiology of specific diseases is discussed within each of the following sections. The key observation, in overview, is that aging is by far the most important risk factor for the neurodegenerative diseases causing dementia, although the underlying biological mechanisms are only incompletely understood. Genetic predispositions and cumulative damage from oxidative stress are among the leading hypothesized mechanisms of age-related cognitive decline. Reduced function of normal cellular degradative systems, such as the endosomal–lysosomal and ubiquitin–proteasome systems, also appears to be important. Indeed, neurodegenerative diseases of aging are increasingly linked with the aberrant accumulation of cellular deposits. Altered age-related cardiovascular function also appears to contribute significantly to the age-related decline in cerebral function.

Dementia typically implies a progressive decline in function over months to years, although dementia can be precipitated acutely, such as from a stroke or head trauma. Dementia can present as a more focal alteration of cognitive function other than memory, as occurs with subcategories of FTD, such as progressive aphasia. AD generally presents with more localized symptoms attributable to short-term memory loss (amnesia), which in its earliest stages can be defined as amnestic MCI. Word-finding difficulties and orientation difficulties in relation to previously familiar surroundings are other common early features of AD. AD is often described as a temporal–parietal dysfunction, since the medial temporal lobes (memory consolidation, language) and parietal lobes (visuospatial function) are involved relatively early in the disease process, compared to FTD, in which frontal (executive function) and lateral temporal (language) lobes are preferentially involved early on. Evidence for selective, compared to more general, cerebral dysfunction can be obtained from the medical history of a patient and from neurobehavioral testing and brain imaging.

The care of demented patients scales in accord with functional ability, often categorized as mild, moderate, or severe. In mild dementia, self-care activities of daily living, such as eating, dressing, and personal hygiene, are not impaired. In moderate dementia, these activities are impaired, though they can still be accomplished with verbal cuing or other simple assistance. In severe dementia, the caregiver provides almost complete support. In such cases, meticulous nursing care is an important determinant of survival. The cause of death in severe dementia is typically from infections, due especially to aspiration, urosepsis, or decubitus ulcers.

The broad differential diagnosis that needs to be considered in the clinical workup of dementia stems from the multiple potential causes for reduced cognitive function in an individual. Normal cognition depends on the maintained connectivity and functioning of multiple interconnected neural circuits. Structural and chemical alterations can impair normal cognitive function. Alterations in the chemical milieu of neurons, from altered levels of electrolytes, oxygen, pH, or toxins, among others, can affect normal neuronal physiology and synaptic functioning. Side effects of medications often can affect cognition, and when they occur for months can be a cause of dementia, especially in the elderly.

Alzheimer's Disease

AD is the leading cause of dementia, estimated to occur in 50–75% of those afflicted by dementia. AD is characterized by an incipient and gradually progressive decline in cognitive function, with deficits especially in memory retrieval, semantic fluency, and visuospatial function. Loss of insight into cognitive deficits is a characteristic of AD. Although deficits in attention are uncommon and deficits in short-term memory are characteristic, impairment in other domains of cognition can be quite heterogeneous in AD. The diagnosis of AD is definitively confirmed only at

autopsy, when the typical dementia is found to be accompanied by sufficient numbers of cerebral amyloid plaques and neurofibrillary tangles. Loss of synapses is considered the best pathological correlate of cognitive decline in AD. Progressive cerebral atrophy and loss of neurons also characterize the disease. It is estimated that clinical diagnosis is accurate up to 95% of the time at specialized memory disorder clinics. Remarkably, age-related deposition of plaques and tangles is seen also in cognitively normal, aged individuals, and considerable neuropathology can be seen at the time of amnestic MCI. Functional imaging in genetically at-risk individuals indicates premorbid reductions in cortical activity that can precede a diagnosis of dementia by years to decades, suggesting that the prodrome of AD begins prior to overt symptoms.

Accumulation of β-amyloid peptides (Aβ), the principal component of plaques, is increasingly viewed as playing an important role in the pathogenesis of the disease (Figure 1). Several lines of evidence support that Aβ accumulation precedes and can induce tangle pathology. Genetic mutations in familial forms of early-onset AD (*PSEN1*, *PSEN2*, and *APP*, encoding presenilin 1, presenilin 2, and β-amyloid precursor protein; see Table 2) and genetic risk factors for late-onset AD (*apolipoprotein E* ε4 allele) have been shown to increase levels of Aβ in the brain. Aβ is a short, hydrophobic peptide that is cleaved from the larger type I transmembrane precursor protein, β-amyloid precursor protein (APP). Aβ peptides aggregate easily, especially the disease-linked Aβ42 form. The molecular mechanism whereby Aβ is involved in AD pathogenesis remains unclear, although increasing evidence links Aβ to alterations at synapses. More recently, accumulation of Aβ42 has been shown to occur within neurons (i.e., intracellularly) prior to the characteristic extracellular Aβ plaques. For example, there is early intracellular Aβ accumulation within disease-prone neurons in individuals with Down syndrome (trisomy 21), who invariably develop β-amyloidosis and dementia in their 30s and 40s, due to triplication of the *APP* gene located on chromosome 21. One can speculate that effects of intraneuronal Aβ accumulation may play a role in the observations of premorbid reductions in cerebral activity noted in individuals at risk for AD. Emerging experimental treatment strategies for AD are increasingly targeting reduction in Aβ.

Environmental risk factors for the development of AD include a history of head trauma associated with loss of consciousness, reduced educational attainment, and diet-related cardiovascular risk factors. Epidemiological studies suggest that increased physical and cognitive activities reduce the risk of developing AD.

Current treatment in AD has centered on findings from the 1970s and 1980s demonstrating that acetylcholine levels are especially reduced in AD brains. Cholinesterase inhibitors are the main class of currently employed US Food and Drug Administration (FDA)-approved medications for the symptomatic treatment of AD, although benefits, when present, tend to be modest, at best. The most recently approved medication for AD is an atypical *N*-methyl-D-aspartate (NMDA) antagonist; although the mechanism whereby it can

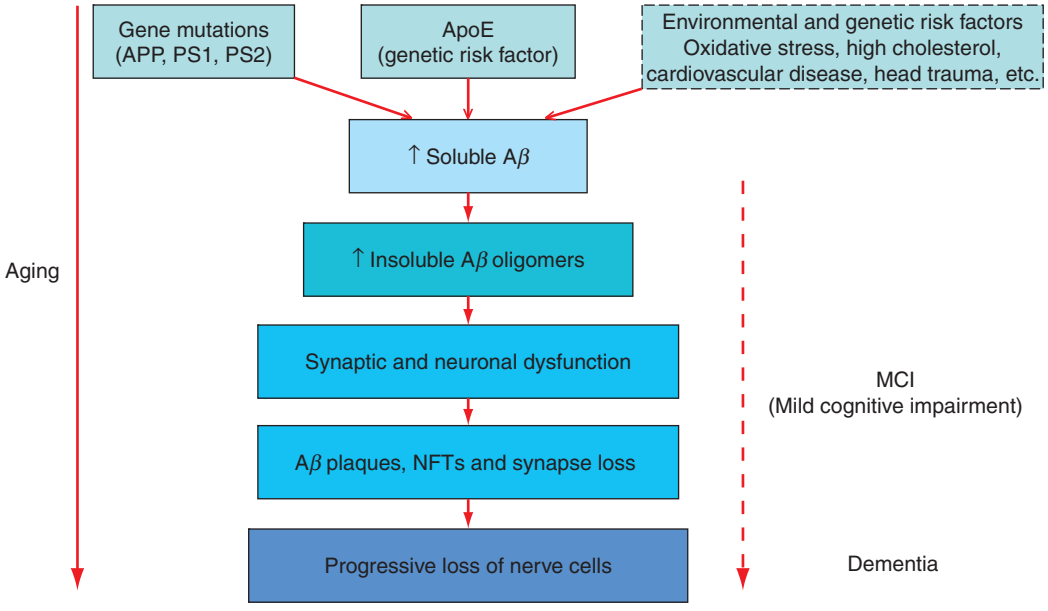

Figure 1 The amyloid cascade hypothesis. APP, amyloid precursor protein; PS1, PS2, presenilin proteins 1 and 2; Apo E, apolipoprotein E; Aβ, β-amyloid; NFTs, neurofibrillary tangles.

Table 2 Genetics of Alzheimer's disease

Familial Alzheimer's disease (FAD) – typically early onset (prior to age 65)
Gene mutations (autosomal dominant)
Amyloid beta (A4) precursor protein (*APP*; chromosome 21)
Presenilin 1 (*PSEN1*; chromosome 14)
Presenilin 2 (*PSEN1*; chromosome 1)
Genetic risk factors for late-onset Alzheimer's disease (after age 65)
Genetic polymorphisms (risk factor)
Apolipoprotein E (*APOE*, ε4 allele; chromosome 19)

provide benefit in the disease currently is poorly understood.

Vascular Dementias

Vascular-related damage to the brain can occur from different etiologies and probably is the second most common cause of dementia. A large embolic stroke can result in an acute onset of dementia. In contrast, multiple small 'lacunar' strokes can cause a more gradual, although usually stepwise, vascular dementia. Transiently reduced perfusion to the brain, such as occurs during cardiac arrest, can cause ischemic damage to vulnerable areas of the brain, including the hippocampus, thereby resulting in an amnestic dementia. Strategic infarct dementia is caused by a relatively small stroke in a part of the brain that is especially critical for cognition.

Vascular damage is also commonly observed in AD, and epidemiological studies are increasingly linking established cardiovascular risk factors, such as elevated cholesterol and diabetes, to an increased risk of AD. Progressive microvascular ischemic disease, especially of the periventricular white matter, formerly called Binswanger's disease, is another form of vascular dementia. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) due to mutations in Notch3 manifests with psychiatric disturbances, migraine, and strokes, and can result in dementia. Hereditary cerebral amyloid angiopathies (CAA) cause amyloidosis within the cerebral blood vessels and can cause dementia, especially from lobar hemorrhages. CAA is a common co-occurrence in AD, although specific mutations in *APP* or *BRI* genes can cause more localized amyloidosis of the cerebral vasculature in the relative absence of parenchymal amyloidosis.

Frontotemporal Dementia

FTD includes a group of previously separately classified diseases that have clinical, pathological, and genetic overlap. Formerly known as Pick's disease and frontotemporal lobar degeneration, FTD also includes semantic dementia and progressive nonfluent aphasia. Other studies are incorporating corticobasal (cortical–basal ganglionic) degeneration and even progressive supranuclear palsy (PSP) in an extended classification of FTD. FTDs are often misdiagnosed as AD. Early symptoms in FTD that can help differentiate it from AD are behavioral in nature; these include especially changes in personality, such as the development of inappropriate behavior or apathy. FTD is generally separated into types that affect behavior or language early on. Subtypes of FTD can be associated with concomitant parkinsonism or motor neuron disease. Common neuropathology features include the presence in many cases of neurofibrillary tangles and/or Pick bodies in the absence of Aβ plaques. Familial mutations in the principal component of tangles, the microtubule-associated protein tau, have been linked with some forms of FTD.

In contrast to AD, consolidation and retrieval of new memories are relatively spared in the early stages of FTD. At the same time, the early involvement of frontal lobes, with resultant behavioral alterations and/or impairment in executive functions (planning, reasoning, fluency, set shifting) can lead to apparent reduction in memory function, though recall is often assisted by cueing. Individuals with FTD can show signs of early dysfunction in complex attention and concentration, which are important for remembering. Structural imaging can show focal lobar atrophy of frontal and lateral temporal lobes, or with corticobasal degeneration, of superomedial frontal and parietal lobes.

Parkinson's Disease

Dementia is a frequent component of Parkinson's disease (PD). PD neuropathology is characterized by eosinophilic cytoplasmic inclusions within neurons containing α-synuclein and ubiquitin, called Lewy bodies. Lewy body diseases encompass PD and Lewy body dementia (or diffuse Lewy body disease), which can be viewed as a clinical–pathological spectrum differentiated by initial clinical symptoms and site of pathology. Initial Parkinsonism, characterized by psychomotor slowing, resting tremor, rigidity, and hunched posture/gait disturbance, in the setting of Lewy body pathology predominantly in the substantia nigra, pars compacta, is characteristic of PD. Initial cognitive decline, followed later by Parkinsonism (usually without tremor) and Lewy bodies predominantly in the cerebral cortex, is characteristic of Lewy body dementia. Lewy body dementia is characterized by visual hallucinations early in the course of the disease (within the first 2 years). The course of Lewy

body dementia may also include episodes of delirium not explained by infection or other toxic–metabolic derangement. Treatment with cholinesterase inhibitors can provide modest benefit. The neuropathology of AD is a frequent co-occurrence when prominent cortical Lewy bodies are present, prompting use of the term 'Lewy body variant of Alzheimer's disease.'

Psychiatric Diseases

Dementia of depression is a leading diagnosis among patients presenting with complaints of cognitive decline. Depression can cause subtle to profound cognitive deficits in multiple areas of cognition. The cognitive findings are typically characterized by prominent impairment in attention and concentration. Amnesia and other focal cognitive deficits, such as aphasia, are not characteristic of depression. Continued social stigma associated with psychiatric diseases can make diagnosis difficult. In contrast to AD, in which individuals often deny significant cognitive dysfunction, cognitive complaints are often pronounced in depression. At times, individuals may not be aware that they are depressed. Neurovegetative symptoms, particularly involving sleep and appetite, weight changes, loss of interest in previously enjoyable activities, and increased crying, can provide clues to the presence of depression. Cognitive dysfunction can markedly improve and normalize upon treatment of depression. Depression and AD are not mutually exclusive, but in fact co-occur frequently. New-onset depression in the elderly has been associated with an increased incidence in developing subsequent AD. Interpretation of the mechanism whereby depression might relate to AD is complicated by AD neuropathology also selectively affecting neurotransmitter systems involved in depression, such as the serotonergic system. Other psychiatric conditions, such as schizophrenia, formerly referred to as dementia praecox, can also cause dementia.

Toxic–Metabolic Encephalopathies

Although more typically a cause of delirium, various insults that can transiently alter cognitive function can also be a cause of dementia, when the insult persists (i.e., medication side effect) or damage is irreversible (i.e., toxins). Toxic–metabolic encephalopathies are generally characterized by a more global cognitive dysfunction, although toxins and nutritional deficiency can also cause more focal damage. Toxins that can cause dementia are numerous and most prominently include alcohol abuse. Chronic alcoholism can result in Wernicke–Korsakoff's syndrome, which is characterized by amnesia and confabulation. Acute (i.e., herpes encephalitis) and chronic (i.e., HIV/AIDS, syphilis, Lyme disease, chronic meningitis) infections can cause dementia. Prion diseases can be categorized either among the toxic–metabolic encephalopathies due to prion infectivity or as neurodegenerative diseases, with which they share many features. Prion protein, similar to β-amyloid in AD, aberrantly accumulates in the brain in genetic (i.e., Gerstmann–Sträussler–Scheinker disease) or sporadic prion diseases. Hormonal alterations, especially from thyroid disease, need to be considered in the workup of dementia. Nutritional and vitamin deficiencies (i.e., B_{12}, folate, and thiamine) can be causes of dementia. Various systemic illnesses, including collagen vascular diseases; pulmonary and hepatic disease, uremia, and anemia, need to be considered in the differential diagnosis of dementia. In contrast to most other causes of dementia, treatment of the underlying cause in toxic–metabolic dementias can halt or even reverse cognitive decline.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: MRI Studies; Alzheimer's Disease: Neurodegeneration; Alzheimer's Disease: Transgenic Mouse Models; Animal Models of Alzheimer's Disease; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Cognition in Aging and Age-Related Disease; Cognition: An Overview of Neuroimaging Techniques; Dementia and Language; Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Down Syndrome; Executive Function and Higher-Order Cognition: Neuroimaging; Functional Neuroimaging Studies of Aging; Neuroimaging.

Further Reading

Bertram L, McQueen MB, Mullin K, et al. (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: The AlzGene database. *Nature Genetics* 39(1): 17–23.

Cummings JL (2004) Alzheimer's disease. *New England Journal of Medicine* 351(1): 56–67.

Gouras GK, Almeida CG, and Takahashi RH (2005) Intraneuronal Aβ accumulation and origin of plaques in Alzheimer's disease. *Neurobiology of Aging* 26(9): 1235–1244.

Hou CE, Carlin D, and Miller BL (2004) Non-Alzheimer's disease dementias: Anatomic, clinical, and molecular correlates. *Canadian Journal of Psychiatry* 49(3): 164–171.

Knibb JA, Kipps CM, and Hodges JR (2006) Frontotemporal dementia. *Current Opinion in Neurology* 19(6): 565–571.

Kawas CH (2003) Clinical practice. Early Alzheimer's disease. *New England Journal of Medicine* 349(11): 1056–1063.

Lin MT and Beal MF (2006) Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature* 443(7113): 787–795.

Lleo A, Greenberg SM, and Growdon JH (2006) Current pharmacotherapy for Alzheimer's disease. *Annual Review of Medicine* 57: 513–533.

Petersen RC (2004) Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine* 256(3): 183–194.
Sampson EL, Warren JD, and Rossor MN (2004) Young onset dementia. *Postgraduate Medical Journal* 80(941): 125–139.
Walsh DM and Selkoe DJ (2004) Deciphering the molecular basis of memory failure in Alzheimer's disease. *Neuron* 44(1): 181–193.

Relevant Websites

http://www.alzforum.org – Alzheimer Research Forum.
http://www.alz.org – Alzheimer's Association.
http://www.nia.nih.gov – National Institute on Aging (National Institutes of Health) (see Alzheimer's Disease Information).

Dementia and Language

S Kemper, University of Kansas, Lawrence, KS, USA
L J P Altmann, University of Florida, Gainesville, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Effects of Dementia on Language

Language production begins with the formulation of a message and includes steps of discourse planning, lexical selection, and syntactic encoding. Language comprehension involves the analysis of discourse, syntactic, and lexical representations. Because of the many levels of analysis and the necessity of integrating information from all levels, both aspects of language use are compromised to some extent in all types of dementia beyond the small but significant effects of normal aging. Errors can arise at any stage of production or comprehension as a result of attentional lapses, processing limitations, or execution problems. For example, normative aging processes may exacerbate production or comprehension problems due to general slowing of cognitive processes, reductions of working memory capacity, or a breakdown of inhibitory processes. In dementia, these deficits are exacerbated by the breakdown of cortical connectivity within and between language processing regions, depending on the locus and severity of damage. This article examines what is currently known about the way dementia affects language use and attempts to distinguish pathological changes from normative age-related changes to language.

One issue has tended to dominate theoretical discussions of the effects of dementia on language: the modularity of phonological, syntactic, semantic, and/or discourse processes. Modularity, or autonomy, theory holds that language production or comprehension involves a sequence of processing stages or levels. Typically, four properties are ascribed to this sequence of processing stages: (1) each stage corresponds to a distinct linguistic rule system; (2) each stage involves a discrete representational system; (3) processing flows unidirectionally through the sequence of stages; and (4) each stage may be characterized by independent constraints on the rules, representations, or processing operations at that stage. Most models assume that there are a number of autonomous levels of processing, including a phonological level, which specifies the sequence of sounds composing each word; a morphological level, at which part-of-speech information is determined; a syntactic level, in which individual words are formed into phrases and clauses; a semantic level, at which meaning relations within and between words and phrases are determined; and a discourse, or pragmatic level of message structure and function. The degree of interactivity among these levels of representation and the possibility of bidirectional flow of information between levels are at the center of this controversy. Specific forms of dementia can attack any or all these levels, perhaps reflecting the mapping of linguistic functions onto cortical or subcortical structures.

A variety of specific tests for the effects of dementia on language are commonly included as part of the assessment of dementia, including tests of verbal fluency, picture naming, and immediate and delayed recall of word lists and short passages. For example, the Mini-Mental State Examination includes tests of sentence production, object naming, immediate and delayed verbal memory, and the ability to follow verbal commands, in addition to tests of orientation, attention, and constructional praxis. Psycholinguists and neuropsychologists have assessed the effects of dementia on various language-related tasks, such as picture naming, word-picture and sentence-picture verification, metalinguistic judgments, and reading and listening comprehension.

Many linguistic tests are relatively complex, involving a variety of component processes that may be affected by normal aging or other age-related conditions; some of these tests are often used to assess other cognitive processes, such as executive function. One example is verbal fluency. Tests of verbal fluency, such as the Controlled Oral Word Association Test, allow individuals a set amount of time, often 1 min, to generate words in response to a cue, such as words beginning with a target letter or phoneme (e.g., words beginning with an 'f') or members of a semantic category (e.g., the names of animals). Because the task draws on both semantic memory and executive function, deficits are found in nearly all dementias. For example, it is well established that individuals with Parkinsonian dementia and Alzheimer's dementia have verbal fluency deficits, particularly on tests involving retrieval of members of semantic categories. These deficits in individuals with Parkinsonian dementia have typically been attributed to nonsemantic problems associated with executive function, such as problems of inhibitory control, whereas similar deficits in individuals with Alzheimer's dementia have been attributable to a breakdown of semantic memory or semantic retrieval processes. Some attempts have been made to develop separate semantic and executive-function

components of verbal fluency performance by tracking the time course of responses, requiring responses to alternate or switch between two categories, and varying other task parameters.

Differential Effects

Dementias differentially affect language processes, primarily because of the location and scope of damage from the particular disease. Thus, some symptoms may be similar to those resulting from stroke affecting the same cortical region. However, dementias are degenerative, so the initial effects often are much milder than those resulting from stroke, more difficult to detect in early stages, and progressive. Alzheimer's dementia has been studied in some depth, whereas research into other dementias has been more limited.

Alzheimer's Dementia

Early in the disease, Alzheimer's dementia primarily affects posterior association cortex and only rarely extends to the perisylvian region. Consequently, phonology and phonetics, the sounds of words and the ability to articulate them, are typically affected only late in the disease, if at all. Similarly, morphological processing (e.g., appropriate use of verb tenses and sentence structure) is also considered a primarily perisylvian function that is usually preserved during early stages of the disease, although grammatical errors increase in frequency as the disease progresses.

In contrast, Alzheimer's dementia has a more significant effect on other aspects of language that rely on semantic knowledge (i.e., word meanings) and conceptual knowledge (i.e., knowledge about items and their attributes) and on the ability to integrate the meanings of several words. Often an early sign of Alzheimer's dementia is a marked increase in word-finding problems during conversation, as individuals experience frequent 'tip of the tongue' events, substitute generic references (e.g., "that thing"), and rely on metalinguistic references (e.g., "oh, you know who I mean"). At diagnosis, most Alzheimer's dementia patients score below normal on picture-naming tests, and this impairment increases over time. Errors are typically substitutions of words from the same semantic category, (e.g., saying 'horse' for 'giraffe') or category names (e.g., 'animal' or 'tool'), although other types of errors, such as visual errors and failures to respond, also occur. However, the structure of semantic knowledge remains intact until quite late in the disease; thus, patients have little difficulty sorting pictures by semantic category or verifying the category membership of an item. Performance on word–picture matching tasks in which the individual must identify a spoken or written target within an array of pictures is also more preserved than picture naming, particularly if there is little visual similarity among distracter pictures. Visual-perceptual difficulties may also contribute to the difficulties of Alzheimer's dementia patients in such tasks.

Extensive research has suggested that word-finding problems in conversation and picture-naming tasks are due to both a degradation of the meanings of words and difficulty accessing word meanings. Preserved partial meanings are evident in 'circumlocutions'; for example, someone unable to recall 'horse' may say, "Oh, you ride it. You put a leather thing on its back and you ride it. They race them too." The disease appears to impair the ability to focus word searches, resulting in activation of conceptually similar words (e.g., horse, deer, cow, donkey). These alternatives may compete for production, contributing to a breakdown in word retrieval. As the disease progresses, word meanings themselves appear to deteriorate. Even so, individual words, particularly high-frequency, highly familiar words, may be preserved, and comprehension is somewhat spared relative to production throughout the disease process.

Oral reading is preserved in mild Alzheimer's dementia, although comprehension of written sentences and texts is impaired. Thus, they may be able to read connected text that they cannot understand. In contrast, writing is impaired relatively early in Alzheimer's dementia and is characterized by simplified grammatical structures, loss of content, and regularized spelling.

Although the speech of individuals with mild or moderate Alzheimer's dementia is often described as 'empty speech,' it is quite distinct from the empty speech produced by individuals with fluent aphasia following stroke. Individuals with Alzheimer's dementia do not produce strings of words with little relevance to context, perseverate, or use ungrammatical phrases or 'word salad' jargon until extremely late in the disease, if at all. Instead, the conversational speech of individuals with Alzheimer's dementia is characterized by the overuse of general nouns (e.g., 'thing,' 'stuff') and verbs (e.g., 'get,' 'do') and by an overreliance on pronouns. These phenomena decrease the amount of information conveyed and can make following their conversations difficult. Even so, in Alzheimer's dementia, the fundamental building blocks of sentences are relatively intact; sentences contain noun and verb phrases and may also include complex grammatical constructions. Thus, conversational speech is often relatively coherent until moderate and late stages of the disease.

Understanding conversational speech requires listeners to combine the meanings of many words, often

across sentence boundaries; keep these in memory; and integrate the new information with past knowledge. Thus, it is an extremely demanding task. Patients with Alzheimer's dementia often miss a good deal of the information in a conversation; however, they may understand the gist of a conversation by relying on contextual cues. As the disease progresses, the appropriateness of language declines as behavior becomes increasingly disinhibited. Even mild-mannered Alzheimer's dementia patients may become verbally abusive. Typically, speech output also declines as the need to communicate dwindles. Gradually, conversational speech disappears, eventually ending in mutism.

Parkinsonian Dementia

Parkinson's disease primarily affects the ability to smoothly coordinate and perform speech, resulting in slow, dysarthric articulation in combination with a soft, breathy voice. The ability to produce consonants that require tongue movements, such as /t/, /d/, /l/, /n/, /k/, and /g/, is most impaired. Individuals with Parkinsonian disease may evidence 'articulatory undershoot,' in which the tongue does not quite reach the appropriate spot in the mouth or does not achieve full closure to block the flow of air when necessary. Further, if they do achieve adequate closure, they may either release it too fast or maintain it too long, decreasing intelligibility. Speech rate varies among patients; some may speak faster than normal, others at a normal rate. Often, there is a lack of awareness of these deficits; individuals with Parkinson's disease may feel that they are shouting when they are barely audible to others.

Recent research has revealed that other aspects of language use are also impaired in Parkinsonian dementia, suggesting either that subcortical structures play a role in language use or that the reduced dopamine production in Parkinson's disease impairs frontal lobe function, secondarily affecting language use. Converging evidence for the latter argument comes from findings that individuals with Parkinsonian dementia often score poorly on tests of immediate memory, executive function, and processing speed, all of which have been found to be associated with performance on language tasks in healthy older adults. Consequently, performance on all language tasks that rely on memory or executive function, such as reading comprehension and the production of complex sentences, is likely to be impaired. Micrographia, extremely small writing, is characteristic of Parkinson's disease. However, there are no studies on the content of writing in Parkinsonian dementia and very few on reading connected text or producing connected speech.

Frontotemporal Dementias

Frontotemporal dementia (FTD) is usually described as having three variants that differ in their behavioral manifestation depending on the locus of damage. These include primary nonfluent aphasia, semantic dementia, and frontal-variant FTD. In contrast to Alzheimer's disease, damage in FTD is relatively focal; consequently, episodic memory, spatial cognition, and orientation are usually preserved until late in the disease. Frontal-variant FTD manifests primarily as social and behavioral impairments. Language is relatively unimpaired in this group, although there are mild deficits affecting category and letter fluency and a very mild impairment of picture naming. Most language impairments in this subgroup seem to be related primarily to deficits in executive function.

Primary nonfluent aphasia Primary nonfluent aphasia, also known as primary progressive aphasia, is characterized by increasing difficulty programming speech movements, which leads to severely impaired spoken output, or dysarthria, from early in the disease. Speech becomes slower and more effortful over time and is marked by long word-finding pauses, phonemic paraphasias, and distortions. Intonation patterns also deteriorate, possibly because of the increasing effort needed to produce speech or the decreasing length of utterances.

Word-finding difficulties are often among the initial complaints of individuals with primary nonfluent aphasia; these difficulties are evident in both picture naming and spontaneous speech and become increasingly severe over time. As mentioned above, the difficulty seems to be in determining and producing the sounds of the words. In contrast, comprehension of single words as measured by word–picture matching tests is largely intact until late in the disease. Thus, in individuals with primary nonfluent aphasia (about 30% at the mild stage of the disease), there is a large discrepancy between the ability to comprehend words and the ability to name pictures, and this discrepancy increases as the disease progresses. Based on these findings, several researchers consider that the primary impairment in primary nonfluent aphasia affects the phonological output representations.

Primary nonfluent aphasia may affect different parts of speech in different ways; phonemic paraphasias may occur primarily with nouns, and semantic paraphasias primarily with verbs. There is mounting evidence that in some patients, primary nonfluent aphasia degrades the semantic or conceptual representations of verbs and that this degradation, in turn, disrupts sentence production. As in acquired aphasias, verb impairments are often accompanied by

impaired use of verb endings and difficulties comprehending complex or reversible sentences. Some individuals with primary nonfluent aphasia have been described as producing agrammatic speech. Very little is currently known about how primary nonfluent aphasia affects written production or the comprehension of sentences or discourse in either spoken or written modalities.

Semantic dementia Semantic dementia, also called progressive fluent aphasia, results when FTD primarily affects the anterior inferior temporal lobe. The defining feature in semantic dementia is a progressive loss of semantic and conceptual knowledge. Thus, semantic dementia affects performance on all tasks that require the access to or manipulation of word meanings, leading to extremely poor picture naming and severely impaired verbal fluency. In contrast, sentence grammar and discourse are relatively preserved although disrupted by the prevalence of word-finding difficulties.

Because the disorder impairs conceptual knowledge as well as semantic knowledge, individuals with semantic dementia may not be able to demonstrate how to use common items, such as a hammer or comb. Visual-perceptual aspects of meaning (e.g., shape and relative size) can be especially impaired, particularly knowledge of biological categories, such as fruits, vegetables, or animals. Comprehension of sentences is also impaired, although the patients may be able to abstract a 'gist' meaning based on whatever partial semantic information is available to them. Due to the deterioration of conceptual knowledge, reading comprehension and written communication are also impaired. Reading aloud may resemble surface alexia, in which irregularly spelled words are pronounced as they are written. Similarly, writing may also be impaired, with regularized spelling and low information content.

Other dementias: vascular dementia and Lewy body disease The effects of other dementias on language have not been well studied. Two of the more common types of dementia on which there is some language research are vascular dementia and Lewy body disease. Vascular dementia is difficult to differentiate from Alzheimer's dementia without reliance on neuroimaging. Both syndromes show impairments in word finding, picture naming, and letter and category fluency. However, individuals with vascular dementia are particularly poor at letter fluency relative to category fluency, and the reverse is true for individuals with Alzheimer's dementia. Individuals with vascular dementia also outperform those with Alzheimer's dementia on verbal memory tasks, such as free verbal recall, delayed verbal recall, and recognition memory of verbal material. Both vascular dementia and Alzheimer's dementia patients show a range of attention, executive function, and working memory impairments. Thus, the cognitive and linguistic characteristics of vascular dementia and Alzheimer's dementia are very similar, and to date, there are no 'cut-off' scores available to distinguish between these groups.

Lewy body disease primarily affects visual-spatial skills and the extrapyramidal systems rather than language, although it frequently accompanies both Parkinson's disease and Alzheimer's dementia. While those with Lewy body disease and Alzheimer's dementia are similar with respect to performance on picture-naming and word fluency tasks, attention and executive function are usually more impaired in Lewy body disease. On the other hand, individuals with Lewy body disease typically have better verbal memory, especially with respect to delayed recall and recognition memory, than do individuals with Alzheimer's dementia. Individuals with Lewy body disease also exhibit profound visuospatial and visuoconstructional impairments very early in the disease; for example, they are often unable to copy the overlapping pentagons in the Mini-Mental State Examination, even at an early stage of the disease. The spontaneous speech of the two groups differs qualitatively. Conversational speech in Alzheimer's dementia is characterized by a lack of content and relatively intact grammar; conversational speech in Lewy body disease is marked by confabulation.

Unfortunately, few studies have examined sentence and discourse level language production or comprehension in either spoken or written modalities in individuals with vascular dementia or Lewy body disease.

Interactions and Interventions

Family members and other caregivers of individuals with dementia typically cite communication problems as a major source of caregiver stress. Breakdowns in communication, whether from word-finding problems, dysarthria, or repetitive questions, may lead to other problem behaviors, such as agitation and aggression, and may impair quality of care. Little systematic research has assessed intervention and training programs targeted at alleviating communication problems with dementing individuals – largely as a result of the perception that, because these are

degenerative conditions, language treatment would be a waste of resources. Two studies of semantic training in semantic dementia have shown immediate gains in naming ability but no maintenance of these gains at follow-up. At least one study has shown maintenance of treatment gains in picture naming in Alzheimer's disease 2 months after treatment. However, most intervention programs target caregivers, emphasizing the use of functional communication strategies, such as making eye contact, repeating key words, using yes–no questions and lexical prompts, and providing memory books and other conversational stimuli. Other programs have targeted the development of techniques for minimizing repetitive questions and other disruptive behaviors.

A special speech register, sometimes termed 'elderspeak,' is often recommended as a way to enhance or facilitate communicating with older adults, especially those with dementia. Elderspeak has been characterized as involving a simplified speech register with exaggerated pitch and intonation, simplified grammar, limited vocabulary, and slow rate of delivery. Many of the characteristics of elderspeak, such as its slow rate and simplified syntax and vocabulary, may be helpful for ensuring comprehension in impaired populations, although there has been little research assessing its effectiveness. On the other hand, certain characteristics of elderspeak, such as the use of exaggerated prosody, collective pronouns (e.g., saying 'we' instead of 'you'), diminutives (e.g., 'Chrissy' instead of Chris or Mrs. Smith), and terms of endearment (e.g., 'honey' and 'darling'), resemble another register of speech, 'baby talk.' These characteristics are often perceived as demeaning or controlling. Alternative strategies for enhancing communication with adults with dementia must be developed by combining clinical observation with empirical research.

Implications and Conclusions

There has been considerable interest in using linguistic measures to predict risk, aid in diagnosis, and track the progression of dementia, particularly Alzheimer's dementia. Performance on tests of verbal memory, such as immediate and delayed word list recall, as well as tests of verbal fluency, appear to be particularly sensitive indicators of risk for dementia and its onset and seem to covary with other known risk factors, such as apolipoprotein E genotype. However, deficits in verbal fluency are not specific to Alzheimer's disease, as discussed above. The relationship between linguistic ability and the risk for Alzheimer's disease and longevity has been the focus of the Nun Study, an ongoing longitudinal, epidemiological study of aging. Low linguistic ability in young adulthood, determined from an analysis of language samples written by the nuns at age 18 to 32 years, was associated with increased risk for poor performance on cognitive and memory tests in late adulthood, increased neuropathology characteristic of Alzheimer's disease, and increased all-cause mortality among participants. These findings suggest that linguistic ability may offer a general measure of cognitive and neurological development. Low linguistic ability in young adulthood may reflect suboptimal neurocognitive development, which in turn may increase or reflect susceptibility to Alzheimer's and other dementing diseases.

See also: Aging of the Brain and Alzheimer's Disease; Aging and Memory in Humans; Alzheimer's Disease: An Overview; Cognition in Aging and Age-Related Disease; Dementia; Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Language in Aged Persons; Language: Cortical Processes; Language: Aphasia; Semantic Memory; Speech Production: Adult; Speech Production: Development; Stroke; Word Production; Word Recognition.

Further Reading

Altmann LJP, Kempler D, and Andersen ES (2001) Speech production in Alzheimer's disease: Re-evaluating morphosyntactic preservation. *Journal of Speech, Language, and Hearing Research* 44: 1069–1082.

Bayles KA and Kaszniak AW (1987) *Communication and Cognition in Normal Aging and Dementia*. Boston: College-Hill.

Grossman M (1999) Sentence processing in Parkinson's disease. *Brain and Cognition* 40: 387–413.

Grossman M and Ash S (2004) Primary progressive aphasia: A review. *Neurocase* 10: 3–18.

Hamilton HE (1994) *Conversations with an Alzheimer's Patient*. Cambridge, UK: Cambridge University Press.

Kemper S, Thompson M, and Marquis J (2001) Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging* 16: 600–614.

Ripich DN (1994) Functional communication with AD patients: A caregiver training program. *Alzheimer's Disease and Associated Disorders* 8: 95–109.

Rogers TT, Ivaniou AD, Patterson K, and Hodges JR (2006) Semantic memory in Alzheimer's disease and the frontotemporal dementias: A longitudinal study of 236 patients. *Neuropsychology* 20: 319–335.

Ryan EB, Giles H, Bartolucci G, and Henwood K (1986) Psycholinguistic and social psychological components of communication by and with the elderly. *Language and Communication* 6: 1–24.

Savundranayagam MU, Hummert ML, and Montgomery RJV (2005) Investigating the effects of communication problems on caregiver burden. *Journal of Gerontology: Social Sciences* 60: S48–S55.

Snowdon DA, Kemper SJ, Mortimer JA, Greiner LH, Wekstein DR, and Markesbery WR (1996) Cognitive ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Journal of the American Medical Association* 275: 528–532.

Williams KN (2006) Improving outcomes of nursing home interactions. *Research in Nursing and Health* 29: 121–133.

Relevant Websites

http://www.alz.org – Alzheimer's Association research page.
http://www.alzforum.org – Alzheimer Research Forum.
http://www.alzheimers-research.org – Alzheimer's Research Foundation.

Demyelinating Diseases

A Javed and B G W Arnason, University of Chicago, Chicago, IL, USA

© 2009 Published by Elsevier Ltd.

Introduction

The demyelinating diseases are important because of their frequency and the disability that they cause. Common to all is a patchy destruction of myelin sheaths, usually secondary to an inflammatory response. Myelin loss occurs in other conditions: examples include (1) genetically determined defects in myelin metabolism, (2) toxin exposures, and (3) infections of oligodendrocytes, the myelin producing cells of the central nervous system (CNS). These entities are usually not classified as demyelinating diseases.

The demyelinating diseases can be divided into those affecting CNS myelin and those affecting peripheral nervous system myelin. CNS diseases include multiple sclerosis (MS), acute disseminated encephalomyelitis (ADEM), neuromyelitis optíca ((NMO) Devic's disease), and acute necrotizing hemorrhagic encephalomyelitis (ANHE). Peripheral nervous system diseases include acute inflammatory demyelinating polyradiculoneuropathy, chronic inflammatory demyelinating polyradiculoneuropathy, anti-myelin-associated glycoprotein (MAG) neuropathy, and POEMS syndrome (so named due to polyneuropathy, organomegaly, endocrinopathy, M protein, and skin changes). The major demyelinating diseases are considered here, with emphasis directed toward mechanisms of disease causation and to treatments directed against them.

CNS Demyelinating Diseases

Multiple Sclerosis

MS presents with recurrent attacks of focal neurological disorder involving brain, spinal cord, and optic nerves. Disease usually begins in young adults. In the United States, one person in 500 has this disease by age 50. During episodes of relapsing–remitting MS, neurologic deficits worsen over a period of several days to 3–4 weeks, after which they remit. Remission may be complete after early attacks but more often is incomplete, and as relapses accumulate, stepwise progression leads to increasing deficit. In 15% of cases the illness manifests as a slowly progressive disease, so-called primary progressive MS. This is particularly likely when onset begins after age 40. In cases with earlier onset, and after multiple relapses over many years, the disease usually transitions into a slowly but continuously progressive worsening of motor deficit called secondary progressive MS. The cause of MS is unknown, but there is agreement that abnormal immune system function characterizes the disease. Evidence for this is summarized in the following discussion.

Certain histocompatibility antigen alleles are overrepresented in MS. Notable is the DR2 allele of the HLA histcombatibility complex. Possession of DR2 increases risk for MS fourfold. Many diseases are associated with histocompatibility abnormalities. Such diseases usually have (1) a chronic or recrudescent course, (2) a moderate tendency to be inherited that does not obey classical Mendelian rules, (3) an inflammatory component, and (4) an infectious or autoimmune basis in almost all instances in which cause is known. The first three features apply to MS. By analogy, the fourth likely applies as well.

Twenty percent of MS patients have a first- or second-degree blood relative with the disease. This could reflect genetic endowment, a shared environmental factor, or some combination of the two. Concordance for MS in identical twins approximates 25% and for siblings, 1–3%. Thus, there must be a strong genetically determined propensity to develop the disease, but there must also be some environmental initiator. The genetic component includes histocompatibility, but almost surely additional genetically determined characteristics, at present unknown, are involved. All attempts to identify the environmental factor that sets MS in motion have failed. Most researchers believe it will prove to be an infectious agent. One candidate is the Epstein–Barr virus that causes infectious mononucleosis. All MS patients have serologic evidence of prior infection with this virus at time of diagnosis.

MS is most common in north Europeans and in those whose forebears were north Europeans. MS is almost unknown among blacks in Africa. Among African Americans who develop MS, there is an increased representation of the HLA-DR2 allele, an uncommon allele among blacks in Africa. Thus, the propensity of African Americans to develop MS, and its rarity among blacks in Africa, may reflect both genetic and environmental differences. MS is uncommon in China and India. MS remains uncommon among descendents of Chinese immigrants to America, but not among descendents of immigrants from India, again indicating both regional environmental factors and genetic endowment as determinants for development of MS.

MS lesions consist of foci of demyelination, known as plaques, in CNS white matter. The earliest lesions

consist of minute pervenular foci within white matter. Over time these accumulate, enlarge, and coalesce to form plaques of considerable size, many of which abut the lateral ventricles. Axons coursing through regions where myelin has been lost are relatively but not completely spared, at least initially, although they are denuded of their insulation. Naked axons may continue to conduct nerve impulses, albeit imperfectly. Stripped axons can cease functioning when sustained effort is attempted, or if body temperature is elevated. Fatigue on effort, and worsening of deficits with fever, are characteristics of MS.

Active MS plaques contain T lymphocytes. These can be subdivided into two major categories known as CD4 and CD8 cells. CD4 and CD8 are surface proteins that facilitate interactions between T cell receptors and antigenic peptides and antigen-presenting cells such as macrophages. Antigenic peptides recognized by T cells are contained within clefts of major histocompatibility complex proteins expressed on the surfaces of antigen-presenting cells. CD4 cells recognize peptide fragments presented by class II histocompatibility alleles. CD8 cells recognize fragments presented by class I alleles. DR2, overrepresented in MS, is a class II allele, indicating a major role for CD4 cells in lesion formation in the early stages of the disease. CD4 cells can be further subdivided into T helper Th1 and Th2 subtypes. Th1 cells are responsible for delayed-type hypersensitivity responses. They secrete numerous cytokines (signaling proteins of the immune system), notably interleukin-2 (a stimulator of T cell proliferation), interferon-γ (an activator of macrophages), and lymphotoxin. The latter two proteins can damage oligodendrocytes. Elevated numbers of interferon-γ-secreting cells are detected in the blood during MS attacks and in MS plaques when disease is active. Interferon-γ, together with interleukin-2, activates macrophages that then secrete tumor necrosis factor, a protein which shares with interferon-γ and lymphotoxin the ability to damage oligodendrocytes. There is a yin-yang between Th1- and Th2-type T cells. Preferential activation of Th2-type T cells inhibits Th1 cells, and in this way may be protective in MS. Activation of Th2-type T cells is one goal of MS therapy.

MS plaques also contain macrophages. Macrophages often predominate within plaques even early on, and invariably predominate once disease has become progressive . Macrophages lack immunological specificity but can be armed to attack tissues under the influence of cytokines released by T cells, or by antibodies (immunoglobulins) released by B cells. Cytokines clearly activate macrophages during MS attacks. Antibodies may further facilitate macrophage activation during MS attacks and are, in all probability, the main activator of macrophages in progressive forms of MS. Thus, macrophages are the final vectors of myelin destruction throughout the course of MS. They also scavenge debris.

CD8 cells can be subdivided into cytotoxic cells and immunoregulatory (suppressor) cells. CD8 cells are found within MS plaques, but whether they are cytotoxic, have suppressor function, or both, has not been established. Interferon-β activates cytotoxic T cells yet protects MS patients (see later), making a role for cytotoxic T cells in MS problematic. In contrast, the function of circulating CD8 suppressor T cells, as measured by several tissue culture assays, is grossly deficient when MS is active, and treatment with interferon-β (a drug approved for treatment of MS) restores circulating suppressor T cell function to near normal levels in relapsing–remitting MS patients. Interferon-β also inhibits production of interferon-γ, lymphotoxin, and tumor necrosis factor, at least *in vitro*.

While the details of the immune response during MS relapses are imperfectly worked out, one possible scenario is as follows: failed regulatory T cell function (Th2 cells, CD8 cells, or both) permits reactivation of Th1 cells with specificity for an unknown brain antigen or antigens; the Th1 cells then stimulate macrophages, which attack and destroy myelin and oligodendrocytes. Restoration of regulatory cell function ends the relapse.

The disease produced by MS has prompted development of therapies directed toward correcting the immune system abnormalities. Interferon-β, an agent approved for the treatment of relapsing–remitting MS, reduces MS attack frequency by 35–55%. In addition to activating CD8 suppressor cell function, and hence inhibition of proinflammatory cytokine production, Interferon-β also inhibits T cell entry into the CNS. Glatiramer acetate, another approved drug for treatment of MS, reduces MS attack frequency by 30%. It has been proposed that it activates regulatory CD8-type T cells, and promotes Th2 cell development. Natalizumab, a third class of drug, is recommended for treatment of relapsing–remitting MS patients who fail to respond to other therapies. Natalizumab drastically curtails T cell trafficking into the CNS. Trials with other agents are ongoing.

A set of diseases that share numerous clinical and pathologic features with MS can be induced in laboratory animals. These diseases are known collectively as experimental autoimmune encephalomyelitis (EAE). They are induced by immunization with any of several CNS myelin proteins. The most commonly employed are myelin basic protein (MBP), proteolipid protein (PLP), and myelin oligodendrocyte protein

(MOG). The pathogenesis of most forms of EAE is a T-cell-mediated delayed hypersensitivity response directed against the individual proteins used to immunize the animals. The T cells responsible for the activation of macrophages that eventuates in myelin destruction during EAE are analogous to human Th1-type T cells, suggesting that T cells of this lineage are the prime movers in acute attacks of MS as well. Antibodies directed against proteins expressed on the myelin surface contribute to the demyelination seen in subacute and progressive forms of EAE, suggesting, by analogy, an ancillary role for antibodies in early MS and a substantial role, possibly a dominant one, in progressive MS. That antibodies are likely to have a major role in the axonal damage of progressive MS is supported by the observation that T cell ablation has minimal effect on progressive motor disability once MS has evolved into its progressive form.

T cells from animals with EAE proliferate when exposed to MBP, PLP, or MOG, depending on the protein used to induce disease. T cells from MS patients seldom proliferate when exposed to these same proteins, or at least to no greater extent than do T cells from controls, so that an immune response directed against them is unlikely to be implicated in MS. Nonetheless, given the similarities between EAE and MS, it is possible that there may be some unique antigen against which an immune response is directed in MS. None has been found with certainty to date despite assiduous search. This could mean that the EAE models provide a false lead. Alternatively, and more likely, the antigen (or antigens) for MS simply awaits discovery.

Mature oligodendrocytes are depleted in MS plaques. For this reason some have proposed that the immune attack in MS may be directed against oligodendrocytes rather than against their product, myelin, inasmuch as compromised oligodendrocytes might be expected to divest themselves of their myelin. Remyelination can follow myelin loss. This occurs to a more limited extent in MS than in other demyelinating processes. Preoligodendrocytes, the precursors of oligodendrocytes, are responsible for remyelination in the human brain. A substantial pool of them is present in the adult brain. During the events that eventuate in remyelination, preoligodendrocytes mature to become oligodendrocytes. Curiously, preoligodendrocytes are present in substantial numbers in MS plaques that are depleted of mature oligodendrocytes, yet remyelination fails. What prevents preoligodendrocytes from repairing myelin in MS is unknown.

Spinal fluid of MS patients contains immunoglobulin G (IgG) in increased amount, a finding that provides useful corroborating evidence when a diagnosis of MS is suspected. Spinal fluid IgG from MS patients, subjected to isoelectric focusing, exhibits oligoclonal banding. Oligoclonal bands bespeak expansion of specific B cell clones that produce monoclonal (i.e., single antigen-specific) antibodies within the brain, the basis for which is not understood. Much of the monoclonal antibody made in the brain in MS is irrelevant to disease pathogenesis. This does not negate the possibility that some may be directed against MS-relevant antigens, yet undiscovered.

Spinal fluid in MS contains T cells. Many are activated even when disease seems quiescent (i.e., during clinical remissions). The activated cells belong to the delayed hypersensitivity response–effector cell lineage. The data point to more persistent disease activity than is apparent clinically. Similarly, when magnetic resonance imaging (MRI) scans are obtained monthly, gadolinium-positive lesions are detected in five scans obtained between overt MS attacks for every gadolinium-positive scan obtained during a relapse. Gadolinium transport into the CNS is a marker for disease activity. The finding again indicates that the illness is far more active than it appears to be clinically, and supports the strategy of continuous long-term immunomodulatory treatment for the illness.

To sum up, immune mechanisms are implicated in MS, but no disease-specific antigen against which an immune-mediated attack is directed has been proved to be the cause. Similarly, attempts to implicate a virus in disease pathogenesis have failed. Yet, immunomodulatory therapies favorably affect the long-term course of MS.

Acute Disseminated Encephalomyelitis

ADEM, a rare disease, is an acute multifocal demyelinating process that follows an infectious illness. Formerly, ADEM occurred particularly often in children with measles. Nowadays the illness most often follows a nondescript viral infectious illness. Although encountered in adults, most cases occur in children. Children seldom get MS, indicating that ADEM constitutes a distinct entity. ADEM occurs throughout the world, and may even be more common in less developed countries where MS is rare, than in developed ones where MS is common. In ADEM, the white matter of the brain and spinal cord is peppered with multiple foci of acute demyelination of varying size. The foci contain T cells and macrophages and most, sometimes all, are simultaneously gadolinium positive on MRI scanning. The illness is usually monophasic; recurrences after a delay are uncommon, although recrudescences may be encountered when glucocorticoid treatment (see later) is abbreviated.

In animals, an acute monophasic form of EAE, with T cell sensitivity to MBP, mimics the human

disease. Sensitivity of lymphocytes to MBP, and antibodies to MBP, have been demonstrated in ADEM, and the disease is a reasonably faithful human counterpart of acute monophasic MBP-induced EAE. Animals, recovered from acute monophasic EAE, are refractory to attempts to induce recurrences. Refractoriness is believed to reflect expansion, in recovered animals, of MBP-specific suppressor cells that damp down any future response to MBP. In MS, unlike the situation in ADEM, recurrences are the rule. Perhaps failed suppressor cell function favors these recurrences – that is, if a means to persistently augment suppressor function were at hand, MS relapses might be prevented.

ADEM begins abruptly and evolves rapidly. Headache and delirium at onset that give way to lethargy and coma are common in childhood cases, less so in adults. Seizures are not infrequent, particularly in children. Later, multiple focal signs emerge, flaccid paralysis of the limbs and brain stem signs being particularly common. The symptoms observed at the onset of ADEM are associated with inflammation alone. The multiple focal findings that follow coincide with onset of demyelination. In former times, mortality was 20%, with 50% of survivors having residual deficits. The prognosis has subsequently become much more favorable.

Treatment of choice is high-dose glucocorticoid administered systemically for several days, followed by an extended oral glucocorticord taper over 6–8 weeks. Controlled trials are lacking, but recrudescences are fewer when this regimen is employed than when it is not, strongly suggesting a benefit from glucocorticoid treatment, as might be expected for a T-cell-mediated process. Remyelination during recovery from ADEM is more extensive than that seen in MS, and the majority of patients recover completely. MRI scanning during the acute phase reveals subcortical white-matter lesions with fuzzy margins, some quite large, with, unlike the situation in MS, periventricular sparing. Follow-up MRIs often reveal complete resolution of prior lesions and, again unlike MS, no new ones. The spinal fluid often shows a florid lymphocytosis at disease onset.

Neuromyelitis Optica (Devic's Disease)

NMO is an uncommon inflammatory disease in which inflammation, restricted to the optic nerves and spinal cord, occurs either sequentially or simultaneously. There has been debate regarding the etiology of this entity given its topographical specificity. Until recently NMO was usually viewed as an MS variant.

NMO has features that permit it to be set apart from MS or ADEM (Table 1).The clinical course is monophasic in one of five of instances. Four-fifths of patients have recurrent bouts of inflammation involving the spinal cord and/or the optic nerves. These relapses are reminiscent of relapsing–remitting MS. In patients with monophasic NMO, visual and motor involvement is severe initially, but recovery, although delayed, is usually substantial. In contrast, patients with relapsing disease have less severe involvement initially, but as attacks accumulate progressive disability follows. Disability can become devastating, with total blindness and complete motor paralysis sometimes being the end result. There is a 30% incidence of respiratory failure, usually fatal, in recurrent NMO due to upper cervical cord and medullary function compromise. Respiratory failure is rare in MS or in ADEM.

NMO lesions have unique features. Spinal cord lesions exhibit prominent vascular fibrosis and hyalinization. Spinal cord NMO lesions involve gray as well as white matter and there may be prominent necrosis, cavitation, and axonal loss. Extensive demyelination is seen, as is extensive oligodendrocyte loss. Lesions in the spinal cord tend to be centrally located in NMO rather than peripheral (i.e., restricted to white matter), as is the case in MS. NMO lesions, seen on MRI scans of the spinal cord, span a greater longitudinal extent than do the spinal cord lesions of MS. Macrophages predominate; T cells are rare. Distinct to NMO is prominent infiltration of neutrophils, eosinophils, and ringlike perivascular deposits of immunoglobulin M (IgM) and IgG. Complement products are also seen in NMO lesions.

The incidence of other autoimmune diseases in NMO is 15-fold increased over that in the general population. Frequent associations are with Sjögren's disease, systemic lupus erythematosus, rheumatoid arthritis, mixed connective tissue disease, and myasthenia gravis. Evidence pointing to other autoimmune processes may be subtle. Rare associations of NMO with tuberculosis, cancer (especially small-cell lung carcinoma), and lymphoma have been described.

An IgG antibody specific for NMO has been discovered. The antibody has a sensitivity of 73%, a specificity of 91%, and permits a clear distinction to be made between NMO and MS. The antibody is directed against aquaporin 4, a water channel protein expressed in astrocyte foot processes.

NMO patients fail to respond to standard MS therapies. Acute attacks are typically treated with high-dose systemic glucocorticoids plus courses of plasmapheresis or of intravenous immunoglobulin. Pulse treatments with plasmapheresis or with intravenous immunoglobulin given every 4–6 weeks, and/or long-term maintenance on an immunosuppressive drug such as azathioprine, methotrexate, or mycophenolate

Table 1 Features of the major CNS demyelinating diseases

Feature	*NMO*	*Multiple sclerosis*	*ADEM*
Median age of onset (years)	40	30	15
Sex ratio (F:M)	8:1 (relapsing)	2:1 (relapsing)	1:1
Ethnicity	↑ Incidence in non-Caucasians	Primarily Caucasian	All populations at risk
Clinical course			
Monophasic	20%	NA	>95%
Relapsing	80%	85% initially	<5%
Primary progressive	NA	15%	NA
MRI features			
Brain white-matter lesions	Absent at onset	Oval, usually perpendicular to ventricles	Large, disseminated, seldom periventricular
Spinal cord lesions	Longitudinally extensive	Punctate	Variable
Optic nerve lesions	Unilateral or bilateral with poor recovery	Usually unilateral with good recovery	Unilateral or bilateral with good recovery
Cerebrospinal fluid findings			
White blood cell count	↑↑↑ Lymphocytes, neutrophils	↑ Lymphocytes	↑↑ Lymphocytes, occcasionally neutrophils
Total protein	Elevated	Typically normal	elevated
Oligoclonal bands	Usually absent	>90% present	Usually absent
Pathology			
Gray matter	Necrosis	Limited demyelination	Limited demyelination
White matter	Necrosis, demyelination	Demyelination	Demyelination, edema
Lesion immunopathology			
Macrophages	+++	+++	+++
T cells	+/−	+++	+++
Eosinophils	++	None	None
Neutrophils	+++	None	Rare
Antibodies	+++	Variable	Variable
Complement	++	Variable	Variable
Coexisting autoimmune diseases	30–50%	Infrequent	Not seen
Disability progression	Relapse related	Independent of relapses	Relapse related

NMO, Neuromyelitis optica; ADEM, acute disseminated encephalomyelitis.

mofetil, have been given in attempts to lessen the likelihood of future relapses. Rituximab has been used recently in recurrent NMO prophylaxis. Rituximab, a monoclonal antibody directed against the CD20 protein expressed on the surface of B cells, potently depletes B cell number and lessens antibody production.

Acute Necrotizing Hemorrhagic Encephalomyelitis

ANHE is a rare disease that resembles ADEM save for its apoplectiform onset and unique pathologic features. There is massive destruction of white matter to the point of liquefaction, and the lesions are flecked with multiple small hemorrhages. As in ADEM, regions of tissue damage contain lymphocytes and macrophages, but, additionally, necrotic vessels. Polymorphonuclear cells infiltrate regions of necrosis and are found in spinal fluid, as are red blood cells. A model for ANHE can be induced by giving endotoxin to animals about to develop acute EAE. This 'double hit' superimposes a Shwartzman reaction onto the animal model for ADEM. A few cases of ANHE complicating MS have been reported.

Peripheral Nervous System Demyelinating Diseases

Demyelination in the peripheral nervous system may be caused by immune mechanisms that involve T cell- and macrophage-mediated mechanisms, by antimyelin antibodies, or by cytokines released at a distance from nerves. Mixed mechanisms are also seen. Examples are presented in the following sections.

Guillain–Barré Syndromes

Demyelinating and axonal forms of Guillain–Barré syndromes are recognized. The demyelinating form, often designated as acute inflammatory demyelinating neuropathy (AIDP), predominates in North America, Europe, and Australia. Axonal forms predominate in other countries, most notably in China. Annual frequency of AIDP is 1.5 persons per 100 000 across all

age groups, so that one person in 1000 will contract this illness at some point over their lifetime. AIDP follows an acute infectious illness in two-thirds of patients. Some 10% of cases follow infection with cytomegalovirus, another 5% follow Epstein–Barr virus infection (both are herpes viruses). Some 20–30% of cases follow infection with *Campylobacter jejeuni*, a bacterium that causes diarrhea. An additional 5% of cases follow infection with *Mycoplasma pneumoniae*. This agent can also incite a concurrent ADEM to give a mixed CNS and peripheral nervous system presentation. Weakness, leading to frank paralysis, that requires respiratory support in 20% of AIDP cases evolves subacutely over several days for up to 4 weeks; 5% of patients succumb and the rest recover, usually completely, although 15% have residual disability. AIDP is usually monophasic. Relapses occur months or years later in fewer than 5% of cases.

Pathologic examination reveals peripheral nerve infiltration with lymphocytes and monocytes/macrophages. The latter strip myelin to leave denuded axons. Segmental demyelination occurs throughout the peripheral nervous system but more so in nerve roots and distal nerve terminals. Remyelination, which can be extensive, correlates with relatively rapid restoration of function. In severe cases axons may be interrupted, followed by either slow regrowth once the inflammatory process has ceased, and hence delayed restoration of function, or no regrowth and permanent deficit.

Complement deposition on the surface of Schwann cells is seen early on, suggesting a potential role for antibody directed against a myelin antigen that remains unknown. It is also likely that nerve-invading CD4-type T cells instruct macrophages to attack myelin. In cases with a protracted course, a late-stage role for CD8-type cytotoxic T cells has also been proposed. Courses of either plasmapheresis or of intravenous gamma globulin have been shown, in controlled clinical trials, to provide substantial benefit in AIDP.

Animal models that mimic AIDP with varying degrees of fidelity exist. Experimental autoimmune neuritis (EAN) is an acute inflammatory neuritis induced by immunization with proteins of peripheral nerve myelin. T cells from such animals proliferate when challenged with myelin proteins and the disease can be transferred to naive recipients with T cells from EAN donors, indicating that EAN is mediated largely by T cells. EAN mimics the histological and clinical features of AIDP, yet sensitivity of T cells from AIDP patients to myelin proteins can be shown in only a minority of cases.

The axonal forms of the Guillain–Barré syndrome are characterized by purely motor involvement (acute motor axonal neuropathy, or AMAN), a mixed motor and sensory neuropathy (acute sensorimotor axonal neuropathy, or ASMAN), and the Miller Fisher syndrome, in which oculomotor paralysis, ataxia, and areflexia are cardinal features. These illnesses are mediated by antibodies to gangliosides situated on axonal membranes. Most cases follow infections with *C. jejeuni*. The antibodies are directed against oligosaccharides of the *Campylobacter* bacterium, but cross-react, because of molecular mimicry, with peripheral nerve gangliosides.

In AMAN, antibodies to anti-GM1/anti-GD1a gangliosides are the rule, whereas in the Miller Fisher syndrome antibodies to GQ1b ganglioside are usually detected. GM1 is enriched at nodes of Ranvier and at motor nerve terminals. Deposition of antibodies and of complement at these sites explains motor nerve functional compromise. GQ1b is enriched in extraocular nerves so that anti-GQ1b antibodies preferentially localize to extraocular nerves in the Miller Fisher syndrome. The axonal forms of the Guillain–Barré syndrome involve nerve impulse conduction block when mild, and irreversible axonal transection when severe. They are not demyelinating. Nonetheless, the finding that antiganglioside antibodies cause paralysis has led to speculation that AIDP may likewise be mediated by antibodies directed against some substance expressed on the surface of peripheral nerve myelin. The issue remains open.

Chronic Inflammatory Demyelinating Neuropathy

Chronic inflammatory demyelinating neuropathy (CIDP) may be an AIDP variant. The illness is characterized by progressive loss of motor and sensory function that evolves over many years. The process is roughly symmetrical, weakness of proximal and distal muscles predominating. Sensory dominant forms are uncommon. The onset is usually insidious without, unlike AIDP, any evident precipitating event. Some patients have a series of relapses prior to developing a progressive course. CIDP can occur at any age but is more common with advancing years. CIDP predominantly affects spinal roots, major plexuses, and proximal nerve trunks. The process is demyelinating, as readily documented by electrophysiologic testing. This reveals dispersed slowing of nerve conduction velocity. The spinal fluid protein is elevated, in keeping with leakage from damaged spinal roots. Tendon reflexes are usually lost.

Histologic examination of active lesions reveals lymphocytic infiltrates, usually modest, and macrophage-mediated demyelination. Macrophages insert processes between myelin lamellae, phagocytose myelin, and denude axons. The process is morphologically

identical to that seen in AIDP, albeit more indolent. Repair occurs to some extent as witnessed by thinned myelin and shortened internodal distances, cardinal features of successful remyelination. As the illness evolves, interruption of naked axons becomes engrafted onto the demyelinating process. Failed attempts at remyelination are evidenced by layers of Schwann cell cytoplasm separated by collagen fibers forming concentric rings, or so-called onion bulbs.

The process is held to be primarily T cell mediated, but T cell reactivity to characterized myelin proteins is seldom found, and the putative antigen, or antigens, against which the T cell response is directed remains unknown. A role for antibody in CIDP has also been proposed since a small proportion of patients have antibody to the Po protein of peripheral nerve myelin.

Most patients respond modestly to oral glucocorticoids when these are first given. This treatment is not helpful in AIDP. There is also evidence that intravenous IgG and plasma exchange may offer modest benefit for a time. Treatment for this condition remains unsatisfactory over the longer term.

Anti-Myelin-Associated Glycoprotein Neuropathy

MAG neuropathy is a demyelinating neuropathy mediated by monoclonal IgM antibody that binds to an oligosaccharide shared by MAG, a minor myelin protein, and two glycolipids of peripheral nerve, sulfated 3-glucuronyl lactosaminyl paragloboside and sulfated 3-glucuronyl paragloboside. The clinical picture is one of a primarily sensory distal and symmetric neuropathy. IgM antibodies have low affinities, but this is compensated for by high avidities due to their pentameric structure. MAG is primarily expressed in uncompacted myelin of the periaxonal and paranodal regions and in Schmidt–Lantermann clefts. Anti-MAG antibody and complement are deposited at these sites and MAG is selectively lost. MAG is implicated in the maintenance of axonal function and its loss compromises axonal function. Anti-MAG antibodies injected into the sciatic nerves of cats cause demyelination, proving that anti-MAG antibody causes the neuropathy. A blood test for anti-MAG antibody is reliable and confirms the diagnosis.

Electrophysiologic studies reveal a demyelinating process with distal accentuation of reduced conduction velocity. Treatment with antimetabolic drugs such as the purine analog fludarabine has been reported as helpful, as has rituximab, a humanized murine monoclonal antibody directed against the CD20 antigen expressed on B cells. Rituximab exerts a cytotoxic effect on B cells with, as a consequence, reduced anti-MAG antibody titers.

POEMS Syndrome

The POEMS syndrome, a demyelinating neuropathy, was originally described as a complication of sclerosing myeloma, an immunoglobulin-secreting plasma cell tumor. Most myelomas are osteolytic and multifocal, but 3% of them are solitary tumors that cause bone to sclerose. Fifty percent of patients with a sclerosing myeloma develop the POEMS syndrome. Radiation therapy of a sclerosing myeloma can be followed by remission of the neuropathy, indicating that some product released by the tumor causes the POEMS syndrome. Those sclerosing myelomas that cause the POEMS syndrome invariably secrete monoclonal IgG or IgA immunoglobulin into the blood. Documentation of this secretion is required for diagnosis, although the monoclonal immunoglobulin may be secreted in such miniscule amounts as to prove difficult to detect. The light chain of the IgG or IgA released by the tumor is of the λ type in almost all instances.

Other B cell dyscrasias can be complicated by the POEMS syndrome, including multicentric angiofollicular lymph node hyperplasia (Castelman's disease), an entity characterized by gross lymph node hyperplasia, and polyclonal plasma cell proliferation. In cases complicated by the POEMS syndrome, focal accumulations of monoclonal plasma cells secreting IgG λ are detectable, pointing to a communality with sclerosing myeloma-associated POEMS syndrome.

The POEMS acronym emphasizes several of its cardinal features: polyneuropathy, the dominant clinical feature; organomegaly; endocrinopathy, most often gonadal dysfunction; M (i.e., myeloma) protein; and skin changes, including hyperpigmentation and hypertrichosis. There are additional features not described by the acronym. These are elevated intracranial pressure with or without papilledema, edema of the extremities, effusions in body cavities, and, ultimately, renal failure and/or failure of multiple other organs. The process that begins as a neuropathy ultimately becomes widespread.

The neuropathy is symmetrical and begins with distal sensory abnormalities, numbness, and dysesthesias predominating. The process then ascends and weakness supervenes. Nerve conduction studies reveal evidence for both demyelination and axonal loss, as do nerve biopsies. Importantly, there is no inflammatory cell infiltrate nor is immunoglobulin deposition within nerves seen. The latter finding indicates that monoclonal IgG or IgA does not cause the neuropathy. Endoneurial blood vessels show endothelial cell hypertrophy. Endothelial cell processes extend into the lumen, reduce luminal diameter, and hence promote hypoxia. Disruption of the tight

junctions that ordinarily join nerve endothelial cells to one another is also seen with, as a consequence, leaky vessels. Vascular endothelial growth factor (VEGF) is an angiogenic hormone. Elevated levels of VEGF are observed in the blood in POEMS syndrome and levels correlate with the tempo of worsening. VEGF levels return to normal with successful treatment of solitary myelomas by radiation, or with alkylating agents and glucocorticoids when the process is more widespread. These findings have led to the postulate that excessive VEGF production by plasma cells causes endothelial cell proliferation and leaky vessels. VEGF, acting far from its site of origin to compromise blood flow throughout the body, fits well with the multiorgan involvement seen in the later stages of the POEMS syndrome.

See also: Autoimmune Autonomic Neuropathy; Axonal Injury in Demyelinating Disease and CNS Injury; Demyelination and Demyelinating Antibodies; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Neuropathy: Peripheral; Transplantation of Myelin Forming Cells.

Further Reading

Anlar B, Basaran C, Kose G, et al. (2003) Acute disseminated encephalomyelitis in children: Outcome and prognosis. *Neuropediatrics* 34: 194–199.

Dale RC, de Sousa C, Chong WK, et al. (2000) Acute disseminated encephalomyelitis, multiphasic disseminated encephalomyelitis and multiple sclerosis in children. *Brain* 123: 2407–2422.

Dispenzieri A, Kyle RA, Lacy MQ, et al. (2003) POEMS syndrome: Definitions and long-term outcome. *Blood* 101: 2496–2506.

Hughes RAC, Allen D, Makowska A, et al. (2006) Pathogenesis of chronic inflammatory demyelinating polyradiculoneuropathy. *Journal of the Peripheral Nervous System* 11: 30–46.

Keegan MB and Noseworthy JH (2002) Multiple sclerosis. *Annual Review of Medicine* 53: 285–302.

Lennon VA, Kryzer TJ, Pittock SJ, et al. (2005) IgG marker of optic-spinal multiple sclerosis binds to the aquaporin-4 water channel. *Journal of Experimental Medicine* 202(4): 473–477.

Scarlato M, Previtali SC, Carpo M, et al. (2005) Polyneuropathy in POEMS syndrome: Role of angiogenic factors in the pathogenesis. *Brain* 128: 1911–1920.

Steck AJ, Stadler AK, and Renaud S (2006) Anti-myelin-associated glycoprotein neuropathy. *Current Opinion in Neurology* 19: 458–463.

Willison HJ (2005) The immunobiology of Guillain–Barré syndromes. *Journal of the Peripheral Nervous System* 10: 94–112.

Demyelination and Demyelinating Antibodies

F J Seil, Oregon Health and Science University, Portland, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004 Elsevier B.V. All rights reserved.

Demyelinating antibodies are generally defined by their destructive effects on myelin *in vitro*. Such antibodies are of interest because of a possible role in demyelinating diseases *in vivo*, including multiple sclerosis (MS), and because of their potential use as probes for the study of myelination.

Experimental allergic encephalomyelitis (EAE) is a demyelinating disease that can be induced in animals by intradermal injection of Freund's complete adjuvant (FCA) in combination with whole central nervous system (CNS) tissue, white matter, or purified CNS myelin. It was discovered that sera from animals with EAE induced by these antigens produced demyelination upon application to myelinated CNS tissue cultures while sparing axons. The demyelinating activity was complement dependent and associated with the immunoglobulin G (IgG) fraction of serum, and additional evidence was presented to support the antibody nature of these serum factors. The demyelinating reaction was reversible, and removal of the antisera, even after several days, was followed by the reappearance of myelin *in vitro*. If similar antisera were applied to CNS cultures before myelin formation, myelination and oligodendrocyte differentiation were inhibited, even at much lower concentrations of serum than required to produce demyelination. Antisera directed against whole CNS tissue, white matter, or purified CNS myelin were specific for centrally myelinated fibers and did not demyelinate peripheral nervous system (PNS) myelin *in vitro*.

The primary encephalitogenic agent in CNS tissue is myelin basic protein (MBP), and EAE can be induced with MBP or with MBP fragments plus FCA. In early attempts to identify the specific myelin antigens that induced antimyelin antibodies, attention was directed to encephalitic proteins and peptides. Sera from animals with EAE induced with encephalitogenic peptides or MBP did not demyelinate CNS cultures, and anti-MBP sera did not inhibit myelin formation. The major CNS myelin protein, proteolipid protein (PLP), which can induce a chronic form of EAE, also did not provoke the production of demyelinating or myelination-inhibiting antibodies. Galactocerebroside (GC), a glycolipid component of myelin, did not induce EAE, but tissue culture demeyelinating and myelination-inhibiting activity were found in sera from rabbits inoculated with GC and FCA. However, the antimyelin activity was not confined to CNS myelin, because PNS cultures were also demyelinated and myelination-inhibited.

Other myelin components were investigated. Myelin-associated glycoprotein (MAG), a minor component of CNS myelin localized at the myelin membrane surface, and thus a good candidate for antibody induction, induced neither EAE nor antimyelin antibodies. Immunization of rabbits with gangliosides, glycolipid components of neuronal and myelin membranes, had been reported to induce a form of demyelinating disease. No demyelinating or myelination-inhibiting activity was evident in CNS–PNS cultures exposed to rabbit antisera to GMI, a major ganglioside. By immunoabsorption and fractionation experiments and complement fixation studies, tissue culture demyelinating factors in sera from whole CNS- or CNS myelin-sensitized guinea pigs were characterized as antibodies to an antigen designated M2. M2, which was absent in PNS myelin and was different from MBP and GC, was identified as a glycoprotein component of oligodendrocyte membrane.

Extraction of CNS myelin with chloroform-methanol removes PLP, almost all of the MBP, and most of the lipids. The remaining chloroform-methanol insoluble protein (CMIP) fraction contains a large number of low- and high-molecular-weight proteins, including the glycoproteins, among which are MAG, M2, and the myelin oligodendrocyte glycoprotein (MOG), the last a minor protein localized at the external surfaces of myelin sheaths and oligodendrocyte membranes. Rabbits inoculated with the CMIP fraction and FCA did not develop clinical EAE but did have occasional focal perivascular mononuclear inflammatory infiltrates, sometimes with spread into the surrounding CNS parenchyma, and their sera demyelinated and inhibited myelination of centrally myelinated fibers while sparing peripheral myelin. Immunization of rodents with MOG induced EAE and application of a monoclonal anti-MOG antibody to myelinated reaggregating brain cell cultures resulted in reversible, complement-dependent demyelination. Other characteristics of anti-MOG antibody remain to be defined. However, because MOG is present in the CMIP fraction of myelin, it is probable that such features as specificity for CNS myelin and myelination-inhibiting capability will be the same as for anti-CMIP antibodies. It is also probable that MOG and M2 are the same glycoproteins.

A PNS equivalent of EAE is experimental allergic neuritis (EAN), which can be induced with FCA plus

whole PNS tissue or the P2 protein fraction of PNS myelin. The P2 protein is also a basic protein but is structurally different from MBP and is found only in PNS myelin. Antisera directed against whole PNS tissue consistently demyelinated and inhibited PNS myelin formation in tissue cultures, while having variable results against CNS myelin. Antisera to the neuritogenic P2 protein, as with the encephalitogenic counterpart, MBP, demonstrated no antimyelin activity *in vitro*.

A summary of the various antigens evaluated for their ability to induce antimyelin antibodies is presented in Table 1. Also indicated in the table is whether or not the antigens induced experimental demyelinating disease. It is apparent from this table that induction of EAE and EAN can be completely dissociated from induction of antimyelin antibodies, raising the question of the relevance of these antibodies to the pathogenesis of disease. Although such antibodies may not be necessary for disease induction, it has been suggested and some supporting evidence has been presented that antimyelin antibodies have an augmenting role in demyelination, their presence being correlated with more severe demyelination in experimental animals. The notion that combined action of cellular and humoral mechanisms to produce inflammatory demyelinating lesions in EAE was reinforced by the finding that MOG, which induces EAE with both inflammatory and demyelinating components, evoked a T-cell-mediated immune reaction and a B-cell-secreted demyelinating antibody response. On the other hand, extensive demyelination has been described in the absence of antibodies that bind to myelin, indicating that antibody augmentation is not always necessary for demyelination. Antimyelin antibodies have also been reported to opsonize myelin for phagocytosis by macrophages.

Sera from some subjects with MS, a human CNS demyelinating disease, and from human subjects with idiopathic polyneuritis (Guillain–Barré syndrome), in which peripheral nerves and nerve roots are segmentally demyelinated, also demonstrated demyelinating activity in CNS and PNS tissue cultures. The MS serum factors paradoxically did not inhibit myelination. Sera from subjects with MS specifically demyelinated centrally myelinated fibers, and sera from patients with idiopathic polyneuritis were active against both CNS and PNS myelin. Demyelinating factors from the latter group of human subjects have not been systematically studied, but attempts have been made to further characterize MS sera.

There was a degree of correlation between the presence of serum demyelinating factors and disease activity in MS. Demyelinating activity of MS sera was removed by absorption with brain tissue and a nonmyelin CNS tissue pellet that included oligodendrocytes but not with purified myelin. No immunostained myelin was evident when demyelinating MS sera were applied to CNS cultures and fixed and treated with peroxidase-conjugated antihuman globulin, suggesting that the demyelinating activity of MS sera was not associated with a myelin-binding

Table 1 Induction of experimental demyelinating disease and antimyelin antibodies by neural antigens

Antigens	*Disease EAE/EAN*	*Antibodies*			
		Demyelinating		*Myelination inhibiting*	
		CNS	*PNS*	*CNS*	*PNS*
CNS antigens					
Whole CNS	+	+	−	+	−
CNS myelin	+	+	−	+	ND
MBP	+	−	−	−	ND
PLP	+	−	−	−	ND
GC	−	+	+	+	+
MAG	−	−	−	−	ND
GMI	±[a]	−	−	−	−
M2	ND	+	ND[b]	ND	ND
CMIP	±[c]	+	−	+	ND
MOG	+	+	ND	ND	ND
PNS antigens					
Whole PNS	+	±	+	+	+
P_2 protein	±	−	−	−	−

[a]Several reports of induction of an experimental demyelinating disease in rabbits.
[b]Sera containing anti-M2 antibodies did not bind to PNS myelin.
[c]No clinical EAE but occasional perivascular inflammatory infiltrates histologically.
ND: not determined. Other abbreviations as in text.

immunoglobulin. Depletion of γ-globulin from MS sera did not decrease their tissue culture-demyelinating capability, and isolated immunoglobulin fractions from most MS sera did not demyelinate CNS cultures, further supporting the concept that demyelinating activity in most MS sera is associated with a serum factor other than γ-globulin.

More recently, anti-MOG antibodies bound to disintegrating myelin sheaths were detected in acute lesions from three human subjects with MS, suggesting that MOG could be a target for antibody-mediated demyelination in some cases of MS. However, given that 60–70% of subjects with active MS have serum demyelinating activity and that only 12.5% of demyelinative MS sera are associated with immunoglobulins, 8–9% of MS cases at most would appear to have serum demyelinating activity related to immunoglobulins, such as anti-MOG antibody, which could be consistent with the currently espoused concept that MS is a disease of heterogeneous pathogenesis. Also, serum demyelinating factors are not specific for MS, as demyelinating activity has been reported in sera from a significant proportion of human subjects with other neurological diseases, further complicating interpretation of the significance of serum demyelinating factors in human diseases.

There are other reasons for studying antibodies to specific myelin antigens, however, including their use as cell markers (e.g., anti-GC antibody as a marker for oligodendrocytes) and as probes for dissecting some of the steps involved with myelination. It has been shown, for example, that antisera to whole CNS tissue can, in the presence of complement, inhibit oligodendrocyte differentiation *in vitro*, whereas decomplemented anti-CNS sera allow oligodendrocyte differentiation but prevent oligodendrocytes from successfully wrapping axons and forming compact myelin. Similar probes may define still other aspects of the complex interactions between neurons and supporting cells that result in myelination.

See also: Demyelinating Diseases; Mitochondrial Encephalomyopathies; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath.

Further Reading

Archelos JJ, Storch MK, and Hartnung H-P (2000) The role of B cells and autoantibodies in multiple sclerosis. *Annals of Neurology* 47: 694–706.

Bradl M and Linington C (1996) Animal models of demyelination. *Brain Pathology* 6: 303–311.

Brosnan C and Raine CS (1996) Mechanisms of injury in multiple sclerosis. *Brain Pathology* 6: 243–257.

Cuzner ML and Norton WT (1996) Biochemistry of demyelination. *Brain Pathology* 6: 231–242.

Seil FJ (1981) Tissue culture studies of neuroelectric blocking factors. In: Waxman SG and Ritchie JM (eds.) *Demyelinating Disease: Basic and Clinical Electrophysiology*, pp. 281–288. New York: Raven Press.

Seil FJ (1982) Demyelination. In: Fedoroff S and Hertz L (eds.) *Advances in Cellular Neurobiology*, vol. 3, pp. 235–274. New York: Academic Press.

Seil FJ (1994) Effects of humoral factors on myelin in organotypic cultures. In: Herndon RM and Seil FJ (eds.) *Multiple Sclerosis: Current Status of Research and Treatment*, pp. 33–50. New York: Demos.

Seil FJ (2003) Tissue culture studies of demyelination and remyelination. In: Herndon RM (ed.) *Multiple Sclerosis: Immunology, Pathology, and Pathophysiology*, pp. 105–112. New York: Demos.

Storch M and Lassmann H (1997) Pathology and pathogenesis of demyelinating diseases. *Current Opinion in Neurology* 10: 186–192.

Storch M, Stefferl A, Brehm U, et al. (1998) Autoimmunity to myelin oligodendrocyte glycoprotein mimics the spectrum of multiple sclerosis pathology. *Brain Pathology* 8: 681–694.

Dendrite Development, Synapse Formation and Elimination

H Cline, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the central nervous system of adult animals, many neurons have functionally polarized architecture in which dendrites, which receive information from other neurons, are physically separate from axons, which send information to other neurons. In general, axons and dendrites both have complex treelike structures, called arbors. The area covered by the axons and dendrites and the density of branches within the dendritic and axonal arbors govern the type and number of neurons which are connected within a circuit. Consequently, the mechanisms that control the development and maintenance of neuronal structure critically affect the ability of the neuron to function within the brain circuit. For instance, the importance of neuronal structure in brain function is clear from studies of brains of children with mental retardation, in which neurons are dwarfed in size compared to those in healthy children.

The acquisition of mature neuronal structure is classically described as being governed by both 'intrinsic' and 'extrinsic' factors, although in fact extrinsic factors, such as growth factors and synaptic inputs, affect intrinsic events such as gene transcription and trafficking of guidance molecules to the cell surface. Several recent studies using modern molecular genetic, imaging, and electrophysiological methods now provide strong evidence that excitatory synaptic inputs control the development of neuronal structure in the intact brain.

Synaptogenesis and Synapse Maturation

The formation and maturation of synapses can be distinguished into several steps:

1. *Establishment of an adhesive contact*. Dynamic filopodial processes from growing presynaptic axons and postsynaptic dendrites come into contact and form an initial adhesive contact, possibly mediated by integrins, cadherins, or Wnt/Frizzled cell surface adhesive molecules.
2. *Conversion of the adhesive contact to a nascent synapse*. In the case of glutamatergic synapses in the vertebrate central nervous system, nascent synapses are characterized by the predominance of *N*-methyl-D-aspartate (NMDA)-type glutamate receptors. NMDA receptors (NMDARs) require postsynaptic depolarization at the same time as ligand (i.e., glutamate) binding in order to permit conductance through the channel. As such, the NMDAR acts as a coincidence receptor. In the context of synapse formation, this ensures that transmission at nascent synapses occurs when other inputs to the postsynaptic neuron surpass a threshold synaptic strength. NMDARs are permeable to calcium. Intracellular calcium signaling may be required for maturation of the developing synapse.
3. *Synapse maturation*. Glutamatergic synapse maturation is characterized by the recruitment of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type ionotropic glutamate receptors into synapses. This renders the synapses functional at resting potentials. The assembly of the complex postsynaptic density, including scaffolding proteins and signaling proteins, occurs as glutamate receptors are trafficked into developing synapses.

Features of Nascent and Mature Synapses

Studies in many experimental systems indicate that synaptic transmission at newly formed synapses is mediated by the NMDA type of glutamate receptors and that AMPA receptors (AMPARs) are added to synapses as they mature. Synapses with only NMDARs are 'silent' at resting potential due to the voltage-dependent block of the NMDAR channel, and the addition of AMPARs to synapses renders them functional at resting potentials. Similarly, the fraction of silent synapses, in which transmission is mediated solely by NMDARs, is high in early stages of synapse formation and decreases as synapses and neurons mature, due to the insertion of AMPARs at synaptic sites. Consequently, the fraction of silent synapses and the ratio of AMPA to NMDA receptor-mediated transmission can be used as indicators of synaptic maturity (see Figure 1). Although some glutamatergic synapses reportedly develop without NMDARs, a sequence in which transmission at new synapses is mediated principally by NMDARs followed by the addition of functional AMPARs to synaptic sites appears to occur at the majority of glutamatergic synapses. Indeed, trafficking of AMPARs into developing synapses is required for their maturation and is required for synaptic plasticity in adult animals under a variety of conditions, including experience-dependent sensory plasticity and fear-conditioned learning. Although these data suggest

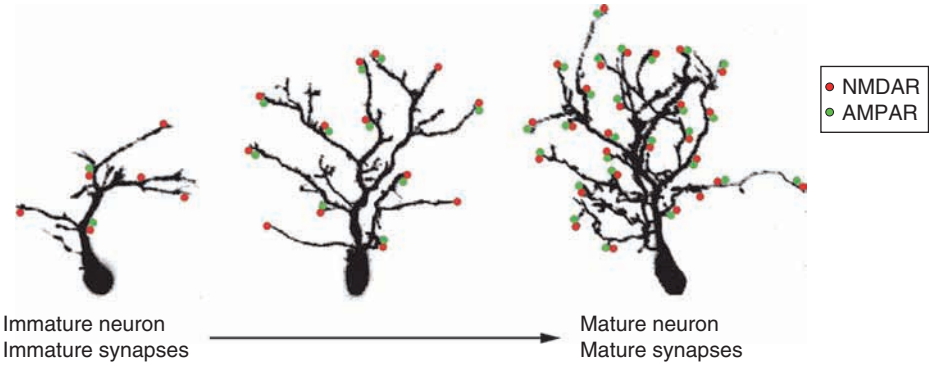

Figure 1 Synaptogenesis and dendrite development are concurrent. *In vivo* images of a neuron from the optic tectum of *Xenopus laevis* tadpoles were collected once a day over 3 days. The dendritic arbor increases in complexity over this time period through the net addition of arbor branches. Images collected over shorter intervals (not shown) demonstrate that net arbor growth occurs as a result of rapid branch addition and retraction. Synapses are located throughout the arbor. New excitatory synapses are mediated principally by *N*-methyl-D-aspartate (NMDA)-type glutamate receptors, shown as red dots. α-Amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type receptors, shown as green dots, are added to these synapses as they mature. Even after 3 days of growth, the dendritic arbor continues to show structural changes and to add new (NMDA receptor (NMDAR) only) synapses.

that AMPAR trafficking is the key element required for synapse maturation and synaptic plasticity, AMPAR trafficking into synapses may escort other proteins into the postsynaptic density, and it is possible that these other proteins in combination with AMPARs are required for synaptic plasticity.

Developing glutamatergic synapses can be distinguished from mature synapse by several other features. Synapse maturation includes changes in the presynaptic element, including recruitment of presynaptic machinery and proteins into the axon terminal and recruitment of synaptic vesicles. Ultrastructural studies indicate that developing synapses have few, sparsely packed synaptic vesicles, and that the most reliable ultrastructural indicator of synaptic maturation is the density of synaptic vesicles.

In vertebrate neurons the subunit composition of synaptic NMDARs switches during development, from receptors containing mostly NR2B subunits to those including NR2A subunits. This change in subunit composition can be detected pharmacologically and by a faster decay of the synaptic responses. Furthermore, immature synapses may show greater spillover of transmitter from the synaptic cleft to activate extrasynaptic receptors, due to relatively poor envelopment of nascent synapses by glia.

Synapse maturation and developmental plasticity of the strength of synaptic communication are thought to be regulated by changing the efficacy of transmitter release from the presynaptic terminal and by changing the response of the postsynaptic neuron to the input signal. The mechanisms underlying these changes are similar to those which occur during learning and memory in the adult animal. For instance, calcium influx through NMDARs, Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) activation, and AMPAR trafficking underlie both synapse maturation and changes in synaptic strength associated with learning.

Process of Dendritic Arbor Development

In vivo time-lapse imaging experiments have shown that dendrites of central nervous system neurons grow by the highly dynamic addition and retraction of fine branches. Although most newly added branches retract within about 10 min of being added to the arbor, a small fraction of branches is maintained and these extend to become more stable components of the arbor. The newly added branches may sample the local environment for appropriate presynaptic contact sites. Establishment and maintenance of synapses may then confer a longer lifetime on the branches that form stable synapses and permit the dendritic arbor to enlarge over time (see Figure 1).

Classical studies using fixed anatomical preparations have led to the idea that axonal and dendritic arbors go through a period of exuberant overgrowth followed by pruning of branches based on competitive mechanisms (see Figure 2). This model of neuronal growth and pruning suggests that neurons in the adult brain have limited capacity to recover following injury, and this in turn has discouraged scientists and clinicians from exploring the potential of the brain to regain function after trauma.

Modern *in vivo* imaging methods demonstrate that neurons in the brain grow very differently than thought from classical studies. Specifically, branch additions and retractions are concurrent, so that the neuron has an ongoing capacity to add new branches and to refine its connections within the

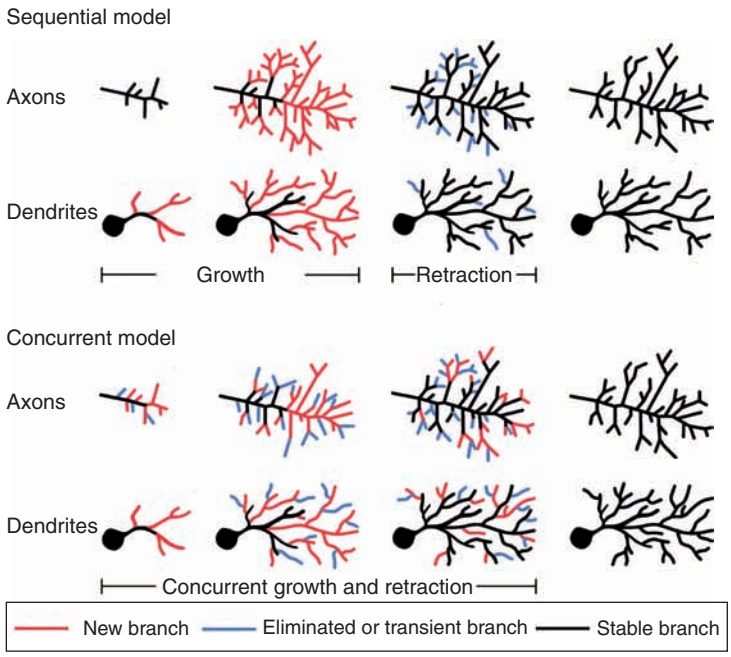

Figure 2 Models of dendritic and axon arbor elaboration. Dendritic and axon arbor elaboration and pruning are concurrent, not sequential. Top: Diagrams of the patterns of growth of axons and dendrites according to the traditional model, in which neurons undergo an exuberant growth phase followed by a temporally distinct phase of net branch retraction. Bottom: Diagrams of the patterns of growth of axons and dendrites in which branch additions and retractions are concurrent. Adapted from Hua JY and Smith SJ (2004) Neural activity and the dynamics of central nervous system development. *Nature Neuroscience* 7: 327–332.

circuit (Figure 2). The refinement of the arbor structure occurs in response to signals from the environment and from other neurons in the circuit. These observations provide critical insight into the cellular mechanisms governing dendritic and axonal arbor development and clearly indicate that learning and recovery from trauma can occur even in the adult brain by tapping into the cellular mechanisms that shape the neuronal structure during development of the brain.

Synaptic Inputs Increase Dendrite Arbor Growth

Several time-lapse imaging experiments now provide convincing support for the hypothesis that the formation and maturation of synaptic contacts stabilize dendritic arbor structure (Figure 3). Direct *in vivo* time-lapse imaging reveals that visual stimulation increases the growth of optic tectal dendritic arbors *in vivo* by promoting the stabilization of newly added branches. Pharmacological blockade of either AMPA- or NMDA-type glutamate receptors decreases dendritic arbor growth *in vivo* and blocks visual stimulation-induced dendritic arbor growth. Similarly, deafferentation or blocking inputs within the auditory system, as well as other sensory systems, has severe effects on dendritic arbor development of neurons that receive and process sensory information. *In vivo* imaging experiments in which the maturation of glutamatergic synapses was blocked show that AMPAR trafficking into developing synapses is required for the stabilization of newly added dendritic arbor branches and the cumulative elaboration of the complex dendritic arbor. Together, these data indicate that the iterative process of dendritic arbor development requires the coordinate formation and stabilization of glutamatergic synaptic inputs (Figures 1–3).

Conclusion

Synapse formation is characterized by the assembly of a complex protein machine that spans a specialized junction, the synapse, which forms between two neurons. Recent multidisciplinary experiments combining time-lapse *in vivo* imaging, molecular manipulations, and electrophysiological recordings demonstrate that synapse formation and maturation are required for the normal development of neuronal structure, including the axon and dendrite. These experiments have also shown that the development of axons and dendrites occur as a result of simultaneous addition and retraction of branches within the neuron. These observations

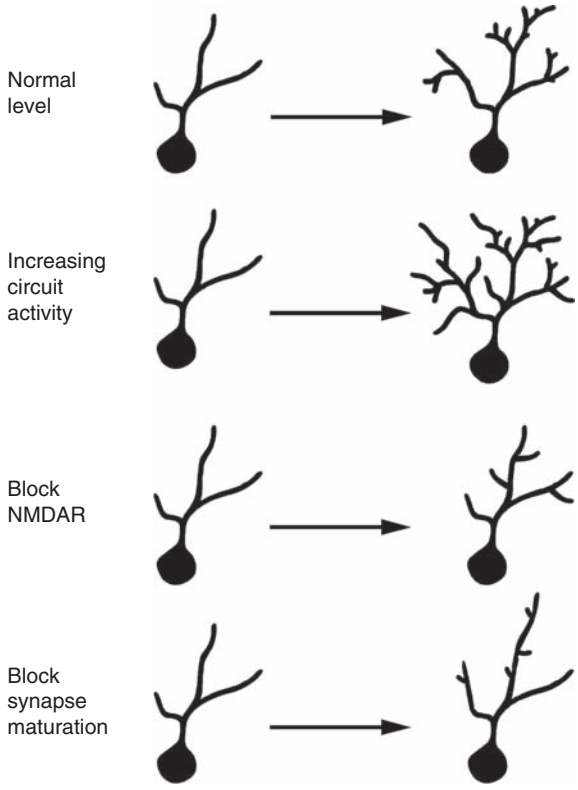

Figure 3 Synaptic input drives dendrite development. The normal rate of dendritic arbor development is increased by enhanced synaptic input and is decreased when glutamate receptor activity or synapse maturation is blocked. The extent of dendritic arbor growth under 'normal' levels of circuit activity in the brain is depicted at the top. The effect of increasing activity on dendritic arbor development is shown in the second row; the effects of blocking *N*-methyl-D-aspartate receptor activity or blocking synapse maturation are shown in the third and fourth rows. Based on data from Sin WC, Haas K, Ruthazer ES, et al. (2002) Dendrite growth increased by visual activity requires NMDA receptor and Rho GTPases. *Nature* 419: 475–480; Rajan I and Cline HT (1998) Glutamate receptor activity is required for normal development of tectal cell dendrites *in vivo*. *Journal of Neuroscience* 18: 7836–7846; and Haas K, Li J, and Cline HT (2006) AMPA receptors regulate experience-dependent dendritic arbor growth *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* 103: 12127–12131.

overturn two previous ideas about brain development, that neuronal growth can occur in the absence of synapse formation and synaptic activity, and that neurons undergo a phase of exuberant growth followed by a separate period of pruning. This modern view of neuron development suggests that brain activity will increase synapse formation as well as the establishment and maintenance of optimal neuronal circuits. This view also lends hope to the idea that brain exercises will aid recovery from trauma even in adults.

See also: Cell Adhesion Molecules at Synapses; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin; Interstitial Axon Branching/Collateral Elimination; Postsynaptic Development: Neuronal Molecular Scaffolds; Presynaptic Development and Active Zones; Receptor Trafficking; Retrograde Transsynaptic Influences; Spine Plasticity; Spontaneous Patterned Activity in Developing Neural Circuits; Synaptic Precursors: Filopodia; Topographic Maps: Molecular Mechanisms.

Further Reading

Cline HT (2001) Dendritic arbor development and synaptogenesis. *Current Opinion in Neurobiology* 11: 118–126.

Haas K, Li J, and Cline HT (2006) AMPA receptors regulate experience-dependent dendritic arbor growth *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* 103: 12127–12131.

Hua JY and Smith SJ (2004) Neural activity and the dynamics of central nervous system development. *Nature Neuroscience* 7: 327–332

Niell CM, Meyer MP, and Smith SJ (2004) *In vivo* imaging of synapse formation on a growing dendritic arbor. *Nature Neuroscience* 7: 254–260.

Rajan I and Cline HT (1998) Glutamate receptor activity is required for normal development of tectal cell dendrites *in vivo*. *Journal of Neuroscience* 18: 7836–7846

Rumpel S, LeDoux J, Zador A, et al. (2005) Postsynaptic receptor trafficking underlying a form of associative learning. *Science* 308: 83–88.

Sin WC, Haas K, Ruthazer ES, et al. (2002) Dendrite growth increased by visual activity requires NMDA receptor and Rho GTPases. *Nature* 419: 475–480.

Takahashi T, Svoboda K, and Malinow R (2003) Experience strengthening transmission by driving AMPA receptors into synapses. *Science* 299: 1585–1588.

Van Aelst L and Cline HT (2004) Rho GTPases and activity-dependent dendrite development. *Current Opinion in Neurobiology* 14: 297–304.

Wong RO and Ghosh A (2002) Activity-dependent regulation of dendritic growth and patterning. *Nature Reviews Neuroscience* 3: 803–812.

Wu G-Y and Cline HT (1998) Stabilization of dendritic arbor structure *in vivo* by CaMKII. *Science* 279: 222–226.

Dendrites: Localized Translation

H Wang, The Rockefeller University, New York, NY, USA
H Tiedge, State University of New York Health Science Center at Brooklyn, Brooklyn, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In eukaryotic cells, translational control of gene expression is an important means in the long-term regulation of cellular structure and function. For complex and highly elongated cells such as neurons, translational control may be at least as significant as transcriptional control in the nucleus. In such cells, reliance on translational control is a consequence of their complex geometry which demands that gene expression be regulated at sites far away from the DNA in the nucleus and in a manner that is specific to the dynamic requirements of numerous and mosaic local sites. On-site translational control mechanisms at the synapse are thus well suited for the long-term, input-specific management of local protein repertoires in response to physiological stimuli. They are now thought to constitute one of the fundamental pathways that underlie neuronal plasticity and, consequently, various higher brain functions. We present a synopsis of the current state of the field, highlighting past achievements as well as key areas of current investigation.

Translational Components in Dendrites

Since the discovery of dendritically localized polyribosomes more than 20 years ago, evidence supporting the concept of localized translation in dendrites has been steadily increasing. Such evidence includes the identification in dendrites of components of translational machinery, posttranslational organelles, and diverse types of RNAs.

Translational Machinery

Polyribosomes are located beneath postsynaptic sites at the base of dendritic spines. Polyribosomes are redistributed from dendritic shafts to spines after induction of long-term potentiation (LTP), and spines with polyribosomes have larger synapses, indicating that local protein synthesis may serve to stabilize stimulation-induced postsynaptic growth. In addition to polyribosomes, other components of the translational machinery have been identified in dendrites. Such components include eukaryotic initiation factor (eIF) 2α, eIF4E, eIF4G, eIF4A, eukaryotic elongatiofactor (eEF) 2, and poly(A) binding protein (PABP). Transfer RNA (tRNA) and aminoacyl-tRNA synthetase are also localized to dendrites.

Posttranslational Organelles

Organelles that participate in cotranslational and posttranslational processing – that is, the rough endoplasmic reticulum and Golgi apparatus – are located in dendrites. The entire endoplasmic reticulum-to-Golgi secretory pathway has been delineated in dendrites. Newly synthesized membrane or secretory proteins are sorted locally to associate with dendritic Golgi outposts.

RNAs

In 1988, microtubule-associated protein 2 (MAP2) mRNA became the first transcript to be detected in dendrites. Others soon followed. Proteins encoded by dendritic mRNAs belong to a variety of functional groups. One important group includes receptors such as the inositol 1,4,5-trisphosphate (InsP3) receptor type 1, several glutamate receptor subtypes, glycine receptors, and TrkB as well as its ligand, brain-derived neurotrophic factor (BDNF). Another group of proteins encoded by dendritic mRNAs includes cytoskeleton-associated proteins such as MAP2 and activity-regulated cytoskeleton-associated protein (ARC/ARG 3.1). Kinase mRNAs have also been reported in dendrites, including those for the α subunit of Ca^{2+}/calmodulin-dependent protein kinase II (CaMKIIα) and for protein kinase Mζ. It is estimated that the number of dendritically localized mRNA species may be several hundred.

In addition to mRNAs, untranslated RNAs are also localized to dendrites. Such RNAs include ribosomal RNAs (rRNAs), tRNAs, microRNAs (miRNAs), and small brain-specific cytoplasmic RNAs (BC RNAs) such as BC1 and BC200. As discussed later, some of these small RNAs have been implicated in translational control mechanisms at the synapse.

RNA Transport in Dendrites

The targeted delivery of select RNAs to synaptodendritic domains is a functional prerequisite for local protein synthesis at the synapse. A general picture has been emerging, from work with neurons as well as with other eukaryotic cell types, that targeting *trans*-acting factors (TAFs) recognize *cis*-acting dendritic targeting elements (DTEs) in dendritic RNAs. DTEs contain spatial codes that are recognized by TAFs:

Such decoding results in the assembly of RNP complexes which in turn are propelled along cytoskeletal elements by molecular motors.

Local Protein Synthesis in Dendrites

The dendritic location of translational machinery and the active transport of specific RNAs to dendritic destination sites are indicative that proteins are locally synthesized in dendrites. This concept is well supported by experimental evidence, and mechanisms of local translational control are being investigated.

Activity-Dependent Local Translation

Dendritic proteins may be locally synthesized in response to various synaptic activations. Activation of glutamate receptors appears to regulate dendritic translation. For example, application of the metabotropic glutamate receptor (mGluR) agonist (RS)-3, 5-dihydroxyphenylglycine (DHPG) increases local levels of fragile X mental retardation protein (FMRP) by twofold within 5 min. Blockade of *N*-methyl-D-aspartate receptor (NMDAR) activation with antagonist 3-(2 carboxypiperazin-4yl) propyl-1-phosphonic acid prevents the increase in CaMKIIα protein levels in synaptic fractions from dark-reared, light-exposed animals. Tetanic stimulation of the Schaffer collateral pathway, which induces NMDAR-dependent LTP, increases autophosphorylation of CaMKIIα both in cell bodies and in apical dendrites, whereas total CaMKIIα levels increase only in apical dendrites.

Synaptic stimulation by neurotrophins such as BDNF and NT-3 has been suggested to upregulate dendritic synthesis of specific proteins. Local protein synthesis is required for neurotrophin-induced synaptic plasticity in CA1 dendritic regions. BDNF was also shown to induce translation of Arc mRNA in synaptosome preparations as well as dendritic translation of a GFP reporter mRNA that contained both 5′ and 3′ untranslated regions (UTRs) of CaMKIIα mRNA.

Mechanisms of Translational Control at the Synapse

Translational control mechanisms that have been implicated in local synaptic regulation can broadly be categorized into three groups: mechanisms that target the initiation step, those that target the elongation step, and those that operate at the RNA level. We discuss each category in turn (Figure 1).

Translation initiation Translation is mechanistically subdivided into the three steps of initiation, elongation, and termination. Initiation is often rate limiting and the target of regulation. Among translation initiation factors, eIF4F plays a strategic role in the recruitment of 43S preinitiation complexes to mRNAs. eIF4F is a heterotrimeric complex that consists of three subunits: eIF4E, a cap binding protein; eIF4G, a scaffolding protein providing the binding sites for eIF4A, eIF3, and PABP; and eIF4A, an ATP-dependent RNA helicase. eIF4E serves as a target for translational control in several neuronal signaling pathways. BDNF and the mammalian target of rapamycin (mTOR) pathway have been implicated in regulating eIF4E. mTOR is a serine/threonine kinase that phosphorylates eIF4E binding proteins (eIF4E-BPs), inducing the subsequent dissociation of eIF4E-BPs from eIF4E and thereby activating translation initiation. BDNF-induced synaptic potentiation in hippocampal slices is inhibited by application of mTOR inhibitor rapamycin. In addition, in cortical neurons and other systems, application of BDNF induces phosphorylation of eIF4E, eIF4E-BP, and mTOR. Work with gene arrays has identified a subset of mRNAs the translation of which increases upon exposure to BDNF. Among these mRNAs, postsynaptic density scaffolding protein Homer 2 has been linked to the BDNF-activated mTOR pathway. The previous data suggest that the mTOR pathway is essential for BDNF to activate translation in dendrites. Independently, the mitogen-activated protein kinase (MAPK) signaling pathway has also been linked to eIF4E. Phosphorylation of MAPK/extracellular signal-regulated kinase kinase activates its downstream signaling targets, resulting in the phosphorylation of eIF4E and in the activation of local translation. The combined data suggest coordination between these two pathways and raise the possibility of multilevel translational regulation in dendrites.

Work with *Drosophila* has shown that eIF4E mRNA may be a target of translational control in postsynaptic microdomains. Pumilio, an important factor for the anterior–posterior patterning during early *Drosophila* embryogenesis, has been suggested to inhibit synthesis of eIF4E at the *Drosophila* neuromuscular junction through interactions with the 3′ UTR of eIF4E mRNA.

Translation initiation factor PABP interacts with eIF4F and enhances initiation of both cap-dependent and encephalomyocarditis virus (EMCV)-type internal ribosome entry site (IRES)-mediated translation. Dendritic BC1 RNA regulates translation initiation by interacting with eIF4A and PABP, repressing formation of the 48S translation initiation complex. BC1 RNA represses cap-dependent and EMCV–IRES-mediated initiation. Because BC1 RNA targets unwinding factor eIF4A, mRNAs with highly structured 5′ UTRs would be particularly sensitive to BC1-mediated translational control.

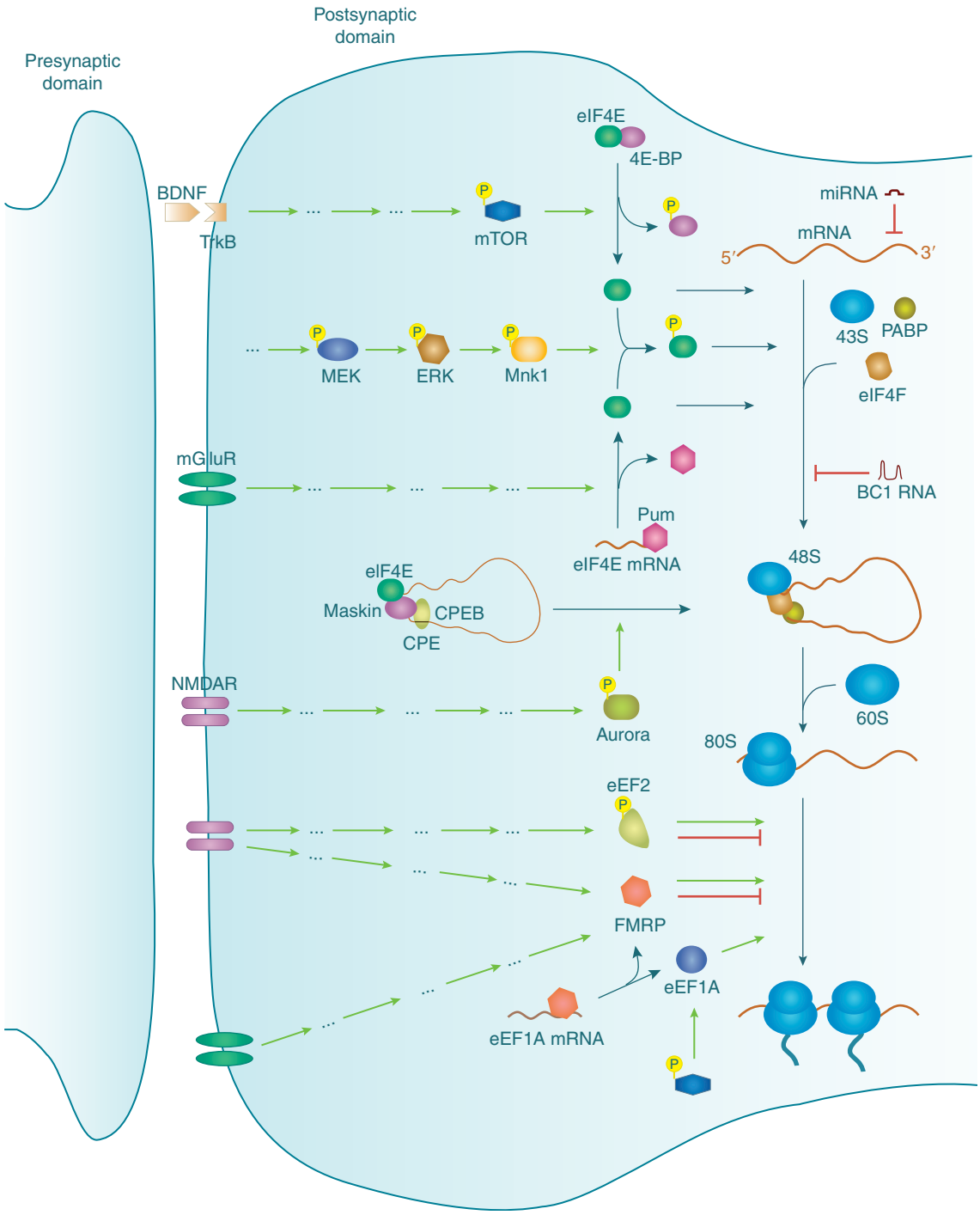

Figure 1 Schematic illustration of translational control pathways in dendrites. Adapted from Kindler S, Wang H, Richter D, and Tiedge H (2005) RNA transport and local control of translation. *Annual Review of Cell and Developmental Biology* 21: 223–245, with permission.

Cytoplasmic polyadenylation is another translational control mechanism that may operate at the synapse. As was first reported with *Xenopus* oocytes, some mRNAs have short poly(A) tails and remain translationally dormant; translation is activated only subsequent to their cytoplasmic polyadenylation. Polyadenylation is initiated through phosphorylation of CPEB by the kinase aurora. Phosphorylated CPEB binds to CPEs that are contained in 3′ UTRs, and it recruits poly(A) polymerase to extend mRNA poly(A) tails. The result is a dissociation of maskin from eIF4E, allowing eIF4E to interact with PABP through eIF4G and thus to initiate translation. Translation of dendritic CaMKIIα mRNA has been suggested to

involve the cytoplasmic polyadenylation mechanism. Two CPE-like sequences have been localized to the 3′ UTR of CaMKIIα mRNA. In addition, factors involved in the polyadenylation mechanism, such as aurora, CPEB, poly(A) polymerase, but not maskin, have been identified at neuronal synapses. NMDAR activation is required for the induction of aurora phosphorylation and thus for the initiation of dendritic CaMKIIα mRNA translation. Similarly, the translation of dendritic plasminogen activator (tPA) mRNA has also been reported to depend on the polyadenylation mechanism: After mGluR simulation, tPA mRNA is rapidly polyadenylated and translated into its cognate protein.

Translation elongation Elongation may also be the target of local translational control in neurons. Phosphorylation of eEF2 can be induced by visual stimulation and is dependent on the activation of NMDARs in the amphibian tectum. Overall protein synthesis is reduced upon phosphorylation of eEF2, although individual neuronal mRNAs may be differentially receptive to such elongation control. FMRP has also been suggested to regulate translation at the level of elongation. Unphosphorylated FMRP is associated with actively translating polyribosomes, whereas a fraction of phosphorylated FMRP is associated with translationally inactive polyribosomes.

Elongation factors can be targets of regulation as well. Upon induction of hippocampal LTP, the protein levels of eEF1A increase within 5 min. Application of rapamycin blocks this effect, indicating that the previously discussed mTOR pathway is involved in this regulation. The expression of eEF1A has also been shown to be increased after DHPG-induced LTD or after induction of LTP by high-frequency stimulation of the perforant path; however, only the former appears to be sensitive to protein synthesis inhibitors.

RNA level Several translational control mechanisms operate at the RNA level. For example, in addition to the aforementioned mechanism, FMRP has been suggested to repress protein synthesis by binding to mRNAs, and activation of both mGluRs and NMDARs is important for the FMRP pathway at the synapse. One of these targets is eEF1A mRNA because FMRP represses its translation by direct binding. Independently, hippocampal mGluR-dependent long-term depression (LTD) triggers rapid translation of FMRP-interacting mRNAs. In addition, mGluR-dependent LTD also induces a transient, translation-dependent increase in FMRP; increased FMRP is in turn rapidly degraded through the ubiquitin–proteasome pathway. These data implicate yet another mechanism in the management of local protein repertoires: protein degradation.

miRNAs have also been associated with dendritic translational control. miRNAs are small RNAs (21–24 nt) that repress translation through partial base pairing with their target mRNAs. miRNAs have been found enriched in polyribosome fractions in mammalian brain preparations. Work with *Drosophila* neurons suggests that CaMKIIα contains two putative miRNA binding sites, indicating the possibility of translation repression through miRNA interaction. Components of the RNA-induced silencing complex have been found at the synapse, where one of them, SDE3 helicase Armitage, degrades during learning. Brain-specific miRNA-134 inhibits the translation of Limk1, a protein kinase that controls dendritic spine development, thus negatively regulating spine size. This effect can be alleviated by exposure to extracellular stimuli such as BDNF. In addition, FMRP has been proposed to interact with miRNAs as well as components of the miRNA pathway, including Dicer and the mammalian ortholog of Argonaute 1 (AGO1).

Translational Repression during Transport

Although the coordination of mRNA transport with translational repression has been well characterized in other systems such as *Drosophila* oocytes and early embryos, only recently have the mechanisms of transport-coupled translational repression been described in more detail in neuronal systems. ZBP1, the aforementioned β-actin mRNA TAF, has been shown to play an additional role as a translational repressor during transport. In a neuroblastoma cell line and in cultured hippocampal neurons, ZBP1 associates with β-actin mRNA and prevents its translation at the initiation step; when ZBP1–RNA complexes reach their destination, translation of β-actin mRNA is activated through phosphorylation of a key tyrosine residue in ZBP1 by protein kinase Src. These data provide clear evidence of a direct link between translational repression and RNA transport in neurons.

A similar coordination between RNA transport and translational repression has been discussed for transport RNPs. CPEB, shown to facilitate dendritic transport of CPE-containing mRNAs, is co-localized with the translational repressor maskin during transport. Neuronal RNP particles have also been found to contain translational repressors FMRP and RNA granule protein 105. These data suggest the possibility of additional functional links between RNA transport and translational repression.

Perspective

Local translational control of gene expression in synaptodendritic microdomains has been firmly established as a key mechanism in the long-term regulation of synaptic form and function. By transferring decision-making authority to the synapse, this mechanism in essence decentralizes control of gene expression, thus fulfilling a *conditio sine qua non* for experience-dependent, input-specific long-term modulation of synaptic strength.

Future efforts will certainly continue to address the two most fundamental open questions in local translational control. First, we urgently need to understand the spatial codes that RNAs use to specify their synaptodendritic destination sites. Such codes are most likely expressed as three-dimensional architectural RNA motifs. In contrast to linear information coding in DNA, RNA information coding is, except for open reading frames, often three-dimensional. It will be important to be able to read such codes and to understand how they are being decoded by targeting TAFs. Second, despite much progress in recent years, translational control mechanisms at the synapse remain poorly understood. Thus, although it has become clear that most, if not all, local translational modulators are in fact repressors, little is known about how local translation is de-repressed in response to appropriate input stimuli.

It will also be important to learn which proteins are locally synthesized upon stimulation – that is, how mRNAs are selected for local translation in a given synaptodendritic microdomain. Are such decisions implemented through selective translational activation (i.e., de-repression) of only a few mRNAs out of a larger locally available subset, or is this selection made at an earlier stage (i.e., through the selective recruitment of only a few specific mRNAs to that microdomain)? These two options are not mutually exclusive, and specificity may thus be achieved at the level of selective mRNA availability at local sites, at the level of selective local translational activation/de-repression, or both. Given Nature's resourcefulness, it would indeed be surprising if examples of either mode of regulation would not be identified in due course.

See also: Axonal mRNA Transport and Functions; Axonal and Dendritic Identity and Structure: Control of; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Translational Regulation at the Synapse.

Further Reading

Cao X, Yeo G, Muotri AR, Kuwabara T, and Gage FH (2006) Noncoding RNAs in the mammalian central nervous system. *Annual Review of Neuroscience* 29: 77–103.

Garner CC, Tucker RP, and Matus A (1988) Selective localization of messenger RNA for cytoskeletal protein MAP2 in dendrites. *Nature* 336: 674–677.

Horton AC and Ehlers MD (2003) Dual modes of endoplasmic reticulum-to-Golgi transport in dendrites revealed by live-cell imaging. *Journal of Neuroscience* 23: 6188–6199.

Huang YS, Jung MY, Sarkissian M, and Richter JD (2002) *N*-methyl-D-aspartate receptor signaling results in Aurora kinase-catalyzed CPEB phosphorylation and αCaMKII mRNA polyadenylation at synapses. *EMBO Journal* 21: 2139–2148.

Hüttelmaier S, Zenklusen D, Lederer M, et al. (2005) Spatial regulation of beta-actin translation by Src-dependent phosphorylation of ZBP1. *Nature* 438: 512–515.

Ishizuka A, Siomi MC, and Siomi H (2002) A *Drosophila* fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes and Development* 16: 2497–2508.

Job C and Eberwine J (2001) Localization and translation of mRNA in dendrites and axons. *Nature Reviews Neuroscience* 2: 889–898.

Kang H and Schuman EM (1996) A requirement for local protein synthesis in neurotrophin-induced hippocampal synaptic plasticity. *Science* 273: 1402–1406.

Kapp LD and Lorsch JR (2004) The molecular mechanics of eukaryotic translation. *Annual Review of Biochemistry* 73: 657–704.

Kelleher RJ, Govindarajan A, Jung HY, Kang H, and Tonegawa S (2004) Translational control by MAPK signaling in long-term synaptic plasticity and memory. *Cell* 116: 467–479.

Kindler S, Wang H, Richter D, and Tiedge H (2005) RNA transport and local control of translation. *Annual Review of Cell and Developmental Biology* 21: 223–245.

Martin KC (2004) Local protein synthesis during axon guidance and synaptic plasticity. *Current Opinion in Neurobiology* 14: 305–310.

Richter JD and Lorenz LJ (2002) Selective translation of mRNAs at synapses. *Current Opinion in Neurobiology* 12: 300–304.

Steward O and Levy WB (1982) Preferential localization of polyribosomes under the base of dendritic spines in granule cells of the dentate gyrus. *Journal of Neuroscience* 2: 284–291.

Steward O and Schuman EM (2001) Protein synthesis at synaptic sites on dendrites. *Annual Review of Neuroscience* 24: 299–325.

Tiedge H (2005) RNA reigns in neurons. *Neuron* 48: 13–16.

Wells DG, Dong X, Quinlan EM, et al. (2001) A role for the cytoplasmic polyadenylation element in NMDA receptor-regulated mRNA translation in neurons. *Journal of Neuroscience* 21: 9541–9548.

Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression

J B Dictenberg, Hunter College, New York, NY, USA
R H Singer, Albert Einstein College of Medicine, Bronx, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Dendrites are master integrators of information flow in the brain and the sites of postsynaptic structural modifications that regulate the plasticity of synapse strength in response to neurotransmission. Recent data suggest that dendrites are also an important site for the dynamic regulation of localized gene expression in the nervous system, with subsets of synapses possessing the ability to independently alter synapse strength through the local synthesis of proteins. An important component to this regulation is the active transport of a select group of messenger RNAs (mRNAs) from the cell body into dendrites and targeting of these mRNAs to specific translation sites. Since several long-lasting forms of plasticity require protein synthesis, synaptic stimulation appears to be an important mechanism for the regulation of mRNA trafficking within dendrites.

Discovery of Dendritic mRNA Transport

Early speculations about the nature of synaptic connections led to the idea that cytoskeletal proteins played an important structural role in neuronal function, and the identification of translational machinery in dendrites but not axons by ultrastructural methods and actin protein localized to dendritic spines spawned an era of investigation into the nature of localized gene expression that is still ongoing. As it was found that microtubule-associated protein 2 (MAP2) was expressed selectively in the dendritic compartment, markers for neuronal polarity became the subject of great curiosity. Subsequent work tracked total mRNA movements in cultured neurons by pulse labeling with radioactive nucleotides to visualize time-dependent migration of mRNA into dendrites and showed the asymmetric localization of cytoskeletal mRNAs to dendrites, such as MAP2. More recently, eloquent molecular approaches have demonstrated the presence of numerous diverse mRNAs by directly aspirating cytoplasm from transected dendrites and identification by polymerase chain reaction. These studies served as the foundation for the idea that many regulatory and cytoskeletal mRNAs may be selectively transported into neuronal dendrites on microtubules and subsequently activated for translation at synapses in response to neuronal stimulation, an idea that currently has enormous support and has led to breakthroughs in visualizing specific transcript motility in living cells.

Structure and Composition of Dendritic Transport mRNPs

mRNAs are transported into dendrites in the form of large, electron-dense RNA- and protein-containing granules. Since these transport granules are distinct from stress granules, they will be referred to as transport messenger ribonucleoproteins (mRNPs). Neuronal cytoplasmic transport mRNPs appear quite heterogeneous in size and have been estimated by ultrastructural methods to be between 200 and 600 μm in diameter and greater in mass than polyribosome complexes (>500 S). They are formed initially when nucleocytoplasmic shuttling RNA-binding proteins (RBPs) associate with nascent transcripts in the nucleus to prevent premature ribosomal loading and translation and to aid in nuclear export of mRNAs. While the composition of proteins in transport mRNPs may change on exit to the cytoplasm, exchanging factors specific to export, for example, several RBPs appear to remain associated with transcripts until they are delivered to distal dendritic sites for translation. Numerous biochemical and proteomic experiments suggests that these cytoplasmic neuronal transport mRNPs are not homogeneous in composition, containing many different RBPs, such as Staufen, zipcode-binding protein 1 (ZBP1), heteronuclear RNP-A2 (hnRNP-A2), Pur-alpha and the fragile X mental retardation protein (FMRP), as well as many mRNA species, such as calcium/calmodulin-dependent protein kinase II (CaMKII)-alpha, activity-regulated cytoskeletal-associated (Arc), beta-actin, and the noncoding BC1. Included in the hundreds of diverse mRNAs that have been isolated from dendrites biochemically are several subunits of glutamate receptor subtypes and translational machinery components, such as ribosome subunits and elongation factor 1a. In addition, transport mRNPs contain multiple cytoskeletal motors such as conventional kinesin (Kif5) and dynein, which contribute to plus end- and minus end-directed transport along microtubules within dendrites. While the protein content may be diverse, the RBPs in transport mRNPs all appear to be important for repression of translation of mRNAs until reaching their distal site, where they are derepressed by regulatory signals. Both biochemical work and ultrastructural work have shown that they contain many densely packed ribosomes but are not translationally

competent as they lack proteins essential for initiation (e.g., eIF4). It is interesting to note that these mRNPs are dramatically remodeled on neuronal depolarization, presumably releasing ribosomes and mRNAs into local active translational complexes. The composition of transport mRNPs suggests that they represent motile pretranslation units that undergo dynamic remodeling in response to signals that induce their dendritic transport and subsequent translation.

Sequence and Structural Characteristics of Localization Signals in Neuronal Transport mRNPs

A unifying theme in the analysis of many localized mRNAs is the requirement for a *cis*-acting sequence for targeting the transcript to specific sites, or 'zipcodes.' These zipcodes are mostly localized to the 3′ untranslated regions (UTRs) of transcripts, where they are bound by *trans*-acting RBPs involved in the localization process (see Table 1). The zipcodes may contain several units that function in a multistep pathway to mRNA localization. One example is the localization of myelin basic protein (MBP) mRNA to the myelin-rich oligodendrocyte compartment. This is a two-step process involving an 11 nt element (the RNA transport signal, RTS) sufficient to direct a reporter mRNA into processes and a distinct 342 nt element (the RNA localization region) required for MBP mRNA targeting to or retention at the myelin-specific subcompartment in oligodendrocytes. The 11 nt RTS has been shown to mediate granule formation and dendritic trafficking with hnRNP-A2 in primary hippocampal cultures as well; however, a requirement for this 11 nt sequence in the transport of the native MBP mRNA has not been established. It is interesting to note that a conserved version of the RTS has also been found in several other dendritic mRNAs, such as MAP2, CaMKII-alpha, Arc, and γ-aminobutyric acid-A-receptor subunit and therefore may play an important role in mRNA targeting using similar machinery in neurons. However, the fact that a 640 nt element of MAP2 mRNA 3′UTR was found to be necessary for dendritic targeting suggests that there may be additional required elements present in the 3′UTR that control localization or stability. Another example of a bipartite element is the beta-actin mRNA, which contains both a 54 nt and a 43 nt zipcode, required for efficient localization to neuronal processes and growth cones. While some zipcodes in localized mRNAs such as beta-actin are necessary for localization of the native mRNAs, most zipcodes have been tested using reporter mRNAs, and therefore their requirement in the localization of native mRNAs is not known. Future work using genetic deletions or antisense technology can more directly address these shortcomings.

Some mRNAs, such as *bicoid* in *Drosophila*, may contain zipcode elements that are repeated several times in the 3′UTR to effect a more efficient localization as the deletion of one or more redundant elements is not sufficient to completely abolish targeting but rather diminishes the amount of mRNA localized over time. This is a feature shared by the beta-actin mRNA 3′UTR zipcode, which has a consensus sequence of ACACCC repeated twice within one zipcode element (the 54 nt chicken zipcode). Although some zipcodes may be conserved between mRNAs, there is no clear indication yet that the same proteins are involved in their localization. Rather diverse collections of RBPs have been found to associate with transcripts that have conserved zipcodes. This has inspired a thorough analysis of secondary structural features in zipcodes that may be more important than primary sequence conservation

Table 1 RNA-binding proteins and their corresponding *cis*-acting sequences in brain

RNA-binding protein	*mRNA*	*Sequence and location*	*Required for localization*[a]
ZBP1	beta-actin	ACACCC, within 54 nt zipcode in 3′UTR	+
hnRNPA2	MBP	GCCAAGGAGCC, 3′UTR	
	MAP2	GCCAAGGAGUC, coding	
CPEB	CaMKII-alpha	UUUUUUAUU X 2 (separated by 82 nt), 3′UTR	
HuD	tau	240 nt, 3′UTR	+
MARTA1/2	MAP2	640 nt, 3′UTR	
Unknown	CaMKII-alpha	3.0 kb, 3′UTR	+/−
		94 nt, 3′UTR	

[a]Plus sign (+) indicates *cis*-acting sequences shown to be required for full-length (either endogenous or exogenous) mRNA localization in dendrites either by deletion of element or using antisense to the endogenous element; no entry symbolizes that the *cis*-acting elements were tested only for localization of a heterologous reporter mRNA; +/− indicates conflicting reports for this mRNA (one study indicated that this 94 nt sequence can localize a heterologous reporter to dendrites (−), and another showed significantly diminished dendritic localization of this endogenous mRNA with a 3.0 kb deletion in the 3′UTR that left the 94 nt sequence intact (+)). CaMKII-alpha, calcium/calmodulin-dependent protein kinase II; CPEB, cytoplasmic polyadenylation element binding protein; hnRNPA2, heteronuclear ribonucleoprotein-A2; HuD, one of the embryonic lethal abnormal vision family of RNA-binding proteins; MAP2, microtubule-associated protein 2; MARTA, MAP2-RNA *trans*-acting protein; MBP, myelin basic protein; UTR, untranslated region; ZBP1, zipcode-binding protein 1.

since RNA can assume diverse three-dimensional structures and the 3′UTRs can possess structure without interfering with translation. Indeed many zipcodes appear to contain stem–loop structures where the essential sequences present binding sites for RBPs, as has been shown for ZBP1 binding to the 54 nt zipcode in beta-actin mRNA and for Staufen binding to the 3′UTR of *bicoid* mRNA. It is important to note that individual nucleotide changes that conserve the stem hydrogen bonding structure do not alter binding to the *trans*-acting RBPs or localization of the mRNAs whereas nucleotides that do disrupt the structure diminish binding and localization. A more thorough analysis of structural motifs common to localized mRNAs, may reveal a class of conserved structures present within mRNAs that share *trans*-acting RBP recognition sites. Indeed more-detailed structural information may explain how the dendritic targeting element (DTE) of MAP2 mRNA is recognized by the diverse RBPs Staufen, hnRNP-A2, MAP2-RNA *trans*-acting protein (MARTA)1, and MARTA2, or how MARTA1 recognizes both the DTE and the zipcode of beta-actin mRNA, which show little primary sequence homology.

Trans-Acting Factors Involved in Dendritic mRNA Transport

The persistent association of RBPs with mRNAs from birth at the transcription site until death after mRNA translation and degradation suggests that RBPs are critical to the spatiotemporal regulation of mRNA expression. At least 300 RBPs are expressed in vertebrate brain during early developmental stages, defined by the presence of conserved RNA-binding motifs, such as the K-homology domain of hnRNP-K family of RBPs. Other common RNA-binding motifs include the RGG-box, the RNA-recognition motif, and the double-stranded RNA-binding domain. These RBPs usually have nuclear shuttling domains (nuclear export and import) to facilitate entry and exit from the nucleus, where they first associate with transcripts in their journey to distal sites for regulated translation. An important theme is that the mRNA must be translationally repressed by these RBPs to facilitate temporally and spatially localized gene expression and that this repression must be linked to the process of mRNA transport. Therefore RBPs may function to repress translation and initiate assembly of the transport-competent higher-ordered mRNP granules before being recognized by the transport machinery. This idea is supported by the observation that exogenous MBP mRNA injected into oligodendrocytes induces the formation of larger RNP particles before their transport into processes. Several RBPs involved in the regulation of mRNA transport and translation in neurons, such as cytoplasmic polyadenylation element binding protein 1 (CPEB1), ZBP1, and Staufen, have been implicated in this dual functional role already.

While a causal role for RBPs in the transport and translational regulation of dendritic mRNAs appears concrete, there is surprisingly little direct evidence for an essential role of these proteins in RNA localization (see Table 1), most likely because many of these proteins may play an essential role in early development and therefore knockout animals may not be viable. However, a viable knockout mouse has been generated for CPEB1, and neurons from these mutant mice have shown reduced dendritic localization of a reporter mRNA containing cytoplasmic polyadenylation elements (CPEs) within the 3′UTR. The absence of data demonstrating a reduction in the dendritic localization of endogenous mRNAs containing CPEs suggests that several RBPs, such as the other recently added members to the CPEB family of RBPs, may play a role in targeting different mRNAs containing CPEs to dendrites along with CPEB1. Alternatively, distinct RBPs may solely regulate specific stimulus-induced pathways for mRNA localization, as has been shown for the neurotrophin-induced localization of beta-actin mRNA to neurites and growth cones. This localization requires the beta-actin mRNA zipcode recognition by ZBP1 because uncoupling of the mRNA from ZBP1 using antisense oligonucleotides that are complementary to the zipcode has caused a reduced localization of beta-actin mRNA by this stimulation paradigm. Other proposed functions for RBPs in dendritic transport include a role for connection of the mRNA to molecular motors, such as that proposed for CPEB. The trafficking of MBP mRNA by hnRNP-A2 in oligodendrocytes and CaMKII-alpha mRNA in hippocampal neurons was reduced by antisense oligonucleotide knockdown of conventional kinesin (Kif5). Recent proteomic approaches have identified numerous RBPs in complex with this form of kinesin from brain extracts, including proteins already implicated in RNA localization: Staufen, Pur-alpha, and hnRNP-U. However, the molecular evidence for a connection to kinesin is still lacking, and further work is needed to understand the mechanism behind this connection, as well as connections to the minus end-directed microtubule motor dynein that contributes to the oscillatory and retrograde trajectories of transport mRNPs.

Dynamic Regulation of Transport mRNP Motility

Early studies in the transport of mRNAs to dendrites showed that on synaptic stimulation, immediate early genes such as Arc are turned on, and the mRNAs are rapidly transported into dendrites. This transport

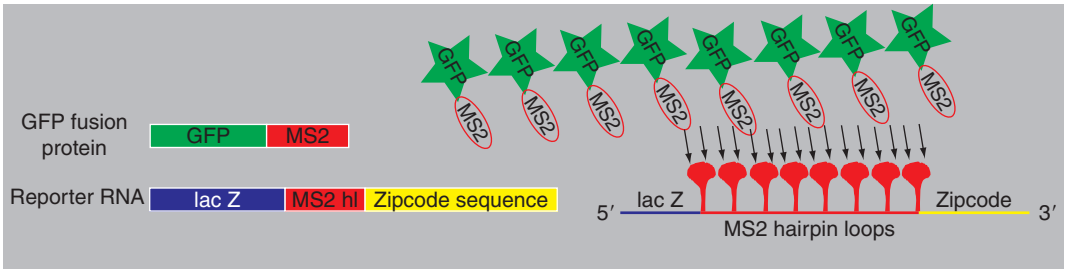

a

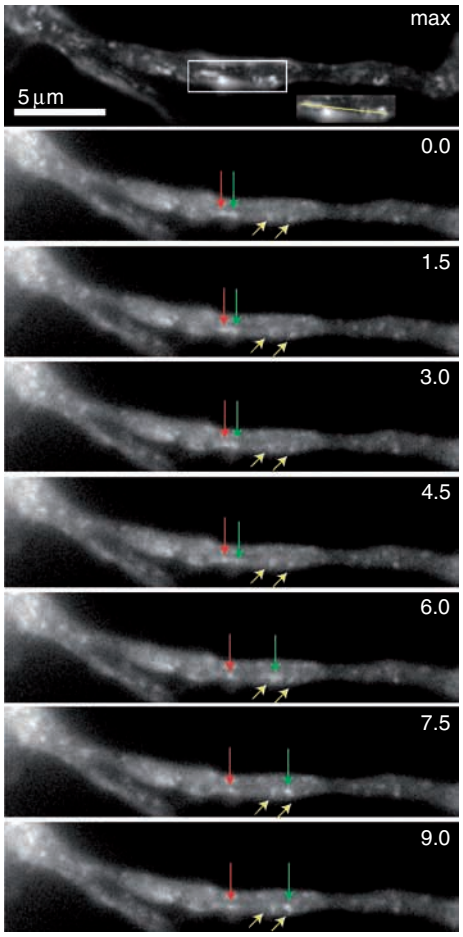

b

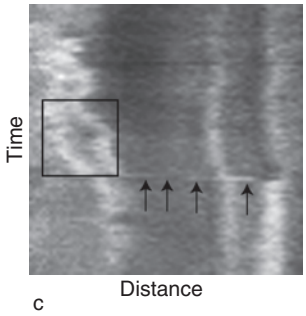

c

Figure 1 Analysis of rapid dendritic transport of calcium/calmodulin-dependent protein kinase II (CaMKII)-alpha messenger RNA (mRNA) in living hippocampal neurons. Cultured mouse hippocampal neurons (10 days *in vitro*) were transfected with the indicated constructs and imaged 18 h later using epi-fluorescence microscopy with filter sets for green fluorescent protein (GFP; a 100×, 1.35 numerical aperture lens was used here) and a closed-chamber heated coverslip apparatus (Bioptechs, USA). (a) Schematic of constructs used to visualize CaMKII-alpha mRNA in living cells. One construct is a fusion of GFP to the viral protein MS2 (GFP fusion protein) and contains a nuclear localization sequence. The second construct is the reporter mRNA, which consists of a fusion of the lac Z coding region (lac Z) to 8 repeats of the MS2-binding hairpin loops (MS2 hl) and the zipcode sequence, which here is the 3′UTR of murine CaMKII-alpha (both the MS2 loops and the zipcode are in the 3′UTR of this construct). Since the MS2 hairpin loops each contain two binding sites for MS2 protein, each mRNA can theoretically be linked to 16 GFP molecules. (b) Analysis of CaMKII-alpha reporter mRNA movements in neurons using Image J. A time-lapse movie stack was converted to a maximum projection image (max), which displays the maximum pixel for each frame with stationary particles as dots and moving particles as lines, highlighting their trajectories over the sequence. A dendritic region of interest (ROI) was highlighted (white box) for further analysis (see **Figure 1(c)**, kymograph), and the line used to analyze a particle trajectory was drawn over that ROI (yellow; see inset). Still frame images of the highlighted particle movements are shown in sequence with the corresponding time stamps (every 1.5 s; upper right corners). An mRNA particle emerges from a stationary particle (red arrow) and traverses another stationary particle (green arrow), which then moves rapidly in the anterograde (right; cell body is at left) direction together with the previously moving particle. The particles move in parallel toward two stationary particles positioned anterograde in the dendrite (yellow arrows), then traverse the first stationary particle (left yellow arrow) and come to a stop at the second, more distal particle (right arrow), where they remain throughout the duration of the imaging series. (c) A kymograph of the rapid CaMKII-alpha

requires *N*-methyl-D-aspartate (NMDA) stimulation but not protein synthesis, and activation of dendritic domains in the medial dentate gyrus by specific stimulation of the medial entorhinal cortex caused a selective increase in Arc mRNA to those sites, implying input-specific localized synaptic activation and subsequent delivery of mRNA to those synapses. This increase in trafficking to dendrites is thought to require the cytoskeleton and interaction of the transport mRNPs with anterograde molecular motors. How the mRNAs recognize the sites at which they are anchored remains unknown, although there is work from other systems such as *Drosophila* that shows a role for microtubule-based motors in anchoring mRNAs. However, given the abundant actin-rich regions surrounding the synapse, one may speculate that myosin motors may also be involved in the anchoring process. In fact, recent data supports a role for MyoVa in RBP movement into dendritic spines.

Neuronal stimulation can cause the localization of mRNAs for neurotrophic factors themselves, such as brain-derived neurotrophic factor (BDNF) and its receptor kinase mediator trkB in cultured neuronal dendrites. In addition, stimulation can enhance the extent of localization, with BDNF and trkB mRNAs enriched only in the proximal dendrites under basal conditions, and upon depolarization, a significant fraction of mRNAs move into the distal dendrite. NMDA and neurotrophin stimulation results in an increase of ZBP1 and beta-actin mRNA localization to dendrites of cultured neurons. With an emerging technology of *in vivo* labeling of mRNAs using green fluorescent protein (GFP)-tagged MS2 reporter construct fusions (see Figure 1), rapid analysis of specific transport mRNP dynamics in living neurons became possible. Particles moved with both oscillatory and persistent trajectories, both in the anterograde and retrograde directions, and neuronal activity modified these dynamics. Neuron-wide depolarization (using KCl) induced the net anterograde transport of CaMKII-alpha mRNA in dendrites and caused a repositioning of the population of mRNAs with respect to synapses for those mRNAs already localized to dendrites. In addition, high-frequency stimulation leads to the rapid increase in CaMKII-alpha mRNA in synapse fractions isolated from dentate gyrus, a process independent of NMDA activation. The results suggest that a fine-tuning mechanism may exist to regulate the local population of mRNAs in shafts and spines, modifying the content of proteins present within those synapses on stimulation in order to effect long-term changes in synapse plasticity. Consistent with this idea, the localization of RBPs involved in plasticity, such as FMRP, is altered in dendrites on stimulation. Both the protein and the mRNA encoding FMRP are trafficked into dendrites of mature hippocampal cultures in response to stimulation of group I metabotropic glutamate receptors (mGluRs). Similarly, alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor subunit mRNAs are increased in their dendritic localization by mGluR, but stimulation of NMDA causes a selective decrease in the abundance of dendritic AMPA subunit mRNAs. This suggests that activity can bidirectionally control the relative abundance of excitatory ionotropic receptor mRNAs within the vicinity of synapses, possibly contributing to changes in synaptic efficacy during developmental plasticity.

Function of Dendritic mRNAs in Synapse Plasticity and Neuronal Polarity

The first evidence of a specific requirement for dendritic mRNA targeting in mediating synaptic plasticity came from studies on CaMKII-alpha mRNA. This mRNA harbors a zipcode within the 3′UTR that is sufficient for dendritic localization of reporter constructs bearing this zipcode. In a mouse that contained a deletion of most of the 3′UTR in the endogenous CaMKII-alpha allele, physiological experiments showed defects in long-term potentiation at synapses, and behavioral experiments showed alterations in several tests for spatial and contextual conditioning that are hallmarks for learning and memory. The neurons from these mice showed significantly diminished CaMKII-alpha protein levels at synapses, suggesting that targeting of the mRNA to dendrites is essential for local protein expression to support long-term plasticity. What remains unclear is at which stage during brain development the requirement for this local synthesis occurs. Also, other studies have indicated that the 5′-most 94 nt of the 3′UTR sequence that was left intact in the above-mentioned deletion mutant was sufficient to mediate dendritic localization of reporter constructs into

mRNA particle trajectory outlined in (b) (max inset, yellow line), where distance moved (in pixels, *x*-axis) is plotted against time (seconds, 1.5 s interval per line). This analysis shows stationary particles as vertical lines and moving particles as nonvertical with velocities equal to the slopes of the lines. An example of the calculation of trajectory velocities is shown (black box) for two particles moving in parallel within the dendrite. These particles appear to merge and then move rapidly (black arrows) toward several other stationary particles, traversing the first stationary particle and then stopping at the second stationary particle. The increased intensity of the right-most vertical line (second stationary particle) after the moving particle stops there confirms this event.

dendrites, suggesting that reporter studies may give misleading results. However, it is interesting that associates with translational regulator CPEB, an RBP that associates with CaMKII-alpha mRNA, plays a role in the modification of synaptic plasticity and is itself phosphorylated by the CaMKII-alpha enzyme. This feedback loop regulation may be an important paradigm for mechanisms that underlie activity-induced changes in synapse strength.

Fragile X syndrome (FXS) is a neuronal disease characterized by mental retardation and spine dysgenesis due to loss of function of FMRP, a key RBP thought to regulate translation of mRNAs in dendrites. Physiological experiments suggest that one form of long-term depression, which depends on mGluR and rapid protein synthesis, is altered in mouse hippocampus bearing a null-mutation for FMRP. Therefore the control of dendritic mRNA processing (transport, translation, or stability) by FMRP may be important for this form of plasticity and may link changes in plasticity to dendritic morphological transformations that are fine-tuned during development. Indeed one of the mRNAs regulated by FMRP is MAP1b, an important cytoskeletal protein that may contribute to the phenotypic alteration of dendritic protrusions that are hallmarks of mental retardation in general and FXS specifically. Another RBP in brain, Staufen2, appears to be involved in regulation of dendritic mRNA transport and causes reduced dendritic spine numbers with concomitant increases in immature-like protrusions when its expression is reduced in neurons. It is interesting to note that one of the target mRNAs that is reduced in dendrites of Staufen2-depleted neurons is beta-actin. The delivery of beta-actin mRNA into dendrites is linked to the growth of dendritic filopodia and formation of filopodial synapses in response to BDNF, a process that requires ZBP1 and its interaction with the zipcode of beta-actin mRNA. These studies implicate a role for transport mRNPs and the delivery of cytoskeletal-encoding mRNAs to dendrites in the regulation of morphological plasticity during neuronal development.

A role for mRNA delivery in establishment of neuronal polarity is under investigation. Two major candidate mRNAs, MAP2 in dendrites and tau in axons, are found in these specific subcellular compartments exclusively, and their local translation there may be important for establishment and maintenance of neuronal polarity. These mRNAs are bound by distinct RBPs that are transported with these mRNAs in dendrites and axons: MAP2 by the MARTA1/2 proteins and tau by one member of the embryonic lethal abnormal vision family of RBPs, HuD. While the molecular mechanism of their differential transport is not well characterized, the details are beginning to emerge. MAP2 mRNA contains a 640 nt zipcode, the DTE mentioned earlier, which is sufficient to confer localization of a reporter mRNA to the dendritic compartment of neurons. An 11 nt sequence within the coding region of MAP2 mRNA has been suggested as a targeting element for localization of a GFP reporter mRNA to oligodendrocyte processes and hippocampal dendrites; however, a reporter mRNA containing the entire coding region of MAP2 is not able to localize efficiently in neurons, suggesting that this element may not be sufficient for dendritic targeting of endogenous transcripts and that inhibitory elements within other regions may serve to regulate transport.

Tau mRNA contains a 240 nt zipcode that is required for targeting to the axonal compartment of differentiated P19 cells that segregate dendritic and axonal markers similarly to primary neurons. This zipcode is also sufficient to redirect the dendritic MAP2 mRNA to axons when substituted for the MAP2 3′UTR. Conversely, the dendritic 640 nt zipcode located in the 3′UTR of MAP2 mRNA is sufficient to redirect tau mRNA to dendrites when substituted for the tau 3′UTR, suggesting that these zipcodes are primary determinants of localized gene expression in the establishment of neuronal polarity. Other factors that may contribute to the distinct localization of MAP2 and tau mRNAs include the association of these transport RNPs with two distinct plus end-directed microtubule motors, Kif5 and Kif3, respectively. Future work may reveal how the localized expression of these genes maintains segregation of downstream determinants of neuronal polarity.

Dendritic mRNP Transport and Neuronal Disease

Several diseases of the nervous system have been attributed to defects in the function of RBPs and in the processing of RNA. FMRP is an RBP highly expressed in brain that is silenced in FXS, the leading cause of inherited mental retardation in humans and a known cause of autism. FMRP binds to numerous mRNAs in brain to repress translation, and loss of FMRP results in elevated mRNA translation and excessive mGluR-dependent long-term depression of synaptic strength in the hippocampus and cerebellum of a mouse model of FXS. In addition, neurons lacking FMRP show an increase in the number and length of dendritic protrusions. Although several identified mRNA targets of FMRP are localized to dendrites and have important functions in neuronal homeostasis, the mechanisms that give rise to altered dendritic morphology and plasticity are not well understood. Similarly, translin is an RBP that is expressed in brain

and regulates mRNA transport and translation in dendrites. Mice lacking translin show altered expression of several mRNAs in brain and have impaired learning and memory function in behavioral tests. One target mRNA of translin is CaMKII-alpha, and since its dendritic transport has been shown to be critical for normal developmental plasticity, this aspect may be altered in neurons lacking translin.

Another RBP is the spinal motor neuron (SMN) protein that is not expressed in the progressive neurodegenerative disease spinal muscular atrophy. SMN functions in the assembly of small nuclear RNP complexes that participate in RNA splicing, and recent work suggests that SMN may play a role in the localization of mRNP complexes to neuronal processes as well. Nova proteins also function in the splicing of neuronal mRNAs, and the autoimmune disorder paraneoplastic opsoclonus myoclonus ataxia (POMA) results in the antagonism of Nova RBP function. One important target of Nova is the inhibitory glycine receptor, and the misregulated splicing of this mRNA may contribute to the excessive motor function in POMA patients. It is interesting to note that Nova-null mice are deficient in a novel form of synaptic plasticity that may result from defective splicing, long-term potentiation of the slow inhibitory postsynaptic current. The localization of Nova target mRNAs in neurons lacking this RBP remains to be determined. Together these alterations in mRNA processing and expression underlie the important function of localized gene expression in the nervous system for the maintenance of synaptic plasticity and normal signal transmission.

Analysis of mRNA Dynamics in Living Neuronal Dendrites

Recent advances in molecular techniques that aid in the visualization of mRNA movements in living neurons have led to the ability to tag mRNAs with fluorescent fusion proteins such as GFP. By insertion of a viral hairpin repeat sequence in the 3′UTR of a given mRNA, one can co-express a GFP fusion containing a small protein, MS2, that binds to these hairpin repeats with nanomolar affinity. With this system, one can follow the movements of a CaMKII-alpha reporter mRNA containing the complete 3.2 kb 3′UTR shown to be sufficient for dendritic targeting in neurons, when fused to eight MS2 hairpin loops. The same system has been used to analyze the transport of CaMKII-alpha reporter mRNAs in living hippocampal neurons with improved camera sensitivity, enabling faster sampling rates (Figure 1).

The analysis of one movie shows several transport mRNP particles (herein termed ‘particles’) moving rapidly in the anterograde and retrograde directions, as well as stationary and oscillatory particles, with an obvious heterogeneity in the intensity of both moving particles and stationary ones, suggesting that the number of mRNAs in these particles varies over a considerable range. To analyze the velocities and trajectories of these particles, the program ImageJ, which is a software that runs on most computer platforms and is available at no cost from the National Institutes of Health, may be employed. Using this program, one can readily identify moving particles by first generating a maximum projection of a time-lapse stack of images (Figure 1(b)). This analysis shows stationary particles as dots and motile particles as lines. The trace of one trajectory with reslicing generates a kymograph that displays time (sequential frames in the stack) on the y-axis and distance (pixels) on the x-axis (Figure 1(c)).

One can notice from the kymograph that the trace line reveals two particles moving in parallel (within the black box) at the same velocity that join together with a stationary particle to form a larger particle (Figure 1(b), red arrow). Then one of these particles rapidly moves out of the stationary particle, almost unnoticeable on the still frame analysis panel (Figure 1(b), green arrow) but very distinct on the kymograph, where it shows a long, almost horizontal line (Figure 1(c), black arrows) that traverses another stationary particle (Figure 1(b), left yellow arrow) and comes to a stop at a second stationary particle (Figure 1(b), right yellow arrow). Overall, this analysis showed that persistent particles containing the CaMKII-alpha reporter mRNA moved at $0.8\,\mu m\,s^{-1}$ on average with a maximal rate of $2.1\,\mu m\,s^{-1}$. In comparison with previous work, this faster sampling rate reveals that transport mRNPs move at rates at least tenfold faster (previous average velocities were estimated at $0.05–0.10\,\mu m\,s^{-1}$ and maximal velocities at $0.1–0.2\,\mu m\,s^{-1}$). It also reveals that many trajectories involve pauses or oscillatory movements that contribute to lower average velocities when sampling is done at lower frequencies. Future work may determine whether classes of transport mRNPs move at homogeneous rates and how neuronal activity may differentially modulate these velocities within dendrites.

See also: Adult Cortical Plasticity; Axonal mRNA Transport and Functions; Axonal and Dendritic Identity and Structure: Control of; Cytoskeleton in Plasticity; Dendrites: Localized Translation; Developmental Disability and Fragile X Syndrome: Clinical Overview; Fragile X Syndrome; Neuronal Plasticity after Cortical Damage; Plasticity and Activity-Dependent Regulation of Gene Expression; RNA Binding Protein Methods; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Job C and Eberwine J (2001) Localization and translation of mRNA in dendrites and axons. *Nature Reviews Neuroscience* 2: 889–898.

Kiebler MA and Bassell GJ (2006) Neuronal RNA granules: Movers and makers. *Neuron* 51: 685–690.

Kosik KS and Krichevsky AM (2002) The message and the messenger: Delivering RNA in neurons. *Science's STKE : Signal Transduction Knowledge Environment* 126: pe16.

Martin KC and Kosik KS (2002) Synaptic tagging: who's it? *Nature Reviews Neuroscience* 3: 813–820.

Schuman EM (1999) mRNA trafficking and local protein synthesis at the synapse. *Neuron* 23: 645–648.

Singer RH (2003) RNA localization: Visualization in real-time. *Current Biology* 13: R673–R675.

Steward O (1997) mRNA localization in neurons: A multipurpose mechanism? *Neuron* 18: 9–12.

St Johnston D (2005) Moving messages: The intracellular localization of mRNAs. *Nature Reviews Molecular Cell Biology* 6: 363–375.

Tiedge H, Bloom FE, and Richter D (1999) RNA, whither goest thou? *Science* 283: 186–187.

Relevant Website

http://rsb.info.nih.gov – ImageJ, Image Processing and Analysis in Java.

Dendritic Signal Integration

N Spruston, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview: Questions

Most neurons have elaborately branching dendritic trees that receive thousands of synaptic inputs (Figure 1). A majority of dendritic synapses are excitatory, but inhibitory synapses also contact dendrites. A number of simple questions can be asked about dendritic integration:

- How many excitatory synapses need to be activated before an action potential will occur in the axon?
- To what degree does this number depend on the location of the synapses on the dendritic tree?
- How does the spatial relationship between excitatory synapses affect dendritic integration?
- How does inhibition affect the integration of excitatory synaptic inputs?
- How does dendritic inhibition differ from somatic or axonal inhibition?
- How does the spatial relationship between excitatory and inhibitory synapses affect dendritic integration?
- How does the presence of voltage-gated channels in dendrites affect dendritic integration?
- Can spikes be initiated in dendrites and/or propagate along them?
- How do dendritic spines affect synaptic integration?

The answers to these questions depend on several important factors, such as dendritic structure and excitability, as well as the pattern of synaptic activation and the modulatory state of the neural network. Dendritic integration clearly depends on dendritic structure and ion channel expression, so the dramatic variations between cell types (Figure 1) will certainly lead to cell-specific differences in dendritic integration. One goal of dendritic physiologists is to determine common principles that apply to all cell types, while another goal is to determine cell-specific specializations of dendritic integration. Similarly, dendritic integration of fast excitatory and inhibitory synapses is influenced by the modulatory state of the network, which is determined by slower, neuromodulatory transmitters acting on G-protein-coupled receptors, in turn affecting both synaptic function and voltage-gated channels. Another goal of dendritic physiologists is to determine how dendritic integration is affected by these modulatory neurotransmitters.

Approaches

The study of dendritic integration draws on several approaches. Basic anatomical methods define the architecture of dendritic trees, while electron microscopic analysis provides detailed information about the fine structure of dendrites, spines, and synapses, as well as organelles that influence the function of dendrites. Antibodies against ion channels, receptors, and other proteins can provide important information about the molecular composition and organization of dendrites. Immunolocalization of these molecules can be visualized at the light level, but electron microscopy provides the best way of determining which molecules are expressed in dendrites, as opposed to the presynaptic and glial structures that are closely associated with dendrites.

Dendritic physiology has primarily been studied *in vitro* using sharp-microelectrode or patch-clamp recording from neurons in brain slices. The ability to use several variants of the patch-clamp method to record from dendrites in slices has been particularly helpful for investigating dendritic function. Imaging neurons with calcium or voltage-sensitive fluorescent probes has also provided valuable information about dendritic integration, especially in regions of the dendrites that are currently inaccessible to patch-clamp recording. The development of multiphoton imaging and uncaging has improved the resolution and expanded the scope of optical approaches to studying dendritic integration. Although slice physiology has dominated the study of dendritic function because of technical advantages, *in vivo* studies have also been important, because they allow dendrites to be studied in a more natural state. Dendritic recordings have been obtained *in vivo*, and have been complemented by studies using multiphoton imaging. Finally, computational methods have been used extensively in the study of dendritic integration. Computer models of realistic or abstracted dendritic trees allow data from the variety of experimental approaches to be brought together in a quantitative way. Such models are powerful tools for developing theories of dendritic function and for making experimentally testable predictions that lead to new insights.

Synaptic Integration in a Simple Neuron: Cerebellar Granule Cells

A good starting point for understanding synaptic integration is the cerebellar granule cell (Figure 1(e)). These cells are the most numerous types of neurons in

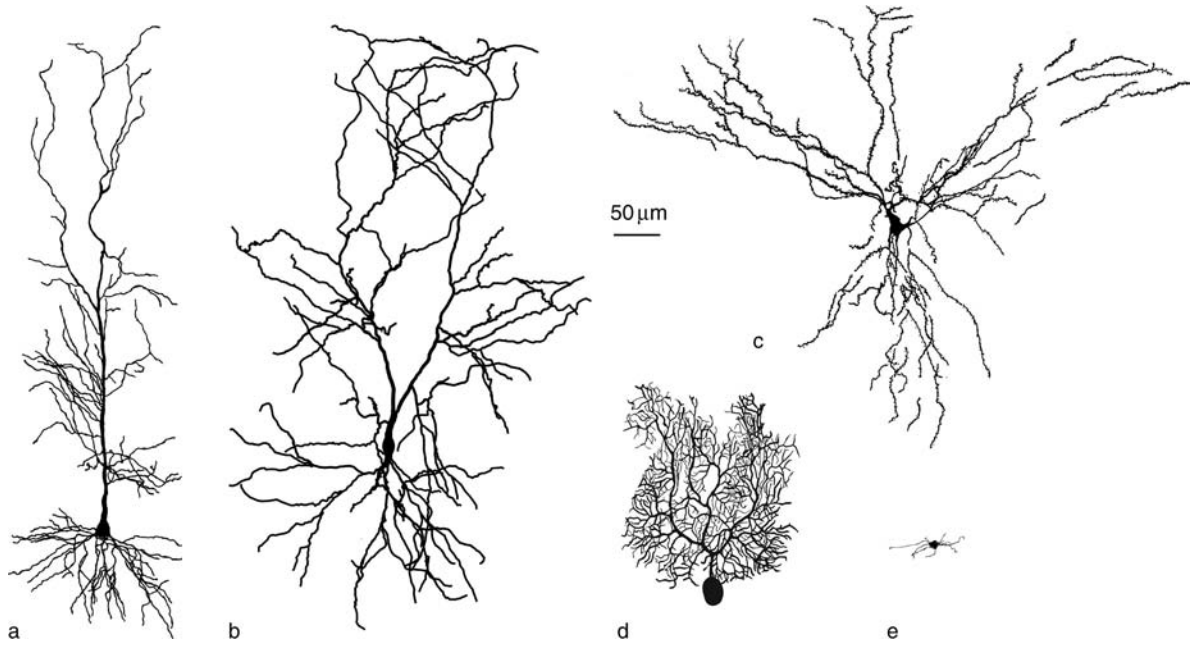

Figure 1 Neurons exhibit dendritic structure. (a) A hippocampal CA1 pyramidal neuron. (b) An interneuron from stratum radiatum of CA1. (c) A stellate cell from layer II of the rat entorhinal cortex. (d) A cerebellar Purkinje cell. (e) A cerebellar granule cell. Neurons in all panels are from the rat; axons have been truncated or removed. The scale bar applies to all panels. (a) Adapted from Golding NL, Jung H, Mickus T, et al. (1999) Dendritic calcium spike initiation and repolarization are controlled by distinct potassium channel subtypes in CA1 pyramidal neurons. *Journal of Neuroscience* 19: 8789–8798; (b) from http://www.northwestern.edu/neurobiology/faculty/spruston/sk_models; (c) adapted from Klink R and Alonso A (1997) Morphological characteristics of layer II projection neurons in the rat medial entorhinal cortex. *Hippocampus* 7: 571–583; (d) adapted from Roth A and Häusser M (2001) Compartmental models of rat cerebellar Purkinje cells based on simultaneous somatic and dendritic patch-clamp recordings. *Journal of Physiology* 535: 445–472; (e) adapted from Cathala L, Brickley S, Cull-Candy S, et al. (2003) Maturation of EPSCs and intrinsic membrane properties enhances precision at a cerebellar synapse. *Journal of Neuroscience* 23: 6074–6085.

the brain, and they are also arguably the simplest. Each granule cell receives on average only four excitatory inputs, each made onto a very short dendrite. Because the dendrites are so short, their influence on synaptic integration is minimal. This also offers technical advantages for studying synaptic physiology with patch-clamp recordings from the cell body.

Studies of these and other glutamatergic synapses have revealed that each activated synapse results in an excitatory postsynaptic current (EPSC), which reaches a peak in a fraction of a millisecond and decays in about 1 ms. This unitary EPSC (i.e., one synapse) results in an excitatory postsynaptic potential (EPSP) of a few millivolts. The EPSP, however, decays more slowly than does the EPSC. When current flows though glutamate-gated channels, the charge is stored on the membrane due to its capacitance, thus producing an EPSP. The EPSP continues to grow as long as the synaptic current exceeds the leak current, so the peak of the EPSP corresponds roughly with the end of the EPSC (**Figure 2(a)**). The amplitude of the EPSP depends on both the amplitude and duration of the synaptic current. When the synaptic current has decayed, the charge leaves the cell through leak channels, resulting in approximately exponential decay of the EPSP. The time constant of this decay is governed by the membrane time constant (τ_m) of the cell, which is determined by the product of the membrane capacitance C and membrane resistance R ($\tau_m = RC$). The leakier the cell (lower R) and the less its ability to store charge (lower C), the lower the time constant, and hence the faster the decay of the EPSP. Thus, resting membrane properties affect synaptic integration as much as do the properties of the synapses. In cerebellar granule cells, the EPSC decays with a time constant of about 1 ms (a property of the synapse), while the EPSP decays with a time constant of about 5 ms (a property of the cell membrane).

The difference between the resting potential and action potential threshold in granule cells is about 15 mV, so more than one excitatory synapse needs to be activated simultaneously in order for an action potential to occur in the axon. This is referred to as 'spatial summation' of EPSPs (**Figure 2(b)**), because the synapses are spatially distinct (even though granule cells are very small). Alternatively, one or more synapses may be activated repeatedly, leading to 'temporal summation' of EPSPs (**Figure 2(c)**). In order for

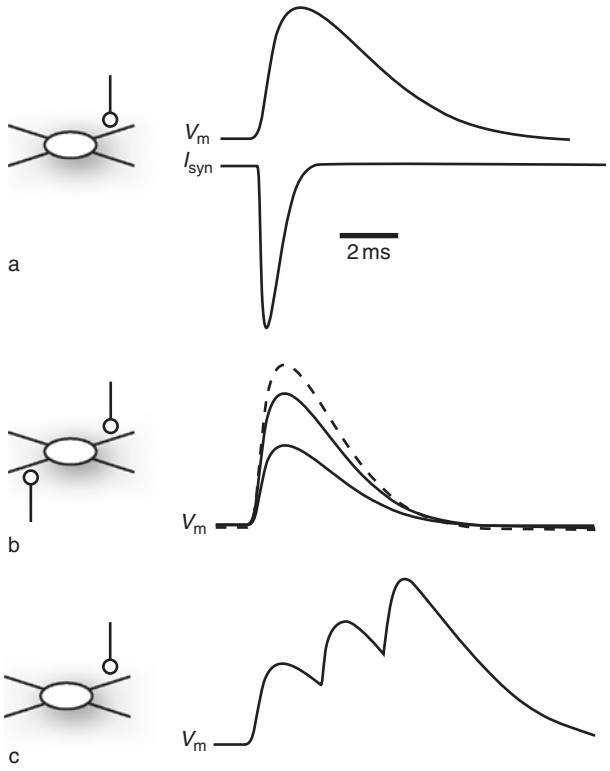

Figure 2 Integration of EPSPs. (a) Activation of a single synaptic input (as shown schematically for a synapse on a cerebellar granule cell) results in a fast inward current (excitatory postsynaptic current; EPSC) that produces an excitatory postsynaptic potential (EPSP). The magnitude of the EPSP depends on the amplitude and duration of the EPSC; the EPSP peaks approximately when the EPSC decays to zero. The magnitude of the EPSP also depends on the postsynaptic capacitance and input resistance, as does the final decay rate of the EPSP. See text for details. (b) Simultaneous activation of two synapses results in spatial summation, which is slightly less than the linear sum of the two inputs alone (dashed line). (c) Repeated activation of a synapse results in temporal summation (also sublinear), provided that the activation interval is shorter than the EPSP decay time. All traces are schematic (additional details are provided in the text).

this to occur, the frequency of this repeated activation must be high enough for each EPSP to occur before the preceding EPSP decays.

A complication of spatial or temporal summation in these neurons is that the EPSP generated by one synapse reduces the current flow at other synapses; thus, EPSP summation is slightly sublinear. This sublinear summation is caused by the effect of membrane potential on current flow at the synapse, which is governed by the driving force for synaptic current. Driving force is the difference between the membrane potential (e.g., at rest, $V_m = -70\,\text{mV}$) and the synaptic reversal potential, which is determined by the ionic permeability of the channels activated at the synapse. For glutamate-gated channels, which are roughly equally permeable to Na^+ and K^+, the synaptic reversal potential (E_{rev}) is 0 mV, so the driving force ($V_m - E_{rev}$) is 70 mV. When one EPSP has depolarized the membrane, the driving force is accordingly reduced, thus reducing the charge that enters at the synapse. For example, a 7 mV EPSP would reduce the driving force by 10%.

This is the basis for sublinear EPSP summation in a passive system. Most neurons, however, also contain voltage-activated channels that are activated in the voltage range between rest and the action potential threshold. The effects of these channels depend on their ionic permeability. For example, subthreshold activation of K^+ channels tends to increase the sublinearity of EPSP summation, while subthreshold Na^+ or Ca^{2+} channel activation can reduce sublinearity or even lead to supralinear EPSP summation. As these channels are voltage sensitive, their effects are dependent on the size of the EPSP. Similarly, the *N*-methyl-D-aspartate (NMDA) type of glutamate receptor has voltage-dependent properties that can amplify EPSPs.

Basic Effects of Dendrites on Synaptic Integration

Because granule cell dendrites are so short, they have a minimal effect on synaptic integration. Most neurons, however, have considerably longer and elaborately branching dendritic trees, which introduces further complexities to the process of dendritic integration. The simplest of these effects is that EPSPs are attenuated as current flows from the dendrites (where most excitatory synapses are located) toward the soma. This attenuation is a significant consideration, because the action potential is initiated in the axon, which typically emanates from the soma. Thus, the soma and axon are often considered to be the final site of synaptic integration.

When charge enters a dendrite through synaptically activated ion channels, much of the charge is deposited onto the membrane locally, due to its capacitance, thus producing a local depolarization. All of the charge eventually flows away from the vicinity of the synapse, however, causing the local membrane potential to decay. Some of this charge leaves the cell via the leak channels that are open at rest in the membrane. Most of the charge, however, flows along the dendrites, with only some of it flowing toward the soma. Furthermore, as charge flows toward the soma, some of it continues to leak out of the cell through the resting leak channels. As a result, the charge that reaches the soma is just a fraction of the original synaptic charge that entered the dendrite. Accordingly, the synaptic potential is smaller in the soma than it is in the dendrite (Figure 3).

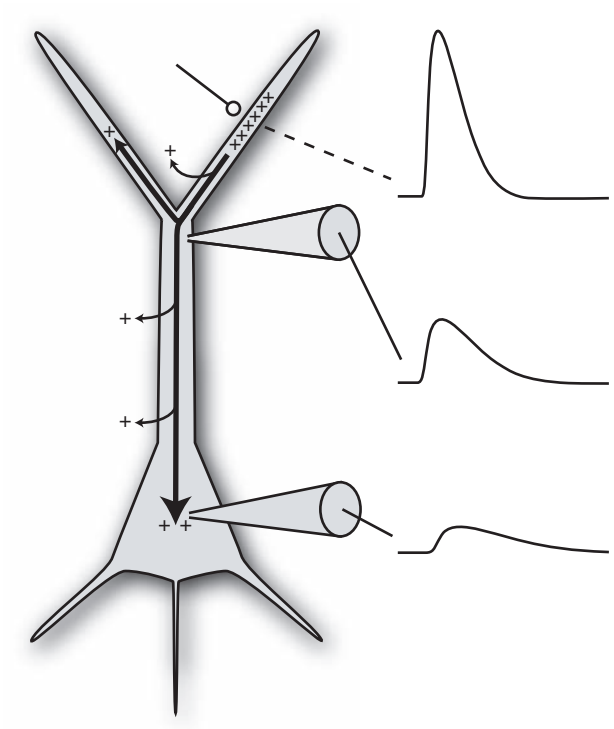

Figure 3 Dendritic attenuation and filtering of an excitatory postsynaptic potential (EPSP). A schematic diagram of a pyramidal neuron is shown, containing a pyramidal soma, apical dendrites (from soma apex), basal dendrites (corners of soma base), and an axon (from center of soma base). Activation of a synapse on an apical dendrite produces a local EPSP (top) that is larger and faster than the EPSP recorded at more proximal locations (as indicated by recording electrodes). EPSP attenuation is caused by loss of charge between the synapse and other locations, as well as changes in local input impedance. See text for details.

The attenuation of EPSPs between the synapse and the soma raises the important question of whether synapses that are far from the soma compensate in some way for this distance-dependent disadvantage. Although there is evidence that some synapses, in some neurons, may have larger conductances at more distal locations, different results have been obtained in other types of neurons, so the issue remains unresolved and somewhat controversial.

Another factor increases the difference between the local synaptic EPSP and the somatic EPSP. Because dendrites are small, they have a high input impedance (high resistance); thus, the current flowing at the synapse produces a large voltage change. By contrast, the soma is much larger and therefore has a much lower impedance; the voltage change resulting from an equivalent current is therefore smaller than in the dendrite. Thus, two factors result in EPSP attenuation between the dendritic synapse and the soma: charge attenuation and impedance differences. Both of these factors become more pronounced for synapses located at greater distances from the soma, but the magnitude of the effects depends on many factors, including dendritic structure (diameter, length, and branching) and membrane properties, as well as the kinetics of the EPSP (faster EPSPs attenuate more).

A mathematical theory for analyzing current flow in dendrites and the resultant effects on EPSP attenuation was provided by Wilfrid Rall in the 1950s and 1960s. This theory was called 'cable theory,' because it treated dendrites as conductive cables with leaky membranes immersed in a conductive medium (the extracellular space). The theory drew extensively on insights provided by the analysis of current flow in transatlantic telephone cables. Rall's cable theory now serves as the foundation for modern computer modeling of neurons with dendritic trees. Such models are now being used to provide insight into how charge spreads in elaborately branching dendritic trees containing many different types of ion channels.

Another way that dendrites influence EPSPs is via their effect on spatial summation. If two synapses on the same dendrite are activated together, they will sum sublinearly for the same reasons as described for the simple case of a cerebellar granule cell. If the two synapses are on different dendrites, however, the sublinearity of spatial summation will be reduced, because each synapse has a reduced influence on the other, due to the attenuation from one location to the other. For synapses located on different dendrites this effect can be quite large.

Dendrites also influence EPSPs by affecting temporal summation. Because of the resistance and capacitance of dendrites, EPSP time course is filtered between the dendrites and the soma. At the site of the synapse, the decay of the EPSP is governed primarily by the flow of current along the dendrites, away from the synapse. Because this pathway has a resistance lower than the resistance of the membrane to current flow, most of the local, dendritic EPSP decays more rapidly than does the membrane time constant. By the time the EPSP reaches the soma, however, the EPSP decay is governed mostly by the membrane time constant, which is slower. In addition, the intervening dendrites filter the rise time of the somatic EPSP. As a result of this difference in time course, temporal summation is greater in the soma than it would be in the dendrites. Furthermore, synapses located more distally are filtered more, so in a passive system their temporal summation in the soma is greater.

Effects of Inhibition on Dendritic Integration

Some subtypes of inhibitory interneurons specifically target dendrites. Although relatively little is known

about the differences between these dendritic inhibitory synapses and their counterparts on the soma and axon, an understanding of cable theory informs speculation about the possibly specialized role of dendritic inhibition.

Just as EPSPs are attenuated and filtered between the dendrites and the soma, inhibitory postsynaptic potentials (IPSPs) generated in the dendrites will be smaller and slower in the soma. However, the mechanism of action of most inhibitory synapses is activation of Cl^- channels (e.g., $GABA_A$ receptors). In most cells, the reversal potential of synapses using Cl^- channels is very close to the resting potential, so the driving force at rest is very small. Therefore, activation of these synapses may not produce significant IPSPs; instead, a major effect of Cl^- channel activation is 'shunting inhibition.' If the conductance is large, shunting inhibition can be very effective. When positive charge from an activated excitatory synapse arrives at the inhibitory synapse, it attracts Cl^- ions through the activated Cl^- channels, thus reducing the EPSP. Thus, inhibitory synapses can 'shunt' current from excitatory synapses.

The effectiveness of shunting inhibition is particularly dependent on the relative location of the excitatory and inhibitory synapses activated. Inhibitory synapses located on the soma can shunt excitatory current originating anywhere in the dendritic tree. By contrast, dendritic inhibition will shunt excitatory synapses on the same (or nearby) dendrites more effectively than it will those on other (or more distal) dendrites. Shunting inhibition also has a greater effect on somatic membrane potential (and therefore on action potential initiation in the axon) if it is located on the path between the excitatory synapses and the soma (Figure 4). Inhibitory synapses located off this path are less effective at shunting current as it flows from the synapse to the soma.

Effects of Voltage-Gated Channels on Dendritic Integration

Most of the foregoing analysis of dendritic integration has been based on passive dendrites (i.e., lacking voltage-gated channels). However, it is well known that dendrites contain voltage-gated conductances, which dramatically influence dendritic function. Dendritic patch-clamp recordings have provided direct evidence for the presence of Na^+, Ca^{2+}, K^+, and other voltage-gated channels in the dendrites of many cells. There is also compelling evidence for differences between the density and properties of these channels in dendrites as compared to the soma of the same cell type. For instance, in hippocampal pyramidal neurons, A-type K^+ channels have a higher density and lower activation threshold at increasingly distal dendritic locations. In the same cells, HCN channels (hyperpolarization-activated Na^+/K^+ channels) have the highest density in the most distal apical dendrites. Although some explanations for the functions of these nonuniform channel distributions have been reported, they remain under investigation. In

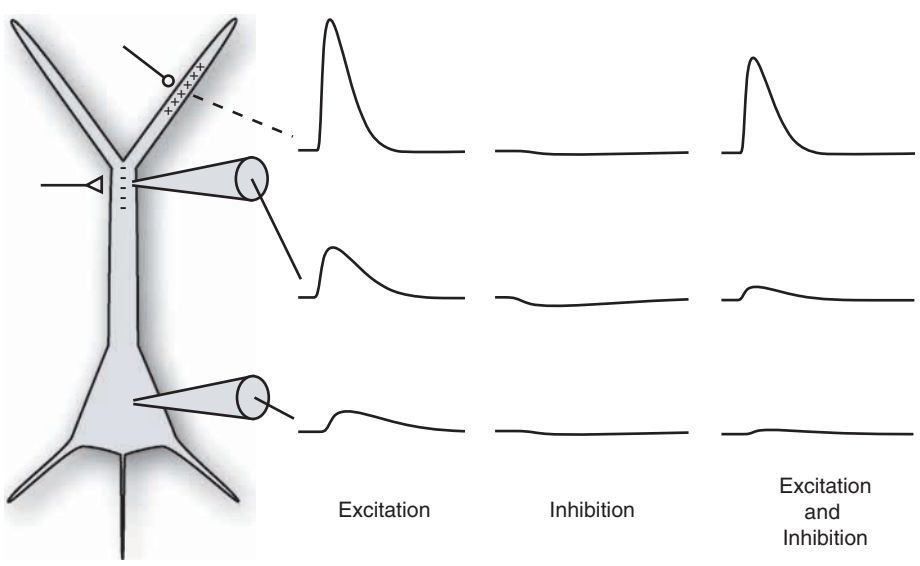

Figure 4 Dendritic inhibition. A single excitatory synapse (circle) on an apical dendrite is shunted by an inhibitory synapse (triangle) on the path between the synapse and the soma. In response to activation of both synapses (right), the somatic excitatory postsynaptic potential (EPSP) is much smaller than the sum of the EPSP on its own (left) and the inhibitory postsynaptic potential on its own (middle). See text for details.

addition to these region-specific differences in channel densities and properties, there are also cell-specific differences. For example, while expression of dendritic voltage-gated Na^+ channels has been observed in most cell types, these channels are virtually absent from the dendrites of cerebellar Purkinje cells.

Voltage-gated channels expressed in the dendrites may support some of the same functions as do channels expressed in the soma or the axon. For example, there is some evidence that subthreshold activation of dendritic Na^+, Ca^{2+}, or K^+ channels can affect EPSP summation. More compelling evidence exists for two functions of dendritic Na^+, Ca^{2+}, and K^+ channels: dendritically initiated spikes and backpropagating action potentials.

Spikes mediated by voltage-gated Na^+ and Ca^{2+} channels can be initiated in dendrites, usually in response to relatively strong excitatory synaptic activation (Figure 5(a)). NMDA receptors can also contribute to the initiation of spikes in dendrites because of their highly nonlinear voltage dependence. Dendritically initiated spikes may serve several functions, including the induction of plasticity and the release of neurotransmitters or growth factors from dendrites. In addition, dendritic spikes are likely to be important for dendritic integration. Synapses located very far from the soma may not be able to exert a significant influence over action potential initiation in the axon unless they trigger dendritic spikes locally. The ability of dendritic spikes to influence axonal action potential initiation, however, depends on their ability to propagate toward the soma. In pyramidal neurons, where dendritic spikes have been studied most extensively, the forward propagation of dendritic spikes is unreliable, often leading to propagation failure (Figure 5(a)).

Given that forward propagation of dendritic spikes is unreliable, there are two possible scenarios for the function of dendritically initiated spikes. One is that they serve a primarily local function, regulating the induction of synaptic plasticity or release of neurotransmitters from dendrites. Another possibility, however, is that forward propagation of spikes can be regulated by other factors, such as synaptic activity or the modulatory state of the neuron. Indeed, these possible functions are not mutually exclusive and there is some evidence for both of these functions.

The end result of synaptic integration is an action potential in the axon, but these signals can also propagate back into the dendrites. Backpropagating action potentials have been studied in several cell types and in all cases the amplitude of the backpropagating action potential gets smaller as it travels further from the soma (Figure 5(b)). This distance-dependent attenuation of the backpropagating action potential is caused by a relatively low Na^+ channel density, compared to the axon, where action potential propagation is reliable over long distances. Some dendrites also have high K^+ channel density, especially in distal regions. The attenuation of backpropagating

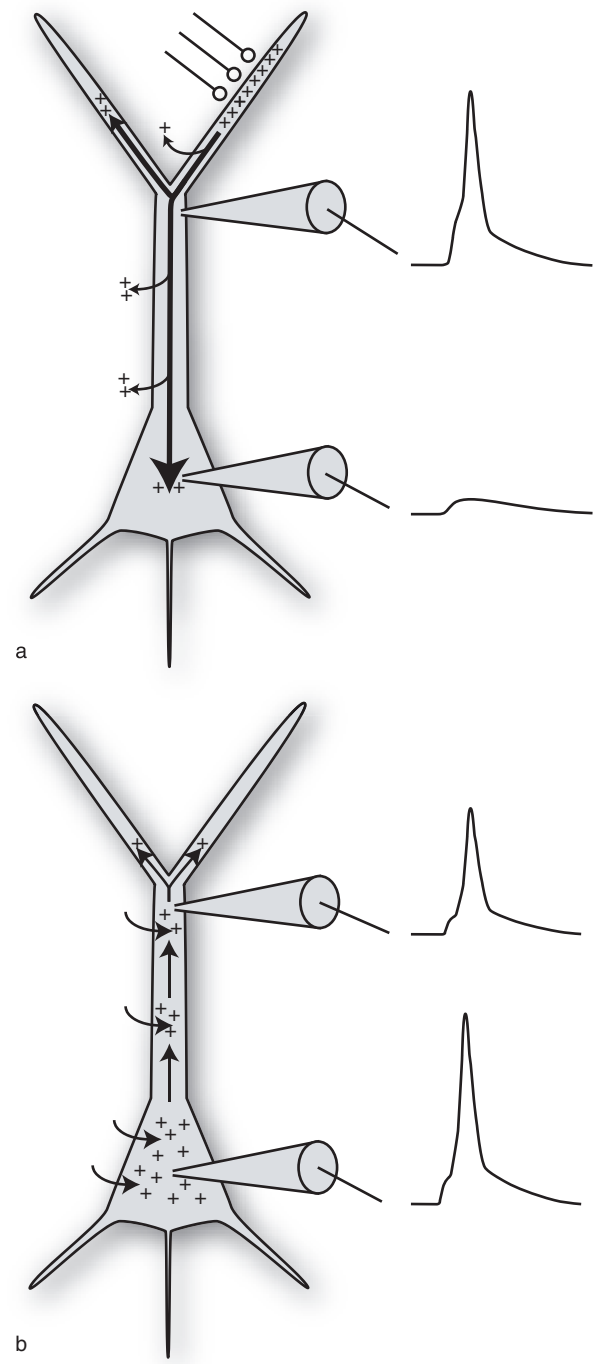

Figure 5 Dendritic excitability. (a) Strong synaptic stimulation can lead to spike initiation in the dendrites. Whether or not the dendritic spike fails (as shown) or propagates successfully to the soma (not shown) depends on several factors (see text for details). (b) An action potential that begins in the axon and soma propagates actively back into the apical dendrites, but with some amplitude attenuation. Backpropagation into basal dendrites may also occur, but is not shown.

action potentials also depends on dendritic structure, with dendritic branching tending to reduce backpropagation. Because both channel density and dendritic architecture vary across cell types, the attenuation of the backpropagating action potential is also cell-type specific. At one extreme are cerebellar Purkinje cells, which lack a significant backpropagating action potential because of the absence of Na^+ channels, as well as extensive dendritic branching. At the other extreme are the dopamine cells of substantia nigra, which show almost no attenuation of the backpropagating action potential, because of high Na^+ channel density and very limited dendritic branching. Even within one cell type, however, backpropagating action potentials are influenced by a variety of factors, including modulatory state and synaptic activity.

The functions of backpropagating action potentials are currently being explored, but there is evidence that they may be involved in the induction of synaptic plasticity and the dendritic release of neurotransmitters or growth factors. Voltage-gated Ca^{2+} channels in dendrites are activated by backpropagating action potentials. Although these channels are not essential for backpropagation, their activation is likely to be important for triggering biochemical events crucial for the functions mentioned above.

Backpropagating action potentials, though attenuated, are always detectable in dendritic recordings. The reverse is not true, however, for dendritically initiated spikes, which sometimes fail to propagate, thus resulting in diminishingly small potentials in the soma. This intriguing difference between backward and forward propagation is actually expected from the branching structure of pyramidal neuron dendrites. When axonal and somatic Na^+ current produces an action potential, a large current is required to support the action potential in the soma, because of its large surface area (and hence low impedance). As this current flows into the dendrites, which are smaller (and have higher impedance), it is relatively easy to depolarize the dendrite to activate the voltage-gated channels required for action potential backpropagation. However, when a spike begins in a dendrite, only a small current is required because of the small size (and high impedance) of the dendrite. As the spike propagates from small dendrites to larger dendrites and eventually into the very large soma, the small current initially generated by the dendritic spike will produce less depolarization of the larger structures (lower impedance). Thus, progressively fewer Na^+ channels may be activated as the spike propagates forward, eventually causing propagation failure.

Although dendritically initiated spikes and backpropagating action potentials are distinct dendritic events, they are not mutually exclusive; a few interesting interactions have been reported. In regions of the dendrites where dendritic spikes are large, Na^+ channels will be inactivated for a period of time following the spike, thus reducing the amplitude of the backpropagating action potential in the same region. The converse situation can also occur, where backpropagating action potentials result in Na^+ channel inactivation and hence a higher threshold for dendritic spike initiation. Another intriguing interaction has been observed in cortical neurons, where backpropagating action potentials – either alone or in combination with synaptic input – can lead to the generation of large dendritic spikes involving sustained activation of voltage-gated Ca^{2+} channels.

Contributions of Dendritic Spines to Dendritic Integration

One of the big mysteries of dendrites is the function of dendritic spines. Aside from relatively simple functions such as increasing dendritic surface area and space for synapses, two functions have been proposed for spines: biochemical and electrical compartmentalization. Biochemical compartmentalization may limit the spread of second messengers and other regulatory factors, thus increasing the specificity of synaptic plasticity. Compartmentalization of Ca^{2+} by the spine has been demonstrated by numerous studies. Electrical compartmentalization is more complicated and has also been more difficult to demonstrate. Membrane potential changes in the dendrite are not likely to attenuate significantly as they spread into spines. This is because the small size (high impedance) of spines makes them very easy to depolarize. On the other hand, EPSPs in spines may attenuate significantly as they spread into the larger (lower impedance) parent dendrite. Such attenuation, if significant, could mean that EPSPs in spines are very large, possibly resulting in activation of voltage-gated channels in spines, such as Ca^{2+} channels known to be present in the spine head. Quantitatively, the magnitude of the electrical compartmentalization imposed by spines depends on the resistance of the spine neck. This parameter is not well known, but imaging studies are beginning to yield important data that will address these and other issues, leading to an improved understanding of the function of dendritic spines.

Concluding Remarks

The basic principles of synaptic integration are well understood, and cable theory provides a solid foundation for using computational models to predict

how dendrites affect synaptic integration in a passive system. However, many questions remain about how excitation and inhibition interact in the dendritic tree and how dendritic integration is influenced by voltage-gated channels. Different types of neurons have different dendritic structure and appear to be equally diverse in the types of voltage-gated channels they contain. In addition, synaptic activity and modulatory neurotransmitters strongly affect how dendritic integration proceeds in any given neuron. Thus, dendritic integration needs to be studied in many different types of neurons and under different conditions. One of the most important issues will be to determine how dendritic integration works in behaving animals. This will require an understanding not only of the structural and functional properties of dendrites, but also knowledge of the patterns of synaptic activation and neuromodulatory background that occurs during various behaviors. Ultimately, this will require refinements, improvements, and extensions of the methods used to study dendrites *in vitro* and *in vivo*.

See also: Action Potential Initiation and Conduction in Axons; Activity-Dependent Remodeling of Presynaptic Boutons; Axonal and Dendritic Identity and Structure: Control of; Dendrite Development, Synapse Formation and Elimination; Dendritic Spine History; Eph Receptor Signaling and Spine Morphology; Ion Channel Localization in Cell Bodies and Dendrites; LIM Kinase and Actin Regulation of Spines; Rho GTPases and Spines; Translational Regulation at the Synapse.

Further Reading

Häusser M, Spruston N, and Stuart G (2000) Diversity and dynamics of dendritic signaling. *Science* 290: 739–744.

Jonas P and Spruston N (1994) Mechanisms shaping glutamate-mediated excitatory postsynaptic currents in the CNS. *Current Opinion in Neurobiology* 4: 366–372.

Lai HC and Jan LY (2006) The distribution and targeting of neuronal voltage-gated ion channels. *Nature Reviews Neuroscience* 7: 548–562.

London M and Hausser M (2005) Dendritic computation. *Annual Review of Neuroscience* 28: 503–532.

Magee J, Hoffman D, Colbert C, et al. (1998) Electrical and calcium signaling in dendrites of hippocampal pyramidal neurons. *Annual Review of Physiology* 60: 327–346.

Segev I and London M (2000) Untangling dendrites with quantitative models. *Science* 290: 744–750.

Segev I, Rinzel J, and Shepherd GM (eds.) (1994) *The Theoretical Foundation of Dendritic Function.* Cambridge, MA: MIT Press.

Stuart G, Spruston N, and Häusser M (eds.) (2007) *Dendrites,* 2nd edn. Oxford: Oxford University Press.

Stuart G, Spruston N, Sakmann B, et al. (1997) Action potential initiation and backpropagation in neurons of the mammalian central nervous system. *Trends in Neurosciences* 20: 125–131.

Williams SR and Stuart GJ (2003) Role of dendritic synapse location in the control of action potential output. *Trends in Neurosciences* 26: 147–154.

Relevant Websites

http://neuromorpho.org – Curated inventory of digitally reconstructed neurons.

http://senselab.med.yale.edu – Searchable database of neuronal properties.

http://www.northwestern.edu – The Spruston Lab (Northwestern University, Department of Neurobiology and Physiology).

Dendritic Spine History

A Matus, Friedrich Miescher Institute, Basel, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Discovery

Dendritic spines were first seen by microscopists of the late nineteenth century in brain sections stained using the Golgi method (Figure 1), named after the Italian neuroanatomist Camillo Golgi who introduced it in 1873. The beauty of this technique is that it impregnates only a small proportion of neurons, typically less than 1%, but those that are stained are revealed in their entirety, including the fine structure of the smallest branches of both axons and dendrites. Initially some authors, including Golgi himself, dismissed the minute protrusions we now know as dendritic spines as artifacts produced by nonspecific precipitation of silver, a reasonable conclusion considering the frequent presence of silver precipitates in Golgi-stained material. It was not until the Spanish doyen of early microscopic neuroanatomy, Santiago Ramon y Cajal, took up the Golgi method some 14 years after its discovery that the consistent features of spine morphology and their relationship to dendrites were recognized.

Cajal is generally considered to be the first to suggest dendritic spines as structures specialized for forging functional contacts between independent nerve cells in the central nervous system, an idea that remained controversial for many decades and about which he and Golgi famously disagreed, even though they shared the Nobel prize in 1912 for their discoveries in neuronal morphology. An article by Javier DeFelipe on the International Brain Research Organization (IBRO) Website with micrographs and drawings from Cajal himself gives an excellent account of this early period. Cajal's initial description of spines in the avian cerebellum, describing dendrites as "bristling with thorns (punta) and spines (espinas)," was published in Spanish in his own privately funded journal, but several useful descriptions and discussions of his work are available in more recent literature, especially in articles by contemporary Spanish neuroscientists. Cajal is also credited with originating, in his Croonian Lecture to the British Royal Society in 1894, the idea that dendritic spines may be sites of anatomical plasticity in the neuronal circuits of the cerebral cortex.

Fine Structure and Proof of Synaptic Status

Despite the constant improvement of light microscopes and histological techniques, proof that dendritic spines are component parts of synapses only became possible with the invention of the electron microscope, with which cellular structure can be resolved down to nanometers and, crucially, cell surface membranes can be visualized. Although synaptic ultrastructure is so familiar today, limitations inherent in early techniques for preparing nervous tissue for electron microscopy meant that some years elapsed before the essential morphological elements of chemical synapses were established. The limited penetrating power of osmium tetroxide, then used as a primary fixative, was a major problem, so that visualizing critical features such as the existence of a gap between the surface membranes of the pre- and postsynaptic cells – the synaptic cleft – and the presence of synaptic vesicles in the presynaptic compartment remained in doubt. Because of these difficulties in fixation and staining, these structures were first described in tissue preparations chosen for their thinness, such as the frog sympathetic ganglion and earthworm nerve cord. In a particularly influential study, the American neuroanatomist Sanford Palay achieved good preservation of mammalian brain synapses by injecting 2 or 3% osmium tetroxide directly into the fourth ventricle of anesthetized rats. By showing that the cell surface membranes of the pre- and postsynaptic cells were separate at the synaptic junction, Palay's micrographs finally resolved the historic question of whether neuronal processes are fused or separate in favor of Cajal's interpretation. However, with the revelation that the width of the synaptic cleft is only 20 nm, it became clear just how close the old controversy about state of contact between the fine branches of axons and dendrites in the brain neuropil had been. At that time, when the status of glutamate and γ-aminobutyric acid (GABA) as major neurotransmitters was not yet established and electrotonic transmission still seemed a viable mechanism for neurotransmission at brain synapses, the existence of an intracellular cleft at the synaptic junction was an important discovery in establishing that neurotransmission in the central nervous system was essentially chemical.

Palay's article contained neither evidence nor mention of the putative synaptic nature of dendritic spines.

Figure 1 Dendritic spines as revealed by the Golgi method. (a) A drawing by Santiago Ramon y Cajal of pyramidal cell neurons in the cerebral cortex prepared from material stained using the Golgi method. For identification, one spine near the bottom right of the picture is labeled 'b'. This illustration is remarkable for the acute observation and accurate rendering of dendritic spines. (b) A modern photomicrograph showing segments of two apical dendrites from pyramidal cell neurons in rat cerebral cortex. Part (a) copyright Herederos de D. Santiago Ramón y Cajal.

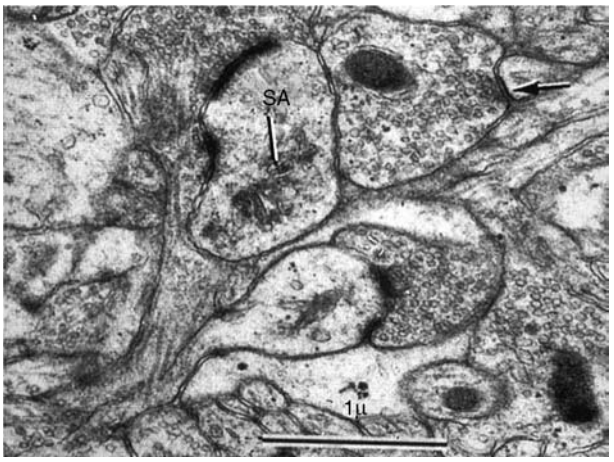

Figure 2 The fine structure of dendritic spines. This is an electron micrograph showing dendritic spines in the central nervous system of the rat. The labeling emphasizes the spine apparatus (SA) in the large postsynaptic spine near the top of the picture. Below and slightly to the right, a particularly fortunate plane of section includes an entire spine synapse and shows the presynaptic terminal, packed with synaptic vesicles, attached to a round-headed spine emerging from a narrower neck. Courtesy of George Gray.

That step was taken by the British electron microscopist George Gray, who in 1959 published detailed descriptions of spine synapses in two now-classic papers, one reporting data from slices of rat visual cortex fixed with osmium tetroxide and the other presening a detailed description and interpretation of the ultrastructural features of spine synapses. Included in these papers are the first clear descriptions of the band of dense-staining material applied to the cytoplasmic face of the postsynaptic membrane, now familiar as the postsynaptic density (PSD), and the stacks of membranous sacs, thought to be involved in regulating postsynaptic calcium levels in the spine cytoplasm, which Gray named the spine apparatus (Figure 2). Gray also called attention to the significant configuration of the pre- and postsynaptic elements, with the presynaptic bouton being applied to the tip of dendritic spine, and to the presence of a uniform gap of 20 to 30 nm between the pre- and postsynaptic cell surface membranes. Together, these features left no doubt that the contact between pre- and postsynaptic cells at dendritic spines forms a synaptic junction whose structure is highly specialized for the business of chemically mediated signal transmission.

These papers initiated an outpouring of further detailed studies of synaptic ultrastructure in different brain areas during the 1960s and 1970s demonstrating that dendritic spines are the postsynaptic contact sites for the majority of synapses on categories of neurons ranging from pyramidal calls of the cortex and hippocampus to the eponymous spiny stellate cells in the basal ganglia.

Association with Excitatory Transmission

The physiological relationship between dendritic spines and excitatory synaptic transmission emerged gradually, beginning with studies by Gray and others demonstrating differences in ultrastructure between synapses known to be associated with either excitatory or inhibitory transmission. These distinctions were aided by improvements in tissue preparation for the electron microscope, particularly fixing the brain by transcardiac perfusion with a mixture of formaldehyde and glutaradehyde. In brief, following aldehyde fixation, excitatory synapses were found to have round presynaptic vesicles, whereas those of inhibitory terminals were often flattened. In addition, excitatory synapses were distinguished by a prominent postsynaptic thickening, the PSD, which was

much reduced or absent at inhibitory synapses. Arguing from relationships between morphology and physiology in synaptic connections between Mauthner cells and motor neurons in the fish spinal cord, Diamond, Gray, and Yarsagil introduced two important ideas. First, based on their location and physiology, they pointed out the likely excitatory nature of spine synapses. Second, they hypothesized that "the dendritic spine provides a postsynaptic region which is effectively isolated from other synapses in the neuron" and pointed out the potential role of spines in dendritic integration. This provided the foundation for much later work on spines as physiological compartments involved in synaptic integration.

Ideas about Dendritic Spine Function

An early suggestion, based on the enormous abundance of dendritic spines, was that they serve to increase the area of dendrites so that a greater density of synaptic connections can be accommodated along their length. However, electron microscopy also shows that few synaptic contacts are made onto the smooth surfaces of dendrites between spines, indicating that their numbers are not limited by lack of surface area. Indeed, in mouse cerebral cortex the density of synaptic contacts is greater on nonspiny than on spiny dendrites. More convincing is the notion, debated in the recent literature, that the advantage of spines lies in their ability to make synaptic contacts at a distance from the dendrite surface. Mathematical modeling supports this hypothesis, showing that, by expanding the reach of dendrites without increasing their length, spines significantly reduce the network size required to achieve a given degree of neuronal connectivity. Perhaps even more important, the extended reach of spines also provides a mechanism for individual dendrites to sample a wider array of inputs than would be possible without these protrusions, thus potentially increasing the richness of local signal processing.

Another proposed function of spines concerns their putative electrical properties. In its simplest form, this envisages spine heads acting passively to provide a restricted environment where electrical signals can trigger local changes in synaptic transmission. More recently it has been proposed that spines have active electrical properties and are capable of synaptic amplification, although the extent to which this occurs remains controversial. A more established idea is that dendritic spines act as miniature cellular compartments that partition postsynaptic signals from the bulk of the dendrite. This has been shown clearly in the case of calcium ions, in which increases in calcium ion levels in dendrites following synaptic stimulation can be visualized using calcium-sensitive fluorescent dyes. Such studies have established that, depending on the strength of synaptic signaling, increases in calcium ion levels may be limited to individual spines, may extend to include neighboring spines, or, with more intense stimuli, may spread to encompass entire branches of dendrites. This property enables spines to operate as integrators of synaptic signals impinging on the dendrites of large neurons, such as cortical pyramidal cells, from diverse sources.

Morphological Plasticity and Cytoskeletal Structure

Morphologically dendritic spines are a curious mixture of uniformity and variety. While the 'typical' spine of the textbook diagram consists of a globular head attached to the dendrite shaft by a narrow neck, in the brain itself spines occur in a wide variety of shapes. In the past these were often loosely grouped as 'stubby' (no neck), 'mushroom' (large head with thin neck), and 'thin' (small heads with a long neck), but in fact each of the parameters underlying spine shape diversity varies along a continuum.

The reason for this diversity became apparent when it was discovered that dendritic spines contain high concentrations of cytoplasmic actin which are generally associated with changes in cell shape such as the ruffling movements of crawling fibroblasts and neuronal growth cones. The suspicion that changes in the actin cytoskeleton were responsible for the morphological variety of spines was confirmed when time-lapse recordings were made from cultured neurons expressing actin tagged with green fluorescent protein (GFP) (**Figure 3**). This direct observation of cytoskeletal dynamics showed not only that spines making synaptic contact are capable of changing shape but also that, as a result of the constant dynamics of actin filaments, dendritic spines are in a continuous state of rapid motility. More recently, imaging studies using GFP to visualize dendrites in living mice have shown that in the intact brain spines change not only shape but also turn over, appearing and disappearing over days and weeks in manner that, in sensory areas of the cerebral cortex, is dependent on sensory input.

In vitro studies suggest that changes in spine morphology are regulated by synaptic transmission and depend particularly on receptors for glutamate, the neurotransmitter substance at spine synapses. Activation of postsynaptic glutamate receptors blocks the actin-based motility of spines and, depending on the pattern of activity applied, may lead to an increase or decrease in size of the spine head which is in turn associated with changes in the number of glutamate

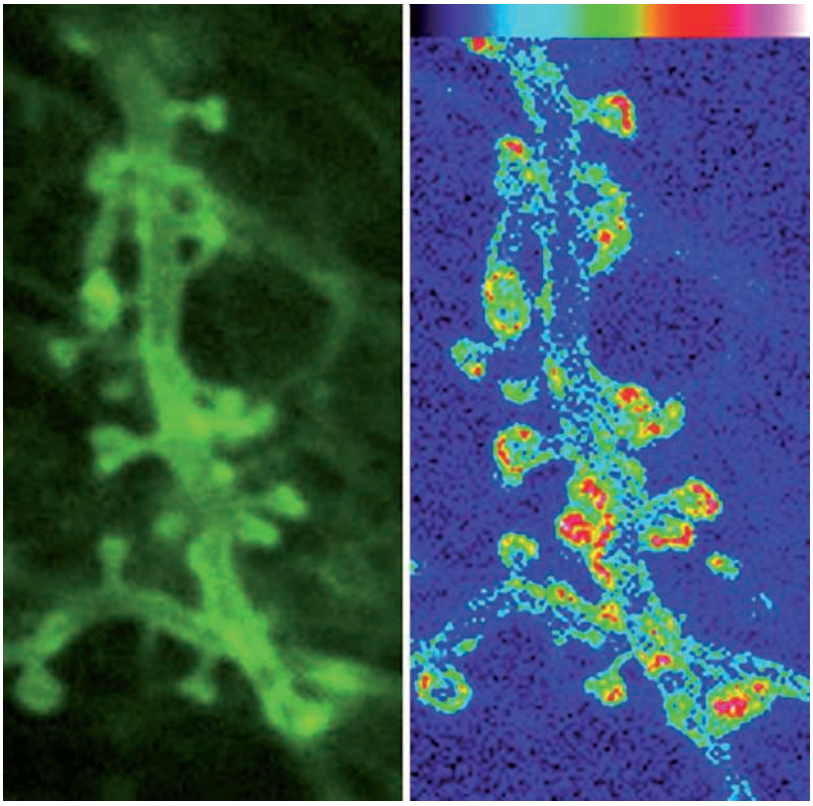

Figure 3 Dendritic spines are morphologically plastic. The left panel shows a confocal microscope image of a dendrite segment in a brain slice from a transgenic mouse in which the surface membranes of neurons are labeled with green fluorescent protein (GFP). The entire surface of the dendrite, including the spines, is evenly labeled. This image is part of a time-lapse series made to study changes in spine morphology over a period of 30 min. The image on the right is a motility plot made from the time-lapse series using a computer routine, in which the degree of motility is color coded such that regions of low motility appear blue and regions of high motility are colored red. This reveals dendritic spines as the most morphologically plastic structures on these dendrites. Reproduced from Roelandse M, Welman A, Wagner U, Hagmann J, and Matus A (2003) Focus motility determines the geometry of dendritic spines. *Neuroscience* 121: 39–49, with permission from Elsevier.

receptors present at the synapse. The stimulation patterns that induce these actin-dependent changes on dendritic spine morphology are the same ones that induce long duration changes in synaptic signal strength – long-term potentiation (LTP) or depression (LTD) – that are associated with learning and memory. Indeed, it has been shown that drugs which suppress actin dynamics interfere with the establishment of LTP. Very recently it has been shown that a specific actin-binding protein, profilin, which *in vitro* studies implicate in regulating spine motility, accumulates in dendritic spines in the appropriate area of the brain, the amygdala, during fear-conditioned learning. Together, these discoveries of the past few years support a link between actin dynamics in dendritic spines and learning and memory.

Dendritic Spines and Cognition

The conjuction of *in vivo* studies of morphological plasticity and *in vitro* exploration of molecular mechanisms is currently revolutionizing our knowledge of how brain circuits adapt to changing behavioral demands. However, it is also possible that dendritic spines may contribute to the most enigmatic of brain functions, namely, cognition. Detailed anatomical analysis shows that dendritic spines are not evenly distributed in the brain but vary in abundance between different cell types and different brain areas. As mentioned previously, spine density is especially high on dendrites of large neurons, such as pyramidal cells, that integrate inputs from many sources. In addition, these cells are especially rich in spines in areas of the brain that are associated with advanced cognitive function such as the frontal lobes of the cerebral cortex. A link between dendritic spines and cognitive function is also suggested by evidence that defects in spine morphology are a common feature in diverse forms of mental retardation.

Dendritic spines are also implicated in the processes of hibernation. It has been shown that the entry and reawakening of ground squirrels from the hiberation state is accompanied by significant loss and subsequent recovery of spine numbers. Interestingly, spine loss also occurs in brain slices of rodents when they are cooled to temperatures equivalent to those of

hibernating animals and subsequently recovers when the slices are rewarmed to physiological temperature. Finally, it has been shown that a variety of volatile anesthetics blocks spine motility *in vitro* at concentrations equivalent to those at which they are used to induce loss of consciousness in clinical anesthesia. Although there is no direct mechanism that might explain how the morphological plasticity of dendritic spines may influence animal cognition, the variety of evidence implying that such a link may exist suggests that the history of dendritic spines is far from over.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Cajal's Place in the History of Neuroscience; Dendritic Signal Integration; Glutamate Regulation of Dendritic Spine Form and Function; Hibernation; History of Neuroscience: Early Neuroscience; Postsynaptic Density/Architecture at Excitatory Synapses; Ultrastructural Analysis of Spine Plasticity.

Further Reading

DeFelipe J (2002) Santiago Ramón y Cajal. http://www.ibro.info/Media/pdf/si-his-pdf-pdf22.pdf (accessed Jun. 2007).

DeFelipe J and Jones EG (1992) Santiago Ramón y Cajal and methods in neurohistology. *Trends in Neuroscience* 15: 237–248.

Diamond J, Gray EG, and Yasargil GM (1970) The function of then dendritic spine: An hypothesis. In: Andersen P and Jensen JKS (eds.) *Excitatory Synaptic Mechanisms*, pp. 213–222. Oslo: Universitets Forlaget.

Elston GN, Benavides-Piccione R, and DeFelipe J (2001) The pyramidal cell in cognition: A comparative study in human and monkey. *Journal of Neuroscience* 21: RC163.

Fiala JC, Spacek J, and Harris KM (2002) Dendritic spine pathology: Cause or consequence of neurological disorders? *Brain Research Reviews* 39: 29–54.

Gray EG (1959) Axo-somatic and axo-dendritic synapses of the cerebral cortex: An electron microscopic study. *Journal of Anatomy* 93: 420–433.

Gray EG (1959) Electron microscopy of synaptic contacts on dendritic spines of the cerebral cortex. *Nature* 183: 1592–1593.

Harris KM and Kater SB (1994) Dendritic spines: Cellular specializations imparting both stability and flexibility to synaptic function. *Annual Reviews in Neuroscience* 17: 341–371.

Holtmaat A, Wilbrecht L, Knott GW, Welker E, and Svoboda K (2006) Experience-dependent and cell-type-specific spine growth in the neocortex. *Nature* 441: 979–983.

Jones EG (1994) Ramón y Cajal and the Croonian Lecture, March 1894. *Trends in Neuroscience* 17: 190–192.

Kaech S, Brinkhaus H, and Matus A (1999) Volatile anesthetics block actin-based motility in dendritic spines. *Proceedings of the National Academy of Sciences of the United States of America* 96: 1043–1047.

Lamprecht R, Farb CR, Rodrigues SM, and Ledoux JE (2006) Fear conditioning drives profilin into amygdala dendritic spines. *Nature Neuroscience* 9: 481–483.

Majewska AK, Newton JR, and Sur M (2006) Remodeling of synaptic structure in sensory cortical areas in vivo. *Journal of Neuroscience* 26: 3021–3029.

Matus A (2000) Actin-based plasticity in dendritic spines. *Science* 290: 754–758.

Popov VI, Bocharova LS, and Bragin AG (1992) Repeated changes of dendritic morphology in the hippocampus of ground squirrels in the course of hibernation. *Neuroscience* 48: 45–51.

Sotelo C (2003) Viewing the brain through the master hand of Ramon y Cajal. *Nature Reviews in Neuroscience* 4: 71–77.

Spacek J and Fiala J (2002) http://synapses.mcg.edu/learn/visualize/golgi1.stm

Stepanyants A, Hof PR, and Chklovskii DB (2002) Geometry and structural plasticity of synaptic connectivity. *Neuron* 34: 275–288.

Depression and the Brain

W C Drevets and M L Furey, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

A major depressive episode (MDE) is a pathological mood state characterized by persistently depressed mood and/or loss of interest and incentive, which results in clinically significant distress and social/occupational impairment. Persons suffering from an MDE also experience anxiety, impaired attention, fatigue, reduced ability to experience pleasure, self-deprecating thoughts, and disturbances of sleep and appetite. When an MDE arises as primary, idiopathic disorder unrelated to any clear medical or psychiatric antecedents, the condition is called 'major depressive disorder' (MDD). When an MDE occurs alternately with manic or hypomanic episodes, the condition is called 'bipolar disorder' (BD). MDEs may also arise secondarily to other neurological, medical, or psychiatric disorders, or in response to specific pharmacological substances. Of the depressive disorders, MDD is the most common. The World Health Organization ranks MDD as the leading cause of years lived with disability.

Patients suffering from MDEs experience prominent physical manifestations, psychological symptoms, and emotional changes. In addition to sadness or depressed mood, the latter may include anxiety, phobic fears, panic attacks, irritability, apathy, anhedonia (inability to experience pleasure or reward), or 'flatness' (lacking in emotional reactivity). The physical manifestations, including fatigue, anorexia, weight change, insomnia, headache, or psychomotor slowing (decrease in the rate and/or amount of motor or mental activity), often eclipse emotional features as the presenting complaint. The psychological manifestations of MDEs confer the greater clinical risk, however, because they include preoccupation with death, hopelessness, guilt, self-deprecation, and suicide. About 15% of patients hospitalized for MDEs ultimately die by suicide, and about one-half of completed suicides occur within the context of an MDE.

Epidemiologists estimate that between 6% and 18% of individuals will develop MDD during their life span. The variability in these prevalence rates partly reflects the lack of objective diagnostic tests and the uncertainties involved in establishing diagnosis in mild cases. This lack of precise and biologically verifiable definition of illness also presumably contributes to inconsistencies in the literature pertaining to neurobiological abnormalities associated with MDD. So long as psychiatric nosology depends on syndrome-based classifications, diagnoses for MDD will likely encompass patients with a wide range of conditions. Illness onset can occur throughout the life span, but cases are grouped, by age at onset, into early-onset depression, when the first MDE occurs before age 40 years, and late-onset depression. The heritability of early-onset MDD is much higher than for late-onset MDD. Moreover, late-onset mood disorders frequently are associated with neuroimaging and clinical evidence of cerebrovascular disease or other neurological disorders.

The course of mood disorders is highly variable. In some cases episodes of depression alternate with periods when function returns to premorbid levels. For such cases the average frequency of episodes is about five to six per 20 years. Many patients, however, experience multiple recurrences per year or develop chronic or intermittent illness without clear returns to the premorbid baseline. Antidepressant pharmacotherapy is thought to shorten the duration of MDEs, and, if continued, to decrease the risk of recurrence.

Finally, MDEs commonly occur comorbidly with anxiety disorders (such as panic disorder, social phobia, and obsessive–compulsive disorder), eating disorders such as bulimia nervosa, and substance abuse/dependence disorders. These comorbid conditions generally worsen or improve in tandem with the severity of depressive symptoms, but in some cases (e.g., social phobia and obsessive–compulsive disorder) they persist despite remission of depressive symptoms. The presence of comorbid anxiety syndromes confers a less favorable prognosis on the course of MDD.

Neurobiological Correlates of Mood Disorders

Despite many psychological and biological theories regarding the pathogenesis of mood disorders, the etiologies of MDD and BD remain unknown. The intrusive, spontaneous, and perseverative nature of MDE symptoms and their responsiveness to antidepressant drugs suggest that abnormal brain processes underlie and maintain these symptoms. Moreover, neurophysiological, neuropathological, and neurochemical abnormalities have been discovered within neural systems that modulate emotional behavior in MDD and BD, although none of these has shown sufficient sensitivity and specificity to prove useful as a diagnostic test.

Twin and other family studies indicate that genetic factors contribute substantially to the liability for developing MDD and BD, particularly in cases characterized by recurrent illness and early age-at-onset status. These data suggest that the transmission of mood disorders involves complex interactions between multiple genes, each exerting relatively small effects on vulnerability. A variety of single-nucleotide polymorphisms are reportedly associated with MDD or BD, although few findings have been replicated. The genetic data also support a role for acquired factors. For example, a gene–environment interaction has been described in which a functional polymorphism of the serotonin transporter promoter region increases the risk for MDE within the context of stress.

Although stressful events may constitute 'acquired factors' that interact with genetic susceptibility in the development of mood disorders, once the illness has been instantiated, the potential links between stressors and subsequent MDE become progressively less evident. Patients generally report that their pattern of depressive symptoms is unrelated to stressful events. The life event most clearly associated with higher risk is the cessation of pregnancy, as the postpartum period constitutes the epoch of greatest risk for developing MDD or BD in females.

A variety of neurological disorders, medical conditions, and pharmacological substances also increase the risk for developing MDE, providing clues about the neurobiological systems in which dysfunction may induce depressive symptoms. For example, Parkinson's disease (PD) is associated with a two- to fourfold increase in the risk for MDEs relative to other similarly disabling conditions. In about one-half of cases, the MDE onset precedes the onset of motor manifestations in PD. These findings have led to hypotheses that degeneration of dopaminergic projections and/or ascending serotonergic projections associated with PD impairs neural processes underlying reward processing and mood modulation, yielding depressive, anhedonic, and anxiety symptoms.

Other medical diseases that increase the risk for MDE or mania include endocrine disturbances (e.g., hypothyroidism, Cushing's syndrome), degenerative basal ganglia disorders (e.g., Huntington's or Wilson's diseases), temporal lobe epilepsy, and structural lesions (tumors, infarcts, injury) involving the striatum, frontal lobe, or mesiotemporal cortex. Pharmacologic agents that may precipitate MDEs include agents that induce neuroendocrine or biochemical changes similar to those found in primary MDD. For example, administration of corticosteroids (e.g., prednisone) or monoamine-depleting agents (e.g., reserpine) can induce an MDE, and also result in neurochemical changes that resemble, respectively, the hypercortisolemia and reduced monoamineregic neurotransmitter function evident in MDD.

Neural Circuits Affected by Mood Disorders

Neuroimaging, neuropathological, and lesion analysis data implicate brain networks that normally regulate the evaluative, expressive, and experiential aspects of emotional behavior in the pathophysiology of mood disorders. These networks include limbic–cortical–striatal–pallidal–thalamic (LCSPT) circuits, formed by the orbital and medial prefrontal cortex (OMPFC), amygdala, hippocampal subiculum, ventromedial striatum, mediodorsal thalamic nucleus, and ventral pallidum. The LCSPT circuits initially were related to emotional behavior on the basis of their anatomical connectivity with limbic structures that mediate emotional expression, such as the hypothalamus and periaqueductal gray (PAG). They initially were implicated in the pathophysiology of depression by the observations that degenerative basal ganglia diseases and lesions of the striatum and orbitofrontal cortex (OFC) increased the risk for MDEs. Because these conditions affect synaptic transmission through the LCSPT circuitry in diverse ways, it appears that dysfunction that alters transmission through these circuits in various ways can produce depressive symptoms.

The structural and functional brain abnormalities associated with mood disorders also implicate the LCSPT circuits and the extended anatomical network formed by neural projections linking the LCSPT components to the mid- and posterior cingulate cortex, superior and medial temporal gyrus, parahippocampal cortex, and midline thalamic nuclei. This extended 'visceromotor' network functions to regulate endocrine, autonomic, neurotransmitter, and behavioral responses to aversive and rewarding stimuli and contexts, by modulating neuronal activity within limbic and brain stem structures that mediate and organize emotional expression (e.g., amygdala, bed nucleus of the stria terminalis, hypothalamus, PAG). Impaired function within this network thus could disinhibit or alter emotional behavior, conceivably giving rise to the signs and symptoms of depression or mania. Compatible with this hypothesis, pharmacological, neurosurgical, and electrical stimulation treatments for MDEs appear to inhibit pathological activity within visceromotor network structures such as the amygdala and subgenual anterior cingulate cortex.

The visceromotor network also is implicated by neuroimaging studies of pathological anxiety syndromes that occur comorbidly with MDD and BD. Notably, antidepressant drugs are the first-line treatments for both depressive and anxiety disorders, suggesting that

inhibition of pathological limbic function may underlie treatment mechanisms for both syndrome types.

Brain Structural Abnormalities

Patients with mood disorders show abnormalities of brain structure in many visceromotor network regions. The extent or prevalence of these abnormalities depends on clinical characteristics, including age at illness onset, capacity for developing mania or psychosis, and evidence for familial aggregation of illness. For example, MDD patients with late-onset depression show an increased prevalence of neuroimaging and neuropathological correlates of cerebrovascular disease, relative both to age-matched, healthy controls and to elderly depressives with early-onset depression. Similarly, in MDD and BD cases in which there is psychosis (delusions and/or hallucinations) or late-life illness onset, there are nonspecific signs of atrophy, such as lateral ventricle enlargement, that are absent in early-onset, nonpsychotic cases.

Nevertheless, early-onset, nonpsychotic MDD and BD cases show volumetric abnormalities in some PFC, cingulate, temporal lobe, and striatal regions (Table 1). The most prominent abnormality reported to date is a 20–50% reduction in gray matter in the left anterior cingulate cortex ventral to the corpus callosum genu (i.e., 'subgenual; sgACC), which is evident in MDD and BD samples with familial clustering of illness or psychotic features. This reduction exists early in illness and in young adults at high familial risk for MDD. Gray-matter volume also is reduced in the OFC (BA11, BA47) and ventrolateral PFC (VLPFC; BA45, BA47) in MDD and BD, frontal polar/dorsal anterolateral PFC (BA9, BA10) in MDD, and posterior cingulate cortex and superior temporal gyrus in BD. Additionally, white matter is decreased in the corpus callosum genu in adults with MDD or BD and their high-risk child or adolescent female offspring, and also is decreased in the corpus callosum splenium in adults with MDD or BD.

Some studies also report reduced hippocampal volume in MDD, ranging in magnitude from 8% to 19%, although other studies have not replicated these differences. The discrepant results across studies may reflect clinical heterogeneity, as one study reported that this reduction was limited to depressed women who suffered early-life trauma, and others reported that hippocampal volume correlated inversely with time spent depressed. In addition, volumetric reductions in the anterior subiculum/ventral CA1 region appear specific to BD.

Conventional antidepressant drug treatment and symptom remission do not alter the reductions in gray-matter volume in sgACC, but may arrest further gray-matter decrements in hippocampus. In contrast, chronic lithium treatment, which exerts robust neurotrophic effects in animal models, has been associated with increasing sgACC gray-matter volume in treatment responders.

Table 1 Neuroimaging and histopathological abnormalities evident in the visceromotor network in early-onset, recurrent major depressive disorder and/or bipolar disorder

Brain region	*Gray-matter volume*	*Cell counts, cell markers*	*Glucose metabolism, CBF*	
	Dep vs. con	*Dep vs. con*	*Dep vs. con*	*Dep vs. rem*
Dorsal medial/anterolateral PFC (BA9)	↓	↓	↓	↑
Frontal polar C (BA10)		↓	↑	↑
Subgenual anterior cingulate C	↓	↓	↓/↑[a]	↑
Pregenual anterior cingulate C	↓	↓	↑	↑
Orbital C/ventrolateral PFC	↓	↓	↑	↑
Posterior cingulate	↓		↑	↑
Parahippocampal C	↓	↓BD	↑	↑
Amygdala	↓/↑[b]	↓MDD	↑	↑
Ventromedial striatum	↓		↑	↑
Hippocampus	↓	↓BD	n.s.	n.s.
Superior temporal G/temporopolar C	↓			↑
Medial thalamus			↑	↑

[a]In the subgenual anterior cingulate cortex, the apparent reduction in CBF and metabolism, seen in positron emission tomographic images of depressed patients, is thought to be accounted for by the reduction in tissue volume in the corresponding cortex, because after partial volume correction for the reduction in gray matter the metabolism appears increased relative to controls.

[b]The literature is in disagreement with respect to the amygdala volume in mood disorders. In major depressive disorders, the volume appears reduced in cases in which major depressive episodes show a chronic or intermit course.

C, cortex; CBF, cerebral blood flow; Dep vs. con, unmedicated depressives versus healthy controls; Dep vs. rem, unmedicated depressives versus themselves in either the medicated or unmedicated remitted phases; G, gyrus; n.s., differences generally not significant; PFC, prefrontal cortex. Empty cells indicate insufficient data.

Neurophysiological Imaging Abnormalities

Many regions where structural abnormalities are evident in mood disorders also contain abnormalities of cerebral glucose metabolism and cerebral blood flow (CBF) (Table 1 and Figure 1). In some visceromotor network structures the physiological activity appears increased in the depressed phase relative to the remitted phase of MDD, based upon longitudinal studies of depressed patients imaged before versus after treatment, and in remission versus during depressive relapse.

Nevertheless, the gray-matter volumetric reduction in some structures is sufficiently prominent to produce partial volume effects in positron emission tomographic (PET) images because of their low spatial resolution, yielding complex relationships between measured physiology and depression severity. For example, relative to controls, depressed MDD and BD patients show metabolic activity that appears reduced in sgACC. When the reduction in gray-matter volume in MDD and BD is taken into account by correcting the metabolic data for partial volume averaging effects, however, metabolism during depression instead appears increased in sgACC. The volumetric reductions in the OFC and VLPFC also may account for the complex relationships observed between PET measures of metabolism and illness severity, as metabolism appears elevated in depressed samples of mild-to-moderate severity, but reduced in more severe, treatment-refractory cases.

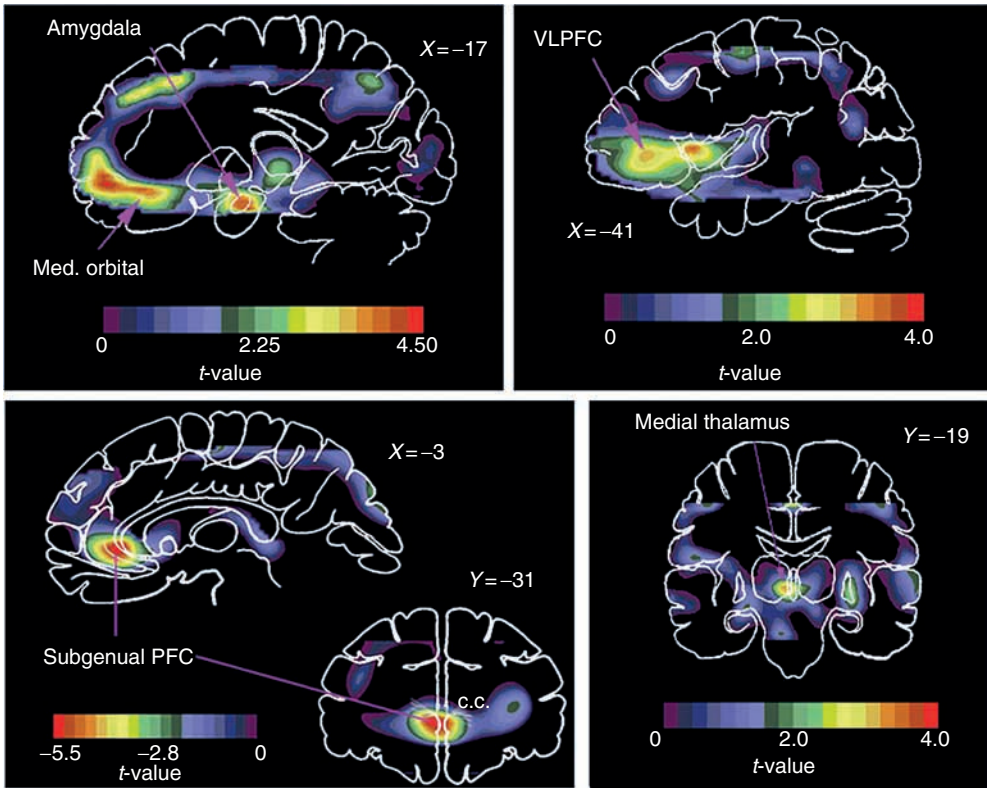

Figure 1 Areas of abnormally increased physiological activity in familial major depressive disorder, shown as images of unpaired *t*-values, which were computed using a statistical parametric mapping approach to compare activity between depressives and controls. Upper left: The positive *t*-values in this sagittal section at 17 mm left of midline ($x = -17$) show areas where cerebral blood flow is increased in the depressives in the amygdala and medial (MED) orbital cortex. Anterior is to the left. Upper right: Positive *t*-values in a sagittal section 41 mm left of midline ($x = -41$) show areas where cerebral blood flow is increased in the depressives in the left ventrolateral prefrontal cortex (VLPFC), lateral orbitofrontal cortex, and anterior insula. Lower right: Positive *t*-values in a coronal section 19 mm posterior to the anterior commissure ($y = -19$) show an area where cerebral blood flow is increased in the depressives in the left medial thalamus. Lower left: Coronal (31 mm anterior to the anterior commissure; $y = 31$) and sagittal (3 mm left of midline; $x = -3$) sections showing negative voxel *t*-values where glucose metabolism is decreased in depressives relative to controls. The reduction in activity in this prefrontal cortex (PFC) region located in the anterior cingulate gyrus ventral to the genu of the corpus callosum (i.e., subgenual) appeared to be accounted for by a corresponding reduction in cortex volume (Table 1). Anterior or left is to the left. Upper left: Reproduced from Price JL, Carmichael ST, and Drevets WC (1996) Networks related to the orbital and medial prefrontal cortex; a substrate for emotional behavior? *Progress in Brain Research* 107: 523–536. Upper right: Reproduced from Drevets WC, Gadde K, and Krishnan KRR (2004) Neuroimaging studies of depression. In: Charney DS and Bunney BJ (eds.) *The Neurobiological Foundation of Mental Illness*, 2nd edn., pp. 461–490. New York: Oxford University Press. Lower left: Reproduced from Drevets WC, Price JL, Simpson JR, et al. (1997) Subgenual prefrontal cortex abnormalities in mood disorders. *Nature* 386(6627): 824–827.

Regions where metabolism and depression severity correlate positively include the amygdala, sgACC, and ventromedial frontal polar cortex. Physiological activity decreases in these regions during effective antidepressant drug treatment or deep brain stimulation of the sgACC. Conversely, under experimental conditions involving serotonin or catecholamine depletion, metabolism increases in these regions in recovered MDD patients who experience depressive relapse.

In some regions, abnormalities of physiological activity appear specific to depressive subtypes. In the amygdala, for example, abnormal elevations of resting metabolism are evident in depressed samples categorized as having BD, familial pure depressive disease, or MDD-melancholic type. In such cases amygdala metabolism decreases during effective antidepressant treatment. In contrast, a broader range of depressed individuals show abnormal hemodynamic responses in the amygdala to emotional stimuli. For example, hemodynamic responses of the left amygdala were reduced in depressed children and adults while viewing fearful faces, and prolonged in duration in depressed adults while viewing sad words.

In the accumbens, medial thalamus, and posterior cingulate cortex, metabolism is abnormally elevated in the depressed phase of MDD and BD. In functional magnetic resonance imaging (fMRI) studies, regional hemodynamic responses are attenuated in the ventral striatum in reward-processing tasks, and in both the ventral striatum and posterior cingulate in tasks involving negative feedback.

Neuropathological Correlations

Postmortem studies have characterized the histopathological correlates of gray-matter volumetric abnormalities found in mood disorders. The reductions in sgACC and OFC volumes are associated with reductions in glia, synaptic markers, and neuron size, and elevations in neuron density. Reductions in glial cell counts, density, and markers (especially of oligodendroglia and their product, myelin basic protein), and in glia-to-neuron ratios, also were found in MDD in the pregenual anterior cingulate cortex (pgACC), frontal polar cortex (BA10), dorsal anterolateral PFC (BA9), and amygdala. The density of nonpyramidal neurons was decreased in the ACC and hippocampus in BD, and in the dorsal anterolateral PFC (BA9) in MDD. Reductions in synapses and synaptic proteins were evident in BD patients in the hippocampal subiculum/ventral CA1 region.

Correlations with Rodent Models of Chronic and Repeated Stress

In regions that appear homologous to areas where gray-matter reductions exist in depressed humans (i.e., medial PFC, hippocampus), repeated stress causes dendritic atrophy and reductions in glial cell counts or proliferation in rats. In the basolateral amygdala, chronic, unpredictable stress also produces dendritic atrophy, but chronic immobilization stress instead increases dendritic branching. This dendritic remodeling depends on interactions between the *N*-methyl-D-aspartate (NMDA) receptor stimulation and glucocorticoid secretion associated with repeated stress.

Dendritic atrophy would decrease the volume of the neuropil, which occupies most of the gray-matter volume. The similarities between the histopathogical changes accompanying stress-induced dendritic atrophy in rats and those found in depressed humans suggest that such a process may underlie gray-matter volume reductions in hippocampal and PFC structures in MDD and BD. In rats the stress-induced dendritic atrophy in the medial PFC is associated with impaired extinction of behavioral responses to fear-conditioned stimuli. Notably, the volumetric reductions in depressed humans are associated with greater severity and illness chronicity.

Neurochemical Systems Implicated in Depression

Mood disorders have been associated with abnormalities in serotonergic, dopaminergic, noradrenergic, cholinergic, glutamatergic, γ-aminobutyric acid (GABA)ergic, and glucocorticoid and peptidergic (e.g., corticotropin-releasing factor) function. The monoamineregic systems have received particular attention in research because most antidepressant drugs exert their primary pharmacological actions through these systems. Nevertheless, the delayed onset of antidepressant effects (about 3 weeks) after initiating monoamine reuptake-inhibiting agents suggests that secondary changes in gene expression underlie the therapeutic mechanisms of these drugs. Mechanisms hypothesized to serve as final common pathways for antidepressant responses include (1) increased expression of brain-derived neurotrophic factor and other neurotrophic/neuroprotective factors, (2) enhanced postsynaptic serotonin type 1A receptor (5-HT_{1A}R) function, and (3) reduced sensitivity or transmission of NMDA–glutamatergic receptors.

Serotonergic System

Selective serotonin reuptake inhibitors (SSRIs) exert antidepressant effects, and most other antidepressant drugs also increase 5-HT neurotransmission. This effect was thought to compensate for deficient serotonergic function in MDD, since the density and sensitivity of some 5-HT receptor types were reduced in

MDD. Compatible with this hypothesis, many remitted MDD patients who are unmedicated or are taking SSRI agents experience depressive relapse under tryptophan depletion (which putatively decreases central serotonergic function).

The most promising evidence for a serotonin system deficiency that is compensated by antidepressant pharmacotherapy involves postsynaptic 5-HT_{1A}R. Chronic administration of antidepressant drugs with diverse primary pharmacological actions enhances postsynaptic 5-HT_{1A}R function. Postsynaptic 5-HT_{1A}R binding in the insula, cingulate, parieto-occipital, and orbital/ventrolateral prefrontal cortices is abnormally decreased in most studies of MDD and BD. Depressed individuals also show blunted thermic and endocrine responses to 5-HT_{1A}R agonist challenge, suggesting that 5-HT_{1A}R function also is reduced.

It remains unclear whether reduced 5-HT_{1A}R function and expression constitute a neurodevelopmental or acquired abnormality in mood disorders. This issue is of importance because interruption of 5-HT_{1A}R function during neurodevelopment persistently alters the function of emotion-modulating systems in genetically engineered mice. Nevertheless, the reduction in 5-HT_{1A}R binding and expression in depression may arise secondarily to cortisol hypersecretion, as 5-HT_{1A}R mRNA expression and density are under tonic inhibition by glucocorticoid receptor stimulation. In experimental animals, the increased corticosterone secretion during repeated stress reduces 5-HT_{1A}R mRNA expression and density. Moreover, mood and anxiety-disordered subgroups with reduced 5-HT_{1A}R binding appear limited to those with a diathesis to hypersecrete cortisol.

Altered serotonin transporter (5-HTT) function may also play a role in the pathophysiology of the MDE. Neuroimaging and postmortem studies of MDD or BD report abnormalities of 5-HTT binding that differ in direction across regions. Moreover, genetic polymorphisms involving 5-HTT regulatory sites increase the vulnerability for developing mood disorders. For example, the common 'short' (s) allele of the functional 5-HTT promoter region length polymorphism increases the vulnerability for developing MDD within the context of stress.

Catecholaminergic Systems

The central dopaminergic and noradrenergic systems are also implicated in the pathophysiology of depression and the mechanisms of antidepressant drugs. Administration of reserpine or α-methyl-*para*-tyrosine, which reduces the availability of catecholamines, induces depressive syndromes in vulnerable individuals. Conversely, selective norepinephrine reuptake inhibitors (e.g., reboxetine), dopamine reuptake inhibitors (e.g., nomifensine), and dopamine receptor agonists (e.g., pramipexole) exert antidepressant effects in placebo-controlled studies. In MDD, cerebrospinal fluid (CSF) and jugular vein plasma levels of the dopamine (DA) metabolite, homovanillic acid, are abnormally decreased, consistent with reduced DA turnover. In contrast, levels of norepinephrine metabolite concentrations are decreased in some, but increased in other studies. Neuroimaging studies of MDD show reduced striatal DA transporter binding and [^{11}C]L-DOPA uptake across the blood–brain barrier, consistent with reduced DA neurotransmission. Finally, degeneration of DA neurons in Parkinson's disease increases vulnerability for MDE.

The mesolimbic DA projections from the ventral tegmental area (VTA) to the nucleus accumbens shell and medial PFC play major roles in learning associations between operant behaviors or sensory stimuli and reward, and in mediating the reinforcing properties of drugs of abuse and natural rewards. Thus, reduced DA neurotransmission in depression is hypothesized to underlie the anhedonia, amotivation, and psychomotor slowing seen in MDEs.

Cholinergic System

The cholinergic system has also been implicated in the pathophysiology of mood disorders, with evidence generally suggesting that the muscarinic cholinergic system is hyperresponsive in depression. Increasing cholinergic activity using physostigmine (an acetylcholine esterase inhibitor) results in rapid induction of depressive symptoms in currently manic bipolar patients, and in a worsening of symptoms in MDD patients. Depressed individuals also show exaggerated polysomnographic, neuroendocrine, and pupillary responses to cholinomimetic agents or muscarinic receptor agonists relative to healthy controls, and aspects of sleep abnormalities observed in depression, including decreased rapid eye movement (REM) latency and increased REM density, are thought to reflect muscarinic receptor hypersensitivity. These data contribute to the hypothesis that a hyperresponsive cholinergic system partially accounts for mood symptoms associated with mood disorders.

The preclinical literature also implicates the muscarinic receptor system in mediating the cholinergic effects on emotional behavior. In rats genetically bred for increased sensitivity of muscarinic receptors, cholinergic agonists enhance the behavioral analogs of depression, while antimuscarinic drugs improve these behaviors. Consistently, the muscarinic cholinergic receptor antagonist, scopolamine, exerts rapid and robust antidepressant effects in depressed MDD and BD patients.

Several studies more specifically implicate the M_2 receptor (M_2R) in mood regulation and depression. Multiple M_2R gene polymorphisms have been associated with increased risk for developing an MDE. In humans, administration of the M_2R antagonist, procaine, elicits emotional responses ranging from sadness, fear, and severe anxiety to euphoria, and results in physiological activation of the cingulate cortex, which receives dense cholinergic projections. A PET neuroreceptor imaging study demonstrated decreased M_2R binding in the cingulate cortex in BD.

In experimental animals, excessive cortical cholinergic activity results in 'hyperattentive impairments', whereby overactivity of acetylcholine (ACh) reduces the threshold for signal detection and leads to overprocessing of stimuli. Given the central role of ACh in emotional and sensory processing (see later), the overactivity of the cholinergic system in mood disorders may result in the overrepresentation (through altered signal/noise processing) of emotionally laden information, conceivably producing the emotional processing bias and correlated cognitive deficits associated with MDD and BD.

Glutamatergic and GABAergic Systems

The major excitatory and inhibitory neurotransmitters, glutamate and γ-aminobutyric acid, respectively, are also implicated in mood disorders. Early studies reported abnormal reductions in plasma and CSF GABA concentrations in MDD. In contrast, postmortem studies of the NMDA receptor complex found evidence of increased glutamatergic transmission in the PFC, and of altered NMDA and metabotropic glutamatergic receptors in depression and suicide. Moreover, antidepressant and mood-stabilizing drugs which have diverse primary pharmacological actions generally reduce NMDA receptor sensitivity and/or transmission, and many of these agents also increase GABA levels or transmission.

Elevated glutamatergic function is thought to support the neurophysiological activation of visceromotor circuits in depression. Anatomical connections between the PFC, thalamus, and amygdala implicated in mood disorders are formed by predominantly excitatory neurons. Cerebral glucose metabolism predominately reflects the energetic requirements associated with glutamatergic transmission, so the elevated metabolism in limbic–thalamic–cortical circuits during MDEs implies that glutamatergic neurotransmission is increased in these circuits. During effective antidepressant drug or electroconvulsive therapy, metabolic activity decreases in these regions (Table 1), compatible with evidence that these treatments result in desensitization of NMDA receptors in the PFC. Elevated glutamatergic transmission within discrete anatomical circuits may explain the targeted nature of gray-matter changes within mood disorders (e.g., affecting left, but not right, sgACC), as the pathogenesis of dendritic atrophy during repeated stress depends partly on NMDA receptor stimulation.

Magnetic resonance spectroscopic (MRS) studies of MDD also demonstrate abnormalities of glutamate (measured together with glutamine as the combined 'Glx' peak in MRS spectra) and GABA concentrations in depressed patients. These MRS spectra reflect the combined intracellular and extracellular pools of glutamate, glutamine, and GABA, but are dominated overwhelmingly by intracellular pools. Depressed MDD patients show abnormally reduced GABA levels in the dorsomedial/dorsal anterolateral PFC. The majority of the GABA pool exists within GABAergic neurons, so the reduction in GABA in the dorsal anterolateral PFC appears compatible with reports of reduced GABAergic neurons in this region (BA9) in MDD. Similarly, the abnormally reduced Glx levels in the dorsomedial/dorsal anterolateral and ventromedial PFC in MDD are compatible with histopathological evidence of reduced glia in these regions, as glia play prominent roles in glutamate–glutamine cycling.

Glucocorticoid System

The hypothalamic–pituitary–adrenocortical (HPA) axis appears overactive in depression. Thus, severe depression is associated with hypersecretion of cortisol, with pituitary and adrenal gland enlargement, and with CSF–CRF levels that are increased to an extent that is inappropriate to plasma cortisol levels (i.e., implying negative feedback systems are impaired, and/or excessive central drive on release of CRF or other adrenocorticotropic hormone secretagogues). Depressed patients also show blunted adrenocorticotropic hormone (ACTH) responses to CRF *in vivo*, reduced CRF receptor density in the PFC, and increased corticotropic cell size and mRNA levels in the pituitary gland postmortem, indicating chronic HPA axis activation. Some depressives show reduced sensitivity to dexamethasone negative feedback and/or to glucocorticoid 'fast feedback.' During successful – but not unsuccessful – treatment, serum cortisol and CSF–CRF levels normalize, and cortisol secretion becomes suppressible by dexamethasone. The chronic elevation of cortisol secretion in depression is thought to contribute to mitochondrial dysfunction, neuropathological changes, elevated body temperature, premature osteoporosis and aging, and other medical morbidity.

Autonomic Nervous System Function in Mood Disorders

Relative to nondepressed controls matched for cardiac disease, depressed patients show a two- to fourfold increase in sudden death and a sevenfold increase in ventricular arrhythmia. This excess risk of cardiac events in MDD is hypothesized to result from elevated noradrenergic and sympathetic autonomic function, coupled with reduced parasympathetic tone on the heart rate.

Disturbances in Sleep Physiology

Clinical complaints of sleep regulation including insomnia, hypersomnia, and restless sleep are among the most common features of an MDE. Polysomnographic studies of sleep architecture during depression demonstrate reduced latency to the onset of rapid eye movement sleep, increased density of REM sleep, reduced time spent in slow-wave sleep, and decreased sleep efficiency (i.e., less time spent asleep due to increased nocturnal awakenings). Some of these abnormalities are evident early in the illness course, and also are evident in female children at high familial risk for mood disorders.

Alterations of Cognitive and Emotional Processing in Mood Disorders

Cognitive and neuropsychological impairments are characteristics of MDE, and were incorporated into the diagnostic criteria for MDD as "an impaired ability to think or concentrate." Nonetheless, behavioral studies designed to characterize these symptoms in mood disorders have produced discrepant results. Some studies reported wide-ranging deficits in mood disorders that include impairments in early information processing, attention, memory, and executive functions, although other studies failed to identify such deficits. Several factors may contribute to the discrepancies in the literature, including heterogeneous patient groups, medication status of patients, and differences in task paradigms that theoretically assess the same cognitive domain. Nonetheless, sufficient evidence points consistently to the presence of cognitive impairment in MDD and BD. Moreover, residual deficits in remitted MDD and BD patients demonstrate that some cognitive features occur independently of mood state.

Motor Function

The diagnostic criteria for MDEs also include psychomotor agitation or retardation, but these symptoms manifest as prominent clinical features in a minority of depressed cases. Most studies of unmedicated, depressed patients do not find increased response latency on most cognitive tasks. Some studies report psychomotor slowing on tasks such as pattern recognition and match-to-sample paradigms, and others report slowed response latency in unmedicated, remitted patients with a history of seasonal depression. While psychomotor slowing may affect a minority of depressed patients, the effects of motor function on performance measures must be considered when interpreting behavioral changes as cognitive effects in mood disorders.

Early Information Processing

Impairments in early information processing are evident in patients with mood disorders. Inspection time, measured as the minimum duration of stimulus presentation necessary for near-perfect performance on a two-choice visual discrimination task, assesses the speed of early information processing independent of motor speed or cognitive strategy. Inspection time is found to be longer in unmedicated MDD patients than in both medicated MDD and healthy controls. In a visual backward masking task, whereby a stimulus is shown briefly and then masked using a noninformative stimulus, medicated BD patients had deficits in the identification of masked stimuli consistent with an early information processing impairment. Thus, evidence suggests that early information processing is altered in mood disorders.

Emotional Processing Bias

A consistently reported finding in mood disorders is the presence of a mood-congruent processing bias, defined as a bias to process negative information as compared to positive or neutral information. In memory studies, depressed patients recall more negatively toned than positively toned information. In the context of attention paradigms, depression-related negative words produce more interference on emotional Stroop tasks than do happy or neutral words. In an affective attention-shifting task (i.e., affective go/no-go), medicated and unmedicated depressed patients are faster when responding to the presentation of sad word targets, as compared to happy word targets, consistent with the idea that sad words are processed more readily. In a face dot-probe task designed to assess the allocation of attention between a face with a neutral expression and one with an emotional expression, individuals with MDD consistently attend preferentially to faces with sad expressions. Finally, medicated depressed patients are more negative in their interpretation of ambiguous words and ambiguous situations than are healthy participants. Thus, the

literature suggests that a bias exists in stimulus processing that may result in a preferential representation of negatively toned information.

Results from functional neuroimaging studies are consistent with behavioral findings, showing altered neurophysiological responses in brain regions that process emotional information during tasks that utilize emotional stimuli. The amygdala, a brain region central to the processing of emotional stimuli, shows elevated levels of activity in MDD during exposure to sad faces. Similarly, the amygdala shows increased activity during the presentation of masked fearful and sad faces in MDD patients versus healthy controls; and lateral OFC shows higher activity to negative emotional stimuli in an affective go/no-go task. Similar increases in amygdala activity are observed in response to sad faces in remitted patients with MDD, suggesting that this abnormality is traitlike in MDD. In another study, healthy volunteers showed a linear increase in activity in fusiform visual cortex and ventral striatum as the intensity of a happy facial expression increased, while MDD patients showed a similar pattern as the intensity of a sad expression increased, but the pattern extended into hippocampus and amygdala in patients. This literature suggests that neural responses within regions underlying emotional processing are altered, and these altered responses may represent the neurophysiological correlates to the emotional processing bias observed in mood disorders.

Attention, Memory, and Learning

Subjective impairments of concentration and attention are among the most commonly reported symptoms and sources of functional impairment associated with MDEs. Although disagreement remains, traditional neuropsychological test batteries with standard measures of attention and memory have failed to identify clear, consistent patterns of deficits in MDD and BD, and this likely is due partially to the lack of tasks designed specifically for this clinical population. In set-shifting tasks, including Trail Making A and digit symbol substitution, medicated depressed patients reportedly perform more poorly than do controls, requiring more time to perform the tasks and making more errors; other studies report performance on set shifting to be spared in medicated depressed patients. The effect is more pronounced when using emotionally toned material, as in the aforementioned affective attention-shifting task. Selective attention and working memory also are reported to be impaired in depression, but again these findings are inconsistent, and the results are complicated further by medication status during study participation. Verbal memory deficits in currently depressed patients and in unaffected twin and nontwin siblings of BD patients have been shown, raising the possibility that impaired verbal memory may constitute an endophenotype for BD. Impairments of spatial recognition memory and delayed match-to-sample tasks have been reported in medicated MDD and BD patients, but not in unmedicated MDD or BD patients.

Executive Function

Studies assessing executive functions, which include domains such as decision making, planning, and problem solving, also produce conflicting findings in the literature. While complicated by diagnostic heterogeneity and medication status, most studies that utilize the Wisconsin Card Sorting Task and the Tower of London task, which require planning and problem solving, report deficits in MDD. Other tasks that assess executive function, such as Trail-Making Part B and Controlled Oral-Word Association tests, produce more variable results. However, these tasks are more attention based and potentially are less demanding tests of executive function, and thus may not be sufficiently sensitive to identify deficits in mood disorders.

Implications for Circuit-Based Models of Depression

The neuropathological, neurophysiological, and neurochemical abnormalities extant within the extended visceromotor network may impair this network's modulation of neuroendocrine, autonomic, neurotransmitter, behavioral, and cognitive responses to aversive and reward-related stimuli or contexts, potentially accounting for disturbances within these domains in mood disorders. For example, as depicted in Figure 2, the basolateral amygdala sends anatomical projections to the central nucleus of the amygdala (ACe) and the bed nucleus of the stria terminalis (BNST), and these structures send projections to the hypothalamus, PAG, nucleus basalis, locus coeruleus, raphe, and other diencephalic and brain stem nuclei, which together mediate the neuroendocrine, neurotransmitter, autonomic, and behavioral responses to stressors and emotional stimuli. The OMPFC sends to each of these structures, as well as to the amygdala, overlapping projections, which function to modulate each component of emotional expression. The neuropathological changes evident in the OMPFC in mood disorders may thus impair the PFC's modulatory role over emotional expression, disinhibiting or dysregulating limbic responses to stressors and emotional stimuli.

Reciprocal, functional relationships between structures that mediate emotional expression (e.g., amygdala) and those which modulate emotional expression (e.g.,

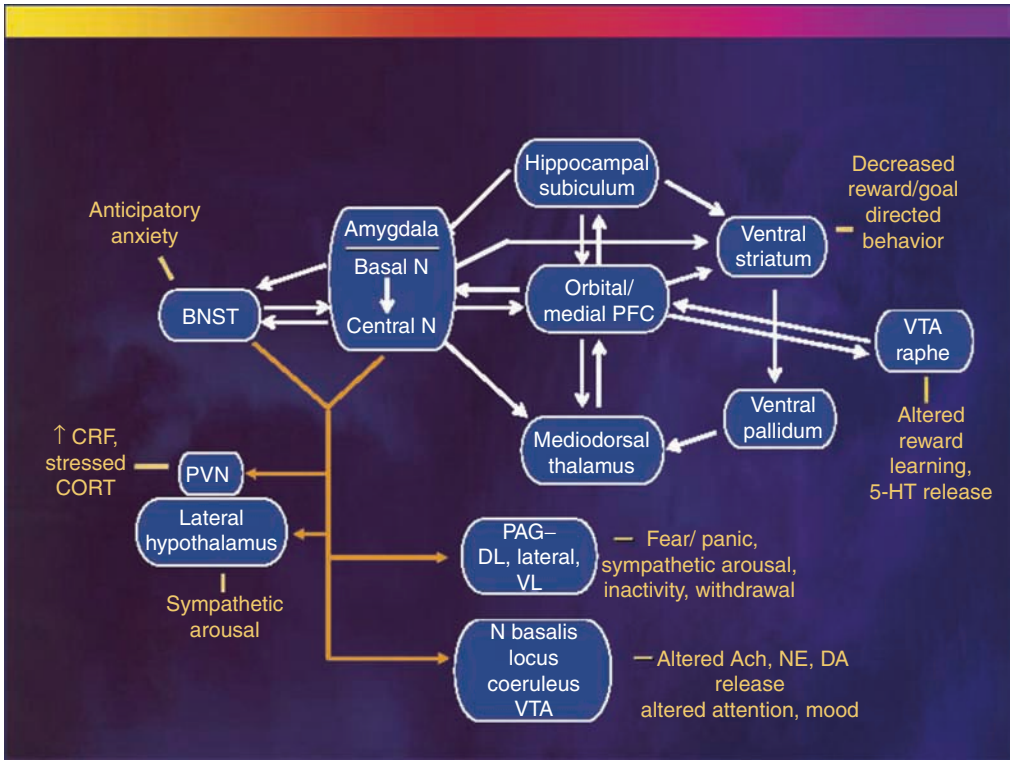

Figure 2 Anatomical circuits involving the orbitomedial prefrontal cortex (PFC) and amygdala, reviewed within the context of a model in which orbitomedial PFC dysfunction results in disinhibition of limbic activity through the amygdala, yielding the emotional, cognitive, endocrine, autonomic, and neurochemical manifestations of depression. The basolateral amygdala sends efferent projections to the central nucleus of the amygdala and the bed nucleus of the stria terminalis (BNST). The efferent projections from these structures to the hypothalamus, periaqueductal gray (PAG), nucleus basalis, locus coeruleus, raphe, and other diencencephalic and brain stem nuclei then organize the neuroendocrine, neurotransmitter, autonomic, and behavioral responses to stressors and emotional stimuli. The orbitomedial PFC shares reciprocal projections with all of these limbic, hypothalamic, and brain stem structures (although only the connections with the amygdala are shown), which function to modulate each component of emotional expression. Impaired orbitomedial PFC function may thus disinhibit or dysregulate the limbic outflow of the central nucleus of the amygdala and BNST. Solid white lines indicate some of the major anatomical connections between structures, with arrowheads indicating the direction of projecting axons. Solid yellow lines show efferent pathways of the central nucleus of the amygdala and BNST, which generally are monosynaptic, but in some cases are bisynaptic connections. 5-HT, 5-hydroxytryptamine (serotonin); ACh, acetylcholine; CORT, cortisol; CRF, corticotropin-releasing factor; DA, dopamine; DL, dorsolateral column of PAG; N, nucleus, NE, norepinephrine; PVN, paraventricular nucleus of the hypothalamus; VL, ventrolateral column of PAG; VTA, ventral tegmental area.

parts of OMPFC) are evidenced by electrophysiological and lesion analysis studies in experimental animals. In rats, stimulation of projections from the amygdala to the medial prefrontal cortex (mPFC) inhibits neuronal ensemble activity in mPFC, and stimulation of projections from the mPFC to the amygdala excites GABAergic projections, which inhibit neuronal activity in ACe. Moreover, rats exposed to fear-conditioned stimuli show reduced mPFC neuronal firing activity, and the magnitude of this decrement correlates inversely with the corresponding increase in amygdala neuron activity and fear behavior. Conversely, lesions of the right or bilateral mPFC enhance behavioral, sympathetic, and endocrine responses to stressors or fear-conditioned stimuli.

Reciprocal relationships between the effects of OMPFC and amygdala function on emotional behavior also are evident in humans. In MDD, depression severity correlates positively with amygdala and sgACC activity, but negatively with activity in the left VLPFC/lateral OFC. Lesions of the left OFC increase the risk for an MDE, whereas electrical stimulation of the amygdala can produce fear, anxiety, dysphoria, sympathetic autonomic arousal, social withdrawal, and cortisol release in humans and/or nonhuman primates. Drugs and deep brain stimulation, which reduce activity in the amygdala and/or sgACC, exert antidepressant effects, while interpersonal psychotherapy enhances activity in the VLPFC/anterior insula and can reduce depressive symptoms in MDD. Finally, higher activity in the pgACC holds positive prognostic significance in MDD, as depressives who improve during antidepressant treatment show higher pgACC metabolism and

electrophysiological activity than do treatment-nonresponsive cases and healthy controls. Taken together, these data suggest that during MDEs, activation of some OMPFC areas reflects a compensatory response which modulates depressive symptoms, and impaired function of these regions results in more severe and treatment-refractory illness.

Impairment of OMPFC function that disinhibits amygdala activity may also contribute to the neuroendocrine, autonomic, neurotransmitter, attention, and reward-seeking abnormalities associated with MDEs. The amygdala mediates the stressed component of glucocorticoid hormone secretion by disinhibiting CRF release from the hypothalamic paraventricular nucleus. Conversely, the glucocorticoid response to stress is inhibited by stimulation of glucocorticoid receptor concentrations in the ventral ACC, such that lesioning this cortex in rats increases ACTH and cortisol (CORT) secretion during restraint stress. Excessive amygdala activity combined with reduced sgACC gray matter may thus contribute to the excessive cortisol response to stress seen in depressed humans.

The autonomic imbalance associated with depression could also reflect OMPFC dysfunction coupled with excessive amygdala activity. Amygdala stimulation of the locus coeruleus, lateral hypothalamus, and PAG increases sympathetic autonomic arousal. In contrast, parasympathetic tone on the heart rate is partly regulated by projections from infralimbic cortex (a putative homolog of the human posterior sgACC) to the nucleus tractus solitarius of the vagus nerve, and lesions of this cortex reduce parasympathetic tone on the heart. The effects of increased amygdala activity and reduced infralimbic cortex function conceivably could account for the increased ratio of sympathetic-to-parasympathetic tone on the heart rate seen in depression.

Dysfunction of the OMPFC coupled with amygdala hyperactivity may also contribute to the anhedonia, amotivation, and inattention manifest in depression. Cognitive and behavioral response patterns supporting goal- or reward-directed activity are thought to depend upon PFC inputs to the ventral striatum, which are maintained or 'gated' by projections from the hippocampal subiculum to the ventral striatum. Stimulation of projections from the amygdala to the ventral striatum interrupts the maintenance of these cortically driven response patterns, providing an adaptive, limbic override that interrupts goal-directed thought or behavior in response to unexpected threats or novelty. Nevertheless, in MDEs, excessive amygdala activity conceivably may interfere with cognitive performance and goal-directed behavior to an extent that is maladaptive.

Summary

Convergent results from studies applying neuroimaging, lesion analysis, and postmortem techniques support models in which the signs and symptoms of MDEs emanate from dysfunction within an extended visceromotor network, which interferes with this system's modulatory role in emotional behavior. Antidepressant therapies may compensate for this dysfunction by attenuating the pathological limbic activity which putatively mediates depressive symptoms, and by increasing expression of neurotrophic/neuroprotective factors that preserve OMPFC function. Cognitive–behavioral strategies for managing depressive symptoms may instead rely upon enhancing the function of OMPFC corticolimbic circuits to facilitate their role in modulating emotional behavior.

See also: Anxiety Disorders; Bipolar Disorder; Chronic (Repeated) Stress: Consequences, Adaptations; Emotional Disorders: Treatment; Emotional Hormones and Memory Modulation; Emotional Influences on Memory and Attention; Stress and Suicide; Stress Response: Genetic Consequences; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness; Suicide.

Further Reading

Austin MP, Mitchell P, and Goodwin GM (2001) Cognitive deficits in depression: Possible implications for functional neuropathology. *British Journal of Psychiatry* 178: 200–206.

Carney RM, Freedland KE, and Veith RC (2005) Depression, the autonomic nervous system, and coronary heart disease. *Psychosomatic Medicine* 67(supplement 1): S29–S33.

Davis M and Shi C (1999) The extended amygdala: Are the central nucleus of the amygdala and the bed nucleus of the stria terminalis differentially involved in fear versus anxiety? *Annals of the New York Academy of Sciences* 877: 281–291.

Detera-Wadleigh SD and McMahon FJ (2004) Genetic association studies in mood disorders: Issues and promise. *International Review of Psychiatry* 16(4): 301–310.

Drevets WC and Price JL (2005) Neuroimaging and neuropathological studies of mood disorders. In: Licinio WM and Wong M-L (eds.) *Biology of Depression: From Novel Insights to Therapeutic Strategies*, pp. 427–466. Weinheim: Wiley-VCH Verlag GmbH & Co.

Drevets WC and Todd RD (2005) Depression, mania and related disorders. In: Rubin E and Zorumski C (eds.) *Adult Psychiatry,* 2nd edn., pp. 91–129. Oxford: Blackwell Publishing.

Everitt BJ and Robbins TW (1997) Central cholinergic systems and cognition. *Annual Review of Psychology* 48: 649–684.

Furey ML and Drevets WC (2006) Antidepressant efficacy of the antimuscarinic drug scopolamine: A randomized, placebo-controlled clinical trial. *Archives General Psychiatry* 63(10): 1121–1129.

Furey ML, Pietrini P, and Haxby JV (2000) Cholinergic enhancement and increased selectivity of perceptual processing during working memory. *Science* 290(5500): 2315–2319.

Gold PW, Drevets WC, and Charney DS (2002) New insights into the role of cortisol and the glucocorticoid receptor in severe depression. *Biological Psychiatry* 52(5): 381–385.

Goto Y and Grace AA (2005) Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. *Nature Neuroscience* 8(6): 805–812.

Herman JP and Cullinan WE (1997) Neurocircuitry of stress: Central control of the hypothalamo-pituitary-adrenocortical axis. *Trends in Neurosciences* 20(2): 78–84.

Izquierdo A, Wellman CL, and Holmes A (2006) Brief uncontrollable stress causes dendritic retraction in infralimbic cortex and resistance to fear extinction in mice. *Journal of Neuroscience* 26(21): 5733–5738.

Janowsky DS and Overstreet DH (1995) *The Role of Acetylcholine Mechanisms in Mood Disorders*. New York: Raven Press.

Krystal JH, Sanacora G, Blumberg H, et al. (2002) Glutamate and GABA systems as targets for novel antidepressant and mood-stabilizing treatments. *Molecular Psychiatry* 7(supplement 1): S71–S80.

LeDoux J (2003) The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology* 23(4–5): 727–738.

López JF, Chalmers DT, Little KY, et al. (1998) A.E. Bennett Research Award. Regulation of serotonin1A, glucocorticoid, and mineralocorticoid receptor in rat and human hippocampus: Implications for the neurobiology of depression. *Biological Psychiatry* 43(8): 547–573.

Manji HK, Drevets WC, and Charney D (2001) The cellular neurobiology of depression. *Nature Medicine* 7(5): 541–547.

Mayberg HS, Lozano AM, Kennedy SH, et al. (2005) Deep brain stimulation for treatment-resistant depression. *Neuron* 45(5): 651–660.

McEwen BS and Magarinos AM (2001) Stress and hippocampal plasticity: Implications for the pathophysiology of affective disorders. *Human Psychopharmacology* 16(S1): S7–S19.

Nestler EJ and Carlezon WA, Jr. (2006) The mesolimbic dopamine reward circuit in depression. *Biological Psychiatry* 59(12): 1151–1159.

Ottowitz WE, Dougherty DD, and Savage CR (2002) The neural network basis for abnormalities of attention and executive function in major depressive disorder: Implications for application of the medical disease model to psychiatric disorders. *Harvard Review of Psychiatry* 10(2): 86–99.

Phillips ML, Drevets WC, Rauch SL, et al. (2003) Neurobiology of emotion perception II: Implications for major psychiatric disorders. *Biological Psychiatry* 54(5): 515–528.

Sarter M, Nelson CL, and Bruno JP (2005) Cortical cholinergic transmission and cortical information processing in schizophrenia. *Schizophrenia Bulletin* 31(1): 117–138.

Stockmeier CA (2003) Involvement of serotonin in depression: Evidence from postmortem and imaging studies of serotonin receptors and the serotonin transporter. *Journal of Psychiatric Research* 37(5): 357–373.

Sullivan PF, Neale MC, and Kendler KS (2000) Genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry* 157(10): 1552–1562.

Swaab DF, Bao AM, and Lucassen PJ (2005) The stress system in the human brain in depression and neurodegeneration. *Ageing Research Review* 4(2): 141–194.

Tavares JV, Drevets WC, and Sahakian BJ (2003) Cognition in mania and depression. *Psychological Medicine* 33(6): 959–967.

Development of Behavior

K A French and A Marin-Burgin, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Anyone who watches the development of behavior in a human infant during his or her first 2 years of life must be struck by the enormous changes that take place during this time. Most newborn humans are capable of producing a number of behaviors: they can suck, they can cry, and they can move their arms and legs, albeit rather randomly. Then, during the first 2 years of life their behavioral repertoire greatly increases in breadth, skillfulness, and focus. However, the behavior of a human infant does not suddenly begin at birth. Embryonic and fetal animals are capable of a range of behaviors beginning well before birth. For example, human embryos make movements and respond to sounds when they are still *in utero*, chick embryos move and even make noises before they hatch out of the shell, and fly embryos wiggle and squirm before emerging from the egg case. In some species, these behaviors play a critical role in the individual's survival: to survive, a baby chick must actively break out of its shell, and components of hatching behavior arise many days before the chick would normally make its well-coordinated escape. In other cases, the ability to produce particular behaviors may not ever be used under ordinary conditions: For example, human infants who are born very prematurely respond behaviorally to light although a baby would never experience this stimulation *in utero*.

Factors That Shape the Development of Behavior

The progression of behavioral development depends on a number of different, but interacting, factors (Figure 1).

Behavior Depends on the Nervous System: As the Nervous System Develops, Behavior Changes

Behavior is the observable consequence of activity within the nervous system of an individual animal, regardless of the type of animal. The individual can produce only behaviors that are initiated, coordinated, or modulated by neurons, so behavioral development reflects the changing state of the nervous system. In many animals, including humans, the nervous system contains at birth most of the nerve cells it will ever have, yet many newborn animals produce a very small behavioral repertoire compared with adults of the same species. What is different between the two stages? It is the complexity of connections among the neurons. As a result, observing the progress of behavioral development is, in a very real sense, observing the progress of 'wiring up' the neurons in the nervous system. Once neurons have been generated, which typically happens quite early in development, synaptogenesis – the formation of the synapses that connect neurons with one another – is the major developmental focus, and executing this phase of development precisely and appropriately exerts an overwhelming effect on the future behavior of an animal. In humans, the first year of life is a period of active synaptogenesis and of refining synaptic connections, making a stimulating environment particularly important at this time.

As the nervous system develops, some behaviors arise and are then maintained into adulthood, although the detailed nature of the behavior may change (human babies usually crawl before they walk, and they walk before they run). Other behaviors appear at one stage of development and later disappear (human infants make sucking motions when they are touched on the cheek, but human children rarely exhibit this response). Perhaps the most dramatic examples of this second category of behaviors are seen in animals that metamorphose in their normal life cycle. In these animals, a single individual progresses through stages, each of which is so different from the others that they appear to be entirely different organisms with different methods of feeding, breathing, locomoting, and acquiring food. In such an individual, the nervous system undergoes major remodeling as the animal makes each transition: some neurons are retained in approximately the same form, some neurons die, some are retained but are strikingly modified, and some new neurons are added. The net effect is a single nervous system that is appropriately matched to very different lifestyles as the individual progresses through its life cycle. For example, a tobacco hornworm moth begins life as a leaf-eating caterpillar that crawls on plants, it then becomes a burrowing pupa, and finally, as an adult, it acquires wings and begins to fly. Similarly, a frog starts as an aquatic tadpole that swims and eats algae, but after metamorphosis it is a terrestrial animal that walks and hops and hunts the invertebrates that make up its carnivorous diet.

Synaptic Connections Are Established in General; Then Later They Are Refined Based on Neuronal Activity

Synapses can be divided into two major types: electrical synapses, in which the two neurons are connected

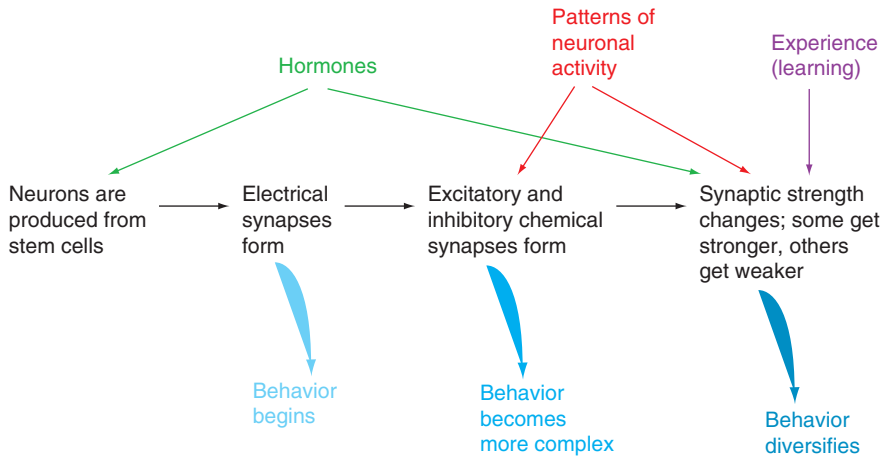

Figure 1 Behavior arises and matures as the nervous system develops. Development is affected by events within the nervous system and by signals from other parts of the body and in the environment.

via protein complexes called gap junctions, and chemical synapses, in which a signal between the two neurons is carried by specific molecules. Electrical synapses are nearly always excitatory; that is, when one of the neurons is excited, the other one becomes excited, too. Chemical synapses may be either excitatory or inhibitory, depending on the nature of the neurotransmitter and the receptors and channels in the postsynaptic cell. In many locations, it appears that neurons are initially connected by gap junctions, which form transiently. Subsequently, under the control of signals not yet understood, some of these electrical synapses are replaced by chemical synapses. Characteristic behavior may be generated during even the earliest stages when neurons are connected only through electrical synapses.

In many systems, construction of synapses begins with a stage in which connectivity is broader and more diffuse than the eventual adult pattern of the connections, and in many instances activity within the nervous system plays an important role in fine-tuning the connections. This increase in specificity among connections improves the skill, coordination, and precision of behavior. Some of these processes begin before birth; for example, in a newborn kitten, neurons in the retina of the eye are connected with neurons in the brain in an approximation of the eventual adult pattern, but once the kitten's eyes open and the circuitry is actively used, the connections are modified based on that activity until the final adult pattern is produced. During the process, the kitten's visually guided behavior becomes more precise; if the process is interrupted by occluding vision in one or both eyes, the adult pattern fails to develop and visually guided behavior remains compromised.

Hormones Modify Some Parts of the Developing Brain

Hormonal signals can exert very strong effects on the development of behavior, beginning as early as embryonic or fetal life. One particularly striking example is the development of sex and reproductive behavior in vertebrates. Sexually reproducing animals typically belong to one of two categories that are usually called 'male' and 'female,' and in many species the two sexes display a specific range of behaviors that are related to their reproductive roles. The sex of an individual is typically determined genetically, which might be thought of as sex at the cellular level. However, the development of reproductive behaviors, which might be thought of as sex at the organismic level, is subject to control by hormonal signals. For example, in birds that sing as part of courtship and territorial defense, it is typically the males that produce the song, whereas the females remain silent. In one species, the zebra finch, a genetically female bird can be induced to sing by treating her with exogenous gonadal hormone early in development and then again once she has reached adulthood. Both treatments are required; neither suffices alone, suggesting that the development of the brain is permanently shaped by the bird's hormonal status early in life.

Learning Changes Behavior

Behavior is shaped by the processes of learning; that is, behavior changes based on an individual's own experience. As it acquires feedback from its environment and from the effects of its own behavior, a human baby increases its skill in a vast array of

actions from focusing on objects within its visual field to reaching for an object, producing vocal sounds, or executing the movements of crawling or walking. The physiological basis of learning is thought to depend at least in part on modifications in the synaptic connections among neurons, so the modification of behavior based on experience is one subset of the larger collection of changes in connectivity that take place during the development of behavior. Interestingly, experiments in a variety of mammals suggest that the shaping of behavior based on learning begins well before birth. Many mammals, including humans, emerge from the uterus already able to recognize their mother based on her smell or the sound of her vocalizations. Thus, neuronal activity is subject to the guidance of experience from surprisingly early in development.

An Individual's Responsiveness to Specific Stimuli Varies during Development

In many instances, there are particular developmental stages during which the nervous system – and consequently developing behavior – is unusually and exquisitely sensitive to environmental conditions. Such a period is called a 'critical period' or 'sensitive period.' In many animals, including humans, exposure to ethanol (e.g., alcoholic beverages ingested by a pregnant woman) can produce babies with a constellation of physical and behavioral abnormalities called fetal alcohol syndrome. This condition is largely irreversible; exposure to ethanol before birth produces behavioral abnormalities that cannot be remedied after birth. However, critical periods are not just important in abnormal development. Many normal behaviors are also established during a critical period: young male zebra finches must be exposed to the song of an adult male during a particular window of time or they will never sing a normal song characteristic of the species. Also, young ducks and geese become attached to (i.e., imprinted on) the first moving object that they see after they hatch (normally their mother). This attachment lasts a lifetime: not only do these imprinted birds follow that object as chicks but also, in adulthood, males will court objects like the one on which they imprinted as chicks.

Behavior as a Window into the Developing Nervous System

Typically, behavior is produced in response to some kind of sensory input, so a simple reflex arc is frequently used to illustrate how behaviors arise: a sensory stimulus causes sensory neurons to become active; that activity is transmitted into the central nervous system, where it is processed by interneurons; interneurons cause activity in motor neurons, which control the contraction of muscles; and contractions of muscles cause parts of the body to move (i.e., cause behavior). A nervous system can execute this transmission of information along a chain of neurons only if during development sensory neurons found and constructed synapses onto interneurons, interneurons found and constructed synapses onto motor neurons, and motor neurons found and constructed synapses onto appropriate muscles.

Although in the definitive circuitry the transmission of information begins at sensory receptors and ends at motor neurons, which affect muscles, the process of constructing the pathway may follow a different logic. In several species, the first neurons to become connected are the motor neurons, and behavior begins as soon as these connections form. This developmental pattern is reflected in the onset and nature of the first observable behaviors, and thus it illustrates how observing changes in behavior during development can provide insight into underlying developmental processes in the nervous system.

Motor Behaviors

Several experiments in different species indicate that the earliest muscle contractions are driven by input from motor neurons before those neurons receive input from sensory receptors. Some of the first experiments addressing this question were carried out in chick embryos while they were still in the egg. Chicks produce movements of the head and legs prior to hatching, and these experiments were designed to reveal the source of embryonic movement. The dorsal spinal cord (which contains the sensory neurons) was surgically removed, leaving the ventral (motor) region of the spinal cord intact. Following this surgery, the embryos moved their legs normally, indicating that in chicks at that developmental stage, sensory input was unnecessary for generating these normal motor behaviors. Similarly, when all sensory neurons were genetically ablated from *Drosophila* (fruit fly) embryos and larvae, a similar result was obtained: embryonic or larval flies performed normal behavior – that is, peristaltic movements of the body – in the absence of any sensory input. These results, together with other observations, suggest that the neuronal circuitry responsible for the first spontaneous behaviors in a variety of species normally develops in the complete absence of sensory input.

However, in the long term, effective behavior must produce an appropriate response to environmental cues, and this matching of behavior to circumstances requires that information from sensory receptors must

be integrated into control of the motor system: synaptic connections must develop to connect sensory input with motor output. The number of connections that link sensory and motor neurons may be small and the arrangement relatively simple, as in simple vertebrate spinal reflexes, or they may be extremely complex, such as when a student reads an exam question and then writes an answer in the form of an essay.

The mechanisms and consequences of sequential development within the nervous system are illustrated by the genesis of a simple adult behavior, called local bending (LB), in an annelid worm, the medicinal leech. LB is a response to moderate touch: within the body segment that was touched, longitudinal muscles close to the touch contract while at the same time longitudinal muscles on the opposite side of the body relax. The net effect is a subtle response that moves the skin away from the site of the touch, and both the identities of neurons that produce LB and their synaptic connections have been characterized in detail.

This simple behavior goes through a developmental sequence that illustrates several of the principles listed previously (Figure 2). In fact, leech embryos produce two closely related behaviors using the same set of motor neurons. The first behavior to emerge is called circumferential indentation (CI); it consists of a ring of longitudinal contraction that encircles the body. This behavior is initially generated spontaneously; touching the body fails to elicit it. During this period, motor neurons are interconnected with one another only through electrical synapses, but they receive no synapses carrying sensory information. A few days later, touching the body wall elicits CI, and experiments have revealed that by this time sensory neurons have formed electrical connections onto interneurons, which have formed electrical connections onto motor neurons, but the behavior is still not adultlike because there is no relaxation on the side opposite the touch. (Similar observations have been obtained from zebra fish embryos and chicks: Initially, locomotory movements occur spontaneously when motor neurons are connected through electrical synapses, and later in development they are elicited by sensory input.) The final adult form of LB emerges when inhibitory chemical synapses form within the set of motor neurons (unlike the vertebrates, many invertebrates have inhibitory, as well as excitatory, motor neurons). The appearance of inhibitory chemical synapses modifies a response in which all longitudinal muscles within the segment are excited by touch into a response in which some of the muscles contract, whereas others relax.

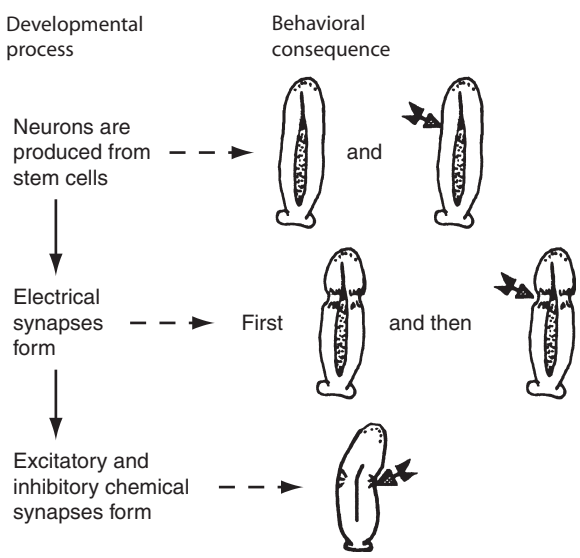

Figure 2 Sequential development of local bending behavior in an embryonic leech. Neurons begin to develop before behavior is produced. During this period, there is neither spontaneous nor evoked behavior. Once electrical synapses form, behavior begins. Initially, only motor neurons are connected, and circumferential indentation (CI; an embryonic behavior) is produced spontaneously, but not in response to touch. Later, when sensory information is carried by way of electrical synapses to the motor neurons, touch elicits CI. Later still, inhibitory chemical synapses appear, permitting expression of adult local bending in which touching the skin elicits contraction of longitudinal muscles ipsilateral to the touch, while contralateral longitudinal muscles relax.

How Are Spontaneous Embryonic Behaviors Generated?

Spinal cords isolated from embryos of several different species have been used to identify the source of early, spontaneous behavior. Conceptually, the first movements might stem from myogenic contractions of muscle fibers, or they might arise only after motor neurons have made functional synapses onto the muscle fibers. Several experiments suggest that even the earliest behaviors are neurogenic. Recordings made from chick spinal cords reveal bursts of action potentials from extensor and flexor motor neurons. Similar bursts of activity have been observed in spinal cords from fetal mice and rats, and several properties appear to be critical for the expression of spontaneous activity by spinal networks. One important feature is that the embryonic motor neurons are hyperexcitable, which may result from one or more of the following:

- functional recurrent connectivity (i.e., the activity of a neuron feeds back to maintain or even heighten its excitation);
- all major neurotransmitters may be excitatory in early spinal development (even γ-aminobutyric acid, which is generally inhibitory in adult nervous

systems, is excitatory in vertebrate embryos due to cellular properties of embryonic neurons);
- overproduction of neuronal connections (which are later pruned back); or
- transient overexpression of some neurotransmitter receptors, which increases the response to released neurotransmitters.

In leech embryos, the transition from CI to LB depends on the appearance of inhibition in a circuit that is initially purely excitatory. Similar results have been observed in mouse embryos. Spinal cords isolated from mouse embryos produce bursts of spontaneous activity that are initially synchronized bilaterally across the cord, which might cause a mouse to hop like a rabbit if it were executing locomotion. However, as development continues, the timing of the bursts shifts until they alternate from side to side, which would correspond to a walking gait. This shift from synchrony to alternation takes place when glycinergic inhibitory synapses appear, and the alternating pattern is then maintained into adulthood.

Acquiring Complexity by Combining Behaviors

In some cases, behavior develops as individual modules become assembled. A striking example of this pattern is displayed in the development of swimming behavior in larval salamanders. A very early spontaneous behavior of salamander tadpoles is a bending of the head to one side or the other. Initially, only the head bends, but as development progresses, muscles further along the body join in, causing the bend to extend posteriorly. Eventually, the bend is so extreme that it produces a behavior called coiling. Even later, coiling can be elicited by a touch to the skin on the side opposite the contraction, presumably when connections are formed that deliver sensory information to the motor neurons. The timing of these unilateral bends then becomes modified to produce contractions that alternate along the body, generating swimming.

Some of the embryonic behaviors may not have an adaptative role; instead, they reflect function in a partially completed neuronal circuit. In other cases, embryonic behaviors can have a clear adaptive purpose. An example is the hatching movement of the bird embryo, which allows the embryo to break open and escape from the shell.

Development of Behaviors That Depend on Endocrine Function and Learning

Although development of the nervous system – that is, generation of neurons and production of synaptic connections – is necessary for the normal development of behaviors, in some cases it is not sufficient. Additional factors are required to make the nervous system capable of producing normal behavior. The acquisition in many bird species of the ability to produce normal, species-specific song provides an excellent example. Most of the research on this topic has focused on canaries and zebra finches, both of which are highly tractable in a laboratory. In both species, only male birds sing, and anatomical study reveals that several brain centers are much larger in male birds than they are in the female brain. These centers are intimately involved in the production of song, and their large size depends on the presence of gonadal hormones during a critical period early in the bird's life. Some of the enlarged structures are necessary for the production of song – that is, they are motor or premotor centers; ablating them from the brain of an adult male bird that could sing normally prior to the ablation renders him unable to sing. Other enlarged regions are unimportant for the production of adult song, at least in the short term, but they are critical for the development of normal singing behavior because these birds must learn to sing – much as a human child must learn to talk.

During early life, a male zebra finch goes through a phase during which he listens to the song of one or more conspecific males around him (in the natural world this 'tutor' is his father) and memorizes the song. At this time, the bird does not himself sing, so memorization does not require active practice, and this learning period is of limited duration. That is, there is a critical period for the learning of the song. Later in life, when environmental cues cause a surge in reproductive hormones, the bird begins to develop his own song, which will resemble – but not precisely reproduce – his tutor's song. During this time, he practices his song and must be able to hear himself; deafened birds sing poorly if at all. Male birds that are reared in isolation during the critical period eventually do sing, but the structure of the song is very simple and lacks the characteristics of normal zebra finch singing.

The development of language in human infants has many features in common with song development in zebra finches. Language learning begins early, and in the initial phase the human infant listens to and memorizes some features of the language he or she encounters, but does not speak. Later, the infant begins to practice sounds that eventually coalesce into language, followed by a period of rapid linguistic development. Around puberty, the critical period for learning language seems to close, and although adults are capable of learning new languages, the process

calls on different regions of the brain and operates differently from acquisition of language by children.

One Nervous System, Very Different Lifestyles: Metamorphosis

In animals in which the life cycle includes distinctly different life stages (e.g., insects and amphibia), the nervous system must be radically reconfigured during transitions to allow the individual to produce critically important behaviors during each stage, even though the lifestyle during each stage may require different sensory capabilities, locomotory behaviors, and central processing. The transitions between stages are typically associated with changes in the levels of one or more hormones, and variations in hormone levels act directly on the identified neuron, which then changes the behavioral output. For example, in the tobacco hornworm moth, during the larval stage the caterpillar crawls on the plants that provide its food using pro-legs, which are not precursors of the legs in the adult. The neurons that drive and coordinate the muscles of the pro-legs survive the transition from the caterpillar to the legless pupal stage, but the morphology and synaptic input changes. In fact, generally the motor neurons survive from the caterpillar stage, through pupation, and into the adult stage, changing their morphology and function with each transition. In contrast, sensory systems in adult insects may include new neurons.

Summary

The development of behavior depends on the intricate interaction between the cellular processes of development in the nervous system, other systems in the body (e.g., endocrine function), and the individual's environment. It begins well before hatching or birth (depending on the species) and continues throughout life. Many of the processes that underlie behavioral development have been studied in other contexts, such as the mechanisms of learning and the hormonal regulation of target organs. This complex web of developmental processes equips the individual to respond appropriately and effectively to the many challenges it will encounter.

See also: Behavioral Hierarchies; Birdsong: The Neurobiology of Avian Vocal Learning; Computational Approaches to Motor Control; Fetal Alcohol Syndrome; Hormones and Behavior; Language Development; Motor Control of Feeding and Drinking; Speech Production: Development.

Further Reading

Bateson P and Martin P (1999) *Design for a Life: How Behaviour Develops*. London: Jonathan Cape.

Doupe AJ and Kuhl PK (1999) Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience* 22: 567–631.

Hamburger V (1963) Some aspects of the embryology of behavior. *Quarterly Review of Biology* 38: 342–365.

Kiehn O and Tresch MC (2002) Gap junctions and motor behavior. *Trends in Neurosciences* 25: 108–115.

Marin-Burgin A, Kristan WB Jr., and French KA (in press) From synapses to behavior: Development of a sensory–motor circuit in the leech. *Developmental Neurobiology.*

Roerig B and Feller MB (2000) Neurotransmitters and gap junctions in developing neural circuits. *Brain Research: Brain Research Reviews* 32: 86–114.

Sanes DH, Reh TA, and Harris WA (eds.) (2000) *Development of the Nervous System*. San Diego: Academic Press.

Suster ML and Bate M (2002) Embryonic assembly of a central pattern generator without sensory input. *Nature* 416: 174–178.

Weeks JC and Levine RB (1995) Steroid hormone effects on neurons subserving behavior. *Current Opinion in Neurobiology* 5: 809–815.

Development of Drosophila Neuromuscular Junctions

B Berke and H Keshishian, Yale University, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The *Drosophila* neuromuscular junction (NMJ) ranks as one of the best-understood genetic model-systems for examining synaptic development and function. Most research involves the motor neurons and muscle fibers of the embryonic and larval body wall. The connections are precise and relatively invariant, simplifying the analysis of mutant phenotypes. Furthermore, the NMJ's peripheral location aids experimental access. Developing motor neurons may be observed *in vivo,* and electrophysiological studies are possible at all developmental stages.

Drosophila generates its motor systems twice, first as an embryo and then again as a pupa. The adult motor system is more complex than the larval one, with segment-specific specializations for flight, walking, and reproductive behavior. Although many useful insights have emerged from studies of the developing pupa, most current research is focused on the more accessible embryonic and larval synapses.

General Features of the Neuromuscular Junction

The *Drosophila* larval body-wall muscles generate the movements involved in locomotion and postural control. Most research has examined the first seven abdominal segments (A1–A7). These share the same muscle pattern of 30 muscle fibers on each side, with only a minor difference in segment A1 (Figure 1(a)). The terminal (A8) segment has a distinct muscle pattern, as do the three thoracic segments (T1–T3).

The major source of innervation for the hemisegment is the ganglionic peripheral nerve (PN), which emerges laterally from each central nervous system (CNS) segment. In segments A1–A7, the PN contains between 80 and 90 axons, of which fewer than half are efferent. The PN elongates throughout development, reaching 3 mm by the end of larval life. At the body wall the PN divides into five branches. All dorsal and most mid-body-wall muscle fibers are innervated by the intersegmental nerve branch (ISN), whereas the remaining mid-body-wall muscles are innervated by the dorsal branch of the segmental nerve (SN)a. All ventral longitudinal muscle fibers are innervated by the SNb branch (also referred to as the ISNb), whereas the remaining oblique and superficial ventral muscles are innervated by the SNc and SNd (or, alternatively, ISNd) branches. A minor source of innervation is provided by the transverse nerve (TN), which exits from the dorsal midline of each abdominal CNS segment to innervate ventral and mid-body-wall targets.

All motor neurons are glutamatergic. A subset also expresses synaptic cotransmitters, including the neuropeptides proctolin and leukokinin I, and the neuromodulatory transmitter octopamine. Unlike many invertebrates, there are no inhibitory GABAergic motor neurons. The NMJ also responds to neuromodulatory molecules that circulate in the hemolymph, such as the FMRFamide peptides.

The neuromuscular junction consists of multiple branches bearing variously sized synaptic boutons. The NMJs lie on the muscle fiber side facing the animal's interior (Figure 2). Individual boutons contain multiple neurotransmitter release sites, characterized by T-bar ribbons with synaptic vesicles and soluble *N*-ethylmaleimide-sensitive factor attached protein receptor (SNARE) proteins. Surrounding each release site is a periactive region where vesicle endocytosis occurs. By the end of larval life, each bouton may contain up to 40 release sites. The postsynaptic membrane opposite the bouton is enriched for glutamate receptors (GluRs), adhesion proteins, and ion channels. The proteins are anchored by multiple adaptor proteins, including the postsynaptic density (PSD-)95 homolog Discs large (Dlg) and the 4.1 superfamily protein Coracle. The postsynaptic site has a highly folded membrane known as the subsynaptic reticulum (SSR).

There are two postsynaptic ionotropic GluRs at the NMJ. These multimeric receptors are assembled from the five receptor subunits expressed by muscles. The two receptors differ by whether they contain the GluR-IIA or GluR-IIB subunit, but both have GluR-IIC (also referred to as dGluR-III), GluR-IID, and GluR-IIE subunits. The GluR-IIA and GluR-IIB subunits form the channel pore and confer the single-channel conductance, antagonist sensitivity, and desensitization characteristics. On the presynaptic side, there is also a single GluR-A metabotropic GluR (also referred to as DmGluR-A) that contributes to synaptic augmentation and facilitation.

The Motor Neurons and Their Projections

There are approximately 35 motor neurons that innervate each abdominal hemisegment. Of these, some 30 type Ib motor neurons synapse with either a single muscle fiber or muscle fiber pair. The Ib neurons provide direct control of individual muscle fibers during locomotion. At the NMJ, the Ib motor

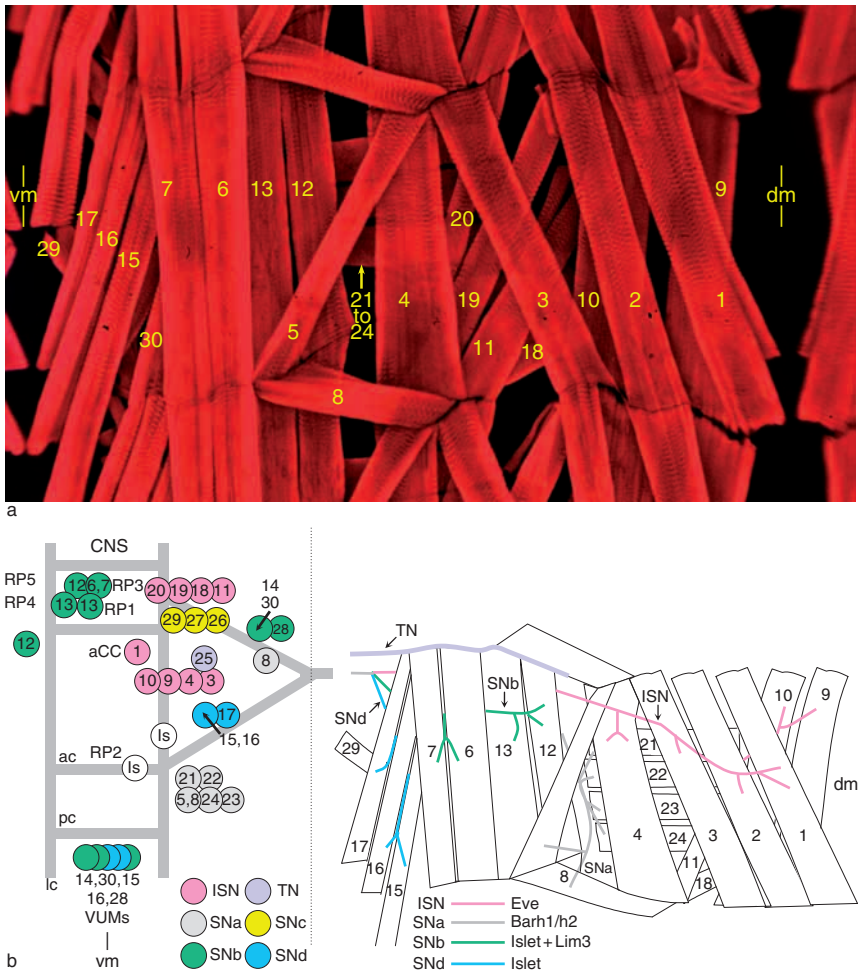

Figure 1 The motor neurons and muscle fibers of the abdominal body wall: (a) muscles of a single abdominal hemisegment displayed from the ventral midline (vm) to the dorsal midline (dm) (rhodamine phalloidin fluorescence; anterior at the top); (b) diagram of the central nervous system (CNS) (left) and muscle map showing the major nerve branches (right). In (a), each muscle fiber is designated by a number from 1 to 30 (several superficial muscle fibers are not visible in this dissection). In (b), left, the diagram shows the motor neurons that innervate a single abdominal hemisegment. The gray bars indicate the anterior commissures (ac), posterior commissures (pc), and the longitudinal connectives (lc). Each motor neuron's preferred muscle target is indicated by the number within the cell body. All the motor neurons shown are type Ib, except for two type Is cells. The color code indicates the nerve branch the axon takes. In (b), right, the muscle map is colored to indicate the transcription factor code that defines motor-neuron-targeting preference. ISN, intersegmental nerve; RP, radula protractor motor neuron; SN, segmental nerve; TN, transverse nerve; VUM, ventral unpaired median motor neuron. Panel (b, left) on Chiba A (1999) Early development of the Drosophila neuromuscular junction: a model for studying neuronal networks in development. *International Review of Neurobiology* 43: 1–24.

neurons have large boutons with diameters ranging from 2 to 6 μm (Figure 2). There are, in addition, two type Is motor neurons that function as common excitors. These cells each innervate large subsets of dorsal and ventral muscles, respectively, and have medium-sized 1–4 μm round boutons. Bouton diameters for both classes remain fairly constant throughout larval development. The size differences between Ib and Is motor neuron boutons may be due to differences in action potential firing rates. Type Ib boutons become smaller and resemble Is boutons when action potential firing rates are experimentally reduced using either Na^+ channel mutations or sublethal doses of tetrodotoxin (TTX) during development.

There are also two type II efferent neurons that express octopamine and are probably neuromodulatory. They each innervate multiple dorsal and ventral muscle fibers, respectively, with boutons smaller than 1 μm. Finally, a single type III motor neuron with oblong boutons innervates muscle fiber 12 in four abdominal segments. A typical muscle fiber receives three inputs: from a type Ib, a type Is, and a type II motor neuron.

Each type Ib motor neuron synapses with its target muscle fibers with high fidelity. The motor neuron RP1 innervates the ventral longitudinal muscle fiber 13 with only a 3% error rate. The preference of a motor neuron for a specific muscle fiber is retained

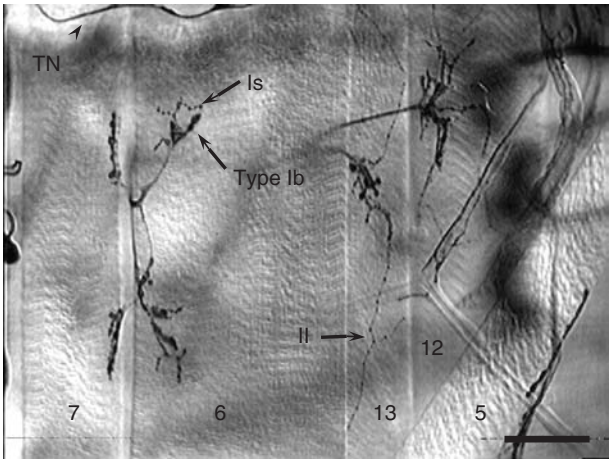

Figure 2 The neuromuscular junctions of the ventral longitudinal muscle fibers 7, 6, 13, and 12 of a third instar larva, visualized with neuron-specific labeling (anti-HRP). The three major motor neuron types (Ib, Is, and II) are distinguishable by their characteristic bouton morphology. Scale bar = 50 μm. From Chang TN and Keshishian H (1996) Laser ablation of Drosophila embryonic motoneurons causes ectopic innervation of target muscle fingers. *Journal of Neuroscience* 16: 5715–5726.

even when muscle patterning is altered. When muscle fiber 13 is duplicated, the RP1 motor neuron faithfully innervates both duplicated cells, whereas motor neurons to neighboring muscle fibers remain unaffected. This implies the presence in motor neurons of a cell-specific muscle-recognition mechanism.

Cellular Determination and Axon Guidance

Motor neuron growth cones make multiple guidance decisions during development. They must first take the correct neuropilar tracts to exit the CNS and then follow an appropriate nerve branch to either a dorsal or ventral muscle field. These choices require multiple receptors operating at specific times during development. The dorsal or ventral muscle field choice is governed by the combinatorial expression of several transcription factors, which, in turn, control the expression of specific receptors and signaling pathways. Significantly, some of the transcription factors involved in ventral muscle field preference are conserved evolutionarily between insects and mammals.

Factors Regulating Motor Neuronal Identity

Two transcription factors, Hb9 and Nkx6, contribute to ventral muscle targeting (Figure 1(b)). The proteins regulate the expression of at least two additional transcription factors, Islet and Lim3. Both Islet and Lim3 are expressed by motor neurons that take the SNb nerve branch to ventral longitudinal muscle fibers, whereas Islet alone is expressed by motor neurons projecting to the nearby ventral oblique muscle fibers innervated by the SNd branch. In *lim3* loss-of-function mutants, the SNb motor neurons behave like those of SNd. Conversely, overexpression of Lim3 in SNd motor neurons switches them to SNb targets. The homeodomain transcription factors Barh1/h2 play a similar essential role for SNa-projecting axons to the mid-body wall (Figure 1(b)).

Dorsally projecting motor neurons express the transcription factor Even skipped (Eve). Mutations of *eve* cause dorsal motor neurons to project to ventral regions. When Eve is expressed in ventral motor neurons, the neurons switch their identity and project to dorsal targets. Because Eve functions as a transcriptional repressor, it is probable that this guidance switch is due to the inhibition of factors required for ventral identity, such as Hb9 and Lim3.

The identity of the downstream effectors regulated by these transcription factors remains incompletely understood. For the ventrally projecting motor neurons, Islet and Lim3 positively regulate the expression of the Ig superfamily homophilic adhesion molecule Fasciclin III (FasIII). They also negatively regulate the IgSF protein beaten path (Beat1c), a molecule involved in axonal defasciculation at specific choice points along the body wall. Among the dorsal class of motor neurons, Eve indirectly regulates the expression of the Ig superfamily homophilic adhesion molecule Fasciclin II (FasII) and Unc5, one of the receptors for the Netrin chemotropic factors.

The Decision to Defasciculate

The next task faced by a motor neuron is to decide where to leave the axon fascicle. The decision to defasciculate depends on a balance between interaxonal adhesion and repulsion. All motor neuron axons express FasII, as well as the chemorepulsive molecule Semaphorin Ia (SemaIa) and its receptor PlexinA. Subsets of motor neuron axons also express additional cell adhesion molecules, such as Fasciclin I, FasIII, and Connectin (Con). Increased expression of FasII in motor neuron axons suppresses defasciculation, as does the reduction of either SemaIa or PlexinA. The *SemaIa* mutant phenotype can be suppressed by reducing levels of either FasII or Con. This is consistent with the idea that defasciculation is regulated by the relative levels of adhesion and repulsion between axons.

Defasciculation is also regulated by locally secreted molecules, such as Beaten path. *Beat* mutants fail to defasciculate, bypassing their muscle targets. Genetic interaction studies indicate that the Beat

protein tips the balance toward defasciculation by selectively interfering with FasII-mediated axonal adhesion, possibly by directly binding to the FasII ectodomain.

Additional insight into defasciculation has come from examining the neurally expressed receptor tyrosine phosphatases (RPTPs) dLAR, DPTP66D, and DPTP99A. In general, mutation of the RPTPs leads to a failure to defasciculate, although with mild phenotypic penetrance. The phenotype is more severe when multiple RPTPs are eliminated, suggesting redundancy or cooperative functions among the proteins. Although the identity of the ligands and intracellular targets of most RPTPs remains unresolved, dLAR is thought to interact with the extracellular heparin sulfate proteoglycans Syndecan and Dally-like. The intracellular targets of dLAR include the cytoplasmic tyrosine kinase Abelson (Abl), and its phosphorylation target Enabled (Ena). dLAR may contribute to a growth cone's chemotropic response by regulating actin cytoskeleton assembly.

Embryonic axons fail to defasciculate when muscle fibers are genetically eliminated from the body wall, suggesting that defasciculation also depends on muscle-derived cues. One candidate for a muscle-derived signal is Sidestep (*side*), a chemotropic ligand that is expressed by all muscle fibers. In *side* mutants, motor neurons bypass their muscle targets, but unlike *beat* mutants, the phenotype is not suppressed by reduced levels of FasII. This suggests that defasciculation depends on at least three mechanisms: (1) a local reduction in interaxonal adhesion, combined with (2) elevated interaxonal repulsion, and possibly (3) one or more chemotropic signals derived from the musculature.

Target Selection

As growth cones explore the body-wall musculature, they extend rapidly moving filopodia that sample cell surfaces. A filopodium may exist for as little as a few minutes and extend as far as two to three muscle fibers ahead of the growth cone. Time-lapse studies in *Drosophila* embryos show that the efferent growth cones direct themselves toward their destined targets by following trajectories pioneered by specific filopodial contacts. The earliest contacts between the motor neuron growth cone and the target muscle fiber occur at muscle protrusions known as myopodia. These processes resemble neuronal filopodia and increase the probability of interactions between the growth cone and muscle fiber. The myopodia may also present key cell-surface molecules to the motor neuron, promoting target recognition.

Molecular Recognition of Synaptic Targets

The molecular recognition of specific muscles by motor neurons involves both attractive and repulsive signaling (Figure 3). Some key molecules are expressed by all or large subsets of motor neurons and muscles, whereas others have narrower, cell-specific expression patterns. The molecules include secreted as well as cell-surface proteins. During embryonic synaptogenesis, all muscle fibers and motor neurons express FasII. In addition, all muscle fibers secrete the chemorepellant protein Semaphorin IIa (SemaIIa), and all motor neurons in turn express PlexinB, the SemaIIa receptor. Gain- and loss-of-function tests indicate that the

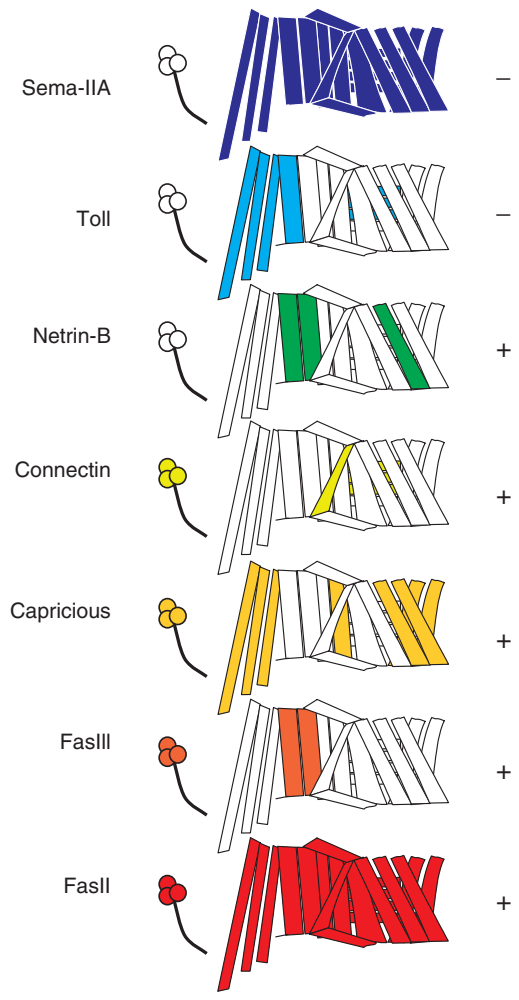

Figure 3 Multiple attractive (+) and chemorepulsive (–) molecules are expressed by individual muscle fibers and motor neurons at the time of innervation. A motor neuron's affinity for a specific muscle target depends on balancing multiple attractive and repulsive molecular cues. From Nicholson N and Keshishian H (2006). Metamorphosis and the formation of the adult musculature. In: Sink H (ed.) *Muscle Development in Drosophila,* pp. 113–120. New York: Springer.

affinity of motor neurons for muscle fibers is adjusted in part by balancing the relative levels of FasII, SemaIIa, and/or PlexinB. Although these broadly expressed proteins cannot define specific target preferences, their balance modulates the stability of neuromuscular contacts as synaptogenesis proceeds.

Specific targeting decisions depend on the expression of molecules within subsets of motor neurons and/or muscles (Figure 3). For example, the IgCAM FasIII is expressed by both motor neuron RP3 and its two synaptic targets, the muscle fibers 7 and 6. In loss-of-function *FasIII* mutants, RP3 shows low-frequency but statistically significant targeting errors. This indicates that other guidance molecules are available to guide RP3 to its targets with some accuracy. However, RP3 can be misdirected when adjacent muscles are made to express FasIII. Significantly, the ectopic expression does not degrade or alter the guidance of other motor neurons, indicating that muscle-expressed FasIII is a specific recognition cue for the RP3 motor neurons. Results similar to these have been obtained for several other proteins involved in target selection, including Connectin and Capricious. Probably every muscle fiber has a distinct and characteristic profile of molecules involved in recognition and targeting. For example, muscle fibers 7 and 6 express FasII, FasIII, NetrinB, Toll, and SemaIIa, whereas the nearby ventral longitudinal muscle fiber 12 expresses FasII, SemaIIa, and Capricious (Figure 3).

A combinatorial model involving attractive versus repulsive signaling has been tested by analyzing the phenotypes of *netrinB* (*netB*), *FasII*, and *SemaIIa* mutant embryos. In *netB* mutants, the RP3 motor neuron fails to reliably innervate muscle fibers 7 and 6, suggesting a role for NetrinB in guiding the growth cone to those muscles. The guidance defect is largely suppressed when *SemaIIa* mutations are combined with *netB*, suggesting that NetrinB-mediated attraction is in balance with repulsive signaling by SemaIIa. Because adjacent nontarget muscle fibers do not express NetrinB, this favors RP3's innervation of its NetrinB-positive target muscle fibers over neighboring ones. Similar scenarios exist for motor neurons projecting throughout the body wall, involving a balance between attractive and repulsive signals to define specific muscle targets against a generally repulsive nontarget background.

Synaptic Development and the Role of Activity

Although relatively little is known about how synaptogenesis is initiated, mutations of genes that delay this step have been informative. Tetraspanins are evolutionarily conserved proteins with widespread functions in cell signaling and motility. Three tetraspanin genes are transiently expressed by motor neurons as synapses form. In mutants of the tetraspanin gene Late bloomer, growth cones correctly identify and contact their target muscle fibers, but then delay synaptogenesis for hours. The phenotype is enhanced when all three tetraspanins are deleted, suggesting that the proteins function in a cooperative fashion to trigger synapse formation. How this happens remains unknown.

The transformation of an embryonic growth cone to an anatomically recognizable presynaptic terminal takes only 3–4 h. On the presynaptic side, the immature contact forms multiple round varicosities that prefigure synaptic boutons. Electron microscopy (EM) studies show that postsynaptic electron-dense regions arise exclusively at the contacts made between the growth cone and its preferred muscle fiber.

Early Roles for Activity at the Neuromuscular Junction

Both the localization of postsynaptic receptors and the refinement of synaptic connections depend on neural activity. Small glutamatergic potentials can be recorded from muscle as soon as the motor neuron growth cone arrives on site. GluRs are initially expressed throughout the muscle fiber. Following growth cone contact, the receptors accumulate at the site of the developing synapse. Although it is not known whether this early localization is due to the trapping of existing receptors (as occurs at larval synapses), the localization to the synapse is disrupted when motor neuron activity is suppressed.

From the time motor neurons first contact the musculature to the time of hatching, the embryo executes 1–2 strong peristaltic contractions/min. The contractions are driven by motor neuron action potentials activating NMJs throughout the body wall. The significance of this periodic, wavelike motor activity has been tested by suppressing motor neuron electrical activity. Normally, motor neurons withdraw contacts made with nontarget muscle fibers during the first few hours of synaptic development. The loss of evoked NMJ activity results in a failure to withdraw these incorrectly placed contacts throughout the body wall. The ectopic contacts are removed only if activity is restored during an early critical period that ends soon after the embryo hatches. The remaining contacts mature into functional but miswired NMJs. The mechanisms governing activity-dependent refinement remain largely unknown.

The Growth of the Neuromuscular Junction

After hatching, the larval NMJ undergoes dramatic growth. In hatchling larvae, the NMJs on muscle fibers 7 and 6 have fewer than 20 synaptic boutons, with each bouton averaging only two release sites. By the end of larval development 100 h later, the muscle fiber surface area has increased 100-fold and the NMJ contains 80–100 boutons, with up to 40 release sites per bouton.

Neurotransmission is also modified as the NMJ grows. Spontaneous miniature currents are initially small and infrequent, but become steadily larger and more regular as the embryo develops. By the time of hatching, these quantal events are skewed toward larger sizes. Quantal size (the postsynaptic response to a single vesicle) thereafter remains constant as the larval NMJ grows. The amplitude of the evoked excitatory junctional potential (EJP) also remains constant despite a fall in membrane resistance as the muscle grows. By contrast, quantal content (the average number of vesicles released per action potential) steadily increases as the NMJ enlarges. This is due to the large increase in the number of release sites at the NMJ and probably from increases in release efficiency.

In vivo imaging of NMJ development indicates that more than half of new boutons are added by intercalation within existing strings of boutons. In the remaining cases, the boutons are added at branch tips. The new boutons have the same density of vesicles, active zones, and T-bar ribbons as mature boutons and precede the development of postsynaptic specializations and SSR folding. Synaptic boutons are also eliminated from the NMJ. Because the SSR develops only after boutons have formed, isolated SSRs that lack boutons probably indicate elimination events. Nearly 18% of muscle fiber 7 and 6 NMJs in late first and early second instar larvae show these tell-tale bouton retraction footprints, compared to 5% in the third instar. Whereas bouton formation can occur anywhere along an NMJ branch, elimination is usually restricted to branch termini.

Time-lapse studies of green fluorescent protein (GFP)-tagged GluRs show that larval GluRs first appear in the muscle membrane as extrasynaptic hot spots. Fluorescence recovery studies of tagged GluR-IIA indicate that the extrasynaptic receptors are preferentially recruited into the expanding PSD that develops opposite the presynaptic boutons. After they enter the PSD, the receptors become immobilized, presumably by binding to adaptor proteins, with only a 20% turnover every 24 h.

The periactive zone surrounding the release site is enriched for the adhesion molecule FasII, a protein that plays an important role in NMJ growth. Newly budded boutons have lower levels of FasII than parental boutons. This is an intriguing observation because mutations that reduce FasII levels generally have enlarged NMJs. The actions of FasII on NMJ expansion depend in part on presynaptic signaling via the amyloid precursor protein (APPL) and through interactions with the PSD-95, Dlg and ZO1 (PDZ)-containing adaptor protein dX11/Mint/Lin-10. The periactive zone is also enriched for Bruchpilot (also referred to as NC82), a coiled-coiled domain protein with homology to the human active zone protein ELKS/CAST. Bruchpilot has a key role in defining the boundaries of the active zone and in the clustering of Ca^{2+} channels.

The Roles of Activity in Neuromuscular Junction Growth and Function

The NMJ's physical size is correlated to the level of neuromuscular activity. Mutations that increase membrane excitability, as well as treatments that increase locomotor activity, result in larger and more highly branched NMJs. These NMJs have up to twice as many boutons as controls and also have larger EJPs and more frequent spontaneous activity.

In genetically hyperactive larvae, there is a substantial reduction of the cell-adhesion molecule FasII at the NMJ. Hypomorphic *FasII* mutations, which directly reduce FasII to the levels observed in hyperactive larvae, have similarly enlarged NMJs. This implies that the activity-dependent downregulation of FasII is sufficient to account for the NMJ expansion seen with neuromuscular hyperactivity. Significantly, the expected NMJ expansion of hyperactive larvae is effectively blocked when FasII levels are held constant.

Cyclic adenosine monophosphate (cAMP) levels also influence NMJ growth. *dunce* mutants have elevated levels of cAMP and significantly expanded NMJs. When *dunce* is combined with hyperactivity mutations such as *Shaker,* there is a strong synergizing effect that further promotes NMJ expansion. Conversely, mutations that reduce cAMP levels suppress activity-dependent NMJ growth. Significantly, FasII levels at the NMJ are reduced in *dunce* mutants, and the overgrowth caused by increased cAMP levels is blocked when FasII levels are held constant. Thus, FasII is a common endpoint in multiple regulatory cascades affecting NMJ expansion.

Activity-dependent NMJ growth is also dependent on transcriptional regulation. The expression of the cAMP responsive element-binding protein (CREB) transcriptional activator enhances the NMJ expansion phenotype of FasII mutants, whereas the expression of

a negative repressor of CREB blocks the increase in presynaptic neurotransmission seen in *dunce* mutants. This indicates that presynaptic transmitter release and NMJ structural plasticity may be regulated in different ways. This idea is supported by the actions of AP-1, an immediate-early transcription factor composed of a Fos and Jun heterodimer. AP-1 functions through CREB to regulate NMJ growth but independently, via Jun kinase, to modulate synaptic function.

Activity influences the expression and localization of GluRs by increasing local subsynaptic translation during larval development. The NMJ expansion associated with elevated motor neuron activity is partially blocked by mutations that decrease the number of available receptors or suppress their translation. This indicates that receptor availability limits the formation of new boutons, which is supported by the finding that reducing GluRIIs to relatively low levels leads to smaller NMJs.

Finally, functional plasticity at the NMJ can arise soon after short bouts of hyperactivity; 40 min of vigorous locomotion is sufficient to cause long-term facilitation of evoked potentials at the NMJ. This is due to an increase in neurotransmitter release, surprisingly through the recruitment of physically larger vesicles to the release sites, resulting in increased quantal size. The effect depends on presynaptic protein kinase A activity and is independent of postsynaptic receptor activation. Similarly, when postsynaptic sensitivity is acutely decreased using GluR toxins, there is a rapid compensatory elevation of quantal release, possibly through the modulation of presynaptic Ca^{2+}-channel activity. The relationships between these acute forms of physiological plasticity and long-term growth changes at the NMJ are not yet known.

Transsynaptic Signals Involved in Neuromuscular Junction Development

Retrograde signaling, from muscle to motor neuron, plays several important roles during NMJ development. For example, GluR-II mutations lead to a compensatory increase in the number of release sites, elevating quantal release. This homeostatic mechanism holds the amplitude of postsynaptic potentials constant despite changes in muscle receptor sensitivity. Similarly, when muscle depolarization is suppressed by experimentally elevating the potassium leak, the motor neuron responds by increasing transmitter release. Elevating CREB transcriptional activity in muscle, achieved by expressing the CREB co-activator dCBP, also results in a retrograde inhibitory signal. The motor neuron responds by reducing both evoked neurotransmitter release and short-term facilitation.

The best-understood retrograde signaling system regulating NMJ growth involves the transforming growth factor (TGF)-β/bone morphogenetic protein (BMP) ligand Glass bottom boat (Gbb). Gbb is expressed in body wall muscle fibers and regulates presynaptic development through both canonical and noncanonical pathways (Figure 4). In the canonical mode, Gbb binds to a presynaptic TGF-β receptor complex containing the type II BMP receptor Wishful thinking (Wit) and at least one of two type I BMP receptors. These kinases stimulate transcription by activating the R-Smad transcription factor Mad and its co-Smad Medea. Disrupting any of these results in smaller NMJs with reduced neurotransmission. Presynaptic TGF-β/BMP signaling also depends on proteins in the early and late endosomal compartments, as well as on retrograde axonal transport, suggesting

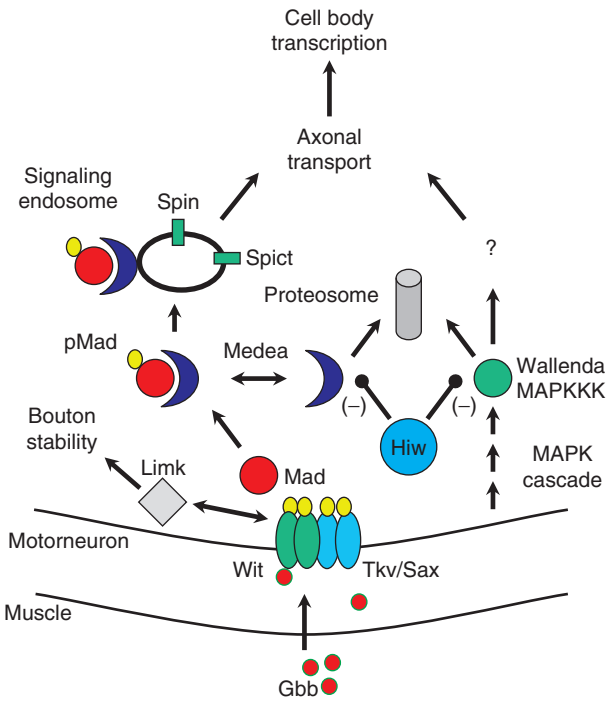

Figure 4 Retrograde signaling at the *Drosophila* NMJ. The TGF-β/BMP growth factor Gbb is expressed by muscle and is required for normal NMJ growth and bouton formation. The ligand stimulates a motor-neuronal receptor complex composed of the type II BMP receptor Wit and the type I receptors Tkv and/or Sax. The activated complex phosphorylates the R-Smad transcription factor Mad. Phospho-Mad associates with the co-Smad Medea and stimulates transcription at the cell body. Regulation is, in part, by ubiquitination of Medea via the E3 ligase (Hiw) and by regulation within the signaling endosome. Hiw also negatively regulates a MAPK cascade that promotes bouton addition through its action on the MAPKKK Wallenda. Gbb binding also stimulates a noncanonical signaling pathway involving Limk, leading to the stabilization of new boutons. BMP, bone morphogenetic protein; MAPK, mitogen-activated protein kinase; MAPKKK, MAPK kinase kinase; NMJ, neuromuscular junction; TGF, transforming growth factor.

the physical transport of one or more proteins from the NMJ to the cell body.

The mechanisms that regulate Gbb release from muscle remain largely unknown. Inhibition of calmodulin-dependent protein kinase (CaMK)II in the muscle fiber stimulates a retrograde signal (possibly Gbb) that depends on the Wit receptor to increase the number of T-bar ribbons and the quantal content of motor neurons. Similarly, postsynaptic Dystrophin regulates a retrograde signal that influences presynaptic short-term facilitation, again through the activation of Wit in the motor neuron.

Finally, there is evidence that Wit can activate noncanonical signaling pathways independent of the Smads. For example, the C-terminus of Wit binds directly to presynaptic Lim kinase (Limk), leading to the stabilization of newly formed boutons. The loss of function mutation of *limk* results in a substantial expansion of the NMJ. Surprisingly, the increased number of NMJ boutons in *limk* mutants does not result in increased EJP amplitudes, as is the case when NMJs expand following elevated neuromuscular activity.

Mutations affecting Highwire *(Hiw)*, a large multifunctional protein with E3 ubiquitin ligase and scaffolding activities, reveal something of the complexity of presynaptic signaling during NMJ growth. Mutant larvae have significantly more boutons than normal, suggesting that Hiw functions to inhibit bouton addition. Hiw negatively regulates at least two signaling systems. It inhibits the TGF-β/BMP cascade involving Wit through a direct interaction with the transcriptional cofactor Medea, and it attenuates a mitogen-activated protein kinase (MAPK) cascade involving the MAPK kinase kinase (MAPKKK) protein Wallenda, the MAPK c-Jun N-terminal kinase (JNK), and the transcription factor Fos (**Figure 4**).

NMJ development also requires anterograde signaling from motor neuron to muscle. The Wnt family morphogen Wingless (Wg) is secreted by motor neurons, and its receptor is located on both pre- and postsynaptic membranes. In *wg* mutants, bouton formation is suppressed. The boutons have unbundled microtubules, as revealed by labeling the microtubule binding protein Futsch. In addition, the terminal NMJ branches do not elongate, resulting in less separation between boutons. *wg* larvae also have extensive postsynaptic phenotypes, including aberrant expression of the GluR-IIA, the Dlg adaptor protein, and overall reduced development of the SSR. Significantly, similar phenotypes are observed when a dominant negative Wg receptor is expressed in the body-wall muscles. These results suggest that Wg may function as both an anterograde as well as an autocrine signal to regulate NMJ growth and development.

Concluding Observations

Several areas of NMJ formation and maturation remain especially promising for further study. These include: (1) to decipher the combinatorial transcription factor codes that define motor neuronal identity, which control the receptors and signaling cascades involved in axon guidance; (2) to define the molecular signals that trigger synaptogenesis, transforming a growth cone into a synapse; and (3) to determine how multiple transsynaptic signals are integrated to control synaptic growth and function on time scales ranging from minutes to days. Finally, we need to acquire a deeper understanding of the motor system from a functional standpoint, including the roles that cotransmitters and circulating substances play in modulating motor output. Fortunately, the molecular-genetic and physiological tools are available to illuminate these areas of neuromuscular development and function in a highly accessible system of synapses.

See also: Comparative Biology of Invertebrate Neuromuscular Junctions; Learning and Memory in Invertebrates: *Drosophila*; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction: Neuronal Regulation of Gene Transcription at the Vertebrate; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): Aging; Schwann Cells and Plasticity of the Neuromuscular Junction.

Further Reading

Broadie KS and Bate M (1993) Development of the embryonic neuromuscular synapse of *Drosophila melanogaster*. *Journal of Neuroscience* 13: 144–166.

Budnik V and Ruiz-Canada C (eds.) (2006) *Special Issue: The Fly Neuromuscular Junction: Structure and Function. International Review of Neurobiology 75*.

Chiba A, Snow P, Keshishian H, and Hotta Y (1995) Fasciclin III as a synaptic target recognition molecule in Drosophila. *Nature* 374: 166–168.

Collins CA and DiAntonio A (2007) Synaptic development: Insights from Drosophila. *Current Opinion in Neurobiology* 17: 1–8.

Griffith L and Budnik V (2006) Plasticity and second messengers during synapse development. *International Review of Neurobiology* 75: 237–265.

Marques G and Zhang B (2006) Retrograde signaling that regulates synaptic development and function at the Drosophila neuromuscular junction. *International Review of Neurobiology* 75: 267–285.

McCabe BD, Marques G, Haghighi AP, et al. (2003) The BMP homolog Gbb provides a retrograde signal that regulates synaptic growth at the Drosophila neuromuscular junction. *Neuron* 39: 241–254.

Packard M, Koo ES, Gorczyca M, Sharpe J, Cumberledge S, and Budnik V (2002) The Drosophila Wnt, wingless, provides an essential signal for pre- and postsynaptic differentiation. *Cell* 111: 319–330.

Schuster CM, Davis GW, Fetter RD, and Goodman CS (1996) Genetic dissection of structural and functional components of synaptic plasticity, II: Fasciclin II controls presynaptic structural plasticity. *Neuron* 17: 655–667.

Thor S and Thomas J (2002) Motor neuron specification in worms, flies and mice: Conserved and 'lost' mechanisms. *Current Opinion in Genetics and Development* 12: 558–564.

Winberg ML, Mitchell KJ, and Goodman CS (1998) Genetic analysis of the mechanisms controlling target selection: Complementary and combinatorial functions of netrins, semaphorins, and IgCAMs. *Cell* 93: 581–591.

Zhong Y, Budnik V, and Wu CF (1992) Synaptic plasticity in Drosophila memory and hyperexcitable mutants: Role of cAMP cascade. *Journal of Neuroscience* 12: 644–651.

Relevant Websites

http://flybase.bio.indiana.edu – Flybase. Drosophila genes.
http://flybrain.neurobio.arizona.edu – Flybrain. Drosophila anatomy.
http://www.its.caltech.edu – Information Technology Services, California Institute of Technology. Motor axon development.

Developmental Disability and Fragile X Syndrome: Clinical Overview

D Hessl, C Green, D Aguilar, L Cordeiro, C McCormick, and J Yuhas, University of California at Davis, Sacramento, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Fragile X syndrome (FXS) is the most common of many known X-linked mental retardation syndromes, and it is the most prevalent inherited form of mental retardation. (Down syndrome is more common but not inherited.) This article provides a clinical overview of FXS, including the distinctive physical features, cognitive profile, and behavioral and psychiatric characteristics that have been identified through empirical study. To set the stage for this overview, and to guide a better understanding of the genotype–phenotype association studies, a brief review of the genetic basis of the disorder is presented.

Genetics of FXS

FXS is caused by an expansion of a trinucleotide (cytosine–guanine–guanine or CGG) repeat in the 5′ untranslated region of the fragile X mental retardation 1 (*FMR1*) gene located on the X chromosome at Xq27.3. CGG expansions larger than 200 repeats typically cause hypermethylation and subsequent transcriptional silencing of the gene, thereby inhibiting the production of the *FMR1* protein (FMRP). The absence or reduction of FMRP production causes FXS, which has an estimated prevalence of 1 in 4000 to 1 in 6000 live births. Individuals with 55–200 CGG repeats have the fragile X premutation, characterized by instability in parent–offspring transmissions, with a bias to expansion related to maternal repeat size, as well as increased risk for premature ovarian failure in women and fragile X-associated tremor/ataxia syndrome (FXTAS), a recently described neurodegenerative condition in aging men and women. Normal individuals have between ~5 and 54 repeats, usually with stable transmission from parent to offspring.

Physical Features

Classic physical features of FXS include a long, narrow face and prominent ears (Figure 1); however, these features may not be present until after puberty. Additional facial features are prominent jaw, midface hypoplasia, narrow palpebral fissures, epicanthal folds, strabismus, a highly arched palate, and general macrocephaly. It is important to recognize that many females and some males, even after puberty, do not develop these features. Many males present with macroorchidism (large testicles), although this has no documented effect on reproduction ability. Other common features include flat feet, joint laxity, scoliosis, tall stature, and connective tissue dysplasia. Individuals with the premutation can present with the FXS physical (and behavioral) phenotype, presumably due to a reduction in FMRP (Figure 1).

Cognitive Profile

As described previously, *FMR1* refers to 'fragile X mental retardation 1,' and it is perhaps not surprising that a high proportion of males with the full mutation (approximately 80%) have cognitive deficits falling within the mental retardation range (IQ below 70). However, there is considerable variability, spanning from the profound mental retardation range to scores falling within, or even above, the average range. 'High functioning' males with FXS usually have some degree of FMRP expression due to repeat size or methylation mosaicism; alleles with an unmethylated *FMR1* site or alleles with fewer than 200 CGG repeats will typically express FMRP normally. More variability in level of functioning is found among females with FXS due to the presence of two X chromosomes (one X with full mutation and one X with normal CGG repeat size) and effects of X inactivation, whereby only one X in each cell is active. Females with a higher proportion of cells with the normal X active are typically less affected. Genotype–phenotype association studies have documented that cognitive ability as measured by IQ is associated with FMRP variation due to the genetic factors discussed above (Figure 2). These association estimates are imprecise because the currently available FMRP measure relies on immunocytochemistry which simply reports the percentage of blood lymphocytes expressing FMRP, which is only a proxy for FMRP levels present in brain tissue. IQ is not associated with CGG repeat size within the full mutation range.

Of course, other factors, such as background genetic and environmental effects, contribute to cognitive ability in FXS, and this has been demonstrated in large family studies. For example, mean parental IQ is correlated with the cognitive ability in children with FXS, suggesting that background genes contributing to

Figure 1 (a) A young boy with the fragile X full mutation demonstrating classic features of prominent ears and a long, narrow face with prominent jaw. (b, c) Two girls with the full mutation showing similar, but perhaps less prominent, features of FXS. (d) A boy with the fragile X premutation (66 CGG repeats, FMRP expression 77% of normal, and *FMR1* mRNA 3.5 times above normal). He has a full-scale IQ of 97 but presents with prominent ears, hyperextensible joints, learning difficulties, and symptoms of ADHD and autism spectrum disorder.

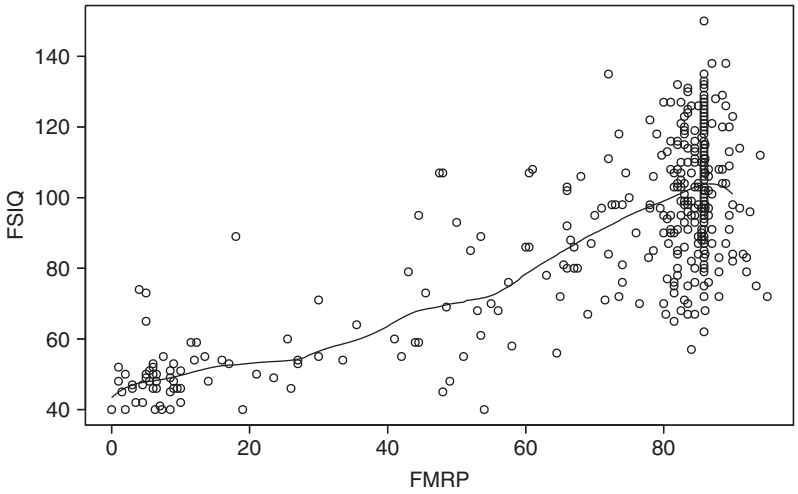

Figure 2 Scatterplots of full-scale IQ on FMRP expression for fragile X full mutation and premutation females. Reprinted from Loesch DZ, Huggins RM, and Hagerman RJ (2004) Phenotypic variation and FMRP levels in fragile X. *Mental Retardation and Developmental Disabilities Research Reviews* 10: 31–41, with permission from Wiley-Liss, Inc.

heritability of intelligence as well as family environmental factors appear to play a role in determining cognitive outcome. Family environmental factors are also related to adaptive and maladaptive behaviors in these children; for example, children living in higher quality family environments with frequent learning opportunities, enrichment, and increased organization tend to have higher adaptive levels and fewer autistic behaviors. These findings indicate that intervention at the level of the family environment, in addition to the typical behavioral and medication treatment strategies, may be effective.

Almost all studies examining the longitudinal trajectory of cognitive ability in FXS have demonstrated substantial IQ decreases across childhood and adolescence. This has been attributed to a slower

acquisition of skills relative to typical children over time rather than a loss of function. In addition, measurement of IQ in younger children appears to capture relative strengths in children with FXS, such as in visual matching and labeling, whereas tests designed for adolescents and adults tend to focus more on abstract reasoning, which is a known area of weakness.

Neuropsychological studies show that the fragile X mutation contributes to a cognitive profile of relative strengths and weaknesses, suggesting that particular brain regions are more affected than others, perhaps due to regional differences in *FMR1* expression during development. For example, it is well established that individuals with FXS show more profound impairment in mathematics computation, visual–spatial organization, and executive functions such as working memory, attention, and response inhibition. Functional magnetic resonance imaging (fMRI) studies have elucidated the brain basis for the math reasoning deficit; females with FXS demonstrate less brain activation in regions that are known to mediate math computation in typically developing individuals, and this reduction in activation is associated with FMRP expression (Figure 3). Individuals with

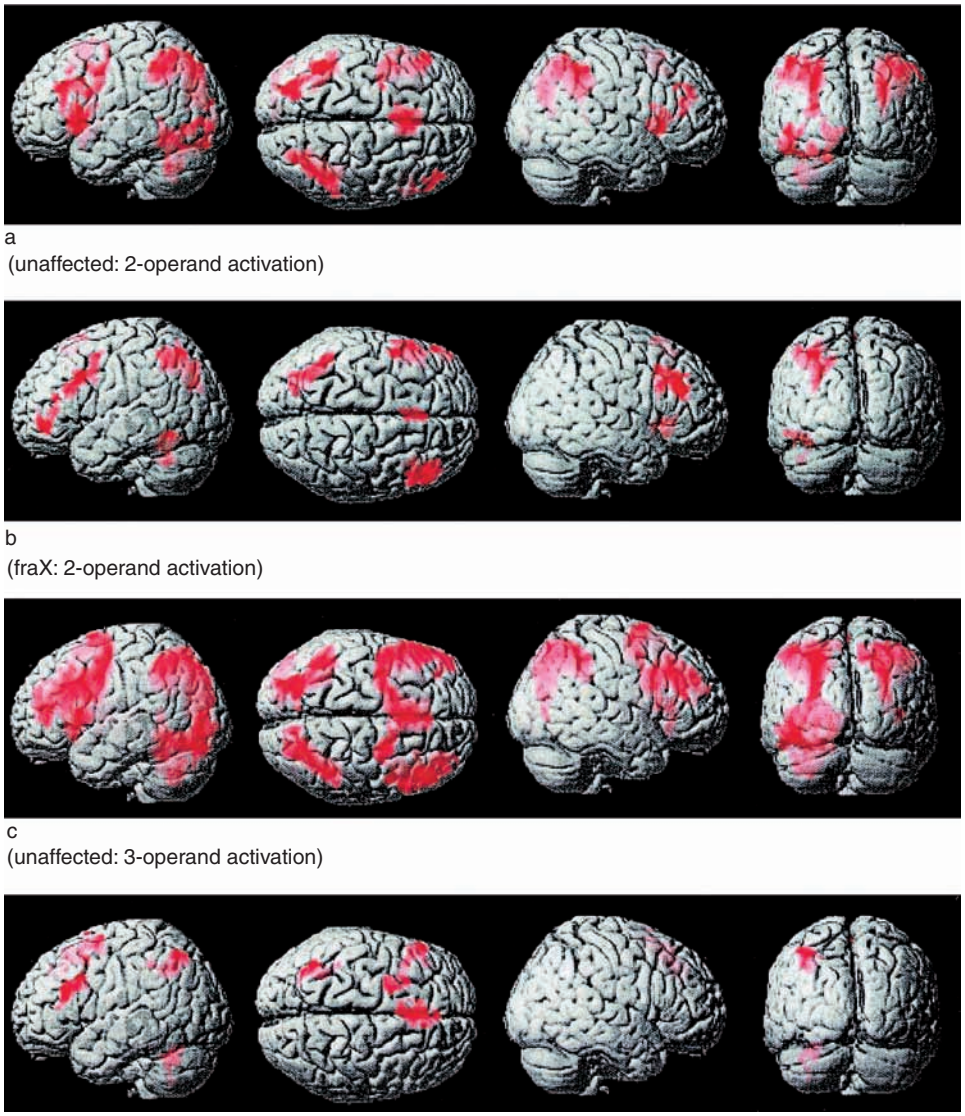

Figure 3 Brain areas that show significantly greater activation during two-operand arithmetic equations for (a) unaffected subjects and (b) females with FXS and during three-operand arithmetic equations for (c) unaffected subjects and (d) females with FXS. Reprinted from Rivera SM, Menon V, White CD, Glaser B, Glover G, and Reiss AL (2002) Functional brain activation during arithmetic processing in females with fragile X syndrome is related to FMR-1 protein expression. *Human Brain Mapping* 16(4): 206–218, with permission from Wiley-Liss, Inc.

FXS typically have speech and language difficulties, including articulation, rate, prosody, and tone problems. Speech is often rapid, cluttered, tangential, perseverative, and repetitive.

Executive functions include such processes as planning, initiation, inhibition, cognitive flexibility, behavioral and attentional regulation, working memory, and conceptual problem solving. Both males and females with FXS demonstrate deficits in executive functioning, and many studies have documented that problems in this domain are present even in comparison to IQ- and mental age-matched comparison groups. Furthermore, brain fMRI studies have been effective in demonstrating executive functioning deficits associated with FXS and, specifically, with *FMR1* gene expression. A notable example is a study of females with FXS that examined brain activation in response to a response inhibition, or 'go/no-go' task. The study found that subjects with FXS demonstrated less brain activation during the inhibition task as opposed to typical developing controls in several brain regions, including the supplementary motor area, anterior cingulate and midcingulate cortex, basal ganglia, and hippocampus. The most important finding demonstrated a strong association between FMRP expression and activation in the ventrolateral prefrontal cortex and the left and right striatum (**Figure 4**). Other fMRI studies have suggested that individuals with FXS have different neural processing organization but also posses the capacity to overcome executive functioning weaknesses by recruiting alternative brain regions to accomplish cognitive tasks in a compensatory fashion. These findings have been hopeful in terms of design of more targeted intervention to address deficit areas. These and other neurogenetic studies have been useful not only in understanding the brain basis of unique phenotypic features of FXS but also in shedding light on more common conditions such as attention-deficit/hyperactivity disorder (ADHD).

The Behavioral Phenotype

Although individuals with FXS may present with a wide range of behavioral and psychiatric difficulties, there are a number of symptoms that are very common and therefore are considered to be phenotypic of the disorder. These symptoms include inattention, hyperactivity, hand flapping, hand biting, poor eye contact, shyness, and other features of social anxiety, tactile defensiveness, aggression, and autistic behaviors such as social withdrawal, stereotypic speech, and gaze avoidance.

A number of studies have investigated the overlap between the behavioral symptoms of FXS and autism. An estimated 15% of females and 33% of males with FXS meet *DSM-IV* diagnostic criteria for autistic disorder, and the vast majority present with clinically significant symptoms on the autism spectrum – up to 67% of males and 23% of females. For example, many individuals have stereotypic behaviors such as hand flapping, whereas core social abilities are preserved and developmentally appropriate. These children are motivated to interact socially with others and recognize and respond to social and emotional cues from others. Obsessive–compulsive behaviors are also common and may co-occur with more significant autism symptoms. There is evidence that lower functioning individuals with FXS present with more significant autistic symptoms. Autism is not as common in females with FXS; however, social problems and anxiety are prominent. Screening studies show that approximately 2–6% of individuals with autism test positive for FXS; these studies have led to the recommendation that all males and females with autism be screened for the *FMR1* mutation with DNA testing. Interestingly, multiple studies have been unable to document an association between FMRP and autism in FXS, suggesting that environmental, modifying secondary genetic factors or function of genes regulated by FMRP might play a role. For example, work by our group documents that the serotonin transporter gene appears to modify behavioral phenotype expression in males with FXS. Furthermore, previous studies have shown that family environment effects are also associated with autism severity in males. Thus, although FXS is a single gene disorder with a well-described phenotype, it is becoming clear that considerable variability exists and that many environmental, genetic, and biological factors known to affect typically developing children also have a measurable impact on children with FXS.

Shyness and social anxiety are common features of FXS, and these symptoms can overlap with autism. Individuals often demonstrate signs of autonomic hyperarousal and panic symptoms when confronted by a change in usual routine, a novel social situation, or in larger social gatherings. Selective mutism is another manifestation of anxiety in FXS, and it seems to occur more often in girls. Classically, individuals with FXS, especially males, turn their heads and look away when introduced to others. Over time, the anxious responses to these stimuli often can lead to significant social withdrawal and isolation. There is also evidence that individuals with FXS demonstrate increased sympathetic nervous system responses to basic sensory stimuli. Aggression toward others and self-injury appear to occur following anxiety-provoking situations or in response to

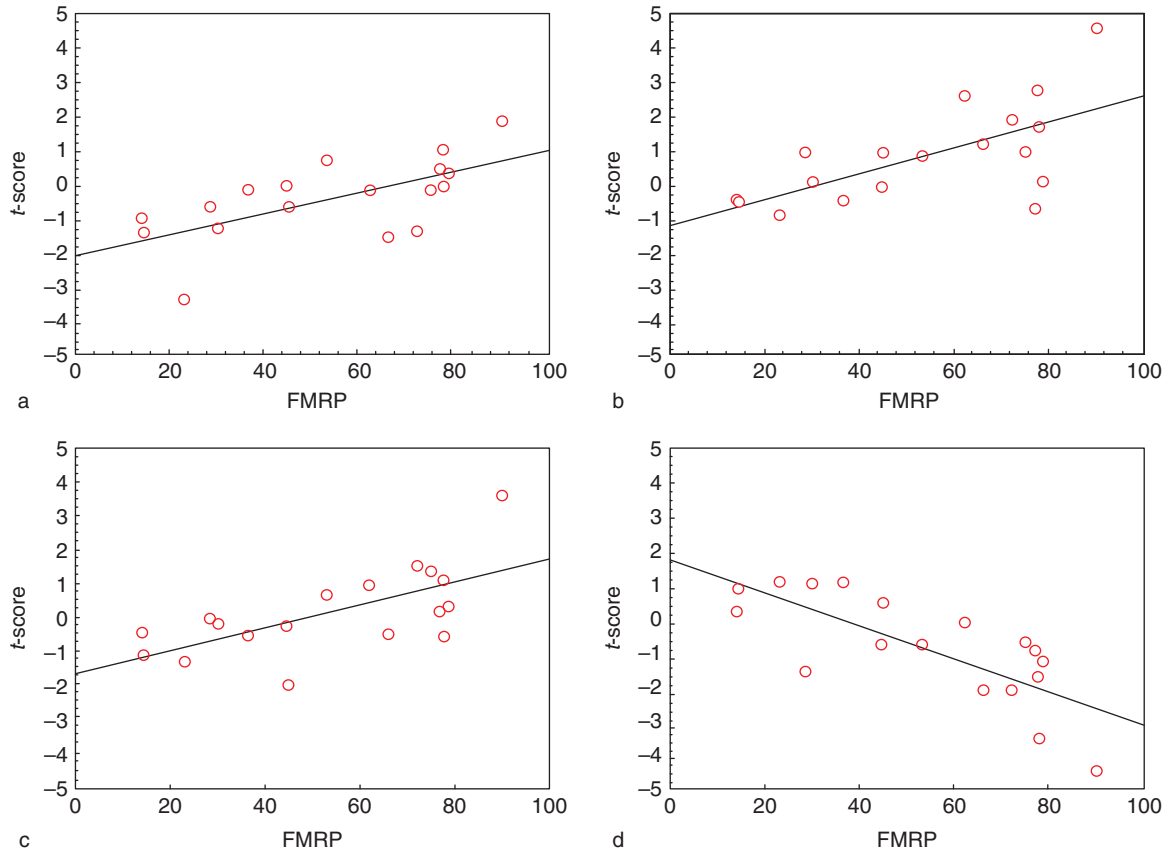

Figure 4 Positive correlation between FMRP and go/no-go-related activation in females with FXS in the left dorsolateral prefrontal cortex (a), right ventrolateral prefrontal cortex (b), and basal ganglia (c). Negative correlation between FMRP and activation in the ventromedial prefrontal cortex (d). Reprinted from Menon V, Leroux J, White CD, and Reiss AL (2004) Frontostriatal deficits in fragile X syndrome: Relation to FMR1 gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 101 (10): 3615–3620, with permission.

aversive sensory input. Some internalizing behaviors that are prominent in FXS are associated with elevated levels of the stress hormone cortisol. Interestingly, it was found that the glucocorticoid receptor mRNA is one of several mRNAs regulated by FMRP. Although multiple clinicians and investigators have emphasized the prominence of anxiety symptoms in FXS, the prevalence of full-blown anxiety disorders such as social phobia or obsessive–compulsive disorder among these individuals is unknown.

A number of empirical studies and clinical evidence suggest that hyperarousal and poor modulation of sensory input contribute to a number of behavioral difficulties in FXS, including aggression, social withdrawal, and some of the autistic symptoms. There are likely to be links between biological systems regulating stress, sensory processing, and hyperarousal and these social–emotional difficulties. For example, studies have shown that cortisol and sympathetic responses decrease following social gaze aversion in individuals with FXS. Males with FXS also have high and stable heart rates and poor vagal modulation of heart rate, further suggestive of autonomic dysfunction. One study showed that prepulse inhibition of startle, a measure of sensorimotor gating, was significantly impaired or completely abolished in males with FXS and highly correlated with IQ, adaptive behavior, autistic symptoms, and inattention (Figure 5).

Psychosis has not been carefully studied in FXS; however, general behavioral studies and clinical experience suggest that it is relatively uncommon. Schizotypal and paranoid features can develop in females and higher functioning males, for example, when social phobia and anxiety is especially pronounced. Identification of psychotic symptoms in FXS is especially difficult because of communicative problems and insight limitations, and classic behaviors related to FXS such as self-talk and unusual gesturing or hand mannerisms may be misinterpreted as evidence of hallucination.

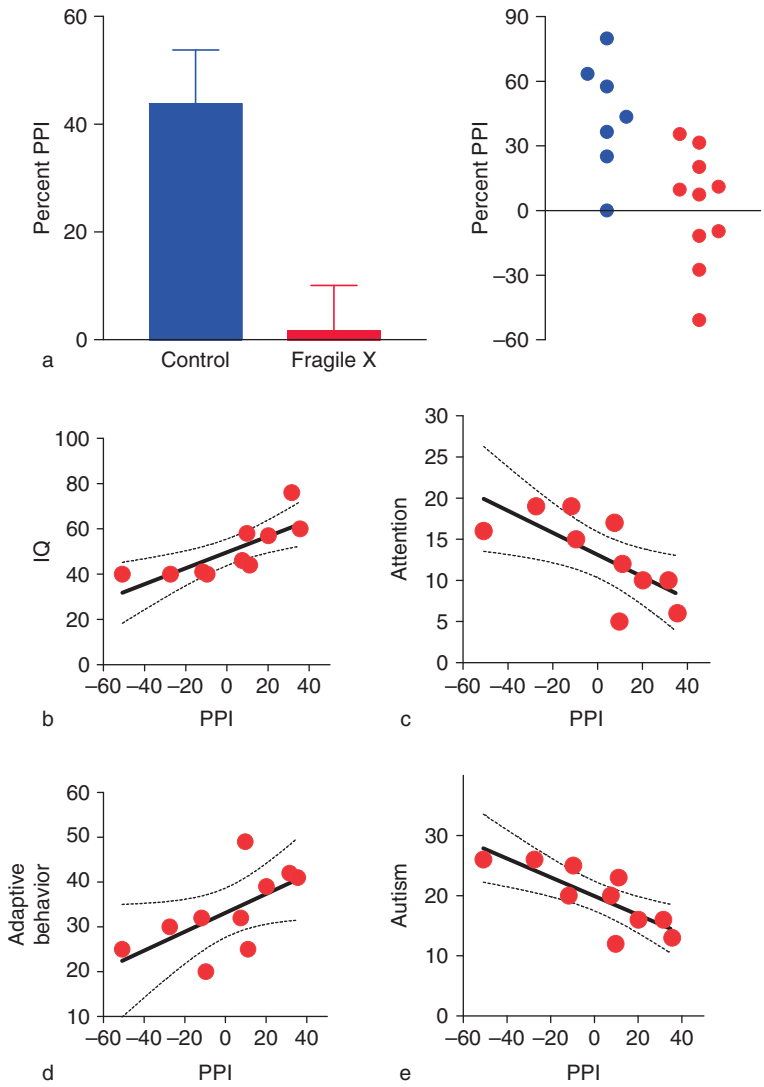

Figure 5 Prepulse inhibition of startle (PPI) is disrupted in young males with FXS. (a) Mean PPI is shown for control (blue) and FXS (red) subjects. Scatterplot showing range of PPI scores in control (red) and FXS (blue) subjects. (b) Scatterplot of IQ vs. PPI magnitude for FXS subjects. Lower IQ scores were associated with reduced levels of PPI. (c) Scatterplot of attention scores vs. PPI magnitude for FXS subjects. Attention deficits were associated with reduced PPI. (d) Scatterplot of adaptive behavior scores vs. PPI magnitude for FXS subjects. Maladaptive behavior was associated with reduced PPI. (e) Scatterplot of autism scores vs. PPI magnitude for FXS subjects. Subjects scoring higher on the autism scale were more likely to have reduced PPI levels. Courtesy of Edward Ornitz and reprinted from Frankland PW, Wang Y, Rosner B, et al. (2004) Sensorimotor gating abnormalities in young males with fragile X syndrome and Fmr1-knockout mice. *Molecular Psychiatry* 9: 417–425, with permission.

Fragile X Premutation: Occasional Features of FXS

This article focuses on FXS who typically have the *FMR1* full mutation. Some clinical aspects of the fragile X premutation are reviewed here because an FMRP deficit, without methylation, can occur in individuals with high CGG repeats within the premutation range and because these individuals can present with features of FXS. The fragile X premutation is defined by CGG repeat expansions between 55 and 200. The premutation occurs more often than the full mutation, in an estimated 1/813 men and 1/259 women in the general population. Those with expansions approaching the full mutation range (upper 100s) may have decreased transcriptional efficiency leading to FMRP reduction without any methylation (Figure 6). In addition, increased *FMR1* mRNA, ranging from 2 to 10 times higher than normal, is present, which may lead to a toxic effect in the brain. Two conditions unique to the premutation population have been specifically attributed to the toxic effects of the premutation mRNA excess. A proportion of older

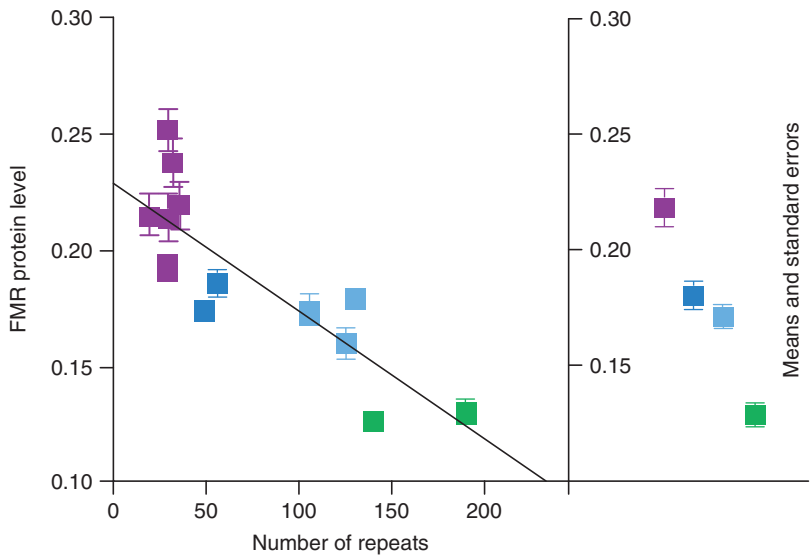

Figure 6 FMRP levels as a function of CGG repeat tract length. Molar ratio (mean and SE) of FMRP vs. number of repeats for 15 cell lines. The means and standard errors of each of the four groups are depicted on the left: normal (purple), intermediate (dark blue), premutation (light blue), and penetrant premutation (green) alleles. Reprinted from Kenneson A, Zhang F, Hagedorn CH, and Warren ST (2001) Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. *Human Molecular Genetics* 10(14): 1449–1454, with permission.

male and, less often, female carriers of the premutation develop the neurodegenerative disease FXTAS, a condition characterized by cerebellar ataxia, progressive intention tremor, and cognitive decline. A second condition thought to be caused by mRNA excess is premature ovarian failure, defined as the cessation of menses before the age of 40 years. This occurs in approximately 20% of female carriers, in contrast to 1% of the general population. The protein deficit found in some premutation carriers has been identified as the cause of clinical involvement similar to that in the full mutation, though less severe (Figure 1(c)). Most fundamentally, some of the same prototypical dysmorphology and behavioral features found in individuals with FXS full are found in an unknown proportion of premutation carriers with protein deficits. Finally, there are reports of children with the premutation having autism spectrum disorder (ASD), although the cause may be a combination of factors rather than a protein deficit alone. A study of non-clinic-referred children showed that 14% of males and 5% of females with the premutation were classified as having clinically significant symptoms on the autism spectrum. Indeed, recruitment bias problems in some of these studies have made it difficult to conclude with certainty that the premutation actually contributes to expression of autism. Carefully controlled studies are needed in this area to document the prevalence and molecular–biological mechanisms involved in expression of ASD symptoms in those with the premutation.

Summary

This article has summarized the primary physical, cognitive, and behavioral clinical features of. Because the brain and molecular genetic basis of FXS is relatively well understood through numerous human studies and animal models, it provides a useful model for understanding links between genes, brain, and behavior; specific conditions such as autism, ADHD, and anxiety disorders; and for testing pharmacological, behavioral, and gene therapies. For example, the discovery that *FMR1* knockout mice (the animal model for FXS) have enhanced long-term depression (LTD; weakening of synapses) in the hippocampus has led to a treatment model involving the role of a drug to block the metabotropic glutamate receptor that is critical in LTD and thereby mitigate symptoms of FXS. Some of the behavioral and physiological measures shown to be abnormal in patients with FXS may be useful in future clinical trials.

See also: Autism; Development of Behavior; Fragile X Syndrome; Fragile X Associated Tremor Ataxia Syndrome; Genetic Influence on CNS Gene Expression: Impact on Behavior.

Further Reading

Crnic LS and Hagerman R (2004) Preface: Fragile X syndrome: Frontiers of understanding gene–brain–behavior relationships. *Mental Retardation and Developmental Disability Research Reviews* 10(1): 1–2.

Frankland PW, Wang Y, Rosner B, et al. (2004) Sensorimotor gating abnormalities in young males with fragile X syndrome and Fmr1-knockout mice. *Molecular Psychiatry* 9: 417–425.

Hagerman RJ (2006) Lessons from fragile X regarding neurobiology, autism, and neurodegeneration. *Journal of Developmental and Behavioral Pediatrics* 27(1): 63–74.

Hagerman RJ and Hagerman PJ (2002) *Fragile X Syndrome: Diagnosis, Treatment, and Research,* 3rd edn. Baltimore: Johns Hopkins University Press.

Hessl D, Dyer-Friedman J, Glaser B, et al. (2001) The influence of environmental and genetic factors on behavior problems and autistic symptoms in boys and girls with fragile X syndrome. *Pediatrics* 108(5): E88.

Hessl D, Glaser B, Dyer-Friedman J, et al. (2002) Cortisol and behavior in fragile X syndrome. *Psychoneuroendocrinology* 27(7): 855–872.

Kenneson A, Zhang F, Hagedorn CH, and Warren ST (2001) Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. *Human Molecular Genetics* 10(14): 1449–1454.

Loesch DZ, Huggins RM, Bui QM, et al. (2003) Effect of fragile X status categories and FMRP deficits on cognitive profiles estimated by robust pedigree analysis. *American Journal of Medical Genetics A* 122(1): 13–23.

Loesch DZ, Huggins RM, and Hagerman RJ (2004) Phenotypic variation and FMRP levels in fragile X. *Mental Retardation and Developmental Disabilities Research Reviews* 10: 31–41.

Menon V, Leroux J, White CD, and Reiss AL (2004) Frontostriatal deficits in fragile X syndrome: Relation to FMR1 gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 101(10): 3615–3620.

Miller G (2006) Biomedical research: A fix for fragile X syndrome? *Science* 312(5773): 521.

Miller LJ, McIntosh DN, McGrath J, et al. (1999) Electrodermal responses to sensory stimuli in individuals with fragile X syndrome: A preliminary report. *American Journal of Medical Genetics* 83(4): 268–279.

Reiss AL and Dant CC (2003) The behavioral neurogenetics of fragile X syndrome: Analyzing gene–brain–behavior relationships in child developmental psychopathologies. *Development and Psychopathology* 15(4): 927–968.

Rivera SM, Menon V, White CD, Glaser B, Glover G, and Reiss AL (2002) Functional brain activation during arithmetic processing in females with fragile X syndrome is related to FMR-1 protein expression. *Human Brain Mapping* 16(4): 206–218.

Tassone F, Hagerman RJ, Ikle DN, et al. (1999) FMRP expression as a potential prognostic indicator in fragile X syndrome. *American Journal of Medical Genetics* 84(3): 250–261.

Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination

D E Feldman, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Neural activity, in the form of spontaneously generated or sensory-evoked spiking, is required for early postnatal development of synapses and brain circuits. Neural activity drives two types of changes in developing glutamatergic synapses: physiological changes that alter the function of existing synapses, and morphological changes that alter synapse structure. Physiological changes include long-term potentiation (LTP) and long-term depression (LTD), which strengthen and weaken synapses, respectively. Structural changes include synapse formation and elimination, which change the physical wiring of developing circuits. Structural plasticity can also include changes in dendritic spines and axonal boutons, and large-scale changes in axonal and dendritic branch patterns.

Functional and structural plasticity are thought to be highly coordinated, so that activity drives physiological strengthening and anatomical stabilization of a subset of synapses, and weakening and physical removal of others. By selecting appropriate synapses to strengthen and retain, developmental synaptic plasticity leads to the emergence of optimized neural circuits. In mature animals, some of these developmental plasticity mechanisms are downregulated, but others are retained, and are thought to contribute to information storage and learning in adults.

Activity-Dependent Maturation of Glutamate Receptor Content

Immature glutamate synapses contain a different mixture of α-amino-3-hydroxy-5-methyl-4-isoxazole-propionic acid (AMPA), *N*-methyl-D-aspartate (NMDA), and kainate (KA) subtypes of ionotropic glutamate receptors than mature synapses. Glutamate receptor composition gradually matures during early postnatal development, as a result of activity-dependent synaptic plasticity, and this maturation contributes to developmental strengthening of synapses.

Three aspects of glutamate receptor maturation have been studied in detail:

- progression from 'silent' synapses to 'functional' synapses,
- insertion of AMPA receptors to strengthen synapses, and
- developmental maturation of NMDA and AMPA receptor subtypes.

'Silent' Synapses Containing Only NMDA Receptors

Many immature glutamate synapses exhibit NMDA receptor-mediated currents without any detectable AMPA receptor currents. These are termed 'silent' synapses, because they do not generate synaptic current when the postsynaptic neuron is at rest, as a result of the voltage dependence of NMDA receptors. Silent synapses are proposed to represent synapses that contain surface NMDA receptors, but no surface AMPA receptors, although alternate interpretations exist. Silent synapses are highly prevalent early in postnatal development, and are downregulated the first few postnatal weeks in many brain areas, as AMPA currents become more widely expressed (Figure 1).

Insertion of AMPA Receptors Converts Silent Synapses into Functional Synapses, and Is a Basic Mechanism for Strengthening Synapses

Strong, correlated firing drives LTP at silent synapses, which is manifest by the rapid appearance of AMPA receptor currents. This appearance of AMPA receptor currents renders silent synapses functional (nonsilent), and is due to activity-dependent insertion of AMPA receptors. Thus, the developmental loss of silent synapses is thought to reflect the gradual accrual of LTP by endogenous activity, which converts early, silent synapses into functional, nonsilent synapses by adding AMPA receptors. The trigger for LTP is calcium entry through NMDA receptors, which detect correlated pre- and postsynaptic firing. Silent synapses have therefore been proposed to represent 'test synapses' that remain nonfunctional unless correlated activity between pre- and postsynaptic neurons renders them functional. At many nonsilent synapses, LTP is also thought to involve insertion of AMPA receptors, which increases AMPA content of synapses and enhances the postsynaptic response to released glutamate. This process has been observed directly in developing cortical synapses, and can be driven by sensory experience. Thus, AMPA receptor insertion is a major means of synaptic strengthening during development.

Developmental Maturation of NMDA Receptor Subunits

NMDA receptors are heteromultimers of NR1 and NR2 subunits, of which four types exist (NR2A-D).

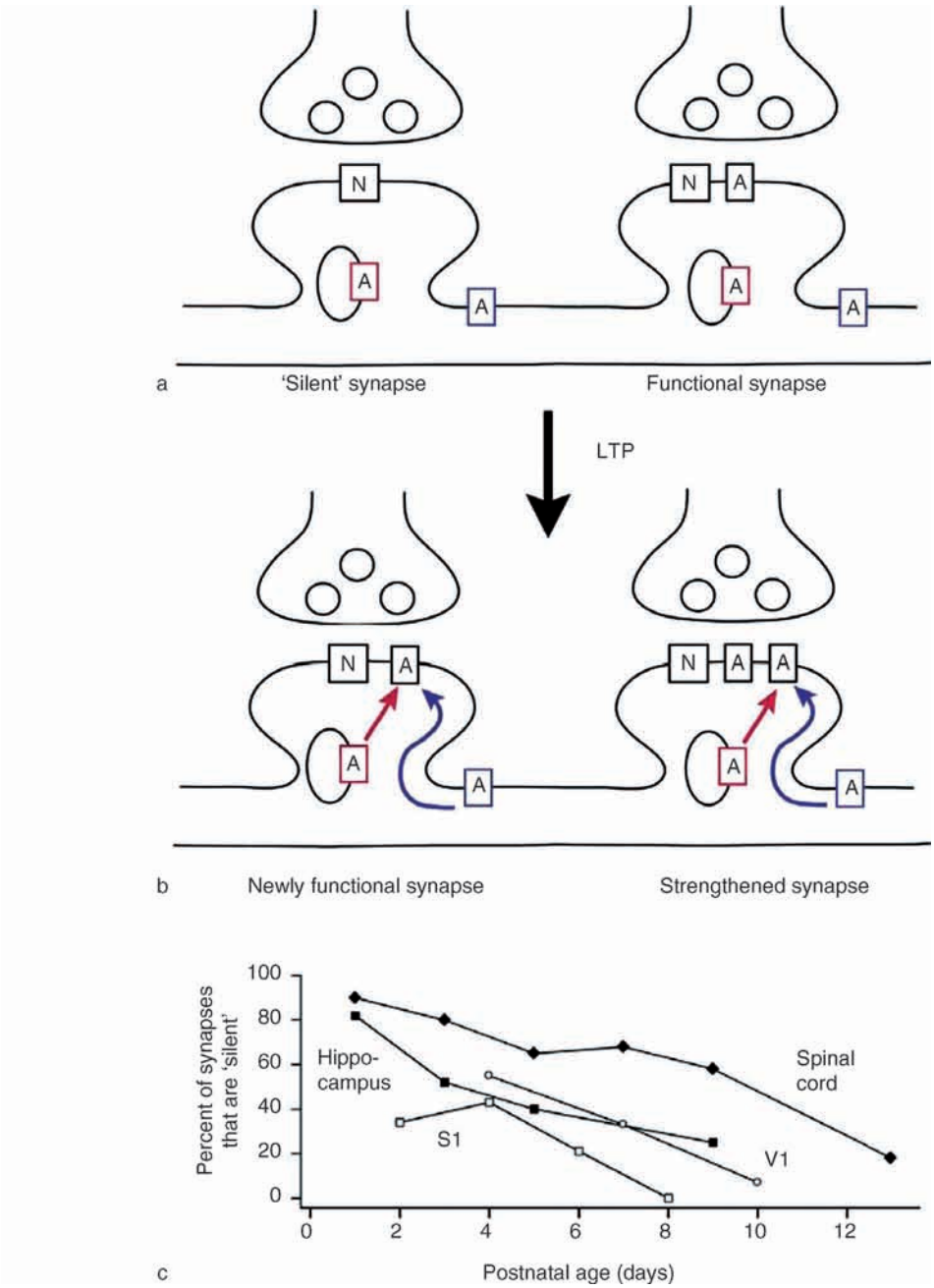

Figure 1 'Silent' synapses. (a) Newly formed glutamate synapses are a mixture of silent synapses, which contain NMDA receptors but not surface AMPA receptors, and standard functional synapses, which contain both receptor types. (b) Correlated activity drives LTP via insertion of AMPA receptors into silent synapses, either from subsurface stores (red) or extrasynaptic locations (blue). This causes silent synapses to become functional. Similar mechanisms act to strengthen previously functional synapses. (c) The proportion of silent synapses decreases during early development. Data are from dorsal horn of spinal cord, CA1 hippocampus, S1 cortex, and V1 cortex.

Immature synapses in neocortex and hippocampus express primarily NR1 and NR2B subunits, which produce long-lasting NMDA receptor currents. During the first few postnatal weeks, NR2A expression increases, resulting in NR2A-containing NMDA receptors that exhibit much shorter-duration currents. This developmental subunit switch is activity dependent, since sensory deprivation and activity blockade prevent the switch, while sensory experience hastens it. The functional importance of this subunit switch remains unclear, but one possibility is that faster NMDA receptor current kinetics enable faster temporal resolution for synaptic plasticity and information processing in mature versus developing circuits.

Changes in AMPA Receptor Subunits during Development

Developmental changes also occur in AMPA receptor subunits. AMPA receptors are composed of subunits GluR1–4, with GluR2 controlling calcium permeability of the AMPA receptor. Mature AMPA receptors contain GluR2 that has undergone RNA editing to make AMPA receptors calcium impermeable. In embryonic development, AMPA receptors lack GluR2, or contain an unedited version of GluR2, both of which render AMPA receptors calcium permeable. During the first two postnatal weeks, edited GluR2 becomes plentiful, and nearly all AMPA receptors on excitatory cells become calcium impermeable. This may serve to restrict calcium-dependent synaptic plasticity and alter information processing in mature networks.

LTP and LTD at Developing Excitatory Synapses

Developing synapses are capable of rapid, activity-driven changes in synaptic strength, including LTP and LTD. LTP and LTD are defined as a rapid, long-lasting increase or decrease, respectively, in strength of an existing synapse. LTP and LTD implement Hebbian plasticity, and are widely hypothesized to help mediate activity-dependent refinement of developing circuits.

The dominant theory is that initial synapses are generated relatively imprecisely, and that appropriate synapses are then strengthened by LTP, while inappropriate synapses are weakened by LTD, and ultimately eliminated. In support of this theory, initial glutamatergic synapses innervating a single postsynaptic neuron are often many and weak, and become fewer and stronger with maturation. This developmental refinement of connectivity occurs not only at glutamate synapses, but also at some inhibitory synapses, and at the neuromuscular junction. This synapse strengthening is hypothesized to reflect LTP at a subset of initially weak synapses, while the reduction in synapse number is thought to represent activity-dependent, anatomical withdrawal of selected synapses, perhaps following LTD (Figure 2).

LTP and LTD Occur Robustly at Developing Synapses

LTP and LTD appear to be near-universal properties of developing glutamate synapses (at least for synapses on excitatory neurons). LTP and LTD have been observed in developing neocortex, hippocampus, amygdala, striatum, spinal cord, nucleus accumbens, cerebellum, and other structures. LTP and LTD are often most robust during early critical periods of postnatal development. However, they can often still be induced, albeit sometimes with more difficulty, in adults. At many developing synapses, LTP is triggered by calcium entry via NMDA receptors, and involves

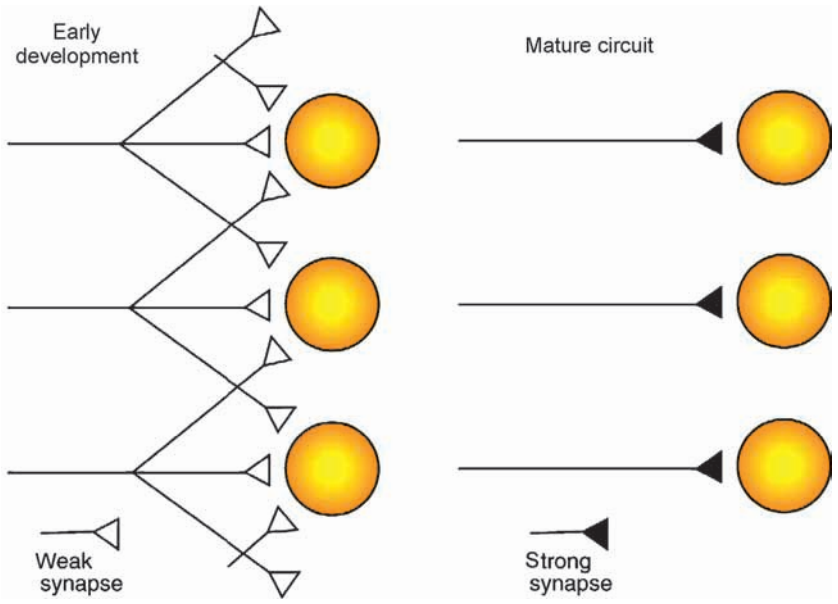

Figure 2 Synaptic refinement in developing networks. Immature networks are composed of weak, imprecisely targeted synapses. Many more presynaptic neurons synapse on each postsynaptic neuron than in mature circuits. During maturation, some synapses are eliminated, while others are strengthened and stabilized. This leads to the maintenance of fewer (sometimes only one), stronger inputs on each postsynaptic cell. Adapted from Hooks BM and Chen C (2006) A model for synaptic refinement in visual thalamus. In: Guido W, Erzurumlu R, and Molnar Z (eds.) *Development and Plasticity in Sensory Thalamus and Cortex*, pp. 226–244. New York: Kluwer/Plenum.

activity-dependent insertion of AMPA receptors, though other mechanisms also exist. LTD is more heterogeneous, existing in NMDA receptor-dependent and independent forms, and can involve either internalization of AMPA receptors or decreases in presynaptic release probability. The intracellular signaling mechanisms for LTP and LTD can be distinct in developing versus mature synapses.

Critical Periods for LTP and LTD during Development

LTP and LTD are often most robust, or entirely restricted to, specific critical periods of development. In sensory cortex, these critical periods for NMDA receptor-dependent LTP and LTD occur in the first postnatal weeks, and overlap with critical periods for experience-dependent plasticity of visual and somatosensory maps, suggesting that LTP and LTD may contribute to map plasticity (Figure 3). Why LTP and LTD are restricted to these early critical periods is not known. The developmental switch in NMDA receptor subunits, which coincides with the end of critical periods in some systems, does not seem to be the causative factor, because interference with this subunit switch in transgenic mice does not alter the critical period for LTP or experience-dependent plasticity. Another hypothesis, still unproven, is that GABAergic inhibition, which increases strongly during this same developmental period, prevents effective LTP induction.

Experience Drives LTP and LTD *In Vivo*

Studies have shown not only that developing synapses are capable of LTP and LTD, but also that sensory experience drives LTP and LTD at these synapses *in vivo*. For example, pharmacological or genetic blockade of LTP and LTD can interfere with experience-dependent map plasticity in developing somatosensory and visual cortex, suggesting that sensory map plasticity involves LTP and LTD. In addition,

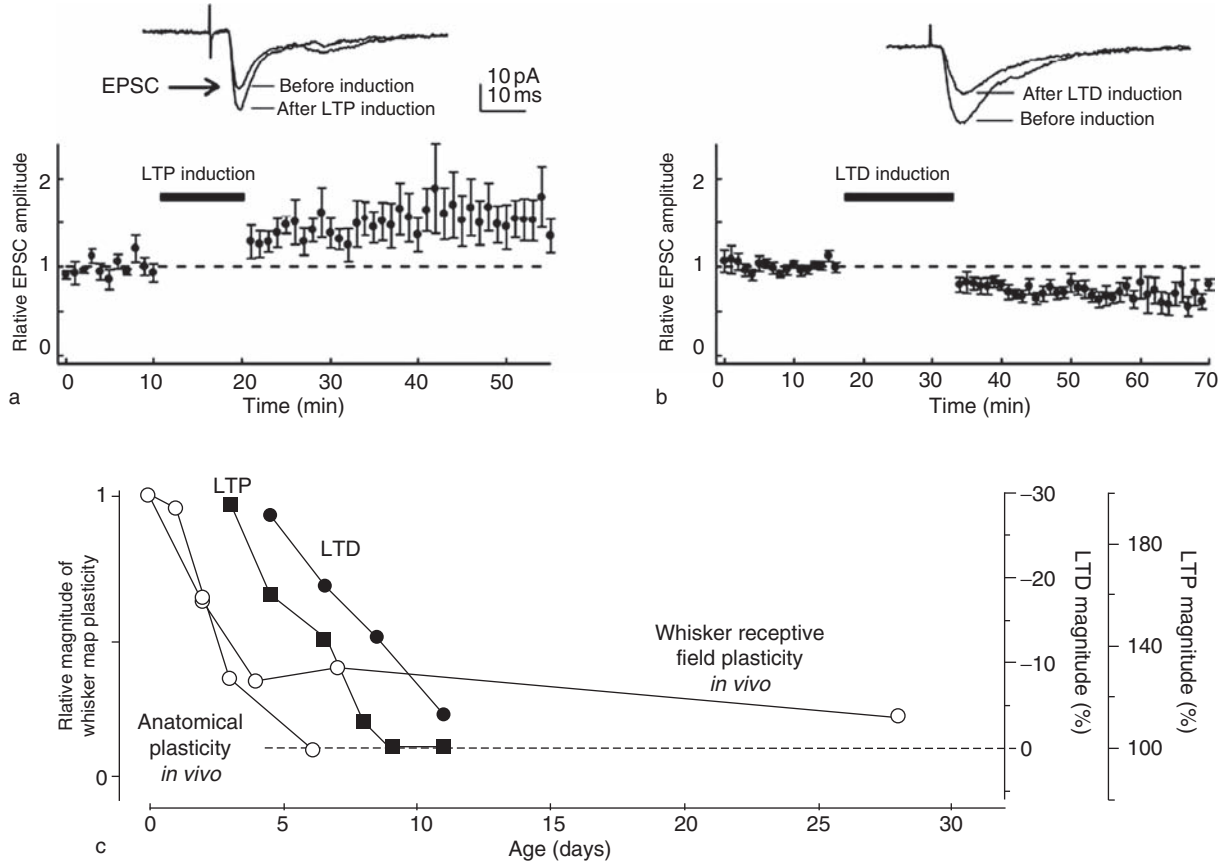

Figure 3 LTP and LTD at developing synapses. (a, b) Examples of LTP and LTD in developing somatosensory cortex *in vitro*. LTP and LTD are apparent as rapid, long-lasting changes in the amplitude of excitatory postsynaptic current (EPSC) evoked by a single presynaptic spike. LTP and LTD were induced after a stable baseline period by correlated activation of pre- and postsynaptic neurons (black bars). (c) Critical period for LTP and LTD at thalamocortical synapses in somatosensory cortex. The critical period for LTP and LTD overlaps with the critical period for somatosensory cortex plasticity *in vivo*, measured by experience-dependent changes in whisker map anatomy or receptive fields. After Feldman DE, Nicoll RA, and Malenka RC (1999) Synaptic plasticity at thalamocortical synapses in developing rat somatosensory cortex: LTP, LTD and silent synapses. *Journal of Neurobiology* 41: 92–101.

sensory deprivation-induced LTD has been directly measured at specific cortical synapses, and sensory use has been shown to drive LTP by AMPA receptor insertion into synapses in somatosensory cortex. It is therefore widely assumed that during normal development, network activity drives LTP and LTD and that this contributes to developmental refinement of circuits and response properties.

Other Forms of Synaptic Plasticity during Development

Developing synapses also exhibit other forms of plasticity, besides LTP and LTD. One example is homeostatic synaptic scaling. Synaptic scaling acts on a slower timescale than LTP or LTD to maintain neural activity at a stable, set-point level. Thus, increased network activity drives slow, cell-wide decreases in excitatory synapse strength, and decreased network activity drives increases in excitatory synapse strength. Synaptic scaling occurs in developing sensory cortex *in vivo*, and has been proposed to function to maintain stable network activity in the face of developmentally changing input levels. Scaling has also been proposed to prevent instabilities inherent in Hebbian plasticity. Developing synapses also exhibit another form of slow plasticity, termed metaplasticity, in which activity patterns change the future ability of cells to undergo LTP and LTD.

Experience-Dependent Structural Plasticity at Developing Synapses: Synapse Formation, Elimination, and Morphological Change

Sensory experience and neural activity also alter the structure of developing synapses. Such structural plasticity includes synaptogenesis, synapse elimination, changes in the morphology of existing synapses and dendritic spines, and larger-scale changes in axonal and dendritic branch morphology. Structural plasticity can be quite rapid (on the timescale of minutes), and has been proposed as a major basis for experience-dependent developmental plasticity and information storage. A key question, currently unanswered, is how structural plasticity is related to functional plasticity like LTP and LTD.

During normal development, synapses are continuously formed, altered, and eliminated, as demonstrated by live, time-lapse imaging of axons, dendrites, and dendritic spines. This ongoing structural remodeling is strongly regulated by sensory experience and neural activity, suggesting that it represents experience-dependent synapse formation and elimination, and decreases markedly in the adult. Recent studies have begun to reveal how activity drives structural plasticity and how this guides development of functional connectivity in developing networks.

Three major features of structural plasticity have been investigated in detail:

- developmental changes in morphology of dendritic spines,
- relationship between structural plasticity of spines and LTP and LTD, and
- structural plasticity of axonal arbors.

Developmental Changes in Morphology of Dendritic Spines

Dendritic spines are the site of most excitatory synapses on excitatory neurons. The earliest dendritic spines are thin and highly motile filopodia, and carry synapses which are thought to dynamically sample the nearby neuropil for appropriate synaptic partners. As development proceeds, filopodia are replaced by less motile spines with large bulbous heads, that are thought to receive mature, stable synapses. Because filopodia and small spines often contain detectable NMDA receptors but not AMPA receptors, it has been proposed that these structures represent newly formed, 'silent' synapses, while larger spines represent mature functional synapses. During postnatal critical periods of plasticity, spine motility persists, and is likely to reflect ongoing synapse elimination and formation. Spine motility decreases with age, in parallel with the decreased capacity for plasticity in older animals (Figure 4).

Relationship between Structural Plasticity of Spines and LTP and LTD

LTP and LTD are not unitary processes, but consist of distinct early and late temporal components. Early theories proposed that rapid components of LTP and LTD represented purely functional synaptic changes (e.g., AMPA receptor insertion), without macroscropic structural correlates. Slower, long-lasting components (which were known to be dependent on protein synthesis) were proposed to reflect synapse formation or elimination. Time-lapse imaging of living spines has recently provided qualified support for this hypothesis. LTP causes a rapid increase in spine head size, particularly for small spines, followed later by formation of new spines, while LTD causes a rapid shrinkage of spine heads. These rapid changes in spine head size may be structural correlates of the earliest changes in synaptic strength, perhaps AMPA receptor insertion or removal, while new spine formation may be related to later components of plasticity. These findings support an emerging model that immature synapses on small spines (with few or no AMPA

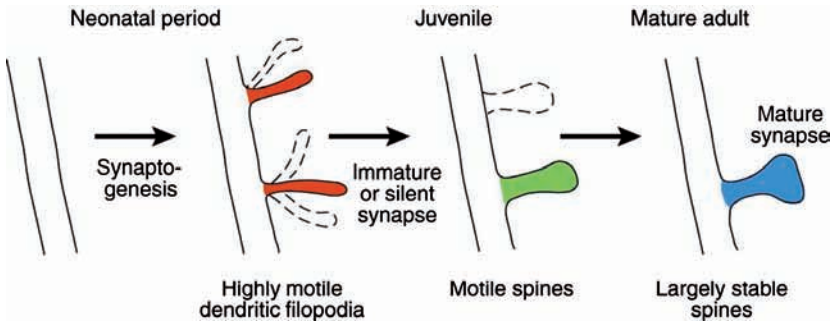

Figure 4 Development of spine morphology. The earliest dendritic protrusions are highly motile, thin filopodia, which are thought to carry weak or silent synapses. During postnatal development, many filopodia are lost, and are replaced by spines. In juveniles, spines dynamically appear and disappear, perhaps reflecting activity-dependent synaptogenesis and synapse elimination. In mature animals, spines become morphologically stable and carry mature synapses. Dashed outlines indicate transient filopodia or spines.

receptors) are converted by LTP into functional synapses with a full complement of AMPA receptors, carried on larger spines. However, additional experiments are needed to test this model fully. In particular, direct evidence that synaptogenesis and synapse elimination contribute to late phases of LTP and LTD remains elusive.

Structural Plasticity of Axonal Arbors

A final question is how activity-dependent synaptic plasticity shapes larger-scale structure of neurons, including axonal and dendritic branch patterns. Studies of the developing retinotectal projection in *Xenopus laevis* have shed light on this issue. Retinal ganglion cells (RGCs) from each eye innervate the opposite optic tectum of *Xenopus* to form a retinotopic map. During retinotectal development, RGC axons and tectal cell dendrites are highly dynamic, adding and retracting small synapse-bearing branches continously. To determine how activity guides axonal development, a surgical manipulation has been used that forces both eyes to innervate a single tectum. This causes activity-dependent segregation of RGC axons into eye-specific territories within the tectum, reminiscent of ocular dominance columns in V1. Time-lapse imaging reveals that RGC axons from one eye form new synapse-bearing branches uniformly in both eye-specific territories, but that branches are preferentially stabilized in the territory dominated by the same eye, and are preferentially eliminated in the territory dominated by the other eye. This differential stabilization/elimination requires NMDA receptor signaling, suggesting that it is guided by Hebbian synaptic plasticity.

These findings indicate that during development, new synapses are formed everywhere, but are selectively retained only when NMDA receptor-dependent Hebbian processes determine that they are appropriate. This supports the common model that activity strengthens and stabilizes some synapses and their associated spines and axonal/dendritic branches, while others are weakened and eliminated, causing their spines and axonal/dendritic branches to be removed.

Summary: Toward a Comprehensive Model of Functional and Structural Synapse Development

Many questions remain to be answered about how activity guides development of glutamate synapses and networks. However, the existing literature, summarized above, suggests a unified model for development of synapse function and structure. In this emerging model, many initial glutamate synapses are weak or silent, and contain an immature complement of glutamate receptors including NMDA receptors but few or no AMPA receptors. These immature synapses are carried on highly motile small spines or filopodia, which sample surrounding neuropil searching for appropriate synaptic partners. During development, activity causes glutamate receptor subtypes to mature. A subset of synapses will experience correlated pre- and postsynaptic firing, which activates NMDA receptors and drives synapse strengthening via Hebbian LTP, due in part to insertion of AMPA receptors. Synapse strengthening also causes structural maturation of spines, making them larger and less motile, and stabilizes the axonal and dendritic branches bearing these 'effective' synapses. Unreinforced, 'ineffective' synapses may be weakened by LTD, structurally destabilized, and removed.

In this model, activity engages Hebbian synaptic plasticity rules to both regulate synapse strength and to guide the formation of mature synaptic morphology and circuit-level connectivity. This allows initial networks comprised of many weak, imprecisely targeted synapses to be refined into

networks built from a smaller number of strong, appropriately targeted synapses.

See also: BDNF in Synaptic Plasticity and Memory; Dendrite Development, Synapse Formation and Elimination; Endocannabinoid Role in Synaptic Plasticity and Learning; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Glutamate Regulation of Dendritic Spine Form and Function; Long-Term Depression: Cerebellum; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Depression; Synaptic Plasticity and Place Cell Formation; Synaptic Plasticity: Short-Term Mechanisms; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Abraham WC and Bear MF (1996) Metaplasticity: The plasticity of synaptic plasticity. *Trends in Neurosciences* 19: 126–130.

Chklovskii DB, Mel BW, and Svoboda K (2004) Cortical rewiring and information storage. *Nature* 431: 782–788.

Cline HT, Wu GY, and Malinow R (1996) *In vivo* development of neuronal structure and function. *Cold Spring Harbor Symposia on Quantitative Biology* 61: 95–104.

Durand GM, Kovalchuk Y, and Konnerth A (1996) Long-term potentiation and functional synapse induction in developing hippocampus. *Nature* 381: 71–75.

Feldman DE and Brecht M (2005) Map plasticity in somatosensory cortex. *Science* 310: 810–815.

Feldman DE, Nicoll RA, and Malenka RC (1999) Synaptic plasticity at thalamocortical synapses in developing rat somatosensory cortex: LTP, LTD and silent synapses. *Journal of Neurobiology* 41: 92–101.

Fox K and Wong ROL (2005) A comparison of experience-dependent plasticity in the visual and somatosensory systems. *Neuron* 48: 465–477.

Hensch TK (2005) Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience* 6: 877–888.

Hooks BM and Chen C (2006) A model for synaptic refinement in visual thalamus. In: Guido W, Erzurumlu R, and Molnar Z (eds.) *Development and Plasticity in Sensory Thalamus and Cortex*, pp. 226–244. New York: Kluwer/Plenum.

Isaac JT, Crair MC, Nicoll RA, and Malenka RC (1997) Silent synapses during development of thalamocortical inputs. *Neuron* 18: 269–280.

Isaac JTR (2003) Postsynaptic silent synapses: Evidence and mechanisms. *Neuropharmacology* 45: 450–460.

Katz LC and Shatz CJ (1996) Synaptic activity and the construction of cortical circuits. *Science* 274: 1133–1138.

Li P and Zhuo M (1998) Silent glutamatergic synapses and nociception in mammalian spinal cord. *Nature* 393: 695–698.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* 25: 103–126.

Matus A (2000) Actin-based plasticity in dendritic spines. *Science* 290: 754–758.

Molnar E and Isaac JTR (2001) Developmental and activity dependent regulation of ionotropic glutamate receptors at synapses. *The Scientific World* 1: 1–20.

Nimchinsky EA, Sabatini BL, and Svoboda K (2002) Structure and function of dendritic spines. *Annual Review of Physiology* 64: 313–353.

Rumpel S, Hatt H, and Gottmann K (1998) Silent synapses in the developing rat visual cortex: Evidence for postsynaptic expression of synaptic plasticit. *Journal of Neuroscience* 18: 8863–8874.

Ruthazer ES and Cline HT (2004) Insights into activity-dependent map formation in the retinotectal system: A middle-of-the-brain perspective. *Journal of Neurobiology* 59: 134–146.

Turrigiano GG and Nelson SB (2004) Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience* 5: 97–107.

Wyatt RM and Balice-Gordon RJ (2003) Activity-dependent elimination of neuromuscular synapses. *Journal of Neurocytology* 32: 777–794.

Yuste R and Bonhoeffer T (2001) Morphological changes in dendritic spines associated with long-term synaptic plasticity. *Annual Review of Neuroscience* 24: 1071–1089.

Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus

A Convit, Nathan Kline Institute, Orangeburg, NY, USA
M Rueger, New York University School of Medicine, New York, NY, USA
O T Wolf, University of Bielefeld, Bielefeld, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The aim of this article is to present our understanding of how the abnormalities in peripheral glucose control and elevations in cortisol levels associated with type 2 diabetes mellitus (T2DM) may give rise to medial temporal lobe (MTL) pathology and dysfunction. The article falls into four main sections: Memory and aging, effects of glucocorticoid functioning on the brain, impact of diabetes on the brain, and, finally, an integrative model that attempts to explain how the specific MTL dysfunction may come about. First, we give a brief introduction on normal and pathological changes in the aging brain before we explain how hypothalamic–pituitary–adrenal (HPA) axis functioning, that is, cortisol secretion and control, can affect brain and cognition. Next, we summarize evidence that insulin resistance and type 2 diabetes can contribute to a reduced integrity of the MTL memory system. In this context we highlight the multiple interactions between the glucoregulatory system and the neuroendocrine stress system. The last section presents an explanatory model of how we conceptualize the above-mentioned processes give rise to the MTL impairments, and provides a framework for the generation of specific hypotheses for future research in this area.

Memory in Aging: Normal Changes, Mild Cognitive Impairment, and Dementia

Cross-sectional studies of human aging have consistently reported that, relative to younger individuals, normal elderly show performance reductions in memory, attention, visual–spatial function, abstraction, and problem solving. Longitudinal studies provide a refinement by demonstrating that over time the most salient declines are in tests sensitive to frontal lobe function, namely working memory (requiring the short-term storage and manipulation of information prior to responding), psychomotor efficiency, word knowledge, attention, learning, and problem solving. Although the majority of this literature suggests that most cognitive deficits are related to frontal dysfunction, some research suggests that MTL, which includes structures such as the hippocampus, amygdala, and entorhinal cortex, may also be involved. In normal aging cognitive deficits do not generally result in impairments in everyday functioning.

Memory impairments have been recognized for many years as the most reliable early clinical symptom in dementia of the Alzheimer's type, and both the pathology – namely the distribution of neurofibrillary tangles (one of the two types of brain pathology that are necessary for a brain diagnosis of Alzheimer's disease (AD)) – and the *in vivo* imaging literature indicate that the earliest lesions are in the MTL. In addition, among elderly individuals who demonstrate subtle impairments in functioning and who may be considered at higher risk for significant cognitive decline and AD, those who have so-called mild cognitive impairment (MCI), demonstrate predominantly MTL pathology in the earliest stages of cognitive deterioration. With that being said, most middle age and elderly individuals who present with memory and other cognitive impairments, but with no impairments in overall functioning, will not go on to develop dementia.

Links between Diabetes and Dementia

There is evidence from longitudinal population-based studies that the risk of dementia is higher among individuals with diabetes than among those without it. The risk seems to be higher for individuals with the Apo E ε 4 genotype or receiving insulin treatment. The Apo E gene, which has a polymorphic distribution in the population, codes for a protein that is important in lipid transport and metabolism. Individuals with one or two copies of the Apo E ε 4 allele have a higher risk for cardiovascular disease. However, findings have been inconsistent as to whether this increased risk is specific for AD, multi-infarct dementia (MID), or dementia in general. Although the associations between diabetes and dementia are intriguing, none of these studies has directly assessed the impact of T2DM-related factors. For example, the impacts of hypertension and cardiovascular disease, conditions associated with T2DM, cognitive decline, and dementia, have not been evaluated. Although diabetes is often associated with vascular pathology ranging from micro-vessel disease to hypertension and stroke, studies to date have not provided sufficiently detailed data on those diabetes-associated factors to better ascertain how they may mediate or modulate the risk for dementia associated with diabetes.

The magnitude of the risk for AD related to diabetes is only modest and smaller than that of some nonspecific risk factors, such as low education or history of head trauma. It is possible that the impact of diabetes, analogous to that of low education and head trauma, is to reduce the brain reserve of affected individuals, and thus the associations observed between diabetes and AD may also be nonspecific. This position is supported by the extant pathology literature. Out of the autopsy series available, most fail to find increased Alzheimer's pathology in those individuals with diabetes (with some studies actually finding reduced AD pathology among diabetics), and only one series found a nonsignificant increase among Apo E ε 4 diabetics. However, based on indirect evidence some investigators have speculated that diabetes may be 'mechanistically' linked to AD. This hypothesis is based on observations that insulin acutely enhances memory in healthy older subjects, but is less able to do so in AD patients. In addition, insulin might increase levels of inflammatory cytokines and amyloid beta in the brain. Future research will test these hypotheses.

The Impact of Cortisol on Cognition and the Brain

Stress is conceptualized as a perceived challenge to the organism's balance or homeostasis, and the resulting processes intended to return the organism to equilibrium are termed 'allostasis.' The most prominent of these processes is the activation of the HPA axis. This activation results in the release of glucocorticoids (GCs) and the activation of the sympathetic nervous system (SNS), leading to the release of catecholamines (epinephrine and norepinephrine). In the presence of an acute stress these responses protect and allow the organism to adapt to its changed internal or external environment. However, when the stress becomes chronic or the organism responds inappropriately (e.g., poor habituation or a failure of negative feedback), multiple systems, including the brain, are negatively impacted. This state has been called 'allostatic load.'

GCs, depending on their timing, can have positive or negative acute effects on memory. For example, stress (or an acute increase in GCs) during learning, or immediately thereafter, leads to enhanced memory consolidation. In contrast, an equivalent rise in GCs shortly before delayed retrieval causes impaired retrieval. Animal and human studies suggest that both the positive and negative GC effects are mediated by interactions between amygdala and hippocampus, two limbic structures central in emotional processing and memory formation.

Chronic stress leads to negative consequences in both animals and humans. In animals, chronic stress causes structural alterations of neurons (dendritic atrophy or dendritic remodeling) in hippocampus and prefrontal regions. In addition, stress negatively impacts neurogenesis and alters cholinergic, serotonergic, and dopaminergic systems. These changes are associated with impaired memory performance, although the mediating mechanisms remain unclear. For example, both rats and tree shrews subjected to repeated stress develop impairments in hippocampal-based memory tasks. In contrast, amygdala-mediated tasks, such as fear conditioning, are enhanced. Age affects the organism's response to stress. Although there is large individual variability, aged individuals have increased GC levels as well as impairments in HPA axis negative feedback, both of which have been associated with cognitive impairments. For example, rodents with age-associated increases in HPA activity had impairments in memory performance, whereas those with normal levels did not. Moreover, pharmacological or behavioral interventions that stabilized HPA activity throughout life offered protection from those age-associated memory problems.

In the human, conditions leading to substantially elevated cortisol levels are characterized by impairments in memory, attention, and mood. For example, the hypercortisolemia present in Cushing's disease is associated with reductions in memory performance and hippocampal atrophy, deficits that are at least partly reversible upon successful reduction of the supraphysiological cortisol levels. Elevations in basal GC levels within the normal range have also been associated with cognitive dysfunction; older individuals with higher cortisol levels show poorer memory and faster cognitive decline. However, the causal order of these associations remains unsettled. The behavioral impairments may be secondary to cortisol-induced hippocampal atrophy. However, given the importance of the hippocampus on cortisol feedback inhibition, the cortisol elevations may be the result rather than the cause of hippocampal atrophy. Nevertheless, some recent studies offer indirect support for primacy of cortisol elevations leading to hippocampal damage. For example, self-reported stress proneness is associated with an increased risk for dementia, a disease that involves hippocampus and other MTL structures in its initial stages. Further indirect support is offered by the finding that individuals with genetically determined increases in tissue GC bioavailability have a higher risk for AD.

HPA axis dysregulation, in addition to possible direct effects on the hippocampus, may affect memory through its impact on other related systems also known to affect memory. For example, GCs can

impair insulin sensitivity or lead to enhanced visceral fat deposition and altered food intake, which in turn can lead to problems with glucose control, conditions known to be associated with reductions in memory performance. Interestingly, the adverse consequences of stress on the hippocampus occur faster in diabetic animals, suggesting an increased vulnerability of the diabetic brain. The interactions between the HPA axis and the glucoregulatory system will be illustrated in the following section.

Interactions between Cortisol, Insulin Function, and Type 2 Diabetes

Research conducted during the past decade has demonstrated that the HPA axis and the glucoregulatory system interact at multiple levels. Examples can be found for interactions at the behavioral and system level as well as on particular aspects of neuronal or transporter functioning.

Cortisol secretion and peripheral glucose regulation are closely linked. For example, we know that physiologically increased cortisol acutely inhibits insulin release, and that pharmacological doses of GCs cause reductions in glucose disposal (insulin resistance). In addition, in animal models GCs reduce the amount of insulin transported across the blood–brain barrier (BBB), and in large doses inhibit neuronal and glial glucose uptake in the hippocampus but not in other brain regions (e.g., hypothalamus, cerebellum, or cortex). Similar effects have been noted in the human; we have shown that acute cortisol administration leads to a specific *in vivo* hippocampal reduction in glucose utilization as seen with positron emission tomography. It is noteworthy that the hippocampus is the brain area with greatest co-localization of cortisol and insulin receptors. It is possible that the GC-mediated reductions in memory performance are due to their impact on hippocampal energy status.

Chronic stress in animals leads to effects at a behavioral level. Investigators have documented that chronic stress, through a facilitating effect on the HPA axis, leads to changes in eating habits. These animals seek more 'comfort' food and the GC elevations promote visceral fat deposition. In the human, chronic stress, with the associated hyperactivity of the SNS and HPA systems, can lead to the metabolic syndrome, which is characterized by central obesity, impairments in glucose regulation, hypertension, and abnormal lipid profiles.

Problems with peripheral glucose control may potentiate the deleterious effects that cortisol exerts on the brain. For example, when diabetic rats are stressed, they develop extensive hippocampal damage in one-third the time it takes nondiabetic animals to develop equivalent stress-mediated damage. It has also been shown that diabetic rats have chronic elevations of basal GCs and have greater and more prolonged responses to stress. We also know that reduced hippocampal integrity may lead to impaired cortisol feedback inhibition, and thus to elevated cortisol secretion. Adding support to the notion that the stress and the glucoregulatory systems interact in their influence of cognition during aging are observations from the MacArthur Study on Successful Aging. These investigators reported that summary measures of allostatic load, consisting of a combination of stress hormone measures and metabolic syndrome measures, were a superior predictor of cognitive and functional decline in older individuals.

Although it is difficult to ascertain cause and effect, it may be that abnormal glucose tolerance causes hippocampal dysfunction, which then affects cortisol feedback control, which in turn may result in elevated cortisol levels, thus contributing to a vicious cycle of further hippocampal damage.

Impact of Diabetes on the Brain

Neurons neither synthesize nor store glucose and are hence dependent on external sources to meet their energy demands. For a long time it was thought that the brain sat in a privileged position and that its glucose supply was independent of what was happening in the periphery. Only in recent years has this view changed, and it has become apparent that glucose transport into the brain is affected by the peripheral environment and that insulin, insulin receptors, and insulin-sensitive glucose transporters (GLUTs) are all present in brain, although their function remains unclear.

Under normal physiological circumstances the metabolism of glucose borne by the blood accounts for 99% of the brain energy needs. Although glia and astrocytes contain glycogen, the contributions of this glycogen to the overall energy demands of the brain are not known. (Note: It is our opinion that understanding the role of glycogen in brain will be an area of great interest in the near future. This opinion is based on the fact that upon restoration of euglycemia after an episode of hypoglycemia, glycogen deposition in brain goes up markedly, perhaps protecting the brain from subsequent hypoglycemic episodes. In addition, experimental animals subjected to weekly episodes of significant hypoglycemia perform better than sham animals on memory tests during euglycemia, and this improved memory performance is preserved in these animals in old age. Whether these memory-enhancing and protective effects are due to hypoglycemia-based improved transport of glucose

across the BBB or to increased deposition of brain glycogen, or both, has yet to be established.)

Despite the fact that the brain depends on peripheral glucose, up until 10 years or so ago, it was commonly believed that the brain was able to obtain all the metabolic substrate it needed, irrespective of what was happening in the periphery. Although the transport of glucose across the BBB remains poorly understood, we now know that the peripheral environment can exert significant influences on BBB transport. For example, the BBB glucose transporter can be up- or downregulated, dependent on the sustained peripheral glucose levels. Additional potential links between systemic and brain glucose control and metabolism are offered by the presence of insulin and insulin-sensitive glucose transporters (GLUT4) in the brain. In peripheral tissues insulin regulates glucose homeostasis through effects on GLUT4, the glucose transporter highly expressed in fat and muscle. GLUT4 is also expressed in brain, including the hippocampus. However, it is not known if insulin affects glucose uptake into the hippocampus. Moreover, it remains to be seen how brain GLUT4 behaves in a state of peripheral insulin resistance. Regardless, it is clear is that glucose uptake and utilization in the brain are complex and that their regulation is yet to be understood.

Because impaired insulin function (or absence of sufficient insulin) and the resulting hyperglycemia are at the core of diabetes, it is likely that these fluctuations will have an impact on the brain and its functioning. There is mounting evidence that diabetes, whether it is type 1 or type 2, has a negative impact on both cognition and brain. However, the resultant adverse effects may vary by the type of diabetes. A straightforward assessment of the differential impact of type 1 and type 2 diabetes on cognition and brain is complicated by the fact that most studies only contrast one type of diabetic patient against controls and do not control for comorbid conditions or possible mediating factors such as hypertension, vascular disease, and depression, which are often associated with diabetes.

Adults with type 1 diabetes mellitus (T1DM), when contrasted to nondiabetic controls, show reduced performance on several cognitive domains, including psychomotor efficiency, intelligence, visual and sustained attention, and speed of information processing. The majority of studies of adults with T1DM demonstrate that the presence of macrovascular disease, but not glycemic control or the number of hypoglycemic episodes, contributed to the lower cognitive performance described. Cognitive deficits are also observed in children and adolescents with T1DM, but there's no consensus in the literature as to whether age of onset or disease duration affects the severity of those cognitive deficits. Furthermore, it remains unclear whether the cognitive impairments worsen with age. In addition, the nature of the impact on cognitive functioning of episodes of either hypoglycemia or hyperglycemia remains unclear. Different studies have evaluated study participants of various ages and have used different test batteries, which do not permit a straightforward comparison across studies.

In contrast to the findings in T1DM, studies of T2DM show mostly impairments in verbal memory and processing speed, whereas other cognitive domains, such as visuospatial function and attention, remain preserved. There is also evidence that in T2DM, glycemic control might influence the cognitive deficits; improvements in glucose control by either pharmacological or behavioral interventions have been associated with attenuation of the cognitive deficits. Additionally, cognitive deficits seem to be augmented by factors associated with T2DM, such as hypertension, depression, and vascular disease. Some studies suggest that severe memory impairments are only seen in older T2DM populations and that the cognitive deficits seen in T2DM represent an acceleration of normal brain aging.

There are only a few *in-vivo* imaging studies in diabetes, and existent studies have not contrasted T1DM and T2DM. The overall findings are that type 1 diabetes leads to general brain atrophy, whereas T2DM seems to preferentially involve MTL structures, and that these findings may be independent of vascular complications. In our own work we have demonstrated that among relatively young individuals with well-controlled T2DM, only reductions in memory, performance, and hippocampal volume separate them from age-, gender-, and education-matched normal controls. In addition, we have described very similar cognitive and imaging findings among normal middle-aged and elderly nondiabetic individuals with impairments in glucose tolerance. Few studies evaluate the brain in T1DM, and there is only one study of adults with T1DM that looked at both cognition and brain with magnetic resonance imaging (MRI). One small pilot study in adults with longstanding T1DM observed reductions in psychomotor speed and selective attention which was accompanied by global cerebral atrophy (increase in cerebrospinal fluid and decrease in total brain volume). However, no evidence for hippocampal volume reductions and memory impairments were detected.

In recent years, paralleling the obesity epidemic, the prevalence of T2DM within the population has risen even among children and adolescents. To date there are no reports in the literature evaluating cognition and brain status among children with

T2DM. Preliminary pilot data from our group shows widespread cognitive impairments among obese adolescents with T2DM when contrasted with age-, gender-, and education-matched obese controls.

The precise mechanisms by which diabetes and impaired glucose regulation affect memory are not clear, but the risk factors that are often associated with insulin resistance and that are part of the metabolic syndrome, namely hypertension and dyslipidemia, might play a role. In the next section we outline a model of how diabetes directly affects brain structures that are central for memory process, that is, the hippocampus.

Possible Mechanisms for the Cognitive Impairments Seen in T2DM

Glucose is transported across the BBB by GLUT1, a glucose transporter highly expressed in the vascular endothelial cells of the BBB. Although the regulation of glucose transport into the brain is not well understood, GLUT1 has very high affinity for glucose and is saturated at normal blood glucose levels. Previous work in animal models has established that long-term elevations in peripheral glucose result in decreased transport of glucose across the BBB, whereas the opposite is true for chronic hypoglycemia, in which there is an increase in transport across the BBB. In both these circumstances it is believed that the amount of GLUT1 expressed at the BBB is responsible for the changes in glucose transport.

During brain activation, there are measurable shifts in the local concentrations of available glucose. For example, visual stimulation reduces visual cortex glucose concentrations as seen with MRI-based spectroscopy. Brain glucose levels can be measured directly with microdialysis probes in free-moving rats. Using these techniques there is evidence that glucose is compartmentalized in brain, and that different brain areas control their glucose levels locally. For example, glucose drops are restricted to the hippocampus when the animal is performing a memory test, and are present despite constant peripheral glucose levels.

The BBB glucose transporter, GLUT1, is saturated at normal physiological glucose concentrations. Therefore, to acutely increase net glucose flux across the BBB, the number of GLUT1 proteins exposed to the blood must be increased. Dilatation of the capillary bed will expose more endothelial cells, with their corresponding GLUT1, to the blood, and thus increase the acute transport of glucose into the brain. However, the mechanisms responsible for regulating the rapid changes in local blood flow that are associated with brain activation are not yet well characterized. What is currently well described is that both individuals with T2DM and individuals with insulin resistance short of diabetes show impairments in endothelial-dependent vasodilatation. Consequently, given that cognitive testing leads to activation (and drops in the interstitial glucose levels in the areas activated), this endothelial dysfunction may lead to a 'functional hypoglycemia' and contribute, at least in part, to the cognitive deficits associated with diabetes and insulin resistance.

It is not known whether elevated cortisol levels affect glucose transport at the BBB (Figure 1). However, we know from animal studies that glucocorticoid exposure inhibits glucose transport into hippocampal neurons and glia. We also know that individuals with T2DM (and insulin resistance) have abnormalities in the HPA axis, particularly in feedback control. Consequently, we propose that, analogous to the reduced ability to locally increase glucose flux across the BBB with activation due to the reduced vascular reactivity in T2DM and insulin

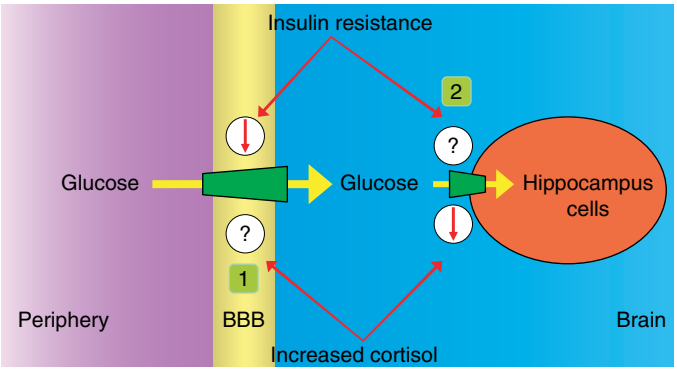

Figure 1 Representation of the blood and the brain compartments separated by the BBB. The green trapezoids represent the GLUTs. Reduced glucose transport across the BBB (step 1) and/or into the cells in the hippocampus (step 2), particularly during activation, may result in functional hypoglycemia and impaired memory performance. The sections of the bidirectional arrows pointing to the down red arrows represent known negative influences on glucose transport, and those pointing to the question marks represent a proposed negative influence that remains to be shown.

resistance, among individuals with elevated cortisol levels, there may also be a 'functional hypoglycemia,' this time at the neuronal level, also contributing to the cognitive deficits.

We propose that although this aggregate 'functional hypoglycemia' may be widespread in the brain, its damaging consequences may be at first more apparent in the hippocampus because of its high vulnerability to damage from hypoglycemia or other noxious influences. This hypothesized chronic low-grade 'functional hypoglycemia' may, in the long run, lead to hippocampal damage and volume loss.

It is also possible that chronic hyperglycemia is toxic to the microvascular endothelial cells, thus impairing transport of glucose across the BBB. Also, chronic hyperglycemia may be toxic to hippocampal cells directly by the increased production of damaging oxidative species resulting from increased metabolism based on increased intracellular glucose. However, this model, proposed by Brownlee to explain the tissue damage resulting from diabetes, does not offer an explanation for the damage associated with milder forms of insulin resistance, when glucose levels are still normal and there is only an elevation in fasting insulin levels. Future research will need to approach the problem more comprehensively, by evaluating multiple systems in the same individuals so as to better determine the relative contribution of each in the phenomenology expressed.

Summary

Americans are living longer, which has contributed to the rising rate of age-associated diseases such as type 2 diabetes and dementia. This, coupled with the epidemic of obesity in industrialized nations, which will further add to an increasing number of T2DM cases within the overall population, makes improving our understanding of how type 2 diabetes affects the aging brain indispensable. In this article, we have reviewed the existing literature on brain aging, diabetes, and the stress regulatory system. There is increasing evidence that dysregulation of the stress and glucoregulatory systems interacts at multiple levels to exert a negative impact on human health in general, and on the aging brain in particular. Although a consensus has not yet gelled within the literature on the exact brain areas affected by the dysregulation of those systems, it is now clear that the MTL is likely the major brain region affected. Within the MTL, the hippocampus, a structure crucially involved in recent memory, appears to be particularly sensitive to these neuroendocrine abnormalities. Since MTL atrophy is also of relevance for MCI and AD, this might explain some of the recently reported associations between dementia and type 2 diabetes. An enhanced understanding of the underlying pathological processes will allow the development of behavioral and pharmacological approaches targeted to protect the integrity of the hippocampus in older individuals. But there is much work to be done and additional progress in this research area is clearly warranted.

See also: Aging of the Brain and Alzheimer's Disease; Aging and Memory in Humans; Chronic (Repeated) Stress: Consequences, Adaptations; Cognition in Aging and Age-Related Disease; Dementia; Hippocampus; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Metabolic Syndrome and Sleep; Stress and Neural Involvement in Metabolism; Stress and Cognition; Stress and Vulnerability to Brain Damage; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Awad N, Gagnon M, and Messier C (2004) The relationship between impaired glucose tolerance, type 2 diabetes, and cognitive function. *Journal of Clinical and Experimental Neuropsychology* 26: 1044–1080.

Biessels GJ, Staekenborg S, Brunner E, Brayne C, and Scheltens P (2006) Risk of dementia in diabetes mellitus: A systematic review. *The Lancet Neurology* 5(1): 64–74.

Brownlee M (2001) Biochemistry and molecular cell biology of diabetic complications. *Nature* 414: 813–820.

Burke SN and Barnes CA (2006) Neural plasticity in the ageing brain. *Nature Reviews* 7: 30–40.

Cervos-Navarro J and Kater SB (1991) Selective vulnerability in brain hypoxia. *Critical Reviews in Neurobiology* 6: 149–182.

Convit A (2005) Links between cognitive impairment in insulin resistance: An explanatory model. *Neurobiology of Aging* 26S: S31–S35.

Convit A, Wolf OT, Tarshish C, and de Leon M (2003) Reduced glucose tolerance is associated with poor memory performance and hippocampal atrophy among normal elderly. *Proceedings of the National Academy of Sciences of the United States of America* 100(4): 2019–2022.

Dallman MF, Pecoraro N, Akana SF, et al. (2003) Chronic stress and obesity: A new view of "comfort food." *Proceedings of the National Academy of Sciences of the United States of America* 100: 11696–11701.

Den Heijer T, Vermeer SE, van Dijk EJ, et al. (2003) Type 2 diabetes and atrophy of medial temporal lobe structures on brain MRI. *Diabetologia* 46: 1604–1610.

Desrocher M and Rovet J (2004) Neurocognitive correlates of type 1 diabetes mellitus in childhood. *Child Neuropsychology* 10: 36–52.

Gold SM, Dziobek I, Sweat V, et al. (2007) Specific hippocampal damage and memory impairments as possible early brain complications of type 2 diabetes. *Diabetologica* 50(4): 711–719.

Lobnig BM, Kromeke O, Optenhostert-Porst C, and Wolf OT (2005) Hippocampal volume and cognitive performance in long-standing type 1 diabetic patients without macrovascular complications. *Diabetic Medicine* 23: 32–39.

McEwen BS (2002) Sex, stress and the hippocampus: Allostasis, allostatic load and the aging process. *Neurobiology of Aging* 23: 921–939.

McEwen BS, Magarinos AM, and Reagan LP (2002) Studies of hormone action in the hippocampal formation: Possible relevance to depression and diabetes. *Journal of Psychosomatic Research* 53: 883–890.

McNay EC, McCarty RC, and Gold PE (2001) Fluctuations in brain glucose concentration during behavioral testing: Dissociations between brain areas and between brain and blood. *Neurobiology of Learning and Memory* 75(3): 325–337.

Pardridge WM, Triguero D, and Farrell CR (1990) Downregulation of blood-brain barrier glucose transporter in experimental diabetes. *Diabetes* 39: 1040–1044.

Roozendaal B, Okuda S, de Quervain DJ, and McGaugh JL (2006) Glucocorticoids interact with emotion-induced noradrenergic activation in influencing different memory functions. *Neuroscience* 138: 901–910.

Rosmond R (2005) Role of stress in the pathogenesis of the metabolic syndrome. *Psychoneuroendocrinology* 30: 1–10.

Tooke JE and Goh KL (1998) Endotheliopathy precedes type 2 diabetes. *Diabetes Care* 21(12): 2047–2049.

Vivian EM (2006) Type 2 diabetes in children and adolescents: The next epidemic? *Current Medical Research and Opinion* 22(2): 297–306.

Wolf OT (2003) HPA axis and memory. *Best Practice and Research: Clinical Endocrinology and Metabolism* 17: 287–299.

Diabetic Neuropathy

H C Powell and A P Mizisin, University of California at San Diego, La Jolla, CA, USA

© 2009 Published by Elsevier Ltd.

Although conventional medical treatment of diabetes mellitus markedly prolongs life, chronic long-term complications involving blood vessels, kidneys, eyes, and nerves affect a large proportion of the more than 10 million people believed to have diabetes in the United States. The frequent occurrence of neurological complications in diabetic patients likely contributed to the erroneous belief held by nineteenth-century physicians that diabetes was a primary disorder of the nervous system. Although peripheral neuropathy remains the major neurological complication of diabetes mellitus, the impact of this endocrine disturbance affects all parts of the nervous system and findings involving the central nervous system have become the focus of increased attention. Mild cognitive impairment and increased susceptibility to Alzheimer's disease are increasingly recognized manifestations of neurological dysfunction in diabetes mellitus. Furthermore, vascular disorders historically linked to kidney, retinal, and nerve complications are now implicated in cerebrovascular disease and stroke etiology in diabetes mellitus. Similar to renal and ocular complications, central and peripheral neural deficits represent major complications of both insulin-dependent (type 1) and noninsulin-dependent (type 2) diabetes mellitus, as well as in diabetes associated with nonalcoholic pancreatic disease, hemochromatosis, and pancreatectomy.

Clinical Overview

Epidemiology

Estimates of the prevalence of neuropathy in the diabetic population are varied and range from 5% to nearly 60%. A lack of agreement as to what constitutes diabetic neuropathy has no doubt contributed to the varied estimates of its prevalence. Indeed, subclinical features of neuropathy can be demonstrated by electrophysiologic tests in nearly every diabetic patient. Both the prevalence and the incidence of neuropathy increase with the duration of diabetes, suggesting a relationship to hyperglycemia or metabolic consequences of hyperglycemia. Information about the geographic distribution of diabetic neuropathy is incomplete and available only from areas where epidemiologic surveys have been conducted, namely North America, Western Europe, Scandinavia, and Nigeria.

Classification

Attempts to classify the types of diabetic neuropathy have been hampered by an incomplete knowledge of pathogenic mechanisms, the diverse neuropathic manifestations observed, and the fact that mixed forms of neuropathy are often encountered. Nevertheless, classifications are clinically useful because individual syndromes occur with sufficient frequency and because classification may help to better predict the natural history and outcome of diabetic neuropathy and, therefore, guide treatment. Although various classifications have been proposed, a simple scheme based on the symmetry of the neuropathy and the type of nerves involved is shown in Table 1.

Somatic and autonomic peripheral nervous system manifestations The most common diabetic neuropathy is a symmetric, predominately sensory, neuropathy that affects the distal regions of the lower limbs first, gradually spreads proximally, and later similarly affects the upper limbs. Symptoms include paresthesias, numbness, and, in some patients, burning and lancinating pain in the lower limbs. Once established, sensory diabetic neuropathy is persistent even if hyperglycemia is controlled.

Acute painful diabetic neuropathy is an infrequent syndrome that often follows precipitous weight loss. Patients with this neuropathy experience unremitting, burning pain in the lower limbs. Cutaneous hyperesthesia that results from contact with clothing or sheets is especially troublesome to these patients. Their symptoms often resolve over the course of a few months, after weight gain accompanying diabetic control.

Severe autonomic neuropathy is rare in diabetic patients, and symptoms usually occur in conjunction with a distal symmetric sensory neuropathy in insulin-dependent diabetes mellitus. A wide variety of symptoms affecting all aspects of autonomic function can be present, including heart rate abnormalities and postural hypotension; distal anhidrosis and gustatory sweating; diabetic diarrhea and gastric, duodenal, and colonic atony; bladder atony and impotence; disturbances of respiratory control; and unawareness of hypoglycemia. Once established, autonomic neuropathy is usually irreversible.

A poorly reversible symmetric lower limb motor neuropathy is sometimes present in older diabetic patients, often accompanied by a distal sensory neuropathy. Proximal lower limb neuropathy can have the natural history of an acute or chronic inflammatory polyradiculoneuropathy. The disorder tends to be

Table 1 Classification of diabetic neuropathy based on the symmetry of the neuropathy and the type of nerves involved

Classification	*Clinical manifestations*
Symmetric	Sensory or sensorimotor polyneuropathy Autonomic neuropathy Proximal lower limb motor neuropathy
Focal or multifocal	Cranial neuropathy Trunk and limb mononeuropathy Asymmetric lower limb neuropathy Mixed

From Llewelyn JG, Tomlinson DR, and Thomas PK (2005) Diabetic neuropathies. In: Dyck PJ and Thomas PK (eds.) *Peripheral Neuropathy*, pp. 1951–1991. Philadelphia: Saunders.

painless and may represent inflammatory neuropathy concurrent with diabetes.

Focal or multifocal neuropathic manifestations Focal and multifocal peripheral nerve lesions appear to occur with increased frequency in the diabetic population. With respect to cranial nerves, painful lesions of the oculomotor nerve are the most frequent. Truncal mononeuropathies that result from lesions of the spinal or thoracic nerves may present as girdle pain or abdominal wall cutaneous hyperesthesia. At sites of entrapment, isolated limb lesions may occur. Asymmetric lower limb neuropathy is a multifocal neuropathy that, unlike proximal lower limb neuropathy, is not usually associated with a distal sensory neuropathy. Although the onset can be rapid and accompanied by proximal pain, most patients recover.

A form of proximal diabetic neuropathy linked to painful myopathy has been referred to as Bruns–Garland syndrome, diabetic amyotrophy, or diabetic lumbosacral radiculoplexus neuropathy. The clinical presentation includes acute onset of asymmetric weakness or paralysis of a lower limb. Symptoms are predominantly motor, but sensory disturbances, including altered nociception, may occur. Pain, usually worse at night, is a key symptom and often the presenting abnormality. The typical syndrome may affect the legs, perineum, and back, but radiculoplexus neuropathy has also been linked to other regions including cervical roots. The inflammatory process appears to be local rather than part of a systemic vasculitis, and in this respect, radiculoplexus neuropathies appear to be specific complications of diabetes. They occur mainly in males with type 2 diabetes mellitus. The mechanism responsible for the inflammatory process has not been identified.

Central nervous system complications The term 'diabetic encephalopathy' has been introduced to describe a syndrome of mild to moderate cognitive deficits, neurophysiological changes, and morphological abnormalities afflicting diabetics, especially elderly patients. Increased risk of stroke and poorer outcomes from cerebrovascular accidents represent additional complications resulting from diabetic angiopathy. Endocrine dysfunction manifesting as either hyper- or hypoglycemia constitutes an additional mechanism for neurological impairment. Radiological imaging studies provide evidence for brain atrophy, which is viewed as part of the phenomenon of accelerated aging in diabetic encephalopathy.

Pathophysiology

Analysis of human nerves and animal models has revealed two distinct phases of diabetic neuropathy. In the early metabolic phase, neuropathy is characterized by biochemical abnormalities and electrophysiological and blood flow deficits without significant morphologic changes. During the later phase, morphologic changes are apparent and chronic neuropathy is associated with degenerative changes of nerve fibers as well as microvascular disease.

Physiological Disturbances

Electrophysiological disturbances are extremely common in diabetic patients and are almost universal in animal models of diabetes mellitus. The earliest changes involve sensory nerve conduction and include a reduction in velocity and a decrease in amplitude as well as an increase in the temporal dispersion of compound action potentials. Motor nerve conduction is similarly affected. Nerve conduction abnormalities are often present in patients lacking symptoms of neuropathy; however, there is a greater reduction in conduction velocity in those with neuropathy. The severity of nerve conduction abnormalities increases with the duration of diabetes and is greater in newly diagnosed patients with poor control than in those with well-controlled disease. Improvement of conduction abnormalities in newly diagnosed patients has been observed following glycemic control. Improvements have also been reported following treatment with aldose reductase inhibitors (ARIs).

Diabetic nerve from patients or experimental animals possesses an ability to function during ischemia. Thus, diabetic nerve retains motor nerve conduction velocity, vibration, and tactile senses, and appreciation of pain for longer than normal nerve. Resistance to ischemic conduction failure reduces the readiness with which the person's limbs 'go to sleep' when kneeling or crossing the legs and appears to be independent of the presence of clinical neuropathy or subclinical electrophysiological changes. Resistance to ischemic conduction failure is not specific to

diabetes mellitus but is also observed in a number of other disorders, including uremia, motor neuron diseases, and chronic liver disease.

Paradoxically, although diabetic nerves are resistant to ischemic conduction failure, experimental studies have established susceptibility to nerve blood flow deficits and endoneurial hypoxia. Endoneurial hypoxia has also been reported in the sural nerves of diabetic patients and, interestingly, the hypoxia was greater in the nerves of those patients with evidence of clinical neuropathy.

Morphological Changes

The structural basis for diabetic neuropathy in humans and in animal models involves injury to all cellular elements of the nerve.

Myelinated fibers With respect to myelinated nerve fibers, the earliest structural change manifests as a diminution of the caliber of axons (axonal dwindling) that is associated with the early metabolic phase of diabetic neuropathy and occurs without other signs of fiber degeneration or loss. Axonal dwindling is thought by some to be the result of maturational arrest induced by the diabetic state. However, others have demonstrated that dwindling occurs in diabetic animals in which skeletal growth is similar to that of controls, and that it is associated with the loss of neurofilaments, the main cytoskeletal determinant of axonal caliber.

With chronic diabetic neuropathy, dystrophic and degenerative changes of axons occur (Figure 1). Dystrophic changes involve massive swelling due to accumulation of swollen organelles, microtubules, neurofilaments, dense bodies, and vesicles. Glycogen accumulation in the outer compartment of axonal mitochondria causes these organelles to swell and ultimately disintegrate. Degenerative material in axons is sometimes seen engulfed by adaxonal processes of the Schwann cell. These dystrophic changes are associated with overall loss of myelinated and unmyelinated fibers in the peripheral nerve trunks that is usually more marked distally but can involve the spinal roots. Axonal loss has been documented in symmetric, autonomic, and focal and multifocal diabetic neuropathies. Regenerative clusters of axonal sprouts provide evidence of regeneration that occurs coincident with axonal degeneration. Increased regeneration has been reported following treatment with an ARI in at least one clinical trial.

Schwann cell injury, including accumulation of lipid droplets, Pi granules of Reich (lysosomal inclusions), intermediate filament accumulation, and enlarged mitochondria, is also evident and can progress to complete dissolution of the Schwann cell with consequent demyelination. Proliferative Schwann cell changes, such as 'onion bulb' formation, are a feature of diabetic neuropathy and are indicative of repeated episodes of demyelination and remyelination (Figure 2). Interestingly, increases in the incidence of remyelination in sural nerve biopsies from diabetic patients with neuropathy have been claimed in a clinical trial following ARI treatment.

Morphologists have not resolved whether segmental demyelination or axonal degeneration is the primary lesion in diabetic neuropathy. Demyelination

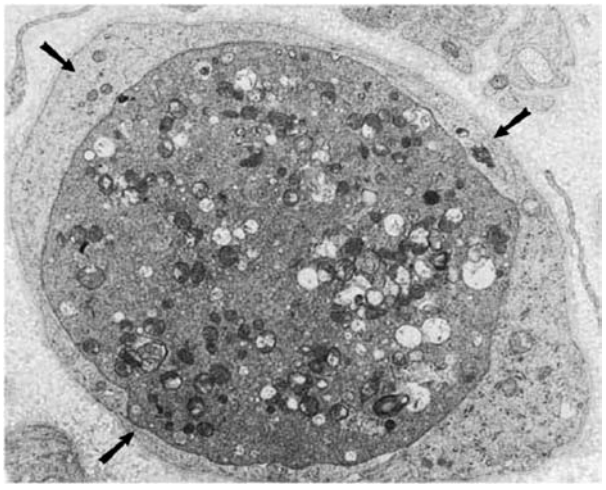

Figure 1 Chronic human diabetic neuropathy: Axonopathy. Electron micrograph of nerve shows axonal dystrophy. The swollen axon is filled with cytoplasmic debris, electron-dense bodies, organelles, and filaments. Although surrounded by Schwann cell cytoplasm (arrows), the fiber is completely demyelinated.

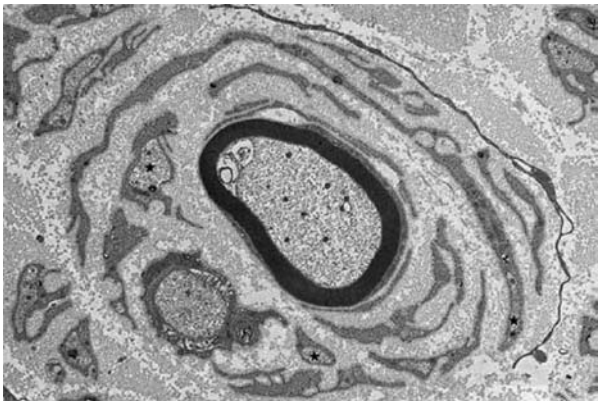

Figure 2 Chronic human diabetic neuropathy: Schwann cell injury. Electron micrograph shows a myelinated nerve fiber surrounded by supernumerary Schwann cell processes ('onion bulb' formations). The abnormal proliferative changes emphasize the response of the Schwann cell to a chronic metabolic insult. Sprouting axons (stars) appear within some of the proliferating Schwann cells.

may be primary (i.e., resulting from a direct effect of diabetes on myelin or Schwann cell function) or secondary to axonal degeneration. Moreover, studies of other metabolic and toxic neuropathies have shown that either condition, when long established, can induce the other as a secondary change. Nevertheless, evidence is accumulating that both primary demyelination and primary axonopathy can occur independently in this complex neuropathy.

Unmyelinated fibers Early electron microscopic studies noted a distinct loss of unmyelinated fibers. Characteristic degenerative changes of these fibers include shrinkage of axons, accumulation of enlarged vesicular elements, and deterioration of tubular and filamentous elements of the cytoskeleton. Edematous Schwann cell cytoplasm has also been observed as well as hyperplasia of surrounding basal lamina. Complete degeneration results in empty or denervated Schwann cell subunits surrounded by a basal lamina. Loss of distal unmyelinated axons of the epidermal innervation has been documented using immunostaining techniques. Using similar techniques, comparable loss of intraepidermal unmyelinated fibers has also been documented in streptozotocin diabetic rats and mice. These methods are attractive not only because the biopsy is minimally invasive but also because they permit quantitative assessment of neuropathic changes and the regenerative response that may ensue.

Blood vessels A distinctive morphological change in human diabetic neuropathy is microangiopathy of the intrinsic vasa nervorum, the endoneurial blood supply of peripheral nerves. The severity of microangiopathy increases with the duration of diabetes, and reductions in endoneurial capillary density parallel reductions in nerve fiber density. Electron microscopic studies of human biopsies reveal that the marked thickening and hyalinization of capillary walls evident by light microscopy are the result of endothelial cell and pericyte basal lamina reduplication in conjunction with endothelial cell proliferation (Figure 3). Vessel wall thickening is associated with a reduction in the luminal area of endoneurial capillaries and, in one study, an increase in the number of 'closed' capillaries.

Inflammatory changes of the extrinsic vasa nervorum, ranging from epineurial vasculitis to microvasculitis, have been described in diabetic lumbosacral radiculoplexus neuropathy. Epineurial vasculitis occurs less frequently and involves larger arteries, arterioles, veins, and venules, which may have fibrinoid necrosis of their walls. More frequently, inflammatory changes manifest as microvasculitis, which involves mild perivascular and intramural infiltrates in the smaller epineurial vasculature.

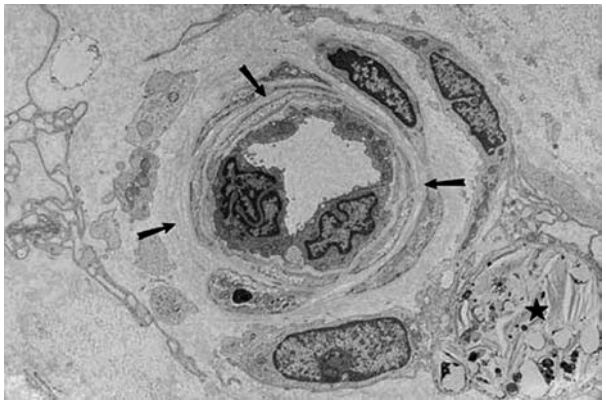

Figure 3 Chronic human diabetic neuropathy: Microangiopathy. Electron micrograph of nerve shows an endoneurial capillary with mural thickening due to supernumerary basal laminae (arrows) and endothelial proliferation. A macrophage (star) filled with lamellar and crystalline debris is also visible.

Structural changes of cerebral blood vessels are part of the normal aging process, with loss of mural flexibility as the vessel walls acquire more collagen and basal laminae also thicken. Diabetic vessels appear less reactive to physiological stimuli since endothelial responses to nitric oxide-mediated dilation of cerebral vessels in streptozotocin diabetic rats is reportedly impaired.

Other changes With respect to morphological changes in other components of the nerve fascicle, thickening of the perineurial basal laminae, as well as calcification localized predominately to the outer lamellae of the perineurium, has been reported. Increases in endoneurial area have been documented in nerve biopsies from patients with diabetic neuropathy. Fibrosis is extensive, with collagen being found in the interstitium as well as within the original Schwann cell–basal lamina tube of degenerated nerve fibers. The persistence of Schwann cell basal lamina in circular profiles may reflect increased rigidity induced by nonenzymatic glycation.

Evidence for endoneurial edema had been found only in biopsy specimens until the use of magnetic resonance spectroscopy to examine the water content of the sural nerves of diabetic patients. Increased amounts of water were found in the sural nerves of patients with symptoms of diabetic neuropathy who had not received treatment but not in those treated with one of two structurally diverse ARIs. Mast cell proliferation and degranulation have been reported in animal models of diabetes as well as in diabetic patients. Mast cell granules, which contain biogenic amines capable of increasing vascular permeability and degranulation, could be either a cause or a consequence of endoneurial edema in diabetes.

Brain and spinal cord In the central nervous system, both radiological and morphological studies emphasize degenerative changes characterized by atrophy. Loss of cerebral cortical tissue accompanied by enlargement of lateral ventricles in aging diabetic patients support the idea of accelerated aging in type II diabetes. White matter changes also appear and may be related to vascular disease. Macrovascular disease involving atherosclerosis, hypertension, and impaired recovery from stroke are accompanied by microangiopathic changes in the brain and spinal cord.

Pathogenesis

It is unlikely that a single pathogenic mechanism will serve to explain the diverse clinical manifestations of diabetic neuropathy because multiple factors probably interact to produce this complex neuropathy. Current therapeutic approaches rarely normalize carbohydrate homeostasis, and both insulin-dependent and noninsulin-dependent diabetics develop neuropathy. Therefore, attention has focused on the metabolic and vascular consequences of hyperglycemia as a potential instigator of neuropathy.

Metabolic Consequences of Hyperglycemia

In diabetic complications, mechanisms of hyperglycemia-induced damage include polyol pathway flux, advanced glycation end product (AGE) formation, activation of protein kinase C (PKC), and hexosamine pathway activity. During hyperglycemia, reduction of intracellular glucose by aldose reductase results in the formation of sorbitol and concomitant oxidation of nicotinamide adenine dinucleotide phosphate (NADPH). The sorbitol formed is subsequently oxidized to fructose by sorbitol dehydrogenase with the reduction of NAD^+. Proposed detrimental effects of increased polyol pathway flux include sorbitol-induced osmotic stress, decreased Na^+-K^+-ATPase activity, increased cytosolic $NADH/NAD^+$ and consequent inhibition of glyceraldehydes-3-phosphate dehydrogenase (GAPDH), and decreased cytosolic NADPH. In nerve, aldose reductase has been localized to myelinated fiber Schwann cell cytoplasm. Extraneural sites of aldose reductase localization include skeletal muscle and endothelial cells of blood vessels. Although an unequivocal role for polyol flux in the development of diabetic neuropathy has not been established, accumulation of sorbitol and changes in cofactor levels may compromise Schwann cell metabolism, initiate reduction of Schwann cell-derived neurotrophic factors, and ultimately promote the destruction of this cell. Other consequences include nerve conduction defects, increased triose phosphate concentrations, and decreased reduced glutathione levels.

The primary initiating event in the formation of intracellular and extracellular AGEs is thought to be the generation of glucose-derived carbonyl precursors by auto-oxidation of glucose to glyoxyl, decomposition of Amadori product to 3-deoxyglucosone, and fragmentation of triose phosphates to methylglyoxal. Potential detrimental effects of AGE production are wide ranging. Proteins modified by AGE formation have altered function and structure, with negative consequences for enzyme function, and interaction of extracellular matrix ligands and receptors. Activation of AGE receptors on endothelial cells and macrophages also induces receptor-mediated production of reactive oxygen species.

PKC is a family of 11 isoforms, the majority of which are activated by the triose phosphate lipid second messenger diacylglycerol (DAG). In diabetes, the β and δ isoforms of PKC are activated by hyperglycemia-induced increases in DAG, which are thought to be the consequence of the cytosolic $NADH/NAD^+$ imbalance and consequent inhibition of GAPDH that results in elevated triose phosphate levels. Many of the proposed effects of PKC activation in diabetes are vascular and include blood flow abnormalities resulting from altered expression of endothelial nitric oxide synthase and endothelin-1; changes in vascular permeability and angiogenesis mediated by enhanced vascular endothelial growth factor expression; capillary and vascular occlusion resulting from increased collagen and fibronectin expression; and decreased fibrinolysis due to altered expression of transforming growth factor (TGF)-β and plasminogen activator inhibitor-1, respectively.

Aside from increases in triose phosphates, another consequence of hyperglycemia and hyperglycemia-induced inhibition of GAPDH is increased activity of the hexosamine pathway. Shunting excess intracellular glucose into the hexosamine pathway results in conversion of fructose-6-phosphate diverted from glycolysis into glucosamine-6-phosphate via the rate-limiting enzyme, glutamine:fructose-6-phosphate amidotransferase. The glucosamine-6-phosphate is subsequently converted into UDP-*N*-acetylglucosamine, which is believed to modulate transcription factors responsible for the production of TGF-β and plasminogen activator inhibitor-1 implicated in the vascular abnormalities alluded to previously.

Specific inhibitors of polyol pathway activity, AGE formation, PKC activation, and hexosamine flux have been shown to ameliorate early hyperglycemia-induced biochemical and functional nerve disorders, emphasizing the potential role of these mechanisms in the pathogenesis of diabetic neuropathy. Studies have

implicated oxidative stress as an element common to these mechanisms and emphasized their interrelationship. Others have suggested that oxidative stress is the common element linking these mechanisms of hyperglycemia-induced damage, with each of the four mechanisms reflecting overproduction of superoxide by mitochondrial electron-transport chains. However, despite these contentions, there is little direct evidence supporting an oxidative stress-mediated link between early metabolic events and subsequent structural injury characteristic of diabetic neuropathy. Such evidence is critical for evaluating the overall importance of oxidative stress in the pathogenesis of diabetic neuropathy.

Vascular Consequences

Reduced nerve blood flow and an associated endoneurial hypoxia have been reported in both acute and chronic experimental diabetes and in diabetic patients. In diabetic animals, improvements in electrophysiological deficits occur with any of a variety of agents that improve nerve blood flow. These agents include vasodilators such as α_1-adrenoreceptor blocking agents, calcium channel blockers, drugs that inhibit the renin–angiotensin system, and vasomodulator prostanoids. Other treatments used in experimental animal models of diabetes that can act via vascular-related effects to improve nerve conduction include omega-6 essential fatty acids, antioxidants, aminoguanidine, and ARIs. The fact that so many agents can alter blood flow in both animal models of diabetes and patients with diabetes suggests that dysfunction of blood vessel endothelial cells underlies nerve blood flow deficits. Moreover, the efficacy of ARIs and aminoguanidine in preventing nerve blood flow and conduction deficits argues for a role for the polyol pathway and nonenzymatic glycation in the vascular consequences of endothelial cell dysfunction and the etiology of diabetic neuropathy, as indicated previously.

See also: Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Insulin-Like Growth Factor Signaling and Actions in Brain; Neuroendocrine Aging: Pituitary Metabolism; Neuropathy: Peripheral.

Further Reading

Akkina SK, Patterson CL, and Wright DE (2001) GDNF rescues nonpeptidergic unmyelinated primary afferents in streptozotocin-treated diabetic mice. *Experimental Neurology* 167: 173–182.

Behse F, Buchthal F, and Carlsen F (1977) Nerve biopsy and conduction studies in diabetic neuropathy. *Journal of Neurology, Neurosurgery & Psychiatry* 40: 1072–1082.

Bianchi R, Buyukakilli B, Brines M, et al. (2004) Erythropoietin both protects from and reverses experimental diabetic neuropathy. *Proceedings of the National Academy of Sciences of the United States of America* 101: 823–828.

Biessels GJ, van der Heide LP, Kamal A, Bleys RL, and Gispen WH (2002) Ageing and diabetes: Implications for brain function. Review. *European Journal of Pharmacology* 441: 1–14.

Bischoff A (1980) Morphology of diabetic neuropathy. *Hormone Metabolism Research Supplement* 9: 18–28.

Brownlee M (2001) Biochemistry and molecular cell biology of diabetic complications. *Nature* 414: 813–820.

Dyck PJB, Norell JE, and Dyck PJ (1999) Microvasculitis and ischemia in diabetic lumbosacral radiculoplexus neuropathy. *Neurology* 53: 2113–2121.

Fagerberg S-E (1959) Diabetic neuropathy: A clinical and histological study on the significance of vascular affections. *Acta Medica Scandinavica* 164: 1–97.

Kennedy WR, Wendelschafer-Crabb G, and Johnson T (1996) Quantitation of epidermal nerves in diabetic neuropathy. *Neurology* 47: 1042–1048.

Llewelyn JG, Tomlinson DR, and Thomas PK (2005) Diabetic neuropathies. In: Dyck PJ and Thomas PK (eds.) *Peripheral Neuropathy*, pp. 1951–1991. Philadelphia: Saunders.

Mizisin AP and Powell HC (2003) Histopathology. Diabetic peripheral neuropathy. In: Gries FA, Cameron NE, Low PA, and Ziegler D (eds.) *Textbook of Diabetic Neuropathy*, pp. 83–87. Stuttgart: Thieme.

Obrosova IG (2002) Glucose-generated stress in peripheral nerve. *International Review of Neurobiology* 50: 3–35.

Pirart J (1978) Diabetes mellitus and its degenerative complications: A prospective study of 4400 patients observed between 1947 and 1973. *Diabetes Care* 1: 168–188.

Polydefkis M, Hauer P, Sheth S, et al. (2004) The time course of epidermal nerve fiber regeneration: Studies with normal controls and in people with diabetes with and without neuropathy. *Brain* 127: 1606–1615.

Thomas PK and Lascelles RG (1966) The pathology of diabetic neuropathy. *Quarterly Journal of Medicine* 35: 489–509.

Relevant Websites

http://www.diabetes.org – American Diabetes Association.
http://pns.ucsd.edu – Peripheral Nerve Society.
http://www.neuropathy.org – The Neuropathy Association.

Dichotic Listening Studies of Brain Asymmetry

K Hugdahl, University of Bergen, Bergen, Norway

© 2009 Elsevier Ltd. All rights reserved.

Dichotic Stimuli

The stimuli typically used in dichotic listening (DL) studies consist of presentations of pairwise combinations of consonant–vowel (CV) syllables that are made up of the six stop-consonants /b/, /d/, /g/, /p/, /t/, and /k/, and the vowel /a/.Thus, examples of DL stimulus pairs are /ba/–/pa/, /ga/–/pa/, and so on. Another often-used variant of the DL test is the so-called fused rhymed words test. These stimuli are computer synthesized and consist of consonant–vowel-consonant (CVC) pairs. The stimuli are synthesized so that the nondistinctive components of each pair are identical (e.g., /aba/–/aka/). The temporal and auditory spectral overlap between numbers of each pair in this test is so great that they fuse into a single auditory percept.

The Dichotic Test Situation

Preparation of dichotic stimulus materials requires computer editing on each trial of the two CV syllables. The most important aspect of the editing procedure is the synchronization of the onset of the energy release in the two syllables in a pair. The syllables are synchronized on both the consonant and vowel segments, which usually requires access to speech synthesis analysis capability. Figure 1 shows a computer display of the syllables /ba/ (top) and /pa/ (bottom) that are synchronized on both the consonant and vowel onset.

DL Paradigms

The most commonly used paradigms in DL laterality studies are the free-report paradigm, the forced-attention paradigm, the fused-rhymed paradigm, and the target-monitoring paradigm. In the free-report paradigm, individuals are required to report the items that they heard on each trial as accurately as possible. Only those trials in which a person can identify one item, but not the other, correctly reveal information about brain laterality. Thus, as a general rule, only single-correct trials should be analyzed. An alternative is to instruct the person to answer only for one item on each trial. The experimenter then calculates number of correct reports from the right and left ears, separately. The free-report paradigm requires that the person has intact speech and can give an oral answer. Thus, the free-report paradigm is less well suited for the study of laterality in clinical populations with speech dysfunctions.

In the target-monitoring paradigm, some of the items in the list to be presented are selected as 'targets,' and the person is instructed to indicate (manually or orally) whenever he/she detects a target. Usually CVC syllables or words are used in the target-monitoring paradigm. The target-monitoring paradigm has the advantage that it can be used also with persons who lack expressive speech.

The fused-rhymed paradigm is similar to the free-report paradigm except for the important distinction that the two sounds presented at the ears fuse into a coherent perceptual unit. A characteristic of the fused-rhymed paradigm is that although individuals subjectively report that they hear only one sound, they report this to be the right-ear item.

The forced-attention paradigm allows for the study of attentional influences on brain laterality. The paradigm involves instructing the individual to pay attention only to the right ear, and report only from that ear in half of the trials, and to pay attention to and report only from the left ear in the other half of the trials. Often there is a third condition included whereby the person is not given any specific instructions. In that case, each instructional condition then applies to a third of the trials.

The Right-Ear Advantage

Irrespective of whether nonfused CV syllables or fused CVC syllables are used, the typical outcome in a standard DL test is a greater percentage of correct reports from the right ear, as compared to the left ear. This is called a right-ear advantage (REA) and is a robust empirical finding in both right- and left-handed persons. The REA is most easily seen in response to the consonant, and is difficult to observe in response to the vowel in a CV syllable. It has thus been argued that the REA might reflect hemisphere specialization for rapidly changing auditory stimuli, such as the rapid formant transition seen in the stop consonants. Figure 2 shows the distributions for correctly reported items from the right-ear (red) and left-ear (blue) stimuli from 694 right-handed adults. Note the apparent shift to the right for the right-ear stimulus distribution.

Attentional Factors

DL also indicates dynamic laterality factors, such as attention, which may modulate a structurally based

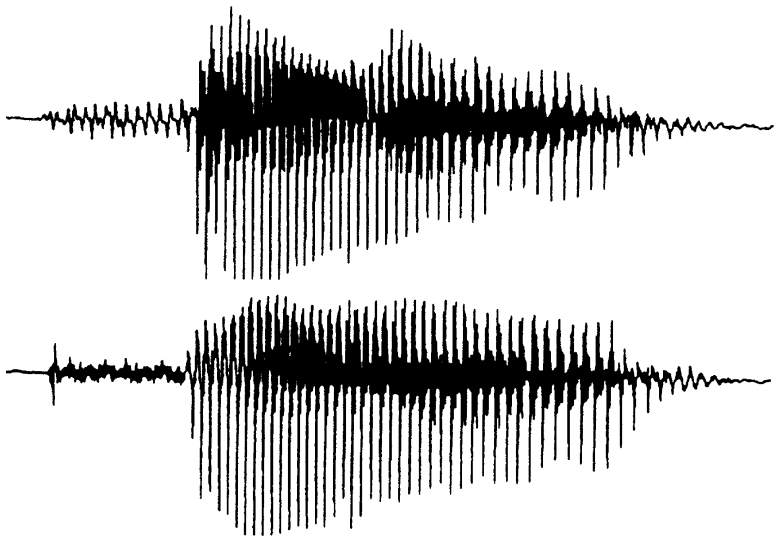

Figure 1 Computer display of the CV syllables /ba/ (top) and /pa/ (bottom) synchronized at the energy release at both consonant and vowel onset segments.

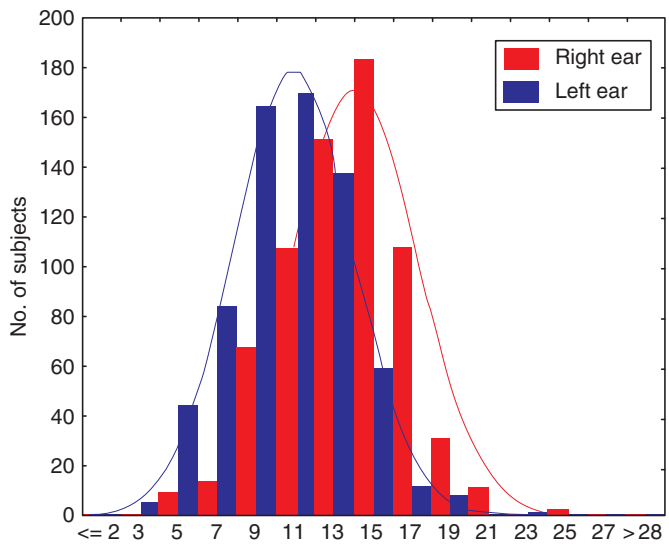

Figure 2 Number of adult study participants ('subjects'; *y*-axis) plotted against correctly reported CV syllables presented in the right ear (red distribution) and left ear (blue distribution), respectively. Data from 694 right-handed adults.

laterality when the individual shifts attention to the right or left side in auditory space. In the typical DL test, one-third of the trials involve a 'nonforced' attentional condition, where individuals are not given any particular instructions regarding deployment of attention. In one-third of the trials the person is instructed to attend to and report from the right ear (forced-right attention), and in one-third of the trials the person is instructed to attend to and report from the left ear (forced-left attention). The REA is increased during the forced-right attention condition, and a left-ear advantage (LEA) is observed during the forced-left attention condition. Thus, attentional factors can sometimes override the basic REA asymmetry, providing evidence for a 'top-down' instruction-driven modulation of a 'bottom-up' stimulus-driven laterality effect. Studies in our laboratory have, however, shown that top-down modulation is reduced in children, particularly before literacy, and in psychiatric patients. The switching of an REA to an LEA during forced-left attention is also dramatically reduced in patients with right-sided brain damage.

Neuroanatomical Basis

The REA is believed to be caused by the fact that although auditory input is transmitted to both auditory cortices in the temporal lobes, the contralateral projections are stronger and more preponderant, interfering with the ipsilateral projections. The advantage for the contralateral auditory projections means that the language-dominant left hemisphere receives a stronger signal from the right ear. The contralateral signal from the left ear to the right hemisphere must first pass the corpus callosum in other to be processed in the left hemisphere. Following the same logic, an LEA indicates the right hemisphere to be language dominant, and a no-ear advantage (NEA) indicates a bilateral language dominance.

In an O^{15}-positron emission tomography (PET) study in our laboratory using a DL target monitoring paradigm, we compared brain activation to CV syllables and musical stimuli that were presented dichotically. The CV syllables and musical stimuli showed opposite activation asymmetries in the temporal lobe. Greater activation was observed in the left superior temporal gyrus to the CV syllable stimuli, while greater activation was observed in corresponding right temporal lobe areas to the musical stimuli. Interestingly, overall activation and the intensity of activation were greater in response to the CV syllable stimuli compared with the musical stimuli. This may indicate specialized cortical networks in the left hemisphere for the processing of linguistic stimuli, also including the planum temporale area. The PET data are shown in Figure 3.

Validity and Reliability

The REA to dichotic presentations of CV syllables is usually about 20% in order of magnitude for oral reports, with about 50–70% correct reports from the right ear and 30–50% correct reports from the left ear. The REA is observed in about 85–90% of right-handed persons, and in about 65% of left-handed persons. The magnitude of the REA may, however, vary considerably among individuals within a given ear advantage direction. Scatterplots of DL correct reports from the right and left ears for more than 1000 persons, both children and adults, are seen in Figure 4.

The 85% of right-handed subjects to be identified as left-hemisphere-language dominant by the DL technique is about 10% lower than what is obtained

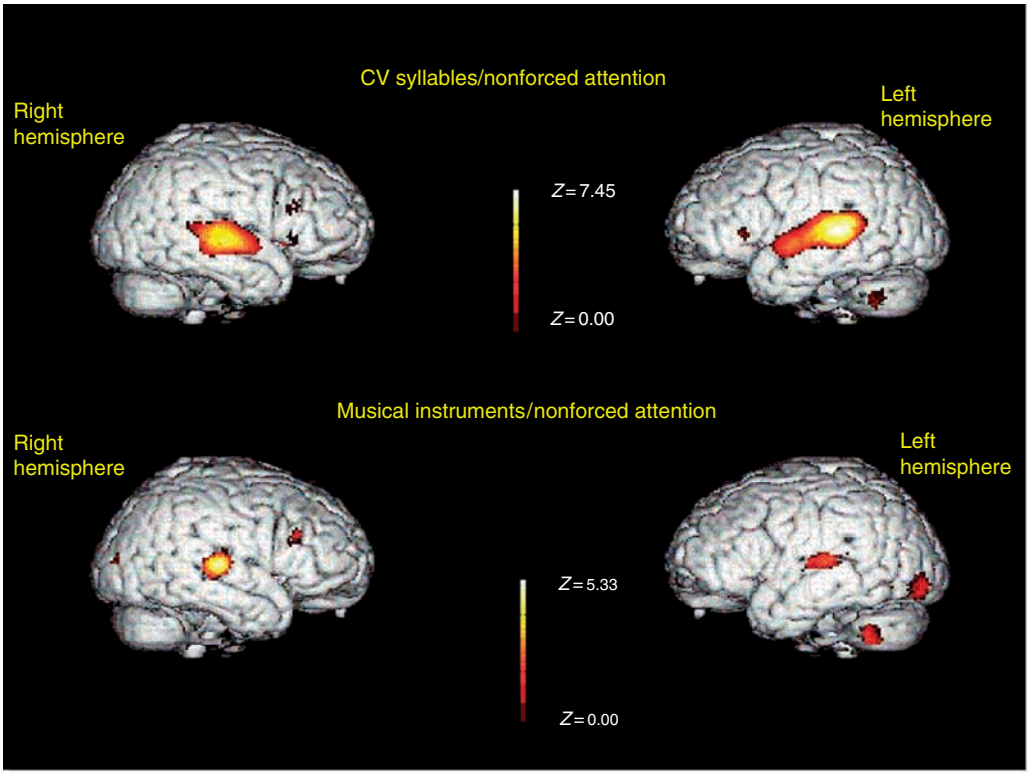

Figure 3 ^{15}O-PET brain activation data in response to CV syllables and musical stimuli. Note the leftward asymmetry for the CV syllable stimuli, and rightward asymmetry for the musical stimuli. Reproduced from Hugdahl K, Brønnick K, Law I, et al. (1999) Brain activation during dichotic presentations of consonant–vowel and musical instruments stimuli: A ^{15}O-PET study. *Neuropsychologia* 37: 431–440, with permission from Elsevier.

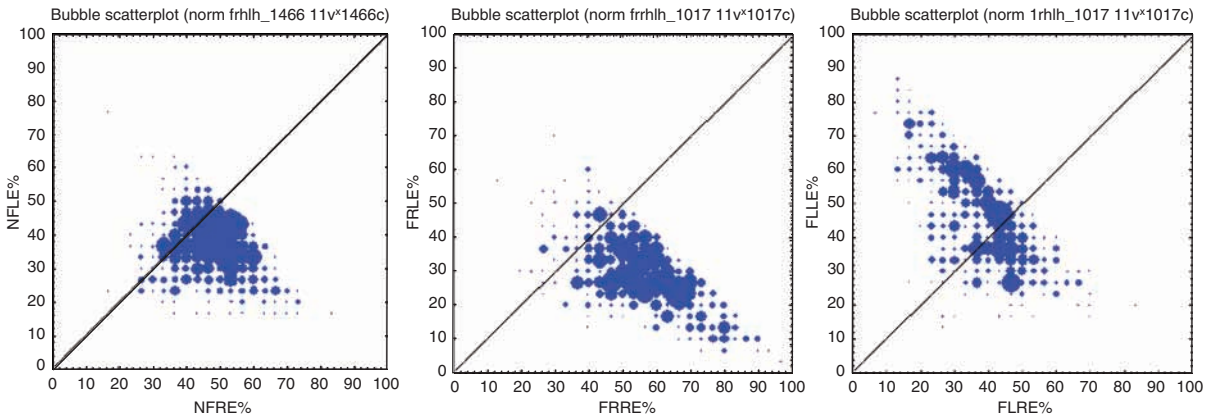

Figure 4 Scatterplots of left- and right-ear correct reports in normal right- and left-handed persons. RE%, % correct right ear reports; LE%, % correct left ear reports, NF, nonforced attention; FR, attention forced to right-ear stimulus; FL, attention forced to left-ear stimulus.

when invasive techniques, like the Wada test, are used. The Wada test involves selective injection of a barbiturate into the right or left hemisphere, 'silencing' the injected hemisphere for about 10 min. The 10% discrepancy between DL and the Wada test probably reflects that DL, in addition to indicating cortical function, also reflects subcortical thalamic influences on language lateralization. DL also picks up perceptual language laterality in addition to expressive speech laterality, whereas the Wada test is an exclusive test of speech laterality. Measures of test–retest reliability vary between 0.70 and 0.90 across different studies. The variation across studies may stem from differences in the reliability index used. However, a study from our laboratory showed that of 17 epileptic patients undergoing both the Wada test for language dominance and DL, 94% of the cases were correctly classified with the DL technique from a stepwise discriminant analysis. The DL results are shown in Figure 5, together with an indication of left- (blue dots) or right- (red dots) hemisphere speech according to the Wada test.

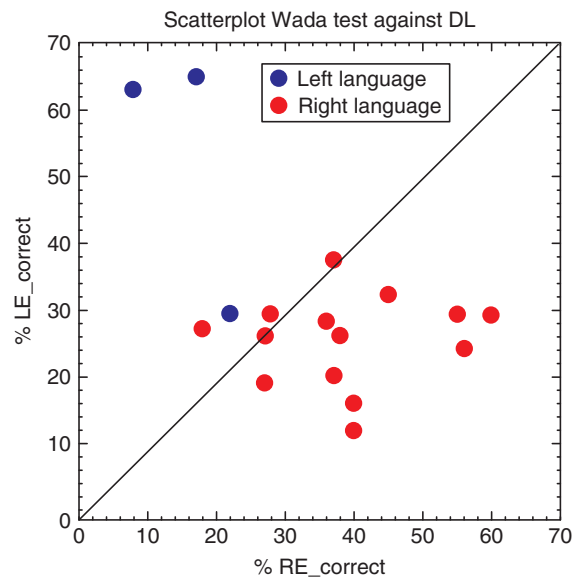

Figure 5 Wada test validation of DL scores from 17 adolescent epileptic patients. Overall, 94% of the group was correctly classified. Data courtesy of Göran Carlsson and Paul Uvebrant, Kiel, Germany, and Göteborg, Sweden.

Calculation of DL Scores

The ear advantage in DL may be calculated in different ways. The most straightforward and simple method is to calculate the number of correct reports from each ear. Since this measure does not compensate for differences in overall performance (i.e., total correct reports) among individuals, some authors favor an index score, where right-ear minus left-ear scores are divided by the sum total of left-ear plus right-ear scores. Still another way of handling DL scores is to display the number of individuals that show a particular ear advantage in scatterplots, as shown in Figure 4.

Arousal and Activation

DL scores may be affected by factors other than brain laterality, and it has been shown that individual differences in arousal or activation may attenuate the REA. Specifically, experimental manipulations whereby the persons are made to believe that they will receive an electric shock to their hand whenever they make an error have been shown to drastically attenuate the REA. This is probably caused by an increase in both brain stem reticular and right-hemisphere activation (right hemisphere being

particularly sensitive for emotional events), which interferes with left-hemisphere processing of the contralateral right-ear signal.

Studies with the forced-attention paradigm have generally revealed that the REA is not an attentional artifact, with right-handed persons attending to the right side in space, which causes the right-ear signal to be better perceived and reported. Experimental manipulations whereby the person is explicitly instructed to turn his/her head and eyes toward the right or left side in space do not seriously affect the REA.

Developmental Effects

The REA to CV syllables can be observed in children as young as 3 years of age, and is frequently reported in 5-year-old children. The magnitude of the REA does not change with age, but stays rather constant. However, the overall level of performance increases with increasing age. Furthermore, preliterate children also report more items from the right ear during testing following instructions to focus attention to the left ear, as in the forced-attention paradigm. The constancy of the REA across different age levels is taken as an indication that brain laterality is not subject to ontogenetic development, and that it is present already at the infant stage.

Sex Differences and Handedness Effects

Males and females differ regarding several cognitive processes (language, visuospatial skills) that are differentially mediated by the left and right hemispheres. The sexes also differ in disorders related to hemisphere differences in brain function. Various disorders such as dyslexia, hyperactivity, and stuttering occur more frequently among boys than girls. Aphasia is also more commonly seen in males than in females after unilateral left-sided lesions. All this evidence points to the hypothesis that females are either less lateralized than males for certain cognitive functions, or that they have a more diffuse cortical organization for cognitive function.

Similar patterns of responding have been obtained in DL tests, with a greater proportion of males than females showing a REA to CV syllables. Males as a group also show a greater mean REA in the standard free-report DL paradigm. It is known that there are differences in cortical organization between right- and left-handers. Only about 65–70% of left-handers are left-hemisphere dominant for language, compared to 95–99% of right-handers. It then follows that right- and left-handers should also differ in their DL performance to CV syllables. Although not all studies have found significant group-mean differences in ear advantage magnitudes between handedness groups, almost every study has reported a tendency for attenuation of the REA in left-handers compared to right-handers. Usually the difference in the magnitude of the REA between right- and left-handers is in the order of 10–15%. The higher proportion of right-handers showing an REA also seem to be independent of the testing procedure or the kind of paradigm used.

Clinical Populations

Commissurotomized patients with sectioning of the corpus callosum show almost complete suppression of the left-ear signal. This is termed 'paradoxical left-ear extinction' and is explained with reference to Kimura's model of suppression of the ipsilateral auditory pathways under dichotic competition. Thus, in a patient with a split brain, the right-ear signal reaches the left temporal cortex unopposed by the ipsilateral right-ear signal. However, the left-ear signal reaches the right hemisphere, and because of the sectioning of the corpus callosum, it cannot be relayed to the left hemisphere for processing.

DL has been used to assess anomalous brain laterality patterns in, for example, stuttering and dyslexia. Although several studies have reported reduced REA magnitudes in these groups, the evidence is not convincing and more research is needed. The DL technique has also been used in studies of schizophrenia and in affective disorders, under the assumption that these disorders are related to left- and right-hemisphere dysfunctions, respectively. The DL data support a notion that schizophrenia may be associated both with left-hemisphere overactivation and with an interhemispheric transfer deficit, although the findings to some extent are controversial.

Nonverbal Stimuli and Lateralization of Affect

Studies of lateralization of affect have involved both the expression of an emotion, as in facial emotional expressions, and the perception of affect, as in the presentation of an emotional stimulus. To summarize, most studies have revealed a right-hemisphere superiority for both expression and perception of affective, or emotional, events. One question that still arouses controversy in the literature is whether the right-hemisphere superiority for emotional processing is valid for both positive and negative emotions.

DL studies of affect have employed either verbal stimuli differing in prosody (vocal intonation), or nonspeech sounds, such as crying or laughing. DL

studies with nonspeech sounds have in general revealed a left-ear advantage in most persons, supporting right-hemisphere processing superiority for these kinds of stimuli. DL studies of differences in prosody to speech sounds have also revealed a larger proportion of persons with an LEA than an REA. However, it is at present unclear whether this is specifically tied to the prosodic component in speech, or whether it is related to an overall right-hemisphere specialization for emotional processing.

See also: Brain Asymmetry: Evolution; Cognitive Neuroscience: An Overview; Hemispheric Specialization and Cognition; Neuropsychological Testing; Psychophysics of Attention; Split-Brain Patients.

Further Reading

Bruder GE (1988) Dichotic listening in psychiatric patients. In: Hugdahl K (ed.) *Handbook of Dichotic Listening: Theory, Methods, and Research*, pp. 527–564. Chichester, UK: Wiley.

Bryden MP (1988) An overview of the dichotic listening procedure and its relation to cerebral organization. In: Hugdahl K (ed.) *Handbook of Dichotic Listening: Theory, Methods, and Research*, pp. 1–44. Chichester, UK: Wiley.

Clark CR, Geffen LB, and Geffen G (1988) Invariant properties of auditory perceptual asymmetry assessed by dichotic listening. In: Hugdahl K (ed.) *Handbook of Dichotic Listening: Theory, Methods, and Research*, pp. 71–83. Chichester, UK: Wiley.

Green MF, Hugdahl K, and Mitchell S (1994) Dichotic listening during auditory hallucinations in schizophrenia. *American Journal of Psychiatry* 151: 357–362.

Hugdahl K (1995) Dichotic listening: Probing temporal lobe functional integrity. In: Davidson RJ and Hugdahl K (eds.) *Brain Asymmetry*, pp. 123–156. Cambridge, MA: MIT Press.

Hugdahl K (ed.) (1988) *Handbook of Dichotic Listening: Theory, Methods, and Research*. Chichester, UK: Wiley.

Hugdahl K and Andersson L (1986) The 'forced-attention paradigm' in dichotic listening to CV-syllables: A comparison between adults and children. *Cortex* 22: 417–432.

Hugdahl K, Brønnick K, Law I, et al. (1999) Brain activation during dichotic presentations of consonant–vowel and musical instruments stimuli: A ^{15}O-PET study. *Neuropsychologia* 37: 431–440.

Hugdahl K, Carlsson G, Uvebrant P, et al. (1997) Dichotic listening performance, intracarotid amobarbital injections in children/adolescent: Comparisons pre- and post-operatively. *Archives of Neurology* 54: 1494–1500.

Hugdahl K, Helland T, Færevåg MK, et al. (1995) Absence of ear advantage on the consonant–vowel dichotic listening test in adolescent and adult dyslexics: Specific auditory–phonetic dysfunction. *Journal of Clinical and Experimental Neuropsychology* 17: 833–840.

Hugdahl K, Nordstrand L, and Engstrand O (1986) A graphic-interactive CAD system for dichotic stimulus alignment (CADDIC). *Psychological Reports, 7, no. 3*. Bergen University of Bergen.

Hugdahl K and Wester K (1992) Dichotic listening studies of brain asymmetry in brain damaged patients. *International Journal of Neuroscience* 63: 17–29.

Wexler BE and Halwes T (1983) Increasing the power of dichotic methods: The fused rhymed words test. *Neuropsychologia* 21: 59–66.

Differentiation: The Cell Cycle Instead

A Philpott, University of Cambridge, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Generation of the central nervous system (CNS) requires a fine balance between the processes of cell proliferation, cell fate determination, and differentiation, but the coordination between these processes has only relatively recently begun to be understood. One can envisage models in which cell cycle status influences cell differentiation or, conversely, in which the process of differentiation has effects on the cell cycle. However, it is becoming increasingly clear that there are complex multidirectional links between processes at many stages (Figure 1). To understand these links, we must first examine what is known about cell division in the CNS.

Cell Division in the Nervous System

The entire nervous system must be generated from division of a relatively small number of initial precursors during development, and, not surprisingly, a strategy of both symmetrical and asymmetrical cell divisions is adopted. First, neuroepithelial (NE) cells, the stem cells of the mammalian nervous system, can divide symmetrically to generate two identical daughters in what are known as proliferative divisions. Alternatively, they can divide asymmetrically to produce one NE cell and one non-stem cell progenitor or differentiating neuron. Non-stem cell precursors may also divide again to form neurons and/or glia, generating enough cells to populate the nervous system. Neuroepithelial cell progeny can adopt any neural or glial cell fate. However, as development progresses, NE cells divide to produce stem cells with a more restricted range of fate known as radial glial (RG) cells, and most brain neurons are derived from RG cell divisions. Finally, cell proliferation is coordinated with regulated cell death to generate the nervous system.

The progeny of neurogenic divisions (i.e., divisions that generate differentiating neurons) must adopt the correct fate in response to both internal and external cues, and it is becoming increasingly apparent that this process is both coordinated with and influenced by the cell cycle. Indeed, this is obviously essential if the correct number and distribution of neurons and glia are to form in the animal, although the complexity of the relationships involved has meant that, until recently, relatively little has been known about the links between cell cycle and differentiation in the nervous system.

Interestingly, the transition from NE to RG cells and their subsequent progression from proliferative to neurogenic divisions may be influenced by cell cycle length; in particular, G_1 phase elongates as this process proceeds. Experimentally, when a chemical cyclin-dependent kinase (cdk) inhibitor olomucine was used in whole mouse embryo cultures at a dose sufficient to lengthen but not block the cell cycle, this resulted in premature neurogenesis. The 'cell cycle length hypothesis' has been proposed to explain these and other results, where it is suggested that a longer G_1 phase allows factors that promote progression from NEs to RGs to neurons greater time to act. Thus, a lengthened cell cycle will promote these transitions.

Cell Cycle and Cell Fate in the *Drosophila* Nervous System

Important information about the molecular links between cell cycle and cell fate has emerged from studying less complex organisms, which are amenable to genetic manipulation. This is perhaps most vividly illustrated by the differentiation of *Drosophila* neuroblasts, which divide and differentiate in a stereotypical manner that makes them particularly amenable to experimental study. Neuroblasts, the stem cells of the *Drosophila* nervous system, can self-renew but also generate a variety of different neurons and glia; a neuroblast division results in generation of another neuroblast and a smaller ganglion mother cell, which will itself divide again to generate two postmitotic neural or glial daughters. The number of times a neuroblast will divide depends on the expression of a sequential series of transcription factors, which are intimately involved in determining the identity of their progeny. If the cell cycle is blocked prior to neuroblast cell division using a mutant of the cell cycle activator String (cdc25), cells maintain expression of the first transcription factor in the series, hunchback. Therefore, it is passage of the cell cycle and not just passage of time that is needed for the changing pattern of transcription factor expression. However, when both String and Hunchback are knocked out, the cell cycle-arrested neuroblasts do go on to express the normal sequence of transcription factors, so this later phase must be cell cycle independent, indicating the complexity of the situation.

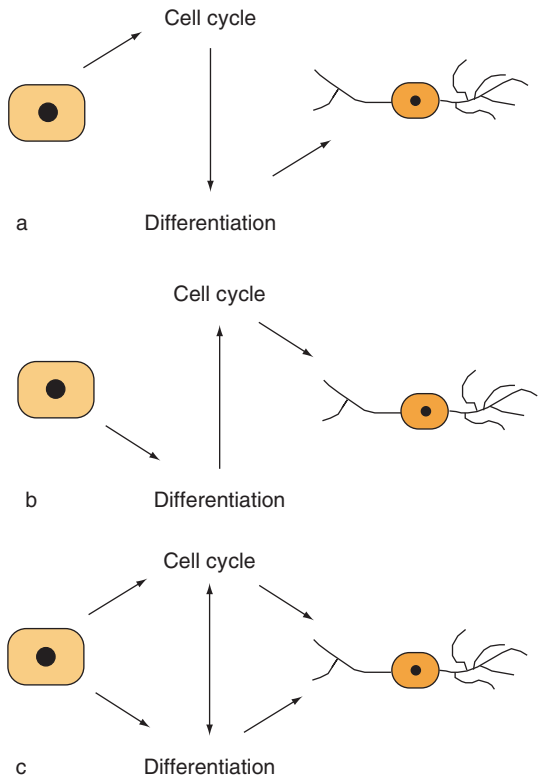

Figure 1 Models linking cell cycle and differentiation processes on the transition from neural precursor (left) to differentiated neuron (right).

Cell Cycle Stage and Cell Fate

The fate of the neurons produced from ganglion mother cell (GMC) divisions is generally determined by expression of transcription factors during the last cell cycle. The same phenomenon may be observed in *Xenopus* retinal progenitors, when transcription factors which determine specific cell fates are upregulated concomitant with a downregulation of cell cycle activator molecules.

Cortical development in mammals shows an unusual 'inside-out' morphology in which the older neurons are found nearer to the ventricular zone, whereas younger-born neurons migrate through to the upper cortical layers, and cells in different layers ultimately adopt specific fates. McConnell and Kaznowski investigated the influence of cell cycle phase on laminar identity in mice. Neuronal precursors from young animals were labeled with tritiated thymidine and then transplanted to an older host. If the cells were transplanted in S phase of the final cell cycle, they were found in layer 3, indicative of an earlier than expected fate, whereas if they were transplanted in mitosis, the cells were ultimately found in layer 6, their expected site. Therefore, it appears that laminar fate is determined in S phase of the final cell cycle and that cell cycle status influences differentiation. There are several ways in which cell cycle stage could influence cell fate, and these may be dependent on events specific to that stage.

Exit from the cell cycle, transiting from G_1 into a G_0 phase, is required for neuronal differentiation. Therefore, it is intuitive that factors influencing the cell cycle will influence differentiation; indeed, cell cycle regulators can directly influence the activity of determination factors. Moreover, there could be cell cycle-dependent expression of transcription factors that influence determination. This is clearly seen, for example, with the homeobox transcription factor prox1, required for determination of horizontal cells in the retina, which is maximally expressed at G_2 phase.

Other characteristics of specific cell cycle phases can also be exploited to coordinate cell fate determination and differentiation. For instance, as described previously, cortical neuronal cell fate seems to be largely determined in S phase. This may be a result of chromatin remodeling which can occur most efficiently at this phase, when complex chromatin structures are opened up to allow DNA replication. This suggestion is supported by evidence that chromatin remodeling complexes play a crucial role in neuronal differentiation. For instance, mice lacking Srg3, a core subunit of the SWI/SNF chromatin remodeling complex, have severe defects in neuronal proliferation and differentiation. Moreover, neurogenin is a central basic helix–loop–helix (bHLH) transcription factor that influences cell fate in many areas of the nervous system, at least in part, by influencing chromatin structure via recruitment of remodeling complexes.

Cell fate can also be determined in mitosis by asymmetrical distribution of determining factors, which is affected by orientation of the mitotic machinery. For instance, in *Drosophila*, Prospero is a transcription factor that is preferentially passed to GMCs and inhibits neuroblast (NB)-specific gene expression. Failure to divide results in repression of NB identity. In addition, blocking the function of the mitotic kinases cdc2 and Aurora A leads to impairment of cell fate determinant localization. Moreover, cyclin E expression may influence symmetry of neuroblast division in the *Drosophila* abdomen, independent of its effects on the cell cycle.

Cell Cycle Genes Regulate Cell Fate and Differentiation in the Nervous System

Differentiation of neurons is likely to be influenced by the proliferation status of the cells. In addition, many instances are now emerging of cell cycle regulators playing a direct role in influencing cell fate and differentiation in the nervous system and even playing a role in some postmitotic aspects of neuronal

function. Although our understanding is far from complete, some interesting themes are emerging, particularly the occurrence of two independent but complementary functions residing within the same cell cycle regulator. Such an approach allows precise coordination between cell cycle exit and differentiation processes and, as such, makes good developmental sense.

The retina is a good system in which to study whether cell cycle influences cell fate because a wide range of different cell types are produced from a single neurogenic precursor type. Normally, early born cell types arise from precursors that exit the cell cycle at an early stage, whereas the converse is true for late-born cell types. Cycling, *per se*, is not essential for producing cell diversity; blocking DNA replication early in retinogenesis still results in production of cells of all the expected histogenic cell types. However, the proportions of these cells are influenced by the cell cycle. For example, atonal-type bHLH transcription factors can influence histogenesis; overexpression of Xath 5 promotes early fates such as the production of retinal ganglion cells. Early cell cycle exit, brought about by co-overexpression of the *Xenopus* cdki Xic1, potentiates this effect, whereas inhibiting the cell cycle by overexpression of cyclin E does the opposite – inhibiting early and biasing toward late cell fates.

Cell Cycle Progression

The cell cycle is driven forward by the action of cdks (Figure 2) that phosphorylate key targets such as initiators of DNA replication, proteins that regulate nuclear envelope breakdown and spindle formation, and so on to bring about cell cycle transitions. Different cdks act at different phases of the cell cycle, and their activity depends on their binding to specific cyclins. Cyclins, in turn, are regulated not only by transcription but also at the level of protein stability.

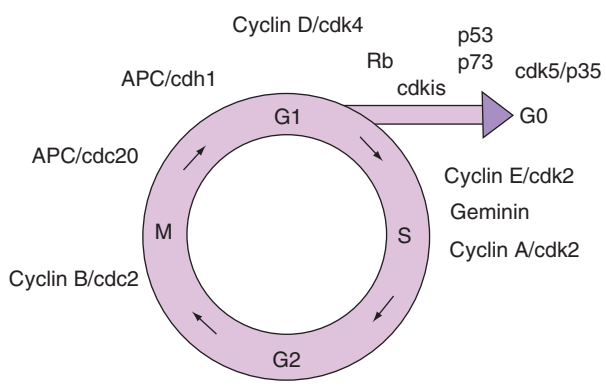

Figure 2 The cell cycle. Cell cycle regulators act at different phases of the cell cycle and also influence cell fate and differentiation in the nervous system.

For instance, the transition from G_2 to M phase is brought about primarily by the activity of cylin B/cdc2. To exit mitosis, cells must degrade cyclin B, and this is effected by ubiquitin-mediated proteolysis controlled by the anaphase-promoting complex (APC), bound to cdc20. Cyclin B levels are kept low in G_1 phase by the continuing activity of the APC, this time bound to cdh1.

Another key transition point in the cell cycle occurs in G_1 and is known as the restriction point (R). After cells have passed R, they can traverse the rest of the cell cycle without any further external stimulus, and passage through R corresponds to the phosphorylation and inactivation of the retinoblastoma protein. Alternatively, cells can exit the cell cycle before R and enter G_0, a process that is facilitated by the activity of cyclin-dependent kinase inhibitors (cdkis). After cell cycle exit into G_0, cells undergo differentiation (Figure 2).

Cyclin-Dependent Kinase Inhibitors

The role of cdkis in neuronal differentiation of mammals has been reasonably well studied, particularly using an oligodendrocyte model system. O-2A cells are bipotential progenitors that have the ability to differentiate into type 2 astrocytes or oligodendrocytes. Differentiation in the presence of mitogens is regulated by an internal clock that measures time rather than cell divisions, and the cdki p27Kip1 (Kip1) is a component of this clock; Kip1 levels rise progressively through O-2A cell divisions and then reach a plateau at the time of cell cycle exit and differentiation. Experimentally, the speed of the clock can be altered by growing cells at different temperatures. This works by changing the levels of Kip1, a protein whose stability is controlled by ubiquitin-mediated proteolysis. Interestingly, the cdki p21Cip (Cip1) is also involved in O-2A differentiation; Kip1 is needed for cell cycle arrest, but Cip1 may have a role in differentiation independent of its ability to control cell cycle progression. In mice, it has been shown that Kip1 can independently promote neural differentiation and migration in the cerebral cortex and at least some of these functions are independent of its ability to inhibit cdks. A third mammalian cdki, p57 Kip2 (Kip2), can regulate differentiation of neurons where it has been shown to play two roles in amacrine cell differentiation – promoting cell cycle exit and then postmitotically influencing amacrine subpopulation development.

Indeed, although the most well-established function of cdkis is to inhibit the kinase activity of cyclin-dependent kinases, many studies have demonstrated that cdk activity can directly affect neuronal

differentiation as well as affect postmitotic neuronal functions. Nerve growth factor (NGF)-induced differentiation of PC12 cells into neurons is associated with a decrease in cdc2 and cdk2 activity, and strikingly, inhibition of cdc2 and cdk2 activity alone is sufficient to trigger PC12 cell differentiation. Normally, Cip1 levels rise dramatically on PC12 cell differentiation, and indeed, ectopic expression of Cip1 alone is enough to mimic NGF-induced differentiation and trigger the formation of neurons.

The Cdki *Xenopus* Xic1 has complementary but independent roles in regulating other functions independent of its ability to arrest the cell cycle. Müller glial cells are the last cells to be born in the *Xenopus* embryonic retina, and this coincides with accumulation of the highest levels of Xic1. Overexpression of Xic1 in the retina strongly promotes formation of Müller glial cells at the expense of neurons. Mutational analysis of Xic1 has demonstrated that this activity resides near its N-terminus and is independent of its ability to arrest the cell cycle or inhibit overall cdk kinase activity. Similarly, Xic1 is required for differentiation of primary neurons in *Xenopus*, the first neurons to differentiate out of the neural plate. Again, this activity is independent of cell cycle effects and found near the N-terminus. Interestingly, in this case Xic1 has been shown to act at an early stage of neurogenesis, in parallel with the bHLH transcription factor neurogenin, where it may influence neurogenin protein stability. Such cell cycle-independent functions of cdkis in the mammalian nervous system have also recently emerged.

Cyclins and Cyclin-Dependent Kinases

Different levels of cyclins have interesting influences on neuronal cell fate and differentiation. Increasing the level of cyclin E in the developing *Xenopus* retina promotes late-born cell fate types. In mammals, cyclins D1 is found at high levels in proliferating cells of the retina and cerebellum, whereas cyclin D3 is found specifically in differentiating retinal Müller glial cells. Cyclin D2 knockout mice have a decrease in granule cell number and have no stellate interneurons. However, basket and Golgi interneurons are unaffected in the mice even though they come from the same precursor, suggesting that cyclin D2 influences cell type as well as cell cycle. Cyclin D1 null mice have reduced thickness of all retinal cell layers, possibly the result of a reduction in proliferation and specific apoptosis of photoreceptors. Surprisingly, cyclin D1 levels rise on differentiation of PC12 cells, but because there is apparently no increase in associated cdk4 kinase activity, the role of this is unclear.

Cdk5 is unusual among the cdks in that it does not seem to play a direct role in cell cycle progression. Instead, it has important functions in postmitotic neurons, in which it complexes with the p35 and p39 proteins to regulate differentiation and neurite outgrowth. However, because cdk5 can also complex with cyclin D and Kip1, it may also have influences on the cell cycle, although these have been poorly investigated. However, one study has shown that cdk5 can phosphorylate and stabilize Kip1 protein in postmitotic neurons, in which it is required to regulate neuronal migration.

The Anaphase Promoting Complex

Cell cycle progression is driven forward by the action of cdks, which are activated by binding to cyclins. To drive the cell cycle in one direction, cyclins must be cyclically destroyed and this is brought about via their ubiquitin (Ub)-mediated proteolysis. The APC is an E3 ligase required to facilitate conjugation of Ub to a variety of substrates, including cyclin B. In mitosis, APC is activated in part by the presence of a cdc20 subunit to target mitotic substrates, whereas this is replaced by cdh1 in G_1, in which the APC still has E3 Ub ligase activity. Remarkably, APC/cdh1 is highly expressed in postmitotic neurons, and a cdh1 knockdown has been shown to promote neurite outgrowth. Moreover, APC/cdh1 controls synaptic size in *Drosophila* via its ability to control stability of the liprin protein. Again, we see a protein complex with dual roles in controlling the cell cycle and other aspects of neuronal differentiation/function.

The Retinoblastoma Protein

One of the key targets for phosphorylation by cdks that allows them to control cell cycle progression is the retinoblastoma protein (Rb). Rb is required to regulate the passage through G_1 phase. At its most simple level, underphosphorylated Rb protein, found in early G_1, binds to and represses the transcriptional activity of the E2F family of transcription factors, whose targets include many genes that are essential for progression into S phase. As G_1 proceeds, Rb becomes progressively phosphorylated first by cyclin D/cdk4 and then by cyclin E/cdk2. Phosphorylated Rb is no longer able to bind to E2Fs, so it releases them to allow transcription of E2F-dependent genes and transition into S phase. However, it should be emphasized that this is a very simplified view of Rb function; Rb has been shown to bind numerous proteins besides E2F, where the function of these interactions is often unknown, and there are several members of the Rb family with at least partially overlapping functions.

Rb was first identified as a gene inherited in families with children who develop multifocal retinoblastoma

tumors of the eye, indicating an important role in cell proliferation and/or differentiation of the retina. Because Rb is thought to be a crucial cell cycle regulator in most, if not all, cell types, it is perhaps surprising that we still do not understand why the greatest tumor predisposition is for cancers of the nervous system. Mutations of Rb family members have also been found in astrocytomas, glioblastomas, and gliosarcomas, but it is also clear that Rb heterozygous individuals are susceptible to other nonneurological tumors later in life, particularly osteosarcomas. Indeed, some component of the Rb pathway, although not necessarily Rb, is thought to be mutated in the vast majority of human tumors.

An early function for Rb in the development of the nervous system was inferred from the severe neurological phenotype seen in Rb null mice, which die by embryonic day 14. However, it has been shown that many of these problems arise as a result of placental insufficiency; if Rb null embryos are rescued with wild-type placentas, then they have a relatively normal nervous system. However, Rb is strongly expressed in the CNS and, in common with Kip1, its overexpression alone is enough to induce differentiation of neuroblastoma cells, indicating specific roles in neural tissue.

It seems likely that Rb may influence neuronal differentiation independently of its ability to control cell cycle, and indeed, in some cultured cell systems, regulation of cell cycle and differentiation by Rb can be separated. At a molecular level, our best understanding of how Rb can directly influence differentiation comes from the study of myogenesis. Interestingly, studies mainly of cultured cells have shown that Rb not only facilitates cell cycle exit but also cooperates with the bHLH myogenic factor MyoD to activate the MEF2 transcription factor. It has also been reported that Rb may directly influence bHLH factor function by binding Id2, an inhibitor of bHLHs with which it can dimerize but these resultant dimers can no longer bind to DNA and activate transcription. Moreover, a study has shown that Rb may be a coactivator for the neurogenic bHLH NeuroD1.

The p53 Family

p53 is a transcription factor whose activity is disrupted in a wide range of tumors. p53 can transactivate both Cip1 and genes involved in apoptosis, and it has been implicated in neurogenesis, although its precise function has been difficult to identify. p53 does seem to accumulate in the nucleus of neuronal cells at an early stage of differentiation. Moreover, a dominant-negative form of p53 inhibits the cell cycle exit and differentiation of O-2A cells on treatment with thyroid hormone. p73, another p53 family member, can also induce the expression of Cip1. p73 null mice are missing the earliest sort of neurons, Cajal–Retzius cells. Moreover, retinoic acid treatment induces the differentiation of a retinoblastoma cell line and this results in upregulation of p73. Overexpression of p73 alone will promote differentiation of these cells, whereas a dominant negative form of p73 blocks retinoic acid-induced differentiation; therefore, it seems that p73 can also regulate both cell cycle and differentiation.

Geminin

Geminin is an excellent example of a protein with dual functions in cell cycle and differentiation. The gene was initially identified in two independent screens, one for proteins that could perturb early *Xenopus* neural development and the other searching for proteins that are specifically degraded during mitosis. It has subsequently become apparent that geminin has at least two functions: (1) to regulate the loading of the mini-chromosome maintenance proteins onto prereplicative DNA prior to S phase and (2) to maintain neuronal progenitors in a proliferative precursor-like state. It performs this second function by apparently binding to and blocking the function of Brg1, a component of the SWI/SNF chromatin remodeling complex, which is required to act with the bHLH neurogenic transcription factors Neurogenin and NeuroD to promote neural differentiation. Geminin levels decrease in differentiating neurons and this may be required to trigger neuronal differentiation. Moreover, other studies have shown that geminin can also interact with Six- and Hox-related transcription factors, so its ability to modulate transcription both within and outside the nervous system may be widespread.

Differentiation and Patterning Factors Regulate the Cell Cycle

Based on the evidence described previously, it is clear that both cell cycle status and cell cycle regulators influence many aspects of cell fate choice and differentiation in the nervous system. Conversely, much evidence is emerging that molecules controlling cell fate choice and differentiation also have direct effects on the cell cycle.

A paradigm for what we might expect to occur in nerves is seen during the differentiation of skeletal muscle: the bHLH transcription factor MyoD not only drives expression of muscle structural genes but also upregulates cdkis to promote cell cycle exit. This is complemented by interactions with Rb and

cyclinD/cdk4 that regulate MyoD function. Similar interactions may occur in differentiating neurons. Mouse P19 embryonal carcinoma cells can be differentiated into neurons on overexpression of several neurogenic bHLH proteins, including neurogenin 1 and neuroD2, and this is accompanied by upregulation of Kip1 and cell cycle exit. Moreover, achaete-scute, another bHLH protein, transcriptionally regulates the mitotic phosphatase and CDC25 homolog string. Also in *Drosophila*, Prospero, an asymmetrically distributed homeodomain protein, controls the expression of cyclin E, cyclin A, and string. In addition, it can work indirectly via the bHLH protein deadpan to control levels of dacapo, a *Drosophila* cdki.

Expression of other tissue-specific transcription factors in the nervous system influences cell proliferation but only in the regions where they are usually expressed, indicating a requirement for cofactors; overexpression of Optx2, Six3, and Rx1 results in very large eyes due to extra cell proliferation, whereas loss of Optx2 results in the upregulation of two cdkis. Interestingly, the winged helix transcription factor XBF1 has dual roles, both recruiting cells to a neural fate and controlling their proliferation via regulation of cdki transcription, having different effects at high and low doses.

The internal clock that drives the sequential expression of a series of transcription factors in *Drosophila* neuroblasts, leading their progeny to adopt different fates, has been described previously. This transcription factor series may also be used as a way for NB cells to 'count' their number of cell divisions; prolonged expression of early genes such as Hunchback and Kruppel results in the generation of extra ganglion mother cells of an early type. A similar counting mechanism may exist for NBs in postembryonic flies, in which loss of a coiled coil protein, Mushroom body defect, results in generation of extra progeny.

Signal transduction pathways responsive to extracellular ligands are also known to regulate both cell fate and differentiation and cell cycle coordinately, although the molecular details of these interactions have only recently come to light. The two pathways about which probably the most is known are those of Notch and Hedgehog (Hh) signaling.

The Notch signaling pathway is used widely throughout the nervous system, and its role in allowing scattered neurons to differentiate out from a field of equivalent neural precursors has been particularly well studied. In general, active Notch signaling keeps neural cells in an undifferentiated state and, in several systems, also favors gliogenesis at the expense of neurogenesis. In some contexts, such as the mouse nervous system, Notch signaling keeps cells in cycle, and this has been proposed as a mechanism whereby cells remain undifferentiated. However, in other systems, Notch activation promotes cell cycle exit, such as in the *Xenopus* retina and neural plate. How does Notch signaling influence the cell cycle? Again, this seems to be context dependent. Hes1, a component of the Notch signaling pathway, can repress Cip1 expression in PC12 cells and inhibit NGF-induced differentiation, whereas in the skin, Notch induces Cip1 and promotes differentiation. Confusingly, in the *Xenopus* neural plate, Notch signaling inhibits cdki expression but nevertheless promotes cell cycle exit, possibly by downregulation of cyclin A and cdk2 mRNAs. This may reflect an effect of Notch required to inhibit the differentiation function of the cdki, rather than to influence its cell cycle function. Moreover, Notch may control levels of cdkis post-translationally via effects on the level of E3 ubiquitin ligase SCF Skp2, which can target Kip1 and Cip1 for proteolysis.

The Hh pathway regulates cell proliferation in several areas of the vertebrate CNS, including the hippocampus, cerebellum, forebrain, spinal cord, and retina. Again, the effects of Hh signaling on proliferation seem to be context and species dependent. For instance, upregulation of Hh signaling in the mouse retina results in proliferative cells persisting inappropriately, whereas in the zebrafish loss of Sonic Hh results in inhibition of Kip2 and increased proliferation.

Other signaling pathways influence cell cycle in the CNS. Wnt signaling is also likely to play an important role in precursor cell proliferation by regulating levels of D-type cyclins and c-myc. If the downstream Wnt effector β-catenin is put under a neuron-specific enhancer, these mice develop enlarged brains. Moreover, transforming growth factor-α has a dose-dependent effect on proliferation in the retina. Other pathways, such as fibroblast growth factor and bone morphogenetic protein signaling, influence cell fate and differentiation in the neuroectoderm, and it seems likely that in the future we will obtain a better understanding of their influence over cell cycle events in the CNS.

Conclusions

Neural precursors must proliferate, adopt the correct fate, and then differentiate at the appropriate time. Thus, it seems obvious that there must be constant feedback between these processes for development of the exquisitely patterned CNS, yet our understanding of these links has only recently begun to emerge. One key point has been the realization that several cell cycle regulators have complementary but separable roles in the processes of fate determination, differentiation, and cell cycle control, providing important insight into the

elegant ways that have evolved to coordinate these events. From this base, our understanding is likely to grow considerably in the near future.

See also: Cerebral Cortex: Symmetric vs. Asymmetric Cell Division; Helix–Loop–Helix (bHLH) Proteins: Proneural; Helix–Loop–Helix (bHLH) Proteins: Hes Family; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Olfaction in Invertebrates: Drosophila; Terminal Differentiation: REST.

Further Reading

Becker EBE and Bonni A (2005) Beyond proliferation – cell cycle control of neuronal survival and differentiation in the developing mammalian brain. *Seminars in Cell & Developmental Biology* 16: 439–448.

Cremisi F, Philpott A, and Ohnuma S (2003) Cell cycle and cell fate interactions in neural development. *Current Opinion in Neurobiology* 13: 26–33.

Fichelson P, Audibert A, Simon F, and Gho M (2005) Cell cycle and cell-fate determination in *Drosophila* neural cell lineages. *Trends in Genetics* 21: 413–420.

Galerisi U, Jori FP, and Giordano A (2003) Cell cycle regulation and neural differentiation. *Oncogene* 22: 5208–5219.

Götz M and Huttner WB (2005) The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology* 6: 777–788.

Maurange C and Gould AP (2005) Brainy but not too brainy: Starting and stopping neuroblast divisions in *Drosophila*. *Trends in Neurosciences* 28: 30–36.

McConnell SK and Kaznowski CE (1991) Cell cycle dependence of laminar determination in developing neocortex. *Science* 254: 282–285.

Ohnuma S and Harris WA (2003) Neurogenesis and the cell cycle. *Neuron* 40: 199–208.

Ohnuma S, Philpott A, and Harris WA (2001) Cell cycle and cell fate in the nervous system. *Current Opinion in Neurobiology* 11: 66–73.

Diffusion Tensor Imaging (DTI)

S Mori and J Zhang, The Johns Hopkins University, Baltimore, MD, USA

© 2009 Published by Elsevier Ltd.

Introduction

The complex anatomy of the human brain poses a challenge for neuroscientists to study its functions and for clinicians to diagnose neurological diseases. There are an array of imaging techniques that can examine neuroanatomy at different levels of resolution, from the optic or electron microscope at the cellular and molecular levels to other modern imaging techniques at the levels of tissue or organ. These imaging techniques offer diverse image contrasts that can reveal gene expression, cellular composition, metabolism, and so on. For judging whether an imaging technique is appropriate for a particular application, the resolution and contrast are among the most important criteria. The imaging contrast determines whether we can obtain the information we need, and the resolution determines the amount of detail.

Magnetic resonance imaging (MRI) has been widely used in neuroscience research and clinical diagnosis. The resolution of MRI varies from the millimeter level for clinical imaging to the submillimeter level for animal imaging. It can visualize brain structures and detect lesions in the central nervous system (CNS) noninvasively. One of the advantages of MRI over other imaging techniques, such as computed tomography and ultrasound, is that it provides versatile soft tissue contrasts, which can reveal underlying tissue anatomy and physiological states. Figure 1 shows four conventional MRIs acquired from a healthy volunteer, each demonstrating unique tissue contrasts in the brain. In the proton-density, T1-, and T2-weighted images, the contrast between the white matter (WM) and gray matter (GM) reflects the underlying difference in tissue proprieties, such as water and myelin content. Although the contrast between GM and WM in the diffusion-weighted image is not as sharp as in other images, previous reports have shown that diffusion-weighted images can provide unique information in several pathological conditions, such as stroke and edema in the CNS.

In this article, we assume that readers already have some knowledge of MRI, and we focus on a novel MR-based imaging technique, diffusion tensor imaging (DTI), which provides new tissue contrasts that conventional MRI lacks. First introduced in the early 1990s, it has been quickly adopted in clinics and for neuroscience research to study the macroscopic brain connectivity and pathology of several neurological diseases, such as multiple sclerosis and brain damage following stroke. Many review articles cover the basic and technical aspects of DTI. In this article, we give a brief introduction to the principles and applications of DTI.

What Is Water Molecule Diffusion?

Water molecules undergo random translational motion driven by internal kinetic energy. This phenomenon is referred to as self-diffusion or simply diffusion. Diffusion is different from flow, which is often driven by mechanical pressure or concentration gradient. For flow, most water molecules are moving in a determined direction similar to blood flows in the blood vessels with a pressure gradient. For diffusion, movements of water molecules are random, with no determined directions. Because of this randomness, we use the so-called diffusion coefficient (D) to characterize the diffusion process. With a given diffusion coefficient, the root mean squared distance traveled by a water molecule in one dimension during a measurement time period t_{diff} is $\sqrt{2Dt_{\mathrm{diff}}}$, where t_{diff} is the diffusion time. In three-dimensional space, the root mean squared distance will be $\sqrt{6Dt_{\mathrm{diff}}}$. The Strokes–Einstein equation tells us that the diffusion coefficient in an unbounded solution environment (free diffusion) without exchange is related to the temperature (T) and the molecular friction coefficient inside the medium (f) as $D = kT/f$, where k is the Boltzmann constant, and the friction coefficient depends on the radius of the molecule and the viscosity of the medium. At body temperature, the diffusion coefficient of water molecules in an unbounded environment is approximately $3.0 \times 10^{-5}\,\mathrm{cm^2\,s^{-1}}$, which means that for a time period of 10 ms, the root mean square distance traveled by a water molecule is approximately 5 μm, comparable to the radius of a neuron.

Inside biological tissue, many structures, such as the cell membranes and myelin sheath, can restrict or hinder the movement of water molecules. There also exist exchanges between different tissue water compartments. These changes from free diffusion to restricted or hindered diffusion mean that the root mean square distance traveled by a water molecule depends on its environment (i.e., the existence of barriers to diffusion) and should be shorter than in the case of free diffusion. To characterize the restricted or hindered water molecule diffusion in biological tissues, the so-called apparent diffusion coefficient (ADC) is used.

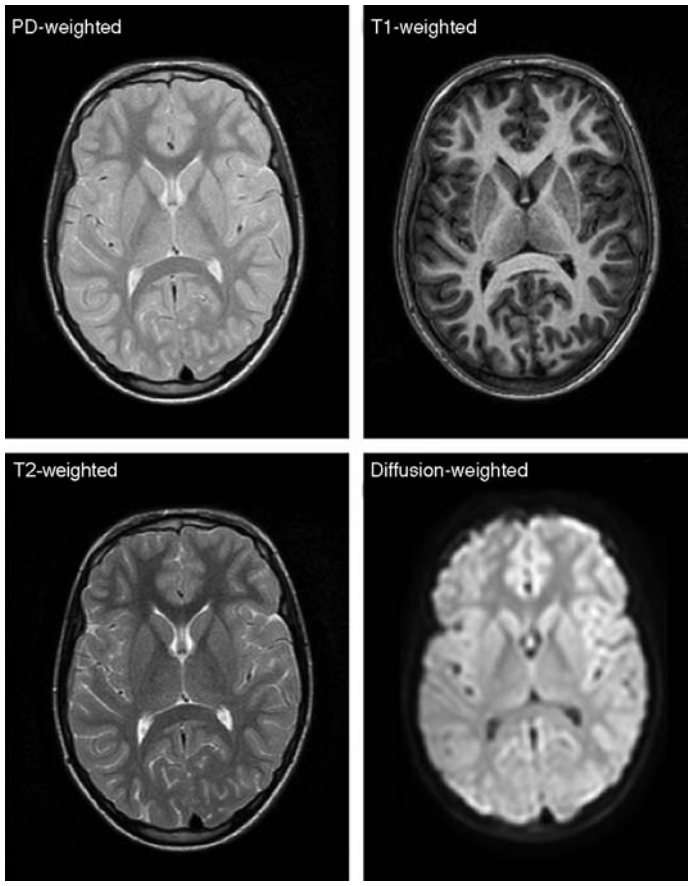

Figure 1 Axial proton density-weighted (PD-weighted), T1-weighted, T2-weighted, and diffusion-weighted images acquired from one healthy adult volunteer. The white and gray matter can be easily distinguished in these images. For the diffusion-weighted image, the resolution is $2.5 \times 2.5 \times 2.5$ mm. For the other images, the resolution is $1 \times 1 \times 1$ mm.

How to Measure Diffusion by MRI?

MRI can measure water diffusion, or ADC. The effect of water diffusion on nuclear magnetic resonance signals in the presence of magnetic field gradients has been described by Hann, Carr, and Purcell. In his paper on spin echo, Hahn showed that the diffusion of water molecules (spins) attenuates the observed signal amplitude. Carr and Purcell further presented the equations that describe the relationships between spin echo signal attenuation and the diffusion process. Later, Torrey incorporated the effect of diffusion into the Block–Torrey equation. Based on these theories, it is possible to measure diffusion using the Stejskal and Tanner sequence, which adds a pair of diffusion-sensitizing gradient pulses to the standard spin echo sequence.

Figure 2 explains how to use a standard spin echo with the addition of diffusion-sensitizing gradient pulses to measure water molecule diffusion. After a 90° radio frequency (RF) excitation pulse, spins start to rotate at a frequency ω ($\omega = \gamma B$, where γ is the gyromagnetic ratio and B is the main magnetic field). The diffusion-sensitizing gradient pulse after the RF pulse has duration of δ and a magnitude of G. It adds a spatial-dependent phase shift to the spins so that the spin at the position $+r$ along the gradient axis (Figure 2, red circle) is leading the spin at the origin (green circle) in phase, and the spin at the position $-r$ along the gradient axis (blue circle) is behind the spin at the origin in phase. Following the first gradient pulse, the 180° RF pulse along the y axis flips the sign of the spins' phases so that the spins with positive phase (leading) before the 180° RF pulse become spins with negative phase afterwards (lagging), whereas lagging spins will become leading spins. The second gradient pulse will have the same effect on the spins as the first gradient pulse (i.e., adding a spatial-dependent phase shift to the spins). If the spins have not undergone any translational motion between the two gradient pulses (Figure 2(d)), the phase shift introduced by the second gradient pulse will cancel the initial phase shift, and all spins will have the same phase and add coherently to form a signal at the receiver. However, if the spins undergo translational motion along the gradient axis between the two gradient pulses (Figure 2(e)),

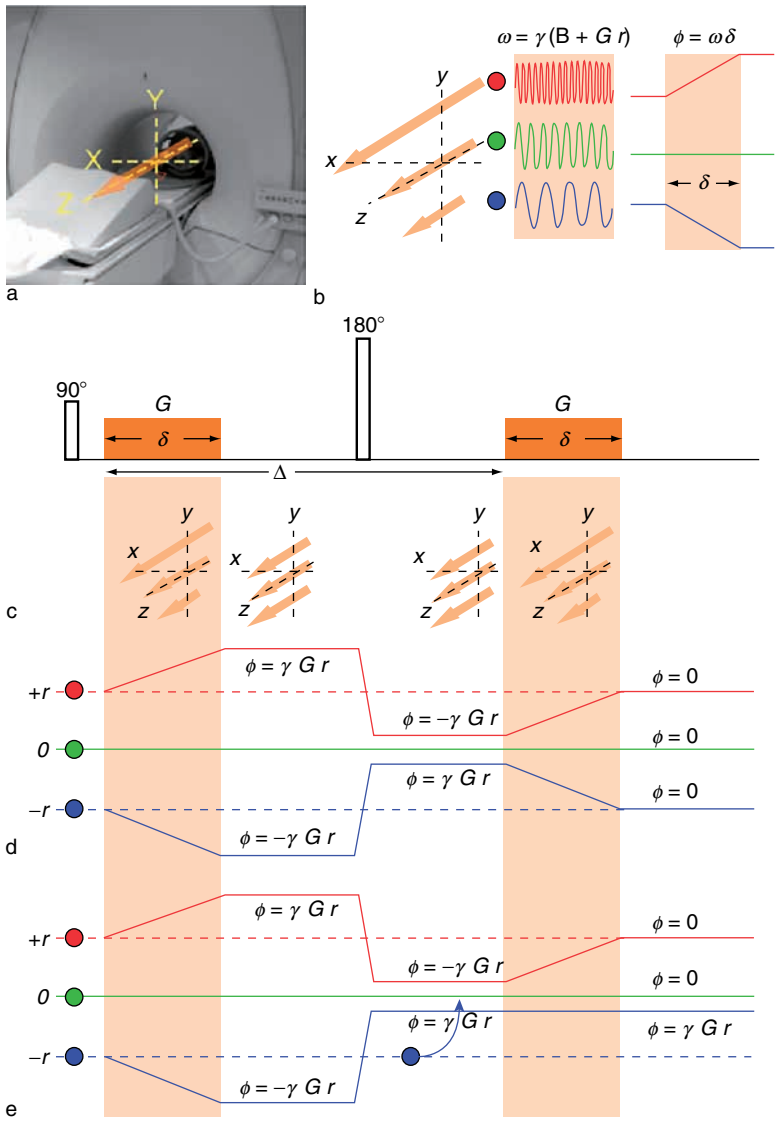

Figure 2 Schematic diagram of MR measurement of water molecule diffusion. (a) In a typical MR scanner, the long axis of the magnet (i.e., the direction of the main magnetic field) is designated as the *z* axis. (b) A magnetic gradient along the *y* direction can affect the signals from three spins (red, green, and blue circles; their frequencies and phases) located along the *y* axis. (c) A standard spin echo sequence with diffusion-sensitizing gradients (orange boxes) around the 180° RF pulse. The gradients are placed along the *y* direction. The magnetic field as the gradients are on and off are illustrated below the sequence. (d) In the case of three static spins positioned at $+r$, 0, and $-r$ along the *y* axis, the phase evolution of the three spins is plotted. In the end, all three spins have the same phase. (e) In the case of three nonstatic spins at the same position as in (d), where the spin located at $-r$ (blue circle) diffuses from $-r$ to 0 between the two diffusion-sensitizing gradients, the phase evolution of the three spins is plotted. In the end, the three spins have different phases.

there will be extra phase shifts that cannot be completely canceled by the second gradient pulse, and the additional phase shifts depend on the distance of spin movements (diffusion). As a result, signals from diffused spins (blue circle) will have different phase from those of the static spins (red and green circles) and cannot be added coherently. The acquired signal in Figure 2(e) will have smaller magnitude than the signal acquired in Figure 2(d) (i.e., signal attenuation). If there are no diffusion-sensitizing gradients, spins will all have the same phase disregard of their movement.

Figure 2 is a highly simplified illustration of what happens during diffusion measurement. In a routine imaging experiment, the region covered by a pixel contains millions of spins, and the measured signal attenuation reflects the statistical distribution of spin displacement in the pixel. The images acquired with diffusion gradient pulses are called diffusion-weighted images, and the images acquired without diffusion gradient pulses are called non-diffusion-weighted images. By comparing the diffusion-weighted and non-diffusion-weighted images, we can obtain a map of ADC.

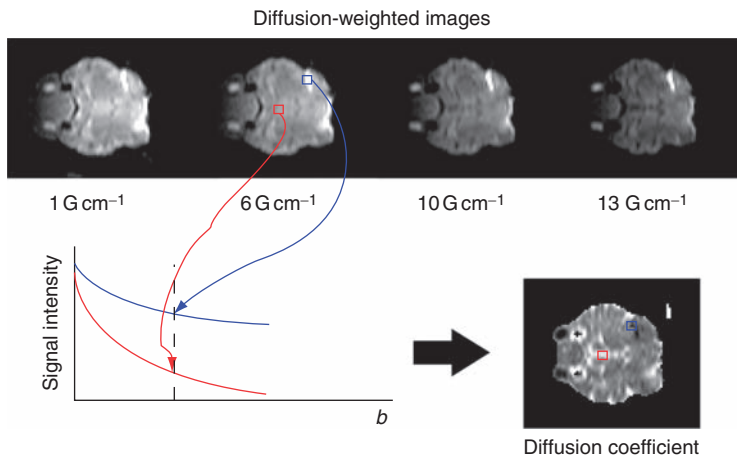

Figure 3 An example of measuring tissue ADC. Diffusion-weighted images with different diffusion weighting (b) were acquired. In these images, the increase in b was achieved by increasing diffusion gradient magnitude from 1 to 13 G cm^{-1}. As b increases, there is increasing attenuation of the image intensity. For each pixel, the signal intensity acquired with different b is fitted to a monoexponential model, which results in a map of ADC. The signal at the red pixel decays more rapidly than that at the blue pixel; therefore, the ADC value at the red pixel is higher than the ADC value at the blue pixel.

Figure 3 illustrates how to measure ADC using the MR sequence described previously. In a simplified monoexponential model, the signal decay follows the equation $S = S_0 e^{-b\mathrm{ADC}}$, where S and S_0 are the amplitudes of signals in diffusion-weighted and non-diffusion-weighted images, respectively, and b is the factor of 'diffusion weighting.' A simplified equation for the b value is $b = \gamma^2 G^2 \delta^2 (\Delta - (\delta/3))$, where γ is the gyromagnetic ratio. According to this equation, the b value, or the diffusion weighting, is determined by the magnitude of the diffusion-sensitizing gradient (G), the duration of the gradient (δ), and the duration between the start of the first gradient and the start of the second gradient (Δ). The contrast in the diffusion-weighted images depends on the tissue diffusion property, relaxation time, and proton density. In order to obtain contrast that depends only on diffusion, it is common to fit the intensity values acquired with different diffusion weightings, or b values, to amonoexponential model (Figure 3). The result is an ADC map.

The sequence shown in Figure 2 is only sensitive to displacements of water molecules along the axis of the diffusion-sensitizing gradients (i.e., the direction with changing magnetic field). An analogy of this property is an altimeter, which is only sensitive to changes in altitude, not changes in latitude or longitude. Figure 4 shows how the sequence can separate displacements in different directions. Suppose the gradient axis is along the x axis (i.e., the magnetic field is changing along the x axis). If spins move along the x axis between the two diffusion gradients (Figure 4(a)), they will have incoherent phases. If spins move along the y axis instead (Figure 4(b)), the phases will be coherent. The bottom panel of Figure 4 demonstrates differences in tissue contrasts in diffusion-weighted images when diffusion gradients are placed along the x, y, and z axes – a phenomenon caused by anisotropic diffusion.

In biological tissues, the ADC along different axes may be different; that is, the ADC depends on the geometrical configurations of diffusion barriers that water molecules encounter. Diffusion which has the same ADCs measured along different axes is called isotropic diffusion; otherwise, it is called anisotropic diffusion. For example, in neuronal fiber tracts, the ADC measured along the trajectory of such tracts is greater than the ADCs measured perpendicular to the trajectory. It is not clear exactly which structure causes anisotropic diffusion. It is possible that the axonal membrane and myelin sheath (in myelinated fiber tracts) act as barriers that restrict water molecule diffusion in the direction perpendicular to the fiber tracts.

What Is Diffusion Tensor?

We have shown that water molecule diffusion along a particular axis can be characterized by ADC measured along the axis. To characterize water molecule diffusion in three dimensions, especially for anisotropic diffusion, it is necessary to measure ADC along multiple axes. Questions arise as to how many axes should we measure and how to quantify and visualize diffusion anisotropy. In order to answer such questions, the concept of diffusion tensor was introduced. A diffusion tensor, in mathematical terms, is a three-by-three, symmetric, positive definite matrix as shown here:

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix}$$

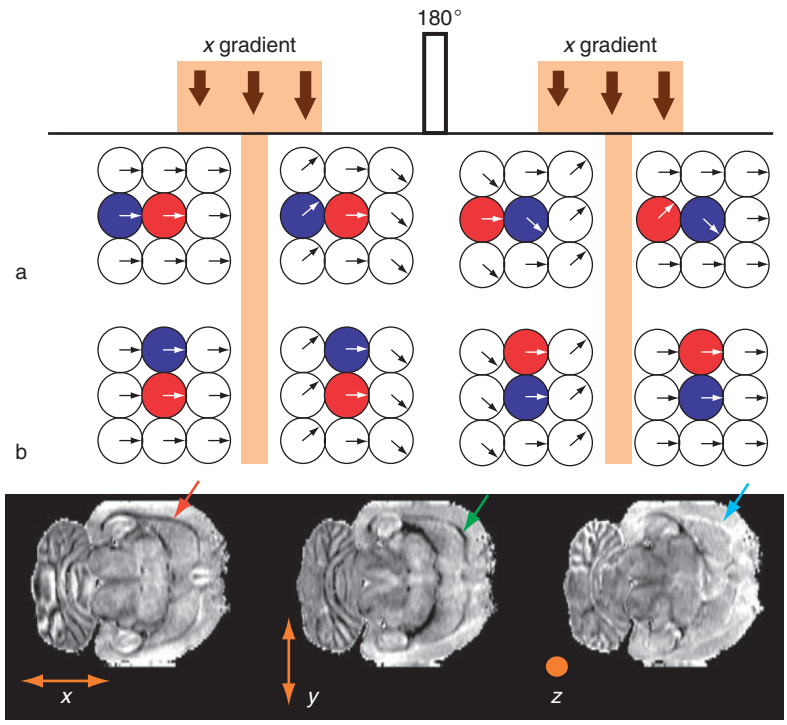

Figure 4 Measurement of direction-dependent ADC. The arrows at the top represent the magnetic field along the *x* axis. Spins are represented by circles, and their phases are visualized by arrows inside the circles. (a) Displacement of spins (red and blue circles) along the *x* axis will produce incoherent phases, as shown by the arrows in the circles. (b) Displacement along the *y* axis will not produce any phase incoherence. (Bottom) Diffusion-weighted images of a postmortem rat brain, with diffusion-sensitizing gradients along the *x*, *y*, and *z* directions. In these images, the contrast between the external capsule, a WM structure in the rat brain, and surrounding GM changes with diffusion gradients along different axes.

Diffusion tensors can be estimated from a minimal set of seven diffusion-weighted images using regression methods. The set of diffusion-weighted images should include one non-diffusion-weighted image and at least six diffusion-weighted images whose gradient directions are noncollinear and non-coplanar. In clinical scanners, imaging protocols that acquired diffusion-weighted images along multiple directions (>7) are commonly used. Because it is symmetric ($D_{xy} = D_{xy}$, $D_{zy} = D_{yz}$, and $D_{zx} = D_{xz}$), we only need to know the diagonal elements (D_{xx}, D_{yy}, and D_{zz}) and three off-diagonal elements (D_{xy}, D_{xz}, and D_{yz}). The diagonal elements are the ADC along the *x*, *y*, *z* axes, and the off-diagonal elements show the coupling between the three axes. For each diffusion tensor, we can compute three real, positive eigenvalues λ_1, λ_2, and λ_3, which are usually arranged in decreasing order (i.e., $\lambda_1 \geq \lambda_2 \geq \lambda_3$), and three eigenvectors, $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ that correspond to λ_1, λ_2, and λ_3, respectively. Here, λ_1, λ_2, and λ_3 are the ADCs along the axes defined by $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, respectively. $\mathbf{v}_1$ is the orientation with maximal diffusion coefficient, and $\mathbf{v}_3$ is the orientation with minimal diffusion coefficient.

One way to visualize diffusion tensor is to use the diffusion ellipsoid. The distance between any point on the ellipsoid's surface to the center of the ellipsoid is proportional to the root mean squared distance that water molecules travel along the line between the points and the center. The $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ of the diffusion tensor make up the three axes of the ellipsoid, and the length of each axis is the eigenvalue of the corresponding eigenvector. As the ratio between the three eigenvalues changes, the shape of the ellipsoid will also change. When the three eigenvalues have similar values (i.e., $\lambda_1 \approx \lambda_2 \approx \lambda_3$), the shape of the ellipsoid is spherical. When $\lambda_1 \approx \lambda_2 > \lambda_3$, the shape of the ellipsoid is called planar. When $\lambda_1 > \lambda_2 \approx \lambda_3$, the shape of the ellipsoid is called tubular. The diffusion ellipsoid of a voxel that contains a single population of axonal fibers appears as elongated along the orientation of the fiber trajectory ($\lambda_1 \gg \lambda_2$), as shown in Figure 5(a), and therefore is considered as a tubular tensor. In this case, the $\mathbf{v}_1$ of the tensor will have the same orientation as the fibers. When a voxel contains two groups of crossing neuronal fibers or two parallel surface barriers with low permeability, there may be two or more orientations within a plane that show maximal diffusion coefficient. In these cases, the corresponding diffusion ellipsoid has a planar shape (Figures 5(b) and 5(c)), as suggested by several authors.

In more complex cellular environments (Figures 5(d) and 5(e)), anisotropy of water diffusion may disappear and result in a sphere-like diffusion tensor.

Although it is possible to use diffusion ellipsoid to visualize tensor data in three dimensions, it is convenient to have scalar or vector measurements that are easier for two-dimensional display and easy quantification. Diffusion anisotropy is an important measurement because it reflect and existence of well-organized tissue structures. The degree of diffusion anisotropy can be measured by various anisotropy indices. Table 1 lists a few commonly used indices.

In the CNS, WM tracts have a high degree of diffusion anisotropy, and the orientation of fiber trajectory can be depicted by the orientation of maximal diffusivity ($\mathbf{v}_1$) from the tensor. The information on WM diffusion anisotropy and orientation obtained from diffusion tensor can generate unique tissue contrast in the CNS. Figure 6 compares a conventional axial T1-weighted image of human brain with results from DTI. In the T1-weighted image, the region of WM appears rather homogeneous. It is difficult to distinguish a single WM tract in the T1-weighted image. In contrast, the image of diffusion anisotropy reveals considerable contrast within the WM (i.e., WM tracts have varying degrees of diffusion anisotropy). Major WM tracts, such as the corpus callosum and corona radiata, have a high degree of anisotropy and appear as bright structures in the anisotropy image. Adding orientation information into the image, there is additional contrast to distinguish individual WM tracts due to their difference in orientation. To visualize orientation information, we use the direction-encoded color map, or simply color map, in which the anterior–posterior orientation is represented by green, left–right orientation is represented by red, and superior–inferior orientation is

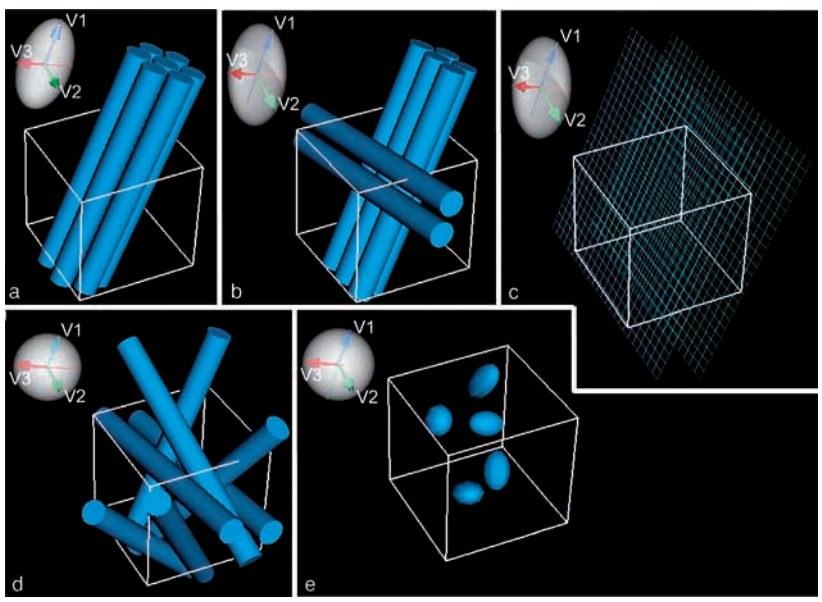

Figure 5 The shape of diffusion tensors and local tissue microstructures. (a) If there is only one group of fibers (blue cylinders for fibers) inside an imaging voxel (the cubic frame), the diffusion tensor acquires a tubular shape. (b, c) If there are two crossing fiber groups (b) or barriers with low permeability (c) blue wireframe surfaces), the diffusion tensor will have a planar shape. (d, e) For more complex local tissue structures, for example, multiple crossing fiber groups (d) or no fibers (e), the diffusion ellipsoids become spherical.

Table 1 Frequently used indices of diffusion anisotropy, their definition using eigenvalues, and references

Anisotropy indices	*Definition*	*Reference*
Fractional anisotropy (FA)	$\mathrm{FA} = \sqrt{\frac{(\lambda_1-\lambda_2)^2+(\lambda_1-\lambda_3)^2+(\lambda_2-\lambda_3)^2}{2.(\lambda_1^2+\lambda_2^2+\lambda_3^2)}}$	Basser and Pajevic (2000)
Relative anisotropy (RA)	$\mathrm{RA} = \sqrt{\frac{(\lambda_1-\lambda_2)^2+(\lambda_1-\lambda_3)^2+(\lambda_2-\lambda_3)^2}{(\lambda_1+\lambda_2+\lambda_3)^2}}$	Alexander et al. (2000)
Linear, planar, and spherical anisotropy indices (CL, CP, and CS, respectively)	$\mathrm{CL} = \frac{\lambda_1-\lambda_2}{\lambda_1+\lambda_2+\lambda_3}, \mathrm{CP} = \frac{2(\lambda_2-\lambda_3)}{\lambda_1+\lambda_2+\lambda_3}, \mathrm{CS} = \frac{3\lambda_3}{\lambda_1+\lambda_2+\lambda_3}$	Westin et al. (2002)

represented by blue, as shown by the color arrows in Figure 6, and the overall intensity proportional to diffusion ansiotropy. WM structures, which are now colored by their orientations and shaded by their degree of anisotropy, can be easily separated from each other. For example, the cingulum, a fiber bundle that travels in the anterior–posterior orientation, is green. The corpus callosum, which connects the left and right hemisphere, is red. The corticospinal tract, which connects the cortex and spinal cord and travels in the superior–inferior orientation, is blue. Furthermore, using this orientation information and diffusion anisotropy measurements, it is possible to reconstruct axonal bundles in three dimensions. The DTI tractography uses streamlines that follow, pixel by pixel, the directions of $\mathbf{v}_1$'s to reconstruct WM pathways (Figure 7). Results from tractography are difficult to validate in humans but have been achieved in animals with manganese. This technique enables clinicians and neuroscientist to see, *in vivo*, the path of major WM tracts and to study the neuronal connectivity between different parts of the CNS. A recent DTI-based atlas on human brain WM may be a useful tool for interested readers.

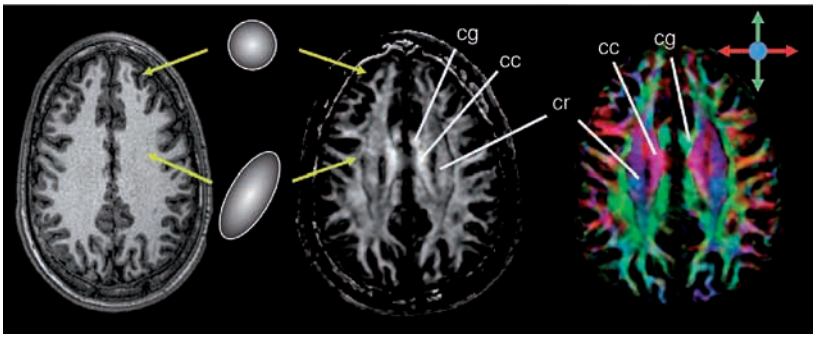

Figure 6 Axial T1-weighted (left), fractional anisotropy map (middle), and direction-encoded color map (right) from a healthy volunteer. In the fractional anisotropy map, regions of the GM appear dark because the diffusion in the GM is mainly isotropic. Regions of the WM appear bright. The color map, in which orientation of WM structures is encoded in red, green, and blue, provides additional image contrast to distinguish individual WM tracts. cc, corpus callosum; cg, cingulum; cr, corona radiata.

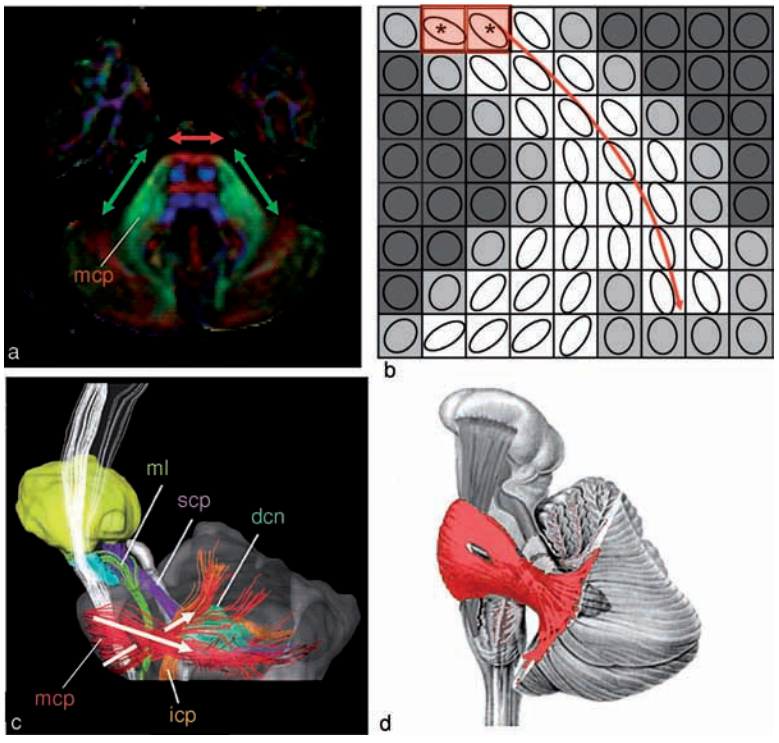

Figure 7 (a) An axial DTI color map image of human brain stem. The medial cerebellar peduncle (mcp) contains axonal fibers that travel along the directions of the arrows. (b) Schematic diagram of fiber tracking. It begins with a number of seed pixels and follows the principal direction of diffusion tensors within regions with values of diffusion anisotropy above a threshold. (c) Fiber tracking results of various white matter tracts in the brain stem. Notice the similarity between mcp reconstructed from DTI results and illustrated in a brain atlas (d). dcn, deep cerebellar nuclei; icp, inferior cerebellar peduncle; ml, medial lemniscus; scp, superior cerebellar peduncle.

Applications and Limitations of DTI

The advantages of DTI make it an ideal tool to study WM structures in the CNS. Because DTI is sensitive to the organization and microstructure of WM, it can be used to detect abnormalities caused by diseases that affect WM and follow the progression longitudinally. It has been shown that DTI can detect Wallerian degeneration in WM caused by stroke and axonal injuries due to multiple sclerosis and amyotrophic lateral sclerosis. Sun et al. showed that loss or damage to myelin can also result in a decrease in diffusion anisotropy. Although both axonal injury and myelin injury will result in loss of anisotropy, detailed examinations show that the axial diffusivity (i.e., λ_1 from the tensor) is sensitive to axonal injury, whereas the transverse and radial diffusivity (i.e., the average of λ_2 and λ_3) is sensitive to myelin injury. One possible explanation is that loss of myelin will enable water molecules to move with less restriction in the transverse directions, whereas loss of axons will pose barriers to water molecules to move in the axial direction.

Another area in which DTI has shown great potential is the study of brain development and maturation. Conventional MRI often produces poor contrast in premature brains, mainly due to lack of myelin at the early stages. This poses a major challenge for neuroscientists to study brain maturation and for clinicians to diagnose injuries in pediatric patients. In comparison, DTI can provide superior WM contrast in neonatal brains. Even without myelin, the early WM tracts possess a relatively high degree of anisotropy. The DTI contrast can be used for delineation of early WM structures and, more important, the measured diffusion anisotropy provides information on WM microstructure and its maturation.

As with any imaging modalities, DTI has its limitations. First, DTI can only provide the path of WM tracts, not the direction of signal transduction. Second, DTI is limited by imaging resolution, which is approximately 2 or 3 mm in humans and 0.08 mm in postmortem animal studies. Therefore, if one does not see a certain WM tract in DTI, this does not mean that it is missing. Validation from histology or other modalities is often necessary. Third, diffusion tensor is a simplified model. It assumes that water diffusion is Gaussian. In the case of two or more fibers crossing in a pixel, DTI cannot distinguish these fibers. To overcome this problem, various data acquisition and postprocessing approaches have been proposed, such as diffusion spectrum imaging, Q-ball imaging, and high angular resolution diffusion imaging.

See also: Cognition: An Overview of Neuroimaging Techniques; Executive Function and Higher-Order Cognition: Neuroimaging; Neuroimaging; Perfusion MRI; Single Photon Emission Computed Tomography (SPECT): Technique.

Further Reading

Alexander AL, Hasan KM, Kindlmann G, Parker DL, and Tsuruda J (2000) A geometeric analysis of diffusion tensor measurements of the human brain. *Magnetic Resonance in Medicine* 44: 283–291.

Basser PJ and Jones DK (2002) Diffusion-tensor MRI: Theory, experimental design and data analysis – A technical review. *NMR in Biomedicine* 15: 456–467.

Basser PJ and Pajevic S (2000) Statistical artifacts in diffusion tensor MRI (DT-MRI) caused by background noise. *Magnetic Resonance in Medicine* 44: 41–50.

Beaulieu C (2002) The basis of anisotropic water diffusion in the nervous system – A technical review. *NMR in Biomedicine* 15: 435–455.

Hermoye L, Saint-Martin C, Cosnard G, et al. (2006) Pediatric diffusion tensor imaging: Normal database and observation of the white matter maturation in early childhood. *NeuroImage* 29: 493–504.

Huppi P, Maier S, Peled S, et al. (1998) Microstructural development of human newborn cerebral white matter assessed *in vivo* by diffusion tensor magnetic resonance imaging. *Pediatric Research* 44: 584–590.

Kingsley PB (2006) Introduction to diffusion tensor imaging mathematics. *Concepts in Magnetic Resonance* 28A: 101–179.

Le Bihan D (1995) Molecular diffusion, tissue microdynamics and microstructure. *NMR in Biomedicine* 8: 375–386.

Le Bihan D (2003) Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience* 4: 469–480.

Mori S, Wakana S, van Zijl PCM, and Nagae-Poetscher LM (2005) *MRI Atlas of Human White Matter.* New York: Elsevier.

Mori S and Zhang J (2006) Principles of diffusion tensor imaging and its applications to basic neuroscience research. *Neuron* 51: 527–539.

Neil JJ (1997) Measurement of water motion (apparent diffusion) in biological systems. *Concepts in Magnetic Resonance* 9: 385–401.

Pajevic S and Pierpaoli C (2000) Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: Application to white matter fiber tract mapping in the human brain. *Magnetic Resonance in Medicine* 43: 921.

Pierpaoli C, Barnett A, Pajevic S, et al. (2001) Water diffusion changes in Wallerian degeneration and their dependence on white matter architecture. *NeuroImage* 13: 1174–1185.

Price WS (1997) Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion. Part 1: Basic theory. *Concepts in Magnetic Resonance* 9: 299–336.

Sotak CH (2002) The role of diffusion tensor imaging in the evaluation of ischemic brain injury – A review. *NMR in Biomedicine* 15: 561–569.

Westin CF, Maier SE, Mamata H, Nabavi A, Jolesz FA, and Kikinin R (2002) Processing and visualization for diffusion tensor MRI. *Medical Image Analysis* 6: 93–108.

Wiegell MR, Larsson HB, and Wedeen VJ (2000) Fiber crossing in human brain depicted with diffusion tensor MR imaging. *Radiology* 217: 897–903.

Dopamine

J D Elsworth and R H Roth, Yale University School of Medicine, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Anatomy

Mesencephalon

The majority of central dopaminergic projections arise in the midbrain from the ventral tegmental area (VTA; A10), substantia nigra (SN; A9), and retrorubral area (A8) to innervate three principal sets of targets: the striatum (the caudate and putamen), the limbic cortex (medial prefrontal, cingulate, piriform, and entorhinal areas), and other limbic structures (regions of the septum, olfactory tubercle, nucleus accumbens, and amygdaloid complex). These systems are frequently referred to as the nigrostriatal, mesocortical, and mesolimbic dopamine (DA) projections, respectively. Although the A9 cells preferentially innervate the striatum, and the A10 cells mainly target the limbic and cortical regions, Figure 1 illustrates that the projections of the midbrain DA cells are not restricted to these regions. Many of the SN DA neurons give off collateral branches and ramify within more than one region. The projection fields of the A8, A9, and A10 cell groups are collectively known as the mesotelencephalic DA system. This term has been used to refer to the projections from the midbrain DA neurons (A8, A9, and A10) to the striatum and nucleus accumbens. The A8–A10 groups comprise approximately 20 000–30 000 neurons on each side of the rat brain. Also located in the mesencephalon, the rostral linear nucleus is sometimes considered part of the VTA. DA neurons of the rostral linear nucleus and periaqueductal gray region project locally or to different brain regions, such as the central amygdaloid nucleus, bed nucleus of the stria terminalis, sublenticular extended amygdala, and substantia innominata.

It is now known that heterogeneities of DA innervation occur within a region. Thus, the nucleus accumbens can be divided into two cytoarchitectonically, physiologically, and pharmacologically distinct compartments, termed core and shell. The core region and the shell region receive preferential inputs from the A9 and A10 cell groups, respectively. Furthermore, two histochemically distinct compartments within the striatum have been defined, known as 'patch' (or 'striosome') and 'matrix,' which contain different densities of a number of markers for dopaminergic, cholinergic, and peptidergic neurons and represent another level of processing in the basal ganglia. Evidence suggests that the mesostriatal DA input to these compartments arises from distinguishable subsets or 'tiers' of the DA-containing midbrain neurons.

Diencephalon

These systems include the A11–A15 cell groups, which are located in the diencephalon, including regions such as the hypothalamus, caudal thalamus, zona incerta, and preoptic region. These neurons project principally within the diencephalon and have been divided into periventricular, incertohypothalamic, and tuberohypophyseal systems. There are approximately 2000–4000 DA neurons on each side of the brain in the rat diencephalic A11–A15 groups. An additional long-length system is the descending projections from the A11 group in the diencephalon to the spinal cord, where a dopaminergic innervation is apparent in the dorsal horn at all levels of the cord.

Other Systems

Among the other DA systems are the short-length interplexiform amacrine-like neurons, which link inner and outer plexiform layers of the retina. There is also a population of periglomerular DA neurons in the olfactory bulb, which link mitral cell dendrites in separated adjacent glomeruli.

Function

The different locations of the various DA systems in the brain dictate that they have different afferent and efferent connections, which in turn determine the roles they play in brain function. These are reviewed here. It should also be appreciated, however, that it is not possible to completely segregate the motor, motivational, and cognitive behaviors that are often attributed to the nigrostriatal, mesolimbic, and mesocortical DA systems, respectively.

Nigrostriatal System

Dopaminergic innervation of the basal ganglia plays an essential role in many aspects of motor control, cognition, and emotion. The striatum is one of the principal input structures of the basal ganglia. The importance of the DA input to the striatum is evident from the motor abnormalities (e.g., bradykinesia, rigidity, and tremor) exhibited by patients with Parkinson's disease, which is characterized by a marked loss of nigrostriatal DA neurons. DA innervations to limbic and cortical regions are also affected in Parkinson's disease but to a lesser extent than the

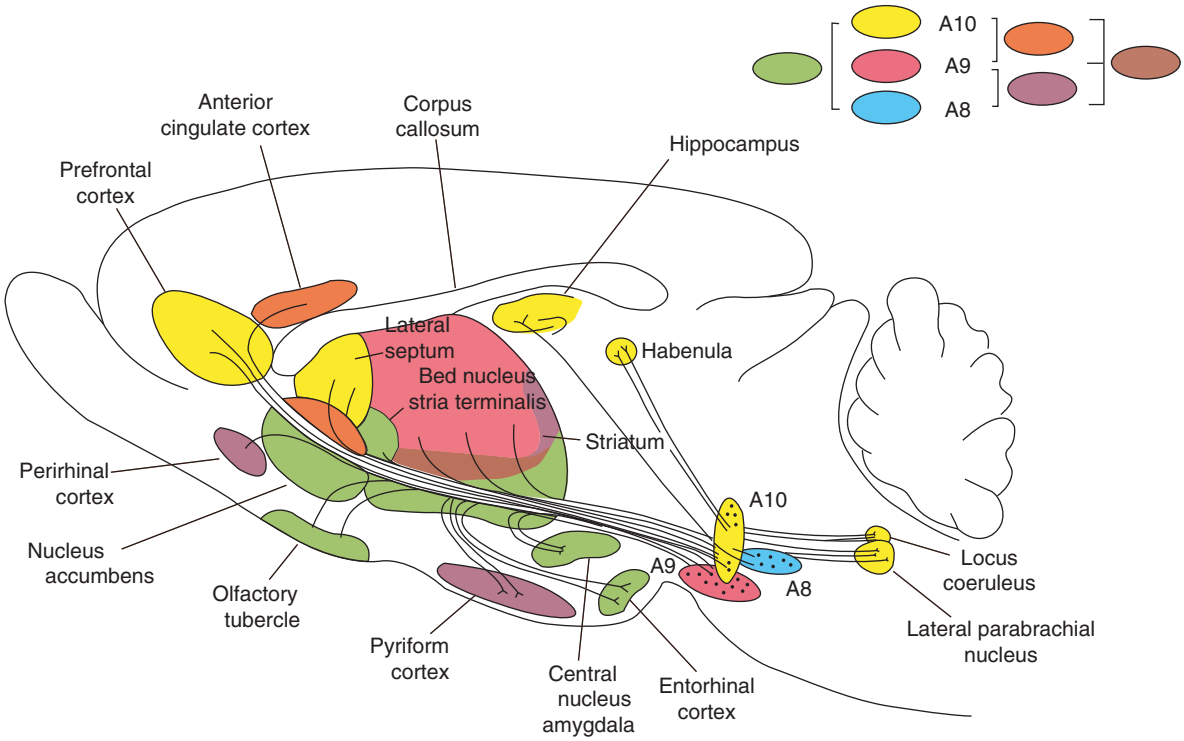

Figure 1 Schematic diagram illustrating the distribution of the main neuronal pathways containing DA in the central nervous system. The cell groups are named according to the 1965 nomenclature of Dahlstrom and Fuxe. Different colors denote the terminal field projections of the A8, A9, and A10 cell groups. Recent investigations have established that collaterals of ascending DA neurons also provide innervation of the subthalamic nucleus, thalamus, and globus pallidus.

striatal input. DA release in the striatum modulates activity in two striatopallidal circuits ('direct' and 'indirect'), which in turn exert control over thalamocortical circuits essential for voluntary control of movement. The signs of Parkinson's disease do not appear until a large majority (approximately 80%) of nigrostriatal DA neurons have been lost, as surviving DA neurons efficiently increase their activity to compensate for the damage. It is now apparent that nigrostriatal DA neurons do not merely allow motor behavior to occur but that they also play an important role in the selection and initiation of actions and establishing motor skills and habits. Current treatments for Parkinson's disease typically involve administration of the DA precursor, dihydroxyphenylalanine (L-dopa), or DA receptor agonists. Future treatments are likely to include restoration of dopaminergic tone by transplantation of fetal DA neurons or stem cells or by gene therapy.

Mesolimbic System

The function of the mesolimbic dopaminergic system, particularly the projection from the VTA to the nucleus accumbens, has been strongly implicated in goal-oriented (motivated) behaviors, in addition to reward, attention, and pharmacologically induced locomotion. Enhancement of DA transmission in this system has been linked with the addicting, reinforcing, and sensitizing effects of repeated exposure to psychostimulant drugs of abuse. Furthermore, the therapeutic effects of antipsychotic drugs used in the treatment of schizophrenia may depend on the inhibition of mesolimbic DA neuron activity that these drugs induce. The ability of antipsychotic drugs to block DA receptors and thereby reduce DA transmission is central to the 40-year-old DA hypothesis of schizophrenia, which posits that the disease is related to excessive central DA activity. Imaging studies in patients have provided support for the existence of disrupted DA transmission in schizophrenia. A revision to the hypothesis is the concept of underactivity in cortically projecting DA neurons together with overactivity in subcortical DA systems. Persuasive evidence implicates developmental, functional, and structural abnormalities in schizophrenia, involving other transmitter systems besides DA, such as glutamate and γ-aminobutyric acid (GABA).

Mesocortical System

The prefrontal cortex has rich connections with other neocortical regions, limbic regions, and other subcortical regions. The prefrontal cortex has been implicated in a wide variety of cognitive functions, and it particularly appears to be involved in directing

appropriate attention, prioritizing the significance of stimuli, monitoring the temporal sequence of stimuli, referencing stimuli to internal representations or cues, and devising abstract concepts. The well-defined DA projection to the prefrontal cortex has been suggested to be involved in short-term and 'working' memory. It has been hypothesized that the mesoprefrontal DA system is involved in the pathophysiology of schizophrenia and that defects in this system, or its associated neuronal connections, may be responsible for the negative and cognitive deficits that characterize the disorder.

Diencephalon Systems

DA neurons in posterior dorsal hypothalamus and periventricular gray of central thalamus (A11) project to spinal cord and appear to have a role in sensory and nociceptive processing and sensory integration. Tuberoinfundibular DA neurons (A12) are located in the arcuate nucleus and periventricular nucleus of mediobasal hypothalamus and project to the external layer of median eminence. They play an important role in inhibiting the release of prolactin from the anterior lobe of the pituitary. Incertohypothalamic DA neurons (A13) reside in the rostral portion of medial zona inserta and terminate in the central nucleus of amygdala, horizontal diagonal band of Broca, and paraventricular nucleus of the hypothalamus. Their functions are not clear but probably involve integration of autonomic and neuroendocrine responses to sensory stimuli. Two subpopulations of A14 neurons have been distinguished. The periventricular–hypophysial (tuberohypophyseal) DA neurons terminate in the intermediate lobe of the posterior pituitary and inhibit secretion of α-melanocyte-stimulating hormone and pro-opiomelanocortin-derived peptides such as β-endorphin. DA neurons also innervate the neural lobe, but their origin is controversial. The other defined group of A14 DA neurons is located in the periventricular hypothalamic nucleus. Fibers of these neurons extend into the medial preoptic nucleus and anterior hypothalamus. Females possess greater number of these neurons than males, and these neurons are believed to play a role in gonadotrophin secretion in females and reproductive behavior in males. DA neurons in the ventrolateral hypothalamic comprise the A15 cell group, and they extend processes to the lateral hypothalamus and caudal supraoptic nucleus, suggesting a role in oxytocin and/or vasopressin secretion.

Other Systems

Retinal DA neurons in vertebrates appear to be particularly involved in contrast sensitivity and light adaptation. In 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-treated parkinsonian monkeys, a loss of retinal DA occurs, along with spatial frequency-dependent abnormalities in both the pattern electroretinogram and the visual evoked potential. In Parkinson's disease, abnormalities occur in ocular motor control, visual evoked potentials, electroretinograms, color vision, and visual contrast sensitivity. Changes are also possible in the acquisition and perception of visual information.

In addition to the DA systems in the brain and pituitary, DA appears to have a restricted role in the periphery. The best defined role for peripheral DA neurons is in renal regulation of sodium homeostasis, and impairment of this function may underlie development of hypertension.

Life Cycle

Figure 2 is a schematic model of a dopaminergic nerve terminal illustrating the life cycle of DA and the regulatory mechanisms that modulate the synthesis, storage, and release of this important neurotransmitter. A brief description of this idealized nerve terminal and model synapse will help to summarize the present state of knowledge concerning the life cycle and regulatory control of DA in a typical DA synapse and possible sites at which drugs and chemicals can influence dopaminergic transmission.

Synthesis

Blood-borne tyrosine, derived from dietary proteins and from phenylalanine metabolism, enters the brain by a low-affinity amino acid transport system. Tyrosine in brain extracellular fluid is taken up into DA neurons by high- and low-affinity amino acid transporters. The relative circulating levels of tyrosine and phenylalanine can affect central catecholamine metabolism because these amino acids compete for transport into the brain and for transport into the neuron. DA is synthesized from tyrosine in two enzymatic steps. The first reaction is catalyzed by the enzyme tyrosine hydroxylase in the cytosol. Because this step is rate limiting, it sets the pace for the conversion of tyrosine to DA and is the step most susceptible to physiologic regulation and pharmacologic manipulation. Short-term activation of tyrosine hydroxylase involves phosphorylation of the regulatory domain by protein kinases. The activated form of tyrosine hydroxylase is thought to have a lower K_m for its pterin cofactor and a higher K_i for DA, which effectively reduces end product inhibition. The activity of the enzyme dihydropteridine reductase is indirectly linked to DA biosynthesis because this enzyme catalyzes the recycling of the

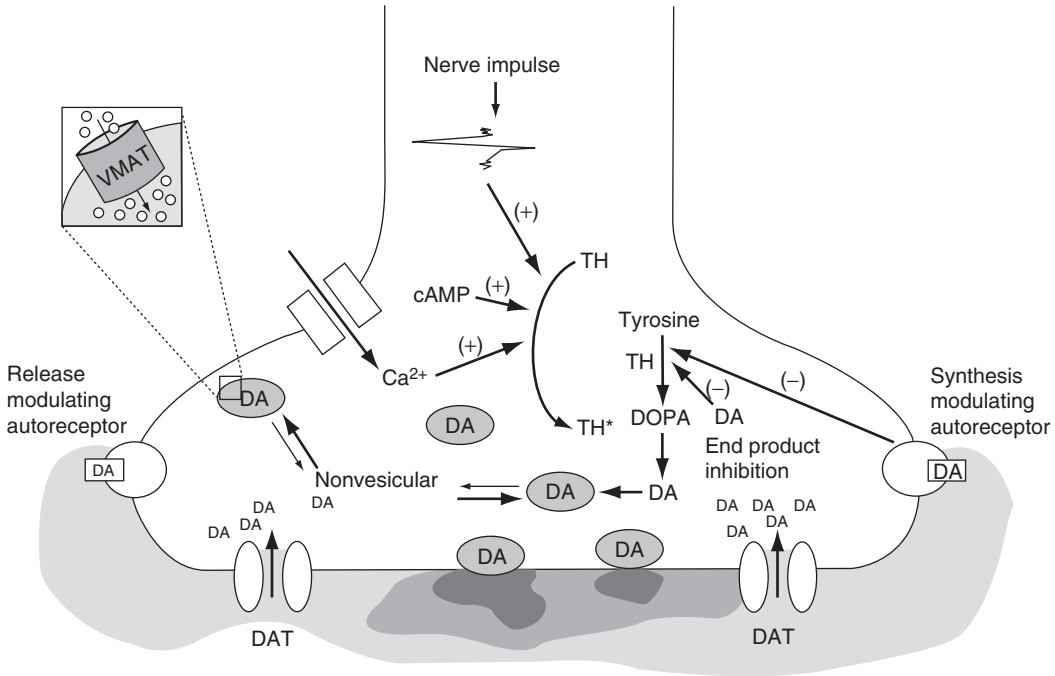

Figure 2 Schematic model of a dopaminergic nerve terminal illustrating the life cycle of dopamine (DA) and the mechanisms that modulate its synthesis, release, and storage. Invasion of the terminal by a nerve impulse results in the Ca^{2+}-dependent release of DA. This release process is attenuated by release-modulating autoreceptors. Increased impulse flow also stimulates tyrosine hydroxylation. This appears to involve the phosphorylation of tyrosine hydroxylase (TH), resulting in conversion to an activated form with greater affinity for tetrahydrobiopterin cofactor and reduced affinity for the end product inhibitor DA. The rate of tyrosine hydroxylation can be attenuated by (1) activation of synthesis-modulating autoreceptors, which may function by reversing the kinetic activation of TH, and (2) end product inhibition by intraneuronal DA, which competes with cofactor for a binding site on TH. Release- and synthesis-modulating autoreceptors may represent distinct receptor sites. Alternatively, one site may regulate both functions through distinct transduction mechanisms. The plasma membrane DA transporter (DAT) serves an important physiological role in the inactivation and recycling of DA release into the synaptic cleft. The vesicular monoamine transporter (VMAT) transports cytoplasmic DA into storage vesicles, decreasing the cytoplasmic concentration of DA and preventing metabolism by MAO. VMAT modulates the concentration of free DA in the nerve terminals. Reproduced from Cooper JR, Bloom FE, and Roth RH (2003) *The Biochemical Basis of Neuropharmacology*, 8th edn. New York: Oxford University Press, with permission from Oxford Press.

quinonoid dihydrobiopterine to tetrahydrobiopterine, which is an essential cofactor of tyrosine hydroxylase. In addition, the synthesis of tetrahydrobiopterine is dependent on the activity of another enzyme, GTP-cyclohydrolase-1. The second step is catalyzed by L-dopa decarboxylase in the cytosol. This enzyme is also referred to as aromatic amino acid decarboxylase because it catalyzes the decarboxylation of several endogenous aromatic amino acids. This enzyme so avidly decarboxylates L-dopa that the levels of this amino acid in brain are very low under normal conditions.

Metabolism

DA can be metabolized in brain by several enzymatic reactions, which are summarized in Figure 3. The major mammalian enzymes of importance in the metabolic degradation of catecholamines are monoamine oxidase (MAO) and catechol *O*-methyltransferase (COMT). MAO converts catecholamines to their corresponding aldehydes. This aldehyde intermediate is rapidly metabolized, usually by oxidation by the enzyme aldehyde dehydrogenase to the corresponding acid. In some circumstances, the aldehyde is reduced by aldehyde reductase. MAO is located on the outer membranes of mitochondria and thus, in brain, is present primarily in nerve terminals and glia. Separate genes encode two isoforms of MAO (types A and B), which can be distinguished by substrate specificity and sensitivity to the irreversible selective inhibitors. In brain, MAO-A is preferentially located in dopaminergic and noradrenergic neurons, whereas MAO-B appears to be the major form present in serotonergic neurons and glia. MAO is a particle-bound protein localized largely in the outer membrane of mitochondria. Usually considered to be an interneuronal enzyme, it also occurs in abundance extraneuronally.

The second enzyme of importance in catabolism of catecholamines is COMT. This is a relatively nonspecific enzyme that catalyzes the transfer of methyl groups from *S*-adenosylmethionine to the *m*-hydroxyl

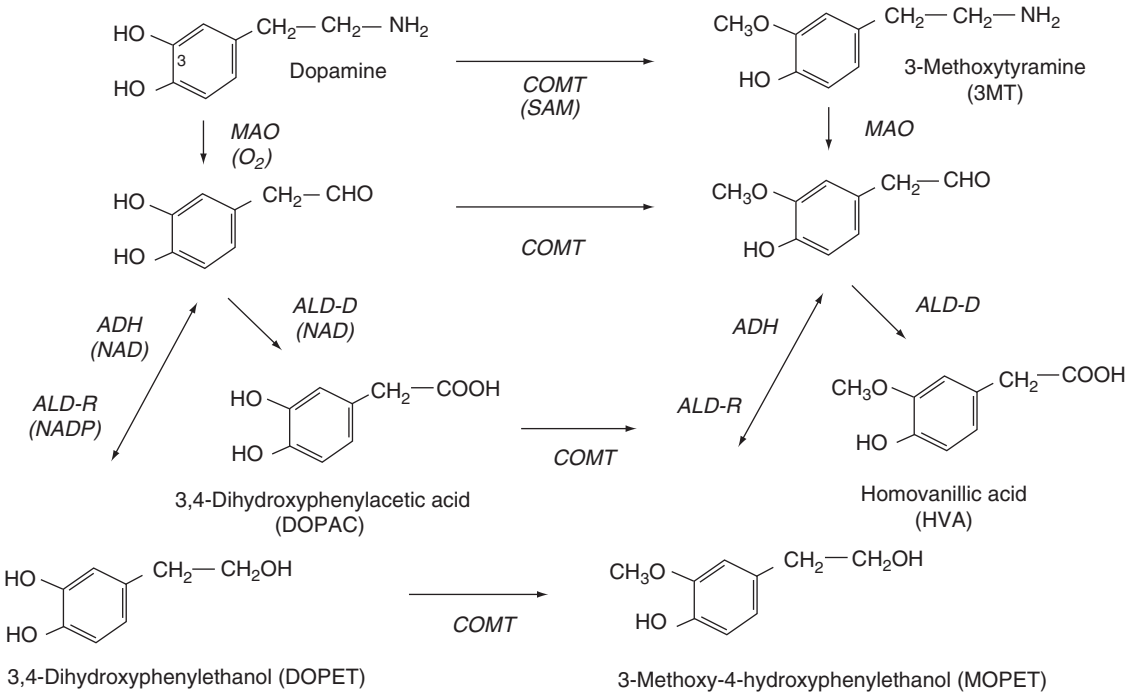

Figure 3 Main pathways of catabolism of DA. Enzymes involved (shown in italics) are monoamine oxidase (MAO), catechol-*O*-methyltransferase (COMT), aldehyde dehydrogenase (ALD-D), aldehyde reductase (ALD-R), and alcohol dehydrogenase (ADH). Cofactors for the enzymes are shown in parenthesis; SAM, *S*-adenosyl-L-methionine. In addition, dopamine (DA) and metabolites are substrates for phenolsulfotransferase; sulfation occurs on the 3-position, indicated on the structure for DA. In brain, DA is metabolized mainly to acidic (DOPAC and HVA) metabolites, with less formation of alcoholic metabolites (DOPET and MOPET); in the periphery, the alcohol metabolites may be metabolized further by alcohol dehydrogenase. In the periphery, glucoronyltransferase is able to form glucoronide conjugates of DA and metabolites.

group of catecholamines and various other catechol compounds. There are two isoforms of COMT, a membrane-bound form and a soluble form. Membrane-bound COMT is the major form found in the central nervous system (CNS), has a higher affinity for catecholamines, and is located principally in neurons. The soluble form has a lower affinity for catecholamines and is the major form expressed in the periphery, but it is present also in CNS glia. The membrane-bound isoform of COMT, having a high affinity for DA, is expressed at neuronal dendritic processes in cortex and striatum, but the expression and activity of COMT are higher in frontal cortex than in striatum. Relevant to this is evidence that in the prefrontal cortex DA transporter (DAT) is rarely expressed within synapses and exerts minimal influence on DA flux in the prefrontal cortex. Thus, COMT activity appears to have a more important function in regulating DA neurotransmission in the frontal cortex than in other regions. In fact, it has been estimated that the flux of DA through the COMT pathway exceeds 60% in the prefrontal cortex but only 15% in the striatum, where DAT is the chief mechanism for terminating the action of DA. The importance of COMT to the actions of DA in the prefrontal cortex is strongly supported by the finding that compared with wild-type mice, COMT knockout mice have increased DA levels in prefrontal cortex, but not in striatum, and that such mice perform better on prefrontal cortex-dependent behavioral tasks. Another interesting facet to the role of COMT in frontal cortex is the finding that the COMT gene contains a polymorphism (Val158Met) that affects the *in vivo* activity of the enzyme. Met158 homozygotes have approximately one-third less COMT enzyme activity in prefrontal cortex than Val158 homozygotes. Consistent with its role in modulating prefrontal cortex DA levels, the Val158Met polymorphism is associated with performance on tests of working memory and executive function, which depend on prefrontal cortex function. Thus, the high-activity Val158 allele is linked with relatively poorer performance on such tasks, relative to the Met158 allele, presumably as a result of increased DA metabolism. The COMT polymorphism has been implicated in a number of neuropsychiatric phenotypes. In particular, there is evidence to support an association between COMT allele frequency and the genetic risk of schizophrenia. Working memory is highly dependent on DA function in the prefrontal cortex, and working memory dysfunction is a cardinal feature of schizophrenia.

The major DA metabolites found in brain differ, depending on the species under study. In general, however, acidic rather than neutral metabolites predominate. In rodents, the primary metabolites of DA found in the CNS are 3,4-dihydroxyphenylacetic acid (DOPAC) and homovanillic acid (HVA) and a small amount of 3-methoxytramine. DOPAC usually predominates. In addition, a considerable proportion of these metabolites in rodent brain are found in the form of conjugates. In both human and nonhuman primates, however, the major DA metabolite is HVA, and only a small amount is found in the conjugate form. It is well documented that in the rat, short-term fluctuations in DOPAC and HVA concentrations in the striatum, nucleus accumbens, and prefrontal cortex provide a useful index, respectively, of alterations in impulse flow in the nigrostriatal, mesolimbic, and mesocortical DA pathways.

Transporters

Presynaptic vesicles or granules present in the nerve terminals are specialized for the uptake and storage of DA, thus protecting it from degradation by the enzyme MAO found in mitochondria present in the nerve terminal. The storage granules contain a vesicular monoamine transporter (VMAT2) that has a pharmacology distinct from the plasma membrane DAT found on the nerve terminals. The Na^+- and Cl^- dependent DAT actively transfers the DA from the synaptic cleft back into the presynaptic cell, where it can be stored and reused. DA is concentrated to approximately 0.1 M in the storage vesicles, 10–1000 times higher than the level in the cytosol.

The tuberoinfundibular, tuberohypophysial, olfactory bulb, mesoprefrontal, and mesoamygdala DA neurons appear to differ from the other DA systems described by the absence, or diminished expression, of a high-affinity DA uptake (transport) mechanism. High-affinity DA uptake serves to recapture released DA and limits the concentration and actions of DA in the synapse. Importantly, neurons equipped with a high density of DATs are especially susceptible to drugs that target this site, such as cocaine and the parkinsonian neurotoxins, 6-hydroxydopamine and MPTP. In fact, the toxicity of MPTP is dependent on the function of MAO-B, DAT, and VMAT2. Once inside the brain, MPTP is metabolized by MAO-B to the actual toxic species, MPP+, which is a substrate for DAT. Whereas DAT is responsible for transport of the toxin into DA neurons, VMAT2 aids in sequestering the toxin and diminishing its detrimental impact on mitochondria in the neuron. Thus, the DAT-to-VMAT2 ratio should be a determinant of the susceptibility of a given brain region to MPTP-induced toxicity.

Since the original studies on DA uptake, it has been known that DA is a good substrate for the NE transporter. However, not until relatively recently has the importance of this been appreciated. While in regions, such as striatum and nucleus accumbens, in which DA neurons express a high density of DAT and there is a relatively sparse NE innervation, DA uptake depends primarily on DAT. However, in regions such as the prefrontal cortex, in which DA neurons express low levels of DAT and there is a rich NE innervation, DA uptake depends strongly on the norepinephrine transporter. In fact, it has been suggested that DA can be co-released with norepinephrine by norepinephrine neurons as a result of nonspecific uptake of DA by the norepinephrine transporter and/or because DA is a precursor in the biosynthesis of NE.

Regulation

Much effort has been directed toward elucidating the regulatory mechanisms that control the function of the various DA systems because this may lead to an understanding of the vulnerability of certain DA systems to disease and may also enable drugs to be developed that target specific DA projections. Some aspects of regulation were discussed in relation to the life cycle of DA. Here, other important controls of DA neurotransmission are reviewed.

Firing Pattern

Electrophysiological studies have shown that DA cells fire in two different modes, either single spiking or burst firing, which have a marked impact on the release of DA. Phasic levels of DA are mediated primarily by bursting events at the level of the cell body and lead to a much larger DA release (150–400 nM in striatum) than when these neurons fire in a slow, irregular single spike mode associated with tonic levels (5–10 nM in striatum). Because cells are able to switch between these levels, the transition in activity is a mechanism for altering the impact of DA neurotransmission on receptive cells. Bursting activity of DA neurons is thought to represent a key component of reward circuitry, signaling a reward, or indicating to what extent a reward occurs differently than predicted. Alterations in tonic levels of DA efflux occur on a much slower time scale than changes in phasic levels and enable function on a variety of motor, cognitive, and motivational processes, which are deficient in Parkinson's disease and which can be restored by DA replacement therapy.

Autoreceptors

A receptor sensitive to the transmitter secreted from the cell on which the receptor is located is termed an

'autoreceptor.' Autoreceptors exist on most portions of DA cells, including the soma, dendrites, and nerve terminals, so that autoreceptors are responsive to both terminal and dendritic DA release. Stimulation of DA autoreceptors in the somatodendritic region slows the firing rate of DA neurons, whereas stimulation of autoreceptors located on DA nerve terminals results in an inhibition of DA synthesis and release. Somatodendritic autoreceptors may also regulate DA release and synthesis by changing impulse flow. Thus, somatodendritic and nerve terminal autoreceptors work in concert to exert feedback regulatory effects on dopaminergic transmission. Three types of autoreceptors can be defined according to their functional effects: impulse-modulating, release-modulating, and synthesis-modulating autoreceptors. In general, all DA autoreceptors can be classified as D2-like DA receptors. The terminals of all midbrain DA neurons examined to date have been found to possess release-modulating autoreceptors. This is not the case for synthesis- and impulse-modulating autoreceptors. Whereas most DA neurons located in the SN and many DA neurons in the VTA possess somatodendritic impulse-modulating and nerve terminal synthesis-modulating autoreceptors, the DA neurons that project to the prefrontal and cingulate cortices and amygdala appear to have either a greatly diminished number, or none, of these receptors. Some of the differences in responsiveness among various midbrain DA neurons may be explained, in part, by differences in their autoreceptor function. Data suggest distinct differences in the autoreceptor regulation of hypothalamic DA neurons. Thus, in contrast to the incertohypothalamic DA neurons, tuberoinfundibular DA neurons appear to lack synthesis-modulating autoreceptors. The activity of tuberoinfundibular DA neurons is regulated by circulating levels of prolactin.

Co-localized Peptides and Proteins

Another interesting intrinsic feature that may contribute to the variable susceptibility of DA neurons to toxicity is the presence of intracellular calcium-binding proteins, such as calretinin and calbindin. These proteins are preferentially expressed in DA neurons that are spared from degeneration in Parkinson's disease and MPTP-treated animals, possibly by buffering calcium overload and protecting cells against excitotoxic damage. DA is co-localized in some neurons with neuropeptides and nontransmitter proteins. Certain midbrain DA neurons contain the peptide cholecystokinin, whereas another subpopulation of mesencephalic DA neurons contains the peptide neurotensin, and a third group of DA neurons in the VTA contains cholecystokinin, neurotensin, and DA. Similarly, DA is co-localized in certain mesencephalic neurons with nontransmitter proteins, including acetylcholinesterase, vesicular glutamate transporter, protein *O*-carboxymethyltransferase, cytochrome P450 reductase, and a vitamin D-dependent calcium-binding protein. The functional significance of such co-localization is not clear. DA neurons in which neuropeptides are co-localized may be regulated by peptide autoreceptors in a fashion analogous to DA autoreceptors. Furthermore, because it appears that release mechanisms for DA and co-stored peptides may be dissociated under certain conditions (e.g., dependent on the firing pattern and firing frequency of the neuron), co-localized peptides may be part of a cascade of regulatory features that exist in dopaminergic neurons. For example, nerve terminal autoreceptors in the prefrontal cortex have been shown to exert reciprocal effects on DA and neurotensin release. Stimulation of DA autoreceptors diminishes DA release and enhances neurotensin release, whereas blockade of DA receptors augments DA release and diminishes neurotensin release. Such differential release of DA and neurotensin could conceivably allow the prefrontal cortex DA neurons to differentially modulate the physiological activity of cortical postsynaptic follower cells.

It appears likely that co-localization of peptides or nontransmitter proteins and DA will distinguish certain subpopulations of DA neurons. For example, cholecystokinin–DA co-localized neurons of the VTA project to the caudal, but not rostral, nucleus accumbens. Such distinctions may have important implications for the regionally specific function of DA in psychiatric and neurological disorders and also for the response of specific DA systems in these pathological conditions to pharmacological treatment.

Extrinsic

The different DA systems also vary in the nature of their interaction with other neurotransmitter systems that impinge on the DA neurons. At the cell body level, afferent inputs into the VTA and SN have been most extensively investigated. Many studies have suggested that these cells can be differentially modulated by afferent inputs. DA neurons are also regulated by neurotransmitter interactions that occur at the level of the nerve terminals. However, a detailed description of such interactions is outside the scope this article.

Postsynaptic Signaling

There are known to be at least six different forms of the DA receptor. The D_1 class of DA receptor has been divided into D_1 and D_5 receptor subtypes, and

the D_2 class comprises D_{2short}, D_{2long}, D_3, and D_4 receptor subtypes. The pharmacological profile and regional distribution in brain of each subtype are different, with the exception of D_{2short} and D_{2long}, which appear to be identical in these respects. Although these different receptors are reviewed elsewhere, certain interesting or novel aspects of DA signaling are discussed here.

Neuromodulation

It is now generally accepted that DA is not a classical, fast ionotropic neurotransmitter like glutamate acting at α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) or *N*-methyl-D-aspartate (NMDA) receptors or GABA acting at GABA-A receptors that elicit excitatory or inhibitory postsynaptic potentials that last a few milliseconds. All the DA receptors are G-protein-coupled receptors that are slow, metabotropic receptors with response times exceeding 100 ms and function to modulate other receptor systems and/or ion channels. Thus, DA is thought of as a neuromodulator and cannot be classified as either an excitatory or an inhibitory neurotransmitter, which may explain why DA does not yield identical effects under all experimental conditions.

Volume Transmission

Neurons may operate via two modes of communication, the synaptic mode and the mode of communication involving short- to long-distance diffusion of chemical signals in the extracellular fluid pathways (volume transmission). The process of volume transmission has been defined as a functionally significant association of release and receptor sites via extrasynaptic diffusion. It is now clear that in certain regions DA functions by volume transmission. The retina provides the clearest example of DA-mediated volume transmission. In the SN, DA receptors are located extrasynaptically, indicating that volume transmission is the predominant form of intercellular communication mediated by DA in this region. In the striatum, volume transmission appears not to play a major role in DA transmission since most DA terminals establish synaptic contacts, and the high density of DA terminals express a high density of DAT. Volume transmission may be more important in maintaining dopaminergic signaling in the striatum after partial loss of nigrostriatal DA neurons, as occurs in Parkinson's disease. With the degeneration of dopaminergic terminals and consequent loss of DAT, DA that is released from the remaining neurons may diffuse in the striatal extracellular fluid to reach supersensitive high-affinity DA receptors.

Somatodendritic Release

There is good evidence that DA can be synthesized and released from dendrites and cell bodies in the SN. The release is vesicular and mediated by exocytosis, as it is in axon terminals. Somatodendritic release is calcium and depolarization dependent in a variety of paradigms. In dendrites, DA appears to be stored both in classical vesicles and in saccules of smooth endoplasmic reticulum, consistent with the expression of VMAT2 in both of these organelles. The function of DA released in the SN in this manner is probably to interact with autoreceptors to control the firing rate of DA neurons, in addition to signaling via DA receptors located on non-DA neurons in the region.

Long-Term Potentiation and Long-Term Depression

DA has a critical role in two distinct forms of plasticity, called long-term depression (LTD) and long-term potentiation (LTP). In regions such as the striatum and prefrontal cortex where DA modulates glutamate signaling, combined stimulation of both glutamate and DA receptors is able to induce either enhancement or weakening of synaptic transmission, depending on the parameters of the stimulation and the receptor subtypes that are activated. These mechanisms may lead to more permanent changes, such as the construction of new connections (synapses). Such plastic changes are thought to have an important role in learning and memory (hippocampus and prefrontal cortex) and acquisition of complex motor skills (striatum).

Pyramidal Cell Dendritic Spines

DA also plays a critical role in the modulation of pyramidal cell dendritic spine synapses and spine density. Several studies have demonstrated that the loss of DA in the striatum results in decreased dendritic length and spine density. Employing multiphoton imaging, selective elimination of glutamatergic synapses on striatopallidal neurons has been observed in Parkinson disease models. It has clearly been shown that DA depletion leads to a rapid and profound loss of spines and glutamatergic synapses on striatopallidal medium spiny neurons but not on neighboring striatonigral medium spiny neurons. This loss of connectivity is triggered by an interesting mechanism – dysregulation of intraspine Cav1.3 L-type Ca^{2+} channels. The disconnection of striatopallidal neurons from motor command structures is likely to be a key step in the emergence of pathological activity that is responsible for symptoms in Parkinson's disease.

DA modulation of pyramidal cell dendritic spines and synapses has been extended to the prefrontal cortex, where work has shown that selective lesion of the DA innervation of the prefrontal cortex caused a decrease in dendritic spine density in pyramidal cells in the prefrontal cortex. It is noteworthy that atypical antipsychotic drugs which cause a large increase in synaptic DA reverse the DA depletion-induced dendritic spine loss in prefrontal cortical pyramidal neurons. In schizophrenia, the prefrontal cortex DA innervation is compromised and postmortem studies have reported decreased dendritic spine density. Taken together, these data suggest that disruption of dopaminergic signaling may be the proximate cause of dendritic remodeling in the prefrontal cortex in schizophrenia. Excessive DA input may have the opposite effect because chronic cocaine or amphetamine use has been shown to increase spine density in nucleus accumbens and prefrontal cortex and increase synapse number in the prefrontal cortex.

See also: Cotransmission; Dopamine Control of Arousal; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopamine Neurons: Reward and Uncertainty; Dopamine: Cellular Actions; Dopaminergic Agonists and L-DOPA.

Further Reading

Akil M, Kolachana BS, Rothmond DA, et al. (2003) Catechol-O-methyltransferase genotype and dopamine regulation in the human brain. *Journal of Neuroscience* 23: 2008–2013.

Cooper JR, Bloom FE, and Roth RH (2003) *The Biochemical Basis of Neuropharmacology,* 8th edn. New York: Oxford University Press.

Day M, Wang Z, Ding J, et al. (2006) Selective elimination of glutamatergic synapses on striatopallidal neurons in Parkinson disease models. *Nature Neuroscience* 9: 251–259.

Dunnett SB, Bentivoglio M, Bjorklund A, and Hokfelt T (2005) *Handbook of Chemical Neuroanatomy,* vol. 21. Elsevier: Amsterdam.

Eisenhofer G, Kopin IJ, and Goldstein DS (2004) Catecholamine metabolism: A contemporary view with implications for physiology and medicine. *Pharmacological Reviews* 56: 331–349.

Elsworth JD and Roth RH (1997) Dopamine autoreceptor pharmacology and function: Recent insights. In: Neve KA and Neve RL (eds.) *The Dopamine Receptors*, pp. 223–265. Humana: Totowa, NJ.

Grace AA (2002) Dopamine. In: Davis KL, Charney D, Coyle JT, and Nemeroff C (eds.) *Neuropsychopharmacology: The Fifth Generation of Progress*, pp. 120–132. Philadelphia: Lippincott Williams & Wilkins.

Jay TM (2003) Dopamine: A potential substrate for synaptic plasticity and memory mechanisms. *Progress in Neurobiology* 69: 375–390.

Jentsch JD, Roth RH, and Taylor JR (2000) Role for dopamine in the behavioral functions of the prefrontal corticostriatal system: Implications for mental disorders and psychotropic drug action. *Progress in Brain Research* 126: 433–453.

Joel D and Weiner I (2000) The connections of the dopaminergic system with the striatum in rats and primates: An analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience* 96: 451–474.

Lookingland KJ and Moore KE (2005) Functional neuroanatomy of hypothalamic dopaminergic neuroendocrine systems. *Handbook of Chemical Neuroanatomy* 21: 435–523.

Miller GW, Gainetdinov RR, Levey AI, and Caron MG (1999) Dopamine transporters and neuronal injury. *Trends in Pharmacological Sciences* 20: 424–429.

Rice ME (2000) Distinct regional differences in dopamine-mediated volume transmission. *Progress in Brain Research* 125: 277–290.

Seamans JK and Yang CR (2004) The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology* 74: 1–58.

Tzschentke TM (2001) Pharmacology and behavioral pharmacology of the mesocortical dopamine system. *Progress in Neurobiology* 63: 241–320.

Dopamine – CNS Pathways and Neurophysiology

A A Grace, D J Lodge, and D M Buffalari, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The dopamine (DA) systems of the brain have been the topic of intensive investigation since the first report of this neurochemical as an independent neurotransmitter in the brain nearly 50 years ago. The DA systems have drawn the attention of investigators due to their known role in a variety of neurological and psychiatric disorders, including Parkinson's disease, schizophrenia, and drug addiction. In this article, we review the principal electrophysiological characteristics of DA neurons, as well as the intrinsic and extrinsic regulation of their activity and firing patterns.

Anatomy of DA Neuron Projections

As with most monoamine systems, individual DA neurons are believed to exhibit dense collateralizations, with single neurons giving rise to 500 000 to 1 000 000 synaptic terminals. However, unlike other monoamine neurons, midbrain DA neurons have discrete, topographic projections to their target regions. Three major DA neuronal systems have been identified that project to forebrain regions: (1) the nigrostriatal DA neuron projection, which arises from the substantia nigra (SN) and projects to the dorsal striatum; this system is involved in motor control and is known to degenerate in Parkinson's disease; (2) the mesolimbic DA projection, which arises from the ventral tegmental area (VTA) and projects to limbic structures such as the ventral striatum, nucleus accumbens, amygdala, and other regions involved in control of affect and likely play a role in schizophrenia and drug abuse; and (3) the mesocortical DA system, which also arises from the VTA and projects primarily to frontal cortex; this projection system is believed to be involved in higher processes related to the control of executive function.

Electrophysiology

Identification

Since the ventral mesencephalon is a heterogeneous region, the first challenge in examining the neurophysiology of DA systems is to ensure that the neurons recorded are indeed DAergic in nature. The first extracellular recordings from putative DA neurons were performed in the 1970s. Through an examination of the neurophysiological characteristics of DA neurons and subsequent histochemical verification, it was determined that DA neurons can be identified based solely on their electrophysiological characteristics. As such, one of the most distinct features of DA neurons is the broad extracellular spike waveform they produce (Figure 1). Such waveforms are biphasic (+/−), with an inflection in the rising phase (representative of the initial segment spike). They have a pronounced negative phase to the action potential, as well as a long time course (>2 ms). More specifically, action potentials consist of a short-duration, smaller-amplitude initial segment spike, followed by a larger and longer duration somatodendritic spike. Much of this unique waveform is a consequence of the active membrane properties of the neuron, causing a train of action potentials to be variable from one waveform to the next, and readily distinguishable by the long-duration negative phase, although such differences can be obscured if improper filter settings are employed.

Passive Membrane Properties

Intracellular recordings from identified DA neurons have provided important insights into the mechanisms underlying action potential generation in these neurons. Interestingly, DA neurons recorded from mesencephalic slices after severing of afferent processes continue to exhibit spontaneous activity that is derived from the active membrane properties of the neuron. One factor associated with spike generation in DA neurons is a large-amplitude (10–15 mV), pacemaker-like slow depolarizing potential that brings the membrane potential from rest (*c.* –55 mV) to the comparatively high action potential threshold (*c.* –40 mV). This slow depolarization is mediated by both sodium and calcium conductances, as it can be blocked by tetrodotoxin (TTX) or cobalt, both of which also inhibit spontaneous spike firing. The resultant spike is followed by a negative shift in membrane potential accompanied by an inhibition of spike firing, or an afterhyperpolarization. This afterhyperpolarization is voltage dependent and mediated by a calcium-activated potassium conductance, as it can be blocked by tetraethylammonium (TEA) and attenuated by the calcium chelator ethylene glycol tetraacetic acid (EGTA). Furthermore, it is believed that the rebound from this hyperpolarization triggers the calcium and sodium conductances comprising the slow depolarization. It is this cycle that has

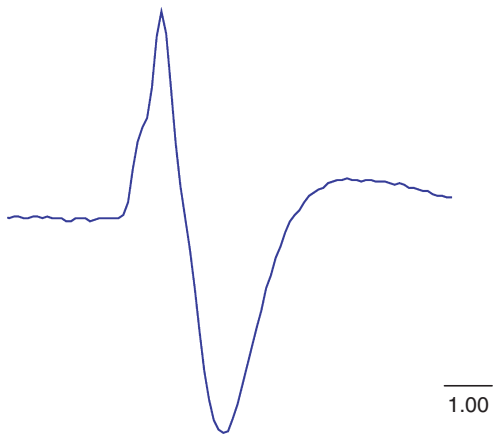

Figure 1 Electrophysiological trace of an action potential recorded extracellularly from an identified VTA DA neuron. The unique characteristics are the biphasic (+/−) waveform with an inflection in the rising phase, pronounced negative phase, as well as a long time course (>2 ms). Scale bar = 1 ms.

been proposed to underlie spontaneous pacemaker firing of midbrain DA neurons observed *in vivo* and *in vitro*.

DA Neuron Activity States

DA neurons of the brain typically display four different states of activity. The first is an inactive, hyperpolarized state suggested to result from GABAergic inhibition (see the section titled 'Intrinsic regulation of DA neuron firing'). These spontaneously inactive neurons have a higher resting membrane potential (*c.* –65 to –75 mV) and can be activated by direct glutamate application or administration of haloperidol (IV). A separate inactive state has been demonstrated following chronic neuroleptic treatment. This state has been termed 'depolarization block,' caused by hyperexcitation and characterized by a depolarized membrane potential. In contrast to hyperpolarized inactive DA neurons, glutamate application will not activate DA neurons following depolarization block, and the restoration of spontaneous activity is achieved by the direct application of the 'inhibitory' transmitter, γ-aminobutyric acid (GABA).

Spontaneously active DA neurons, recorded *in vivo*, typically display two main firing patterns: a slow, irregular single-spike firing pattern and a faster bursting mode. It is the alternation between these two states that is thought to result in different levels of DA release in terminal regions, with irregular spike firing regulating extracellular, 'tonic' DA levels and burst firing leading to transient synaptic 'phasic' DA levels. Accordingly, much research has focused on the regulation of these two types of firing patterns and the mechanisms that lead to the transition from one pattern to the other.

Irregular, Single-Spike Firing

DA neurons recorded *in vivo* commonly display firing rates between 2 and 8 Hz, with an average of approximately 4 Hz, and seldom fire at rates lower than 1 Hz or exceeding 10 Hz. While in single-spike firing mode, DA neuron interspike intervals (ISIs) are fairly long (200–250 ms) and follow a normal distribution. It is thought that the slow single-spike firing pattern of DA neurons is driven by the intrinsic pacemaker depolarization, as well as the prolonged afterhyperpolarization discussed above. Interestingly, although DA neurons recorded *in vitro* display a highly regular pacemaker firing pattern, DA neurons recorded *in vivo* possess a significantly more irregular pattern of activity. This pattern of activity is likely to be associated with variations in calcium entry into the neuron during spontaneous spike firing, since the intracellular injection of EGTA into an irregularly firing neuron causes a transition to a pacemaker pattern.

Burst Firing

The firing rates of DA neurons fall into a fairly limited range, usually 2–8 Hz, which consequently might limit the flexibility of DA neurons to release differential amounts of DA in terminal regions. However, this is overcome by a change in firing pattern from single-spike firing to burst firing. Burst firing in DA neurons induces increases in DA release that are 2–3 times that of increases in tonic firing rates of the same magnitude. Moreover, DA released during burst firing has been demonstrated to be localized to the synaptic cleft and is considered the functionally relevant signal sent to postsynaptic sites. These bursts are comprised of a series of spikes that display accommodation, with decreasing amplitude and increasing action potential duration (Figure 2). DA neuron bursts are typically comprised of two to eight spikes which ride on top of a membrane depolarization. While the first two of these spikes are separated by an ISI of 80 ms or less, following spikes tends to have increasing ISIs of up to 160 ms. As such, DA neurons firing in burst tend to display a bimodal ISI distribution (Figure 3). Bursts of DA neurons tend to be different from bursting described for neurons in other regions in three primary ways: (1) they cannot be elicited by short, depolarizing current pulses; (2) they have a much larger ISI than is typical for bursting neurons; and (3) they can be blocked by intracellular calcium chelators such as EGTA.

The transition from irregular single-spike firing to burst firing is dependent on an excitatory amino acid, since activation of glutamatergic afferents or direct microiontophoretic application of glutamate induces burst firing in DA neurons *in vivo*. Furthermore, direct application of competitive *N*-methyl-D-aspartate

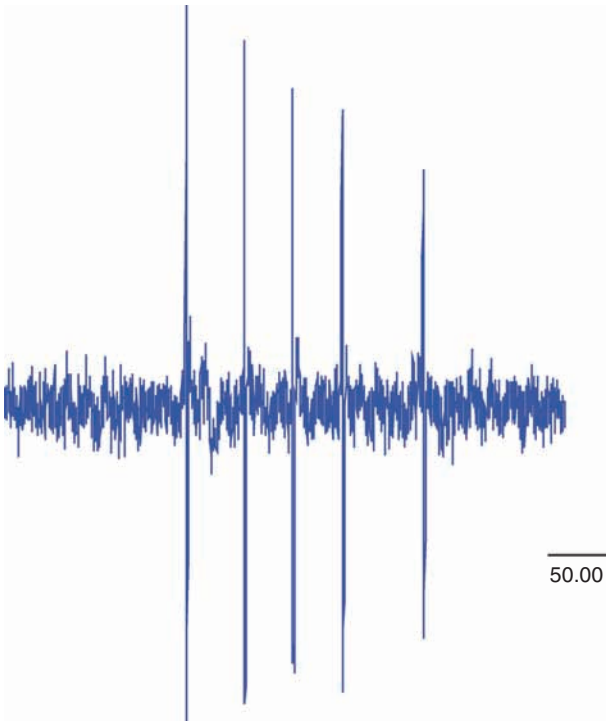

Figure 2 Extracellular recording of a single burst of action potentials recorded from a burst firing DA neuron. These bursts are comprised of a series of spikes that display accommodation, with decreasing amplitude and increasing action potential duration. Scale bar = 50 ms.

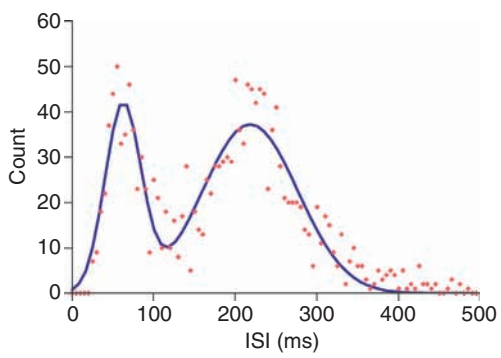

Figure 3 An ISI histogram from a burst-firing DA neuron depicting the characteristic bimodal distribution. Diamonds represent individual bin values, and the line is the fitted average curve. The early peak is the ISI within the burst, and the later peak is the time between bursts.

(NMDA) receptor antagonists potently blocks spontaneous burst firing. Taken as a whole, there is a significant literature demonstrating a primary role for NMDA receptor-mediated glutamatergic transmission in the regulation of DA neuron burst firing. On the other hand, glutamate alone is not sufficient to mediate burst firing. Thus, DA neurons recorded from mesencephalic slices obtained from adult rats, in which afferent input has been severed, display a regular pacemaker firing pattern and cannot be made to fire in bursts in response to glutamate agonist administration or alterations in membrane potential alone.

Based on the role of Ca^{2+} and K^{+} conductances in the generation of burst firing *in vivo*, the first reported instance in which burst firing was observed *in vitro* was in the presence of apamin, a toxin derived from bee venom and a potent and irreversible inhibitor of small conductance (SK) Ca^{2+}-activated K^{+} channels (gKCa). Interestingly, in studies that include apamin in the perfusate, DA neuron activity more closely resembles that seen *in vivo* with a decreased regularity of firing and the potential to display burst firing in response to depolarizing current injections and NMDA application. Taken as a whole, these data suggest that DA neuron burst firing is regulated by at least two mechanisms, the first being a permissive inactivation of the SK Ca^{2+}-activated K^{+} channels that by itself does not necessarily induce burst firing but is required for burst firing to occur, and the second being an afferent-derived glutamatergic drive. Interestingly, these mechanisms are not necessarily mutually exclusive, since recent reports have demonstrated a Ca^{2+}-mediated feedback loop whereby SK channels can regulate NMDA-dependant Ca^{2+} influx within individual dendritic spines. Afferent inputs that may be responsible for altering membrane conductances and allowing for the shift from single-spike, irregular firing mode to bursting mode are reviewed below.

Electrical Coupling

Electrical coupling between DA neurons was suggested following the observation that intracellular injection of Lucifer yellow dye into a single DA neuron can result in the labeling of adjacent DA neurons. Further evidence for electrical coupling stems from the fact that some DA neurons located in close proximity have been shown to fire in nearly synchronous firing patterns (with ISIs on the order of 2–3 ms). Administration of D2 receptor blockers such as haloperidol is reported to increase the incidence of simultaneous spike discharge among DA neurons. Furthermore, near-synchronous firing seems to occur more frequently during burst firing patterns. More recently, direct evidence for DA neuron coupling has been observed by simultaneous patch-clamp recordings from DA neuron pairs. This study demonstrated that depolarization of a DA neuron can result in an increased firing frequency of a neighboring electrically coupled DA neuron. In summary, there is significant evidence for electrical coupling between adjacent DA neurons that may result in synchronous firing under appropriate conditions.

Afferent Input to Midbrain DA Neurons

Given the importance of the VTA in encoding reward prediction or indicating incentive salience, it is perhaps not surprising that this region is a site of extensive integration of afferent information. As such, early anatomical studies demonstrated a widespread convergent input to the VTA. This afferent input has since been confirmed and extended using more selective and sensitive techniques. Thus, afferent inputs to the VTA from a large number of regions have been described, which include the prefrontal cortex, nucleus accumbens, ventral pallidum (VP), lateral preoptic area, lateral hypothalamus, and brain stem regions including the pedunculopontine tegmentum (PPTg) and laterodorsal tegmentum (LDTg), among other areas. In addition, the VTA is a heterogeneous structure containing GABAergic interneurons and dendritodendritic synaptic appositions. As such, the output of the VTA is a function of its intrinsic organization as well as regulation by extrinsic factors.

Intrinsic Regulation of DA Neuron Firing

Autoreceptor-Mediated Inhibition

DA neurons display an autoregulatory mechanism important in the determination of their activity levels. DA neuron dendrites in the VTA contain tyrosine hydroxylase (the rate-limiting enzyme in DA synthesis), contain D2 autoreceptors located near the cell body, and DA has been shown to be released in a calcium-dependent manner from the somatodendritic region of DA neurons. Furthermore, investigations using a number of different methods have shown that somatodendritic DA release in the VTA is altered under different behavioral and pharmacological conditions. This local DA release provides a potent autoinhibitory role in the VTA by decreasing DA cell firing rate via activation of D2 receptors. Iontophoresis of DA onto DA neurons causes inhibition of neural activity. Moreover, low doses of DA agonists administered systemically will preferentially activate these highly sensitive somatodendritic D2 autoreceptors. Moreover, application of D2 antagonists enhances DA neuronal firing even following transection of feedback pathways from postsynaptic targets, suggesting a tonic level of autoreceptor activation of DA neurons. D2 antagonists also block DA-induced suppression of DA neuron firing. These data therefore demonstrate a potent autoinhibitory role for DA within the nigrostriatal and mesolimbic DA system.

GABA

The midbrain DA neuron regions are heterogeneous structures containing not only DAergic and non-DAergic projection neurons, but also a number of GABAergic interneurons. Investigations into the effects of these intrinsic GABAergic neurons on DA neuron firing patterns demonstrated the somewhat surprising observation that systemic administration of a GABA agonist (muscimol) actually increases DA neuron firing rate *in vivo*. Further study determined that this 'paradoxical excitation' was correlated with a decrease in firing rates of non-DA neurons. Moreover, DA neurons recorded intracellularly *in vivo* have been shown to be under a constant bombardment with large-amplitude GABAergic inhibitory postsynaptic potentials (IPSPs). These data demonstrate that the increase in DA neuron firing induced by systemic muscimol is secondary to the inhibition of tonically active GABAergic interneurons, and thus demonstrates a tonic role for local inhibitory regulation of DA neuron firing.

Afferent Connectivity

GABAergic Inputs

As stated above, intracellular recordings from identified DA neurons of the rat midbrain *in vivo* have demonstrated that these neurons are constantly bombarded by GABAergic IPSPs. Indeed, it has been suggested that up to 50% of the midbrain DA neurons are quiescent due to GABAergic-mediated hyperpolarization. Moreover, these IPSPs are not observed *in vitro*, suggesting a prominent and tonically active GABAergic tone to this region in addition to that supplied by intrinsic interneurons. One region providing such input is the pallidal complex; a GABAergic structure that receives GABAergic inputs from the striatum and nucleus accumbens (among other regions) displays relatively high rates of spontaneous activity and exerts a powerful tonic inhibitory influence over efferent structures. Since anatomical studies have demonstrated a substantial projection from the globus pallidus to the SN DA neurons and from the VP to the VTA, the pallidum is well positioned to influence DA neural activity. Investigations have demonstrated that inactivation of the VP results in a dramatic increase in the number of spontaneously active DA neurons observed per electrode track (a standard measure of DA neuron population activity) that is correlated with a significant increase in extracellular 'tonic' DA release in the nucleus accumbens. Taken as a whole, these data demonstrate a prominent tonic inhibition of DA neuron firing by the VP.

Glutamatergic Inputs

As mentioned above, the transition from irregular single-spike firing to a burst firing pattern has been

largely attributed to increased glutamatergic input. As such, direct microiontophoretic application of glutamate has been demonstrated to increase burst firing in DAergic neurons. This effect has largely been attributed to actions at the ionotropic NMDA receptor, since it is blocked by MK-801, a highly potent and selective noncompetitive NMDA receptor antagonist. Furthermore, it has been suggested that glutamatergic transmission is required for burst firing *in vivo* since both glutamate-induced and natural burst events can be blocked by inactivation of glutamatergic afferents or the administration of glutamatergic or NMDA receptor antagonists. There is significant literature suggesting a primary role for NMDA receptor-mediated glutamatergic transmission in the regulation of DA neuron burst firing. Given that DA neuron burst firing is considered to be the functionally relevant signal sent to postsynaptic sites to encode reward prediction or indicate incentive salience, glutamatergic inputs to the VTA have garnished significant attention. Such inputs have been demonstrated to arise from the prefrontal cortex, PPTg, and lateral preoptic-rostral hypothalamic area.

The prefrontal cortex is a region associated with planning complex cognitive behaviors such as executive function and expression of appropriate social behavior. A significant literature has been accumulated implicating the medial prefrontal cortex (mPFC) in the regulation of DA neuron burst firing. Thus, chemical stimulation of the mPFC increases burst firing in a proportion of DA neurons, whereas inactivation of this area has been found to reduce spontaneous burst activity. Furthermore, electrical stimulation of the mPFC induces events that resemble natural burst firing in a subset of DA neurons. It is important to note that recent anatomical studies have demonstrated that terminals from the mPFC synapse exclusively onto mesocortical DA neurons, suggesting that glutamatergic inputs from the mPFC provide direct regulatory control over an important subset of VTA DA neurons.

Another major excitatory input to the VTA arises from the PPTg, a glutamatergic/cholinergic region driven by limbic afferents, including the prefrontal cortex and extended amygdala. The PPTg is activated by auditory, visual, and somatosensory stimuli, and has been demonstrated to directly regulate burst firing of DAergic neurons. Thus, electrical stimulation of the PPTg produces time-locked bursts of action potentials in DA neurons that resemble natural bursts. Furthermore, chemical activation of the PPTg induces a significant increase in DA neuron burst firing. Taken as a whole, the PPTg serves as a site of convergence whereby a variety of sensory inputs can modulate DA neuron burst firing.

Association between GABA and Glutamatergic Inputs

Given that the midbrain DA system is a site of extensive integration of afferent inputs, it is perhaps not surprising that in addition to providing independent control over distinct DA neuron activity states, afferent pathways also interact to modulate DA neuron responsivity. For example, it has been suggested that activation of excitatory inputs to the VTA results in burst firing only in DA neurons that are already spontaneously active. Thus, DA neurons that are inactive, presumably due to GABA-mediated hyperpolarization, are unresponsive to activation of glutamatergic afferents, likely attributable to Mg^{2+} blockade of the NMDA receptor-channel complex. This suggests the presence of a unique interdependence of the GABAergic and glutamatergic VTA inputs, that is, only neurons that are not under GABA-mediated hyperpolarization are capable of entering a burst firing mode in response to a glutamatergic input. Indeed, we have previously reported that the hippocampus can gate phasic DA transmission by regulating the number of spontaneously active DA neurons that can be further modulated by excitatory inputs to induce a graded phasic response. Hence, although independent inputs to the VTA can regulate distinct DA neuron activity states, these afferents also interact to regulate DA neuron responsivity.

Regulatory Inputs

Although there is significant literature demonstrating a primary role for NMDA receptor-mediated glutamatergic transmission in the regulation of DA neuron burst firing, glutamate alone is not sufficient to induce burst firing in DA neurons recorded from mesencephalic slices *in vitro* (in which afferent processes have been severed). As reviewed above, in contrast to what is seen *in vivo*, DA neurons recorded *in vitro* from mesencephalic slices do not exhibit burst firing under baseline, depolarization, or glutamate-stimulated conditions. Thus, although glutamatergic transmission is critical, it is not the sole determinant of the generation of burst firing *in vivo*.

Therefore, we have recently demonstrated that a critical factor required to allow spontaneous and glutamate-driven burst firing *in vivo* depends on tonic drive from the LDTg. After baclofen/muscimol inactivation of this region, DA neuron discharge patterns more closely resemble those observed *in vitro*; as pacemakers that lack spontaneous burst firing. Interestingly, neither activation of the PPTg nor direct glutamate iontophoresis *in vivo* are capable of inducing burst firing following LDTg inactivation. Taken

as a whole, these data demonstrate that a functional input from the LDTg is essential for DA neuron burst firing *in vivo*.

Neuromodulatory Inputs

As discussed above, the predominant excitatory and inhibitory neurotransmitters, glutamate and GABA, respectively, potently and dramatically regulate DA neuron activity states. It is important to recognize, however, that there are also significant inputs to the VTA that contain biogenic amines, neuropeptides, or nonclassical transmitters (i.e., nitric oxide or arachadonic acid derivatives). Moreover, a number of these have been shown to modulate DA neuron firing patterns. These include cholinergic, adrenergic, serotonergic, and peptidergic inputs arising from a number of regions. Given the large number of neuromodulators present in the VTA, it would not be possible to discuss them all in this review; as such, we focus on the two that have been at the center of recent significant advances, namely, the cannabinoids and the orexins.

Cannabinoids

Research into the neurobiological actions of cannabis has demonstrated that the active ingredient of marijuana, Δ9-tetrahydrocannabinol (Δ9-THC), increases extracellular DA efflux in the accumbens. In addition, further research demonstrated that systemic administration of Δ9-THC (and other cannabinoid agonists) increases the firing rate and burst firing of identified VTA DA neurons. Interestingly, in contrast to the mechanism of action of classical neurotransmitters, endocannabinoids appear to act in a retrograde manner. As such, it is believed that endocannabinoids are released from postsynaptic neurons due to membrane depolarization, and that they migrate to adjacent presynaptic targets where they can modulate neurotransmitter release by activating presynaptic CB1 receptors. This retrograde neurotransmitter function of endocannabinoids has been identified recently in the VTA. Therefore, the activity-dependent release of endocannabinoids in the VTA has been suggested to act as a regulatory feedback mechanism to modulate presynaptic glutamatergic and GABAergic release. As a result, it is likely that endogenous cannabinoids act to provide an additional level of autoregulatory feedback in the VTA.

Orexin

Orexins are neuropeptides that are found in neurons of the lateral hypothalamus associated with arousal, feeding, and appetitive behaviors. Due to the significant orexigenic input from the LH to DA neurons of the VTA, it has been suggested that orexin modulates DAergic neurotransmission. Indeed, the direct application of orexin increases the firing rate of DA neurons recorded *in vitro*, suggesting a role for the orexin system in the modulation of DA neuron activity. More recent data have expanded previous observations and demonstrated a unique role for these neuropeptides in the regulation of DA neuronal activity. Specifically, the application of orexin *in vitro* has been shown to augment NMDA neurotransmission by inducing the insertion of NMDA receptors into the membrane of DA neurons. Such findings provide evidence for a role of the endogenous orexin system in the regulation of DA neuronal activity and plasticity.

Conclusion

The functional output of a brain region is dependent on both the intrinsic properties of the neurons and its regulation by extrinsic inputs. Here we illustrate that VTA DA neurons display distinct patterns of activity and furthermore that these activity states can be dynamically regulated by intrinsic and extrinsic inputs. This ability to integrate a variety of incoming signals and dynamically regulate DA neural firing patterns therefore provides the flexibility essential for the role of the VTA in the regulation of cognitive and motivational behaviors.

See also: Addiction: Neurobiological Mechanism; Animal Models of Parkinson's Disease; Dopamine; Dopamine in Perspective; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopamine Neurons: Reward and Uncertainty; Dopamine: Cellular Actions; Dopaminergic Differentiation; Dopaminergic Agonists and L-DOPA; Drug Addiction: Behavioral Neurophysiology; Parkinsonian Syndromes; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Substance Abuse and Dependence.

Further Reading

Borgland SL, Taha SA, Sarti F, Fields HL, and Bonci A (2006) Orexin A in the VTA is critical for the induction of synaptic plasticity and behavioral sensitization to cocaine. *Neuron* 49: 589–601.

Chergui K, Charlety PJ, Akaoka H, et al. (1993) Tonic activation of NMDA receptors causes spontaneous burst discharge of rat midbrain dopamine neurons *in vivo*. *European Journal of Neuroscience* 5: 137–144.

Floresco SB, West AR, Ash B, Moore H, and Grace AA (2003) Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nature Neuroscience* 6: 968–973.

Geisler S and Zahm DS (2005) Afferents of the ventral tegmental area in the rat-anatomical substratum for integrative functions. *Journal of Comparative Neurology* 490: 270–294.

Grace AA (1991) Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: A hypothesis for the etiology of schizophrenia. *Neuroscience* 41: 1–24.

Grace AA (2002) Dopamine. In: Charney D, Coyle J, Davis K, and Nemeroff C (eds.) *Neuropsychopharmacology: The Fifth Generation of Progress*, pp. 119–132. New York: Lippincott, Williams & Wilkins, Raven Press.

Grace AA and Bunney BS (1984) The control of firing pattern in nigral dopamine neurons: Burst firing. *Journal of Neuroscience* 4: 2877–2890.

Grace AA and Bunney BS (1984) The control of firing pattern in nigral dopamine neurons: Single spike firing. *Journal of Neuroscience* 4: 2866–2876.

Hyland BI, Reynolds JN, Hay J, Perk CG, and Miller R (2002) Firing modes of midbrain dopamine cells in the freely moving rat. *Neuroscience* 114: 475–492.

Kitai ST, Shepard PD, Callaway JC, and Scroggs R (1999) Afferent modulation of dopamine neuron firing patterns. *Current Opinions in Neurobiology* 9: 690–697.

Lodge DJ and Grace AA (2006) The laterodorsal tegmentum is essential for burst firing of ventral tegmental area dopamine neurons. *Proceedings of the National Academy of Sciences of the United States of America* 103: 5167–5172.

Riegel AC and Lupica CR (2004) Independent presynaptic and postsynaptic mechanisms regulate endocannabinoid signaling at multiple synapses in the ventral tegmental area. *Journal of Neuroscience* 24: 11070–11078.

Sesack SR, Carr DB, Omelchenko N, and Pinto A (2003) Anatomical substrates for glutamate–dopamine interactions: Evidence for specificity of connections and extrasynaptic actions. *Annals of the New York Academy of Sciences* 1003: 36–52.

Dopamine Control of Arousal

P M Fuller and J Lu, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The traditional account of the central dopaminergic system includes the important role for dopamine (DA), a catecholamine neurotransmitter, in the regulation of a myriad of neurobiologic, physiologic, and pathophysiologic processes, including cognition, motivation, memory, salience detection, motor disturbances of Parkinson's disease (PD), depression, schizophrenia, and hypophyseal function. The central dopaminergic system comprises three major and well-described tracts, all of which originate in ventral mesencephalic neurons. The mesolimbic and mesocortical pathways both originate in DA neurons of the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) and project to structures in the ventral striatum, hypothalamus, nucleus accumbens and other limbic structures, and the prefrontal association cortex. These projections are implicated in mediating many motivated behaviors, addiction, and salience detection. The nigrostriatal pathway connects the substantia nigra with the striatum (caudate and putamen) and is implicated in movement and the pathogenesis of PD. In addition to the three major central dopaminergic tracts described here, several other dopaminergic systems exist, including the diencephalic A11–A15 cell groups, which include the tuberoinfundibular pathway that regulates the release of prolactin from the anterior pituitary; the retinal interplexiform cells (A16); the periglomerular cells of the olfactory bulb (A17); and the wake-active DA neurons located in the ventral periaqueductal grey (vPAG) (Figure 1).

In contrast to other monoaminergic and cholinergic systems, the central DA system has historically been ascribed only a limited role in the regulation of sleep–wakefulness and cortical arousal. Similarly, in the past, DA has been assigned only a marginal role in mediating the effect of wake-promoting therapeutics. Here, we review recent work, including our own, that has challenged this popular conception. In addition to firmly establishing an important role for DA in the regulation of both sleep–wake behavior and electrocortical arousal, this work has identified, for the first time, the neuroanatomical locus of wake-active DA neurons.

Sleep Regulation: Wake-Active Dopamine Neurons

The role of DA in the regulation of sleep–wakefulness as well as the neuroanatomical locus of wake-active DA neurons has only recently been established. As described previously, the DA neurons of origin for the mesocorticolimbic and nigrostriatal projections are distributed in several midbrain areas (designated A8–A10) that include the VTA (A10), the SN (A9, including the pars reticulata and pars compacta), and the retrorubral area (A8). Whereas the A8 and A9 cell groups form the origin of the projections to the striatum (predominantly the caudate and putamen), the A10 group projects most heavily to limbic forebrain areas, including the medial prefrontal cortex (PFC) and septal area. Because A10 (and to some extent A9) neurons respond to alerting stimuli, mesocorticolimbic DA transmission originating in these cell groups has been implicated in both behavioral and EEG arousal. Yet, surprisingly, neurotoxic lesions of the A10 cell group do not decrease behavioral wakefulness. Moreover, recording studies have revealed that the firing patterns of mesencephalic DA neurons do not correlate with overall levels of behavioral wakefulness. By contrast, all other monoaminergic cell groups exhibit robust state-dependent activation. Although these results seem to suggest a limited role for midbrain DA neurons and mesocorticolimbic DA transmission in sleep regulation, an important role for DA in the regulation of sleep–wakefulness is nevertheless indicated by several findings. First, it has been demonstrated that DA transporter (DAT) knockout mice have approximately 20% more wakefulness than control mice and are refractory to psychostimulant-induced (presumably DA-mediated) arousal. A similar phenotype (i.e., increased wakefulness) results from mutation of the *Drosophila* DAT gene. Second, administration of exogenous dopaminomimetics affects the sleep–wake state in a complex dose- and receptor-dependent manner. For example, in general, lower dopaminomimetic doses have a soporific effect, presumably mediated by presynaptic D2-like inhibitory autoreceptors, whereas higher doses (which typically promote an attendant increase in locomotor activity and suppress rapid eye movement (REM) sleep) enhance arousal, likely via postsynaptic D1-like receptors. Third, patients with PD who have extensive loss of DA neurons in the SN (A9), and less extensive loss of DA neurons in the VTA (A10), often demonstrate

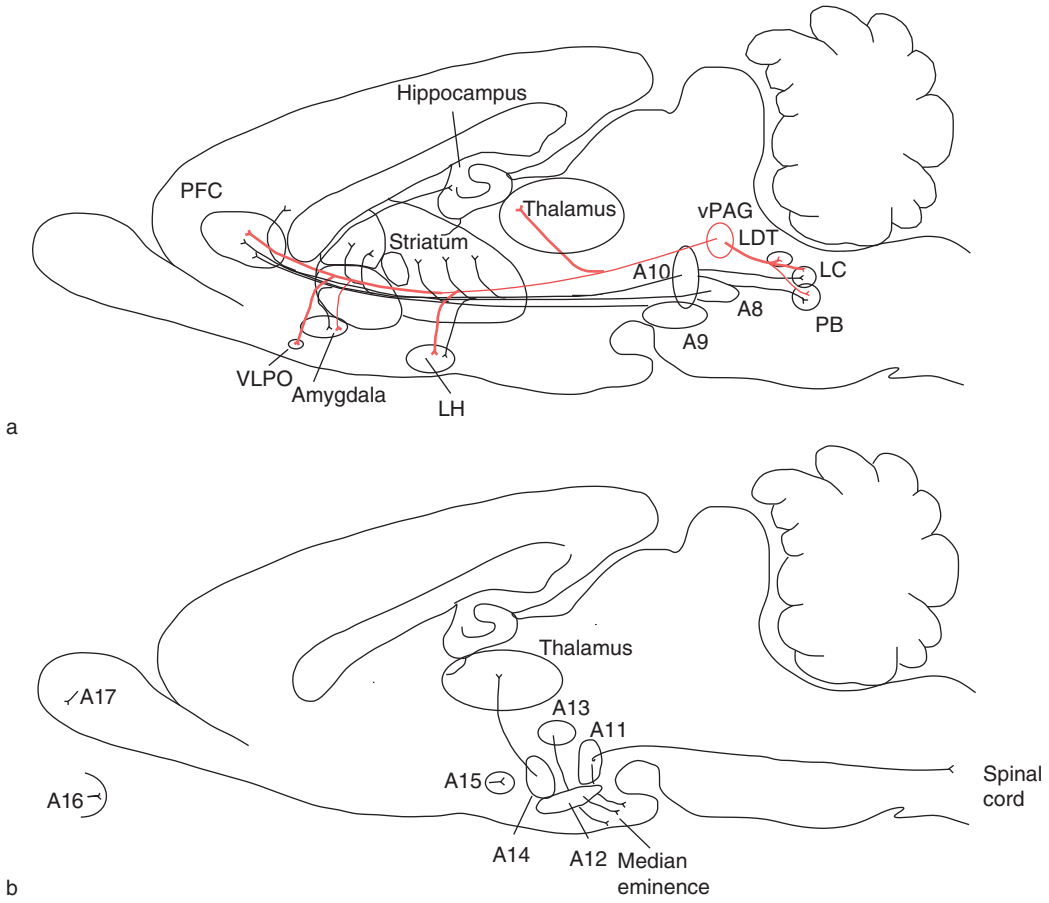

Figure 1 (a) The afferent and efferent projections of the midbrain dopaminergic (A8–A10) system, including the ventral periaqueductal area (vPAG) dopaminergic neurons. The wake-active vPAG DA neurons are reciprocally connected with several components of the sleep regulatory and arousal systems, including the ventrolateral preoptic nucleus (VLPO), the locus coeruleus (LC), the pontine laterodorsal tegmental nucleus (LDT), the lateral hypothalamic (LH) orexin neurons, the midline and intralaminar thalamus, the basal forebrain (BF) cholinergic cells, and the prefrontal cortex (PFC). (b) The projections of the diencephalic dopaminergic system, including the descending subparafascicular thalamus (A11) projection to the dorsal horn at all levels of the spinal cord.

excessive daytime sleepiness (EDS), which is made worse by D2 receptor agonist (which activates inhibitory presynaptic autoreceptors on DA neurons and therefore inhibits the firing of DA neurons). Finally, arousal and waking behaviors are associated with increased forebrain DA secretion.

Work has uncovered a previously unrecognized group of wake-active DA neurons in the vPAG that may provide the long-sought ascending dopaminergic waking influence (Figure 2). In this study, 6-hydroxydopamine-induced lesions of vPAG DA neurons (sparing intermingled dorsal raphe serotonergic neurons, the VTA, and SNc) resulted in an increase in total daily sleep (non-REM and REM) by approximately 20%. Importantly, the magnitude of this increase in sleep seen following vPAG DA lesions is significantly larger than that produced by lesions of other monoaminergic and cholinergic cell groups thought to be important in arousal, including the locus coeruleus (LC), lateral dorsal tegmentum (LDT), dorsal raphe (DRN), lateral hypothalamic (LH) orexin neurons, basal forebrain cholinergic neurons, and the histaminergic tuberomammillary nucleus (TMN). Interestingly, DA neurons in the vPAG have been identified previously in humans and rats. Yet because the vPAG DA neurons share many efferent projections with A10 DA neurons, such as the ventral striatum and PFC, they have long been considered a rostral extension of the A10 DA group. Anatomical and physiological characteristics of the vPAG DA neurons suggest, however, that these cells constitute a functionally distinct neuronal population. For example, unlike VTA DA neurons, vPAG DA neurons project heavily to the central and extended nucleus of the amygdala as well as to the ventrolateral preoptic nucleus (VLPO), a critical

Figure 2 Photomicrograph showing the distribution of DA neurons in the ventral periaqueductal area (vPAG). Aq, aqueduct.

sleep-promoting center. These vPAG DA neurons also project heavily to (and receive reciprocal innervations from) most of the major components of the sleep–wake and arousal systems, including the VLPO, basal forebrain, LH orexin cells, LC, the LDT cholinergic cells, and the midline and intralaminar thalamus. Also, as mentioned previously, unlike the vPAG DA neurons, VTA DA neurons do not demonstrate state-dependent changes in firing. Taken together, these findings indicate that vPAG DA neurons contain a functionally distinct group of DA neurons which likely form the origin of a potent dopaminergic arousal system. Thus, for example, loss/dysfunction of vPAG DA neurons may underlie EDS in PD. Although it remains unclear how DA influences sleep regulation and arousal, it is likely through both inhibition of sleep active neurons in the VLPO and activation of the basal forebrain and monoaminergic systems (i.e., as part of the extrathalamic cortical arousal system).

A role for DA in REM sleep (also called 'paradoxical sleep' and 'active sleep') control has also been indicated by several studies. REM sleep is a behavioral state characterized by activation of the cortical and hippocampal EEG, REM, and muscle atonia, although the neurobiologic significance of this behavioral state remains unresolved. Nevertheless, electrophysiological work has revealed increased bursting activity of VTA (A10) DA neurons during REM sleep. DA levels are also elevated in the PFC during REM sleep. It remains unclear, however, if the A10 bursting is linked to the increase in DA in the PFC. Additional work suggests that DA may also directly regulate the activity of REM active neurons located in the peri-LC alpha. Specifically, Sakai and colleagues have shown that microdialysis of DA in the peri-LC alpha inhibits REM sleep in cats, likely via adrenergic α_2 receptors. Finally, numerous cholinergic receptors have been found in the midbrain DA neurons, indicating that PPT cholinergic neurons may influence the midbrain DA system, particularly during wake and REM sleep states. Collectively, these findings provide further evidence linking changes in the activity of DA neurons with changes in behavioral state.

Dopaminergic Disorders and Sleep: PD

Patients with PD suffer a progressive loss of DA neurons, largely in the SNc and VTA, leading to a marked reduction in DA content in the basal ganglia and ultimately manifesting in motor abnormalities that include akinesia, rigidity, resting tremor, and postural instability. In addition to motor disturbances, PD patients often complain of sleep disturbances ranging from sleep fragmentation to abnormal motor activity and EDS. These problems tend to worsen with disease progression. During the early stages of PD, dopaminomimetic drugs such as L-3,4-dihydroxyphenyalamine (L-DOPA) can treat these sleep disturbances; unfortunately, patients become refractory to L-DOPA treatment during the more advanced stages of PD.

Ascertaining the contribution of central DA dysfunction to the sleep disturbances of PD is complicated by several factors. First, patients with PD also exhibit neurochemical changes in cholinergic and monoaminergic systems, both of which are implicated in sleep regulation. Second, medications used to treat PD may alter sleep. Third, nocturnal tremors and inappropriate phasic motor bursting may produce fragmented sleep. Nevertheless, several compelling lines of evidence provide support for the concept that DA dysfunction significantly contributes to the pathological sleep disturbances of PD, the most common of which are EDS and nocturnal sleep disruption in the form of involuntary motor disturbance (e.g., REM sleep behavior disorder (RBD)).

Significant hypersomnia in PD, manifesting as EDS, is common and yet underrecognized with respect to diagnosis and treatment. Clinic-based objective

measurements of sleepiness employing the standardized Multiple Sleep Latency Test has revealed a high rate of EDS (approximately 20–50%) in PD patients. As indicated previously, the determinant(s) of EDS in PD is difficult to resolve because EDS is likely secondary to severe sleep fragmentation in PD, which itself may be attributable to other motor (e.g., abnormal nocturnal movements) or respiratory (e.g., central/obstructive sleep apnea) disturbances that accompany the pathology of PD.

Perhaps a more insidious form of sleep disruption that has been tacitly linked to central DA dysregulation is RBD. RBD is a parasomnia that typically manifests as 'dream enactment' behavior – that is, involuntary nocturnal movements that include kicking, punching, shouting, and screaming during REM sleep. RBD may represent an early pathophysiologic manifestation of evolving PD and other Lewy body diseases (LBDs) (e.g., dementia with Lewy bodies, multiple systems atrophy, and pure autonomic failure). Indeed, RBD typically manifests a decade prior to the motor and cognitive sequela of PD, and thus the diagnosis of RBD may provide an early therapeutic window for delaying or preventing the full development of PD.

Of note, because Lewy bodies and Lewy neurites are composed of α-synucleinopathies, these disorders are termed 'synucleinopathies.' Although the brain stem is clearly implicated in RBD pathogenesis, the identity of the neural networks that become dysfunctional to manifest RBD is unknown. Because approximately 50% of people with RBD develop PD or dementia with Lewy bodies and the majority of PD patients have RBD, it has been suggested that the intrinsic pathology of PD (i.e., severe nigrostriatal DA neuron loss) may contribute to the development of excessive nocturnal movement, particularly RBD. Despite the intuitive appeal of this hypothesis, however, studies have linked RBD to dysfunction of brain stem REM sleep circuitry. These structures include the subcoeruleus region (SC; equivalent to the sublateral dorsal nucleus (SLD) in rats) and, possibly, the intermediate region of the ventromedial medulla; both structures appear critical for generating atonia during REM sleep. SC/SLD dysfunction, as the neuropathologic substrate for RBD, is seemingly consistent with the temporal pattern of neuronal degeneration in PD and other LBDs, which starts in the brain stem and includes the coeruleus–subcoeruleus complex in stages I and II of PD and progresses inexorably rostrally toward the forebrain. The temporal pattern of lesions (i.e., caudal-to-rostral progression) is also consistent with RBD (secondary to SC/SLD degeneration) as an early manifestation of these neurodegenerative conditions. Nevertheless, although SC/SLD dysfunction has been demonstrated in PD, it remains unclear if RBD in evolving PD is caused by SC/SLD degeneration or loss of critical inputs such as DA nigrostriatal projections.

Dopamine and Stimulant-Induced Arousal

Drugs such as amphetamines and amphetamine-like stimulants (e.g., cocaine, methylphenidate, and methamphetamines) have potent wake-promoting effects. These drugs are thought to produce their arousal effects by blocking DA reuptake/transport (i.e., acting on the cell membrane DAT and/or stimulating DA release), resulting in increased synaptic DA concentrations. Determining DA's contribution to behavioral arousal in this context has nevertheless been complicated by the fact that psychostimulant administration also promotes the synaptic accumulation of other monoamines, particularly norepinephrine (NE). This is particularly true at higher doses, where amphetamines interact with vesicular monoamine transporter-2 to increase the cytoplasmic pool of monoamines. In fact, for many years, the wake-promoting effect of amphetamine-like stimulants was attributed almost exclusively to NE mechanisms since adrenergic signaling is known to modulate arousal state. Further reinforcing this notion is the fact that LC neurons (the major source of brain NE) display robust state-dependent activity and discharge most rapidly during enhanced arousal. Although, taken together, these observations suggest NE signaling might form the molecular basis for the wake-promoting actions of psychostimulants, several studies have suggested a more dominant role for DA than NE in this regard. Centrally acting DA antagonists, for example, cause drowsiness, and drugs that selectively block the DAT are more effective in promoting wakefulness than drugs that selectively block the NE transporter (NET). Indeed, both desipramine and nisoxetine, two potent and selective NET blockers, have minimal effects on EEG arousal in narcoleptic canines (even at high doses which suppress REM sleep). The wake-promoting effect of amphetamine-like stimulants is also abolished in mice with deletion of the DAT gene. Conversely, the wake-promoting effect of amphetamines is preserved following lesions of the LC (and hence a dramatic reduction in central NE) in cats and following chemical ablation of NET-bearing NE forebrain projections from the LC in mice. The potency of most wake-promoting psychostimulants is also best predicted by binding affinity to DAT. Finally, elevated arousal levels correlate with increases in synaptic DA but not NE. Taken together, these findings support the concept that presynaptic modulation of DA transmission (at least at the level of the forebrain) is the key pharmacological mechanism

by which amphetamines and their derivatives mediate cortical and behavioral arousal.

The anatomical substrates for DA's effects on arousal also remain unresolved, although several lines of experimental evidence suggest this may occur through inhibition of sleep-promoting systems. For instance, as previously indicated, a mutually inhibitory circuit between wake-active vPAG DA neurons and sleep-active neurons of the VLPO has been elucidated. Moreover, *in vitro* administration of DA inhibits neurons of the VLPO. Curiously, however, DA-mediated inhibition of the VLPO is blocked by α_2 receptor antagonists but not by D2 receptor antagonists. Although these findings appear difficult to reconcile, they likely indicate, at a minimum, cross-talk between these central nervous system catecholaminergic systems. Thus, for example, the drug modafinil (which shares the wake-promoting properties of traditional psychostimulants) hyperpolarizes VLPO neurons *in vitro*, presumably by blocking the NET. Yet DAT knockout mice are unresponsive to modafinil treatment, suggesting instead that DA neurotransmission (i.e., blockade of the DAT) underlies the wake-promoting effect of modafinil. Combined, these findings suggest a possible dual mode of action for modafinil (i.e., interference with both NE and DA uptake), and this may explain why modafinil can exert its wake-promoting effects without inducing dopaminergic side effects such as addiction. Indeed, data indicate that modafinil may produce its arousal effects through dopaminergic-dependent adrenergic signaling. For example, DA may function as a physiological ligand at adrenergic receptors. In support of this concept, DA stimulation of NE receptors has been documented previously in the pontine brain stem. However, the mechanism of action by which DA produces its effects on cortical and behavioral arousal remains unresolved.

In addition to inhibiting VLPO neurons, amphetamine-like stimulants may also promote wakefulness by activating, for example, the wake-promoting basal forebrain cholinergic neurons and LH orexin neurons. To this end, DA has been reported to stimulate cholinergic cells in the basal forebrain by D1-like receptors (i.e., D1 and D5) *in vitro*. Research has also uncovered an extended network of basal forebrain and preoptic sites that mediate amphetamine-induced increases in behavioral and electrocortical arousal. Remarkably, however, although the anatomical resolution of this mapping study was limited, close inspection of the data reveals that infusion of amphetamine into the region best approximating the location of the VLPO produced one of the largest increases in arousal, providing further support for the concept that the VLPO may be a critical structure for mediating the arousal-promoting effects of amphetamines and amphetamine-like stimulants.

Summary

Work during the past decade has firmly delineated an important role for DA in sleep–wake regulation (both normal and pathologic) and behavioral and electrocortical arousal. Wake-active DA neurons of the vPAG likely exert their potent arousal influence through a mutually inhibitory interaction with the VLPO as well as less well-defined interaction with components of the ascending arousal system (e.g., basal forebrain, LC, and lateral hypothalamus). Based on recent findings, it is tempting to speculate that alterations in DA neurotransmission may underlie the excessive sleepiness of evolving PD as well as the manifestation of abnormal nocturnal movements in the form of RBD. Finally, a critical role for DA in mediating the wake-promoting effects of psychostimulants has begun to emerge, although the indirect or direct mechanism by which this occurs remains to be clarified.

See also: Dopamine – CNS Pathways and Neurophysiology; Sleep and Sleep States: Thalamic Regulation; Sleep and Sleep States: Hypothalamic Regulation; Sleep and Sleep States: Histamine Role; Sleep and Sleep States: Cytokines and Neuromodulation; Sleep–Wake State Regulation by Noradrenaline and Serotonin; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Aston-Jones G and Bloom FE (1981) Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep–waking cycle. *Journal of Neuroscience* 1: 876–886.

Berridge CW, O'Neil J, and Wifler K (1999) Amphetamine acts within the medial basal forebrain to initiate and maintain alert waking. *Neuroscience* 93: 885–896.

Crochet S, Onoe H, and Sakai K (2006) A potent non-monoaminergic paradoxical sleep inhibitory system: A reverse microdialysis and single-unit recording study. *European Journal of Neuroscience* 24(5): 1404–1412.

Dahan L, Astier B, Vautrelle N, Urbain N, Kocsis B, and Chouvet G (2007) Prominent burst firing of dopaminergic neurons in the ventral tegmental area during paradoxical sleep. *Neuropsychopharmacology* 32(6): 1232–1241.

Gallopin T, Luppi PH, Rambert FA, Frydman A, and Fort P (2004) Effect of the wake-promoting agent modafinil on sleep-promoting neurons from the ventrolateral preoptic nucleus: An *in vitro* pharmacologic study. *Sleep* 27: 19–25.

Jones BE, Harper ST, and Halaris AE (1977) Effects of locus coeruleus lesions upon cerebral monoamine content, sleep–wakefulness states and the response to amphetamine in the cat. *Brain Research* 124: 473–496.

Kume K, Kume S, Park SK, Hirsh J, and Jackson FR (2005) Dopamine is a regulator of arousal in the fruit fly. *Journal of Neuroscience* 25: 7377–7384.

Lu J, Chou TC, and Saper CB (2006) Identification of wake-active dopaminergic neurons in the ventral periaqueductal gray matter. *Journal of Neuroscience* 26(1): 193–202.

Miller JD, Farber J, Gatz P, Roffwarg H, and German DC (1983) Activity of mesencephalic dopamine and non-dopamine neurons across stages of sleep and waking in the rat. *Brain Research* 273: 133–141.

Wisor JP and Eriksson KS (2005) Dopaminergic–adrenergic interactions in the wake promoting mechanism of modafinil. *Neuroscience* 132: 1027–1034.

Wisor JP, Nishino S, Sora I, Uhl GH, Mignot E, and Edgar DH (2001) Dopaminergic role in stimulant-induced wakefulness. *Journal of Neuroscience* 21: 1787–1794.

Dopamine in Perspective

P Foley, Prince of Wales Medical Research Institute, Sydney, NSW, Australia

© 2009 Elsevier Ltd. All rights reserved.

Dopamine (systematic: 1,2-benzenediol, 4-(2-aminoethyl)-9CI; 3-hydroxytyramine or 3,4-dihydroxyphenylethylamine; $C_8H_{11}NO_2$; abbreviated DA) is a biogenic amine, catecholamine neurotransmitter, and pituitary hormone; it is also an intermediate in the synthesis of norepinephrine and epinephrine. DA was long known as 3-hydroxytyramine; Russell Dale proposed the name 'dopamine' in 1952 in order to avoid confusion with 5-hydroxytryptamine, as well as to underscore the relationship between the amino acid (DOPA) and the amine. 'Dopamine' did not, however, become the standard designation until the mid-1960s.

DA is the principal transmitter of several central nervous system (CNS) pathways, particularly those originating in the midbrain, and also acts as a neurotransmitter in several peripheral neurovegetative systems. In the CNS, dopaminergic (DAergic) systems are critically involved in the control and modulation of motor activity, learning, and motivation, and regional disruptions of these systems have been implicated in a number of disorders, including Parkinson's disease and schizophrenia. Several important classes of therapeutic agents are correspondingly active in CNS DAergic systems. DA itself is unable to cross the blood–brain barrier, for which reason L-DOPA and, more recently, DA receptor agonists are employed to increase CNS DAergic tone.

DA Synthesis and Catabolism

DA is synthesized in nervous tissue and adrenal gland by aromatic amino acid decarboxylase, a cytosolic enzyme which requires the cofactor pyridoxal phosphate (vitamin B_6), from the amino acid L-DOPA (3,4-dihydroxyphenylalanine); the rate-limiting step in DA synthesis, however, is the conversion of the amino acid tyrosine to L-DOPA by tyrosine hydroxylase (TH; Figure 1). DA is then sequestered in synaptic vesicles or particles following uptake by the relatively nonspecific ATP-dependent vesicular membrane transporter 2 (VMAT-2), which can be irreversibly blocked by the alkaloid reserpine (see below). Depolarization of the nerve terminal elicits a calcium-dependent fusion of vesicles with the terminal membrane and release of DA into the synaptic cleft (exocytosis), where it interacts with a variety of postsynaptic membrane protein receptors, initiating diverse biochemical responses in the postsynaptic cell. The released transmitter also interacts with presynaptic autoreceptors, thereby inhibiting depolarization-induced DA synthesis and release. DA synthesis and release in the CNS are tightly coupled; free DA (and other catecholamines) inhibit TH activity directly, competing for the cofactor binding site, while depolarization and activation of DAergic autoreceptors leads to phosphorylation of TH by a variety of kinases, reducing its sensitivity to catecholamine-related inhibition. Reduced nerve stimulation and release can thus lead to increased DA synthesis in cells bearing DA autoreceptors (Figure 2).

The principle mechanism for termination of DA-mediated transmission involves transmitter reuptake along its concentration gradient by a specific presynaptic Na^+- and Cl^--sensitive membrane DA transporter (DAT). Diffusion of DA away from the synapse and stimulation of more distant receptors is termed volume transmission, and a local extracellular DAergic 'tone' has been described in certain brain regions.

There are two major enzymes involved in DA catabolism in mammalian nervous tissue: mitochondrial outer membrane monoamine oxidase (MAO) in the nerve terminal, and extracellular catechol-*O*-methyl transferase (COMT), whereby the latter is not regarded as playing a major role in the regulation of CNS DA (Figure 3). Mammalian MAO exists in two forms; MAO-A is primarily responsible for DA metabolism in rodents, whereas type B is more significant in this capacity in humans, a difference attributable more to regional enzyme abundance than to differences in enzyme specificity. The first specific MAO-B inhibitor (selegiline) was synthesized in the USA and Hungary in the early 1960s, but this specificity was recognized only in the early 1970s. The major DA metabolite in humans is homovanillic acid (HVA), whereas dihydroxyphenylacetic acid (DOPAC) is more abundant in rodents. Levels of these metabolites have been monitored in the laboratory as indicators of DA turnover in specific brain regions of experimental animals; in humans, DA metabolites, as measures of CNS DA turnover, are assayed in cerebrospinal fluid to exclude products of peripheral DA catabolism.

DA Biochemistry: History

DA played a particularly significant role in the history of neurochemistry: following a series of investigations between 1957 and 1965, commencing in the laboratory and completed in the clinic, chemical

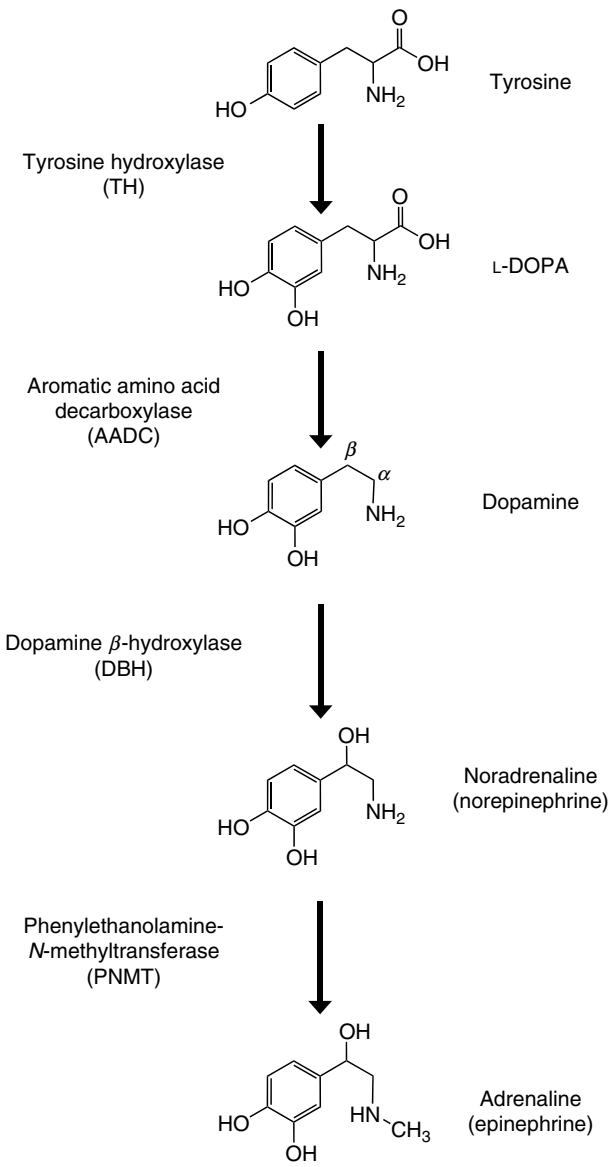

Figure 1 The major catecholamine synthetic pathway. The basic catecholamine structure is most easily seen in DA, consisting of the 1,2-dihydroxlated benzene ring (catechol or, more properly, pyrocatechol), to which is attached at the 4-position the aminoethyl group. The conventional numbering of the carbons in catecholamines, however, is referenced to phenylethylamine; DA is thus 3,4-dihydroxyphenylethylamine. The designation of the carbons in the aminoethyl chain is also shown. TH, the rate-limiting enzyme in catecholamine synthesis from tyrosine, is nearly saturated by the amino acid at normal physiological concentrations. AADC (also referred to as DOPA decarboxylase), which also decarboxylates the serotonin precursor 5-hydroxytryptophan, is found not only in DAergic and serotonergic nerve tissue, but also in kidney and vascular tissue.

neurotransmission and its significance for normal CNS function was first demonstrated in the case of DA. While chemical (as opposed to electrical) neurotransmission in peripheral nervous pathways had been recognized since the 1920s, the central role of so-called neurohumoral agents in the CNS remained a matter of vigorous dispute until the early 1960s. But before this unparalleled chain of investigations was conceptually possible, DA had principally been invoked in the context of general catecholamine research stretching back to the beginning of the twentieth century.

DA was first synthesized (from vanillin) in 1909–10 during investigations of the adrenal medulla in Berlin and London. In comparison with the 'major catecholamines,' adrenaline and noradrenaline, DA ('aminoethyl-catechol') was a weak vasopressor agent, and consequently sank into relative obscurity until the 1930s.

The DA precursor L-DOPA was first isolated from the broad bean (*Vicia faba*) by Torquati in 1911, but Markus Guggenheim (Hoffmann-La Roche, Grenzach) published the first major paper on its physiological properties in 1913. Guggenheim discovered by self-administration that it was "not completely innocuous," and for many years he sought a physiological role for L-DOPA as part of his investigation of what he termed the "biogenic amines"; ironically, he died in the year that L-DOPA was commercially launched as a medication for parkinsonism. Bruno Bloch (Basle) reported in 1917 that L-DOPA was involved in the synthesis of melanin from tyrosine in pigmented cells; he also discovered that it was the substrate for an oxidative reaction catalyzed by 'dopaoxydase' (tyrosinase).

Naturally occurring DA itself was first isolated from the broom plant (*Sarthamnus scoparius*) by two Hamburg chemists in 1931. By this time, it was widely assumed that L-DOPA was a precursor in adrenaline synthesis – elucidating the catecholamine synthetic pathway was one of the central problems of physiological chemistry until the late 1950s – but ignorance of how the amino acid might be converted to an amine *in vivo* left the question open. Much of the work which determined the biochemical pathway involved was published in a series of landmark papers by Peter Holtz (Rostock; after 1945: Frankfurt am Main). In 1937, he demonstrated that 'DOPA decarboxylase' in kidney rapidly decarboxylates L-DOPA (but not D-DOPA) to DA, and by 1939 had proposed the now-accepted catecholamine synthetic pathway (Figure 1). Proof that amino acid decarboxylation actually occurred *in vivo* was delivered by Holtz in 1941 and 1942 by administering L-DOPA to guinea pigs and humans and then isolating DA from their urine. In 1952, Holtz reported that pyridoxal-5-phosphate activates adrenal cortex DOPA decarboxylase, and that a similar activator was present in cortex, thus identifying the cofactor for the enzyme. By 1956, the presence of DA in a number of tissues

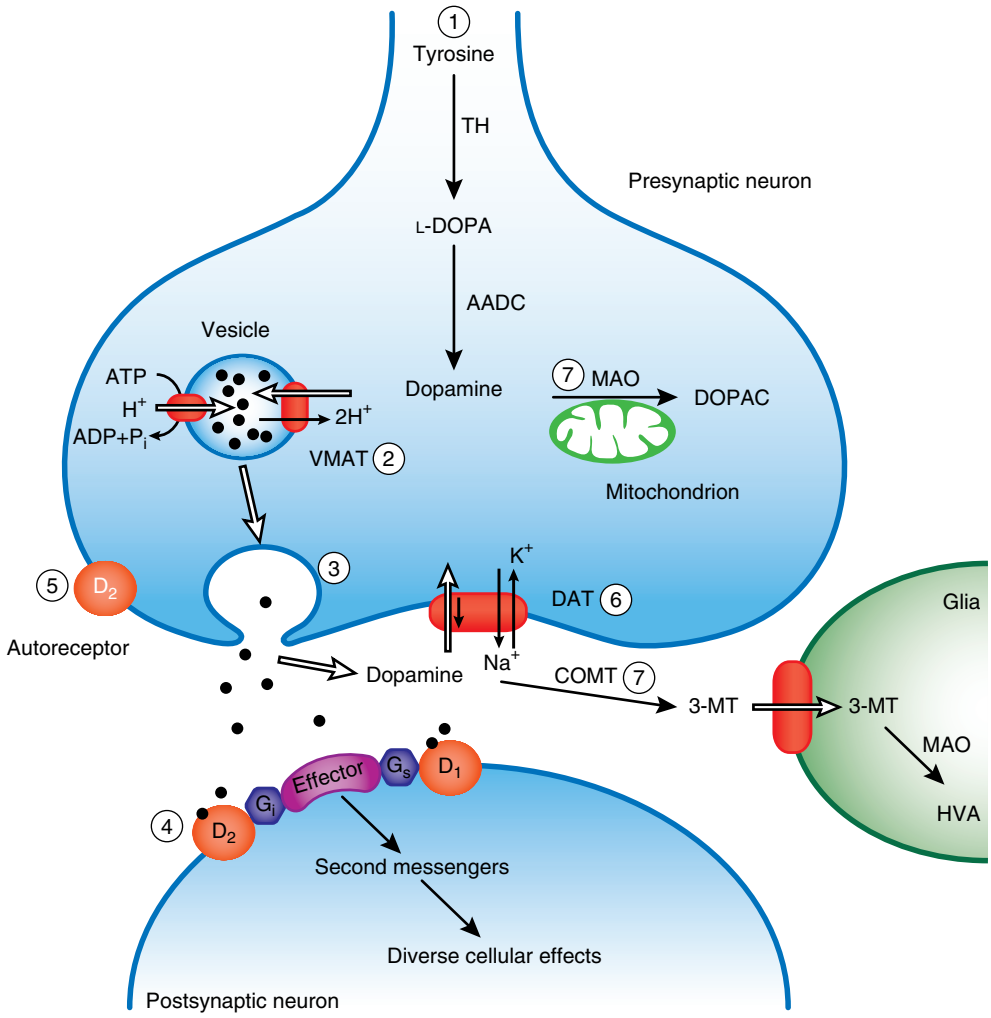

Figure 2 Schematic depiction of the synthesis, storage, release, and metabolism of DA in a DAergic nerve terminal. ① DA synthesis. TH activity can be inhibited by α-methyl-p-tyrosine, or elevated by increasing levels of its cofactor tetrahydrobiopterin. ② Vesicular storage effected by VMAT, the activity of which can be inhibited irreversibly by reserpine or irreversibly by tetrabenazine. ③ Exocytotic, Ca^{2+}-dependent DA release into synaptic cleft stimulated by depolarization of the nerve terminal. D-Amphetamine at high doses increases effective DA release. ④ DA binds postsynaptic G-protein-associated DA receptors, the varieties of which are broadly classified into the D_1 and D_2 receptor families, and thereby elicits a variety of specific intracellular biochemical events in the postsynaptic cell, including activation of adenylyl cyclase, various kinases, and modulation of ion channel states. Agonists and antagonists of varying specificity for particular DA receptors are employed experimentally and clinically to modify the activity of DAergic systems. ⑤ Binding of DA to presynaptic D_2 autoreceptors inhibits DA synthesis (by inhibiting TH activity, which is also reduced by nerve depolarization) and release. ⑥ Reuptake of DA into the presynaptic terminal by DAT is the primary means for ending DAergic stimulation of the postsynaptic cell. DAT activity can be inhibited by D-amphetamine, cocaine, and nomifensine. ⑦ Intracellular catabolism of DA by MAO and extracellular metabolism by COMT can be reduced by enzyme inhibitors of varying specificity.

had been demonstrated, but not its active accumulation. Most significantly, Schümann, working in Holtz' laboratory in Frankfurt, had identified DA in sympathetic nerves and ganglia, and he suggested that much of what had hitherto been recognized as adrenaline in these tissues was probably DA.

Nagatsu, Levitt, and Udenfriend (Bethesda) provided in 1964 the final essential link in the catecholamine synthetic pathway when they reported the first direct evidence for the presence of tyrosine hydroxylase in brain. Further, Kenneth Shaw and his group (Salt Lake City) had established in 1957 that HVA was present in normal human urine, and their experiments led them to propose the now-accepted metabolic pathway for DA (Figure 3).

DA Function in the CNS: History

By the mid-1950s, the possibility of at least a limited role for chemical transmission in the CNS was being discussed, with acetylcholine, histamine, and substance P seen as the likeliest candidates, while

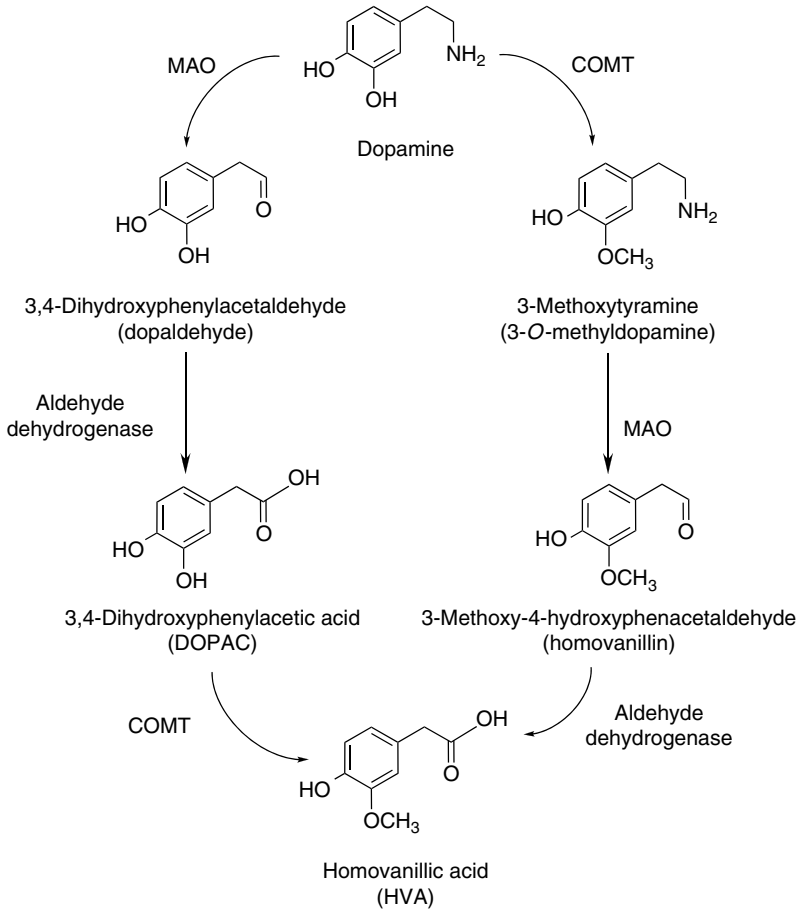

Figure 3 The major catabolic pathways of DA. Both DOPAC and HVA can be further metabolized to sulfate conjugates prior to excretion in urine. MAO, monoamine oxidase; COMT, catechol-*O*-methyltransferase.

Bernard Brodie and his group (NIH, Bethesda) believed that serotonin (5-HT) was particularly significant. Most researchers, however, believed that the physiological significance of DA lay not in its circulatory or sympathomimetic capacity, but rather in its role as precursor for noradrenaline, adrenaline, and melanins.

This view was challenged about this time by Hermann ('Hugh') Blaschko (Oxford). In 1937, he had discovered (mono)amine oxidase and found that it metabolized both DA and noradrenaline, and in 1939 had independently proposed the catecholamine synthetic sequence. In 1956, Blaschko noted that DA was not stored in adrenal chromaffin tissue, consistent with its putative role as a metabolic intermediate, but that this was not true for adrenergic nerve endings, so that DA here might possess in its own right hitherto unrecognized regulatory functions. The investigations which would be employed to explore this hypothesis were made possible by the development of more sensitive assays which exploited the fact that DA is liable to oxidation under controlled laboratory conditions, and in particular by adaptations of the fluorometric techniques introduced by Brodie's laboratory for the assay of serotonin.

In 1956, the Swedish pharmacologist Arvid Carlsson and colleagues reported that the plant alkaloid reserpine, then employed in the clinic as an antipsychotic agent, depleted catecholamine stores in adrenal medulla and brain tissue. Further, they demonstrated that stimulation of sympathetic nerves after reserpine treatment no longer elicited catecholamine release, and Carlsson hypothesized that this was relevant to the central hypotensive effect of reserpine. This view placed him at odds with his respected mentors Brodie and Shore, with whom he had recently spent a year as postdoctoral fellow, who attributed its effects to its release of serotonin. Carlsson sought to resolve the issue by replenishing central monoamine stores in reserpinized animals by administrating their amino acid precursors; peripherally administered DA and serotonin do not appreciably penetrate the blood–brain barrier. In experiments which would prove to among the most important in catecholamine history, Carlsson found that treatment with racemic

DOPA, but not 5-hydroxytryptophan, quickly reversed reserpine-induced tranquilization of mice and rabbits, but without elevation of noradrenaline levels. He therefore concluded that DA depletion underlay reserpine sedation, and that DA was thus critically involved in CNS neurotransmission.

The history of neurochemistry is, like the history of brain function since the nineteenth century, largely a history of the localization of particular functional elements. It was necessary, before any function could be ascribed to DA or the other catecholamines in the CNS, to establish not only that they occurred in brain, but also that they were localized to specific regions of the brain. It is possible that CNS DA was first detected by the cardiologist Wilhelm Raab (Vermont) in 1947 as 'encephalin.' The unknown catechol substance detected in 1957 by Katharine Montagu (Wickford, Essex) in brain, 'substance X,' may have partially consisted of DA, but Montagu did not appear to recognize the significance of this finding. In contrast, Carlsson's group concluded in 1958 that the levels of DA measured in rabbit brain were inconsistent with its being merely an intermediate. Two of Carlsson's doctoral students, Bertler and Rosengren, reported that in eight mammalian species CNS DA was predominantly localized in the striatum, distinctly different from the distribution of noradrenaline, also examined using the fluorescent techniques then being introduced in Carlsson's laboratory. On the basis of this distribution, and the fact that drugs which modulated its levels also affected motor performance, the authors concluded that "dopamine is concerned with the function of the corpus striatum and thus with the control of motor function." At the First International Catecholamine Symposium (1958), Carlsson also compared the effects of reserpine-induced DA depletion with the symptoms of parkinsonism, and suggested that the deficit could be compensated by administration of either the transmitter precursor or of a sympathomimetic amine. In early 1959, the neurologist Isamu Sano (Osaka) published a detailed analysis of DA distribution in human brain which indicated that the highest concentrations here were also in the striatum.

At the CIBA Foundation Symposium on Adrenergic Mechanisms (March 1960), many doyens of physiology remained unconvinced that catecholamines played a major role in CNS neurotransmission. Marthe Vogt, for example, who in 1954 had published a landmark paper on the localization of catecholamines in the dog brain, argued that any central catecholamine effects were attributable to general or local hypertensive effects. During a vigorous discussion of the question at the end of the symposium, where Carlsson advanced the role of catecholamines and Brodie that of serotonin in modulating behavior, Vogt concluded that any theories linking central DA or 5-HT levels with behavior would not have a long life.

On the contrary, results accrued from various laboratories during the next few years which confirmed the significance of CNS biogenic amines. But the conclusive evidence for the role of DA was ultimately provided by a series of clinical experiments. During the 1950s, reserpine was a commonly employed sedative, particularly in psychiatric patients, but extrapyramidal reactions ranging to a parkinsonian-like state were an occasional but unwelcome side effect. In 1958, Rudolf Degkwitz and colleagues (Frankfurt am Main) commenced experiments in which they administered L-DOPA to reserpine-treated patients, and reported in 1960 that the amino acid rapidly abolished the lethargy and feelings of exhaustion which had hitherto plagued their patients. Similarly, Sano reported in February 1960 the experimental treatment of parkinsonian patients with infused D,L-DOPA. Sano had found the DA levels in specific brain regions of a deceased parkinsonian patient to be significantly reduced when compared with normal levels, most dramatically in substantia nigra and striatum. The results of his experimental D,L-DOPA therapy were promising, but Sano did not pursue this phenomenon further.

Convincing, direct evidence for disturbed monoamine function in Parkinson's disease was provided in Vienna by the pharmacologist Oleh Hornykiewicz and the neurologist Walther Birkmayer. Hornykiewicz had 18 months' valuable experience under the tutelage of Blaschko in Oxford, where he had acquired the best available methodologies for measuring catecholamines in biological tissues. After reading the 1959 Bertler/Rosengren paper, Hornykiewicz and Herbert Ehringer found the highest human CNS DA concentrations in the caudatus and putamen, with moderately high levels in the substantia nigra and area postrema, confirming the results of Carlsson and Sano. More exciting was what he observed in brains from parkinsonian patients, as captured by Hornykiewicz' 1992 reminiscence of the moment of discovery: he could literally see with the naked eye the basal ganglia DA deficiency in Parkinson's disease, as the total absence of the pink reaction elicited by DA in control brains.

Shortly after Hornykiewicz had identified the basal ganglia DA deficit in the parkinsonian brain, he proposed that Birkmayer treat parkinsonian patients with L-DOPA. In July 1961, Birkmayer intravenously administered 50 mg L-DOPA to a female postencephalitic parkinsonian patient, with dramatically positive effects on her parkinsonian symptoms, particularly her akinesia. The simplest explanation was that the hypothesis that the striatal DA deficit was responsible

for parkinsonian akinesia and that L-DOPA administration was capable of ameliorating this deficit, was correct. But the initial response to these findings was lukewarm, and Hornykiewicz designated the period 1960–67 the "seven lean years" of DA research. Despite the promising start at the end of the 1950s, interest in DA remained restricted in the early 1960s to a small group scattered across the world.

Hornykiewicz wished to improve the new therapy by expanding the theoretical basis upon which it was built, while Birkmayer applied his observational skills to determining whether theoretical expectations were reflected by clinical reality. At this point Hornykiewicz was invited to write a review article on DA and Parkinson's disease for the leading German medical journal, the *Deutsche Medizinische Wochenschrift*, the first review concerning DA in a major medical journal.

The foundations for studies of the anatomical basis of CNS DAergic transmission were also laid in the early to mid-1960s. André Barbeau and Theodore L Sourkes (Montréal) had independently investigated L-DOPA therapy for parkinsonian patients in 1961, basing their approach on reduced urinary DA levels in parkinsonian patients. More significantly, Sourkes and Louis J Poirier reported in 1964–65 that interruption of nigrostriatal fibers reduced striatal catecholamine levels. It is easy to forget that the concept of 'dopaminergic' and other transmitter-defined pathways was remarkable at the time; neural tracts were spatially defined by histological techniques according to biochemical reactions unrelated to transmitter content. Some residual suspicion persisted regarding the Sourkes–Poirier results in light of the originality of their integration of experimental neuroanatomy, neurochemistry, and behavioral studies; further, the existence of the nigrostriatal pathway contradicted long-standing neurohistological orthodoxies. These doubts were soon allayed by functional investigations which provided further evidence for a direct DAergic nigrostriatal connection.

The visualization of the pathway was also achieved in 1965, by Carlsson, Nils-Åke Hillarp, and Kjell Fuxe (Karolinska Institute, Stockholm), using the sensitive condensation-fluorescent detection method which had been developed since 1959 by Hillarp, Bengst Falck, and Georg Thieme. The perfection of this method and its employment to demonstrate CNS catecholaminergic neurons was of fundamental importance to modern neurochemistry. As Carlsson noted in 1987, the discoveries of DA deficits in parkinsonism and the beneficial effects of L-DOPA were important, but the real breakthrough came with the histochemical techniques visualizing the cellular localization of the monoamines by means of the fluorescence microscope.

The importance of this work for neurochemistry cannot be overestimated, not only for its contribution to the theoretical analysis of Parkinson's disease but also because, as Poirier and Sourkes noted in retrospect, it stimulated studies aimed at the identification of neurochemically defined pathways and their role in behavioral and psychoneuroendocrinological phenomena.

The first unequivocal demonstration of a unity between physical lesion, neurochemical changes, and behavioral abnormalities was a major step forward in establishing neurochemical investigations as the dominant research paradigm in the further investigation of CNS function and disease. For his particularly significant contributions of his group to these developments from 1956 onward, Carlsson was one of the three recipients of the Nobel prize for physiology or medicine in 2000.

The outcome of these investigations was that residual skepticism regarding the role as CNS transmitters of catecholamines, and of DA in particular, was largely assuaged. In subsequent years, increasingly sophisticated neuroanatomical techniques were employed to precisely map CNS pathways associated with DA and other transmitters, revealing functional connections not apparent on the basis of classical, structural neurohistology, as well as interactions between these systems and the integration of CNS information and its exchange. The further development of visualization techniques has similarly permitted the elucidation of the structural characteristics of DAergic neurons and synaptic structures. Further, increasingly sensitive electrochemical procedures have been applied to the assay of DA in biological fluids, both *ex vivo* and in specific brain regions of living research animals. Finally, animal studies and the correlation of neurochemical and neurohistological findings with clinical data have allowed the illumination of the functional significance of DA for human emotion, behavior and motor performance.

Functional Significance of CNS DA

DA was initially thus identified as being involved in the regulation of certain motor functions. The link between deterioration of the nigrostriatal DAergic pathway and the consequent loss of striatal DA and the akinesia of parkinsonism represented the first association of a specific neurochemical deficit with a CNS disorder, and thus the first conclusive implication of a chemical neurotransmitter in CNS function. Pharmacological treatment of Parkinson's disease has since been directed at compensating for reduced extrapyramidal DA release by administration of the DA precursor L-DOPA (usually in

combination with a blood/brain barrier-impermeable decarboxylase inhibitor, in order to minimize peripheral L-DOPA metabolism), of direct-acting DA receptor agonists, or of inhibitors of DA-metabolism, particularly MAO-B inhibitors.

In the ensuing four decades, further CNS DAergic pathways and systems have been identified, and the palette of functions linked with DA correspondingly expanded. The major DAergic CNS pathways are the following: (1) the nigrostriatal system, projecting from the substantia nigra to the striatum, involved in control of voluntary movement, as discussed above; (2) the mesocortical–mesolimbic system, projecting from the ventral tegmentum to the nucleus accumbens, tuberculum olfactorium, amygdala, area entorhinalis, and septum, involved in control of motivational behavior and probably also of learning and memory; and (3) the tuberoinfundibular system, projecting from the hypothalamic nucleus arcuatus and inhibiting hormone release (principally of prolactin) from the anterior pituitary.

That excessive DAergic activity is involved in the etiology of psychosis and schizophrenia was first explicitly hypothesized by Carlsson and Lindqvist in 1963, who demonstrated that phenothiazine-class antipsychotic agents acted principally by blocking DA receptors. This hypothesis, since rendered more precisely as hyperactivity of the mesolimbic DAergic system, is consistent with the observation that releasing agents which increase extracellular DA levels, such as cocaine (an uptake blocker) and amphetamine, can elicit psychotic symptoms. On the other hand, reduced DAergic function in the prefrontal cortex, where DAergic neurons contribute to regulating the flow of information from other CNS regions, is associated with deficits in memory and attention processes, and has been implicated both in the negative symptoms of schizophrenia as well as in attention-deficit disorders.

The mesocortical–mesolimbic DAergic system, particularly the ventral tegmentum-nucleus accumbens projection, has similarly been implicated in phenomena relevant to reward, motivation, and drug reinforcement/addiction, and thus also in problems of psychostimulant drug withdrawal. Although the concept of DA as a reward chemical is on overly simplistic interpretation of these phenomena, the appropriate regulation of DAergic function in this pathway is undoubtedly important for proper emotional performance as well as for goal-directed behavior. DA appears to be particularly critical for establishing a "conditioned tendency to re-approach environmental stimuli that have been associated with most primary rewards, and for maintaining habit strength once a task has been learned" (Wise, 2004).

The role of DA outside the CNS is not as clear cut. DA-containing neurons and DA receptors have been identified in a variety of tissues, and DA is the most abundant plasma catecholamine, presumably related to peripheral autonomic activity, its levels being increased by a number of physiological stressors. DA elicits hypotension subsequent to direct vasodilation and reduced vascular resistance (via stimulation of D_1-type receptors) and indirect vasodilation (D_2-type receptor-mediated reduction of sympathetic tone) in renal, pulmonary, coronary, and mesenteric vascular beds, and stimulates cardiac contractility (via activation of β_1-adrenergic receptors). Identification of specific DA receptor types by molecular biological techniques also suggests that DAergic activity plays roles in the regulation of immune activity, in blood lymphocytes, and in the lung. DA plays an important role in renal function, increasing perfusion as well as enhancing natriuresis and diuresis, and defective renal DA function has been suggested to underlie age-related declines in kidney function, with implications for the pathogenesis of hypertension and diabetes.

Intravenous DA is employed in the clinic for its direct (β_1-adrenergic receptor agonist) and indirect (noradrenaline releasing) sympathomimetic effects. At low ('renal') doses it is administered as a renal and visceral vasodilator, while at higher dosages it is utilized as an inotropic agent. It has thus been used in cardiac emergencies, hypotension, shock, and kidney failure. The utility of 'renal dose' DA, however, has been increasingly questioned in recent years; critics note that evidence for its clinical effectiveness from large-scale controlled trials has not been forthcoming, and that the potential deleterious effects of systemic DA administration (impaired rather than enhanced mucosal blood flow via activation of α_1 receptors, aggravation of reduced gastric motility, impairment of pulmonary function) also suggest caution, particularly as the precise mechanisms of DAergic renal regulation, for instance, remain to be defined. The significance of DA-induced negative changes in endocrine (central hypothyroidism following suppression of anterior pituitary function) and immune function also requires clarification.

See also: Animal Models of Parkinson's Disease; Dopamine; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopamine Neurons: Reward and Uncertainty; Dopamine: Cellular Actions; Dopaminergic Differentiation; Dopaminergic Agonists and L-DOPA; Parkinsonian Syndromes; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome.

Further Reading

Bennett MR (1998) Monoaminergic synapses and schizophrenia: 45 years of neuroleptics. *Journal of Psychopharmacology* 12: 289–304.

Blaschko H (1959) The development of current concepts of catecholamine formation. *Pharmacological Reviews* 11: 307–316.

Carlsson A (1987) Perspectives on the discovery of central monoaminergic neurotransmission. *Annual Review of Neuroscience* 10: 19–40.

Carlsson A (2003) A half-century of neurotransmitter research: Impact on neurology and psychiatry. In: Jörnvall H (ed.) *Nobel Lectures. Physiology or Medicine 1996–2000. Nobel Lectures, Including Presentation Speeches and Laureates' Biographies*, pp. 303–322. Singapore: World Scientific.

Crossland J (1957) The problem of non-cholinergic transmission in the central nervous system. In: Richter E (ed.) *Metabolism of the Nervous System*, pp. 523–541. London: Pergamon.

Foley PB (2003) *Beans, Roots and Leaves. A History of the Chemical Therapy of Parkinsonism*. Marburg: Tectum.

Guggenheim M (1961) Historische Betrachtungen über Dioxyphenylalanin, Catecholamine und andere biogene Amine. *Bulletin der Schweizerischen Akademie der Medizinischen Wissenschaften* 17: 309–316.

Holtz P and Credner K (1942) Die enzymatische Entstehung von Oxytyramin im Organismus und die physiologische Bedeutung der Dopadecarboxylase. *Archiv für experimentelle Pathologie und Pharmakologie* 200: 356–388.

Hornykiewicz O (1966) Dopamine (3-hydroxytyramine) and brain function. *Pharmacological Reviews* 18: 925–964.

Hornykiewicz O (1992) From dopamine to Parkinson's disease: A personal research record. In: Samson F and Adelman G (eds.) *The Neurosciences: Paths of Discovery II*, pp. 125–147. Boston: Birkhäuser.

Rice ME (2000) Distinct regional differences in dopamine-mediated volume transmission. In: Agnati LF, Fuxe K, Nicholson C, and Syková E (eds.) *Volume Transmission Revisited (Progress in Brain Research 125)*, pp. 277–290. Amsterdam: Elsevier.

Schenarts PJ, Sagraves SG, Bard MR, et al. (2006) Low-dose dopamine: A physiologically based review. *Current Surgery* 63: 219–225.

Sourkes TL (2000) How dopamine was recognised as a neurotransmitter: A personal view. *Parkinsonism and Related Disorders* 6: 63–67.

Wise RA (2004) Dopamine, learning and motivation. *Nature Reviews Neuroscience* 5: 483–494.

Dopamine Neurons: Reward and Uncertainty

W Schultz, University of Cambridge, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The classic view of dopamine functions derives from the clinical neurologist examining a human parkinsonian patient who suffers from severe problems with ocular, facial, and skeletal movements, resting tremor, and muscular rigidity. Likewise, experimental dopamine depletions by lesions or drugs, or dopamine receptor blockade, result in severe behavioral deficits of movement initiation and execution. Further investigations of humans and animals with dopamine depletions have revealed an even larger spectrum of behavioral deficits, including many cognitive and motivational disabilities. However, electrophysiological studies of the impulse activity of single dopamine neurons have revealed primarily a role in the processing of information related to reward and uncertainty. These studies could not confirm active dopamine involvement in the large variety of functions deficient after their lesion, including movements, suggesting that the crucial role of dopamine in these processes consists of enabling postsynaptic neurons that actively encode these specific functions. This article describes the electrophysiological activity of single dopamine neurons in relation to reward and uncertainty.

Behavioral Theories of Reward and Uncertainty

Public perception associates rewards primarily with happiness and special gratification, but behavioral research suggests wider functions. Rewards are any objects or events that generate approach behavior and consumption, produce learning of such behavior, and constitute outcomes of economic decision making. Rewards are of crucial importance for individual and gene survival, and they support elementary processes such as drinking, eating, and reproduction. Rewards are polysensory and do not engage specialized reward receptors, and the brain extracts the reward information from visual, auditory, somatosensory, olfactory, and other sensory information. Thus, the functions of rewards are not defined by the physics and chemistry of their inputs but by the behavioral reactions they induce.

The laws of mechanics, optics, acoustics, and biochemistry define the key functions of primary sensory systems in the brain. The dedicated physical and chemical receptors of these systems translate environmental energy and information into neural language. By contrast, there are no dedicated receptors for reward, and the information enters the brain through the touch, taste, visual, and auditory receptors of primary sensory systems. The functions of rewards cannot be derived entiely from the physics and chemistry of input events and are based primarily on behavioral effects. Thus, the investigation of reward functions requires behavioral theories that can conceptualize the different effects of rewards on behavior. Animal learning theory and economic decision theories provide coherent frameworks for the investigation of neural reward mechanisms. The central tenets of these theories are based on observable behavior and, superficially, on the behaviorist approach, although mental states of representation and prediction are included.

Learning

Rewards induce changes in observable behavior and serve as positive reinforcers by increasing the frequency of the behavior that results in reward. In Pavlovian, or classical, conditioning, the outcome follows the conditioned stimulus irrespective of any behavioral reaction, and repeated pairing of stimuli with outcomes leads to a representation of the outcome that is evoked by the stimulus and elicits the behavioral reaction. Thus, Pavlovian conditioning produces outcome predictions. By comparison, instrumental, or operant, conditioning requires the subject to execute a behavioral response; without such response, no reward occurs. Instrumental conditioning increases the frequency of those behaviors that are followed by reward by reinforcing stimulus–response links. Instrumental conditioning allows subjects to influence their environment and determine their rate of reward. However, there is general agreement that the stimuli used in instrumental conditioning have become reward predictors through Pavlovian learning.

Associative learning depends crucially on the discrepancy between the occurrence of a reward and its prediction. The importance of such prediction errors is derived from Kamin's blocking effect, which postulates that a reward that is fully predicted does not contribute to the learning of a stimulus or action, even when it has been repeatedly paired with the stimulus or action. This is conceptualized in the associative learning rules, according to which learning advances only to the extent to which a reinforcer is unpredicted and slows progressively as the reinforcer becomes more predicted. The omission of a predicted

reinforcer reduces the strength of the conditioned stimulus and produces extinction of behavior. So-called attentional learning rules in addition relate the capacity to learn (associability) in certain situations to the degree of attention evoked by the conditioned stimulus or reward.

Approach Behavior and Decision Making

Rewards elicit approach and consummatory behavior. This is due to the objects' being labeled with appetitive value through innate mechanisms (primary rewards) or, in most cases, classical conditioning, after which these objects constitute, strictly speaking, conditioned reinforcers. Nutritional rewards derive their value in addition from hunger and thirst states (drive), and satiation of an animal reduces the reward value and consequently the behavioral reactions.

Conditioned, reward-predicting stimuli also induce preparatory or approach behavior toward the reward. In Pavlovian conditioning, subjects automatically show nonconsummatory behavioral reactions that would otherwise occur after the primary reward and which increase the chance of consuming the reward, as if a part of the behavioral response has been transferred from the primary reward to the conditioned stimulus (Pavlovian response transfer).

In instrumental conditioning, the action becomes associated with reward and thus obtains a value. Decision-making mechanisms are based on the action values of the different available options. Furthermore, a reward can become a goal for instrumental behavior if the reward is represented at the time the behavior is being prepared and executed and if there is a systematic relationship between the specific action leading to the reward.

When more than one option is available, animals show preferences for specific options, expressed as the probability of choosing one option over all others. Choices are crucially based on predictions of outcomes. Without predictions, agents can only guess, as they do not know what they will get when making the choice. This gives the Pavlovian learning of reward predictors a key role in decision making. The reward function in decision making is thus indirect: rewards serve for predictive Pavlovian learning, and the acquired prediction directs the choice. Reward itself is the outcome of the overt choice, and its evaluation is used to update the prediction used for the next decision.

Quantification of Reward Value and Uncertainty

Decision mechanisms maximize the outcome of choices by comparing the values of all available options and choosing the option with the highest value. The values of options are compared on the basis of their probability distributions. The expected value of a probability distribution denotes the summed product of each reward value weighted by its probability and provides a single numeric measure of the outcome. However, additional factors play important roles, and outcomes are better quantified by the utility they have for the individual decision maker. Utilities are assessed by preferences in overt choice behavior, resulting in the quantification of outcomes as single scalar variable: expected utility $= \sum u(x) \times p$ (u is utility, x is value, and p is probability).

A major factor influencing utility is uncertainty. In risk averse subjects, uncertainty reduces the utility of a reward (a reward has less value for me if I am afraid that I won't get it), whereas risk seekers find higher utility in an uncertain reward (I like the uncertainty and thus prefer risky rewards over safe ones), all other reward parameters being equal. The influence of uncertainty on outcome utility becomes apparent when doing the Taylor series expansion of expected utility.

Pleasure

Subjective feelings of pleasure and the resulting positive emotion represent key functions of rewards. However, it is unclear whether the pleasure can have a reinforcing causal effect for behavior (I feel good because of the outcome I got and therefore will do again what produced the pleasure) or is simply an epiphenomenon (my behavior gets reinforced and, in addition, I feel good because of the outcome). Although recent theories propose a distinction between the (unconscious, unpleasurable) wanting and the (conscious, pleasurable) liking of rewards, animal experiments investigating the issues of subjective feelings of pleasure run into obvious problems because of the lack of common language. Other than for specific investigations, the issue may be left aside when studying the neural mechanisms of reward in controlled behavioral neurophysiological experiments on animals.

Motivational Valence

Punishers have valence opposite that of rewards, induce withdrawal behavior, and act as negative reinforcers by increasing the behavior that results in decreasing the aversive outcome. Avoidance can be passive when subjects increasingly refrain from doing something that is associated with a punisher, or it can be active by increasing an instrumental response that is likely to reduce the impact of a punisher. Punishers induce negative emotional states of anger, fear, and panic.

Reward Signals in Dopamine Neurons

A prime goal in the investigation of neural processes of reward is to identify an explicit neuronal reward signal, just as retinal responses to visual stimuli constitute starting points for investigating the neuronal processes underlying visual perception. The search for a 'retina of the reward system' has located brain signals related purely to reward value irrespective of sensory and motor attributes in midbrain dopamine neurons and in select neurons of orbitofrontal cortex, dorsal and ventral striatum, and possibly amygdala. Reward signals influence neural decision processes in cortical and subcortical structures underlying behavioral actions and thereby contribute to economic choices.

Most midbrain dopamine neurons show phasic activations following primary food and liquid rewards; visual, auditory, and somatosensory reward-predicting stimuli; and physically intense visual and auditory stimuli. They are briefly depressed in their activity by reward omission and by stimuli predicting the absence of reward. They respond very little to aversive stimuli and not at all to inedible objects and known neutral stimuli unless these stimuli are very intense or large. The subsecond dopamine reward response may be responsible for the reward-induced dopamine release seen with voltammetry but would not easily explain the 300–9000 times slower dopamine fluctuations with rewards and punishers seen in microdialysis.

Reward Prediction Error

The response to reward appears to code the discrepancy between the reward and its prediction ('prediction error') such that an unpredicted reward elicits an activation (positive prediction error), a fully predicted reward elicits no response, and the omission of a predicted reward induces a depression (negative error, Figure 1). The prediction error response is sensitive to both the occurrence and the time of the reward, as a delayed reward induces a depression at its original time and an activation at its new time.

The positive and negative reward prediction error responses are graded such that a partial prediction error induces a smaller error response. Prediction errors covary in a monotonic fashion with both reward magnitude and probability and thus reflect the expected value of the probability distribution of reward magnitudes relative to prediction.

The reward prediction error response appears to normalize to the predicted range or standard deviation of the prediction error. For example, when three visual stimuli predict different binary distributions of equiprobable reward magnitudes, the larger magnitude always elicits the same positive response, even with a tenfold difference in prediction error,

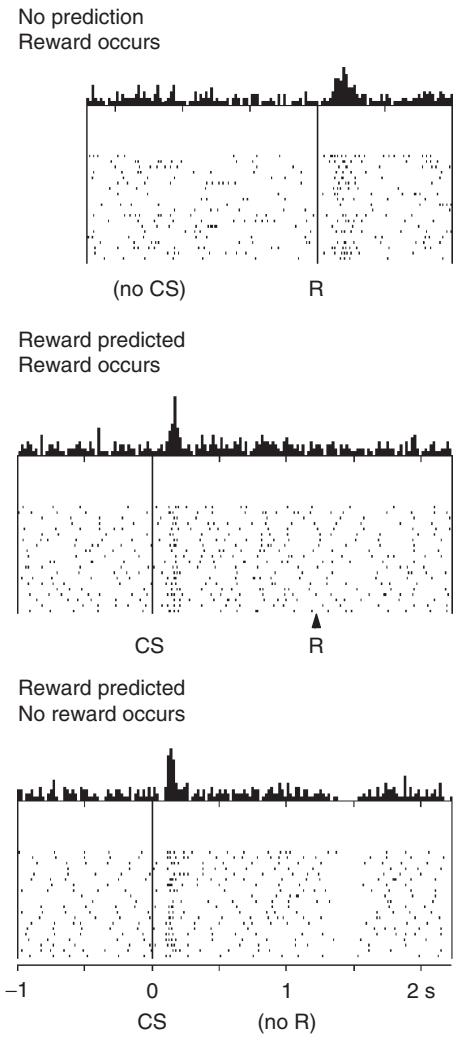

Figure 1 Reward prediction error response of a single dopamine neuron. CS, conditioned stimulus; R, reward. Reprinted from Schultz W, Dayan P, and Montague RR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599, with permission from AAAS.

although the same neurons are sensitive to unpredicted magnitudes. As a result of this gain adaptation, the neural response discriminates between the two likely outcomes equally well, regardless of their absolute magnitude difference.

The hypothesis that dopamine neurons report reward prediction errors can be tested formally by paradigms developed by animal learning theory, using the Rescorla–Wagner learning rule. In the blocking test, a stimulus is not learned when it is paired with an already fully predicted reward, indicating the importance of prediction errors for learning. After pairing with a fully predicted reward, the blocked stimulus does not come to predict a reward. Accordingly, the absence of a reward following the blocked stimulus does not produce a response in dopamine neurons as no prediction error is elicited, and the delivery of a

reward does produce a positive prediction error response. By contrast, after a well-trained reward-predicting stimulus, reward omission produces a depressant neural response, and reward delivery does not lead to a dopamine response.

In the conditioned-inhibition paradigm, a test stimulus is presented simultaneously with an established reward-predicting stimulus, but no reward is given after the compound, making the test stimulus a conditioned inhibitor which predicts the absence of reward. Reward omission after a well-learned conditioned inhibitor does not produce a prediction error response in dopamine neurons, even when a well-learned reward-predicting stimulus is added. By contrast, the occurrence of reward after the inhibitor produces an enhanced dopamine response, as the prediction error represents the difference between the actual reward and the negative prediction from the inhibitor.

Taken together, the data from these formal tests suggest that dopamine neurons show bidirectional coding of reward prediction errors, described by the following equation:

$$\text{Dopamine response} = \text{reward occurred} - \text{reward predicted}$$

Thus the dopamine response seems to convey the crucial learning term (λ-V) of the Rescorla-Wagner learning rule and complies with the principal characteristics of teaching signals of efficient reinforcement models.

Neuronal computations using prediction errors may contribute to the self-organization of behavior (Figure 2). Brain mechanisms establish predictions, compare current inputs with predictions from previous experience, and emit a prediction error signal once a mismatch is detected. The error signal may act as an impulse for synaptic modifications that lead to subsequent changes in predictions and behavioral reactions. The process is reiterated until behavioral outcomes match the predictions and the prediction error becomes nil. In the absence of a prediction error, there would be no signal for modifying synapses, and synaptic transmission remains unchanged and stable.

Reward-Predicting Stimuli

Dopamine neurons acquire responses to reward-predicting visual and auditory stimuli that covary with the expected value of reward but occur irrespective of spatial position; sensory stimulus attributes; and arm, mouth, and eye movements. The responses are modulated by the motivation of the animal, the time course of predictions, and the animal's choice among rewards. Activations do not occur when the conditioned stimulus (CS) is predicted by another stimulus in time ranges of seconds. These responses conform to temporal difference learning models that conceptualize prediction errors irrespective of primary or conditioned reinforcers and view a CS response as reflecting an error in the prediction of the CS. Although discriminating between reward-predicting CSs and neutral stimuli, dopamine activations have a non-negligible propensity for generalization.

Physically intense stimuli induce activations in dopamine neurons which are enhanced by stimulus novelty, although smaller visual stimuli that become effective for driving dopamine neurons after pairing with reward do not induce saliency or novelty responses before conditioning, suggesting that novelty *per se* is not sufficient to activate dopamine neurons. As intense stimuli have attention-inducing and reinforcing functions, the dopamine responses to rewards themselves might be due to the attention-inducing rather than the reinforcing functions of the rewards. However, other strongly attention-inducing stimuli, including primary and conditioned aversive events, reward omission, and conditioned inhibitors, are rather ineffective in producing the characteristic, short-latency activation of dopamine neurons. Thus, the dopamine responses to rewards and intense novel stimuli do not appear to be due to a general

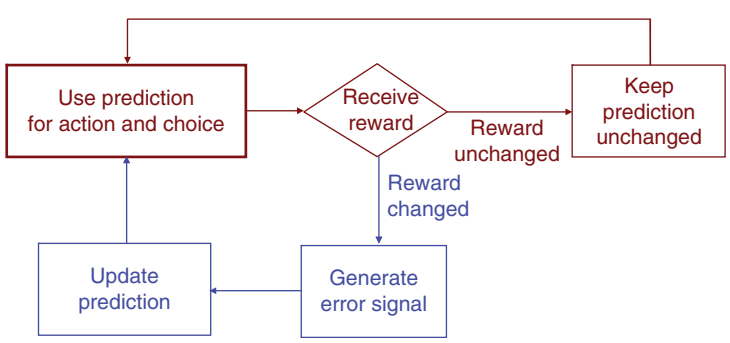

Figure 2 Learning viewed as change of action following change of prediction. The error signal would be transmitted by dopamine neurons and would influence synaptic processes in dopamine target structures such as striatum and orbitofrontal cortex.

alerting or attention-generating function of these stimuli. The occurrence of intensity responses may be due to a combined sensitivity of dopamine neurons to both rewarding and intense events, may reflect the known rewarding and approach-generating functions of intense stimuli, or may be related to less-well-known forms of attention attached to rewards and intense novel stimuli but not to punishers.

Uncertainty Signal in Dopamine Neurons

Rewards occur in most natural situations with some degree of uncertainty. The uncertainty of reward can be tested by delivering rewards with different probabilities and separating expected value (linearly increasing from $p = 0$ toward $p = 1$) from uncertainty expressed as the variance of the probability distribution of magnitudes (inverted U function with peak at $p = 0.5$). More than one-third of dopamine neurons show a relatively slow, sustained, and moderate activation between the reward-predicting stimulus and the reward which covaries with variance. This activation occurs in single trials in single neurons and does not propagate during learning from reward back to the conditioned stimulus, as assumed by some implementations of temporal difference reinforcement models. The uncertainty-related activation contrasts with the responses to reward-predicting stimuli and reward prediction errors that covary with expected value and occur with different time courses (Figure 3). The two responses are uncorrelated in strength across individual neurons. In a separate experiment, the delivery of two equiprobable rewards with different magnitudes influences the variance while keeping the entropy constant at 1 bit. Increasing variance in this procedure leads to an increased uncertainty response, suggesting that variance is an effective measure of uncertainty for dopamine neurons.

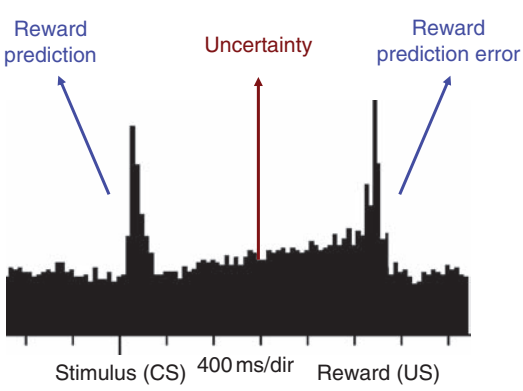

Figure 3 Separate uncertainty and reward signals in dopamine neurons. Data show averages responses from a population of dopamine neurons. CS, conditioned stimulus; US, unconditioned stimulus.

The distinct neural coding of reward value and uncertainty is consistent with the separation of expected utility into these two components suggested by financial decision theory and found in human brain imaging. However, the separation contrasts with the combined coding as scalar variable generally assumed by classic microeconomic expected utility theory.

Comparisons with Other Reward Centers

Information about rewards is processed also in a number of other brain structures and in different ways. Such reward signals are found in brain structures to which the axons of dopamine neurons project, including the orbitofrontal cortex, striatum, and amygdala. Rewards influence the action-related activity of neurons in additional structures, including the prefrontal and parietal cortex. Additional brain structures influenced by reward include the supplementary motor area in the frontal lobe, the rhinal cortex in the temporal lobe, the pallidum and subthalamic nucleus in the basal ganglia, and a few others. These activations are seen in neuronal recordings not only from behaving animals, mostly monkeys, but also in human brains, via functional magnetic resonance imaging.

Orbitofrontal Cortex

Orbitofrontal neurons show activations following reward-predicting stimuli, during the expectation of reward, and after reward reception. Orbitofrontal responses to rewards and reward-predicting stimuli are related to the motivational value rather than the more sensory properties of reward objects, as satiation with specific rewards reduces the neuronal responses. They constitute pure reward signals by reflecting only reward and not spatial or visual object features. Orbitofrontal reward signals distinguish between reward and punishment, change with reversal of stimulus–reward associations, discriminate between different volumes of liquid reward, and encode the economic value of rewards for decision making irrespective of the actual reward objects. Different neurons in this structure show more-sustained activations preceding the expected delivery of liquid or food reward. Besides these pure reward-related responses, a few other orbitofrontal neurons respond to visual or spatial object properties or are activated in relation to movements.

Orbitofrontal neurons do not appear to be specialized for particular reward objects but seem to discriminate between different rewards depending on their current availability. A reward that is effective for activating an orbitofrontal neuron may lose its

efficacy when the distribution of available rewards changes and the initially effective reward changes its level of preference. By encoding economic value rather than specific reward objects, these responses appear to adapt to the current probability distribution of reward values. A change in this distribution changes the neuronal responses. The apparent dependence of responsiveness on a set point corresponds to a basic tenet of prospect theory indicating that outcomes are valued relative to movable references rather than absolute physical characteristics.

Pure Reward Signals in Striatum and Nucleus Accumbens

The slowly firing medium spiny neurons in striatum and nucleus accumbens and the tonically active striatal neurons (TANs) respond to the reception of food and liquid rewards. Other striatal and accumbal neurons show phasic activations following visual reward-predicting stimuli and more-sustained activations during the expectation of rewards. Changes of existing reward expectation during learning lead to adaptations of reward expectation-related activity to the currently valid expectation in parallel with the animal's behavioral differentiation. The TANs discriminate between rewards and air puff punishers, and many slowly firing striatal neurons distinguish reward from no reward and discriminate between different rewards and reward magnitudes irrespective of visual object properties, spatial information, and movements. Neurons in the ventral striatum show a higher incidence of reward responses and reward expectation activities compared with caudate and putamen neurons and, their larger spectrum of task-related activity. Thus, subpopulations of striatal neurons appear to process pure reward signals.

Reward Influences on Action-Related Activity in Striatum

Similar to prefrontal neurons, the action-related activity of a population of neurons in the striatum (caudate and putamen) is influenced by predicted rewards. These neurons are activated during the preparation and execution of specific arm and eye movements toward different spatial targets and discriminate between movement and nonmovement reactions. At the same time, these action-related activities are differentially influenced by the predicted presence versus absence of reward and by different predicted types, magnitudes, and probabilities of reward. This activity can predict the animal's choice toward a rewarding outcome. The reward-differentiating nature of the activations develops and adapts during learning while differential reward expectations are being acquired. Similar specific action–reward activities exist also in another basal ganglia structure, the subthalamic nucleus.

The activations in the striatum and cortex mentioned above do not simply represent outcome expectations; they differentiate in addition between different behavioral reactions for the same outcome (e.g., movement vs. nonmovement, different spatial targets), and they not only reflect different behavioral reactions but differentiate also between the expected outcomes. The combined action and reward coding by striatal neurons complies with theoretical notions of action values by which specific behavioral actions become associated with a rewarding outcome through operant learning. These activities may constitute neuronal correlates of goal-directed behavior as they appear to reflect neuronal representations of the reward for the specific action while this action is being prepared and executed.

The combined coding of action and reward contrasts with the earlier described pure reward signals in dopamine neurons and in some neurons of orbitofrontal cortex and striatum, which reflect the predicted or received reward irrespective of other stimulus or behavioral components. In demonstrating the influence of predicted reward on action-related activity, action–outcome coding may represent the next processing stage downstream from pure reward signals toward overt choices. Action–outcome coding may be a component of mechanisms by which reward signals are translated into behavioral choices for obtaining reward through action. Information about the value of each possible action in choice situations would constitute important inputs for decision-making mechanisms.

Prefrontal Cortex

In addition to generating specific signals, rewards also influence ongoing action-related activity. The prediction of different food or liquid rewards modifies the typical, spatially discriminating delay activity of neurons in dorsolateral prefrontal cortex and influences movement-specific cue responses in medial prefrontal cortex. These prefrontal neurons carry signals related to the memory of movement cues and the preparation of movement and at the same time encode the expected reward. Only a small population of prefrontal neurons is preferentially activated by aversive stimuli.

Other Cortical Areas

Predicted rewards also influence activity related to arm and eye movement in other cortical areas, including parietal cortex and anterior and posterior

cingulate. Similar reward effects in premotor and supplementary motor cortex may reflect the motivating functions of rewards on movements coded in this part of the motor system.

Amygdala

In rats, reward-predicting stimuli and unpredicted liquid, food, cocaine, and intracranial electrical stimulation elicit responses in central and basolateral amygdala, the responses being differentiated against aversive and neutral outcomes. Responses discriminate between reward magnitudes and change with outcome reversal. Responses correlate with orbitofrontal responses during early discrimination learning and decrease after orbitofrontal lesions. Reward studies on monkeys demonstrate satiety-sensitive gustatory responses and responses to liquid or food-predicting visual stimuli which decrease with increasing behavioral requirements.

See also: Basal Ganglia: Internal Organization; Basal Ganglia: Physiological Circuits; Dopamine – CNS Pathways and Neurophysiology; Dopamine: Cellular Actions; Representation of Reward; Reward and Learning; Reward Decision-Making; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex; Reward Processing: Human Imaging; Transcription and Reward Systems.

Further Reading

Critchley HG and Rolls ET (1996) Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. *Journal of Neurophysiology* 75: 1673–1686.

Fiorillo CD, Tobler PN, and Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299: 1898–1902.

Hikosaka O, Sakamoto M, and Usui S (1989) Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *Journal of Neurophysiology* 61: 814–832.

Hollerman JR, Tremblay L, and Schultz W (1998) Influence of reward expectation on behavior-related neuronal activity in primate striatum. *Journal of Neurophysiology* 80: 947–963.

Kahneman D and Tversky A (1984) Choices, values, and frames. *American Psychologist* 4: 341–350.

Lauwereyns J, Watanabe K, Coe B, and Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. *Nature* 418: 413–417.

Matsumoto K, Suzuki W, and Tanaka K (2003) Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science* 301: 229–232.

Padoa-Schioppa C and Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441: 223–226.

Paton JJ, Belova MA, Morrison SE, and Salzman CD (2006) The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* 439: 865–870.

Ravel S, Legallet E, and Apicella P (2003) Responses of tonically active neurons in the monkey striatum discriminate between motivationally opposing stimuli. *Journal of Neuroscience* 23: 8489–8497.

Rescorla RA and Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH and Prokasy WF (eds.) *Classical Conditioning II: Current Research and Theory*, pp. 64–99. New York: Appleton Century Crofts.

Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology* 57: 87–115.

Schultz W, Dayan P, and Montague RR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

Sutton RS and Barto AG (1998) *Reinforcement Learning*. Cambridge, MA: MIT Press.

von Neumann J and Morgenstern O (1944) *The Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Waelti P, Dickinson A, and Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412: 43–48.

Watanabe M (1996) Reward expectancy in primate prefrontal neurons. *Nature* 382: 629–632.

Dopamine Receptors and Antipsychotic Drugs in Health and Disease

P Seeman, University of Toronto,Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The discovery of dopamine receptors is intertwined with the discovery and development of antipsychotic drugs. The research in this area started with the development of antihistamines after World War II, particularly with H Laborit using such compounds to enhance surgical analgesia. In patients receiving these medications, Laborit noticed a 'euphoric quietude,' and that the patients were "calm and somnolent, with a relaxed and detached expression." Of this series of Rhône-Poulenc compounds, RP4560, now known as chlorpromazine, was the most potent.

Chlorpromazine was tested by many French physicians for use in various medical illnesses. Although Sigwald and Bouttier were the first to use it as the sole medication for a psychotic patient, their work was not reported until 1953, after a 1952 report by Delay and colleagues that chlorpromazine alleviated hallucinations and stopped internal 'voices' in eight patients. A significant aspect of the action of chlorpromazine was that it was effective within 3 days. This rapid improvement, especially during the first week of antipsychotic treatment, has been observed in many studies and summarized in reviews by S Kapur and O Agid.

The clinically successful action of chlorpromazine stimulated the search to identify chlorpromazine's mode of action. The assumption then, as now, was that the discovery of such a mode of action would open the avenue to uncovering the biochemical cause of psychosis and possibly of schizophrenia.

Early Days: Before Discovery of Dopamine Receptors

In searching for the mechanism of chlorpromazine action in the 1960s and 1970s, many types of electrophysiological and biochemical experiments were done. Because high doses of chlorpromazine and other antipsychotics (or 'neuroleptics,' as they were then called) also elicited parkinsonism as a side effect, the basic science quickly focused on the action of antipsychotics on dopamine pathways in the brain. The rationale for examining brain dopamine regions was based on the finding by H Ehringer and O Horniekiewicz that the parkinsonism of Parkinson's disease was associated with a massive loss of brain dopamine. It was felt, therefore, that the unwanted side effect of chlorpromazine-induced parkinsonism, as well as the antipsychotic action itself, might arise by antipsychotics interfering with dopamine neurotransmission. The working assumption was that if the brain targets for antipsychotics could be found, then perhaps it could be determined whether these sites were overactive or underactive in psychosis or schizophrenia.

A variety of mechanisms were explored for the mode of action of chlorpromazine, including its action on mitochondrial enzymes, sodium–potassium-ATPase, and related enzymes, and its membrane-stabilizing action, such as its strong potency to inhibit membrane action potentials and to stabilize cellular and subcellular membranes from releasing their contents. It also became clear in 1963 that all antipsychotics were surface active, readily explaining their hydrophobic affinity for membranes. Some of these nonreceptor-related findings, such as the surface activities of the antipsychotics, showed an astonishingly excellent correlation with clinical antipsychotic potencies.

Therapeutic Concentrations of Antipsychotics

All of the early experiments in the 1960s revealed that the *in vitro* active concentrations of the antipsychotics were generally between 20 and 1000 nM. These concentrations, however, were found in 1971 to be far in excess of the nanomolar concentrations (e.g., 1–2 nM for haloperidol) that exist in the spinal fluid in patients being successfully treated with these medications.

In Vivo Experiments

In parallel with the *in vitro* experiments, there were many *in vivo* experiments by F Bloom, by G Aghajanian, and by B Bunney, showing that dopamine agonists can excite or inhibit neurons in the nigrostriatal dopamine pathway. Moreover, other workers (WD Heiss, J Hoyer) showed that direct application of dopamine on neurons also stimulated or inhibited snail neurons, and that haloperidol or fluphenazine could block these actions (H Struyker Boudier). These studies provided evidence for the existence of distinct dopamine receptors on neurons.

Additional work *in vivo* showed that chlorpromazine and haloperidol increased the production of normetanephrine and methoxytyramine, metabolites of epinephrine and dopamine, respectively. To explain the increased production of these metabolites,

Carlsson and Lindqvist suggested that "the most likely [mechanism] appears to be that chlorpromazine and haloperidol block monoaminergic receptors in brain; as is well known, they block the effects of accumulated 5-hydroxytryptamine.…" In other words, they proposed that antipsychotics might block all three types of receptors for noradrenaline, dopamine, and serotonin, but they did not identify which receptor was selectively blocked or how to identify or test any of these receptors directly *in vitro*. This study in 1963 by Carlsson and Lindqvist is often mistakenly cited as discovering 'the dopamine receptor' and that antipsychotics are selectively acting on this receptor. However, N-E Andén, a student of A Carlsson, had a different view, and proposed that "chlorpromazine and haloperidol delay the elimination of the (metabolites).…" Moreover, 7 years later Andén reported that antipsychotics increased the turnover of both dopamine and noradrenaline, but he could not show that the antipsychotics were selective in blocking dopamine; for example, chlorpromazine enhanced the turnover of noradrenaline and dopamine equally. Therefore, it remained for *in vitro* radio-receptor assays to detect the dopamine receptor directly and to demonstrate antipsychotic selectivity for the dopamine receptor.

The Dopamine D1 Receptor

With the advent of assays for adenylate cyclase in the 1960s, J Kebabian found that dopamine stimulated adenylate cyclase in the superior cervical ganglion. This receptor was later named the dopamine D1 receptor, selectively labeled by [^{3}H]SCH23390, and subsequently cloned by three research groups in 1990.

The dissociation constants at D1 for dopamine agonists and antagonists of medical therapeutic interest are given in Tables 1 and 2. There is no correlation between the antipsychotic clinical doses and the dissociation constants of the antipsychotic antagonists at D1, as illustrated in Figure 1. These data suggested that D1 was not the major or common target for antipsychotics, in addition to the fact that the antipsychotic molarities at D1 are between 10 and 10 000 nM, far in excess of the therapeutic concentrations in the spinal fluid of treated patients.

In addition to the lack of targeting D1 receptors by clinical doses of the common antipsychotics, D1-selective compounds have not been found to be effective as antipsychotics (Figure 2).

Discovery of the Antipsychotic Dopamine Receptor, or the Dopamine D2 Receptor

In 1974 and 1975, in order to detect and discover the dopamine receptors on which the antipsychotics presumably acted, it was essential to label a receptor with a ligand, such as radioactive haloperidol, having an affinity (or dissociation constant) of ~1 nM, because this was the haloperidol therapeutic concentration found in the spinal fluid or plasma water of treated patients. For this to occur, the specific activity of [^{3}H]haloperidol would have to be at least 10 Ci mmol^{-1}. Although the [^{3}H]haloperidol donated in 1971 by Janssen Pharmaceutica (J Heykants, J Brugmans) had a low specific activity of 32–71 mCi mmol^{-1}, I R E Belgique (M Winand) custom synthesized [^{3}H]haloperidol at even lower specific activity (10.5 Ci mmol^{-1}) for Seeman's laboratory by June 1974. Specific binding of this new [^{3}H]haloperidol to brain striatal tissue was readily detected in 1975, and the following concentrations of compounds were found to inhibit the binding of [^{3}H]haloperidol by 50%: 2 nM for haloperidol, 20 nM for chlorpromazine, 3 nM for (+)butaclamol, and 10 000 nM for (–) butaclamol.

The stereoselective action of butaclamol and the good correlation between the $IC_{50\%}$ values and the clinical doses, as shown in Figure 3, indicated that the 'antipsychotic receptor' had finally been discovered. Equally important, of the endogenous compounds tested, dopamine was the most potent in inhibiting the binding of [^{3}H]haloperidol, indicating that the antipsychotic receptor was actually a dopamine receptor (see Table 1). Using sequences related to the β-adrenoceptor, the antipsychotic/dopamine receptor, now named the dopamine D2 receptor, was finally cloned in 1988; its amino acid sequence is shown in Figure 4. Using the cloned D2 receptor, values for the dissociation constants for agonists and antagonists can be determined (Table 1).

It is now known that therapeutic levels of antipsychotic drugs occupy 60–80% of brain D2 receptors, as shown by L Farde and colleagues. In fact, using the antagonist dissociation constants in Table 1, the concentrations necessary to occupy 75% of D2 receptors can be calculated and are essentially identical to the free molarities of the antipsychotics found in either the plasma water or the spinal fluid of treated patients, as shown in Figure 5.

Because radiolabeled raclopride is less tightly bound to D2 receptors than radiolabeled spiperone, which in turn is less tightly bound to D2 receptors than radiolabeled nemonapride, the dissociation constant for any antagonist at D2 receptors is lowest when using radiolabeled raclopride, but rises when using radiolabeled spiperone, or rises higher still when using radiolabeled nemonapride (see Table 1). This dependence of the dissociation constant on the ligand is also seen in positron emission tomography studies, where it has been found, for example,

Table 1 Agonist potencies at dopamine receptors (in 120 mM NaCl)

	D1		*D2*		*D3*		*D4*		*D5*
Tissue [Ref.]:	*Rat striatum or human D1 [f]*		*Human D2 Long clone [b]*		*Human D3 clone [b]*		*Human D4 clone*		*Human D5 clone*
Radioligand:	*[^{3}H]SCH 23390*		*[^{3}H]SCH 23390*		*[^{3}H]SCH 23390*		*[^{3}H]spiperone*		*[^{3}H]SCH 23390*
	K^{High} (*nM*)	K_{50} (*nM*)	K^{High} (*nM*)	K_{Low} (*nM*)	K^{High} (*nM*)	K_{50} (*nM*)	K^{High} (*nM*)	K_{50} (*nM*)	K_{50} (*nM*)
Neurotransmitters									
Dopamine	10.1	2340 [g]	2.4 D 1.4 R;1.3 A	1400 D 729 R	25 D 3.9 I [a]	25 I [a] ~22 S	28 [f]	148 [f]	228 [g]
Noradrenaline	No high	700	9.8 D	n.i.	n.i.	n.i.	n.i.	2200 [f]	12 000 [g]
Adrenaline	No high	1700	No high	805 D	n.i.	n.i.	n.i.	n.i.	n.i.
Serotonin	No high	9700 [g]	No high	4300 D	n.i.	n.i.	n.i.	4180 [f]	3000 [g]
Dopamine agonists									
N-Propyl-norapomorphine-R-(−)	1.2	1816 [g]	0.14 D	54 D rat	0.25 D	~0.8	n.i.	4 f	1136 [g]
(+)PHNO	39	n.i.	0.4 D	62 D	0.6 D	n.i.	n.i.	47 f	n.i.
(±)Quinagolide.HCl, or CV205 502	No high	5400	0.5 D	1400 D	3.2 D	n.i.	n.i.	n.i.	n.i.
Bromocriptine.base	No high	672 [g]	0.8 D	20 D	1.3 D	7.4 I [a]	n.i.	250 [e]	720 [e]
Apomorphine-R-(−).HCl	3.5	360 [e]	0.75 D	140 D	2.6 D	73 I [a]	4.1 [f]	5 [e]	13 [e]
Pergolide mesylate [LY 127,809]	No high	330 [e]	1.3 D	19 I [a]	0.9 D	2.3 I [a]	n.i.	62 [e]	39 [e]
(–)Quinpirole, or (−)-LY 171,555.HCl	No high	2600	1.7 D	202 R[d]	0.45 R[d]	256R [d];50 I [e]	n.i.	31 [e]	>1000 [g]
Pramipexole monohydrate	No high	8200	12 D	2600 D	9 D	11 I [e]	n.i.	120 [e]	>1000 [e]
SKF 38393	1.1 [c]	~380 [c]	~150 S[c]	~8800 S[c]	No high	5000 [a]	n.i.	1800 [f]	100 [g]
(+)-7-OH-dipropylaminotetralin	10 [c]	~4700 [c]	1. 1 R [d]	42 R [d]	0.06 R [d]	0.2 R [d]	n.i.	650 [a]	n.i.

[a], Sokoloff et al. (1992); [b], Seeman et al. (2005); [c], Seeman and Niznik (1988); [d], Malmberg and Mohell (1995); [e], Millan et al. (2002); [f] P. Seeman; [g], Sunahara et al. (1991); K_{50} (dissociation constant), C50% (conc. to inhibit 50% of binding)/(1 + C*/K_d), where C* – concentration of radioligand in competition with agonist, and K_d – dissociation constant obtained by saturation with radioligand (Scatchard analysis); A, [^{3}H]dopamine (K_d = 1.3 nM); D, [^{3}H]domperidone (Ref. b or P. Seeman, unpublished; K_d = 0.43 nM); I, [^{125}I]Iodosulpride (K_d = 0.6 nM); R, [^{3}H]raclopride (K_d = 1.9 nM); S, [^{3}H]spiperone (K_d = 65 pM); rat, rat striatum; no high, no high-affinity state recognized by competing compound; n.i., no information available.

Table 2 K values (dissociation constants)

Human clone:	*M1 nM*	*D1 nM*	*D2 nM*	*D2 nM*	*D2 nM*	*D3 nM*	*D4 nM*	*D4 nM*
[^{3}H]ligand used:	*QNB*	*Sch.*		*Raclo.*	*Spip.*	*Raclo.*		*Spip.*
[^{3}H]Amisulpride			1.7	1.8	4.6		—	
Amoxapine	49		—	21	56		—	5.5
Aripiprazole			—	1.8				
Butaclamol-(+)			—	0.14	0.9		—	70
Chlorpromazine	3.4	16.5	—	1.2	4.6	1.4	—	9.6
[^{3}H] Chlorpromazine			0.77	—	—	—	1.2	—
Clozapine	9.5	90	—	75	180	190	—	22
[^{3}H]Clozapine			51	—	—	—	2	—
Clozapine-*iso*	14.4	120	—	15	60	21	—	21
Cyproheptadine			—	24			—	
Droperidol			—	0.54	2.3		—	2
Epidepride			—	0.036	0.06		—	
Flupentixol-*cis*		1.8	—	0.38	0.7	0.7	—	15
Flupentixol-*trans*				151				
Fluphenazine		2.6	—	0.55	1.2	0.17	—	30
Haloperidol	794	55	—	0.74	2.7	8.8	—	2
[^{3}H]Haloperidol			0.4	—	—	—	0.85	—
Iloperidone (HP873)	109	5.6	—	5.4	10	20	—	9.6
Loxapine	117	18	—	9.2	22.7	7.2		8
Loxapine-*iso*	203	16	—	22	86	18		11
Melperone		148	—	152	375	315	—	720
Molindone		1558	—	4.9	15	44	—	3900
Norclozapine	17	73	—	180	300		—	120
Olanzapine	2.1	9.2	—	7.4	21	14	—	15
[^{3}H]Olanzapine			2.7	—	—	—	1.6	—
Perphenazine		4.5	—	0.27	0.47	0.23	—	32
Pimozide	470		—	1.4	0.95			
Prochlorperazine		7.7	—	1.7	4		—	89
Quetiapine	135	290	—	140	680	240	—	2000
[^{3}H]Quetiapine			104	—	—	—	—	—
Raclopride			—	1.6	7.1	2.9	—	2400
[^{3}H]Raclopride			1.9	—	—	1.6	—	—
Remoxipride	>10 μM	4900	—	67	800	960	—	2400
Risperidone	>10 μM	42	—	1.09	4	3.5	—	4.4
Risperidone-9-OH				1.6				
Sertindole	400	22	3	1.9	6.5	3	—	11
[^{3}H]Sertindole			1.2	—	—	—	0.85	—
Spiperone			—	0.018	0.06		—	
[^{3}H]Spiperone			0.065	—	—	0.32	0.086	—
Sulpiride-*S*			—	9.9	8	10	—	1000
Trifluperazine		2.9	—	1.4	3.8	0.7	—	39
Ziprasidone	265	9	—	2.7	6	1.5	—	8

Blank cells indicate 'not done'.
Sch., Schering 23390; Raclo., Raclopride; Spip., spiperone.

when monitoring patients with [^{11}C]raclopride, that therapeutic doses of clozapine occupy ~50% of D2 receptors but that there is much less occupancy of D2 receptors when using [^{11}C]methylspiperone or [^{18}F]fluorethylspiperone, both of which bind more tightly to D2 receptors than raclopride (see Table 2).

The correlations in Figures 3 and 5 remain a cornerstone of the dopamine hypothesis of psychosis or schizophrenia, and the dopamine hypothesis is still the major contender for an explanatory theory of schizophrenia etiology.

Two Classes of Dopamine Receptors

The D1 site and the [^{3}H]haloperidol/dopamine receptor binding site were soon considered as distinct, because B Roufogalis found that the sulpiride antipsychotic did not block dopamine-stimulated adenylate cyclase. Two general classes of dopamine receptors were recognized,

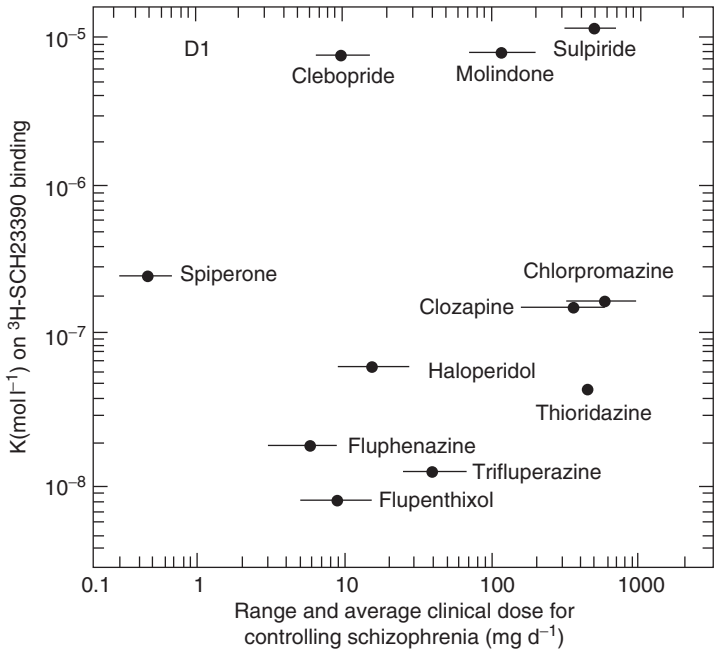

Figure 1 There is no correlation between the clinical antipsychotic doses and the antipsychotic dissociation constants (or concentrations) that inhibit the binding of a D1 ligand ([^{3}H]SCH23390) at dopamine D1 receptors in homogenized striatal tissue. The high concentrations inhibiting the D1 receptor are far higher than those found clinically in the plasma water or spinal fluid. Adapted from Seeman P (1987) Dopamine receptors and the dopamine hypothesis of schizophrenia. *Synapse* 1: 133–152, with permission from John Wiley & Sons Inc.

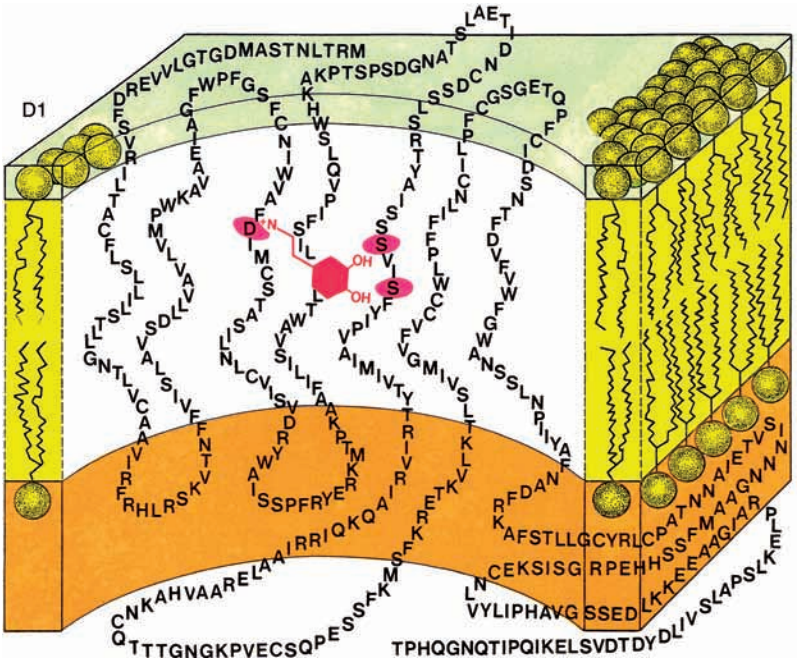

Figure 2 Amino acid sequence of the dopamine D1 receptor. The two hydroxyls of dopamine are presumed to be associated with the two serine residues, while the tertiary nitrogen atom is presumed to be associated with the aspartic acid residue (D in transmembrane 3).

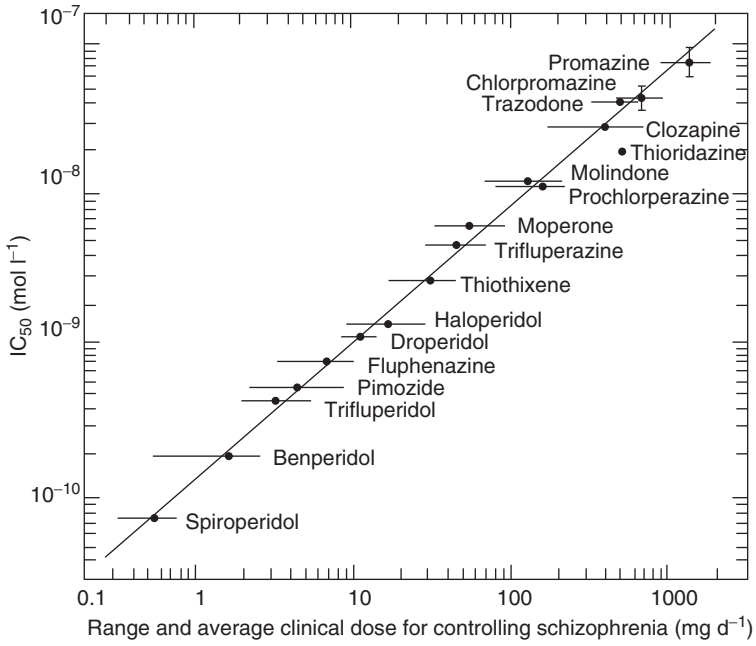

Figure 3 The clinical antipsychotic doses correlate with the concentrations that inhibit by 50% the specific binding of [^{3}H]haloperidol in homogenized caudate nucleus tissue (calf). These concentrations are similar to those found in the plasma water or spinal fluid in patients treated with antipsychotic drugs. Adapted from Seeman P, Lee T, Chau-Wong M, et al. (1976) Antipsychotic drug doses and neuroleptic/dopamine receptors. *Nature (London)* 261: 717–719, with permission from Nature.

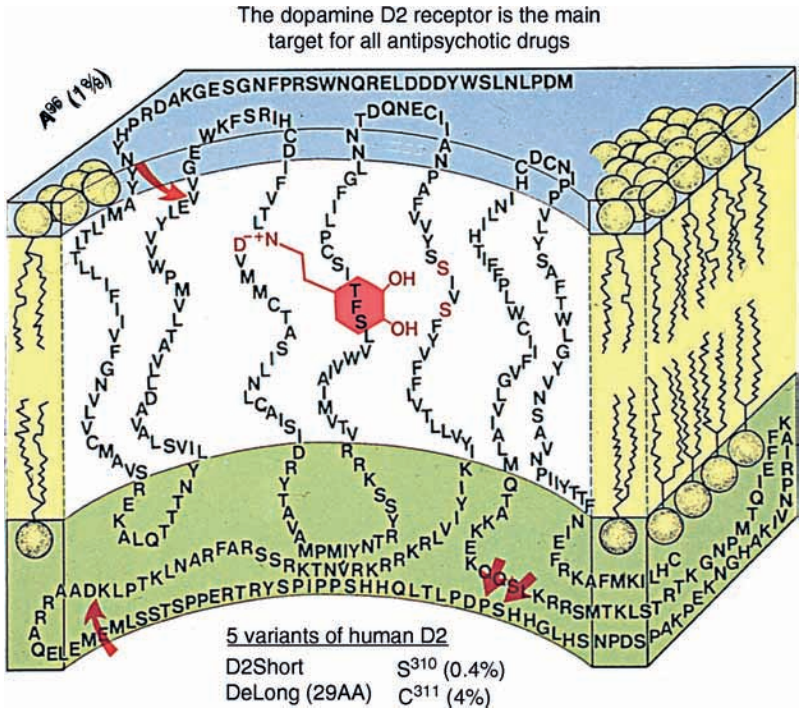

Figure 4 Amino acid sequence of the dopamine D2 receptor. The variants or polymorphisms of D2 include an alanine instead of a valine at position 96 (~1% of population), an insertion of a 29-amino-acid polypeptide (D2Long) at the position shown by the lower left arrow, a serine instead of a proline at position 310, and a cysteine instead of a serine at position 311. Dopamine is presumed to be associated with the two serines (S) in transmembrane 5 and the aspartic acid (D) in transmembrane 3 (see **Figure 2**).

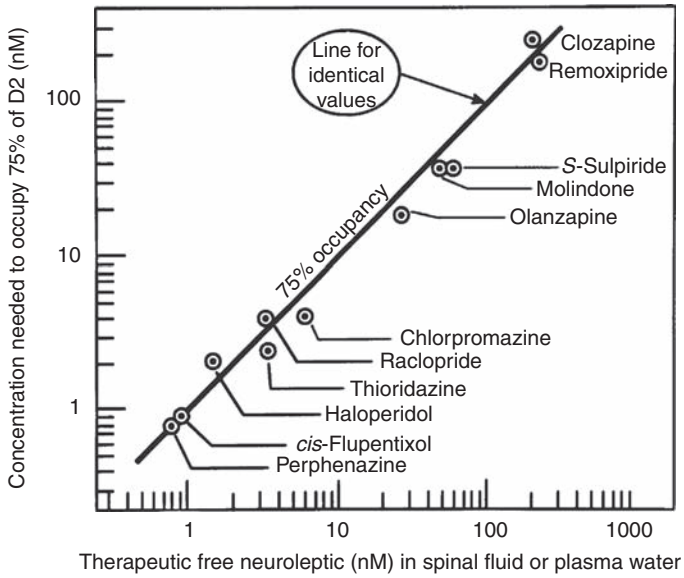

Figure 5 The therapeutic antipsychotic concentrations in the spinal fluid or in the plasma water in treated patients are essentially identical to the antipsychotic concentrations that occupy approximately 75% of the D2 receptors *in vitro*. The concentrations in the plasma water were obtained by correcting for the amount bound to the plasma proteins. The concentrations to occupy 75% of D2 were calculated as being three times higher than the dissociation constant at D2. Adapted from Seeman P (2002) Atypical antipsychotics: Mechanism of action. *Canadian Journal of Psychiatry* 47: 27–38.

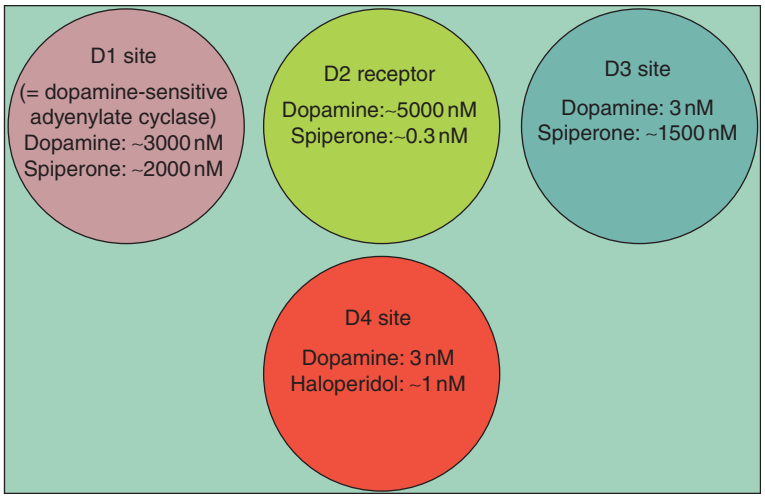

Figure 6 Early version of dopamine receptors before clones of receptors became available. The D1 receptor (dopamine-stimulated adenylate cyclase) was stimulated by ~3000 nM dopamine and inhibited by high concentrations of butyrophenones such as 2000 nM spiperone. The D2 receptor, or the [^{3}H]haloperidol binding site, was highly sensitive to spiperone, but required ~5000 nM dopamine to inhibit adenylate cyclase. The D3 site was a site labeled by [^{3}H]dopamine and, therefore, very sensitive to dopamine at 3 nM, but required very high concentrations of antipsychotics to be inhibited. The D4 site was defined as being sensitive to both the agonists and the antipsychotic antagonists. While these definitions for the D3 and D4 sites are no longer used, the D1 and D2 properties are still valid for the cloned D1 and D2 receptors, with the additional point being that both D1 and D2 have high- and low-affinity states.

therefore, coupled or uncoupled to adenylate cyclase. These two classes were named D1 and D2 by J Kebabian and D Calne.

Nomenclature of Dopamine Receptors

The data for the pattern of binding of [^{3}H]haloperidol identifying the antipsychotic/dopamine D2 receptor were very different from those for the pattern of [^{3}H] dopamine binding described by studies in the mid-1970s. For example, the binding of [^{3}H]haloperidol was inhibited by ~5000 nM dopamine, while that of [^{3}H]dopamine was inhibited by ~3 nM dopamine, as summarized in Figure 6. For several years, this latter [^{3}H]dopamine binding site was termed the 'D3 site,' a

term which is not to be confused with the later discovery of the dopamine D3 receptor.

Dopamine D2 Receptor Variants

As noted in Figure 4, there are several variants of D2, the most important of which are the short form and the long form of D2, the latter having an additional 29 amino acids. There is also a D2Longer form where a dipeptide, valine–glutamine, is inserted into the intracellular loop, as shown in Figure 7.

There are at least three polymorphisms in D2 (Figure 4): alanine replaces valine at position 96 in about 0.8–1% of some populations, serine replaces proline at position 310 in ~0.4% of people, and cysteine replaces serine at position 311 in approximately 3–4% of the population. The variants at 310 and 311 are markedly less effective in inhibiting the synthesis of cyclic AMP than is the more common form of D2. Silent polymorphisms have also been found in the DNA code for D2, but there is no change in the amino acid (leucine at 141, histidine at 313, and proline at 319).

There are also polymorphisms in the noncoding regions of D2, including an A-to-G at position –241, a C insert at position –141, an A-to-G before transmembrane 1 (*Taq*1B polymorphism), an A-to-G in the intron within transmembrane 2 (*Taq*1D polymorphism), an A-to-C in the intron before transmembrane 4, a G-to-A in the intron before transmembrane 6, and an A-to-G situated ~10 kb beyond transmembrane 7 (*Taq*1A polymorphism). A further mutation in D2 occurs in individuals with hereditary autosomal dominant myoclonus dystonia, with valine replaced by isoleucine at position 154 at the beginning of the fourth transmembrane segment of D2 (see Figures 4 and 7).

D2 Function and Distribution

A wide variety of psychomotor functions, including biochemical, physiological, and pathological, have been attributed to D2. Specifically, D2 inhibits action potentials by eliciting a prolonged 'inhibitory postsynaptic potential' (IPSP), inhibits adenylate cyclase, and inhibits the entry of calcium ions into cells, thereby inhibiting many aspects of stimulus–response coupling in a variety of neurons and cells. D2 receptors are located on presynaptic terminals and, therefore, can readily inhibit the release of dopamine. These presynaptic receptors appear to be predominantly D2Short, while D2Long receptors are mostly postsynaptically located on dendritic spines. The genetic absence of D2 receptors leads to animals that are akinetic and Parkinson-like.

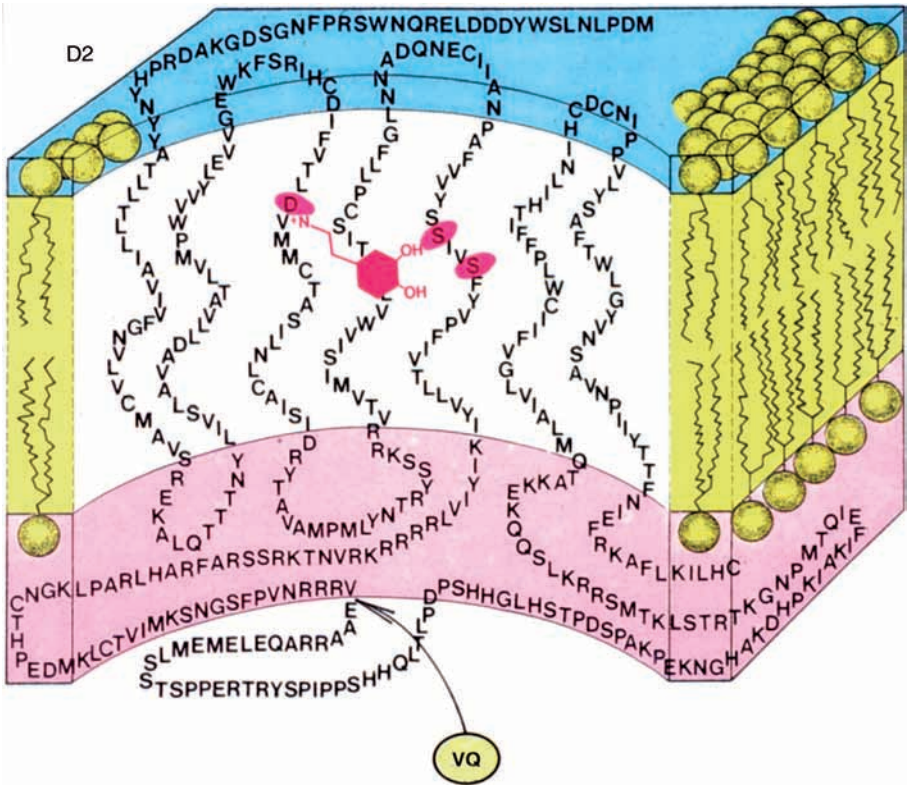

Figure 7 Amino acid sequence of D2Longer. Compared to D2Short and D2Long (Figure 4), D2Longer has an extra valine–glutamine dipeptide (as indicated by the arrow) that is usually spliced out in D2Short and D2Long.

In alcoholics, it has been reported that the prevalence of the *Taq*1A polymorphism is about twofold higher than in control subjects. In schizophrenia, it has been found in 27 studies, comprising 3707 patients and 5363 controls, that the serine311cysteine polymorphism was significantly associated with schizophrenia. Furthermore, the number of D2 receptors in the caudate-putamen is elevated in schizophrenia, as measured *in vivo* (Corripio) or in postmortem samples *in vitro* (Figure 8). Although Figure 8 shows that the density of D2 receptors in postmortem human schizophrenia tissues is elevated, some of this elevation may have resulted from the antipsychotic administered during the lifetime of the patient. The postmortem tissues from half of the patients with schizophrenia revealed elevated densities of [^{3}H]spiperone-labeled D2 receptors in the caudate-putamen tissue. The other half of the postmortem schizophrenia tissues were normal in D2 density, even though most of the patients were known to have been treated with antipsychotics during their lifetime.

Such findings have long been controversial, because the D2 density is not elevated in schizophrenia when using [^{11}C]raclopride. It should be noted, however, that the number of D2 receptors is significantly elevated in healthy identical co-twins of individuals with schizophrenia, suggesting that an elevation of D2 may be a necessary but not sufficient requirement for schizophrenia.

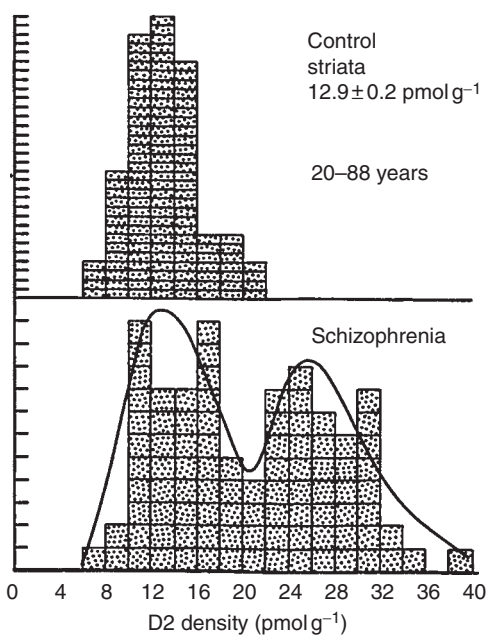

Figure 8 Bimodal pattern of D2 receptors in postmortem striata from individuals who had schizophrenia during life. The control individuals had died of nonneurological disorders. The bimodal pattern was found in the schizophrenia tissues, regardless of whether the individuals had received antipsychotics or not. Each square indicates a different postmortem human brain (caudate nucleus or putamen regions). Adapted from Seeman P and Niznik HB (1990) Dopamine receptors and transporters in Parkinson's disease and schizophrenia. *FASEB Journal* 4: 2737–2744, with permission from Federation of American Societies for Experimental Biology.

The distribution of D2 receptors within the various brain regions is reflected in the gene expression pattern of D2, as shown in Figure 9. An additional unique feature is that D2 receptors are organized in bands in the normal human temporal cortex, but these bands are not found in brains of patients with Alzheimer's disease. Finally, the density of D2 is not constant over one's lifetime, but slowly falls by about 2% per decade, as shown in the postmortem human tissues in Figure 10. There is a sharp transient rise during ages 1–3, but this is followed by a gradual pruning of neurons with D2 receptors.

The Dopamine D3 Receptor

Using methods similar to those used for cloning the D2 receptor, the D3 receptor was cloned in 1990; its sequence is shown in Figure 11. D3 has several polymorphisms, including a serine replacing glycine at position 9 in ~28% of the population (see Figure 11). In addition, there are nonfunctional forms of D3, where the amino acid chain stops after transmembrane 2, after transmembrane 3, or before transmembrane 6, the latter being found in, but not diagnostic for, Alzheimer's disease and schizophrenia.

Although BP897 is a partial agonist at D3, with a selectivity for D3 of about 100-fold higher than that for D2, this drug (at 10 mg day^{-1}) did not appear effective against schizophrenia symptoms. Other drugs moderately selective for D3, such as S33138 and A437,203, are currently being tested in schizophrenia patients. The highly D3-selective drug, FAUC 365, has not yet been tested in this disease. It is possible that the D3-selective compounds may be helpful in treating drug abuse.

The Dopamine D4 Receptor

The dopamine D4 receptor was cloned in 1999; its sequence is shown in Figure 12. The D4 receptor probably has more polymorphisms than any other protein in the body. For example, the intracellular loop contains repeat forms of a 16-amino acid polypeptide, the number of repeats varying from person to person. Most people have four such repeats, but up to ten repeats are known. Moreover, the precise sequence within each repeat usually varies from person to person, with at least 20 different types of repeat units known, thereby resulting in a massive number of polymorphisms in the human population.

An unusual polymorphism in D4 occurs at position 194, where glycine replaces valine in 13% of Africans

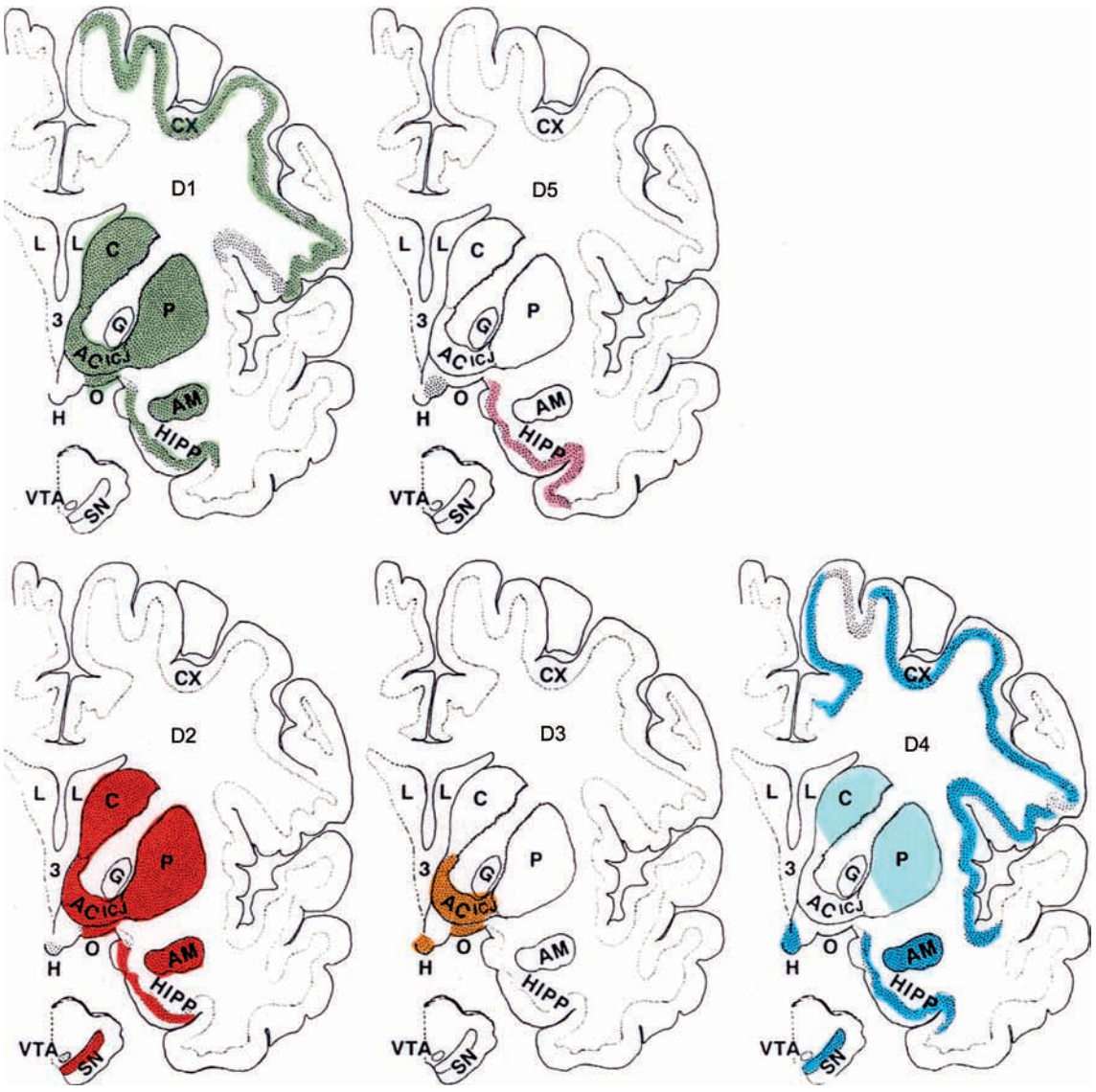

Figure 9 Anatomical location for gene expression of dopamine receptor genes in human brain. abbreviations: CX, cerebral cortex; L, lateral ventricles; C, caudate nucleus; P, putamen; G, globus pallidus; AC, nucleus accumbens; O, olfactory tubercle; H, hypothalamus; AM, amygdala; Hipp, hippocampus; VTA, ventral tegmental area; SN, substantia nigra; ICG, Islands of Calleja.

and Caribbeans (Figure 13), but not in Caucasians. This mutant form of D4 markedly reduces its sensitivity to dopamine. One young man in the Caribbean was found to be homozygous for this V194G polymorphism, but no medical abnormalities were found. Another D4 polymorphism occurs when the GASA sequence is missing at position 21.

Interestingly, clozapine has a higher affinity at D4 than at D2, as shown in Table 2. Nevertheless, despite clozapine's selectivity for D4, clozapine occupies the necessary 60–70% of brain D2 receptors at clinical doses (~400 mg day^{-1}), compatible with the idea that D2 is the therapeutic target for clozapine, as with all the other antipsychotics. It may be noted that isoclozapine causes catalepsy, in contrast to clozapine, which does not elicit catalepsy. Both drugs have dentical affinity for D4, but isoclozapine has higher affinity for D2 (see Table 2), and, therefore, causes catalepsy.

Although the gene expression of D4 was found to be elevated in the frontal cortex of schizophrenia tissues, selective D4 antagonists, such as sonepiprazole and L-745,870, did not have any antipsychotic action. Some evidence suggests that the longer forms of D4, such as D4.7, with seven repeats, or D4.9, are found in hyperactive individuals or in those persons who take unusual risks, but this is controversial.

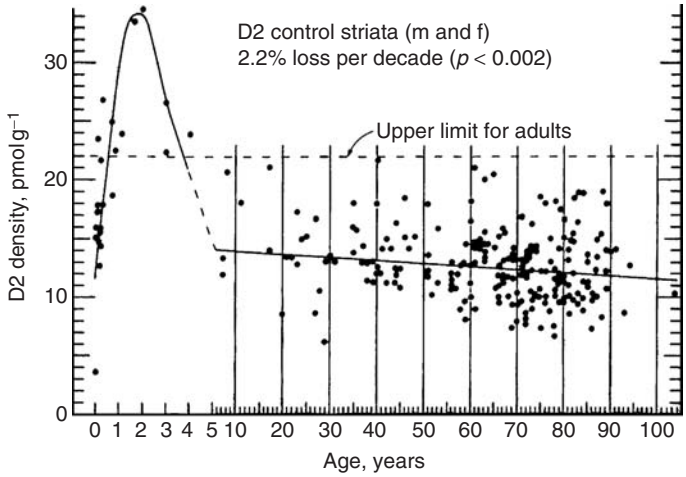

Figure 10 Postmortem human D2 densities in the striatum. After an initial rapid growth period in the first 3 years of life, the D2 receptor density is rapidly pruned before age 10, and thereafter decays by 2.2% every 10 years. Adapted from Seeman P, Bzowei NH, Guan H-C et al. (1987) Human brain dopamine receptors in children and aging adults. *Synapse* 1: 399–404, with permission from John Wiliy & Sons Inc.

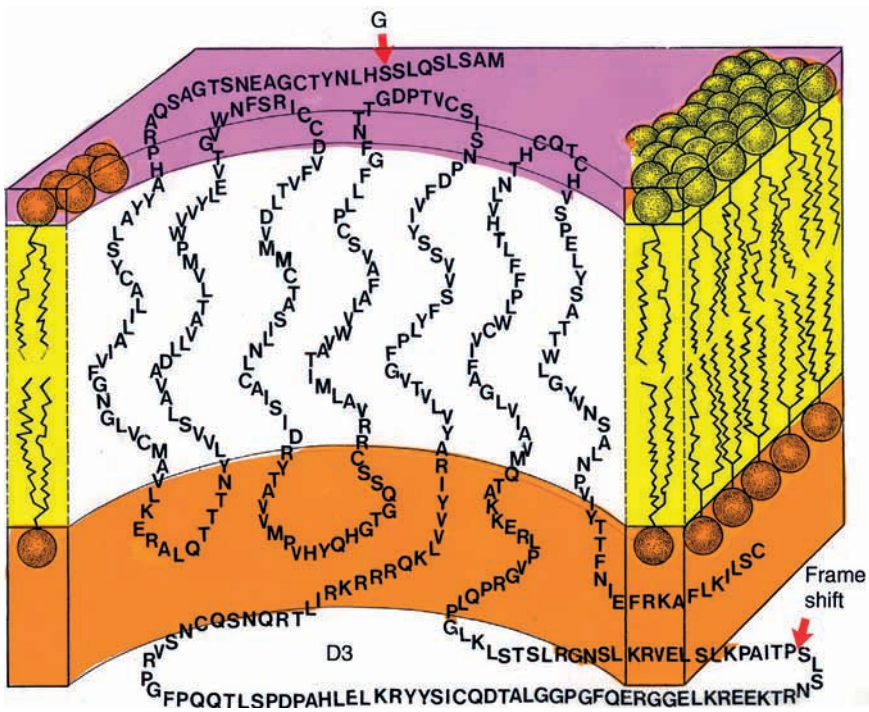

Figure 11 Amino acid sequence of the human cloned dopamine D3 receptor. A polymorphism occurs at position 9, where glycine replaces serine. When a frame shift occurs in the cytoplasmic loop, as shown, the receptor is nonfunctional.

The Dopamine D5 Receptor

The dopamine D5 receptor was cloned in 1991 (Sunahara), and its sequence is shown in Figure 14. There are two pseudogenes of D5, where the amino acid sequence stops at position 154. Although D5 is essentially D1-like in sequence and function, the characteristic feature of D5 is that it is more sensitive to dopamine than D1 is, as indicated in Table 1 for the K50 values.

Regulation of Dopamine Receptors

Each of the five dopamine receptors has a state of high affinity and a state of low affinity for dopamine, an example of which is shown in Figure 15 for the D2 receptor in anterior pituitary tissue. Dopamine receptors belong to a group of more than 1000 receptors known to be associated with G-proteins. The binding of an agonist to such a G-linked receptor occurs in two concentration ranges. Low nanomolar

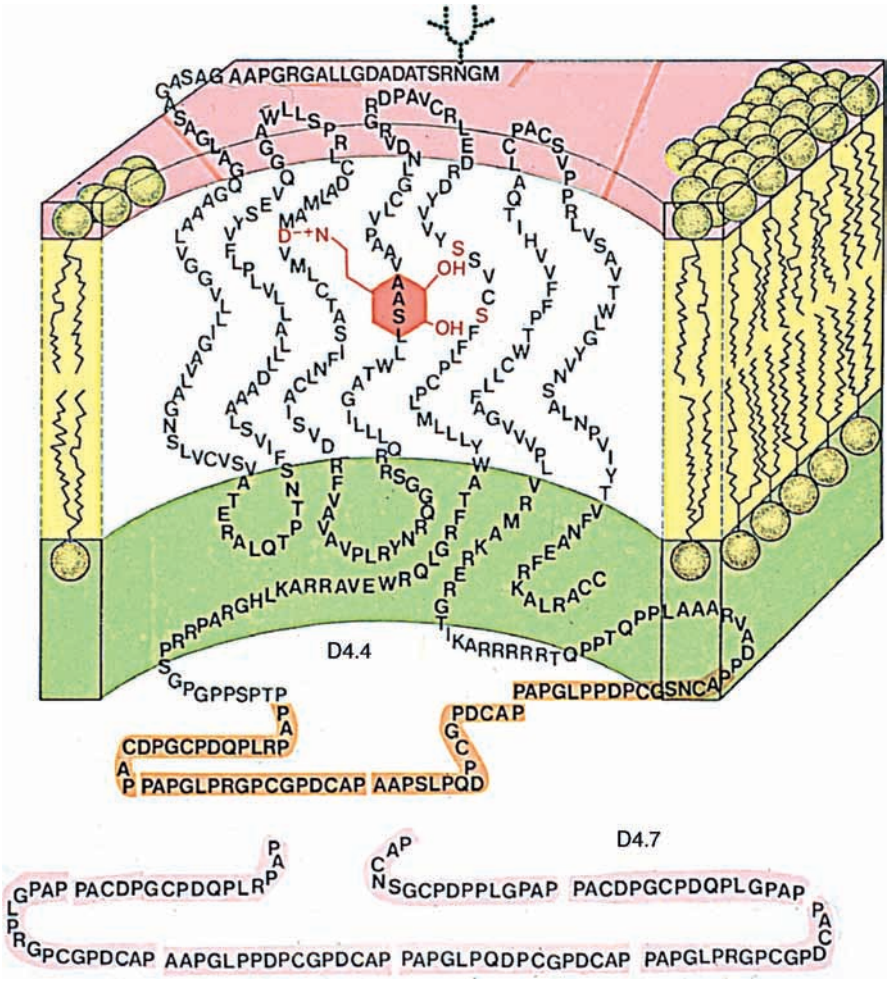

Figure 12 Amino acid sequence of the human dopamine D4 receptor and its polymorphisms. The cytoplasmic loop has repeat units of 16 amino acid polypeptides. Different humans have different numbers of repeats. Shown are four repeats (D4.4), the most common, and seven repeats (D4.7).

concentrations of the agonist binds to the high-affinity state of the receptor, while high micromolar concentrations bind to the low-affinity state of the receptor. Generally, it is the high-affinity state of the receptor that is the functionally active state of the receptor, because the agonist affinities for the high-affinity state are usually identical to the concentrations that elicit the physiological action of the agonists. This holds for many neurotransmitter receptors, including dopamine D2 receptors, cholinergic muscarinic receptors, α_2-adrenoceptors, and β_2-adrenoceptors.

$D2^{High}$ is the functional state in the anterior pituitary, upon which dopamine and other dopamine-like drugs (bromocriptine) act to inhibit the release of prolactin. $D2^{High}$ is presumably also functional on the terminals of the dopamine-containing terminals, and these receptors are usually referred to as presynaptic receptors. Although it has been reported that 90% of the D2 receptors in brain slices are in the $D2^{High}$ state, the proportion of D2 receptors in the high-affinity state in homogenized striatum *in vitro* is generally between 15% and 20%. The $D2^{High}$ state can be quickly converted into the $D2_{Low}$ state by guanine nucleotide, as illustrated in Figure 15 for anterior pituitary tissue and in Figure 16 for the striatum. Furthermore, as shown in Figure 16, the high-affinity state of D2 is most readily detected by dopamine competing with [^{3}H]domperidone, but not [^{3}H]raclopride or [^{3}H]spiperone, which are less sensitive to the competitive action of dopamine. In fact, it is known that the physiological concentration of dopamine in the synaptic space (between neurons) is 2–4 nM, matching the known dissociation constant of 2 nM for dopamine at the $D2^{High}$ receptor. This latter value of 2 nM is obtained from the dopamine/[^{3}H]domperidone competition curve in Figure 16, using the standard Cheng–Prusoff equation to correct for the ligand concentration and the [^{3}H]domperidone K_d.

There are at least two views of the physical existence of the high-affinity state. The traditional view is that

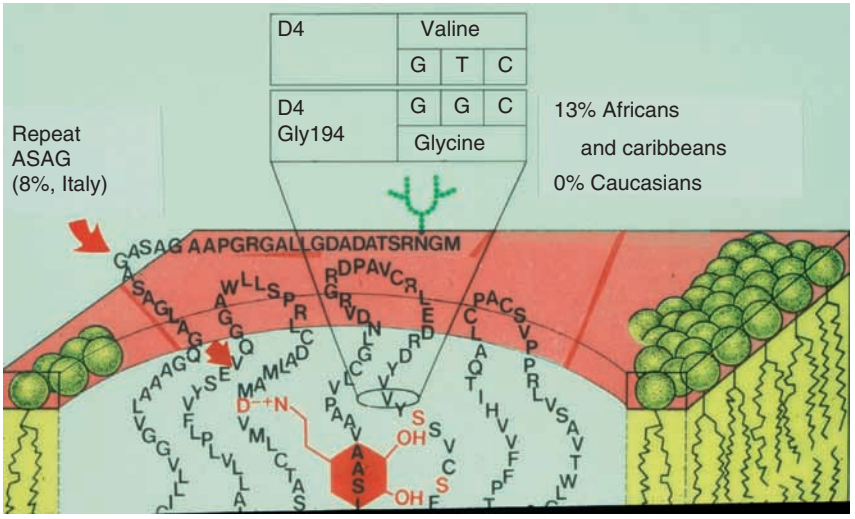

Figure 13 Additional polymorphisms of the human dopamine D4 receptor. Approximately 13% of Africans and Caribbeans have a glycine replacing valine at position 194. Approximately 8% of Italians have an additional ASAG peptide inserted at the position shown.

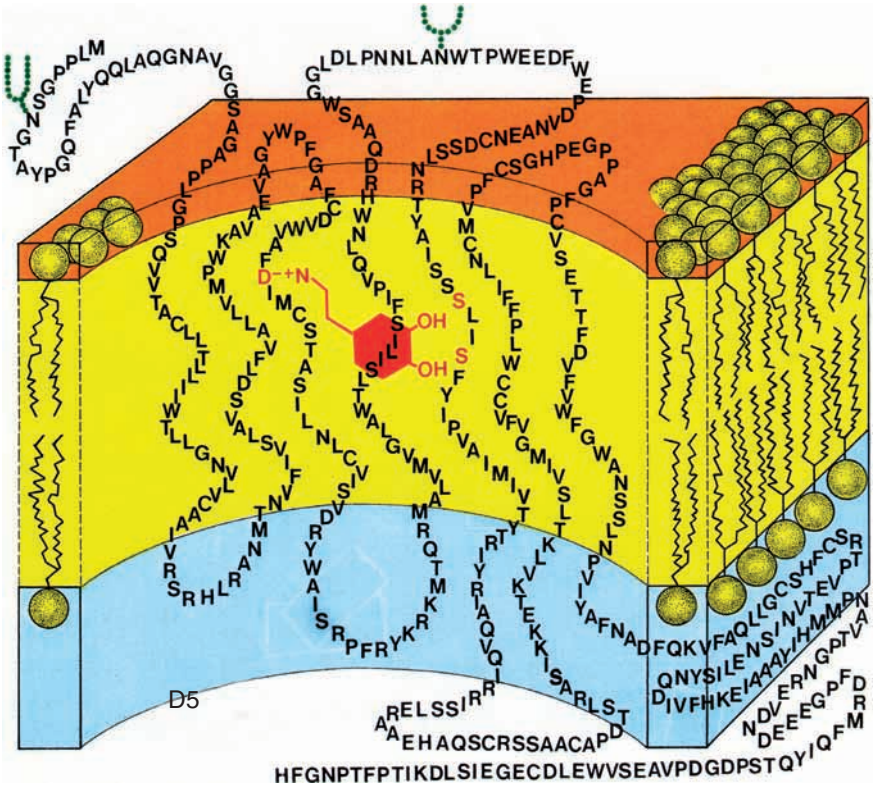

Figure 14 Amino acid sequence of the human dopamine D5 receptor. Two polymorphisms of D5 exist where the sequence abruptly stops to create a sequence of 154 amino acids instead of the full-length sequence of 477 amino acids.

the high-affinity state of the receptor exists when the receptor, R, is associated with the G-protein, and the agonist, D, binds to this high-affinity state to form the 'ternary complex,' namely DRG. This view of the receptor proposes that the low-affinity state occurs when the G-protein is not associated with the receptor. However, there are many significant shortcomings with this view of the high-affinity state of the

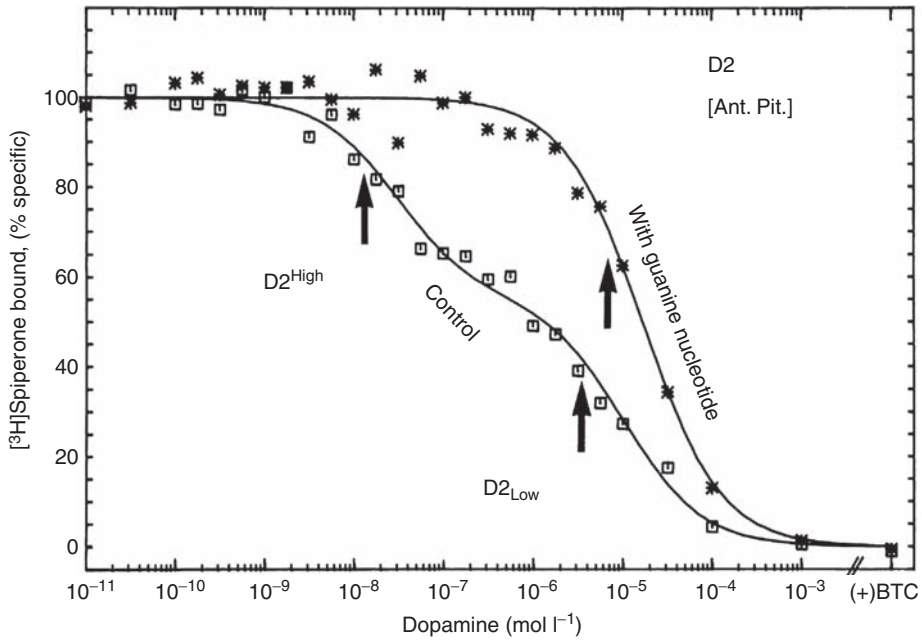

Figure 15 The high-affinity state of the dopamine D2 receptor, or $D2^{High}$, occurs at low nanomolar concentrations of dopamine in inhibiting the binding of [^{3}H]spiperone to anterior pituitary tissue (Ant. Pit.; porcine). The low-affinity state of the D2 receptor, or $D2_{Low}$, occurs at high concentrations of dopamine, as shown. However, the presence of 200 μM guanilylimidodiphosphate converts all the $D2^{High}$ receptors into the $D2_{Low}$ form.

receptor in the ternary complex model. For example, the ternary complex suggests that RG should have a transient existence. This is not the case, however, because it has been found that the purified muscarinic RG is stable. Moreover, the purified muscarinic receptor, free of G and GDP, clearly shows high-affinity and low-affinity states.

An alternate view of the high-affinity state of the receptor is the 'cooperativity' model, as worked out by J Wells and colleagues. The cooperative model proposes that the receptor cooperates with other receptors to form either a dimer, a tetramer, or a larger oligomer. The receptor is in the high-affinity state when it is vacant and unoccupied by the agonist. However, when the agonist binds to the vacant receptor, the occupied receptor interacts or 'cooperates' with the other receptors (within the tetramer) such that the affinity of the other receptors for the agonist is markedly reduced. This reduced affinity for the agonist is a result of 'negative cooperativity' between the receptors, and corresponds to the low-affinity state of the receptor.

In other words, if there is very strong negative cooperativity, then the second, third, and fourth receptors (within the tetramer) would hardly bind the agonist, and only the high-affinity sites would be observed in the competition between, say, dopamine and [^{3}H]domperidone, all taking place at the first receptor. These events are depicted in a diagram in Figure 17.

D2 Interactions with Other Receptors

The D2 receptor is known to interact with the D1 receptor as well as with other receptors, such as the adenosine A2 receptor. Many D1/D2 interactions occur at all levels, including the molecular level, where D1 and D2 can form functional heterodimers with one another. Such interaction also occurs at the cellular level, where the block of D1 (by SCH23390) unmasks the high-affinity state of the D2 receptor, $D2^{High}$, as shown in Figure 18. This experiment shows that D1 normally inhibits or suppresses the high-affinity state of D2.

Psychosis and the $D2^{High}$ Basis of Dopamine Supersensitivity

The dopamine hypothesis of psychosis or schizophrenia was first outlined by J. Van Rossum in 1967:

> The hypothesis that neuroleptic drugs may act by blocking dopamine receptors in the brain has been substantiated by preliminary experiments with a few selective and potent neuroleptic drugs. There is an urgent need for a simple isolated tissue that selectively responds to dopamine so that less-specific neuroleptic drugs can also be studied and the hypothesis further tested. . . . When the hypothesis of dopamine blockade by neuroleptic agents can be further substantiated it may have fargoing consequences for the pathophysiology of schizophrenia. Overstimulation of dopamine receptors could then be part of the aetiology.

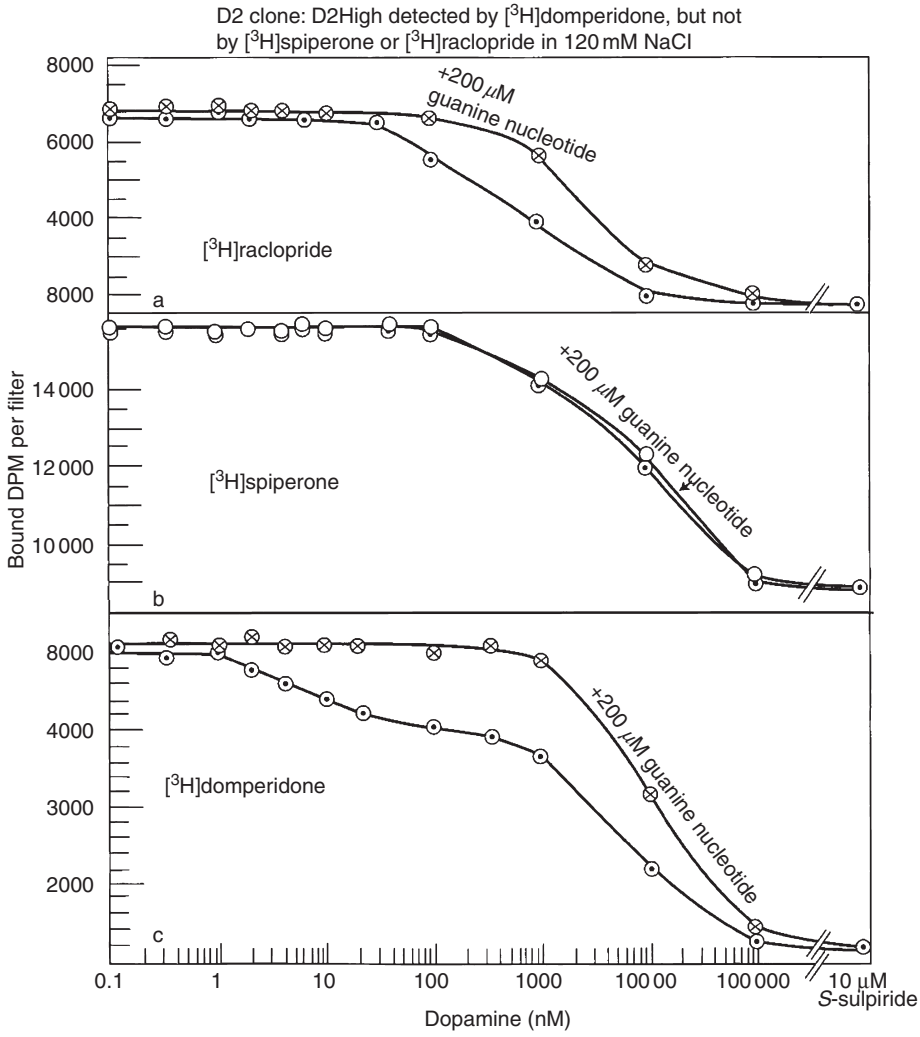

Figure 16 Low nanomolar concentrations of dopamine readily inhibit the binding of [^{3}H]domperidone at the high-affinity state of dopamine D2 receptors (between 1 and 100 nM), in contrast to [^{3}H]raclopride and [^{3}H]spiperone, which are less sensitive to competition by dopamine. The high-affinity state, $D2^{High}$, is converted to the low-affinity state, $D2_{Low}$, by guanine nucleotide. Adapted from Seeman P, Tallerico T, and Ko F (2003) Dopamine displaces [^{3}H]domperidone from high-affinity sites of the dopamine D2 receptor, but not [^{3}H] raclopride or [^{3}H]spiperone in isotonic medium: Implications for human positron emission tomography. *Synapse* 49: 209–215, with permission from John Wiley & Sons Inc.

As noted earlier, it has not been clearly established that D2 receptors are elevated in psychosis or schizophrenia, although studies using brain imaging of healthy co-twins of schizophrenia individuals, as well as single-photon brain imaging of nonmedicated psychotic patients in at least one study, have shown significant elevations of D2.

At present, the most promising direction in this field is to examine the molecular basis of dopamine supersensitivity, because up to 70% of patients are supersensitive to either methylphenidate or amphetamine at doses which do not affect controls. Moreover, a wide variety of brain alterations in animals (lesions, birth injury by C-section, amphetamine or phencyclidine treatment, knockouts of a variety of receptors) all lead to the final common finding of behavioral dopamine supersensitivity and elevated proportions of D2 receptors in the $D2^{High}$ state in the striatum. For example, repeated administration of amphetamine to animals or humans leads to behavioral dopamine supersensitivity. While the density of D2 receptors in the striatum does not change in such studies, it is remarkable that the density of $D2^{High}$ receptors increases dramatically by several fold, as shown in Figure 19.

A similar situation occurs in animals that receive hippocampal lesions neonatally. Such animals, as adults, reveal behavioral dopamine supersensitivity, and the striatum contains a marked increase in the proportion of D2 receptors in the high-affinity state, as shown in Figure 20. Therefore, the molecular control of the high-affinity state of D2 is emerging as a

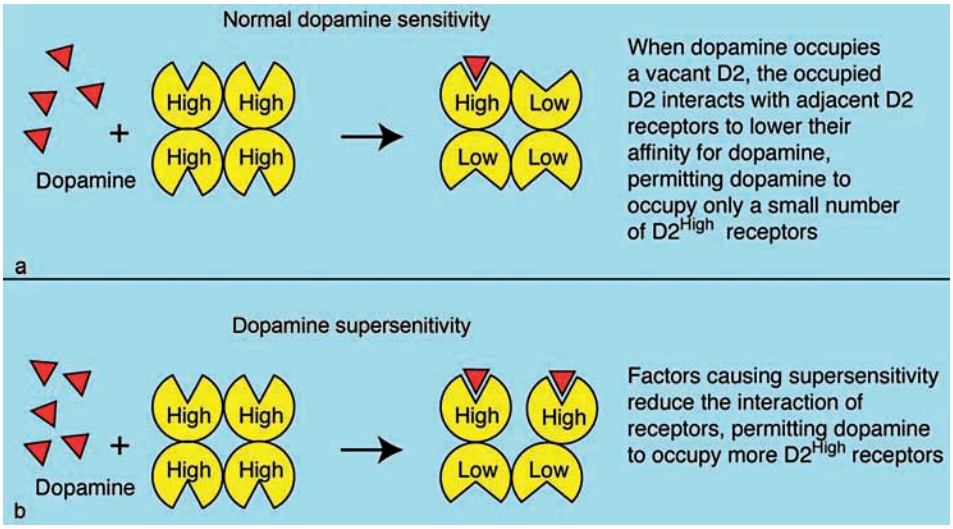

Figure 17 (a) The cooperativity model for the dopamine D2 receptor, according to J Wells and colleagues. The receptor is proposed to exist as an oligomer of D2 receptors, such as a tetramer of D2. Each of the vacant receptors is in the high-affinity state, $D2^{High}$. When dopamine first attaches to one of the vacant D2 receptors in the tetramer, the occupied D2 then interacts with the other three receptors within the tetramer to reduce their affinity for dopamine (i.e., negative cooperativity). (b) One molecular explanation for dopamine supersensitivity is that an unknown factor may reduce the negative interaction between the D2 receptors, thereby allowing more dopamine to occupy more $D2^{High}$ receptors.

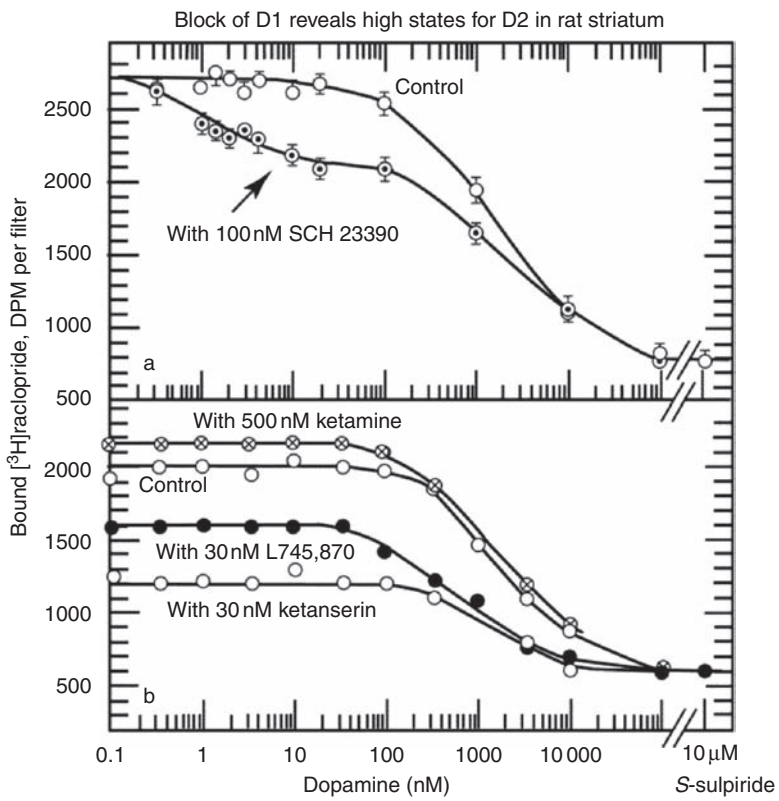

Figure 18 Regulation of $D2^{High}$ by the D1 receptor. (a) Dopamine inhibited the binding of [^{3}H]raclopride to D2 receptors at dopamine concentrations higher than 100 nM. However, in the presence of SCH23390 to block D1 receptors, the binding of [^{3}H]raclopride was readily inhibited by 1–100 nM, corresponding to the presence of $D2^{High}$. These data suggest that D1 actively suppresses the existence of the functional $D2^{High}$ state. (b) In comparison to 100 nM SCH23390 unmasking the $D2^{High}$ state (a), other drugs do not lead to such unmasking of $D2^{High}$. L745,870 is a dopamine D4 receptor antagonist. Adapted from Seeman P and Tallerico T (2003) Link between dopamine D1 and D2 receptors in rat and human striatal tissues. *Synapse* 47: 250–254, with permission from John Wiley & Sons Inc.

central problem in this field. At present, there is uncertainty as to whether this high-affinity state of D2 is controlled through G_o or one of the G_i proteins, because this varies from cell to cell.

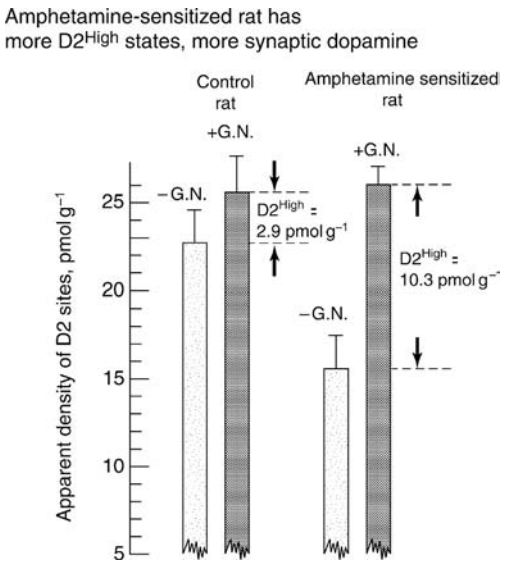

Figure 19 Repeated administration of amphetamine to rats leads to behavioral dopamine supersensitivity. While the total density of D2 receptors was normal in the striata of such supersensitive rats (about 26 pmol g^{-1}), the density of $D2^{High}$ receptors was markedly elevated by 355%, from a control value of 2.9 pmol g^{-1} to the elevated level of 10.3 pmol g^{-1}. Nonspecific binding of [^{3}H]raclopride was done in the presence of 10 μM *S*-sulpiride. In the presence of 200 μM guanilylimidodiphosphate (G.N.), all the $D2^{High}$ receptors were converted to $D2_{Low}$. Adapted from Seeman P, Tallerico T, Ko F, et al. (2002) Amphetamine-sensitized animals show a marked increase in dopamine $D2^{High}$ receptors occupied by endogenous dopamine – Even in the absence of acute challenges. *Synapse* 46: 235–239, with permission from John Wiley & Sons Inc.

According to the negative cooperativity model (Figure 17), the increased number of D2 receptors in the high-affinity state, $D2^{High}$, found in the striata of supersensitive animals, may be attributed to a reduction in the overall negative cooperativity between the receptors, as illustrated in Figure 17. Thus, in order to determine the molecular mechanism of dopamine supersensitivity, it will be essential to determine the factors that reduce negative cooperativity among the D2 receptors or that alter the association of the receptor with its G-protein. The role of guanine nucleotides in regulating the overall sensitivity of the dopamine D2 receptors would be to alter the extent of the receptor–receptor negative cooperativity.

Current Clinical and Basic Research on Dopamine Receptors

Of the five dopamine receptors and their many variants, the D2 receptor and its properties continue to be most actively investigated, because D2 is the main clinical target of antipsychotics and of dopamine agonist treatment of Parkinson's disease. The D1 receptor, however, also has an important clinical role in treating Parkinson's disease because the stimulation of D1 synergizes with the stimulation of D2, possibly via D1/D2 heterodimers or cell–cell interactions. A current active area of clinical research on dopamine receptors is to measure the occupancy of D2 receptors both in the striatum and outside the striatum in individuals taking antipsychotic medications. Some researchers find that the same D2 occupancy occurs in both striatal and limbic regions, while others find a lower occupancy in the limbic regions.

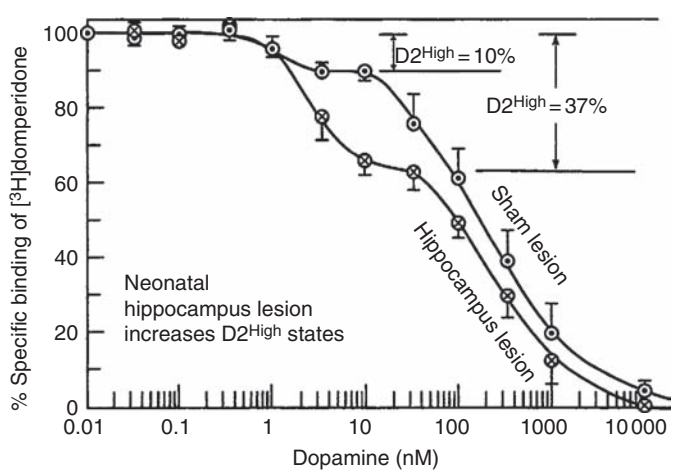

Figure 20 Rats with lesions made neonatally reveal behavioral dopamine supersensitivity when they become adults. Such supersensitive animals reveal a marked increase of 3.7-fold in the proportion of D2 receptors in the high-affinity state, $D2^{High}$. Nonspecific binding of [^{3}H]domperidone was done in the presence of 10 μM *S*-sulpiride. Adapted from Seeman P, Weinshenker D, Quirion R, et al. (2005) Dopamine supersensitivity correlates with $D2^{High}$ states, implying many paths to psychosis. *Proceedings of the National Academy of Sciences of the United States of America* 102: 3513–3518.

As previously noted, therapeutic doses of antipsychotics occupy 60–80% of the D2 receptors in psychotic patients, while D2 occupancies higher than 80% are associated with elevated serum prolactin and parkinsonism. The new aripiprazole antipsychotic tends to occupy more than 80% of D2 but does not seem to cause parkinsonism at these higher levels. This may be a result of the fact that aripiprazole, like clozapine and quetiapine, quickly desorbs from the D2 receptor (in under 60 s *in vitro*), minimizing prolactin elevation and parkinsonism. Rapid desorption permits the dopamine neurotransmission to proceed more normally. However, if the antipsychotic drug dose or antipsychotic concentration in the plasma remains high, then the antipsychotic drug will readsorb. In effect, fat-soluble antipsychotics remain adsorbed to D2 in patients for 1 or 2 days or even more, while the plasma concentration falls. Less fat-soluble antipsychotics such as clozapine or quetiapine remain attached to D2 in patients for only 6–12 h, immediately falling in D2 occupancy as the plasma concentration falls.

Probably the most central question in determining the basis of psychosis or schizophrenia is to determine the molecular basis of dopamine supersensitivity and to determine which proteins or genes regulate the maintenance of D2 receptors in their high-affinity state.

See also: Antipsychotic Drugs; Dopamine; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopamine Neurons: Reward and Uncertainty; Dopamine: Cellular Actions; Dopaminergic Agonists and L-DOPA; Parkinsonian Syndromes; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome.

Further Reading

Agid O, Kapur S, Arenovich T, et al. (2003) Delayed-onset hypothesis of antipsychotic action: A hypothesis tested and rejected. *Archives of General Psychiatry* 60: 1228–1235.

Baumeister AA and Francis JL (2002) Historical development of the dopamine hypothesis of schizophrenia. *Journal of the History of Neurosciences* 11: 265–277.

Bressan RA, Erlandsson K, Jones HM, et al. (2003) Is regionally selective D_2/D_3 dopamine occupancy sufficient for atypical antipsychotic effect? An *in vivo* quantitative [^{123}I]epidepride SPET study of amisulpride-treated patients. *American Journal of Psychiatry* 160: 1413–1420.

Farde L, Nordstrom AL, Wiesel FA, et al. (1992) Positron emission tomographic analysis of central D1 and D2 dopamine receptor occupancy in patients treated with classical neuroleptics and clozapine. Relation to extrapyramidal side effects. *Archives of General Psychiatry* 49: 538–544.

George SR, Watanabe M, Di Paolo T, et al. (1985) The functional state of the dopamine receptor in the anterior pituitary is in the high affinity form. *Endocrinology* 117: 690–697.

Hagberg G, Gefvert O, Bergström M, et al. (1998) N-[^{11}C]methylspiperone PET, in contrast to [^{11}C]raclopride, fails to detect D_2 receptor occupancy by an atypical neuroleptic. *Psychiatry Research* 82: 147–160.

Hirvonen J, van Erp TG, Huttunen J, et al. (2005) Increased caudate dopamine D2 receptor availability as a genetic marker for schizophrenia. *Archives of General Psychiatry* 62: 371–378.

Kapur S and Seeman P (2001) Does fast dissociation from the dopamine D2 receptor explain the action of atypical antipsychotics? – A new hypothesis. *American Journal of Psychiatry* 158: 360–369.

Seeman P (1987) Dopamine receptors and the dopamine hypothesis of schizophrenia. *Synapse* 1: 133–152.

Seeman P (2002) Atypical antipsychotics: Mechanism of action. *Canadian Journal of Psychiatry* 47: 27–38.

Seeman P, Bzowej NH, Guan H-C, et al. (1987) Human brain dopamine receptors in children and aging adults. *Synapse* 1: 399–404.

Seeman P, Chau-Wong M, Tedesco J, et al. (1975) Brain receptors for antipsychotic drugs and dopamine: Direct binding assays. *Proceedings of the National Academy of Sciences of the United States of America* 72: 4376–4380.

Seeman P, Lee T, Chau-Wong M, et al. (1976) Antipsychotic drug doses and neuroleptic/dopamine receptors. *Nature (London)* 261: 717–719.

Seeman P and Niznik HB (1990) Dopamine receptors and transporters in Parkinson's disease and schizophrenia. *FASEB Journal* 4: 2737–2744.

Seeman P and Tallerico T (1999) Rapid release of antipsychotic drugs from dopamine D_2 receptors: An explanation for low receptor occupancy and early clinical relapse upon drug withdrawal of clozapine or quetiapine. *American Journal of Psychiatry* 156: 876–884.

Seeman P and Tallerico T (2003) Link between dopamine D1 and D2 receptors in rat and human striatal tissues. *Synapse* 47: 250–254.

Seeman P, Tallerico T, and Ko F (2003) Dopamine displaces [^{3}H] domperidone from high-affinity sites of the dopamine D2 receptor, but not [^{3}H]raclopride or [^{3}H]spiperone in isotonic medium: Implications for human positron emission tomography. *Synapse* 49: 209–215.

Seeman P, Tallerico T, Ko F, et al. (2002) Amphetamine-sensitized animals show a marked increase in dopamine $D2^{High}$ receptors occupied by endogenous dopamine – Even in the absence of acute challenges. *Synapse* 46: 235–239.

Seeman P, Weinshenker D, Quirion R, et al. (2005) Dopamine supersensitivity correlates with $D2^{High}$ states, implying many paths to psychosis. *Proceedings of the National Academy of Sciences of the United States of America* 102: 3513–3518.

Wilson AA, McCormick P, Kapur S, et al. (2005) Radiosynthesis and evaluation of [^{11}C]-(+)-PHNO as a potential radiotracer for in vivo imaging of the dopamine D2 high affinity state with positron emission tomography (PET). *Journal of Medicinal Chemistry* 48: 4153–4160.

Dopamine: Cellular Actions

G Bernardi and N B Mercuri, Università di Roma 'Tor Vergata', and IRCCS Fondazione Santa Lucia, Rome, Italy

© 2009 Elsevier Ltd. All rights reserved.

There is a plethora of experimental and clinical evidence indicating that dopamine (DA) is a key element in promoting and regulating the reward and motivation processes occurring in the brain. Therefore, the state of pleasure and well-being that one might derive from everyday activities such as playing, eating palatable food, drinking, listening to a melody, smelling a perfume, and having sex certainly rely on the correct operation of this catecholamine at the neuronal level. On the other hand, DA plays a significant role when an altered behavior is instituted. Hence, drug abuse and craving, compulsive acts, and affective problems could depend on maladaptive changes in DA neurotransmission. This catecholamine is synthesized by discrete groups of cells (mesencephalic, ipothalamic, and retinal). It is noteworthy that the main source of DA for the central nervous system arises from the population of neurons originating in the ventral mesencephalon (ventral tegmental area, substantia nigra pars compacta) and mainly terminating in the cerebral cortex, accumbens, and striatum.

The diffuse projections of the dopaminergic neurons suggest that DA is able to act in different brain areas to induce motivation and reward. Accordingly, behavioral, pharmacological, and clinical studies have demonstrated that the cerebral regions principally involved in incentive and behavioral activation are the prefrontal cortex, the ventral/dorsal striatum, and the ventral mesencephalon itself. Indeed, these regions are the richest in DA.

Therefore, in order to better understand the reward processes, it is of crucial importance to understand the functions of DA at the cellular and synaptic level within the different areas indicated previously. Another important aspect to be considered for the clarification of DA action is that the catecholamine stimulates subtypes of specific receptors that cause distinct effects on neurons. There are two families of G-protein-linked DAergic receptors, D1 and D2. While the D1 family (constituting the D1 and D5 receptors), linked to Gαs, stimulates adenylate cyclase, the D2 family (constituting the D2, D3, and D4 receptors), linked to Gα0/i, inhibits adenylate cyclase. Usually, the D1-mediated effects are cyclic adenosine monophosphate (cAMP)- and protein kinase A-dependent, while the D2-mediated effects depend on the inhibition of cAMP and protein kinase A (PKA), eventually the activation of phospholipase C-inositol triphosphate (IP3), and changes in intracellular Ca^{2+}. The dichotomy in the formation of cAMP results in opposite consequences on neuronal excitability. In addition, by stimulating either similar or different receptors, DA might exert competing presynaptic and/or postsynaptic actions to modify neuronal responsiveness.

In the present article we reexamine the electrophysiological actions of DA on neurons located in those brain areas that are linked to the reward processes.

Dopamine Actions in the Frontal Cortex

DA in the frontal cortex has implications in reward, the psychomotor and hedonic effects of drug abuse, drug seeking, and certain forms of memory formation linked to salience. Electrophysiological data have extensively shown a variety of effects caused by DA and DAergic agonists in this structure. Although there is a consensus that DA inhibits cortical neurons, several studies have also reported DA-mediated neuronal excitation. DA might have different effects on the excitability of pyramidal neurons by reducing either persistent voltage-dependent sodium currents or voltage-dependent potassium currents. It has been also reported that the stimulation of D1 receptors can activate firing rate by modulating synaptic strength. In fact, some authors have shown that DA significantly enhances excitatory postsynaptic current (EPSC) amplitudes in the pyramidal cells. Recent data have also suggested that DA potentiates the late postsynaptic responses in an excitatory postsynaptic potential (EPSP) train evoked by a sustained presynaptic stimulation. However, a combined activation of D1 and D2 receptors could also determine a reduction of the efficacy of excitatory synaptic transmission at synapses onto pyramidal cells. Moreover, contrasting effects of DA on neuronal responsiveness to glutamate agonists have been reported. Thus, bath application of *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), and the D1 agonist SKF 38393 induces concentration-dependent excitability increases, whereas the D2 receptor agonist quinpirole causes a concentration-dependent excitability decrease. Furthermore, opposing D1- or D2-mediated actions, depending on DA concentrations, have been shown to modify the inhibitory drive on cortical neurons. As a result, a nanomolar concentration of DA might enhance the inhibitory postsynaptic currents (IPSCs) via D1 receptors while a micromolar concentration could decrease the IPSCs via D2 receptors. In contrast, it has been reported that the stimulation of presynaptic D1-like receptors

decreases the IPSCs in layer II–III of neocortical pyramidal cells.

In addition, DA could induce important changes in the local network connections of pyramidal cells with γ-aminobutyric acid (GABA)ergic interneurons. Thus, the activation of D1 receptors reduces the amplitude of the EPSPs or causes a reversible membrane depolarization in the fast-spiking (FS) interneurons. On the other hand, the stimulation of D4 receptors, limiting interneuronal excitability, favors multiple spike discharge of pyramidal cells.

Furthermore, DA has been shown to sustain enduring changes of the synaptic signals. Thus, the long-term potentiation (LTP) of the excitatory transmission on pyramidal neurons requires D1 receptor stimulation. This could activate PKA and increase the surface expression of GluR1 in prefrontal cortex neurons. On the other hand, a constant presence of DA in the extracellular space might facilitate long-term depression (LTD). It is supposed that these DA-mediated persistent changes in synaptic responses (LTP-LTD) could be important in processes linked to reward.

Dopamine Actions in the Ventral and Dorsal Striatum

A sizable body of evidence considers the ventral striatum (nucleus accumbens) as a fundamental station for the DA-regulated reward processes. In fact, accumbens DA modulates motivation and cognitive aspects related to incentives. In addition, DA effects in this structure are particularly important in mediating the activating and addictive effects of psychostimulants.

As in the cerebral cortex, accumbal DA might change the activity of medium spiny neurons by involving multiple mechanisms. It has been shown that this catecholamine causes a membrane hyperpolarization that is due to the activation of D1 receptors and is associated with an increase in potassium conductance. However, DA also causes a membrane depolarization that is generated by the stimulation of D2 receptors and sustained by a decrease in potassium conductance. In addition, D2 receptors seem to reduce the direct excitability of medium spiny accumbal neurons by inhibiting evoked Na^+ spikes through the involvement of slow A-type K^+ currents, while reported effects of DA on inward rectification and 'leak' K^+ currents could favor excitability. D2 receptors could also cause an enhancement of voltage-sensitive sodium currents mediated by the suppression of the cyclic AMP/PKA cascade and the facilitation of intracellular Ca^{2+} signaling. In addition, D1 receptor stimulation appears to suppress N- and P/Q-type Ca^{2+} currents by activating a cAMP/protein kinase A/protein phosphatase signaling system, presumably leading to channel dephosphorylation.

Moreover, a coactivation of D1 and D2 receptors and signaling through G-protein $\beta\gamma$ subunits and PKA could enhance spike firing in nucleus accumbens shell medium spiny neurons.

Referring to the DA-induced changes of synaptic functions, there are data supporting an inhibitory role of this catecholamine on the EPSP-IPSP sequences The depression of the EPSP appears to be mainly mediated by a D2 receptor-induced decrease of AMPA currents, while the inhibition of the EPSCs progressively diminishes; the inhibition of the IPSCs seems to be persistent. The resultant effect could cause an increased excitability of these neurons.

Activation of D3 receptors is also reported to suppress the efficacy of the inhibitory synaptic transmission in the nucleus accumbens by increasing the phospho-dependent endocytosis of GABA-A receptors.

Interestingly, DA also attenuates the neuronal responses to repetitive activation of glutamatergic afferents and thereby could block LTP. On the other hand, the accumbal LTD requires, among other factors, dopamine D1 receptor stimulation and cAMP-PKA activation. DA-mediated plastic adaptations of the neural synaptic functions in the nucleus accumbens caused by psychostimulants might control the development and long-term maintenance of sensitization to abused drugs.

The increase of extracellular DA in the dorsal striatum of humans (caudoputamen) related to self-reported measures of liking and the feeling of a 'high' (euphoria) implies that this structure is an essential element of the circuitry responsible for the control of motivated behavior and reward. However, in spite of the recognized importance of DA in reward-related locomotor activation, the precise nature of the modulation that DA exerts on striatal neurons remains largely elusive.

There are early electrophysiological studies demonstrating a predominant inhibitory effect of DA on low-firing medium spiny striatal cells. Thus, the most common responses produced by DA on these cells are the decrease of voltage-dependent inward conductances and the modulation of the corticostriatal synaptic transmission. The DA-mediated inhibition of firing activity seems to be mediated by the activation of D1 receptors. The stimulation of D1-like receptors initiates a cascade of intracellular events, including cAMP formation and activation of cAMP-dependent PKA that reduce excitability and possibly activate the neuronal changes linked to reward. An additional stimulation of D2 receptor reduces glutamate- and GABA-mediated currents.

Some authors have also described a reduction of AMPA-mediated postsynaptic current caused by D2 receptors in medium-sized striatal neurons. In contrast, the stimulation of D1 receptors could potentiate the NMDA-induced currents and reduce GABA-evoked currents by activating a PKA/dopamine- and cAMP-regulated phosphoprotein (DARPP-32)/protein phosphatase 1 signaling cascade. It must also be considered that the cholinergic interneurons are affected by DA. While the stimulation of D1 receptors depolarizes and excites the cells, the stimulation of D2 receptors reduces the GABAergic and cholinergic inputs on these interneurons. With regard to the synaptic function, the two main subclasses of DA receptors contribute in the formation of the long-term changes (LTD/LTP) occurring at the corticostriatal terminals. Therefore, D1-like and D2-like receptors interact synergistically to allow LTD formation. A form of LTD in the striatum is also bidirectionally modulated by D2 receptors and requires the regulation of mGluR-dependent endocannabinoid release. The LTP is blocked by dopamine depletion but is not affected by a D2 antagonist. Furthermore, the use of knockout mice with the ablation of D1 receptors has demonstrated a disrupted corticostriatal LTP, whereas pharmacological blockade of D5 receptors prevented LTD in these animals. Interestingly, DA-dependent stimulation of DARPP-32, a small protein expressed in the spiny neurons which acts as a potent inhibitor of protein phosphatase-1, is necessary for the expression of both corticostriatal LTD and LTP and is required for the full expression of the behavioral response to cocaine and amphetamine. On this basis, a role for neostriatal DA in the formation of a pathological drug habit has been hypothesized.

Dopamine Actions in the Ventral Mesencephalon

There is homogeneous experimental material demonstrating that DA, released from the dendrites, has an inhibitory action on the dopaminergic cells located in the ventral tegmental area and substantia nigra pars compacta. This inhibition is D2-mediated and is due to a membrane hyperpolarization depending on a G-protein (Go/i)-activated potassium current (GIRK). Therefore, by modifying firing discharge of the DAergic cells, the catecholamine might encode the salience of stimuli linked to natural and artificial rewards. Although there is evidence that the dopaminergic neurons fire during reward anticipation, encoding errors in reward prediction, the inhibition of firing mediated by locally released DA could play part of the reward processes. Therefore, a delicate balance of activity in different subsets of DA cells could regulate the timing of the various expressions of reward. In addition, DA, by modifying the strength of the excitatory transmission in the ventral mesencephalon, might contribute to enhancing memory processes linked to salience. Thus, DA could block the induction of LTD in the ventral midbrain via activation of D2-like receptors. This DA-dependent modulation of the synaptic code could be a substrate of memory traces linked to reward. Of note is the fact that the DA-mediated long-term alterations in the strength of the excitatory transmission have been reported to occur after treatments with psychomotor stimulants. Since the development of addictive behaviors shares common features with traditional learning models, these DA-dependent modifications of synaptic plasticity could represent an important substrate for the acquisition of reward-related behaviors.

Conclusions

It is believed that the multiple and, in some instances, opposite effects of DA on cerebral neurons could account for the whole range of behavioral modifications linked to reward processes. However, in spite of the large amount of experimental work so far available, we have to admit that the overall picture describing the actions of DA on single cells and groups of cells is rather complex. Undoubtedly, there is a certain unpredictability in the neuronal responses to DA that could likely depend on competing presynaptic, postsynaptic, and nonsynaptic mechanisms, the type of receptor involved, and the type of neuron. Notwithstanding these limitations, future effort in studying the function of DA should be concentrated on individuating and eventually correcting the processes that control reward at the cellular level.

See also: Dopamine Control of Arousal; Dopamine; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopamine Neurons: Reward and Uncertainty; Dopaminergic Differentiation; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Bernardi G, Cherubini E, Marciani MG, et al. (1982) Responses of intracellularly recorded cortical neurons to the iontophoretic application of dopamine. *Brain Research* 245: 267–274.

Beurrier C and Malenka RC (2002) Enhanced inhibition of synaptic transmission by dopamine in the nucleus accumbens during behavioral sensitization to cocaine. *Journal of Neuroscience* 22: 5817–5822.

Bonci A, Bernardi G, Grillner P, et al. (2003) The dopamine-containing neuron: Maestro or simple musician in the orchestra of addiction? *Trends in Pharmacological Science* 24: 172–177.

Centonze D, Grande C, Saulle E, et al. (2003) Distinct roles of D1 and D5 dopamine receptors in motor activity and striatal synaptic plasticity. *Journal of Neuroscience* 23: 8506–8512.

Cepeda C, Colwell CS, Itri JN, et al. (1998) Dopaminergic modulation of NMDA-induced whole cell currents in neostriatal neurons in slices: Contribution of calcium conductances. *Journal of Neurophysiology* 79: 82–94.

Chen G, Kittler JT, Moss SJ, et al. (2006) Dopamine D3 receptors regulate GABAA receptor function through a phospho-dependent endocytosis mechanism in nucleus accumbens. *Journal of Neuroscience* 26: 2513–2521.

Kreitzer AC and Malenka RC (2005) Dopamine modulation of state-dependent endocannabinoid release and long-term depression in the striatum. *Journal of Neuroscience* 25: 10537–10545.

Lacey MG, Mercuri NB, and North RA (1987) Dopamine acts on D2 receptors to increase potassium conductance in neurones of the rat substantia nigra zona compacta. *Journal of Physiology* 392: 397–416.

Law-Tho D, Hirsch JC, and Crepel F (1994) Dopamine modulation of synaptic transmission in rat prefrontal cortex: An *in vitro* electrophysiological study. *Neuroscience Research* 21: 151–160.

Perez MF, White FJ, and Hu XT (2006) Dopamine D(2) receptor modulation of K(+) channel activity regulates excitability of nucleus accumbens neurons at different membrane potentials. *Journal of Neurophysiology* 96: 2217–2228.

Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36: 241–263.

Seamans JK and Yang CR (2004) The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology* 74: 1–58.

Thomas MJ, Malenka RC, and Bonci A (2000) Modulation of long-term depression by dopamine in the mesolimbic system. *Journal of Neuroscience* 20: 5581–5586.

Uchimura N, Higashi H, and Nishi S (1986) Hyperpolarizing and depolarizing actions of dopamine via D-1 and D-2 receptors on nucleus accumbens neurons. *Brain Research* 375: 368–372.

Wise RA (2004) Dopamine, learning and motivation. *Nature Reviews Neuroscience* 5: 1–12.

Dopaminergic Agonists and L-DOPA

A Pisani and N B Mercuri, Università di Roma 'Tor Vergata,' Rome, Italy

© 2009 Elsevier Ltd. All rights reserved.

Pharmacological Treatment of Parkinson's Disease

Parkinson's disease (PD) is a common neurodegenerative disorder, affecting approximately 3% of people older than 65 years. Considerable advances have been made in identifying the etiopathogenesis of PD. In recent years, different gene mutations have been identified as causing familial forms of PD, accounting for nearly 2% of cases. Despite this, the large majority of cases are 'sporadic–idiopathic' and are thought to be the outcome of complex interactions between the environment and the genetic background. Cardinal features of PD are represented by resting tremor, rigidity, bradykinesia, and impaired balance, often complicated by autonomic, cognitive, and emotional disturbances. The main neuropathological characteristic of the disease is a loss of pigmented dopaminergic cells in the substantia nigra pars compacta that produces a massive depletion of dopamine in the basal ganglia (mainly striatum). Neuropathological findings have also demonstrated degenerative processes in the serotoninergic, noradrenergic, cholinergic, and peptidergic pathways.

Although there is currently no cure for PD, because no therapy has been definitely shown to halt its progression, there are medications available to treat the motor symptoms, which can certainly improve the quality of life of affected patients. These are mainly based on a dopamine replacement strategy.

L-DOPA in PD

The dopamine precursor L-DOPA (L-3,4-dihydroxyphenylalanine, also called levodopa) is still the 'gold standard antiparkinsonian agent' because a drug superior to it has not been produced after more than four decades of use. Long-lasting observations in clinical practice have undoubtedly demonstrated that orally administered L-DOPA is a more 'potent' symptomatic antiparkinsonian drug than any dopaminergic agonist thus far introduced in the therapeutic armamentaria. It has been also reported that it increases life expectancy in PD patients.

The superiority of L-DOPA is very likely due to its special feature that allows *de novo* synthesis of dopamine mainly in the surviving dopamine-containing neurons, thus raising the level of this catecholamine in the striatum. L-DOPA is converted in dopamine by the enzyme aromatic L-amino acid decarboxylase (AADC) that is present not only in the brain but also in many peripheral organs and vessels. It has been recognized that the peripheral transformation of L-DOPA by AADC only permits a small percentage of L-DOPA to reach the dopaminergic neurons in the brain. Therefore, in order to obtain valid therapeutic concentrations, a huge dose of L-DOPA was administered. This caused important side effects, including hypotension, nausea, and vomiting. Opportunely, the coadministration of L-DOPA with two compounds (benserazide and carbidopa), which block AADC but do not cross the blood–brain barrier, has increased the availability of this dopamine precursor to the brain, thus diminishing its dosage and reducing side effects.

It has been suggested that the initial excellent therapeutic response to L-DOPA depends on the fact that newly synthesized dopamine can be released mainly at the synaptic contacts under a stimulus-dependent control. Thus, the newly formed catecholamine might be, at least initially, released respecting neurochemistry and the cytoarchitecture of the remaining dopaminergic neurons. As a result, despite a neuronal degeneration in PD, a phasic and a tonic release of dopamine could still be sustained by oral administrations of L-DOPA. Thus, this L-DOPA-induced dopamine release adequately stimulates dopaminergic receptors. At least six different dopamine receptor subtypes have been cloned. These receptors are divided in two classes: the D1 class comprises D1 and D5 receptor subtypes, whereas the D2 class includes D2 (D2S and D2L isoforms), D3, and D4 receptors. It has been demonstrated that the D2 receptors are mainly involved in the regulation of striatal motor functions. Interestingly, all dopaminergic receptors cloned from the brain are also present in tissues outside the central nervous system (CNS), mainly in the cardiovascular system, the kidney, as well as sympathetic ganglia and the adrenal glands.

Whereas the phasic release of dopamine might control the transitory activation of the motor system during movement, the tonic release might prepare and set the motor programs. Therefore, both the phasic and the tonic release of dopamine induced by L-DOPA should correct the disturbances of muscle tone and the alterations of motor performances in Parkinsonian patients. Another important aspect, which might account for L-DOPA's superiority with respect to other drugs, is the possibility that L-DOPA is not only the precursor of dopamine's synthesis but also regulates the formation of noradrenaline and, eventually, the release of serotonin and trace

amines in the extracellular space. Unfortunately, the 'optimal' activation of different receptor systems obtained with L-DOPA in an early phase fades. Thus, the 'honeymoon' with the use of the drug vanishes as more dopaminergic neurons degenerate. Therefore, despite L-DOPA administration, because the surviving dopaminergic cells cannot produce enough dopamine in advanced Parkinsonian patients, not only a phasic but also a tonic stimulation of dopaminergic receptors is no longer possible. Consequently, the Parkinsonian symptoms worsen.

Important negative side effects also develop during long-term L-DOPA treatment of PD (3–5 years), including attenuation of drug efficacy, motor fluctuations, and psychoses. Along with disease progression, and the consequent loss of dopaminergic terminals, the duration of clinical benefit derived from each L-DOPA dose shortens and patients begin to experience motor fluctuations such as 'end of dose deterioration' or the 'wearing off' phenomena. The 'off' periods do not necessarily respond to an increased frequency of L-DOPA intake. Indeed, the clinical response to L-DOPA becomes unpredictable and the patient may frequently switch from 'on' condition to off states. Accordingly, the plasma levels of L-DOPA fluctuate as well and exceed the therapeutic range, producing excessive, involuntary movements (dyskinesias). With further progression of the disease, dyskinesias become even more complex in pattern, with the occurrence of 'peak-dose dyskinesias,' and also 'end-of-dose dyskinesias.' Currently, the pathogenesis of these motor fluctuations is elusive, although both pre- and postsynaptic mechanisms likely play a critical role. Similarly, psychoses are a common side effect of long-term use of L-DOPA, affecting nearly 20% of treated patients. Managing psychotic symptoms is a very difficult challenge in terms of pharmacological treatment and for daily nursing.

For these reasons, some authors claim that it is preferable to avoid treating patients in the early phase of PD with L-DOPA and that, instead, it should be used later when the administration of dopaminergic agonists no longer controls symptoms. Nevertheless, there is not a consensus on this therapeutic approach because the optimal time for initiating L-DOPA therapy is still debated.

Variations in the L-DOPA's pharmacokinetic and alterations of the presynaptic and postsynaptic components in the dopaminergic system might account for the unpredictable responses to L-DOPA in individual patients. The unwanted effects of L-DOPA can be reduced by modifying the therapeutic schedule. For instance, it is important to avoid excessive fluctuations of L-DOPA concentrations in the blood and consequently in the brain. Thus, continuous stimulation of dopaminergic receptors should be obtained in order to attain an acceptable therapeutic outcome.

Several attempts have been made to reduce the oscillations in the extracellular concentrations of L-DOPA: (1) the number of daily administrations of L-DOPA has been increased; (2) the dietary intake of proteins has been reduced to favor a better penetration in the brain (the amino acids and L-DOPA compete on the same carrier at the blood–brain barrier); (3) although not practical, a continuous duodenal or intravenous infusion of L-DOPA has been used to diminish 'on–off' events; and (4) sustained-release preparations of L-DOPA have been used to avoid excessive plasmatic and cerebral peak concentrations of L-DOPA. This has been done to prolong the actions of each single dose of L-DOPA. However, there is no clear evidence that these formulations of L-DOPA reduce side effects.

It is also important to mention that established clinical studies have recognized that L-DOPA has a short-duration response (hours) and a long-duration response (days). It has been reported that the long-duration response is observed when consistent single doses of this drug are orally administered. Therefore, therapeutic regimens that use small and divided dosage during the day, as is currently done in clinical practice, might not achieve the beneficial long-duration response.

Other strategies, via the use of different pharmacological agents, have also been used to enhance the availability of L-DOPA in the brain. All these approaches reduce the degradation of L-DOPA, thus prolonging its effects.

It has been shown that inhibition of the catechol-*O*-methyltransferase enzyme potentiates L-DOPA efficacy. In fact, two inhibitors have been coadministered with L-DOPA to parkinsonian patients – tolcapone, which is centrally and peripherally acting, and entacapone, which does not cross the blood–brain barrier. These inhibitors prolong the effects of the drug, increasing on-time and reducing off-time. Of note, tolcapone has been associated with severe liver toxicity and hepatic failure, which represent a significant limitation to its clinical use. Thus, hepatic functions have to be carefully monitored during treatment.

It has been also shown that inhibition of the monoamine oxidase B mitochondrial enzyme, caused by the irreversible and selective inhibitor L-deprenyl (selegiline), reduces the on–off side effects of L-DOPA and produces modest symptomatic positive actions on Parkinsonian symptoms.

Some authors have also claimed that L-DOPA is toxic to the dopaminergic cells due to its oxidative metabolism. On the contrary, others have suggested that it might be neuroprotective. There is clear clinical

evidence demonstrating that L-DOPA is not toxic and possibly increases the lifetime of Parkinsonian patients.

L-DOPA and L-DOPA-Sensitive Dystonia

It has been shown that L-DOPA has a dramatic therapeutic action in reducing the symptoms of a form of primary dystonia, the DOPA-responsive dystonia (DYT5). This disease is caused by a deficit in the synthesis of dopamine. This fact explains the remarkable positive response to L-DOPA therapy. Interestingly, the therapeutic response to L-DOPA is maintained for years with a low incidence of motor fluctuations, which instead are often observed after long-term treatment in Parkinsonian patients.

Dopaminergic Agonists

Use of dopamine receptor agonists to treat PD has increased dramatically in the past two decades. In the CNS, dopaminergic receptors are involved in a wide range of functions, including locomotion, cognition, emotion, and endocrine responses. Whereas the D1 receptors are mainly distributed in the medium spiny striatal cells, giving rise to the so-called direct pathway projecting to the globus pallidus interna and substantia nigra pars reticulata, the D2 receptors are preferentially localized in the striatal neurons, giving rise to the indirect pathway projecting to the globus pallidus externa. Experimental and clinical studies have provided evidence that activation of D2 receptors exerts a central role in mediating the beneficial effects of dopaminergic agonists, although stimulation of both classes (D1 and D2) of receptors has been shown to be necessary for optimization of the clinical response. The dopaminergic agonists introduced in PD therapy exert their pharmacologic effects by directly activating dopaminergic receptors with different affinities (Table 1).

The use of dopamine receptor agonists has proven effective in the management of motor symptoms occurring in PD as adjunctive therapy to L-DOPA in advanced phases of disease. Moreover, compelling evidence supports the valuable role of dopaminergic agonists as a monotherapy in the early stages of PD. Although still a matter of debate, it is believed that early treatment with dopaminergic agonists could prevent or substantially reduce the neurochemical changes underlying the development of the disabling motor fluctuations commonly observed in the course of long-term L-DOPA therapy. Indeed, a significant advantage of dopaminergic agonists, compared to L-DOPA, is that they do not require metabolic conversion to an active compound to produce effects. Rather, they directly stimulate postsynaptic dopaminergic receptors, thus circumventing the synthetic processes of dopamine at degenerating nigrostriatal terminals. Moreover, these dopaminergic compounds have a significantly longer half-life compared to both standard and controlled-release formulations of L-DOPA, allowing a more continuous, rather than pulsatile, stimulation of target dopaminergic receptors. Another advantage compared to L-DOPA is that dopamine receptor agonists do not compete for both absorption and transport across the blood–brain barrier with circulating plasma amino acids. Prolonged stimulation of the D3 receptors by a selective dopaminergic agonist has been reported to activate neurogenesis of the dopaminergic system in a rat model of PD. These experimental results may have very important clinical relevance if confirmed in PD patients.

The concept of an initial monotherapy with a dopaminergic agonist, or an early combination of an agonist with L-DOPA, in an attempt to postpone disabling motor complications has been shown to be, at least to some extent, valid. Prospective long-term clinical trials have shown that early symptomatic treatment with a dopamine receptor agonist is associated with a significantly lower risk of developing motor

Table 1 Affinity of agonists for dopamine receptor subtypes

	D1-like		*D2-like*		
	D1	*D5*	*D2*	*D3*	*D4*
Bromocriptine	−	+	++	++	+
Lisuride	+	?	+++	+++	?
Pergolide	++	+	+++	+++	?
Cabergoline	0	?	+++	+++	?
Pramipexole	0	?	+++	+++	++
Ropinirole	0	0	+++	+++	+
Apomorphine	++	+	+++	+++	++
Rotigotine	+	?	++	+++	?

+/+++, agonist; −, antagonist; 0, no activity.

complications compared to L-DOPA therapy. Despite these encouraging results, the clinical efficacy of dopaminergic agonists has been shown to decrease over time, along with disease progression, so that the number of patients receiving these agents as monotherapy rapidly decreases to 50% after 2–4 years of treatment and to less than 20% after 5 years. In such circumstances, it is mandatory to introduce L-DOPA. In some cases, switching dopaminergic agonists could be a strategy to maintain a clinical response or to avoid specific adverse events. Moreover, another aspect deserving consideration is the combination of two dopamine receptor agonists, which appears to be an attractive alternative to delay initiation of L-DOPA therapy. This dual agonist's treatment has been shown to reduce the time spent in 'off' conditions and to improve the motor scores by nearly 20%. However, only few controlled and well-designed trials have been conducted in an attempt to compare the efficacy of the different marketed dopaminergic agonists in clinical practice.

Indeed, the way in which dopamine receptor agonists stimulate dopaminergic receptors appears to produce a tonic stimulation of the dopaminergic system (although this is an oversimplification); thus, a phasic stimulus is still necessary. Therefore, small amounts of L-DOPA (that might provide a phasic stimulus on dopaminergic receptors and, eventually, activation of other receptorial systems) are eventually required in patients treated with only direct dopaminergic agonists. This clinical observation highlights the importance of a phasic (neuronal-driven) stimulation of dopaminergic receptors that could be provided by L-DOPA, but it also suggests that the best way to activate striatal (and perhaps extrastriatal) dopaminergic receptors is by combining L-DOPA and dopaminergic agonists in the treatment of PD.

It is worth mentioning that a central dopaminergic dysfunction has also been implicated in a disease (restless leg syndrome) characterized by periodic limb movements during sleep that lead to severe insomnia. Therefore, this syndrome has an excellent response to low doses of direct dopaminergic agonists.

Ergoline Derivatives

Dopamine receptor agonists are classified into two major categories: ergolines, which are derivatives of ergot alkaloids that may be referred to as first-generation agents, and non-ergolines, more recently introduced in the pharmacotherapy of PD. Ergolines were introduced into clinical practice in the early 1970s. Bromocriptine, a bromo-substituted derivative of the natural peptide ergot alkaloid ergocryptine, was introduced for the treatment of hormonal disturbances, particularly for the treatment of prolactinomas, and was found to have anti-PD properties at higher dosages. It was first utilized as adjunctive therapy to L-DOPA in advanced PD patients experiencing motor fluctuations, and later it was proposed as monotherapy in the early stages of disease. A large controlled, 10 year trial directly compared bromocriptine with L-DOPA in early PD. At 3 years, patients receiving bromocriptine had lower disability scores than those in the L-DOPA arm of the trial; however, the difference was not significant at 9 years. Also, in two Cochrane reviews, patients receiving L-DOPA for 3 years were more likely to develop dyskinesias than patients receiving bromocriptine, although several patients from the bromocriptine group had to drop out because of intolerable side effects.

The second agonist introduced, lisuride, had initially been employed in the prevention of migraine attacks and then was approved for the treatment of PD. Similar to bromocriptine, lisuride has been utilized both as adjunct to L-DOPA and as monotherapy. Notably, subcutaneous lisuride infusion by means of micropumps has been shown to provide benefit in advanced PD patients. In a 4 year study comparing patients receiving subcutaneous lisuride and those receiving conventional L-DOPA and/or dopaminergic agonists, a significant reduction in both 'off' time and motor complications was obtained in the group treated with lisuride, with a long-term beneficial effect for the duration of the study.

The clavine derivative pergolide is a potent dopamine receptor agonist used as adjunct to L-DOPA therapy to retard the clinical progression of PD or to delay the occurrence of L-DOPA-induced motor complications. Several studies have been conducted comparing the efficacy of the ergot derivatives (bromocriptine and pergolide) as adjunctive therapy to L-DOPA. Despite many methodological differences within the study design, these trials have confirmed the efficacy of the combined approach in reducing the appearance of motor fluctuations. In fluctuating PD patients, these dopaminergic agonists may shorten the 'off' periods and increase the time spent in 'on' condition. Comparative studies among different dopamine receptor agonists have revealed that pergolide is more effective than bromocriptine as adjunctive therapy to L-DOPA in advanced PD patients. Both drugs proved to be effective compared to L-DOPA monotherapy, although pergolide treatment resulted in a more sustained improvement in terms of motor activity and quality of life, probably because of its longer half-life. The efficacy of pergolide has also been shown as monotherapy in early stages of PD.

Dihydroergocryptine is another ergoline derivative, showing affinity for D2 and partially for D1 receptors.

Controlled trials have been conducted both as monotherapy and as adjunctive to L-DOPA treatment.

The newest marketed ergoline derivative is the lysergic acid amide derivative cabergoline, which shows a higher affinity for D2 receptors. The main feature of this drug is its longer half-life (nearly 65 hours) compared to those of all other agonists, allowing a single daily intake. In adjunct to L-DOPA, cabergoline has been proven effective in ameliorating motor symptoms, compared to placebo, and in reducing the daily dosage of L-DOPA. Moreover, a controlled clinical trial conducted in patients in an early stage of PD showed that the incidence of motor fluctuations was significantly lower in the cabergoline group compared to the group treated with L-DOPA.

Non-Ergoline Derivatives

Two dopaminergic agonists with no relation to ergot derivatives have been introduced in the pharmacotherapy of PD: the indole derivative ropinirole and the benzothiazol derivative pramipexole.

Ropinirole has high affinity for both D3 and D2 dopaminergic receptors, with little affinity for D1 receptors. Rapidly absorbed and well tolerated, it was reported to be as efficacious as bromocriptine in reducing motor fluctuations and decreasing L-DOPA daily dosage. As monotherapy, ropinirole was proven effective in treating tremor in PD patients. Several studies have demonstrated that ropinirole is effective both in the early and in the advanced phases of PD. A 5 year controlled clinical trial in newly diagnosed PD patients has been conducted comparing the effects of ropinirole to L-DOPA. After 5 years, the occurrence of dyskinesias was 20% in the group of patients receiving ropinirole compared to 45% in the L-DOPA group.

Pramipexole, a non-ergoline derivative, features a preferential affinity for D3 over D2 dopamine receptor subtypes. Its efficacy and safety have been investigated in several clinical trials, which have demonstrated that pramipexole utilized both as adjunct to L-DOPA and as monotherapy has beneficial effects on motor symptoms, significantly reduces the periods spent in 'off' condition, and allows a reduction in the daily dosage of L-DOPA. In particular, patients with pronounced resting tremor derived a clear benefit from pramipexole use compared to L-DOPA.

Among non-ergot derivatives, the most recent agent introduced in the pharmacotherapy of PD is the aminotetraline derivative rotigotine, considered as a D3/D2/D1-preferring dopamine receptor agonist. Indeed, rotigotine showed agonistic activity on dopaminergic receptors with higher affinity (approximately 20-fold) for the D3 compared to the D2 (approximately 100-fold) and the D1 receptor. Despite the short duration of rotigotine efficacy on oral administration, the ability of the drug to act after transdermal application has provided a means of developing its clinical potential. Based on the pharmacologic profile of the agent and its effect in experimental models, subsequent clinical investigation has shown that transdermal rotigotine is an effective and long-acting treatment for the control of motor symptoms in PD. In two large clinical trials, early PD patients receiving rotigotine experienced a significant improvement in motor symptoms compared to placebo. Moreover, patients in more advanced disease stage, utilizing rotigotine as adjunctive therapy with L-DOPA resulted in a significant reduction in the time spent in 'off' condition.

Concerning the achievement of satisfactory results with dopaminergic agonists, the route of delivery is gaining importance more than their supposed receptor affinity. A 'continuous dopaminergic stimulation' concept of the treatment of PD suggests that the use of long-acting dopaminergic agonists, such as transdermal patches or continuous subcutaneous infusion, may indeed produce effective control of motor symptoms, accompanied by a reduction of the motor fluctuations induced by L-DOPA.

In the past decade, there has been renewed interest in apomorphine. The tetracyclic apomorphine is a morphine derivative but without opiate-like effects. It was introduced in the late nineteenth century for the treatment of intoxications because of its powerful emetic effects. It is a short-acting dopaminergic agonist with receptor affinity similar to that of dopamine, thereby representing the only agonist with a pharmacological profile comparable to that of dopamine. Apomorphine is commonly prescribed for subcutaneous use as bolus injection or continuous infusion by micropumps. The subcutaneous injection is commonly used in clinical practice as a confirmatory, diagnostic test for PD. Moreover, apomorphine is currently utilized for the management of severe, refractory 'off' periods in fluctuating PD patients because of its rapid onset and good efficacy. Several studies have also demonstrated that continuous subcutaneous infusion of apomorphine results in reduced 'off' periods, reduced occurrence of dyskinesias, and reduced daily L-DOPA dosage. Despite compelling evidence of both efficacy and safety, the validity of apomorphine as an important adjuvant in the advanced stages of PD is still underestimated.

Side Effects

Despite a significantly different affinity for the distinct classes of dopamine receptor subtypes, the beneficial

Table 2 Side effects of dopamine receptor agonists utilized in the therapy of Parkinson's disease

	Bromocriptine	*Lisuride*	*Pergolide*	*Cabergoline*	*Pramipexole*	*Ropinirole*	*Apomorphone*
Gastrointestinal	+	+	+	+++	±	n.r.	++
Nausea	+++	++	++	+	+++	++++	+++
Emesis	+	+	+	±	±	++	+++
Hypotension	++	+	++	+++	+++	++	++
Dyskinesia	++	+	+++	+++	+++	+++	++
Psychiatric	±	+	+	n.r.	n.r.	±	±
Hallucinations	+	+	+++	+++	+++	++	±
Confusion	+++	++	±	+++	+++	+	±
Sleepiness	++	+	±	++	++	+	+
Sleep attacks	n.r.	±	n.r.	±	±	±	++
Edema	±	+	±	+	+	+	+

Reported incidence: ±, 0–1%; +, 1–5%; ++, 5–10%; +++, 10–25%; ++++, >25%; n.r., no specific reports of effect.

clinical and side effects of dopaminergic receptor agonists are remarkably similar, independent of their chemical structure (Table 2). Indeed, expectations of clinical improvements claimed on the basis of their biochemical and pharmacological profile have not been met.

Experimental and clinical evidence indicates a central role for the pharmacokinetics of dopaminergic agonists in the development of side effects. The commonly observed undesired effects, such as nausea, dizziness, somnolence, emesis, and orthostatic hypotension, are not in fact directly related to the effective plasma levels but, rather, seem to be triggered by the rapid increase in drug level in the blood, probably related to drug bioavailability and individual susceptibility. Although in some PD patients the occurrence of these side effects requires the discontinuation of the drugs, in a much larger number of patients ongoing treatment with slowly increasing dosages of agonists permits the development of tolerance for the undesired effects while preserving the beneficial actions. To some extent, the use of antagonists of peripheral dopamine D2 receptors, such as domperidone, is helpful in preventing some of the side effects caused by the dopaminergic agonists and enables a more rapid titration of the drugs within a few weeks. However, despite these therapeutic precautions, serious side effects may occur, and they may be comparable to those observed with L-DOPA therapy. These also include psychiatric symptoms, such as hallucinations, paranoia, vivid dreams, mood alterations, compulsive gambling, and compulsive sexual behavior, and they appear to be more frequent in elderly patients or in patients with preexisting cognitive impairment. Similarly, excessive daytime somnolence is very common and often complicated by sudden sleep attacks. Sleep attacks have been shown to occur with virtually all of the dopaminergic agonists.

Ergoline derivatives have also been associated with pleuropulmonary and retroperitoneal fibrosis, so careful patient selection is required. Thus, during treatment patients should be monitored for persistent cough, dyspnea, chest and abdominal pain, and cardiac failure. Pergolide and cabergoline have been associated with a significantly higher incidence of valvular heart disease, which will certainly limit their use in clinical practice. Thus, in the course of ergot-derived agonist treatment, an annual screening with chest x-rays, blood tests, and echocardiographic examination should be warranted.

Conclusion

Considering all the points discussed in this article, we believe that the dosage of administered L-DOPA and dopaminergic agonists should be customized to the needs of the individual patient on the basis of the clinical response and the profile of adverse events.

See also: Animal Models of Parkinson's Disease; Dopamine; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopamine Neurons: Reward and Uncertainty; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Restless Leg Syndrome.

Further Reading

Bonuccelli U and Pavese N (2006) Dopamine agonists in the treatment of Parkinson's disease. *Expert Review of Neurotherapeutics* 6: 81–89.

Calne DB (1993) Treatment of Parkinson's disease. *New England Journal of Medicine* 329: 1021–1027.

Colosimo C, Merello M, and Albanese A (1994) Clinical usefulness of apomorphine in movement disorders. *Clinical Neuropharmacology* 17: 243–259.

Colzi A, Turner K, and Lees AJ (1998) Continuous subcutaneous waking day apomorphine in the long term treatment of levodopa induced interdose dyskinesias in Parkinson's disease. *Journal of Neurology, Neurosurgery, & Psychiatry* 64: 573–576.

Fahn S, Shoulson I, Kieburtz K, et al. (2004) Levodopa and the progression of Parkinson's disease. The Parkinson Study Group. *New England Journal of Medicine* 351: 2498–2508.

Hutton JT, Metman LV, Chase TN, et al. (2001) Transdermal dopaminergic D(2) receptor agonist therapy in Parkinson's disease with N-0923 TDS: A double-blind, placebo-controlled study. *Movements Disorders* 16: 459–463.

Jankovic JJ (2002) Therapeutic strategies in Parkinson's disease. In: Jankovic JJ and Tolosa E (eds.) *Parkinson's Disease and Movement Disorders,* 4th edn., pp. 116–151. Philadelphia: Lippincott Williams & Wilkins.

Mercuri NB and Bernardi G (2005) The 'magic' of L-DOPA: Why is it the gold standard Parkinson's disease therapy? *Trends in Pharmacological Sciences* 26: 341–344.

Olanow CW, Agid Y, Mizuno Y, et al. (2004) Levodopa in the treatment of Parkinson's disease: Current controversies. *Movement Disorders* 19: 997–1005.

Parkinson's Disease Research Group in the United Kingdom (1993) Comparisons of therapeutic effects of levodopa, levodopa and selegiline, and bromocriptine in patients with early, mild Parkinson's disease: Three year interim report. *British Medical Journal* 307: 469–472.

Parkinson Study Group (2000) Pramipexole vs. levodopa as initial treatment for Parkinson disease. A randomized controlled trial. *Journal of the American Medical Association* 284: 1931–1938.

Poewe W and Wenning G (2002) Levodopa in Parkinson's disease: Mechanisms of action and pathophysiology of late failure. In: Jankovic JJ and Tolosa E (eds.) *Parkinson's Disease and Movement Disorders,* 4th edn., pp. 104–115. Philadelphia: Lippincott Williams & Wilkins.

Rascol O, Payoux P, Ory F, et al. (2003) Limitations of current Parkinson's disease therapy. *Annals of Neurology* 53: S3–S15.

Stocchi F, Ruggieri S, Vacca L, et al. (2002) Prospective randomized trial of lisuride infusion versus oral levodopa in patients with Parkinson's disease. *Brain* 125: 2058–2066.

Zanettini R, Antonini A, Gatto G, et al. (2007) Valvular heart disease and the use of dopamine agonists for Parkinson's disease. *New England Journal of Medicine* 356: 39–46.

Relevant Websites

http://www.movementdisorders.org – The Movement Disorder Society.

http://www.wemove.org – WE MOVE.

Dopaminergic Differentiation

A Rosenthal and J C Lin, Rinat Neuroscience Corp., South San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Dopaminergic (DA) neurons in the mammalian midbrain region are critically affected in several neuropsychiatric disorders, such as schizophrenia, and they are specifically lost in Parkinson's disease (PD). This article reviews how the DA neurons are generated through the interactions of multiple extrinsic and intrinsic factors during embryogenesis. The molecular identities and the mechanisms of action of an increasing number of such factors have been elucidated. Many of the same factors have been successfully used to induce efficient differentiation of DA neurons *in vitro* from embryonic or neural stem cells. These advances promise to have significant impacts on science and medicine in the future.

Overview: The Medical Importance of DA Neuron Development

Dopamine is a major catecholamine neurotransmitter of the mammalian central nervous system (CNS). It is also present in noradrenergic and adrenergic neurons as a biosynthetic precursor to noradrenalin and adrenaline. Several distinct populations of DA neurons can be identified in the CNS. They include a subset of interneurons in the olfactory bulb, a subset of amacrine interneurons of the retina, distinct neuronal groups in the hypothalamus, and the ventral mesencephalon, which includes substantia nigra (SN, the so-called nigrostriatal system), and ventral tegmental area (VTA, the so-called mesolimbic system). Medically, the most important group of DA neurons, which is also the largest population by number, of the mammalian CNS is in the ventral mesencephalon. The nigrostriatal DA neurons are selectively lost in PD, whereas dysfunction of the mesolimbic DA neurons is implicated in obsessive–compulsive disorder, schizophrenia, and substance abuse.

Since PD is marked by progressive loss of the nigrostriatal DA neurons, 'dopamine supplementation' therapy such as L-dopa and dopamine receptor agonists has been the mainstay of the symptomatic management of this most common movement disorder. However, the standard dopamine supplementation gradually becomes ineffective after the first 2–4 years of treatment, presumably due to the inexorable loss of nigrostriatal DA neurons and eventually their synaptic targets. Even when effective, the current DA agonist medications are frequently accompanied by fluctuations, dyskinesias, and psychotic side effects, which can be attributed at least in part to improperly regulated and ectopic dopamine transmission. Consequently, the concept of 'DA cell replacement' gradually emerged as a potential alternative to dopamine supplementation. Transplantation of DA neurons or their precursors in the striatum or in the substantia nigra may restore physiologically regulated and anatomically precise dopamine release. However, there are severe technical and ethical constraints on the use of human tissue for transplantation. As a result, attention has been directed increasingly toward deriving DA neurons from embryonic stem (ES) cells as well as from adult-derived neural stem cells. Understanding the mechanism by which normal mesencephalic DA neurons develop, therefore, has become important, not only as a model of neuronal subtype differentiation in neural development, but also for its medical implications. This article focuses on the developmental mechanism of DA neurons in the ventral midbrain.

Extrinsic Factors in Mesencephalic DA Neuron Development

Mesencephalic DA neurons are generated in the mid–hindbrain boundary (MHB) under the influence of two major signaling centers, the floor plate and the isthmus, of developing embryos. Sonic hedgehog (Shh) and fibroblast growth factor-8 (FGF-8) are the principal secreted signals underlying the activity of each of these signaling centers, respectively. Shh determines the location of DA neurons along the dorsal–ventral axis, whereas FGF-8 positions the mesencephalic DA neurons along the anterior–posterior axis of the neural tube. The Gli proteins are a conserved family of zinc finger transcription factors that function as both activators and repressors in transducing the Shh signal inside the cells. Forced Gli-1 expression in transgenic mice induced ectopic mesencephalic DA neurons while Gli-2-deficient mouse embryos failed to develop mesencephalic DA neurons. These results further underscore the importance of the Shh signaling pathway in the development of mesencephalic DA neurons.

How do these early patterning signals originate during embryogenesis? Shh expression in the floor plate was shown to be controlled by the forkhead transcription factor FoxA1/HNF3β in mouse and zebra fish, the basic helix-loop-helix (bHLH) and PAS domain containing transcription factors Sim-1 and -2 in the mouse, and directly by TGFβ family factor Cyclop in

the zebra fish. Likewise, FGF-8 expression in the MHB is controlled by a complex network of regulatory genes including a key homeodomain transcription factor Pax2. These early developmental events that initiate, refine, and maintain the pattern and signaling centers along the anterior–posterior and dorsal–ventral axes in the MHB have been reviewed extensively elsewhere.

In addition to the patterning signals intersecting at the ventral MHB of the early embryo, several other secreted factors can induce the differentiation or support the survival of DA neurons *in vitro* and *in vivo*. Among them, glia-derived neurotrophic factor (GDNF) is the most potent trophic factor for mesencephalic DA neurons. The GDNF−/− mice initially develop a full complement of midbrain DA neurons, but these mice die soon after birth, precluding the analysis of later survival and function of DA neurons in the absence of GDNF. When transplanted into the striatum of a 1-methyl-4-phenyl-1,2,3, 6-tetrahydropyridine-lesioned wild-type mouse, the GDNF mutant (−/−) donor tissues exhibit a greatly reduced survival of DA neurons compared with the wild-type (+/+) donor tissues. Moreover, such a defect can be rescued by immersion of the GDNF−/− donor tissues with a high concentration of GDNF prior to transplantation. Thus GDNF is needed for continued survival of mesencephalic DA neurons in the adult. This is particularly significant in view of early promising clinical studies in which continuous and localized delivery of recombinant GDNF into the SN by a pump led to a rapid and significant rescue of the DA neurons of PD patients. Significant technical hurdles, such as efficient drug diffusion in the striatum, remain to be resolved in order to achieve optimal outcome in such applications of a potentially neuroprotective or regenerative agent.

Intrinsic Factors in Mesencephalic DA Neuron Development

Once the DA neuronal fate is induced by the extrinsic signals in the ventral MHB region, several cell-autonomous transcription factors are required for initiation and maintenance of the expression of DA cell-specific genes as well as for survival of DA neurons. They include several homeodomain transcription factors (En1, En2, Msx1, Lmx1a, Lmx1b), a bHLH transcription factor (Ngn-2), and a nuclear receptor family transcription factor (Nurr1) as described below in more detail.

Two homeodomain transcription factors, En-1 and En-2, are highly expressed in the developing and mature mesencephalic DA neurons as well as other cell types in the MHB. Although the midbrain DA neurons are initially specified in single- or double-En knockout mice, their survival requires the En1 and -2 genes in a dosage-dependent manner. In the En1/En2 double knockout mice, a very small population of TH-positive mesencephalic DA neurons remains, but the expression of α-synuclein, a gene mutated in a familial form of PD, is completely lost in the ventral midbrain, suggesting that En genes may directly regulate α-synuclein expression in the DA neurons. The loss of α-synuclein expression, however, is unlikely to be responsible for the loss of DA neurons in the En1/En2 double knockout mice since the α-synuclein knockout mice maintain a full complement of midbrain TH neurons.

A LIM homeodomain transcription factor, Lmx-1a, and another homeodomain protein, Msx-1, are specifically turned on by Shh in the ventral MHB. Lmx-1a is sufficient and required to induce DA neuron differentiation *in vivo*. One function of Lmx-1a is to induce the expression of Msx-1, which further turns on the proneural gene Neurogenin-2. The absence of Neurogenin-2 in mice resulted in progressive loss of DA neurons in both the SN and the VTA. It is important to note that forced expression of Lmx-1a can drive robust DA cell differentiation in ES cell culture. It is intriguing to note that Lmx-1a and Msx-1 are both expressed in dorsal neural tube and required for the development of the roof plate. These two dorsal factors are somehow co-opted to serve in the development of a ventral cell type, that is, the DA neurons, specifically in the MHB region.

Lmx-1b, another LIM homeodomain transcription factor, is initially highly expressed in the MHB and the dorsal neural tube and later expressed in developing and adult mesencephalic DA neurons. The early expression of Lmx-1b is induced and/or maintained by FGF-8, and retroviral-mediated Lmx-1b expression is able to induce ectopic Wnt-1 expression, another secreted signal of the MHB. In the absence of Lmx-1b, the ventral mesencephalic DA neurons retain expression of Nurr1 and TH, but most of them lose the expression of a homeodomain transcription factor Pitx3. These results suggest that Lmx-1b and Pitx3 form a regulatory cascade independent of the expression of Nurr1 or TH in DA neurons.

Another gene that attracted significant attention is Nurr1, an orphan nuclear receptor that is widely expressed in the CNS including in the mesencephalic DA neurons. In the absence of Nurr1, mice initially still develop a group of Pitx-3-, Lmx-1b-, and HNF3β-expressing cells in the ventral MHB region but fail to initiate TH expression. These 'prospective' DA cells eventually degenerate and disappear, resulting in a complete loss of mature DA neurons. Conversely, forced expression of Nurr1 in the adult hippocampus-derived progenitors can induce TH expression without affecting other DA-specific genes,

suggesting a direct regulation of TH expression by Nurr1. Furthermore, overexpression of Nurr1 can induce the differentiation of a mature DA neuronal phenotype by a cerebellum-derived immortalized cell line in the presence of the type 1 astrocytes. Therefore, Nurr1 appears to be a transcriptional regulator of TH expression and is essential for DA neuronal differentiation and survival.

Development of Distinct Subtypes of Mesencephalic DA Neurons

The mesencephalic DA neurons are mainly distributed in two distinct anatomical structures, SN DA neurons, innervating the striatum, and VTA DA neurons, innervating the nucleus accumbens and the prefrontal cortex. Little is known about how these two groups of DA neurons develop since they largely express a common set of genes. Raldh-1 (class 1 aldehyde dehydrogenase) is one of the first genes characterized to exhibit a subgroup-specific expression in mesencephalic DA neurons. The Raldh-1-expressing axonal terminals form a gradient in the striatum and nucleus accumbens such that its density is highest in the rostral and dorsal region of the basal forebrain. Such a graded distribution suggests that Raldh-1 is preferentially expressed in the DA neurons of SN instead of those of the VTA. Given the importance of retinoic acid in early embryogenesis, hindbrain patterning, and spinal motor neuron development, such a striking expression pattern of a retinoic acid-synthesizing enzyme suggested that retinoic acid and/or its derivatives may play a role in the development of subsets of DA neurons.

Some other genes, while not necessarily exclusively expressed in any specific subset of DA neurons, are nevertheless required for the development or maintenance of only a subpopulation of mesencephalic DA neurons. Genetic deletion of mouse transforming growth factor-α (TGF-α), for instance, leads to a selective reduction of DA neurons in SN, but not in VTA. This appears to reflect an early, possibly a developmental, requirement of TGF-α since such a defect was already present in the newborns. Neither the site of expression nor the mechanism of TGF-α action in a subset of the SN DA neurons is known.

The brain-derived neurotrophic factor (BDNF) and its receptor trkB are expressed in the mesencephalic DA neurons. BDNF can promote DA neuronal survival and differentiation in experimental culture conditions. When researchers conditionally deleted the BDNF gene by crossing the floxed allele of BDNF and a Wnt–Cre line, BDNF expression was nearly completely abolished in the MHB region. Such mouse mutants exhibit a ~23% reduction in tyrosine hydroxylase (TH)+ DA neurons in the substantia nigra compacta but not in the VTA, indicating a preferential requirement of BDNF in DA neuron development of the SN.

One transcription factor that regulates the differentiation or maintenance of a subset of DA neurons is Pitx-3/Ptx-3, a paired-type homeodomain transcription factor that is expressed almost exclusively in all the mesencephalic DA neurons. It is interesting that the number of DA neurons fell dramatically in the SN, but not the VTA, in the Aphakia mutant mice that are deficient in the Pitx-3 gene. This is accompanied by ~90% loss of dopamine in the dorsal striatum and significant deficits in sensorimotor coordination behaviors. These deficits can be reversed by peripheral administration of L-dopa. Finally, the Aphakia mutant mice also displayed denervation hypersensitivity to L-dopa in the dorsal striatum, making Aphakia one of the best mouse models with all the major symptomatic and pharmacologic characteristics of the early clinical stages of PD.

The Control of DA Axonal Development

DA neurons project their axons in the rostral direction to form the medial forebrain bundle (MFB) before reaching the final targets in the forebrain, which include the corpus striatum, nucleus accumbens, and the prefrontal cortex. MHB and DA neuron co-culture experiments have shown that the MHB tissue alone is not repulsive to the DA axons, suggesting that no localized repulsive signal from the MHB and rostral hindbrain region prohibits the DA axons from entering the hindbrain. In contrast, the DA axons do turn abruptly when forced to grow into a stratum of the reverse polarity, suggesting that locally distributed directional cues guide the rostral projection of DA axons to their precise positions in their physiological targets. The absence of repulsive signals in the hindbrain suggests that the DA axons may turn rostrally in response to attractive guidance cue(s) distributed in the vicinity of ventral midbrain.

Unlike many ascending axons originating from the spinal cord and elsewhere, the mesencephalic DA axons never cross the midline. Once turned rostrally, the DA axons join additional ascending axons to form the MFB, a symmetrical pair of axonal bundles running in parallel to the ventral midline of the forebrain. Studies in the *Drosophila* embryo have shown that longitudinal axonal pathways are kept from crossing the midline and maintained at a constant distance from the midline by the repulsive signal Slit. Similarly, mammalian Slit-1 and Slit-2 are both expressed in the ventral midline of the CNS. The TH-positive axons, as well as other axons in the MFB, are displaced ventrally in Slit2−/− mutant mice

and frequently enter the hypothalamus in Slit1−/−; Slit2−/− double mutants. Therefore, the repulsive Slit signals contribute to the maintenance of DA axonal trajectory in the ventral forebrain.

Finally, DA axons invade selected targets in the forebrain, that is, the striatum and the prefrontal cortex, in a topographic manner. The SN DA axons preferentially innervate the dorsal striatum, while the VTA DA axons the ventral striatum (mainly nucleus accumbens) and the prefrontal cortex. The Eph receptor–ligand system has been implicated in the formation of the DA topographical map. EphB1 is preferentially expressed in the SN DA neurons, while its ligand ephrin-B2 shows a graded expression in the DA axonal targets, with the highest level in the ventral striatum. Furthermore, ephrin-B2 specifically inhibits neurite outgrowth of SN but not VTA, suggesting that repulsive interaction between EphB1 and ephrin-B2 may underlie at least in part the formation of the DA topographic map. This proposal awaits confirmation by gain and loss of function experiments *in vivo*.

DA Neuron Differentiation *In Vitro*

Current understanding of the mechanisms by which the mesencephalic DA neurons are generated during normal development has been successfully extended to useful methods of promoting DA neuronal differentiation from the totipotent ES cells. Addition of Shh and FGF-8 to the ES cell culture increases the efficiency of DA cell differentiation from ~5% to 20–30%. Similar results were also obtained with clonal ES cell lines derived from nuclear transplantation of adult donor cells (ntES cells). It is interesting that different ntES cell lines produced DA neurons at highly variable frequencies, with one line generating 50% DA neurons, suggesting that the inherent potential of DA neuron differentiation is ES cell line-dependent.

Such intrinsic factors with the ability to influence DA neuron differentiation have been identified, including Lmx-1a and Nurr1. As mentioned earlier, Lmx-1a expression in mouse ES cells can drive up to 60% TH+ neuronal differentiation in the presence of basic fibroblast growth factor, FGF-8, and Shh. Essentially all these TH+ neurons also express the correct mesencephalic DA neuronal phenotype, including dopamine transporter, Lmx1a, En, Nurr1, and Pitx3. A separate study found that the efficiency of DA differentiation can be increased up to 80% by adding Shh and FGF-8 to an ES cell line forced to express Nurr-1. The DA cells derived from these experimental protocols express not only dopamine-synthesizing enzymes and dopamine receptor and transporter but also molecular characteristics that are specific to the differentiating and mature mesencepahlic DA neurons (e.g., cRet, Pitx-3, En-1). When transplanted into the striatum, these cells can correct the amphetamine-induced ipsilateral rotation of rats whose mesencephalic DA neurons were ablated by 6-hydroxydopamine in one side. Moreover, animals grafted with Nurr-1-expressing ES cells not only exhibited contralateral turning on amphetamine stimulation but also manifested spontaneous contralateral rotations. This phenomenon, which results from overproduction of dopamine in one side of the brain, demonstrates a robust differentiation of Nurr-1 ES cells into functional dopamine-producing DA neurons *in vivo*.

Given the success in inducing ES cells to the DA lineage, promoting the differentiation of adult or fetal mesencephalic neural progenitor cells would seem conceptually straightforward. This approach, however, has generally achieved a lower percentage of DA neuron production compared with ES cells. One possible explanation for this discrepancy relates to the fact that the tissue oxygen pressure in the developing and adult brain ranges from 1–5% whereas the standard techniques expose the neural progenitor cells to the ambient 20% of oxygen pressure. Lowering the tissue culture condition to 3% oxygen significantly increases the proliferation of mesencepahlic progenitor cells, decreases their death, and most important, enhanced efficiency of their differentiation into DA neurons by up to 56% (compared with 18% at the regular 20% oxygen condition). Nurr-1 and Shh expression levels were not altered by low oxygen, but FGF-8, vascular endothelial growth factor (VEGF), and erythropoietin (EPO) messenger RNA levels are significantly increased by low oxygen tension. Neither adding nor blocking VEGF affected DA differentiation. However, FGF-8 exposure enhanced mesencephalic progenitor proliferation and delayed the differentiation of TH+ cells. It is interesting that recombinant EPO protein increased the yield of DA neurons from mesencephalic progenitors at 20% oxygen environment in a dose-dependent manner, and an EPO-neutralizing antibody partially abolished the efficiency of DA differentiation at the 3% oxygen condition. It will be interesting to learn whether and how EPO acts as an inducer of DA neurons *in vivo* as well as in the ES cell culture system.

Conclusion

In the past decade, considerable advances have been made at an ever accelerating pace in understanding the cellular and molecular mechanisms controlling the development of mesencephalic DA neurons. Indeed the extrinsic and intrinsic signals underlying native DA neuron development have turned out to be

highly relevant and useful in efforts at directing DA differentiation from ES cells *in vitro*. Conversely, new observations from *in vitro* culture systems have revealed unexpected molecular mechanisms influencing DA neuron differentiation, which will inspire further studies of their roles *in vivo*. While the recent progress was fueled by the expectation that *in vitro* production of DA neurons would be useful as a form of cell replacement therapy for PD, it is quite possible that 'neuroprotective' or 'regenerative' medicines for PD can also derive from the knowledge of DA neuron development.

An update on latest advances is as follows:

1. DA neurons were found to be greatly reduced in mice in which Dicer, a critical enzyme essential for micro-RNA formation, has been conditionally deleted in DA neurons. Furthermore, a specific micro-RNA, miR133b, is enriched in the midbrain DA neurons and can negatively regulate Pitx3 expression and the maturation of dopaminergic neurons.
2. As a definitive confirmation of the importance of GDNF in DA neuronal development and maintenance, conditional deletion of GDNF in adult mice resulted in severe hypokinesia and loss of DA neurons.

See also: Dopamine in Perspective; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopamine Neurons: Reward and Uncertainty; Drosophila Apterous Neurons: from Stem Cell to Unique Neuron; Motor Neuron Specification in Vertebrates; Neural Crest Cell Diversification and Specification: Melanocytes; Neural Crest Cell Diversification and Specification: ErbB Role; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation.

Further Reading

Andersson E, Tryggvason U, Deng Q, et al. (2006) Identification of intrinsic determinants of midbrain dopamine neurons. *Cell* 124: 393–405.

Baquet ZC, Bickford PC, and Jones KR (2005) Brain-derived neurotrophic factor is required for the establishment of the proper number of dopaminergic neurons in the substantia nigra pars compacta. *Journal of Neuroscience* 25: 6251–6259.

Blum M (1998) A null mutation in TGF-alpha leads to a reduction in midbrain dopaminergic neurons in the substantia nigra. *Nature Neuroscience* 1: 374–377.

Gash DM, Zhang Z, Ai Y, Grondin R, Coffey R, and Gerhardt GA (2005) Trophic factor distribution predicts functional recovery in parkinsonian monkeys. *Annals of Neurology* 58: 224–233.

Grimm J, Mueller A, Hefti F, and Rosenthal A (2004) Molecular basis for catecholaminergic neuron diversity. *Proceedings of the National Academy of Sciences of the United States of America* 101: 13891–13896.

Hwang DY, Fleming SM, Ardayfio P, et al. (2005) 3,4-dihydroxyphenylalanine reverses the motor deficits in Pitx3-deficient aphakia mice: Behavioral characterization of a novel genetic model of Parkinson's disease. *Journal of Neuroscience* 25: 2132–2137.

Hynes M, Stone DM, Dowd M, et al. (1997) Control of cell pattern in the neural tube by the zinc finger transcription factor and oncogene Gli-1. *Neuron* 19: 15–26.

Kele J, Simplicio N, Ferri AL, et al. (2006) Neurogenin 2 is required for the development of ventral midbrain dopaminergic neurons. *Development* 133: 495–505.

Kim J, Inoue K, Ishii J, et al. (2007) A microRNA feedback circuit in midbrain dopamine neurons. *Science* 317: 1220–1224.

Kim JH, Auerbach JM, Rodriguez-Gomez JA, et al. (2002) Dopamine neurons derived from embryonic stem cells function in an animal model of Parkinson's disease. *Nature* 418: 50–56.

Morrison SJ, Csete M, Groves AK, Melega W, Wold B, and Anderson DJ (2000) Culture in reduced levels of oxygen promotes clonogenic sympathoadrenal differentiation by isolated neural crest stem cells. *Journal of Neuroscience* 20: 7370–7376.

Nunes I, Tovmasian LT, Silva RM, Burke RE, and Goff SP (2003) Pitx3 is required for development of substantia nigra dopaminergic neurons. *Proceedings of the National Academy of Sciences of the United States of America* 100: 4245–4250.

Pascual A, Hidalgo-Figueroa M, Piruat JI, Pintado CO, Gomez-Diaz R, and Lopez-Barneo J (2008) Absolute requirement of GDNF for adult catecholaminergic neuron survial. *Nature Neuroscience* 11: 755–761.

Smidt MP, Asbreuk CH, Cox JJ, Chen H, Johnson RL, and Burbach JP (2000) A second independent pathway for development of mesencephalic dopaminergic neurons requires Lmx1b. *Nature Neuroscience* 3: 337–341.

Wagner J, Akerud P, Castro DS, et al. (1999) Induction of a midbrain dopaminergic phenotype in Nurr1-overexpressing neural stem cells by type 1 astrocytes. *Nature Biotechnology* 17: 653–659.

Ye W, Bouchard M, Stone D, et al. (2001) Distinct regulators control the expression of the mid-hindbrain organizer signal FGF8. *Nature Neuroscience* 4: 1175–1181.

Ye W, Shimamura K, Rubenstein JL, Hynes MA, and Rosenthal A (1998) FGF and Shh signals control dopaminergic and serotonergic cell fate in the anterior neural plate. *Cell* 93: 755–766.

Dorsal Root Ganglion Neurons

A M Rush, NeuroSolutions Ltd., Coventry, UK
S G Waxman, Yale University School of Medicine, New Haven, CT, USA; and VA Connecticut Healthcare System, West Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The dorsal root ganglia (DRG) are a collection of cell bodies of the afferent sensory fibers, which lie between adjacent vertebrae. Unlike motor neurons, the soma is on a side branch of the main axon and receives little synaptic input and has sometimes been written off as unimportant and uninteresting, perhaps simply working as the metabolic factory of the neuron. Access to the soma for techniques such as patch clamp is relatively easy and can provide excellent recording conditions, making it a convenient model neuron. A substantial body of work during the past two decades has shown that changes in excitability of DRG neurons are representative of functional changes in the associated fibers, and that these alterations are of importance in many pathological conditions. In addition, these studies have increased our understanding of the very different roles particular ion channels play in influencing excitability of DRG and also how they may alter firing behavior in other neurons in which they are expressed or misexpressed. One field of research that has been especially revealing is the study of sodium channel subtypes in DRG, and this article draws on examples from this area to illustrate the control of excitability in DRG by plasticity of these proteins.

Expression of Multiple Ion Channel Subtypes

It has been known for some time that it is possible to link the conduction velocity of the dorsal root fibers with DRG cell size and to classify these neurons into four main groups of decreasing velocity: Aα, 30–55 m s^{-1} Aβ, 14–30 m s^{-1}; Aδ, 2.2–8 m s^{-1}; and C, <1.4 m s^{-1}. These fibers relay information from muscle and skeletal mechanoreceptors (Aα and Aβ), cutaneous and subcutaneous mechanoreceptors (Aβ and Aδ), and nociceptors (Aδ and C). The slower velocity fibers are generally associated with small cells, and the action potential duration is broader, often displaying a hump on the falling phase. Some of these characteristics are obviously related to the intrinsic behavior of, for instance, a myelinated axon versus a nonmyelinated one, but it has become increasingly clear that the excitability of these fibers, and the plasticity of that excitability, is directly related to the subtypes and combinations of ion channel that they express. In particular, differential expression and modulation of sodium channels, which are responsible for subthreshold depolarizations, as well as the upstroke of the action potential, have been demonstrated to be important modulators of excitability.

Initial clues came from studies of C-fibers (e.g., from sciatic nerve) that generated a slow sodium-dependent spike that was resistant to tetrodotoxin (TTX). Subsequently, this phenomenon was linked to the presence of TTX-R sodium channels and later to the specific isoform responsible for the current, $Na_v1.8$. Multiple studies have now shown that there are large variations in sodium current parameters in DRG neurons linked to expression of a heterogeneous population of sodium channels. Apart from pharmacological intervention using TTX, one can further dissect the channels by using parallel methods of immunohistochemistry and electrophysiological recording, giving insight into the expression patterns and possible roles. A convenient classification method is to use cell size. Using immunocytochemical techniques, large DRG cells have been shown to predominantly express TTX-S channels, such as $Na_v1.6$ and $Na_v1.1$, with some TTX-R $Na_v1.8$ expression. Small cells, which are likely to be nociceptive in nature, predominantly express the TTX-S $Na_v1.7$ channel, in conjunction with TTX-R $Na_v1.8$ and $Na_v1.9$ channels. These findings have been further corroborated using intracellular recording, together with immunohistochemistry, to show the expected distribution of channels in particular fiber types, such as $Na_v1.9$, which is almost exclusively confined to nociceptive fibers (Figure 1).

Several studies have shed light on the distinct roles of the different sodium channel isoforms in DRG neurons. The TTX-S currents of large DRG neurons tend to have relatively fast recovery from inactivation (repriming) characteristics, which correlates with the predominant expression in these neurons of $Na_v1.6$, an isoform that displays this characteristic when expressed and studied in isolation in DRG. In contrast, small DRG neurons tend to have slow repriming, which corresponds to the expression of $Na_v1.7$. $Na_v1.6$ is a channel that has been linked in some cell types with production of a resurgent current, an inward current that occurs upon repolarization. In line with this is the observation that many large DRG neurons can produce resurgent current, whereas small neurons do not.

As mentioned previously, the action potential of many small DRG neurons is broad and the upstroke has been associated with the presence of the kinetically

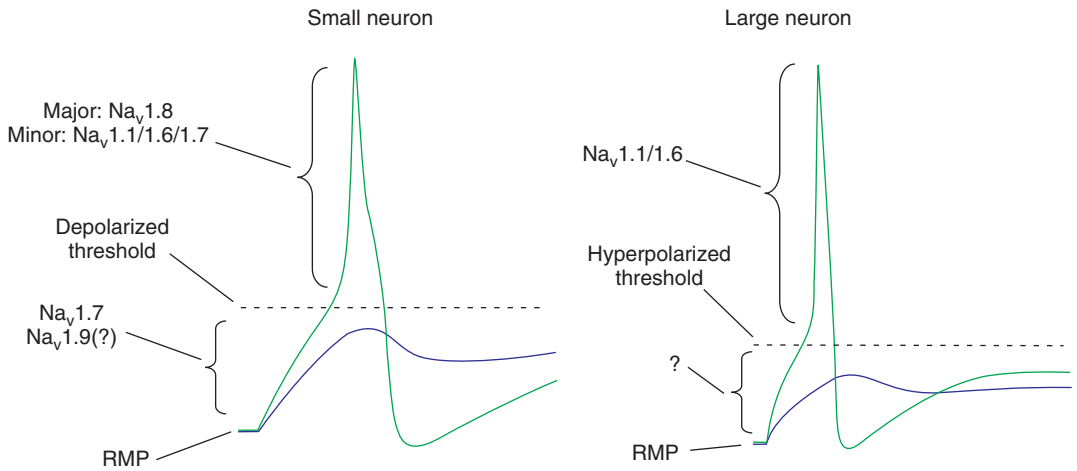

Figure 1 Putative roles of sodium channel subtypes in DRG action potentials. Subthreshold responses and overshooting action potentials are shown from small and large DRG neurons. In the small neuron, Na$_v$1.7 (and Na$_v$1.9 in some cells) is likely to bring the neuron toward threshold, whereas Na$_v$1.8 is the main channel responsible for the overshooting action potential. Note the large depolarization necessary for action potential activation, due to the depolarized characteristics of Na$_v$1.8. In contrast, larger DRG neurons are likely to have a more hyperpolarized threshold for action potential overshoot, in line with the predominance of TTX-S (Na$_v$1.1 and Na$_v$1.6) channels requiring less depolarization for activation. RMP, resting membrane potential.

slower TTX-R Na$_v$1.8 current. Thus, the majority of nociceptive fibers predominantly rely on current via Na$_v$1.8 for the upstroke of the action potential, rather than completely on TTX-S currents, as is the norm in the majority of excitable cells. Na$_v$1.8 carries a current which has very depolarized voltage dependence of activation and inactivation and is extremely resistant to depolarization. However, TTX-S currents do still play a role in the action potential, especially around threshold. One of the major candidates is Na$_v$1.7, which has been suggested to amplify subthreshold generator potentials. This is due to one specific property – the slow onset of inactivation. It is thought that under conditions in which there is slow depolarization, this particular channel would tend to remain available for activation, in comparison with other channels. This would make it a prime candidate to be able to produce ramp currents (responses to small, slow depolarizing stimuli), and this isoform has indeed been shown to produce a robust ramp current in response to slow depolarizations. Mutations in this Na$_v$1.7 have helped to demonstrate its role as a threshold channel that can interact with Na$_v$1.8 to produce repetitive firing behavior, as discussed in further detail later.

The role of Na$_v$1.9 in firing behavior of nociceptive DRG is unclear. Na$_v$1.9 produces a TTX-R, persistent current, which may be active around the resting membrane potential (RMP) due to hyperpolarized midpoints of steady-state inactivation and activation that lead to a large predicted window current. Modeling studies suggested that Na$_v$1.9 can depolarize the RMP of a neuron and may amplify depolarizing inputs. Subsequent *in vitro* experiments supported the predictions, in which the current was activated via G-proteins, and also showed that the presence of Na$_v$1.9 tends to lower threshold. In contrast, studies from knockout animals suggest that RMP may be unaffected or even hyperpolarized by a mechanism that is not yet understood, although the current may play a role in the sustained depolarization required for repetitive firing involving Na$_v$1.8 since this was eliminated in many cells of the knockout animals.

Excitability of Injured DRG Neurons

Chronic constriction injury (CCI) and axotomy of the sciatic nerve both cause a downregulation in Na$_v$1.8 and Na$_v$1.9 and an upregulation in Na$_v$1.3 expression in nociceptive DRG neurons. Na$_v$1.3 is normally only expressed during early stages of development and is practically undetectable in the adult. These changes are thought to be related to changes in access of axotomized DRG neurons to neurotrophic factors such as nerve growth factor and glial-derived neurotrophic factor because these effects can be reversed by increasing their levels.

The effect of replacement of Na$_v$1.8 and Na$_v$1.9 by Na$_v$1.3 in injured DRG neurons on firing behavior provides an excellent example of the effects that plasticity of ion channels can have on a neuron. As previously stated, the TTX-S current of small, nociceptive DRG normally recovers from inactivation slowly, and this current has largely been correlated to Na$_v$1.7 expression. The current produced when Na$_v$1.3 is expressed in a cell line or in DRG neurons reprimes rapidly and may boost subthreshold responses due

to a robust ramp current. Therefore, previously subthreshold inputs may generate action potentials in the presence of $Na_v1.3$ and, in addition, high frequency firing may be more sustained.

Inflammatory Pain

Additional factors may also be involved in inflammatory pain. Here, long-term chronic changes in expression may occur in addition to shorter term acute alterations in behavior of existing channels. Models of inflammatory pain, such as carrageenan injection into the hindpaw, produce an upregulation of TTX-S channels $Na_v1.3$ and $Na_v1.7$ in DRG after a few days, together with a parallel increase in the TTX-S current. This is in addition to an increased expression of $Na_v1.8$ and a boosted TTX-R current. In contrast, $Na_v1.6$ and $Na_v1.9$ are unaffected in the inflammatory response. These changes would amplify previously subthreshold inputs ($Na_v1.3$ and $Na_v1.7$) and help in sustaining repetitive firing under prolonged depolarization ($Na_v1.3$ and $Na_v1.8$). $Na_v1.8$ is also likely to be involved in acute inflammatory pain because administration of agents such as prostaglandin E_2 (PGE_2) and 5-HT produces an increase in current amplitude and a hyperpolarizing shift in activation, both likely to increase excitability. Evidence also points toward the involvement of $Na_v1.9$ in the inflammatory response. Acute administration of the inflammatory mediator PGE_2 causes a large increase in the steady-state current amplitude of $Na_v1.9$, a channel which, when knocked out in the mouse, produces deficits in the inflammatory response. These data suggest that $Na_v1.9$ may play a significant role in the increased excitability of nociceptive fibers during inflammation, either by RMP effects or by sustaining depolarization. Taken together, the previous findings indicate that changes in multiple ion channels due to either levels of expression or modulation of the existing isoforms can have profound effects on the firing behavior of DRG and suggest that particular isoforms may be more important targets for treatment at different stages of a pathology.

Demyelination and Multiple Sclerosis

Misexpression of ion channels may also be of importance in demyelinated peripheral axons, such as that which occurs during multiple sclerosis, leading to slowing of conduction and conduction block. $Na_v1.6$ is normally concentrated at nodes of Ranvier along myelinated axons, but subsequent to demyelination, the channel can be found expressed along the long lengths of some axons, including the centrally projecting axons of DRG neurons within the dorsal columns. With normal development, $Na_v1.2$ is often the initial sodium channel at immature nodes of Ranvier, which is then replaced by $Na_v1.6$ at later stages. $Na_v1.2$, which is normally expressed along the lengths of premyelinated axons, also reemerges along some demyelinated axons, including the axons of DRG neurons within the dorsal columns. When the properties of these channels have been compared by expression studies in DRG neurons, $Na_v1.6$ has been found to have a greater persistent current. This current may provide a sustained sodium conductance that can drive the reverse Na^+/Ca^{2+} exchanger which contributes to axonal degeneration. Consistent with this, the Na^+/Ca^{2+} exchanger has been found to be co-localized with $Na_v1.6$ in injured axons, whereas it was co-localized with $Na_v1.2$ in axons that did not display signs of injury. In addition, $Na_v1.2$ was found to have more depolarized voltage dependence of activation and inactivation than $Na_v1.6$ in DRG. $Na_v1.2$ may therefore allow a damaged and depolarized neuron to continue to fire action potentials, thus sustaining function. These observations suggest that specific expression patterns of these two isoforms can have profound effects on the survival or degeneration of particular axons and also on maintenance of the signal-carrying function following demyelination.

Ion Channel Mutations Associated with DRG

As mentioned previously, an important characteristic of $Na_v1.7$ is its behavior in response to slow ramp depolarizations which mimic natural small depolarizations such as generator potentials. $Na_v1.7$ displays a larger ramp current than other sodium channel isoforms so that it may function as a threshold channel.

Inherited erythromelalgia (also known as erythermalgia), a disorder in which patients suffer burning pain in their extremities, has been linked to mutations in $Na_v1.7$. These mutations tend to cause a negative shift in the voltage dependence of activation, often accompanied by an increased ramp current. When the effects of these mutations on firing behavior of DRG were examined, there was a lowering of threshold for single action potentials and high-frequency firing. These mutations in $Na_v1.7$ may also cause a depolarization of a neuron in which they are expressed, possibly due to the larger persistent current around the RMP. This depolarization could inactivate many TTX-S channels with relatively hyperpolarized steady-state characteristics, a factor that would tend to decrease excitability. However, $Na_v1.8$ is responsible for the overshooting action potentials in the majority of small DRG neurons and has very depolarized activation and

inactivation properties compared to other isoforms. This suggests that even if some of the erythromelalgia mutations caused a depolarization of DRG neurons, the essential $Na_v1.8$ channel would still be available for action potential firing and, in fact, the neuron would be closer to the activation of those channels, which would decrease the threshold for firing. Conversely, this implies that neurons which do not express $Na_v1.8$ would display hypoexcitability in response to mutant $Na_v1.7$ expression rather than hyperexcitability.

This hypothesis has been tested directly in a study in which the effects on firing behavior of expression of an erythromelalgia mutated $Na_v1.7$ channel were compared to those of wild type in DRG neurons and in superior cervical ganglion (SCG) sympathetic neurons (the other type of neuron in which $Na_v1.7$ is expressed), which do not express $Na_v1.8$. The mutation caused an $\sim$5 mV depolarization in resting membrane potential in both types of neurons. The threshold for firing of single action potentials and high-frequency firing was decreased in DRG neurons (hyperexcitability). However, the opposite occurred in SCG neurons, in which an overall reduction in excitability was seen, presumably due to inactivation of sodium channels essential for the action potential upstroke. Subsequent addition of $Na_v1.8$ into SCG neurons allowed these cells to fire action potentials despite the depolarization. This suggests that channelopathies in general may have very different effects on firing properties, dependent on the other channel isoforms present where the mutations are expressed.

Modulation of Excitability by Cofactors and Phosphorylation

Thus far, we have mainly discussed alterations in firing behavior caused by differential expression of specific sodium channel isoforms in a specific cell type, DRG neurons. Studies have revealed that the cellular background in which particular channels are expressed is an important determinant of behavior of the isoform. These differences are likely caused by selective expression of particular associated proteins.

One such factor is contactin, a cell adhesion molecule that has been shown to associate with $Na_v1.2$, $Na_v1.3$, and $Na_v1.9$, probably by the intracellular C- and N-termini, via a linker protein such as β_1 so that it enhances the channel density at the plasma membrane. The interaction with $Na_v1.3$ may play a role in accumulation of the channel within the neuroma and DRG of a transected sciatic nerve and therefore the emergence of a rapidly repriming sodium current. In a nonpathological setting, recordings from DRG of contactin knockout animals suggest that contactin does not affect the channel density of the TTX-S currents present (likely $Na_v1.6$ and $Na_v1.7$) but does increase the amplitude of both the slow and persistent TTX-R currents (produced by $Na_v1.8$ and $Na_v1.9$) in a subset of neurons. These data demonstrate that contactin is likely to influence the trafficking and expression of particular sodium channel isoforms to selected compartments and thus influence their firing properties.

Other proteins have been shown to not only boost the channel density at the plasma membrane but also to change the properties of the channel. This has been shown, for example, for fibroblast growth factor homologous factor 2A (FHF2A) and -2B (FHF2B). These proteins associate with $Na_v1.6$ via the C-terminus of the channel, and both cause an increase in current density in cells in which they are co-expressed. FHF2B causes a moderate and FHF2A a large positive shift in the voltage dependence of inactivation, but the largest difference is demonstrated with recovery from inactivation. FHF2A drastically inhibits repriming of $Na_v1.6$, leading to a large accumulation of inactivation during repetitive stimulation. In contrast, FHF2B allows a quicker recovery of channels. These effects would likely influence the response of a neuron to sustained depolarization, where FHF2A would be expected to inhibit repetitive firing. Interestingly, FHF2B is selectively expressed at nodes of dorsal root sensory but not ventral root motor axons, whereas FHF2A is localized in unmyelinated fibers which originate from small DRG neurons. The differential expression of FHF2B may contribute to differences in the electrogenic properties of motor versus sensory fibers. The presence of FHF2A in small DRG neurons may help explain the anomalous observation that TTX-S current is slowly repriming in small DRG, despite $Na_v1.6$ as well as $Na_v1.7$ being present in these cells. A protein such as FHF2A may be responsible for slowing the recovery from inactivation of $Na_v1.6$ current in these cells.

Sodium channels of DRG can also be modulated by phosphorylation. The current density of $Na_v1.6$ is reduced upon activation of p38 mitogen-activated protein (MAP) kinase, which has been shown to be upregulated in certain pain conditions. Phosphorylation is via a single site on the intracellular linker connecting domains I and II of the channel, but the mechanism of the change in current density is not known. One study found that $Na_v1.8$ may also be modulated by p38 but in the opposite direction. Tumor necrosis factor-α (TNF-α) is a proinflammatory cytokine that has been linked to enhanced sensitization during inflammation. Mechanical sensitization due to TNF-α was found to likely occur via an increased TTX-R ($Na_v1.8$) current in a p38 MAP kinase-dependent manner. As previously mentioned, $Na_v1.8$

current can also be increased via the protein kinase A cascade (whereas $Na_v1.2$ is decreased); thus, TTX-S and TTX-R channels may display differential modulation by phosphorylation.

Conclusions

Excitability of DRG can be affected by changes during development and by a variety of pathologies and injuries. However, the control of firing behavior has increasingly been linked to alterations in expression of specific isoforms with different biophysical characteristics or modulation of existing ion channels. Sodium channels, which were once thought to be uniform in terms of biophysical properties and responsible for a single characteristic (the rising phase) of an action potential, have been shown to be able to play multiple roles in electrogenesis and thus can finely tune the behavior of excitable cells. Not unexpectedly, alterations in expression or physiological properties of sodium channels thus contribute to altered function of neurons such as DRG neurons following injury.

See also: Demyelinating Diseases; Ion Channel Localization in Axons; Ion Channel Localization in Cell Bodies and Dendrites; Neuromodulation of Sodium Channels; Sodium Channels.

Further Reading

Baker MD and Wood JN (2001) Involvement of Na^+ channels in pain pathways. *Trends in Pharmacological Sciences* 22: 27–31.

Cantrell AR and Catterall WA (2001) Neuromodulation of Na^+ channels: An unexpected form of cellular plasticity. *Nature Reviews Neuroscience* 2: 397–407.

Catterall WA, Goldin AL, and Waxman SG (2005) International Union of Pharmacology: XLVII. Nomenclature and structure–function relationships of voltage-gated sodium channels. *Pharmacological Reviews* 57: 397–409.

Dib-Hajj SD, Rush AM, Cummins TR, and Waxman SG (2006) Mutations in the sodium channel $Na_v1.7$ underlie inherited erythromelalgia. *Drug Discovery Today: Disease Mechanisms* 3: 343–350.

Gold MS (1999) Tetrodotoxin-resistant Na^+ currents and inflammatory hyperalgesia. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7645–7649.

Isom LL (2000) Pathobiology of visceral pain: Molecular mechanisms and therapeutic implications: I. Cellular and molecular biology of sodium channel beta-subunits: Therapeutic implications for pain? *American Journal of Physiology – Gastrointestinal and Liver Physiology* 278: G349–G353.

Kostyuk PG, Veselovsky NS, and Tsyndrenko AY (1981) Ionic currents in the somatic membrane of rat dorsal root ganglion neurons – I. Sodium currents. *Neuroscience* 61: 2423–2430.

Lai J, Porreca F, Hunter JC, and Gold MS (2004) Voltage-gated sodium channels and hyperalgesia. *Annual Review of Pharmacology and Toxicology* 44: 371–397.

Rush AM, Brau ME, Elliott AA, and Elliott JR (1998) Electrophysiological properties of sodium current subtypes in small cells from adult rat dorsal root ganglia. *Journal of Physiology* 511: 771–789.

Rush AM, Cummins TR, and Waxman SG (2007) Multiple sodium channels and their roles in electrogenesis within dorsal root ganglion neurons. *Journal of Physiology* 579(1): 1–14.

Rush AM, Dib-Hajj SD, Liu S, Cummins TR, Black JA, and Waxman SG (2006) A single sodium channel mutation produces hyper- or hypoexcitability in different types of neurons. *Proceedings of the National Academy of Sciences of the United States of America* 103: 8245–8250.

Waxman SG (2001) Transcriptional channelopathies: An emerging class of disorders. *Nature Reviews Neuroscience* 2: 652–659.

Waxman SG, Cummins TR, Black JA, and Dib-Hajj SD (2002) Diverse functions and dynamic expression of neuronal sodium channels. *Novartis Foundation Symposium* 241: 34–51.

Waxman SG and Dib-Hajj SD (2005) Erythermalgia: Molecular basis for an inherited pain syndrome. *Trends in Molecular Medicine* 11: 555–562.

Waxman SG, Dib-Hajj SD, Cummins TR, and Black JA (1999) Sodium channels and pain. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7635–7639.

Wood JN (2007) Ion channels in analgesia research. *Handbook of Experimental Pharmacology* 177: 329–358.

Relevant Website

http://www.ionchannels.org – Ion Channel Media Group.

Double Cortex

P Gressens and S Passemard, Inserm, U676, Paris, France
G Sebag and F Chalard, Radiology, Hôpital Robert Debre, Paris, France
A Laquerriere, Pathology, CHU Rouen, Rouen, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Diffuse subcortical band heterotopias (SBHs) are neuronal migration disorders that have been described pathologically as well as on clinical and imaging basis under the names of double cortex, bicortical lissencephalies, partial lissencephalies, subcortical laminar heterotopia, or laminar band heterotopias (OMIM #300067). In this brain malformation, two cortical plates, separated by white matter, coexist: one cortical plate is in a normal superficial position while the other one is located in a deep heterotopic position and is separated from the lateral ventricle by another area of white matter. This malformation is generally isolated without any other malformation although in a few occasions it can be associated with pachygyria. A large spectrum of clinical manifestations can be observed in this condition: reported patients have usually moderate to severe developmental delay and an early onset of medically refractory seizures but other patients can have relatively mild clinical manifestations. Mutations in DCX gene (or XLIS gene; OMIM #300121) are responsible for an X-linked neuronal migration disorder where females display SBH (or double cortex syndrome) (Figure 1) and males have lissencephaly (Figure 2). This gene encodes a protein, doublecortin, which is a microtubule-associated protein (Map) expressed by migrating and differentiating neurons. The majority of cases of SBH have been found to be linked to abnormalities in DCX. Some genotype–phenotype correlations for SBH patients are being proposed.

Genetic Bases

Recent studies suggest that almost all familial cases of SBH are linked to mutations of DCX, while up to 80–85% of sporadic female cases and about 25% of sporadic male cases of SBH are linked to mutations in the same gene. DCX mutations are generally associated with anterior greater than posterior migration abnormalities. DCX is located on the long arm of the X-chromosome (Xq22.3–q23) and is inherited in an X-linked dominant manner. In general, affected hemizygous males have a more severe phenotype (lissencephaly) than affected heterozygous females (SBH). This observation, which is classical for X-linked dominant diseases, could suggest that in female patients, due to random X-inactivation, neocortical neurons expressing the mutated DCX will fail to migrate properly while neurons expressing the nonmutated DCX will migrate to the superficial cortical plate. In contrast, in male patients, due their obligate homozygous status for DCX, all neocortical neurons will fail to migrate properly.

In addition, DCX mutations have been described in male patients with SBH and in female patients with anterior greater than posterior lissencephaly. Male patients with SBH and associated DCX mutations were recently shown to exhibit somatic mosaicism of these DCX mutations in hair root cells. Although brain mosaicism for DCX mutation remains to be confirmed in these patients, one can make the hypothesis that this somatic mosaicism underlines the less-severe migration disorder in these male patients.

For what concerns female patients with lissencephaly associated with DCX mutations, one potential explanation would be a skewed inactivation of DCX in migrating neurons. Analysis of lymphocytes has failed to support such a skewed inactivation of DCX but analysis of brain tissue would allow testing this hypothesis directly.

Mutations of LIS1 gene are associated with isolated type 1 lissencephaly or with Miller–Diecker syndrome. Lissencephaly observed in male patients with DCX mutations has an anterior greater than posterior distribution while type 1 lissencephaly due to mutations of LIS1 gene has a posterior greater than anterior distribution. Of note, some male patients with posterior greater than anterior SBH have been shown to have mosaic mutations of LIS1.

Reported mutations in DCX include deletions, nonsense mutations, and missense mutations, leading to truncated protein or amino acid changes. Most of these reported mutations cluster in two hot spot regions of DCX called R1 and R2. These R regions code internal tandem repeats which are both required for assembly of Dcx with microtubules.

Imaging Studies

As mentioned above, migration disorders associated with mutations in DCX have an anterior-predominant distribution while migration disorders associated with mutations in LIS1 have a posterior-predominant distribution.

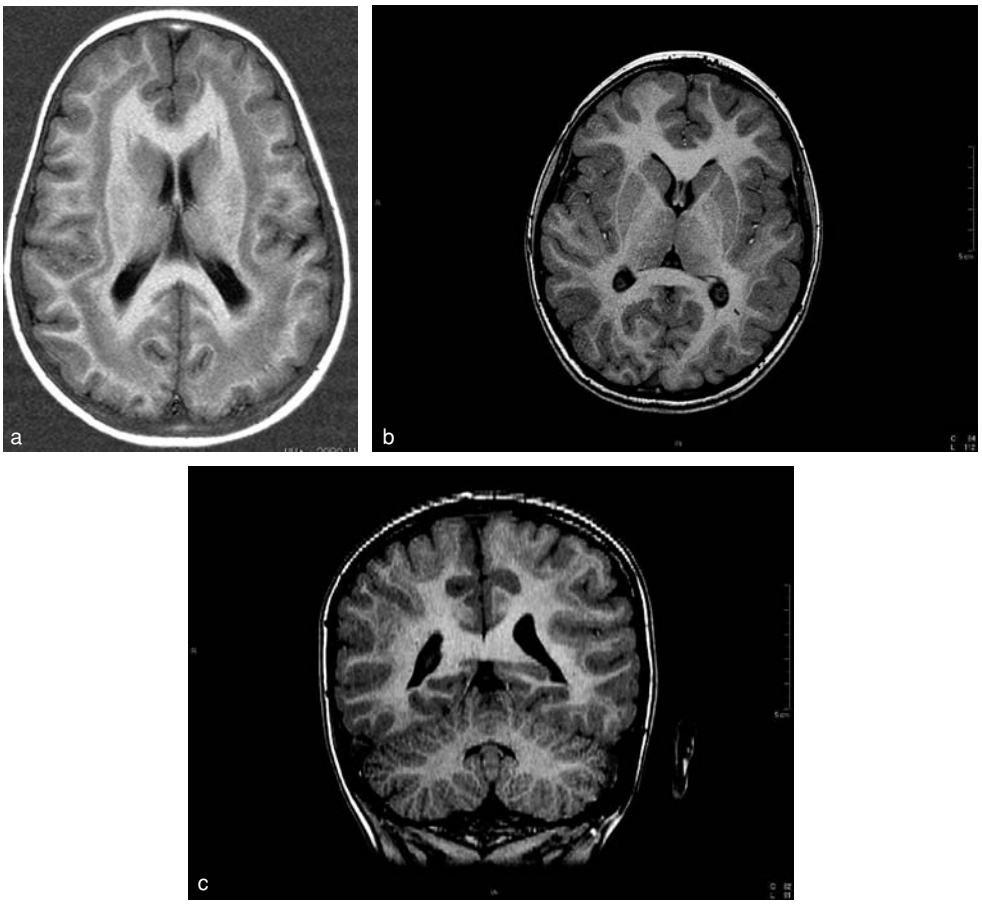

Figure 1 Diffuse subcortical band heterotopia imaged by MRI in three separate female patients with mutations in DCX gene. (a) Courtesy of Dr. Thierry Duprez, Service de Radiologie, St-Luc Hospital, University of Louvain Medical School at Brussels, Belgium.

Figure 2 Lissencephaly imaged by MRI in a male patient with mutation in DCX gene.

SBH associated with DCX mutations can be divided into three different classical groups according to severity: (1) generalized thick SBH with an anterior greater than posterior gradient; (2) generalized thin SBH with an anterior greater than posterior gradient; and (3) thin SBH restricted to frontal lobes. Interestingly, patients with generalized (thin or thick) SBH tend to have more severe mutations in DCX than patients with SBH restricted to frontal lobes. A recent imaging comparison of male and female patients with SBH revealed that some female patients displayed more frequently classical anatomical types such as diffuse SBH or SBH with anterior predominance while some more unusual anatomical subtypes such as posterior SBH associated with anterior pachygyria or diffuse SBH with posterior predominance were more frequently or exclusively observed in male patients.

Functional magnetic resonance imaging (fMRI) has been performed in a few patients with SBH, yielding interesting results. Using simple motor tasks and/or sensory stimulus, different groups were able to show that motor and sensory tasks induce activation of the

SBH. In addition, when compared to controls, these tasks induce the recruitment of bilateral primary areas and more numerous higher-order association normotopic cortices, suggesting the need for a more widespread network to perform simple tasks in patients with SBH. Positron emission tomography (PET) scan with [18F]-fluorodeoxyglucose performed in a patient with SBH revealed that band heterotopia had the same degree of glucose metabolism as that of the overlying cortex.

Neuropathology

At the macroscopic level, SBH appears in females as nodular or confluent masses of gray matter forming a contiguous subcortical band. This laminar band is clearly separated from the overlying cortex (which looks normal except in cases associating SBH and pachygyria) and the lateral ventricle by white matter tracts. The extent of this band can be variable with bilateral or unilateral laminar heterotopia and diffuse or local laminar heterotopia. Also, the thickness of these laminar heterotopia can be thick or thin. As mentioned above, in DCX mutations, the distribution of this laminar heterotopia shows an anterior versus posterior predominance.

At the microscopic level, SBH does not exhibit any typical histological feature except for an increased number of neurons in the superficial layer, and a smaller appearance of pyramidal neurons in some cases. Between both layers, a large zone of hypomyelinated brain parenchyma containing neurons can be observed. Neuronal nodules appear surrounded by normally myelinated fiber tracts.

Most of male patients with DCX mutations display lissencephaly rather than SBH. Until recently, it was considered that the DCX-linked migration lissencephaly was similar to the lissencephaly associated with mutations in LIS1 and observed in isolated type 1 lissencephaly or Miller–Diecker syndrome. However, imaging studies comparing patients with DCX and LIS mutations have revealed different distribution patterns of lissencephalies (anterior vs. posterior predominance). Similarly, neuropathological studies comparing patients with DCX and LIS mutations have revealed different distribution patterns of agyria–pachygyria (anterior vs. posterior predominance). Sylvian fissures are abnormal, due to hypoplasia of the frontal opercula, and sometimes of the frontal lobes. The lack of primary and secondary sulci predominates in the frontoparietal areas. On coronal sections, the corpus callosum is short and thick, but the basal ganglia appear as normal. The cortical ribbon is poorly demarcated from the white matter (Figure 3).

Neuropathological studies have shown that, although brains with LIS1 deletions exhibited the classic inverted four-layer lissencephalic architecture, male patients with DCX mutations and lissencephaly had unique cytoarchitectural findings distinct from LIS1 including a roughly ordered six-layer lamination (Figure 3). The molecular layer is paucicellular, the layer II (external granular cell layer) is absent, and replaced by a highly cellular and thick band of medium-sized and large, often inverted, pyramidal neurons devoid of laminar organization. Under a paucicellular layer III, layer IV is also thick and composed of small-sized neurons. In the underlying white matter, multiple clusters of heterotopic neurons arranged in a columnar fashion are observed (layer V), as well as nodules (layer VI) located in the vicinity of the ventricular wall. The striatum is disrupted by the columns. Brain stem structures are generally normal, although decreased basis pontis neuronal populations and olivary pachygyria have been described. The deep cerebellar nuclei are dysplastic and the cerebellar white matter contains heterotopic nests of Purkinje cells.

Clinical Aspects

A large spectrum of clinical manifestations can be observed in this condition: reported patients have usually moderate to severe developmental delay and an early onset of medically refractory seizures but other patients can have relatively mild clinical manifestations. Epileptic manifestations, if present, are often partial complex seizures or atypical absences with focal or regional onset.

A study comparing male and female patients with SBH showed that the range of clinical phenotypes in male patients greatly overlapped that in female patients. However, female patients tend to have mild to moderate clinical symptoms while male patients tend to have less severe or more severe clinical symptoms than females. This gender-related distribution of clinical signs is globally correlated with the gender-related distribution and severity of anatomical abnormalities.

Genotype–phenotype correlations for SBH are emerging but remain weaker than those observed for lissencephaly. More severe mutations tend to be observed in sporadic cases when compared to familial cases.

Genetic counseling remains a difficult area in SBH. In SBH cases with DCX mutations, a variable penetrance (potentially due to random X-inactivation-induced somatic mosaicism, modifying genes and/or environmental factors) has been observed. Furthermore, silent carriers of DCX mutations have been reported. In the case of silent carrier, it is suggested

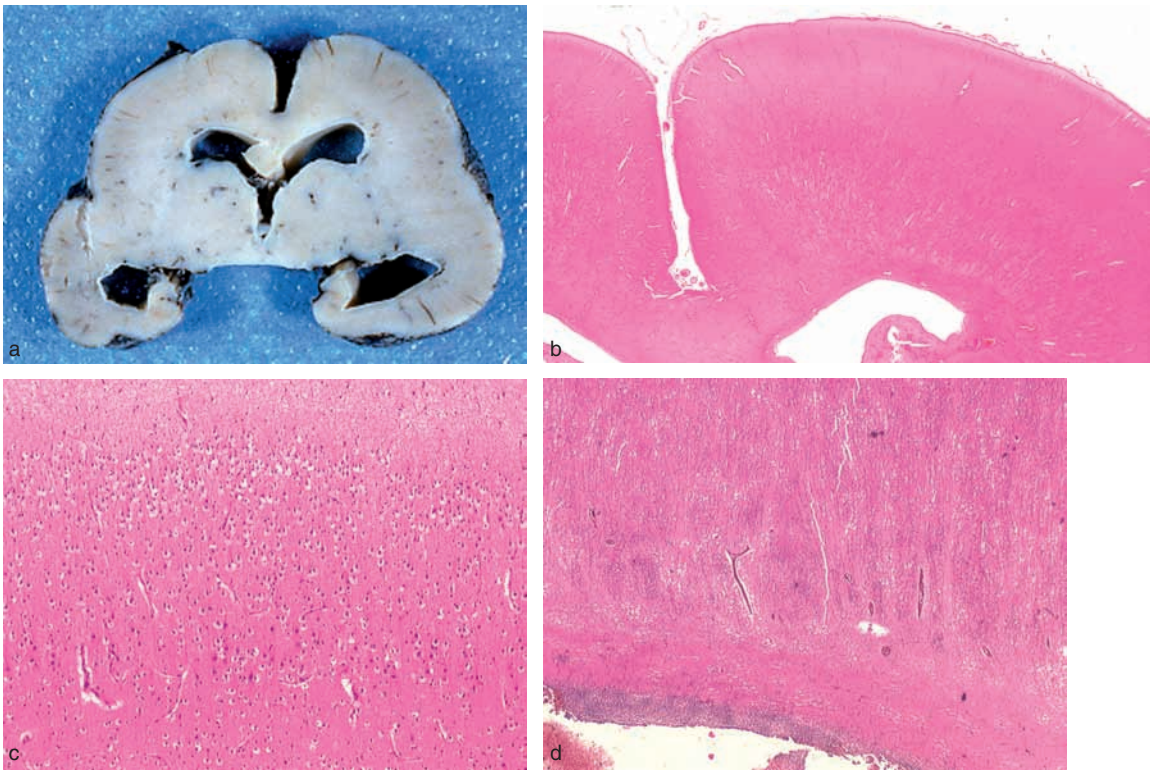

Figure 3 Lissencephaly in a male patient with mutation in DCX gene. (a) Macroscopic coronal section of the telencephalon passing through hippocampic gyri, showing the poor delineation between gray and white matter, with an inverted ratio gray/white matter. (b) Representative microphotograph of the frontal lobe: the cortical ribbon is extremely thick as well as the corpus callosum. (c) Higher magnification showing the superficial cortical layers. The external granular cell layer is replaced by more or less differentiated pyramidal neurons. (d) Higher magnification showing the deep cortical layers with periventricular heterotopic neuronal nodules (right) and neuronal columnar heterotopias (left).

that there is a 10% risk to have a second child affected after the diagnosis of the first, the real risk being linked to the level of mosaicism and whether germ cells are involved by the mutation. Finally, in male patients, DCX mutations are only responsible of a fraction of cases of SBH, suggesting the existence of other genes involved with potentially different transmission and penetrance than DCX mutations.

Pathophysiology

Studies over the last decade have identified several molecules involved in the control of neuronal migration and targeting of neurons to specific brain regions. These molecules can be divided into four broad categories (Figure 4): (1) molecules of the cytoskeleton which play an important role in the initiation (filamin-A and Arfgef2) and ongoing progression (Dcx, Lis1, 14–3-3epsilon, Map1B, Map2, and Tau) of migration; (2) signaling molecules playing a role in lamination (Reelin, Disabled-1 (Dab1), ApoE receptor 2, and very low density lipoprotein receptor (Vldlr); p35, cdk5, and Brn1/Brn2); (3) molecules modulating glycosylation (protein *O*-mannosyltransferase, protein *O*-mannose β-1, 2-*N*-acetylglucosaminyltransferase, and Fukutin) which seem to provide stop signals for migrating neurons; and (4) other factors which have been shown to modulate neuronal migration (glutamate, γ-aminobutyric acid (GABA), brain-derived neurotrophic factor (BDNF), thyroid hormones, molecules deriving from peroxisomal metabolism, and environmental factors such as ethanol and cocaine). Recent studies have highlighted significant cross talk between the groups of molecules.

Neuronal migration involves two distinct phases: extension of the leading process and movement of the cell body and nucleus (nucleokinesis). Movement of the centrosome (microtubule organizing center) in the leading process, in front of the nucleus, precedes movement of the nucleus toward the centrosome. Nucleokinesis requires, among others, the dynein/Lis1/Ndel1 complex, Map's including Dcx, signaling molecules (cdk5, which phosphorylates Dcx and Fak), and stop molecules (Fak). The positioning of the centrosome in the leading process and the dynamic coupling and uncoupling of the nucleus to the microtubule network and centrosome seem to be key events of neuronal migration.

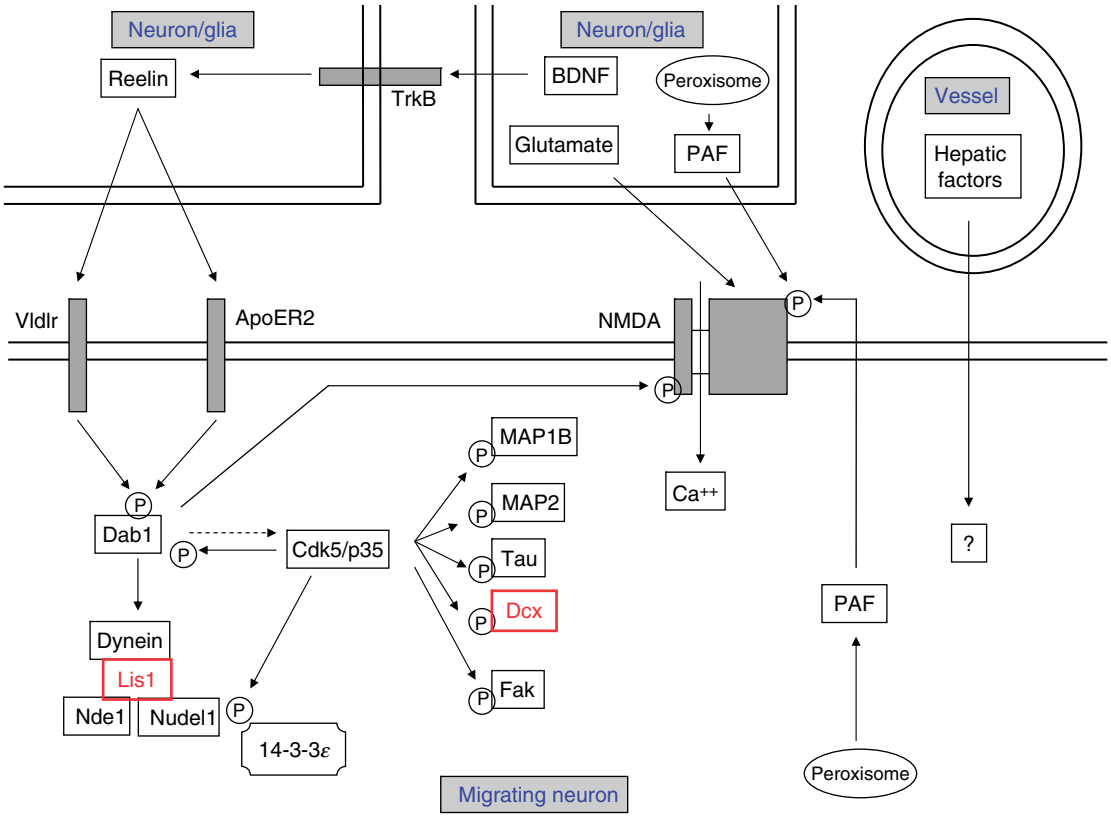

Figure 4 Schematic representation of some critical molecular modulators of neuronal migration illustrating the cross talk between different groups of molecules including cytoskeletal proteins, signaling molecules of the Reelin pathway, NMDA receptor-mediated pathway, and peroxisome-derived factors. This scheme highlights the central position of Dcx and Lis1 in the cytoskeletal machinery controlling migration.

Several recent studies have focused on the mechanisms of action of Dcx. Although mutations in DCX are responsible for SBH, the precise role of Dcx in the control of neuronal migration is not yet totally understood. Dcx could act as a molecular link between microtubule and actin cytoskeletal filaments. As mentioned above, two R (R1 and R2) regions coding internal tandem repeats are required for assembly of Dcx with microtubules. Dcx microtubule affinity is regulated by a balance of kinase and phosphatase activity. In addition, two other reports suggest roles of Dcx which are not directly related to microtubules. Dcx interacts with c-Jun N-terminal kinase (Jnk) and Jnk interacting protein (Jip), suggesting the involvement of Dcx in a signaling pathway. Dcx was also shown to interact with the ubiquitin-specific protease *Drosophila* fat facets related on X-chromosome (Dffrx), which prevents degradation of different molecules including β-catenin by the proteasome, suggesting a potential role of Dcx in the regulation of cell adhesion.

Further confirming the role of Dcx in neuronal migration, *in utero* electroporation of plasmids encoding short hairpins RNA to create interference (RNAi) of Dcx protein in embryonic rats induces two types of neocortical neuronal migration defects: many neurons stop their migration in the intermediate zone (prospective white matter) and form an SBH; in addition, many neurons migrate into inappropriate cortical layers within the normotopic cortical plate. Interestingly, SBH contains transfected pyramidal neurons as well as nontransfected cell types, suggesting that affected cells also influence the migration of neighboring neurons expressing normal levels of Dcx. Cells in the SBH display a multipolar stage, which occurs as bipolar progenitor cells become radially migrating neurons. This multipolar stage seems to be a critical checkpoint of migration and therefore a vulnerable stage for disruption of neuronal migration.

Two other animal models of SBH are currently available: the ibotenate-induced laminar heterotopia in the newborn hamster and the tish (telencephalic internal structural heterotopia) rat. Focal injections of ibotenate (a glutamate analog mainly acting on *N*-methyl-D-aspartate (NMDA) receptors) into the brain of hamsters on the day of birth (a stage where neurons destined to the superficial cortical layers are migrating) induces different types of migration abnormalities, including SBH, focal periventricular

nodular heterotopia, or intracortical arrests of migrating neurons. The tish rat is a spontaneously mutant animal exhibiting a bilateral laminar band heterotopia which predominates in the frontal and parietal cortex. The laminar heterotopia seems to be secondary to the combination of abnormal neuronal migration and heterotopic neurogenesis. Affected rats exhibit spontaneous recurrent electrographic and behavioral seizures.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Evoked Potentials: Recording Methods; Lissencephaly; Lissencephaly Type I and Periventricular Heterotopia; Microtubules: Organization and Function in Neurons.

Further Reading

Bai J, Ramos RL, Ackman JB, et al. (2003) RNAi reveals doublecortin is required for radial migration in rat neocortex. *Nature Neuroscience* 6: 1277–1283.

D'Agostino MD, Bernasconi A, Das S, et al. (2002) Subcortical band heterotopia (SBH) in males: Clinical, imaging and genetic findings in comparison with females. *Brain* 125: 2507–2522.

des Portes V, Francis F, Pinard JM, et al. (1998) Doublecortin is the major gene causing X-linked subcortical laminar heterotopia (SCLH). *Human Molecular Genetics* 7: 1063–1070.

Gleeson JG (2000) Classical lissencephaly and double cortex (subcortical band heterotopia): LIS1 and doublecortin. *Current Opinion in Neurology* 13: 121–125.

Gleeson JG, Allen KM, Fox JW, et al. (1998) Doublecortin, a brain-specific gene mutated in human X-linked lissencephaly and double cortex syndrome, encodes a putative signaling protein. *Cell* 92: 63–72.

Gleeson JG, Luo RF, Grant PE, et al. (2000) Genetic and neuroradiological heterogeneity of double cortex syndrome. *Annals of Neurology* 47: 265–269.

Gressens P (2006) Pathogenesis of migration disorders. *Current Opinion in Neurology* 19: 135–140.

Kato M, Kanai M, Soma O, et al. (2001) Mutation of the doublecortin gene in male patients with double cortex syndrome: Somatic mosaicism detected by hair root analysis. *Annals of Neurology* 50: 547–551.

Leventer RJ (2005) Genotype–phenotype correlation in lissencephaly and subcortical band heterotopia: The key questions answered. *Journal of Child Neurology* 20: 307–312.

Loturco JJ and Bai J (2006) The multipolar stage and disruptions in neuronal migration. *Trends in Neurosciences* 29: 407–413.

Matsumoto N, Leventer RJ, Kuc JA, et al. (2001) Mutation analysis of the DCX gene and genotype/phenotype correlation in subcortical band heterotopia. *European Journal of Human Genetics* 9: 5–12.

Monuki ES and Ligon KL (2004) Cerebral heterotopia. In: Golden JA and Harding BN (eds.) *Pathology and Genetics: Developmental Neuropathology*, pp. 52–60. Basel: ISN Neuropath Press.

Viot G, Sonigo P, Simon I, et al. (2004) Neocortical neuronal arrangement in LIS1 and DCX lissencephaly may be different. *American Journal of Medical Genetics A* 126: 123–128.

Down Syndrome

A Salehi and W C Mobley, Stanford University, Stanford, CA, USA
B Pohlman, University of California at San Francisco, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Down syndrome (DS) is caused by trisomy (i.e., the presence within the genome of three copies) of chromosome 21 (HSA21). Infants with DS constitute about 1 in every 700 live births in the United States. Not surprisingly, the presence of an extra copy of nearly 300 genes results in many abnormalities. DS affects many organ systems: cardiovascular, digestive, endocrine, immune, reproductive, musculoskeletal, and nervous systems (Figure 1).

DS is a lifelong condition, often resulting in multiple medical interventions and ongoing management of chronic disease and disability. Many individuals with DS need special education and residential, medical, and social services for the duration of their lives. They have often been stigmatized and marginalized in society. Even during the twentieth century, they were subject to harsh institutionalization practices and medical neglect. Deinstitutionalization and the development of community care models, coupled with medical and scientific advances, have dramatically increased the quality of medical care, education, and the life expectancy for people with DS.

In addition to its consequences for the health and well-being of affected individuals, DS exacts a significant economic cost to society through the provision of medical and social services. The total yearly cost (i.e., both medical and nonmedical) is estimated to be $54–123 billion for the USA (350 000 cases at $156–353 000 per case). Furthermore, the increase in life expectancy for people with DS has been quite remarkable. The median age at death for people with DS has increased from 2 years in the 1950s to 50 years in the 1990s. While much has been accomplished in improving not only the longevity but also the quality of life for people with DS, increased life expectancy has given rise to medical and social concerns related to the aging of this population.

In the sections that follow, we first trace some of the historical, epidemiological, and etiological aspects of DS. We then discuss its clinical manifestations. Focusing on the nervous system, we point to recent research advances in pathophysiology and outline some prospects for future research and therapeutic interventions.

History

Although DS was recognized and represented in the arts as long ago as 1500 BC, its conception as a distinct and natural entity was formed within the emergence of specialty asylums in England and Wales in the mid-nineteenth century. These types of institutions were built in a time of optimism with regard to rehabilitation and training of those deemed cognitively deficient.

J. Langdon Down of the Earlswood Asylum (UK), in the 1860s, is credited for having identified the syndrome that later carried his name. He formalized his concept of 'Mongolian imbecility' in 1866. He characterized the population according to his perception of how they looked in ethnic and racial terms, identifying those who are now regarded as afflicted with DS not only as looking remarkably alike one another, but also resembling Mongols. Drawing from the social evolutionary theory of his day, Down theorized that these people were evidence of regression to a lower racial state in a hierarchy of races, a spontaneous reversion of individuals to earlier 'premoderns.'

Down's racial argument held sway for several decades until successfully challenged by Lionel Penrose. Penrose argued that there was no evidence of increased 'Mongol' blood in persons with DS, using statistical analysis and extensive blood investigation. He was perhaps the first to reject the name 'Mongolism' in the 1930s; however, it was not until 1966 that the World Health Organization officially replaced the name with 'Down syndrome' (note: per a 1974 *Lancet* article, the singular, rather than possessive, form of 'Down' is preferred).

The latter half of the 1950s was a watershed for genetic inquiry and for the conceptualization of DS. In 1959, Jérôme Lejeune of France published his observations that there was an extra chromosome present in the smooth muscle tissue of people with DS. Immediately following, Patricia Jacobs also linked the disorder to an extra copy of chromosome 21. This model of the DS chromosomal anomaly was used as a foundation in identifying other autosomal disorders and was a landmark discovery for medical genetics. The research on DS reached its climax in 2000 when the complete sequence of human chromosome 21 was published by Hattori and colleagues.

Over the past 40 years, parental and professional advocacy have dramatically affected the lives of persons with DS, impacting institutional practices, education, work opportunities, and clinical care. Advances in pediatric medicine have been especially

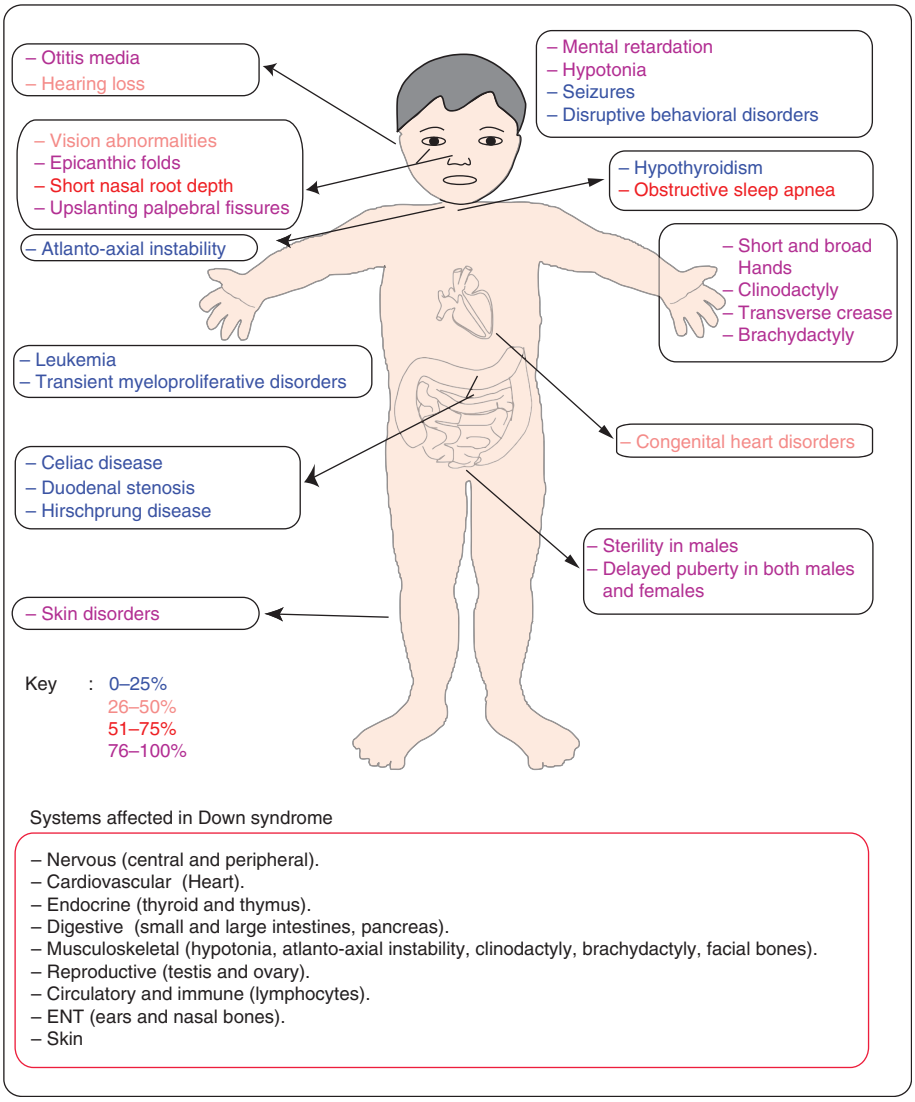

Figure 1 Schematic representation of systems affected in Down syndrome. Text colors indicate the assigned frequency of occurrence of each feature.

important in identifying and treating the conditions associated with DS, facilitating survival and longevity.

Epidemiology

Humans are highly susceptible to numerical abnormalities in chromosomes (i.e., to aneuploidy). It has been estimated that around 20% of human oocytes and 1–2% of sperms are aneuploid. As a result, it is estimated that approximately one-third of all miscarriages in humans involve abnormal numbers of fetal chromosomes. This mechanism of natural elimination leads to a drastic reduction in the number of aneuploid liveborn babies). Only 0.3% of newborns are aneuploid.

The most important risk factor for aneuploidy in humans is the age of the mother. At the maternal age of 20 years, around 0.058% of liveborn infants are diagnosed with DS. This elevates to 4.12% at the maternal age of 46 years – a 70-fold increase in incidence rate. Unlike sperm cells, after entering into meiotic prophase I before birth, oocytes enter into an extended meiotic arrest which lasts till ovulation. This makes meiosis I very vulnerable to endogenous and exogenous factors. This is supported by the fact that in the majority of cases (65%), human trisomy 21 is due to errors in maternal meiosis I.

According to the United States Department of Health and Human Services, Centers for Disease Control and Prevention, the prevalence of DS from

1983–90 was found to be 9.2 per 10 000 liveborn infants in the 17 states studied. It estimated that 3000–5000 children are born with DS in the USA each year. The annual per-capita cost of $27 000 for children with DS between 0 and 1 year of age, approximately $135 million, will be spent for DS each year due to new cases.

The causes of death for people with DS include congenital circulatory defects, dementia, hypothyroidism, digestive system abnormalities, epilepsy, leukemia, and respiratory illnesses. In the United States, the number of deaths of people with DS is estimated to be 5.6 deaths per 10 000 total deaths (1983–97). As mentioned, life expectancy has increased dramatically for people with DS and is now nearly 60 years. From 1983 to 1997, the average median age at death for people with DS has increased 1.7 years per year – from 25 to 49. The death rate for children with DS presenting with heart abnormalities was reduced from 10.6 per 1000 in years 1988–93 to 1.7 per 1000 in 1994–99 in California.

Clinical Features

The most frequent DS manifestation in newborns is hypotonia. The clinical features of DS are dysmorphic changes involving craniofacial, neck, and limb abnormalities. Manifestations include, but are not limited to, brachycephaly, oblique palpebral fissures, epicanthic folds, flat nasal bridge, folded or dysplastic ears, open mouth, narrow palate, protruding and furrowed tongue, abnormal teeth, short neck with loose skin on nape, short and broad hands, incurved fifth finger, transverse palmar creases, gap between first and second toe, and hyperflexibility. Brushfield spots on the iris and hypotonia are also characteristic for people with DS. Cardiac malformations, thyroid dysfunction, digestive system abnormalities, and seizures are less prevalent (Figure 1), but have greater consequence for individuals with DS. Acute megakaryoblastic leukemia is the most common form of myeloid leukemia in DS and is present at rates reflecting a 466-fold increase in this population. Most males with DS are infertile, probably due to defects in spermatogenesis. Although females with DS often have reproductive system dysfunction, it is possible for them to become pregnant.

Individuals with DS face many medical problems. Perhaps the most important among them is nervous system dysfunction. In addition to hypotonia and seizures, virtually all individuals manifest intellectual disabilities (commonly known as 'mental retardation') to varying degree. The severity of intellectual disability is typically moderate, with Intelligent Quotient (IQ) averaging 40, with a range of 20–70. Children with DS often experience delays in the acquisition of developmental milestones as well as delayed disappearance of primitive reflexes (e.g., the grasp reflex and the Moro response). There are cognitive abnormalities attributable to the abnormal function of the hippocampus and prefrontal cortex (e.g., organizing verbal material according to its categorical structure and active retrieval of stored information); these are significantly impaired in DS. This is in contrast to implicit or nondeclarative memory, which is normal in DS. Pain and temperature sensations are abnormal for these individuals, as the peripheral nervous system is also affected. It is unclear to what extent peripheral nervous system involvement contributes to the burden of the disorder.

Neuroimaging studies in children and nondemented adults with DS complement clinical observations by revealing that their brains are often smaller in relation to their overall body weight. After accounting for the decrease in the size of the brain, the volumes of the hippocampus and cerebellum were smaller while those of the parahippocampal gyrus and subcortical gray matter were increased.

Adults with DS show additional manifestations of central nervous system (CNS) dysfunction. After the age of 40 years, virtually all DS patients develop a neuropathology indistinguishable from that of Alzheimer's disease (AD). Many, but perhaps not all, also show progressive cognitive decline resulting in dementia. It has been shown that the prevalence of dementia in institutionalized DS patients is about 8% in patients 35–40 years old, 55% in patients 55–59 years old, and 75% in DS patients older than 60 years. With increasing life expectancy, the disorders that affect brain aging with DS will become a significant source of medical and social concern.

Etiology

The genetic foundation for DS is HSA21 trisomy. Three hypotheses have been suggested to explain the link between genes and DS phenotypes: (1) the 'gene dosage effect' hypothesis proposes that dosage imbalance of specific genes or sets of genes on HSA21 causes specific DS phenotypes, (2) the 'sensitization model' states that specific phenotypes arise from one triplicated gene, or a small set of specific triplicated genes, but only when many other triplicated genes are also present and active, and (3) the 'developmental instability' hypothesis predicts that small, nonspecific effects of large numbers of genes at increased dose are responsible for DS phenotypes.

Neurobiological Bases for Nervous System Dysfunction in Down Syndrome

Human Studies

Due to the scarcity of postmortem material from humans with DS, the structural and functional bases for intellectual disability in DS are yet to be fully explored. Nevertheless, a number of observations have been made. Macroscopically, prenatal brain weight and shape, fronto-occipital diameter, and the sizes of the brain stem and cerebellum do not distinguish DS brains from controls. At birth, brain weight is generally decreased by approximately 12% in relation to controls and by age 3 to 5 months it is decreased even further. At this age there is shortening of the anteroposterior diameter due to a relative reduction in the size of the frontal lobes, flattening of the occipital poles, and a smaller brain stem and cerebellum. Other postnatal changes in DS include a narrow superior temporal gyrus and diminished sulcation. *In vivo* magnetic resonance imaging studies confirm these changes.

Microscopically, DS brain exhibits changes in synaptic structure and function and in basal forebrain cholinergic neurons.

Synaptic structure and function It has long been believed that synaptic dysgenesis contributes to the cellular basis for mental retardation in DS. If so, the consequences for neuronal circuit function would be significant. Focusing on an important set of structural indicators of synapses, dendritic spine density and morphology, dendritic length, and dendritic complexity have been examined. One caveat to the interpretation of such studies is the paucity of well-preserved tissue for analysis, resulting in relatively few careful studies and a still incomplete characterization of the changes present. Nevertheless, decreased synaptic density in the sensorimotor cortex was detected as early as 32–34 weeks of gestation and included a shift from intermediate contacts to primitive ones in the DS brain. A reduced number of spines, including in hippocampus and cortex, has been reported in both children and adults with DS. Sachio Takashima and colleagues studied a large number of pre- and postnatal DS cases. Although they failed to detect differences in the number of spines in fetal cases, they reported a significant reduction in the number of spines in visual cortex of infants with DS. Echoing this finding are reports on dendritic length in the neocortex: while there are no obvious changes in DS fetal cases, shorter lengths have been recorded in children with DS.

Using the Sholl analysis, Laurence E Becker studied the effects of age on the growth of dendritic length in DS and normal children. It was found that the total dendritic length per cell in layer 3 of the visual cortex was increased by 65% in the control group in transition from the infantile (6 months to 2 years old) to juvenile (>2 years old) period. However, the DS group showed a 50% decrease in the same period. Structural alterations in synapses are assumed to have functional consequences. Unfortunately, the only published physiological studies of human tissue are those of Brain S Scott and colleagues on dorsal root ganglion cells of the peripheral nervous system.

Basal forebrain cholinergic neurons Cholinergic neurochemical markers appear to be intact in prenatal stages of DS. Examining the brains of 3- to 12-month-old DS infants, the activities of the cholinergic enzymes choline acetyltransferase (ChAT) and acetylcholinesterase (AChE) are found to be normal. Remarkably, however, in DS cases ranging in age from 16 to 56 years, there is a significant reduction in the number of neurons and in nucleolar size in the nucleus basalis of Meynert. Of note, these changes are most marked in the oldest patient, who also showed AD-like pathology, drawing a further parallel between the two disorders. Early, preliminary tests of the cholinergic hypothesis have been carried out clinically through treatment with donepezil, an AChE inhibitor. In one trial, adults aged 20–41 years were entered in an open-label treatment with the AChE inhibitor donepezil. While some improvement in expressive language was recorded, the study suffered from a very small number of patients and from lack of a control group. In a more recent trial in children, aged 8–13 years, the same investigators examined the effects on language function of donepezil in an open-label study for 16 weeks in a small number of patients. Improved language performance has also been found in one test.

Mouse Models of Down Syndrome: Defining Nervous System Abnormalities

The development of animal models for disorders with strong genetic components is extremely useful for understanding underlying genetic and molecular mechanisms for pathogenesis. Studies in humans have helped to suggest chromosomal regions and genes that may contribute, but the ability to resolve the basis of important neurological phenotypes is limited. In contrast, recent studies in mouse models are providing important new insights.

The models are of three types. In the first, mice have been created that are trisomic for segments of the mouse chromosomes that are syntenic to HSA21.

In the majority of cases, DS is caused by triplication of the entire HSA21. At least 300 known and predicted genes have been mapped to HSA21. A comparison with the mouse genome shows that a majority of HSA21 orthologs are mapped to the distal end of MMU16 (about 23.2 Mb). Smaller numbers of genes are mapped to mouse chromosomes 17 (1.1 Mb) and 10 (2.3 Mb). It is noteworthy that some human genes are not present in the mouse genome and, correspondingly, that some mouse genes within these regions are not present in the human genome. Given this, it is clear that mouse genomic sequences cannot mirror precisely the genetics of DS. The second type of model consists of mice that carry an extra copy of all or part of HSA21. While of obvious benefit for research studies, such mice are much more difficult to produce and much less data exist concerning DS-relevant phenotypes. The third model includes transgenic mice overexpressing individual genes mapped to HSA21 or MMU16 (e.g., *App*, *Dyrk1A*, *Sod1*, *Sim2*, and *S100β*.

Mice Trisomic for Mouse Orthologs of HSA21

Ts16 and Ts65Dn mice Initial studies were carried out in mice trisomic for the entire sixteenth chromosome. In the first step toward developing a mouse model for DS, mice trisomic (Ts16) for the entire MMU16 were generated. These mice demonstrated a number of biochemical and structural changes similar to those of DS, including basal forebrain cholinergic neuron abnormalities. However, Ts16 mice do not survive beyond birth (Figure 2).

To produce a better model, Muriel T Davisson and colleagues, using γ-radiation, induced reciprocal translocation of T(16C3–4;17A2) and generated a mouse segmentally trisomic for MMU16 called Ts65Dn. These mice, which have normal life spans, have three copies of a fragment of MMU16 extending from *Znf295* to *Mrp139*. In behavioral analyses, Ts65Dn mice reveal significant deficits in hippocampally mediated cognition – spatial and nonspatial learning disabilities, as shown by the hidden platform

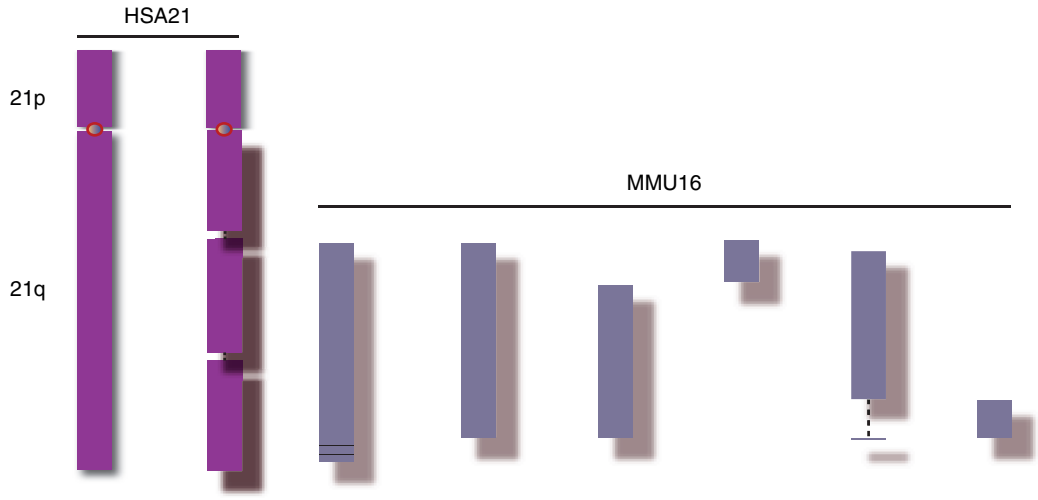

	Down syndrome	Tc1	Ts16	Ts65Dn	Ts1Cje	Ms1Cje/Ts65Dn	Ms1Rh1/Ts65Dn	Ts1Rh1
Number of triplicated genes	261–364	240–334	731*	104–132	81–85	22–46	70–98	33
Viability	+	+	–	+	+	+	+	+
Cognitive deficits	Moderate	Severe	ND	Severe	Moderate	Mild	ND	ND
Change in brain structure/function	+	+	+	+	+	ND	ND	ND

Figure 2 Schematic representation of triplicated segments of chromosome 21 (HSA21) in Down syndrome and MMU16 in different mouse models of DS. The asterisk indicates predicted genes for Ts16 (includes many nonorthologs to HSA21); ND, not determined. Based on data from Antonarakis SE, Lyle R, and Dermitzakis ET (2004) Chromosome 21 and Down syndrome: From genomics to pathophysiology. *Nature Reviews Genetics* 5: 725–738; O'Doherty A, Ruf S, Mulligan C, et al. (2005) An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes. *Science* 309: 2033–2037; Mural RJ, Adams MD, Myers EW, et al. (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* 296: 1661–1671; and Olson LE, Richtsmeier JT, Leszl J, et al. (2004) Direct testing does not support a chromosome 21 critical region as the cause of Down syndrome phenotypes when triplicated. *Science* 306: 687–690.

and probe tests in the Morris water maze (MWM). Furthermore, Ts65Dn mice recapitulate a variety of morphological changes detected in the DS brain and are increasingly used as mouse models of DS.

Ts1Cje and Ms1Ts65 mice A second segmentally trisomic mouse was generated by CJ Epstein and colleagues. Ts1Cje mice are trisomic for a fragment of MMU16 extending from *Sod1* to *ZNF295,* but not including a functional copy of *Sod1*. In contrast to Ts65Dn mice, to which they can be compared, Ts1Cje mice are hypoactive; they do show abnormalities in spatial learning and memory but these are less severe than in Ts65Dn mice. Ms1Ts65 mice are produced by crossing Ts65Dn mice with T(12:16)1Cje mice. They are segmentally trisomic for the genetic segment from *Mrp139* to *Sod1,* but not including a functional copy of *Sod1*. Thus, they represent a genetic complement to the Ts1Cje mice. Though little characterized, they show changes in spatial cognition that are significantly less severe than in Ts65Dn mice. This suggests that genes in both the centromeric and telomeric regions of Ts65Dn contribute to changes in hippocampal-dependent learning.

Ts1Rhr and Ms1Rhr mice A recent study examined the hypothesis suggested by human studies that triplication of a 5.4-Mb specific region of chromosome 21, referred to as the Down syndrome critical region (DSCR), causes many DS phenotypes. Using chromosomal engineering, Lisa E Olson and colleagues generated two lines of mice – Ms1Rhr and Ts1Rhr; the former is monosomic for the region of interest (about 33 genes conserved between mice and humans) while the latter is trisomic for it. Ms1Rhr mice crossed to Ts65Dn mice produced animals that were trisomic for most of the genes trisomic in Ts65Dn, but had only the normal two copies of the critical segment. No results of nervous system function have yet been reported, but studies on craniofacial features showed that the DSCR was neither necessary nor sufficient to produce the phenotypes present in Ts65Dn. These studies are important because they chart the direction that future studies will take in examining the contribution of gene segments and individual genes.

Mice Trisomic for HSA21

The limitations encountered by genetically modeling DS through triplication of mouse genes led A. O'Doherty and colleagues to attempt to produce a mouse genome containing an extra copy of HSA21, using the irradiation microcell-mediated chromosome transfer method. The mice appear to carry three copies of about 90% of HSA21 genes. Abnormalities in hippocampal-dependent learning and memory were detected and decreased long-term potentiation (LTP) was detected in physiological studies in dentate gyrus. In addition, there was a reduction in neuron density in the cerebellum. One shortcoming of this model is a high degree of mosaicism within tissues (e.g., in the brain, approximately one-third of cells were not trisomic).

Mice Transgenic for Individual Genes

In order to link DS features to specific genes, transgenic mice have been developed that overexpress single human HSA21 genes or their mouse orthologs. To date, mice overexpressing one of the following genes have been examined: *App*, *Cbs*, *Dscr1*, *Dyrk1A*, *Ets2*, *Hmg14*, *PfKl*, *S100β*, *Sim2*, and *Sod1*. Nervous system phenotypes including defects in learning and memory have been detected in mice transgenic for *App*, *Dyrk1A*, *S100β, Sim2*, and *Sod1*. For instance, mice overexpressing *App* show decreased working memory and learning in the water maze task.

The shortcoming in studies using only one transgene is that the genetic interactions characteristic of DS may not be replicated. Nevertheless, such studies are informative with respect to the possible contributions of individual genes. They make an even more compelling contribution when included in studies in which phenotypes of interest are examined both in the presence of a single gene as well as in mice in which only the third copy of the gene of interest is deleted. Using this paradigm, it may be possible to demonstrate whether an individual gene makes a contribution to a phenotype that is necessary and sufficient.

Defining and Characterizing Nervous System Abnormalities in Mouse Models of Down Syndrome

Synaptic Structure and Function

Pre- and postsynaptic structural and functional abnormalities have been detected in both Ts65Dn and Ts1Cje mice. Although less severe in Ts1Cje mice, both models exhibit significant alterations in spine structure, including a reduction in the density of dendritic spines, enlargement of spine heads, and reduced length of spine necks. Mara Dierssen and colleagues have studied dendrite morphology in 10-month-old Ts65Dn and control mice. They found a significant reduction in the area and number of dendritic branches, length of dendritic branches, and spine density, and in the basal dendritic arbors of layer III pyramidal cells in Ts65Dn mice. Alterations

in spines in the hippocampus were associated with failure to induce LTP in the dentate gyrus and CA1 areas of the hippocampus in Ts65Dn mice.

Basal Forebrain Cholinergic Neurons

Basal forebrain cholinergic neurons (BFCNs, the principle source of cholinergic innervation to the entire cortex and hippocampus) undergo degeneration in DS patients. Studies of BFCNs in Ts65Dn mice and Ts1Cje mice have been carried out. Using the p75 neurotrophin receptor ($p75^{NTR}$), a reliable marker of BFCNs, there was a significant reduction in the number and size of $p75^{NTR}$-positive BFCNs in Ts65Dn versus diploid (2N) controls at 12 months of age, and this was more marked at 18 months. These data have been replicated in several laboratories and using different markers for BFCNs. Nerve growth factor (NGF) plays a critical role in the maintenance of BFCN phenotype, and the changes in BFCN morphology replicate those in which BFCNs are deprived of NGF. NGF signaling was examined in Ts65Dn and Ts1Cje mice and their 2N controls. A series of experiments showed that failed NGF signaling was due to decreased axonal transport rather than to failure in binding or internalization of NGF–NGF receptor complexes. Very recent studies show that the defect in transport is not general, since no decrease in the retrograde transport of fluorogold is detected. This suggests that a specific failure in NGF retrograde mechanisms is causative in Ts65Dn mice.

One gene that is found in three copies in Ts65Dn mice and that has been linked to neurodegeneration is that for the amyloid precursor protein (APP); the *App* gene, the mouse ortholog of the human gene *APP*, is present in the centromeric portion of the trisomic segment in Ts65Dn. Examining NGF retrograde transport in these mice showed a drastic reduction in Ts65Dn mice. A significantly greater level of transport was detected in Ts1Cje mice and the changes in BFCNs were not detected. Interestingly, Ts65Dn mice with only two copies of *App* (Ts65Dn:$App^{+/+/-}$) achieved transport levels for NGF that were more than 50% of those in 2N mice. Thus, deleting one copy of *App* markedly reduced the severity of diminishment of NGF transport, similar to that in Ts1Cje mice. Morphological analysis of BFCNs revealed that while the area of BFCN cell bodies in Ts65Dn (with three copies of *App*) was markedly reduced, in Ts65Dn with only two copies of *App*, it was indistinguishable from that of 2N mice. These results are compelling evidence that increased gene dose for *App* contributes significantly to the disruption in NGF transport and atrophy of BFCNs in Ts65Dn mice. They suggest that the human gene plays the parallel role in people with DS.

Hippocampal Functions

LTP and long-term depression (LTD) are commonly used as models of synaptic plasticity involved in learning and memory. These parameters have been measured in several subregions of the hippocampus in Ts65Dn (e.g., dentate gyrus and CA1 areas). It appears that failed LTP induction in the dentate gyrus is mainly due to increased inhibition, since suppressing inhibition with picrotoxin, a γ-aminobutyric acid (GABA)-A receptor antagonist, LTP elicitation could be restored.

Cerebellum

DS patients show significantly smaller cerebellar volumes as compared to normal individuals. Olson and colleagues found about a 12% reduction in Ts65Dn mice compared to controls. This may explain failed motor coordination reported in Ts65Dn mice.

Basis of DS Therapy

At present there are no proved treatments for cognitive dysfunction in people with DS. Linking individual genes to particular manifestations of DS is relevant to the goal of understanding and treating its cognitive dysfunction. The goal of linking specific genes to neurological phenotypes is now much closer to realization. In the future, it may be possible to define many more such linkages, thus enhancing opportunities for both understanding the genetic and cellular mechanisms of nervous system dysfunction in people with DS and for treating them effectively.

See also: Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Cognitive Control and Development; Development of Behavior; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Down Syndrome: A Disorder of Histogenesis; Gene Therapy: Genetically Modified Cells; Language: Learning Impairments; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Antonarakis SE and Epstein CJ (2006) The challenge of Down syndrome. *Trends in Molecular Medicine* 12: 473–479.

Antonarakis SE, Lyle R, and Dermitzakis ET (2004) Chromosome 21 and Down syndrome: From genomics to pathophysiology. *Nature Reviews Genetics* 5: 725–738.

Berg JM and Korossy M (2001) Down syndrome before Down: A retrospect. *American Journal of Medical Genetics* 102: 205–211.

Down JL (1995) Observations on an ethnic classification of idiots. 1866. *Mental Retardation* 33: 54–56.

Epstein CJ (1990) The consequences of chromosome imbalance. *American Journal of Medical Genets* supplement 7): 31–37.

Epstein CJ (2001) Down syndrome (trisomy 21). In: Scriver CR, Beaudet AL, Valle D, et al. (eds.) *The Metabolic and Molecular Bases of Inherited Disease*, pp. 1223–1256. New York: McGraw-Hill.

Gardiner K, Fortna A, Bechtel L, et al. (2003) Mouse models of Down syndrome: How useful can they be? Comparison of the gene content of human chromosome 21 with orthologous mouse genomic regions. *Gene* 318: 137–147.

Hattori M, Fujiyama A, Taylor TD, et al. (2000) The DNA sequence of human chromosome 21. *Nature* 405: 311–319.

Jacobs PA, Baikie AG, Court Brown WM, and Strong JA (1959) The somatic chromosomes in mongolism. *Lancet* 1: 710.

Kevles DJ (1985) *In the Name of Eugenics: Genetics and the Uses of Human Heredity.* Cambridge: Harvard University Press.

Lejeune J, Turpin R, and Gautier M (1959) Mongolism; a chromosomal disease (trisomy) [French, transl.]. *Bulletin of the National Academy of Medicine* 143: 256–265.

Mural RJ, Adams MD, Myers EW, et al. (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* 296: 1661–1671.

O'Doherty A, Ruf S, Mulligan C, et al. (2005) An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes. *Science* 309: 2033–2037.

Olson LE, Richtsmeier JT, Leszl J, et al. (2004) Direct testing does not support a chromosome 21 critical region as the cause of Down syndrome phenotypes when triplicated. *Science* 306: 687–690.

Salehi A, Delcroix JD, and Mobley WC (2003) Traffic at the intersection of neurotrophic factor signaling and neurodegeneration. *Trends in Neuroscience* 26: 73–80.

Salehi A, Delcroix JD, Belichenko PV, et al. (2006) Increased App expression in a mouse model of Down's syndrome disrupts NGF transport and causes cholinergic neuron degeneration. *Neuron* 51: 29–42.

Sofroniew MV, Howe CL, and Mobley WC (2001) Nerve growth factor signaling, neuroprotection, and neural repair. *Annual Review of Neuroscience* 24: 1217–1281.

Wright D (2001) *Mental Disability in Victorian England: The Earlswood Asylum 1847–1901.* Oxford: Clarendon.

Down Syndrome: A Disorder of Histogenesis

T F Haydar, George Washington University School of Medicine, Washington, DC, USA
L Chakrabarti, Children's National Medical Center, Washington, DC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

With an incidence of 1 in 800 live births, Downs syndrome (DS) is the most common congenital cause of mental retardation. DS is caused by trisomy of chromosome 21, and it is believed that overexpression of chromosome 21 genes is responsible for the mental retardation and other phenotypic abnormalities of DS. Individuals with DS are typically microcephalic and have characteristic facial features, hypotonia, and small stature. Other phenotypes include congenital heart defects, premature aging, and leukemia. Although its severity may vary, mental retardation remains the most striking and consistent feature of DS. Recent advances in mapping the human genome and constructing mouse models have accelerated progress toward understanding the etiology of the neurobehavioral phenotype in DS. Evidence suggests that mental retardation in DS is associated with both pre- and postnatal events and that neurogenesis, cell division, and survival may be common abnormalities in the DS central nervous system (CNS).

DS children have different degrees of developmental brain abnormalities associated with CNS maturation defects and cortical and hippocampal dysgenesis in particular. Although a wide range of studies have found no differences in terms of shape, weight, and morphology of the DS brain at birth, gross neuropathological differences between DS and control brains have been consistently observed postnatally. For example, the anterior–posterior diameter of brain hemispheres is smaller in postnatal DS brain. In early infancy, narrowing of the superior temporal gyrus is noted together with diminished size of the brain stem and cerebellum. Architectonic abnormalities of the cerebral cortex are also found in infant and adolescent DS brains. Since normal cortical development requires the generation and migration of cells from the ventricular and subventricular zones to the cortical mantle, delayed and disorganized cortical lamination observed in fetal DS brain are believed to be due, in part, to changes in neuronal proliferation and migration. In addition, areal measurements of the dentate gyrus and pyramidal cell layers in the fetal and adult hippocampus indicate that these structures are significantly smaller in DS brain than those in control subjects. For example, a recent high-resolution magnetic resonance imaging (MRI) study gathered quantitative neuroimaging data in children and young adults with DS and reported significantly smaller hippocampal volume in DS groups. These system-level defects reflect abnormalities in histogenesis, axonal and dendritic arborization, and synaptic development which may underlie many of the cognitive deficits seen in DS individuals.

Although conspicuous morphological abnormalities begin to be apparent in postnatal DS brain, reduced expression of synaptosomal-associated proteins in the fetal DS brain indicates that fetal life in DS starts with impaired dendritic spines and synaptosomal structures; these anomalies are manifested in postnatal brain as shortened basilar dendrites and a decreased number of spines with altered morphology. These changes during early life culminate in delayed myelination, fewer neurons, and lower neuronal and synaptic density in the adult DS brain. Although the main underlying factors for these synaptic changes remain to be fully identified, a major hypothesis is that prenatal abnormalities in corticogenesis precede the synaptic abnormalities. Brain development occurs via highly regulated stages of proliferation, migration, differentiation, and synaptogenesis, and each stage must be successfully and expeditiously completed before onset of the next stage to result in proper development and cognitive function. As such, the anatomical and cytological findings in DS and the progressive changes in cognitive function are all consistent with abnormalities during the early stages of embryonic brain development. It is entirely possible that the changes in the histogenesis of the DS brain may be unrelated to the synaptic abnormalities which develop later. Alternatively, it is possible that these two phenomena are linked. This issue has not yet been resolved due to the lack of appropriately staged DS and control tissue. However, considerable evidence in support of the early embryonic etiology of mental retardation in DS has been provided by analyses in several mouse models of DS (Figure 1). Therefore, the relationship(s) between prenatal histogenetic abnormalities and later cognitive dysfunction can now be investigated in animal models of DS.

Ts16

The trisomy 16 (Ts16) mouse model of DS, developed using spontaneous Robertsonian translocations to generate aneuploidy for mouse chromosome 16, take advantage of the substantial homology between human chromosome 21 and mouse chromosome 16. Ts16 fetuses have a number of phenotypic characteristics

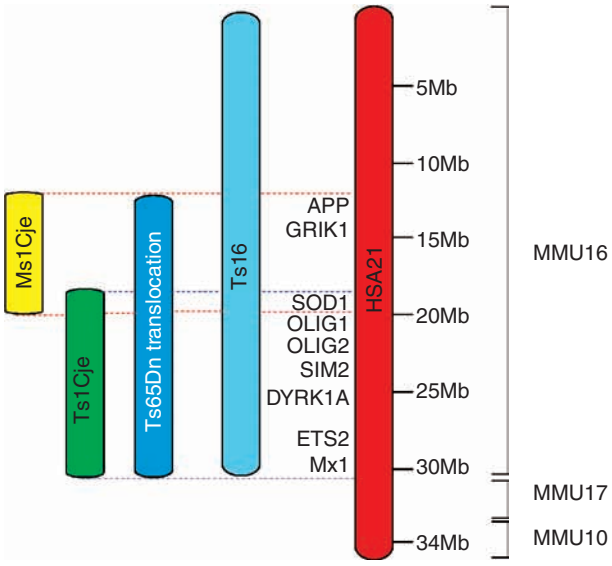

Figure 1 Comparison of triplicate chromosome regions in mouse models of DS with the location of specific genes on human chromosome 21 (HSA21). Regions of HSA21 are syntenic with MMU16, MMU17, and MMU10 (mouse chromosomes 16, 17, and 10). The Ts16 model contains an extra copy of the entire MMU16, whereas Ts65Dn, Ts1Cje, and Ms1Cje have smaller segmental translocations.

similar to those of DS subjects including growth retardation, edema, cardiac defects, and anasarca. A wide range of studies on prenatal brain development demonstrate that Ts16 fetuses exhibit altered corticogenesis. In particular, Ts16 neocortical development is described by a delay in radial expansion of the dorsal telencephalic wall and altered timing of neurogenesis of subplate pathfinding neurons. These delays are primarily due to abnormal proliferation of the ventricular zone and subventricular zone stem and progenitor cells, including longer duration of S-phase of the cell cycle and a delay in neuron-generating divisions. Moreover, reduced numbers of neocortical founder stem cells and slowed tangential expansion of the neuroepithelium in Ts16 result in microcephaly. Consequent to the delayed development of the Ts16 neocortex, a reduction in the number of thalamocortical axons is observed just prior to birth in Ts16 mouse embryos. Thus, disturbances in the rate of neocortical histogenesis, found to be due to abnormalities in the proliferating stem cell populations, result in delayed connectivity between the neocortex and subcortical structures.

Intriguingly, these histogenetic changes, which occur during the initial neurogenesis period in Ts16, are compensated for before birth; there is a massive production of neurons during the late stages of neurogenesis, eventually yielding the proper number of neurons. This observation of a histologically normal neocortex at birth is highly reminiscent of findings in DS. However, despite the fact that the Ts16 cortex is of normal thickness at birth, the transient delay during neurogenesis may predispose trisomic embryos to abnormalities in early synaptogenesis. Taken together, studies in Ts16 demonstrate that perturbations in the timing of neuron generation before the final numbers of neurons are produced can lead to subsequent neocortical abnormalities similar to those found in DS. Unfortunately, Ts16 embryos die perinatally, so research on this model has been limited to fetal stages of development. In addition, the Ts16 model also includes genes on mouse chromosome 16 that are not syntenic with human chromosome 21, opening the possibility that some phenotypes observed in Ts16 may not be relevant to DS. Together, these drawbacks of the Ts16 model limit the investigation of links between transient histogenetic delays and mental retardation in DS.

Ts65Dn

The Ts65Dn mouse may be a more relevant trisomy mouse model of DS as it has only a segmental trisomy of chromosome 16 that is orthologous to the DS critical region (DSCR) on human chromosome 21. In contrast to Ts16 animals, Ts65Dn mice live into adulthood and thereby provide the opportunity to test the embryonic theories of DS etiology. Although the genes triplicated in Ts65Dn represent most of those in the DSCR, these mice do not show all of the gross abnormalities frequently observed in DS, including cardiovascular abnormalities and senile plaques or neurofibrillary tangles in aged animals. Nevertheless, Ts65Dn mice display many phenotypic similarities with DS including altered craniofacial structure, obesity, growth retardation, premature aging, and behavioral deficits. Cognitively, there are significant deficits in the development of a number of sensory and motor tasks and defects in learning and memory mediated by the hippocampus in Ts65Dn. These cognitive deficits are analogous to the mental retardation seen in DS individuals. Complementing the behavioral findings, physiological analyses demonstrate a marked suppression of long-term potentiation in the CA1 and dentate gyrus of Ts65Dn hippocampus. In addition, long-term depression is augmented in Ts65Dn CA1 hippocampal neurons. Alterations in inhibitory tone have been hypothesized to underlie these changes in synaptic plasticity.

Stereological morphometric studies demonstrate irregularities in selected regions of the postnatal Ts65Dn mouse brain. Reductions in the volume of CA2, neuronal density in the CA3 and dentate gyrus, and synaptic density in CA1, CA3, and dentate gyrus suggest that altered neurogenesis and/or survival in the Ts65Dn hippocampus may combine with impaired synaptic connectivity to cause the cognitive

impairments. Intriguingly, aged (13–15 months) mice show reduced proliferative capacity in the dentate gyrus, suggesting that postnatal neurogenesis may also be affected in Ts65Dn. In the neocortex, Ts65Dn pyramidal cells are smaller and exhibit less dendritic branching compared to control littermates and these effects parallel those obtained in humans. High-resolution MRI and histological analyses also reveal neuroanatomical parallels between DS and Ts65Dn in the cerebellum. For example, cerebellar volume is decreased in Ts65Dn mice due to fewer granule cells in both the internal granule layer and the molecular layer. Recently, the origin of the granule cell deficit in Ts65Dn cerebellum has been traced to precursors in early postnatal development. These precursors show a substantially reduced mitogenic response to sonic hedgehog (Shh) signaling protein. Purified cultures of trisomic granule cell precursors show a reduced but dose-dependent response to Shh *in vitro*. Interestingly, systemic treatment of newborn Ts65Dn mice with an agonist of Shh pathway increases mitosis and restores granule cell precursor populations *in vivo*. Thus, there is substantial evidence for histogenetic changes being observed during both embryonic and postnatal development in both of these mouse models of DS. Future work will determine whether delays similar to those found during corticogenesis in Ts16 are also present in Ts65Dn, and this will allow a frank examination of the hypothetical links between early histogenetic abnormalities and later cognitive deficits in DS. Altogether, it is now clear that developmental abnormalities in histogenesis and synaptogenesis precede the synaptic and functional deficits which eventually present as mental retardation in DS. These animal model studies now provide a system to test the hypothesis that early abnormalities in CNS formation may underlie the later functional, synaptic, and cognitive deficits found in models of DS.

Basal Forebrain Deficits

Significant abnormalities are observed during development and maturation of the basal forebrain in DS, and in Ts16 and Ts65Dn animal models. DS individuals have fewer cholinergic neurons in the nucleus basalis of Meynert compared to age-matched controls. Similarly, Ts16 mice exhibit a reduction in cortical cholinergic activity and a reduced number of cholinergic cells in the basal forebrain. These cells appear to cease proliferation prematurely, particularly in those regions projecting to the hippocampal formation. Although the number and size of basal forebrain cholinergic neurons (BFCNs) is similar in Ts65Dn and diploid mice during development, adult Ts65Dn mice show an atrophy and progressive loss of BFCNs. Moreover, Ts65Dn mice develop septohippocampal cholinergic neuron degeneration at 12 months of age. These cellular changes are compensated for by increases in choline acetyltransferase (CHAT) enzyme activity of the remaining cholinergic terminals in the hippocampus. Nevertheless the loss of medial septum projections to the hippocampus in Ts65Dn mice may be a causative factor in the learning and context discrimination deficits found in Ts65Dn. Similar to Ts65Dn mice, DS infants appear to have normal septohippocampal cholinergic system and degeneration of this system only appears with aging.

Apoptosis in DS

Substantial evidence suggests that increased apoptosis is involved in many of the abnormalities in DS. DS neurons display an increased rate of apoptosis *in vitro* which may be due to overexpression of the chromosome 21 transcription factor Ets2. Similar to DS, neuronal cell loss is found to be accelerated in Ts16 neuronal cultures as cultured Ts16 hippocampal neurons display an inherited defect in survival response mediated by low concentrations of glutamate; this accelerated neuronal death in Ts16 occurs by caspase-mediated apoptosis. *In vivo* evidence also suggests an increased rate of prenatal apoptosis in DS and Ts16 neocortex. Since normal CNS development involves the sequential differentiation of multipotent stem cells, any alteration of the numbers of stem cells, their self-renewal ability, or their proliferative capacity is likely to have major effects on the appropriate development of the nervous system.

Ts1Cje and Ms1Cje Mouse Models of DS

Ts1Cje and Ms1Cje are recently described partial trisomy 16 mouse models of DS that contain smaller regions of chromosome 16 than the Ts16 or Ts65Dn models. Although few studies on histogenesis and neuron proliferation have been conducted on these newer mouse models, detailed investigation of cerebellar phenotypes has revealed a reduction in granule cell density in both Ts1Cje and Ms1Cje. Moreover, the Ts1Cje hippocampus has reduced long-term potentiation and increased long-term depression, suggesting that abnormal synaptic plasticity may be a common feature of mouse models of DS.

See also: Basal Forebrain and Memory; Cholinergic System Imaging in the Healthy Aging Process and Alzheimer Disease; Developmental Disability and Fragile X Syndrome: Clinical Overview; Down Syndrome; Genomic Disorder and Gene Expression in the Developing CNS; Hippocampus; Programmed Cell Death; Spines and Mental Disorders.

Further Reading

Belichenko PV, Masliah E, Kleschevnikov AM, et al. (2004) Synaptic structural abnormalities in the Ts65Dn mouse model of Down syndrome. *Journal of Comparative Neurology* 480: 281–298.

Galdzicki Z, Siarey R, Pearce R, Stoll J, and Rapoport SI (2001) On the cause of mental retardation in Down syndrome: Extrapolation from full and segmental trisomy 16 mouse models. *Brain Research Reviews* 35: 115–145.

Galdzicki Z and Siarey RJ (2003) Understanding mental retardation in Down's syndrome using trisomy 16 mouse models. *Genes Brain Behavior* 2: 167–178.

Golden JA and Hyman BT (1994) Development of the superior temporal neocortex is anomalous in trisomy 21. *Journal of Neuropathology and Experimental Neurology* 53: 513–520.

Haydar TF, Blue ME, Molliver ME, Krueger BK, and Yarowsky PJ (1996) Consequences of trisomy 16 for mouse brain development: Corticogenesis in a model of Down syndrome. *Journal of Neuroscience* 16: 6175–6182.

Haydar TF, Kuan CY, Flavell RA, and Rakic P (1999) The role of cell death in regulating the size and shape of the mammalian forebrain. *Cerebral Cortex* 9: 621–626.

Haydar TF, Nowakowski RS, Yarowsky PJ, and Krueger BK (2000) Role of founder cell deficit and delayed neurogenesis in microcephaly of the trisomy 16 mouse. *Journal of Neuroscience* 20: 4156–4164.

Korenberg JR (1995) Mental modeling. *Nature Genetics* 11: 109–111.

Kuida K, Haydar TF, Kuan CY, et al. (1998) Reduced apoptosis and cytochrome c-mediated caspase activation in mice lacking caspase 9. *Cell* 94: 325–337.

Pinter JD, Eliez S, Schmitt JE, Capone GT, and Reiss AL (2001) Neuroanatomy of Down's syndrome: A high-resolution MRI study. *American Journal of Psychiatry* 158: 1659–1665.

Reeves RH, Irving NG, Moran TH, et al. (1995) A mouse model for Down syndrome exhibits learning and behavior deficits. *Nature Genetics* 11: 177–184.

Wisniewski KE (1990) Down syndrome children often have brain with maturation delay, retardation of growth and cortical dysgenesis. *American Journal of Medical Genetics Supplement* 7: 274–281.

Dream Function

D Kuiken, University of Alberta, Edmonton, AB, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Theories of dream function posit a correspondence between the effects of dreaming and what enables the dreamer, in some sense, to thrive. Perhaps the most common proposal is that dreams metaphorically represent disturbing life events and allow the covert review of self-protective or self-restorative responses. Another long-standing hypothesis is that dreaming distinctively reorganizes memories for newly learned tasks and enhances the retrieval of related knowledge and skills. And, still another conception is that dreaming directly alters attention, evokes feeling, and enhances the flexibility and emotional relevance of postdream thought. As is evident from these broad themes, models of dream function vary considerably in their scope and in the nature of the evidence required to substantiate them.

Contrasting Conceptions of 'Function'

Theories of dream function differ markedly in their articulation of the temporal locus of what is beneficial about dreaming. Researchers have sometimes taken a broad perspective, hypothesizing that dreaming facilitates long-term adaptation (e.g., species survival). In contrast, some investigators have identified immediate and transient dream effects (e.g., postawakening attentional flexibility). Also, theories of dream function are usually presented as though all dreams fulfill a common function. Perhaps dreaming is a sufficiently uniform phenomenon to serve consistently some function (or functions), but, by abandoning strict assumptions about highly general dream effects, the contrasting effects of different dream types – for example, nightmares versus night terrors, or rapid eye movement (REM) dreams versus non-REM (NREM) dreams – have also been fruitfully considered.

Contrasting conceptions of the temporal locus and uniformity of dream effects became important when evaluating the efficacy of REM sleep research protocols during the past half century. Although researchers originally posited an isomorphism between REM sleep and dreaming, it is now widely acknowledged that dreaming in some form occurs elsewhere in the sleep cycle, most notably during sleep onset and stage 2 NREM sleep. The implications of these observations remain controversial. Even though dreaming occurs during NREM sleep, it is still unclear whether dreams reported upon awakening from REM sleep are qualitatively or merely quantitatively different from those reported upon awakening from NREM sleep. If REM dreams are qualitatively different from NREM dreams, comparison of the consequences of REM and NREM sleep deprivation may help to establish the contrasting functions of REM and NREM dreams. On the other hand, if NREM dreams are only somewhat more intense but of the same basic type as REM dreams, the psychological functions of 'typical' dreaming can only be weakly clarified by comparing the effects of REM and NREM sleep deprivation.

Moreover, even if the differences between REM and NREM dreams were clearly established, and REM sleep physiology could confidently be regarded as a useful index of REM dreaming, the use of REM sleep research protocols in studies of dream function would leave unresolved an even more basic issue: the psychological function of REM dreaming is not reducible to the physiological function of REM sleep. Thus, even if REM sleep were reliably associated with a distinct form of dreaming, the challenge of differentiating the physiological effects of REM sleep from the psychological effects of REM dreaming would remain. This issue is far from resolved and contributes to very different perspectives on the significance of REM sleep research protocols for studies of dream function.

Dream Carryover Effects

Awakenings from REM and NREM sleep are followed by a brief period ($\approx$20 min) during which thoughts and feelings are systematically altered. Observations during this period are often thought to reflect the nature of mental activity during the sleep stage that immediately precedes awakening. However, these observations are also appropriately considered as the direct influence of the preceding sleep stage on subsequent waking experience.

Carryover Effects of REM Dreaming

In cats, it has been observed that (1) high-amplitude pontogeniculooccipital (PGO) waveforms accompany the orienting response during waking, (2) comparable high-amplitude PGO waves occur spontaneously during REM sleep, and (3) novel auditory stimuli that elicit the orienting response during waking reliably induce PGO waves in both REM and NREM sleep. Such observations suggest that REM sleep is characterized by persistent and spontaneous activation of the orienting response. Correspondingly, in humans, there is evidence that an automatic, stimulus-induced shift of

attention toward unexpected stimuli occurs more quickly during the period immediately following awakening from REM sleep than it does following awakenings from NREM sleep.

Fluidity of thought Such attentional flexibility helps to explain dreamers' post-REM readiness to entertain infrequently considered possibilities. For example, individuals awakened from REM sleep show greater priming by weakly associated words (e.g., thief–wrong) than by strongly associated words (e.g., hot–cold), whereas individuals awakened from NREM sleep show greater priming by strongly associated than by weakly associated words. Also, individuals awakened from REM sleep use strategies that differ from those used during wakefulness to solve simple problems (e.g., anagrams), whereas, after awakening from NREM sleep, strategies for solving these problems more nearly resemble those used during wakefulness. Finally, after REM deprivation, individuals awakened from REM sleep score higher on fluidity and flexibility in traditional tests of divergent thinking than do those deprived of NREM sleep. In sum, numerous findings support the contention that awakenings from REM sleep are more likely than awakenings from NREM sleep to be followed by a brief period of fluid thinking. Whether this fluidity reflects the lingering effects of REM dreaming, rather than the lingering effects of REM physiology, remains, as indicated earlier, unresolved.

Felt presence The fluid thinking that follows REM dreaming acquires additional significance when it is considered that carryover effects also involve the readiness to see motion where it objectively is not. Individuals awakened from REM sleep are more likely than those awakened from NREM sleep to (1) perceive one dot moving between two positions when two stationary dots are alternately presented (the beta phenomenon) and (2) perceive movement in a static figure after a period of concentration on a rotating spiral (the spiral aftereffect). This 'feeling' of motion is complemented by enhanced 'feeling for' objects. After awakenings from REM sleep, performance on a left-handed tactile-kinesthetic recognition task is better than after NREM awakenings (even though analogous differences are not evident with the enumeration of dot patterns). Together these results suggest that the felt presence of objects – rather than simply their visual appearance – is intensified during the period that immediately follows REM dreaming.

Felt position There is also evidence that REM dreaming accentuates the felt sense of the dreamer's position in relation to the external world. It has been proposed that, during REM sleep, the initiation of bodily movement without corollary proprioceptive feedback (a pattern associated with intensification of the orienting response) precipitates perceived adjustments in bodily position. This possibility is consistent with evidence that REM deprivation increases REM density, the associated sense of self-participation (i.e., active participation in the dream scenario), and also the perception of human movement in ambiguous figures after awakening.

Emotional feeling The carryover of emotional dream feelings might seem to follow directly from the preceding findings. However, emotional feeling is attenuated immediately following awakening from a typical REM sleep period, either because of the lingering atonia of that sleep stage or, more likely, because of the mood-regulating function of REM dreams. Regardless of its origin, imaginative expression (e.g., creating stories in response to ambiguous stimuli) immediately after REM awakenings seems to evoke emotional feeling more than does straightforward dream reflection.

Auxiliary effects Carryover effects of typical REM dreaming are subtle and transient. Nonetheless, post-REM fluidity of thought is sometimes not only greater than after NREM sleep but is also greater than during ordinary wakefulness. Correspondingly, there is evidence that reflection on remembered REM dreams is more likely than reflection on remembered NREM dreams to facilitate feeling expression during insight-oriented therapy. Although the carryover effects of typical REM dreaming may exert their adaptive influence especially through timely and vivid recollection, this auxiliary function of dreaming has considerable practical import.

Carryover Effects of Intensified Dreaming

Sometimes dream carryover effects are intensified during the transition from REM sleep to wakefulness and during the period immediately following REM dreaming. These extreme forms of dreaming sometimes have lasting significance.

Sleep paralysis episodes In sleep paralysis episodes, the general atonia of REM sleep persists despite the capacity for external perception and the presence of consciousness. In this liminal state, the accompanying hallucinatory experiences are of three types. One type involves the felt presence of a spatially distant but threatening figure (accompanied by assorted noises, humanoid apparitions, and sensations of being touched), a second type involves the felt presence

of a spatially near and disturbing figure (accompanied by breathing difficulties, feelings of suffocation, and thoughts of impending death), and a third type involves sensations of linear and angular acceleration (e.g., flying, floating), fictive motor movements (ranging from simple arm movements to apparent locomotion), and feelings of bliss. Consistent with laboratory research on REM sleep carryover effects, these three types of transitional dream experience involve the distinctly felt but nonvisual presence of an 'other' and a keenly felt but anomalous sense of 'self.'

The effects of sleep paralysis episodes on waking thoughts and feelings are often direct and persistent. The lingering 'reality' of these dream experiences, traceable to the compellingly felt presence of external objects and the experience of an unembodied but uniquely positioned self, contribute to beliefs about spirit figures and self-transcendence. Whether these effects are lasting and whether they are beneficial depend upon the cultural context – even when these dreams challenge, rather than merely reflect, the surrounding cultural milieu.

Impactful dreams Significant features of sleep paralysis episodes correspond to empirically established types of impactful dreams. Impactful dreams are defined as those that directly influence postdream thoughts and feelings, most likely through the intensification of REM carryover effects. One type, the nightmare, with features such as intense fear, the avoidance of threatening others, and physical metamorphoses, is followed by externally directed vigilance and apprehension. A second type, the existential dream, with features such as intense sadness, rejection by or separation from significant others, and spontaneous feeling change, is followed by sadness, autobiographical review, and changed self-perception. A third type, the transcendent dream, with features such as surprise, magical accomplishment, and shifts in visual–spatial orientation (floating, flying), is followed by astonishment and reported spiritual change. In forms that are analogous to sleep paralysis episodes, nightmares involve the threatening presence of distant external objects, existential dreams involve the disturbing presence of near external objects, and transcendent dreams involve an anomalously positioned self.

The effects of impactful dreams also seem direct. Dreamers are often reluctant to have them interpreted, and there is evidence that, although they may alter the way dreamers understand their daily lives, they do not provide optimal material for therapeutic dream reflection. However, the auxiliary processes by which such intense dream experiences do exert influence on the dreamer's personal, social, or cultural life have not been systematically investigated.

Memory Consolidation

The long-standing hypothesis that memories are reorganized and stabilized during sleep became more focused during the 1970s when researchers began to examine specifically whether REM sleep was the psychobiological state within which memory consolidation occurs. At that time, the apparent isomorphism between REM sleep and dreaming suggested that the processes by which dreams are formed are also the processes by which memories are consolidated. So, REM deprivation became a common means by which memory consolidation was hypothetically disrupted, and assessing the effects of prior learning on REM sleep parameters was often used to assess increased pressure for REM-dependent consolidation. Although research initially supported the memory consolidation hypothesis among animals but not humans, in recent years, studies of human memory during sleep support the role of REM sleep, and by implication REM dreaming, in memory consolidation among humans.

Memory Consolidation and REM Sleep

A turning point in this research occurred with the realization that REM sleep may play a particularly important role in the consolidation of procedural, as opposed to declarative, memory. Review of the literature indicated that, when a task requires the development of perceptual-procedural, verbal-procedural, or cognitive-procedural skills, REM sleep facilitates learning; when the task requires the retrieval or recognition of generic semantic information, REM sleep does not facilitate learning.

Effects of learning on REM sleep In one commonly used research paradigm, a learning task is presented prior to sleep and then REM sleep parameters (e.g., REM density, percentage of sleep in stage REM) are assessed. For example, when people are required to adapt to lenses that distort their visual field, there is a subsequent increase in the percentage of stage REM sleep. Comparable effects have been found after learning verbal-procedural skills (e.g., second language learning) and cognitive-procedural skills (e.g., solving a conceptual problem that involves recursive steps).

Effects of REM deprivation on consolidation In the frequently used REM deprivation paradigm, people are awakened whenever they enter REM sleep, or, in comparison conditions, whenever they enter NREM sleep. In general, as hypothesized, when people are deprived of REM sleep after learning perceptual-procedural skills (e.g., a complex visual discrimination task) or cognitive-procedural skills (e.g., the recursive cognitive

problem), their performance during retesting is impaired. These findings have sometimes been challenged because deprivation procedures are often confounded with the stress precipitated by repeated awakenings from REM sleep. However, some studies of REM deprivation have demonstrated that performance disruption is still evident several days after the termination of REM deprivation and, hence, after recovery from the related stress or fatigue. Also, REM deprivation during specific temporal windows, sometimes days after the learning task, sometimes disrupts performance even more than does REM deprivation immediately after learning, precluding the possibility that stress associated with REM deprivation accounts for impaired performance.

Convergent findings from other paradigms Other research methods have provided converging evidence that REM sleep facilitates perceptual-procedural and cognitive-procedural memory. Research using functional neuroimaging procedures indicates that specific brain areas activated while learning a serial reaction time task are significantly more active during REM sleep among those who have previously trained on the task than among those who have not. Also, when people who have learned a recursive cognitive skill in the presence of a clicking sound are presented clicks coincident with their rapid eye movements during REM sleep, they perform better on retesting than do people who have learned the skill and then are presented clicks during a period of REM sleep without rapid eye movements or during a period of NREM sleep. These findings suggest a special role for the phasic events of REM sleep in memory consolidation, a possibility bolstered by evidence that hippocampal theta rhythms, which have been implicated in memory consolidation during wakefulness, are not only prominent during REM sleep but also are precipitated by the PGO waveforms that generate the REMs, twitches, and other phasic features of that sleep stage.

However, because hippocampal theta rhythms are especially implicated in the consolidation of declarative rather than procedural memory, the association between hippocampal theta and REM sleep is inconsistent with the notion that REM-mediated memory consolidation is specific to procedural learning. Moreover, there is evidence from a partial sleep paradigm that declarative memory involving emotional information is enhanced by REM sleep. Specifically, when memory is assessed after periods of undisturbed sleep during either the first half of the night (when NREM sleep predominates) or the second half of the night (when REM sleep predominates), the consolidation of emotional declarative memories (e.g., for emotional stories), but not of neutral declarative memories (e.g., for nonemotional stories), is enhanced by late-night (REM-dominated) sleep. These findings suggest that declarative memory for emotional events, although dependent upon hippocampal theta activity, is enhanced during REM sleep by amygdalar modulation of hippocampal function. Since, within any particular learning task, declarative and procedural learning are often mixed, there clearly is need for more refined characterization of the particular kinds of memories that are consolidated during REM sleep.

Memory Consolidation and NREM Sleep

Although there is evidence that sleep stages other than stage REM are implicated in memory consolidation, identification of their distinctive contributions has only begun. There are replicable findings indicating that NREM sleep facilitates performance on motor tasks (e.g., a pursuit rotor task). And, there is evidence that a word-pair association task influences some parameters of early-night (NREM-dominated) sleep (e.g., the number of sleep spindles) and that the disruption of early-night (NREM-dominated) sleep interferes with this type of declarative memory. This might be expected from dual-process theories, which suggest that REM and NREM sleep have specialized memory consolidation functions. However, other evidence (e.g., correlations between retesting performance and time spent in both NREM and REM sleep) suggest that NREM and REM sleep independently contribute to memory consolidation. According to sequential process theories, the orderly succession of NREM and REM sleep is required for memory consolidation, regardless of the nature of the task.

Dream Content and Memory Consolidation

On a few occasions, researchers have directly linked REM-dependent memory consolidation to dream content. For example, individuals who have successfully adapted to lenses that distort their visual field report dreams including visual and motor difficulties, as well as misfortune and confusion. Similarly, among people learning a second language, those performing successfully are among the first to incorporate that second language into their own dream communications. However, there is ample room in this literature for closer examination of the relations between the incorporation of the presleep task (into both REM and NREM dreams) and subsequent task performance.

Such research must take into account the extended temporal reach of dream formation processes. The incorporation of challenging life events into dreams appears to be modulated by a 7-day rhythm. One peak in incorporation occurs for dreams following the event by 1 day (the day-residue effect), and a

second peak occurs 6–7 days later (the dream-lag effect). Dream-lag incorporations (1) seem related to interpersonal problem-solving (interpersonal relationships, positive emotions, and resolved personal problems), (2) are most pronounced among individuals who report that their dreams influence waking thoughts and feelings, and (3) are at least compatible with evidence that REM deprivation during specific, delayed temporal windows disrupts REM-dependent memory consolidation. Documenting the complex processes by which presleep tasks and challenging events are incorporated into dreaming may help clarify the forms of memory reorganization and elaboration that support memory consolidation.

Emotional Adaptation

Studies of dream function often are couched in the language of problem solving. Usually, the proposal is that dreams incorporate aspects of a waking problem, the process of dream formation actively reorganizes material related to that problem, and the dream effect is articulation of a novel solution to the problem, either within the dream or during immediate waking reflection upon the dream images. Whereas research supports the notion that REM dreaming, as a form of fluid thinking that briefly persists after awakening, contributes to the formulation of solutions that are not characteristic of waking thought, generalization from transient nonemotional problems (e.g., anagrams) to enduring emotional problems (e.g., bereavement) is common but problematic. Research procedures suited to the investigation of short-term dream effects cannot simply be extended to studies of the influence of dreaming on long-term emotional well-being. Longitudinal (or quasi-longitudinal) research designs are required.

Within longitudinal designs, the contribution of REM sleep research protocols continues to be significant for two reasons. First, laboratory studies of the immediate effects of REM dreaming often guide articulation of the hypotheses examined in longitudinal studies. Second, sleep laboratory procedures remain crucial when the research question concerns changes in dreaming during the course of the night, contrasts between REM dreams and NREM dreams, etc.

Mood Regulation

The hypothesis that REM sleep has a mood regulatory function is consistent with evidence of an overnight attenuation of either positive or negative mood intensity. An extension of this hypothesis is that nightmares reflect the failure of this regulatory mechanism to subdue disturbing emotions, a proposal that is reminiscent of Freud's suggestion that the function of dreaming is to protect sleep. A recent and modified version of the mood-regulation hypothesis has influenced studies of the role of REM dreaming in the alleviation of depression.

REM sleep and mood regulation during depression There is evidence that early-night and late-night REM sleep stages play different roles in mood regulation during depression. It is well established that depressed individuals show reduced latency to the first REM sleep episode of the night and an increase in REM density during the first REM sleep period, indicating a phase advance of the diurnal cycle. Also, during depression, REM density during the first REM period is negatively correlated with the attenuation of negative mood during the night, whereas REM density during the last REM period is positively correlated with the attenuation of negative mood. Finally, depressed individuals reporting more negative dreams at the beginning of sleep, and fewer negative dreams toward the end of sleep, are more likely to be in remission 1 year later than are those with fewer negative dreams at the beginning of the night and more negative dreams toward the end of the night. Taken together, these findings indicate that, among depressed individuals, the progressive attenuation of negative affect during successive REM periods and the associated attenuation of negative mood in the morning alleviate depression. While it is not clear what normally brings this about, REM deprivation is one intervention that seems to do so.

REM deprivation and depression When acutely depressed individuals (e.g., responding to divorce) undergo selective REM deprivation over several weeks, the phase advance of the diurnal cycle is reversed, leading to sustained remission from depressive symptoms. Preliminary evidence that dreaming is implicated in these improvements includes (1) the attenuation of negative mood on REM deprivation nights, (2) the maintenance of this attenuation during the following day, (3) the vividly dreamlike character of REM mentation early in the treatment period, and (4) dream reports that incorporate emotionally significant aspects of the life event that precipitated the depressive episode (e.g., the former spouse; the children). The preceding pattern is best described as emergent; its various components remain somewhat unstable from study to study. However, the prevalence of depression (including as a complement to other distressing life events, e.g., trauma) makes clarification of this pattern especially important.

This research is a reminder of the procedural complexity that is required in studies of long-term dream function. Among those who are depressed, it is demonstrably important to (1) differentiate the types of dreaming that occur early in the night from those that

occur late in the night, (2) examine the blending of and variations in dream themes between early-night and late-night REM periods, (3) monitor not only the immediate mood carryover effects, but also the waking events that support or obstruct stabilization of these mood changes during wakefulness, and (4) assess individual differences in the patterns of dreaming and dream effects.

Conflict Resolution

In recent years a number of investigators have focused on repeated themes, and their variations, in individual dream series. These repeated themes vary on a continuum, from the reappearance of particular images in a dream series, to the recurrence of analogous dream scenarios, to the repeated replication of overwhelming traumatic moments. In a variety of theories, these repeated themes are understood as indicating the presence of unrecognized and unresolved personal conflicts. Correspondingly, the repetition and variation of these themes are understood as fostering their recognition and resolution.

There is some evidence from research using quasi-longitudinal research designs that is consistent with this formulation. When adults who are presently experiencing recurrent dreams are compared with those who experienced recurrent dreams in the past and with those who have not experienced recurrent dreams, recurrent dreamers report dreams that are more anxious, dysphoric, and conflicted laden; they also have lower scores on several indices of personal well-being (e.g., anxiety, depression, life stress). Past recurrent dreamers report greater personal well-being, as well as more positive dream content, than do recurrent dreamers or nonrecurrent dreamers. These results have been replicated with younger adults among whom past recurrent dreamers were defined as those who had recurrent dreams during late childhood. In this case, past recurrent dreamers did not differ from nonrecurrent dreamers. These studies indicate that the cessation of adult – but not childhood – recurrent dreams is associated with enhanced well-being.

Vestigial Defense

There are two conceptions of dream function that posit their historical rather than current adaptive significance. In an ontogenetic version of this possibility, the capacity of REM dreaming to enhance the consolidation of memories for emotional events may fix traumatic memories in a way that is not adaptive in the dreamer's current life. A prototypic example occurs in posttraumatic stress disorder (PTSD), with its hyperemotional memories, flashbacks, and nightmares. Hypothetically, the enhanced consolidation of declarative memory for overwhelming events (through amygdalar modulation of hippocampal function) consolidates those memories so that they remain easily evoked long after the threat associated with trauma has passed (e.g., in posttraumatic nightmares). Generalization of this hypothesis to more mundane emotional memories may explain the repetition principle – that is, the persistence in dreams of a variety of recurrent themes that are unrelated to current concerns.

In an analogous evolutionary model, dreaming purportedly simulates threatening events of the kind that early humans were likely to encounter frequently (e.g., aggressive interaction with animals) but that contemporary humans typically do not. In hunter–gatherer societies, such dreamed threat simulations may have provided covert rehearsal of a defensive response to actual threatening circumstances, even though they have no contemporary ecological validity. So, when contemporary threats evoke dreams with those primordial characteristics, the threat simulation is likely to be unrelated to, and perhaps even incompatible with, adaptive responses. These conceptions of vestigial dream function may help to explain why dreams related to trauma, or even nightmares that do not have origins in identifiable trauma, often disrupt sleep, prompt hypervigilance, and create lingering distress.

Conclusion

It sometimes seems that theories of dream function are strongly conflicting – as though if one is valid then the others cannot be. However, since short-term dream effects may serve auxiliary functions within more obviously 'adaptive' activities, long-term dream function may be understood as the capacity of a complex self-organizing system that depends upon integration of numerous simpler capacities. From this perspective, the short-term functions of dreaming, when coordinated according to the system's superordinate integrative principles, may subserve a long-term function that is irreducible to any particular dream effect. This integrative possibility remains undeveloped but increasingly plausible.

See also: Dreams and Nightmares in PTSD; Dreams and Dreaming: Incorporation of Waking Events; Dreams, Dreaming Theories and Correlates of Nightmares; Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep-Dependent Memory Processing; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Blagrove M (1992) Dreams as the reflection of our waking concerns and abilities: A critique of the problem-solving paradigm in dream research. *Dreaming* 2: 205–220.

Cartwright R, Baehr E, Kirkby J, et al. (2003) REM sleep reduction, mood regulation and remission in untreated depression. *Psychiatry Research* 121: 159–167.

Cheyne JA, Rueffer SD, and Newby-Clark IR (1999) Hypnagogic and hypnopompic hallucinations during sleep paralysis: Neurological and cultural construction of the night-mare. *Consciousness and Cognition* 8: 319–337.

Greenberg R and Pearlman C (1993) An integrated approach to dream theory: Contributions from sleep research and clinical practice. In: Moffitt A, Kramer M, and Hoffman R (eds.) *Functions of Dreaming*, pp. 363–377. Albany: SUNY Press.

Kramer M (1993) The selective mood regulatory function of dreaming: An update and revision. In: Moffitt A, Kramer M, and Hoffman R (eds.) *Functions of Dreaming*, pp. 139–195. Albany: SUNY Press.

Kuiken D and Sikora S (1993) The impact of dreams on waking thoughts and feelings. In: Moffitt A, Kramer M, and Hoffman R (eds.) *Functions of Dreaming*, pp. 419–476. Albany: SUNY Press.

Maquet P, Laureys S, Peigneux P, et al. (2000) Experience-dependent changes in cerebral activation during human REM sleep. *Nature Neuroscience* 3: 831–836.

Rauchs G, Desgranges B, Foret J, et al. (2005) The relationships between memory systems and sleep stages. *Journal of Sleep Research* 14: 123–140.

Revonsuo A (2003) The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. In: Pace-Schott EF, Solms M, Blagrove M, et al. (eds.) *Sleep and Dreaming: Scientific Advances and Reconsiderations*, pp. 85–109. New York: Cambridge University Press.

Smith C (2001) Sleep states and memory processes in humans: Procedural versus declarative memory systems. *Sleep Medicine Review* 5: 491–506.

Stickgold R (2005) Sleep-dependent memory consolidation. *Nature* 437: 1272–1278.

Wagner U, Degirmenci M, Drosopoulos S, et al. (2005) Effects of cortisol suppression on sleep-associated consolidation of neutral and emotional memory. *Biological Psychiatry* 58: 885–893.

Dreams and Dreaming: Incorporation of Waking Events

E J Wamsley and R Stickgold, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

We all have the intuition that our dreams in some way relate to experiences from waking life. Even in cases in which no specific event presents itself as the likely source of a particular dream, still, our nightly reveries often involve places, persons, or activities which are familiar to us. Indeed, empirical investigations of the effects of presleep experience on dream content support the generally held presumption that mentation during sleep is constructed, at least in part, from memories of both recent and remote experiences. But what does this tell us? Despite decades of research effort, we still have surprisingly little understanding of the cognitive mechanisms controlling incorporation of waking events into sleep mentation, and virtually no understanding of the neural events underlying these processes.

What algorithms determine which recent and remote memories will appear in sleep mentation and what form incorporation of these memories will take? 'Does the incorporation of waking events into dreaming rely on the engagement of particular memory systems, to the exclusion of others? Is the appearance of recent events in dreaming related to neural processes facilitating memory processing during sleep? Might the reexpression of waking events in sleep mentation serve some adaptive function?' Answering these questions not only will provide us with a better understanding of the neural basis of the dreaming process, but may also provide us with insights into the nature of human cognition and memory processing in general.

The definition of 'dreaming' remains a contentious subject, and one that we do not address here. For our purposes, we refer to 'dreaming' as being all sleep mentation of which an individual is aware either during sleep or upon awakening, and we refer to a 'dream' as being the content of that mentation. In addition, we will sometimes use the term 'dream' to refer to the waking report of dream content, and 'dream content' to refer to the content of such reports. We do so only for the sake of clarity in reading, and not to stake out a particular position in regard to the nature of dreaming.

Effects of Presleep Experience on Dream Content

Manipulation of Presleep Experience in the Sleep Laboratory

During the 1960s and 1970s, a considerable amount of research effort was devoted to understanding the relation of waking events to dreaming through laboratory studies manipulating participants' presleep experience. Unfortunately, much of this research was plagued by methodological weaknesses, including frequent failure to employ blind raters and/or reliance on subjective interpretation of symbolic relationships between waking experience and dream content. Overall, the most consistently observed direct effect of waking experience on laboratory-collected dream reports to emerge from these studies was the powerful influence of the laboratory setting on dream content. In an analysis of 813 rapid eye movement (REM) dream reports collected across several studies, Dement and co-workers reported that 22% of reports unambiguously incorporated either isolated elements (e.g., the experimenter, the sleep laboratory itself, electrodes) of the laboratory situation (in 10% of the reports) or more complete representations (e.g., a combination of the previously mentioned elements) of the experimental setting (in 12%).

In contrast to the strong influence of the laboratory setting, experimental manipulations of the presleep experience have often failed to demonstrate clear effects on dream content. In fact, the most striking outcome of such investigations has been the conspicuously rare appearance of content directly related to a particular presleep intervention. Even the use of emotionally salient stimuli (e.g., violent films) or the manipulation of basic biological drives (e.g., the induction of thirst) has rarely resulted in dream content obviously related to the manipulation. In studies in which unambiguous incorporations of experimental stimuli have been reported, the influence of the laboratory environment often greatly overshadows these effects.

In retrospect, this is hardly surprising, given the relative salience of sleeping in a strange place and being awakened during the night to report dreams. The salience of any particular film, activity, etc. introduced to individuals as a part of an experimental protocol likely pales in comparison. At the same time, these findings suggest that emotional salience may be a major determinant of dream content. This conclusion must be

tempered, however, by the fact that the individuals were receiving sensory input from the laboratory environment (e.g., from the recording electrodes) even as they slept, and such concurrent sensations might have dictated their incorporation.

Well-designed experimental investigations utilizing appropriate control groups have had relatively more success in demonstrating statistically significant effects of experimental manipulations when formal or qualitative dream features (e.g., length, affect), rather than content of dreams *per se*, were used as the outcome measure. In short, it has proved to be difficult, if not impossible, to reliably manipulate the content of dreams through presleep interventions administered in a laboratory setting.

Correspondence between Presleep Experience and Dreaming in the Home

Other, nonexperimental approaches, however, clearly show that waking experience is a primary source of sleep mentation content. In an analysis of the possible waking memory sources of 299 home-collected dream reports, Magdalena Fosse and colleagues demonstrated that fragments of recent experience are often seen in dream reports. Of 299 reports, 51% analyzed in this study were judged by individuals to contain at least one feature bearing strong similarity to a recent waking event. In this particular study, participant-identified incorporations were typically direct in nature. As in the case of the aforementioned experimental studies, however, naturally experienced waking events appear to sometimes affect qualitative dream features, as opposed to resulting in direct incorporations of the experience. In an analysis of home dreams collected in the weeks prior to and following the terrorist attacks of 11 September 2001, Ernst Hartmann reported that, although the events of September 11 were not typically incorporated into participants' dreams in the period following the attacks, the emotional intensity of central dream images was significantly increased.

Other lines of research have explored the more general correspondence between sleep mentation and waking life. Calvin Hall, an early cognitive dream theorist, pioneered the use of content analysis to quantitatively assess the content of large sets of mentation reports. This classification system, developed in collaboration with Robert Van de Castle as a response to subjective interpretations of dream content generally employed by psychoanalytic dream theorists, counts the occurrences of various categories of characters, settings, objects, social interactions, etc., as explicitly described in the text of dream reports. Using this system, Hall was able to create surprisingly accurate psychological profiles and histories for psychiatric patients, based solely on blind content analysis of their dream reports. Later work employing this same system of quantifying dream content has reported consistent, statistically significant differences in the dream content of groups of individuals with divergent waking experience (e.g., between males and females, children and adults, and blind and sighted individuals). These investigations have contributed to our understanding of the dreaming process particularly inasmuch as they remind us that, while attempts to predict or control the waking source of specific dream elements have often failed, a broad correspondence between dream content and waking experience is nonetheless transparently obvious. In short, people clearly tend to dream of the same persons, activities, etc. which dominate their waking lives. While psychoanalysts have long presumed that dreaming conceals a hidden message demanding the application of a complex 'decoding' process, this research, on the contrary, suggests that the relationship between dream content and waking experience may be, if not straightforward, at least discoverable without resorting to symbolic interpretation or reference to hidden 'latent' content.

Incorporation of Presleep Experience into Sleep Onset Dreaming

In contrast to the relative difficulty of manipulating dream content later in the night, experimentally administered tasks appear to more clearly affect mentation during the first few minutes of sleep. In a series of studies employing engaging video games as a presleep task, Stickgold has demonstrated a powerful influence of presleep experience on dreaming at sleep onset. In one such study, participants played the video game Tetris in the morning and evening, several hours prior to retiring for the night. Dream reports were then repeatedly elicited from each participant after just 30 s to 3 min of sleep (in Tetris, the gamer must manipulate falling block shapes and fit them together; problematic subjective interpretations of dream imagery were not necessary in this study, as gamers reported clearly related content elements, such as "Tetris shapes" and "tiles fall[ing] down"). All told, 64% of participants reported isolated, unambiguous, Tetris-related images in at least one sleep onset report. In a related investigation utilizing a downhill skiing arcade game, a surprising 88% of participants reported direct incorporation of game features into at least one mentation report collected during the 3-day study. The frequency of direct incorporation in these sleep onset studies, which is dramatically higher than that observed in any previous overnight investigation, suggests that the first minutes of sleep represent the ideal conditions for reexpression of waking experience in dreams. The instance of direct Alpine Racer incorporation, however, is greatly decreased when reports

are collected from a delayed sleep onset period following 2 h of sleep, even when these report collections are matched for time of night with initial sleep onset controls. It may be, then, that direct incorporation of these experiences into initial sleep onset reports is dependent upon the position of the report collection period within the sleep phase, rather than on either simple proximity to waking events or general neurobiological features of the sleep onset state.

Taken together, these data strongly support the common presumption that waking experiences influence dreaming both at sleep onset and later in the night. Yet, the question of 'whether or not' waking experience affects dream content may not be a particularly meaningful one. While empirical verification of the folk psychology presumption that dreams incorporate experience-related content is a necessary step, its confirmation is hardly a revelation, and, more importantly, it tells us little about the neural processes which create dreaming, nor about a possible function for the dreaming process. Instead, a more detailed examination of the nature of the incorporation waking experience, grounded in neurobiological data regarding the engagement of memory systems across states of consciousness, is needed if we are to gain clues to the neural and cognitive processes supporting incorporation of waking experience into sleep mentation.

Form of Waking Experience Incorporations

Even in cases in which mentation reports clearly do include elements originating from a specific spatiotemporal episode, these dreams almost never take the form of an exact replay of a life event. Instead, sleep mentation typically incorporates only isolated elements, or fragments, of a waking episode. These fragments, in turn, may be interleaved with remote and semantic memory material. For example, although Fosse's group found that more than half of the reports in their sample contained individual features with strong similarity to a recent experience, less than 2% of these preserved all of the key elements of the original episode, including the setting, characters, objects, and actions found in the memory of the waking event. Mentation reports following Tetris and Alpine Racer exposure exhibit similar characteristics, with individuals describing seeing clearly identifiable isolated features of the game (e.g., Tetris shapes) without the larger context of the activity (e.g., the computer screen, keyboard, or desk).

The following report from Fosse's group provides an illustrative example of a waking memory source and corresponding report, in which the sleep mentation clearly incorporates fragments of a waking experience, without replicating the original context in which these fragments were embedded:

> Waking Experience: "*When I left Starbucks, we had so many leftover pastries and* ***muffins*** *to throw away or take home.* ***I couldn't decide which muffins to take*** *and which to toss…*"
> Corresponding Sleep Mentation: "*My dad and I leave to go shopping. We go from room to room, store to store. One of the stores is filled with* ***muffins, muffins, muffins*** *from floor to ceiling, all different kinds,* ***I can't decide which one I want****…*"

The various components of a waking episode, then, do not appear to be 'bound' together in the manner that characterizes the 'mental time travel' (as described by psychologist Endel Tulving) of episodic memory recall in waking life. This phenomenological disparity between the appearance of memories in dreams and the recall of recent episodes in waking life suggests that divergent, though likely overlapping, processes underlie these two phenomena. We consider in the following discussion how the fragmented appearance of recent memory in dream reports, interleaved with remote and semantic content, might reflect the consolidation of recent memories during sleep as they become integrated into older cortical memory networks.

Note that it may be precisely the fragmentary nature of event incorporation which frustrated the efforts of so many early presleep-stimulus studies. Investigators have often, for example, exposed study participants to multidimensional film stimuli prior to sleep, expecting to see veridical 'replay' of this stimulus in subsequent mentation reports. The stimuli employed might contain a number of relatively well-learned features (e.g., persons or commonplace objects and activities). If only isolated features of these experiences appeared later in dream reports, their ubiquitous nature may have prevented the recognition as having originated in the presleep manipulation of interest. In contrast, the novel and easily identifiable features of Tetris, for example, provide the optimal conditions for tracking isolated features of waking events in subsequent dreams.

It may be important to note also that, as in the preceding example, isolated dream elements originating from the identified/manipulated presleep experience clearly appear in combination with additional elements derived either from other memory sources or from generalized semantic memory networks (e.g., 'shopping with dad' in the preceding example). Indeed, although participants often cite relatively recent experiences as the source of their mentation content, remote memories from several years or longer ago, as well as generic semantic memories, are

also frequently cited as the mnemonic referent of dream material. In sum, the incorporation of recent experiences into dreams is manifested in the intrusion of isolated, salient features of a presleep experience, intermingled with other content.

Sleep Stages, Waking Memory Sources, and Memory Systems

There appear also to be systematic differences across various sleep states in the types of memories incorporated into dream content. This distribution of memory sources, in turn, may inform us about the engagement of particular memory systems across sleep stages and their participation in the dream creation process. In a series of studies, primarily conducted by researchers at the University of Bologna, Italy, participants identified possible waking memory sources of mentation collected from non-REM (NREM), REM, and sleep onset. Blind raters classified each participant-identified memory source as either 'episodic' (autobiographical episodes taking place in a specific spatiotemporal context) or 'semantic' (elements of general knowledge not tied to a particular spatiotemporal context). Perhaps the most interesting result emerging from this work is that NREM reports, rather than REM reports, are associated with the highest proportion of episodic memory sources. This observation fits well with what we know about regional cerebral blood flow (rCBF) patterns during sleep. Acquisition and retrieval of recent episodic memories rely crucially on the engagement of the hippocampal formation, which is more active during NREM than in either waking or REM sleep. In addition, the low levels of acetylcholine characterizing NREM sleep are hypothesized to preferentially facilitate output from the hippocampus to the cortex.

Consistent with the data of Stickgold, however, reports at sleep onset tend to contain an even greater proportion of episodic memory sources than do NREM reports, as well as to contain a very high proportion of 'day residues' (episodic memory sources originating from an experience during the previous day). Comparatively little is known about the activity of the hippocampal system at sleep onset, although at least one investigation has reported that rCBF in the hippocampus is not significantly decreased at sleep onset relative to waking. However, there is cognitive evidence that the hippocampus is not required for incorporation of recent experience into sleep onset mentation (see later).

When participants, returning to sleep following awakening after the first complete REM cycle, provide dream reports from a delayed sleep onset period, episodic sources in general are still proportionally greater than in either REM or NREM reports. But 'day residues' are significantly less frequent. Though proximity to waking experience has been postulated to account for this decline in incorporation of very recent events into later sleep onset periods, recall that Stickgold et al. observed a similar decline in Alpine Racer game-related content when mentation was collected from a delayed sleep onset phase, even though time of night and time since last game play were both kept constant. Again, this suggests that the disproportionate appearance of recent memories in sleep onset reports is not a function of simple proximity to the daytime experience. Rather, memory processing occurring during the initial hours of sleep may alter the incorporation of recent experiences into subsequent dreams, such as into delayed, as compared to initial, sleep onset periods.

Emotional Value and Waking Experience Incorporation

But what processes determine 'which' memory fragments will appear in subsequent sleep mentation? Though doubtless others operate in tandem, the emotional value of waking experiences appears to be one relevant factor. Analyses of the correspondence between dreamed and waking activities, for example, demonstrate that while some relatively mundane activities consuming a great deal of our waking time (e.g., reading and writing) are underrepresented in dreams, other activities of greater personal significance (e.g., socializing with friends) are overrepresented relative to their frequency during waking. Other lines of research suggest a direct link between the emotional value of specific waking events and their appearance in dream reports. Rosalind Cartwright's work on the adaptive function of dreaming in women experiencing divorce directly addresses personal relevance as a mediator of the incorporation of waking experiences into dream content. Cartwright found that individuals' self-rated concern about their spouse is strongly correlated with the percentage of dream reports containing explicit mention of the spouse. Meanwhile, DeKoninck and Koulack have reported similar findings in a group of individuals viewing a film of work-related accidents prior to sleep. Those individuals reporting increased levels of anxiety following morning film viewing reported significantly more film-related content in their subsequent dream reports. Each of these studies suggests that degree of emotional involvement in waking experiences modulates their incorporation into subsequent dreams.

Recall also that in laboratory studies of the effects of presleep experience, the experimental situation was often more strongly incorporated into mentation reports than any was particular experimentally manipulated stimulus. Presumably, in these situations, the

experience of sleeping in a strange place, being wired with electrodes, and participating in an experiment is far more salient and more important than is any specific presleep experience (e.g., a film or other activity) of interest to the investigators. Again, this general set of findings indicates that those experiences in which an individual is more emotionally involved are more likely to appear in subsequent dreams. We also see this effect dramatically, of course, in the dreams of trauma survivors, who tend to replay the traumatic emotional event in their dreams to such a degree that the disturbing dreams themselves become a clinically significant symptom of posttraumatic stress disorder (PTSD).

Time Course of Incorporations

Incorporation of recent experiences into dream content may follow an organized temporal pattern, examination of which could provide clues to a possible function of dreaming. First of all, very recent events may more often be incorporated into mentation occurring early in the sleep phase. Though only a small handful of studies have examined the effect of time of night on such incorporations, both recency of subject-identified memory sources and the similarity of sleep mentation reports to presleep thoughts have been reported to correlate negatively with time since sleep onset. Meanwhile, Stickgold has reported that, during the first few minutes of sleep, incorporation of Tetris imagery into sleep onset dreams declines linearly as a function of time since sleep onset. Recall also that, overall, very recent experiences are more often incorporated into initial sleep onset mentation than into delayed sleep onset mentation or dream reports later in the night. As discussed earlier, simple proximity to waking events appears to be inadequate to explain these effects, as incorporation of recent events declines later in the sleep cycle, even when time of night is held constant.

Examining the incorporation of waking experience into sleep mentation across longer timescales also suggests that mere proximity to a waking event is a relatively poor predictor of incorporation. Although some studies conducted across multiple nights have reported incorporations occurring most frequently on the night immediately following a presleep experience, others have found that frequency of incorporations peaks only after a delay of $\geq$1 day. Stickgold, for example, found that sleep onset reports of novice game players contained far more Tetris imagery on the second night following initiation of game play than on the first. Cartwright has similarly reported increased incorporation of a sexually arousing film on the second night following viewing the film, though these incorporations were most often 'symbolic' in nature. Roffwarg and colleagues have reported even longer incorporation delays, as in the case of the influence of extended wearing of colored goggles on the perceptual quality of dreams. Here, the incorporation effect was seen most strongly several days after participants began wearing the goggles.

In addition, Tore Nielsen has conducted a series of careful studies that find two distinct postevent time frames in which incorporation is most likely to be identified. In accordance with the findings of several Roffwarg and colleagues, the first of these is on the night immediately following a presleep experience (the 'day residue' effect). Nielsen also finds evidence, however, for a 'dream lag' effect, in which waking experiences are again likely to be incorporated into dreams recalled approximately 7 days after the initial experience. Though broadly consistent with the long incorporation delays reported by Roffwarg et al., this is a relatively novel finding and the cognitive and neurobiological mechanisms that might underlie it remain unknown. Nonetheless, this 'dream lag' effect has remained relatively robust across several studies, and controls have ruled out a number of potential confounds.

On an even broader timescale, the temporal distribution of participant-identified dream memory sources is quite similar to that of experimentally elicited waking autobiographical memory samples. In each case, the largest proportion of memory references is drawn from more recent experiences, with references to more remote memories declining in concert with increased temporal distance. This observation is consistent with the idea that some overlap exists between the neurobiological basis for waking recall of autobiographical episodes and their appearance in dream content, and, contrary to Freudian theories, dreams are relatively infrequently associated with memories from the distant past.

Recent Memories, Dreaming, and Memory Consolidation

A growing body of evidence suggests that sleep, and in particular, NREM sleep, facilitates the offline processing of recent declarative memories (e.g., memories of events and facts). Behavioral studies in humans, for example, suggest that periods of NREM sleep are beneficial for declarative memory performance, relative to periods of wakefulness or REM sleep. Meanwhile, both human and animal studies demonstrate that patterns of neural activity associated with recent hippocampally dependent learning are reexpressed during subsequent NREM sleep. This neural 'replay' of recent episodes during NREM sleep is thought to support consolidation processes that mediate the

transformation of recent memories from an initial hippocampus-dependent labile form into more permanent long-term memories in the cortex.

Naturally, the observation that recent declarative memories are reexpressed in the sleeping brain has prompted speculation that these processes may be related to the content of dreams. Although very little work has directly addressed the possible relationship between dreaming and sleep-dependent memory consolidation, a small group of studies do support the notion that such memory processing is reflected in the content of sleep mentation. For example, Carlyle Smith reported preliminary evidence of dream reports related to a mirror-tracing task learned prior to sleep. These task-related dreams were present to a greater degree in individuals 'cued' during sleep using a sound associated with the learning task, compared to noncued individuals. Fiss has provided evidence of a link between overnight changes in memory performance and dreaming, reporting that dreaming of stories presented to individuals prior to sleep was associated with enhanced morning recall of those stories. Thus, the available cognitive evidence is at least consistent with the notion that sleep mentation may reflect consolidation of memories in sleep.

Neural Basis for Dream Incorporation of Recent Events: Connection to Declarative Memory Consolidation?

Research into the neurophysiological basis of sleep-dependent memory consolidation may shed light on neural mechanisms mediating the relationship between recent experience and dreaming. Models of memory consolidation are grounded in evidence that, although declarative memory recall is initially dependent on the hippocampus, over time, repeated memory trace reactivation enables the cortex to support retrieval without the need of hippocampal involvement. Repeated memory reactivation is thought to be crucial to the process of memory consolidation, and, as noted previously, evidence suggests that this occurs optimally during NREM.

Supporting this general model, a number of studies have demonstrated that neural activity initially established during waking experience is 'replayed' during subsequent sleep. These data primarily derive from studies in rodents utilizing multiple single-neuron recordings. Analogous observations have been reported in humans, however, as patterns of regional brain activation associated with specific learning experiences in wakefulness are again seen during subsequent NREM sleep.

Several key features of the replay of learning-associated neural activity in sleep exhibit interesting parallels with the preceding cognitive data on sleep mentation:

1. The neural replay of recent experiences has most often been reported to occur in NREM sleep. As described earlier, NREM is also the sleep stage during which episodic memories are most often incorporated into sleep mentation.
2. Within NREM sleep, neural replay of recent experience is expressed most strongly immediately after learning. The strength of this reactivation effect then tends to decay quickly across time. Similarly, as already described, sleep mentation may be most strongly related to recent experience early in the sleep cycle.
3. Recent learning is not 'replayed' during sleep in its original form. Following learning, firing sequences established during wakefulness are reexpressed only intermittently during rodent NREM, with relatively low fidelity, and on a faster timescale than the original experience. Similarly, as already described, recent experiences are not 'replayed' in dream content in their original form. Instead, intermittent fragments of recent experience are recalled from sleep, intermingled with other content.

Although far from conclusive, this correspondence between features of neuronal-level experience replay during sleep and dream content is consistent with the hypothesis that subjective experience is derived, at least in part, from memory reactivation supporting consolidation processes in sleep. Though these generalizations are largely based on neural replay of recent events observed in the hippocampus, it is important to note that reexpression of learning-related activity has also been observed in cortical structures. Matt Wilson, for example, recently observed sequential replay of a recent experience in the rodent primary visual cortex. Though the inference that an observed pattern of brain activity is associated with conscious awareness always presents a problematic conceptual leap, experimental stimulation of visual cortex in humans does result in associated visual experience.

Despite these arguments, it must be pointed out that little evidence directly links subjective experience recalled from sleep to neurophysiological and cognitive data on memory replay and consolidation. In fact, the study of Tetris imagery in sleep onset dreaming suggests that, at least at sleep onset, this incorporation likely does not rely on neural processes involved in sleep-dependent declarative memory consolidation. Replay and consolidation of recent

declarative memory during sleep is presumed to depend crucially on involvement of the hippocampus. Remarkably, however, Stickgold found that amnesiac patients with complete loss of hippocampal function incorporated Tetris images into their sleep onset dream reports at rates similar to normal persons. These images appeared even though the patients could neither identify the source of the Tetris imagery nor recall ever having played the game. This observation suggests that, at least at sleep onset, dream incorporation of recent experiences does not rely on the hippocampal processes presumed to underlie consolidation of declarative memory. It seems likely that this is exclusively a sleep onset phenomenon, as sleep onset is a state with distinct neurophysiological and phenomenological features. Alternatively, as described in early studies of dreaming in hippocampal patients, there may be subtle differences in the quality of memory content seen in the dreams of amnesiacs. Certainly, more research is needed before a definitive statement can be made about the relationship between dreaming and memory consolidation in sleep.

A Function for Dreaming?

To the extent that dreaming does turn out to reflect memory consolidation processes, neural activity supporting the dreaming process may serve an important adaptive function. Presumably, the function of memory consolidation, in sleep or in wakefulness, is not only to facilitate long-term retention of information, but also to identify those memories of recent events worth retaining and then to integrate these with memories of related older experiences and with general semantic information, so that this new knowledge can be most adaptively utilized to guide future action. It may be that dreaming is associated both with the identification of valuable memories and with the integration of these memories with remote and semantic information stored in cortical networks.

Supposing that dreaming is indeed associated with such a function, might there be cases in which this function is disrupted? Although we have characterized the incorporation of recent experiences into dream content as a fragmentary process, with elements of the waking event interleaved with remote and semantic content, there are cases in which more veridical, complete replay of experiences is reported. In patients suffering from PTSD, the traumatic experience is often replayed in an abnormally direct manner, and the patients' frequent, near-veridical replay is a persistent symptom of the disorder, which can continue for years. It is conceivable that this abnormal replay of traumatic experiences in the dreams of PTSD patients is reflective of the dysfunctional consolidation of the traumatic memories, perhaps involving a failure to successfully integrate the traumatic experience with existing cortical memory networks.

In closing, it is important to remember that despite this or any speculation on a possible 'function' for dreaming, the question of whether the conscious experience of dreaming *per se*, as opposed to the neural activity associated with it, supports these processes is essentially unapproachable until more is understood about the neural basis of consciousness in general.

See also: Amygdala: Contributions to Fear; Dream Function; Dreams and Nightmares in PTSD; Dreams, Dreaming Theories and Correlates of Nightmares; Posttraumatic Stress Disorder: Overview; Posttraumatic Stress Disorder as an Emotional Disorder; Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Research and Sleep Medicine in Historical Perspective; Sleep-Dependent Memory Processing; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Baylor GW and Cavallero C (2001) Memory sources associated with REM and NREM dream reports throughout the night: A new look at the data. *Sleep* 24(2): 165–170.

Buzsaki G (1996) The hippocampo–neocortical dialogue. *Cerebral Cortex* 6(2): 81–92.

Cartwright R, Agargun MY, Kirkby J, et al. (2006) Relation of dreams to waking concerns. *Psychiatry Research* 141(3): 261–270.

De Koninck JM and Koulack D (1975) Dream content and adaptation to a stressful situation. *Journal of Abnormal Psychology* 84(3): 250–260.

De Koninck J and Brunette R (1991) Presleep suggestion related to a phobic object: Successful manipulation of reported dream affect. *Journal of General Psychology* 118(3): 185–200.

Dement WC, Kahn E, and Roffwarg HP (1965) The influence of the laboratory situation on the dreams of the experimental subject. *Journal of Nervous and Mental Disease* 140: 119–131.

Fiss H, Kremer E, and Lichtman J (1977) The mnemonic function of dreaming. *Sleep Research* 6: 122–136.

Fosse MJ, Fosse R, Hobson JA, et al. (2003) Dreaming and episodic memory: A functional dissociation? *Journal of Cognitive Neuroscience* 15(1): 1–9.

Grenier J, Cappeliez P, St-Onge M, et al. (2005) Temporal references in dreams and autobiographical memory. *Memory & Cognition* 33(2): 280–288.

Hall C and Nordby V (1972) *The Individual and His Dreams*. New York: New American Library.

Hartmann E (2000) We do not dream of the 3 R's: Implications for the nature of dream mentation. *Dreaming* 10: 103–110.

Hasselmo ME (1999) Neuromodulation: Acetylcholine and memory consolidation. *Trends in Cognitive Science* 3(9): 351–359.

McClelland JL, McNaughton BL, and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102(3): 419–457.

Nielsen TA, Kuiken D, Alain G, et al. (2004) Immediate and delayed incorporations of events into dreams: Further replication and implications for dream function. *Journal of Sleep Research* 13(4): 327–336.

Stickgold R, Malia A, Maguire D, et al. (2000) Replaying the game: Hypnagogic images in normals and amnesics. *Science* 290 (5490): 350–353.

Dreams and Nightmares in PTSD

A Germain, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
A Zadra, Université de Montréal, Montréal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Clinical Significance of Nightmares in Posttraumatic Stress Disorder

Posttraumatic stress disorder (PTSD) refers to symptoms of reexperiencing, avoidance, and hyperarousal that persist for more than 1 month after exposure to a traumatic event. Violent crimes, including rape, physical assaults, and combat exposure, constitute traumatic events that involve threat to integrity of the self or to others and are accompanied by intense fear, helplessness, or horror. Epidemiological studies indicate that community prevalence estimates of PTSD are in the range of 1–10%, with higher estimates reported for victims of interpersonal violence (20–30%) and combat veterans (15–30%). Recommended first-line treatments of PTSD include selective serotonin reuptake inhibitors (SSRIs) and cognitive-behavioral interventions such as prolonged exposure and cognitive restructuring.

Nightmares are a core feature of PTSD. As many as 90% of trauma-exposed individuals who develop PTSD report disturbing dreams that bear varying degrees of resemblance to the actual traumatic event. Growing evidence suggests that trauma-related nightmares may directly contribute to the pathophysiology of PTSD. The persistence of nightmares and sleep disruption 1 month posttrauma increases the likelihood of developing PTSD at follow-up assessments. Moreover, a personal history of nightmares prior to trauma exposure can predict the severity of PTSD and other posttraumatic psychiatric symptoms. Frequent recurrent nightmares about a traumatic event can persist for years and even decades after the trauma. Trauma-related nightmares disrupt sleep and independently contribute to increased severity of daytime PTSD symptoms and increased alcohol use in trauma survivors, and adversely affect quality of life and perceived physical health.

Nightmares and Sleep following Trauma Exposure and in Chronic PTSD

In healthy persons, elaborate dream recall is more prominent following awakening from rapid eye movement (REM) sleep than from awakening from non-REM (NREM) sleep. In trauma-exposed adults, two preliminary studies suggest that disruption of REM sleep, as indicated by increased sleep stage transitions and arousal, as well as increased sympathovagal tone during REM sleep are associated with increased PTSD symptom severity and may increase the likelihood of developing PTSD at follow-up assessments. Thus, REM sleep disruption and increased arousal during REM sleep may increase one's vulnerability to experience vivid nightmares following trauma exposure and to develop PTSD.

REM sleep disturbances have been most consistently reported in patients suffering from chronic PTSD, but other sleep alterations, such as reduction in slow-wave sleep, have also been documented. More general sleep disturbances, such as increased sleep latency, total sleep time, and duration of and number of nocturnal awakenings, have been reported in some, but not all, studies. Polysomnographic studies based on refined objective measures of central and peripheral physiological arousal, such as quantitative electroencephalography and quantitative electrocardiography, have observed indices of heightened central and peripheral arousal during sleep in patients suffering from PTSD. PTSD in patients who report trauma-related nightmares show increased duration of wakefulness after sleep onset in comparison to those who report non-trauma-related nightmares. However, REM sleep parameters, or more global sleep measures, do not correlate with specific aspects of nightmare content. These observations suggest that in chronic PTSD, trauma-related nightmares may represent expressions of a more global sleep disturbance related to heightened arousal.

Dream Recall and Nightmare-Related PTSD

Studies on dream recall frequency following trauma exposure have yielded inconsistent results. Increases in dream recall frequency following a traumatic event have been reported and the hypothesis put forth is that dream recall is facilitated by the intensity, affective charge, and references to the victims' trauma depicted in their dreams. Conversely, significant decreases in dream recall have also been documented. These data, however, are largely based on retrospective assessments conducted many months or even several years following trauma exposure. It is possible that dream recall increases immediately following an individual's exposure to trauma and subsequently diminishes over time. One comparison of high-functioning Holocaust survivors, average-functioning Holocaust survivors, and controls found that the high-functioning survivors reported fewer dreams from REM sleep (33.7%), compared to the average-functioning survivors (50.5%)

and controls (80.8%). A reduction in dream recall may thus reflect improved posttrauma adaptation in this population.

Repetition of Trauma in Dreams

The recall of dreams that replicate the traumatic event experienced by trauma victims has been described in many populations, including war veterans, adults and children exposed to war, witnesses of violence and abuse, burn victims, and individuals exposed to natural disasters. A classification system for trauma-related dreams has been put forth. Posttraumatic dreams refer to the replication of the traumatic event during sleep or the introduction of the encapsulated memory of the trauma during sleep. Modified dreams present distorted themes, elements, or emotions from the trauma. Finally, disguised dreams represent the traumatic event symbolically or metaphorically. These categories form a continuum of replication (according to the extent to which the traumatic event is represented in the dream) and of repetition (i.e., the dream's recurrence). Empirical studies support this classification system. For example, in one study of Vietnam War veterans, 304 of 316 veterans reported combat-related nightmares. An examination of their dream narratives revealed that over half of the veterans reported realistic combat dreams, 21% reported plausible war sequences that they nevertheless had not experienced, and 26% reported dreams that alluded to the war, but also included fantastical and everyday elements. In addition, a laboratory study of war veterans found that only 21% of dreams exactly replicated the traumatic event, while the majority contained distortions related to the traumatic event. Similarly, one study of patients hospitalized following an accident or assault found that 46% of reported dreams replicated the traumatic event, 33% were dissimilar to the traumatic event while containing high levels of distress, and 19% showed little similarity to the trauma or distress.

Taken together, the results reveal changes in dream content with the passage of time and with improvement in the posttrauma reaction. Dreams recalled during the initial posttrauma phase tend to include some type of repetition of the trauma while subsequent phases are characterized by more symbolic representations of the trauma and greater integration of the individual's everyday life in recalled dream content.

Trauma and Dream Content

Several studies have reported salient alterations in the content of trauma victims' dreams. Some data suggest that when compared to the dream reports from controls, dreams recalled by trauma victims tend to be more ordinary and realistic and less salient, bizarre, and imaginative. The 'ordinary' quality of trauma victims' dreams may reflect an alteration in the process of dream construction or may be related to a protective mechanism which keeps traumatic images from intruding into everyday dreams.

Intense emotions have been documented in several studies of trauma victims' dreams. Feelings of horror, anxiety, rage, sadness, and frustration are among the most frequently reported. PTSD patients rate their nightmares as being significantly more distressing than do non-PTSD patients. Emotional preoccupations prior to bedtime may also play a role in the affective content experienced during subsequent sleep. The finding that one's mood prior to sleep is inversely related to positively or negatively experienced dream affect supports the idea that dreams may help regulate emotional equilibrium in traumatized individuals.

Some research has been directed at delineating relationships between measures of dream content and specific types of traumas.

Victims of Physical and Sexual Abuse

Themes of attack, pursuit, and of one's own death are more frequently found in the dreams recalled by victims of sexual abuse than in those recalled by control participants. Episodes of verbal and physical abuse are more evident in the dreams of physically abused women when compared to the dreams of controls. Although themes of violence and aggression often characterize women's nightmares, the nightmares reported by victims of sexual abuse are more likely to include unique themes of blood and dismemberment as well as a higher occurrence of male strangers. In addition, the dreams and nightmares of sexually abused women contain more frequent references to negative sexual activity, including themes concerning lack of trust and shame, guilt, jealousy, anger, and violence. These dreams often depict situations wherein elements of sex and aggression are confounded, in which sexuality is unpredictable and often results in fear. Victims of sexual abuse also frequently report the presence of serpents and worms in their dreams as well as references to both sexual and nonsexual body parts.

Victims of War and Violence

Exposure to an environment of war or violence is associated with dreams containing high proportions of aggressive and hostile human interactions. Themes involving immediate dangers to one's life appear more frequently in the dreams of Holocaust survivors, children that have been kidnapped, and children involved in car accidents than in nontrauma populations. It is not unusual for war veterans to report that their

traumatic dreams entail being killed in the place of their compatriots. A study of war-exposed Palestinian children found that traumatized children had more dreams with threatening strangers. Their dreams were often composed of attacks, anxiety, persecution, hostility, and nondesirable endings. In addition, nontraumatized children report more dreams of school and of their peers than do traumatized children.

That said, conflicting results exist regarding the impact of trauma and PTSD on dream content. For example, in addition to having more dreams with threatening strangers, victims of violence also report more dreams referring to their family, house, affiliations, and human connections than do controls.

Frequency of Dream-Related Disorders

Dream-related disorders, such as posttraumatic dreams, nightmares, bad dreams, and recurrent dreams, are the most frequently reported and most persistent symptoms exhibited by trauma victims. Several factors appear to mediate this association. First, a positive relationship exists between the degree and severity of trauma exposure and the frequency of dream-related disorders. Second, the victim's psychological reaction to trauma (e.g., level of anxiety) positively influences the occurrence of nightmares and other dysphoric dream experiences. Third, the time elapsed since trauma exposure appears to play an important role, with reductions in dream-related disorders occurring over time. However, it should be noted that dream disturbances can nevertheless persist for many years following the trauma.

Functional Hypotheses of PTSD-Related Nightmares

Many contemporary dream theorists suggest that dreaming is functionally significant and may subserve a biologically important function, but some argue that dreams are epiphenomenal to neurophysiological activity during REM sleep and have no value in and of themselves. While this overarching question on the function of dreaming is of phenomenological interest, the resolution of this debate is of little clinical relevance for chronic, frequent, and distressing nightmares in PTSD patients. However, understanding the possible role(s) of nightmares as one of the normal initial reactions to trauma exposure, which may also become a perpetuating condition contributing to maladaptive, chronic trauma response, has direct clinical implications regarding the development and refinement of effective nightmare-specific treatments. The following sections describe the main hypotheses put forth to explain the possible dual role of nightmares occurring posttrauma.

Nightmares and Emotional Adaptation to Trauma

Several neuropsychophysiological hypotheses of the function of dreaming have been adapted to explain the occurrence of nightmares after exposure to a trauma or other significant life events, and the role of nightmares in the pathogenesis and maintenance of PTSD. Freud proposed that nightmares reflect attempts to master anxiety and guilt associated with a traumatic experience, and to integrate traumatic experiences into one's psyche during sleep. Contemporary hypotheses are based on a similar premise – that dreaming serves a function of emotional adaptation to emotionally salient or traumatic events – and combine empirical observations supporting the role of REM sleep (and by association, dreaming) in learning, memory consolidation, emotional processing, and adaptation to stress. Given that REM sleep is associated with muscle atonia, and given the lack of detectable physiological activation during REM sleep episodes preceding awakening from nightmares, dreaming and REM sleep may provide a unique psychophysiological milieu during which traumatic memories can be reexperienced in the absence of physiological arousal. This process could facilitate the attenuation of emotional and physiological responses associated with these memories, leading to desensitization to the trauma-related memories, including a reduction of nightmares. Prospective dream collection studies in combination with repeated objective assessment of sleep patterns and other physiological parameters of central and peripheral arousal in people at high risk for trauma exposure (e.g., disaster workers, police officers, firefighters, deployed military personnel) are necessary to assess the role of nightmares in the recovery of trauma response and PTSD.

Stickgold proposed an explicit REM sleep model of the neurobiological substrates potentially subserving emotional processing and integration of trauma-related memories. Specifically, it is suggested that REM sleep provides a unique neurochemical and neurobiological brain state that allows the transfer of hippocampally mediated episodic traumatic memories and amygdala-dependent salient affect into cortically distributed semantic networks. It should be noted that the possibility has not been evaluated that PTSD-related nightmares may also reflect the failure of compensatory mechanisms aimed at preserving processes such as emotional adaptation, memory consolidation, and sleep continuity.

Sensitization Hypothesis of PTSD Nightmares

While the adaptation process may be successful in individuals who recover from trauma exposure, the persistence of nightmares in those who develop PTSD may, instead, reflect a failure of dreaming and REM

sleep to fulfill their role of emotional adaptation and memory integration. In this perspective, nightmares are the outcome of unsuccessful processing during dreaming and REM sleep. In addition, nightmares may directly contribute to the failure of the adaptation and integration process by promoting sensitization (rather than desensitization) to trauma-related memories. PTSD-related nightmares are often associated with awakenings and are accompanied by intense emotional and physiological arousal. Trauma-related nightmare content, sleep disruption induced by awakening from a nightmare, and heightened physiological and emotional arousal levels that follow such awakenings may contribute to reinforce the occurrence of nightmares and related distress during both wakefulness and sleep. In other words, the recall of nightmares can expose patients to their prior trauma and may even induce retraumatization, since PTSD nightmares often repeat a traumatic situation in whole or in part, whereas habituation to physiological arousal elicited by these memories does not occur.

There is evidence that frequent and distressing PTSD-related nightmares are associated with increased severity of daytime reexperiencing, arousal, and avoidance PTSD symptoms. Alternatively, some PTSD patients who show clinically significant improvements in PTSD following exposure-based therapy also report improvements in trauma-related nightmares, suggesting that habituation to trauma memories during wakefulness can also attenuate nightmare frequency, intensity, and related distress. Desensitization therapy has also been used successfully for the treatment of trauma-related nightmares. These observations support the hypothesis that chronic nightmares without physiological habituation may indeed contribute to the maintenance of PTSD by sensitizing patients to trauma memories.

Potential Psychophysiological and Neurobiological Underpinnings of PTSD-Related Nightmares

While several polysomnographic studies have been conducted in PTSD samples, these sleep measurement methods do not allow the identification of neurobiological underpinnings of trauma-related nightmares or, more generally, of PTSD during sleep. Such an endeavor requires the use of functional sleep neuroimaging approaches. Sleep neuroimaging has not yet been conducted in PTSD patients. However, animal models of the effects of fear conditioning on sleep, and neuroimaging sleep studies in humans, have provided valuable insights into the potential neurobiological underpinnings of altered REM and NREM sleep mechanisms following stress exposure PTSD, and into the neurobiology of normal sleep, respectively. While animal models and sleep imaging in healthy persons limit possible extrapolation of neurobiological underpinnings of trauma-related nightmares, findings derived from these two distinct fields provide evidence for functionally significant overlaps between the neurobiology of fear and sleep regulation.

Functional neuroimaging studies conducted during wakefulness in PTSD patients are consistent with animal models of fear conditioning. PTSD is associated with hyperresponsiveness of the amygdala to threat-related stimuli, and/or a with blunted response of medial prefrontal cortical regions, which exert inhibitory control over the amygdala. Other functional findings include increased thalamic activity with trauma-related stimuli in PTSD compared to non-PTSD subjects, and hyperresponsiveness of the noradrenergic system. Of note, neuronal activity in these structures cannot be captured by conventional sleep measurement methods, which further reinforces the need to use state-of-the-science sleep neuroimaging methods in patients with PTSD.

In rodents, cued fear conditioning increases REM sleep latency, decreases REM sleep duration and number of REM bouts, and increases pontogeniculooccipital (PGO) waves, a marker of alerting mechanisms during sleep and wakefulness analogous to REMs in humans. The effects of fear conditioning on REM sleep in rodents are mediated by amygdalar projections to brain stem regions involved in alerting response and REM sleep generation. Anatomically, the interconnections that the amygdala shares with the basal forebrain, hypothalamus, preoptic area of the anterior hypothalamus, brain stem reticular formation, and solitary tract nucleus allow the amygdala to influence both wakefulness-promoting and sleep-promoting areas. Neuronal firing of the amygdala varies across the sleep–wake cycle, and stimulation of the amygdala during REM sleep increases PGO waves in REM and NREM sleep, whereas inactivation of the amygdala with tetrodotoxin decreases sleep latency and increases slow-wave activity during wakefulness, REM sleep, and NREM sleep. Conversely, ablation of the amygdala in rhesus monkeys is associated with increased sleep consolidation and total sleep time. These findings indicate that the amygdala is an important modulator of sleep and wakefulness regulation mechanisms.

Sleep neuroimaging studies conducted in healthy persons have reliably shown that REM sleep is a natural activator of the amygdala and anterior paralimbic areas and of the medial pons and thalamus. REM sleep is also associated with relative deactivation of lateral prefrontal parietal regions and primary sensory cortices. These selective activation and deactivation

patterns during REM sleep relative to wakefulness have yielded the hypothesis that dreams may reflect the mental representations of high limbic activations in conjunction with deactivation of high-order cortical regions. NREM sleep is a natural deactivator of arousal-promoting structures.

The functionally significant interconnections between the amygdala and wakefulness- and sleep-promoting areas, the unique limbic activation and cortical deactivation patterns observed during sleep in healthy persons, and the functional neuroanatomical findings in PTSD patients during wakefulness all support a role for a heightened amygdalar activity during sleep in PTSD, and provide preliminary guidance for the investigation of the neurobiology of PTSD during REM and NREM sleep. While the relationship between heightened amygdalar activity and dream content cannot yet be ascertained, the role of the amygdala in acquisition of the fear response and the emotional salience of stimuli raises the possibility that heightened amygdalar activity during REM sleep also contributes to salient, negative aspects of dreams and nightmares.

Treatments of PTSD-Related Nightmares

Nightmares are closely related to sleep disturbances, including increased sleep latency, decreased total sleep time, and increased number and duration of awakenings after sleep onset and restless or fitful sleep. Nightmares often produce an insomnia-like pattern of sleep disturbance, if not outright insomnia. Sleep avoidance and fear of darkness, frequent features associated with PTSD-related nightmares, can further reinforce the maintenance of both nightmares and insomnia.

Recommended (by the Expert Consensus Guideline Series on the treatment of PTSD) first-line treatments include Food and Drug Administration (FDA)-approved SSRIs such as sertraline and paroxetine, and cognitive-behavioral interventions such as prolonged exposure and cognitive therapy. However, nightmares are often resistant to such first-line PTSD treatments, and adjunctive treatments are often required to alleviate PTSD-related nightmares.

Several psychological and pharmacological nightmare-focused treatments have been shown to effectively reduce and eliminate nightmares in PTSD patients. Successfully treated nightmare patients often report improvements in sleep quality, feeling more rested upon awakening and having more daytime energy, and reduction in nightmares is a significant predictor of sleep improvement. In addition, nightmare reduction is associated with marked improvements in daytime PTSD symptoms, quality of life, depression and anxiety symptoms, and overall functioning. These findings suggest that at least in some cases, chronic nightmares may represent a primary sleep disorder rather than a symptom of a psychiatric disorder, and as such can benefit from targeted, adjunctive cognitive-behavioral or pharmacological treatments.

Imagery Rescripting and Rehearsal

Numerous techniques have been proposed to treat PTSD-related nightmares, including hypnosis, lucid dreaming, eye movement desensitization and reprocessing, desensitization, and imagery rescripting and rehearsal (IRR). However, only desensitization and IRR have been the objects of controlled studies. Only IRR has been shown to effectively reduce nightmare frequency and nightmare-related distress in patients suffering from idiopathic, recurrent, or PTSD-related nightmares.

IRR for nightmares is composed of two general elements, each of which utilizes several steps in the therapeutic process. The first component is an educational/cognitive restructuring element, focused on helping nightmare sufferers to consider their disturbing dreams as a learned behavior. The second component is an imagery education/training element, which teaches nightmare patients about the nature of human imagery and how to implement a specific set of imagery steps to decrease nightmares. Generally, patients are shown how a nightmare can be effectively treated as a problem in and of itself, without any discussion or emphasis on prior traumatic events or nonsleep-related PTSD symptoms. As such, IRR seeks to minimize the use of exposure therapy as an ingredient in its procedures.

When learning IRR, patients are instructed to modify an original nightmare in any manner of their preference, and to create a new dream scenario. The new dream scenario is then rehearsed in session, and practiced each day until the next session. The rationale is that the rescripting and rehearsal of new dream imagery during the day can replace or alter nightmare scenarios experienced during sleep. Generally, patients report reductions in nightmare frequency and intensity within 6–12 weeks, and improvements are maintained for as long as 30 months.

Several variations of IRR have now been described. The distinguishing features between variations of IRR generally revolve around the degree of exposure used during treatment sessions and/or the specific application of the technique during the sessions. IRR can be delivered in group or individual formats. The number of sessions required for IRR can vary from one to six sessions. Since IRR involves minimal exposure and abreaction, its effectiveness is most likely related to the patient's ability to increase mastery over distressing dream elements by generating new dreams. Research on the scripting of new dreams linked to IRR indicates that mastery is in fact a key

ingredient in this treatment. The observed increase in mastery is all the more remarkable given that IRR does not include instructions to increase mastery when the technique is described during treatment. Although these data are consistent with the view that a dysfunctional imagery system is involved in traumatic nightmares, more work is required to elucidate the relationship between variables related to patient characteristics (e.g., traumatized vs. nontraumatized, presence of psychiatric and sleep disorders, dispositional factors, degree of distress), nightmare content and type (e.g., replicative, recurrent, lifelong, idiopathic), and therapeutic effects (e.g., enhanced self-efficacy, perceived control, modes of emotion expression and representation).

Prazosin

Prazosin is a central α1 adrenoreceptor blocker that is FDA-approved for the treatment of hypertension. Prazosin is currently the most efficacious treatment for PTSD-related nightmares in military veterans and civilian populations with PTSD. Promising results for the use of prazosin to reduce PTSD-related nightmares became available after the Expert Consensus Guideline Series on the treatment of PTSD were established. Results from controlled trials indicate that prazosin effectively reduces chronic PTSD-related nightmares in military and civilian sample groups with PTSD. Reduction in nightmares with prazosin is associated with clinically meaningful improvements in sleep quality, overall PTSD severity, and daytime functioning in combat veterans with chronic PTSD. Most patients experience a recurrence of insomnia and nightmares upon prazosin discontinuation. The drug is well tolerated by patients, is associated with minimal side effects, and no adverse events have been reported with long-term use.

Serotonin-Potentiating Drugs

Randomized controlled trials on the efficacy of FDA-approved SSRIs for the treatment of PTSD (i.e., sertraline and paroxetine) either did not measure improvements in sleep using validated measures, or reported minimal improvements in nightmares or overall sleep quality. In fact, some SSRIs may increase the number of arousals during sleep and decrease total sleep time. Open-label trials with fluvoxamine suggest moderate improvements in PTSD-related nightmares, but randomized controlled trials have not been conducted and side effects may impede tolerability. Open-label trials and case reports suggest that serotonin-potentiating non-SSRIs, such as nefazodone and trazodone, can be associated with reductions in PTSD-related nightmares and sleep disruption. However, double-blind randomized controlled trials in PTSD samples are not available. The need for rigorous randomized controlled trials is highlighted by the example of cyproheptadine. Initial case series suggested that cyproheptadine was a potent agent for the alleviation of PTSD-related nightmares. However, a subsequent double-blind, randomized controlled trial in Vietnam veterans with PTSD found that cyproheptadine could in fact exacerbate nightmares and sleep disruption. A few randomized controlled trials have evaluated the efficacy of monoamine oxidase inhibitors and tricyclic antidepressants for PTSD, and slight to moderate improvements were reported in some, but not all, studies. Optimal doses, durability of therapeutic gains, side effect profiles, and tolerability have not been assessed specifically for PTSD-related nightmares and sleep disruption with non-SSRIs, monoamine oxidase inhibitors, or tricyclic antidepressants. It should be noted that very few trials targeting clinically meaningful reductions in PTSD-related nightmares and sleep disturbances have used specific nightmare measures (e.g., questionnaires, sleep or dream logs) as primary outcome measures. Such measures are necessary to determine the effects of targeted treatments on specific aspects of nightmares, such as frequency, intensity, nightmare-induced sleep disruption, and associated distress.

Other Pharmacological Agents

Several medications that have sedative properties have been used in PTSD patients. Antipsychotics such as olanzapine and quetiapine have been tested in PTSD and have been associated with mixed findings as to their potential for reducing PTSD-related nightmares and insomnia. Despite the lack of evidence for their efficacy in reducing both daytime and nighttime PTSD symptoms, benzodiazepines are often prescribed to patients with PTSD. An open trial with zolpidem, a nonbenzodiazepine hypnotic, reported improvements in sleep and some improvements in nightmares in military veterans with PTSD. Case series reports on other pharmacological agents, such as topiramate, respiridone, and gabapentin, among others, are also available. While most cases report some degree of improvement in nightmares and insomnia in PTSD patients, proper randomized controlled trials are necessary to fully assess the efficacy of these agents as adjunctive PTSD treatments for nightmares and insomnia, and to determine associated health risks in light of side effects such as tolerance, weight gain, and glucose dysregulation.

See also: Amygdala: Contributions to Fear; Dream Function; Dreams and Dreaming: Incorporation of Waking Events; Dreams, Dreaming Theories and Correlates of Nightmares; Nightmares; Posttraumatic Stress Disorder: Overview; Posttraumatic Stress Disorder as an

Emotional Disorder; Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep and Sleep States: PET Activation Patterns; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Barrett D (ed.) *Trauma and Dreams*. Cambridge: Harvard University Press.

Foa EB, Davidson JRT, and Frances A (eds.) (1999) The Expert Consensus Guideline Series: Treatment of posttraumatic stress disorder. *Journal of Clinical Psychiatry* 60(supplement 16).

Harvey AG, Jones C, and Schmidt DA (2003) Sleep and posttraumatic stress disorder: A review. *Clinical Psychology Review* 23: 377–407.

Krakow B, Hollifield M, Johnston L, et al. (2001) Imagery rehearsal therapy for chronic nightmares in sexual assault survivors with posttraumatic stress disorder: A randomized controlled trial. *JAMA: Journal of the American Medical Association* 286: 537–545.

Maher MJ, Rego SA, and Asnis GM (2006) Sleep disturbances in patients with post-traumatic stress disorder: Epidemiology, impact and approaches to management. *CNS Drugs* 20: 567–590.

Mellman TA, David D, Bustamante V, et al. (2001) Dreams in the acute aftermath of trauma and their relationship to PTSD. *Journal of Traumatic Stress* 14: 241–247.

Mellman TA and Hipolito MM (2006) Sleep disturbances in the aftermath of trauma and posttraumatic stress disorder. *CNS Spectrums* 11: 611–615.

Raskind MA, Peskind ER, Kanter ED, et al. (2003) Reduction of nightmares and other PTSD symptoms in combat veterans by prazosin: A placebo-controlled study. *American Journal of Psychiatry* 160: 371–373.

Rothbaum BO and Mellman TA (2001) Dreams and exposure therapy in PTSD. *Journal of Traumatic Stress* 14: 481–490.

Stickgold R (2002) EMDR: A putative neurobiological mechanism of action. *Journal of Clinical Psychology* 58: 61–75.

Wood J, Bootzin R, Rosenhan D, et al. (1992) The effects of the 1989 San Francisco earthquake on frequency and content of nightmares. *Journal of Abnormal Psychology* 101: 219–224.

Relevant Websites

http://www.aasmnet.org – American Academy of Sleep Medicine.

http://www.istss.org – International Society for Traumatic Stress Studies.

http://www.ncptsd.org – National Center for PTSD (US Department of Veterans Affairs).

Dreams, Dreaming Theories and Correlates of Nightmares

T Nielsen, Université de Montréal, Montreal, QC, Canada
R Levin, Ferkauf Graduate School of Psychology, Bronx, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Nightmares Defined

Nightmares are currently defined in both the *Diagnostic and Statistical Manual of Psychiatric Disorders*, *Fourth Edition*, *Text Revision* (DSM-IV-TR) and the *International Classification of Sleep Disorders*, *Second Edition* (ICSD-2) (Table 1). These definitions are consistent in linking nightmares to abrupt rapid eye movement (REM) sleep awakenings with a clear recall of primarily fearful dream content. The more recent ICSD-2 criteria acknowledge that nightmares may involve dysphoric emotions other than fear and anxiety and distinguish idiopathic nightmares from the more severe and highly distressing nightmares associated with posttraumatic stress disorder (PTSD). However, the ICSD-2 does not subscribe to a criterion of subjective distress as does the DSM-IV-TR, even though converging evidence suggests that one's habitual manner of affective responding (nightmare distress) is an important determinant of nightmares as a clinical problem.

Nightmare Prevalence and Correlates

Large community-based epidemiological studies demonstrate a high prevalence of occasional nightmares: about 85% of adults report at least one within the past year, whereas 8–29% report at least one per month. These same studies indicate that 2–6% of respondents report weekly nightmares, a frequency widely considered to be of moderately severe pathology. The 2–6% range is robust cross-culturally, with similar rates reported in Canada, Europe, Japan, the Middle East, and America. This range may nonetheless underestimate the real prevalence of nightmares for a number of reasons, such as reliance on retrospective measures and undergraduate samples, inconsistent operational definitions, and failure to distinguish traumatic from nontraumatic nightmares.

Frequent nightmares are up to three to four times as prevalent in childhood and adolescence as in adulthood. According to the DSM-IV-TR, 10–50% of children aged 3–5 years of age have disturbing dreams, with a prevalence that increases through early adolescence. Nightmare 'problems' (duration >3 months) occur with prevalences in the same order of magnitude: 24%, 41%, and 22% for children between age ranges of 2–5, 6–10, and 11–12 years, respectively. Evidence from the Finnish Twin Cohort, a community-based sample of 1298 monozygotic and 2419 dizygotic twin pairs aged 33–60 years, suggests that nightmares are a common and stable trait from childhood to middle age and are substantially affected by genetic factors. In contrast, nightmare prevalence and frequency in elderly populations are considerably lower than rates found in younger adults.

Females at all ages consistently report nightmares at significantly higher rates than do males, although some portion of this difference may be attributed to trauma-induced nightmares. This consistent gender gap in nightmares appears to first emerge around early adolescence. Nightmares are more frequent and more prevalent in psychiatric populations and are associated with a diversity of symptoms: anxiety, neuroticism, and global symptom reporting, schizophrenia-spectrum symptoms; heightened risk for suicide; dissociative phenomena; health behavioral problems; and sleep disturbances. Nightmares are also linked to a diversity of psychopathological traits. They are particularly prevalent in PTSD, for which they are considered a hallmark symptom. A broad range of traumatic events may trigger nightmares: combat exposure, motor vehicle accidents, natural disasters, crime victimization, and rape.

Nightmares are also reactive to intense stress. More frequent nightmares are reported during increased life stress and are associated with specific stressors, such as anticipated surgery, experimental pain stimulation, menstruation, pregnancy, miscarriage, new motherhood, preparation for exams, sham intelligence testing, stock market downturns, and the viewing of disturbing movies. Their onset is often immediately preceded by stressful events, such as death of someone close, interpersonal conflict, news of a disaster (e.g., an earthquake), and major life events in general.

Nightmares also accompany several chronic health problems, including migraine, bronchitis/asthma, chronic obstructive airways disorder, cardiac disease, and substance abuse. In addition, individuals who report frequent nightmares are characterized by a number of personality variables strongly associated with waking emotional distress: heightened physical and emotional reactivity, imagery vividness, fantasy proneness, maladaptive coping, and thin boundaries.

The association of nightmares with this wide spectrum of pathological symptoms and conditions, all of which are marked by considerable waking emotional

Table 1 DSM-IV-TR and ICSD-2 diagnostic criteria for nightmare disorder[a]

Component	*DSM-IV-TR nightmare disorder (2000)*	*ICSD-2 nightmare disorder (2005)*
Nature of dream recall	A. Repeated awakenings from the major sleep period or naps with detailed recall of extended and extremely frightening dreams, usually involving threats to survival, security, or self-esteem	A. Recurrent episodes of awakenings from sleep with recall of intensely disturbing dream mentation, usually involving fear or anxiety but also anger, sadness, disgust, and other dysphoric emotions
Nature of awakening	B. On awakening from the frightening dreams, the person rapidly becomes oriented and alert (in contrast to sleep terror disorder and some epilepsies)	C. Alertness is full immediately on awakening, with little confusion or disorientation Recall of sleep mentation is immediate and clear
Nature of associated distress	C. The dream experience, or the sleep disturbance resulting from the awakening, causes clinically significant distress or impairment in social, occupational, or other important areas of function	D. Associated features may include: • Return to sleep after the episodes is typically delayed and not rapid
Timing of event	A. The awakenings generally occur during the second half of the sleep period	D. Associated features may include: • The episodes typically occur in the latter half of the habitual sleep period
Physiological criteria	None	None
Differential diagnosis	D. The nightmares do not occur exclusively during the course of another mental disorder (e.g., a delirium, posttraumatic stress disorder) and are not due to the direct physiological effects of a substance (e.g., a drug of abuse, a medication) or a general medical condition	Nightmares are distinguished from: seizure disorder, arousal disorders (sleep terrors, confusional arousal), REM sleep behavior disorder, isolated sleep paralysis, nocturnal panic, posttraumatic stress disorder, acute stress disorder

[a]DSM-IV-TR, *Diagnostic and Statistical Manual of Mental Disorders*, *Fourth Edition, Text Revision*; ICSD-2, *International Classification of Sleep Disorders, Second Edition*. Items A–D refer to specific definitional criteria described in the manuals; REM, rapid eye movement. Data from American Psychiatric Association (2000) *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Arlington, VA: American psychiatric association; and American Academy of Sleep Medicine (2005) *International Classification of Sleep Disorders: Diagnostic & Coding Manual*, 2nd edn. Westchester, IL: American Academy of Sleep Medicine.

distress, supports the contention that nightmare production is related to a personality style characterized by intense reactive emotional distress.

Models of Nightmare Production

Despite a literature that has proliferated over the past two decades, relatively little is known about the etiology and pathogenesis of nightmares. There is a surprising paucity of working models of nightmare production, with most existing attempts being based on clinical observation but little experimental verification. Many existing models address nightmares only indirectly within the broader context of theories about dream function or personality. Further, some models are concerned almost exclusively with elucidating a nightmare function but not how nightmares are formed.

Psychoanalytic and Neopsychoanalytic Models

It is widely accepted that Freud's dream theory failed to deal adequately with nightmares. To incorporate them into his theoretical framework, Freud resorted to the notion of masochistic wish-fulfillment dreams, within which anxiety feelings stemmed from the transformation of libido. Despite the unpopularity of this idea, twentieth-century theories of nightmare production were nonetheless largely informed by Freud's work. For example, various neo-Freudian models proposed that nightmares were expressions of repressed incestuous impulses, were residues of unresolved psychological conflicts, were attempts to transform shame into fear, and so forth. Some of these views (e.g., incestuous impulse theory) have held little currency in contemporary deliberations over processes of nightmare production, whereas others have continued to wield considerable influence. For example, one contemporary model embraces an explicitly psychoanalytic framework in describing a three-stage neurological mechanism responsible for normal dreaming and nonrecurring nightmares. Other, more recent, mastery and adaptation models of dream function (see later) are also broadly consistent with neopsychoanalytic orientations.

Personality and Evolutionary Models

Boundary permeability Based on extensive clinical work with frequent lifelong nightmare sufferers, Hartmann proposed the personality dimension of 'boundary permeability' to explain nightmare pathology. Nightmare sufferers fall to one extreme of this dimension ('thin' boundaries) by virtue of their striking openness, sensitivity, and vulnerability to cognitive and emotional intrusions – and possibly by biologically based differences as well. Hartmann reasoned that thin-boundary individuals were susceptible

to internal events not usually perceived by most others as threatening or traumatic, that they had difficulty discerning internal fantasy from external reality, and that they were prone to source monitoring disturbances across various states of consciousness. Their boundary 'thinness' thus consists in a vulnerability to spikes of heightened emotional distress; during dreaming this vulnerability leads to nightmares.

A modest body of empirical work supports the validity of the Boundary Questionnaire created to measure an individual's boundary thinness or thickness. As predicted by the model, thin-boundary persons have high levels of dream recall, dream bizarreness, and nightmare recall. Relationships have been found for adolescents as well as for adults, but not for healthy elderly individuals. However, failures to find relationships between boundary thinness and nightmare recall have also been reported. In general, despite the generally positive findings and the publication of a book on the boundary model, this approach has not been elaborated into a complete theory of nightmares and has not enjoyed widespread application as such by the research community.

Image contextualization Unlike most other models that consider nightmares to result from a breakdown in emotional regulation processes, Hartmann's contextualization model proposes that (with the exception of nonrepetitive PTSD nightmares) nightmares serve the function of contextualizing, or finding a picture context for an individual's predominant emotional concerns. Contextualization proceeds by establishing a wide swath of new associations to the emotion, the result of which is emotionally therapeutic. Contextualization is presumed to be a function of dreaming more generally, but the contextualizing processes in normal dreams are often difficult to discern when the underlying emotions are too weak or diffuse. However, after exposure to stressful or traumatic events that engender high levels of emotionality, the processes are more evident. Emotion is thus the central instigating force in the formation of dreaming and nightmares, and integration of this emotion, via the establishment of broad memory associations, is the ultimate goal of dreaming.

Contextualization occurs via contextualizing images, which are powerful central images of a dream whose emotions are consistent with a central concern but whose content may differ from this concern. For example, an image of being swept up in a tornado may contextualize an individual's feelings of helplessness, fear, and foreboding that stem from a prior assault. Hartmann's group has demonstrated that such images are, in fact, more frequent after trauma. Research has also revealed that the presence and intensity of contextualizing images are related to thinness on the Boundary Questionnaire, to a history of trauma, and to the impact of a dream on waking life. Independent validation of some features of this model was also recently reported.

Emphasis on a contextualizing mechanism likens this model to connectionist models of memory and emotional processing; both of which are thought to be central to REM sleep function. Recent evidence does support the notion that memory systems become hyperassociative and more flexible during REM sleep. Hartmann's focus on context formation during dreaming also presages recent theoretical trends in neuroscience – specifically, the view that sleep-related alterations in hippocampal context-building functions are key in determining how dreaming transmogrifies episodic memories into oniric worlds and relationships. Thus, the contextualization model anticipates both recent speculations that an REM sleep function is to create contextual memories and recent research demonstrating sleep-related facilitation of the consolidation of implicit, contextual (hippocampal-dependent) learning.

In sum, the contextualization model is a relatively new proposal which is still in the process of validation but which is consistent with several emerging trends in the neuroscience of memory.

Threat simulation The threat simulation model of nightmares is an evolutionary theory that considers nightmares to be virtual representations that enable the self to engage in behaviorally realistic responses to subjectively real threatening events. Active 'rehearsal' of responses during such simulations enhances threat-avoidance skills in the waking world and confers behavioral and survival advantages. Thus, Revonsuo suggests that nightmarish dreams are prime examples of dreams that fully realize this biological function. The suggestion that ancestral threats are more apparent in children's dreams than in those of adults underlines the importance of assessing the dreams of these two populations comparatively.

This model has generated substantial interest and empirical test in the scientific community but remains controversial. The evolutionary assumption of the model, that nightmares are heritable, has been supported to some extent. The only available study on this question used structural equation modeling of the responses of subjects from the Finnish Twin Cohort to reveal persistent genetic effects on the disposition to nightmares both in both childhood (44–45% of phenotypic variance) and adulthood (36–38%).

However, heritability *per se* is insufficient to establish function; many psychiatric disorders bear a genetic component. As described earlier, nightmares

are associated with increased vulnerability, psychopathology, and dysfunctional adaptation to the environment. A functional role for nightmares is not supported by the finding that sleep disturbances (including nightmares) increase the risk of developing PTSD upon subsequent trauma exposure and that nightmares following trauma are associated with more severe PTSD. PTSD is associated with a host of sleep and waking state abnormalities and does not appear to be an adaptive condition in any sense.

While this model has stimulated little research on the content of nightmares *per se*, assessment of dream content has provided several consistent findings. Revonsuo's group reported evidence that severely traumatized children living in environments of threat report dreams with more threatening events and dream threats that are more severe than do less traumatized or nontraumatized children. Similarly, the earliest dreams remembered by adults (i.e., dreamed when they were children) contain a very large proportion of threat themes, also as predicted by the model. Revonsuo also reported that the dreams of college students contained threats that were frequent (66% of reports), severe (39% nontrivial), realistic (aggression, failures, misfortunes, etc.), and directed toward the self (73%), and to which the self responded with relevant defensive behaviors (56%) – all findings consistent with the model. Mixed support comes from an independent study of recurrent dreams in which six of eight predictions from the model were judged to be supported. However, less than 15% of these dreams contained realistic and probable threats critical for physical survival or reproductive success. Further, less than 2% of the dreams supported all of the predictions. Similarly, only 8% of undergraduate students' most recent dreams contain realistic life-threatening events (and escape is depicted in only a third of these), whereas severe life-threatening events are experienced in real life by 45% of them. It bears noting, however, that the presence of threats *per se* in dream content does not necessarily prove that dreams are adaptive in an evolutionary sense; these threats may simply reflect daily reality, as stipulated by the continuity hypothesis of dreaming.

In sum, the threat simulation model remains a provocative but largely unproved explanation of nightmares. As an evolutionary theory, many of its tenets are difficult, if not impossible, to test empirically. However, it provides a context, which places nightmare studies within the broader field of evolutionary biology, and has generated many novel hypotheses for empirical testing.

Neurobiological Models

Neurotransmitter imbalance Hartmann and colleagues early proposed a model of nightmare etiology based upon the effects of pharmacological treatments for chronic nightmare sufferers. An imbalance in neurotransmitter systems, characterized by diminished levels of norepinephrine and/or serotonin and elevated levels of dopamine (or a combination of these), leads to repeated experiences of nightmares. Hartmann reviewed much of the evidence available at the time but appears to have abandoned this work in favor of the boundary and contextualization models. Evaluation of this complex model today would require an extensive review of the side effects and interactions of several classes of drugs and is not attempted here. Interested readers are directed to the 'Further reading' section for relevant reviews of this literature.

REM sleep desomatization Fisher and colleagues are among very few groups to have recorded polysomnographic variables during spontaneous nightmares. They found nightmares to be associated with smaller than expected levels of autonomic activation during REM sleep and, in some cases, stage 2 sleep. The extent of autonomic activation recorded with measures of heart (HR), respiratory rate (RR), and eye movement (EM) activity was low, was limited to the last few minutes of preawakening sleep, and was, in 60% (12 of 20) of the nightmares recorded, absent altogether. Even lower levels of activation were found in a more recent study. This apparent separation of subjective fear (fearful dream imagery) from its normal psychophysiological concomitants (low or no autonomic arousal) was referred to as 'desomatization,' that is, an REM sleep mechanism for modulating anxiety by abolishing or diminishing its physiological concomitants. Such a mechanism was thought to help preserve REM sleep, to prevent the self-perpetuation of anxiety, and to contribute to the mastery of traumatic experiences. Nightmares result when the anxiety exceeds a certain threshold and the REM desomatization mechanism breaks down, allowing autonomic activation to occur.

Findings from this study, though influential in clinical research, remain somewhat questionable because an indeterminate number of patients with borderline psychosis, prior trauma, and comorbid sleep terrors participated in the study sample. Nonetheless, some of the findings have been replicated with a sample of idiopathic nightmare cases. Little new empirical evidence has been brought to bear on this model, although some laboratory findings are consistent with the possibility that dream emotion is inhibited by REM sleep processes related to the orienting response. Fisher's findings and speculations, however, have had an impact. Several investigators have also suggested that components of REM sleep may be responsible for desomatization or desensitization – that

is, the eye movements of REM sleep desensitize affect in a way similar to what occurs in eye movement desensitization and reprocessing (EMDR) therapy.

In sum, the desomatization model is based upon a limited number of polysomnographic recordings and has not been tested systematically, but has generated much further speculation. The specificity of its proposed mechanism remains an intriguing explanation of nightmare function that is compatible with other theories of dream and nightmare function.

Mood regulation Kramer's mood regulatory theory of dreaming is premised on laboratory findings consistent with the claim that REM sleep is characterized by a 'surge' of affective arousal (e.g., a progressive increase and plateau in limbic system, eye movement, and heart and respiratory activity across the REM period). Dream content is proposed to serve the adaptive function of containing these surges by decreasing the intensity and variability of the associated emotion. This is achieved by a particular pattern of dream content that unfolds across the night and is referred to as 'progressive-sequential' (P-S) in nature. P-S dream series enable a form of emotional problem solving that ameliorates mood. P-S dreaming is distinguished from a repetitive-traumatic pattern, during which an emotional conflict is simply stated and restated without evidence of adaptive change. Nightmares presumably occur when the capacity of dreaming to assimilate the emotional surge in this fashion is exceeded.

While the physiological description of REM sleep as surgelike remains debatable, Kramer has marshaled some empirical support for the proposed function of dream content. In general, evidence that dreams are influenced by one's immediate presleep thoughts and emotional experiences, and that one's waking state mood is related to the previous night's dreams, is consistent with the notion that intervening dream activity regulates mood. More specific evidence that dreaming mediates this regulation is that pre- to post-decreases in mood scores, the 'unhappiness subscale' especially, are correlated with intervening dream content scores, scales involving the number of dream characters especially. While one study by Kramer's group failed to replicate this relationship, consistent findings have been reported by independent researchers. In this case, presleep depression scores were found to be correlated negatively with emotional tone in the dreams of the first REM period of the night (higher depression associated with more negative dreaming) but not with sleep physiology variables. There is also evidence for a mood regulatory function of dreaming in studies of persons undergoing marital separation, who report more negative dreams at the beginning and fewer at the end of the night and who prove more likely to be in remission a year later, compared to those with the opposite pattern. These findings may indicate that negative dreams early in sleep reflect a within-sleep mood regulation process, while negative dreams late in sleep reflect a failure of regulation.

The functional claims of the mood regulation model parallel those of several other nightmare models (e.g., image contextualization, REM sleep desomatization) which claim that dreaming functions to adaptively modify emotions. This claim is supported by a small but growing body of experimental evidence. However, there is as yet no convincing evidence that the P-S dream pattern *per se* is the active agent of mood regulation.

Affective network dysfunction Our neurocognitive model of nightmare formation is based upon a synthesis of findings in the areas of sleep physiology, PTSD, and anxiety disorders. It stipulates that nightmares reflect dysfunction in a network of processes that, during normal dreaming, serves the adaptive function of fear memory extinction. At the cognitive level, this extinction function depends upon three imagery processes that operate on the constituent elements of fear memories: (1) element activation, or the increased availability of isolated features of fear memories removed from their episodic (real-world) contexts, (2) element recombination, or the reorganization of these features into novel, virtual, 'here-and-now' simulations of reality, and (3) emotional expression in reaction to the recombined features, which allows new, fear extinction, memories to be formed and maintained.

At the neural level (see Figure 1), the fear extinction function is supported by a network of four limbic, paralimbic, and prefrontal regions that constitute the control center for a number of emotional processes, including the perception and representation of emotional stimuli and the expression and regulation of emotional responses. These four brain regions, which we refer to as the 'AMPHAC model,' include the amygdala (Am), the medial prefrontal cortex (mPFC), the hippocampal (Hip) complex, and the anterior cingulate cortex (ACC). The AMPHAC regions operate in a coordinated manner to influence other perceptual, cognitive, memorial, and affective brain events. However, general correspondences between each region and cognitive processes are also postulated (e.g., Am, emotional activation; Hip, control of memory context; mPFC, control of extinction memories; ACC, regulator of affect distress).

The cognitive and neural explanatory levels together define an affective network within which

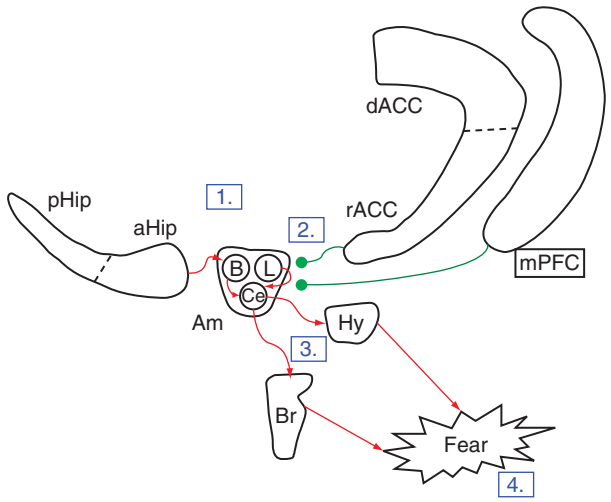

Figure 1 Schematic representation of a network of brain regions hypothesized to be implicated in the production and extinction of fear during normal dreaming; dysfunction in this network is responsible for nightmares. (1) Context is relayed in realistic (virtual) form via anterior hippocampus (aHip) to basal nucleus (B) of the amygdala (Am) and is further processed by the central (Ce) nucleus. (2) Medial prefrontal cortex (mPFC) and dorsal and rostral anterior cingulate cortex (dACC, rACC) afferents to the amygdala regulate the output of Ce neurons to induce extinction and signal distress and maintain appropriate levels of fear. (3) Ce nucleus signals brain stem (Br) and hypothalamus (Hy) to produce (4) the autonomic and behavioral correlates of fear. Excitatory connections are shown in red, inhibitory connections in green. For readability, only connections judged most pertinent to the consolidation and extinction of fear are indicated. Reproduced from Levin R and Nielsen TA (2007) Disturbing dreams, posttraumatic stress disorder, and affect distress: A review and neurocognitive model. *Psychological Bulletin* 133(3): 482–528, with permission from APA.

perturbations may produce a variety of dysphoric dreaming types – from occasional bad dreams to nontraumatic nightmares to recurring posttraumatic nightmares. Further, network activity is 'cross-state,' meaning that it is active during both sleep and wakefulness, which means that its functions – and especially its dysfunctions – should be measurable with various psychophysiological instruments both during sleep and waking states.

Two cross-state factors are proposed for which substantial supporting evidence exists. One is a situational or state factor, affect load, and reflects the combined influence of stressful and emotional negative events (e.g., interpersonal conflicts, daily hassles) on an individual's capacity to effectively regulate emotions. Increases in affect load result in intensified dysphoric dreaming in general. A second factor is a dispositional or trait-like factor, affect distress, and is characterized as a long-standing tendency to experience heightened distress and negative affect in response to increases in affect load, and to react. Affect distress is influenced by events such as prior neglect and trauma, is akin to both the negative affect and negative emotions (personality dimensions), and is associated with a wide variety of psychopathological conditions (anxiety disorders, health behaviors, PTSD, etc.). Affect distress is related to early development and acts as a risk factor diathesis for emotional dysregulation in the face of subsequent stress.

Thus, during nightmares, high levels of affect load and/or affect distress interact with the neurophysiological state of REM sleep so as to favor the activation of fear memories that are coherent (i.e., that resist feature recombination). These memories are akin to those occurring in waking, fear-based pathological conditions such as phobias or social anxiety. Nightmare-related fear memories are highly resistant to extinction, overly weighted with response elements (usually involving escape or avoidance), and, in more severe instances, corrupted by elevated affect distress. A fuller delineation of this model can be found in recent work by Levin and Nielsen.

As a new model of nightmare formation, the affective network dysfunction model has not yet been tested empirically. However, its tenets are consistent with much literature; its neural propositions, especially, are consistent with recent neuroimaging studies of REM sleep, PTSD, anxiety disorders, and personality factors. To illustrate, several studies indicate that REM sleep is characterized by high levels of activation in Am, mPFC, ACC, and the hippocampal complex. Additionally, PET studies have found that glucose metabolic rates in mPFC during REM sleep are highly correlated with elevated anxiety in the content of dreams sampled during these REM periods. Neuropsychological evidence from brain-lesioned patients also supports the model, demonstrating a link between temporolimbic brain regions and frequent nightmares of both recurring and nonrecurring types. Eight of nine patients who reported recurring nightmares were found to have temporal or frontotemporal lesions, including, in some cases, the Hip or ACC. Nightmares and 'dream–reality confusions' (DRCs) were also habitually associated with limbic lesions (nine of ten patients); six of these patients had lesions affecting the mPFC or ACC or both.

Findings from PTSD patients also support this model. All four of the designated brain regions are affected in PTSD patients. Further, heightened affect distress may be expressed in the form of several sleep-related hyperarousal symptoms, including increased awakenings, wake after sleep onset (WASO), and insomnia, as well as nightmares in stages other than REM sleep and at times other than the habitual last third of the night (e.g., stage 2 nightmares occurring early in the sleep episode). Hyperarousal is also suggested by the expression of motor activity in sleep,

including REM-related twitches in leg muscles, more periodic leg movement during sleep in all stages, frequent large body movements, and more REM-related motor activity and vocalization. Explosive motor activity can be elicited from any stage of sleep in some patients with war-related PTSD.

In sum, the affective network dysfunction model is new and still untested but consistent with a large and recent empirical literature. The proposed fear extinction function, while novel, is nonetheless modeled after a well-established research literature on the nature, learning, and unlearning of fear memories. This function is compatible with adaptive emotional functions proposed by other dream and nightmare theorists and may serve as an integrative foundation for resolving these diverse theories.

See also: Dream Function; Dreams and Nightmares in PTSD; Dreams and Dreaming: Incorporation of Waking Events; Nightmares; Parasomnias; Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Research and Sleep Medicine in Historical Perspective; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Arkin AM and Steiner SS (1978) The effects of drugs on sleep mentation. In: Arkin AM, Antrobus JS, and Ellman SJ (eds.) *The Mind in Sleep*, pp. 393–415. New York: Lawrence Erlbaum Associates.

Braun AR, Balkin TJ, Wesensten NJ, et al. (1998) Dissociated pattern of activity in visual cortices and their projections during human rapid eye movement sleep. *Science* 279: 91–95.

Cartwright R (2005) Dreaming as a mood regulation system. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 565–572. Philadelphia, PA: Elsevier Saunders.

Davis M, Myers KM, Chhatwal J, et al. (2006) Pharmacological treatments that facilitate extinction of fear: Relevance to psychotherapy. *NeuroRx* 3: 82–96.

Hublin C, Kaprio J, Partinen M, et al. (1999) Nightmares: Familial aggregation and association with psychiatric disorders in a nationwide twin cohort. *American Journal of Medical Genetics* 88: 329–336.

Hull AM (2002) Neuroimaging findings in post-traumatic stress disorder. Systematic review. *British Journal of Psychiatry* 181: 102–110.

Kim JJ and Jung MW (2006) Neural circuits and mechanisms involved in Pavlovian fear conditioning: A critical review. *Neuroscience and Biobehavioral Reviews* 30: 188–202.

Kramer M (1993) The selective mood regulatory function of dreaming: An update and revision. In: Moffitt A, Kramer M, and Hoffmann R (eds.) *The Functions of Dreaming*, pp. 139–196. Albany, NY: State University of New York.

Levin R and Fireman G (2002) Nightmare prevalence, nightmare distress, and self-reported psychological disturbance. *Sleep* 25: 205–212.

Levin R and Nielsen TA (2007) Disturbing dreams, posttraumatic stress disorder, and affect distress: A review and neurocognitive model. *Psychological Bulletin* 133(3): 482–528.

Maquet P, Ruby P, Maudoux A, et al. (2005) Human cognition during REM sleep and the activity profile within frontal and parietal cortices: A reappraisal of functional neuroimaging data. *Progress in Brain Research* 150: 219–227.

Nielsen TA (2005) Disturbed dreaming in medical conditions. In: Kryger M, Roth N, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 936–945. Philadelphia, PA: Elsevier Saunders.

Nielsen TA and Stenstrom P (2005) What are the memory sources of dreaming? *Nature* 437: 34–38.

Nielsen TA and Zadra AL (2005) Nightmares and other common dream disturbances. In: Kryger M, Roth N, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 926–935. Philadelphia, PA: Elsevier Saunders.

Nutt DJ and Malizia AL (2004) Structural and functional brain changes in posttraumatic stress disorder. *Journal of Clinical Psychiatry* 65(supplement 1): 11–17.

Pace-Schott EF, Gersh T, Silvestri R, et al. (2001) SSRI treatment suppresses dream recall frequency but increases subjective dream intensity in normal subjects. *Journal of Sleep Research* 10: 129–142.

Revonsuo A (2000) The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences* 23: 877–901.

Solms M (1997) *The Neuropsychology of Dreams*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stickgold R (2005) Why we dream. In: Kryger M, Roth N, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 579–587. Philadelphia, PA: Elsevier Saunders.

Thompson DF and Pierce DR (1999) Drug-induced nightmares. *Annals of Pharmacotherapy* 33: 93–98.

Drosophila Apterous Neurons: From Stem Cell to Unique Neuron

S Thor, Linköping University, Linkoping, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Animals contain large numbers of different cell types, and this cellular diversity is profound in the nervous system, where estimates in mammalians are as high as 10 000 unique cell types. Understanding how this remarkable cellular diversity is generated, and with such precision, remains one of the challenges of neurobiology. Although tremendous progress has been made during the past two decades with respect to our understanding of the molecular genetic mechanisms controlling nervous system development, due to the complexity of this system, in many cases research has not progressed beyond the descriptive level. For instance, in more complex tissues such as the mammalian, or even *Drosophila*, adult brains, we are far from even completing a survey of the cellular complexity, let alone understanding the genetic mechanisms involved in generating that diversity. Moreover, although one may argue that it is not necessary to understand the regulatory details underlying the specification and differentiation of each and every *Caenorhabditis elegans* and *Drosophila* neuron, this is obviously not true for the human nervous system. Here, the ultimate goal is to characterize such mechanisms for every one of the 10 000 different neuronal subtypes since they are all likely involved in some intriguing aspect of human biology and are likely linked to various human diseases. As exemplified here, by the regulatory complexity underlying specification of a few cell fates in the powerful *Drosophila* model system, decoding the molecular genetic mechanisms controlling the generation of the complete developing human nervous system remains a major challenge.

This article reviews studies on neuronal cell-type specification of a unique neuronal subset, the apterous (Ap) neurons in the *Drosophila* ventral nerve cord. The focus is on a subset of Ap neurons – the Ap cluster of four cells located in each thoracic hemisegment – and studies that address the origin, generation, specification, differentiation, and pupal remodeling of these neurons are reviewed.

The Apterous Neurons of the *Drosophila* Ventral Nerve Cord

The *Drosophila* central nervous system can be subdivided into the brain and the ventral nerve cord (VNC). These tissues are functionally equivalent to the mammalian brain and spinal cord, respectively. The VNC is segmentally organized into three thoracic and eight abdominal segments, and it contains approximately 10 000 cells, the majority of which are neurons. In the VNC, only 90 cells express the LIM-homeodomain gene *apterous* (*ap*), but these neurons, the Ap neurons, are remarkably diverse. They differ in axon pathfinding; most Ap neurons extend their axons in a common fascicle, whereas the Tv neuron innervates a peripheral secretory gland, the dorsal neurohemal organ (DNH), which is a specialized glial-derived structure present in the three thoracic VNC segments (Figure 1).

Ap neurons also differ in neurotransmitter expression and approximately half are peptidergic. In this peptidergic subclass, the Tv cell selectively expresses the neuropeptide gene *FMRFamide*, whereas dAp and Tvb cells selectively express the neuropeptide gene *Nplp1* as well as a dopamine type I receptor gene, *Dop-R*. Importantly, for these terminal differentiation genes, their expression within the VNC is confined to these subsets of Ap neurons; *FMRFa* is only expressed in 6 and *Nplp1/Dop-R* in 28 of the 10 000 cells present in the VNC, respectively. How is this remarkably specific gene expression controlled?

Genetic Mechanisms of Apterous Neuron Specification

Studies have identified a number of regulatory genes that act postmitotically in the specification and differentiation of the Ap neurons. These include the transcription factors *ap* (mammalian *Lhx2a-b*), the Col/Olf-1/EBF family member *collier/knot* (*col*) (mammalian *ebf1–3*), the zinc finger gene *squeeze* (*sqz*) (mammalian *CIZ*), and the basic helix–loop–helix gene *dimmed* (*dimm*) (mammalian *Mist1*) as well as the transcriptional cofactors *eyes absent* (*eya*) (mammalian *Eya1–4*) and *dachshund* (*dac*) (mammalian *Dach1–2*). In addition, FMRFa expression in the Tv neuron is critically dependent on a target-derived transforming growth factor-β (TGF-β)/bone morphogenetic protein (BMP) signal – the first identified case of a retrograde instructive signal. These regulatory genes act in a combinatorial manner such that a combinatorial code of *ap/sqz/dac/eya/dimm* and target-derived TGF-β/BMP signaling dictates Tv neuron identity and activates *FMRFa* expression, and *ap/eya/dimm/col* dictates Tvb/dAp identity and activates *Nplp1/Dop-R* expression (Figure 2).

As outlined previously, using the neuropeptides FMRFa and Nplp1 as terminal cell identity markers, as well as using independent axonal markers for these

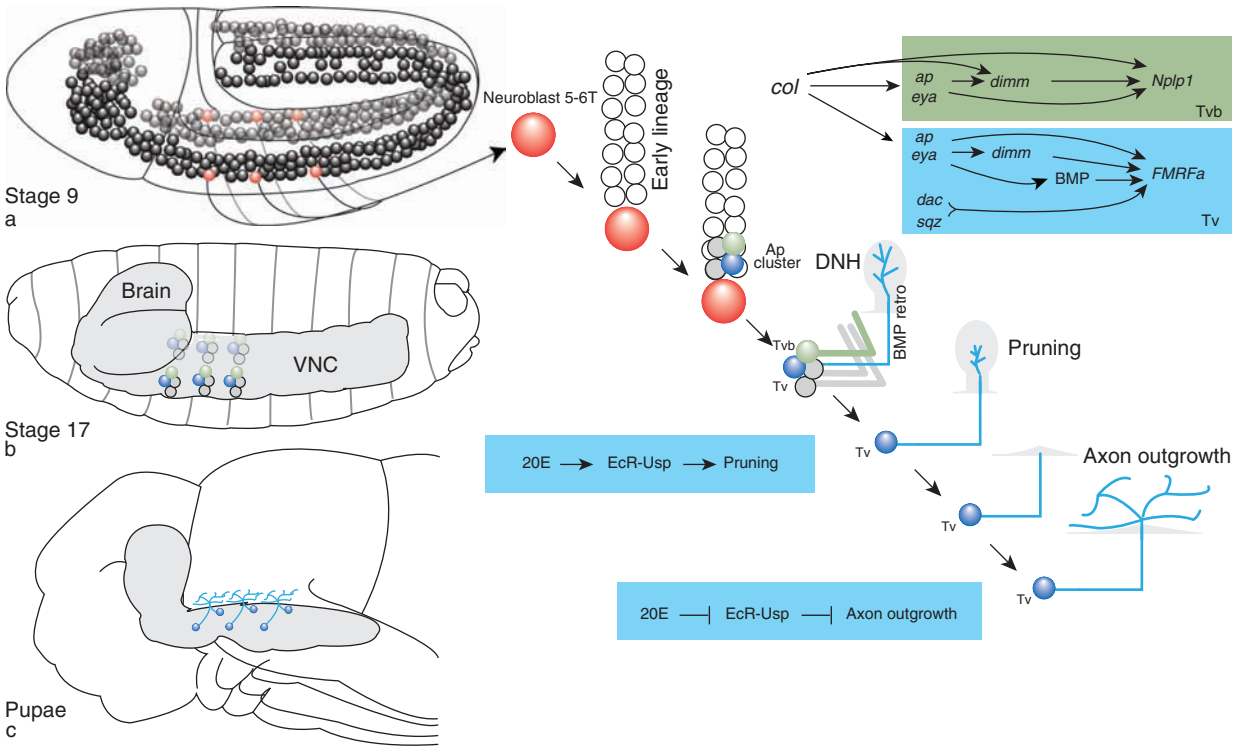

Figure 1 (a) The *Drosophila* central nervous system is generated from a defined set of neuroblasts (gray circles) that delaminate from the ectoderm during embryonic stages 9–11. The six NB 5-6T neuroblasts are located laterally in the three thoracic hemisegments (red circles). NB 5-6T undergoes a series of divisions to generate an early lineage of neurons and glia, after which the four Ap cluster neurons are generated. A complex regulatory cascade, unique to the Tvb (Nplp1 neuropeptide expressing; green circles) and Tv (FMRFa expressing; blue circles) neurons, ensures that these two neurons acquire proper identity. (b) After completion of embryonic neurogenesis, the *Drosophila* VNC is segmentally organized into the three thoracic and eight abdominal segments. The VNC contains an estimated 10 000 cells, of which only 90 express the *ap* gene, with the prominent Ap cluster of four cells locates in each of the six thoracic hemisegments. All Ap cluster neurons have now extended their axons, and the Tv neurons (blue) have received the TGF-β/BMP ligand Glass bottom boat that finally triggers expression of the *FMRFa* neuropeptide gene, whereas the Tvb neurons (green) now express the *Nplp1* neuropeptide gene. (c) During metamorphosis, the Tv neurons undergo elaborate changes in their axonal projections, involving pruning, axon outgrowth, and branching. These events are under the control of the steroid hormone 20E that acts via the EcR-Usp nuclear receptor signaling system. The function of the 20E-EcR-Usp system during these events is rather complex, and a simplified version if depicted here.

neurons, a number of genes have been identified that are important for specification of Tv/FMRFa and/or Tvb/Nplp1 cell fate. However, how do these regulators act – in parallel or in a genetic cascade? Using a multitude of markers and experimental approaches, such as loss of function, rescue (gene A of mutant A), cross-rescue (gene A of mutant B), misexpression, co-misexpression, gene expression analysis, and temporal analysis of gene function using transgenic RNAi (*UAS-RNAi*), it has been determined that the identified regulators act in rather complex genetic cascades (Figure 1). It is noteworthy that not all regulators participate in all events during Ap cluster specification and differentiation. For instance, *ap* and *eya*, but not *sqz*, *dimm*, or *dac*, are involved in axon pathfinding of Ap neurons. Similarly, *sqz* plays a unique role in determining the proper composition of Ap cluster neurons, and *sqz* mutants display both ectopic Ap cluster neurons and ectopic Tvb neurons. Thus, there exists a clear division of labor, with most regulators participating in more than one, but not all, events during Ap cluster specification and differentiation.

The Apterous Cluster Is Generated by the Stem Cell Neuroblast 5-6T

Studies have led to the identification of the stem cell that generates the lateral four-cell clusters of Ap neurons located in the three thoracic segments that include the Tv and Tvb neurons (Figure 1(a)). A large body of work during the past 20 years has been dedicated to understanding the development of the *Drosophila* VNC, revealing how a specific region of the ectoderm acquires neurogenic potential; how this neuroectodermal sheet is patterned; and how precursor cells, central nervous system stem cells

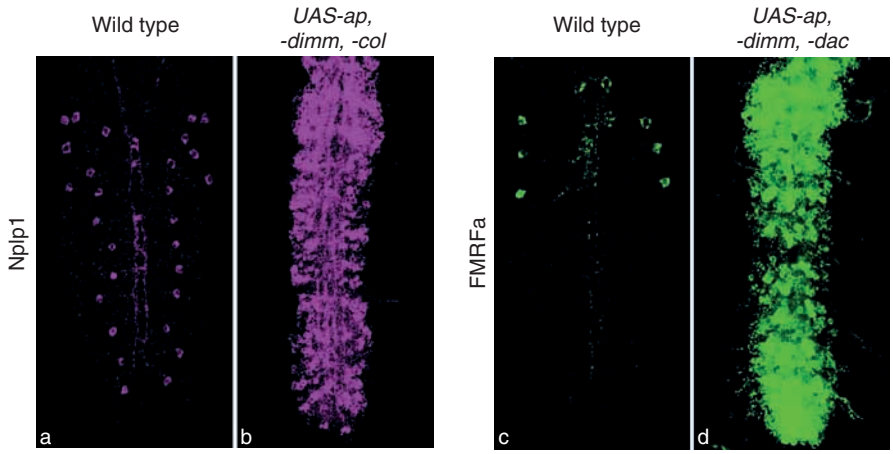

Figure 2 (a, c) Expression of the neuropeptides Nplp1 and FMRFa is confined to 28 and 8 of 10 000 cells in the wild-type VNC, respectively. (b, d) Misexpression of specific combinatorial codes of regulators in all neurons, by crossing a transgenic *elav-Gal4* driver to transgenic lines carrying *UAS-cDNA* constructs, triggers widespread ectopic neuropeptide expression. Adapted from Baumgardt M, Miguel-Aliaga I, Karlsson D, Ekman H, and Thor S (2007) Specification of neuronal identities by feedforward combinatorial coding. *PLoS Biology* 5: 295–308, *Public Library of Science*.

(denoted neuroblasts in *Drosophila*), are determined within this sheet. These early patterning events result in the generation of a set of 30 neuroblasts in each developing hemisegment, each expressing a unique combinatorial code of regulatory genes – a code that will dictate a stereotyped number of asymmetric divisions that generate a unique set of neurons and glia. By taking advantage of this previously generated information, it has been determined that the four Ap cluster cells are generated from the thoracic neuroblast 5-6 (NB 5-6T) (Figures 1(a) and 1(b)). Having identified NB 5-6 as the stem cell generating the Ap cluster, now the goal is to decipher the mechanisms by which global regulatory mechanisms, acting along the anterior–posterior, dorsal–ventral, and segmental axes, integrate with temporal transitions in stem cells to generate the four Ap cluster neurons.

Remodeling of Apterous Tv Neurons during the Pupal-to-Adult Transition

After 3 or 4 days of life, *Drosophila* larvae leave the food, attach themselves to a rigid surface, and undergo a fascinating transformation – the metamorphosis of holometabolous insects. During this stage, many larval tissues undergo histolysis (tissue degradation) and many new, adult-specific structures are formed. At this stage, studies using live imaging time-lapse videos to follow the pruning and extension events in great detail have revealed that one of the Ap neurons, the Tv neuron, undergoes a dramatic remodeling of its axonal arbors, involving an initial retraction (pruning) followed by axon outgrowth and extensive branching onto the surface of the VNC (Figure 1(c)). Steroid hormones play a central role during insect metamorphosis and have been found to play an important role also in the Tv neuron. One of the key players in this regulatory system is the hormone ecdysone and its metabolite, 20-hydroxyecdysone (20E). 20E binds to a nuclear hormone receptor consisting of a heterodimer of two nuclear receptors, EcR and the RXR-ortholog Ultraspiracle (USP), and this heterodimer binds to target sites in the DNA to regulate gene expression. By analyzing EcR mutants, misexpressing dominant negative EcR forms, and by blocking EcR by transgenic RNA interference (*UAS-RNAi*), it has been demonstrated that EcR plays an important role, particularly during the pruning phase. For example, by driving expression of a dominant negative EcR construct from the FMRFa enhancer (*FMRFa-Gal4/UAS-EcRDN*), it was found that the pruning was severely impaired and axon branch extension delayed and reduced. Since 20E-EcR-Usp signaling likely affects a multitude of neurons in different ways, it would be interesting to learn how this system integrates with neuronal subtype-specific regulatory mechanisms. Moreover, whether other Ap neurons also undergo remodeling during metamorphosis has not been addressed.

Important Themes during Neuronal Subtype Specification

Studies of Ap neurons have revealed several interesting mechanisms of neuronal subtype specification, and the more pertinent ones are highlighted in this section.

Combinatorial Coding versus Master Regulators

In-depth genetic studies, particularly of the Tv and Tvb neurons, have revealed an intriguing division of labor between the identified regulators. Notably, whereas some features of both Tv/FMRFa and Tvb/Nplp1 neuron identity are under strict combinatorial control – expression of FMRFa requires the combined action of *ap/sqz/dac/eya/dimm* and target-derived TGF-β/BMP signaling and expression of *Nplp1* depends on *ap/eya/dimm/col* (Figure 2) – other features, such as neuropeptide processing capacity, are under discreet control of one regulator, namely *dimm*, which is necessary and sufficient to activate gene expression of the neuroamidase gene *PHM* in both Tv and Tvb. In fact, *dimm* can activate *PHM* expression in most, if not all, neurons and can thus be viewed as a 'master' regulatory gene, perhaps better termed a 'selector' gene, for one specific neuronal subtype property – neuropeptidergic identity.

Feed-Forward Loops

Genetic network studies in single cell systems such as *Escherichia coli* and yeast have identified a common regulatory loop denoted the 'feed-forward loop.' In this loop, gene A will activate gene B, and A/B then act together to activate C. Studies of the Ap neurons have revealed the existence of such feed-forward loops during both Tv and Tvb specification. This is clearly exemplified by the action of *col* in the early postmitotic Tvb neuron, in which it initially activates *ap* and *eya*, subsequently acts with *ap/eya* to activate *dimm*, and, finally, acts with all three regulators to activate *Nplp1* expression. In adjacent Ap cluster neurons, *col* plays only an early postmitotic role to activate *ap* and *eya*. Thus, the nature of *col* expression has an informative value: Transient *col* expression activates *ap/eya* and generates a 'generic' Ap cluster fate, whereas persistent *col* expression leads to subsequent activation also of *dimm* and, finally, *Nplp1*, and thus it has a Tvb-specific value.

Target-Derived Instructive Signals

The results of many studies of mammals during the past few decades have led to the development of the well-known neurotrophic model, in which mammalian neurons are strongly dependent on target-derived signals for their survival. These signals primarily revolve around two different receptor-tyrosine kinase signal transduction pathways – the neurotrophin and glial cell line-derived neurotrophic factor systems. Genetic studies in *Drosophila*, however, have revealed that the TGF-β/BMP signal transduction system is also used in a target-derived manner, specifically to modulate the arbor size of developing motor neurons. However, studies of the Tv neurons have revealed that the TGF-β/BMP system is also used in an instructive manner to dictate terminal identity of unique neurons, and they have found that expression of FMRFa is strictly dependent on a TGF-β/BMP signal from the DNH target gland. It is tempting to speculate that such instructive target-derived signals may indeed be quite common also in the mammalian nervous system, and that their identification may have been hindered by the prevailing target-derived survival signals – that is, malfunction of axon transport or axon misrouting may often lead to cell death, thus occluding any effects on neuronal cell fate determination.

Remodeling

When is a neuron terminally differentiated? The Tv neurons illustrate this issue: Their axons innervate the DNH target gland in the mid- to late embryo and begin expressing the FMRFa neuropeptide, only to go through a complete remodeling after 3 or 4 days of larval life. It is tempting to speculate that this morphological change may be accompanied by other changes in its function and neurotransmitter profile, and it will be interesting to relate the genetic mechanisms involved in generating the Tv neuron in the embryo to the mechanisms acting to remodel this cell in the pupae.

Conclusion

As revealed by the Ap neurons, there exists an amazing complexity of gene regulatory networks that act to specify unique neuronal cell identity. Similar highly complex networks have been identified that act to specify other well-studied neuronal subtypes, such as the ASER/L neurons of *C. elegans* and the spinal motor neurons of the chick and mouse spinal cord. The following are some of the more pressing questions that may now be addressed in this and other models: (1) How are the correct numbers of cells specified (i.e., only 6 Tv and 28 Tvb neurons per VNC)? (2) Are early specification genes used again during later remodeling and 'homeostasis' of mature neurons? (3) How is context specificity achieved – that is, most of the regulators identified that act to specify Ap neuron identity also act in other developmental events – how is their distinct role in each event determined? It is hoped that answering these questions will help us address neuronal subtype specification in humans.

See also: Differentiation: the Cell Cycle Instead; Motor Neuron Specification in Vertebrates; Neural Crest Cell Diversification and Specification: Melanocytes; Oligodendrocyte Specification; Terminal Differentiation: REST.

Further Reading

Aberle H, Haghighi AP, Fetter RD, McCabe BD, Magalhaes TR, and Goodman CS (2002) Wishful thinking encodes a BMP type II receptor that regulates synaptic growth in *Drosophila*. *Neuron* 33: 545–558.

Allan DW, Park D, St. Pierre SE, Taghert PH, and Thor S (2005) Regulators acting in combinatorial codes also act independently in single differentiating neurons. *Neuron* 45: 689–700.

Allan DW, Pierre SE, Miguel-Aliaga I, and Thor S (2003) Specification of neuropeptide cell identity by the integration of retrograde BMP signaling and a combinatorial transcription factor code. *Cell* 113: 73–86.

Baumgardt M, Miguel-Aliaga I, Karlsson D, Ekman H, and Thor S (2007) Specification of neuronal identities by feedforward combinatorial coding. *PLoS Biology* 5: 295–308.

Benveniste RJ, Thor S, Thomas JB, and Taghert PH (1998) Cell type-specific regulation of the *Drosophila* FMRF-NH2 neuropeptide gene by Apterous, a LIM homeodomain transcription factor. *Development* 125: 4757–4765.

Brown HL, Cherbas L, Cherbas P, and Truman JW (2006) Use of time-lapse imaging and dominant negative receptors to dissect the steroid receptor control of neuronal remodeling in *Drosophila*. *Development* 133: 275–285.

Gauthier SA and Hewes RS (2006) Transcriptional regulation of neuropeptide and peptide hormone expression by the *Drosophila* dimmed and cryptocephal genes. *Journal of Experimental Biology* 209: 1803–1815.

Hewes RS, Park D, Gauthier SA, Schaefer AM, and Taghert PH (2003) The bHLH protein Dimmed controls neuroendocrine cell differentiation in *Drosophila*. *Development* 130: 1771–1781.

Jessell TM (2000) Neuronal specification in the spinal cord: Inductive signals and transcriptional codes. *Nature Reviews Genetics* 1: 20–29.

Lundgren SE, Callahan CA, Thor S, and Thomas JB (1995) Control of neuronal pathway selection by the *Drosophila* LIM homeodomain gene *apterous*. *Development* 121: 1769–1773.

Mangan S and Alon U (2003) Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America* 100: 11980–11985.

Marques G, Bao H, Haerry TE, et al. (2002) The *Drosophila* BMP type II receptor Wishful Thinking regulates neuromuscular synapse morphology and function. *Neuron* 33: 529–543.

Marques G, Haerry TE, Crotty ML, Xue M, Zhang B, and O'Connor MB (2003) Retrograde Gbb signaling through the Bmp type 2 receptor wishful thinking regulates systemic FMRFa expression in *Drosophila*. *Development* 130: 5457–5470.

Miguel-Aliaga I, Allan DW, and Thor S (2004) Independent roles of the dachshund and eyes absent genes in BMP signaling, axon pathfinding and neuronal specification. *Development* 131: 5837–5848.

Park D, Han M, Kim YC, Han KA, and Taghert PH (2004) Ap-let neurons – a peptidergic circuit potentially controlling ecdysial behavior in *Drosophila*. *Developmental Biology* 269: 95–108.

Poole RJ and Hobert O (2006) Early embryonic programming of neuronal left/right asymmetry in *C. elegans*. *Current Biology* 16: 2279–2292.

Schmid A, Chiba A, and Doe CQ (1999) Clonal analysis of *Drosophila* embryonic neuroblasts: Neural cell types, axon projections and muscle targets. *Development* 126: 4653–4689.

Schubiger M, Tomita S, Sung C, Robinow S, and Truman JW (2003) Isoform specific control of gene activity *in vivo* by the *Drosophila* ecdysone receptor. *Mechanisms of Development* 120: 909–918.

Schubiger M, Wade AA, Carney GE, Truman JW, and Bender M (1998) *Drosophila* EcR-B ecdysone receptor isoforms are required for larval molting and for neuron remodeling during metamorphosis. *Development* 125: 2053–2062.

Skeath JB and Thor S (2003) Genetic control of *Drosophila* nerve cord development. *Current Opinion in Neurobiology* 13: 8–15.

Drug Addiction: Behavioral Neurophysiology

R M Carelli, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Drug addiction in humans is a chronic illness characterized by repeated cycles of compulsive drug intake, abstinence from drug taking, and resumption of drug consumption (relapse). In order to develop effective pharmacotherapies to treat drug addiction, it is critical to elucidate the underlying neurobiological mechanisms mediating this behavior. Research using a variety of techniques supports the view that the 'brain reward system,' most notably the nucleus accumbens (NAc) and associated regions, is crucially involved in the drug addiction process. Electrophysiological recording in behaving animals is one technique that has provided important insight into cellular mechanisms underlying drug-seeking behaviors. Some of the most informative studies have used electrophysiological recordings in conjunction with the drug self-administration procedure. In this case, extracellular neuronal activity is recorded in animals that are implanted with an in-dwelling catheter in the jugular vein, and trained to press a lever for intravenous infusion of a drug. Drug self-administration is considered one of the best animal models of human drug addiction since the drug is not given to the animals by the experimenter, but instead animals give the drug to themselves. In fact, virtually all drugs that are abused by humans are readily self-administered by animals including cocaine, heroin, nicotine, morphine, and alcohol. One advantage of this combined approach is that it allows for the characterization of cell firing within the brain reward circuit during specific stages of drug-seeking behavior (e.g., initiation of drug-seeking behavior including approach toward the lever vs. lever depression for drug infusion). Moreover, this experimental method allows for the investigation of other critical factors, such as the role of environmental stimuli or drug abstinence periods, in controlling and/or modulating cellular activity underlying drug addictive behaviors.

Anatomy of the Nucleus Accumbens: A Key Neural Substrate in Addiction

In order to appreciate the contribution of electrophysiological investigations in behaving animals to our understanding of the neurobiology of addiction, it is important to consider the anatomic organization of the NAc, a key neural substrate that has been extensively studied with respect to drug abuse. In the rodent, the NAc is composed of two primary subregions that differ with respect to their anatomic organization and functional properties: the core and shell. Extensive studies completed in rodents show that the subregions of the NAc receive afferent projections from a variety of cortical and subcortical structures including the basolateral amygdala (BLA), the prefrontal cortex (PFC), the subiculum of the hippocampus, and the ventral tegmental area (VTA). In turn, the major efferent projections of the NAc involve motor-related regions although some differences exist with respect to efferent projections of the core and shell. Specifically, the core parallels basal ganglia circuitry sending outputs through the ventral pallidum (dorsolateral district), subthalamic nucleus, and substantia nigra. These outputs in turn project via the motor thalamus to premotor cortical areas. In contrast, the shell projects preferentially to subcortical limbic regions including the lateral hypothalamus, ventral pallidum (ventromedial district), and VTA.

The vast majority of neurons in the NAc are of the medium-spiny type (MSN) that comprise the majority of the projection neurons from this region. Importantly, it has been shown that NAc afferents make 'convergent' synaptic contacts onto MSNs. For example, studies using immunocytochemistry in conjunction with electron microscopy showed that hippocampal and dopaminergic inputs converge onto the same NAc neuron. Similarly, a convergence of inputs from the medial PFC and the ventral subiculum on NAc neurons has been identified.

Virtually all drugs of abuse increase levels of the neurotransmitter dopamine in the NAc; dopamine is considered a key chemical substrate involved in drug addiction. In the NAc, it is believed that dopamine functions as a neuromodulator, influencing the activation of NAc cell firing by specific NAc afferents including the hippocampus, BLA, and PFC. As noted above, anatomic studies showing the convergence of synaptic input onto NAc neurons from various neural structures support the view that dopamine is in a unique position to modulate afferent inputs from cortical and subcortical areas onto specific NAc cells and ultimately influence behavior.

In summary, the NAc receives convergent afferent inputs from limbic structures involved in processing information related to memory (i.e., hippocampus), drive, and motivation (i.e., BLA), higher cortical processing (i.e., PFC), as well as a rich dopaminergic input from the VTA, and in turn sends efferent

projections to areas involved in movement (e.g., ventral pallidum and associated regions). Given this anatomic organization, it was proposed by Gordon Mogenson and elaborated upon by others that the NAc functions as a site for the integration of limbic information related to memory, drive, and motivation, and the generation of goal-directed motor behaviors (termed 'limbic-motor integration'). Within this model, dopamine plays a key role in this process by functioning to modulate or 'gate' the transfer of these neural signals and thereby influence goal-directed behaviors including those involved in drug addiction.

Neurophysiological Investigations of Drug Addiction: Overview

The majority of electrophysiological studies on drug addiction examined cell firing within the NAc and associated regions during cocaine self-administration, although heroin and ethanol self-administration have also been studied to a lesser degree. Investigations initially examined NAc activity relative to the operant response (typically a lever press) for intravenous cocaine. However, since drug addiction in humans is often characterized by periods of drug intake followed by abstinence and relapse, additional studies have sought to examine the characteristics of cell firing related to this cycle of addictive behavior. Moreover, animal studies show that the resumption of drug-taking behavior following removal of drug availability can be elicited by factors such as re-exposure to the drug, stress, and exposure to drug-associated cues (e.g., audiovisual cues previously paired with drug infusion). As such, the responsiveness of NAc neurons to cocaine-associated stimuli and the effects of drug abstinence on NAc activity have also been addressed. These later studies are important since it is well known that in human drug addicts, exposure to drug-associated cues (such as drug paraphernalia) following periods of drug abstinence often lead to intense craving for the drug and resumption of drug taking (relapse). These topics are considered in the following sections.

Nucleus Accumbens Cell Firing during Cocaine Self-Administration

Initial electrophysiology studies sought to determine the firing characteristics of NAc neurons during drug-seeking behavior, particularly involving cocaine. In rodents, the beginning of cocaine self-administration sessions typically involves an initial drug 'load-up' period during which time animals press the lever 2–5 times in rapid succession within the first 5 min of the session. This is followed by a more stable pattern of responding for the remainder of the session, in which rats press the lever once approximately every 5 min (depending on cocaine dose) to maintain optimal levels of drug in their system. Electrophysiology studies revealed that following the initial load-up phase, NAc neurons exhibit patterned discharges (increases and/or decreases in cell firing) time-locked to lever-press responding for intravenous cocaine during self-administration sessions within two distinct time domains. Specifically, NAc cells exhibit significant changes in activity within seconds of the reinforced response for intravenous cocaine (termed short-term patterned discharges) and/or cycle on the order of minutes within the entire inter-infusion interval (termed long-term cyclic alterations).

With respect to short-term patterned activity, NAc neurons increase firing within seconds before the lever-press response, or increase or decrease activity immediately after response completion. These findings indicate that distinct populations of NAc neurons encode information about the important features of drug-seeking behavior including response initiation and completion. Interestingly, similar types of neuronal firing patterns have been observed relative to lever-press responding for 'natural' rewards such as water, food, or a sweet sucrose solution. Given the anatomic difference between the core and shell subregions of the NAc, investigators examined whether short-duration patterned discharges were restricted to either subregion. Studies revealed that similar types of neuronal firing patterns are observed within the core and shell of the NAc. However, West and colleagues have shown that neurons in the core exhibit a greater change in firing rate within seconds preceding the lever-press response for cocaine. These findings indicate that the core may be more involved in processing information about the execution of drug-seeking behaviors.

In addition to the patterned discharges observed within seconds of the reinforced response for cocaine, other studies have shown that some NAc neurons exhibit patterned activity across a much longer time period. For neurons exhibiting long-term cyclic alterations, cell firing is time-locked to the reinforced response for the drug but cycles on the order of minutes within the entire inter-infusion interval. This type of activity has been described in detail in a series of reports by Peoples and West. The authors revealed that a subset of NAc cells exhibits changes (primarily decreases) in cell firing approximately 1–2 min after the reinforced response and then reverse activity

(i.e., increase firing) later in the inter-infusion interval. For the majority of neurons, this reversal is gradual and varies positively with the inter-infusion interval (termed 'progressive reversal') and is not related solely to locomotion.

Since dopamine is a key neurotransmitter linked to cocaine addiction, the relationship between this chemical signal and NAc cell firing has been considered. Microdialysis studies have shown that dopamine levels rise in the NAc at the start of cocaine self-administration sessions and fluctuate within a smaller range during the inter-infusion interval. It has been proposed that the decline in extracellular dopamine during the interval may trigger subsequent self-administration behavior in order to maintain optimal levels of dopamine in the NAc. Interestingly, the time course of the progressive reversal described by Peoples and West mirrored the interval in which dopamine levels were shown to decrease in microdialysis studies. As such, it was proposed that the progressive reversal of NAc cells may constitute a neurophysiological mechanism by which decreasing drug levels (and presumably decreasing NAc DA levels) results in subsequent self-administration behavior. This is an intriguing proposal because it suggests that particular subsets of NAc neurons may be involved in the maintenance of cocaine self-administration behavior.

Dopamine may also play an important role in modulating the activity of NAc neurons that exhibit changes in cell firing within seconds of the reinforced response for cocaine (short-term patterned discharges). As noted above, anatomy studies show that dopamine nerve terminals are in a unique position to modulate cortical and subcortical input onto NAc cells. Furthermore, studies employing an electrochemical technique termed fast-scan cyclic voltammetry revealed rapid dopamine release within seconds of the reinforced response for cocaine, and relative to presentation of cocaine-associated cues (i.e., a tone–houselight stimulus that was always paired with cocaine infusion). These findings show that dopamine release occurs within the same time domain as short-duration patterned discharges of NAc neurons and may thereby function to modulate NAc activity during drug-seeking behavior. In this regard, one possibility that has been proposed is that dopamine functions to decrease background firing rates of NAc cells, thereby increasing the signal-to-noise ratio of those neurons. The net result would be an increase in the patterned discharge of NAc neurons relative to the lever-press response for the drug, or stronger neural encoding of drug-seeking behavior.

Role of Drug Abstinence on Accumbens Cell Firing during Cocaine Seeking

Cocaine addiction in humans is characterized by cycles of drug taking followed by periods of abstinence and relapse. One factor that has been linked with relapse in humans is exposure to environmental cues associated with prior drug-taking behavior. These cues have the ability to elicit intense craving for the drug and resumption of drug use. In order to better model this aspect of addiction, researchers have begun to examine additional factors controlling NAc cell firing as it relates to drug-seeking behaviors. For example, electrophysiology studies have determined that distinct populations of NAc neurons are activated by a discrete audiovisual (tone–houselight) stimulus that was always paired with the intravenous infusion of cocaine during self-administration in rats. As noted above, rapid dopamine release has also been observed relative to presentation of these same cocaine cues.

Other studies have examined the activity of NAc cell firing following periods of experimenter-imposed cocaine abstinence. For example, West and colleagues established a behavioral task incorporating a tone that signaled cocaine availability during self-administration sessions (i.e., each lever press resulted in cocaine delivery only during tone presentation). Next, cocaine availability was removed for 3–4 weeks. Thereafter, presentation of the same tone resulted in cocaine-seeking behavior, and significantly activated neurons in the shell but not the core of the NAc. Importantly, NAc shell neurons were not activated during presentation of a neutral tone never associated with cocaine. These findings reveal a persistent processing in the shell region of the NAc regarding information about cocaine-associated cues lasting several weeks after the last drug exposure.

Likewise, research completed in our laboratory examined NAc neuronal firing properties during the resumption of cocaine self-administration following 1 month of drug abstinence. We showed that the percentage of neurons that exhibit patterned discharges dramatically increased (nearly twofold) following 1 month abstinence. These findings were specific to the core but not the shell of the NAc. Likewise, the strength of the neural correlates (determined via signal-to-baseline ratios) also increased as a function of abstinence. Collectively, the results show that removal of drug availability (abstinence) from cocaine self-administration results in a dramatic increase in the number and strength of NAc neurons that encode cocaine-related information. Thus, neurons in the NAc, a key neural substrate in addiction,

become much more active following drug removal and subsequent resumption of drug-taking behavior.

Distinct Populations of Nucleus Accumbens Neurons Selectively Encode Goal-Directed Behaviors for Drug versus 'Natural' Rewards

A long-held view is that the brain reward system evolved to process information about 'natural' reinforcers such as food, water, and sex and that abused substances exert their reinforcing properties by 'hijacking' this neural circuit. Importantly, electrophysiological studies have shown that NAc neurons exhibit patterned discharges relative to reinforced responding for natural rewards such as food and water, similar to that observed during cocaine self-administration sessions. However, those initial studies were completed in different sets of animals; a critical issue addressed in a subsequent series of studies in our laboratory is whether the same neural circuit (i.e., the same neurons in the NAc) similarly process information about goal-directed behaviors for cocaine versus natural rewards. To address this issue, experiments were conducted to determine if NAc neurons that encode information about drug-seeking behaviors are the same neurons that are responsive during goal-directed behaviors for 'natural' rewards. This was accomplished by recording NAc activity during a multiple schedule of reinforcement for a natural reinforcer (water or food) and cocaine. Animals were trained to press one lever for a natural reward, followed by responding on another lever for intravenous cocaine during the same behavioral session. It was observed that in well-trained animals, cocaine activated a population of neurons in the NAc that were generally not responsive during lever-press behaviors for water reinforcement (only 8% showed similar firing patterns). Interestingly, this finding was observed from the very first day of cocaine exposure indicating that separate neural circuits for the encoding of goal-seeking behaviors involving cocaine versus natural rewards are not related to chronic drug experience.

Additional studies support the notion that separate neural circuits (or microcircuits) differentially encode information about drug versus natural rewards, at least at the level of the NAc. Differential firing patterns have also been observed during a multiple schedule for food reinforcement and cocaine self-administration in both rodents and monkeys. Furthermore, presentation of cocaine-associated stimuli has the capacity to selectively activate the subset of NAc neurons that function to process information related to cocaine-seeking behaviors but not those cells that encode information about water reinforcement. Moreover, the selective encoding of goal-directed behaviors is not unique for cocaine versus natural rewards but has also been observed during behavioral sessions involving consecutive self-administration of cocaine and heroin. Future studies are needed to establish whether NAc neurons that encode information about drug reward (cocaine or heroin) are the same neurons that process information about other types of 'natural' reinforcers such as those involving sexual behaviors.

Insights into the Functional Organization of the Nucleus Accumbens

Electrophysiological examination of single-cell activity in behaving animals allows for the discovery of a level of cellular organization in the NAc not revealed by other experimental approaches. As noted above, electrophysiological studies employing multiple schedules of reinforcement revealed that distinct populations of NAc neurons differentially process information about goal-directed behaviors for cocaine versus 'natural' (e.g., water or food) reward. Likewise, different types of abused substances (heroin and cocaine) appear to activate different groups of cells within the NAc. These findings indicate that discrete microcircuits in the NAc selectively process information about drug versus 'natural' reward. One view consistent with these findings that was proposed by Pennartz and colleagues is that the NAc is comprised of a collection of neuronal 'ensembles' or groups of cells with different functional properties. A particular ensemble is temporarily and synchronously activated by a particular set of excitatory inputs from the PFC, hippocampus, and/or BLA that in turn activate distinct NAc outputs and ultimately influence behavior. The activation of specific ensembles is modifiable depending upon reward-related learning processes. Clearly, electrophysiological recordings of single NAc neurons in animals engaged in goal-directed behaviors for drug versus natural rewards support this view.

Neurophysiological Investigations during Self-Administration of Other Abused Substances

The activation of NAc neurons during drug-seeking behavior is not unique to cocaine, but has also been observed during behaviors directed toward other abused substances including heroin and ethanol. NAc neurons exhibit increases or decreases in cell

firing relative to lever press responding for intravenous heroin with the time course and types of neuronal firing patterns similar to the short-term changes observed during self-administration of cocaine. Interestingly, similar types of patterned discharges have been recorded even though the behavioral changes induced by heroin are dramatically different from psychomotor stimulant administration (i.e., cocaine). It is not clear from studies employing heroin whether long-term cyclic alterations in cell firing are evident during the entire inter-infusion interval, as reported during cocaine self-administration. Nevertheless, these findings are consistent with the notion that the NAc is involved in mediating goal-directed behaviors for drug rewards including both psychomotor stimulants and opiates.

NAc neurons also exhibit patterned activation during ethanol self-administration. In this case, animals typically consume the drug orally, as opposed to intravenous infusion. It has been reported that a subset of NAc neurons exhibit changes (increases or decreases) in cell firing within seconds before and/or following the reinforced response for ethanol. Reinforced responses for alcohol typically involve either lever-press responses or nose-poking behavior (in which rats have to break a photo beam with their snout within a specific slot in the chamber). In addition, neurons showed patterned discharges during presentation of a tone stimulus that signaled the availability of ethanol. Collectively, these findings show that NAc neurons encode information about goal-directed behaviors across different drug classes (cocaine, heroin, ethanol) independent of type of response requirement (i.e., lever press vs. nose poke).

Behavioral Electrophysiology of the PFC and BLA

As noted above, the NAc is one component of a larger brain reward circuit and additional studies have been completed that examined cell firing during drug self-administration in regions that give rise to NAc afferents. For example, the NAc receives projections from the PFC and studies have considered how this input may influence NAc cell firing during drug-seeking behavior. It has been shown that distinct subsets of neurons in the PFC exhibit patterned discharges relative to the operant response for cocaine or heroin, similar to NAc activity. Moreover, Woodward and colleagues applied cross-correlational analysis of simultaneously recorded PFC, and NAc cells during cocaine self-administration sessions show that NAc neurons discharge before medial PFC neurons in the majority of cases. This correlated firing suggests that the NAc and PFC may participate in the control of cocaine self-administration with the NAc playing a key role in the initiation of this behavior. We have also shown that neurons in the BLA exhibit similar types of patterned discharges during cocaine self-administration to that observed for NAc cells, and are also activated by cocaine-associated cues. Collectively, these studies suggest that the encoding of drug-directed behaviors in the brain extends beyond the NAc to a more distributed, interconnected neural network involving additional structures such as the PFC and BLA. Since these neural substrates send afferents to the NAc, it is likely that the activation of NAc neurons during drug-seeking behavior is influenced by particular cortical and subcortical inputs onto NAc neurons.

Concluding Remarks

In summary, electrophysiological recordings conducted in behaving animals reveal important information about how neurons in the brain reward circuit process information about drug-seeking behaviors. When used in combination with the drug self-administration procedure, studies revealed that distinct populations of neurons in the NAc and associated regions encode the key features of drug-seeking behaviors. Moreover, studies show that neurons process information about environmental cues associated with drug intake and are influenced by factors such as drug abstinence known to affect drug addiction in humans. Collectively, these studies set the foundation for future work that can provide additional insight into neural processes underlying drug addiction in humans, perhaps leading to new pharmacotherapies to treat this chronic illness.

See also: Addiction: Neurobiological Mechanism; Alcoholism; Drug Addiction: Behavioral Pharmacology of Drug Addiction in Rats; Drug Addiction: Cellular Mechanisms; Drug Addiction: Neuroimaging; Drugs Addiction: Actions; Prefrontal Contributions to Reward Encoding; Reward Systems: Human; Substance Abuse and Dependence.

Further Reading

Carelli RM (2002) Nucleus accumbens cell firing during goal-directed behaviors for cocaine versus 'natural' reinforcement. *Physiology and Behavior* 76(3): 379–387.

Carelli RM and Wightman RM (2004) Functional microcircuitry in the accumbens underlying drug addiction: Insights from real-time signaling during behavior. *Current Opinion in Neurobiology* 14(6): 763–768.

Carelli RM, Williams JG, and Hollander J (2003) Basolateral amygdala neurons encode cocaine self-administration and

cocaine-associated cues. *Journal of Neuroscience* 23: 8204–8211.

Chang JY, Janak PH, and Woodward DJ (1998) Comparison of mesocorticolimbic neuronal responses during cocaine and heroin self-administration in freely moving rats. *Journal of Neuroscience* 18(8): 3098–3115.

Chang JY, Janak PH, and Woodward DJ (2000) Neuronal and behavioral correlations in the medial prefrontal cortex and nucleus accumbens during cocaine self-administration by rats. *Neuroscience* 99(3): 433–443.

Everitt BJ and Robbins TW (2005) Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience* 8(11): 1481–1489.

Ghitza UE, Fabbricatore AT, Prokopenko V, Pawlak AP, and West MO (2003) Persistent cue-evoked activity of accumbens neurons after prolonged abstinence from self-administered cocaine. *Journal of Neuroscience* 23: 7239–7245.

Ghitza UE, Fabbricatore AT, Prokopendo VF, and West MO (2004) Differences between accumbens core and shell neurons exhibiting phasic firing patterns related to drug-seeking behavior during a discriminative-stimulus task. *Journal of Neurophysiology* 92: 1608–1614.

Hollander JA and Carelli RM (2005) Abstinence from cocaine self-administration heightens neural encoding of goal-directed behaviors in the accumbens. *Neuropsychopharmacology* 30: 1464–1474.

Janak PH, Chang JY, and Woodward DJ (1999) Neuronal spike activity in the nucleus accumbens of behaving rats during ethanol self-administration. *Brain Research* 817(1–2): 172–184.

Mogenson GF (1987) Limbic-motor integration. *Progress in Psychobiology and Physiological Psychology* 12: 117–169.

Pennartz CM, Groenewegen HJ, and Lopes da Silva FH (1994) The nucleus accumbens as a complex of functionally distinct neuronal ensembles: An integration of behavioral, electrophysiological and anatomical data. *Progress in Neurobiology* 42(6): 719–761.

Peoples LL and West MO (1996) Phasic firing of single neurons in the rat nucleus accumbens correlated with the timing of intravenous cocaine self-administration. *Journal of Neuroscience* 16 (10): 3459–3473.

Rebec GV and Sun W (2005) Neuronal substrates of relapse to cocaine-seeking behavior: Role of prefrontal cortex. *Journal of Experimental Analysis of Behavior* 84(3): 653–666.

Zahm DS (2000) An integrative neuroanatomical perspective on some subcortical substrates of adaptive responding with emphasis on the nucleus accumbens. *Neuroscience and Biobehavioral Reviews* 24(1): 85–105.

Relevant Website

http://www.nida.nih.gov – National Institute on Drug Abuse.

Drug Addiction: Behavioral Pharmacology of Drug Addiction in Rats

A Badiani, University of Rome Sapienza, Rome, Italy
T E Robinson, University of Michigan, Ann Arbor, MI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The enormous advances made possible by developments in molecular pharmacology, and the availability of refined preparations that provide exquisite control over many variables, make it very tempting to try and reduce the effects of psychoactive drugs to their direct actions on a receptor, following which a fixed series of intracellular reactions is initiated. The thesis of this article, however, is that the behavioral and psychological effects of psychoactive drugs cannot be reduced to the straightforward consequences of ligand–receptor binding, and attempts to do so neglect the complexity of drug actions. This is because, as put forward by Falk and Feingold, nearly two decades ago, "the behavioral action of a drug [is due to] a cluster of interactions among the pharmacological substance, the individual set of the organism, and the environmental setting"; that is, the effects of drugs on behavioral and psychological functions are not only a consequence of the pharmacological effects of the drug, but also on the context in which the drug is experienced.

Of course, it has long been known that the physical characteristics of the environment (e.g., in laboratory settings, the size and shape of cage, type of bedding, etc.) can have a large effect on the behavioral actions of drugs. However, of particular importance here is the notion that even when the physical characteristics of the environment are held constant, apparently subtle differences in the 'psychological context' powerfully modulate both the acute effects of psychoactive drugs on brain and behavior and, with repeated administration, the ability of these drugs to change the brain. This general theme is illustrated here by discussing a series of studies by Aldo Badiani, Terry Robinson, and colleagues concerning how the neurobehavioral effects of psychostimulant and opioid drugs are modulated by the circumstances surrounding their administration.

Environmental Modulation of the Behavioral Effects of Addictive Drugs

In most of the studies discussed here, doses of amphetamine, cocaine, or morphine were administered to rats in one of two conditions. Rats in one group were transported each day from distinct cages in which they lived to test cages, at which time they received drug treatments. Rats in another group lived in the test cages in which they received the drugs. Thus, rats in both groups received drug treatments in physically identical environments, but for one group this was a relatively 'novel' environment (so we call this the 'novel group') and for the other group it was 'home' (so we call this the 'home group'). This apparently minor difference in environmental context has a profound influence on the initial (acute) psychomotor-activating effects of amphetamine (Figures 1 and 4), morphine, and heroin, but relatively little influence on cocaine effects. The psychomotor-activating effects of these drugs are much greater when they are given in a novel environment than when they are given in the home cage.

Perhaps even more importantly, environmental context not only modulates the acute psychomotor-activating effects of drugs, but also alters the ability of repeated exposure to these drugs to produce behavioral sensitization, which may be crucial in the transition to addiction, as hypothesized by Terry Robinson and Kent Berridge. Behavioral sensitization refers to an increase in a drug effect with repeated drug administration and is thought to be due to drug-induced neuroplastic adaptations in brain systems that mediate drug sensitization effects, which include psychomotor-activating and incentive-motivational effects. When given in association with environmental novelty, repeated treatments with amphetamine, cocaine, morphine, or heroin produce much greater psychomotor sensitization than when these drugs are administered in the home cage. This effect is particularly striking when the treatments are administered via intravenous catheters activated by remote control, so that the animals in the home groups are almost completely deprived of any cue that could signal drug delivery. Under these conditions, repeated administration of low doses of amphetamine, cocaine, or morphine produced robust sensitization only in the novel group (Figure 1). At higher doses, however, these drugs produce psychomotor sensitization regardless of environmental context, indicating that the effect of environmental context is not to gate sensitization in an all-or-nothing way, but rather to modulate susceptibility to this particular form of drug-induced behavioral plasticity. It is not clear what accounts for this effect of environment on susceptibility to sensitization, but it does not appear to be due to group differences in associative learning.

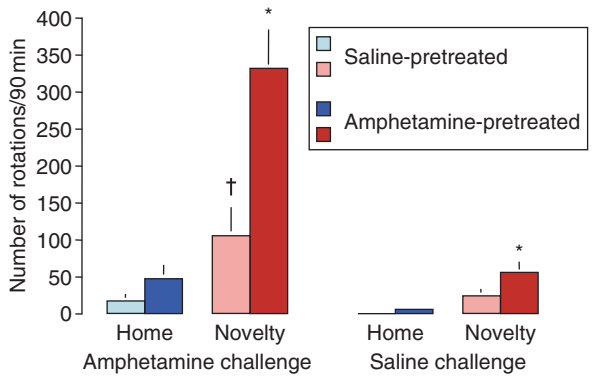

Figure 1 Environmental modulation of the psychomotor response to an intravenous (i.v.) challenge with 1.0 mg kg^{-1} of amphetamine or saline in rats that had previously received repeated i.v. infusions of either saline or 1.0 mg kg^{-1} of amphetamine and were then withdrawn from treatment for 4 days. For the differences between home and novelty groups see text. Data from Crombag HS, Badiani A, and Robinson TE (1996) Signalled versus unsignalled intravenous amphetamine: Large differences in the acute psychomotor response and sensitization. *Brain Research* 722: 227–231.

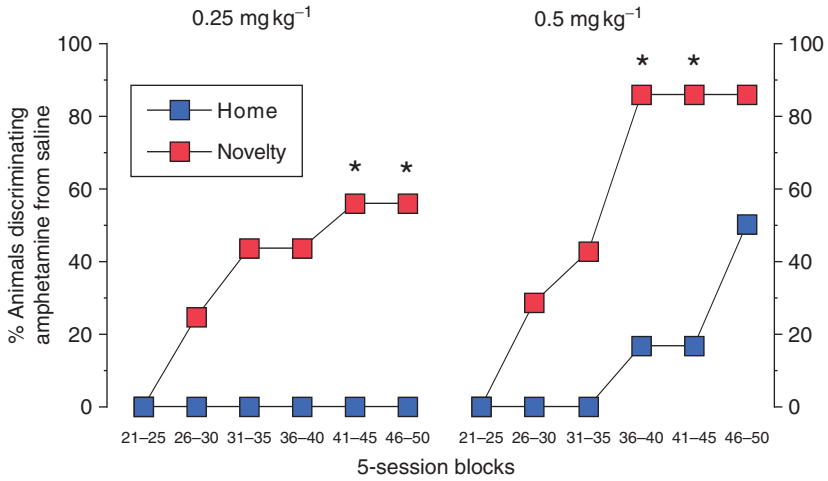

Figure 2 Environmental modulation of the ability of rats to discriminate amphetamine (0.25 or 0.5 mg kg^{-1}, intraperitoneal (i.p.)) from saline in a water-reinforced operant procedure (FR30). For the differences between home and novelty groups see text. Data from Paolone G, Palopoli M, Marrone MC, et al. (2004) Environmental modulation of the interoceptive effects of amphetamine in the rat. *Behavioural Brain Research* 152: 149–155.

This environmental manipulation not only influences the psychomotor effects of drugs, but also modulates the discriminative effects of amphetamine (Figure 2). It is important, however, to point out that other drug effects do not appear to be modulated in the same way by environmental context. For example, environmental novelty has only a modest effect on morphine-induced eating, and virtually no effect on the ability of morphine to suppress phasic pain from radiant heat in the tail-flick test (Figure 3).

Environmental Modulation of Immediate-Early Gene Expression

A simple difference in environmental context presumably has a large effect on the acute psychomotor-activating effects of some drugs and on their ability to induce behavioral sensitization because environmental context alters in some way the neurobiological impact of drugs. The neuropharmacological effect of amphetamine, cocaine, and morphine that is thought to be responsible for their psychomotor-activating effects is their ability to increase dopamine neurotransmission in the dorsal and ventral striatum. However, despite the large effects of environmental context on behavior, it had no effect on the ability of amphetamine to increase dopamine release in the dorsal or ventral striatum, as indicated by microdialysis studies (Figure 4).

Addictive drugs, such as morphine, cocaine, and amphetamine, also induce the expression of the gene c-*fos* (encoding the Fos protein) and other immediate-early genes (IEGs) in a number of forebrain regions, including the caudate nucleus and the nucleus

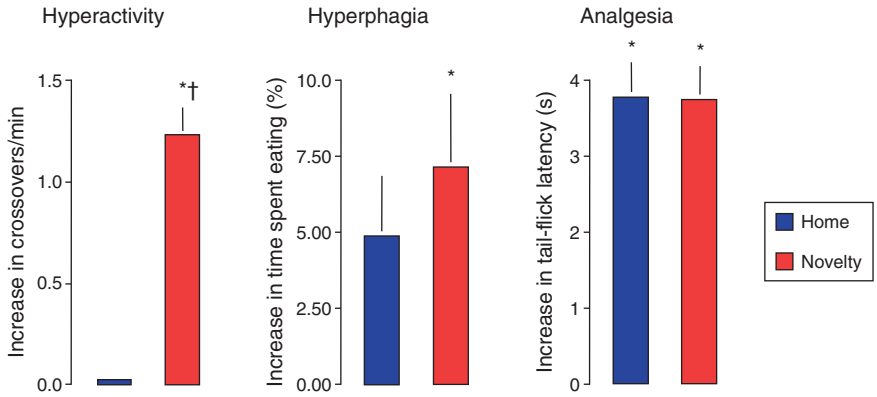

Figure 3 Environmental modulation of morphine-induced hyperactivity, hyperphagia, and analgesia. Animals received repeated i.p. injections of either saline or morphine (4.0 mg kg^{-1}). Each bar indicates the average difference between the morphine and the saline group. The animals were tested under either home or novelty conditions (see text for details). Data from Paolone G, Palopoli M, Marrone MC, et al. (2004) Environmental modulation of the interoceptive effects of amphetamine in the rat. *Behavioural Brain Research* 152: 149–155.

accumbens. This is important because IEGs can serve as indicators of neuronal activity and are also thought to represent an initial step in initiating experience-dependent plasticity. Interestingly, studies on the influence of environmental context on drug-induced IEG expression using *in situ* hybridization have revealed that the ability of amphetamine, cocaine, or morphine to induce c-*fos* or *arc* mRNA expression varies greatly, depending on the environmental context (home vs. novel) in which the drugs are administered. Although a relationship between drugs and environmental novelty is detected in almost every cortical and subcortical structure that has been examined, the most interesting changes are noted to occur in the striatum and in the amygdala, two brain regions implicated in incentive motivation, reward, and learning.

Striatum

Environmental novelty enhances the ability of amphetamine and cocaine to induce c-*fos* or *arc* mRNA expression in the caudate-putamen, and the interaction is maximal in the caudal portions of the caudate, where c-*fos* mRNA levels are nearly two times greater than would be predicted by the simple addition of the effects of amphetamine and mere exposure to a novel environment (Figure 5). In the ventral striatum the effects of novelty and amphetamine (or cocaine) are infra-additive and there is no significant rostrocaudal gradient. Interestingly, in the nucleus accumbens the effect of mere exposure to a novel environment on Fos mRNA expression is greater than that produced by an injection of amphetamine or cocaine in the home setting, suggesting that the nucleus accumbens is especially responsive to environmental novelty.

Most importantly, in the striatum the effect of environmental context on drug-induced IEG expression does not simply represent a quantitative effect, but the context in which these drugs are administered determines which specific cells and circuits are engaged. In the striatum over 90% of the neurons are γ-aminobutyric acid (GABA)ergic medium spiny projection neurons. Approximately half of these neurons co-express mRNAs for pre-prodynorphin, substance P, and the dopamine D1 receptor and project directly to the output nuclei of basal ganglia (substantia nigra pars reticulata and the external globus pallidus), forming the so-called direct pathway. The other half of these projection neurons co-express pre-proenkephalin and D2 receptor mRNAs and modulate basal ganglia output nuclei via an 'indirect pathway' by providing an inhibitory input to the internal globus pallidus, which in turn inhibits the subthalamic nucleus (STN). When given at home, amphetamine, cocaine, or morphine induces robust Fos mRNA expression almost exclusively in dynorphin/D1-positive neurons. In contrast, when amphetamine (or cocaine) is given in association with environmental novelty it induces robust Fos mRNA expression in both dynorphin/D1 and enkephalin/D2 neurons, as well as in the STN (Figure 6). These findings indicate that which cells and circuits are engaged with psychostimulant drugs (and morphine) depends to a large extent on the environmental context in which the drugs are experienced.

The mechanism by which environmental novelty modulates the cells and circuits engaged by drugs is not known. It is known that mere exposure to a novel environment powerfully induces Fos mRNA expression in most cortical areas, and activation of corticostriatal projections induces c-*fos* expression

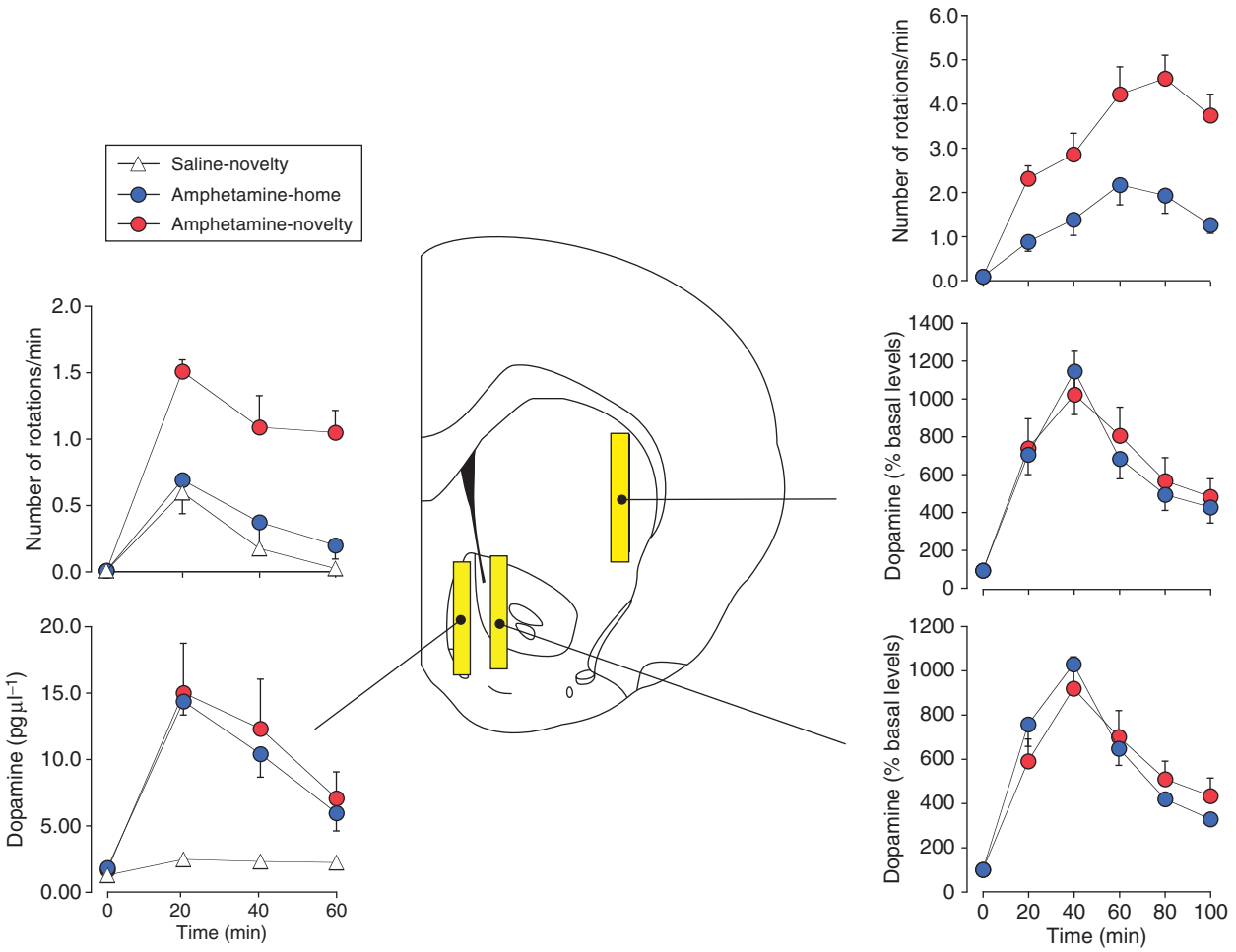

Figure 4 Environmental modulation of the psychomotor and neurochemical (dopamine release) response to an acute administration of amphetamine (0.5 mg kg^{-1}, i.v. for the left panels; 200 mg kg^{-1}, i.p. for the right panels). Three groups of rats were implanted with microdialysis probes aimed at one of the three regions of the striatal complex. Amphetamine was administered under either home (amphetamine group) or novelty (amphetamine + novelty group) conditions (see text for details). Some animals received saline in association with novelty (novelty group). Data from Badiani A, Oates MM, Day HEW, et al. (1998) Amphetamine-induced behavior, dopamine release and c-*fos* mRNA expression: Modulation by environmental novelty. *Journal of Neuroscience* 18: 10579–10593; and Badiani A, Oates MM, Fraioli S, et al. (2000) Environmental modulation of the response to amphetamine: Dissociation between changes in behavior and changes in dopamine and glutamate overflow in the rat striatal complex. *Psychopharmacology (Berlin)* 151: 166–174.

specifically in enkephalin/D2-positive neurons in the striatum. Thus, a working hypothesis is that novelty modulates drug-induced c-*fos* expression in the striatum via glutamatergic projections from the cortex to the caudate-putamen and to the STN. In support of this, *N*-methyl-D-aspartate (NMDA) receptor antagonists preferentially reduce amphetamine-induced Fos mRNA expression in enkephalin/D2-positive neurons, as does transection of corticostriatal fibers.

Amygdala

Other areas of brain regions where novelty and addictive drugs interact in a very complex manner are the amygdala and the bed nucleus of the stria terminalis (BST). Novelty, but not amphetamine or cocaine, induces Fos mRNA expression in the medial amygdala (MEA) and medial posterior nucleus of the BST (BSTmp). Induction of c-*fos* expression in these areas has been previously observed in response to some stressors but it is unlikely that the hypothalamic–pituitary–adrenal (HPA) axis is implicated in this effect of novelty, because amphetamine, at the same doses that fail to enhance c-*fos* expression in the MEA/BSTmp, produces plasma corticosterone levels comparable to those produced by novelty.

Both novelty and psychostimulants induce Fos mRNA expression in the basolateral amygdala, which has been implicated in the response to affective stimuli

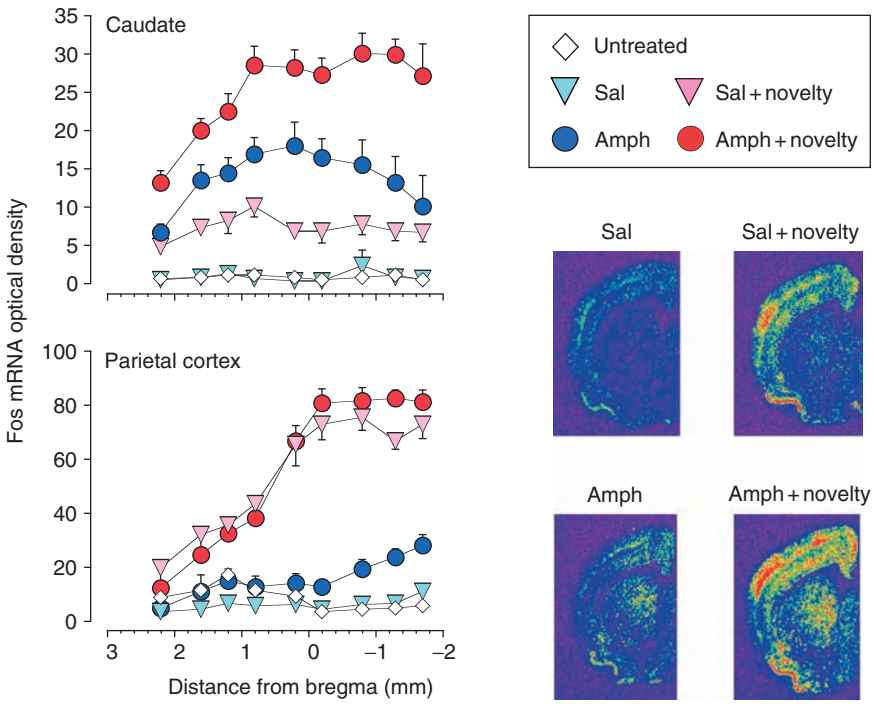

Figure 5 Environmental modulation of saline- and amphetamine-induced (2.0 mg kg^{-1}, i.p.) Fos mRNA expression in the rat caudate and parietal cortex. The figure also shows representative densitograms. The animals were tested under either home (untreated, sal, and amph groups) or novelty (sal + novelty and amph + novelty groups) conditions (see text for details). Progressively greater intensity of the signal from the hybridized ^{35}S-riboprobe for Fos mRNA is indicated by the transition from blue to yellow to red. Data from Badiani A, Oates MM, Day HEW, et al. (1998) Amphetamine-induced behavior, dopamine release and c-*fos* mRNA expression: Modulation by environmental novelty. *Journal of Neuroscience* 18: 10579–10593.

and in conditioned reinforcement. When amphetamine or cocaine is given in association with novelty, Fos mRNA expression is further enhanced, a pattern similar to that observed in the caudate.

Amphetamine or cocaine administered in the home setting induces robust Fos mRNA expression in the central amygdala (CEA) and oval subnucleus of the BST (BSTov), whereas novelty produces only a modest effect. Quite surprisingly, when amphetamine (or cocaine) is given in association with novelty, the Fos response in these structures is much smaller than that produced by amphetamine (or cocaine) at home. This most unusual pattern is fundamentally different from what is observed in all other brain regions examined so far, and the reason for it is still unclear.

Neocortex

At least from a purely quantitative point of view the interaction between novelty and the effect of amphetamine in the neocortex and septum appears to be relatively simple. In these regions, as seen for the MEA and BSTmp, novelty produces much more robust c-*fos* expression than does exposure to amphetamine alone, and the effect of novelty is comparable in magnitude to that of amphetamine plus novelty (Figure 5). Double *in situ* hybridization studies are required to determine whether the lack of quantitative differences in c-*fos* expression between the novelty and amphetamine plus novelty groups masks a differential activation of distinct populations of neurons.

Environmental Modulation of Drug-Induced Structural Neuroplasticity

The influence of environmental context on the ability of drugs to induce psychomotor sensitization is not only interesting with respect to understanding how drug–environment interactions modulate drug effects, but it also provides a unique tool to explore whether any given drug-induced change in the nervous system is associated with drug-induced sensitization, or occurs as a function of mere drug history. Indeed, in the extensive literature on the neurobiology of sensitization there are almost no examples of attempts to dissociate the effects of mere drug history from the ability of drugs to induce sensitization. In sensitization experiments typically drugs are given

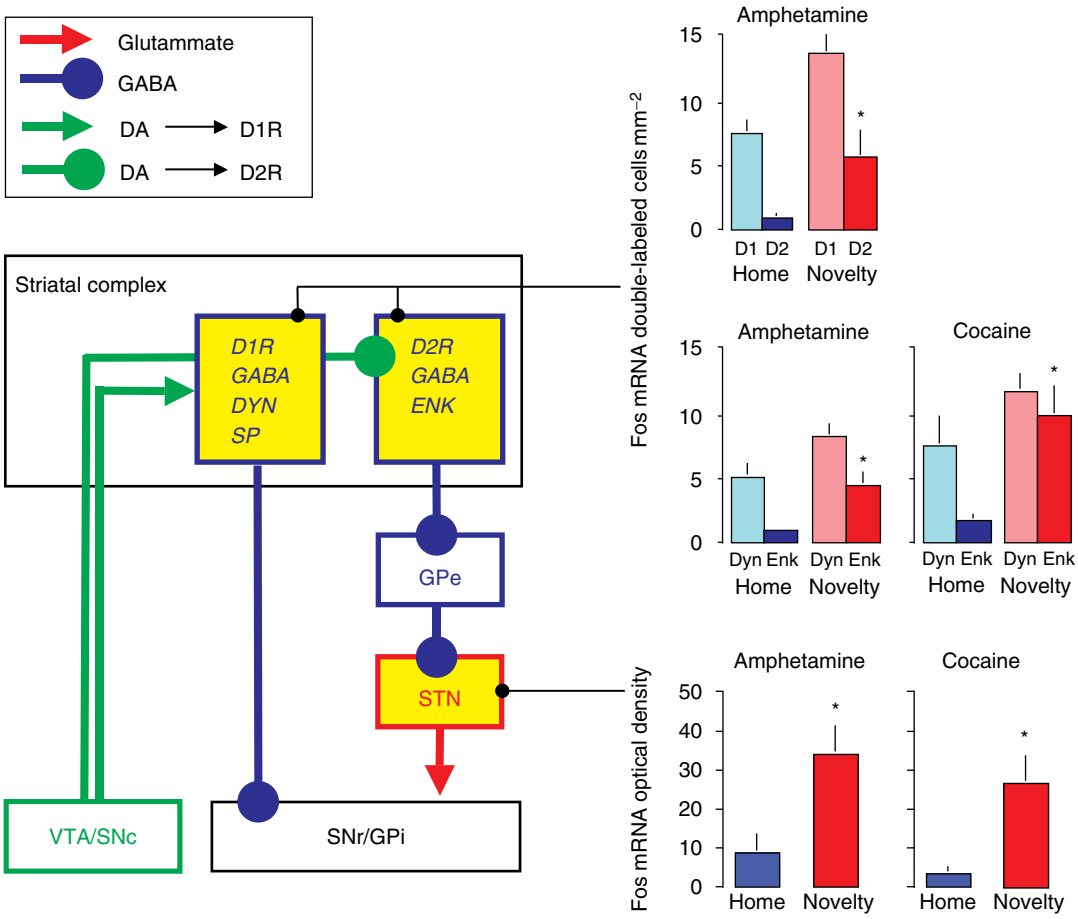

Figure 6 Effect of environment on amphetamine- and cocaine-induced Fos mRNA expression in the neurons of the direct vs. the indirect projection system of the rat caudate. These two subpopulations of striatal neurons were identified by examining the co-expression of markers for either D1 vs. D2 receptor mRNAs or prodynorphin vs. proenkephalin mRNAs. For the differences between home and novelty groups see text. Data from Badiani A, Oates MM, Day HEW, et al. (1999) Environmental modulation of amphetamine-induced c-*fos* expression in D1 versus D2 striatal neurons. *Behavioural Brain Research* 103: 203–209; and Uslaner JM, Badiani A, Norton CS, et al. (2001) Amphetamine and cocaine induce different patterns of c-fos mRNA expression in the striatum and subthalamic nucleus depending on environmental context. *European Journal of Neuroscience* 13: 1977–1983.

repeatedly in a manner that induces behavioral sensitization, and then brains are examined to determine if the drug treatment regimen altered the neurobiological measure of interest. However, such studies do not reveal whether drug treatment would have altered the molecule/measure of interest whether drug treatment induced sensitization or not.

However, the effect of environmental context as described herein provides an interesting way to hold drug history constant while varying drug treatment, to see whether drug treatment induces sensitization. The drug history of an animal in its home setting versus in a novel setting is constant, but within a given dose range only drug treatments in the novel setting induce the observed form of behavioral plasticity. If any given drug-induced neuroadaptation is related to the development of sensitization, it should be seen in the novel group but not the home group. This procedure has been used recently to determine whether one neural correlate of sensitization is specifically related to the development of sensitization or to mere drug history.

Terry Robinson, Bryan Kolb, and their colleagues have reported that repeated intermittent treatments with amphetamine or cocaine, treatments that induce behavioral sensitization, also alter synaptic organization in the nucleus accumbens and the prefrontal cortex, as indicated by changes in dendritic branching and in the number of dendritic spines on cells in these brain regions. It is not clear from these studies, however, whether psychostimulant drugs would produce this form of structural plasticity even if they failed to produce behavioral sensitization. This question was examined by treating animals with repeated injections of cocaine, either in the home environment or in a relatively novel test environment (as described

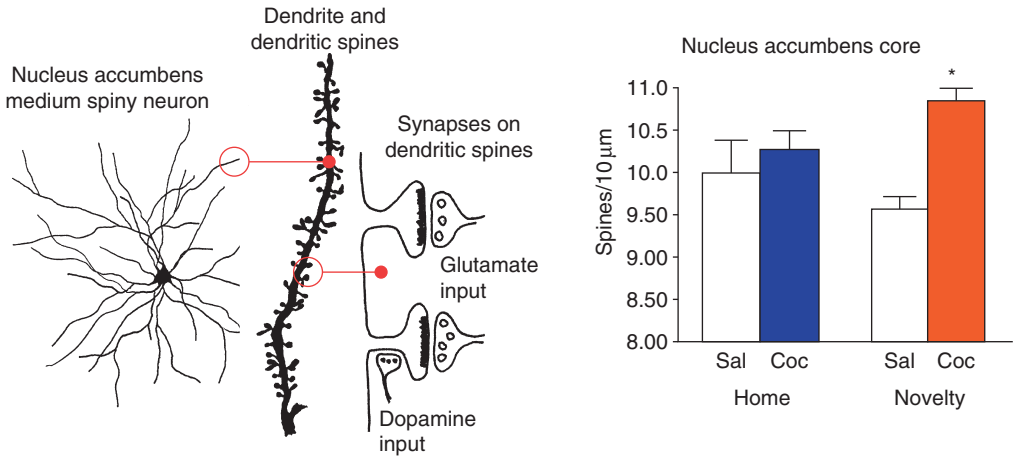

Figure 7 Effect of environment on the ability of repeated cocaine to induce structural plasticity in the shell of the nucleus accumbens. For the differences between home and novelty groups see text. Data from Li Y, Kolb B, and Robinson TE (2003) The location of persistent amphetamine-induced changes in the density of dendritic spines on medium spiny neurons in the nucleus accumbens and caudate-putamen. *Neuropsychopharmacology* 28: 1082–1085.

previously). Under the conditions of one experiment a cocaine treatment regimen (15 mg kg^{-1} daily for 8 days) was used that induced robust behavioral sensitization in the novel condition, but no significant sensitization in the home condition. Two weeks after the last injection the brains were examined for evidence of structural plasticity.

It was found that cocaine treatment increases spine density on medium spiny neurons in the 'core' of the nucleus accumbens only in the novel group (i.e., only in the group that was sensitized) (Figure 7). This is consistent with the hypothesis that drug-induced structural plasticity in the core of the accumbens is related to the development of psychomotor sensitization. In contrast, cocaine treatment increases spine density in the 'shell' of the accumbens in both the novel and the home groups (i.e., independent of whether drug treatment produced sensitization). Thus, in the shell this particular form of structural plasticity is associated with mere drug history, and does not require that drug treatments also induce sensitization. Finally, as indicated previously, the effect of environmental context is dose dependent, and as dose is increased amphetamine and cocaine induce sensitization even when given at home. It is important, therefore, that repeated treatments with a higher dose of cocaine (30 mg kg^{-1}), a dose that induces psychomotor sensitization even at home, produce structural plasticity in the core of the accumbens, even though treatments are given in the home cage.

Thus, the context in which cocaine is experienced not only modulates its ability to induce behavioral sensitization, but also its ability to induce structural plasticity in the core (but not shell) of the nucleus accumbens. Varying the context in which drugs are experienced provides a powerful way to dissociate those drug-induced neuroadaptations that are associated with the development of this form of behavioral plasticity from those that occur as a function of mere drug history.

Conclusions

Two important conclusions can be drawn from the findings on the environmental context of drug addiction. The first one is a cautionary note for neuroscientists who may be tempted to interpret behavioral data obtained under conditions similar to those of our 'novel' groups (as is customary in behavioral pharmacology) on the basis of molecular biology data obtained under conditions that are similar to those of our 'home' groups. The fact is, the behavioral and neurobiological effects of drugs vary qualitatively depending on the context in which they are experienced; therefore, to understand how drugs influence behavior and psychological function, they need to be studied 'in context.'

The second conclusion is that environmental context can exert a powerful modulatory influence on the ability of drugs to induce the neuroplastic adaptations implicated in drug addiction. Thus, an understanding of how repeated drug use sometimes leads to addiction will also require considering the environment in which drugs are experienced. Recent data from Aldo Badiani's laboratory indicating that environmental context has a powerful influence on the acquisition of cocaine self-administration behavior in the rat (Figure 8) further emphasize the importance of this issue for addiction research.

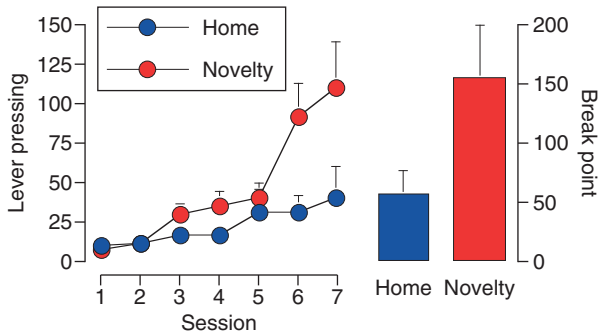

Figure 8 Environmental modulation of cocaine self-administration (0.2–0.4 mg kg^{-1} per infusion). The number of consecutive responses required to obtain, on a fixed ratio schedule, a single infusion was raised from 1 (sessions 1–2) to 2 (sessions 3–5) and then to 5 (sessions 6–7). The right panel illustrates the maximum number of responses emitted in a break-point procedure for the same doses of cocaine. The number of responses required to obtain a single infusion was increased within the session according to the following progression: 5, 10, 20, 30, 50, 70, 100, 150, 200, 300, 500, etc. Data from Caprioli D, Paolone g, Celentano M, Testa A, Nencini P, and Badiani A (2007) Environmental modulation of cocaine self-administration in the rat. *Psychopharmacology* 192: 397–406.

See also: Addiction: Neurobiological Mechanism; Amphetamines; Appetitive Systems: Amygdala and Striatum; Drug Addiction: Behavioral Neurophysiology; Drug Addiction: Neuroimaging; Drugs Addiction: Actions; Stimulant and Wake-Promoting Substances; Striatum: Internal Physiology; Substance Abuse and Dependence.

Further Reading

Badiani A and Robinson TE (2004) Drug induced neurobehavioral plasticity: The role of environmental context. *Behavioural Pharmacology* 15: 327–339.

Badiani A, Anagnostaras SG, and Robinson TE (1995) The development of sensitization to the psychomotor stimulant effects of amphetamine is enhanced in a novel environment. *Psychopharmacology (Berlin)* 117: 443–452.

Badiani A, Oates MM, Day HEW, et al. (1998) Amphetamine-induced behavior, dopamine release and c-*fos* mRNA expression: Modulation by environmental novelty. *Journal of Neuroscience* 18: 10579–10593.

Badiani A, Oates MM, Day HEW, et al. (1999) Environmental modulation of amphetamine-induced c-*fos* expression in D1 versus D2 striatal neurons. *Behavioural Brain Research* 103: 203–209.

Badiani A, Oates MM, Fraioli S, et al. (2000) Environmental modulation of the response to amphetamine: Dissociation between changes in behavior and changes in dopamine and glutamate overflow in the rat striatal complex. *Psychopharmacology (Berlin)* 151: 166–174.

Caprioli D, Paolone G, Celentano M, Testa A, Nencini P, and Badiani A (2007) Environmental modulation of cocaine self-administration in the rat. *Psychopharmacology (Berlin)* (in press).

Crombag HS, Badiani A, and Robinson TE (1996) Signalled versus unsignalled intravenous amphetamine: Large differences in the acute psychomotor response and sensitization. *Brain Research* 722: 227–231.

Falk JL and Feingold DA (1987) Environmental and cultural factors in the behavioral actions of drugs. In: Meltzer HY (ed.) *Psychopharmacology: The Third Generation of Progress*, pp. 1503–1510. New York: Raven Press.

Ferguson SM, Norton CS, Watson SJ, et al. (2003) Amphetamine-evoked c-fos mRNA expression in the caudate-putamen: The effects of DA and NMDA receptor antagonists vary as a function of neuronal phenotype and environmental context. *Journal of Neurochemistry* 86: 33–44.

Li Y, Kolb B, and Robinson TE (2003) The location of persistent amphetamine-induced changes in the density of dendritic spines on medium spiny neurons in the nucleus accumbens and caudate-putamen. *Neuropsychopharmacology* 28: 1082–1085.

Paolone G, Burdino R, and Badiani A (2003) Dissociation in the modulatory effects of environmental novelty on the locomotor, analgesic, and eating response to acute and repeated morphine in the rat. *Psychopharmacology (Berlin)* 166: 146–155.

Paolone G, Conversi D, Caprioli D, et al. (2007) Modulatory effect of environmental context and drug history on heroin-induced psychomotor activity and Fos protein expressing in the rat brain. *Neuropsychopharmachology* (in press).

Paolone G, Palopoli M, Marrone MC, et al. (2004) Environmental modulation of the interoceptive effects of amphetamine in the rat. *Behavioural Brain Research* 152: 149–155.

Robinson TE and Berridge KC (1993) The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews* 18: 247–291.

Robinson TE and Kolb B (2004) Structural plasticity associated with exposure to drugs of abuse. *Neuropharmacology* 47(supplement 1), 33–46.

Uslaner JM, Badiani A, Norton CS, et al. (2001) Amphetamine and cocaine induce different patterns of c-fos mRNA expression in the striatum and subthalamic nucleus depending on environmental context. *European Journal of Neuroscience* 13: 1977–1983.

Drug Addiction: Cellular Mechanisms

A Bonci and F W Hopf, University of California at San Francisco, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Addiction and Synaptic Plasticity

Glutamate Receptors as Targets for Addictive Drugs

There is abundant evidence that dopamine represents a key molecule for the development and expression of addictive behaviors. However, a growing body of evidence suggests that the excitatory neurotransmitter glutamate plays a central role during drug-dependent learning and other adaptive processes that ultimately lead to substance abuse, to modulate both ventral tegmental area (VTA) neurons and the limbic system. When glutamate is released from presynaptic terminals, it activates both ionotropic receptors – for example, receptors (Rs) for α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), kainate, and *N*-methyl-D-aspartate (NMDA) – and metabotropic glutamate receptors, and a variety of studies have shown that inhibition of any of these receptors interferes with both the development and the expression of addictive behaviors. For example, AMPAR antagonists or NMDAR antagonists impair lever pressing for cocaine. Further, morphine- and psychostimulant-induced place preference are disrupted as well. Similarly, AMPAR antagonists inhibit lever pressing for cocaine, amphetamine, and alcohol administration and impair amphetamine- and morphine-induced place preference. Along the same line, the development of behavioral sensitization (a behavioral index of craving) to cocaine, amphetamine, and morphine is blocked by NMDAR antagonists whereas AMPAR antagonists affect the maintenance of sensitization. These data support the hypothesis that all major drugs of abuse initiate abnormal learning mechanisms via activation of NMDARs and AMPARs, which ultimately leads to drug-dependent behaviors.

Long-Term Potentiation: A Cellular Substrate for Abnormal Learning and Memory Processes

Long-term potentiation (LTP) and long-term depression (LTD) were first described in the hippocampus many years ago. Unlike classical synaptic events that last from a few milliseconds to a few hundreds of milliseconds at most, LTP and LTD of AMPA receptors can last for hours, days, and perhaps longer, and represent the most studied forms of synaptic plasticity. The main reason LTP has received so much attention over the past 20 years is because it has been suggested as a key substrate of normal learning and memory processes. LTP and LTD typically have a rapid induction phase – which is in many cases, but not exclusively, NMDAR dependent – and a maintenance phase involving changes in glutamate release and/or AMPAR turnover and sensitivity. The temporal properties of LTP and LTD make them contenders for mediating rapid onset and persistent neural adaptations in the reward circuitry.

Both LTP and LTD have been identified in VTA dopaminergic cells. LTP is NMDAR dependent, whereas LTD is NMDAR independent, although LTD is blocked by activation of somatodendritic D_2 dopamine receptors present on VTA dopaminergic neurons. Subsequent studies have shown that amphetamine blocks LTD, whereas *in vivo* cocaine exposure as well as exposure to most other major drugs of abuse promotes LTP-like potentiation of AMPAR-mediated synaptic responses in VTA dopaminergic cells (Figure 1). Thus, the consequence of even a single exposure to an addictive drug is a persistent increase in the efficacy of glutamatergic synaptic transmission in VTA dopaminergic cells via potentiation of AMPARs. In the first study to provide evidence that *in vivo* exposure to an addictive drug (cocaine) leads to potentiation of AMPAR synaptic responses, the effect was observed for at least 5 days but less than 10 days following a single exposure. Importantly, the magnitude of cocaine-induced potentiation of AMPAR responses correlates with the magnitude of cocaine-induced locomotion, and, as with behavioral sensitization, cocaine-induced potentiation can be blocked by NMDAR antagonists. However, although these studies suggest that cocaine induces transient synaptic changes in VTA neurons, the results of the studies also imply that plasticity at excitatory synapses in the VTA perhaps plays a role in the initiation, but not the maintenance, of the addiction process, based on examination of synaptic plasticity in the VTA by using noncontingent, experimenter-administered cocaine procedures.

Because drug taking is a learned process, it is important to understand whether voluntary drug intake and passive drug exposure produce similar durations of synaptic plasticity. Further, since drugs of abuse are hypothesized to usurp learning processes in the motivational circuit, it is also important to determine whether voluntary intake of natural rewards produces changes in cellular functions similar to those produced by drugs of abuse. For this reason, our laboratory examined glutamatergic synaptic transmission in the VTA of animals trained to self-administer cocaine or food, as well as in the VTA of animals

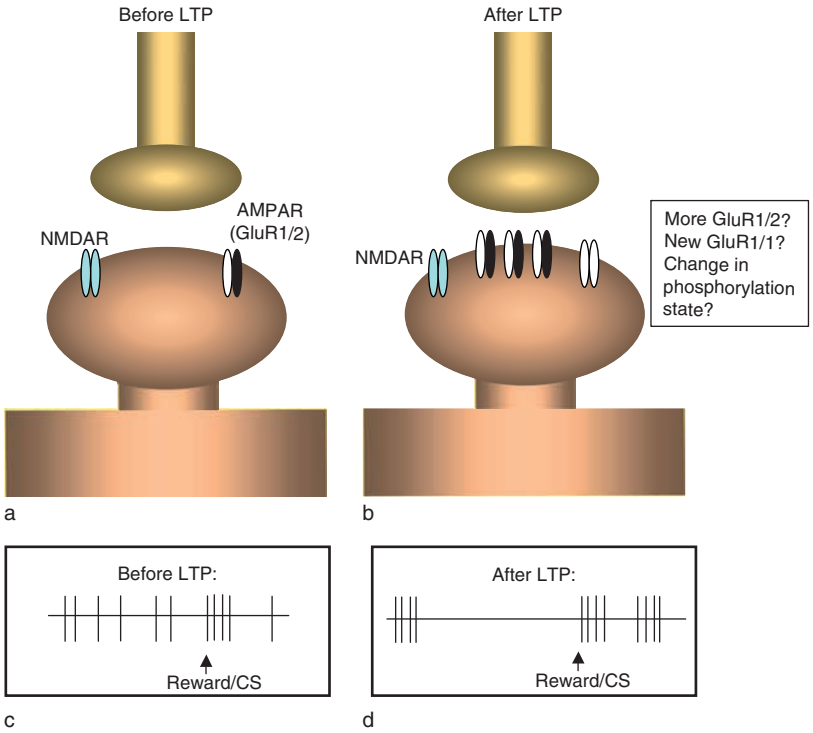

Figure 1 Hypothetical α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor (AMPAR) modifications as a consequence of long-term potentiation (LTP). A ventral tegmental area excitatory synapse is expanded to indicate potential changes in glutamate receptor (GluR) subunit expression. (a) GluR1 subunits before LTP (NMDAR, *N*-methyl-D-aspartate receptor). (b) GluR subunit expression following LTP in response to addictive drugs. (c) Representation of typical dopaminergic cell firing pattern, both under control conditions of background firing (irregular tonic firing) and with a firing pattern in response to a reward or to a reward-associated (conditioned) stimulus (CS; phasic firing of high-frequency bursts). (d) Proposed changes in firing patterns following drug-induced LTP, when less tonic irregular firing and more burst firing might be observed, particularly in response to salient reward or reward-related (CS) stimuli. However, as mentioned in the text, it is possible that depolarization block might occur owing to increased AMPAR function. Depolarization block would lead to a temporary reduction of dopamine release. Adapted from Jones S and Bonci A (2005) Synaptic plasticity and drug addiction. *Current Opinion in Pharmacology* 5: 20–25.

administered noncontingent cocaine via a yoked design in which cocaine exposure is similar to that received by the animals self-administering cocaine. In this study, longer-lasting plasticity (3–4 weeks) at AMPARs was specifically observed during abstinence from cocaine, but not during food self-administration. Similarly, the excitatory synaptic inputs to the VTA are strengthened as well, but this only suggests that the learned association of the operant task with cocaine caused a long-lasting synaptic enhancement in the glutamatergic projections onto VTA dopamine (DA) neurons.

Plasticity at excitatory synapses has been demonstrated in brain regions receiving dopaminergic innervation, including prefrontal cortex, dorsal striatum, nucleus accumbens, hippocampus, and amygdala. The expression of LTP in these brain regions could contribute to acquisition of instrumental and Pavlovian learning under physiological conditions as well as being a consequence of chronic exposure to addictive drugs. Thus, it will be important to identify specific synaptic inputs that undergo plasticity under different learning conditions, since inputs to the VTA come from a variety of brain regions activated during reward-related learning. In addition, it will be interesting to examine the time course of onset and duration of synaptic plasticity in different parts of the reward circuit to ascertain the temporal role – and hence relative contribution – of each region.

In addition to evidence that LTP and LTD are expressed in the VTA, there is strong behavioral evidence that plasticity at glutamatergic synapses in the VTA is a possible key player in sustaining substance abuse. First, VTA dopaminergic cells are responsive to rewards and to cues learned by classical association to be predictors of reward. Second, place preference induced by cocaine shows dependence on VTA glutamate receptors. Third, inactivation of the VTA impairs cocaine-seeking behavior maintained by a drug-associated stimulus. Fourth, VTA glutamate receptor inhibition blocks stress and cue-induced relapse to cocaine seeking. Finally, repeated or even

a single administration of addictive drug (including cocaine, amphetamine, morphine, and nicotine) directly into the VTA initiates sensitization to systemic drug challenge, and NMDAR antagonists, delivered directly to the VTA, block the development of sensitization. Taken together, these findings strongly suggest that glutamatergic synapses in VTA dopaminergic cells constitute the sites of some of the earliest neural adaptations in response to addictive drugs. In summary, synaptic plasticity occurs at glutamatergic synapses in the VTA in response to several addictive drugs. Thus, potentiation of glutamatergic synapses might trigger persistent changes in dopaminergic cell signaling, thereby contributing to the persistence of pathological addictive behaviors.

Drug-Induced Synaptic Plasticity and VTA Dopaminergic Cell Activity

Thus far, we have discussed the evidence that drug-induced synaptic plasticity occurs in VTA dopaminergic cells, and that glutamate receptor activation onto VTA neurons is required for learning mechanisms in reward and addiction. However, a key question remains to be answered: What are the consequences of glutamatergic synaptic plasticity in the VTA for dopaminergic cell activity and dopamine release? VTA dopaminergic cells exhibit two main patterns of firing activity *in vivo*, single-spike firing and burst firing. Studies in rodents and primates have indicated that burst firing of VTA dopaminergic cells signals the occurrence of salient reward-associated stimuli. When spikes are clustered into bursts, the increase in extracellular dopamine in projection areas would likely produce a significantly larger amount of DA release than that observed for regularly spaced spike trains at the same mean spike frequency, as a consequence of nonlinear summation. Further, burst firing might promote co-release of a variety of peptides, as well as favor early gene expression. These results suggest that NMDAR activation might play two crucial roles in fine-tuning dopaminergic cell activity. First, NMDARs would promote induction of LTP of AMPAR responses. Second, NMDARs increase the efficiency of dopamine release by promoting burst firing of dopaminergic cells.

In vivo studies have shown that repeated cocaine or amphetamine administration causes an increase of VTA dopaminergic cell firing in response to AMPA exposure *in vivo*. Thus, it is plausible to hypothesize that LTP of AMPAR responses after a single or repeated drug exposure produces an increase in the incidence of burst firing, thus enhancing dopamine release. However, it is possible that excessive activation of AMPAR-mediated current might favor the occurrence of a depolarization block, thereby causing a reduction of action potential firing in VTA dopaminergic cells, and hence a reduction of dopamine release. It is very plausible that both phenomena might occur, ultimately causing an increase of the signal-to-noise ratio, where periods of depolarization blocks would be followed and/or preceded by periods of burst activity. Therefore, following drug exposure, plasticity at VTA excitatory synapses could be instrumental in determining a dramatic change of the firing pattern of VTA neurons, as reward-related stimuli are received and processed by VTA neurons. Such changes in the patterns of dopaminergic cell activity might represent an important mechanism by which dopamine signaling modulates the limbic circuit and sustains the formation of pathological drug-dependent learning that ultimately leads to substance abuse.

In conclusion, forms of synaptic plasticity sensitive to addictive drugs have been identified in the VTA. Addictive drugs administered either passively or voluntarily *in vivo* cause a rapid and persistent potentiation of AMPAR-mediated synaptic responses. Changes in AMPAR subunit expression are also reported to occur, but how such changes relate to the functional plasticity and behavioral effects of drugs is still a matter of debate. Ultimately, the questions are whether, and how, synaptic plasticity underlies specific forms of learning and long-lasting neural adaptations associated with addiction. These questions are fundamental in studies of the neurobiology of behavior and have yet to be answered.

Addiction and Ion Channels

Phasic alterations in action potential firing are a predominant mechanism through which neurons transmit information and drive behavioral expression. Firing changes are observed in several mesolimbic areas in association with many aspects of drug-related responding, including for operant responding and for stimuli that predict drug reward. Thus, any functional change that alters the firing properties of a neuron, either through glutamate receptor activation or ion channels, could significantly influence how the particular neuron and brain region contribute to behavioral expression.

Here we focus on investigations of ion channel modulation of acute drug effects, self-administration or reinstatement, and of modulation of ion channel function by chronic drug exposure. In this way, we hope to validate the potential *in vivo* role of a particular ion channel, and also identify consistencies or differences across different drugs of abuse. The contributions of intracellular signaling pathways to addiction have been addressed in recent reviews, and are not addressed here. In addition, acute effects of

drugs of abuse on ion channels through activation of G-protein-coupled receptors are extensive and varied, and are generally not addressed here.

We should note that there is much we do not yet know about regulation of addictive behaviors by particular ion channels; there are few channel types for which a broad range of drug-related behaviors has been investigated or for which the underlying brain regions have been identified. Furthermore, ion channel modulators could alter drug-related behavior by acting in brain regions different from those where a drug has its acute effects, thus altering the activity of the brain circuit underlying the behavior. This is a particularly important caveat when using transgenic animals or systemic channel modulators. Nonetheless, a channel modulator may actually be of great benefit for the treatment of addiction, even if therapeutic investigations outpace our understanding of its role at the cellular and regional level.

Calcium Channel Modulation of Addictive Behaviors

Perhaps the most widely studied example of ion channel regulation of an addictive behavior is the contribution of L-type voltage-gated calcium channels (LVGCCs), which can regulate firing, plasticity, and activation of intracellular regulatory molecules, to morphine tolerance and withdrawal (Figure 2). Concurrent administration of LVGCC antagonists with morphine significantly reduces the development of both tolerance and behavioral withdrawal symptoms, and also prevents the development of cellular changes thought to contribute to the expression of withdrawal (including upregulation of LVGCCs and reduced basal dopamine levels in many brain areas). Increased LVGCC levels after chronic morphine likely contribute causally to withdrawal, since acute inhibition of LVGCCs only, at the time of withdrawal, also significantly reduces signs of withdrawal.

A pattern of increased LVGCC density after chronic drug exposure, and reduction of the development of tolerance and withdrawal after treatment with LVGCC antagonists, are observed after chronic exposure to other drugs of abuse, including ethanol and nicotine. Chronic ethanol upregulates LVGCC levels, and acute LVGCC antagonists decrease ethanol withdrawal symptoms and the depression in DA levels associated with ethanol withdrawal. Reduced dopamine levels likely contribute to ethanol seeking, since self-administration of ethanol during withdrawal continues until nucleus accumbens (NAcb) dopamine levels are normalized. Thus, LVGCC dysregulation could represent a common mechanism underlying early withdrawal from abused drugs. However, not all drugs produce similar effects on LVGCCs; repeated amphetamine exposure upregulates VTA LVGCCs for at least 14 days.

The exact mechanism through which chronic exposure to LVGCC antagonists reduces withdrawal symptoms is uncertain. Opiate receptors can exert their cellular effects through calcium channels (e.g., for analgesia), and inhibition of LVGCCs might prevent these acute actions of opiate receptors and the development of compensatory neuroadaptations. LVGCC upregulation might also represent a more general homeostatic response to opiate-induced decreases in excitability – for example, through activation of G-protein-coupled inwardly rectifying potassium channels (GIRKs). Also, since low dopamine levels are associated with morphine withdrawal, it has been proposed that increased LVGCC levels reduce DA release in ventral tegmental area target regions, leading to some withdrawal symptoms; however, LVGCCs within the VTA enhance firing of DA neurons, so the cellular mechanism that could underlie this effect remains unclear.

In addition to regulation of withdrawal, LVGCC antagonists reduce the reinforcing effects of a number

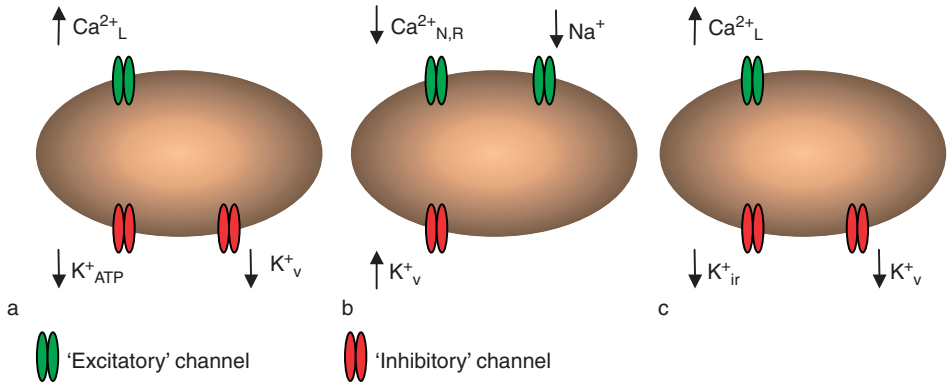

Figure 2 A schematic demonstrating examples of ion channels reported to be altered after (a) chronic morphine administration (brain region not specified) or (b, c) chronic cocaine administration, with differential changes in the nucleus accumbens (b) and prefrontal cortex (c), indicating potassium channels and N-, R-, and L-type voltage-gated Ca^{2+} channels.

of abused drugs. LVGCC antagonists reduce self-administration of cocaine, ethanol, and opiates, and, in one study, LVGCC antagonists did not alter sucrose self-administration except at high doses. LVGCC antagonists also reduce conditioned place preference (CPP) for several drugs, as well as nicotine reinstatement and amphetamine sensitization. Intra-NAcb infusion of LVGCC antagonists reduces the development of behavioral sensitization to cocaine, although this enhances cocaine CPP. We should note that not all studies agree, and this could reflect the differential effectiveness of the various LVGCC antagonists examined or the differential action on particular VGCC subtypes.

Considerable human preclinical evidence suggests that LVGCC antagonists may ameliorate withdrawal, tolerance, and craving in human addicts. LVGCC antagonists decrease adverse symptoms during early withdrawal from ethanol, opiates, and amphetamine, although effects on cocaine maybe more moderate. LVGCC antagonists decrease the euphoric effect of opiates, while ethanol intoxication is not altered in humans, although LVGCC antagonists enhance the sedative and ataxic effects in rodents. LVGCC antagonists enhance the analgesic effect of opiates by reducing the development of tolerance, in agreement with studies showing decreased opiate tolerance in rodent models. Taken together, studies from humans and rodents generally concur regarding the antiaddictive properties of LVGCC antagonists. However, a study in monkeys found that LVGCC blockers did not alter cocaine self-administration or cocaine modulation of food self-administration. Thus, preliminary investigations have suggested, but have not fully established, the clinical utility of LVGCC antagonists in the treatment of some aspects of addiction.

In general, the cellular mechanisms and brain regions where LVGCCs modulate withdrawal or reinforcement are unknown. Calcium influx could contribute to memory formation, since LVGCCs are critical for some types of synaptic plasticity. In addition, LVGCCs enhance firing – in particular, burst firing – of a number of brain regions, including the nucleus accumbens and VTA. LVGCC can also enhance DA release from terminals. However, a recent study suggests that LVGCC antagonists can directly inhibit nicotinic acetylcholine receptors at more moderate doses, underscoring the need for caution when working to develop human therapies based on rodent models. In addition, some results regarding LVGCC and the cellular mechanisms seem paradoxical, since LVGCC antagonists can prevent memory impairments, in apparent opposition to any LVGCC contribution to memory formation through enhanced plasticity, firing, or gene expression. Thus, it will ultimately be critical to identify the particular brain regions where LVGCC antagonists modulate different addiction- and memory-related behaviors. Nonetheless, even if LVGCC antagonists are found to inhibit addictive responses through nonspecific or peripheral effects, this may not diminish their clinical utility for treatment of human addiction and alcoholism.

Modulation of Drug Reinforcement by Other Ion Channels

In addition to LVGCCs, a number of other channels have been investigated in relation to the reinforcing effects of drugs (see Table 1). Some studies also have observed changes in firing in which the underlying channel changes are unknown. Here we address G-protein-coupled inwardly rectifying potassium channels in detail, since GIRK agonists may represent a potential therapy for addiction.

Many receptors inhibit neuronal firing through activation of GIRKs, and GIRK2 knockout studies also support a role for GIRKs in some acute effects of ethanol. GIRK2 and GIRK3 knockout mice also exhibit reduced cocaine self-administration, perhaps suggesting enhanced cocaine reinforcement due to enhanced neuronal excitability. However, simultaneous knockout

Table 1 Examples of channels influencing drug-related behavior

Channel	*Behavior*
ATP-activated potassium channels	Forebrain levels decreased by chronic morphine; systemic agonist reduces morphine withdrawal symptoms
Kv1.1 A-type potassium channel	Kv1.1 antisense (intracerebroventricular administration) reduces amphetamine but not cocaine or basal locomotion, and also food intake
N-Type voltage-gated calcium channels	Transgenic knockout reduces ethanol self-administration and place preference as well as hypnotic effects
R-Type voltage-gated calcium channels	Reduces cocaine-induced locomotion without affecting cocaine sensitization or conditioned place preference
BK (large-conductance, calcium-dependent potassium channel)	BK overexpression increases ethanol tolerance in *Drosophila*, while ethanol tolerance in the hypothalamic–neurohypophysial system is associated with decreased BK activity

of GIRK2 and GIRK3 surprisingly increases cocaine self-administration, serving as an example of the caution required when addressing the cellular mechanisms underlying drug reinforcement after brain-wide changes in ion channel function.

Although removal of GIRK channels may prevent certain drug actions, GIRK activators, such as baclofen, which activates GIRK through $GABA_B$ receptors, may represent a powerful therapeutic intervention to counteract craving and drug seeking. In animals, baclofen prevents the expression of sensitization or reinstatement for several abused drugs, including morphine and psychostimulants. In preclinical human studies, baclofen reduced withdrawal symptoms and craving in cocaine, ethanol, and opiate addicts. Baclofen may reduce drug seeking by decreasing neuronal excitability in multiple brain regions, perhaps reducing neuronal hyperexcitability during withdrawal, and also decreasing cue-related firing and drug seeking.

Interesting as well is the increasing number of studies using DNA microarray or proteomic techniques to uncover proteins associated with differential preference for abused drugs, including ethanol. Some of these studies compare different ethanol-preferring and nonpreferring rodent strains, while others examine the consequence of ethanol exposure on gene expression. Of course, many different genes and gene products, including some channels, show differential expression and could contribute to ethanol preference.

Changes in Ion Channel Function after Chronic Drug Exposure

There is considerable interest in understanding the persisting changes in cellular function that may underlie craving and relapse, even after months or years of abstinence, following chronic self-administration of an abused drug. Changes in ion channel regulation of firing could dramatically alter sensitivity to relapse-inducing stimuli. However, with the exception of studies of psychostimulant alteration of ion channel function (Figure 2), relatively few studies have examined whether chronic drug exposure persistently alters ion channel function.

Repeated cocaine administration followed by 3 days of withdrawal reduces NAcb excitability due to increased potassium channel activity, decreased R- and N-type VGCC activity, and decreased sodium channel activity. Repeated cocaine administration has opposite effects in frontal cortical neurons, with enhanced excitability due to increased activity of LVGCCs and decreased activity of IRK and voltage-gated potassium channels. No changes are observed in the motor cortex, suggesting specific neuroadaptations in limbic areas. In contrast to cocaine, repeated amphetamine administration reduced sodium channel activity and excitability of subicular and frontal cortical neurons. Interestingly, some channel changes may be secondary to alterations in tonic activity of intracellular signaling molecules. For example, in the NAcb, increased tonic protein phosphatase activity and increased tonic protein kinase A (PKA) activity likely underlie the reduced calcium and sodium channel function, respectively. Also, importantly, enhanced excitability of frontal cortical neurons persists for 2–3 weeks after withdrawal from cocaine, suggesting that this activity may be long-lasting and thus may contribute to the expression of behavioral sensitization.

Chronic morphine also alters several ion channels other than LVGCCs, with different effects, depending on the brain region. The striatum shows reduced potassium channel (Kv1 and Kv2) mRNA, likely enhancing firing, while P/Q- and L-type VGCCs are reduced in trigeminal neurons but not in the locus coeruleus. Chronic morphine can also alter opiate channel coupling, with a change at the level of the receptor instead of the channel. Finally, after 10 months of ethanol exposure and 2–4 weeks of withdrawal, the only change in intrinsic excitability in hippocampal neurons was a decrease in the afterhyperpolarization, suggesting reduced K^+ channel function.

Conclusions and Future Directions

Although there are interesting and tantalizing clues pointing toward the effect of chronic drug exposure on channel function, many questions remain. Few studies have examined channel function beyond the period of early withdrawal, while changes persisting after weeks or months of abstinence may be critical in long-term susceptibility to craving and relapse. It is also clear that self-administration and passive administration can differentially alter gene expression and protein function, perhaps due to learned cues that come to be associated with the act of voluntary seeking or self-administration of drugs. Thus it is also critical to understand whether any changes in ion channel function are specific to active self-administration. A recent study has identified a change in glutamatergic function in animals that self-administered cocaine, but not in those that received cocaine passively, which perhaps represents a cellular correlate of learned associations, possibly between the drug experience and drug-associated cues, which can drive cocaine-seeking behavior.

Identifying the most critical cellular changes that may underlie the persisting neuroadaptations that drive relapse of drug seeking will likely require a number of different important techniques. Clearly, biochemical methods (such as Western blots and local knockdown or overexpression of particular signaling

molecules) provide tremendous power in identifying protein products altered by chronic drug exposure and prolonged abstinence, and which contribute to susceptibility to relapse. Similarly, large-scale screening for changes in protein levels (using proteomics) or mRNA levels (using DNA microarrays) hold great promise for indicating potential changes in channel function. In addition, we believe that brain slice experiments examining the electrophysiological consequences of chronic drug exposure and prolonged abstinence are critical, since there are many mechanisms that can alter receptor or channel function without changing total protein. Such experiments examining electrophysiological properties in brain slices from adult animals are laborious (due to the amount of time needed for training the animals) and technically challenging. However, the information gained is potentially of great value to the understanding of the mechanisms underlying drug self-administration and relapse.

See also: Addiction: Neurobiological Mechanism; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Drug Addiction: Behavioral Neurophysiology; Drug Addiction: Neuroimaging; Drugs Addiction: Actions; Glutamate Receptor Organization: Ultrastructural Insights; Ion Channel Localization in Axons; Ion Channel Localization in Cell Bodies and Dendrites.

Further Reading

Cardinal RN and Everitt BJ (2004) Neural and psychological mechanisms underlying appetitive learning: Links to drug addiction. *Current Opinion in Neurobiology* 14: 156–162.

Cooper DC (2002) The significance of action potential bursting in the brain reward circuit. *Neurochemistry International* 41: 333–340.

Dong Y, Nasif FJ, Tsui JJ, et al. (2005) Cocaine-induced plasticity of intrinsic membrane properties in prefrontal cortex pyramidal neurons: Adaptations in potassium currents. *Journal of Neuroscience* 25: 936–940.

Gerdemen GL, Partridge JG, Lupica CR, et al. (2003) It could be habit forming: Drugs of abuse and striatal synaptic plasticity. *Trends in Neurosciences* 26: 184–192.

Harris GC and Aston-Jones G (2003) Critical role for ventral tegmental glutamate in preference for a cocaine-conditioned environment. *Neuropsychopharmacology* 28: 73–76.

Kauer JA (2004) Learning mechanisms in addiction: Synaptic plasticity in the ventral tegmental area as a result of exposure to drugs of abuse. *Annual Review of Physiology* 66: 447–475.

Kobayashi T, Washiyama K, and Ikeda K (2004) Modulators of G protein-activated inwardly rectifying K^+ channels: Potentially therapeutic agents for addictive drug users. *Annals of the New York Academy of Sciences* 1025: 590–594.

Martin M, Chen BT, Hopf FW, et al. (2006) Cocaine self-administration selectively abolishes LTD in the core of the nucleus accumbens. *Nature Neuroscience* 9: 868–869.

Rajadhyaksha A, Husson I, Satpute SS, et al. (2004) L-Type Ca^{2+} channels mediate adaptation of extracellular signal-regulated kinase 1/2 phosphorylation in the ventral tegmental area after chronic amphetamine treatment. *Journal of Neuroscience* 24: 7464–7476.

Robbins TW and Everitt BJ (2002) Limbic-striatal memory systems and drug addiction. *Neurobiology of Learning and Memory* 78: 625–636.

Rodd ZA, Bertsch BA, Strother WN, et al. (2006) Candidate genes, pathways and mechanisms for alcoholism: An expanded convergent functional genomics approach. *The Pharmacogenomics Journal* (published online 10 October 2006).

Shulman A, Jagoda J, Laycock G, et al. (1998) Calcium channel blocking drugs in the management of drug dependence, withdrawal and craving. A clinical pilot study with nifedipine and verapamil. *Australian Family Physician* 27(supplement 1): S19–S24.

Schultz W (2004) Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Current Opinion in Neurobiology* 14: 139–147.

Ungless MA, Whistler JL, Malenka RC, et al. (2001) Single cocaine exposure *in vivo* induces long-term potentiation in dopamine neurons. *Nature* 411: 583–587.

Wise RA (2004) Dopamine, learning and motivation. *Nature Reviews Neuroscience* 5: 483–494.

Drug Addiction: Neuroimaging

R Z Goldstein and N Alia-Klein, Brookhaven National Laboratory, Upton, NY, USA
N D Volkow, National Institute on Alcohol Abuse and Alcoholism, Rockville, MD, USA; and National Institute on Drug Abuse, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

Many substances of abuse exist; they are classified according to their psychoactive effects on the individual's behavioral, psychological, and neurobiological functioning. For example, central nervous system depressants include ethanol, opiates, and sedative hypnotics. Stimulants include cocaine and methamphetamine. All substances of abuse are reinforcing. That is, they increase the probability of a subsequent response through increasing pleasure/reward or decreasing displeasure/pain processes. All drugs of abuse thus share a powerful psychoactive ('mind' altering) impact that enhances a learning process whereby the individual develops anticipatory responses to the drugs. These drug-induced effects are believed to contribute to the progression from drug abuse to drug addiction in susceptible individuals (see Table 1 for current definitions of drug abuse and addiction by the *Diagnostic and Statistical Manual for Mental Disorders*, fourth edition (*DSM-IV*)).

Addiction is a chronic disease characterized by repeated periods of drug craving, intoxication, bingeing, and withdrawal. This cycle culminates in the escalated preoccupation with the attainment and consumption of the substance. In particular, the compulsive consumption of the drug occurs at the expense of the attainment of other rewarding experiences and despite detrimental consequences to the individual's functioning (encompassing physical health and other personal, social, and occupational goals). This pattern continues despite attempts of the drug-addicted individual to stop or curtail drug use and even when the rewarding experiences from the drug are markedly reduced.

Note that not all individuals who consume or abuse psychoactive substances become addicted to them; it is estimated that approximately 10% of those who use drugs develop drug addiction/dependence. The precise statistic is uncertain; nevertheless, it depends on the drug of abuse, ranging from approximately 4% of marijuana users to approximately 32% of those who ever try tobacco (heroin, 23%; cocaine, 16–17%; and alcohol, 15%). This actual risk for developing drug addiction is determined by genetic (it is estimated that 50% of the vulnerability is genetically determined), developmental (the onset of progression to addiction is more likely to occur during adolescence and early adulthood), other individual (e.g., psychiatric comorbidity) and environmental factors (e.g., exposure to stress and availability of drugs), and their interactions. In this article, we concentrate on mapping drug addiction and the unique symptomatology of this chronic disease process to its underlying neurobiological mechanisms as revealed by neuroimaging studies.

Major Brain Circuits and Neurotransmitters Studied in Drug Addiction

Animal and human studies show that drugs of abuse exert their reinforcing and addictive effects by activating and ultimately co-opting the brain's reward system, a phylogenetically ancient circuitry normally involved in the reward sensations that are essential to learning and survival. This system was first discovered in the 1950s when it was observed that animals would work for electrical stimulation to specific brain regions, subsequently found to include myelinated fiber tracts in the medial forebrain bundle, which descend from the cortex and terminate at or near the cells bodies of dopamine (DA) neurons in the ventral tegmental area (VTA). The hub of the system consists of the mesolimbic and mesocortical DA fibers, which originate in the VTA and terminate in the nucleus accumbens (NAc), amygdala, hippocampus, and prefrontal cortex (PFC). This circuit's drug-induced stimulation occurs directly by triggering DA action or indirectly by modulating other neurotransmitters. The latter encompass, among others, glutamate (the main excitatory neurotransmitter in the brain, involved in the molecular changes that encompass long-term potentiation associated with learning), gamma aminobutyric acid (GABA; the main inhibitory neurotransmitter in the brain), endogenous opioids (involved in hedonic responses and pain regulation), acetylcholine (predominantly via its effects in nicotine receptors which reside in DA neurons in the VTA), cannabinoids (the cannabinoid CB1 receptor modulates DA cell firing), and serotonin (involved in impulsivity and mood).

The response of this DA mesolimbic circuit to drugs of abuse occurs in all organisms (it is not specific to addiction) and therefore cannot solely explain the progression to the complex phenomenology of drug addiction that occurs in susceptible individuals. Indeed, advances in our understanding of the neurobiology of drug addiction have expanded our current

Table 1 *DSM-IV* criteria for substance abuse and dependence

A maladaptive pattern of substance use leading to clinically significant impairment or distress, as manifested by one or more (abuse) or at least three (dependence) of the following, occurring within a 12 month period:

Abuse

1. Recurrent substance use resulting in a failure to fulfill major role obligations at work, school, or home (e.g., repeated absences or poor work performance related to substance use; substance-related absences, suspensions, or expulsions from school; neglect of children or household)
2. Recurrent substance use in situations in which it is physically hazardous (e.g., driving an automobile or operating a machine when impaired by substance use)
3. Recurrent substance-related legal problems (e.g., arrests for substance-related disorderly conduct)
4. Continued substance use despite having persistent or recurrent social or interpersonal problems caused or exacerbated by the effects of the substance (e.g., arguments with spouse about consequences of intoxication, physical fights)

These symptoms have never met the criteria for substance dependence for this class of substance.

Dependence

1. Tolerance, as defined by either of the following:
   (a) A need for markedly increased amounts of the substance to achieve intoxication or desired effect
   (b) Markedly diminished effect with continued use of the same amount of the substance
2. Withdrawal, as manifested by either of the following:
   (a) The characteristic withdrawal syndrome for the substance (refer to criteria A and B of the criteria sets for withdrawal from the specific substances)
   (b) The same (or a closely related) substance is taken to relieve or avoid withdrawal symptoms
3. The substance is often taken in larger amounts or over a longer period than was intended
4. There is a persistent desire or unsuccessful efforts to cut down or control substance use
5. A great deal of time is spent in activities necessary to obtain the substance (e.g., visiting multiple doctors or driving long distances), use the substance (e.g., chain smoking), or recover from its effects
6. Important social, occupational, or recreational activities are given up or reduced because of substance use
7. The substance use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by the substance (e.g., current cocaine use despite recognition of cocaine-induced depression, or continued drinking despite recognition that an ulcer was made worse by alcohol consumption)

view beyond the reward process and its underlying mesolimbic brain circuitry to include the cortical brain regions that are associated with higher order cognitive–behavioral functions such as attribution of emotional salience to stimuli in the environment and the adaptive control of behavior – two interrelated and core neuropsychological dimensions impaired in addiction (Figure 1). The first impairment refers to reward processing/reinforcement associative learning, whereby as addiction progresses the reinforcement value of drugs of abuse becomes more potent than most other non-drug-related/common stimuli in the environment (e.g., sex and food). The disruption of the following interconnected brain circuits is involved: (1) reward, mapped in the NAc and the ventral pallidum, involved in the processing of and sensitivity to drug and non-drug-related reward and possibly also negative reinforcers; (2) motivation/drive, represented in the orbitofrontal cortex (OFC) and the subcallosal cortex, involved in the behavioral drive to seek a reinforcer and in salience attribution; and (3) memory and learning, modulated by the hippocampus, amygdala, and dorsal striatum, involved in the formation and storage of drug-related memories, conditioned responses, and habits, respectively. Note the interaction between these circuits: attribution of salience to a given stimulus, which is a function of the OFC, depends on the relative value of a reinforcer compared to simultaneously available reinforcers, which requires knowledge of the strength of the stimulus as a reinforcer – a function of the hippocampus and amygdala. The second impairment refers to inhibitory control (e.g., impulsivity), which is suggested to be compromised, especially in a drug-related context, and is executed in the PFC, the anterior cingulate cortex (ACC), and the lateral OFC. Thus, as described later, these 'impairments in response inhibition and salience attribution' (I-RISA) and their neurobiological underpinnings are thought to provide a meaningful cutting-edge exploration of drug addiction.

Major Neuroimaging Findings in Human Drug Addiction

The imaging of DA *in vivo* is possible through the use of positron emission tomography (PET) and single photon emission computed tomography (SPECT). Specifically, through the use of radioactively labeled radiotracers such as C^{11}raclopride (for imaging DA D2/D3 receptors) and C^{11}cocaine, C^{11}altropane, or I^{123}B CIT (for imaging the DA transporters (DATs)), PET can be used to image DA function of the brain's subcortical basal ganglia (including the caudate and putamen). PET has also been used to evaluate the

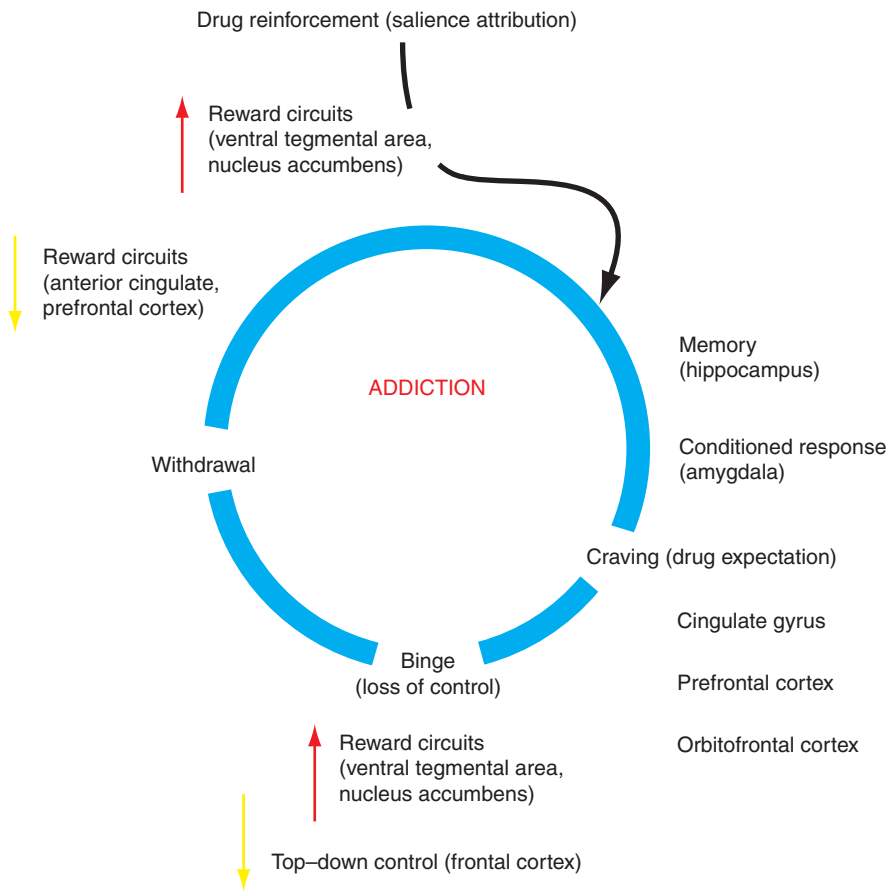

Figure 1 Impaired response inhibition and salience attribution (I-RISA) model of drug addiction. Here, core clinical characteristics of the drug addiction cycle (encompassing intoxication, craving, bingeing, and withdrawal) are mapped vis-à-vis the functioning of the subcortical and cortical brain regions forming the mesolimbic and mesocortical dopamine circuit that is implicated in drug addiction and its underlying neuropsychological mechanisms. Reprinted from Goldstein RZ and Volkow ND (2002) Drug addiction and its underlying neurobiological basis: Neuroimaging evidence for the involvement of the frontal cortex. *American Journal of Psychiatry* 159: 1642–1652, with permission from the American Journal of Psychiatry, Copyright 2002. American Psychiatric Association.

pharmacokinetics of drugs at their site of action. For example, using C^{11}cocaine, studies showed a close correspondence between the fast uptake and clearance of cocaine in the human brain with the fast and short-lasting but intense 'high.' A parallel comparison with the pharmacokinetics of methylphenidate (a stimulant drug that, similar to cocaine, blocks DAT), measured with C^{11}methylphenidate, revealed that it was the fast uptake but not the presence of the drug in the brain that was linked to its rewarding effects as assessed by self-reports of 'high.' This study highlighted the relevance of the rate of drug delivery into the brain as one of the variables affecting the rewarding properties of drugs of abuse. This study also identified the importance of the rate of clearance of a drug from the brain in its reinforcing effects since the slower clearance of methylphenidate than that of cocaine would interfere with its frequent and fast administration, providing an explanation for the differential abuse rates of both drugs (low for methylphenidate and high for cocaine).

In addition to measures of neurochemistry, PET can also quantify regional brain function such as glucose metabolism (with ^{18}F-fluorodeoxyglucose (FDG)) and cerebral blood flow (with ^{15}oxygen). These measures can be assessed at a resting baseline and during a behavioral or pharmacological challenge. Functional magnetic resonance imaging (fMRI) provides temporally and spatially enhanced images of brain function. The fMRI signal is based on differences in the magnetic properties of oxygenated versus deoxygenated hemoglobin, producing a blood oxygen level-dependent (BOLD) contrast when brain tissue is stimulated. Because the neural and behavioral effects of acute drug administration tend to be of short duration, fMRI is especially well suited to identify the brain circuits activated during drug intoxication and their association with the drug's behavioral effects.

Both PET and fMRI have facilitated the study of the brain circuits underlying observable abnormalities in behavior, cognition, emotion, and personality in

psychopathology (including drug addiction). A particularly productive strategy has been the use of functional neuroimaging in conjunction with neuropsychological measures (including self-reports) during scanning and also outside of the restrictive imaging environment. The simultaneous recordings allow the mapping of the real-time drug effects (e.g., self-reported euphoria and craving) to their underlying neural basis; the assessment of neuropsychological functioning obtained outside of the imaging environment allows for a more in-depth analysis of trait measures, such as personality.

Association of Brain Imaging with the Core Clinical Drug Addiction Characteristics

Brain Imaging of Acute Pharmacological and Nonpharmacological Effects

Intoxication Intoxication occurs when a person consumes doses that are large enough to produce significant behavioral, physiological, or cognitive impairments. Early in the process of drug addiction, the initial drug-induced response is that of a sense of reward. Later, when drug addiction has developed, the impact of positive reinforcement decreases (e.g., feelings of 'high' or 'rush'), whereas that of negative reinforcement increases (e.g., alleviation of the stress of withdrawal). Preclinical (*in vivo* and *in vitro*) studies have consistently shown that the initial reinforcing drug effect is associated with increases in DA in the NAc. The precise mechanism of the drug-induced DA increases depends on the actual drug of abuse. Thus, cocaine increases DA by blocking the DAT, which is the main mechanism for recycling DA back into the nerve terminals. Indeed, PET studies have linked the level of DAT blockade by cocaine with the intensity of the self-reported 'high' (Figure 2). Depressant drugs such as benzodiazepines, barbiturates, and alcohol increase DA indirectly, in part via their effects on the GABA/benzodiazepine receptor complex. Opioid drugs such as heroin, oxycontin, and vicodin act by stimulating mu opiate receptors, some of which are located in DA cells and others in the GABA cells that regulate the DA cells and their terminals. Nicotine is believed to exert its reinforcing effects in part by activation of the $\alpha_4\beta_2$ nicotine receptors, which have also been identified in DA cells. Nicotine (similar to heroin and alcohol) also appears to release endogenous opioids, and this is also likely to contribute to its rewarding effects. Imaging studies have also documented that cigarette smoking results in inhibition of the monoamine oxidase (MAO) B and MAO A enzymes, which are involved in the metabolism of the catecholamines (including DA). This MAO inhibition is not due to nicotine but, rather, to some of the chemicals in the cigarette smoke; it is not limited to the brain but is observed throughout multiple organs in the body (Figure 3). Finally, marijuana exerts its

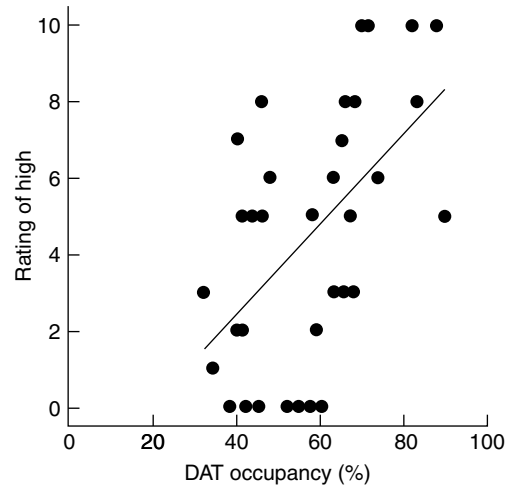

Figure 2 Correlation between dopamine transporter (DAT) occupancy by cocaine and subjective reports of 'high.' Ratings of 'high' significantly correlate with DAT occupancy in cocaine-addicted subjects: DAT occupancies >60% are consistently associated with significantly increased ratings of 'high' relative to baseline. Reproduced from Volkow ND, Wang GJ, Fischman MW, et al. (1997) Relationship between subjective effects of cocaine and dopamine transporter occupancy. *Nature* 386: 827–830, with permission.

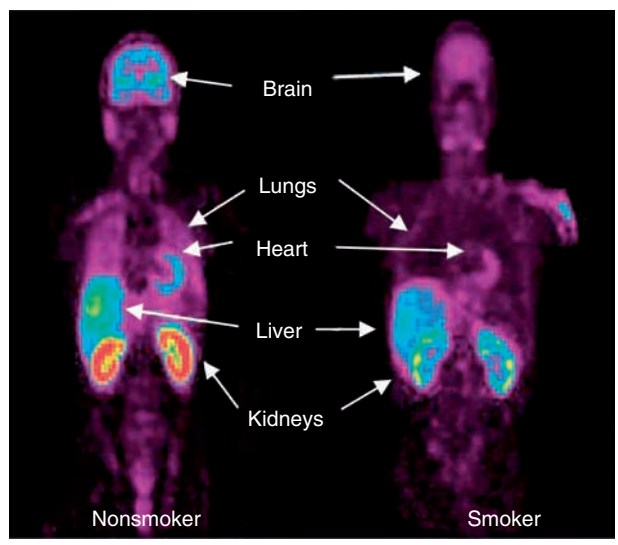

Figure 3 Monoamine oxidase (MAO) B activity in cigarette smokers (right) showing dramatic reductions in the brain, kidneys, heart, and lungs. These data indicate that the MAO-inhibiting chemicals in smoke end up not only in the lungs but also throughout the entire body, providing an explanation as to why a wide variety of cancers (bladder, pancreas, and kidney) are much more frequently present in cigarette smokers than in nonsmokers. From Fowler JS, Logan J, Wang GJ, et al. (2003) Low monoamine oxidase B in peripheral organs in smokers. *Proceedings of the National Academy of Sciences of the United States of America* 100: 11600–11605.

effect by activating cannabinoid 1 receptors, which modulate DA cells as well as postsynaptic DA signals. Moreover, there is increasing evidence for the involvement of cannabinoids in the reinforcing effects of other drugs of abuse, including alcohol, nicotine, cocaine, and opioids.

Activation of the PFC, including the ACC (Figure 4), and the mesolimbic DA subcortical brain areas has been reported during drug intoxication in drug-addicted subjects – an effect that is proportional to the intensity of the subjective perception of intoxication, the reinforcing effects of the drug, or enhanced mood. It is also intriguing that in the case of marijuana or methylphenidate, activation of the PFC regions is predominantly observed in users but not in non-using subjects and in subjects in whom the drug induces intense craving or enhances mood. This suggests that PFC regions (including the OFC and ACC) are involved in the intoxication process, and that their response to drugs is in part related to previous drug experiences. Therefore, the intact functioning of memory circuits is crucial since these circuits set the expectations and prepare the drug-addicted individual for the drug's acute effects (indeed, these memory circuits are also activated during drug intoxication). This preparation is essential for activating homeostatic neurobiological and other physiological (e.g., cardiovascular) responses to counteract the acute effects of the drugs (note that overdosing is frequently reported in situations in which the drug-addicted individual consumes drugs in unfamiliar environments or circumstances).

Salience attribution/reward processing. Neuroimaging findings have provided the first evidence that the brain's threshold to responding to abstract rewards is modified in drug-addicted individuals and that it is directly linked to changes in the responsiveness of the PFC. In a fMRI study of the I-RISA model (Figure 1), cocaine-addicted individuals and matched healthy control subjects were given monetary reward for performance on an attention task. Results indicated that the OFC responded to different monetary amounts in a graded fashion (the higher the potential reward, the greater the response) and the dorsolateral PFC and

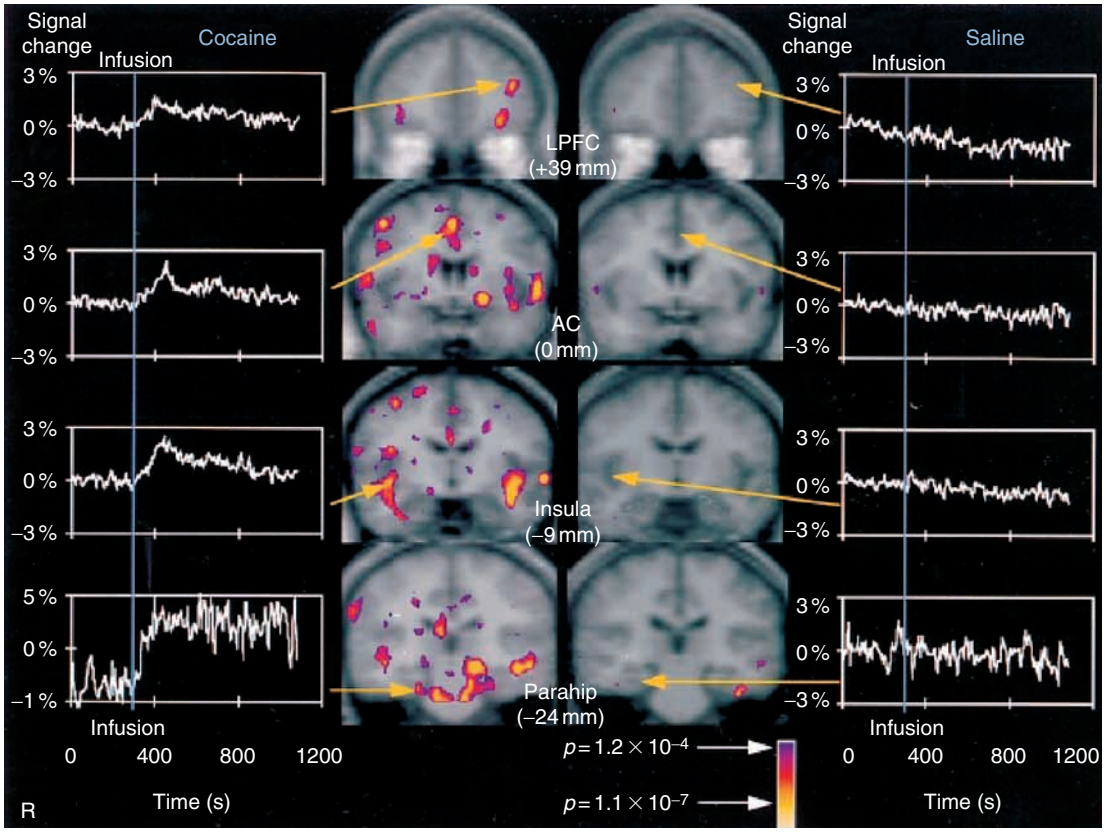

Figure 4 Functional magnetic resonance imaging and its application to measuring brain responses to intravenous cocaine (vs. saline) in addicted individuals. Intravenous administration of cocaine (but not saline) in ten cocaine-addicted individuals resulted in activations in a distributed brain network including the lateral prefrontal cortex and anterior cingulate cortex. Other activated regions were the insula, parahippocampal gyrus, nucleus accumbens, ventral tegmental area, and the amygdala. Reproduced from Breiter HC, Gollub RL, Weisskoff RM, et al. (1997) Acute effects of cocaine on human brain activity and emotion. *Neuron* 19: 591–611, with permission from Elsevier.

ACC responded to any money only in the healthy subjects but not in drug-addicted individuals. Instead, within the cocaine group, the higher the sensitivity to money in the dorsolateral PFC and ACC, the higher the self-reported motivation and ability to control behavior (Figure 5). Data further suggested that in the cocaine abusers, there was a 'disconnect' between subjective measures of motivation (how much they said they were engaged in the task) and the objective measures of motivation (how fast and accurately they performed on the task), implicating a disruption in the ability to perceive or translate into appropriate action inner motivational drives. Overall, this altered brain–behavior reward processing, which encompassed the role of the PFC in sensitivity to reward, awareness of inner motivation, and the ability to

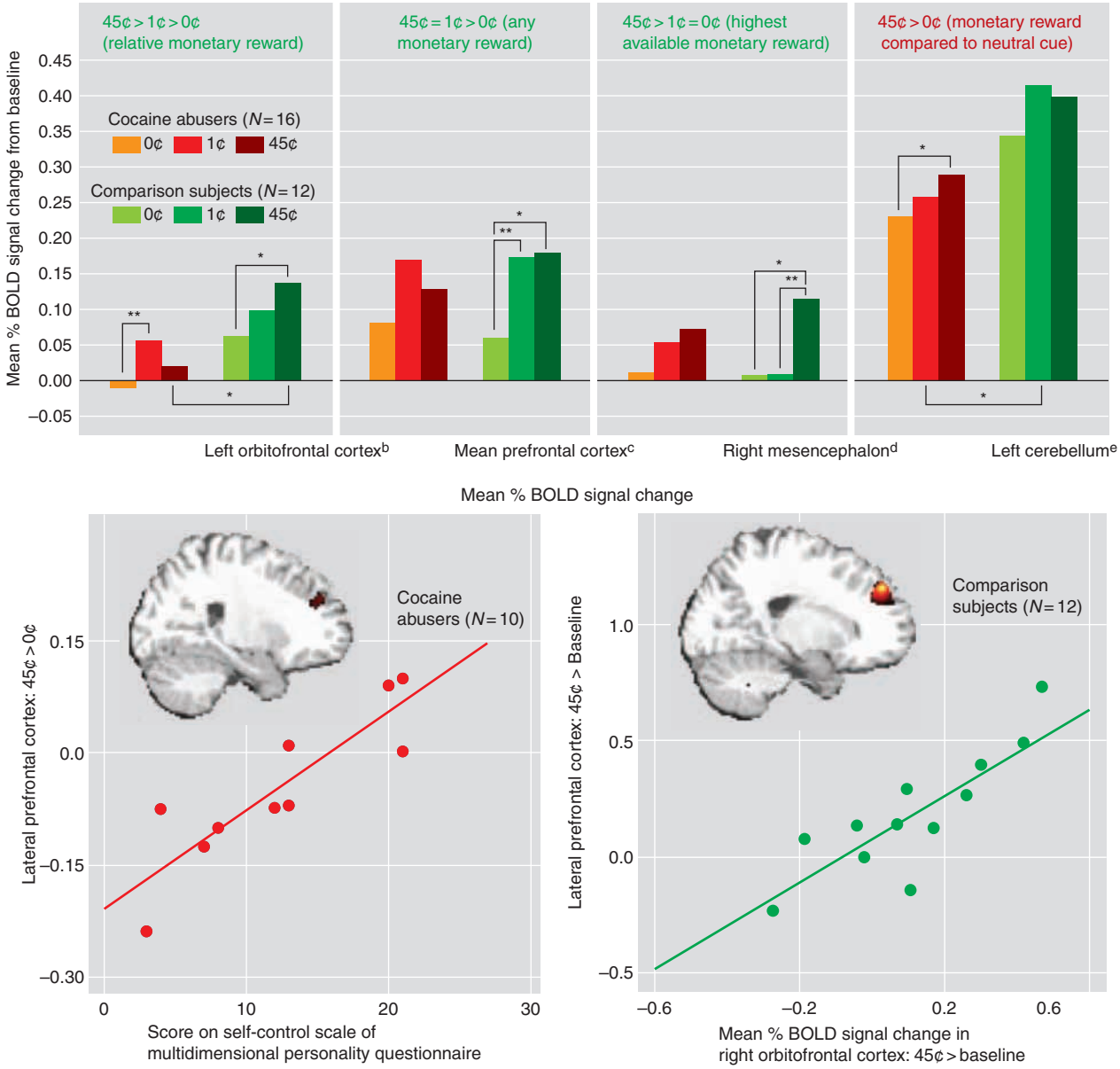

Figure 5 Functional magnetic resonance imaging and its application to measuring brain responses to a secondary reward (money) in addicted individuals versus healthy control subjects. Whereas 12 control subjects had a complex neural response to monetary reward (e.g., monotonically positive in the orbitofrontal cortex and response to any money in the dorsolateral prefrontal cortex and anterior cingulate cortex), this complexity was compromised in 16 addicted subjects. Furthermore, sensitivity to monetary reward in the dorsolateral prefrontal cortex in the addicted subjects was associated with better self-reported ability to control behavior at a trait level. Reprinted from Goldstein RZ, Alia-Klein N, Tomasi D, et al. (2007) Is decreased prefrontal cortical sensitivity to monetary reward associated with impaired motivation and self-control in cocaine addiction? *American Journal of Psychiatry* 164: 43–51, with permission from American Journal of Psychiatry, Copyright 2007. American Psychiatric Association.

control behavior, may contribute to the chronically repeating pattern of impulsiveness and lack of control over drug use in addiction, as further described next.

Inhibitory control. Inhibitory control is a neuropsychological construct that refers to the capacity to control the inhibition of harmful and/or inappropriate emotion, cognition, or behavior. Critically, the disruption of self-controlled behavior is likely to be exacerbated during drug use and intoxication (culminating in bingeing, as further described later) as modulated by a compromise in an essential function of the PFC: its inhibitory effect on subcortical limbic regions (e.g., the NAc, basal ganglia, and the amygdala). This impairment in top–down control (a core function of the PFC) would release behaviors that are normally kept under close monitoring, simulating stresslike reactions in which control is suspended/inhibited and stimulus-driven behavior is facilitated. This PFC compromise is consistent with the accounts of drug-addicted individuals who claim that once they start taking the drug, they cannot stop, even when the drug is no longer pleasurable and clearly exposing the individual to physical or psychological risk. The use of psychoactive substances impairs mechanisms of inhibitory control not just during acute drug intoxication; this effect spreads to all stages of the addiction process, possibly contributing (together with the changed valuation system as described previously) to the initiation or escalation of undesirable or disadvantageous decisions and behaviors (e.g., trading a car for a couple of hits). Indeed, one of the most consistent findings from fMRI and PET studies is that of abnormalities (e.g., hypoactivations) in the PFC, including the ACC, and increased impulsivity on neuropsychological measures of behavior (including compromised detection and regulation of error responses on inhibitory control tasks such as the go/no-go task) in drug-addicted individuals who are current but also past users.

Craving It is well recognized that drug effects are modulated by nonpharmacological variables, including context and conditioned responses. These variables encompass learned responses that link the drug and its environment to a pleasurable or an intensely overpowering experience. This learning shapes an individual's expectations of the effects of a drug, which in turn modify the neural and behavioral responses to that drug, as mentioned previously. Thus, when addicted individuals are exposed to drug cues such as places, people, or paraphernalia associated with drug intake, an intense and urgent desire (drug wanting or craving) to consume the drug frequently develops, as documented by self-report and by concomitant DA increases in the dorsal striatum (caudate and putamen) (Figure 6). This DA activation of the dorsal striatum could result in the readiness to engage in the behaviors necessary to procure the desired object. Thus, variations in DA D2 receptor levels may be one of the neurobiological mechanisms that underlie the effect of environmental cues on the propensity to abuse drugs or even relapse; this possibility is supported by findings showing that other environmental factors, such as stress and social hierarchy, can affect baseline levels of DA D2 receptors.

In addition, craving by itself is sufficient in activating the PFC regions that are activated during intoxication in individuals previously exposed to the drug of abuse. Thus, higher levels of brain activation (measured by cerebral blood flow, FDG, or BOLD) in PFC areas, primarily in the OFC and ACC, have been demonstrated in cocaine users exposed to videotapes and other cues depicting drug-related stimuli (Figure 7), and these neural changes correlate with self-reports of craving. Similar cue-reactivity effects have been observed in other types of drug addiction (e.g., opiate and nicotine). The pathways between the PFC (including OFC and ACC but also the dorsolateral PFC) and mesolimbic subcortical (e.g., amygdala) regions may underlie the phenomenology of craving that encompasses recall of emotionally laden previous experiences and anticipation of a future drug reward. Moreover, the involvement of these pathways in drug craving suggests the intriguing hypothesis that drug addiction may engage the same neurobiological processes that underlie motivated behaviors required for survival (e.g., eating).

Bingeing Experience with a reinforcing substance (especially psychostimulants but also alcohol and heroin) could lead to bingeing, in which, for a discrete period of time, there is an extreme loss of control during which the individual engages in the repeated consumption of the substance. These bingeing episodes – or 'runs' in which the drug(s) is consumed for longer periods or in greater amounts than intended – can override survival actions (e.g., feeding, sleeping, and physical safety). Termination of these extreme periods usually occurs upon exhaustion or external circumstances such as lack of drug availability.

Bingeing is associated with DA, serotonergic, and glutamatergic circuits and probably involves the activation of the thalamo-OFC circuit and the ACC. Experimentally studying the neurobiology of bingeing in humans is not feasible for understandable reasons, although elegant laboratory designs can circumvent these difficulties. For example, studies of non-treatment-seeking drug-addicted individuals evaluated when and how often drug administration occurred within a medically supervised time-limited

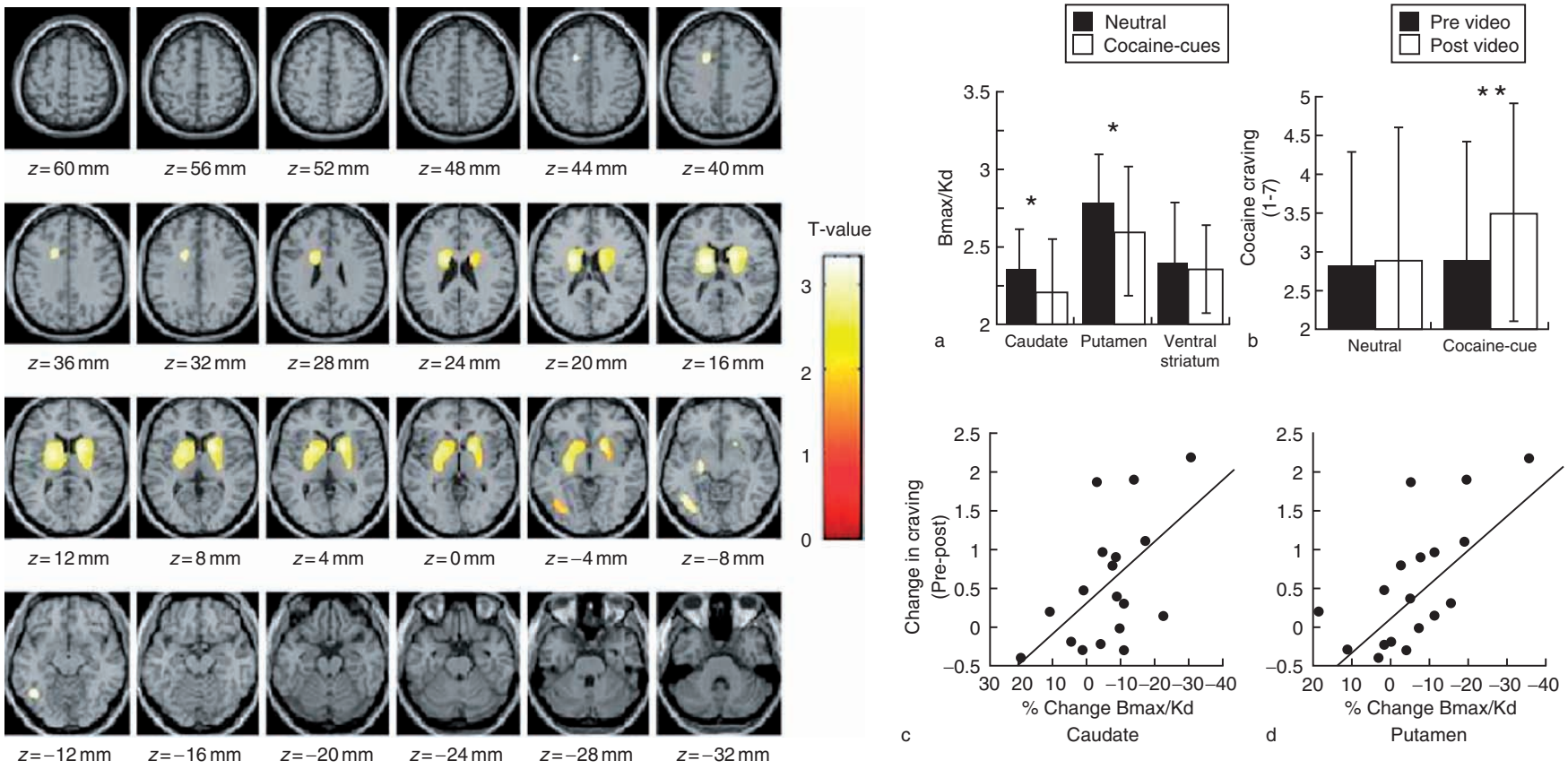

Figure 6 Positron emission tomography with C^{11}raclopride to measure brain responses to cocaine cues in cocaine-addicted individuals. Image on the left and top right shows increased dopamine receptor availability in the caudate and putamen (but not in the ventral striatum) while the 18 subjects were watching a video of addicted individuals procuring and using cocaine versus a video of neutral scenes. This difference was positively associated with the abusers' reported cravings after the cocaine-cue video compared to the neutral video as shown in the bottom right image. From Volkow ND, Wang GJ, Telang F, et al. (2006) Cocaine cues and dopamine in dorsal striatum: Mechanism of craving in cocaine addiction. *Journal of Neuroscience* 26: 6583–6588.

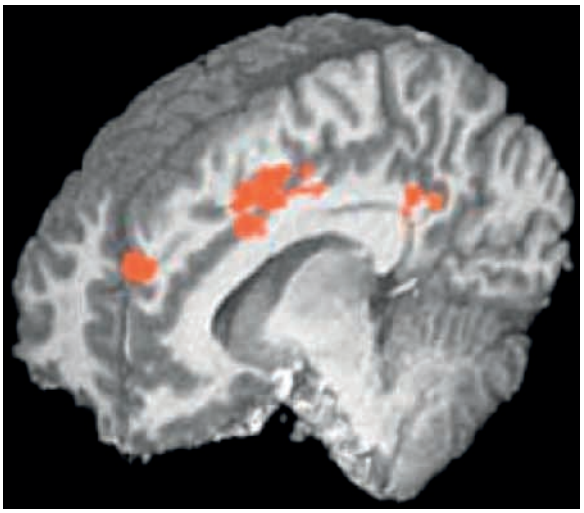

Figure 7 Functional magnetic resonance imaging and its application to measuring brain responses to cocaine cues in cocaine-addicted individuals. Presented are the functional brain activations in the left medial frontal gyrus and the anterior and posterior cingulate cortices that were associated with cue-induced (by video depicting individuals smoking crack-cocaine) cocaine craving in 17 experienced cocaine users. Reprinted from Garavan H, Pankiewicz J, Bloom A, et al. (2000) Cue-induced cocaine craving: Neuroanatomical specificity for drug users and drug stimuli. *American Journal of Psychiatry* 157: 1789–1798, with permission from the American Journal of Psychiatry, Copyright 2000. American Psychiatric Association.

self-administration procedure. Such fMRI and behavioral studies in cocaine-addicted individuals have shown that repeated drug self-administration-induced 'high' correlated negatively with activity in limbic, paralimbic, and mesocortical regions including the OFC and ACC, whereas craving correlated positively with activity in these regions. To circumvent technical difficulties in such drug self-administration fMRI paradigms (e.g., the effect of motion), paradigms that simulate compulsive behavior (e.g., gambling when it is clearly no longer beneficial) or studies of other compulsive disorders (e.g., obsessive–compulsive disorder) might offer invaluable insight into the circuits underlying this extreme form of loss of control in addiction.

Withdrawal Withdrawal occurs upon terminating drug use and is accompanied by a variety of symptoms, including excessive fatigue, irritability, anxiety, and anhedonia. Withdrawal symptoms vary in intensity and quality depending on the type of the drug (more intense for heroin and alcohol than for cocaine) and the amount of time that has passed since last drug use. As such, they are typically characterized as 'early' versus 'protracted' withdrawal symptoms.

In general, results of PET studies in drug-addicted individuals show long-lasting changes (mostly reduced sensitivity) in regional neural responsiveness to pharmacological challenges that share common properties with the drug of abuse. However, during early withdrawal, activity in certain regions (e.g., the OFC and striatum) may be higher in the drug-addicted than in healthy comparison subjects, which has been linked with the subjective perception of craving. During protracted withdrawal (1–6 weeks since last use), brain activity in these regions is usually lower, an effect that is most accentuated in the PFC and that persists for at least several months of drug-free periods.

Chronic Effects and Other Issues

Chronic drug effects Whereas acute drug administration temporarily increases DA neurotransmission, chronic drug consumption results in a marked and sustained decrease in DA activity, which appears to be long-lasting and to persist months after protracted detoxification, and which is associated with dysregulation of frontal brain regions. Thus, PET and SPECT studies have revealed lower DA D2 receptor availability at baseline in the striatum (including ventral striatum, where the NAc is located) in cocaine-, heroin-, alcohol-, and methamphetamine-addicted individuals compared to control subjects (Figure 8). In animal studies, the drug-induced changes in the dorsal striatum are observed after longer drug exposures than those observed in the NAc and have been interpreted to reflect further progression into the addicted state. Studies have also shown changes in mu opioid receptors as well as endogenous opioids in cocaine abusers.

Relapse. Individuals who have developed drug addiction are prone to relapse even after prolonged periods of abstinence. For example, at 3 months post-treatment, 40–60% of individuals in treatment for alcohol problems relapse to a first drink, whereas by 12 months this rate increases to 70–80%. These rates are similar for heroin users. In stimulant (cocaine or methamphetamine)-dependent patients, rates of relapse over 6 months to 1 year following treatment have typically ranged from 21% to 33%. Among cigarette smokers, of the 70% who want to quit smoking, only 4.7% are able to quit for at least 3 months. The recovery rates for other types of drugs of abuse are also poor. Contributing factors to relapse include exposure to stress, drug cues, poor coping strategies, and disrupted cognitive and emotional functioning. Rates of relapse are also related to the availability of the drug and the availability of controlled environments, support systems, and medications.

Although most devastating, the neurobiology that underlies relapse in addiction is not well understood. A promising line of investigation uses fMRI to link cognitive or emotional activation of the striatal–PFC circuit with subsequent relapse in treatment-seeking

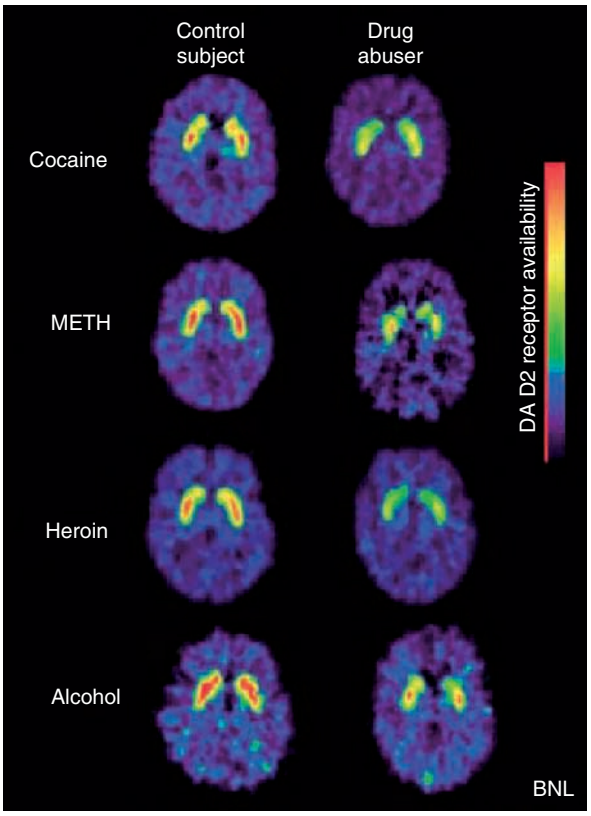

Figure 8 Positron emission tomography and its application to measuring dopamine receptor availability. This image shows lower striatal dopamine D2 receptor availability in drug users during withdrawal from cocaine, methamphetamine, heroin, and alcohol than in healthy comparison subjects. Reproduced from Volkow ND, Fowler JS, Wang GJ, and Swanson JM (2004) Dopamine in drug abuse and addiction: Results from imaging studies and treatment implications. *Molecular Psychiatry* 9: 557–569, with permission.

drug-addicted individuals; preliminary results suggest that stronger drug cue- or stress-induced activations in the putamen, ACC, and adjacent medial PFC in the early abstinence period are predictive of earlier (e.g., time to relapse) or more severe (e.g., consumption of the drug in greater amounts or for longer durations) relapse at follow-up in initially abstinent drug-addicted (cocaine, alcohol, and methamphetamine) subjects. These studies generally demonstrate that the regional fMRI activations are more sensitive in predicting relapse than task-related behavioral performance. This involvement of the PFC and ACC (and also the striatum and insula) in relapse prediction suggests that the neurobiological underpinnings of emotional and autonomic regulatory processes (e.g., suppression of craving, stress, or general negative affect) play a central role in the addicted individual's ability to remain abstinent.

Neurotoxicity *Brain structure*. Drugs of abuse have substantial neurotoxic effects as imaged in drug-addicted individuals. For example, studies using MRI have provided evidence of disruption in the OFC and ACC and in a broad network of related regions in addicted subjects compared to non-drug-using control subjects. Specifically, reductions in gray matter in frontal cortical regions (ACC and medial and lateral OFC) and limbic and paralimbic structures (including anteroventral insula and superior temporal cortex) were documented in abstinent cocaine-addicted individuals, methamphetamine users, alcoholics, and polysubstance users (Figure 9; see also Figure 10 for increases in striatal volumes in methamphetamine users). Studies using various MRI technologies have also provided evidence of white matter abnormalities in drug-addicted individuals, encompassing white matter hyperintensities, reduced white matter integrity, or overall increased volumes of white matter. In general, it appears that the frequency of white matter changes increases with age and with longer duration and amount of drug use. Studies have revealed changes specifically in frontal white matter integrity among stimulant-dependent subjects. Indeed, preclinical studies show that chronic administration of cocaine or amphetamine results in significant increases in dendritic branching and the density of dendritic spines in the PFC. Together, these changes in the synaptic connectivity and size of the PFC may underlie the changes in decision making, judgment, and cognitive control (e.g., increased impulsivity) that characterize drug addiction. Nevertheless, it is not clear whether these changes predate the development of addiction, are precipitated by the disease, or, probably, an interaction of the two.

Cerebrovascular events. The abuse of drugs such as cocaine, methamphetamine, and alcohol can lead to transient cerebral ischemia (through vasoconstriction but also through thrombosis formation), stroke, and hemorrhages. Indeed, imaging studies have provided evidence of marked regional decreases in cerebral blood flow present even in cocaine abusers who are neurologically asymptomatic. The vasoconstricting properties of these drugs are also likely to be implicated in their cardiotoxic effects that can result in myocardial infarction even in relatively young individuals.

The special case of methamphetamine. Of important mention here is methamphetamine, a highly addictive and potentially neurotoxic drug. In humans, imaging studies with PET have documented marked decreases in DAT, which were associated with psychomotor slowing and impaired learning and with incomplete recovery even after drug-free periods of 9 months. Furthermore, a hypermetabolic activity in the brain of the methamphetamine abusers was documented, which was interpreted to be consistent with inflammation/gliosis. Studies of methamphetamine abusers using magnetic resonance spectroscopy indeed have

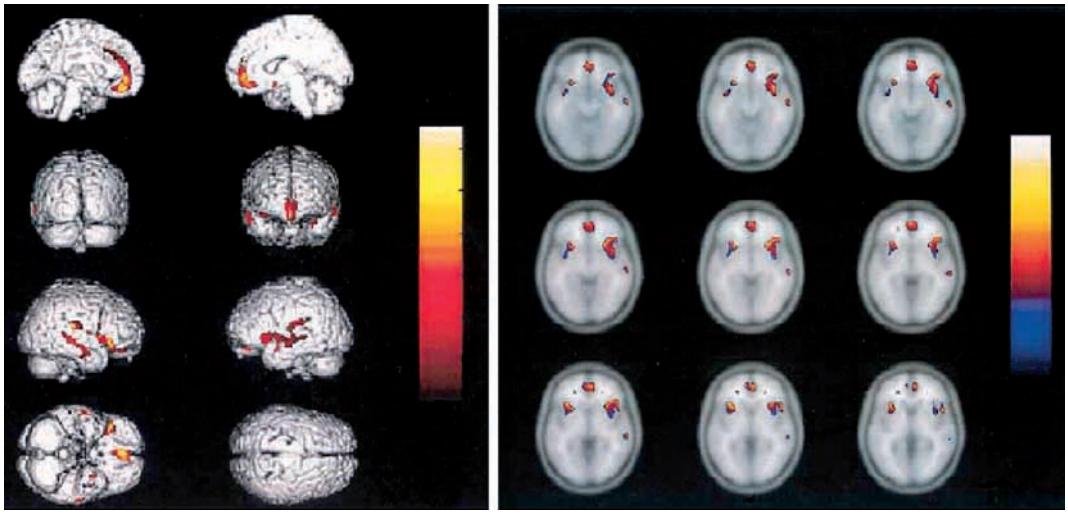

Figure 9 Magnetic resonance imaging and its application to measuring brain structure/volume differences between drug-addicted individuals and control subjects. (Left) Regions of decreased gray matter density (including the ventromedial orbitofrontal cortex through the anterior cingulate cortex) in 13 cocaine-dependent subjects compared to 16 cocaine-naive individuals. (Right) The percentage (0–14%) of decreased gray matter density in the cocaine group relative to the control group. From Franklin TR, Acton PD, Maldjian JA, et al. (2002) Decreased gray matter concentration in the insular, orbitofrontal, cingulate, and temporal cortices of cocaine patients. *Biological Psychiatry* 51: 134–142.

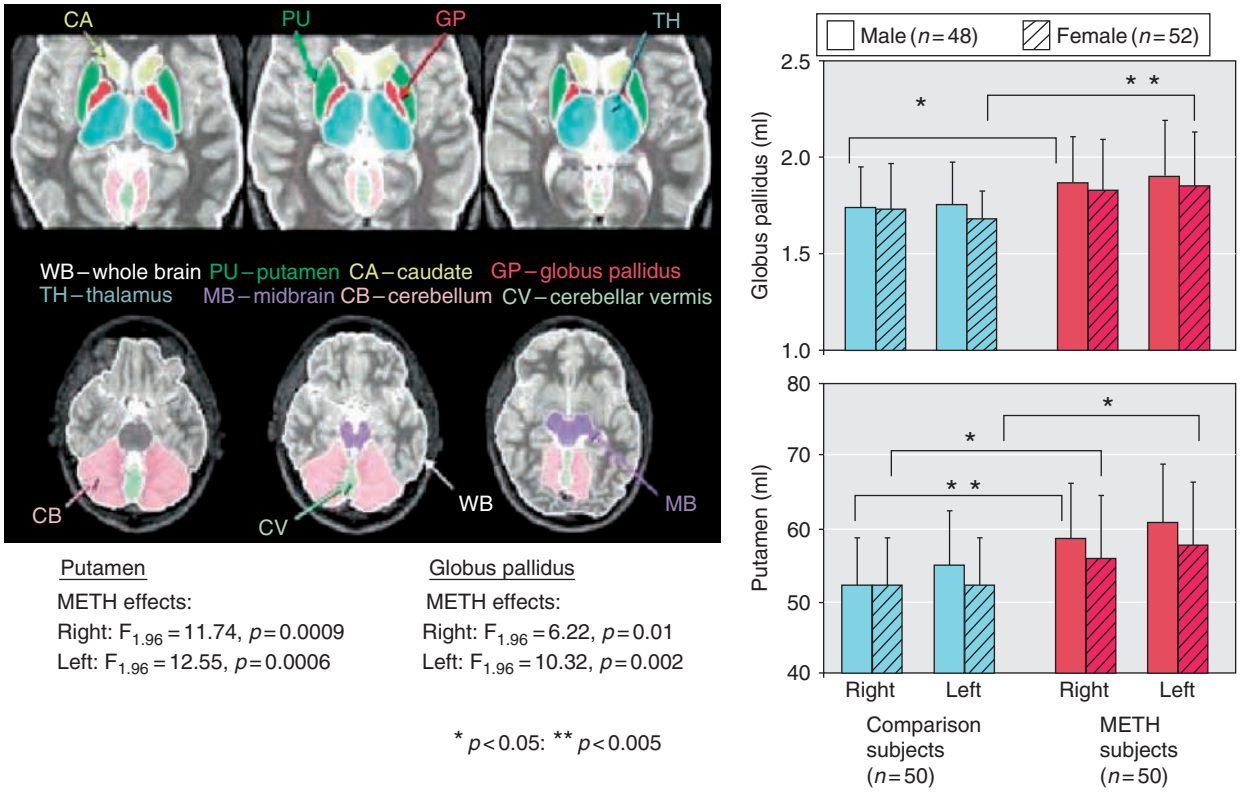

Figure 10 Magnetic resonance imaging and quantitative morphometric analyses applied to measuring brain structure/volume differences between drug-addicted individuals and control subjects. Fifty methamphetamine users showed enlarged putamen (10%) and globus pallidus (7–9%) compared to 50 age- and sex-matched comparison subjects. Smaller striatal structures were associated with poorer cognitive performance and greater cumulative drug use in the methamphetamine users. From Chang L, Cloak C, Patterson K, et al. (2005) Enlarged striatum in abstinent methamphetamine abusers: A possible compensatory response. *Biological Psychiatry* 57: 967–974.

shown increases in the concentration of myoinositol, a glial marker. Of note is that the hypermetabolic activity was most prominent in the parietal cortex (association cortex), a brain region that is poorly innervated by DA terminals. Therefore, these results most likely reflect the effects of methamphetamine on neuronal types or circuits other than DA, including degeneration of myelinated processes, damage to serotonergic terminals, damage to pyramidal glutamatergic cells, and an impact on the glutamate receptors (including hypersensitivity to glutamate excitotoxicity).

Imaging alcohol neurotoxicity. Consumption of very large amounts of alcohol can lead directly to death by depressing the respiratory center in the medulla. Chronic alcohol-related neuropathology encompasses peripheral neuropathy which results from multiple vitamin B deficiencies and which is usually reversible with adequate nutrition. Other neurological complications include anterior lobe cerebellar degenerative disease and Wernicke–Korsakoff syndrome (resulting from thiamine deficiency and characterized by severe short-term memory loss). In addition to the gray and white matter changes as reviewed previously, evidence for global brain atrophy in alcohol-dependent subjects has also been accumulating.

Effects of exposure to drugs *in utero* Prenatal exposure to drugs can cause permanent changes in the brain, and these changes affect behavior. Fetal alcohol syndrome, for example, is associated with exposure to high doses of alcohol in the prenatal development period. Brain alterations seen in these individuals include reduced brain size; alterations in shape, tissue density, and symmetry; and volumetric reductions and abnormalities in the cerebellum, basal ganglia, and corpus callosum. These shape and tissue abnormalities highlight the regional nature of brain morphological differences and brain functional abnormalities in addiction and suggest that assessment of brain–behavior relationships should be regionally specific. In addition to alterations in brain structure, individuals with histories of heavy prenatal alcohol exposure face long-lasting deficits that affect many aspects of cognitive and behavioral functioning (e.g., visuospatial functioning, verbal and nonverbal learning, and memory, attention, and executive functioning). The mapping of the regional brain differences to the neuropsychological abnormalities remains to be established.

Studies of prenatal exposure to other substances provide evidence that they too may affect the developmental trajectory of the brain. For example, preliminary prospective studies suggest that prenatal exposure to active maternal cigarette smoking is associated with alterations in medial temporal lobe function and concomitant deficits in visuospatial memory; these are mostly observed in adolescent tobacco smokers and are further associated with early adolescence attention and conduct problems. Nevertheless, it is important to recognize that other factors, encompassing the parent's socioeconomic status and history of antisocial personality disorder, are more powerful predictors of these impairments in the offspring.

A novel approach emphasizes the combination of MRI and PET technologies to assess pharmacokinetics of drugs of abuse on the fetal brain *in vivo*. This approach has been applied in a nonhuman primate model. Measurement of radioisotope distribution in maternal and fetal organs demonstrated that although the uptake of radioactivity into the fetal brain after the injection of C^{11}cocaine is lower and slower than in the maternal brain, C^{11}cocaine (or its labeled metabolites) readily crosses the placenta; a measurable quantity of C^{11}cocaine (or its labeled metabolites) accumulates in the fetal brain early after injection, although the highest accumulation occurs in the fetal liver.

Imaging predisposition and vulnerability Differences in DA circuits may be one mechanism underlying the variability in responsiveness to drugs of abuse, which in turn could influence vulnerability to drug addiction. Baseline measures of striatal DA D2 receptors with PET in nonaddicted subjects have been shown to predict subjective responses to the reinforcing effects of intravenous stimulants. Thus, individuals with lower DA levels reported liking methylphenidate more than individuals with higher DA levels, and this lower baseline DA D2 level may predispose to drug abuse through the initial hedonic reactions to the drug. A PET study suggested that it is the higher than normal DA D2 receptor availability that may be driving the differential predisposition to drug addiction: Nonalcoholic members of alcoholic families may be protected from developing alcoholism by their higher DA D2 receptor availability (in caudate and ventral striatum), possibly through its effect on the ability to regulate emotion. It is important to note that although tempting, causality cannot be inferred from these neuroimaging studies. Nevertheless, preclinical studies showed that upregulation of DA D2 receptors in the NAc dramatically reduced alcohol intake in animals previously trained to self-administer alcohol.

Note that other neurotransmitters have also been implicated in the vulnerability for developing drug addiction. For example, the endogenous opioid and

serotonergic systems have been implicated not only in alcohol action (consumption and reinforcement) and dependence but also in the vulnerability to alcoholism. Similarly, genetic studies have linked the opioid system with alcoholism and heroin addiction. Abnormalities in these neurotransmitter systems are also related to anxiety and depression, which may further increase the risk for developing drug addiction.

Summary

Neuroimaging studies are beginning to highlight the role of neurobiology in the core characteristics of drug addiction, encompassing drug intoxication, craving, bingeing, withdrawal, and relapse. These studies point to alterations in drug-addicted individuals in their brain structure, encompassing cortical and subcortical brain regions, and brain neurochemistry, encompassing DA and other neurotransmitters such as GABA, glutamate, serotonin, and opioids. A promising line of functional neuroimaging research has focused on the role of the OFC and ACC in understanding how the value of the drug as a reinforcer progresses to become more important than other reinforcers, which over time can no longer compete with the enhanced saliency value of the drug for the drug-addicted individual. Together, the goal of brain imaging studies is to offer biomarkers for particular outcomes in susceptible and drug-addicted individuals. It would be helpful, for example, to be able to identify at early adolescence those individuals who could benefit from more intensive drug addiction prevention efforts. It would similarly be helpful to identify among treatment-seeking drug-addicted individuals who would benefit from more tailored interventions to prevent relapse. Beyond highlighting the role of specific brain regions and neurotransmitters in particular drug-related behaviors, these studies promise to highlight the neurobiology underlying complex human phenomenology, encompassing motivation and free-will, or the 'mind.'

See also: Addiction: Neurobiological Mechanism; Alcoholism; Drug Addiction: Behavioral Neurophysiology; Drug Addiction: Behavioral Pharmacology of Drug Addiction in Rats; Drug Addiction: Cellular Mechanisms; Drugs Addiction: Actions; Substance Abuse and Dependence.

Further Reading

Ahmed SH and Koob GF (1998) Transition from moderate to excessive drug intake: Change in hedonic set point. *Science* 282: 298–300.

Benveniste H, Fowler JS, Rooney W, et al. (2005) Maternal and fetal 11C-cocaine uptake and kinetics measured *in vivo* by combined PET and MRI in pregnant nonhuman primates. *Journal of Nuclear Medicine* 46: 312–320.

Breiter HC, Gollub RL, Weisskoff RM, et al. (1997) Acute effects of cocaine on human brain activity and emotion. *Neuron* 19: 591–611.

Di Chiara G and Imperato A (1988) Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. *Proceedings of the National Academy of Sciences of the United States of America* 85: 5274–5278.

Fischman MW and Foltin RW (1991) Utility of subjective-effects measurements in assessing abuse liability of drugs in humans. *British Journal of Addiction* 86: 1563–1570.

Fowler JS, Logan J, Wang GJ, et al. (2003) Low monoamine oxidase B in peripheral organs in smokers. *Proceedings of the National Academy of Sciences of the United States of America* 100: 11600–11605.

Goldstein RZ, Alia-Klein N, Tomasi D, et al. (2007) Is decreased prefrontal cortical sensitivity to monetary reward associated with impaired motivation and self-control in cocaine addiction? *American Journal of Psychiatry* 164: 43–51.

Goldstein RZ and Volkow ND (2002) Drug addiction and its underlying neurobiological basis: Neuroimaging evidence for the involvement of the frontal cortex. *American Journal of Psychiatry* 159: 1642–1652.

Koob GF and Bloom FE (1988) Cellular and molecular mechanisms of drug dependence. *Science* 242: 715–723.

Leshner AI (1997) Addiction is a brain disease, and it matters. *Science* 278: 45–47.

O'Brien CP (2005) Anticraving medications for relapse prevention: A possible new class of psychoactive medications. *American Journal of Psychiatry* 162: 1423–1431.

Ritz MC, Lamb RJ, Goldberg SR, and Kuhar MJ (1987) Cocaine receptors on dopamine transporters are related to self-administration of cocaine. *Science* 237: 1219–1223.

Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36: 241–263.

Tremblay L and Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. *Nature* 398: 704–708.

Volkow ND and Fowler JS (2000) Addiction, a disease of compulsion and drive: Involvement of the orbitofrontal cortex. *Cerebral Cortex* 10: 318–325.

Volkow ND, Fowler JS, and Wang GJ (2003) The addicted human brain: Insights from imaging studies. *Journal of Clinical Investigation* 111: 1444–1451.

Volkow ND, Fowler JS, Wang GJ, and Swanson JM (2004) Dopamine in drug abuse and addiction: Results from imaging studies and treatment implications. *Molecular Psychiatry* 9: 557–569.

Volkow ND, Wang GJ, Begleiter H, et al. (2006) High levels of dopamine D2 receptors in unaffected members of alcoholic families: Possible protective factors. *Archives of General Psychiatry* 63: 999–1008.

Volkow ND, Wang GJ, Fowler JS, et al. (1997) Decreased striatal dopaminergic responsiveness in detoxified cocaine-dependent subjects. *Nature* 386: 830–833.

Volkow ND, Wang GJ, Fischman MW, et al. (1997) Relationship between subjective effects of cocaine and dopamine transporter occupancy. *Nature* 386: 827–830.

Volkow ND, Wang GJ, Telang F, et al. (2006) Cocaine cues and dopamine in dorsal striatum: Mechanism of craving in cocaine addiction. *Journal of Neuroscience* 26: 6583–6588.

Zink CF, Pagnoni G, Martin-Skurski ME, Chappelow JC, and Berns GS (2004) Human striatal responses to monetary reward depend on saliency. *Neuron* 42: 509–517.

Drugs Addiction: Actions

R A Wise, National Institute on Drug Abuse, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The immediate actions of most addictive drugs are at identified binding sites on various neurotransmitter receptors or transporters. These immediate actions are varied. The psychomotor stimulants amphetamine and cocaine act in circuitry involved in sleep–wakefulness, arousal, and attention, circuitry that is thought to be out of balance in depression, schizophrenia, and Parkinson's disease. The opiates heroin and morphine act in circuitry of thermoregulation, pain modulation, feeding, and control of intestinal contractions. Nicotine acts in circuitry at all levels of the nervous system, from the neuromuscular junction to the cerebral cortex. Alcohol destabilizes membranes and acts at receptors in multiple brain regions. Most of the actions of addictive drugs have little to do with the fact that the drugs are addictive. The addictive actions of addictive drugs, on the other hand, are a subset of each drug's actions, a subset presumed to result in activation of a common forebrain circuit that is involved in appetitive motivation.

The common denominator of addictive drugs is their ability to form compulsive response habits. The brain mechanisms through which addictive drugs come to establish response habits and dominate behavior are the brain mechanisms through which the natural pleasures of life – food, water, shelter, sexual interactions – have established response habits and dominated our behavior through most of our evolutionary history. Thus, it is within the brain circuitry of natural motivated behavior – circuitry dubbed 'reward circuitry' historically – that the brain mechanisms of addiction have been sought.

The concept of specialized reward circuitry in the brain arose from the discovery by Olds and Milner that direct electrical stimulation of certain portions of the brain could be more powerfully rewarding than food, sex, or the retrieval of a mother's pups. Brain stimulation was used to map anatomical sites contributing to this circuitry, and neurotransmitter antagonists and drug rewards – most notably a psychomotor stimulant and an opiate reward – were used to identify the chemical coding of this circuitry. Eventually, central microinjections of drugs and neurotransmitters were used to pinpoint synaptic junctions where addictive drugs and other substances – neurotransmitter agonists, antagonists, or enzymes – have rewarding actions. Most of our current knowledge of the actions of drugs of addiction comes from such studies. This line of investigation has implicated as core elements in the brain mechanisms of drug reward the mesolimbic and mesocortical dopamine systems, GABAergic neurons of the nucleus accumbens that are the efferent targets of the mesolimbic system, and GABAergic neurons of the midbrain that are both input neurons to the mesolimbic and mesocortical dopamine neurons and output neurons of the ventromedial subdivision of the extrapyramidal system. Various drugs of addiction act at each of these targets.

Psychomotor Stimulants

The psychomotor stimulants amphetamine and cocaine are habit-forming, largely because of actions at monoamine transporters causing elevations of extracellular dopamine. Blockade of either D1 type or D2 type dopamine receptors blocks the habit-forming effects of cocaine. Cocaine blocks the plasma membrane dopamine transporter, limiting the reuptake of extracellular dopamine and thus prolonging its actions at receptors on target neurons in a variety of forebrain structures. The olfactory tubercle, the medial prefrontal cortex, and the nucleus accumbens shell appear to be the most important dopamine terminal fields for cocaine reward. Rats can be trained to lever-press for cocaine injections restricted to each of these regions; the olfactory tubercle appears to be the most sensitive to the rewarding effects of cocaine. Cocaine blocks the noradrenergic and serotonergic transporters as well as the dopamine transporter, and each of these can take up dopamine under the right conditions. While blockade of the dopamine transporter appears to be most critical for cocaine reward in normal animals, cocaine remains habit-forming in mice lacking the dopamine transporter. In these animals, the blockade of the serotonin transporter appears critical for cocaine self-administration; cocaine is not rewarding in mice with both the dopamine and the serotonin transporters deleted. The mechanism for cocaine reward in these mice is not firmly established, but it is known that serotonin can inhibit γ-aminobutyric acid (GABA) release in the ventral tegmental area (VTA), which, in turn, should disinhibit the dopamine system. Cocaine is known to elevate nucleus accumbens dopamine levels in the nucleus accumbens of mice with the dopamine transporter deletion.

Whereas cocaine merely blocks dopamine uptake, amphetamine acts to reverse the dopamine transporter and cause release of cytosolic dopamine even in the

absence of nerve impulses. Amphetamine alters the subcellular localization of the vesicular monoamine transporter, apparently elevating cytosolic dopamine levels for release by reverse transport. Amphetamine also reverses the noradrenergic and serotonergic transporters. Amphetamine is self-administered directly into the nucleus accumbens, but not into the medial prefrontal cortex; self-administration into the olfactory tubercle has not yet been tested. Like amphetamine, a mixture of D1 type or D2 type dopamine agonists is rewarding when injected into the nucleus accumbens. Blockade of either D1 type or D2 type dopamine receptors blocks the habit-forming effects of amphetamine.

Opiates

The opiates morphine and heroin are habit-forming because of actions at receptors for the endogenous opioids enkephalin, β-endorphin, endomorphin, and dynorphin. Three classes of opioid receptors have been identified: μ opioid receptors activated by the endogenous opioids endomorphin, enkephalin, and β-endorphin; δ opioid receptors activated by enkephalin and β-endorphin; and κ opioid receptors activated by dynorphin. Of these the μ and δ opioid receptors appear to be most important for the rewarding actions of opiates; activation of κ opioid tends to antagonize reward function.

Morphine is rewarding when directly injected into two brain regions, the VTA and the nucleus accumbens. In the VTA morphine acts at μ opioid receptors on neurons containing the inhibitory neurotransmitter GABA. GABAergic neurons in this region normally hold their dopaminergic neighbors under inhibitory control; thus, morphine's action at μ receptors on these GABAergic neurons results in disinhibition of the dopamine system. These or adjacent GABAergic neurons also project to distal targets, and opiates have dopamine-independent as well as dopamine-dependent rewarding actions in the VTA. The selective μ opioid D-Ala^2,*N*-MePhe^4,Gly^5-ol enkephalin (DAMGO) is rewarding and activates the mesolimbic dopamine system when injected into this region, presumably acting on μ opioid receptors expressed by GABAergic neurons that normally inhibit their dopaminergic neighbors.

Morphine also acts at δ opioid receptors and the selective δ opioid D-Pen^2,D-Pen^5 enkephalin (DPDPE) is also rewarding. It also activates the mesolimbic dopamine system when injected into this region; as an activator of the dopamine system and also as a reward, it is approximately two orders of magnitude less effective than DAMGO. This may be due to the much more sparse expression of the δ opioid receptor in relation to that of the μ opioid receptor in this region.

The relative effectiveness of selective μ and δ opioid injections into the nucleus accumbens is not known, but again the presumed target is GABAergic neurons: in this case, the medium spiny output neurons of nucleus accumbens that receive dopaminergic input from the VTA. The κ opioids antagonize reward function, in part by inhibiting ventral tegmental dopamine neuron cell firing and in part by increasing dopamine uptake at dopaminergic axon terminals.

Nicotine

Nicotine acts at a subtype of acetylcholine receptor, the nicotinic acetylcholine receptor. This receptor is found widely distributed in the brain and also in the neuromuscular junction and the autonomic ganglion. Nicotine infused into either the VTA or the nucleus accumbens elevates nucleus accumbens dopamine levels, and nicotine injections directly into at least one of these regions – the VTA – are habit forming. Acetylcholine is the primary neurotransmitter in pedunculopontine and laterodorsal tegmental nuclei that provide input to the ventral tegmental dopamine neurons involved in reward function. Dopaminergic neurons in the VTA express both nicotinic and muscarinic acetylcholine receptors, the former is important for the rewarding effects of nicotine and the latter is important for the rewarding effects of direct electrical stimulation of reward sites in the lateral hypothalamus and adjacent portions of the medial forebrain bundle.

Alcohol

Ethyl alcohol (ethanol), like most other addictive drugs, elevates dopamine levels in the nucleus accumbens and other terminal fields of the mesocorticolimbic dopamine system. Unlike most other addictive drugs, the mechanism of action of ethanol does not seem to involve the binding of the drug to an identified neurotransmitter receptor or transporter. Ethanol disturbs lipid membranes, and while ethanol is not known to bind to a known receptor, it nonetheless alters ion movement through channels associated with known receptors. The neurotransmitter receptors most strongly implicated at present are $GABA_A$ and *N*-methyl-D-aspartate (NMDA) receptors; however, it is possible that $GABA_A$-associated effects are due to a presynaptic influence on GABA release rather than a postsynaptic effect on the receptor itself. Ethanol also appears to modulate ion flow through the purinergic P2X receptor and the serotonergic 5HT3 receptor.

Whatever the transduction mechanism, ethanol increases dopaminergic cell firing and nucleus accumbens dopamine levels, and is habit forming, at least in part, for that reason. Localized injections of ethanol

into the VTA are habit forming, and this is believed to be dopamine dependent.

Benzodiazepines and Barbiturates

Benzodiazepines and barbiturates act to enhance $GABA_A$ receptor function. Their sites of rewarding action are not known, but systemically administered benzodiazepines decrease dopamine levels in the nucleus accumbens. $GABA_A$ agonists and antagonists are rewarding or reward-inhibiting in the VTA, depending on the site of injection. The $GABA_A$ agonist muscimol is rewarding when administered into the posterior but not the anterior VTA; the $GABA_A$ antagonist picrotoxin is rewarding when administered into the anterior but not the posterior VTA. GABAergic neurons are ubiquitous in the brain, however, and the VTA may play a minor role relative to the nucleus accumbens in the habit-forming actions of benzodiazepines and barbiturates. Inasmuch as the mesolimbic dopamine system projects to GABAergic neurons in the nucleus accumbens, these drugs may well act more importantly at sites efferent to the dopamine system, within an anatomical cascade of GABAergic projections from the nucleus accumbens to the ventral pallidum, from the ventral pallidum and the nucleus accumbens to the VTA, and from the VTA to the medial thalamus, superior colliculus, and mesencephalic locomotor region.

Addictive Consequences

The mechanisms by which drugs become habit-forming are presumed to be mechanisms embedded in the circuitry the drugs activate and not in the immediate actions of the drugs at their receptors. The brain circuitry of reward function is not well understood, but the midbrain dopamine systems appear to play a major part. Injection of dopamine or dopamine agonists is habit-forming, and the psychomotor stimulants are not habit-forming when dopamine synaptic function is pharmacologically blocked or when forebrain dopamine is depleted. The fact that opiates, alcohol, nicotine, and cannabis elevate forebrain dopamine levels implicates dopamine in the rewarding effects of these drugs as well. The fact that blockade of dopamine receptors does not block all the habit-forming properties of morphine, nicotine, or phencyclidine indicates that dopamine does not play an exclusive, but merely a central, role in drug reward.

The ability of rewarding stimuli to establish response habits and to make otherwise neutral stimuli attractive is termed 'reinforcement.' Reinforcement is the process of 'stamping in' associations between stimuli and between stimuli and responses. The essential consequence of reinforcement is the enhanced consolidation of short-term memory for the rewarded response and the associated stimuli into a long-term memory that increases the probability that the response will recur and the associations remain established. Dopamine release appears to accomplish this. Thus the posttraining administration of the dopamine releaser amphetamine enhances the memory for whatever was learned in the training session, regardless of its relevance to amphetamine. For example, amphetamine given after an animal learns the location of a hidden underwater platform increases performance when the animal is subsequently required to find the platform. The mechanism by which dopamine enhances memory consolidation is not fully understood, but the effects of dopamine on intracellular guanine nucleotide binding proteins (G-proteins) are essential in most brain regions for two cellular models of synaptic plasticity: long-term potentiation and long-term depression. The role of dopamine in response reinforcement (habit-learning) appears to involve the metabotropic receptor-mediated action of dopamine on intracellular G-proteins in neurons receiving dopaminergic synapses or local nonsynaptic dopamine release.

A question of major recent interest is whether addictive drugs cause unique neuroadaptations in brain reward circuitry that are above and beyond the cellular correlates of learning and memory. Both cocaine and morphine cause a number of neuroadaptations, some of which result in tolerance to the drugs and some of which result in sensitization to the drugs. Some of the neuroadaptations are in dopamine neurons and others are in their GABAergic target neurons in the nucleus accumbens (and probably elsewhere). These neuroadaptations are common to the two drug classes, despite the fact that the drugs have opposite effects on the dopamine neurons while having common effects on the target neurons (morphine disinhibits the dopamine neurons while cocaine, by elevating dopamine levels at somatodendritic autoreceptors, inhibits them). It has been postulated that neuroadaptations caused by addictive drugs are responsible for the compulsive nature of drug self-administration and the fact that cravings for drug come to dominate cravings for more natural rewards such as food or sexual interaction. It remains to be determined whether addictive drugs are unique in the ability to cause these neuroadaptations; recent evidence suggests that some such neuroadaptations can be caused by intermittent exposure to concentrated sucrose solutions or to sex.

See also: Addiction: Neurobiological Mechanism; Alcoholism; Basal Ganglia: Habit; Drug Addiction: Behavioral Pharmacology of Drug Addiction in Rats; Drug Addiction: Cellular Mechanisms; Drug Addiction:

Neuroimaging; Prefrontal Contributions to Reward Encoding; Reward Systems: Human; Stimulant and Wake-Promoting Substances; Substance Abuse and Dependence.

Further Reading

Bals-Kubik R, Shippenberg TS, and Herz A (1990) Involvement of central μ and ∂ opioid receptors in mediating the reinforcing effects of β-endorphin in the rat. *European Journal of Pharmacology* 174: 63–69.

Berke JD and Hyman SE (2000) Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* 25: 515–532.

Corrigall WA, Franklin KB, Coen KM, et al. (1992) The mesolimbic dopaminergic system is implicated in the reinforcing effects of nicotine. *Psychopharmacology* 107: 285–289.

Devine DP and Wise RA (1994) Self-administration of morphine, DAMGO, and DPDPE into the ventral tegmental area of rats. *Journal of Neuroscience* 14: 1978–1984.

DiChiara G and Imperato A (1988) Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. *Proceedings of the National Academy of Sciences of the United States of America* 85: 5274–5278.

Gardner EL (2005) Endocannabinoid signaling system and brain reward: Emphasis on dopamine. *Pharmacology, Biochemistry and Behavior* 81: 263–284.

Goeders NE and Smith JE (1983) Cortical dopaminergic involvement in cocaine reinforcement. *Science* 221: 773–775.

Hoebel BG, Monaco AP, Hernandez L, et al. (1983) Self-injection of amphetamine directly into the brain. *Psychopharmacology* 81: 158–163.

Ikemoto S and Wise RA (2004) Mapping of chemical trigger zones for reward. *Neuropharmacology* 47(supplement 1), 190–201.

Jones S and Bonci A (2005) Synaptic plasticity and drug addiction. *Current Opinion in Pharmacology* 5: 20–25.

Lupica CR, Riegel AC, and Hoffman AF (2004) Marijuana and cannabinoid regulation of brain reward circuits. *British Journal of Pharmacology* 143: 227–234.

Olds ME (1982) Reinforcing effects of morphine in the nucleus accumbens. *Brain Research* 237: 429–440.

Rodd ZA, Melendez RI, Bell RL, et al. (2004) Intracranial self-administration of ethanol within the ventral tegmental area of male Wistar rats: Evidence for involvement of dopamine neurons. *Journal of Neuroscience* 24: 1050–1057.

Wise RA (1987) Intravenous drug self-administration: A special case of positive reinforcement. In: Bozarth MA (ed.) *Methods of Assessing the Reinforcing Properties of Abused Drugs*, pp. 117–141. New York: Springer.

Wise RA (2001) Neural circuitry of desire. In: Hetherington M (ed.) *Food Cravings and Addiction*, pp. 33–58. London: Leatherhead Press.

Wise RA (2004) Dopamine, learning and motivation. *Nature Reviews Neuroscience* 5: 483–494.

D-Serine: From Its Synthesis in Glial Cell to Its Action on Synaptic Transmission and Plasticity

S H R Oliet and J-P Mothet, Institut National de la santé et de la Recherche Médicale U862, Bordeaux, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When thinking about brain signaling, neurophysiologists traditionally refer to the transfer of information from one neuron to another. Accordingly, the chemical synapse is viewed as the structure in which such communication occurs between the pre- and postsynaptic elements and where neurotransmitters are released from the presynaptic bouton to activate specific receptors localized on the postsynaptic neuron. This model, however, needs to be revised in view of the findings describing the role of glial cells as the third element of the chemical synapse. At the anatomical level, astrocytic processes enwrap up to 60% of the synaptic volume. Despite such an intimate structural relationship, astrocytes were believed to ensure only a housekeeping function at synapses. There is now accumulating evidence indicating that glial cells, and particularly astrocytes, contribute actively to synapse development, synaptic transmission, and neuronal excitability. These results fuel the emerging concept of the tripartite synapse, which considers astrocytes as an integral part of central and peripheral synapses. According to this model, glial cells first sense synaptic activity through a broad variety of ion channels, transporters, and receptors expressed at their surface. Synaptic activation of glial cells then triggers intracellular second-messenger pathways, including Ca^{2+} increases that vary according to the synaptic inputs solicited and the glial receptors involved. In turn, activation of these second-messenger pathways induces the release of active substances termed gliotransmitters (by analogy to neurotransmitters). These gliotransmitters mediating astrocyte-to-neuron signaling include glutamate, taurine, and adenosine triphosphate (ATP), which can be hydrolyzed in adenosine.

Another major gliotransmitter which could have a major role in brain signaling is D-serine. The discovery of this amino acid in the brain has forced us to reconsider the dogma that only L-isomers of amino acids occur in mammals. Organic molecules, and therefore biological molecules such as D-amino acids, are based on the chemistry of the carbon atom. Carbon atoms can have up to four bonded groups attached to them in three-dimensional space, forming a tetrahedron. Because of this structure, carbon-containing molecules can have the same four constituents, yet can differ in their structure by the location of the four groups in space. This property of carbon-based molecules is known as chirality. As an example, our hands are mirror images of one another, but they cannot be superimposed on one another. Similarly, carbon atoms with four different groups attached occur in two forms, known as enantiomers, that are not superimposable mirror images of one another (see Figure 1).

Although the chemical and physical properties of L-amino acids and D-amino acids are extremely similar, only L-amino acids seemed to have been selected from the origin of life on the primitive Earth. In this chemical evolutionary step, D-amino acids were eliminated, and it was thought that all living organisms were composed only of L-amino acids. This asymmetry in biology is assumed to be a feature of fundamental physics because it turns out that natural L-amino acids are more stable than their 'unnatural' mirror images, D-amino acids. Until the last 30 years, it was believed that D-amino acids were excluded from living systems except for D-amino acids in the cell walls of microorganisms. Now, biologists have discovered that nature can deal with at least two D-amino acids, D-serine and D-aspartic acid, in higher organisms. Of these two atypical amino acids, D-serine is potentially very important for the nervous system because it is likely be the major endogenous ligand for the strychnine-insensitive glycine-binding site of *N*-methyl-D-aspartate receptors (NMDARs), key receptors for excitatory transmission and cognitive functions. Since the late 1980s, it has been established that activation of NMDARs required glutamate and a co-agonist that was first identified as glycine. Yet significant levels of D-serine are present in rodent brain areas enriched in NMDARs, suggesting that D-serine might be an alternative to glycine for NMDAR activation. D-Serine and serine racemase (SR), an enzyme that synthesizes D-serine from L-serine, are mainly, if not exclusively, localized to astrocytes. That glial-derived D-serine is an endogenous ligand of NMDARs and was unambiguously demonstrated by the use of D-amino acid oxidase (DAAO), an enzyme that selectively degrades D-serine but not glycine. Under conditions in which D-serine levels are dramatically reduced with DAAO, NMDAR activity is markedly impaired in different brain regions. It is now obvious that glial cells, through the release of D-serine, contribute actively to NMDAR function in the mammalian brain.

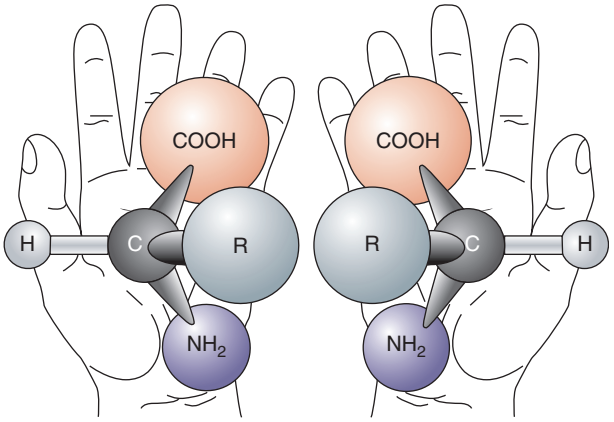

Figure 1 The chirality in your hands. The term chiral (from the Greek word for hand) is used to describe an object which is not superimposable on its mirror image, for example, our two hands. This property is based on the fact that carbon atoms (C) bind four groups – COOH, NH_2, H, or any other radical (R) – in a three-dimensional space. Such a tetrahedral hybridized carbon with four different groups always forms a chiral center. Conversely, a substrate carbon atom is not chiral if two out of four of its groups are identical, as it is in the case of the amino acid glycine.

Localization of D-Serine

D-Serine is present in significant amounts in the brain of rodents and humans, where its levels (~500 μM) are up to one-third of the total free serine pool. Its distribution is heterogeneous with the highest concentrations in the telencephalon and the developing cerebellum. This pattern of distribution throughout the rat brain resembles that of NMDARs. In contrast, glycine immunoreactivity does not correspond to NMDAR distribution excepted for some regions where D-serine is also present. Detailed analysis of D-serine staining shows that it is mostly present in astrocytes that ensheath synapses. In the developing cerebellum, D-serine is localized to Bergmann glia, whereas in adults it declines to negligible levels. Recent investigations have found that microglia cells as well as Schwann cells contain significant amount of D-serine and SR, which are also found in Müller glial cells in the retina and in the supporting cell of the vestibular sensory epithelium. Whether D-serine and SR are also present in other types of glial cells, such as oligodendrocytes, pituicytes, tanycytes, or ependymal cells, remains unknown. Although at very low levels, neuronal immunoreactivity for D-serine has been still observed in the neurons of the cerebral cortex, the brain stem, and the olfactory bulb.

Synthesis and Degradation of D-Serine

The idea that D-amino acids and particularly D-serine serves specific roles in the brain was strengthened by the discovery of its metabolic pathway. High levels of D-serine in the mammalian brain are generated by the activity of a pyridoxal 5′-phosphate (vitamin B_6)-dependent enzyme, SR. This enzyme not only converts L- into D-serine, but also converts D- into L-serine, with a lower affinity. The enzymatic-catalyzed racemization of L-serine proceeds by removal of a proton from the asymmetric C–H bond of the amino acid to form a carbanion intermediate. The trigonal carbon atom of the carbanion, having lost the original asymmetry, then recombines with a proton to regenerate as an inverted tetrahedral structure corresponding to D-serine. The distribution of SR closely resembles that for D-serine, with the highest expression in the hippocampus and corpus callosum and very low levels in the amygdala, subthalamic nuclei, and the thalamus. Regarding cellular distribution, detailed analysis indicates that SR, like D-serine, is mostly confined to astrocytes, although substantial expression of SR is present in some neurons. However, due to the numeric preponderance of astroglia cells over neurons, glial cells remain the principal source of D-serine in the brain.

SR activity can be negatively regulated by a series of cellular compounds. Thus, glycine and a series of metabolites related to L-aspartic acid competitively inhibit the enzyme. Because glycine concentrations in astrocytes are approximately 3–6 mM, glycine would constitutively inhibit SR activity unless glycine and SR show different compartmentalizations within the astrocyte cytosol. The metabolic pathway for D-serine degradation remains more elusive. Mammalian D-serine can be metabolized by the peroxisomal flavoprotein DAAO, an enzyme highly present in astrocytes of the hindbrain and cerebellum. Adult DAAO-deficient mice display increased D-serine levels, especially in areas where it normally occurs at low levels. However, although DAAO protein is present in the forebrain, an area enriched in D-serine, levels of the D-amino acid appear relatively unchanged in this region in DAAO-deficient mice. This result suggests that other mechanisms are probably implicated in regulating D-serine concentrations in this brain area.

SR and DAAO may not work in isolation because SR and DAAO activities are controlled in opposite ways by nitric oxide (NO). Although NO enhances SR activity, it decreases that of DAAO, thus down-regulating the intracellular levels of D-serine. In turn, D-serine inhibits NO synthase in glial cells and stimulates the production of NO in neurons. Neuronal NO could thus represent an inhibitory feedback mechanism tightly regulating D-serine metabolism in astrocytes and thereby preventing its overproduction and excessive stimulation of NMDARs.

Release of D-Serine

Glial cells detect changes in their environment through a broad variety of ion channels, transporters, and receptors expressed at their surface. Specific activation of metabotropic and non-NMDA glutamate receptors on astrocytes induces the efflux of D-serine from these cells in culture (Figure 2).

Whether activating these glutamate receptors is the only stimulatory pathway for D-serine release remains to be determined. Similarly, the conditions during which the activation of glial glutamate receptors, and, consequently, D-serine release occur in response to afferent synaptic activity are unknown. Interestingly, release of glutamate from glial cells also occurs when metabotropic and non-NMDA glutamate receptors are stimulated. An important issue for D-serine release, and for gliotransmission in general, is the identification of the molecular pathway responsible for the efflux of active substances from cells that were long thought to be passive and/or unexcitable. Astrocytes, like all eukaryotic cells, use secretory lysosomes to transport new membrane and proteins to the plasma membrane during constitutive exocytosis. Most cell types also possess a pathway of regulated exocytosis in which secretory vesicles undergo Ca^{2+}-dependent fusion with the plasma membrane. Until recently, Ca^{2+}-regulated exocytosis was considered as a hallmark of neurons in the nervous system. Both constitutive and regulated secretory pathways require specialized proteins to bring together the membranes of the vesicles with the plasma membrane. The soluble *N*-ethylmaleimide-sensitive factor (NSF) attachment protein receptors (SNAREs) are leading candidates for mediating membrane fusion and most of them are present in glial cells. In addition, glial cells contain also synaptic-like microvesicles (SLMV) and large dense-core vesicles (LDCV), as revealed by electron microscopy and confocal microscopy analyses. Results obtained over the last decade using electrophysiological, biochemical, and imaging techniques suggest that astrocytes use a Ca^{2+}-regulated SNARE-dependent exocytosis to release glutamate from SLMVs but also ATP and peptides from LDCVs. In cultured astrocytes, activation of metabotropic and non-NMDA receptors effectively triggers a Ca^{2+}- and SNARE-dependent release of D-serine. In this experimental model, modifying extracellular or intracellular Ca^{2+} concentrations or depleting thapsigargin-sensitive Ca^{2+} stores, considerably impairs D-serine release. These results reveal that activation of non-NMDARs is the major source of extracellular Ca^{2+} and that mobilization of Ca^{2+} from internal stores is necessary for D-serine release. This is consistent with the observation that both inositol triphosphate (IP_3)- and caffeine/ryanodine-sensitive Ca^{2+} stores control the release from glial cells of glutamate, another gliotransmitter. Furthermore, inhibiting exocytosis with clostridial toxins or preventing amino acid uptake into vesicles with concanamycin or bafilomycin, impairs the release of D-serine from astrocytes. These results are thus consistent with a vesicular storage and release of the D-amino acid. Interestingly, astrocytic glutamate also appears to be stored in, and released from, vesicles. If glutamate and D-serine were co-stored in the same vesicles, this would be the perfect gliotransmitter cocktail to activate NMDARs. This hypothesis is indeed supported by the observation that some D-serine immunoreactivity is found in vesicles expressing the vesicular transporter for glutamate. The final proof that D-serine and glutamate are co-stored and co-released, however, awaits further investigations. The existence of a vesicular pathway does not exclude the possibility that other routes for D-serine release coexist in glial cells. Such nonvesicular pathways could involve processes that have been linked already to gliotransmitter release such as hemichannels, $P2X_7$ receptors, volume-sensitive channels, and reversal of transporters (Figure 3).

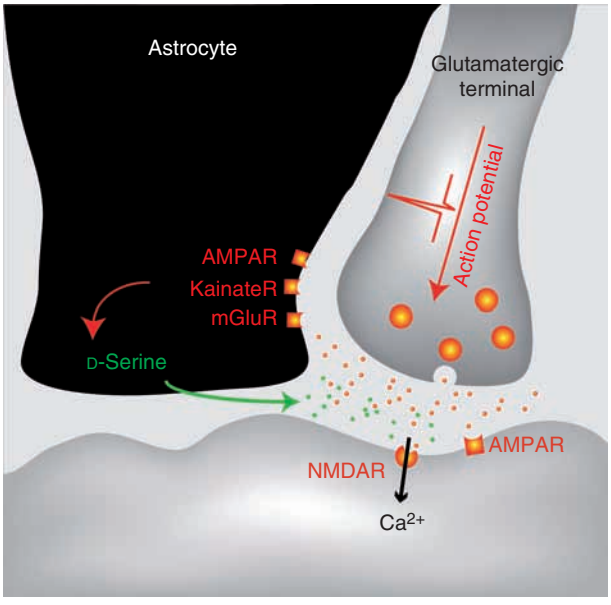

Figure 2 Induction of D-serine release at glutamatergic synapses. Synaptically released glutamate diffuses locally to neighboring glial processes and binds to AMPA (AMPAR), kainate (kainaiteR) and/or metabotropic (mGluR) receptors. This triggers the release of D-serine within the synaptic cleft, enabling NMDAR activation on the postsynaptic neuron. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; AMPAR, AMPA receptor; kainaiteR, kainaite receptor; mGluR, metabotropic glutamate receptor; NMDAR, *N*-methyl-D-aspartate receptor.

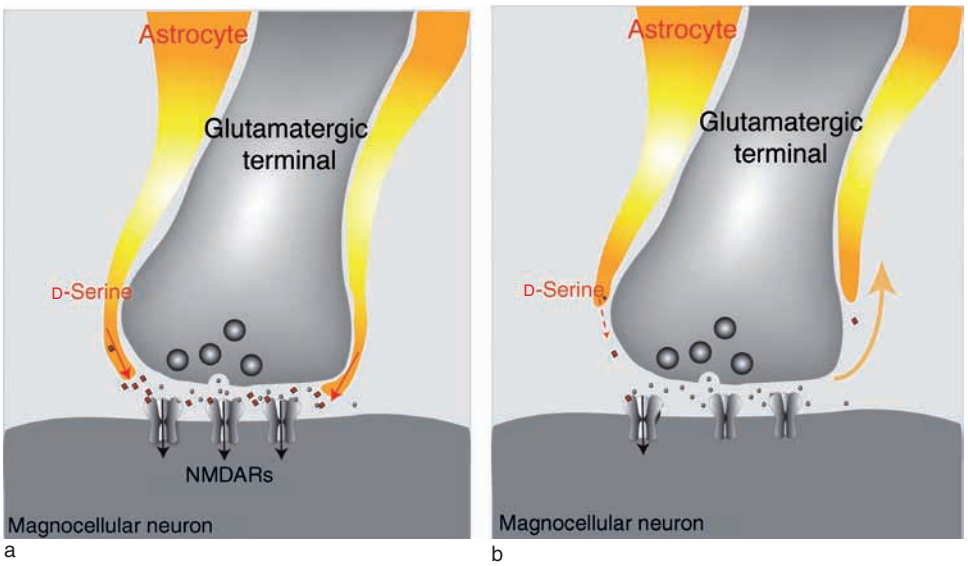

Figure 3 Concentration of D-serine in the synaptic cleft and the number of NMDARs available: (a) control, (b) with glial withdrawal in lactating rats. In (a), the concentration of D-serine is sufficient to make most NMDARs on magnocellular neurons available for activation by synaptically released glutamate. In (b), glial withdrawal causes a reduction of synaptic D-serine levels, thereby diminishing the number of NMDARs available. NMDAR, *N*-methyl-D-aspartate receptor.

Another important issue regarding D-serine release, and more generally gliotransmission, is to know whether specialized sites for its release exist in astrocytes. Regarding D-serine, it is not clear whether its release is restricted to the fine processes apposed to NMDARs or whether it can occur from any compartment of the glial cell. Interestingly, astroglial cells have functionally distinct compartments, known as microdomains, where localized high Ca^{2+} increases occur and which tightly enwrap synapses. Furthermore, electron microscopy analysis has revealed that glutamate-containing vesicles are found in glial processes that contact dentritic spines bearing NMDARs. It is thus tempting to speculate that D-serine release occurs from such localized glial microdomains in close vicinity to the synapses and to NMDARs. This hypothesis is supported by electrophysiological recordings obtained in the hippocampus and in the hypothalamus demonstrating that the level of occupancy of the NMDAR glycine site is higher at synaptic than at extrasynaptic receptors. Although this could simply reflect the proximity of synaptically released glutamate, which is the putative trigger for D-serine release, it could also indicate the existence of microdomains for gliotransmitter release localized at the synapses.

D-Serine Clearance

Like for most neurotransmitters, the signaling action of D-serine should be terminated by its clearance from the extracellular space. This could occur through metabolic breakdown, as it is the case with ATP and its hydrolysis by specific ectonucleotidases located in the extracellular compartment, and/or through cellular uptake as occurring for glutamate. Injection of D-serine into the lateral ventricle results in a preferential accumulation of the amino acid in glial cells, as expected if glial uptake is the main process involved in the clearance of the D-amino acid. Several putative candidates for D-serine transport have been identified on the membrane of glial cells but also on neurons. Glial cells express a Na^{+}-dependent transporter with low affinity for D- and L-serine and with characteristics similar to those of the alanine–serine–cysteine transporter (ASCT) system, which carries D-serine in cultured astrocytes as well as in isolated retina. Another neutral amino acid transporter, the ASC-1, has also been identified in neurons, and its cellular localization on axon terminals and dendrites suggests that it could contribute to the synaptic clearance of D-serine. Finally, a novel Na^{+}/Cl^{-} sensitive transporter has been described in rat brain synaptosomes which, in contrast to the ASCT system, has limited affinity for other neutral amino acids, including cysteine and alanine. It is thus conceivable that multiple transport systems, some of which may still be awaiting identification, contribute concomitantly to the regulation of D-serine concentrations in the extracellular space.

D-Serine Contribution to Synaptic Transmission and Plasticity

NMDARs are peculiar ionotropic receptors in the sense that they require not one but two different

agonists to be activated: glutamate and glycine. NMDARs are tetramers formed by the association of two NR1 and two NR2 subunits. On the NR1 subunits, there is a strychnine-insensitive binding site for glycine. This so-called glycine site can be activated not only by glycine but also by D-serine, with a higher potency according to the subunit composition of NMDARs. The discovery that D-serine occurs naturally in the mammal brain has thus completely changed our view regarding the identity of the endogenous ligand for NMDAR glycine-binding site in the brain. Although glycine was first identified as such, it is obvious today that in several brain areas the endogenous ligand of the NMDAR glycine site is D-serine and not glycine. And because D-serine is released from glial cells, these cells are likely to play an active, if not essential, role in regulating NMDAR-dependent processes, including synaptic transmission, synaptic plasticity, rhythmic activity, activation of second messenger pathways, gene expression, and pathophysiological phenomenon such as excitoxicity and neurodegeneration.

The first indication that D-serine could play a role in NMDAR function came from the observation that the D-amino acid was present in astrocytes ensheathing neurons bearing NMDARs. *In vitro* studies then revealed that D-serine was released from astrocytes on activation of their glutamatergic receptors. These findings strongly suggested that, in some regions of the brain, glutamate released from nerve terminals triggered D-serine release from glial cells, which, in turn, could modulate NMDARs localized on adjacent neurons. The functional role of D-serine was first studied in the hippocampus, where high densities of D-serine and NMDARs occur in the subiculum as well as in the CA1 and CA3 regions. In hippocampal neurons co-cultured with atrocytes, specific enzymatic degradation of D-serine with exogenous DAAO considerably reduces agonist-evoked and spontaneous NMDAR-mediated currents. Because DAAO does not affect glycine levels, this result unambiguously demonstrates that endogenous D-serine is required for NMDAR activity in this brain area. Similar results were obtained in the retina, where Müller glial cells appear to control NMDAR-mediated neurotransmission through the release of D-serine, as revealed by the inhibitory effect of DAAO on the NMDAR-mediated responses evoked by an agonist or light in retinal ganglia cells. The ability of D-serine to control NMDAR-dependent synaptic transmission has been confirmed through the use of a naturally occurring mouse strain lacking DAAO activity. In DAAO-deficient mice, the levels of D-serine are very high in the brain stem and spinal cord. As expected, NMDAR-mediated excitatory postsynaptic currents recorded from dorsal horn neurons in the spinal cord are significantly augmented in these animals. Knockout mice for the transporter ASC-1 provides another experimental model for studying the relevance of D-serine in glutamatergic neurotransmission; these mice display NMDAR-dependent hyperexcitability, presumably resulting from elevated extracellular D-serine associated with a deficient clearance of the amino acid.

Because in many brain areas NMDARs are responsible for the induction of long-term potentiation (LTP) and long-term depression (LTD), the cellular substrates for learning and memory in the mammalian brain, it is then important to know whether D-serine, and thus astrocytes, contributes to such forms of synaptic plasticity in the brain. The contribution of D-serine in LTP was shown in hippocampal cell cultures and brain slices where reducing D-serine levels with DAAO dramatically compromised the ability of high-frequency stimulation to induce LTP in CA1 pyramidal cells. This result confirms that D-serine, rather than glycine, is the endogenous ligand of NMDARs in this brain region, not only during basal transmission but also when synaptic inputs are stimulated intensively. Interestingly, it is commonly believed that senescence is associated with impaired NMDAR-dependent synaptic plasticity and notably LTP. In agreement with this hypothesis, a deficient LTP is observed in the hippocampus of a senescence-accelerated mouse strain. This senescent-related deficiency in synaptic plasticity appears to be due to an impaired D-serine metabolism, as indicated by the reduced production of the D-amino acid measured in the hippocampi of senescent rats. In agreement with this observation, the hippocampal LTP is completely rescued when the D-serine is supplied to the tissue. The key role of D-serine in governing NMDAR activity in the hippocampus is further supported by the demonstration that it is the major, if not the only, endogenous ligand involved in NMDAR-mediated neurotoxicity in this structure. Neuronal death is indeed prevented when slices are treated with serine deaminase, another enzyme that degrades D-serine, whereas supplying the slices with glycine oxidase (GO), an enzyme that degrades specifically glycine, does not affect NMDAR-dependent neurotoxicity.

Taken together, these findings suggest that astrocytic D-serine modulates NMDAR-dependent neurotransmission and synaptic plasticity in different brain regions. Because D-serine is synthesized and released from astrocytes, its action in these different regions depends on the astrocytic environment of neurons. Indeed, it is now accepted that glial coverage of neurons is not static and that it undergoes profound reversible anatomical remodeling in different

areas under different physiological and/or pathological conditions. The hypothalamo-neurohypophysial system (HNS) constitutes certainly the most striking example of such anatomical remodeling that can be observed physiologically because it occurs during lactation, parturition, and chronic dehydration. This system is made of magnocellular neurons, localized in the hypothalamic supraoptic and paraventricular nuclei, whose axons project to the neurohypophysis where their hormonal content, namely oxytocin and vasopressin, can be directly released in the bloodstream. The morphological plasticity of the HNS is characterized by a pronounced reduction in astrocytic coverage of oxytocin-secreting magnocellular neurons, which is entirely reversible on the cessation of the stimulation. Both SR and D-serine are found at high levels in this brain structure, and their expression is strictly restricted to astrocytes. The HNS thus provides a remarkable model for studying the physiological impact of glial-derived D-serine in the context of synaptic transmission and synaptic plasticity.

As reported for the hippocampus and the retina, treating hypothalamic slices with the highly specific enzyme DAAO dramatically reduced NMDAR-mediated synaptic responses recorded in magnocellular neurons, whereas degrading specifically glycine with GO has no effect, thereby indicating that D-serine is the endogenous co-agonist of NMDAR in the HNS. Accordingly, D-serine levels within the synaptic cleft, and thus the level of occupancy of the glycine site of synaptic NMDA receptors, is reduced under conditions in which the astrocytic coverage of neurons is diminished. This results in NMDAR-mediated synaptic responses of smaller amplitude in lactating than in control animals, a difference that disappears when saturating concentrations of D-serine are supplemented. Because the reduction of D-serine concentrations within the synaptic cleft caused by glial withdrawal lowers the number of synaptic NMDARs available for activation, the activity-dependence of phenomena such as LTP and LTD, whose induction depends on NMDAR activation, is modified. The neuron–glia remodeling causes a shift of the activity-dependence of long-term synaptic changes toward higher activity values, similar to what can be observed when NMDARs are partially blocked with pharmacological agents. Therefore, the glial environment of neurons, through its capacity to provide D-serine, has a profound impact not only on NMDAR-mediated synaptic transmission but also on the direction and magnitude of NMDAR-dependent long-term synaptic plasticity. Such astrocyte-mediated metaplasticity is likely to exist at all synapses where endogenous D-serine is involved in regulating NMDARs.

Conclusion

It is now clear that D-serine fulfils all criteria that identify it as the leader of a new class of brain messengers, the D-amino acids. Glial-derived D-serine plays essential roles in the mammalian brain. By modulating NMDARs, this amino acid contributes actively to the transfer and storage of information. Because of this role in governing NMDAR function, glial cells may become prime targets in pathological events that cause neurons to degenerate. It is well documented that over- or downstimulation of NMDARs are implicated in a large number of acute and chronic degenerative disorders, including stroke, epilepsy, peripheral neuropathies, Parkinson's disease, Alzheimer's disease, Huntington's disease, and in psychiatric disorders such as schizophrenia. Whereas previous therapeutic approaches centered directly on NMDARs have been associated with deleterious side effects, the discovery of D-serine as the predominant endogenous ligand of the NMDARs thus provides new strategies for the development of drugs that target NMDAR function indirectly through the modulation of D-serine metabolism, for example, by inhibiting SR or DAAO.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Excitotoxicity in Neurodegenerative Disease; Glial Plasticity and Neuroendocrine Regulation; Glial Influence on Synaptic Transmission; Glutamate; Glycine Receptors: Molecular and Cell Biology; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; NMDA Receptors and Development; NMDA Receptors, Cell Biology and Trafficking.

Further Reading

Araque A, Parpura V, Sanzgiri RP, and Haydon PG (1999) Tripartite synapses: Glia, the unacknowledged partner. *Trends in Neuroscience* 22(5): 208–215.

Boehning D and Snyder SH (2003) Novel neural modulators. *Annual Review of Neuroscience* 26: 105–131.

Fields RD and Stevens-Graham B (2002) New insights into neuron-glia communication. *Science* 298(5593): 556–562.

Fujii N (2002) D-Amino acids in living higher organisms. *Origins of Life and Evolution of Biospheres* 32(2): 103–127.

Haydon PG and Carmignoto G (2006) Astrocyte control of synaptic transmission and neurovascular coupling. *Physiological Review* 86(3): 1009–1031.

Hirrlinger J, Hulsmann S, and Kirchhoff F (2004) Astroglial processes show spontaneous motility at active synaptic terminals *in situ*. *European Journal of Neuroscience* 20(8): 2235–2239.

Johnson JW and Ascher P (1987) Glycine potentiates the NMDA response in cultured mouse brain neurons. *Nature* 325(6104): 529–531.

Laming PR, Kimelberg H, Robinson S, et al. (2000) Neuronal-glial interactions and behaviour. *Neuroscience and Biobehavioral Reviews* 24(3): 295–340.

Mothet JP, Parent AT, Wolosker H, et al. (2000) D-Serine is an endogenous ligand for the glycine site of the *N*-methyl-D-aspartate receptor. *Proceedings of the National Academy of Sciences of the United States of America* 97(9): 4926–4931.

Mothet JP, Pollegioni L, Ouanounou G, Martineau M, Fossier P, and Baux G (2005) Glutamate receptor activation triggers a calcium-dependent and SNARE protein-dependent release of the gliotransmitter D-serine. *Proceedings of the National Academy of Sciences of the United States of America* 102(15): 5606–5611.

Panatier A, Theodosis DT, Mothet JP, et al. (2006) Glia-derived D-serine controls NMDA receptor activity and synaptic memory. *Cell* 125(4): 775–784.

Shleper M, Kartvelishvily E, and Wolosker H (2005) D-Serine is the dominant endogenous coagonist for NMDA receptor neurotoxicity in organotypic hippocampal slices. *Journal of Neuroscience* 25(41): 9413–9417.

Theodosis DT (2002) Oxytocin-secreting neurons: A physiological model of morphological neuronal and glial plasticity in the adult hypothalamus. *Frontiers in Neuroendocrinology* 23(1): 101–135.

Volterra A and Meldolesi J (2005) Astrocytes, from brain glue to communication elements: The revolution continues. *Nature Reviews Neuroscience* 6(8): 626–640.

Dynamin

V Anggono and P J Robinson, Children's Medical Research Institute, Sydney, NSW, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons communicate with one another via the release of neurotransmitters from synaptic vesicles (SVs) by exocytosis from presynaptic nerve terminals. Exocytosis is a fast membrane-fusion event. A typical nerve terminal in the brain contains approximately 200 SVs, yet individual exocytic events can occur with high frequency in the range of 5–50 Hz and for prolonged lengths of time (greater than 30 s). Therefore, SVs must be rapidly recycled to prevent the depletion of their finite number; otherwise, synaptic fatigue occurs. After exocytosis, synaptic vesicle endocytosis (SVE) retrieves them. Away from the synapse, a similar endocytic pathway, known as receptor-mediated endocytosis (RME) functions in many cellular processes, including receptor downregulation and intracellular signaling. Both SVE and RME involve plasma membrane budding (or invagination) and pinching off new SVs by membrane fission at the vesicle neck. As well as maintaining the SV supply, the process is responsible for the initial formation (or biogenesis) of mature SVs. The protein dynamin plays a crucial role at an early point in vesicle fission and at the final step.

Dynamin was originally identified as a nerve terminal phosphoprotein and later as a GTPase that can associate with microtubules *in vitro*. Sequencing of the *dynamin* gene revealed that it contained a GTP-binding site homologous to that of other GTPases (or G-proteins). Later, a superfamily emerged from large dynamin-like GTPases comprising proteins present across all eukaryotic species. The biological importance of dynamin was first discovered with the identification of temperature-sensitive mutants in *Drosophila* called *shibire* that give rise to a pleiotropic defect in endocytosis, leading to rapid paralysis at a nonpermissive temperature of 29 °C. When the *shibire* gene was cloned in 1991, it was found to be 70% identical to rat brain dynamin 1. To better understand endocytosis, research since then has focused on the molecular mechanisms underlying dynamin's ability to bind, manipulate, and fragment lipid.

Dynamin Superfamily

Dynamin is a 100 kDa GTPase enzyme belonging to a family of large GTP-binding proteins (Figure 1). Dynamins are also known as the large GTPases to distinguish them from four other families of G-proteins: the small Ras superfamily of small monomeric GTPases (20–25 kDa), the α-subunits of heterotrimeric G-proteins, the septins, and other regulatory GTPases. All these G-proteins have a conserved tripartite GTP-binding motif within their N-terminal domains. Based on their structural similarity, dynamin superfamily proteins can be grouped into three broad categories: the classical dynamins (dynamin 1, 2, and 3), the dynamin-related proteins (dynamin-like protein 1 (Dlp1), optic atrophy 1 (OPA1), mitofusins, Mx proteins), and the guanylate-binding proteins (GBPs)/atlastins. All dynamin superfamily members appear to share two extraordinarily unique properties of self-assembly, to which their mechanism of GTP hydrolysis is uniquely linked, and the ability to bind and tubulate lipids. This adapts them to membrane fission and fusion functionality. Members of the dynamin superfamily have diverse roles in multiple forms of vesicle fission from the plasma membrane (SVs, clathrin-coated vesicles, caveolae, and phagocytosis), as well as in membrane budding events within intracellular organelles (such as the mitochondria, Golgi, endoplasmic reticulum (ER), and peroxisomes).

Dynamin Domains

Dynamin superfamily members share three conserved domains in common: an N-terminal GTPase domain, middle domain (MD), and GTPase effector domain (GED) (Figure 1). The three domains distinguish them from other G-protein families and define the unusual enzymatic and self-assembly properties of these large GTPases. Most superfamily members contain additional domains tailored to their specific cellular function or subcellular location. For example, three groups have transmembrane domains and are integral membrane proteins: OPA1, mitofusins, and atlastins. Dynamin contains two additional functional domains: a pleckstrin-homology (PH) domain and a proline-rich domain (PRD) at the C-terminus.

GTPase domain The GTPase domain is the most highly conserved region. Compared with the Ras superfamily, dynamin has an unusually high GTPase activity and low affinity for GTP. Thus, it has a low basal GTPase activity which is controlled by self-assembly or lipid binding. Mutations in the GTPase domain such as K44A, K44E, and S45N are defective in GTP binding. These mutations are dominant-negative, and DNA constructs harboring them have

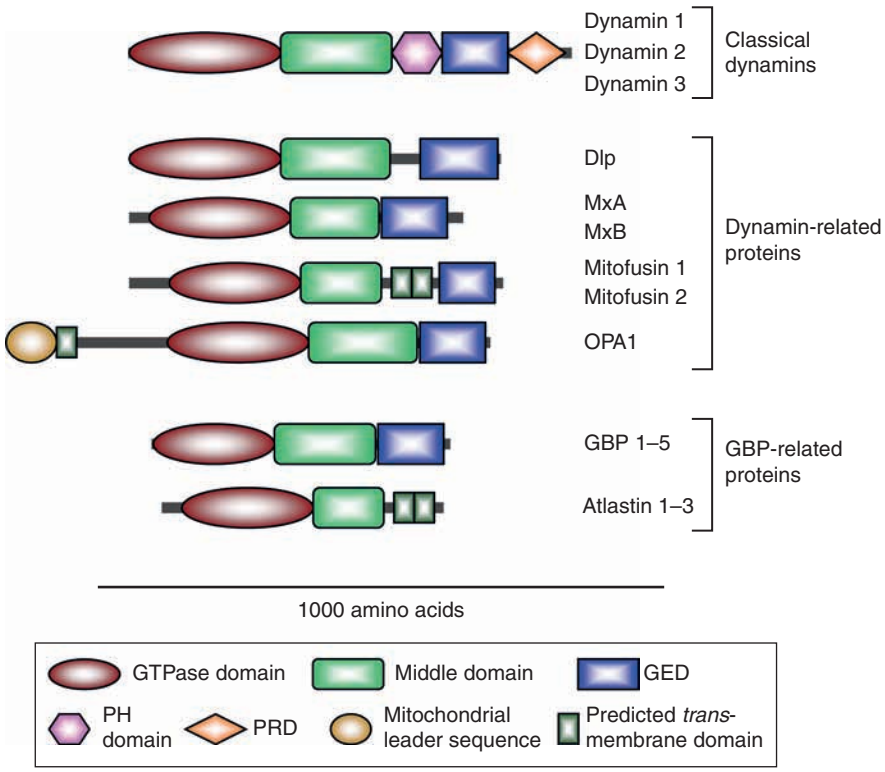

Figure 1 Domain structure of the human dynamin superfamily. All dynamin superfamily members contain a GTPase domain that binds and hydrolyses GTP, a middle domain (MD) and a GTPase effector domain (GED) that are involved in higher-order self-assembly and stimulation of GTPase activity. Classical dynamins also have a pleckstrin-homology (PH) domain that interacts with lipid membranes and a C-terminal proline-rich domain (PRD) that interacts with Src-homology 3 (SH3) domain-containing proteins. Other family members may have divergent sequences at their N-termini or between the MD and GED. Human dynamin superfamily members are grouped according to their domain structures, which are represented by color-coded symbols. GBP, guanylate-binding protein; GTP, guanosine triphosphate.

been widely used to inhibit endocytosis in cells by transfection. However, GTP binding alone is not sufficient to support endocytosis, and GTP hydrolysis is also required to complete fission. In particular, dynamin-dependent GTP hydrolysis is required for SVE at fast synapses in the nervous system.

Middle domain and GTPase effector domain Dynamin MD is involved in the formation of dynamin tetramers and then for the higher-order self-assembly of tetramers into rings and helices. It lacks sequence homology to any known protein structural motif. A synthetic peptide derived from the N-terminal half of the MD of dynamin 1 adopts an α-helical coiled-coil structure and forms stable tetramers in solution. It also makes extensive contacts with the self-assembling GED. The GED is also known as the assembly or coiled-coil domain. It binds to the GTPase domain, to the MD, and to itself to form homodimers. These interactions have been implicated in self-assembly of dynamin which leads to an increase of GTPase catalytic activity by up to 100-fold. Cryo-electron microscopy (EM) studies have suggested that the GED and the MD may form a stalk which mediates intermolecular interactions in self-assembled dynamin.

Pleckstrin-homology domain Among the dynamin family members, only the classic dynamins contain the PH domain. It mediates binding to lipid membranes via interaction with a family of phosphoinositide phospholipids. The PH domain of dynamin 1 has very low lipid specificity, but favors binding to phosphatidylinositol (4,5) bisphosphate ($PI(4,5)P_2$). Their interaction produces the highest stimulation of dynamin. A burst of $PI(4,5)P_2$ is generated at the plasma membrane early in endocytosis, which recruits dynamin to these sites of active endocytosis.

Proline-rich domain The PRD is present only in the classic dynamins and is a protein–protein interaction domain for a variety of signaling or cytoskeletal proteins. Dynamin 1 contains 13 PxxP motifs (where P is proline and x is any amino acid) which bind numerous SH3 domain-containing proteins that play important roles in endocytosis (Table 1). The PRD does not

Table 1 Proteins of the endocytic machinery that bind the dynamin proline-rich domain

Protein	*Membrane deformation*	*Actin binding*
Syndapin 1–3	F-BAR	N-WASP
Endophilin 1–3	N-BAR	
Amphiphysin 1 and 2	N-BAR	
FBP17	F-BAR	N-WASP
Tuba	BAR	N-WASP
Intersectin		N-WASP
Cortactin		F-actin and N-WASP
Actin-binding protein 1		F-actin
Profilin 1 and 2		G-actin
SPIN-90		
SNX-9, 18	BAR	

BAR, bin-amphiphysin-rvs; F-BAR, FER-CIP4 homology-BAR; FBP17, formin-binding protein 17; N-BAR, N-terminal BAR; N-WASP, neural Wiskott–Aldrich syndrome protein; SNX, sorting nexin; SPIN-90, SH3 protein interacting with Nck, 90 kDa.

appear to be important for dynamin assembly into rings because its deletion (Dyn1ΔPRD) has no effect. A major function of the PRD is to target dynamin to its site of action in the cell. Overexpression of Dyn1ΔPRD results in the failure of dynamin to accumulate at clathrin-coated pits. The PRD also allows dynamin to be recruited by key partners such as syndapin that may be involved in membrane tubulation. The dynamin 1 PRD is also the site for its phosphorylation and regulation of its activity in nerve terminals.

Dynamin Mediates Endocytosis

Dynamin in Flies

Drosophila has only a single *dynamin* gene, *shibire*, for multiple cellular functions of SVE, RME, and cytokinesis. The *shibire* temperature-sensitive mutant fly exhibits a rapid and reversible paralysis at a nonpermissive temperature, which was later identified as a block in endocytosis. Within a few minutes after exposure to the nonpermissive temperature, the nerve terminals of this fly are remarkably empty of SVs and show an accumulation of endocytic pits at the plasma membrane (**Figures 2(a)** and **2(b)**). At the base of these pits was one or two electron-dense rings; hence, they were called collared pits (**Figure 2(c)**). Tubular invaginations without collars accumulate in a number of other cell types in these flies, indicating a general defect in endocytosis (**Figure 2(d)**). Endocytosis is inhibited in all mammalian cells overexpressing a dynamin mutant (K44A or K44E) deficient in GTP binding (**Figure 2(e)**). As in *shibire*, large tubular invaginations are observed which are connected to the plasma membrane. Thus, mammalian dynamin has a similar role.

Mammalian Classic Dynamins

Unlike flies, mammals have three *dynamin* genes, *dynamin 1, 2,* and *3*. They are each approximately 100 kDa in size, and they appear to work together for endocytosis in the brain. Dynamin 1 is the most studied isoform and is essentially specific to the central nervous system (CNS), being highly enriched in presynaptic nerve terminals. Dynamin 2 is ubiquitously expressed in all tissues, including neurons. Dynamin 3 is predominantly expressed in the testis and in some regions of the brain, where it is found both presynaptically and postsynaptically. Within the brain, expression levels of dynamin 1 are approximately 100-fold greater than those of the others and at least five times higher than synapsin 1. In fact, dynamin 1 is among the most abundant neuronal proteins in the brain.

All three dynamins play a role in vesicle scission during the specialized neuronal function of rapid synaptic vesicle recycling. Dynamin 1 mediates the bulk of SVE, particularly all the endocytosis that occurs during a depolarization stimulus. A knockout of dynamin 1 in mice produces animals that are normal at birth but then die within the first week. Lack of dynamin 1 does not completely inhibit SVE, and at birth synapses from these mice are filled with SVs. Therefore, dynamin 1 is largely dispensable for the biogenesis and recycling of SVs during development. It is, however, essential for recycling during high or prolonged synaptic activity. SVE is severely impaired during nerve depolarization but resumes efficiently on stimulus termination. Overexpression of dynamin 3, but not dynamin 2, rescues the phenotype, suggesting that dynamin 3 has a homologous role to that of dynamin 1 but is much lower in abundance. Dynamin 3 is also found in dendritic spine tips, where it regulates their morphogenesis in developing neurons. Dynamin 2 is multifunctional. It plays a major role in SVE, but this appears to be mainly in the slow endocytosis that occurs after nerve stimulation has ceased. In neurons and nonneuronal cells, it mediates the conventional clathrin-mediated uptake of surface receptors or membrane. In addition, dynamin 2 mediates endocytosis of caveolae and is required for phagocytosis. It also plays a role in endocytosis-independent cellular functions, such as actin dynamics, cytokinesis, centrosome cohesion, apoptosis, and mitogen-activated protein kinase (MAPK) signaling.

Function of Other Superfamily Members

Dynamin-related proteins Dynamin-related proteins are large in size (70–100 kDa) and have a relatively low affinity for GTP and a high GTPase activity. However,

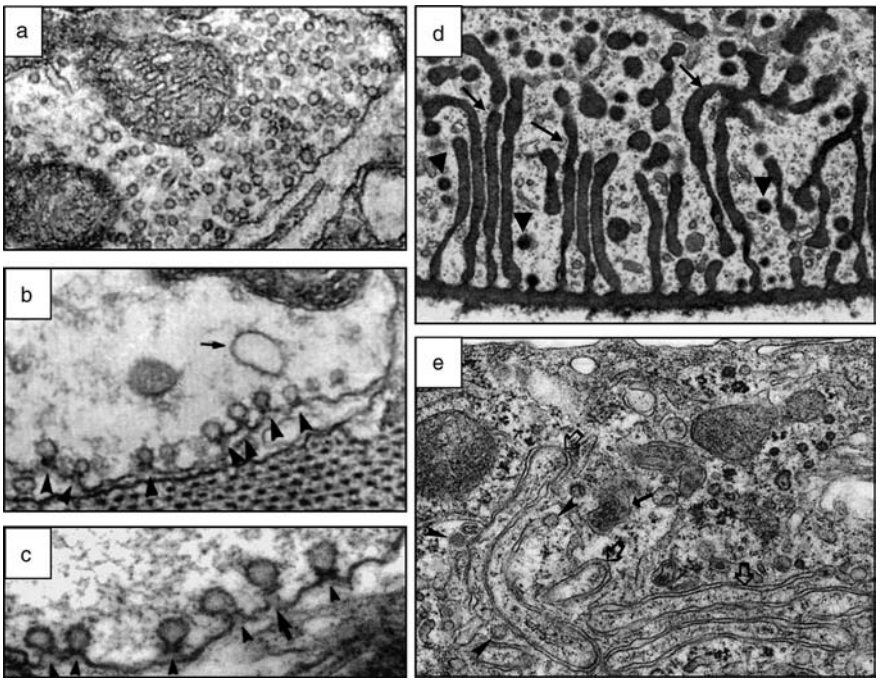

Figure 2 Endocytosis is structurally altered in the *Drosophila* mutant *shibire* and dynamin-K44A mutant: (a) electron micrograph of a typical coxal synapse of a *shibire* fly at the permissive temperature of 19 °C, which is full of synaptic vesicles (SVs) and two mitochondria; (b) a similar synapse at the nonpermissive temperature of 29 °C is characterized by the loss of SVs and the appearance of small invaginations of the plasma membrane, which possess an electron-dense ring or collar around their necks (arrowheads); (c) a higher magnification of a synapse containing collared pits at 29 °C, showing single (arrowheads) or double rings (arrow); (d) in contrast to neuronal cells, long labyrinthine tubules that lack collars are formed at the cortical region near the surface of a *shibire* garland cell at the nonpermissive temperature; cells were stained with tannic acid to show the structures of the tubules (arrows) and coated pits (arrowheads) that are continuous with the extracellular spaces; (e) an accumulation of long tubular structures (open arrows) are also apparent in the HeLa cell overexpressing dynamin 1-K44A mutant, reminiscent of the *shibire* phenotype; many of these tubular structures have coated ends (arrowheads). (a–c) Reproduced from Koenig JH and Ikeda K (1989) Disappearance and reformation of synaptic vesicle membrane upon transmitter release observed under reversible blockage of membrane retrieval. *Journal of Neuroscience* 9: 3844–3860. (d) Reproduced from Kosaka T and Ikeda K (1983) Reversible blockage of membrane retrieval and endocytosis in the garland cell of the temperature-sensitive mutant of *Drosophila melanogaster*, *shibire*^{ts}. *Journal of Cell Biology* 97: 499–507. (e) Reproduced from Damke H, Baba T, Warnock DE, et al. (1994) Induction of mutant dynamin specifically blocks endocytic coated vesicle formation. *Journal of Cell Biology* 127: 915–934.

they lack both the PH domain and the PRD. They share a common ability to self-assemble into oligomers and show elevated assembly-dependent GTPase activity. Most of the dynamin-related proteins are involved in the morphogenesis and maintenance of intracellular organelles such as the mitochondria (Dlp1, OPA1, and mitofusin) and ER (Mx, Dlp1, and atlastin) (Table 2). All members tested to date (Dlp1 and MxA) are able to self-assemble into rings and to tubulate lipids. Like classic dynamin, Dlp1 and Mx proteins appear to be involved in organelle fission events. For example, Dlp1 is responsible for the mitochondrial fission required for mitochondrial division. In contrast, OPA1 and mitofusin are integral membrane proteins responsible for the fusion of mitochondrial inner and outer membranes, respectively. The biological role of Mx proteins is less clear. Their expression is induced by interferon, and they provide cells with strong protection against viral infection.

Table 2 Functions of dynamin-related proteins

Protein	*Localization*	*Function*
Dlp1 (Drp1)	Cytoplasmic, mitochondrial outer membrane, peroxisome and endoplasmic reticulum	Organelle fission
OPA1	Mitochondria inner membrane	Mitochondrial fusion
Mitofusin 1 and 2	Mitochondrial outer membrane	Mitochondrial fusion
MxA and MxB	Smooth endoplasmic reticulum	Viral resistance

Dlp1, dynamin-like protein 1; OPA1, optic atrophy 1.

Guanylate-binding proteins/atlastins The crystal structure of human GBP1 (hGBP1) places the GBPs within the dynamin superfamily, despite low amino-acid sequence homology. hGBP1 has a low GTP-binding affinity and relatively fast GTP hydrolysis.

It is unique among the dynamin superfamily members in its ability to convert GTP to guanosine monophosphate (GMP), cleaving off phosphate from GTP in two successive steps. GBPs contain only the GTPase domain, MD and GED. Within cells, hGBP1 in its GTP-bound form appears to target the Golgi. Like Mx, GBPs are induced in response to interferons and play roles in protective immunity against microbial and viral pathogens.

Atlastins are the most recently described superfamily members. Like mitofusin and OPA1, atlastins are integral membrane proteins, with the same predicted membrane topology (N-terminal GTPase and short C-terminal sequences in the cytoplasm). This suggests a role in membrane fusion rather than fission. Atlastin 1 appears to be involved in vesicle budding and trafficking in the early secretory pathway between the ER and the Golgi. It is found at highest levels in the brain and is highly enriched in neurites and growth cones, where it plays a role in neurite outgrowth and axon formation.

Mechanisms of Dynamin-Mediated Membrane Fission

Role of Unassembled Dynamin Early in Endocytosis

Dynamin appears to act twice in the mechanism of endocytosis: early in the constriction of the invaginating vesicle and late in its scission. The first step involves a regulatory GTPase-like function of unassembled dynamin, and the second involves self-assembly and assembly-stimulated GTP hydrolysis. The two models are not mutually exclusive. It has been proposed that unassembled dynamin plays a role in the early, rate-limiting steps of endocytosis, prior to self-assembly. It has been suggested that dynamin in the GTP-bound state or in the course of steady-state GTP turnover serves as a kinetic timer that regulates and controls coated pit assembly and maturation. In this model, basal GTP hydrolysis of unassembled dynamin is controlled by an interaction of GED with the switch 2 region of the GTPase domain. One suggestion is that it acts as a molecular switch that recruits downstream effectors in a GTP-dependent fashion. For example, GTP-bound dynamin interacts with Hsc70 and auxilin, proteins that uncoat clathrin from vesicles, but no major partner has yet been found to interact with dynamin specifically in its GTP-bound state that accounts for its early role in endocytosis.

Ring Assembly on Lipid Templates

The second step of dynamin action in late stages of endocytosis involves assembly-stimulated GTP hydrolysis driving a twisting helix mechanism of action. Dynamin 1 exists as a tetramer that can self-assemble *in vitro* into two types of higher-order structures. In the absence of any factor, and on removal of salt, dynamin self-assembles into rings or stacks of rings (Figures 3(a) and 3(b)). In the presence of liposomes, microtubules, or other suitable template, dynamin forms spiral helices (Figure 3(c)). Such self-assembly results in an up to 1000-fold increase in GTPase activity. These higher-order structures resemble the collared pits seen at the base of endocytic pits in *shibire* (Figure 3(d)) and similar shaped helical spirals are seen in lysed nerve terminals treated with GTPγS, a nonhydrolyzable GTP analog (Figure 3(e)). Dynamin also has a membrane-deforming ability. When added to liposomes or lipid monolayers, purified dynamin 1 drives their transformation into tubules (Figure 3(f)). The addition of GTP to dynamin-coated lipid tubes *in vitro* causes their fragmentation into small vesicles (Figure 3(g)). The protein most likely mediates vesicle fission using the energy obtained from GTP hydrolysis.

Helical Tension and Twist Produce Fission

Two models were initially proposed to explain the mechanism of dynamin-mediated vesicle fission: the constriction model and the expansion model. A more recent model incorporates both concepts. In the constriction model, dynamin forms a collar around vesicle necks and constricts the lipid to induce fission (also called a pinchase activity; Figures 4(a) and 4(b)). This is supported by three-dimensional cryo-EM structures of dynamin in the constricted and nonconstricted states. Recombinant Dyn1ΔPRD forms a T-shaped dimer consisting of three prominent densities corresponding to the leg (GTPase domain), stalk (GED and MD), and head (PH domain). The addition of GTP causes the head and stalk to rearrange, potentially generating a force on the underlying lipid bilayer and resulting in rings of smaller diameter.

In the helical expansion model, GTP hydrolysis provides the energy for rapid expansion of the dynamin helix at the neck of the vesicle, which propels it from the presynaptic membrane into the cytosol (also called a poppase activity). This model is based on a dramatic increase in helical pitch of dynamin spirals formed on rigid lipid nanotubes following GTP hydrolysis (Figures 4(c)–4(e)).

A single twisting model consolidates both the previous models and is based on observations that dynamin-coated lipid tubules twist and supercoil on addition of GTP (Figures 4(f) and 4(g)). The twisting activity of assembled and activated dynamin produces a longitudinal tension that is released by tubule breakage. Hence, late-stage endocytosis may involve cooperation between mechanisms that generate membrane tension and those that mediate constriction (Figure 4(g)).

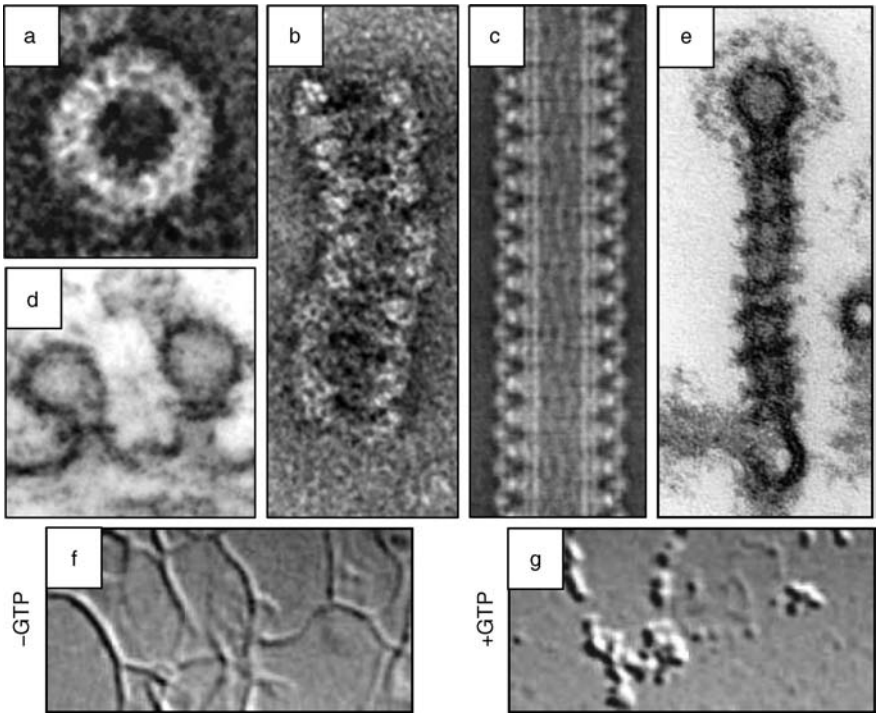

Figure 3 Dynamin forms rings and helices *in vitro* and *in vivo*: (a) purified dynamin self-assembles into a ring in the presence of GDP and γ-phosphate analogs; (b) purified dynamin self-assembles into and spirals in the presence of GDP and γ-phosphate analogs; (c) dynamin forms helices around lipid tubules when added to phosphatidylserine liposomes in the absence of any nucleotide; (d) the rings have an outside diameter of 50 nm and an inner diameter of 30 nm and resemble the collared structures seen in *shibire* mutant nerve terminals; (e) long invaginating tubules decorated with a helical pattern and a terminal clathrin-coated bud are observed in isolated nerve terminals (synaptosomes) treated with GTPγS; (f) dynamin also has the ability to deform liposomes into narrow tubules; (g) the addition of GTP to dynamin lipid tubules result in the fragmentation of the tubes. GDP, guanosine diphosphate; GTP, guanosine triphosphate; GTPγS, nonhydrolyzable GTP analog. (a) Reproduced from Hinshaw JE and Schmid SL (1995) Dynamin self-assembles into rings suggesting a mechanism for coated vesicle budding. *Nature* 374: 190–192. (b) Reproduced from Carr J and Hinshaw JE (1997) Dynamin assembles into spirals under physiological salt conditions upon the addition of GDP and γ-phosphate analogues. *Journal of Biological Chemistry* 272: 28030–28035. (c) Reproduced from Chen YJ, Zhang P, Egelman EH, and Hinshaw JE (2004) The stalk region of dynamin drives the constriction of dynamin tubes. *Nature Structural and Molecular Biology* 11: 574–575. (d) Reproduced from Koenig JH and Ikeda K (1989) Disappearance and reformation of synaptic vesicle membrane upon transmitter release observed under reversible blockage of membrane retrieval. *Journal of Neuroscience* 9: 3844–3860. (e) Reproduced from Takei K, McPherson PS, Schmid SL, et al. (1995) Tubular membrane invaginations coated by dynamin rings are induced by GTPX in nerve terminals. *Nature* 374: 186–190. (f, g) Reproduced from Roux A, Uyhazi K, Frost A, and De Camilli P (2006) GTP-dependent twisting of dynamin implicates constriction and tension in membrane fission. *Nature* 441: 528–531.

Dynamin Recruitment to Sites of Endocytosis

The lipid-deforming property of dynamin is also shared with bin-amphiphysin-rvs (BAR) domain-containing proteins. BAR domain proteins have a binding preference for lipid curvature and are therefore proposed as shaping the neck of invaginating presynaptic plasma membrane during SVE. Syndapin 1, amphiphysin 1, and endophilin 1 are three such proteins that are abundant in the brain. These three also contain SH3 domains that bind dynamin 1 through its PRD, and these interactions are important for SVE. Similarly, formin-binding protein 17 (FBP17) is a BAR domain protein that binds dynamin 2 in the brain and testis. It is likely that these proteins function along with PI(4,5)P_2 in the recruitment of dynamin to endocytic sites. Some BAR-domain dynamin-binding proteins, and several others, also bind actin or actin assembly regulatory proteins (Table 1). It is thought that these interactions link the endocytosis machinery to the actin cytoskeleton.

Regulation of Dynamin 1 Function

The role of dynamin 1 in SVE is controlled by its phosphorylation. Dynamin 1 undergoes cycles of phosphorylation and dephosphorylation on nerve-terminal depolarization (Figure 5(a)). Stimulus-dependent dephosphorylation in synaptosomes (isolated nerve terminals) is very rapid (<1 s) and is mediated by the same Ca^{2+} influx through voltage-gated Ca^{2+} channels that initiates exocytosis and transmitter release. Dephosphorylation is mediated

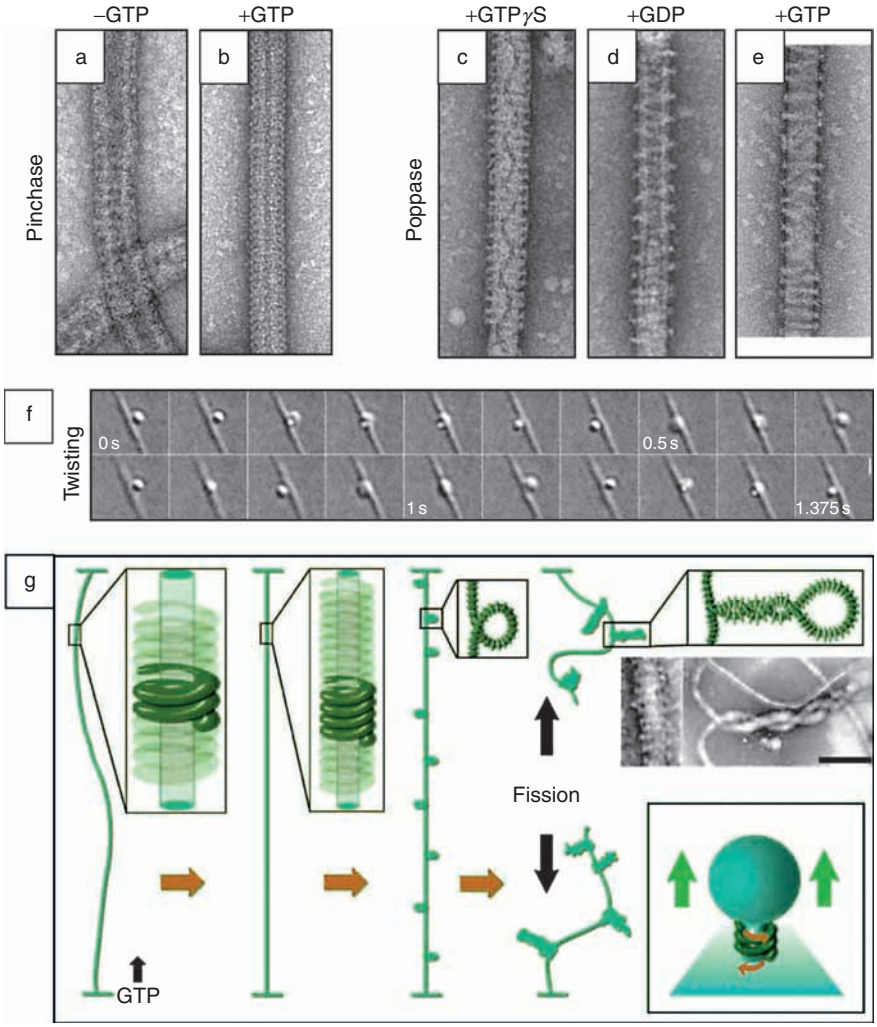

Figure 4 Molecular mechanisms of vesicle fission by dynamin: (a) dynamin induces the tubulation of liposomes that are decorated with helical oligomers with a diameter of ~50 nm; (b) the addition of a nonhydrolysable GTP analog to the dynamin–lipid tubes caused the tubules to constrict to ~40 nm (these data gave rise to the constrictase or pinchase hypothesis); (c) dynamin binding to lipid nanotubes assembles to an oligomer in the form a helix with a pitch (interring spacing) of ~11 nm; (d, e) on GTP hydrolysis, the helix becomes much more uncoiled, with a dramatic increase in pitch to ~20 nm (these data gave rise to the poppase hypothesis); (f) biotinylated dynamin induces the tubulation of liposomes in the presence of a streptavidin-coated polystyrene bead; on addition of GTP, the bead rotates around the longitudinal axis of the tubule, suggesting the GTP-dependent twisting activity of dynamin; (g) current hypothesis of the mechanical action of dynamin in membrane fission. In (g), dynamin tubulates liposomes and forms a helical structure around the lipid tube. On the addition of GTP, the dynamin helix undergoes an increase in torsion, leading to straightening and twisting of the tubule to form supercoil structures (insets). The twisting activity contributes to the longitudinal contraction to establish membrane tension around the vesicle neck. In combination with the constricting activity of dynamin, vesicle fission is then achieved. GDP, guanosine diphosphate; GTP, guanosine triphosphate; GTPγS, nonhydrolyzable GTP analog. (a, b) Adapted from Zhang P and Hinshaw JE (2001) Three-dimensional reconstruction of dynamin in the constricted state. *Nature Cell Biology* 3: 922–926. (c, e) Adapted from Stowell MHB, Marks B, Wigge P, et al. (1999) Nucleotide-dependent conformational changes in dynamin: Evidence for a mechanochemical molecular spring. *Nature Cell Biology* 1: 27–32. (f, g) Adapted from Roux A, Uyhazi K, Frost A, and De Camilli P (2006) GTP-dependent twisting of dynamin implicates constriction and tension in membrane fission. *Nature* 441: 528–531.

by the Ca^{2+}-dependent protein phosphatase, calcineurin, which can be blocked by calcineurin antagonists such as cyclosporin A (Figure 5(b)). After the stimulus is removed, dynamin 1 is slowly (~40 s) rephosphorylated by its endogenous protein kinase, cyclin-dependent kinase 5 (cdk5). Rephosphorylation can be blocked by a specific cdk5 antagonist, roscovitine. Synaptosomal dynamin 1 has seven phosphorylation sites, five of which reside in the PRD and two of which are in the MD (Figure 5(c)). Ser-774 and Ser-778 are the major cdk5 sites, representing 70% of the total dynamin 1 phosphorylation *in vivo*. Dynamin 3, but not dynamin 2, is also phosphorylated in nerve terminals at homologous sites. This supports the

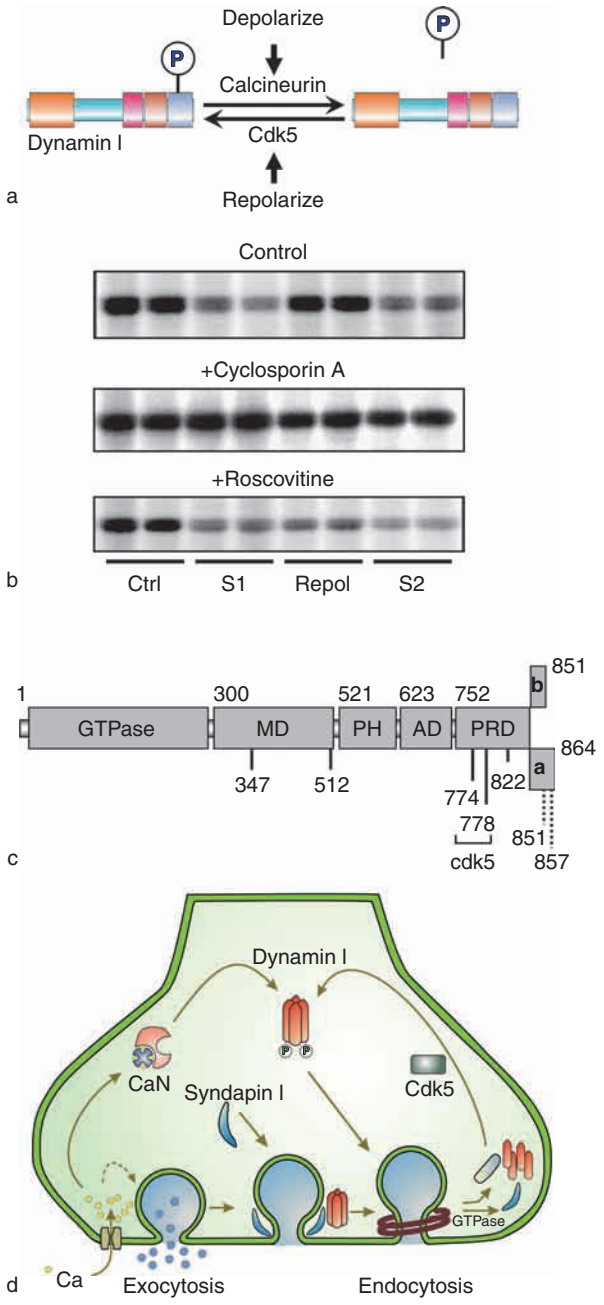

Figure 5 Dynamin 1 phosphorylation in nerve terminals: (a) model for the regulation of dynamin 1 dephosphorylation by calcineurin on nerve terminal depolarization and rephosphorylation by cdk5 on repolarization; (b) phosphorylation, dephosphorylation, and rephosphorylation of dynamin 1; (c) domain structures of rat dynamin 1 (*a* and *b* are long and short splice variants) showing the location of all *in vivo* phosphorylation sites (the dominant sites are shown in bold); (d) a proposed model outlining how the dynamin 1 dephosphorylation and rephosphorylation cycle might regulate synaptic vesicle endocytosis. In (b), dynamin 1 isolated from ^{32}P-labeled synaptosomes is constitutively phosphorylated in resting nerve terminals (Ctrl), is dephosphorylated after stimulation for 5 s with 41 mM potassium chloride (S1), is rephosphorylated after potassium chloride removal (Repol), and is dephosphorylated on second depolarizing stimulus (S2, top panel). Dynamin 1 dephosphorylation is inhibited by specific calcineurin blocker, cyclosporin A (middle panel), whereas dynamin 1 rephosphorylation is inhibited by specific cdk5 antagonist, roscovitine (lower panel). As shown in (d), in resting nerve terminals dynamin 1 is constitutively phosphorylated in cytosol. During nerve stimulation, an influx of Ca^{2+} stimulates vesicle fusion (exocytosis) as well as calcineurin-dependent dephosphorylation of dynamin 1. This allows dynamin 1 to form a complex with the BAR domain-containing protein syndapin 1. The formation of dynamin 1–syndapin 1 complex may help localize dynamin 1 to the site of endocytosis, and it is essential for membrane retrieval. Other BAR domain proteins are also involved. After endocytosis, cdk5 rephosphorylates dynamin 1 and promotes the disassembly of the endocytic complex with syndapin 1. AD, assembly domain; BAR, bin-amphiphysin-rvs; cdk5, cyclin-dependent kinase 5; CaN, calcineurin; GTP, guanosine triphosphate; MD, middle domain; PH, pleckstrin-homology; PRD, proline-rich domain. (b; top and bottom) reproduced from Tan TC, Valova VA, Malladi CS, et al. (2003) Cdk5 is essential for synaptic vesicle endocytosis. *Nature Cell Biology*. 5: 701–710.

hypothesis that dynamin 1 and 3 share common control mechanisms and may have redundant functions in SVE.

Dynamin 1 phosphorylation and dephosphorylation have a number of effects in their functions that may explain their role in SVE; phosphorylation increases its GTPase activity *in vitro*, inhibits binding of the BAR domain containing protein syndapin 1, and localizes dynamin 1 to the cytosol of nerve terminals. Mutation of the main two dynamin 1 phosphorylation sites (Ser-774 and Ser-778) inhibits SV recycling in cultured neurons. A model has been proposed to explain the role of the phosphorylation cycle in the SVE cycle at the synapse (**Figure 5(d)**). On neuronal depolarization, Ca^{2+} influx stimulates SV fusion with the plasma membrane (exocytosis), followed by SVE to retrieve the empty SV. During the onset of SVE, Ca^{2+} activates calcineurin, which in turn dephosphorylates dynamin 1, along with seven other endocytic phosphoproteins (amphiphysin 1/2, synaptojanin 1, PIPKIγ, epsin 1, eps15, and AP180), collectively termed the dephosphins. Dephosphorylated dynamin 1 relocates to the plasma membrane, presumably through interaction with PI(4,5)P_2, which is enriched at sites of endocytosis, or through an interaction with syndapin 1, which is proposed to shape the neck of the budding vesicle. Dynamin 1 then assembles into helices around the neck of the budding vesicle. Via its twisting action and its assembly-dependent and lipid-stimulated GTPase activity, dynamin 1 releases the vesicle from the plasma membrane. At the completion of SVE, cdk5 rephosphorylates dynamin 1, which is dissociated from the plasma membrane and the syndapin 1 complex back into the cytosol. This process may reset dynamin 1 availability for the next cycle of endocytosis.

Dynamin and Synaptic Transmission

SV recycling in nerve terminals is a fundamental part of maintaining sustained synaptic transmission. It becomes rate limiting for neurotransmitter release under conditions of only moderate or prolonged stimulation. The role of dynamin in sustaining synaptic transmission was first evident from studies of *shibire* flies, which are paralyzed within 2 min at the nonpermissive temperature. The paralysis is brought about by SV depletion, resulting in synaptic transmission impairment. Targeted expression of the temperature-sensitive *shibire* mutant in different regions of the *Drosophila* CNS has been used to study the neural mechanisms of behavior. For example, its expression in the mushroom body revealed that synaptic transmission is required for retrieval of olfactory memory but not for memory acquisition.

In the mammalian CNS, presynaptic loading of GTPγS or a dynamin 1 PRD peptide abolishes endocytosis at the calyx of Held giant nerve terminal. These treatments have no immediate effect on exocytosis, but cause a use-dependent rundown of exocytosis at this fast CNS synapse. Similarly, when rat brain synaptosomes are incubated with a dynamin 1 phospho-mimetic peptide that specifically blocks binding of syndapin 1 (Figure 6(a)), endocytosis is inhibited. The peptide has no effect on exocytosis during the first 60 s of stimulation, but causes an activity-dependent rundown of glutamate release when the stimulation is maintained (Figure 6(b)). These observations establish a requirement of dynamin 1 and SVE in the overall control of synaptic transmission.

Dynamin Inhibitors

An activity-dependent rundown of exocytosis due to restricting dynamin 1 function might have clinical applications in treating neurological disorders such as epilepsy. Epilepsy is characterized by recurrent unprovoked seizures caused by excessive neuronal activity in the brain. Current pharmacological treatments target the ion channels and glutamate or γ-aminobutyric acid (GABA) receptors that control synaptic excitability. Targeting the GTPase activity of dynamin 1 is a more attractive candidate for an anti-epileptic drug because it would not affect synaptic transmission at lower frequency but would reduce synaptic transmission during excessive brain activity. The dynamin 1 phospho-mimetic peptide is a good candidate for such a dynamin 1 inhibitor.

A small number of dynamin inhibitors have been reported. The first was a small series of long-chain amines based on myristoyl trimethyl ammonium bromide (MiTMAB), and the second was a large series of bis-tyrphostin (Bis-T; Figure 6(c)) analogs. Both drugs inhibit lipid-stimulated dynamin 1 GTPase activity with a median inhibitory concentration (IC_{50}) in the low micromolar range, and MiTMAB inhibits endocytosis in exteriorized kidneys. A new small-molecule dynamin inhibitor, dynasore, also inhibits GTPase activity (Figure 6(c)). All these drugs reversibly inhibit both SVE and RME in cultured neurons or nonneuronal cells. Dynasore has no effect on SV exocytosis, but diminishes synaptic responses with repeated stimulation in hippocampal neurons expressing the fluorescent tracer synaptopHluorin (Figure 6(d)). This illustrates the potential for dynamin inhibitors to regulate synaptic transmission *in vivo*. Further development of these compounds may prove useful in the future treatment of synaptic transmission disorders.

Dynamin and Disease

There are several emerging links between the dynamin superfamily of proteins and human diseases, particularly neurodegeneration. Among the classic dynamins, only dynamin 2 has been directly implicated to date. Frame-shift and missense mutations affecting the PH domain lead to a dominant intermediate form of Charcot–Marie–Tooth disease (CMT) called DI-CMTB. CMT is a hereditary motor and sensory neuropathy, characterized by various progressive distal sensory maladies such as numbing of the extremities, deafness, skeletal deformities, and muscle weakness. Missense mutations in the MD cause autosomal dominant centronuclear myopathy, a slowly progressive congenital myopathy characterized by abnormal centrally located nuclei in muscle fibers and a slowly progressing muscle weakness. It remains unclear how dynamin 2 mutations cause these neurodegenerative diseases. Mutations in mitofusin 2, mostly in the region encoding the GTPase domain, also lead to CMT type 2A. Missense mutations in another mitochondrial dynamin-related protein, OPA1, cause dominant optic atrophy (DOA), the most commonly inherited optic atrophy, which presents in childhood as loss of visual acuity associated with optic nerve degeneration. Thus, mitochondrial fusion and fission are important for mitochondrial function and for the overall health of neurons. The observations underscore the importance of dynamin-related protein in maintaining organelle morphology and biogenesis.

At least 20 mutations in atlastin 1 (*SPG3A* gene) lead to dominant pure hereditary spastic paraplegia. Symptoms start in early childhood and can impair the function of the lower limbs. Atlastin 1 is enriched in the layer V pyramidal motor neurons in the

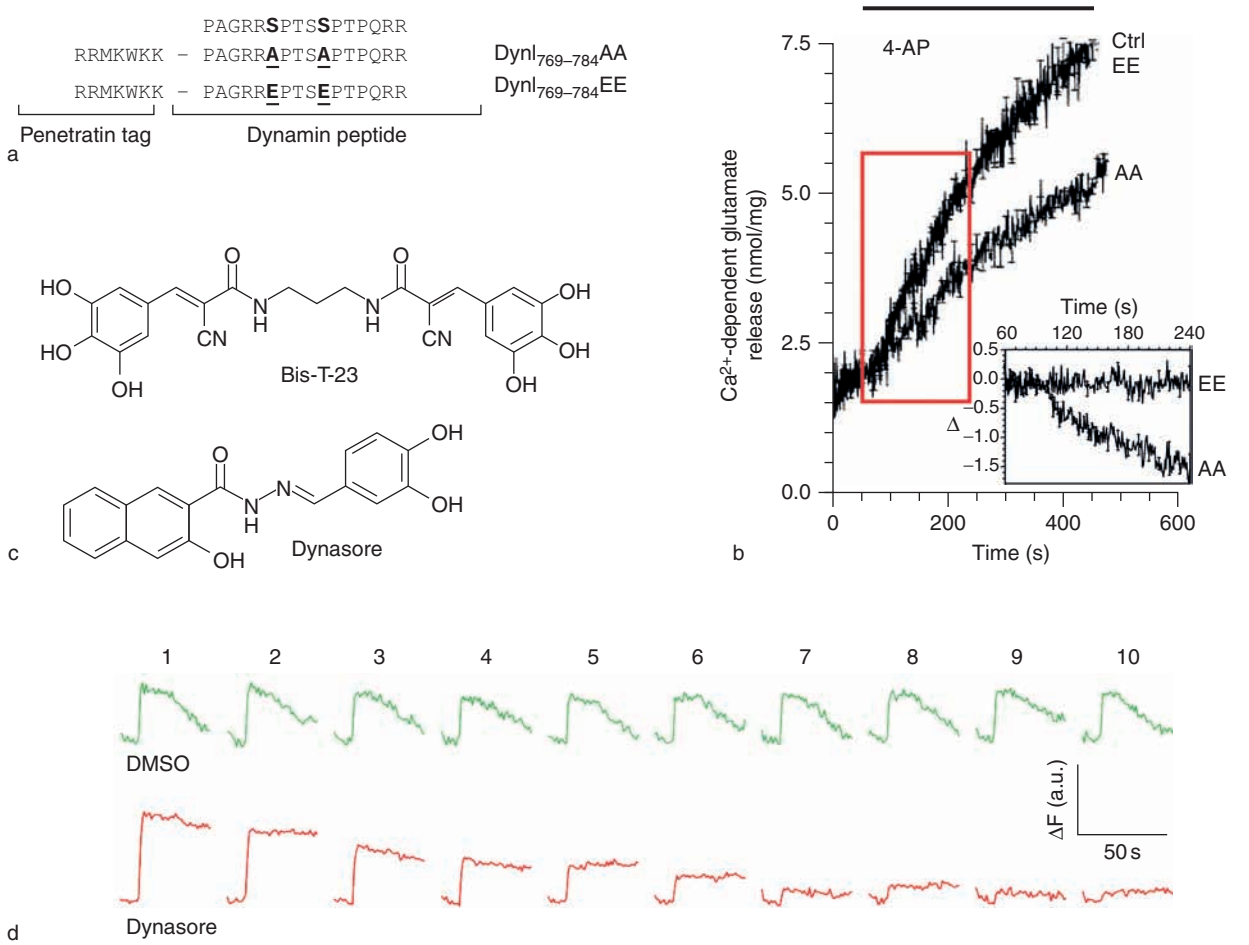

Figure 6 Dynamin controls endocytosis and synaptic transmission: (a) sequences of two penetratin-linked synthetic peptides made from the sequences in the dynamin 1 phospho-box; the $DynI_{769–784}AA$ peptide selectively inhibits syndapin 1 binding to dynamin 1, hence blocking synaptic vesicle endocytosis; (b) an activity-dependent rundown of synaptic transmission in the presence of $DynI_{769–784}AA$ peptide; (c) the chemical structure of two dynamin inhibitors, bis-T-23 and dynasore, which block the GTPase activity of dynamin; (d) rundown of synaptic responses in the presence of dynasore. In (b), synaptic vesicle exocytosis is monitored by the release of glutamate from synaptosomes stimulated with 4-aminopyridine (4-AP). The inset shows the subtracted value for the AA or EE peptides from the control exocytosis (Δ) and is zoomed on the first 3 min from 4-AP addition (which was at 60 s) to highlight the lack of effect during the first 45 s. In (d), synaptic vesicle exocytosis and endocytosis were measured in the synapses of hippocampal neurons expressing the reporter construct synaptopHluorin. An increase in fluorescence intensity indicates exocytosis, whereas a decrease indicates endocytosis. Note the lack of endocytosis in the presence of dynasore, which subsequently causes synaptic responses to diminish with repeated stimulation. In the control condition (DMSO), however, the responses remain stable. AA, $DynI_{769–784}AA$; Ctrl EE, $DynI_{769–784}EE$; DMSO, dimethyl sulfoxide; GTP, guanosine triphosphate. (a, b) Reproduced from Anggono V, Smillie KJ, Graham ME, et al. (2006) Syndapin I is the phosphorylation-regulated dynamin I partner in synaptic vesicle endocytosis. *Nature Neuroscience* 9: 752–760. (d) Reproduced from Newton AJ, Kirchhausen T, and Murthy VN (2006) Inhibition of dynamin completely blocks compensatory synaptic vesicle endocytosis. *Proceedings of the National Academy of Sciences of the United States of America* 103: 17955–17960.

cerebral cortex, which have among the longest axons in the CNS. Mutations in an atlastin 1 binding partner, spastin (*SPG4* gene), also cause hereditary spastic paraplegia. Atlastin mutations affect vesicle trafficking from the ER to Golgi and impair Golgi morphogenesis, suggesting that this underlies the degeneration or dysfunction of the motor neurons. hGBP1 is the first GTPase known to be secreted from endothelial cells. Although no human mutations are known yet, increased hGBP1 concentrations are found in the cerebrospinal fluid of patients with bacterial meningitis. GeneChip arrays further suggest that the expression of the *GBP1* gene may play a role in the development of multidrug resistance.

See also: Apoptosis in Neurodegenerative Disease; Endocytosis and Presynaptic Scaffolds; Epilepsy; Epilepsy: Neuronal Death; Hereditary Spastic Paraplegia; Neuropeptides: Epilepsy; Presynaptic Endosomes; Synaptic Vesicles; Synaptic Transmission: Models; Synaptosomes.

Further Reading

Fergusson SM, Brasnjo G, Hayashi M, et al. (2007) A selective activity-dependent requirement for dynamin 1 in synaptic vesicle endocytosis. *Science* 316: 570–574.

Hinshaw JE (2001) Dynamin and its role in membrane fission. *Annual Review of Cell and Developmental Biology* 16: 483–519.

Koenig JH and Ikeda K (1989) Disappearance and reformation of synaptic vesicle membrane upon transmitter release observed under reversible blockage of membrane retrieval. *Journal of Neuroscience* 9: 3844–3860.

Liu JP, Sim AT, and Robinson PJ (1994) Calcineurin inhibition of dynamin I GTPase activity coupled to nerve terminal depolarization. *Science* 265: 970–973.

Macia E, Ehrilich M, Massol R, et al. (2006) Dynasore, a cell-permeable inhibitor of dynamin. *Developmental Cell* 10: 839–850.

Narayanan R, Leonard M, Song BD, et al. (2005) An internal GAP domain negatively regulates presynaptic dynamin in vivo: A two-step model for dynamin function. *Journal of Cell Biology* 169: 117–126.

Praefcke GJK and McMahon HT (2004) The dynamin superfamily: Universal membrane tubulation and fission molecules? *Nature Reviews Molecular and Cell Biology* 5: 133–147.

Roux A, Uyhazi K, Frost A, and De Camilli P (2006) GTP-dependent twisting of dynamin implicates constriction and tension in membrane fission. *Nature* 441: 528–531.

Sweitzer SM and Hinshaw JE (1998) Dynamin undergoes a GTP-dependent conformational change causing vesiculation. *Cell* 93: 1021–1029.

Tan TC, Valova VA, Malladi CS, et al. (2003) Cdk5 is essential for synaptic vesicle endocytosis. *Nature Cell Biology* 5: 701–710.

Dysautonomia: Familial

G Gold-von Simson and F B Axelrod, New York University Medical Center, New York, NY, USA
S A Slaugenhaupt, Massachusetts General Hospital, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Clinical Phenotype

Familial dysautonomia (FD) was initially termed Riley–Day syndrome after the first two authors that originally described the disorder. It is now appreciated that FD is one example of a group of rare disorders termed hereditary sensory and autonomic neuropathies (HSANs). Within this classification, FD is HSAN type III. The HSANs are variable disorders in which normal migration and maturation of neural crest-derived cells has been impeded, especially those destined to evolve into the sensory and autonomic populations. The genetic defect affects prenatal neuronal development so that symptoms are present from birth, although individual expression varies widely. Because the entire autonomic nervous system is affected, there is a pervasive effect on the functioning of other systems.

The phenotypic expression of FD is extremely variable. Some manifestations of the disorder, such as hypotonia and poor suck, are apparent at birth, and others become more prominent and problematic as a function of age. In addition to decreased sensitivity to temperature and pain and consistent lack of overflow emotional tears, there are prominent gastrointestinal symptoms that include discoordinated swallows and gastroesophageal reflux resulting in aspiration and recurrent lung infections. The most prominent manifestation of abnormal gastrointestinal dysmotility in FD individuals is the propensity to vomit. Vomiting can occur intermittently as part of a systemic reaction to physical or emotional stress, or it can occur daily in response to the stress of arousal. Because vomiting is often associated with hypertension, tachycardia, diffuse sweating, and even personality change, this constellation of signs has been termed the dysautonomic crisis. Furthermore, FD patients have devastating cardiovascular perturbations that include postural hypotension without compensatory tachycardia, episodic hypertension, and arrhythmia with risk of sudden death. Insensitivity to hypoxia increases mortality, as it induces profound circulatory responses consistent with sympathetic denervation. In contrast to controls, when FD subjects are exposed to hypoxia during rebreathing, there is a marked decrease in heart rate and systemic blood pressure. Other frequent features of the disorder are delayed developmental milestones, short stature, dysarthria, and orthopedic problems such as kyphoscoliosis. Many symptoms worsen with advancing age, such as gait and vision due to spinocerebellar and optic atrophy, respectively. Fertility has been demonstrated in both sexes, and intelligence is generally within the normal range, although an increased incidence of learning disorders has been noted. With supportive care, morbidity and mortality have been reduced. Currently, approximately 54% of patients are over 15 years of age, and the probability of survival to age 40 is 50%.

Diagnosis

For more than 50 years, diagnosis was determined solely by clinical criteria, with confirmation in questionable cases supported by neuropathological data from sural nerve biopsy specimens. These clinical criteria consisted of absence of overflow emotional tearing, decreased deep tendon reflexes, absence of fungiform papillae on the tip of the tongue, and an absent axon flare following intradermal histamine in an individual of Ashkenazi Jewish or Eastern European Jewish extraction. However, with the recent identification of the genetic mutations causing the disorder, DNA diagnosis is now available. Screening programs for the Ashkenazi Jewish population are now globally available. The impact of these programs on the FD patient population is already appreciated as the number of new births has declined steadily. As a result, there is increasing attention on maximizing the function and improving the prognosis for the existing patient population.

Pathology

Although clinical manifestations suggest both central and peripheral nervous system involvement, consistent neuropathological lesions have only been described for the peripheral nervous system. Most of the symptoms in FD can be explained by the moderate to substantial lack of neurons in sensory, sympathetic, and some parasympathetic ganglia, as well as the continued attrition of these neurons throughout life. For example, superior cervical ganglia in FD patients have on average 12% of the number of neurons as ganglia from age-matched controls. Marked neuronal losses have also been documented in dorsal root ganglia and the parasympathetic sphenopalatine ganglia, while there is only a 20% mean loss of neurons in parasympathetic ciliary ganglia. The only consistent pathological change reported in the central

nervous system (CNS) is in the brain stem, where elevated levels of substance P are found. In addition, recent studies using functional neuroimaging provide new evidence for central dysfunction in FD, as there is hyperperfusion of temporal regions during the dysautonomic crisis.

Biochemical Abnormalities

Consistent with neuropathological descriptions of a decreased sympathetic neuronal population, FD patients have abnormal urinary catecholamine metabolite ratios. Homovanillic acid (HVA), the end product of dopamine, is increased and vanillyl mandelic acid (VMA) is decreased, resulting in elevated urinary HVA:VMA ratios. In addition, plasma levels of noradrenaline and dopamine β-hydroxylase (DBH) do not increase with standing. In fact, FD patients appear to have a distinctive pattern of plasma levels of catechols. Regardless of posture, plasma levels of dopa are disproportionately high and plasma levels of 3,4-dihydroxyphenylglycol (DHPG) are low, resulting in elevated plasma dopa:DHPG ratios, which are not seen in other disorders associated with neurogenic orthostatic hypotension. The high plasma dopa levels are consistent with FD subjects having an increased proportion of tyrosine hydroxylase in superior cervical ganglia. As a result of decreased sympathetic and parasympathetic innervation of peripheral tissues, there is hypersensitivity to adrenergic and cholinergic agents.

Genetics

FD is a recessive genetic defect with a remarkably high carrier frequency in the Ashkenazi Jewish population. The average carrier frequency is 1 in 30; however, it has been reported to be as high as 1 in 18 in Ashkenazi Jews of Polish descent. In 1993, the FD gene was mapped to an 11 cm region of human chromosome 9q31–33 using genetic linkage analysis in 26 families with more than two affected members. The closely linked markers showed strong evidence of allelic association with a single major haplotype present on >99.5% of FD chromosomes, suggesting that the overwhelming majority of FD patients shared a single founder mutation. Positional cloning strategies were used to identify mutations in a candidate gene known as *IKBKAP*. To date, three FD-causing mutations have been identified in *IKBKAP*: an intronic noncoding point mutation, IVS20 + 6T > C, which leads to variable skipping of exon 20, and two missense mutations, R696P (four patients) and P914L (one non-Jewish patient).

Importantly, all FD patients possess at least one copy of the major IVS20 + 6T > C mutation; 99.5% are homozygous, and five patients are compound heterozygotes for the missense mutations. The major mutation results in variable tissue-specific skipping of exon 20 in the *IKBKAP* transcript and thus weakens, but does not completely inactivate, the 5′ splice site of exon 20 (Figure 1). FD patient cell lines express both wild-type and mutant *IKBKAP* mRNA and are therefore capable of producing functional IKAP protein. This has also been shown in FD patients, making the disease quite unique in that patients have variable levels of IKAP protein in different tissues. The lowest levels of normal *IKBKAP* are produced in tissues from the central and peripheral nervous system (Figure 2). These observations suggest that the cellular level of IKAP protein in neurons falls below the threshold level required for normal development and maintenance.

IKAP Function

The *IKBKAP* gene encodes IKAP, which was originally identified as an IκB kinase complex-associated protein. However, this name is now a misnomer, as subsequent work has disproved the role of IKAP in this pathway. IKAP is a subunit of the highly conserved complex Elongator, which is involved in transcription elongation. Recently, RNA interference studies have shown that depletion of IKAP, and therefore Elongator, results in reduced transcriptional

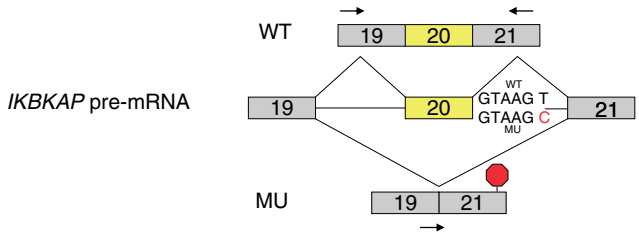

Figure 1 Schematic showing the mRNA isoforms produced in familial dysautonomia (FD). The first six base pairs of the intron 20 sequence are shown, with the FD mutation (T > C) shown in red. WT indicates the normal *IKBKAP* mRNA containing exon 20. MU indicates the mutant message that lacks exon 20 and leads to a downstream stop codon in exon 21. The arrows indicate the locations of the PCR primers used to amplify the mRNA.

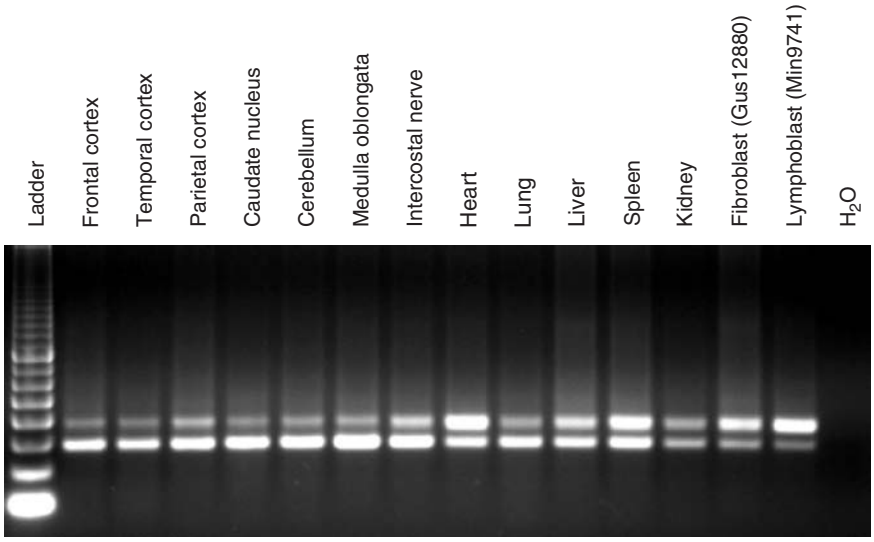

Figure 2 PCR amplification of wild-type and mutant *IKBKAP* mRNA in patient tissues showing tissue-specific variability in splicing. The top band is wild-type (contains exon 20), the bottom band is mutant (skips exon 20). Primers used are shown in Figure 1. Nervous system tissues show a marked reduction of normal mRNA.

elongation of several target genes via histone H3 hypoacetylation. A subset of these target genes is required for cell motility, and FD patient cells were shown to be defective in cell migration assays, suggesting that defective cellular motility may underlie the developmental neuropathology of FD. IKAP has also been implicated in a variety of cellular functions in addition to its proven role in transcription. Data in yeast suggest that the Elongator complex is involved in exocytosis and tRNA modification, while studies in mammalian cells implicate IKAP as a scaffold protein involved in cytoplasmic c-Jun N-terminal kinase (JNK) activation in response to extracellular stress. IKAP's role in these processes, and how they relate to the FD phenotype, is not yet understood.

Treatment of FD

At present, treatment of FD remains preventive, symptomatic, and supportive and must be directed toward specific problems. Dysphagia is initially treated by thickening feeds and feeding therapy. However, the insertion of a gastrostomy is frequently required. The gastrostomy ensures a route to maintain adequate hydration yet avoids aspiration. Esophageal dysmotility and gastroesophageal reflux are also common and increase the risk of aspiration. When medical management with prokinetic agents, thickening of feeds, and positioning are not successful, then surgical intervention (fundoplication) is performed. Failure of medical management is defined as radiographic evidence of lower respiratory disease, hematemesis, apnea, or failure to maintain adequate weight gain, growth, and hydration.

Treatment of the dysautonomic vomiting crisis is perhaps one of the most challenging aspects of care. In addition to treating the precipitating cause, for example, the stress of an infection, the crises are managed with a combination of intravenous or rectally administered diazepam and avoidance of dehydration and aspiration. The remarkable response to diazepam is consistent with the hypothesis that the crisis may be a type of autonomic seizure. If there is associated refractory hypertension, then clonidine via the gastrostomy or as a transdermal patch may be tried.

Postural hypotension can be treated with standard methods such as adequate hydration, lower extremity exercises, elastic stockings, fludrocortisone, and midodrine. Hypertension is usually transitory, so treatment should be directed to factors precipitating the hypertension such as anxiety or visceral pain. If the hypertension is persistent, diazepam and clonidine have been found to be particularly effective.

To compensate for alacrima, artificial tears are instilled on a regular basis. Corneal complications have been decreasing with regular use of artificial tear solutions and maintenance of normal body hydration. In addition, cautery of the tear duct puncta has been used in refractory situations. Tarsorrhaphy has been reserved for unresponsive and chronic situations. Corneal transplants have had limited success.

The nature of the major FD mutation suggests an alternative route toward the development of therapies for FD. It has been speculated that by raising the amount of the normal protein product, IKAP, the progression of FD will be slowed and the clinical symptoms reduced. RNA splicing mutations are

sometimes referred to as 'leaky' since they produce both mutant and wild-type transcripts from the same allele, as in FD. There are several examples in both murine and human disease which demonstrate that subtle differences in the level of normal protein produced can lead to striking differences in disease phenotype and severity. Therefore, splicing defects have become attractive candidates for the development of therapeutics since small shifts in splicing efficiency may lead to drastic phenotypic improvement. Targeting splicing defects as a means of treating genetic disease is a young field that has shown some initial promise. A variety of strategies have been used to alter splicing, including oligonucleotide-based technologies, RNA-mediated mechanisms, and pharmacologic agents. Screening strategies using FD patient cells have identified three compounds that alter the expression or splicing of the *IKBKAP* gene. Tocotrienols have been shown to increase the expression of *IKBKAP* in cultured cells. Two compounds, epigallocatechin gallate (EGCG) and kinetin, have been reported to alter splicing of *IKBKAP* and enhance inclusion of exon 20, thereby increasing the amount of normal *IKBKAP* mRNA. The bioactivity of kinetin is significantly higher than that of EGCG, and kinetin has been shown to increase the amount of IKAP protein. To date, efficacy of these compounds in patients has not been shown and clinical trials are necessary.

Summary

Although FD is genetically and pathologically homogenous, there is great variability in clinical expression resulting in protean manifestations. Thus, FD serves as a model to enhance our knowledge of the autonomic nervous system and the various ways it can influence other systems. Furthermore, the most important piece of information to result from the discovery of the FD gene is that, despite the presence of a homozygous mutation, cells from individuals with FD can produce wild-type *IKBKAP* message and IKAP protein. This offers the hope that protein expression and primary and secondary gene functions may be manipulated by potential therapeutic agents.

See also: Autoimmune Autonomic Neuropathy; Autonomic Disorders; Autonomic Failure; Autonomic Nervous System; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Anderson SL, Coli R, Daly IW, et al. (2001) Familial dysautonomia is caused by mutations of the IKAP gene. *American Journal of Human Genetics* 68: 753–758.

Anderson SL, Qiu J, and Rubin BY (2003) Tocotrienols induce IKBKAP expression: A possible therapy for familial dysautonomia. *Biochemical and Biophysical Research Communications* 306: 303–309.

Anderson SL, Qiu J, and Rubin BY (2003) EGCG corrects aberrant splicing of IKAP mRNA in cells from patients with familial dysautonomia. *Biochemical and Biophysical Research Communications* 310: 627–633.

Axelrod FB (2002) Hereditary sensory and autonomic neuropathies: Familial dysautonomia and other HSANs. *Clinical Autonomic Research* 12: 2–14.

Axelrod FB (2007) Autonomic and sensory disorders. In: Rimoin DL, Connor JM, Pyeritz RE, and Korf BR (eds.) *Principles and Practice of Medical Genetics,* 5th edn., pp. 2802–2816. Philadelphia: Churchill Livingstone.

Axelrod FB, Goldberg JD, Ye XY, and Maayan C (2002) Survival in familial dysautonomia: Impact of early intervention. *Journal of Pediatrics* 141: 518–523.

Axelrod FB, Goldstein DS, Holmes C, Berlin D, and Kopin IJ (1996) Pattern of plasma levels of catecholamines in familial dysautonomia. *Clinical Autonomic Research* 6: 205–209.

Axelrod FB, Zupanc M, Hilz MJ, and Kramer EL (2000) Ictal SPECT during autonomic crisis in familial dysautonomia. *Neurology* 55: 122–125.

Bernardi L, Hilz M, Stemper B, Passino C, Welsch G, and Axelrod FB (2002) Respiratory and cerebrovascular responses to hypoxia and hypercapnia in familial dysautonomia. *American Journal of Respiratory and Critical Care Medicine* 167: 141–149.

Close P, Hawkes N, Cornez I, et al. (2006) Transcription impairment and cell migration defects in elongator-depleted cells: Implication for familial dysautonomia. *Molecular Cell* 22: 521–531.

Hims MM, Ibrahim EC, Leyne M, et al. (2007) Therapeutic potential and mechanism of kinetin as a treatment for the human splicing disease familial dysautonomia. *Journal of Molecular Medicine* 85: 149–161.

Ibrahim EC, Hims MM, Shomron N, Burge CB, Slaugenhaupt SA, and Reed R (2007) Weak definition of *IKBKAP* exon 20 leads to aberrant splicing in familial dysautonomia. *Human Mutation* 28: 41–53.

Slaugenhaupt SA, Blumenfeld A, Gill SP, et al. (2001) Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. *American Journal of Human Genetics* 68: 598–605.

Slaugenhaupt SA, Mull J, Leyne M, et al. (2004) Rescue of a human mRNA splicing defect by the plant cytokinin kinetin. *Human Molecular Genetics* 13: 429–436.

Dyslexia: Neurodevelopmental Basis

G F Eden, Georgetown University Medical Center, Washington, DC, USA
D L Flowers, Wake Forest University School of Medicine, Winston-Salem, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Developmental dyslexia, also referred to as reading disability, is an unexpected reading problem that cannot be attributed to poor intellectual ability or educational shortcomings. Unlike acquired reading disorders such as alexia (acquired dyslexia), this reading disability is developmental in nature and accounts for about 80% of all learning disabilities. The 5–12% of the population affected with developmental dyslexia struggles to attain proficient levels of reading accuracy and fluency, without which they are prevented from deriving meaning from text. The identification of individuals with dyslexia and research into its etiology have been complicated by variations among researchers and educators in placing boundaries between dyslexic and typical readers. Further, although individuals with dyslexia have traditionally been distinguished from those poor readers with lower cognitive capacities or those lacking appropriate reading instruction, avenues for improving reading in these populations may follow similar paths. The most popular approach to counter dyslexia entails explicit instruction in phonological representation (i.e., phonological code). This approach has been reinforced by the positive relationship observed between phonological skills and reading proficiency in the majority of children during the early school grades. Further, functional brain imaging studies have revealed lower brain activity in dyslexic people during the performance of skills invoking phonological representations, and more recently, brain imaging techniques have been employed to study the functional anatomy of reading interventions. Linkage studies continue to reveal candidate genes, pointing to the multigenetic nature of dyslexia and its heterogeneous phenotype. The behavioral manifestations of dyslexia also include deficits in sensorimotor performance which, much like the language-based deficits, have been tied to perturbations in brain function. Together, this information about the behavioral manifestations and biological markers of dyslexia is being used to ensure better identification of dyslexia and more effective treatment strategies.

Behavioral Profile of Developmental Dyslexia

Definition of Developmental Dyslexia

There are numerous definitions of dyslexia. One formulated in 2002 by a working group commissioned by the National Institutes of Health and the International Dyslexia Association specifies the following:

> Dyslexia is a specific learning disability that is neurological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities and the provision of effective classroom instruction. Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede growth of vocabulary and background knowledge.

This definition serves researchers and practitioners. It reflects four decades of behavioral research indicating that the word-decoding and reading-fluency deficiencies observed in dyslexia are the symptoms of an underlying deficit in phonological processes. The definition is therefore now more specific than earlier definitions were in identifying one of the cognitive deficits and less specific about the role of IQ. Previous versions of the definition operated using a discrepancy definition. Under that model, IQ measures provided a benchmark of expected achievement against which to measure actual reading achievement. Typically in the United States, a discrepancy of least one standard deviation (15 standard score points) defines dyslexia for inclusion of a student in research studies or allocation of resources by school systems. However, the existence of a strong correlation between IQ and reading achievement biases against identifying individuals with dyslexia, in part because responses to verbal portions of the IQ test are influenced by reading experience. In addition, some of the core deficits exhibited by individuals with dyslexia are shared by poor readers with lower IQ (known as garden-variety poor readers), and these deficits are typically the focus of applied interventions (e.g., phonological coding). Hence IQ is no longer included in the above definition. From an educational perspective, these interventions are now administered in the school systems under a 'response to intervention' model, which in itself selects for those children who fail to benefit from standard instruction, with the next stage being implemented for those students

who fail to show sufficient gains at the previous level. Finally, the more recently derived data from neuroscience research indicating anatomical differences in the brains of individuals with dyslexia has motivated the inclusion of the now widely accepted notion that developmental dyslexia has a neurobiological origin. However, measures of brain function are not currently used to diagnose dyslexia. Hence the current definition of developmental dyslexia has been shaped by a convergence of research findings and the practical aspects of educational reality.

Cognitive Characteristics of Developmental Dyslexia

For the majority of children, early phonological coding skills predict the course of their later reading development. These skills encompass sound manipulation at the phoneme level (phonemic awareness), rapid naming of letters and numbers (phonological retrieval), and short-term verbal memory (phonological recoding). Children who fail to show awareness of the phonemic structure of spoken words at the kindergarten level are more likely to struggle in acquiring grapheme–phoneme correspondence rules, and this affects their long-term ability to decode words. These observations have led to ways by which to identify children with dyslexia early, as those children with poor phonological representation are considered to be at higher risk for reading disability, even prior to the advent of formal schooling. This longitudinal research, together with the fact that treatments aimed at promoting the phonological code are fairly successful in helping children with reading disabilities, has given rise to the notion that a phonological core-deficit hypothesis represents the best explanation for dyslexia.

However, while phonological processes are impaired in many poor readers, these children may also exhibit deficits in other language-related processes, including vocabulary, morphology and syntax, and text comprehension. Likewise, measures of vocabulary and grammar attained at kindergarten are strongly associated with reading outcome in second grade. Finally, skills in the domain of orthography, such as the ability to name words based on visual information or appreciating letter patterns that are legitimate, affect reading fluency and reading comprehension in the later grades.

In sum, the most prominent cognitive deficit associated with dyslexia is that pertaining to phonological coding. However, it should not go unnoticed that other language skills make strong predictions for later reading development and that the process of reading itself promotes phonological representation. Hence it would be fair to say that dyslexia is associated with a number of language-based deficits, with poor decoding serving as the identifying feature. However, because the definition of dyslexia is largely exclusionary, it is not surprising that the result is the identification of heterogeneous populations of dyslexics. This in turn results in variability in the description of the dyslexic phenotype. Although the phonological and other language deficits are prominent and most are directly related to the process of reading acquisition, other behavioral manifestations have been observed and will be described next.

Sensorimotor Characteristics of Developmental Dyslexia

While the behavioral measures of phonological coding described above have provided a strong theoretical framework at the cognitive level, this approach has not been able to address other diverse symptoms observed in dyslexia outside the domain of language. Psychophysical studies in the domain of vision and audition have suggested aberrations within lower-level, perceptual systems, and impaired motor performance in individuals with dyslexia also suggest the possibility of motor system involvement. The hypotheses put forward in this area of research are sometimes based on systems, such as the magnocellular-based visual perceptual hypothesis, or on concepts (such as temporal integration) or are directly linked to locations in the brain (e.g., parietal and cerebellar deficits).

It was first demonstrated by Lovegrove and colleagues using psychophysical experimental techniques that magnocellular visual system function is altered in individuals with dyslexia. These studies were largely confirmed by data acquired using neuroanatomical (altered magnocellular layers in the lateral geniculate nucleus at postmortem), electrophysiological, and functional neuroimaging techniques comparing dyslexics with nondyslexics. For example, the perception of visual motion, which relies on magnocellular function, usually induces a robust activation response in the dorsal visual stream, including parietal cortex and area MT/V5. However, in individuals with dyslexia, this activity is reduced or absent (see **Figure 1**). These results raise the question of the nature of the relationship between dorsal stream function, which is highly developed for visuospatial perception in human and nonhuman primates, and reading, a skill that is uniquely human and has come into use very recently. The dual symptomatology of impaired visual motion perception and phonological processing could be due to the close proximity of brain regions functionally specialized for these divergent tasks, resulting in equal susceptibility under circumstances of brain perturbation. Another possibility is that visual perception that requires tracking of rapidly changing information is impaired because of systematic differences in sensory systems that handle rapidly changing

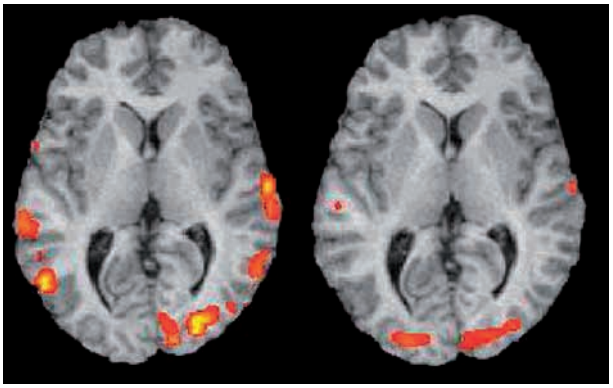

Figure 1 Horizontal sections of the brain reveal that typical adult readers (left) display activity in visual cortex, including area MT/V5, during the perception of visual motion. Adults with dyslexia (right) lack this activity. From Eden GF, VanMeter JW, Rumsey JW, Maisog J, and Zeffiro TA (1996) Abnormal processing of visual motion in dyslexia revealed by functional brain imaging. *Nature* 348: 66–69.

information. While neither of these hypotheses has been confirmed, this latter idea has some parallels with a hypothesis that has emerged from the study of auditory perception in children with reading problems.

In this case, the poor readers were children with a primary diagnosis of specific language impairment (SLI; see the section titled 'Comorbid disorders') and were found to require longer intervals in order to accurately perceive the sequence of rapid successive auditory inputs. This work by Tallal and colleagues also suggested a link between early sensory processing of nonverbal sounds and phonological processing. This led to an intervention strategy in which children completed exercises devised to improve processing of rapidly presented successive acoustic stimuli and to improve speech perception. However, while initial laboratory work showed promise for improving language function in children with SLI, independent, randomized controlled studies have not yet demonstrated any long-term benefits to reading. Nevertheless, the temporal-deficit hypothesis continues to be actively investigated and provides an interesting model by which to examine whether the remediation of behavioral deficits observed in dyslexia generalize to improvements in other domains, specifically reading.

On the other hand, a locus-specific theoretical framework for understanding the behavioral manifestations of dyslexia has been built around the cerebellum. Nicolson and Fawcett, for example, have argued that reading and phonological deficits arise in parallel with poor skill automaticity, the latter being dependant on cerebellar integrity. While this work is supported by behavioral and functional brain imaging results, the claims that reading gains can be brought about via regimens that target functions carried out by the cerebellum have been controversial.

The appeal of most brain-based research is that it ties into well-characterized neural functions and frequently offers an explanation for the numerous deficits noted in dyslexia. However, a common weakness of this research is that the causal relationship between sensorimotor deficits and reading has not been demonstrated. To do so would require the examination of behavioral deficits over a wide age range or longitudinally, in large participant groups. Only in this way can it be established whether a perceptual deficit plays a causal role in disrupting phonological coding abilities. Research of this type is ongoing and will undoubtedly provide a clearer picture of the direction of causality among these factors.

Biological Basis of Developmental Dyslexia

Brain-Based Evidence

A landmark study by Galaburda and colleagues in 1985 revealed microscopic anomalies in the brains of individuals with dyslexia examined at postmortem (see Figure 2). As the appropriate brain imaging tools became available, this finding provided the impetus for a large number of investigations into the anatomical and physiological integrity of reading-impaired participants *in vivo*. Since then, an abundance of investigations using functional neuroimaging (positron emission tomography and functional magnetic resonance imagery) have reported anomalous patterns of activation associated with reading and phonological representation in adults with persistent and compensated developmental dyslexia. Although these studies have varied in their use of tasks (e.g., the detection or judgment of rhymes, nonword reading, and implicit reading) and in terms of the languages spoken by the participants, their respective observations have largely been consistent. The most common finding for differential brain activity between dyslexic and nondyslexic participants falls into left-hemisphere perisylvian regions. Differences in task-related signal change in the left parietal cortex are commonly reported in children and adults with dyslexia. It is important that these differences are not the result of lower task performance by the dyslexic participants or simply the consequence of impoverished reading experience. In addition, there are reports of altered involvement of the left inferior frontal gyrus in dyslexia (some report hypo- and others hyperactivation) and underactivity in the left occipitotemporal junction (BA 37), the latter being a common finding. For example, during the performance of reading

tasks, dyslexic adults from different countries showed common underactivity in this region despite their differences in language experience (see Figure 3).

Considering the number of brain regions identified in these studies, it is conceivable that dyslexia may arise in a number of different ways. Further, the loci at which functional differences are observed do not necessarily represent the original site of perturbation. However, recent MRI studies measuring fractional anisotropy (estimating white matter integrity) do concur with the physiological differences observed in left temporal parietal in dyslexia. Therefore, not only do functional and anatomical findings converge, but also data emerging from genetic research are beginning to offer a gene-based explanation for the neuronal migrations that may lie behind these neuroanatomical and neurophysiological differences observed in dyslexia.

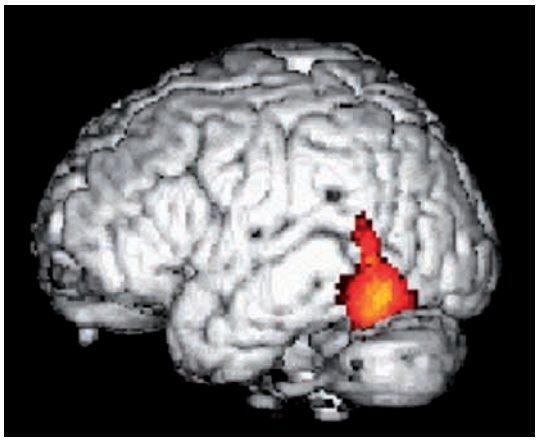

Figure 2 An example of a brain examined at postmortem. The individual was diagnosed as having had dyslexia. Developmental anomalies consist of neuronal ectopias and architectonic dysplasias. Adapted from Galaburda AM, Sherman GF, Rosen GD, Aboitiz F, and Geschwind N (1985) Developmental dyslexia: Four consecutive patients with cortical anomalies. *Annals of Neurology* 18: 222–233.

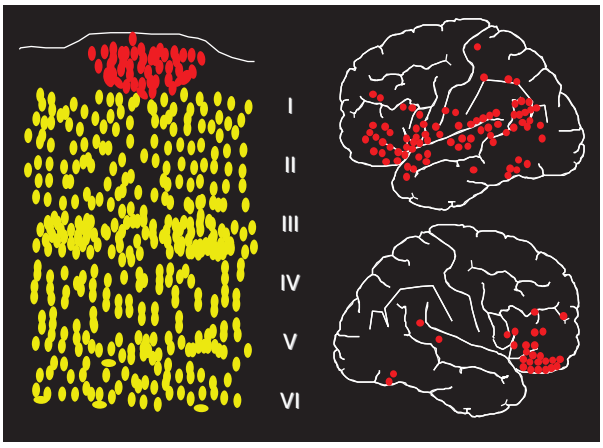

Figure 3 During word processing, dyslexics in French-, Italian-, and English-speaking countries, despite their inherent differences in language systems, all exhibit less activity than controls at the occipitotemporal junction of the left hemisphere. From Paulesu E, Demonet J-F, Fazio F, et al. (2001) Dyslexia: Cultural diversity and biological unity. *Science* 291(5511): 2165–2167.

Genetic Evidence

There is abundant evidence that component reading skills which influence reading acquisition are strongly influenced by genes; however, it is already clear that the elusive 'dyslexia gene' is a myth. That is, at least seven regions of susceptibility have been identified by linkage studies (on chromosomes 1, 2, 3, 6, 15, 18, and X), and four candidate genes (on chromosomes 3, 6, and 15) have been associated with dyslexia phenotypes. This complexity may help explain the co-occurrence of dyslexia with other learning and developmental disorders. In any case, there is hope that genetic screening, together with other information, will help to identify individuals at risk for dyslexia well before reading problems emerge in the classroom.

Consistent with human postmortem reports of migration anomalies as well as with animal models, the candidate genes implicate abnormalities in the coding of axonal guidance and neuronal migration. In the present case, they are believed to be involved in the abnormal organization of cortical networks that contribute to the component skills of reading acquisition. As a caveat, however, it must be noted that the known functions of the candidate genes might be expected to lead to a general rather than to a specific disorder. Furthermore, it seems likely at this stage of research that the culprits are mutations in regions that regulate the candidate genes.

Dyslexia in Different Languages and Orthographies

How many of the reading problems encountered by individuals with dyslexia are unique to the English language? Although most of the language-based deficits described above have been reported for dyslexia in a variety of different countries and the neurobiological basis for these problems is shared in large part among users of the alphabet (see Figure 3), there are differences in the rate of reading acquisition that can be attributed to variations in language characteristics. For example, it has been argued by Ziegler and Goswami that the differences in syllable structure and the consistency by which phonology is mapped to orthography will affect the rate at which children reach reading proficiency (e.g., faster in German than in English). It is interesting that skills such as phonological coding also play a role in logographic writing

systems, such as Chinese, although in this case the role of orthography clearly is more dominant. Not surprisingly, Chinese reading, compared with English reading, evokes greater responses in the visual system and less activity in the superior temporal lobe regions that are associated with phonological processing in the alphabetic languages. Struggling Chinese readers show less activity in frontal systems thought to be responsible for the mapping of orthography to phonology and orthography to semantics (see Figure 4). In other words, dyslexia is omnipresent, and while there are common elements that support reading across languages and writing systems, there are also unique differences in the brain signatures for reading, each of which is most likely susceptible to changes that could result in reading disability.

Comorbid Disorders

Developmental Dyslexia and Attention-Deficit/Hyperactivity Disorder

It has been suggested that as many as 40% of children with dyslexia also have attention-deficit/hyperactivity disorder (ADHD). Because these disorders have been seen both alone and together, there has been some debate as to whether they were entirely independent. More recently, a common etiology to their co-occurrence has been supported by genetic evidence of susceptibility regions on chromosomes 2, 8, 14, and 15. Separately, either disorder can affect learning or either can affect focus of attention, thus making their independent and collective contributions particularly difficult to identify. Obviously, the importance of correct diagnosing has obvious treatment implications, whether academic, behavioral, or pharmacological. Few neuroimaging studies have focused on the question of a common etiology, as opposed to independent etiologies, of ADHD and dyslexia, with the former focusing on frontal and subcortical function and the latter on language substrates.

Developmental Dyslexia and SLI

Approximately 50% of children diagnosed with SLI in early childhood will develop learning difficulties, particularly in reading, especially if the language impairment is severe and is still present on school entry. Dyslexia is not the inevitable outcome of SLI, however, and dyslexic individuals do not necessarily have a history of SLI – evidence that the two disorders are independent. Brain imaging studies support the proposition that phonological deficits and language comprehension are served by different neural networks, further support for independent etiologies. Like ADHD, SLI has been studied alone and in combination with dyslexia. Although phonological-awareness problems might be expected to be associated with SLI, studies have found only mild deficits in that group on measures of phonological-coding skills. In contrast, dyslexic participants with or without comorbid SLI have poorer phonological skills. The separate and combined aspects of these often coexisting disorders make it prudent to adopt specific intervention strategies to address their respective deficits, highlighting phonological processes on one hand and listening comprehension on the other.

Therapeutic Approaches to Developmental Dyslexia

Given the plurality of weaknesses in addition to reading problems that are observed in individuals

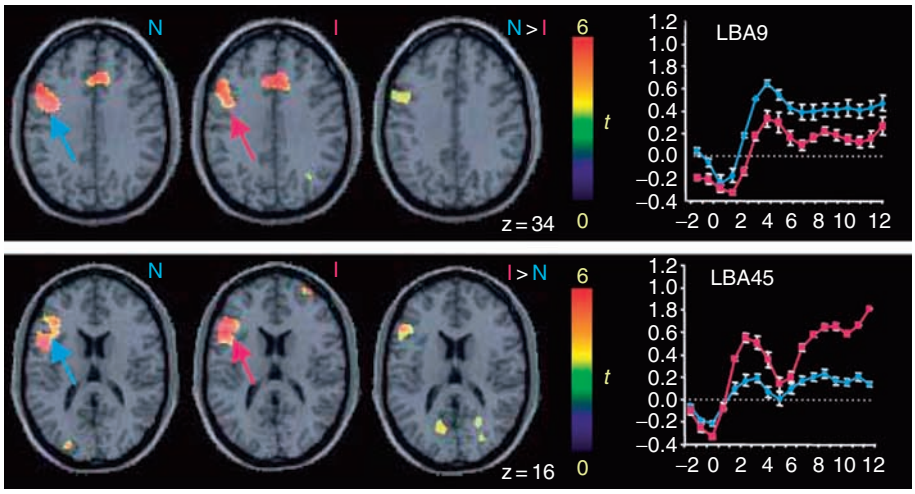

Figure 4 During a homophone judgment task, Chinese normal (blue) and impaired readers (red) differ in brain activity in left-middle frontal gyrus (top panel) and inferior frontal gyrus. From Siok WT, Perfetti CA, Jin Z, and Tan LH (2004) Biological abnormality of impaired reading is constrained by culture. *Nature* 431(7004): 71–76.

with dyslexia, the necessity for multiple treatment strategies may seem pressing. As discussed above, programs that emphasize phonological coding are known to facilitate reading acquisition. However, targeting just this one domain is not sufficient. Although gains in phonological representation can bring about growth in multiple aspects of reading, they do not generalize to all aspects of reading. For example, promoting reading fluency and reading comprehension in addition to phonology requires other types of intervention approaches. Studies targeting domains of reading other than phonological coding, such as orthography, are less numerous. Increasingly, successful treatment methods represent a combination of training approaches (combining phonics, phonemic awareness, fluency, vocabulary, and reading comprehension), which, although logical from an educational perspective, complicates any experimental approach to identify independent contributors of reading recovery. It is interesting that the methods by which these instructions are delivered in clinics or classrooms often involve 'multisensory' enhancing techniques, whereby the same principles are taught by reliance on several sensory systems, through speaking (audition), visual imagery or visual representation touch, and motor system involvement (writing, air writing, or body movements that correspond to the number of syllables or phonemes in a given word). These observations from the classroom are ripe for neuroscientific inquiry, based on our understanding of neural responses under conditions of single versus multiple sensory stimulation in the auditory and visual systems.

Evaluating the benefits of various remediation strategies for dyslexia is beyond the scope of this article; a thorough review of representative therapies has been published by Alexander and colleagues. From the neuroscientific perspective, it is of value to establish the neural correlates of successful reading interventions as these provide important insight not only into brain mechanisms of dyslexia but also into behavioral plasticity. Based on a rich animal and human literature of training-induced brain plasticity and functional reorganization following rehabilitation from cerebral injury, one can consider several paths that would lead to changes following successful reading intervention in developmental dyslexia. One is a mechanism by which brain regions with known anomalies in dyslexia are shown to establish a more typical response. An alternative brain response could involve the manifestation of brain activity in brain regions not typically associated with the reading process. Such compensatory strategies have been described for stroke patients in brain regions contralateral to the injury (and which encompassed the site typically activated by the now relearned skill). Functional brain imaging technology has allowed these models to be subjected to the test. For example, in a study comparing dyslexic adults with persistent reading problems following 120 h of intensive reading intervention to a matched group of dyslexics receiving no intervention at all, the intervention group was found to have increased activity in bilateral parietal and right perisylvian cortex following the intervention (see Figure 5). These results suggest that the significant gains in reading enjoyed by the intervention group could be attributed to right hemisphere compensation combined with increases in a more typically reading-related left hemisphere brain region. Several functional brain imaging studies have also been conducted in children receiving reading interventions, and a review of this small literature suggests a role for age (or brain maturation) in directing the pattern of treatment-induced brain plasticity. This, however, does not mean that changes in brain activity necessarily guarantee behavioral gains, and results from these studies are meaningful only if the studies follow rigorous research practices, such as the use of controlled, randomized experimental designs.

Summary

Developmental dyslexia has been studied in the context of learning, development, experiential factors, genetic influences, and brain function. For all children, a mastery of reading requires the integration and coordination of sensorimotor and cognitive processes. How these fail to operate at an optimal level to bring about accurate and fluent reading in the context of normal cognitive function and appropriate teaching methods continues to be questioned from the viewpoint of behavior and brain function. Avenues of remediation within both cognitive and sensorimotor

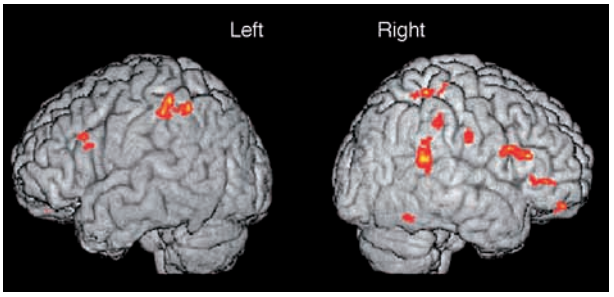

Figure 5 Functional anatomy of phonological manipulation following reading remediation (Group X Session interaction) revealed increases during phonological manipulation in left parietal cortex and fusiform gyrus. Right hemisphere increases included posterior superior temporal sulcus/gyrus and parietal cortex. Reprinted from Eden GF, Jones KM, Cappell K, et al. (2004) Neural changes following remediation in adult developmental dyslexia. *Neuron* 44(3): 411–422.

domains are under active investigation to establish the efficacy of these approaches as well as to discover their neuronal targets. Investigations into the hereditary mechanisms of reading deficits hold the promise of strengthening the understanding of the neural mechanisms of dyslexia. Together, behavioral and physiological studies continue to aid in the early identification of dyslexia and best avenues for its treatment.

See also: Alexia; Attention Deficit Hyperactivity Disorder; Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine; Connectionist Models of Language Processing; fMRI: BOLD Contrast; Language Evolution; Language Development; Language: Cortical Processes; Language: Learning Impairments; Positron Emission Tomography (PET); Reading.

Further Reading

Alexander AW and Slinger-Constant AM (2004) Current status of treatments for dyslexia: Critical review. *Journal of Child Neurology* 19(10): 744–758.

Bradley L and Bryant P (1983) Categorizing sounds and learning to read: A causal connection. *Nature* 301: 419–421.

Eckert M (2004) Neuroanatomical markers for dyslexia: A review of dyslexia structural imaging studies. *Neuroscientist* 10(4): 362–371.

Eden GF, Jones KM, Cappell K, et al. (2004) Neural changes following remediation in adult developmental dyslexia. *Neuron* 44(3): 411–422.

Eden GF and Moats L (2002) The role of neuroscience in the remediation of students with dyslexia. *Nature Neuroscience* 5(supplement): 1080–1084.

Fisher SE and DeFries JC (2002) Developmental dyslexia: Genetic dissection of a complex cognitive trait. *Nature Review Neuroscience* 3(10): 767–780.

Galaburda AM, LoTurco J, Ramus F, Fitch RH, and Rosen DG (2006) From genes to behavior in developmental dyslexia. *Nature Neuroscience* 9(10): 1213–1217.

Lovegrove WJ, Bowling A, Badcock D, and Blackwood M (1980) Specific reading disability: Differences in contrast sensitivity as a function of spatial frequency. *Science* 210(4468): 439–440.

Nicolson R, Fawcett AJ, and Dean P (2001) Dyslexia, development and the cerebellum. *Trends in Neuroscience* 24(9): 515–516.

Paulesu E, Demonet J-F, Fazio F, et al. (2001) Dyslexia: Cultural diversity and biological unity. *Science* 291(5511): 2165–2167.

Rutter M, Caspi A, Fergusson D, et al. (2004) Sex differences in developmental reading disability: New findings from 4 epidemiological studies. *Journal of the American Medical Association* 291(16): 2007–2012.

Schatschneider C, Fletcher JM, Francis DJ, Carlson CD, and Foorman BR (2004) Kindergarten predictions of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology* 96(2): 265–282.

Siok WT, Perfetti CA, Jin Z, and Tan LH (2004) Biological abnormality of impaired reading is constrained by culture. *Nature* 431(7004): 71–76.

Vellutino FR, Fletcher JM, Snowling MJ, and Scanlon DM (2004) Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry* 45(1): 2–40.

Ziegler JC and Goswami U (2005) Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin* 131(1): 3–29.

Relevant Website

http://www.interdys.org – International Dyslexia Association.

Dystonia: Classification, Genetics and Therapeutics

P Shashidharan, Mount Sinai School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Dystonia is a neurological syndrome characterized by involuntary muscle contractions, which may cause twisting and repetitive movements or abnormal postures. Dystonia can affect any part of the body, including the arms and legs, trunk, neck, eyelids, face, or vocal cords. Dystonia can involve any voluntary muscle or muscle groups. In 1911, Oppenheim was the first to use the term dystonia to describe a childhood-onset syndrome, characterized by twisted postures, muscle spasms, abnormal walking with bending, twisting of the torso, and development of sustained fixed postural defects also known as dystonia musculorum deformans. This is one of the most severe forms of dystonia affecting most regions of the body. Because the movements and resulting postures are often unusual, and the condition is rare, it is one of the most frequently misdiagnosed neurological conditions.

Dystonia is the third most common movement disorder after Parkinson's disease and tremor, affecting more than 300 000 people in North America. The prevalence estimates for primary dystonia range from 2 to 50 cases per million for early onset dystonia and from 30 to 7320 cases per million for late-onset dystonia. From an analysis of methodological information from 14 selected studies, it was concluded that all of the studies based on treatment settings or record-linkage systems, and two population-based surveys, were probably flawed by incomplete ascertainment; a third population-based study provided the largest prevalence for late-onset dystonia but probably overestimated the prevalence of the disorder. Age and ethnic differences among study populations further biased comparisons of the estimates. On the basis of methodologically more robust service-based studies and the likely percentage of underdiagnosis in a given area, more accurate prevalence estimates may be 111 per million for early onset dystonia in Ashkenazi Jews from the New York area, 600 per million for late-onset dystonia in northern England, and 3000 per million for late-onset dystonia among Italians older than 50 years of age.

Classification of Dystonia

Dystonia is classified into three categories based on the cause of dystonia, the location of symptoms in the body, and the age of onset (see Figure 1).

Classification Based on Cause

Based on the cause, dystonia is classified as either primary, if the sole symptom is dystonia, or secondary, if the dystonia is not the sole feature observed. Primary dystonias (PTDs) are largely genetic in origin, and the dystonia is the sole symptom of the patient. On the other hand, secondary dystonias result from the effects of medications, trauma, toxins, infections, or other disorders. Spinal cord, head, and peripheral injuries are also believed to be contributors to secondary dystonia. Other examples of secondary dystonias include levodopa-induced dystonia in the treatment of parkinsonism; acute and tardive dyskinesia due to antipsychotic agents; and dystonias associated with cerebral palsy, cerebral hypoxia, cerebrovascular disease, cerebral infections and postinfectious states, stroke, encephalitis, brain tumor, and toxicants such as manganese, cyanide, and 3-nitropropionic acid.

Tardive dystonia and tardive dyskinesia are caused by exposure to antipsychotic drugs. The term tardive means 'late,' indicating that the condition occurs after drug exposure, and the terms dyskinesia and dystonia describe the types of movements involved. Tardive dystonias and dyskinesias may also develop as a symptom after prolonged treatment with levodopa in some Parkinson's patients. These symptoms may develop after weeks or years of drug exposure.

Classification Based on Affected Body Regions

Dystonias that are localized to a certain body part are referred to as focal dystonias. When more than one muscle group is affected, the dystonia is called multifocal, and when adjacent body parts are involved it is referred to as segmental dystonia. When many regions are affected, it is referred to as generalized dystonia.

Focal dystonias

- *Blepharospasm*. Increased blinking and involuntary closing of the eyes characterize this disorder. Blepharospasm affects the eye muscles and usually begins gradually with excessive blinking and/or eye irritation. In the early stages, it may only occur with specific precipitating stressors, such as bright lights, fatigue, and emotional tension. It is almost always present in both eyes. Blepharospasm can occur with dystonia affecting the mouth and/or jaw and is referred to as Meige's syndrome. In such cases, spasms of the eyelids are accompanied by jaw clenching or mouth opening, grimacing, and tongue protrusion.
- *Cervical dystonia*. Cervical dystonia is also known as spasmodic torticollis, characterized by the

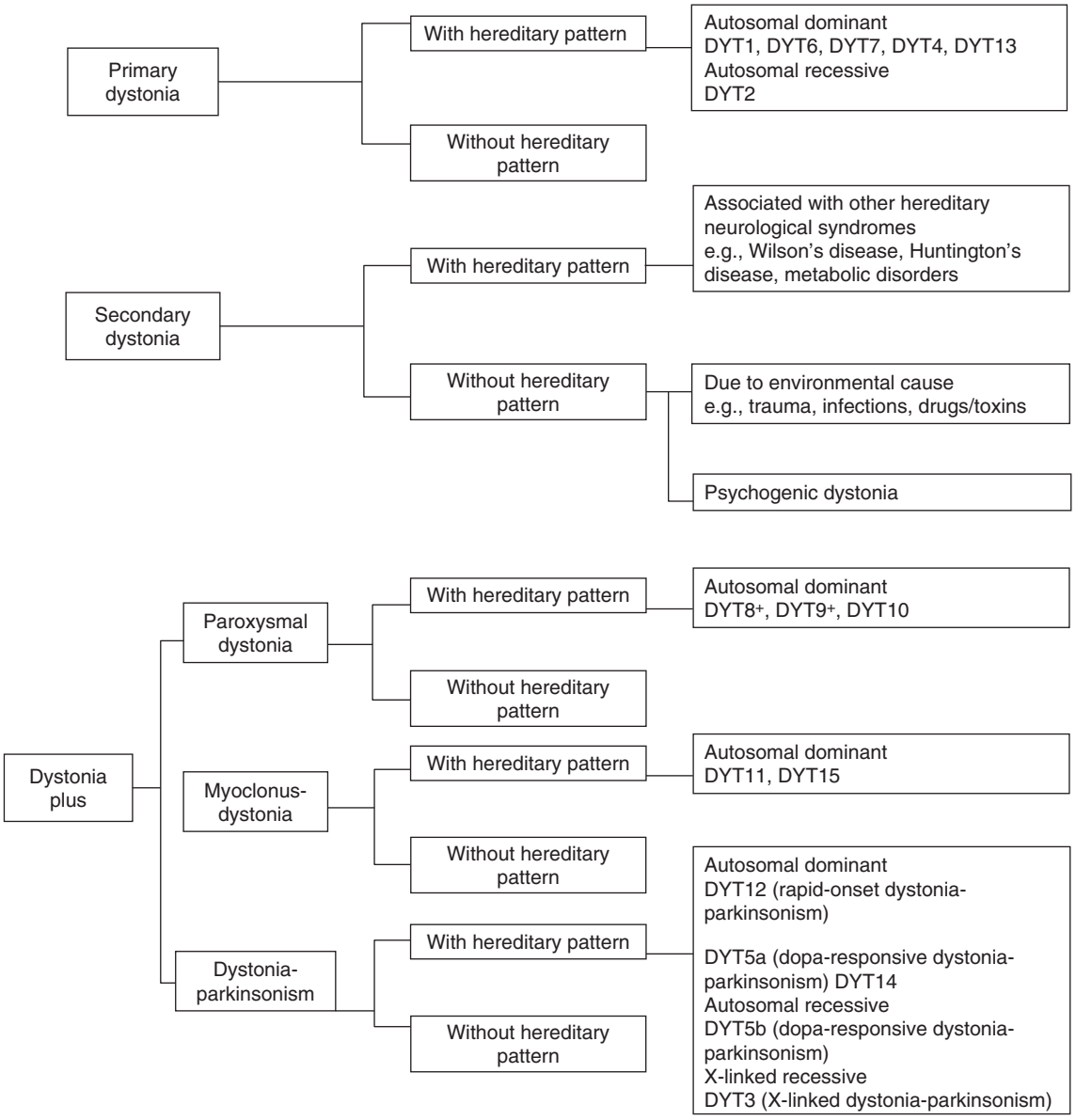

Figure 1 General classification of dystonia.

neck muscles contracting involuntarily, causing abnormal movements and posture of the head and neck. Cervical dystonia characterized by forward tilting of the head is referred to as anterocollis, backward tilting is called retrocollis, and sideways tilting is called torticollis. The contractions in the neck muscles may be sustained or jerky. Spasms in the muscles or pinching of nerves in the neck can result in considerable pain and discomfort, causing the abnormal posture of the head and neck.

- *Oromandibular dystonia*. Oromandibular dystonia is characterized by forceful contractions of the jaw and tongue causing difficulty in opening and closing of the mouth and often affecting chewing and speech. This syndrome is often associated with torticollis, blepharospasm, or spasmodic dysphonia. When oromandibular dystonia is combined with blepharospasm, the condition is referred to as Meige's syndrome. The symptoms usually begin between ages 40 and 70 years and appear to be more common in women than men. Difficulty swallowing is a frequent clinical feature of oromandibular dystonia.
- *Spasmodic dysphonia*. Spasmodic dysphonia is characterized by involuntary spasms of the vocal cords causing interruptions of speech and affecting the quality of voice. Two types of spasmodic dysphonia are generally observed. In the adductor type of dysphonia, speaking causes abnormal involuntary excessive contraction of the muscles that bring the vocal cords together. This causes a tight voice pattern,

often with abrupt initiation and termination of the voice, resulting in a broken speech pattern. In the abductor type of dysphonia, increased contraction of the muscles that separate the vocal cords results in a breathy, whispering voice pattern. In addition, three other subtypes of spasmodic dysphonia have been identified. One is a combination of adductor and abductor symptoms in which an individual may demonstrate both types of spasms when speaking. In a second subtype, spasmodic dysphonia symptoms are accompanied by a tremor in the voice. A third subtype involves a primary voice tremor that is so severe the patient experiences adductor voice interruption during the tremor. Although it can begin at any time during life, spasmodic dysphonia seems to occur frequently in 40- to 50-year-olds.

- *Writer's cramp*. This is a task-specific focal dystonia of the hand and the symptoms usually appear when a person is attempting a task that requires fine motor movements. The symptoms may appear only during a particular type of movement, such as writing or playing the piano, but the dystonia may spread to affect many tasks. Two types of writer's cramp have been described: simple and dystonic. People with simple writer's cramp have difficulty with only one specific task. In dystonic writer's cramp, symptoms will be present not only when the person is writing but also when the person is performing other task-specific activities, such as shaving, using eating utensils, and applying makeup. The symptoms usually begin between the ages of 30 and 50 years and affect both men and women.
- *Paroxysmal dystonia*. Paroxysmal dystonia is characterized by brief attacks of dystonic movements and postures and a return to normal posture between the attacks. The paroxysmal dyskinesias are classified as action-induced (kinesigenic) or non-action-induced (nonkinesigenic). The most common paroxysmal dyskinesias are paroxysmal kinesigenic choreoathetosis (PKC) and paroxysmal dystonic nonkinesigenic choreoathetosis (PDC), also known as paroxysmal dystonia. The majority of cases of PKC and PDC are idiopathic or familial (either autosomal dominant or occasionally sporadic).
- *Myoclonic dystonia*. Myoclonic dystonia is characterized by rapid jerklike movements alone or in combination with the sustained muscular contractions and postures of dystonia. Myoclonic movements are the most prominent clinical feature in the majority of people affected. The distribution of these rapid 'jerks' is variable, but most frequently the central part of the body is affected, including the shoulders, arms, neck, and trunk. Rarely are the face and legs affected. The jerks may occur synchronously or asynchronously in different muscle groups. The dystonic symptoms of prolonged muscle spasms or abnormal posture may also be present. The distribution of symptoms of myoclonic dystonia is different from that of typical childhood-onset dystonia; myoclonic dystonia more often affects the upper body, whereas typical childhood-onset dystonia usually affects the legs. The age of onset is in the first or second decade of life, although some cases of adult onset have been reported.
- *Musicians with dystonia*. It has been observed that musicians are prone to task-specific focal dystonia or musician's dystonia. The vigorous practice and constant repetition of precise movements over the course of many years that are required for performing at an expert level can cause task-specific focal dystonia.

Dystonia-plus syndrome To describe dystonias manifesting with other neurological disorders, the classification of dystonia-plus was coined to distinguish some dystonic syndromes from both PTD and secondary disorders. The dystonia-plus group includes dystonic syndromes with neurological features in addition to dystonia, in which clinical and laboratory findings suggest neurochemical disorders but with no evidence of neurodegeneration. Included in this category are disorders characterized by dystonia with parkinsonism (i.e., DYT5 and DYT12) as well as myoclonus–dystonia (DYT11).

- *Rapid-onset dystonia parkinsonism*. Rapid-onset dystonia parkinsonism (RDP), a rare hereditary form of dystonia, is characterized by the abrupt onset of slowness of movement (parkinsonism) and dystonic symptoms. The classic features of RDP include involuntary dystonic spasms in the limbs, prominent involvement of the speech and swallowing muscles, slowness of movement, and poor balance. Onset of the combined dystonic and parkinsonian symptoms can be sudden, occurring over hours to days. Symptoms usually stabilize in less than 4 weeks, after which there is reported to be little progression. RDP usually occurs in adolescence or young adulthood (age range, 15–45 years), but the onset of mild dystonia–parkinsonism has been reported in individuals up to the age of 58 years.
- *X-linked dystonia parkinsonism*. X-linked recessive dystonia parkinsonism (XDP) was first described in 1976 by Lee and colleagues as endemic in Panay, Philippines, affecting predominantly the male population. It is an adult-onset disorder characterized by progressive and severe dystonia often affecting other body regions and progressing to generalized dystonia. Pathology of the XDP corpus striatum demonstrated that the neurons of the matrix compartment are

relatively spared, whereas the striosomes are severely depleted during the dystonic phase of the disorder. However, the matrix compartment also loses neurons as the disease progresses into parkinsonism.

Classification Based on Age of Onset

Based on the age at which the symptoms of dystonia develop, it is classified as early onset and late onset. Early onset dystonia occurs between the ages of 2 and 28 years (median age, 14.5) and the symptoms usually develop mostly in the limbs and spread to the rest of the body and become stabilized within 5 years. Late-onset dystonia usually begins after age 28 years, and the symptoms generally remain focal.

Early onset generalized dystonia Early onset generalized dystonia is a PTD and the most common hereditary form of dystonia; it is characterized by twisting of the limbs, specifically the foot/leg or hand/arm. Early onset generalized dystonia has also been referred to as dystonia musculorum deformans. The following are symptoms of early onset generalized dystonia: twisted postures; turning in of the foot or arm; muscle spasms; unusual walking with bending and twisting of the torso; rapid, sometimes rhythmic, jerking movements; and progression of the symptoms to sustained or fixed postures. The spasms may spread to involve twisting contractions of other parts of the body. Because the legs and trunk are commonly affected in early onset generalized dystonia, abnormal gait is generally noticed in patients. The age of onset varies from 2 to 28 years, but the peak period is between the ages of 7 and 10 years, with symptoms progressing and stabilizing within a 5-year period.

Early onset dystonia is caused by a mutation in the DYT1 gene, which codes for a protein called torsinA. The mutation involves deletion of a single glutamic acid residue at position 302/303 in the amino acid sequence of torsinA. The precise function of torsinA is not known; however, it belongs to the family of AAA+ proteins involved in a variety of cellular functions. Several invertebrate and vertebrate animal models have been developed to better understand the impact of mutation in torsinA on these model animals.

Dopa-responsive dystonia Dopa-responsive dystonia (DRD) is a hereditary form of dystonia characterized by progressive difficulty in walking. The symptoms resemble those of early onset generalized dystonia (DYT1). However, DRD differs from DYT1 in that patients with DRD respond very well to L-DOPA therapy. Although the symptoms of DRD usually present at approximately the age of 7 years, the disorder may appear at any age. Symptoms may appear minor (e.g., muscle cramps after exercise) or present later in life in a form that more closely resembles Parkinson's disease.

Pathophysiology of Dystonia

There is no distinct pathology in dystonia. Studies of DYT1 dystonia demonstrated that the overall immunostaining pattern of torsinA was not different in dystonia and normal brains in the midbrain and neostriatal regions. However, a later study demonstrated the nigral dopaminergic neurons to be larger in both DYT1 and non-DYT1 dystonic brains compared to normal, and they may be more closely spaced in DYT1 brains. Other than these changes in neuronal size and spacing in dystonic brains, there was no indication of neuronal loss, inflammation, or altered distribution of torsin-like immunoreactivity. One study demonstrated brain stem pathology and showed the presence of perinuclear inclusion bodies in the midbrain reticular formation and periaqueductal gray in four clinically documented and genetically confirmed DYT1 patients but not in controls. The inclusions were located within cholinergic and other neurons in the pedunculopontine nucleus, cuneiform nucleus, and griseum centrale mesencephali and stained positively for ubiquitin, torsinA, and the nuclear envelope protein lamin A/C. Similar results were demonstrated in transgenic and knockin models of DYT1 dystonia. Studies of X-linked dystonia parkinsonism have revealed a loss of cells in the striasome compartment of the striatum at the dystonic phase of the disorder, and upon progression to parkinsonism, a loss of cells in the matrix was also observed.

Imaging Studies in Dystonia

No structural abnormalities of the brain have been demonstrated by performing magnetic resonance imaging (MRI) on patients with primary focal dystonia. However, functional neuroimaging with positron emission tomography (PET) revealed differences in DYT1 carriers versus affected individuals. PET, single-photon emission computer tomography, MRI, and magnetoencephalography are powerful tools to elucidate anatomical and functional impairment underlying movement disorders, including dystonia. They have revealed that presymptomatic cerebral abnormalities may be a common feature in dystonia no matter the clinical status. In addition, techniques using specific markers have focused on the type of receptors that may be disinhibited and/or on the kind of neurotransmitters that may be dysregulated in dystonia.

Genetics of Dystonia

In the past two decades, genetic studies of families of affected individuals have led to the identification of several chromosomal loci linked to dystonia. A genetic basis has been implicated in 15 dystonias. However, the gene mutations responsible are known for only a few of them. Table 1 presents the different types of dystonias known so far to have a genetic basis:

- *DYT1*. This designation refers to the gene responsible for most cases of early onset and limb-onset primary dystonia. Also known as Oppenheim dystonia, DYT1 dystonia is an autosomal dominant disorder with reduced penetrance and variable expressivity. The DYT1 gene, which has been mapped to chromosome 9q34.1, encodes for a protein called torsinA. Mutation in torsinA results in the loss of a single glutamic acid residue at position 302/303 in the amino acid sequence of the protein.
- *DYT2*. This genetic designation was reserved for a possible autosomal recessive primary dystonia. However, no such form has been confirmed.
- *DYT3*. This refers to the gene for a neurodegenerative, X-linked recessive disorder that is characterized by dystonia and parkinsonism. Known as X-linked dystonia–parkinsonism or Lubag, the disorder has been reported primarily in young adult Filipino men. The DYT3 gene has been mapped to chromosome Xq13.1.
- *DYT4*. This genetic designation is for an autosomal dominant primary dystonia in a multigenerational Australian family. DYT4 dystonia is primarily characterized by abductor ('whispering') spasmodic dysphonia (laryngeal dystonia).
- *DYT5a,-b*. DYT5a refers to the gene known as GTP cyclohydrolase I (GCH1), which has been mapped to chromosome 14q22.1. Mutations of the GCH1 gene may result in the dystonia-plus syndrome called dopa-responsive dystonia (Segawa syndrome), which is transmitted as an autosomal dominant trait. DYT5b is inherited in an autosomal recessive mode and is due to mutation in the TH gene localized to chromosome 11p15.5. The gene encodes for tyrosine hydroxylase, the enzyme that synthesizes the neurotransmitter dopamine.
- *DYT6*. This is for an autosomal dominant primary dystonia mapped to chromosome 8 (8p21–q22). Described in two Mennonite families (kindreds), this form of dystonia may be of childhood or adult onset; initially it affects limb, cranial, or neck muscles and typically extends to involve multiple body regions.
- *DYT7*. This genetic designation is for an autosomal dominant, adult-onset primary dystonia of the neck region (familial torticollis) that has been mapped to chromosome 18p in a German kindred.
- *DYT8*. This designation is for paroxysmal nonkinesigenic dyskinesia (PNKD) that may be inherited as an autosomal dominant trait. PNKD is characterized by sudden, transient episodes of excessive or disorganized (hyperkinetic or dyskinetic) movements that may occur spontaneously or be triggered or worsened by stress, fatigue, or other factors. Abnormal movements may include any combination of

Table 1 Genetically defined forms of dystonia

Dystonia locus	*Dystonia type*	*Mode of inheritance*	*Chromosomal location*	*Gene product*
DYT1	Early onset generalized dystonia	AD	9q34	TorsinA
DYT2	Generalized dystonia	AR	UNK	UNK
DYT3	X-linked dystonia parkinsonism; 'Lubag'	XR	Xq13.1	UNK
DYT4	Non-DYT1 dystonia	AD	UNK	UNK
DYT5a	Dopa-responsive dystonia	AD	14q22.1	GCH1
DYT5b	Infantile	AR	11p15.5	TH
DYT6	Adolescent-onset dystonia of mixed type	AD	8p21–q22	UNK
DYT7	Adult-onset focal dystonia	AD	18p	UNK
DYT8	Paroxysmal dystonic choreoathetosis	AD	2q33–q35	UNK
DYT9	Paroxysmal choreoathetosis with episodic ataxia and spasticity	AD	1p21	UNK
DYT10	Paroxysmal kinesigenic choreoathetosis	AD	16p11.2–q12.1	UNK
DYT11	Myoclonus–dystonia	AD	7q21, 11q23	ε-Sarcoglyan, DRD2
DYT12	Rapid-onset dystonia–parkinsonism	AD	19q	Na,K-ATPase
DYT13	Cervical/cranial/brachial	AD	1p36.11–p36.32	UNK
DYT14	DOPA-responsive dystonia	AD	14q13	UNK
DYT15	Myoclonus dystonia	AD	18p11	UNK

AD, autosomal dominant; AR, autosomal recessive; DRD2, dopamine receptor 2; GCH1, GTP cyclohydrolase 1; TH, tyrosine hydroxylase; UNK, unknown; XR, X-linked recessive.

dystonia, chorea, athetosis, and/or ballismus. DYT8 has been mapped to chromosome 2q33–q35.

- *DYT9*. This genetic designation is for a family affected by an autosomal dominant disorder known as 'choreoathetosis spasticity, episodic.' It is characterized by sudden, transient (paroxysmal) attacks of dyskinesia – such as chorea, athetosis, limb dystonia, imbalance, and difficulty speaking (dysarthria) – that, in some cases, may be associated with constant weakness and stiffness of the legs (spastic paraplegia). Paroxysmal episodes may be precipitated by physical exercise, lack of sleep, alcohol consumption, stress, or other factors. DYT9 has been mapped to chromosome 1p.
- *DYT10*. This designation is for paroxysmal kinesigenic dyskinesia (PKD), a neurologic movement disorder that may be transmitted as an autosomal dominant trait. In patients with PKD, sudden episodes of dyskinesia, including dystonic movements, may be triggered by sudden voluntary movement or unexpected stimuli (startle). DYT10 has been mapped to chromosome 16p11.2–q12.1.
- *DYT11*. This is an autosomal dominant disorder known as myoclonus–dystonia. Genetic analysis demonstrated mutations of the DRD2 (dopamine receptor 2) gene, which has been mapped to chromosome 11q23. Mutations in the gene for ε-sarcoglycan may also give rise to this disorder.
- *DYT12*. This is characterized by abrupt onset of slow, dystonic movements and prominent bulbar features. The gene responsible for rapid-onset dystonia–parkinsonism, an autosomal dominant disorder, has been mapped to chromosome 19q13. Mutations in the ATP1A3 gene are the cause of this disorder.
- *DYT13*. This is an autosomal dominant disorder with incomplete penetrance (58%). Clinical presentation and age at onset are more variable than in DYT1, and the neck is involved in many affected individuals. The chromosome locus for this condition has been mapped to 1p36.11–p36.32.
- *DYT14*. This is an autosomal dopa-responsive dystonia linked to chromosome 14q13. The substantia nigra and locus ceruleus showed a normal abundance of severely hypomelanized dopaminergic neurons and no Lewy body, similar to that observed in DYT5a.
- *DYT15*. This is associated with a large Canadian kindred in which 12 members over four generations had alcohol-responsive myoclonic dystonia characterized mainly by jerky movements of the upper limbs, hands, and axial muscles. Four members also had dystonia of the upper limbs, and 1 had dystonia of the leg. The chromosomal locus was mapped to 18p11; however, the gene responsible for this syndrome has not been identified.

Therapy

There are no adequate treatments for dystonia. Current treatments are palliative only, guided by the age of onset, cause, and medication history. Early onset dystonia is usually treated with L-DOPA, which helps in the differential diagnosis of DRD versus DYT1; DRD responds very well to L-DOPA, whereas DYT1 does not. However, because of the identification of a causative mutation in DYT1, L-DOPA treatment can be avoided if the patient is first subjected to genetic testing.

Medications

Dopaminergics Primary dystonias respond to drugs such as levodopa that increase dopamine levels; however, some patients respond to agents that block or deplete dopamine, although caution must be taken because the medications may evoke tardive dystonia.

Anticholinergics Drugs that block acetylcholine, thereby deactivating muscle contractions, are among the most commonly used medications for the treatment of dystonia. The dosage is limited by central side effects, such as confusion, drowsiness, hallucination, personality change, and memory loss, and by peripheral side effects, such as dry mouth, blurred vision, urinary retention, and constipation.

Benzodiazepines Benzodiazepines block $GABA_A$ receptors in the central nervous system and may relax muscles and ease symptoms associated with dystonia. The primary side effect is sedation, but others include depression, personality change, and drug addiction. Rapid discontinuation can result in a withdrawal syndrome.

Baclofen Baclofen is used for the treatment of spasticity. It is a GABA agonist that activates the $GABA_B$ receptor. The primary site of action of this drug is the spinal cord, where it reduces the release of neurotransmitters that stimulate muscular activity. Baclofen can be administered orally or via a surgically implanted (intrathecal) pump that delivers the drug directly to the spinal cord.

Botox Treatment

For focal dystonias, intramuscular injection of botulinum toxin substantially reduces the involuntary movements of the muscles, thereby providing substantial relief from dystonic symptoms. Botox is injected into specific muscles to weaken muscle activity sufficiently and reduce spasm but not enough to cause paralysis. Botox prevents the release of acetylcholine, a neurotransmitter that causes muscle contractions. However,

treatment with Botox is not permanent; it needs to be repeated every few months.

Surgical Treatment

Only patients with intractable dystonias that do not respond to any therapies are considered for surgical procedures. Ablation of parts of the thalamus and pallidum was originally done for Parkinson's disease, the rationale being that abnormal firing in these brain regions was considered to result in either Parkinson's disease or dystonia. Therefore, surgical removal of this malfunctioning brain region is believed to correct this abnormal firing. Primary early onset generalized dystonias have been treated successfully with surgical removal of either the pallidum or parts of the thalamus. This is highly invasive surgery and the process cannot be reversed. This procedure is being replaced with deep brain stimulation that is comparatively less invasive than surgical removal of brain regions.

Deep Brain Stimulation

Early onset generalized dystonias that are not treatable using any other available treatments are considered for deep brain stimulation. In this procedure, electrodes are implanted in the globus pallidus using stereotaxic methods. The deep brain stimulation electrodes are connected by wires to a type of pacemaker device (called an impulse generator) implanted under the skin of the chest, below the collarbone. The electrodes are then stimulated at a certain frequency and send electrical pulses into the brain region. Although the mechanism by which deep brain stimulation works is not clear, the procedure has successfully treated many primary generalized dystonias.

Other Therapies

Physiotherapy, yoga, and stretching exercises have been found to be beneficial to dystonic patients. In some patients, the use of sensory tricks has provided temporary relief from their dystonic symptoms. These sensory tricks frequently involve touching or stroking a particular spot on the skin. The exact mechanism of relief obtained by these sensory tricks is not understood. It is believed that stimulation of the skin may interrupt abnormal nerve impulses that cause involuntary contraction of muscles.

Molecular Approaches

In DYT1 dystonia for which the gene defect is known, intervention with siRNA technology has been successful in selectively silencing the production of mutant torsinA message in cell culture models. It is hoped that the development of such technology can lead to prevention or a cure for DYT1 dystonia. However, this technology is not developed enough to be applied to human patients.

See also: Cranial–Cervical Dystonia; Dystonia: Myoclonus–Dystonia; Segawa Dopa Responsive Dystonia; Torsion Dystonia.

Further Reading

Bressman SB (2004) Dystonia genotypes, phenotypes, and classification. *Advances in Neurology* 94: 101–107.

Carbon M, Trost M, Ghilardi MF, and Eidelberg D (2004) Abnormal brain networks in primary torsion dystonia. *Advances in Neurology* 94: 155–161.

de CarvalhoAguiar PM and Ozelius LJ (2002) Classification and genetics of dystonia. *Lancet Neurology* 1(5): 316–325.

Fahn S, Bressman SB, and Marsden CD (1998) Classification of dystonia. *Advances in Neurology* 78: 1–10.

Ford B, Greene P, Louis ED, et al. (1996) Use of intrathecal baclofen in the treatment of patients with dystonia. *Archives of Neurology* 53(12): 1241–1246.

Gonzalez-Alegre P, Bode N, Davidson BL, and Paulson HL (2005) Silencing primary dystonia: Lentiviral-mediated RNA interference therapy for DYT1 dystonia. *Journal of Neuroscience* 25(45): 10502–10509.

Jankovic J (1998) Medical therapy and botulinum toxin in dystonia. *Advances in Neurology* 78: 169–183.

Jinnah HA, Hess EJ, Ledoux MS, Sharma N, Baxter MG, and Delong MR (2005) Rodent models for dystonia research: Characteristics, evaluation, and utility. *Movement Disorders* 20(3): 283–292.

Klein C and Ozelius LJ (2002) Dystonia: Clinical features, genetics, and treatment. *Current Opinion in Neurology* 15(4): 491–497.

Krause M, Fogel W, Kloss M, Rasche D, Volkmann J, and Tronnier V (2004) Pallidal stimulation for dystonia. *Neurosurgery* 55(6): 1361–1370.

Kupsch A, Kuehn A, Klaffke S, et al. (2003) Deep brain stimulation in dystonia. *Journal of Neurology* 250(supplement 1): I47–I52.

Ondo WG, Desaloms JM, Jankovic J, and Grossman RG (1998) Pallidotomy for generalized dystonia. *Movement Disorders* 13(4): 693–698.

Perlmutter JS and Mink JW (2004) Dysfunction of dopaminergic pathways in dystonia. *Advances in Neurology* 94: 163–170.

Vidailhet M, Vercueil L, Houeto JL, et al. (2005) Bilateral deep-brain stimulation of the globus pallidus in primary generalized dystonia. *New England Journal of Medicine* 352(5): 459–467.

Vitek JL and Giroux M (2000) Physiology of hypokinetic and hyperkinetic movement disorders: Model for dyskinesia. *Annals of Neurology* 47(4 supplement 1): S131–S140.

Walker RH and Shashidharan P (2003) Developments in the molecular biology of DYT1 dystonia. *Movement Disorders* 18(10): 1102–1107.

Dystonia: Myoclonus–Dystonia

C Klein, Medical University of Lübeck, Lübeck, Germany
L J Ozelius, Mount Sinai School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Definition, Terminology, and Diagnostic Criteria of Myoclonus–Dystonia

Definition

Myoclonus is a rapid, brief contraction ('fast lightning jerk') of one muscle or a group of muscles which may or may not result in observable movement at a joint. The movements can involve any muscle group but most commonly occur in proximal muscles of the limbs and in the neck and trunk musculature. Dystonia, on the other hand, is characterized by sustained twisting and repetitive movements, resulting in abnormal postures. Myoclonus and dystonia may occur together in several diseases. Some patients with primary torsion dystonia display fast dystonic movements which may resemble myoclonus, and occasional myoclonic jerks may occur superimposed on dystonic movements. Conversely, myoclonus may be associated with dystonic features.

Terminology

A large variety of different terms has been used in the past to describe familial myoclonus or myoclonus in combination with dystonia, including 'paramyoclonus multiplex,' 'hereditary essential myoclonus,' 'familial essential myoclonus,' 'essential familial myoclonus,' 'dominantly inherited myoclonic dystonia with dramatic response to alcohol,' 'hereditary myoclonic dystonia,' 'myoclonus–dystonia (M-D),' and 'inherited myoclonus–dystonia syndrome.' Lang proposed to use a hyphenated combination of the two terms 'myoclonus' and 'dystonia.' The order of presentation was used to emphasize the more constant of the two features – 'dystonia–myoclonus' describes dystonia with some myoclonic jerks, while 'M-D' was suggested as the appropriate term for a predominantly myoclonic syndrome associated with dystonia. Accordingly, the term 'M-D' shall be used herein for all cases of myoclonus associated with dystonia. However, it has to be noted that in several M-D families, affected individuals may display myoclonus or, rarely, dystonia as the sole symptom of the disease. A patient with clinically typical M-D is shown in Figure 1.

Myoclonus or dystonia is considered primary if either is the only neurological feature, and the etiology is unknown or genetic. M-D is often referred to as a 'dystonia-plus' syndrome. Owing to the advances in the molecular genetics of the dystonias in the past 10 years, a genetically based categorization has been introduced. Currently, at least 15 different types of dystonias can be distinguished on a genetic basis, designated dystonia *DYT1–DYT15*. M-D linked to chromosome 7q is synonymous with dystonia 11 (or *DYT11*, OMIM 159900 – now *epsilon-sarcoglycan*, or *SCGE*, OMIM 604149), and a second M-D locus has been described on chromosome 18p (*DYT15*, OMIM 607488). Finally, based on the identification of the *epsilon-sarcoglycan* gene as the major M-D gene, M-D may be grouped under the sarcoglycanopathies. (Note: The Online Mendelian Inheritance in Man (OMIM) database is a catalog of human genes and genetic disorders from the National Center for Biotechnology Information (NCBI).)

Diagnostic Criteria

Although several similar descriptions had appeared in the literature before 1967, Mahloudji and Pikielny were the first to define diagnostic criteria for M-D, which they referred to as 'hereditary essential myoclonus.' Original diagnostic criteria were (1) onset of myoclonus in the first or second decade, (2) males and females equally affected, (3) a benign course, often variable but compatible with an active life of normal span, (4) dominant mode of inheritance with variable severity, (5) absence of seizures, dementia, gross ataxia, and other neurological deficits, and (6) normal electroencephalogram (EEG). These criteria have been modified by Gasser and more recently by Klein based on findings in genetically proved (*DYT11*) M-D and are subdivided into cardinal and optional diagnostic criteria for M-D. Since members from at least three mutation-positive families have been found to have epileptic seizures and/or EEG changes, absence of seizures and an abnormal EEG are no longer considered exclusion criteria for M-D. Adapted diagnostic criteria are as follows.

- Cardinal criteria
  1. Onset of myoclonus usually in the first or second decade of life; dystonic features are observed in more than half of the affected patients in addition to myoclonus and may rarely be the only manifestation of the disorder.
  2. Males and females about equally affected.
  3. A relatively benign course, often variable, but compatible with an active life of normal span in most cases.

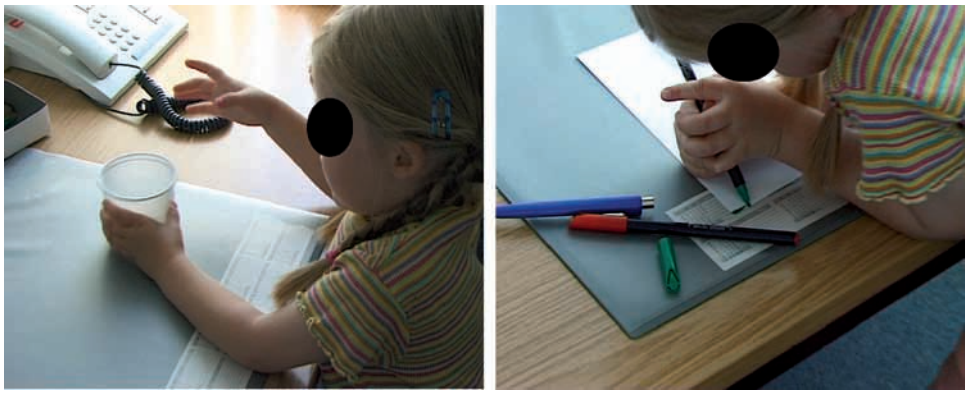

Figure 1 Four-year-old girl with clinically typical M-D and a proved *SGCE* mutation. (Left) Dystonic posturing of the right hand when attempting to lift a cup with the left hand. (Right) Both hands are needed to hold a pen when drawing, which provokes writer's cramp and myoclonus of the upper body half.

4. Autosomal dominant mode of inheritance with variable severity, and incomplete penetrance which is dependent on the parental origin of the disease allele; affected individuals usually inherit the disease from their father.
5. Absence of dementia, gross ataxia, and other neurological deficits.
6. Normal somatosensory evoked potential (SSEP); normal results of computed tomography (CT) or magnetic resonance imaging (MRI) neuroimaging studies.

- Optional criteria
  1. Response of symptoms (particularly of the myoclonus and to a lesser degree of the dystonia) to alcohol.
  2. Various personality disorders and psychiatric disturbances.

Clinical Features of M-D

The myoclonic jerks typical of M-D are brief, lightninglike movements, usually beginning in and most often affecting the neck, trunk, and upper limbs, with legs involved less prominently. Less common distributions of myoclonus include laryngeal, distal upper limb, and truncal. Approximately half of the affected individuals have focal or segmental dystonia that presents as cervical dystonia and/or writer's cramp. In contrast to primary torsion dystonia, involvement of lower limbs is rare and usually does not occur at onset. In addition, the dystonia does not tend to generalize in the course of the disease. Rarely, dystonia is the only manifestation and sometimes develops years after the onset of myoclonus. Age of onset is usually in childhood or adolescence. In many cases, myoclonus is the presenting symptom. For example, in a study of 23 mutation carriers, the average age of onset for myoclonus was 5.4 years, whereas the mean age of onset of dystonia was 8.8 years.

The involuntary movements are frequently precipitated or worsened by active movements of the affected body part. Other factors eliciting or enhancing the hyperkinetic movements include certain postures, emotional stress, anxiety, nervousness, fright, excitement, fatigue, sound, startle, touch, and caffeine. In female patients, the involuntary movements have been described to worsen premenstrually. The myoclonus and dystonia disappear during sleep. However, single violent jerks may occur at night and may be severe enough to wake up the patient, a phenomenon previously described by Friedreich. Response of the myoclonus and, to a lesser extent, of the dystonia to alcohol is a very frequent finding but not invariably present. Some patients experience a rebound phenomenon after alcohol withdrawal. By contrast, M-D is relatively unresponsive to medical treatment.

The course of the disease is often relatively benign. An initial worsening of symptoms is almost always followed by stabilization of the condition in adulthood, with most patients remaining completely independent in the activities of daily living. The disease is usually compatible with an active life and normal life span in most patients and even spontaneous remission of M-D has been reported in selected cases. In some patients, however, M-D may be gradually progressive and may lead to considerable functional disability and result in early retirement.

Psychiatric disorders are a frequent finding in M-D patients. Many reports can be found of alcohol abuse and addiction to alcohol, as well as of personality disorders and psychiatric abnormalities. Psychiatric disorders include obsessive–compulsive disorder, major affective and anxiety disorder, and panic attacks.

Although psychiatric problems are probably part of the M-D phenotypic spectrum, they are not invariably present (and often not specifically assessed).

Unusual features have been described in selected M-D families with identified mutations in the *SGCE* gene and include postural and other forms of tremor and mild mental retardation. Epileptic seizures have been documented in some patients or may occur due to alcohol withdrawal.

Clinical Features of M-D Patients with Documented *SGCE* Mutations

Most descriptions of patients with proved *SGCE* mutations include details on the mean age of onset, the presence and distribution of myoclonus and dystonia, and the response to alcohol (Table 1). Although most cases are familial, information on affected family members is not systematically given. Conversely, some reports describing larger families do not include details on individual family members but rather summarize mean ages of onset, and percentages of certain M-D signs or clinical features cannot be unequivocally linked to the pedigrees and mutations. Therefore, the data presented here are from mutation-positive index patients, based on data published between 2002 and 2006, with detailed clinical information.

A total of 44 patients had an average age of onset of 6.9 ± 6.5 years with a range of 0.3–30 years. Information on gender was provided for 42 of the 44 index patients, 14 (33.3%) of whom were male and 28 (66.7%) female. All 44 index patients (100%) had myoclonus, whereas dystonia was present in only 36 (81.8%). Both myoclonus and dystonia affected mainly, and sometimes exclusively, the upper body half, with predominant involvement of the neck and upper limbs. Additional lower limb involvement was described in almost half of the index patients on whom information was available (20/42; 47.6%). A response to alcohol was positive in 21 of the 44 index patients (47.7%), negative in three of 44 (6.8%), and unknown in the remaining 20 patients (45.5%). It should be kept in mind, however, that these frequencies are probably biased by the fact that only index patients have been included here, resulting in likely oversampling of clinically typical cases.

A previous review (in 2003) by the authors included available information on relatives and revealed that 84 of 92 (91.3%) affected had myoclonus, and only 52 of 92 (56.5%) showed dystonia. Thus, in 40 of the 92 (43.5%) affected individuals, myoclonus was the only movement disorder, while in eight of 92 (8.7%), dystonia was the sole sign of M-D.

Differential Diagnosis

The clinical picture with myoclonus, dystonia, or both, which often occurs in a familial fashion, along with the absence of any other neurological signs or symptoms, is quite distinct. In most cases, the diagnosis of M-D can be established on clinical grounds with a relatively high degree of diagnostic certainty.

Most other conditions in which myoclonus is a prominent feature are characterized by a variety of

Table 1 Clinical features of M-D patients with documented *SGCE* mutations

Reference	*No. of mutation-positive index patients*	*Mean (range) onset age*	*Male/female ratio*	*Myoclonus*	*Dystonia*	*Upper/lower body*	*Response to alcohol (+/?/−)*
Klein et al. (2002) (Dohney et al. (2002))	2	10 and 6 years	n.r.	$n=2$	$n=1$	n.r.	$n=0/2/0$
Muller et al. (2002) (Kock et al. (2004))	1	2 years	$n=0/1$	$n=1$	$n=1$	$n=1/0$	$n=0/1/0$
Marechal et al. (2003)	1	2 years	$n=0/1$	$n=1$	$n=1$	$n=1/1$	$n=0/1/0$
Hjermind et al. (2003)	1	4 years	$n=1/0$	$n=1$	$n=1$	$n=1/1$	$n=1/0/0$
Foncke et al. (2003)	1	17 years	$n=1/0$	$n=1$	$n=1$	$n=1/0$	$n=0/1/0$
DeBerardinis et al. (2003)	1	1 year	$n=1/0$	$n=1$	$n=1$	$n=1/0$	$n=0/1/0$
Hedrich et al. (2004)	2	7 and 6 years	$n=0/2$	$n=2$	$n=2$	$n=2/0$	$n=2/0/0$
Schüle et al. (2004)	2	9 and 7 years	$n=0/2$	$n=2$	$n=2$	$n=2/0$	$n=1/1/0$
O'Riordan et al. (2004)	1	3 years	$n=0/1$	$n=1$	$n=1$	$n=1/1$	$n=0/1/0$
Cif et al. (2004)	1	1 year	$n=1/0$	$n=1$	$n=1$	$n=1/0$	$n=0/1/0$
Valente et al. (2005)	6	8.2 (0.3–3.0) years	$n=1/5$	$n=6$	$n=5$	$n=6/3$	$n=4/1/1$
Asmus et al. (2005)	2	1 and 14 years	$n=0/2$	$n=2$	$n=2$	$n=2/0$	$n=2/0/0$
du Montcel et al. (2006)	16	5.8 (1.5–23) years	$n=5/11$	$n=165$	$n=11$	$n=14/7$	$n=6/9/1$
Gerrits et al. (2006)	7	10.0 (1–17) years	$n=4/3$	$n=7$	$n=7$	$n=7/5$	$n=5/1/1$

n.r., not reported.

other neurologic signs and symptoms not compatible with a diagnosis of M-D (e.g., ataxia). Genetically defined conditions with myoclonus as a major component include progressive myoclonus epilepsy, also known as Baltic myoclonus, caused by mutations in the *cystatin B* gene (formerly *EPM1*; now *CSTB*); myoclonus epilepsy of the Lafora type, associated with mutations in the *EPM2A* gene; and myoclonus epilepsy associated with ragged-red fibers (MERRF), due to mutations in mitochondrial genes. Finally, patients with the trinucleotide repeat disorder dentato rubro pallido luyisian atrophy (DRPLA) may also present with myoclonus.

Genetic Findings in M-D

M-D is inherited in an autosomal dominant fashion. However, it should be noted that, for two main reasons, 'pseudo'-sporadic cases may occur despite the presence of a mutation in the *SGCE* gene. First, *de novo* mutations have been described, and second, *SGCE* mutations show reduced penetrance, meaning that some family members do not develop the disease although they are mutation carriers. M-D due to *SGCE* mutations is one of the rare examples in which a molecular mechanism has been discovered that explains the majority of reduced penetrance.

Reduced Penetrance Due to Imprinting of the *SGCE* Gene

Zimprich et al. were the first to point out that reduced penetrance was found particularly in the offspring of affected females, suggestive of possible maternal imprinting (Figure 2). Parental imprinting is defined as the epigenetic marking of certain subregions of the parental genome in mammals – that is, imprinted genes are expressed from one allele only, according to their parent of origin. Analysis of the distribution of parental alleles in eight informative families revealed that almost 90% inherited the mutated allele from their father, whereas about 90% of the unaffected carrier family members inherited the mutated allele from their mother. Interestingly, an imprinting mechanism has also been described for the orthologous *SGCE* mouse gene, which is predominantly expressed from its paternal allele in all adult mouse tissues, with the exception of weak expression from the maternal allele in mouse brain. Subsequently, paternal transmission of the human *SGCE* mutant allele was demonstrated in affected individuals, as well as DNA methylation differences, supporting maternal imprinting. However, the reasons for reversal of the maternal imprint in a minority (about 10%) of cases remain unknown.

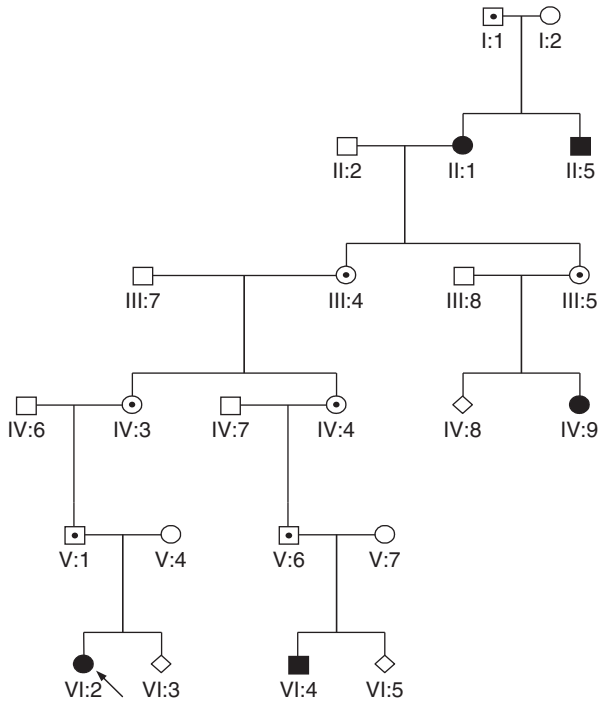

Figure 2 Pedigree demonstrating reduced penetrance of *SGCE* mutations due to paternal transmission of the mutations. This may result in transmission of the mutation through several generations of unaffected individuals (generations III–V in this example). Of note, there is one case of maternal inheritance of the disease (Patient IV:9), as occurs in about 10% of the described M-D patients with *SGCE* mutations. The reason for these exceptions from paternal transmission remains unclear. Squares represent males, circles female, and diamonds individuals of unspecified sex.

Mutational Screening of the *SGCE* Gene

Testing to date has identified mutations in the *epsilon-sarcoglycan* gene at the *DYT11* locus in many individuals with familial M-D.

Frequency of *SGCE* mutations in tested patient samples A comprehensive review of 22 articles on results of screening for *SGCE* mutations reveals a total of 558 index patients and 233 affected family members with a diagnosis of M-D, other forms of dystonia, or other movement disorders. Fourteen of the 22 publications were case reports describing 30 M-D index patients and their affected family members ($n = 113$). These index cases were all mutation-positive, and 25 of them had a positive family history (83%). Therefore, the majority of these cases were preselected for positive family history and/or had previously been used in linkage studies to the *DYT11* locus on chromosome 7 and, thus, were biased toward finding *SGCE* mutations. Three out of 15 available pedigrees showed exceptions from paternal transmission (Table 2). In the eight remaining studies, 528 index patients with various

movement disorders were genetically investigated. This included 149 patients with diagnosed M-D (28.2%), 335 with different forms of dystonia (63.5%), 32 with Tourette's syndrome (6.1%), eight with Ramsay–Hunt syndrome (1.5%), and four with benign hereditary chorea (0.8%). Among these were 34 apparently unrelated mutation carriers, 23 of whom had a positive family history (67.6%). An *SGCE* mutation was found in 22.8% (34/149) of the M-D index patients. In contrast, no mutations were detected in patients affected with other forms of dystonia or other movement disorders (Table 3). For the nonmutation carriers, a positive family history was reported in 25.3% (117/462) of the cases. Exceptions from an inheritance pattern characteristic of maternal imprinting appeared in 11.8% (2/17) of the described mutation-positive families and in 36.1% (22/61) of the mutation-negative families.

Types and localization of *SGCE* mutations More than 40 different *SGCE* mutations have been reported to date (Table 4 and Figure 3; http://www.dmd.nl). Thirty-two of these are exonic, including three exon rearrangements (c.exon5del, c.exon6del, and c.exon1_exon12del), 14 single base pair substitutions, and 15 small deletions or insertions of one or several base pairs. Additionally, nine splice-site mutations were found. Nine of all known mutations are recurrent. A majority of mutations result in premature protein truncation before or within the *SGCE* transmembrane domain. Of note, the recent identification of exonic deletions in four families with M-D but without otherwise identifiable *SGCE* mutations suggests that some percentage of individuals with M-D and as yet undetected mutations may have this type of mutation.

Sarcoglycan Genes and SGCE Protein

SGCE, the fifth sarcoglycan gene to be identified, consists of 12 exons and is highly homologous to the gene encoding α-sarcoglycan. Both gene sequences predict a hydrophobic signal sequence followed by an extracellular domain with a conserved consensus site for asparagine-linked glycosylation, and four highly conserved cysteine residues. The overall gene structure is conserved, which implies that both genes may have arisen through a gene duplication event. However, the introns in the gene encoding α-sarcoglycan are smaller than those in *SGCE*, which is estimated to span 70.9 kb. *SGCE* has a hydrophobic transmembrane domain and a short cytoplasmic domain. An alternative cytoplasmic domain, which was seen in only one of four isolated cDNAs, resulted from the inclusion of exon 10. Exon 10 is of particular note, as it is homologous to the Alu class of repetitive DNA sequences. The murine gene for ε-sarcoglycan maps to the syntenic region on mouse chromosome 6. Other alternative splice variants in mouse brain that affect the C-terminal end of the encoded protein have been identified and at least one of these (exon 11a) is conserved in the human sequence. Finally, a processed pseudogene, *pseudo-ε-sarcoglycan*, is located on human chromosome 2. *SGCE* is widely expressed in both embryonic and adult tissues, including smooth muscle, Schwann cell membranes of peripheral nerves, and brain, and is widely expressed in various brain regions both during development and adulthood.

The SGCE protein consists of 405 amino acids and has a predicted molecular mass of 47 kDa (without exon 10 and posttranslational modifications). The function of the SGCE protein is still largely elusive. The other members of the sarcoglycan family, α-, β-, γ-, and δ-sarcoglycan, compose the heterotetrameric sarcoglycan complex which is an important component of the dystrophin-associated glycoprotein assembly in striated muscles. This complex and its associated components are thought to link the cytoskeleton with the extracellular basement membrane, thus stabilizing the muscle and protecting it from contraction-induced damage. SGCE appears to be functionally similar to α-sarcoglycan in skeletal muscle; however, its function in the brain is presently unknown. It has been speculated that, because *SGCE* is maternally imprinted and because the vast majority of affected individuals inherit their disease gene from their fathers, the disease is caused by loss of function of SGCE protein. However, up to 5% of cases inherit their mutated allele from their mothers and presumably also express the wild-type allele from their fathers. Further, co-expression of both alleles has been documented in one patient. Therefore, the mechanism of disease is not entirely clear.

Genetic Heterogeneity of M-D

Mutations in two other genes have been associated with M-D: it remains unclear if a missense mutation in the *dopamine receptor D2* (*DRD2*) gene found in a single family (in combination with an *SGCE* mutation) is disease-causing or a rare polymorphism. An 18 bp deletion in the *DYT1* gene that is usually associated with early-onset primary torsion dystonia was found in another single family. However, mutations in *SGCE* were subsequently identified also in this family and the significance of this combination of mutations is unknown.

Many simplex and familial cases without identifiable *SGCE* mutations have been reported, suggesting locus heterogeneity. A large Canadian family with clinically similar M-D does not have a mutation in the *SGCE* gene but rather shows linkage to markers

Table 2 Case reports of M-D patients with documented *SGCE* mutations

No.	*Case report reference (original clinical description)*	*No. of patients, all M-D: index (add. aff. FMs)*	*No. of mut. carriers: index (add. aff. FMs avail. for testing)*	*Pos. family history of mut. carriers*	*No. of mut.-pos. families w/exceptions from imprint (family code)*	*Atypical features in mut. carriers (no. of cases)*	*Psychiatrics (no. of cases)*	*Comments*
1/2	Zimprich et al. (2001); Grabowski et al. (2003); Scheidtmann et al. (2000); Asmus et al. (2001)	6 (31)	6 (22)	6/6	1/6 (MD7); aff. daughter inherited mut. from mother	None	Alcohol abuse (3), panic attacks (1)	All families previously known to be linked to the *SCGE* region
3	Asmus et al. (2002)	9 (15)	9 (15)	7/9	0/1	None	Alcohol abuse (2), panic attacks (2)	In three families, linkage previously known
4	Klein et al. (2002) (Dohney et al. (2002))	2 (10)	2 (9)	2/2	n/a; gender disguised	None	Depression (7), anxiety (5), panic attacks (1), OCD (3), alcohol/drug abuse (2), attention deficit disorder (1), personality disorder (1)	Non-mut. carrier had possible dystonia only; several members of both families carry add. mut. in other genes (*DRD2* or *DYT1*)
5	Dohney et al. (2002)	2 (14)	2 (12)	2/2	n/a; gender disguised	None	Alcohol/drug abuse (2), depression (10), schizophrenia (2), anxiety (3), panic attacks (3), phobic disorder (2), OCD (3), nonspecific psychotic disorder (1)	Family F3 included in Foncke et al. (2003)
6	Muller et al. (2002) (Kock et al. (2004))	1 (0)	1 (0)	0/1	n/a	None	n.r.	First evidence of maternal imprinting of *SCGE*; family V in Schüle et al. (2004)
7	Marechal et al. (2002)	1 (5)	1 (3)	1/1	0/1	No response to alcohol (3)	OCD, depression	None

8	Hjermind et al. (2003)	1 (8)	1 (8)	1/1	0/1	Positive response to L-dopa (1); prominent lower limb involvement	n/a	n/a
9	Foncke et al. (2003)	1 (9)	1 (3)	1/1	1/1 (F3); aff. son inherited disease from mother	Epilepsy and EEG changes	Depression	None
10	DeBerardinis et al. (2003)	1 (0)	1 (0)	0/1	n/a	Microcephaly, short stature, dysmorphic face, language delay	n.r.	Interstitial deletion aff. chromosome 7q21
11	Hedrich et al. (2004)	2 (4)	2 (2)	1/2	0/1	None	n.r.	Report of *de novo* mut.; families A and C included in Gerrits et al. (2006)
12	O'Riordan et al. (2004)	1 (3)	1 (3)	1/1	0/1	Epilepsy (3)	n.r.	Two mut. carriers w/o any myoclonus
13	Cif et al. (2004)	1 (7)	1 (0)	1/1	1/1; aff. son inherited disease from mother	None	None	Treatment of an M-D patient w/deep brain stimulation
14	Asmus et al. (2005)	2 (7)	2 (5)	2/2	0/2	None	None	First report of a single exonic deletion in *SCGE*
	Sum of patients	30 (113)	30 (82)	25/30	3/15			

add., additional; aff., affected; EEG, electroencephalogram; FMs, family members; mut., mutation; n/a, not/available; n.r., not reported; OCD, obsessive–compulsive disorder.

Table 3 Larger studies screening for *SGCE* mutations

No.	*Study reference*	*No. of patients: index (add. aff. FMs)*	*No. of M-D patients: index (add. aff. FMs)*	*No. of mut. carriers: index (add. aff. FMs avail for testing)*	*Pos. family history of mut. carriers*	*Pos. family history of non-mut. carriers*	*No. of mut.-pos. families with exceptions from imprint (family code)*	*No. of mut.-neg. families with exceptions from imprint (family code)*	*Other forms of movement disorders (no. of cases)*	*Atypical features in mut. carriers (no. of cases)*	*Atypical features in non-mut. carriers (no. of cases)*	*Psychiatrics (no. of cases)*	*Comments*
1	Han et al. (2003)	7 (37)	7 (37)	3 (1)	3/3	3/4	1/3 (family 3); unaff. son inherited mut. from father	1/3 (family 1); disease inherited through mother	n/a	None	3/4 non-mut. carriers had no dystonia	n.r.	Families 4 and 6 are mixed up
2	Valente et al. (2003)	16 (15)	16 (15)	0 (0)	n/a	6/16	n/a	2/3 (families 5 and 16); disease inherited through mothers	n/a	n/a	No response to alcohol (1); age of onset >20 yrs (4); unusual phenotype (1); prominent lower limb involvement	Depression (2)	None
3	Schüle et al. (2004)	10 (19)	10 (19)	2 (3)	2/2	5/8	1/2 (family 1); unaff. son inherited mut. from father	2/5 (families 4 and 5); disease inherited through mothers	n/a	Prominent lower limb involvement	n/a	n.r.	None
4	de Carvalho Aguiar et al. (2004)	32 (0)	0 (0)	0 (0)	n/a	n.r.	n/a	n/a	Tourette's syndrome	n/a	Non-M-D phenotype	OCD	None

5	Grundmann et al. (2004)	298 (0)	2 (0)	0 (0)	n/a	0/2 (M-D); 44/296 (other types of dystonia)	n/a	n/a	Generalized dystonia (13); multifocal dystonia (11); segmental dystonia (56); focal dystonia (216)	n/a	Non-M-D phenotypes (296/298)	Unspecified psychiatric features (26)	None
6	Valente et al. (2005)	58 (n.r)	29 (n.r)	6 (n.r)	4/6	27/52	0/4	6/27	'Jerky' dystonia (18); Ramsay–Hunt syndrome (8); benign hereditary chorea (3)	No response to alcohol (1)	M-D phenotype (23/58): no response to alcohol (3/23), prominent lower limb involvement (2/23); non-M-D phenotypes (29/58)	n.r.	None
7	du Montcel et al. (2006)	76 (16)	54 (16)	16 (16)	8/16	23/60	0/8	11/23; disease inherited through mothers	Dystonia w/ tremor (5); D-M (13); generalized dystonia (3); benign hereditary chorea (1)	No response to alcohol (3)	Prominent lower limb involvement	None	DYT1 mut. found in 1 patient w/myoclonic dystonia w/excellent response to alcohol
8	Gerrits et al. (2006)	31 (33)	31 (33)	7 (33)	6/7	9/24	n.r.	n.r.	n/a	No response to alcohol (1)	n/a	Unspecified psychiatric features	None
	Sum of patients	528 (120)	149 (120)	34 (53)	23/34	117/462	2/17	22/61	379				

add., additional; aff., affected; D-M, dystonia-myoclonus; FMs, family members; mut., mutation; n/a, not/available; n.r., not reported; OCD, obsessive–compulsive disorder.

Table 4 Known mutations in *SGCE*

No.	*Local*	*Mutation*	*Protein*	*Origin*	*No. of reports*	*Reference*
1	1–12	exon1_exon12del	None	Caucasian; Caucasian	2	DeBerardinis et al. (2003); unpublished data
2	Intr.1	109+1:IVS1+1G>A		British	1	Dagvadorj et al. (2002)
3	Ex.2	Exon2del		German	1	Unpublished data
4	Ex.2	164delG	Gly55/ Stop85	British	1	Hedrich et al. (2004)
5	Ex.2	179A>G	His60Arg	German, Serbian	1	Schüle et al. (2004)
6	Ex.2	179A>C	His60Pro	Dutch; Dutch	2	Hedrich et al. (2004); Gerrits et al. (2006)
7	Ex.2	208G>T	Glu70Stop	French, Caucasian	1	du Montcel et al. (2006)
8	Ex.2	221delA	Y74fsX85	French, Caucasian	1	du Montcel et al. (2006)
9	Ex.2	232G>A	Gly78Ser	Bolivian	1	Dagvadorj et al. (2002)
10	Intr.2	233–1:IVS2–1G>A		French; Caucasian	2	Asmus et al. (2002); du Montcel et al. (2006)
11	Intr.2	232+2:IVS2+2T>C		French, Caucasian	1	du Montcel et al. (2006)
12	Intr.2	232+1:IVS2+1G>A		French, Caucasian	1	du Montcel et al. (2006)
13	Ex.3	274T>C	M92T	French, Caucasian	1	du Montcel et al. (2006)
14	Ex.3	276delG	Met92/ Stop131	German	1	Asmus et al. (2002)
15	Ex.3	289C>T	Arg97Stop	German; German-Italian; Irish; German	6	Zimprich et al. (2001); Dagvadorj et al. (2002); O'Riordan et al. (2004); Valente et al. (2005); unpublished data
16	Ex.3	300G>A	W100X	French, Caucasian	3	du Montcel et al. (2006)
17	Ex.3	304 C>T	Arg102Stop	German; Dutch; French; Arab	12	Zimprich et al. (2001); Asmus et al. (2002); Han et al. (2003); Hedrich et al. (2004); Gerrits et al. (2006); unpublished data
18	Ex.3	344A>G	Y115C	French, Caucasian	1	du Montcel et al. (2006)
19	Intr.3	391–3:IVS3–3T>C		Caucasian	1	Valente et al. (2005)
20	Ex.4	444_447delTAAT	N149fsX149	French, Caucasian	2	du Montcel et al. (2006)
21	Ex.4	484_487del (neu:416_419del)		French	1	Marechal et al. (2003)
22	Ex.4	488_497del (neu:391_406del)		German	1	Zimprich et al. (2001)
23	Intr.4	463+6:IVS4+6T>C		British	1	Asmus et al. (2002)
24	Ex.5	564_576del	fs192X		1	Valente et al. (2005)
25	Ex.5	565delA	Asn189/ Stop196	German	1	Zimprich et al. (2001)
26	Ex.5	587T>G	Leu196Arg	German	1	Klein et al. (2002)
27	Ex.5	625insG	Gly209/ Stop216	German	1	Müller et al. (2002)
28	Ex.5	Exon5del		German	1	Asmus et al. (2005)
29	Ex.5	619_620delAG	R207fsX215	Dutch	1	Gerrits et al. (2006)
30	Intr.5	662+5:IVS5+5G>A		German	1	Asmus et al. (2002)
31	Ex.6	Exon6del		Danish	1	Asmus et al. (2005)
32	Ex.6	709C>T	R237X	German	1	Unpublished data
33	Ex.6	733_737delAATT	Gln245/ Stop254	French	1	Asmus et al. (2002)
34	Ex.6	771_772delAT	T257fsX258	Caucasian	1	Unpublished data

Continued

Table 4 Continued

No.	*Local*	*Mutation*	*Protein*	*Origin*	*No. of reports*	*Reference*
35	Ex.6	745_746insTGTA	S249fsX250	French, Caucasian	1	du Montcel et al. (2006)
36	Ex.6	808T>C	W270R	Dutch	1	Gerrits et al. (2006)
37	Ex.6	812G>A	C271Y	French, Caucasian	1	du Montcel et al. (2006)
38	Intr.6	907+1: IVS6+1G>A (neu:825+1: IVS6+1G>A)		German	1	Zimprich et al. (2001)
39	Ex.7	832_836delAAAAC	Lys278/ fs>295X	French; Caucasian	3	Han et al. (2003); Klein et al. (2002); du Montcel et al. (2006)
40	Ex.7	885insT	L295fsX297	Dutch; Dutch	2	Foncke et al. (2003); Gerrits et al. (2006)
41	Ex.7	856C>T	Gln286Stop	French; Dutch	2	Asmus et al. (2002); Gerrits et al. (2006)
42	Ex.7	974delC	Arg325X	Caucasian	1	Hjermind et al. (2003)
43	Ex.7	966delT	Ala322/ Stop333	German; Serbian	2	Müller et al. (2002); Schüle et al. (2004)
44	Intr.7	1037+2:IVS7+2T>C		Dutch	1	Gerrits et al. (2006)
45	Intr.7	1037+5:IVS7+5G>A		German	1	Asmus et al. (2002)
46	Ex.9	1114C>T	R372X	French; Caucasian; German	4	Valente et al. (200); du Montcel et al. (2006); unpublished data

References and mutations can be found at http://www.dmd.nl.

on chromosome 18p (locus *DYT15*). Two other families also show possible linkage to this chromosomal region. However, the overall contribution of this locus to M-D cannot be determined until the gene is identified.

Phenotype–Genotype Correlations

An association between the localization or type of the reported mutations and the described phenotype of the patients is not obvious. This is not surprising since almost all *SGCE* mutations are nonsense mutations causing premature protein truncation upstream of the transmembrane domain. Therefore, the mutated proteins – regardless of the exact location of the mutation – will not be able to localize to the membrane and it is thus unable to execute its function. Since only the paternal *SGCE* allele is transcribed due to maternal imprinting, SGCE mutations likely act through a loss-of-function mechanism, though, as described earlier, this is not entirely clear-cut.

Rare evidence for an association of genotype and phenotype has been reported in a patient with a complex clinical picture comprising M-D and other features. Due to a large deletion on chromosome 7 (q21.2-q21.3), this patient was affected with microcephaly, short stature, dysmorphic face, and language delay in addition to the myoclonus and dystonia. We recently observed a similar patient with a 7q21.3-q22.1 deletion, encompassing the entire *SGCE* gene as well as the adjacent *COL1A2* (*collagen, type I, alpha 2*) gene. Interestingly, several mutation carriers reported skin and joint problems that were likely explained by haploinsufficiency of the COL1A2 protein. Although such patients are probably rare, these findings stress the importance of a careful and comprehensive clinical examination with special emphasis on potential additional (nonmotor) signs. This will facilitate genetic testing and, in turn, help in providing the optimal counseling and treatment to patients.

Animal Models

Animal models of myoclonus (but not of M-D) have long been available but studies have mostly focused on posthypoxic myoclonus and other secondary forms of myoclonus. Recently an *sgce* knockout mouse model was created and heterozygous mice with a paternally inherited knockout allele were examined for behavioral and neurochemical abnormalities. These mice exhibited myoclonus, deficits in motor coordination, hyperactivity, and psychiatric alterations, including anxiety. In addition, significantly higher levels of striatal dopamine and its metabolites, as well as an inverse correlation between the dopamine and serotonin metabolites, were found. These results suggest that a single *SGCE* mutation can lead to alterations of dopaminergic and serotonergic systems resulting in the phenotypic spectrum seen in M-D patients.

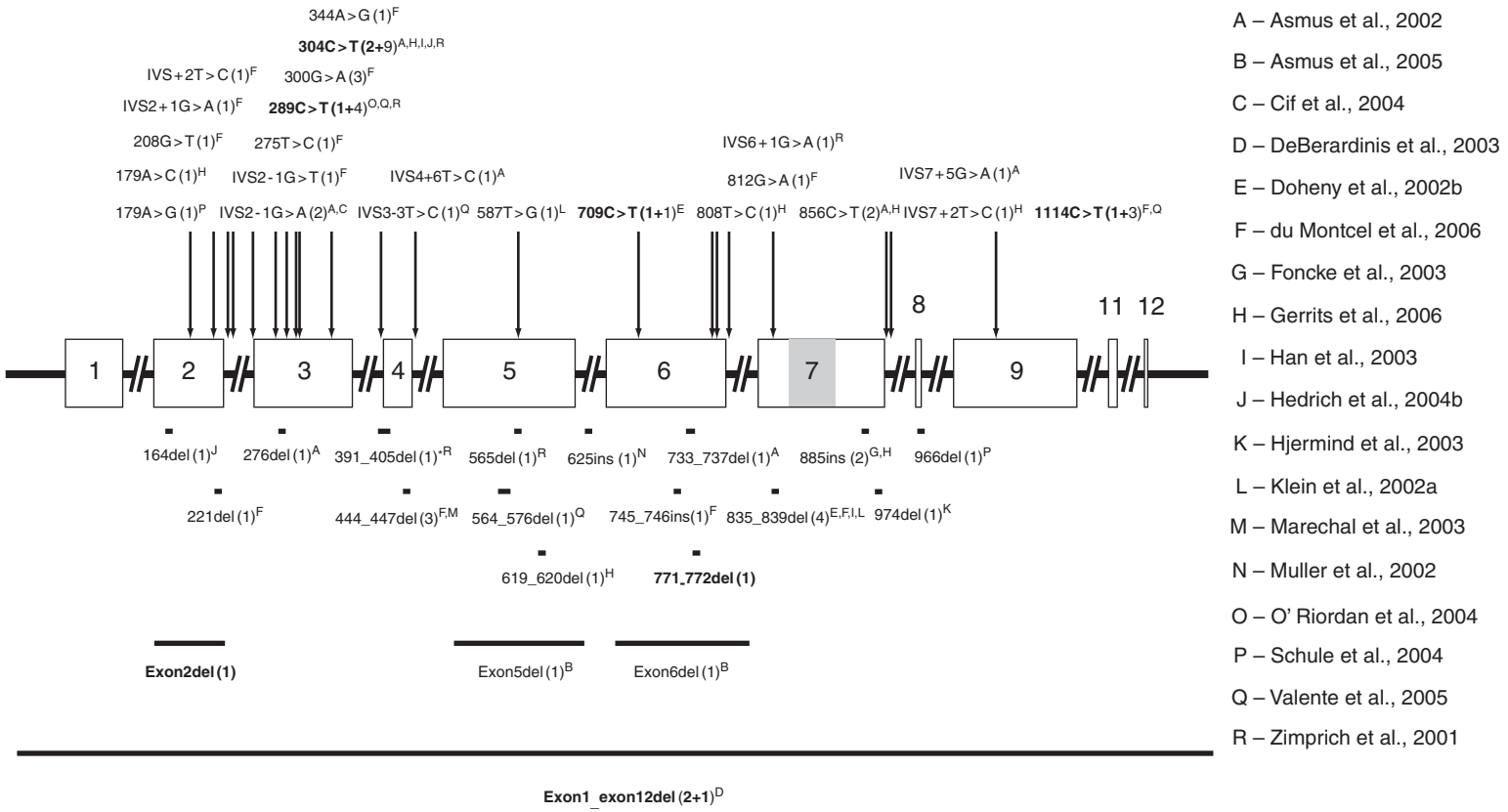

A – Asmus et al., 2002
B – Asmus et al., 2005
C – Cif et al., 2004
D – DeBerardinis et al., 2003
E – Doheny et al., 2002b
F – du Montcel et al., 2006
G – Foncke et al., 2003
H – Gerrits et al., 2006
I – Han et al., 2003
J – Hedrich et al., 2004b
K – Hjermind et al., 2003
L – Klein et al., 2002a
M – Marechal et al., 2003
N – Muller et al., 2002
O – O' Riordan et al., 2004
P – Schule et al., 2004
Q – Valente et al., 2005
R – Zimprich et al., 2001

Figure 3 Schematic representation of the *SGCE* gene and identified mutations (for detailed references, see http://www.dmd.nl). Point mutations are indicated above the gene, and small deletions and exon rearrangements are shown below the gene. Mutations detected in the current study are emphasized by bold characters; frequency of the mutation is given in parentheses; respective references are indicated on the right. Numbering of the mutation according to the transcription start as +1. The sequence encoding the transmembrane domain of SGCE is highlighted in gray; *, mutation also includes 82 bp of adjacent intron 3.

Diagnostic and Ancillary Tests

The diagnosis of M-D is established on clinical grounds and can be confirmed by genetic analysis.

DNA Tests

Genetic testing for M-D – that is, a screen for mutations in the *SGCE* gene – is currently available commercially or on a research basis in selected laboratories specializing in the molecular analysis of movement disorders, but does not usually include a screen for exon rearrangements. A listing of laboratories offering clinical genetic testing can be found at http://www.geneclinics.org.

Other Laboratory Tests

By definition, all laboratory tests are normal in primary M-D. However, routine laboratory testing should be performed to exclude secondary causes of the movement disorder. Liver function tests may be abnormal in patients who regularly consume alcohol.

Neuroimaging

Likewise, neuroimaging studies, including CT and MRI, are entirely unrevealing in M-D, with the possible exception of degenerative changes due to chronic alcohol intake.

Treatment

In general, medical treatment of M-D has proved to be rather ineffective. Uniformly, the best 'treatment' results are achieved with alcohol. However, the risk of addiction to alcohol and the unacceptable side effects of this 'medication' render it an unacceptable treatment option. Initial promising results with γ-hydroxybutyric acid, as well as its cousin sodium oxybate, have been reported, but potential for abuse exists for both of these drugs.

Selected patients have been reported to respond to various benzodiazepines or antiepileptic drugs such as valproate. However, most patients with a good treatment response have been described on a case report basis only. Drugs reported to have a beneficial effect include baclofen, and L-5-hydroxytryptophan. Finally, Korten et al. successfully applied very large doses of levodopa ($5\,\mathrm{g\,day^{-1}}$), which, however, cannot be recommended at such high doses and has otherwise been shown to be ineffective. The following drugs have been tested without any good effect on the involuntary movements: amphetamine, heroin, carbamazepine, propranolol, chlorazepate, and haloperidol.

Stereotactic thalamotomy has improved myoclonus, but caused dysarthria in one patient and mild hemiparesis in another. In two other M-D patients, myoclonus responded to thalamotomy but did not result in any significant gain in function. Deep brain stimulation of the medial globus pallidus was shown to ameliorate both myoclonus and dystonia. Finally, neurostimulation of the ventral intermediate thalamic nucleus has been demonstrated to be a safe and efficacious treatment in several patients with medically intractable and progressing M-D.

See also: Cranial–Cervical Dystonia; Dystonia: Classification, Genetics and Therapeutics; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): Presynaptic Stretch Effects on Neuromuscular Transmission; Obsessive–Compulsive Disorder; Opsoclonus/Myoclonus and Neuroblastoma; Psychiatric Disorders: Functional Genomics; Segawa Dopa Responsive Dystonia; Torsion Dystonia; Writer's Cramp.

Further Reading

Asmus F, Salih F, Hjermind LE, et al. (2005) Myoclonus-dystonia due to genomic deletions in the epsilon-sarcoglycan gene. *Annals of Neurology* 58: 792–797.

Asmus F, Zimprich A, Naumann M, et al. (2001) Inherited Myoclonus-dystonia syndrome: Narrowing the 7q21-q31 locus in German families. *Annals of Neurology* 49: 121–124.

Asmus F, Zimprich A, Tezenas Du Montcel S, et al. (2002) Myoclonus-dystonia syndrome: Epsilon-sarcoglycan mutations and phenotype. *Annals of Neurology* 52: 489–492.

Cif L, Valente EM, Hemm S, et al. (2004) Deep brain stimulation in myoclonus-dystonia syndrome. *Movement Disorders* 19: 724–727.

de Carvalho Aguiar P, Fazzari M, Jankovic J, and Ozelius LJ (2004) Examination of the SGCE gene in Tourette syndrome patients with obsessive-compulsive disorder. *Movement Disorders* 19: 1237–1238.

DeBerardinis RJ, Conforto D, Russell K, et al. (2003) Myoclonus in a patient with a deletion of the epsilon-sarcoglycan locus on chromosome 7q21. *American Journal of Medical Genetics A* 121: 31–36.

Doheny D, Danisi F, Smith C, et al. (2002) Clinical findings of a myoclonusdystonia family with two distinct mutations. *Neurology* 59: 1244–1246.

Doheny DO, Brin MF, Morrison CE, et al. (2002) Phenotypic features of myoclonus-dystonia in three kindreds. *Neurology* 59: 1187–1196.

du Montcel, Clot F, Vidailhet M, et al. (2006) Epsilon sargoglycan mutations and phenotype in French patientes with myclonic syndromes. *Journal of Medical Genetics* 43: 394–400.

Foncke EM, Klein C, Koelman JH, et al. (2003) Hereditary myoclonus-dystonia associated with epilepsy. *Neurology* 60: 1988–1990.

Friedreich N (1881) Neuropathologische Beobachtungen: Paramyoklonus multiplex. *Virchows Archiv* 86: 421–434.

Gasser T (1998) Inherited myoclonus-dystonia syndrome. *Advances in Neurology* 78: 325–334.

Gerrits MC, Foncke EM, de Haan R, et al. (2006) Phenotype-genotype correlation in Dutch patients with myoclonus-dystonia. *Neurology* 66: 759–761.

Grabowski M, Zimprich A, Lorenz-Depiereux B, et al. (2003) The epsilon-sarcoglycan gene (SGCE), mutated in myoclonus-dystonia syndrome, is maternally imprinted. *European Journal of Human Genetics* 11: 138–144.

Grimes DA, Han F, Lang AE, et al. (2002) A novel locus for inherited myoclonus-dystonia on 18p11. *Neurology* 59: 1183–1186.

Grundmann K, Laubis-Herrmann U, Dressler D, et al. (2004) Lack of mutations in the epsilon-sarcoglycan gene in patients with different subtypes of primary dystonias. *Movement Disorders* 19: 1294–1297.

Han F, Lang AE, Racacho L, Bulman DE, and Grimes DA (2003) Mutations in the epsilon-sarcoglycan gene found to be uncommon in seven myoclonus-dystonia families. *Neurology* 61: 244–246.

Hedrich K, Meyer EM, Schule B, et al. (2004) Myoclonus-dystonia: detection of novel, recurrent, and de novo SGCE mutations. *Neurology* 62: 1229–1231.

Hjermind LE, Werdelin LM, Eiberg H, Krag-Olsen B, Dupont E, and Sorensen SA (2003) A novel mutation in the epsilon-sarcoglycan gene causing myoclonusdystonia syndrome. *Neurology* 60: 1536–1539.

Klein C (2003) Myoclonus and myoclonus-dystonias. In: Pulst SM (ed.) *Genetics of Movement Disorders*, pp. 451–471. Amsterdam: Academic Press.

Klein C, Liu L, Doheny D, et al. (2002) Epsilon-sarcoglycan mutations found in combination with other dystonia gene mutations. *Annals of Neurology* 52: 675–679.

Kock N, Kasten M, Schule B, et al. (2004) Clinical and genetic features of myoclonus-dystonia in 3 cases: A video presentation. *Movement Disorders* 19: 231–234.

Lang AE (1997) Essential myoclonus and myoclonic dystonia. *Movement Disorders* 12(1): 127.

Mahloudji M and Pikielny RT (1967) Hereditary essential myoclonus. *Brain* 90(3): 669–674.

Marechal L, Raux G, Dumanchin C, et al. (2003) Severe myoclonus-dystonia syndrome associated with a novel epsilon-sarcoglycan gene truncating mutation. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics* 119: 114–117.

Muller B, Hedrich K, Kock N, et al. (2002) Evidence that paternal expression of the epsilon-sarcoglycan gene accounts for reduced penetrance in myoclonus-dystonia. *The American Journal of Human Genetics* 71: 1303–1311.

O'Riordan S, Ozelius LJ, de Carvalho Aguiar P, Hutchinson M, King M, and Lynch T (2004) Inherited myoclonus-dystonia and epilepsy: Further evidence of an association? *Movement Disorders* 19: 1456–1459.

Scheidtmann K, Muller F, Hartmann E, and Koenig E (2000) Familial myoclonusdystonia syndrome associated with panic attacks. *Der Nervenarzt* 71: 839–842.

Schule B, Kock N, Svetel M, et al. (2004) Genetic heterogeneity in ten families with myoclonus-dystonia. *Journal of Neurology, Neurosurgery, and Psychiatry* 75: 1181–1185.

Quinn NP, Rothwell JC, Thompson PD, et al. (1988) Hereditary myoclonic dystonia, hereditary torsion dystonia and hereditary essential myoclonus: An area of confusion. *Advances in Neurology* 50: 391–400.

Valente EM, Edwards MJ, Mir P, et al. (2005) The epsilon-sarcoglycan gene in myoclonic syndromes. *Neurology* 64: 737–739.

Valente EM, Misbahuddin A, Brancati F, et al. (2003) Analysis of the epsilon-sarcoglycan gene in familial and sporadic myoclonus-dystonia: Evidence for genetic heterogeneity. *Movement Disorders* 18: 1047–1051.

Yokoi F, Dang MT, Li J, et al. (2006) Myoclonus, motor deficits, alterations in emotional responses and monoamine metabolism in epsilon-sarcoglycan deficient mice. *Journal of Biochemistry* 140(1): 141–146.

Zimprich A, Grabowski M, Asmus F, et al. (2001) Mutations in the gene encoding epsilon-sarcoglycan cause myoclonus-dystonia syndrome. *Nature Genetics* 29: 66–69.

Relevant Websites

http://www.dystonia-foundation.org – Dystonia Medical Research Foundation.

http://www.geneclinics.org – GeneTests (a medical genetics information resource, funded by the National Institutes of Health).

http://www.dmd.nl – Leiden Muscular Dystrophy Pages (listing of all allelic variants in the *SGCE* database).

http://www.ninds.nih.gov – National Institute of Neurological Disorders and Stroke (Disorders; in the disorders index, see Dystonias).

http://www.wemove.org – WE MOVE (Worldwide Education and Awareness for Movement Disorders).

Dystrophin, Associated Proteins, and Muscular Dystrophy

H Lidov, Children's Hospital Boston, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Duchenne Muscular Dystrophy: Clinical Features, Pathology

Duchenne muscular dystrophy (DMD) is the most common myopathy, muscular dystrophy, and lethal X-linked disease in humans. Elucidation of the molecular genetics and pathobiology of this disease is the cornerstone of our understanding of all muscular dystrophies and congenital myopathies, and it is therefore the appropriate starting point for an overview of the muscular dystrophies.

Duchenne dystrophy presents in males as progressive proximal weakness, usually noted at about the age of 2 years, when the onset of walking is delayed. Affected boys may rise from sitting on the floor by 'climbing' up their legs, in what is called Gower's maneuver. This consists of first raising the hips high, with hands and feet on the floor, then using the arms to compensate for the weakness of proximal hip muscles by pressing hands successively upward against the feet, shins, and thighs, thus bringing the torso and head to an upright position. The typical course of Duchenne dystrophy is relentless loss of muscle strength, losing ambulation (wheelchair bound) by the age of 8–10 years, developing musculoskeletal deformity, scoliosis, and respiratory compromise, and usually succumbing to the accumulated effects of weakness by the second to third decades of life. Cardiomyopathy is also a feature and fatal arrhythmias may occur. The pseudohypertrophy in the calves and, less prominently, shoulders is due to fibrosis and not to true muscular hypertrophy. Steroids and respiratory and physical therapy may ameliorate some of these effects, although no effective long-term therapy is available.

The disease affects one in 3000 boys (boys carry only a single copy of the X chromosome, so X-linked disorders are usually only seen in males). Nevertheless, extremely rare abnormalities in the X chromosome may manifest as DMD in girls. A significant finding is markedly elevated serum creatinine phosphokinase (CPK), a muscle enzyme that is released in the systemic circulation by the breakdown of myocytes. Although an elevation in CPK is only indicative of myocyte destruction from many possible causes, the elevations in DMD, in the range of 2000–20 000 IU, is rarely seen except in muscular dystrophies. After initial examination and laboratory studies, most patients suspected of having a myopathy undergo a skeletal muscle biopsy, which in DMD has diagnostic features. The features do, however, vary with the clinical stage of disease at the time of the biopsy. In young boys at the age of ~2 years, when the diagnosis of DMD may first be entertained, the muscle biopsy shows excessive variation in fiber size, and occasional degenerating or regenerating muscle fibers and very subtle fibrosis. The biopsy has remarkably few pathological changes. By the age of 8 years, when the patient is likely severely clinically affected, possibly nonambulatory, the biopsy shows marked variation in fiber diameter, with both tiny atrophic fibers and some enlarged hypertrophic fibers, and a dramatic increase in degenerating and regenerating myocytes. There may be scattered chronic inflammatory cells and macrophages. Fibrosis is present, and there will be variable amounts of infiltration by adipose tissue (fatty replacement). Biopsies in the second decade of life, or seen at autopsy, will show 'end-stage' muscle, typically with only a few atrophic fibers remaining, and muscle fibers largely replaced by fibroadipose tissue. At this stage the muscle will appear to the eye fibrous and yellow, hardly recognizable and without the normal pink appearance of muscle (Figure 1).

The diagnostic test that can be performed with clear result at any of these stages is immunocytochemical staining for dystrophin (see later). In current medical practice it is now frequently possible to unequivocally diagnose DMD based on clinical features and molecular genetic studies performed on patient genomic DNA from several sources, such as a blood sample or buccal swab. In a small percentage of patients these tests are not diagnostic and further testing on a muscle biopsy is necessary.

Dystrophin: Identification, Features of Gene and Protein

DMD was described clinically in the late nineteenth century, and efforts to identify a pathogenetic process were repeatedly made in the twentieth century, using histological and biochemical techniques of ever-increasing sophistication and sensitivity. Nevertheless, no abnormality that would explain the relentless death of myocytes could be found. Starting in the mid-1980s a purely genetic approach was harnessed to study the genomic DNA of an unfortunate boy with a relatively large deletion on the X chromosome; he had three diseases, McLeod's anemia, chronic granulomatous disease, and DMD, suspected to reflect three contiguous genes lost in the single deletion.

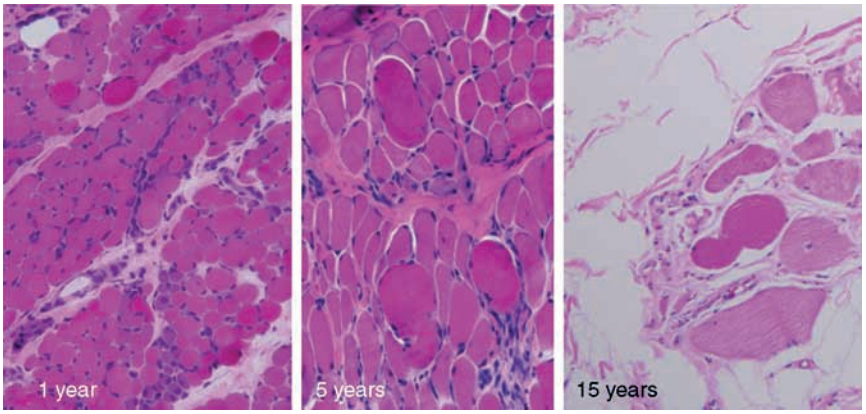

Figure 1 Muscle biopsies, stained with hematoxylin and eosin, from patients with Duchenne muscular dystrophy at three ages. Note the progressive fibrosis, noticeable at the age of 5 years, and the profound loss of myocytes and replacement by adipose tissue at 15 years.

Based on this, Kunkel and co-workers successively localized the deleted DMD gene, cloned the complete cDNA, and developed antibodies based on the predicted amino acid sequence. The last effort facilitated characterization of the protein product of the deleted DMD gene. This protein, which is absent in DMD, was named dystrophin. The 'reverse genetic' approach, a term that refers to locating the disease-causing protein by working 'backward' from the mutation in an affected individual, rather than forward from an abnormal protein, was groundbreaking and a forerunner of what is now the commonly used method of positional cloning.

The dystrophin gene is 2.4 megabases (Mb; million base pairs) long and in terms of genomic span is the largest in the human genome. It has been speculated that this may contribute to the high frequency of mutations, since a random process might be expected to affect a large target with a proportionately high frequency. The precise genetic locus is Xp21.2. The cDNA is 14 kilobases (kb; thousand base pairs) and is composed of 74 exons. Whereas the gene spans a uniquely large region of genomic DNA, the cDNA, like the protein which it encodes, is large, but by no means the largest cDNA or protein (the great size of the dystrophin gene reflecting the large stretches of intronic DNA).

Certain predictions about dystrophin can be made from study of the cDNA and recognition of homologies with other well-studied proteins. Dystrophin is composed of 3684 amino acids and has a high degree of homology with the spectrin family of actin-binding cytoskeletal proteins. The N-terminal nucleic acid sequence predicts a canonic actin-binding domain. The central portion of the protein contains 24 approximately repeated sequences, characteristic of spectrin family members. This portion of spectrin, like cytoskeletal proteins, has the biophysical properties of a stiff rod, and in the case of dystrophin is interrupted by two proline-rich regions which predict a hinge. There are two additional proline-rich hinge regions at each end of the rod domain. The C-terminal region is unique and bears no homology to any known protein. Its structure suggests a calcium-binding domain, although further study has indicated that there is no functional calcium binding. The C-terminal region was initially speculated to be associated with unspecified integral membrane proteins, and further work has confirmed that it is the moiety which is involved in protein–protein interactions with a variety of other proteins. The C-terminal domain is also subject to a complicated assortment of tissue-specific alternative splice variations, and these in turn affect the possible repertoire of protein–protein interactions.

A milder phenotypic variant of DMD is Becker muscular dystrophy (BMD), which is also X linked and about one-tenth as common. The pattern of weakness is similar, but the clinical features are considerably more variable in age of onset, progression, and severity. It was initially predicted on purely genetic grounds that this was not a distinct disorder, but was an allelic variant at the DMD locus. Molecular genetic studies have confirmed this. In essence, DMD is marked by the nearly complete absence of dystrophin, whereas BMD reflects the presence of dystrophin which is abnormal in molecular weight (either decreased or increased, reflecting deletions or duplications) or is reduced in amount. An arbitrary threshold that has been used is a reduction to 5% or less of normal for DMD.

As a general rule, frame-shifting mutations in any part of the dystrophin cDNA will produce DMD, while mutations which preserve the reading frame

will produce BMD. There are exceptions: massive deletions will produce DMD, and deletions in the C-terminal region tend to produce DMD irrespective of reading frame.

With the identification of dystrophin, the development of polyclonal and monoclonal antibodies to dystrophin became possible and this in turn created a powerful tool for both the exploration of normal biology and pathologic diagnosis. Dystrophin is normally found precisely associated with the sarcolemma of the myocyte, but not with any sarcoplasmic or internal structures (Figure 2). In the standard transverse plane this appears as a uniform circumferential line, but confocal or other imaging techniques show that this is in fact a reticulated network, with a spacing that roughly corresponds to the sarcomeric or Z-line spacing. Immunoelectron microscopy has demonstrated that the localization is more precisely subsarcolemmal. The density of dystrophin appears to be increased at the two areas of sarcolemmal specialization, the neuromuscular junction (NMJ) and the myotendinous junction. Although the context is different, dystrophin maintains this sarcolemmal distribution in smooth muscle, cardiac muscle, and extraocular muscle.

Standard immunohistochemistry techniques using a dystrophin antibody applied to frozen tissue sections of a muscle biopsy show an essentially complete absence of dystrophin reactivity, diagnostic of DMD (Figure 3). The term 'essentially' is used because, by a variety of not completely understood processes, some mutations can be circumvented by alternative splicing or perhaps second mutations, with the result that a vary rare minority of fibers are seen by immunocytochemistry to express normally localized dystrophin, so-called revertant fibers. The amount is minute, undetectable by Western blot, and of no physiologic significance. Nevertheless, augmentation of these fibers is considered as a therapeutic approach.

Less commonly, patterns of uniform weak and or irregular, but still sarcoplasmic, dystrophin immunoreactivity are seen, and these are characteristic of BMD. It is prudent to confirm the diagnosis with Western blot or molecular genetic analysis.

In women who carry one defective copy of the dystrophin gene, a mosaic pattern may be seen. In such 'carrier females' rare fibers – and it should be noted, only in some regions ('nuclear domains') in the multinucleate syncytial myocyte – are negative for dystrophin. In most cases this is a small proportion of fibers and there is no clinical myopathy, although serum CPK may be elevated. In rare female carriers there will be a skewed X inactivation, for reasons that remain speculative, and a larger proportion, possibly even a majority, of fibers will be dystrophin negative, because there has been inactivation of the X chromosome carrying the wild-type dystrophin gene. These unfortunate women will manifest varying degrees of myopathy and, of particular concern, cardiomyopathy.

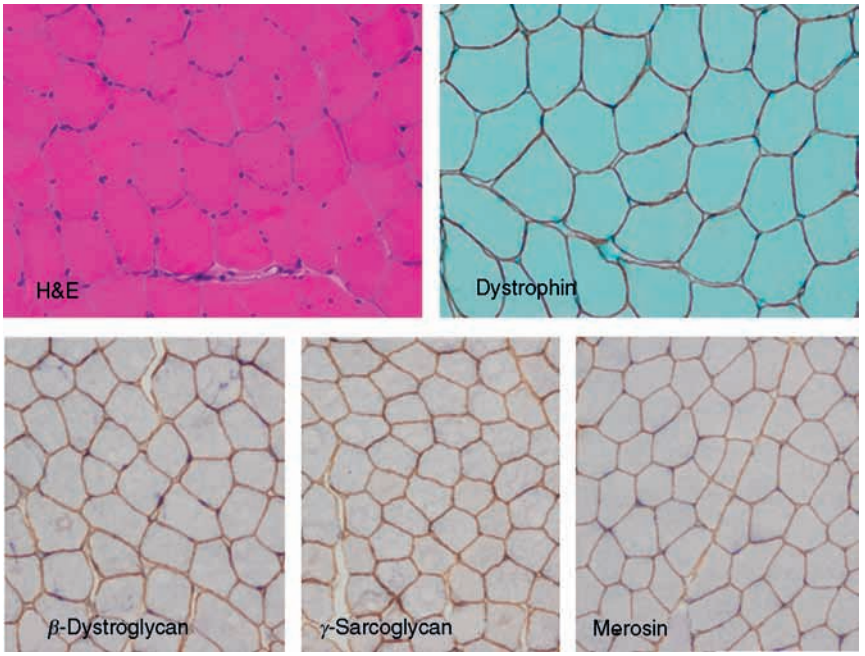

Figure 2 Hematoxylin and eosin (H&E) stain of normal skeletal muscle (upper left) and immunocytochemical labeling of muscle for dystrophin, β-dystroglycan, γ-sarcoglycan, and merosin (α2-laminin subunit). Note that labeling for each of the dystrophin and dystrophin-associated proteins is restricted to the sarcolemma.

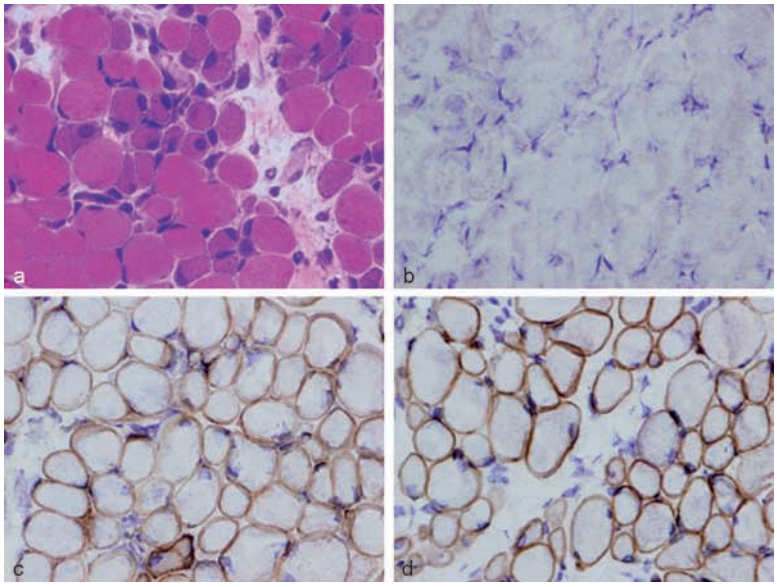

Figure 3 A muscle biopsy from a boy with presymptomatic Duchenne muscular dystrophy. Hematoxylin and eosin-stained section (a) shows abnormal variation in fiber diameter, and small rounded basophilic fibers which are regenerating (b). Dystrophin immunostaining is absent. β-Dystroglycan (c) and α-sarcoglycan (d) immunolabeling is uneven, but present and normally localized to the sarcolemma.

Autosomal Limb-Girdle Muscular Dystrophies

Although DMD constitutes the large majority of progressive proximal myopathic syndromes with dystrophic features based on muscle biopsy, clinical neurologists have long recognized that there were phenotypically similar diseases which were most distinctive for not being sex linked (i.e., occurring in females), and in families that were ascertained to show autosomal recessive or, less commonly, dominant inheritance. The clinical courses of these diseases were quite variable or heterogeneous, although in some instances as severe as DMD. These diseases were referred to as types of limb-girdle muscular dystrophy (LGMD). One geographic region in which an autosomal recessive LGMD is particularly prominent is North Africa; the disease there was initially called severe childhood autosomal recessive muscular dystrophy (SCARMD), or muscular dystrophy of the Magrebian type (LGMD2C).

Genetic linkage studies have been the backbone of defining or rationalizing the now three dominant and 12 recessive forms of this disease. Linkage studies which establish a chromosomal localization (or at least exclude all of the known localizations) are now the standard for establishing a new member of this class (Table 1).

Congenital Muscular Dystrophy

In addition to DMD and the autosomal variants which present in childhood, there are patients identified in infancy who are typically hypotonic and whose biopsies show a dystrophic myopathy. Rather than being characterized by markedly elevated serum CPK, the CPK is usually normal to moderately elevated (<1000 IU). Disease progression is usually slow, with survival into adulthood. Within this group of diseases some patients have additional features. Some showed an associated rigidity or contractures, sometimes at birth and sometimes developing over time. Others showed associated, often severe, central nervous system (CNS) and ocular malformations, referred to as Walker–Warburg syndrome, muscle–eye–brain (MEB) disease, and Fukuyama congenital muscular dystrophy. All are relatively rare, autosomal recessive diseases and show some geographic predilection. MEB is found in Finland, and Fukuyama muscular dystrophy is most commonly found in Japan.

The Dystrophin-Associated Complex

Following the identification of dystrophin there was rapid progress in biochemical, genetic, and molecular biological unraveling of the LGMDs. Sarcolemmal membrane preparations were subjected to isolation, solubilization, and separation of proteins in elution fractions enriched in dystrophin (i.e., dystrophin-associated proteins). The precise number of proteins was initially not clear, but consensus has settled on a canonic seven proteins: α- and β-dystroglycans (43 and 156 kDa); α-, β-, γ-, and δ-sarcoglycans (50, 43, 32, and 35 kDa, respectively); and sarcospan (25 kDa). The four sarcoglycans and α-dystroglycan are glycoproteins. Posttranslational cleavage of a

Table 1 Tabulation of the muscular dystrophies

Disease	*Inheritance*	*Gene location*	*Gene product*
DMD/BMD	XL	Xp21	Dystrophin
LGMD-1A	AD	5q31	Myotilin
LGMD-1B	AD	1q11–q21	Lamin A/C
LGMD-1C	AD	3p25	Caveolin 3
LGMD-1D[a]	AD	6q23	?
LGMD-1E[a]	AD	5q31	?
LGMD-1F[a]	AD	7q32.1	?
LGMD-2A	AR	15q15–q21	Calpain 3
LGMD-2B	AR	2p13	Dysferlin
LGMD-2C	AR	13q12	γ-Sarcoglycan
LGMD-2D	AR	17q21	α-Sarcoglycan
LGMD-2E	AR	4q12	β-Sarcoglycan
LGMD-2F	AR	5q33	δ-Sarcoglycan
LGMD-2G	AR	17q11–q12	Telethonin
LGMD-2H	AR	9q31–q34	E3 ubiquitin ligase
LGMD-2I	AR	19q13.3	FKRP
LGMD-2J	AR	2q31	Titin
LGMD-2K	AR	9q34.1	POMT1
LGMD-2L	AR	11P13	?
LGMD-2M	AR	9q31	Fukutin
MDC-1A	AR	6q2	Merosin
MDC-1B	AR	1q42	?
MDC-1C	AR	19q1	FKRP
MDC-1D	AR	22q12.3–q13.1	LARGE
Walker–Warburg	AR	19q13.3, 14q24.3, 9q34.1, 9q31	FKRP POMT1, POMT2 Fukutin
Muscle–eye–brain	AR	19q13.3, 1p34–p33	FKRP POMGnT1
Fukuyama MD	AR	9q31	Fukutin

[a]Conflicting reports unresolved.

The genetic linkage, and proteins harboring causative mutations where known, are listed. AD, autosomal dominant; AR, autosomal recessive; BMD, Becker muscular dystrophy; DMD, Duchenne muscular dystrophy; FKRP, fukutin-related protein; LGMD, limb-girdle muscular dystrophy; MDC, merosin-deficient congenital muscular dystrophy; POMGnT1, protein *O*-mannose β-1,2-*N*-acetylglucosaminyltransferase 1; POMT, protein *O*-mannosyltransferase; XL, X linked.

single peptide produces α- and β-dystroglycans, and the α component is glycosylated. Refinement of the solubilization conditions suggested that although the group as a whole formed a complex – the dystrophin-associated glycoprotein complex (DAG complex) – the dystroglycans and sarcoglycans were each more tightly associated into two subcomplexes.

Each of these proteins has been shown to have complicated but selective and specific patterns of association with other muscle proteins, with some but not all other members of the DAG complex, and with other cytosolic muscle proteins. Of particular importance, a dystroglycan is found to bind to merosin, the α_2 subunit of laminin, a component of the muscle extracellular matrix.

The DAG Complex: A Functional Model

Based on the initial details, a working functional model of the DAG complex evolved; this model has been refined, with major aspects supported by considerable further experimental testing. At its simplest functional core, it appears that there is a bridge from cytoskeletal actin, through dystrophin, to β-dystroglycan, to α-dystroglycan, and ending in merosin (Figure 4). This bridge is presumed to provide mechanical support for the sarcolemma, which must withstand a lifetime of cycles of contraction and relaxation. The complex of sarcoglycans is tightly attached to β-dystroglycan and, based on its potential for causing LGMD, appears to contribute critically to the function of this bridge, though a precise role has been elusive. The most common speculation is that the sarcoglycan subcomplex plays a facilitating signal transduction role, but experimentally tested details are wanting.

Two lines of research, the genetic characterization of the LGMD and the molecular genetic characterization of proteins closely interacting with dystrophin, rapidly converged. Simply, the proposed mechanistic model suggested that mutations rendering elements other than dystrophin in the bridge connecting the actin cytoskeleton to the extracellular matrix ought to produce diseases phenotypically similar to DMD/BMD. It was quickly recognized that the North African

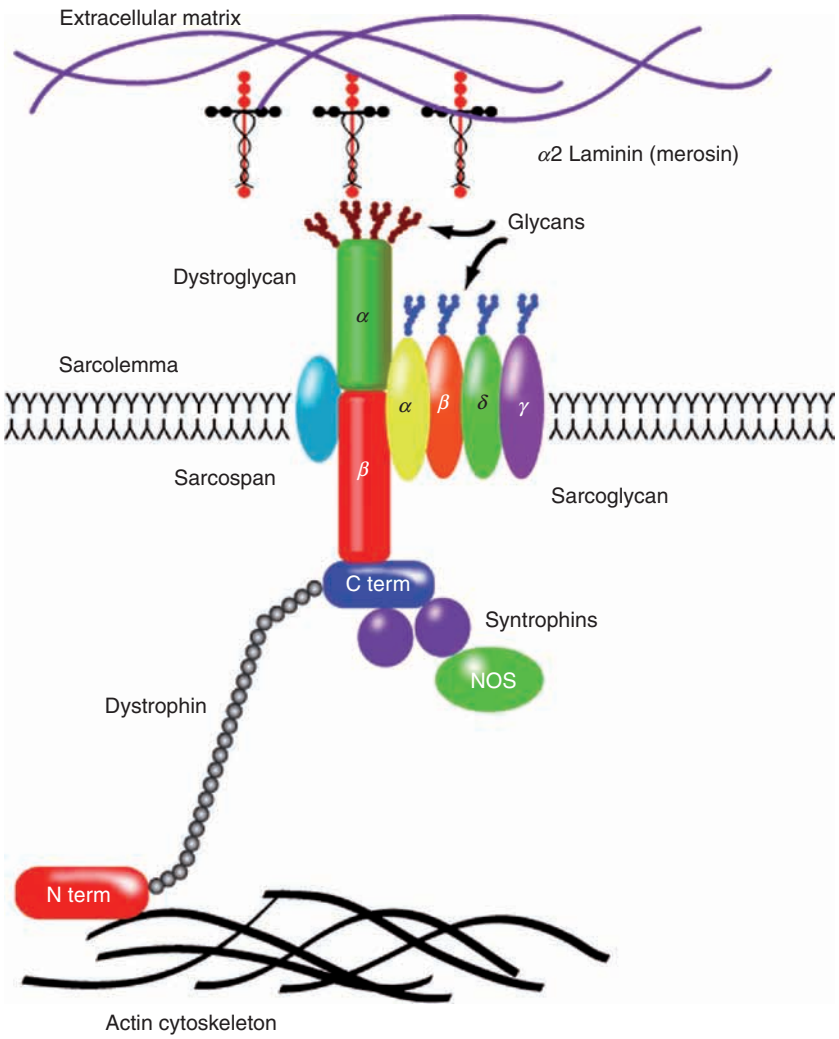

Figure 4 Diagram of dystrophin and the dystrophin-associated complex. C term, C-terminal end; N term, N-terminal end; NOS, nitric oxide synthetase.

SCARMD reflected mutations in α-sarcoglycan or, somewhat more frequently, γ-sarcoglycan. Shortly thereafter, families with type 2E LGMD (LGMD-2E) were recognized to have mutations in β-sarcoglycan, and somewhat later the very uncommon LGMD-2F was found to reflect a δ-sarcoglycan mutation. Each of these mutated proteins could be reliably identified by immunohistochemistry with specific antibodies performed on muscle biopsies. The attractive suggestion – that the presence of a null mutation in any one of the sarcoglycans destabilized the whole complex and as a result there was an essentially complete secondary loss of all sarcoglycans – does not appear to be true. The levels are reduced, but the secondary reduction is variable and not absolute.

Certain other predictions of this model present apparent or real paradoxes. The model suggests that dystroglycan mutations should manifest as an LGMD, yet no dystroglycan-null patients have been identified. Since α and β forms are produced by the cleavage of a single peptide, they might not be lost independently. The surprising absence of pathogenic dystroglycan mutations was explained when a targeted deletion in the mouse was created and shown to be an early embryonic lethal event. A tissue-specific conditional mutant and fortuitous mosaics do confirm that loss of dystroglycan in muscle does produce a severe dystrophic myopathy. This highlights the further consideration that although mutations of each of the components of dystrophin-associated proteins may produce a myopathy, these proteins may be expressed in nonmuscle tissues, and thus mutations produce a phenotype of 'myopathy plus,' depending on expression in other tissues. In the case of dystroglycan, Reichert's membrane is such an additional critical site. In contrast, sarcospan is a 25 kDa tetraspan

integral membrane protein which is a component of the dystrophin-associated complex. This is lost secondary to the loss of sarcoglycans, but so far it is not itself associated with any recognized disease phenotype; naturally occurring mutations have not been identified but a targeted null mutation in mouse has no apparent pathology.

Microscopic analyses have shown that a large number of proteins have the same sarcolemmal localization as do dystrophin and the core dystrophin-associated proteins, or, based on biochemical analyses, are present in sarcolemmal membrane fractions enriched in dystrophin or associate with one or more of the DAG proteins *in vitro*. Some of these other proteins, but not all, when mutated, cause forms of LGMD. Whether these should be regarded as part of the dystrophin-associated complex is unclear. Viewed in this more expansive fashion, the complex consists of ten or more proteins, some of which have predominant localizations elsewhere in the myocyte. The following proteins are some of the better studied examples.

Calpain 3 is a calcium-activated neutral protease which is mutated in patients with LGMD-2A. Although it is tempting to speculate that its substrate is a component of the DAG complex (DAGC), or a protein that interacts with the DAGC, the substrate or substrates are not know. Caveolin 3 is the muscle-specific member of the family of proteins, caveolins, which coat caveolae, the 50–100 μm vesicles and invaginations that lie below the plasma membrane of myocytes, the sarcolemma. Caveolins have been suggested to have a variety of functions, and it seems likely that they are involved in membrane trafficking or recycling. Mutations in caveolin 3 cause LGMD-1C.

α-Dystrobrevin is the 94 kDa mammalian homolog of an 87 kDa phosphoprotein, originally isolated as a component of the *Torpedo* electric organ, and localized to the NMJ. A homolog, β-dystrobrevin, is encoded by a separate gene and is found in several tissues (kidney and brain, but not skeletal muscle). α-Dystrobrevin is highly alternatively spliced, giving rise to at least five variants. This protein proved to be one of the components of the dystrophin-enriched muscle membrane fraction, and associates with the C-terminal end of dystrophin as well as with syntrophin. Mice with a deletion in the dystrobrevin gene have a mild skeletal and cardiac myopathy (albeit much less dramatic than when there is dystrophin or sarcoglycan deficiency) and structural changes at the NMJ. One family with cardiomyopathy has been found to have a missense mutation in the dystrobrevin gene.

Syntrophins comprise a family of three distinct but highly homologous and conserved 53 kDa intracellular proteins. They have a postsynaptic density/disc/zonula occludens-1 (PDZ) domain which facilitates interacting with a variety of signaling and membrane proteins, including nitric oxide synthetase (NOS), aquaporin-4, and potassium channels, among others. They also bind to the C-terminal domain of dystrophin, exon 74, so they may couple dystrophin to a variety of functionally different molecules. In muscle, β_2-syntrophin is found predominantly at the NMJ, while the α_1 form is found at the extrajunctional sarcolemma. It addition to the range of proteins interacting with syntrophin, syntrophin binds to an alternatively spliced exon in dystrophin, possibly the mechanism for further precise control of the location and type of protein–protein interaction. Thus far no myopathy or other disease has been associated with syntrophin mutations. Mice with a target null mutation of a syntrophin do not have a dystrophic phenotype and only show structural alterations of the NMJ.

The spatially specific distribution of syntrophins raises a more general distinction. There may be two independent systems involving dystrophin and several of the dystrophin-associated proteins: one at the sarcolemma, primarily supporting mechanical stability, and an independent role at the NMJ, involved in organization of the NMJ postsynaptic apparatus. Only the former has a clearly established causal role in a group of diseases, the muscular dystrophies. There is no recognized pathology or pathophysiology of the NMJ in Duchenne dystrophy.

Filamin C is a striated muscle-specific member of a family of actin-binding proteins. It is localized to the sarcolemma, as well as to cytosol and sarcomeric Z-lines, is associated with γ- and δ-sarcoglycan, as well as multiple Z-line proteins, and its distribution seems to be altered in some patients with muscular dystrophy. There are no recognized diseases with filamin C mutations and it has not been identified in dystrophy patients, but mice with a targeted deletion have lethally hypomorphic skeletal muscle and die at birth due to respiratory failure. Filamin C seems a good candidate to play an intracytoplasmic signaling role, one element of which is the dystrophin–DAG complex.

A remarkable finding is that neuronal nitric oxide synthetase (nNOS) is associated with the dystrophin–DAG complex. This association appears to be mediated through syntrophin, which in turn is associated with dystrophin. The cellular localization is sarcolemmal. nNOS is lost in the *mdx* mouse model of dystrophin deficiency, as well as in DMD biopsies. Clearly a host of possible cytotoxic and cytoprotective roles might be mediated through nNOS, although a specific role in dystrophic muscle remains speculative.

Glycosylation-Dependent Muscular Dystrophies; Muscular Dystrophies with CNS Involvement

Although CNS abnormalities are not found with most muscular dystrophies, there is a distinct group of diseases with dystrophic myopathy and typically neuronal migration defects. In this group are Fukuyama muscular dystrophy, muscle–eye–brain disease, Walker–Warburg syndrome, LGMD-2K, and MDC-1C. These are associated with various constellations of structural ocular abnormalities and cerebral malformations, including lissencephaly, cerebellar and pontine hypoplasia, and agenesis of the corpus callosum. Seizures and mental retardation are common.

The explanation of this conjunction of muscular dystrophy and CNS malformation is found in the answer to a molecular genetic paradox. It is apparent in the list of genetically distinct LGMDs that although there is a group of dystrophies that are caused by mutations in dystrophin-associated proteins which function consistent with the proposed mechanistic model, this group consists of only six of 20 dystrophies. Others do not easily fit this model. Some of these mutated dystrophin-associated proteins, including fukutin, fukutin-related protein (FKRP), LARGE, POMT1, and POMGnT1, have the feature in common of facilitating the glycosylation of other proteins, specifically a dystroglycan. All of the disorders associated with the genes encoding these proteins are characterized by hypoglycosylation of a dystroglycan, and the hypoglycosylated form has been shown to be unable to bind normally to laminin (and possibly also to other interacting proteins such as neurexin, perlican, and agrin).

POMT1, POMT2, and POMGnT1 have demonstrated *O*-mannosyltransferase activity. LARGE is predicted to be a glycosyltransferase based on sequence homology, although activity has not been demonstrated; however, LARGE does appear to be able to bypass the absence of POMT in glycosylating α-dystroglycan. Fukutin is localized to the Golgi apparatus, and but it is unclear that it has enzymatic activity; it may exert its effect by associating with or regulating a glycosyltransferase. FKRP similarly may function as a chaperone rather than having autonomous enzymatic activity. Substrate specificities for these proteins are poorly understood, and their effects may be exerted by glycosylating proteins other than or in addition to dystroglycan.

The glycosylation of sarcoglycans remains to be studied. In the absence of a clear understanding of the role of the sarcoglycans it is difficult to say what effect failure of normal glycosylation might have. Nevertheless, it seems reasonable to speculate that hypoglycosylated sarcoglycans may not be able to play their normal role, and like the absence of sarcoglycans, hypoglycosylation may disrupt the function of the structural bridge from actin through dystrophin and dystroglycan to the extracellular matrix, leading to a dystrophic myopathy.

A striking feature of the muscular dystrophies, which appear to be mediated through the failure to glycosylate components of the DAG complex, is their propensity to also affect the CNS. This reflects a role for dystroglycan at the blood–brain barrier (BBB), and in targeted deletions of dystroglycan, using tissue-specific conditional deletion (to circumvent the embryonic lethal phenotype of the nonconditional deletion), the dystroglycan mutations produce cerebral malformations. This effect is certainly independent of dystrophin, since the localization is totally different from the localization of dystrophin in the CNS. It is also independent of α-, γ-, and δ-sarcoglycans, since these do not occur in the CNS. It does, however, parallel the distribution of utrophin, so this may be the cytoplasmic component of a transmembrane complex, and might also involve the CNS-specific short dystrophin gene products, Dp71 and Dp140 (see later).

Dysferlin: Sarcolemmal Repair Mechanism

Dysferlin presents a novel but paradigmatic extension of thinking about sarcolemmal proteins and LGMD. Dysferlin is a 230 kDa protein with a high degree of homology to the fer-1 family of proteins, initially found in *Caenorhabditis elegans*. It was identified by positional cloning as the gene responsible for LGMD-2B as well as Miyoshi (distal) myopathy. Initially it was supposed, given the series of sarcolemmal proteins which were causal for LGMD, that dysferlin, also sarcolemmal, and causative of LGMD-2B, should also be associated with the DAG complex. Indeed, a small amount of early data suggested an association with caveolin 3. However, further studies showed that dysferlin-deficient myocytes were slow to seal calcium leaks after membrane damage. The membrane 'patching' process involves recruitment of membrane vesicles, possibly associated with caveolin 3, to a region of damage, and involves an association of dysferlin with annexin A1 and annexin A2, a connection first suggested by relevance network analysis of expression array data, and then confirmed by standard biochemical methods. The process is calcium dependent. Dysferlin therefore is the mirror of the dystrophin-associated proteins; whereas the latter appear to function by protecting the sarcolemma from mechanical stress damage, dysferlin functions independently to repair such damage.

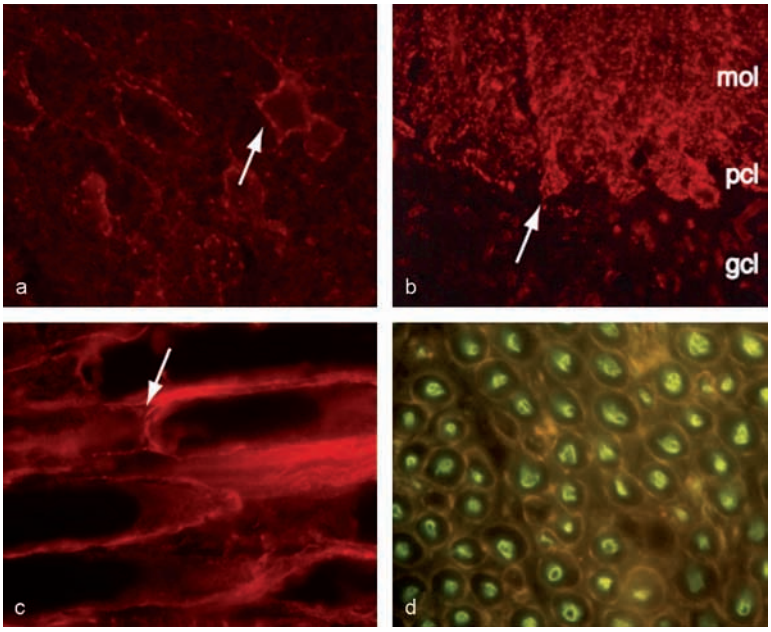

Figure 5 Dystrophin in murine (a) neurons of frontal cortex (arror indicates neuronal cell body) and (b) cerebellar Purkinje cells (mol, molecular; pcl, Purkinje cell layer; gcl, granule cell layer), arrows indicates Purkinje cell body, and (c, d) in peripheral nerve (c: longitudinal section – arrow indicates a node of Ranvier; d: double-labeled cross section – neurofilament-labeled axons are green, Dp116 is red). In neurons, the dystrophin labeling is a punctate distribution along perikaryon and dendrites. In peripheral nerve, the short dystrophin isoform Dp116 is localized to the Schwann cell membrane.

Dystrophin and Dystrophin-Associated Proteins in the Central and Peripheral Nervous Systems

Although the vast majority of detailed understanding of dystrophin and the dystrophin-associated proteins is derived from study of skeletal muscle, and patently this is the locus of the pathologic process that results in muscular dystrophy, these proteins are expressed in nonmuscle tissue. This is the basis for some nonmyopathic features of muscular dystrophies, raising many novel possibilities about the functions of these proteins.

Dystrophin is present predominantly in skeletal, cardiac, and smooth muscle, but is also found at about 100-fold lower levels in CNS tissue. This is still greater than the low background level of transcript and protein found in most other tissues, and is presumed to reflect the endogenous vascular smooth muscle. Tissue culture and neuroanatomical studies have shown that CNS dystrophin is localized to neurons, and further to specific classes of neurons, such as those of the cerebral cortex, the hippocampus, and the Purkinje cells of the cerebellum. Furthermore, the distribution is in a punctuate pattern along dendrites and neuronal perikarya, a pattern that would be consistent with a localization to synaptic junctions (Figures 5(a) and 5(b)). Immunogold ultrastructural studies suggest localization to synaptic junctions, and fractionation of synaptosomes and postsynaptic densities show that these are progressively more enriched in dystrophin. This enrichment also parallels the development of synapses in the early postnatal period. Studies of the retina show similar punctuate distribution of dystrophin in the outer plexiform layer, which is predominantly synaptic neuropil, and some studies report localization at synaptic junctions.

A functional presence of dystrophin at synapses in the CNS, and particularly in the neocortex, would fit well with the early observation of a mild static cognitive impairment that is a well-documented feature of DMD and BMD. Repeated studies have shown that the IQ in cohorts of boys with DMD is normally distributed (arguing against the co-deletion of a second contiguous gene) and the mean is depressed by about 1 standard deviation. It follows that about one-third of boys with DMD will have an IQ that falls below the arbitrarily chosen threshold for 'mental retardation' of 70, but it is important to recognize that this does not imply a distinct cognitively affected population, but rather that there is some effect on cognitive function in all affected individuals. In contrast to the musculoskeletal impairment, the mild cognitive effect is nonprogressive and does not correlate with the degree of musculoskeletal disease.

Dystrophin Isoforms

Following the identification of full-length, 427 kDa dystrophin, it was recognized that there were several transcripts, as well as a very homologous autosomal protein, utrophin, and a series of small homologous utrophin-related transcripts. The function of these proteins remains unclear, although their tissue-specific distribution suggests or constrains their possible roles. The smaller dystrophin transcripts are derived from additional tissue-specific promoters located in introns of the dystrophin gene. Each of the smaller transcripts, like dystrophin, may have some additional variation as the result of alternative splicing. Some but not all of the small transcripts have short, novel first exons, although the majority of the coding sequence, and thus the composition of the protein, is identical to the cognate region of dystrophin. In terms of domains, all of the small transcripts are composed of varying amounts of the dystrophin rod domain, and all preserve the novel C-terminal domain. Apparently this precludes that any of the short forms are actin-binding proteins, but all are involved in protein–protein interactions, likely with integral membrane proteins, either dystroglycan, or other unidentified ligands. The specific shorter transcripts are as follows:

- Dp71 (71 kDa), found in many tissues.
- Dp116 (116 kDa), found in Schwann cells.
- Dp140 (140 kDa), found in CNS, possibly glia, and the epithelium of the ascending loops in the kidney.
- Dp260 (260 kDa), found in the retina.

In addition, full-length dystrophin transcription is driven by three relatively tissue-specific promoters, the skeletal muscle promoter, a cerebral promoter (brain promoter), and a Purkinje cell promoter. The transcripts and protein products are identical, and at least in the case of the muscle and brain promoter the brain promoter will drive some, albeit a lower than normal, level of transcription in muscle, in the absence of a functional muscle promoter.

It follows that in the CNS the situation in regard to dystrophin transcripts is considerably more complicated than in skeletal muscle. There are three promoters driving the expression of full length dystrophin in neurons, as well as Dp260, predominantly in retinal tissue; Dp140, possibly in astrocytes or ependymal cells; and Dp71, in greater abundance than any of the foregoing, although the localization is unclear, but likely to be glial. Although a strict genotype–phenotype correlation between cognitive impairment and dystrophin mutations has not been established, there is a suggestion that deletions toward the C-terminal region may be overrepresented in the more cognitively impaired DMD patients. Consistent with this, the more C-terminal mutations would disrupt more of the different dystrophin transcripts, and hence perhaps produce greater perturbations in CNS function. Nevertheless, loss of 427 kDa dystrophin alone is sufficient to produce cognitive impairment. Although no correlated physiologic abnormality has been identified in the neurons of the CNS, a distinctive retinal physiologic abnormality, decreased amplitude of the electroretinographic scotopic b-wave, has been correlated with loss of dystrophin and Dp260.

The question of a neuronal equivalent of the DAG complex is also complicated and very incompletely understood. Dystroglycan is present in the CNS, but appears to have a completely different and nonoverlapping distribution compared to dystrophin. Dystroglycan is found at the pia limitans along the Virchow–Robin perivascular boundary. This most closely approximates the distribution of utrophin, the autosomal homolog of dystrophin. The localization of the short dystrophin isoforms (Dp140 and Dp71) is too poorly demonstrated to say if there is a better correlation with either of these. The α-, γ-, and δ-sarcoglycans are not expressed in the CNS; the β-sarcoglycan transcript is present, although the protein has not been localized. A gene encoding the fifth sarcoglycan, epsilon, a protein which does replace α-sarcoglycan in some sarcoglycan complexes in cardiac and smooth muscle, has been identified by positional cloning as the gene responsible for familial myoclonus-dystonia, but the localization or function of the protein in the CNS remains unknown.

Nevertheless, targeted disruptions of dystroglycan in mice, or abnormalities of its glycosylation in mice or humans, produce migrational abnormalities, particularly adjacent to the pia limitans. CNS abnormalities, in addition to the dystrophic myopathy, are a feature of just those LGMD which appear to result from mutations of putative glycosyltransferases and reduction or loss of their function. However, this effect is independent of dystrophin, and these malformations are not a part of the DMD phenotype. Conversely, the cognitive impairment associated with DMD is not a feature of LGMD-2C–LGMD-2F. It seems clear that dystroglycan is involved in a complex in the CNS that does not involve dystrophin, but possibly substitutes for utrophin, and that dystrophin is likely associated with one or more integral membrane proteins or complexes, but is different (and at present unknown) from those in muscle. One of these associations may be known; a growing body of evidence suggests an association of Dp71 with aquaporin-4, which mediates a role in blood–brain barrier function and water homeostasis.

Exploration of the functional roles of proteins first identified in skeletal muscle in association with dystrophin, as they occur in the CNS, promises many more such unexpected findings.

The DAG Complex in Peripheral Nerve

As mentioned, Schwann cells which ensheathe peripheral nerve axons express Dp116, localized to the outer wrap of the Schwann cell (abaxonal Schwann cell membrane) and at the node of Ranvier (Figures 5(c) and 5(d)). There is a parallel localization of dys troglycan, the ε-, β-, δ-, and ζ-sarcoglycans, and the α_2 subunit of laminin (merosin). These appear to form a membrane-spanning structural complex with considerable similarity to that present in skeletal muscle. Loss of the most peripheral component, the α_2 subunit of laminin (merosin-deficient congenital muscular dystrophy; MDC-1A), is characterized by slowing of the peripheral motor nerve conduction, both in humans and in mice with the similar mutation. The murine model, the *dy/dy* mouse, has marked pathology, including bundles of unmyelinated axons in the spinal roots, hypomyelination, abnormalities at the node of Ranvier, and abnormal ion channel distribution; descriptions of the equivalent human nerve biopsies are lacking. Nerve conduction abnormalities are not reported in association with C-terminal dystrophin mutations, which would affect Dp116 and β-sarcoglycan or δ-sarcoglycan mutations (LGMD-2D, LGMD-2F), but since nerve conductions studies are not routinely part of the diagnostic evaluation of suspected muscular dystrophy this may be simply an absence of evidence.

Dystrophin and Animal Models of Muscular Dystrophy

Dystrophin is remarkably conserved across the evolutionary spectrum. Dystrophin homologs have been identified in metazoans (*Drosophila* and *C. elegans*), zebra fish, shark, fugu, *Torpedo*, *Xenopus*, rodents (mice and rats), dogs, cats, horses, and primates. There is an actin-binding protein or gelation factor with high homology to dystrophy in *Dictyostelium*. Components of the dystrophin complex are also found in *C. elegans*, Annelida (leech), *Drosophila*, *Xenopus*, *Torpedo*, and zebra fish, as well as in more complex organisms. Homologs of the LARGE glycosyltransferase have been reported in several vertebrates, including mouse, dog, chicken, zebra fish, and fugu. Nonetheless these homologs do have alterations at the nucleic acid and amino acid levels (and these increase with evolutionary distance), and as a practical matter specific reagents may or may not recognize a specific homolog across species boundaries.

It follows that naturally occurring or targeted mutations in each of these species will provide potentially useful models of the dystrophic myopathies. However the histologic dystrophic myopathy may be difficult to recognize, or may not occur in species more remote evolutionarily from vertebrates. Additionally, some models, notably the murine models, precisely model the genetics and molecular biology of human dystrophy but are limited in modeling the histologic or pathologic sequelae. This is in itself informative, but indicates that careful consideration must be given to choosing an appropriate model system for a particular question or study. Certain models, such as *Torpedo*, are particularly apt for study of the role of the NMJ, but only the canine model provides a realistic equivalent of human disease for testing functional reconstitution or compensation in a potential therapy, and zebra fish, *C. elegans*, or *Drosophila* may be most useful for screening approaches.

Mouse Models

Far and away the most important series of animal models of dystrophin loss, or loss of the DAG complex, have been the mouse models. Of these, the first and most important is the *mdx* mouse, a naturally occurring mutant which was identified as a model of Duchenne dystrophy based on genetics and histology, well before the identification of dystrophin. The *mdx* mouse shows a maximum of necrosis and regeneration before 3 months of age and active damage declines thereafter, although internalized nuclei remain increased. The mice have normal appearance, size, and life span and they are susceptible to increased damage with exercise. There is damage to the diaphragm that is comparable to that seen in human DMD, although respiratory function is apparently uncompromised; CPK is elevated. In addition to findings directly related to skeletal muscle organization and degeneration, the synaptic folds at NMJs were found to be flattened.

Soon after the identification of dystrophin, it was recognized that the *mdx* mouse is a dystrophin-null mutant with a stop codon in exon 23. Among other results, an upstream stop codon is quite amenable to exon-skipping mechanisms, and the *mdx* mouse has a high rate of rare, immunocytochemically detected 'revertant' (normally localized dystrophin-positive) fibers. This phenomenon increases with age, or with exposure of the muscle to high levels of ionizing radiation.

In short, the *mdx* mouse is an excellent genetic and molecular biological model of DMD, a poor

physiological model (although biophysical measures of sarcolemmal parameters are probably valid), and a very inadequate model of the pathology of human DMD. Some improvements in the *mdx* model have been made by the 'mdx^{cv}' series of dystrophin mutants. These strains are the product of a mutagenesis screen, and represent a series of mutations spread down the length of the dystrophin molecule. Some, such as mdx^{4cv} and mdx^{5cv}, have deletions that are less prone to revertant fibers, a technical advantage in monitoring an intentional restoration of dystrophin expression. Another mutation, mdx^{3cv}, is a leaky splice mutation, but due to its location at exon 66 it lowers the level of all seven of the dystrophin gene transcripts. Researchers have used other technologies to generate additional target mutations of the dystrophin gene. Series of such mutants have been useful for attempting to establish genotype–phenotype correlations between specific dystrophin mutations and particular molecular genetic or physiologic abnormalities.

Mouse models of DAG complex mutations All of the components of the human DAG complex and the full repertoire of associated proteins and short dystrophin isoforms, utrophin, and the utrophin isoforms are found in the mouse. The corollary targeted mouse null mutants of many of these molecules have been made and are frequently available for research.

Targeted mutations of all four core sarcoglycans, as well as ε-sarcoglycan, have been created. The mutant mice show a range of phenotypes, including cardiomyopathy (δ-sarcoglycan). The attempt to produce a mouse deficient in dystroglycan led to the first insight that this was a very early embryonic lethal mutation, but subsequent strategies, primarily Cre–*lox* technology, have permitted the tissue-specific deletion of dystroglycan. The *myd* mouse has a mutation of the LARGE gene, a glycosyltransferase gene that recapitulates many of the biochemical and histologic features of POMT1 mutations. A naturally occurring murine model, the *dy/dy* mouse, has a defect in merosin, with severe pathology similar to merosin-negative congenital muscular dystrophy (MDC1A). At least four (two naturally occurring, two engineered) dysferlin-null mouse strains have been reported. Like the dystrophin-null mice, the dysferlin-null mice have considerably less pathology than is seen in the human equivalent, Miyoshi distal myopathy and LGMD2B, but are important molecular biological surrogates.

Dystrophin-Deficient Feline; Lethal Muscle Hypertrophy

Several strains of domestic cat are dystrophin deficient and exhibit lethal muscular hypertrophy; these cats have decreased activity and stiff gait and run abnormally. Histologically they have a dystrophic appearance, with hypertrophy and dystrophic calcification. The cats succumb (they are euthanized for humane reasons) at 18 months to 2 years due to complications of hypertrophy of the pharyngeal muscles and diaphragm. At this stage fibrosis is not prominent. This defect has been termed 'hypertrophic feline muscular dystrophy' (HFMD).

Utrophin, α-sarcoglycan, and β-dystroglycan are present in feline muscle. Merosin (laminin $\alpha 2$) deficiency has been reported in several strains of American and European cats and is manifest as muscular dystrophy and neuropathy similar to that seen in human MDC1A.

Dystrophin and Muscular Dystrophy in Dogs

A golden retriever and a beagle with X-linked dystrophin deficiency have been identified and maintained as breeding colonies. The offspring exhibit elevated CPK from birth. Diaphragmatic weakness is present at birth and is more severe than limb muscle involvement. As they grow, the dogs exhibit gait disturbance, limb and temporal muscle weakness, and eventually macroglossia, dysphagia, drooling, and contracture of the mandibular joint. There is hypertrophy of the sublingual muscles by 1 year of age, similar to the cat model. The dystroglycan gene has been characterized in these dogs.

Dystrophin and DAG in Zebra Fish

The zebra fish, a vertebrate with many desirable features for genetic manipulation and molecular biological analysis, has been shown to have homologs of dystrophin, as well as the components of the DAG complex. Strains have been identified with mutations that have been mapped to the dystrophin gene, and these fish have abnormal skeletal muscles with progressive degeneration and short survival. Phenotypically similar strains will likely prove to have mutations in components of the DAG complex. In addition, it is possible to reduce expression of these proteins using morpholino technology. The zebra fish may be the simplest organism which has a muscle phenotype that is reasonably comparable to that seen in tetrapods.

Dystrophin and DAG in *Drosophila*

Drosophila melanogaster has been shown to have homologs of dystrophin, several of the smaller dystrophin isoforms, and the core components of the dystrophin-associated complex: four sarcoglycans, dystroglycan, and sarcospan. Disruptions of either the dystrophin homolog or the dystroglycan homolog

produce decreased mobility and progressive degeneration of muscle. In addition, there are abnormalities of cell polarity and defective photoreceptor development.

Dystrophin and Muscular Dystrophy in *C. elegans*

Caenorhabditis elegans has a dystrophin Dys-1 like protein, expressed in muscle. It also has homologs of the entire DAG complex, and dystrobrevin. Mutant strains have been identified with truncated Dys-1 protein. The mutants are hyperactive, but muscle appears to be structurally normal. The physiologic abnormality is thought to reflect alterations in cholinergic transmission. A mutation in a homolog of the dystrophin-associated protein dystrobrevin, called dyb-1, produces a similar phenotype.

Several interesting observations have come out of the phylogenetic perspective on dystrophin, in addition to the practical experimental aspects. As might be expected, the onset of manifestations of dystrophin deficiency increases in proportion to the longer development and life span of species. Perhaps of more significance, the severity of dystrophic pathology in skeletal muscle appears to increase with species size. Only the canine model fully recapitulates the later stages of histologic myofiber degeneration and musculoskeletal disease seen in humans. Mice, in contrast, have normal life span, and unless stressed essentially are free of overt musculoskeletal disease. It seems likely that fiber length and biomechanical factors are paramount, but the precise relationship has not been defined.

It is also apparent that dystrophic myopathy and indeed muscular manifestations change dramatically in the simpler organisms. The predominant phenotype is abnormality at the NMJ, synaptic vesicle trafficking, and cell polarity at an embryonic stage. Nonetheless, the underlying biochemical process may be informative and the facility of genetic manipulations may make these apt models for specific studies.

Conclusion

Insight into the severe progressive myopathies known as muscular dystrophies, of which the archetype is DMD, has evolved from two convergent lines of study. The underlying classification has been put on a precise rational basis by progress from subjective phenotypic (clinical) characterization to classification based on the Mendelian inheritance of mutations at specific chromosomal loci. This has been complemented by the molecular genetic identification first of dystrophin, and then of a series of interacting dystrophin-associated proteins. In several instances these have matched the genetic loci identified in affected families. Positional cloning in affected families has identified other proteins. Out of this a model has emerged of a cytoskeletal framework which extends from the actin cytoskeleton, through dystrophin, to the integral membrane protein β-dystroglycan and the peripheral membrane protein α-dystroglycan, and which is anchored to the basement membrane component merosin, the α_2 chain of laminin. The complex of four sarcoglycans (glycosylated integral membrane proteins) is an intrinsic component of this complex and is critical for its function, although the precise mechanism is unclear. Further study of this core complex has identified several more proteins which, when mutated, are responsible for additional forms of muscular dystrophy, several because they have a critical role in glycosylating dystroglycan, without which dystroglycan is not functional. Yet the proteins responsible for several forms of muscular dystrophy remain to be identified, and the pathogenetic mechanism for several of the identified proteins remain to be understood.

A host of questions remain to be answered about the role of 'muscular dystrophy' proteins in the non-muscle tissue in which they occur, and the additional phenotypic features when they are lost or nonfunctional. A wide range of organisms have homologs of dystrophin and the dystrophin-associated proteins, and provide to varying degrees genetic, molecular biological, and pathological models of the muscular dystrophies. These provide a wide array of situations in which to explore normal and abnormal function and to test potentially therapeutic strategies.

See also: Cytoskeletal Interactions in the Neuron; Cytoskeleton in Plasticity; Demyelinating Diseases; Demyelination and Demyelinating Antibodies; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Transplantation of Myelin Forming Cells.

Further Reading

Ahn AH and Kunkel LM (1993) The structural and functional diversity of dystrophin. *Nature Genetics* 3: 283–291.

Barresi R and Campbell KP (2006) Dystroglycan: From biosynthesis to pathogenesis of human disease. *Journal of Cell Science* 119: 199–207.

Chamberlain JS, Slightom JL, Adams EG, et al. (1996) Differential expression of dystrophin isoforms in strains of mdx mice with different mutations. *Human Molecular Genetics* 5: 1149–1153.

Dalkilic I and Kunkel LM (2003) Muscular dystrophies: Genes to pathogenesis. *Current Opinion in Genetics & Development* 13: 231–238.

Ervasti JM, Ohlendieck K, Kahl SD, et al. (1990) Deficiency of a glycoprotein component of the dystrophin complex in dystrophic muscle. *Nature* 345: 315–319.

Guyon JR, Steffen LS, Howell MH, et al. (2007) Modeling human muscle disease in zebrafish. *Biochimica et Biophysica Acta* 1772: 205–215.

Haenggi T and Fritschy JM (2006) Role of dystrophin and utrophin for assembly and function of the dystrophin glycoprotein complex in non-muscle tissue. *Cellular and Molecular Life Sciences* 63: 1614–1631.

Hoffman EP, Hudecki MS, Rosenberg PA, et al. (1987) Cell and fiber-type distribution of dystrophin. *Neuron* 1(5): 411–420.

Kang PB and Kunkel LM (2006) Muscular dystrophies. In: Scriver CR, Beaudet AL, Sly WS, et al. (eds.) *The Online Metabolic and Molecular Bases of Inherited Disease*, ch. 216. New York: McGraw-Hill.

Koenig M, Beggs AH, Moyer M, et al. (1989) The molecular basis for Duchenne versus Becker muscular dystrophy: Correlation of severity with type of deletion. *American Journal of Human Genetics* 45: 498–506.

Koenig M, Hoffman EP, Bertelson CJ, et al. (1987) Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* 50: 509–517.

Lennon NJ, Kho A, Bacskai BJ, et al. (2003) Dysferlin interacts with annexins A1 and A2 and mediates sarcolemmal wound-healing. *Journal of Biological Chemistry* 278: 50466–50473.

Lidov HGW (1996) Dystrophin in the nervous system. *Brain Pathology* 6: 63–77.

Longman C, Brockington M, Torelli S, et al. (2003) Mutations in the human LARGE gene cause MDC1D. A novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of alpha-dystroglycan. *Human Molecular Genetics* 12: 2853–2861.

Moore SA, Shilling CJ, Westra S, et al. (2006) Limb-girdle muscular dystrophy in the United States. *Journal of Neuropathology & Experimental Neurology* 65: 995–1003.

Ozawa E, Mizuno Y, Hagiwara Y, et al. (2005) Molecular and cell biology of the sarcoglycan complex. *Muscle & Nerve* 35: 563–576.

Ozawa E, Noguchi S, Mizuno Y, et al. (1998) From dystrophinopathy to sarcoglycanopathy: Evolution of a concept of muscular dystrophy. *Muscle & Nerve* 21: 421–438.

Shcherbata HR, Yatsenko AS, Patterson L, et al. (2007) Dissecting muscle and neuronal disorders in a *Drosophila* model of muscular dystrophy. *EMBO Journal* 26: 481–493.

Relevant Websites

http://www.dmd.nl – Leiden Muscular Dystrophy pages.

http://genetics.accessmedicine.com – McGraw-Hills Access Medicine.

http://www.neuromuscular.wustl.edu – Neuromuscular Disease Center.

Eating Disorders

W Kaye, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Anorexia nervosa (AN) and bulimia nervosa (BN) are disorders that tend to occur in women and typically have an onset during adolescence but can occur in males as well as at other ages. Both AN and BN are often chronic, disabling conditions characterized by aberrant patterns of feeding behavior and weight regulation, and deviant attitudes toward and perceptions of body weight and shape. The etiology of AN and BN is presumed to be influenced by developmental, social, and biological processes; however, understanding of its etiology and pathophysiology remains incomplete. AN and BN are often considered to be a consequence of cultural attitudes toward standards of ideal body shape. However, although dieting behavior and the drive toward thinness are common in industrialized countries, AN and BN affect only an estimated 0.3–0.7% and 1.7–2.5%, respectively, of females in the general population. Moreover, numerous descriptions of a syndrome reminiscent of AN date from the late seventeenth century or earlier, suggesting that factors other than current culture play an etiologic role. In addition, these syndromes, particularly AN, have a relatively stereotypic clinical presentation, sex distribution, and age of onset, underscoring the role of intrinsic biologic vulnerability and diagnostic validity of the syndrome.

Description

AN is divided into two subtypes. In the restricting subtype, subnormal body weight and an ongoing malnourished state are maintained by unremitting food avoidance. In the bulimic subtype of AN, there is comparable weight loss and malnutrition, yet the course of illness is marked by supervening episodes of binge eating, usually followed by some type of compensatory action such as self-induced vomiting or laxative abuse. Individuals with BN remain at normal body weight, although many aspire to ideal weights far below the range of normalcy for their age, gender, and height. The core features of BN include repeated episodes of binge eating followed by compensatory self-induced vomiting, laxative abuse, or pathologically extreme exercise or fasting, as well as excessive concern with weight and shape. Although abnormally low body weight is an exclusion for the diagnosis of BN, 25–30% of bulimics have a prior history of AN. In addition, some individuals with AN convert to BN. Finally, persons afflicted with AN and BN, despite differences in the topography of their feeding behavior, share equivalent concerns with weight and shape. Likewise, common to individuals with AN and BN are low self-esteem, depression, anxiety, obsessionality, and perfectionism.

Core Symptoms

AN and BN are syndromes consisting of disturbances of a range of behaviors, including feeding, mood, impulse control, body image perception, as well as other domains. The most striking features are related to eating and body image. In AN, an inexplicable fear of weight gain and unrelenting obsession with fatness, even in the face of increasing cachexia, accounts for a protracted course, extreme medical and psychological morbidity, and standardized mortality rates exceeding those of all other psychiatric disorders. People with AN have stereotypic patterns of food choices. They severely restrict food intake, particularly fats and carbohydrates, but rarely stop eating completely; rather, they restrict their caloric intake to a few hundred calories per day. They tend to be vegetarians, have monotonous choices in food intake, select unusual combinations of foods and flavors, and have ritualized eating behaviors. BN usually emerges after a period of food restriction, which may or may not have been associated with weight loss. Binge eating is followed by either self-induced vomiting or some other means of compensation for the excess of food ingested. The majority of people with BN have irregular feeding patterns, and satiety may be impaired.

It has long been debated whether symptoms in individuals with AN and BN are cause or consequence of malnutrition. Confounding this understanding is the reality that most studies of symptoms have been done when individuals are ill. Particularly common in individuals with AN are personality traits of marked perfectionism, conformity, obsessionality, constriction of affect and emotional expressiveness, and reduced social spontaneity. Studies have shown that the majority of individuals with AN and BN exhibit childhood perfectionism, obsessive–compulsive personality patterns, and anxiety (particularly obsessive–compulsive disorder and social phobia), and these symptoms predate the onset of AN and BN. Because

these characteristics typically are present in childhood, prior to the onset of the eating disorder, and persist even after long-term weight recovery, they may be traits and not merely epiphenomena of acute malnutrition and disordered eating behavior. Individuals with BN are also more likely to exhibit histories of behavioral dyscontrol, substance abuse, and overt family conflict in comparison to those with the restricting subtype of AN.

It is important to note that other 'core' symptoms are common in AN and BN. One of the most puzzling symptoms of AN is the severe and intense body image distortion in which emaciated individuals perceive themselves as fat. Individuals ill with AN in particular have stereotyped hyperactive motor behavior, anhedonic and restrictive personalities, reduced novelty seeking, and a lower incidence of alcohol and drug use. Neuropsychological findings, using a broad range of assessments, have been somewhat inconsistent. A more focused, hypothesis-driven approach, pioneered by Tchanturia, Treasure, and others, has been based on the observation that people with AN show, as core features, high harm avoidance, inflexibility of thinking, rigid behaviors around eating, and high perfectionism. In addition, there is inflexibility on cognitive tasks and an inability to change past patterns of thinking, which can be conceptualized as difficulties in set shifting. These symptoms are evident in individuals with AN both during the acute phase of the illness and after weight restoration. In addition, such difficulties in set shifting are shared by the healthy sisters of those with AN, raising the intriguing possibility that set shifting represents a potential endophenotype for genetic studies of AN. Individuals with AN show a rigid pattern of responding and an inability to change a pattern of response once a behavior has been initiated. They also seem to have difficulty accessing and using conventional strategies for problem solving, such as learning from making mistakes. Rather, they obsessively repeat the same strategies despite the fact that such strategies are maladaptive and are not productive. Moreover, they often fail to accurately recognize and incorporate affective and social stimuli in the environment. Most strikingly, denial of emaciation or being ill, lack of insight, and resistance to treatment are often present. These characteristic styles raise the question of whether there is some physiologic disturbance of executive brain mechanisms that detect errors or plan and verify actions.

Familial/Genetic

Findings from case–control studies suggest a 7- to 12-fold increase in the prevalence of AN and BN in female relatives of eating-disordered probands compared to relatives of never ill controls. Despite the phenotypic differences between AN and BN, the two disorders appear to share at least some liability features in common. Convergent findings from family and twin studies indicating an increased coaggregation of AN and BN among relatives of AN and BN probands argue that a confluence of shared and unique liability features characterize the broader spectrum of eating disorder phenotypes. This significant familial occurrence of AN and BN provides impressive evidence of the heritability of both phenotypes.

Twin studies can differentiate genetic from environmental effects by comparing concordance for a trait, or disorder, between identical (monozygotic; MZ) and fraternal (dizygotic; DZ) twins. Although relatively few twin studies of AN and BN have been performed, they consistently show that approximately 50–80% of the liability to eating disorders is genetic. The transmitted liability to eating disorders may possibly be underpinned by a more diffuse phenotype of continuous traits. Specifically, evidence from population-based twin studies indicates heritability of disordered eating attitudes, weight preoccupation, dissatisfaction with weight and shape, dietary restraint, binge eating, and self-induced vomiting, with 32–72% of the variance in these behavioral features potentially accounted for by genetic factors. In addition, subthreshold forms of eating disorders have also been shown to coaggregate in families with full syndrome AN and BN, consistent with the possibility of a continuum of transmitted liability in at-risk families manifesting a broad spectrum of eating disorder phenotypes.

Identification of specific genes that play a role in AN and BN is in its infancy. Still, several studies of AN have replicated findings of a region on the short arm of chromosome 1 and potential involvement of the serotonin (5-HT) 1D (HTR1D) and opioid delta 1 (OPRD1) receptor genes in this region. In addition, traits thought to be related to liability to eating disorders, such as drive for thinness, obsessionality, anxiety, and concern over mistakes, have been associated with a number of linkage signals in BN and AN. A growing number of studies using a case–control design have examined the potential association of various candidate genes with AN and BN. Evidence linking AN and BN to monoamine functioning has led researchers to target 5-HT and dopamine (DA)-related genes, but findings have tended to be inconsistent or negative. In addition, the primary role of feeding, energy expenditure, or gender in the pathology of AN and BN has led researchers to examine genes related to these processes. In summary, no consistent evidence for association of a candidate gene

with either AN or BN has accrued to date, although inadequate sample sizes severely constrain these efforts. Additional research is necessary to clarify conflicting findings and replicate initial results.

Pathophysiology

There is growing acknowledgment that neurobiological vulnerabilities make a substantial contribution to the pathogenesis of AN and BN. Still, we have little understanding of how such vulnerabilities result in disturbances of brain pathways or what systems are primarily involved. For example, are there disturbances of pathways related to the modulation of feeding behaviors, mood, temperament, obsessionality, or impulse control? Are there primary disturbances of pathways that may modulate some factors related to body proprioception and thus result in body image distortions? The past decade has seen the introduction of tools, such as brain imaging, which hold the promise of being able to characterize complex neurocircuits and their relationship to behavior in living humans. In fact, these tools have rapidly advanced knowledge to the point where we can begin to make educated guesses about the pathophysiology of AN and BN and start to model mechanisms that may be used to test hypotheses.

In general, imaging studies have been relatively consistent in that many studies have findings in frontal, cingulate, temporal, and/or parietal regions. Thus, it can be stated that individuals with eating disorder, when ill and after recovery, have alterations in brain activity compared to matched controls. However, it should be noted that these studies have not consistently identified regions, pathways, or behavioral correlates. Sample sizes have been small, and imaging technologies and methods vary widely. Moreover, investigations have tended to assess relatively large regions of brain that vary widely between studies. Moreover, articles to date indicate alterations of brain function. Because brain pathways are highly complex, the neuroanatomy of AN and BN has only begun to be characterized.

When ill, people with AN, and to a somewhat lesser extent BN, commonly have neuroendocrine, autonomic, and metabolic disturbances and an exaggeration of comorbid psychiatric symptoms. Neuropeptides have been of interested because they could play a role in disturbances of appetite regulation, mood, impulse control, endocrine function, and many of the other physiological disturbances. Determining whether such symptoms are a consequence or a potential cause of pathological feeding and other altered behavior is a major methodological issue in the field. It is difficult to study AN prospectively due to the young age of onset and difficulty in premorbid identification of people who will develop AN. An alternative strategy is to identify behavioral phenotypes that are independent of the confounding effects of malnutrition by studying women who have recovered from AN and BN. The majority of neuroendocrine and neuropeptide alterations present during symptomatic episodes of AN and BN tend to normalize after recovery. These neuropeptides include corticotropin-releasing hormone, opioid activity, neuropeptide Y, and leptin. This observation suggests that most of the disturbances are consequences rather than causes of malnutrition, weight loss, and/or altered meal patterns.

In contrast, considerable studies have reported disturbances of brain 5-HT and DA function during the ill state and persisting after recovery. Studies using positron emission tomography (PET) brain imaging with [^{18}F]altanserin binding potential (BP) have characterize the 5-HT2A receptor. This receptor is of interest because it has been implicated in the regulation of feeding, mood, and anxiety and in antidepressant action. Ill and recovered AN and BN patients have reduced [^{18}F]altanserin BP in limbic regions. Several groups have assessed the 5-HT1A receptor in AN because this receptor may play a role in anxiety, mood and impulse control, feeding behavior, as well as in selective serotonin reuptake inhibitor (SSRI) response. PET imaging with the radioligand [^{11}C]WAY100635 found increased 5-HT1A binding in individuals who were ill and recovered from BN, whether or not they ever had AN. Finally, several lines of evidence support the possibility that disturbances of DA function could contribute to alterations of weight, feeding, motor activity, and reward in AN. PET imaging with [^{11}C]raclopride to assess DA D2/D3 receptor binding showed recovered AN patients had significantly higher [^{11}C]raclopride BP in the anteroventral striatum. Importantly, these studies have consistently found relationships between receptor binding and harm avoidance, a construct which is highly elevated in recovered eating-disorder individuals and characterizes overconcern about future events. Together these studies support earlier findings of persistent 5-HT and DA function in AN and BN.

Studies have used functional magnetic resonance imaging (fMRI) to assess symptoms. Studies have used either pictures of food or tastes of sugar and other substances, and they have found persistent alterations of brain pathways in individuals recovered from AN. These studies have implicated frontal as well as the insular cortex and related regions, pathways known to modulate feeding behavior. Brain imaging studies done in the past decade in human and nonhuman primates show that a primary taste cortex resides in

the rostral insula and adjoining frontal operculum. Some studies argue that these regions provide a representation of food in the mouth that is independent of hunger and, thus, of reward value. The responsiveness of taste neurons in secondary regions, such as the orbitofrontal cortex, may compute the hedonic value of food. These studies raise the provocative possibility of disturbances of higher order modulation of feeding in AN. Other fMRI studies have addressed body image distortions and have found altered response in left parietal regions, which have also been seen in PET studies. Theoretically, body image distortion might be related to the syndrome of neglect which may be coded in parietal, frontal, and cingulate regions that assign motivational relevance to sensory events. It is well-known that lesions in the right parietal cortex may not only result in denial of illness or anosognosia, and the numerous misidentification syndromes but also produce experiences of disorientation of body parts and body image distortion. Several brain imaging studies have reported altered left parietal cortex function. It is intriguing to raise the possibility of whether left hemisphere disturbances of this pathway contribute to body image distortion.

Treatment

A number of consensus publications have reviewed treatment options for AN and BN. These include the American Psychiatric Association guidelines and the National Collaborating Centre for Mental Health guidelines. These publications should be consulted for more specific details about therapy. In addition, websites such as that of the National Eating Disorders Association can be accessed for more information and referrals for treatment.

Bulimia Nervosa

Controlled outpatient trials have provided convincing evidence for the efficacy of both psychological and pharmacological treatments for BN. The largest number of controlled trials have focused on manual-based cognitive behavioral treatment (CBT). These studies have shown a decrease in the frequency of binge eating and purging episodes of approximately 60–70%, with complete abstinence from these behaviors in 30–50% of cases. There is also evidence that CBT results in improvements in other psychological symptoms of the disorder, including preoccupation with body shape and weight, and that CBT is more effective than more narrow behavioral approaches in addressing bulimic symptoms. Other data suggest that manual-based interpersonal therapy may achieve reduction in symptoms similar to results achieved with CBT, although the rate of response may be somewhat slower. Preliminary reports indicate that structured self-help approaches may be of assistance to some individuals, consistent with suggestions regarding the potential usefulness of 'stepped-care' algorithms in treatment planning.

The efficacy of a wide variety of antidepressant medications in the treatment of BN has been demonstrated in a number of double-blind, placebo-controlled trials. The average decrease in binge frequency for patients receiving the antidepressant medication was approximately 55%, approximately twice the widely variable rate of response for patients randomized to placebo administration. Only a minority of patients in the medication treatment trials actually achieved full abstinence from binge eating and purging behaviors. Whereas antidepressant administration may contribute to reduction in symptoms of depression, the reduction in bulimic symptoms does not appear to be correlated with the magnitude of the antidepressant response. With similar response rates seen across different classes of antidepressants, initial choice of medication is often based on a range of clinical considerations including typical side effect profile.

Several studies have compared response to psychotherapy and pharmacotherapy or to combination treatment. In general, these trials have demonstrated a greater decrease in frequency of bulimic episodes for psychotherapy (usually CBT) in comparison to pharmacotherapy. Although combined treatment was not notably superior to either treatment alone, this approach may be helpful in decreasing symptoms of depression, anxiety, and eating-related symptoms in some individuals. Currently, there are no established predictors of which patients might respond preferentially to psychotherapy or pharmacotherapy, although it has been suggested that individuals who show little response during the early sessions of CBT may be candidates for alternate treatment interventions.

Anorexia Nervosa

Progress in developing new and more effective treatment interventions for patients with AN has been slower than for BN. Individuals with AN are often resistant to treatment because they tend to have ego syntonic symptoms or may frankly deny being underweight or in need of treatment. Few control trials of any therapy have been performed, in part because it has been difficult to enlist cooperation of individuals with AN and in part because psychological and pharmacological strategies that have been successful in other disorders appear to be less effective in this

illness. For severely emaciated patients, hospitalization for supportive medical care and weight restoration may be useful or necessary. Still, relapse is common after discharge. Controlled trials suggest that the Maudsley family therapy techniques appear to be particularly helpful in younger patients with AN. Follow-up studies suggest that these interventions may also help prevent relapse in patients who achieve weight restoration. Studies have begun more systematic evaluation of the potential benefits of manualized cognitive behavioral and family therapies for this disorder. A range of medications, including SSRIs and neuroleptics, have been tried in AN, but there are few controlled trials in ill, malnourished patients. Some data have raised the possibility that fluoxetine may be useful in reducing relapse after weight restoration particularly in those with restricting-type AN compared to bulimic-type AN.

Summary

The etiology of AN and BN remains unknown. It may be that individuals with AN and BN may have a genetically transmitted alteration of neurotransmitter function that contributes to anxiety and obsessional symptoms in childhood. Perhaps brain changes caused by pubertal female gonadal steroids contribute to triggering the onset of an eating disorder. It is possible that people who develop eating disorders may respond in an overanxious manner to stimuli or have exaggerated overconcerns about the future. Because of these temperament traits, and perhaps traits related to disturbed body image, people with eating disorders may be more vulnerable to focusing on and incorporating certain societal messages, or they may be particularly susceptible to stress. It may be that people with AN and BN find that extremes of food intake produce neurochemical changes that temporarily reduce dysphoric moods, and thus they are prone to restrictive eating for AN and overeating for BN.

See also: Anxiety Disorders; Depression and the Brain.

Further Reading

American Psychiatric Association (2006) *Practice Guideline for the Treatment of Patients with Eating Disorders*, 3rd edn. Arlington, VA: American Psychiatric Association.

Bulik CM (2005) Exploring the gene–environment nexus in eating disorders. *Journal of Psychiatry Neuroscience* 30(5): 335–339.

Jimerson DC and Wolfe BE (2004) Neuropeptides in eating disorders. *CNS Spectrums* 9(7): 516–522.

Kaye W, Strober M, and Jimerson D (2004) The neurobiology of eating disorders. In: Charney DS and Nestler EJ (eds.) *The Neurobiology of Mental Illness*, pp. 1112–1128. Oxford University Press: New York.

Kaye WH, Wagner A, Frank GK, and Bailer UF (2006) Review of brain imaging in anorexia and bulimia nervosa. In: Mitchell J, Wonderlich S, Steiger H, and de Zwaan M (eds.) *AED Annual Review of Eating Disorders Part 2*, pp. 113–130. Northbrook, IL: Academy for Eating Disorders.

Klein DA and Walsh BT (2003) Eating disorders. *International Review of Psychiatry* 15(3): 205–216.

Klump KL and Gobrogge KL (2005) A review and primer of molecular genetic studies of anorexia nervosa. *International Journal of Eating Disorders* 37: S43–S48 (discussion S87–S89).

Lock J and le Grange D (2005) Family-based treatment of eating disorders. *International Journal of Eating Disorders* 37: S64–S67 (discussion S87–S89).

National Collaborating Centre for Mental Health (2004) *Core Interventions in the Treatment and Management of Anorexia Nervosa, Bulimia Nervosa and Related Eating Disorders*. London: National Institute for Clinical Excellence.

Tchanturia K, Campbell IC, Morris R, and Treasure J (2005) Neuropsychological studies in anorexia nervosa. *International Journal of Eating Disorders* 37: S72–S76 (discussion S87–S89).

Treasure J and Schmidt U (2004) Anorexia nervosa. *Clinical Evidence* 11: 1192–1203.

Yager J and Andersen AE (2005) Clinical practice. Anorexia nervosa. *New England Journal of Medicine* 353(14): 1481–1488.

Relevant Websites

http://www.edreferral.com – Eating Disorder Referral and Information Center.

http://www.edap.org – National Eating Disorders Association.

Echolocation I: Behavior

A D Grinnell, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Echolocation is the active detection, localization, identification, tracking, and interception or avoidance of targets, based on information obtained from the echoes of emitted sounds. Several groups of animals have independently evolved this adaptation, but the behavior was first demonstrated and is most highly developed in bats of the suborder Microchiroptera. One genus of Megachiroptera (*Rousettus*), toothed whales, two genera of birds (*Steatornis*, the South American oil bird, and *Collocalia*, the cave swifts), and shrews are known to echolocate. Echolocating animals other than Microchiropterans usually have excellent night vision and use echolocation facultatively.

Study of echolocation began in the late 1700s. Spallanzani, in Italy, and his Swiss counterpart, Jurine, showed that blinded bats could orient, avoid obstacles, and feed while flying, and that hearing was necessary for these capabilities. Not until 1938, however, did Donald Griffin, then an undergraduate at Harvard, establish that flying bats emit trains of brief, high-frequency sounds and use their echoes to detect and avoid obstacles. The study of echolocation has since become one of the most challenging and fascinating areas of neuroethology.

With their remarkable adaptations for flight and echolocation, bats are among the most diverse and successful orders of mammals. They evolved echolocation more than 50 Ma and have since differentiated widely. Echolocating bats inhabit all climatic zones in all continents except Antarctica. They range in weight from <2 to >100 g and in wing span from <15 cm to almost 1 m. About 70% are insectivorous, catching flying insects in their tail or wing membranes at rates up to 2 s^{-1}. Prey range in size from mosquitoes to large moths and beetles. Some species eat small birds or other bats: others lap blood from roosting birds and large mammals. Some use both prey-generated sounds and echolocation to glean vegetation or the ground for prey like cicadas, walking sticks, crickets, katydids, and scorpions. At least one genus of bats eats frogs and several catch fish or floating insects by skimming the water surface where they have detected movement. Many bats feed on fruits. Others are specialized for nectar feeding, with elongated snouts, long tongues, and the ability to hover (Figure 1).

All bats must also detect and avoid obstacles such as tree branches, leaves, wires, and buildings; drink by flying low over a body of water, dipping in just the lower jaw; land by flipping around at the last moment to grip a surface irregularity with their hind claws; and find their way (probably with the aid of vision and olfaction) throughout a home territory that may encompass many square miles. To accomplish these feats, bats must simultaneously localize many echo sources in three-dimensional space and discriminate the shapes, dimensions, and surface textures of each. In short, bats transform the cacophony of echoes they encounter when flying in their nocturnal environments into a series of acoustic three-dimensional 'snapshots' of the environment approaching the detail obtained visually by day-flying insectivorous birds.

Echolocation Sounds and Foraging Niches

Within any given habitat, different bat species exploit different niches. Independent of their often quite distant phylogenetic relationships, bats hunting in similar acoustic environments use similar repertoires of sounds that differ depending on the demands of each environment. A major determining factor is the amount of echo 'clutter' interference the bat faces while foraging. Although all bats must be able to fly and orient in cluttered space, most tend to hunt selectively (1) in open air (no clutter), (2) near borders between open space and foliage, water surface, or the ground (background clutter), or (3) in dense vegetation (highly cluttered space). Each niche is further refined by the size of prey; the minimum distance from clutter a bat can capture flying insects; whether prey can be gleaned from vegetation, water surface, or ground; and whether prey-generated sounds are required for detection. Bats exploiting these niches, or combinations thereof, apportion the resources of an environment and are said to form different guilds (Figure 2).

Search Phase Sounds and Detection

Bats use relatively simple, stereotyped signals, but there is considerable facultative flexibility in duration, frequency structure, intensity, and repetition rate, all of which vary in ways that optimize information retrieval in a given environment and stage of echolocation. The sounds, produced in the larynx, are mostly ultrasonic (>20 kHz) and only a few milliseconds long. Their intensity, measured 10 cm in front of the bat, ranges from <60 dB sound pressure level (SPL) ('whispering bats') to >130 dB SPL. Broadband signals, mainly downward frequency

Figure 1 Representative bats. (a) *Myotis velifer* (family Vespertilionidae), a bat that hunts in open air; (b) *Antrozous pallidus*, a gleaning bat, with prey; (c) *Rhinolophus ferrumequinum*, a long CF/FM-emitting bat that hunts in highly cluttered environments and often echolocates for prey from a perch; and (d) *Rousettus*, the only Megachiropteran that echolocates. Photographs by Merlin Tuttle, Bat Conservation International, Inc.

modulation (FM) sweeps through an octave or more, often with multiple harmonics to further broaden the bandwidth, can be used in any of the three clutter-defined niches. However, bats that hunt in open space have little concern about obstacles in their flight path so tend to optimize the range of detection by using loud, long (5–25 ms) sounds that concentrate the energy into a relatively narrow band of frequencies to which the bat is highly sensitive. These may have almost constant frequency (CF) or shallow FM sweeps (quasi-CF) (Figure 2, *Nyctalus noctula* and *Pipestrellus kuhlii*; see also Figure 1(a)). Search calls are typically emitted at rates of 2–10 s^{-1}, often in phase with wing beats.

In open air, insectivorous bats can detect medium-sized moths at 5–10 m. They dart at ~1 cm diameter artificial targets tossed into the air when they are within about 2 m (but almost invariably stop pursuit before intercepting the target). Echoes from small targets may be attenuated by 90 dB or more. This attenuation reflects the spherical spreading loss from the point of emission and again from the reflecting target (echo intensity falls with the third to fourth power of distance). For small targets echo intensity varies inversely with the fourth power of sound wavelength; hence, some bats that forage on small insects emit dominant frequencies as high as 150 kHz. However, atmospheric attenuation also

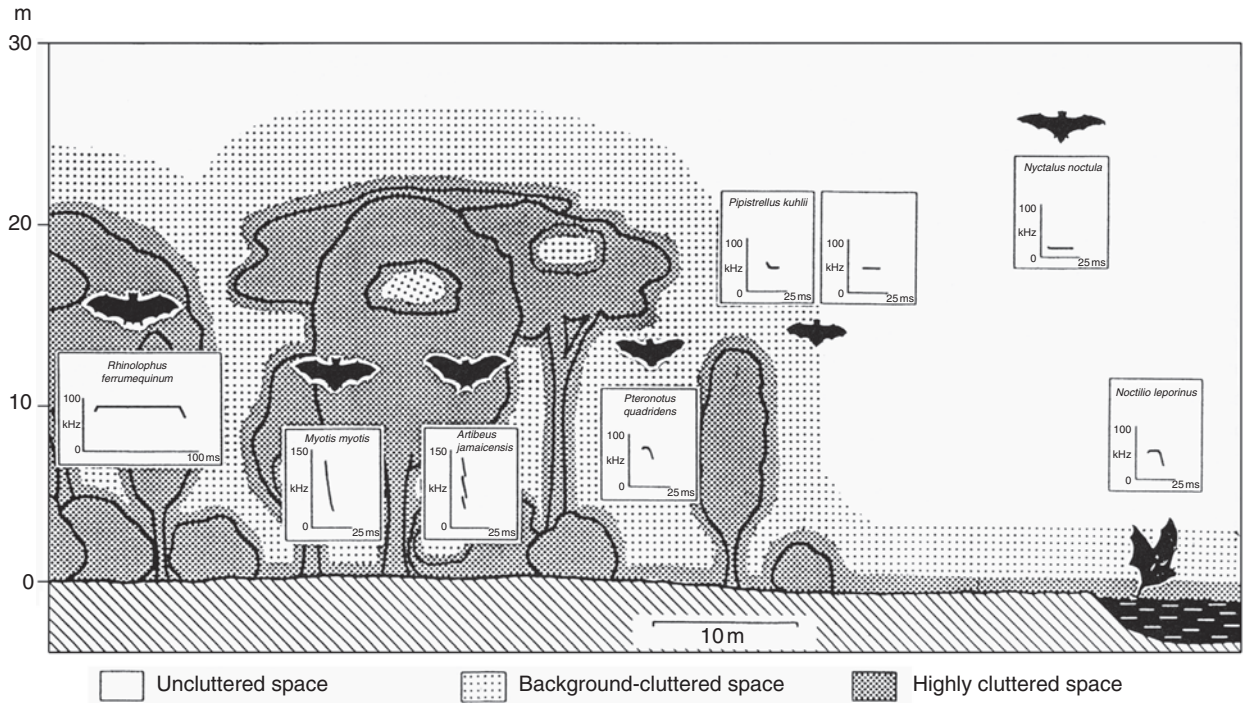

Figure 2 Schematic design showing different echolocation foraging niches and the flight silhouettes and signal designs used during search flight by representative species hunting in each niche. The niches are defined by the amount of acoustic clutter interference. In open air (no stippling), bats tend to be fast fliers and search with CF or quasi-CF signals, which are good for long-distance detection but not for localization. After detection, emitted sounds become broadband. Bats that hunt within about 5 m of the edge of the ground or vegetation or that trawl the water surface (light stipple) tend to use sounds with both CF and FM components, emphasizing FM as they face greater echo clutter. Finally, those that hunt amid vegetation, facing severe echo clutter, either use very short broadband signals or long CF/FM signals. The latter are specialized for 'flutter detection.' From Schnitzler H-U and Kalko EKV (2001) Echolocation by insect-eating bats. *BioScience* 51: 557–569, copyright, American Institute of Biological Sciences.

increases with frequency (5.6 dB m^{-1} at 150 kHz), restricting the range of echolocation in very high frequency bats.

Bats that hunt within about 5 m of vegetation or over water, where targets must be detected against background echo clutter, search with pulses that contain a prominent FM component, either alone or with a CF component (CF/FM or FM/CF) (e.g., Figure 2, *Pteronotus quadridens*, *Noctilio leporinus*). Bats that move from open air to background-cluttered space shift from CF or CF/FM to largely FM signals, and the reverse (Figure 2, *Pi. kuhli*). The CF component facilitates detection and the FM component allows localization of background obstacles.

Within each guild, further ecological distinctions are correlated with subtle differences in emitted pulses. For example, five sympatric European species of *Myotis* all use brief FM sweeps to hunt for insects in the presence of background clutter. The greater the bandwidth of the FM signals used by a given species, the closer to a cluttered background it can recognize and capture stationary tethered insects (from 2–3 cm with a 120 kHz bandwidth to 14 cm with a 40 kHz

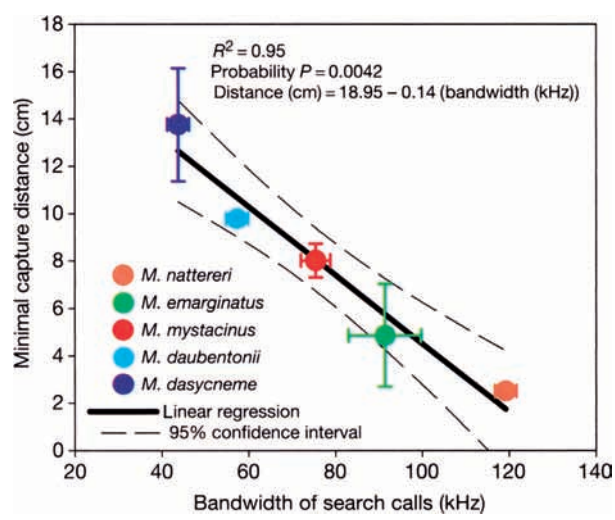

Figure 3 Correlation between the 'minimal capture distance' and the bandwidth of the FM search calls used by five sympatric *Myotis* species. Minimal capture distance was defined as the distance from a background 'clutter screen' at which 50% of tethered targets (mealworms) were captured within 1 min of searching flight. Mean ±SD for each species is indicated. From Siemers BM and Schnitzler H-U (2004) Echolocation signals reflect niche differentiation in five sympatric congeneric bat species. *Nature* 429: 657–661.

bandwidth, Figure 3). Gleaning bats typically use faint broadband signals. Often they use prey-generated sounds to detect and localize prey, but some bats can detect and capture quiescent insects on vegetation by echolocation alone.

In highly cluttered space, many bats that hunt on the wing use steep FM sweeps (e.g., Figure 2; *Myotis myotis* and *Artibeus jamaicensis*). However, others use long CF signals terminating in a brief FM sweep (Figure 2; *Rhinolophus ferrumequinum*). Long CF/FM bats are of particular interest because their auditory systems are highly specialized to utilize information in the CF component of echoes (see below).

Pursuit and Capture

After detection, bats must determine the distance, direction, and nature of the echo source(s). Obtaining this information requires broadband sounds. Almost universally, after detection, the rate and extent of the FM sweep increase, the repetition rate increases, and pulse duration decreases (Figure 4). Use of broadband FM sweeps allows redundant measurement of pulse–echo delay by many parallel frequency-tuned channels, thereby providing accurate information about distance. At the same time, binaural comparison of echo intensity at different frequencies yields

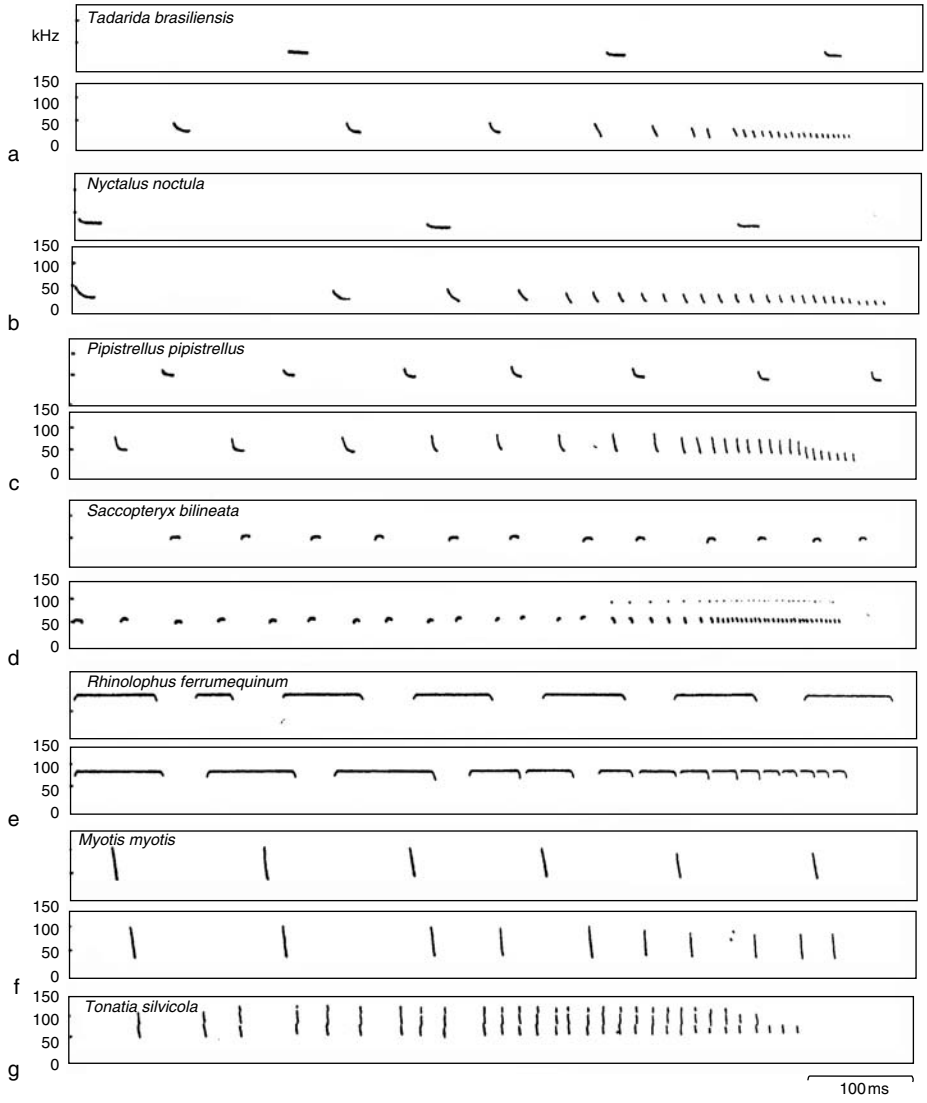

Figure 4 Sound spectrographic records of the emitted sounds of representative species as they search for, detect, pursue, and capture an insect in flight (a–e), and glean an insect from a surface after detection during flight (f) and after detection from a perch (g). *T. brasiliensis* and *N. noctula* are typical open-air foraging bats. *P. pipistrellus* and *S. bilineata* hunt close to the edge of foliage, and *R. ferrumequinum* hunt, often from perches, in a highly cluttered environment. Note the changes in signal structure and repetition rate associated with insect pursuit and capture in flight. From Schnitzler H-U and Kalko EKV (2001) Echolocation by insect-eating bats. *BioScience* 51: 557–569, copyright, American Institute of Biological Sciences.

information about direction, and the relative intensity of different frequencies in the echo and the location of spectral notches provide information about target shape. The decrease in emitted pulse duration appears intended to avoid overlap of outgoing and returning echo: a severe form of forward masking. Echoes from multiple objects in the environment do overlap, however, and each must be independently localized and identified.

Finally, as bats close in on a target, especially a flying insect, the repetition rate reaches 100–200 s^{-1} (the 'terminal buzz'). Pulse duration decreases to a fraction of a millisecond, and any remaining CF component tends to disappear. Intensity is reduced, and the starting frequency of the FM sweep often decreases (Figure 4(e)). The high pulse repetition rate during insect pursuit probably maximizes information acquisition for tracking and capturing moving targets. Such 'buzzes' are seldom observed during avoidance (or capture) of stationary targets (Figures 4(f) and 4(g)).

Long CF/FM Bats and Doppler-Shift Compensation

These bats, which hunt in cluttered environments, include members of the Old World families Rhinolophidae and Hipposideridae and one New World representative, *Pt. parnellii* (family Mormoopidae). The best-studied Rhinolophid, *R. ferrumequinum*, can hunt on the wing but often perches and hunts by acoustically scanning the environment (Figure 1(c)). Its search sounds have a CF component of about 83 kHz that can last 100 ms, ending in a brief FM sweep to ~70 kHz. *Pt. parnellii*, which always hunt in flight, emit a dominant CF component of ~61 kHz that lasts about 30 ms and terminates in an FM sweep to ~45 kHz. In both species, the CF is regulated to within ~50 Hz, or 0.06%. After the bats detect a target, the CF duration first increases to maximize the duration of the CF component of the echo. It then decreases as the repetition rate increases and the depth of the FM sweep increases (Figure 2(e)). Although the duty cycle (percentage of time sounds are emitted) may be close to 100%, only the CF component of echoes overlaps with outgoing sounds. This long CF is critical to their echolocation strategy.

Echoes returning from objects in front of a bat are Doppler-shifted upward in frequency in proportion to the bat's flight speed. Moving targets return echoes at a slightly different CF frequency than surrounding stationary objects. As echoes return upward-shifted in frequency, long CF/FM bats exhibit 'Doppler-shift compensation.' They lower their emitted CF just enough to maintain the returning echo at a constant 'reference' frequency (Figure 5). In both *Rhinolophus*

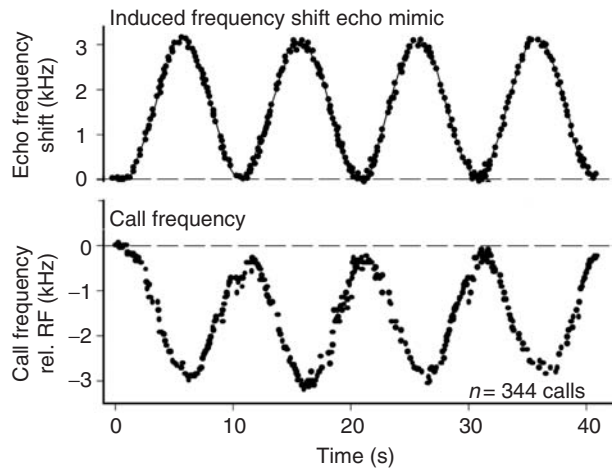

Figure 5 Doppler-shift compensation in *Rhinolophus ferrumequinum*. A stationary bat, trained to track a phantom target, emitted a constant stream of orientation sounds. A microphone picked up the emitted sounds and played them back at a frequency that ranged sinusoidally from the same as the emitted frequency (CF approximately 83 kHz) to a frequency upward shifted by 3 kHz (simulating a target moving toward the bat at 12 m s^{-1}). The bat reacted by lowering the emitted CF of each emitted sound by just enough to compensate for the upward Doppler-shift of the preceding echo. As a result the echo returned at an almost CF. The bat compensates very little to downward Doppler-shift, representing a target moving away from the bat. Kindly provided by Dr. Walter Metzner, UCLA.

and *Pteronotus*, the cochlea and nervous system are highly specialized – relatively insensitive to the downward-shifted emitted CF, but exquisitively sensitive to the reference frequency of the echo CF and to small fluctuations around it. The long CF appears to be used mainly for detection and discrimination of flying insects based on wing beat-specific modulations in the frequency and amplitude of the echo CF. Perching *Rhinolophus* hunt with sounds in which the CF is in the range of insensitivity and echoes of approaching insects are shifted upward into the range of high sensitivity.

The similar use of Doppler-shift compensation by *Rhinolophus* and *Pteronotus* is a remarkable example of convergent evolution. However, details of cochlear mechanisms and the neural circuitry for analysis of echoes are different in the two species. *Rhinolophus* does not use multiple harmonics in various pulse–echo combinations to obtain information, a central feature of the *Pteronotus* nervous system. Moreover, *Rhinolophus* emits sounds through the nostrils and *Pteronotus* by mouth.

In bats that emit short (<10 ms) CF/FM signals, Doppler-shift compensation is less accurate than in long CF/FM bats, and their auditory systems are less sharply tuned around the CF. A short CF component is less useful than a long CF for obtaining Doppler-shift information, especially about target wing beats.

The short CF component probably serves primarily to improve target detection, while the FM sweep provides most of the information about distance, direction, and the nature of targets. *Noctilio*, a pisciverous New World bat, often emits pulses in pairs or triplets, one of which is an 8–10 ms pure CF; the others are CF/FM.

Laboratory Studies of the Limits of Echolocation Skills

Detection

Early experiments using changes in pulse emission rate as a criterion of detection showed that the FM bat *M. lucifugus* reacted to an array of 1 mm diameter wires at about 2 m and to 0.18 mm wires at almost 1 m. In a major advance in the methodology for studying echolocation, Simmons used a food reward to train bats to sit on a platform, emit orientation pulses, and choose between alternative targets by crawling to the left or right of a Y-shaped platform. The threshold distance for *Eptesicus fuscus* to detect a 19.1 mm-diameter sphere was 5.1 m, while a 4.8-mm-diameter sphere was detected at 2.9 m. In these and similar experiments with other bats, the absolute echo intensities at threshold ranged from 8 to 40 dB SPL, that is, 50–85 dB fainter than the emitted sounds. Thresholds are lowest in long-eared 'whispering bats,' such as *Plecotus*.

These threshold values are comparable to those of most mammals at their best frequencies. However, bats must detect and analyze these sounds within milliseconds of emitting much louder sounds and in the presence of interfering echoes from other objects. In most mammals, responsiveness to echoes is actively suppressed. Bats, on the other hand, suppress responsiveness to outgoing sounds – by synchronized middle-ear muscle contraction and efferent neural inhibition coupled to pulse emission – and they have neural mechanisms for exaggerating responsiveness to returning echoes.

Range (Distance) Discrimination

To catch an insect, a bat must localize its prey within 1–2 cm in three-dimensional space. Sound travels 34.4 cm ms^{-1} in air. Hence, the delay between emission of a pulse and return of an echo is an accurate measure of distance. Simmons showed that several species of bats using broadband FM signals could determine which of two targets was closer if the difference was 8–15 mm or greater (45–85 μs, about 3% of the overall delay). The same bats seamlessly made the same discrimination when real targets were replaced by 'phantom' targets – playbacks of the bat's emitted sounds picked up by a microphone and broadcast from two equidistant loudspeakers at the angle of the original targets but at different delays corresponding to the different distances tested with real targets (Figure 6). This confirmed that pulse–echo delay is the criterion used for distance discrimination. In similar range discrimination tests, the long CF/FM bat, *Rhinolophus*, was less accurate (30–41 mm). The CF component is comparatively useless for distance discrimination, and their FM component is less broadband than the sweeps used by FM bats. The emitted sound appears to open a 'ranging window' of 20–30 ms duration during which echoes arriving at different delays excite different populations of 'delay-tuned' neurons.

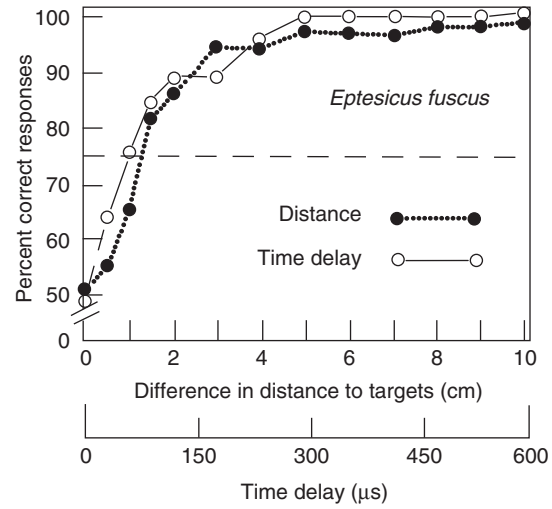

Figure 6 Range-difference discrimination by two *Eptesicus fuscus* in a two-alternative forced-choice behavioral test, where each bat was asked to distinguish which of two real targets was the closer. The correct choice was a target at 30 cm distance, the wrong choice was at a greater distance as shown in the upper abscissa. The targets were 20° to each side of midline. The real targets were then replaced by 'phantom targets', simulated by playing back the bat's emitted sounds from loudspeakers in the directions and at delays corresponding to the distances tested with real targets. The behavior was the same for both real and phantom targets. The dashed line indicates 75% correct choices, the usual criterion for successful discrimination. From Simmons JA (1973) The resolution of target range by echolocating bats. *Journal of the Acoustical Society of America* 54: 157–173.

Time-Intensity Trading, Middle-Ear Muscles, and Echo 'Gain Control'

In general, the latency of a neural response decreases as stimulus intensity increases. In behavioral experiments, bats consistently perform as if they interpret a 1 dB increase in echo intensity as a 15–17 μs decrease

in delay. When a bat approaches a target, the increase in echo intensity (12 dB per each halving of distance) would be expected to introduce a systematic error in distance perception.

This problem may be minimized by changes in sensitivity imposed by the middle-ear muscles, which contract during each pulse emission, sharply decreasing sensitivity. As the muscles relax during the few milliseconds after pulse emission, sensitivity recovers. The closer the target is, the louder the echo, but also the greater the attenuation due to middle-ear contraction. This modulation of sensitivity after each emission, combined with a decrease in emitted sound intensity as a bat nears its target, may compensate almost completely for the increase in echo intensity due to the decrease in distance. If so, echoes would return at almost constant intensity during the last 1–2 m of pursuit. This echo 'gain control' would reduce errors otherwise resulting from the intensity-latency trade-off.

Fine Temporal Resolution/Range Jitter Discrimination

Early behavioral measurements unexpectedly suggested that the threshold for distance discrimination was about the same over a wide range of absolute distances (30–240 cm). One would expect the discernible difference to increase in proportion to absolute distance. The behavioral measurements might instead have tested the uniformity with which a bat held its head at a constant distance from the microphone as it moved on its Y-shaped platform emitting pulses in the direction of each target.

In order to minimize this potential possibility, Simmons trained *Eptesicus* to discriminate between two phantom targets. One target always returned echoes at the same delay; the other returned echoes that alternated ('jittered') between slightly shorter and slightly longer delays (Figure 7). Only one phantom target returned an echo to any given pulse, determined by which of two microphones received the louder emitted sound. The 'jitter' was reduced until the bat could no longer distinguish which target appeared to be jittering in space. The results indicated that *Eptesicus* could easily detect 1–2 μs jitter and even as little as 10–100 ns if the signal-to-noise level was high! Although this astonishing 'hyperacuity' is highly controversial, it has been confirmed in part by others who found that *Eptesicus* resolved jitter almost perfectly down to 400 ns – the smallest difference tested. This is for jitter between successive echoes separated by ~100 ms! Subsequent research indicates that reduction in signal bandwidth, especially by high-pass filtering (excluding lower frequencies in the echo), degrades performance. With full bandwidth, the jitter thresholds approach those expected for an ideal receiver performing coherent cross-correlation between pulse and echo (Figure 7).

Additional experiments support the idea that bats can resolve temporal differences in the microseconds range. Taking advantage of the time–intensity trade-off, Simmons showed that reducing the amplitude of the earlier of two sounds by 1 dB or increasing the intensity of the later sound by 1 dB eliminated the perception of jitter when the interval between echo delays was 15–17 μs. This result coincides with the prediction that the bat interprets the louder sound as arriving 15–17 μs earlier. Spectral cues would not yield this result. If instead of alternating the delay between pairs of sounds in a jitter experiment the phase of one sound was changed by 90° or 180° compared to the other, bats clearly detected the difference.

'Glints,' Multiple–Wave Front Resolution, and Target Discrimination

To extract the information encoded in multiple overlapping echoes, bats must be able to identify which objects produce which echoes. This identification likely involves recognition of target dimensions based on wing beat detection, on spectral transformations due to differential reflection of different frequencies in the pulse, and on analysis of multiple 'glints' returning from different features of the target, overlapping in time.

The ability of bats to resolve multiple wave fronts in echoes has been tested with both real and phantom targets. When presented with two Plexiglas plates having arrays of holes drilled to a different depth, bats discriminated differences in depth of <1 mm. *Rhinolophus* can discriminate between pairs of phantom targets that differed from the positively reinforced reference separation of 7.7 μs (1.3 mm) by only 1 μs (equivalent to a spatial difference 0.17 mm). Such overlapping wave fronts create spectral notches in the compound echo. The position of the notches depends on the extent of overlap of echoes from the two surfaces and how individual frequency components reinforce and cancel each other. Consequently, two-wave front discrimination is thought to be based on subtle spectral cues. Mathematically, however, the cues could equally well be two separate but overlapping echoes slightly displaced in time. When bats are asked to distinguish a two-wave front echo from a single-wave front echo, their performance deteriorates when one of the two wave fronts is at the same delay as the single wave front.

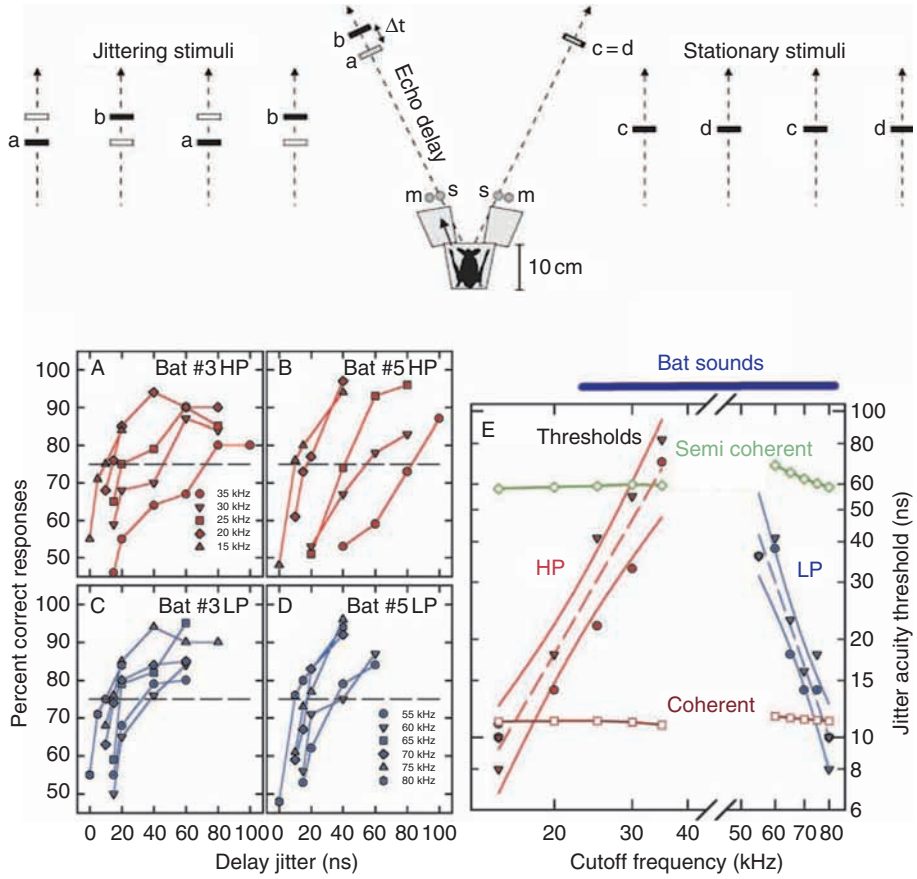

Figure 7 Range 'jitter' discrimination by *Eptesicus fuscus*. (Top) Experimental design. Microphones (m) picked up emitted sounds and played them back from one or the other loudspeaker (s) at delays corresponding to a constant distance for one target (c and d) and a distance that alternated back and forth by Δt (between a and b) in successive echoes from the other target (Bottom) (A–D): Performance of two bats to jitter between 0 and 100 ns, when the echoes (playback of the bats' own cries) were high-pass (HP, red) or low-pass (LP, blue) filtered (–24 dB per octave) to reduce the FM bandwidth. *Eptesicus* sounds contain nearly equal sound levels between 90 and 25 kHz. Symbols in panels A and D indicate cutoff filter settings. Note that removal of low-frequency components of the echoes degraded jitter discrimination more severely than removal of high-frequency components. (E) Plot of bats' jitter acuity thresholds at 75% correct response levels for different HP and LP filter settings (horizontal axis). Dashed red and blue lines show linear regression and solid lines show 99% confidence limits. Predicted thresholds for optimal receivers with coherent (open squares, brown line) and semicoherent (green) processing FM echoes are shown for each filter condition. (Top) From Simmons JA, Ferragamo MJ, Sanderson MI (2003) Echo delay versus spectral cues for temporal hyperacuity in the big brown bat, *Eptesicus fuscus. Journal of Comparative Physiology A* 189: 693–702. (Bottom) Simmons JA, Neretti N, Intractor N, et al. (2004) Delay accuracy in bat sonar is related to the reciprocal of normalized echo bandwidth, or *Q. Proceedings of the National Academy of Sciences* 101: 3638–3643.

Angular Resolution in the Horizontal and Vertical Dimensions

To intercept and catch an insect, a bat must know where the insect is in three-dimensional space within ±2–5°. Laboratory experiments have confirmed that bats can discriminate between two targets separated by as little as 1.5° in the horizontal plane (azimuth) and about 3° in the vertical plane. The cues for horizontal and vertical angular discrimination are less obvious than for distance (pulse–echo delay). Unlike vision, the auditory input from either ear does not provide a map of space. The map must be computed in the brain by comparing inputs at the two ears.

Pulse emission is somewhat beamed. Sound intensity falls off by ~10 dB within 30–45° off-axis in both horizontal and vertical planes. High frequencies are more narrowly beamed than low frequencies. Auditory sensitivity tends to be more directional than emitted intensity, again especially at the higher frequencies in each bat's range. The large and complex pinnae and the prominent tragus are essential to establishing directionality. In long CF/FM bats, and to a lesser degree in some short CF/FM bats, the distal portion of one pinna flicks forward and the other backward in synchrony with each pulse emission. During the next pulse the movements are reversed.

These ear movements probably are important for processing the CF component, determining target direction, or both.

Cues for horizontal localization of targets are interaural time (ITD) and intensity (IID) differences. For most echolocating bats, ITD is a maximum of 25–30 μs. IIDs, which depend on ear structure and sound shadows cast by the head, vary with frequency and can be as great as 40 dB at high frequencies in the echo (Figure 8). Typically, at 40 kHz and higher, sensitivity changes by an average 0.5–1 dB per degree over the ±30° on either side of midline. Thus a target 5° on one side of midline generates an echo that arrives at one ear about 2.5 μs earlier than at the other and is perceived as being 5–10 dB louder on the side that hears the echo first. Since a 1 dB difference in intensity translates into a difference of 15–17 μs in response latency, bats most likely use IIDs to discriminate direction in the horizontal plane.

Vertical displacement above or below the horizontal plane does not affect the phase or arrival time of echoes at the two ears. However, spectral peaks and notches in the echo result from interference between multiple wave fronts of an echo bouncing off different parts of the elaborate structures of the pinna and tragus. Thus, changes in vertical angle can dramatically

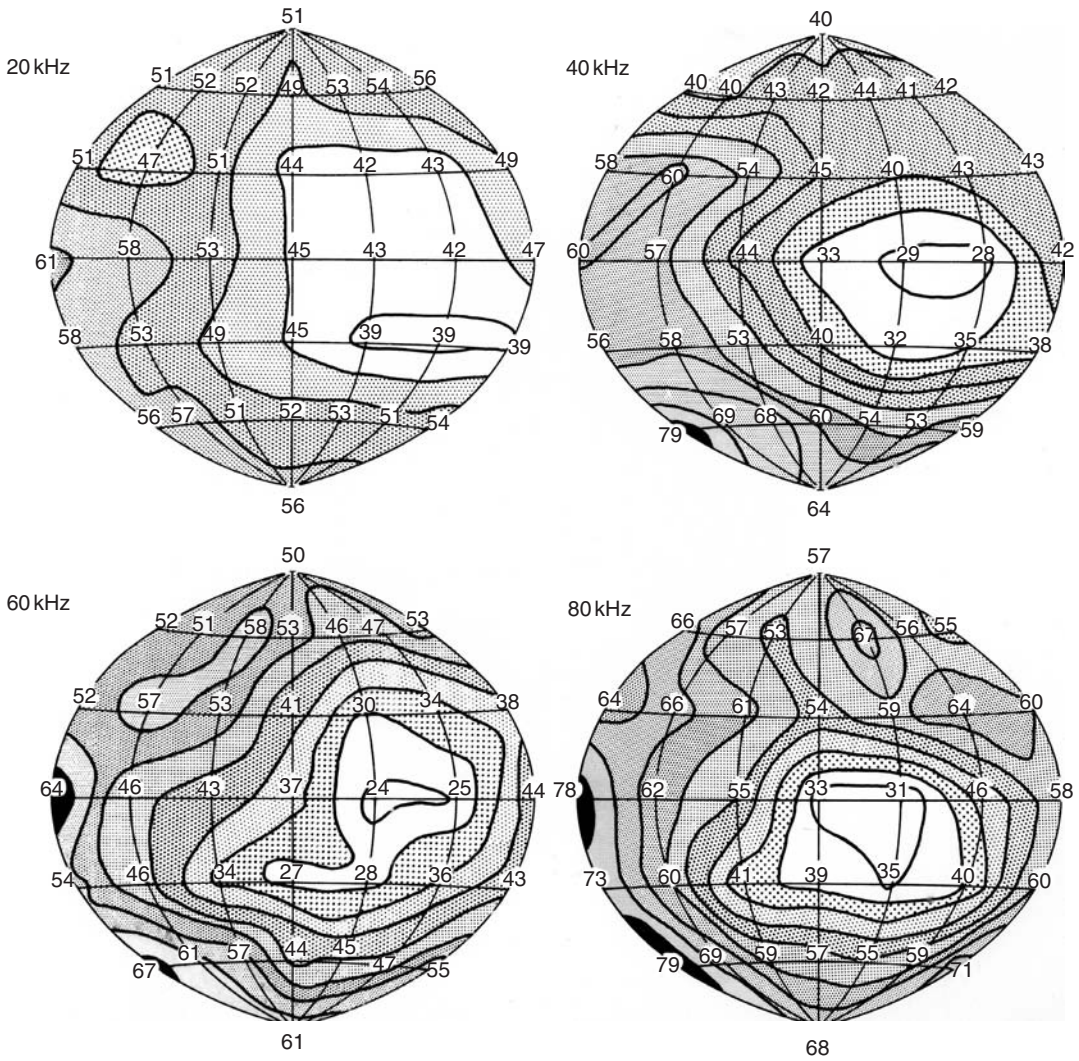

Figure 8 Directionality of hearing in a representative little brown bat, *Myotis lucifugus* (family Vespertilionidae). Diagrams showing the sensitivity of the evoked response from one posterior colliculus at 20, 40, 60, and 80 kHz to sounds coming from different positions in the hemisphere in front of the bat. Numbers are decibel SPL at threshold. Contour lines are drawn at every 5 dB. Sensitivity was greatest to sounds coming from the contralateral side. The colliculus on the other side would respond in mirror-image fashion, so binaural comparison would approximately double the perception of change in intensity with change in direction in the horizontal plane. Single units in the auditory pathway show much sharper directionality than this. From Grinnell AD and Grinnell VS (1965) Neural correlates of vertical localization by echo-locating bats. *Journal of Physiology* 181: 830–851.

alter sensitivity at different frequencies (Figure 8). It is thought that vertical localization can be done monaurally based on multiple-wave front interference patterns. However, it is questionable whether these spectral cues can be separated from the large intensity differences that result from changing horizontal direction. Therefore, bats may learn to interpret IIDs at different frequencies for both horizontal and vertical directionality. Even slight distortion of the configuration of the pinnae in the long-eared bat *Plecotus* seriously degrades wire avoidance skills.

Echolocation by Other Animals

The Megachiropteran *Rousettus* and the birds that echolocate do so facultatively, with broadband tongue clicks. They have been little studied. However, *Rousettus* avoid mist nets and appear to be quite skilled at echolocating. In contrast to other Megachiropterans, which orient entirely visually, *Rousettus* are maximally sensitive to ultrasonic frequencies and show the temporal resolution needed for echo detection and analysis (Figure 1(d)).

Dolphins and other toothed whales also have good vision but have developed echolocation to a level rivaling that of Microchiropteran bats. Broadband clicks are generated by forced movements of air through nasal passages and focused by the waxy melon of the forehead. A low searching rate is punctuated by foraging 'buzzes' of up to 600 clicks per second. Both the frequency spectrum (much of the energy >100 kHz) and intensity can be regulated depending on the echolocation task. Hearing is mediated through a fat body inside the lower jaw that extends to the temporal bone of the ear. Despite the animals' overall size, sensitivity is maximal in the ultrasonic frequency range. The bottle-nosed dolphin *Tursiops* can detect a water-filled 7.6 cm-diameter steel sphere at >100 m; discriminate a 3 mm difference in wall thickness of equal-sized aluminum spheres; distinguish targets composed of bronze, glass, and stainless steel; and distinguish spherical from cylindrical targets, independent of echo strength. They can visually recognize targets in air that they have learned to discriminate using echolocation when blindfolded underwater.

See also: Audiovocal Communication in Bats; Echolocation II: Neurophysiology; Hearing and Echolocation in Dolphins.

Further Reading

Au WWL (1993) *The Sonar of Dolphins*. New York: Springer.

Fay RR and Popper AN (eds.) (1995) *Hearing by Bats*, pp. 1–36. New York: Springer (see especially papers by Grinnell, Moss and Schnitzler, and Simmons, et al.).

Grinnell AD and Grinnell VS (1965) Neural correlates of vertical localization by echo-locating bats. *Journal of Physiology* 181: 830–851.

Griffin DR (1986) *Listening in the Dark*. Ithaca, NY: Cornell University Press.

Metzner W, Zhang S, and Smotherman M (2002) Doppler-shift compensation behavior in horseshoe bats revisited: Auditory feedback controls both a decrease and an increase in call frequency. *The Journal of Experimental Biology* 205: 1607–1616.

Moss CF and Schnitzler H-U (1989) Accuracy of target ranging in echolocating bats: Acoustic information processing. *Journal of Comparative Physiology A* 165: 383–393.

Neuweiler G (2000) *Biology of Bats*. New York: Oxford University Press.

Neuweiler G (2003) Evolutionary aspects of bat echolocation. *Journal of Comparative Physiology A* 189: 245–256.

Schnitzler H-U and Kalko EKV (2001) Echolocation by insect-eating bats. *BioScience* 51: 557–569.

Schnitzler H-U, Moss CF, and Denzinger A (2003) From spatial orientation to food acquisition in echolocating bats. *Trends in Ecology and Evolution* 18: 386–394.

Siemers BM and Schnitzler H-U (2004) Echolocation signals reflect niche differentiation in five sympatric congeneric bat species. *Nature* 429: 657–661.

Simmons JA (1973) The resolution of target range by echo-locating bats. *Journal of the Acoustical Society of America* 54: 157–173.

Simmons JA, Ferragamo M, Moss CF, Stevenson SB, and Altes RA (1990) Discrimination of jittered sonar echoes by the echo-locating bat, *Eptesicus fuscus*: The shape of target images in echolocation. *Journal of Comparative Physiology A* 167: 589–616.

Simmons JA, Ferragamo MJ, and Sanderson MI (2003) Echo delay versus spectral cues for temporal hyperacuity in the big brown bat, *Eptesicus fuscus*. *Journal of Comparative Physiology A* 189: 693–702.

Simmons JA, Neretti N, Intrator N, et al. (2004) Delay accuracy in bat sonar is related to the reciprocal of normalized echo bandwidth, or *Q*. *Proceedings of the National Academy of Sciences* 101: 3638–3643.

Thomas JA, Moss CF, and Vater M (eds.) (2004) *Echolocation in Bats and Dolphins*. Chicago: The University of Chicago Press.

Echolocation II: Neurophysiology

N Suga, Washington University, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The order *Chiroptera* (bats) comprises one-fifth of mammalian species and consists of two suborders: Megachiroptera (173 species) and Microchiroptera (813 species). All microchiropterans thus far studied echolocate, but only one megachiropteran (genus *Rousettus*) does. The morphology and ecology of bats are so diverse that echolocation behavior and orientation sounds (biosonar pulses or, simply, pulses) are quite different among different species of bats. Bats that emit biosonar pulses which always consist of a constant frequency (CF) component followed by a frequency-modulated (FM) component are called 'CF–FM' bats, whereas bats that emit biosonar pulses which always consist of an FM component only are called 'FM' bats. Accordingly, the auditory system differs among species. As is well known, toothed whales (about 70 species), such as dolphins, echolocate. In addition, two species of birds, *Steatornis* (oilbird) and *Collocalia* (swiftlet), have been known to echolocate. The biosonar pulses of the whales, birds, and *Rousettus* are clicks. Microchiropterans are not blind; their eyes are suited for scene analysis at night and function for long-range navigation, whereas echolocation is suited for short-range scene analysis in the dark and functions for object avoidance, hunting insect, etc.

Studies in the neuroscience of echolocation have been mainly performed with five different species of microchiropterans: *Pteronotus parnellii* (mustached bat), *Rhinolophus ferrumequinum* (horseshoe bat), *Myotis lucifugus* (little brown bat), *Eptesicus fuscus* (big brown bat), and *Antrozous pallidus* (pallid bat). The first two species are CF–FM bats and the last three are FM bats, all being insectivorous. Among them, the parallel and hierarchical processing of biosonar information was first and best explored in *Pteronotus*. Thus, the neurophysiology of the auditory system of *Pteronotus* is mainly described in this article.

Properties of Biosonar Signals

The biosonar pulses of *Pteronotus* always consist of a long CF component followed by a short FM component. Since each biosonar pulse contains four harmonics ($H_{1–4}$ in Figure 1(a)), there are potentially eight major components ($CF_{1–4}$, $FM_{1–4}$). The second harmonic (H_2) is always predominant, with CF_2 at ~61 kHz and FM_2 sweeping from 61 to ~49 kHz. The CF_2 frequency slightly differs among individual bats and is sexually dimorphic: CF_2 in males is ~1.04 kHz lower than in females. In target-directed flight, a CF–FM pulse varies in duration and emission rate (Figure 1(b)), but its spectrum changes little. *Rhinolophus* also emits CF–FM sounds with CF_2 at ~83 kHz. In comparison, *Myotis* and *Eptesicus* emit FM pulses that sweep downward about one octave within a range between 100 and 15 kHz. The properties of FM pulses vary depending on the species and echolocation situation (Figures 1(c) and 1(d)). Such species-specific biosonar pulses can be elicited by electrical stimulation of the cingulate cortex, periaqueductal gray, or the midbrain reticular formation.

The middle ear muscles start to contract a few milliseconds prior to sound emission, then maximally contract during, and relax within, a few milliseconds after a sound emission, in order to attenuate vocal self-stimulation by ~20 dB. Such a muscle contraction is well synchronized with a sound emission even at the emission rate of 200 s^{-1}. The stapedius muscle plays the essential role in this attenuation of vocal self-stimulation. Its tetanus-fusion frequency is ~320 s^{-1}. When the bat emits a sound, its vocal system sends signals (efferent copy) to the auditory system and attenuates vocal self-stimulation by ~15 dB, perhaps at the nucleus of the lateral lemniscus. These muscular and neural attenuations of vocal self-stimulation facilitate the detection of echoes returning after the emission of the strong biosonar pulses. *Rhinolophus* moves its pinnae synchronously with the emission of biosonar pulses, which go up to ~100 s^{-1}, to scan the auditory field in front of the animal. A nose-leaf bat such as *Rhinolophus* radiates biosonar pulses through the nostrils. The radiation pattern of the pulses is modulated by the nose leaf (leaflike attachment at the nose), which moves synchronously with pulse emissions.

Echoes that elicit behavioral responses in *Pteronotus* usually overlap with the emitted pulse, so that biosonar information is extracted from a complex sound potentially containing up to 16 components. Different components or parameters of echoes carry different types of target information (Table 1). The long CF component is an ideal signal for target detection and velocity measurements (relative velocity and velocity of insect wing-beat), while the short FM component is more appropriate for ranging, localization, and characterization of a target. *Pteronotus* and *Rhinolophus* optimize the acquisition of target information by an acoustic behavior called 'Doppler-shift

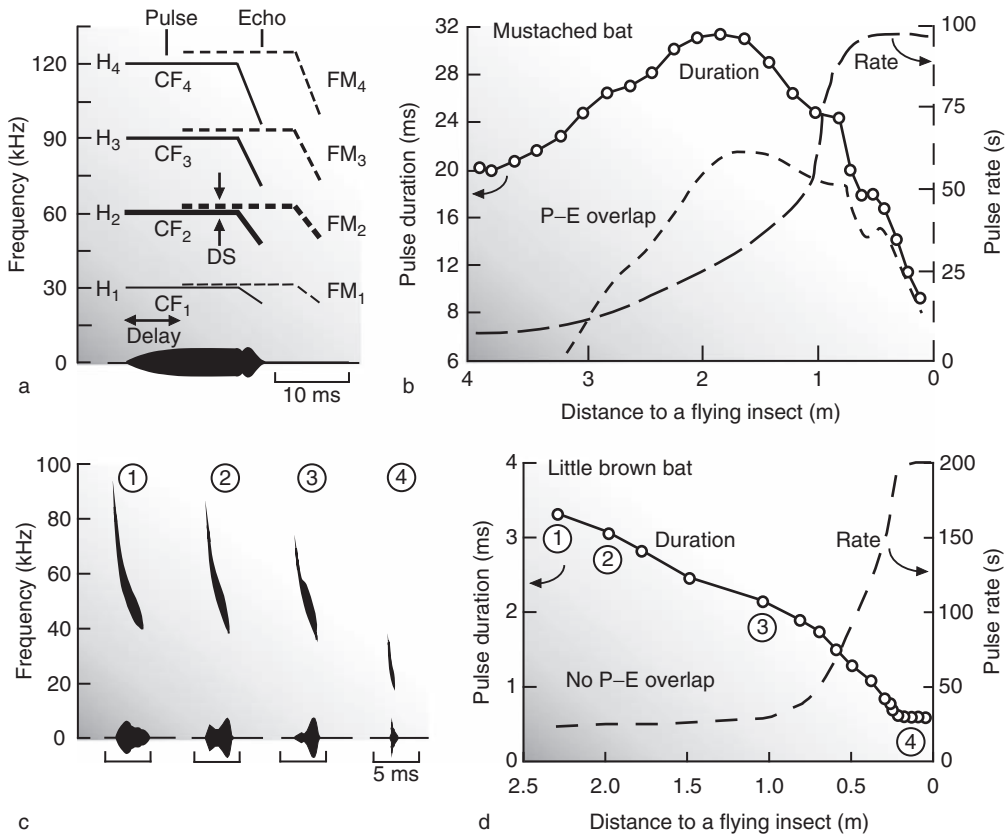

Figure 1 Biosonar pulses of the mustached bat, *Pteronotus parnellii* (a, b) and the little brown bat, *Myotis lucifugus* (c, d). (a) The four harmonics (H_{1-4}) of a *Pteronotus* biosonar pulse (P) each consist of a long CF and a short FM component, so that there are eight components: CF_{1-4} and FM_{1-4}. Most of the sound-energy is in the second harmonic (H_2), with CF_2 at about 61 kHz and FM_2 sweeping from 61 to about 49 kHz. As the bat approaches a target, a reflected echo (E) is Doppler-shifted (DS) in frequency, and its delay from the pulse becomes shorter. (b) During target-directed flight, *Pteronotus* increases the rate of its pulse emissions. The bat initially lengthens the duration of the pulse and then systematically shortens it. Pulse–echo overlap is maximal when the pulses emitted are longest in duration. (c) *Myotis* emits short downward-sweeping FM pulses. The duration and frequency sweep range of the pulse change during target-directed flight. (d) During target-directed flight, *Myotis* increases the rate of pulse emissions and shortens pulse duration so as to avoid pulse–echo overlap. Circled numbers 1–4 correspond to those in (c). CF, constant frequency; FM, frequency modulation. (b) Adapted from Novick and Vaisnys (1964). (d) Adapted from Friend, Suga, and Suthers (1966).

Table 1 Different parameters of an echo relative to the biosonar pulse at the ear carry different types of target information. Bats detect small objects and discriminate small differences between objects

Echo	*Target*	*Just noticeable difference*
Detection	Long wire	80 μm diameter (Rf)
	Long wire	100 μm diameter (Ml)
	Disk	4.2 mm diameter at 127 cm (Ml)
	Disk	10 mm diameter at 200 cm (Ef)
Doppler shift	Velocity	
Steady shift	Relative velocity	0.06–0.36% (Rf)
Periodic shift	Insect wing-beat	0.01–0.07% (Rf)
Periodic shift	Wing-beat rate	30% (Ef)
Amplitude	Subtended angle	16% (Ef)
Delay	Range	14 mm (81 μs) (Ef)
Jittered delay	Jittered range	10 ns (1.7 μm) (Ef)
Amplitude + delay	Size	16% (Ef)
Amplitude spectrum	Fine characteristics	87 μm in texture (Ef)
Binaural cues	Azimuth	1.5° (Ef)
Pinna-tragus cue (spectral notch)	Elevation	3° (Ef)

Ef, *Eptesicus fuscus*; Ml, *Myotis lucifugus*; Rf, *Rhinolophus ferrumequinum*.

compensation,' by which the frequency of the Doppler-shifted echo CF_2 is stabilized by changing the pulse CF_2 frequency at about 61 and 83 kHz, respectively, the frequency to which peripheral neurons are extremely sharply tuned. Echoes from flying insects are periodically Doppler-shifted, that is, frequency modulated, by beating wings. Such periodic Doppler shifts evoke hunting behavior of hungry bats. They also show an acoustic behavior called 'intensity compensation,' by which the echo intensity is stabilized at their preferred intensities.

The Auditory System of CF–FM Bats

Subcortical Specialization

The structure of the auditory system of a bat is basically the same as that of other mammals, but it shows a specialization in a species-specific way, reflecting the properties of its biosonar pulse and echolocation behavior. In particular, *Pteronotus* shows a unique functional organization for processing biosonar information. The width and thickness of the basilar membrane in the cochlea vary along its longitudinal axis, differently from those in primates. An innervation density of auditory neurons to hair cells also varies along the basilar membrane (Figure 2). One of the most striking features is the extremely sharp tuning of the auditory periphery for fine frequency analysis of the CF_2 component at about 61 kHz (about 83 kHz in *Rhinolophus*). Accordingly, auditory nerve fibers are also very sharply tuned to 61 kHz. They can code a 6-Hz frequency shift from 61 kHz, so that they can easily code the small FM of echoes that would be evoked by the wings of flying insects. The CF_2 processing channel is disproportionately large from the cochlea through the auditory cortex.

The other striking feature of the auditory system of *Pteronotus* is the parallel and hierarchical organization to extract certain types of biosonar information by 'combination-sensitive' neurons. All the CF and FM components of the biosonar pulse differ in frequency (Figure 1(a)), so that they are separately analyzed in the cochlea and are sent in parallel into the brain by different auditory nerve fibers. In the CF_{1-3} processing channels, many neurons at higher levels show 'level-tolerant' sharp frequency tuning for further analysis of the CF signals. A level-tolerant tuning curve has a very narrow bandwidth regardless of stimulus level and plays a role in level-tolerant fine frequency analysis. Inhibition in the central auditory system plays many different roles in auditory signal processing. Lateral inhibition, which occurs at different levels of the central auditory system, creates the level-tolerant sharp frequency tuning and amplitude tuning (Figures 3(A)–3(D)). Therefore, some neurons are tuned to a specific combination of frequency and amplitude. Sharpening of frequency-tuning curves by lateral inhibition is most dramatic in the CF_2 processing channel, which mostly projects to the Doppler-shifted constant frequency (DSCF) processing area of the primary auditory cortex (Figure 3(f)).

The parts of the CF_{1-3} processing channels are first integrated in the inferior colliculus for the extraction of velocity information from biosonar pulse–echo pairs. As a result, 'CF/CF combination-sensitive neurons' are produced. They are tuned to a specific combination of pulse CF_1 and echo CF_2 or CF_3, showing strong facilitation when pulse CF_1 is combined with echo CF_2 or CF_3. Their frequency-tuning curves for facilitation are very sharp and level-tolerant (Figure 4(c)). CF_1 is ~36 dB weaker than CF_2 and ~24 dB weaker than CF_3. According to an interpretation, CF/CF neurons extract velocity information from Doppler-shifted CF_2 or CF_3 referring to an emitted CF_1. The CF/CF neurons project to the CF/CF and dorsal intrafossa (DIF) areas in the auditory cortex via the medial geniculate body (Figure 4(d)).

In the FM_{1-4} processing channels, some neurons respond preferentially or selectively to FM sounds. FM sensitivity or selectivity is explained by the sequence of stimulation of excitatory and inhibitory tuning curves by an FM sweep or a neural circuit incorporating 'disinhibition' and/or 'facilitation.' The inferior colliculus has frequency versus latency coordinates. Part of the latency axis is used as delay lines. The parts of the FM_{1-4} processing channels are integrated into the inferior colliculus to create 'FM–FM combination-sensitive neurons' which are tuned to specific echo delays for the extraction of target-distance information from biosonar pulse–echo pairs. The neural mechanisms for creating FM–FM neurons consist of at least four physiological components: phasic/constant latency neurons, delay lines, coincidence detectors, and amplifiers. An FM sound sequentially stimulates an array of neurons tuned to different frequencies by sweeping across their frequency-tuning curves. Phasic/constant latency neurons are suited for coding that moment of the stimulus, that is, for coding the exact timing of the pulse and its echo (i.e., the echo delay). They are mostly created in the subcollicular nuclei, whereas the other three components are created in the inferior colliculus. Delay lines shift the response to the pulse FM_1 in time. A coincidence detector (FM–FM neuron) has two inputs: one input carries nerve impulses evoked by the pulse FM_1 through a delay line, and the other input carries nerve impulses evoked by the echo FM_n ($n = 2$, 3, or 4) without delay lines (Figure 5(b)). An echo always delays acoustically from the pulse

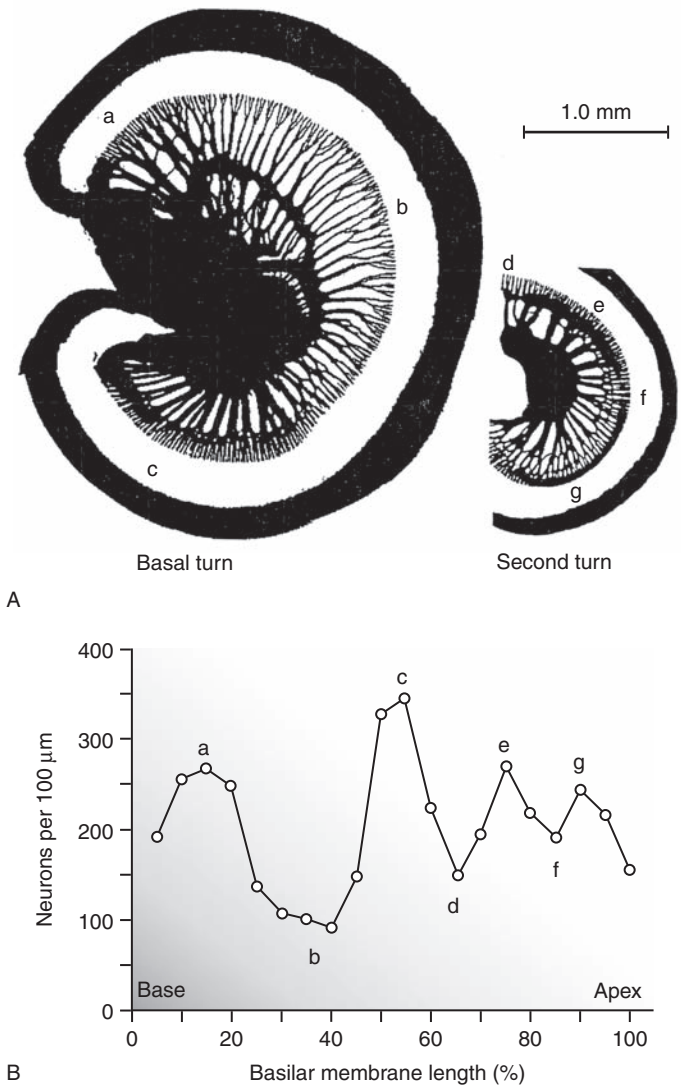

Figure 2 Nerve fiber innervation densities as illustrated by whole-mount preparations of basal and second turns of the cochlea of the mustached bat (A) and a graphic display of neuron density as a function of basilar membrane length (B). (A) The large basal turn (left) and the small second turn (right). Three prominent densely innervated regions (a, c, e) are identified on the basis of the size and density of myelinated nerve trunks. Each of these regions is bounded by sparsely innervated regions (b, d, f). (B) The distribution of the neural density along the cochlea shows peaks which closely correspond to the nerve densities shown in (A). The densities are shown from base (0%) to apex (100%). (A) Reproduced from Henson MM and Henson OW Jr. (1991) Specilizations for sharp tuning in the mustached bat: The rectorial membrane and spiral limbus. *Hearing Research* 56: 122–132, with permission. (B) Reproduced from Zook JM and Leake PA (1989) Connections and frequency representation in the auditory brainstem of the mustache bat, *Pteronotus pamelli. Journal of Comparative Neurology* 290: 243–261, with permission.

emitted by the bat, and the amount of echo delay is linearly related to a distance between the bat and an echo source, for example, a target. At a coincidence detector whose neural delay is equal to an echo delay, the excitatory response to the echo FM_n arrives at the same time as the delayed excitatory response evoked by the pulse FM_1. Then, the facilitation of the auditory response of the coincidence detector (i.e., FM–FM neuron) occurs. FM–FM neurons produced in this way are thus tuned to specific echo delays, that is, specific target distances. They project to thalamic FM–FM neurons, which show stronger facilitative responses to pulse–echo pairs and sharper delay tuning than the collicular FM–FM neurons. These thalamic FM–FM neurons project to the FM–FM, dorsal fringe (DF) and ventral fringe (VF) areas in the auditory cortex (Figure 5(c)). As a consequence of such parallel and hierarchical processing of complex sounds, there are many functional divisions in the auditory cortex (Figure 6).

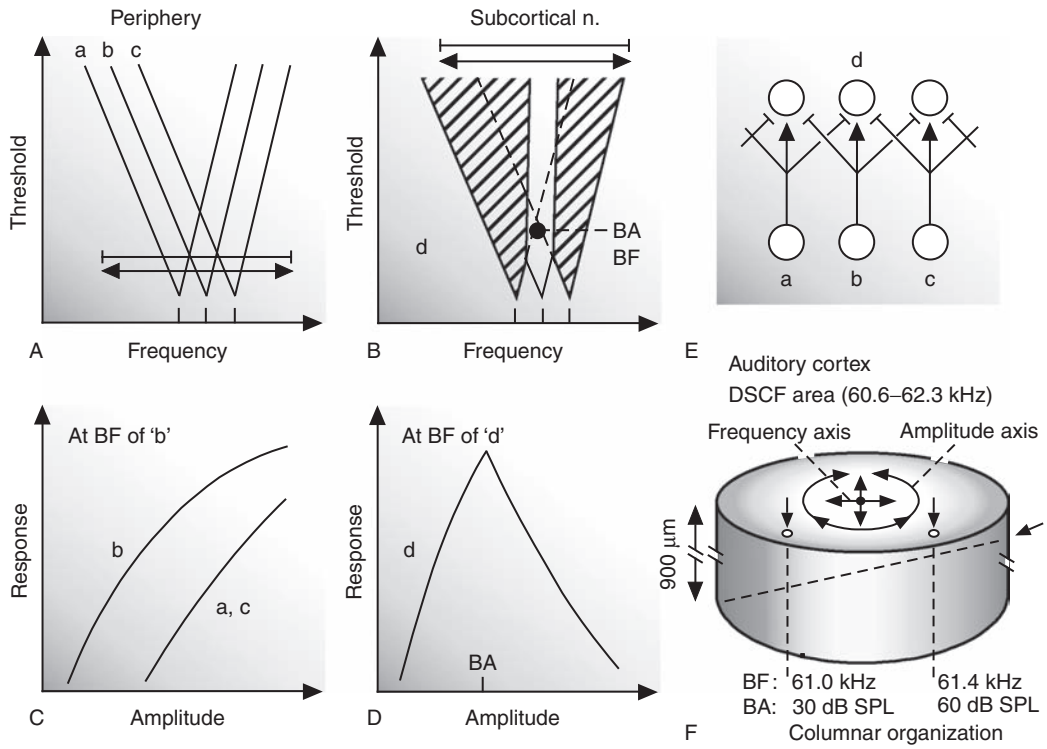

Figure 3 Sharpening of frequency-tuning curves (A and B) and creation of amplitude tuning (C and D) by lateral inhibition (E) in the Doppler-shifted constant frequency (DSCF) processing channel. (A and C) Frequency-tuning curves (A) and amplitude (intensity)–response (number of impulses) functions (C) of three peripheral neurons (a, b, and c in (E)). The amplitude–response functions are measured with a tone burst at a best frequency of neuron b. The double-headed arrow in (A) indicates that the neurons respond to upward- and downward-sweeping FM sounds with a sweep range shown by the length of the arrow shaft and that the threshold of the response is slightly higher than that to a CF tone at a neuron's best frequency. The horizontal bar in A indicates that these neurons respond to a noise burst with a specific bandwidth shown by the bar length and that the threshold of the response is slightly higher than those to the FM sounds. (B) A subcortical neuron (d in (E)) has a level-tolerant sharp frequency-tuning curve sandwiched between inhibitory-tuning curves (shaded) as a result of lateral inhibition from neurons a and c in (E). Neuron d is tuned to a specific combination of a best frequency (BF) and a best amplitude (BA) because of the overlapping inhibitory areas. (D) The amplitude–response function measured with a tone burst at the BF of neuron d is highly nonmonotonic. The double-headed arrow and horizontal bar in (B) indicate that neuron d shows very high thresholds to FM sounds and noise bursts and does not respond to them. (E) Neural circuit for lateral inhibition. The arrow- and bar-heads represent excitatory and inhibitory synapses, respectively. (F) In the DSCF area of the auditory cortex, each cortical column is characterized by a specific combination of BF and BA for the excitation of neurons (vertical arrows). With an oblique electrode penetration, both BF and BA systematically vary (oblique arrow). It has the frequency-versus-amplitude coordinates. (D) Based on Suga (1965a, 1965b, 1969, 1973). (E, F) Based on Suga (1965b, 1977) and Suga and Manabe (1982).

Cortical Specialization

In the primary auditory cortex of *Pteronotus*, high-frequency-sensitive neurons are located anteriorly and low-frequency-sensitive neurons are located posteriorly. Part of the auditory cortex has binaural bands which consist of either excitatory–excitatory neurons (EE; excitatory inputs from both ears) or excitatory–inhibitory neurons (EI; an excitatory input from one ear and an inhibitory input from the other ear). The EE band is callosally connected, but the EI band is not. Therefore, the auditory cortex of *Pteronotus* shares its functional organization with that of the cat and ferret. However, the auditory cortex of the mustached bat shows at least four unique aspects: (1) About 30% of the auditory cortex (called the DSCF area) is devoted to the representation of 60.6–62.3 kHz sound, although the total range of frequency representation in the auditory cortex is from 7 to 120 kHz (Figure 6(b)). Neurons in the DSCF area are combination-sensitive. (2) Another 30% of the auditory cortex (called the CF/CF and FM–FM areas) is primarily devoted to extract target velocity and range information from a biosonar pulse–echo pair by comparing the two signal elements. (3) The auditory cortex shows multiple representations of information-bearing elements such as velocity and range. (4) The auditory cortex shows no hemispheric asymmetry for biosonar signal processing, but appears to show asymmetry for processing communication calls. The DSCF, CF/CF, and FM–FM areas are highly specialized for

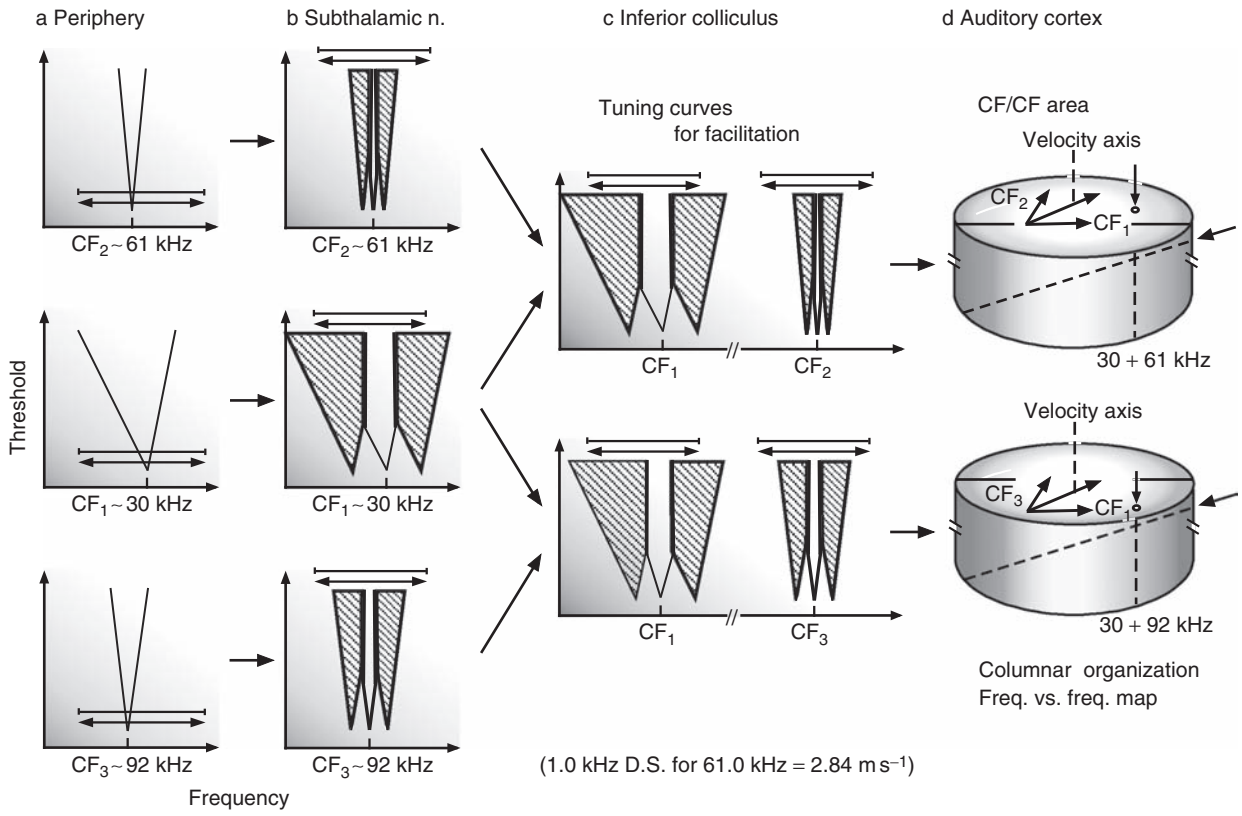

Figure 4 Neural mechanisms for creating Doppler-shift (velocity) axis in the CF/CF velocity-processing channel. Triangular frequency-tuning curves at the periphery (a) are changed into level-tolerant sharp frequency-tuning curves in the subthalamic auditory nuclei by lateral inhibition (b). Then, the CF_1 and CF_2 or CF_3 channels are integrated in the inferior colliculus to create two types of velocity-tuned CF/CF combination-sensitive neurons: CF_1/CF_2 and CF_1/CF_3 (c). The CF/CF neurons show a strong facilitative response to a specific combination of CF_1 and CF_n in the frequency domain. These neurons project to the CF/CF area in the auditory cortex and form the CF_1/CF_2 and CF_1/CF_3 subdivisions (d). In each subdivision, a cortical column is characterized by a specific combination of two frequencies (vertical arrows). Each subdivision has frequency-versus-frequency coordinates, in which relative target-velocities from –2 to 9 ms are systematically mapped (oblique arrow) (Suga et al., 1983a). See **Figure 3(B)** for the horizontal bars and double-headed arrows (Suga, 1965b).

the processing of biosonar information. The functional organization of these areas is described in the following paragraphs.

Neurons in the DSCF area are tuned to specific frequencies and amplitudes and form 'frequency-versus-amplitude coordinates' (i.e., tonotopic vs. ampliotopic representations). The frequency-tuning curves of the great majority of DSCF neurons are extremely sharp and level tolerant, that is, the widths of the tuning curves are narrow, even at high stimulus levels. Therefore, they are suited for level-tolerant frequency (velocity) discrimination. Both level-tolerant frequency tuning and amplitude selectivity result from lateral inhibition that occurs in the subcortical auditory nuclei. The frequency representation in the DSCF area is extremely fine. Inactivation of this area reversibly impairs the bat's ability to detect fine frequency differences (0.08–0.8%), but not its ability to detect differences larger than 3%. The DSCF area represents the frequency and amplitude modulations of echoes from a flying insect by the spatiotemporal pattern of neural activity occurring in its frequency-versus-amplitude coordinates. DSCF neurons are FM_1–CF_2 combination sensitive and, perhaps, provide information to the bat for maintaining appropriate distances from background objects for optimal hunting.

Neurons in the CF/CF area respond poorly to biosonar pulses and Doppler-shifted echoes when presented alone, but show a remarkable facilitative response to a pulse–echo pair. The essential elements in the pair are the CF_1 of the pulse and the CF_2 or CF_3 of the echo. For maximum facilitation, a specific frequency relationship between the two CF signals is required regardless of stimulus levels, because the majority of CF/CF neurons show level-tolerant sharp frequency tuning for facilitation. They are tuned to specific amounts of Doppler shifts, that is, relative target velocities. CF_1/CF_2 neurons are clustered dorsally and CF_1/CF_3 neurons are clustered ventrally (**Figure 6(b)**). The CF_1/CF_2 and CF_1/CF_3 subdivisions each have frequency-versus-frequency coordinates for the systematic representation of target velocities from 9 to –2 m s^{-1}. About one-third of CF/CF

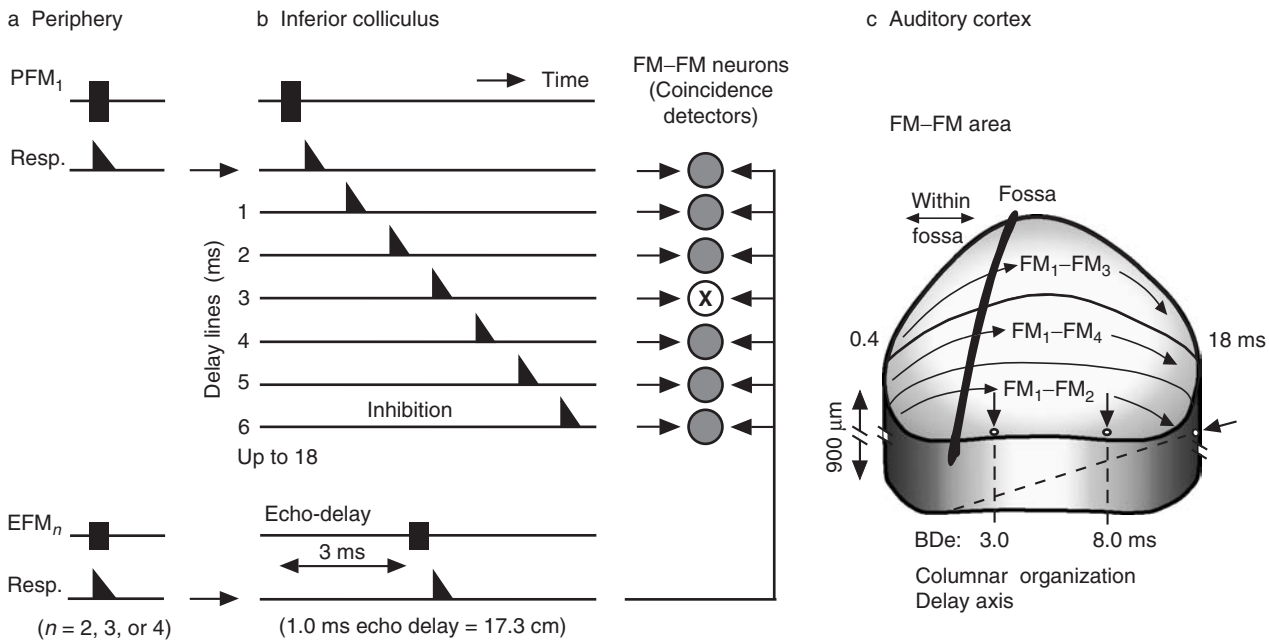

Figure 5 Neural mechanisms for creating delay-tuned neurons and a delay (range) axis in the FM–FM range-processing channel. Delay-tuned neurons utilize delay lines (0.4–18 ms long) that are mostly created at inferior collicular neurons tuned to frequencies swept by FM_1 (top portion of (b)). The delay-tuned neurons do not use delay lines created at neurons tuned to frequencies swept by FM_n ($n = 2$, 3, or 4) (bottom portion of (b)). The filled triangles in (a) and (b) are neural responses. An array of delay-tuned neurons in the inferior colliculus receives signals from the neurons responding to either FM_1 or FM_n channels (right portion of (b)). In a delay-tuned neuron (circle with the X) where an echo (EFM_n) delay from the pulse (PFM_1) is equal to the delay line associated with it, the signals from FM_1 and FM_n arrive at the same time. Then, the neuron shows a strong facilitative response. The neuron acts as a coincidence detector and multiplier. Three types of collicular delay-tuned neurons separately project to the FM–FM area of the auditory cortex through the thalamus, forming three subdivisions within the FM–FM area (c). Each subdivision of the FM–FM area shows columnar organization characterized with a specific echo delay, that is, best delay (BDe) for the excitation of neurons (vertical arrows). It has an echo delay axis from 0.4 to 18 ms (i.e., range axis). For simplicity, phasic-on response (onset detector), amplitude tuning, and FM sensitivity are not included in this model. Adapted from Suga (1990).

neurons show phase-locked responses to the periodic frequency modulations of the echo CF component that would be evoked by the wings of a flying insect when the biosonar pulse is paired with a Doppler-shifted echo. As for DSCF neurons, CF/CF neurons are tuned in amplitude as well as in frequency. However, amplitotopic representation is not known in the CF/CF area. CF_1/CF_2 and CF_1/CF_3 neurons are also clustered in the DIF area (Figure 6(b)). These CF/CF neurons are tuned to unusually high speed, if the interpretation is correct.

Neurons in the FM–FM area respond poorly or not at all to either a biosonar pulse or echo alone, but show a strong facilitative response to a pulse–echo pair with a specific echo delay, the essential elements being the FM_1 of the pulse and one of the $FM_{2–4}$ components of the echo. FM–FM neurons are tuned to specific echo delays, that is, target ranges. FM_1–FM_2, FM_1–FM_3, and FM_1–FM_4 neurons are separately clustered, so that the FM–FM area consists of three subdivisions (Figure 6(b)). Each subdivision has an echo-delay axis from 0.4 to 18 ms for the systematic representation of 7- to 310-cm target ranges (range axis). About 25% of FM–FM neurons are tuned to echo delays shorter than 4 ms (69-cm target range) and respond vigorously to each pulse–echo pair delivered, even at a rate of 100 s^{-1}. Therefore, the FM–FM area is involved in information processing even during the terminal phase of echolocation. Reversible inactivation of the FM–FM area temporarily impairs the bat's ability to detect a small difference in echo delay.

FM–FM neurons are tuned not only to specific echo delays but also to specific echo amplitudes. Therefore, they would respond best to targets of specific cross-sectional areas at specific distances. FM_1–FM_4 neurons are theoretically better suited for fine characterization of small targets than FM_1–FM_2 neurons, because FM_4 has a much shorter wavelength and broader bandwidth than FM_2. Therefore, the distribution of neural activity perpendicular to the range axis on the cortical plane represents the amplitude spectra of all FM components of an echo, which are directly related to target characteristics such as shape and size.

The three types of FM–FM neurons (FM_1–FM_2, FM_1–FM_3, and FM_1–FM_4) are also clustered in the DF and VF areas (Figure 6(b)). The DF area is smaller than the FM–FM area and consists of three clusters of

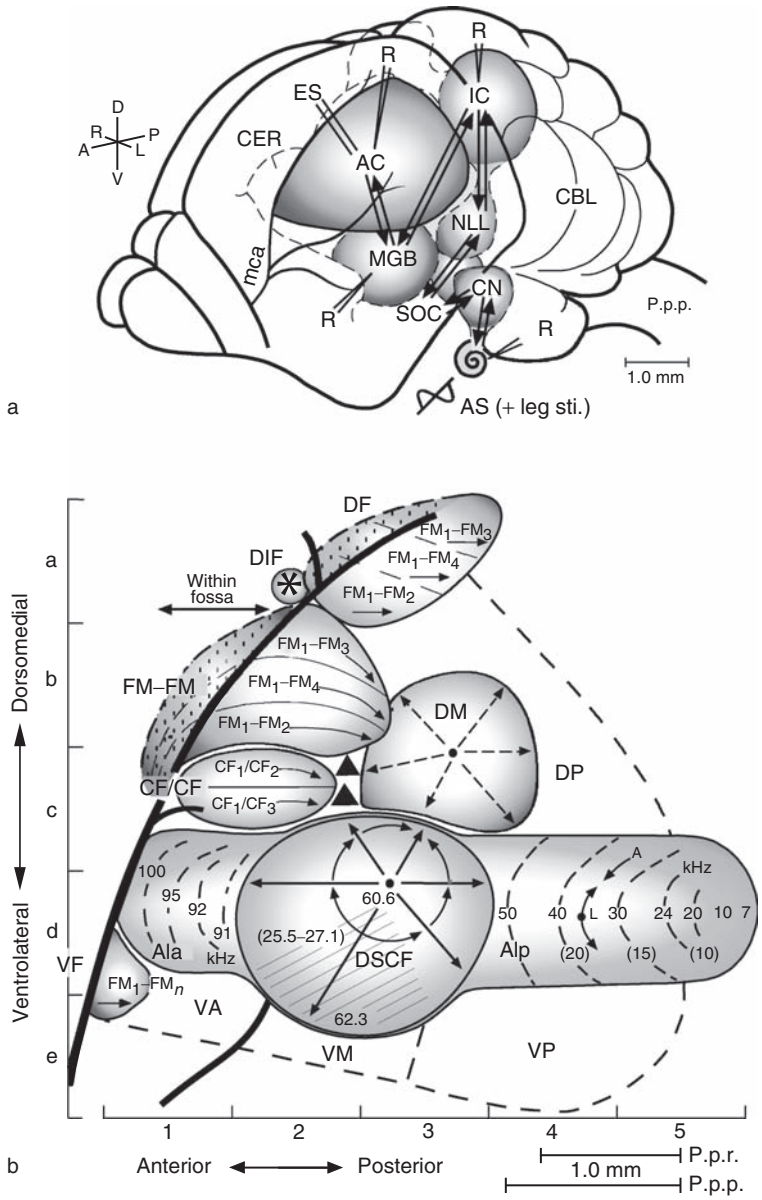

Figure 6 The functional organization of the auditory cortex of the mustached bat. (a) Dorsolateral view of the left cerebral hemisphere and the branches of the median cerebral artery (mca). The long branch is on the fossa. The ascending and descending auditory pathways are indicated by the arrows. The auditory cortex consists of several electrophysiologically distinct areas. (b) The frequency (tonotopic) representation of the primary auditory cortex (AIa, DSCF, and AIp) and the functional organization of the other areas are indicated by lines and arrows. The DSCF area has axes representing target-velocity information (echo frequency: 60.6–62.3 kHz) and subtended target-angle information (echo amplitude: 13–98 dB SPL). It consists of two subdivisions suited for either target detection (shaded) or target localization (unshaded) which are occupied mainly by EE or EI neurons. Neurons in the DSCF area are sensitive to an FM_1–CF_2 combination. The FM–FM area consists of three major types of FM–FM neurons (FM_1–FM_2, FM_1–FM_3, and FM_1–FM_4), which form separate clusters. Each cluster has an axis representing target ranges from 7 to 310 cm (echo delay: 0.4–18 ms). The CF/CF area consists of two major types of CF/CF neurons (CF_1/CF_2 and CF_1/CF_3), which are also found in separate clusters. Each cluster has two frequency axes and represents a target velocity from –2 to $+9\,m\,s^{-1}$ (echo Doppler-shift: –0.7 to +3.2 kHz for CF_2, and –1.1 to +4.8 kHz for CF_3). The DIF area also contains two types of CF/CF neurons. The DF area also consists of the three clusters of FM–FM neurons, and has an axis representing target ranges from 7 to 140 cm. The VF area also contains FM–FM neurons, and represents target ranges up to 80 cm. The DM area has an axis representing the azimuthal location of the target on the contralateral side in front of the animal. This azimuthal representation is incorporated with frequency representation. The functional organization of the DP, VA, VM, and VP areas remains to be studied further. AC, auditory cortex; CBL, cerebellum; CER, cerebrum; CN, cochlear nucleus; DM, dorsomedial area; IC, inferior colliculus; MGB, medial geniculate body; NLL, nucleus of lateral lemniscus; P.p.p, *Pteronotus parnellii parnelli*; SOC, superior olivary complex; sti., stimulation; VA, ventroanterior; VM, ventromedial area; VP, ventroposterior. Adapted from Suga N (1984) The extent to which biosonar information is represented in the bat auditory cortex. In: Edelman GM, Gall WE, and Cowan WM (eds.) *Dynamic Aspects of Neocortical Function*, pp. 315–373. New York: Wiley.

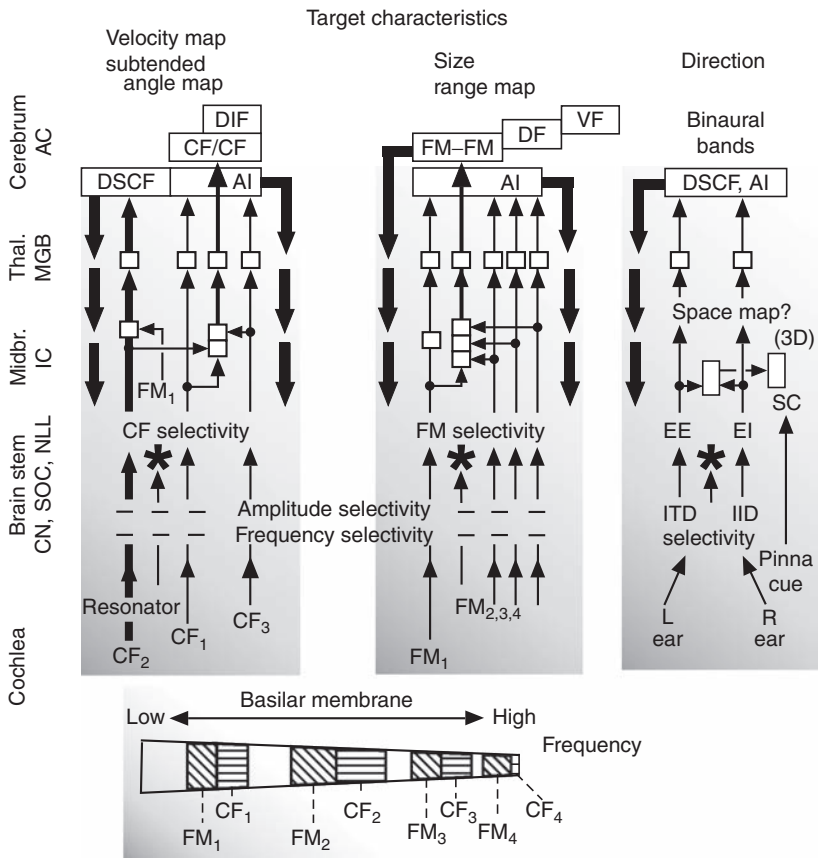

Figure 7 Parallel and hierarchical processing of different types of biosonar information carried by complex biosonar signals. The eight components (CF_{1-4} and FM_{1-4}) of the biosonar pulse and its echo are analyzed at different locations on the basilar membrane in the cochlea (bottom). These signal elements coded by auditory nerve fibers are sent up to the auditory cortex (AC) through a number of auditory nuclei (left margin): cochlear nucleus (CN), superior olivary complex (SOC), nucleus of lateral lemniscus (NLL), inferior colliculus (IC), and medial geniculate body (MGB). During the ascent of the signals, frequency selectivity is increased in some neurons, and amplitude, constant-frequency (CF), and frequency-modulation (FM) selectivities are added to some neurons (arrows with a star). Each star indicates that the addition of selectivity takes place in the auditory nuclei and cortex as well as in the nucleus where the arrow starts. The CF_2 channel is associated with a sharply tuned local resonator and is disproportionately large. It projects to the DSCF area of the AC. In certain portions of the IC, two channels processing different signal elements (e.g., CF_1 and CF_n or FM_1 and FM_n; $n = 2$, 3, or 4) converge to produce two major types of combination-sensitive neurons: velocity-tuned CF/CF neurons and delay-tuned FM–FM neurons. Velocity-tuned neurons project to the CF/CF area of the AC, where target-velocity information is systematically represented (left top). Velocity-tuned neurons also cluster in the DIF area. Delay-tuned neurons project to the FM–FM, DF, and VF areas of the cortex where target-range information is systematically represented (center top). The delay lines utilized by them are mostly created in the inferior colliculus. The DSCF area has frequency-versus-amplitude coordinates for systematic representation of the properties of an echo CF_2 and consists of two subdivisions, mainly containing excitatory–inhibitory (EI) or excitatory–excitatory (EE) binaural neurons (right column). Intraural intensity difference- or intraural time difference-tuned neurons (ITD, IID) are created in the SOC. The superior colliculus (SC) has a cluster of 3D-tuned neurons. Revised from Suga (1988).

FM–FM neurons as in the FM–FM area. Each cluster has an echo delay axis up to 9 ms for systematic representation of target ranges up to 155 cm. The VF area is much smaller than the DF area and has an echo delay axis up to 5 ms for systematic representation of target ranges up to 86 cm.

The parallel and hierarchical processing of biosonar information has been best explored in *Pteronotus* (Figure 7). DSCF, CF/CF, and FM–FM neurons are undoubtedly specialized for the feature extraction from biosonar signals. However, they also respond to certain types of communication sounds which are comparable to biosonar signals, but contain a first harmonic much stronger than that of the biosonar pulses.

The Auditory System of FM Bats

Eptesicus and *Myotis* are so-called FM bats. Their auditory cortices have no areas comparable to the DSCF, CF/CF, and FM–FM areas of *Pteronotus*. Their auditory systems, however, also create neurons tuned to behaviorally relevant values of parameters characterizing biosonar signals through excitatory (facilitatory)–inhibitory interactions taking place at different

subcortical levels of the auditory system. All of the neurons discussed in the following list are frequently found in the inferior colliculus and above.

1. *Latency-constant, phasic on-responding neurons*. Clustered in the 'columnar' region of the ventral nucleus of the lateral lemniscus, they can mark the onset time of the emitted pulse and may be used for the processing of target-distance information carried by echo delays.
2. *Paradoxical latency-shift neurons*. Their response latency is long for a strong sound, but shifts short for a weak sound. When the strong sound (e.g., emitted biosonar pulse) is followed by the weak sound (e.g., echo) with a time delay equal to the latency shift, the excitation evoked by the weak sound superimposes on that evoked by the strong sound and evokes a strong facilitative response. Therefore, the paradoxical latency-shift neurons are suited for processing target-distance information carried by an echo delay.
3. *Duration-tuned neurons*. They are mostly tuned to short durations like biosonar pulses. In the auditory cortex, neurons with short best durations are ventrally located to those with long best durations.
4. *FM-sensitive or-specialized neurons* (*simply, FM neurons*). For the excitation of FM neurons, the direction, sweep rate, and duration of FM sounds are important. Short FM sounds similar to the biosonar pulses are best for exciting them.
5. *Delay-tuned neurons*. They are different from the delay-tuned neurons of the CF–FM bats, because they are sensitive to the combination of FM_1–FM_1. An echo-delay map has not yet been found in the FM bats.
6. *Level-tolerant, sharply frequency-tuned neurons*. They are tuned not only in frequency but also in amplitude. During the search phase of echolocation, *Eptesicus* lengthens its biosonar pulse by adding a quasi-CF component to the end of the FM pulse to optimize target detection. The analysis of the quasi-CF component by level-tolerant, sharply frequency-tuned neurons is suited for enhanced target detection.
7. *FM or AM rate-tuned neurons*. They may be suited to respond to the repetitive biosonar signals during target-directed flight and/or FM or AM sounds which are commonly found in the communication calls of some bats.

Signal Processing for Sound Localization

As with other mammals, bats can use interaural intensity differences (IIDs) and interaural time differences (ITDs) for the localization of a sound (echo) source in azimuth, and the 'pinna–tragus' (or simply 'pinna') cue for the localization of the sound source in elevation. The pinna cue is a spectral notch (or notches) created by interference between sound waves directly incidenting on the eardrum and indirectly incidenting on it through the reflection by the pinna and tragus. The large pinna of the bat amplifies sound waves and facilitates sound localization. The pinna cue perhaps starts to be processed in the cochlear nucleus, whereas the binaural cues start to be processed in the lateral and medial superior olivary nuclei. IID is the primary cue to localize high-frequency sounds. It is encoded by 'EI' neurons that are excited by one ear stimulation and inhibited by the other ear stimulation. This binaural interaction makes the EI neurons sensitive to IIDs: their response magnitude drastically changes as a function of IID, that is, of azimuthal location of a sound source. They are not tuned to a specific azimuth degree. Instead they quickly increase their response, reaching a plateau, when a sound source is moved laterally from front 0°. The cut-off azimuth degree at which the response reaches the plateau differs from neuron to neuron. The EI neurons are first created in the lateral superior olivary nucleus and project to the inferior colliculus. Neurons in the dorsal nucleus of the lateral lemniscus (DNLL) are inhibitory and are EI. They bilaterally project to the inferior colliculus. DNLL neurons either modify collicular EI neurons or create EI neurons in the inferior colliculus. In the inferior colliculus, EI, EE (binaurally excited), and EO (monaural) neurons are separately clustered. EE neurons are most sensitive to a sound source at front 0°.

The inferior colliculus projects to the superior colliculus where visual, auditory, and somatosensory signals are integrated and where sensory–motor interface occurs. The superior colliculus plays an important role in head and pinna orientation. It contains neurons tuned to a specific combination of azimuth and elevation ('2D neurons') and those tuned to a specific combination of echo delay, azimuth, and elevation ('3D neurons'). The 2D neurons are located throughout the superior colliculus, whereas the 3D neurons are located in the anterior half and near the midline. EI, EE, and 2D neurons exist in all mammalian brains and are not unique to bats; however, 3D neurons are found only in the bat brain.

In the auditory cortex of the *Eptesicus* and *Pteronotus*, low-frequency-tuned neurons are sensitive to more laterally located sound sources, whereas high-frequency-tuned neurons are sensitive to more medially located ones. Therefore, the frequency (tonotopic) axis is also the azimuthal axis. EI and EE neurons are separately clustered in the DSCF area of the auditory cortex.

Protection of Biosonar Information Processing from Jamming by Biosonar Pulses of Conspecifics

Bats, such as the mustached bat, live in large colonies. Consequently, one important problem faced by individual bats is how to protect their echolocation system from the jamming effect of biosonar pulses emitted by conspecifics. Bats' echolocation is hardly ever jammed by sounds delivered from multiple loudspeakers. There are several possible mechanisms that could work in combination to reduce jamming: (1) the sharp directionality of biosonar pulses, (2) the sharp directional sensitivity of the ear and binaural hearing, (3) the signature of biosonar pulses and the personalized auditory system, (4) an auditory time gate for processing echoes, (5) heteroharmonic sensitivity, (6) variation in emitted signals and the sequential processing of echoes, and (7) efferent copy originating from the vocal system. Items (4) and (5) are unique and are related to the response properties of FM–FM and CF/CF neurons (Figures 4 and 5). These mechanisms are explained in the following paragraphs.

CF/CF and FM–FM neurons do not respond or poorly respond to echoes which delay by more than 20 ms from the emitted pulse. That is, they have an auditory time gate for processing velocity or distance information by comparing the higher harmonics of the Doppler-shifted echo with the first harmonic of an emitted pulse. The first harmonic (CF_1 and FM_1) is the weakest component (Figure 1(a)). It has less than 1% of the total energy of the emitted pulse. However, it is used as the reference to measure relative velocity and distance of a target.

Why does the auditory system create such heteroharmonic-sensitive neurons instead of homoharmonic-sensitive neurons such as CF_2/CF_2 and FM_2–FM_2? Let us consider the situation in which hundreds of bats are flying in a cave and the first harmonic of their pulses is completely suppressed by the antiresonance of the vocal tract. Then, in the air, there are many second, third, and fourth harmonics produced by conspecifics. The combinations of these harmonics cannot excite the heteroharmonic-sensitive FM–FM and CF/CF neurons because the first harmonic is absent. When the animal itself emits pulses, however, the first harmonic is present at the larynx and stimulates both of the ears by bone conduction. The FM–FM and CF/CF neurons are then turned on by the first harmonic to process target-distance and velocity information, respectively. In this way, the neural computation of biosonar information in the auditory system is protected from the jamming that would otherwise be caused by the sounds emitted by many conspecifics flying nearby.

Corticofugal Feedback for Auditory Signal Processing

The parallel and hierarchical processing of auditory signals in the ascending auditory system greatly depends on the descending (corticofugal) auditory system, which forms multiple feedback loops (Figure 7). The corticofugal system modulates subcortical neurons for sharpening and shifting of neural tuning curves in the frequency, amplitude, time, and spatial domains and plays a key role in reorganization (plasticity) and, therefore, organization of the auditory system. In the DSCF and FM–FM processing channels of the mustached bat, inhibition is strong so that corticofugal modulation is different from that in the big brown bat, rodents, and cats.

See also: Audiovocal Communication in Bats; Echolocation I: Behavior; Hearing and Echolocation in Dolphins; Social Communication in Whales and Dolphins; Sound Localization: Neural Mechanisms.

Further Reading

Covey E and Casseday JH (1999) Timing in the auditory system of the bat. *Annual Review of Physiology* 61: 457–476.

Fay RR and Popper AN (1995) *Hearing by Bats*. New York: Springer.

Fenton MB, Racey P, and Rayner JMV (1987) *Recent Advances in the Study of Bats*. Cambridge, MA: Cambridge University Press.

Fuzessery ZM (1994) Response selectivity for multiple dimensions of frequency sweeps in the pallid bat inferior colliculus. *Journal of Neurophysiology* 72: 1061–1079.

Henson MM and Henson OW Jr. (1991) Specilizations for sharp tuning in the mustached bat: The rectorial membrane and spiral limbus. *Hearing Research* 56: 122–132.

Kobler JB, Wilson BS, Henson OW, Jr., et al. (1985) Echo intensity compensation by echolocating bats. *Hearing Research* 20: 99–108.

Metzner W, Zhang S, and Smotherman M (2002) Doppler-shift compensation behavior in horseshoe bats revisited: Auditory feedback controls both a decrease and an increase in call frequency. *Journal of Experimental Biology* 205: 1607–1616.

Neuweiler G (1990) Auditory adaptations for prey capture in echolocating bats. *Physiological Reviews* 70: 615–641.

Pollak GD, Burger RM, and Klug A (2003) Dissecting the circuitry of the auditory system. *Trends in Neurosciences* 26: 33–39.

Schuller G, Beuter K, and Schnitzler H-U (1974) Response to frequency shifted artificial echoes in the bat. *Rhinolophus ferrumequinum*. *Journal of Comparative Physiology* 89: 275–286.

Simmons JA (1989) A view of the world through the bat's ear: The formation of acoustic images in echolocation. *Cognition* 33: 155–199.

Suga N (1984) The extent to which biosonar information is represented in the bat auditory cortex. In: Edelman GM, Gall WE, and Cowan WM (eds.) *Dynamic Aspects of Neocortical Function*, pp. 315–373. New York: Wiley.

Suga N and Ma X (2003) Multiparametric corticofugal modulation and plasticity in the auditory system. *Nature Reviews Neuroscience* 4: 783–794.

Suga N, O'Neill WE, Kujirai K, et al. (1983) Specificity of combination-sensitive neurons for processing of complex biosonar signals in auditory cortex of the mustached bat. *Journal of Neurophysiology* 49: 1573–1626.

Suga N and Tsuzuki K (1985) Inhibition and level-tolerant frequency tuning in the auditory cortex of the mustached bat. *Journal of Neurophysiology* 53: 1109–1145.

Valentine DE and Moss CF (1997) Spatially selective auditory responses in the superior colliculus of the echolocating bat. *Journal of Neuroscience* 17: 1720–1733.

Zook JM and Leake PA (1989) Connections and frequency representation in the auditory brainstem of the mustache bat, *Pteronotus pamelli*. *Journal of Comparative Neurology* 290: 243–261.

Electrical Perception and Communication

C D Hopkins, Cornell University, Ithaca, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Electric Fishes

This article concerns electrical communication in fish, an unusual modality of communication found only in fishes that are capable of both generating and receiving weak electric signals in water. Electrogenesis in fish arose independently in six groups of marine or freshwater fishes (Table 1). Electric organs are derived from modified muscle or nerve cells. Some of these fish are strongly electric and have been known since antiquity; others are weakly electric and went unnoticed until the 1950s. Communication is most highly developed among weakly electric fishes.

Occurrence of Electroreceptors

For electric communication to occur, an electric fish must have the ability to sense electric signals in water. The electric sense was discovered in only 1961 but has since been found among most primitive orders of fishes, even those that do not possess electric organs. There is good evidence for electroreception in lampreys, sharks, rays, sawfishes, chimeras, lobe-finned fishes, and primitive ray-finned fishes including the bichirs, polypteriformes, sturgeons, and paddlefishes. It may have been present in many extinct groups of fishes. However, this special 'sixth sense' does not occur in most species of modern ray-finned fishes (bowfins, gars, and teleosts) except for two separate phylogenetic lineages or clades of teleosts: the catfishes worldwide plus gymnotiforms of South America and the notopterid plus mormyriform fishes from Africa. In the gymnotiforms and mormyriforms, a novel submodality of tuberous electroreceptors evolved independently. Tuberous sensors are physiologically adapted for detecting high-frequency components of electric organ discharges (EODs). One strongly electric teleost group, the Uranoscopidae, lacks electroreceptors altogether.

The Signaling Environment and Signal Transmission

Electric communication is restricted to aquatic environments where the conductivity of the medium is sufficient to transmit electric signals. When the signaler discharges its electric organ, a dipole-shaped electric field causes current to flow through the water parallel to the electrostatic field lines. Currents are strongest near the source and diminish rapidly with distance. Electric signals are detected by distant receivers as current passes through the electroreceptors embedded in the skin (Figure 1).

Advantages of Electric Communication

Electric communication in fish is used for a variety of behavioral functions, including advertisement of species, gender, and individual identity; courtship; aggression; threat; alarm; and appeasement. Although these are the same functions as acoustic, vibratory, visual, or chemical signals, the special physical constraints on the electric modality have resulted in the evolution of unique design features which have shaped the form of signals, the mechanisms for their reception, and their function.

Little is known of electric communication among the Torpedo rays, the electric skates, the malapterurid catfishes of Africa, or the several groups of catfishes with weak electric discharges (Mochokidae, Clariidae). Although fish in these taxa emit both weak and strong discharges in social contexts, they also do so in isolation, during prey capture and defense. The signals are irregular and difficult to study. Little is known about the social context and behavioral consequences of signals. Most of the information about electric communication comes from studies of gymnotiform and mormyriform fishes (Figures 2 and 3), in which the communication modality is highly complex and well developed.

Dual Functions

Fishes that communicate electrically also use electroreception for detecting objects in the environment, sensed as distortions in the self-generated fields from their own organs. The dual functions of 'electrolocation' and communication help explain the adaptive significance of weak electric organs, which must have been intermediate stages in the evolution of more-powerful electric organs used for prey capture and defense.

Electric Signal Production and Signal Transmission

Electric Organs

Electric organs (Table 1 and Figure 2) are specialized for generating electric currents outside the animal's body. They are usually located in the fish's tail, which prevents current shunting by the body tissues and maximizes signal range. Electric organs are composed

Table 1 The six known groups of electric fishes, arranged taxonomically

Electric fishes	*Picture*
Rajiformes (electric skates): marine, worldwide Rajidae (weak, pulse discharges, ampullary receptors) (14 genera, 200 species) including *Raja*, all with electric organs	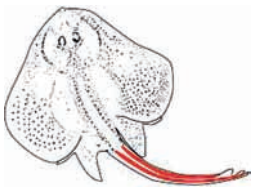
Torpediniformes (electric rays): marine, worldwide Narcinidae (strong, pulse discharges, ampullary receptors) (9 genera, 24 species) including *Narcine* Torpedinidae (strong, pulse discharges, ampullary receptors) (2 genera, 22 species) *Hypnos, Torpedo*	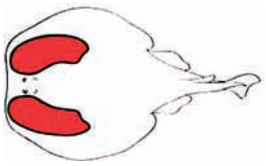
Mormyriformes (African electric fishes): sub-Saharan Africa and Nile River Gymnarchidae (weak, wave discharges, ampullary ND tuberous electroreceptors) (1 genus, 1 species) *Gymnarchus*	
Mormyridae (weak, pulse discharges, ampullary and tuberous electroreceptors) Petrocephalinae (1 genus, 30 species) *Petrocephalus*	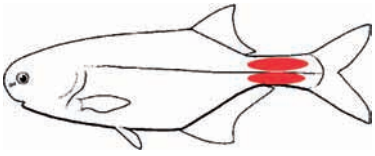
Mormyrinae (16 genera, 177 species) *Boulengeromyrus*, *Brienomyrus*, *Campylomormyrus*, *Genyomyrus*, *Gnathonemus*, *Heteromormyrus*, *Hippopotamyrus*, *Hyperopisus*, *Isichthys*, *Ivindomyrus*, *Marcusenius*, *Mormyrops*, *Mormyrus*, *Paramormyrops*, *Pollimyrus*, *Stomatorhinus*	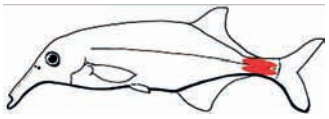
Gymnotiformes (neotropical electric fishes): South and Central America Apteronotidae (weak, wave discharges, ampullary and tuberous electroreceptors) (15 genera, 54 species): *Adontosternarchus*, *Apteronotus*, *Compsaraia*, *Magosternarchus*, *Megadontognathus*, *Orthosternarchus*, *Parapteronotus*, *Pariosternarchus*, *Platyurosternarchus*, *Porotergus*, *Sternarchella*, *Sternarchogiton*, *Sternarchorhamphus*, *Sternarchorhynchus*, *Tembeassu*	
Gymnotidae (strong, weak, pulse discharges, ampullary and tuberous electroreceptors) (2 genera, 33 species): *Electrophorus*, *Gymnotus*	
Hypopomidae (weak, pulse discharges, ampullary and tuberous electroreceptors) (7 genera, 16 species): *Brachyhypopomus*, *Hypopomus*, *Hypopygus*, *Microsternarchus*, *Steatogenys*, *Stegostenopos*	
Rhamphichthyidae (weak, pulse discharges, ampullary and tuberous electroreceptors) (3 genera, 15 species): *Iracema*, *Gymnorhamphichthys*, *Rhamphichthys*	
Sternopygidae (weak, wave discharges, ampullary and tuberous electroreceptors) (5 genera, 30 species): *Archolaemus*, *Distocyclus*, *Eigenmannia*, *Rhabdolichops*, *Sternopygus*	

Continued

Table 1 Continued

Electric fishes	*Picture*
Siluriformes (catfishes): Africa Malapteruridae (strong, pulse discharges, ampullary receptors) (two genera, 18 species) *Malapterurus*, *Paradoxoglanis*	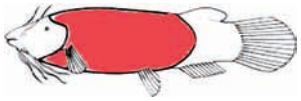
Mochokidae (weak, pulse discharges, ampullary receptors; one genus known to be weakly electric, at least three species with weak electric discharges) *Synodontis*	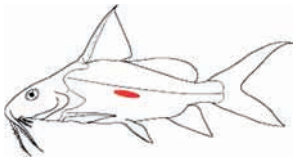
Clariidae (weak, pulse discharges, ampullary receptors; one genus known to be weakly electric, at least one species) *Clarias*	
Perciformes (Stargazers)marine Uranoscopidae (strong, pulse discharges, electroreception absent; two genera, unknown number of species) Astroscopus (three species), *Uranoscopus*	

Each of the species in this list is capable of either weak or strong electrogenesis using specialized electric organs (indicated in red). Some of the electrogenic species have weak electric discharges, and others have strong electric discharges, as indicated. Weak, weak electric organ discharge (EOD); strong, strong EOD; wave, wave-discharging species; pulse, pulse-discharging species.

of cells called electrocytes, derived from nerve or muscle. Weak or strong discharges are produced by the synchronized action potentials from hundreds or thousands of cells acting in series and in parallel. The discharge is initiated in the command nucleus in the medulla and transmitted via a medullary relay nucleus to the electromotor nucleus in the spinal cord. Electromotor neurons innervate the individual electrocytes in the electric organ. For Apteronotid fishes, the electric organ is neural in origin and is simply composed of the spinal electromotor neurons which are anatomically specialized. The discharges of individual electrocytes summate when they are arranged in series, and the currents from individual cells summate when they are arranged in parallel. The voltage of the electric eel, *Electrophorus electricus*, can be an astonishing 500–600 V at the peak of the pulse discharge because of the linear summation of thousands of electrocytes in series. Most of the gymnotiforms and mormyriforms generate only weak discharges of about 10 V – sufficient for short-range communication and active electrolocation but not for prey capture or defense. The discharges of a few notable large mormyrids such as *Mormyrus* and *Mormyrops* can be felt by humans, but most cannot. The discharges were unknown until 1951, when Hans Lissmann from Cambridge University, using electronic amplifiers, discovered continuous 300–500 Hz wavelike discharges emanating from the tail of *Gymnarchus niloticus*, a mormyriform (Figure 2). Several years later, Lissmann demonstrated that this fish could electrolocate objects in its environment if the objects differed in conductivity from the surrounding water. By the 1960s, MVL Bennett, H Grundfest, C Coates, A Fessard, and others had documented EOD diversity in fishes, and T Szabo and TH Bullock and colleagues had made the first electrophysiological recordings from electroreceptors in weakly electric fish. By the 1970s, P Black-Cleworth, P Moller, B Kramer, W Harder, C Hopkins, M Westby, J Bastian, C Bell, TH Bullock, W Heiligenberg, and others had provided strong evidence for the existence of electric communication in weakly electric fishes. Today, researchers working in Africa and South America and in laboratories continue to search for new species, novel signals, and distinctive patterns and functions for electric communication. Neurobiologists have discovered novel mechanisms for electrogenesis and its control, as well as important general mechanisms of information processing and sensory–motor

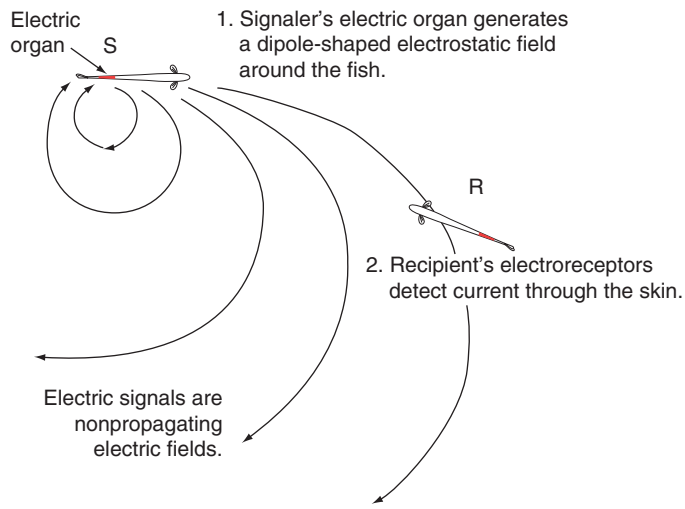

Figure 1 In electric communication, one individual, the signaler, S, generates an electric organ discharge by firing its electric organ, usually located in the tail. This generates a nonpropagating electrostatic field around the fish in three dimensions, approximating the electric field from a dipole. Electric current is sensed by the receiver, R, at a distance, using electroreceptors embedded in its skin. Signalers use pulse waveforms to advertise their gender, age class, and species, and they use patterns of discharges to threaten, show submission, approach, or court other fish.

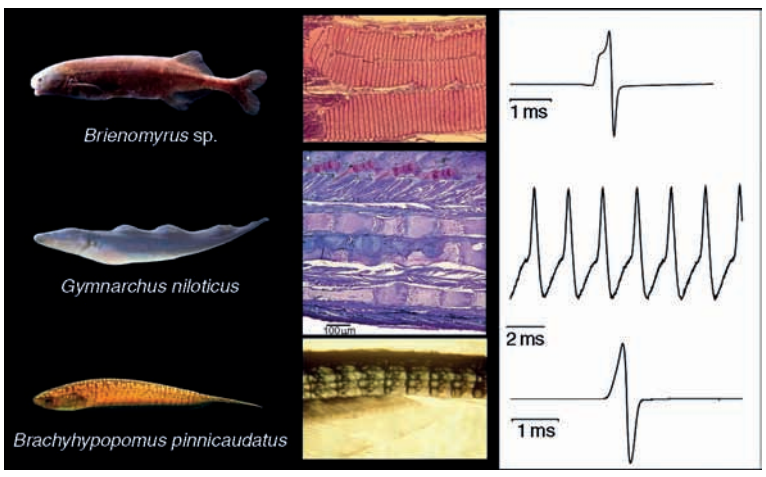

Figure 2 Electric fish, electric organs, and electric organ discharges (EODs) of a mormyrid fish, *Brienomyrus* sp. (top); a mormyriform, *Gymnarchus niloticus* (middle); and a gymnotiform, *Brachyhypopomus pinnicaudatus* (bottom). *Brienomyrus* and *Brachyhypopomus* are both pulse-discharging species whereas *Gymnarchus* is a wave-discharging species. The center column shows histological sections of the electric organ for *Brienomyrus* and *Gymnarchus* and cleared and stained whole mount of Sudan black-stained electrocytes of *Brachyhypopomus*. The EODs are recorded head positive upward in each case. Photo of *Gymnarchus* by Masashi Kawasaki; photo of *Brachyhypopomus* electric organ by JP Sullivan.

integration. Although unusual, electric fish have become an important model system in neuroethology.

The Nature of Electric Signals

There are two components of the electric discharge signals of fish (Figure 4). The first is the stereotyped waveform of the electric organ discharge (EOD), which is fixed for an individual but varies between individuals, genders, and species. The second is the sequence of pulse intervals (SPI), which is generated by the pacemaker or command nucleus in the medulla. Complex neural networks in the midbrain and thalamus modulate activity in the command nucleus, producing the patterns used for communication displays. Both EODs and SPIs are important for this function. EODs are especially important for tonic advertisement of identity, and SPIs are key elements of threats, alarm signals, courtship patterns, and signals of submission. Some patterns and processes that generate them are summarized below.

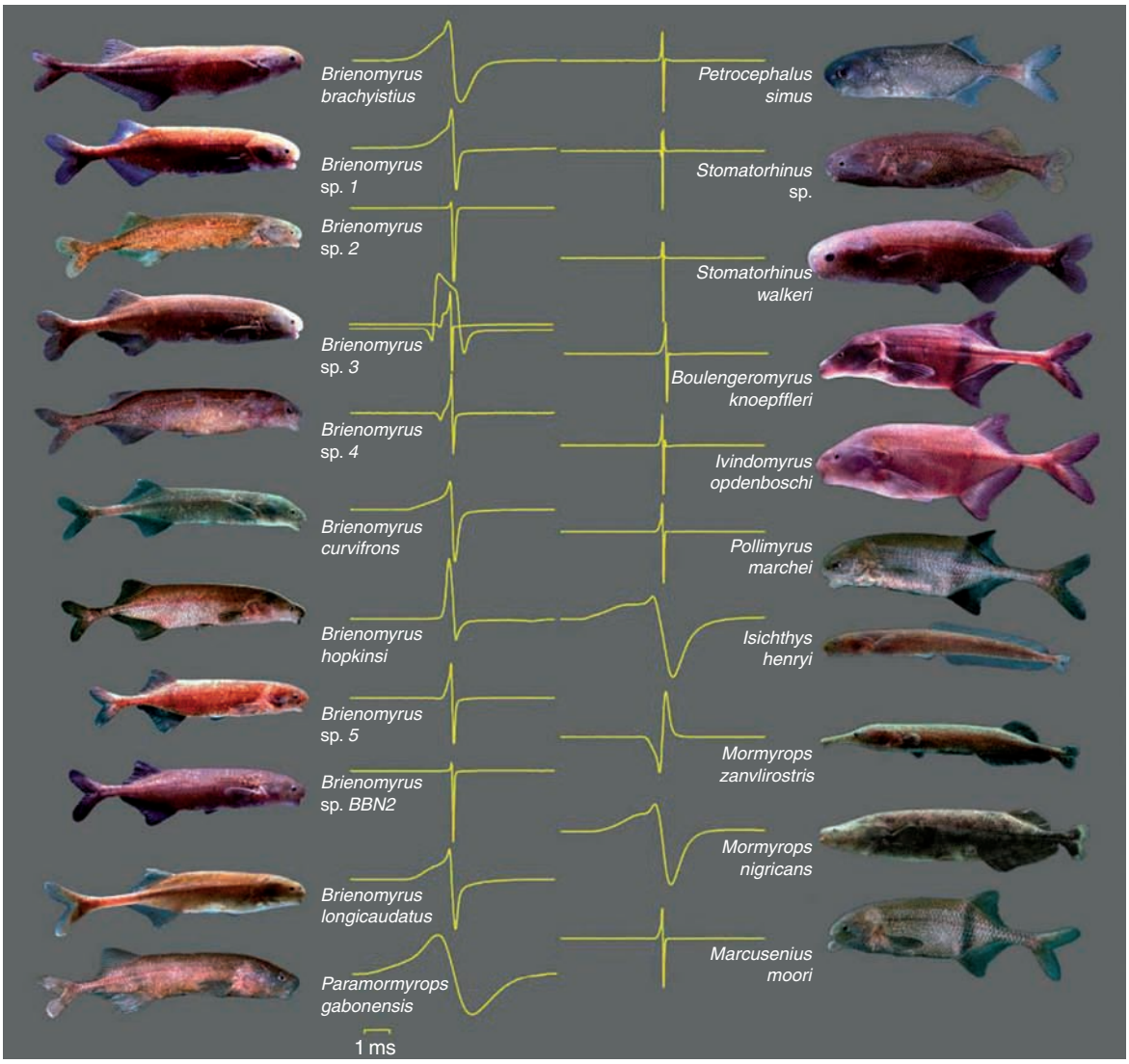

Figure 3 Electric organ discharge (EOD) waveform diversity of mormyrid fishes from the Ogooué and Ivindo River basins of Gabon, West Central Africa. A single EOD waveform typical of the species is shown for each of 21 species. From Sullivan JP and Hopkins CD (2001) Quand les poissons apportent leur pierre et leurs signaux électriques. *Canopée Bulletin sur l'Environment en Afrique Centrale* 19: 17–20.

Signal Transmission: The Geometry of Electric Fields in a Volume Conductor

EODs are transient pulses with durations ranging from 100 μs to 20 ms. They produce nonpropagating electrostatic fields, not propagating electromagnetic waves. This feature contrasts with acoustic communication, in which signals propagate as waves with a finite velocity, generating echoes and reverberations. Electric signals are transmitted instantaneously, with no echoes. Geometric spreading in electrogenesis also differs from sound. Acoustic signal intensity decreases according to the inverse square of distance from the source; electric field magnitude decreases according to the inverse cube of distance.

Figure 5 shows isopotential contours surrounding the gymnotiform *Eigenmannia* at one phase of its EOD cycle. The potential reverses sign at the midplane between the two poles of the electric organ. The electric field, which is computed as the vector derivative of potential, follows a circular arc from positive to negative pole, declining in magnitude in proportion to the inverse cube of distance from the dipole center.

With movable mutielectrode arrays, EOD fields can now be digitally mapped in both time and space. *Eigenmannia* produces a very simple electric field which closely approximates an oscillating dipole with one pole located at the midpoint of the body, the other at the tip of the tail (Figure 6). The location of the

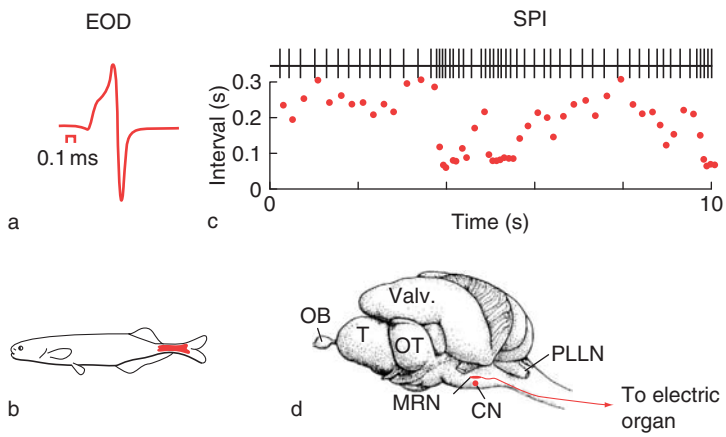

Figure 4 There are two components to electric signals from fish: the fixed electric organ discharge (EOD) waveform and the variable sequence of pulse intervals (SPI). (a) The EOD of *Brienomyrus brachyisitus*, a mormyrid, is an invariant three-phase waveform shown with head-positivity upward. Calibration is in milliseconds. (b) The EOD is generated in the electric organ, which consists of four parallel columns of approximately 100 electrocytes located in the caudal peduncle. (c) The SPI is shown both as a sequence of pulses (upper trace) and as a plot of interpulse interval versus time (lower). (d) The SPI is generated in a command nucleus (CN) in the medulla and sent to the electric organ via the medullary relay nucleus (MRN), and then to the spinal electromotor nucleus. SPIs are highly variable in mormyrids. Other brain structures seen from the side: T, telencephalon; OB, olfactory bulb; OT, optic tectum; Valv., valvula of cerebellum; PLLN, posterior branch of the lateral line nerve, which carriers electroreception to the brain.

poles remains constant throughout the EOD cycle. Other species of gymnotiform fishes are spatially and temporally more complex. *Apteronotus leptorhynchus* generates an electric dipole that appears to propagate from the head to tail and then back again during each cycle of the EOD. The electric field vector recorded at a distance rotates once per EOD cycle. The pulse-discharging hypopomid, *Brachyhypopomus*, has a complex multipole EOD with one dipole located in the head and a second in the tail. This pattern is also seen in several species of *Gymnotus* in which the more anterior parts of the organ contain columns of electrocytes with novel innervation that anticipate the firing of electrocytes in the rest of the organ by a fraction of a millisecond. Other pulse-discharging gymnotiforms have 'accessory' electric organs in the head in addition to a main electric organ in the tail. Both organs fire together, but slightly out of phase. One remarkable Hypopomid, *Steatogenys elegans*, has two accessory electric organs composed of parallel columns of approximately 100 electrocytes in superficial grooves on the underside of head. The communicative significance of accessory organs and the spatial patterns that result is unknown.

Active Space

The active space of a communication signal is the volume of space within which a signal evokes a response from a receiver. The following factors determine its size: the amplitude of the signal at the source, the rate of signal decrease due to geometric spreading, the amplitude of masking or background noise in the environment, and the sensitivity of the receiver. Empirical measurements of the active space surrounding a 10–20 cm weakly electric gymnotiform or mormyrid reveal an ellipsoidal volume approximately 1 m in diameter. Electric signaling is limited to short range because of the rapid falloff of signal amplitude with distance. For an electric fish to double the distance of communication, it must increase the magnitude of the discharge at the source eightfold. This brings high costs to communication and places limits on how far an electric fish can signal. Many species of electric fish exhibit energy-conserving adaptations for signaling, including the slowed pulse repetition rates, reduced amplitude, and shortened pulse waveforms during the daytime, when the fish are inactive. Males also reduce their seasonally elongated pulse widths during the nonbreeding season, when they are not displaying to attract females, and they turn their discharges off when hiding. Some fish even maximize the external signal energy by matching the impedance of the electric organ to that of the environment, discussed next.

The Economy of Impedance Matching

The active space of communication is greatly influenced by water resistivity, as might be expected from the Thevenin equivalence theorem, which states that any electrical source can be modeled as an ideal voltage source in series with a fixed internal resistance. In water, the electric organ is under load, so the external resistance which is in series with the internal resistance experiences a voltage drop proportional to resistance. For a fish in high-resistivity water, the external voltage drop will consequently be higher

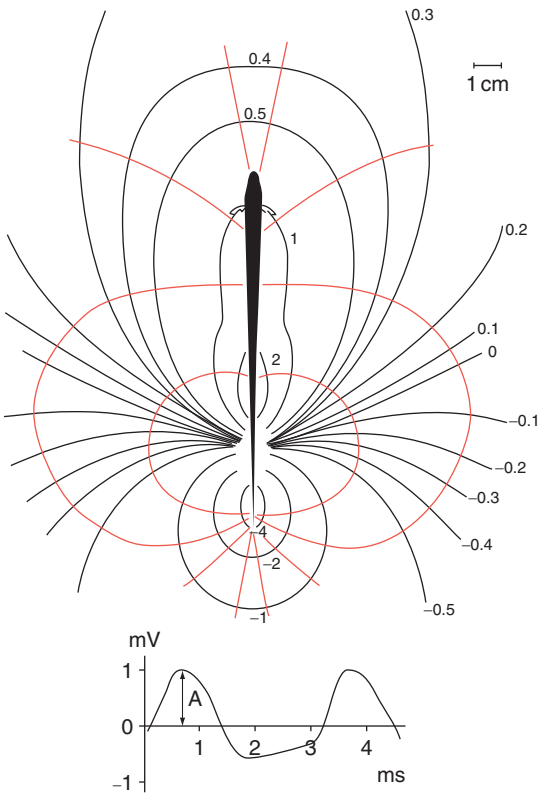

Figure 5 Isopotential lines (black) and current lines (color) at the peak of the electric organ discharge (EOD) of an *Eigenmannia virescens* seen from above and recorded in the midhorizontal plane of the fish. The lower insert shows 1.5 EOD cycles from the head region with a peak voltage amplitude, A, measured relative to an electrode located at infinity. Values of the isopotential lines indicated in the figure are in millivolts. Adapted from Heiligenberg W (1977) Principles of electrolocation and jamming avoidance in electric fish: A neuroethological approach. In: Braitenberg V (ed.) *Studies in Brain Function*, pp. 1–85. New York: Springer.

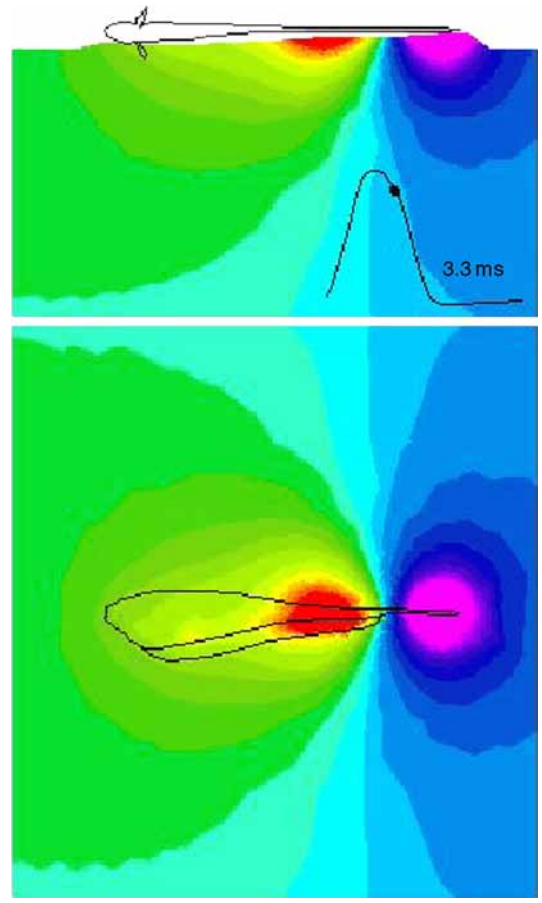

Figure 6 The spatial geometry of the electric organ discharge (EOD) potential from the gymnotiform fish, *Eigenmannia virescens*, as seen from above (above) and from the side (below) as measured using a robotic electrode array. The scalar magnitude of the voltage at one phase of the EOD cycle (shown as inset) is plotted as color contours. The spatial pattern of potential resembles that from an electric dipole. One pole is located at the tip of the tail, the other approximately a third of the body length from the end of the tail. The EOD is a wavelike quasi-sinusoidal discharge. The anterior pole goes positive for approximately 1.5 ms, then negative for the remainder of the EOD. The electric field is computed as the gradient of the potential. Current lines run perpendicular to the isopotential lines shown here. From Assad C, Rasnow B, Stoddard PK, and Bower JM (1998) The electric organ discharges of the gymnotiform fishes: II. Eigenmannia. *Journal of Comparative Physiology. A, Sensory, Neural, and Behavioral Physiology* 183(4): 419–432.

than the internal, so that most of the signal is external, not across the skin. In low-resistivity water, most of the voltage drop will be across the skin, not the environment. A fish adapted to high-resistivity water will experience a serious reduction in active space if it is transferred to water with low resistivity.

Some species of electric fish in the gymnotiform family Hypopomidae appear to be adapted to a narrow range of water resistivity (Figure 7). Those found in relatively low-resistivity (high conductivity) water have short but thick caudal filaments containing the electric organ, with five or more longitudinal columns of electrocytes in parallel. These organs have low internal resistance and are good for generating current, even in low-resistance environments. Fish from high-resistivity (low conductivity) water have extremely elongated caudal filaments containing only three longitudinal columns in parallel. These organs have higher internal resistance but also higher voltage capability, and both are adaptations for generating stronger signals in high-resistivity water. Fish from intermediate resistivity water have intermediate-length caudal filaments containing an intermediate number of columns in parallel. These hypopomids thus appear to have evolved electric organs that are impedance matched to their external environment. Since impedance matching appears to involve rigid restructuring of electric organs, a most interesting consequence may be that these fish might confront invisible barriers when they attempt to disperse into river systems that differ in water

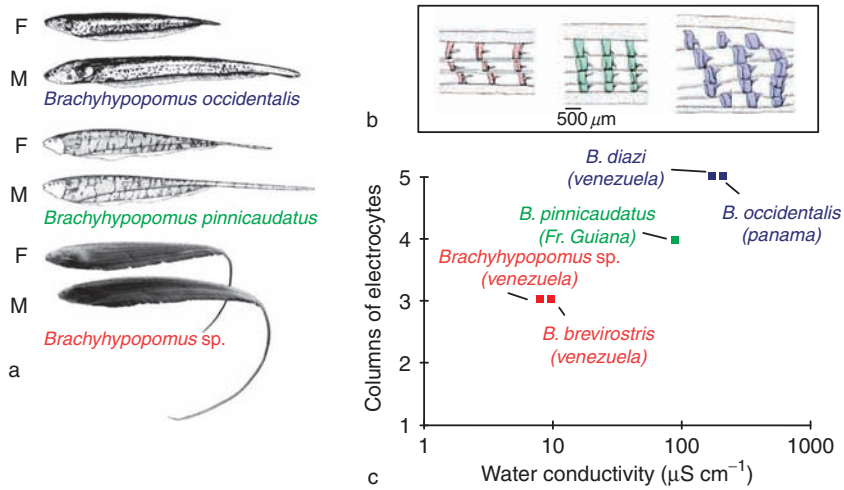

Figure 7 Fishes in the genus *Brachyhypopomus* are impedance-matched to the conductivity of their aquatic environment. (a) Outlines of females (F) and males (M) of three different species adapted to differing-conductivity water. (b) Cross sections through the electric organ in the caudal filament for each of the three species (color coded same as text). (c) Number of longitudinal columns of electrocytes for five *Brachyhypopomus* species as a function of the water conductivity of their environment. The fishes in low-conductivity water tend to have long slender tails with many electrocytes in series but few in parallel. Those in high-conductivity water have many electrocytes in parallel but few in series, and those in intermediate-conductivity water have intermediate characteristics. In all three cases, the length and width of the tail is exaggerated in the male compared with the female of the same body size. From Hopkins CD (1999) Design features for electric communication. *Journal of Experimental Biology* 202(10): 1217–1228.

conductivity from that to which their electric organs are most efficiently matched.

Nonpropagated Signals

Because electric communication signals are nonpropagating electrostatic fields, signals are unaffected by the echoes and reverberations that corrupt acoustic signals during transit. Electric receivers, therefore, may be able to make use of precisely encoded properties of signals, including zero crossing times, details in waveforms on a submillisecond timescale, signal polarity, and minor inflection points on signals that might be altered in transit if propagated as a wave. Electric fish exhibit remarkable sensory specializations for encoding, preserving, and recoding the temporal features of communication signals, often with a precision on the order of microseconds.

Since there is no delay to an electric communication signal, receivers will not be able to use signal arrival times at different parts of the body surface to localize electric sources. Nevertheless, research has shown that electric fish can align their body axis parallel to the local electric field vector as an alternative strategy for finding signal sources even if the pathway leading there might be curved.

Noise

The greatest source of environmental electrical noise in the signal bandwidth of electric fish comes from distant lightning activity, which propagates globally over hundreds or thousands of kilometers as electromagnetic waves. The signature clicks, 'chirps', 'tweeks', and other disturbances from lightning can be recorded continuously using dipole electrodes immersed in water. Although there is no evidence that electric fish respond to this noise, quantitative measurements suggest that its magnitude in different environments correlates with the sensory threshold of electroreceptors of the inhabitants. In marine environments, where the magnitude of noise from lightning is 0.01 that found in freshwater, electroreceptor thresholds tend to be 100 times more sensitive than those of freshwater fish. Many resting mormyrid seem to take advantage of this noise, blending into it by discharging at irregular intervals when they are alone and standing out with more regular intervals of discharge in the presence of conspecifics.

Signal Reception

Electric communication depends on sensitive lateral-line-derived electroreceptors located on the head and trunk of certain fishes. These sense organs direct the flow of current through low-resistance canals and through loosely layered patches of epithelial cells to a sensory epithelium where specialized transducer cells carrying voltage-gated ion channels in their cell membranes convert outside electrical stimulation into internal neural responses. Receptor organs filter the stimulus waveform, impart directionality to signal reception, adjust the sensitivity of the receptor, and

ensure that the sensory epithelium is ideally positioned for maximal sensitivity.

Electrosensory cells make synaptic contacts with afferent nerve fibers which project to the electrosensory lateral-line lobes in the medulla via the acusticolateralis or VIIIth nerves. Ampullary receptors are found in every electroreceptive species; tuberous receptors are unique to gymnotiforms and mormyriforms. Receptors resembling tuberous organs were once reported from a species of blind catfish, *Pseudocetopsis*, from South America, but this has not been confirmed physiologically. Figure 8 shows the morphology of ampullary and tuberous electroreceptors in mormyrids, *Gymnarchus*, and gymnotids.

Ampullary Electroreceptors

Ampullary receptors have a sensory epithelium at the base of a blind, mucus-filled, low-resistance canal leading from the external environment. Receptor cells are sensitive to electric stimuli in the range of 0.05–50 Hz. The receptor cells contact tonically active afferent nerve fibers and modulate their firing rate upward or downward depending on stimulus polarity.

Ampullary receptors typically function in the passive electrolocation of predators and prey. For example, sharks and rays detect fish buried in sand by sensing the weak electric fields of a fraction of 1 μV cm^{-1}. In tanks they dig up buried electrodes playing electric potentials from prey. Male round stingrays

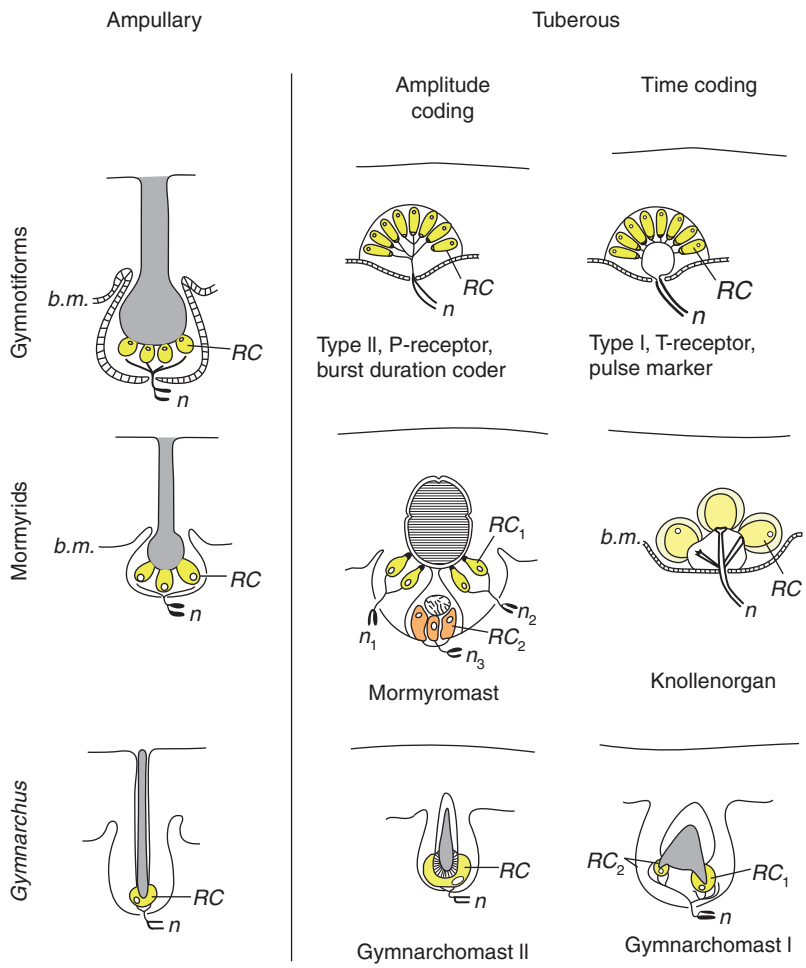

Figure 8 Anatomy of ampullary and tuberous electroreceptors in the freshwater teleosts: Gymnotiformes, Mormyrids, and *Gymnarchus*. Ampullary receptors have a highly conducting jelly-filled canal leading from the receptor epithelium at the base of the ampoule to the skin surface; tuberous receptors have a loose plug of epithelial cells over the receptor organ. Mormyromasts have two receptor subclasses, RC_1 and RC_2, which are innervated by separate nerves: two nerve axons innervate RC_1 receptors (also called Type A receptors); one axon innervates RC_2 receptors (Type B). Tuberous receptors function in amplitude coding or in time coding. RC, receptor cells; n, nerve axon; b.m., basement membrane. Adapted from Szabo T (1965) Sense organs of the lateral line system in some electric fish of the Gymnotidae, Gymnarchidae, and Mormyridae. *Journal of Morphology* 117: 229–250 and Hopkins CD (1999) Design features for electric communication. *Journal of Experimental Biology* 202(10): 1217–1228.

(*Urolophus*) use ampullary receptors to find females which produce low-frequency electric signals during ventilation. Electric rays, skates, and catfishes use ampullary receptors to sense conspecific EODs.

In addition to prey sensing, wave gymnotiforms use ampullary organs for sensing brief discharge interruptions (see below). A resting-wave EOD lacks energy in the range of 0.05–50 Hz and will not stimulate ampullary receptors, but during discharge cessations, a residual head-negative offset strongly stimulates them.

Tuberous Electroreceptors

Tuberous organs have an interepidermal cavity lined with sensory-receptor cells but lack an open canal to the exterior. Instead, a plug of loose epithelial cells covers the receptor and behaves like a series capacitor to block direct currents. Tuberous organs are sensitive to stimuli in the range of 100–10 000 Hz and are used for communication and for active electrolocation of objects. They are electrically tuned to the most prominent frequencies of the species-typical EOD (Figure 9). Mormyrid fishes have two types of tuberous organs: Knollenorgans, which are used for communication, and mormyromasts, which are used for active electrolocation (Figures 8 and 9). Gymnotiforms also have two physiological types of tuberous organs (Figure 8), but each is used for both communication and electrolocation. One receptor type is specialized for detecting stimulus amplitude; the other is specialized for timing.

Three lines of evidence suggest that Knollenorgan electroreceptors are used exclusively for communication. First, they have low thresholds, so they are ideally suited for detecting the weak signals from distant fishes; second, the spikes from Knollenorgan afferents are inhibited, or 'blanked,' in the hindbrain by neurons carrying a corollary discharge of the EOD command so that a mormyrid never 'hears' its own EOD on the Knollenorgan pathway; and third, lesions to the nucleus *exterolateralis* in the midbrain, which receives all Knollenorgan receptor inputs, blocks responses to all electric communication signals while having no effect on a fish's ability to electrolocate.

Tuberous receptors are electrically tuned to the peak of the power spectrum of the species-specific EOD waveform both in gymnotiforms and mormyrid fishes (Figure 9).

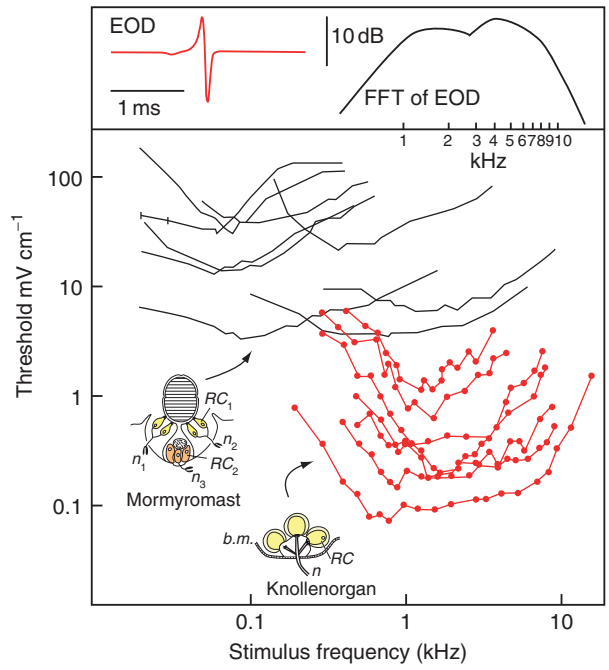

Figure 9 Tuberous electroreceptors in mormyrids are electrically tuned to frequencies characteristic of the fast Fourier transform (FFT) of the electric organ discharge (EOD). Tuning curves are generated by measuring the spiking threshold of single receptor afferents of Mormyromasts (solid lines) or Knollenorgans (lines with filled symbols in color). Each curve represents a separate electroreceptor. The EOD waveform is shown in the upper left and the FFT power spectrum of the EOD in the upper right. Brienomyrus brachyistius. From Hopkins CD (1983) Functions and mechanisms in electroreception. In: Northcutt RG and Davis RE (eds.) *Fish Neurobiology*, vol. 1, pp. 215–259. Ann Arbor: University of Michigan Press.

Temporal Dimensions of Social Communication Signals

Electric signals have both tonic features, which are invariant for an individual over a period of days, months, or years, and phasic features, which change from moment to moment, depending on social conditions.

Tonic Signals in Electric Communication

An individual's signature EOD, typically a pulse waveform lasting only 1 ms or more, is one example of a tonic signal. It is typically invariant for days or months and serves to advertise the species, sex, and even individual identity of signalers. The EODs from different species diverge from one another (Figure 10). Pulse duration, number of phases or peaks, order polarity of peaks, and characteristic inflection points all contribute to a signaler's signature waveform.

EOD waveforms have communicative significance. Males will court electrodes playing female EODs while ignoring the same electrodes playing male EODs or heterospecific female EODs. They will ignore female EODs played backwards, as well as phase-shifted EODs (Figure 11), suggesting that the precise form of the EOD is important, not the spectrum of the signal. In some genera, such as *Brienomyrus*, *Campylomormyrus*, *Marcusenius*, and *Mormyrops*,

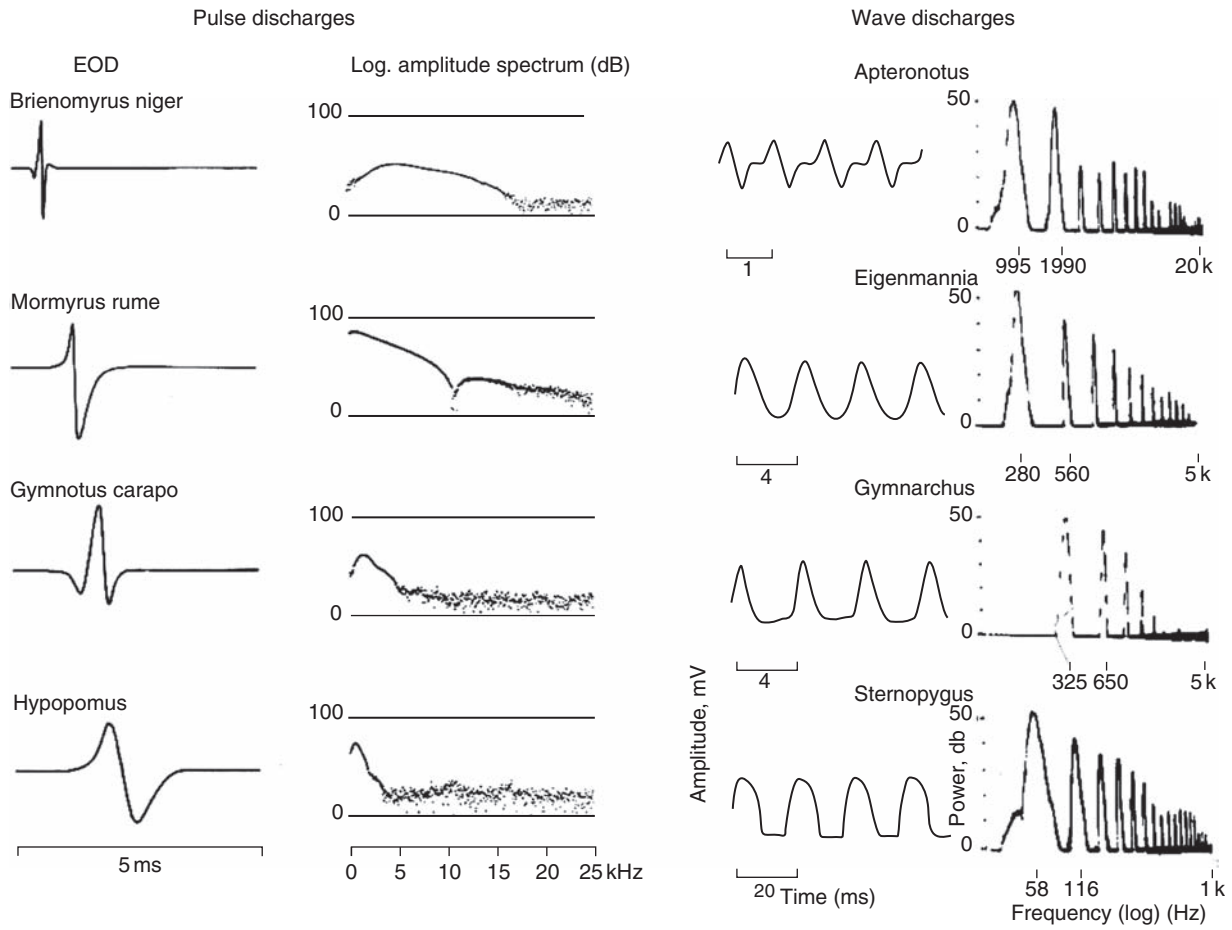

Figure 10 Pulse-discharging species of electric fishes in the families Mormyridae, Gymnotidae, and Hypopomidae all produce electric discharges which have a low duty cycle and variable intervals. Wave-discharging species in the families Apteronotidae, Sternopygidae, and Gymarchidae produce quasi-sinusoidal discharges with high duty cycle and extremely regular interpulse intervals. The waveform of the discharge is shown to the left in each case, and the amplitude spectrum of the discharge is shown to the right. From Heiligenberg W (1977) Principles of electrolocation and jamming avoidance in electric fish: A neuroethological approach. In: Braitenberg V (ed.) *Studies in Brain Function*, pp. 1–85. New York: Springer.

EOD waveforms are highly divergent within close relatives (congeners) living in sympatry. In some species of *Mormyrops* and *Brienomyrus* from Central Africa, EODs may even be polymorphic, exhibiting more than one stable EOD type within a population.

Of course the EOD waveform is also used for active electrolocation of objects, so its properties may be adapted to navigation as well as to advertisement and identification. Consider, for example, that many mormyrid fish sense the electrical capacitance of objects, and the duration of the EOD waveform duration sets the range of capacitances over which they are sensitive. By adopting an EOD waveform with a short rather than a long duration, a fish may be better at detecting low-capacitance objects. Consider also that most mormyriforms and gymnotiforms produce pulse discharges in which positive and negative phases of the EOD are balanced so that there is very little signal energy below 10 Hz. This appears to be an important adaptation for avoiding exposure to predatory catfishes, which have ampullary electroreceptors sensitive to low-frequency or DC components of signals. For electrolocation, which must function continuously, electric fish must avoid signals with DC components, but for communication displays which are infrequent, DC may be used. In several species of Hypopomidae (pulse-discharging gymnotiforms), juvenile and female EODs are balanced and have little energy at low-frequency while breeding males have seasonal EODs with stronger DC components. Males of the gymnotid *Brachyhypopomus pinnicaudatus* modulate their EOD waveform over 24 h so that DC components of the discharge are present only at night, when the fish are active, and only during the breeding season. A similar pattern is seen in mormyrids. In *Marcusenius macrolepidotus* from South Africa, males have a stronger DC component to their discharge than females do, and catfish predators

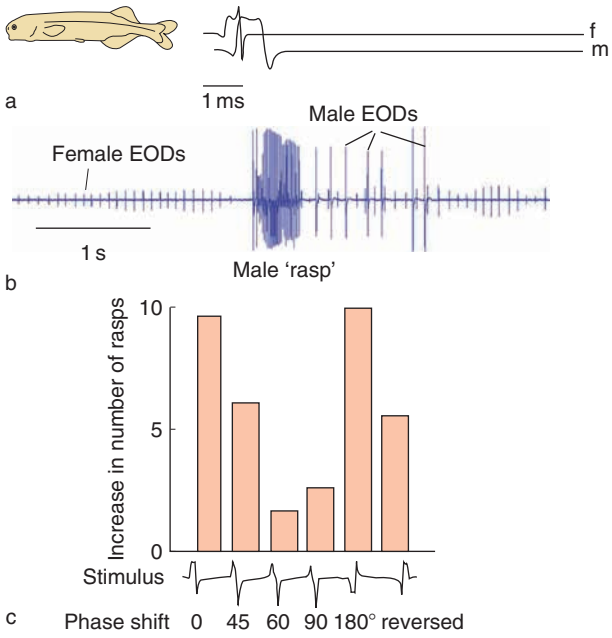

Figure 11 (a) A species of mormyrid in the genus *Brienomyrus* from Gabon exhibits a gender difference in the electric organ discharge (EOD) waveform, the pulse duration for the male (m) being nearly double that for the female (f). During the breeding season, which starts with the seasonal rains, males court passing females by producing bursts of EODs at high frequency, making stereotyped displays called 'rasps.' (b) In experiments, males produce rasps in response to digitized EOD waveforms of females, as shown in the histograms of responses in (c). If the female EOD is phase-shifted by 45°, 60°, or 90° or if the EOD is reversed (played backwards in time), the male shows reduced responsiveness, but if the female EOD is phase shifted by 180°, responses are as strong as for the normal female EOD. This is not surprising given that a 180° phase shift is equivalent to inversion of polarity of the waveform, which would happen if the a real female were to turn around in space. These experiments strongly suggest that males attend to the temporal properties of EOD waveforms, not the amplitude spectral characteristics, which are all identical. In addition to courting digitized female EODs, males will also court a highly simplified square wave as long as the duration of the square wave is 0.5 ms, a close approximation to the duration of the head-positive phase of the female EOD (not shown). This can be explained by showing that female EODs and 0.5 ms square waves evoke from Knollenorgan electroreceptors the same neural code based on spike timing. (c) Reprinted from Hopkins CD and Bass AH (1981) Temporal coding of species recognition signals in an electric fish. *Science* 212: 85–87, with permission from AAAS.

seem to take more of the long-duration EOD males than females, as revealed by analysis of stomach contents of *Clarias gariepinus* in South Africa. *Clarias* appears to use its ampullary receptors to detect the DC component of the male EOD.

The physiological basis of EOD waveform diversity – whether between species or between sexes – traces to the type, the density, and the distribution of ion channels in electrocyte membranes; to the anatomy and pattern of innervation of electrocytes within the electric organ; and to the diversity of accessory electric organs in the fish. Much of this diversity is understood at a gross anatomical level but has yet to be fully explored at the molecular level. This is an area of active research that will undoubtedly reveal interesting patterns of evolution, especially within species flocks which show a large diversity of EOD waveforms.

Figure 12 shows variation in anatomy of electrocytes of selected species of mormyriforms arranged on a cladogram which is derived from analysis of mitochondrial and nuclear DNA sequences. Superposition of the pattern of electrocyte morphology on the phylogenetic tree gives insight into the evolution of the electric organs. Complex EOD waveforms emerge in the family Mormyridae very early with the evolution of penetrating stalked electrocytes. Reversions to nonpenetrating electrocytes with simpler EOD waveforms occur frequently. There are also cases in which electrocyte innervation is reversed anterior for posterior and others in which stalks are doubly penetrating electrocytes. All this diversity illustrates how signal diversity parallels species diversity in this remarkable radiation (Figure 12).

A second component to tonic diversity arises from the firing pattern of the pacemaker or command nucleus in the medulla. Wave discharges differ from pulse discharges in both the duty cycle of the discharge waveform and the variance in interpulse intervals. Wave discharges have a high duty cycle and low interval variance; pulse discharges have a low duty cycle and a high interval variance. Figure 10 shows wave versus pulse discharges of Gymnotiforms and Mormyriforms.

Wave or pulse discharges vary by taxonomic lineage. For example, all the gymnotiform species in the families Apteronotidae and Sternopygidae have wave discharges (Figure 10), whereas all others have pulse discharges. *Gymnarchus niloticus* has a wave discharge, but all other mormyriforms have pulse discharges (Figures 10 and 13). Wave discharges are a derived character in gymnotiform signal evolution; only the Apteronotidae and Sternopygidae produce waves. It is unknown whether the common ancestor of Gymarchidae (wave) and Mormyridae (pulse) was pulse or wave.

The average frequency of a pulse or wave discharge is another important tonic signal. It is controlled in the pacemaker in the medulla. Discharge frequency varies between species and sometimes between sexes. The first documented case of a sex difference in electric signaling was the firing rate of the gymnotiform wave species, *Sternopygus macrurus*. Adult males discharge in the range of 50–90 Hz, and females discharge in the range of 100–150 Hz (Figure 14(a)). This tonic feature has communicative significance, as shown by playing pure sine waves to male *Sternopygus*

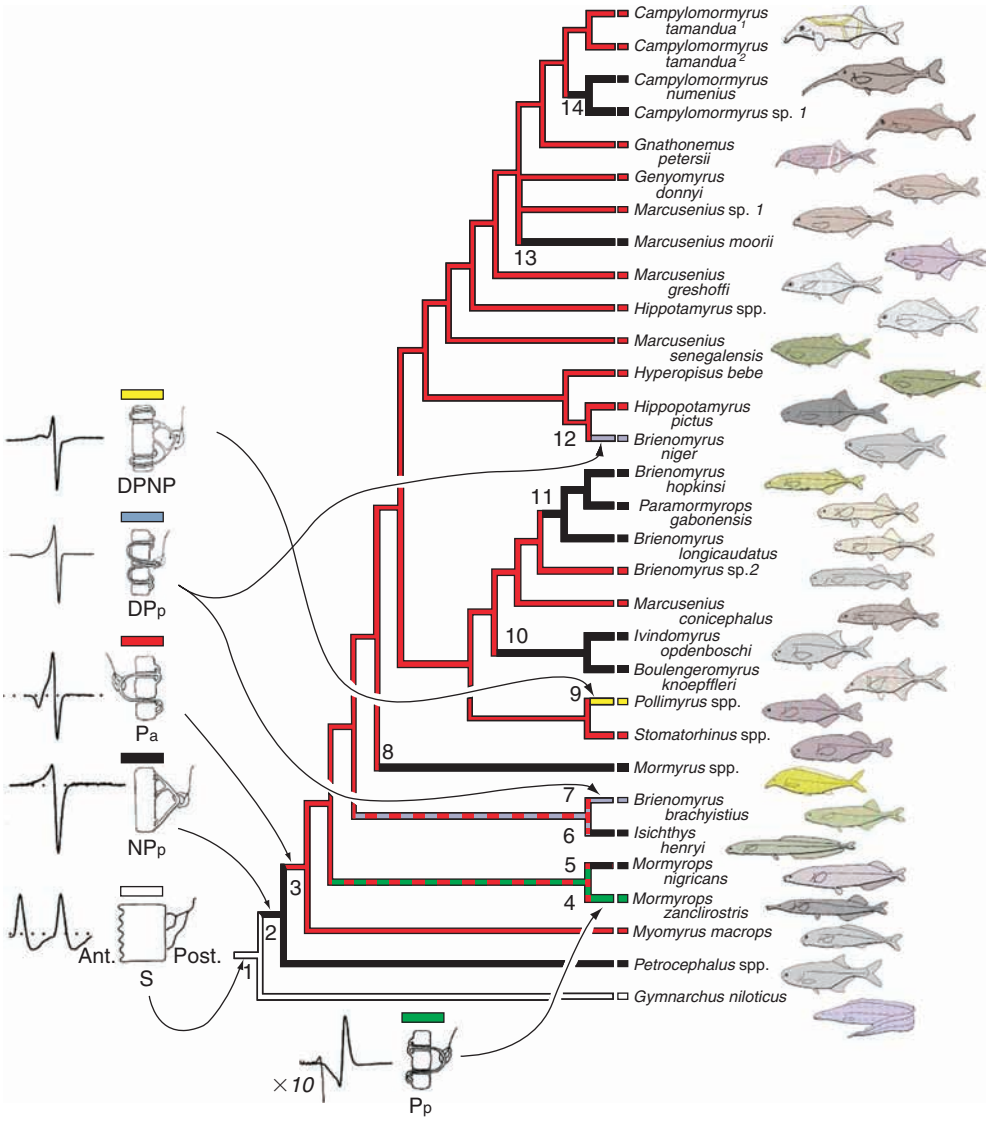

Figure 12 A hypothesis for the evolution of electrogenesis among mormyroid fishes of Africa based on a phylogenetic tree of *Gymnarchus* and 31 species of mormyrids. The tree, which is derived from mitochondrial and nuclear DNA sequence data, is largely consistent with classifications based on osteology and other morphological characters. Taxa sampled are shown as outlines to the right. The structure of electrocytes in the electric organs is shown to the left, color coded. Type S electrocytes are 'stalkless.' They are found in adult *Gymnarchus niloticus* and in larval stages of all known mormyrids. The innervated faces of these electrocytes are nonspecialized, lacking the 'stalk' system which characterizes electrocytes in adult organs of mormyrids. Type S electrocytes are innervated on the posterior face (post.), which fires an action potential that produces a head-positive monophasic electric organ discharge (EOD). The anterior face (ant.) is electrically unexcitable. *Gymnarchus* has a wave type EOD in which the monophasic potentials are superimposed on a head-negative DC potential. All the fishes in the family Mormyridae have electrocytes with 'stalks.' The electrocytes are located in an adult electric organ in the caudal peduncle. All the species in the genus *Petrocephalus*, toward the base of the tree, have stalked electrocytes in which the stalks are 'nonpenetrating' and are innervated on the posterior side (type NP_p). Many of the Mormyridae have electrocytes with type P_a 'penetrating' stalks in which an elaborate stalk system, innervated on the anterior side, penetrates through the electrocyte to fuse with the posterior face. Action potentials are generated in the stalk, in the posterior face, and in the anterior face. The EOD is typically triphasic. Penetrating stalked electrocytes appear first in the genus *Myomyrus* at the base of the phylogenetic tree. Penetrating stalked electrocytes appear throughout the family tree, although there are multiple paedomorphic reversals to nonpenetrating stalked electrocytes. Two additional morphologies are known: doubly penetrating with nonpenetrating stalked electrocytes (DPNP) and penetrating stalks with posterior innervation (P_p). These occur in isolated lineages. Tracing the evolution of electrocyte diversity gives insight into the origin of signal complexity in this group of fishes. In this reconstruction, arrows point to the hypothesized origin of a novel electrocyte morphology. Representative EOD waveforms accompany each type of electrocyte. Alternating color bars to the left of a clade indicate that the ancestral type of penetrating stalk electrocyte is equivocal. From Sullivan JP, Lavoué S, and Hopkins CD (2000) Molecular systematics of the African electric fishes (Mormyroidea: Teleostei) and a model for the evolution of their electric organs. *Journal of Experimental Biology* 203(4): 665–683.

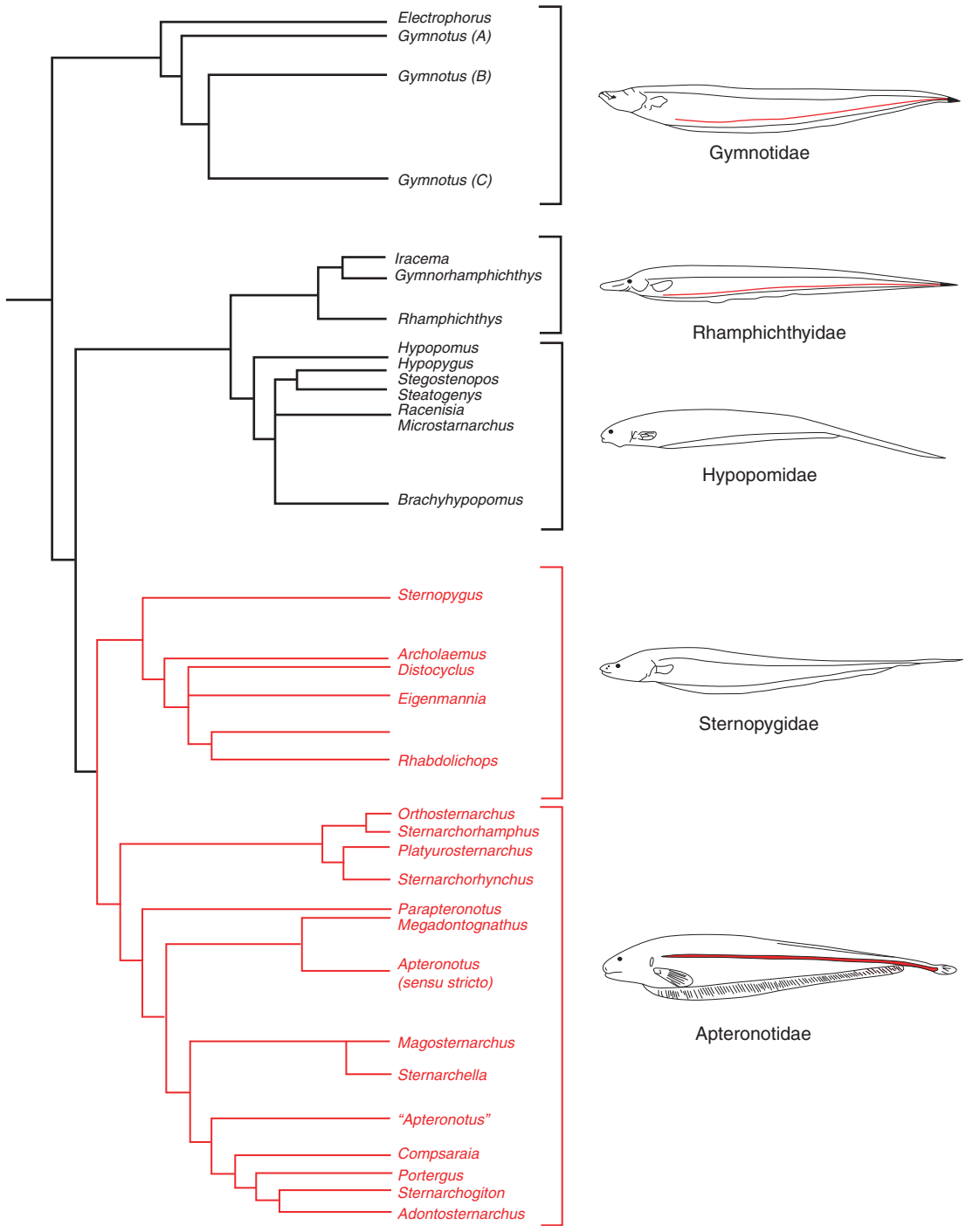

Figure 13 Phylogenetic hypothesis for gymnotiforms from South and Central America derived from morphological and molecular sequence data and representing 88 of the more than 170 species of gymnotiforms that have been described. Those fishes with wave discharges are indicated in red, those with pulse discharges in black. From Albert JA and Crampton WGR (2005) Diversity and phylogeny of neotropical electric fishes (Gymnotiformes). In: Bullock TH, Hopkins CD, Popper AN, and Fay RR (eds.) *Electroreception*, vol. 21, pp. 360–409. New York: Springer.

at night during their breeding season in small creeks in Guyana. The males responded with electrical courtship songs (Figure 14(b)) whenever the stimulus frequencies were in the females' range of 130–150 Hz (see Figure 14(c)). The males did not sing to electrodes playing sine waves in the male range or to signals in the ranges of two other sympatric wave species, *Eigenmannia* and *Apteronotus* (Figure 14(c)).

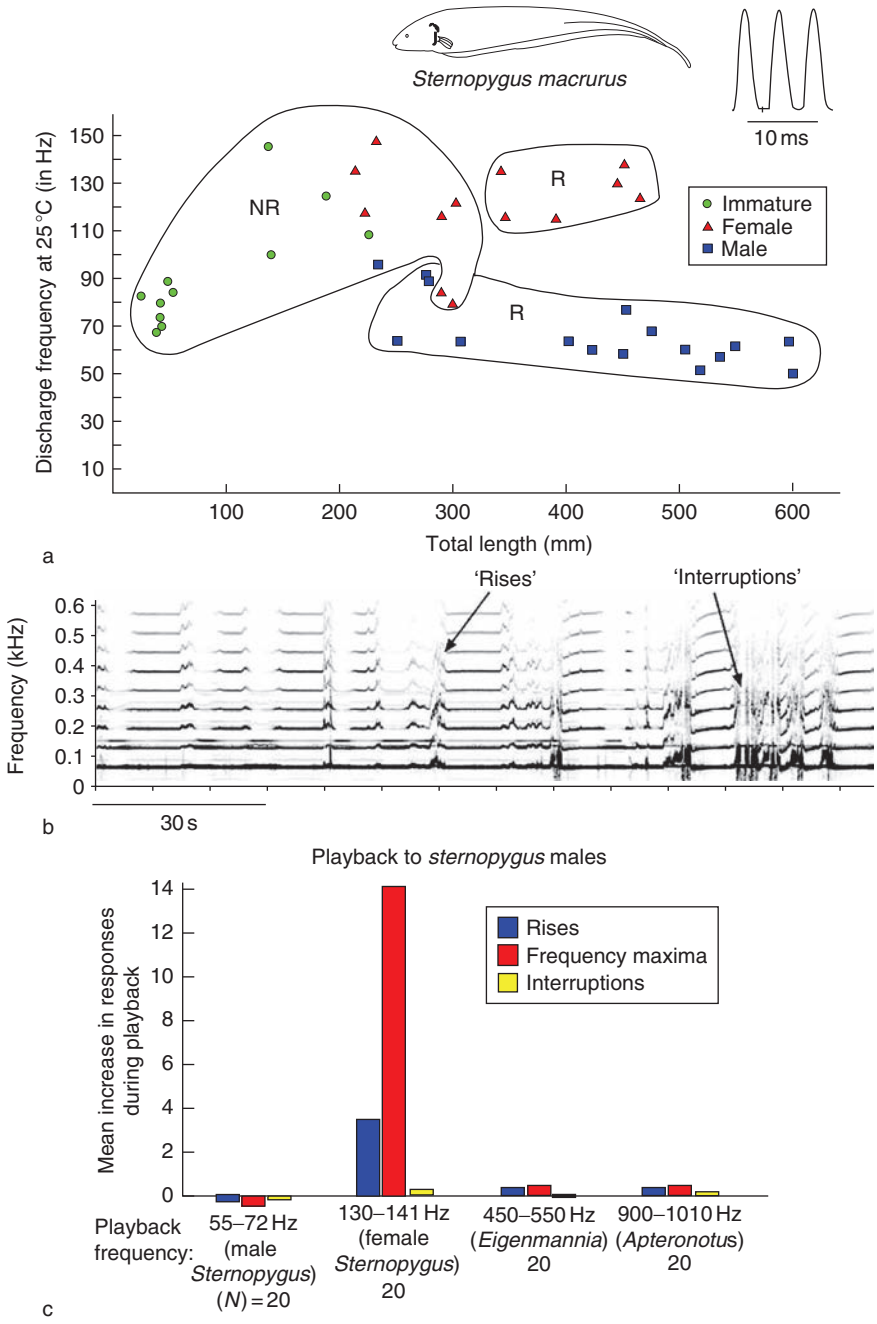

Figure 14 Sex differences in the discharge frequency of *Sternopygus macrurus* observed under field conditions in Guyana. (a) A plot of discharge frequency versus standard length for males (squares), females (triangles), and juveniles (circles) shows that sex differences emerge in individuals longer than 300 mm. The inset shows the fish's outline and a sample of the electric organ discharge waveform. From Hopkins CD (1974) Electric communication in the reproductive behavior of Sternopygus macrurus (Gymnotoidei). *Zeitschrift für Tierpsychologie* 35: 518–535. (b) Sound spectrogram of a male *Sternopygus* producing electrical courtship song in presence of a female. The male's discharge is a wave discharge in the 50–70 Hz frequency range. During courtship, males produce 'rises' in discharge frequency and 'interruptions' in the discharge song. (c) When presented with an electrical sine wave stimulus, delivered via electrodes placed in the water near the fish, the male *Sternopygus* gives courtship rises and interruptions if the frequency of the sine wave is in the 130–141 Hz range of a female *Sternopygus*. The male does not court sine wave frequencies in the 55–72 Hz range typical of other males, nor in the 450–550 Hz range typical of *Eigenmannia*, nor in the 900–1010 Hz range typical of Apteronotus, two other species that are sympatric to Sternopygus. (C) Reprinted from Hopkins CD (1972) Sex differences in electric signaling in an electric fish. *Science* 176: 1035–1037, with permission from AAAS.

Phasic Signals

Phasic signals vary rapidly in response to changes in the social environment, typically over a period of seconds or minutes. Phasic signals are produced by modulating the firing rate of the command or pacemaker nucleus in the medulla or by interrupting or silencing the discharge in the pacemaker or in the descending neural pathway from the command nucleus. Pulse and wave fish both produce two types of phasic signals: frequency modulations and discharge interruptions. Frequency modulations are either upward or downward shifts in the pacemaker frequency, last seconds to minutes, and consist of small shifts in frequency (long rises, low-frequency modulations, warbles, and jamming avoidance responses) or large shifts in rate (sudden increases in frequency, followed by decrease, 'chirps,' 'rasps,' 'scallops,' 'accelerations,' and a variety of other types of pulse bursts). Discharge interruptions are either short cessations in the discharge, lasting less than 0.5 s, or long-lasting periods of silence. All these phasic signals contrast with the constant rate of the tonic discharge.

Figure 14(b), showing the courtship singing of a male *Sternopygus macrurus*, illustrates both types of signals. There are short interruptions in the 50–90 Hz wave discharge, lasting less than 0.5 s, and there are brief modulations in the fundamental frequency upward by as much as 50 Hz above base rate. Interruptions and frequency modulations make the complex patterns of song, which may continue for hours at night during the breeding season.

There are striking parallels in the phasic displays of *Gymnarchus* and *Eigenmannia*, which independently evolved wave discharges in the range of 300–500 Hz on separate continents (Figure 15). Both produce short interruptions during attacks, approaches, head butts, and other aggressive acts. These acts are most common from dominant fish. They are rare from subordinates.

They both also produce subtle frequency modulations as displays of submission. Subordinate *Eigenmannia* produce long rises, which are elevations in frequency in the range of 5–10 Hz, when being attacked (Figure 15), and subordinate *Gymnarchus* produce similar displays if attacked, but the frequency decreases by 5–10 Hz rather than increasing (Figure 15). Subordinate *Gymnarchus* turn their discharges off altogether for many minutes or hours when attacked. The similarity in both form and function of these displays is striking, considering their independent evolution in Africa and South America.

Pulse fish also produce phasic displays, as illustrated in Figures 11 and 16, and these can be evoked by playback of the appropriate stimulus. Figure 11

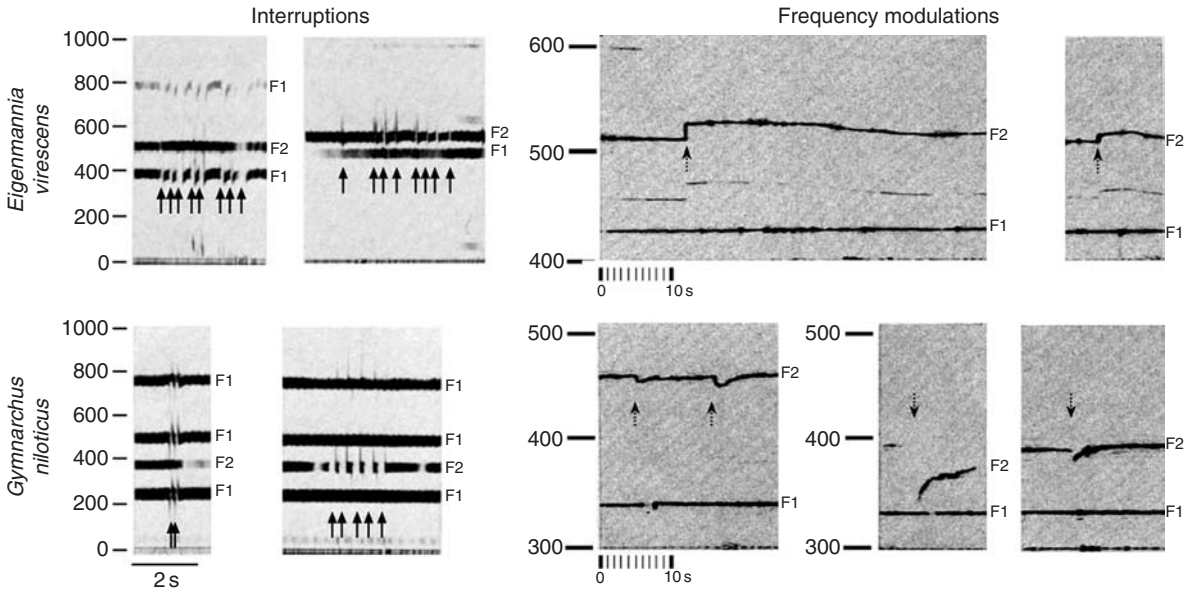

Figure 15 Sound spectrograms comparing the communication displays of *Eigenmannia virescens* from South America with those of *Gymnarchus niloticus* from Africa. Both species have wave discharges in the same range (300–500 Hz), and both produce discharge interruptions as aggressive threat displays (solid arrows) and frequency modulations (dashed arrows) as signals of submission. The sound spectrograms on the left are on an expanded timescale and broad bandwidth so that the rapid temporal features of the displays may be seen. The spectrograms on the right have a compressed timescale and narrow bandwidth to better visualize the small but slow changes in frequency over time. There are two fish in each interaction; the lower-frequency fish is labeled F1, the higher, F2. Fundamental and higher harmonics are clearly evident for fish 1 on left. Adapted from Hopkins CD (1977) Electric communication. In: Seboek TA (ed.) *How Animals Communicate*, pp. 263–289. Bloomington: Indiana University Press.

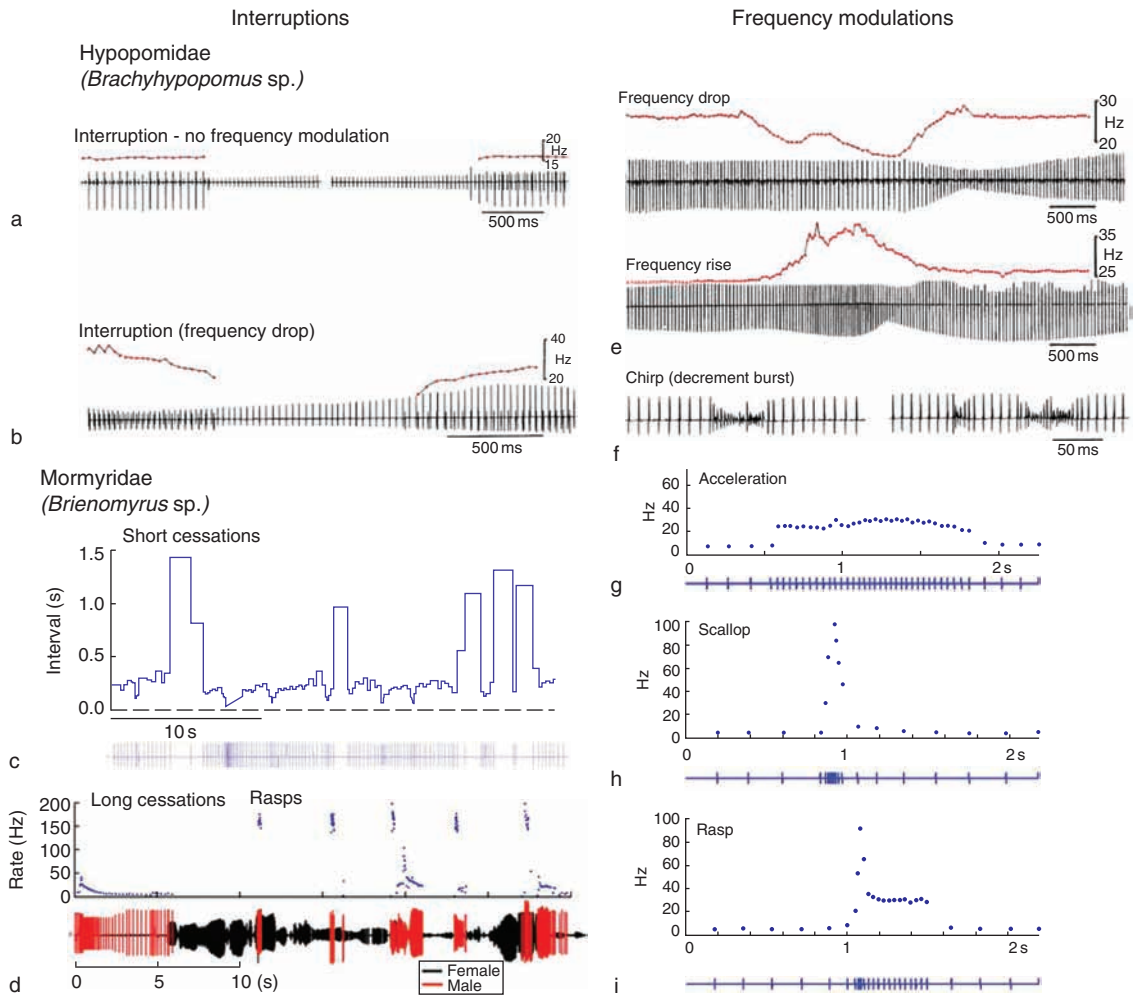

Figure 16 Two types of temporal displays, 'interruptions' and 'frequency modulations,' produced by pulse-discharging gymnotiforms in the family Hypopomidae (above) and Mormyridae (below) as illustrated by waveforms and pulse rate plots. The gymnotiform *Brachyhypopomus brevirostirs* produces discharge interruptions by turning its discharge off for brief or extended periods. Shown here are two types: one (a) in which the discharge is turned off suddenly without a decrease in pacemaker frequency, and the other (b) preceded by a frequency drop. Similar patterns are seen for mormyrids in the genus *Brienomyrus* (c) and (d). In (c), pulse intervals are plotted as a function of time; in (d), pulse rate is plotted against time. During agonistic behavior, a fish may turn its discharge off for short periods (less than 1–2 s) as an aggressive threat. During courtship, males may turn their discharges off for several seconds, in alternation with rasp displays. Cessations contrast with the regular resting discharge intervals of 250–500 ms. Both groups of pulse fish also modulate the frequency of their discharge. (e), (f) *Brachyhypopomus* also produces frequency drops, frequency rises, and 'chirps' or 'decrement bursts.' Chirps are male courtship signals. *Brienomyrus brachyistius* (g)–(i) produces accelerations, scallops, and rasps as communication displays, with the rasp display used for male courtship calling. (a, b, e, f) From Kawasaki M and Heiligenberg W (1988) Distinct mechanisms of modulation in a neuronal oscillator generate different social signals in the electric fish Hypopomus. *Journal of Comparative Physiology A* 165: 731–741. (g, h, i) From Carlson BA and Hopkins CE (2004) Central control of electric signaling behavior in the mormyrid *Brienomyrus brachyistius*: Segregation of behavior-specific inputs and the role of modifiable recurrent inhibition. *Journal of Experimental Biology* 207: 1073–1084.

shows the 'rasp' display from a pulse-discharging mormyrid. The displays are strong, sudden bursts of pulses, up to 100 s^{-1}, lasting 200–400 ms. Males make these high-frequency bursts when courting females.

Some phasic displays in mormyrids consist of 'regularization' of the irregular tonic discharges. Female *Pollimyrus adspersus* regularize their discharge at low rates when entering a male's territory for spawning. This signal appears to indicate that the female is ready to spawn, and in response, males allow the females to enter their territory without attacking them.

Jamming Avoidance Responses (JAR)

Several species of electric fish which electrolocate objects in their environment have evolved behavioral

strategies for avoidance of the jamming effects of EODs from other fish. These complex reflexlike responses were studied extensively by WH Heiligenberg and others, especially in the wave species *Eigenmannia*, which uses its 500 Hz discharge for electrolocation of objects. *Eigenmannia* is jammed when a second individual's EOD is in the range of 4–10 Hz above or below its discharge frequency, and it is unable to sense objects. It shifts its frequency upward or downward by as much as 10–20 Hz to prevent frequency overlap with the external stimulus. Electrolocation performance is then restored, and the jamming effects avoided by this behavior. Later, on removal of the stimulus, the jammed fish relaxes its EOD back to its original resting frequency. In some species, which move in shoals and are routinely jammed by other fish, prolonged jamming avoidance may produce a long-term elevation in frequency that persists for many hours or days.

Species with different tonic discharges have differing avoidance strategies. Wave fish avoid overlapping discharge frequency by moving upward or downward in frequency; pulse fish avoid temporal overlap of EODs by delaying or advancing their pulse times with transient increases in pacemaker firing rate or by making rapid, short-latency echo responses to jamming pulses.

Jamming avoidance responses resemble some of the phasic signals used in social communication, and they are generated by the same neural circuits in the brain. The 'long rises' of *Eigenmannia* and the frequency decreases in *Gymnarchus* are both used by subordinate fishes as signals of submission, and the transient rate increases seen in jamming avoidance responses among the pulse gymnotiforms also resemble the burst discharges and rate increases seen in agonistic behavior. The close resemblance of the jamming avoidance responses to social communication displays may not be accidental but may be yet another example of ritualized behavior – adopting physiological reflexes as displays for communication. This may help explain the remarkable convergence in signal form and function in unrelated groups of electric fish.

Brain Mechanisms of Signal Generation

Gymnotiform and mormyriform fishes generate patterns of discharges using a hierarchy of control nuclei in the medulla, the midbrain, and the thalamus. There are many parallels between the electromotor control areas in the brains of these two groups of fishes. Each time the electric organ fires, it is triggered by the firing of neurons in the pacemaker or command nucleus in a ventral medulla. In gymnotiform fishes, this ventral nucleus is called a pacemaker nucleus since it is spontaneously active and generates a regular repetitive discharge frequency. In mormyrids, the equivalent nucleus is called a command nucleus because these cells make the decision to fire each EOD pulse after integrating synaptic inputs from a variety of higher centers in the brain, but they do not show pacemaker-like regularity. In both groups, the pacemaker or command nucleus is located on the ventral midline and is composed of cells which are electrotonically coupled, ensuring synchrony of firing. Command or pacemaker neurons project to cells in a nearby relay nucleus containing very large neurons that project down the spinal cord to the electromotor nucleus near the electric organ.

The primary inputs to the command and pacemaker nuclei come from a column of cells located in the dorsal thalamus. For gymnotiforms, the central posterior nucleus projects to the pacemaker and causes smooth rises in EOD firing frequency. In mormyrids, the dorsal posterior nucleus produces smooth discharge accelerations or bursts. These two thalamic nuclei are heterogeneous along their lengths such that the firing of cells in the rostral regions produces smooth accelerations in discharge rate while the firing of cells in the caudal regions produces sharp transient bursts of pulses.

A secondary input comes from smaller groups of cells in the midbrain. In gymnotiforms, the sublemniscal prepacemaker nucleus (sPPn) causes frequency decreases or discharge interruptions; in mormyrids, the ventral part of the ventral posterior nucleus within the torus semicircularis projects to the command nucleus, but its effects are unknown.

Both these nuclei receive inputs from sensory and motor nuclei which get inputs from a variety of sensory modalities, including electric, auditory, lateral line, and visual pathways. These inputs are important means by which other sensory systems affect the discharge of the electric organ. In gymnotiforms, the nucleus *electrosensorius* sends inputs to both the sPPn and to the PPn and the central posterior nucleus; in mormyrids the *tectum mesencephali* sends fibers to the dorsal part of the ventral posterior nucleus in the torus and to the dorsal posterior nucleus in the thalamus.

Physiological work spanning four decades has resulted in good understanding of the electromotor control system of electric fishes. Building on earlier work by Bennett, Szabo, Fessard, and others, which described command and relay nuclei and electromotor nuclei in spinal cord, more recent work by M Kawasaki, C Keller, WF Heiligenberg, G Zupanc, BA Carlson, K Grant, G von der Emde and others has uncovered many of the higher electromotor control centers in the brain that generate many of the social communication signals discussed here.

Electric communication is a fascinating example of a novel modality of communication that has undergone independent adaptive radiations in two separate lineages of fishes. Both groups exploit electroreception to the fullest, using tuberous electroreceptors for active electrolocation and reception of communication signals. Both have diverse electric organs and diverse signal types. Together, their systems of electric communication provide insight into the evolution of signals; strategies for coping with noise, predation, and jamming; and energetic costs of signaling while maintaining signal diversity. Work on electric fish has led to the development of several model systems for the comparative study of nervous systems and behavior.

See also: Electrocommunication; Electrolocation; Gap Junctions and Electrical Synapses; Signal Transmission in Natural Environments.

Further Reading

Bennett MVL (1971) Electric organs. In: Hoar W and Randall DJ (eds.) *Fish Physiology*, pp. 347–491. New York: Academic Press.

Bennett MVL (1971) Electroreception. In: Hoar WS and Randall DJ (eds.) *Fish Physiology*, pp. 493–574. New York: Academic Press.

Bullock TH, Hagiwara S, Kusano K, and Negishi K (1961) Evidence for a category of electroreceptors in the lateral line of gymnotid fishes. *Science* 134: 1426–1427.

Bullock TH and Heiligenberg W (eds.) (1986) *Electroreception*. New York: Wiley.

Bullock TH, Hopkins CD, Popper AN, and Fay RR (eds.) (2005) *Electroreception*. New York: Springer.

Fessard A (1958) Les organes électrique. In: Grassé PP (ed.) *Traité de Zoologie*, pp. 1143–1238. Paris: Masson.

Fessard A (1974) *Electroreceptors and Other Specialized Receptors in Lower Vertebrates*. New York: Springer.

Fessard A and Szabo T (1961) Mise en évidence d'un récepteur sensible à l'électricité dans la peau d'un mormyre. *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 253: 1859–1860.

Heiligenberg W (1977) Principles of electrolocation and jamming avoidance in electric fish: A neuroethological approach. In: Braitenberg V (ed.) *Studies in Brain Function*, pp. 1–85. New York: Springer.

Heiligenberg W (1991) *Neural Nets in Electric Fish*. Cambridge, MA: MIT Press.

Hopkins CD (1972) Sex differences in electric signaling in an electric fish. *Science* 176: 1035–1037.

Hopkins CD (1973) Electric communication in fish. *American Scientist* 62: 426–437.

Hopkins CD (1973) Lightning as background noise for communication among electric fish. *Nature* 242: 268–270.

Hopkins CD (1974) Electric communication in fish. *American Scientist* 62: 426–437.

Hopkins CD (1986) Behavior of Mormyridae. In: Bullock TH and Heiligenberg TH (eds.) *Electroreception*, pp. 527–576. New York: Wiley.

Hopkins CD (1999) Design features for electric communication. *Journal of Experimental Biology* 202(10): 1217–1228.

Kalmijn AJ (1988) Detection of weak electric fields. In: Atema J, Fay RR, Popper AN, and Tavolga WN (eds.) *Sensory Biology of Aquatic Animals*, pp. 151–186. New York: Springer.

Kalmijn AJ (1988) Hydrodynamic and acoustic field detection. In: Atema J, Fay RR, Popper AN, and Tavolga WN (eds.) *Sensory Biology of Aquatic Animals*, pp. 83–130. New York: Springer.

Kawasaki M (1993) Comparative studies on the motor control mechanisms for electric communication in gymnotiform fishes. *Journal of Comparative Physiology A* 173: 726–728.

Kramer B (1990) *Electrocommunication in Teleost Fishes: Behavior and Experiments*. Berlin: Springer.

Lavoué S, Bigorne R, Lecointre G, and Agnese JF (2000) Phylogenetic relationships of electric fishes (Mormyridae: Teleostei) inferred from cytochrome b sequences. *Molecular Phylogenetics and Evolution* 14: 1–10.

Moller P (1995) *Electric Fishes: History and behavior*. London: Chapman & Hall.

Sullivan JP, Lavoué S, and Hopkins CD (2000) Molecular systematics of the African electric fishes (Mormyroidea: Teleostei) and a model for the evolution of their electric organs. *Journal of Experimental Biology* 203(4): 665–683.

Szabo T (1965) Sense organs of the lateral line system in some electric fish of the Gymnotidae, Gymnarchidae, and Mormyridae. *Journal of Morphology* 117: 229–250.

Relevant Websites

http://nelson.beckman.uiuc.edu – Bibliographies of electric fish.

http://alumnus.caltech.edu, http://www.fiu.edu – Electric fish electric organ discharge movies.

http://nelson.beckman.uiuc.edu, http://nelson.beckman.uiuc.edu – Electrolocation.

http://www.nbb.cornell.edu – Field studies on weakly electric fish.

http://www.fishbase.org – Fishbase, a general catalog of fish biodiversity.

http://www.npa.uiuc.edu – Jamming Avoidance Response.

http://tolweb.org – Tree of Life web project, Gymnotiformes.

http://tolweb.org – Tree of Life web project, Osteoglossiformes.

Electrical Self-Stimulation

R A Wise, National Institute on Drug Abuse, Baltimore, MD, USA

Published by Elsevier Ltd.

Introduction

James Olds and Peter Milner discovered in the early 1950s that rats will perform arbitrary responses that result in direct electrical stimulation of the septal area, the lateral hypothalamus, and related regions of the brain. At about the same time, Russel Monroe and Robert Heath discovered that humans found stimulation in this region pleasant. Subsequent studies have revealed that many vertebrate species will learn to work for what has come to be called brain stimulation reward (BSR). While stimulation at some sites is only weakly rewarding or can cause aversive or motoric side effects, stimulation in the lateral hypothalamic and more caudal regions of the medial forebrain bundle can be very powerfully rewarding. Rats will lever-press several thousand times per hour for half-second trains of lateral hypothalamic stimulation. Rats given unlimited access to such stimulation have been known to lever-press for 2 or 3 days without stopping; one rat averaged almost 30 responses per minute over a continuous 20-day period, despite occasional 2- to 3-h naps and two long periods of equipment failure. Experienced rats accept painful levels of footshock to gain access to such stimulation, levels they would not accept to get to food or to retrieve their pups. If forced to choose between food and lateral hypothalamic stimulation during the single hour a day when each is available, the animals 'self-starve,' choosing stimulation over food to the point of death. Olds termed the lateral hypothalamus a 'pleasure center' in the brain, and argued at the time that it was the site of convergence of reward-related fibers from a variety of other brain sites.

Multiple Reward Sites

There are perhaps two dozen brain regions where stimulation is rewarding. However, the fact that stimulation is rewarding in many brain regions does not mean that there are many independent systems supporting BSR. The boundaries of the various regions are not well defined, stimulation is not equally effective in each of the different regions, and stimulation activates fibers passing through the regions more effectively than it activates the cell groups for which the regions are usually named.

Positive sites for BSR include the olfactory bulb, a sensory structure, the region of the motor branch of the trigeminal nerve, a motor pathway, and many structures ambiguously labeled 'associational.' Medial forebrain bundle reward sites include the anterior, lateral, and posterior hypothalamus, the ventral tegmental area, and ventral portions of the bundle that extend caudal to the ventral tegmental area on the midline of the pons. Cortical reward sites include the medial prefrontal, sulcal, cingulate, and entorhinal cortex. Limbic sites include hippocampus, amygdala, and medial and lateral septal regions. Stimulation of the deep cerebellar nuclei and the region around the decussation of the brachium conjunctivum is rewarding. Stimulation near several monoamine nuclei is rewarding: these include the regions of the noradrenergic locus coeruleus and nucleus of the solitary tract, the serotonergic dorsal and medial raphe nuclei, and the dopaminergic ventral tegmental area and zona compacta and pars lateralus of the substantia nigra.

An early theory of BSR was that noradrenergic fibers from nucleus locus coeruleus are the substrate of reward in most or perhaps all of these structures. Locus coeruleus fibers ascend to the diencephalon and telencephalon, descend to the brain stem and spinal cord, and invade the cerebellum. While subsequent studies have ruled out the noradrenergic fibers of the locus coeruleus and the bed nucleus of the lateral hypothalamus as primary substrates of reward function, Olds' notion that reward fibers from many structures converge on a final common path involving (passing through) the lateral hypothalamus remains widely accepted.

Role of Monoamines

Early evidence for involvement of monoamine systems in BSR came from pharmacological studies: chlorpromazine and reserpine attenuate intracranial self-stimulation (ICSS) while amphetamine enhances it. Following the discovery that dopamine is a transmitter in its own right, not just a precursor of noradrenaline, it became clear that it was the ability of these agents to alter dopamine rather than noradrenaline function that explained their effects on self-stimulation. Mapping studies revealed that the boundaries of the reward system did not correspond precisely with the boundaries of the noradrenergic system but did correspond precisely to the boundaries of the dopamine system (including the boundaries of the pars compacta and pars lateralis of the substantia nigra). Selective pharmacological blockade of dopamine function

confirmed its importance for the rewarding effects of lateral hypothalamic self-stimulation, and the fact that the meso-corticolimbic and nigrostriatal dopamine fibers pass through the lateral hypothalamus (as part of the medial forebrain bundle) led to the view, now largely discredited, that it was direct activation of dopamine fibers that made lateral hypothalamic stimulation rewarding.

These conclusions were reached slowly, largely because pharmacological studies of BSR were hampered by the possibility that the drugs disrupted the ability of animals to perform the required response, rather than the rewarding effects of the stimulation *per se*. Of particular concern were the possibilities that noradrenergic antagonists might be altering arousal or attention, and that dopaminergic antagonists might be disrupting motor capacity. Two approaches confirmed that blockade of dopaminergic but not noradrenergic neurotransmission attenuates the rewarding effects of medial forebrain bundle stimulation. First, the gradual attenuation of responding in animals pretreated with selective dopamine antagonists parallels that of animals tested with reduced stimulation current or frequency. The animals initiate responding at normal rates, but decrease their responding progressively as they gain access to the reduced stimulation or to the normal stimulation under the drug condition. After responding has ceased, normal performance can be reinstated, temporarily, by presentation of conditioned stimuli associated with normal reward. Thus the animals remain capable of responding at times when the stimulation fails to maintain that responding. In contrast, responding in animals pretreated with selective noradrenergic antagonists is impaired even before the animals sample the stimulation and does not deteriorate progressively as the animal experiences earned stimulation in the drugged condition.

The second approach to separating the effects of drugs on reward function from their effects on performance capacity is known as the 'curve-shift' approach. In this method, rate of responding is determined as a function of the strength of the rewarding stimulation. Reward strength can be controlled by varying the frequency, intensity, train length, or pulse width of the stimulation. For pharmacological studies, stimulation frequency is varied because stimulation pulses given at different frequencies always reach the same subpopulation of fibers near the electrode tip. Like the dose-response function in pharmacology, the function relating response rate to stimulation frequency (the 'rate-frequency' function) is an ogive (Figure 1): chance performance is seen with a range of low frequencies, asymptotic performance is seen with a range of high frequencies, and

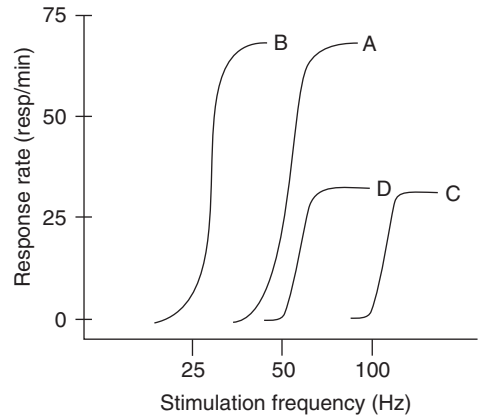

Figure 1 Lever-pressing rate as a function of stimulation frequency for BSR under various drug conditions. (A) Baseline condition. (B) Morphine injection into the ventral tegmental area reduces the amount of stimulation necessary for normal performance. (C) A high dose of a dopamine antagonist causes two independent effects; it decreases the performance capacity of the animal, reducing the maximum rate of responding (lower doses need not have this effect) and it increases the amount of stimulation necessary for normal performance. (D) When the treatments are combined, morphine cancels the effect of the dopamine antagonist on the rewarding effect of stimulation but does not restore the normal performance ceiling.

increasing performance is seen over a range of intermediate frequencies. Reducing the reward by reducing the stimulation intensity shifts the rate-frequency function to the right; increasing the reward by increasing the intensity shifts the function to the left. Making the task harder or easier – for example, by adding or subtracting weight from the response lever – shifts the asymptote down or up, respectively. Dopamine antagonists at low doses shift the rate-intensity function to the right (high doses shift it to the right and down), while indirect dopamine agonists like amphetamine and cocaine shift the function to the left. The effects of a dose of amphetamine that halves the stimulation frequency needed to maintain normal responding are exactly offset by a dose of a dopamine antagonist that doubles the required frequency. These findings not only establish that dopamine antagonists reduce the effectiveness of BSR, but also establish a metric for comparing the relative effectiveness of various drug treatments on reward function and performance capacity.

Fast Descending Reward-Relevant Fibers in the Medial Forebrain Bundle

While pharmacological studies establish the importance of the dopamine system for the rewarding effects of medial forebrain bundle stimulation, electrophysiological studies challenge the assumption that the relevant dopamine fibers are directly

depolarized at the electrode tip. While high levels of stimulation are required to activate dopamine fibers, low levels of stimulation are rewarding in the medial forebrain. A series of sophisticated parametric studies suggest that the bulk of the 'first-stage' fibers (the fibers that are depolarized at the electrode tip) are fast myelinated fibers that descend the medial forebrain bundle to activate the dopamine system trans-synaptically. In contrast, the dopamine fibers are slow, unmyelinated, and ascend the bundle.

The first set of parametric studies that identify fast descending fibers as the primary component of the first stage of medial forebrain bundle BSR involved studies in which the distribution of refractory periods for the populations of directly activated populations are estimated. For this determination, the intervals between stimulation pulses are manipulated to form trains of pulse 'pairs,' where the interval between the first ('C') and second ('T') pulses of each pairs is systematically varied. When the intervals between the C and T pulses are long, each pulse triggers an action potential and the animal responds frequently for the stimulation. When the intervals are shorter than the refractory period of the first-stage neurons, however, each C pulse triggers an action potential but each T pulse fails to do so. In this case, the animal receives half the reward and performance is much weaker. The distribution of refractory periods for the first-stage medial forebrain bundle fibers suggests two subpopulations, one with refractory periods ranging between 0.4 and 0.5 ms and one with refractory periods ranging from 0.6 to about 2 ms. These refractory period estimates are consistent with large myelinated first-stage fibers.

The second set of parametric studies also utilized paired-pulse stimulation; here the C pulses were delivered at the tip of one electrode, and the T pulses were delivered at the tip of a second electrode aligned a distance down the same axon trajectory. The goal is to stimulate the same first-stage fibers at two sites along their length to add estimates of conduction velocity to estimates of refractory period. When two electrodes are properly aligned along the same fiber path, paired pulse stimulation again proves normally effective, so long as the interval between C and T pulses is sufficiently long. Here the necessary interval between the two pulses of each pair must be long enough not only to allow recovery from the C pulse; it must also be long enough to allow the action potential triggered by the C pulse to pass the site where the T pulse will be delivered. With short C–T intervals, the action potential triggered by the C pulse will collide with an action potential triggered at the second electrode by the T pulse, and only one action potential will survive to reach the synapse (while this is not intuitively obvious, the collision test works regardless of which electrode is closer to the synapse). Since the refractory period can be estimated by paired pulse stimulation at either of the two electrodes, the conduction time between electrodes, and thus the conduction velocity, can be determined. The first-stage medial forebrain bundle conduction velocities are in the range of 1–11 m s^{-1}. The faster fibers implicated by such studies are again consistent with large myelinated first-stage fibers, not small unmyelinated dopamine fibers.

The two-electrode experiment lends itself to a third determination, that of directionality. If cathodal pulses are given at one electrode and anodal pulses are given at the other, stimulation-induced conduction will fail at the anodal electrode and reward signals will only propagate to the synapse when the anode is between the cell body and the cathode. Such studies confirm that at least the major portion of the first-stage reward fibers in the medial forebrain bundle conduct reward signals in the rostral–caudal direction, whereas the dopamine fibers project in the caudal–rostral direction.

Taken together, the parametric studies suggest that the bulk of the first-stage reward-relevant medial forebrain bundle fibers are not monaminergic, as was suggested by early pharmacological studies. Dopaminergic fibers are the wrong size and conduct in the wrong direction to be first-stage fibers. Nonetheless, the pharmacological studies implicate dopamine fibers as a final common path for medial forebrain bundle reward, and the current view is that the dopamine fibers are activated trans-synaptically by the descending myelinated fibers of the first stage.

Second-Stage and Third-Stage Candidates

Two hypotheses have been advanced for how the descending reward fibers might activate the ascending dopamine fibers of the medial forebrain bundle. The first hypothesis rests on evidence that the dorsal–ventral and medial–lateral boundaries of the descending fiber system correspond to the dorsal–ventral and medial–lateral boundaries of the dopamine cell layer of the ventral tegmental area and substantia nigra. Thus the descending fibers or collaterals from them may synapse on the dopamine cells directly, making the dopamine cells the second-stage neurons of medial forebrain bundle reward.

The second hypothesis is that descending medial forebrain bundle fibers synapse on cholinergic pedunculopontine tegmental neurons known to project back to the ventral tegmental area and substantia nigra, which would then serve as the third

stage. Rewarding lateral hypothalamic stimulation causes acetylcholine release in the ventral tegmental area, and muscarinic antagonists in this region cause an approximate doubling of reward thresholds. Muscarinic blockade of cholinergic autoreceptors at the level of the cholinergic cell bodies increases dopamine release and augments lateral hypothalamic BSR. Thus cholinergic neurons of the pedunculopontine and adjacent laterodorsal pontine tegmental nucleus are strong candidates for a link between first-stage reward fibers of the medial forebrain bundle and high-threshold third-stage dopamine reward fibers that project back past the stimulated region, but are too insensitive to respond to the stimulation directlyt.

Interaction with Addictive Drugs

Several addictive drugs potentiate the rewarding effects of lateral hypothalamic brain stimulation. This is reflected in leftward shifts of the rate-frequency function, as illustrated for morphine in Figure 1. Strong potentiation is seen with amphetamine, cocaine, and nicotine. While opiates potentiate the rewarding effects of the stimulation, particularly if given in the ventral tegmental area of the brain, opiates have sedative side-effects in rats that can impair responding when the drugs are given systemically. Cannabis and ethanol potentiate BSR in some strains and some individuals but not in others, and the potentiation is generally less robust with these drugs. The findings with barbiturates and benzodiazepines are mixed, and caffeine seems to inhibit BSR.

The brain sites at which addictive drugs act to enhance brain stimulation are often sites where the drugs are rewarding in their own right. Thus, amphetamine potentiates BSR, establishes conditioned place preferences, and is directly self-administered when injected into the nucleus accumbens. Morphine potentiates BSR, establishes conditioned place preferences, and is self-administered when injected into the ventral tegmental area. Both μ and δ opioids potentiate BSR and are self-administered when injected into the ventral tegmental area, and are effective as reward potentiators and rewards in their own right in proportion to their ability to cause nucleus accumbens dopamine release. Thus it is widely held that the pathways identified by BSR studies are pathways activated pharmacologically by habit-forming drugs.

Electrical Self-Stimulation in Perspective

BSR circuitry in general and the dopaminergic link in that circuitry in particular must be viewed in the context of millions of years of evolution. This circuitry cannot be specialized for a laboratory artifact. Natural rewards exert dramatic control over behavior patterns in vertebrates, and this control must be mediated by brain circuitry in some way or another. The fact that electrical stimulation of some regions of the brain is rewarding tells us that those regions are likely candidate substrates for the control of behavior by natural rewards like food, water, and sexual interaction. Indeed, food and sexual interaction, like rewarding brain stimulation and most addictive drugs, are known to elevate extracellular dopamine levels in nucleus accumbens and other brain regions. How this circuitry modifies behavior is not yet understood, but it is known that the lateral hypothalamic stimulation that animals will work for has the ability to enhance the consolidation of learning and memory, even when the stimulation is unearned and bears no relation to the learning and memory in question. Thus, when response-independent stimulation is given after a training trial, it enhances consolidation of whatever was learned in that training trial. Presumably the same mechanism is involved when the animal learns to self-stimulate its own brain.

See also: Basal Ganglia: Habit; Connectivity of Primate Reward Centers; Deep Brain Stimulation; Dopamine: Cellular Actions; Hypothalamic Structure–Function Relationships; Neuropsychology of Primate Reward Processes.

Further Reading

Bielajew C and Shizgal P (1986) Evidence implicating descending fibers in self-stimulation of the medial forebrain bundle. *Journal of Neuroscience* 24: 919–929.

Gallistel CR, Shizgal P, and Yeomans J (1986) A portrait of the substrate for self-stimulation. *Psychological Review* 88: 228–273.

Huston JP, Mueller CC, and Mondadori C (1986) Memory facilitation by posttrial hypothalamic stimulation and other reinforcers: A central theory of reinforcement. *Neuroscience and Biobehavioral Reviews* 1: 143–150.

Miliaressis E, Rompré P-P, Laviolette LP, et al. (1986) The curve-shift paradigm in self-stimulation. *Physiology and Behavior* 37: 85–91.

Monnroe RR and Heath RG (1986) Psychiatric observations. In: Heath RG (ed.) *Studies in Schizophrenia*, pp. 345–382. Cambridge: Harvard University Press.

Olds J and Milner PM (1986) Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology* 47: 419–427.

Olds ME and Olds J (1986) Approach-avoidance analysis of rat diencephalon. *Journal of Comparative Neurology* 120: 259–295.

Routtenberg A and Lindy J (1986) Effects of the availability of rewarding septal and hypothalamic stimulation on bar pressing

for food under conditions of deprivation. *Journal of Comparative and Physiological Psychology* 60: 158–161.

Wise RA (1986) Addictive drugs and brain stimulation reward. *Annual Review of Neuroscience* 19: 319–340.

Wise RA and Rompré P-P (1986) Brain dopamine and reward. *Annual Review of Psychology* 40: 191–225.

Yeomans J (1990) *Principles of Brain Stimulation*. Oxford: Oxford University Press.

Yeomans J, Forster G, and Blaha C (1986) M5 muscarinic receptors are needed for slow activation of dopamine neurons and for rewarding brain stimulation. *Life Sciences* 68: 2449–2456.

Electrocommunication

G K H Zupanc, Jacobs University Bremen, Bremen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Several taxonomic groups of fishes generate, by means of specialized organs, electric discharges. Based on the voltage of the signals produced, two groups of electric fish can be distinguished. Strongly electric fish generate powerful discharges of up to several hundred volts; in contrast, the electric signals produced by weakly electric fish rarely exceed a few volts. Whereas the powerful discharges of the strongly electric fish are used for prey hunting, the electric signals of the weakly electric fish play important roles in orientation and localization of objects in the closer vicinity of the fish (electrolocation) and in species recognition and intraspecific communication. This article focuses on the latter two functions.

Electric Organs and Electroreceptors

The discharges of the electric fish are produced by specialized electric organs, which, in the vast majority of species, are derived from various types of muscles and hence are termed myogenic (Figure 1). The only exceptions are members of the family Apteronotidae within the order Gymnotiformes, in which the electric organs are formed by modified spinal motor axons. The organs are, therefore, referred to as neurogenic. The cells making up the electric organs are commonly called electrocytes. In many species, they are stacked behind one another along the anterio-posterior body axis, thus forming a column of cells. Depending on the species, between a few columns and several hundreds of such columns, they are enclosed within tubes of connective tissue, running parallel to one another – thus resembling rolls of coins wrapped in paper, all facing the same way.

The synchronous activity of the electrocytes defines the output of the electric organ, called the electric organ discharge (EOD). The number of electrocytes arranged in series determines the voltage; the number of electrocytes arranged in parallel determines the current of the EOD. Typically, individual electrocytes generate potentials of the order of 100 mV.

The electrogenic system is complemented by the ability to sense electric signals. This capability is found in many fishes, both electric and nonelectric. It is mediated by specific sensory cells, called electroreceptors. Together with support cells, the electro receptor cells form encapsulated end organs known as electroreceptor organs (Figure 2). These receptor organs are located in invaginations of the fish's epidermis, but communicate with the body's surface through canals filled with mucous substrate or loosely packed epithelial cells. Typically, electroreceptor organs are distributed over most of the fish's body surface, with especially high densities in the head region.

Two anatomically distinct classes of electroreceptor organs can be distinguished: ampullary and tuberous. Ampullary receptors are maximally sensitive to direct current (DC) and low-frequency electric fields of both biological and nonbiological sources. In the context of electrocommunication, ampullary receptors have been shown to be involved in the detection of low-frequency components associated with modulatory signals used for communication among conspecifics. Tuberous receptors, on the other hand, are tuned to frequencies within or near the fundamental discharge frequency of the respective species, a feature that plays an important role in species recognition and male–female discrimination (see the sections titled 'Sexual dimorphisms in the electric organ discharge,' 'Species recognition,' and 'Electrocommunication').

Patterns of Electric Organ Discharges

In the two best-examined orders of electric fish, the Gymnotiformes from South and Central America, and the Mormyriformes from Africa, the EODs are produced either continuously, resulting in a wavelike appearance of the signals on the oscilloscope screen, or intermittently, so that the individual discharge pulses are separated by relatively long periods of silence. These two groups of electric fish are, therefore, referred to as wave-type and pulse-type fish, respectively.

As examples of wavelike discharges, the EODs of two gymnotiform species, the brown ghost knifefish (*Apteronotus leptorhynchus*) and the knifefish (*Eigenmannia* sp.), are shown in Figure 3. Both species belong to the group of weakly electric fish. Typically, the number of discharges produced within 1 s (defined as the EOD frequency) is highly regular in an individual fish, even over longer time periods of hours and days. On the other hand, the EOD frequencies may differ among different individuals within a certain range, thus defining the species-specific frequency band. Also quite characteristically, different species may occupy different frequency bands, often with no or little overlap. Like the discharge frequency, the waveform of the EOD is extremely constant in an individual fish, but may differ among different individuals and different species.

In pulse-type species, the individual discharge pulse is highly stereotyped in a given individual, but the

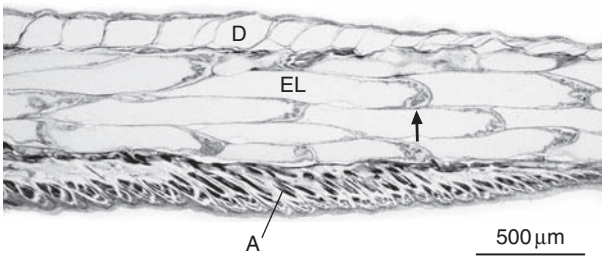

Figure 1 Electric organ of the knifefish, *Eigenmannia* sp. The parasagittal section through the tail shows, in addition to muscle fibers of the anal fin (A) and dermis (D) with scales, electrocytes (EL), which are distinguished by their long slender shape. These cells are arranged in series along the longitudinal axis of the body, thus constituting the electric organ. At the caudal end, where the membrane exhibits pronounced evaginations (arrow), the electrocytes are innervated by axons of spinal motor neurons. Courtesy of GKH Zupanc and F Kirschbaum.

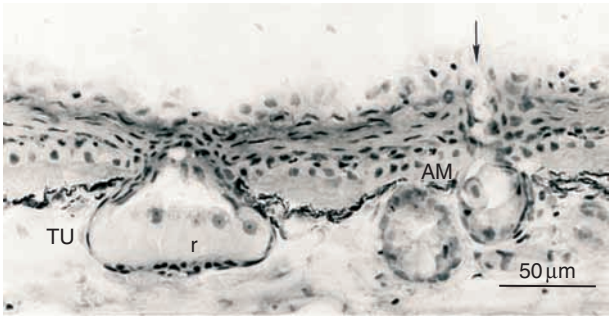

Figure 2 Electroreceptor organs in the skin of the knifefish, *Eigenmannia* sp. The photomicrograph shows two different types of these receptor organs: a tuberous (TU) electroreceptor organ with its individual receptor cells (one is indicated by r) and an ampullary electroreceptor organ (AM). The latter type is characterized by a long large-diameter canal (arrow). Courtesy of GKH Zupanc and F Kirschbaum.

intervals between the pulses are often quite variable, resulting in a nonperiodic pattern of discharge (Figures 4(a) and 4(b)). Thus, in addition to the average number of pulses produced per second (commonly referred to as the pulse repetition rate), the frequency distribution of the interpulse interval lengths is usually analyzed and plotted in the form of interpulse interval histograms to characterize this variability (Figure 4(c)). Such histograms have revealed that the mode of discharge pattern (for instance, a highly rhythmic discharge pattern vs. a more random discharge pattern) changes with the behavioral context and may be used by the fish to encode information important for communication.

Sexual Dimorphisms in the Electric Organ Discharge

Sexual dimorphism in the EOD has been found in a number of electric fish species. In *Sternopygus macrurus* (a wave-type gymnotiform species) sexually mature males discharge on average at 67 Hz, whereas sexually mature females discharge on average at 120 Hz. The mean EOD frequency of sexually immature fish is somewhat intermediate, at 93 Hz. The development of this sexual dimorphism is controlled by steroid hormones. Treatment of *Sternopygus* with the androgen testosterone or 5α-dihydrotestosterone leads to a pronounced decrease in EOD frequency, whereas the administration of estradiol causes a slight increase in EOD frequency. Hormonal control of the sexually dimorphic EOD also occurs in *Apteronotus leptorhynchus*. In this wave-type gymnotiform fish, mature males discharge at higher frequencies than sexually mature females. Implantation of estrogens causes a gradual decrease in EOD frequency, whereas treatment with 5α-dihydrotestosterone results in a slight increase (Figure 5).

Sexual dimorphism in the EOD may also involve the waveform of the electric discharge. Such a phenomenon has been found in both wave-type and pulse-type species. In the pulse-type mormyriform fish *Brienomyrus brachyistius* (long biphasic), the duration of the EOD pulse produced by mature males is, on average, 2.3 ms long, compared to 0.9 ms in mature females and immature specimens. The treatment of females with the androgen 17α-methyltestosterone or 5α-dihydrotestosterone causes a marked increase in EOD duration over a 2- to 3-week period.

In the pulse-type gymnotiform fish *Hypopomus occidentalis*, mature males are distinguished from females and immature males not only by morphological characteristics, such as their larger body size and the thicker tail, but also by differences in the waveform of the individual pulse (Figure 6). Both males and mature females produce biphasic EOD pulses, the first phase of which is head positive and the second phase head negative. However, whereas in mature females and immature males the durations of the first and second phases are equal, thus resulting in a symmetrical pulse, in mature males the EOD pulse is asymmetrical, which is caused by a longer duration of the second phase compared to the first phase. Treatment of females with 5α-dihydrotestosterone leads to the development of broader tails and asymmetric EOD pulses, caused by an increase in the size of individual electrocytes.

The hormonal effect on the EOD is paralleled by alterations in receptor tuning. In *Sternopygus*, in which the tuberous receptors are closely tuned to the EOD frequency of the individual fish, treatment with steroid hormones also leads to alterations in receptor tuning such that the best frequency of the receptors tracks the EOD frequency.

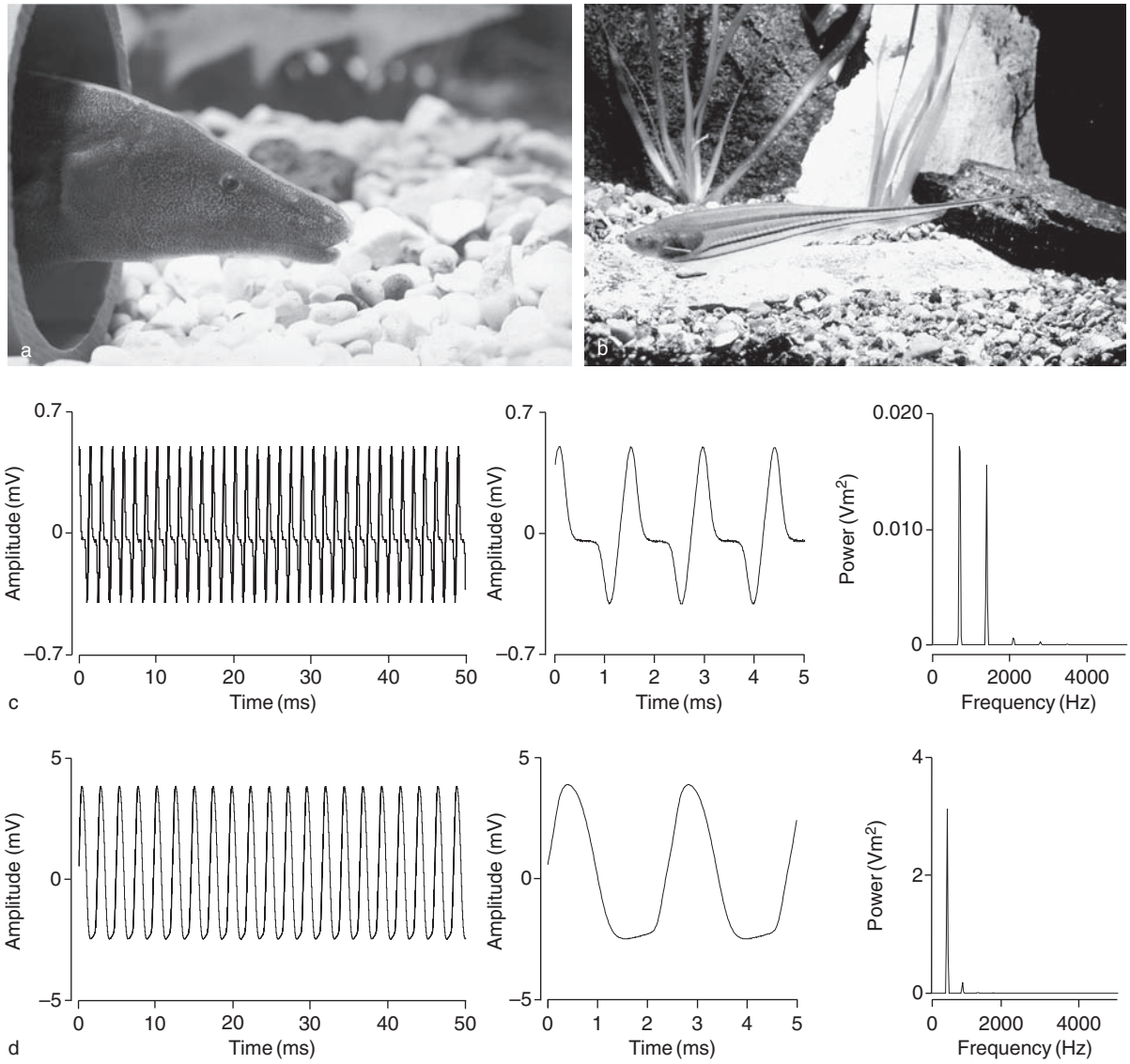

Figure 3 Two wave-type fish: (a) *Apteronotus leptorhynchus*; (b) *Eigenmannia* sp.; (c) electric organ discharges for *A. leptorhynchus*; (d) electric organ discharges for *Eigenmannia* sp. As is evident from the oscilloscope traces at compressed timescales ((c,d), left) and at expanded time scales ((c,d), middle) and from the Fourier spectra ((c,d), right), the discharges of these two species of gymnotiform fish differ in frequency, waveform, and spectral composition. Courtesy of GKH Zupanc and JR Banks.

Species Recognition

Field studies have revealed an enormous diversity of the EODs of sympatric species. This diversity involves the pattern of discharge (wave-type vs. pulse-type); the waveform and the frequency of the EOD in wave-type species, as well as the waveform and the peak spectral frequency of the pulse; and the pulse repetition rate in pulse-type species. These divergent features of the EOD may form the basis for the recognition by the weakly electric fish of their own species. In terms of perception, this ability is mediated by an increased sensory sensitivity to cues of the EOD of their own species while filtering out unwanted EODs. In *Hypopomus*, a pulse-type species, two out of five electroreceptor types respond maximally to spectral frequencies characteristic of the peak power of the species-specific EOD, thus acting as EOD filters. In wave-type fishes, tuberous receptors are commonly tuned to frequencies in the range of their own species.

The notion that weakly electric fish are capable of recognizing their own species based on invariant features of the EOD is corroborated by behavioral studies. Conditioning experiments showed that *Eigenmannia* sp. can discriminate electric signals differing in frequency or waveform. Moreover, *Sternopygus*

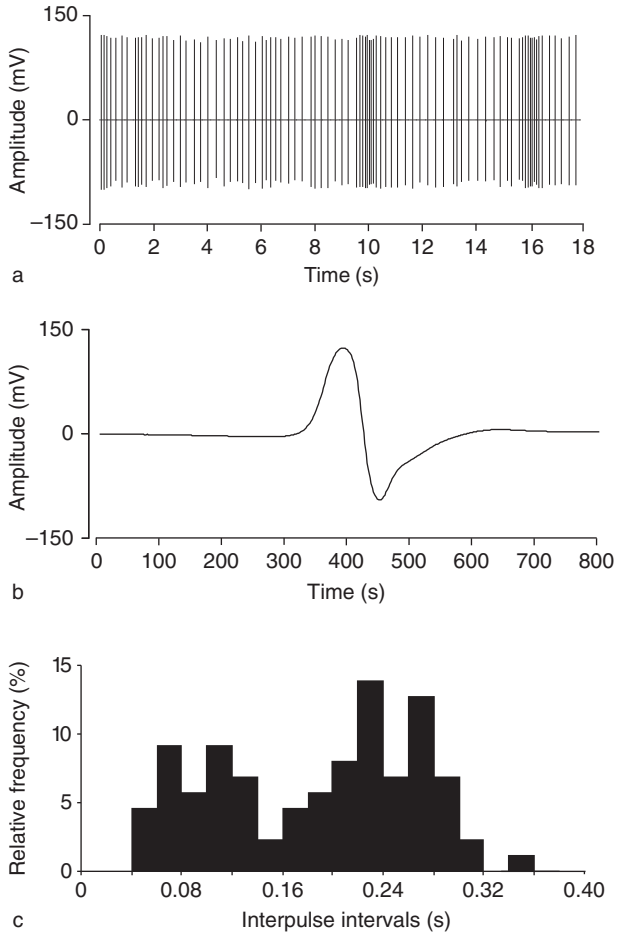

Figure 4 Electric organ discharge of a pulse-type weakly electric fish, the elephant nose (*Gnathonemus petersii*): (a) oscillogram at a compressed time scale; (b) expanded time scale; (c) frequency distribution histogram. In (a), which shows the individual discharge pulses as vertical lines, there is a rather irregular discharge rhythm. In (b), the details of the waveform of an individual pulse are resolved. The frequency distribution histogram (c) is based on analysis of the interpulse intervals of the discharge sequence shown in (a). This histogram suggests two major classes of interpulse intervals: short intervals between 0.04 and 0.16 s, and long intervals between 0.16 and 0.36 s. Courtesy of GKH Zupanc and JR Banks.

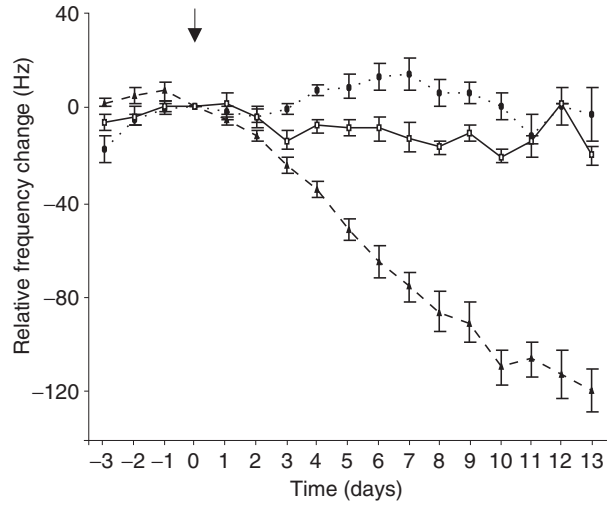

Figure 5 Control by steroid hormones of the sexual dimorphism in frequency of the electric organ discharge. Individuals of the wave-type fish *Apteronotus leptorhynchus* received implants of 5α-dihydrotestosterone ($n = 13$ fish; circles and dotted lines), received implants of 17β-estradiol ($n = 12$ fish; triangles and dashed lines), or were sham-treated (controls; $n = 12$ fish; squares and solid lines). In each of the three groups, the frequency of the electric organ discharge was measured on four consecutive days before the implantation (days 3–0) and on the 13 days following the surgery (days 1–13). The fish received the implants immediately following the measurement of the discharge frequency on day 0 (indicated by arrow). The frequency measured on each day was then normalized relative to the frequency determined on day 0 and averaged over the number of fish examined. Also indicated is the standard error of the mean. The diagram shows a pronounced decrease in the discharge frequency in the estradiol-treated fish and a slight frequency increase in the 5α-dihydrotestosterone-treated fish compared to the control fish. Courtesy of GKH Zupanc.

macrurus males produce electric courtship signals preferentially in response to electric sine waves in the frequency range of females of their own species, but less to sinusoidal signals outside this range.

Electrocommunication

Wave-Type Weakly Electric Fish

In wave-type fishes, both the normal EOD and transient modulations appear to play roles in communicatory behavior. As already mentioned, the normal EOD is extremely stable in a given fish, but may differ among individuals. Often, such differences reflect sexually dimorphic signals. In the brown ghost knifefish (*Apteronotus leptorhynchus*; Figure 3(a)), males discharge at a higher frequency than females. In the knifefish (*Eigenmannia* sp.; Figure 3(b)), the EOD frequency of females is higher than the frequency of the males. Moreover, in the latter species, females generate a quasi-sinusoidal signal, whereas the EOD waveform of males is more distorted, with the positive half-wave being markedly shorter than the negative half-wave. Conditioning experiments have demonstrated that *Eigenmannia* sp. is capable of distinguishing electric signals differing in frequency and/or waveform, including male and females EODs. These results are in agreement with the hypothesis that the fish identify the sex of a conspecific based on sexually dimorphic characteristics of the signal.

Transient modulations of the EOD are quite rarely produced spontaneously. In *Apteronotus leptorhynchus*, the most common form of such modulations are chirps (Figure 7). This name was given to this behavior to reflect the quality of the modulation, as

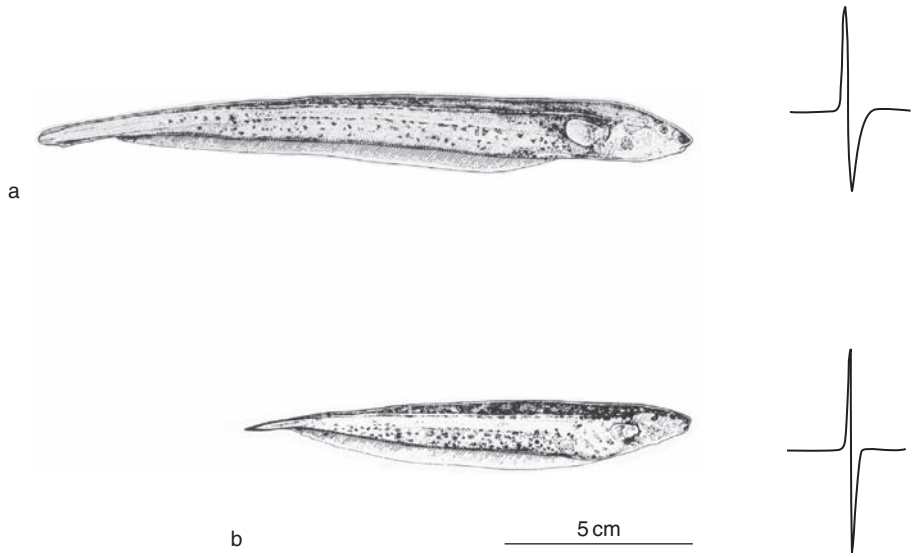

Figure 6 Sexual dimorphisms in morphology and electric organ discharge of the pulse-type fish, *Hypopomus occidentalis*: (a) males; (b) females. Males are larger and have a thickened tail ((a), left). Their discharge pulse is asymmetrical, which is caused by the longer duration of the second, head-negative phase, compared to the first, head-positive phase ((a), right). In contrast, females are smaller and have a slender tail ((b) left). Their discharge pulse is of shorter duration and exhibits a symmetrical form ((b), right). Adapted from Hagedorn M and Carr CE (1985) Single electrocyctes produce a sexually dimorphic signal in South American electric fish, *Hypopomus occidentalis* (Gymnotiformes, Hypopomidae). *Journal of Comparative Physiology A* 156: 511–523.

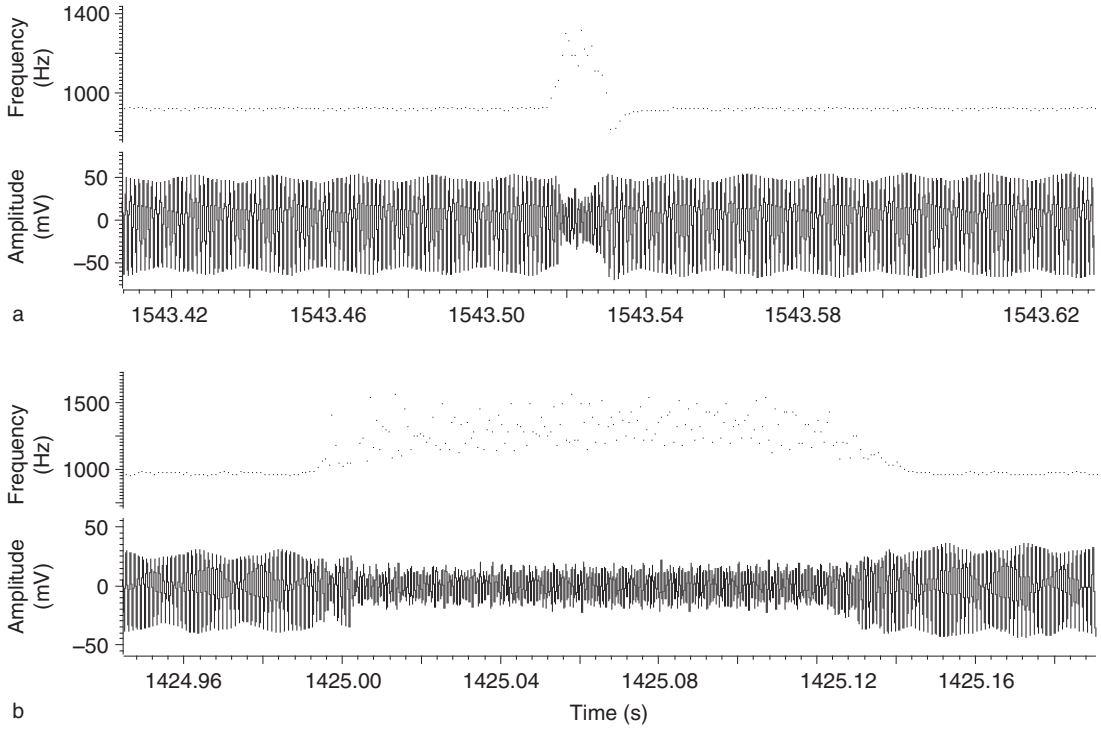

Figure 7 Chirplike modulations of the electric organ discharge in the brown ghost knifefish (*Apteronotus leptorhynchus*) during electric interaction with a conspecific: (a) type 1 chirp; (b) type 4 chirp. In these two graphs, a frequency–time plot is shown in the upper trace and the corresponding oscillogram is shown in the lower trace. Type 1 chirps are characterized by a rapid and pronounced increase in frequency, followed by a similarly rapid decrease, which finally results in an undershoot of the discharge frequency (a). By contrast, type 4 chirps are distinguished by their long duration of the frequency increase (b). In each of the two types of chirps, the amplitude is markedly reduced during the modulatory phase. Adapted from Zupanc GKH, Sîrbulescu RF, Nichols A, and Ilies I (2006) Electric interactions through chirping behavior in the weakly electric fish, *Apteronotus leptorhynchus*. *Journal of Comparative Physiology A* 192: 159–173.

perceived by the human observer when transforming the electric signal into an audible one. In isolated individuals of *Apteronotus leptorhynchus*, chirps occur at an average rate of just 1 chirp per 10 min. During social interactions, however, the chirp production may increase 1000-fold to rates of over 2 chirps s^{-1}. Chirps are predominantly generated by males and involve complex frequency and/or amplitude modulations. They last between some tens and a few hundred milliseconds. Thus far, six distinct types of chirps, differing in various physical parameters of the signal, have been described. It is thought that these different chirp types subserve different behavioral functions.

Closer analysis has shown that the chirps produced by two electrically interacting fish are not independent of one another. Instead, the chirps generated by one fish follow the chirps of the other individual with a preferred latency of roughly 500–1000 ms. This coordinated production of alternating chirps is called echo response. One function of this behavior may be to avoid coincident discharges, which could compromise the fish's electrolocation ability. An alternative explanation is that the echo response serves a communicatory function during encounters, such as aggression and courtship.

Pulse-Type Weakly Electric Fish

Studies have demonstrated that both the waveform of the individual pulse and the sequence of the interpulse intervals play important roles in intraspecific communication. The functional significance of the pulse waveform for sex recognition was first examined in *Brienomyrus brachyistius*. During the breeding season, males of this species occupy territories in shallow streams and produce courtship EODs consisting of a burst of EOD pulses at a pulse repetition rate of 100–150 Hz. This behavior is called rasp. In playback experiments, it can be used as an indicator of the male's response to the presumptive presence of a female because EODs recorded from females are more effective in eliciting rasps than discharges from males or other electric fish living in the same habitat. Experiments in which digitized EOD pulses were presented at various sequences of interpulse intervals demonstrated that males use the EOD waveform, rather than the sequence of interpulse intervals, to recognize a female signal. Because sex differences in EOD waveform have been found in many other species of weakly electric fish, presumably sex recognition based on EOD waveform is a widespread, and possibly even a universal, phenomenon.

Whereas the waveform of the individual EOD pulse is highly stereotyped and undergoes alterations only in the course of weeks and months (see the section titled 'Sexual dimorphisms in the electric organ discharge'), the sequence of interpulse intervals is highly variable and subject to modulations, particularly in the context of social interactions between two or more fish. This can be demonstrated by the following experiment. Two elephant nose fish (*Gnathonemus petersii*), a pulse-type mormyriform species, are kept isolated in two separate tanks (**Figure 8(a)**). Their EODs are recorded by electrodes placed close to each fish and, on the closure of an electric switch, transmitted into the other tank. Thus, only electric cues, but no others, are exchanged during stimulation. During the experiment, the discharges of the two fish are monitored through a second pair of electrodes installed in one of the two tanks. Through this single pair of recording electrodes, not only can the discharges of the fish in the recording tank be monitored but also those of the fish in the neighboring tank because the discharges of the latter are transmitted via the electric circuit used for stimulation. In the recordings, the EODs of the fish in the recording tank are distinguished by their larger amplitude, whereas the amplitude of the EODs of the fish in the neighboring tank are markedly smaller (**Figure 8(b)**).

The results of such an electric interaction experiment are shown in **Figures 8(b) and 8(c)** (an expanded time plot). Within a few seconds after transmission of the first fish's signals from the neighboring to the recording tank, the second fish, in the recording tank, started to modulate the discharge pattern. First, it reduced its pulse repetition rate to almost half the rate before stimulation. Second, this brief initial phase was followed by a complete cessation of the EOD for more than 1 min. Third, the return to EOD production was characterized by an initial generation of a burst of pulses, followed by a slowing down of the discharge rhythm. A similar discharge pattern is frequently observed during aggressive encounters. This suggests that the second fish interprets the signals of the neighboring fish as those of an intruder trying to enter its territory.

In general, a variety of modulation patterns that occur during social interactions have been identified. They include complete cessations of the EODs, transient acceleration of the discharge rate, long-term shifts in the baseline EOD repetition rate, changes in the regularity of the discharge pattern, and coupling between the two discharge pattern in that the EOD of one fish affects the EOD of another individual. An example of the last type of pattern is the echo response of pulse-type fish, which is characterized

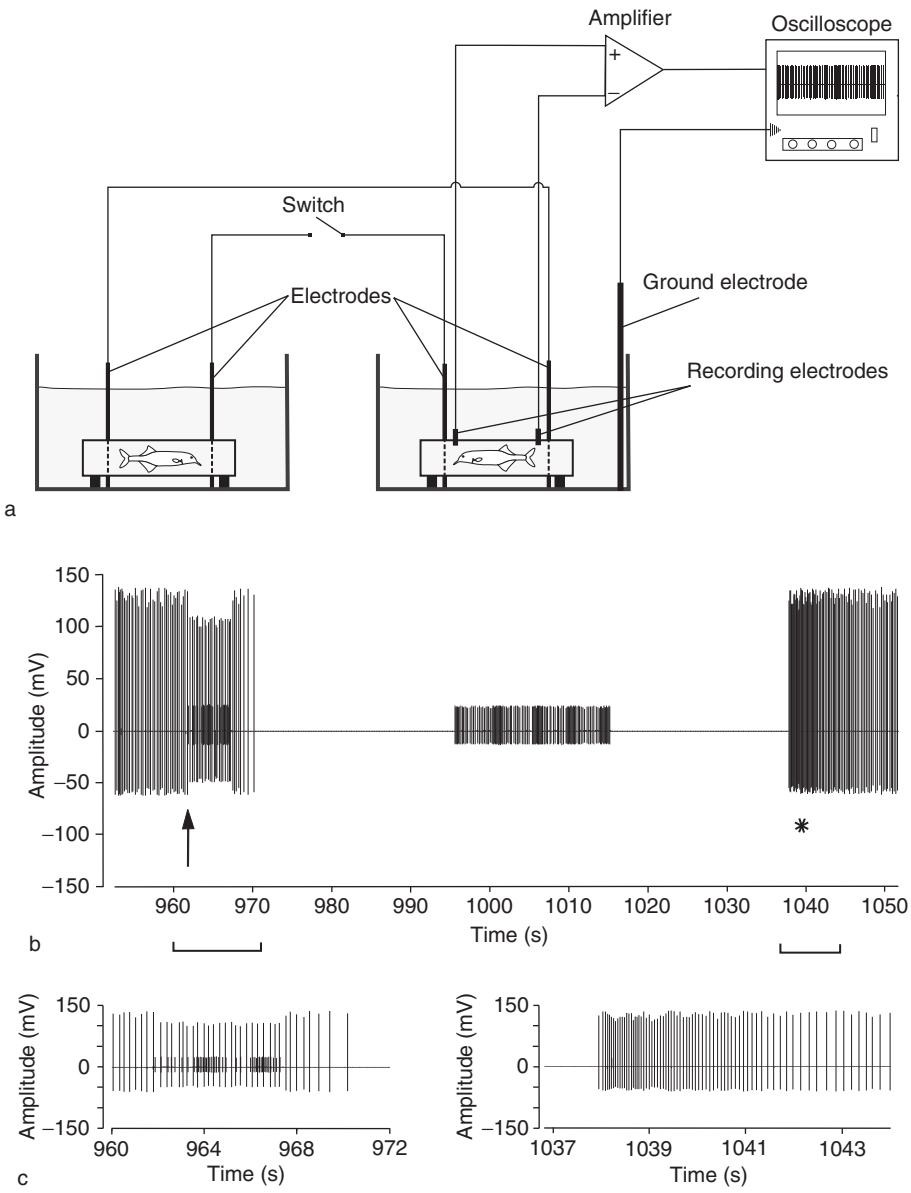

Figure 8 Electrical interactions in the elephant nose fish (*Gnathonemus petersii*): (a) experimental setup to demonstrate electrical interactions between two fish; (b) modulation of the discharge rhythm of *Gnathonemus petersii* by the electric organ discharges of a conspecific; (c) expanded timescales of bracketed areas in (b). As shown in (a), a single fish is kept in each of the two tanks. The two aquariums are electrically connected via wires attached to electrodes, which are placed near the fish. The connection can be controlled by an electrical switch. In addition, the discharges of one of the two fish are monitored through recording electrodes. The electric organ discharges of this fish are intermingled with the discharge of the second fish, which are visible as pulses of significantly lower amplitude. These discharges are transmitted into the tank of the first fish as long as the electric circuit connecting the two tanks is closed. As shown in (b), a few seconds after closing the electrical circuit (indicated by arrow), and thus stimulating the first fish with the discharges of the second, the first fish slows down its discharge rhythm and, another few seconds later, even completely ceases discharging. During this period of silence, the neighboring fish generates a salvo of discharge pulses, to which the first fish responds only with a delay of some 20 s (indicated by *). This response of the first fish is distinguished by the production of discharges at an extremely high pulse-repetition rate within the first few seconds. Courtesy of GKH Zupanc and D Meissner.

by a short (12–14 ms in *Gnathonemus petersii*) latency of pulse generation in response to the EOD pulse of a conspecific.

The production of specific EODs displayed in different situations of social interactions suggests a role in intraspecific communication. This hypothesis has been tested by the use of electric fish dummies. Isolated fish are stimulated with recorded EOD sequences or artificially generated EOD-like patterns via electric dipoles, and the behavioral response is

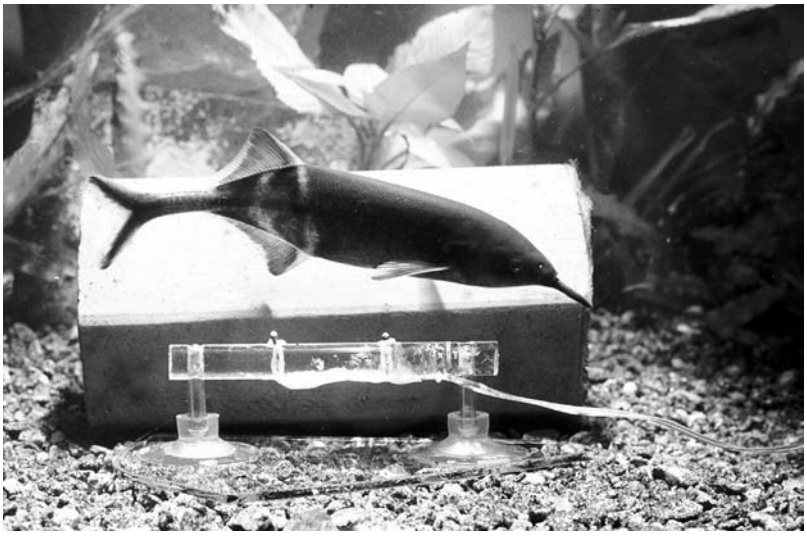

Figure 9 The behavioral significance of various parameters of the electric organ discharge can be explored by the use of electric fish dummies, consisting of electric dipole electrodes. Here, an elephant nose fish (*Gnathonemus petersii*) approaches such an electric fish dummy, via which recorded electric signals are played. Courtesy of GKH Zupanc.

analyzed (Figure 9). In *Gnathonemus petersii*, such fish dummy experiments have demonstrated clear differences in the response to signal trains recorded from resting and attacking fish. In contrast to the resting discharge, the attack discharge, characterized by its higher mean pulse repetition rate and the production of short sequences of EOD bursts, elicits a larger number of attacks on the dipole electrodes.

Neural Control of Chirping Behavior

Weakly electric fish have proven to be excellent model systems to study the neural basis of specific behavioral patterns, including those occurring during electrocommunication. One particularly well-studied example of such a behavior is the previously described chirping of the gymnotiform fish *Eigenmannia* sp. and *A. leptorhynchus*. In each of the two fish, the EOD is driven by command pulses originating from the pacemaker nucleus, an endogenous oscillator in the medulla oblongata. The major cell types forming the pacemaker nucleus are pacemaker cells and relay cells. The pacemaker cells, which remain intrinsic to the pacemaker nucleus, converge on to relay cells. Cells of the latter type, in turn, send axons down the spinal cord to command, through their spiking, the discharges of the electric organ in a one-to-one fashion. The neurons of the pacemaker nucleus fire both *in vivo* and *in vitro* at a very constant frequency, a phenomenon that parallels the regularity of the EOD.

Chirplike modulations of the EOD are mediated by input to the pacemaker nucleus arising from the prepacemaker nucleus (Figure 10). Because this diencephalic nucleus forms part of the central posterior nucleus, it is commonly referred to as the central posterior/prepacemaker nucleus. Stimulation experiments have indicated that the chirping behavior is controlled by a ventrolaterally located portion of the central posterior/prepacemaker nucleus. This subnucleus comprises large multipolar cells projecting to the pacemaker nucleus, where they synapse predominantly on the relay cells. Signal transmission at these synapses appears to involve a non-*N*-methyl-D-aspartate (NMDA) type of glutamate receptors. Iontophoresis of L-glutamate into the region, or electrical stimulation, of the ventrolaterally located large multipolar neurons of the central posterior/prepacemaker nucleus evokes EOD frequency modulations resembling natural chirps. This effect is thought to be mediated by a rapid depolarization of the relay cells in the pacemaker nucleus. Because pacemaker cells are intensively connected with one another and also with relay cells via mixed chemical and electrical synapses, this depolarization can spread within the pacemaker nucleus and, finally, lead to an acceleration of the spike frequency of both the pacemaker neurons and the relay cells.

Chirps can be elicited by sensory stimulation of the intact fish through an electric dipole with recorded discharges or sinusoidal electric signals in the frequency range of the fish's own EOD frequency. The perception of these EOD-like stimuli is mediated by the electroreceptors on the fish's body surface, which project via primary afferents to the somatotopically ordered electrosensory lateral line lobe in the

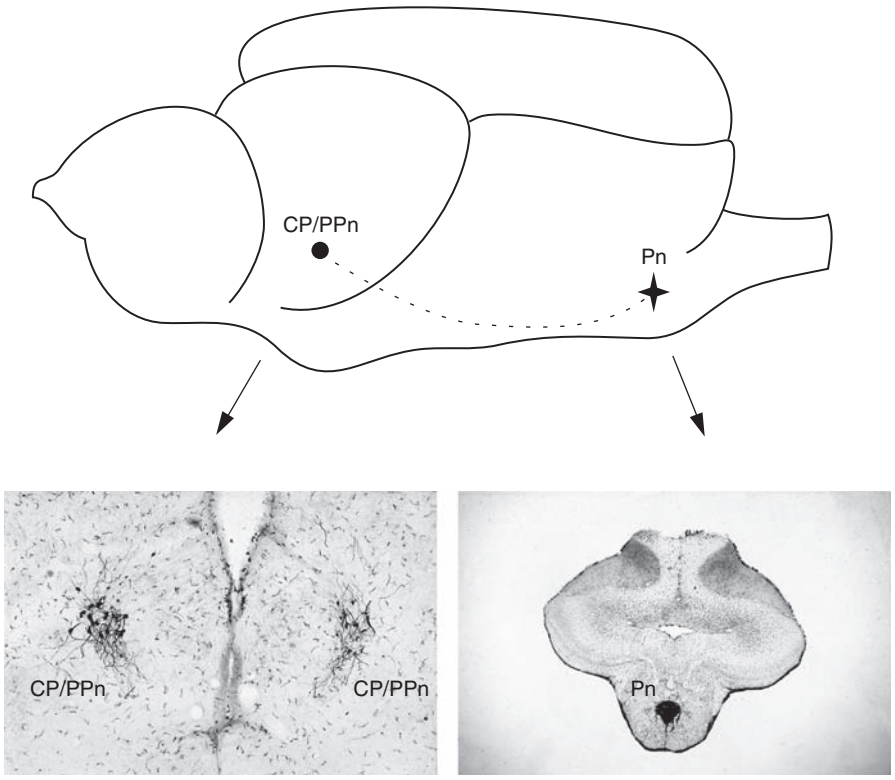

Figure 10 Neural circuitry involved in the neural control of chirp production in the weakly electric fish *Eigenmannia* sp. The electric organ discharge is driven by the activity of the pacemaker nucleus (Pn). This endogenous oscillator in the brain stem area receives input from neurons of the central posterior/prepacemaker nucleus (CP/PPn) in the diencephalon. The side view (top) shows the location of these two nuclei in a drawing. The connection between pacemaker nucleus and CP/PPn has been revealed by neuronal tract tracing. As shown in the two microphotographs, the application of the enzyme horseradish peroxidase to the pacemaker nucleus (bottom right) results in retrogradely labeled neurons in the CP/PPn (bottom left). Courtesy of GKH Zupanc.

hindbrain. This brain region, in turn, sends axons to the torus semicircularis in the midbrain. Axons from the latter innervate, among other targets, the nucleus electrosensorius in the diencephalon. This brain region projects to the central posterior/prepacemaker nucleus, which, therefore, can be regarded as an interface between the electrosensory system and the electromotor system.

See also: Electrical Perception and Communication; Electrolocation; Sexual Selection and the Evolution of Animal Signals; Signal Design Rules in Animal Communication.

Further Reading

Bass AH (1986) Electric organs revisited: Evolution of a vertebrate communication and orientation organ. In: Bullock TH and Heiligenberg W (eds.) *Electroreception*, pp. 13–70. New York: Wiley.

Bass AH and Hopkins CD (1983) Hormonal control of sexual differentiation: Changes in electric organ discharge waveform. *Science* 220: 971–974.

Bennett MVL (1971) Electric organs. In: Hoar WS and Randall DJ (eds.) *Fish Physiology, Vol. 5: Sensory Systems and Electric Organs*, pp. 347-491. New York: Academic Press.

Bullock TH and Heiligenberg W (eds.) (1986) *Electroreception*. New York: Wiley.

Bullock TH, Hopkins CD, Popper AN, and Fay RR (eds.) (2005) *Electroreception*. New York: Springer.

Hagedorn M and Carr CE (1985) Single electrocyctes produce a sexually dimorphic signal in South American electric fish, *Hypopomus occidentalis* (Gymnotiformes, Hypopomidae). *Journal of Comparative Physiology A* 156: 511–523.

Heiligenberg W (ed.) (1991) *Neural Nets in Electric Fish*. Cambridge, MA: MIT Press.

Hopkins CD (1972) Sex differences in electric signaling in an electric fish. *Science* 176: 1035–1037.

Hopkins CD (1976) Stimulus filtering and electroreception: Tuberous electroreceptors in three species of gymnotoid fish. *Journal of Comparative Physiology A* 111: 171–207.

Hopkins CD (1986) Behavior of Mormyridae. In: Bullock TH and Heiligenberg W (eds.) *Electroreception*, pp. 527–576. New York: Wiley.

Kramer B (1979) Electric and motor responses of the weakly electric fish, *Gnathonemus petersii* (Mormyridae), to play-back of social signals. *Behavioral Ecology and Sociobiology* 6: 67–79.

Metzner W (1999) Neural circuitry for communication and jamming avoidance in gymnotiform electric fish. *Journal of Experimental Biology* 202: 1365–1375.

Zakon HH (1986) The electroreceptive periphery. In: Bullock TH and Heiligenberg W (eds.) *Electroreception*, pp. 103–156. New York: Wiley.

Zupanc GKH (2002) From oscillators to modulators: Behavioral and neural control of modulations of the electric organ discharge in the gymnotiform fish, *Apteronotus leptorhynchus*. *Journal of Physiology* (Paris) 96: 459–472.

Zupanc GKH and Maler L (1997) Neuronal control of behavioral plasticity: The prepacemaker nucleus of weakly electric gymnotiform fish. *Journal of Comparative Physiology A* 180: 99–111.

Zupanc GKH, Sîrbulescu RF, Nichols A, and Ilies I (2006) Electric interactions through chirping behavior in the weakly electric fish, *Apteronotus leptorhynchus*. *Journal of Comparative Physiology A* 192: 159–173.

Electroencephalography (EEG)

F H Lopes da Silva, University of Amsterdam, Amsterdam, The Netherlands
S I Gonçalves and J C De Munck, VU Medical Center, Amsterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Definitions

The electroencephalogram (EEG) is traditionally defined as the electrical activity of the brain recorded from the human scalp. This is the common translation of *elektrenkephalogram*, the term coined by Hans Berger (1873–1941). However, this term may also be applied to the electrical activity of neuronal masses as recorded anywhere at the surface of the cortex (i.e., the electrocorticogram) or within the brain, both in humans and in animals.

Despite the important recent developments of other powerful techniques capable of imaging functional states of the brain (positron emission tomography, single photon emission computed tomography, and functional magnetic resonance imaging (fMRI)), the scalp EEG occupies an important position as a diagnostic and research tool because of four main characteristics: (1) it has an exceedingly high time resolution since it can monitor the activity of neuronal populations at the millisecond time scale, (2) it is noninvasive, (3) it is simple to record and it is inexpensive, and (4) it is the only technique by means of which the dynamics of brain activity can be monitored in the freely moving subject. Electroencephalography shares the first two of these characteristics, but not the last one, with the recently developed technique of magnetoencephalography (MEG).

Neuronal Sources of the EEG/MEG

The electrical activity of neuronal populations consists of ionic currents generated by cellular sources, flowing in a volume conductor formed by the brain and surrounding tissues. These ionic currents cause electric and magnetic fields that can be measured at a distance from the cellular sources, and thus they form the sources of the EEG (and MEG). In order to understand EEG phenomena, one has to take into consideration both the organization in space of the cellular sources and their time-dependent properties.

Synchronously active neurons arranged spatially in an ordered way can form macroscopic sources of potential fields measurable at a distance (i.e., EEG signals). A favorable spatial organization in this respect is the arrangement of the neurons in palisades, where they are aligned with the main axes of the dendritic trees oriented parallel to each other and perpendicular to the cortical surface. When such a neuronal population is activated in a more or less synchronous way by synapses organized in well-defined layers, the resulting extracellular currents flow mainly parallel to the main axes of the neurons. In this way, a well-organized distribution of sinks and sources is formed (Figure 1).

The characteristics of EEG signals depend on the amount of correlation between the active neurons. This correlation is determined by the synchrony of the incoming signals and by the synaptic organization of the local circuits. Typically, a population of neurons sharing the same type of inputs and strongly interconnected can form a macroscopic source of an EEG signal. An essential feature of such a population is that the neurons are interconnected by way of feedback loops, involving excitatory and inhibitory synapses. This fact, in conjunction with the intrinsic oscillatory properties of some types of neurons, is responsible for the occurrence of brain rhythms, that is, oscillatory phenomena of EEG.

EEG Rhythms: Phenomenology and Functional Significance

Since EEG signals consist of time series that can vary from low to high frequencies, it is the general consensus to divide the frequency range within which EEG signals can occur in a number of frequency bands that have been named after Greek letters: δ (delta) = below 3.5 Hz, θ (theta) = 4–7.5 Hz, α (alpha) = 8–13 Hz, β (beta) = above 13 Hz, and the term γ (gamma) is often used for frequencies above 30–35 Hz. There are different α rhythms depending on brain area and behavioral state. The posterior (occipital, parietal, and posterior temporal areas) α rhythm is the classical brain rhythm that occurs in a state of relaxed wakefulness, particularly evident when the eyes are closed. The α rhythm is strongly attenuated, or even blocked, at the opening of the eyes, and it is also attenuated as the state of alertness is enhanced, such as by sensory stimulation or mental activities. The term μ (mu) is used for a special rhythm occurring over the rolandic or central area within the α frequency range. Most conspicuous is the blocking of the μ rhythm with contralateral movement (Figure 2), or even with the mental intention to perform a movement. The latter property can be used to activate a brain–computer interface with the aim of controlling robotic devices in paralyzed patients. In addition to these two rhythms in the α range, a

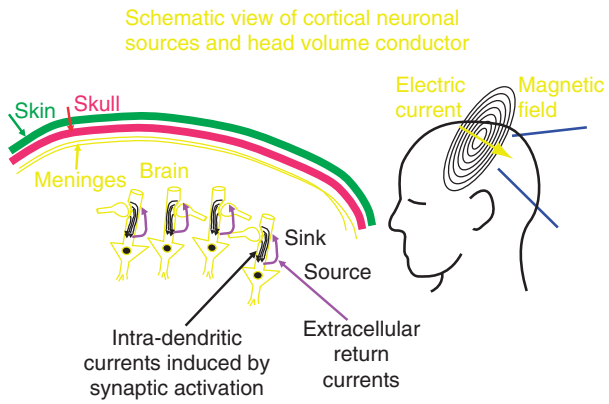

Figure 1 (Left) Schematic representation of a palisade of cortical pyramidal neurons and of the tissues enveloping the brain. These neurons receive synaptic activation at the dendrites. This results in the flow of an intracellular current along the dendrites toward the soma. These currents are compensated by volume currents in the extracellular medium, resulting in the formation of extracellular sources and sinks. (Right) Drawing of a head with the representation of the current dipole that is responsible for the EEG signal and the corresponding MEG field. The latter is oriented perpendicularly to the electric field.

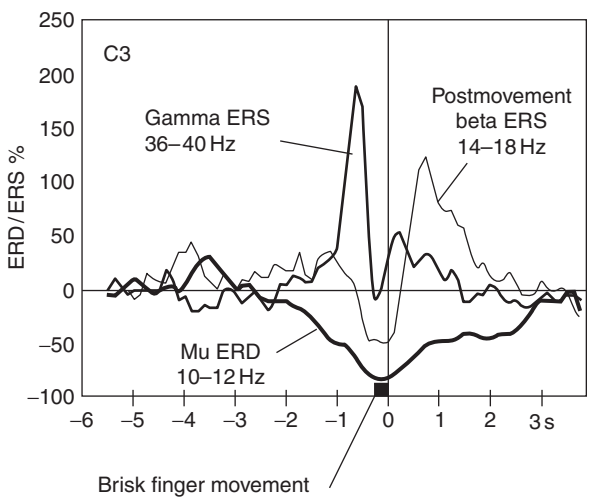

Figure 2 Dynamic changes in EEG signals recorded from the central scalp area in a normal subject correlated to a brisk finger movement in three frequency bands: mu (10–12 Hz), beta (14–18 Hz), and gamma (36–40 Hz). Vertical scale indicates the change in power within a given frequency band relative to the baseline recorded seconds before the event. ERD, decrease in power or event-related desynchronization; ERS, the opposite. Note the μ ERD starting approximately 2 s before the movement, the γ burst just before the movement, and the postmovement β rebound.

midtemporal α rhythm was described over the temporal lobe in MEG recordings, although it is difficult to detect in EEG recordings unless there are bone defects. This rhythm is attenuated by sound stimuli but not by opening of the eyes.

Some authors consider that there is a basically functional similarity between the rolandic μ and the posterior α rhythms since both would represent 'resting states' of the corresponding cortical areas: the sensorimotor cortex in the case of the former and the visual cortex in the case of the latter. This would correspond to a resting state during which the underlying neuronal networks are not actively processing specific stimuli; for this reason, this state was described in the classic literature as 'idling' by Adrian and Mathews in 1934. Nevertheless, the functional significance of α rhythms, which has been a matter of some heated arguments, cannot be simply described as corresponding to an 'idling state' of some brain systems. The functional significance of α rhythms has to be derived from basic neurophysiology. In this respect, it is known from detailed electrophysiological investigations and comprehensive models that oscillations within the α frequency range, in general terms, can be generated by two interacting populations; in the thalamus such populations consist of the thalamocortical relay neurons (TCR) and the reticular nucleus neurons (RN), which form a feedback loop. The TCR neurons form excitatory synapses on the RN, whereas these form inhibitory synapses on the TCR neurons. During an oscillatory bout of activity in the α frequency range, the RN neurons are in an active state, although this results in inhibition of TCR neurons. The latter display oscillations around a mean membrane potential relatively hyperpolarized, but this hyperpolarized state does not imply that these neuronal populations are idling. On the contrary, they can play an active inhibiting role in gating the transfer of information in certain neuronal pathways. In summary, the α rhythm state may fulfill, in itself, a relevant functional role, namely a 'gating function.' In this context, it is also interesting to consider what additional information about the state of excitability of the neuronal population underlying α power enhancement may be gained by combining EEG recordings with other methods. Experiments in which EEG recordings and transcranial magnetic stimulation (TMS) of the sensorimotor cortex were combined showed an association between the inhibition of a motor response and a significant increase of 11–13 Hz oscillatory activity. How this is reflected in the blood oxygen level-dependent (BOLD) fMRI signal is still not absolutely clear.

Discussion about the relationship between EEG/MEG rhythms and the level of activity of the underlying neuronal populations has to take into consideration that different brain rhythmic activities may have different kinds of relationships. In the case of the relation between rolandic beta activity (approximately 15–30 Hz) and levels of neuronal activity, a clear

temporal relationship with peripheral electromyography (EMG) activity during isometric contractions was found using MEG recordings. The MEG signals in the beta frequency range lead the EMG signals in time, whereas the time lag increases with increasing brain–muscle distance. This means that this beta rhythmic activity corresponds to an active state of the populations of the motor rolandic cortex in terms of efferent bursts of action potentials that constitute the central drive to spinal motor neurons. Nevertheless, there are also conditions in which beta rhythms do not represent an active state in the sense defined previously. This is the case for the postmovement beta event-related synchronization (ERS) that occurs after the completion of a finger (16–21 Hz) or foot (19–26 Hz) movement over the contralateral cortical area when the muscles relax (Figure 2).

The ERS in the beta band (around 20 Hz) found after a movement corresponds to a period of decreased corticospinal excitability determined by applying TMS. Thus, this beta rebound ERS represents a state of cortical disfacilitation that takes place at the cessation of the movement. Indeed, a generalized statement regarding the correspondence between beta EEG activity at the scalp and the activity state of the underlying cortical neuronal populations, without further specification of cortical area, frequency band, and behavioral conditions, is not adequate.

In the case of activities within the gamma frequency range (above 30 Hz), it was found that bursts of gamma activity may occur just before a brisk finger movement that is associated with motor cortex activation (Figure 2); this was further demonstrated using direct electrocorticography. The underlying neuronal mechanism responsible for these cortical gamma oscillations was further unraveled by finding that synchronized gamma frequency oscillations are enhanced with arousal and attention or by electrical stimulation of the mesencephalic reticular formation, and they are mediated by cholinergic muscarinic receptors. Interestingly, a close correlation between hemodynamic signals and the power of local field potentials in the gamma frequency range in the visual cortex of the cat was reported. Gamma ERS currently plays a prominent role in neurocognitive investigations.

In general, we may state that information processing in the brain occurs in a parallel fashion, where distinct brain areas are active either simultaneously or in short succession. The corresponding activities can be reflected in the local EEG signals. However, these networks are very often not anatomically contiguous; that is, they may occupy quite different locations within the cortex, such that the corresponding EEG signals form a complex spatiotemporal pattern distributed over the scalp. To determine the functional interactions between the distinct cortical areas involved in a given cognitive task, the spatiotemporal relationships of different EEG signals have to be determined. In this context, several analytic measures of signal similarity and synchronization are used. When a subject actively processes information, β or γ activity is enhanced in the EEG. It has been hypothesized that these β or γ rhythms may represent the activity of integrative processes responsible for binding the activities of distributed neuronal networks in such a way that integrated percepts may be constructed or coordinated motor activity may occur.

EEG Signals in Vigilance and Sleep Stages

With the onset of drowsiness (sleep stage 1), the EEG changes gradually: the classic α rhythm becomes less continuous (α dropout), the amplitude declines, and the dominant frequency decreases toward the θ range. With the deepening of drowsiness, there is an enhancement of slow-wave components, and trains of 2 or 3 Hz and 4–7 Hz tend to appear. Typical of deep drowsiness is the appearance of sharp transients that are particularly evident in recordings from the vertex (vertex waves and K-complexes). These transients represent impulse responses of the neuronal networks when the δ mode of activity begins to dominate. In sleep stage 2, sleep spindles, or sigma waves, with frequencies between 12.5 and 15.5 Hz, also appear. In sleep stage 3, δ frequencies in the 0.5- to 3-Hz range are dominant, particularly over the anterior areas, whereas sleep spindles become less conspicuous and K-complexes occur frequently. In very deep sleep (stage 4), the slow waves within the δ frequency range are even more abundant than in the preceding stage. A most conspicuous phenomenon in the evolution of sleep is the sudden change of the deep-sleep EEG to the pattern of rapid eye movement sleep, characterized by a dramatic decrease of the EEG amplitudes and by the overwhelming presence of high frequencies. The neuronal mechanisms underlying different features of sleep are becoming better known. The discovery of the cortically generated very slow oscillation (0.5–1 Hz) that groups both low-frequency (spindles and delta) and high-frequency (beta/gamma) rhythms by Steriade's group is particularly interesting. These investigations showed that the low-frequency rhythms are generated by the synchronous firing of cortical neurons during the depolarizing phase of the slow oscillation, which impacts on the thalamic circuitry to generate spindles and delta EEG activity. These findings of the generation of sleep rhythms, initially described in intracellular studies of neocortical and thalamic neurons, are now supported by EEG/MEG studies during human

sleep. The activity of neocortical neurons during slow-wave sleep is likely associated with neuronal plasticity that may play a role in consolidating memory traces acquired during the state of waking.

Spatial Resolution of EEG

Both for research and for clinical applications, the general use of microprocessors has made possible the development of comprehensive digital (paperless) EEG systems. These systems offer the possibility of data acquisition, artifact rejection, editing and processing of the data, brain mapping, and statistical treatment of the EEG and derived data such as frequency spectra. Brain maps represent the spatial distribution of a given variable, such as the peak amplitude of an evoked potential, a spectral component, or a correlation measure. Although brain maps suffer from several handicaps due to the limited sampling in space commonly applied in clinical settings, and due to the ambiguity introduced by the montage used, they are becoming very popular. The poor spatial resolution of the EEG is an important handicap in the application of this methodology to the study of brain functions. Research is being carried out in an attempt to improve this situation. An essential aspect is that the EEG has to be recorded with a sufficiently large number (>64) of scalp electrodes to improve its spatial resolution. In addition, some forms of spatial deblurring need to be applied. One way is to compute the Laplacian derivation (LD). Another methodology uses a 'finite element deblurring' algorithm to correct the EEG potential field for distortion due to the transfer from the cortical surface to the scalp, and thus to estimate the electrical potentials that would have been measured at the cortical surface, given the EEG scalp distribution. These new methodological advances permit using EEG signals with a higher spatial resolution in order to obtain objective correlates with cognitive functional states.

Still another methodology used to improve the localization power of EEG is to make a parametric model of its underlying generators and to estimate the model parameters from the observed EEG distribution using standard parameter estimation techniques. Such models generally describe the activated synapses as current dipoles, embedded in a conductor consisting of different compartments, with different electrical conductivities: the brain, the skull, and the skin. Various assumptions can be made to parameterize the temporal variations of the current dipoles: either all dipole parameters vary over time (moving dipole models) or only the amplitude (stationary dipole models), or all dipole moments vary over time (rotating dipole model).

To reduce systematic errors in the estimated dipole parameters, it is of crucial importance to use the appropriate geometric description of the conductor and the appropriate electric properties. Currently, several volume conductor models are in use: concentric spherical models or, recently, realistic models of the subject's heads obtained from magnetic resonance imaging (MRI) scans and considering usually at least three head compartments – brain, skull, and scalp. Because the dipole position parameters appear in a nonlinear way in the regression model, equivalent dipolar sources are found using iterative numerical methods. The electric resistivities of the head compartments can be estimated *in vivo* and individually for each subject using a parameterized iterative nonlinear model. This procedure highly decreases systematic errors due to the use of inadequate electric parameters describing the head compartments based on literature values derived from *in vitro* measurements containing large systematic errors. Thus, it also reduces the systematic errors associated with the estimation of dipole position.

This dipole modeling approach has been most useful in the analysis of transient EEG phenomena localized to restricted brain areas, such as evoked potentials and epileptiform spikes. Estimating equivalent dipolar sources of such events using realistic head models based on MRIs can be used to analyze spatial relations between those sources and epileptogenic zones associated with a lesion visible on the MRI (Figure 3).

The Combined Recording of EEG and fMRI

Despite several sophisticated modeling approaches, the use of EEG for localizing sources in the brain is restricted to situations in which it is known *a priori* that the number of sources is limited and that their geometry is relatively simple. In many of these cases, the sources can be described approximately by equivalent dipoles. Such conditions typically occur at short latencies of evoked potentials (when mainly primary cortices are activated) or when isolated events such as interictal spikes are considered. For the analysis of phenomena such as spontaneous rhythms, dipole models should be applied with care. New technological developments partly circumvent the limitations of dipole modeling. It has become possible to combine EEG and fMRI scans in order to enhance the possibilities of dynamical brain imaging, by recording EEG inside the MR scanner, during constant acquisition of echo-planar imaging (EPI) images. This mode of recording is identical to that used in fMRI studies, in which the presence and absence of a stimulus or task given to the subject is

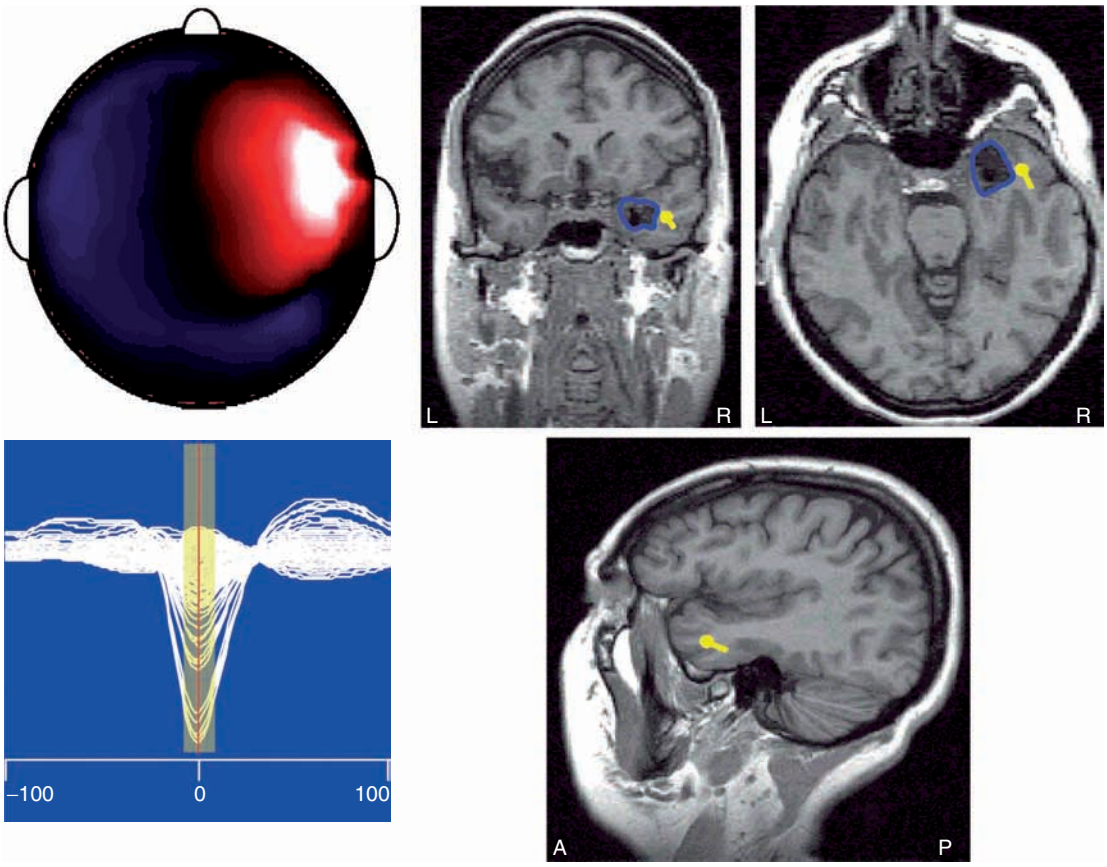

Figure 3 (Left) Spatial MEG map of the average of a cluster of interictal epileptiform spikes. (Right) Corresponding equivalent dipole (yellow dots and pointers) projected on three MRI slices (sagittal, horizontal, and coronal) and location of epileptogenic lesion. Note the location of the equivalent dipoles on the edge of the lesion.

correlated to the time variations of the MR signal at each pixel element of the EPI images. The localization power of fMRI is based on the fact that the fMRI signal from brain regions that are involved in processing the stimulus or task tends to correlate to the presence/absence of the stimulus/task, in contrast to brain regions that are not involved. The same underlying principle can be used in studies in which EEG and fMRI are recorded simultaneously. Under these conditions, the reference function correlated to fMRI is not necessarily the presence/absence of a task, but it can be derived from the EEG signal instead – for example, an epileptiform event. Combined EEG–fMRI was developed during the late 1990s with such clinical applications in mind, but the technique can also be applied to study other types of phenomena, such as sleep spindles and other spontaneous rhythms, whether or not they are in relation to cognitive tasks (Figure 4).

Although the simultaneous registration of EEG and fMRI is a very powerful combination with great potential, especially to study unpredictable phenomena that are thus impossible to study consistently with each technique separately (e.g., spontaneous brain rhythms), it should be realized that several technical precautions are necessary before it can be applied in practice. One point of concern is the safety of the subject, which is an issue because the long copper wires can pick up radiofrequency (RF) signals used in the MR scanner, and this can result in harmful currents if no safety resistors are used to reduce them. Other points of concern are the fact that standard EEG equipment does not have enough shielding and causes additional noise in the fMRI images. Finally, both the RF pulses and the gradient switches of the MR scanner cause large artifacts in the EEG that must be removed with dedicated software algorithms during postprocessing of the raw data.

Toward a New Era of Dynamical Brain Imaging

The EEG is an important aid to diagnosis in clinical neurology, mainly of irritative functional disturbances and particularly in epilepsy. Both the interictal and ictal EEG patterns of seizure disorders can yield important information for the diagnosis and

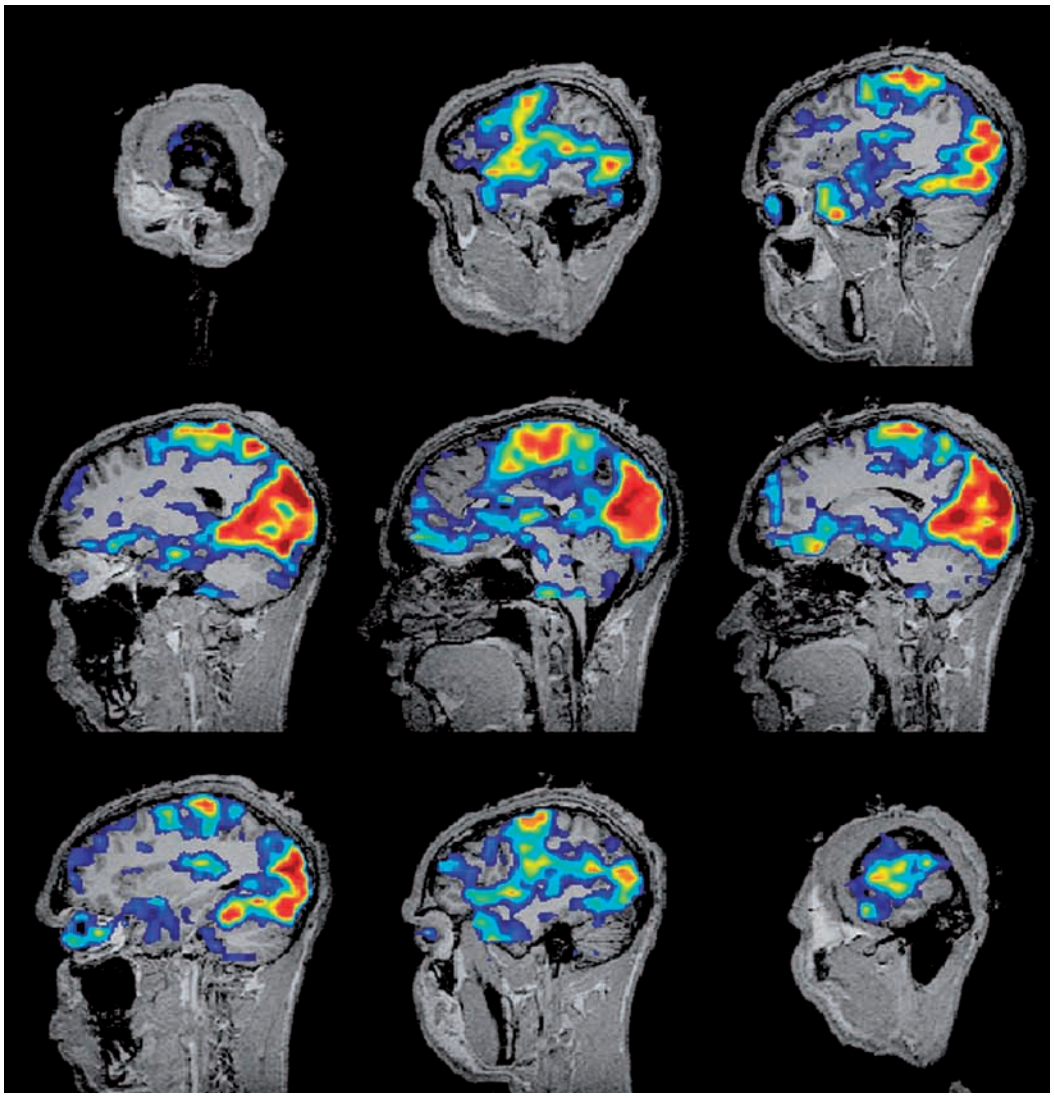

Figure 4 Correlation maps of the fMRI BOLD signal and the simultaneously recorded EEG (alpha rhythm frequency band). Correlation maps corresponding to the spontaneous variations of the alpha rhythm. The maps were obtained by computing the alpha power (occipital derivations) during each of the 600 coregistered EPI scans and correlating it with each of the BOLD signals, thereby accounting for confounders as motion, trends, and heartbeat effects. In the regression model, the alpha profile was shifted in time and multiplied with parameters, which constitute the hemodynamic response function. The significance of the partial correlation was tested using an *F* test, and the significances for all voxels were converted into a false detection rate (FDR). The map presents a color scale (cool colors – low; hot colors – large correlation values) of the (partial) correlation coefficient, which is thresholded at a correlation value of 0.22, corresponding to an FDR of 1%. The estimated hemodynamic responses (not shown) are mainly negative for voxels at the cortex and positive at the thalamus. The physiological meaning of this finding is that the thalamus is activated during generation of alpha band activity, whereas the cortex is relatively deactivated during these periods of time.

follow-up of these patients. The EEG is indispensable for the study of sleep disturbances and to characterize states of altered consciousness, such as in monitoring toxic effects of drugs, uremia, hypoglycemia, and, in general terms, coma. It is also of interest in monitoring some forms of anesthesia and in intensive care situations. Another interesting application of the EEG is to construct brain–computer interfaces, with the aim of controlling robotic devices in paralyzed patients using EEG-derived signals that can be modulated by mental imagery. Although the usefulness of the EEG in the diagnosis of brain lesions, such as tumors and stroke, has become less important because of the general availability of fine structural brain imaging techniques, mainly CAT and MRI scans, it still maintains its power in estimating the concomitant functional disturbances of the brain tissue. In addition, the modern quantitative EEG methodology is of increasing utility

in cases of difficult differential diagnosis and in the evaluation of functional disturbances related to psychiatric conditions.

The coregistration of EEG and fMRI opens new doors to the fundamental knowledge of the brain. In particular, it offers novel perspectives in the study of phenomena such as the spontaneous brain rhythms which play an important role as signaling elements associated with neurocognitive processes. This knowledge will eventually increase our understanding of neurocognitive processes in general and their disturbances in particular. However, the coregistration of EEG and fMRI also has the potential to shed more light into the fundamental relationship between electric and hemodynamic phenomena. The clinical relevance of this latter aspect is illustrated by the application of coregistered EEG and fMRI in epilepsy.

Technological developments of the EEG show that it is not a technique of the past but, on the contrary, holds the promise of contributing to a better understanding of the dynamics of the working of the brain.

See also: Aging: Brain Potential Measures and Reaction Time Studies; Electromyography (EMG) and Nerve Conduction Studies; Electrophysiology: EEG and ERP Analysis; Executive Function and Higher-Order Cognition: EEG Studies; fMRI: BOLD Contrast; Magnetic Resonance Spectroscopy; Perfusion MRI; Positron Emission Tomography (PET).

Further Reading

Basar E, Hari R, Lopes da Silva FH, and Schürmann M (eds.) (1996) *Alpha Activity: Cognitive and Sensory Behaviour.* Boston: Birkhäuser.

Freeman WJ (2006) Origin, structure, and role of background EEG activity. Part 4: Neural frame simulation. *Clinical Neurophysiology* 117(3): 572–589.

Gevins A, Leong H, Smith ME, Le J, and Du R (1995) Mapping cognitive brain function with modern high-resolution electroencephalography. *Trends in Neuroscience* 18: 429–436.

Gonçalves SI, de Munck JC, Pouwels PJW, et al. (2006) Correlating the alpha rhythm to BOLD using simultaneous EEG/fMRI: Inter-subject variability. *Neuroimage* 30(1): 203–213.

Hari R (2005) Magnetoencephalography in clinical assessment of human cortical functions. In: Niedermeyer E and Lopes da Silva FH (eds.) *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields,* 5th edn., pp. 1165–1197. Philadelphia: Lippincott Williams & Wilkins.

Lachaux JP, George N, Tallon-Baudry C, et al. (2005) The many faces of the gamma band response to complex visual stimuli. *Neuroimage* 25(2): 491–501.

Lopes da Silva FH (1991) Neural mechanisms underlying brain waves: From neural membranes to networks. *Electroencephalography and Clinical Neurophysiology* 79: 81–93.

Muller-Putz GR, Scherer R, Pfurtscheller G, and Rupp R (2005) EEG-based neuroprosthetics control: A step towards clinical practice. *Neuroscience Letters* 382: 169–174.

Niedermeyer E and Lopes da Silva FH (eds.) (2004) *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields,* 5th edn. Philadelphia: Lippincott Williams & Wilkins.

Nunez P (ed.) (1995) *Neocortical Dynamics and Human EEG Rhythms*. New York: Oxford University Press.

Pfurtscheller G and Lopes da Silva FH (1999) Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology* 110(11): 1842–1857.

Rodriguez E, George N, Lachaux JP, et al. (1999) Perception's shadow: Long-distance synchronization of human brain activity. *Nature* 397(6718): 430–433.

Salek-Haddadi A, Friston KJ, Lemieux L, and Fish DR (2003) Studying spontaneous EEG activity with fMRI. *Brain Research Reviews* 43: 110–133.

Steriade M (2006) Grouping of brain rhythms in corticothalamic systems. *Neuroscience* 137(4): 1087–1106.

Suffczynski P, Kalitzin S, Pfurtscheller G, and Lopes da Silva FH (2001) Computational model of thalamo-cortical networks: Dynamical control of alpha rhythms in relation to focal attention. *International Journal of Psychophysiology* 43(1): 25–40.

Wolpow JR and McFarland DJ (2004) Control of a two-dimensional movement signal by a non-invasive brain–computer interface in humans. *Proceedings of the National Academy of Sciences of the United States of America* 101: 17849–17854.

Electrolocation

L Maler, University of Ottawa, Ottawa, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Electric organs, found in several species of fish, produce weak electric discharges by the use of a modified muscle or nerve (electric organ discharge (EOD)). These feeble discharges (<1 V) are sensed by specialized electroreceptors (tuberous electroreceptors) distributed over the skin of the fish. Together, the electric organ and electroreceptors comprise the active electrosensory system; these fish also have a small number of ampullary-type electroreceptors sensitive to the low-frequency electric fields produced by other organisms (not discussed in this article). It is believed that electroreceptors evolved from lateral line receptors. However, unlike lateral line receptors, electroreceptors do not receive efferent input from the brain.

Active electroreception has independently evolved in two freshwater teleost families: the African mormyrid fish and the South American gymnotiform fish. The electrosensory system, like most other sensory systems, is used for two very different purposes: electrolocation and electrocommunication.

The patterning of EOD pulses can be either 'pulse' or 'wave' type (Figure 1). For a pulse fish, the pulse duration is far shorter than the interpulse interval, and the interpulse interval is, in many species, quite variable (i.e., an irregular electric discharge). For a wave species, the pulse duration and interpulse durations are similar and very constant in duration; the EOD of wave species therefore resembles a sine wave. Almost all mormyriform fish have pulse-type discharge (*Gymnarchus niloticus* is the one exception and is not discussed). Both pulse- and wave-type species are common in the gymnotiform family. There are many parallels in the electromotor/electrosensory circuitry of mormyrid and gymnotiform fish. Perhaps the greatest difference between African and South American electric fish is the presence of a corollary discharge or efference copy in the mormyrid, but not the gymnotiform, fish. Thus, in mormyrid fish, the medullary command nucleus sends a 'copy' of each EOD command spike to the electrosensory system, where it can be used to modify sensory processing. In gymnotiform fish, such a corollary discharge is not present, although a specialized type of time-coding electroreceptor can also signal the occurrence of each EOD.

Electrolocation

The EOD creates an electric field surrounding the fish. This field is distorted by the presence of conductors or nonconductors (Figure 2). For example, a crustacean such as *Daphnia* is more conductive than the water that electric fish inhabit. The EOD current will preferentially flow through the *Daphnia* and therefore activate a patch of electroreceptors in the skin below this tiny crustacean. A nonconductive object such as a rock will have the opposite effect: it will reduce the electric current over a localized patch of electroreceptors. The EOD is therefore a carrier; the signals important for electrolocation are amplitude modulations (AMs) of the EOD carrier. Electric fields fall off very rapidly with distance so that electrolocation is effective over only a few centimeters. Furthermore, unlike the visual system, the electric field has no focusing mechanism so that electric images are blurry. Despite these apparent disadvantages, the localized distortions of the electric field are still a sufficient cue for navigation and the electrolocation of small prey in total darkness.

Electric fish can locate their prey, as well as obstacles, in the dark. Like the visual system, and unlike the somatosensory system, this requires localization in three dimensions; accurate three-dimensional localization has in fact been behaviorally demonstrated in electric fish. Localization with respect to the fish's body axes (nose to tail and back to belly) is presumably carried out, as in the visual and somatosensory systems, with the aid of topographic maps. Work in both mormyrid and gymnotiform fish has suggested that the localization in the third dimension (distance from the fish's body) requires that the electrosensory system estimate the amount of blur of the electric image. The computations underlying this ability are not known and are the subject of current research.

There are now accurate models of the electric field for a number of species of electric fish. These models can be used to predict the effects of perturbations of the field by conductors and nonconductors at different distances from the fish's body. There are two major conclusions of these analyses: (1) the perturbations caused by prey are minute (<1 μV), and (2) at the swimming speeds used during prey capture, the perturbations caused by prey will produce low-frequency AMs (<20 Hz).

Gymnotiform fish have a ribbon fin that permits them to hover or swim forward and backward with ease. Excellent models of the mechanics of swimming in these fish are available. Combined, the models of electrosensory input and motor output provide insight into general questions of sensorimotor control during navigation and prey capture.

Electrocommunication

Electric fish can detect the EOD of nearby conspecifics; depending on the species, electrocommunication

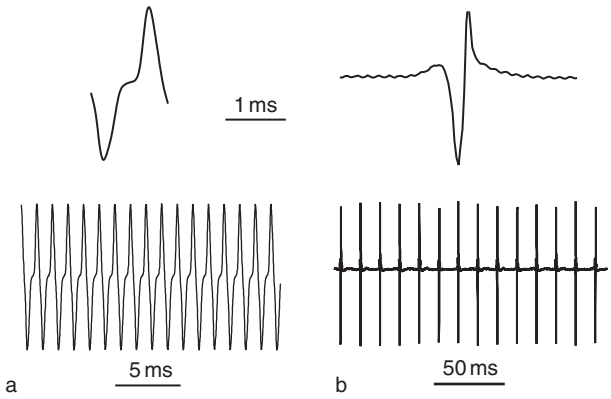

Figure 1 (a) The top panel shows a single EOD pulse of the high-frequency species, *Apteronotus leptorhynchus*, and the bottom panel illustrates the typical wavelike appearance of its EOD. This fish's frequency was 770 Hz. (b) The top panel shows a single EOD pulse of the low-frequency species, *Gymnotus carapo*, and the bottom panel illustrates the typical pulse train appearance of its EOD. The resting EOD frequency was 70–75 Hz for this fish. Data for this figure were provided by Jason Middleton.

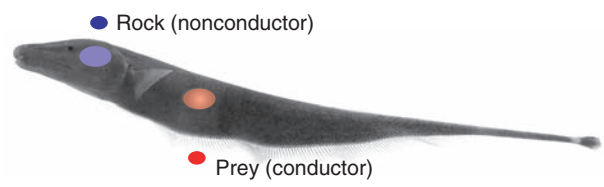

Figure 2 Electric images. An gymnotiform fish, *Apteronotus leptorhynchus*, is illustrated. The electric organ of these fish, located in the caudal half of the trunk and stretching back through the tail, forms an electric dipole at each electric organ discharge. The electric field created surrounds the fish and creates electric current flow through the cutaneous electroreceptors. Objects distort this electric field. A conductive object such as prey will cause an increase of local current density; this results in a blurred electric image (red) of increased electric current that increases the firing rate of the affected electroreceptors. In contrast, a nonconductor such as a rock will decrease the local current density. This results in a blurred shadow of decreased current flow (blue) and a reduction in firing rate of the affected electroreceptors. This is the basis of electrolocation.

is thought to be effective at distances up to at least 50 cm. The EOD is typically, at least over a short time scale, almost constant in amplitude. Electric communication signals are therefore based on changes in the EOD rate, and either increases or decreases in rate have been hypothesized to be important signals. A great deal of research in this area has focused on a 'wave-type' species of gymnotiform fish, *Apteronotus leptorhynchus*; the EOD of this species resembles a nearly constant high-frequency sine wave (species range, 650–1000 Hz), with males having a higher frequency than females. The frequency difference between males and females is an important signal for these fish. Some wave species forage in groups, and it has been proposed that the complex patterns (envelopes) produced by the summation of their EODs might act as an additional signal for identifying individual fish in the group.

Two major types of active electric communication signals have been described for this species: chirps and rises. Chirps are transient (~20 ms) increases in EOD frequency that are emitted mainly by mature males. Therre are two types of chirps: small (~100 Hz frequency increase) and large (~400 Hz frequency increase). Small chirps occur in response to the EODs of other males (EOD frequencies similar to those of the chirping fish) and are believed to be an aggressive signal. Large chirps are elicited by the EODs of females (EOD frequencies lower than those of the chirping fish) and are believed to be involved in courtship behavior. In contrast to electrolocation signals, chirps cause high-frequency amplitude modulations of the carrier EOD. The population of electroreceptors codes for small and large chirps with very different patterns of discharge. The extraction of these different patterns and final identification of the chirp type by higher order neural processing is an important problem under investigation.

Rises are long, slow increases in EOD frequency made primarily by females in response to the EODs of other females. Again, there may be several varieties of rises. It has been suggested that different types of rises represent submissive signals and female–female aggressive signals.

The neural circuits controlling the EOD have been worked out in some detail for both mormyrid and gymnotiform fish. A medullary command or pacemaker nucleus initiates each EOD and projects to the spinal electromotor nucleus. Diencephalic and mesencephalic prepacemaker nuclei in turn project to the command nucleus and control modulations of EOD frequency. As might be expected, the prepacemaker nuclei receive highly processed electrosensory input. However, these nuclei also receive substantial serotonergic input from the medullary raphe nuclei as well as multiple peptidergic inputs (enkephalin, galanin, substance P, and somatostatin) from hypothalamic sources. Presumably, these inputs mediate 'motivational' control over electric communication signals. It is therefore not surprising that the sexual dimorphism of chirp signals seen in some species is paralleled by the sexual dimorphism of serotonin and some peptidergic input to the prepacemaker nucleus. In addition, hormones such as testosterone, cortisol, and vasotocin can modulate electric communication signals, although the site of their action is not known. The ease of recording and manipulating electric communication signals has made electric fish a useful model system for studying the neural basis of animal communication.

Electroreceptor Coding

There are many types of electroreceptors and each utilizes different signal encoding strategies. In general, there are different types of electroreceptors tuned for conveying precise information on the timing versus the amplitude of electrosensory signals; the time and amplitude coding pathways are kept separate up to the midbrain. The most detailed computational analyses have focused on an amplitude coding electroreceptor found in wave-type electric fish termed the P-unit; P-units discharge in a probabilistic manner in response to the unperturbed EOD. Thus, for a fish with an EOD frequency of 1000 Hz, a P-unit with a p value of 0.15 would discharge at 150 spikes s^{-1}. The firing rate of a P-unit is controlled by the intensity of the electric field it senses; thus, a P-unit with a rock (nonconductor) above it might decrease its discharge to less than 10 spikes s^{-1}, whereas the presence of a *Daphnia* (conductor) near the receptor would increase its discharge to more than 300 spikes s^{-1}. The spatial pattern of the P-unit firing rate is one key coding strategy used by electric fish, as well as many other sensory systems, to gain information about spatial features of their environment.

The spontaneous discharge pattern of P-unit receptors is highly irregular, but careful statistical analysis reveals that it is not entirely random (not a renewal process). The statistical structure of the P-unit spike train optimizes these receptors for the detection of both low-frequency (electrolocation) and high-frequency (communication) signals. The mechanisms responsible for detecting such a wide frequency range of signals include biophysical reduction of noise levels by spike frequency adaptation and also frequency-dependent synchronization across the population of P-units. There are now very accurate models of P-unit electroreceptors; combining these models with those of the electric field permits researchers in this field to accurately predict the electroreceptor input to the brain in response to environmental signals such as rocks and prey as well as to communication signals such as chirps. Electroreceptors have been useful for investigating neural coding since the 1960s and are a useful model system for this purpose.

Central Pathways for Electroreception

Electroreceptors project to a hypertrophied medullary nucleus, the electrosensory lateral line lobe (ELL) (Figure 3). The ELL is a layered structure with separate layers for electroreceptor input, interneurons, principal cells, and feedback/descending input. The principal cells are mainly pyramidal cells with both basal and apical dendrites: The basal dendrites can receive input from electroreceptors, the soma receive input from interneurons, and the prominent apical dendrites receive extensive descending and feedback input. The ELL is also segmented. In mormyrid fish, there is one segment in receipt of ampullary input and two segments in receipt of different types of tuberous electroreceptors. In gymnotiform fish, one segment also receives ampullary input, whereas three segments receive identical collateral input from tuberous receptors including P-units. Thus, in gymnotiform fish, all three tuberous segments receive identical inputs; the specialized roles that these segments might play in electrosensory processing are under investigation. In all cases, the electroreceptor input is topographically arranged (all maps in both mormyrid and gymnotiform ELL). Thus, the familiar topographic principle of sensory representation seen in the somatosensory (somatotopy), visual (retinotopy), and auditory (tonotopy) is also expressed in the electrosensory system.

There are many subtypes of interneurons found in the ELL; a subset of these are GABAergic and, as in other sensory systems, engaged in regulating neural processing in both the temporal and spatial domains. In the temporal domain, inhibition allows the electrosensory system to respond to transient and high-frequency input.

Pyramidal cells in ELL have typical receptive fields: input to electroreceptors in a small patch of skin is capable of exciting these cells. In the spatial domain, specific types of inhibitory interneurons participate in the generation of surround receptive fields. Thus, many pyramidal cells have receptive fields with the center-surround organization very similar to that described for retinal ganglion cells. There is an even more remarkable similarity between retina and ELL. One morphologically defined population of ELL pyramidal cells is functionally equivalent to ON center–OFF surround retinal ganglion cells (receptive field center responsive to conductive objects); a morphologically distinct population is functionally equivalent to OFF center–ON surround retinal ganglion cells (receptive field center responsive to nonconductive objects). The presence of multiple topographic maps with different tuning curves is highly relevant to current theoretical analyses of the effect of tuning curves (center–surround organization in this case) on population coding.

Electroreceptors utilize glutamate as a neurotransmitter, as do most of the feedback pathways to the ELL. Both AMPA- and NMDA-type ionotropic glutamate receptors have been localized in the ELL as well as other regions of the electric fish brain (both gymnotiform and mormyrid). The NMDA receptors of gymnotiform fish have been cloned and found to

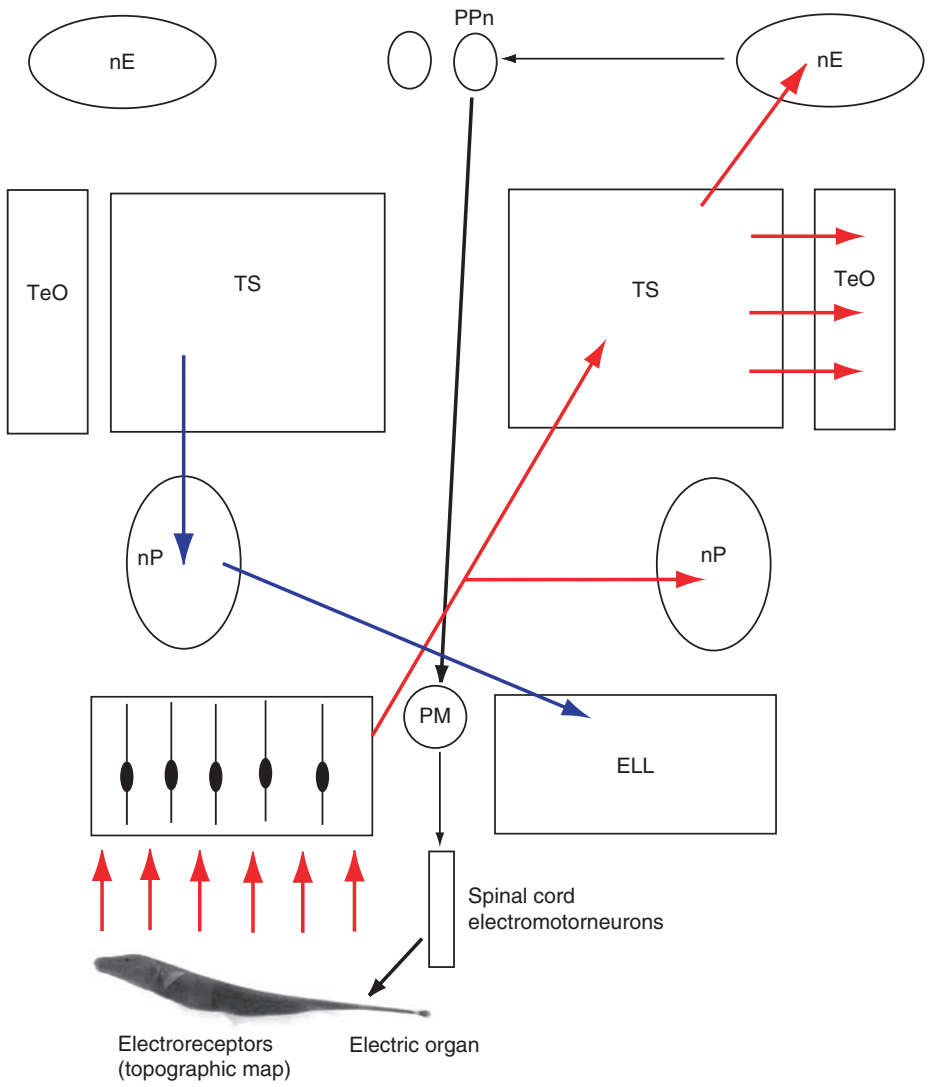

Figure 3 Simplified schematic of major neuronal pathways of a gymnotiform fish; there are similar pathways in mormyrid fish, but there are additional complex pathways related to the corollary discharge to sensory regions that are not illustrated here. Ascending electrosensory pathways are shown in red, feedback pathways in blue, and electromotor pathways in black. Only the most prominent contralateral projections are indicated. Electroreceptors project topographically onto the ipsilateral electrosensory lateral line lobe (ELL); the presence of three maps is not indicated here. The structure of a major projection neuron, the basilar pyramidal cell, is also indicated; the basal dendrite is in receipt of electroreceptor input. The ELL projects to the nucleus praeminentialis (nP) and the midbrain torus semicircularis (TS). The TS in turn projects topographically to the optic tectum (TeO) for control of movements associated with prey capture and navigation. The TS also projects diffusely to the nucleus electrosensorius (nE), a diencephalic region that extracts species-specific electric communication signals. The nE projects to the prepacemaker nucleus (PPn). The PPn also receives extensive additional input from hypothalamic and other brain regions and initiates electric communication signals by modulating the activity of the pacemaker nucleus. The PPn projects to the pacemaker nucleus (PM); cells in the PM cause the baseline (unmodulated) electric organ discharge (EOD). The PM projects to electromotor neurons in the spinal cord; these in turn project to the electric organ in the tail of the fish to generate the electric field that stimulates the electroreceptors. The nP receives both ascending (ELL) and feedback (TS) electrosensory input. In turn, it provides feedback projections to the ELL; these terminate predominantly on the apical dendrites of ELL pyramidal cells. The direct and indirect feedback pathways are not shown.

be highly conserved; ELL pyramidal cells express high levels of NMDA receptor. As discussed later, it may be critically important in adaptive plasticity of descending and feedback projections to ELL.

The dendritic specialization of ELL pyramidal cells (basal and apical dendrites receive different inputs) has prompted extensive work on the possible molecular and biophysical correlates of these anatomical differences. At least in the gymnotiform ELL, there appear to be important differences in the distribution of at least Na^+ and K^+ voltage-gated channels on the basal versus apical dendritic trees and perhaps in the distribution of various second messengers as well. Such differential dendritic distribution of ion channels has frequently

been reported for pyramidal cells of mammalian cortex, but its function remains unknown. In ELL, some of these dendritic ion channels appear to mediate and regulate dendritic backpropagating potentials that are important for spike bursts. Spike bursts have specific coding properties in the pyramidal cells (they detect low-frequency input). Thus, in ELL, it may be possible to directly link dendritic biophysics to sensory computations.

The ELL projects to predominantly contralateral higher brain centers via the lateral lemniscus, a thick myelinated fiber track. There are two main targets: a hindbrain nucleus termed nucleus praeminentialis (nP) and the midbrain torus semicircularis (TS). The nP is involved solely in feedback projections to ELL and is discussed later. The TS is similar to the inferior colliculus of mammals; there are far more neurons in TS than in ELL. The TS of gymnotiform fish has a complex layered structure, whereas that of mormyrids is organized into discrete nonlayered nuclear groups. For both families of electric fish, the different ELL maps converge in TS, as do the time and amplitude coding pathways.

The TS of gymnotiform fish has been studied in detail. It projects mainly to three target regions: the nP, the optic tectum, and the nucleus electrosensorius. The projections to the nP imply that higher order electrosensory processing can influence first-order processing in ELL (since nP is a feedback nucleus projecting back to ELL); this is a common theme in mammalian sensory systems in which cortical sensory areas provide strong feedback to brain stem sensory regions. The exact role of TS to nP feedback projections is not known.

The optic tectum (superior colliculus in mammals) is a highly conserved midbrain nucleus of vertebrates that typically receives retinal input and directs the eyes and body toward salient visual features. It is also typical that other senses, such as touch or audition, project to the deeper layers of the superior colliculus and can therefore also initiate orientation toward these stimuli. It is not surprising that TS (electrosensory) projects topographically to the deeper layers of the gymnotiform tectum and that there are tectal cells which respond to both electrosensory and visual input. Although this has not been studied in-depth, it is assumed that the tectum acts as the interface between the electrosensory and motor systems.

The TS projections to the nucleus electrosensorius (a diencephalic brain region) appear to be diffuse (not topographic). Electrophysiological recordings from this nucleus demonstrate that it responds to various types of communication signals, including chirps. Thus, there appears to be a segregation of neural processing by which electrolocation is mediated via the optic tectum, whereas electrocommunication is mediated by the nucleus electrosensorius. The nucleus electrosensorius projects to both the prepacemaker nuclei and the hypothalamic regions; the hypothalamic projections presumably mediate motivational functions of electrosensory input but have not been investigated in detail.

The TS of gymnotiform fish has been intensely investigated with respect to both its morphology and its electrophysiology. There are approximately 50 different cell types in TS, and a subset is tuned for specific frequency modulations of the EOD. The frequency selectivity is dependent on both voltage-gated ion channels and synaptic plasticity. Furthermore, some of these neurons are also very sensitive to object motion, and in some cases they are directionally selective (i.e., they may respond preferentially to head-to-tail movements or vice versa). Like the ELL pyramidal cells, different neurons are selectively responsive to conductive (prey) or nonconductive (rocks) objects. The TS clearly has far more specific tuning than its input ELL pyramidal cells (more sparse coding); the role of synaptic plasticity in sparsifying the ELL input is a key topic of research for the TS.

Feedback and Descending Pathways and the Cancellation of Redundant Signals

The nP of gymnotiform fish projects back to the ELL via three pathways: direct inhibitory and excitatory feedback and an indirect excitatory feedback.

A small group of GABAergic neurons project back directly to ELL pyramidal cells. These cells appear to be responsible for gamma range oscillations (30–40 Hz); the role of such oscillations in sensory processing remains to be firmly established, but it has been suggested that they may enhance directional movement detection.

A subset of glutamatergic nP cells (stellate cells) project directly back to the proximal apical dendrites of ELL pyramidal cells (both mormyrid and gymnotiform fish). This is a topographic, excitatory projection and has been hypothesized to mediate a 'searchlight' of attention to movement of objects such as prey.

Several different types of glutamatergic nP cells project back as mossy fibers to the granule cells of the caudal cerebellum. The recipient granule cells then project their parallel fibers down to the distal apical dendrites of ELL pyramidal cells. These parallel fibers make excitatory synaptic connections onto dendritic spines that utilize both AMPA and NMDA receptors. Studies have demonstrated that this pathway (in both mormyrid and gymnotiform fish) supports anti-Hebbian plasticity, likely implemented by NMDA receptors. This plasticity in turn mediates the cancellation of predictable sensory input.

In mormyrid fish, the corollary discharge is a predictive signal arising from the electromotor command nucleus. The corollary discharge signal reaches the caudal cerebellum; when descending corollary discharge input to the pyramidal cell's apical dendrites is consistently paired with electroreceptor input, the descending input is modified so as to form a 'negative image' that cancels the electroreceptor input. The corollary discharge represents a very brief signal (each EOD), and in this case the anti-Hebbian plasticity was shown to require spike-timing-dependent plasticity (STDP). This was one of the first demonstrations of STDP and is still the only one linked to a well-defined function.

In gymnotiform fish, a similar circuit is also present. In this case, however, the predictable signal comes from sources such as tail bending. The rhythmic tail bending of many gymnotiform fish causes predictable amplitude modulations of the EOD. These low-frequency AMs are readily detected by the electroreceptors but are ignored by most ELL pyramidal cells. The cancellation of such redundant low-frequency signals is also mediated by the cerebellar feedback fibers to distal apical dendrites of ELL pyramidal cells; again, the mechanism is that of anti-Hebbian synaptic plasticity. In this case, the circuitry has also been analyzed in detail and involves nonadaptive pyramidal cells driving the plasticity of the adaptive ones. Presumably, the cancellation of such redundant signals improves the detection of more important unpredictable signals such as moving prey.

In both mormyrid and gymnotiform fish, NMDA receptors associated with the cerebellar feedback/descending input are critically involved in the cancellation of predictable input. It is also likely that second messenger cascades, including protein kinase A (cAMP-dependent phosphorylation) and calcium/calmodulin-dependent kinase, are essential for this process. Further work in this area promises to link molecular and cellular mechanisms directly to systems processes such as redundancy reduction.

More Complex Learning

The forebrains of gymnotiform and (especially) mormyrid fish are both larger and better differentiated than those of most other teleosts. Furthermore, there are high levels of both NMDA receptors and protein kinases in the dorsal (pallial) forebrain of gymnotiform (and likely mormyrid) fish. This suggests that the weakly electric fish might be capable of learning complex tasks. Work in this area is only in its initial stages but does suggest that the electric sense can support the learning of spatial shapes and patterns.

See also: Electrical Perception and Communication; Electrocommunication; Optic Tectum: Development and Plasticity.

Further Reading

Benda J, Longtin A, and Maler L (2006) A synchronization–desynchronization code for natural communication signals. *Neuron* 52: 347–358.

Bullock TH and Hopkins C (eds.) (2005) *Handbook of Auditory Research, Vol. 21. Electroreception*: New York: Springer.

Fortune ES (2006) The decoding of electrosensory systems. *Current Opinion in Neurobiology* 16: 474–480.

Graff C, Kaminski G, Gresty M, and Ohlman T (2004) Fish perform spatial pattern recognition and abstraction by exclusive use of active electrolocation. *Current Biology* 14: 818–823.

Krahe R and Gabbiani F (2004) Burst firing in sensory systems. *Nature Reviews Neuroscience* 5: 13–23.

Maler L (2007) Neural strategies for optimal processing of sensory signals. Symposium on Computational Neuroscience. *Progress in Brain Research* 165: 137–157.

Middleton JW, Longtin A, Benda J, and Maler L (2006) The cellular basis for parallel neural transmission of a high-frequency stimulus and its low-frequency envelope. *Proceedings of the National Academy of Sciences of the United States of America* 103: 14596–14601.

Perkel DH and Bullock TH (1968) Neural coding. *Neuroscience Research Program Bulletin* 6: 223–344.

Ramcharitar JU, Tan EW, and Fortune ES (2006) Global electrosensory oscillations enhance directional responses of midbrain neurons in Eigenmannia. *Journal of Neurophysiology* 96: 2319–2326.

Rose GJ (2004) Insights into neural mechanisms and evolution of behavior from electric fish. *Nature Reviews Neuroscience* 5: 943–951.

Sawtell NB, Williams A, and Bell CC (2005) From sparks to spikes: Information processing in the electrosensory system. *Current Opinion in Neurobiology* 15: 437–443.

Zakon H (2003) Insight into the mechanisms of neuronal processing from electric fish. *Current Opinion in Neurobiology* 13: 744–750.

Electromyography (EMG) and Nerve Conduction Studies

J M Shefner and L Simionescu, SUNY Upstate Medical University, Syracuse, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Definitions

Electromyography (EMG) is a technique whereby extracellular current fluxes are recorded from muscles at rest and during voluntary activation. Most commonly, signals are recorded from an intramuscular concentric needle electrode, but recordings can be made from surface electrodes placed on the skin overlying the muscle to be studied. Signals are amplified and filtered digitally and displayed in real time. When needle electrodes are used, action potentials from single muscle fibers that are spontaneously active (fibrillation potentials) may be recorded, as may the combined response of all muscle fibers innervated by a single motor axon (motor unit potential). Motor unit potentials may occur spontaneously (fasciculation potentials) but usually are seen in response to voluntary, reflex, or electrically stimulated activity of motor neurons within the central nervous system. With appropriate processing, motor unit potentials can sometimes be isolated from signals recorded with surface electrodes, but responses of individual muscle fibers and motor units in deeper muscles usually cannot be seen without intramuscular electrodes.

Electromyographic signals are recorded for research purposes to assess the timing and level of voluntary, reflex, or other muscle activation; simultaneous recordings from multiple muscles can also be used to investigate central control of motor activity. However, EMG is primarily a clinical tool used to evaluate a variety of diseases in which the peripheral neuromuscular system is affected. Motor unit potential morphology and discharge patterns are altered by both muscle and nerve disease, usually in opposite ways. Because only muscle membrane activity is recorded (and not force), disorders that primarily affect the contractile apparatus within muscle will not affect individual motor unit potentials but may alter discharge rates.

In clinical laboratories, EMG is usually performed in conjunction with nerve conduction studies (NCSs); these studies usually involve electrical stimulation of a mixed sensory and motor nerve bundle with recording of compound sensory responses from cutaneous fibers, or they involve compound motor action potentials from muscle fibers innervated by the nerve being stimulated. Because only current fluxes rather than transmembrane potentials are being recorded, sensory NCSs assess conduction only in the largest diameter myelinated fibers that produce the greatest current flows, whereas motor studies reflect activity in most or all of the alpha motor axons directed to the muscle being recorded. NCSs are particularly useful to evaluate generalized and focal disorders of nerves rather than primary muscle disorders. Although NCSs have evolved primarily as a clinical tool, research applications continue to appear. One current research application is in the area of excitability testing, in which NCSs are used to evaluate the local sources and time courses of ionic currents that occur in response to an evoked action potential. Finally, neuromuscular transmission can be studied using both EMG and NCS.

With a specially designed EMG needle and filters that attenuate spatially remote signals, action potentials from two muscle fibers within the same motor unit may be recorded simultaneously (single-fiber EMG). The precise temporal relationship between the firing patterns of the two fibers provides information about the effectiveness of neuromuscular transmission. In addition, if motor nerves are stimulated repetitively, the extent to which the compound motor action potential amplitude declines with repeated stimulation is a measure of the inability of the neuromuscular junction to sustain transmission. Although EMGs and NCSs are typically used to evaluate processes peripheral to the spinal cord, more central systems may also be studied. The mono/oligosynaptic stretch reflex may be studied electrically by stimulating muscle primary (Ia) afferents and recording motor potentials from the corresponding muscle (H reflex). Prior stimulation or concurrent sensory stimulation to evaluate, in particular, spinal circuitry subserving recurrent, reciprocal, and presynaptic inhibitory pathways may modify this reflex. In addition, the recently developed technique of transcranial magnetic stimulation can be used to evaluate the integrity of motor pathways from cortex to muscle.

Studies of Nerve Conduction

Many aspects of axon conductivity may be assessed with nerve conduction studies; in the clinical laboratory, however, only conduction velocity and response amplitude are measured. Motor axons and sensory axons can be evaluated separately, and different subtypes of sensory fibers can be selectively studied. Axon numbers are reduced in a variety of clinical disorders. In generalized neuropathies, axons of many or all modalities are lost, often with a distal-to-proximal gradient, whereas focal neuropathies reduce axon numbers in a particular nerve distribution. More proximal disorders of nerve roots or spinal cord will

also result in axonal dropout; however, if the site of the disease process is proximal to the dorsal root ganglion, only motor fibers will be affected. Thus the ability to distinguish between loss of motor or sensory axons is of clear clinical importance.

Measurements of conduction velocity are useful to evaluate the pathogenesis of different neuropathic conditions. When the primary site of pathology is the nerve cell body or axon itself, the number of functioning axons will be reduced but conduction will be unaffected in those axons remaining. In contrast, other disease processes affect the Schwann cell or the myelin sheath that surrounds large-diameter, rapidly conducting axons. Disruption of the myelin sheath may produce slowing of conduction in some or all myelinated fibers, sometimes with relatively little associated axon loss. In addition, focal areas of demyelination can render an axon locally unexcitable, even though conduction may persist both proximal and distal to the area of damage. Such an axon would be unable to transmit impulses between the spinal cord and the appropriate target organ. Such focal dysfunction is known as conduction block.

Clinicians perform nerve conduction studies by applying an electrical current pulse to the skin overlying a mixed (sensory and motor) peripheral nerve. With such stimulation, individual axons can be stimulated to produce action potentials. Nerve fibers are much more sensitive to electrical stimulation than muscle fibers are, so even when a nerve lies beneath muscle it is almost always preferentially stimulated by surface electrodes. So-called direct electrical stimulation of muscle with current pulses less than a few milliseconds in duration actually produces contraction indirectly by stimulating motor axons as they pass through the muscle. Within the nerve itself, certain axons are more apt than others to give rise to action potentials. In general, larger-diameter axons have higher sensitivity (lower thresholds) to external electrical stimulation; therefore, when a mixed nerve is stimulated, large-diameter sensory fibers are activated most easily, followed by axons of alpha motor neurons. Geometric considerations are also very important; fascicles lying closer to electrodes on the skin surface will be stimulated more effectively than will be deeper-lying fascicles. With appropriate stimulus intensities, it is usually possible to stimulate all large-diameter (sensory and motor) axons; to stimulate smaller myelinated sensory fibers and any unmyelinated fibers, extremely strong (and therefore quite painful) stimuli are required.

Motor Conduction Studies

A stimulus of appropriate intensity and location will cause action potentials to arise in all motor axons present in a mixed nerve, and a compound muscle action potential (M response) can then be recorded using a surface electrode placed over the motor point of a muscle innervated by the stimulated nerve, with an indifferent electrode placed over a bony prominence. The recorded potential can be characterized in a number of different ways; the two clinically most useful measures are (1) onset latency and (2) baseline to peak amplitude. The onset latency reflects conduction in only the fastest few fibers and is a function of nerve conduction velocity, speed of neuromuscular transmission, and conduction along the muscle membrane. If, however, the same nerve is stimulated at a second location, the difference in onset latencies can be used to calculate nerve conduction velocity if the distance between the stimulating electrodes is known. Obviously, the accuracy of such measurements depends on the ability to determine the length of nerve between the two locations.

Because motor axons conduct with a fairly restricted range of velocities, and the duration of muscle action potentials is much longer than nerve action potentials, the amplitude of the compound response directly assesses the number of muscle fibers activated. This number is not a direct measure of the number of motor axons responding; in conditions of chronic axon loss, surviving axons reinnervate muscle fibers that have lost synaptic contact. Thus a reduced number of motor axons may still contact all available muscle fibers. Therefore, motor action potential amplitude is not a sensitive measure of axon loss.

The compound motor action potential can be studied in a way that more directly assesses both axon loss and resultant terminal reinnervation. This technique, known as motor unit number estimation (MUNE), was first proposed more than 30 years ago, but continues to be refined and is increasingly employed clinically and as an outcome measure in clinical trials. A number of MUNE techniques have been developed and remain in use; all rely on the same basic premise: the response amplitude of an average single motor unit potential (SMUP) is calculated, then it is divided into a maximum response, the compound muscle action potential (CMAP), to derive a number that estimates the total number of units present (the MUNE). Of the MUNE techniques that have been devised (incremental, multiple point stimulation spike-triggered averaging, F wave, and statistical methods), only the method of sampling the SMUP responses varies. Although differences of opinion remain among MUNE investigators as to which method should be used, the statistical and the multiple point stimulation (MPS) MUNEs currently enjoy the most widespread use.

In diseases characterized by progressive motor unit loss, such as amyotrophic lateral sclerosis (ALS),

MUNE can be used reliably in the longitudinal evaluation of patients. Multiple studies have demonstrated a decline of the number of the motor neurons with disease progression. MUNE has been employed as an outcome measure in several multicenter ALS treatment trials. MUNE can be used for assessing the effects of both neuroprotective therapies and therapies designed to enhance collateral reinnervation as it enables quantitation of motor unit numbers while simultaneously estimating the collateral reinnervation. MUNE may also have a significant potential in clinical decision making, as suggested by studies showing that a single MUNE evaluation and a demonstration of rapid decline of MUNE are both strong predictors of survival in ALS. In addition, it was shown that MUNE could be used to detect motor neuron loss prior to onset of symptoms.

In most situations, amplitude and shape of the compound muscle action potential is invariant to changes in stimulating electrode position, so that proximal and distal stimulations produce responses that differ in onset latency only. However, diseases that affect myelin may cause dispersion in conduction velocities, so that the amplitude of the compound response may decrease with proximal stimulation even though the number of stimulated fibers remains the same. This may be confused with true conduction block; however, because the pathophysiology of both events is segmental demyelination, the distinction is not critical.

Sensory Conduction Studies

Sensory conduction studies require recording of nerve action potentials directly. Because electrical current fluxes surrounding nerve are much smaller than those surrounding muscle, up to 1000 times greater amplification is often required to record from the former. Typically a mixed nerve is stimulated, and recordings are made from electrodes placed on the skin over the cutaneous field supplied by the nerve. Conductive gels and mild skin abrasion are used to reduce the impedance between skin and recording electrode. The compound sensory action potential is a more direct reflection of nerve function than is the motor action potential just discussed, because no intervening neuromuscular junction exists and the response amplitude is a more direct reflection of total number of nerve fibers responding. Thus conduction velocity can be calculated from a single recording site. However, given that the distribution of axon diameters is considerably greater for sensory than for motor axons, both onset latency and response amplitude are indicators of conduction in only the largest diameter, fastest conducting axons.

Conduction in slower-conducting sensory fibers can be studied in several different ways. Instead of recording from electrodes applied to the skin surface, needle electrodes may be placed within several millimeters of the nerve (near-nerve recording technique). With appropriate amplification and filtering, recordings from these needle electrodes reflect activity in axons with diameters as small as one-fourth of the largest diameter sensory axons (small myelinated fibers in Aδ range). Although many myelinated and all unmyelinated fibers are still left unrecorded, the ability to analyze activity in smaller myelinated sensory axons is very useful in the evaluation of a variety of different neuropathic conditions.

To be able to record the response of unmyelinated fibers (C fibers) the technique of microneurography was developed. As unmyelinated fibers have a diameter of less than 1 μm and conduct as slow as 0.5 m s^{-1}, the current fluxes generated by action potentials are too small to be recorded by electrodes outside the nerve bundle. With microneurography, a tungsten electrode is inserted directly into the nerve bundle being studied and single unmyelinated C fiber responses can be recorded. Important applications of this technique include analysis of sympathetic efferent activity, as well as investigation of the receptive field properties of different afferents, such as receptive field size and preferred stimulus.

Small-diameter sensory and motor axons can also be evaluated using tests of autonomic function. Changes in skin resistance can be measured; they are mediated by activity in unmyelinated autonomic efferent sudomotor fibers to the hand or foot after one or other innocuous stimulus. The latency of this sympathetic skin response is very long, because efferent conduction velocity is about 1 m s^{-1}. This response (sometimes called the psychogalvanic reflex) is absent in conditions that interrupt conduction in unmyelinated efferents. Because conduction in unmyelinated fibers cannot be slower than it is normally, abnormalities are manifested by absence of the response rather than increases in its latency. In addition, autonomic afferent and efferent pathways can be studied by recording the variability of heart rate as a function of controlled deep breathing.

Axons may be affected by disease or external stress in ways that are reflected by changes in membrane excitability. A noninvasive battery of excitability measures has been developed that yield information about specific ion conductance channels, as well as how such channels are affected by disease processes. Threshold tracking measures, rheobase, and activity-dependent axonal conductance measures have been obtained in normal individuals, as well as in patients with neuropathy and motor neuron disease. Results in patients with diabetic neuropathy suggest a defect in inward rectification, while ALS patients tend to

show a pattern of abnormalities suggesting a persistent sodium conductance not seen in normal persons. Such insights are not available using traditional methods of clinical neurophysiology; thus, excitability testing holds the promise for more detailed study of membrane processes in a variety of peripheral nerve disorders.

H Reflex and Late Responses

To study conduction in more central segments of peripheral nerve, including plexus and roots, volleys of impulses produced by peripheral stimulation pass centrally into the spinal cord and then back out to muscles where the activity can be recorded. Tendon reflexes are produced by activity traveling toward the cord in large-diameter Ia sensory fibers and then back to muscles in alpha motor axons. Their neurophysiologic counterpart is the H reflex, in which afferent volleys are produced by electrical stimulation of spindle primary afferents rather than by mechanical stimulation of their intramuscular endings, and the compound muscle action potential is recorded rather than the clinically observed joint movement. Absence of the H reflex or marked prolongation of its latency usually reflects abnormalities in peripheral nerve conduction, but it may not be possible to separate sensory from motor abnormalities. Supramaximal stimulation of a mixed nerve produces, in addition to impulses in sensory axons and in motor axons conducted distally to the muscle, impulses in the same motor axons that travel antidromically through the ventral roots into the motor neuron pool. An impulse in some of these axons, once it reaches its motor neuron, fails to die out and, in that motor neuron, gives rise to another impulse that is then conducted orthodromically back out, where it can be recorded from muscle as an F wave. Neural activity responsible for F waves travels both to the spinal cord and back out again through motor axons: F waves are therefore not reflexes – their afferent arcs are not made up of sensory fibers. The F wave latencies are nevertheless useful, because they are inversely proportional to conduction velocity throughout the entire length of a motor axon.

Clinical EMG

EMG provides invaluable information about the functional status of individual motor units, including, of course, their spinal motor neurons. Electromyographic evaluation of a single muscle involves investigation of three distinct properties. First, electrical activity present during complete relaxation is studied. Normal muscles are silent; that is, no significant spontaneous activity occurs in the motor neuron or muscle fibers. However, any process that causes loss of motor axons will, at least acutely, result in the presence of some denervated muscle fibers. Initially, such fibers are silent; however, over the course of 2–3 weeks (and associated with the proliferation of postsynaptic acetylcholine receptors along the sarcolemmal surface), they become spontaneously active and generate action potentials called fibrillation potentials. Once established, fibrillations from a denervated muscle will persist until contact with another motor axon is established. In addition, if a motor axon is damaged but not destroyed, spontaneous activity may also occur, the origin of which is the axon itself; such potentials are called fasciculations. Muscle diseases usually do not involve destruction of motor axons, so it might be expected that spontaneous activity would be absent. However, in many (particularly inflammatory) myopathic processes, muscle fibers split. Under those conditions, part of the fiber may remain connected to a motor axon but the other part is functionally denervated. If an axonal connection is not rapidly established, fibrillation potentials will occur in this latter part.

In addition to analyzing spontaneous activity, EMG involves the study of voluntarily activated motor units. Mild voluntary contraction usually results in the activation of one to three different motor units within the region being recorded by the needle electrode. The morphology of an individual motor unit potential is a product of the number of muscle fibers in that unit and the spatial distribution of those fibers with respect to the recording electrode. Within a muscle, a normal motor unit includes some but not all muscle fibers in a specific location (Figure 1). The amplitude of the recorded potential from that unit will be a reflection of both the number of fibers activated and their proximity to the electrode. The duration will reflect primarily the spatial extent of the motor unit. Diseases that produce a loss of motor axons will, after 2–3 months, also be associated with remodeling of existing motor units, so that more muscle fibers are innervated by a single axon. This will produce the electrical changes shown in Figure 1; motor unit potentials will be larger and more prolonged than normal. In contrast, diseases that are primarily directed at muscle itself will not alter the number of motor units but will reduce the number of functioning muscle fibers; motor unit potentials will then be abnormally small in amplitude and short in duration.

In most clinical laboratories, evaluation of motor unit morphology remains a subjective task; electromyographers view the potentials in real time on a computer screen and derive an impression of whether potentials are small or large, brief or prolonged.

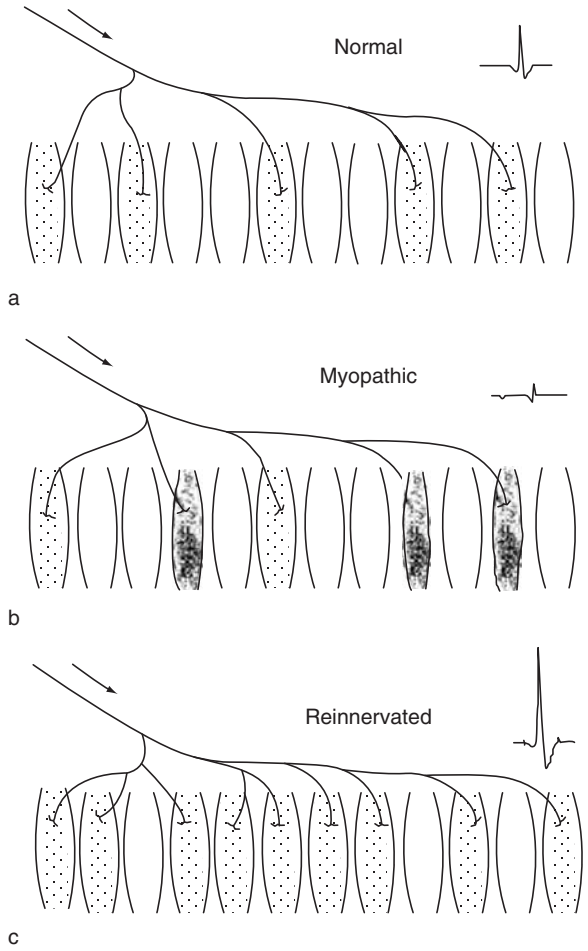

Figure 1 A single motor unit branches terminally to innervate some but not all muscle fibers within an area of muscle. The action potential produced by each motor unit is seen in the upper right of each section. The normal-appearing but unshaded fibers belong to other motor units. (a) In a normal motor unit, some but not all muscle fibers in a local area are innervated by the same axon. (b) In a myopathic unit, terminal axons are attempting to innervate dysfunctional muscle fibers, resulting in an action potential that is reduced in amplitude and duration. (c) In a reinnervated motor unit occurring as a consequence of chronic axon loss, a single axon innervates more muscle fibers than in the normal situation, resulting in an action potential that is larger in amplitude and longer in duration.

Usually the signals are stabilized by a triggering mechanism. Quantification of duration and amplitude using either weighted averaging or template-matching algorithms is available in most EMG machines, but is not often used in routine studies. In addition to analysis of spontaneous activity and motor unit morphology, EMG also evaluates the pattern of motor unit activation. The central nervous system encodes an increase in muscle activation in two ways: (1) motor units that are initially active at low levels of force production increase their firing rates and (2) motor units that had previously been silent are recruited into activity. Thus an EMG needle recording from a muscle with increasing activity should note these two parallel processes. If the number of motor units is reduced because of motor axon loss, the EMG should show the normal increase in firing rate in those motor units available to discharge, but a reduction in the number of additional motor units recruited as force is increased. This reduced motor unit recruitment occurs immediately after nerve injury preceding the development of fibrillation potentials or changes in motor unit morphology. In contrast, if the number of motor units is normal but the number of muscle fibers is reduced because of muscle disease, an attempt to increase muscle activation will result in both the increase in firing rate of previously active units and a rapid involvement of other units, because the force generated by a single unit will be less than normal. Thus muscle disease is characterized by early recruitment of smaller than normal motor units.

To further obtain accurate data regarding motor unit potential (size, configuration, and firing characteristics), a number of quantitative EMG techniques have been developed. An automated method of decomposition-based quantitative electromyography (DQEMG) can be applied to interference patterns recorded simultaneously from intramuscular and surface electrodes. The EMG-interference pattern can be decomposed into its constituent motor unit potentials. Both motor unit morphology and firing pattern may be studied. Surface-recorded motor unit potentials derived from decomposition may be averaged to yield an average motor unit size, from which a motor unit number estimate may be calculated. Currently, decomposition techniques are primarily research tools; however, the potential exists for DQEMG to provide a reliable method for evaluating the disease progression as well as the response to treatment interventions.

The MUPs recorded via concentric needle electrodes reflect the electrical activity of a portion of a motor unit. To obtain information regarding the entire motor unit, another technique using macro needle electrodes was developed. By using a modified single-fiber EMG needle electrode, a single-fiber potential is first recorded and then used as a trigger for recording and time-locking surface potentials, which are averaged to determine the surface electrical activity of a whole motor unit. The parameters measured are the amplitude, area, and duration of the averaged SMUPs. Comparing the macro-EMG with a concentric-needle EMG, regarding their sensitivity in detection of neuromuscular disorders, it was shown that the techniques appear to have similar sensitivities and that by combining the results of both methods additional increase in sensitivity can be achieved.

Studies of Neuromuscular Transmission

Single-Fiber EMG

A standard concentric needle EMG electrode has a large recording surface, designed to record potentials from at least 20 neighboring muscle fibers. However, by reducing the size of the recording area and by filtering out distant potentials, it is possible to record from a small number of individual muscle fibers within the same motor unit (single-fiber EMG).

The average number of fibers from the same motor unit seen with repeated electrode penetrations is called the fiber density, and is a measure of the geographic concentration of single muscle fibers belonging to a given unit. In addition, when two to three fibers from one motor unit are recorded simultaneously, variability in the temporal relationship between activation of both fibers can be studied. This variability is called jitter and is extremely small under normal conditions. However, diseases of the neuromuscular junction, such as myasthenia gravis, cause variability in the time taken by neuromuscular transmission and therefore increase the jitter. Occasionally, neuromuscular transmission is so tenuous that one of the fibers being recorded fails to fire at all; this is called blocking and is a sign of severe disease. Jitter is the most sensitive measure of dysfunction at neuromuscular junctions but it is not specific, and increased jitter can be seen with neuropathic processes as well.

Repetitive Stimulation

Compound motor action potential response area is positively correlated with the number of muscle fibers activated and, if neuromuscular transmission is intact, with the number of motor axons activated. Repetitive supramaximal stimulation of motor nerves is used clinically to assess function at neuromuscular junctions. With supramaximal electrical stimuli (all alpha motor axons being stimulated), changes in M amplitude, duration, or area with repetitive stimulation reflect disturbances in neuromuscular transmission. Disorders such as myasthenia gravis produce a typical decremental pattern of M response amplitude/area, but only in an affected muscle, when the motor nerve is stimulated at 2–3 Hz. The first M response is usually not as large as normal, but each subsequent one becomes smaller until the fourth or fifth response is reached, at which time the decrement ceases; subsequent responses often become slightly larger. This pattern (because of reduction in number of functional postsynaptic acetylcholine receptors) reflects normal quantal rundown; the number of quanta of acetylcholine released by each successive impulse decreases, as does the postsynaptic response, until presynaptic potentiation increases acetylcholine release. Because many more quanta are normally released than are necessary for maximal postsynaptic depolarization (a large safety factor), quantal rundown is only seen when neuromuscular transmission is compromised. In contrast, presynaptic neuromuscular junction disorders (such as botulism or Eaton–Lambert (myasthenic) syndrome) may show some decrement with 2–3 Hz of stimulation but show striking increases in response amplitude at either rapid rates (30–50 Hz) of stimulation or immediately after a short period of maximum voluntary exertion.

Central EMG/NCSs

Historically, EMG and NCSs have been used as aids to the diagnosis of peripheral nervous system disorders. However, they can provide insight into central processes as well. By recording either needle or surface EMG signals from multiple muscle groups simultaneously, movement disorders such as tremor and dystonia can be studied. Different types of tremors are characterized by distinct patterns of EMG activity in agonist and antagonist muscle groups. For example, parkinsonian tremor is electromyographically alternating; that is, one set of muscles is active during relative silence of its antagonists. Enhanced physiologic tremor and essential/familial tremor, in contrast, are characterized by co-contraction; agonist and antagonist muscle groups are active simultaneously. EMG also aids analysis of dystonias, allowing clinicians to determine which muscle groups are active during the assumption of specific postures. This information is of critical importance in planning treatment, which frequently consists of weakening individual muscle groups with local injections of botulinum toxin.

In addition to evaluation of conduction in peripheral nerve, it has recently become possible to stimulate brain or spinal cord directly and record peripheral responses. This can be done with high-voltage electrical stimuli applied to the scalp or spinal column; perhaps not surprisingly, this is an uncomfortable procedure. Magnetic stimulation is a less painful alternative, with stimulators that produce brief, powerful magnetic fields. When coils producing these magnetic fields are placed over the skin near nervous tissue, these fields in turn induce in that tissue an electrical field that can stimulate neighboring cortical, spinal cord, or peripheral nerves, therefore providing a method for testing the functional integrity of the corticospinal tract as well as the motor cortex excitability. By using transcranial magnetic stimulation (TMS), motor-evoked potentials can be recorded from arm or leg muscles and, after estimates of peripheral conduction time are subtracted, an estimate of central conduction time can be made. Central conduction

has been demonstrated to be abnormal in a variety of neurologic diseases, including multiple sclerosis and amyotrophic lateral sclerosis, even prior to onset of symptoms. Therefore, TMS represents a useful investigational tool, allowing early diagnosis, monitoring of disease progression, and estimating prognosis. In addition, a newer TMS technique, repetitive transcranial magnetic stimulation (rTMS), has been developed as a promising therapeutic tool, by modifying the excitability of the cerebral cortex. Tentative therapeutic applications at this time include Parkinson's disease, epilepsy, stroke, and depression.

Spasticity, another central disorder, has also proved amenable to study with methods adapted from routine NCSs. As discussed previously, the H reflex is analogous in many ways to a deep tendon reflex; as such, it provides a qualitative measure of upper motor neuron abnormalities. More information about spinal cord processing can be obtained by conditioning the H reflex in a variety of ways. Rapid tendon vibration effectively stimulates primary afferent fibers and acts as a potent stimulus for presynaptic inhibitory pathways; H reflexes conditioned by tendon vibration therefore vary in size as a function of amount of presynaptic inhibition. Prior activity in motor neurons activates Renshaw cells; H reflexes obtained at certain intervals after such activity can therefore be used to evaluate inhibitory activity in recurrent pathways. Thus although EMG/NCSs are primarily tools for evaluation of the peripheral nervous system, they also provide information about central nervous system function.

See also: Cognition: An Overview of Neuroimaging Techniques; Cognitive Neuroscience: An Overview; Electroencephalography (EEG); Glutamate Regulation of Dendritic Spine Form and Function; Neuromuscular Junction (NMJ): Presynaptic Stretch Effects on Neuromuscular Transmission; Transcranial Magnetic Stimulation.

Further Reading

Aggarwal A and Nicholson G (2002) Detection of preclinical motor neuron loss in SOD1 mutation carriers using motor unit number estimation. *Journal of Neurology, Neurosurgery & Psychiatry* 73(2): 199–201.

Alisauskiene M, Truffert A, Vaiciene N, et al. (2005) Transcranial magnetic stimulation in clinical practice. *Medicina* 41(10): 813–824.

Armon C and Brandstater ME (1999) Motor unit number estimate-based rates of progression of ALS predict patient survival. *Muscle and Nerve* 22: 1571–1575.

Boe GS, Stashuk DW, and Doherty TJ (2004) Motor number estimation by decomposition-enhanced spike-triggered averaging: Control data, test–retest reliability and contractile level effects. *Muscle and Nerve* 29: 693–699.

Boe GS, Stashuk DW, and Doherty TJ (2006) Within-subject reliability of motor unit number estimates and quantitative motor unit analysis in a distal and proximal upper limb muscle. *Clinical Neurophysiology* 117: 596–603.

Bostock H, Cikurel K, and Burke D (1998) Threshold tracking techniques in the study of human peripheral nerve. *Muscle and Nerve* 21(2): 137–158.

Buchthal F and Rosenfalck A (1966) Evoked action potentials and conduction velocity in human sensory nerves. *Brain Research* 3: 1–122.

Burke D, Kiernan M, and Bostock H (2001) Excitability of human axons. *Clinical Neurophysiology* 112: 1575–1585.

Doherty TJ and Stashuk DW (2003) Decomposition-based quantitative electromyography: Methods and initial normative data in five muscles. *Muscle and Nerve* 28: 204–211.

Eisen A (2001) Clinical electrophysiology of the upper and lower motor neuron in amyotrophic lateral sclerosis. *Seminars in Neurology* 21(2): 141–154.

Finsterer J and Fuglsang-Frederiksen A (2001) Concentric-needle versus macro EMG. II Detection of neuromuscular disorders. *Clinical Neurophysiology* 5: 853–860.

Mogyoros I, Kiernan MC, and Burke D (1996) Strength–duration properties of human peripheral nerve. *Brain* 119(2): 439–447.

Olney R, Yuen E, and Engstrom J (1999) The rate of change in motor unit number estimates predicts survival in patients with amyotrophic lateral sclerosis. *Neurology* 52(supplement 2): A3.

Shefner JM (2001) Excitability testing in clinical neurophysiology – what, why and when? *Muscle and Nerve* 24: 845–847.

Shefner JM (2002) Peripheral sensory conduction. In: Brown WR, Bolton CP, and Aminott MJ (eds.) *Neuromuscular Function and Disease*, vol. 1, chapter 6. Philadelphia: WB Saunders.

Shefner JM and Gooch C (2003) Motor unit number estimation. *Physical Medicine and Rehabilitation Clinics of North America* 14: 243–260.

Shefner JM, Preston DC, and Logigian EL (1996) Activity-dependent conduction in single motor units. *Neurology* 46(5): 1387–1390.

Stalberg E (1980) Macro EMG, a new recording technique. *Journal of Neurology, Neurosurgery & Psychiatry* 43: 475–482.

Stalberg E (1990) Macro EMG. *Methods in Clinical Neurophysiology* 1: 1–14.

Vallbo AB, Hagbarth KE, and Wallin G (2004) Microneurography: How the technique developed and its role in the investigation of the sympathetic nervous system. *Journal of Applied Physiology* 96: 1262–1269.

Electronic Nose

J S Kauer and J White, Tufts University School of Medicine, Boston, MA, USA; and CogniScent, Inc., Weston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Relationship to Neuroscience

This article focuses on electronic noses as they relate to the neuroscience of olfaction, on which they are based. Less emphasis is placed on details of the differences among artificial or 'e-nose' systems, with more attention paid to what aspects of the olfactory pathway have been, and may in the future be, useful for enhancing e-nose function. As in development of all biomimetic engineering, the relationship between the biological system being mimicked and construction of the device which uses the biology model is a reciprocal one in which biology informs the engineering, while, the engineering process can be a useful heuristic tool to inform further analysis of the biology.

As discussed later, the concept of an artificial nose developed directly from analyses of the biological olfactory system. At the time of the first e-nose development and for many years before that, it was widely appreciated in the olfactory research community that monomolecular odors were encoded by processes widely distributed across the circuits of the pathway. Physiological studies had shown that the sensory elements in the nose, the second-order neurons in the circuits of the olfactory bulb, and the neurons of olfactory cortex routinely respond to ranges of chemically unrelated odorants, indicating that odors are not obviously encoded by neurons or circuits dedicated exclusively to single odorous compounds at any level of the pathway. Based on these observations, a number of hypotheses were formulated to explain how combinatorial and pattern recognition processes could be the basis for these findings.

Combinatorial, disseminated representation of specific odorants by relatively nonspecific sensing elements was not a concept that readily resonated with the classic chemical view of highly specific ligand/receptor interactions. Despite this, the data available at the time supported the prevailing view that the extraordinary sensitivity and discriminability widely recognized in animals such as dogs were, in fact, emergent processes that arose from activation of large populations of neuronal elements. The specific molecular signatures of the vapor phase stimuli were thus encoded in these populations by spatial and temporal activity that was then interpreted at various levels of the pathway to give rise to perception.

In 1982 Persaud and Dodd used these two ideas – activity patterns across arrays of relatively nonspecific primary sensors and pattern recognition algorithms for associating the response patterns with odorants – as the basis for the first artificial nose. These principles remain as guiding paradigms underlying virtually all modern e-nose devices. It is notable that Persaud and Dodd set out to implement these fundamentals of olfactory function before many of the biochemical mechanisms that underlie the primary sensory process were well understood.

Principles of Operation

Defining Odorant Stimuli

To understand how artificial noses use biological principles to capture odor information, the major steps in the volatile detection process need to be defined since these need to be captured by both natural and engineered systems (Figure 1).

The essence of all sensing systems is to convert different forms of sensory information in the external world (visual, auditory, mechanical, chemical) into information encoded by the biological world of neuronal circuits in the brain or by the electronic circuits of the biomimetic system. In the brain, this encoding process usually takes the form of transformation (transduction) of sensory information into trains of neuronal impulses propagated in the biological circuits of a neural pathway. In engineered devices, this takes place as conversion of odorant chemistry into electrical charges moving within analytical computer circuits.

For odor detection systems, the target information in the outside world is the vapor phase molecules arising from odorous substances volatilized to a degree dependent on their vapor pressures. The essence of the odorant encoding problem then is to convert the chemical and physical information about these gas phase molecules (structure, charge, size, molecular weight, polarity, solubility, volatility, concentration, and so on) into neuronal information that can be manipulated by brain circuits to ultimately give rise to odor perception and/or behavioral output. A single, volatile molecular species is usually referred to as an odorant (e.g., acetone, propionic acid, ethyl acetate, or methyl mercaptan), while volatiles perceived by an organism (usually a complex mixture of many odorants, for example, coffee, jasmine, orange, and chocolate) are

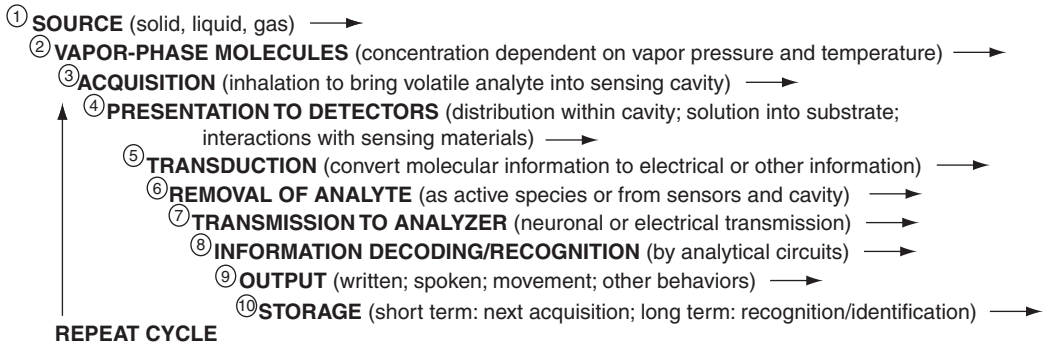

Figure 1 Steps in detecting vapor-phase compounds in the environment: encoding their molecular information by a detection system, recognizing the vapor-phase signatures, storing the results, and resetting the system for another detection event.

defined as an odor. An odor perception can, of course, also be generated by single, pure odorants (single compounds) such as acetone; thus, one can perceive the odor of chocolate or the odor of acetone.

Getting Odorant Stimuli to the Detectors

Vapor phase odorant detection requires that molecules arising from a volatile source come into physical contact with either the receptors of a biological olfactory system or detectors of an electronic nose. There is no distance detection in either biological or artificial olfaction – the source giving rise to the vapor may be distant, but vapor phase molecules from the source must reach the sensors. From a complex source such as coffee, odorant molecules are released into the air in different concentrations based on their individual vapor pressures and may arrive at the sensory detectors either by movement of the carrier air (wind) to the detector or by movement of the detector system through the air. Odorant plumes are invariably heterogeneous, chaotic, turbulent, and temporally variable, and therefore their physical structure is impossible to define completely. Insects and certain other animals such as sharks rely on their own movement through the medium in which they live to bring odorants to their sensing elements, whereas most air-breathing animals have active inhalation (sniffing) mechanisms to bring odorants to receptors that are often sequestered in complex nasal cavities that provide another aerodynamic dimension to the odorant delivery process.

For both biological and engineered odor-sensing systems, this access process and the subsequent, usually underappreciated, removal process are both crucial to the effective function of the systems. An additional important component of the access process is the relationship between the complex turbulent nature of odorant plumes and the physical structure of the sensory surfaces with which the plumes interact. The sensing surfaces in biological systems are usually physically complex structures consisting of amplified surfaces (cilia or microvilli) and elaborate turbinate folds, whereas sensing surfaces in engineered devices have, so far, been virtually flat and uniform. How to best deploy sensors on surfaces in order to optimally capture turbulent odorant plumes is still poorly understood and not usually a rationally designed feature of engineered devices.

Converting Chemical Information into Electrical Information

Coding paradigms Over the past 20 years, much has been learned about the biochemical cascades and receptors underlying the initial transduction events in biological olfactory systems. For example, many details about receptors, downstream amplification mechanisms, and changes in membrane conductance triggered by odor binding have been elucidated. Precise descriptions of how specific structural odorant features interact with the receptors have just begun to evolve. Relatively little is known about the binding constants of odorants to various receptors, and as yet there is no information on how receptor structure revealed by X-ray crystallography relates to odorant binding.

What is firmly established is that biological olfactory systems, in general, do not encode odor information by activation of specific neurons by specific odorants. Thus, the system does not operate by encoding odorant 'A' with cell type 'A' and does not appear to function by having so-called 'labeled lines,' one for each odorant of interest to the organism. Rather, with the exception of some apparently highly specific pheromone pathways in insects and certain vertebrate olfactory neurons, both first order sensory receptor cells in the nose and neurons in the higher order brain circuits tend to respond to variable, broad ranges of odorants, leading to an encoding process (even for

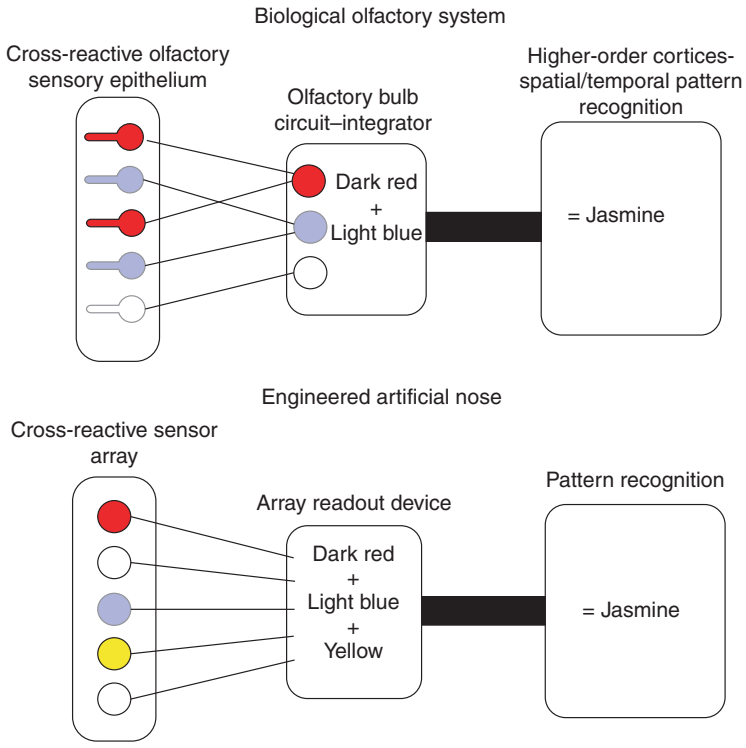

Figure 2 Schematic diagrams of the major components of biological and artificial olfactory systems: cross-reactive receptors or sensors, integration levels that receive the spatially and temporally distributed outputs from the sensor layer, and a pattern recognition level that categorizes, compares with stored information, and ultimately carries out identification.

monomolecular compounds) that is highly distributed at each level of the pathway (see Figure 2). It is the essence of these distributed events that is captured by artificial nose devices.

E-nose sensing mechanisms At the first level of e-nose design are sensing materials that interact with multiple vapor phase odorants. These interactions show moderate, overlapping specificities, thereby serving as cross-reactive sensors. The e-nose interrogates the sensing materials to read out the degree to which they interact with each odorant material and then categorizes and analyzes their response patterns.

In most e-nose systems the sensing materials are some form of hydrocarbon polymers either chosen from catalogues of commercially available polymeric materials, explicitly synthesized for detecting specified vapors, or made from DNA polymers composed of deoxynucleotide bases. E-nose sensors also are made from thin tin oxide films impregnated with various precious metals and from chemically modified metalloporphyrins. The degree to which polymer-, tin oxide-, or metalloporphyrin-based sensors interact with odorant compounds depends on their relative affinity for the vapor, which is in turn dependent on solvation, polarity, charge affinity, and other chemical and physical properties of the sensing materials. The exact mechanisms that impart the responsivity to different odorant vapors are not usually understood to the degree that one can predict *a priori* how a particular material will respond to a particular odorant vapor. Thus, selections tend to be empirical, operating on some general principles, although some highly specific detectors have been rationally constructed. Polymer responses are read out using a variety of methods, including measurement of changes in conductivity, mass (surface acoustic wave, quartz resonance, and microbalance), or optical properties (colorimetric, fluorescence).

Figure 3 illustrates responses from polymers that change their conductivity (left panels), an array of metalloporphyrins that change color (right panels), and arrays of other polymers labeled with dyes that show changes in fluorescence (bottom panels) upon odorant exposure. In most cases, the response arrays are allowed to come to equilibrium with vapor exposure (Figure 3, left and right panels), while in systems with rapidly responding sensors, the time course as well as differential sensor response can be used as part of the encoding mechanism (Figure 3, bottom panels, z-dimension going into the page). Tin oxide responses consist of changes in conductivity dependent on the affinity of the analyte vapor for the tin oxide film. See Table 1.

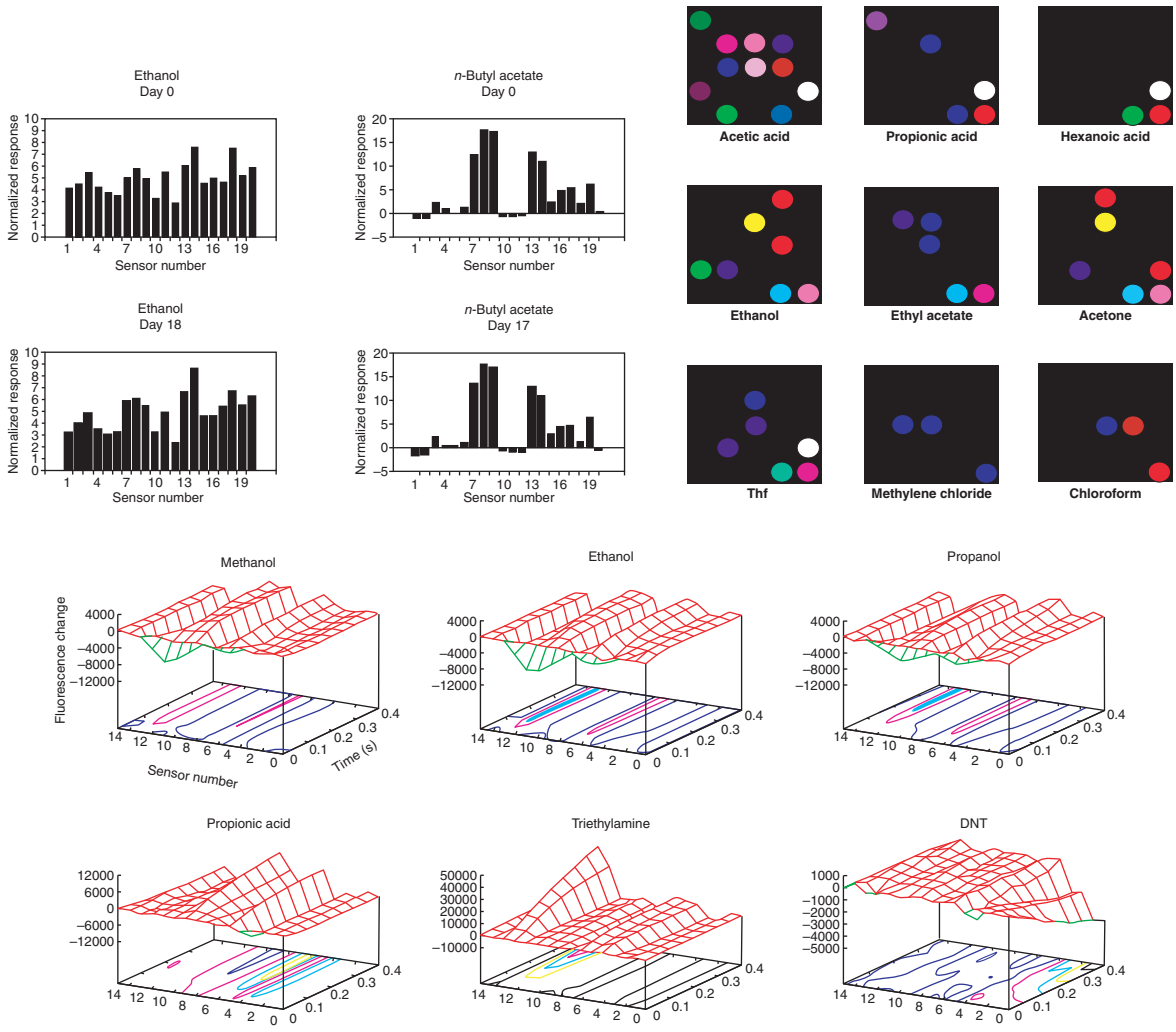

Figure 3 Responses from three types of sensing arrays. (Top left) Changes in conductivity from 20 individual polymer-based sensors (Osmetech: Persaud). (Top right) Colorimetric changes in 25 individual metalloporphyrin-based sensors (ChemSensing: Suslick). (Bottom) Changes in fluorescence (*y*-axis) over time (*z*-axis) from 16 individual polymer-based sensors (*x*-axis) (CogniScent: White/Kauer).

Although they are sometimes referred to as artificial noses, devices with single detectors that rely on separation of vapor phase molecules by affinity along a column (rapid gas chromatographs) or in a chamber by charge (rapid ion mobility spectrometers) are not considered here because they do not use the principles of cross-reactive sensor arrays and distributed pattern recognition based on the biological principles described earlier.

Pattern Analysis Leading to Odor Recognition

The second essential element of e-nose design is analytical. Algorithms have been developed to associate the patterned output arising from the sensor arrays with the applied odorous analytes. In the biological olfactory system, the circuits of the olfactory bulb (shown in Figure 4, top) carry out the first level of analysis of the highly parallelized input from the olfactory receptor neurons in the nasal cavity, and the output of the bulb is transmitted to a number of higher olfactory circuits such as the piriform cortex. Although little is known about how olfactory cortex works in detail, it is presumed that pattern recognition, comparison with stored memories, and subsequent storage of new information reside at this level.

In artificial devices, pattern recognition has been implemented in a number of different algorithms. In most cases recognition is carried out by a process that compares the input pattern from the test analyte with patterns that have been stored in computer memory during 'training' sessions. Training usually takes the form of generating sensor array responses to known odorants and storing the array responses in memory. Input patterns generated by unknown test odorants

Table 1 Examples of sensing materials used in artificial noses

Sensing material	*Deployment*	*Readout*
Tin oxide	SnO_2 films doped with precious metals, run at 300–400 °C	Conductivity; works like n-type semiconductor
Polymers	Onto surfaces appropriate for readouts below	• Conductivity detected by resistance measurement • Swelling detected optically or by resistance measurement • Attached to fluorophores, change in fluorescence • Changes in mass due to differential adsorption, detected by surface acoustic wave, quartz resonance, or microbalance
Metalloporphyrins	Colored spots onto paper substrates	Colorimetric

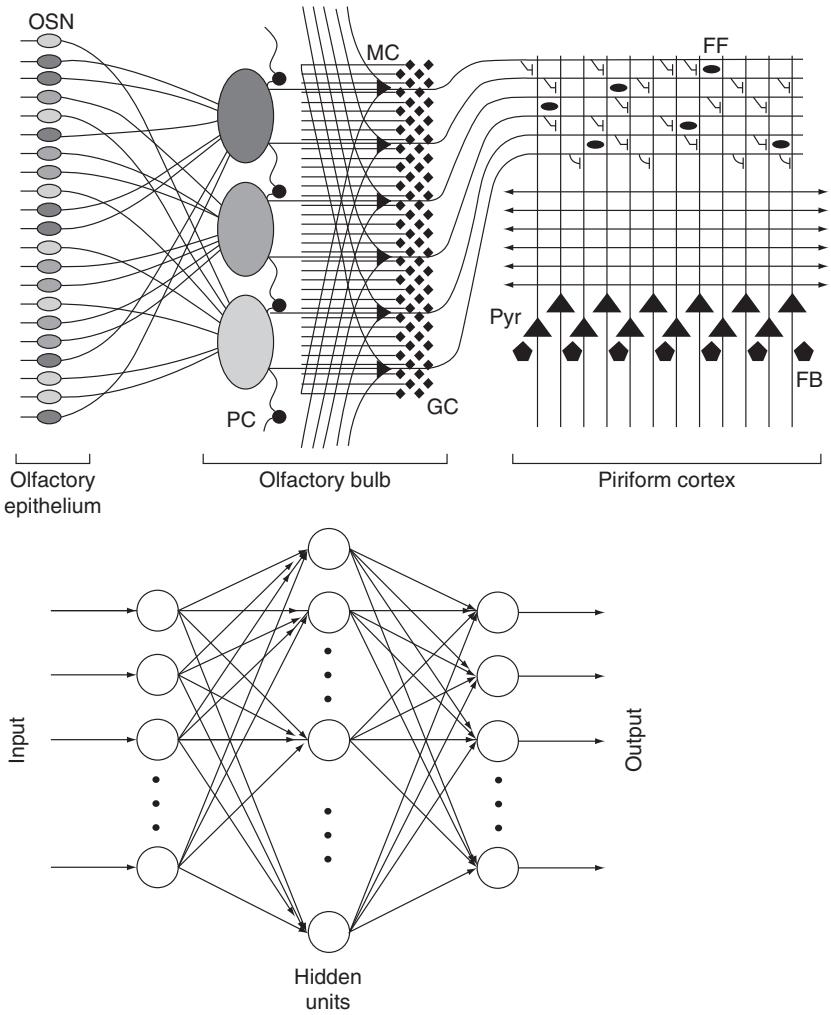

Figure 4 (Top) Integrative circuits of the mammalian olfactory system including olfactory receptors and bulbar receiving convergent and divergent inputs from receptor neurons in the nasal cavity (White and Kauer). (Bottom) An example of a simple neural network model (Rummelhart) with convergent and divergent connections among the layers of a type that has been used for analysis of responses from artificial nose detector arrays.

are then compared with the stored data using a variety of methods, including neural networks (Figure 4, bottom), principal-component analysis, self-organizing maps, and simple Euclidian distance evaluations. In principle these approaches work under many conditions, although identification can be compromised by the difficulties of differences in training/test analyte concentrations, the presence of interferent odorants that were not included during training sessions, and changes in temperature and relative humidity.

Utility for Solving Real-World Problems

The goal of artificial nose development is to exploit the intrinsic flexibility and broad responsivity of arrays of cross-reactive sensing materials deployed in arrays to detect as many different vapor phase analytes in as many different interferent backgrounds as possible. Theoretically, the combinatorial power of such sensing devices scales exponentially with the number of sensors and offers the potential for detecting molecules arising from virtually any substance with some level of vapor pressure. One can imagine perfected devices being able to detect, quantify, and identify single odorants and complex mixtures associated with medical diagnostic problems, environmental monitoring, food and beverage quality control, analytical chemistry associated with a myriad of industries, and so on. Whether the promise of detection of vast numbers of different vapor-phase species will be realized still remains to be seen. Based on theoretical grounds it would appear that combinatorial systems are the only way to go.

Strengths and Weaknesses of Artificial Noses

The strengths of artificial noses are as follows:

- simple
- potentially broad spectrum sensitivity with small numbers of sensors
- relatively inexpensive
- small, portable, low power
- flexible, trainable, improvements in detection sensitivity are under continuous development and can be enhanced by both engineering solutions (e.g., by increasing sensor surface area) and by algorithmic development.
- with explicit control over sensor chemistry, there is the potential for modifying sensor arrays in order to tailor the devices to defined problems.

The weaknesses of e-noses are the following:

- often responsive to humidity and temperature
- must be calibrated for stability
- rational ways to select sensing materials for particular target volatiles not yet well defined
- algorithms and other aspects of device design need to be fitted to the target detection task.

The Future

E-noses are only about 25 years old, and there are a number of areas in which rapidly occurring changes in technology are likely to improve functionality. Miniaturization of electronics and the increasing computational power of small circuits will be further exploited in the evolution of e-nose analytical circuitry, as will sensing elements and circuits that take advantage of nanotechnology.

Finally, artificial olfactory technology has not yet successfully entered the consumer marketplace. One can imagine that a more mature e-nose technology, improvements in stability, miniaturization, and lower costs would lead to useful devices for detecting vapor phase compounds related to many diverse aspects of human life. These might include, for example, an emergence of digitally based olfactory 'assistants' just as we have digitally based 'assistants' for scheduling (Palm Pilots), communication (cell phones), imaging (digital cameras), and acoustics (iPods), thus opening up an awareness of the unappreciated richness of the odor universe that envelops us.

See also: Olfactory Neuron Patterning and Specification; Olfactory Coding; Olfactory Receptors; Olfactory System Theory.

Further Reading

Alkasab TK, White J, and Kauer JS (2002) A computational system for simulating and analyzing arrays of biological and artificial chemical sensors. *Chemical Senses* 27: 261–275.

Buck LB (1996) Information coding in the vertebrate olfactory system. *Annual Review of Neuroscience* 19: 517–544.

Freund MS and Lewis NS (1995) A chemically diverse conducting polymer-based "electronic nose". *Proceedings of the National Academy of Sciences of the United States of America* 92: 2652–2656.

Gardner JW and Bartlett PN (eds.) (1999) *Electronic Noses: Principles and Applications*. Oxford, UK: Oxford University Press.

Hildebrand JG and Shepherd GM (1997) Mechanisms of Olfactory discrimination: Converging evidence for common principles across phyla. *Annual Review of Neuroscience* 20: 595–631.

Kauer JS (2002) On the scents of smell in the salamander. *Nature* 416: 20–26.

Kress-Rogers E (ed.) (1997) *Handbook of Biosensors and Electronic Noses: Medicine, Food, and the Environment*. Boca Raton, FL: CRC Press.

Montag S, Frank M, Ulmer H, Wernet D, Gopel W, and Rammensee H-G (2001) "Electronic nose" detects major histocompatibility complex-dependent prerenal and postrenal odor components. *Proceedings of the National Academy of Sciences of the United States of America* 98: 9249–9254.

Pearce TC (1997) Computational parallels between the biological olfactory pathway and its analogue 'The Electronic Nose': Parts I & II. *BioSystems* 41: 43–90.

Persaud K and Dodd G (1982) Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature* 299: 352–355.

Rakow NA and Suslick KS (2000) A colorimetric sensor array for odour visualization. *Nature* 406: 710–714.

Rumelhart DE, Hinton GE, and Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE and McClelland JL (eds.) *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, pp. 318–362. Cambridge, MA: MIT Press.

Schild D (ed.) (1990) *Chemosensory Information Processing*. Berlin: Springer.

Settles GS (2005) Sniffers: Fluid dynamic sampling for olfactory trace detection in nature and homeland security – The 2004 Freeman Scholar Lecture. *Journal of Fluids Engineering* 127: 189–218.

White J, Mall S, and Kauer JS (2001) Using biology to guide development of an artificial olfactory system. In: Ayers J, Davis J, and Rudolph A (eds.) *Neurotechnology for Biomimetic Robots*, pp. 97–114. Cambridge, MA: MIT Press.

White J, Williams LB, AtKisson MS, and Kauer JS (2004) DNA-based fluorescent chemosensors for direct detection of volatile compounds in an artificial nose. *Association for Chemoreception Sciences Abstract*.

Yang J-S and Swager TM (1998) Fluorescent porous polymer films as TNT chemosensors: Electronic and structural effects. *Journal of the American Chemical Society* 120: 11864–11873.

Electrophysiology: EEG and ERP Analysis

S Makeig, University of California at San Diego La Jolla, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Although studies of event-related electroencephalograph (EEG) dynamics were the first modalities to image brain activity linked to the cognitive interpretation of stimulus and other events, from the first availability of averaging computers in the early 1960s, the field was dominated by averaging of epoch time locked to some set of events considered to be equivalent, generating what was first known as the (stimulus) evoked potential, later the event-related potential (ERP), and still later the equivalently derived magnetoencephalograph (MEG) 'event-related field' time course. Currently, a wide array of techniques from spectral and time-frequency analysis, information-based signal processing, and machine learning are being introduced. Applied to currently available high-density scalp recordings of up to 256 or more channels, the study of event-related brain dynamics reveals much more about how EEG and MEG source signals, representing partial synchrony of local cortical field activities, are linked to the experience and control of behavior.

Strengths and Weaknesses of Response Averaging

Any useful method of analysis of complex signals involves some set of assumptions about the generation of the data that, if fulfilled, makes the results of the analysis truly interpretable. For response averaging of EEG (or similarly of MEG) data epochs, these assumptions are that (1) the exact times of occurrence of a set of time-locking events are known, (2) that these events produce essentially equivalent EEG responses, (3) that these responses each consist of a fixed response time course that is added linearly to the ongoing EEG, and (4) that the ongoing EEG is wholly unrelated to and unperturbed by the time-locking events. In the case that these assumptions are true, then the averaged response (ERP) reveals the approximate time course of activity added to the EEG signal time locked to the time-locking events of interest with a signal-to-noise ratio inversely proportional to the square root of the number of epochs averaged.

Difficulties arise when, as in this case, the data to which the analysis method is applied do not match the assumptions under which the results of the analysis can be truly interpreted. In the case of EEG response averaging, it is undeniably the case that the ongoing EEG may be and is typically affected by sensory and other events, contrary to assumption 4 above. This was amply demonstrated by Pfurtscheller and colleagues as early as 1979 in the form of reliable event-related changes in the strength of prominent ongoing 8–12 Hz alpha band rhythms recorded over posterior and central cortex. These changes, termed event-related synchronizations (ERSs; alpha power increases) and event-related desynchronizations (ERDs; alpha power decreases), typically did not appear in ERP averages of many similar trials. This is because, contrary to ERP assumption 3 above, they resulted from a multiplicative modulation of ongoing EEG activity, not the addition of a fixed response time series to the unchanged, ongoing EEG.

From the perspective of time–frequency analysis, response averaging retains only those portions of the event time-locked data epochs that are also reliably phase locked. This means that when a brief sinusoidal signal at a frequency of interest is fit to the single trials at a event-locked latency of interest, the phase values of the best-fitting sinusoids for each of the trials are not randomly distributed around the phase circle. To create a nonzero ERP signal, this nonrandom phase distribution requirement is necessary and nearly sufficient (except in the case that the nonrandom phase distribution is radially symmetric).

At zero frequency or DC, the concept of data phase collapses to the sign of the data, positive or negative, with respect to some baseline value. Thus, for example, if prominent signal peaks 300 ms following the sudden, unexpected but awaited occurrence of rare tones against a silent or contrasting sonic background were not all positive, but instead half positive and half negative, the resulting ERP average 300 ms following tone onsets would be zero, even though each trial contained a large potential fluctuation undeniably evoked or induced by the tone presentations. In practice, this does not happen. Rather, a predominantly positive peak, often termed P300, is produced in ERP averages over, most prominently, the central posterior scalp. However, from the ERP average, it is not possible to tell whether the event-related positivity occurred in every trial with comparable form and amplitude or whether the ERP feature was produced by a slight positive-favored imbalance in the numbers of much larger positive and negative peaks at 300 ms latency in the single trials. In practice, in such experiments the P300 positivity is recorded with highly variable amplitudes in only about 80% of trials, calling into question any interpretation, from the average

ERP record, that the P300 feature is an invariant consequence of these tone occurrences.

Lack of phase consistency explains why even strong alpha band ERD and ERS phenomena may produce no trace in the average ERP. In these cases, the multiplicative modulation of ongoing activity is independent of the phase (e.g., positive-going or negative-going) of the alpha activity 300 ms after the tone onset. Therefore, averaging a large enough set of epochs in which strong and consistent blocking or augmenting of alpha band activity takes place at a fixed latency with respect to the time-locking events may reveal no alpha activity at all, either before or after stimulus onsets, since at neither latency is the alpha activity phase distribution unbalanced. To make an analogy, say someone set up a motion camera at the beach and recorded an hour of film images and then averaged them. The average image would not reveal any waves but instead only a continuous blur near water's edge. Here, interpretation of the mean photo could be quite inaccurate, as even brief viewing of the original footage would reveal.

Event-Related Phase Resetting of Ongoing Brain Activity

In some continuously rhythmic systems, such as circadian cycles in body temperature and the cardiac rhythm, infrequently and briefly pausing the rhythm, say by an aptly timed burst of bright light or an electric pulse, is observed to first stop the ongoing rhythm briefly and then to allow its restart at a fixed phase of its cycle a fixed latency following the interrupting event. These activities, once resumed, then continue at their original frequency and amplitude. Consider, in these cases, the appearance of the stimulus time-locked event-related averages. In both cases, they would show a flat prestimulus baseline period, followed, at some time after the interrupting events, by an appearance of the rhythm, which would continue for some time. Again, a simple interpretation of the event-related average as representing the addition of a rhythmic activity to an otherwise flat record would be quite an error. In fact, no increase in the amplitude of the rhythmic activity occurs, quite contrary to the impression delivered by the average ERP.

In these cases, the term 'phase resetting' to a reliable (re)starting value has been used. Can it apply to event-related EEG dynamics? The answer to this question is not simple or uniform. The primary complication comes from the temporal instability or nonstationarity of EEG activity, either recorded from the scalp or, in the form of the electrocorticogram, from the cortical surface directly. If, for example, an arousing stimulus is delivered to an EEG subject only at infrequent moments when the subject's EEG at a channel of interest contains a prominent and steady ongoing alpha band rhythm, then an average ERP containing a flat baseline and a prominent poststimulus alpha band rhythm may indeed be recorded, and if so, this phenomenon might be said to represent phase resetting similar to the biological examples above.

More commonly, however, the delivery of stimuli in event-related experiments is not dependent on the occurrence of any EEG feature in particular, and even at posterior scalp electrodes, the alpha rhythm may or may not be observable in single trials. Here, a consistent change in average alpha band power in the single-trial signals, time-locked to the arousing events, may not genuinely represent a phase resetting, since in many trials there is actually no definable alpha phase to reset! Rather, in such a case the appearance of an ERD or ERS feature may represent only a change in the relative propensity of the revelant cortical regions to synchronize local alpha band activities so as to produce alpha band EEG signals on the scalp. The average ERD and ERS measures proposed by Pfurtscheller are also averages, as is their generalization to a whole-spectrum measure of average power at a set of event latencies, proposed later by Makeig and termed the event-related spectral perturbation (ERSP) or spectral power time/frequency image. Spectral power averages, like time–domain ERP averages, fail to accurately distinguish between different types of temporally unstable phenomena in the data and therefore are highly susceptible to inaccurate interpretation.

In the case of ERP averages, this susceptibility to misinterpretation comes into play not only for fairly rare examples of true phase resetting (e.g., in cases dubbed 'alpha ringing'). A more common and also serious misinterpretation of ERP features comes in cases when brief event-related bursts of quasi-rhythmic activity occur reliably time locked but only partially phase locked to events. In this case, the shape of the remnant ERP feature may not well represent the bursts that actually occur in single trials, depending on the exact distribution of relative phases and amplitudes of the phenomenon across the single trials. This is a more serious problem for interpretation since the average ERP in this case may give an erroneous impression that a fixed series of indistinct peak events occur in each trial, whereas in reality these may be but the shadow of much more robust dynamics occurring with variable phase and amplitude in the single trials. This instability violates ERP assumption 2 above.

A more serious form of violation of assumption 2 (temporal stability of the response) comes when the

brain shows quite different modes of responding to instances of the 'same' class of stimuli, for instance, theta bursts versus high-frequency activity. Here, all the event-related averages (ERP, ERD/ERS, and ERSP) fail to reveal the form of the nonstationarity.

Spatial Mixing of Brain and Nonbrain Potentials and Field Activities

Another source of confusion in the interpretation of event-related EEG (and MEG) data is the fact that scalp electrodes (or magnetic sensors) sum field activities reaching them through electrical volume conduction (or radial flux). This summation is linear, and the propagation is near instantaneous. But the degree of spatial mixing is quite strong, leading to the widely accepted opinion that EEG (and MEG) data are noisy or contain relatively little spatial information. In particular, a positive or negative value observed at a certain scalp electrode (with respect to the reference electrode or electrodes used in the recording) does not mean that the observed activity represents activity in cortex immediately below the electrode. Simple biophysical examples of such cases are easy to construct.

Continuous Experience versus Tachistoscopic Stimulus Presentation

Finally, assumption 1 above, that the exact times of relevant events are known to the experimenter, limits analysis to experiments in which events of interest can be known precisely. Typically, this has meant that event-related EEG and MEG experiments use unnatural tachistoscopic event presentations and do not study the more natural order of events which develop over time with no instantaneous onset or time of recognition by the subject.

New Approaches to Measurement and Modeling of Event-Related Brain Dynamics

Currently, a wealth of new approaches to the analysis of complex, high-dimensional data has been and is continuing to be developed in the fields of signal processing and machine learning or artificial intelligence. The ERD/ERS and ERSP measures are simple examples of application of the quite large, modern field of time–frequency analysis to EEG or MEG data. Of increasing importance is the attempt to directly locate and model sources of information in complex data rather than focusing on characterizing only its mean and most prominent directions of overall variance. Information-based signal processing, typified by the development over the last two decades of independent component analysis (ICA), is beginning to touch every aspect of signal processing, including analysis of event-related EEG and MEG dynamics.

Applied directly to the recorded multichannel data under favorable circumstances, ICA produces maximally independent signal components that in many cases account neatly for activity from either (a) a particular type of nonbrain artifact (eye blinks or eye movements, activity produced by the heart or a particular scalp muscle, line noise, electrode movement, etc.) or (b) synchronous local field activity within a single, oriented patch of cortex. In some cases, the pattern of projection of an independent component resembles the projection of synchronous activity at two noncontiguous cortical patches, but in this case the model patches appear to be located in bilateral symmetry with respect to the brain midline, compatible with locally synchronous field activities synchronized between patches by dense bidirectional white matter connections, possibly in the corpus callosum.

Thus, ICA provides a method for spatially filtering activity recorded at many EEG or MEG channels into the overlapping projections of brain and nonbrain sources. Although ICA does not itself identify the natures or exact locations of the separated sources, it can increase significantly their signal-to-noise ratio in single trials or the continuous data record, allowing application of many types of analysis that do not require that the brain's response to events be highly determined or similar in response to superficially similar events.

An alternative method of separating EEG or MEG source signals, termed beamforming, relies on having an accurate model of electrical conduction or magnetic flux from any source area in the head to the scalp sensors. As such a head model is difficult to obtain, and more difficult or impossible to make accurate, and the number of predetermined beamforming source locations is generally limited to the number of sensors, beamforming is not necessarily preferable to ICA, which does not require a head model and determines the nature, and ultimately the locations, of the sources (both brain and nonbrain) solely from the higher-order statistics or information content of the data.

However, spatial filtering or separation is best viewed only as a preprocessing step for event-related EEG or MEG data. Once separated, both the consistency and the variability of the individual source signals across event-related trials are to be determined and modeled, and during this analysis all forms of signal nonstationarity should be considered,

tested, and when present, characterized. In addition, EEG dynamics should not be thought of as occurring completely independently in different brain (or nonbrain) source areas, even when the putative source signals are determined to be maximally independent by some ICA definition. Rather, coordinated dynamics in some or many sources should be checked and modeled since they may reveal the action of distributed brain systems to optimize experience and behavior – the goal of the brain as a whole.

See also: Electroencephalography (EEG); Event-Related Potentials (ERPs) and Cognitive Processing; Event-Related Potentials (ERPs); Executive Function and Higher-Order Cognition: EEG Studies; Single Cell PCR Coupled with Electrophysiology.

Electroretinography

L J Frishman, University of Houston, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The electroretinogram (ERG) is an electrical potential generated by the retina of the eye in response to light. It can be recorded noninvasively, using electrodes placed on the cornea, and it reflects the summed response of all the cells in the retina. The ERG arises from extracellular currents that flow in the retinal tissue as a result of neuronal signaling and in some cases spatial buffering of potassium by glial cells. With appropriately applied stimuli and analyses, the ERG can provide useful, objective information about the functional integrity of the cells and retinal circuits that contribute to its generation. This information can be used in diagnosing and defining retinal diseases in the clinic, in monitoring the effectiveness of neuroprotective treatments and gene therapies in human patients and animal models, and in studying normal aspects of retinal processing and altered function in genetically manipulated retinas.

Recording ERGs

ERG recordings can be made from the corneal surface, *in vivo*, from intact humans or other animals under physiological or nearly physiological conditions (e.g., anesthesia). A wide variety of electrodes have been used to record ERGs: contact lenses with conductive metal electrodes embedded in them, thin fibers with silver particles, or gold foil electrodes, for instance. For rodent recordings, many labs use metal wires, often in a loop that rests on the corneal surface that is kept moist with a lubricating conductive solution. The reference electrode can be positioned under the eyelid, as the speculum of a contact lens electrode (Burian-Allen is a common type), or remotely from the eye, on the temple, for example. The signal picked up by the electrode is amplified using conventional electrophysiological amplifiers, preferably with low noise because ERG signals in mammals are small, ranging from a microvolt or less up to a millivolt or more, peak to peak, for responses to high-intensity stimuli. Amplifiers and/or digital filter settings in the software are important so that relevant signals are not lost but low-frequency drift, line noise, and other noise out of the frequency range of the signal are eliminated.

Full-Field (Ganzfeld) Flash ERGs

The most common stimulus used in ERG studies is a brief (< 10 ms) flash of light that uniformly fills the entire visual field, called a full-field, or ganzfeld, flash ERG. The use of a uniformly illuminated field prevents differential local adaptation of various retinal elements that can contaminate the response of interest. In dark-adapted eyes, the full-field ERG is generated by the rod photoreceptors and the circuits that transmit rod (scotopic) signals through the retina. Figure 1 shows a rod-dominated response in humans to a moderately intense light (Figure 1(a)) and to a light 100 times brighter (Figure 1(b)). As marked in the figure, the initial negative-going wave of the ERG, occurring in the first few milliseconds after light onset for the brighter stimulus, is designated as the a-wave, and the large positive potential that follows it is called the b-wave. The small wavelets superimposed on the b-wave are called oscillatory potentials (OPs).

The rod-driven a-wave primarily reflects the rod photocurrent, and it can be used to study photoreceptor function and the effects of disease or transgenic manipulations on the response. A model of phototransduction can be fit to the leading edge of the a-wave to assess photoreceptor function (rod or cone – see Figure 2), and the entire time course of the photoreceptor response can be mapped by means of a paired flash approach. At late times in the leading edge, as shown in Figure 2, the a-wave includes contributions from postreceptoral cells in the retinal off pathways. The presence of the postreceptoral contributions was established in experiments in macaque monkeys, whose retinas (and ERGs) are very similar to those of humans, using intravitreally injected pharmacological agents such as glutamate analogs and antagonists to block specific retinal circuits. The primate scotopic b-wave, as well as that of other mammals (e.g., mice), is generated largely by the light-driven activity of the second cell in the primary rod system, the 'rod bipolar cell.' Rod signals that have entered cones via gap junctions and travel via cone bipolar cells to inner retina contribute to the response as stimulus intensity and background light are increased. The origin of the OPs has been traced to inner retina, primarily amacrine cells (although high-frequency OPs in the cone-driven ERG depend on integrity of ganglion cells, at least in primates).

To study the cone-driven, photopic flash ERG, the rods must be suppressed or unable to respond by means of the stimuli selected. Rods can be rendered unresponsive by imposition of a background illumination sufficient to saturate the rod response

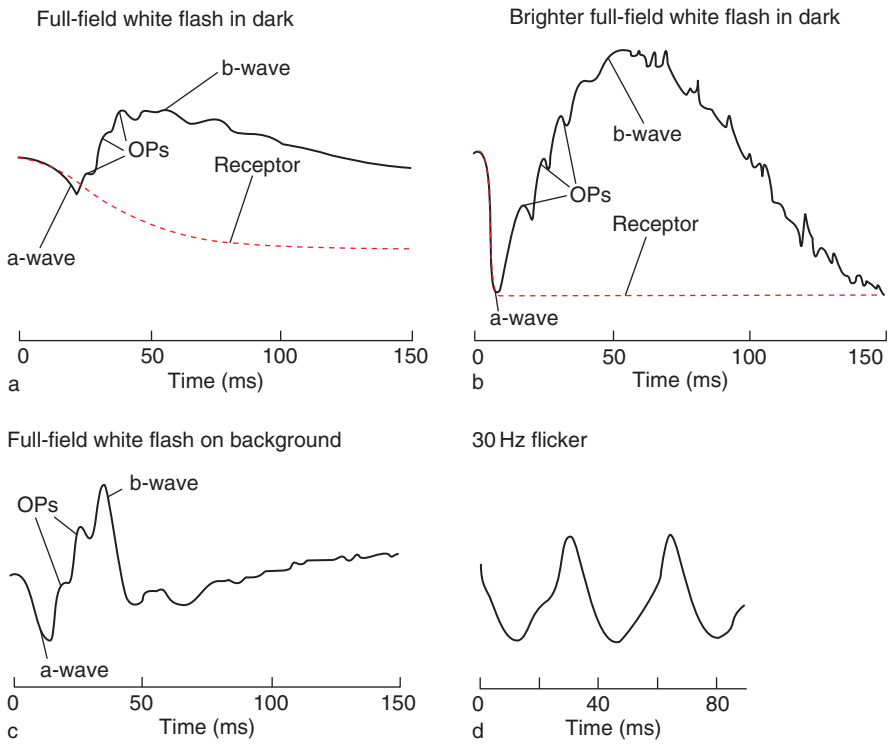

Figure 1 Full-field electroretinograms (ERGs). (a) Dark-adapted ERG recorded in response to a brief test flash of about 2 log scotopic trolands seconds (scot td-sec). The dashed curve is the response of the rod photoreceptors. (b) Dark-adapted ERG recorded in response to a test flash of about 4 log scot td-sec. The dashed curve predicts the response of the rod photoreceptors. (c) ERG recorded in response to a test flash of about 2.3 log photopic td-sec presented on a background of about 30 cd m^{-2}. (d) ERG recorded in response to 30 Hz flicker. From Hood DC (2004) Electroretinogram, electrophysiology. In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience*, 3rd edn. Amsterdam: Elsevier.

(i.e., >25 cd m^{-2}). A photopic ERG in human eyes is shown in Figure 1(c). Pharmacologic blockade studies in macaque retina have shown that the cone-driven a-wave is a mix of cone receptor and postreceptor responses from the off pathway. In contrast, the photopic b-wave is mostly the sum of responses from on-cone bipolar cells, but it is modulated by contributions from off-cone bipolar cells, off inner retinal circuits, and horizontal cell feedback to cone photoreceptors. When long-duration (e.g., 150–200 ms) stimuli are used, a positive response at light offset, called the d-wave, emerges in most mammalian retinas, although not in mice and rats. The d-wave reflects off-cone bipolar cell responses, added to the positive-going offset of the cone photoreceptors. Cone-driven responses also can be isolated from rod responses by the use of flicker presented at temporal frequencies that are too high for rods to resolve, that is, 30 Hz in humans, as illustrated in Figure 1(d). Although this flicker frequency can be used in primates to separate rod- and cone-driven responses, the flicker approach is less useful in mice, whose frequency response ranges for the different photoreceptor types and their respective retinal circuits overlap substantially.

Rod and cone signals can also be separated spectrally. In primates the long-wavelength L-cone can be stimulated with long-wavelength light presented at an intensity that hardly affects the less sensitive rods. However, mice and rats do not have an L-cone, and the spectral sensitivity of their medium-wavelength M-cone overlaps with that of rods. In these animals, it is possible to selectively stimulate the short-wavelength UV cones.

Amacrine and ganglion cells in the inner retina do not contribute as strongly to the ERG as the bipolar cells, but some signals specific to these cells can be identified. For example, in the light-adapted ERG in macaques and humans (and cats), a slow negative wave after the b-wave in the light-adapted ERG called the photopic negative response (PhNR), reflects ganglion cell activity. In rodents, with relatively fewer ganglion cells, it is likely that this same wave is generated by amacrine cells. In the dark-adapted ERG, the most sensitive responses, near absolute psychophysical threshold, named the scotopic threshold responses (STRs), also arise from amacrine and/or ganglion cells, depending on the species. In macaques, a negative-going STR wave following the b-wave is

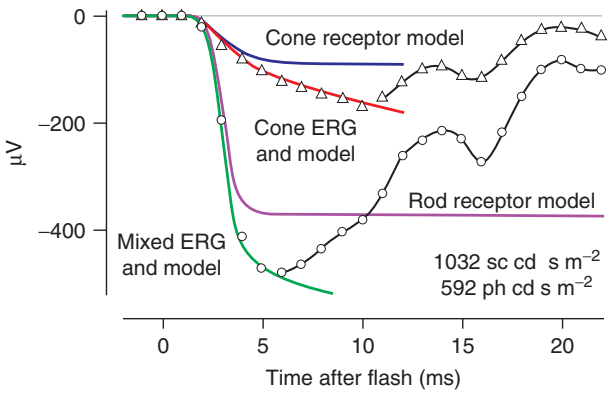

Figure 2 Full-field ERG from a macaque monkey in response to a high-energy brief flash. The data points show the mixed rod–cone ERG (circles), the small cone-driven ERG obtained when rods were suppressed (triangles). The purple and dark blue lines plot models described in Robson JG, Saszik SM, Ahmed J, and Frishman LJ (2003) Rod and cone contributions to the *a*-wave of the electroretinogram of the dark-adapted macaque. *Journal of Physiology* 547: 509–530 for rod and cone photoreceptor responses. The red line shows the model responses for the cone-driven ERG (photoreceptor plus postreceptoral off-pathway contributions), the green line shows the mixed rod-cone ERG (modeled rod photoreceptor response plus modeled cone-driven ERG). Later portions of the response that were not modeled are connected by black lines. Oscillatory postreceptoral contributions are obvious in the late portions of the responses. sc, scotopic; ph, photopic. Adapted from Robson JG, Saszik SM, Ahmed J, and Frishman LJ (2003) Rod and cone contributions to the *a*-wave of the electroretinogram of the dark-adapted macaque. *Journal of Physiology* 547: 509–530.

generated at least in part by retinal ganglion cells, whereas in rodents, a positive-going STR at the time course of the b-wave appears to be generated by retinal ganglion cells, and the negative-going STR primarily by AII amacrine cells.

Pattern ERG

A more common approach to assess the ganglion cell response is to record a pattern ERG (PERG). The stimulus for the PERG stimulus is a contrast-reversing checkerboard or grating pattern for which the mean luminance remains constant. The PERG response represents, in large part in humans, macaques, and mice as well, a response generated by the retinal ganglion cells. A drawback of the PERG is that the response is quite small, only a few microvolts, and for primates, good optical focus is desirable.

Multifocal ERGs

The development of the multifocal ERG has provided a technique for recording localized ERG responses from many small retinal regions simultaneously. Previously, study of local retinal responses was limited because of the low amplitude of focal ERG responses elicited by small stimuli and the time-consuming nature of the task, in which only one small region could be stimulated at a time. Thus the multifocal stimulus is a valuable tool for topographical analyses of normal and abnormal retinas and central visual pathways. For human and macaque studies, as shown in **Figure 3**, the stimulus display often contains 103 hexagons spread over the central 50° or so of the retina. The stimulus presentation is determined by a random binary 'm-sequence' that changes contrast at the frame rate of the stimulator (commonly at 75 Hz). In its simplest form, at any given transition, roughly half of the hexagons will be white, the other half black (**Figures 3(b) and 3(c)**). A single continuous record (**Figure 3(d)**) is obtained after about 8 min of recording, and the software computes the 103 local responses (**Figure 3(e)**) by means of a correlation approach. The results can also be plotted as a three-dimensional plot (**Figure 3(f)**) that shows the response density (the response size divided by the stimulus area). For the commonly used light levels and resolution of the hexagons, the multifocal ERG is a photopic response, dominated by cone bipolar cells, but not fully formed conventional ERG b-waves, because of the rapid transitions in contrast that occur more frequently than the integration time of the response. Some cone photoreceptor contribution can be seen, particularly in central hexagons. If blank frames are interleaved to slow the presentation so that flashes occur only every 100–200 ms, the multifocal stimulus can produce responses that are similar to flash ERGs, and this can be used to study local contributions to a more recognizable ERG waveform.

Applications of Electroretinography

ERG recording is noninvasive and is useful clinically for input to the diagnosis of retinal disease. The multifocal approach has improved the utility of electrodiagnostic testing because it has allowed clinicians to identify local functional problems objectively and to correlate them with subjective measures, such as perimetry, as well as with structural measures rendered by modern imaging approaches. Useful information for clinical testing can be found in the standards and guidelines of the International Society for Clinical Electrophysiology in Vision.

Normal retinal function in both humans and animals can be studied with the ERG. With appropriate selection of stimuli, the rod and cone systems can be examined separately, and specific levels of processing can be assessed. The ERG is particularly useful for tracking, noninvasively, the development of retinal function and the loss of retinal function as retinas in humans and animal models degenerate and modify their circuitries as a result of disease and, in the case

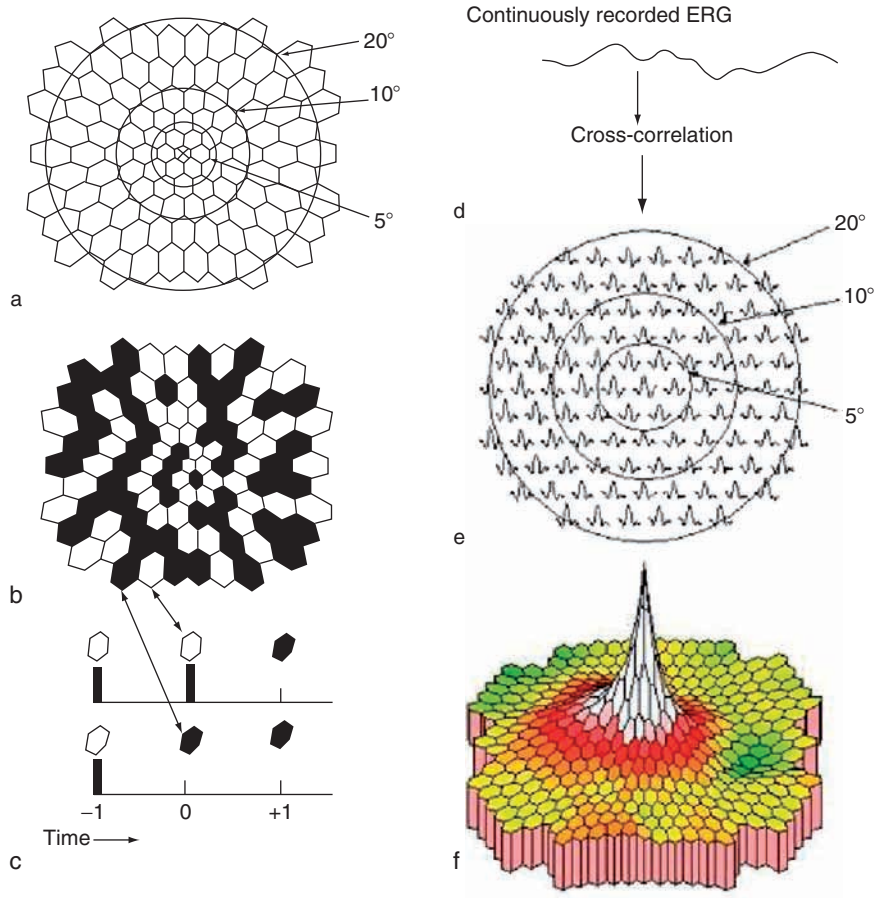

Figure 3 Multifocal electroretinogram (mfERG). (a) The mfERG stimulus display. (b) The display as it appears at a given point in time. (c) The sequence of random presentations for two of the hexagons. (d) The continuous record that is cross-correlated with the sequence in C. (e) The resulting mfERG responses. (f) The three-dimensional response density plot for the responses in (e). From Hood DC (2004) Electroretinogram, electrophysiology. In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience*, 3rd edn. Amsterdam: Elsevier.

of animals, genetic manipulations. Because the origins of ERGs are better understood than previously, ERGs now are useful for noninvasive, *in vivo* evaluation of retinal function in physiologically intact animals, including humans, and thus can provide a useful complement to functional information obtained from studies of single cells.

See also: Cone Photopigment Evolution; Fovea: Primate; Photoreceptors: Physiology; Phototransduction; Retina: An Overview; Retinal Bipolar Cells; Retinal Color Mechanisms; Retinal Development: An Overview; Retinal Glia; Retinal Models; Retinitis Pigmentosa.

Further Reading

Fishman GA, Birch DG, Holder GE, and Brigell MG (2001) *Electrophysiologic Testing*. San Francisco: Foundation of the American Academy of Ophthalmology.

Frishman LJ (2005) Electrogenesis of the ERG. In: Ryan SJ and Hinton DR (eds.) *Retina, Vol. I: Basic Science and Inherited Retinal Diseases*, 4th edn., pp. 103–135. Philadelphia: Elsevier/Mosby.

Frishman LJ (2006) Origins of the ERG. In: Heckenlively J and Arden GB (eds.) *Principles and Practices of Clinical Electrophysiology of Vision*, 2nd edn., pp. 139–183. Cambridge, MA: MIT Press.

Heckenlively J and Arden GB (eds.) (2006) *Principles and Practices of Clinical Electrophysiology of Vision*, 2nd edn. Cambridge, MA: MIT Press.

Holder GE (2001) Pattern electroretinography (PERG) and an integrated approach to visual pathway diagnosis. *Progress in Retinal and Eye Research* 20: 531–561.

Holder GE, Brigell MG, Hawlina M, Meigen T, Vaegan, and Bach M (2007) for the International, Society for Clinical, Electrophysiology of Vision. ISCEV standard for clinical pattern electroretinography: 2007 update. *Documenta Ophthalmologica. Advances in Ophthalmology* 114: 111–116.

Hood DC (2000) Assessing retinal function with the multifocal ERG technique. *Progress in Retinal and Eye Research* 19: 607–646.

Hood DC (2004) Electroretinogram, electrophysiology. In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience*, 3rd edn. Amsterdam: Elsevier.

Hood DC and Birch DG (2006) Measuring the health of the human photoreceptors with the leading edge of the a-wave. In:

Heckenlively J and Arden GB (eds.) *Principles and Practices of Clinical Electrophysiology of Vision,* 2nd edn., pp. 487–502. Cambridge, MA: MIT Press.

Hood DC and Greenstein VC (2003) Multifocal VEP and ganglion cell damage: Applications and limitations for the study of glaucoma. *Progress in Retinal and Eye Research* 22: 201–251.

Marmor MF, Holder GE, Seeliger MW, and Yamamoto S (2004) International Society for Clinical Electrophysiology of Vision Standard for clinical electroretinography (2004 update). *Documenta Ophthalmologica. Advances in Ophthalmology* 108: 115–123.

Porciatti V, Saleh M, and Nagaraju M (2007) The pattern electroretinogram as a tool to monitor progressive retinal ganglion cell dysfunction in the DBA/2J mouse model of glaucoma. *Investigative Ophthalmology & Visual Science* 48: 745–751.

Pugh EN Jr., Falsini B, and Lyubarsky A (1998) The origin of the major rod- and cone-driven components of the rodent electroretinogram and the effects of age and light-rearing history on the magnitude of these components. In: Williams TP and Thistle AB (eds.) *Photostasis and Related Phenomena*, pp. 93–128. New York: Plenum.

Robson JG, Saszik SM, Ahmed J, and Frishman LJ (2003) Rod and cone contributions to the *a*-wave of the electroretinogram of the dark-adapted macaque. *Journal of Physiology* 547: 509–530.

Sutter EE and Tran D (1992) The field topography of ERG components in man. I. The photopic luminance response. *Vision Research* 32: 433–466.

Relevant Website

http://webvision.med.utah.edu – Webvision: The Organization of the Regina and Visual System.

Emotion and Vigilance

F C Davis, J A Oler, and P J Whalen, Dartmouth College, Hanover, NH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A common strategy invoked by scientists who aim to understand the neural substrates of emotion is to focus on one emotion exemplar. For instance, we have learned much about the role of the amygdala and its interaction with reciprocally connected prefrontal cortical regions by studying their reactivity during fear states. Emotional states such as fear have lasting impacts on neural organization, as plastic changes in brain regions such as the amygdala, hippocampus, and prefrontal cortex subserve memories that allow us to avoid fear-inducing stimuli in the future.

Fear states also serve to protect us in the moment. For example, a scared organism often initially 'freezes' its movement, a behavioral response that is orchestrated by the amygdala. In addition to being the quintessential manifestation of a state of fear, somatomotor arrest is also useful if the organism believes that it has more to learn before resuming action. Such a response in the face of a cue that predicts a biologically relevant outcome allows an organism to physically orient sensory organs (e.g., eyes, ears) toward a stimulus, thereby optimizing sensory information processing. This strategy is particularly adaptive if the outcome predicted by this cue is not entirely clear, as it aids in processing additional environmental information that might facilitate the interpretation of a predictively ambiguous cue. Thus, a state of fear includes amygdala responsivity, which influences nonspecific attentional systems that support learning about the environment in order to facilitate behavioral responses now and in the future. A critical question currently under investigation concerns whether an extreme state of fear is necessary to engage amygdala modulation over this sort of acute information gathering. Research on this topic suggests that the amygdala is involved along a continuum of state changes, from subtle changes in the state of vigilance, to more punctate moments of surprise, to a full-blown state of horror.

The purpose of this article is to review data demonstrating that a partially shared prefrontal–amygdala neural circuit subserves more subtle fluctuations in vigilance as well as more extreme fear states. This network is responsible for monitoring the environment for biologically relevant events, learning about the cues that predict those events, and modifying behavior in order to appropriately react to those cues in the future. Thus, affective information processing is part of a greater system that facilitates biologically relevant learning, especially when learning contingencies are uncertain. This system is capable of, but does not necessitate, the generation of extreme emotional states.

The Amygdala

The anatomy of the amygdala is not detailed here. Rather, we focus on three important aspects of amygdala anatomical organization relevant to the ensuing discussion (see Figure 1). First, the heaviest feedback projections from sensory and prefrontal cortical regions are received by the basolateral amygdala (BLA). This design allows for the convergence of multimodal cortically processed information with the latest sensory stimuli being detected. Second, in addition to the known role of the central nucleus (CN) in modulating autonomic and motor responsivity, the CN also projects to all major neuromodulatory centers (e.g., cholinergic, dopaminergic, serotonergic, and noradrenergic source neurons). It is these connections that can more globally modulate the 'readiness' of cortical and subcortical sensory neurons, thereby setting an appropriate level of vigilance. Third, whether BLA activity will provoke activity within the CN (thereby affecting vigilance) is largely controlled by the gamma-aminobutyric acid (GABA)ergic intercalated cell masses (ICMs), which also receive direct prefrontal input. Thus, the functional role of the ICMs provides an answer to a long-standing conundrum concerning amygdala function: how can a detection system (e.g., BLA) effectively continue to represent the present value of a predictive stimulus (e.g., in terms of valence) without necessarily calling the action system (e.g., CN) into action? The answer lies in the ICMs. By acting as a sort of 'moat' the ICMs provide an inhibitory barrier between the BLA and the CN. In this way, the BLA can continue to represent potential threats, while the ICMs inhibit CN output, thereby inhibiting overt responses to potentially threatening stimuli that have proved to pose no threat in the present context (i.e., habituation). Thus, habituation of responsivity of the amygdaloid complex to repeated presentation of biologically relevant stimuli is an active process. It does not necessitate a time-limited role for the amygdala, but instead represents the conservation of resources (e.g., responsivity of the CN), while the BLA continues

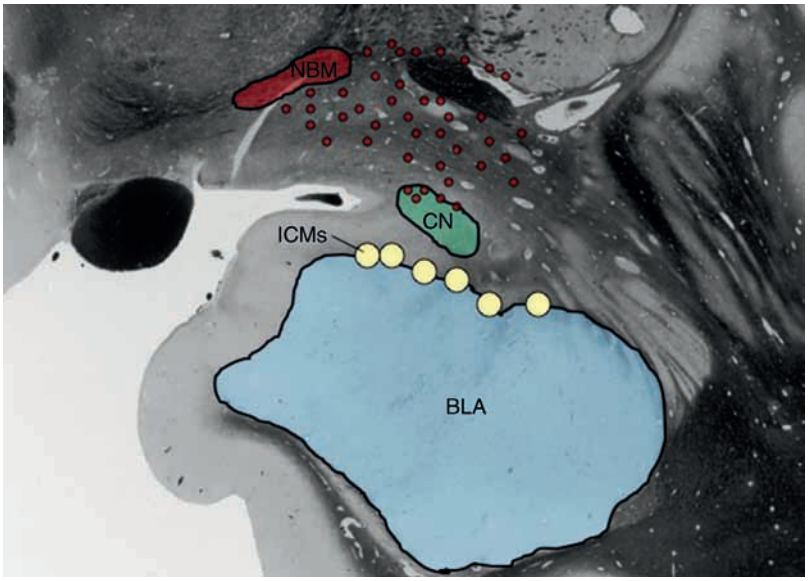

Figure 1 Coronal depiction of the human amygdala: BLA, basolateral amygdala; ICM, intercalated cell masses; CN, central nucleus; NBM, nucleus basalis of Meynert. Adapted from Mai JK, Assheuer J, and Paxinos G (2004) *Atlas of the Human Brain*. Amsterdam: Elsevier Academic press.

to represent the present, potential, and/or future significance of these stimuli.

Nonspecific Arousal Facilitates Attention

The CN serves two primary functions. The most widely recognized role of the CN is that of an output station for the amygdala. With direct connections to the hypothalamus and numerous brain stem motor neuron targets, the CN can drive autonomic nervous system reactivity and affect the coordination of behavioral responses to biologically relevant stimuli. However, the CN also serves another primary purpose during biologically relevant learning, which is to enhance nonspecific arousal and facilitate attention to allow the animal to detect and appropriately respond to stimuli in its environment. Indeed, stimulation of the CN enhances reflexes such as the nictitating membrane reflex, startle reflex, pupil dilation, eye widening, and somatomotor arrest; all responses that indicate that an organism is gathering information. During Pavlovian conditioning, an animal is first exposed to a surprising event (e.g., shock), which automatically and nonspecifically enhances this group of attentional reflexes. This enhanced level of vigilance allows the animal to detect any additional environmental events that may be related to the occurrence of shock, specifically, another event that may aid in predicting its reoccurrence. One CN projection that aids in producing this increased level of vigilance is the projection from the CN to the nucleus basalis of Meynert (NBM). The NBM is directly superior to the CN within the substantia innominata (SI) region of the ventral basal forebrain. While the NBM comprises a compact, 'nucleus-like' portion medially, this nucleus also includes a more distributed lateral portion where individual NBM neurons can be found immediately superior to and even within the CN (see Figure 1). NBM neurons provide the primary source of acetylcholine (ACh) for the entire cortex. ACh release from the NBM serves to potentiate neuronal responsivity (i.e., decrease response thresholds) and thereby facilitates information processing throughout cortical systems. Stimulation of the CN in the rabbit produces activation of cortical neurons (as measured by cortical EEG), and cholinergic antagonists can block this effect. Driven by inputs from the BLA signaling the detection of important sensory stimuli, the CN can directly influence the NBM, thereby augmenting sensory information processing by increasing ACh release throughout the cortex. An organism in this heightened state of vigilance becomes a more efficient processor of information. Any new information collected during this state of arousal can then be communicated back to the BLA via cortical feedback projections. Such feedback serves to update information stored in the BLA, which impacts the response of the animal the next time a similar activating stimulus is encountered.

The BLA–CN Connection

Lesion data from appetitive learning paradigms support a role for the BLA in assigning motivational significance to neutral stimuli and for the CN/SI

in influencing widespread attentional mechanisms. While neither BLA nor CN/SI lesions impair the ability to associate rewarding outcomes with conditioned stimuli, more subtle deficits have been observed that help to elucidate the differential roles of these amygdaloid subregions during biologically relevant learning. For example, during second-order appetitive conditioning a neutral stimulus (e.g., a light) initially predicts a reward (e.g., food). Once the animal has learned this association, the light becomes rewarding because of its predictive value. Subsequent pairing of the light with a tone results in second-order associative learning, as the animal learns that the tone will also predict a reward. Lesions of the BLA block acquisition of this second-order association, suggesting that the light never becomes reinforcing in BLA-lesioned animals. Another phenomenon that has helped to elucidate the role of the BLA during associative learning is that BLA-lesioned animals do not show stimulus devaluation. In intact animals, it is possible to devalue a food reward by mixing it with a toxin, and later presentation of the light (conditioned stimulus – CS) no longer elicits a conditioned response, suggesting that the light has also been devalued. However, BLA-lesioned animals do not show this type of learning. Taken together, these data suggest that the BLA is necessary to assign reinforcing value to a CS, such that the light assumes some of the reinforcing qualities of food.

Lesions of the CN have no effect on second-order conditioned responding or stimulus devaluation, but do impact attentional mechanisms during appetitive learning. Intact animals orient to novel stimuli (e.g., a light), and this orienting reflex habituates as the animal learns that the novel stimuli are not biologically significant. However, if you then pair light with a food reward, orienting to the light returns. This sort of orienting is referred to as 'associative' orienting. CN-lesioned animals do not show potentiated associative orienting to light-food pairings, even though they still show basic orienting responses to novel stimuli. Thus, the CN is necessary for associative orienting, a process that facilitates attention to predictive stimuli in anticipation of a biologically relevant event.

The aforementioned research emphasizes projections from the CN/SI in increasing cortical ACh release, which lowers neuronal response thresholds, thereby facilitating attention during associative learning paradigms. However, it is important to note that other neurotransmitter systems are also modulated by amygdala function and play a role in behaviors such as associative orienting. For example, CN modulation of the nigrostriatal dopaminergic pathway is necessary for the acquisition and expression of conditioned orienting responses. This same pathway is also implicated in enhancement of attention to surprising (i.e., unpredictable) stimuli during associative learning tasks. Thus, the amygdala's influence on attention during biologically relevant learning is diffuse and likely involves multiple neurotransmitter systems.

The Amygdala–Prefrontal Connection

Once an animal has learned that a particular stimulus predicts a biologically relevant event, the motivational significance of this stimulus can be represented in the BLA. But complex representations of stimuli (e.g., contextually dependent representations) and the coordination of these representations with appropriate behavioral choices appear to be subserved through connectivity between the amygdala and prefrontal cortex. Anatomical tracing studies in nonhuman primates document reciprocal connections between the BLA and the medial prefrontal cortex (mPFC), and functional data in both animals and humans suggest that these connections support fundamental context-dependant valence calculations. Moreover, stimulation studies in the rat show that BLA stimulation inhibits mPFC, and that mPFC stimulation has inhibitory effects on both the BLA and the CN. These effects could be mediated by direct inhibition of the BLA, and/or via direct mPFC projections to GABAergic ICMs, which exert inhibitory control over CN output. Indirect routes could also support modulatory effects between these regions. For example, the amygdala projects to ventral tegmental brain stem neurons that provide the mPFC with dopamine. Interestingly, activation of these dopaminergic neurons inhibits mPFC activity, suggesting that similar effects seen following BLA stimulation could be mediated indirectly via the brain stem.

The animal literature documents that the mPFC provides an important regulatory input to the amygdala, communicating cortical valence representations as they relate to predicted outcomes. Much information about this circuitry has been elucidated through the study of the extinction process following Pavlovian conditioning, which indicates that a specific ventral portion of the mPFC (vmPFC) is necessary for the consolidation and expression of extinction memories (i.e., tone now predicts no shock). This region is referred to as the infralimbic cortex (IL) in the rat but will be generally referred to as vmPFC as we discuss data in rats and humans here. Lesions of the vmPFC do not disrupt initial extinction learning, but on the day following extinction, vmPFC-lesioned animals show no memory for their previous extinction training (i.e., continue to freeze in response to the tone, though it now predicts no shock). Single-unit

recording studies indicate that vmPFC neurons respond to conditioned stimuli, and their responses are positively correlated with extinction memory (i.e., reductions in conditioned freezing). Stimulation of vmPFC neurons during extinction training accelerates learning, and evidence suggests this active inhibition of conditioned responding works via a vmPFC–ICM–CN pathway. Thus, vmPFC inputs to the amygdala can dynamically inhibit amygdala activity, signaling either a new, more positive interpretation of a once negative predictive stimulus, or, perhaps, a more general override message that the valence of the outcome predicted by the cue has changed.

Initial neuroimaging studies of the extinction process in humans have reported results consistent with these data in animals. While amygdala activation predicts expression of a conditioned response during both conditioning and extinction training (while stimulus contingencies are changing), activity in the vmPFC only predicts a conditioned response during extinction trials 1 day later. Moreover, recent work indicates that the cortical thickness of the vmPFC is significantly correlated with retention of extinction memory. These data are consistent with the notion that the amygdala is involved in biologically relevant learning while stimulus contingencies are changing, but that the vmPFC is involved in the consolidation and expression of extinction memories via regulatory inputs to the amygdala.

Thus, consistent with the modulatory role the vmPFC appears to play during extinction, interactions between the vmPFC and amygdala may be especially important for producing appropriate responses to changes in the predictive meaning of stimuli, especially when these changes include a shift in predicted valence. These animal and human findings resonate with observations of clinical patients with vmPFC lesions, who show intact psychophysiological responses to positive and negative outcomes (e.g., GSR responses), but lack the ability to choose advantageously when these situations are subsequently encountered.

Amygdala–vmPFC Interactions to Predictive Uncertainty

Thus far we have summarized just a small portion of what we have learned about amygdala–prefrontal function through paradigms that necessarily induce a strong fear state in animal subjects. However, many neuroimaging studies in humans have also sought to assess the role of the amygdala in processing more innocuous biologically relevant predictive stimuli that likely induce more subtle fluctuations in attention. For example, static pictures of facial expressions of emotion consistently activate the amygdala during human neuroimaging studies. This is because facial expressions are salient, biologically relevant stimuli that have predicted important events for you in the past. In other words, facial expressions are conditioned stimuli. A happy facial expression likely predicts reinforcing or affiliative behavior from a conspecific, whereas an angry facial expression predicts social threat. Fearful facial expressions are of particular interest to the amygdala, presumably because they tell the viewer that something important and potentially dangerous is happening in the environment, but, unlike angry faces, do not reveal the source of this threat.

A recent neuroimaging experiment in humans delineates a functional spatial dissociation in the amygdala using facial expressions of emotion that differ in their predictive value. In an experimental environment, static pictures of facial expressions have no biological relevance to viewers, but, because of the past predictive nature of the expressions, the amygdaloid system does respond to these conditioned stimuli, at least initially. As predicted by the animal data outlined thus far, both fearful and angry facial expressions activate the BLA, which represents these faces as negatively valenced. However, these two facial expressions differ on one important dimension: angry faces are inherently embody of negativity, while in the case of fearful faces, more information is needed to determine the exact nature of the source of threat. Such a model predicts that angry expressions can be handled via BLA outputs related to the enactment of an appropriate response plan (e.g., are you going to fight or flee?). In contrast, the meaning of a fearful facial expression is context dependent and thus requires the CN and its connections to neuromodulatory centers, such as the corticopetal cholinergic neurons of the NBM, to call upon the cortex to process additional information in the environment. This assertion was validated recently in a neuroimaging study showing that both fearful and angry faces activate the ventral amygdala where the BLA is located, but only fearful faces activate the dorsal amygdala/SI region where the CN and NBM are located. Thus, facial expressions call for parallel processing across a number of different dimensions, as two different circuits through the amygdaloid complex simultaneously compute valence (fear and anger are both clearly negative and activate the ventral amygdala) and uncertainty (only fear requires greater processing of contextual information and is related to dorsal amygdala/SI activation) processes. These data predict that information processing should be enhanced following the presentation of fearful facial expressions. Indeed, recent data show that humans show enhanced

perceptual discrimination abilities following presentations of fearful faces.

One way to further understand amygdala response to fearful facial expressions would be to test an expression that, like fear, was ambiguous in terms of its eliciting event but, unlike fear, could predict either a negative or positive event. Facial expressions of surprise are similar to facial expressions of fear both configurally (e.g., wide eyes) and in terms of their information value in that they tell you that something important is happening in your environment, but you need more information to determine what it is and how to react to it. However, they differ from fearful facial expressions on one important dimension: surprised expressions can predict both positive and negative outcomes. When viewers are asked to rate surprised faces using a valence scale (see Figure 2), some viewers interpret the expression positively, while others interpret it negatively. Consistent with the previously discussed data comparing amygdala response to fear and anger, surprised faces produce greater ventral amygdala activation only in viewers who interpret them negatively. Further, this effect within the amygdala appears to depend on inputs from the prefrontal cortex. In this same study amygdala activity was inversely correlated with vmPFC activity. That is, viewers with a more positive interpretation of surprised faces show greater vmPFC and lower amygdala activity. Thus, viewers showing a stronger vmPFC reaction in an area known to send inhibitory projections to the amygdala interpret the ambiguous facial expressions more positively.

1------- 2------- 3------- 4 ------- 5------- 6------- 7------- 8-------- 9
VP P NN N VN

Figure 2 Valence scale. Study participants view surprised facial expressions and are asked to rate how positive or negative the expression on the face is: VP, very positive; P, positive; NN, neither negative nor positive; N, negative; VN, very negative. Reproduced from Ekman P and Friesen WV (1976) *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press. Face image used by permission from PaultEkman.com.

Linking Amygdala–vmPFC Response and Surprised Faces to the Fear Extinction Literature

The amygdala is necessary for the acquisition and expression of learned fear associations. Typically, animals learn to fear a CS (e.g., tone) that predicts an unconditioned stimulus (US; e.g., shock). Thalamic and cortical afferents communicate CS and US information to the BLA, where CS–US associations are thought to be formed. As described earlier, the behavioral and autonomic expression of conditioned fear, however, is thought to result from excitatory influences of the BLA on the coordinated efferent systems of the CN.

As already discussed, extinction involves the attenuation of a conditioned response when a tone CS is repeatedly presented in the absence of a US. Behavioral studies have demonstrated that extinction does not expunge the CS–US association, but rather a new memory trace is formed that inhibits the previously established conditioned response. After extinction, therefore, the CS is inherently ambiguous, potentially predicting either shock or the absence of shock.

Relevant to the present discussion, lesions of the vmPFC in the rat affect the ability to use this new information learned about the CS during extinction, when the CS is subsequently encountered. For example, while neurons of the vmPFC are not responsive to a tone CS during conditioning or extinction training, these same cells show robust CS-evoked firing when the animal is recalling extinction the following day. In addition, electrical stimulation of the vmPFC is sufficient to inhibit the expression of freezing, mimicking the effects observed following extinction training. This suggests that the vmPFC can regulate the response of the amygdala to extinguished tones that now have ambiguous predictive value.

It is hoped that the reader sees the parallel between this discussion (vmPFC control of the amygdala to extinguished tones in rats) and the preceding discussion (vmPFC–amygdala response to surprised faces) (see Figure 3). Like extinguished tones, surprised faces also have an inconsistent reinforcement history, predicting both positive and negative outcomes. Viewers who show strong vmPFC activation to extinguished tones, or surprised faces, are readily able to recall the more positive interpretation of the presented stimulus. While these effects have been shown to be causal in

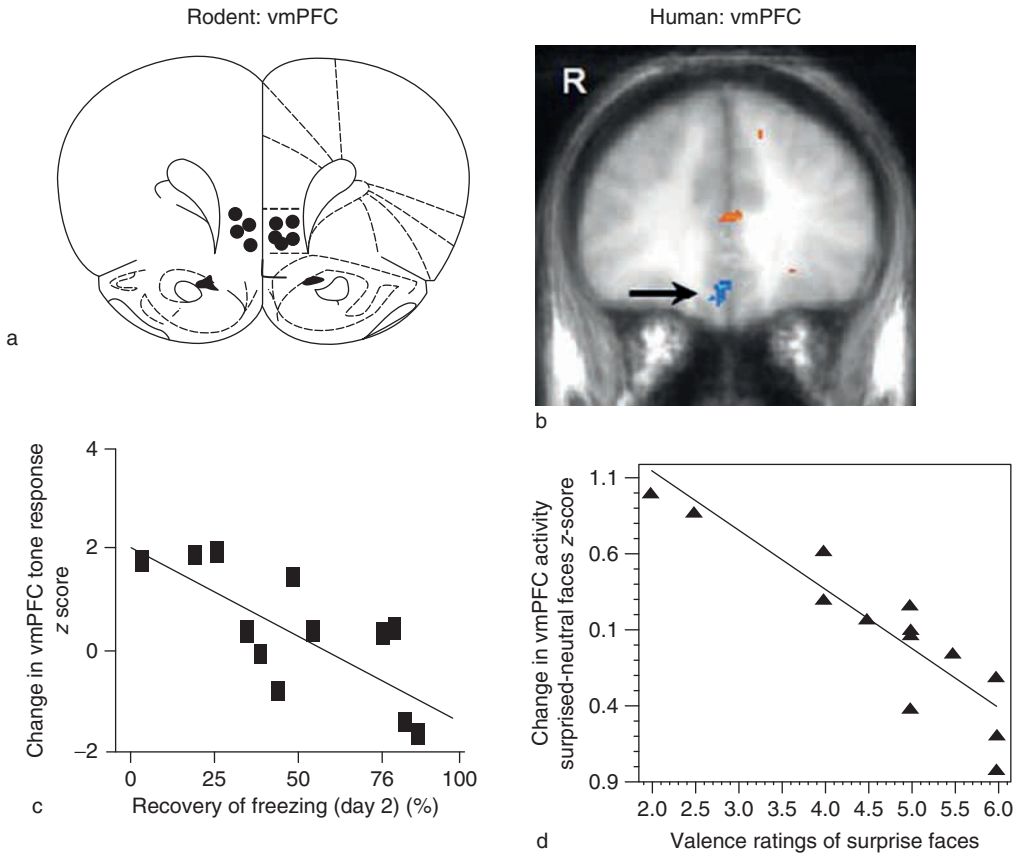

Figure 3 Ventromedial prefrontal cortex (vmPFC) activity and behavior are inversely correlated in the rodent and human. (a) Single-cell recording sites in the rodent vmPFC. (b) Functional magnetic resonance imaging data in the human; blue color indicates surprised < neutral faces. (c) Scatter plot showing that less activity in vmPFC is related to increased recovery of freezing on day 2 (i.e., reduced extinction memory) in rats. (d) Scatter plot showing that less activity in vmPFC is related to more negative interpretations of surprised facial expressions in humans. Study participants view surprised facial expressions and are asked to rate how positive or negative the expression on the face is (see **Figure 2**; higher ratings indicate more negative interpretations). Adapted from Milad MR and Quirk GJ (2002) Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature* 420: 70–74; and Kim H, Somerville LH, Johnstone T, et al. (2003) Inverse amygdala and medial prefrontal cortex responses to surprised faces. *NeuroReport* 14: 2317–2322.

the animal model and necessarily remain correlational in the human model, we believe a vmPFC–amygdala circuitry similar to that which extinguishes conditioned fear accounts for subjective interpretations of valence in humans who encounter an ambiguously valenced facial expression (such as surprise).

The animal and human data lead us to the following working hypothesis: upon encountering the expression of surprise, the amygdala initially sends out a 'first pass' message, identifying the uncertain nature of these faces with respect to valence, and likely categorizes these faces as 'potentially threatening' in all viewers. The vmPFC, based upon additional inputs from multiple brain regions providing information about past experiences or present context, could then communicate alternative hypotheses back to the amygdala, including the 'potential positivity' of these faces. Individual differences in the strength of this vmPFC 'override' message would account for the observed differences in valence ratings. This scenario is consistent with numerous models and examples of inverse amygdala–vmPFC interaction, and suggests a mechanism by which complex representations of valence, communicated via the vmPFC, can influence the amygdala's representation of the predictive value of a stimulus. These valence representations are used to calculate appropriate behavioral responses within a given environmental context and can change the way that the amygdala influences cognitive functions such as attention and memory.

The notion that the amygdala is sensitive to environmental uncertainty *per se* and that it will first test whether this uncertainty is potentially negative was demonstrated in a recent study using mice and humans. The amygdala was shown to be more active

during segments of time when unpredictable tones were played compared to segments when tones were played in a predictable pattern. Critically, unpredictable tones also biased the study participants to treat other biologically relevant environmental events as potentially threatening. Thus, unpredictablity *per se*, even for events that are not biologically relevant, is enough to engage the amygdala. This activity creates a vigilance state that augments an animal's reaction to subsequent potential danger.

Role of the Amygdala in Vigilance: Resolving Associative Ambiguity

Resolving associative ambiguity is a complex process requiring access to memory, temporal contexts, spatial contexts, and visceral contexts in order to make a probabilistic 'guess' as to the best course of action. More than simple uncertainty, when one does not yet have a working hypothesis about predicted outcomes, the term 'associative ambiguity' very specifically refers to a situation in which a predictive stimulus has more than one potential meaning (based upon prior learning) that must be selected in a given situation. Here we consider the role of vmPFC–amygdala interactions in determining which valence representation (positive or negative) will be invoked in a particular circumstance. Stimuli with inconsistent reinforcement histories (such as extinguished tones and surprised faces) are inherently ambiguous, potentially predicting either a positive or negative outcome, and this vmPFC-amygdala circuit appears to subserve the calculation and/or retrieval of this information.

The vmPFC is ideally situated for this function, given its access to working memory, spatial (hippocampal) function, and visceral inputs, as well as to outputs able to override subcortical fear expression centers. Each of the divergent brain stem autonomic and somatic targets of the amygdala CN also receives a projection from these medial prefrontal areas, most of which are inhibitory. This arrangement, in a healthy human brain, effectively gives the vmPFC 'veto power' over subcortical conditioned responses, allowing the organism to respond appropriately based on recent experience.

Failure to resolve associative ambiguity appropriately can lead to pathological outcomes, and the vmPFC has been implicated in a variety of affective disorders, as well as the genetic predisposition to inappropriate or pathological expression of affect. For example, posttraumatic stress disorder (PTSD) is characterized by an inability to suppress fear responses to stimuli that were once associated with trauma but now no longer predict danger. Consistent with impaired cortical regulation of the amygdala, PTSD patients show decreased volume and decreased activity in vmPFC, coupled with increased amygdala activity. As mentioned previously, a similar area of the vmPFC was recently shown to correlate (in thickness) with memory for fear extinction in normal study participants, suggesting that PTSD patients may be deficient in the neural circuitry underlying extinction. Thus, similar to those persons showing a negative bias in the interpretation of surprised faces, persons with smaller vmPFC volume tend to interpret conditioned stimuli that had been extinguished (and were therefore ambiguous) as dangerous.

Conclusions

The amygdala coordinates emotional responses and also modulates cognitive functions such as attention that affect subsequent learning and memory. The amygdala is consistently implicated in learning about, and assigning emotional meaning to, biologically relevant events and the stimuli that predict those events. Interactions between the amygdala and vmPFC provide flexible, context-dependent representations of emotional valence. If a stimulus predicts a certain aversive or appetitive event, then the amygdala can coordinate an appropriate behavioral response to that stimulus. If the predictive value is uncertain, because more contextual information is needed, or perhaps because of a stimulus' inconsistent reinforcement history, vigilance networks are recruited to gather additional information from the environment. Here we have highlighted the amygdala's influence over this process through the NBM cholinergic system. But the amygdala projects to all major neuromodulatory centers. Indeed, amygdala projections to substantia nigra dopaminergic corticopetal neurons have been specifically implicated in response to surprising outcomes. Through these networks, the amygdala influences cognition in the service of protecting us from danger or leading us toward reward. This not only happens in extreme emotional circumstances, but it also happens on a more subtle, moment-by-moment basis. By constantly monitoring the environment for biologically relevant cues, the amygdala maximizes our ability to avoid danger or avail ourselves of reward.

The complementary animal and human findings discussed in this article are consistent with a role for the vmPFC in the regulation of amygdala responsivity during the subjective interpretation of ambiguous stimuli. Indeed, these results join a diverse list of experimental paradigms implicating the vmPFC in regulatory control when stimuli are associatively ambiguous. Associative ambiguity is high following extinction, when conflicting facts about a CS have been

learned. Associative ambiguity is also high during reversal learning, set-shifting, and other situations requiring behavioral flexibility. Taken together, these studies support the more generalized involvement of frontotemporal interactions in behavioral flexibility. As more unified theories of limbic function advance, such theories must include a role for vmPFC–amygdala integration and regulatory control of biologically relevant information processing, particularly when contingencies are ambiguous.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Structure and Circuitry in Rodents and Felines; Amygdala: Contributions to Fear; Emotion: Neuroimaging; Emotion: Computational Modeling; Emotional Influences on Memory and Attention; Pharmacology of Fear Extinction; Posttraumatic Stress Disorder: Overview; Social Emotion: Neuroimaging.

Further Reading

Aggleton JP and Saunders RC (2000) The amygdala – what's happened in the last decade? In: Aggleton JP (ed.) *The Amygdala: A Functional Analysis,* 2nd edn., pp. 1–30. New York: Oxford University Press.

Amaral DG, Price JL, Pitkanen A, et al. (1992) Anatomical organization of the primate amygdaloid complex. In: Aggleton JP (ed.) *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*, pp. 1–66. New York: Wiley-Liss.

Bouton ME (1994) Context, ambiguity, and classical conditioning. *Current Directions in Psychological Science* 3: 49–53.

Davis M and Whalen PJ (2001) The amygdala: Vigilance and emotion. *Molecular Psychiatry* 6: 13–34.

Ekman P and Friesen WV (1976) *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press.

Heimer L and VanHoesen GW (2006) The limbic lobe and its output channels: Implications for emotional functions and adaptive behavior. *Neuroscience & Biobehavioral Reviews* 30: 126–147.

Herry C, Bach DR, Esposito F, et al. (2007) Processing of temporal unpredictability in human and animal amygdala. *Journal of Neuroscience* 27: 5958–5966.

Holland PC and Gallagher M (1999) Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences* 3: 65–73.

Kapp BS, Wilson A, Pascoe JP, et al. (1990) A neuroanatomical systems analysis of conditioned bradycardia in the rabbit. In: Gabriel M and Moore J (eds.) *Learning and Computational Neuroscience: Foundations of Adaptive Network*, pp. 53–90. Cambridge, MA: MIT Press.

Kim H, Somerville LH, Johnstone T, et al. (2003) Inverse amygdala and medial prefrontal cortex responses to surprised faces. *NeuroReport* 14: 2317–2322.

LeDoux JE (1987) Emotion. In: Blum F, Geiger SR, and Mountcastle VB (eds.) *Handbook of Physiology, Section 1, The Nervous System,* vol. 5, pp. 419–459. Bethesda, MD: American Physiological Society.

Likhtik E, Pelletier JG, Paz R, et al. (2005) Prefrontal control of the amygdala. *Journal of Neuroscience* 25: 7429–7437.

Mai JK, Assheuer J and Paxinos G (2004) *Atlas of the Human Brain.* Amsterdam: Elsevier Academic Press.

Milad MR and Quirk GJ (2002) Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature* 420: 70–74.

Phelps EA, Ling S, and Carrasco M (2006) Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological Science* 17: 292–299.

Weisz DJ, Harden DG, and Xiang Z (1992) Effects of amygdala lesions on reflex facilitation and conditioned response acquisition during nictitating membrane response conditioning in rabbit. *Behavioral Neuroscience* 106: 262–273.

Whalen PJ (1998) Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science* 7: 177–188.

Emotion in Speech

J Bachorowski, Vanderbilt University, Nashville, TN, USA
M J Owren, Georgia State University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Emotion-Related Vocal Acoustics

Speech sounds are imbued with indexical or personal cues, meaning nonlinguistic aspects of the signals that provide acoustic correlates of talker characteristics such as sex, individual identity, age, and emotional state. Indexical cues are provided in nonlinguistic vocalizations, such as laughter, as well. Regardless of the type of vocal signal being analyzed, the source–filter model of speech production has been central to selecting which acoustic measurements to make. In this model, vocal acoustics arises from the combination of an underlying energy source and filtering effects that are due to the resonance properties of the supralaryngeal vocal tract. Although source energy is noisy for some speech sounds, those of greatest interest involve quasi-periodic vocal-fold vibration, which produces phonated or voiced phonemes such as vowels. The basic rate of vibration in the sounds is referred to as the fundamental frequency (F_0), which largely contributes to their perceived pitch. Associated acoustic properties examined in searching for cues to vocalizer emotional state often include mean F_0 and a number of measures associated with F_0 variability. The latter are, for instance, overall range, F_0 contour, and moment-to-moment perturbations in vocal-fold vibration. Physical amplitude or perceived loudness of vocal signals has also been shown to be important.

Supralaryngeal resonances are also studied, with the frequency of these formants being of specific interest. Supralaryngeal filtering exerts a shaping effect on the frequency of voiced sounds, resulting in prominent energy bands at resonance locations and attenuated source energy elsewhere in the frequency spectrum. Close tracking of formant changes has the potential to reveal the effects of momentary emotional states as they occur, although there has to date been little evidence for this kind of effect. However, global frequency characteristics of emotion-laden speech sounds have been described by calculating a long-term average spectrum (LTAS) over speech samples of 30 s or more. LTAS is quick, robust, and well-suited to capturing longer-term trends in vocal expression, but is less likely to capture the fleeting effects of brief emotional experiences as they occur.

Vocal Expression of Emotion in Speech

The most common approach to understanding vocal expression of emotion in speech can be called the cue-configuration perspective. In this view, discrete affective states experienced by the vocalizer are reflected in specific patterns of acoustic cues in the speech being produced. In contrast, the dimensional perspective holds that emotional cues in speech acoustics more closely reflect arousal and valence dimensions. Although the relative merits of the two perspectives are debated, there are few if any empirical studies that have compared them directly.

It has been challenging to make that kind of direct comparison, with most studies examining predictions derived specifically from one or the other of the two perspectives. One common approach is to record speech meant to portray particular emotions, as performed by trained actors or naïve experimental participants. Trained or naïve listeners are then asked to judge the intended emotions, followed by acoustic analysis of the best exemplars (and sometimes further perceptual testing). Of the many acoustic properties that can be measured, F_0 and mean acoustic amplitude are most often found to show clear connections to the emotions being portrayed. These findings are important because, regardless of theoretical position, researchers widely agree that these features probably index vocalizer arousal rather than more specifically valenced states. Although other acoustic measures also show links to particular emotional portrayals, they typically account for much less of the variance involved. When many acoustic features are included in the analysis, classification based on discrete emotion categories (e.g., joy, sadness, anger, and fear) is approximately 40% correct, on average. This moderate but reliable classification success might be expected based on findings about other kinds of indexical cueing in speech. For example, classification of talkers from neutral speech segments results in classification accuracies of approximately 70%. The classification task is arguably more difficult when operating on the variability inherent to emotional utterances. On the other hand, these recordings are often carefully screened before classification begins, which probably improves apparent success. In addition, because emotions are also expressed emblematically both in everyday life and in acting, stimuli based on simulated rather than spontaneous emotional states may not be capturing the kind of cueing that occurs under natural circumstances.

As already noted, the most reliable empirical outcomes in testing the acoustics of emotional speech have been arousal related. Numerous studies have, for example, shown that both anger and joy are associated with increased F_0 and higher amplitude. Associations between valence and vocal acoustics are less clear-cut and often emerge only in complex interactions among variables such as sex of talker, task-dependent positive and negative feedback, and trait differences in emotional intensity.

Two conclusions that can be drawn are that arousal plays a noticeably more important role in shaping speech acoustics than does valence and that it is difficult to demonstrate clear links to discrete emotional states. However, it would be premature to therefore assume that the acoustics of emotional speech reflects only arousal or can be fully accounted for by a dimensional approach alone. Such an inference requires testing alternative predictions from the two perspectives head-to-head within the same empirical framework. The strongest conclusion that can be drawn at this point is that the acoustics of emotional speech is influenced by a variety of factors, not all of which are neatly linked with either the cue-configuration or dimensional positions. However construed, arousal and valence are not sufficient to account for the available data concerning vocal emotion effects. Other important factors appear to involve sex of talker, individual differences in emotional traits, strategic shaping of vocal acoustics in pursuit of social goals, and the social context in which the speech is being produced.

Perception of Emotion from Speech

The theory and findings associated with perception of emotion from speech acoustics generally parallel those described for production. Specifically, some researchers adopt a cue-configuration perspective, whereas others emphasize a dimensional view. When listeners are asked to identify the intended emotion in utterances produced by actors, accuracy is again moderate, approximately 55% across studies. Similar outcomes and confusions are observed across different cultures and language groups, although identification errors also reflect the degree of disparity between vocalizer language and listener language.

The standard experimental strategy has been to use forced-choice identification paradigms in which listeners select one from a small set of emotion words to best describe the affect being conveyed. Stimulus sets often include only a small number of talkers and emotions while relying on prototypical instances of the emotions in question. The strategy of using acted vocal samples and then selecting only a subset for testing may in and of itself account for some of the evidence of differentiated perception of emotion.

When listeners rate vocal samples using dimensional scales, by far the strongest associations between the ratings and vocal acoustics are arousal related. Just as in vocal production, the primary acoustic cues are F_0 and amplitude. For example, listeners report hearing stress when vocalizers show significant F_0 and amplitude variation, but otherwise fail to perceive it. The link between speech acoustics and valence perception is generally weaker, again paralleling outcomes observed in production. Moreover, rather than revealing categorical responses, listeners' responses often vary continuously in accordance with continuous changes in F_0. Overall, listeners are most likely to make errors when stimuli reflect similar arousal levels and similarly valenced members of emotion families.

Taken together, the perceptual evidence shows that listener accuracy is typically moderate. Many investigators argue that achieving approximately 55% correct identification indicates both that vocalizer emotion is associated with differentiated acoustics and that listeners perceive these cues effectively. On the other hand, the studies show incorrect identification nearly half the time, in spite of relying on stimuli that are specifically selected to be rich in emotional cueing. Under more natural circumstances, the emotional speech that humans hear is often produced by familiar individuals and is embedded in a rich social context. In real life, listeners attempting to infer a talker's emotional state are therefore likely to benefit from a history of interaction with that individual and a host of nonacoustic situational cues. Both factors are likely to significantly boost accuracy relative to that observed in laboratory testing. Even so, however, it would be difficult to conclude that the cues involved are either related to discretely organized emotions or invariant across different vocalizers.

Methodologically, it becomes important to recognize that when listeners are correct in assessing emotion from voices, they may be depending more on their own inferential processes than on veridical acoustic cues provided in the speech of vocalizers. This theme is elaborated next in an approach that shifts the focus from links between emotional states and vocalizer acoustics to the effects that sounds have on listeners. Although it is compatible with aspects of both discrete and dimensional-emotions views, this perspective argues that emotion-related cuing in vocalizers is secondary, whereas the primary function of nonlinguistic vocalization is to influence a listener's state and behavior.

Affect Induction as an Alternative to Representation

A central theme so far is that investigators have expected to find veridical links among vocalizer

affect, vocal acoustics, and listener perception. An implicit assumption of this approach is that vocalizers are signaling in order to convey to listeners that particular affective states are occurring. But simply informing listeners about such states is not necessarily beneficial. Instead, the evolution of signaling is necessarily shaped by the effect that signals have on listeners' behavior. The most fundamental selection pressure acting on signalers is to influence others in ways that benefit themselves. This premise leads to viewing affect-related acoustics more broadly as a means of influencing listeners by any means possible using sound, not just by conveying encoded information. Arguably the most fundamental way to sway a listener's behavior is to produce sounds that in and of themselves have an effect on that individual's emotional state.

Affect Induction through Sound

Everyday experience shows that auditory events of all kinds can influence listener emotion, from spectacular thunder claps to quietly dripping faucets. In humans, crying and laughter readily come to mind as examples of potent, affect-inducing vocal signals. Although speech acoustics can also elicit emotion in listeners, the affect-induction properties of sound have received little overall consideration in research examining vocal expression of emotion. Instead, both the cue-configuration and dimensional perspectives adopt the approach that emotional vocalizations carry affective information in coded form and that listeners decode the signals to recover that information. Conceptually, emotion expression is thus treated as an almost linguistic-like, representational communication system.

The affect-induction model of vocalization presents an alternative view. Originally proposed for nonhuman primate vocalizations, the approach has been shown to be applicable to human affective signaling, too. Rather than construing emotional communication as a process of encoding and decoding, the model argues that the primary function of nonlinguistic vocal signals is to change listeners' affect and thereby influence their behavior. Whereas representational accounts of communication implicitly rely on information encoding and decoding processes, the affect-induction approach argues that vocal signals can affect listeners by eliciting low-level affective responses that then influence listeners' behavioral decisions. The effects can be direct, meaning that signal acoustics itself has an impact, or indirect, meaning that listeners experience a learned affective response to a sound as a result of previous experience. For direct effects, acoustic salience is critical, with aspects such as F_0 range and amplitude being primary. For indirect effects, learning occurs through social interactions in which individually distinctive sounds produced by a particular vocalizer are paired with affect being experienced by the listener.

Comparisons between the affect-induction and representational perspectives can be illustrated by describing some of the similarities and differences between them. One important point of contact is that both approaches assume an association between signaler state and signal acoustics. The representational approach argues that this is a strong association, with the connection between a particular internal state and corresponding vocal signal being sufficiently specific that the listener can infer that it is occurring. The affect-induction approach also proposes that the vocalizer's internal state is critical in triggering signal production, but it proposes that the function is not to convey emotion so much as to induce emotion in the listener. As a result, signal features are not expected to be tightly linked to differentiated vocalizer states because it may benefit the signaler to induce similar responses in the listener across a variety of situations. Conversely, producing a diverse set of sounds in a given situation might serve a common function in influencing listener arousal and emotion. In this view, relationships between signaler state and physical signal are probabilistic rather than being highly specific.

A second point of contact is that listeners are able to draw inferences about vocalizer states or likely upcoming behaviors. From a representational perspective, such inferences are part and parcel of why communication occurs; listeners receive encoded information about vocalizer state, decode that information, and act on it. The affect-induction perspective instead views listener inference as a secondary outcome, with the primary function of vocalizing being to influence listener affect and behavior. Regardless of what particular mechanism may be involved, signaling must first and foremost benefit the signaler. Vocal behavior is therefore expected to show substantial variability, although general patterns will be discernible given that similar vocal strategies will be used by a variety of signalers. Listeners who attend to those patterns will therefore be able to draw some inferences about vocalizer states and likely behavior, both consciously and unconsciously. However, those inferences will be of a probabilistic, rather than highly specific, sort. A critical difference between the affect-induction and representational views thus concerns what is primary versus what is secondary to signal function.

The affect-induction perspective proposes that, although emotion-related signals are not representational *per se*, they can still be informative in the sense that listeners can draw inferences about their significance. Those inferences are based on a number of factors, including not only the acoustic attributes of the sound but also listener affective state and familiarity with both the signal and the vocalizer, as well as

the overarching context in which signaling is occurring. A corollary is that affective responses to highly similar sounds may be quite variable. For example, the same high-pitched, high-amplitude laugh might induce positive affect when heard at a party but elicit a strong negative response when heard while walking down a dark isolated street. In both cases, the acoustics of the loud laugh will probably capture the listener's attention, increase arousal, and thereby exacerbate whatever positive or negative affective state is already being experienced. In addition, if the sound has distinctive acoustic features that have previously been paired with either positive or negative affect in the listener, the sound will also activate corresponding learned responses. Finally, sounds can trigger more complex inferences, for instance, when hearing a loud laugh signals to the listener that the vocalizer is inebriated and will need a ride home.

Parallels between Speech and Laughter

Human laughter provides a good example of a nonlinguistic, emotion-related vocalization that is typically interpreted as a representational signal through which vocalizers convey information about their emotional states to listeners. Laugh acoustics is therefore expected to vary in accordance with variation in those emotions, which have been proposed to include emotional states as diverse as joy, amusement, anxiety, embarrassment, aggression, and submissiveness. There are as yet no data that specifically test this hypothesis, but the data that are available show laughter to be acoustically variable, even in situations involving only comparatively uniform positive states. Rather than being differentially associated with different emotions, laugh acoustics may be more closely associated with factors such as the sex and familiarity of the laughers' social partners. Males, for instance, laugh most and use acoustics likely to have the greatest direct impact when the listener is a friend, especially a male friend. Females, on the other hand, are more likely to produce high-impact sounds when the listener is a male than a female. The upshot is that both males and females appear to use laughter strategically, producing high-impact sounds when the likely effect on listeners would be positively toned arousal and using laughter rich in identity cues with friends who have learned positive affective associations with these sounds.

In both cases, F_0 emerges as a critical component of laughter's affect-induction effects. Males, for instance, report being more interested in female strangers when these women produce more laughter. Voiced laughter in particular has this effect, meaning that the laugh sounds are vowel-like, produced through regular vocal-fold vibration and with a prominent F_0. Noisy, unvoiced laughter that lacks vocal-fold vibration has little influence on listeners, even though this kind of laughter is extremely common. In perceptual testing, listeners consistently rate voiced laugh sounds as eliciting positive emotion, whereas unvoiced laughter is reported to have a slightly negative effect. F_0-related acoustics thus plays a crucial role in shaping listener responses to laughter, just as in emotional speech. In the affect-induction view, vocal-fold vibration and F_0 characteristics exert their impact in at least two ways. First, the F_0 component of voiced sounds is extremely salient, and a high or strongly modulated F_0 has direct impact on listener attention and arousal. Second, research on indirect learned responses to laughter has also highlighted the importance of voicing as being critical in cuing the vocalizer's identity.

Vocal Expressions of Emotion as Social Tools

Both speech and laughter are inherently social events and play critical roles in the communicative processes that are central to human relationships. There also appear to be some telling commonalities in how acoustic features of these two signals function in the emotional aspects of vocal communication. A central conclusion for speech is that F_0- and amplitude-related features play a primary role. On the one hand, speech produced while experiencing salient emotion shows significant changes in these particular characteristics, which also clearly impact listener perception and attributions of emotion. On the other hand, there is little indication that these or other acoustic features of speech can be linked to either discrete emotional states or valenced affect, and when listeners are asked to assess vocalizer states, performance based on speech acoustics alone is only moderate. Although less information is available for laughter, outcomes are similar. For example, the degree of variability occurring in laugh sounds produced during positive states alone argues against the possibility of emotion-specific laugh acoustics. Laugh sounds nonetheless have an unambiguous impact on listener emotion, with F_0, in particular, being found to play a central role.

It has thus been difficult to demonstrate differentiated, state-specific acoustic effects for either speech or laughter, but quite easy to show that acoustics has general and robust effects on listener emotions and evaluations of vocalizers. These results raise substantial questions about the adequacy of representational perspectives concerning emotion-related vocal signaling, and they are more compatible with the view that vocalizers use relatively undifferentiated acoustics that nonetheless has potent effects on listener affect.

The evidence is far from definitive, particularly because much of the research conducted on vocal expression of emotion in speech has assumed rather than tested the proposition that representational signaling is involved. It is therefore noteworthy that this research has largely failed to find compelling evidence of representational value in either production or perception. Although listeners are able to discern emotional content to a moderate degree, their judgments are constrained and context dependent. An alternative interpretation is that these inferences primarily reflect the listeners' own emotional responses, richly interpreted in the context of other about the vocalizer and the situation. Affective communication in both speech and laughter thus becomes a process of attributing emotion to vocalizers rather than one of recovering encoded information.

A larger implication of this view is that emotional expressions primarily function as nonconscious strategies of social influence rather than as veridical representations of internal state. Rather than informing, emotion-related signals are designed to shape the perceiver's affect and attitude in ways that ultimately benefit the signaler. The information value of a signal thus depends on a perceiver's previous experience, relationship with the signaler, and ability to incorporate signaling context.

However, much more data are needed to settle these issues than are currently available. It will be especially useful to focus on vocalizations acquired under controlled but naturalistic circumstances in which vocalizers are experiencing actual rather than simulated emotions. Further, it will be important to contrast the various alternatives as directly as possible within the same experimental framework, for instance, including comparing cue-configuration versus dimensional approaches and representational versus affect-induction perspectives. Finally, the scope of inquiry should be broadened to include the inferential processes that listeners use to make attributions about vocalizer states. Taking these steps will lead to a deeper understanding of how humans communicate through emotion-related vocal signals.

See also: Genetics of Human Anxiety and Its Disorders; Social Emotion: Neuroimaging; Speech Production: Adult; Speech Production: Development; Speech Perception: Adult; Speech Perception: Development; Speech Perception: Neural Encoding; Speech Perception: Cortical Processing.

Further Reading

Bachorowski J-A and Owren MJ (1995) Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science* 6: 219–224.

Bachorowski J-A and Owren MJ (2001) Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science* 12: 252–257.

Banse R and Scherer KR (1996) Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 70: 614–636.

Bargh JA and Chartrand TL (1999) The unbearable automaticity of being. *American Psychologist* 54: 462–479.

Dawkins R and Krebs JR (1978) Animal signals: Information or manipulation? In: Krebs JR and Davies NB (eds.) *Behavioral Ecology: An Evolutionary Approach*, pp. 282–309. London: Blackwell Scientific.

Johnson K (1997) *Acoustic & Auditory Phonetics*. Cambridge, MA: Blackwell.

Johnstone T and Scherer KR (2000) Vocal communication of emotion. In: Lewis M and Haviland J (eds.) *Handbook of Emotions*, 2nd edn., pp. 220–235. New York: Guilford.

Kappas A, Hess U, and Scherer KR (1991) Voice and emotion. In: Rime B and Feldman R (eds.) *Fundamentals of Nonverbal Behavior*, pp. 200–238. Cambridge, UK: Cambridge University Press.

Owren MJ and Bachorowski J-A (2003) Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior* 27: 183–200.

Owren MJ and Bachorowski J-A (2007) Measuring vocal acoustics. In: Coan J-A and Allen JB (eds.) *The Handbook of Emotion Elicitation and Assessment*, pp. 239–266. New York: Oxford University Press.

Owren MJ, Rendall D, and Bachorowski J-A (2003) Nonlinguistic vocal communication. In: Maestripieri D (ed.) *Primate Psychology*, pp. 359–394. Cambridge, MA: Harvard University Press.

Russell JA, Bachorowski J-A, and Fernandez-Dols J-M (2003) Facial and vocal expressions of emotion. *Annual Review of Psychology* 54: 329–349.

Scherer KR (1986) Vocal affect expression: A review and model for future research. *Psychological Bulletin* 99: 143–165.

Stevens KN (1998) *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Emotion Systems and the Brain

J LeDoux, New York University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In 1884, William James asked, "What is an emotion?" Although his answer was not universally accepted, the way James framed the problem continues to influence how we think about emotions today. Particularly important was James' distinction between the bodily expression of emotions and conscious emotional experiences or feelings.

James' goal was to explain feelings. But to do so, he felt he had to first dispel the common sense notion that feelings cause us to behave in certain ways. He argued that this was flat out wrong – in fact, backwards. In his view, the expression of an emotion in behavior and other physiological responses of the body actually determine what we feel. James' paper gave rise to two questions that have occupied emotion researchers and theorists ever since. One is, "how do emotional stimuli elicit emotional responses?" The other is, "what are feelings?" The first is a relatively straightforward empirical question. The second is considerably more complicated. In this overview of emotion concepts, this article begins by summarizing progress that has been made on these two questions. This article also discusses how the two questions sometimes get confused.

How Do Emotional Stimuli Elicit Emotional Responses?

James put his finger on a very important point when he rejected the idea that feelings cause emotional responses. Because different emotions (say, fear vs. joy) are elicited by different kinds of stimuli, the brain must have already determined the emotional nature of the situation before the feeling or the response occurs. James did not dwell on this, since he wanted to know what causes feelings. But implicit in his formulation was the notion that emotional responses are elicited on the basis of unconscious processing of emotional stimuli, since the conscious aspect of emotion, the feeling, does not emerge until the responses are expressed. Regardless of whether you accept James' theory, unless you accept the common sense view, that feelings cause responses, unconscious emotional processing must be the basis of the responses. The following sections, outline how emotional stimuli, via unconscious processes, lead to the expression of emotional responses. Then, research aimed at revealing the brain mechanisms is summarized through which emotional responses are controlled by external stimuli.

Unconscious Emotional Processing

In the first half of the twentieth century, behaviorists eliminated 'consciousness' from psychology and only allowed the description of psychological processes in terms of observable behavior. With the rise of cognitive science, the possibility arose for viewing psychological functions in terms of internal representations, without equating those internal processes with consciousness. That is, psychological functions involve information processing. Such processing could, but does not have to, lead to conscious experiences. Information processing, in other words, is by definition unconscious. This notion was originally about cognitive processes such as perception and memory, but it was subsequently extended to emotional processing.

The processing of emotional stimuli is typically referred in psychology to as 'evaluative processing,' or appraisal. Appraisal is said to underlie both the emotional responses and the conscious feelings that occur in the presence of emotional stimuli. Nevertheless, appraisal theorists have traditionally focused on the role of appraisals in feelings, and have viewed appraisals in terms of higher cognitive processes. However, as we will see, basic evaluative or appraisal mechanisms that operate unconsciously are also involved in the elicitation of emotional responses by external stimuli.

Everyone has, in some situation, found his or her heart pounding before recognizing the nature of the eliciting event. Considerable scientific evidence also supports the idea that unconsciously processed stimuli can control emotional responses. This does not mean that conscious appraisal never occurs. It simply means that conscious appraisal is not necessary for the elicitation of emotional responses. The idea that emotional responses are based on unconscious appraisal processes is actually of tremendous value from the point of view of brain research. It allows us to do research on emotional behavior without having to first solve the mind–body problem, and allows us to treat the control of emotional responses in animals and humans similarly.

There are two basic ways in which emotional stimuli can elicit emotional responses by activating appraisal mechanisms. Some appraisals are hardwired by evolution and others are learned through individual experience. All organisms, including humans, have innate appraisal capacities that help ensure their survival. Stimuli that signal sexual partners, palatable

and poisonous foods, painful stimulation, predators, or aggressive members of one's own species are common examples. Such stimuli activate innate appraisal mechanisms and lead directly to the expression of species-typical responses and accompanying autonomic responses. For example, a rat that encounters a cat for the first time will express signs of fear, even if it was reared in a laboratory and never previously exposed to a cat. When organisms encounter stimuli that innately activate appraisal mechanisms they often form associations (through Pavlovian conditioning) with other stimuli present, since these may be predictive of harm or benefit in the future. These conditioned stimuli acquire the capacity to elicit species-typical emotional responses.

Species-typical behavioral and autonomic responses are often just the first stage of emotional responding to an innate or conditioned emotional stimulus. In addition to these reactions, organisms also often perform instrumental actions. Avoidance behavior is a canonical example of a learned aversive instrumental action. Thus, when you encounter an innate or conditioned danger signal, your first reaction may be to freeze. But you may also try to escape from or avoid the danger. Sometimes avoidance of danger becomes pathological, as when a patient with panic disorder stays inside her home in order to reduce the chances of having an attack. Positive instrumental responses are also important. We seek stimuli and situations that lead to desirable outcomes. But these too can have pathological consequences. Eating disorders and drug addiction are common examples.

Instrumental responses are often learned through trial-and-error processes in which the probability of the response changes depending on the rewarding or punishing consequences of the behavior. Once learned, these are performed as habits. However, instrumental responses can also be based on spontaneous decisions; cognitive estimation of outcomes eliminates the need for trial-and-error leaning.

In summary, the brain unconsciously appraises the value of external stimuli. Some stimuli are significant because of innate wiring, while the significance of others is learned through experience. Stimuli that are appraised as significant then lead to the expression of responses. Both innate and learned stimuli can lead to the expression of hard-wired species-typical emotional responses. These reactions are then often followed by instrumental responses that reflect past learning or spontaneous decisions.

Brain Mechanisms of Emotional Processing

William James not only speculated about the nature of emotion, he also proposed the first modern brain theory of emotion. By the late nineteenth century researchers had discovered sensory and motor regions in the cortex. James therefore asked whether an emotion 'center' was yet to be discovered or whether it might be possible to explain emotions as functions of cortical sensory and motor areas. He preferred the latter idea, arguing that emotional stimuli elicit emotional responses by way of sensory cortex activation of motor cortex. This idea was then challenged by Walter Cannon on the basis of studies showing that emotional reactions were preserved in decorticate animals, but were disrupted after damage to the hypothalamus. On the basis of these and other observations, Papez, in 1937, proposed a circuit theory of emotion involving the hypothalamus, anterior thalamus, cingulate gyrus, and hippocampus. Later, MacLean then named the structures of the Papez circuit, together with several additional regions (amygdala, septal nuclei, orbitofrontal cortex, portions of the basal ganglia), the limbic system, which he viewed as a general-purpose system involved in behaviors that ensure survival of the individual and the species.

MacLean's writings were very persuasive and for many years the problem of relating emotion to brain mechanisms seemed solved at the level of anatomical systems. However, the limbic system concept is now believed to suffer from imprecision at both the structural and the functional levels. For example, it has proved impossible to provide unequivocal criteria for defining which structures and pathways should be included in the limbic system. A standard criterion is connectivity with the hypothalamus. However, because studies have shown that the hypothalamus is connected with structures at all levels of the central nervous system, from the neocortex to the spinal cord, the limbic system would occupy much of the brain. Further, classic limbic areas, such as the hippocampus and mammillary bodies, have proved to be far more important for cognitive processes, such as declarative memory, than for emotional processes.

The post-Jamesian theories of the emotional brain were fairly consistent in their view of how sensory stimuli lead to the expression of emotional responses. Sensory information transmitted from the thalamus and/or cortex to specialized subcortical emotion circuits regulates the expression of emotional responses. In these theories, the subcortical circuits were centered on the hypothalamus. Electrical stimulation studies provided strong support for the role of the hypothalamus in the control of emotional responses, including autonomic responses.

Today, the emphasis has shifted from the hypothalamus to the amygdala, especially in the control

of the bodily responses that occur during fear reactions in both animals and humans. The animal studies indicate that one region of the amygdala, the lateral nucleus, receives sensory information from the thalamus and cortex. It then connects with the central amygdala, which in turn connects with lower areas, including the hypothalamus, that then connect with specific motor neuron groups that control emotional responses. Circuits within the amygdala in essence mediate the appraisal of danger, including learning about stimuli associated with danger. The amygdala has also been implicated in appraising positive emotional stimuli, but the circuit details are less well understood.

The scenario just described mainly applies to aversive emotional responses that are elicited automatically in response to a stimulus. But, as noted earlier, in addition to reacting in an automatic fashion to emotional events, organisms also perform instrumental responses. The amygdala is also involved in instrumental responses, but in a different way than for automatically elicited responses. Specifically, the central amygdala and its connectivity with lower brain regions is not involved. Instead, connections from the lateral nucleus to the basal amygdala, and from there to the striatum, seem to be important.

A prominent idea in recent years has been that the prefrontal cortex, especially the ventromedial prefrontal cortex, interacts with the amygdala in the regulation of emotional responses. The amygdala is often thought of as an emotional accelerator and the prefrontal cortex as the brakes. Prefrontal cortex is also likely to play an important role in the decision-making processes through which instrumental behavior is regulated in emotional situations.

In summary, much has been learned about the neural circuits through which sensory stimuli come to elicit emotional responses. Progress has been made for both negative and positive emotions, and for both automatic and instrumental responses.

What Are Feelings?

As noted previously, what William James really cared about was the nature of feelings. And throughout the following century, up to this day, there has been much debate about just what feelings are, both in psychology and brain science.

The Psychology of Feelings

After several decades of acceptance, James' theory of feelings was dethroned by Walter Cannon. He rejected the idea that feelings are due to feedback from responses, since, he said, the bodily responses would be too slow and too vague to define the difference between fear and joy and love and grief.

The debate stalled for decades, but took a new turn in the 1960s when Stanley Schachter and Jerome Singer proposed that both James and Cannon were right. Cannon was right that body arousal is not specific enough. James was right that body arousal was important. The solution, they said, was that through cognitive processes we interpret nonspecific body arousal in light of the context which exists at the moment. Later theories eliminated the need for bodily responses at all, emphasizing the importance of cognitive processing.

By the 1980s, emotion had become a full-fledged cognitive process. Some, such as Robert Zajonc, felt this went too far. On the basis of the results of clever experiments he performed, Zajonc argued that feelings are the products of unconscious affect. Because he objected to the cognitive take over of emotion, he avoided the use of the term unconscious processing. Nevertheless, his work made the idea that conscious feelings might be based on unconscious appraisal palatable.

Feelings and the Brain

The various aforementioned theories of the emotional brain each proposed that feelings are products of information processing in the cerebral cortex. For Cannon, feelings occurred in this way: sensory information split at the thalamus, with some going to the hypothalamus and some to the cortex. The hypothalamus then communicated with the cortex. When hypothalamic information interacted with sensory information in the cerebral cortex, a feeling was the result. Papez had a similar theory, but with more anatomical specificity, proposing that feelings occurred when the cingulate cortex integrated external sensory information with information processed by the hypothalamus. For MacLean, feelings resulted in the limbic system when the hippocampal cortex integrated information from the internal and external environments.

After several decades of dormancy, research and theory on the brain mechanisms of emotion began to pick up in 1980s. For the most part, this work was focused on emotional learning and the control of emotional behavior. However, attention was also returning to questions about the nature of feelings. William James, you'll recall, proposed that feelings result when the cortex perceives body sensations produced during an emotional reaction. Thus, fear feels different from love because these two emotions have different effects on the body. Although James emphasized the entire bodily response, critics such

as Cannon focused on the autonomic nervous system response. While autonomic responses are indeed slow and nonspecific, feedback from the muscles through proprioception can be fast and specific. A contemporary theory that emphasizes such responses is the facial feedback hypothesis. This theory proposes that the specific facial responses associated with certain emotions provide feedback to the brain that determines what we feel. Another feedback theory was proposed in 1994 by Antonio Damasio. Building on James, Damasio argued that we use 'gut feelings' to figure out what we are feeling, and that this information plays an important role in the way we make decisions. For example, when trying to decide between two courses of action, we play these out in our minds and the body sensations that result in one versus the other, according to Damasio, help us decide which to choose (which one 'feels' better). According to Damasio, the orbitofrontal cortex and the body-sensing regions of the cortex (somatosensory and insular regions) are involved in feeling states.

From authors viewpoint, feedback from the body such as that described by James and Damasio adds intensity and duration to emotional experiences, but the quality of the experience is a product of working memory. Because working memory is believed to be the cognitive operation that makes consciousness possible, it may be the basis of conscious feelings as well as cognitive states of consciousness. The basic idea is that we have one mechanism of consciousness, and qualitatively different states of consciousness are defined by the kind of information that we are conscious of at the moment. Through working memory we are able to integrate the way something looks and sounds with past memories of this thing, and with the way this thing is making us feel at the moment. If you encounter a rattlesnake on a path, for example, the following kinds of information will likely come to be present in working memory: a visual representation of the snake, an auditory representation of the rattling sound, information retrieved from long-term memory (things known about snakes as well as experiences with them), and information about the state of your brain (the fact that a certain emotion system is activated and that you are aroused). These will combine to produce an integrated representation in working memory of the entire experience. The cognitive appraisal of this experience in working memory is the initial feeling. At the same time, bodily responses, which are slower to occur, are expressed. These then feed back to the brain and add to the feeling that is evolving. The difference between an emotional conscious experience and a nonemotional one is simply that working memory is working with additional information in the case of the emotional experience – namely, information about the emotional aspects of the experience. The difference between different kinds of emotional experiences (fear vs. joy) is due to the particular emotion system that is active at the moment and monopolizing consciousness.

Considerable evidence implicates the dorsolateral prefrontal cortex (PFC) in working memory. While the brain networks of working memory were once narrowly viewed as mainly involving the dorsolateral PFC, this is no longer the case. The various medial areas of PFC, including the orbital cortex and anterior cingulate, as well as insular cortex, are now viewed as being involved, especially in emotional situations. This broader view of working memory overlaps with Damasio's somatic hypothesis but without placing too much burden on peripheral feedback.

Can Responses Tell Us about Feelings?

A final issue to consider is the relation of responses and feelings. This issue has two parts. One has already been discussed: do responses cause feelings? James and Damasio emphasize this view. Others take issue with it. But the part that we want to consider here is this: can responses tell us what other organisms are feeling?

Some years ago, Nagel pointed out that each species is likely to have its own kind of subjective experience of world. Because working memory plays a key role in human conscious awareness of the world, the brain mechanisms of working memory may give us clues about the kinds of experiences different organisms can have. Interestingly, the brain regions involved in working memory are more highly developed in humans than in other primates and more highly developed in other primates than in other mammals. Adding in the fact that language is likely to change the way experiences are interpreted (and appraised), and that only humans have natural language, we are led to the conclusion that whatever kinds of experiences other organisms have, they are likely to be distinct from those typical of humans.

The bottom line is that two organisms may have identical behavioral and autonomic responses but have completely different conscious experiences. You can trust to some extent that when another human is acting fearful or joyful they are likely to be feeling what you refer to as fear or joy. The reason we can trust this conclusion is because we have the same basic brain plan as do other members of our species. We can therefore take what we know from our own experiences and apply it to others. We do this all the time in social interactions. People also often do this for pets and other animals. The reason this 'anthropomorphic' reasoning is inappropriate was

explained earlier: because the brain mechanisms of consciousness are likely to be different in different organisms, mainly because of the development of the prefrontal cortex in primates and language in humans, conscious experiences are likely to be different.

Conclusion

This brief survey, much of it including historical material, provides an overview of some of the major concepts of emotion that are relevant to the pursuit of brain mechanisms of emotion. Many of these concepts have been explored in depth in the large body of literature on emotion theory.

See also: Aggression: Hormonal Basis; Aggression: Neurochemical and Molecular Mechanisms; Amygdala: Contributions to Fear; Aversive Emotions: Molecular Basis of Unconditioned Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Emotion and Vigilance; Emotion in Speech; Emotion: Neuroimaging; Emotion: Computational Modeling; Emotional Disorders: Treatment; Emotional Hormones and Memory Modulation; Emotional Influences on Memory and Attention; Emotional Control of the Autonomic Nervous System; Genetics of Human Anxiety and Its Disorders; Social Emotion: Neuroimaging.

Further Reading

Aggleton JP and Mishkin M (1986) The amygdala: Sensory gateway to the emotions. In: Plutchik R and Kellerman H (eds.) *Emotion: Theory, Research and Experience*, vol. 3, pp. 281–299. Orlando: Academic Press.

Amorapanth P, LeDoux JE, and Nader K (2000) Different lateral amygdala outputs mediate reactions and actions elicited by a fear-arousing stimulus. *Nature Neuroscience* 3: 74–79.

Arnold MB (1960) *Emotion and Personality*. New York: Columbia University Press.

Baars BJ (1997) In the theatre of consciousness. *Journal of Consciousness Studies* 4: 292–309.

Bandler RJ (1982) Induction of 'rage' following microinjection of glutamate into midbrain but not hypothalamus of cats. *Neuroscience Letters* 30: 183–188.

Bard P (1929) The central representation of the sympathetic system: As indicated by certain physiological observations. *Archives of Neurology and Psychiatry* 22: 230–246.

Blanchard CD and Blanchard RJ (1972) Innate and conditioned reactions to threat in rats with amygdaloid lesions. *Journal of Comparative Physiological Psychology* 81: 281–290.

Brodal A (1982) *Neurological Anatomy*. New York: Oxford University Press.

Cannon WB (1929) *Bodily Changes in Pain, Hunger, Fear, and Rage*. New York: Appleton.

Cohen NJ and Eichenbaum H (1993) *Memory, Amnesia, and the Hippocampal System*. Cambridge, MA: MIT Press.

Damasio A (1994) *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Gosset/Putnam.

Damasio AR and van Hoesen G (1983) Emotional disturbances associated with focal lesions of the limbic frontal lobe. In: Heilman KM and Satz P (eds.) *Neuropsychology of Human Emotion*, pp. 85–110. New York: Guilford.

Davis M and Whalen PJ (2001) The amygdala: Vigilance and emotion. *Molecular Psychiatry* 6: 13–34.

Ekman P (1993) Facial expression and emotion. *American Psychologist* 48: 384–392.

Erdelyi MH (1985) *Psychoanalysis: Freud's Cognitive Psychology*. New York: Freeman.

Everitt BJ, Parkinson JA, Olmstead MC, et al. (1999) Associative processes in addiction and reward. The role of amygdala-ventral striatal subsystems. In: McGintry J (ed.) *Advancing from the Ventral Striatum to the Extended Amygdala*, pp. 412–438. New York: New York Academy of Sciences.

Fanselow MS and Poulos AM (2005) The neuroscience of mammalian associative learning. *Annual Review of Psychology* 56: 207–234.

Frijda N (1986) *The Emotions*. Cambridge: Cambridge University Press.

Gray JA (1982) *The Neuropsychology of Anxiety*. New York: Oxford University Press.

Hilton SM and Zbrozyna AW (1963) Amygdaloid region for defense reactions and its efferent pathway to the brainstem. *Journal of Physiology* 165: 160–173.

Holland PC and Gallagher M (1999) Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences* 3: 65–73.

Irwin W, Davidson RJ, Lowe MJ, et al. (1996) Human amygdala activation detected with echo-planar functional magnetic resonance imaging. *NeuroReport* 7: 1765–1769.

Isaacson RL (1982) *The Limbic System*. New York: Plenum.

Izard CE (1971) *The Face of Emotion*. New York: Appleton-Century-Crofts.

James W (1884) What is an emotion? *Mind* 9: 188–205.

James W (1890) *Principles of Psychology*. New York: Holt.

Kaada BR (1960) Cingulate, posterior orbital, anterior insular and temporal pole cortex. In: Field J, Magoun HJ, and Hall VE (eds.) *Handbook of Physiology: Neurophysiology II*, pp. 1345–1372. Washington, DC: American Physiological Society.

Kapp BS, Pascoe JP, and Bixler MA (1984) The amygdala: A neuroanatomical systems approach to its contributions to aversive conditioning. In: Buttlers N and Squire LR (eds.) *Neuropsychology of Memory*, pp. 473–488. New York: Guilford.

Kihlstrom JF (1987) The cognitive unconscious. *Science* 237: 1445–1452.

Lazarus R and McCleary R (1951) Autonomic discrimination without awareness: A study of subception. *Psychological Review* 58: 113–122.

Lazarus RS (1984) On the primacy of cognition. *American Psychologist* 39: 124–129.

LeDoux JE (1984) Cognition and emotion: Processing functions and brain systems. In: Gazzaniga MS (ed.) *Handbook of Cognitive Neuroscience*, pp. 357–368. New York: Plenum.

LeDoux JE (1987) Emotion. In: Plum F (ed.) *Handbook of Physiology. 1: The Nervous System. Volume V, Higher Functions of the Brain*, pp. 419–460. Bethesda: American Physiological Society.

LeDoux JE (1996) *The Emotional Brain*. New York: Simon and Schuster.

LeDoux JE (2000) Emotion circuits in the brain. *Annual Review of Neuroscience* 23: 155–184.

LeDoux JE (2002) *Synaptic Self: How Our Brains Become Who We Are*. New York: Viking.

MacLean PD (1949) Psychosomatic disease and the 'visceral brain': Recent developments bearing on the Papez theory of emotion. *Psychosomatic Medicine* 11: 338–353.

MacLean PD (1952) Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain). *Electroencephalography and Clinical Neurophysiology* 4: 407–418.

Maren S (2001) Neurobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience* 24: 897–931.

Mishkin M and Aggleton J (1981) Multiple functional contributions of the amygdala in the monkey. In: Ben-Ari Y (ed.) *The Amygdaloid Complex*, pp. 409–420. Amsterdam: Elsevier/North-Holland Biomedical Press.

Mowrer OH (1956) Two-factor learning theory reconsidered, with special reference to secondary reinforcement and the concept of habit. *Psychological Review* 63: 114–128.

Nader K and LeDoux JE (1997) Is it time to invoke multiple fear learning system? *Trends in Cognitive Sciences* 1: 241–244.

Nagel T (1974) What is it like to be a bat? *Philosophical Review* 83: 4435–4450.

Neisser U (1967) *Cognitive Psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Norman DA and Shallice T (1980) Attention to action: Willed and automatic control of behavior. In: Davidson RJ, Schwartz GE, and Shapiro D (eds.) *Consciousness and Self-Regulation*, pp. 1–18. New York: Plenum.

Ohman A and Mineka S (2001) Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review* 108: 483–522.

Ono T and Nishijo H (1992) Neurophysiological basis of the Kluver–Bucy syndrome: Responses of monkey amygdaloid neurons to biologically significant objects. In: Aggleton JP (ed.) *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*, pp. 167–190. New York: Wiley-Liss.

Ortony A, Clore GL, and Collins A (1988) *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Panksepp J (1982) Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences* 5: 407–467.

Panksepp J (1998) *Affective Neuroscience*. New York: Oxford University Press.

Papez JW (1937) A proposed mechanism of emotion. *Archives of Neurology and Psychiatry* 79: 217–224.

Phelps EA and LeDoux JE (2005) Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* 48: 175–187.

Quirk GJ, Russo GK, Barron JL, et al. (2000) The role of ventromedial prefrontal cortex in the recovery of extinguished fear. *Journal of Neuroscience* 20: 6225–6231.

Rolls ET (1999) *The Brain and Emotion*. Oxford: Oxford University Press.

Schachter S (1975) Cognition and centralist–peripheralist controversies in motivation and emotion. In: Gazzaniga MS and Blakemore CB (eds.) *Handbook of Psychobiology*, pp. 529–564. New York: Academic Press.

Schachter S and Singer JE (1962) Cognitive, social, and physiological determinants of emotional state. *Psychological Review* 69: 379–399.

Shallice T (1988) Information processing models of consciousness. In: Marcel A and Bisiach E (eds.) *Consciousness in Contemporary Science*, pp. 305–333. Oxford: Oxford University Press.

Siegel A and Edinger H (1981) Neural control of aggression and rage behavior. In: Morgane PJ and Panksepp J (eds.) *Handbook of the Hypothalamus, Vol. 3: Behavioral Studies of the Hypothalamus*, pp. 203–240. New York: Dekker.

Sotres-Bayon F, Cain CK, and Ledoux JE (2006) Brain mechanisms of fear extinction: Historical perspectives on the contribution of prefrontal cortex. *Biological Psychiatry* 60: 329–336.

Squire LR (1987) Memory: Neural organization and behavior. In: Plum F (ed.) *Handbook of Physiology, Section 1: The Nervous System. Vol. V: Higher Functions of the Brain*, pp. 295–371. Bethesda: American Physiological Society.

Swanson LW (1983) The hippocampus and the concept of the limbic system. In: Seifert W (ed.) *Neurobiology of the Hippocampus*, pp. 3–19. London: Academic Press.

Whalen PJ, Rauch SL, Etcoff NL, et al. (1998) Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience* 18: 411–418.

Zajonc RB (1984) On the primacy of affect. *American Psychologist* 39: 117–123.

Emotion: Computational Modeling

J-M Fellous, University of Arizona, Tucson, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Understanding the neural basis of human emotions is difficult because emotions are, to a large extent, subjective and nondeterministic. The same stimulus may create different emotions in different individuals, and the same individual may express different emotions in response to the same stimulus, at different times. However, in spite of this variability, it is assumed that there are basic principles, perhaps even basic neural mechanisms, that make a particular event 'emotional.' To find these principles and their underlying mechanisms, neuroscientists typically study specific emotions, using specific tasks. As is appropriate, they use a combination of animal and human preparations, yielding various types of data, from single neuron firing patterns, to activation levels of a whole brain area. The approach, while rigorous, is slow and yields an increasingly complex body of often conflicting data. An integrative approach is needed. Computational models of emotion have emerged as a promising tool for integration. Because they require that all assumptions be made explicit, they offer a new language in which to express and test hypotheses, as well as explain and predict neural mechanisms.

It is paradoxically the advances in cognitive neuroscience that have allowed for the emergence of emotion research. The paradox is in the fact that usually, the 'cognitive' is opposed to the 'affective.' We now know from experimental and modeling work that both cognition and emotion develop and work hand in hand. Fortunately, as a result of the cognitive neuroscience push to understand emotion, the techniques used in that field have carried over to emotion research. Models once used to understand cognitive functions, such as perception or memory, are now extended and sometimes fundamentally modified, to address issues pertaining to emotion.

The past 10 years has seen an explosion of artificial intelligence (AI) models that have directly addressed issues on the role of emotion. Though generally not neurally motivated, these models have emphasized several important functional aspects of emotions, such as their role in communication (human–machine interface), in decision making, in general body homeostasis, and in environmental adaptation (in autonomous robots). Several books have been published, one of the first being that of Rosalind Picard, *Affective Computing*. Artificial intelligence focuses on basic principles, without necessarily the need to find neural mechanisms that support them (e.g., chess playing). Neural modeling, on the other hand, focuses on mechanisms and hopes to emerge basic computational principles. The elucidation of the underlying mechanisms is powerful in that it explains how a system works and fails, and predicts the behavior of the system under conditions not yet tested experimentally. This aspect of modeling is particularly interesting for the understanding and treatment of emotional disorders; the focus here is on such neural modeling approaches, leaving AI approaches for further reading.

Building neurally motivated models of emotions or specific aspects of emotion processing can be achieved at many levels of abstractions. At the one end of the spectrum are models that implement a specific neural pathway, involving specific neural populations. These biophysical models typically describe neurons at the level of individual action potentials. Their explanatory and predictive power comes from the emergence of interesting dynamical network behaviors (e.g., oscillation, competition between neuronal pools) on millisecond timescales. At the other end are models that are mainly concerned with neural processing principles, where emphasis is brought on the connectivity and level of activation of typical neurons, or groups of neurons. They are primarily designed on the basis of anatomical and imaging data. These connectionist models can effectively give a larger view of the computations achieved by a set of brain areas. The vast majority of connectionist models implement some form of learning, whereby interesting network behavior emerges from the reconfiguration of the synaptic connection patterns. They can often explain and predict neural activity on behavioral time scales. There is of course a wide range of hybrid models in between those two extremes, as will be seen in the following discussion.

There has been a flurry of experimental work producing data on various aspects of the neuroscience of emotions in the past few decades. These data have given rise to many anatomical models, and have been reviewed by Tim Dalgleish. However, computational models of emotions have been possible because of the emergence of complementary mechanistic theories of emotions. Four of those conceptual models (as seen in the work of Joseph LeDoux, Edmund Rolls, Antonio Damasio, and Klaus Sherer) are at the basis of several neurally inspired models of emotions. Negative emotions are approached through the modeling of various aspects of fear processing in the amygdala. Positive

emotions are addressed through the study the neural substrate of reward processing and decision making.

Fear and the Amygdala

According to LeDoux's conceptual model of auditory fear conditioning, the amygdala neurons receive fast and broadly tuned inputs from the thalamus, and slower and more refined inputs from the cortex. Armony and co-workers built a computational model of auditory fear processing whereby synaptic connections were modified so that each amygdala neuron 'learned' to represent a restricted range of tone frequencies; together, the neurons cover the audible spectrum. Interestingly, when the learning was conducted under conditions in which fear was associated with certain tones but not with others, a subset of neurons retuned and shifted their preference to those frequencies that were paired with fear elicitation. This fear-driven remapping of neural tone preference was later confirmed experimentally by recording from the amygdala of animals trained to associate tones and fearful electrical footshocks. Surprisingly, the model also showed that the removal of cortical inputs to the amygdala cells did not affect their stimulus preference, as would be expected from computational units that have access to less incoming information. This result, again confirmed experimentally by lesion studies, predicted that noncortical inputs (thalamic in origin) may have more influence than previously expected. The model, constrained by experimental data, was therefore able to make experimentally testable, and later confirmed, predictions about fear processing, and lent support to the notion of dual routes of fearful stimulus processing.

Other more abstract models using genetic algorithms and artificial neural networks showed how these fast and slow routes of processing could emerge from the necessity for survival when discrimination between predator and prey (i.e., positive and negative 'emotions') was made difficult by environmental circumstances.

Beyond its involvement in fear, the amygdala has also been modeled as a component of the general motivational system. One of Stephen Grossberg's models was designed to explore the cognitive–emotional interactions at play during various learning paradigms. This model included a sensory (visual) perceptual system that consisted of the thalamus, primary visual cortices, and temporal lobes, a motivational system that represented the activity of the hypothalamus and amygdala, and a generic motor system that included overall cortical and cerebellar motor outputs. The amygdala acted as a motivational bias to the sensory stream, and through its interaction with the prefrontal cortex influenced action selection. The activity of the amygdala relied on the dynamical interplay between opponent emotional representations such as fear and relief. With the proper tuning, arousal-like inputs to these representations could lead to a characteristic inverted U-shaped behavioral response, whereby too little or too much arousal decreased motor output. Mechanistically, the model predicted that a large arousal input may yield a depression of amygdala activity that in turn caused a decrease in prefrontal cortex activation (hypofrontality). This chain of event lead to the emergence of negative symptoms characteristic of schizophrenia, such as flat affect (modeled as low amygdala activation), decrease of motivation (modeled as low prefrontal activity), and attentional deficits (modeled as a lack of prefrontal cortex control on sensorimotor information flow).

The idea of emotion as a component of the motivational system, and of emotional behavior as a result of the competition between different motivational systems, is also at the basis of a connectionist model of anxiety in rats. In the Salum model, spatial exploration was the result of a dynamical conflict between exploratory behavior (curiosity) and avoidance for open/high places (anxiety generating). While based on experimental behavioral data obtained in rats, the model did not, however, explicitly address the role of the amygdala, but rather, as in Grossberg's model, illustrated the basic principle of dynamical competition between conflicting motivational biases.

Other connectionist-like models by Taylor and Fragopanagos have attempted to clarify the link between emotion and attention. Their models relied on a fast, preattentive, processing route inspired by LeDoux's conceptual model. The amygdala evaluated the negative nature of incoming stimuli, and exerted a gain enhancing, excitatory effect on sensory processing which moved the focus of attention on the emotional stimulus (bottom-up attentional modulation). Alternatively, and perhaps complementarily, the focus of attention could also be modified by an amygdala-triggered interaction between ventromedial and dorsolateral prefrontal cortices modules that eventually changed the active cognitive goal (top-down attentional modulation). This model was used to explain various emotion–attention experiments and confirmed the view that the amygdala may have a fast, subattentive modulatory role in attention through both the production of saliency and the cognitive control of behavior. This model, as with that of Grossberg, points to a crucial interaction between the amygdala and prefrontal cortices. This interaction

has been further studied in the context of reward processing.

Reward Processing and the Prefrontal Cortex

The involvement of the orbitofrontal cortex (OFC) in emotion relies on a large body of data on reinforcement learning. For Rolls, this type of learning is at the basis of the processing of emotions in general, and the orbitofrontal cortex is in charge of decoding the reinforcement value of the stimuli and the rapid associations and reassociations between the stimuli and reinforcers. His experiments have shown that normal OFC function mediated one-trial reversal of stimulus–reward association. This phenomenon is mechanistically problematic because its time course is incompatible with learning mechanisms such as typically proposed by connectionist models, whereby synaptic modifications occur after multiple stimulus–response associations (i.e., training). Deco and Rolls' model needed to account for fast time scales, and had therefore to rely on spiking (biophysical) neurons. The model was based on 'biased competition' between pools of explicitly simulated neurons. A particular stimulus–reward association was represented by a pool of actively spiking neurons, and, interestingly, a change in association was simply a change in which pool of neuron fired (and not a change in synaptic strength). The change was dynamically achieved by a small input bias coming from a specialized OFC group of neurons called 'rule-specific' neurons. These neurons were context dependent and allowed for the same object to be perceived as rewarding or punishing, depending on the task contingency (context). They responded to punishment or the nondelivery of an expected reward by delivering an inhibitory bias to the currently active neuronal pool (a particular stimulus–reward association). When inactivated, the pool became momentarily refractory, and a new pool (a new stimulus–reward association) emerged rapidly, in one trial. This model made a number of testable predictions, one of which was the existence of specialized pools of OFC neurons that fired when a particular object–value combination (e.g., rewarded square, but not punished square) was presented. Such neurons may be used for general emotional reactions to a variety of cognitive contexts. Some experimental evidence indeed points to the existence of such neurons.

The idea that emotional signals bias the activity of competing active circuits is also at the basis of Damasio's somatic marker hypothesis. According to this theory, the perception of ones' own action in response to a particular stimulus has the potential to activate emotional memories that were previously associated with that stimulus. These memories ('somatic markers') in turn bias further cognitive processing, decision making, and action selection mechanisms. Several brain areas are known to be involved in this type of emotional bias, and special attention has been drawn to the ventromedial prefrontal cortex (VMPFC). The VMPFC is part of the executive system of the brain whereby actions are selected and plans are made. Because of its connection to the amygdala, the VMPFC also contains associations (the 'somatic markers') between previous actions and their expected consequences ('emotional' in nature). In the Wagar and Thagard model, when a stimulus is perceived, several potential actions are activated in the VMPFC. Some of these actions are part of somatic markers, and activate amygdala-dependent representations that 'reenact' the bodily state predictive of the action's consequences. These unconscious representations (representing emotional valence) in turn bias decision-making processes. The model further proposes that this bias depends on the context (hippocampus activation) in which the stimulus is perceived, and that it is gated by the nucleus accumbens.

The nucleus accumbens receives inputs from the VMPFC, the amygdala, and the hippocampus, and is thought to be key to high-level cognitive processes and action selection. However, because the nucleus accumbens is generally inhibited by dopamine (from the ventral tegmental area), these inputs are not sufficient to make it fire. The model showed that firing occurred only if the amygdaloid inputs (appraisal of a stimulus) arrived first, followed by a VMPFC input (a possible action to perform), and provided that the hippocampus is active throughout (the correct contextual information). In a simulated 'Iowa gambling' task, the model correctly showed that in normal conditions, choices are made according to long-term predicted outcomes (VMPFC activity), while in VMPFC lesions conditions (as in the case of the famous Phineas Gage), choices are made instead on the basis of immediate outcomes (amygdala activity). In the specific case of positive and negative stimulus evaluation taken as different 'emotional states,' this model offered insights on the role of the nucleus accumbens in the integration of cognitive, contextual, and emotional information flow.

One major issue in both the Deco and Rolls and the Wagar and Thagard model is that learning is assumed. The connectivity patterns among the various components of the model are configured 'manually' and emphasis is brought on the dynamical aspects of the network. Frank and Claus attempted to explicitly model the learning phase, and their model focused on the interactions between OFC and the

dopaminergic system of the basal ganglia. They showed that the probability of reward/punishment was likely to be encoded by the basal ganglia. In contrast, the variations in the magnitude of reward and punishment were both encoded by the amygdala and actively maintained in the medial and lateral portions of the OFC, respectively. The model was tested on the Iowa gambling task, reversal learning, and devaluation paradigms. They explained various aspects of decision-making deficits in OFC-damaged patients, Parkinson's disease patients, and in drug abusers, and offered a partial explanation for individual differences in risk-averse and risk-seeking normal individuals.

Modeling Emotions

Mainly because of the availability of experimental data, the focus of neural models of emotions has been on fear and reward processing; there have been relatively few attempts at using neural networks to model emotions in general. One step in this direction was taken by Scherer and co-workers, on the theoretical basis that emotions may be understood as the result of a complex sequence of appraisals. According to Scherer, emotions have five major components: a cognitive component (rational evaluation of objects and events), a peripheral efference component (homeostasis, body regulation), a motivational component (preparation for action), a motor component (expression of emotion, communication to others), and a subjective 'feeling' component (monitoring of emotional expressions). Formulated in this framework, an emotion is the dynamical coactivation of these components in response to the evaluation (by, e.g., the amygdala) of an internal or external stimulus that is particularly important to the subject. The architecture of the model is that of a feed-forward network, with lateral connections. The stimulus activates a set of interacting representational units that are specialized in the recognition of predefined aspects of the stimulus, such as its novelty, intrinsic pleasantness, or outcome probability. Together these units determine the significance of the stimulus, and activate a second layer of interacting units that prepare the organism for action. These units assess the relevance of the action, its appropriateness with respects to a set of standards, its potential consequences on the world, and the potential for coping with those consequences.

Emotion is therefore a pattern of activity, an emergent property of the interaction of these units, rather than the weighted activation of some set of basic and independent emotion 'nodes,' or modules. Unfortunately, no implementation of this connectionist model has yet been performed, so while conceptually useful, the model has not quantitatively explained or predicted emotional phenomena. In particular, such a model would be useful to understand the appraisal basis of emotional disorders, as well as the extent to which differences in appraisal processes explain individual differences in emotional processing in normal individuals.

For Mischel and Shoda, the emotional system is indeed at the core of what defines personality, individual differences, and the way we interact with others. In their connectionist model, abstract units representing cognitive and affective knowledge are randomly interconnected and interact dynamically. In some conditions, the response of the network to different situations (activation of a subset of the units) displays the same variability as that of human behaviors when individuals are repeatedly presented with a fixed set of situations. In other conditions, the network shows response consistencies across different situations, as would also be observed in humans. According to their model, individual differences and variability of behavior reside in the different connections between cognitive and affective units, rather than depending on the exact nature of the units.

Conclusion

Computational models of emotion, being biophysical or connectionist, point to several important conclusions regarding the neural bases of emotions. First, it has become clear that there are no 'emotional centers' in the brain. Models show that individual modules (being neurons, populations of neurons, or brain areas) do not work in a sequential manner. They are part of complex recurrent networks, so much so that their activity intimately depends on the activity of many other modules, being emotional or not. No region is specifically dedicated to the generation of emotions in general, or of a specific emotion in particular. Furthermore, models show that the dichotomy between cognitive and emotional processing is probably too simple, if not fundamentally incorrect. Brain areas such as the prefrontal cortex or amygdala are inherent parts of the 'computing' networks of stimulus perception and decision making, yet also mediate 'emotional' functions such as stimulus evaluations and emotional expressions.

Second, emotions are not simple 'states,' or patterns of activations of a few brain areas. While previous stimulus–emotion–action associations are a large part of what triggers an emotional behavior, it has become clear that emotions are dynamical modes of functioning. They temporally wax and wane as internal (e.g.,

memory driven) or external (e.g., perceptions) events occur. They develop over short timescales, as in fear, or over long timescales, as in moods or depression. As such, perhaps it is more useful to think about emotional flow, rather than emotional states.

Third, emotions have a functional role, and are not simple by-product of neural activity. The current computational models of emotions have so far focused on the notion of computing bias, whereby information flow is redirected or modulated in various ways. This view is in contrast with the notion of a parallel stream of 'emotional computations' that would simply be added as an additional information channel. The functional aspects of emotion are also clearly apparent in the AI literature.

In sum, the convergence of conceptual, experimental, and computational models seems to be key to the advancement of our understanding of the various aspects of emotional processing, emotional expression, and emotional experience.

See also: Amygdala: Contributions to Fear; Aversive Emotions: Molecular Basis of Unconditioned Fear; Emotion Systems and the Brain; Emotion: Neuroimaging; Emotional Learning in Humans; Emotional Control of the Autonomic Nervous System; Learning, Action, Inference and Neuromodulation; Pharmacology of Fear Extinction; Prefrontal Contributions to Reward Encoding; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Armony JL, Servan-Schreiber D, Cohen JD, et al. (1997) Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning. *Trends in Cognitive Sciences* 1: 28–34.

Armony JL, Servan-Schreiber D, Romanski LM, et al. (1997) Stimulus generalization of fear responses: Effects of auditory cortex lesions in a computational model and in rats. *Cerebral Cortex* 7: 157–165.

Dalgleish T (2004) The emotional brain. *Nature Reviews Neuroscience* 5: 583–589.

Damasio A (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace.

Davidson RJ (2003) Affective neuroscience and psychophysiology: Toward a synthesis. *Psychophysiology* 40: 655–665.

Deco G and Rolls ET (2005) Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. *Cerebral Cortex* 15: 15–30.

den Dulk P, Heerebout BT, and Phaf RH (2003) A computational study into the evolution of dual-route dynamics for affective processing. *Journal of Cognitive Neuroscience* 15: 194–208.

Fellous J-M and Arbib MA (eds.) (2005) *Who Needs Emotions? The Brain Meets the Robot*. New York: Oxford University Press.

Frank MJ and Claus ED (2006) Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review* 113: 300–326.

LeDoux J (1996) *The Emotional Brain*. New York: Simon & Schuster.

Mischel W and Shoda Y (1998) Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology* 49: 229–258.

Picard RW (1997) *Affective Computing*. Cambridge, MA: MIT Press.

Rolls ET (2005) *Emotion Explained*. New York: Oxford University Press.

Salum C, Morato S, and Roque-da-Silva AC (2000) Anxiety-like behavior in rats: A computational model. *Neural Networks* 13: 21–29.

Sander D, Grandjean D, and Scherer KR (2005) A systems approach to appraisal mechanisms in emotion. *Neural Networks* 18: 317–352.

Taylor JG, Scherer K, and Cowie R (2005) Emotion and brain: Understanding emotions and modelling their recognition – introduction to special issue. *Neural Networks* 18: 313–316.

Trappl R, Petta P, and Payr S (eds.) (2002) *Emotions in Humans and Artifacts*. Cambridge, MA: MIT Press.

Wagar BM and Thagard P (2004) Spiking Phineas Gage: A neurocomputational theory of cognitive–affective integration in decision making. *Psychological Review* 111: 67–79.

Emotion: Neuroimaging

R J Dolan, Welcome Trust Centre for Neuroimaging, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Background

Emotions, of which fear is a prime exemplar, are unique among mental states in having in equal measure psychological and somatic referents. The former embody the subjectivity of all psychological states, whereas the latter are evident in stereotyped behavioral patterns of facial expression, comportment, and states of autonomic arousal. Emotional states exert global effects on virtually all aspects of cognition, including attention, perception, and memory. They also exert biasing influences on high-level cognition, including decision-making processes. An informed neurobiological account of emotion must incorporate how these wide-ranging effects are mediated.

Historically, emotional behavior and experience have been attributed to the functions of the limbic system. But an increasing body of evidence suggests that there is widespread propagation of emotional signals within the brain, and this has diminished the general utility of the concept of a limbic-based emotional system. Consequently, contemporary perspectives on emotion emphasize a dynamic brain configuration of emotional states, which to a large extent are determined by an emotion-eliciting context. Within this framework, emotional states are not unique to any single brain region but are expressed in widespread patterns of brain activity, including within early sensory cortices.

Notwithstanding these considerations, the structure most closely affiliated with emotional processing is the amygdala, an anatomically and functional heterogeneous bilateral collection of nuclei in anterior medial temporal cortex. The importance of the amygdala in emotional and behavioral control was first highlighted by reports that rhesus monkeys with bilateral temporal lobe ablations no longer showed appropriate fear or anger responses. The importance of the amygdala derives in part from its extensive anatomical connections with the sensory-processing cortices, hippocampus, basal ganglia, limbic cortex and homeostatic regulatory regions of the hypothalamus and brain stem. This connectivity means that processing within this structure can exert diffuse modulatory influences, including influences on effector autonomic and motor output systems.

Although much of what we can infer about emotional processing in the human brain is derived from the type of studies already described, the advent of high-resolution, noninvasive, functional neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), have greatly expanded this knowledge base.

Imaging Emotional Stimuli

Studies of human emotional processing using functional neuroimaging techniques have invested intensely in an experimental approach that involves presenting to healthy subjects stimuli that represent basic emotions as expressed in facial expressions of fear, anger, sadness, disgust, happiness, and surprise. The key finding from these studies is an enhanced amygdala response to the presentation of fearful faces, as well as enhanced activation in the sensory-processing regions specialized for faces, the fusiform gyrus. A notable observation is that an emotional response to fear in facial expressions is conveyed principally by a specific feature, namely the eyes, such that the presentation of fearful eyes alone can elicit an amygdala response.

The selectivity of amygdala for processing fearful faces, as suggested in neuropsychological studies, is challenged by neuroimaging findings of an amygdala response to a range of emotional faces (Figure 1(a)). Thus, although fearful faces provide reliable stimuli for eliciting amygdala responses, other emotional expressions, including sad, anger, and disgust, also evoke increased amygdala activity (Figure 1(b)). Indeed, when brain responses to multiple emotions are studies within the same experiment, a reliable amygdala activity is elicited to an entire range of emotions.

A widely held view is that emotion is reducible to dimensions of arousal and valence. A key question arising from this conceptualization is whether amygdala activation due to the presentation of emotional stimuli, seen using functional neuroimaging, reflects a specialization for processing emotional intensity (a surrogate for arousal) or, alternatively, for processing valence. There is good evidence for an amygdala response profile that suggests a more complex profile involving an intensity-by-valence interaction. Consequently, it seems that the amygdala codes neither intensity nor valence *per se* but, instead, an interaction between intensity and valence that reflects the overall emotional value of a stimulus.

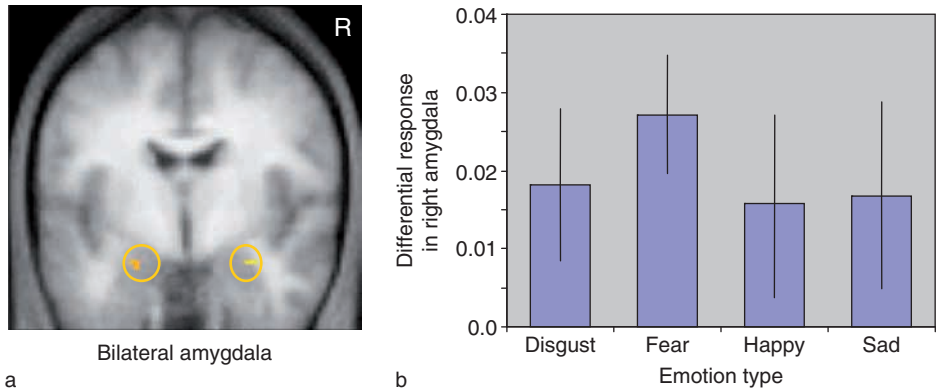

Figure 1 Amygdala responses to emotion: (a) bilateral amygdala activation in response to the presentation of multiple exemplars of emotion; (b) response magnitude for four exemplars of emotion taken from a voxel in the left amygdala.

Imaging the Acquisition of Emotion

In classical or Pavlovian conditioning, a neutral stimulus, through pairing with an emotional stimulus (e.g., an aversive noise in fear conditioning), acquires the ability to predict its future occurrence. Associative learning provides a highly conserved means by which organisms acquire knowledge regarding the causal structure of the environment. Such knowledge endows an organism with the ability to anticipate future events of value, such as the likelihood of food or danger, on the basis of predictive sensory cues. The amygdala plays a central role in this type of causal learning. Thus, amygdala activation is seen when a previously neutral item (the conditioned stimulus (CS+)) acquires predictive significance through its pairing with a biologically salient reinforcer (the unconditioned stimulus (UCS). This pattern of activation is seen both for aversive and appetitive events. Thus, across a wide range of experiments, an enhanced amygdala response is evident when subjects learn that a neutral stimulus predicts the occurrence of an emotion-inducing event, for example, an aversive event.

Theoretical accounts of associative learning based on the Rescorla–Wagner rule and their real-time extensions, such as temporal difference (TD) reinforcement models, provide plausible descriptions of the computational processes underling associative learning. Crucial to these models is the expression of a teaching signal, referred to as the prediction error. A prediction error is used to direct acquisition and refine expectations relating to predictive cues, and it records a change in expected affective outcomes, being expressed whenever predictions are generated, updated, or violated. Functional neuroimaging studies have demonstrated that such a prediction error is expressed in the human brain in structures such as amygdala, striatum, and orbital-prefrontal cortex (OFC) during aversive and appetitive associative emotional learning.

Imaging Updating of Emotional Value

Predictive contingencies acquired through associative learning are limited in their ability to provide a basis for optimizing future behavior. For example, the value of the states associated with predictive cues can change in the absence of further pairing with these cues. Thus, a cue that predicts that the occurrence of food that is valued when a person is hungry has diminished value when the person is sated with that same food. Consequently, maintaining an updated representation of the current value of a sensory-predictive cue that does not depend on new learning in relation to that cue greatly enhances behavioral flexibility.

In human functional neuroimaging studies, neuronal responses elicited by predictive stimuli (such as a CS+) that have been subject to devaluation are associated with parallel response decrements in the left amygdala and OFC. This response pattern suggests these regions contribute to updating the reward value of predictive stimuli. This pattern of value-related response expressed in these regions may explain why pathologies that impact on the medial temporal and basal orbital-frontal lobes, such as fronto-temporal dementias, are associated with behavioral aberrations such as indiscriminate eating, food cramming, change in food preference, and hyperorality. In these contexts, damage to the OFC and amygdala entails that subjects can no longer recruit motivationally appropriate representations of food-based reward value, either appetitive or aversive, in response to predictive cues.

Imaging Emotion in Social Contexts

The amygdala may be important for social cognition and a functional dysregulation within this structure is often suggested as important in the genesis of disorders of social cognition, such as autism. A role for the amygdala in social cognition is suggested by observations of increased activity to the presentation of faces that subjects rate as untrustworthy. An increased amygdala response in White American subjects viewing pictures of unfamiliar Black (African American) faces has also been reported. Increased amygdala activity was not seen when White Americans viewed familiar positively regarded Black (African American) faces, suggesting that this response pattern reflects a learned, culturally specific phenomenon and not innate processes. Observations such as these have led to a view that increased amygdala responsivity to facial expressions actually indexes the occurrence of potential threat. This idea has gained support from other functional neuroimaging studies which show that threat, as embodied in threatening words, aversive tastes, and aversive odors, all evoke an amygdala response. Thus, a stimulus within any modality perceived as a potential threat evokes increased amygdala activity.

Imaging Modulatory Effects of Emotion on Cognition

The domain in which the modulatory influences of emotion are best characterized is episodic memory. The benefit of emotion to episodic memory function is confirmed across numerous studies which show a mnemonic enhancement for material that encompasses personal autobiographical and picture- and word-based items, an effect best seen in free recall tasks. The critical role played by the amygdala in this modulation is evidenced within functional neuroimaging experiments in which amygdala activity during encoding predicts a benefit in the later recall of emotional material. Thus, enhanced amygdala activity at encoding for both positive and negative stimuli is predictive of later episodic memory function, as determined in free recall tasks.

The role of the amygdala in modulating episodic memory function is not restricted to the encoding of emotional events. This structure is also engaged during the retrieval of emotional items and contexts. Connectivity analysis performed on neuroimaging data indicates that contextual recall of emotional events leads to enhanced amygdalar–hippocampal coupling under a controlling influence from OFC. This finding of enhanced interactions between the amygdala and hippocampus is also supported by observations from patients who have lesions to either the amygdala alone or jointly to the amygdala and hippocampus. During the encoding of emotional items, there are bidirectional interactions between these structures, such that an enhanced amygdala response to emotional items is dependent on an intact hippocampus, whereas, conversely, an enhanced hippocampal response to emotional items is dependent on an intact amygdala.

The neurochemical mediation of emotional memory effects has been a subject of increasing interest. Emotional memory enhancement in human subjects can be blocked by the administration of the β-adrenoreceptor blocker propranolol prior to the study. In subjects who are pretreated with propranolol, functional imaging experiments show an absence of emotion-evoked activation of the amygdala. Intriguingly, the encoding of negative emotional items has anterograde and retrograde amnesic effects such that, for example, in word-list presentations there is a relative amnesia for antecedent and preceding items.

Imaging Emotional Enhancement of Perception

An evolutionary perspective on emotion suggests that environmental events of value should be susceptible to preferential perceptual processing. There appear to be two distinct mechanisms by which emotion influences perception. One of these is through emotion interacting with attention, leading to enhanced deployment of attention, leading to the preferential detection of emotional events. For example, in spatial-orienting tasks there are faster responses to targets appearing on the same side as an emotional cue (e.g., faces, spiders, threat words, or conditioned shapes) and a slower response to those appearing on the opposite side. Neuroimaging data, using spatial-orienting paradigms, indicate that the OFC is the site of the expression of this type of interaction between attention and emotion.

Attention is not the sole means by which emotional stimuli influence perception. There is now evidence for mechanisms that are independent of attention. In visual backward-masking paradigms, a target presented for a brief instance can be rendered invisible if it is immediately followed by a second masking stimulus. In situations in which the hidden target stimulus is an emotional item (e.g., a conditioned angry face or a spider), there is preserved processing indexed by differential skin conductance responses (SCRs) to fear-relevant compared to fear-irrelevant targets. Similar findings are evident using the attentional blink paradigm which refers to a situation in which the detection of an initial target stimulus in a stimulus stream leads to impaired awareness, or

inattentional blindness, for a successive second target. Neuroimaging data show that the processing of emotional stimuli outside the window of awareness is mediated by the amygdala.

Studies of patients with focal brain lesions provide additional evidence for a degree of independence between emotional processing and attentional mechanisms. Following brain damage to the right inferior parietal cortex, patients frequently fail to perceive a stimulus presented in their contralesional hemifield (spatial neglect). This contralesional deficit is greatly attenuated for emotional stimuli (e.g., faces with happy or angry expressions or images of spiders). Nonconscious processing of emotion has also been demonstrated in the blindfield of patients with damage to the primary visual cortex. In terms of the anatomical substrates of these effects, functional neuroimaging evidence implicates the amygdala. In functional neuroimaging experiments, using visual backward-masking paradigms, an amygdala response discriminates between unseen emotional and unseen nonemotional targets. In other experiments that involve overt stimulus presentation but in which attention is systematically manipulated, such that emotional items are presented out of the window of attention, an amygdala response to fearful faces is independent of concurrent attentional focus. Likewise, patient studies of subjects with either blindsight or visual extinction demonstrate an amygdala response to emotional stimuli presented out of awareness in the damaged hemifield. Interestingly, residual processing abilities for the unaware presentation of emotional stimulus are associated with the engagement of a subcortical retino–collicular–pulvinar pathway specific to unaware emotional stimulus processing. The involvement of this pathway is of considerable interest because it is implicated in residual visual processing evident in patients with blindsight. The suggestion here is that certain classes of emotional stimuli, for example fearful faces, can be processed by a noncortical pathway to allow rapid adaptive responses to danger.

How the pre-attentive processing of emotional events influence, and enhance, perception is a question addressed by studies in normal subjects and patients with discreet lesions. One possibility is that inputs from emotional-processing regions, in particular the amygdala, modulate the very regions involved in object perceptual processing, specifically those related to the emotion-eliciting object. Anatomically the amygdala receives visual inputs from ventral visual pathways and sends feedback projections to all levels of this pathway. Neuroimaging data provide evidence for the enhancement of functional connectivity between the amygdala and extra-striate visual regions, expressed during the processing of an emotional visual input. Crucially, neuropsychological data from patients with amygdala damage indicate that a perceptual enhancement for emotional items is abolished following damage to this structure. Furthermore, in patients with amygdala lesions, the enhancement of activity seen in early extra-striate visual areas during the encoding of emotional items (e.g., faces) is no longer expressed. This type of evidence is consistent with a proposal that early sensory cortex boosting of activity reflects a direct modulatory influence from the amygdala. Indeed, an emotion-induced enhanced connectivity between the amygdala and extra-striate regions associated with boosting of perception has been demonstrated in a visual detection.

Imaging Feeling States

Human emotion research often conflates neurobiological mechanisms that index the perception or occurrence of an emotional event with their subjective experiential counterparts, referred to as feeling states. Feelings can be defined as mental representations of physiological changes that characterize, and are consequent on, processing an emotion-eliciting objects or image. This definition assigns an important causal role in the genesis of subjective feeling states to afferent feedback to the brain, both sensory and neurochemical. At a broader level, feeling states can be subsumed within homeostatic mechanisms that underlie the survival of the organism. In a recent theoretical model, based on neurological observations, prime emphasis is given to the cerebral representation of bodily states as providing the substrate for emotions and conscious awareness.

The key neurobiological question is whether brain systems supporting emotional perception are distinct from those supporting feelings states. Candidate structures that mediate feeling states encompass those that receive inputs regarding the internal milieu and include the brain stem periacqueductal gray (PAG) and parabrachial nuclei, tegmentum, hypothalamus, and insular, somatosensory and cingulate cortices. Functional neuroimaging provides strong evidence that feelings are mediated by neuronal systems distinct from those that support emotional perception. Thus, volunteer studies have shown that the generation and rerepresentation of peripheral autonomic states involve structures such as the anterior cingulate and insular cortex. Recall of subjective feeling states associated with past emotional experiences engages regions encompassing upper brain stem nuclei, hypothalamus, somatosensory, and insular and orbito-frontal cortices. In subjects with pure autonomic failure (PAF), where there is an absence of visceral afferent information regarding the peripheral body state due to selective

peripheral autonomic damage, there is attenuation of subjective emotional feelings and of emotion-elicited neuronal activity in regions implicated in mediating feeling states. A notable feature of these studies is an absence of any deficit in amygdala activity, in keeping with a more critical role for this structure in emotional perception.

Among the regions most strongly implicated in mediating feeling states is the insular cortex. Direct evidence for its role in representing subjective feeling states comes from a study that used a task that taps awareness of internal bodily states, namely a heartbeat detection task. When subjects detect their own heartbeat, there is enhanced activity in the anterior insular cortex. The proposal that this area mediates feelings is also bolstered by evidence that empathetic awareness of the subjective feeling states of others, for example the awareness engendered when we observe another person receiving pain, is reflected in enhanced activity within the anterior insular cortex and anterior cingulate.

Imaging Emotional Influences on Decision Making

Real-life decision making often involves choices between actions which yield rewards with varying probabilities and magnitudes. Decisions that seek to optimize rewards require an estimation of an expected future reward that will follow from choosing a particular action and the subsequent use of these values to guide action selection. In its simplest form, reward prediction based on expected value can be studied through classical conditioning, in which an arbitrary cue (or CS+) takes on predictive value by association with the subsequent delivery of an affectively significant stimulus or UCS (which can be a reward or punishment or, strictly speaking, an appetitive or aversive stimulus). Neuroimaging studies implicate the OFC alongside other structures such as the amygdala and ventral striatum in reward prediction. Enhanced responses to cues predictive of delivery of either a pleasant or aversive odor result in significant orbitofrontal responses suggesting the OFC maintains predictions for negatively valenced as well as positively valenced stimuli. Human neuroimaging studies of classical conditioning for reward report prediction error signals in prominent target areas of the dopamine neurons, including the OFC. The finding that neural prediction error signals are present in the OFC and throughout the reward network is consistent with the possibility that this mechanism is implemented in the human brain to mediate flexible learning and updating of stimulus–reward associations, a function frequently ascribed to the OFC.

Theoretical accounts of human decision making emphasize a rationalistic perspective of utility maximization, which invokes analytic processes mediated by the executive prefrontal cortex. An emotional or value-based contribution to high-level decision making is evident following ventromedial prefrontal cortex damage where, despite the absence of intellectual deficits, patients make real-life decisions that are disadvantageous. The somatic marker hypothesis arising from observations with these type of lesions suggests that this region provides access to feeling states in relation to past decisions during contemplation of future decisions of a similar nature. Thus, the evocation of past feeling states biases the decision-making process toward or away from a particular behavioral option. However, alternative frameworks that might explain the behavioral deficits seen following damage to this region include an inability to represent the value of competing options for action or a discounting of future rewards. Neuroimaging evidence indicates that this region is activated both during anticipatory states in relation to rewards or punishments and by actual outcomes associated with reward or punishment.

To make effective behavioral decisions between different types of reward requires that the rewards be encoded in a common representational space or currency. A distinct representation of value for different rewards can facilitate goals pertaining to specific motivational states. A candidate region that is activated to different types of reward is the anterior sector of the medial OFC, and this region is a potential candidate for mediating a representation of unitary stimulus–reward value.

It is well recognized that human decision making does not always accord with rationalistic perspectives of utility maximization. An influence of prior outcomes on decision processes is captured by regret theory, in which counterfactual thinking involved in comparing an obtained and forgone outcome indicate that the latter, if chosen, would have been more advantageous (Figure 2(a)). It is known that subjects who experience regret show a subsequent bias away from the maximization of expected value in their choices. This behavioral bias is associated with engagement of the amygdala and orbito-frontal cortices, regions that are also engaged by the actual experience of regret (Figure 2(b)). This pattern is consistent with theories that suggest that the evocation of past emotions in the context of decision making provides a biasing influence on rational decision processes.

A tenet of rational behavior is invariance or extensionality, in which human decisions are held to be consistent regardless of how choices are presented. A deviation from this axiom is described within

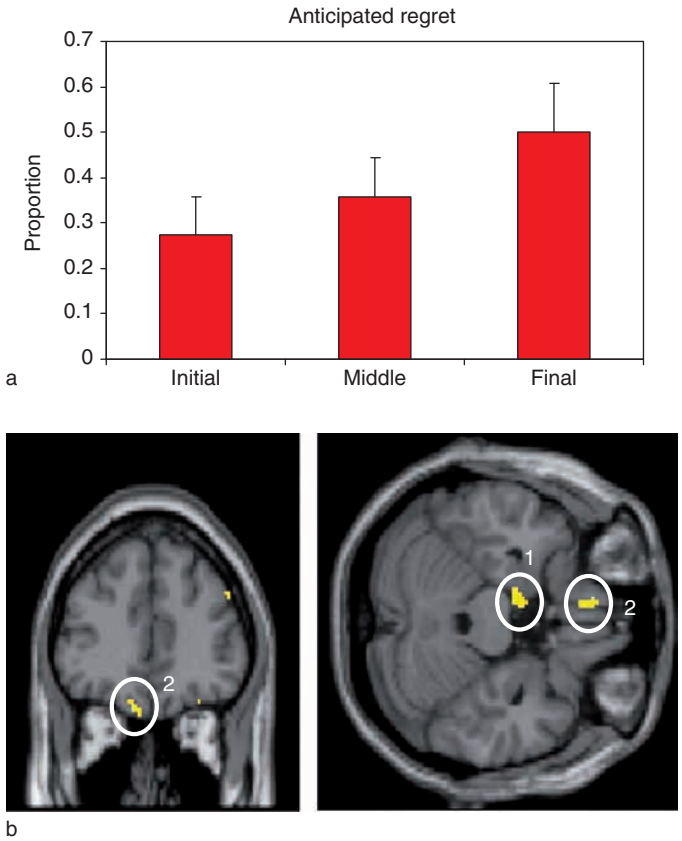

Figure 2 Influence of regret on choices in a gambling task: (a) proportion of subject choices that can be attributed to experiencing regret; (b) brain activity during choice. In (a), the graph shows an increasing attribution of choices to prior experiences of regret; In (b), amygdalar (1) and orbito-prefrontal (2) engagement correlates with the increasing influence of regret on choice.

prospect theory as the framing effect. In simple terms, the framing effect describes a bias in decision making observed when a gamble is presented in terms of gain, leading to choices of a sure option, versus the same gambles presented as a loss leading to risk seeking, when subjects are biased to choose a gamble option. Functional neuroimaging data show that a framing-engendered bias in human decision making, risk aversiveness in the gain frame and risk seeking in the loss frame, is associated with enhanced amygdala activity. The suggestion here is that an emotional heuristic, mediated via key emotion-processing brain regions, is invoked when humans make decisions in situations in which information is incomplete or overly complex.

One common situation that can engender emotion is when an individual makes a decision under risk (the probability distribution of possible outcomes is known) or under ambiguity (where the probability distribution of outcomes is unknown). Subjective utility theory embodies the idea that the outcome probability should influence choice but also that there should no influence on confidence about these probabilities. The Ellsberg paradox shows that subjects behave irrationally when presented with a choice between a risky and ambiguous option, when choices can be shown to depend on subjective probabilities and on ambiguity of events. The possibility that this influence on rational decision making reflects an input from an emotional system is supported in neuroimaging data which show that the level of ambiguity in choices correlates positively with activation of the amygdala.

Conclusion

Neuroimaging has radically altered our conceptualization of emotion. Key observations include evidence that emotion-related activity is expressed across widespread brain regions. A fine distinction between regions that process sensory emotional signals and regions that contribute to the actual experience of emotion has become possible. Furthermore, it has become possible to identify activity in the brain that reflects key computational strategies implemented during emotional learning and emotion-guided decision making.

See also: Aggression: Hormonal Basis; Amygdala: Contributions to Fear; Aversive Emotions: Molecular Basis of Unconditioned Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Emotion Systems and the Brain; Emotional Hormones and Memory Modulation; Emotional Influences on Memory and Attention; Genetics of Human Anxiety and Its Disorders; Neuroimaging; Social Emotion: Neuroimaging.

Further Reading

Critchley HD, Mathias CJ, and Dolan RJ (2001) Neuroanatomical basis for first- and second-order representations of bodily states. *Nature Neuroscience* 4: 207–212.

Damasio AR (1995) *Descartes Error.* London: Picador.

Damasio AR, Grabowski TJ, Bechara A, et al. (2000) Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience* 3: 1049–1056.

Gottfried JA, O'Doherty J, and Dolan RJ (2002) Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. *Journal of Neuroscience* 22: 10829–10837.

O'Doherty JP, Dayan P, Friston K, Critchley H, and Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38: 329–337.

Phelps EA, O'Connor KJ, Gatenby JC, Gore JC, Grillon C, and Davis M (2001) Activation of the left amygdala to a cognitive representation of fear. *Nature Neuroscience* 4: 437–441.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, and Frith CD (2004) Empathy for pain involves the affective but not sensory components of pain. *Science* 303: 1157–1162.

Vuilleumier P, Armony JL, Driver J, and Dolan RJ (2001) Effects of attention and emotion on face processing in the human brain: An event-related fMRI study. *Neuron* 30: 829–841.

Winston JS, O'Doherty J, and Dolan RJ (2003) Common and distinct neural responses during direct and incidental processing of multiple facial emotions. *Neuroimage* 20: 84–97.

Emotional Control of the Autonomic Nervous System

P Carrive, University of New South Wales, Sydney, NSW, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Fear and anxiety are the two main forms of aversive emotion in human and animals. These emotions are basically stress responses with a psychological origin. They are triggered by environmental stimuli that would otherwise be innocuous but that the individual perceives as potentially harmful, either because they are linked to unpleasant memories or because the individual is in an altered state of mind. Like stress responses of physical origin (e.g., pain, cold), stress responses of psychological origin can be associated with marked autonomic and cardiovascular changes. Presumably, these changes prepare the organism for an active response in case it may have to confront or escape from the expected danger (i.e., the traditional 'fight or flight' response). The responses are preprogrammed, and as we all know, they can be difficult to control.

This article focuses on the structures that link the limbic forebrain (where emotions are processed) with the thoracic cord (where sympathetic preganglionic neurons are located), with a particular emphasis on blood pressure and heart rate. It is divided into two parts, for the two experimental approaches that have been used in studies of emotional control of the autonomic nervous system. The first part describes the central autonomic circuitry as revealed by direct activation of brain regions presumed to mediate aversive responses. This approach is used mainly in the anesthetized preparation, which is convenient for the recording of physiological variables, but is limited because of the confounding effect of anesthesia. It sprung up from the discovery in 1943 that hypothalamic stimulation evokes integrated defensive behavior. The second part deals with the same circuitry but as revealed by the effect of intracerebral lesions on naturally evoked aversive emotional responses. This other approach is used in awake animals, which is obviously the only way that emotional responses can be evoked, but it is limited experimentally by the difficulties of recording physiological variables in conscious animals.

The Centrally Evoked Defense Reaction

Early Work with Electrical Stimulation

The term 'defense reaction' (German, Abwehrreaktion) was first introduced by Hess and Brugger in 1943 to describe the effect of electrical stimulation in the hypothalamus of conscious cats. The responses they observed were well coordinated, had a clear emotional component, and looked similar to the characteristic patterns of defensive behaviors seen in normal cats. The most responsive area of the hypothalamus was in the dorsal tuberal hypothalamus, around the fornix. Further work in the 1950s and 1960s showed that this hypothalamic defense region extended caudally into the periaqueductal gray (PAG) and rostrally into the amygdala and bed nucleus of the stria terminalis, leading to the idea that defensive behavior was integrated along a system descending from the amygdala to the midbrain via the hypothalamus.

Then, in 1960, it was demonstrated that the same sites that produced the defense reaction in the conscious cat also produced a characteristic cardiovascular response under anesthesia. The response consisted of marked increases in blood pressure and heart rate associated with a fast vasodilatation in skeletal muscles and vasoconstriction in the gut and skin, all sympathetically mediated. This characteristic pattern of blood flow redistribution to the limbs was consistent with the pattern originally proposed by Cannon in 1929 for what he called emergency reactions – that is, reactions necessary for vigorous responses, such as fight and flight.

Attention then focused on the descending pathways of this response. The muscle vasodilatation, which was mediated by cholinergic sympathetic fibers in the cat, could be traced along a descending pathway from the dorsomedial/perifornical/lateral hypothalamus (DMH/PeF/LH) region and PAG, down into the ventrolateral tegmentum of the midbrain, pons, and ventral medulla. However, it was not clear if the projection was direct to the thoracic cord or if it involved a relay in the lower brain stem. Anatomical tracing has since revealed direct projections from the PeF and LH to the thoracic cord, but the discovery in 1973 of a vasomotor area near the ventrolateral surface of the rostral medulla (now known as the rostral ventrolateral medulla, or RVLM) raised the possibility of a relay in this region. Indeed, this was shown to be the case when it was found that blockade of the RVLM with glycine (an inhibitory neurotransmitter) or kynurenic acid (a glutamatergic antagonist) or lesions with kainic acid markedly reduced all components of the cardiovascular response, including the muscle vasodilatation. The tachycardic component, however, could not be reduced.

The cardiovascular response evoked from the hypothalamic defense area would later be known as the

visceral alerting response. For Hilton and co-workers, the response was a set pattern, a preparatory response for fight or flight, which was activated from the alerting stages to the full expression of the defense response. Although simple and attractive, this concept was based on studies carried out in anesthetized preparations using intracerebral electrical stimulation. Studies conducted in the conscious cat during natural defensive responses did not fully agree with this view. In particular, other investigators argued that muscle vasodilatation was only evoked with movement of the limbs (i.e., during fight or flight). Immobile confrontation with an angry cat, for example, did not evoke muscle vasodilatation but muscle vasoconstriction instead. As will be described next, the idea that there is more than one pattern of cardiovascular response was actually correct.

Excitatory Amino Acid Activation and Shift to the PAG

To circumvent the problem of electrical stimulation which activates both cell bodies and fibers *en passage*, Dampney and Bandler introduced the excitatory amino acid (EAA; e.g., glutamate) injection method in 1982 to try to identify the exact location of the neuronal cell groups mediating the defense reaction in the hypothalamus and PAG. This led to the surprising finding that EAA injection into the hypothalamus or amygdala could not evoke defensive responses nor the full visceral alerting response. Only EAA injection in the PAG could evoke the fully integrated response. Attention then shifted to the PAG.

To understand how PAG neurons could integrate both somatic and cardiovascular components of the defense reaction, Carrive, Dampney, and Bandler recorded behavioral and cardiovascular changes at the same time in an unanesthetized, unparalyzed precollicular decerebrate cat preparation. This approach proved to be very powerful and led to the concept of columnar organization in the PAG in the early 1990s. The most important finding was that different types of basic defensive behavior were topographically represented along a longitudinal column of neurons extending along the lateral PAG. Thus, EAA activation of the middle third of this lateral PAG column evoked a strong facio-vocal response and tonic extension of the limbs, but without stepping movement. This was associated with a rise in blood pressure and heart rate, a moderate renal vasoconstriction, a carotid vasodilatation, and a hind limb vasoconstriction – in other words, the same response that had been observed in cats during immobile confrontation with an aggressive conspecific. In contrast, EAA activation of the caudal third of the lateral PAG column evoked vigorous stepping movements associated with a rise in blood pressure and heart rate, a strong renal and carotid vasoconstriction, and a marked, cholinergically mediated iliac vasodilatation. This was a flight response and it was associated with the classic visceral alerting response of the defense reaction. Thus, contrary to Hilton's view of one defense reaction and a unique visceral alerting response, there were at least two different defensive responses associated with perfectly adapted cardiovascular adjustments (vasodilatation in active parts of the body coupled with vasoconstriction in inactive parts). Further work showed that lateral PAG neurons project to the RVLM in a viscerotopic fashion – that is, the hind limb and renal constrictor parts of the lateral PAG project to their respective vasoconstrictor parts in the RVLM. This work was extended to the rat, in which a similar PAG organization was revealed, although not as clear-cut as in the cat.

Thus, within 10 years of the introduction of the EAA microinjection technique, a column of neurons had been found in the PAG that could mediate not one but at least two different patterns of coordinated defensive responses. In contrast, EAA activation had failed to reveal neurons integrating defensive responses in the hypothalamus. This suggested at the time that the final integrating center for defensive behavior and possibly other forms of emotional behavior was the PAG. However, it gradually became apparent that EAA activation also had limitations and that it had failed to reveal the true properties of hypothalamic neurons.

GABA Disinhibition and Renaissance of the Hypothalamus

One reason for initially dismissing the hypothalamus as a defense integrating center was that EAA microinjections evoked depressor rather than pressor effects in the anesthetized preparation. One problem may have been anesthesia, because pressor and tachycardic effects could actually be evoked from the DMH and PeF in the conscious animal. Another problem may have been the density of output pressor neurons. EAA activation may be less effective in areas where output neurons are scattered, especially if they are mixed with inhibitory neurons, because EAA coactivation of inhibitory interneurons would result in no net effect or even an inhibition of pressor neurons.

In fact, a more effective way to activate hypothalamic neurons in the anesthetized preparation is by disinhibition with microinjection of γ-aminobutyric acid (GABA)-A antagonists (e.g., bicuculline), as was first shown by DiMicco in the 1980s. It is now well established that bicuculline injections in the DMH

and PeF produce marked tachycardia, pressor responses, and increases in renal sympathetic nerve activity. When tested in the conscious animal, the same injections produce running and jumping. Thus, the hypothalamic defense area contains neurons that can mediate flight responses associated with appropriate cardiovascular changes, and these neurons are under tonic GABAergic inhibition.

Current work is now focusing on the output relays of the DMH in the anesthetized animal. Recent studies by Dampney and Morrison show that the RVLM is a necessary relay for the renal sympathetic nerve response because it is blocked by bilateral inhibition of this region with muscimol, a GABA-A agonist. However, the tachycardia is not reduced and the pressor response is reduced only partially. The tachycardia and pressor response may in part be mediated by the rostral medullary raphe nuclei (MRN; MRN will be used here to refer to nuclei raphe magnus and raphe pallidus, not raphe obscurus), especially the raphe pallidus, because inhibition of this region with muscimol reduces these two components. In addition, there is growing evidence from independent work by Blessing, McAllen, and Morrison that MRN is as important premotor center for the sympathetic control of cutaneous vasomotor tone and brown adipose tissue. Brown adipose tissue is a thermogenic organ. Activation of the DMH reduces skin blood flow and increases core temperature, and these effects are prevented by muscimol blockade of MRN. Because stress responses, including those of psychological origin, also cause skin vasoconstriction in the extremities and increase core temperature, it is likely that the same descending DMH–MRN–sympathetic preganglionic neuron (SPN) pathway that mediates tachycardia also mediates these two additional sympathetic responses. Thus, there is good evidence that premotor sympathetic neurons in the RVLM and MRN mediate the vasopressor, tachycardic, hyperthermic, and skin vasoconstrictor responses to activation of the DMH. This is consistent with the fact that the DMH does not have direct projections to the spinal cord. The RVLM and MRN may also be relays of the PeF and LH; however, direct projections to the spinal cord, which originate from these more lateral regions of the dorsal tuberal hypothalamus, could also contribute. This has not yet been tested.

GABA disinhibition with bicuculline was also found to be effective in the amygdala, where it could evoke clear pressor and tachycardic responses when injected in the basolateral nucleus, but not the central nucleus, of conscious rats. This effect appears to be relayed in the dorsal tuberal hypothalamus as was originally suggested.

Thus, the past 15 years have seen a renaissance of the hypothalamus, in large part due to the introduction of GABAergic antagonists. So far, chemical stimulation experiments have confirmed the general organization of the defense reaction neural substrate as originally defined with electrical stimulation. There is now a better picture of the location, organization, and functions of some of the neurons that compose this network, from the amygdala to the hypothalamus, PAG, and rostral ventral medulla (RVLM and MRN). However, as pointed out in the introduction, these studies have two main limitations. One is the artificial means of evoking the emotional behavior. The other is the use of anesthetics. The risk is that this approach may become too reductionist and reveal details of organization that are not relevant to the expression of aversive emotions. Therefore it is important to consider other approaches that do not have these limitations.

Naturally Evoked Forms of Psychological Stress

The other approach that has been used in studies of emotional control of the autonomic nervous system consists of studying the effect of permanent lesions or temporary neuronal blockade on the expression of naturally evoked psychological stress responses. This approach can also be combined with maps of the expression of the immediate-early gene c-*fos* to reveal activated areas. The advantage of this approach is that a brain structure or pathway can be directly linked to an integrated response known to be part of the behavioral repertoire of the animal. A major limitation, however, is that psychological stress can only be studied in the conscious animal, which imposes serious technical challenges to the physiologist.

Different forms of psychological stress have been studied. Some can be considered pure forms of psychological stress (e.g., conditioned fear or psychosocial stress), because they depend purely on the salience of the stimuli that trigger them. Others, such as air-jet and restraint, are mixed forms of psychological stress, because the stimuli that trigger them also have a strong physical (but not noxious) component. We now describe the cardiovascular responses and what is known of the associated circuitry of these four types of psychological stress.

Conditioned Fear

Conditioned fear, probably the best studied form of psychological stress, is evoked by reexposing an animal to a tone, light, or context that has previously

been associated with an aversive stimulus, usually electric footshock. The cardiovascular response consists of a rise in mean arterial pressure and heart rate. The associated pattern of peripheral blood flow changes has not been characterized in the rat, but in the baboon it includes renal vasoconstriction and hind limb vasodilatation, as in the visceral alerting response described earlier. In the rat, the behavioral component of this response is a characteristic tense immobile posture known as freezing. Studies in the 1980s by Smith in the baboon and by LeDoux in the rat have shown that excitotoxic lesions (which destroy cell bodies, not fibers) restricted to the PeF/LH abolish the cardiovascular response of conditioned fear to a tone. LeDoux and co-workers also showed that the same hypothalamic lesions did not affect the freezing response. In contrast, excitotoxic lesions of the caudal PAG reduced the freezing response but not the blood pressure response, while large excitotoxic lesions of the amygdala reduced both the blood pressure and freezing responses. More recently, Carrive showed that large neurotoxic lesions of the DMH/PeF/LH abolished the cardiovascular responses and reduced the freezing responses of conditioned fear to context, while inactivation of the caudal ventrolateral PAG only reduced the freezing response. To summarize, although there may be differences with respect to the role of the hypothalamus in the expression of the behavioral component, these studies confirm that the hypothalamic defense area is necessary for mediating the cardiovascular response of conditioned fear, while the PAG is not. In addition, LeDoux's studies also showed that the same hypothalamic lesions did not affect the cardiovascular response to the footshock itself (i.e., the unconditioned response), suggesting that the dorsal tuberal hypothalamus may be necessary for psychological stressors but not for physical stressors.

It is not known how the effects of the hypothalamic defense area are transmitted to the sympathetic preganglionic neurons of the thoracic spinal cord during conditioned fear. Relays in the RVLM and MRN are most probably involved, but this has not been tested. Because the effective lesions appeared restricted to the PeF/LH, it is possible also that the direct spinal projections that arise from this area could play a role.

Psychosocial Stress

Psychosocial stress can be studied in the animal with the resident-intruder test. This is a relatively pure form of psychological stress that does not require conditioning. Its cardiovascular component also consists of a rise in arterial pressure and heart rate. A recent study by Kuwati and colleagues shows that orexin knockout mice have a markedly reduced cardiovascular response when confronted to an intruder. Orexin (also known as hypocretin) is a neuropeptide produced by a restricted group of neurons located precisely and exclusively in the DMH/PeF/LH area. Orexin has been implicated in the control of arousal and of many autonomic functions, including cardiovascular ones. Orexinergic neurons project not only to the RVLM and MRN but also to the thoracic cord, and in each region orexin evokes pressor and tachycardic responses. Interestingly, the pressor and tachycardic responses to disinhibition of the PeF area are also reduced in these orexin knockout mice, but the cardiovascular response to a noxious cutaneous pinch is not. This supports the idea that the hypothalamic defense area is involved in the cardiovascular response to psychological but not physical stressors and strongly suggests that orexin neurons are implicated. It also supports the idea that the response could be mediated, at least in part, by direct spinal projections.

Air-Jet Stress

Application of a jet of pressurized air to the face (air-jet stress) is another form of psychological stress, although not as 'pure' as conditioned fear or psychosocial stress. The cardiovascular response also consists of a rise in arterial pressure and heart rate. Work by DiMicco in the rat has shown that bilateral temporary blockade of the DMH with muscimol or kynurenic acid markedly reduces this response, while similar injection of muscimol in the adjacent paraventricular nucleus of the hypothalamus has no effect. Furthermore, neuronal blockade of MRN with muscimol abolishes the tachycardic response but does not significantly reduce the pressor response. The role of RVLM in the rat has not been investigated, but in the rabbit, its blockade with kynurenic acid reduces the increase in renal sympathetic nerve activity evoked by air-jet stress. These are the first demonstrations that the MRN and RVLM relay at least parts of the cardiovascular response to a mixed psychological stressor.

Restraint

Restraint or immobilization is also a mixed stressor with an even stronger physical component. The cardiovascular response is similar to that of air-jet stress and, as with that stressor, bilateral temporary blockade of the DMH with muscimol abolishes the response. However, neurotoxic lesions that abolish the cardiovascular response of conditioned fear to context do not reduce the cardiovascular response to restraint. This intriguing finding suggests that although the DMH can contribute to the cardiovascular response of restraint, it is in fact not necessary. Other centers, presumably located at lower levels of the neuraxis,

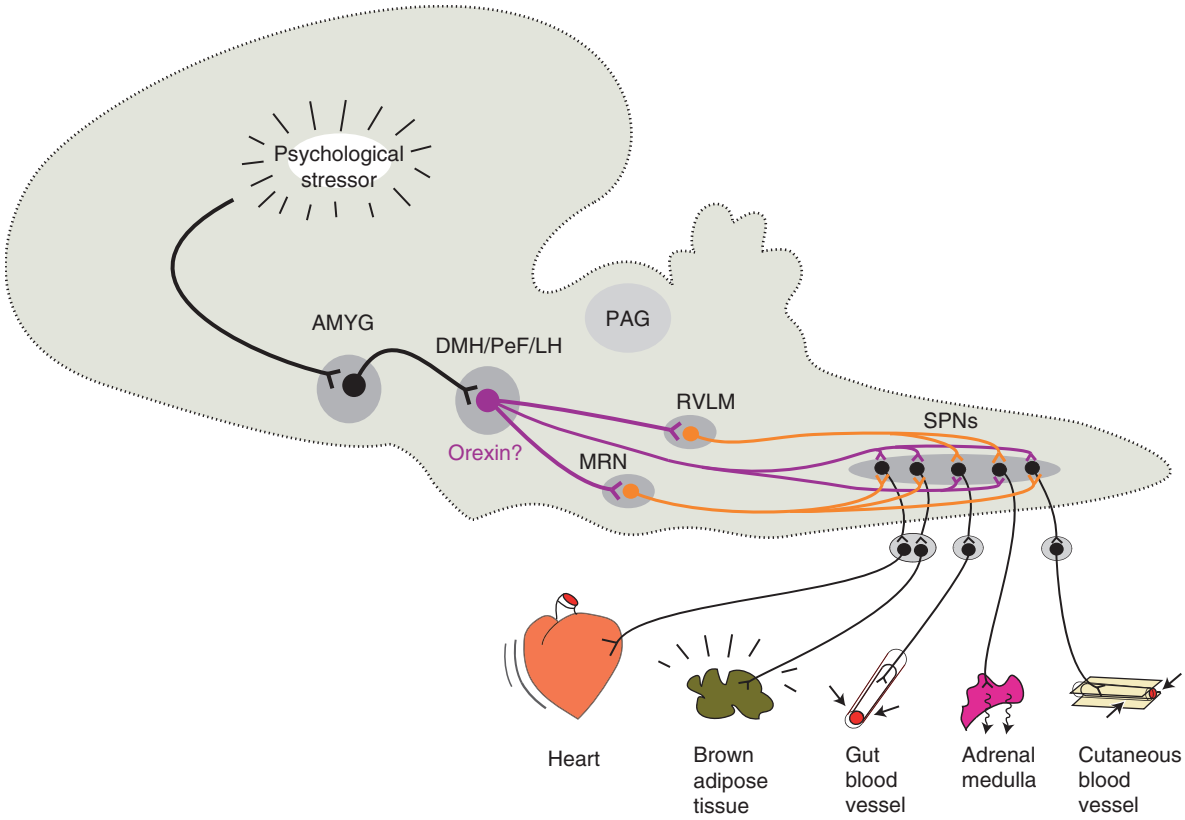

Figure 1 Schematic diagram of the descending pathways that could mediate some of the sympathetic responses to aversive emotions. Activations of the heart (tachycardia), brown adipose tissue (hyperthermia), gut blood vessels (increased peripheral resistance), adrenal medulla (release of noradrenaline and adrenaline), and cutaneous blood vessels (skin vasoconstriction) are mediated by different pathways, all originating from the dorsal tuberal hypothalamus (DMH/PeF/LH). Hypothalamic control of sympathetic preganglionic neurons (SPNs) may be direct as well as indirect via relays in the rostral ventrolateral medulla (RVLM) and medullary raphe nuclei (MRN; rostral ones – i.e., raphe magnus and raphe pallidus), which control different end organs. The efferent connections of the dorsal tuberal hypothalamus may contain the neuropeptide orexin (hypocretin). AMYG, amygdala; DMH, dorsomedial hypothalamus; LH, lateral hypothalamus; PAG, periaqueductal gray; PeF, perifornical hypothalamus.

perhaps the lateral PAG or possibly the RVLM and MRN, can take over this role if DMH neurons have been permanently lesioned. It would be interesting to know if the same is true for air-jet stress, since it is also a mixed psychological stressor. In any case, this is in sharp contrast to conditioned fear. Thus, it would appear that the dorsal tuberal hypothalamus is necessary for the purer forms of psychological stress but not for the other more physical ones, such as restraint, or for the purely physical ones, such as electric footshock and noxious pinch.

Conclusion

Periventricular structures of the forebrain and midbrain play an important role in the integration of the somatic and autonomic components of complex behaviors, including aversive emotional responses. In this article, an attempt has been made to put together a hypothetic model of the lower pathways mediating the autonomic responses to aversive emotions (Figure 1). Much remains to be done to test this model, but it is already clear that the dorsal tuberal hypothalamus (i.e., the DMH, PeF, and LH) plays a necessary role in the expression of the cardiovascular response to the 'purer' forms of psychological stress. The PAG, on the other hand, does not appear to play a critical role in these cardiovascular responses, although it can mediate integrated behavioral and cardiovascular patterns of basic defensive responses.

During aversive emotions, the dorsal tuberal hypothalamus is most probably under the control of the amygdala, which has a more integrative role. The dorsal tuberal hypothalamus in turn transmits its output to the preganglionic sympathetic neurons of the thoracic cord; however, the exact pathways mediating these effects during emotional responses have not been fully resolved. Experiments in anesthetized animals suggest relays in the lower brain stem, in the RVLM and MRN. The RVLM may mediate the

kidney and gut vasoconstrictor responses and most probably activation of the adrenal medulla as well, while the MRN may mediate the tachycardia, skin vasoconstrictor, and hyperthermic responses. Both structures may contribute to the pressor response by acting on different effectors (i.e., heart, arteries, and release of circulating catecholamines). In addition to RVLM and MRN there are also direct spinal projections originating from the more lateral parts of the hypothalamus (PeF and LH) that could make a significant contribution. This direct projection to the thoracic cord raises the possibility of a premotor sympathetic pathway that could be activated during the purer forms of psychological stress in the conscious animal. Such a direct line could mediate sympathetic changes without integration in the homeostatic centers of the lower brain stem, an idea consistent with the view that autonomic responses to aversive emotions are emergency responses.

See also: Amygdala: Contributions to Fear; Anxiety Disorders; Autonomic Disorders; Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Gastrointestinal Control; Autonomic Nervous System: Respiratory Control; Aversive Emotions: Molecular Basis of Unconditioned Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Emotion and Vigilance; Fear Conditioning and Synaptic Plasticity; Social Stress in Adult Primates; Sympathoadrenal System: Neural Arm of the Stress Response.

Further Reading

Abrahams VC, Hilton SM, and Zbrozyna A (1960) Active muscle vasodilatation produced by stimulation of the brainstem: Its significance in the defense reaction. *Journal of Physiology (London)* 154: 491–513.

Bandler R and Shipley MT (1994) Columnar organization in the midbrain periaqueductal gray: Modules for emotional expression? *Trends in Neuroscience* 17: 379–389.

Cao WH, Fan W, and Morrison SF (2004) Medullary pathways mediating specific sympathetic responses to activation of dorsomedial hypothalamus. *Neuroscience* 126: 229–240.

Carrive P (1993) The periaqueductal gray and defensive behavior: Functional representation and neuronal organization. *Behavioral Brain Research* 58: 27–47.

Carrive P (2007) Central circulatory control: Psychological stress (and the defense reaction). In: Llewellyn-Smith I and Verberne AJM (eds.) *Central Regulation of Autonomic Function,* 2nd edn. New York: Oxford University Press (in press).

DiMicco JA, Sarkar S, Zaretskaia MV, et al. (2006) Stress-induced cardiac stimulation and fever: Common hypothalamic origins and brainstem mechanisms. *Autonomic Neuroscience: Basic and Clinical* 126–127: 106–119.

Hilton SM (1982) The defense-arousal system and its relevance for circulatory and respiratory control. *Journal of Experimental Biology* 100: 159–174.

Kayaba Y, Nakamura A, Kasuya Y, et al. (2003) Attenuated defense response and low basal blood pressure in orexin knockout mice. *American Journal of Physiology* 285: R581–R593.

LeDoux JE (2000) Emotion circuits in the brain. *Annual Reviews in Neuroscience* 12: 155–184.

LeDoux JE, Iwata J, Cicchetti P, et al. (1988) Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *Journal of Neuroscience* 8: 2517–2529.

McDougall SJ, Widdop RE, and Lawrence AJ (2005) Central autonomic integration of psychological stressors: Focus on cardiovascular regulation. *Autonomic Neuroscience: Basic and Clinical* 123: 1–11.

Nalivaiko E and Blessing WW (2001) Raphe region mediates changes in cutaneous vascular tone elicited by stimulation of amygdala and hypothalamus in rabbits. *Brain Research* 891: 130–137.

Smith OA and DeVitto JL (1984) Central integration for the control of autonomic responses associated with emotions. *Annual Reviews in Neuroscience* 7: 43–65.

Emotional Disorders: Treatment

L R Stines and N C Feeny, Case Western Reserve University, Cleveland, OH, USA
E B Foa, University of Pennsylvania, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Anxiety Disorders: Social Phobia, Obsessive–Compulsive, and Posttraumatic Stress

Social Anxiety Disorder

Social anxiety disorder, also referred to as social phobia, is characterized by a persistent fear of negative evaluation in one or more contexts involving performance and/or interaction with others. Social phobia is increasingly recognized as among the most common of psychiatric disorders in the United States. The recently published National Comorbidity Study–Revised (NCS-R) reported a 12 month prevalence rate of 6.8%, and similarly, the original NCS data reported 12 month and lifetime rates of 7.9% and 13.3%, respectively, in the community. In clinical samples, the incidence of social phobia is much higher. Individuals with social anxiety fear embarrassment (e.g., concern that others will know they are anxious) and judgment (e.g., worry that others will think they are stupid), and this anxiety is elicited consistently when confronted with the feared situations. As a result, the feared situations, commonly including public speaking, eating, or writing, are often avoided or endured with significant distress. Adults with social anxiety recognize that this fear is excessive or unreasonable. In order to meet diagnostic criteria based on the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV), this anxiety must significantly interfere with their work, daily living, or social functioning. The diagnosis varies slightly in children, such that they are not required to identify the fears as excessive or unreasonable and their difficulties must persist for at least 6 months before a diagnosis can be made.

Obsessive–Compulsive Disorder

Obsessive–compulsive disorder (OCD) is an anxiety disorder characterized by recurrent obsessions or compulsions that are time consuming (i.e., more than 1 h per day) or cause significant distress or impairment in functioning. Approximately 1–2% of adults in the United States meet criteria for OCD at any one time. Obsessions are defined as persistent, intrusive thoughts or images that are experienced as being inappropriate, unexpected, and outside the individual's control. Common obsessions include thoughts of contamination (e.g., fear of catching disease by shaking hands) and order (e.g., arranging objects in symmetry). Compulsions are conceptualized as repetitive behaviors or mental acts that are designed to suppress or neutralize the obsessive thought or image in order to reduce anxiety. Common compulsive behaviors include hand washing, counting, and checking (e.g., locks on doors and appliances). In order to meet diagnostic criteria for OCD, an individual must experience either obsessions or compulsions, but not necessarily both. However, epidemiological data suggest that most individuals with OCD do have both. As with social phobia, the diagnostic criteria for OCD vary slightly for children, who may not identify their thoughts or behaviors as excessive or unreasonable.

Posttraumatic Stress Disorder

In the United States, it is estimated that 50–60% of the population will experience a potentially traumatic event, such as a physical or sexual assault or motor vehicle accident, during their lifetime. Following exposure to traumatic events, most people do not develop long-lasting psychological difficulties. Among those who do, however, the most common constellation of such difficulties is called posttraumatic stress disorder (PTSD). With regard to prevalence, the NCS-R data showed 12 month rates of PTSD to be 3.5%. In the community, the lifetime prevalence of PTSD is estimated to be 9%. Overall, rates of PTSD among those who are exposed to trauma are higher, of course, with lifetime prevalence estimated at 24%, though rates vary according to the kind of trauma experienced. The symptoms of PTSD fall into three clusters: reexperiencing the event (e.g., intrusive and distressing thoughts); avoidance of trauma-related reminders (e.g., avoidance of trauma-related thoughts or situations); and hyperarousal (e.g., sleep disturbance). In order to meet DSM-IV diagnostic criteria for PTSD, these difficulties must persist for at least 1 month and cause significant impairment in functioning. The diagnostic criteria for PTSD are the same for adults and children.

Treatments for the Anxiety Disorders

Psychosocial Treatments for the Anxiety Disorders

Several cognitive behavioral therapies (CBTs) have been developed to treat these anxiety disorders and have been tested in prospective, randomized studies.

Exposure therapies, in particular, have been among the most investigated and have received some of the strongest empirical support in the treatment of persistent, pathological anxiety. The chief aim of this type of treatment is to help patients approach feared and avoided objects, situations, memories, and images until anxiety reduces, a process called habituation. Exposure therapy for the anxiety disorders is based on the notion that once-neutral, or unconditioned, stimuli have become conditioned to evoke a fear response, or anxiety. In an effort to reduce anxiety, one escapes from, or otherwise avoids, the fear-eliciting stimuli. Avoidance interferes with new learning by preventing opportunities to habituate to objectively harmless stimuli and to disconfirm erroneous cognitions, thus maintaining unrealistic fear. Exposure, then, provides opportunities for this learning. The evidence for the efficacy of these treatments for social phobia, OCD, and PTSD will be summarized in the next sections.

CBT for social phobia Cognitive behavioral treatments for social anxiety disorder typically include *in vivo* exposure to avoided or dreaded social situations. During therapy, patients are encouraged to engage in, rather than avoid, situations like talking with neighbors, making phone calls, attending social functions, and speaking up in meetings. Individual hierarchies of avoided situations are developed with the patient, and *in vivo* exposure occurs in a graded fashion, often through homework assignments. Some treatment packages for social phobia include cognitive therapy with an emphasis on changing negative social cognitions, and some deliver exposure therapy without cognitive therapy.

Multiple randomized controlled trials have supported the efficacy of CBT for social phobia. CBTs such as cognitive behavioral group therapy have been shown to be effective in several studies. Exposure therapy without cognitive restructuring is also efficacious in the treatment of social phobia. Moreover, at least eight controlled studies have compared CBT (exposure therapy plus cognitive therapy) to exposure therapy without any explicit focus on cognitions. In only two of these trials did CBT prove superior to exposure therapy alone, immediately following treatment. Results of a dismantling study suggest that exposure without cognitive therapy is as effective as exposure plus cognitive therapy. In addition, several meta-analyses of the social phobia treatment outcome literature have shown that both CBT and exposure therapy produce comparable results with regard to improvement and dropout. With regard to the magnitude of change experienced by patients, four meta-analyses have found the average improvements in social anxiety following CBT to be large, with pre–post effects sizes ranging from 0.80 to 1.07. Overall, the evidence for the efficacy of exposure therapy with or without cognitive therapy in this disorder is quite strong.

Exposure and response prevention for OCD Exposure and response prevention (EX/RP) is widely regarded as the treatment of choice for OCD. In this treatment approach, patients are exposed to the feared stimuli in a hierarchical manner, beginning with stimuli that evoke less anxiety and gradually moving to more-distressing stimuli. During exposure therapy for OCD, patients are encouraged to refrain from engaging in any neutralizing thoughts or behaviors, which is the response prevention component of the treatment. *In vivo* exposures are commonly used, though imaginal exposure is implemented when *in vivo* exposure would not be possible or should not be implemented.

Empirical support for the efficacy of EX/RP is strong. Several randomized controlled clinical trials have investigated the efficacy of EX/RP in the treatment of OCD and have found that EX/RP results in a significant reduction in OCD symptoms that is generally maintained at follow-up, though some residual symptoms often remain. EX/RP has demonstrated superior outcomes in the treatment of OCD when compared with psychopharmacological interventions (e.g., clomipramine) and comparable outcomes when compared with CBT. Support for the effectiveness of EX/RP for OCD was reported by Franklin and colleagues, who found similar rates of improvement as a result of EX/RP among patients treated in a fee-for-service clinic, with no exclusionary criteria based on comorbidity, treatment history, or concomitant treatments, compared with those reported in published controlled clinical trials.

Exposure therapy for PTSD Exposure therapy programs for clients with PTSD also typically include both imaginal and *in vivo* exposure and aim to help the person emotionally process the trauma. Imaginal exposures are typically conducted across multiple treatment sessions and between them in the form of homework. In this procedure, the patient is asked to vividly imagine the traumatic event and to describe it aloud, along with the thoughts and feelings that occurred during the event. *In vivo* exposures are typically done outside treatment sessions as homework and involve systematic, prolonged, and repeated confrontation with external situations, places, or activities that trigger trauma-related fear and anxiety. Prolonged exposure, one exposure-based treatment for PTSD, consists of a variety of therapeutic components, including

psychoeducation, breathing retraining, repeated recounting of the traumatic event (i.e., imaginal exposure), and encouragement to systematically confront trauma-related reminders (i.e., *in vivo* exposure).

The efficacy of exposure therapy for PTSD is well established. Several well-controlled studies demonstrate that patients who receive exposure therapy experience a significant reduction in symptoms (with approximately 60% of patients being PTSD diagnosis-free following CBT) and maintain their gains through follow-up. To date, four randomized controlled trials have demonstrated the efficacy of prolonged exposure for sexual and nonsexual assault victims, though exposure therapy has also been shown to be effective for PTSD resulting from various traumas in men and women. In the recent treatment guidelines developed by experts in the field in conjunction with International Society for Traumatic Stress Studies, exposure therapy was identified as the most supported intervention for PTSD.

Impact of Treatments on Brain Functioning

In the past two decades, improved technology has made possible the identification of the brain systems involved in anxiety. Noninvasive functional imaging techniques, including positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), have been used to detect differences between adults with psychopathology and healthy controls. Recent research has investigated the role of treatments, both psychosocial and psychopharmacological, on changes in brain structure and functioning among adults with anxiety disorders. In this section, we will briefly review and compare the impact of CBT and selective serotonin reuptake inhibitors (SSRIs) on brain abnormalities among adults with anxiety disorders.

Symptom provocation is one primary methodology that has been employed to examine brain functioning across psychiatric disorders. In symptom provocation studies, stimuli are used (e.g., imaginal exposure to anxiety-producing scenarios) to illicit symptoms characteristic of the disorder under investigation. Rauch and colleagues used individually tailored symptom provocation to assess brain functioning during a PET scan for adults with OCD. When compared with resting state (i.e., neutral stimuli presented), the OCD symptomatic state (i.e., anxiety-inducing stimuli presented) was associated with significantly increased regional cerebral blood flow (rCBF) in the right caudate nucleus, left anterior cingulate cortex, and bilateral orbitofrontal cortex. Similarly, fMRI found activation in several brain areas, including the medial orbitofrontal cortex, the anterior cingulate cortex, and the amygdala, among adults with OCD during symptom provocation, whereas control participants had no areas of brain activation. A recent meta-analysis of the OCD imaging literature found reliable differences between adults with OCD and healthy controls in activity in the orbital gyrus and the head of the caudate nucleus, but no other differences in brain activation.

In the first published study to evaluate treatment effects on brain activity in OCD, Baxter and colleagues found that patients who responded to treatment by either fluoxetine or CBT (EX/RP) experienced a decrease in right caudate metabolic activity, and these results were replicated in a subsequent study of only CBT. In a study of behavior therapy for treatment-refractory adults with OCD, clinical improvement measured as global functioning was correlated with decreased activity, measured pre- and posttreatment, in the right caudate nucleus.

In imaging studies of social phobia, symptom provocation tasks have typically included the presentation of human faces expressing various emotional states. These studies have consistently demonstrated amygdala activation in adults with social phobia when they are presented with anxiety-producing stimuli.

Only one published study to date has compared the impact of CBT and SSRIs on brain functioning in social anxiety disorder. Furmark and colleagues examined the brain activation in adults with social phobia using PET and compared those who received CBT, those who received citalopram (an SSRI), and a waiting list control. The researchers used a symptom provocation methodology and evaluated rCBF during a public speaking task both pre- and posttreatment. There were no differences between the CBT and citalopram groups on any outcome measures posttreatment, both improved significantly more than controls, and across both groups, symptom reduction was associated with reduced rCBF in the amygdala, hippocampus, and anterior and medial temporal cortex.

Brain neuroimaging in PTSD is a burgeoning area of research, and a number of brain areas have been implicated in the disorder. Several studies have found increased amygdala activity and decreased activation in the medial prefrontal cortex during symptom provocation tasks in adults with PTSD. Discrepant results have been found with regard to hippocampal volume in studies that have used imaging to assess brain structure in adults with PTSD; some studies observed smaller hippocampal volume although others found no difference between PTSD patients and trauma-exposed controls.

In contrast to the breadth of studies on brain activation and structure in PTSD, research on brain

changes associated with treatment for PTSD is in its infancy. To date, only two published studies have examined the impact of psychotropic medications on brain functioning. In the study by Vermetten and colleagues, adults who received 1 year of treatment for PTSD with an SSRI (paroxetine) experienced on average 4.6% increase in hippocampal volume from pre- to posttreatment, although there was no correlation between change in hippocampal volume and PTSD symptom reduction. Bremner and colleagues examined brain structure pre- and posttreatment using a novel, anticonvulsant (phenytoin) treatment for PTSD and found that the treatment resulted in a 5% increase in right hippocampal volume although the treatment did not produce a statistically significant improvement in PTSD. Lindauer and colleagues, however, examined the impact of psychotherapy (incorporating some CBT elements) on hippocampal volume in adults with PTSD. They found that adults with PTSD had significantly smaller hippocampal volume than controls at baseline, and there were no changes in hippocampal volume as a function of treatment.

It is important to highlight that studies comparing the impact of CBT and pharmacotherapy on changes in brain functioning have demonstrated that the treatments produce similar results, suggesting that there may be a common biological mechanism through which these treatments affect the brain, though certainly further research is necessary to determine common and distinct pathways of treatment response.

Conclusion

The prevalence of the anxiety disorders reviewed here – social phobia, OCD and PTSD – is quite high, particularly in clinical populations. Fortunately, however, effective interventions have been identified for their treatment. In particular, exposure-based treatments are generally recommended as the first-line treatment for the anxiety disorders. Emerging evidence suggests that CBT may actually lead to changes in brain activation and functioning, thus broadening the scope of the very powerful impact that these types of treatments may have on individuals. It is likely that publication of the DSM-V, slated for 2011, will reflect significant changes in the classification of, and potentially the diagnostic criteria for, anxiety disorders and influence the ways in which these disorders are treated.

See also: Anxiety Disorders; Anxiety: Drug Therapy; Emotion Systems and the Brain; Emotion and Vigilance; Genetics of Human Anxiety and Its Disorders; Obsessive–Compulsive Disorder; Phobia and Human Evolution; Posttraumatic Stress Disorder: Overview; Posttraumatic Stress Disorder: Neurobiology; Psychotherapeutic Approaches to Psychiatric Disorders.

Further Reading

Abramowitz JS (1996) Variants of exposure and response prevention in the treatment of obsessive-compulsive disorder: A meta-analysis. *Behavior Therapy* 27(4): 583–600.

Foa EB, Hembree EA, Cahill SP, et al. (2005) Randomized trial of prolonged exposure for posttraumatic stress disorder with and without cognitive restructuring: Outcome at academic and community clinics. *Journal of Consulting and Clinical Psychology* 73: 953–964.

Kessler RC, Chiu WT, Demler O, and Walters EE (2005) Prevalence, severity, and comorbidity of 12-Month *DSM-IV* disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry* 62: 617–627.

Lindauer RJL, Vlieger EJ, Jalink M, et al. (2005) Effects of psychotherapy on hippocampal volume in out-patients with post-traumatic stress disorder: A MRI investigation. *Psychological Medicine* 35(10): 1421–1431.

Taylor S (1996) Meta-analysis of cognitive-behavioral treatments for social phobia. *Journal of Behavior Therapy and Experimental Psychiatry* 27(1): 1–9.

Emotional Hormones and Memory Modulation

J L McGaugh and B Roozendaal, University of California, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Emotionally arousing experiences are generally well remembered. We know that from our own experiences as well as from extensive research findings. Memories of automobile accidents, fires, muggings, rapes, wartime battles, terrorists' bombings, and other intensely emotional experiences are often reported as being 'deeply etched in the brain.' We also remember significant details of special occasions such as birthdays, bar mitzvahs, weddings, graduation ceremonies, and funerals as well as horrific events such as those of 11 September 2001. Milder emotional experiences such as those induced by praises, insults, or embarrassments also tend to be well remembered.

There are some obvious and partly correct explanations of why it is that emotional experiences (of either positive or negative valence) create strong memories. We may pay more attention to details during the experiences and we may repeatedly recall the experiences and, thus inadvertently, or purposefully, rehearse them. However, in addition, as discussed below, there is extensive evidence that neurobiological processes initiated by emotionally arousing experiences act automatically to strengthen memories of the experiences. And, it certainly seems highly adaptive to have such automatic processes that enable the significance of events to regulate their remembrance.

However, stress and emotional arousal not only induce strong memories of new information but can also impair our remembering. We all have, probably, experienced situations such as stressful examinations or job interviews during which we were simply not able to recall information. The evidence from many animal and human studies now indicates that the mechanisms that enhance the consolidation of new, emotionally arousing experiences are also responsible for the impairment of memory retrieval and working memory.

Stress Hormone Effects on Memory Consolidation

Emotionally arousing experiences induce the release of stress hormones from the adrenal glands. The adrenergic hormone epinephrine (adrenaline) is released into the circulatory system from the adrenal medulla, and the glucocorticoid cortisol (corticosterone in rodents) is secreted from the adrenal cortex. The extensive evidence that stimulant drugs (e.g., amphetamine) administered to rats or mice shortly after they are trained on a task enhance the consolidation of memory of the training suggested that stress hormones released by the training experience may act as endogenous modulators of memory consolidation. Considerable evidence supports this hypothesis. Epinephrine or glucocorticoid receptor (GR) agonists administered within minutes after training produce dose-dependent enhancement of the long-term retention of many different kinds of training experiences typically used in animal memory studies, including inhibitory avoidance, contextual and cued fear conditioning, spatial discrimination, conditioned taste aversion, and object recognition. Further, adrenoceptor antagonists or drugs that disrupt glucocorticoid functioning impair memory when administered post-training. As stress hormones administered an hour or longer after training do not affect subsequent memory, the effects obtained with administration shortly after training provide strong evidence that the hormones influence the consolidation of long-term memory that is initiated by the training experiences. There is also extensive evidence that epinephrine and glucocorticoids, as well as stressful conditions that stimulate their release, enhance memory consolidation in human subjects when administered shortly before or after learning.

Although epinephrine and glucocorticoid stress hormones have similar effects in enhancing memory consolidation, their effects are initiated through different mechanisms. Epinephrine does not directly affect brain activity as it does not pass the blood–brain barrier. The effects are mediated by the stimulation of β-adrenoceptors located on the ascending vagus nerve that projects to noradrenergic cells in the nucleus of the solitary tract (NTS) located in the brain stem. These noradrenergic cells directly innervate brain regions involved in memory consolidation, including the amygdala. The NTS may also influence noradrenergic activation via its projection to the locus coeruleus (LC), which has noradrenergic cells with more widespread projections to forebrain regions, including the hippocampus, cortex, and amygdala. The key role of the NTS in mediating epinephrine effects on memory is indicated by the finding that brief inactivation of cells in the NTS with lidocaine (to block sodium channels) prevents epinephrine-induced release of norepinephrine (NE) in other brain regions and blocks epinephrine influences on memory consolidation. Further, stimulating the NTS immediately after training

by infusing β-adrenoceptor agonists directly into the NTS enhances memory. Thus, activation of receptors in the NTS plays a critical role in mediating the transmission of adrenergic signals from the periphery to brain systems that process memory for emotionally significant experiences.

Glucocorticoids are lipophilic and readily enter the brain when released or administered peripherally. Thus, they can directly activate adrenal steroid receptors located on brain neurons, where their effects on memory involve specific activation of low-affinity GRs. In contrast to the high-affinity mineralocorticoid receptors, GRs are activated only during emotional arousal. Glucocorticoid binding to GRs can affect gene transcription by direct binding of receptor homodimers to DNA. Glucocorticoids may also act postsynaptically on membrane-bound GRs to influence the β-adrenoceptor-cAMP pathway and work synergistically with noradrenergic influences to regulate activity of downstream molecular mechanisms, including that of the transcription factor cAMP response-element binding (CREB) protein. These effects may ultimately result in structural changes, for example, via modifications of cell adhesion molecules, and strengthen cell–cell interactions.

The findings of many experiments indicate that administration of glucocorticoids to animals shortly after training on learning tasks produces dose-dependent enhancement of memory consolidation highly comparable to that found with epinephrine. Importantly, memory is also enhanced by posttraining infusions of GR agonists administered directly into specific brain regions, including the NTS, cortex, hippocampus, or amygdala. Disruption of glucocorticoid synthesis by administration of the drug metyrapone prevents the memory-enhancing effects of epinephrine. Thus, although epinephrine and glucocorticoid effects are initiated by different mechanisms, these adrenal stress hormones influence brain systems involved in processing memory and induce effects, as noted briefly above and discussed further below, that interact in modulating memory consolidation.

Involvement of the Basolateral Amygdala

It is well established that emotional experiences that induce the release of adrenal stress hormones also activate the amygdala. Further, there is considerable evidence that these stress hormones affect memory consolidation either directly by activating the amygdala or indirectly through influences that subsequently activate the amygdala. Lesions of the amygdala prevent the memory-enhancing effects of peripherally administered stress hormones. The basolateral amygdala (BLA) is the region of the amygdala that is critical for stress hormone influences on memory consolidation. The BLA is a cortical-like structure that has many reciprocal connections with other higher brain regions. Selective lesions of the BLA, but not lesions of other amygdala regions such as the central amygdala, prevent the memory-modulating effects of adrenal stress hormones. Further, memory consolidation is modulated by infusions of drugs affecting adrenoceptors or GRs administered selectively into the BLA. Adrenoceptor and GR agonists produce dose-dependent memory enhancement and antagonists of these receptors impair memory when infused directly into the BLA shortly after training. Infusions of these adrenoceptor and GR agonists and antagonists administered into other amygdala nuclei are ineffective.

Role of Noradrenergic Activation of the BLA

Noradrenergic activation of the amygdala by projections from the NTS and LC plays an essential role in the modulation of memory consolidation induced by both adrenergic and glucocorticoid effects. Infusions of β-adrenoceptor antagonists administered into the amygdala block the memory-enhancing effects of peripherally administered epinephrine that, as discussed above, are known to be mediated by activation of the noradrenergic cells of the NTS and LC. Further, β-adrenoceptor antagonists infused selectively into the BLA also block glucocorticoid-induced memory enhancement. Studies using *in vivo* microdialysis techniques to assess NE levels in the amygdala provide additional evidence that noradrenergic activation within the amygdala is involved in influencing memory. In rats, footshock stimulation of the kind that is typically used in fear-based training induces the release of NE in the amygdala, and the amount of NE released varies directly with the footshock intensity. Moreover, drugs and hormones known to enhance memory consolidation potentiate the increase in NE levels in the amygdala, induced by footshock stimulation. Further and importantly, in rats given inhibitory avoidance training, the magnitude of training-induced increases in amygdala NE assessed shortly after training correlates very highly with the rats' subsequent long-term retention performance, providing additional evidence suggesting that amygdala NE regulates memory consolidation.

Noradrenergic activation of the BLA also appears to be essential in enabling glucocorticoid modulation of memory consolidation. Animal experiments investigating stress hormone and BLA involvement in memory consolidation typically use highly arousing training conditions, such as inhibitory avoidance or contextual or cued fear conditioning, that are known to induce amygdala NE release. There is evidence that such NE release is critical for glucocorticoid

enhancement of memory. When arousing training conditions are used, β-adrenoceptor antagonists administered peripherally or infused directly into the BLA block glucocorticoid-induced memory enhancement. Further, posttraining administration of corticosterone does not enhance retention of training experiences (e.g., object recognition) that induce relatively low emotional arousal. However, with such low-arousing conditions, administration of yohimbine, a drug that enhances NE release by blocking noradrenergic autoreceptors, enables glucocorticoid-induced memory enhancement.

Other findings indicate that corticosterone also does not directly induce activation of the BLA but, rather, influences BLA activity only when there is sufficient noradrenergic activation within the BLA. Such evidence strongly suggests that arousal-induced noradrenergic activation of the BLA is essential in enabling glucocorticoid-induced memory enhancement. These findings fit well with evidence from studies of human subjects indicating that glucocorticoid enhancement of memory consolidation may be limited to memories of information that is emotionally arousing. Such findings are also consistent with evidence that glucocorticoid-induced memory enhancement is blocked by intra-BLA infusions of β-adrenoceptor antagonists or Rp-cAMPS, a drug that blocks the NE signaling cascade, as well as evidence that GR antagonists do not prevent the memory-enhancing effects induced by intra-BLA infusions of the cAMP analog 8-Br-cAMP. This evidence clearly indicates that, in the BLA, cAMP activation is downstream from the interaction of glucocorticoids with β-adrenoceptors. The interaction of glucocorticoids with noradrenergic influences in the BLA is summarized in Figure 1.

Most studies of the involvement of glucocorticoid and noradrenergic influences on memory consolidation have investigated the effects of hormones or drugs administered after the initial learning of tasks. As extinction consists of the learning of new consequences of cues, for example, that a cue that previously predicted footshock no longer predicts footshock, effects of these treatments on extinction should be comparable to those found with studies of

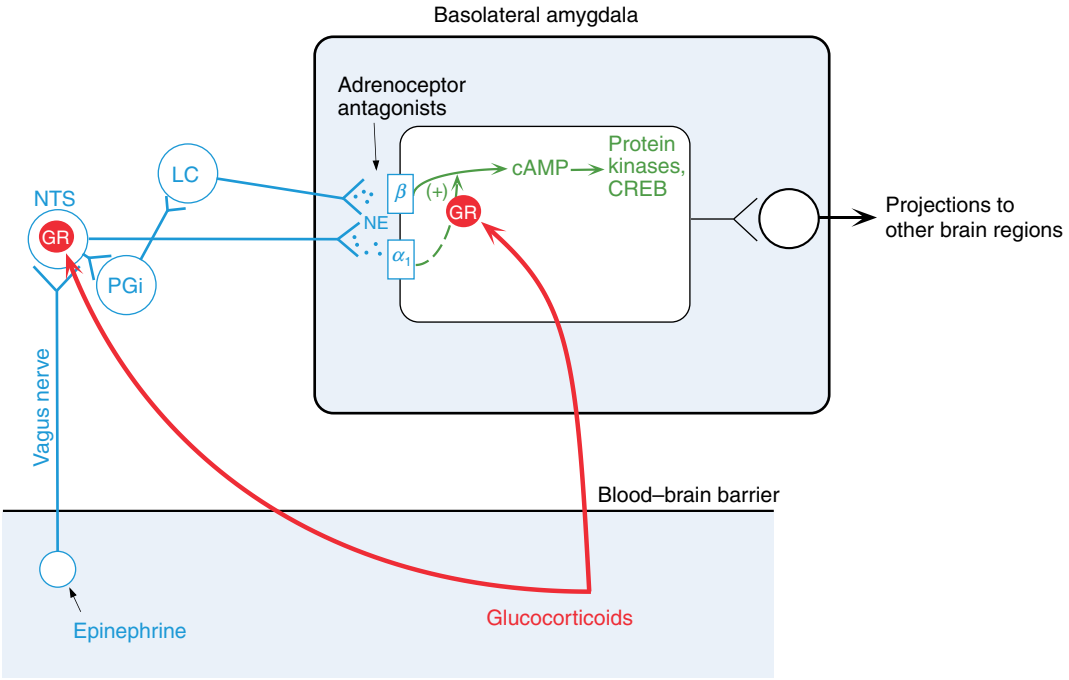

Figure 1 Interactions of adrenal stress hormones with the noradrenergic system in the BLA in modulating memory consolidation. Adrenal stress hormones are released during training in emotionally arousing tasks and are known to enhance memory consolidation. Epinephrine, which does not cross the blood–brain barrier, induces the release of norepinephrine (NE) in the BLA by activating vagal afferents to the nucleus of the solitary tract (NTS). Noradrenergic neurons in the NTS project directly to the BLA, and indirectly via the locus coeruleus (LC). NE binds to both β-adrenoceptors (β) and α_1-adrenoceptors (α_1) at postsynaptic sites and activates cAMP and protein kinase A formation. Glucocorticoids freely enter the brain and bind to glucocorticoid receptors (GRs) in brain stem noradrenergic cell bodies to potentiate NE release in the BLA, as well as postsynaptically in BLA neurons to facilitate the NE signal cascade. Glucocorticoids may influence the β-adrenoceptor-cAMP system via a coupling with α_1-adrenoceptors. These stress hormone effects on noradrenergic activation in the BLA are required for regulating memory consolidation in other brain regions. PGi, nucleus paragigantocellularis; CREB, cAMP response-element binding protein. Reproduced from Roozendaal B (2000) Glucocorticoids and the regulation of memory consolidation. *Psychoneuroendocrinology* 25: 213–238, with permission from Elsevier.

original learning if the same modulatory systems are involved. The findings support this implication. When administered after extinction training, systemically or intra-BLA-administered glucocorticoids or NE enhance the consolidation of extinction memory.

Effects of Corticotropin-Releasing Hormone (CRH) in the BLA on Memory Consolidation

Although most studies examining emotional arousal or stress effects on memory consolidation investigated the effects of adrenal stress hormones, emotional arousal is known to induce the release of a variety of neuromodulatory hormones within the brain, including opioid peptides, γ-aminobutyric acid (GABA), vasopressin, and adrenocorticotropic hormone (ACTH). Importantly, many findings have shown that intra-BLA infusions of drugs that either mimic or block the action of these transmitter systems enhance or impair memory for emotionally arousing training. Moreover, disruption of BLA activity prevents these drugs from influencing memory consolidation. One transmitter system that is of particular interest is CRH. CRH is a neuropeptide that is released by emotional arousal into the hypothalamus and plays a crucial role in regulating activity of the hypothalamic–pituitary–adrenocortical axis and circulating glucocorticoid levels. However, CRH is also released in extra-hypothalamic sites, including the amygdala as well as several other brain regions, in response to arousing or stressful stimulation. Blockade of endogenous CRH in the BLA with infusions of a CRH receptor antagonist impairs memory for emotionally arousing training, whereas infusions of CRH dose-dependently enhance memory consolidation. Furthermore, CRH is known also to interact with noradrenergic mechanisms in the BLA in influencing memory consolidation. Thus, this evidence indicates that BLA activation by emotional arousal is a general gateway in mediating stress hormone and neurotransmitter effects on memory consolidation.

BLA Interactions with Other Brain Regions in Modulating Memory Consolidation

Although the evidence summarized above clearly indicates that the BLA is a critical brain site for integrating adrenergic, glucocorticoid, and other stress-activated influences on memory consolidation, other evidence clearly indicates that the BLA is not the locus of the long-term memory processes modulated by stress hormones. Rather, the evidence indicates that BLA activation acts via efferent projections to influence memory processing in other brain regions. The BLA is richly interconnected with very many brain regions, including the hippocampus, caudate nucleus, nucleus accumbens, and cortical regions known to participate in different aspects or forms of memory. Posttraining noradrenergic activation of the amygdala influences memory in tasks known to involve functioning of different brain regions, including water-maze spatial and cued tasks, inhibitory avoidance, contextual and cued fear conditioning, conditioned taste aversion, and object recognition. Lesions of BLA efferent pathways block the memory-modulating effects of BLA stimulation. Additionally and importantly, lesions of the BLA, the nucleus accumbens or stria terminalis (a fiber pathway that connects the BLA with the nucleus accumbens) block the effects of glucocorticoids administered systemically after training.

The finding of many studies indicating that noradrenergic inactivation of the BLA or lesions of the BLA or its efferent pathways block the memory-modulating influences of drugs administered posttraining to other brain regions provides additional strong evidence that interactions with efferent brain regions are critical in mediating BLA influences on memory consolidation. The memory enhancement induced by a GR agonist infused into the hippocampus after training is blocked by selective lesions of the BLA or infusions of a β-adrenoceptor antagonist into the BLA. An intact and functioning BLA is also required for inducing memory modulation by adrenergic and muscarinic cholinergic drugs administered after training directly into cortical regions, including the entorhinal cortex, anterior cingulate cortex, prefrontal cortex, and insular cortex. The BLA also projects directly to the nucleus basalis (via the stria terminalis) that provides cholinergic activation of the cortex. Such activation is essential for learning-induced cortical plasticity. The evidence that stimulation of the BLA enhances cortical activation and that reversible inactivation of the nucleus basalis blocks the activation suggests that BLA activation may influence memory via such cholinergic activation of the cortex or that such activation is a co-requirement. Findings indicating that lesions of the nucleus basalis-cortical projections block the memory enhancement induced by NE infused into the BLA posttraining support this hypothesis.

There is also considerable evidence indicating that the BLA regulates neuroplasticity in efferent brain regions. Stimulation of the amygdala enhances long-term potentiation (LTP) in thalamocortical synapses. Lesions of the BLA or infusions of β-adrenoceptor antagonists into the BLA prevent the induction of LTP into the dentate gyrus of the hippocampus as well as stress-induced influences on hippocampal and cortical LTP. In contrast, stimulation of the BLA facilitates LTP induction in these brain areas, whereas blockade of adrenergic or glucocorticoid mechanisms reduces this BLA effect on hippocampal neuroplasticity. Other

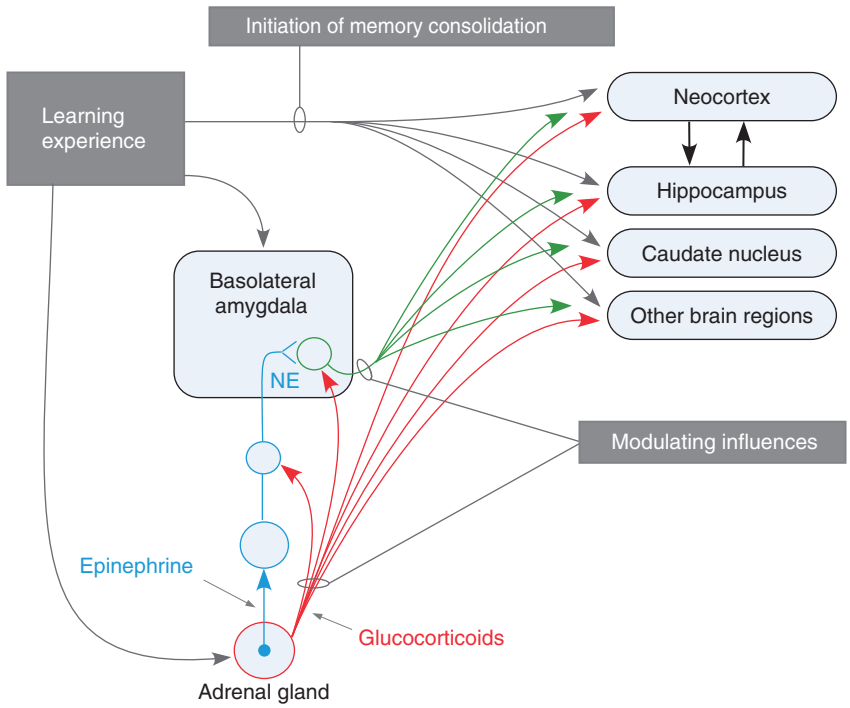

Figure 2 Emotional arousal-induced modulation of memory consolidation. Experiences initiate memory storage in many brain regions involved in the forms of memory represented. Emotionally arousing experiences also release adrenal epinephrine and glucocorticoids and activate the release of noradrenaline in the basolateral amygdala (BLA). The BLA modulates memory consolidation by influencing neuroplasticity in other brain regions. NE, norepinephrine. Reproduced from McGaugh JL (2000) Memory: A Century of consolidation. *Science* 287: 248–251.

evidence suggests that synchronized oscillatory activity within the BLA may facilitate cortical processes involved in memory consolidation. A single footshock of the kind typically used in training increases the duration and synchrony of firing of BLA neurons lasting for up to 2 h. Such findings suggest that the increased firing may act to facilitate neuroplasticity underlying memory consolidation in efferent brain regions, including the cortex. Figure 2 summarizes the roles of adrenal stress hormones and BLA activation in modulating memory consolidation in other brain regions.

Influence of Emotional Arousal on Human Memory: Stress Hormones and Amygdala Activation

The findings of studies of the influence of stress hormones on human memory are highly comparable to those obtained in experiments with animals. Several studies have reported that epinephrine and glucocorticoids enhance memory and suggest that the effects are obtained selectively with emotionally arousing material. Cortisol administered to subjects prior to presentation of emotionally arousing words or pictures enhanced long-term memory. Similarly, epinephrine or cold pressor stress (known to induce the release of epinephrine and cortisol) administered shortly after subjects view emotionally arousing pictures enhanced their long-term memory of the pictures. In hospitalized patients undergoing intensive care treatment, the strength of subsequent memory of the experiences varied directly with the amount of cortisol and epinephrine they received during the treatment. Additionally, administration of a β-adrenoceptor antagonist prior to subjects' viewing of an emotionally arousing series of pictures blocked the enhancing effect of emotional arousal on long-term memory.

There is also considerable evidence from human studies supporting the conclusions of animal studies indicating that the enhancing influence of emotional arousal on memory involves activation of the amygdala. In human studies, however, the experiments have not as yet investigated the possible selective involvement of the BLA. Evidence that in human subjects with amygdala lesions, unlike subjects with an intact amygdala, emotionally arousing stimulation does not enhance long-term memory supports the view that amygdala activation may be critical for emotionally enhanced memory. Alzheimer's patients' memory of a serious earthquake assessed months after the experience varied directly with the sizes of their amygdala, as examined by magnetic resonance imaging (MRI).

The involvement of amygdala activation in emotionally influenced memory has also been investigated in many studies using positron emission tomography (PET) and functional MRI (fMRI) in healthy humans. These studies reported that activity of the amygdala assessed during the presentation of emotionally arousing stimuli correlated highly with memory of the stimuli tested weeks later. Further, the relationship between amygdala activity during encoding and subsequent long-term memory was greatest for the most emotionally arousing stimuli and for subjects who showed a large increase in endogenous cortisol in response to the arousing stimuli. Importantly, β-adrenoceptor antagonists blocked both the increase in amygdala activity and the enhanced retention induced by emotional stimuli. Other findings are consistent with evidence from animal experiments, suggesting that amygdala activity modulates memory processing in other brain regions. Amygdala activity is highly correlated with the activity of hippocampal and parahippocampal regions during the encoding of emotional stimuli.

Stress Hormone Effects on Memory Retrieval and Working Memory

Although most studies investigating stress hormone effects on memory have focused on the neurobiological mechanisms underlying the consolidation of recent experiences, evidence indicates that these hormones also influence other memory functions. Stress exposure or glucocorticoids administered immediately after a learning session impair retention performance tested shortly after the session, that is, at a time when the memory trace has not yet been consolidated into long-term memory. Extensive evidence now indicates that glucocorticoids can affect such retention performance by impairing the retrieval of previously learned information. Stress exposure or glucocorticoids administered peripherally to rats shortly before retention testing, 24 h after training, impair retention performance. Likewise, administration of stress-level doses of glucocorticoids to human subjects shortly before memory testing impairs delayed recall on episodic memory tasks. Importantly, in contrast to stress hormone effects on memory consolidation, the impairing effects of emotional arousal on memory retrieval are temporary and subside when the hormone levels return to baseline. Extensive cognitive and neurobiological research indicates that the hippocampus is an important brain region implicated in memory retrieval. Glucocorticoid-induced memory retrieval impairment depends, in part, on GR activation in the hippocampus. Additionally, evidence from a PET study of memory in human subjects indicates that a stress-level dose of cortisone reduces activity of the parahippocampal gyrus, an effect that correlates with impaired episodic memory retrieval. The effects of epinephrine on memory retrieval have not been investigated.

The effects of glucocorticoids on memory retrieval are highly comparable to those obtained in studies of memory consolidation in that the effects depend critically on an interaction with noradrenergic mechanisms. Systemic administration of a β-adrenoceptor antagonist blocks the memory retrieval impairment induced by concurrent injections of corticosterone. Other evidence from animal studies indicates that the BLA interacts with the hippocampus in mediating glucocorticoid effects on memory retrieval impairment. BLA lesions or a β-adrenoceptor antagonist infused into the BLA block hippocampal glucocorticoid impairment of memory retrieval.

Such evidence clearly indicates that the role of BLA noradrenergic activity in regulating emotional arousal effects on hippocampus-dependent cognitive processes is not restricted to modulating memory consolidation but extends to memory retrieval. The evidence from animal studies is consistent with that from human studies examining stress hormone effects on memory retrieval. Glucocorticoids or psychosocial stress only impair retrieval of emotionally arousing information or during emotionally arousing test conditions. Further, the glucocorticoid-induced impairment of memory retrieval in humans is blocked by concurrent administration of a β-adrenoceptor antagonist. Findings of imaging studies indicate that successful retrieval of emotionally arousing information induces greater activity in and connectivity between the amygdala and hippocampus than retrieval of emotionally neutral information.

Stress exposure or glucocorticoid administration also impairs working memory, which is known to rely on the integrity of the prefrontal cortex. Mild uncontrollable stress impairs performance of rats on a delayed alternation task, a task commonly used to assess working memory in rodents. Such stress also increases NE (and dopamine) turnover in the prefrontal cortex. Excessive levels of NE in the prefrontal cortex are known to induce working memory impairment. Like stress, glucocorticoid administration impairs working memory. Stress doses of corticosterone or a GR agonist administered either peripherally or directly into the prefrontal cortex impair delayed alternation performance in rats. In addition, cortisol administration impairs working memory performance in human subjects. Importantly, glucocorticoids appear to interact with noradrenergic mechanisms in inducing working memory impairment. A β-adrenoceptor antagonist administered peripherally blocks the impairing effect of corticosterone on working memory in rats.

Further, animal studies have shown that glucocorticoid effects on working memory depend on functional interactions between the BLA and the prefrontal cortex. Disruption of BLA activity blocks the effects on working memory of a GR agonist administered into the prefrontal cortex. This evidence provides strong support for the hypothesis that BLA activity modulates stress or emotional arousal effects on working memory in other brain regions. There is also evidence from human studies supporting the hypothesis that glucocorticoids interact with noradrenergic mechanisms in inducing working memory impairment. Psychosocial stress impaired working memory only during conditions inducing concurrent activation of glucocorticoids and the sympathetic nervous system.

Emotional Arousal and Traumatic Memory

It is, of course, essential for our adaptation and survival that we record and retain lasting memories of our significant experiences. But can the activation of stress hormones and the amygdala by emotional arousal strengthen memories more than is necessary? Several kinds of studies have addressed this issue. In children who experienced traumatic injuries, elevated levels of epinephrine and cortisol measured immediately after the children were admitted to the hospital correlated with symptoms of posttraumatic stress disorder (PTSD) assessed 6 weeks later. Moreover, the administration of a β-adrenoceptor antagonist to recently traumatized patients attenuated the development of PTSD symptoms. Propranolol administered to patients (for 10 days) within 6 h after they experienced an acute psychologically traumatic event reduced the symptoms of PTSD assessed 3 months after the trauma. As with humans exposed to other types of extreme trauma, patients who have been treated in an intensive care unit often report traumatic memories. Extremely traumatic memories from intensive care treatment in some of these patients are associated with the development of PTSD. Many patients with stress-associated diseases, such as PTSD, show sustained neuroendocrine abnormalities, which include increased catecholaminergic activity and impairment in glucocorticoid signaling. Severely ill patients also often show insufficient endogenous glucocorticoid signaling which has recently been termed critical illness-related corticosteroid insufficiency (CIRCI). In several controlled trials in patients with suspected CIRCI, the administration of glucocorticoids (stress doses of cortisol) during intensive care treatment resulted in a significant reduction of PTSD symptoms in long-term survivors. Cortisol administration may help to surmount impaired glucocorticoid signaling from CIRCI during critical illness resulting in a downregulation of the stress response. Sustained cortisol administration might also prevent PTSD by interfering with memory retrieval.

As discussed above, high levels of glucocorticoids impair memory retrieval. Thus, low cortisol levels such as found in some patients after trauma at risk for PTSD or with established PTSD could lead to excessive retrieval of traumatic information. As it has been proposed that PTSD develops over time because of positive feedback mechanisms in which traumatic memories are constantly retrieved and restored; the administration of stress doses of glucocorticoids after a massive stress exposure could inhibit the recall of traumatic information and thereby reduce the probability of PTSD. In addition, glucocorticoid administration might also reduce the development of PTSD by facilitating the extinction of traumatic memories. Some evidence indicates that administering stress doses of cortisol might also be useful for the treatment of established PTSD as a result of major traumatic experiences of nonmedical origin. Thus, the findings of studies investigating memory of traumatic experiences offer support for the hypothesis that emotions can and do strengthen memories more than is necessary. And, the findings of the experimental studies provide some understanding of the neurobiological processes that underlie such strengthening as well as some possible implications for therapeutic intervention to ensure that significant events are well remembered, but not excessively remembered.

See also: Amygdala: Contributions to Fear; Hormones and Memory; Posttraumatic Stress Disorder: Overview; Posttraumatic Stress Disorder: Neurobiology; Stress and Cognition; Stress and Neuronal Plasticity.

Further Reading

Buchanan TW and Adolphs R (2004) The neuroanatomy of emotional memory in humans. In: Reisberg D and Hertel P (eds.) *Memory and Emotion*, pp. 42–75. New York: Oxford University Press.

Cahill L (2000) Modulation of long-term memory in humans by emotional arousal: Adrenergic activation and the amygdala. In: Aggleton JP (ed.) *The Amygdala*, pp. 425–446. London: Oxford University Press.

de Kloet ER, Oitzl MS, and Joëls M (1999) Stress and cognition: Are glucocorticoids good or bad guys? *Trends in Neurosciences* 22: 422–426.

de Wied D (1991) The effects of neurohypophyseal hormones and related peptides on learning and memory processes. In: Frederickson RCA, McGaugh JL, and Felten DL (eds.) *Peripheral Signaling of the Brain*, pp. 335–350. Toronto: Hogrefe and Huber.

Koob GF (1987) Neuropeptides and memory. In: Iversen LL, Iversen SD, and Snyder SH (eds.) *Handbook of Psychopharmacology, Vol. 19: New Directions in Behavioral Pharmacology*, pp. 531–573. New York: Plenum.

LaBar KS and Cabeza R (2006) Cognitive neuroscience of emotional memory. *Nature Review Neuroscience* 7: 54–64.

Lupien SJ, Fiocco A, Wan N, et al. (2005) Stress hormones and human memory function across the lifespan. *Psychoneuroendocrinology* 30: 225–242.

McGaugh JL (2000) Memory: A Century of consolidation. *Science* 287: 248–251.

McGaugh JL (2003) *Memory and Emotion: The Making of Lasting Memories*. New York: Columbia University Press.

McGaugh JL (2004) The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience* 27: 1–28.

McGaugh JL and Gold PE (1989) Hormonal modulation of memory. In: Brush RB and Levine S (eds.) *Psychoendocrinology*, pp. 305–339. New York: Academic Press.

Paré D (2003) Role of the basolateral amygdala in memory consolidation. *Progress in Neurobiology* 70: 409–420.

Richter-Levin G and Akirav I (2003) Emotional tagging of memory formation-in the search for neural mechanisms. *Brain Research Review* 43: 247–256.

Roozendaal B (2000) Glucocorticoids and the regulation of memory consolidation. *Psychoneuroendocrinology* 25: 213–238.

Roozendaal B (2002) Stress and memory: Opposing effects of glucocorticoids on memory consolidation and memory retrieval. *Neurobiology of Learning and Memory* 78: 578–595.

Roozendaal B, Okuda S, de Quervain DJ-F, and McGaugh JL (2006) Glucocorticoids interact with emotion-induced noradrenergic activation in influencing different memory functions. *Neuroscience* 138: 901–910.

Schelling G, Roozendaal B, and de Quervain DJ-F (2004) Can posttraumatic stress disorder be prevented with glucocorticoids? *Annals of the New York Academy of Sciences* 1032: 158–166.

Emotional Influences on Memory and Attention

E A Phelps, New York University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The interaction of emotion and cognition is obvious to most people. It is not uncommon for traffic to slow because the attention of the drivers ahead is distracted by the view of an accident. It is also not uncommon for someone to proclaim that his or her memory for an event is vivid or detailed because the event is highly emotional, or what is often called 'memorable'. In spite of the common recognition that emotion and cognition are intertwined, the study of emotion in the scientific investigation of cognitive processes, such as memory and attention, has only recently blossomed. The reasons for this have to do with both the philosophical foundations of the discipline of cognition within psychology and the emergence of cognitive neuroscience as a dominant approach in understanding cognitive functions. When one examines brain function, it is apparent that the neural systems of emotion and cognition are intertwined from perception to reasoning and decision making. This recent emphasis on understanding the neural representation of human behavior has resulted in a new appreciation of the important role of emotion in cognitive functions.

One brain region that plays a prominent role in the interaction of emotion and cognition is the amygdala. The amygdala is a small structure in the medial temporal lobe that sits adjacent and anterior to the hippocampus (see Figure 1). The amygdala is known to be important in a range of emotional behaviors, most notably the ability to learn that stimuli in the environment predict potential aversive consequences, referred to as fear conditioning. The amygdala is critical for the acquisition, storage, and expression of fear conditioning. In addition to its role in emotional learning, the amygdala is widely connected to other brain regions that mediate a range of cognitive functions. It has been suggested that through these connections, the amygdala can influence cognition in the presence of potentially important stimuli. This additional role for the amygdala in the tuning of cognition by emotion helps assure that events that may be important for our survival receive priority in processing. This article reviews what has been learned from the study of the amygdala concerning the interaction of emotion, attention, and memory.

Emotion's Influence on Attention and Perception

Attention and perception are the first stages of stimulus processing. Factors that influence these processes will also influence downstream cognitive functions, such as memory. An early example of emotion's impact on attention is the classic 'cocktail party effect' described by Cherry in 1953. This effect gets its name from the observation that when we focus our attention on one stream of information and ignore others, such as paying attention to a single conversation at a cocktail party, we might inadvertently pick up on some of the unattended information if it is salient or important, such as when one's own name is mentioned in another conversation at the party. This finding and others suggest that emotion can facilitate awareness for emotionally salient stimuli in situations in which attention is limited.

Recent cognitive neuroscience research indicates the amygdala may mediate the facilitation of attention with emotion. This was investigated using a paradigm that tests the temporal limitations of attention, called the attentional blink. In this paradigm, stimuli, such as words, are presented in rapid succession – so quickly it is difficult to identify any words. However, if participants are told that they can ignore most of the words presented and selectively attend to a few target words, such as those printed in a different color ink, they can usually identify the target words. However, this ability to selectively attend to and identify specific words in the rapid stream of words is limited by the temporal relation between the targets. If the second target word is presented a few items after the first target word, participants will often miss it. It is as if noticing and encoding the first target word results in a temporary refractory period during which it is difficult to notice and encode a second target. When this occurs, it is as if attention 'blinked'. However, if the second target word is emotional and arousing, this temporal limitation of attention is not as great. During the 'blink' period, when attentional resources are limited, arousing words are more likely to be identified than are neutral words. This attenuation of the attentional blink effect is not apparent for patients with damage to the amygdala. Unlike normal controls, patients with damage to the left amygdala experience a similar decrement in the ability to report emotional and neutral words. This finding is consistent with previous studies indicating that in situations when one has limited attentional resources, emotional stimuli are more likely than neutral stimuli to reach

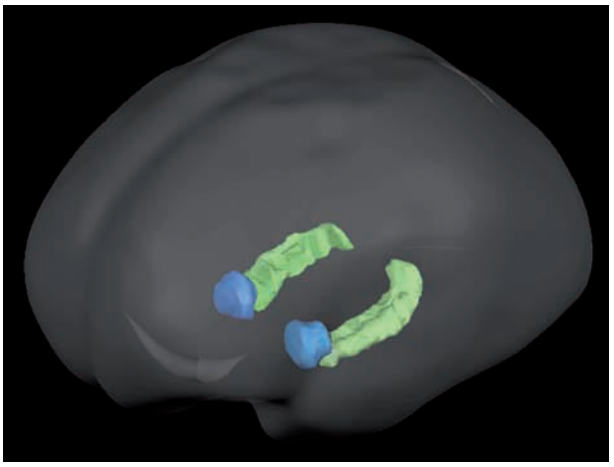

Figure 1 The human amygdala (blue) and the hippocampus (green).

awareness. It also indicates that the amygdala plays a critical role in this facilitation of attention with emotion.

It is suggested that the amygdala has its influence on attention by modulating processing in brain regions known to mediate perception. Two distinct mechanisms have been explored that may explain the amygdala's facilitation of attention. The first was proposed by Weinberger and colleagues, who demonstrated that fear conditioning can tune the representation of emotional stimuli in sensory cortices. In studies with rats, they showed that the representation of simple tones in the auditory cortex is changed through fear conditioning when a tone is paired with, and comes to predict, an aversive shock. This change in the cortical representation creates a heightened sensitivity to the tone that predicts shock. This sensory tuning with emotional learning depends on the amygdala. If the amygdala is damaged, fear learning does not occur, and the representation of the tone in the auditory cortex is unchanged. It is interesting that once fear learning has occurred, this change in the auditory cortex will last, even if new learning diminishes the fear response to the tone or the amygdala is damaged and the learned fear is no longer expressed. In other words, the amygdala is needed to create this change in the auditory cortex to an emotional stimulus, but it is not required to maintain it. A recent study that used functional magnetic resonance imaging (fMRI) in humans showed a similar lasting change in the auditory cortex in response to a tone that predicted a shock. Although it has been suggested that a change in the cortical representation of stimuli in the visual cortex might also occur with emotional learning, this has yet to be explored. It is clear, however, that one way that emotion can influence attention is by creating a lasting change in the cortical representation of stimuli that have acquired emotional properties through learning.

The second mechanism that has been proposed to underlie emotion's influence on attention is more transient and has been demonstrated primarily in studies of visual processing. Anatomical studies have shown that there are reciprocal connections between the amygdala and the sensory cortices, including the visual cortex. It is proposed that early in stimulus processing, the amygdala receives input about the emotional significance of a stimulus and, through projections to the visual cortex, modulates further attentional and perceptual processes. Three lines of evidence from brain imaging studies in humans support this model: (1) studies demonstrating that emotional stimuli are processed automatically, (2) findings of enhanced responses in visual cortical regions for emotional stimuli, and (3) findings demonstrating that the amygdala mediates this enhanced cortical response.

The idea that emotional stimuli are processed automatically is consistent with suggestions in the psychological literature that the emotional qualities of stimuli are detected quickly and prior to awareness. Similarly, it is proposed that the amygdala receives input about the emotional significance of a stimulus quickly and prior to awareness. In support of this hypothesis, fMRI studies have demonstrated robust amygdala activation in response to subliminally presented images of faces with fear expressions, in contrast to the response to neutral expressions. In addition, a number of studies have shown that attention and awareness have little impact on the amygdala's response to fearful stimuli. It is suggested that this early, automatic amygdala response to fear or threat stimuli is an important factor in the amygdala's ability to modulate further cognitive processes and responses to potential danger.

One of the first brain imaging studies to examine emotion demonstrated that visual cortical regions show enhanced activation to novel emotional scenes compared with novel neutral scenes. This finding has also been observed in comparisons of responses in the visual cortex to faces with fear versus faces with neutral expressions. The studies of facial expression have also demonstrated an enhanced amygdala response to fearful versus neutral faces, the magnitude of which is correlated with the enhanced visual cortex activation to these same stimuli. This line of research is consistent with the proposed mechanism by which the amygdala detects the emotional significance of a stimulus quickly and modulates further perceptual and attentional processing. However, one drawback of these studies is that most brain imaging findings are correlational in nature and thus cannot determine the precise function of a given brain region or the

direction of the communication between two regions. In an effort to demonstrate that the amygdala's input mediates the enhanced response observed in the visual cortex to emotional stimuli, Vuillimier and colleagues combined functional imaging and lesion methodologies. During brain imaging, patients with varying degrees of damage to the amygdala were shown pictures of faces with fear and neutral expressions. The amount of amygdala damage predicted the magnitude of the response in the visual cortex: the more damage to the amygdala, the smaller the enhanced visual cortex response to fear faces. This finding strongly supports the suggestion that the amygdala mediates the enhanced response in the visual cortex.

If the amygdala is responding to the emotional significance of a stimulus in the environment quickly and prior to awareness, and modulating further processing in the visual cortex, then certain properties of emotion's influence on attention should emerge. First, any enhanced visual response should be transient and evoked by the presence of an emotional stimulus. Second, the enhanced response with emotion should be apparent in perception, since this is the job of the visual cortex. This enhanced perception may underlie the observed facilitation of attention in that it requires less attentional resources to detect something that is easier to perceive. Finally, the enhanced processing with emotion is not necessarily tied to the emotional stimulus itself, but should extend to neutral stimuli that are cued by emotional stimuli and presented close in time.

To explore whether these properties emerge, researchers have used a psychophysical paradigm to examine the effect of a fearful face cue on sensitivity to changes in contrast – that is, the subtle gradations of gray that help us differentiate edges of lines. The ability to detect contrast is an early perceptual function that has been linked to primary visual cortex. In this study, a fear or neutral face cue was presented in one or four quadrants of a computer screen. Then 40 ms later, four patches with lines appeared in the corners of the screen (adjacent to the location of faces). The lines in one of the four patches were tilted, and the participants had to indicate whether the lines were tilted to the right or to the left. If only one face was presented, then the patch with tilted lines was presented in the same quadrant as the face. This single-face condition allowed participants to covertly attend to the location of the subsequent patch. Across trials, the contrast in the patches varied so that the ability to detect the orientation of the lines was easier or harder. Two effects for the emotional expression of the face cue were observed. First, irrespective of whether the face cue directed covert attention, a fear face enhanced contrast sensitivity. That is, less contrast was needed to detect the orientation of the lines if the lines were preceded by a fear face rather than by a neutral face. Second, if a fear face cue directed covert attention, contrast sensitivity was enhanced more than would have been predicted by the independent effects of a fear face cue and covert attention on perception. In other words, emotion potentiates the perceptual benefit of covert attention. These psychophysical results are consistent with a model in which the emotion, via the amygdala, modulates processing in early visual regions in a transient manner for stimuli cued by emotion.

The amygdala, through its extensive connectivity with sensory processing regions, is ideally situated to influence perception. The existing evidence supports two independent means by which emotion influences perception, one of which is long lasting and affects perception of an emotional stimulus that predicts an aversive outcome, and the other of which is transient and can affect the perception of stimuli cued by emotion. It is suggested that the amygdala's influence on perception may underlie emotion's facilitation of attention. Discussions of the adaptive function of emotion's facilitation of attention emphasize the preferential detection of stimuli that signal potential importance or threat and suggest a primary role for the amygdala is the modulation of vigilance in the presence of these stimuli. The current evidence of emotion's influence on attention and perception is largely consistent with this interpretation.

What is less clear at this point is whether these mechanisms that lead to the facilitation of perception with emotion also play a role in any perceptual cost that may emerge for other stimuli. It has been suggested that emotion can capture attention. When attention is captured by emotion, there is impaired processing of nonemotional aspects of the stimulus or event. A number of behavioral studies have demonstrated the capture of attention with emotion, and it is suggested that this effect is due to difficulty disengaging attention from the emotional qualities of a stimulus. Only a few studies have examined the neural systems underlying the capture of attention with emotion, and these suggest that emotion may decrease activity in regions of the parietal lobe thought important for the allocation of attention. It is not known how this effect may or may not be related to the facilitation of attention and perception observed when people are processing emotional aspects of a cue or stimulus.

Emotion's Influence on Memory

"An impression may be so exciting emotionally as almost to leave a scar upon the cerebral tissues." This

phrase is attributed to William James in his classic textbook, *Principles of Psychology*, which was first published in 1890. Although James did not have the benefit of experimental research to support his claims, his intuition was that emotion should strongly influence how impressions are remembered. This intuition has been largely supported by scientific research since 1890. The questions remain, how does the brain enable this 'scar' to form, and what are its qualities?

Memory can be broken down into a series of stages. The first, called encoding, occurs when a stimulus is first encountered. After the encoding stage, the memory is stored. Although storage may seem like a passive stage, it is known that during storage, processes occur over time that allow the memory to become more or less stable. This time-dependent storage process is called consolidation. Finally, when a memory is called to mind, it is retrieved. There is evidence that emotion might influence all three of these stages of memory processing. Although episodic memory critically depends on other brain regions, most notably the hippocampus and surrounding cortices (see **Figure 1**), the amygdala appears to play an important role in modulating the neural circuitry of memory with emotion.

To start with the first stage, emotion can influence the encoding of to-be-remembered stimuli through its modulation of attention and perception. Research has indicated that manipulations of attention will significantly affect memory encoding. Therefore, the changes in attention and perception that occur with emotion, outlined earlier, should lead to differences in memory performance. In support of this hypothesis, a recent study found that patients with amygdala damage show impairments only in memory for details of emotional scenes that are central to the event, with intact memory for details that are more peripheral. This study suggests that the amygdala may be involved in highlighting and narrowing attention around central, emotional details, leading to enhanced encoding of these details in normal study participants. However, most studies to date have failed to differentiate the amygdala's influence on encoding and consolidation. In fact, with many studies, it is not clear how to differentiate emotion's impact on encoding and consolidation since both are assessed with retrieval and occur close in time. Although a number of brain imaging studies have demonstrated that amygdala responses during encoding can predict later recognition or recall for emotional stimuli, these studies often attribute this effect to emotion's influence on the modulation of memory storage rather than on attention and perception.

The primary neural mechanism that has been explored in an effort to understand emotion's influence on episodic memory is the amygdala's modulation of consolidation. Evidence across species suggests that the consolidation of episodic memory critically depends on the hippocampus. Emotion, specifically arousal, is proposed to enhance hippocampal-dependent consolidation. Research by James McGaugh and colleagues conducted in rats has identified the neural mechanisms underlying the effect of arousal on memory consolidation. These studies have shown that physiological arousal results in activation of β-adrenergic receptors in the amygdala. The amygdala, in turn, modulates hippocampal processing, resulting in enhanced consolidation or storage for events that elicit an arousal response. Damage to the amygdala does not impair episodic memory for a stimulus, but it eliminates any enhancement observed with physiological arousal. Evidence that this effect is due to the modulation of consolidation, as opposed to encoding, comes from studies demonstrating that manipulations of amygdala function after stimulus encoding will alter arousal's influence on episodic memory. McGaugh suggests that one adaptive function of having a slow consolidation process is to allow for the emotional reaction to a stimulus, which occurs after presentation, to influence memory strength. In this way, events that result in an emotional response, and are more likely to be important for future survival, are less likely to be forgotten.

Consistent with the proposed mechanisms outlined by McGaugh and colleagues, a number of studies in humans have reported that arousal enhances the retention of episodic memories so that they are less likely to be forgotten over time. Evidence that the human amygdala plays a role in this enhanced memory with arousal comes from studies using a range of cognitive neuroscience techniques. As mentioned earlier, brain imaging studies have reported that activation of the amygdala at encoding can predict later retention for emotional stimuli. The amygdala has direct projections to the anterior portion of the hippocampus. A recent study found that the activation of the amygdala and anterior hippocampus is correlated during the encoding of emotional scenes that were later remembered. In addition, patients with amygdala damage fail to show the normal enhancement of episodic memory with arousal. Consistent with a role for the amygdala in modulating storage or consolidation, patients with amygdala damage show similar forgetting curves for arousing and neutral stimuli, in contrast to normal controls, who show enhanced retention for arousing stimuli. In addition, pharmacological manipulations in humans that block β-adrenergic receptors also block arousal's impact on episodic memory, consistent with animal models. Taken together, these studies provide strong support for the mechanism proposed by McGaugh and

colleagues suggesting the amygdala modulates hippocampal consolidation with arousal.

Although there is substantial evidence that emotion can influence later memory accuracy, a few limitations should be mentioned. First, the magnitude of the effect of emotion on episodic memory varies depending on the paradigm. It can be quite subtle and is not always observed. Second, in addition to arousal, other factors that vary with emotion, such as the similarity of to-be-remembered stimuli, can also influence later memory. These other factors most likely achieve their influence on memory through different neural mechanisms. Finally, more-extreme stress can have an opposite effect; that is, an impairment of hippocampal function and episodic memory accuracy. In spite of these caveats, there is robust evidence that the human amygdala, through its modulation of hippocampal consolidation, plays a critical role in situations in which physiological arousal leads to enhanced episodic memory.

The final stage of memory is retrieval, when the stored information is called to mind. Processes that occur at retrieval reflect the outcome of encoding and storage, as well as additional functions related to judgment and memory monitoring. Although most studies investigate memory accuracy, there is also research indicating that the retrieval of emotional stimuli may be marked by a difference in subjective quality. Studies of episodic memory of real-life, public, emotional events have suggested that emotion may influence the subjective experience of memory retrieval, irrespective of memory accuracy. These studies examining 'flashbulb' memories have found that for highly emotional, public events, the confidence that a memory is accurate and the sense that it is detailed and vivid may not reflect actual accuracy for details. One of first examples of this effect examined memory of the explosion of the space shuttle *Challenger*. Within a few days of this tragedy, study participants were asked to record their memory for the circumstances in which they become aware of this event. Two and a half years later, the participants were asked to report the same memory. Even though the participants gave detailed recollections and were highly confident in their accuracy, most of these memories were inaccurate. More recently, a study examining memory for the terrorist attack of September 11, 2001, found that accuracy for memories of learning about this event did not differ from other, nonemotional events that occurred around the same time. However, memories of the terrorist attacks, in comparison to the more mundane events, were rated as more confident, vivid, detailed, and recollected. These studies of flashbulb memories suggest that emotion has an independent effect of enhancing the subjective sense of remembering during retrieval.

In an effort to explore the mechanisms underlying the enhanced subjective sense of remembering with emotion, laboratory studies have used the 'remember–know' paradigm. Recognition memory judgments are thought to rely on two independent processes: recollection, which includes the retrieval of contextual details, and familiarity, which is a sense that a stimulus is familiar in the absence of contextual details. The remember–know procedure is a subjective measure of these two processes. During recognition judgments, study participants view old and new stimuli and are asked to judge whether each stimulus is 'new' (not presented before), 'known' (familiar, but there is no specific recollection of details for the encoding context), or 'remembered' (recollected with details of the encoding context). For neutral stimuli, it has been demonstrated that the subjective judgment of 'remember' is correlated with accuracy for memory of the contextual details of the stimulus, suggesting that the accurate recollection of details and the subjective judgment of remembering rely on the same processes. For emotional stimuli, it has been found that the subjective judgment of remembering is enhanced, but this subjective judgment of recollection is not as closely linked to a corresponding increase in accuracy of memory of the contextual details. In other words, emotion may enhance the feeling of remembering independent of its impact on memory accuracy.

Using a similar paradigm with fMRI, researchers have demonstrated that different neural mechanisms underlie the subjective judgment of remembering for pictures of emotional as opposed to neutral scenes. Consistent with previous studies, responses in a part of the hippocampal complex, the posterior parahippocampus, differentiated 'remember' and 'know' judgments for neutral scenes. In contrast, signals in the amygdala differentiated 'remember' and 'know' judgments for emotional scenes. The posterior parahippocampus has previously been linked to memory accuracy for details of visual scenes, which is the type of information that might be expected to result in a judgment of 'remember' if this subjective judgment reflects accurate memory for scene details. However, responses in this region were not similarly enhanced for 'remembered' emotional stimuli. These results suggest that the neural mechanisms underlying the subjective judgment of recollection differ between emotional and neutral scenes. The amygdala is specifically linked to judgments of recollection for emotional scenes.

In a recent study, a similar pattern was observed when participants who were in Manhattan on September 11, 2001, retrieved autobiographical memories from that day or the summer of 2001. Although

memory accuracy could not be verified in this study, it was found that those participants who were closer to the World Trade Center (WTC) on that day reported September 11 memories that were rated as more vivid, detailed, and confident. Proximity to the WTC was not only correlated with the subjective sense of remembering but also increased activation of the amygdala and decreased activation of the posterior parahippocampus during memory retrieval. These findings are consistent with previous studies of flashbulb memories, which indicate that unique processes may underlie the judgment of recollection during the retrieval of highly emotional events.

Studies of memory retrieval suggest that for emotional stimuli, judgments of the subjective sense of remembering may be influenced by the emotional qualities of the stimulus, with less emphasis on mnemonic details. Consistent with behavioral reports of flashbulb memories suggesting emotion enhances the feeling of retrieval accuracy irrespective of actual accuracy, brain imaging results indicate emotion may alter the neural mechanisms underlying the subjective judgments of remembering. The impact of emotion on subjective quality of memory retrieval likely reflects emotion's influence on processes that occur at encoding and storage, as well as retrieval.

Although there is significant evidence that emotion interacts with episodic memory, and that the amygdala plays an important role in this interaction, an understanding of the complexity of this relationship is just beginning to emerge. Most studies in humans examining the neural systems underlying emotion's influence on episodic memory have emphasized arousal's impact on memory consolidation, perhaps inspired by the elegant animal models outlining this mechanism. However, only a few studies have clearly documented arousal's specific influence on memory consolidation in humans, independent of its effect on encoding. Disentangling the range of mechanisms by which emotion interacts with episodic memory will be a challenge for future studies. An exploration of the underlying neural systems will aid in determining the complex components of the "scar upon the cerebral tissues" that characterizes emotion's impact on human memory.

Conclusions

This article has reviewed what is known about two cognitive processes, attention and memory, that are influenced by emotion. What is clear from this limited review is that the relation between emotion and cognition is complex and nuanced. One topic that was not mentioned is the complexity of the conceptualization of emotion. Although emotion is often described as a single construct, this is clearly an oversimplification. Not only does emotion influence cognition at several stages of stimulus processing, but different facets of emotion likely have different effects on cognitive functions. In the psychological literature, the relation between emotion and cognition has often been described as two separate processes that interact at specific stages of processing. Emerging evidence from studies of human neuroscience suggest this distinction may not be sufficient. When brain function is examined, it is difficult to identify a clear dividing line between the representation of processes that are generally referred to as either emotion or cognition. Emotion affects cognition at every stage of stimulus processing, and the neural mechanisms of emotion and cognition appear to be inherently intertwined.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Structure and Circuitry in Rodents and Felines; Amygdala: Contributions to Fear; Attention: Models; Attentional Functions in Learning and Memory; Cognition: An Overview of Neuroimaging Techniques; Emotion and Vigilance; Emotion: Neuroimaging; Emotional Hormones and Memory Modulation; Hormones and Memory; Neuroimaging; Social Emotion: Neuroimaging.

Further Reading

Amaral DG (2003) The amygdala, social behavior, and danger detection. *Annals of the New York Academy of Science* 1000: 337–347.

Davis M and Whalen PJ (2001) The amygdala: Vigilance and emotion. *Molecular Psychiatry* 6: 13–34.

Dolan RJ and Vuilleumier P (2003) Amygdala automaticity in emotional processing. *Annals of the New York Academy of Sciences* 985: 348–355.

Hamann S (2001) Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences* 5: 394–400.

LeDoux JE (1996) *The Emotional Brain*. New York: Simon and Schuster

McEwen BS and Sapolsky RM (1995) Stress and cognitive function. *Current Opinions in Neurobiology* 5: 205–216.

Phelps EA (2004) Human emotion and memory: Interactions of the amygdala and hippocampal complex. *Current Opinion in Neurobiology* 14: 198–202.

Phelps EA (2006) Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology* 57: 27–53.

Phelps EA and LeDoux JE (2005) Neural systems underlying emotion behavior: From animal models to human function. *Neuron* 48: 175–187.

Pourtois G and Vuilleumier P (2006) Dynamics of emotional effects on spatial attention in the human visual cortex. *Progress in Brain Research* 156: 67–91.

Whalen PJ (1998) Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science* 7: 177–188.

Emotional Learning in Humans

L Cahill, University of California at Irvine, Irvine, CA, USA
A Anderson, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term 'emotional learning' typically refers to one of two quite different concepts. First, it may refer to the possibility that emotional responses themselves are learned. In recent years, this view was investigated most extensively in the context of what is called Pavlovian 'fear conditioning.' It has been suggested that, in such a learning situation, an emotional response (in this case, fear) is learned. Succinctly put, by this view emotion 'is' the memory. However, extensive research documents that what is learned in a typical Pavlovian fear-conditioning experiment is not an association between a stimulus and a response such as fear (termed 'S–R' learning), but a predictive relationship between two stimuli ('S–S' learning). Thus the emotion of fear appears not to be learned in fear-conditioning situations; rather, it is a response to what is learned, namely, an S–S association. Thus this first conceptualization of emotional learning appears to be a misnomer.

The second common conception of emotional learning refers to the idea that learning can occur during, and is influenced by, emotional excitement, or arousal. Broadly speaking, neurobiological research has focused on two primary factors by which emotional arousal may modulate (i.e., enhance or impair) memory storage in the brain: (1) via an 'online' modulation of attentional processing of emotional events and (2) via 'off-line' modulation of postlearning memory consolidation processing of the events. Evidence suggests that the amygdala complex, a collection of inter-connected nuclei located in the medial temporal lobe, is crucial to both types of modulation. This article reviews key findings regarding both of these modulatory factors, in particular as they concern human participants.

Emotional Modulation of Attention

Selective online processing is required for an organism to sift the wheat from the chaff of the immense amount of environmental information impinging on its nervous system. As observers can neither consciously perceive everything in their perceptual field nor remember every waking moment of their lives, information acquisition and storage must be modulated by its importance. Thus, the focusing of attention represents a prioritizing of what is important for the organism to process. It would be adaptive for the allocation of central cognitive resources to be tightly coupled with information of sufficient biological import to mobilize peripheral metabolic resources. Indeed, the classic definition (e.g., from the work of Sokolov) of the 'orienting response' to novel and important events refers to the parallel evocation of both central attentional and peripheral arousal responses. Research from the past decade has shown this coupling of arousal and attentional processes converges on the amygdala complex. Consistent with the proposed role of the central nucleus of the amygdala in arousal and attentional orienting in rodent models, recent evidence from human studies indicates that the amygdala is essential to the way emotional arousal alters attentional allocation and that this influence in turn enhances the encoding of emotionally significant events.

Emotional Arousal Influences Attention

In addition to enhancing the durability and accessibility of representations of one's past experiences in memory, it is now known that emotional arousal, by modulating the allocation of attentional resources, is also critically associated with enhancing accessibility of representations of one's present experience. Perception and memory are constructive processes. Perceivers do not provide equal access to awareness for all stimuli. This subjectivity of perception is often associated with the psychological construct of attention. The primary means by which a stimulus reaches awareness is through the allocation of attention toward its encoding. It has thus long been proposed that emotions powerfully modulate attentional processes (e.g., by narrowing the 'spotlight' of attention, as proposed in 1959 by Easterbrook). In particular, research has shown that events associated with emotional arousal acquire greater attentional resources.

Evidence showing that emotional arousal modulates attention comes from studies of what is called the 'attentional blink' (AB). Following identification of a single target stimulus, there exists a short (up to 500 milliseconds) period of impaired perceptual awareness, evidenced by the fact that a second target presented during this period may not be perceived, a phenomenon referred to as the AB. The AB impairment is substantially diminished for emotionally arousing events of both negative and positive valence and is correlated with individual differences in the subjective

arousal response to the stimuli. Also, the degree of AB reduction is found to be independent of confounding effects of familiarity, unexpectedness, semantic and perceptual distinctiveness, or altered decision criteria. Thus evidence indicates that the likelihood that a stimulus will receive attention and reach awareness is commensurate with the strength or intensity of emotional arousal induced by the stimulus, irrespective of its emotional valence. Emotionally arousing events are thus not only associated with later enhanced memory but also enhanced conscious perceptual awareness during initial encoding.

It should also be mentioned that, in addition to its effects on detection and awareness of stimuli during encoding, emotional arousal may also change the subjective phenomenological experience of a stimulus being retrieved. For example, the oft-quoted vivid 'flashbulb' quality of memories associated with highly arousing events may reflect in many instances, not the accuracy of the event, but the vividness of perceptual experience during event encoding. Similarly, some evidence suggests that 'flashbacks' that occur in posttraumatic stress disorder (PTSD) may reflect enhanced attentional allocation during the initial traumatic event combined with hyperarousal during later reexperience. Evidence from healthy people also indicates that highly arousing scenes are associated with a special visual quality that renders them more subjectively vivid.

Finally, malfunction of the mechanisms by which emotion influences attention may underlie, at least in part, various anxiety disorders. Evidence supporting this view comes in large part from studies using the 'emotional Stroop' paradigm, which reveals exaggerated attentional biases to emotional stimuli in anxiety disorders such as PTSD. It has thus been argued that the normally adaptive role of emotion in enhancing attention to emotionally salient events is rendered maladaptive when chronically activated, as in PTSD, resulting in a hypervigilant state.

Amygdala and Motivated Attention

Arousal-modulatory influences on attention and awareness are thought to depend on the integrity of the amygdala and its sensory cortical efferents. There is a striking asymmetry in amygdalo-fugal and amygdalo-petal projections, whereby the amygdala sends many more projections throughout neocortical regions, including perceptual cortices, than it receives. In fact, it has been known since the early 1950s that stimulation of the amygdala activates the cortical electroencephalograph as effectively as does the reticular activating system. The amygdala also has projections to subcortical structures important for attention and eye movements – the thalamic pulvinar nucleus and the superior colliculus, and receives input from sensory thalamic nuclei. This arrangement is consistent with the view that the amygdala receives a relatively rapid and coarse representation of the state of the world and, via its robust feedback projections to the cortex, alters the perception of stimuli, upregulating cortical encoding of stimuli of emotional importance.

Rodent models supporting the emotional modulation of perceptual processes have shown that after a few auditory fear conditioning trials, the receptive field tuning curves of neurons in the medial geniculate nucleus and primary auditory cortex are modulated such that the maximal responses are shifted toward the conditioned stimulus auditory frequency. This rapid change in neuronal tuning is consistent with the allocation of attention toward the conditioned stimulus frequency range, suggesting that the amygdala may modulate early and later perceptual processes, augmenting attention to emotionally significant events.

In humans, evidence from functional imaging studies has pointed toward a specialized attentional orienting network centered on the amygdala. For example, multiple studies have shown that emotionally arousing scenes engage both primary and secondary visual cortex significantly more than do neutral scenes. This phenomenon has been referred to as 'motivated attention' and may underlie the effects of emotion on attention discussed above. Event-related potential studies with much higher temporal resolution have similarly demonstrated that emotional significance can alter the course of perceptual encoding early in the visual processing stream, influencing visual components within the first 100 ms of stimulus processing.

These modulatory effects of arousal on attention appear to require the amygdala. For example, enhanced extrastriate cortical responses to facial expressions of fear are correlated with magnitude of amygdala response, and these enhanced extrastriate responses are abolished following amygdala damage. Direct evidence that the amygdala modulates attention to promote enhanced encoding and awareness of emotionally arousing events comes from studies of the AB, which demonstrate that emotional influences on the AB are abolished in patients with bilateral amygdala lesions. Whereas healthy people demonstrate a reduced AB to emotionally arousing events, patients with bilateral amygdala damage do not. Amygdala contributions to perceptual encoding appear to be selective for the influence of emotional arousal, as amygdala damage does not disrupt enhanced awareness associated with other manipulations of perceptual salience (e.g., increased stimulus discriminability). Furthermore, impaired arousal modulation of attentional processes in

patients with amygdala lesions is evident despite their possessing a normal ability to evaluate the arousing value of a stimulus, which suggests the amygdala is critical, not for the appraisal of event significance *per se*, but for the modulation of attention by emotional arousal. Collectively, experiments to date indicate that the amygdala facilitates perceptual encoding by influencing responsiveness of the neocortex to stimuli, thereby allowing emotional arousal to shape conscious perceptual experience.

Amygdala Modulation of Attention: Implications for Memory

Evidence suggests the brain possesses multiple attentional orienting systems with distinctive neuroanatomical and functional characteristics. Neuroimaging suggests, for example, that enhanced attention to neutral events typically activates the highest level of the cortical hierarchy, the prefrontal cortices, in conjunction with posterior parietal cortices. By contrast, as discussed, the amygdala has been associated with enhanced attention to emotionally arousing events. One critical functional difference between these frontoparietal and amygdala attentional orienting systems may be the degree to which they are voluntarily or reflexively engaged. Frontoparietal systems appear engaged during volitional attentional control, providing the flexible capacity for individuals to select what information is desired for further processing or is to be ignored. By contrast, amygdala recruitment has often been cast as relatively automatic, being recruited more readily during inattention, although the degree of this automaticity is actively debated. The amygdala may then act independently of volitional attention to ensure that emotionally salient events are encoded even when higher-order, prefrontally based attention networks are preoccupied.

The reflexive nature of amygdalar modulatory influences on attention may have important implications for memory. Human models of 'declarative' (conscious) memory suggest that attention engagement during encoding is one of the most critical predictors of the success of subsequent memory retrieval. When attention to stimuli is reduced (as, e.g., under conditions of 'divided attention'), conscious memory for neutral events is drastically reduced. In contrast, the costs to memory of reduced attention are relatively minimal for emotionally arousing events. That is, when prefrontally based attention networks are preoccupied with a concurrent task, encoding of emotionally arousing events is largely preserved. Indeed, neuroimaging evidence suggests that the influence of emotional arousal on memory formation does not necessarily depend on recruitment of the prefrontal cortices but depends more on amygdala function.

On the surface, these results may suggest that attention is not required for emotional enhancement of memory. However, as we have reviewed above, emotionally arousing events capture attention even when ignored, and this enhanced processing is orchestrated by the amygdala. Thus amygdala-based enhancement of attention at encoding, and its subsequent benefit to declarative memory, may occur even when prefrontal attentional systems are unavailable.

Emotional Modulation of Memory Consolidation

The second manner in which emotional arousal influences memory is via off-line modulation of postlearning consolidation processes. There is a very large and cohesive body of evidence documenting the ability of the amygdala (and in particular its basolateral nuclei) to modulate memory storage processes when it is stimulated soon after learning. The activity of neurons in the basolateral nuclei is elevated for hours after a single emotionally arousing event. Indeed, some experiments suggest that the amygdala may be involved in modulating memory storage for up to 2 days after an emotionally significant event. Extensive evidence also documents that the amygdala interacts with endogenous stress hormones released during and after emotional events to influence memory storage. Circulating stress hormones feed back, directly or indirectly, to the brain after an emotionally arousing event, affecting in particular the basolateral amygdala, which in turn is thought to modulate memory storage processes occurring elsewhere in the brain. This mechanism may provide an evolutionarily adaptive way to achieve memory strength that is, in general, proportional to memory importance. While this view receives its primary support from animal research, strongly convergent support comes from studies of human participants.

A few studies to date have examined the effects of postlearning stress hormone activation on memory consolidation in humans. In one study, epinephrine infusions given immediately after participants viewed a series of images enhanced long-term (1-week) memory for some of the images, in particular those associated with a small degree of arousal at initial encoding. These findings suggested the novel concept that postlearning stress hormone action interacts with the degree of arousal of experiences at initial encoding to modulate memory consolidation. This conclusion was confirmed in a study using cold pressor stress (forearm immersion in ice water) to activate the stress response immediately after learning in humans, as well as in a study of the enhancing effects of cortisol in rats. Conversely, impairment of the

adrenergic stress response (with the adrenergic blocking drug propranolol) impairs memory for emotional material.

In contrast to their generally enhancing effects on memory consolidation, endogenous stress hormones, in particular cortisol, impair retrieval of already-stored information. Studies with both animals and humans have convincingly shown that the adrenal hormone cortisol can impair the retrieval of well-stored information. This impairing effect of cortisol on retrieval, in contrast to its generally enhancing effects on consolidation, nicely illustrates the 'modulatory' concept that should be kept in mind when considering stress hormone influences on memory.

The Amygdala and Modulation of Memory Consolidation

As mentioned, a key neural element in our present understanding of emotional modulation of consolidation learning in humans is the amygdala, a structure uniquely positioned among cortical structures to widely modulate other brain regions. Many studies now indicate that, consistent with the 'memory modulation' view of the amygdala derived from animal research, amygdala activity during learning relates significantly to subsequent memory for emotionally arousing material but not to memory for nonarousing material. These studies include glucose positron emission tomography (PET), blood-flow PET, and functional magnetic resonance imaging investigations.

Studies from several labs also indicate, again consistent with the prediction from animal research, that the degree of arousal associated with stimuli (as opposed to their pleasant or unpleasant nature, or valence) is the key to amygdala engagement. In all studies to date that have directly manipulated the arousal and valence qualities of stimuli, amygdala responsiveness was significantly related to the arousing quality of the stimuli. This fact fits very well with the view that the amygdala functions especially in the context of sympathetic nervous system-activating (i.e., -arousing) situations to influence memory.

If the relationship of the amygdala to subsequent memory for emotional material is indeed dependent on adrenergic activation, then it should be eliminated in participants who view emotional material while under the influence of propranolol. And in fact, confirming evidence has come from both the studies to date that have tested this hypothesis.

Although the imaging studies mentioned to this point strongly support a selective role for the human amygdala in memory for emotional events, they do not address the possibility suggested by animal research that the amygdala influences memory storage in other brain regions. One way to produce such evidence in humans is to use analytical techniques that allow estimates of information flow between brain regions. One such method is called structural equation modeling, or 'path analysis.' Only one study to date has used path analysis to examine how the amygdala relates to other brain regions during the encoding of emotional versus neutral material. The key finding was an increased outflow from the amygdala to the ipsilateral hippocampal region in emotional compared with neutral learning conditions. Findings from other studies indicate an increased correlation between amygdala activity and that of other key brain regions, such as the hippocampus, in emotional compared with neutral learning situations. Thus the available evidence from human participants to date appears to support the memory modulation concept of amygdala function in memory.

Uncovering Influences of Sex

Finally, it is increasingly apparent that pronounced, but largely unexplored, influences of participant sex exist on brain mechanisms of emotional learning. For example, imaging studies have consistently revealed a sex-related hemispheric lateralization of amygdala activity in relation to long-term memory for emotional material. Activity of the left hemisphere amygdala while viewing emotional stimuli is preferentially related to subsequent memory for those stimuli in women, but activity of the right hemisphere amygdala is preferentially related to subsequent memory for the same stimuli in men. Other studies have revealed that sex influences stress hormone effects on memory, electroencephalograph responses to emotional stimuli, and even the nature of what is retained from an emotional experience. A straightforward, but highly important, conclusion from these studies is that investigations of the neural bases of emotional learning in humans can no longer safely assume that participant sex will not significantly influence the findings, and hence conclusions.

Summary

Emotional learning, that is, memory storage for emotionally significant events, appears to involve at least two general processes: an online modulation of attentional resources, and off-line modulation of memory consolidation processes after the event is over. Both processes seem to depend critically on the amygdala, which may be necessary, not for emotional experience *per se*, but for the effects of emotional arousal on encoding (via attentional influences) and memory consolidation. The postevent modulatory processes also involve stress hormone influences on memory

mediated by the amygdala. Finally, while these processes are likely to be broadly similar in both sexes, they also appear to differ in important ways that merit careful consideration in future work.

See also: Amygdala: Contributions to Fear; Attentional Functions in Learning and Memory; Conditioning: Theories; Declarative Memory System: Anatomy; Emotion Systems and the Brain; Emotion: Neuroimaging; Emotional Hormones and Memory Modulation; Emotional Influences on Memory and Attention; Fear Conditioning and Synaptic Plasticity; Procedural Learning: Classical Conditioning; Stress and Cognition.

Further Reading

Anderson AK (2005) Affective influences on the attentional dynamics supporting awareness. *Journal of Experimental Psychology: General* 134: 258–281.

Anderson AK and Phelps EA (2001) Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature* 411: 305–309.

Cahill L (2006) Why sex matters for neuroscience. *Nature Reviews Neuroscience* 7: 477–484.

Cahill L (2000) Emotional modulation of long-term memory storage in humans: Adrenergic activation and the amygdala. In: Aggleton J (ed.) *The Amygdala: A Functional Analysis*, pp. 425–444. Oxford, UK: Oxford University Press.

Gallagher M, Graham PW, and Holland PC (1990) The amygdala central nucleus and appetitive Pavlovian conditioning: Lesions impair one class of conditioned behavior. *Journal of Neuroscience* 10: 1906–1911.

Kapp BS, Supple WF, and Whalen PJ (1994) Effects of electrical stimulation of the amygdaloid central nucleus on neocortical arousal in the rabbit. *Behavioral Neuroscience* 108: 81–93.

Kensinger EA and Corkin S (2004) Two routes to emotional memory: Distinct neural processes for valence and arousal. *Proceedings of the National Academy of Sciences of the United States of America* 101: 3310–3315.

McGaugh JL (2004) The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience* 27: 1–28.

Pare D (2003) Role of the basolateral amygdala in memory consolidation. *Progress in Neurobiology* 70: 409–420.

Strange BA, Hurlemann R, and Dolan RJ (2003) An emotion-induced retrograde amnesia in humans is amygdala- and beta-adrenergic-dependent. *Proceedings of the National Academy of Sciences of the United States of America* 100: 13626–13631.

Van Stegeren AH, Goekoop R, Everaerd W, et al. (2005) Noradrenaline mediates amygdala activation in men and women during encoding of emotional material. *Neuroimage* 24: 898–909.

Vuilleumier P, Richardson M, Armony JL, Driver J, and Dolan RJ (2004) Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature Neuroscience* 7: 1271–1278.

Young MP and Scannell JW (1994) Analysis of connectivity: Neural systems in the cerebral cortex. *Reviews in the Neurosciences* 5: 227–250.

Encephalitis

J F Bale Jr., University of Utah School of Medicine, Salt Lake City, UT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Encephalitis denotes inflammation or infection of the brain parenchyma. Because meningeal inflammation commonly accompanies encephalitis, many authorities employ the term 'meningoencephalitis' when discussing encephalitis. Encephalitis has many distinct etiologies, including infections with parasites, bacteria, fungi, and viruses, and the clinical manifestations of encephalitis can be mimicked by several noninfectious conditions, including acute stroke, metabolic or inflammatory disorders, and neoplasia. This article focuses on the most common cause of encephalitis, viral infections. Despite remarkable advances in preventing many viral causes of encephalitis (such as polio, measles, and mumps) virus encephalitis remains a major potential source of morbidity and mortality among the world's peoples.

Many viruses can cause encephalitis (Table 1), although an etiologic agent is never identified in as many as 50% of encephalitis cases presumed due to virus infections. The incidence of encephalitis in young children, perhaps the most vulnerable age group, is as high as 18 per 100 000 child-years. Prior to the recent outbreaks of West Nile virus (WNV) infections, the Centers for Disease Control and Prevention (CDC) received reports of up to 2000 cases of viral encephalitis annually, corresponding to 0.5 cases of encephalitis per 100 000 persons living in the United States. More than 4000 human cases of WNV infection, an Old World flavivirus first observed in the United States in 1999, were identified in 2003; many were associated with neuroinvasive disease. Approximately 2500 cases of herpes simplex virus type 1 (HSV-1) encephalitis, a nonreportable disorder, occur annually among persons of all ages in the United States. Asymptomatic infections with many neurotropic viruses exceed symptomatic cases by ratios of 50 to 1 or greater.

Virus encephalitis begins with virus entry and replication at sites distant from the nervous system. Enteroviruses invade via the nasopharynx or gastrointestinal tract, herpesviruses enter the host via the mucosa of oropharynx or genital tract, and arboviruses replicate initially in the subcutaneous or lymphoid tissues near the site of virus inoculation by the tick or mosquito. Local factors, such as pH, mucosal immune responses, or the integrity of skin and mucosal barriers, influence whether a specific viral pathogen will cause encephalitis.

Many viruses reach the central nervous system (CNS) hematogenously and invade the CNS through the choroid plexus or vascular endothelial cells. Rabies virus infects neuromuscular junctions, enters nerve endings, and travels retrograde by axonal transport to the neurons of the dorsal root ganglia and spinal cord. On reaching the spinal cord, the rabies virus then ascends axonally to the brain, producing behavioral changes, seizures, and coma. Transmission via axons also has a role in pathogenesis of HSV-1 encephalitis and in reactivations of HSV-1, HSV type 2 (HSV-2), and varicella-zoster virus.

Adsorption to and penetration of human cells, the initial events of viral infection, depend on virus contact with the cell and attachment of the virus proteins to receptors on the cell surface. Although the precise molecular determinants of neuroinvasiveness and neurotropism – the capacity to infect cells of the nervous system – have not been determined for all viral pathogens, neurovirulence – the capacity to cause neurologic disease – has been linked to several viral genes and gene regions. Viruses enter host cells by phagocytosis, direct release of viral nucleic acids, or fusion of the virus envelope with the host cell membrane. Viruses utilize host cell molecular pathways to produce virus-encoded nucleic acids and proteins. Specific replicative strategies vary among viruses, depending on the DNA or RNA composition of the infecting virus. Production of viral progeny involves assembly of virions in the cell nucleus or cytoplasm and release of mature virions by budding or cell lysis. Cell death can result from stimulation of apoptotic pathways.

Clinical Manifestations

The majority of infants, children, or adults who experience virus encephalitis lack recognized predisposing factors. The incubation period for virus-induced neurologic disorders ranges from a few days to 3 or more weeks. Encephalitis begins with nonspecific, systemic features such as malaise, anorexia, myalgia, low-grade fever, vomiting, or headache. Certain viral infections can be associated with rash, diarrhea, arthralgia, pharyngitis, cough, or abdominal pain. Persons with rabies, a rare disorder in the United States and other developed countries, may have pain or paresthesia near the site of virus inoculation.

Persons with viral encephalitis have alterations of consciousness varying from mild somnolence to deep coma. Fever, a common but not invariable feature of

Table 1 Selected encephalitis-causing viruses and preferred diagnostic studies

Suspected pathogen	*Diagnostic study*
Herpesviruses	
Herpes simplex virus type 1	PCR: CSF
Herpes simplex virus type 2	PCR: CSF, serum (neonates)
Epstein–Barr virus	Serology; PCR: CSF
Varicella-zoster virus	PCR: CSF; antibody studies: CSF
Cytomegalovirus	PCR: CSF or urine; urine or saliva culture
Human herpes virus 6	PCR: CSF
Human herpes virus 7	PCR: CSF
Arthropod-borne viruses (alphaviruses, flaviviruses, bunyaviruses, orbiviruses)	
Western and Eastern equine encephalitis viruses	Serology; CSF IgM
St. Louis encephalitis virus	Serology
Japanese encephalitis virus	Serology; RT-PCR: CSF
West Nile virus	Serology; CSF IgM
LaCrosse virus	Serology; CSF IgM
Tick-borne encephalitis virus group	Serology
Colorado tick fever	Serology
Nonpolio enteroviruses	
Echo and Coxsackie viruses	RT-PCR: CSF, serum
Enterovirus 71	Culture of stool and oropharyngeal secretions; PCR: CSF
Other	
Rabies virus	Serology; IFA
Measles	Serology: RT-PCR: CSF

CSF, cerebrospinal fluid; IFA, immunofluorescence antibody staining; PCR, polymerase chain reaction; RT-PCR, reverse transcription PCR.

viral encephalitis, ranges from low grade to 40 °C or higher. Seizures, a frequent sign of encephalitis, can be partial or generalized. Although focality can be an important feature of encephalitis due to HSV-1, partial seizures accompany several types of encephalitis, including relatively mild cases caused by the La Crosse virus or the nonpolio enteroviruses.

The neurologic examination of individuals with encephalitis reveals behavioral or cognitive changes, hyperreflexia, altered muscle tone, ataxia, or focal deficits such as aphasia, hemiparesis, or ophthalmoplegia. Not all features are present in every patient. Focal neurologic features should increase the clinician's suspicion of HSV encephalitis, especially in adults older than 50 years. In severe cases of encephalitis, signs of increased intracranial pressure, including pupillary abnormalities, ophthalmoplegia, abnormal respiratory patterns, or flexor or extensor posturing, may be present.

Diagnosis

The patient with suspected virus encephalitis requires a thorough evaluation that includes neuroimaging studies, cerebrospinal fluid (CSF) examination, electroencephalography, and microbiological assays for suspected infectious pathogens. The clinician must search for abnormalities that indicate a specific, treatable viral disorder, such as HSV encephalitis, and remain alert for the clinical or laboratory findings that suggest noninfectious disorders or infections with bacteria, fungi, or parasites that may require pathogen-specific treatment.

Computed Tomography and Magnetic Resonance Imaging

Imaging studies should be performed in all patients with presumed encephalitis, especially when HSV encephalitis is suspected. The computed tomography (CT) of patients with encephalitis can be normal, have nonspecific features, or demonstrate focal or generalized cerebral swelling manifested by loss of the gray–white interface or effacement of the basal cisterns. In patients with HSV encephalitis, CT may detect focal edema of the temporal lobe, compression of adjacent CSF spaces, focal hemorrhage, or abnormal contrast enhancement.

Magnetic resonance imaging (MRI), the preferred imaging modality for patients with suspected encephalitis, has superior sensitivity, especially in the early stages of virus encephalitis. T2 prolongation (Figures 1 and 2) or enhancement of the mesial temporal lobe, insular cortex, and cingulate gyrus strongly suggests HSV encephalitis. Administration of paramagnetic contrast agents, such as gadolinium-DTPA, can enhance brain lesions in HSV and other forms of viral encephalitis (Figure 3), but the absence of enhancement does not exclude these disorders. MRI detects diffuse

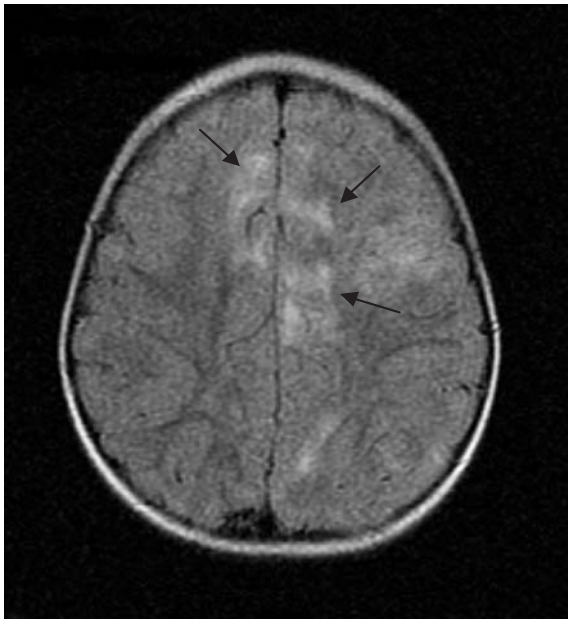

Figure 1 Axial FLAIR magnetic resonance imaging in varicella-zoster virus encephalitis shows several areas of T2 prolongation compatible with acute disseminated encephalomyelitis (arrows).

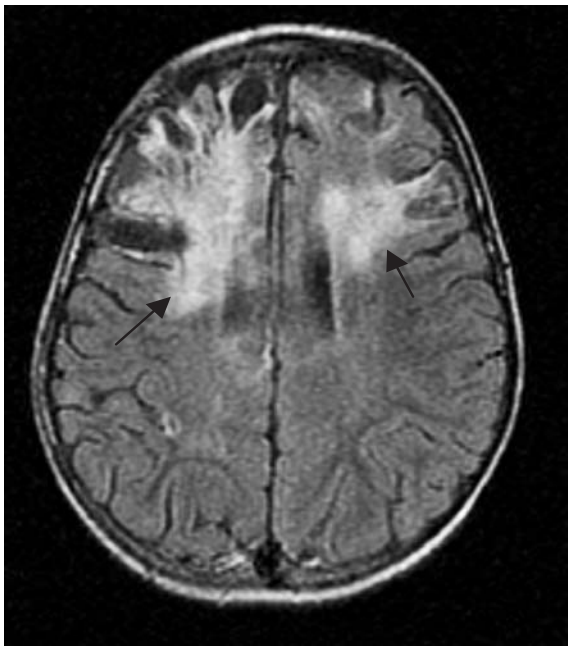

Figure 2 Axial FLAIR magnetic resonance imaging in a child who recovered from Epstein–Barr virus encephalitis shows residual gliosis (arrows) and cystic encephalomalacia.

or focal abnormalities in other forms of virus encephalitis, and in certain disorders, such as Japanese encephalitis, the MRI features can be diagnostic. MRI also has utility in distinguishing viral encephalitis from acute disseminated encephalomyelitis (ADEM), a condition that accounts for approximately 15% of

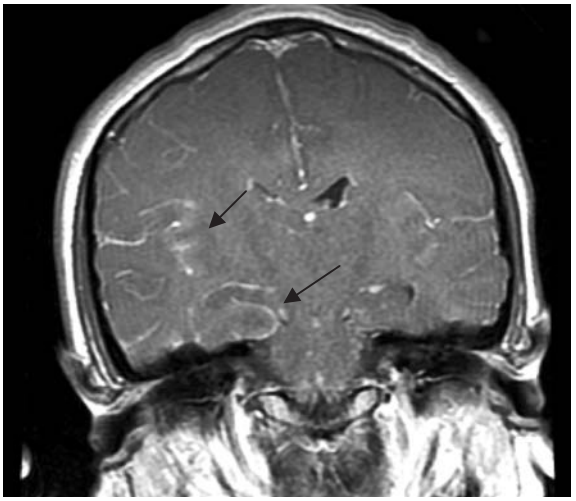

Figure 3 Gadolinium-enhanced coronal T1-weighted magnetic resonance imaging in acute herpes simplex virus type 1 encephalitis shows meningeal enhancement adjacent to the insular and mesial temporal regions (arrows).

encephalitis cases in the United States. MRI in ADEM shows multiple areas of T2 prolongation in the white matter of the cerebrum, brain stem, or cerebellum and occasionally in the cerebral cortex, basal ganglia, and thalamus.

Patients with encephalitis commonly have abnormal electroencephalograms (EEGs) that demonstrate slowing of background rhythms or the presence of focal or diffuse epileptiform discharges. Detecting focal features that may suggest HSV encephalitis is an important reason to obtain an EEG during the early stage of the patient's encephalitic illness. Of patients older than 5 months studied by the National Institute of Allergy and Infectious Diseases-supported Collaborative Antiviral Study Group (NIAID-CASG), 81% with biopsy-proven HSV encephalitis had focal slowing or repetitive epileptiform discharges (periodic lateralizing epileptiform discharges). In neonates with HSV infection, EEGs are less frequently abnormal. In nonherpetic encephalitis, the EEG usually shows diffuse slowing, but focal features can be present. Approximately 40% of the children with California encephalitis exhibit focal slowing or epileptiform discharges, and nearly 50% of patients with Eastern equine encephalitis (EEE) can have focal EEG slowing. In severe cases of encephalitis, regardless of the etiologic agent, the EEG can reveal marked attenuation of background activity or a burst suppression pattern.

CSF

The CSF in virus encephalitis usually shows a lymphocytic pleocytosis, modestly elevated protein content, and normal glucose content. Total CSF leukocyte

ranges from 10 per cubic millimeter to 1000 per cubic millimeter or higher. However, patients exhibit variability in their CSF responses, and the CSF can be entirely normal during the early phase of encephalitis. Approximately 5–15% of patients with HSV encephalitis have normal CSF cell counts (<5 per cubic millimeter) initially, although pleocytosis, as well as protein elevation, typically appears. In certain viral disorders, such as EEE, the CSF may show a neutrophilic pleocytosis, thus resembling bacterial infection. Other leukocytes (such monocytes, plasmacytoid cells, and eosinophils) may be present, and ependymal cells can occasionally be detected. Erythrocytes can also be observed, although a hemorrhagic CSF has little diagnostic specificity. In some patients, the CSF examination must be deferred or omitted due to increased intracranial pressure.

In typical cases of virus encephalitis, the protein content is normal or modestly elevated (50–100 mg dl^{-1}), but it can exceed 500 mg dl^{-1}. Although the glucose content tends to be normal in most viral CNS infections, modest hypoglycorrachia can be seen occasionally in infections due to HSV and mumps, lymphocytic choriomeningitis, or EEE viruses. CSF abnormalities, especially leukocytosis and elevated protein content, may persist for several weeks after symptoms subside.

Microbiological Studies

Although molecular methods of detection have replaced conventional virus isolation in the diagnosis of many forms of encephalitis, viruses can be isolated in certain conditions from throat washings; urine; whole blood or buffy coat leukocytes; CSF; fluid from any skin lesion; or swabs of the nasopharynx, throat, or rectum. Tissues obtained by biopsy or autopsy can also be sent for virus culture, and such approaches should be considered in persons with HIV/AIDS or unusual neurologic disorders of suspected infectious etiology. Specimens for detection of bacterial, fungal, or parasitic pathogens may also be appropriate, depending on the clinical manifestations of the neurologic condition or immune status of the host.

The polymerase chain reaction (PCR) can detect the nucleic acid of numerous viral pathogens, including HSV, HIV, cytomegalovirus, the nonpolio enteroviruses, adenoviruses, Epstein–Barr virus, varicella-zoster virus, human herpes virus type 6, and certain alpha- and flaviviruses. Consequently, PCR has become the 'gold standard' for the detection of many encephalitis-causing viral pathogens. Analysis of CSF by PCR for DNA viruses is particularly useful in suspected cases of HSV encephalitis, and reverse transcription (RT)-PCR can be used to detect RNA viruses, such as the nonpolio enterovirus. Multiplex PCR methods have been developed to detect the nucleic acids of several neurotropic viruses simultaneously. The sensitivity of CSF PCR exceeds 90% in HSV encephalitis in children or adults but may be lower in neonatal HSV infections. PCR detects HSV in the CSF of approximately 70% of infants with HSV encephalitis and in serum of approximately 50%. Detection of HSV in serum reflects the hematogenous route of HSV transmission in neonatal encephalitis. The sensitivity and specificity of CSF or serum PCR in other viral encephalitides are less well established, although positive results should generally be considered strong evidence for infection. Conversely, false negative results indicate that clinicians must exercise sound clinical judgment when making therapeutic decisions in persons with suspected virus encephalitis (Figure 2).

Serologic studies remain the recommended diagnostic methods for identifying viruses that cannot be readily cultured or detected with the current molecular methods. Lymphocytic choriomeningitis, Epstein–Barr virus, and many arboviruses require serologic studies for diagnosis. Serologic diagnosis relies on detecting seroconversion, pathogen-specific IgG or IgM in serum or CSF, or fourfold or greater changes in virus-specific IgG in sera obtained during the acute and convalescent stages of infection. Infections with certain viral pathogens, especially the arthropod-borne agents, including the WNV, can be established by detecting agent-specific IgM in a CSF sample obtained during the acute stage of the illness.

Many nonviral and noninfectious disorders produce neurologic signs and symptoms that overlap with those of virus encephalitis. Consequently, the clinician must occasionally obtain studies for nonviral disorders. Such disorders include stroke, tumor, bacterial infections (e.g., *Mycoplasma pneumoniae* and Legionnaire's disease), inflammatory disorders (e.g., systemic lupus erythematosis), paraneoplastic syndromes, spirochetal infections (e.g., Lyme disease and syphilis), rickettsial disease (e.g., Rocky Mountain spotted fever), erhlichiosis, cat scratch disease, and parasitic infections (e.g., toxoplasmosis or amebic meningoencephalitis).

Treatment and Prognosis

Although many viral infections cannot be treated with specific therapeutic agents, increasing numbers of medications may benefit children or adults with viral CNS disorders (Table 2). All patients with viral CNS disorders require supportive care that must be tailored to the severity of their illness. Seizures, increased intracranial pressure (ICP), and acute respiratory failure should be anticipated as potential complications of viral encephalitis.

Table 2 Antiviral therapies

Virus	*Drug*
Herpes simplex virus	Acyclovir
Varicella-zoster virus	Acyclovir, valacyclovir, famciclovir
Cytomegalovirus	Ganciclovir, foscarnet, cidofovir
Human herpes virus types 6 and 7	Ganciclovir, foscarnet
Influenza	Amantadine
Measles	Ribavirin[a]
Enteroviruses	Pleconaril[b]
La Crosse virus	Ribavirin[a]

[a]Benefit not established definitively.
[b]Experimental.

Table 3 Approximate morbidity and mortality rates for selected viral encephalitides

Virus[a]	*Mortality*	*Morbidity*
HSV-1 (child, adult)	5%, 10%	50%
HSV-1 (neonatal)	15%	>50%
HSV-2 (neonatal)	15%	>50%
Epstein–Barr virus	<5%	20%
La Crosse virus	1% or less	10%
West Nile virus	10%	40%
Rabies	nearly 100%	

[a]HSV, Herpes simplex virus.

ICP contributes to the potential morbidity of viral encephalitis in children and adults. Measures that may be necessary to treat elevations of ICP include mechanical ventilation or intravenous administration of osmotic diuretics. Severely ill patients may require invasive procedures for monitoring of intracranial, arterial, and central venous pressures. Intravenous fluid and electrolyte therapy must be managed meticulously to maintain cerebral perfusion but avoid iatrogenic elevations of ICP. Seizures may require intravenous therapy with benzodiazepines, fosphenytoin, or barbiturates.

The prognosis in virus encephalitis depends on the etiologic agent and host factors, such as age, underlying medical conditions, and immune status (Table 3). La Crosse virus encephalitis has very low rates of mortality or morbidity. By contrast, human rabies encephalitis virtually always results in death despite vigorous supportive measures. Mortality in untreated HSV encephalitis ranges from 40% to 70%, whereas therapy with acyclovir reduces mortality to 20% or less. Unfortunately, many survivors of HSV encephalitis have neurologic sequelae (Figure 4) despite appropriate management and treatment with acyclovir. Survivors of virus encephalitis can have epilepsy, paralysis, movement disorders, and cognitive or behavioral dysfunction.

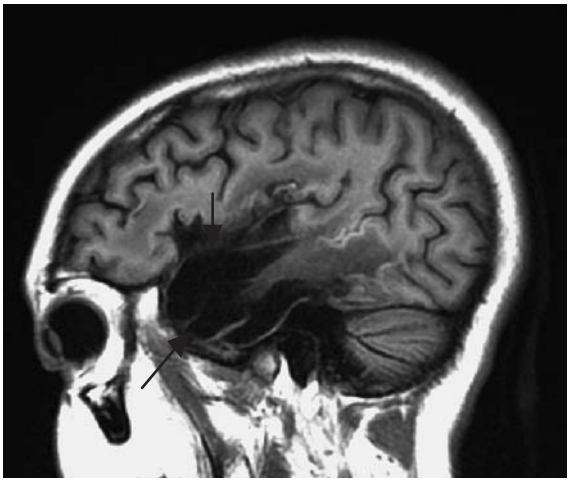

Figure 4 Follow-up sagittal T1-weighted magnetic resonance imaging in herpes simplex virus type 1 encephalitis shows residual cystic encephalomalacia of the right temporal lobe (arrows).

Selected Disorders

The following material summarizes the key clinical and therapeutic features of several representative disorders, including La Crosse virus encephalitis, HSV encephalitis, WNV encephalitis, and rabies. These disorders have been selected because they provide useful examples of the clinical features, diagnosis, and management of virus encephalitis.

Lacrosse Virus Encephalitis

The prototype California encephalitis virus was originally isolated in Kern County, California, but the La Crosse virus, endemic to the midwestern and eastern United States, causes the vast majority of encephalitis cases due to North American bunyaviruses. *Aedes triseriatus*, a mosquito species that inhabits stagnant water in tree holes or abandoned automobile tires, is the vector for La Crosse virus. Small mammals, such as chipmunks and squirrels, serve as amplifying hosts. La Crosse encephalitis (LE) typically affects previously healthy young children and occurs principally during the summer and early fall, when mosquitoes are abundant. Children with LE often live adjacent to or have recently entered mosquito habitat. As many as 30% or more of people living in endemic areas possess antibody to the La Crosse virus, indicating that infections are frequently asymptomatic. Until the recent emergence of WNV neuroinvasive disease, La Crosse virus was the most commonly encountered arboviral encephalitis in the United States.

LE typically occurs in children between the ages of 5 and 10 years; LE after 18 years of age is unusual. After an incubation period of 3–7 days, fever, vomiting, headache, and abdominal pain begin, and CNS

signs appear 2–4 days thereafter. Approximately, 60% of children with LE have seizures, 50% have altered mental status, and 20% have focal deficits, consisting of aphasia, mild extremity paralysis, or partial seizures that can mimic HSV encephalitis. Other than fever, systemic signs are uncommon, although cough, arthralgia, and transient maculopapular rash have been reported. Neurologic abnormalities and fever continue for 3–7 days; most patients recover within 10–14 days of disease onset.

The child with LE can have an elevated sedimentation rate and peripheral leukocytosis, often with a left shift. The CSF findings consist of a mild lymphocytic pleocytosis, normal or modestly elevated protein content, and normal glucose content. The EEG shows diffuse slowing of background activity, although focal or epileptiform features can be detected. CT or MRI is usually normal. The diagnosis of LE can be confirmed by detecting La Crosse virus-specific IgM in serum or CSF. Though the virus is not typically found in CSF or body fluids, it has been isolated from brain tissue in rare instances. The RNA of La Crosse and other California serogroup viruses can be detected by RT-PCR.

Therapy generally consists of supportive care, treatment of fever or seizures, and anticipation of syndrome of inappropriate antidiuretic hormone or increased ICP, rare complications of LE. Because LE can mimic HSV encephalitis, however, therapy with acyclovir is required until a specific diagnosis can be confirmed. Ribavirin can inhibit La Crosse virus *in vitro* and has been used in severe cases. The majority of children survive LE without neurological sequelae, although occasional children have persistent emotional lability or seizure tendencies.

HSV Encephalitis

HSV-1 encephalitis, the most common nonepidemic encephalitis worldwide, with an annual frequency of approximately one in 250 000 persons, exhibits no seasonal, gender, or racial predilections. Encephalitis can complicate primary, recurrent, or reactivated HSV-1 infections. Reactivation of HSV-1 and retrograde transmission of the virus to the CNS are assumed to cause most cases of human encephalitis after the neonatal period. In children and adults, HSV-1 exhibits tropism for the frontotemporal regions of the brain. By contrast, neonatal encephalitis, usually due to HSV-2, involves the brain diffusely as the result of hematogenous viral transmission.

Approximately 25% of HSV-infected neonates have isolated encephalitis. Encephalitis can also be a feature of disseminated neonatal HSV infections in which infants have pneumonitis, hepatitis, or disseminated intravascular coagulopathy. Nonspecific clinical manifestations, consisting of fever, lethargy, and poor feeding, usually appear during the second or third week of life but can be present as early as day 1. As neonatal HSV encephalitis progresses, lethargy, seizures (partial or generalized), paralysis, or respiratory failure predominate. Vesicular rash is observed in approximately two-thirds of HSV-infected infants.

The initial clinical features of HSV encephalitis in children and adults consist of a nonspecific prodrome of headache, fever, malaise, and vomiting. Focal neurologic signs appear subsequently in most patients and can include dysphasia, hemiparesis, visual field deficits, or partial seizures. The clinical course can be rapidly progressive, manifested by repetitive focal seizures, coma, increased ICP, and death within 2 weeks or by indolence, with hallucinations, headache, memory loss, or behavioral disturbances for several days before seizures appear. Children and adults rarely have systemic signs of HSV infection.

Infants with HSV encephalitis and disseminated infection can have thrombocytopenia, hemolytic anemia, elevated serum transaminases, and metabolic acidosis. By contrast, routine laboratory studies have no diagnostic specificity in children or adults with suspected HSV encephalitis. The CSF usually has a mononuclear pleocytosis (as high as1000 white blood cells or more per cubic millimeter), an increased protein content, and a normal or modestly depressed glucose content. Red blood cells or xanthochromia can be seen, but these findings have little diagnostic specificity.

Infants with suspected HSV infections should undergo virus cultures at multiple sites, including the conjunctiva, oropharynx, rectum, circulating leukocytes, CSF, and skin. Neonatal infection can be confirmed by isolating HSV-1 or -2 from any of these sites. Conjunctival cultures have the highest yield (>90%); the yield of oropharyngeal cultures is less (~40%). HSV can be detected in CSF samples in 40–50% of infants with CNS involvement. For maximum sensitivity, cultures should be obtained prior to acyclovir treatment (HSV CSF PCR can remain positive 24–48 h or longer after initiation of acyclovir therapy). By contrast, viral cultures rarely detect HSV in the blood or body fluids, including CSF, of children or adults with HSV encephalitis.

PCR can detect HSV DNA in the CSF of persons of any age with HSV encephalitis and has become the 'gold standard' for the diagnosis of HSV encephalitis. HSV DNA was detected in the CSF of 98% of the children and adults with biopsy-proven HSV encephalitis who participated in NIAID-CASG studies. The sensitivity of PCR is lower, however, in cases of

neonatal HSV encephalitis. HSV DNA was detected in only 26 of 34 (76%) infants with HSV-induced CNS disease in the NIAID-CASG studies.

Infants and young children with HSV encephalitis require treatment with acyclovir, 60 mg kg^{-1} daily in equally divided doses every 8 h. Older children and adolescents should receive 1500 mg kg^{-1} daily, and adults should be given at least 30 mg kg^{-1} daily in equally divided doses every 8 h. Dose reductions are necessary in persons with impaired renal function because of the drug's potential nephrotoxicity. Patients with confirmed HSV encephalitis require therapy for 21 days. Many authorities recommend repeat lumbar puncture and CSF HSV PCR at 21 days to confirm eradication of HSV. Longer courses of acyclovir may be necessary when the CSF HSV PCR result remains positive. A small number of patients with HSV encephalitis experience relapses and require retreatment with acyclovir. Corticosteroids can be considered when the relapse appears to be secondary to ADEM.

WNV

WNV, first identified in Uganda in 1937, caused occasional cases in East Africa and the Middle East until the mid-1990s, when neuroinvasive disease and fatal cases became more frequent. In 1999 WNV first appeared in the Western Hemisphere, causing seven deaths in adults during an encephalitis outbreak in the New York City metropolitan area. The WNV strain causing this outbreak was closely related to a strain identified in Israel in 1998.

In 2000 WNV activity was detected in 12 states surrounding New York State, and by the end of 2001, 27 states and the District of Columbia reported the presence of WNV. WNV spread rapidly among North American birds and mosquitoes, leading to a massive epidemic in 2003 (Figure 5) that affected more than 4000 inhabitants of 44 states from New York to California. By 2005 nearly 17 000 human cases of WNV infections had been reported in the United States, and approximately 40% were neuroinvasive. Several mosquito species serve as the natural vectors for human infection, but transmission among humans has been reported in association with breast feeding, blood transfusion, and organ transplantation.

Asymptomatic infections with WNV outnumber cases by approximately five to one. Neuroinvasive disease affects approximately one of every 150 infected persons. Although persons of all ages can be affected by WNV, adults comprise the majority of cases. To date, most persons with WNV neuroinvasive disease have been >25 years of age. After an incubation period of 2–14 days, systemic symptoms of fever, malaise, headache, nausea, and vomiting appear in symptomatic infections. Symptomatic patients can also have maculopapular rash or lymphadenopathy.

Encephalitis, affecting approximately 60% of persons with WNV neuroinvasive disease, begins gradually with a systemic prodrome or abruptly with high fever and coma. Manifestations of WNV encephalitis include altered mental status, coma, diffuse muscle weakness (proximal more than distal), respiratory paralysis, tremor, myoclonus, and parkinsonian features. Less common neurologic complications of WNV include aseptic meningitis, myelitis, optic neuritis, or the Guillain–Barré syndrome. Involvement of the anterior horn cells can produce a poliomyelitis-like disorder. A single case report suggests that WNV can also cause congenital CNS infection.

The CSF features in WNV encephalitis consist of a mild, lymphocytic pleocytosis (30–100 cells mm^{-3}), elevated protein content (80–100 mg dl^{-1}), and normal glucose content. Cranial CT is normal, whereas MRI can reveal leptomeningeal enhancement, periventricular inflammation, or abnormalities of the spinal cord. The diagnosis is confirmed by serologic studies that detect WNV-specific IgM in serum or CSF. WNV-specific IgM persists in serum for at least 16 months in more than 50% of infected persons. Infected patients experience brief periods of viremia, but infectious virus is rarely isolated from serum or CSF.

Therapy of WNV encephalitis consists of supportive care with close observation for respiratory failure. Ribavirin possesses activity against the WNV *in vitro*, but there are no controlled trials indicating efficacy in WNV encephalitis. There were 284 deaths among the 4156 cases reported to the CDC in 2003, corresponding to a mortality rate of approximately 7%. Death is more likely among older patients. A substantial proportion of the survivors of neuroinvasive disease have permanent cognitive, behavioral, or motor sequelae.

Rabies

Rabies poses a threat to persons in many regions, especially Africa and Asia, where the incidence of one case of rabies per 100 000–1 000 000 inhabitants corresponds to thousands of deaths annually. In these regions, domestic dogs or wild canines transmit rabies, and children constitute the majority of human cases. Human and animal rabies has been eradicated in a few regions as a result of quarantine and vaccination of domestic animals.

In the United States, rabies is endemic in raccoons, skunks, foxes, and bats. In the northeastern United States, the number of rabies virus-infected raccoons has risen steadily during the past 30 years. This epizootic constitutes a potential health hazard and

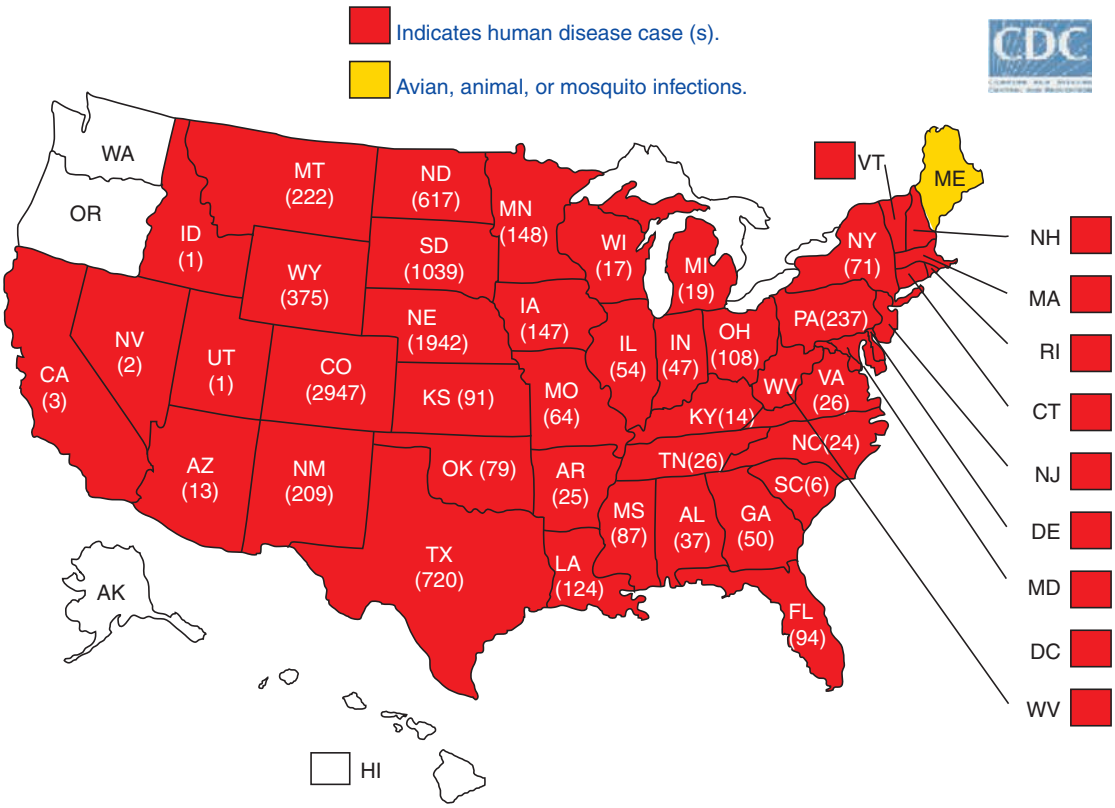

Figure 5 West Nile virus activity in the United States, 2003. From the Centers for Disease Control and Prevention, http://www.cdc.gov/ncidod/dvbid/westnile/Mapsactivity/surv&control03Maps.htm.

a substantial economic burden for postexposure prophylaxis among exposed humans. The CDC receives reports of one to five indigenous cases of human rabies annually in the United States. Most have no apparent history of animal bites; many are linked by molecular analysis to the rabies strain carried by the insectivorous silver-haired bat.

After an incubation period averaging 20–60 days, human rabies begins with chills, fever, headache, sore throat, malaise, nausea, or abdominal pain. Pain or itching at the site of inoculation is common. Neurologic symptoms and signs begin thereafter, being cerebral in 80% and Guillain–Barré syndrome-like in the remainder. Patients with cerebral symptoms have agitation, hypersalivation, delirium, and occasionally opisthotonic posturing. Hydrophobia, present in the range of 20–50% of patients, is associated with spasms of the laryngeal muscles, diaphragm, and accessory respiratory muscles. Seizures ensue in many cases. Eventually, patients with rabies encephalitis lapse into a fatal coma. Rabies-induced Guillain–Barré syndrome-like disease produces bulbar dysfunction, facial diplegia, and often respiratory arrest. Cardiac dysrhythmias and irreversible coma are additional causes of death in paralytic cases.

Because of its rarity in the United States and Canada, human rabies is often unsuspected, especially given the nonspecific clinical presentation. CSF may be normal or reveal nonspecific changes with a mixed or lymphocytic pleocytosis and mild protein elevation. CT and MRI are usually normal. Rabies is often diagnosed at postmortem examination. Premortem diagnosis can be established by immunofluorescence studies of skin from the nape of the neck, detection of rabies virus antibody in CSF or serum (assuming the patient was not vaccinated previously), or isolation of rabies virus from saliva or CSF. RT-PCR can be used to detect rabies virus and determine the molecular profile and animal origin of the rabies virus strain.

Because rabies cannot be treated effectively with the currently available antiviral agents, management of symptomatic human rabies consists of supportive care in an intensive care unit. The prognosis remains very grim, with few reported survivors, and most patients die within 12 days of disease onset. Because rabies virus can be present in saliva, urine, and other body fluids, patients with suspected rabies must be maintained in isolation.

Because of the strong epidemiologic relation between bat rabies strains and recent rabies cases in

humans, prophylaxis should be administered after contact with bats unless testing of the bat proves negative for rabies virus infection. Human-to-human transmission of rabies virus has not been reported except after corneal or organ transplantation. Nonetheless, persons with suspected rabies should be maintained in strict isolation to minimize contact with rabies virus-contaminated saliva and CSF and to reduce the number of healthcare workers who require postexposure prophylaxis.

Local public health officials or infectious disease experts should be consulted regarding recommendations for prophylaxis. Prevention of human rabies requires strict adherence to postexposure guidelines. The potential for rabies transmission depends on the animal species, the nature of the exposure, and the prevalence of rabies in the geographic region. All postexposure prophylaxis should begin with a thorough cleansing of the wound with soap and water. Current human diploid cell rabies vaccines are safe and effective.

See also: Glial Responses to Virus Infection; Hepatic Encephalopathy; Neuroimaging; Parasitic Diseases: Nervous System Effects; Perfusion MRI; Viral Vectors in the CNS.

Further Reading

Calisher CH (1994) Medically important arboviruses of the United States and Canada. *Clinical Microbiology Reviews* 7: 89–116.

Espy MJ, Uhl JR, Sloan LM, et al. (2006) Real-time PCR in clinical microbiology: Applications for routine laboratory testing. *Clinical Microbiology Reviews* 19: 165–256.

Hayes EB, Sejvar JJ, Sherif SR, Lanciotti RS, Bode AV, and Campbell GL (2005) Virology, pathology and clinical manifestations of West Nile virus disease. *Emerging Infectious Diseases* 11: 1174–1179.

Hemachudha T, Wacharapluesedee S, Laothamatas J, and Wilde H (2006) Rabies. *Current Neurology and Neuroscience Reports* 6: 460–468.

Johnson RT (1998) *Viral Infections of the Nervous System,* 2nd edn., New York: Lippincott-Raven.

Kimberlin D and Whitley R (2005) Neonatal herpes: What have we learned. *Seminars in Pediatric Infectious Diseases* 16: 7–16.

Koskiniemi M, Korppi M, Mustonen K, et al. (1997) Epidemiology of encephalitis in children: A prospective multicenter study. *European Journal of Pediatrics* 156: 541–545.

Leake JA, Albani S, Kao AS, et al. (2004) Acute disseminated encephalomyelitis in childhood: Epidemiologic, clinical and laboratory features. *Pediatric Infectious Disease Journal* 23: 756–764.

McJunkin JE, de los Reyes EC, Irazuzta JE, et al. (2001) La Crosse encephalitis in children. *New England Journal of Medicine* 344: 801–807.

Romero JR and Kimberlin DW (2003) Molecular diagnosis of viral infections of the central nervous system. *Clinics in Laboratory Medicine* 23: 843–865.

Tyler KL (2004) Update on herpes simplex encephalitis. *Reviews in Neurological Diseases* 1: 169–178.

Whitley RJ (1990) Medical progress: Viral encephalitis. *New England Journal of Medicine* 323: 242–250.

Whitley RJ, Soong SJ, Linneman C, Liu C, Pazin G, and Alford CA (1982) Herpes simplex encephalitis: Clinical assessment. *Journal of the American Medical Association* 247: 317–320.

Endocannabinoid Role in Synaptic Plasticity and Learning

B Lutz, Johannes Gutenberg University, Mainz, Germany
G Marsicano, Johannes Gutenberg University, Mainz, Germany; and U862 Centre de Recherche INSERM François Magendie Université Bordeaux 2, Bordeaux, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The records of cannabinoids date back more than 5000 years, to when the therapeutic and psychotropic actions of extracts of the hemp plant *Cannabis sativa* were first documented. The long history of this plant has been very vivid and has not ceased to elicit a lot of attention in our society. As with other natural products with strong pharmacological effects in humans, there has been great interest in understanding what the active components of extracts from *C. sativa* are and how they act on the human body, but success was hampered for a long time because of technical reasons. Only in 1964 was the group of Raphael Mechoulam able to chemically identify Δ^9-tetrahydrocannabinol (Δ^9-THC) as the major psychoactive constituent of *C. sativa* (Figure 1(a)). However, it took another 25 years to identify the mechanisms underlying these effects. Specific binding proteins for Δ^9-THC, called cannabinoid receptors, were discovered. Potent derivatives of Δ^9-THC such as CP55940 and HU-210 (Figures 1(c) and 1(d)) were instrumental in the pharmacological and molecular identification of these novel receptors, which belong to the G-protein-coupled receptor family. It is worth mentioning that in addition to Δ^9-THC, *C. sativa* extracts contain many other nonpsychoactive cannabinoids, such as cannabidiol (Figure 1(b)), which do not bind to cannabinoid receptors but which have interesting pharmacological effects with therapeutic potential. The molecular mechanisms underlying these effects have not yet been elucidated, but further investigation may lead to interesting insights, possibly even novel endogenous physiological mechanisms. Finally, the cloning of the cannabinoid receptor enabled the identification of the body's own molecules that can elicit effects similar to those of Δ^9-THC and that bind to and activate cannabinoid receptors. These endogenous ligands represent a novel class of lipid signaling molecules, called endocannabinoids (Figure 2).

Endocannabinoids are derivatives of the polyunsaturated eicosanoid arachidonic acid and can act as versatile neuromodulators at the synaptic level in both the central and peripheral nervous systems. Multiple roles in nonneuronal cells (such as the regulation of lipid metabolism, bone mass, and immune responses) have also been reported. Discovered in 1992 and 1995, respectively, the two major endocannabinoids are *N*-arachidonoyl ethanolamine (AEA), named anandamide, and 2-arachidonoyl glycerol (2-AG) (Figures 2(a) and 2(b)). Both of them bind to and activate cannabinoid receptors. Recently, other endocannabinoids (such as 2-arachidonoyl glycerol ether and *N*-arachidonoyl dopamine (Figures 2(c) and 2(d))) have been isolated. To date, two receptors have been characterized, both belonging to the class of seven-transmembrane receptors, coupled to the pertussis toxin-sensitive G-protein, $G_{i/o}$. Cannabinoid type 1 (CB1) receptors are predominantly expressed in neurons, but important sites of expression have also been found in peripheral nonneuronal cells, such as adipocytes and hepatocytes. Cannabinoid type 2 (CB2) receptors are present mainly in immune cells, but recent investigations were able to show functional receptors in central nervous system (CNS) neurons, too. This article focuses mainly on the CB1 receptor and its role in synaptic plasticity and memory processing as the present knowledge of CB2 receptors in the CNS is sparse. However, it would not come as a surprise if in the near future, the relevance of CB2 receptors in the CNS is better understood and receives more attention than is the case to date.

Endocannabinoids and cannabinoid receptors, together with the enzymes of endocannabinoid biosynthesis and degradation, constitute the endocannabinoid system. This system has recently emerged as an important regulator of a plethora of physiological processes in the CNS, including pain, appetite, reward, memory processing, and neuronal excitability. These functions appear to involve the fine-tuned regulation of synaptic transmission, which ultimately ensures homeostatic conditions of the nervous system. During pathological processes in the nervous system, dysregulations of the endocannabinoid system may occur. The endocannabinoid system in this state may be a therapeutic target for the treatment of particular nervous system diseases by agents that either stimulate or inhibit the activity of the endocannabinoid system, depending on the pathological conditions. During the past few years, this concept has drawn a lot of attention to the endocannabinoid system.

Cannabinoid Receptor CB1 and the Modulation of Intracellular Signaling

The prototypical signaling after activation of CB1 receptor by both plant-derived cannabinoids and

endocannabinoids is the inhibition of adenylate cyclase via $G_{i/o}$ protein, resulting in a decrease of intracellular cyclic adenosine monophosphate (cAMP) production (Figure 3), although coupling to G_s was also reported under some experimental conditions. CB1 receptor-dependent inhibition of the cAMP signaling pathway can stimulate A-type potassium channels, which are negatively regulated by phosphorylation provided by the cAMP-dependent protein kinase A (PKA). A direct coupling of the $G_{i/o}$ protein to inwardly rectifying potassium channels (GIRK1 and GIRK2) and the $G_{i/o}$-dependent inhibition of N-type and P/Q-type voltage-gated Ca^{2+} channels lead to an enhanced K^+ efflux and to an inhibition of Ca^{2+} influx, respectively. Thus, CB1 receptor activation leads to a general decrease of neuronal excitability. In addition, a multitude of other intracellular signaling systems are activated, of which many have been implicated in synaptic plasticity and memory processing. Finally, immediate early genes, including brain-derived neurotrophic factor and the transcription factors zif268 and c-fos, are switched on after CB1 receptor activation.

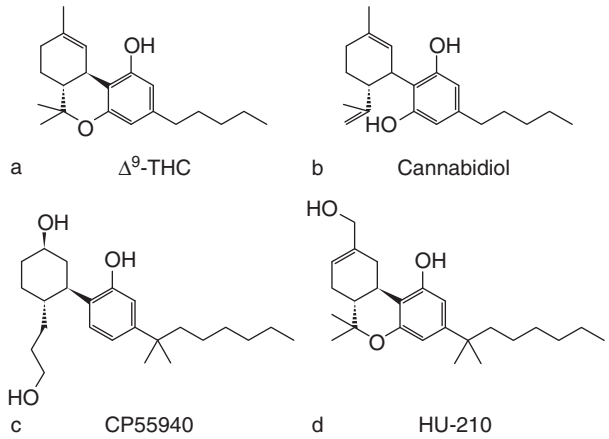

Figure 1 Chemical structures of plant-derived and synthetic cannabinoids. (a) Δ^9-Tetrahydrocannabinol (Δ^9-THC), identified in 1964 by Raphi Mechoulam, is the major psychoactive component in extracts from the hemp plant *C. sativa*. It binds to and activates the cannabinoid receptors. (b) Cannabidiol, a major constituent of *C. sativa*, is not psychoactive, as it does not bind to cannabinoid receptors. However, it is able to elicit distinct pharmacological effects by mechanisms that have not been elucidated in detail. (c) CP55940, a synthetic derivative of Δ^9-THC developed by Pfizer in 1974, is 40 times more potent than Δ^9-THC on cannabinoid receptors. The radiolabeled CP55940 was pivotal in the discovery and molecular characterization of the long-sought endogenous Δ^9-THC receptor, named cannabinoid receptor type 1 (CB1 receptor), as published in 1990. (d) HU-210, developed by the Hebrew University, is the most potent activator of cannabinoid receptors known to date (100 times more potent than Δ^9-THC).

Biochemistry of Endocannabinoids

Endocannabinoids are remarkable signaling molecules with multiple faces. As they are biosynthetic derivatives from arachidonic acid, they belong to the class of bioactive eicosanoids, such as prostaglandins, thromboxanes, and leukotrienes. These three well-known classes of lipophilic signaling molecules have been shown to act locally, in either an autocrine or a paracrine manner. For endocannabinoids, similar modes of actions also occur both in the CNS and in peripheral tissues.

Different from 'classical' neurotransmitters, endocannabinoids cannot be stored in vesicles. Therefore,

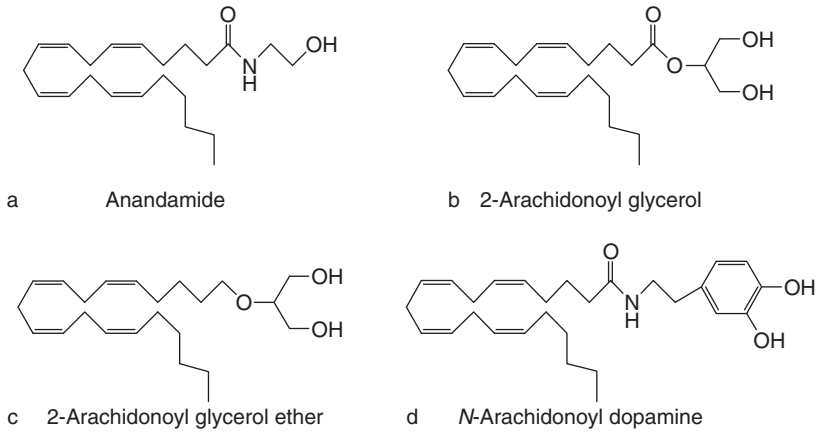

Figure 2 Structure of endocannabinoids. (a) Anandamide (*N*-arachidonoyl ethanolamine, AEA), the first endocannabinoid, discovered in 1992 and named from the Sanskrit word *ananda*, meaning 'internal bliss.' (b) 2-Arachidonoyl glycerol (2-AG). (c) 2-Arachidonoyl glycerol ether (noladin ether). (d) *N*-Arachidonoyl dopamine (NADA). All these four compounds bind to and activate cannabinoid receptors, but with different efficacies. AEA and 2-AG are the major endocannabinoids; the relevance of noladin ether in physiological processes has still to be elucidated. NADA represents a novel category of compounds as it is a condensation product of arachidonic acid with an amino acid derivative. Other lipids with similarity to AEA and 2-AG, such as *N*-arachidonoyl serine, dihomo γ-linolenoyl ethanolamide, docosatetraenyl ethanolamide, and oleamide, have been characterized, but these compounds do not bind or bind only with low affinity to the cannabinoid receptors CB1 and CB2.

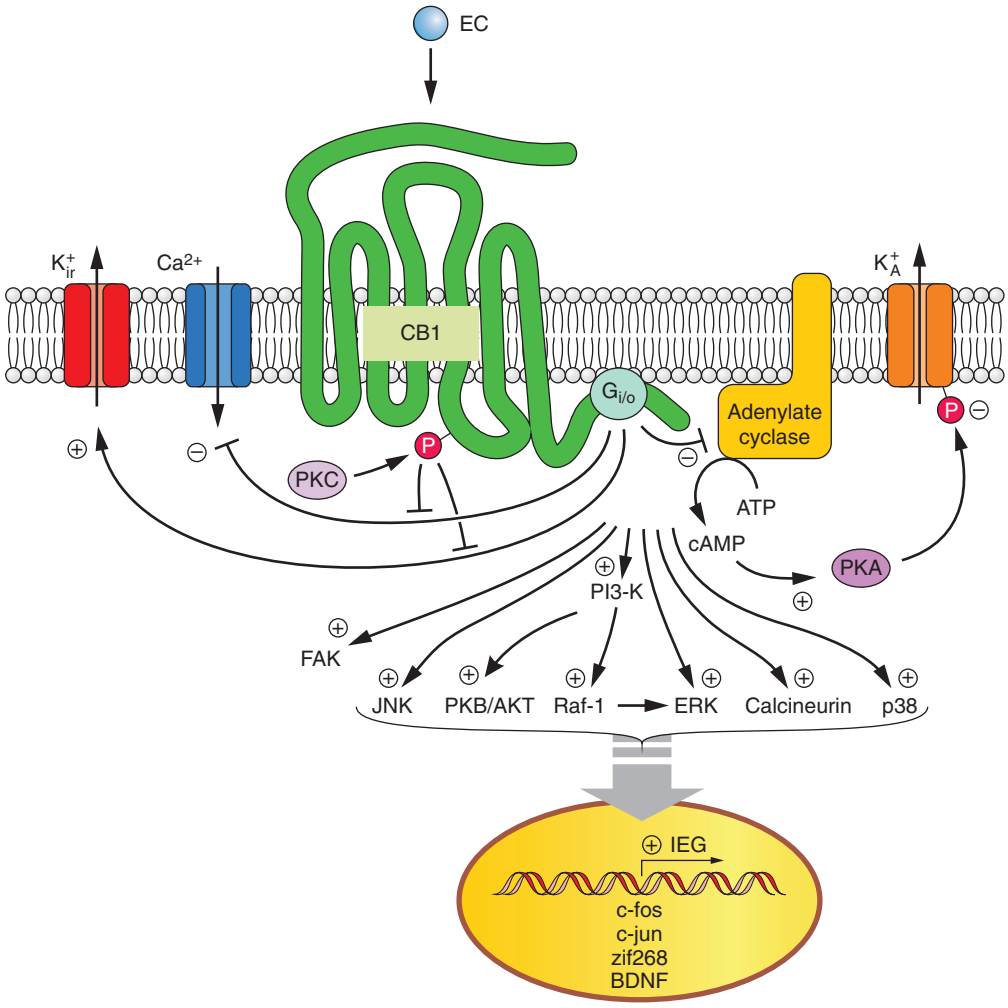

Figure 3 Schematic representation of the possible cannabinoid type 1 (CB1) receptor-mediated signaling pathways. Activation of CB1 receptors by both Δ^9-THC and endocannabinoids (EC) leads to the stimulation of $G_{i/o}$ proteins that, in turn, inhibit the adenylate cyclase-mediated conversion of adenosine triphosphate (ATP) to cyclic adenosine monophosphate (cAMP). Then cAMP molecules can bind the regulatory subunits of protein kinase A (PKA) and induce the liberation of the catalytic subunits. Activated PKA can phosphorylate A-type potassium (K^+_A) channels, causing a decrease of the K^+ efflux. Given the negative effect of CB1 receptor activation on adenylate cyclase, the final result is an activation of K^+_A channels. Activated $G_{i/o}$ can also directly inhibit voltage-gated N- and P/Q-type Ca^{2+} channels and activate inwardly rectifying potassium (K^+_{ir}) channels. The latter two effects are controlled by protein kinase C (PKC), which, after activation, can phosphorylate CB1 receptors in the third cytoplasmatic loop and uncouple the receptor from the ion channels. CB1 receptor activation can also stimulate several intracellular kinases, such as focal adhesion kinase (FAK), phosphatidyl inositol-3-kinase (PI3-K) and its downstream effectors protein kinase B (PKB) (also known as AKT) and Raf-1, extracellular signal-regulated kinases (ERKs), c-Jun N-terminal kinase (JNK), and p38 mitogen-activated kinase. CB1 receptor activation can also lead to increased expression levels of calcineurin, also called protein phosphatase 2B. Stimulation of cytoplasmic kinases can also mediate CB1 receptor-induced expression of immediate early genes (IEG), such as the transcription factors c-fos, c-jun, and zif268 and brain-derived neurotrophic factor (BDNF). It is important to note that several of these CB1 receptor-regulated signaling pathways have been shown to be involved in the modulation of synaptic plasticity and memory processing. P, phosphorylation.

biosynthesis (Figures 4 and 5) and degradation pathways of these compounds are the central regulatory steps in the activation and termination of endocannabinoid signaling. These pathways are currently being characterized, but some features are known: (1) Biosynthetic precursor molecules are phospholipids stored in the membrane. These phospholipids constitute a large pool and are not only specific for endocannabinoid synthesis but are also used for other metabolic processes in the cell. (2) The ondemand biosynthesis of endocannabinoids can be regulated by several intracellular pathways, including Ca^{2+}-phospholipase C (PLC)- and/or cAMP-dependent mechanisms. (3) Biosynthesis of the two major endocannabinoids, anandamide and 2-AG, can occur independently of each other in different pathways, and regulation of the on-demand synthesis of anandamide and 2-AG can also differ. (4) Biosynthetic pathways

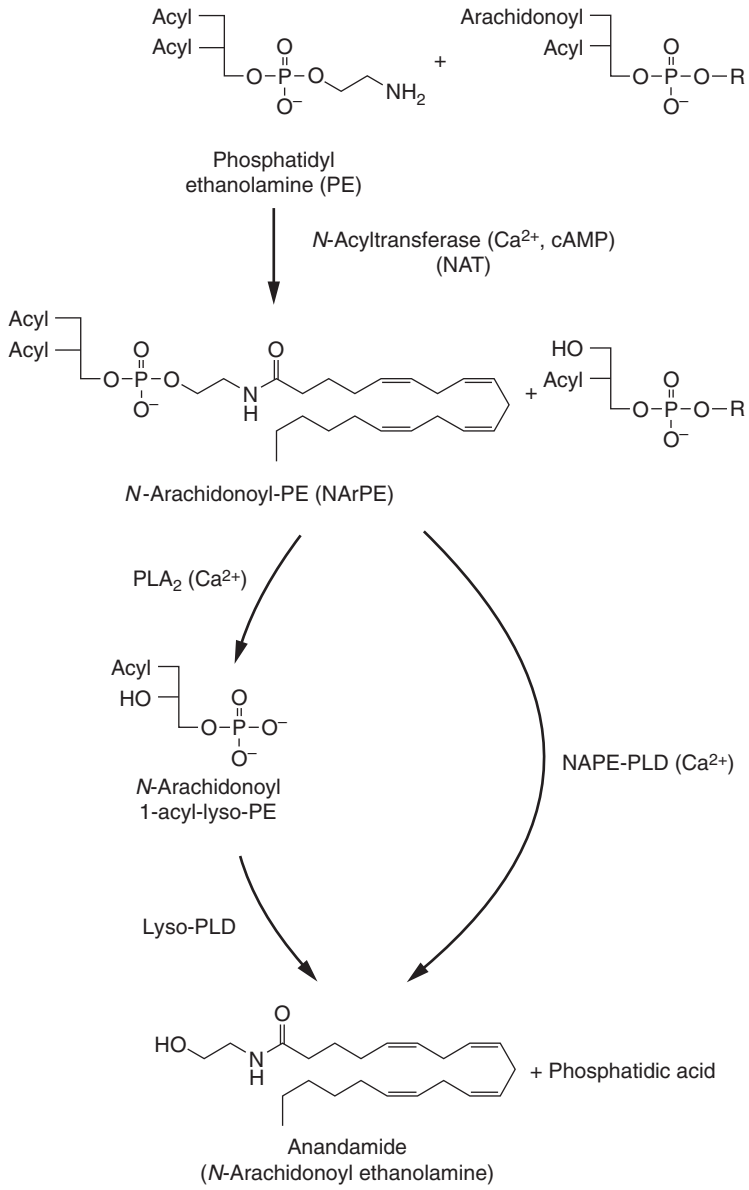

Figure 4 Schematic representation of the biosynthetic pathways leading to anandamide (*N*-arachidonoyl ethanolamine, AEA). As a first Ca^{2+}-dependent (and possibly also cyclic adenosine monophosphate (cAMP)-regulated) reaction catalyzed by *N*-acyltransferase (NAT), *N*-arachidonoyl-PE (NArPE) is generated from two membranous precursor molecules, phosphatidyl ethanolamine (PE) and an arachidonoyl phosphoglyceride. Then, anandamide can be synthesized by two different Ca^{2+}-dependent pathways, either by *N*-acyl-PE-phospholipase D (NAPE-PLD) or by phospholipase A_2 (PLA_2) and lyso-PLD. Note the redundancy of the pathways.

appear to be redundant, as gene inactivation experiments have not yet been able to pinpoint conclusively the crucial enzymes. Pharmacological experiments using inhibitors have provided good evidence for particular pathways, but such inhibitors may not be completely specific for one particular enzyme. (5) Degradation enzymes (FAAH and MAGL for anandamide and 2-AG, respectively), although clearly able to degrade the respective compounds, appear not to be completely specific. In particular, 2-AG and its precursors are at the crossroads of several metabolic pathways and are important precursors or/and degradation products of phospho-, di- and triglycerides. FAAH is also able to degrade many other acyl ethanolamides, which are not considered endocannabinoids. (6) After action, endocannabinoids can be taken up into the cells by facilitated transport mechanisms, although the endocannabinoid membrane transporter has not been cloned yet. (7) The differential physiological functions of anandamide and 2-AG are not completely understood, although many electrophysiological studies have identified 2-AG as the most likely to act as a retrograde neurotransmitter. (8) Anandamide is promiscuous and able to modulate the activity of

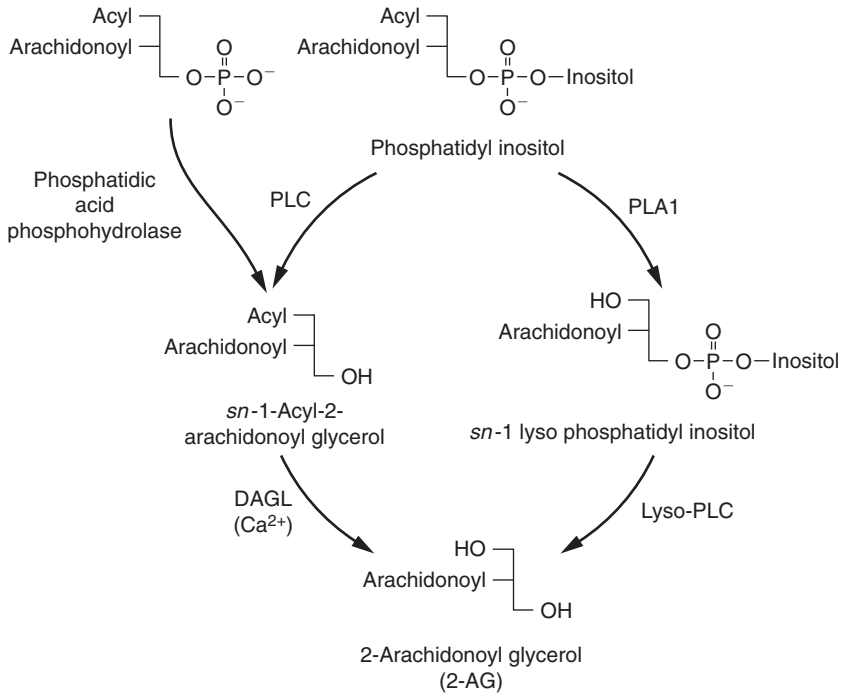

Figure 5 Schematic representation of the biosynthetic pathways leading to 2-arachidonoyl glycerol (2-AG). Membranous arachidonoyl phosphoglycerides serve as precursors to react with the enzymes phosphatidic acid phophohydrolase, phospholipase C (PLC, which can be regulated by G-protein-coupled receptors), and/or phospholipase A_1. 2-AG is then formed by the enzymes diacylglycerol lipase (DAGL) and lyso-PLC. The pathways are redundant, as are the pathways for anandamide synthesis. PLA_1, phospholipase A_1.

several other receptors, in particular vanilloid TRPV1 receptor (anandamide stimulates Ca^{2+} influx, cation current, and depolarization), TASK-1 (anandamide inhibits outward K^+ current), low-voltage-activated T-type Ca^{2+} channels (anandamide blocks Ca^{2+} current), 5-hydroxytryptamine type 3 serotonin receptor (anandamide acts as a negative allosteric modulator), and glycine receptor (anandamide and 2-AG act as inhibitors).

In conclusion, the current level of knowledge about endocannabinoids is not yet comparable with what is known about 'classical' neurotransmitters, and further intensive investigations are necessary to fully illuminate the on-demand biosynthesis of endocannabinoids. This is exemplified by the observation that genetic deletion of the anandamide-synthesizing enzyme NAPE-PLD does not lead to a deficit in either basal or stimulated anandamide levels. Thus, either NAPE-PLD is not the important enzyme, or redundancy and/or efficient compensatory pathways exist to synthesize anandamide.

During past years, several other bioactive lipid mediators have been described which have structural similarities with endocannabinoids and which act, at least in some cases, via cannabinoid receptors CB1 and/or CB2. These compounds include dihomo γ-linolenoyl ethanolamide, docosatetraenoyl ethanolamide, *N*-arachidonoyl serine, and oleamide. Specific pharmacological effects *in vivo* have been described for some of them, but for many of these compounds, endogenous functions remain to be deciphered, and possible corresponding high-affinity receptors have yet to be identified. Therefore, the term endocannabinoid may not be correct in a strict sense, but it is important to note that the prototypical structures of anandamide and 2-AG have initiated the discovery of novel endogenous compounds that may be components of yet-to-be-discovered signaling systems in the body. This is exemplified by recent observations that GPR55, a G-protein-coupled receptor, may be considered a novel 'cannabinoid' receptor. Recent data indicate that endogenous peptides could also modulate the activity of cannabinoid receptors, thus opening new perspectives in the field of the endocannabinoid system.

Distribution of the Endocannabinoid System in the Nervous System

In the adult nervous system of rodents and humans, CB1 receptors are abundantly and widely expressed in neurons of numerous brain areas, in the spinal cord, and in peripheral neurons such as dorsal root ganglia and enteric nervous system. The various expression sites in the brain correspond very well with

the different physiological functions of the endocannabinoid system, including, to mention only a few, in hippocampus, cerebral cortex, and amygdala for memory processing; in striatum for motor activity; in the nucleus accumbens, cerebral cortex, and ventral tegmental area for reward processing; and in hypothalamus for the regulation of the neuroendocrine axis, feeding behavior, and energy balance. CB1 receptors are also expressed in glial cells, such as astrocytes and microglial cells. The latter expression sites have not been implicated to date in the regulation of synaptic transmission, but they are likely involved in CB1 receptor-mediated neuroprotection and repair mechanisms. CB1 receptors are also highly expressed in the developing nervous system of the embryo, where they are particularly highly present in proliferating neuronal precursor cells. Matching this embryonic expression, CB1 receptors are also described in neuronal precursors in the dentate gyrus and subventricular zone of the adult brain, promote the proliferation of these neuronal precursor cells, and endorse the switch of these newly born cells toward a glial fate, as shown in the dentate gyrus.

In neurons of the adult brain, functional CB1 receptors are shown to be localized predominantly at presynaptic sites and, as recently shown for particular cortical γ-aminobutyricacidergic (GABAergic) neurons, on the soma. Some CB1 receptor protein can also be detected intracellularly in the cytoplasma of the soma and in axonal projection, but the functional significance of these expression sites is still a matter of investigation, and the possibility exists that these CB1 receptors are merely proteins that have been newly synthesized and transported to their final sites of action.

CB1 receptors present at the presynapse mediate the endocannabinoid-mediated retrograde suppression of neurotransmitter release, while their expression at the soma is likely responsible for an autocrine endocannabinoid-mediated self-inhibition of neuronal activity. These features will be discussed later as they represent remarkable mechanisms for the short- and long-term modulation of excitatory and inhibitory transmission in the nervous system and in synaptic plasticity, with profound implications for the quest to understand how phenomena such as memory processing function.

The complexity of CB1 receptor expression is further enhanced by the fact that CB1 receptors and other components of the endocannabinoid system are found in diverse neuronal subpopulations. CB1 receptor messenger RNA (mRNA) and protein are highly abundant in GABAergic interneurons (in particular, in cholecystokinin-containing GABAergic interneurons). The expression levels in GABAergic neurons may be counted at the highest levels known for any receptor in the brain. This feature contrasts very much with the expression in cortical glutamatergic neurons. Here, the expression levels of both mRNA and protein may be estimated to be as much as 50–100 times lower than in GABAergic neurons (**Figure 6**). It is, however, remarkable that CB1 receptor expression at these low levels in glutamatergic neurons is necessary and sufficient to mediate several functions, such as protection from abnormal neuronal activities and some of the central effects of Δ^9-THC.

All other components of the endocannabinoid system also show a differential distribution. For the two major endocannabinoids, anandamide and 2-AG, levels differ depending on brain area and the diurnal cycle, but they can also grossly change in response to physiological stimuli or during pathological conditions of the brain. To terminate the signaling, the enzymes for the degradation of endocannabinoids, fatty acid amide hydrolase (FAAH, degrading anandamide) and monoacylglycerol lipase (MAGL, degrading 2-AG), are also present in the brain. It is interesting that both enzymes exhibit differential expression patterns in cortical areas, in particular in the hippocampal formation. FAAH is present in pyramidal neurons of CA1 and CA3 and in granule cells of the dentate gyrus, but not in GABAergic interneurons. MAGL is expressed similarly to CB1 receptor, present both in GABAergic interneurons and pyramidal neurons. Remarkably, FAAH is expressed specifically at the postsynapse, while MAGL is present in the presynaptic terminal. The differential distribution of FAAH and MAGL might underlie the different functions of the two major endocannabinoids, anandamide and 2-AG, in neurons. The regional distribution of the putative enzymes for the biosynthesis of endocannabinoids has not been studied in detail yet.

Cannabinoids and the Suppression of Neurotransmitter Release

Exogenously applied cannabinoids were shown to suppress GABAergic transmission in several brain regions, including hippocampus, amygdala, striatum, nucleus accumbens, and substantia nigra. These effects were inhibited by specific CB1 receptor antagonists (e.g., SR141716, also called rimonabant) and were absent in slices derived from CB1 receptor-deficient mice. Similar results were obtained for glutamate transmission. Furthermore, cannabinoids were also shown to inhibit the evoked release of other neurotransmitters, such as noradrenalin and acetylcholine in the hippocampus, dopamine in the striatum, and serotonin in the cerebral cortex. An understanding of the underlying mechanisms has been advanced by insights into the biochemistry and neuroanatomy of

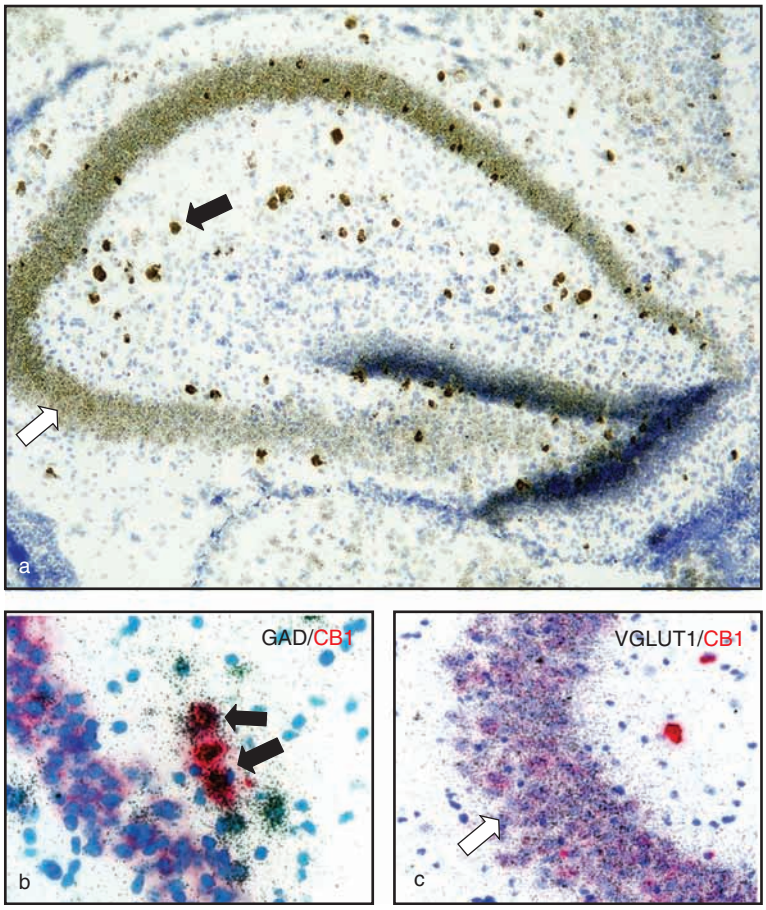

Figure 6 Distribution of messenger RNA (mRNA) encoding CB1 receptor visualized by *in situ* hybridization. (a) View of the hippocampal formation of an adult mouse brain. Brown-colored precipitation, CB1 receptor mRNA; blue staining, cell nuclei. Note the very high expression in scattered neurons (GABAergic interneurons, black arrow) and the low but distinct expression in pyramidal cells (glutamatergic neurons, white arrow). (b) Co-expression (black arrow) of CB1 receptor mRNA (red staining) with the GABAergic marker γ-aminobutyric acid decarboxylase (GAD; silver grains). (c) Co-expression (white arrow) of CB1 receptor mRNA (red staining) with the glutamatergic marker vesicular glutamate transporter type 1 (VGLUT1; silver grains).

the endocannabinoid system. The principal mechanisms by which the endocannabinoid system can modulate synaptic transmission are depicted in Figure 7. Several specific modulatory mechanisms of synaptic transmission and the involvement of the endocannabinoid system in short- and long-term synaptic plasticity are discussed briefly in the following sections. This field of study is expanding rapidly.

Endocannabinoids and Short-Term Synaptic Plasticity

When a neuron is activated postsynaptically, it produces endocannabinoids, which travel in a retrograde direction to the presynaptic site and transiently (<1 min) suppress presynaptic neurotransmitter release. The phenomenon of short-term depression (STD) was first observed about 15 years ago, when a brief depolarization of hippocampal principal neurons was shown to trigger a transient suppression of GABAergic synaptic input. This was named depolarization-induced suppression of inhibition (DSI). Critical features of DSI were defined as follows: (1) DSI induction requires an increase in postsynaptic Ca^{2+}. (2) DSI has a presynaptic location of expression. (3) Therefore, a retrograde messenger must exist mediating the inhibition of neurotransmitter release. The molecular nature of such a retrograde messenger has remained elusive for a long time. However, in 2001, endocannabinoids, which are stimulated by Ca^{2+} and act at presynaptic CB1 receptors, were discovered to be the postulated retrograde transmitter (Figure 8). Several areas besides the hippocampus were found to express DSI, including substantia nigra, dentate gyrus, and cerebellum. In addition, this phenomenon has been observed for glutamatergic transmission, is called depolarization-induced suppression of excitation (DSE), and is found in areas, including hippocampus, cerebral cortex, hypothalamus, ventral tegmental area, and cerebellum.

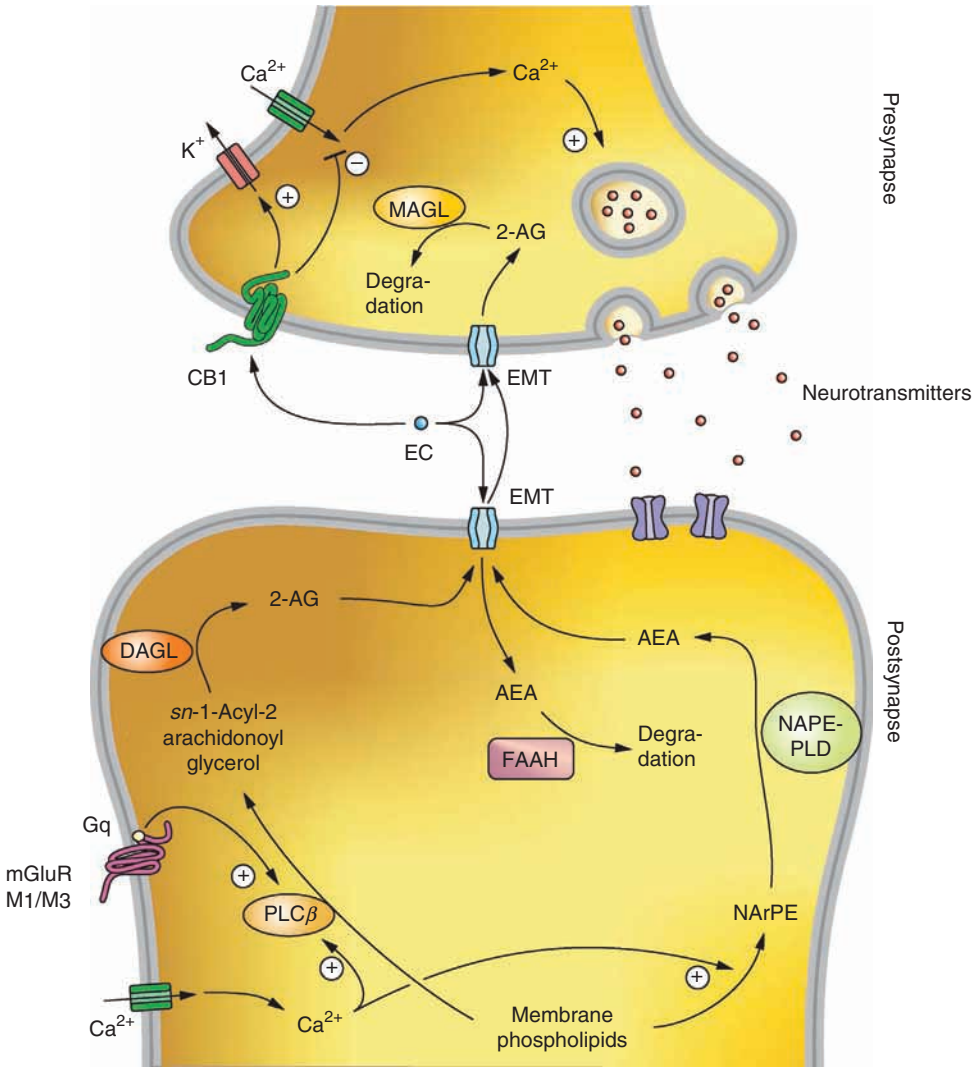

Figure 7 Representation of the synthesis, signaling, and degradation of endocannabinoids at the synaptic site. Key steps and mechanisms are depicted. In the postsynapse, Ca^{2+} and G-protein-coupled receptor-induced activation of G_q protein can initiate the synthesis of endocannabinoid (EC), which reaches the synaptic cleft by facilitated diffusion through the EC membrane transporter (EMT). Cannabinoid type 1 receptor (CB1) is localized at the presynaptic site, where it is activated by EC and then reduces Ca^{2+}-induced release of neurotransmitter. AEA, anandamide; 2-AG, 2-arachidonoyl glycerol; DAGL, diacylglycerol lipase; FAAH, fatty acid amide hydrolase; MAGL, monoacylglycerol lipase; mGluR, metabotropic glutamate receptor; M1/M3, muscarinic acetylcholine receptor types M1 and M3; NAPE-PLD, *N*-acyl phosphatidyl ethanolamine phospholipase D; NArPE, *N*-arachidonoyl phosphatidyl ethanolamine; PLCβ, phospholipase Cβ.

Clearly, the number of different brain areas with proven CB1 receptor-dependent DSI/DSE expression will grow even further in the near future.

Synthesis and release of endocannabinoids is a complex process. For the initiation of endocannabinoid production, two pathways have been described in the hippocampus. (1) Postsynaptic increase of Ca^{2+} is necessary and sufficient to trigger endocannabinoid production. (2) Activation of postsynaptic metabotropic glutamtate receptors (mGluRs) or muscarinic acetylcholine receptors with exogenous agonists is also sufficient to trigger endocannabinoid release. It is interesting that metabotropic induction of DSI does not require a large increase in Ca^{2+}. It is suggested that the metabotropic pathway triggers endocannabinoid production by PLCβ and that PLC activity is facilitated by increases in Ca^{2+}. In this model, PLC acts as a coincidence detector, integrating two different postsynaptic processes: metabotropic receptor activation and Ca^{2+} increase. Two contributions for the intracellular Ca^{2+} increase are described: Influx through voltage-gated Ca^{2+} channels, and release from intracellular Ca^{2+} stores via ryanodine-sensitive receptors.

After their synthesis, endocannabinoids travel to the presynapse. This process is not understood at all. It is not clear how highly lipophilic molecules such as

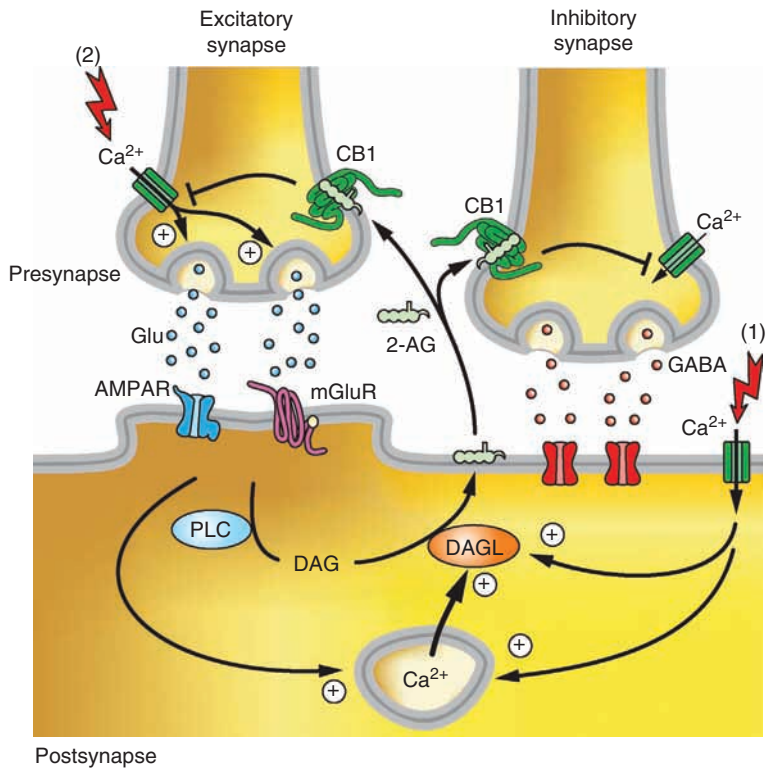

Figure 8 Schematic summary of endocannabinoid-mediated short-term plasticity, called depolarization-induced suppression of excitation (DSE; depicted on the left side) and depolarization-induced suppression of inhibition (DSI; depicted on the right side). Two events can induce endocannabinoid production. (1) Postsynaptic step depolarization (experimentally by a shift from −60 mV to 0 mV) or action potential at the postsynapse induces Ca^{2+} influx via voltage-gated calcium channels. Ca^{2+} influx may be amplified by enhancement of Ca^{2+} release from intracellular stores via ryanodine receptors. (2) Brief tetanic stimulation of excitatory afferents or group I metabotropic glutamate receptor (mGluR) activation by agonists can initiate endocannabinoid synthesis. Alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor (AMPAR) activation may also contribute to this induction process. Phospholipase C (PLC) and diacylglycerol lipase (DAGL) are required for mGluR-mediated processes, indicating 2-arachidonoyl glycerol (2-AG) as an important candidate for this pathway; 2-AG will activate presynaptic cannabinoid type 1 (CB1) receptors, leading to a short depression of neurotransmitter release (<1 min). Depending on the synaptic input that is measured, either DSI or DSE will be observed. DAG, diacyl glycerol; GABA, γ-aminobutyric acid.

endocannabinoids are able to diffuse so rapidly across the synaptic cleft. Binding proteins may help to increase the solubility in the hydrophilic environment of the synaptic cleft.

In the cholecystokinin-expressing GABAergic interneurons of the hippocampus, the CB1 receptor-dependent effects at the presynaptic site are mediated mainly by an inhibition of N-type Ca^{2+} channels by a direct interaction with the CB1 receptor-induced liberation of $G_{i/o}$ $\beta\gamma$ subunits, but a direct effect on the vesicle release machinery , and/or K^+ channel activation also have to be considered.

In synapses of hippocampal GABAergic interneurons onto principal neurons, the time-to-onset of inhibitory postsynaptic current in the voltage step protocol (−60 mV to 0 mV in a particular frequency) to induce DSI was determined to be about 410 ms. In experiments using caged anandamide that can be photoactivated to exert its action, the times for three steps were determined: Time for the Ca^{2+} induction was 60 ms, time for endocannabinoid synthesis, 190 ms, and time to activate CB1 receptors, 160 ms. When mGluRs were activated to induce DSI, time for endocannabinoid synthesis was reduced to 75 ms. These observations suggest that different pathways can lead to the on-demand synthesis of endocannabinoids and that the time to activate CB1 receptors is an important rate-limiting step.

The detailed mechanisms for post- and presynaptic events can vary in different brain regions, as shown for the cerebellum in particular. There are also indications that, depending on the stimulus, variations between one synapse and another within a distinct brain region exist. These features thus provide a high level of complexity to endocannabinoid-mediated modulation of synaptic plasticity.

Endocannabinoids and Long-Term Plasticity

CB1 receptor-mediated processes also have a long-term impact on synaptic transmission. Long-term depression (LTD) of synaptic transmission can occur via a presynaptic CB1-dependent mechanism. Once established, LTD does not require continuous CB1 activation. The postsynaptic trigger is similar to the one inducing short-term depression, usually Ca^{2+} increase and/or activation of mGluRs. As compared with the 'common' LTD forms on excitatory synapses, endocannabinoid-mediated LTD is independent of postsynaptic glutamate receptors of the *N*-methyl-D-aspartate (NMDA) type. The wide occurrence of endocannabinoid-mediated LTD in the brain indicates that the endocannabinoid system is able to make long-term modifications to neural circuits and behaviors such as learning and memory.

Homosynaptic LTD on Excitatory Synaptic Input

Homosynaptic LTD on excitatory synaptic input was first discovered in corticostriatal and corticoaccumbal projections. For example, at glutamatergic cortical inputs of dorsal striatal GABAergic medium spiny neurons, high-frequency stimulation of this synapse (100-Hz trains) induces a form of LTD shown to depend on the activation of CB1 receptors. This process follows a scheme similar to the one described for DSI (Figure 8) and presents the following hallmarks: (1) requirement of postsynaptic intracellular Ca^{2+} increase; (2) postsynaptic loading with anandamide-suppressed synaptic transmission via a presynaptic mechanism, which is CB1 receptor dependent and which suggests that anandamide can travel from the post- to the presynaptic site; and (3) dependence on a convergence of several input signals, including group I mGluRs, D2 dopamine receptors, and L-type Ca^{2+} channels.

The dorsal striatum and nucleus accumbens are relevant for a variety of goal-oriented behaviors, dorsal striatum for motor planning and reward-driven motor learning and nucleus accumbens in processing of rewarding environmental stimuli and in drug addiction. Therefore, the involvement of endocannabinoids on this LTD is very relevant to understanding several motor- and reward-related behaviors.

Heterosynaptic LTD on Inhibitory Synaptic Input

This form of CB1-dependent synaptic plasticity was first discovered in amygdala (called LTD_i) and hippocampus (called I-LTD), two brain areas critically involved in emotional and declarative memory. They are heterosynaptic forms of plasticity, induced by repetitive activity of excitatory synaptic inputs onto glutamatergic neurons, resulting in LTD of nearby inhibitory synapses of CB1 receptor- and cholecystokinin-expressing GABAergic interneurons and thus generating a long-term decrease in GABA release (Figure 9). As GABAergic inputs regulate the balance between excitatory and inhibitory drive in the neural network, this durable disinhibition may have long-term effects on excitability and on the synaptic plasticity of excitatory synapses, finally facilitating the formation of new memories by use-dependent processes. However, similar to the continuing discussion of whether and in which manner long-term potentiation (LTP) on excitatory synapses is an electrophysiological correlate of memory processing, it is likely to be very difficult to establish a functional link between heterosynaptic endocannabinoid-dependent LTD and memory processing and behavior.

In contrast to DSI and homosynaptic LTD, heterosynaptic LTD does not require postsynaptic Ca^{2+} increase, but activation of group I mGluRs is both necessary and sufficient for the induction of LTD_i/I-LTD. However, there are distinct differences between LTD_i and I-LTD. In the hippocampus, I-LTD is induced by moderate- to high-frequency stimulation (10–100 Hz) and by theta-burst stimulation. PLC and diacylglycerol lipase are involved in generating 2-AG, as postsynaptic inhibition of these enzymes inhibited I-LTD. In the amygdala, LTD_i is induced by low-frequency stimulation (1 Hz, 100 stimuli), and inhibitors of adenylate cyclase and PKA, but not of PLC and diacylglycerol lipase, did affect the induction of LTD_i. It is suggested that a particular isoform of mGluRs may be linked to the adenylate cyclase-PKA pathway, which then activates *N*-acyltransferase, the first step in the synthesis of anandamide. Genetic inactivation of the anandamide-degrading enzyme FAAH facilitates the formation of LTD_i, further suggesting a role of anandamide but not 2-AG in this process. However, more experiments will be necessary to fully reveal this cAMP-regulated pathway of endocannabinoid production.

Time-Dependent LTD

This form of plasticity of glutamatergic transmission has been observed in pyramidal neurons in layer V of the visual cortex. Here, LTD on glutamatergic synapses is triggered by pairing of presynaptic stimulations with postsynaptic depolarizations or action potentials. The time schedule of the pairing of the stimuli is important for LTD formation; thus, it is also called timing-dependent LTD. Increases of postsynaptic Ca^{2+} lead to the production of endocannabinoids, which then inhibit glutamate release from the presynapse. Presynaptic, but not postsynaptic,

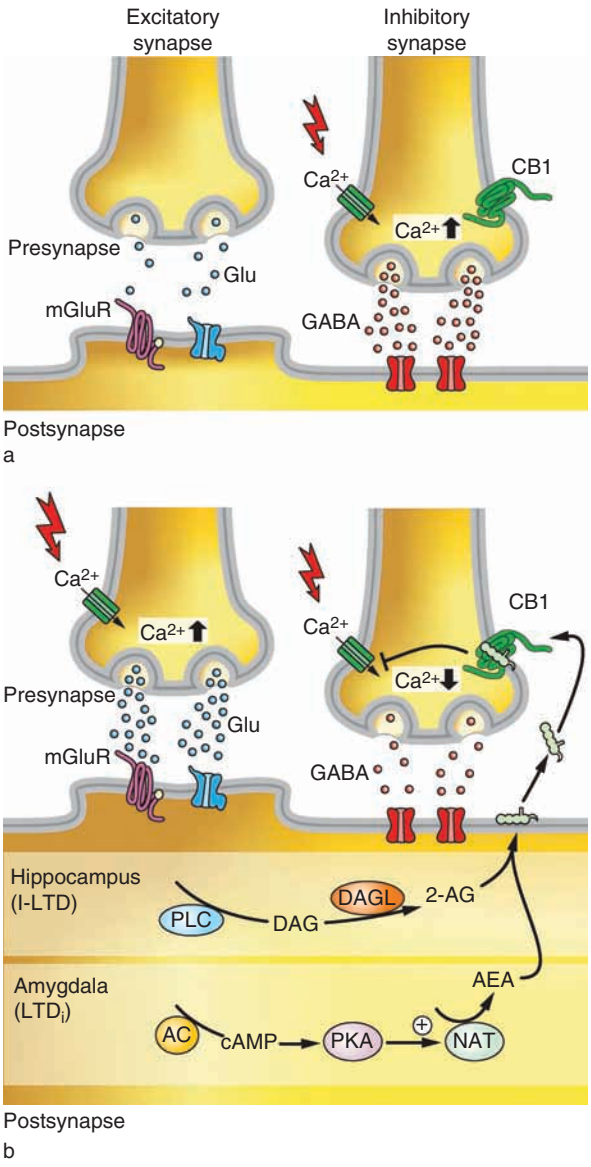

Figure 9 Heterosynaptic endocannabinoid-mediated long-term depression (LTD) of inhibitory current. (a) Gamma-aminobutyricacidergic (GABAergic) input onto a principal glutamatergic neuron is activated and γ-aminobutyric acid (GABA) is released. (b) In heterosynaptic LTD of GABAergic transmission, a coincident stimulation of an excitatory input onto the same principal glutamatergic neuron provokes the synthesis of endocannabinoids at the postsynaptic site. Cannabinoid type 1 (CB1) receptor-expressing GABAergic terminals receive the endocannabinoid signaling and reduce GABA transmission. This effect persists as long as 30 min. Metabotropic glutamate receptor (mGluR) activation is necessary and sufficient to induce LTD. The induction of endocannabinoid synthesis may differ in hippocampus and amygdala. First, while medium- to high-frequency stimulation is required for hippocampal I-LTD, only low-frequency stimulation induces LTD_i in amygdala. Second, hippocampal I-LTD requires phospholipase C (PLC) and diacylglycerol lipase (DAGL) activity and involves 2-arachidonoyl glycerol (2-AG), while LTD_i involves cyclic adenosine monophosphate (cAMP) pathway to generate presumably anandamide (AEA). AC, adenylate cyclase; DAG, diacylglycerol; Glu, glutamate; NAT, *N*-acyltransferase; PKA, protein kinase A.

glutamate receptors of the NMDA type are required for this process, representing a coincident signal paired with the postsynaptic endocannabinoid signal reaching the presynapse via the CB1 receptor. This mechanism is very interesting with respect to memory processing as the timing of neuronal activities induced by distinct behaviors may certainly be highly relevant and may ultimately influence synaptic strength and memory formation.

Long-Lasting Self-Inhibition

While in DSI/DSE and endocannabinoid-dependent LTD, retrograde signaling from the post- to the presynapse occurs, an autocrine mechanism has recently been described in low-threshold-spiking (LTS) neocortical GABAergic interneurons (Figure 10). This function occurs after sustained firing of the interneuron. Intracellular Ca^{2+} rises after opening of voltage-gated Ca^{2+} channels in the somatodendritic compartment, and 2-AG is synthesized from membranous precursors. Then 2-AG most likely travels laterally within the membrane and activates CB1 receptors, leading to a $G_{i/o}$-dependent inhibition of K^+ channels and a subsequent hyperpolarization of LTS GABAergic interneurons, which lasts at least 30 min. As GABAergic interneurons are centrally involved in oscillatory processes in neural networks, this endocannabinoid-mediated modulation of synaptic transmission may also be relevant for memory processing.

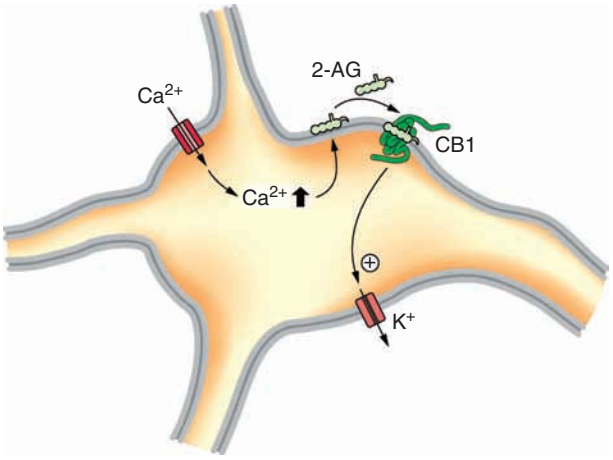

Figure 10 Mechanism of endocannabinoid-mediated slow self-inhibition. Depolarization of distinct neocortical γ-aminobutyricacidergic interneurons (so-called low-threshold spiking interneurons, which also express cholecystokinin) opens somatodendritic Ca^{2+} channels and induces the synthesis of 2-arachidonoyl glycerol (2-AG), which binds to cannabinoid type 1 (CB1) receptors present on the same neurons. CB1 receptor activation inhibits inwardly rectifying K^+ channels, finally leading to a hyperpolarization of the neuron.

Endocannabinoids and Memory Processing

The quest to understand the roles of the endocannabinoid system is impelled by the well-known observation that exogenous cannabinoids (e.g., from smoking marijuana) impair certain forms of memory. However, application of exogenous cannabinoids to animal models and subsequent behavioral analysis have not been particularly successful in elucidating this process of impairment. An important reason is that the endocannabinoid system is activated on demand, whereas exogenous application of agonists lacks temporal and spatial selectivity. As discussed earlier, the endocannabinoid system is involved in many different forms of modulation of synaptic transmission. These modulations can be either of short duration or long-lasting and represent important mechanisms of synaptic plasticity. Short-term and long-term suppression of both glutamate and GABA transmission is a major effect, but the endocannabinoid system is also able to influence LTP at excitatory synapses, an electrophysiological phenomenon hotly discussed in memory research. In hippocampal slices, an induction of DSI with subsequent subthreshold high-frequency stimulation (which would normally not induce LTP) enables the formation of LTP. Thus, DSI transiently reduces the inhibitory drive in the system and makes synapses more excitable and can thereby strengthen synaptic connections. Thus, the endocannabinoid system provides multiple possibilities for influencing memory processing.

To produce results that enable one to draw solid conclusions about the role of the endocannabinoid system in memory processing, the most powerful approach has been the application of specific CB1 receptor antagonists to animals, in particular to rodents, and the use of mice deficient in CB1 receptor. In several behavioral paradigms, distinct phenotypes have been observed.

Fear Conditioning

In this behavioral test, the mouse hears a loud but *per se* harmless tone. Then, the tone co-terminates with a mild electric foot shock. The animal will form an association of the tone and the context (the conditioning chamber) with the electric shock. In subsequent exposures to the tone or the context, the mouse will show a fear reaction (e.g., freezing, an innate defense response in rodents) as a harmful shock is expected. If the shock does not occur anymore, the mouse will gradually start to decrease the freezing reaction, a phenomenon called extinction. It has been shown that the endocannabinoid system is required for the appropriate extinction of both tone freezing and contextual freezing. It is noteworthy that acquisition and consolidation are the same in wild-type and CB1 receptor-deficient mice. As fear conditioning depends on the amygdala, electrophysiological properties have been characterized. CB1 receptor-deficient mice show increased LTP, normal LTD, and abolished LTD_i. How these changes can be understood in relation to impaired extinction remains to be elucidated.

Morris Water Maze

This assay investigates learning abilities in a spatial memory task. Mice have to learn the location of an invisible escape platform in a small water tank by using contextual cues to orientate themselves in the space. Efficiency and accuracy in learning this location do not depend on the endocannabinoid system, but in the reversal test, when the position of the platform is transferred into the opposite quadrant of the tank, CB1 receptor-deficient mice have shown impaired ability to find the new position and keep on searching in the original position much longer than control mice do. This behavior can be interpreted as impaired extinction of the previously acquired memory trace and impaired flexibility toward a new environment. Thus, the phenotype is similar to the one observed for fear conditioning. A possible interpretation is that the Morris water maze represents a strong aversive event for mice, and thus systems similar to those for fear conditioning may be required in a CB1 receptor-dependent manner.

Object Recognition and Social Recognition

In this assay, rodents have to learn whether a new object or another rodent can be classified as familiar or unfamiliar. Normal animals spend more time with unfamiliar than with familiar objects or animals. CB1 receptor antagonist-treated rodents and CB1 receptor-deficient mice show an enhanced ability to prefer novel objects and animals, at least when mice are young and middle-aged. This means that the endocannabinoid system acts to inhibit the formation and stability of memory traces in this period of age. In old animals, however, CB1 receptor-deficient mice are inferior to wild-type mice and show neurodegeneration in the hippocampus. This may be due to the observation that the endocannabinoid system has important functions in the regulation of stress responses, mediated by the hypothalamic–pituitary–adrenal axis. Lifelong absence of CB1 receptors may

cause sustained increases of corticosterone, leading finally to neurodegenerative effects in the hippocampus. Regarding electrophysiological properties, it is interesting that LTP is enhanced in CB1 receptor-deficient mice compared with wild-type controls, as analyzed in young and middle-aged animals.

Appetitively Motivated Learning Task

In this task, in order to obtain a food reward, mice have to learn to poke a hole at the location where a light appears. After learning, mice are tested after 1 week to evaluate memory consolidation. After this, extinction is induced by omitting the food reward when they poke the correct hole. Neither the learning phase nor consolidation nor extinction depends on the endocannabinoid system in this task. Thus, if there is no aversive component, the endocannabinoid does not seem to be involved in extinction processes.

Conclusion

During the past few years, the endocannabinoid system has emerged as an important modulator of synaptic plasticity. Both short-term and long-term effects on synaptic strength and efficacy have been reported at GABAergic and glutamatergic synapses. Numerous brain circuits are modulated by endocannabinoids. The effects are mediated either by presynaptically expressed CB1 receptors and retrograde endocannabinoid transmission or by autocrine mechanisms.

The endocannabinoid system is involved in several memory systems, particularly in the extinction of emotional memories and the acquisition of object and social recognition memory. However, further investigation is needed to bridge the gap between the functions of this neuromodulatory system in electrophysiological processes *in vitro* and the behavior of the intact organism.

See also: Calcium-Calmodulin Kinase II (CaMKII) in Learning and Memory; Cyclic AMP (cAMP) Role in Learning and Memory; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Energy Homeostasis: Endocannabinoid System; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Fear Conditioning and Synaptic Plasticity; Long-Term Depression: Cerebellum; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Depression (LTD): Endocannabinoids and Cerebellar LTD; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Presynaptic Facilitation; Presynaptic Inhibition; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Depression; Synaptic Plasticity: Short-Term Mechanisms; Synaptic Plasticity: Cerebellum; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Alger BE (2002) Retrograde signaling in the regulation of synaptic transmission: Focus on endocannabinoids. *Progress in Neurobiology* 68: 247–286.

Bilkei-Gorzo A, Racz I, Valverde O, et al. (2005) Early age-related cognitive impairment in mice lacking cannabinoid CB1 receptors. *Proceedings of the National Academy of Sciences of the United States of America* 102: 15670–15675.

Chevaleyre V, Takahashi KA, and Castillo PE (2006) Endocannabinoid-mediated synaptic plasticity in the CNS. *Annual Review of Neuroscience* 29: 37–76.

Di Marzo V, Bifulco M, and De Petrocellis L (2004) The endocannabinoid system and its therapeutic exploitation. *Nature Reviews. Drug Discovery* 3: 771–784.

Gerdeman GL, Ronesi J, and Lovinger DM (2002) Postsynaptic endocannabinoid release is critical to long-term depression in the striatum. *Nature Neuroscience* 5: 446–451.

Heinbockel T, Brager DH, Reich CG, et al. (2005) Endocannabinoid signaling dynamics probed with optical tools. *Journal of Neuroscience* 25: 9449–9459.

Howlett AC, Barth F, Bonner TI, et al. (2002) International Union of Pharmacology. XXVII. Classification of cannabinoid receptors. *Pharmacological Reviews* 54: 161–202.

Mackie K (2005) Distribution of cannabinoid receptors in the central and peripheral nervous system. *Handbook of Experimental Pharmacology* 168: 299–325.

Marsicano G, Wotjak CT, Azad SC, et al. (2002) The endogenous cannabinoid system controls extinction of aversive memories. *Nature* 418: 530–534.

Monory K, Massa F, Egertova M, et al. (2006) The endocannabinoid system controls key epileptogenic circuits in the hyppocampus. *Neuron* 51: 455–466.

Nicoll RA and Alger BE (2004) The brain's own marijuana. *Scientific American* 291: 68–75.

Pagotto U, Marsicano G, Cota D, Lutz B, and Pasquali R (2006) The emerging role of the endocannabinoid system in endocrine regulation and energy balance. *Endocrine Reviews* 27: 73–100.

Piomelli D (2003) The molecular logic of endocannabinoid signalling. *Nature Reviews Neuroscience* 4: 873–884.

Varvel SA, Anum EA, and Lichtman AH (2005) Disruption of CB (1) receptor signaling impairs extinction of spatial memory in mice. *Psychopharmacology (Berl)* 179: 863–872.

Wilson RI and Nicoll RA (2001) Endogenous cannabinoids mediate retrograde signalling at hippocampal synapses. *Nature* 410: 588–592.

Wotjak CT (2005) Role of endogenous cannabinoids in cognition and emotionality. *Mini Reviews in Medicinal Chemistry* 5: 659–670.

Endocrine Function During Sleep and Sleep Deprivation

J M Mullington, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

Frequently measured hormonal profiles in healthy volunteers have included the chronobiological pineal hormone melatonin; hypothalamic anterior pituitary hormones, including cortisol, growth hormone (GH), prolactin, and thyroid-stimulating hormone (TSH); and more recently metabolic hormones related to nutrient homeostasis, leptin and ghrelin. The diurnal patterns of some important and frequently measured hypothalamic anterior pituitary hormones along with the rhythm of core body (rectal) temperature and a hypnogram of a typical normal sleep period are presented Figure 1.

Melatonin: The Chronobiological Pineal Hormone

Melatonin, the hormonal timekeeper, is produced by pinealocytes in the pineal gland. The pineal is situated in the depression between the colliculi of the midbrain on the dorsal aspect, projecting into the subarachnoid, outside the blood–brain barrier. It receives a large blood supply originating from the posterior cerebral artery and eventually drains into the internal cerebral veins, which flow in turn into the great cerebral vein. The human retinal ganglion cells make bilateral monosynaptic connection with the superchiasmatic nucleus of the anterior hypothalamic nucleus via the retinohypothalamic tract. These fibers continue and synapse at the paraventricular nucleus of the hypothalamus, project through the medial forebrain bundle, and synapse at the intermediolateral cell column of the spinal cord. This fiber tract then synapses on preganglionic cells that synapse at the superior cervical ganglion.

In humans, light inhibits melatonin secretion during the day. Melatonin is a strongly circadian (from the Latin phrase *circa diem,* meaning about a day) hormone that increases regularly at the dim light-onset time of day, peaking during the nocturnal sleep period, and diminishing again with the onset of bright light. In humans, melatonin increases at night independently of sleep and is inversely related to the rectal temperature rhythm. During early sleep deprivation studies, it was thought that at night light could not suppress melatonin in humans, but this was shown to be dependent on the light level. Later studies that applied bright light (>2000 lux) when melatonin levels were high during the sleep period were able to reduce melatonin to daytime or near-daytime levels, although they returned to presuppression levels within 40 min after resumption of darkness. Studies that followed found that the application of light is more or less effective at delaying or advancing the onset of melatonin, depending on the biological time (or circadian phase) at which it is applied. Bright light applied late in the day delays the melatonin onset, and light applied early in the morning advances it. Although studies have not shown melatonin to have clinically significant hypnotic qualities, the presence of melatonin may facilitate the coordination of sleep-onset processes.

Studies of the responsiveness of the melatonin rhythm to different intensities, spectrums, and timing of light administration have laid the foundation for therapeutic approaches to the treatment of jet lag, chronobiologically related mood disorder (seasonal affective disorder), and sleep disorders such as delayed sleep-phase syndrome, in which individuals are unable to keep a consistent sleep-onset time and instead go to bed later and later each night.

Key Hormones of the Anterior Pituitary

Anterior pituitary hormones are released in a pulsatile fashion and vary in peripheral concentration throughout the day. The determining factors are in part circadian (sleep-independent biological rhythms) and in part related to sleep and other behavioral processes. Cortisol and TSH are more strongly circadian in their diurnal secretion profiles and maintain their diurnal rhythms in the absence of sleep. GH and prolactin are more sleep-dependent and are generally suppressed in the absence of sleep, and they move with the shift of the sleep period in experimental studies of the effects of shifting sleep opportunity. As can be seen in Figure 1(a), cortisol reaches its normal daily low point, or nadir, in the nocturnal sleep period, typically an hour or two after sleep onset, corresponding to the time of maximal nocturnal GH release. Cortisol increases in the early hours of the morning, peaking shortly after the time of rising.

The paraventricular nucleus of the hypothalamus synthesizes and releases corticotropin-releasing hormone (CRH), which in turn regulates the release of adrenocorticotropin (ACTH) in the anterior pituitary, leading to the production and secretion of cortisol by

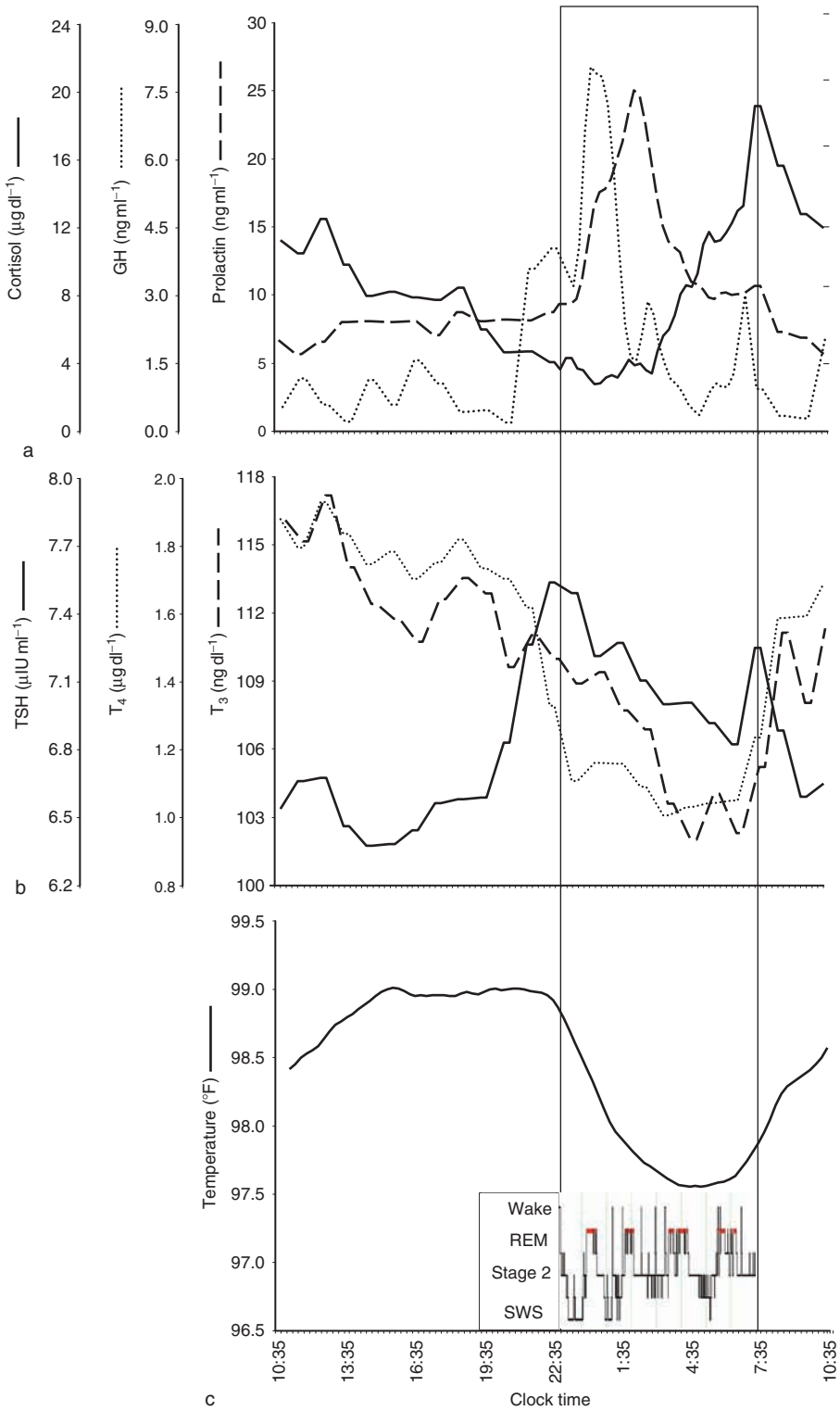

Figure 1 Data from 40 healthy men and women ages 21–40 combined in smoothed tracings: (a) diurnal hormonal profiles for cortisol (based on hourly samples) and for growth hormone (GH) and prolactin (based on half-hourly samples) through the night; (b) normal diurnal rhythm of thyroid-stimulating hormone (TSH), thyroxine (T_4), and triiodothyronine (T_3); (c) normal diurnal rhythm of temperature. The inset in (c) shows a representative sleep hypnogram from one of the subjects in the group of 40, illustrating the descent into slow-wave sleep (SWS) followed by a rapid eye movement (REM) period and subsequent cycling through non-REM and REM periods of sleep for the sleep period. which spans 11 p.m.–7 a.m.

the adrenal cortex. Growth hormone releasing hormone (GHRH) is produced by the arcuate nucleus, which induces the production and release of GH by the anterior pituitary. Somatostatin downregulates the production of GHRH. In addition, insulin-like growth factor (IGF-)I is stimulated by GH. IGF-I indirectly inhibits GH through its stimulation of somatostatin in the hypothalamus.

Cortisol

Cortisol, the most important and abundant glucocorticoid, is essential for proper water, carbohydrate, protein, and fat metabolism, and without this hormone, animals die within a week. In addition to these functions, cortisol is also important for the regulation of vascular reactivity, immune cells and inflammatory mediators, and cellular transcription processes. In addition to having a diurnal rhythm under pleasant relaxed conditions, the hypothalamic anterior pituitary system acts to increase circulating cortisol in response to unpleasant or potentially unpleasant stimuli. Although the reason is not well understood, the ability to withstand noxious stimuli is critically reliant on cortisol. In humans, there is far more cortisol than ACTH measured in the periphery, with a ratio of 7:1. Stimuli which activate ACTH secretion also elevate cortisol, and it is cortisol which is more readily measured in the periphery.

Although ACTH and cortisol are considered to be under strong circadian control, sleep loss and other behaviors do influence portions of the amplitude and shape of the curve. For example, waking causes a surge of ACTH and cortisol. This increase in ACTH is not simply a response to rising, but may also play a role in preparing for rising in that it has been shown that the anticipation of an earlier than normal awakening causes ACTH to rise earlier. Cortisol is also influenced by meals, the first morning exposure to light, stress, and noxious and other stimuli.

During sustained extended wakefulness under controlled laboratory conditions, an increase in the nadir of cortisol during total sleep deprivation has been observed. An afternoon elevation of cortisol has also been described under conditions of sustained reduced sleep. Although the significance of these modifications to the cortisol rhythm under conditions of insufficient sleep is not completely understood, it is known that during aging the nadir of cortisol increases, leading some to consider sleep loss a model of aging.

Cortisol levels are increased in depression and also have been seen to be elevated in individuals suffering from sleep disturbance. Sleep that is disturbed, either by sleep disorders such as sleep apnea or insomnia or disrupted by experimental means that cause fragmentation of the nocturnal sleep period, may lead to blunted GH and an elevated afternoon and early nighttime cortisol secretion.

Growth Hormone

GH is critical for growth and development, in which it is responsible for linear bone formation during early development, as well as growth and maintenance of tissues through ontogeny. During normal sleep, slow-wave sleep is predominant in the first half of the night with the hormonal release of GH at its highest level of the day. In fact, 50–70% of GH is released during the first half of the nocturnal sleep period in men, although the tightness of this association is somewhat lower in men younger than 25 years of age and higher in men older than middle age. Of particular note, the associative relationship between slow-wave sleep and GH release is also lower in women than it is in men.

Prolactin

Prolactin is usually secreted early after sleep onset and continues to rise through the night, peaking around the middle of the sleep period. Similar to GH, prolactin is suppressed during sleep deprivation and shifts with a delayed bed period. Brief awakenings during nocturnal sleep are associated with decreasing prolactin levels.

Rapid eye movement (REM) sleep predominates in the second half of sleep period and is associated with and accompanied by a rise in cortisol. A reciprocal interaction between the CRH and GHRH systems is thought to be important for the regulation of sleep, with the GHRH system promoting sleep and the CRH system stimulating ACTH and cortisol production and promoting arousal and wakefulness. In men, there is a positive relationship between the amount of slow-wave sleep and GH measured through the sleep period. Men with lower cortisol nadirs have been found to have higher amounts of REM sleep.

Thyroid-Stimulating Hormone

TSH is produced by parvocellular neurons in the paraventricular and anterior hypothalamic nuclei. TSH is increased during the nocturnal bed period and the production of thyroxine (T_4) and triiodothyronine (T_3) by the thyroid gland is concomitantly decreased. During acute total sleep deprivation of one night, TSH and free T_3 are increased above sleep control conditions and free T_4 is unchanged. In contrast to the elevation seen during total acute sleep

deprivation, nocturnal TSH has been reported to be reduced in the context of prolonged periods of reduced sleep amount relative to TSH levels seen during extended recovery sleep.

Consolidated nocturnal sleep, initiated at a regular time, when melatonin is rising and ended after the sleep need is satiated, also at a regular time, should facilitate optimal hormonal release and support homeostasis most efficiently. Consistent with this assumption, there is evidence that disruption of sleep, whether due to experimental manipulation or to sleep disorders, leads to nocturnal elevation of cortisol and TSH and a decrease in prolactin; all patterns that are in contrast with the entrained healthy-sleep hormonal profile.

Energy Metabolism and Adipose Tissue Hormones

Because hormonal systems play key roles in the maintenance of energy balance and metabolism and because consolidated sleep is essential for the appropriate timing and secretion amplitude of these hormonal systems, it is perhaps not surprising that sleep itself appears to play a role in metabolism. Under conditions of acute total sleep deprivation and also under conditions of prolonged shortened sleep, glucose metabolism is slowed and young healthy research participants begin to respond in a prediabetic fashion to glucose tolerance testing.

Leptin

Leptin is a hormone produced by adiposites; it signals energy sufficiency and satiety to the brain. It shows a regular diurnal rhythm, reaching maximal levels at night during sleep. Leptin's primary site of action is located in the arcuate nucleus of the hypothalamus. Leptin levels drop during fasting, and it is a hormone known to be important for reproductive capacity. In women, menses ceases when leptin levels drop below normal levels and resume when leptin is supplemented as treatment. Leptin is also integrally involved in energy homeostasis and appetite regulation. Under conditions of total acute and prolonged partial sleep deprivation, leptin amplitude is reduced, with a lowered nocturnal peak.

Ghrelin

Ghrelin is primarily produced by the stomach; it signals hunger to the brain. During reduced sleep, ghrelin levels and appetite are elevated. Insufficient or disrupted sleep leads to a negative energy balance and increased appetite, leading to a higher caloric intake. Furthermore, there is evidence that the food preferences under these conditions favor a high-carbohydrate diet.

Conclusion

The field of sleep and hormonal regulation is a rapidly developing one. New peptides and hormones are currently being discovered that are important for integrating metabolic function. Sleep is critical for the homeostasis of energy systems, and much remains to be learned about the mechanisms involved in translating sleep into health maintenance.

See also: Autonomic Dysregulation During REM Sleep; Circadian Function and Therapeutic Potential of Melatonin in Humans; Clock Gene Regulation of Endocrine Function; Immune Function During Sleep and Sleep Deprivation; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Metabolic Syndrome and Sleep; Pineal Gland and Melatonin; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Born J, Hansen K, Marshall L, Mölle M, and Fehm HL (1999) Timing the end of nocturnal sleep. *Nature* 397: 29–30.

Czeisler CA, Allan JS, Strogatz SH, et al. (1986) Bright light resets the human circadian pacemaker independent of the timing of the sleep-wake cycle. *Science* 223: 667–671.

Holsboer F, von Bardeleben U, and Steiger A (1988) Effects of intravenous corticotropin-releasing hormone upon sleep-related growth hormone surge and sleep EEG in man. *Neuroendocrinology* 48: 32–38.

Lewey AJ, Wehr TA, Goodwin FK, Newsome DA, and Markey SP (1980) Light suppresses melatonin secretion in humans. *Science* 210: 1267–1269.

Obál F Jr., Payne L, Opp M, Alföldi P, Kapás L, and Krueger JM (1992) Growth-hormone releasing hormone antibodies suppress sleep and prevent enhancement of sleep after sleep deprivation. *American Journal of Physiology* 263: R1078–R1085.

Parker DC and Rossman LG (1971) Human growth hormone release in sleep: Nonsuppression by acute hyperglycemia. *Journal of Clinical Endocrinology* 32: 65–69.

Sassin JF, Parker DC, Mace JW, Gotlin RW, Johnson LC, and Rossman LG (1969) Human growth hormone release: Relation to slow wave sleep and sleep-waking cycles. *Science* 165: 513–515.

Spiegel K, Leproult R, and Van Cauter E (1999) Impact of sleep debt on metabolic and endocrine function. *Lancet* 354: 1435–1439.

Weitzman ED, Nogeire C, Perlow M, et al. (1974) Effects of a prolonged 3-hour sleep-wake cycle on sleep stages, plasma cortisol, growth hormone and body temperature in man. *Journal of Clinical Endocrinology & Metabolism* 38: 1018–1030.

Zee PC and Turek FW (eds.) (1999) *Lung Biology in Health and Disease, Vol. 133: Regulation of Sleep and Circadian Rhythms.* New York: Dekker.

Endocrinology of Animal Communication: Behavioral

G F Ball, Johns Hopkins University, Baltimore, MD, USA
J Balthazart, University of Liège, Liège, Belgium

© 2009 Elsevier Ltd. All rights reserved.

What Is Animal Communication and Why Is It Important to Neuroscientists?

Basics of Communication

This article concerns the hormonal regulation of animal communication. The basics of communicative interactions for the purposes of this article are summarized in Figure 1. In Figure 1(A), the essential components are illustrated as a signal-generating source that sends a signal via a transmission channel. During transmission, there are several possible fates of the signal: It might miss the receiver (a), be blocked somehow while in the transmission channel (b), reach the receiver directly (c), or reach the receiver via an indirect route (d). This theoretical notion is shown with an actual example in Figure 1(B), in which a female deer is illustrated releasing a pheromone that is transmitted via the air to the male.

Significance of Animal Communication for Neuroscientists

The study of animal communication addresses many questions of fundamental importance to the field of neuroscience. Communication signals are often produced by specialized effector organs. This fact gives neuroscientists a clear starting point for neural circuitry studies and therefore facilitates the nearly complete identification of neural circuits that project to these effector organs. In many species, there are also clear co-evolutionary processes between signals and their receiver systems that also facilitate the identification of sensory specializations. Communicative interactions often involve intentional behaviors that are profoundly influenced by social context and experience. This means that systems that mediate communication are linked to forebrain circuits regulating cognitive processing in interesting ways. Communication signals often need to be coordinated with other physiological and behavioral events to function properly. One of the best examples of such coordination concerns courtship signals used in the context of reproduction. Reproduction is generally periodic so that animal populations can adjust breeding to times when conditions are favorable for the successful rearing of progeny. As a result, the timing of signal production and any specialized changes in effector systems needed for proper signal production must be regulated to ensure that they occur at the appropriate time. Reproductive hormones, such as the gonadal sex steroids, orchestrate these diverse events so that they are temporally configured in an appropriate manner. Also, neural systems controlling communication are often regulated by hormones. An important theme of this article is a consideration of the hormonal regulation of communication in the reproductive context.

Hormonal Regulation of Signal Production

To illustrate how hormones can facilitate the production of communication signals, we consider whether gonadal androgens such as testosterone enhance high rates of signaling in the context of courtship and reproduction. A causal relationship between a hormone and a behavior can be established with the use of three criteria: (1) Removal of the source of the hormone (in the case of testosterone, the testis) should lead to a decline in the production of the behavior; (2) reinstatement of plasma concentrations of the hormone via exogenous administration should also reinstate high rates of behavioral production; and (3) behavioral frequency, at least at the population level, should be positively correlated with natural variation in the plasma concentration of the hormone. Avian studies of courtship vocalizations clearly illustrate the value of these criteria. In many temperate zone seasonally breeding songbird species, the rate of male song is positively correlated with the vernal increase in plasma testosterone that is induced in part by increasing photoperiods. As illustrated in Figure 2(a), in Corsican blue tits (*Parus caeruleus*), plasma testosterone concentrations (as determined by serial blood sampling and subsequent measurement via radioimmunoassay) increase dramatically in mid- to late March just prior to an equally dramatic increase in the rate of male song in the sampled population.

Studies that have manipulated the availability of endogenous testosterone by removal of the testis also support its importance in enhancing male vocal production. For example, male Japanese quail (*Coturnix japonica*) produce a species-typical vocalization called the 'crow' when trying to attract females during the breeding season. Castration of captive male quail results in a complete cessation of this behavior (Figure 2(b)). However, replacement of testosterone via the administration of silastic capsules filled with crystalline testosterone can reinstate the behavior. As illustrated in Figure 2(b), groups of castrated male quail were implanted with capsules of increasing length (the

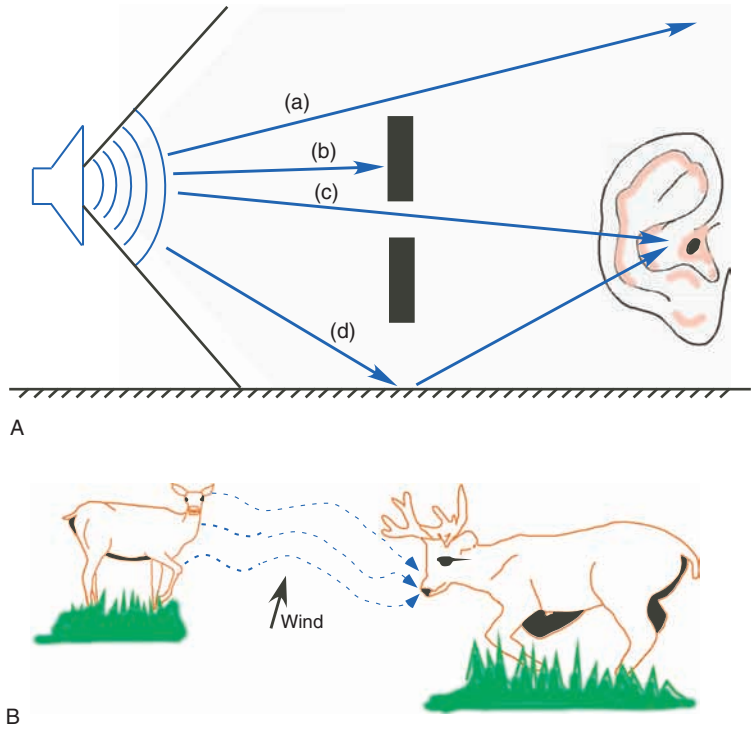

Figure 1 Animal communication, like every form of communication, involves signal generation, transmission, and perception. Transmission can be affected by a variety of factors. (A) Schematic representation of sound transmission from a loudspeaker to an ear. Sound can be lost in space (a), blocked by an obstacle (b), or reach its target directly (c) or indirectly (d). (B) Similarly in animal communication, a female deer in this example emits various olfactory signals originating in several skin glands. Signals will reach the receiver's nose (male deer) after diffusion in the transmission space, eventually affected by environmental factor (wind in this case). Similar concepts apply for olfactory, acoustic, and visual communication.

length of the capsules determines the concentration of testosterone in the plasma, with longer implants leading to higher concentrations of the steroid). The percentage of males in which crowing was restored increased in a dose-dependent manner. This study illustrates another principle of behavioral activation by hormones, namely that there are often individual differences within a study population in the behaviorally effective concentration of a hormone. As illustrated in Figure 2(b), 10-mm-long implants could restore crowing behavior in more than 50% of the castrated males, but to ensure behavioral restoration in the entire sample implants four times longer were required. However, results such as these only explain part of the variance related to the hormonal regulation of behavior. Variation in the target sensitivity to hormone action, in this case the central nervous system, is also important.

In the case of testosterone and male-typical courtship signals, the ability of the target tissue to metabolize testosterone to another hormone is an important factor. It has been recognized since the early 1970s that testosterone is essentially a 'prohormone,' meaning that the form of the hormone released into the blood from the testis is not the same as the form that acts on the target tissue. Testosterone, an androgen, is routinely metabolized both in the brain and in the periphery to either estrogenic or androgenic metabolites. For example, brain aromatase, which is very active in the male brain in many vertebrate species, can convert testosterone to the estrogen 17β-estradiol, whereas the enzyme 5α-reductase can convert testosterone to the potent androgenic metabolite 5α-dihydrotestosterone. Therefore, testosterone effects on the activation of male-typical signaling systems may require its conversation to another androgen, an estrogen, or both. In male zebra finches (*Taeniopygia guttata*), castration reduces the rate of courtship song production, and this is reinstated by both testosterone and another androgen, androstenedione, both of which can be metabolized into androgens and/or estrogens (Figure 2(c)). The administration of androgenic or estrogenic metabolites of testosterone alone is not effective in reinstating high rates of song. However, the combination of androgenic and estrogenic metabolites (i.e., 5α-dihydrotestosterone and 17β-estradiol) is effective in restoring song behavior to a rate similar to that induced by exogenous testosterone in the castrates (Figure 2(c)).

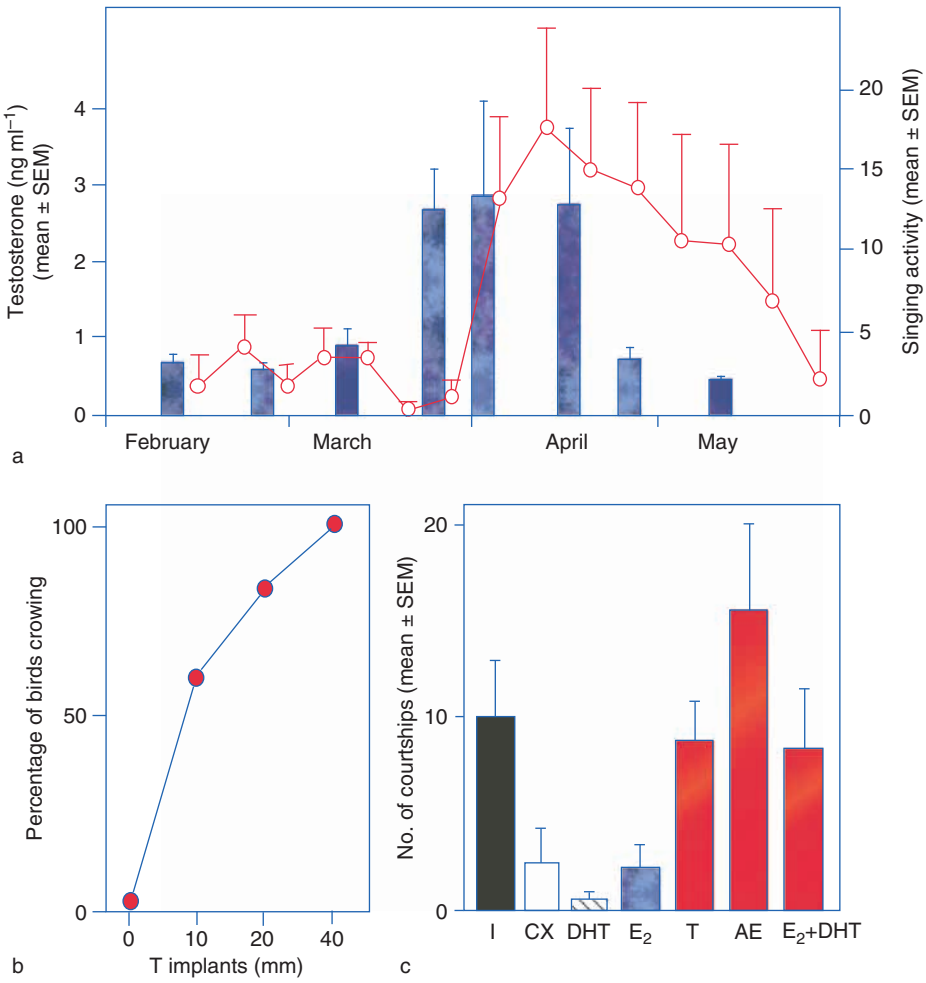

Figure 2 Hormones modulate the rate at which social signals are emitted by animals and this phenomenon has been particularly well documented in birds. (a) The number of songs (curve with standard errors, SEM) emitted by Corsican blue tits detected during systematic field recordings varies markedly during the spring. These behavioral changes are paralleled, actually slightly preceded and thus presumably caused to a large extent, by changes in plasma testosterone concentration (bars). (b) Treatment of castrated male quail with increasing amounts of exogenous testosterone (T) provided by subcutaneous implants of increasing length augments in a dose–response manner the percentage of males that produce crows, a call attracting females in naturalistic conditions. (c) The number of courtship songs produced by sexually mature, gonadally intact (I) male zebra finches exposed to a female is markedly decreased by castration (CX) and restored to nearly intact levels by treatments with exogenous steroids that simultaneously provide androgens and estrogens, such as the aromatizable androgens testosterone (T) and androstenedione (AE) or the combined administration of 5α-dihydrotestoeterone (DHT) and estradiol (E_2). DHT, a nonaromatizable androgen, and E_2 alone have little or no effect.

These studies also illustrate another principle of behavioral endocrinology: Hormones such as testosterone do not cause the activation of a communication signal or any other behavior in a simple stimulus–response relationship. In all these studies, vocal behavior was elicited by the presence of an appropriate conspecific female. The presence or absence of testosterone did not absolutely determine the occurrence of the behavior. In the case of the blue tits in the nonbreeding season and castrated zebra finches, the presence of low testosterone concentrations did not result in a complete decline in song behavior. Rather, the probability and intensity of the vocal response to a female are modulated by the presence of testosterone in high concentrations in the blood.

Hormonal Regulation of Signal Quality

Often, the effectiveness of a signal is not simply the result of whether it is produced at a high rate but also is influenced by aspects of its quality. For example, one of the most important functions of reproductive signals is for conspecifics to recognize that a signaler is a member of a particular species and is of a particular sex. In addition to this fundamental information essential for a successful mate choice, individuals are

sensitive to variation in the signal that seems to predict the mate's quality. Both of these aspects of signal quality can be regulated by hormones. For example, in some electric fish, males produce a markedly lower frequency of electric organ discharges (EODs) than females (Figure 3(a)). The frequency of these EODs can be significantly lowered into the male range by administering to adult females either testosterone or another gonadal androgen quite common in fish, 11-keto-testosterone (Figure 3(a)).

Male canaries (*Serinus canaria*) produce a complex song whose rate of production is enhanced in a seasonal manner by the presence of testosterone. Females also sing, but generally it is a simpler song that is not used in a reproductive context. Males sing at high rates during the breeding season to attract females. Administering testosterone to adult females induces song behavior that is much more male-like than the naturally occurring female song. Playback studies show that one component of male canary song that females find particularly attractive consists of syllables repeated at a very high rate. When females are treated with testosterone, they produce these attractive syllables. Interestingly, this particular component of male song is selectively regulated by estrogenic metabolites of testosterone. When exogenous testosterone is administered simultaneously with an inhibitor of the enzyme aromatase that results in a blockade of the production of estrogenic metabolites, the resulting song is more male-like overall but it lacks the syllables that are particularly attractive to females. Thus, testosterone can do more than just increase the rate of song production; it can also selectively regulate song quality via estrogenic metabolites. These examples of testosterone and male signaling can be generalized to a variety of species and hormone systems.

Effects of Signal Perception on Hormone Release

Another important principle is that there are reciprocal interactions between hormones and signal production. Being on the receiving end of a signal can lead to change in hormone concentrations in the receiver. This change in the receiver's physiology can then affect subsequent and ongoing social interactions. The ability of signals to regulate hormone secretion in the receiver is one criterion for assessing the signal's effectiveness. For example, if a sexual signal from a male or a female fails to induce a hormonal change in the receiver, then the reproductive sequence cannot be successfully completed. Furthermore, producing a signal can change the hormonal state of the sender as well. Thus, studies of the physiological consequences of signal production are not complete unless the effects of the signal on hormonal secretion in both the receiver and the sender are assessed, for example, by measuring physiological responses to signal production.

In Gulf toadfish (*Opsanus beta*), breeding males emit a 'boatwhistle' vocalization in territorial encounters. Playbacks of pure tones that mimicked the duration of the boatwhistle vocalization were able to elicit a behavioral response, whereas a series of other sounds were ineffective. Furthermore, this sound only elicited

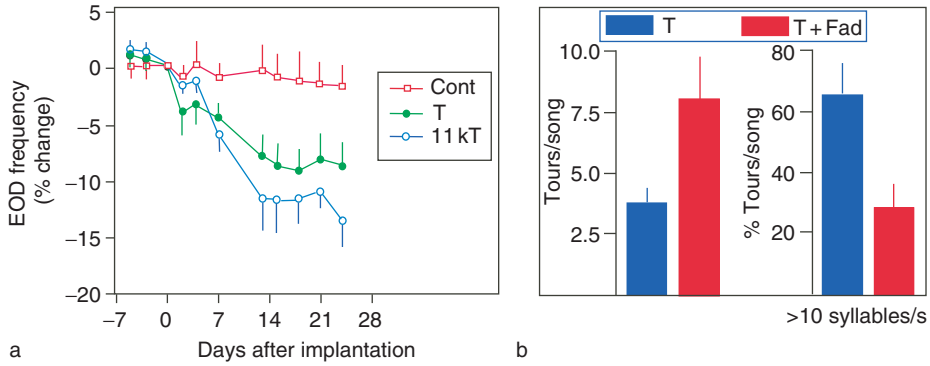

Figure 3 Hormones affect the quality of the social signal emitted by animals. (a) Gymnotiform electric fish produce electric organ discharges (EODs) used for sex recognition that display a lower frequency (in Hz) in males compared to females. The lower frequency of these electric signals in males is controlled by androgens, as demonstrated by the fact that a treatment for 1 month of control (Cont) fish (groups of three males and three females in each case) with testosterone (T) or 11-keto-testosterone (11kT), a very active androgen in many fish species, markedly lowers the EOD frequency in a time-dependent manner. (b) In canaries, song quantity and quality are affected by androgens and estrogens. Song is induced in females (who rarely sing) by a treatment with testosterone (T) that provides both androgenic and estrogenic stimulation (through central aromatization of T into estradiol). Concurrent treatment with the aromatase inhibitor fadrozole (Fad) markedly changes the structure and organization of songs produced. The number of tours (syllables) per song is increased but there is a lower proportion of tours with a syllable repetition rate higher than 10 syllables/s. These modified songs produced by T + Fad females are known to be sexually less attractive.

a rapid increase in plasma 11-keto-testosterone concentrations (Figure 4(a)). This rapid increase appears to facilitate responses to territorial challenge that males of this species exhibit during the breeding season. Many species of frogs in the temperate zone exhibit explosive breeding assemblages that are tied to spring rains. Males produce species-typical vocalizations in a chorus that has profound consequences for the endocrine physiology of the participants. Studies of tree frogs (*Hyla cinerea*) reveal that the playback of a chorus of conspecific calls for ten nights to isolated males results in a marked increase in plasma testosterone, whereas in control males the presentation of a tone has no effect (Figure 4(b)). These two studies clearly illustrate how a specialized signal can have an effect on endocrine physiology.

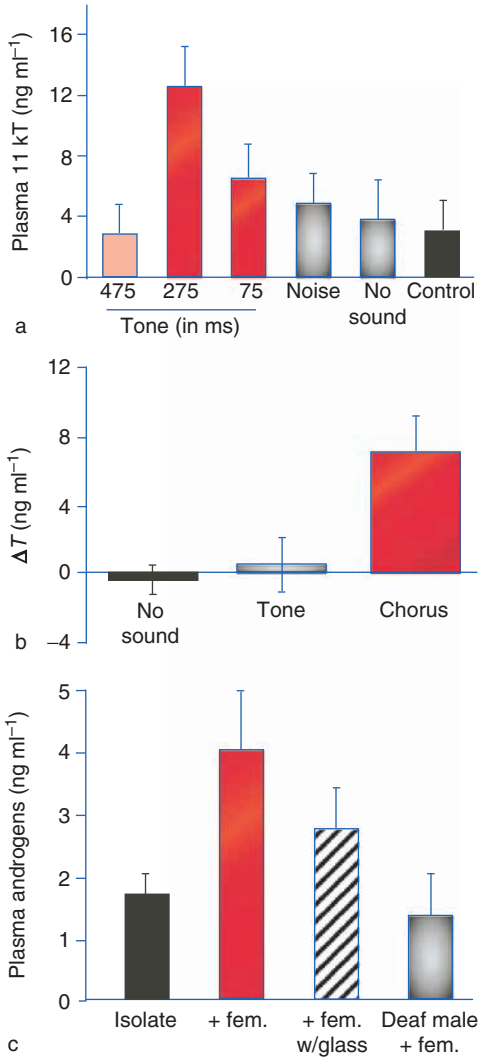

Figure 4 Social signals affect the endocrine physiology of many species in all classes of vertebrates. (a) In Gulf toadfish (*Opsanus beta*), the playback of sounds that closely resemble vocalizations emitted by reproductively active males (tone with a duration of approximately 275 ms) markedly increases plasma concentration of 11-keto-testosterone (11kT), the major androgen in these fishes. Other tones with a similar structure but different duration or other acoustic signals do not have similar effects. (b) Exposure for ten nights of frogs (*Hyla cinerea*) to the chorus of conspecific calls markedly increases plasma testosterone (T) levels in comparison to controls or to frogs exposed to a control sound (tone) that does not share acoustic similarities with the species-typical calls. (c) In ring doves (*Streptopelia risoria*), exposure of an adult male to a sexually receptive female increases plasma levels of androgens (essentially T and DHT) within 24 h. This effect is partly masked if a physical barrier (glass partition) separates the male from the female and is completely absent if the male has been experimentally deafened, which suggests a role for both visual and acoustic stimuli in the control of androgen concentration by stimuli emanating from the female.

Communication can involve signals that are multimodal. Male and female ring doves court each other with displays that involve both visual and auditory components. These courtship signals result in physical contact between the members of the pair when they mutually preen each other and then copulate. Classic studies on captive ring doves (*Streptopelia risoria*) by Daniel Lehrman in the 1960s provided some of the first demonstrations that courtship interactions have marked effects on the endocrine physiology of the receiver of a signal. Subsequent studies by his colleagues were able to tease apart the stimulus factors that are able to elicit these endocrine responses. For example, the presentation of a female to a male that has been maintained in social isolation for several weeks results in a marked increase in his plasma testosterone concentrations that is apparent within 1 day. The ability of a female to stimulate a male is dependent on her endocrine state, which determines whether she will engage in the appropriate reproductive behaviors. This can be easily demonstrated by pairing a male with an ovariectomized female that fails to perform reproductive displays in the presence of a male. There is a significant increase in testosterone even if the male is separated from an intact behaving female via a glass partition, but if the male can see but not hear the female, the androgen rise is greatly attenuated (Figure 4(c)). Thus, experiencing just the appropriate visual and especially auditory stimuli without tactile contact is sufficient to stimulate androgen concentrations in the male in response to a female.

Two additional aspects relevant to our understanding about how the perception of communication signals can regulate endocrine physiology have been clarified by studies of ring doves. For a displaying male to stimulate endocrine activity in female ring doves, it is not enough for the female to see a displaying male: It must be clear that the male is directing his display at the female. If a female sees a displaying male through a one-way glass so that he cannot see her and direct his attention at her, the effect of this experience on her endocrine physiology is significantly reduced compared with that of a female seeing the same male and

that he can see and direct his attention toward. A second phenomenon is that the behavior of the receiver in response to the displaying male can be important in enhancing physiological responses to a sender. Female ring doves that were devocalized and experienced a displaying male exhibited a markedly reduced endocrine response compared with females who could vocalize. Furthermore, this decline in hormonal response could be reversed by playing back the cooing calls that the female usually produces in response to courtship signals from the male.

The previous two studies raise important issues for studies of changes in endocrine physiology in response to a signal. First, the social context and the experience that the receiver has with both the signal and the signaler can have a significant effect on the receiver's endocrine response. Second, the receiver's behavioral response can influence the magnitude of its endocrine response to the receiver's signal.

Neural Mechanisms of Steroid Hormone Action on Signal Production and Perception

General Principles of the Cellular Actions of Steroid Hormones on Behavior

Steroids and other classes of hormones can activate communicatory behaviors by acting on sensory inputs, neural pathways in the central nervous system, and/or peripheral effector organs that produce the signal or behavior of interest. Whether or not any of these different processes will be mediated by hormone action is determined by whether the target tissue expresses the appropriate receptors. Steroid hormones are lipophilic compounds that readily cross the cell membrane and bind to a receptor protein that is a member of the steroid/thyroid receptor family of intracellular receptors. This ligand–receptor complex then acts in the cell nucleus to induce gene transcription, which is necessary for the observed biological responses that are induced by the hormone. The hormone–receptor complex must recruit a number of other proteins in order to induce efficient gene transcription, but a discussion of the molecular basis of hormone-induced gene transcription is beyond the scope of this article.

Receptor proteins can be localized in the brain or in peripheral tissues by chemical neuroanatomical methods, such as autoradiography and immunohistochemistry. An alternative method that is commonly utilized is to localize the mRNA for the receptor gene of interest with *in situ* hybridization histochemistry methods. These methods provide behavioral endocrinologists with useful insight into the localization of hormone action in the brain (Figure 5). These types of studies can be followed up with experimental investigations such as precise stereotaxic implantation of hormones to the potential sites of interest in gonadectomized animals. In this way, the functional significance of the hormone expression can be investigated. The connectivity of hormone-sensitive brain sites of functional significance can then be described so that a neural circuit regulating the behavior of interest can be discerned. In the next section, three such circuits are discussed.

Examples of Neural Circuits Regulating the Activation of Communication Signals That Are Regulated by Steroid Hormones

Three examples of hormone-regulated neural circuits mediating the expression of male-typical vocal signals used in courtship and aggression are illustrated in Figure 5, which diagrammatically illustrates the neural control of vocalization in the midshipman fish, *Porichthys notatus* (a), in the African clawed frog, *Xenopus laevis* (b), and the generalized circuit (c) that seems to apply to all male songbirds (i.e., species in the suborder Passeres of the Passeriformes avian order). Brain areas in the circuit that are 'steroid sensitive' are defined by the stippling pattern. In this case, the term steroid sensitive means that the area expresses either androgen receptors (AR) or one of the two types (α or β) of estrogen receptor (ER). As discussed previously, the activation of many male-typical behaviors by testosterone requires the conversion in the brain of testosterone to an estrogen such as 17β-estradiol or another more potent androgen, such as 5α-dihydrotestosterone. Therefore, AR and both ERα and ERβ are relevant. The localization in the brain of the relevant metabolizing enzymes has been mapped to varying degrees in these three species, but they are omitted in order to minimize the complexity of the figure. It should also be noted that in all these species there is evidence that the relevant effector organ producing the signal (i.e., the larynx in *Xenopus* and the syrinx in the songbirds) is a target of testosterone action because cells in these structures also express AR.

Several generalizations emerge from a comparison of these figures. In all three species, there are midbrain-to-brain stem projections that ultimately synapse on motor neurons that innervate the vocal-producing effector organ (i.e., the torus semicircularis to midbrain vocal acoustic integration center and ultimately to sonic motor nucleus in the midshipman fish; the anterior thalamus ultimately to the motor nucleus of the glossopharyngeal nerve in *Xenopus* and nucleus intercollicularis to the motor nucleus of the hypoglossal nerve, tracheosyringeal part (nXIIts) in the songbirds;

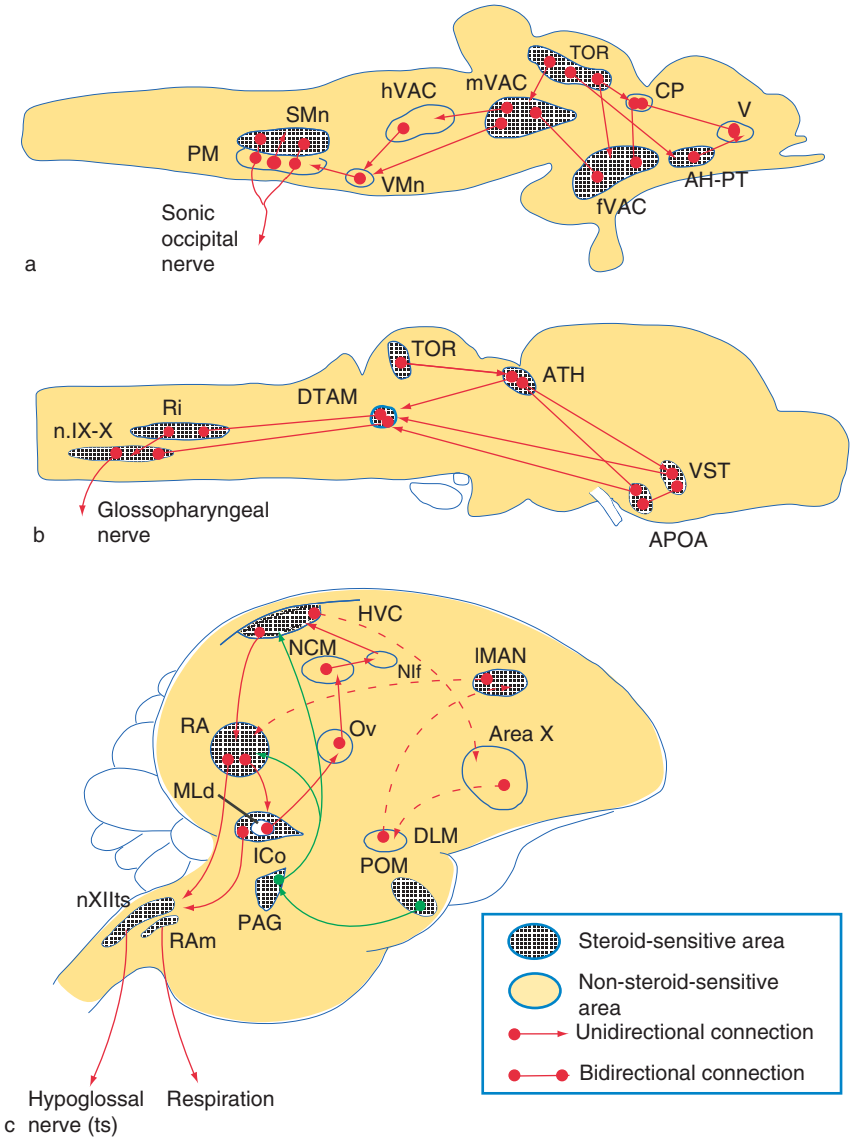

Figure 5 Schematic representation of the neural circuits controlling vocalizations in batrachoidid fishes such as the midshipman fish, *Porichthys notatus* (a), in the African clawed frog *Xenopus laevis* (b), and in songbirds (c). Key parts of the vocal pathways as well as auditory pathways are illustrated. For the auditory pathways, only the last relays that interface with vocal structures are indicated and are labeled 'auditory' in the list of abbreviations. Aspects of the auditory pathways and the vocal pathways were shown to be steroid sensitive (the presence of estrogen and/or androgen receptors) and are represented in the figure by stippled areas. In songbirds, the tentative connections between the medial preoptic nucleus and the song system via the periaqueductal gray are indicated in green. The anterior forebrain pathway mediating song learning is indicated by dashed arrows. (a) AH–PT, anterior hypothalamus–posterior tubercle; CP, central posterior nucleus (auditory); f-, h-, and mVAC, forebrain, hindbrain, and midbrain vocal acoustic integration centers, respectively; PM, pacemaker nucleus; SMn, sonic motor nucleus; TOR, torus semicircularis (auditory); V, ventral telencephalon; VMn, ventral medullary nucleus. (b) APOA, anterior preoptic area; ATH, anterior thalamus (auditory); DTAM, dorsal tegmental area of the medulla; n.IX-X, motor nucleus of the glossopharyngeal nerve; Ri, nucleus reticularis inferior; TOR, torus semicircularis (auditory); VST, ventral striatum. (c) DLM, nucleus dorsolateralis anterior thalami, pars lateralis; HVC, used as proper name (formerly high vocal center); ICo, nucleus intercollicularis; IMAN, lateral nucleus of the anterior nidopallium; MLd, nucleus mesencephalicus lateralis, pars dorsalis (auditory); NCM, caudal medial nidopallium (auditory); NIf, nucleus interfacialis; nXIIts, motor nucleus of the hypoglossal nerve, tracheosyringeal part; Ov, nucleus ovoidalis (auditory); PAG, periaqueductal gray (also called substantia grisea centralis in birds); POM, nucleus preopticus medialis; RA, nucleus robustus of the arcopallium; RAm, nucleus retroambigualis.

Figure 5). In all these cases, both the midbrain and the brain stem nuclei express AR. Thus, in all these examples, neural areas exerting immediate control over the effector organ are potentially regulated by androgens. Another generalization is that these midbrain–brain stem areas in turn receive projections from diencephalic and telencephalic regions that are also steroid sensitive in that they express either AR or ER. There

are interesting taxonomic differences in this case. In songbirds, the telencephalic structure RA (robust nucleus of the arcopallium) expresses AR but sends a direct projection down to the motor nucleus nXIIts that innervates the avian vocal production organ the syrinx. Such a direct projection from the telencephalon to the brain stem is unusual in vertebrates and is not observed in *Xenopus* or the midshipman, in which a multisynaptic pathway links the telencephalon and/or diencephalon with the relevant motor nuclei. In all cases, one can observe links between the preoptic area/hypothalamus and the motor production brain regions. Again, these areas also express either AR or ER of various types. Brain areas such as subregions of the preoptic area and the hypothalamus are important sites for the hormonal activation of reproductive behaviors. They tend to integrate relevant sensory inputs, resulting in an organized motor response. Thus, the connections in all three species between these brain areas and midbrain and brain stem areas involved in signal production are consistent with general views of hormonally regulated neural circuits controlling sexual and parental behaviors.

There is still much to be learned about exactly how steroid hormones act in these target areas to activate communication signals. Although it is known that the steroid–receptor complex regulates gene expression, the field is only beginning to identify the relevant genes. One candidate gene for the control of social signal production codes for the neuropeptide arginine vasopressin (AVP) or, in nonmammalian vertebrates, the homologous neuropeptide arginine vasotocin (AVT). There is evidence in *Xenopus* as well as in Syrian hamsters, in which testosterone induces flank marking used by males as an aggressive signal in a reproductive context, that testosterone upregulates this peptide system and that is a necessary step for signal production. In this case, the evidence is particularly compelling that AVP acting in the anterior hypothalamus can reliably induce high rates of flank marking. In songbirds, testosterone acting via estrogenic and androgenic metabolites induces marked changes in the volume of key forebrain song nuclei so that they are larger in the spring and early summer, when males are producing song in a courtship and/or territorial context, and much smaller in the fall, when song is not produced at all or produced in a different context. Song, unlike the calls produced by the midshipman and frogs, is learned, and it was initially hypothesized that these changes might be related to seasonal variation in song learning. However, not all seasonally breeding songbird species change their repertoire due to song learning, yet the volume of these nuclei does clearly change with season. Therefore, these changes in the volume of the song nuclei seem to be most reliably linked to changes in song performance, such as the quantity or the quality (stereotypy or consistency) of the signal.

The physiological or biochemical cascade of events induced by steroid hormones will almost certainly be different in forebrain areas compared to motor neurons. An important future goal is therefore to identify specific genes regulated by steroids in the different brain areas that make up the communication circuit. In weakly electric fish, the rate and quality of EODs are regulated by steroid hormones, as mentioned previously. In particular, dihydrotestosterone has been shown to slow down while 17β-estradiol speeds up the rate of inactivation of Na^+ currents. This effect seems to result from steroid hormone effects on the expression of genes coding for specific ion channels or on enzymes that could influence the state of the channels such as kinases. Future studies will seek to identify the specific genes regulated by hormones that result in changes in signaling ability.

Hormone Effects on Sensation and Perception

Although most work on hormones and signaling has stressed the production of signals, hormones can also influence communication by tuning the reception of signals. In neotropical electric fish that exhibit wave-like EODs, the electrosensory receptors are precisely tuned to perceive the EOD that an individual fish produces. When exogenous steroid hormones are administered, the EOD changes, as does the tuning of the electroreceptors. This change in tuning is a direct effect of steroids on the tuning properties of electroreceptors. There are other examples of peripheral sensory organs being regulated by hormones. In nonreproductive female midshipman fish, the administration of 17β-estradiol increases the temporal encoding of frequency by the inner ear such that the phase locking matches the frequency of the male call that is produced in the breeding season. Thus, changes in the female's reproductive state mediated by sex steroid hormones result in her matching the tuning of her auditory periphery to the male's call. In one form or another, regulation of auditory tuning may be widespread. There is evidence, for example, that the accuracy in judging octave frequency varies over the menstrual cycle in women, being at the lowest level of accuracy during the midluteal phase and peaking during ovulation when plasma 17β-estradiol is at its highest concentration. Similarly, auditory brain stem responses in birds exhibit a seasonal change in sensitivity that could be related to seasonal changes in tuning to male song. Finally, hormone-induced adjustments in sensory processing in rats have been

observed in the olfactory system with 17β-estradiol enhancing sensory perception of male-typical cues based on recordings in the female preoptic area. There is also evidence that estrogen enhances the size of the genital sensory field in female rats to facilitate tactile stimulation from males that results in female receptive responses.

Overall, there is good evidence that the peripheral tuning of sensory systems relevant to the sensation and perception of specific signals can be adjusted by hormonal fluctuations, although the exact mechanisms are unclear. However, work on estradiol effects on the inner ear of midshipman fish is a promising model system for the investigation of the cellular bases of adjustments in auditory tuning that can result in an organized hormonal response, as previously discussed. The neural pathways by which sensory information is conveyed to the hypothalamic–pituitary–gonadal axis to induce an endocrine change are also not well understood. However, work in frogs and in ring doves has described projections from midbrain and thalamic auditory areas to the preoptic region that may mediate the effects of auditory stimulation on hormonal secretion in these taxa.

Conclusions

In this article, we highlighted studies of hormone, brain, and behavior interrelationships that are relevant to the study of communication. The production of specialized signals for communicatory interactions is controlled by neural circuits that can be regulated by hormones at the level of sensory inputs, central processing, and motor outputs. The perception of signals can in turn result in a marked change in the endocrine physiology of the receiver. We provided examples of all these processes that have been investigated in a variety of vertebrate taxa. One current focus of the field is to identify specific genes regulated by steroids and other hormones that result in defined changes in behavior. Some effects of hormones on signaling systems were not discussed in this article. For example, signaling systems are sex typical in many species, and such sex differences in neural organization result from early ontogenetic effects of steroid hormones. We focused on adult activational effects. Also, there is clear evidence that in addition to the well-characterized genomic effects of steroid hormones on communication circuits, relatively rapid nongenomic effects can also occur. Such rapid effects are undoubtedly important for the regulation of social interactions related to communication, but little is known about them.

See also: Agonistic and Affiliative Signals: Resolutions of Conflict; Birdsong Learning: Evolutionary, Behavioral, and Hormonal Issues; Communication Networks and Eavesdropping in Animals; Localizing Signal Sources; Neuroendocrinology of Social/Affiliative Behavior; Pheromones and other Chemical Communication in Animals; Signal Production and Amplification in Birds; Signal Identification: Peripheral and Central Mechanisms; Signal Transmission in Natural Environments; Vocal Communication in Birds.

Further Reading

Adkins-Regan E (2005) *Hormones and Animal Social Behavior.* Princeton, NJ: Princeton University Press.

Ball GF (1999) The neuroendocrine basis of seasonal changes in vocal behavior among songbirds. In: Hauser MD and Konishi M (eds.) *The Design of Animal Communication*, pp. 213–254. Cambridge, MA: MIT Press.

Bass AH and Zakon HH (2005) Sonic and electric fish: At the crossroads of neuroethology and behavioral neuroendocrinology. *Hormones and Behavior* 48: 360–372.

Hauser MD (1996) *The Evolution of Communication.* Cambridge, MA: MIT Press.

Hopp SL, Owren MJ, and Evans CS (eds.) (1998) *Animal Acoustic Communication.* Heidelberg: Springer.

Kelley DB and Brenowitz EA (2002) Hormonal influences on courtship behaviors. In: Becker JB, Breedlove SM, Crews D, and McCarthy MM (eds.) *Behavioral Endocrinology,* 2nd edn., pp. 289–329. Cambridge, MA: MIT Press.

Kelley DB and Tobias ML (1999) Vocal communication in *Xenopus laevis*. In: Hauser MD and Konishi M (eds.) *The Design of Animal Communication*, pp. 9–36. Cambridge, MA: MIT Press.

Remage-Healey L and Bass AH (2005) Rapid elevations in both steroid hormones and vocal signaling during playback challenge: A field experiment in Gulf toadfish. *Hormones and Behavior* 47: 297–305.

Schlinger B and Brenowitz EA (2002) Neural and hormonal control of birdsong. In: Paff DW, Arnold AP, Etgen AM, Fahrbach SE, and Rubin RT (eds.) *Hormones, Brain and Behavior,* vol. 2, pp. 799–840. San Diego: Academic Press.

Wingfield JC (2006) Communicative behaviors, hormones–behavior interactions, and reproduction in vertebrates. In: Neil JD (ed.) *Knobil and Neill's Physiology of Reproduction,* vol. 2, 3rd edn., pp. 1995–2040. San Diego: Academic Press.

Zakon HH and Smith GT (2002) Weakly electric fish: Behavior, neurobiology, and neuroendocrinology. In: Paff DW, Arnold AP, Etgen AM, Fahrbach SE, and Rubin RT (eds.) *Hormones, Brain and Behavior,* vol. 2, pp. 349–374. San Diego: Academic Press.

Zeigler HP and Marler P (eds.) (2004) Behavioral neurobiology of birdsong. *Annals of the New York Academy of Sciences* 1016.

Endocytic Traffic in Spines

S S G Ferguson, University of Western Ontario, London, ON, Canada
Y T Wang, University of British Columbia, Vancouver, BC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Mechanisms of Spine Endocytosis

Endocytosis of most plasma membrane proteins is mediated primarily by a mechanism involving clathrin and several clathrin adaptor proteins. Clathrin coats are recruited to the plasma membrane and linked to receptors by adaptor proteins. Most integral plasma membrane proteins, including growth factor receptors and ligand-gated ionotropic neurotransmitter receptors, are recruited to clathrin-coated pits through the adaptor protein AP-2, a heterotetrameric complex containing $\alpha\beta_2\mu_2\sigma_2$ subunits. Most metabotropic G-protein-coupled receptors (GPCRs) employ a family of monomeric adaptor proteins known as β-arrestins. In addition to binding the serine-phosphorylated carboxyl tails of the GPCRs, the β-arrestin C-terminal domains interact with both the clathrin heavy chain and the β2-adaptin subunit of the heterotetrameric AP-2 adaptor complex, thereby targeting the activated receptor to clathrin-coated pits. Binding of adaptor proteins to receptors and to clathrin initiates clathrin coat assembly and membrane invagination, and the vesicles then pinch off via a dynamin guanosine triphosphatase-dependent process. The resultant clathrin-coated vesicle is internalized from the plasma membrane to the intracellular compartment, where it is either recycled to the plasma membrane or delivered to lysosomes for degradation. Recent electron microscopy (EM) of hippocampal neurons has shown that most of the endocytic machinery, including clathrin, AP-2, and dynamin, is expressed within spines, consistent with the ability of a spine to undergo clathrin-dependent endocytosis. Indeed, both EM and light microscope studies have been able to demonstrate coated pits and vesicles within spines. It is interesting that these studies found that the coated vesicles are in close association with, but rarely directly at, the postsynaptic density (PSD), suggesting that just lateral to the PSD exist some 'hot spots' where spine endocytosis may usually take place.

Spine Endocytosis in Controlling the Density of Postsynaptic Glutamate Receptors and Synaptic Plasticity

One general function of clathrin-mediated endocytosis is to regulate the density of plasma membrane proteins via constitutive and/or regulated endocytotic retrieval (recycling) of these proteins into intracellular compartments. Glutamate is the major excitatory neurotransmitter in the mammalian brain and acts primarily on two large families of receptors, the ionotropic and the metabotropic glutamate receptors. Ionotropic glutamate receptors are organized into three main classes: α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors, which comprise various combinations of glutamate receptor subunits GluR1–4; kainate receptors, which consist of GluR5–7, KA1, and KA2 subunits; and the *N*-methyl-D-aspartate (NMDA) receptors, which comprise NR1 and NR2A–D subunits. The metabotropic glutamate receptors are a family of GPCRs comprising mGluR1 to mGluR8. Most of these glutamate receptors are expressed in postsynaptic spine membranes, and their density in the membrane is subject to regulation by clathrin-mediated spine endocytosis.

AMPA Receptor Endocytosis and Its Role in Synaptic Plasticity

Among all glutamate receptor subfamilies, the spine endocytosis of the AMPA receptor subtype is, both mechanistically and functionally, the best characterized. Structurally, AMPA receptors are presumed to be heterotetrameric structures assembled by combining homologous subunits. The current topology model suggests that each AMPA receptor subunit (GluR) consists of three transmembrane regions, membrane domain (MD) 1, MD3, and MD4, and a reentrant membrane segment corresponding to MD2. The major intracellular C-terminal region contains potential sites for protein phosphorylation and binding motifs for GluR-interacting proteins, including adaptor proteins involved in forming clathrin-coated endocytic vesicles. Three distinct types of clathrin-dependent AMPA receptor endocytosis have been proposed: constitutive, ligand-induced, and activity-dependent (regulated) endocytosis.

Constitutive endocytosis occurs across all AMPA receptors and requires a motif in the most juxtamembrane region of the C-terminus that is 100% conserved among all AMPA receptor GluR subunits. Interference with the constitutive pathway does not lead to detectable alterations in the number of AMPA receptors on the cell surface, likely because of the existence of a constitutive AMPA receptor exocytosis pathway that balances the constitutive endocytosis.

Like most GPCRs, AMPA receptors also undergo clathrin-dependent ligand-induced endocytosis. Since both agonists and antagonists evoke AMPA receptor endocytosis, this form of endocytosis seems solely dependent on ligand–receptor interactions and does

not require receptor activation. Similar to constitutive AMPA receptor endocytosis, ligand-induced endocytosis applies to all GluR subunits and requires the conserved sequences in the proximal membrane region of the GluR C-terminus. The physiological significance of this receptor endocytosis remains unknown, but analogous to what is observed for GPCRs, it may represent an important feedback mechanism minimizing receptor overactivation under certain physiological and pathological conditions.

Regulated endocytosis is specifically associated with AMPA receptors containing the GluR2 subunit. This GluR2 specificity is mediated by several sequence motifs uniquely present within the C-terminal region of the subunit, including a middle domain that binds to both *N*-ethylmaleimide-sensitive factor and AP-2 clathrin adaptor and a terminal PDZ-binding domain that interacts with a number of PDZ-containing proteins, such as GRIP and PICK1. In addition, systematic mutational analysis has revealed a novel tyrosine phosphorylation-dependent endocytic motif in the GluR2 C-terminus that is absolutely required for regulated AMPA receptor endocytosis. Given the fact that interfering with any of these motifs blocks activity-dependent AMPA receptor endocytosis, all these motifs appear to be functional and may be indispensably involved in regulated AMPA receptor endocytosis in neurons. The mechanisms by which these multiple internalization signals act in concert during various forms of synaptic plasticity remain to be illustrated.

Functionally, unlike constitutive endocytosis, regulated AMPA receptor endocytosis leads to a rapid and persistent reduction in the number of receptors expressed on the postsynaptic membrane surface of the spine and hence plays an important role in controlling synaptic strength. The ability to reduce cell surface AMPA receptors makes regulated endocytosis an ideal candidate mechanism for the expression of certain forms of long-term depression (LTD), one of the best-characterized forms of synaptic plasticity believed to be involved in learning and memory. Consistent with this hypothesis, interfering with these GluR2 C-terminal endocytic motifs by application of short 'interference peptides' comprising the tyrosine-containing motif or AP-2- or PICK1-interacting sequences can prevent the expression of hippocampal CA1 homosynaptic LTD, both in brain slices and in intact, anesthetized rats. Moreover, although the facilitated endocytosis of postsynaptic AMPA receptors is principally observed in conjunction with hippocampal CA1 LTD, it appears to be a common mechanism involved in the expression of certain forms of LTD in many other areas of the brain, including the cerebellum and the nucleus accumbens. These studies have firmly established a critical role for AMPA receptor endocytosis as a common mechanism of LTD expression. The results also suggest that these three GluR2–C-terminal endocytic signals are indispensably involved in regulated AMPA receptor endocytosis and hence LTD expression and that peptides interfering with the function of these signals may be used as specific inhibitors to probe the functional roles of LTD. Indeed, using one of these membrane-permeable interfering peptides, researchers were recently able to provide the first evidence for a causative role of LTD in the expression of behavioral sensitization induced by repeated intermittent administration of drugs of abuse, such as amphetamine, in an animal model of drug addiction. Thus, endocytosis of AMPA receptors in spines may be the principle mechanism of regulating the strength of synaptic transmission and may be critically important to both brain function and dysfunction.

NMDA Receptor Endocytosis

Because of the voltage-dependent blockade by physiological concentrations of extracellular Mg^{2+}, the NMDA subtype of ionotropic glutamate receptors contributes little to basal synaptic transmission. However, because of its high calcium permeability, activation of the receptor is thought to be fundamental to many normal brain functions, such as cognition and memory, and its overactivation is believed to be critically involved in mediating neuronal death during both acute brain injuries and chronic brain degenerative disorders. Therefore, the trafficking of NMDA receptors has recently been the subject of intensive investigation. From what is known so far, the distinctions between constitutive and regulated trafficking seen with AMPA receptors are less obvious with NMDA receptors. However, several mechanisms of clathrin-mediated NMDA receptor endocytosis have been characterized and are presumed to contribute to the maintenance of steady state receptor levels in spine membranes. Native NMDA receptors are tetrameric complexes comprised of NR1 subunits with at least one type of NR2 subunit. Different NR2 subunits (NR2A–D) confer distinct properties during spine endocytosis. At the postsynaptic spine membrane, NMDA receptors are stabilized by interacting with postsynaptic density protein 95 (PSD-95). This interaction is mediated via the PDZ-binding domain present in the C-terminus of either NR2A or NR2B and can result in rapid internalization via clathrin-mediated endocytosis in response to the binding of agonist glutamate or coagonist glycine. The NMDA receptor–PSD-95 interaction and subsequent receptor endocytosis appear to be subject to regulation by receptor phosphorylation, tyrosine phosphorylation in particular. Phosphorylation of the NR2B C-terminus

interferes with its interaction with PSD-95, thereby facilitating clathrin-mediated endocytosis via a mechanism of promoting the binding of AP-2 to NR2B. There is also evidence that dephosphorylation of NR1 and NR2A on tyrosine residues may promote clathrin-dependent endocytosis in a similar manner. Few studies have investigated activity-dependent NMDA receptor endocytosis, but a facilitated clathrin-mediated endocytosis of NMDA receptors has recently been reported to be associated with increased AMPA receptor endocytosis after the induction of CA1 homosynaptic LTD in a hippocampal slice preparation. Detailed understanding of the mechanisms underlying this phenomenon remains to be achieved, but preliminary data suggest a mechanism distinct from that involved in AMPA receptor endocytosis. Thus, there is evidence that the density of postsynaptic NMDA receptors and hence the strength of the receptor-mediated synaptic transmission is also under the regulation of spine endocytosis.

mGluR Endocytosis

The density of postsynaptic mGluRs is also under the control of spine endocytosis. Like most other GPCRs, mGluRs are subject to ligand-induced and clathrin-dependent receptor endocytosis. For instance, agonist application causes rapid internalization of mGluR1 in a β-arrestin- and dynamin-dependent manner. However, emerging evidence suggests that not all GPCRs internalize via a β-arrestin-dependent mechanism. The agonist-independent or constitutive endocytosis of mGluR1 and mGluR5 appears to be dependent on the formation of a multimolecular complex that includes phospholipase D2, Ral, and RalGDS. However, the detailed mechanisms underlying spine endocytosis of mGluRs and their roles in mediating structural and functional synaptic plasticity remain unclear and await further investigation.

Spine Endocytosis in Signal Transduction

Emerging data suggest that in addition to controlling the density of postsynaptic plasma membrane proteins such as neurotransmitter receptors via internalization, spine endocytosis may have an additional function, mediation of intracellular signal transductions ranging from the formation of signaling complexes to synapse–nucleus signal transduction. It is well characterized that in addition to mediating the G-protein-uncoupling and internalization of GPCRs, β-arrestins link GPCRs to G-protein-independent signal transduction pathways. The β-arrestins interact with a wide variety of signaling proteins, including kinases (e.g., Src, Raf-1, JNK, ERK1/2, Akt), phosphatases (PP2A), guanine nucleotide exchange factors (ARNO and RalGDS), and small G-proteins (Arf6 and Rho). The diversity of β-arrestin interacting proteins potentially links β-arrestin function to a wide variety of neuronal phenotypes, such as alterations in the actin cytoskeleton reorganization that are required for neurite outgrowth and gene transcription. With regard to spine endocytosis at excitatory synapses, activation of NMDA receptors can lead to rapid clathrin-mediated endocytosis of AMPA receptors, resulting in LTD, and can also cause excitotoxic neuronal death. It is interesting that blocking the NMDA receptor-mediated clathrin-dependent AMPA receptor endocytosis, while having no effect on NMDA-induced Ca^{2+} responses, prevents both NMDA-induced activation of caspase-3 and neuronal apoptosis in hippocampal neurons maintained in primary culture. The detailed mechanisms by which endocytosed AMPA receptors contribute to the initiation of downstream apoptotic cascades remain to be determined. However, analogous to the recently reported retinal cell apoptosis induced following stimulated endocytosis of rhodopsin and activation of epidermal growth factor receptor-mediated mitogen-activated protein kinase signaling by endocytosed epidermal growth factor receptors, it is possible that endocytosed AMPA receptors, in response to NMDA stimulation, may form, with some of the clathrin-coat protein components, a signaling complex that initiates the apoptotic cascade. These results not only identify AMPA receptor endocytosis as a critical link between NMDA-induced Ca_i^{2+} overload and intracellular cascades leading to apoptosis but also suggest that, in addition to a decrease in receptor-mediated synaptic transmission, spine endocytosed AMPA receptors play previously unsuspected roles in intracellular signaling pathways. The mediation of NMDA-induced apoptosis is one example.

Spine Endocytosis in Structural Plasticity of Excitatory Synapses

Neuronal spines dynamically change their size, shape, and number throughout the life of an animal. Synapse elimination (pruning) is a fundamental process during early development and also an important mechanism for activity-dependent fine-tuning of neuronal networks during learning and memory in the mature brain. In addition, synaptic loss is a common pathological feature of a broad range of psychiatric diseases and neurological disorders. The precise way in which synapse elimination occurs remains a controversial topic, likely involving multiple mechanisms that include synaptic activity, membrane trafficking, and spine cytoskeletal rearrangement. Evidence that spine endocytosis plays a critical role in synapse pruning in

mature neurons has come from two recent studies of hippocampal CA1 LTD in slice preparations. In one, two-photon time-lapse imaging of dendritic spines in acute hippocampal slices from neonatal rats showed that the expression of LTD is accompanied by a marked shrinkage of spines. Since facilitated endocytosis of postsynaptic AMPA receptors is the primary mechanism of LTD expression, the study suggests a potential role of spine endocytosis in spine shrinkage. Indeed, the causative link between the endocytosis of AMPA receptors and spine removal (synapse loss) was established in a second study. The researchers found that amyloid beta (Abeta), a peptide believed to underlie the pathophysiology of Alzheimer's disease, facilitates endocytosis of synaptic AMPA receptors during LTD. This facilitated AMPA receptor endocytosis is both sufficient and necessary for the loss of dendritic spines, thereby indicating the central role played by AMPA receptor trafficking in Abeta-induced modification of synaptic structure and function. Together, these studies demonstrate that spine endocytosis may have a critical role in synapse pruning or elimination, thereby controlling the number of glutamatergic excitatory synapses in both the developing and the mature brain.

Concluding Remarks

Endocytosis is a fundamental cellular process in all eukaryotic cells. In neurons, endocytosis can occur within the spine. Mechanistically, spine endocytosis is likely mediated by both clathrin-dependent and clathrin-independent pathways, but the best-characterized spine endocytic mechanism occurs via clathrin-coated vesicles. Functionally, spine endocytosis allows retrieval of postsynaptic plasma membranes, thereby modifying the size, shape and even number of synapses, and recycling of integral plasma membrane proteins, including neurotransmitter receptors, to control their density (number) at the postsynaptic membrane, thereby regulating the strength of synaptic transmission. Emerging evidence also suggests that endocytosed spine plasma membrane proteins such as AMPA receptors may form part of intracellular signaling complexes that initiate cell signaling processes, introducing a third functionality of spine endocytosis.

See also: AMPA Receptors: Molecular Biology and Pharmacology; AMPA Receptor Cell Biology/Trafficking; Dendritic Spine History; Endocytosis and Presynaptic Scaffolds; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation; Postsynaptic Specialization Assembly; Spines and Mental Disorders.

Further Reading

Collingridge GL, Isaac JT, and Wang YT (2004) Receptor trafficking and synaptic plasticity. *Nature Reviews Neuroscience* 5: 952–962.

Dale LB, Bhattacharya M, Seachrist JL, Anborgh PH, and Ferguson SSG (2001) Agonist-dependent and -independent internalization of metabotropic glutamate receptor 1a: β-Arrestin isoform-specific endocytosis. *Molecular Pharmacology* 60: 1243–1253.

Hsieh H, Boehm J, and Sato C (2006) AMPAR removal underlies Aβ-induced synaptic depression and dendritic spine loss. *Neuron* 52: 831–843.

Kennedy MJ and Ehlers MD (2006) Organelles and trafficking machinery for postsynaptic plasticity. *Annual Review of Neuroscience* 29: 325–362.

Leftowitz RJ and Shenoy SK (2005) Transduction of receptor signals by beta-arrestins. *Science* 308: 512–517.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* 25: 103–126.

Morishita M, Marie H, and Malenka RC (2005) Distinct triggering and expression mechanisms underlie LTD of AMPA and NMDA synaptic responses. *Nature Neuroscience* 8: 1043–1050.

Nong Y, Huang YQ, and Ju W (2003) Glycine binding primes NMDA receptor internalization. *Nature* 422: 302–307.

Roche KW, Standley S, McCallum J, Dune LYC, Ehlers MD, and Wenthold RJ (2001) Molecular determinants of NMDA receptor internalization. *Nature Neuroscience* 4: 794–802.

Song I and Huganir RL (2002) Regulation of AMPA receptors during synaptic plasticity. *Trends in Neurosciences* 25(11): 578–588.

Vissel B, Krupp JJ, Heinemann SF, and Westbrook GL (2001) A use-dependent tyrosine dephosphorylation of NMDA receptors is independent of ion flux. *Nature Neuroscience* 4: 587–596.

Waites CL, Craig AM, and Garner CC (2005) Mechanisms of vertebrate synaptogenesis. *Annual Review of Neuroscience* 28: 251–274.

Wang Y, Ju W, and Liu L (2004) α-Amino-3-hydroxy-5-methylisoxazole-4-propionic acid receptor endocytosis is essential for N-methyl-D-aspartate-induced apoptosis. *Journal of Biological Chemistry* 279: 41267–41270.

Zhou Q, Homma KJ, and Poo M (2004) Shrinkage of dendritic spines associated with long-term depression of hippocampal synapses. *Neuron* 44: 749–757.

Endocytosis and Presynaptic Scaffolds

V Haucke, Freie Universität Berlin, Berlin, Germany
E D Gundelfinger, Leibniz Institute for Neurobiology, Magdeburg, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons communicate with each other by the temporally and spatially controlled release of secretory molecules (the so-called neurotransmitters) via regulated exocytosis. Following diffusion across the synaptic cleft the released nonpeptide neurotransmitters and neuroactive peptides bind to and activate postsynaptic receptors, which then elicit a response within the postsynaptic cell. In the case of most fast-acting transmitters, such as glutamate, γ-aminobutyric acid (GABA), or acetylcholine, signaling is elicited by ligand-gated ion channels, which are clustered at morphologically discernible zones specialized for chemical neurotransmission termed postsynaptic densities (PSDs). The PSD is opposed to a corresponding presynaptic element including the 'active zone' at which presynaptic neurotransmitter-bearing vesicles (SVs) are clustered. Presynaptic active zones are characterized by an electron-dense grid of scaffolding proteins interconnected with an actin-rich cytoskeleton, which among other functions helps to maintain a pool of vesicles docked at the presynaptic plasmalemma. To sustain chemical neurotransmission under conditions of high activity and to counter-balance net insertion of membrane by exocytic vesicle fusion, SVs undergo activity-driven cycles of calcium-triggered exocytosis and endocytosis within nerve terminals, commonly referred to as the SV cycle. Cycling of SVs must allow them to retain their specific biochemical identity, including the ability to store neurotransmitter by proton pump-driven neurotransmitter transporters, and to undergo further rounds of calcium-induced fusion with the presynaptic plasmalemma. The observed tight coupling between exocytic neurotransmitter release by vesicle fusion and compensatory endocytosis has resulted in a long and still unresolved debate regarding the precise molecular mechanisms involved in SV cycling. Here, we provide a brief summary of the pathways of SV cycling, the role of clathrin and its partner proteins in maintaining SV pools, and the temporal and spatial cues provided by scaffolding proteins and membrane lipids in maintaining presynaptic exocytic–endocytic membrane traffic.

Pathways of SV Cycling

Clathrin-Mediated Endocytic Cycling of Presynaptic Vesicles

Landmark studies in the early 1970s by Heuser and Reese showed that following fusion by complete collapse into the plasmalemma, SVs are retrieved by compensatory 'clathrin-dependent endocytosis' at specialized endocytic areas just outside the active zone. An overwhelming amount of genetic, morphological, biochemical, and physiological data suggests that clathrin-mediated endocytosis indeed constitutes an essential pathway of SV recycling, at least on the organismic level and over extended periods of time. Mutants in clathrin coat components, including the AP-2 complex or accessory and adaptor proteins such as dynamin, AP180, stoned B (the *Drosophila* ortholog of mammalian stonins 1 and 2), eps15, synaptojanin, endophilin, amphiphysin, or intersectin/DAP160, all display defects in neurotransmission owing, at least in part, to impaired SV endocytosis (Table 1). The most dramatic phenotypes have been observed following injection of dominant-negative domains or inhibitory peptides into the giant reticulospinal synapse of the lamprey. Following electrical stimulation, distinct endocytic intermediates and vacuolar structures accumulate within and around the active zones in conjunction with a partial or complete depletion of the recycling vesicle pool. Some of these intermediates resemble structures seen in neuromuscular junctions following intense stimulation or after genetic perturbation of protein function in temperature-sensitive alleles. The mechanistic details of this pathway will be discussed later.

Kiss-and-Run Mode

Based on an apparent lack of correlation between the number of morphologically distinct stable endocytic intermediates and the synaptic endocytic activity, Ceccarelli and colleagues proposed an alternative model according to which SVs release their contents through the controlled opening of a narrow fusion pore, followed by rapid closure and refilling with neurotransmitter (Figure 1). This 'kiss-and-run' mode of regulated secretion has been convincingly demonstrated to occur by combined electrophysiological recordings and membrane capacitance measurements in neuroendocrine cells, which mostly secrete peptide hormones or biogenic amines from large secretory granules (SGs; also termed large dense core vesicles (LDCVs)). In the case of SG exocytosis, flickering of

Table 1 Endocytic proteins involved in presynaptic vesicle cycling and their interaction partners

Endocytic protein	*Domains and motifs*	*Interaction partner(s)*	*Proposed function(s)*
Abp1	ADF homology domain	F-actin	Linking actin cytoskeleton with endocytosis and CAZ
	SH3 domain	PRDs of synaptojanin, dynamin, Piccolo, and synapsin1	
AP-2 (α, β2, μ2, σ2 subunits)	Trunk	PI(4,5)P_2	Plasmalemmal recruitment
	α-Appendage	FXDXF/DXW, WVXF motif proteins	Coat assembly
	β2-Appendage	$[DE]_nX_{1-2}$FXX[FL]XXR motif proteins	Cargo selection and coat assembly
	β2-Hinge	Clathrin terminal domain	Scaffolding
	C-μ2	Yxxϕ and basic motif membrane cargo; PI(4,5)P_2	Membrane cargo selection
Amphiphysins	N-BAR domain	(Acidic) phospholipids; dimerization	Membrane curvature induction/sensing; scaffolding
	PWXXW, LLDLD motifs	Clathrin terminal domain	Coat assembly
	FXDXF, DXF motifs	AP-2 via α-appendage	
	SH3 domain	PRDs of dynamin (& synaptojanin)	Vesicle fission
AP180	ANTH domain	PI(4,5)P_2	Plasmalemmal recruitment
	FXDXF, DXF motifs	AP-2 via α-appendage	Coat assembly
	DLL motifs	Clathrin	Scaffolding
Auxilin	DnaJ domain	Hsc70	Stimulation of ATPase
	DPF, WDW motifs	AP-2 via α-appendage	Coated pit recruitment
	WDW, DLL motifs	Clathrin	
Clathrin (HC, LCa/b)	HC-terminal domain	PWXXW, LLDLD, DLL motif proteins	Assembly of scaffold
	LC	HIP1, calmodulin, Hsc70	Diverse
Dynamin 1	GTPase domain	GDP/GTP	Control of membrane fission
	GED domain	Dynamin-GED	Self-assembly; stimulation of GTPase activity
	PH domain	PI(4,5)P_2	Plasmalemmal recruitment
Endophilin	N-BAR domain	(Acidic) phospholipids; dimerization	Membrane curvature induction/sensing
	SH3 domain	PRD of dynamin and synaptojanin	Coated pit maturation; fission and uncoating

Epsins	ENTH domain	$PI(4,5)P_2$	Membrane curvature induction; coat assembly
	DPW motifs	AP-2 via α-appendage	
	LLDLD motifs	Clathrin terminal domain	Scaffolding
	UIMs	Ubiquitin (Ub)	Ub-dependent cargo endoctyosis
Eps15, Eps15R	EH domains	NPF motif proteins	Endocytic protein network formation
	DXF motifs	AP-2 via α-appendage	Coat assembly (edges of CCPs)
	UIMs	Ubiquitin (Ub)	Ub-dependent cargo endoctyosis
HIP1, HIP1R	ANTH domain	$PI(4,5)P_2$	Plasmalemmal recruitment
	FXDXF, DXF motifs	AP-2 via α-appendage	Coat assembly
	LLDLD motif	Clathrin terminal domain	Scaffolding; links actin with endocytosis
Intersectin	EH domains	NPF motif proteins	Endocytic protein network formation
	SH3 domains	PRD domain proteins	CDC42-mediated actin polymerization
	Rho-GEF domain	CDC42	
Numb, Numb-like	PTB	$PI(4,5)P_2$	Plasmalemmal recruitment; coat assembly
	DXF motifs	AP-2 via α-appendage	
PIPK type Iγ	PIP kinase domain	PI(4)P, ATP; Arf6-GTP	Localized $PI(4,5)P_2$ formation; regulation of
	WYSPL tail peptide	FERM domain of talin	$PI(4,5)P_2$ synthesis at cell adhesion sites
Stonin 2	WVXF motifs	AP-2 via α-appendage	Coat assembly; coated pit recruitment
	Stonin-homology domain	?	?
	μ-Homology domain	Synaptotagmin (1,2,9)	AP-2-dependent recycling
Synaptojanin	SAC1 domain	4′-phosphate-containing phosphoinositides	4′-phosphoinositide phosphate hydrolysis
	5′-Phosphatase domain	$PI(4,5)P_2$	$PI(4,5)P_2$ hydrolysis, CCV uncoating coat assembly; coated pit recruitment
	FXDXF, WVXF motifs	AP-2 via α-appendage	
Synaptotagmin	C2A	Ca^{2+}, acidic phospholipids	Ca^{2+}-triggered membrane fusion membrane fusion; SV recycling
	C2B	$PI(4,5)P_2$; AP-2 via subdomain B of μ2	
	C2AB	Stonin 2 via μ-homology domain	SV recycling
Syndapins	BAR	(Acidic) phospholipids; dimerization	Membrane curvature sensing actin and dynamin-mediated fission
	SH3	PRD of N-WASP and dynamin	

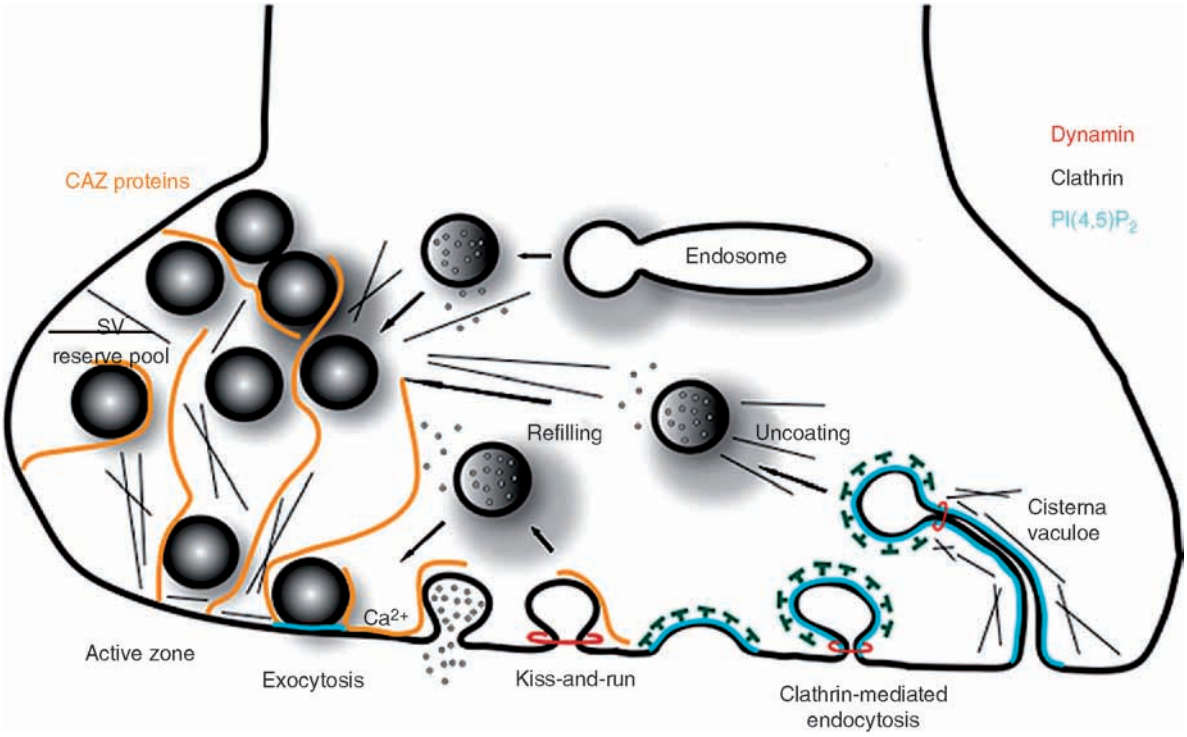

Figure 1 Pathways of SV recycling. Schematic depiction of various proposed modes of synaptic vesicle (SV) recycling: a fast 'kiss-and-run' mechanism, where the vesicle connects only briefly to the plasma membrane without full collapse ('kiss-and-run') or a slow clathrin-mediated pathway, which either operates from large vacuolar infoldings (cisternae) or by direct recovery of vesicle membrane from plasmalemmal CCPs ('clathrin-mediated endocytosis'). SVs may also arise from endosomes. Components of the cytomatrix assembled at the active zone (CAZ) together with actin function as molecular scaffolds in the spatial organization of the active zone. The SV cycle is paralleled by a cycle of phosphorylation and dephosphorylation of phosphoinositides, including $PI(4,5)P_2$, that couple exocytosis and endocytosis.

a transient fusion pore precedes complete degranulation. However, in contrast to small clear SVs that undergo local recycling, SGs need to pass through the *trans*-Golgi network in order to allow refilling with secretory peptides generated from larger precursor proteins. The recent development of lipophilic fluorescent styryl dyes (FM dyes) that rapidly partition into membranes and exhibit a large increase in fluorescence within this hydrophobic environment and of pH-sensitive fluorescent proteins (so-called 'pHluorins'; Figure 2) has provided the means to follow exocytic–endocytic cycling of SVs in real-time. FM1–43 dye-based single-vesicle tracking in dissociated hippocampal neurons in culture has revealed the existence of at least two types of release: small-amplitude events that show tightly clustered rate constants of dye release, and larger events with a more scattered distribution. The small-amplitude partial release events have been attributed to a pool of vesicles that undergoes cycling by rapid closure of a narrow, approximately 1 nm diameter, fusion pore. One would therefore have to assume that vesicles are targeted for partial release by specific factors that prevent the dilation and thus the complete opening of the fusion pore. Whether vesicles undergoing transient opening and closure of the fusion pore remain docked ('kiss-and-stay') or undergo local cycling (as depicted in Figure 1) is under debate. The balance between partial kiss-and-run-type and full fusion events that may be followed by clathrin-dependent compensatory endocytosis can be shifted depending on the frequency of stimulation. While kiss-and-run exocytosis may prevail under conditions of low activity, high-frequency stimulation results in predominantly complete fusion events. Membrane capacitance measurements of giant terminals (e.g., goldfish retinal bipolar cells or the calyx of Held) have also provided evidence for two kinetically distinguishable cycling vesicle pools. However, all of these studies suffer from the lack of information on specific factors that target SVs for fast kiss-and-run exocytosis–endocytosis and that allow the application of genetic or biochemical tools to molecularly distinguish the proposed kiss-and-run mode from other pathways of SV endocytosis.

Vacuolar Bulk Retrieval and Synaptic Endosomes

Extensive stimulation of the presynaptic neuron results in the massive insertion of SV membrane into

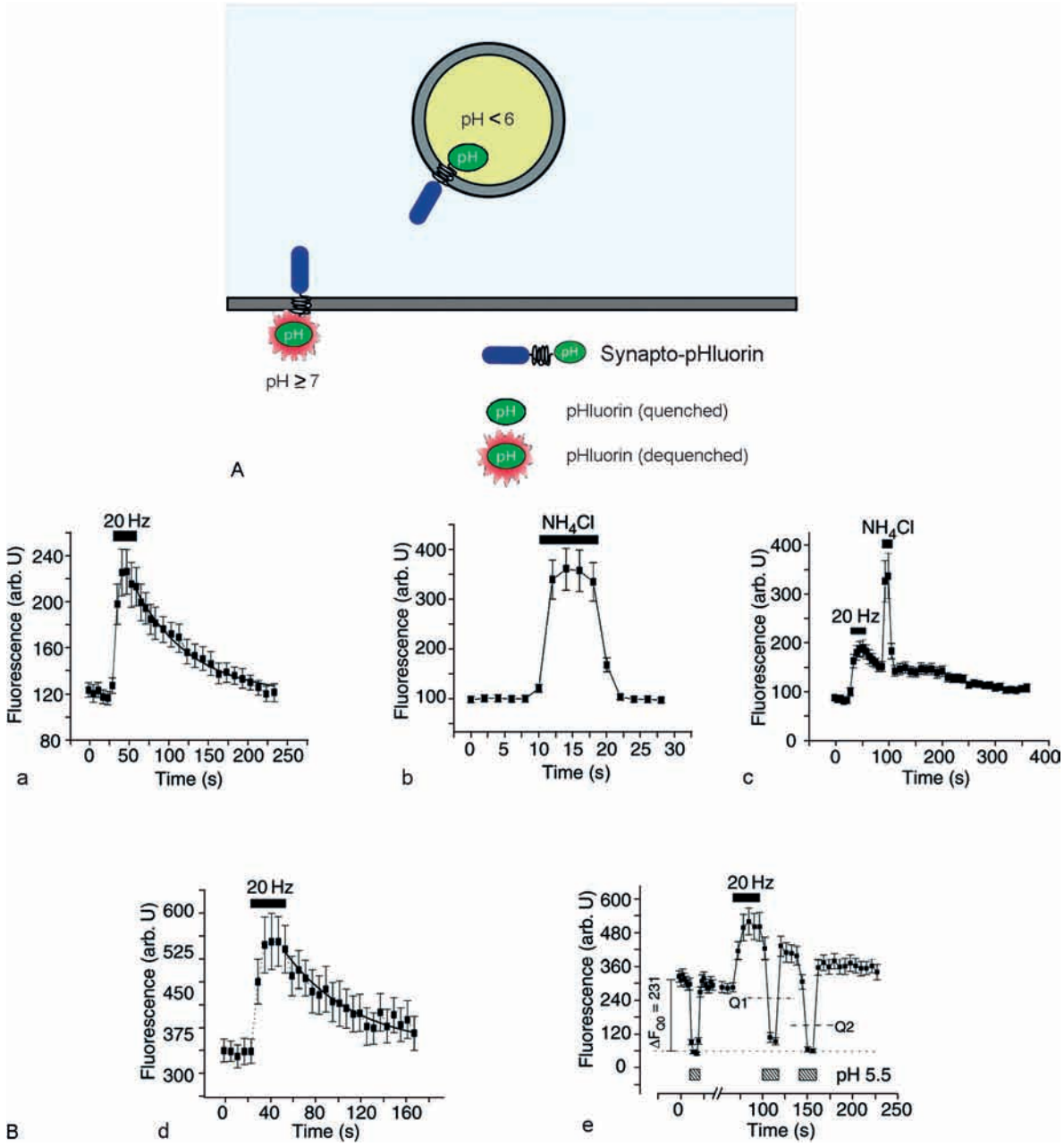

Figure 2 Real-time measurement of SV cycling using pHluorins. Schematic illustration (A) of how synapto-pHluorin can be used to probe SV cycling. Its fluorescence is quenched in the acidic vesicular lumen, but not when residing at the plasmalemma. Synapto-pHluorin signals (B) during firing of action potentials. (a)–(c) synapto-pHluorin (spH) is recycled at boutons. (a) Time course of fluorescence intensity, averaged over 13 boutons expressing spH, following stimulation with a train of 600 action potentials at 20 Hz. The dark bar shows the duration of the stimulus. The decay of fluorescence was fit by a single exponential (solid line) with $\tau = 74$ s. (b) Time course of average spH fluorescence in the same boutons as those used for (a) during alkalinization with NH_4Cl (dark bar). (c) Time course of fluorescence intensity at the same boutons as in (a, b) during train of 600 action potentials followed by exposure to NH_4Cl 30 s after the end of the electrical stimulus. NH_4Cl-induced changes are completely reversible. (d) and (e) Endocytosis, not reacidification, is rate limiting during fluorescence decay. (d) Time course of the fluorescence intensity of spH-positive boutons ($n = 20$) following electrical stimulation (dark bar). Exocytosis of spH causes a rapid increase in fluorescence, followed by a slow decay (solid line is a single exponential fit to the average fluorescence decay, $\tau = 68$ s). (e) Time course of fluorescence intensity during brief exposures to acidic solution (hatched bars below trace) before and after electrical stimulation (dark bar). Exposure to acid during resting periods led to decreases in fluorescence (quenching), indicating the presence of a resistant surface pool of spH. The fluorescence after acid quenches was similar before and after electrical stimulation, indicating that most of the newly endocytosed vesicles were rapidly reacidified. Reprinted by permission from Macmillan Publishers Ltd: *Nature Cell Biology* (Sankaranarayanan S and Ryan TA (2000) Real time measurements of vSNARE recycling in CNS synapses. *Nature Cell Biology* 2: 197–204), copyright 2000.

the presynaptic plasmalemma. It therefore may not be surprising that at least in some experimental systems, such as the neuromuscular junction of frogs and snakes, parts of the presynaptic membrane can be internalized via large vacuolar structures and cisternae, in particular after chemical induction of neurotransmitter release by application of high concentrations of K^+ and calcium. Some of these vacuoles may still exhibit a narrow tubular connection with the plasma membrane and are sometimes seen to contain clathrin-coated buds at their cytoplasmic ends (Figure 1). Whether such cisternal invaginations are eventually consumed by clathrin- and/or dynamin-dependent processes remains unclear. Once having undergone fission from the plasmalemma, cisternae could undergo additional budding steps and thereby constitute a form of a specialized presynaptic endosome. In fact, early endosomal markers including the small GTPase Rab5 and the SNARE protein Vti1aβ are present on SVs. Rab5 mutations in *Drosophila* interfere with efficient release during repetitive stimulation, suggesting that presynaptic endosomes could play an important functional role in maintaining SV pools.

Clathrin-Mediated SV Endocytosis

Clathrin was first purified by Barbara Pearse more than 30 years ago, using coated vesicles isolated from pig brain. In fact, clathrin is most abundantly expressed in the central nervous system, where it is found to be particularly concentrated in presynaptic nerve terminals. The importance of clathrin for SV recycling is further underscored by the fact that clathrin-coated vesicles (CCVs) isolated from nerve terminals are highly enriched in SV proteins. Clathrin, the heterotetrameric adaptor complex (AP-2), and monomeric adaptors and accessory proteins (including epsin, eps15, AP180, HIP1/HIP1R, amphiphysin, endophilin, stonin 2, etc.) play an early role in coat formation. Recruitment of AP-2 to the plasma membrane is a cooperative and presumably highly regulated process involving interactions with phosphoinositides, membrane cargo, and a variety of AP-2α ear domain-binding partners. Many of these adaptor and accessory proteins also display higher expression levels in brain than in other tissues, perhaps owing to their increased half-lives. In addition, neurons contain endocytic protein isoforms, including splice variants of clathrin light chains and α_A-adaptin, AP180, auxillin, intersectin, and dynamin 1. Much of what we know about the mechanism of CCV formation has been learned from nonneuronal systems or from structural studies on clathrin, adaptor, and accessory proteins or domains thereof. In the following sections, we summarize these data and provide a tentative model for how clathrin, dynamin, and their binding partners could act at nerve terminals.

Early Steps of Clathrin-Coated Pit Formation

CCVs are formed by the coordinated assembly of clathrin triskelia built from three tightly linked heavy and associated light chains onto the plasma membrane. The recruitment and polymerization of the outer clathrin layer is assisted by mono- and heterotetrameric adaptor proteins, which simultaneously bind to clathrin, to membrane lipids, and in many cases to transmembrane cargo proteins. In addition, there is a large reservoir of preassembled flat hexagonal clathrin lattices at the plasma membrane that, however, need to undergo a structural transition involving the formation of clathrin pentagons in order to accommodate a curved membrane bud. The most important clathrin adaptor is the heterotetrameric AP-2 complex comprising two large subunits (α and β2), a medium subunit (μ2), and a small subunit (σ2). The two large subunits together with σ2 and the amino-terminal domain of μ2 (N-μ2) form the trunk or core domain of AP-2, and are joined by extended, flexible 'hinges' to the appendage or ear domains of α- and β2-adaptins. Since AP-2 associates with clathrin, a variety of accessory endocytic proteins, phosphatidylinositol 4,5-bisphosphate [PI(4,5)P_2], and membrane cargo proteins, it has been postulated to serve as a main protein interaction hub during coated pit assembly. Many accessory proteins, such as epsins, AP180/CALM, and amphiphysin, also have an adaptor function by linking clathrin assembly to membrane bud formation. These mono- or dimeric adaptors possess a folded lipid-binding domain linked to a more flexible portion of the protein harboring short clathrin- and AP-2-binding motifs, which may aid stabilization of nascent clathrin-coated pits (CCPs) during the assembly process. During CCP assembly transmembrane cargo proteins are recognized by adaptor proteins, most notably the AP-2 complex, which bind to endocytic sorting motifs within their cytoplasmic tails. These motifs include tyrosine-based Yxxø (where ø is a bulky hydrophobic residue) and acidic cluster di-leucine motifs, which bind directly to distinct sites within the AP-2 core domain. Yxxø motifs have been co-crystallized with the carboxy-terminal portion of the AP-2 μ-subunit (C-μ2), to which they bind in an extended conformation. Cargo recognition by AP-2 requires the presence of PI(4,5)P_2, which stabilizes the protein in an open conformation that enables cargo recognition by its μ2-subunit. Consistent with this, clathrin/AP-2-coated pits were

shown to become stabilized in living cells upon encounter of cargo receptors, suggesting that the process of AP-2 recruitment and initiation of plasmalemmal CCPs is highly cooperative.

Despite intense efforts, evidence regarding the presence of canonical Yxxø- or di-leucine-type endocytosis signals within SV protein cytoplasmic domains remains scarce. However, C-μ2 harbors a structurally unresolved binding site for basic internalization motifs found in a variety of multimeric membrane proteins, including the presynaptic vesicle protein synaptotagmin, the presumed calcium sensor in neuroexocytosis. Neuronal synaptotagmin isoforms also interact with the AP-2-binding μ-homology domain containing adaptor protein stonin 2, which is capable of targeting synaptotagmin for clathrin/AP-2-dependent internalization in neurons as well as in transfected fibroblasts. In addition to its interaction with the ear-domain of AP-2α, stonin 2 can bind to eps15 and intersectin, thereby linking synaptotagmin with other components of the clathrin-dependent endocytic machinery in neurons. Stonin 2 (and by analogy, stoned B in *Drosophila*) thus represents the first endocytic adaptor protein identified that is specifically dedicated to the endocytic internalization of a SV protein. Synaptotagmin may thus regulate both the exocytic and endocytic limbs of the SV cycle. In support of this hypothesis, it has been observed that genetic or chemical perturbation of synaptotagmin function by fluorophore-assisted light inactivation in mice, flies, or worms results in recycling defects and a partial depletion of SVs. Whether other SV proteins also interact with specific endocytic adaptor proteins or get co-sorted with synaptotagmin, that is, as part of a membrane microdomain, remains an open question.

CCP Maturation and Vesicle Fission

After the clathrin lattice is formed, endophilin, epsin, eps15, amphiphysins, and other proteins are involved in membrane bending and clathrin rearrangements as coated pits progressively invaginate and mature. Partitioning of the amino-terminal amphipathic helix of the ENTH domain protein epsin and perhaps other components (i.e., the small GTPase Arf6) drives the acquisition of membrane curvature. Bin–amphiphysin–Rvs (BAR) domain proteins, such as amphiphysin and endophilin, may aid membrane bending, and function as curvature sensors that signal completion of the process. Through their SH3 domains, both amphiphysin and endophilin also interact with and recruit accessory enzymes, including the large GTPase dynamin and the phosphoinositide phosphatase synaptojanin, to the nascent vesicular bud. Dynamin is required for fission of endocytic membrane vesicles by mechanochemically constricting ('pinchase') the vesicle neck. The eminent role of dynamin for SV recycling is best illustrated by the dramatic phenotype seen in *shibire*^{ts} mutants in *Drosophila*, which exhibit temperature-sensitive paralysis due to the accumulation of unbudded membrane infoldings and endocytic intermediates.

Observation of CCPs dynamics using evanescent wave microscopy indicates that during fission, dynamin recruitment to coated pits is rapidly followed by recruitment of actin. Moreover, perturbation of actin disrupts the endocytic reaction with accumulation of coated pits with wide necks, suggesting a role of actin, actin-binding factors, and dynamin-interacting accessory proteins, such as Abp1 or syndapin, in promoting constriction of the neck and removal of endocytosed vesicles from the membrane. In lamprey, snake, and fly neuromuscular synapse the invagination of the membrane into pits, occurs at distinct 'endocytic zones' surrounding the active zones of exocytosis (termed the peri-active zone). FM1–43 photoconversion and serial section electron microscopy analysis revealed that labeled clathrin-coated endocytic vesicles were clustered significantly near active zones, consistent with local exocytic–endocytic recycling vesicle pools at this synapse. Together with the regulated turnover and synthesis of membrane phosphoinositides, actin and actin-binding proteins may thus provide spatial and temporal landmarks for SV endocytosis (see the next section).

Protein Scaffolds as Spatial Regulators of Vesicle Cycling

Morphologically and functionally, the active zone can be divided into two parts: the core active zone, where regulated exocytosis (and kiss-and-run-type SV retrieval) takes place; and the peri-active zone, where clathrin-mediated endocytosis occurs (see the preceding section). At the ultrastructural level, the core active zone is characterized by the more or less regular array of electron-dense material, called the presynaptic grid, presynaptic particle web, or 'cytomatrix assembled at the active zone' (CAZ). During recent years, multiple molecular components – both CAZ-specific ones and those that are recruited through interaction with CAZ scaffolding proteins – have been identified and characterized.

Functionally, the CAZ is thought to define the site of regulated neurotransmitter release by localizing presynaptic membrane proteins, including voltage-gated calcium channels and cell adhesion molecules, to organize steps of the SV cycle, including tethering and priming of SV, and to link the exocytic machinery

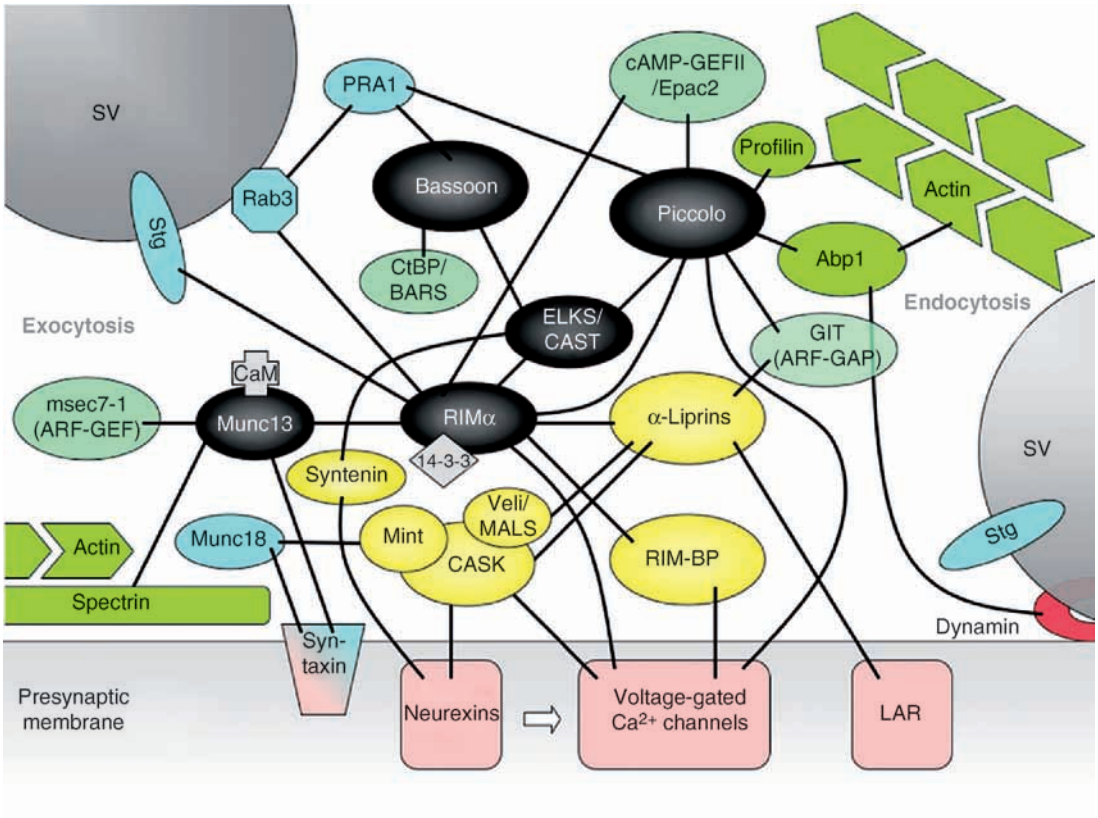

Figure 3 Molecular organization of the cytomatrix at the active zone (CAZ). The scheme depicts observed physical interactions between active zone-specific scaffolding proteins (black), associated proteins with putative structural functions (yellow), effector proteins (blue-green), actin cytoskeletal and associated elements (green), small modulatory molecules (gray), proteins involved in SV exocytosis (blue) and endocytosis (red), as well as presynaptic membrane proteins (pink). Some of the interactions for Piccolo and RIM were discovered in pancreatic β-cells and will have to be confirmed for the CAZ. For further details see **Table 2**. Note: the diagram neither reflects the relative sizes of the proteins nor their exact topographic localization within the presynaptic bouton. The arrow indicates that neurexins (α-forms) are involved in the localization of calcium channels.

with elements of the endocytic zone and with the surrounding actin cytoskeleton (Figure 3). In addition, CAZ elements turned out to be essential mediators of presynaptic plasticity.

Molecular Organization of the CAZ

Relatively few CAZ-specific structural and effector proteins have been identified to date that are believed to constitute the scaffold of the CAZ and to mediate its structural and functional organization (black in Figure 3). These proteins belong to four different protein families: the Rab3-interacting molecules (RIMs), the mammalian Unc13 proteins (Munc13s), the two related giant CAZ scaffolding proteins Bassoon and Piccolo, and the ELKS/CAST proteins (Figure 3; Table 2).

RIMs are multidomain proteins that were identified as effectors of Rab3, a small GTPase associated with SVs. In particular, the α-forms of RIM1 and RIM2 are important scaffolding molecules that interact with multiple other presynaptic proteins. These include isoforms of Munc13, a liaison that might be involved in making SVs fusion-competent, as well as ELKS/CAST and Piccolo. Analysis of RIM1α-deficient mouse mutants revealed that this protein is involved in short- and long-term forms of presynaptic plasticity. For example, long-term potentiation at hippocampal mossy fiber terminals or parallel fiber synapses of the cerebellum involve protein kinase A-dependent phosphorylation of RIM1α and phosphorylation-dependent binding of 14-3-3 proteins. Further interactions of αRIMs point to a central role of these proteins in active zone organization. They can bind presynaptic voltage-gated Ca^{2+} channels, either directly or via their binding to RIM-binding proteins, and the Ca^{2+} sensor synaptotagmin, which is involved in both exocytosis and endocytosis of SVs. The interaction with α-liprins, originally identified as a cytoplasmic adaptor for the receptor tyrosine phosphatase LAR, might serve the formation and maintenance of active zones, as suggested by work in invertebrates.

Munc13 isoforms are involved in SV priming and the regulation of synaptic plasticity. They can link

Table 2 Protein components of the cytomatrix assembled at the active zone

Protein family *Relevant members and synonymous names*	*Domains and motifs*	*Interacting protein*	*Proposed function(s)*
RIMs – Rab3-interacting molecules	Zn finger	Rab3	Link to SV, tethering of SV, presynaptic plasticity
		Munc13-1, ubMunc13-2	SV priming
RIM1α; RIM2α,β,γ; RIM3γ; RIM4γ; UNC10 (*C. elegans*)	Region between Zn finger and	14-3-3	Presynaptic plasticity
	PDZ domain	cAMP-GEFII / Epac2	Insulin secretion in pancreatic β-cells
	PDZ domain	ELKS / CAST	CAZ scaffolding
	C2A	Piccolo	CAZ scaffolding
		N-type calcium channels	Channel anchoring and/or clustering
	Pro-rich motif btw. C2A & B	RIM-BPs	CAZ scaffolding
	C2B	α-Liprin,	CAZ scaffolding,
		N-type calcium channels	Anchoring and/or clustering of Ca channels
		synaptotagmin	CAZ scaffolding, organization of Ca sensing
RIM-binding proteins	SH3 domains 2 and/or 3	RIMs	CAZ Scaffolding
RIM-BP1		N-type ($Ca_v2.2$) and L-type ($Ca_v1.3$) Ca channels	Channel anchoring and/or clustering
RIM-BP2			
Piccolo–Bassoon family	Zn fingers	PRA1 (prenylated Rab3 acceptor)	Link to SV?
Bassoon and Piccolo	Between CC1 and CC2	CtBP/BARS	Membrane trafficking?
		Ribeye/CtBP2	Scaffolding of synaptic ribbons
	CC3	ELKS / CAST	CAZ scaffolding
Piccolo/Aczonin only	N-terminal Q domain	Abp1 (actin-binding protein 1)	Link to actin cytoskeleton, endocytosis
	Between CC1 and CC2	GIT (ARF-GAP)	Membrane trafficking
	Pro-rich region btw CC1 and 2	Profilin	Actin regulation
	PDZ domain	cAMP-GEFII/Epac2	Insulin secretion in pancreatic β-cells
	C2A	RIMs	CAZ scaffolding, organization of SV cycle
		Piccolo	Homophilic interaction, scaffolding
	C2A and C2B	L-type Ca channel	Channel anchoring in β-cells

Continued

Table 2 Continued

Protein family *Relevant members and synonymous names*	*Domains and motifs*	*Interacting protein*	*Proposed function(s)*
UNC13 proteins	M13-1 and ubM13-2 N-term.	α-RIMs	CAZ scaffolding,
Munc13-1		Calmodulin	Ca^{2+}-dependent plasticity
bMunc13-2	Conserved homology region (MUN)	Spectrin β-spIIIΣ	Anchoring to actin/spectrin cytoskeleton
ubMunc13-2		msec7-1 ARF-GEF	Regulation of actin cytoskeleton
Munc13-3		DOC2α (double C2 domain protein)	?
Munc13-4			
UNC13 (C.e.)	C-terminus	Syntaxin (in debate)	SV priming, control of SNARE complex formation
Dunc13 (D.m.)			
ELKS/CAST proteins	CC regions	Bassoon, Piccolo	CAZ scaffolding
ELKS1A/ERC1A		α-Liprins	CAZ scaffolding
ELKS1B/ERC1B	C-terminus	RIMs	CAZ scaffolding
ELKS2/CAST/ERC2		Syntenin	Active zone organization?
α-Type Liprins	N-term CC region	RIMs	CAZ scaffolding
Liprin-α1–α4		ELKS/CAST	CAZ scaffolding
SYD-2 (C.e.)		GIT (ARF-GEF)	Regulation of membrane trafficking and actin cytoskeleton
		Kif1A (kinesin motor)	Transport
	SAM domains	LAR receptor protein tyrosine Phosphatase	Regulation of membrane protein anchoring
		β-Liprin	
		CASK (MAGuK)	Regulation of transsynaptic adhesion?
		MALS/Veli	Regulation of transsynaptic adhesion?
	C-terminal PDZ binding site	GRIP	Receptor transport and clustering/postsynaptic

to the actin-spectrin cytoskeleton via binding to a β-spectrin isoform and to presynaptic membrane-trafficking processes via the ARF guanine nucleotide exchange factor msec7-1.

ELKS/CAST proteins display a very high content of coiled-coil structures and are considered as a major structural component of the CAZ. They can physically interact with RIMs as well as with Bassoon and Piccolo and thus might interconnect the major scaffolding proteins of the CAZ. Bruchpilot, an ELKS/CAST-related protein in *Drosophila*, is responsible for the proper anchoring of particular specializations of the CAZ, so-called T-bars, to the active zone membrane. Via syntenin, ELKS/CAST proteins also connect to neurexins, which are specific cell adhesion molecules of the presynaptic membrane occurring as long α- and short β-forms involved in Ca^{2+} channel localization and linkage to postsynaptic neuroligins, respectively. Neurexins are additionally anchored to the presynaptic cytomatrix via a trimeric protein complex of CASK, Mint, and MALS/Veli, which in turn links to Ca^{2+} channels and α-liprins (Figure 3).

Bassoon and Piccolo are considered as very large multidomain scaffolding molecules of the CAZ. Most of their interaction partners have still to be discovered. They both bind ELKS/CAST and the small prenylated Rab3 acceptor (PRA1), which potentially links into the SV cycle. In addition, Bassoon has been reported to interact with CtBP/BARS and its homolog RIBEYE, a specific component of synaptic ribbons in retinal photoreceptors and inner ear hair cells. The RIBEYE–Bassoon interaction was suggested to be essential for anchoring of ribbons to the presynaptic plasmalemma. The interaction of Bassoon (and potentially also Piccolo) with CtBP/BARS is of interest, as this protein has been implicated in vesicular fission from the *trans*-Golgi complex. By analogy, a similar function might be envisioned in the presynapse.

Links of the CAZ to the Endocytic Machinery and the Actin-Based Cytoskeleton

While multiple interactions of the previously described CAZ proteins underscore the role of the presynaptic cytomatrix in organizing the apparatus for regulated exocytosis, a link to clathrin-mediated endocytosis is less clear. Specific interaction partners for Piccolo might be of particular interest in this context, as they can provide the basis for the physical linkage of exocytic and endocytic presynaptic processes. On the one hand, Piccolo can bind directly to Ca^{2+} channels and the guanine nucleotide exchange factor cAMP-GEFII/Epac2, as revealed from studies on pancreatic β-cells. Moreover, the C2A domain of Piccolo is discussed as a candidate low-affinity Ca^{2+} sensor. These characteristics argue for a role in active zone organization and exocytosis. On the other hand, Piccolo can bind to the ARF-GTPase-activating protein GIT, which has been implicated in endocytic processes such as receptor internalization. The N-terminus of Piccolo specifically interacts with Abp1, an actin-binding factor directly regulating the GTPase dynamin. Indeed, the N-terminal Q-domain of Piccolo can interfere with endocytic processes in heterologous systems. Yet another link between Piccolo and the actin cytoskeleton is profilin, a small G-actin- and phosphoinositide-binding protein that is involved in local remodeling of the actin cytoskeleton. Thus, Piccolo might indeed be an important mediator between the neurotransmitter release apparatus and the neighboring endocytic machinery.

Regulation of the SV Cycle by Membrane Lipids

Phosphoinositide Regulation of SV Cycling

Cycling of presynaptic vesicles requires the precise spatial and temporal regulation of protein–lipid interactions. Many synaptic proteins – including synaptotagmin1, calcium-dependent activator protein for secretion (CAPS), the Munc18-interacting proteins Mint-1 and Mint-2, voltage-gated P/Q-type calcium channels, and a variety of endocytic proteins such as AP-2, AP180, epsin, and dynamin (Table 1) – directly bind to and are regulated by [PI(4,5)P_2]. SV cycling thus appears to be nested into a local cycle of phosphoinositide phosphorylation and hydrolysis. Accordingly, PI(4,5)P_2 acts at multiple stages of the vesicle cycle. Knockout mice lacking PIP kinase type Iγ, the major PI(4,5)P_2-synthesizing enzyme at synapses, display defects in neurotransmitter release and endocytic recycling that cause synaptic depression. The activity of PIPK Iγ is regulated by a variety of factors, including phosphatidic acid (PA), the small GTPases Arf6 and Rac1, and the actin cytoskeleton-associated adhesion protein talin. Association of PIPK Iγ with these factors is dependent on their phosphorylation status, providing a means for the temporal and spatial regulation of phosphoinositide metabolism. PI(4,5)P_2 is eventually consumed by synaptojanin-mediated dephosphorylation, resulting either in formation of PI(4)P and perhaps PI, cleavage via phospholipase C, or PI3 kinase-dependent synthesis of PI(3,4,5)P_3. At least in neuroendocrine PC12 cells, PI(4,5)P_2 appears to be concentrated within cholesterol-enriched microdomains near release sites, where it may aid vesicle docking and/or fusion. The recent observation that SV proteins remain clustered during their exocytic–endocytic journey, together

with the extremely high cholesterol content of SV membranes, suggests that cholesterol-enriched microdomains could serve to spatially organize the SV cycle, perhaps in part by locally concentrating PI(4,5)P_2.

At present we can only speculate about the exact mechanism of action of PI(4,5)P_2 during vesicle fusion, but its role in clathrin-mediated synaptic vesicle endocytosis is much better understood. PI(4,5)P_2 is an important factor in recruiting endocytic adaptor and accessory proteins to the membrane where these form a network of protein–protein interactions. The stability of this network critically depends on the PI(4,5)P_2 content of the membrane, as suggested by the observation that CCVs accumulate in nerve terminals of synaptojanin knockout mice. These observations also indicate that PI(4,5)P_2-hydrolysis may normally occur concomitantly with or directly after dynamin-mediated membrane fission. CCVs at presynaptic sites of synaptojanin knockout mice become trapped in a meshwork of actin filaments, consistent with the fact that PI(4,5)P_2 regulates actin polymerization and drives the formation of actin comet tails that may help to propel endocytic vesicles away from the plasmalemma.

Lipids and Membrane Deformation

SV fusion and the subsequent formation of endocytic clathrin-coated buds at the presynaptic peri-active zone involve radical geometric remodeling of the membrane in order to generate areas of different membrane curvature. While lipids with bulky polar headgroups and saturated or single fatty acid tails, including lysolipids and many glycolipids, promote positive curvature, lipids with compact headgroups and space-filling hydrophobic tails, such as PA and diacylglycerol (DAG), favor negative curvature. For example, lysophosphatidic acid (LPA) and PA, which are interconverted by LPA-acyl transferase and phospholipase A_2 activities, respectively, favor opposite curvatures.

Although differential distribution of distinct types of lipids between the two membrane leaflets may contribute to membrane deformation, it is generally assumed that membrane bending requires the action of proteins. During endocytic SV recycling, the forming bud must adopt a positive curvature at the bud center and negative curvatures at the edges. Endocytic proteins may act by one of several mechanisms to bend membranes. As stated previously, epsin, a PI-(4,5)P_2-binding clathrin accesssory protein, is able to partition into the cytoplasmic leaflet of the plasma membrane via PI(4,5)P_2-induced formation of an extra α-helical segment that drives acquisition of positive curvature. Dimeric BAR domain-containing proteins, including amphiphysin or endophilin (Table 1), induce curved membranes by an additional amphipathic helix at their amino-terminal end, largely via electrostatic interactions of their concave surface with negatively charged membrane phospholipids. Endophilin has originally been proposed to be an LPA-acyl transferase, but this activity has recently been called into question. Highly curved membranes may become stabilized in addition by other scaffolding proteins, including clathrin itself, which forms a rigid basket around the emanating vesicular membrane bud. Moreover, the transmembrane domains of synaptic vesicle proteins could provide a barrier that prevents local areas of high curvature from lateral diffusion and thereby contribute to maintaining vesicle identity. Finally, protein–lipid interactions and the formation of microdomains might also underlie the choice between fast and slow modes of SV cycling that would require a tight control of fusion pore expansion and constriction.

See also: Activity-Dependent Remodeling of Presynaptic Boutons; Clathrin and Clathrin-Adaptors; Dynamin; Endocytic Traffic in Spines; Endocytosis: Kiss and Run; Liprins, ELKS, and RIM-BP Proteins; Munc13 and Associated Molecules; Piccolo and Bassoon; Presynaptic Endosomes; Rab3A Interacting Molecules (RIMs); Synaptic Vesicles; Vesicle Pools.

Further Reading

Dresbach T, Qualmann B, Kessels MM, et al. (2001) The presynaptic cytomatrix of brain synapses. *Cellular and Molecular Life Sciences* 58: 94–116.

Galli T and Haucke V (2004) Cycling of synaptic vesicles: How far? How fast! *Science Signal Transduction Knowledge Environment* 264: re19.

Gundelfinger ED, Kessels MM, and Qualmann B (2003) Temporal and spatial coordination of exocytosis and endocytosis. *Nature Reviews Molecular Cell Biology* 4: 127–139.

Murthy VN and De Camilli P (2003) Cell biology of the presynaptic terminal. *Annual Review of Neuroscience* 26: 701–728.

Royle SJ and Lagnado L (2003) Endocytosis at the synaptic terminal. *Journal of Physiolology* 553: 345–355.

Shankaranarayanan S and Ryan TA (2000) Real time measurements of vSNARE recycling in CNS synapses. *Nature Cell Biology* 2: 197–204.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Ziv NE and Garner CC (2004) Cellular and molecular mechanisms of presynaptic assembly. *Nature Reviews Neuroscience* 5: 385–399.

Endocytosis: Kiss and Run

E T Kavalali, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Multiple Pathways for Synaptic Vesicle Recycling

The synaptic vesicle cycle is initiated by a rise in intrasynaptic Ca^{2+} leading to vesicle fusion and neurotransmitter release at the active zone. According to the conventional view of this process, upon fusion, synaptic vesicles completely collapse onto the plasma membrane, and vesicle membrane components (lipids as well as proteins) intermix with their plasma membrane counterparts. Subsequently, clathrin, through its adaptor proteins, is recruited to the membrane (typically within the periphery of the active zone) and forms coated vesicles by reclustering vesicle membrane components which eventually bud off the plasma membrane with the help of the GTPase dynamin. In addition to the plasma membrane, endocytosis of clathrin-coated vesicles may also occur from membrane infoldings or endosomal cisternae, which form upon accumulation of fused synaptic vesicles. A V-type ATPase then lowers the pH in these vesicles and subsequently the vesicles are refilled with neurotransmitter. The reacidification and neurotransmitter filling steps are thought to be rapid, and the entire process for exocytosed vesicles to become reavailable for release occurs within 40–90 s (Figure 1).

Kiss-and-run refers to an additional fast pathway of vesicle recycling during which synaptic vesicles retain their identity and do not intermix with the plasma membrane or endosomal compartments. During this process, the fusion pore opens transiently without complete collapse of the vesicle and without intermixing of synaptic vesicle membrane components with the plasma membrane. In several models of cellular secretion, the vesicle fusion process is easier to access technically through direct visualization or via membrane capacitance measurements. In these systems, there are several examples of transient vesicle fusion events that resemble what would be expected from kiss-and-run fusion. In central synapses, however, limited experimental access to the fusion process restricts our ability to assess the preponderance of kiss-and-run-type fusion. Nevertheless, studies have used elaborate experimental settings to address this question.

In addition to limited morphological and functional characterization of the kiss-and-run process, the actual molecular mechanism of this pathway remains to be determined. There is evidence that this rapid component of endocytosis may operate independent of clathrin and may require the function of the synaptic vesicle protein synaptobrevin, which is also critical for rapid Ca^{2+}-dependent fusion. However, it is a matter of debate whether kiss-and-run represents a reversible SNARE-mediated fusion event or almost instantaneous fission of vesicle–plasma membrane interface upon fully executed SNARE-mediated fusion.

Clathrin-mediated endocytosis comprises a ubiquitous means for vesicle recycling in most cell types. This pathway typically possesses well-defined morphological markers (coated pits, endosomal intermediates, etc.), and adequate molecular tools are available to probe its properties in synapses (biochemical purification of coated vesicles and isolation of endocytic intermediates). The kiss-and-run pathway, in contrast, may not employ the same molecular players and structural intermediates and is more difficult to examine morphologically and molecularly. Therefore, most evidence in support of a kiss-and-run-type fast retrieval and recycling mechanism for synaptic vesicles relies on electrophysiological and optical techniques with rapid time resolution.

Cell Biological Implications of Kiss-and-Run Endocytosis

The proposal for kiss-and-run-type endocytosis emerged soon after the vesicular hypothesis for neurotransmission had been widely accepted. In the early 1970s, Bruno Ceccarelli and colleagues performed elegant experiments on the neuromuscular junction using a combination of electrophysiology and electron microscopy to examine the synaptic vesicle recycling process. The term 'kiss-and-run,' however, was introduced by Meldolesi and colleagues in 1994. A key feature of early experiments was morphological preservation of synaptic vesicle curvature during fusion events. In addition, there was limited or negligible depletion of synaptic vesicles during sustained stimulation, which released substantial neurotransmitter quanta exceeding the number of vesicles available in the terminal. Furthermore, horseradish peroxidase or dextran labeling during stimulation revealed that several vesicles are formed from the same axolemmal membrane where fusion occurs. This proposal was quite attractive because it introduced an economical means to retrieve synaptic vesicles upon fusion rapidly and with high fidelity, which would match the rapid pace of synaptic transmission compared to other constitutive vesicle

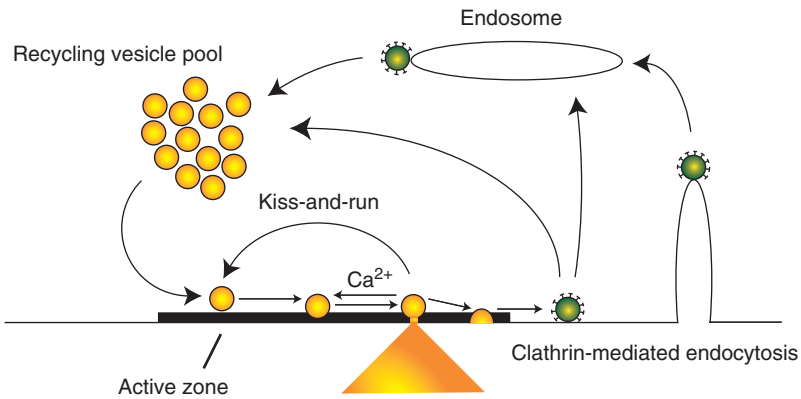

Figure 1 The synaptic vesicle cycle. During neurotransmission, a rise in Ca^{2+} level leads to vesicle fusion and neurotransmitter release at the active zone, after which vesicles completely collapse onto the plasma membrane. During this process, vesicle membrane components are thought to intermix with their plasma membrane counterparts. Subsequently, clathrin, through its adaptor proteins, is recruited to the membrane (typically at the periphery of the active zone) and forms coated vesicles by reclustering vesicle membrane components which eventually bud off the plasma membrane with the help of the GTPase dynamin. In addition to the plasma membrane, endocytosis of clathrin-coated vesicles may also occur from membrane infoldings or endosomal cisternae. This classical pathway is thought to operate on a 40–90 s timescale. During an alternative fast (within seconds) pathway of vesicle recycling, synaptic vesicles retain their identity and do not intermix with the plasma membrane or endosomal compartments. This pathway is also referred to as the 'kiss-and-run' pathway, in which the fusion pore opens transiently without complete collapse of the vesicle and without intermixing of synaptic vesicle membrane components with the plasma membrane.

trafficking events. These observations were also consistent with the later identification and exhaustive characterization of synaptic vesicles as distinct organelles in terms of their membrane lipid and protein composition relative to the plasma membrane. If synaptic vesicle components could freely diffuse on the plasma membrane upon fusion, it would be a daunting task to retrieve all the necessary components in sufficient and adequate quantities and reassemble a synaptic vesicle. Furthermore, most synaptic vesicle proteins do not contain distinct molecular tags that would dictate their targeting to synaptic vesicles. Therefore, *de novo* assembly of synaptic vesicles following fusion is difficult to envision in the limited time frame required to sustain neurotransmission and prevent disfiguration of synaptic membranes (Figure 2).

Studies using fluorescently tagged synaptic vesicle proteins, however, challenged this notion and showed that at least the synaptic vesicle proteins synaptobrevin and synaptotagmin can diffuse from the fused vesicle membrane and endocytosis involves recruitment of a distinct set of the same proteins previously resident on the plasma membrane. Although most studies to date have tested synaptobrevin or synaptotagmin in this context, recent work has examined if this observation is applicable to other synaptic vesicle proteins. In these experiments, different synaptic vesicle proteins gave different results with respect to their background fluorescence (i.e., surface exposure) and retrieval. For instance, both synaptophysin (a four transmembrane domain synaptic vesicle protein) and the vesicular glutamate transporter show

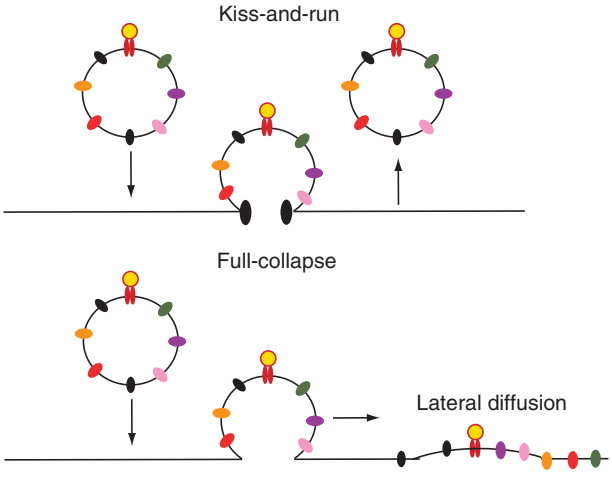

Figure 2 Preservation of vesicle identity upon kiss-and-run endocytosis. Synaptic vesicles are distinct organelles in terms of their membrane lipid and protein composition from the plasma membrane. According to the kiss-and-run model, synaptic vesicles retain their composition upon fusion and are retrieved intact. The alternative view suggests that synaptic vesicle components can freely diffuse onto the plasma membrane upon full collapse of the vesicle membrane during fusion. However, most synaptic vesicle proteins do not contain distinct molecular tags that would dictate their targeting to synaptic vesicles. Therefore, in this alternative model it is difficult to envision how all the necessary components can be retrieved in sufficient quantities to reassemble a synaptic vesicle.

very little surface exposure (thus fluorescence) in the absence of stimulation. Following stimulation, synaptophysin and the vesicular glutamate transporter show a rapid increase in fluorescence without significant

lateral diffusion upon fusion, which was a salient feature of earlier synaptobrevin-based measurements. The dependence of the properties of these GFP-based measurements on the identity of the tagged protein complicates a straightforward interpretation of the earlier synaptobrevin-based experiments in terms of synaptic vesicle recycling.

Since the early 1990s, fluorescence probes such as styryl dyes and pH-sensitive GFP-tagged synaptic vesicle proteins have revealed that synaptic vesicle endocytosis occurs within seconds in a manner tightly coupled to synaptic vesicle fusion. Classical clathrin-mediated endocytosis mechanisms operate on a slower timescale. In cellular imaging experiments in which capture of a single cargo molecule is monitored, the time course of clathrin coat assembly and mean residence time on the plasma membrane typically varies between 30 s and 1 min (mean lifetime of 46 s). The rapid and high-fidelity retrieval of a complex organelle such as a synaptic vesicle results in significant constraints to the classical clathrin-mediated endocytosis machinery, especially if *de novo* assembly of synaptic vesicles is involved in the process. However, the rapidity of endocytosis cannot be considered as proof of kiss-and-run endocytosis because clathrin-mediated endocytosis may be modified to adjust to the rapidity of the synaptic vesicle recycling. According to one hypothesis, synaptic vesicles may be preassembled at the membrane in a 'readily retrievable pool' through a clathrin-mediated process and endocytosed rapidly upon membrane fusion. During this process, patches of endocytosed membrane and fused membrane are not identical. This premise requires a means of communication between the spatially separate exocytosis and endocytosis events. One possibility for this messenger is the rise in intrasynaptic Ca^{2+} that triggers exocytosis. However, the exact role of Ca^{2+} in synaptic vesicle endocytosis remains to be tested. It is unclear if Ca^{2+} is absolutely required for endocytosis or if it is merely regulatory. If Ca^{2+} turns out to be essential for synaptic vesicle retrieval, then it would be critical to determine if it is permissive (it merely needs to be present) or instructive (it is required) for this process.

The Mechanism of Kiss-and-Run

In addition to limited morphological and functional characterization of the kiss-and-run process, the actual molecular mechanism of this pathway remains to be determined. As indicated previously, kiss-and-run endocytosis may not require clathrin and may involve the synaptic vesicle protein synaptobrevin, which is a critical component of the fusion machinery. However, it is unknown whether kiss-and-run represents a reversible SNARE-mediated fusion event.

The interaction of synaptic vesicle protein synaptobrevin 2/vesicle-associated membrane protein-2 (VAMP-2) with the plasma membrane proteins syntaxin 1 and SNAP-25 (25 k Da synaptosomal-associated protein) is an essential step for vesicle fusion. These proteins are collectively called SNAREs. During fusion, synaptobrevin 2, syntaxin 1, and SNAP-25 form an exceptionally stable complex that is resistant to treatment with the detergent sodium dodecyl sulfate, which would normally be sufficient to dissociate other macromolecular complexes. Under physiological conditions, this tight complex, referred to as the 'core complex,' can only be disassembled by the ATPase *N*-ethylmaleimide-sensitive factor together with SNAPs. Biochemical stability and molecular structure of the core complex suggest that formation of the complex can be sufficient to overcome the energy barrier to membrane fusion by bringing vesicular and target membranes into nanometer proximity or even driving the fusion reaction. In this scheme, the association of synaptobrevin with corresponding syntaxin and SNAP-25 supplies the energy required for the fusion reaction in a process accordingly termed 'zippering.' Because SNARE complex formation is an energetically highly favorable reaction, it is difficult to envision how this reaction can be reversible. Nevertheless, kiss-and-run may still be due to rapid reversibility of the fusion reaction mediated after SNARE complex disassembly presumably requiring rapid action of the GTPase dynamin for fission of vesicles from the plasma membrane. In this way, vesicles can be endocytosed at the spot where they have fused without losing their curvature. This reversible fusion reaction may also involve regulation of core complex disassembly, providing a switch between complete fusion and fusion pore flickering.

The Impact of Kiss-and-Run Endocytosis on Neurotransmission

In addition to preservation of vesicle identity and composition upon fusion and retrieval, the kiss-and-run type of synaptic vesicle recycling offers additional avenues for regulation of neurotransmission. First, rapid retrieval of synaptic vesicles from the plasma membrane without significant loss of their protein composition can make them quickly reavailable for fusion during stimulation. During sustained stimulation, neurotransmission can be maintained either by constant supply of vesicles from a reserve pool or by rapid recycling of fused vesicles. Indeed, most endocytic protein mutants as exemplified by dominant negative or null mutants of dynamin (shibire mutant in *Drosophila* or mouse knockout) manifest an extremely rapid synaptic depression phenotype

consistent with this premise. Furthermore, in hippocampal synapses, inhibition of vesicle reacidification (using inhibitors of vacuolar ATPase) leads to a selective increase in the rate of synaptic depression without significantly impairing other parameters of neurotransmitter release. Therefore, rapid synaptic vesicle recycling, possibly mediated by a kiss-and-run-like mechanism, may potently regulate the kinetics of short-term synaptic plasticity. Furthermore, the rate of recycling can be a critical determinant of a synapse's release rate at steady state during maintained stimulation once the releasable vesicles are depleted and neurotransmission solely relies on vesicle reuse. The second aspect of kiss-and-run endocytosis is its potential ability to regulate fusion pore openings to limit diffusion of neurotransmitter substances to the synaptic cleft. For instance, limited diffusion of neurotransmitter molecules can be clearly detected in carbon fiber measurements of catecholamine release from adrenal chromaffin cells. In central synapses, there is evidence that similar fusion pore-restricted release of glutamate from vesicles may only activate high-affinity NMDA receptors without evoking currents through low-affinity AMPA receptors.

Evidence for Kiss-and-Run Endocytosis during Synaptic Transmission

The prevalence of kiss-and-run endocytosis in central synapses is still an open question. This section discusses the evidence for kiss-and-run from five distinct modalities of measurement, commonly applied to the study of neurotransmission in central synapses.

Retention of Styryl Dyes upon Fusion

FM1-43 is a partially lipophilic dye, which fluorescence increases 100-fold when it partitions into membranes and thus it is virtually invisible in aqueous solution. The chemical structure of FM1-43 makes it difficult to cross lipid membranes, thus limiting the fluorescence labeling to outer membranes and recycling vesicles. Using this methodology, one can detect activity-dependent vesicular dye uptake and release in synaptic connections in multiple preparations including central neuronal cultures. Studies using FM dyes have revealed that activity-induced dye uptake and release operates with high fidelity. Specifically, the extent of fluorescent dye endocytosed during a particular stimulation matches the magnitude of fluorescence loss due to exocytosis in response to the same stimulation. This observation suggests that synaptic vesicles recycle without intermixing with endosomal compartments and thus the fluorescence probes do not get diluted.

The properties of styryl dyes also make them useful tools to reveal quantitative information on vesicle fusion within a presynaptic terminal. These dyes can be rapidly inserted into membranes (~3 ms); however, with the increasing lengths of their hydrocarbon chains, they have decreasing rates of dissociation from membranes. Measurements have shown that FM1-43 departitions from lipid to aqueous environment with a time constant of 2.5 s, whereas its analogs, FM2-10 and FM1-84, departition with time constants of 0.6 and 4.8 s, respectively. Since the introduction of FM dyes, several studies have reported partial release of FM1-43 upon synaptic vesicle fusion during multiple forms of stimulation. In some studies, destaining rates of synaptic boutons stained with FM2-10, FM1-43, and FM1-84 were slower in proportion to the increasing time constants of departitioning. These initial observations were later extended by experiments examining FM dye loss from single vesicle fusion events. The common interpretation of partial FM dye release is the presence of a fast endocytotic mechanism, which preferentially traps the slowly departitioning dyes, FM1-43 and FM1-84, in synaptic vesicles. Partial dye release is only possible if lateral diffusion of FM dye following the formation of the vesicle–plasma membrane interface is delimited by a barrier (presumably a protein-based fusion pore) between the two membranes. However, in other systems such as goldfish bipolar cells, all FM dyes have been reported to laterally diffuse out of fused vesicles, thus losing fluorescence irrespective of their departitioning rates in aqueous solution (Figure 3).

Rapid Changes in SynaptopHluorin Fluorescence

SynaptopHluorin (spH) is another fluorescent probe that is commonly used to monitor exocytosis and endocytosis of synaptic vesicles. spH is a fusion construct of the synaptic vesicle protein synaptobrevin (VAMP) with a pH-sensitive enhanced green fluorescent protein (EGFP) at its C-terminal (located in the synaptic vesicle lumen). The synaptic vesicle lumen normally has an acidic pH of approximately 5.5 at which spH fluorescence is quenched. When vesicles fuse, lumenal EGFP is exposed to the extracellular pH, resulting in a marked increase in its fluorescence. Studies using spH have provided valuable information on the tight coupling between exocytosis and endocytosis. One advantage of the spH probe is the molecular specificity of the fluorescence signal since the signal originates from a synaptic vesicle protein as opposed to nonspecific lipid-based fluorescence of FM dyes. Proton diffusion from a synaptic vesicle is not expected to be limited, and the time course of vesicle reacidification is thought to be fast without

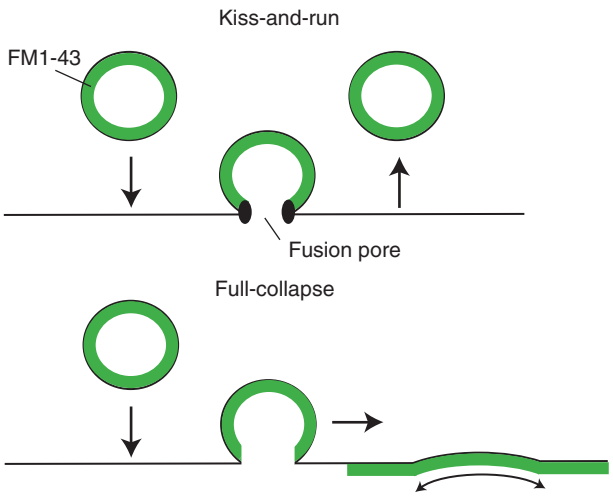

Figure 3 Optical monitoring of kiss-and-run endocytosis using FM dyes. FM1-43 and its analogs are partially lipophilic dyes which fluorescence increases almost 100-fold when they partition into membranes. Therefore, they are virtually invisible in aqueous solution. The chemical structure of FM dyes makes it difficult for them to cross lipid membranes, thus limiting the fluorescence labeling to outer membranes and recycling vesicles. During a typical experiment, stimulation in the presence of extracellular dye leads to uptake of dye molecules into vesicles that have undergone exocytosis followed by endocytosis. Subsequent washout of extracellular dye uncovers fluorescent staining specific to actively recycling vesicles during the stimulation paradigm. Consequent stimulation results in fusion of dye-loaded vesicles and loss of fluorescence trapped in synapses. This fluorescence loss is due to departitioning of the dye into aqueous solution or lateral diffusion of dye in neuronal membrane. During kiss-and-run endocytosis, a protein barrier that makes up the fusion pore restricts FM dye lateral diffusion. If vesicle retrieval is faster than the dye departitioning rate, then the endocytosed synaptic vesicle may still contain residual dye.

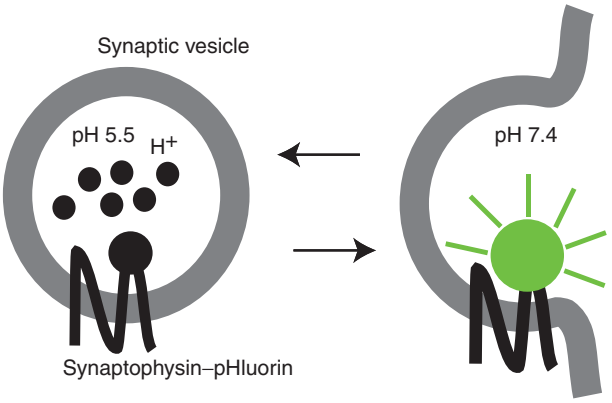

Figure 4 Synaptophysin–pHluorin is a fusion construct of the synaptic vesicle protein synaptophysin (a four transmembrane domain synaptic vesicle protein) with a pH-sensitive EGFP at its C-terminal (located in the synaptic vesicle lumen). The synaptic vesicle lumen normally has an acidic pH of approximately 5.5 at which spH fluorescence is quenched. When vesicles fuse, lumenal EGFP is exposed to the extracellular pH, which results in a marked increase in its fluorescence. During endocytosis, pHluorin fluorescence is requenched as the vesicle lumen becomes acidic.

imposing a significant delay in the fluorescence signal during spH endocytosis. Although most studies have used synaptobrevin as a carrier for the pHluorin tag, some studies have taken advantage of other synaptic vesicle proteins as pHluorin attachments. As previously indicated, in these experiments, tagging synaptophysin (a four transmembrane domain synaptic vesicle protein) and the vesicular glutamate transporter with pHluorin revealed very little surface exposure (as fluorescence) (Figure 4).

Experiments have shown that once the surface fluorescence is quenched, it is possible to monitor single vesicle fusion events using spH as a probe. Gandhi and Stevens detected single vesicle fusion events with square pulse-like rapid alkalinization and reacidification behavior akin to single ion channel openings. These findings suggest that the number of spH molecules per vesicle is preserved during the exocytosis and endocytosis cycle in a manner consistent with the kiss-and-run hypothesis. However, the same results were not reproducible with a pHluorin fusion construct of synaptophysin, suggesting that rapid reversal of pHluorin signal may be due to lateral diffusion of synaptobrevin because synaptophysin–pHluorin does not laterally diffuse upon vesicle fusion.

Capacitance Measurements

Synaptic vesicles are membranous organelles which release their neurotransmitter content after fusion with the plasma membrane. This process can be directly monitored via capacitance measurements. Alteration in total membrane capacitance can be used as a reporter of membrane fusion or endocytosis as long as the magnitude of the capacitance change is above the background noise. A critical requirement for employment of this method is direct electrical access to the secretory cellular component. This requirement significantly impairs the applicability of this method to small synapses of the central nervous system, which typically cover a 1 μm^2 surface area. Therefore, most capacitance measurements have been limited to secretion from cell bodies of mast cells, adrenal chromaffin cells, or large pituitary nerve terminals. Studies by Ling-Gang Wu and colleagues on a large presynaptic terminal called the calyx of Held, located within the medial nucleus of the trapezoid body region of the brain stem, have employed capacitance measurements in examining the coupling between synaptic vesicle exocytosis and endocytosis.

Despite its limited applicability to synaptic vesicle fusion, the capacitance measurement method, when available, provides several advantages due to its ability to report the kinetics of exocytosis and endocytosis and, in some cases, the properties of fusion pore openings. Wu and colleagues extended their earlier work using the cell-attached configuration of the patch clamp technique on or near the active zones of the calyx of Held nerve terminals. This configuration has several advantages over traditional whole terminal recording configuration due to electrical isolation of a small membrane area, thus limiting background noise. This setting allowed resolution of single vesicle fusion events and revealed several rapidly reversible fusion events consistent with kiss-and-run fusion. These rapidly reversible fusion events were previously detected only in models of cellular secretion (e.g., chromaffin cells and mast cells). These findings are considered to be the most conclusive evidence for the occurrence of kiss-and-run fusion in central synapses albeit with limited information regarding the regulation and prevalence of this type of fusion.

Carbon Fiber Measurements

The chemical structures of catecholamine or monoamine transmitters make them susceptible to rapid oxidation. These oxidation current signals can be detected using carbon fiber electrodes held at DC voltage above the redox potential of these neurotransmitters. This method provides an accurate depiction of vesicular neurotransmitter content free of nonlinear properties of postsynaptic ionotropic receptors. Furthermore, detailed analysis of individual fusion event profiles can provide information regarding diffusion kinetics of neurotransmitters out of fused vesicles. For instance, initial application of this method to catecholamine secretion of adrenal chromaffin cells revealed prominent slowly rising 'foot' signals preceding the actual rapid rise of neurotransmitter concentration after fusion. The characteristics of this foot signal have been analyzed in multiple settings to examine mechanisms underlying fusion pore expansion. Sulzer and colleagues measured dopamine release from small synaptic vesicles of cultured ventral midbrain neurons using carbon fiber amperometry. Their data indicate that small synaptic vesicle fusion pores flicker either once or multiple times in rapid succession, with each flicker releasing approximately 25–30% of vesicular dopamine. Rapidly flickering fusion pore openings are consistent with kiss-and-run-type fusion events. However, limited applicability of the amperometric recording method restricts generalizations to other types of central synapses that use amino acid neurotransmitters.

Alterations in Postsynaptic Receptor Activation due to Kiss-and-Run Fusion Events

As previously indicated, a key aspect of the kiss-and-run hypothesis is its potential for leading to postfusion regulation of chemical substance diffusion from the vesicle lumen through regulation of fusion pore opening duration and diameter. Catecholamine or dopamine secretion detected via carbon fiber measurements clearly indicates that vesicle fusion events can be limited by the rate and size of fusion pore openings. In the case of glutamatergic synaptic transmission in the brain, however, this process has been more difficult to detect. Nevertheless, in the literature there are several lines of evidence that argue for a limitation in the fusion pore opening kinetics altering the extent of glutamate diffusion into the synaptic cleft. Glutamatergic synapses typically possess two types of ionotropic postsynaptic receptors. AMPA receptors have a low-affinity for glutamate, and they mediate most of the electrical signaling in response to glutamate release from presynaptic terminals. NMDA receptors, on the other hand, have a very high affinity for glutamate, although they require postsynaptic depolarization and glutamate binding to become active. Experiments performed in hippocampal synapses have shown that activation of postsynaptic AMPA or NMDA receptors can be determined by the kinetics of fusion pore opening and the release profile of glutamate. In young nerve terminals, glutamate diffusion occurs through a narrow fusion pore leading to exclusive activation of NMDA receptors because they have a higher affinity for glutamate. Synapse maturation, in turn, leads to an increase in the preponderance of full fusion events, thus activating NMDA and AMPA receptors.

Under these conditions, one can detect synaptic transmission with the same apparent features of an NMDA-only synapse (also referred to as a 'silent synapse'), but the NMDA-only nature of these synapses is explained by a presynaptic mechanism and not by lack of postsynaptic AMPA receptors. This model may explain some forms of synaptic plasticity; therefore, it has been proposed as an attractive presynaptic alternative to the more commonly observed postsynaptically silent synapses (i.e., synapses with no AMPA receptors).

Conclusion

In the past decade, an increasing number of studies using multiple tools have examined the synaptic vesicle fusion process and uncovered fusion events consistent with the kiss-and-run endocytosis mechanism. Nevertheless, none of these observations are without caveats; therefore, despite extensive research, the existence of kiss-and-run fusion in central synapses

is still in question. As stated previously, work on kiss-and-run endocytosis has been dominated by functional approaches without significant molecular insight. However, future molecular studies on the exact role of dynamin in synaptic vesicle endocytosis (since kiss-and-run is likely to be a dynamin-dependent process) as well as the role of SNARE complex components in synaptic vesicle endocytosis will inevitably lead to a more accurate picture of exocytosis–endocytosis coupling in synapses and help elucidate the mechanism of kiss-and-run.

See also: Active Zone; Clathrin and Clathrin-Adaptors; Dynamin; Endocytosis and Presynaptic Scaffolds; Fusion Pore; SNAREs; Synaptic Depression; Synaptic Vesicles; Vesicle Pools.

Further Reading

Ceccarelli B, Hurlbut WP, and Mauro A (1973) Turnover of transmitter and synaptic vesicles at the frog neuromuscular junction. *Journal of Cell Biology* 57: 499–524.

Ehrlich M, Boll W, Van Oijen A, et al. (2004) Endocytosis by random initiation and stabilization of clathrin-coated pits. *Cell* 118: 591–605.

Fesce R, Grohovaz F, Valtorta F, and Meldolesi J (1994) Neurotransmitter release: Fusion or 'kiss-and-run'? *Trends in Cell Biology* 4: 1–4.

Gandhi SP and Stevens CF (2003) Three modes of synaptic vesicular recycling revealed by single-vesicle imaging. *Nature* 423: 607–613.

Harata NC, Aravanis AM, and Tsien RW (2006) Kiss-and-run and full-collapse fusion as modes of exo-endocytosis in neurosecretion. *Journal of Neurochemistry* 97: 1546–1570.

Kavalali ET, Klingauf J, and Tsien RW (1999) Properties of fast endocytosis at hippocampal synapses. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 354: 337–346.

Murthy VN and De Camilli P (2003) Cell biology of the presynaptic terminal. *Annual Review of Neuroscience* 26: 701–728.

Staal RG, Mosharov EV, and Sulzer D (2004) Dopamine neurons release transmitter via a flickering fusion pore. *Nature Neuroscience* 7: 341–346.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Voglmaier SM and Edwards RH (2007) Do different endocytic pathways make different synaptic vesicles? *Current Opinion in Neurobiology* 17: 374–380.

Energy Homeostasis: Adiposity Signals

S C Woods, University of Cincinnati, Cincinnati, OH, USA

© 2009 Published by Elsevier Ltd.

Introduction

Two major categories of peripherally arising signals influence food intake. The first consists of peptides secreted from the gastrointestinal tract during meals in response to ingested food. These acutely acting signals contribute to the perception of fullness and the termination of an ongoing meal and are collectively called satiety signals. The prototypical example is cholecystokinin (CCK). Messages conveyed by satiety signals are either relayed to the hindbrain by sensory nerves or else stimulate the hindbrain directly. The second category of signals includes hormones, particularly insulin from the pancreatic islets and leptin from adipose tissue, stomach, and elsewhere; the secretion of these hormones is proportional to the amount of fat in the body. These adiposity signals are transported from the circulation into the hypothalamus in the forebrain. The actions of adiposity signals and their interactions with other controllers of food intake are the topic of this article.

Overview of the Controls over Food Intake

For body weight (body fat, actually) to remain stable over time, energy intake (food intake) must match energy expenditure (the sum of excreted energy, heat loss, and physical work). Any prolonged deviation from this balance will result in weight gain or loss. Humans and most mammals take in energy in discrete bouts or meals. For most of us, the impetus to begin a meal is rarely, if ever, based on a biological deficit or need, such as insufficient glucose or other energy source, in some critical tissue. Rather, evidence suggests that the timing of when meals are initiated is based on psychological factors such as habit, time of day, social situation, and convenience, among others. Thus, if body weight is indeed regulated via a linking of energy intake to energy expenditure, then the regulatory process must logically be manifest as a control over how many calories are consumed once a meal begins (i.e., meal size). Consistent with this hypothesis, many secretions of the gastrointestinal tract that are generated during meals are proportional to the number of calories consumed, and some of these secretions function as satiety signals to the central nervous system to help limit meal size.

In contrast to satiety signals that are secreted mainly during meals, adiposity signals are tonically active, providing an ongoing message to the brain proportional to total body fat. When an individual's weight changes, the amounts of insulin and leptin secreted into the blood change in parallel, and this is reflected as an altered adiposity signal reaching the brain. As discussed in detail in the following sections, the net effect is a change in the brain's sensitivity to satiety signals. For example, when an individual is food restricted or voluntarily reduces intake sufficiently to lose weight, the consequent reduction of leptin and insulin renders brain circuits that control meal size less sensitive to meal-generated satiety signals. This means that more calories than normal must be consumed before a sufficient signal is generated to stop a meal. This situation persists until body weight returns to normal. Conversely, when an individual gains excess weight, the increased insulin and leptin signal in the brain results in increased sensitivity to satiety signals. Smaller meals are consumed until the excess weight is lost.

Adiposity Signals

Insulin and leptin are hormones that are secreted in direct proportion to the amount of white fat in the body, and each is transported via a receptor-mediated process through the blood–brain barrier into the hypothalamus and other brain areas (see Figure 1). Neurons sensitive to insulin and/or leptin thus receive a signal directly proportional to the amount of fat in the body. As discussed in the following sections, when exogenous insulin or leptin is added locally into the brain, food intake is reduced and body weight is lost. Conversely, when either the leptin or the insulin signal is reduced within the brain, more food is eaten and the individual gains weight. Consistent with their common actions, there is evidence that insulin and leptin, acting via their respective receptors, activate a common intracellular signaling cascade in neurons. In obesity, although circulating levels of both insulin and leptin are elevated, a disproportionately small amount of circulating insulin and leptin is transported into the brain. Analogously, the brain of obese individuals is also relatively insensitive to leptin and insulin.

To summarize, insulin and leptin provide an integrated signal to the brain regarding the amount of body fat. This signal influences the activity of integrative networks controlling energy homeostasis. When incremented locally in the brain, leptin and insulin have a net catabolic action, causing individuals to eat less and lose weight. Conversely, when

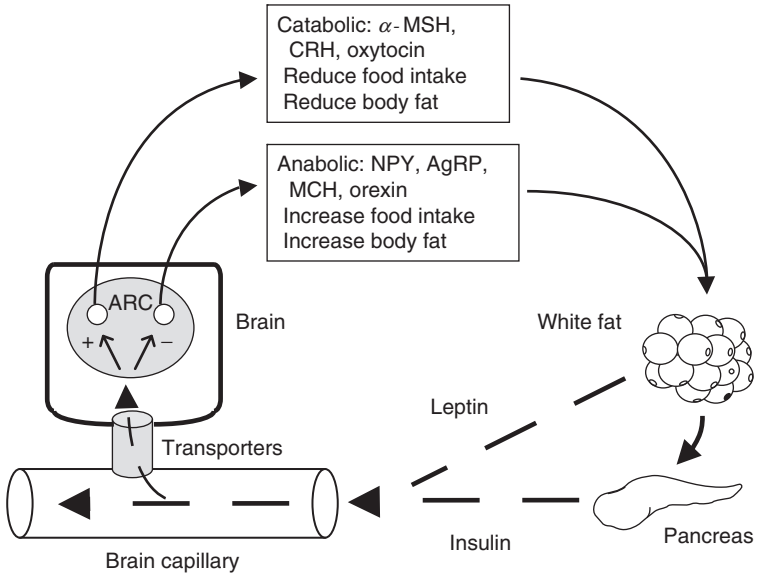

Figure 1 Diagram of the negative feedback control of body weight by insulin and leptin. Leptin from white fat and insulin from the pancreas are secreted in direct proportion to adiposity. Each penetrates through the blood–brain barrier by a receptor-mediated process. Receptors for insulin and leptin are expressed by two groups of hypothalamic arcuate (ARC) neurons. These in turn project to other brain areas to elicit net anabolic or catabolic actions. Catabolic transmitters in these circuits include α-melanocyte-stimulating hormone (α-MSH), corticotropin-releasing hormone (CRH), and oxytocin; anabolic transmitters include neuropeptide Y (NPY), agouti-related protein (AgRP), melanin-concentrating hormone (MCH), and the orexins.

the adiposity signal is reduced in the brain, food intake and body weight are increased (Figure 1).

Insulin as an Adiposity Signal

The best-known action of insulin is to increase glucose uptake in most peripheral tissues, consequently lowering blood glucose level. Diabetic patients take exogenous insulin to reduce their hyperglycemia, and a not uncommon side effect of this routine practice is to lower blood glucose to the point of hypoglycemia. However, because hypoglycemia *per se* can elicit hunger and induce eating, the best characterized behavioral effect of insulin administration is increased food intake, which is a consequence of insufficient glucose reaching the brain. This is not a direct effect of insulin, since administering glucose along with insulin does not result in eating, and since reducing the ability of the brain to utilize glucose for energy by administering nonmetabolizable glucose analogs also increases eating.

Insulin Acts within the Brain to Decrease Food Intake

Plasma insulin is low during fasting (basal condition) and increases during and immediately after meals (prandial condition) or after glucose administration (stimulated condition). Basal, prandial, and stimulated insulin levels are all directly proportional to the amount of white adipose tissue in the body, leaner individuals having lower levels and more obese individuals having higher levels. The amount of insulin in the plasma therefore conveys a signal indicating the degree of adiposity to any insulin-sensitive tissue. Insulin enters the brain via insulin receptor-facilitated transport through capillary endothelial cells. The process is saturable and selective for insulin. The transport of insulin into the brain is reduced in fasted animals, in animals maintained on a high-fat diet, and in genetic and dietary-induced obesity.

Insulin receptors are expressed in many areas of the brain, with high concentrations in the mediobasal hypothalamus, especially the arcuate nucleus, which also expresses the most active form of the leptin receptor. This region is thought to mediate most of insulin's catabolic action. When exogenous insulin is administered directly into the brain, near or directly within the mediobasal hypothalamus, animals behave as if they have excess fat in the body (i.e., they reduce their food intake and lose weight). The response is dose dependent, has been documented in numerous species including nonhuman primates, is not secondary to illness or incapacitation, and is more robust in males than in females. Animals fed a high-fat diet have central insulin resistance, perhaps contributing to the obesity that often develops. When insulin antibodies are administered in or near the mediobasal hypothalamus, animals overeat and gain weight.

Although insulin has not been administered directly into the brains of humans, certain formulations of insulin have been administered intranasally to humans, with a consequent increase of cerebrospinal fluid but not plasma insulin. Humans receiving insulin in this way eat less food and lose body fat.

To summarize, increased insulin in the mediobasal hypothalamus provides a signal that ample or excess energy is available in the body, and one consequence is a reduction of the amount of energy subsequently ingested. Simultaneously, the increased hypothalamic insulin signal also elicits a vagal reflex to the liver to suppress the production of glucose. Increased insulin (as well as leptin) is therefore analogous to other signals, indicating a surfeit of energy, such as an increase of certain lipids or glucose locally within the hypothalamus; there is evidence that overlapping intracellular signaling pathways mediate the overall catabolic response to these diverse metabolic signals.

Leptin as an Adiposity Signal

Since its discovery in 1994, the peptide hormone leptin has been identified as a key signal indicating the level of adiposity to the brain. Leptin is secreted primarily from adipocytes in direct proportion to adipose mass, although the actual stimulus is related to glucose metabolism within the fat cell. Analogous to the situation with insulin, leptin circulates in proportion to adiposity, is transported through the blood–brain barrier, and interacts with leptin receptors in the mediobasal hypothalamus and elsewhere in the brain. Once formed, the leptin receptor can be differentially spliced to yield fragments of different lengths. Only the complete or long form of the receptor, which is called ObRb, has intracellular signaling capacity, and it is localized in hypothalamic areas important in the regulation of energy homeostasis. Although ObRb are expressed in many tissues throughout the body, only those in the hypothalamus are necessary for the control of energy homeostasis.

Decades before leptin was discovered, both its action as a hormone and its receptor were inferred from a series of clever experiments on genetically obese mice with naturally occurring mutations. Douglas Coleman identified one strain of mouse that lacked a single gene and that was consequently hyperphagic and obese; he called them *obob* (obese) mice. He identified a second strain with a similar phenotype (hyperphagic and obese) that he called *dbdb* (diabetic) mice. He further discovered that normal control mice as well as *dbdb* mice had a factor in their blood that would cause *obob* mice to eat less and lose weight. He correctly concluded that *obob* mice lack the key hormone which was subsequently identified as leptin. He also discovered that blood from control mice, as well as blood from *obob* mice, had no effect on the eating or body weight of *dbdb* mice, and correctly concluded that *dbdb* mice, while producing the secreted hormone (leptin), lack the gene for the receptor for the hormone and are thus blind to its actions. The subsequent identification of leptin and its receptor verified Coleman's work and confirmed his prediction that administration of pure leptin (once it was discovered) would cure the obesity of *obob* but not *dbdb* mice.

Leptin is considered to be a gatekeeper signal with regard to activities controlled by the hypothalamus. Low leptin signaling, which typifies the lean state, attenuates activities not essential to acute survival such as reproduction. As body fat increases and leptin levels rise, a threshold is passed that disinhibits these activities, and further increments of leptin above the threshold appear to have little consequence.

Administration of leptin systemically or directly into the brain in or near the ventral hypothalamus of experimental animals causes reduced food intake and body weight. At the same time, leptin acts in the brain to increase energy expenditure by activating the sympathetic nervous system. Because of these actions, it was initially thought that exogenous leptin or other ligands for the leptin receptor could be used to treat human obesity. However, administering leptin systemically has not proved beneficial for weight loss in humans with frank obesity, although it might provide some benefit for individuals with type 2 diabetes. Because of this, the focus of trying to find therapeutic targets for leptin or for insulin to tackle body weight has shifted to circuits within the hypothalamus.

Central Neural Networks Influenced by Insulin and Leptin

Leptin receptors that are critical for leptin's catabolic actions are located mainly in the mediobasal hypothalamus and especially the arcuate nucleus (ARC), an area that contains parallel anabolic and catabolic circuits that are tonically active and influenced by adiposity as well as many other signals (**Figure 1**). The presence of these opposing hypothalamic circuits enables rapid and fine-tuned control over energy homeostasis, since the brain can simultaneously turn up one system (e.g., catabolic or anabolic) while turning down the other. Insulin and leptin cause a net reduction of activity in hypothalamic anabolic circuits and a net increase of activity in hypothalamic catabolic circuits.

Insulin and leptin interact with two specific populations of ARC neurons. Activity of neurons that synthesize proopiomelanocortin (POMC), the precursor molecule of the melanocortins, is stimulated by insulin and leptin, whereas activity of neurons that

synthesize neuropeptide Y (NPY) plus agouti-related protein (AgRP) is inhibited by insulin and leptin. α-Melanocyte-stimulating hormone (α-MSH) is a melanocortin that is derived from ARC POMC and is an agonist at melanocortin receptors (MC3R and MC4R) in several hypothalamic nuclei; AgRP is an endogenous antagonist of these same melanocortin receptors. Central administration of insulin or leptin, or the administration of α-MSH or synthetic agonists for melanocortin receptors, or the application of treatments that reduce ARC NPY (e.g., antisense oligonucleotides administered directly into the ARC), all result in reduced food intake and body weight. Conversely, the administration of NPY, AgRP, or synthetic melanocortin antagonists, or the absence of endogenous insulin or leptin signaling, all result in a net increase of food intake and body weight.

In sum, the ARC contains at least two types of neurons, the integrated activity of which determines body fat. ARC neurons respond directly to adiposity signals, as well as to local levels of energy-rich fuels, including glucose and long-chain fatty acids. The ARC also receives satiety information from the hindbrain. ARC neurons project to diverse brain areas, including several nearby regions of the hypothalamus. The paraventricular nuclei (PVN) expressing MC3R/MC4R and various Y receptors are targets for α-MSH, AgRP, and NPY. PVN neurons in turn secrete neuropeptides that exert a net catabolic action, including corticotropin-releasing hormone (CRH) and oxytocin. Central administration of CRH or oxytocin reduces food intake. Hence, excess body fat is signaled to the ARC via increased insulin and leptin activity, eliciting increased α-MSH and decreased NPY and AgRP activity. This in turn elicits increased activity in second-order neurons of the PVN, leading to reduced food intake and increased energy expenditure.

Lateral hypothalamic (LHA) neurons also receive axons from the ARC, and they secrete the anabolic peptides melanin-concentrating hormone (MCH) and the orexins. Central administration of either peptide increases food intake and body weight. The presence of opposing hypothalamic circuits for controlling energy homeostasis enables rapid and fine-tuned control over energy homeostasis, since the brain can simultaneously turn up one system (e.g., catabolic or anabolic) while turning down the other.

Central Resistance to Adiposity Signals

Obese individuals have high circulating insulin and leptin, yet they do not respond to these signals by reducing their caloric intake and losing weight. They are therefore said to be insulin and leptin resistant. Evidence suggests that the mechanism(s) underlying the resistance is manifest at several levels. First of all, as obesity increases, the proportionate amount of circulating insulin and leptin that is able to cross the blood–brain barrier decreases, such that the signal reaching key brain receptors is lower than is reflected in the blood. Second, the number of insulin receptors in the brain of several animal models of obesity is reduced. Finally, there is evidence that the intracellular signaling pathways linked to insulin and leptin receptors on brain neurons are compromised in obesity. Thus, in obesity, less insulin and less leptin arrive at brain neurons, the neurons have reduced numbers of insulin and leptin receptors, and when those receptors bind with insulin or leptin, the effect is reduced.

Recent evidence suggests that consumption of a diet containing relatively high amounts of saturated fats produces central insulin resistance independent of developing obesity *per se*. Finally, although the brains of both males and females respond to both insulin and leptin, there is an important sex difference. Females are relatively more sensitive to leptin whereas males are relatively more sensitive to insulin. One implication is that treatment strategies for obesity based on these adiposity signals might have to be gender specific.

Integration of Satiety and Adiposity Signals

Total daily food intake is the sum of intake in each individual bout or meal (including snacks). As discussed earlier, meal initiation is thought to be under the control of psychological as opposed to homeostatic signals. Regulatory control of energy intake based on homeostatic demands must therefore be exerted on how much is eaten during each meal, and meal size is at least partly under the influence of satiety signals. There is compelling evidence that the ability of satiety signals to terminate a meal varies with body fat, as signaled to the brain by leptin and insulin. When an individual has been calorie restricted (i.e., been on a diet) and loses weight, a reduced adiposity signal reaches the hypothalamus. This in turn reduces the sensitivity of the individual to satiety signals. The consequence is that larger meals are eaten. Conversely, individuals who have overeaten and gained weight have an increased adiposity signal, resulting in enhanced sensitivity to satiety signals.

Thus, a regulatory system exists in the brain that functions to maintain weight relatively stable over long periods of time. If circumstances dictate that body fat is reduced (or gained), the change is signaled to the brain with the result that more (or less) calories

are consumed at each meal. This situation persists until body fat is restored to its original level.

Conclusion

The brain – in particular the arcuate nucleus of the hypothalamus – receives and integrates diverse information pertinent to the maintenance of adequate stored energy as well as to the availability of ready-to-use energy in the blood in the form of glucose. The adiposity signals insulin and leptin provide a background tone in the hypothalamus, and this tone in turn determines the sensitivity of the individual to satiety signals, influencing how much food is eaten at any one time, as well as the sensitivity of the liver to signals that control glucose production. Malfunction of the signaling pathway for either leptin or insulin in the brain leads to increased fat deposition, peripheral insulin resistance and hyperglycemia, and/or other symptoms of the metabolic syndrome.

See also: Food and Water Intake: Regulation; Gastrointestinal Signals: Satiety; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Neuropeptides: Food Intake.

Further Reading

Clegg DJ, Brown LM, Woods SC, et al. (2006) Gonadal hormones determine sensitivity to central leptin and insulin. *Diabetes* 55 (4): 978–987.

Flier JS (2004) Obesity wars: Molecular progress confronts an expanding epidemic. *Cell* 116: 337–350.

Friedman JM (2002) The function of leptin in nutrition, weight, and physiology. *Nutrition Review* 60(10, part 2): S1–S14; discussion S68–S84, S85–S87.

Langhans W (1996) Metabolic and glucostatic control of feeding. *Proceedings of the Nutrition Society* 55: 497–515.

Levin BE, Routh VH, Kang L, et al. (2004) Neuronal glucosensing: What do we know after 50 years? *Diabetes* 53(10): 2521–2528.

Niswender KD and Schwartz MW (2003) Insulin and leptin revisited: Adiposity signals with overlapping physiological and intracellular signaling capabilities. *Frontiers in Neuroendocrinology* 24: 1–10.

Obici S, Feng Z, Morgan K, et al. (2002) Central administration of oleic acid inhibits glucose production and food intake. *Diabetes* 51(2): 271–275.

Porte D Jr., Baskin DG, and Schwartz MW (2005) Insulin signaling in the central nervous system: A critical role in metabolic homeostasis and disease from *C. elegans* to humans. *Diabetes* 54(5): 1264–1276.

Schwartz MW, Woods SC, Porte D Jr., et al. (2000) Central nervous system control of food intake. *Nature* 404(6778): 661–671.

Strader AD and Woods SC (2005) Gastrointestinal hormones and food intake. *Gastroenterology* 128(1): 175–191.

Strubbe JH and Woods SC (2004) The timing of meals. *Psychological Review* 111: 128–141.

Woods SC, Seeley RJ, Porte D Jr, et al. (1998) Signals that regulate food intake and energy homeostasis. *Science* 280(5368): 1378–1383.

Woods SC, Seeley RJ, Baskin DG, et al. (2003) Insulin and the blood–brain barrier. *Current Pharmaceutical Design* 9: 795–800.

Woods SC (2005) Signals that influence food intake and body weight. *Physiology & Behavior* 86(5): 709–716.

Energy Homeostasis: Endocannabinoid System

G Marsicano, AVENIR INSERM, Bordeaux, France
U Pagotto, C Cervino, and R Pasquali, University of Bologna, Bologna, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Cannabis sativa preparations, such as marijuana, hashish, and dagga, have been used in medicine for many centuries. *Cannabis* derivatives have also been proposed for patients in whom it is necessary to promote feeding and to reduce vomiting. Furthermore, the orexigenic properties of *Cannabis* have been anecdotally noted for causing the 'munchies' after marijuana smoking. Following the discovery of Δ^9-tetrahydrocannabinol (THC) as the main component of *Cannabis*, it became clear that this compound was responsible for the food intake promotion induced by marijuana. THC, also known as dronabinol, has been largely and rather successfully used to stimulate appetite in human syndromes characterized by wasting diseases, such as cancer or acquired-immunodeficiency-syndrome-associated anorexia, or forms of dementia, including Alzheimer's disease. More recently, the pharmacological blockade of cannabinoid type 1 (CB1) receptors by CB1 antagonists has emerged as a key component of comprehensive treatment of abdominal obesity and its multiple cardiometabolic abnormalities.

In this context, two notions confirm the importance of the endocannabinoid system in the regulation of food intake and energy metabolism. On the one hand is the finding of a high degree of evolutionary conservation of the role of this system in the regulation of feeding responses. On the other hand, high levels of endocannabinoids in maternal milk are critically important for the initiation of the suckling response in newborns.

The Endocannabinoid System

Endocannabinoid Synthesis and Degradation

After the identification of the chemical structure of THC, cannabinoid research received a further important boost by the discovery of the CB1 receptor, the binding site of exogenous cannabinoids. In the early 1990s, together with the characterization of another cannabinoid receptor named CB2, the first endogenous ligands for CB1 and CB2 receptors, anandamide (*N*-arachidonoylethanolamine) (AEA) and 2-arachidonoylglycerol (2-AG), were identified. Both are nonoxidative derivatives of arachidonic acid. Over the past few years, several other derivatives of long-chain polyunsaturated fatty acids have been described; they appear to act, at least in part, through CB1 and CB2 receptors. Their functions are, however, less characterized as compared to AEA and 2-AG.

Endocannabinoids are lipophilic substances and their synthesis derives from phospholipid precursors. Unlike other neurotransmitters or hormones, they are not stored in vesicles but rather are released 'on demand.' This peculiar characteristic needs, therefore, a strict regulation of the different phases of their release, uptake, and degradation. In general, the synthesis of endocannabinoids is triggered by elevated intracellular concentrations of Ca^{2+}, such as during membrane depolarization. However, Ca^{2+}-independent processes have also been proposed to induce endocannabinoid synthesis.

The formation of AEA occurs in two steps. Initially, the precursor phosphatidylethanolamine, an abundant lipid present in the cell membrane, exchanges the ethanolamine moiety with an arachidonic acid moiety to yield *N*-arachidonoyl phosphatidylethanolamine using the enzyme *N*-acyltransferase. Thereafter, AEA is synthesized from this intermediate by the recently cloned *N*-acylphosphatidylethanolamine-selective phospholipase D. In addition, AEA can also be synthesized from *N*-arachidonoyl phosphatidylethanolamine via an additional intermediate, *N*-arachidonoyl phosphatidyl lysoethanolamine. The second major endocannabinoid, 2-AG, is synthesized in two steps from phosphatidylinositol, another lipid precursor abundantly present in membranes. Two different pathways have been described for the synthesis of 2-AG: 1,2-diacylglycerol and lysophosphatidylinositol were found to be intermediate products. A phospholipase C and a phospholipase A1, respectively, are involved in the generation of the two intermediate compounds. Two diacylglycerol lipase (DAGL) enzymes (DAGL-α and DAGL-β) catalyzing the synthesis of 2-AG from 1,2-diacylglycerol have recently been cloned. The other enzymes involved have not yet been clearly identified.

In a way similar to mechanisms involving 'classical' neurotransmitters, it is proposed that endocannabinoids – after exerting their effects in the extracellular space – are taken up into the cell by a transport mechanism. Transport is not driven by transmembrane ion gradients, but it does appear to be facilitated by membrane diffusion of these lipophilic compounds. The molecular identity of a putative endocannabinoid transporter(s) has not yet been

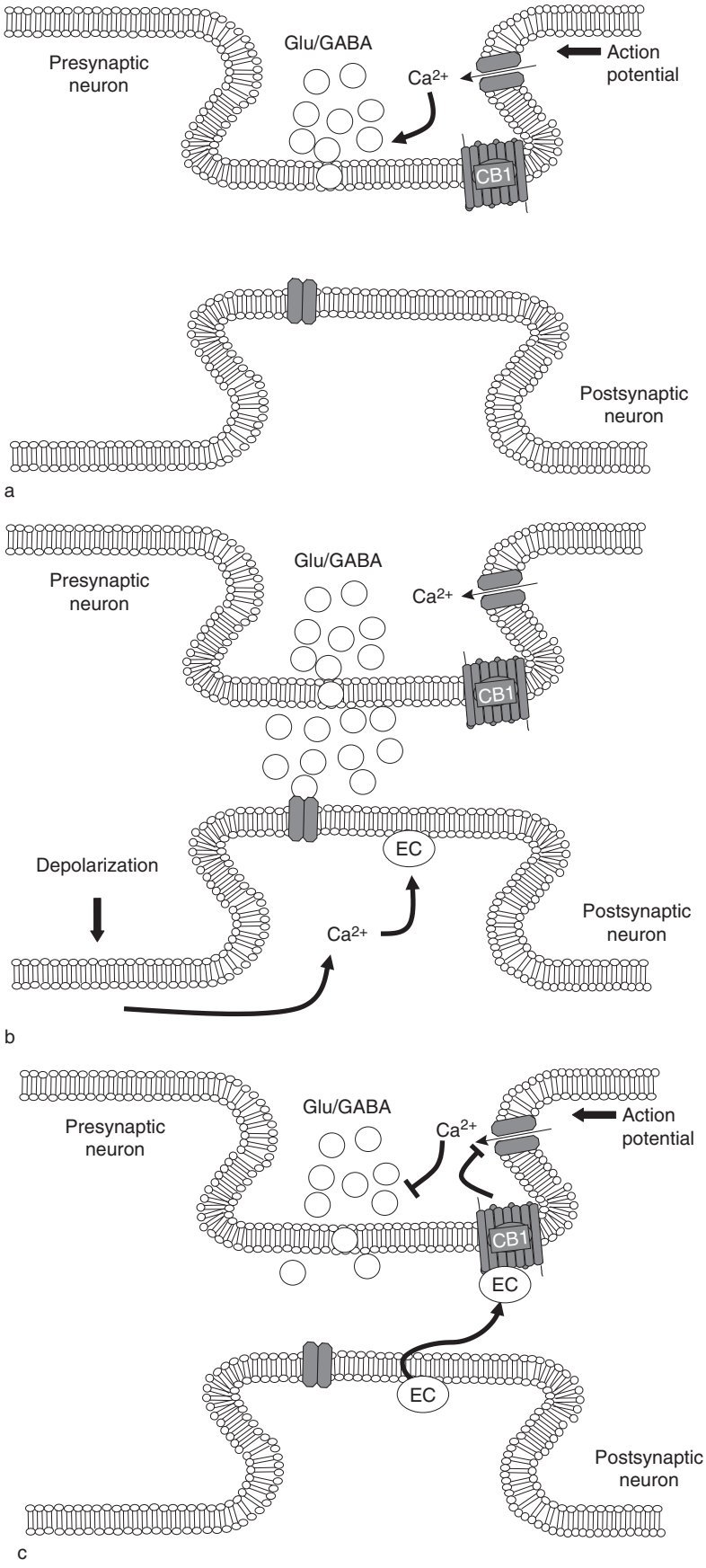

Figure 1 Continued.

ascertained. However, it has become clear that this process is complex and likely involves several proteins and mechanisms.

The degradation of endocannabinoids is better understood than is the biosynthesis. AEA is hydrolyzed to arachidonic acid and ethanolamine by the enzyme fatty acid amide hydrolase (FAAH). Degradation of 2-AG likely involves at least two enzymes. Monoacylglycerol lipase appears to be responsible for about half of the 2-AG-hydrolyzing activity, suggesting that additional 2-AG-hydrolyzing enzymes may exist.

Signal Transduction of the CB1 Receptor

In the central nervous system (CNS), endocannabinoids can act as neurotransmitters, transferring information from one neuron to the next. Here, postsynaptically released endocannabinoids travel to the presynaptic site where they activate CB1 receptors. They thus mediate a retrograde signal (Figure 1). The overall effect is a decrease in the presynaptic release of neurotransmitters such as glutamate and γ-aminobutyric acid (GABA). This phenomenon is present in synaptic connections of many brain regions, thus representing an important modulatory mechanism of neuronal transmission. The signal transduction of cannabinoid receptors might lead to the stimulation of different intracellular pathways, depending on the cell type involved and the experimental conditions. For instance, the CB1 receptor, which normally inhibits adenylate cyclase, can also stimulate the cyclic adenosine monophosphate (cAMP) pathway in some conditions. Moreover, recent results suggest the possibility of functional interactions of CB1 receptors with other receptors – for instance, with type 1 orexin receptors, 5-hydroxytryptamine (5-HT_2; serotonin) receptors, and dopamine D2 receptors. The possibility that such interactions depend on hetero-oligomerization processes might represent a very interesting novel aspect that will expand the view of the pharmacology and physiology of the endocannabinoid system. Activation of the CB1 receptor is also able to inhibit N- and P/Q-type voltage-activated inward Ca^{2+} channels in a $G_{i/o}$-dependent, but adenylate cyclase-independent, manner. CB1 receptor stimulation is also coupled to the opening of K^+ channels in a $G_{i/o}$-dependent manner. A-Type potassium ($K^+{}_A$) channels are activated by the CB1 receptor through dephosphorylation processes, probably mediated by the inhibition of adenylate cyclase, whereas inwardly rectifying K^+ (K_{ir}) channels are stimulated by direct coupling to the $G_{i/o}$ proteins. Cannabinoids acting at CB1 receptors are able to decrease the Ca^{2+} influx by inhibiting N- and P/Q-type Ca^{2+} channels and to increase the efflux of K^+ ions by activating K_{ir} and $K^+{}_A$ channels in neurons. Stimulation of K^+ and inhibition of Ca^{2+} currents are generally considered as the main effects of CB1 receptors on intracellular ion levels. However, contrary effects have also been described in certain neuronal systems. CB1 receptor stimulation activates extracellular signal-regulated kinases (ERKs), and the phosphorylation of ERKs in the hippocampus appears to physiologically depend on the 'on-demand' activation of the endocannabinoid system. Another important kinase pathway modulated by CB1 receptor is the phosphoinositol 3-kinase (PI3K) and its downstream effector protein kinase B (PKB; also known as Akt). Interestingly, at least in *in vitro* systems, the activation of the PI3K pathway appears to be upstream to the stimulation of ERK phosphorylation. Therefore, CB1 receptor stimulation might activate the PI3K pathway, which in turn activates Raf-1, and subsequently stimulates ERK phosphorylation. Thus, activation of ERKs might be linked to several intracellular effects, such as activation of Na^+/H^+ exchanger and activation of phospholipase A_2. Moreover, ERK stimulation might link CB1 receptor to the expression of immediate-early genes (IEGs). Several transcription factors, such as zif268, c-fos, and c-jun, were shown to be activated by exogenous CB1 receptor agonists. In analogy to the activation of ERKs, the endocannabinoid system has been proposed to control expression of zif268 and c-fos in pathophysiological conditions. Brain-derived neurotrophic factor (BDNF) is encoded by another immediate-early gene that is regulated *in vivo* and *in vitro* both by pharmacological treatment with CB1 receptor agonists and by endogenously produced cannabinoid. In summary, endocannabinoids appear to be very versatile signaling mediators.

Distribution of Cannabinoid Receptors in Tissues and Cells Controlling Energy Balance

Expression studies show that the CB1 receptor is one of the most abundant G-protein-coupled receptors in the mammalian brain. CB1 receptors have been found in all brain regions involved in the control of

Figure 1 Retrograde signaling mediated by endocannabinoids (ECs) and cannabinoid type 1 (CB1) receptors. (a) A presynaptic neuron is activated by an action potential, inducing an increase of intracellular Ca^{2+}, which, in turn, activates the machinery for the release of neurotransmitter – for example, glutamate or γ-aminobutyric acid (Glu/GABA; empty circles). (b) During synaptic neurotransmission, a stimulus in the postsynaptic neuron (e.g., depolarization) can induce the synthesis of ECs from the postsynaptic membrane. (c) Postsynaptically synthesized ECs are released into the synaptic cleft and activate presynaptic CB1 receptors. CB1 activation induces a decrease of presynaptic Ca^{2+} and, thereby, reduces the release of neurotransmitter. Note: This is a simplified representation of the retrograde signaling of ECs. Different mechanisms have been described to be involved in these events.

energy intake, including the nucleus accumbens, the paraventricular (PVN), and the ventromedial hypothalamic nuclei, in the lateral hypothalamus, the nodose ganglion, and the vagus nerve, and in the myenteric neurons and epithelial cells of the large intestine. In some regions of the brain, the ratio between the estimated amount of CB1 receptor and the level of G-protein activation is not always constant, thus indicating regional differences in receptor-coupling efficiencies. This is important to consider, because sometimes the endocannabinoid system seems to exert functions involving regions where the density of CB1 receptors is relatively low (e.g., in the hypothalamus, modulation of food intake). Therefore, the activity of cannabinoids on CB1 receptors cannot be predicted based solely on the relative receptor density, and other factors, such as receptor-coupling efficiency, should be taken into account. Finally, recent studies have highlighted CB1 receptor presence in various peripheral organs involved in metabolic control, including adipocytes, hepatocytes, endocrine pancreatic cells, and skeletal muscle cells.

Neuronal Endocannabinoid Control of Food Intake

Analogous to what has been described for THC, endocannabinoids also display orexigenic properties, and after the synthesis of specific CB1 receptor antagonists, the orexigenic properties of endocannabinoids have been definitively attributed to an activation of CB1 receptor.

Regulation of Rewarding Properties of Food by Endocannabinoids

As a general hypothesis, endocannabinoids gradually increase during intermeal intervals, reaching a critical level in order to trigger the motivation to eat. In favor of this possibility it should be borne in mind that AEA and 2-AG levels are increased in the fasting condition in the nucleus accumbens and are decreased after a meal. The endocannabinoid system is considered as a general stress recovery system and is overall 'silent.' The system exerts its modulatory actions only 'when' and 'where' it is needed to maintain the homeostasis after a stressful perturbation and its activation exhibits tightly regulated spatial and temporal selectivity. If this theory holds true, the elevation of endocannabinoids after fasting to promote food intake fits in this scenario. In fact, fasting remains one of the most challenging stressors for animals. In this view, the longer the time since the last meal, the greater the activity in relevant endocannabinoid circuits, and consequently the higher is the motivation to eat, driven by endocannabinoids. The endocannabinoid system is one of the key players in the modulation of 'reward/reinforcement' circuitry that is represented by an interconnection of brain nuclei associated with the medial forebrain bundle, linking the ventral tegmental area, the nucleus accumbens, and the ventral pallidum. The most relevant pathway of this circuitry is represented by the mesolimbic dopaminergic system. A relevant cross-talk between limbic endocannabinoid and dopaminergic systems to modulate the rewarding properties of tasty food is known to occur, as documented by the ability of a dopamine D1 antagonist to reduce the orexigenic stimulus induced by THC. Therefore, bearing in mind the anatomical distribution of CB1, which is highly expressed in the mesolimbic area, the initial leading hypothesis was in favor of a role of endocannabinoids as modulators of the appetitive value of ingested substances. Therefore, the first CB1 receptor antagonists were tested in many experimental settings aimed at monitoring their ability to reduce the intake of palatable food or alcohol. Whether the endocannabinoids increase the incentive value of the food, regardless of the quality of the macronutrients ('incentive hypothesis'), or, as also derived from experiments in humans with phytocannabinoids, whether endocannabinoids drive the search for highly palatable food ('orosensory reward hypothesis'), is still an unsolved issue. The opioid system is another important neuronal signaling system involved in the modulation of food intake. Endocannabinoids act in a synergistic manner with opioids to stimulate feeding behavior, and this interaction seems to occur not only at the mesolimbic system level but also at the level of the PVN of the hypothalamus.

The cross-talk between endocannabinoids and serotoninergic pathways in the modulation of feeding behavior has been less investigated as compared to the opioidergic one. Nonetheless, it seems that the two pathways stimulate food intake via independent mechanisms of action.

Endocannabinoids and the Neural Hypothalamic Network Controlling Food Intake

At the hypothalamic level, specific neuronal populations exert a strategic control over food intake. By interconnecting different hypothalamic nuclei, this network provides high levels of adaptability of feeding behavior to a multiplicity of central and peripheral signals. Signals coming from various peripheral organs, such as the liver, gastrointestinal tract, and adipose tissue, are conveyed at the hypothalamic level to constantly inform the brain about the state of nutrition via neural and humoral routes (see Figure 2). On the other hand, the hypothalamus appears to be sensitive to a wide range of metabolic perturbations associated with

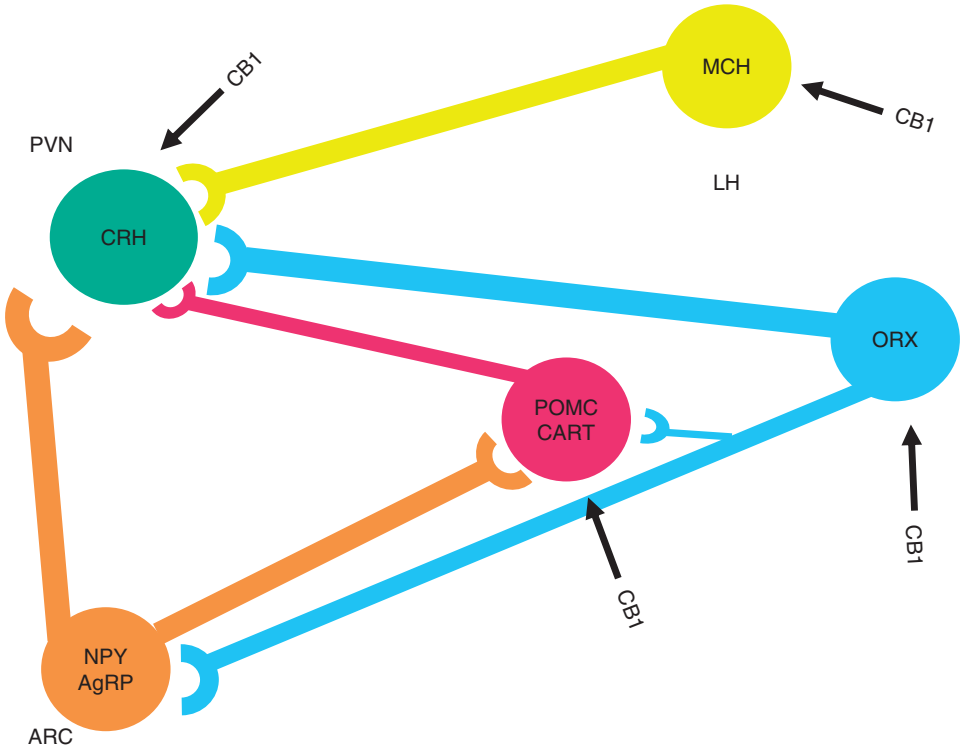

Figure 2 Schematic representation of the putative role of endocannabinoids through cannabinoid type 1 (CB1) receptor activation in the hypothalamic circuits controlling food intake, depicting sites of action of hypothalamic endocannabinoids (black arrows) on corticotropin-releasing hormone (CRH) neurons in the paraventricular nucleus (PVN), on orexin (ORX) and melanin-concentrating hormone (MCH) neurons in the lateral hypothalamus (LH), and on proopiomelanocortin/cocaine- and amphetamine-regulated transcript (POMC/CART)-producing neurons (ARC, arcuate nucleus; AgRP, agouti-related protein; NPY, neuropeptide Y). In addition, the endocannabinoid system has been described to control glutamatergic and GABAergic transmission onto several hypothalamic nuclei.

fuel availability. In turn, this information is used to adjust both caloric intake and the deposition of fuels in the periphery.

Endocannabinoids are variably produced in the hypothalamus of rodents administered a standard diet. Analogous to the endocannabinoid changes at the level of mesolimbic system, 2-AG levels have been shown to increase during acute fasting, decline as the animals are re-fed, and return to normal values in satiated animals in the hypothalamus. Moreover, as demonstrated in adult rodents and in pups from undernourished dams, long-term food deprivation is known to reduce the amount of endocannabinoid production. Therefore, also at the hypothalamic levels endocannabinoids seem to display the ability to quickly respond to changes in nutritional states.

Leptin signaling interacts with endocannabinoids at the level of the hypothalamus. Animal models of impaired leptinergic signaling resulting in an obese and hyperphagic phenotype, such as that exhibited by *ob/ob* or *db/db* mice or *fa/fa* Zucker rats, are characterized by increased levels of endocannabinoids in the hypothalamus. Moreover, a single intravenous injection of leptin in *ob/ob* mice reduces the pathological overproduction of endocannabinoids. In these animals, CB1 antagonist treatment is able to reduce food intake, confirming the anorectic properties of this class of drugs. CB1 receptors present in the lateral hypothalamus co-localize with melanin-concentrating hormone (MCH) and prepro-orexin neurons. Populations of both orexin and MCH neurons have been implicated in the hedonic or motivational aspects of food intake by modulating projections to dopaminergic neurons in the ventral tegmental area. MCH neurons (promoting an increase in food intake) are tonically inhibited by GABAergic inputs coming from the limbic system. On one hand, by decreasing GABA release, endocannabinoids may enhance the excitability of MCH neurons, leading to increased feeding behavior, while on the other hand, by inhibiting the voltage-gated calcium currents in the same neurons, the leptin signal may result in less synthesis and release of endocannabinoids, leading to a reduced excitability of MCH neurons and to a decrease in appetite as a final effect. Therefore, the lateral hypothalamus may represent the site of converging signals between hypothalamic and mesolimbic structures to mediate the orexigenic activity of endocannabinoids.

An interaction between endocannabinoids and corticotropin-releasing hormone (CRH) has been hypothesized, based on the co-expression of CRH and CB1 mRNA in the PVN and the co-expression of CRH type 1 receptors and CB1 receptors in several brain regions. Accordingly, postsynaptically released endocannabinoids acting at presynaptic CB1 receptors are able to decrease glutaminergic transmission onto CRH-producing neurons, resulting in an inhibition of CRH release. This release of endocannabinoids from the parvocellular neurons is stimulated by a nongenomic effect of glucocorticoids. Therefore, it is conceivable that the well-known regulation of food intake by glucocorticoids may partly derive from a functional cross-talk with the endocannabinoid system.

The arcuate nucleus is the production site of neuropeptide Y (NPY), one of the most important hypothalamic orexigenic neuropeptides. NPY-positive neurons send projections to other feeding areas, including the PVN, the ventro- and dorsomedial hypothalamus, and the lateral hypothalamus. The presence of CB1 receptors in the arcuate nucleus is very limited if not absent, and the NPY/agouti-related protein system in the arcuate nucleus does not seem to be directly targeted by endocannabinoid action. However, endocannabinoids have been found to increase, and a CB1 receptor antagonist to decrease, depolarization-induced NPY release in a hypothalamic explant. This evidence does not mean that the effect of cannabinoids on NPY neurons is caused by a direct action, but it is more likely that cannabinoid agonists and antagonists are able to affect an intermediate factor, which in turn alters NPY release.

A functional cross-talk between CB1 receptors and melanocortin receptor type 4 (MCR4) receptors in the control of food intake has been recently highlighted by the finding of the synergistic action of subanorectic doses of a CB1 antagonist and of an MCR4 agonist administered together. Furthermore, CB1 receptor antagonists are able to inhibit the stimulation of food intake induced by MCR4 antagonists, whereas the orexigenic impulse initiated by the administration of CB1 receptor agonists is not blocked by co-stimulation with MCR4 agonists. These findings seem to suggest that melanocortin receptor signaling in the hypothalamic regulation of food intake is upstream of the activation of the endocannabinoid system and that CB1 receptors display an obligatory role in α-melanocortin-stimulating hormone (α-MSH) effects on food intake.

The interaction between cocaine- and amphetamine-related transcript (CART) and endocannabinoids remains controversial. $CB1^{-/-}$ (knockout) mice in basal conditions have reduced CART expression, suggesting that blocking the endocannabinoid signaling impacts negatively on CART expression. In contrast, mice with a genetic deletion of the anandamide-degrading enzyme FAAH, having consequently increased levels of endocannabinoids, exhibit reduced CART immunoreactivity in several hypothalamic regions involved in food intake. Moreover, in the same animals, treatment with CB1 antagonists is able to restore normal levels of CART expression, indicating that inhibition of CART signaling by CB1 receptor activation may be involved in the orexigenic effect of AEA. In conclusion, there is solid evidence confirming the presence of an interaction between endocannabinoids and CART, but the precise physiological significance deserves further study for complete elucidation. In analogy to the interaction with the CART system, cross-talk between endocannabinoids and ghrelin is undergoing intense scrutiny. Ghrelin is a hormone released by the stomach into the bloodstream during food deprivation, to signal to the hypothalamus the need for energy intake. Treatment with CB1 antagonists has been shown to strongly reduce the orexigenic effect stimulated by intracerebroventricular injection of ghrelin. Therefore, ghrelin might promote food intake by elevating endocannabinoid hypothalamic levels, but this hypothesis needs to be confirmed by further evidence.

A growing spectrum of data indicates that, like many peripheral cell types, some hypothalamic neurons integrate information about fuel sources other than glucose, to modulate food intake. An increasing body of evidence points to a critical role for both fatty acid production and utilization in regulating hypothalamic neurons implicated in the control of food intake. Consistent with this, it has recently been proposed that endocannabinoids may modulate the fatty acid synthetic (FAS) pathway in the hypothalamus via CB1 receptors and that the increase in food intake following food deprivation may involve a CB1 receptor modulation of the FAS pathway. CB1 receptor antagonist treatment has been demonstrated to inhibit the increase in FAS hypothalamic expression, and this may explain the anorexigenic properties of cannabinoid antagonists.

Endocannabinoids and the Enteric Nervous System

The CB1 receptor and endocannabinoids are highly expressed in neurons of the gastrointestinal tract. In the small intestine, starvation induces a sevenfold increase in AEA release, and this effect is reversed on re-feeding. Cholecystokinin (CCK) is known to have an important satiety role and the CCK type 1 receptor is expressed in the same vagal afferent neurons which also express the CB1 receptor. These neurons project from the stomach and duodenum to the brain stem. In addition, CB1 receptor expression

in the nodose ganglia is increased by food deprivation and is decreased after re-feeding; this effect has been shown to be blocked by CCK antagonists and mimicked by administration of CCK. Therefore, the endocannabinoid system could also affect food intake acting through CCK signaling.

Peripheral Endocannabinoid Control of Energy Metabolism

The Endocannabinoid System and Peripheral Organs Involved in Control of Metabolic Processes

The endocannabinoid system plays a central role in the neuronal control of food intake and energy balance. However, recent results have added important elements to the scenario of the mode of action of endocannabinoids in the regulation of food intake and energy metabolism, and they indicate that peripheral nonneuronal cell types also contribute to the overall functions of the endocannabinoid system. The advent of the $CB1^{-/-}$ mouse model has provided a step forward to substantiate the hypothesis of an additional effect of endocannabinoids in peripheral organs. Animals lacking CB1 receptors are hypophagic, leaner, and lighter as compared to wild-type littermates. Importantly, pair-feeding experiments show that in adult $CB1^{-/-}$ mice, metabolic factors appear to be the major cause of the lean phenotype. These differences in body weight regulation are amplified when a high-fat diet is administered to the same groups of animals, allowing the conclusion that the lack of CB1 receptors may confer resistance to diet-induced obesity (DIO). These observations converge to support the idea that additional peripheral food intake-independent metabolic functions may participate, or even predominate, in the control of energy balance exerted by the endocannabinoid system. Similar observations were reported in DIO mice using pharmacological CB1 receptor blockade. Long-term treatment with CB1 antagonists produced a transient reduction in food intake in DIO mice followed by the development of tolerance to the anorectic effect of the drug. However, the effect on body weight was sustained until the end of the 5-week experiment as compared to vehicle-treated DIO mice. It is now becoming clear that endocannabinoids interact with all of the organs involved in the control of metabolic functions. CB1 receptors are functionally active in white adipocytes stimulating lipogenesis. Importantly, CB1 antagonist treatment induces adiponectin release from adipocytes in culture, in animals and also in humans. Adiponectin is an adipose tissue-derived peptide known to possess antiatherogenic and antidiabetic properties. The circulating levels of adiponectin are directly correlated to increased insulin sensitivity. Moreover, the peptide is able to decrease hepatic glucose production and to regulate free fatty acid metabolism, via suppression of lipogenesis and activation of free fatty acid oxidation.

The CB1 receptor is also present in skeletal muscles. In *ob/ob* mice, treatment with a CB1 receptor antagonist generates considerable glucose uptake in soleus muscle in comparison with vehicle-treated animals.

Activation of the CB1 receptor in the liver increases gene expression of the lipogenic transcription factor sterol regulatory element binding protein-1c (SREBP-1c) and of its targets, FAS and acetyl-CoA carboxylase-1 (ACC-1), leading to a stimulation of *de novo* lipogenesis. Moreover, prolonged consumption of a high-fat diet in mice increases hepatic levels of AEA and induces an upregulation of CB1 receptor expression that may in turn promote the development of steatosis.

Summary

Energy balance and food intake control represent an extremely exciting area for studies on the regulatory functions of the endocannabinoid system. Strong evidence supports the idea of plasticity in endocannabinoid tone when shifting from conditions of normal to impaired energy balance. In fact, endocannabinoid levels are transiently upregulated in the CNS areas involved in control of food intake following short food deprivation and are relatively chronically downregulated following long-lasting food restriction. Moreover, it seems that a permanent and pathological overactivation of the hypothalamic endocannabinoid tone is associated with obesity, contributing to the hyperphagia often associated to this disorder. In addition to this central mode of action, endocannabinoids also appear to act at a peripheral level, by targeting a large variety of organs involved in metabolic processes. By reducing intestinal motility and energy expenditure and by optimizing nutrient assimilation and storage, the endocannabinoid system may be one of the most important players in the generation of our ancestral precursor, the 'thrifty genotype,' a human who was exceptionally efficient in the intake and/or utilization of food. Nowadays, considering the dramatic differences in ancestral and contemporary human lifestyles, the endocannabinoid system may be one of the key elements to explain the pandemic phenomenon of obesity.

See also: Appetitive Systems: Amygdala and Striatum; Cortical Processing of the Reward Value of Food; Eating Disorders; Endocannabinoid Role in Synaptic Plasticity and Learning; Food and Water Intake: Regulation; Hypothalamic Structure–Function Relationships;

Neuropeptides: Enteric Nervous System; Neuropeptides: Food Intake; Psychopharmacology of Reward and Appetite in Rats.

Further Reading

Cota D, Marsicano G, Tschoep M, et al. (2003) The endogenous cannabinoid system affects energy balance via central orexigenic drive and peripheral lipogenesis. *Journal of Clinical Investigations* 112: 423–431.

Cota D, Tschoep MH, Horvath TL, and Levine AS (2006) Cannabinoids, opioids and eating behavior: The molecular face of hedonism? *Brain Research: Brain Research Reviews* 51: 85–107.

Dallman MF (2003) Fast glucocorticoid feedback favors 'the munchies.' *Trends in Endocrinology and Metabolism* 14: 394–396.

Di Marzo V, Bifulco M, and De Petrocellis L (2004) The endocannabinoid system and its therapeutic exploitation. *Nature Reviews Drug Discovery* 3: 771–784.

Di Marzo V, Goparaju SK, Wang L, et al. (2001) Leptin-regulated endocannabinoids are involved in maintaining food intake. *Nature* 410: 822–825.

Di Marzo V and Matias I (2005) Endocannabinoid control of food intake and energy balance. *Nature Neuroscience* 8: 585–589.

Horvath TL (2003) Endocannabinoids and the regulation of body fat: The smoke is clearing. *Journal of Clinical Investigations* 112: 323–326.

Jo YH, Chen YJ, Chua SC, Jr., Talmage DA, and Rode LW (2005) Integration of endocannabinoid and leptin signaling in an appetite-related neural circuit. *Neuron* 48: 1055–1066.

Marsicano G and Lutz B (2006) Neuromodulatory functions of the endocannabinoid system. *Journal of Endocrinological Investigations* 29(supplement 3): 27–46.

Pagotto U, Marsicano G, Cota D, Lutz B, and Pasquali R (2006) The emerging role of the endocannabinoid system in endocrine regulation and energy balance. *Endocrine Reviews* 27: 73–100.

Peters H and Nahas GG (1999) A brief history of four millennia (BC 2000–AD 1974). In: Nahas GG, Sutin K Harvey D, and Agurell S (eds.) *Marihuana and Medicine*, pp. 3–7. Totowa, NJ: Humana Press Inc.

Piomelli D (2003) The molecular logic of endocannabinoid signalling. *Nature Reviews Neuroscience* 4: 873–884.

Energy Homeostasis: Hypothalamic Development

K L Grove, Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

Arcuate Neurons Are Key Homeostatic Feedback Regulators of Energy Balance

There are three major brain pathways that influence feeding and energy homoeostasis. (1) Hormonal homeostatic feedback from peripheral metabolic signals (including insulin, leptin, adiponectin, glucose, free fatty acids, glucagon-like peptide-1 (GLP-1), ghrelin, peptide YY, and pancreatic polypeptide). Within the brain, these signals primarily act in the hypothalamus and brain stem and can be involved in both stimulation and inhibition of food intake. (2) Vagal-gustatory sensory signals activate the dorsal vagal complex in response to increased gut distention to inhibit food intake. (3) Hedonic or 'reward' circuitry from the cortical/limbic system regulates food based on palatability or rewarding properties.

The arcuate nucleus (ARH) is located at the base of the hypothalamus and is generally considered the key conduit for sensing peripheral signals of energy balance. Although these neurons are within the blood–brain barrier, they do appear to have specialized access to circulating factors. The nature of this specialized access remains to be determined. The neurons in the ARH transmit these homeostatic signals to other hypothalamic sites that integrate the peripheral metabolic signals with incoming signals from the brain stem (vagal-gustatory sensory signals) and the cortical/limbic systems (hedonic or 'reward'-based information). In general, it is thought that all of these signals converge onto the paraventricular nucleus of the hypothalamus (PVH), which in turn generates neuronal signals to initiate appropriate responses to maintain body weight and energy homeostasis (i.e., modulate energy expenditure or initiate food foraging). The exact mechanism by which the hypothalamus initiates feeding behavior or food foraging is poorly understood. The modulation of energy expenditure is through activation of autonomic neurons in the parvocellular region of the PVH. These autonomic neurons project to the brain stem and spinal cord (intermediallateral cell column (IML)) to activate sympathetic and parasympatheic outputs to target organs (i.e., pancreas, adipose tissue, liver).

The two most-studied populations of neurons in the ARH involved in regulating homeostatic feedback circuits are the orexigenic neuropeptide Y/agouti-related peptide (NPY/AgRP) neurons and the anorexic α-melanocyte-stimulating hormone (αMSH – a peptide cleaved from pro-opiomelanocortin (POMC) prohormone) neurons. In the hypothalamus, NPY is thought to stimulate food intake and suppress energy expenditure primarily through the Y1 receptor, a G-protein-linked receptor; however, there is strong evidence that both the Y2 and Y5 receptor subtypes also play a role. Both AgRP and αMSH regulate food intake and energy expenditure through the melanocortin receptors MC3 and MC4; AgRP is an endogenous antagonist, whereas αMSH is an agonist for these G-protein-linked receptors. Thus, ARH-NPY/AgRP neurons provide a powerful orexigenic signal through activation of the Y receptors and through antagonism of the anorexic MC3/4 receptors.

The importance of the ARH-NPY/AgRP and αMSH neurons in the regulation of homeostatic feedback control is conserved in most species, including fish, worms, birds, sheep, rodents, and primates (both human and nonhuman). The most likely species difference in the regulation of food intake is the hedonic- or 'reward'-based feeding, which appears to be more prominent in higher species and is especially developed in humans. It is this hedonic-based feeding that likely contributes the most to the obesity epidemic in humans.

Regulation of Feeding during the Early Postnatal Period in the Rodent

At birth, rodents are relatively undeveloped compared to primates, including the hypothalamic circuitry that controls food intake and energy expenditure (discussed later). During the early postnatal period in the rodent, feeding behavior is thought to be primarily regulated by a combination of maternal behavior and gut distention. Rodent pups spend most of their time sleeping, but grooming stimulation from the dams rouses the pups and initiates nursing. The nursing stimulus initiates kyphosis (a trancelike state) in the dam that allows the pups to feed until their stomachs are full. The increased gut distention activates vagal-gustatory feedback into the dorsal vagal complex in the brain stem to terminate feeding. The brain stem circuitry to the hypothalamus transmits gustatory-vagal sensory information that develops very early in the postnatal period, as do the reciprocal autonomic efferent projections from the PVH to the brain stem that regulate energy expenditure. Rodents get their sole source of nutrients from the dam's milk until

around postnatal day 15–17, when pups initiate independent consumption. At around 21–23 days of age, most rodents in the laboratory setting are weaned from their dams.

Development of ARH Homeostatic Feedback Circuits in the Rodent

The NPY System

Throughout the brain there are numerous populations of NPY neurons that are involved in a variety of functions (reproductive neuroendocrine function, stress responses, fluid balance, learning and memory, etc.). However, NPY is primarily studied as one of the most abundant and potent orexigenic neuropeptides. While NPY neurons are present throughout the brain, the major source of NPY production within the hypothalamus of the normal adult rodent is the ARH-NPY neurons. In specific models of hyperphagia and obesity there is a second population of NPY neurons within the dorsomedial nucleus of the hypothalamus (DMH) that can become activated to play a role in energy homeostasis.

Rodents are born with NPY-expressing neurons within the ARH, with NPY expression in the ARH increasing throughout the postnatal period to peak around postnatal day 15–16. However, during the postnatal period (postnatal days 10–21) there is also expression of NPY in neurons in several other hypothalamic regions that are not present in the normal adult rodent. This includes NPY-expressing neurons in the DMH, both the compact and noncompact zones, PVH, perifornical region (a critical region for NPY-induced food intake), and the lateral hypothalamus. All of these regions are important energy homeostasis areas. The expression of NPY in these regions is low at birth and rapidly increases to peak around postnatal day 15–16 and subsequently decline or disappear before postnatal day 30. The physiological importance of these different populations of hypothalamic NPY neurons during postnatal development, and their role in the regulation of food intake, remains unexplored. It is of interest that the peak expression of NPY in all of these regions coincides with the initiation of independent ingestion. It is possible that NPY is an important signal for motivating maturing pups to search out additional food sources when milk from the dam can no longer provide the proper nutrients or the dam can no longer produce quantities sufficient to support the whole litter.

The role of NPY in the regulation of food intake during the early postnatal period is supported by evidence that central administration of NPY into the PVH as early as postnatal day 2 increases milk consumption in pups. Furthermore, there is an abundance of NPY Y1 receptor gene and protein expression throughout the hypothalamus at birth, including the critical 'feeding centers.' Thus, it makes sense that the NPY system may play an important role in providing the appetitive drive to meet the energy demands of growth and development. It is recognized that NPY is not the sole regulator of food intake during development or in the adult. In fact, mice with mutations of the NPY or NPY receptor genes display near normal growth and development. However, recent studies have shown that while destruction of ARH-NPY neurons early in development allows near normal development, destruction of these neurons in adult rodents causes a wasting disease. This indicates that the absence of NPY during the early postnatal period leads to compensation by other orexigenic systems, but that if present throughout development, the NPY system develops into a critical regulator of food intake in the adult rodent.

The Melanocortin System

The melanocortin system provides one of the most potent anorexigenic signals and is likely the key system for homeostatic feedback regulation of energy balance. Like the NPY system, the main populations of melanocortin neurons (αMSH and AgRP) are located within the ARH. Furthermore, it has been well documented in the adult rodent and primate that AgRP, the endogenous antagonist to the melanocortin system, is exclusively localized to ARH-NPY neurons.

Nearly all of the components of the melanocortin system (αMSH, AgRP, and MC4 receptors) are present in the hypothalamus at birth. However, while there are several populations of NPY neurons in the hypothalamus during postnatal development, only the ARH-NPY neurons co-localize AgRP. Likewise, within the hypothalamus, POMC neurons are also limited to the ARH. αMSH (POMC) neurons are also present in the solitary tract nucleus in the brain stem; however, very little is known about their function or projections. It is well recognized that the melanocortin system is one of the most important regulators of food intake and energy expenditure in adults; this appears to be true in all species from rodents to humans. Genetic deletion or mutations leading to inactivation of the melanocortin system (including the POMC, MC3 or MC4 genes, as well as prohormone converting enzymes) lead to obesity. In humans, mutations of the melanocortin system are the most common genetic cause of obesity. Overexpression of the endogenous melanocortin antagonists (AgRP or agouti) also leads to obesity.

In rodents, deletion of members of the melanocortin system causes a late-onset obesity that does not become evident until after 4 weeks of age. In contrast, in humans, mutations of the melanocortin system cause early-onset obesity, with children developing hyperphagia as infants. The reason for the different obesity ontogeny between rodents and humans likely has to do with the difference in the timing in the development of hypothalamic circuitry (discussed later).

Development of ARH Projections in Rodents

At birth, NPY fibers are present throughout the hypothalamus, but the source of this NPY is unknown. Recent studies have begun to track the sources of the NPY fibers using specific markers for ARH-NPY and brain stem–NPY projections. Prior to postnatal day 5, neither ARH nor brain stem NPY projections are developed, suggesting that local NPY-producing cells in the PVH, DMH, perifornical region, and lateral hypothalamus must be main sources of NPY during the neonatal period. Brain stem–NPY projections to the hypothalamus, which co-express norepinephrine, develop during the latter half of the first postnatal week and likely play a role in the regulation of food intake through modulation of vagal-gustatory feedback, but this remains to be confirmed.

ARH-NPY projections develop during the second and third postnatal week, with projections reaching the DMH around postnatal day 10 and the PVH around postnatal day 15. Other populations of ARH neurons appear to show a similar developmental pattern, but this has not been systematically worked out for all neurons in the ARH. This delayed development of ARH projections has several significant implications: (1) the lack of prominent anorexigenic ARH-αMSH projections to the PVH, combined with the abundance of local NPY, would promote an appetitive drive; (2) postnatal environmental cues, such as nutrition, likely have a significant impact on the development of ARH projections and could have long-term implications on body weight regulation (to be discussed later); and (3) there is not a functional homeostatic feedback circuit during this period.

Ontogeny of Homeostatic Feedback Control of Energy Balance in Rodents

Leptin is the most recognized and studied homeostatic feedback signal. This hormone is produced and secreted from adipose tissue in proportion to adipose mass as a signal to the brain. One of the major sites of action of leptin is the ARH, where leptin inhibits orexigenic NPY/AgRP neurons and activates anorexigenic αMSH neurons. Thus, when body fat stores are high, leptin signals the brain to reduce food intake and increase energy expenditure. When body fat stores are decreasing, such as during weight loss or in response to a fast, the decrease in leptin release is permissive for an increased appetitive drive. The importance of leptin feedback is highlighted by evidence that chronic leptin treatment results in sustained reductions in food intake and weight loss, while mutations of the leptin or leptin receptor genes result in severe obesity (in both rodents and humans).

In the adult rodent, leptin is a potent inhibitor of food intake and stimulator of energy expenditure. However, during the early postnatal period, prior to the development of ARH projections, leptin administration has no effect. What is surprising is that leptin treatment does not become effective at modulating food intake and energy expenditure until the fourth postnatal week, long after the development of ARH projections. This indicates that even though the ARH projections are intact, they are unable to transmit the homeostatic signal related to body fat.

Leptin and Insulin Are Key Signals for the Development of ARH Projections in the Rodent

While leptin does not play a role in the regulation of body weight homeostasis during the early postnatal period, it does play a critical role in the development of ARH projections. In rodents there is an endogenous surge of leptin that occurs during the second postnatal week. What causes this leptin surge is unknown. However, in mouse models that lack functional leptin or leptin receptor genes, ARH circuits fail to fully develop. Without the intact ARH projections the brain loses its major homeostatic feedback regulator and the result is hyperphagia and obesity. Leptin replacement specifically during the second postnatal week in leptin-deficient mice, mimicking the endogenous leptin surge, results in a normal development of ARH projections and partially reverses the obese phenotype in these animals.

While it is evident that the absence of leptin compromises the development of ARH projections, there is also evidence that too much leptin during the postnatal period can also cause abnormalities in hypothalamic function. In rodent models, chronic postnatal overfeeding results in early-onset obesity that persists into adulthood. Accompanying the early-onset obesity is a three- to fourfold increase in leptin levels during the early postnatal period. While there is no evidence to date to indicate that there are abnormalities in the development of ARH projections in animals with hyperleptinemia, there is

clear evidence that the responsiveness of neurons within the hypothalamus to homeostatic signals (leptin and insulin) is permanently altered. Thus, homeostatic feedback regulation of food intake and energy expenditure is likely blunted. Furthermore, in adult animals overfed during the early postnatal period, responsiveness of neurons within the PVH to NPY is dramatically increased, resulting in an increased orexigenic drive.

Insulin, produced from pancreatic β cells, is the key regulator of glucose homeostasis. Insulin receptors are highly expressed in the fetal brain in many species, including rodents and humans, and this expression declines during the postnatal period. Insulin is recognized as a potent neurotrophic factor; however, its role in brain development is not limited to the ARH or the hypothalamus. Both maternal hyperinsulinemia (induced by insulin treatment during pregnancy) and hypoinsulinemia (induced by streptozotocin destruction of the β cells) cause widespread abnormalities in brain development. However, it is difficult to determine what effects are due directly to the insulin, or lack thereof, and what is due to the changes in glucose levels. Genetic deletion of insulin receptors or any of its key downstream signaling components also causes severe abnormalities in overall brain development, such as reduction in neuronal proliferation. Because of the broad range and severity of the effects of knockout of components of the insulin system, it is difficult to discern a specific role in the development of ARH circuits.

Development of ARH Projections in Primates

Most of the neuropeptide and hormonal signals that are important for the regulation of food intake and energy balance in the rodent have also been shown to be important in primates (human and nonhuman), including NPY, AgRP, αMSH, leptin, and insulin. However, there are some basic differences in the neurocircuitry controlling food intake between the rodent and the primate. Within the hypothalamus of the normal adult nonhuman primate there are NPY neurons within the PVH, DMH, and ARH (also referred to as the infundibular nucleus in primates). The NPY neurons within the ARH likely play a role in homeostatic feedback regulation of energy balance similar to that in rodents. The role of NPY neurons in the PVH and DMH is less certain, yet mild metabolic challenges activate the NPY neurons in the PVH, suggesting that they are involved in the regulation of food intake and/or energy expenditure. Another significant difference between rodents and primates is the connections between the brain stem and hypothalamus. The GLP-1 neurons in the brain stem are activated by the vagus in response to gut distention and inhibit food intake through efferent projections to the hypothalamus. In rodents, the PVH and DMH are the major efferent targets of GLP-1 neurons. In contrast, the primary projection of GLP-1 neurons in the nonhuman primate is the ARH. The purpose of the direct interaction between the vagal-gustatory and homeostatic feedback circuits in the primate remains to be determined. Finally, there is a difference in the ontogeny of the brain, including the hypothalamus, between rodents and primates.

In nonhuman primates, NPY, AgRP, and αMSH neurons are evident in the ARH by the end of the second trimester. However, at this early age ARH projections to downstream target sites are sparse. The ARH projections to the DMH, PVH, and other hypothalamic centers develop during the third trimester and appear to be near fully developed by birth. While the development of ARH projections has not been studied in humans, NPY neurons are present in the ARH from the late second trimester fetal brain tissue, suggesting that the ontogeny of these neurons and their projections is likely similar to the nonhuman primate. The development of the ARH circuits during the third trimester suggests that maternal health and diet are likely critical environmental factors. Furthermore, perturbations in energy balance (i.e., over- or undernutrition) or poor maternal health, such as obesity and/or diabetes, during this 'critical period' of neuronal plasticity in the primate could cause abnormal development of this circuit, leading to permanent alterations in body weight management.

See also: Circumventricular Organs; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Hypothalamic Structure–Function Relationships; Sleep and Sleep States: Hypothalamic Regulation.

Further Reading

Flier JS and Maratos-Flier E (1998) Obesity and the hypothalamus: Novel peptides for new pathways. *Cell* 92: 437–440.

Gehlert DR (1999) Role of hypothalamic neuropeptide Y in feeding and obesity. *Neuropeptides* 33: 329–338.

Grove KL, Grayson BE, Glavas MM, Xioa XQ, and Smith MS (2005) Development of metabolic systems. *Physiology and Behavior* 86: 646–660.

Grove KL and Smith MS (2003) Ontogeny of the hypothalamic neuropeptide Y system. *Physiology and Behavior* 79: 47–63.

Horvath T (2005) The hardship of obesity: A soft-wired hypothalamus. *Nature Neuroscience* 8: 561–565.

Kalra SP, Dube MG, Pu S, Horvath TL, and Kalra PS (1999) Interacting appetite-regulating pathways in the hypothalamic regulation of body weight. *Endocrinology Reviews* 20: 68–100.

Rinaman L (2003) Postnatal development of hypothalamic inputs to the dorsal vagal complex in rats. *Physiology and Behavior* 79: 65–70.

Sawchenko PE (1998) Toward a new neurobiology of energy balance, appetite, and obesity: The anatomists weigh in. *Journal of Comparative Neurology* 402: 435–441.

Schwartz MW, Woods SC, Porte D Jr., Seeley RJ, and Baskin DG (2000) Central nervous system control of food intake. *Nature* 404: 661–671.

Smith MS and Grove KL (2002) Integration of the regulation of reproductive function and energy balance: Lactation as a model. *Frontiers in Neuroendocrinology* 23: 225–256.

Swithers SE (2003) Do metabolic signals stimulate intake in rat pups? *Physiology and Behavior* 79: 71–78.

Watts AG (2000) Understanding the neural control of ingestive behaviors: Helping to separate cause from effect with dehydration-associated anorexia. *Hormones and Behavior* 37: 261–283.

Energy Homeostasis: Paraventricular Nucleus (PVN) System

D Richard and E Timofeeva, Laval Hospital Research Center and Laval University, Quebec, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The paraventricular nucleus (PVN), which is located in the dorsorostral hypothalamus adjacent to the third ventricle, represents a major integratory region in the brain with recognized neuroendocrine, autonomic, and behavioral functions. The PVN controls the activity of key hypothalamic–pituitary–target organ systems that include the thyrotropic and corticotropic axes, which are also referred to as the hypothalamic–pituitary–thyroid (HPT) and the hypothalamic–pituitary–adrenal (HPA) axes. The controls exerted by the PVN on the HPT and HPA axes are in large part ensured by cells from the parvicellular neurosecretory system, including the thyrotropin-releasing hormone (TRH) and corticotropin-releasing factor (CRF) neurons. The parvicellular neurosecretory division together with the magnocellular neurosecretory system and the descending division constitute the three major functional parts of the PVN. The magnocellular neurosecretory system, which also has a major neuroendocrine function, hosts neurons that synthesize significant amounts of vasopressin and oxytocin, which are released into the circulation via neuron terminals ending in the posterior pituitary and which control water balance and blood pressure (vasopressin), as well as parturition and lactation (oxytocin). The descending division comprises neurons that innervate the brain stem and spinal cord and coordinates autonomic and behavioral responses.

The PVN, through the complex and integratory involvement of its neuroendocrine, autonomic, and behavioral functions, is in a key position to markedly influence energy homeostasis (or energy balance). Indeed, the parvicellular neurosecretory system, via its control over the HPA and HPT axes, modulates, together with the descending division, the regulation of energy homeostasis, which is achieved through fine-tuned controls exerted on energy intake (food intake) and energy expenditure (thermogenesis).

The PVN and Regulation of Energy Balance

The role played by the PVN in regulation of energy balance and hence of body weight and body fat stores has been acknowledged for years. Experimental lesions of the PVN, as well as *Sim1* haploinsufficiency, which provokes a hypodevelopment of the nucleus, lead to an excessive fat deposition in laboratory rodents. Sim1 is a basic helix–loop–helix Per–Arnt–Sim (bHLH-PAS) transcription factor that has proved to be necessary for the growth of several major cell types in the PVN, including oxytocin-, vasopressin-, TRH-, and CRF-containing neurons. These cells represent key players in the neuroendocrine, autonomic, and behavioral functions of the PVN. Additionally, the PVN is connected with numerous regions involved in the control of food intake and thermogenesis. Specifically, both the parvicellular neurosecretory system and the descending division receive major inputs: (1) those from the brain stem report on visceral signals and on the nutritional status, (2) those from the limbic system convey emotion/reward-related signals, and (3) those from the arcuate nucleus (ARC) plus the surrounding hypothalamic nuclei redirect to the rest of the brain information about the nutritional status and the state of energy stores. The PVN is rich in appetite- and thermogenesis-modifying neuropeptides/neurotransmitters, which exert their effect when directly microinjected within its major divisions. The list of these neuropeptides/neurotransmitters includes CRF, oxytocin, vasopressin, TRH, neuropeptide Y (NPY), melanocyte-stimulating hormone (α-MSH), galanin, cocaine- and amphetamine-regulated transcript (CART), opioid peptides, histamine, serotonin, and noradrenaline.

Data from studies carried out in PVN-lesioned laboratory rodents or in *Sim1* haploinsufficient ($Sim^{+/-}$) mice tend to suggest that the PVN participates in energy balance regulation mainly by controlling energy intake. PVN-lesioned and $Sim^{+/-}$ animals show hyperphagic behaviors without exhibiting marked changes in energy expenditure. However, one cannot exclude some role for the PVN in the control of energy expenditure. Selective excitation of the PVN and PVN microinjections of thermogenesis-inducing substances stimulate heat production in small laboratory animals. A role for the PVN in the control of energy expenditure is also supported by compelling neuroanatomical evidence associating the nucleus to the sympathetic innervation of brown adipose tissue (BAT), which represents an extraordinary thermogenic effector in small mammals. BAT thermogenesis is referred to as adaptive or facultative thermogenesis, as it may occur as a mechanism to increase energy expenditure in the sole purpose of regulating energy balance.

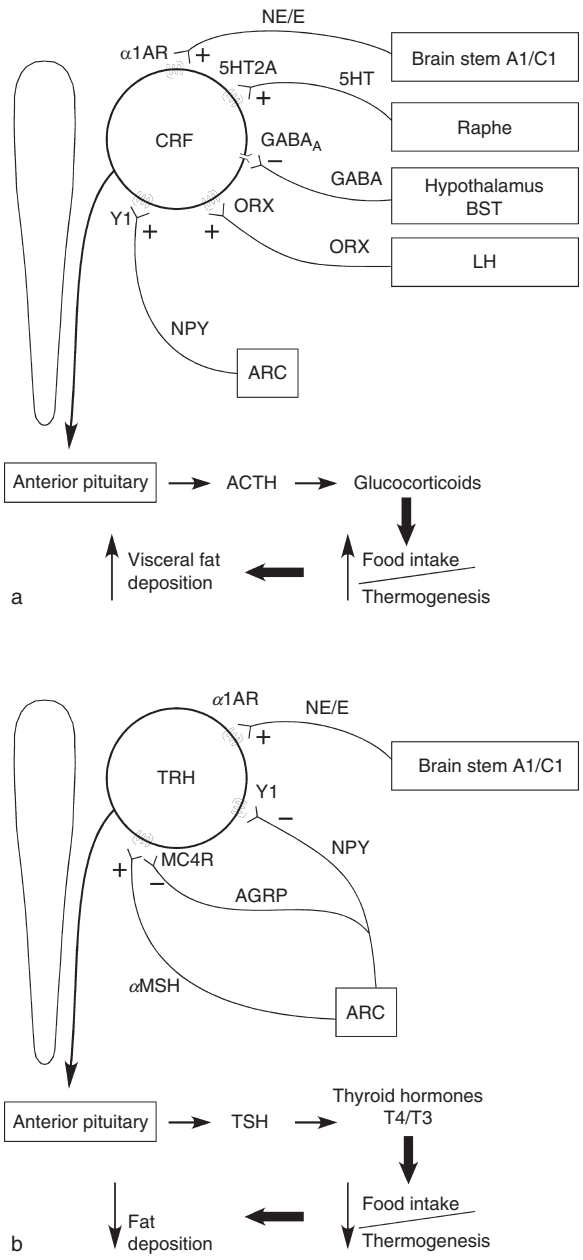

Figure 1 The neuroendocrine paraventricular nucleus and the regulation of energy balance. The paraventricular nucleus (a) corticotropin-releasing factor (CRF) and (b) thyrotropin-releasing hormone (TRH) cells project to the anterior pituitary, where they control the release of glucocorticoids and thyroid hormones. Glucocorticoids enhance visceral fat deposition by increasing food intake and reducing energy expenditure, whereas thyroid hormones reduce fat deposition by decreasing the food intake/thermogenesis ratio. The parvicellular neurosecretory division that contains the CRF and TRH cells is highly innervated by neurons originating from regions strategically positioned to relay signals informing about nutritional status and energy homeostasis. ACTH, adrenocorticotropic hormone; AGRP, agouti-related protein; AR, adrenergic receptor; ARC, arcuate nucleus; BST, bed nucleus of the stria terminalis; E, epinephrine; GABA, γ-aminobutyric acid; 5HT, 5-hydroxytryptamine; LH, lateral hypothalamus; MC4R, melanocortin receptor 4; MSH, melanocyte-stimulating hormone; NE, norepinephrine; NPY, neuropeptide Y; ORX, orexins; OXR, orexin receptors; T3, triiodothyronine; T4, thyroxine; TRH, thyrotropin-releasing hormone; TSH, thyroid-stimulating hormone; Y1, NPY receptor 1.

The Neuroendocrine PVN and Energy Balance Regulation

The parvicellular neurosecretory system of the PVN governs two major neuroendocrine entities, namely the HPA and the HPT axes, which are of particular importance in the regulation of energy balance (Figure 1). The HPA and HPT axes each comprise distinctive neuroendocrine cells located in the parvicellular division of the PVN; these cells control the activities of the axes. The PVN parvicellular CRF neurons, also called the hypophysiotropic CRF cells, govern the HPA axis by commanding the pituitary release of the adrenocorticotropic hormone (ACTH), which stimulates the secretion of corticosteroids from the adrenal cortex via the circulation. Similarly, a selective population of parvicellular TRH neurons controls the activity of the HPT axis and thereby the release into circulation of the thyroid hormones. The activation of the HPT axis is catabolic (decrease in energy/fat deposition), whereas that of the HPA axis generally increases energy/fat deposition.

The HPA axis activity is primarily stimulated by stressful conditions, which excite PVN CRF cells. When stimulated, the HPA axis, which is also referred to as the stress axis, causes the release of corticosteroids (cortisol in human and nonhuman primates; corticosterone in small mammals such as laboratory rodents). In humans as well as in laboratory rodents, excess in corticosteroids is generally associated with visceral/intra-abdominal fat retention. This phenotype is remarkable in patients with Cushing's syndrome, who are not only diagnosed with high circulating levels of corticosteroids but also are characterized by an excessive waist circumference and its associated deleterious health consequences.

The role played by the HPA axis in the regulation of energy balance has further been emphasized in several animal models of obesity, which include the Zucker *fa/fa* rat, a leptin-resistant rat due to a mutation in the gene encoding the long form of the leptin receptor (Ob-Rb), and the *ob/ob* mouse, a leptin-deficient mouse as a result of a mutated leptin gene. In both models, which display hyperactive HPA axes, disruption of the HPA axis through adrenalectomy prevents, blocks, or cures excessive fat deposition by removing corticosterone from the circulation. In laboratory rodents, corticosteroids enhance fat deposition by stimulating energy intake and inhibiting

BAT thermogenesis. Corticosterone is in fact a strong repressor of uncoupling protein 1 (UCP1). UCP1, the activity of which is largely dictated by the sympathetic nervous system (SNS), is a mitochondrial inner membrane protein that confers BAT with an immense thermogenic power. The SNS stimulation of UCP1 uncouples fuel oxidation from ATP synthesis by dissipating the proton gradient that builds up across the mitochondrial inner membrane during the transport of electrons through the mitochondrial respiratory chain during the process of oxidative phosphorylation. In uncoupled BAT adipocytes, ATP production is markedly reduced, which releases a major constraint on energy oxidation and therefore allows BAT mitochondria, upon SNS stimulation, to oxidize fat at an accelerated rate and to produce a considerable amount of heat.

The activity of the HPA axis is modulated by fluctuations in energy stores and by the nutritional status. In laboratory rodents, food restriction/deprivation is associated with high circulating corticosterone levels that result, in most circumstances, from a reduced plasma clearance of the hormone. In fact, in lean rats, food deprivation elevates corticosterone levels while it reduces CRF synthesis within the PVN CRF hypophysiotropic cells. However, in some models of obesity, elevation of corticosteroid levels associated with food restriction/deprivation appears centrally triggered. In the obese Zucker rat, for instance, acute food deprivation leads to a stimulation of the PVN CRF neurons, which readily increases the activity of the pituitary–adrenal complex and the release of corticosterone into the circulation. In this model, the pattern of brain activation elicited by food deprivation shows similarities with the pattern seen following neurogenic (interoceptive or psychological) stresses. In the Zucker rat, food deprivation seems to represent one of the most powerful stimulators of the PVN CRF cells of the HPA axis. This enhanced activity of the HPA axis is immediately suppressed upon eating, revealing the importance of the nutritional status in controlling the activity of the PVN CRF neurons, at least under certain circumstances. The nutritional status also appears to control the HPA axis activity in humans, in whom reductions in cortisol secretion and depressive mood after stress have been reported following the ingestion of a high-carbohydrate diet. There is evidence, in both humans and rodents, that stress may enhance the intake of specific palatable foods, which could be energetically beneficial, rewarding, and potentially 'comforting.'

Alternatively, severe stress has often been associated with anhedonia and anorectic effects. These effects have been attributed to an enhanced activation of the CRF system, which is incidentally not restricted to the population of hypophysiotropic cells found in the PVN, since CRF cells are widely distributed throughout the brain. CRF, through its two G-coupled receptor subtypes 1 and 2, which are also expressed in several brain regions, is capable of a very strong anorectic effect. The brain site for this effect does not appear to be the PVN and remains to be firmly determined. The CRF-mediated anorectic effect of stress does not therefore involve the PVN hypophysiotropic CRF neurons. These cells, which project to the external layer of the median eminence, where they release CRF through numerous terminal boutons into the proximal capillary plexus of the hypothalamic–hypophysial portal system, appear to fulfill a rather selective stimulatory role on corticotrophs. The action of the PVN CRF hypophysiotropic neurons on energy homeostasis is mainly mediated through the HPA axis and hence the release of corticosteroids. Yet, one should bear in mind that the control of the PVN CRF neurons is complex, being typically integrated into a stress response that superimposes over numerous neuronal circuitries that may individually influence food intake. The PVN receives major inputs from the brain stem and hypothalamic nuclei, which release peptides/neurotransmitters such as α-MSH, NPY, orexins, and catecholamines, whose action are mediated through various receptor types located on hypophysiotropic CRF cells, within the parvicellular neurosecretory system.

The HPT axis, which is also under the control of the neuroendocrine PVN, exerts nontrivial effects on energy homeostasis, most probably through the release of thyroid hormones. The activity of this axis is governed by TRH neurons found in the parvicellular neurosecretory division. These neurons, which seemingly also express other peptides, such as CART, project to the external zone of the median eminence, where they are ultimately responsible for the pituitary release of thyroid hormones. These hormones have proved to be of significant influence in the control of energy expenditure, generating up to 30% of the resting metabolic rate. The exact mechanism whereby thyroid hormones act to stimulate thermogenesis has yet to be fully elucidated. In small laboratory rodents, thyroid hormones modulate SNS-mediated BAT adaptive thermogenesis. There is also evidence that thyroid hormones enhance heat production through a regulated mitochondrial proton leak, which has yet to be described in terms of its functionality.

The PVN TRH cells receive major inputs from the brain stem and other hypothalamic nuclei, principally from the ARC. These inputs, as is the case for the CRF hypophysiotropic cells, are capable of

modulating the activity of the TRH cells, which are very sensitive to energy store oscillations. It has been acknowledged for years that a state of negative energy balance, such as that generated by fasting, strongly represses the activity of the TRH cells and downregulates TRH mRNA expression within the hypophysiotropic TRH neurons selectively. These effects lead to a significant reduction of HPT axis activity and hence to a reduction in TRH-mediated thermogenesis. Deprivation-induced hypothyroidism may be an important adaptive response that reduces energy expenditure in an attempt to minimize energy loss. This adaptation to energy shortage, which is achieved in concomitance with an increased activity of the HPA axis, further substantiates the significant implication of the neuroendocrine PVN in the regulation of energy balance.

The PVN Descending Division and Energy Balance Regulation

The descending division of the PVN consists of the dorsal and ventral medial parts of the structure that send projections to the brain stem dorsal vagal complex (DVC) and to the preganglionic autonomic neurons of the spinal cord. The DVC comprises the area postrema, nucleus of the tractus solitarius (NTS), dorsal motor nucleus of the vagus nerve. The descending division, which also innervates the parabrachial nucleus, periaqueductal gray, and ventral medulla, as well as nearby areas, receives, from the ARC and the brain stem, major inputs integrating signals informing about energy homeostasis. Other projections to the descending division include those from the dorsomedial hypothalamus (DMH), ventromedial hypothalamus (VMH), medial preoptic area (MPOA), and bed nucleus of the stria terminalis (BNST).

The PVN descending division consists of a heterogeneous population of cells with different morphology and physiological properties. It mainly comprises oxytocin and vasopressin neurons; the former are mostly located in the dorsal parvicellular subdivision, whereas the latter are predominantly found in the ventral part of the medial parvicellular subdivision. The subset of oxytocin neurons projecting to the brain stem is thought to be of particular significance in the role played by the PVN in the control of food intake and energy expenditure. This group extends to the DVC and to the intermediolateral (IML) column cells of the spinal cord and receives an important contingent of axons from the DVC capable of conveying visceral signals informing about energy stores and nutritional status. It is probable that oxytocin neurons can also influence thermogenesis, as they functionally connect to BAT. The use of retrograde transsynaptic tracers has indeed highlighted a functional anatomical link between the oxytocin neurons of the descending division and BAT. Oxytocin neurons connect to the IML in the spinal cord, which comprises preganglionic sympathetic neurons that project to the stellate ganglion, from which originates postganglionic SNS neurons innervating BAT adipocytes of the interscapular depot. It is noteworthy that the group of preganglionic autonomic PVN neurons includes cells that are also immunoreactive to CRF, TRH, somatostatin, met-enkephalin, and dynorphin. Some PVN CRF neurons projecting to the spinal cord have been identified both in the ventral part of the medial parvicellular subdivision and in the medial posterior parvicellular part of the PVN.

The ARC–PVN Pathway in Energy Balance Regulation

The PVN receives major inputs from the ARC, which currently is one of the most investigated hypothalamic nuclei in connection with energy homeostasis. The ARC is a tiny nucleus found in the basomedial hypothalamus just above the median eminence and adjacent to the third ventricle, in a position slightly caudal to the PVN. It comprises two populations of neurons strongly involved in the control of energy intake and energy expenditure, and capable of integrating peripheral signals that affect energy homeostasis. One population of neurons synthesizes proopiomelanocortin (POMC), whereas the other synthesizes NPY and agouti-related peptide (AGRP). Stimulation of POMC neurons, which also co-localize CART, is associated with a decrease in food intake and an increase in energy expenditure, whereas the activation of NPY/AGRP neurons leads to opposite effects. The POMC and NPY/AGRP neurons innervate both the descending division and the parvicellular neurosecretory area of the PVN, through which they mostly act to influence energy homeostasis (Figure 2).

POMC neurons exert their catabolic effect mainly via α-MSH, a peptidergic fragment resulting from POMC cleavage. Within the brain, α-MSH binds to melanocortin 3 and melanocortin 4 receptors (MC3R, MC4R), with which it forms, together with AGRP, the 'homeostatic' melanocortin system. The functional significance of both the MC3Rs and MC4Rs in energy homeostasis has been validated in *Mc3r* and *Mc4r* knockout mice. The ablation of the gene encoding MC3Rs enhances visceral fat deposition, whereas the targeted disruption of MC4Rs causes a massive and widespread body fat deposition. In

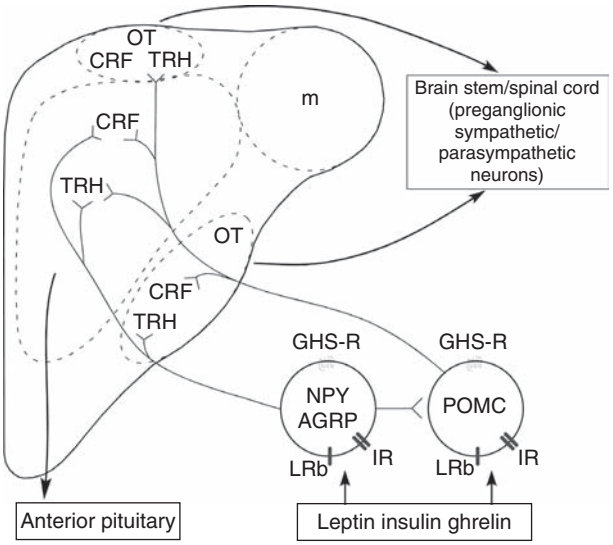

Figure 2 The arcuate nucleus–paraventricular nucleus pathway. The paraventricular nucleus receives an important contingent of fibers from the arcuate nucleus, which comprises two populations of neurons markedly involved in the regulation of energy balance. These neurons produce neuropeptide Y (NPY), agouti-related protein (AGRP), and proopiomelanocortin (POMC), three peptides capable of significantly modulating the effects of the neuroendocrine and autonomic/behavioral paraventricular nucleus on energy balance. The arcuate nucleus expresses receptors that mediate the effect of blood-borne hormones on energy balance. CRF, corticotropin-releasing factor; GHS-R, growth hormone secretagog receptor; IR, insulin receptor; LRb, leptin receptor, long form; m, magnocellular division of the paraventricular nucleus; OT, oxytocin; TRH, thyrotropin-releasing hormone.

humans, the MC4R deficiency represents the most commonly known monogenic form obesity; up to 2–3% of severely obese patients apparently have pathogenic MC4R mutations. The excessive fat deposition occurring in *Mc4r* knockout mice has been found to result from an increase in energy intake and a decrease in energy expenditure. AGRP is an endogenous antagonist of the melanocortin receptors that is solely expressed in the ARC neurons, where it is co-synthesized with NPY. AGRP is overexpressed in obese mice. When injected centrally, it increases food intake and reduces energy expenditure, similar to synthetic melanocortin receptor antagonists.

Unlike the MC3R, which is virtually absent from the PVN, the MC4R is expressed in the three major parts of the structure. Microinjections of MC4R ligands into the PVN reduce food intake, without apparently causing aversive effects. The role of the PVN in the anorectic effect mediated by the MC4R is well accepted. It has recently been demonstrated that the PVN reactivation by Cre-recombinase of the gene encoding the MC4R may correct the hyperphagia seen in *Mc4r* null allele mice (loxTB *Mc4r* mice). Reactivation of *Mc4r* in loxTB *Mc4r* mice was achieved by crossing these mice with Sim1–Cre transgenic mice to allocate the reexpression of *Mc4r* selectively in Sim1-expressing cells, which are mostly found in the PVN.

The neural circuitry involving the PVN MC4R in the control on food intake is currently largely unknown. Several MC4R mRNA-containing cells are located throughout the dorsal and ventral parvicellular parts of the PVN that constitute the descending division. It remains to be firmly established whether these neurons are clearly those that are responsible for the MC4R-mediated control of food intake. It is also not entirely clear whether the PVN MC4R plays a major role in controlling energy expenditure, even though a large proportion of the PVN neurons that are associated with the sympathetic outflow to white adipose tissue and BAT express the MC4R. The recent observation that the rather selective PVN reactivation of the *Mc4r* gene did not rescue the blunted thermogenesis observed in loxTB *Mc4r* mice cast doubts about the real influence of the PVN MC4R in the control of thermogenesis – this is despite the demonstration that microinjection of MC4R agonists within the PVN increases energy expenditure.

Alternatively, the PVN MC4R could influence energy homeostasis through the parvicellular neurosecretory division. There indeed exists a strong anatomical and functional relationship between the central melanocortin system and the PVN TRH neurons. A majority of the hypophysiotropic TRH neurons are dually innervated by POMC and NPY/AGRP neurons from the ARC. MC4R agonism activates the TRH system while it blocks fasting-induced repression of TRH synthesis in the PVN. Also, energy shortage is known to stimulate the expression of NPY and AGRP, which are strong inhibitors of the HPT axis. It is noteworthy that the number of AGRP fibers reaching the vicinity of the TRH neurons exceeds that of α-MSH-containing terminals, which is in line with the observation that AGRP accounts for the variations in HPT axis activity linked to oscillating energy stores.

The PVN also expresses Y1 and Y5 receptors, which are recognized as the main receptors mediating the action of NPY on energy balance. NPY represents a potent orexigenic peptide, the expression of which in the ARC is markedly stimulated in obese and food-deprived animals. In the PVN, Y1 and Y5 receptor mRNAs are expressed in a pattern very similar to that of the MC4R. NPY injected into the PVN not only stimulates food intake but also inhibits thermogenesis by BAT. The functional importance of NPY, as well as that of AGRP, in the regulation of energy balance has

recently finally found a genetic endorsement when mice genetically engineered for a selective ablation at the adult stage of the AGRP/NPY-expressing neurons were shown to exhibit an acute reduction in feeding. The selective ablation of these neurons leads to a significant reduction in the NPY-immunostained axons and terminals in the PVN, further emphasizing the relevance of ARC–PVN pathway in the regulation of energy balance. Ablation of the AGRP/NPY-expressing neurons is seemingly ineffective in neonates, in which compensatory adaptations may occur to rescue the lack of NPY/AGRP cells. It is noteworthy that conventional inactivation of *Npy*, *Npyr1*, *Npyr5*, and *Agrp* genes is largely inefficacious to influence body weight regulation in expected ways.

The neuronal pathways involved in the PVN Y1 and Y5 receptor-mediated effects on food intake and energy expenditure are not fully understood. These receptors are expressed in cells of the dorsal and ventral parvicellular parts of the PVN that host descending division neurons through which NPY may affect the control of food intake and energy expenditure. NPY could also exert a PVN-mediated effect on energy balance through its influence on the parvicellular neurosecretory division. NPY stimulates the HPA axis at the PVN level by stimulating CRF synthesis in hypophysiotropic CRF neurons. The elevated expression of NPY in obese and food-deprived animals has been associated with the stimulated HPA axis activity seen in those animals, which also exhibit a reduced HPT activity involving Y1 and Y5. This control exerted by NPY/AGRP neurons on the neurosecretory parvicellular system of the PVN further illustrates the importance of the neuroendocrine part of PVN in energy homeostasis.

The role played by the ARC–PVN pathway in the regulation of energy balance is modulated by adipostatic signals such as leptin and insulin, as well as by acute satiety/hunger signals such as PYY and ghrelin. These hormones most probably act at the level of the ARC to modify the PVN regulation of energy balance, even though direct actions at the levels of the PVN or indirect action through the brain stem cannot be excluded. The ARC NPY/AGRP and the POMC/CART neurons are sensitive to the action of leptin, insulin, ghrelin, and PYY.

The adipocyte-derived hormone leptin markedly decreases food intake and fat stores when injected in leptin-deficient humans or leptin-deficient *ob/ob* mice. Leptin acts in the brain, where it is actively transported and where it binds to its long-form receptor (Ob-Rb). It exerts its catabolic action mainly (not exclusively) at the level of the hypothalamus, where, through the STAT3 signaling cascade, it reduces the production of NPY and AGRP while stimulating synthesis of POMC. Importantly, leptin also acts early in life as a trophic signal that stimulates the ARC axon outgrowth to the PVN, lateral hypothalamus, and DMH. The effect of leptin on brain plasticity also translates into acute increases in the ratio of inhibitory synapses to the number of excitatory synapses on the ARC NPY and in the ratio of number of excitatory synapses to number of inhibitory synapses on the ARC POMC. Likewise, recent studies have highlighted the importance of adenosine monophosphate-activated protein kinase (AMPK) in the anorectic effect of leptin. AMPK is seen as a gauge for the cellular energy status and its activity is downregulated in the ARC and PVN following leptin administration; melanocortin receptor (MCR) agonism and MCR antagonism, respectively, decrease and increase AMPK activity in the PVN. Leptin also blunts the activity of the HPA axis in leptin-deficient *ob/ob* mice through a suppression of the activity of the PVN CRF hypophysiotropic neurons. Even though it is more and more certain that the effects of leptin on the PVN are mostly exerted through the ARC, the possibility nonetheless remains that the hormone can also act at the levels of the PVN as the nucleus synthesizes the Ob-Rb.

A central role for insulin in the regulation of energy balance has been advocated. Central administration of insulin reduces food intake and stimulates thermogenesis, and mice with a neuron-specific disruption of the insulin receptor gene develop diet-sensitive obesity. Similar to leptin, insulin has been reported to reduce the activity of NPY/AGRP neurons and to stimulate the POMC cells. The effect of insulin on the ARC–PVN pathway is in a certain way comparable to that of leptin.

The gastrointestinal hormones peptide tyrosine-tyrosine (PYY) and ghrelin also influence the activity of the ARC–PVN pathway. PYY, which has two endogenous forms, $PYY_{1–36}$ and $PYY_{3–36}$, acts at the level of the ARC to exert catabolic effects through the Y2 receptor via a mechanism that has yet to be fully delineated. $PYY_{3–36}$ reduces food intake in MC4R and POMC knockout mice, which suggests that the melanocortin system is not essential to the action of the peptide. This observation does not preclude an ARC-mediated action of PYY, since $PYY_{3–36}$ increases neuronal activity in the ARC and since the selective hypothalamic deletion of Y2 receptor, upon which acts PYY, alters the expression of both POMC and AGRP in the ARC. In contrast to PYY, ghrelin stimulates food intake. The orexigenic action of this peptide, which is abundantly produced in the

stomach, is mediated through the growth hormone secretagog receptor, which is found on the NPY/AGRP neurons that project to the PVN. Peripheral administration of ghrelin induces c-*fos*, a neuronal activity marker, in both the ARC in the PVN, suggesting that ghrelin, like leptin, insulin, and PYY, produces ARC effects that may ultimately influence the descending division and the parvicellular neurosecretory system of the PVN.

Conclusion

The PVN is an integratory region that plays a major role in the regulation of energy balance. Through its parvicellular neurosecretory system and through its descending division, this hypothalamic nucleus is indeed capable of significant controls on both energy intake and thermogenesis. The parvicellular neurosecretory system comprises the neuroendocrine TRH and CRF cells, the metabolic roles of which are exerted through the HPA and HPT axes. The activities of these axes in turn are highly coordinated with those of the neurons of the descending division that govern ingestive and other complex behaviors, as well as autonomic controls exerted on the cardiovascular, respiratory, and thermoregulatory/metabolic functions. The parvicellular neurosecretory system and the descending division are highly innervated by neurons originating from regions strategically positioned to relay signals informing about the nutritional status and energy homeostasis. Those regions received projections from the ARC, a prominent player in the regulation of energy balance. The ARC sends the PVN an important contingent of fibers, the terminals of which release in the PVN NPY, AGRP, and α-MSH, three peptides capable of significantly modulating the role of the neuroendocrine and autonomic/behavioral PVN on energy balance. The ARC harbors receptors, which mediate the effect of blood-borne hormones on energy balance.

See also: Blood Pressure: Baroreceptors; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Energy Homeostasis: Visceral Control; Energy Homeostasis: Hypothalamic Development; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Hypothalamo–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH; Motor Control of Feeding and Drinking; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress, the HPA Axis and Depressive Illness; Superior Colliculus.

Further Reading

Balthasar N, Dalgaard LT, Lee CE, et al. (2005) Divergence of melanocortin pathways in the control of food intake and energy expenditure. *Cell* 123: 493–505.

Bartness TJ, Kay Song C, Shi H, et al. (2005) Brain–adipose tissue cross talk. *Proceedings of the Nutrition Society* 64: 53–64.

Bray GA and York DA (1998) The MONA LISA hypothesis in the time of leptin. *Recent Progress in Hormone Research* 53: 95–117, discussion 117–118.

Cannon B and Nedergaard J (2004) Brown adipose tissue: Function and physiological significance. *Physiological Reviews* 84: 277–359.

Caqueret A, Yang C, Duplan S, et al. (2005) Looking for trouble: A search for developmental defects of the hypothalamus. *Hormone Research* 64: 222–230.

Cone RD (2005) Anatomy and regulation of the central melanocortin system. *Nature Neuroscience* 8: 571–578.

Dallman MF, Pecoraro NC, and la Fleur SE (2005) Chronic stress and comfort foods: Self-medication and abdominal obesity. *Brain, Behavior, and Immunity* 19: 275–280.

Elmquist JK, Coppari R, Balthasar N, et al. (2005) Identifying hypothalamic pathways controlling food intake, body weight, and glucose homeostasis. *Journal of Comparative Neurology* 493: 63–71.

Gao Q and Horvath TL (2007) Neurobiology of feeding and energy expenditure. *Annaul Review of Neuroscience* 30: 367–398.

Lechan RM and Fekete C (2006) Role of melanocortin signaling in the regulation of the hypothalamic–pituitary–thyroid (HPT) axis. *Peptides* 27: 310–325.

Oldfield BJ, Giles ME, Watson A, et al. (2002) The neurochemical characterisation of hypothalamic pathways projecting polysynaptically to brown adipose tissue in the rat. *Neuroscience* 110: 515–526.

Richard D, Lin Q, and Timofeeva E (2002) The corticotropin-releasing factor family of peptides and CRF receptors: Their roles in the regulation of energy balance. *European Journal of Pharmacology* 440: 189–197.

Saper CB, Chou TC, and Elmquist JK (2002) The need to feed: Homeostatic and hedonic control of eating. *Neuron* 36: 199–211.

Sawchenko PE, Li HY, and Ericsson A (2000) Circuits and mechanisms governing hypothalamic responses to stress: A tale of two paradigms. *Progress in Brain Research* 122: 61–78.

Swanson LW (1986) Organization of mammalian neuroendocrine system. In: Mountcastle VB, Bloom FE, and Geiger SR (American Physiological Society) (eds.) *Handbook of Physiology, The Nervous System*, pp. 317–362. Baltimore: Waverly Press.

Timofeeva E and Richard D (2001) Activation of the central nervous system in obese Zucker rats during food deprivation. *Journal of Comparative Neurology* 441: 71–89.

Energy Homeostasis: Thermoregulation

K J Kaiyala, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Mammalian and avian core body temperatures (T_{core}) are generally stabilized within impressively narrow limits by intricate and highly adaptive biobehavioral control systems. This stability occurs, often despite marked fluctuations in environmental and metabolic conditions, because powerful behavioral and autonomic effector responses are recruited in the face of anticipated or actual disruptions of T_{core} and/or skin temperature (T_{skin}).

Thermoregulatory strategies fall into two categories, behavioral responses and reflexive autonomic responses. Much of thermoregulation in higher mammals (almost all in humans) is accomplished via behavior, the neural control of which is barely understood. Reflexive autonomic thermoregulatory responses are generated in accordance with the principles of negative feedback control wherein sufficient disruptions of T_{core} and/or T_{skin} are sensed by the central nervous system (CNS) and translated into corrective responses via neuronal and peripheral effectors of body temperature. Accordingly, the recruitment of autonomic thermoeffector responses alters the balance between the rates of body heat production (HP) and heat loss (HL) so as to favor corrective changes of body temperature; to a reasonable approximation, the addition or subtraction of 1 cal of heat energy changes the temperature of 1 g of body mass by 1 °C.

Much progress has been made in understanding autonomic mechanisms of thermoregulation. This article is focused on autonomic effectors of HP and HL in terms of four simplified models designed to identify and synthesize some of the important concepts and anatomical substrates of modern thermoregulatory theory. Each model represents the efferent limb of a particular, relatively independent thermoregulatory control loop, one each for the control of nonshivering thermogenesis, shivering thermogenesis, cutaneous blood flow, and sweating. This organization reflects an overarching theoretical framework according to which multiple, relatively independent control loops govern each of the effector responses that promote HP and HL. The models trace their neuronal protocols to Jack Boulant and owe their inspiration and general architecture largely to Evelyn Satinoff, Shawn Morrison, Kazuyki Kanosue, and William Blessing, and their many fine colleagues. The models embrace the concept that thermoregulation is governed by a diverse set of relatively independent control loops, as opposed to a unitary central integrator which generates functionally coordinated efferent responses. The independence-of-control-circuits-hypothesis was initially propounded by Satinoff in the late 1970s and early 1980s, and is nicely compatible with the Boulant neuronal model. The independent-loops concept has been emphasized recently by Andrej Romanovsky, Kanosue, and other prominent thermoregulatory investigators.

The models emphasize a role for warm-sensitive neurons (which increase their firing rate with increasing temperature) as the thermal integrators involved in the establishment of both the lower and upper body temperature thresholds for the activation of corrective effector responses. The schema presented here invoke three classes of warm-sensitive neurons, one with a firing rate near the upper limit at normal T_{core}, one with a firing rate intermediate between the upper and lower limits at normal T_{core}, and one with a firing rate near the lower limit at normal T_{core}. All three classes have been observed experimentally. The concepts emphasized in this article are congruent with a wealth of experimental data and avoid some vexing theoretical problems faced by traditional set-point models and by models assuming a great deal of reciprocal inhibition between pathways for HP and HL.

Heat Production

Metabolic HP in homeotherms can be divided into obligative and facultative components. Obligative HP corresponds to the basal metabolic rate (BMR), a largely theoretical construct corresponding to the minimum metabolic rate obtained under ideal homeostatically benign conditions in unstressed, mildly fasted, and resting individuals. BMR is technically problematic, and is often replaced by resting metabolic rate (RMR), a slightly higher value measured under standardized conditions (in humans, RMR is the value obtained in a supine individual at rest in a thermoneutral environment, following an 8–12 h fast). BMR (and thus RMR) is sufficient for thermal homeostasis during exposure to ambient temperatures in the thermoneutral zone (TNZ), which, in fact, is often defined as the range of ambient temperatures in which metabolic rate remains at its minimal level (~28–32 °C in rats, ~25–27 °C in humans, corresponding to a BMR of ~70 and 17.5 cal kg^{-1} min^{-1}, respectively). In fact, in the TNZ, both HP

and evaporative heat loss (EHL) are at basal or resting values, and T_{core} is stabilized at the usual defended level (typically ~37 °C in advanced mammals) solely by vasomotor responses. Facultative thermogenesis refers to the increase of HP that occurs when ambient temperature decreases below the lower limit of the TNZ. Autonomic facultative thermogenic mechanisms include nonshivering thermogenesis (NST) and shivering thermogenesis (ST).

Homeotherms also employ behavioral means of increasing metabolic HP, exploiting the modest mechanical efficiency of muscular work (~10–20%) to obtain the heat. This strategy is certainly familiar to readers with outdoor experience, such as legions of hikers, campers, and skiers.

Nonshivering Thermogenesis

NST is an increase of metabolic rate that is not mediated by increased muscular contractile activity. It is customary to define NST as the increase in metabolic heat production above the normal basal level in response to cold exposure. In small mammalian species (e.g., mice and rats), NST involves a remarkable, specialized heat-producing organ, brown adipose tissue (BAT), while in larger species (e.g., adult humans) the role of BAT in NST is contentious. BAT is an organ both specialized to produce heat and vested with a substantial adaptive ability to increase its thermogenic capacity in the face of relevant homeostatic challenges, including cold exposure and chronic caloric excess. Indeed, BAT is a critical organ of survival in conditions of cold exposure in small mammals such as rats and mice, which have a 'dangerously' large surface area:volume ratio. BAT is also essential to thermal homeostasis for neonatal rodent species that are poorly insulated (lacking fur) and largely incapable of shivering at birth. It should be noted that BAT thermogenesis largely or completely replaces shivering as the major mode of heat production during cold acclimation in rodents. BAT is strategically located in the intrascapular, cervical, pericardial, intercostal, and perirenal regions. BAT is highly vascularized, and the vascular architecture is optimized to facilitate heat transfer from BAT to key anatomical areas, including the spinal cord and heart. BAT has an impressive thermogenic capacity, and can generate over 5500 cal kg^{-1} min^{-1} under experimental conditions. BAT is regarded as an important organ of thermoregulation in infant humans, but has been mostly dismissed as a meaningful participant in adult thermoregulation. More recently, however, the concept that BAT may play a meaningful role in adult human thermoregulation (and in body fat regulation) has come to be regarded more favorably by some investigators. It is interesting to note that BAT has been shown to be essential to obesity resistance in various rodent models of body fat regulation, including high-fat feeding in mice.

BAT achieves its remarkable thermogenic capability by way of a high density of large mitochondria which express in their inner membranes a unique guanosine diphosphate (GDP)-binding protein, uncoupling protein 1 (UCP1). This protein prevents the electrical potential energy in the mitochondrial proton gradient (generated by fuel oxidation and the electron transport chain) from being transferred (coupled) into the high-energy chemical phosphate bonds in adensosine triphosphate (ATP), the obligatory currency for most energy-requiring processes in the body. Instead, UCP1 allows protons to leak down their gradient into the mitochondrial matrix such that potential energy is transferred into heat energy rather than high-energy chemical bonds in ATP.

Resting energy expenditure and autonomic adaptive thermogenesis (activation of BAT and perhaps other UCP-expressing tissues, as well as ST) are regulated chiefly by the sympathetic nervous system (SNS). The SNS–BAT–UCP1 axis is relatively well understood in rats, mice, and rabbits, with many of the recent conceptual advances stemming from obesity-related research. NST is also proposed to occur to a lesser extent in other tissues, such as liver and muscle, but several aspects of this proposal await further data for clarification. In humans, NST appears to be chiefly dependent upon thyroid hormones. Interestingly, some neurons in the brain appear to be capable of releasing heat, and this phenomenon has been experimentally manipulated by generating mice that overexpress the putative mechanism for this neuronal heat production, an uncoupling protein analogous to the UCP isoform described for BAT.

Model for the Control of BAT Thermogenesis

Figure 1 illustrates a simplified scheme for the regulation of BAT thermogenesis by thermal and nonthermal factors. The scheme includes a parsimonious subset of the many brain areas implicated in BAT regulation, sufficient to be congruent with some important properties of thermoregulation, including alteration of thermogenic responses during fever, the fact that stress evokes a substantial increase in T_{core} (often called 'stress hyperthermia'), and the facts that feeding and adiposity significantly modulate T_{core} and thermogenic responses (e.g., states of caloric excess promote thermogenesis whereas states of caloric deficiency inhibit thermogenesis and decrease mean daily T_{core}). A key area is the preoptic anterior hypothalamus (POAH), a brain region critical to the

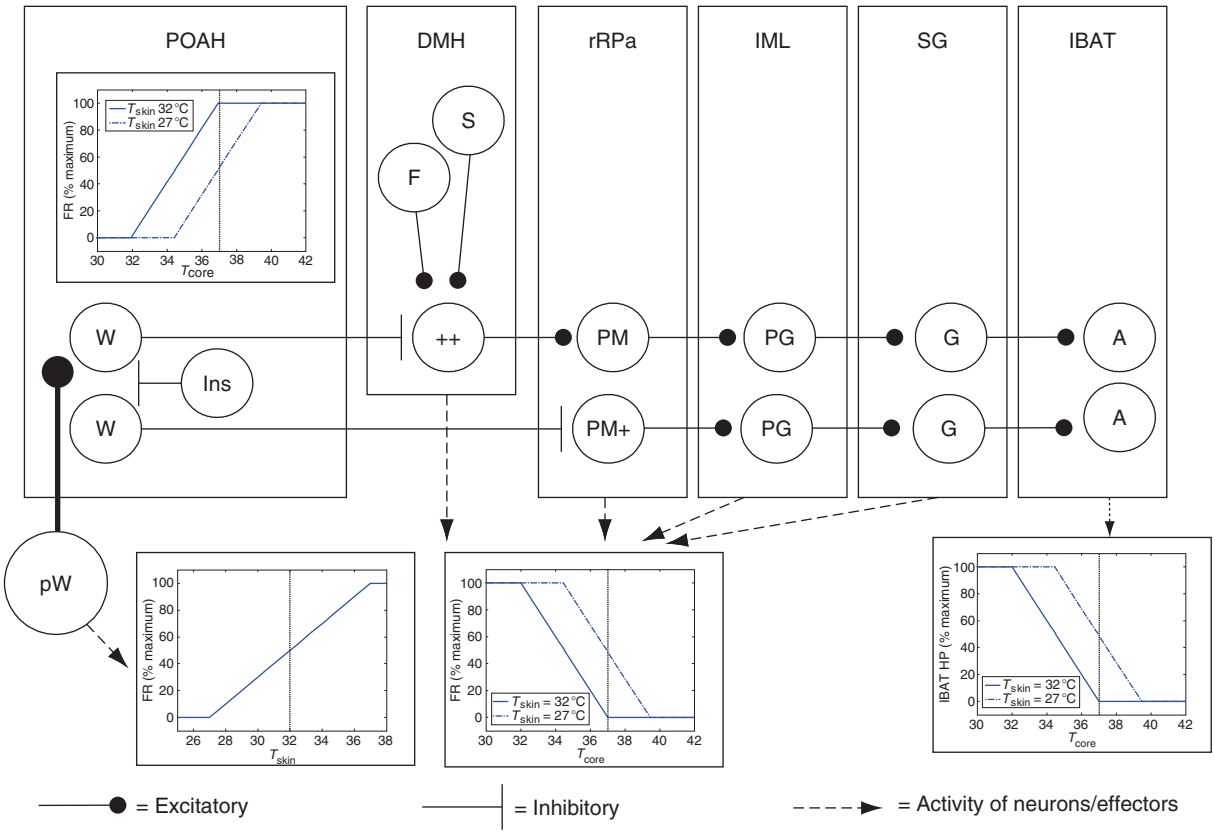

Figure 1 Model for the control of BAT thermogenesis by the sympathetic nervous system. Under thermoneutral conditions, intracapsular brown adipose tissue (IBAT) facultative thermogenesis is fully inhibited. This inhibition reflects a high level of inhibitory drive produced by preoptic anterior hypothalamic (POAH) warm-responsive (W) neurons acting on tonically excited neurons (++ and premotor (PM+)) in the dorsomedial hypothalamus (DMH) and rostral raphe pallidus (rRPa). POAH warm-sensitive neurons exhibit spontaneously high firing rates (FR) due both to an intrinsically high level of activity and strong excitatory input from peripheral warm-responsive (pW) sensors; a pW neuron is shown as providing excitatory input to both POAH neurons. With a lowering of hypothalamic temperature (assumed to correspond to T_{core}) and/or skin temperature, downstream DMH and rRPa neurons are disinhibited, thereby increasing noradrenergic input into IBAT brown adipocytes (A). The graphs depict the idealized effects of reducing pW input, as with skin cooling, illustrating that core and skin temperatures work together in producing IBAT effector responses. Ins, temperature-insensitive neuron; F, feeding-responsive neuron; S, stress-responsive neuron; IML, intermediolateral column of the thoracic spinal cord; PG, prostaglandin E subtype receptor-expressing neuron; SG, stellate ganglion; G, ganglionic neuron; solid lines drawn at normal core or skin temperatures (see text for further details).

precise regulation of T_{core} via autonomic means (effective behavioral regulation can survive ablation of the POAH and it should be noted that this is not the only region with temperature-sensitive neurons that are coupled to autonomic thermoeffector responses; such neurons are, in fact, found in all subdivisions of the neuraxis, including the spinal cord). A high percentage (perhaps ~30%) of the neurons in the POAH are warm-sensitive (W) neurons, meaning that their firing rate increases (by at least 0.8 Hz/°C) with increasing local temperature. Many of these are also responsive to skin and spinal temperature – note that Figure 1 depicts POAH W neurons receiving considerable excitatory input by peripheral warm sensors (pW). POAH W neurons also receive input from temperature-insensitive neurons (Ins), and Figure 1 depicts an inhibitory γ-aminobutyric acid (GABA)ergic prostaglandin (PG) E subtype receptor-expressing neuron, which enables a model for the upward setting of the temperature 'set point' (i.e., fever) during certain inflammatory states (described later). Note also in Figure 1 that the thermosensitivity of the POAH W neuron is such that its basal firing rate (FR) is nearly or fully maximal at the usual normothermic body temperature (37 °C). The high basal FR reflects the combination of a high intrinsic FR and strong excitatory drive provided by pW sensors; as indicated by the upper left-hand inset (in the POAH block), the W receptor FR at a given T_{core} (e.g., 37 °C) decreases when T_{skin} decreases from 32 °C (a thermally comfortable value) to 27 °C (which would usually be perceived as uncomfortably cold except perhaps in

cold-acclimated Aborigines, as described later). This schema agrees with experimental data and accounts for the fact that reflexive thermoeffector responses depend on a combination of core and skin temperatures. For simplicity, FR varies directly with temperature within a range bounded by particular hypothermic-to-normothermic values of core and skin temperature (here, the function for W receptor FR is given as % maximum FR $= 20 \times T_{core} + 10 \times T_{skin} - 960$, with the constraints that $0\% < \text{FR} < 100\%$); in reality, different central and peripheral thermosensors evince different FR functions, including quasilinear, sigmoidal, and quasi-Gaussian. It should be noted that warm sensors (as opposed to cold sensors, i.e., those that increase FR with a decrease in temperature) are hypothesized to fulfill the role of primary functional importance in thermoregulation, although by no means do all authorities agree with this. The % maximum FR equation is in general agreement with a long-accepted functional relationship originally described by Carl Gisolfi, a thermoregulatory and exercise physiologist, holding that a variety of thermoeffector responses can be described by the equation: $R - R_0 = B_1 \times T_{core} + B_2 \times T_{skin} - B_3$, where R is the effector response, R_0 is the basal effector response level, and the coefficients weight the importance of core and skin temperature in driving the change in response. In humans, the ratio of B_1:B_2 for thermoeffector responses ranges between about 9:1 and 5:1 (i.e., considerably higher than that for any of the models presented in this article).

Figure 1 depicts two POAH W neuronal projections that are inhibitory and which project caudally to synapse with tonically active neurons in other brain regions – the dorsal medial hypothalamic region (DMH, containing the neuron labeled ++) and the rostral raphe pallidus (rRPa, containing the premotor neuron labeled PM+) in the medulla oblongata. Neurons in the DMH and rRPa are coupled via additional excitatory relays to BAT. This type of connectivity comports with abundant data indicating that BAT (and also cutaneous vasoconstriction) are subject to considerable tonic excitatory drive that is normally restrained by inhibitory inputs. The DMH is a key region involved in integrating a variety of homeostatic responses, including 'emotional' stress, feeding/adiposity, and thermoregulation. Accordingly, the neuron labeled 'S,' a stress-activated neuron, and the neuron labeled 'F,' a feeding-activated neuron, are each capable of increasing excitatory drive to intrascapular BAT (IBAT). The rRPa region in the medulla oblongata has emerged as a major site of excitatory premotor neurons controlling IBAT in rodents. In addition to inputs from the DMH, there is evidence that some POAH W neurons project monosynaptically to rRPa neurons. An important functional consequence of both POAH-to-RPa circuit pathways is that RPa premotor neuronal firing is normally suppressed almost fully by the high level of POAH W neuron firing associated with normothermia, resulting in a low basal level of drive to IBAT, and this concept fits well with the experimental literature. With the occurrence of hypothalamic and/or skin cooling, the W receptor FR decreases, reducing the inhibitory drive to RPa IBAT premotor neurons, which stimulates HP via premotor neurons in the intermediolateral (IML) column of the spinal cord, ganglionic (G) cells in the stellate ganglion (SG), and, finally, brown adipocytes (A) in IBAT. Consider the genesis of fever, a phenomenon that behaves like a coordinately regulated increase of T_{core}. The temperature-insensitive neuron situated in the POAH represents a prostaglandin E subtype (EP3) receptor-expressing GABAergic cell. The EP3 receptor binds prostaglandin E_2 (PGE_2), a neurochemical product of inflammatory cascades such as that induced by the endotoxin lipopolysaccharide (LPS). LPS administration is a widely used experimental model of fever induction. It is now thought that LPS-induced PGE_2 binds to the EP3 receptor on POAH GABAergic neurons, thereby decreasing the FR of POAH W neurons and reducing the inhibitory drive to the tonically active RPa IBAT premotor neurons, so as to stimulate IBAT thermogenesis and precipitate a state of positive heat balance. Accordingly, this schema provides a putative mechanism for the genesis of fever, a state often ascribed to the upward resetting of the body temperature 'set point.' An important point emerges here: biology does not employ set points designed in the manner of the typical engineered device, although this fiction is still promulgated in myriad contemporary biomedical textbooks and scholarly publications (i.e., the set point equals the value of the reference input from which the value of negative is subtracted to generate the error signal). Starting with the classical neuronal model proposed by Hammel in the mid-1960s, a variety of models have been advanced to explain how the body implements set-point-like mechanisms that are consistent with the facts of neuronal signaling – particularly the fact that neurons do not generate outputs with negative values (e.g., -5 Hz), as would be necessary if the engineered set-point device were to be used – and the ability of many factors to modify the ultimate value of the regulated variable that is governed by the set point (e.g., the ability of LPS to elicit coordinated biobehavioral responses that increase the defended level of body temperature during fever). The reader should appreciate the existence of

biologically plausible schemes for what is variously called 'set point,' 'settling point,' 'set-level determinant,' 'interthreshold zone,' or 'null zone,' and is encouraged to consider the reports by Hammel, Bligh, Boulant, and Kanesoue.

Some additional details of sympathetically mediated BAT thermogenesis merit discussion. The SNS terminals innervating BAT release norepinephrine (NE), which binds to the β-3-adrenergic receptor on BAT adipocytes. This initiates a G-protein/cyclic adenosine monophosphate/protein kinase A (Gs/cAMP/PKA) signaling cascade that produces hormone-sensitive lipase (HSL). HSL increases lypolysis of triglyceride droplets. The resultant increase in free fatty acids both activates UCP1 and enters the β-oxidation cycle in mitochondria, serving as the thermogenic substrate. SNS activation increases UCP1 gene expression by phosphorylation of the transcription factor cAMP response element-binding protein (CREB), another cAMP/PKA-dependent signaling pathway. Chronic SNS stimulation of BAT induces mitochondrial biogenesis and brown adipocyte hyperplasia, and also induces hypertrophy in these cells via the β-1-adrenergic receptor. Conversely, a chronic decrease of SNS activity results in BAT atrophy.

Adaptive BAT thermogenesis also involves a critical role for thyroid hormones. Physiological increases of triiodothyronine (T3) increase oxygen consumption (heat production is proportional to oxygen consumption, and in nonexercising individuals can be estimated with minimal error as $4.85\,\mathrm{cal\,g^{-1}}$ of body mass). In fact, key steps in BAT thermogenesis are thyroid hormone dependent. Brown adipocytes express thyroid hormone receptors TRα-1 and TRβ; the latter mediates T3-induced UCP1 gene expression. In addition to the effects outlined earlier, NE-induced increases in cAMP also increase the conversion of intracellular thyroxine (T4) to T3 by activating type 2 iodothyronine deiodinase (D2), and this confers a thyroid–SNS synergism that is essential for the increase of thermogenesis required by small mammals in cold environments.

While BAT is the main source of thermoregulatory heat production in small mammals such as rats and mice, in human infants BAT may not be the only organ available for NST. Although UCP1 is expressed exclusively in BAT, other tissues, such as muscle and liver, express other variants of the UCP protein, which may confer the capability for NST (here defined, as is customary, as the metabolic heat produced above the basal level in response to cold exposure) in these tissues. The evidence for this proposal is not strong, however. What is clear is that NST in humans can be increased substantially by chronic cold exposure and, in some (lucky) individuals, by overfeeding, as indicated by an increase of resting energy expenditure (REE) during periods of overfeeding, conferring resistance to diet-induced weight gain. The roles of BAT and other tissues in this phenomenon remain unresolved.

A remarkable paradigm of increased NST in humans is provided by Hong's classic studies of traditional ama divers of Korea and Japan, women who harvest plant and animal life in waters reaching 10 °C during the winter. Before the adoption of wet suits, these women wore traditional cotton bathing suits. During winter months, these ama exhibited a ~25% increase of basal metabolic rate (which was associated with a ~50% increase of caloric intake, presumably reflecting additional energetic challenges of winter diving), as well as other adaptive heat-conserving responses, but not including an increase of adiposity.

Shivering Thermogenesis

Shivering thermogenesis occurs when practically all of the skeletal muscle groups are involuntarily driven to contract synchronously (at ~40 and 31 Hz in the mouse and rat, respectively), producing heat (the mechanical efficiency of shivering is 0%).

Heat production during cold exposure can be increased by nearly three- to fourfold by shivering. Skin and core temperatures act synergistically in the activation of shivering (but a simpler additive relationship is assumed in the model depicted in Figure 2). Although shivering can increase T_{core} by ~0.5 °C in humans, this thermogenic mode is inefficient because it increases the convective transfer of body heat away from the core by increasing muscle blood flow, and it increases convective heat loss to the environment via gross bodily movement (wind chill).

Model for the Control of Shivering Thermogenesis

The model depicted by Figure 2 shares several important structures and functional relationships with the model in Figure 1. Additional CNS areas include the posterior hypothalamic (PH) area, an area contiguous with the DMH which has long been implicated in thermoregulation; the midbrain red nucleus (RN) and reticular formation (RF), a diffuse complex of neurons and processes in the midbrain, pons, and medulla; and the ventral horn (VH) and dorsal horn (DH) of the spinal cord. The RN and RF give rise to descending motor pathways that innervate skeletal muscle fibers (SFs) as well as intrafusal muscle fibers (IFs) within muscle stretch receptors (also known as fusiform bodies). As with the model for BAT thermogenesis (Figure 1), reducing the FR of POAH

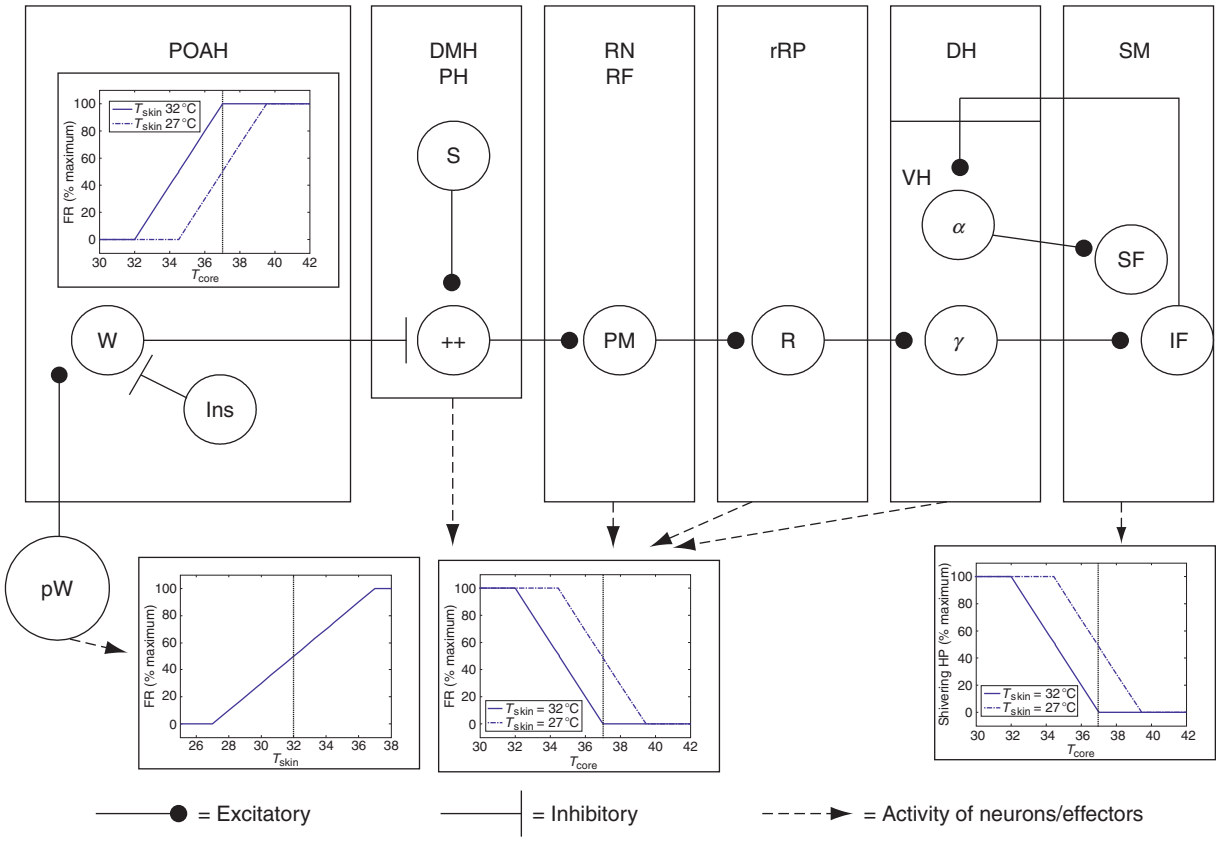

Figure 2 Model for the control of shivering thermogenesis. A decrease in preoptic anterior hypothalamic (POAH) warm-sensitive (W) neuronal activity disinhibits dorsomedial hypothalamus (DMH) and posterior hypothalamic (PH) stress-responsive (S) neurons that signal to premotor (PM) neurons in the red nucleus (RN) and reticular formation (RF). This allows increased excitatory drive to γ (and probably α) motor neurons in the ventral horn (VH) of the spinal cord (DH, dorsal horn of the spinal cord). Increased γ discharge to intrafusal fibers (IF) in skeletal muscle (SM) increases excitatory drive to the α motor neurons (analogous to the stretch reflex), causing involuntary contractions of skeletal muscle fibers (SF). FR, firing rate; HP, heat production; Ins, temperature-insensitive neuron; pW, peripheral warm-responsive sensor; rRP, rostral raphe pallidus; R, rostral raphe pallidus neuron.

W neurons by hypothalamic and/or skin cooling disinhibits tonically excited descending pathways. In this case (Figure 2), this 'excitation via the withdrawal of inhibition' activates γ motor neuronal input to IF muscle fibers. This, in turn, increases the afferent excitatory input from muscle spindles to the α motor neurons in the VH of the spinal cord, leading to increased contractile activity (i.e., shivering), with heat as a by-product. In essence, this mechanism exploits the classical stretch reflex by increasing the loop gain to the point that this subsystem exhibits sustained oscillations between states of contraction and relaxation. Note that Figure 2 includes a role for temperature-insensitive fever-promoting neurons in the POAH, as described for Figure 1, in keeping with the known involvement of shivering in thermogenesis during fever. Also depicted is a temperature-insensitive stress neuron in the DMH/PH, in keeping with the ability of nonthermal stressors to evoke a shivering-like state.

Heat Loss

Autonomic mechanisms of heat loss include pathways regulating cutaneous (skin) blood flow, a major determinant of the delivery of heat from the body core to the shell, and pathways regulating evaporative heat transfer to the environment via the secretion of sweat from skin (critical in humans), panting (critical in canines and other nonsweating homeotherms), or saliva spreading (critical in rodents).

Cutaneous Blood Flow

Sympathetically mediated regulation of cutaneous blood flow is vitally important for controlling the rate of body heat loss. In the rat, it is chiefly the cutaneous vascular bed of the tail that is involved in thermoregulation, while in rabbits the ear pinna is a major site of regulated cutaneous heat dissipation. Humans benefit from thermoregulatory control of the cutaneous vasculature of most of the body's

skin, comprising a surface area of ~1.7 m^2 for a 70 kg person. In humans, there are two branches of the SNS control of cutaneous blood flow, a noradrenergic vasoconstrictor system and a sympathetic cholinergic active vasodilation system (which is also involved in sweating, as described later). In rodents, cutaneous tail blood flow appears to be chiefly regulated by modulation of SNS vasoconstrictor tone, achieved by noradrenergic control of the smooth muscle in arteriolar precapillary sphincters and arteriovenous anastomoses (AVAs), which are both abundant in the rat tail. The tail dissipates ~20–25% of a rat's basal HP, which ranges between approximately 70 and 100 cal kg^{-1} min^{-1} (note: 'calories' as used here means gram-calories, or small calories – not kilocalories, or large calories). During heat stress, the ability to increase tail blood flow contributes to heat tolerance, but not nearly as much as does evaporative HL (see later); for example, in one classic study, rats whose tails were amputated exhibited an 8% reduction in heat tolerance as measured by the time required to reach a T_{core} of 40 °C at an ambient temperature of 40 °C.

In a 70 kg human, skin blood flow is ~0.25 l min^{-1} under resting normothermic conditions, whereas during severe heat stress skin blood flow rises to ~5 l min^{-1}, a 20-fold increase, representing nearly one-fourth of the cardiac output. During cold stress, skin blood flow can drop to ~0.05 l min^{-1}. It is worth noting that the rate of heat loss is proportional to the product of skin blood flow and the difference between core and mean skin temperatures. In addition to underscoring the importance of skin blood flow responses to thermal challenges, this quantitative relationship focuses attention on the importance of skin temperature in thermoregulation. In particular, one may note the value of keeping the skin relatively cool during exposure to hot environments and during exercise, where metabolic heat production can increase by nearly eight- to tenfold in healthy, fit individuals (roughly equivalent to the heat generated while jogging on a level track at 6 and 7.5 mph, respectively). Under such conditions, sweating provides an impressively powerful mechanism of skin cooling and hence of heat loss (see later). It is also of interest to note that one of the adaptations observed in some cold-dwelling humans is the ability to tolerate lower skin and core temperatures. For example, Hammel and his colleagues reported that during a night of controlled cold exposure, Australian Aborigines (who were acclimated to sleeping with little or no clothing in moderate cold, ~4–8 °C) readily tolerated core and skin temperatures that caused considerable discomfort and shivering in a group of Caucasian controls. Since HL is proportional to the difference between skin and ambient temperatures, lowering skin temperature reduces heat loss to the environment. In Hammel's Aboriginal subjects, the lowering of both core and skin temperatures reduced heat conductance to the environment by ~30% during sleep, a significant savings of body energy reserves. Therefore, being able to tolerate lower core and skin temperatures likely confers a significant survival advantage in settings of caloric scarcity. It should be noted that the adapted thermoregulatory state observed in Aborigines is apparently not found in humans living in extremely cold environments (e.g., Eskimos). Such peoples have survived by developing exceptionally effective clothing and shelter. (The traditional Eskimo cold-weather clothing ensemble is a marvel both of design and materials; the success of Roald Amundsen's pioneering expedition to the South Pole, reached on December 15, 1911, owed a great debt to Amundsen's adoption of Eskimo clothing and travel methods, as described in Roland Huntford's book, *The Last Place on Earth*. The interior of the Eskimo igloo rarely decreases below 10 °C.) The point is that if the environment is so severe that autonomic thermoregulatory adaptations are of little consequence to survival, humans simply eschew such adaptations in favor of largely avoiding the cold by way of a broad repertoire of behavioral options.

Model for the Noradrenergic Control of Cutaneous Blood Flow

Figure 3 depicts a model for the control of cutaneous blood flow by the noradrenergic vasoconstrictor system, based primarily on research in rats. Most of the structures depicted in Figure 3 have already been described for Figure 1; the new CNS region is the ventral tegmental area (VTA) of the midbrain, where an 'other' (O) neuron is situated to modulate the activity of a tonically active VTA neuron. Some of these likely are temperature-sensitive neurons. The ultimate peripheral target of this system is the smooth muscle controlling blood flow through the cutaneous vasculature (CV). The general functional scheme of Figure 3 follows that of Figures 1 and 2. Normally active POAH W inhibitory neurons project to tonically active neurons in the DMH and VTA, but in the present model there is room for upward shifts in the POAH W firing rate at normal hypothalamic temperature so as to permit an increase inhibitory drive to tonically active DMH and VTA neurons if skin temperature increases. Accordingly, an increase of pW input to POAH W neurons further inhibits DMH and VTA neuronal firing, leading to a withdrawal of sympathetic smooth muscle tone and a resultant increase in cutaneous blood flow. The function describing the POAH W FR is % maximum

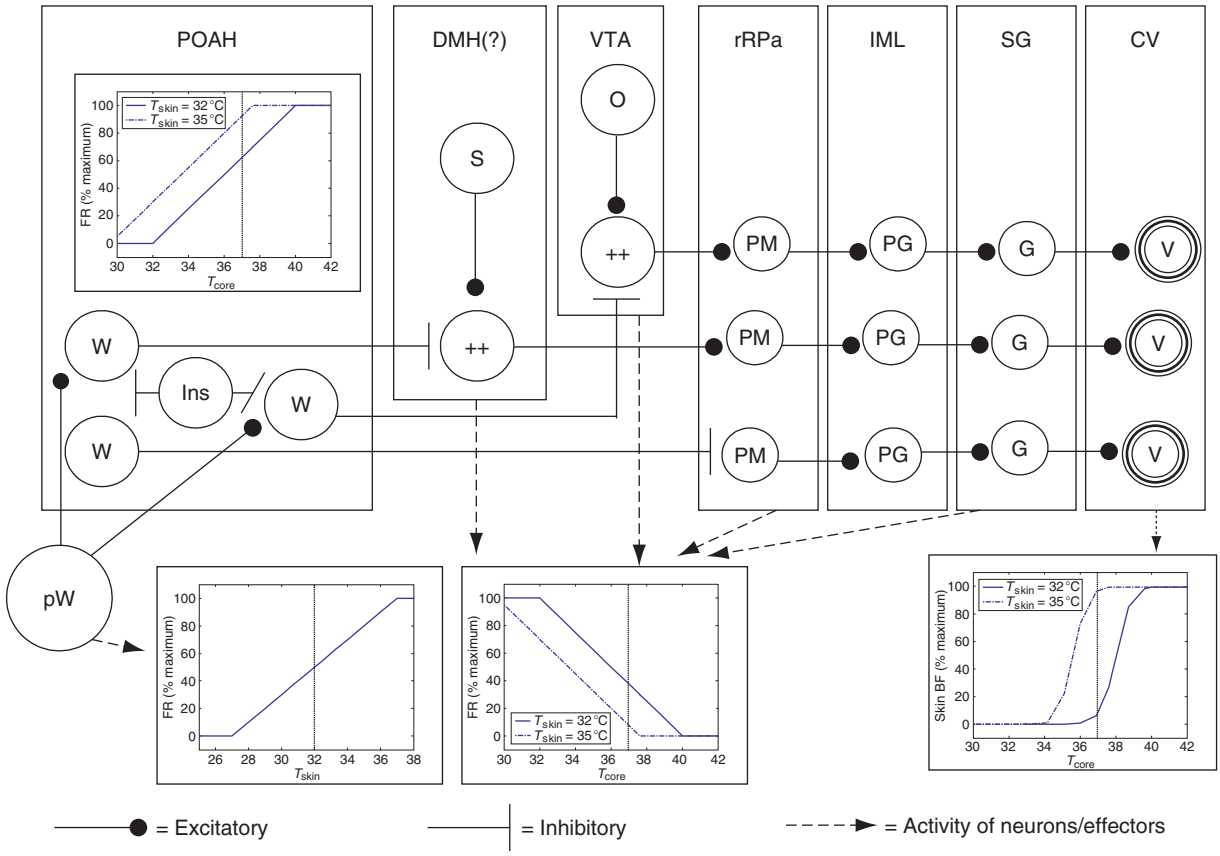

Figure 3 Model for the control of cutaneous blood flow by the sympathetic nervous system. The pathways and control strategy are similar but not identical to those of **Figure 1**. Under thermoneutral conditions, the vasoconstrictor tone of the cutaneous vasculature (CV) is maintained at a fairly high level by tonically excited dorsomedial hypothalamus (DMH) and ventral tegmental area (VTA) neurons (O, other neuron) that are only partially inhibited by preoptic anterior hypothalamic (POAH) warm-sensitive (W) neurons. Increasing T_{core} and/or skin temperature causes an increase in POAH W neuronal activity, leading to greater inhibition of noradrenergic drive to the CV. This withdraws vasoconstrictor tone, allowing an increase of skin blood flow (BF), whereas decreases of POAH W neuronal activity decrease skin BF to very low levels via disinhibition of the tonically active DMH and VTA neurons. The question mark in the DMH block indicates that this area is not yet proved to be in this circuit (see text for further details). FR, firing rate; Ins, temperature-insensitive neuron; pW, peripheral warm-responsive sensor; S, stress-responsive neuron; rRPa, rostral raphe pallidus; IML, intermediolateral column of the thoracic spinal cord; SG, stellate ganglion; G, ganglionic neuron; V, vasoconstrictor neuron.

$FR = 12.5 \times T_{core} + 10 \times T_{skin} - 720$; the sigmoidal function describing skin blood flow is % maximum skin BF $= 100 \times [1 - 1/(1 + \exp(-1 \times (U))]$, where $U =$ (input FR $- 25)/5$ – that is, the response is half-maximal at an input FR of 25 and the width parameter is 5. This sigmoidal response is congruent with data indicating that rat tail blood flow increases exponentially by nearly tenfold from the basal values measured in a thermoneutral environment (e.g., ambient temperature = 30 °C) to the values measured in a hot environment (e.g., ambient temperature = 35 °C). The proportional increase in the rate of heat loss via the tail mirrors the proportional increase of tail blood flow.

Cholinergic Control of Cutaneous Blood Flow

Humans have both a sympathetic noradrenergic vasoconstrictor system and a sympathetic cholinergic active vasodilation system; the latter is responsible in humans for 80–95% of the increase in skin blood flow accompanying heat stress. Relatively little is known about the CNS areas that govern thermoregulation in humans, and it certainly remains to be established to what degree rodents, rabbits, and humans share common central substrates and pathways. (No model is presented in this article for cholinergic active vasodilation.)

Evaporative Heat Loss

Each liter of sweat evaporated from the skin surface requires ~580 kcal of heat energy, an impressive value. Most of this energy is derived from body heat, but thermal energy derived from the environment contributes to the evaporation of sweat, particularly in hot, dry environments where sweating keeps skin temperature below environmental temperature.

In humans, sweat rates can exceed $2\,l\,h^{-1}$, which would equate to $\sim 1200\,kcal\,h^{-1}$ of body heat removal if all of the sweat were to be vaporized by body heat. Normally, however, only about 50% of the sweat produced is evaporated, and the fraction that is not evaporated is of utterly no value in removing heat. The legendary marathoner Derek Clayton could produce sweat at $3.7\,l\,h^{-1}$, a vital factor in his ability to sustain a sub-5-min $mile^{-1}$ pace for 26.2 miles, a feat that typically requires sustaining a metabolic rate ~20-fold higher than resting. Anthropologists have proposed that the human capacity for evaporative heat loss (EHL) conferred a major selection advantage. EHL is so effective in preserving thermal homeostasis during prolonged walking and running in hot climates that early humans benefited greatly in terms of hunting, gathering, and other essential survival activities. In humans, sweating is controlled by sympathetic cholinergic nerve terminals that innervate the ~2 million eccrine ('thermoregulatory') sweat glands distributed in the skin. The cholinergic control of sweat glands is called 'sudomotor regulation.' Sweating and active vasodilation commence at nearly the same time in resting heat-stressed humans, and both processes are thought to involve cholinergic sudomotor activity, although the precise relationship between them remains unclear. It should be noted that much of the impressive human capacity for heat dissipation via sweating would be lost if it were not for the ability to achieve the marked increases in skin blood flow enabled by active cholinergic vasodilation.

Sweat glands are found almost exclusively in humans (the patas monkey has eccrine sweat glands). Lacking sweat glands, however, does not preclude the use of evaporative cooling for thermoregulation in the heat. Thermoregulatory EHL also occurs via panting and the application of saliva, urine, and other sources of moisture to the fur and skin. Panting, defined as a rapid respiratory frequency producing an increased minute volume but decreased tidal volume, is a critical avenue of EHL in canines, felines, and other mammals. Rodents also exhibit a panting-like phenomenon during heat stress, but the role of

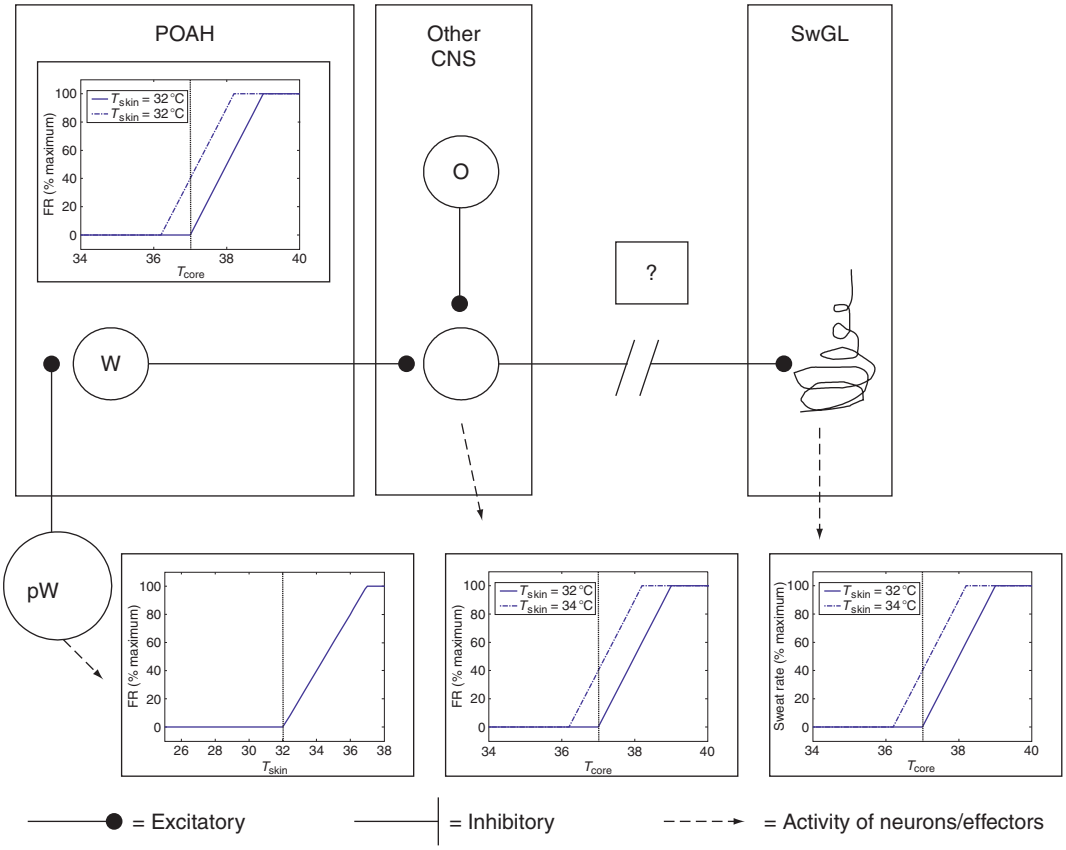

Figure 4 Model for the control of sweating by the sympathetic nervous system. An increase in core and/or skin temperatures increases excitatory cholinergic drive to sweat glands (SwGL). This scheme postulates an important role for preoptic anterior hypothalamic (POAH) warm-sensitive (W) neurons that are excitatory to downstream neurons in other brain areas. FR, firing rate; CNS, central nervous system; O, other neuron capable of modulating sweat rate; pW, peripheral warm-responsive sensor.

panting in thermoregulation appears to differ among rodent species. What is clear is that rodents are very good at increasing EHL by spreading saliva on the fur and other surfaces of the body, such as the paws, scrotum, and base of the tail. Indeed, this mode of EHL is a vital component of maintaining thermal homeostasis in rats (and presumably other rodents), based on studies in which surgically desalivated rats were shown to be highly susceptible to heat stress (one study documented an 81% reduction in the time to reach a T_{core} of 40 °C). Interestingly, just as cholinergic blockade in humans abolishes EHL via sweating, it also prevents EHL via saliva spreading in rats.

Model for the Control of Evaporative Heat Loss

Figure 4 depicts a simple model for the control of EHL via sweating (with a few changes this model is also applicable to panting and saliva spreading). Uncertainties about the central pathways involved in EHL regulation oblige a scheme specifying few constituent parts (unfortunately, rats do not employ thermoregulatory sweating, or we would probably know a great deal more about this topic). The box labeled SwGL refers to sweat glands, and 'Other CNS' simply acknowledges that there must be important CNS areas downstream of the POAH. Note that the thermosensitivity of POAH W neurons is such that the threshold for increasing their firing rate is at the normal T_{core}. We also postulate that the threshold for increased skin (pW) firing is at the normal, preferred T_{skin}. Accordingly, an increase of hypothalamic temperature above 37 °C promptly drives an increase in sweating, and this response can be initiated at a lower T_{core} if skin temperature increases. (In fact, skin and core temperatures act synergistically in the sudomotor response. Here, the equation is % maximum FR = $50 \times T_{core} = 20 \times T_{skin} - 2490$.)

Concluding Remarks

The fact that the typical mammalian T_{core} of 37 °C is poised just 4–5 °C away from the limits of survival is often cited to underscore the vital survival importance of thermoregulation. But perhaps less widely appreciated is that all biobehavioral processes in health and disease depend – often critically – on T_{core}. Accordingly, the biobehavioral basis of thermoregulation has long fascinated students and scientists in a broad array of disciplines. However, much remains to be learned, and since the new knowledge will be of key importance in understanding many other physiological and pathophysiological phenomena, the reader is encouraged to learn more about, and perhaps even contribute, to the field of thermoregulation.

See also: Autonomic Failure; Autonomic Nervous System: Central Thermoregulatory Control; Sympathetic Nervous System; Thermoregulation During Sleep and Sleep Deprivation; Thermoregulation: Autonomic, Age-Related Changes.

Further Reading

Blessing WW (2005) BAT control shows the way: Medullary raphe/parapyramidal neurons and sympathetic regulation of brown adipose tissue. *American Journal of Physiology Regulatory Integrative and Comparative Physiology* 288: R557–R560.

Boulant JA (2006) Neuronal basis of Hammel's model for set-point thermoregulation. *Journal of Applied Physiology* 100: 1347–1354.

Brengelmann GL (1989) Body temperature regulation. In: Patton HD, Fuchs AF, Hille B, Scher AM, and Steiner R (eds.) *Textbook of Physiology,* 21st edn., pp. 1584–1596. Philadelphia: W.B. Saunders.

Cao W-H and Morrison SF (2006) Glutamate receptors in the raphe pallidus mediate brown adipose tissue thermogenesis evoked by activation of dorsomedial hypothalamic neurons. *Neuropharmacology* 51: 426–437.

DiMicco JA and Zaretsky DV (2007) The dorsomedical hypothalamus: A new player in thermoregulation. *American Journal of Physiology Regulatory Integrative and Comparative Physiology* 292: R47–R63.

DiMicco JA, Sarkar S, Zaretskaia MV, and Zaretsky DV (2006) Stress-induced cardiac stimulation and fever: Common hypothalamic origins and brainstem mechanisms. *Autonomic Neuroscience: Basic and Clinical* 126–127: 106–119.

Fan W, Voss-Andrea A, Cao W-H, and Morrison SF (2005) Regulation of thermogenesis by the central melanocortin system. *Peptides* 26: 1800–1813.

Gisolfi CV and Mora F (2000) *The Hot Brain*. Cambridge, MA: MIT Press.

Gordon CJ (1993) *Temperature Regulation in Laboratory Rodents*. Cambridge, UK: Cambridge University Press.

Hammel HT, Jackson DC, Stolwijk JAJ, Hardy JD, and Strømme SB (1963) Temperature regulation by hypothalamic proportional control with an adjustable set point. *Journal of Applied Physiology* 18: 1146–1154.

Romanovsky AA (2007) Thermoregulation. Some concepts have changed. Functional architecture of the thermoregulatory system. *American Journal of Physiology Regulatory Integrative and Comparative Physiology* 292: R37–R46.

Satinoff E (1978) Neural organization and evolution of thermal regulation in mammals. *Science* 201: 16–22.

Satinoff E (1983) A reevaluation of the concept of the homeostatic organization of temperature regulation. In: Satinoff E and Teitelbaum P (eds.) *Handbood of Behavioral Neurology,* vol 6, pp. 443–472. New York: Plenum Press.

Tanaka M, Owens NC, Nagashima K, Kanosue K, and McAllen RM (2006) Reflex activation of rat fusimotor neurons by body surface cooling, and its dependence on the medullary raphe. *Journal of Physiology* 572: 569–583.

Energy Homeostasis: Visceral Control

T L Powley and R J Phillips, Purdue University, West Lafayette, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

Distributed Visceral Networks

The regulation of energy homeostasis is the product of neural networks that operate as physiological control systems. The networks span the central, peripheral, and enteric nervous systems (CNS, PNS, and ENS, respectively). The circuits extend over the entire visceral neuraxis, from the ENS in the walls of the viscera, through the autonomic or PNS pathways and the networks in both brain stem and spinal cord, to the rostral forebrain systems. These networks are interconnected and organized as feed-forward, feedback, and adaptive control loops. The pathways carrying afferent inflows and efferent outflows are dynamically cross-linked, and, importantly, integration of signals occurs at all levels of the visceral neuraxis.

Comparisons with earlier views of the neural mechanisms of energy balance bring this feature of this control-system concept into sharper relief: early views of the neural mechanisms of energy balance focused on the CNS. These older ideas employed hierarchical models of neural organization and construed the hypothalamus and a small number of CNS sites as integrative and executive centers. In such models, the peripheral nerves simply transmitted unintegrated signals from the periphery to diencephalic and forebrain centers and then carried motor commands back to the periphery. As more has been learned about the organization of the CNS, PNS, and ENS, however, such center-based hierarchical views of energy homeostasis have been replaced by a recognition that multiple distributed control loops integrate the regulatory adjustments of the CNS, PNS, and ENS.

Central Integrators and Visceral Networks Organize Short- and Long-Term Adjustments of Energy Homeostasis

Neural mechanisms that determine the energy economy of the body are typically distinguished as either short-term or long-term control loops. Short-term mechanisms execute adjustments that often occur in seconds or minutes. The responses occur so rapidly that they rely predominantly on neural pathways – there is relatively little time for paracrine, endocrine, and humoral responses to mobilize fully. Visceral afferents provide most of the fast, phasic short-term feedback associated with ingesting, digesting, and absorbing nutrients that provide energy.

Long-term mechanisms make adjustments that can often take hours (e.g., circadian fluctuations), days (e.g., estrus or menstrual cycles), or seasons (e.g., seasonal weight gain). Such adjustments are so protracted they can be, and typically are, orchestrated by endocrine and other humoral factors. These longer term signals operate directly on peripheral visceral afferents, but they also directly affect central stations of the visceral neuraxis, including notably the hypothalamus as well as the caudal brain stem, amygdala, hippocampus, and various other limbic regions.

The short-term/long-term distinction is useful in the discussions of function, but it is by no means hard and fast. Neural effects can occur over extended intervals, and paracrine and endocrine effects can occur in minutes, perhaps in seconds. Other factors also can blur the distinction. Energy control systems are cross-linked, the same controller can participate in both short-term and long-term adjustments, and circulating humoral factors modulate or change the gain of the neural elements in the networks. The present survey focuses on the neural elements of these networks.

Peripheral Inputs of the Energy Control Circuitry: Visceral Afferents

The first-order, or primary, visceral afferents that transmit sensory information from the gastrointestinal (GI) tract and organs of digestion to the CNS are organized into two pools of neurons. These two pools interconnect the CNS and the viscera and organize them into control networks. Neurites of one of the afferent pools travel with parasympathetic motor fibers, notably those of the vagus nerve, to innervate the different visceral organs involved in digestion and metabolism. Neurites of the other afferent pool travel with sympathetic nerves, notably the splanchnic nerves, to innervate the visceral organs (as well as adipose tissue depots).

The larger of the two populations of visceral afferent neurons constitutes the neurons of the nodose ganglion. Nodose neurons are pseudounipolar, distributing their peripheral processes through the vagus nerve to the organs of digestion and projecting their central processes into the brain stem nucleus of the solitary tract. Within the nucleus of the solitary tract, the visceral afferents synapse on second-order neurons that relay information rostrally throughout the central visceral neuraxis participating in energy

homeostasis. Some primary nodose neurites also synapse directly on dorsally directed dendrites of preganglionic neurons of the dorsal motor nucleus, thus establishing vago-vagal reflex arcs. Centrifugal outputs from more rostral CNS networks project onto these brain stem circuits to effect adjustments in their outflows and thus energy homeostasis. In addition to incorporating modulatory influences from more rostral mechanisms, the reflex circuits operate directly as mono- or oligosynaptic reflex arcs. Efferent outflows of these loops then reciprocally, in closed-loop fashion, affect or alter afferent inflows.

The second and smaller pool of visceral afferents constitutes a subset of the dorsal root ganglion (DRG) cells situated in association with the thoracic and higher lumbar spinal cord. Like nodose afferents, DRG primary afferent neurons projecting to the GI tract are pseudounipolar neurons, distributing their peripheral processes through the peripheral nerves projecting to the prevertebral autonomic ganglia that supply the viscera. The spinal afferent processes then course still more peripherally with postganglionic efferent axons to innervate the organs. These visceral afferents project their central neurites into laminae I and V of the spinal cord. Second-order neurons in the laminae of the cord project both rostrally into the more rostral visceral neuraxis and locally to sympathetic preganglionic motor neurons of the intromediolateral column.

Figure 1 Vagal afferent innervation of the stomach. A montage of a dorsal stomach whole mount from a rat that received a left nodose injection of wheat germ agglutinin–horseradish peroxidase. The pylorus is in the upper left, and the forestomach is at the right. Bundles of afferents enter from the lower esophageal sphincter (excised at the top) and radiate over the stomach to the greater curvature. In the antrum (upper left), bundles of axons can be seen traversing across the pylorus to innervate the small intestine. At this scale, fiber bundles are the most conspicuous elements; however, afferent terminals in the myenteric plexus and smooth muscle can be recognized. The organ measures approximately 40 mm on the long axis.

Vagal and Spinal Visceral Afferents Transduce Primary Energy Signals

The vagus nerve supplies the majority of the visceral afferents to the GI tract and liver. The nerve densely innervates the upper esophagus and larynx, as well as the more distal esophagus, the stomach, and the duodenum; it also innervates, though less densely, the remainder of the small and large intestines. As Figure 1 illustrates for the stomach, this innervation forms an extensive network distributing throughout the organs. These nodose ganglion afferents detect both the mechanical, chemical, and thermal properties of ingested nutrients and the chyme that is produced in the digestion and absorption of the energy.

Vagal afferents elaborate a heterogeneity of different ending types to transduce and analyze energy-related signals. Vagal mechanoreceptors, for example, are located both in the smooth muscle wall of the GI tract and in the mucosa and submucosal lining of the viscera. Different populations of mechanoreceptors are located in different tissues and sites, have different morphologies, and appear to transduce, depending on the specific population, forces of tension, or stretch. Figures 2 and 3 document two of the morphological specializations of vagal mechanoreceptors found in the stomach wall and illustrate that neurites of visceral afferents are differentiated into specialized terminals. Functional studies implicate the mechanoreceptors in monitoring the contents of the organs, particularly nutrient loads of the stomach.

Vagal chemoreceptors are situated to detect stimuli in the lumen of the organ system and, likely, paracrine factors in the lamina propria and adjacent tissues as well. The better studied of these chemoreceptors detect and respond selectively to carbohydrates, lipids, and proteins. Structural details of receptor location and morphology have not yet been fully established, but both electrophysiological and behavioral experiments indicate that the vagal chemoreceptors of the GI tract constitute the primary short-term negative feedback associated with nutrient ingestion and digestion.

The dorsal root ganglia also supply primary visceral afferents to the gastrointestinal organs. DRG afferent neurites are distributed through the esophagus,

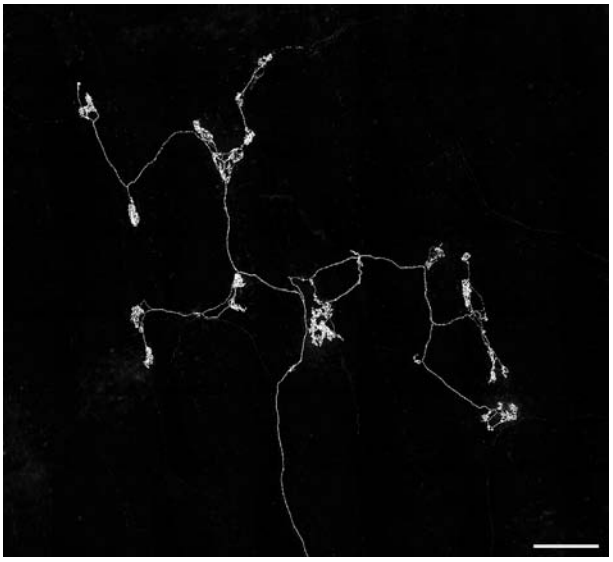

Figure 2 Vagal afferents supply intraganglionic laminar endings to the myenteric plexus of the gastrointestinal tract. An individual vagal afferent that ends in multiple intraganglionic laminar endings is illustrated in a montage from the corpus of the rat stomach. A single afferent neurite labeled with the fluorescent tracer tetramethylrhodamine, which was injected into the nodose ganglion, can be seen entering the field from the bottom of the figure. As the neurite reaches its target area, it repeatedly divides into higher order processes that terminate in plates of laminar puncta that interdigitate with neurons and glia within the ganglia (neurons and glia are not stained in this image). Scale bar = 250 μm.

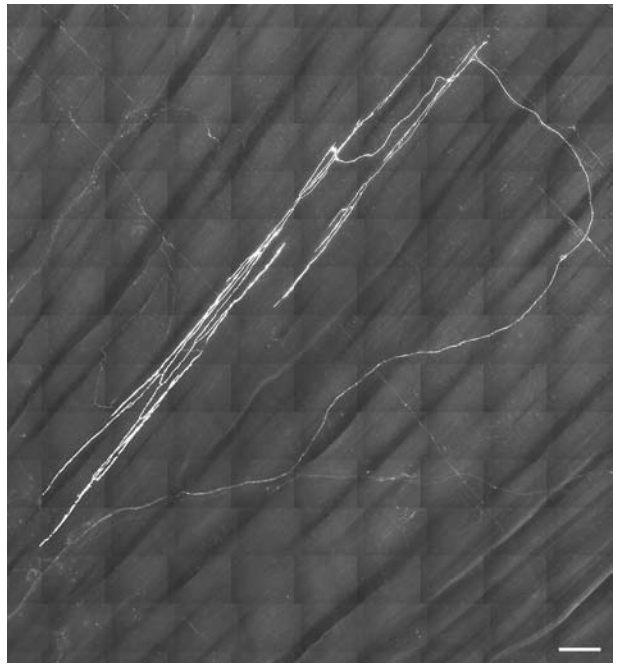

Figure 3 Vagal afferents supply intramuscular arrays (IMAs) to the smooth muscle layers of the GI tract. A single vagal afferent fiber terminating in a complex of IMAs is illustrated in this montage from the forestomach. In this example, a single afferent neurite labeled with the fluorescent tracer tetramethylrhodamine, which was injected into the nodose ganglion, can be seen entering from the bottom left and traveling to the upper right of the figure. As the neurite reaches the site of innervation of the smooth muscle, it gives off multiple collaterals that run in the same orientation as the muscle fibers in which the neurite terminates. The long rectilinear processes of IMAs are distributed parallel to each other and parallel to the muscle fibers. IMAs make contact with both smooth muscle fibers and the intercalated interstitial cells of Cajal (not stained). In the example illustrated, the arrays of neurite terminals that constitute the IMAs run in the circular muscle layer that can be seen running from the upper right of the figure to the lower left of the figure. Scale bar = 250 μm.

stomach, and intestines. These neurites are located in both muscle layers and in the mucosa and submucosa, where many of them associate with the vasculature. In comparison with the situation for vagal visceral afferents, even less is known about structural specializations of different DRG afferent neurites in the GI tract, but electrophysiological observations have established that DRG afferents are predominantly high-threshold receptors that are implicated in nociception associated with mechanical and chemical insults.

Ironically, given their importance in human health and disease processes, much less is known about the vagal and spinal visceral receptors and their transduction mechanisms than is known about the corresponding features of some exotic receptors (e.g., electroreceptors, magnetoreceptors) found in fish and birds. The situation results in the main from the inaccessibility of these receptors and other technical difficulties associated with studying the 'enteroreceptors' of energy homeostasis. The limited observations available suggest, and most investigators adopt the working assumption, that the mechanoreceptors located in the viscera share some, if not all, receptor molecules, channels, and sensory cascades with the mechanoreceptors found in skeletal muscle and skin. Similarly, the chemoreceptors located in the viscera presumably share at least some of the receptor molecules, channels, and molecular cascades found in the chemosenses of taste and possibly olfaction.

Hormones, Neuromodulators, and Metabolites Reflecting the Energy Economy Modulate the Gain of Visceral Afferents

Though the available information is still limited, visceral afferents have been shown to express extensive libraries of receptors for gut hormones, neuropeptides, and other signal factors as well as multiple species of ionotropic and metabotropic channels. Much of the transduction of the mechanical, chemical, and thermal

signals associated with the processing of nutrients apparently involves ionotropic membrane channels. In addition, however, many of the receptors and metabotropic channels are involved in detecting the local hormonal, paracrine, and cytokine milieu produced by the processes of digesting, absorbing, and metabolizing sources of energy. As has been detailed for load-sensitive vagal mechanoreceptors in the GI tract, the activity of many afferents appears to be potentiated or otherwise modulated locally by hormones and metabolites. The sensitivity or gain of gastric mechanoreceptors, for example, is affected by cholecystokinin (CCK), leptin, ghrelin, peptide tyrosine-tyrosine (PYY), and a number of other factors associated with processing and utilizing energy sources.

Fewer analyses have examined whether or to what extent the sensitivity or gain of DRG afferents may also be modulated by GI and metabolic factors, but the principle seems broadly instantiated: observations have recently established that not only vagal, but also gustatory, afferents express receptors for a variety of metabolic hormones.

Visceral Afferents Articulate with Motor Outflows

Visceral afferents elicit and also organize the CNS and autonomic motor outflows and the responses coordinated by central integrators in multiple ways. Depending upon the situation, an afferent input might simultaneously influence responses and integrator adjustments at several levels. Three organizing roles of the afferents are as (1) sources of axon responses or axon reflexes, (2) elements of mono- or oligosynaptic autonomic reflex arcs, and (3) input arms of polysynaptic central relays incorporated into more rostral control-system adjustments.

Visceral afferents appear to establish axon responses or reflexes that can locally modulate peripheral secretion and motility. Most of the evidence that visceral afferents can produce such effector-like responses, through the release of neuropeptides and other products at their peripheral terminals where they transduce visceral signals, comes from other organ systems (e.g., the respiratory system), but limited observations on GI physiology as well as the presence of secretory material in the peripheral terminals of afferents in the GI tract suggest that at least some of the visceral afferents participating in energy homeostasis may generate axon reflexes as well.

Visceral first- and second-order afferent projections in the brain stem and spinal cord establish mono- and oligosynaptic reflex arcs through which the inflows are translated into preganglionic motor outflows that produce or modulate many of the GI motor and secretory responses that orchestrate energy intake and expenditure (see earlier). Vago-vagal reflexes organized in the dorsal vagal complex of the brain stem participate in many responses that support and potentiate energy intake and metabolism. Spinal sympathetic reflexes influence energy expenditure as well as energy intake.

Visceral afferent projections in the CNS are organized not only to establish short-loop autonomic reflexes in the caudal brain stem and spinal cord; they are also organized to project heavily throughout the entire visceral neuraxis, including the more rostral limbic and cortical circuits that constitute the control systems of energy homeostasis. Such circuits in turn project caudally to modulate the autonomic as well as somatic outflows that produce, respectively, visceral and behavioral adjustments.

Visceral Pathways Participate in the Central Integration of Energy Homeostasis

The evolving understanding of visceral afferents and efferents, complemented by a developing appreciation that energy homeostasis can best be understood in terms of control-system networks, changes not only ideas about PNS and ENS roles in energy balance, but also revises perspectives on CNS mechanisms. Many of the early models of neural control of energy intake and expenditure assigned all integrative functions to the brain, with the different inputs (humoral, hormonal, etc.), including visceral afferent inputs as well, as so many static bits of information that were not processed and integrated until they were factored into a central or hypothalamic decision that then generated motor response commands and visceromotor programs. These models explicitly or implicitly emphasized a hierarchical view of neuraxis organization and executive decision-making model of homeostasis.

In contrast to the older view, the control-system view recognizes that the circuitry of energy balance consists of multiple neural loops organized as distributed, in-parallel networks. As energy-related inputs transduced by visceral afferents are relayed to the brain, these enteroceptive signals are modulated by gut hormones and other peripheral factors. Such modulation begins even at the level of the distal primary afferent processes. As the afferents carrying signals that integrate information about both the primary sensory events and the multiplicity of hormonal and humoral conditions in the peripheral milieu reach the brain, they articulate with the brain stem motor

and premotor areas and they also relay more rostrally where they are integrated with the processing occurring in other interconnected loops.

Perhaps most conspicuously, visceral afferent information is projected to hypothalamic sites, where it is integrated with the multiple humoral and endocrine factors, many of them signals of long-term energy regulation, that determine hypothalamic outflows to the somatomotor systems, the pituitary, and the descending autonomic efferent mechanisms. Incoming visceral afferent information is also relayed to visceral/gustatory cortical fields, where it presumably is then integrated with the myriad sensory representations of the external environment that are organized in the cortex. Centrifugal cortical efferents in turn project to the motor and autonomic networks found in the brain stem. Through other projections into multiple limbic sites as well as to cortex, the signals ascending in the visceral neuraxis are integrated with conditioned and learned contingencies that reflect the individual's experiences in procuring, ingesting, and utilizing energy. Additionally, through their proximity to the ventricular system and the circumventricular organs, many of the neuronal relays of the visceral neuraxis have access to the endocrine, metabolic and other factors that accumulate in cerebrospinal fluid.

As enteroceptive signals are relayed centripetally throughout the distributed networks that control energy homeostasis, and then again as the effector signals are distributed centrifugally to orchestrate motor and autonomic adjustments, the relay nuclei appear to serve as stations where reentrant collateral projections and humoral signals can adjust the processing of neural responses. Many of the neurons in the relevant neuronal relay stations of the visceral neuraxis in the brain stem, hypothalamus, and limbic system (e.g., the amygdala) express receptors for metabolic hormones, thus serving as additional nodes at which integration can occur.

In this system, inputs are as integral to the system operation as is any other element, and inputs can be as dynamic and as subject to changes in gain as any can other element of the control system. Though 'central' and 'peripheral' divisions have a logic in terms of descriptions of the nervous system, they have little significance in terms of control systems. Visceral afferents provide many of the key inputs or paths that constitute the feed-forward and feedback loops in the integrative control strategies that the nervous system uses to achieve homeostasis. Central integrators are controllers that adjust fluctuating feed-forward and feedback signals of the closed loops to reflect different targets, goals, and disturbances the system receives. Thus, these controllers operate on and through the feed-forward and feedback loops consisting of the visceral inputs and outputs or controlled variables.

Evolving Views of Visceral Circuits and Their Dynamic Roles in Energy Homeostasis

In summary, older views of the neural mechanisms involved in energy intake and expenditure traditionally adopted models that explained intake in terms of feeding and satiety centers in the hypothalamus and explained energy expenditure in terms of brain centers that controlled activity and the sympathetic efferent outflows to a variety of tissues. In the case of both intake and expenditure, the roles of afferent inputs were assumed to be limited. The afferents were taken to be relatively modest in number and distribution, the information they supplied was assumed to be unintegrated until it reached higher CNS stations, and the linkages of the inputs to the neural centers of execution were relatively tenuous.

Newer neuroscience views of regulation of energy homeostasis are more consonant with control-systems analyses in which the myriad responses that constitute intake and expenditure are the product of neural networks that operate as feed-forward, feedback, and adaptive controllers. These controllers are multiply interconnected or cross-linked by axon collaterals and projection pathways. The control loops are also functionally cross-coupled by shared mechanical, chemical, paracrine, hormonal, and humoral signals. The newer views include the recognition that visceral afferents are more extensive and more extensively distributed than previously assumed. These views also recognize that the visceral afferents dynamically integrate – even as distally as their peripheral neurites in the walls of the viscera – a much more complete array of energy-based signals than traditionally appreciated. And finally, newer perspectives see the visceral afferents participating in energy homeostasis not as remote neural channels conveying limited static information, but as integral components of the neural control systems that participate in the integrative processing of extensive arrays of nutritive and metabolic signals.

See also: Dorsal Root Ganglion Neurons; Enteric Nervous System: Sensory Pathways; Enteric Nervous System: Neural Circuits and Chemical Coding; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Neuroendocrine Control of Energy Balance (Central Circuits/Mechanisms); Neuropeptides: Enteric Nervous System; Neuropeptides: Food Intake.

Further Reading

Berthoud H-R (2002) Multiple neural systems controlling food intake and body weight. *Neuroscience and Biobehavioral Reviews* 26: 393–428.

Blessing WW (1997) *The Lower Brainstem and Bodily Homeostasis*. New York: Oxford University Press.

Brookes S and Costa M (2002) *Innervation of the Gastrointestinal Tract*. London: Taylor & Francis.

Cameron OG (2002) *Visceral Sensory Neuroscience*. Oxford: Oxford University Press.

Cervero F and Morrison JFB (1986) *Visceral Sensation*. Amsterdam: Elsevier.

Furness JB, Kunze WAA, and Clerc N (1999) Nutrient tasting and signaling mechanisms in the gut. II. The intestine as a sensory organ: neural, endocrine, and immune responses. *American Journal of Physiology* 277: G922–G928.

Grill HJ and Kaplan JM (2002) The neuroanatomical axis for control of energy balance. *Frontiers in Neuroendocrinology* 23: 2–40.

Grundy D and Scratcherd T (1989) Sensory afferents from the gastrointestinal tract. In: Schultz SG, Wood JD, and Rauner BB (eds.) *Handbook of Physiology, Section 6: The Gastrointestinal System, Vol. 1, Part 1*, pp. 593–620. Bethesda: American Physiological Society.

Houk JC (1988) Control strategies in physiological systems. *FASEB Journal* 2: 97–107.

Mei N (1985) Intestinal chemosensitivity. *Physiological Reviews* 65: 211–237.

Phillips RJ and Powley TL (2000) Tension and stretch receptors in gastrointestinal smooth muscle: Reevaluating vagal mechanoreceptor electrophysiology. *Brain Research Reviews* 34: 1–26.

Schwartz GJ and Moran TH (1996) Sub-diaphragmatic vagal integration of meal-related gastrointestinal signals. *Neuroscience and Biobehavioral Reviews* 20: 47–56.

Sengupta JN and Gebhart GF (1994) Gastrointestinal afferent fibers and sensation. In: Johnson LR (ed.) *Physiology of the Gastrointestinal Tract,* vol. 1, pp. 483–519. New York: Raven Press.

Taché Y and Wingate D (eds.) (1991) *Brain–Gut Interactions*. Boca Raton, FL: CRC Press.

Undem BJ and Weinreich D (eds.) (2005) *Advances in Vagal Afferent Neurobiology.* Boca Raton, FL: CRC Press.

Wang FB and Powley TL (2000) Topographic inventories of vagal afferents in gastrointestinal muscle. *Journal of Comparative Neurology* 421: 302–324.

Woods SC (2004) Gastrointestinal satiety signals – I. An overview of gastrointestinal signals that influence food intake. *American Journal of Physiology* 286: G7–G13.

Engineering Viruses for CNS studies

R Lansford, California Institute of Technology, Pasadena, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The basic concept of gene therapy is to deliver specific therapeutic nucleic acid sequences to host cells, which results in a clinically relevant phenotypic change. Viruses, nanoassemblies of nucleic acid and protein, have been engineered by evolution to penetrate host cells and to usurp the host-cell cellular machinery to permit their own replication. The basic concept of a viral vector is to harness the highly honed virus infection pathway to deliver specific therapeutic nucleic acid sequences to host cells but to avoid the expression of viral genes that lead to replication, toxicity, and illness. Viral vectors have been found to be more efficient in delivering therapeutic genes than nonviral vectors. The ability to efficiently introduce specific genes into host cells makes viruses ideal tools for studying neural questions and combating neural diseases.

The life cycle of a typical virus involves its attachment to the cell surface, the uptake of the virions into the host-cell cytosol, the liberation of the genetic material from its protein casing, the expression of viral proteins, the replication of the viral genetic material, and the assembly of nascent virions and their egress from the host cells.

Recombinant viral vectors are generated from wild-type viruses by replacing genes that encode proteins necessary for their replication and packaging with foreign transgenes such as green fluorescent protein (GFP) or a therapeutic gene. The deleted viral genes can be stably expressed when integrated into the genome of a packaging cell line or transiently expressed from a plasmid or virus so that the viral gene products can be provided *in trans*. The recombinant viral vectors ideally maintain their ability to infect and transduce the same cell types they would normally have infected.

However, several hurdles still exist for using viral vectors for gene therapy and basic research. These hurdles include the preparation of large amounts of stable high-titer vectors, reducing immune reaction against viral vectors, and improving the ability to target the viral vectors to specific tissues and regulate transgene expression in a tissue-specific manner. The goal of this brief article is to introduce the reader to the value and diversity of viral vectors for gene therapy and basic research and to outline the exciting research currently going on in the field of engineering novel viral vectors.

Deciding on the Right Viral Vector for an Experiment

There are a many choices to consider when deciding the best viral vector to use for a particular experiment (Table 1). These involve determining what is preferred in terms of tissue tropism, tissue-specific gene expression, and the stability of viral genome within the host cell. Other considerations for choosing the best viral vectors are whether it induces an immune reaction within the human patient or animal model, the ease and cost of viral vector generation and purification, and the difficulty of the recombinant DNA techniques.

In general, viral vectors are designed to be replication defective for safety purposes to infect only a founder cell and its direct progeny for fate mapping studies, to introduce more transgenic DNA into their genomes, and to avoid recognition by their host's immune systems. It is vital for those using recombinant viral vectors to consider both the nature of the virus used as a vector and the effects of any transgene, introduced genetic elements, or other modification. For instance, there are distinct differences in working with viral vectors that express a gene that encodes GFP than in working with a gene that expresses a potentially hazardous transgene such as v-src or diphtheria toxin a. Placing a hazardous transgene in a viral vector that has been pseudotyped to efficiently infect a broad host range can be dangerous to investigators and their labmates.

Tissue Tropism

The typical surface of viruses includes numerous copies of an envelope protein that binds to specific receptor molecules on the plasma membrane of the host cell. Once bound to the receptor, viruses can either fuse with the plasma membrane or enter via the endocytic pathway. The tissue tropism or specific cell that a virus can infect is generally restricted to cells that express receptors that interact with the viral envelope glycoprotein. For instance, rabies virus is specific for motor neurons due to the affinity of gC in the virion envelope to heparin sulfate on the surface of the cell.

Table 1 Characteristics of viral vectors

Viral vector	*Genetic material*	*Packaging size (kb)*	*Tropism*	*Non-dividing*	*Integration*	*Titer (virions ml^{-1})*	*Advantages*	*Disadvantages*
Adenovirus	dsDNA	8–36	Broad	Yes	No	10^{11}–10^{12}	Efficient transduction of most tissues	Immune response; short-term expression
AAV	ssDNA	<5	Broad	Yes	Episomal (>90%) Integrated (<10%)		Long-term expression	Small packaging capacity; insertional mutagenesis
Alphavirus	RNA	2–5	Neurotropic	Yes	No		High expression	High expression
HSV-1	dsDNA	40; 152	Strong for neurons, broad	Yes	No	Low	Large transgene capacity	Immune response; short-term expression
Retrovirus	RNA	8.8	Broad	No	Yes	10^{8}–10^{9}	Long-term expression	Insertional mutagenesis
Lentivirus	RNA	9.6	Broad	Yes	Yes	10^{8}–10^{9}	Long-term expression	Insertional mutagenesis
AAV–HSV TN–HSV	DNA	152	Broad	Yes	Yes		Long-term	Insertional mutagenesis

AAV, adeno-associated virus; ds, double-stranded; HSV, herpes simplex virus; ss, single-stranded.

Pseudotyping alters the host range of a virus by exchanging the surface antigens between both DNA and RNA viruses, thus generating chimeric viral vectors. The glycoprotein from vesicular stomatitis virus (VSV-G), a rhabdovirus, is a popular choice for pseudotyping viruses because of its ability to confer a broad tropism and its ability to withstand the mechanical shearing stress present in ultracentrifuges, which permits viral vectors to be concentrated 1000-fold with minimal loss of biological activity. VSV-G can be incorporated into several different classes of viral vectors, including retroviruses, lentiviruses, and herpes viruses.

Murine central nervous system (CNS) transduction patterns of lentiviral vectors pseudotyped with envelope glycoproteins from murine leukemia virus (MuLV), lymphocytic choriomeningitis virus (LCMV), or the rabies-related Mokola virus were compared to a vector pseudotyped VSV-G. Mokola-, LCMV-, and VSV-G-pseudotyped vectors transduced similar populations, including the striatum, thalamus, and white matter, whereas MuLV efficiently transduced the striatum and hippocampal dentate gyrus. These results demonstrate that targeted transduction in the CNS can be achieved using specific envelope glycoproteins to pseudotype lentiviral vectors and offer Mokola-pseudotyped and MuLV-pseudotyped lentivectors as possible alternatives to VSV-G-pseudotyped lentivectors. The ability to biotinylate the adenovirus fiber allows researchers to modify the viral tropism by incorporating a myriad of targeting peptides/proteins into the structural context of the virion.

Genome versus Episome

Integration of the viral vector genome into the host genome guarantees that the genetic material is maintained during the lifetime of the cell and its daughter cells. This is ideal for long-term gene therapy in clinical settings or cellular fate mapping in researcher settings. A potential disadvantage of genomic integration is that the viral DNA may integrate near and possibly deregulate the normal expression pattern of an essential gene, inducing deleterious effects. Episomal vectors, however, are typically lost from the infected cells if they do not contain an origin of replication (ori)-permitted higher copy numbers. The Epstein–Barr virus (EBV) replicon elements (*oriP* and *EBNA-1*) help to stabilize episomal vectors by permitting up to 20 copies per cell.

Promoter Control

Transcriptional enhancers and silencers are typically distant from the promoter sequences that they regulate. Some of these regulatory elements do not operate correctly when they are abutted too close to one another. Undefined insulator elements within the human immunodeficiency virus (HIV)-based vectors appear to permit endogenous gene expression and persistence with certain tissue-specific transcription units; however, this seems to be the exception rather than the rule within small insert viral vectors. Promoters can be silenced due to positional effects related to their chromatin integration site, by the transcription complexes of genes that may have integrated into or adjacent to them, or by DNA methylation by the host cell of their transcription elements. Gene expression can also be downregulated at the level of mRNA stability. Viral promoters and some endogenous transcription units tend to be silenced in viral vectors, especially in cells that undergo cellular differentiation and during germ line passage in transgenic animals.

Herpes simplex virus (HSV)-1 amplicons can accommodate foreign DNA up to 150 kb, which is attractive for both basic and clinical research. Genomic sequences as well as large cDNAs, large transcriptional regulatory sequences for cell-type-specific expression, or multiple transgenes can be inserted in a modular fashion.

Targeted Gene Inactivation Using RNAi

Targeted gene inactivation in mice is a common and powerful technique that researchers use to determine the functional significance of genes. RNA interference (RNAi) has emerged as a powerful technique for generating loss-of-function phenotypes by downregulating the expression of specific genes in plants and animals. The basic concept is that the antisense sequence hybridizes with the normal mRNA in the cell and either blocks translation of the mRNA or increases the degradation of the host mRNA. RNAi typically regulates gene expression in cells via microRNAs, which are abundant in the nervous system.

Several of the ways that neuroscientists are currently using RNAi for both basic and therapeutic research have been recently well reviewed. Long-term RNAi may be efficaciously accomplished with adeno-associated viruses (AAVs), adenoviruses (Ads), and lentiviruses that have been engineered to deliver inhibitory RNAs (siRNAs) to the nervous system. siRNAs can be produced as short hairpin transcripts from RNA polymerase (RNAP) II and III promoters.

Engineering the Principal Viral Vectors

Adenovirus Vectors

Ad infections in humans leads to coldlike symptoms. Ad is a DNA virus containing a linear double-stranded

DNA genome of approximately 36 kb, of which up to 30 kb can be replaced with foreign DNA (Table 1). Neutralizing sera have been used to identify 49 different serotypes of Ad that are divided into six groups (A–F) on the basis of genome size, composition, homology, and organization. Cellular uptake of Ad involves binding of the Ad fiber coat protein with a cell surface receptor, including the major histocompatibility complex (MHC) class I molecule, the Coxsackie B viruses and Ad receptor, and the integrin family of cell surface heterodimers. The Ad fiber coat has been successfully modified using several ligands to alter the tissue tropism and specifically target the viral vector to desired cells and tissues.

The advantages of using Ad vectors include their broad host range, their ability to infect and deliver transgenes to both dividing and nondividing cells, and their ability to replicate episomally within the host nucleus. However, this also means that Ad expression is relatively short-lived (5–10 days) because the proportion of transduced cells decreases with each cell division or, possibly, due to immune intervention. Furthermore, they can be concentrated to high titers (10^{11}–10^{12} pfu ml^{-1}) and manufactured on a large scale.

Ad vectors are typically derived from subgroup C, serotypes 2 or 5. The first generation of Ad vectors contained deletions for one or two early (E) genes (E1 and E3). Deletion of the E1A and part of the E1B gene results in replication-deficient viruses that must be propagated in helper cell lines that provide the deleted E1 gene products *in trans*. The amount of DNA that can be inserted into Ad virions cannot exceed ~105% of the wild-type genome, which permits the insertion of only ~2 kb of foreign DNA into full-length Ad vectors, 5–6 kb of foreign DNA into E1A deleted vectors, or 7–8 kb of foreign DNA into E1- and E3-deleted vectors. Additional vectors have been generated that contain an E2a temperature-sensitive mutant or an E4 deletion. Recently, vectors have been produced that contain only the inverted terminal repeats (ITRs) and a packaging sequence around the transgene, with all the necessary viral genes being provided *in trans* by a helper virus to minimize recognition of the Ad vector by the immune system. Finally, high-capacity (hc) Ad vectors that permit gene transfer into dividing and nondividing cells but lack Ad viral genes that makes them more spacious, safer, and better suited for prolonged transgene expression have been reported.

Two disadvantages of many existing Ad vectors are caused by the induction of a T cell-mediated response against Ad proteins, which causes an inflammatory reaction, resulting in lysis of the transduced cells, and shortens the duration of transgene expression. The transient nature of gene expression and the induction of host immune responses limit the efficacy of Ad gene delivery and oblige new research to restore user confidence.

Adeno-Associated Virus

AAV vectors are promising therapeutic gene carriers because they can transduce nondividing cells, lack viral genes encoding immunogenic proteins, can be prepared at high titers, and achieve long-term transgene expression in animal models. AAV is classified as a separate genus in the family Parvoviridae, designated *Dependovirus* because it requires the presence of a wild-type Ad or herpes virus for productive infection; in the absence of a helper virus, AAV stably integrates into the host cell genome. There are more than 40 AAV serotypes. Although approximately 80% of humans are seropositive for AAV2 (serotype 2), the virus has not been linked to any human illness.

AAV is composed of a linear single-stranded DNA genome of approximately 4.7 kb and associated with a nonenveloped icosahedral capsid of approximately 22 nm (Table 1). Two open reading frames (ORFs), cap (which encodes the structural proteins that form the capsid) and rep (which encodes the regulatory proteins), are flanked by ITRs. The AAV site specifically integrates its DNA into chromosome 19q13-qter of the human genome via the interaction of AAV Rep78 or Rep68 proteins with Rep binding sites in the ITRs. Site-specific integration of AAV is unique among eukaryotic viruses, and its exploitation holds considerable promise for the targeted integration and stable expression of therapeutic DNA. Transduction does not appear to require cell division. The primary limitations of AAV vectors are their small cloning capacity and their elicitation of a humoral immune response.

AAV vectors exhibit a number of properties that have made this vector system an excellent choice for both CNS gene therapy and basic neurobiological investigations. *In vivo*, the preponderance of AAV vector transduction occurs in neurons, where it is possible to obtain long-term, stable gene expression with very little accompanying toxicity. Promoter selection, however, significantly influences the pattern and longevity of neuronal transduction distinct from the tropism inherent to AAV vectors.

AAV vectors have used to manipulate CNS function using a wide variety of approaches, including the expression of foreign genes, expression of endogenous genes, expression of antisense RNA, and expression of RNAi. For instance, AAV-mediated transduction of murine neurons has been seen as early as 7 days after stereotaxic injection of the virus into the hippocampus, striatum, and septum and reported

to last for at least 6 months. Interestingly, AAV can be retrogradely transported efficiently from muscle to the motor neurons of the spinal cord, permitting the long-term delivery of insulin-like growth factor 1 to a mouse model for amyotropic lateral sclerosis (ALS), which prolonged life and delayed disease progression. The different AAV serotypes increase the potential patterns of *in vivo* vector transduction, offering alternatives to the often-used AAV2 serotype that can be silenced by the immune system. These advantageous CNS properties of AAV vectors have fostered a wide range of clinically relevant applications, including Parkinson's disease, lysosomal storage diseases, Canavan's disease, epilepsy, Huntington's disease, and ALS.

Alphavirus Vectors

Alphavirus is a neurotropic positive linear single-stranded RNA virus that can cross the blood–brain barrier. Its plus-strand RNA genomes self-replicate within the cytosol and express high levels of protein. Viral vectors that express transgenes have been generated from the alphaviruses Semliki Forest virus (SFV) and Sindbis virus.

Cloning is straightforward with alphavirus genomes, and infectious RNA can be generated *in vivo* or *in vitro*. The virus releases positive-strand RNA into the host-cell cytosol, permitting rapid, high levels of gene expression, in dividing or nondividing cells (Table 1). Alphavirus replicons can express as much as 25% of the protein of a cell over a period of 72 h. That said, a potential disadvantage of these vectors is that they can induce cell death via apoptosis and inhibit macromolecular synthesis due to such high levels of protein synthesis. However, they may be useful as suicide vectors and may be desirable in certain experimental and clinical scenarios.

Alphaviruses efficaciously infect neuronal cells both *in vitro* and *in vivo* and express high levels of their transgene, such as GFP. The high levels of GFP expression in the cortical and hippocampal neurons permit dynamic imaging of delicate dendrites and neuronal tracing. In addition, new Sindbis-based vectors that delay the inhibition of protein synthesis due to a point mutation in a nonstructural protein delayed the onset of viral-induced cytotoxicity, and higher levels of enhanced GFP (eGFP) expression permit high-resolution dynamic imaging.

Herpes Simplex Virus Type 1

HSV-1 is a neurotropic double-stranded DNA virus that possesses several features suitable for application in the CNS. In an unmodified form, HSV-1 is a human pathogen, causing recurrent cold sores, keratitis, and, infrequently, severe encephalitis. Infection of wild-type HSV-1 on the epithelial cells of the skin or mucous membrane often results in a lytic viral replication cycle. Viruses released at the site of the primary infection may enter local sensory nerve terminals. Enveloped HSV virions can be actively transported anterogradely and retrogradely in neuronal axons. In the neuron, HSV-1 may enter into another lytic replication cycle that leads to neuronal cell death and egress of infectious particles, or the viral DNA may enter into a latent state in which the viral genome persists as stable episomal element.

HSV-1 is the prototype and best-studied virus of the a-herpes virus group. HSV-1 virions consist of four structural components: DNA, capsid, tegument, and envelope. The icosahedral capsid with a diameter of 125 nm surrounds a 153-kb double-stranded viral DNA that encodes over 70 genes (Table 1). The tegument is an amorphous coat of ~20 proteins that is localized between the capsid and the viral envelope, which contains ~12 membrane proteins.

After fusion of the viral envelope with the plasma membrane, HSV-1 capsids are propelled by dyenin along microtubules to the nuclear pore channel (NPC). The capsids of HSV-1 bind to the cytosolic side of the NPC, a vertex in the capsid is opened, and the DNA passes through the NPC, leaving behind an intact empty capsid that is liberated from the NPC. The mechanisms of how the HSV-1 genome is translocated through the NPC and even its genomic state (linear, circular, or concatemer) within the nucleus are poorly understood.

The HSV-1 amplicon vector is a plasmid-based expression system equipped with the full infection machinery of its parental virus. Advances of the helper virus-free packaging systems have eliminated many of the concerns over safety and toxicity. The HSV-1-based system, which rapidly and efficiently delivers and expresses bacterial artificial chromosome (BAC) library clones in human and mouse cells by converting them into infectious amplicon vectors, has been described. An HSV-1 amplicon is a bacterial plasmid that contains one or more transgene cassette (s) and two noncoding viral sequences: an origin of DNA replication (ori) and a DNA cleavage/packaging signal (pac). Homologous recombination methods in *Escherichia coli* permit genetic alterations with nucleotide precision to be carried out in BAC inserts. HSV-1 amplicon vectors possess several advantages as gene delivery vehicles, including (1) their large transgene capacity, up to 150 kb; (2) their ability to infect a wide variety of cell types across a broad range of species, both dividing and nondividing, without being infectious; (3) the ease of vector construction; and (4) their limited cytotoxicity and immunogenicity because of their lack of viral coding sequences. The presence of

prokaryotic DNA sequences within the packaged HSV amplicon triggers transcriptional silencing of the entire vector sequence, suggesting that generating amplicons devoid of bacterial sequences will improve long-term expression.

Wade-Martins and colleagues described a system to rapidly and efficiently deliver and express BAC library clones in human and mouse cells by converting them into infectious amplicon vectors. HSV-1-based amplicons were used to package and deliver into cells genomic inserts of >130 kb with efficiencies of up to 100%, as they demonstrated with a complete human hypoxanthine phosphoribosyltransferase (HPRT) locus contained within a 115 kb BAC insert. HSV-1-based amplicons were also exploited to deliver a complete genomic DNA locus encoding the human low-density lipoprotein receptor (LDLR) into mutant Chinese hamster ovary (CHO) cell lines and into human fibroblasts derived from patients with familial hypercholesterolemia (FH). These infectious genomic locus vectors appear to supply physiologically appropriate levels of the LDLR transgene product. The primary limitations of the HSV amplicon vector for human gene therapy are the transient nature of amplicon-mediated transgene expression and the absence of a scalable method to produce a high-titer vector. A comprehensive discourse that explores the several strategies that employ HSV amplicon vectors as tools for the dissection of neural network function and as therapeutics directed against a variety of neurological disorders has been published.

Lentivirus Vectors

HIV-1 is the etiological agent of acquired immunodeficiency syndrome. The ~9 kb HIV-1 RNA genome contains nine ORFs which encode 15 proteins. Lentiviruses, in contrast to most retroviruses, can infect dividing and nondividing cells when viral proteins guide the proviral complex through the nuclear pore.

After binding to its receptor, the viral capsid containing the diploid RNA genome enters the cell by membrane fusion. The viral RNA genome is next converted into a double-stranded proviral DNA by the viral enzyme reverse transcriptase. The proviral DNA is complexed with the nucleocapsid, reverse transcriptase, and integrase viral proteins. Dynamic imaging of HIV particles in living cells revealed that cytoplasmic dyenin and the microtubule network by HIV to migrate toward the nucleus. The HIV particles accumulate in the perinuclear region and access the nucleus during mitosis or by active transport through the NPC. The HIV-1 preintegration complex has karyophilic properties due to the presence of nuclear localization signals (NLSs) in the viral proteins matrix and integrase, which permit nuclear import in the absence of mitosis. After translocation into the nucleus, the viral enzyme integrase mediates the integration of the provirus into the host-cell genome.

Lentiviral vectors as gene delivery vehicles have been shown to confer efficient and long-lasting, *in vivo* gene transfer into a variety of cell types. What makes lentiviral vectors particularly useful is that they are able to stably express the delivered genes in a tissue-specific manner in differentiating cells and in the germ cells of transgenic animals. Lentiviral vectors have been used to generate transgenic mice, rats, quails, chickens, and monkeys very efficiently, with the vast majority of the animals expressing high levels of GFP. Lentiviral vectors injected into the mouse striatum may induce an immune response against the transgenic, but not viral, antigens. Current viral helper or packaging cell lines use separate plasmids for a pseudotyped *env* gene and for supplying the viral structural and regulatory genes *in trans*.

Several studies have used lentivectors to infect regions of the CNS. Lentiviral vectors have been used to infect cortical neurons *in vivo* with several neuron-specific promoters, including the α-calcium/calmodulin-dependent protein kinase II and the synapsin I promoters. This permitted the structural dynamics of dendritic spines and axonal projections in single cortical cells to be monitored using high-resolution two-photon time-lapsed imaging. In another study, lentivectors were injected into the subventricular zone of the adult mouse brain to mark and prove that neural precursor cells display self-renewal and multipotency, two essential characteristics of neural stem cells (NSCs). Thus, lentivectors may be an efficient way to introduce therapeutic genes into NSCs to ameliorate a broad range of CNS diseases.

Engineering Chimeric Viral Vectors

The pursuit for better viral vectors has led many researchers to combine the best features of different viruses into chimeric vectors. Chimeric viral vectors provide a system for the production of viral vectors whose genome is encoded within another virus with a different life cycle and distinct biological characteristics. Many intriguing chimeric viral vectors merge HSV vectors, due to their large insert capacity, and viral elements that permit genomic integration.

Chimeric HSV-1–AAV Amplicon Vectors

HSV–AAV-based chimeric vectors can mediate highly efficient and precise site-specific integration of large therapeutic transgenes into the human genome. HSV-1–AAV amplicon vectors contain the replication origin (OriS) and the packaging signal

(pac) of HSV-1, along with the integrative elements of AAV, including the rep gene and the ITRs. Vectors based on HSV-1 permit the delivery of transgenes of up to 150 kb while the ITRs and Rep of AAV can confer site-specific integration into the AAVS1 site chromosome 19 in human cells, which allows sustained expression of a transgene.

HSV–AAV chimeric vectors have been used for the infectious transfer of the human lysosomal β-galactosidase (β-gal) gene of 100 kb. Sustained expression of β-gal was reported in cells infected with rep-plus, but not rep-minus, HSV–AAV hybrid vectors. Approximately, 80% of the rep-plus clones displayed β-gal activity above normal levels for up to 4 months of constant growth, and ~33% of the clones were found to have undergone AAVS1-specific integration of the ITR-flanked transgene. These findings are promising in that HSV–AAV chimeric vectors neatly merge the site-specific integration capabilities of AAV elements with the large genome-carrying capacity of HSV1, although the percentage of AAVS1-specific integration needs to be higher to ensure a low probability of insertional mutagenesis.

One remaining hurdle of HSV–AAV chimeric vectors is that HSV-1 replication is inhibited by the rep genetic elements from AAV, resulting in low titers. The data suggest that high levels of Rep protein expression during vector packaging are responsible for the titer reduction. By changing the promoters that drive rep genes in or removing them altogether from HSV–AAV chimeric vectors, Fraefel's lab has obtained higher viral titers.

Chimeric TN–HSV Amplicon Vectors

Another approach taken to confer stable, long-term transgene expression on the HSV amplicon vector DNA has been to incorporate the *Sleeping Beauty* (SB) DNA transposon system into an HSV-1 amplicon vector. The chimeric SB–HSV amplicon vector comprises two separate amplicon vectors: one expressing the SB transposase and the other carrying an expression cassette flanked by an inverted repeat/direct repeat (IR/DR) from the SB transposon. A 25-fold increase in the frequency of stable transduction in cell lines was observed with integration-competent SB–HSV amplicons compared to typical episomal HSV amplicons. They also demonstrated that the vector system efficiently integrated the IR/DR-flanked expression cassette randomly into the genomes of nondividing primary mouse neurons in culture, thereby leading to sustained transgene expression. *In utero* injection of the SB–HSV amplicon vectors into the lateral ventricle of mouse brains on embryonic day (E)14.5 led to widespread neural transgene expression in postnatal-day-90 brains, including the hippocampus, dentate gyrus, subventricular zone, septofimbrial nucleus, and the motor cortices. The authors concluded that the target of vector transduction on E14.5 must have primarily been multipotent neural stem/progenitor cells residing along the ventricular zone.

Chimeric hcAd–AAV Vectors

The hcAd–AAV vector comprises linear monomeric DNA molecules consisting of the Ad origin of replication and packaging signal and the AAV ITRs and Rep; it can confer site-specific integration into the AAVS1 site chromosome 19 in human cells, which allows sustained expression of a transgene. Packaging cells are infected with a helper Ad vector to provide Ad DNA replication machinery, which along with AAV ITR-dependent dimerization leads to the assembly of vector genomes with a tail-to-tail configuration that are efficiently amplified and packaged into Ad capsids. Chimeric hcAd–AAV vectors have been used to express a 14 kb dystrophin cDNA in rat cardiomyocytes *in vitro* and to restore dystrophin synthesis in the muscle tissues of mdx mice *in vivo*.

Chimeric Lentivirus–Ad Vectors

Ad–retroviral chimeric vectors were developed first; however, their efficacy was limited because neighboring cells must be dividing for the second stage of infection by the retrovirus to occur. To overcome this limitation, lentiviral elements are often used instead of retroviral elements. The nonintegrative, high-titer, broad tropism Ad is able to deliver DNA that becomes integrated into neighboring cells by the lentivector intermediary in the current lentivirus–Ad chimeric vectors. Because both the Ad and lentivirus can infect dividing or nondividing cells, these chimeric vectors should permit the efficient integration of a transgene into nondividing neighboring cells *in vivo*.

Conclusion

Considerable progress has been made in generating novel viral vectors that may have profound therapeutic effects in the CNS in the years to come, but considerable advances still need to be made. The ability to temporally and spatially introduce therapeutic genes into a diseased CNS in a highly efficient manner indicates that viral vectors will play a vital role in future gene therapies. Although the highly efficient nature of viral-mediated gene delivery is alluring and intoxicating, the repercussions of a gene therapy experiment gone awry can be severe. A more comprehensive understanding of the basic biology of viruses will lead to better and safer viral

vectors in the future. A wide range of clinically relevant applications, including Parkinson's disease, lysosomal storage diseases, Canavan's disease, epilepsy, Huntington's disease, and ALS, are currently or will soon be in clinical trials. However, each individual treatment presents a unique set of difficulties that must be deciphered in order to attain clinically effective gene therapies.

See also: BAC Transgenesis: Cell-Type Specific Expression in the Nervous System; BAC Use in the Study of the CNS; Gene Therapy: Genetically Modified Cells; Gene Therapy: Direct Viral Delivery; Viral Vectors in the CNS.

Further Reading

Abbas-Terki T, Blanco-Bose W, Deglon N, Pralong W, and Aebischer P (2002) Lentiviral-mediated RNA interference. *Human Gene Therapy* 13: 2197–2201.

Abordo-Adesida E, Follenzi A, Barcia C, et al. (2005) Stability of lentiviral vector-mediated transgene expression in the brain in the presence of systemic antivector immune responses. *Human Gene Therapy* 16: 741–751.

Armentano D, Zabner J, Sacks C, et al. (1997) Effect of the E4 region on the persistence of transgene expression from adenovirus vectors. *Journal of Virology* 71: 2408–2416.

Bergelson JM, Cunningham JA, Droguett G, et al. (1997) Isolation of a common receptor for Coxsackie B viruses and adenoviruses 2 and 5. *Science* 275: 1320–1323.

Blomer U, Naldini L, Kafri T, Trono D, Verma IM, and Gage FH (1997) Highly efficient and sustained gene transfer in adult neurons with a lentivirus vector. *Journal of Virology* 71: 6641–6649.

Bouyac-Bertoia M, Dvorin JD, Fouchier RA, Jenkins Y, and Meyer BE (2001) HIV-1 infection requires a functional integrase NLS. *Molecular and Cellular Biology* 7: 1025–1035.

Bowers WJ, Mastrangelo MA, Howard DF, Southerland HA, Maguire-Zeiss KA, and Federoff HJ (2006) Neuronal precursor-restricted transduction via *in utero* CNS gene delivery of a novel bipartite HSV amplicon/transposase hybrid vector. *Molecular Therapy* 13: 580–588.

Brockstedt DG, Podsakoff GM, Fong L, Kurtzman G, Mueller-Ruchholtz W, and Engleman EG (1999) Induction of immunity to antigens expressed by recombinant adeno-associated virus depends on the route of administration. *Clinical Immunology* 92: 67–75.

Bukrinsky MI, Haggerty S, Dempsey MP, Sharova N, and Adzhubel A (1993) A nuclear localization signal within HIV-1 matrix protein that governs infection of non-dividing cells. *Nature* 365: 666–669.

Bukrinsky MI, Sharova N, Dempsey MP, et al. (1992) Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. *Proceedings of the National Academy of Sciences of the United States of America* 89: 6580–6584.

Burns JC, Friedmann T, Driever W, Burrascano M, and Yee JK (1993) Vesicular stomatitis-virus G glycoprotein pseudotyped retroviral vectors – concentration to very high-titer and efficient gene-transfer into mammalian and nonmammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* 90: 8033–8037.

Campos SK, Parrott MB, and Barry MA (2004) Avidin-based targeting and purification of a protein IX-modified, metabolically biotinylated adenoviral vector. *Molecular Therapy* 9: 942–954.

Cardone G, Winkler DC, Trus BL, et al. (2006) Visualization of the herpes simplex virus portal *in situ* by cryo-electron tomography. *Virology* 361: 426–434.

Cavazzana-Calvo M, Hacein-Bey S, Basile GDS, et al. (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* 288: 669–672.

Chapman SC, Lawson A, MacArthur WC, et al. (2005) Ubiquitous GFP expression in transgenic chickens using a lentiviral vector. *Development* 132: 935–940.

Chen BE, Lendvai B, Nimchinsky EA, Burbach B, Fox K, and Svoboda K (2000) Imaging high-resolution structure of GFP-expressing neurons in neocortex *in vivo*. *Learning & Memory* 7: 433–441.

Chen HH, Mack LM, Kelly R, Ontell M, Kochanek S, and Clemens PR (1997) Persistence in muscle of an adenoviral vector that lacks all viral genes. *Proceedings of the National Academy of Sciences of the United States of America* 94: 1645–1650.

Coffin JM, Hughes SH, and Varmus HE (1997) Retroviruses. Plainview, NY: Cold Spring Harbor Laboratory Press.

Consiglio A, Gritti A, Dolcetta D, et al. (2004) Robust *in vivo* gene transfer into adult mammalian neural stem cells by lentiviral vectors. *Proceedings of the National Academy of Sciences of the United States of America* 101: 14835–14840.

Davidson BL and Breakefield XO (2003) Viral vectors for gene delivery to the nervous system. *Nature Reviews Neuroscience* 4: 353–364.

Davidson BL and Boudreau RL (2007) RNA interference: A tool for querying nervous system function and an emerging therapy. *Neuron* 53: 781–788.

Deglon N and Aebischer P (2002) Lentiviruses as vectors for CNS diseases. *Current Topics in Microbiology and Immunology* 261: 191–209.

Dittgen T, Nimmerjahn A, Komai S, et al. (2004) Lentivirus-based genetic manipulations of cortical neurons and their optical and electrophysiological monitoring *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* 101: 18206–18211.

Dull T, Zufferey R, Kelly M, et al. (1998) A third-generation lentivirus vector with a conditional packaging system. *Journal of Virology* 72: 8463–8471.

Engelhardt JF, Ye X, Doranz B, and Wilson JM (1994) Ablation of E2A in recombinant adenoviruses improves transgene persistence and decreases inflammatory response in mouse liver. *Proceedings of the National Academy of Sciences of the United States of America* 91: 6196–6200.

Fire A, Xu S-Q, Montgomery MK, Kostas SA, Driver SE, and Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391: 806–811.

Flint SJ, Enquist LW, Krug RM, Racaniello VR, and Skalka AM (2000) *Principles of Virology*. Washington, DC: ASM Press.

Frankel AD and Young JA (1998) HIV-1: Fifteen proteins and an RNA. *Annual Review of Biochemistry* 67: 1–25.

Frenkel N (2006) The history of the HSV amplicon: From naturally occurring defective genomes to engineered amplicon vectors. *Current Gene Therapy* 6: 277–301.

Gallay PHT, Chin D, and Trono D (1997) HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. *Proceedings of the National Academy of Sciences of the United States of America* 94: 9825–9830.

Geller AI and Breakefield XO (1998) A defective HSV-1 vector expresses *Escherichia coli* beta-galactosidase in cultured peripheral neurons. *Science* 241: 1667–1669.

Glauser DL, Ackermann M, Saydam O, and Fraefel C (2006) Chimeric herpes simplex virus/adeno-associated virus amplicon vectors. *Current Gene Therapy* 6: 315–324.

Goncalves MA, van Nierop GP, Tijssen MR, et al. (2005) Transfer of the full-length dystrophin-coding sequence into muscle cells by a dual high-capacity hybrid viral vector with site-specific integration ability. *Journal of Virology* 79: 3146–3162.

Gong S, Zheng C, Doughty ML, et al. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425: 917–925.

Heister T, Heid I, Ackermann M, and Fraefel C (2002) Herpes simplex virus type 1/adeno-associated virus hybrid vectors mediate site-specific integration at the adeno-associated virus preintegration site, AAVS1, on human chromosome 19. *Journal of Virology* 76: 7163–7173.

Hong SS, Karayan L, Tournier J, Curiel DT, and Boulanger PA (1997) Adenovirus type 5 fiber knob binds to MHC class I alpha2 domain at the surface of human epithelial and B lymphoblastoid cells. *EMBO Journal* 16: 2294–2306.

Jackson SA and DeLuca NA (2003) Relationship of herpes simplex virus genome configuration to productive and persistent infections. *Proceedings of the National Academy of Sciences of the United States of America* 100: 7871–7876.

Kaspar BK, Vissel B, Bengoechea T, et al. (2002) Adeno-associated virus effectively mediates conditional gene modification in the brain. *Proceedings of the National Academy of Sciences of the United States of America* 99: 2320–2325.

Kaspar BK, Llado J, Sherkat N, Rothstein JD, and Gage FH (2003) Retrograde viral delivery of IGF-1 prolongs survival in a mouse ALS model. *Science* 301: 839–842.

Kim J, Dittgen T, Nimmerjahn A, et al. (2004) Sindbis vector SINrep(nsP2S726): A tool for rapid heterologous expression with attenuated cytotoxicity in neurons. *Journal of Neuroscience Methods* 133: 81–90.

Kootstra NA and Verma IM (2003) Gene therapy with viral vectors. *Annual Review of Pharmacology and Toxicology* 43: 413–439.

Kosik KS (2006) The neuronal microRNA system. *Nature Reviews Neuroscience* 7: 911–920.

Kotin RM, Menninger JC, Ward DC, and Berns KI (1991) Mapping and direct visualization of a region-specific viral DNA integration site on chromosome 19q13-qter. *Genomics* 10: 831–832.

Liljestrom P and Garoff H (1991) A new generation of animal cell expression vectors based on the Semliki Forest virus replicon. *Biotechnology* (New York) 9: 1356–1361.

Lois C, Hong EJ, Pease S, Brown EJ, and Baltimore D (2002) Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. *Science* 295: 868–872.

Lundstrom K (1999) Alphaviruses as tools in neurobiology and gene therapy. *Journal of Receptors and Signal Transduction Research* 19: 673–686.

Maletic-Savatic M, Malinow R, and Svoboda K (1999) Rapid dendritic morphogenesis in CA1 hippocampal dendrites induced by synaptic activity. *Science* 283: 1923–1927.

McCown TJ (2005) Adeno-associated virus (AAV) vectors in the CNS. *Current Gene Therapy* 5: 333–338.

McDonald D, Vodicka MA, Lucero G, et al. (2002) Visualization of the intracellular behavior of HIV in living cells. *Journal of Cell Biology* 159: 441–452.

McGrew MJ, Sherman A, Ellard FM, et al. (2004) Efficient production of germline transgenic chickens using lentiviral vectors. *EMBO Reports* 5: 728–733.

Miller DG, Adam MA, and Miller AD (1990) Gene transfer by retrovirus vectors occurs only in cells that are actively replicating at the time of infection. *Molecular and Cellular Biology* 10: 4239–4242.

Naldini L, Blömer U, Gallay P, et al. (1996) *In vivo* gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* 272: 263–267.

Naldini L, Blomer U, Gage FH, Trono D, and Verma IM (1996) Efficient transfer, integration, and sustained long-term expression of the transgene in adult rat brains injected with a lentiviral vector. *Proceedings of the National Academy of Sciences of the United States of America* 93: 11382–11388.

O'Connor TP and Crystal RG (2006) Genetic medicines: Treatment strategies for hereditary disorders. *Nature Reviews Genetics* 7: 261–276.

Oehmig A, Cortés ML, Perry KF, Sena-Esteves M, Fraefel C, and Breakefield XO (2007) Integration of active human beta-galactosidase gene (100 kb) into genome using HSV/AAV amplicon vector. *Gene Therapy* 14: 1078–1091.

Oehmig A, Fraefel C, Breakefield XO, and Ackermann M (2004) Herpes simplex virus type 1 amplicons and their hybrid virus partners, EBV, AAV, and retrovirus. *Current Gene Therapy* 4: 385–408.

Ojala PM, Sodeik B, Ebersold MW, Kutay U, and Helenius A (2000) Herpes simplex virus type 1 entry into host cells: Reconstitution of capsid binding and uncoating at the nuclear pore complex *in vitro*. *Molecular and Cellular Biology* 20: 4922–4931.

Palfi S, Leventhal L, Chu Y, et al. (2002) Lentivirally delivered glial cell line-derived neurotrophic factor increases the number of striatal dopaminergic neurons in primate models of nigrostriatal degeneration. *Journal of Neuroscience* 22: 4942–4954.

Qin L, Ding Y, Pahud DR, Chang E, Imperiale MJ, and Bromberg JS (1997) Promoter attenuation in gene therapy: Interferon-gamma and tumor necrosis factor-alpha inhibit transgene expression. *Human Gene Therapy* 8: 2019–2029.

Roe T, Reynolds TC, Yu G, and Brown POI (1993) Integration of murine leukemia virus DNA depends on mitosis. *EMBO Journal* 12: 2099–2108.

Saeki Y, Fraefel C, Ichikawa T, Breakefield XO, and Chiocca EA (2001) Improved helper virus-free packaging system for HSV amplicon vectors using an ICP27-deleted, oversized HSV-1 DNA in a bacterial artificial chromosome. *Molecular Therapy* 3: 591–601.

Samulski RJ, Zhu X, Xiao X, et al. (1991) Targeted integration of adeno-associated virus (AAV) into human chromosome 19. *EMBO Journal* 10: 3941–3950.

Scott BB and Lois C (2005) Generation of tissue-specific transgenic birds with lentiviral vectors. *Proceedings of the National Academy of Sciences of the United States of America* 102: 16443–16447.

Smith AE (1995) Viral vectors in gene therapy. *Annual Review of Microbiology* 49: 807–838.

Sodeik B (2000) Mechanisms of viral transport in the cytoplasm. *Trends in Microbiology* 8: 465–472.

Sodeik B, Ebersold MW, and Helenius A (1997) Microtubule-mediated transport of incoming herpes simplex virus 1 capsids to the nucleus. *Journal of Cell Biology* 136: 1007–1021.

Spaete RR and Frenkel N (1982) The herpes simplex virus amplicon: A new eucaryotic defective-virus cloning-amplifying vector. *Cell* 30: 295–304.

Suzuki M, Kasai K, and Saeki Y (2006) Plasmid DNA sequences present in conventional herpes simplex virus amplicon vectors cause rapid transgene silencing by forming inactive chromatin. *Journal of Virology* 80: 3293–3300.

Thomas CE, Ehrhardt A, and Kay MA (2003) Progress and problems with the use of viral vectors for gene therapy. *Nature Reviews Genetics* 4: 346–358.

Tyler CM, Wuertzer CA, Bowers WJ, and Federoff HJ (2006) HSV amplicons: Neuro applications. *Current Gene Therapy* 6: 337–350.

Van Denhaute C, Eggermont K, Nuttin B, Debyser Z, and Baekelandt V (2003) Lentiviral vector-mediated delivery of short hairpin RNA results in persistent knockdown of gene expression in mouse brain. *Human Gene Therapy* 14: 1799–1807.

Wade-Martins R, Saeki Y, and Chiocca EA (2003) Infectious delivery of a 135-kb LDLR genomic locus leads to regulated complementation of low-density lipoprotein receptor deficiency in human cells. *Molecular Therapy* 7: 604–612.

Wade-Martins R, Smith ER, Tyminski E, Chiocca EA, and Saeki Y (2001) An infectious transfer and expression system for genomic DNA loci in human and mouse cells. *Nature Biotechnology* 19: 1067–1070.

Wahlfors JJ, Zullo SA, Loimas S, Nelson DM, and Morgan RA (2000) Evaluation of recombinant alphaviruses as vectors in gene therapy. *Gene Therapy* 7: 472–480.

Watson DJ, Kobinger GP, Passini MA, Wilson JM, and Wolfe JH (2002) Targeted transduction patterns in the mouse brain by lentivirus vectors pseudotyped with VSV, Ebola, Mokola, LCMV, or MuLV envelope proteins. *Molecular Therapy* 5: 528–537.

Weinberg JB, Matthews TJ, Cullen BR, and Malim MH (1991) Productive human immunodeficiency virus type 1 (HIV-1) infection of nonproliferating human monocytes. *Journal of Experimental Medicine* 174: 1477–1482.

Wickham TJ, Mathias P, Cheresh DA, and Nemerow GR (1993) Integrins alpha v beta 3 and alpha v beta 5 promote adenovirus internalization but not virus attachment. *Cell* 73: 309–319.

Wiznerowicz M and Trono D (2003) Conditional suppression of cellular genes: Lentivirus vector-mediated drug-inducible RNA interference. *Journal of Virology* 77: 8957–8961.

Xiong C, Levis R, Shen P, Schlesinger S, Rice CM, and Huang HV (1989) Sindbis virus: An efficient, broad host range vector for gene expression in animal cells. *Science* 243: 1188–1191.

Yang XW, Model P, and Heintz N (1997) Homologous recombination based modification in *Escherichia coli* and germline transmission in transgenic mice of a bacterial artificial chromosome. *Nature Biotechnology* 15: 859–865.

Yates JL, Warren N, and Sugden B (1985) Stable replication of plasmids derived from Epstein–Barr virus in various mammalian cells. *Nature* 313: 812–815.

Zavada J (1982) The pseudotypic paradox. *Journal of General Virology* 16: 15–24.

Zufferey R, Nagy D, Mandel RJ, Naldini L, and Trono D (1997) Multiply attenuated lentiviral vector achieves efficient gene delivery *in vivo*. *Nature Biotechnology* 15: 871–875.

Enteric Nervous System Development

M D Gershon, Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Discovery of the Independent Function of the Enteric Nervous System

The discovery of the enteric nervous system (ENS) dates to the period of the American Civil War, when the presence of ganglia within the external muscle layers of the bowel (Auerbach's or myenteric plexus) and within the submucosa (Meissner's and Schabadasch's or submucosal plexuses) was described. Before the end of the nineteenth century, Bayliss and Starling had established that the 'local neural mechanism' (the ENS) mediates propulsive motility, which they called "the law of the intestine," independent of input from the central nervous system (CNS). While World War I raged, Trendelenburg emphatically confirmed this conclusion by eliciting "the law of the intestine," now renamed the peristaltic reflex, *in vitro*. Despite the clarity of this evidence, and the absence of its refutation, many biologists continue to believe that the ENS is an array of parasympathetic relay ganglia used by vagal and sacral nerves to control the gut. That they should do so is surprising because in his original definition of the autonomic nervous system, JN Langley clearly differentiated between the enteric and parasympathetic nervous systems. The parasympathetic and sympathetic systems were defined, respectively, by their craniosacral and thoracicolumbar outflows from the CNS, while the ENS, which can operate independently of CNS control, is a separate division that contains many neurons that lack a CNS innervation. The misapprehension of the independence of the ENS tends to trivialize its importance, decrease ENS research, and focus explanations of functional gastrointestinal disorders exclusively on the CNS.

Organization of the ENS

The ENS is comprised of myenteric and submucosal plexuses, which interconnect and innervate effectors, smooth muscle, blood vessels, and glands (Figure 1). Extrinsic motor and sensory nerves of vagal and spinal origin link the ENS and CNS. The number of enteric neurons ($>10^8$ in the human small intestine), which approximates that of the spinal cord, dwarfs the number of motor axons ($\sim 10^3$) in the vagus nerves. This disparity critically influenced Langley's view of the enteric innervation but does not mean that the CNS is unable to influence the ENS. It can, and does, with effects that often become all too disturbingly apparent when anxiety strikes. In fact, investigators who selectively visualize the efferent vagal innervation have marveled at its apparent extent. Vagal axons, however, are only a small proportion of the vast numbers of fibers in the ENS. When all axons are demonstrated, fibers can be distinguished individually only by electron microscopy. Intrinsic reflexes, their integration within the ENS, and their coordination with myogenic activity (muscle and interstitial cells of Cajal (ICCs)), not the CNS, are what determine the basic motile and secretory behaviors of the gut. ICCs are electrically coupled to smooth muscle and act as pacemakers and intermediaries, which transmit signals from nerve to smooth muscle.

The sensory component of the extrinsic enteric innervation is very much larger than the motor; about 90% of vagal fibers are afferent. Pain and discomfort are conveyed to the CNS primarily by spinal nerves, while information that the vagi convey to the brain does not necessarily reach consciousness. Vagal sensory fibers are responsible for nausea and, when highly stimulated, also convey pain, but mostly they are involved in 'homeostasis,' the ramifications of which are not completely understood but may be beneficial. Vagal stimulation is utilized to treat depression and epilepsy, and it improves learning and memory in humans and animals. Vagal motor fibers have been postulated to act on enteric command neurons, allowing the CNS to exert a powerful overall effect, despite the intrinsic control of motility and secretion. The CNS thus operates in a role analogous to that of a chief operating officer, who directs overall operations but does not micromanage.

The anatomy, physiology, and neurochemistry of the ENS are complex. The histological organization of the ENS, for example, resembles that of the CNS and lacks the connective tissue sheaths that normally surround peripheral nerves. Because the ENS thus lacks internal collagen, neurons are supported by enteric glia, which are rich in glial fibrillary acidic protein and resemble CNS astrocytes. Enteric neurons, moreover, are phenotypically diverse. Every class of neurotransmitter and neuromodulator that has been described in the CNS is also represented in the ENS. Enteric interneurons, furthermore, are arranged in microcircuits, which are not yet understood.

ENS Control of Motile and Secretory Behaviors

The basic pattern of the peristaltic reflex has been ascertained, but motility is more than just peristaltic reflexes. The gut sometimes propels its content and sometimes mixes them. The ENS contains a library of

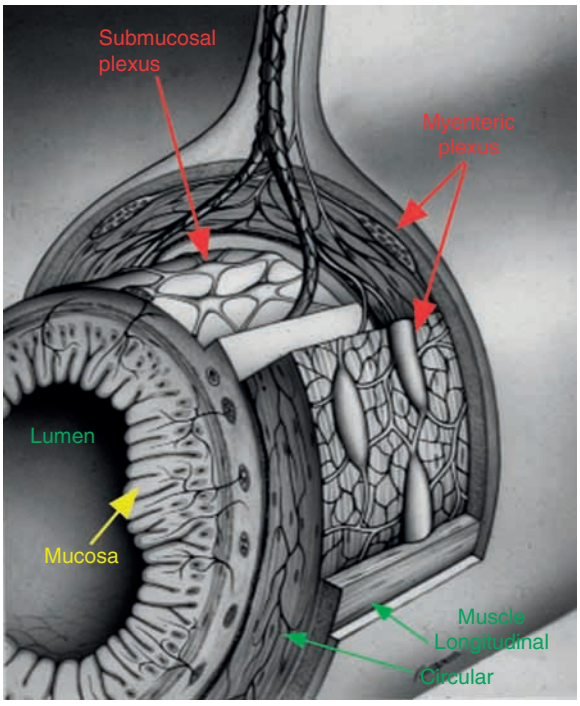

Figure 1 A drawing showing the organization of enteric innervation. Extrinsic nerves are depicted at the top of the figure reaching the bowel as perivascular (sympathetic) and paravascular (sensory, parasympathetic) bundles in the mesentery. The submucosal and myenteric plexuses are indicated. The myenteric plexus lies between the circular and longitudinal layers of the muscularis externa. The submucosal plexus can be seen in the cutaway view and in profile in the cross-sectional view of the gut. Nerve fibers enter the mucosa but are limited to the connective tissue of the lamina propria, and none enter the lumen.

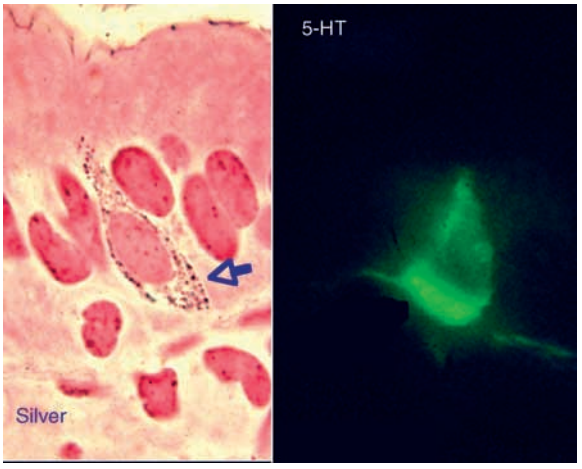

Figure 2 Enterochromaffin (EC) cells, which function as sensory transducers, can be demonstrated with silver (left) or immunocytochemically with antibodies to 5-HT (right). Silver stains depend on the ability of 5-HT to act as an agent that reduces silver salts to metallic silver, which precipitates to reveal the storage granules. Note that these granules (silver) and 5-HT are concentrated in the basolateral cytoplasm. This orientation reflects the basolateral secretion of 5-HT.

motor programs, from which it selects appropriate patterns. This library includes postprandial mixing to facilitate digestion and absorption, aborally directed power propulsion, orally directed power propulsion (associated with vomiting), and physiological ileus (in which motility stops). There is also an interdigestive, housecleaning program consisting of migrating motor complexes that sweep down the bowel. Choosing programs requires that the content of the lumen of the bowel be closely monitored.

No nerves actually enter the lumen (Figure 1); therefore, the ENS must be apprised of luminal conditions transepithelially. Enteroendocrine cells in the mucosal epithelium function as sensory transducers. The best known of these is the enterochromaffin (EC) cell, which contains serotonin, or 5-hydroxytryptamine (5-HT; Figure 2), which is secreted in response to increases in intraluminal pressure or, in the duodenum, applications of acid or glucose to their apical microvillous border. Also, 5-HT stimulates the mucosal processes of intrinsic primary afferent neurons (IPANs), subsets of which exist in the submucosal and myenteric plexuses. These cells do not project to the CNS but activate the intrinsic microcircuits that mediate peristaltic and secretory reflexes. IPANs thus allow the ENS to function independent of CNS input. Common motor neurons of the ENS relax or contract intestinal muscle, probably through the intermediation of ICCs. Excitatory motor neurons are cholinergic and also secrete a neurokinin; inhibitory motor neurons contain nitric oxide synthase but may also utilize adenosine 5′-triphosphate or vasoactive intestinal peptide (VIP) to exert relaxant effects. Additional enteric neurons are secretomotor and cause goblet cells to secrete mucous and crypt epithelial cells to secrete chloride and thus water, and they also regulate gastrointestinal peptide secretion by enteroendocrine cells. Secretomotor neurons are either cholinergic or VIP-secreting, without overlap. The IPANs of both plexuses are cholinergic and may also co-store peptides and a calcium binding protein, calbindin. In fact, about 70% of all enteric neurons are cholinergic; however, most of these neurons co-store additional peptides and/or small molecule neurotransmitters, such as 5-HT. Neurons of the ENS can be identified by a chemical code made up of the transmitter and other molecules that they co-express.

Serotonin Is Important in the Normal and Abnormal Behavior of the Bowel

Not only is 5-HT produced and secreted by EC cells to initiate reflexes, but it is also a neurotransmitter of long descending myenteric interneurons. Although

its CNS activity (seemingly related to everything that makes life worthwhile) is what made 5-HT famous, its concentration in the CNS pales beside that in the gut. The enteric epithelium, moreover, continually turns over. New cells are generated from stem cells in the necks of gastric glands and intestinal crypts, mature as they rise to the gastric surface and villus tips, and eventually die and slough into the lumen. The epithelial cells are thus constantly in motion. Because axons are not adapted to track moving targets, the axons of IPANs cannot make traditional synapses on EC cells but terminate as processes in the lamina propria of the mucosa. This type of paracrine junction is acceptable because the bowel is a gross and imprecise organ. Specificity of signaling is achieved, not by the close apposition of interacting elements, but by the ability of receptors to selectively respond to 5-HT. EC cells overcome the large and variable distance to the nerves they stimulate by secreting massive quantities of 5-HT. In fact, 95% of the body's 5-HT is located in the gut, and the EC cells represent a much larger store of it than do the neurons of the myenteric plexus. The overflow of 5-HT from EC cell secretion reaches the intestinal lumen and stocks platelets with 5-HT. Platelets do not synthesize 5-HT but take it up as they circulate through the bowel. The rate-limiting enzyme in 5-HT's biosynthesis is tryptophan hydroxylase (TpH). There are two TpH isoforms, TpH1 and TpH2. EC cells contain TpH1, and neurons of the ENS and the CNS contain TpH2.

The large quantities of constitutively secreted 5-HT, and its overflow, necessitate an efficient mechanism in the bowel for 5-HT inactivation. Inactivation is accomplished by 5-HT reuptake; this involves a plasmalemmal serotonin transporter (SERT), which is expressed in the gut by enterocytes and serotonergic neurons. Enteric SERT is identical to CNS SERT. Tricyclic antidepressants, serotonin selective reuptake inhibitors (SSRIs), and cocaine, which are in widespread use, and abuse, all inhibit SERT. These compounds potentiate 5-HT, but their enteric effects are limited by the presence of backup transporters, which include organic cation transporters (OCTs) and the dopamine transporter (DAT). The OCTs and DAT have a lower affinity than SERT does for 5-HT, but they have a high capacity, and they prevent catastrophe from occurring when SERT is knocked out. They do not, however, completely protect the bowel from SERT inhibition. Tricyclic antidepressants and SSRIs exert gastrointestinal 'side effects,' which are really direct effects. These include nausea, diarrhea, and perhaps as a result of the desensitization of 5-HT receptors, constipation. Diarrhea and constipation alternate in knockout mice, which lack SERT; these mice also lose water excessively in stool.

Major advances in the treatment of patients undergoing chemotherapy for cancer and in those suffering from irritable bowel syndrome (IBS) have highlighted the importance of enteric 5-HT. EC cells not only trigger intrinsic reflexes by secreting 5-HT and stimulating IPANs, but their 5-HT also activates extrinsic sensory nerves to initiate nausea and discomfort. Different subtypes of 5-HT receptor are utilized to activate, respectively, intrinsic and extrinsic nerve fibers. Submucosal IPANs, but are stimulated by 5-HT_{1P} receptors, which may be heteromultimers of other G-protein-coupled receptors. IPANs (as well as cholinergic motor neurons to smooth muscle) also express 5-HT_4 receptors at their terminals. The 5-HT_4 receptors enhance acetylcholine and calcitonin gene related peptide secretion and thus strengthen synaptic transmission in excitatory (prokinetic) pathways. Further, 5-HT activates extrinsic sensory nerves by stimulating 5-HT_3 receptors. Myenteric IPANs, which appear to drive the giant, sometimes painful contractions associated with the aboral power propulsion pattern of activity, are also stimulated by 5-HT_3 receptors.

The dichotomy of 5-HT receptor subtypes involved in intrinsic and extrinsic reflex stimulation makes therapy possible. Some 5-HT_3 antagonists, including ondansetron and granisetron, provide relief from the nausea that accompanies cancer chemotherapy without at the same time paralyzing the bowel by interfering with 5-HT-driven peristaltic and secretory reflexes, and some 5-HT_3 antagonists, particularly alosetron, also alleviate the discomfort associated with IBS. This is a major advance because the pain and discomfort of IBS are far more troubling to patients than the altered bowel habits that are another hallmark of that disease. Blockade of the giant contractions of power propulsion by 5-HT_3 antagonism may also contribute to the relief of pain and discomfort in IBS but may be counterproductive. Not only do 5-HT_3 antagonists block activation of the IPANs that initiate power propulsion; they also inhibit serotonergic fast excitatory neurotransmission within the ENS. Thus, 5-HT_3 antagonists are constipating and useful only to treat diarrhea-predominant IBS. It is possible that interference with the emergency aboral power propulsive pattern of activity contributes to the rare complication of ischemic colitis that has been linked to the use of 5-HT_3 antagonists in treating IBS.

In contrast to the slowing of motility associated with 5-HT_3 antagonists, the strengthening of neurotransmission in prokinetic pathways by 5-HT_4 agonists, such as tegaserod, is useful in the treatment of chronic constipation and constipation-predominant IBS. These compounds do not initiate reflexes and do not activate pain pathways but instead rely on

natural stimuli to get a refractory bowel moving again, and 5-HT_4 agonism strengthens propulsive motility. Therefore, 5-HT_4 agonists are effective when taken for their approved use.

The ENS Originates from Vagal, Truncal, and Sacral Regions of the Neural Crest

The complexity and unique nature of the ENS suggests that factors that drive its development will be different from those that govern development of the remainder of the autonomic nervous system. Meaningful research on ENS development dates from the demonstration by Yntema and Hammond that deletions of what they called the 'anterior neural crest' caused deficiencies to occur in enteric ganglia of chick embryos. Subsequently, Le Douarin deleted the crest from various regions of chick embryos, but she also replaced it with grafts of crest from quail embryos, creating interspecies chimeras. The grafted quail cells migrated seemingly normally in their new hosts. Quail cells express an identifying pattern of nucleolar-associated heterochromatin that enables them to be reproducibly identified in fields of chick cells (Figure 3). Quail cells can also be recognized immunocytochemically, as has been done in more recent studies, by using quail-selective antibodies. As a result, cells of neural crest origin can be identified in the target organs of chimeric embryos, which in combination with suitable grafts allows the sites of origin of crest-derived émigrés to be determined. The axial levels of origin of the crest-derived enteric neurons and glia have now definitively been mapped in avians through the study of interspecies chimeras. For some time, it was simply presumed that the origin of the mammalian ENS was more or less analogous to that of avians. More recently, vital dyes and genetic mapping techniques have been used to verify the correctness of that assumption. The ENS is derived from the postotic (vagal and rostral truncal axial levels equivalent to somites 1–7) and sacral (axial caudal to somite 28) regions of the crest (Figure 4).

Enteric Neurons and Glia Develop from a Multipotent Precursor Population

Along with knowledge of the sites of origin of crest-derived cells came the counterintuitive discovery that differences between enteric neurons in the neural crest origins of their predecessors do not explain the phenotypic diversity of the ENS. The developmental potential of various regions of the crest, at least with respect to the ENS, does not appear

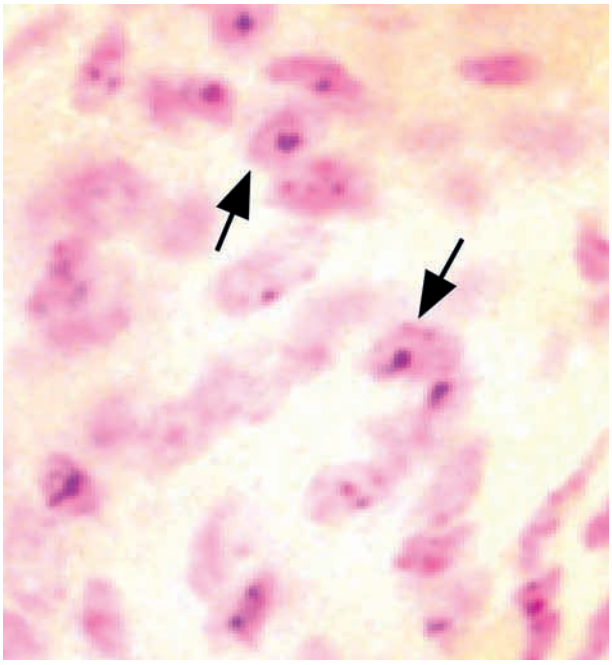

Figure 3 Quail cells (arrows) from a graft of neural crest are demonstrated in the bowel of a chick embryo to which they have migrated. The nucleolar-associated heterochromatin that characterizes the quail cell nucleus has been demonstrated by staining DNA with the Feulgen reaction. Quail cells and thus cells of donor neural crest origin can be distinguished from the chick cells of the host.

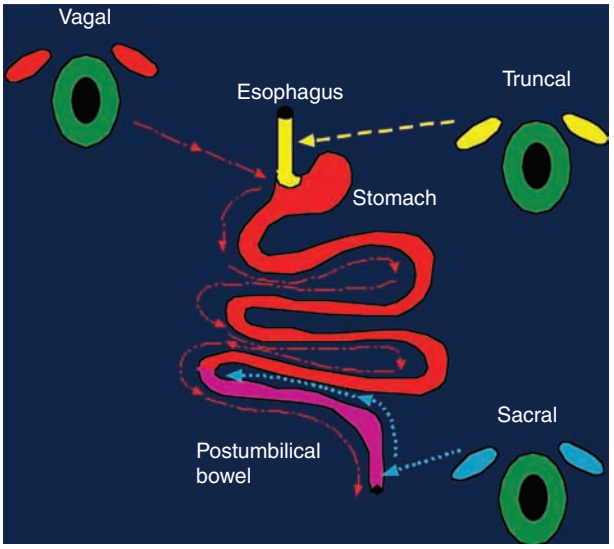

Figure 4 Cells that migrate from the vagal, rostral truncal, and sacral levels of the neural crest colonize the bowel. Vagal crest-derived cells (red) colonize the entire bowel, truncal crest-derived cells (yellow) colonize only the esophagus and immediately adjacent stomach, and sacral crest-derived cells (blue) colonize the postumbilical gut, which thus contains cells of both vagal and sacral origin (purple).

to be preset. For example, crest which, if left unperturbed, would form the adrenal medulla will instead migrate to the gut and give rise to an ENS when it is transplanted to the vagal region of a recipient embryo. This interchangeability of developmental potential of populations of crest cells has led to the concepts that there are defined pathways in embryos along which crest-derived cells migrate and that populations of crest cells are pluripotent; nevertheless, despite the pluripotentiality of the population, individual premigratory crest cells may not be pluripotent. Some crest-derived cells may already be specified before they migrate. For example, crest-derived cells that form sensory neurons migrate away from the neuraxis before those that give rise to melanocytes. The evident pluripotency of populations of crest-derived cells may be due in part, therefore, to their heterogeneity. Even if individual stem cells are not able to give rise to all the derivatives of the crest, a population may contain a sufficient number of determined precursors to allow all these derivatives to develop. According to this idea, if cells migrate to the skin, those precursors that are capable of forming melanocytes will develop, but if the same population finds itself in the bowel, those precursors that are capable of developing as enteric neurons and glia will now differentiate. Actually, cells committed to be melanocytes (which leave the neuraxis late) do not normally migrate to the bowel; they preferentially migrate along the dorsolateral pathway that leads them to the skin. The small population of melanogenic crest-derived cells that do migrate ventrally are filtered out of the population that colonizes the bowel as it passes through the caudal branchial arches so that the gut is nonpigmented. On the other hand, melanocytes develop in the gut when melanogenic precursor cells are experimentally planted there; the gut is also pigmented in chicks with the Silkie Fowl mutation, in which migratory controls fail. The crest-derived precursors of the ENS are thus not normally pluripotent, in the sense of being able to give rise to every type of cell that can be generated by the neural crest. Crest-derived cells that have completed their migration to the bowel, for example, cannot give rise to melanocytes or mesectoderm. Examinations of clones of enteric crest-derived cells, however, have established that they are multipotent, in the sense that they can give rise to glia or any of the many types of neuron that are found in the ENS. In any case, the unique properties of the ENS are determined, not by the site of origin of committed precursors in the neural crest, but by epigenetic factors, especially microenvironmental cues these cells encounter within the gut or while migrating to it.

Development of the ENS Is a Progression of Interacting Cell Autologous and Nonautologous Events

The developmental potential of enteric crest-derived cells decreases progressively as these cells age; moreover, crest-derived cells cannot respond to growth factors or molecules of the extracellular matrix unless they first express the relevant receptors. Cell autologous and nonautologous events thus combine and interact to define differentiation of the ENS. Developmental progress in the formation of enteric neurons and glia is marked by a series of transcription factors that must be expressed, receptors that must be displayed on plasma membranes, and growth factors that must be encountered. The multipotent precursors of the ENS, which are capable of giving rise to both neurons and glia, obligatorily express the transcription factors Pax3 and Sox10, which drive expression of Phox2b and the receptor tyrosine kinase, Ret. Knockout of Sox10 or Phox2b prevents the formation of any enteric neurons or glia, while knockout of Ret blocks the formation of enteric neurons below the level of the esophagus and immediately adjacent stomach. To activate Ret, ligands must first bind to a member of the glial cell-derived neurotrophic factor (GDNF) family receptor alpha (GFRα) molecules. Two Ret ligands function in ENS development to stimulate Ret, GDNF, which binds to GFRα1, and nurturin, which binds to GFRα2. Of these, GDNF is required first and by almost all neural precursors. Nurturin appears to be necessary for the formation of restricted later-arising populations of neurons found mainly in the submucosal plexus. GDNF promotes proliferation of crest-derived precursors, which are extensive in fetal bowel, and is also chemoattractive for crest-derived cells. GDNF is produced by non-crest-derived elements of the enteric mesenchyme and is thought to assist in preventing crest-derived precursors from migrating out to the gut during its colonization. Recent evidence suggests that the basic helix–loop–helix transcription factor, Hand2, is also essential in allowing enteric crest-derived cells to complete the series of steps that take them from precursors to neurons. The gut is colonized, neuronal differentiation commences in the absence of Hand2, but when Hand2 is deleted, the precursors express early neuronal markers but then freeze without terminally differentiating. Perhaps because they begin neuronal differentiation, progenitors do not shift by default toward the glial lineage, nor do the unfinished precursors die. Thus far, animals lacking Hand2 have not survived long enough to determine the fates of the neuronal precursor cells that cannot complete differentiation.

Hirschsprung's Disease Is the Most Obvious Developmental Disorder of the ENS

Defects in the genes required early in ENS differentiation are associated in humans with Hirschsprung's disease, or congenital megacolon. Hirschsprung's disease occurs when a variable length (long segment or short segment) of terminal bowel is aganglionic. The affected region of the gut is not denervated, but when the ganglia, which mediate the complex behaviors of the bowel, are lacking, the gut is as effectively obstructed as if it had been tied off with a suture. The gut dilates proximal to the aganglionic segment. Extrinsic nerves and descending nerve fibers from the proximal ganglionated portion of the gut form coarse trunks in the aganglionic region. *RET*, which is mutated in 3–35% of cases in which a gene has been identified, and the endothelin B receptor (*EDNRB*), which is mutated in 5–15% of cases, are the most common genes found to be mutated in association with Hirschsprung's disease. Others include genes encoding the ligands GDNF, GFRα1, and nurturin for RET and EDN3 (endothelin3) for EDNRB, as well as the transcription factors SOX10, PAX3, PHOX2B, and SIF1.

Hirschsprung's disease is the most visible and highly investigated enteric neuromuscular birth defect. Hirschsprung's disease is not uncommon, occurring in about one per 5000 births, except among Mennonites, where the incidence is one per 500 births because of inbreeding. Because both the development of glia and all types of enteric neurons fail, albeit segmentally, it is not surprising that expression of each of the genes linked to Hirschsprung's disease is required early in ontogeny. If left untreated, Hirschsprung's disease is lethal, but surgical removal of the aganglionic bowel prevents death. Surgery is thus life-saving, but it may not leave patients with a fully functional, inconvenience-free gut. Defects in motility leading to constipation and/or soiling may remain, even after the aganglionic tissue has flawlessly been removed, suggesting that the remaining ostensibly 'normal' ENS is really abnormal. The presence of ganglia thus does not guarantee that the ENS is normal.

Subtle ENS Defects May Arise as a Result of Mutations in Genes Required Late in ENS Development

Because enteric behavior is complex, the presence or absence of enteric neurons, which has historically been used by pathologists to assess ENS function, is inadequate. Intestinal neuronal dysplasia, for example, is a Hirschsprung's-like condition, associated with abnormal ganglia rather than aganglionosis, in affected regions. Intestinal neuronal dysplasia has been somewhat controversial, and some have doubted its existence, but it has been documented by competent pathologists and is mimicked in mice lacking Hox 11L1 and in rats that with a heterozygous endothelin B deficiency, suggesting that intestinal neuronal dysplasia is a real condition. Chronic idiopathic intestinal pseudo-obstruction (CIIP) severe enough to require bowel transplantation can even occur despite the presence of ganglia in the nonfunctioning gut. The occurrence of functional abnormalities of gastrointestinal motility, despite the presence of enteric ganglia, supports a hypothesis that there is a spectrum of functional disorders, from CIIP to IBS, that are the result of defects in genes that are not required early in the development of the ENS but which are needed to complete the development of particular subsets of neurons or to regulate synaptogenesis. Gene products required late in ENS ontogeny include the neurotrophin NT3 and its high-affinity receptor, TrkC. During the period when GDNF is required to drive proliferation of enteric crest-derived cells, NT3 does nothing to them. Crest-derived cells do not acquire TrkC or become NT3-responsive until after these cells have colonized the bowel. NT3, furthermore, is not needed for the development of all enteric neurons but only for subsets. One such NT3-dependent subset is the submucosal IPAN, which initiates peristaltic and secretory reflexes. After the deletion of NT3 or TrkC, therefore, the ENS appears to be normal and ganglia are present; however, the bowel will not function normally because it lacks a critical subset of its neurons.

Late events are important in ENS development. Early crest-derived precursors have been visualized as they migrate through the murine bowel. These cells, of vagal origin, migrate proximodistally in long chains of seemingly attached cells in the outer gut mesenchyme. The driving force for migration appears to be the proliferation of precursors at the leading edge of the colonizing population; however, the evident contact of leaders with stragglers raises the possibility that population pressure from stragglers contributes to the forward movement of émigrés. The location of the migrating cells in the outer gut mesenchyme is appropriate for the formation of the myenteric plexus. Sacral crest-derived cells delay their arrival in the postumbilical bowel until the arrival of their vagal counterparts but then ascend in the bowel in a countercurrent fashion. The submucosal plexus forms after the myenteric. The formation of the submucosal plexus results from a secondary perpendicular migration of crest-derived cells from the outer gut mesenchyme toward the mucosa. This secondary migration occurs because

netrins, which are secreted by the endoderm, attract crest-derived cells. Netrins act on the netrin receptor called 'deleted in colorectal cancer' (DCC), which is expressed by migrating crest-derived cells. Netrins are also secreted by developing pancreatic acinar cells and attract DCC-expressing crest-derived cells from the bowel, which migrate out of the gut and into the developing pancreatic buds. These crest-derived cells ultimately give rise to intrapancreatic ganglia. It is interesting that the enteropancreatic migration of crest-derived cells is followed by the extension of enteropancreatic nerve fibers from neurons in gastric and duodenal myenteric ganglia. These nerves ultimately connect the bowel to the pancreas, where they terminate in ganglia.

Gastrointestinal dysmotility disorders are neither well understood nor rare. These conditions are believed to be congenital when they occur in children, but that belief has not clarified their pathogenesis or provided a means of treatment. Although similar gastrointestinal dysmotilities also occur in adults, the possibility that adult disorders might, like their pediatric counterparts, be congenital has not been tested. Much of the research on ENS development has focused on the pathogenesis of Hirschsprung's disease. This focus is understandable because the aganglionosis of Hirschsprung's disease is so obvious; moreover, the current treatment of Hirschsprung's disease, while life-saving, leaves much to be desired. Still, the paradigm that the presence of ganglia signifies a normally functioning ENS has proven to be false. It is thus possible that future research on the development of the ENS will be not only intellectually satisfying but fruitful in pointing the way toward understanding the pathogenesis of gastrointestinal conditions that are now mysterious, unpreventable, and poorly treated.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Enteric Nervous System: Sensory Pathways; Enteric Nervous System: Neural Circuits and Chemical Coding; Enteric Nervous System: Neurotrophic Factors; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Enteric Nervous System: Physiology; Neuropeptides: Enteric Nervous System; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation.

Further Reading

Berthoud HR, Blackshaw LA, Brookes SJ, and Grundy D (2004) Neuroanatomy of extrinsic afferents supplying the gastrointestinal tract. *Neurogastroenterology and Motility* 16(supplement 1): 28–33.

Bondurand N, Natarajan D, Barlow A, Thapar N, and Pachnis V (2006) Maintenance of mammalian enteric nervous system progenitors by SOX10 and endothelin 3 signalling. *Development* 133: 2075–2086.

Furness JB (2005) *The Enteric Nervous System*. Malden, MA: Blackwell.

Gershon MD (2005) Nerves, reflexes, and the enteric nervous system: Pathogenesis of the irritable bowel syndrome. *Journal of Clinical Gastroenterology* 39: S184–S193.

Gershon MD and Ratcliffe EM (2004) Developmental biology of the enteric nervous system: Pathogenesis of Hirschsprung's disease and other congenital dysmotilities. *Seminars in Pediatric Surgery* 13: 224–235.

Gershon MD and Tack J (2007) The serotonin signaling system: From basic understanding to drug development for functional GI disorders. *Gastroenterology* 132: 397–414.

Gianino S, Grider JR, Cresswell J, Enomoto H, and Heuckeroth RO (2003) GDNF availability determines enteric neuron number by controlling precursor proliferation. *Development* 130: 2187–2198.

Grundy D, Schemann M, and Wood J (2000) A tale of two brains, one little and one big. *Neurogastroenterology and Motility* 12: 105–111.

Kapur RP, Clarke CM, Doggett B, et al. (2005) Hox11L1 expression by precursors of enteric smooth muscle: An alternative explanation for megacecum in HOX11L1−/− mice. *Pediatric and Developmental Pathology* 8: 148–161.

Liu M, Geddis MS, Wen Y, Setlik W, and Gershon MD (2005) Expression and function of 5-HT4 receptors in the mouse enteric nervous system. *American Journal of Physiology. Gastrointestinal and Liver Physiology* 289: G1148–G1163.

Meier-Ruge WA, Bruder E, and Kapur RP (2006) Intestinal neuronal dysplasia type B: One giant ganglion is not good enough. *Pediatric and Developmental Pathology* 9: 444–452.

Newgreen D and Young HM (2002) Enteric nervous system: Development and developmental disturbances – Part 1. *Pediatric and Developmental Pathology* 5: 224–247.

Newgreen D and Young HM (2002) Enteric nervous system: Development and developmental disturbances – Part 2. *Pediatric and Developmental Pathology* 5: 329–349.

Powley TL and Phillips RJ (2002) Musings on the wanderer: What's new in our understanding of vago-vagal reflexes? I. Morphology and topography of vagal afferents innervating the GI tract. *American Journal of Physiology. Gastrointestinal and Liver Physiology* 283: G1217–G1225.

Young HM, Bergner AJ, Anderson RB, et al. (2004) Dynamics of neural crest-derived cell migration in the embryonic mouse gut. *Developmental Biology* 270: 455–473.

Enteric Nervous System: Disorders

D Grundy, The University of Sheffield, Sheffield, UK
R De Giorgio, University of Bologna, Bologna, Italy
M Schemann, Technische Universität München, Freising-Weihenstephan, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The human enteric nervous system (ENS) contains an estimated 10^8 neurons in two major ganglionated plexuses that extend the entire length of the bowel: the myenteric plexus between the longitudinal and circular muscle and the submucosal plexuses associated with the mucosal epithelium. More than 16 phenotypically distinct neuronal populations have been identified and classified by their morphology, transmitter content, and electrophysiological properties into sensory neurons, ascending and descending interneurons, and excitatory and inhibitory motor neurons. These are organized into reflex circuits that control and coordinate motility, blood flow, and secretion to meet the digestion needs of the individual. While enteric reflexes operate independently of any connections to the central nervous system (CNS), extrinsic afferent and efferent fibers running in the parasympathetic and sympathetic connections to the brain stem and spinal cord are responsible for reflexes that help coordinate distant regions of the gastrointestinal (GI) tract, link digestive function to behavioral mechanism that control food intake, and mediate sensations such as pain. Disorders of the ENS are manifested as dysmotility and altered secretory function, which in turn lead to symptoms of dysphagia, diarrhea or constipation, nausea, discomfort, and pain. Disorders arise under a variety of circumstances, from congenital abnormalities in ENS development such as Hirschsprung disease (HSCR) to functional defects in reflexes brought about by inflammation or stress. Some are specific to the ENS, others are enteric manifestation of more widespread systemic disease. Understanding the ENS in health and disease has implications for advances in treatment options and the development of novel pharmaceuticals to treat GI symptoms.

Developmental Disorders

Developmental disorders of the ENS clinically present with signs of intestinal obstruction, usually in neonatal life or early childhood. In about 75% of patients the abnormal segment is restricted to the rectosigmoid of colon. In a small minority of cases (about 5%) the abnormality may involve the whole colon (total colonic aganglionosis) or may be even more widespread, affecting the entire length of the GI tract. Abnormalities extending into the esophagus are associated with difficulty in swallowing. Developmental defects may be familial, with a distinct pattern of inheritance, or sporadic. There is a familial tendency in about 7% of short-segment disease, increasing to about 21% in patients when the whole of the colon is affected. The most common developmental disorder was named after Hirschsprung, who first described the condition in the late nineteenth century. It occurs in about 1 in 4500 live births (3.8:1 males to female), and is characterized by a segment of bowel that is completely devoid of enteric ganglia. The aganglionic segment is constricted and unable to pass contents, leading to the normal bowel above becoming massively dilated. In many cases of HSCR the defect has been localized to chromosome 10q11.2, which specifies the *RET* gene that encodes a tyrosine kinase receptor (see later). Other genes have been linked to HSCR, including *SOX10*, *GFRα1*, and the gene for endothelin receptor type B (*EDNRB*), but, even so, for 50% of Hirschsprung cases the genetic cause has not been identified.

The neurons and glia that make up the ENS are derived from precursor cells in the vagal and sacral neural crest. Vagal crest-derived precursors are multipotent cells that migrate rostrocaudally through the gut before exiting the cell cycle and differentiating into the neuronal types that make up the unique features of the ENS. Sacral neural crest cells enter the gut after it has been colonized by vagal neural crest cells, but the molecular control of sacral neural crest cell development has not been extensively investigated. Migration, proliferation, and differentiation are all cued by the molecules that the crest-derived cells encounter on their journey along the bowel, acting on developmentally regulated receptors that these cells express. Some of the molecules and receptors that determine this process have already been well established. The *RET* knockout (RET −/−) shows a total absence of the ENS, suggesting that normal enteric development depends upon the functional integrity of this tyrosine kinase receptor signaling pathway. The ligand for *RET* receptor is glial-derived neurotrophic factor (GDNF), and knocking out the gene for this factor leads to a similar lack of expression of enteric neurons. GDNF acts as a chemoattractant for vagal neural crest cells (NCCs) expressing the RET/GFRα1 receptor complex. While RET and GFRα1 −/− mice have severe aganglionic phenotype, the heterozygotes have only mild reductions in neuronal counts. Nevertheless, these

animals have striking deficits in intestinal motility and transmitter release, demonstrating the importance of *RET* signaling for both ENS structure and function. Endothelin-3 (EDN3) and its G-protein-coupled EDNRB are another important component of the signaling process, particularly in the development of the distal colon. In the absence of EDNRB or EDN3 the spread of neural precursors from the small to the large bowel is severely impaired.

Neural crest-derived cells differentiate into glial cells and the various subsets of neurons that can be identified by morphology, projection pattern, neurotransmitter content, ion channels, and receptor expression. Intrinsic factors such as transcription factors and extrinsic cues, including molecules in the gut environment, combine to influence the neural fate and can be detected in various combinations during the course of development. Environmental cues appear to change over time while intrinsic cues vary with developmental age, such that enteric neural progenitors are only competent to differentiate at a particular time point in development. The transcription factor SOX10 is required for differentiation of enteric glial cells. SOX10 and EDNRB appear to interact to determine the penetrance and severity of aganglionosis, which is a feature of HSCR where there is phenotypic variability within and between family members. A potential site of action of SOX10 is on an enhancer element of the *EDNRB* locus that is spatiotemporally activated at the time that ENS precursors are populating the distal gut.

In addition to neurons acquiring the correct phenotype, they must assume their correct location within the myenteric and submucosal plexuses. Crest cells migrate predominantly in the outer gut mesenchyme. Recent evidence suggests that cells that form the submucosal ganglia turn away from their proximodistal migration in response to netrins secreted from the GI mucosa. Crest-derived cells acquire netrin receptors, including DCC ('deleted in colorectal cancer') at the time they make their turn. This migration can be abolished by DCC antibodies or in transgenic mice that lack DCC, and these animals lack a submucosal plexus. Neurotrophin-3 is required for development and survival of enteric neurons that express tyrosine kinase C receptor (TrkC), which includes a subset of submucosal neurons.

Sonic hedgehog (Shh) plays a pivotal role in the spatial development of the gut. Recent evidence indicates that Shh may regulate proliferation and differentiation of NCCs in the gut and modulate the responsiveness of NCCs to GDNF. Bone morphogenetic proteins (BMPs) influence many developmental processes under the influence of Shh. Transcripts of BMP-2 and BMP-4 can be detected in the bowel at the time of neuronal differentiation, and coinciding with the expression of their receptors BMPR-1a and BMPR-1b and their antagonists noggin, gremlin, chordin, and follistatin in both crest- and non-crest derived cells. Animals overexpressing noggin under neuron-specific enolase (NSE) control, as a means of avoiding earlier effects on differentiation and migration, had greatly increased numbers of both myenteric and submucosal neurons throughout the gut, suggesting that BMP signaling limits the size of the ENS. BMP also induced the precocious appearance of TrkC-expressing lineages in cultures of crest-derived cells, an observation that correlates with the decrease in TrkC-expressing neurons in the noggin-overexpressing animals. Thus, BMP signaling regulates both neuronal numbers and cell fate differentiation.

Clearly there are a multitude of factors that regulate proliferation, migration, and differentiation that are needed to enable NCCs to populate and subsequently innervate the entire bowel. Hirschsprung disease is typified by aganglionosis. However, it is clear that developmental abnormalities can lead to regions of bowel which have too few (hypoganglionosis) or too many (hyperganglionosis) enteric neurons. In addition, enteric neuronal survival and cell death may be influenced by extrinsic factors and, as such, may contribute to adult disease and alterations in GI function associated with age. One such factor may be as a consequence of inflammatory insult.

Inflammatory Neuropathies

Pathology and Major Clinical Features

The ENS *per se* may be a target of inflammatory/autoimmune reactions altering its structure and function. These forms of enteric neuropathies are referred to as 'enteric ganglionitis' or 'plexitis,' because they are characterized by a dense and composite inflammatory infiltrate of lymphocytes, plasma cells, monocytes, and other elements located within myenteric and/or submucosal ganglia and/or throughout the axonal processes supplying the alimentary tract. Myenteric ganglia, rather than submucosal ganglia, appear more commonly involved, raising the hypothesis that myenteric neurons may express certain constitutive antigens preferentially targeted by the immune system. Furthermore, the inflammatory/immune infiltrate within enteric ganglia is commonly associated with neuronal degeneration, up to complete loss of neurons in the most severe cases (i.e., acquired aganglionosis). As a result, the inflammatory/autoimmune injury of the enteric ganglionated plexuses and nerves causes gut dysfunction, including dysmotility and delayed transit, at any level of the GI tract. Depending on the segment

Table 1 Secondary forms of enteric ganglionitis: possible diseases responsible for enteric ganglia inflammatory changes

Disease	*Explanation*
Paraneoplastic	Small cell lung carcinoma, lung carcinoid, neuroblastoma, thymoma, gynecological tumors
Infectious	Chagas' disease, herpes virus infection
Autoimmune disorders of the CNS	Encephalomyeloneuropathy, multiple sclerosis
Connective tissue disorders	Scleroderma
Inflammatory bowel disorders	Ulcerative colitis and Crohn's disease

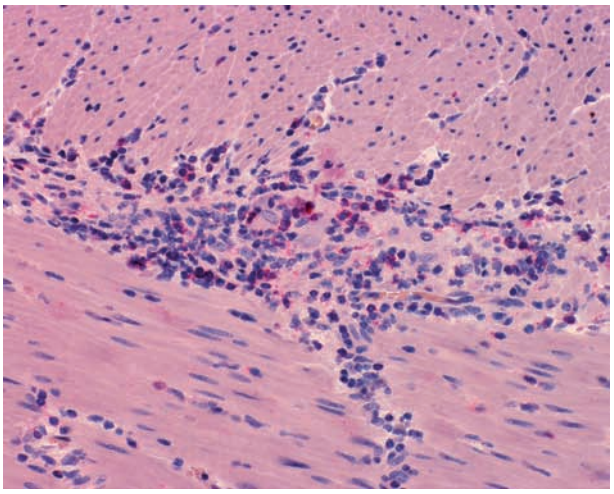

Figure 1 Inflammatory neuropathy characterized by dense infiltrate within the myenteric plexus (ganglionitis) in the colon of a patient with severe gut dysmotility. Lymphocytes are stained (pink) with a mouse monoclonal antibody targeting CD8 receptors. Note the close proximity between CD8 lymphocytes and enteric neurons. Blue staining identifies other components of the immune infiltrate. Anti-alkaline phosphatase immunohistochemistry technique. Original magnification is 100×.

involved in the inflammatory neuropathy, clinical manifestations include achalasia, gastroparesis, intestinal pseudoobstruction, and colonic inertia/megacolon.

Enteric ganglionitis can be classified into primary and secondary forms that show similar histopathological findings (e.g., inflammatory infiltrate within enteric ganglia) and clinical manifestations (e.g., severe dysmotility syndromes). The former are also termed idiopathic, as the search for underlying disorders responsible for inflammatory changes of the ENS is negative. The latter, in contrast, can occur in the context of a variety of diseases (Table 1). Enteric ganglionitis is a rare condition usually associated with severe gut disorders. However, recent findings raise the possibility that ENS inflammation could underlie a wide spectrum of gut sensorimotor abnormalities, including functional disorders such as irritable bowel syndrome (IBS). It is likely that the severity of immune response within myenteric ganglia is dependent upon genetic background influencing the immune system. Two major histopathological forms of enteric ganglionitis have been described – namely, lymphocytic type and eosinophilic type.

Lymphocytic ganglionitis This is characterized by CD3-positive T lymphocytes surrounding ganglion cell bodies mainly confined within the myenteric plexus. A vast majority of these T cells belong to T-helper (CD4) and T-cytotoxic/suppressor (CD8) subtypes (Figure 1). The T-cytotoxic activity likely targets constitutively proteins or *de novo*-expressed antigens on the surface of enteric neurons. This has been described in patients with achalasia and intestinal pseudoobstruction. Other components of the infiltration include CD79α-expressing cells, indicating the importance of mature B lymphocytes and circulating antineuronal antibodies in patients with enteric ganglionitis. The expression of *de novo* antigens and chemokines contributes to the loss of self-versus-nonself function, which in turn triggers an immune/inflammatory response within enteric ganglia (see later).

Further, regarding the activation of immunocytes, patients with inflammatory neuropathies develop a significant humoral response with a wide array of circulating antineuronal antibodies. These autoantibodies, which may be identified in cases of idiopathic and secondary (e.g., paraneoplastic syndrome) lymphocytic ganglionitis and dysautonomia, target different antigens expressed by enteric neurons, in particular the RNA-binding proteins of the Hu family. Yo protein (an antibody targeting a Purkinje cell cytoplasmic antigen), P/Q- and N-type Ca^{2+} channels, and ganglionic nicotinic acetylcholine receptors are also targets.

Eosinophilic ganglionitis This is characterized by a prominent infiltration of eosinophils in enteric ganglia. In contrast to lymphocytic ganglionitis, the eosinophilic forms do not appear to be associated with neuronal cell damage, indicating that eosinophils evoke neuronal dysfunction rather than degeneration. Animal models of gut inflammation confirm the general features of eosinophilic ganglionitis and provide a basis to better understand how eosinophils (and neutrophils) alter enteric neuron function. Mice infected with *Schistosoma mansoni*, a nematode known to evoke mucosal ileitis, granulomas, and diffuse inflammation of the neuromuscular layer, develop eosinophilic and neutrophilic ganglionitis of the myenteric plexus without significant neurodegeneration.

Putative Pathogenetic Mechanisms Leading to Enteric Ganglionitis

Several mechanisms play a role in the pathogenesis of enteric ganglionitis.

De novo Expression of Antigens in the ENS

A number of neurotropic viruses, such as herpes simplex type 1, Epstein–Barr virus, cytomegalovirus, and varicella-zoster, have been associated with gut dysmotility through an intraganglionic immune response secondary to the infection of enteric neurons. For example, varicella-zoster DNA was found within enteric ganglia in a proportion of achalasia patients and in animals infected with the same agent, suggesting a role in inflammatory neuropathies. This viral DNA leads to the expression of viral-related antigens on neurons and glial cells, thereby triggering immune activation. The ENS appears to be a target for scrapie-related prion protein (PrP^{sc}) and is involved in its transmission to the CNS.

Disruption of ENS Microenvironment

The ENS is protected by a blood–ganglion barrier. Glial cells are key players in the protection against noxae. Indeed, enteric glia exert a critical role in ENS maintenance and survival as well as in the cross-talk between ENS and the immune system. *In vitro*, enteric glial cells are stimulated by proinflammatory cytokines and in response release cytokines including interleukin (IL)-6. Mice deficient of enteric glia develop widespread enteritis associated with a progressive decrease in the neuronal population, which emphasizes a critical role for glia in enteric neuronal survival.

Molecular Mimicry

Classic paradigms, which apply to these mechanisms, are parasitic infection induced by *Trypanosoma cruzi*, which is responsible for Chagas' disease, and paraneoplastic syndromes. The GI tract is affected in cases of paraneoplastic syndrome related to neoplasms, including small cell lung carcinoma, thymoma, and gynecological and breast tumors. It is likely that onconeural antigens represent the trigger leading to both antineuronal antibody generation and enteric ganglionitis. Typical target molecules of paraneoplastic syndromes with GI involvement are the Hu and Yo (or cdr2) proteins, along with transmitter G-protein-coupled receptors and channels (see earlier).

Trypanosoma cruzi, a parasite transmitted to humans by Triatominae insects (the Reduviidae beetles), is responsible for Chagas' disease, which is an endemic disorder in South America, especially Brazil. In addition to cardiac and urinary system involvement, patients with Chagas' disease often demonstrate severe motor dysfunction of the GI tract – for example, achalasia (megaesophagus), megaduodenum and megacolon, and extrahepatic biliary tract. The underlying enteric ganglionitis is likely the result of a molecular mimicry leading to immune cross-reactivity between the parasite and enteric neurons. Particularly, the *T. cruzi* flagellar antigen Fl-160 (a 160 kDa surface protein) antigenically mimics a 48 kDa protein expressed by mammalian axons and myenteric neurons.

Postoperative Ileus

Postoperative ileus (POI) is a condition of impaired GI motility that is a frequent consequence of abdominal surgery. Contributory factors include anesthetic agent and postsurgery medication for managing pain. However, an inflammatory mechanism interacting with elements of the ENS and its extrinsic innervation is now considered to be relevant, particularly for the prolonged phase of POI. Resident macrophages play a pivotal role in the initiation of this enteric inflammatory response. Surgical manipulation triggers the release from macrophages of various mediators, including nitric oxide (NO) and prostaglandins, that can directly influence neuromuscular function, leading to inhibition of the entire GI tract. Macrophages also release proinflammatory cytokines, including tumor necrosis factor-α (TNF-α), IL-1β, IL-6, and monocyte chemotactic protein 1, all of which mediate cellular inflammation by recruiting circulating monocytes and neutrophils to the gut wall. These result in the further generation of neurogenic modulators, including NO and prostanoids. Cyclooxygenase-2 is considered to be crucial for activation of inhibitory motor reflexes and postoperative gut dysfunction. Mast cells may also contribute to POI and, like macrophages, can be triggered to release proinflammatory mediators and neuronal modulators, including histamine and proteases. POI may also involve hyperactivity of the sympathetic nervous system, which consequently results in strong inhibition of ENS circuits due to α_2-mediated presynaptic suppression of excitatory transmission as well as to direct postsynaptic inhibition of ENS activity.

Functional Disorders

The term 'functional bowel disease' was originally coined to indicate gut disorders without any detectible structural and biochemical abnormalities. Symptoms from the upper GI tract are referred to as functional dyspepsia (FD) and those in the lower GI tract as IBS. The hallmark of IBS is viscera hypersensitivity together with altered bowel habit (constipation, diarrhea, or alternating). FD is characterized by epigastric pain, nausea, and gastric motility and emptying disorder. Impairment of fundus relaxation associated with

early satiety is another hallmark of FD. Our current understanding is that the ENS, the brain–gut axis, and ENS–immune signaling are implicated in the pathophysiology of FD and IBS. Features include changes in serotonin signaling, altered gut microflora, genetic abnormalities, plasticity in receptor expression, and neuroinflammatory interactions.

Relevance of Immune Nerve Signaling in Functional Bowel Disease

Several studies have suggested a low-grade colonic inflammation as a pathophysiological factor in IBS. Patients with IBS exhibit an enhanced density of intraepithelial lymphocytes and an increased frequency of integrin $\beta 7^+$ T lymphocytes in the peripheral blood. Low-grade infiltration of lymphocytes in the myenteric plexus and neuronal degeneration have been described in patients with severe IBS. Pro- and anti-inflammatory mediators have powerful actions on the ENS.

Mast cell density and activity are significantly increased in biopsies taken from the large bowel of IBS patients. Increased numbers of mast cells in gastric biopsies have also been observed in patients with FD. An elevated mast cell number in IBS patients is associated with symptom severity and increased release of mast cell mediators, including histamine. Analysis of endoscopic biopsies from patients with food allergy and IBS has revealed significantly elevated H1- and H2-receptor mRNA levels. Both receptors are involved in histamine evoked enhanced excitability of enteric neurons, increases in secretion and motility, as well as altered sensory signaling in vagal and spinal afferents. The upregulated immune nerve signaling therefore contributes to symptom generation in IBS patients.

Gut infections may lead to functional bowel disease long after the acute infection has resolved. Some patients develop IBS symptoms after gastroenteritis, but it is not yet known what triggers plasticity in the ENS or chronic alterations in the immune nerve signaling. However, the outcome of such plasticity is observed as increased numbers of enterochromaffin (EC) cells and intraepithelial lymphocytes, altered cytokine profile, and increased gut permeability. The strong link between infectious gut diseases and the later occurrence of IBS and/or FD symptoms has led to a new patient definition, 'postinfectious' IBS (PI-IBS) or FD (PI-FD).

Serotonin and Functional Bowel Disease

Serotonin (5-hydroxytryptamine; 5-HT) is a key signaling molecule released from EC cells and ENS interneurons. It plays a critical role in reflex responses to chemical and mechanical stimuli. 5-HT acts through various receptor subtypes on enteric nerves, muscle, and epithelial cells. The ligand-gated 5-HT_3 receptor and metabotropic 5-HT receptors evoke excitatory responses in the ENS. The 5-HT_4 receptor is localized presynaptically on enteric nerves to facilitate fast synaptic transmission, but is also present on smooth muscle cells, where it seems to have a relaxing effect. In general, 5-HT_4 agonists are clinically used as prokinetic drugs whereas 5-HT_3 antagonists act to reduce neural activity. Thus, 5-HT_3 receptor antagonists improve symptoms in patients with diarrhea-predominant irritable bowel syndrome (D-IBS), whereas 5-HT_4 agonists are beneficial in patients with constipation-predominant IBS (C-IBS). Individuals with PI-IBS are a clinically distinct subgroup characterized by diarrheal symptoms and increased serotonin-containing EC cells, compared to those with non-PI-IBS.

Relaxation of the fundus is important for gastric accommodation during meal ingestion. There is evidence that PI-FD patients have impaired fundus accommodation attributable to dysfunction of enteric inhibitory neurons. Reflex relaxation is also important for coordinated gastric emptying. Activation of serotonin receptors by sumatriptan leads to release of NO from inhibitory motor neurons and as a consequence improves fundic accommodation in FD patients.

Microflora, ENS, and Functional Bowel Disease

There is compelling evidence for the importance of the gut microflora for the maintenance of normal gut functions. Some of theses effects may involve direct or indirect interactions with the ENS. Changes in the gut microflora in IBS patients have been observed and alterations of gut microflora associated with probiotic treatment improve some IBS symptoms. What we know so far is that probiotic treatment is associated with alterations of ENS phenotype, specifically reduced expression of the calcium-binding protein, calbindin, and the neuromediator, substance P (SP). In contrast, antibiotic treatment has been reported to increase SP expression in the ENS. How altered expressions of calbindin and SP are induced remains unknown, but interactions of the microflora with the enteric immune system are likely to be important, possibly by an action of cytokines on neurochemical coding. In this respect, the probiotic bacterium *Bifidobacterium infantis* 35624 alleviates symptoms in IBS patients by normalizing the ratio of anti- and proinflammatory cytokines.

CNS and ENS Interactions

Brain–gut interactions are important in IBS, especially in the context of stress and its effect on the hypothalamic–pituitary–adrenal axis and autonomic outflow, the latter being influenced by hypothalamic mediators such as corticotropin-releasing factor

(CRF). CRF and its receptors are also present within the ENS, and peripheral mechanisms play an important role in stress responses. Basal CRF levels are lower in IBS patients and increase significantly during stress. CRF1 receptor activation leads to acetylcholine release from excitatory motor neurons, and this contributes to alterations in gut function associated with stress.

Conclusion

Enteric disorders range from the rare and relatively rare life-threatening inflammatory and congenital neuropathies to common functional disorders including IBS and FD. Similarly, treatment options vary, from extreme measures of surgical intervention to pharmaceutical approaches that accelerate or delay transit to treat symptoms of constipation or diarrhea. However, with recent advances in understanding comes the prospect for novel treatments – for example, neural stem cells transplants to repopulate the aganglionic bowel in HSCR, or new generations of drugs that normalize the altered enteric reflexes that cause GI motor and secretory dysfunction and visceral hypersensitivity.

See also: Autoimmune Autonomic Neuropathy; Enteric Nervous System: Sensory Pathways; Enteric Nervous System: Neural Circuits and Chemical Coding; Enteric Nervous System Development; Neuropathy: Peripheral.

Further Reading

Barbara G, Stanghellini V, De Giorgio R, et al. (2006) Functional gastrointestinal disorders and mast cells: Implications for therapy. *Neurogastroenterology and Motility* 8: 6–17.

Bauer AJ and Boeckxstaens GE (2004) Mechanisms of postoperative ileus. *Neurogastroenterology and Motility* 16: 54–60.

De Giorgio R, Guerrini S, Barbara G, et al. (2004) Inflammatory neuropathies of the enteric nervous system. *Gastroenterology* 126: 1872–1883.

De Giorgio R and Camilleri M (2004) Human enteric neuropathies: Morphology and molecular pathology. *Neurogastroenterology and Motility* 16: 515–531.

Gershon MD (2004) Review article: Serotonin receptors and transporters – roles in normal and abnormal gastrointestinal motility. *Alimentary Pharmacology & Therapeutics* 7: 3–14.

Iwashita T, Kruger GM, Pardal R, et al. (2003) Hirschsprung disease is linked to defects in neural crest stem cell function. *Science* 301: 972–976.

Kapur RP, Gershon MD, Milla PJ, et al. (2004) The influence of Hox genes and three intercellular signalling pathways on enteric neuromuscular development. *Neurogastroenterology and Motility* 16: 8–13.

Kellow JE, Azpiroz F, Delvaux M, et al. (2006) Applied principles of neurogastroenterology: Physiology/motility sensation. *Gastroenterology* 130: 1412–1420.

Latiano A, De Giorgio R, Volta U, et al. (2006) HLA and enteric antineuronal antibodies in patients with achalasia. *Neurogastroenterology and Motility* 18: 520–525.

Matsuda H, Hirato J, Kuroiwa M, et al. (2006) Histopathological and immunohistochemical study of the enteric innervations among various types of aganglionoses including isolated and syndromic Hirschsprung disease. *Neuropathology* 26: 8–23.

Schemann M and Neunlist M (2004) The human enteric nervous system. *Neurogastroenterology and Motility* 16: 55–59.

Spiller R and Grundy D (2004) *Pathophysiology of the Enteric Nervous System: A Basis for Understanding Functional Diseases*. Oxford: Blackwell Publishing.

Tack J, Talley NJ, Camilleri M, et al. (2006) Functional gastroduodenal disorders. *Gastroenterology* 130: 1466–1479.

Vasina V, Barbara G, Talamonti L, et al. (2006) Enteric neuroplasticity evoked by inflammation. *Autonomic Neuroscience: Basic and Clinical* 126–127: 264–272.

Wood JD (2004) Enteric neuroimmunophysiology and pathophysiology. *Gastroenterology* 127: 635–657.

Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal

A Rühl, Novartis Pharma AG, Basel, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The enteric nervous system (ENS) is a complex neuronal network which extends along the entire axis of the gastrointestinal tract and controls vital functions, such as motility, secretion, nutrient uptake, blood flow, as well as immune and inflammatory processes.

The gastrointestinal tract constitutes a unique environment for neuronal structures in that it is made up of a large variety of nonneuronal cell types, including enteric glial cells (EGCs), epithelial cells, enteroendocrine cells, smooth muscle cells, interstitial cells of Cajal (ICCs), and fibroblasts as well as fibroblast-like cells, all of which are highly heterogeneous in their ontogenetic origin and functions. In addition, the healthy intestine is in constant and immediate contact with bacterial microbiota and food antigens from the gut lumen and therefore contains a resident population of inflammatory cells contributing to a physiological state of basal, low-level inflammation. Due to this distinct environment, the ENS contains some cellular phenotypes which differ substantially from other parts of the autonomic or peripheral nervous system (PNS).

Neural elements in the gut wall are ensheathed by astrocyte-like EGCs rather than Schwann cells, which are the typical glial cells elsewhere in the PNS. Of all cells, EGCs have the closest contact with neural structures in the gut wall.

However, there is another cell type in a number of different locations in the gastrointestinal tract which also forms close associations with nerve terminals, namely the ICC. ICCs are nonneurectodermal cells which receive neuronal input and at the same time are electrically coupled to intestinal smooth muscle cells (ISMCs).

Both EGCs and ICCs can be conceived as cellular interfaces between the neuronal elements of the ENS and the various nonneuronal cell types in the gut wall.

Enteric Glial Cells

Structure, Growth, and Development

Morphology and anatomy The ENS is formed of two major cell populations – neurons and EGCs. Since their first description by Dogiel in 1899, EGCs have been recognized as the most abundant cell type in the ENS, but quantitative data on EGC numbers are still limited.

EGCs form a widespread network at all levels and throughout all layers of the gastrointestinal tract: they lie adjacent to neuronal cell bodies in the enteric ganglia and accompany the extraganglionic nerve strands in which they are the only cellular elements; they form a dense network around the epithelial crypt bases in the intestinal mucosa, where they come into close contact with the epithelial cell layer; and they are also present in the mesenterium, where they accompany the extrinsic nerves to the gut.

Morphologically, EGCs resemble central nervous system (CNS) astrocytes: both cell types are small, stellate, and express similar antigens, such as the intermediate filaments glial fibrillary acidic protein (GFAP) and vimentin, the calcium-binding protein S-100β, the enzyme glutamine synthetase, and Ran-2. Additionally, EGCs transiently express the intermediate filament nestin during development. In contrast to astrocytes, EGCs are mostly regarded as a relatively homogeneous cell population. However, there is evidence to suggest regional heterogeneity of EGCs even though distinct subsets have not been consistently identified, and the functional significance of this is currently unclear.

EGCs have unique ultrastructural features that differentiate them from the satellite cells of autonomic ganglia and Schwann cells: they envelop neuronal cell bodies and axon bundles without producing myelin; they do not form basal laminae; they ensheathe axons not individually but, rather, in groups; they are extensively coupled by gap junctions; and they make numerous synaptoid contacts with vesicle-containing nerve varicosities.

Development, differentiation, and growth factors acting on EGCs Akin to enteric neurons, glial cells in the gut arise from a population of multipotent neural crest precursor cells whose terminal differentiation occurs in the bowel. Most EGCs evolve from vagal neural crest cells which colonize the entire length of the gut in a proximodistal direction, but some evolve from sacral neural crest cells which migrate in a distal–proximal direction. Although the vast majority of EGCs thus originate from early waves of crest-derived cells, a few of them are derived from precursors that enter the gut along with the extrinsic innervation later in development.

Proliferation and differentiation of the early ENS precursor cells is tightly controlled by microenvironmental factors. Early in the ontogeny of the ENS, nestin-immunoreactive neural crest cells emerge

which are capable of giving rise to both EGC and neurons. These cells proliferate in response to the neurotrophic factors glial cell line-derived neurotrophic factor (GDNF) – most likely from mesenchymal cells in the developing gut wall – and neurturin, but not persephin or endothelin-3.

Subsequently, nestin-positive precursors are replaced by separate lineages of neuronal and glial progenitors which express type-specific cytosolic and intermediate filament proteins. At this stage, the glial precursors are no longer supported by GDNF but, rather, by neurotrophin-3 (NT-3) through the high-affinity tyrosine kinase receptor TrkC. This is modulated by bone morphogenetic protein-2 (BMP-2) and BMP-4, both of which are members of the transforming growth factor-β (TGF-β) superfamily, and promoted by the neuropoietic cytokines ciliary neurotrophic factor (CNTF) and leukemia inhibitory factor through the α and β subunits of the CNTF receptor.

Neuregulins (NRGs) from the epidermal growth factor superfamily constitute crucial factors for further EGC growth and differentiation: the NRG1 family members are neuronally derived molecules that interact with the transmembrane tyrosine kinase receptor erbB3 and its erbB2 coreceptor. Glial growth factor is a soluble NRG1 isoform which strongly suppresses neuronal differentiation while promoting differentiation of peripheral glia, including EGCs. Mice with a targeted deletion of *erbB3* lack enteric glia, whereas CNS astrocytes are unaffected, indicating that NRGs are critically important for EGC development and that there are fundamental differences between astrocytes and EGCs.

Finally, the transcription factors Sox10 and Sox8 have a vital role for the proper differentiation of enteric glia: *Sox10*-deficient mice lack peripheral glial cells including EGCs, and the effects of *Sox10* heterozygosity are aggravated by *Sox8* mutations in *Sox10/Sox8* double-mutant mice, suggesting that *Sox8* acts as a modifier gene for Sox10. Because Sox10 controls the expression of ErbB3 in neural crest cells, *Sox10* ablation phenotypically mimics the deletion of NRG1/ErbB3.

Function

Physiological conditions *Enteric glia and neurons.* Historically, EGCs have been viewed as packing material holding the ENS together ('glia' derives from the Greek word for 'glue'). This concept is challenged by evidence for a role of EGCs in the maintenance of gut functions: loss or defunctionalization of EGCs is associated with changes in neurochemical coding and function of enteric neurons, or neuronal degeneration.

It is unknown if EGCs maintain neuronal integrity via provision of structural support, release of neurotrophic or neuroprotective factors, or creation of a protective local microenvironment by buffering the extracellular milieu (e.g., via uptake of excess cations).

EGCs express voltage-activated inward and outward K^+ channels which are related to the astrocytic K^+ channels, confirming that EGCs could play a role in the prevention of extracellular K^+ accumulation in the ENS.

Morphological data suggest that EGCs structurally stabilize the ENS because they are anchored to the surface of enteric ganglia and nerve strands through GFAP-positive filaments. Whether the production of GDNF, nerve growth factor, or NT-3 by EGCs contributes to the maintenance of enteric neurons is unknown. Akin to astrocytes, EGCs may have an active role in information transfer and may modulate neuronal communication and synaptic plasticity.

Evidence for EGC contribution to neurotransmitter synthesis and inactivation comes from, for example, immunohistochemical findings that in the ENS, glutamine synthetase is exclusively expressed by EGCs. EGCs may thus have a pivotal role in glutamatergic signaling through detoxification of waste ammonia and glutamate and also as a source of glutamine for neuronal glutamate and γ-aminobutyric acid (GABA) (re-)synthesis. Intriguingly, immunohistochemical evidence further supports a role for EGCs in GABAergic neurotransmission: the high-affinity GABA transporter GAT2 appears to be predominantly expressed in EGCs, implicating them in GABA detoxification via rapid removal from the extracellular space.

EGCs could also be involved in neuropeptidergic transmission via inactivation and removal of enteric neuropeptide transmitters. In the gut wall, the oligopeptide transporter PEPT2 is almost exclusively expressed by EGCs, which are thus able to rapidly take up di- or tripeptide neuropeptide degradation products.

Evidence for a role of EGCs in nitrergic neurotransmission comes from the observation that enteric L-arginine immunoreactivity is almost exclusively in EGCs, suggesting that EGCs are the main reservoir of L-arginine – an essential precursor for nitric oxide (NO) – so that neuronal NO synthesis may depend on glial L-arginine delivery.

In an animal model of partial EGC ablation, glial loss is accompanied by a decrease in neuronal vasoactive intestinal polypeptide (VIP) and NO expression levels, confirming the role of EGCs in peptidergic and nitrergic neurotransmission. EGCs are further involved in nucleotide signaling: in the ENS, expression

of the ectonucleotidase NTPDase2 is restricted to EGCs, suggesting that they control extracellular ATP and ADP concentrations. EGCs further express purinoceptors of the $P2X_7$ and $P2Y_2$ subtypes and respond to the purinergic agonists ATP and UTP. An exciting study has provided indirect evidence that EGCs are also a source of ATP, which in turn participates in the intercellular propagation of Ca^{2+} waves between EGCs. Because Ca^{2+} wave-induced glial ATP release could elicit neuronal responses, nucleotide signaling may be instrumental for information transfer between enteric neurons and glia and serves to illustrate that enteric glia do not only participate in neurotransmitter synthesis and inactivation but also, akin to astrocytes, may play an active role in neurotransmission via a bidirectional glia–neuron axis. The demonstration of propagated intercellular Ca^{2+} waves adds to the evidence for functional networks between EGCs via gap junctions, allowing for concerted glial activity in the gut.

EGCs are activated by numerous neurotransmitters and neuroligands, such as epinephrine, serotonin, endothelin family members (through an ET_B receptor linked to phospholipase C and intracellular inositol 1,4,5-triphosphate (IP_3) receptors), and vasopressin, as well as by molecules implicated in neuroimmune cross-talk, such as histamine, prostaglandins, and the protease-activated receptors-1 and -2. Central vagal efferent stimulation of the ENS induces glial expression of the immediate early gene Fos through nicotinic acetylcholine receptors, and α_{2a}-adrenoceptors are present on a subset of EGCs. Bioactive lipid mediators such as sphingosine-1-phosphate and lysophosphatidic acid activate EGCs through a G-protein-coupled receptor system. Finally, EGCs respond to stimulation of primary afferent nerves by intraluminal capsaicin and inflammation.

Enteric glia and endothelial/epithelial permeability. In murine models, ablation of EGCs induces vascular lesions (e.g., a submucosal vasculitis), indicating a breakdown of the vascular endothelial barrier, and mucosal disruption, implicating EGCs in the regulation of endothelial and epithelial permeability. However, the interactions between EGCs and intestinal blood vessels that modulate microvascular functions, particularly endothelial permeability, are unclear, and the debate regarding whether an equivalent of the 'blood–brain barrier' exists in the ENS is unresolved.

In contrast, glial interactions with the intestinal epithelium are more thoroughly investigated. The epithelial lining of the gastrointestinal tract is formed by a continuous monolayer of polarized cells which are held together by tight junctions and constitute a first-line defense system against noxious contents of the gut lumen. Soluble factors released from EGC (e.g., TGF-β) appear to enhance the intestinal epithelial seal (e.g., by acting on the epithelial cytoskeleton and tight junction proteins). Furthermore, EGCs are a potential source of neurotrophins such as GDNF which modulate intestinal epithelial cell apoptosis.

Thus, EGCs seem to be crucial components of the mucosal defense system via sustainment of the intestinal epithelial barrier. Furthermore, EGC have been observed to inhibit intestinal epithelial cell proliferation, implicating them in the prevention of epithelial barrier disorganization and hyperproliferation during early inflammatory processes, bacterial infection, or epithelial dys- and neoplasia.

Pathophysiological conditions *(Neuro-)inflammation.* Intestinal inflammation is associated with functional and structural alterations of the ENS, the majority of which are secondary to the inflammatory process. However, in animals with compromised EGC networks, glial alterations occur very early in the course of intestinal inflammation, which causally implicates these cells in intestinal inflammation.

In animal models, EGCs vigorously respond to inflammatory insults, and glial proliferation rate as well as GFAP expression levels are consistently upregulated. In some models, this is associated with signs of neuronal disintegration or decreased neuronal cell counts.

Interestingly, a diminished EGC network with reduced GFAP expression was described in the non-inflamed ileum of Crohn's disease patients and implicated in disease etiopathogenesis. However, this finding was subsequently challenged by another study reporting increased EGC counts in the inflamed human intestine. It is not clear which mechanisms determine EGC proliferation in the inflamed intestine, but published evidence suggests it is differentially modulated by the various cytokines expressed in the inflamed intestine.

Inflammation-induced EGC activation is not restricted to enhanced cell proliferation. Expression of major histocompatibility complex class II (MHC II) antigen on EGCs has been demonstrated in patients with Crohn's disease, accompanied by lymphocytic infiltration and neuronal damage in the ENS, suggesting that EGCs may direct immunologically mediated damage to the ENS.

The concept of functional EGC–immune cell interactions is endorsed by a series of *in vitro* experiments that have ascertained the capacity of purified EGCs for phagocytosis and cytokine-inducible MHC II and intercellular adhesion molecule-1 expression which is fully functional for antigen-specific, MHC-restricted T cell activation. Additionally, EGCs appear to produce and be regulated by cytokines, for example, via functional receptors for the

proinflammatory cytokines interleukin (IL)-1β and IL-6. Also, EGCs from normal mouse mucosa express the neuropeptides neurokinin (NK) A and substance P, both of which have the potential to act on both immune and epithelial cells.

Overall, EGCs actively respond to inflammation. Also, because they interact with enteric neurons, endothelial cells, and the intestinal epithelium, all of which participate in intestinal inflammation, EGCs may act as an integral part of an enteric neuroimmune network that contributes to and regulates inflammatory processes in the intestine.

Neurodegeneration. A substantial loss of EGCs has been described in the aged ENS and in patients with various functional and/or inflammatory disorders, including intractable slow transit constipation and diverticulitis. It is unclear if EGC decay precedes or follows neuronal loss. EGCs could contribute to neurodegeneration by attracting immune cells to the ENS. Alternatively, impaired glial functions could result in insufficient support of enteric neurons and thus cause neurodegeneration, even in the absence of inflammation.

Malign proliferation. Notably, in contrast to astrocytomas and glioblastomas, cancerous proliferation of EGCs is unknown. The reason for this difference is not understood.

Interstitial Cells of Cajal

Structure, Growth, and Development

Morphology and anatomy In the gut wall, ICCs are present as networks associated with ENS plexuses and dispersed inside most circular muscle layers, including the sphincters. ICCs can be recognized by their unique ultrastructure and expression of the proto-oncogene *c-kit* that encodes for the receptor tyrosine kinase Kit. ICCs typically have fusiform cell bodies and multiple long processes, and they express a variety of gap junction proteins, including connexins 40, 43, and 45. ICCs are electrically coupled with each other as well as with the ISMC syncytium via gap junctions, and their intricate networks extend along and around the gastrointestinal organs to all points involved in phasic contractions.

Based on topographical and functional evidence, the following groups of ICCs have been defined: ICC-MY (also called ICC-MP or ICC-AP) are located in the plane of the myenteric plexus between the circular and longitudinal muscle layers; ICC-DMP are located in the small intestine in close association with the nerve terminals of the deep muscular plexus and close to the submucosal surface of the circular muscle layer; ICC-IM, also called intramuscular ICCs, are distributed within the muscle layers of the gastrointestinal tract, where they are interposed between varicose nerve terminals and ISMC and receive input from enteric motor neurons; and ICC-SM are the pacemaker cells along the submucosal surface of the circular muscle layer in the colon. ICC-SM also occur in the gastric antrum and pylorus, but the physiological roles of these ICCs have not been identified. Finally, ICCs can be found in the septa between intestinal smooth muscle bundles; these cells are called ICC-SEP.

ICC-MY and ICC-SM have membrane caveolae that may serve in cellular signaling processes and provide a scaffold for transport and signaling proteins. L-type Ca^{2+} channels, plasma membrane Ca^{2+} pump protein, and Na^{+}/Ca^{2+} exchanger type 1 were found co-localized with caveolin-1 in ICCs. ICCs express an abundance of mitochondria and prominent smooth endoplasmic reticulum (SER), which are often closely associated with each other and lie in close proximity (within 10–20 nm) to the plasma membrane.

Development and differentiation The ontogeny of ICCs was settled in the 1990s when it was eventually consented that ICCs and ISMC arise from common mesenchymal progenitors. From midembryonic stages, ICC precursors express the receptor tyrosine kinase Kit. Both ICCs and many ISMC precursors initially express Kit, and then the cells destined to become ISMC downregulate Kit and upregulate the synthesis of myofilament proteins, whereas cells destined to differentiate into ICCs maintain Kit expression.

In adults, ICCs are dependent on the ligand for the Kit-receptor – namely Kit ligand, stem cell factor, or Steel factor – and disappear in its absence, whereas in mutant mice with blocked Kit activity, normal ICCs are present until soon after birth, suggesting that the Kit signaling pathway is not necessary for the embryonic development of ICCs but, rather, for their postnatal development and maintenance.

Insulin and insulin-like growth factor type-1 have also been implicated as essential growth factors for ICCs.

Function

Physiological conditions ICCs provide several important functions in the gastrointestinal tract: (1) generation of electrical slow wave activity, (2) coordination of pacemaker activity and active propagation of slow waves, (3) transduction of the motor output from the ENS to the ISMC syncytium, and (4) mechanosensation to stretch of gastrointestinal muscles.

Slow waves are rhythmical oscillations of the membrane potential which spread, without decrement, through the ICC network. Release of Ca^{2+} from

IP_3 receptor-operated stores is the fundamental cellular event that initiates pacemaker current. Uptake of Ca^{2+} and cycle-to-cycle maintenance of Ca^{2+} levels in IP_3 receptor-operated stores, mediated by SERCA Ca^{2+} pumps, is critical for Ca^{2+} release events in ICCs. The contribution of additional ionic conductance(s) to pacemaker activity and the steps between Ca^{2+} release and activation of pacemaker currents is currently under investigation.

Because ICCs are electrically coupled to ISMC, their activity is conducted into the ISMC syncytium which responds to slow wave depolarizations with activation of a variety of voltage-dependent ion channels to generate contractions. Thus, ICC electrical slow wave activity organizes excitation–contraction coupling into a pattern of phasic contractions in the gut.

Unlike the heart, the intestinal smooth muscle layers do not have point sources, or nodes, with specialized electrical pacemaker cells but, rather, continuous networks consisting of thousands of spontaneously active, electrically coupled ICCs that function as a unit with the dominant pacing frequency provided by the cells with the most rapid pacing frequency.

In the stomach and small intestine, there are proximal-to-distal pacemaker frequency gradients which are the basis for oral to aboral peristaltic waves. Slow waves do not propagate from the stomach into the small intestine because these regions are separated by an isolating region of electrically quiescent ISMC.

In the small bowel, pacemaker frequencies and propagation properties of slow waves organize motility into segmental contractions which are typical for the postprandial motor pattern in this region of the gut.

The colon displays two pacemaker regions, one at the submucosal surface of the circular muscle layer and the other between the circular and longitudinal muscle layers, and a distal-to-proximal slow wave frequency gradient has been suggested.

The ENS can regulate the excitability of the ISMC syncytium to slow wave depolarizations. However, motor neurotransmission in the gut does not occur through loosely defined synaptic structures between nerves and ISMC but, rather, via synaptic-like connections between varicose nerve terminals and ICC-IM. Since ICC-IM are coupled to ISMC, the electrical responses elicited in ICCs are conducted to ISMC. Thus, the mechanical response of phasic muscles to the ongoing slow wave activity is strongly modulated by the ENS through regulation of ISMC excitability.

ICCs express functional receptors to receive ENS motor input, such as NK-1 and -3, and somatostatin 2A receptors. After exposure to substance P, NK1 receptors on ICCs are internalized and receptor immunoreactivity aggregates in the cytoplasm, indicating receptor-mediated ICC activation. In addition, ICCs express the muscarinic receptors M2 and M3 for cholinergic input and also VIP receptors (VIP-1). They have also been demonstrated to receive nitrergic input.

Morphological evidence also supports a role for ICC-IM as possible integrators for in-series stretch-dependent changes: vagal afferent nerve endings can terminate as intramuscular arrays within the muscle wall of the stomach. These endings form close appositions with ICC-IM, upon which they appear to form multiple spiny appositions or varicosities that may provide a role in stretch perception. Thus, ICCs could play a role as mechanosensitive transducers.

Pathophysiological conditions ICC damage has been implicated in a variety of gastrointestinal motility disorders, such as gastroparesis, pseudo-obstruction, chronic constipation, and posttraumatic or postinfectious dysfunction. Disruption of ICC networks will abolish synchronization of electrical pacemaker activity and allow for localized increase in slow wave frequency as well as the appearance of ectopic pacemaker sites, resulting in disorganized motor patterns.

ICCs are among the first cells that are injured during an inflammatory insult, and maintenance of ICC health or promotion of their regeneration may be crucial in recovery after injury. In humans, there is evidence of injury to the processes of ICCs and nerves of the colonic submuscular plexus in patients with active ulcerative colitis. Therefore, ICCs may be a major factor in inflammation-induced gastrointestinal motility disorders.

Finally, a pathological condition has been described in which pacemaker frequency in the distal stomach – the antrum – is augmented so that the more proximal gastric regions can no longer drive antral activity. This condition is termed tachygastria, and it is associated with delayed gastric emptying.

See also: Autonomic Neuroimmunology; Cajal's Place in the History of Neuroscience; Enteric Nervous System: Sensory Pathways; Enteric Nervous System: Neural Circuits and Chemical Coding; Enteric Nervous System Development; Enteric Nervous System: Disorders; Enteric Nervous System: Neurotrophic Factors; Inflammation in Neurodegenerative Disease and Injury; Nausea and Vomiting.

Further Reading

Aube AC, Cabarrocas J, Bauer J, et al. (2006) Changes in enteric neurone phenotype and intestinal functions in a transgenic mice model of enteric glia disruption. *Gut* 55: 630–637.

Bush TG, Savidge TC, Freeman TC, et al. (1998) Fulminant jejunoileitis following ablation of enteric glia in adult transgenic mice. *Cell* 93: 189–201.

Cornet A, Savidge TC, Cabarrocas J, et al. (2001) Enterocolitis induced by autoimmune targeting of enteric glial cells: A possible mechanism in Crohn's disease? *Proceedings of the National Academy of Sciences of the United States of America* 98: 13306–13311.

Dogiel AS (1899) Ueber den Bau der Ganglien in den Geflechten des Darmes und der Gallenblase des Menschen und der Säugethiere. *Zeitschrift für Naturforschung B*: 130–158.

Gabella G (1981) Ultrastructure of the nerve plexuses of the mammalian intestine: The enteric glial cells. *Neuroscience* 6: 425–436.

Geboes K, Rutgeerts P, Ectors N, et al. (1992) Major histocompatibility class II expression on the small intestinal nervous system in Crohn's disease. *Gastroenterology* 103: 439–447.

Gershon MD and Bursztajn S (1978) Properties of the enteric nervous system: Limitation of access of intravascular macromolecules to the myenteric plexus and muscularis externa. *Journal of Comparative Neurology* 180: 467–488.

Gershon MD and Rothman TP (1991) Enteric glia. *Glia* 4: 195–204.

Goyal RK and Hirano I (1996) The enteric nervous system. *New England Journal of Medicine* 334: 1106–1115.

Jessen KR and Mirsky R (1983) Astrocyte-like glia in the peripheral nervous system: An immunohistochemical study of enteric glia. *Journal of Neuroscience* 3: 2206–2218.

Rühl A (2005) Glial cells in the gut. *Neurogastroenterology and Motility* 17: 777–790.

Rumessen JJ and Vanderwinden JM (2003) Interstitial cells in the musculature of the gastrointestinal tract: Cajal and beyond. *International Review of Cytology* 229: 115–208.

Vanderwinden JM, Rumessen JJ, Bernex F, Schiffmann SN, and Panthier JJ (2000) Distribution and ultrastructure of interstitial cells of Cajal in the mouse colon, using antibodies to Kit and Kit(W-lacZ) mice. *Cell and Tissue Research* 302: 155–170.

Young HM (1999) Embryological origin of interstitial cells of Cajal. *Microscopy Research and Technique* 47: 303–308.

Young HM, Bergner AJ, and Müller T (2003) Acquisition of neuronal and glial markers by neural crest-derived cells in the mouse intestine. *Journal of Comparative Neurology* 456: 1–11.

Enteric Nervous System: Neural Circuits and Chemical Coding

J B Furness, University of Melbourne, Parkville, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The enteric nervous system (ENS) contains the integrating nerve circuitry for control of the movements of the gastrointestinal tract, for transfer of fluids across the mucosal surface in the small intestine and colon, for local blood flow control, and for control of secretion of gut hormones. Enteric neurons also influence the gut immune system. The intrinsic nerve circuits contain 14–20 types of neurons, depending on region and species. In the small intestine of the guinea pig, in which all types of neurons have been identified, 15 neuron types occur. In addition to the intrinsic neurons, the endings of extrinsic neurons (vagal, pelvic, and sympathetic efferent and vagal and spinal afferent neurons) intermingle with the processes of enteric neurons. Along with the fact that the cell bodies of intrinsic primary afferent neurons, interneurons, and motor neurons are not segregated in enteric ganglia, this large number of participating neurons presented a major problem to scientists who tried to unravel the enteric neural circuitry. Part of the solution to this problem came when it was discovered in the late 1970s and the early 1980s that there were a range of chemical markers of enteric neurons. From this discovery arose the chemical coding hypothesis. The hypothesis in its simplest form proposes that neurons of every functional class can be distinguished from neurons of other classes by chemical attributes that constitute a chemical code. In the ENS, the code is contributed to by neuropeptides, which is fortunate because peptides are very suitable for the generation of highly effective antibodies for immunohistochemical studies. Once chemical identities that could be exploited histochemically were established, it became possible to determine the projections and synaptic relations of each class of neuron.

Chemical Coding in the ENS

Table 1 provides a listing of the chemical codes for enteric neurons in the guinea pig small intestine. This reinforces the observation that the chemical codes provide very useful distinctions between neurons of different functional types (Figure 1). Similar, but less extensive, investigations have been made in human, mouse, rat, pig, and sheep ENS. These show that some aspects of the chemical codes are highly consistent between species. This includes codes for excitatory muscle motor neurons that all contain tachykinins and the synthesizing enzyme (choline acetyltransferase) for the transmitter acetylcholine (ACh). Inhibitory muscle motor neurons contain nitric oxide synthase (NOS), implying that NO is a transmitter of these neurons, vasoactive intestinal polypeptide (VIP) and pituitary adenylyl cyclase activating polypeptide (PACAP). These neurons also use adenosine 5′-triphosphate (ATP) as a transmitter. In all species, there are secretomotor neurons that contain VIP. Intrinsic primary afferent neurons (IPANs) contain one or both of the calcium-binding proteins, calbindin and calretinin, and either or both substance P (SP) and calcitonin gene-related peptide (CGRP). They are also characterized by binding the lectin, IB4, which is also a marker of some vagal and spinal primary afferent neurons.

The aspects of codes that are highly conserved are those that relate to essential functions of the neurons – for example, primary transmitters, postsynaptic receptors, and associated enzyme systems. Thus, when a substance is a primary neurotransmitter, it appears to be present in all neurons that have equivalent roles in different species and regions of the gastrointestinal tract. For example, extensive studies in many species indicate that all inhibitory neurons to the muscle contain the synthesizing enzyme for NO, NOS, and the peptide VIP. Beyond this, the codes can vary considerably between species, by the inclusion of chemical markers that do not seem to have significant roles. These have been termed superfluous transmitters, although some of the substances are not transmitters. In relation to enteric neurons, their significance has been discussed elsewhere. It has been postulated that some substances may have no role (they could be considered developmental or evolutionary leftovers) or that some substances may have roles other than a participation in neurotransmission.

Some of the coding is related to subtle differences within enteric neuron populations; for example, in the guinea pig small intestine, inhibitory neurons with short projection have the code NOS/VIP/ENK/GABA/NPY/PACAP, whereas the long-projecting neurons have the code NOS/VIP/BN/DYN/NFP/PACAP. Some of the long neurons are also neuropeptide Y (NPY) immunoreactive. Also in human, a proportion of the VIP-containing neurons that innervate the circular muscle are NPY immunoreactive. No differences in the pharmacology of transmission that relate to these chemical coding differences are known. Another example of species difference is that the noncholinergic

Table 1 Chemical coding of functionally defined neurons in the ENS

Functional definition	*Chemical coding*	*Transmitter/comments*	*Comments on species differences*
Myenteric neurons			
Excitatory circular muscle motor neurons (6)	Short: ChAT/TK/ENK/GABA Long: ChAT/TK/ENK/NFP	To all regions, primary transmitter ACh, cotransmitter TK	Excitatory neurons in all species contain ACh and TKs, which subserve transmitter roles
Inhibitory circular muscle motor neurons (7)	Short: NOS/VIP/PACAP/ENK/NPY/GABA Long: NOS/VIP/PACAP/Dynorphin/BN/NFP	Several cotransmitters with varying prominence: NO, ATP, VIP, PACAP	Inhibitory neurons in all species contain NOS and VIP family peptides. Other markers, such as NPY, differ between species
Excitatory longitudinal muscle motor neurons (4)	ChAT/Calretinin/TK	Primary transmitter ACh, cotransmitter TK	In all species these contain ACh and TKs, but calretinin is not always present
Inhibitory longitudinal muscle motor neurons (5)	NOS/VIP/GABA	Several cotransmitters with varying prominence: NO, ATP, VIP, PACAP	In all species these contain NOS and VIP family peptides
Ascending interneurons (1)	ChAT/Calretinin/TK/ENK	Primary transmitter ACh	Not identified in other species
Descending interneurons (local reflex) (8)	ChAT/NOS/VIP $\pm$ BN $\pm$ NPY	Primary transmitter ACh, ATP probably a cotransmitter	Not identified in other species, but commonly contain VIP
Descending interneurons (secretomotor and motility reflex) (9)	ChAT/5-HT	Primary transmitters ACh, 5-HT (at 5-HT_3 receptors)	In all species there appears to be descending interneurons that contain 5-HT
Descending interneurons (migrating myoelectric complex) (10)	ChAT/SOM	Primary transmitter ACh	Not identified in other species
Myenteric intrinsic primary afferent neurons (2)	ChAT/TK/orexin/IB4/NeuNcyt/NK_3 receptor, most calbindin	Primary transmitters TK and ACh	Calbindin is a good marker in guinea pigs, but calretinin is a better marker in rats, mice, and human. These neurons contain CGRP in many species
Intestinofugal neurons (3)	ChAT/BN/VIP/NOS/CCK/ENK	Primary transmitter ACh, cotransmitter VIP	ACh is primary transmitter in all species. VIP seems to be a reliable across-species marker
Motor neurons to gastrin cells[a]	GRP	Primary transmitter GRP	Same marker (GRP) in all mammals investigated
Submucosal neurons			
Noncholinergic secretomotor/vasodilator neurons (12)	VIP/CART/CRF/GAL/PACAP/NMU	Primary transmitter VIP; a small proportion of these have cell bodies in myenteric ganglia	In all species In human, these neurons may be cholinergic
Cholinergic secretomotor/vasodilator neuron (13)	ChAT/Calretinin/Dynorphin	Primary transmitter ACh	Not specifically identified in other species
Cholinergic secretomotor (nonvasodilator) neurons (14)	ChAT/NPY/CCK/SOM/CGRP/Dynorphin/GAL/NMU	Primary transmitter ACh; a small proportion of these have cell bodies in myenteric ganglia	Not specifically identified in other species
Uni-axonal neurons projecting to the myenteric plexus (15)	VIP (NOS?)	Possibly displaced myenteric interneurons	Neurons projecting to the myenteric plexus also identified in the pig. They probably occur in other species
Submucosal intrinsic primary afferent neurons (11)	ChAT/TK/orexin/IB4/NeuNcyt/calbindin	Calbindin-IR, seen with some antisera only. Primary transmitter ACh, may be TK contribution	Appear to be present in all species. CGRP and calretinin are generally markers of these neurons

Continued

Table 1 Continued

Functional definition	*Chemical coding*	*Transmitter/comments*	*Comments on species differences*
Excitatory motor neurons to the muscularis mucosae[a]	ChAT/TK	Primary transmitter ACh	In all species these contain ACh and TKs
Inhibitory motor neurons to the muscularis mucosae[a]	NOS/VIP	Pharmacology of transmission appears to be similar to other enteric inhibitory muscle motor neurons	Inhibitory neurons in all species contain NOS and VIP family peptides

[a]Motor neurons that are found in other parts of the tubular digestive tract.
ACh, acetylcholine; ATP, adenosine 5′-triphosphate; BN, bombesin (the mammalian form is sometimes referred to as GRP); CCK, cholecystokinin; ChAT, choline acetyltransferase; CART, cocaine and amphetamine-regulated transcript peptide; CGRP, calcitonin gene-related peptide; CRF, corticotrophin-releasing factor; ENK, enkephalin; GABA, gamma aminobutyric acid; GAL, galanin; GRP, gastrin-releasing peptide (mammalian bombesin); 5-HT, 5-hydroxytryptamine; IB4, isolectin B4; MMC, migrating myoelectric complex; NeuNcyt, cytoplasmic immunoreactivity for the neuronal nuclear protein, NeuN; NFP, neurofilament protein; NK, neurokinin; NOS, nitric oxide synthase; NPY, neuropeptide Y; NeuNcyt, cytoplasmic NeuN; PACAP, pituitary adenylyl cyclase activating peptide; SOM, somatostatin; TK, tachykinin; VIP, vasoactive intestinal peptide.
This table lists the neuron types that are found in the guinea pig small intestine and some of their defining characteristics. The numbers in brackets are the identifying numbers for the neurons in Figure 1.

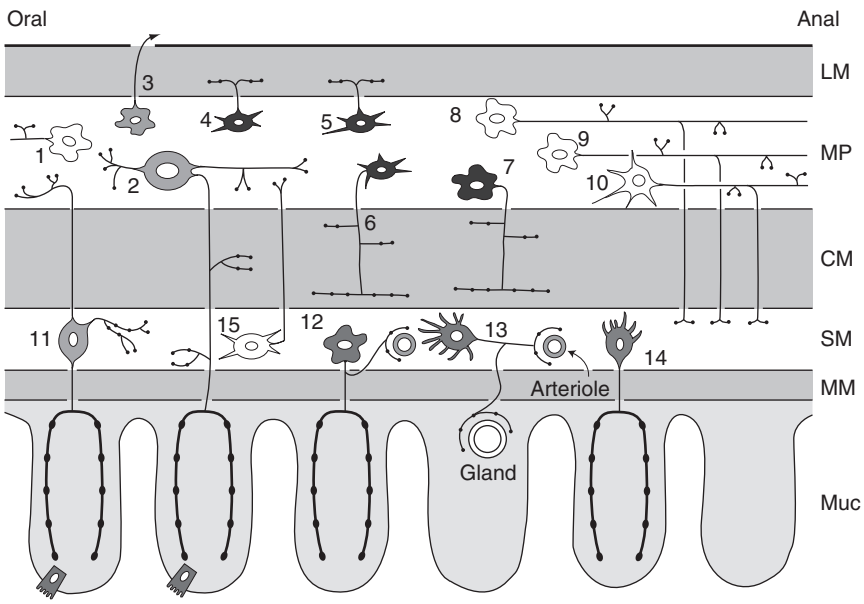

Figure 1 The types of neurons in the small intestine of the guinea pig, all of which have been defined by chemical coding, cell body morphology, key transmitters, projections to targets, and functions. The numbers adjacent to the neurons correspond to the numbers in Table 1, which lists each of the neuron types by their functions and provides data on their chemical coding and transmitters. Projection directions (to oral or to anal) are indicated. Note that the muscularis mucosae (MM) is very thin in the guinea pig small intestine and the neurons that innervate this muscle are not depicted. Motor neurons to the muscularis mucosae are in the submucosal ganglia and are prominent in other species. LM, longitudinal muscle; MP, myenteric plexus; CM, circular muscle; SM, submucosal plexus; Muc, mucosa.

secretomotor neurons in guinea pig do not contain NPY, whereas they do in rats.

Peptide Immunoreactivity

Many of the substances that form the chemical codes are peptides. In most cases, peptides belong to families of similar molecules, not all of which were known when the immunoreactivity or mode of transmission of enteric neurons was first discovered. For example, antibodies raised against SP revealed enteric neurons, and antagonists for SP's actions reduced transmission. Thus, the earlier literature refers to SP as the marker and transmitter of enteric neurons. Later, it

was found that the same gene that gives rise to SP codes for the closely related peptide neurokinin A (NKA), that most anti-SP antibodies recognize NKA, and that both contribute to excitatory transmission from the same neurons. The collective name for these peptides, tachykinins, is now used to describe the contents and transmitters of these neurons. Similarly, immunoreactivity for VIP could also include immunoreactivity for the closely related peptides of the VIP family, PHI and PHM. These problems of exact definition of the peptide contained in and released from enteric neurons extend to enkephalins and the different molecular forms of CCK, and most likely to other peptides.

Chemical Coding and Plurichemical Transmission

Pharmacological studies between the 1960s and 1980s (and even more recent studies) indicated that there were multiple substances that contributed to transmission in the gastrointestinal tract. One of the outcomes of investigations of chemical coding was that investigators were obliged to conclude that the same neuron releases several transmitters. This appears to be the rule in the ENS. Several potential neurotransmitters have been identified in almost all neurons, and there is direct evidence of several neurotransmitters for some of the neurons. Examples of neurons that release several transmitters are included in Figure 2. The type of transmission that occurs in the ENS, which involves several transmitter substances derived from the same neuron, has been called plurichemical transmission. Among the transmitter substances that are released from a single neuron, one usually has a dominant role. This has been called the primary transmitter.

Excitatory Muscle Motor Neurons

These neurons have as their primary transmitter ACh, which acts on muscarinic receptors on the muscle (the dominant receptor subtypes being M_2 and M_3 receptors). Thus, transmission from the neurons is effectively abolished by muscarinic receptor block. However, in all species and regions, these neurons contain tachykinins (SP, NKA, and some related peptides), and if transmission is carefully analyzed, it is found that a small component of transmission is contributed by tachykinins, which act through NK1 and NK2 receptors. The method of choice to locate the nerve terminals of excitatory neurons in the muscle is using tachykinin immunoreactivity. These neurons also contain opiate (enkephalin-related) peptides that act presynaptically on the nerve endings from which they are released to reduce transmitter output.

Inhibitory Muscle Motor Neurons

There was a period of difficulty and substantial debate associated with the identification of the transmitter of inhibitory neurons. It is now recognized that these neurons utilize cotransmitters, including ATP, NO, and peptides of the VIP family.

The primary transmitter is generally NO. NO is an unusual chemical messenger because it is not stored but, rather, created on demand when the synthesizing enzyme, NOS, is activated. NO is membrane permeable, and it crosses the membrane to interact with its receptor, soluble guanylyl cyclase, in the cytoplasm. NOS is readily revealed in tissue, either by enzyme histochemistry or by immunohistochemistry, and it is a very good marker of inhibitory neurons supplying gut muscle, although VIP immunoreactivity is also an excellent label for these neurons.

ATP, which is also released as a transmitter from these neurons, acts on both P2X and P2Y receptors on the muscle. The effect of ATP on the muscle is mediated by the opening of apamin-sensitive K^+ channels (SK channels), and apamin is in many cases an effective antagonist of the ATP-mediated component of inhibitory transmission. It is clear that there must be a link between purine receptor activation and SK channel opening. The other transmitter type of these neurons is the VIP peptide family. There is good evidence that two family members, VIP and PACAP, are both released, but there may be other related peptides, such as PHI. These peptides both act through VPAC receptors. The relative roles of the three transmitters differ between species and gut region, even within the same species. There is also evidence that carbon monoxide could contribute to inhibitory transmission.

One consequence of there being multiple transmitters of the inhibitory neurons may be that a deficiency in one transmitter can be partly compensated. There is little in the way of functional deficit in mice in which neuronal NOS is eliminated by gene knockout. The mice have gastric outlet obstruction, which is due to a failure of the pyloric sphincter to adequately relax, but motility control in the remainder of the bowel is close to normal.

Chemical Coding and Transmission from Secretomotor and Vasodilator Neurons

In the small intestine and colon, there are neurons that have both secretomotor and vasomotor functions (secretomotor/vasodilator neurons). These neurons have axons that branch to supply both the mucosal epithelium and the arterioles in the mucosa. Most prominent among these are neurons that contain the VIP family peptides, VIP and PACAP; these

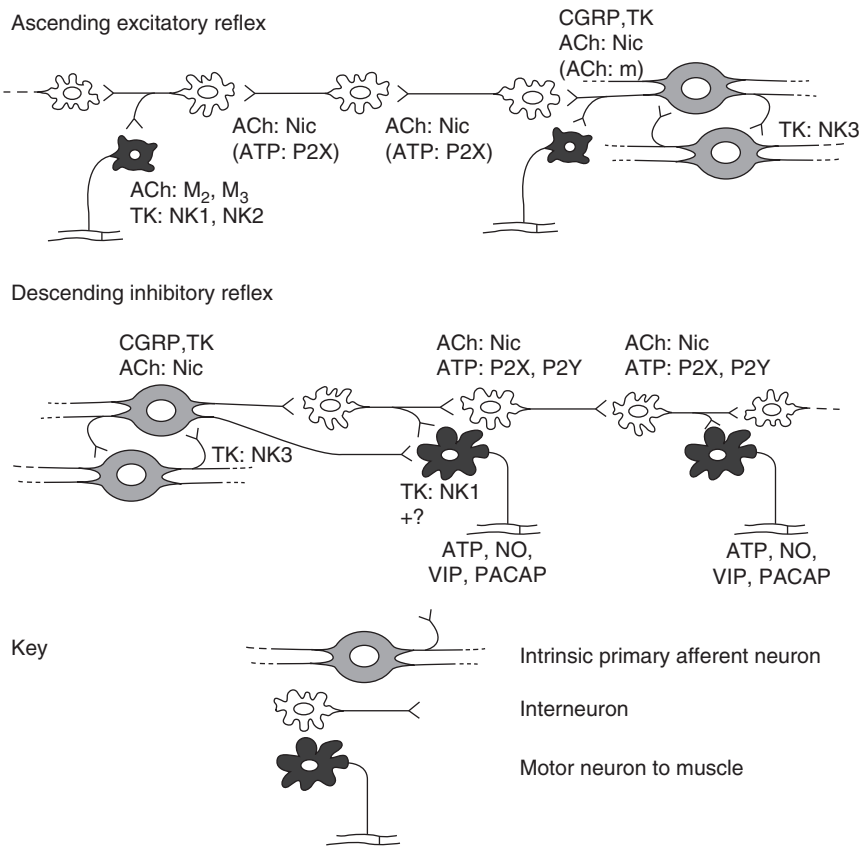

Figure 2 Transmitters of ascending and descending reflex pathways supplying the muscle. Transmitters are indicated adjacent to the synapses where they act. Information flows from the intrinsic primary afferent neurons to motor neurons directly or through one or more interneurons. Primary transmitters and their receptors are indicated (e.g., ACh: Nic, primary transmitter, acetylcholine, acting at nicotinic receptors). In cases in which roles are minor or uncertain, the information is in brackets. See **Table 1** for transmitter abbreviations. Receptors: Nic, nicotinic; m, muscarinic; M2, M3, muscarinic 2, 3; NK1 (2, 3), neurokinin 1 (2, 3); P2X, purine 2X; P2Y, purine 2Y.

are noncholinergic secretomotor/vasodilator neurons. In cats, rats, guinea pigs, and humans (and probably all mammals), they use VIP/PACAP as the principal transmitter, and in humans and rats at least, the peptides act on the mucosal epithelium at VPAC1 receptors. The chemical coding of these neurons has not been worked out in all species; in guinea pig they are coded by VIP/PACAP/CART/CRF/GAL/NMU, in rat by VIP/PACAP/NPY/calbindin (and maybe more substances), and in human by VIP/PACAP/NPY (there are probably other contributors to this code that are yet to be discovered).

Cholinergic secretomotor/vasodilator neurons also occur, but it is the noncholinergic neurons that appear to dominate physiological responses. In guinea pigs, the cholinergic secretomotor/vasodilator neurons are best revealed by their content of calretinin. These neurons have not been adequately characterized in other species.

In addition, there is a population of cholinergic vasodilator neurons that have one of the most extensive neuropeptide codes discovered to date. The code in the guinea pig is ChAT/NPY/CCK/SOM/CGRP/Dynorphin/GAL/NMU (seven different neuropeptides).

Gastrin-releasing peptide (GRP; also known as mammalian bombesin) is a primary transmitter from motor neurons to gastrin-containing endocrine cells of the antral mucosa. Analysis of chemical codes indicates that these neurons also express a cholinergic phenotype (they are ChAT immunoreactive). Moreover, pharmacological analysis indicates that there is also a cholinergic contribution to transmission. This is quite an unusual situation. A dominant idea is that when co-localization of a neuropeptide and a classical transmitter such as ACh occurs, the major transmitter role is taken by the classical transmitter and a minor role is taken by the peptide. Here, that is reversed.

Chemical Coding of Sympathetic, Noradrenergic Neurons That Supply the Gut

The coding of the sympathetic innervation of the gastrointestinal tract has been very useful in the analysis of

the circuitry. The cell bodies of the sympathetic neurons that supply the gastrointestinal tract are in sympathetic chain (paravertebral) ganglia and in the large prevertebral ganglia of the abdominal and pelvic cavities (the coeliac, superior mesenteric, inferior mesenteric, and pelvic ganglia). Although indirect evidence had suggested that three neuron types, controlling three functions (motility, secretion of fluid, and vasoconstriction in the intestine) were separate, this was made much more obvious when it was discovered that the three classes of neuron had different chemical coding. The vasoconstrictor neurons innervate the arteries directly and cause constriction through the release of noradrenaline and ATP; these have the code NA/NPY in guinea pigs (and probably in most other mammals). Secretomotor inhibitory neurons have the code NA/SOM, with the primary transmitter role being taken by noradrenaline. The motility-controlling neurons may also have a neuropeptide code in the guinea pig, but the accompanying peptide is not known.

Enteric Nerve Circuits

The chemical codes that have been uncovered have provided a major impetus to solving the problem of the organization of neural circuits in the intestine (Figure 3). Histochemical identification of neurons according to their chemistry has been combined with intracellular injection of dye, nerve lesion studies, and retrograde tracing to identify the morphologies and projections. To this has been added information of the synaptic connections that are made, which has been determined by electron microscopy, electrophysiological analysis of synaptic transmission, and pharmacological investigation of enteric reflexes. A concurrence between results obtained with each of the methods has led to a very robust depiction of the circuitry for motility control in the small intestine (Figure 3) and also for control of water and electrolyte secretion and local blood flow.

In response to suitable stimuli, the motility-controlling circuits in the small intestine can switch between different functional roles, including propulsive movements, mixing, retropulsion, and the migrating myoelectric complex. How this switching occurs is not known.

The functions and general circuit organization in the large intestine are similar to those in the small intestine, and although the circuits have not been studied in the meticulous detail that they have been in the small intestine, they are presumed to be quite similar. The ENSs of the esophagus and stomach are quite different.

In the esophagus, the lining is protective, not fluid transporting, and there is no submucosal plexus in this part of the gastrointestinal tract and no secretomotor or local vasomotor reflexes. In addition, motor control of the external muscle, where it is composed of striated muscle (which is the majority of the esophagus in most species of mammal), is through circuits in the central nervous system (specifically, in the medulla oblongata) via vagal nerve pathways. Control of gastric motility and acid secretion (its major nerve-controlled functions) is also primarily through the brain stem nuclei, although there is a small degree of regulation by enteric circuits.

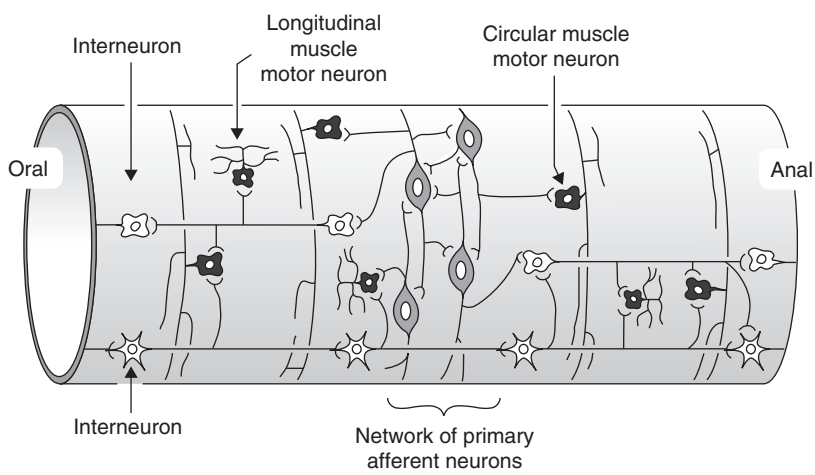

Figure 3 The circuitry for motility reflexes in the small intestine (simplified). Reflex pathways are depicted in a short segment of intestine to illustrate circumferential projections of neurons. The first parts of descending inhibitory and ascending excitatory reflex pathways for propulsive reflexes are depicted. Intrinsic primary afferent neurons (IPANs) are circumferentially oriented and form self-reinforcing networks. They provide outputs to ascending and descending interneurons and monosynaptic connections to motor neurons. One type of descending interneuron (bottom), which is involved in conducting the migrating myoelectric complex along the small intestine, receives few inputs from IPANs but connects with motor neurons.

The brain stem circuits exert their effects through axons that run in the vagus nerve and synapse on enteric neurons.

See also: Cortical Processing of the Reward Value of Food; Enteric Nervous System: Sensory Pathways; Enteric Nervous System Development; Enteric Nervous System: Disorders; Neuropeptides: Enteric Nervous System.

Further Reading

Bowers CW (1994) Superfluous neurotransmitters? *Trends in Neuroscience* 17: 315–320.

Brehmer A (2006) Structure of enteric neurons. *Advances in Anatomy* 186: 1–95.

Brookes SJH and Costa M (2002) Cellular organisation of the mammalian enteric nervous system. In: Brookes SJH and Costa M (eds.) *Innervation of the Gastrointestinal Tract*, pp. 393–467. London: Taylor & Francis.

Costa M, Brookes SJH, Steele PA, et al. (1996) Neurochemical classification of myenteric neurons in the guinea-pig ileum. *Neuroscience* 75: 949–967.

Furness JB (2006) *The Enteric Nervous System*. Oxford: Blackwell.

Furness JB, Morris JL, Gibbins IL, and Costa M (1989) Chemical coding of neurons and plurichemical transmission. *Annual Review of Pharmacology and Toxicology* 29: 289–306.

Furness JB, Young HM, Pompolo S, et al. (1995) Plurichemical transmission and chemical coding of neurons in the digestive tract. *Gastroenterology* 108: 554–563.

Sanders KM and Ward SM (1992) Nitric oxide as a mediator of nonadrenergic noncholinergic neurotransmission. *American Journal of Physiology* 262: G379–G392.

Timmermans J-P, Hens J, and Adriaensen D (2001) Outer submucous plexus: An intrinsic nerve network involved in both secretory and motility processes in the intestine of large mammals and humans. *Anatomical Record* 262: 71–78.

Wattchow DA, Furness JB, and Costa M (1988) Distribution and coexistence of peptides in nerve fibers of the external muscle of the human gastrointestinal tract. *Gastroenterology* 95: 32–41.

Enteric Nervous System: Neurotrophic Factors

M J Saffrey, Open University, Milton Keynes, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Our understanding of the roles of neurotrophic factors in the enteric nervous system (ENS) has lagged behind that in other parts of the peripheral nervous system, largely because of the complexity of enteric neuronal circuitry. Neurons of the ENS are located within the gut wall, in numerous small, irregular-sized ganglia located along the entire length of the gastrointestinal tract. The circuitry of the ENS is difficult to analyze experimentally, because, unlike many parts of the nervous system, where well-identified neurons make clearly defined connections with their target tissues, enteric ganglia, particularly the myenteric ganglia, are composed of a mixture of different neuronal types in varying proportions, with diverse projections to a range of target cell types.

When the prototypic neurotrophic factor, nerve growth factor (NGF), was first identified and its effects on sensory and sympathetic neurons characterized, its possible effects on the ENS were not studied. In the early 1990s, however, when other members of the neurotrophin family, notably brain-derived neurotrophic factor (BDNF) and neurotrophin-3 (NT-3) and their receptors, were identified and their properties and actions began to be elucidated, it was found that several of these proteins were expressed in the gut. At that stage, their localization within the gut wall had not been fully characterized, but analysis of the effects of NT-3 on neural precursors isolated from the embryonic gut indicated that NT-3 promoted differentiation of both enteric neurons and enteric glial cells. These studies provided the first indication that neurotrophins may play a role in the ENS. More recently, supporting evidence for this hypothesis has come from transgenic mice in which NT-3 expression is disrupted; these mice exhibit abnormalities in the ENS (see later).

Soon after identification of the neurotrophins, two major discoveries provided a breakthrough in the study of neurotrophic factors and the ENS. First, in the early 1990s, was the discovery that a mutation in the c-ret gene was associated with Hirshsprung disease, a congenital condition in which varying lengths of the terminal colon are aganglionic, and have no enteric neurons. At this time, although the c-ret gene was known to encode a tyrosine kinase receptor, its ligand was unknown. c-ret −/− knockout mice were found to die shortly after birth, and among other abnormalities were found to lack enteric neurons (and glia) from all regions of the gut except for the esophagus. Soon after this finding, in 1996, transgenic animals lacking the newly discovered neurotrophic factor, glial cell line-derived neurotrophic factor (GDNF), were found to display an identical phenotype to the c-ret −/− mice, and it was discovered that c-ret was the signaling receptor for GDNF. c-ret forms part of a receptor complex, together with a glycosyl phosphatidylinositol (GPI)-linked GDNF binding receptor, now known as GFRα1. GFRα1 knockout mice also display the aganglionic phenotype. Since the discovery of GDNF, three related proteins were discovered, at least one of which, neurturin (NTN), is also active in the ENS.

Since these first seminal studies suggesting roles for GDNF and NT-3 in the developing ENS, much more work has been carried out on the expression and actions of neurotrophic factors in the gut. Most detailed studies have focused on establishing the roles of neurotrophic factors in the development of enteric neurons, a process which is not described in detail here. Increasing efforts, however, are now being put into investigating the patterns of expression and actions of neurotrophic factors in the gut after birth and in maturity, with the aim of determining if these factors are involved in neural plasticity in the adult ENS and if disruption of neurotrophic factor support may be involved in some pathological changes that take place in the ENS, or in ENS aging.

Neurotrophic Factors and Their Receptors in the Gut

The main neurotrophic factors and/or their receptors that have been shown to be present in the gut, or active on enteric neurons, and hence implicated in the development and/or the maintenance of the ENS, are shown in Table 1. It is important to note, however, that in addition to factors classified as neurotrophic factors, other secreted proteins, peptides, and molecules play important roles in the survival, differentiation, and maintenance of the ENS. Examples of such proteins that act on the developing ENS before birth include the hedgehog family members, Indian and Sonic hedgehog, which may be involved in the patterning of the ENS; the bone morphogenetic proteins, BMP-2 and -4, which are involved in early specification of enteric neurons; and endothelins, which inhibit neuronal differentiation and may thus prevent precocious neuronal differentiation in the

Table 1 Neurotrophic factors and neurotrophic factor receptors involved in ENS development and/or maintenance[a]

Neurotrophic factor family	*Receptors*	*Comments*
Neurotrophins	Common low-affinity p75 receptor; high-affinity tyrosine kinase receptors (Trks), with preferential specificity:	NGF, NT-3, and BDNF all expressed in the gut; varied reports of receptor expression by enteric neurons (see text)
Nerve growth factor (NGF)	TrkA	
Brain-derived neurotrophic factor (BDNF)	TrkB	
Neurotrophin-3 (NT-3)	TrkC	Evidence for roles of NT-3 in ENS both pre- and postnatally; some evidence for roles of other factors (see text); no evidence for NT-4/5 expression in the gut, but extrinsic sensory innervation of intestine severely affected in NT-4 −/− mice
Neurotrophin-4/5 (NT-4/5)		
Glial cell line-derived neurotrophic factor (GDNF) family	Common signaling tyrosine kinase receptor c-ret; glycosyl phosphatidylinositol (GPI)-linked binding GFRα family, with preferential specificity:	
GDNF	GFRα1	GDNF and NTN and their receptors expressed in the gut; established roles for GDNF in early ENS development; NTN has later actions
NTN	GFRα2	
Artemin	GFRα3	Some evidence for artemin expression in gut and action on enteric neurons (see text)
Persephin	GFRα4	No evidence for expression or actions of persephin in the gut
Neuropoietic cytokines	Tripartite receptor, comprising:	Evidence that an as yet unknown factor, probably related to CNTF or LIF, acts at a neuropoietic cytokine receptor complex in the ENS; activation of receptor complex implicated in prenatal development, but roles uncertain; some actions postnatally
Ciliary neurotrophic factor (CNTF)	CNTFRα	
Leukemia inhibitory factor (LIF)	Gp130	
Oncostatin M	LIFRβ	
Cardiotrophin-1		

[a]Note: this is not an exhaustive list of all factors/receptors in these families.

developing gut. In several cases these proteins act by influencing the responsiveness of neural precursors to neurotrophic factors such as GDNF.

Other growth factors, proteins, peptides, and even molecules traditionally associated with processes such as neurotransmission may also have neurotrophic actions in the ENS. Examples include fibroblast growth factor 2, which may be expressed by enteric neurons and which promotes neurite outgrowth from postnatal myenteric neurons *in vitro*; adenosine; serotonin, which promotes differentiation of fetal neuronal precursors in culture; and also vasoactive intestinal peptide (VIP) and nitric oxide, which have been found to promote survival of adult myenteric neurons in culture. These molecules are not further considered here.

Roles of Neurotrophic Factors in Early (Prenatal) ENS Development

The most detailed understanding of neurotrophic factor action in the gut is that of GDNF, in particular its roles during early development of the cells of the ENS, which arise from specific regions of the neural crest. Since this topic is covered in detail elsewhere, only an outline is provided here.

Glial Cell Line-Derived Neurotrophic Factor

GDNF is essential for development of the ENS. It is expressed in the early intestinal mesenchyme, just after the arrival of migrating neural crest-derived progenitors in the developing gut; these progenitors express the signaling receptor for GDNF, c-ret, as or just before they arrive in the gut. In cell culture models, GDNF has been found to promote survival, proliferation, and differentiation of ENS progenitor cells as well as their migration along the growing gut. As already mentioned, transgenic mice in which the GDNF gene is disrupted have an aganglionic gut distal to the esophagus and are not viable. Targeted disruption of either the c-ret or the GFRα1 genes produces animals with a similar ENS phenotype.

Neurturin

NTN is also expressed in the developing gut mesenchyme, and GFRα2 is expressed by developing enteric ganglion cells. Both NTN and GFRα2, however, are first detected slightly later than GDNF and GFRα1 in

the developing mouse gut. *In vitro*, NTN promotes the survival and proliferation of enteric neural progenitors; *in vivo*, targeted disruption of the NTN or GFRα2 genes results in mice which exhibit abnormalities of the ENS. The enteric abnormality in NTN and GFRα2 knockout animals results in a reduction in neuronal perikaryon size, a small but consistent reduction in myenteric neuronal numbers, and a reduction in the density of cholinergic nerve fibers in adult animals. GFRα2 knockout mice also display a reduction in the density of substance P-immunopositive fibers. Both NTN −/− and GFRα2 −/− transgenic animals also exhibit abnormalities in gut function, including reduced smooth muscle contractility and reduced responses to electrical stimulation and carbachol. These changes may contribute to the poor growth of NTN −/− and GFRα2 −/− mice seen after birth.

Artemin

A recent study has shown that overexpression of artemin in mice results in hyperplasia of myenteric ganglion cells, and that artemin has a small effect on enteric neurons *in vitro*. Clear evidence for expression of GFRα3 by enteric neuronal precursors or differentiated enteric neurons, however, is lacking, and it is possible that the effects observed in this study were due to an action of artemin binding to GFRα1 or GFRα2.

Neurotrophin-3

NT-3 is expressed in the developing intestinal mesenchyme, and TrkC receptors are expressed by enteric neural progenitors. *In vitro*, NT-3 promotes survival of progenitors and differentiation of enteric neurons and glia. *In vivo*, NT-3 −/− and TrkC −/− transgenic mice have enteric ganglia, although recent evidence has shown that NT-3 null mice have reduced numbers of myenteric and submucosal neurons. Conversely, transgenic mice in which overexpression of NT-3 is targeted to the myenteric plexus have increased numbers of myenteric neurons, of a larger size than those in wild-type mice. Evidence also indicates that NT-3 acts together with other factors, and may be important in later stages of fetal ENS development.

Nerve Growth Factor and Brain-Derived Neurotrophic Factor

Although NGF, BDNF, TrkA, and TrkB transcripts have all been detected in the developing gut, there is no evidence that they have a role in early ENS development.

Neuropoietic Cytokines

The neuropoietic cytokines ciliary neurotrophic factor (CNTF) and leukemia inhibitory factor (LIF) are expressed in the gut; CNTF has been detected in the fetal mouse gut, and LIF is expressed in postnatal rat gut. Neural crest-derived neural precursors immunoselected from embryonic day 14 (E14) fetal rat gut express the receptor, CNTFRα, as do developing neurons in the E16 and E18 rat ENS. Although CNTF and LIF knockout mice appear to have a normal ENS, transgenic mice in which either the CNTF receptor α or LIF receptor β are disrupted die at birth and have reduced numbers of enteric neurons; in particular, they exhibit a loss of substance P-immunoreactive and nitric oxide synthase (NOS)-immunoreactive neurons that innervate smooth muscle.

Although the actions of all these different types of neurotrophic factors have been summarized separately here, evidence suggest that they are likely to interact; ENS development is dependent upon a range of neurotrophic and other factors.

Developmental Disorders of the ENS That Involve Neurotrophic Factors

The main developmental disorder of the ENS, Hirschprung disease, has been much studied, and disruption of the GDNF/c-ret signaling pathway as a major cause of Hirschprung disease is well established. The involvement of disruption to the endothelin signaling pathway in Hirschprung disease has also been well described. The possible involvement of other neurotrophic factors in developmental disorders of the ENS, however, has been little studied. Deficiency in neurotrophin expression has been described in the transitional zone of some Hirschprung disease patients. Dysregulation of neurotrophic signaling may also be involved in the etiology of infantile pyloric stenosis, a condition in which infants have hypertrophic pyloric muscle associated with abnormal innervation.

Neurotrophic Factors and Postnatal Changes in the ENS

The mammalian ENS undergoes a number of changes after birth, during the periods of gut growth and dietary change. Although most enteric neurons differentiate during fetal development and enteric ganglia are formed before birth, neural crest-derived stem cells have been identified in the postnatal gut and a small population of neural precursors has been shown to withdraw from the cell cycle postnatally. The relative proportions of different types of enteric neuron in newborn animals have been found to be different from those in adults. The cholinergic population, in particular, is not fully established at birth, and in rodents, the numbers of neurons expressing VIP, pituitary adenylyl cyclase-activating peptide (PACAP),

and neuronal NOS (nNOS) in the myenteric plexus have been shown to increase during the first few weeks of life. The morphological properties of enteric neurons also change; in addition to an increase in cell soma size, neuronal processes must grow appropriately in parallel with their target tissues and it is likely that 'fine tuning' takes place, to fully establish enteric circuitry at a structural level and also at the neurochemical level – for example, in the expression of neurotransmitter receptors. Neurotrophic factors are likely to play a role in these changes. Evidence in support of a postnatal role for neurotrophic factors and their receptors comes from studies of their expression and actions in the postnatal gut.

Expression of Neurotrophic Factors and Their Receptors in the Postnatal Gut

The neurotrophic factors that are active in the fetal ENS – GDNF, NTN, and NT-3 – continue to be expressed in the postnatal rodent gut. Evidence from immunohistochemical studies of human tissues indicates that BDNF and LIF are also expressed in the gut during the postnatal period. Semiquantitative reverse transcription polymerase chain reaction (RT-PCR) of mRNA extracted from whole gut has been used to analyze possible changes in expression levels of GDNF and NTN and their specific GPI-linked receptors postnatally. This work indicates that there may be a decline in the level of expression of GDNF in the first few postnatal weeks, while that of NTN may increase, reaching a maximum at postnatal day 14 in mice. These changes in expression are mirrored by possible decreases in the levels of GFRα1 and increase in GFRα2 during the same time period. Interpretation of such analysis of mRNA levels in whole gut, which contains a mixture of cell types, however, does not determine if local expression levels of proteins change, and is further complicated by the rapid postnatal growth of the different gut layers. In this context it is worth noting that other studies have reported continued high levels of expression of both GDNF and NTN into adult life. Possible changes in expression levels are nevertheless potentially of importance, since data from both NTN and GFRα2 null transgenic mice indicate that NTN may play a role in the later development of the ENS, particularly of cholinergic neurons, some of which have been reported in other studies to differentiate postnatally.

The expression of other neurotrophic factors and their receptors during postnatal development has been studied using immunohistochemical techniques. In human infants, TrkA and TrkB receptors have been found to be expressed by both enteric neurons and glia, while TrkC receptors are expressed only by enteric neurons. BDNF and NT-3 are also expressed by enteric ganglion cells. Detailed analysis to localize neurotrophic factor and receptor expression to particular neuronal subpopulations, however, has not been performed during the postnatal period. It should also be noted that a different pattern of expression of TrkB has been described in adult rodent gut (see later). Nevertheless, expression of neurotrophic factors by enteric neurons in postnatal animals suggests that these factors could be involved in establishing interneuronal connections within the developing ENS circuitry.

Actions of Neurotrophic Factors on Postnatal Enteric Neurons and Glia

In vitro work supports the hypothesis that neurotrophic factors play a role in the postnatal development of the ENS. Studies of isolated myenteric ganglion cells in dissociated cell culture show that myenteric neurons from postnatal rats respond to exogenously applied GDNF, NGF, NT-3, and CNTF. GDNF and NT-3 promote neuronal survival and neurite extension, NGF promotes neuritogenesis, and CNTF has been reported to also promote neurite sprouting *in vitro*. There is also evidence that GDNF and NT-3 may have effects on the neurochemical properties of cultured myenteric neurons, possibly promoting expression of VIP. Study of the effects of GDNF on neurons at different postnatal ages indicates that older (postnatal day 14) myenteric neurons may be less responsive to GDNF than are those from younger (postnatal day 7) animals, an observation that is consistent with the molecular data on GFRα1 and GFRα2 mRNA levels in postnatal mice.

Neurotrophic Factors and the Adult ENS: Roles in Intestinal Neurophysiology and Neuroprotection?

Expression of Neurotrophic Factors and Their Receptors in the Adult Gut

Expression of neurotrophic factors and their receptors in the gut continues into adult life. GDNF and NTN mRNA expression in the adult rat gut has been confirmed by molecular techniques, and GDNF protein has also been shown by immunoassay to be present at high levels in the adult gut. Analysis of mRNA indicates that GDNF and NTN expression continues even in aged (24 months old) rats. Although relatively few studies have been performed, immunohistochemical evidence indicates that some, but not all, enteric neurons in adult humans, rats, mice, and other vertebrates express NGF, BDNF, or NT-3. Co-expression of the three neurotrophins by enteric neurons has not been examined to date. In rat intestine, NT-3-immunoreactive neurons were found to be abundant in the

submucous plexus, but sparse in the myenteric plexus, and most NT-3-immunopositive neurons also expressed the peptide VIP.

Analysis of receptor expression indicates that adult enteric neurons express c-ret, GFRα1, GFRα2, TrkA, and TrkC, while TrkB expression seems to be confined to glial cells. Recent immunohistochemical analysis in adult rat colon showed that 71% of myenteric neurons are immunopositive for TrkA, and 78% express the low-affinity neurotrophic receptor, p75. More than 60% of myenteric neurons express both receptors. The functional significance of this widespread expression of NGF receptors was demonstrated by application of exogenous NGF, which resulted in TrkA phosphorylation. Double labeling demonstrated that TrkA was expressed by different types of enteric neurons, including choline acetyltransferase-immunopositive and nNOS-immunopositive neurons. Whether or not the level or pattern of receptor expression by individual neurons changes during the life span remains to be determined, although there is evidence that responsiveness of myenteric neurons to NGF, NT-3, and GDNF is maintained throughout life.

Roles of Neurotrophic Factors in the Adult and Aging Gut

What is the role of neurotrophic factors in the mature and aging ENS? Recent evidence shows that there may be both a neuromodulatory action, affecting gut function, and a neuroprotective role, promoting neuronal survival.

Effects of neurotrophic factors on gastrointestinal motility The acute effects of BDNF in the central nervous system have been well documented. The observation that BDNF and other neurotrophins are expressed by enteric neurons therefore led to the suggestion that they may also have rapid and short-lived actions in the gut, as well as, or instead of, longer term neurotrophic effects.

Evidence that neurotrophins may have effects on intestinal physiology came indirectly, when it was noted that patients in clinical trials involving treatment with recombinant BDNF or NT-3 for neurological disorders had an increased frequency of bowel movements and diarrhea. Subsequent trials of the possible therapeutic use of NT-3, administered three times a week subcutaneously, for treatment of patients with idiopathic constipation also resulted in a dose-dependent enhancement of colonic transit and improved passage of stools. Gastric emptying and both intestinal and colonic transit were accelerated in these patients. Studies of the effects of NGF, recombinant BDNF, and recombinant NT-3 infused into the tail vein of rats showed that these factors had similar effects on colonic motility: an increase in the spike frequency, amplitude, and duration. Recombinant NT-3 (but not the other factors) also promoted myoelectrical activity in the stomach and small intestine. The mechanisms by which the neurotrophins exert these effects, however, are unclear. The effect in humans began 24–48 h after onset of treatment, so could possibly be due to a relatively rapid trophic effect, perhaps on the efficacy of particular nerve types or neurotransmitter systems. To date, analysis of the effects of these factors *in vitro* has not provided clear evidence on their mode of action *in vivo*.

Protective roles of neurotrophic factors in the ENS Recent evidence indicates that neurotrophic factors may have protective roles in the ENS during maturity and in aging. Effects of neurotrophic factors in model systems in which enteric neuronal damage was induced either by hyperglycemia or by reactive oxygen species (ROS) have been investigated.

Enteric neuropathy is a well-documented complication of diabetes, and loss of enteric neurons has been demonstrated in humans and in animal models. The causes of enteric nerve damage in diabetes have not been established, although hyperglycemia is likely to be involved. In recent work, the effects of elevated glucose levels on cultured fetal enteric neurons were examined. Glucose produced a dose-dependent increase in apoptosis in these cultures, an effect that was significantly reduced in the presence of GDNF. Importantly, *in vivo*, overexpression of GDNF also protected against apoptosis, specific neuronal loss, and motility changes seen in streptozotocin-induced diabetic transgenic mice.

GDNF and NT-3 have also been shown to reduce generation of ROS and protect against ROS-induced apoptosis *in vitro*. When isolated segments of the muscularis externa containing the myenteric plexus are incubated in the presence of dihydrorhodamine, which fluoresces when exposed to ROS, addition of GDNF and NT-3, but not NGF, reduces the fluorescent signal. Furthermore, when myenteric ganglia are incubated in the presence of the ROS generator menadione, some myenteric neurons undergo apoptosis. More do so in samples from aged animals, fed an *ad libitum* diet. Incubation in the presence of GDNF or NT-3, but not NGF, reduced the number of apoptotic neurons detected in this model. Interestingly, this protective effect was lost in the aged *ad libitum*-fed animals, but not in rats fed a calorically restricted diet.

Are Neurotrophic Factors Involved in Clinical Conditions Affecting the ENS?

The continued expression of neurotrophic factors in the adult gut and their receptors by enteric neurons

also leads to the question of whether disruption of neurotrophic factor signaling is involved in clinical conditions that affect the ENS. Enteric neurons innervate all cell types in the gut, so neurotrophic factors may directly or indirectly influence all types of intestinal pathophysiology. One area in which neurotrophins have been implicated is that of the inflammatory bowel diseases, where, in addition to the inflammatory damage to nonneuronal gut tissues, abnormalities of the ENS have been described.

In experimental colitis, pretreatment with neutralizing antibodies to NGF or NT-3 resulted in exacerbation of the inflammatory response, implying that these factors normally exert a protective effect in this model of inflammatory disease. In humans with either Crohn's disease or ulcerative colitis, increased expression of NGF and NT-3 was seen, at both the mRNA and the protein levels. Selective losses of NGF-sensitive enteric neurons in experimental colitis, however, have also been described, and intestinal mast cells have also been shown to express both NGF and TrkA. Enteric neurons have also been shown to cause mast cell degranulation. Neuroimmune interactions in the gut are therefore complex, and although neurotrophic factors are likely to play a role in inflammatory conditions, much remains to be learned about this and other aspects of their actions in the gut.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Nervous System: Neuroanatomy; Enteric Nervous System Development; Enteric Nervous System: Physiology; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3.

Further Reading

Anitha M, Gondha C, Sutliff R, et al. (2006) GDNF rescues hyperglycemia-induced diabetic enteric neuropathy through activation of the PI3K/Akt pathway. *The Journal of Clinical Investigation* 116: 344–356.

Bolon B, Jing S, Asuncion F, et al. (2004) The candidate neuroprotective agent artemin induces autonomic dysplasia without preventing peripheral nerve dysfunction. *Toxicologic Pathology* 32: 275–294.

Chalazonitis A (1998) Promotion of the development of enteric neurons and glia by neuropoietic cytokines: Interactions with neurotrophin-3. *Developmental Biology* 198: 343–365.

Chalazonitis A (2004) Neurotrophin-3 in the development of the enteric nervous system. *Progress in Brain Research* 146: 243–63.

Chai N-L, Li Z-F, Wang JH, et al. (2003) Effects of neurotrophins on gastrointestinal myoelectric activities. *World Journal of Gastroenterology* 9: 1874–1877.

Coulie B, Szarka LA, Camilleri M, et al. (2000) Recombinant human neurotrophic factors accelerate colonic transit and relieve constipation in humans. *Gastroenterology* 119: 41–50.

Hoehner JC, Wester T, Pahlman S, et al. (1996) Localization of neurotrophins and their high-affinity receptors during human enteric nervous system development. *Gastroenterology* 110: 756–767.

Lin A, Lourenssen S, Stanzel RDP, et al. (2005) Nerve growth factor sensitivity is broadly distributed among myenteric neurons of the rat colon. *Journal of Comparative Neurology* 490: 194–206.

Newgreen D and Young HM (2002) Enteric nervous system: Development and developmental disturbances – part 1. *Pediatric and Developmental Pathology* 5: 224–247.

Parkman HP, Rao SSC, Reynolds JC, et al. (2003) Neurotrophin-3 improves functional constipation. *American Journal of Gastroenterology* 98: 1338–1374.

Sandgren K, Lin Z, Svenningsen AF, et al. (2003) Vasoactive intestinal peptides and nitric oxide promote survival of adult rat myenteric neurons in culture. *Journal of Neuroscience Research* 72: 595–602.

Schafer K-H, Saffrey MJ, and Burnstock G (1995) Trophic actions of 2-chloroadenosine and bFGF on cultured myenteric neurones. *NeuroReport* 6: 937–941.

Thrasivoulou C, Soubeyre V, Ridha H, et al. (2006) Reactive oxygen species, dietary restriction and neurotrophic factors in age-related loss of myenteric neurons. *Aging Cell* 5(3): 247–257.

Vasina V, Barbara G, Talamonti L, et al. (2006) Enteric neuroplasticity evoked by inflammation. *Autonomic Neuroscience: Basic and Clinical* 126–127: 264–272.

VonBoyen GBT, Reinshagen M, Steinkamp M, et al. (2002) Enteric nervous plasticity and development: Dependence on neurotrophic factors. *Journal of Gastroenterology* 37: 583–588.

Enteric Nervous System: Physiology

J D Wood, The Ohio State University, Columbus, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The enteric nervous system (ENS) is a subdivision of the autonomic nervous system (Figure 1). Neural control of the digestive tract is by the autonomic nervous system. Three subdivisions of the autonomic nervous system, known as the sympathetic, parasympathetic, and enteric divisions, make up the autonomic innervation of the alimentary tract. The enteric division, which is also called the ENS, functions like a brain in the gut. It is an independent integrative nervous system which has its neural networks placed in close apposition to the effector systems it controls in the stomach, in the approximately 30 ft of small intestine and in the 8 ft of large intestine as well as the esophagus, gallbladder, and pancreas. The effector systems are the musculature, secretory glands, and blood vasculature. There are approximately 100 million neurons in the ENS, which is about the same number as in the spinal cord. Rather than packing the 100 million neurons required for control of gut functions into the skull as part of the brain and relying on signal transmission over long, unreliable pathways to the gut, evolutionary natural selection distributed the integrative neural networks of the ENS along the gut in immediate proximity to the effector systems.

In a manner similar to the central nervous system, the networks at the effector sites have evolved as an organized array of different kinds of neurons interconnected by chemical synapses. Function in the networks is determined by generation of action potentials within single neurons and chemical transmission of information at synaptic points of contact between individual neuronal elements. Coded information is transmitted from one region of the neuron to another by action potentials. The action potential codes are transformed into chemical codes at synapses with other neural elements within a given network. Action potential codes arriving in presynaptic neuronal terminals are transformed to chemical signals for transmission across the synapse to receptors on postsynaptic neural elements where reconversion into action potential codes occurs in the postsynaptic neurons. Postsynaptic neurons integrate large numbers of synaptic inputs, which may be excitatory or inhibitory. Postsynaptic integration is fundamental for computation and processing of neural information.

The enteric microcircuits in the various specialized compartments of the digestive tract are wired with large numbers of neurons and synaptic sites where information processing occurs. Multisite computation generates output behavior from the integrated circuits that could not be predicted from properties of their individual neurons and synapses. Emergence of complex behaviors is a fundamental property of the neural networks of the ENS, in the same way as in the brain and spinal cord.

In much the same manner as in the brain and spinal cord, the enteric microcircuitry processes information and generates patterns of signals that determine the timing and sequencing of behavior of the muscle groups and secretory glands and the distribution and dynamics of blood flow within the gut wall. Processing of information in the circuits is a function whereby computation of input signals results in output to the effectors that is a meaningful transform of the input. Like electronic calculators, which receive numbers as input and calculate outputs according to specific equations preprogrammed into the circuits, the neural circuits receive sensory signals and compute outputs that control behavior of the effectors in ways adapted for overall performance of the specialized regions of gut.

Processing of sensory signals is among the major functions of the microcircuits in the ENS. The sensory signals are generated by sensory nerve endings and coded in the form of action potentials. The code may represent the status of an effector system (e.g., tension in a muscle) or may signal a change in parameters, such as acidity and osmolarity, in the environment of the lumen of a specialized compartment. Sensory signals are deciphered by the neural networks to generate output signals that initiate homeostatic adjustments in the behavior of each individual effector system and the integrated behavior of the whole organ.

Myenteric and Submucosal Plexuses

The ENS consists of ganglia, primary interganglionic fiber tracts, and secondary and tertiary fiber projections to the effector systems (i.e., musculature, glands, and blood vessels). Two ganglionated plexuses are the most obvious structures of the ENS (Figure 2). The myenteric plexus, also known as Auerbach's plexus, is located between the longitudinal and circular muscle layers of the esophagus, stomach, and small and large intestine. The submucosal plexus, also known as Meissner's plexus, is situated in the submucosal region between the circular muscle and

mucosa. The submucosal plexus is most prominent as a ganglionated network in the small and large intestine. It does not exist as a ganglionated plexus in the esophagus and is sparse in the submucosal space of the stomach. In larger mammals (e.g., humans), the intestinal submucosal plexus consists of an inner submucosal network (Meissner's plexus) located at the serosal side of the muscularis mucosae and an outer plexus (Schabadasch's plexus) adjacent to the luminal side of the circular muscle coat. In human small and large intestine, a third intermediate plexus lies between Meissner's and Schabadasch's plexus.

Most motor innervation of the intestinal circular and longitudinal muscle coats originates from motor neurons that have their cell bodies in the myenteric plexus. Motor innervation of the intestinal secretory glands (i.e., Brunner's glands in the duodenum and crypts of Lieberkühn in both small and large intestines) originates in the submucosal plexus. Neurons in submucosal ganglia send fibers to the myenteric plexus and also receive synaptic input from axons projecting from the myenteric plexus. The interconnections link the two networks into a functionally integrated nervous system.

Structure, function, and neurochemistry of ganglia in the ENS are reminiscent of those in the central nervous system but differ considerable from ganglia of the sympathetic and parasympathetic autonomic divisions. Unlike most sympathetic or parasympathetic autonomic ganglia, where function is mainly as relay-distribution centers for signals transmitted from the brain and spinal cord, enteric ganglia are interconnected to form a nervous system with mechanisms for integration and processing of information like those found in the brain and spinal cord. This is the justification for referring to the ENS as the brain in the gut.

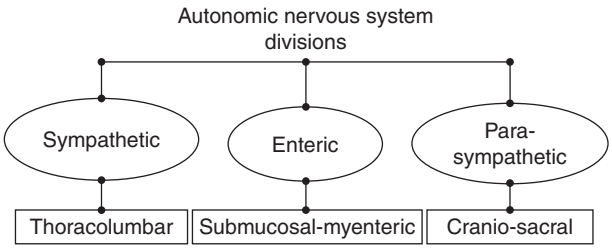

Figure 1 The digestive tract is innervated by three divisions of the autonomic nervous system: the sympathetic, parasympathetic, and enteric divisions. The parasympathetic division is subdivided into cranial and sacral divisions based on the anatomical location of the preganglionic cell bodies in the brain stem and spinal cord. Preganglionic cell bodies of the sympathetic division are in the thoracolumbar regions of the spinal cord. Cell bodies of the enteric division are in its myenteric and submucosal divisions.

Heuristic Model for the ENS

The heuristic model for the ENS is the same as for the brain and spinal cord and, for that matter, any other independent integrative nervous systems, such as those in invertebrate animals (Figure 3). As in any independent integrative nervous system, sensory neurons, interneurons, and motor neurons in the ENS are connected synaptically for flow of information from sensory neurons to interneuronal integrative networks to motor neurons to effector systems. The job of the ENS is to organize and coordinate the activity of each effector system into meaningful behavior of the integrated organ. It is integrated with sympathetic and parasympathetic autonomic centers in the central

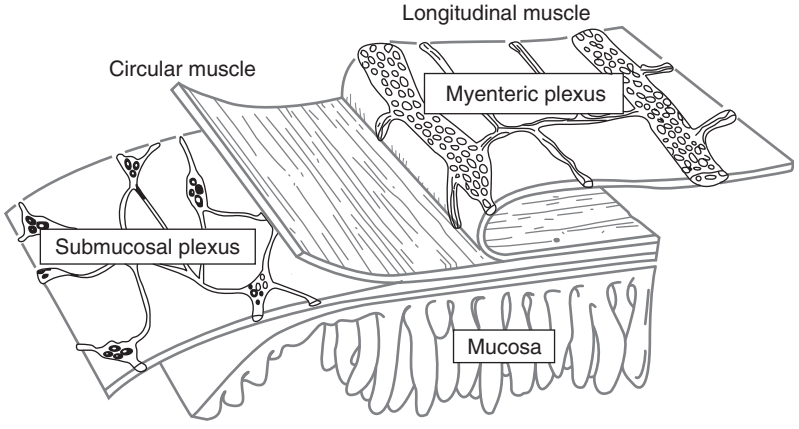

Figure 2 Structural relations of the intestinal musculature and the ENS. Longitudinal and circular muscle coats are the main components of the musculature involved in propulsive motility; secretory glands are in the mucosa. Ganglia and interganglionic fiber tracts of the myenteric and submucosal plexuses are main constituents of the ENS. The myenteric plexus is situated between the circular and longitudinal muscle coats; the submucous plexus is between the mucosa and circular muscle coat.

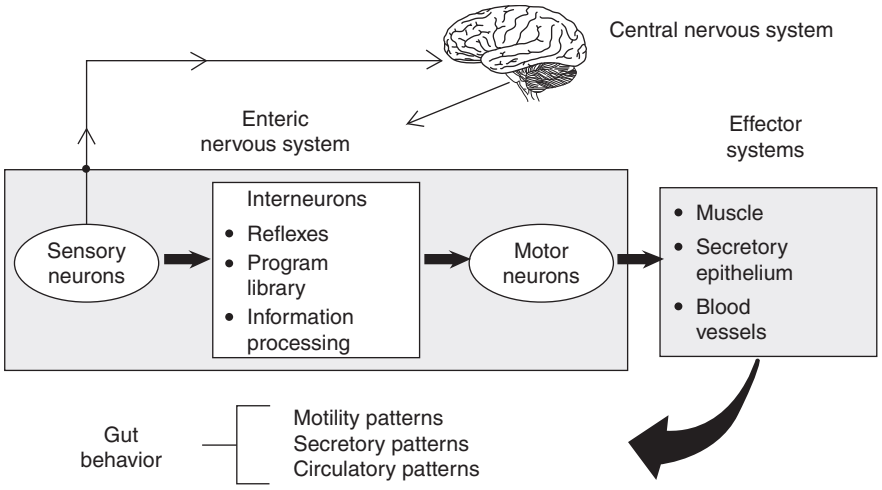

Figure 3 Sensory neurons, interneurons, and motor neurons are synaptically interconnected to form the integrated microcircuits of the ENS. Like the central nervous system, sensory neurons, interneurons, and motor neurons are connected synaptically for flow of information from sensory neurons to interneuronal integrative networks to motor neurons to effector systems. The ENS organizes and coordinates the activity of each effector system into meaningful behavior of the whole organ. Bidirectional communication occurs between the central and the ENS.

nervous system and bidirectional communication occurs between the centers in the central nervous system and the ENS.

Neuronal Types

Dogiel types I and II are the principal morphological classes of neurons in the ENS. The German neuroanatomist, AS Dogiel, described these two morphological types of enteric ganglion cells that were named after him. Dogiel types I and II are distributed in both myenteric and submucosal plexuses. Both types of neurons are distributed in a two-dimensional plane (i.e., the cell bodies are not stacked one on the other). Cell bodies are flat, coin-like disks with multiple neurites exiting around the perimeter, which may adapt the ENS for the changing forces within the gut wall experienced during contraction of the musculature or stretching of the wall as the hollow organs fill.

Dogiel type I neurons have cell bodies with many short processes and a single long process (Figure 4). These are flat neurons with the processes extending from the cell body in the circumferential and longitudinal planes of the wall. The short processes are dendrites, which receive synaptic input; the long process is an axon. The axon of Dogiel type I neurons projects for relatively long distances through interganglionic fiber tracts and many rows of ganglia. Still, the projections are not long in absolute length; the longest known projections of any enteric ganglion cell are only approximately 2–3 cm. Dogiel I neurons projecting in a specific direction often express a specific neurotransmitter. Aborally projecting Dogiel I axons release nitric oxide and vasoactive intestinal polypeptide as neurotransmitters; those projecting in the oral direction release substance P and acetylcholine. Some of the Dogiel type I neurons are motor neurons to the musculature and secretory epithelium; others are interneurons.

The cell bodies of ENS neurons with Dogiel type II morphology have smooth surfaces with long and short processes arising in a variety of configurations (Figure 4). The long processes may extend through interganglionic fiber tracts across several rows of ganglia in the circumferential direction, the oral direction, or aborally. Shorter processes may only project within the home ganglion. Almost all Dogiel II neurons in the myenteric plexus project a process to the submucosa/mucosa. Dogiel II processes in the mucosal regions express the 5-HT_3 serotonergic receptor subtype and are fired by serotonin that is released from mucosal enterochromaffin cells during mechanical and chemical stimulation of the mucosa. Dogiel II neurons express the neurotransmitters substance P and acetylcholine.

A population of Dogiel II neurons, which electrophysiologically are characterized by long-lasting afterhyperpolarization following generation of an action potential, have been proposed to be intrinsic primary afferent neurons referred to by the acronym IPAN. Nevertheless, these neurons behave quite differently from spinal or vagal primary sensory afferent neurons, which faithfully encode changes in stimulus energy into the firing of action potentials. Instead, variability in excitability of enteric afterhyperpolarization (AH)-type neurons and their synaptic interconnections

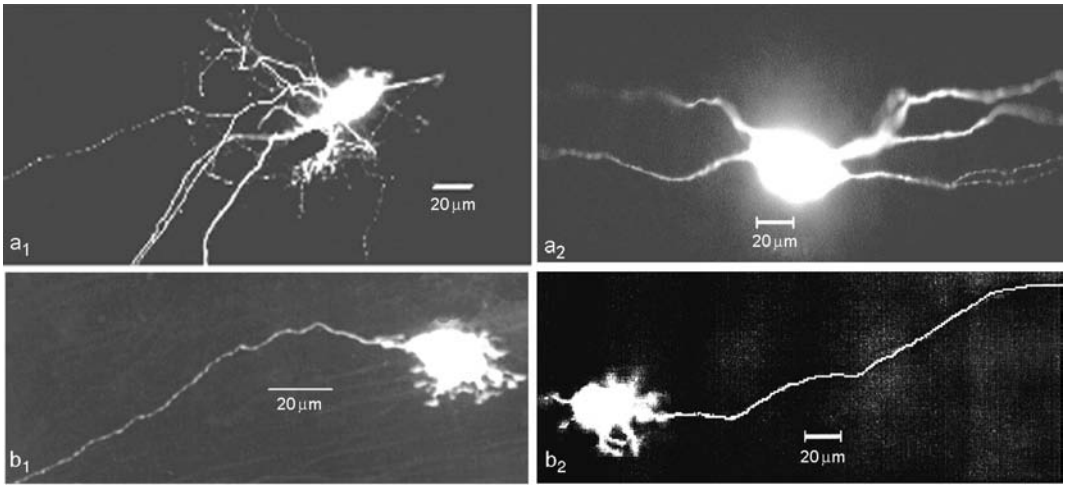

Figure 4 Histoanatomy of Dogiel morphologic types I and II in the ENS. (a_1, a_2) Dogiel morphologic type II neurons in the myenteric plexus of the guinea pig small intestine. Dogiel II neurons have multiple long processes with variable lengths and shapes that ramify throughout the ganglion and into interganglionic fiber tracts. (b_1, b_2) Dogiel morphologic type I neurons in the submucosal plexus of the small intestine of the guinea pig. Dogiel I neurons have multiple short dendrites at the edges of the cell body and a single axon. The neurons were marked by the injection of biocytin from micropipettes that also served as recording microelectrodes.

suggest that they may be multifunctional interneurons and play an important role in determining the output of, as well as the input to, enteric reflex circuits.

Electrical Behavior

Intracellular microelectrodes are used to record the electrical and synaptic behavior of enteric neurons in preparations removed from the various organs of the digestive tract. Differences in electrical and synaptic behavior provide a basis for classification of enteric neurons into two main subgroups termed AH type and submucosal (S) type.

S-type neurons were first identified on the basis of nearly universal occurrence of fast excitatory postsynaptic potentials (Figure 5) and AH-type neurons were identified based on the occurrence of long-lasting AH associated with the action potential (Figure 6). AH-type electrophysiological behavior is found mainly in enteric neurons with Dogiel type II morphology and S-type behavior is found mainly in Dogiel type I neurons.

Repetitive action potential discharge can occur in S-type neurons; repetitive discharge generally does not occur in AH-type enteric neurons (Figure 6). Membrane excitability is higher in S-type enteric neurons than in AH-type neurons. S-type neurons fire repetitively during the intraneuronal injection of long-duration depolarizing current pulses through the recording microelectrode. The frequency of the repetitive spike discharge increases in proportion to the amplitude of the depolarization produced by the injected pulses. AH-type neurons do not fire repetitively during sustained membrane depolarization when they are in a resting state. Irrespective of the strength of a depolarizing pulse, these neurons fire only one or two spikes at the onset of the pulse. During slow synaptic excitation (Figure 5), AH neurons become hyperexcitable and fire repetitively like S-type neurons.

In the guinea pig, which is the most commonly studied model, the occurrence of AH-type neurons is limited to the gastric antrum and small and large intestine. S-type neurons are found throughout the digestive tract, including gastric corpus and antrum, gallbladder, esophagus, and pancreas.

Neurotransmission

Synaptic events in the ENS are generally the same as in the brain and spinal cord. Excitatory postsynaptic potentials, inhibitory postsynaptic potentials, and presynaptic inhibition and facilitation are the principal synaptic events in the enteric microcircuitry. Both slow and fast synaptic mechanisms are expressed in neurons with AH-type electrophysiological behavior and Dogiel II morphology and S-type electrophysiological behavior and uniaxonal Dogiel I morphology. Fast synaptic potentials have durations in the millisecond range; slow synaptic potentials last for several seconds or minutes. Fast synaptic potentials are usually excitatory postsynaptic potentials. The slow synaptic events may be either excitatory postsynaptic potentials or inhibitory postsynaptic potentials.

Fast Excitatory Postsynaptic Potentials

Fast excitatory postsynaptic potentials are membrane depolarizations with durations less than 50 ms

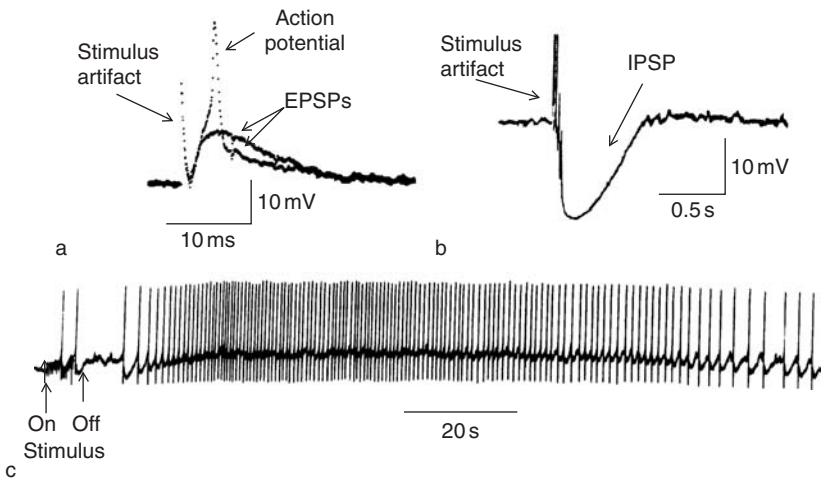

Figure 5 Fast and slow excitatory postsynaptic potentials (EPSPs) and slow inhibitory postsynaptic potentials (IPSPs) are the primary kinds of synaptic potentials in enteric neurons. (a) Two fast EPSPs were evoked by successive stimuli and are shown as superimposed records. Only one of the EPSPs reached threshold for discharge of an action potential. The time course of the EPSPs is in the millisecond range. The fast EPSPs were evoked by single electrical shocks applied to the axon that formed the synapse with the recorded neuron. (b) Slow IPSP evoked by stimulation of an inhibitory input to the neuron. This is a hyperpolarizing synaptic potential that will suppress excitability (i.e., decrease probability of action potential discharge) compared with enhanced excitability during slow EPSPs. (c) Slow EPSP evoked by repetitive electrical stimulation of the synaptic input to the neuron. Slowly activating depolarization of the membrane potential continues for more than 2 min after termination of the stimulus. Repetitive discharge of action potentials reflects enhanced neuronal excitability during the slow EPSP.

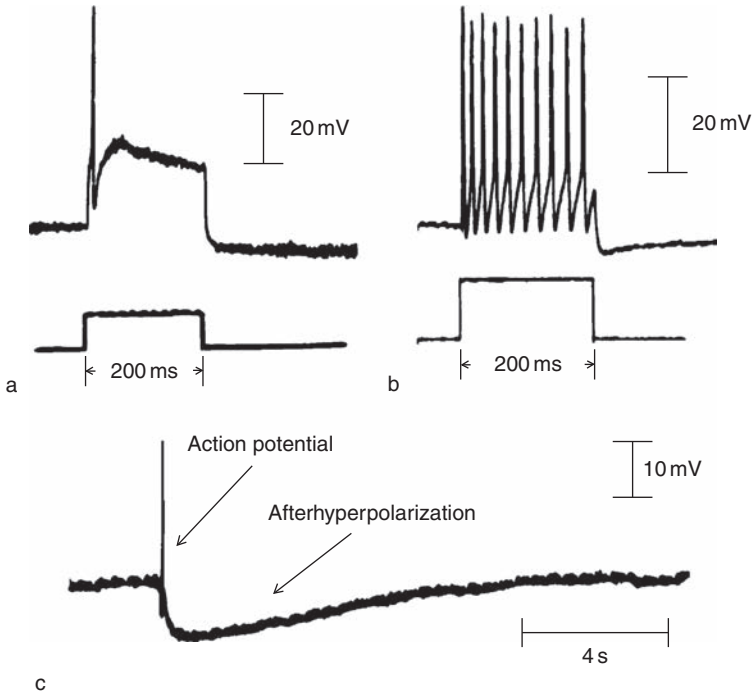

Figure 6 Enteric neurons are classified as S type 1 or AH type 2 based on their electrophysiological behavior. (a) AH type 2 neuron had low excitability reflected by discharge of only a single action potential during intraneuronal injection of a 200 ms depolarizing current pulse. (b) S type 1 neuron had relatively high excitability reflected by repetitive discharge of action potentials during intraneuronal injection of a 200 ms depolarizing current pulse. The bottom traces in (a) and (b) show the current pulses and the top traces show the neuronal responses to the depolarizing current pulses. (c) AH type 2 neurons are characterized by long-lasting membrane hyperpolarization (i.e., afterhyperpolarization) following the discharge of an action potential.

(Figure 5). They are found at synapses in both myenteric and submucosal plexuses, where they appear most prominent in neurons with S-type electrophysiological behavior in all of the specialized organs of the digestive tract. Most of the fast excitatory postsynaptic potentials are mediated by acetylcholine at nicotinic postsynaptic receptors. The actions of serotonin at the serotonergic 5-HT_3 receptor subtype and purine nucleotides at P_{2X} purinergic receptors behave much like fast excitatory postsynaptic potentials.

Slow Excitatory Postsynaptic Potentials

Slow excitatory postsynaptic potentials are involved in neurotransmission in both myenteric and submucosal plexuses and in both AH- and S-type neurons. They are most pronounced in Dogiel morphologic type II neurons with AH-type electrophysiologic behavior where conversion from hypo- to hyperexcitability is involved. Neurons with slow excitatory synaptic inputs are found in the small and large intestine and gastric antrum but not the gastric corpus or gallbladder. They seem to be associated with specialized regions in which propulsive motility is a significant function.

Slowly activating membrane depolarization continuing for several seconds to minutes after termination of release of the neurotransmitter from the presynaptic terminal identifies slow excitatory postsynaptic potentials (Figure 5). Enhanced excitability, reflected by long-lasting trains of impulses, is the hallmark of slow excitatory postsynaptic neurotransmission. Enhanced excitability is apparent experimentally as repetitive spike discharge during depolarizing current pulses. AH-type neurons, which fire only a single spike at the beginning of a depolarizing current pulse in the inactivated state, will fire repetitively in response to depolarizing pulses when the slow excitatory postsynaptic potential is in effect. When activated by slow synaptic inputs, behavior of AH-type neurons is much like that of S-type neurons and may be confused as such if they happen to be in an activated state due to ongoing release of the transmitter or a paracrine mediator that mimics slow synaptic excitation. Postspike hyperpolarization in AH-type neurons is suppressed during slow excitatory neurotransmission. Suppression of the AH is part of the mechanism that permits repetitive spike discharge at increased frequencies during the enhanced state of excitability.

Slow excitatory postsynaptic potentials are a mechanism for long-lasting activation or inhibition of gastrointestinal effector systems. The prolonged neuronal firing during a slow excitatory postsynaptic potential drives the release of neurotransmitter from the neuron's axon for the duration of the spike discharge.

Table 1 Mimetics of slow excitatory postsynaptic potentials

Mimetic (receptor)	*Mimetic (receptor)*
Acetylcholine (muscarinic M_1)	Vasoactive intestinal peptide
Cholecystokinin (A)	Caerulein
Thyrotropin-releasing hormone	Gastrin-releasing peptide
Calcitonin gene-related peptide	Tachykinins (NK_1/NK_3)
Bombesin	γ-Aminobutyric acid
5-Hydroxytryptamine (5-HT_{1P}/5-HT_2)	Motilin
Interleukin-1β	Glutamate
Histamine (H_2)	Interleukin-6
Adenosine (A_2)	Platelet-activating factor
Corticotropin-releasing factor (CRF_1)	Bradykinin (B_2)
ATP ($P2Y_1$)	Proteases ($PARS_{1,\ 2,\ \text{and}\ 4}$)
Pituitary acenylate cyclase Activating peptide	Prostaglandin E
	Leukotrienes C_4, D_4, and E_4

Prolonged inhibition, or excitation at neuronal synapses in the processing circuits and at neuroeffector junctions, is the functional outcome of slow synaptic excitation. This governs the functional behavior of the effector systems. Compared to twitches of skeletal muscles, contractile responses of the gut musculature and glandular secretion are sluggish events that last for several seconds from start to completion. The prolonged trainlike firing during slow excitatory postsynaptic potentials is the neural correlate of long-lasting excitatory or inhibitory responses of the muscle groups in the functioning gastrointestinal tract. Prolonged secretory responses in the intestinal crypts are also related to the sustained firing during slow synaptic excitation.

Several messenger substances found in neurons or endocrine or immune cells of the gut mimic slow synaptic excitation when applied experimentally to enteric neurons (Table 1). Receptors for more than one of the messenger substances may be present on the same neuron.

Slow Inhibitory Postsynaptic Potentials

Slow inhibitory postsynaptic potentials are hyperpolarizing synaptic potentials found in neurons in both myenteric and submucosal plexuses of the small and large intestine and in myenteric neurons of the gastric antrum. Slow inhibitory postsynaptic potentials activate relatively slowly and continue for several seconds after termination of the stimulation (Figure 5). The characteristics of slow synaptic inhibition are the inverse of slow synaptic excitation in that hyperpolarization instead of depolarization of the membrane potential occurs (e.g., a change from -50 to -75 mV). Hyperpolarization of the membrane potential reduces excitability and the

Table 2 Mimetics of slow inhibitory postsynaptic potentials

Mimetic (receptor)	*Mimetic (receptor)*
Acetylcholine	Opioid peptides (μ, δ)
5-Hydroxytryptamine (5-HT_{1A})	Norepinephrine (α_2)
Neurotensin	Cholecystokinin
Somatostatin	ATP
Adenosine (A_1)	Galanin (GAL-1)
Nociceptin	

probability that the neuron will fire impulses. Inhibitory postsynaptic potentials are more readily demonstrated in the submucous than in the myenteric plexus of animal preparations.

Several putative neurotransmitters or paracrine mediators evoke responses similar to slow inhibitory postsynaptic potentials when experimentally applied to enteric neurons (Table 2). Some of these substances are peptides, others are purine compounds, and another is norepinephrine. Receptors for two or more of these substances may be localized to the cell body of the same neuron.

Peristaltic Reflex

The peristaltic reflex is a polysynaptic reflex circuit in the ENS that underlies all of the propulsive motility patterns found in the small and large intestine and esophagus. It is the intestinal analog of spinal motor reflexes (e.g., monosynaptic patellar and Achilles tendon reflexes and polysynaptic withdrawal reflexes). The peristaltic reflex is much the same in that it is a fixed response evoked by investigational stretching of the intestinal wall or stroking of the mucosa. It is like a polysynaptic spinal reflex in that it is a motor response to sensory stimulation that is repeated the same way each time the hardwired reflex circuit is activated. The peristaltic reflex circuit is synaptically connected to evoke relaxation of the circumferentially oriented muscle layer and contraction of the longitudinal muscle below the point of stimulation and contraction of the circumferentially oriented muscle layer above the point of stimulation.

Like spinal reflexes, the peristaltic reflex is positioned at the lowest level of the hierarchical organization of neural control of intestinal motility and undoubtedly underlies each of the various patterns of propulsive motility that impart functionality to the intestine during daily life. As with a spinal motor reflex, the sequencing of the pattern of behavior of the intestinal longitudinal and circular muscles is hardwired into the circuitry, whereas the strength of each motor component of the pattern and the repetition rate of the pattern are adjusted by sensory feedback or other commands to automatically compensate for local loads and higher functional demands on the intestine as a whole. The distance over which propulsion occurs and the direction in which it occurs in the specific patterns of motility, which characterize the various digestive states, are additional factors requiring a higher order of neural control. Short-distance propulsion in the postprandial digestive state, propulsion over intermediate distances during interdigestive motility (i.e., migrating motor complex) and long-distance power propulsion, all in the orthograde direction, and long-distance retropulsion during emesis are neural control requirements that are met by the integrative microcircuitry of the ENS.

The muscle layers of the intestine contract and relax in a stereotypic pattern during peristaltic propulsion (Figure 7). This pattern is determined by the sequence in which the peristaltic polysynaptic reflex circuit activates excitatory and inhibitory motor neurons to the longitudinal and circular muscle layers. During propulsion, the longitudinal muscle layer in the segment ahead of the advancing intraluminal contents contracts in response to activation of its excitatory motor innervation while at the same time the circular muscle layer relaxes in response to activation of its inhibitory motor innervation. The intestinal tube behaves geometrically like a cylinder with constant surface area. Shortening of the longitudinal axis of the cylinder during contraction of the longitudinal muscle is accompanied by a widening of the cross-sectional diameter. The simultaneous shortening of the longitudinal axis and relaxation of the circular muscle results in expansion of the lumen, which prepares a receiving segment (Figure 7) for the forward-moving intraluminal contents during peristaltic propulsion.

Organization of a receiving segment constitutes one-half of propulsive peristaltic reflex behavior. The second half is contraction of the circular muscle in the segment behind the advancing intraluminal contents. The longitudinal muscle layer in this segment relaxes at the same time that the circular muscle contracts, resulting in the conversion of this region to a propulsive segment that propels the luminal contents ahead into the receiving segment (Figure 7). The propulsive segment is formed when neural connections in the reflex circuit shut off the excitatory innervation to the longitudinal muscle and the inhibitory innervation to the circular muscle while activating excitatory motor neurons to the circular muscle.

The heuristic model for peristaltic intestinal propulsion assumes blocks of the basic polysynaptic reflex circuit connected in series along the length of small intestine and large intestine (Figure 8). A block of the basic polysynaptic circuit is formed by synaptic

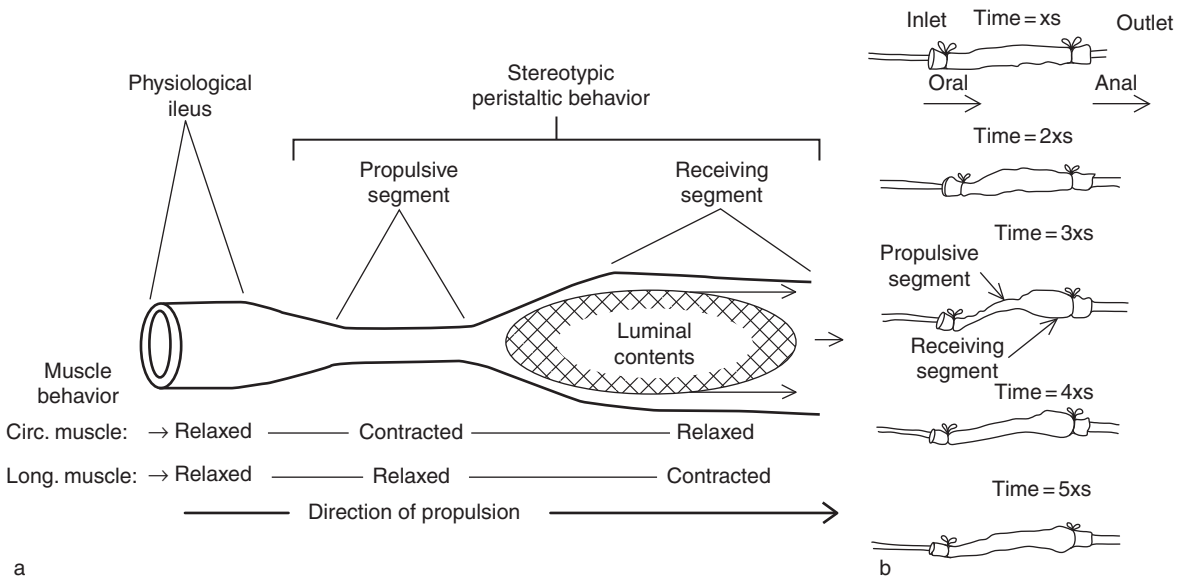

Figure 7 Intestinal peristaltic propulsive motility. (a) The circumferential and longitudinal muscle layers of the intestine behave in a stereotypical pattern during peristaltic propulsion. A hardwired reflex circuit in the ENS determines the pattern of behavior of the two muscle layers. During peristaltic propulsion, the longitudinal muscle layer in the segment ahead of the advancing intraluminal contents contracts while the circumferential muscle layer relaxes simultaneously. Simultaneous shortening of the longitudinal intestinal axis and relaxation of the circumferential muscle in the same segment results in expansion of the lumen, which becomes a receiving segment for the forward-moving contents. The second component of the reflex is contraction of the circular muscle in the segment behind the advancing intraluminal contents. The longitudinal muscle layer in the same segment relaxes simultaneously with contraction of the circular muscle, which results in conversion of this region to a propulsive segment that propels the luminal contents ahead into the receiving segment. (b) Motor behavior of the intestinal wall during the peristaltic reflex in a segment of guinea pig ileum in response to distension by infusion of saline at time = xs. Shortening of the longitudinal axis and expansion of the lumen is seen in the receiving segment and contraction of the circular muscle and reduction in the diameter of the receiving segment is apparent at time = 3xs as propulsion proceeds to empty the lumen through the outlet. Wall behavior was redrawn from photographic images provided by Prof. M Takaki, Department of Physiology, Nara Medical University, Japan.

connections between interneurons and motor neurons (Figure 8). Propulsion occurs over extended lengths of intestine as blocks of the basic circuit are recruited to activity in consecutive segments. In this respect, the intestine is like the spinal cord, in which connections for polysynaptic reflexes remain irrespective of the destruction of adjacent regions of the spinal cord. Resection of an intestinal segment does not alter the reflex circuitry in the two segments remaining on either side of the resection. Consequently, organized propulsion is not impaired after resection of various lengths of bowel.

Central Pattern Generation

Independent integrative nervous systems in vertebrates and invertebrates contain neural networks that are programmed to generate timing and phasing signals for repetitive, rhythmic movements. Walking, swimming, flying, and copulation are examples of repetitive movements controlled by neural networks that have central pattern generators. In the small intestine, repetitive activation of peristalsis within the activity front of the migrating motor complex during the fasting state, mixing movements occurring rhythmically at multiple sites along the small intestine during the digestive state, and rhythmic cycles of secretion from the mucosa in response to food allergens all reflect rhythmic output from pattern generators in neural networks of the ENS.

Earlier concepts of neural networks taught that each kind of somatomotor behavior generated by an integrated muscle group required a different hardwired circuit for each specific behavior. Current concepts hold that a neural network can be polymorphic with the same circuit being capable of generating a variety of different motor behavioral patterns. Circuits of this nature are present in the ENS and can be described as anatomical networks, which hold a library of components that can be reconfigured in different ways to form a variety of different functional circuits. Overlays of neuromodulatory signals select the configuration and thereby determine the output of the network. This is accomplished through selective release of neuromodulatory chemicals that blanket the circuit and act to alter the electrical and

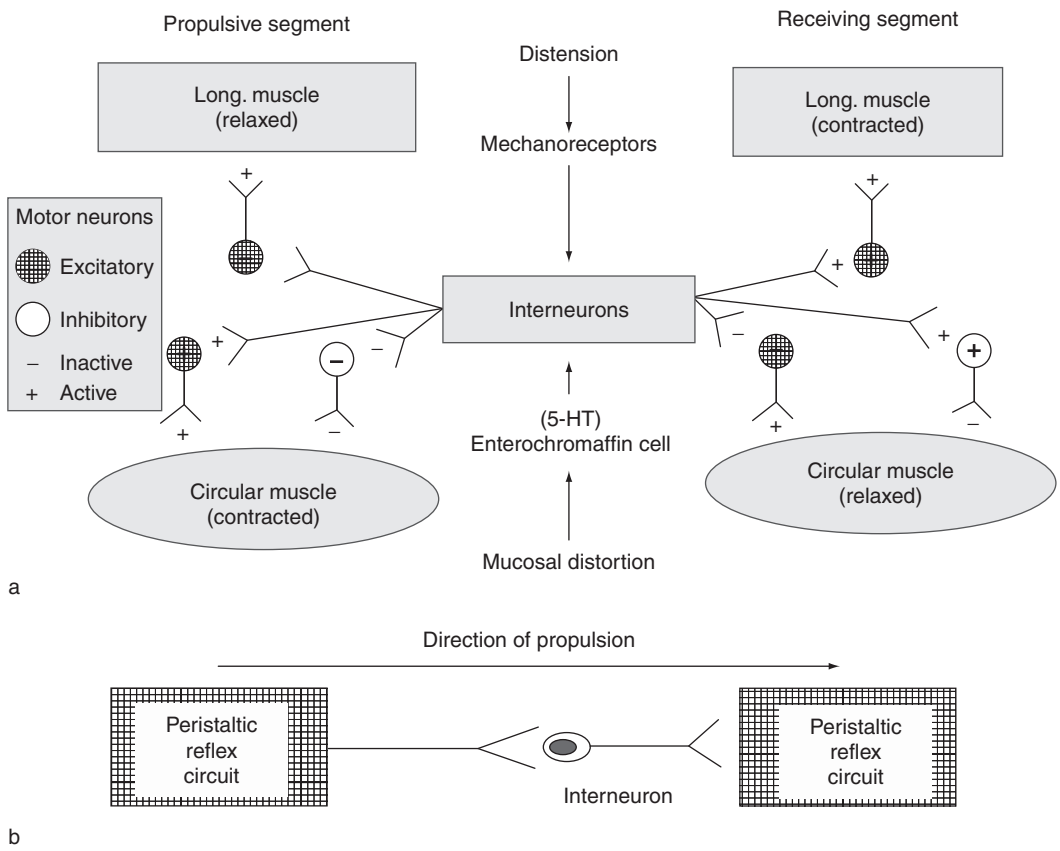

Figure 8 A hardwired polysynaptic peristaltic reflex circuit in the ENS underlies intestinal propulsive motility. (a) Two kinds of input activate the basic peristaltic reflex circuit. One is distension of the intestinal wall and activation of intramural mechanoreceptors. Second is the release of 5-hydroxytryptamine (5-HT) from enterochromaffin cells by mechanical stimulation of the mucosa. When the reflex circuit is functioning, excitatory motor neurons to the longitudinal muscle coat and inhibitory motor neurons to the circular muscle coat are activated to form the receiving segment below the point of stimulation (see Figure 7). At the same time, activation of excitatory motor neurons to the circular muscle coat and inactivation of inhibitory motor neurons to the circular muscle coat occur in the propulsive segment above the point of stimulation. (b) Blocks of the basic polysynaptic peristaltic reflex circuit are continuous along the intestine. Propulsion over extended distances is accomplished when synaptically connected blocks of circuitry are sequentially activated in a specific direction. Activation of sequential blocks generally travels in the aboral direction in the normally functioning bowel; it travels in the oral direction in the upper small intestine during emesis.

synaptic behavior of the neural elements that make up the circuit. Reconfiguration of ENS microcircuits and the resulting output to the effector systems occurs in response to neuromodulatory messenger substances released in paracrine manner from endocrine cells and immune/inflammatory cells within the intestinal wall.

The ENS contains a library of programs for specific patterns of intestinal motor behavior (Figure 9). Each program, when running, adapts the bowel for a specific digestive state or, in some cases, for adverse condition in the intestinal lumen (e.g., inflammation, infectious agents, or food allergens). One or the other program may be called up by endocrine or paracrine release of a neuromodulator that floods over the circuit or by neuromodulatory inputs from the central nervous system.

Neuroimmune Modulation

Histamine release from enteric mast cells in response to a food allergen and the resulting overlay of histamine on the ENS is an example of how a neuromodulator reconfigures the network to generate a rhythmic pattern of coordinated secretion and muscle contraction. Experimental application of histamine, which simulates degranulation of mast cells and mimics antigen-evoked degranulation of mast cells in the intestine of sensitized animal models, evokes cyclical bursts of glandular secretion with each of the secretory bursts linked to contraction of the musculature. The secretory behavior is activated for approximately 1 min and then inactivated for a few minutes prior to the next cycle (Figure 10). As each secretory cycle peaks, contraction of the intestinal muscularis externa occurs.

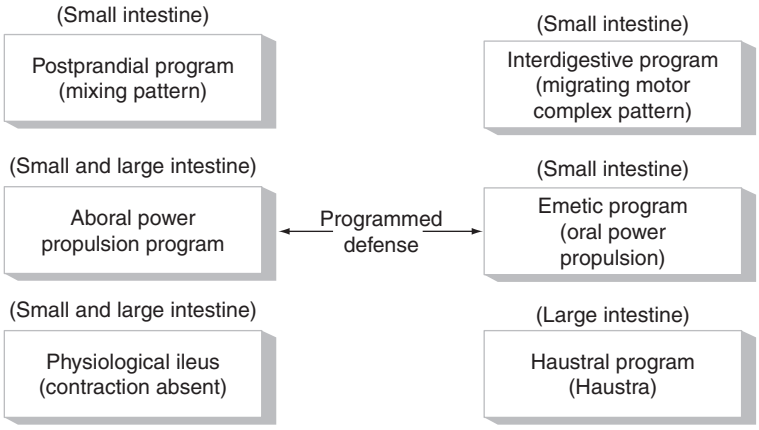

Figure 9 The ENS stores a library of programs for specific patterns of intestinal motor behavior that are adaptive for one or the other of the intestinal digestive states. In the postprandial state, the ENS runs the motor program for mixing movements while other programs are inoperative. At the completion of digestion and absorption of the nutrients of a meal, the postprandial program is terminated and the system calls up the interdigestive program, which, when running, programs for motor behavior described as the migrating motor complex. The ENS runs defensive programs characterized by powerful propulsive motility when threatening agents (toxins, infective agents, and food allergens) find their way into the intestinal lumen.

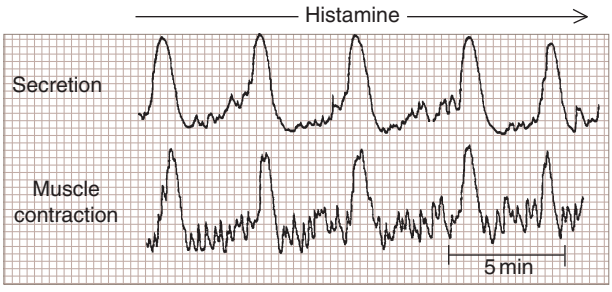

Figure 10 Activation of an ENS program for coordinated mucosal secretion and muscular contraction in the guinea pig colon. An overlay of histamine on the microcircuits of the submucosal plexus evokes repetitive cycles of mucosal secretion and coordinated contractions of the muscularis externa. The full-thickness preparation of intestinal wall was placed in an Ussing flux chamber with a small strain gauge attached to the serosal surface. The upper trace records mucosal secretion reflected as increases in transmucosal short-circuit chloride current; the lower trace is muscle contraction recorded by the strain gauge.

Only two neuromodulatory actions of histamine at the single neuron level of organization account for the emergent pattern of enteric network output that is subsequently transformed into a rhythmic pattern of integrated behavior at the level of the secretory glands and musculature. One neuromodulatory action is an intense increase in excitability, which mimics slow synaptic excitation in the cell bodies of neurons that make up the circuit (Figure 5). This excitatory action of histamine is mediated by histamine H_2 receptors expressed at the surfaces of the enteric neuronal cell bodies in the guinea pig intestine. The second neuromodulatory action of histamine is at synaptic connections between the neurons that form the integrative networks. Histamine acts at inhibitory receptors on the presynaptic terminals to suppress the release of neurotransmitter and thereby suppress transmission at the synapses. The presynaptic inhibitory action of histamine is mediated by the histamine H_3 receptor subtype in the guinea pig. These two actions of an overlay of histamine on two different histaminergic receptor subtypes at the single neuronal level are transformed into reconfigured rhythmic output at the integrated circuit level of organization.

See also: Autonomic Nervous System: Gastrointestinal Control; Autonomic Nervous System: Neuroanatomy; Autonomic Nervous System: Central Control of the Gastrointestinal Tract; Enteric Nervous System: Sensory Pathways; Enteric Nervous System: Neural Circuits and Chemical Coding; Enteric Nervous System Development; Enteric Nervous System: Disorders; Enteric Nervous System: Neurotrophic Factors; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Nausea and Vomiting.

Further Reading

Furness JB (2006) *The Enteric Nervous System*. Oxford: Blackwell.

Gershon M (1998) *The Second Brain*. New York: HarperCollins.

Grundy D and Schemann M (2006) Enteric nervous system. *Current Opinion in Gastroenterology* 22: 102–110.

Wood J, Alpers D, and Andrews PLR (1999) Fundamentals of neurogastroenterology. *Gut* 45: II1–II44.

Wood JD (1994) Physiology of the enteric nervous system. In: Johnson L, Alpers D, Christensen J, Jacobson E, and Walsh J (eds.) *Physiology of the Gastrointestinal Tract*, 3rd edn., vol. 1, pp. 423–482. New York: Raven Press.

Wood JD (2000) Neuropathy in the brain-in-the-gut. *European Journal of Gastroenterology & Hepatology* 12: 597–600.

Wood JD (2004) Enteric neuroimmunophysiology and pathophysiology. *Gastroenterology* 127: 635–657.

Wood JD (2005) Neurobiology of the enteric nervous system. In: Dyck PJ and Thomas PK (eds.) *Peripheral Neuropathy*, pp. 249–277. Philadelphia: Saunders.

Wood JD (2006) Cellular neurophysiology of enteric neurons. In: Johnson LR, Barrett KE, Ghishan FK, Merchant JL, Said HM, and Wood JD (eds.) *Physiology of the Gastrointestinal Tract*, 4th edn., pp. 629–663. San Diego: Elsevier.

Wood JD (2006) Integrative functions of the enteric nervous system. In: Johnson LR, Barrett KE, Ghishan FK, Merchant JL, Said HM, and Wood JD (eds.) *Physiology of the Gastrointestinal Tract*, 4th edn., pp. 665–683. San Diego: Elsevier.

Wood JD and Kirchgessner A (2004) Slow excitatory metabotropic signal transmission in the enteric nervous system. *Journal of Neurogastroenterology and Motility* 16: 71–80.

Enteric Nervous System: Sensory Pathways

N Clerc, CNRS–Université de la Méditerranée, Marseille, France
J B Furness, University of Melbourne, Parkville, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Nature of the Enteric Nervous System

The enteric nervous system (ENS) is involved in many physiological and pathophysiological processes in the gastrointestinal tract. It is essential to normal life, as indicated by the high morbidity or mortality associated with congenital or acquired deficiencies of this system. It generates motility, secretomotor, and vasomotor reflexes triggered by stimuli that indicate the mechanical state of the gastrointestinal tract and the presence, the consistency, and the chemical nature of the luminal contents. The ENS contains all the basic components of polysynaptic reflex arcs – that is, primary afferent neurons, interneurons, and motor neurons. Several types of interneurons (one ascending type and three descending types) and motor neurons (excitatory and inhibitory motor neurons innervating the longitudinal or circular muscle, secretomotor neurons, and vasomotor neurons) have been identified. In the case of the guinea pig, the animal model that has been most extensively used for studying the ENS, 14 functional neuron types, which are distinguished by their shapes, projections, neurochemistry, and cell physiological properties, have been progressively discriminated in the small intestine.

Enteric Primary Afferent and Motor Neurons Are Not Segregated, Unlike Neurons Controlling Other Organs

Remarkably, the primary afferent limbs of the ENS reflex arcs are still not well defined and the identity of the primary afferent neuron(s) is still the object of new investigations. This is due to the fact that the afferent and efferent limbs of the reflex arcs are not topographically and functionally compartmented, with a defined integrative center that would group the interneurons similarly to the dorsal horn, the medulla, or the midbrain. On the contrary, they are integrally embedded in the gut wall that is both the source of the stimuli to be detected by primary afferent neurons and the effector that receives commands elaborated by the motor neurons. The absence of functional topography is also a rule within each myenteric ganglion, although there is a limited topography in submucosal ganglia. These ganglia are organized in two interconnected networks, the myenteric and the submucosal plexuses, with the myenteric ganglia occurring all along the gastrointestinal tract. In total, the ganglia contain as many neurons as the spinal cord.

Different Types of Primary Afferent Neurons in the ENS

In this unique system, five different neuron types are known to act as primary afferent neurons. Two are entirely included in the ENS circuitry within the wall of the gastrointestinal tract: these neurons, which are ENS neurons *sensu stricto*, correspond to the Dogiel type II and some filamentous neurons. Although suggested by experimental studies as early as the beginning of the twentieth century, the fact that enteric neurons could act as primary afferent neurons was established definitively only in the 1990s. This function was attributed exclusively to the Dogiel type II neurons, called intrinsic primary afferent neurons (IPANs), until strong arguments were provided in 2004 showing that some filamentous neurons also subserve a primary afferent role.

A third neuron type has been called an intestinofugal afferent neuron (IFAN). As indicated by this terminology, the cell body of this neurons is located within the ENS, whereas the axon projects to sympathetic ganglia in order to activate sympathetic motor neurons innervating the gut (thus forming enteroenteric reflex circuits). These neurons have been described as being mainly Dogiel type I neurons, although some Dogiel type II neurons also seem to be IFANs.

The other primary afferent neurons that innervate the gastrointestinal tract are the vagal and spinal primary afferent neurons, whose central processes project respectively to the medulla and the dorsal horn. Much less numerous than the ENS primary afferent neurons, they innervate extensive areas of the gastrointestinal tract, compared to the intrinsic primary afferent neurons, because of the broad distribution of their collateral processes. Here, we focus exclusively on ENS primary afferent neurons.

Criteria for the Identification of Enteric Primary Afferent Neurons

Morphological Criteria

The first criterion in the literature is morphological. It needs to be placed in historical context (the end of the nineteenth century) because morphology is not convincing by itself. On the basis of functional

arguments indicating that the enteric reflex arcs were totally included in the gastrointestinal tract, Dogiel sought a presumptive enteric primary afferent neuron and intuitively attributed this function to the neurons that are now called Dogiel type II. These neurons have round or oval smooth cell bodies with several long smooth processes that seem to be axons. Their cell bodies are found in both the myenteric and the submucosal plexuses. In each case, one process innervates the mucosa or muscle, whereas other processes connect with some myenteric neurons.

Functional Criteria

Experiments directed at the functional identification of enteric primary afferent neurons were first successfully conducted in the 1990s using *in vitro* preparations. The first criterion that was advanced was the persistence of neuron activity elicited by mechanical or chemical stimuli in the absence of synaptic input (which was blocked by hexamethonium or lowering the Ca^{2+} concentration of the saline). Such identification was first provided using activity-induced expression of the immediate early gene c-*fos*. A mild mechanical stimulus (puffs of N_2), applied to the mucosa when synaptic transmission had been blocked, activated submucosal Dogiel type II neurons of the small intestine. Further functional studies were performed using intracellular micropipettes to record non-synapse-dependent electrical activity evoked in myenteric Dogiel type II neurons of the small intestine by physiological stimuli. Brief bursts of action potentials were induced when the mucosal endings of the myenteric Dogiel type II neurons were stimulated by acetate. In these neurons, ongoing firing of action potentials was evoked by circumferentially stretching nonparalyzed preparations. The colonic myenteric filamentous neurons were also activated by circumferential stretching that induced an ongoing firing of action potentials in paralyzed preparations.

A second criterion was to demonstrate that action potentials are not generated at the level of the soma but in distal processes. This was demonstrated by recording proximal process potentials (PPPs, also referred to as electrotonic potentials). PPPs are recorded when action potentials are initiated in a process, and conducted to the cell body, but do not invade the cell body. Thus, a PPP in response to a mechanical or chemical stimulus indicates that the terminal of the process was a 'sensory' receptor (directly or indirectly) for the stimulus. PPPs were recorded from ileum myenteric Dogiel type II neurons when 5-hydroxytryptamine (5-HT) was applied on the mucosal endings of these neurons. 5-HT, likely to be released by enterochromafin cells activated by mechanical or chemical stimuli, is a possible intermediate in the sensory transduction process. PPPs were also recorded from ileum myenteric Dogiel type II neurons when actively contracting preparations were stretched circumferentially. They were also recorded from colonic filamentous neurons in paralyzed preparations stretched circumferentially. The fact that action potentials can be generated at the level of the mucosal processes of Dogiel type II neurons in the ileum was confirmed by collision experiments in which the soma and the mucosa were stimulated sequentially.

The initial change in membrane potential of a receptor when it is stimulated, the receptor potential, is a local and nonpropagated event that decreases in amplitude with increasing distance of the recording site from the receptor. For this reason, it cannot be recorded when it is generated far from the cell body. However, in one case such events could be recorded from myenteric Dogiel type II neurons in the small intestine. This was done in paralyzed ileum preparations by focal deformation of the neuronal processes within ganglia in which recordings were made. Depolarizing potentials that could reach the threshold for generating action potentials were recorded.

Mediators of Transduction of Physiological Stimuli

The endings of both intrinsic and extrinsic primary afferent neurons are in the lamina propria, on the opposite side of the mucosal epithelium exposed to chemicals in the lumen. Thus, to signal to the afferent nerve endings, the luminal chemical must either cross the epithelium or cause a chemical messenger to be released, most likely from mucosal endocrine cells. Two mediators have been identified, 5-HT and ATP, both of which can be released from the stimulated mucosa. There are many other products of mucosal endocrine cells that may also have roles.

Additional Functions of Dogiel Type II and Filamentous Neurons

Dogiel type II neurons, like many neurons (e.g., motor (efferent) neurons to skeletal muscle) that, through their collaterals, can act as interneurons, have dual functions. Like some spinal primary afferent neurons that have additional motor effects through transmitter release from their peripheral endings, Dogiel type II neurons probably have motor effects when they release transmitter onto the mucosal epithelium. In addition, it is remarkable that the somas of these neurons receive slow excitatory postsynaptic inputs, and in this context they can be considered to be interneurons.

Filamentous neurons are classically described as being ascending or descending interneurons. Which type of descending interneurons are the mechanosensitive filamentous neurons remains to be established.

Enteric primary afferent neurons also act as nociceptors, whose activation by noxious stimuli triggers protective responses. In the small intestine and colon, protective responses are initiated by irritants that are included in enemas, by bacterial products that cause diarrhea, and by parasitic infestations that induce expulsion of the parasites.

Enteric Primary Afferents and Integrative Functions

The enteric primary afferent neurons detect several aspects of the physiological condition of the intestine, including wall tension (representing the bulk of material in the lumen and the contractile state of the intestine), mechanical forces on the mucosa, and the chemical nature of the luminal contents. This information is integrated (although the details of how the integration is accomplished are unknown) to determine the output of the ENS, which includes mixing and propulsive activity of the musculature, secretion of water and electrolytes, and adjustment of mucosal blood flow. These local enteric controls are in turn integrated with signals from other gut regions and from the central nervous system (CNS).

The fact that the enteric neurons that have been demonstrated to function as primary afferent neurons (the Dogiel type II neurons and the filamentous neurons) are also both interneurons is striking. The Dogiel type II neurons connect with each other in order to communicate circumferentially, whereas the filamentous neurons propagate the messages distally (oral to anal). As stated previously, the ENS contains reflex arcs with no topographically individualized integrative center. We can speculate that this particular condition predestines the interneurons to fulfill the primary afferent function. It is possible that specific 'sensory' signals have to be driven circumferentially or distally in order to produce coordinated reflexes, which would be performed in an 'economic' way if the interneurons transduce the mechanical stimuli.

Terminology: 'Primary Afferent Neuron' and 'Sensory Neuron'

It is necessary to conceptually distinguish between the primary afferent innervation of tissues and organs and their sensory innervation: The activation of visceral receptors (e.g., stretch receptors in the lung and osmoreceptors in the liver) evokes no sensation, and therefore the term primary afferent is most appropriate to designate the neurons that carry this information. On the other hand, some neurons innervating the viscera do give rise to sensation, such as pain and warmth, and these are properly called sensory neurons. The important difference between visceral primary afferent neurons in general and the subset that mediate conscious sensation of the state of the viscera extends to the primary afferent neurons that reside within the intestine.

Ionic Basis of the Excitability of the Enteric Primary Afferent Neurons

The electrophysiology of the Dogiel type II neurons has been studied in detail in the guinea pig small intestine. As yet, comparable investigations of the filamentous neurons have not been undertaken. The Dogiel type II neurons of the small intestine exhibit a suite of channels through which their excitability and patterns of firing are controlled. The patterns of firing and the excitability influence the nature (e.g., propulsion or mixing) and intensity of reflexes. Because of their critical position in the circuits (being the first neurons of enteric reflex pathways), the ion channels of enteric primary afferent neurons are potential targets for therapeutic drugs. The distinctive features of the electrophysiology of Dogiel type II neurons are a large action potential with a hump on the falling phase, a large and prolonged after-hyperpolarizing potential (the slow AHP) that can last 4 to approximately 20 s, a low-input resistance, and a prominent hyperpolarization-activated cation current.

Two inward currents underlie the action potential in the soma, a tetrodotoxin (TTX)-sensitive Na^+ current (I_{NaV}) and a Ca^{2+} current (I_{CaV}). A TTX-resistant Na^+ current (TTX-R I_{NaV}), carried by $Na_v1.9$ channels, also occurs in Dogiel type II neurons, but there is no evidence that this contributes in any significant way to the action potential. I_{CaV} during the action potential is carried by N channels with a small contribution of R and P/Q-type channels.

Under conditions in which background synaptic transmission is suppressed, the Dogiel type II neurons exhibit the slow AHP that identifies them as the so-called AH neurons, but when they are acted upon by neurotransmitters or hormones, the slow AHP can be suppressed. The AHP current (I_{AHP}) is also active at rest. This current is blocked by charybdotoxin and clotrimazole, both of which depolarize the neurons. When AH/Dogiel type II neurons are in a state of low excitability, they fire only one to three action potentials to a 500 ms intracellular depolarizing pulse and they have a resting membrane potential of approximately −60 to −65 mV. Some contribution to the

background conductance is made by the hyperpolarization-activated cation current (I_h) and Cl^- currents. AH/Dogiel type II neurons also have an A-type K^+ current (I_A) that may have a low level of activation at rest.

An early AHP is continuous with the falling phase of the action potential and lasts approximately 50–100 ms. The currents that cause this AHP include a contribution from large conductance potassium channels (BK channels). This early AHP is followed by a depolarization that precedes the slow AHP and is due to a Ca^{2+}-activated depolarizing cation current.

The slow AHP is the result of two opposite currents: the slow AHP current itself (I_{AHP}) and the I_h current that generates a sustained depolarization and that is increased by the I_{AHP}-induced hyperpolarization. I_{AHP} is carried by IK channels whose opening probability is reduced by phosphorylation by protein kinase A (PKA) or protein kinase C (PKC). Consistent with this, the PKC stimulant, phorbol dibutyrate (PDBu), causes excitability increases, membrane depolarization, and increases input resistance. The IK channels are activated by dephosphorylation, particularly via calcineurin.

Synaptically Mediated Changes in Excitability

As previously indicated, AH/Dogiel type II neurons receive slow synaptic inputs that have an important influence on their excitability and pattern of firing. Synaptic inputs arise from other AH/Dogiel type II neurons, from serotoninergic descending interneurons, and from neurons that are still not identified.

In AH/Dogiel type II neurons, slow excitatory postsynaptic potentials (EPSPs), which can be elicited by stimulation of the interganglionic connectives at 10–20 Hz for 1s, are characterized by a depolarization of membrane potential and a transient increase in somatic excitability, accompanied by a reduction of the slow AHP. The effects commonly peak in a few seconds, last for several minutes, and are primarily due to a reduction in resting K^+ conductance, including inhibition of I_{AHP}. This is reported to be combined in some neurons with activation of a Cl^- conductance. The change in the slow AHP is not caused by reduced Ca^{2+} entry. One possible intermediate in causing this change is the activation of protein kinases and subsequent inactivation of the IK channels, as described previously.

Slow EPSPs induced in AH/Dogiel type II neurons by stimulating circumferential pathways (which preferentially stimulates the synaptic inputs arising from other AH/Dogiel type II neurons) are mimicked by neurokinin-1 receptor (NK_1R) and neurokinin-3 receptor (NK_3R) agonists. The activation of each receptor triggers two different types of slow EPSPs, which can be discriminated by the fact that the slow EPSPs mediated by NK_1R have significantly shorter latencies than those due to activation of NK_3R. NK_3R appears to be the predominant mediator of the slow EPSPs evoked by stimulation of circumferential inputs.

Slow EPSPs are also induced in AH/Dogiel type II neurons when the descending pathways, included serotonergic pathways, are activated. These slow EPSPs, with longer time to peak and duration than the neurokinin-mediated slow EPSPs, are mediated via receptors that have the properties of 5-HT7 receptors (they are blocked by SB 269970); 5-HT1P-mediated slow EPSPs have also been described. The 5-HT7 receptor has an impressive similarity with the 5-HT1P receptor, a G-protein-coupled receptor that acts through G_o, PKC, and finally accumulation of cAMP and activation of PKA. Whether 5-HT7 and 5-HT1P are the same entity remains to be established.

Other potential enteric neurotransmitters also cause slow excitation of enteric neurons. These include acetylcholine, acting via muscarinic receptors; cholecystokinin; gastrin-releasing peptide; pituitary adenylyl cyclase-activating peptide; and vasoactive intestinal peptide.

Another synaptic event, more prolonged than slow EPSPs, is also observed in AH/Dogiel type II neurons. This synaptic event, called sustained slow postsynaptic excitation (SSPE), outlasts nerve stimulation by many minutes, and successive SSPEs show substantial facilitation. The increased excitation during a facilitated SSPE can be maintained after stimulation for as long as 4 h.

The SSPE is induced by repeated, low-frequency (0.5–2 Hz) trains of stimulation of synaptic inputs. It involves increased somatic excitability that is facilitated by successive stimuli. During stimulation, the soma excitability of the AH/Dogiel type II neurons slowly increases, they depolarize, and their input resistance increases markedly, often by more than twofold. This magnitude of increase, accompanied by depolarization, implies that K^+ currents are inhibited. During the SSPE, the slow AHP is reduced in amplitude and duration and can eventually be blocked. With sufficient depolarization of the neurons, invasion by antidromic action potentials is suppressed. Spontaneous action potentials are rare and the occurrence of fast EPSPs is not reported during the heightened excitability caused by the SSPE. The suppression of the slow AHP results from IK channels being targeted by phosphorylating enzymes during the SSPE. Consistent with this, the PKC stimulant PDBu suppresses the slow AHP in a manner identical to that seen during the SSPE. In addition, PDBu causes

somatic excitability increases, membrane depolarization, and increased input resistance in a concentration-dependent manner, mimicking the SSPE. In a converging way, PKC inhibitors Ro31–8220 and calphostin C prevent the SSPE. On the contrary, the SSPE is maintained in the presence of the PKA inhibitor H89.

An obvious candidate neurotransmitter for the SSPE could be a tachykinin. However, this is unlikely because the SSPE persists after block of NK_1 and NK_3 tachykinin receptors on enteric neurons.

Influence of Inflammation on the Excitability of Enteric Primary Afferent Neurons

Inflammation of the small intestine or colon causes marked changes in motility and secretion. Effects can long outlast the period of inflammation and cause ongoing pathology (e.g., irritable bowel syndrome). Because of these observations, there has been considerable interest in the reactions of IPANs to inflammation.

Synaptic changes have been described in a neuron population that should include the filamentous neurons; however, this neuron type was not discriminated from the rest of the population. On the contrary, the changes that affect the AH/Dogiel type II neurons have been specifically analyzed following inflammation. Infection with *Trichinella spiralis* in the jejunum and administration of trinitrobenzene sulfonic acid (TNBS) into the lumen of the distal colon both increase the excitability of the AH/Dogiel type II neurons. In these models, in addition to inhibition of the I_{AHP}, different ionic currents were affected. In the *T. spiralis* infection model, the increase of excitability was reversibly suppressed by H_2 histamine receptor blockers and might thus reflect an endogenous release of histamine. Activation of H_2 histamine receptors probably suppresses the I_A current, which not only contributes to the resting potential but also modulates the firing frequency of neurons. Histamine also activates a Cl^- conductance via H_2 receptors. Thus, I_A blockade and Cl^- conductance activation could explain the firing increase, the membrane depolarization, and the increase in input resistance described in this model. The abolition of the slow AHP was accompanied by a decrease in the action potential duration, indicating a decrease of Ca^{2+} entry that can explain the reduction of the amplitude and the duration of the slow AHP. All these changes were accompanied by the presence of fast EPSPs that routinely reached suprathreshold depolarization and caused action potential discharge. In the TNBS model, the increased neuronal excitability and the reduction in the slow AHP was proposed to be principally due to an increased activation of I_h rather than alterations in gK_{Ca}. Action potential shape, and therefore Ca^{2+} entry into the cell, appeared to be unaffected. In this model, the changes of excitability were shown to be mediated by cyclooxygenase-2 (COX-2) because they were abolished in TNBS-treated animals in which COX-2 was inhibited by colony-forming units.

The excitability of these neurons is increased by the inflammatory mediators histamine, prostaglandins, leukotrienes, and interleukins and by activation of proteinase activated receptors and 5-HT. Each of these inhibits the slow AHP current, I_{AHP}, thus causing an increase in somatic excitability, and in most cases they depolarize the membrane potential and increase Rin. The mechanisms leading to a decrease of I_{AHP} differ according to the mediator and also differ from the actions of the SSPE in some respects. The inflammatory mediators may also act on other targets to affect excitability of AH/Dogiel type II neurons.

Are There IPANS in the Esophagus and Stomach?

There is a lack of knowledge about the existence and possible identities of IPANs in the esophagus and stomach. In these regions, many of the controlling reflexes pass through the CNS, notably through vagal control centers in the medulla oblongata. Efforts to identify AH/Dogiel type II neurons by their morphology or electrophysiological properties have led to the conclusion that this type of neuron is rare or absent from the esophagus, is not found in the proximal stomach, and is rare in the gastric antrum. Whether another type of neuron plays the role of IPANs, or whether intrinsic afferent neurons are indeed rare in these regions, has not been resolved.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Enteric Nervous System: Neural Circuits and Chemical Coding; Enteric Nervous System Development; Enteric Nervous System: Disorders.

Further Reading

Brehmer A (2006) Structure of enteric neurons. *Advances in Anatomy* 186: 1–95.

Brookes SJH and Costa M (2002) Cellular organisation of the mammalian enteric nervous system. In: Brookes SJH and Costa M (eds.) *Innervation of the Gastrointestinal Tract*, pp. 393–467. New York: Taylor & Frances.

Clerc N, Furness JB, Kunze WAA, Thomas EA, and Bertrand PP (1999) Long term effects of synaptic activation at low frequency on excitability of myenteric AH neurons. *Neuroscience* 90: 279–289.

Furness JB (2006) *The Enteric Nervous System*. Oxford: Blackwell.

Gershon MD (2005) Nerves, reflexes, and the enteric nervous system. *Journal of Clinical Gastroenterology* 38: S184–S193.

Kirchgessner AL, Tamir H, and Gershon MD (1992) Identification and stimulation by serotonin of intrinsic sensory neurons of the submucosal plexus of the guinea pig gut: Activity-induced expression of Fos immunoreactivity. *Journal of Neuroscience* 12: 235–248.

Kunze WAA, Clerc N, Furness JB, and Gola M (2000) The soma and neurites of primary afferent neurons in the guinea-pig intestine respond differentially to deformation. *Journal of Physiology (London)* 526: 375–385.

Lomax AE, Fernández E, and Sharkey KA (2004) Plasticity of the enteric nervous system during intestinal inflammation. *Neurogastroenterology & Motility* 17: 4–15.

Monro RL, Bornstein JC, and Bertrand PP (2005) Slow excitatory post-synaptic potentials in myenteric AH neurons of the guinea-pig ileum are reduced by the 5-hydroxytrytamine-7 receptor antagonist SB 269970. *Neuroscience* 134: 975–986.

Rugiero F, Gola M, Kunze WAA, et al. (2002) Analysis of whole cell currents by patch clamp of guinea-pig myenteric neurones in intact ganglia. *Journal of Physiology (London)* 538: 447–463.

Rugiero F, Mistry M, Sage D, et al. (2003) Selective expression of a persistent TTX-resistant Na^+ current and $Na_v1.9$ subunit in myenteric sensory neurons. *Journal of Neuroscience* 23: 2715–2725.

Schemann M and Neunlist M (2004) The human enteric nervous system. *Neurogastroenterology & Motility* 16: 55–59.

Spencer NJ and Smith TK (2004) Mechanosensory S-neurons rather than AH-neurons appear to generate a rhythmic motor pattern in guinea-pig distal colon. *Journal of Physiology (London)* 558(2): 577–596.

Starodub AM and Wood JD (2000) Histamine suppresses A-type potassium current in myenteric neurons from guinea pig small intestine. *Journal of Pharmacology and Experimental. Therapeutics* 294: 555–561.

Wood JD (2004) Enteric neuroimmunophysiology and pathophysiology. *Gastroenterology* 127: 635–657.

Entorhinal Cortex

A Solodkin and G W Van Hoesen, University of Iowa, Iowa City, IA, USA

© 2009 Published by Elsevier Ltd.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

The term entorhinal cortex (Brodmann's area 28) is derived from the fact that it lies 'within' the rhinal sulcus, and this is the case in all species including the great apes. However, in humans the rhinal sulcus is often small and anteriorly placed and it is the collateral sulcus instead, that aids in delimiting this cortex laterally. In carnivores and rodents it constitutes the largest part of the piriform lobe, whereas in nonhuman primates it forms the anterior portion of the parahippocampal gyrus. In humans, the anterior and medial portion of the parahippocampal gyrus is called the gyrus ambiens, but this is merely a bulge of entorhinal cortex that sits in the tentorial aperture medial to the point where the free edge of the tentorium cerebelli cuts or grooves the parahippocampal surface (Figure 1(a)).

The entorhinal cortex is a type of periallocortical cortex and is atypical when compared to isocortical areas or for that matter, to any other cortex. For example, the size of the cells composing its layers appears involuted compared to isocortical areas. In area 28, the largest cells are located superficially, whereas in the isocortex they are found more deeply. Additionally, it resembles a hybrid, where the superficial layers are related more to the proisocortices, and the deeper layers to the allocortices. Additionally, an acellular layer IV, or lamina dissecans, formed by tangentially oriented myelinated fibers is readily apparent.

There is variability among species in entorhinal cytoarchitecture, and this seems most apparent in layers II and III. In the rat, the neurons found in these layers have a horizontal or laminar organization, forming a continuous and somewhat homogeneous band at variable distances from the pia. In monkeys, however, these cells are sometimes clumped together forming islands of cells. This discrete organization reaches its maximal expression in the human brain producing clear anatomical modularity within the superficial layers of the entorhinal cortex. This is highly suggestive of a vertical pattern of organization coexisting with the more conventional horizontal multilayered one. Entorhinal modules in the human brain are suggested even in the level of the gross specimen because this cortex can be delimited by the presence of superficial bumps easily detected with the unaided eye (Figure 1(a)). Near the turn of the century the Swedish anatomist Retzius named these verrucae to call attention to their resemblance to skin or epidermal diseases of viral origin. Thus, the entorhinal cortex in humans appears formed by a mosaic of components whose most conspicuous elements are the cell islands of layer II and the myelinated fibers associated with them (Figure 1(b)). Additionally, the distribution and disposition of local circuit neurons as well as cholinergic and dopaminergic afferent fibers, cytochrome oxidase patches and the distribution of its vasculature, highlights and supports the notion of modularity (Figure 1(c)).

In terms of connections, the entorhinal cortex is remarkable since it can be conceptualized as an interface between the hippocampal formation on the one hand and a large variety of association and limbic cortices on the other. In fact, the entorhinal cortex, via the perforant pathway (which originates in cells of layers II and III), is the main source of extrinsic afferent input to the hippocampal formation, targeting the ammonic fields CA1, CA2, and CA3 in addition to the subiculum and dentate gyrus. Nonhippocampal efferents of the entorhinal cortex project to neighboring areas such as perirhinal cortex (Brodmann's area 35), posterior parahippocampal gyrus (areas TH and TF) and inferior and polar temporal cortices. Other efferents from the entorhinal cortex are widely distributed to several association and limbic cortices such as prefrontal and orbitofrontal (areas 9, 25, and 46), retrosplenial (area 29), cingulate (area 24) and posterior parietal cortices (area 7) as well as superior temporal gyrus (area 22). These are of great interest since they arise from the deep layers of the entorhinal cortex, from neurons known to receive powerful hippocampal output.

Afferents to the entorhinal cortex arise from a large variety of cortical and subcortical areas including adjoining areas such as TF and TH in the posterior parahippocampal gyrus, subiculum, perirhinal cortex (area 35), the temporal polar cortex (area 38), periamygdaloid cortex, and the superior temporal gyrus. Other areas sending afferent fibers to the entorhinal cortex are the prefrontal and orbitofrontal cortices (areas 25 and 32); limbic cortices send important afferent input from cingulate (areas 24 and 23), retrosplenial (area 29), piriform (in rodents), and agranular insular cortices. Among the subcortical areas sending afferent projections to the entorhinal cortex are the septal nuclei, anterior and medial thalamus, raphe nuclei, amygdala, locus coeruleus, basal nucleus of Meynert, and anterior olfactory nucleus.

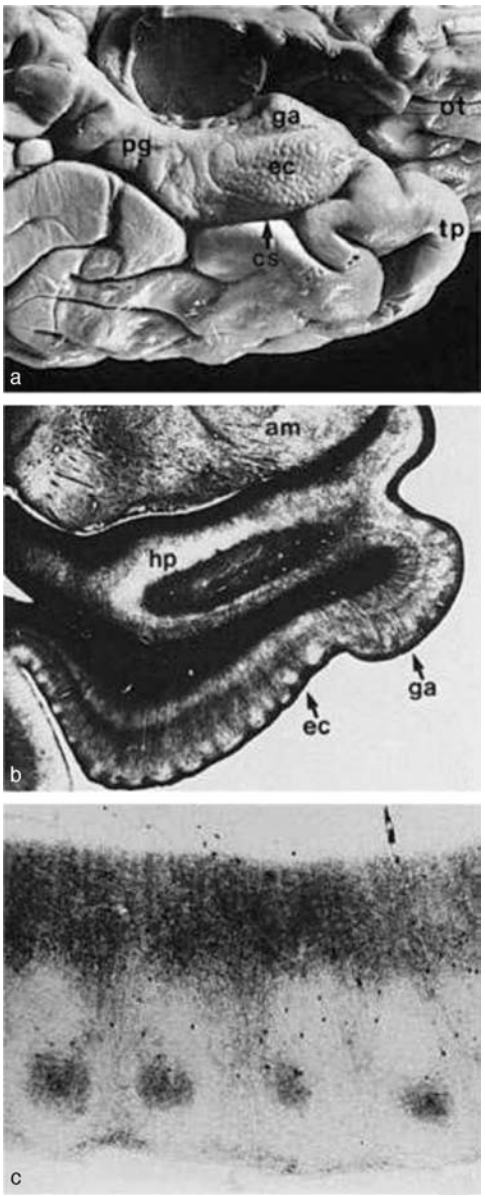

Figure 1 (a) Gross view of the ventral medial temporal lobe in the normal human brain after dissection of the brain stem and cerebellum. Note the position of the parahippocampal gyrus (pg) medial to the collateral sulcus (cs) and posterior to the temporal pole (tp). The anterior part of the parahippocampal gyrus contains the entorhinal cortex (ec) which also covers the gyrus ambiens (ga). Note the corrugated appearance of the entorhinal cortex created by the verrucae or wartlike bumps that demarcate its extent. (b) Cross-section through the entorhinal cortex of the normal human brain stained with the Gallyas fiber stain. The modular appearance of the fiber distribution outlines the verrucae. (c) Parvalbumin immunostaining for this calcium-binding protein in the entorhinal cortex of the normal human brain. The sharply defined patches of neuropil and cellular staining in the superficial layers are coexistent with the verrucae and layer II islands of the entorhinal cortex. The more dense laminar immunostaining for parvalbumin in deeper layers corresponds to layer III. am, amygdala; hp, hippocampus; ot, olfactory tract.

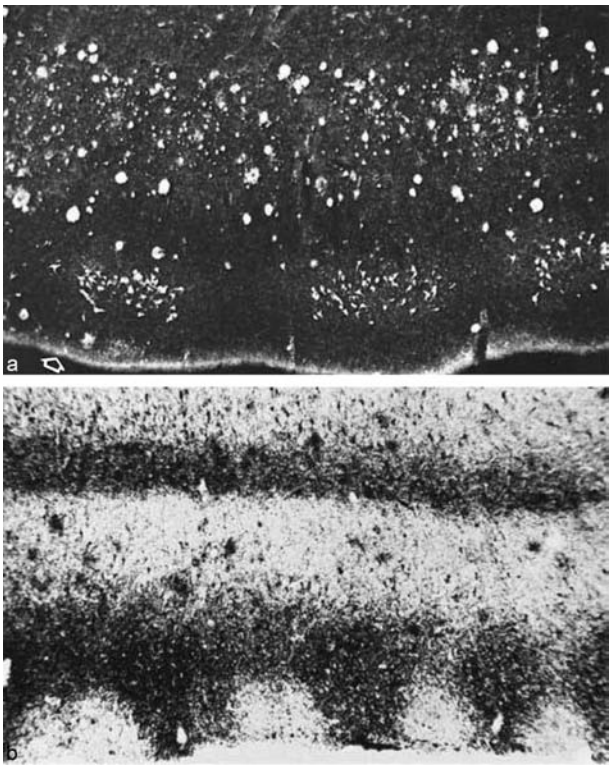

Figure 2 (a) Thioflavine-S fluorochrome staining for neuritic plaques, diffuse plaques, and neurofibrillary tangles in the entorhinal cortex in Alzheimer's disease. Note the islands of neurofibrillary tangles in the superficial layers of the entorhinal cortex and their coincidence with the verrucae as seen along the pial surface (marked by the arrow). The large circular images in the deeper layers of entorhinal cortex are largely neuritic and diffuse plaques. (b) Alz-50 immunostaining for the abnormally phosphorylated tau protein A68 in the entorhinal cortex in Alzheimer's disease. The modules of immunostaining in the superficial layers correspond to the islands of neurons that form layer II and the distribution of their dendrites.

In summary, the entorhinal cortex receives direct inputs from a large number of association and limbic cortices, and in turn, sends widespread neural systems to other parts of the cortex. It is in a unique position to provide input to the hippocampus and then disseminate hippocampal output to other parts of the cortical mantle. Thus, it is not surprising that it is thought to play a role in memory consolidation related to all sensory modalities. The close and two-way relationship established between the entorhinal cortex and hippocampal formation on the one hand and the entorhinal cortex and association and limbic cortices on the other makes its role in memory processing highly plausible.

Clinical and behavioral observations support the above. For example, the entorhinal cortex in humans is the most affected cortical area in Alzheimer's

disease. In this disorder, pathology targets the entorhinal cortex early in the illness and destroys at least three of its five cell layers if the duration of the illness is long. The modules are always targeted (**Figures 2(a) and 2(b)**). Due to the modular organization of this cortex, these changes in all likelihood do not have to be complete to disrupt the functions modules are designed to subserve.

See also: Alzheimer's Disease: An Overview; Brodmann's Areas; Cerebral Cortex; Evolution of the Limbic System; Hippocampus; Perirhinal Cortex.

Further Reading

Amaral DG, Insausti R, and Cowan WM (1987) The entorhinal cortex of the monkey. I. Cytoarchiteconic organization. *Journal of Comparative Neurology* 264: 326–355.

Braak H and Braak E (1992) The human entorhinal cortex: Normal morphology and laminar-specific pathology in various diseases. *Neuroscience Research* 15: 6–31.

Hyman BT, Van Hoesen GW, Damasio AR, and Barnes CL (1984) Alzheimer's disease: Cell-specific pathology isolates the hippocampal formation. *Science* 225: 1168–1170.

Insausti R, Amaral DG, and Cowan WM (1987) The entorhinal cortex of the monkey. II. Cortical afferents. *Journal of Comparative Neurology* 264: 356–395.

Insausit R, Tunon T, Sobreviela T, Insausti AM, and Gonzalo LM (1995) The human entorhinal cortex: A cytoarchitectonic analysis. *Journal of Comparative Neurology* 355: 171–198.

Solodkin Q and Van Hoesen GW (1996) Entorhinal cortex modules of the human brain. *Journal of Comparative Neurology* 365: 610–627.

Suzuki WA and Amaral DG (1994) Perirhinal and parahippocampal cortices of the macaque monkey: Cortical afferents. *Journal of Comparative Neurology* 350: 497–533.

Suzuki WA and Amaral DG (1994) Topographical organization of the reciprocal connections between the monkey entorhinal cortex and the perirhinal and parahippocampal cortices. *Journal of Neuroscience* 14: 1856–1877.

Van Hoesen GW (1982) The parahippocampal gyrus: New observations regarding its cortical connections in monkey. *Trends in Neuroscience* 5: 345–350.

Van Hoesen GW and Solodkin A (1993) Some modular features of temporal cortex in humans as revealed by pathological changes in Alzheimer's disease. *Cerebral Cortex* 3: 465–475.

Witter MP (1993) Organization of the entorhinal-hippocampal system: A review of current anatomical data. *Hippocampus* 3: 33–44.

Witter MP, Van Hoesen GW, and Amaral DG (1989) Topographical organization of the entorhinal projection to the dentate gyrus of the monkey. *Journal of Neuroscience* 9: 216–228.

Relevant Website

http://www.alzforum.org – Alzheimer Research Forum.

Entrainment of Circadian Rhythms by Light

A M Rosenwasser, University of Maine, Orono, ME, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In general, the endogenous, inherited period of the circadian pacemaker is close to – but not exactly equal to – 24 h. Indeed, the term *circadian* ('approximately daily') was coined to emphasize this fundamental feature of biological timing. Thus, endogenous circadian pacemakers depend upon signals derived from exogenous daily cycles in order to synchronize, or entrain, to the exact 24 h periodicity of the environment. It is only through this entrainment process that circadian pacemakers acquire their adaptive utility as biological 'clocks,' responsible for coordinating behavior and physiology with local time. While several periodic environmental factors may contribute to or modulate the entrainment of the circadian pacemaker, the most prominent, widespread, and fundamental source of circadian entrainment is the daily alternation of environmental light and darkness. In nature, light intensity and spectral composition change in a complex manner as a function of time of day, and depend not only on solar illumination but also on lunar phase, starlight, latitude, and local climatic conditions. Further, actual exposure to available environmental light depends dramatically on an animal's ecological niche and lifestyle (e.g., time spent in burrows, caves, or deep in the ocean). Nevertheless, simple square-wave ('on–off') light–dark cycles have proved to exert effective circadian entrainment in the laboratory, and much has been learned about the formal, physiological, and molecular bases of light entrainment from studies conducted under such simplified environmental conditions.

Formal Analysis of Light Entrainment

Parametric and Nonparametric Entrainment

Brief light pulses can evoke rapid and permanent changes in the phase of the free-running circadian pacemaker. The response of the circadian pacemaker to a light pulse is phase dependent, such that photic stimulation near 'subjective, dusk' or early in the subjective night results in phase delays (i.e., shifts to relatively later timing in subsequent cycles), while photic stimulation late in the subjective night or near dawn results in phase advances (i.e., shifts to relatively earlier timing in subsequent cycles, see Figure 1). In contrast, the circadian pacemaker is relatively refractory to photic stimulation during the circadian daytime, indicating that the critical phases of light exposure for entrainment occur near dusk and dawn. The function relating the magnitude and direction of circadian phase shifting to the phase of stimulation for any given stimulus is referred to as the 'phase–response curve' (PRC) for that stimulus (see Figure 2). Extensive research has shown that the general shape of the light-pulse PRC, as just described, is seen consistently across species, and indeed, across orders, and in both nocturnal and diurnal life-forms. Thus, while the seminal vertebrate PRC experiments were done mainly in nocturnal rodents, more recent analyses have shown that the 'typical' diurnal mammalian PRC very closely resembles that of nocturnal mammals in both amplitude and waveform.

Colin Pittendrigh systematized these observations in the highly influential and extensively validated nonparametric theory of light entrainment, which proposed that phasic responses to brief light exposures occurring at specific critical circadian phases (i.e., near dusk and dawn) are responsible for entrainment of the circadian pacemaker, even under more natural light–dark schedules with 'complete' photoperiods (see Figure 3). This model may be contrasted conceptually with the parametric entrainment theory, associated with the work of Jurgen Aschoff, and based on extensive observations indicating that the period of a free-running circadian pacemaker changes systematically (either lengthening or shortening, depending on the species) as a function of continuous light intensity. Thus, parametric entrainment theory proposes that the tonic effects of temporally extended light exposure alter the underlying period of the circadian pacemaker, causing it to match the periodicity of the entraining environment.

Although nonparametric entrainment has proved to be a powerful descriptive and predictive model, recent analyses indicate that both phasic, nonparametric effects and tonic, parametric effects may contribute to circadian entrainment, especially under field conditions, or during exposure to more realistic laboratory simulations of the natural light–dark cycle. Thus, some species appear to entrain successfully in the natural environment without ever being exposed to abrupt, dawn- and dusk-related light transitions, while in laboratory experiments, both brief and extended light exposures result in long-term changes in free-running period as measured in continuous darkness (i.e., 'after-effects' of prior entrainment). Further, quantitative simulations incorporating both nonparametric (phase-shifting) and parametric (period-adjusting) effects

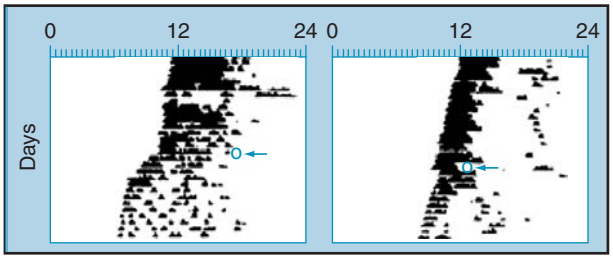

Figure 1 Examples of phase shifting of free-running circadian activity rhythms by brief light pulses. Left panel: Circadian rhythm of wheel running recorded from a hamster in constant darkness. On the day indicated by the arrow, the animal was exposed to light for 30 min (open circle) beginning at circadian time 20. (By convention, the onset of activity in a nocturnal animal is defined as circadian time 12; thus, circadian time 20 is 8 circadian h after activity onset; where 1 circadian hour = circadian period/24.) On circadian cycles following the light pulse, activity onset occurs earlier than would be predicted by extrapolating from the phase of the preceding free-running rhythm. After several days of 'transient' cycles, as is typical for advancing phase shifts, a stable phase advance of several hours is revealed. Right panel: In this experiment, a similar light pulse was presented at circadian time 14, resulting in a phase delay shift of approximately 1 h, which was complete within one circadian cycle. Phase resetting without transients is typical of light-induced phase delays. Reproduced from Mistlberger RE and Rusak B (2005) Circadian rhythms in mammals: Formal properties and environmental influences. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, 4th edn., pp. 1–30. Philadelphia: Elsevier-Saunders, with permission from Elsevier.

result in more stable entrainment than do models based solely on nonparametric effects.

Toward Naturalistic Light–Dark Cycles

The importance of parametric effects for natural entrainment is also highlighted by laboratory experiments utilizing simulated gradual twilight transitions, rather than the more typical square-wave light–dark cycles. Acute exposure to brief twilight-like light pulses and brief square-wave light pulses yields identical phase shifts when the two pulse types are equated for photon numbers, indicating that the nonparametric effects of light pulses are independent of pulse 'shape.' In contrast, inclusion of simulated twilight transitions during steady-state entrainment to extended photoperiods alters dramatically the pattern of circadian entrainment relative to standard square-wave light–dark cycles. In addition to altering the timing of daily activity onsets and offsets, exposure to twilights results in more stable and more effective entrainment, and alters the free-running period as measured during subsequent exposure to constant darkness. Together, these studies strongly suggest that parametric effects of twilight exposure on the underlying free-running period increase the entraining efficacy of a light–dark cycle.

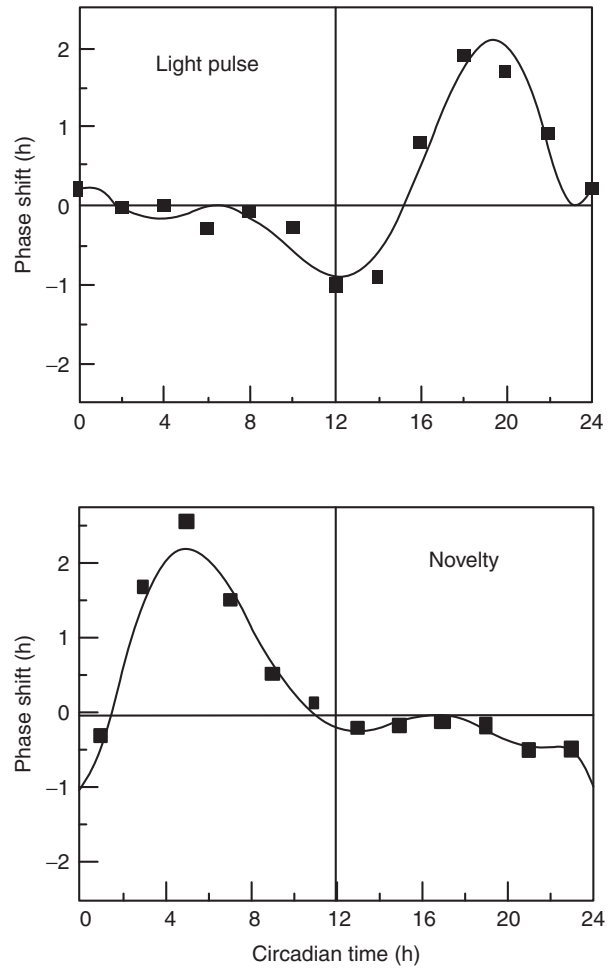

Figure 2 Phase–response curves describing the phase-dependent sensitivity of the hamster circadian pacemaker to photic (top) and nonphotic (bottom) stimulation. (Top) As seen in Figure 1, brief (here, 15 min) light pulses result in phase delays when presented near subjective dusk or early in the subjective night, while similar light pulses evoke phase advances when presented late in the subjective night. While the exact timing and relative amplitude of the delay and advance zones of the phase–response curve differ across species, these general features characterize all known photic phase–response curves in both nocturnal and diurnal animals. (Bottom) Phase advances during mid-subjective day and small phase delays during late subjective night are the result of 2 h of exposure of free-running hamsters to a novel environment containing a running wheel. Such nonphotic phase shifting is typically seen only in animals displaying signs of behavioral arousal or activation during the novelty stimulus. Reprinted from Rosenwasser AM (2003) Neurobiology of the mammalian circadian system: Oscillators, pacemakers, and pathways. In: Fluharty SJ and Grill HJ (eds.) *Progress in Psychobiology and Physiological Psychology*, vol. 18, pp. 1–38. San Diego: Elsevier Academic Press, with permission from Elsevier. Original data replotted from (top) Daan S and Pittendrigh CS (1976) A functional analysis of circadian pacemakers in nocturnal rodents. II. The variability of phase response curves. *Journal of Comparative Physiology A* 106: 253–256 and (bottom) Mrosovsky N, Salmon PA, Menaker M, et al. (1992) Non-photic phase shifting in hamster clock mutants. *Journal of Biological Rhythms* 7: 35–41.

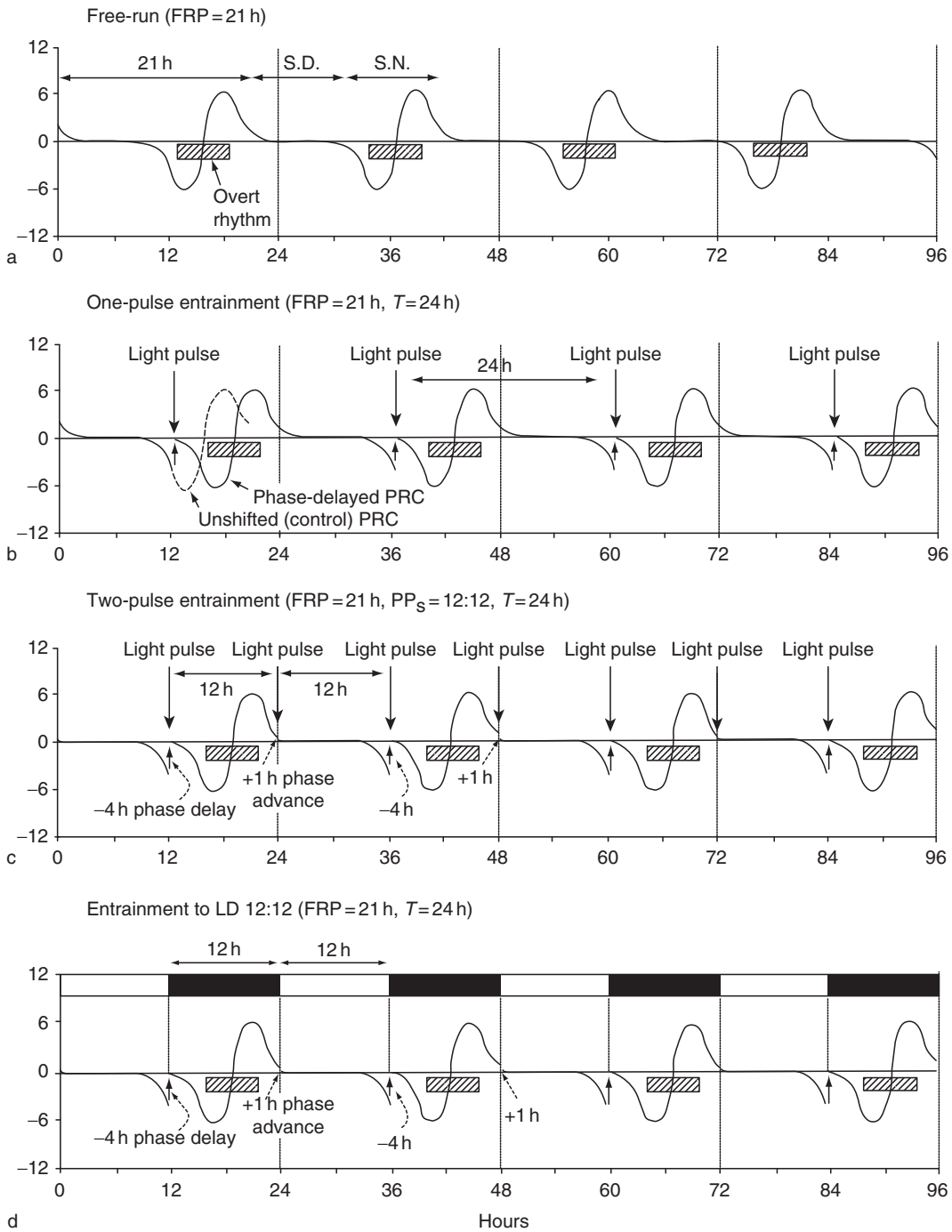

Figure 3 Light entrainment of a free-running circadian pacemaker according to the nonparametric model. In these panels, the phase of the circadian pacemaker is represented by its phase–response curve (PRC) and by the expected timing of locomotor activity. The examples shown in this figure assume a circadian pacemaker of a nocturnal organism that free-runs in constant darkness with a free-running period (FRP) of 21 h. (a) Free-running of PRC and overt rhythm with a period of 21 h. Note that the PRC and activity rhythm recur earlier each day relative to the 24 h objective day. (b) 'One-pulse' entrainment. A brief light pulse is administered to the same organism once every 24 h. During stable entrainment, the light pulse will recur at a circadian phase that results in a daily phase delay of 3 h (in this case, at approximately circadian time 13). (c) Two-pulse entrainment by a 'skeleton' photoperiod (PPs 12:12) in which a brief light pulse is administered once every 12 h. During stable entrainment, the light pulses will occur at circadian phases such that the net steady-state phase (i.e., advance plus delay) will result in a daily 3 h delay (in this case, the net 3 h delay is achieved by a daily 1 h advance combined with a daily 4 h delay). (d) Entrainment to a complete 12 h light–dark photoperiod (LD 12:12). In this case, the dawn transition results in a daily 1 h phase advance, while the dusk transition results in a daily 4 h delay, similar to the effects of the skeleton photoperiod shown in panel (c). According to this model, light occurring between dawn and dusk (i.e., in the daytime) is largely irrelevant for circadian entrainment. From Johnson CH, Elliott JA, and Foster R (2003) Entrainment of circadian programs. *Chronobiology International* 20: 741–774.

Another difference between typical laboratory light–dark cycles and natural light–dark cycles is that the natural nighttime environment includes variable amounts of starlight and moonlight, while the night phase of laboratory light–dark cycles is typically characterized by total darkness. While field studies have frequently concluded that animal activity patterns are altered by such very dim nocturnal illumination, the relevance of these studies for entrainment of the circadian pacemaker is uncertain. Recent laboratory studies have begun to compare circadian entrainment under square-wave light–dark cycles, with or without very dim nocturnal illumination approximating that of starlight or dim moonlight. Even at intensities that fail to evoke circadian phase shifting, dim nocturnal illumination, like twilight transitions, appears to increase the effectiveness of the light–dark cycle, possibly by increasing the pacemaker's sensitivity to daytime light intensities. The possible contribution of parametric modification of the underlying period of the circadian pacemaker to these effects has not been tested directly.

Physiology of Light Entrainment

Circadian Photoreception

In invertebrates and in nonmammalian vertebrates, both ocular and extraocular photoreceptors typically contribute to circadian entrainment. Vertebrate extraocular photoreceptors include both light-sensitive pinealocytes, as well as deep-brain or encephalic photoreceptors. While a number of novel and retinal photopigments have been identified within the vertebrate brain, the functional and anatomical couplings of encephalic photoreceptors to the circadian pacemaker have not been described in any vertebrate model system. In contrast, mammalian circadian photoreception appears to be entirely dependent on retinal mechanisms. Nevertheless, the photoreceptors and the central neural pathways mediating mammalian circadian entrainment comprise a distinct biological system, separate from those mediating visuoperceptual functions.

Studies with mice bearing genetic mutations leading to the progressive loss of nearly all classical rod and cone photoreceptors showed that some strains of retinally degenerate mice continue to display normal circadian photosensitivity. Eventually, similar results were obtained in genetically engineered mice completely lacking rods and cones throughout development, demonstrating unequivocally that nonclassical retinal photoreceptors are sufficient to mediate circadian entrainment. Physiological and biochemical experiments identified a novel retinal photopigment, melanopsin, and showed that this protein was specifically expressed within the small and scattered subpopulation of retinal ganglion cells giving rise to the retinohypothalamic tract (RHT), thus conferring intrinsic photosensitivity on these neurons. The term retinohypothalamic tract has been used traditionally to refer to the near-bilaterally symmetrical retinal projection to the hypothalamic suprachiasmatic nucleus (SCN), site of the primary vertebrate circadian pacemaker. Nevertheless, melanopsin-positive retinal ganglion cells also project to other diencephalic and mesencephalic targets, including other identified components of the circadian timing system (e.g., the intergeniculate leaflet), as well as neural structures mediating oculomotor reflexes and other 'nonvisual' (some authors prefer 'non-image-forming') retinal functions.

While melanopsin-positive photosensitive retinal ganglion cells are sufficient to mediate circadian light entrainment, genetic deletion of melanopsin impairs, but does not fully abolish, photic control of the circadian pacemaker. Indeed, light entrainment is abolished only when both melanopsin and classical photoreceptors are deleted, indicating a degree of redundancy in circadian photoreception. Thus, very recent studies show that photosensitive retinal ganglion cells can also respond to light via rod- and cone-mediated neural signals, and that melanopsin expression levels depend on the integrity of the classical photoreceptors. Further, genetic deletion of RPE65, a protein required for the maintenance of rod and cone photosensitivity, results in much more dramatic disruption of circadian entrainment than does elimination of the classical photoreceptors. Together, these findings indicate that classical retinal photoreceptors, while not necessary, do normally contribute significantly to circadian entrainment.

Entrainment Pathways, Neurotransmitters, and Intracellular Signaling

As mentioned earlier, the RHT projects to the SCN, intergeniculate leaflet (IGL), and other 'nonvisual' brain regions. The terminals of this projection release glutamate as the primary neurotransmitter, as well as specific peptide cotransmitters, including pituitary adenylyl cyclase-activating peptide (PACAP) and possibly substance P. Indeed, double-labeling studies of the retina demonstrate that melanopsin and PACAP provide essentially redundant markers for RHT cell bodies. Light-evoked glutamate release from RHT terminals acts through both *N*-methyl-D-aspartate (NMDA) and non-NMDA glutamate receptors to activate a number of second-messenger systems and downstream signaling pathways, leading eventually to the transcriptional changes underlying

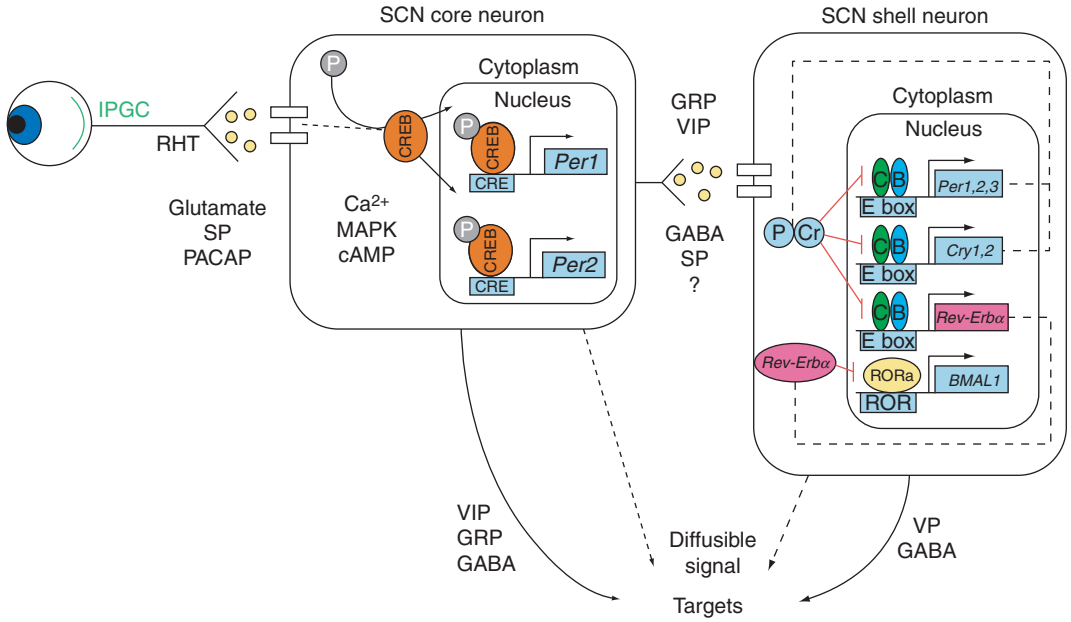

Figure 4 Cellular and molecular bases of light entrainment. Light is transduced into a neural signal by melanopsin-positive intrinsically photoreceptive ganglion cells (IPGCs) in the retina and conveyed to the suprachiasmatic nucleus (SCN) core along the retinohypothalamic tract (RHT), resulting in the release of the neurotransmitter glutamate and the neuromodulators pituitary adenylyl cyclase-activating peptide (PACAP) and substance P (SP) onto retinorecipient cells in the SCN core. Glutamate activates *N*-methyl-D-aspartate (NMDA) receptors, leading to calcium influx and activation of several known calcium-dependent signaling pathways (e.g., mitogen-activated protein kinase; MAPK), resulting in phosphorylation of cyclic AMP (cAMP) response element-binding protein (CREB). Activated CREB binds to the calcium/cAMP response element (CRE) in the promoter regions of both *Per1* and *Per2*, thus activating their transcription. Neurons in the photoresponsive SCN core then communicate with the spontaneously rhythmic SCN shell (and with SCN efferent targets) using a variety of neurotransmitters, including vasoactive intestinal polypeptide (VIP), gastrin-releasing peptide (GRP), SP, and γ-aminobutyric acid (GABA). Cells in the rhythmic SCN shell contain molecular clocks driven by a set of interlocking positive and negative autoregulatory transcription–translation feedback loops (e.g., *Per*, *Cry*, *Rev-Erba*, *BMAL1*). SCN shell neurons communicate with SCN efferent targets using vasopressin (VP) and GABA as neurotransmitters. Additionally, the SCN communicates with some target sites using a diffusible signal. Reprinted from Antle MC and Silver R (2005) Orchestrating time: Arrangements of the brain circadian clock. *Trends in Neurosciences* 28: 145–151, with permission from Elsevier.

phase shifting of the circadian pacemaker (see Figure 4). Analysis of these mechanisms is complicated, however, by the fact that the pacemaker autoregulates its own entrainment by establishing time-limited windows of sensitivity to environmental entrainment stimuli and to activation of neurotransmitter receptors and second-messenger systems. For example, the circadian pacemaker is sensitive to light-induced phase shifting mainly during the subjective night, as described by the photic PRC. Thus, light entrainment is based on a circadian rhythm in the photic sensitivity of the pacemaker, and like other circadian rhythms, rhythmic light sensitivity is controlled by mechanisms intrinsic to the pacemaker. In addition, while photic stimuli evoke similar changes in gene transcription during early and late subjective night, the signaling pathways coupling glutamate receptor activation to transcriptional activation during these two circadian time windows appear to be at least partially nonoverlapping.

Both *in vivo* and *in vitro* studies have demonstrated that light, glutamate, and/or NMDA evoke alterations in gene expression and circadian phase shifting only at night. These night-active phase-shifting stimuli appear to access the molecular clock via multiple calcium-dependent and cyclic GMP (cGMP)-dependent signaling pathways, leading to (1) phosphorylation of the transcription factor calcium and cyclic AMP response element-binding protein (CREB), (2) transcription of specific immediate-early genes (IEGs), including *c-fos*, and (3) alterations in the transcription of specific 'clock' genes such as the *period* genes (*Per1*, *Per2*), the protein products of which represent state variables encoding the oscillation phase of the circadian pacemaker. While not a primary focus of the present discussion, the circadian pacemaker can also be phase shifted and entrained by nonphotic stimuli linked to behavioral arousal and wakefulness (see Figure 2), and by agonists at specific subtypes of serotonin, γ-aminobutyric acid (GABA), and neuropeptide Y (NPY) receptors. In contrast to photic entrainment, the circadian pacemaker is sensitive to

these stimuli primarily during the subjective day, and day-active phase shifting stimuli access the molecular clock via cyclic AMP (cAMP)-dependent signaling pathways. Further, night-active and day-active phase-shifting stimuli are mutually antagonistic. Thus, while the circadian pacemaker is refractory to phase shifting by serotonin, GABA, and NPY agonists during the subjective night, these stimuli effectively block nighttime phase shifting by light and glutamate agonists, while in turn, photic stimulation during the subjective day antagonizes nonphotic phase shifting.

In addition to glutamate, RHT neurons also synthesize and co-localize the neuropeptide PACAP. Numerous studies indicate that PACAP plays a significant role in light entrainment of the circadian pacemaker, in part via modulation of glutamatergic signaling. PACAP–glutamate interactions appear to be rather complex, however, and depend on both circadian phase and PACAP levels. Early studies using the *in vitro* SCN slice preparation suggested that PACAP functions mainly as a daytime entrainment signal; despite its retinal origin, PACAP stimulation of the SCN mimicked the effects of nonphotic stimuli, producing cAMP-mediated phase advances during subjective day. While subsequent studies confirmed these observations at relatively high (micromolar) concentrations, PACAP was also shown to mimic the nighttime effects of light and glutamate on circadian phase shifting and clock gene expression at lower (nanomolar) concentrations. Further, at the higher concentrations that evoke daytime phase shifting, nighttime PACAP administration exerts complex modulation of glutamate-induced phase shifting and gene expression. Thus co-application of PACAP is reported to block glutamate-induced phase advances and *period* expression during late subjective night, but to potentiate glutamate-induced phase delays during early subjective night. Similarly, in behavioral experiments, intra-SCN PACAP mimicked – and a PACAP receptor antagonist blocked – the phase shifting effects of light during early subjective night, but PACAP was without effect during late subjective night. The role of PACAP in light entrainment has also been examined using both PACAP-deficient and PACAP receptor (PAC1)-deficient mice, but the effects of genetic deletion of the peptide and the receptor were different. Thus, PACAP-knockout mice showed decreased light-induced phase delays while PAC1-knockout mice showed increased phase delays during early subjective night. Further, while PACAP-knockout mice also showed reduced light-induced phase advances, PAC1-knockout mice actually showed abnormal light-induced phase delays during late subjective night. Finally, the PAC1-knockout mice also displayed surprising dissociation between light-induced phase shifting and gene expression, in that the potentiated early-night phase delays were accompanied by blunted light-induced *c-fos* and *period* expression.

Processing of Entrainment Signals by the SCN

The initial evidence for the anatomical heterogeneity of the SCN emerged from tract-tracing studies showing that retinal afferents to the rat SCN are generally segregated within the ventral aspect of the nucleus, especially in its more caudal aspects. Subsequently, chemoanatomical studies revealed marked regional heterogeneity in the expression and distribution of specific neuropeptides within the nucleus. Based on the distribution of retinal (and other) SCN afferents, as well as on neuropeptide expression, the rat SCN has been typically characterized as comprising two distinct regions: (1) a ventrolateral zone, defined in part by the presence of retinal terminals along with vasoactive intestinal peptide (VIP)-immunopositive and gastrin-releasing peptide (GRP)-immunopositive cells and (2) a dorsomedial zone, defined mainly by the presence of arginine–vasopressin (AVP)-positive cells. More recent studies, however, have revealed extensive species differences in the intranuclear arrangement of both afferent terminals and neuronal phenotypes, even among nocturnal rodents. In one particularly striking example, Silver and co-workers have identified a small, densely packed and centrally located cluster of calbindin (CB)-positive neurons in the hamster SCN that is not readily apparent in other species. In attempting to better accommodate such species differences, Moore has proposed an alternate view in which the SCN is composed of functionally distinct 'core' and 'shell' subnuclei, which differ across species in their exact extent and placement. Nevertheless, a two-component core and shell scheme may eventually prove insufficiently complex to account for the full range of species-related variation, or even for the functional compartmentalization of the SCN within a species. Thus, retinal, thalamic, and brain stem SCN afferents, generally described as overlapping in the ventral SCN, may exhibit somewhat different terminal fields within the SCN core, while different SCN peptides clearly display more than two distinct intranuclear distribution patterns. On the other hand, given that the SCN appears to mediate two primary circadian functions (rhythm generation and entrainment), it is appealing to hypothesize the existence of two discrete subnuclei, each associated mainly or exclusively with one of these functions.

Since retinal projections to the SCN are generally coextensive with the core subnucleus, it has frequently

been suggested that this SCN region may play a preeminent role in the photic entrainment of the circadian pacemaker (see Figure 4). Indeed, retinal afferents form direct synaptic connections with VIP-, GRP-, and CB-positive SCN core neurons. In addition, both *in vivo* and *in vitro* studies show that direct application of SCN core peptides such as VIP and GRP can mimic the cellular and phase shifting effects of light on the circadian pacemaker, but that similar effects are not seen for direct application of SCN shell peptides. Two other major SCN afferent systems converge within the SCN core to form terminal fields that overlap at least partially with the RHT: (1) a projection arising from retinorecipient neurons of the thalamic IGL, and co-localizing NPY and GABA, and (2) a projection consisting of serotonergic (and possibly nonserotonergic) axons of the midbrain raphe. Projections from the IGL to the SCN provide an indirect route for retinal signals to reach the SCN, and while disruption of this pathway fails to abolish photic entrainment of the circadian pacemaker, it does alter the effects of light on circadian period, phase, and coherence. Further, IGL lesions broadly disrupt the ability of arousal-related nonphotic cues to phase shift the circadian pacemaker, indicating that the IGL plays a major role in the integration of photic and nonphotic entrainment signals. Similarly, an important function of serotonergic raphe projections to the SCN involves arousal-dependent modulation of photic input to the circadian pacemaker, via both pre- and postsynaptic interactions with the photic signaling pathway(s).

In accord with the apparent convergence of both photic and photomodulatory signals, acute light exposure increases the expression of the immediate-early gene, *c-fos*, and of the identified clock genes, *Per1* and *Per2*, specifically within the SCN core (see Figure 4), including in VIP-, GRP-, and CB-containing cells. While some studies have also shown light-induced IEG and *period* expression in the SCN shell, such effects appear to reflect indirect signaling pathways by which photic information is relayed from light-sensitive SCN core neurons to light-insensitive SCN shell neurons. Remarkably, light evoked phase shifts of gene expression in the SCN core appear to occur almost immediately (within hours), while resynchronization of gene expression in the SCN shell tracks closely with the gradual reestablishment of a new steady-state phase at the behavioral level. Both pharmacological evidence and studies with genetically modified mice indicate that VIP and GABA signaling within the SCN are critical for conveying such signals from core to shell, and, further, for maintaining synchrony among individual SCN pacemaker neurons, probably located in both core and shell.

In contrast to the evidence supporting a preeminent role for the SCN core in processing photic entrainment signals, other studies have revealed that the spontaneous expression of cellular and molecular rhythmicity is more widespread and more robust within the SCN shell relative to the core. For example, SCN shell peptides exhibit persisting *in vivo* circadian rhythms at both mRNA and protein levels under constant darkness, while SCN core peptides do not (both core and shell peptides are consistently rhythmic under light–dark cycles). Similarly, the SCN shell displays spontaneous circadian rhythmicity in both *c-fos* and *period* gene expression while the SCN core does not. Surprisingly, however, spontaneous rhythms in normally nonrhythmic SCN core peptides can be 'unmasked' by depletion of other SCN neuromodulators, including somatostatin (generally a shell-associated peptide) and serotonin. In the SCN slice preparation, dorsal (presumably shell) SCN neurons are more likely than ventral (presumably core) SCN neurons to exhibit spontaneous circadian rhythms in firing rate, while SCN slices are more likely to display rhythmic AVP release than VIP release. It is important, however, that circadian firing rhythms may in fact be detected throughout the SCN, and that both shell and core peptides may be released rhythmically by SCN slices. Indeed, cultured SCN slices may display simultaneous circadian rhythms in both AVP and VIP release, with different free-running periods, demonstrating that both core and shell subnuclei must contain potentially autonomous oscillatory networks.

Early studies indicated that microsurgical separation of the dorsomedial (shell) and ventrolateral (core) SCN in a hypothalamic slice preparation spared electrophysiological rhythms in the ventrolateral aspect while abolishing rhythmicity in the dorsomedial aspect of the nucleus. Similarly, fetal tissue transplants appear to restore rhythmicity in SCN-lesioned arrhythmic hosts only when the transplants include surviving core-related peptides, such as VIP and/or CB. While these studies may seem to suggest that critical circadian pacemaker cells are located in the SCN core, they are probably better interpreted as confirming the necessary role for SCN core peptides in coordination and synchronization of critical pacemaker cells within the SCN core and shell. Thus, a reasonable hypothesis for photic information processing within the SCN is that retinorecipient light-entrainable cellular oscillators in the SCN core provide neural signals that are necessary for the entrainment and synchronization of light-insensitive cellular oscillators in the SCN shell. According to this view, much of the dynamic complexity of circadian light entrainment revealed

by formal analysis at the behavioral level can be accounted for by interactions among functionally distinct SCN subnuclei.

Conclusion

Analysis of the mechanisms by which photic signals entrain the circadian pacemaker has provided a rich model system in chronobiology, and more generally, in behavioral neuroscience. Classic and contemporary behavioral analyses have revealed several general quantitative principles, the most powerful of which is the phase-response curve, a complex yet predictable function that characterizes the phase-dependent responsiveness of the circadian pacemaker to light. Recently, an integrative, multidisciplinary approach incorporating physiological, pharmacological, anatomical, and molecular approaches has successfully pried open the 'black box' formalism of PRC analysis and has provided a robust biological understanding of the photic entrainment pathway. These basic scientific studies await further exploitation as applied chronobiologists develop practical interventions that may be effectively deployed in the service of public health and safety.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Organization in Non-Mammalian Vertebrates; Circadian Rhythms: Influence of Light in Humans; Photoreceptor Adaptation; Photoreceptors and Circadian Clocks; Photoreceptors: Physiology.

Further Reading

Antle MC and Silver R (2005) Orchestrating time: Arrangements of the brain circadian clock. *Trends in Neurosciences* 28: 145–151.

Aschoff J (1960) Exogenous and endogenous components in circadian rhythms. *Cold Spring Harbor Symposia on Quantitative Biology* 25: 11–28.

Berson DM, Dunn FA, and Takao M (2002) Phototransduction by retinal ganglion cells that set the circadian clock. *Science* 295: 1070–1073.

Boulos Z, Macchi MM, and Terman M (2002) Twilights widen the range of photic entrainment in hamsters. *Journal of Biological Rhythms* 17: 353–363.

Daan S and Aschoff J (2001) The entrainment of circadian systems. In: Takahashi JS, Turek FW, and Moore RY (eds.) *Handbook of Behavioral Neurobiology, Vol 12: Circadian Clocks*, pp. 7–43. New York: Kluwer Academic/Plenum.

Daan S and Pittendrigh CS (1976) A functional analysis of circadian pacemakers in nocturnal rodents. II. The variability of phase response curves. *Journal of Comparative Physiology A* 106: 253–256.

Foster RG and Hankins MW (2002) Non-rod, non-cone photoreception in the vertebrates. *Progress in Retina and Eye Research* 21: 507–527.

Gorman MR, Evans JA, and Elliott JA (2006) Potent circadian effects of dim illumination at night in hamsters. *Chronobiology International* 23: 245–250.

Hannibal J (2006) Roles of PACAP-containing retinal ganglion cells in circadian timing. *International Review of Cytology* 251: 1–39.

Hattar S, Lucas RJ, Mrosovsky N, et al. (2003) Melanopsin and rod–cone photoreceptive systems account for all major accessory visual functions in mice. *Nature* 424: 76–81.

Johnson CH, Elliott JA, and Foster R (2003) Entrainment of circadian programs. *Chronobiology International* 20: 741–774.

Lee HS, Billings HJ, and Lehman MN (2003) The suprachiasmatic nucleus: A clock of multiple components. *Journal of Biological Rhythms* 18: 435–449.

Meijer JH and Schwartz WJ (2003) In search of the pathways for light-induced pacemaker resetting in the suprachiasmatic nucleus. *Journal of Biological Rhythms* 18: 235–249.

Mistlberger RE and Rusak B (2005) Circadian rhythms in mammals: Formal properties and environmental influences. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, 4th edn., pp. 1–30. Philadelphia: Elsevier-Saunders.

Morin LP and Allen CN (2006) The circadian visual system, 2005. *Brain Research: Brain Research Reviews* 51: 1–60.

Mrosovsky N, Salmon PA, Menaker M, et al. (1992) Non-photic phase shifting in hamster clock mutants. *Journal of Biological Rhythms* 7: 35–41.

Pittendrigh CS and Daan S (1976) A functional analysis of circadian pacemakers in nocturnal rodents IV: Entrainment: Pacemaker as clock. *Journal of Comparative Physiology A* 106: 291–331.

Rosenwasser AM (2003) Neurobiology of the mammalian circadian system: Oscillators, pacemakers, and pathways. In: Fluharty SJ and Grill HJ (eds.) *Progress in Psychobiology and Physiological Psychology,* vol. 18, pp. 1–38. San Diego: Elsevier Academic Press.

Ependymal Cells

H Wolburg, K Wolburg-Buchholz, and A F Mack, University of Tübingen, Tübingen, Germany
A Reichenbach, University of Leipzig, Leipzig, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

As early as 1837 Valentin and Purkinje visualized the lining of the inner cavities of the brain, the brain ventricle system, by the observation of striking cilia on its surface. This lining was called 'ependyma' (Greek meaning 'upon garment'), and its constituent cells are termed ependymal cells. The ependymal cells belong to the family of neuroglial cells (glia) and are thus also called ependymoglial cells or ependymal glia. After its early discovery, it required many decades of research to elucidate this cellular layer(s) as an extremely important component of the developing and mature brain.

During ontogenetic development, the ventricular lining is composed of proliferating cells termed the 'matrix zone,' which generates all types of neuroectodermal brain cells, including neurons and their 'local successors,' the ependymoglial cells. In the adult central nervous system (CNS), the ependymoglial cells contribute to the secretion of the cerebrospinal fluid (CSF) and are an essential component of an important biological barrier system segregating different compartments such as blood, CSF, and CNS interstitial space (blood–brain and blood–CSF barriers; Figure 1). Originally, Paul Ehrlich found that a blood-infused dye failed to stain the brain tissue, and his pupil, Ernst Goldmann, observed complementarily that the very same dye, if applied into the CSF, did stain the brain tissue. This led to the concept of a biological barrier between blood and brain. Due to the free access of the dye from brain ventricle to brain tissue, they concluded that there is no CSF–brain barrier. However, the staining of circumventricular organs and the choroid plexus in Goldmann's experiment applying the dye into the general circulation (Goldmann-I experiment) and the avoidance of staining of these organs in the experiment applying the dye into the CSF (Goldmann-II experiment) suggested the existence of a barrier between the CSF and the blood (Figure 1). The cellular basis of these barriers has been located within the endothelium (endothelial blood–brain barrier in most vertebrates; only in elasmobranchs (sharks and rays) is the blood–brain barrier located in astrocytes), in the epithelial choroid plexus cells, and in the tanycytes of the circumventricular organs (glial blood–CSF barrier).

Definition of Ependymoglial Cells

Macroglial cells, including oligodendrocytes, astrocytes, and ependymoglial cells, may be classified according to the shapes and contacts of their cell processes. Distinctively, only ependymoglial cells establish contact with the ventricular surface (or, in the case of Müller glial cells, with the subretinal space, a remnant of the embryonic ventricle system). Astroglial cells are consistently characterized by the fact that at least one of their processes bears endfeet contacting a basal lamina around blood vessels (glia limitans perivascularis), the pia mater (glia limitans superficialis), or both. Ependymoglial cells share this property; opposite to their ventricular processes, these bipolar cells establish endfeet at a basal lamina (the majority of the subependymal blood vessels become obliterated in later stages of embryogenesis and most of the ependymocytes contact the remaining 'basement membrane labyrinths' rather than true blood vessels or the pia). Generally, we have to distinguish between these 'normal' cilia-bearing ependymocytes lining the major part of the ventricular surface and the choroid plexus epithelial cells and the tanycytes (Figure 1). Both of the latter cell types form the blood–CSF barrier to avoid free diffusion between blood and CSF. This is essentially due to the absence of an endothelial barrier: in the choroid plexus, the endothelium has to be permeable in order to allow the formation of the CSF from the blood by the plexus epithelial cells (Figures 1–3). In the circumventricular organs and in the hypothalamus–hypophyseal system, the vessels are permeabilized in order to receive neurosecretory input from the neurons and to deliver signal compounds from the blood to the neurons (Figures 2 and 3).

Morphology of Ependymoglial Cell Types

The morphology of ependymoglia is diverse (Figure 4). Much of this diversity is related to structural and functional interactions of a given cell with its microenvironment, which includes blood vessels, the pia mater, and/or the ventricular space. Where macroglial cells form a 'border sheath' against the ventricular space, pia, or blood vessels, they form epitheloid aggregates. This is observed in ependymocytes, choroid plexus cells, and retinal pigment epithelial cells which also belong to ependymoglial cells. The term 'radial glia' should be restricted to bipolar ependymoglial

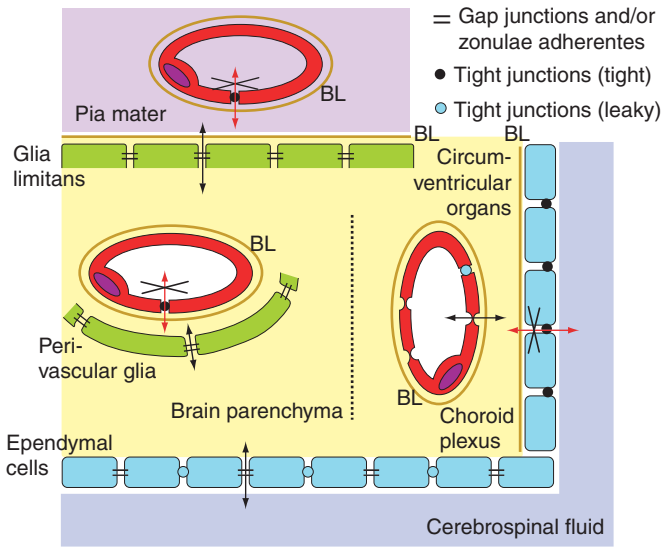

Figure 1 Schematic drawing of the topology of the compartments within the central nervous system (CNS). Within the CNS, there is the neural compartment containing neurons and the neuropil, glial cells, and the vasculature consisting of endothelial cells surrounded by a basal lamina, pericytes, and astroglial endfeet. At the outer surface of the CNS, superficial astroglial endfeet interconnected by gap junctions form the glial limiting membrane, a border toward the meningeal cells which, however, is no physiological barrier. The blood vessels within the pia mater have blood–brain barrier properties. At the inner surface of the CNS, ependymal cells line the ventricular space. Here as well, they have no barrier properties but are interconnected by gap junctions and poorly developed, leaky tight junctions. Due to the fact that the cerebrospinal fluid (CSF) is produced by the plexus epithelium from the blood, there is no barrier between blood and choroid plexus epithelium; however, between blood and cerebrospinal fluid, there is the so-called blood–cerebrospinal fluid barrier. This barrier is formed by tight junctions between choroid plexus epithelial cells and tanycytes in the circumventricular organs. The endothelial cells in these organs and in the choroid plexus are highly permeable and fenestrated.

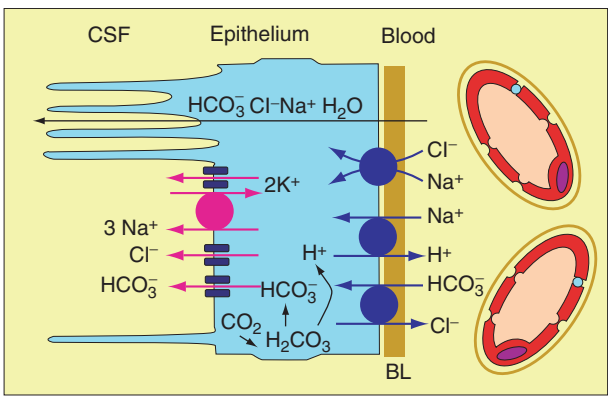

Figure 2 Highly schematic view of the equipment of a choroid plexus epithelial cell with transporter molecules and the Na^+, K^+-ATPase. Prerequisite for cerebrospinal fluid (CSF) production is the fenestration of the choroid plexus blood vessels.

cells that extend long processes throughout the thickness of the tissue. Early in embryonic development, the immature brain (i.e., the wall of the ventricular system) is spanned by a scaffold of many fetal radial glial cells. In the mature CNS, the descendants of these cells are further differentiated into multipolar astrocytes, or they are maintained as tanycytes in the circumventricular organs or as Müller cells in the retina (Figure 4).

Tanycytes

Tanycytes are a common type of macroglia in the CNS of lower vertebrates (and even deuterostomic invertebrates). In adult mammals, they are restricted to certain brain regions where the tissue is thin, such as some circumventricular organs (e.g., the subcommissural organ), the stalk of the pituitary, and the velum medullare, and to the raphe region of the spinal cord. In the circumventricular organs of all vertebrates except sharks, the capillaries are fenestrated; in these regions, the tanycytes (as well as the choroid plexus cells) constitute the blood–CSF barrier by expressing an extensive apical tight junctional network (Figure 3(f)). Some of these tanycytes are specialized for the secretion of signaling molecules and the material constituting the Reissner's fiber.

Ependymocytes, Choroid Plexus Cells, and Retinal Pigment Epithelial Cells

Ependymocytes, choroid plexus cells, and retinal pigment epithelial cells are specialized glial cells lining the ventricle (or the subretinal space, respectively; Figure 4). At their basal pole, most mature ependymocytes contact remnants of embryonic blood vessels (so-called basement membrane labyrinths; Figure 4)

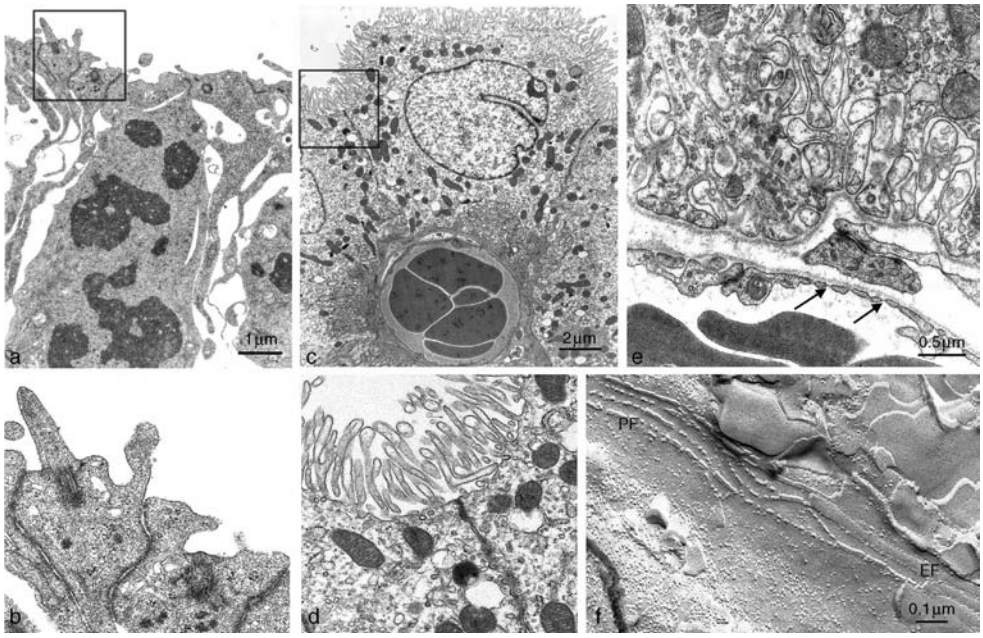

Figure 3 Electron micrographs of the ependyma and choroid plexus. (a) Ultrathin section through a mitotic ependymal cell of a prenatal mouse brain showing the proliferative property of this periventricular matrix zone. To the left of this mitotic cell, a differentiated ependymal cell already has developed a cilium. The framed box is shown in b, showing the cilium with the basal apparatus and the tight and adhesion junctions interconnecting this cell with its neighbors. (c) Ultrathin section of choroid plexus with epithelial cells (top) and a blood vessel (bottom). The framed box of the apical region of the epithelial cell is magnified in d, showing the microvilli and the junctional domain between two cells. (e) Ultrathin section through the basal labyrinth of a choroid plexus epithelial cell underlined by a basal lamina. Below this basal lamina, the fenestrated endothelium of a blood vessel can clearly be recognized. Arrows indicate fenestrations. (f) Freeze-fracture replica of the apical tight junction of a choroid plexus epithelial cell constituting the blood–cerebrospinal fluid barrier and consisting of parallel strands mainly associated with the inner (protoplasmic) leaflet of the membrane (P-face (PF)). EF, E-face, the outer (external) leaflet of the membrane.

rather than intact blood vessels. At their other pole (i.e., the ventricular pole), they possess, in addition to microvilli, kinocilia to support the movement of CSF (Figures 3(a), 4, and 5(c)). The latter is mainly secreted by the choroid plexus cells, characterized by a high density of Na^+, K^+-ATPase molecules at their microvillous membrane and a high concentration of the water channel protein aquaporin-1 (Figures 2 and 5(d)). This secretion requires a high permeability of the fenestrated plexus endothelial cells (Figure 3(e)), which is now known to be induced by the delivery of the vascular endothelial growth factor (VEGF). VEGF is recognized by the endothelial VEGF receptor followed by a signal cascade leading to the formation of fenestrations. To prevent the diffusion of blood-borne substances into the ventricle, the plexus epithelium is equipped with a densely meshed tight junctional apparatus (Figure 3(f)) constituted by the zonula occludens protein 1 (ZO-1), occludin, and claudin family members 1, 2, and 11 (Figure 5) and also probably by unknown molecules.

Retinal pigment epithelial (RPE) cells reveal features which are related to those of choroid plexus epithelial cells. They line the subretinal space opposite to the neuroretina. Their apical surface forms two types of microvilli: (1) long (5–7 µm) thin microvilli maximizing the membrane area available for transepithelial transport and (2) specially arranged shorter microvilli termed photoreceptor sheaths. The basal surface of RPE cells contains numerous invaginations to increase the surface area. Like the choroid plexus cells, RPE cells are in close apposition to many blood vessels, secrete VEGF in order to induce fenestrations in the choroid vessels, are specialized for transmembrane transport, and form the blood–CSF (or, in this case, –subretinal) fluid barrier by their tight junctions. However, there is an important difference between RPE and plexus epithelial cells (and all other ependymoglial cells): RPE cells do not secrete fluid across their apical microvillous membrane but, rather, perform a net fluid uptake from the subretinal space. This water resorption is mediated via Na^+/bicarbonate$^-$ exchange and prevents (together with the intraocular pressure and with cellular adhesion molecules) a detachment of the neuroretina from the pigment epithelium.

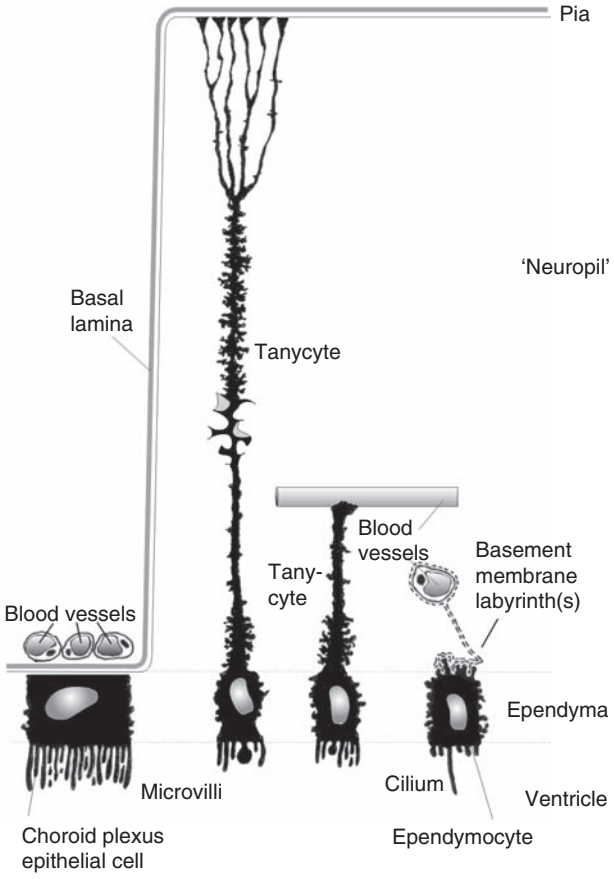

Figure 4 Semischematic survey of the main types of ependymoglial cells and their localization in different regions of the central nervous system. The situation shown at left is also true for the pigment epithelial cells of the eye. The left tanycyte can also be called radial glial cell due to contact to both the pial surface of the brain and the brain ventricle. Another example of this type of tanycyte is the retinal Müller (glial) cell: in this case, the basal lamina corresponds to the inner limiting membrane of the retina. The microvilli-bearing lower part is oriented toward the pigment epithelial microvilli within the subretinal space which is equivalent to the ventricle of the early eye cup. The right tanycyte is the vascular tanycyte, which is the site of the blood–cerebrospinal fluid barrier in the circumventricular organs. The classical ependymocytes line the ventricle and bear cilia on their surface which provide for the steady movement of the cerebrospinal fluid.

Marker Molecules of Ependymoglial Cells

Ependymoglial cells can be visualized by immunocytochemical labeling of certain antigens that are, at least within the CNS, restricted to these cells. The expression of these molecules by certain cell types may change with differentiation or during pathological processes (Table 1). Furthermore, not all members of a given cell population (otherwise considered homogeneous) must express the same antigen (at detectable levels). Table 1 provides an overview of antigens commonly identified by immunolabeling procedures on specific types of astro- and ependymoglial cells.

Ultrastructural Features of Ependymoglial Cells

The somata of some types of ependymoglial cells contain conspicuous melanin pigment granula (RPE cells and choroid epithelial cells). In most of these epitheloid glial cells (including the ependymocytes), the lateral membranes of the somata are interconnected by zonulae adherentes and tight junctions, thus forming the blood–brain (or –retina) barrier (Figures 3 and 5).

In ependymoglial cell nuclei, the nucleoplasm is rather evenly distributed compared to that in oligodendrocytes and microglial cells. In some ependymoglial cells, such as in many tanycytes, the cell nuclei are very irregularly shaped and may display deep incisions. The nuclei (and somata) of Müller cells seem to be 'indented' by neighboring neurons; atomic force microscopy has shown that Müller cell somata are 'more soft' (i.e., possess a lower module of elasticity) than the somata of the neighboring bipolar neurons.

Stem processes are those cellular processes which directly arise from the soma. Typically, they contain bundles of intermediate filaments. Particularly high densities of intermediate filaments are found in the basal (i.e., endfoot-bearing) processes of tanycytes and Müller cells. The stem processes usually contain numerous mitochondria. An interesting exception is Müller cell processes in species with avascular retinae which contain mitochondria only at their apical pole (i.e., close to the choroid, which is the only source of oxygen supply), whereas their stem processes are devoid of these organelles. It has been argued that due to their dominant glycolytic energy metabolism, Müller cells are free to move and place their mitochondria toward sites of high pO_2 rather than toward sites of high energy demand as observed in the neurons with their dominant aerobic metabolism.

The stem processes of tanycytes and Müller cells do not show the regular dichotomic branching pattern characteristic of neuronal dendrites but, rather, are the origins of specialized endings or side branches described in the following sections.

Ependymoglial and astrocytic endfeet cover almost all basal laminae within the CNS (along the blood vessels, the pia mater, and the vitreous body in the eye). In the case of tanycytes and Müller cells, the endfeet are densely filled with smooth endoplasmic reticulum. Bundles of intermediate filaments extend into the endfeet but fail to occupy the cytoplasm close to the basal lamina-contacting endfoot membrane. These filaments consist primarily of vimentin when the endfoot is in contact with CSF or vitreous humor and of glial fibrillary acidic protein (GFAP) when a blood vessel is contacted; however, this rule may be modified in some cases. With the exception of Müller

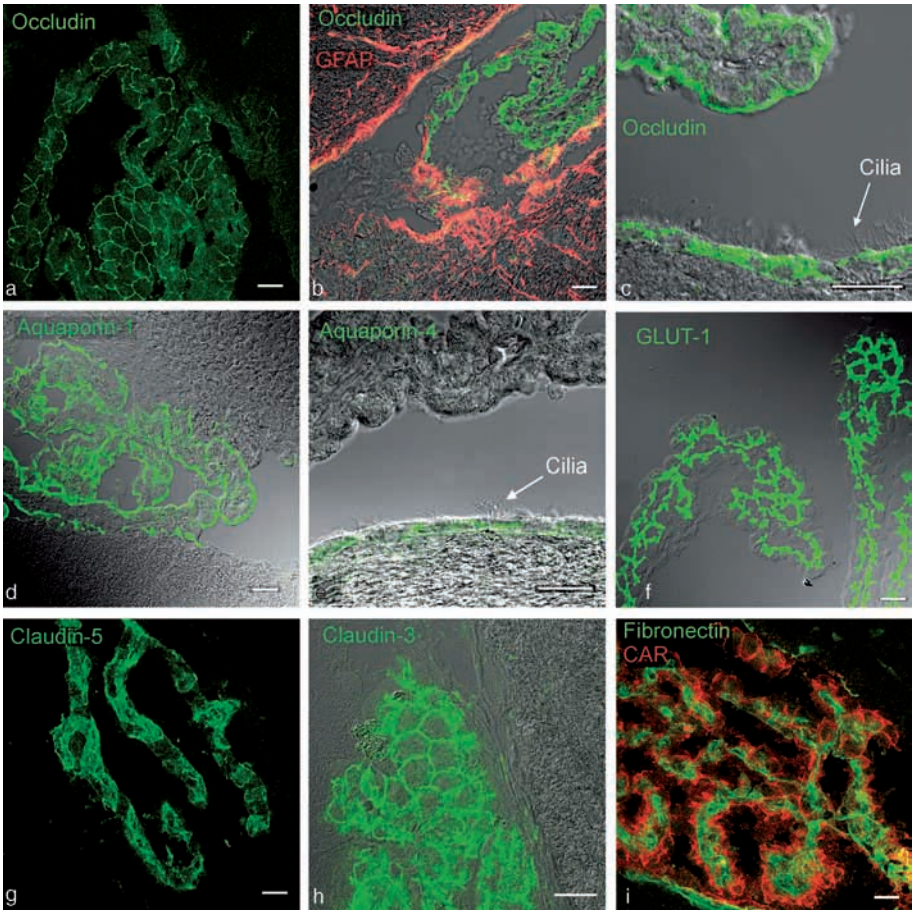

Figure 5 Confocal laser scanning images of the ependyma and the choroid plexus. (a) Occludin was the first known tight junctional component within the membrane. It is very well expressed in the epithelial cells of the choroid plexus. (b) Transitional zone between ependyma and choroid plexus. GFAP (red) is present in astrocytes and ependymal cells but not in choroid plexus epithelial cells, whereas the tight junctional protein occludin (green) is expressed in choroid plexus epithelial cells only and not in astrocytes. However, as shown in (c), occludin is expressed in both ependymal cells (as recognized by cilia) and plexus epithelial cells (top). (d) The choroid plexus epithelium specifically expresses the water channel protein aquaporin-1 (AQP1), which is not found in ependymal cells. (e) Ependymal cells, together with astrocytes, synthesize AQP4, which in turn is not found in the choroid plexus (top). (f) Glucose transporter-1 isoform (GLUT-1) immunostaining. Where the paracellular route is occluded by highly resistant tight junctions as in plexus epithelial cells, glucose has to be transported by specific transporter molecules such as the sodium-independent GLUT-1. Because the ependymal cells have only leaky (occludin-positive) tight junctions (c), the glucose transporter is not needed in ependymal cells (not shown). (g and h) Within the choroid plexus, the endothelial tight junction molecule claudin-5 is restricted to blood vessels, whereas the tight junctions of epithelial cells contain, besides claudin-2 and -11 (not shown), claudin-3 (h). (i) The epithelial and endothelial basal laminae of the choroid plexus can be recognized by antibodies against fibronectin. The epithelial tight junctions can be visualized not only by classical tight junctional components such as occludin and claudins but also by the coxsackievirus–adenovirus receptor (CAR). Scale bar = 20 μm.

cells in species with completely avascular retinae (i.e., missing both intra- and supraretinal blood vessels), the endfeet are rich in mitochondria. The occurrence of caveolae, coated pits, and vesicles concerned with endo-, exo-, or pinocytosis indicates active material exchange with the compartment on the other side of the basal lamina (i.e., blood plasma, vitreous body, or subarachnoidal fluid). A secretory function has been ascribed to the tanycytes of some circumventricular organs.

Enigmatic membrane structures of ependymoglial cells are the so-called orthogonal arrays of intramembranous particles (OAPs), which have been most extensively described in astrocytes. There, they are concentrated in membrane domains that directly contact a basal lamina, particularly the perivascular basal lamina. In periventricular ependymocytes which do not possess a basal lamina, OAPs occur across the whole cellular surface with the preference of the basolateral membrane. Interestingly, the existence of OAPs and tight junctions is inversely correlated: the ependymoglial cells have OAPs but no or less developed tight junctions, whereas the choroid plexus epithelial cells have well-developed tight junctions but no OAPs. It is known that OAPs are mainly composed of the water channel protein aquaporin-4

Table 1 'Marker antigens' suitable to visualize and/or to identify the various types of ependymoglial cells during ontogenetic development, in the normal mature CNS, and during reactive changes in cases of pathology

Cell type	*Antigen*	*Developing*	*Adult*	*Reactive*
Ependymoglia	GFAP	(+)	++	
	Vimentin		++	
	Cytokeratin		+++	
	RAN-2		++	
	Occludin		++	
	AQP4		++	
RPE	Vimentin		++	
	CRALBP		+++	
	R-cadherin[C]	++	–	
	Occludin	+	+++	
	GLUT-1		+++	
Choroid plexus epithelium	Cytokeratin		++	
	GFAP		–	++
	Vimentin		++	
	Neurofilament*		++	
	Occludin		+++	
	Claudin-2,-3,-11		+++	
	CAR		+++	
	AQP1		+++	
	GLUT-1		+++	

Whereas the table basically reflects the situation in mammals, many of the antigens can also be found in the corresponding cell types of other vertebrates. In cases in which an antigen is only found in nonmammalian cells, it is labeled by an exposed letter: C, chicken. An asterisk indicates that the listed antigen (or antibody, respectively) does not label selectively glial cells but (in other regions of the CNS) also labels neurons. Antigens expressed in the cytoplasmic membranes, such as ion channels, receptors, or adhesion molecules, are excluded since immunocytochemistry for these antigens usually results in 'diffuse' labeling of the neuropil at the light microscopical level. However, tight junctional molecules are included (**Figure 4**).
AQP1/4, water channel proteins aquaporin-1 and -4; BLBP, brain lipid-binding protein; CAR coxsackievirus–adenovirus receptor; CRALBP, cellular retinaldehyde-binding protein; GFAP, glial fibrillary acidic protein; GLUT-1, sodium-independent glucose transporter isoform 1; RAN-2, rat neural antigen-2.

(AQP4). In astrocytes, the role of OAP/AQP4 for water homeostasis of the brain and for maintenance of the blood–brain barrier has been widely discussed; in periventricular cells, the role of OAP/AQP4 may be connected to osmotic regulation as well.

The development of the OAP/AQP4-related polarity of astrocytes seems to correlate with the time of agrin expression. Agrin is a heparan sulfate proteoglycan of the extracellular matrix present on ependymoglial endfeet. It has been shown to be essential for the clustering of acetylcholine receptors in the postsynaptic membrane of the motor endplate. Agrin also binds to α-dystroglycan, a member of the dystrophin–dystroglycan complex which localizes at glial endfoot membranes in a similar manner as OAP/AQP4. In mice with an induced deletion of the gene encoding Dp71 (i.e., the Müller cell-specific member of the dystrophins), a redistribution of AQP4 and Kir4.1 in the Müller cell membrane was observed. Kir4.1 is a weakly inward rectifying potassium channel involved in the glial buffering capacity of extracellular K^+ ions. These and other data support the hypothesis that agrin and dystrophins cooperate in modulating the OAP/AQP4- (and Kir4.1-) related polarity of astrocytes and ependymoglial cells, such as retinal Müller cells and tanycytes.

When, under pathological or experimental conditions, apicolateral membranes of RPE cells, Müller cells, or ependymocytes are confronted with mesenchyma, they lose their original features and develop an endfoot-like structure. This response suggests that mesenchymal contact (i.e., collagen/laminin/agrin and/or other molecules) stimulates the insertion of Kir4.1 K^+ channels and the production of a basal lamina by adjacent mesenchymal cells.

Ventricular Contacts

Cell processes contacting the ventricular (or subretinal) space occur in ependymoglial cells such as tanycytes and Müller cells (but not in astrocytes). They always display an enlarged surface area by extending microvilli into the fluid volume; furthermore, the apical pole contains abundant mitochondria. Both features are indicative of a high metabolic activity that is presumably related to active exchange of substances with the luminal fluid.

Neighboring glial ventricular contact processes (and adjacent neuronal cell processes, if present) are connected by various types of apicolateral junctions. These (particularly desmosomes) are general 'markers' of virtually all epithelial cells and occur very early in development. However, their nature varies considerably depending on the local microenvironment. In regions where no endothelial blood–brain barrier exists (e.g., in most circumventricular organs and in the RPE), ependymoglial cells form a CSF–brain barrier by expressing tight junctions. Apicolateral gap junctions normally occur between retinal Müller cells in frogs but not in mammals. However, when rabbit Müller cells form homogeneous cultures *in vitro*, gap and even tight junctions can be observed, suggesting that the retinal microenvironment *in situ* inhibits the establishment of such intercellular contacts. RPE cells proliferating in areas of retinal detachment may lose their basal lamina contact (Bruch's membrane) and face the subretinal fluid around their entire surface. In these cases, they lose their membrane polarity.

The apical surface of typical ependymocytes is characterized by the presence of 12–60 kinocilia, which vary in number according to the species. The cilia are 10–20 μm long and are of the $9+2$ type. This term means that the cilia contain an axoneme structure consisting of two centrally arranged single microtubules and nine circularly arranged doublets of microtubules. These cilia beat rhythmically at a frequency of approximately 200 beats per minute, and they appear to assist the rostrocaudad flow of CSF.

Choroid Plexus and Tanycytes as the Site of the Blood–Cerebrospinal Fluid Barrier and Immunological Aspects of It

As mentioned previously, the choroid plexus epithelium and the tanycytes of the circumventricular organs form a physiologically important barrier between the CSF and the blood. This is necessarily a consequence of the facts that (1) the CSF has free access to the brain parenchyma through the leaky lining of ependymocytes and (2) the neurosecretory cells of the circumventricular organs and the tuberohypophyseal system need the direct contact with leaky blood vessels in order to release their hormones into the vasculature. Likewise, plexus epithelial cells need the direct contact with leaky blood vessels to produce the CSF from the blood. The intense interplay between neurons, ependymoglial cells, and blood vessels may be demonstrated by the observation that hormonal or osmotic stimulation of neurons in the tuberohypophyseal system leads to changes in the covering of blood vessels by tanycytic endfeet, the ultrastructure of which changes with the physiological state. These changes modify the interface area available for neurohumoral secretion. In rhesus monkeys, the structure of hypothalamic tanycytes differs between males and females; moreover, the apical protrusions of tanycytes in females change their size and number depending on the estrous cycle.

Much of the morphological diversity of astrocytes and ependymoglial cells results from the different local microenvironments into which a given cell is born (or migrating). Inevitably, mesenchymal contact will induce the formation of endfeet with OAP-rich membranes, whereas contact with CSF will induce the outgrowth of microvilli and the formation of stabilizing cell–cell junctions. Where neuronal elements are contacted, the glial cells form delicate side branches that end in lamellar sheaths or fingerlike branchlets. The number, size, and shape of these glial 'end structures' are precisely adjusted to the morphological and functional features of the adjacent neuronal elements. This adjustment continues after ontogeny as a lifelong process of plasticity.

A role of the choroid plexus in the pathogenesis of multiple sclerosis (MS) or its animal model experimental allergic encephalomyelitis (EAE) – that is, as an alternative entry site for circulating lymphocytes directly into CSF – has not been seriously considered. The morphological correlate of the blood–CSF barrier is the tight junctions of choroid plexus epithelium. Interestingly, in EAE the adhesion molecules intercellular adhesion molecule-1 (ICAM-1) and vascular cell adhesion molecule-1 (VCAM-1) are upregulated on the choroid plexus epithelium in parallel with their upregulation on blood–brain barrier endothelium. Ultrastructural studies have revealed a polar localization of ICAM-1, VCAM-1, and mucosal addressin cell adhesion molecule-1 (MAdCAM-1) on the apical surface of choroid plexus epithelial cells and their complete absence on the fenestrated endothelial cells within the choroid plexus parenchyma. Furthermore, ICAM-1, VCAM-1, and MAdCAM-1, expressed in choroid plexus epithelium, mediate the binding of lymphocytes via their known ligands.

Massive ultrastructural changes can be observed during EAE within the choroid plexus; however, at the ultrastructural level the tight junctions seem to remain intact. Choroid plexus epithelial tight junctions are morphologically unique (resembling only those of the central myelin sheaths) because they are characterized by their parallel and poorly anastomosing strands (**Figure 3(f)**). This specific type of tight junctions of CNS myelin sheaths was found to be exclusively constituted by oligodendrocyte-specific protein (OSP), which has been identified as a member of the claudin family (claudin-11). A possible role for OSP/claudin-11 in the development of EAE was

demonstrated by the induction of the disease in susceptible mice after immunization with OSP/claudin-11 and by the demonstration of antibodies directed against OSP/claudin-11 in the CSF of patients with MS. Although the presence of anti-claudin-11 antibodies in the CSF of MS patients is still taken as a measure for myelin degradation, it might well reflect a modulation of the blood–CSF barrier localized at the choroid plexus epithelium.

In addition, because choroid plexus epithelial cells have been found to express major histocompatibility complex class I and II molecules on their surface, it will be interesting to determine the contribution of this barrier to autoimmune inflammation within the CNS.

Developmental Aspects of the Ependymoglial Cells

Since the nineteenth century, it has been recognized that the periventricular zone is a matrix zone of growth and differentiation. There was a long-standing debate on whether this matrix contains segregated pools of glial and neuronal progenitors or common progenitors for macroglial cells and neurons. Data suggest that even in adult mammals neural stem cells in the matrix zone generate precursor cells of both the neuronal and the glial lineage, and that the determination of the cleavage plane during mitosis may also determine the fate of the stem cells. Considerable effort is being spent to understand the extremely complex molecular mechanisms that control the differentiation processes. Apparently, there is a complex balance between neurogenic and gliogenic factors and their receptors, such as neurogenin, notch receptor, bone morphogenetic proteins, or leukemia inhibitory factor receptor, which decide together (and/or under the control of the transcription factor Pax6) which one eventually determines the cell fate for a given postmitotic cell. Similarly, the radial glial (Müller) cells in the retina are considered as a potential source of neural regeneration after damage of retinal neurons, at least in the chicken. Again, Pax6 and other transcription factors play an important role in this glio-neuronal transdifferentiation process.

See also: Bergmann Glial Cells; Glial Growth Factors; Macroglial Lineages; Oligodendrocyte Morphology; Retinal Pharmacology: Inner Retinal Layers; Retinal Glia; Retinal Development: Cell Type Specification.

Further Reading

Campbell K and Götz M (2002) Radial glia: Multi-purpose cells for vertebrate brain development. *Trends in Neuroscience* 25: 235–238.

Doetsch F (2003) The glial identity of neural stem cells. *Nature Neuroscience* 6: 1127–1134.

Engelhardt B (ed.) (2001) The choroid plexus in health and disease. *Microscopy Research and Technique* 521.

Fischer AJ and Reh TA (2001) Müller glia are a potential source of neural regeneration in the postnatal chicken retina. *Nature Neuroscience* 4: 247–252.

Götz M and Huttner WB (2005) The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology* 6: 777–788.

Hatton GI (1997) Function-related plasticity in hypothalamus. *Annual Review of Neuroscience* 20: 375–397.

Kröger S (1997) Differential distribution of agrin isoforms in the developing and adult avian retina. *Molecular and Cellular Neuroscience* 10: 149–161.

Leonhardt H (1980) Ependym und circumventriculäre Organe. In: Oksche A (ed.) *Neuroglia I,* vol. 4, part 10 of *Handbuch der mikroskopischen Anatomie des Menschen* (Oksche A and Vollrath L, series eds.), pp. 177–666. Berlin: Springer.

Reichenbach A (1989) Attempt to classify glial cells by means of their process specialization using the rabbit retinal Müller cell as an example of cytotopographic specialization of glial cells. *Glia* 2: 250–259.

Reichenbach A and Robinson SR (1995) Phylogenetic constraints on retinal organization and development: An Haeckelian perspective. *Progress in Retinal Research* 15: 139–171.

Reichenbach A and Wolburg H (2005) Astrocytes and ependymal glia. In: Ransom BR and Kettenmann H (eds.) *Neuroglial Cells*, pp. 19–35. Oxford University Press: Oxford.

Sarthy V and Ripps H (2001) *The Retinal Müller Cell. Structure and Function.* New York: Kluwer Academic/Plenum.

Saunders NR and Dziegielewska KM (eds.) (2000) Barriers in the brain. *Cellular and Molecular Neurobiology* 20: 1–2.

Venero JL, Vizuete ML, Machado A, and Cano J (2001) Aquaporins in the central nervous system. *Progress in Neurobiology* 63: 321–336.

Wolburg H (1995) Orthogonal arrays of intramembranous particles. A review with special reference to astrocytes. *Journal of Brain Research* 36: 239–258.

Wolburg H and Lippoldt A (2002) Tight junctions of the blood–brain barrier: Development, composition and regulation. *Vascular Pharmacology* 20: 1–15.

Eph Receptor Signaling and Spine Morphology

F Irie and Y Yamaguchi, Burnham Institute for Medical Research, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Dendritic spines are small protrusions that are present in large numbers on the surface of dendrites. For example, a single pyramidal neuron in the hippocampal CA1 region possesses as many as 30 000 dendritic spines. A majority of excitatory synapses are formed on the surface of these dendritic spines. Dendritic spines typically have a mushroom-like morphology, with a spherical head that contains the postsynaptic density and a stalk that connects the head to the dendritic shaft. Spines exhibit remarkable morphological plasticity during development as well as in the adult brain. It has been speculated that the morphological plasticity of dendritic spines is important for learning and memory, and spine morphological abnormalities characterize neurological disorders such as mental retardation and schizophrenia. The molecular mechanisms underlying the morphological plasticity of spines have been intensively investigated, and recent studies have uncovered a critical role for Eph receptor signaling in the regulation of spine plasticity.

Dendritic Spine Formation

Young, immature neurons do not have dendritic spines. Instead, the dendrites of these neurons are covered with a large number of thin, filopodia-like protrusions. As neurons mature, these dendritic filopodia are replaced by dendritic spines. It is still not entirely clear how the transition from filopodia to mature spines occurs. In one model, highly motile dendritic filopodia make contact with nearby axons, and these initial contacts evolve into mature synapses as the tips of the filopodia expand to become spine heads. In another model, spines are formed directly from the dendritic shaft as new membrane appendages, independent of the filopodia that initiated the axonal contact. While there is still no consensus as to how dendritic spines develop, it is nevertheless clear that a cell surface recognition event between an axon and a dendrite is involved in initiating spine formation. This surface recognition event likely transduces signals into the postsynaptic neuron to induce the structural changes that lead to dendritic spine formation. As a cell contact-dependent signal transduction system that regulates the actin cytoskeleton, Eph receptors and their ephrin ligands possess the necessary properties for carrying out such a role (see Figure 1).

The Eph Receptor Family

The Eph receptors are transmembrane tyrosine kinases and are classified into two groups. The ten EphA receptors preferentially bind the glycosyl phosphatidylinositol (GPI)-anchored ephrin-A ligands, whereas the six EphB receptors preferentially bind the transmembrane ephrin-B ligands. Binding of ephrin ligands activates the tyrosine kinase activity of the Eph receptors through receptor clustering. Activated Eph receptors phosphorylate themselves (autophosphorylation) and other cytoplasmic proteins. Some of the phosphorylated tyrosine residues within the Eph receptor serve as binding sites for SH2 domain-containing proteins, such as Src, Fyn, and Abl. In addition to these classical signaling mechanisms, the C-terminal tail of the Eph receptors can bind postsynaptic density/disc/zonula occludens-1 (PDZ) domain-containing proteins, many of which have been implicated in synaptic function.

Binding of Eph receptors to ephrins also sends signals into the cells expressing the ephrins. The B-ephrins, which do not have a kinase domain, can be phosphorylated by cytosolic Src family kinases. Phosphorylated B-ephrins interact with SH2 domain-containing adaptor proteins, which may help transduce reverse signals. Furthermore, interaction of the cytoplasmic tail of the B-ephrins with PDZ domain-containing proteins has also been reported to play a role in reverse signaling. Thus, ephrin–Eph receptor signaling is bidirectional and can induce a reciprocal exchange of signals between two cells upon physical contacts.

Eph receptors and ephrins are most highly expressed in the developing nervous system. During development, the main function of these molecules is to regulate axon pathfinding and morphogenesis. For instance, Eph receptors and ephrins show remarkable gradient expression patterns in the retina and optic tectum, which is essential for establishing correct topographic connections between retinal axons and neurons in the tectum. While the roles of Eph receptors and ephrins in the developing nervous system are undeniably large, these proteins are also present in the adult brain. Strong expression of Eph receptor proteins in the adult brain is observed in regions such as the cortex, hippocampus, cerebellum, and olfactory bulbs, where remodeling of neuronal connectivity continues in the adult. Importantly, Eph receptors and ephrins are concentrated in the synaptic regions.

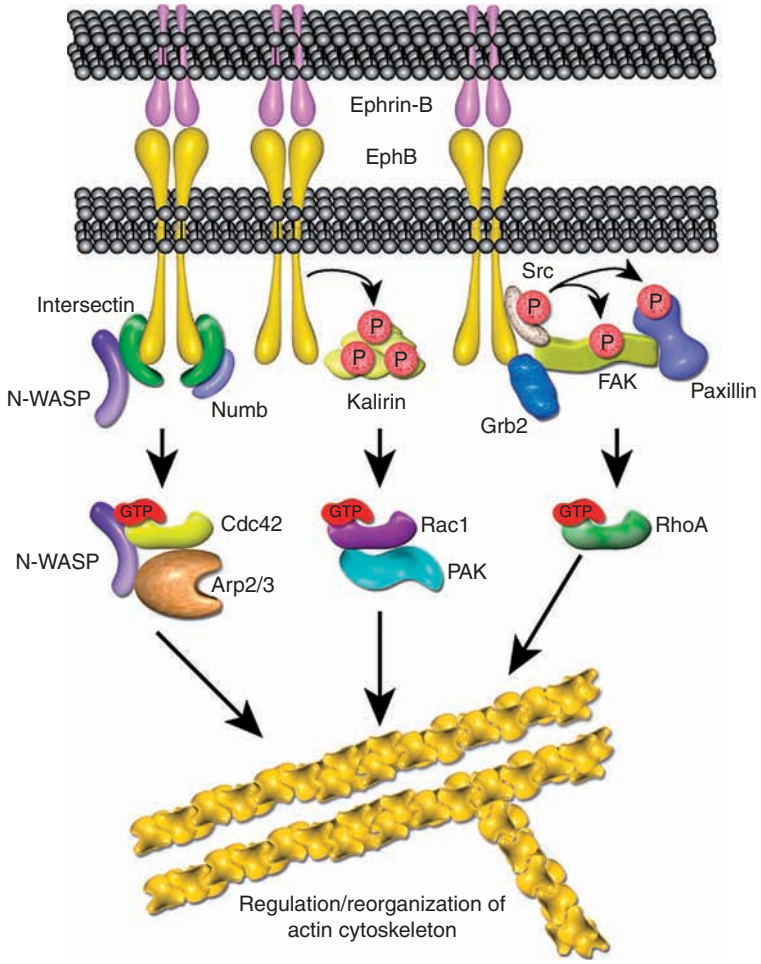

Figure 1 Ephrin ligands contact the Eph receptors, inducing intracellular signaling that controls multiple tissue morphogenic processes. See text for discussion.

Earlier studies suggest a pattern in which Eph receptors are present in the postsynaptic region, whereas ephrins are in axons and presynaptic sites. However, recent data indicate that the actual expression patterns are more complex, depending on the type of synapse. For instance, B-ephrins are presynaptic in the CA3 region of the hippocampus but postsynaptic in the CA1 region. In addition, ephrins and Eph receptors are also expressed at sites of contact between neurons and glial cells (see later).

Eph Receptors Promote Spine Formation during Development

It is well established that Eph receptor signaling plays a physiological role in the morphological maturation of dendritic spines during development. Initial observations were made in cultures of primary neurons in which EphB receptor signaling was perturbed. Transfection of a kinase-inactive EphB2 receptor mutant, which acts on all types of EphB receptors in a dominant-negative manner, inhibits spine formation in hippocampal neurons. Perturbed neurons continue to have immature, filopodia-like protrusions even after 3 weeks in culture. In contrast, treatment with exogenous B-ephrins induces premature spine formation in immature neurons, which usually have only filopodia-like protrusions.

A physiological requirement of EphB receptor signaling in spine development *in vivo* has been demonstrated by the analysis of EphB receptor mutant mice. In *EphB1/EphB2/EphB3* triple-null mice, a significant reduction in spine density and a prevalence of spines with no heads or with abnormally small sizes have been identified. Abnormal spine development was also observed in knockin mice expressing an *EphB2* mutant, in which its kinase domain is replaced with lacZ, on an *EphB1* null background. These results indicate that EphB receptor forward signaling is required for normal dendritic spine formation, a conclusion consistent with the observations from *in vitro* experiments.

In neurons expressing the kinase-inactive EphB2 receptor, the number of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) glutamate receptor clusters is markedly reduced, whereas the number of γ-aminobutyric acid (GABA) receptors, which form directly on the dendritic shaft, remains unchanged. It is not known whether the loss of AMPA receptor clusters is the consequence of the absence of spine formation *per se* or whether EphB receptor signaling is directly involved in AMPA receptor clustering. As for the latter possibility, it is interesting to note that the AMPA receptor-binding protein, glutamate receptor-interacting protein 1 (GRIP), binds to the C-terminal tail of EphB2 receptor.

While EphB receptor forward signaling is required for spine formation, it is not known which ephrin(s) acts as physiological ligands in this context. In the mouse hippocampus, ephrin-B3 is expressed in dentate gyrus granule neurons that project to CA3 neurons. This ephrin is an obvious candidate as the physiological ligand for EphB receptors on dendritic spines of CA3 neurons. Increasing evidence demonstrates that the spatiotemporal expression patterns of ephrins and Eph receptors are complex. It is therefore likely that different ephrin–Eph receptor combinations are involved in spine development in different brain regions.

Although a physiological role for EphB receptors in spine development is supported by both *in vitro* and *in vivo* data, much less is known about the role of EphA receptors. There is evidence that ephrin-A–EphA receptor signaling plays a role in the morphological plasticity of spines through neuron–glia communication. Astrocytes are an integral component of synapses, with their fine membrane processes making contact with extrasynaptic surfaces of dendritic spines. Ephrin-A3 and EphA4 receptor are expressed at sites of neuron–astrocyte contact (ephrin-A3 on astrocyte processes, and EphA4 receptor on dendritic spines). An interesting observation in this regard is that treatment of hippocampal slices with ephrin-A3 induces shortening of dendritic spines, and spines in *EphA4* null neurons are longer than in wild-type neurons, suggesting that ephrin-A3 on astrocytes may regulate spine morphology through activation of EphA4 receptor signaling in dendritic spines.

Molecular Mechanisms of Spine Morphogenesis and the Role of Eph Signaling

The cytoskeleton of dendritic spines is composed mainly of actin filaments. Dynamic reorganization of the actin cytoskeleton is essential for the motility of dendritic spines, as shown by a number of perturbation experiments. In contrast, blocking myosin motors does not inhibit spine motility. These results suggest that reorganization of the actin cytoskeleton is the key molecular event underlying spine formation. Consistent with this notion, recent studies have revealed the essential role of Rho family guanine triphosphate hydrolases (GTPases) in dendritic spine formation. For example, expression of constitutively active Rac1 in transgenic mice causes an increase in number of smaller spines in cerebellar Purkinje cells, whereas expression of dominant-negative Rac1 causes a progressive elimination of spines in hippocampal slices. In a *Drosophila* Cdc42 mutant, spine density is reduced in a subset of neurons (the vertical system neurons). Perturbation of another Rho family GTPase, Rnd1, causes reduction in the spine number and width and shortens the length of headless protrusions.

There is a wealth of data indicating that ephrin–Eph receptor signaling regulates spine formation and other morphological changes in neurons via reorganization of the actin cytoskeleton (see **Figure 1**). The main targets of EphB receptor signaling appear to be Rho family GTPases and their regulators, guanine nucleotide exchange factors (GEFs). A pioneering study on this regard was the identification of the Rho GEF ephexin as an effector for ephrin-A-mediated growth cone collapse. Ephexin binds EphA receptors and the association activates the GEF activity of ephexin, which leads to RhoA activation and growth cone collapse. The role of EphB receptors in spine development appears to involve a similar modulation of GEFs and Rho family GTPases. Treatment of cultured neurons with ephrin-B2 ligand causes activation of Cdc42. The association of EphB2 receptor and the actin-binding neural Wiskott–Aldrich syndrome protein (N-WASP) with the GEF intersectin-l enhances the activity of intersectin toward Cdc42. It is well established that Cdc42 and N-WASP coordinately activate the actin polymerization machinery through the actin-related protein (Arp2/3) complex, leading to the formation of branched actin filaments. Thus, EphB receptor activation on the surface of presumptive spines may cause the expansion of spine heads through the polymerization of branched actin filaments. EphB receptor–intersectin signaling may be modulated by Numb, an adaptor protein that interacts with intersectin-l and Cdc42. Numb promotes the GEF activity of intersectin, and small interfering RNA (siRNA)-mediated Numb knockdown suppresses ephrin-B-dependent spine formation in cultured hippocampal neurons.

Another GEF regulated by EphB receptors is kalirin. Kalirin-7, one of the kalirin isoforms, catalyzes nucleotide exchange on Rac1 and is enriched in the postsynaptic density. Expression of a GEF-inactive

mutant of kalirin-7 reduces the number of dendritic spines in cortical neurons. Ephrin-B2-mediated activation of EphB receptors induces tyrosine phosphorylation and translocation of kalirin to synapses, which in turn activates p21-activated kinase (PAK). PAK is a Cdc42/Rac-activated serine–threonine kinase that promotes actin reorganization. This pathway is also likely to be involved in spine formation, since perturbation of PAK inhibits ephrin-B-induced spine formation.

EphB receptor signaling leading to spine formation has also been recently reported to activate a pathway involving focal adhesion kinase (FAK) and paxillin. Stimulation with ephrin-B activates RhoA through the association of EphB receptors with FAK, Src, Grb2, and paxillin, which is induced by tyrosine phosphorylation of these proteins. Knockout of FAK or dominant-negative perturbation of RhoA blocks dendritic filopodia shortening induced by ephrin-B stimulation in young neurons. These data suggest that RhoA also plays a role in dendritic spine formation.

As described earlier, both *in vitro* and *in vivo* experiments indicate that kinase activity is required for EphB receptors to induce spine formation. This suggests that proteins that are phosphorylated by EphB receptors play a role in the process. Besides kalirin, one of the phosphorylation targets of EphB receptors is the cell surface proteoglycan syndecan-2. Upon transfection into young neurons, syndecan-2 induces spine formation. Although the molecular mechanism underlying syndecan-2-mediated spine formation remains unclear, phosphorylation of its cytoplasmic domain by EphB receptors is necessary for this activity. Other example of EphB receptor phosphorylation targets that are likely to be involved in spine formation are the nonreceptor tyrosine kinase Abl and the Abl-related kinase Arg. These kinases affect actin polymerization via enabled/vasodilator-stimulated phosphoproteins (Ena/VASPs), which associate with the barbed ends of actin filaments and promote continued actin polymerization. In *Drosophila*, Ena promotes branching of dendrites and the formation of spine-like protrusions.

The formation of mature dendritic spines requires not only the reorganization of the actin cytoskeleton but also a supply of membranes to accommodate the increase in surface area. Dendritic spines contain endosomal compartments, which serve as a store of recycling membranes. Manipulating the levels of an endocytic protein (CPG2) has been shown to affect spine size. It is also interesting that ephrin-B–EphB receptor signaling regulates clathrin-mediated endocytosis of AMPA-type glutamate receptors through tyrosine phosphorylation of the endocytic accessory protein synaptojanin 1. Although not yet investigated, activation of EphB receptor signaling may therefore alter the supply of membranes available for enlargement of dendritic spines. Thus, ephrin-B–EphB receptor signaling could regulate dendritic spine formation and plasticity by influencing not only the actin cytoskeleton but also membrane trafficking.

Concluding Remarks

While originally identified as axonal guidance cues operating during embryonic development, the ephrin–Eph receptor system has emerged as an important regulator of synaptic function. As summarized in this article, studies done during the past 5 years have largely established the functional significance of these signaling molecules in dendritic spine morphogenesis. There remain important unanswered questions, though. For instance, our understanding of the role of ephrins and Eph receptors in activity-dependent spine plasticity is still limited. Is morphological plasticity of spines in the adult brain also regulated by ephrins and Eph receptors? We do not know how the function of Eph receptors in *N*-methyl-D-aspartate (NMDA)-dependent synaptic plasticity, which is mediated by their extracellular domains, is related to their function in spine formation. Finally, it will be important to investigate whether dysfunction of the ephrin–Eph receptor system may underlie some human mental disorders. Abnormal spines have been reported in various mental disorders and psychiatric conditions, including mental retardation, autism, schizophrenia, and drug addiction. Although the association of null mutations of Eph receptor and ephrin genes with mental disorders has not been identified, Eph receptor signaling impairment may be caused by other genetic, epigenetic, or pharmacological mechanisms, which would be one of the major focuses in ephrin–Eph receptor research in the brain.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Dendritic Spine History; Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin; LIM Kinase and Actin Regulation of Spines; Rho GTPases and Spines; Spine Plasticity; Spines and Mental Disorders; Synaptic Precursors: Filopodia.

Further Reading

Ethell IM, Irie F, Kalo MS, et al. (2001) EphB/syndecan-2 signaling in dendritic spine morphogenesis. *Neuron* 31: 1001–1013.

Henderson JT, Georgiou J, Jia Z, et al. (2001) The receptor tyrosine kinase EphB2 regulates NMDA-dependent synaptic function. *Neuron* 32: 1041–1056.

Henkemyer M, Itkis OS, Ngo M, et al. (2003) Multiple EphB receptor tyrosine kinases shape dendritic spines in the hippocampus. *Journal of Cell Biology* 163: 1313–1326.

Irie F and Yamaguchi Y (2002) EphB receptors regulate dendritic spine development via intersectin, Cdc42 and N-WASP. *Nature Neuroscience* 5: 1117–1118.

Murai KK, Nguyen LN, Irie F, et al. (2003) Control of hippocampal dendritic spine morphology through ephrin-A3/EphA4 signaling. *Nature Neuroscience* 6: 153–160.

Nimchinsky EA, Sabatini BL, and Svoboda K (2002) Structure and function of dendritic spines. *Annual Review of Physiology* 64: 313–353.

Penzes P, Beeser A, Chernoff J, et al. (2003) Rapid induction of dendritic spine morphogenesis by trans-synaptic ephrinB–EphB receptor activation of the Rho–GEF kalirin. *Neuron* 37: 263–274.

Epilepsy

P Kandula and C Harden, Weill Medical College at Cornell University, New York, NY, USA

Published by Elsevier Ltd.

Introduction

Epilepsy is a chronic neurologic condition marked by the recurrence of unprovoked seizures. Seizures in turn, are paroxysmal events that result from abnormal neuronal discharges. Both the terms seizure disorder and epilepsy are interchangeable and refer to recurrent, unprovoked seizures. Although terminology is a more recent advancement, epilepsy itself is not a newly recognized condition. In fact, mention of epilepsy dates back to the writings of Hippocrates around 700 BC. Nationwide, epilepsy affects anywhere from 0.5% to 3% of the population. The highest incidence occurs in young children and the elderly.

Etiology

As previously mentioned, only recurrent, unprovoked seizures constitute epilepsy. Seizures can also occur in certain 'provoked' settings. Metabolic derangements such as hypoglycemia, hyponatremia, and hypocalcemia could provoke self-limited seizures. In addition, acute alcohol withdrawal, certain medications that lower seizure threshold, and high fevers in young children, can also cause a seizure.

Any condition that alters the structure or function of the central nervous system (CNS) can predispose an individual toward epilepsy. In the elderly, vascular disease (ischemic strokes, hemorrhage, malformations) accounts for a large number of epilepsy cases. Primary CNS tumors, as well as metastatic tumors to the brain, account for epilepsy in mid-life. Infectious (encephalitis) and degenerative disorders (Alzheimer's dementia) are additional causes of epilepsy. Congenital malformations (cortical dysplasias, migrational defects) are more common in children. A few epilepsy syndromes have a known underlying genetic cause. Many of these syndromes occur in children and young adolescents, with an autosomal dominant inheritance, with incomplete penetrance. In reality, however, in approximately 50% of cases, a definitive underlying etiology cannot be found.

Pathophysiology

Epilepsy is a reflection of abnormal firing of neurons that produces an electrophysiologic change, as well as variable clinical symptoms. Simplistically, epilepsy is a mismatch between neuronal excitatory and inhibitory processes. In the CNS, the neurotransmitter that mediates excitation at most synapses is glutamate. However, glutaminergic expression of excitation depends on the specific receptor stimulated. Neuronal glutamate receptors include AMPA (alpha amino-3-hydroxyl-5-methyl-4-isoxale propionic acid), kainic acid, and NMDA (*N*-methyl-D-aspartate). AMPA receptors are responsible for fast excitatory neurotransmission via the sodium-potassium channel. NMDA receptors are found with AMPA receptors but are physiologically blocked by magnesium ions. However, when the AMPA channels are open (with glutamate action), the magnesium blockade is released, and the NMDA channels are open. Calcium can then enter, and the slow, sustained phase of excitatory neurotransmission takes place. On the other hand, GABA (γ-aminobutyric acid) is the main inhibitory neurotransmitter found in the CNS. Activation of the GABA-A receptor subtype is responsible for neuronal inhibition. There is also evidence to suggest that impaired GABA-A receptor function can lead to certain inherited or acquired epilepsies. Thus, lack of GABA-mediated inhibition along with excessive neuronal excitation are the critical mechanisms of epileptogenicity.

Classification and Clinical Features

Over the years, a concerted effort has been made to classify epilepsy syndromes. Today, classification is primarily based on the combination of the 1981 Classification of Epileptic Seizures and the 1989 Classification of Epilepsies and Epileptic Syndromes by the ILAE (International League Against Epilepsy). In large part, however, these distinctions are primarily academic, and the classification of seizures in practical terms is based on clinical symptoms and region of onset on the electroencephalogram (EEG).

Partial seizures are divided into complex partial and simple partial. Complex partial seizures involve altered awareness or impaired consciousness. Complex partial seizures can vary clinically. In the classical complex partial seizures of temporal lobe origin, for example, prominent lip smacking (oral movements) and automatic behavior consisting of nonvoluntary movements of fumbling and picking at objects can occur in the setting of altered consciousness. Patients may often be confused, tired, and unable to follow commands immediately after clinical symptoms cease, constituting a 'post-ictal' phase. In contrast, during simple partial seizures, patients are conscious and awareness is not altered. Most often, focal motor seizures are seen with simple partial seizures. These

movements consist of jerking of the contralateral leg, arm, face, or any combination of the three. Both simple and complex partial seizures have a focal onset on the EEG. However, either seizure type can start focally, but can spread to other regions of the brain and become secondarily generalized. Once a seizure becomes secondarily generalized, there is always associated loss of consciousness.

Generalized tonic-clonic convulsive seizures are the most common pattern in clinical practice. Seizures can be primarily generalized at onset, or instead start off as a complex partial seizure with subsequent tonic-clonic movements of all extremities. Clinically, associated cyanosis, increased salivary secretions, changes in the blood pressure or heart rate, bladder incontinence, or tongue biting can occur. Unlike a partial seizure, there is no preceding aura (symptoms preceding the onset of a seizure). Individuals are usually difficult to arouse for several minutes to hours after the convulsion. In fact, individuals may report complete amnesia for the entire event upon awakening. With regard to the EEG, generalized seizures start with rhythmic ictal discharges in both cerebral hemispheres at the same time. Currently, there are seven types of generalized seizures recognized by the ILAE, consisting of generalized tonic-clonic, tonic, clonic, myoclonic, absence, atypical absence, and atonic seizures.

Generalized tonic-clonic seizures (grand mal) were described earlier. Tonic seizures consist of only the stiffening (tonic) phase of a grand mal seizure. Similarly, clonic seizures consist of only the clonic (jerking) phase of a grand mal seizure. Myoclonic seizures are quick, sudden jerks of the entire body, or predominantly the upper extremities. At times, the movement is profound and patients may fall to the ground or alternatively drop an object out of their hand during the seizure. Absence, or petit mal, seizures are again brief episodes of behavioral arrest, with clinical unresponsiveness, staring, and eyelid fluttering or automatic type activity. Individuals immediately resume prior activity after the absence seizure is over. However, patients are generally amnestic to the event. Absence seizures are more common in childhood and can be mistaken by family and educational providers as attention-deficit disorders. Classically, the EEG shows bihemispheric three cycle per second spike-and-wave discharges during the seizure. Atypical absence seizures are generally longer in duration and include more complicated motor manifestations. In addition, the EEG does not show the classical three cycle per second spike-and-wave discharges, but rather a slower variant, referred to as slow spike and waves. Atonic seizures consist of a sudden loss of muscle tone. Clinical manifestations range from falling to the side while sitting to dropping to the ground, if standing. Patients can injure themselves if not protected. Atonic seizures are usually associated with a complex of other seizure types, seen in severe epilepsies. If atonic seizures are frequent, individuals may wear a helmet to protect themselves during an event.

Evaluation and Treatment

When a patient presents with a new onset paroxysmal event(s), particularly with alteration or loss of consciousness, seizures should be entertained as a possible diagnosis. Careful history and physical can help in distinguishing a seizure from other paroxysmal events. A clinical description from an observer can provide additional clinical information. Detailed questioning regarding the presence or absence of a warning sign (aura), alteration of consciousness, focal movements, automatic behavior, urinary or fecal incontinence, and post-event confusion should be obtained. Duration of the event is also important, as most seizures are typically less than 1.5 min in duration. Potential risk factors for seizures should be identified, such as a history of head trauma, prior history of encephalitis/meningitis, childhood febrile seizures, and family history of seizures. General medical history regarding use of new medications that may lower the seizure threshold (bupropion, lithium, levofloxacin) or cause metabolic derangements (diuretics causing hyponatremia, insulin causing hypoglycemia) that can lead to subsequent seizures should be questioned.

During the physical exam, focal neurologic signs (focal weakness, etc.) as well as systemic signs (fever, altered mental status) should be investigated. Laboratory investigation including electrolyte counts (sodium, calcium, magnesium, serum glucose), liver and renal function test, toxicology screen, and complete blood cell count should be performed. Further testing may include urinalysis, urine cultures, chest X-ray, blood cultures, and lumbar puncture if a source of infection is likely. If encephalitis is suspected, cerebrospinal fluid should be sent for herpes simplex virus polymerase chain reaction (PCR). Empiric treatment with antibiotics should be started in suspected cases of meningitis/encephalitis. Imaging should be performed in all patients, particularly when focal findings are suspected based on neurological examination and/or clinical history. Magnetic resonance imaging (MRI) is the imaging modality of choice. MRI sequences should include adequate sequences to visualize in particular the temporal lobes and mesial structures to identify common conditions such as mesial temporal sclerosis (atrophy of medial temporal lobe structures). Imaging with intravenous contrast should be

performed if a tumor is suspected. Finally, EEG should always be performed in addition to the above serologic and imaging tests. Neuroimaging is a test of structural integrity, while EEG is a test of cortical function. The purpose of the EEG is threefold. First, epileptiform discharges in the form of spikes, sharps, or spike-and-wave complexes may be identified, suggesting potential origin sites for seizures. Second, epileptiform discharges with specific patterns may help to identify epilepsy syndromes, such as the characteristic spike-and-wave discharges seen in childhood absence epilepsy. Finally, stratification of seizures into focal versus generalized onset can help select an appropriate anti-epileptic medication. Although epileptiform discharges are highly suggestive of epilepsy, they are not pathognomonic. The gold standard for localization of seizure onset would be to record an actual ictal (seizure) event. Thus, if further localizing information is needed, a prolonged EEG with video should be performed.

Treatment with an anti-epileptic agent is then based on EEG findings and clinical correlation. Technically, recurrent seizures constitute epilepsy. Thus, a patient with two or more unprovoked seizures should be treated. For some persons, it may be appropriate to treat after only one seizure has occurred, for example, high functioning adults for whom a seizure recurrence could have severe consequences, or those patients who have a structural lesion on brain MRI. The risk of seizure recurrence after one seizure is reduced with anti-epileptic drug treatment; therefore, the decision to treat should take into account the risk of seizure recurrence versus the risk of treatment.

Some individuals will continue to have seizures despite an adequate trial of at least two anti-epileptic agents and fall under the category of refractory epilepsy. Individuals with a focal onset of seizures in noneloquent cortex may be potential candidates for resective epilepsy surgery. In other cases, refractory patients may be considered for vagus nerve stimulation implantation or, more recently, implantable devices for deep brain stimulation for epilepsy.

Prognosis

Ultimate prognosis is dependent on the epilepsy syndrome, underlying etiology, and response to anti-epileptic agents. Some generalized epilepsies, such as childhood absence epilepsy, have a good prognosis. In some cases, children may eventually stop having seizures as adulthood approaches. In others in whom a baseline neurologic deficit exists, seizures may be more difficult to control. In general, patients who do not have a favorable response to a first anti-epileptic agent, unless it was for drug allergy, have a lower chance (around 15%) of completely responding to a subsequent second or third agent. The overall cumulative success rate of controlling epilepsy is on the order of 70%. Thus, 30% of patients would be considered medically refractory. In cases where patients are candidates for temporal lobectomy for mesial temporal sclerosis, epilepsy treatment success rate is nearly 70%.

After several years of complete seizure control on medication, the decision as to whether to stop antiseizure medications is often discussed. While there is no clear point after which discontinuing antiseizure medications is appropriate, most neurologists would not do so unless the patient was seizure-free for at least 2 years. Risk factors for seizure recurrence include the presence of a structural brain lesion, a persistently abnormal EEG, and frequent seizures prior to seizure control with medication.

Conclusions

As in any chronic illness, a multidisciplinary approach including neurologists, epileptologists, and psychiatrists is required for the management of epilepsy. Coexisting mood disorders such as depression, bipolar disorder, and dysthymia are not uncommon in the epilepsy population. There are also studies suggesting that people with epilepsy attempt suicides at higher rates than the general population. Lastly, many patients suffer from the psychosocial aspects of the condition, including cumulative medication side effects, the loss of control with recurrent seizures, and the ongoing cultural and societal stigma associated with the condition.

See also: Coma and Other Pathological Disorders of Consciousness; Epilepsy: Neuronal Death; Gamma-Aminobutyric Acid (GABA); Glutamate; Intracellular Calcium and Neuronal Death; Split-Brain Patients.

Further Reading

Engel J Jr. and Pedley TA (eds.) (1998) *Epilepsy: A Comprehensive Textbook*. Philadelphia, PA: Lippincott-Raven.

Harden CL (2002) The co-morbidity of depression and epilepsy: Epidemiology, etiology, and treatment. *Neurology* 59(6 supplement 4): S48–S55.

Theodore WH (2004) Brain stimulation for epilepsy. *Lancet Neurology* 3(2): 111–118.

Wyllie E (2001) *The Treatment of Epilepsy*. Philadelphia, PA: Lippincott Williams and Wilkins.

Epilepsy: Channelopathies

J A Vidaurre, K M Zamel, and E S Roach, The Ohio State University College of Medicine, Columbus, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Epilepsy can arise from a complex interplay between multiple genes and various environmental factors, but the recent discovery that some idiopathic epilepsies result from specific single gene mutations (monogenetic epilepsies) enhances our understanding of the genetics and molecular mechanisms of the these disorders. That most of these mutations directly affect ion channel function is not surprising, since ion channels are important in the brain for the maintenance of the delicate electrical equilibrium between neuronal networks.

Ion channels are important in a wide variety of cellular functions in the brain and other electrically excitable tissues, such as the heart muscle and the peripheral nerves, and there are numerous examples of neurological diseases that are caused by dysfunctional ion channels (e.g., periodic paralysis, episodic ataxia, hereditary deafness, inherited arrhythmia with congenital long QT syndrome, and migraine). Diseases arising from ion channel dysfunction are termed channelopathies, and to these can now be added several forms of genetically mediated epilepsy.

Function of Ion Channels

In the brain, ion channels are necessary for the generation and propagation of signals between neurons. Alteration in their function could produce a reduction in seizure threshold, causing different epilepsy syndromes. Ion channels typically result from cell membrane proteins that form selective pores so as to allow the passage of sodium (Na^+), potassium (K^+), chloride (Cl^-), or calcium (Ca^{2+}) ions across their respective electrochemical gradients. Basically, there are two types of ion channels:

1. Voltage gated (these are opened and closed by changes in membrane potential).
2. Ligand gated (a transmitter molecule opens and closes the gate by binding to a receptor site).

The gene mutations underlying the different genetic epilepsy syndromes can affect either the voltage- or the ligand-gated ion channels. Each type of ion channel is briefly described in the following sections, after which the clinical features and genetic bases of the epilepsy syndromes caused by mutations affecting ion channels are discussed (see Table 1 and Figure 1).

Voltage-Gated Cation Channels

Voltage-gated Na^+, K^+, Cl^-, or Ca^{2+} channels are formed by several subunits. Four pore-forming homologous subunits (the α subunits) determine the ion gating and permeation in the channel. One or more smaller accessory subunits (identified as β, γ, or δ) have other modifying functions. These ion channels are normally closed at the resting potential but are readily activated once the membrane is depolarized. Activation gates are opened within milliseconds of depolarization, and with sustained depolarization, closure of inactivation gates leads to channel inactivation (i.e., repolarization). The inactivated channel remains refractory to further openings till it recovers. The smaller accessory β, γ, or δ subunits regulate functional proteins in the membrane, thus affecting the kinetics of channel gating.

Na^+ and K^+ channels are mostly involved in the process of depolarization and repolarization. Depolarization occurs with inward Na^+ currents, while repolarization occurs with rapid inactivation of these Na^+ currents and activation of outward K^+ current. Alteration in these channel functions can lead to hyperexcitability and fast repetitive action potentials. Ca^{2+} channels regulate presynaptic neurotransmitter release at the nerve terminals.

Ligand-Gated Ion Channels

Ion channels that are activated by neurotransmitters are called ligand-gated channels. The transmitters activating these channels include acetylcholine (ACh), γ-aminobutyric acid (GABA), glycine, glutamate, and nucleotides. Neurotransmitters can either be excitatory (e.g., ACh and glutamate) or inhibitory (e.g., GABA and glycine). Ligand-gated channels are typically composed of four or five main subunits. The channel opens with transmitter binding and remains desensitized until removal of the transmitter, which leads to closure of the channel, allowing it to be available again for further openings.

Generalized Epilepsy with Febrile Seizures Plus

Febrile seizures are a common childhood seizure disorder affecting from 2% to 5% of children from 6 months to 5 years of age. Febrile seizures are not considered epilepsy, since they are, by definition, associated with an increase in body temperature. Although genetic factors play a role in febrile seizures,

Table 1 Genetic epilepsy syndromes and associated gene mutations

Epilepsy syndrome	*Chromosomal location*	*Gene*	*Channel, subunit*
GEFS+	19q13	*SCN1B*	Na, β1
	2q24	*SCN1A*	Na, α
	2q24	*SCN2A*	Na, α
	5q31–q33	*GABRG2*	GABA A, γ2
	1p36.3	*GABRD*	GABA A, δ
SMEI	2q24	*SCN1A*	Na, α
	5q31–q33	*GABRG2*	GABA A, γ2
BFNS	20q13.2	*KCNQ2*	K, α
	8q24	*KCNQ3*	K, β
BFIS	2q24	*SCN2A*	Na, α
	16p12–p11		
	19q		
BFNIS	2q24	*SCN2A*	Na, α
ADNFLE	20q13	*CHRNA4*	ACh, α4 subunit
	1q21	*CHRNB2*	ACh, β2 subunit
	15q24		
CAE	3q26	*CLCN2*	Cl
	8q24		
	16p13.3	*CACNA1H*	Cav3.2
JAE	3q26	*CLCN2*	Cl
JME	3q26	*CLCN2*	Cl
	2q22–2q23	*CACNB4*	Ca, β4
	5q14		
	5q34–q35	*GABRA1*	GABA A, α2
	6p21		
EGMA	3q27	*CLCN2*	Cl
GEEA	19p13	*CACNA1A*	Ca, α1A subunit

Abbreviations (syndromes and genes): GEFS+, generalized epilepsy with febrile seizures plus; SMEI, severe myoclonic encephalopathy of infancy; BFNS, benign familial neonatal seizures; BFIS, benign familial infantile seizures; BFNIS, benign familial neonatal–infantile seizures; ADNFLE, autosomal dominant nocturnal frontal lobe epilepsy; CAE, childhood absence epilepsy; JAE, juvenile absence epilepsy; JME, juvenile myoclonic epilepsy; EGMA, epilepsy with grand mal on awakening; GEEA, generalized epilepsy with episodic ataxia; SCN, sodium channel; GABR, GABA A receptor; KCN, potassium channel; CHRN, neuronal nicotinic acetylcholine receptor; CLCN, chloride channel; CACN, calcium channel. The different gene mutations encoding ion channels and their respective involved subunits are given (Na, sodium; K, potassium; GABA A, γ-aminobutyric acid; ACh, acetylcholine; Cl, chloride; Cav, calcium channel, voltage dependent; Ca, calcium).

the likelihood of a child with typical febrile seizures developing epilepsy (about 2%) is little different from the epilepsy risk in the general population. In 1997, a dominantly inherited epileptic syndrome was described in a large Anglo-Australian family; this led for the first time to joining febrile and afebrile seizures as part of the same clinical phenotype. Affected individuals frequently have afebrile seizures as well as simple febrile seizures that occur beyond 4 or 5 years of age. This pedigree shows a mild overall phenotype. The term generalized epilepsy with febrile seizures 'plus' (GEFS+) was coined, and many affected families have now been identified.

Clinical Features

GEFS+ is transmitted as an autosomal-dominant trait with reduced penetrance and a highly variable phenotype, even within the same pedigree. The manifestations include (1) febrile seizures, (2) febrile seizures 'plus' (initially used to describe febrile seizures after 5 years of age, but later, to include afebrile generalized tonic–clonic seizures occurring in addition to febrile seizures), (3) afebrile generalized seizures, including generalized tonic–clonic, myoclonic, atonic, absence (when absence occurs, seizures usually occur less frequently than with typical 'childhood absence epilepsy' in the sense that they are less frequent), and myoclonic–astatic epilepsy, and (4) partial seizures (frontal and temporal lobe epilepsy).

The most common seizure patterns within the different pedigrees are febrile seizures (FS) and febrile seizures plus (FS+). Different seizure types can be associated with FS or FS+, but the most common are generalized tonic–clonic. As mentioned earlier, the phenotype can be widely variable even within the same family with an identical mutation. Affected family members may manifest with GEFS+, only FS, or even isolated afebrile generalized tonic–clonic (GTC) or partial temporal seizures, suggesting that isolated afebrile seizures may represent a 'GEFS+

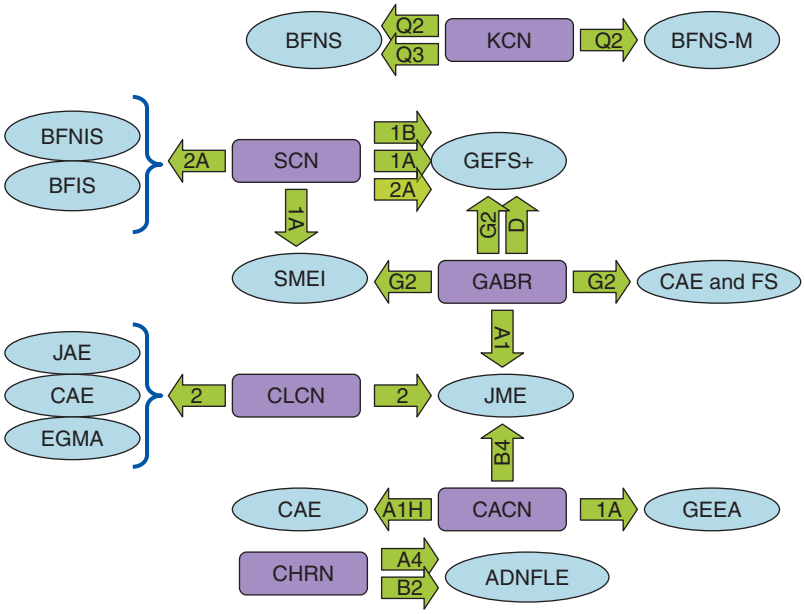

Figure 1 This diagram depicts the complex genotype–phenotype relationship in epilepsy. While a variety of mutations can produce similar phenotypes, a single mutation can produce a range of phenotypes. Abbreviations (syndromes, blue; genes, purple; gene subunits, green): BFNS, benign familial neonatal seizures; BFNIS, benign familial neonatal–infantile seizures; BFIS, benign familial infantile seizures; GEFS+, generalized epilepsy with febrile seizures 'plus'; SMEI, severe myoclonic epilepsy of infancy (Dravet's syndrome); CAE, childhood absence epilepsy; CAE&FS, childhood absence epilepsy and febrile seizures; JAE, juvenile absence epilepsy; JME, juvenile myoclonic epilepsy; ADNFLE, autosomal dominant nocturnal frontal lobe epilepsy; EGMA, epilepsy with grand mal on awakening; GEEA, generalized epilepsy with episodic ataxia; KCN, potassium channel; SCN, sodium channel; GABR, GABA A receptor; CLCN, chloride channel; CACN, calcium channel; CHRN, neuronal nicotinic acetylcholine receptor.

phenotype,' adding to the complex genotype–phenotype relationship. GEFS+ mutation carriers who remain asymptomatic have been identified, suggesting that other disease-modifying genes or possibly environmental factors play a role in the pathogenesis.

Genetics

Genes encoding both voltage-gated and ligand-gated ion channels have been implicated in the etiology of GEFS+. In 1998, linkage to chromosome 19q was demonstrated, and soon after, the site of the $\beta 1$ subunit of the voltage-gated Na^+ channel gene (*SCN1B*) was identified. To date, only a few mutations have been recognized in *SCN1B*, suggesting a minor role as a causative gene. Later, two French families with GEFS+ were mapped to chromosome 2q24 and mutations were identified in the aα1 subunit of the voltage-gated sodium channel (*SCN1A*). Subsequently, more than 100 mutations have been described in this gene. *SCN1A* is probably the most important gene in the pathophysiology of GEFS+.

Missense mutations are the most common defects identified in GEFS+, but *de novo* mutations in *SCN1A* have also been identified in patients with severe myoclonic epilepsy of infancy (SMEI), which also features seizures associated with fever but typically has a more severe phenotype. A single base mutation in *SCN2A* has been found in 1 among 19 unrelated Japanese families. However, the significance of *SCN2A* as a disease-causing gene remains controversial. Mutations in *SCN2A* have also been found in families with benign neonatal–infantile familial seizures and benign infantile familial seizures.

The GABA A receptor, a ligand-gated chloride channel, is a heteromeric complex with five subunits (most native brain receptors contain two α, two β, and one δ subunit). The gene (*GABRG2*) encoding the $\gamma 2$ subunit of the GABA A receptor was identified as a causal gene for GEFS+. The mutations for this pedigree differ from most GEFS+ pedigrees, in the sense that idiopathic generalized epilepsy (IGE) phenotypes, including classical childhood absence epilepsy, can be expressed together in the same pedigree. One *GABRG2* mutation was found recently in a patient with the SMEI phenotype. The family had a bilinear inheritance and it was hypothesized that a second mutation could have been contributed to the phenotype. Another gene encoding the δ subunit of the GABA A receptor (*GABRD*) has been lately identified in the etiology of GEFS+.

Treatment

Given the wide variability in seizure phenotype, treatment can be tailored to best suit the clinical situation.

Patients with isolated febrile seizures may not require treatment, while those with febrile seizures plus other seizure types may benefit from a broad spectrum of antiepileptic drugs.

Severe Myoclonic Epilepsy of Infancy: Dravet's Syndrome

Severe myoclonic epilepsy of infancy (SMEI) is a well-defined syndrome first described by Dravet in 1978. The 1989 International League against Epilepsy classifies this syndrome under undetermined epilepsies and syndromes, depending on whether seizures are focal or generalized.

Clinical Features

Symptoms usually appear before 1 year of age, and the average age of onset was approximately 5 months in Dravet's series. The first seizure is typically precipitated by fever and can be generalized or focal with variable side predominance. Individual seizures are often prolonged and can evolve into status epilepticus. Seizures may become recurrent. These generalized or partial febrile seizures are followed by different types of afebrile seizures, including myoclonic, partial seizures, and absence.

The myoclonic seizures usually occur later in the course of the syndrome, but they may appear earlier. Some children have numerous, even relentless, myoclonic seizures, while others have them only on awakening or during the short interval preceding another seizure. The myoclonias can be generalized or segmental. The generalized myoclonias can be massive, causing the child to fall. In some cases they are focal, involving one limb or the head prior to becoming generalized. Segmental myoclonus affecting distal part of the upper extremities and face can be observed, especially in the periods with more frequent seizures. These myoclonic seizures are accompanied by generalized spike-wave or polyspike-wave discharges seen on electroencephalography (EEG).

The partial seizures are complex partial with adversive or atonic features and can present with autonomic changes, including cyanosis, and they can become secondarily generalized. The absence may present atypical features, sometimes with a myoclonic or atonic component and less rhythmic spike-wave discharges in the EEG. Patients can develop frequent episodes of status epilepticus, either convulsive, precipitated by fever, or 'obtundation status' with erratic myoclonias and increase in muscular tone.

Development is usually normal prior to seizure onset, but later psychomotor slowing and neurological signs, including ataxia and pyramidal signs, are observed. Occasionally, patients whose seizures are treated effectively retain good cognition. The EEG may be normal at the beginning but progressively changes to show generalized and focal or multifocal abnormalities. Photosensitivity is present in more than 40% of cases. Recently, similar phenotypes without myoclonus have been described and are considered to be a clinical variant of SMEI. It has been suggested that SMEI could represent the more severe end of the spectrum of GEFS+.

Genetics

Because fever-induced seizures are known to occur in both GEFS+ and SMEI patients, a mutation analysis of *SCN1A* was performed in seven patients with SMEI. *De novo* dominant mutations were found and the phenotype of *SCN1A* mutations was extended to include SMEI. A majority of the *SCN1A* mutations are truncating alleles and are presumed to result in a loss of Na^+ channel function.

Treatment

Seizures are often refractory to treatment with antiepileptic drugs. Valproate and benzodiazepines are the most useful drugs. It has been observed that carbamazepine and lamotrigine may aggravate the syndrome. Stiripentol (a drug not available in the US) and topiramate have been shown to be effective in the convulsive seizures. Ethosuximide can be helpful for absences and myoclonic seizures.

Benign Familial Neonatal Seizures

Benign familial neonatal seizures (BFNS) represent a type of rare autosomal-dominant inherited epilepsy syndrome with a penetrance as high as 85%. The seizures occur in otherwise well newborns, typically beginning on the second or third day of life (2–8 days in the majority of neonates) and typically remit after a few weeks. Up to 31% of patients may have seizures during the neonatal period after the first week of life. Affected premature babies will have their seizure onset at an older chronological age. Seizures may be focal clonic or may have a tonic or tonic–clonic component and are often accompanied by apnea. Usually they are brief, but the rate of recurrence can be high and they have a tendency to cluster. These seizures often spontaneously remit, but about 11% of patients may develop other types of seizures later in life.

Genetics

BFNS exhibits both clinical and genetic heterogeneity. The disease was first mapped to chromosome 20q, and later a second focus was identified in chromosome 8q. The genes (*KCNQ2* and *KCNQ3*, respectively) were identified as encoding voltage-dependent

potassium channels. Both KNCQ2 and KNCQ3 channels co-assemble to form heteromeric channels (KCNQ2/KCNQ3), expressing larger currents. It has been postulated that the association of these channels is the equivalent of the M current (named for its regulation by muscarinic receptors), a slowly activating and deactivating potassium current that functions as a brake for repetitive action potentials. A slight reduction by approximately 25% in KNCQ2/KNCQ3 currents, caused by mutations, can lead to an increase in neuronal hyperexcitability. *KCNQ2* and *KCNQ3* mutations can produce the same phenotype. Most of the mutations have been identified in *KCNQ2* and recently a mutation in *KCNQ2* combined BFNS patients and affected members who subsequently developed myokymia in the same pedigree, suggesting that other gene–environmental factors could modify the phenotype.

It has been postulated that the seizures occur in newborns because there is a general increase in excitation during the neonatal period. GABA is excitatory early in brain development, and this is probably due to the low expression of a potassium chloride cotransporter (KCC2), so the intracellular concentration of chloride is higher than in the extracellular space, and when chloride channels open, chloride ions exit the cell, causing depolarization. On the other hand, the expression of KCNQ2/KCNQ3 currents probably plays a more important role early in life than they do in the mature brain.

Treatment

Since seizures usually remit spontaneously, the best treatment approach is not clear. Phenobarbital, phenytoin, and valproate have been used. Treatment is usually for a brief period of time. Retigabine, an experimental drug currently in trials, activates KCNQ2/KCNQ3 channels by shifting their voltage dependence to more negative stages.

Benign Familial Infantile Seizures

Benign familial infantile seizures (BFIS) is the name given a newly described syndrome characterized by focal seizures with or without secondary generalization, appearing usually between 4 and 8 months of life. It has been observed that the seizures have a tendency to occur in clusters and they are brief. Slow deviation of head and eyes to one side, hypertonia, limb jerks, and cyanosis have been described in the clinical characteristics. This clinical entity has an autosomal-dominant pattern of inheritance and a benign outcome, since seizures resolve within the first year.

BFIS is a genetically heterogeneous condition, with loci mapped to chromosomes 16 and 19, but recently a mutation in *SCN2A* was identified in a family with BFIS and good outcome, thus expanding further the group of the epilepsies caused by channelopathies.

Benign Familial Neonatal–Infantile Seizures

In recent years a mutation in *SCN2A* was described in two families with a mean seizure onset of 1–9 months and mean offset of 3–8 months of age. This condition was named benign familial neonatal infantile seizures (BFNIS), demonstrating the overlapping clinical features between the genetic epileptic syndromes.

Autosomal-Dominant Nocturnal Frontal Lobe Epilepsy

Clinical Features

Autosomal-dominant nocturnal frontal lobe epilepsy (ADNFLE) is a clinical entity characterized by clusters of nocturnal seizures with a predisposition to occur during non-rapid eye movement (NREM) sleep, especially during stage 2. The attacks consist of sudden arousals with hypermotor manifestations with bipedal or bimanual movements, kicking or cycling, and agitated motor behavior. The seizures can manifest with abnormal dystonic postures or complex motor activities, with expression of fear or shouting, and can be accompanied by autonomic manifestations, in particular tachycardia. There is marked familial variation in semiology and severity.

The majority of patients retain consciousness during the events. The seizures usually start in the first or second decade of life and have a tendency to decrease in frequency during adulthood, but rarely disappear. Diurnal attacks can be present in some patients, especially in those with refractory epilepsy. The seizures can be misdiagnosed as parasomnias, but the epileptic events have a tendency to cluster. Interictal EEGs can be normal, and ictal recordings are difficult to interpret due to movement artifacts, but frontal spikes or rhythmic theta–delta activity in the anterior regions have been reported. It is impossible to distinguish ADNFLE from sporadic NFLE by clinical semiology alone.

Genetics

ADNFLE was the first familial focal epilepsy for which a genetic background was identified. The transmission pattern is autosomal dominant with

reduced penetrance (84%), but depending on the family, the penetrance can vary from 29% to 87%, to 100%. Three loci (chromosomes 20q13, 15q24, and 1q21) but only two genes have been identified. These genes encode for the $\alpha 4$ and $\beta 2$ subunit of the neuronal nicotinic acetylcholine receptor (*CHRNA4* in chromosome 20q13 and *CHRNB2* in chromosome 1q21). Altogether the *CHRNA4* and *CHRNB2* mutations account for the minority of cases, and for the vast majority the causal gene remains unknown.

It is unclear how the *CHRNA4* and *CHRNB2* mutations increase neuronal hyperexcitability specifically affecting the frontal lobes. Mutations can cause either loss or gain of function, and it has been postulated that reduction in the calcium dependence of the receptors could facilitate glutamate release at excitatory synapses, since calcium modulation could act as a negative feedback mechanism, preventing the overstimulation of glutamate release by the presynaptic $\alpha 4$ and $\beta 2$ neuronal acetylcholine receptors (nAChRs) at central excitatory synapses.

Treatment

Carbamazepine has been the drug of choice. One-third of patients become intractable. Recently, good results with nicotine patches and acetazolamide have been reported.

Other Types of Common Idiopathic Generalized Epilepsies Associated with Mutations in Ion Channel-Encoding Genes

The idiopathic generalized epilepsies include a distinctive spectrum of generalized epilepsies that range from rare syndromes to more common phenotypes, such as childhood absence epilepsy (CAE), juvenile absence epilepsy (JAE), juvenile myoclonic epilepsy (JME), and epilepsy with grand mal on awakening (EGMA). A hallmark of generalized epileptic disorders is the generation of synchronous spike-wave discharges (SWDs) in the range of 3–6 Hz recorded bilaterally over both brain hemispheres during the seizure event. Several chromosomes have been implicated in IGEs, including 3q26, 2q36.1, 14q23, 6p21.2, 8p11.2, and 18q21.1. This indicates that IGEs are a group of genetically heterogeneous disorders that are associated with mutations in several genes. Several clinical studies suggest that all IGEs share common genetic etiologies, as all major IGE syndromes have been found to occur within the same families. Recently, mutations in genes encoding different types of ion channels have been found in families affected with the most common forms of IGEs, including CAE, JAE, and JME.

Childhood Absence Epilepsy (Pyknolepsy)

CAE (pyknolepsy) accounts for 5–15% of childhood epilepsies. CAE typically manifests at age 6–7 years of age, with recurrent absence attacks that are characterized by an abrupt alteration of consciousness presenting with brief alteration in responsiveness and awareness. The seizures are accompanied on EEG by stereotyped 3 Hz bihemispheric spike-wave complexes that are most prominent in the frontal regions.

Fourteen of 118 patients with CAE were reported to have 12 putative mutations in the T-type calcium channel gene *CACNA1H*. It was concluded that missense mutations in this gene may predispose to CAE. The role of *CACNA1H* variants in the cause of human epilepsy still remains unclear, as these findings could not be replicated in another study. Furthermore, analysis of some of these putative mutations revealed changes in channel properties which were consistent with epileptogenesis. Interestingly, ethosuximide, the drug of choice for CAE, modulates the thalamocortical network that is synchronized by the voltage-dependent T-type calcium channel (Cav3.2).

GABRG2 mutations have been implicated in families with CAE and febrile seizures. These families probably represent a different subtype of GEFS+ phenotypes. Recently, mutations of *CLCN2* in the 3q26 region were found in families with IGEs, including CAE, JAE, JME, and EGMA. *CLCN2* mutation was estimated to be responsible for 40% of all familial IGEs, although this was challenged in a separate study that found it only in 3 out of 46 families with IGEs. CLCN2 channels play an important role at the GABAergic inhibitory synapses. CLCN2 channels are believed to play a role in maintaining a low intracellular Cl^- concentration. GABAergic stimulation leads to Cl^- flux intracellularly, resulting in hyperpolarization inhibition. Therefore, a dysfunctional CLCN2 channel results in buildup of intracellular Cl^-, leading to reduction in the inhibitory GABA response.

Juvenile Absence Epilepsy

JAE is mainly differentiated from CAE based on age of onset, which is usually between 9 and 13 years of age. Otherwise, JAE and CAE share many clinical similarities. As stated previously, mutations in the *CLCN2* gene have been also associated with JAE. GTC seizures tend to occur more often in JAE, especially in the morning. Like CAE, bursts of generalized spike-wave discharges are also seen in JAE, but usually at a higher frequency of 3–4 Hz. Although clinically a unique syndrome, it is not clear if JAE is a genetically distinct condition different from CAE. Two recent studies suggest that JAE and CAE share a common genetic relationship that is different from

that of JME. Despite these similarities, prognosis of JAE is not as benign as it is for CAE. While seizure control is usually achieved in JAE, most patients require treatment indefinitely.

Juvenile Myoclonic Epilepsy

JME is a common epilepsy syndrome, representing 20–27% of all patients with IGEs and accounting for 5–10% of all epilepsy cases. JME usually presents between 12 and 18 years of age. Early morning myoclonic seizures are a prominent part of this syndrome. These are symmetric shocklike jerks involving mainly the upper extremities, resulting frequently in sudden dropping of objects but almost never causing the patient to fall. These myoclonic jerks can be precipitated by alcohol, fatigue, or menstruation. Unlike CAE and JAE, absence seizures in JME occur only in about 30% of patients, while GTC seizures can occur in over 90% of cases. The EEG usually shows a normal background with intermittent generalized bursts of symmetric 3.5–6 Hz polyspikes and wave discharges that are far less regular than the spike waves seen in CAE. The prognosis for seizure control is excellent. Similar to JAE, JME responds remarkably to antiepileptic treatment, though this treatment may be required lifelong. Neurologic outcome and cognitive outcome are also excellent.

Extensive genetic work has been done to understand the inheritance of JME. It has been linked to several chromosomes, indicating that JME is a genetically heterogeneous disorder. *GABRA1* mutation (A322D) on chromosome 5q34–q35 was implicated in causing autosomal-dominant JME (AD-JME) in a large French-Canadian kindred. This mutation can impair the fast synaptic inhibition of the brain at the GABA A receptor. Involvement of the calcium (Ca^{2+}) channel was suggested when a nonsense mutation of the gene (*CACNB4*), mapped to 2q22–q23, encoding the Ca^{2+} channel β4 subunit was found in a German family with IGE and in a woman with JME. Mutation in the chloride channel-2 gene (*CLCN2*) on chromosome 3q26 was also found in a family with JME.

Generalized Epilepsy with Episodic Ataxia

Generalized epilepsy and episodic ataxia, two distinct paroxysmal disorders, have been recently linked together and related to calcium channel gene mutations. A heterozygous nonsense mutation in *CACNA1A* was reported in a patient with generalized tonic–clonic seizures, absences, and episodic ataxia.

Conclusion

The recognition that some of the 'generalized and partial idiopathic epilepsies' are caused by alterations in ion channel function (channelopathies) not only provides us with a better understanding of the molecular mechanisms implicated in seizure generation, but also improves our present ability as clinicians to make more precise diagnosis. This knowledge increases the possibilities for potential development of newer antiepileptic drugs with more specific mechanisms of action, as is the case of retigabine, designed to operate on the KCNQ2/KCNQ3 channels.

In addition, we have learned of the complex genotype–phenotype relationship between the different genetic epilepsy syndromes. Mutations in different genes may cause a similar phenotype, as is the case for BFNS, in which mutations in *KCNQ2* and *KCNQ3* are associated with seizures in the neonatal period. Conversely, mutations in the same gene can produce different phenotypes, as in *SCN2A* mutations that have been linked to GEFS+, BFNIS, and BFIS. This implies that different modifier genes or environmental factors play a role in the final phenotypic expression (see **Figure 1**).

The genetics of these epilepsy syndromes becomes even more complex with the discovery that epileptic channelopathies can be caused by *de novo* mutations, and a family history may be lacking, as is the case of the *SCN1A de novo* mutation causing the SMEI phenotype. Even if considerable progress has been achieved in this field, the genes for the vast majority of idiopathic epilepsies remain unknown, but we are at the beginning of a new era in the discovery of the genetics behind the different seizure types.

See also: Calcium Channels; Cyclic Nucleotide-Gated and Hyperpolarization-Activated Channels; Epilepsy; Epilepsy: Neuronal Death; Hemiplegic Migraine and Channel Disorders; Inwardly Rectifying Potassium Channels; Large Conductance Calcium-Activated Potassium Channels; Neuromodulation of Sodium Channels; Neuromodulation of Calcium Channels; Neuropeptides: Epilepsy; Paroxysmal Paralysis; Potassium Channel Regulation; Sodium Channels; Transient Receptor Potential (TRP) Channels; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Andermann F, Kobayashi E, and Andermann E (2005) Genetic focal epilepsies: State of the art and paths to the future. *Epilepsia* 46(supplement 10): 61–67.

Audenaert D, VanBroeckhoven C, and DeJonghe P (2006) Genes and loci involved in febrile seizures and related epilepsy syndromes. *Human Mutation* 27(5): 391–401.

Chen Y, Lu J, Pan H, et al. (2003) Association between genetic variation of CACNA1H and childhood absence epilepsy. *Annals of Neurology* 54: 239–243.

Dravet C (1978) Les épilepsies graves de l'enfant. *Vie Médicale* 8: 543–548.

Gardiner M (2005) Genetics of idiopathic generalized epilepsies. *Epilepsia* 46(supplement 9): 15–20.

George A (2004) Inherited channelopathies associated with epilepsy. *Epilepsy Currents* 4(2): 65–70.

Hirose S, Motohiro O, Sunao K, et al. (2005) Genetics of idiopathic epilepsies. *Epilepsia* 46(supplement 1): 38–43.

Kullmann D (2002) The neuronal channelopathies. *Brain* 125: 1177–1195.

Roger J (2002) *Epileptic Syndrome in Infancy, Childhood and Adolescence*, 3rd edn. Montrouge: John Libbey Eurotext.

Scheffer IE and Berkovic SE (1997) Generalized epilepsy with febrile seizures plus. A genetic disorder with heterogenous clinical phenotypes. *Brain* 120: 479–490.

Scheffer IE, Harkin LA, Dibbens LM, et al. (2005) Neonatal epilepsy syndrome and generalized epilepsy with febrile seizures plus (GEFS+). *Epilepsia* 46(supplement 10): 41–47.

Steinlein O (2004) Genes and mutations in human idiopathic epilepsy. *Brain & Development* 26: 213–218.

Epilepsy: Neuronal Death

A Pitkänen, University of Kuopio and Kuopio University Hospital, Kuopio, Finland
K Lukasiuk, The Nencki Institute of Experimental Biology, Warsaw, Poland

© 2009 Elsevier Ltd. All rights reserved.

Occurrence of Neurodegeneration in Epilepsy

Approximately 0.8% of the world population (50 million) has epilepsy, according to the World Health Organization. Epilepsies can be divided into three major categories based on etiology: idiopathic, symptomatic, and presumed symptomatic (previously called 'cryptogenic'). In idiopathic epilepsies, genetic factors causing, for example, channelopathies, are presumed to have a major causative role in the development of seizures. In symptomatic epilepsies, there is an identifiable lesion in the brain that triggers seizures. The lesion can be, for example, a genetically programmed molecular defect leading to cortical dysplasia, or an acquired lesion like traumatic brain injury (TBI). Presumed symptomatic epilepsies are those most likely to be symptomatic, but no lesion can be identified by currently available methods. Neurodegeneration in epilepsy is a multifaceted issue. For example, in some epilepsies progressive neurodegeneration can be caused by genetically encoded enzymatic defect that leads to brain atrophy and mental retardation, and epilepsy is just one component of the clinical phenotype. Here we focus exclusively on acquired symptomatic epilepsies that typically develop in previously normal brain. In these patients epileptic process has three phases: a brain-damaging insult (BDI) (e.g., status epilepticus (SE), TBI, or stroke) leading to epileptogenesis (latency period, during which there are no seizures), followed by recurrent seizures or epilepsy. Most of the available information comes from models where epileptogenesis is triggered by electrically or chemically induced SE. It is noteworthy that neurodegeneration is just one component of reorganization of neuronal circuits that eventually lead to occurrence of spontaneous seizures after epileptogenic BDI. Other changes include axonal and dendritic plasticity, neurogenesis and gliogenesis, vascular alterations, axonal damage, and remodeling of extracellular matrix and cellular membranes (Figure 1). The mechanisms of neurodegeneration are under most intense investigations because of the promise of neuroprotective therapies to improve the long-term outcome after various epileptogenic BDIs.

Where in Epileptic Brain Do Neurons Die?

Analysis of surgically resected hippocampal tissue from patients who have been operated due to drug-refractory temporal lobe epilepsy (TLE) show that about 90% of patients have neuronal loss in the hippocampus (Figure 1). This patient group represents, however, only a small subpopulation of epilepsy patients. Currently, we do not have histological data about newly diagnosed patients with acquired symptomatic epilepsies. Magnetic resonance imaging (MRI) studies have shown that hippocampal or amygdaloid volume reduction can be seen in less than 10% of newly diagnosed patients. Other brain areas showing damage in histologic or MR images include the medial temporal cortex (entorhinal cortex, perirhinal cortex, parahippocampal cortex), thalamus, and cerebellum. All these studies point out that structural alterations in epileptic brain are not limited to the hippocampus.

Timing of Neurodegeneration in Epilepsy

In acquired symptomatic epilepsies the neurodegeneration can relate to acute damage caused by epileptogenic BDI (like TBI, stroke, infection, SE) or its progression over time. The contribution of recurrent seizures on neurodegeneration in epilepsy is still under dispute. Animal studies have shown that an electrical stimulation of the ventral hippocampus lasting for about 70 s can trigger apoptosis in the hippocampal CA1 subfield. Further, studies in kindling models suggest that recurrent evoked brief seizures can trigger progressive neurodegeneration (kindling refers to a phenomenon in which an initially subthreshold stimulus, when given repeatedly, eventually evokes partial and secondarily generalized seizures). This observations has not, however, been reproduced in all kindling studies. Also, studies in spontaneously seizing rodents have not been able to demonstrate seizure-associated neurodegeneration but rather indicate that most of the damage is caused by initial BDI. Also, there is no direct evidence that epileptic seizures cause neurodegeneration in human epilepsy. However, indirect evidence available from imaging studies in humans suggests that recurrent seizures can associate with progressive reduction in brain (and hippocampal) volume in a subpopulation

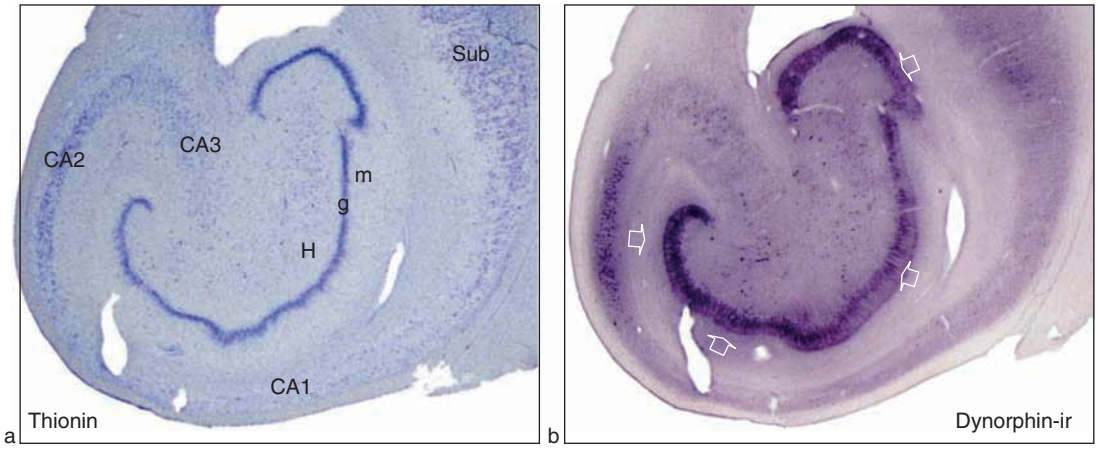

Figure 1 Brightfield photomicrographs from the hippocampus of a patient who underwent epilepsy surgery due to drug-refractory TLE. (a) A Nissl-stained section demonstrating severe neuronal loss in the hilus (H) of the dentate gyrus, as well as in the CA3 and CA1 subfields of the hippocampus proper. The CA2 subfield is better preserved. (b) An adjacent section stained with an antibody raised against dynorphin to demonstrate the sprouting of granule cell (g) axons or mossy fibers to the inner molecular layer (m) of the dentate gyrus (open arrows). ir, immunoreactivity; Sub, subiculum.

of patients. Whether volume reduction is exclusively due to neurodegeneration remains to be studied.

Mechanisms of Neurodegeneration in Epilepsy

Neuronal death in epileptogenic BDI can be caused by direct mechanical forces to the brain (like TBI), as well as by vascular, cellular, or molecular alterations that result in acute and delayed neuronal death or injury. Neurodegeneration can continue for months after BDI both in humans and experimental models. Importantly, the mechanisms can differ depending on the time delay from BDI with necrosis predominating at early time points. There are number of other factors that can influence neurodegeneration. For example, the severity of neurodegeneration following SE in mice heavily depends on the genetic background, with some strains more susceptible than others. It also depends on age, that is, young animals are less susceptible to pathologic insults, including SE as compared to adult rodents. There is also evidence that gender, type, and severity of injury, metabolic state of the brain, and other conditions (e.g., other diseases, medication) associated with BDI can affect the severity and/or mechanisms of neurodegeneration.

Morphologic Features of Neurodegeneration: Necrosis and Apoptosis

Morphologically, neuronal death can be divided into two major types: necrosis and apoptosis. The two categories provide some clues about the underlying mechanisms of neuronal death even though usefulness of morphological criteria as a starting point for understanding biochemical cascades has been challenged recently. Electron microscopic morphology of necrosis consists of swollen cellular appearance with swollen mitochondria, vacuolated cytoplasm, dilation of endoplasmic reticulum (ER), pycnotic nuclei, and plasma membrane rupture. Necrosis can be initiated, for example, by mechanical damage that leads to neuronal membrane failure, Ca^{2+} influx, and disruption of ionic homeostasis which triggers the release of glutamate. In addition, spillover of lysosomal cathepsins contributes to necrotic death. Inflammatory responses caused by the extracellular release of cytoplasmic components, and/or breakdown of the blood–brain barrier is another factor that exposes adjacent neurons to deadly signals and secondary damage (e.g., activation of caspase 1). Necrotic appearance is most common for cells dying within minutes to few days after epileptogenic BDI.

Morphologic features of apoptosis include nuclear and cytoplasmic condensation, internucleosomal DNA cleavage, and packing of the cell into apoptotic bodies that are engulfed by phagocytes, preventing release of intracellular components that is typically energy and transcription dependent. Apoptosis is rarely detected in degenerating neurons in SE models of epilepsy if morphological criteria are used. Nevertheless, such neurons display other apoptotic features, for example, DNA fragmentation has been found in the brain subjected to prolonged seizures. Also, activation of number of programmed cell death pathways via caspase-dependent or-independent mechanisms has been described (Table 1 and Figure 2).

Table 1 Summary of components of cell death pathways known to be involved in neurodegeneration in human or experimental TLE

	Human TLE	*SE*	*Kindling*
Caspase-dependent cell death			
Mitochondria-mediated pathway			
Cytochrome *c* release		⇑	
Apaf 1/cytochrome *c* complex formation		⇑	
Caspase-9		⇑ Activity	
Caspase-3	⇑ Protein expression	⇑ mRNA expression	⇑ Activity
	⇑ Proteolysis	⇑ Protein expression	
		⇑ Activity	
Caspase-6		⇑ Protein expression	
		⇑ Activity	
Caspase-7		⇑ Protein expression	
		No change in activity	
Death receptor-mediated pathway			
TNFR-1		⇑ Protein expression	
		⇑ Activity	
TRADD		⇑ Protein expression	
Fas		⇑ mRNA expression	
		⇑ Protein expression	
DAP kinase	⇑ Protein expression	⇑ Protein expression	
	Translocation	⇑ Proteolysis	
FADD		⇑ Protein expression	
SODD		⇑ Proteolysis	
Caspase-2		⇑ Protein expression	
		⇑ Activity	
Caspase-8		⇑ Protein expression	
		⇑ Activity	
ER-mediated pathway			
TRAF2		⇑ Activity	
Caspase-independent cell death			
Calpains			
Calpain		⇑ Activity	
Cathepsins			
Cathepsin D		⇑ mRNA	
Mitochondrial proteins			
AIF	⇑ Protein expression	⇑ Protein expression	
	Translocation	Translocation	
Bcl family proteins			
Bcl-2	⇑ Protein expression	⇑ ⇓ mRNA expression	
		⇑ Protein expression	
Bcl-xs		⇑ mRNA expression	
		⇑ Protein expression	
Bcl-xl	⇑ Protein expression		
Bcl-w		⇑ Protein expression	
Bax	⇑ Protein expression	⇑ ⇓ mRNA expression	⇑mRNA expression
		⇑ Protein expression	
		Translocation	
Bad		No change	
Bid		⇑ Proteolysis	
		Translocation	
Bim	⇓ Protein expression	⇑ Protein expression	
IAPs			
XIAP		⇑ Protein expression	
		⇑ Proteolysis	
PARP		⇑ Proteolysis	
Caspase-1	⇑ Proteolysis		
ICAD		⇑ Degradation	
CAD	⇑ Protein expression		
p53 pathway			

Continued

Table 1 Continued

	Human TLE	*SE*	*Kindling*
MDM2		⇑ Protein expression ⇑ Proteolysis	
p53		⇑ mRNA expression ⇑ Protein expression ⇑ Proteolysis Translocation	

AIF, apoptosis-inducing factor; Apaf-1, apoptosis protease activating factor-1; Bad, Bcl-2-associated death promoter; Bax, Bcl-2-associated X protein; Bid, BH3-interacting domain death agonist; Bim, Bcl-2-interacting mediator of cell death; CAD, caspase-activated DNase; FADD, Fas-associated death domain protein; ICAD, inhibitor of caspase-activated DNase; MDM2, murine double minute gene 2; SODD, silencer of death domains; TRADD, TNFR-associated death domain protein; TRAF2, (TNFR)-associated factor 2.

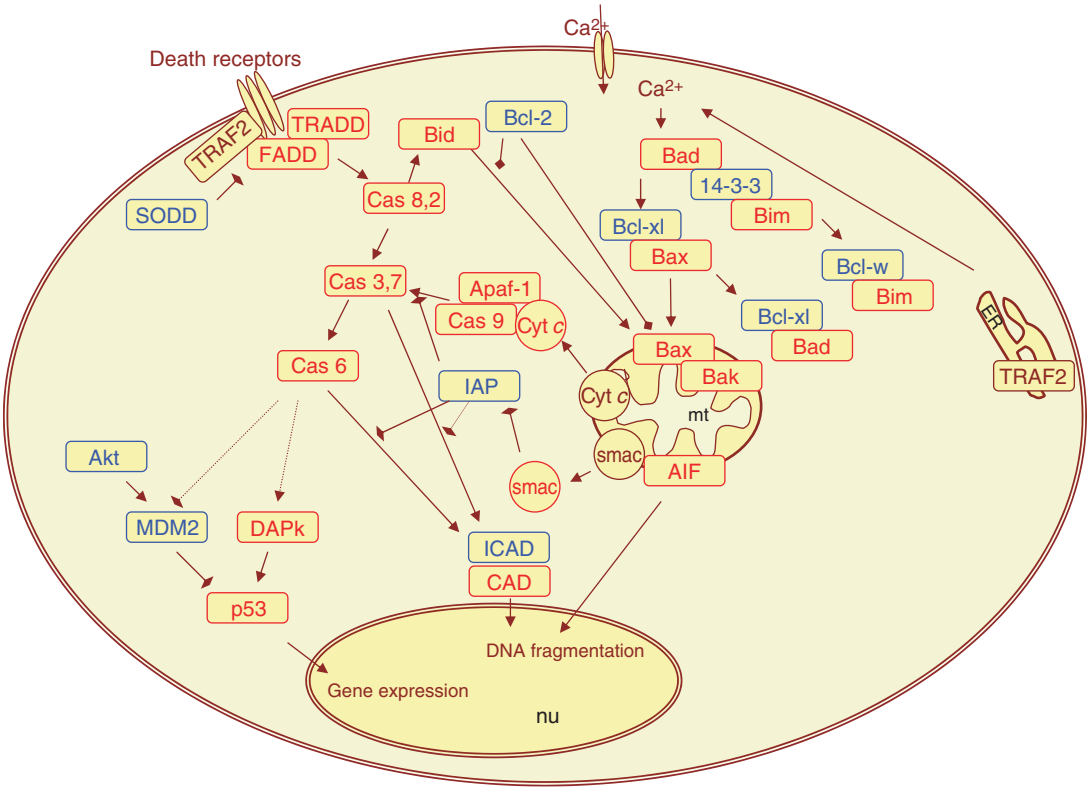

Figure 2 Summary of cell death pathways known to participate in SE-induced neurodegeneration (see the text for description). AIF, apoptosis-inducing factor; Apaf-1, apoptosis protease activating factor-1; Bad, Bcl-2-associated death promoter; Bax, Bcl-2-associated X protein; Bid, BH3-interacting domain death agonist; Bim, Bcl-2-interacting mediator of cell death; CAD, caspase-activated DNase; Cas, caspase; Cyt *c*, cytochrome *c*; DAPk, death-associated protein kinase; ER, endoplasmatic reticulum; FADD, Fas-associated death domain protein; ICAD, inhibitor of caspase-activated DNase; MDM2, murine double minute gene 2; mt, mitochondrium; nu, nucleus; smac, second mitochondria-derived activator of caspase; SODD, silencer of death domains; TRADD, TNFR-associated death domain protein; TRAF2, TNFR-associated factor 2.

Caspase-Dependent Cell Death in Epilepsy

A number of molecular pathways can be involved in execution of apoptosis. Here we focus exclusively on pathways that are of relevance for epilepsy-related neurodegeneration.

Apoptosis can be triggered by activation of death receptors localized in the cell surface (extrinsic pathway) or by dysfunction of mitochondria or ER (intrinsic pathway). Studies indicate involvement of both intrinsic and extrinsic pathways in SE-induced neurodegeneration (Table 1).

Extrinsic, death receptor-mediated pathway is activated by binding a ligand tumor necrosis factor alpha (TNF-α or FasL) to its receptor (TNFR-1 and Fas, respectively). This leads to formation of a scaffold complex (DISC), composed of several proteins such as TRADD, FADD, Rip, TRAF2, and results in activation of caspase 2, 8, or 10. TNFR-1 and Fas death receptors, as well as components of death receptors signaling, such as TRADD and FADD, are constitutively expressed in the brain. In addition, Fas and FADD are overexpressed following kainate-induced SE. Involvement of death receptor signaling pathway in SE-induced neurodegeneration is supported by evidence of formation in the scaffold complex. Such molecular complexes, isolated from the hippocampal tissue after SE, contain TNFR1, TRADD, FADD, and caspase 8. Also, TRAF2 and RIP can be recruited to this scaffold complex. Activity of above complexes can be blocked by FLIP or by SODD protein. Elimination of SODD protein from hippocampal neurons that was shown following kainate-induced SE would lead to activation of death receptor signaling. Finally, processing of caspase 8 and 2 – targets for death receptor signaling pathway – has been observed following SE.

Activation of mitochondrial intrinsic pathways involves release of cytochrome *c* from the mitochondria. In the cytoplasm, cytochrome *c* creates so-called apoptosome by binding to APAF-1 and subsequent recruitment and activation of caspase 9. Activity of this complex can be blocked by inhibitors of apoptosis protein family members (IAPs). Inhibitory action of IAPs can be blocked by smac/DIABLO, released from mitochondria. The release of cytochrome *c* from the mitochondria following prolonged seizures is a widely observed phenomenon. Also, formation of apoptosome and activation of caspase 9 have been detected in the hippocampus following intrahippocampal kainate-induced SE. Seizures can influence progression of intrinsic pathway acting on IAPs as well. For example, XIAP is initially upregulated and then degraded in hippocampal neurons dying in a consequence of kainic acid-induced SE.

In addition to mitochondria, ER too can participate in seizure-induced neurodegeneration, for example, by releasing Ca^{2+} to the cytoplasm.

Both, caspases 8 and 2 activated in death receptor cascades, as well as caspase 9 activated in mitochondrial intrinsic pathway are able to cleave, and therefore activate executioner caspases 3, 6, and 7. Expression and activation of caspase 3 in epilepsy-related neurodegeneration have been extensively studied in experimental models and in human tissue. Increase in mRNA expression, and protein expression or activity has been detected in hippocampus, temporal lobe, and thalamus. Activation of caspase 3 leads to degradation of its cellular targets like spectrin or inhibitor of caspase-activated DNase (ICAD), that was detected following kainic acid-induced SE. Degradation of ICAD is of special interest, since it results in disinhibition of its molecular partner, Caspase-activated DNase (CAD), translocation of CAD to the nucleus, and subsequent DNA fragmentation. Translocation of CAD to the nucleus has been shown in tissue resected from patients with TLE. It should be mentioned however, that some studies did not reveal activation of caspase 3 after seizures, and its role in seizure-induced neurodegeneration is still a matter of controversy.

Changes in protein expression of the other executioner caspase, caspase 7, have been found in the rat hippocampi following intra-amygdala kainate injection. However, it was not accompanied by caspase 7 activation, and therefore its role in seizure-induced neurodegeneration is not clear. Increase in protein expression and activity of caspase 6 has been observed in hippocampi of animals subjected to prolonged seizures as well.

Caspase-Independent Cell Death in Epilepsy

The Bcl-2 family comprises large number of proteins that are of great importance in regulation of apoptosis progression. Some of the members belonging to this family promote apoptosis (Bax, Bid, Bim), others have a opposite function (Bcl-2, Bcl-xl, Bcl-w). Activity of these proteins often depends on cellular localization and interactions with other proteins. Involvement of Bcl-2 family proteins in epilepsy-related neurodegeneration has been studied extensively. Increase in expression of proapoptotic Bax and Bcl-xs proteins was detected following prolonged seizures in animals, as well as in human TLE. Seizures cause dissociation of Bad from 14-3-3 protein and binding and neutralization of antiapoptic Bcl-l. Proteolysis of Bid occurs following seizures induced by intra-amygdalar injection of kainic acid. It results in translocation of processed fragment to mitochondria and promotion of cell death. In the same animal model increased expression of Bim was also observed. Interestingly, Bim protein expression decreases in human TLE. Also, expression and activity of antiapoptotic proteins can be altered by prolonged seizures. Increased expression of Bcl-2 and Bcl-w has been found in experimental models. Expression of Bcl-2 and Bcl-xl is elevated in human epileptic tissue.

Recently, the p53 pathway has emerged as one of importance in regulation of neuronal death. The p53 protein is a nuclear phosphoprotein acting as a transcription factor. It is a key player in cell response to

DNA damage and oxidative stress and it has been implicated in neurodegeneration. There are several mechanisms postulated for the function of p53 in neurodegeneration including activation of transcription of Bcl-2 family members or Fas. Both increased expression and activation of p53 in dying neurons have been observed as a result of pilocarpine or kainic acid-induced SE. Promotion of p53 function also occurs by SE-induced inactivation of MDM2, a protein which regulates p53 degradation. Number of other proteins regulate MDM2 or p53 activity in dying neurons, such as Akt or DAP kinase are influenced by prolonged seizures.

In addition to caspase-mediated proteolysis, proteolytic activity of calpains during neuronal death is frequently implicated. There is, however, only incidental information about involvement of this pathway in seizure-induced neurodegeneration.

Does Treatment with Antiepileptic Drugs Affect Neurodegeneration?

As there is ample evidence that neurodegeneration can be one component of progressive neuropathology in symptomatic epilepsies, a critical question is: can we affect the neurodegenerative process by treatment? In theory, number of antiepileptic drugs (AEDs) have the potential to influence cell death, for example, by inhibition of voltage-activated calcium channels (felbamate, phenytoin, topiramate), by inhibiting N-and/or P-type calcium channels (gabapentin, pregabalin, lamotrigine, oaxacarbazepine, phenytoin, and topiramate), or triggering antiapoptotic signalling pathways (valproate). There is some evidence from cell culture studies that valproic acid and lamotrigine reduce experimentally induced apoptosis.

The neuroprotective potential of AEDs following epileptogenic stimuli, such as SE, TBI, or ischemia have been studied in animal models of epilepsy. Beneficial action of selected drugs has been observed at least in some studies summarized in Table 2.

Number of studies have been directed at preventing neurodegeneration following ischemic stroke, which is one of the common epileptogenic insults. For example, carbamazepine, oxacarbazepine, or lamotrigine administered at 0.5 and 24 h after insult improved outcome in permanent middle cerebral artery occlusion (MCAO). However, in another study that used transient MCAO model, no effect of carbamazepine was observed following application at 30 min before and 90 min after insult. Clonazepam, felbamate, and lamotrigine are effective in the bilateral occlusion model in gerbil when administered shortly after vessel clamping. There are inconsistent data on the action of lamotrigine in MCAO. Both, reduction of infarct volume, as well as lack of effect have been reported. A neuroprotective effect of tiagabine was observed in global ischemia in gerbil. In rat, MCAO protective effects of tiagabine were visible even if drug was given within 2 h after reperfusion. Also, topiramate is protective if applied within 2 h after MCAO in rats. Administration of topiramate reduced hipppocampal damage in gerbils too. Vigabatrin was neuroprotective in bilateral carotid artery occlusion in gerbil, when assessed 1 week after insult, but this effect was lost by 28 days after ischemia. Finally, zonisamide is protective when administered up to 4 h after MCAO in rat, but no protective effect was seen in global forebrain ischemia in gerbil. Altogether, some AEDs alleviate ischemia-induced damage when severity of damage is assessed within few days after insult.

Most of the AEDs have been studied in SE models. These reports show that some of the compounds are neuroprotective when given during and/or after SE. For example, felbamate reduces damage when administered 10 min after beginning the perforant path stimulation. Lamotrigine was beneficial when given for 2 weeks after perforant path stimulation, but not after amygdala stimulation. No effect was noted

Table 2 Effect of AEDs on neurodegeneration caused by epileptogenic brain damaging insults. Only the studies in which the AED was administered 'during' *or* 'after' the insult induced *in vivo* are included

Drug	*SE*	*Ischemia*	*TBI*
Carbamazepine	n.d.	±0 or ⇓	n.d.
Clobazam	n.d.	n.d.	n.d.
Clonazepam	n.d.	⇓	n.d.
Ethosuximide	n.d.	n.d.	n.d.
Felbamate	⇓	⇓	n.d.
Gabapentin	⇓	n.d.	n.d.
Lacosamide	⇓	n.d.	±0
Lamotrigine	±0 or ⇓	±0 or ⇓	n.d.
Levetiracetam	±0 or ⇓	n.d.	n.d.
Oxcarbazepine	n.d.	⇓	n.d.
Phenobarbital	±0 or ⇓	n.d.	n.d.
Phenytoin	n.d.	±0 or ⇓	n.d.
Pregabalin	⇓ or ±0	n.d.	n.d.
Primidone	n.d.	n.d.	n.d.
Retigabine	⇓	n.d.	n.d.
Stiripentol	n.d.	n.d.	n.d.
Talampanel	n.d.	n.d.	⇓
Tiagabine	n.d.	±0 or ⇓	n.d.
Topiramate	±0 or ⇓	⇓	n.d.
Valproate	±0 or ⇓	n.d.	n.d.
Valrocemide	n.d.	n.d.	n.d.
Vigabatrin	±0 or ⇓	⇓	n.d.
Zonisamide	n.d.	±0 or ⇓	n.d.

AED, antiepileptic drug; n.d., no data available; SE, status epilepticus; ±0 no effect; ⇓, reduction in the severity of damage.

when lamotrigine was applied 30 min before kainic acid-induced SE. Application of levetiracetam for 21 days starting 30 min after pilocarpine-induced SE was not effective when damage was assessed 3 weeks later. Chronic application of lamotrigine in the same model had a mild neuroprotective effect on hipocampal principal cells and interneurons. Chronic treatment with pentobarbital, which started 24 h after kainic acid-induced SE and lasted for the subsequent 100 days did not influence hippocampal damage. Pentobarbital was, however, effective when administered simultaneously with kainic acid. In the Li-pilocarpine model of SE, pregabalin treatment that started 20 min after pilocarpine injection and continued for 6 days, reduced damage in the piriform and entorhinal cortices, but not in the hippocampus. Effect of retigabine was investigated following SE induced by kainic acid and terminated with diazepam. The protective effect was observed in the piriform cortex, but not in the hippocampus. Topiramate neuroprotection has been shown in the hippocampus following unilateral hippocampal stimulation when administered 140 min after SE and in Li-pilocarpine-induced SE when administered 1 and 10 h after SE. Topiramate has been also shown to improve sensorimotor outcome but the effect was not associated with a neuroprotective effect. Application of valproate for 1 day starting 2 h after perforant pathway stimulation did not influence hippocampal damage. Treatment with valproate for 21 days starting 30 min after pilocarpine-induced SE also did not reduce hippocampal damage. Vigabatrin was neuroprotective in pilocarpine model when treatment started 10 min after induction of SE and continued for 45 days. On the other hand, no protective effect of vigabatrin was detected if drug was applied 1 h or 1 week after kainic acid-induced SE. Vigabatrin treatment was not, however, neuroprotective when started 2 days after amygdala stimulation-induced SE and continued for 10 weeks. Also, a 4-week treatment starting 2 days following hippocampal stimulation-induced SE was ineffective.

The effect of AEDs on neuroprotection following TBI has not been extensively studied. The effect of talampanel was studied in fluid-percussion injury model. The damage was alleviated when the compound was given at 30 min but not at 3 h after trauma.

Overall, neuroprotection evoked by AEDs is mild and in most cases can be detected only within few days after the insult. Frequently, the data available are inconsistent, which can be attributed to animal models or methodologies used to quantify the severity of neurodegeneration. Neuroprotective effects do not relate to the mechanism of drug action. One can speculate that in many cases the favorable effects likely associate with suppression of seizure activity at acute phase. Moreover, neuroprotection does not imply favorable functional outcome. Importantly, there is no hard experimental evidence that successful neuroprotective treatment would delay or prevent epileptogenesis.

The relevance of experimental data for human condition is difficult to assess. Clinical studies that have evaluated neuroprotective effects of AEDs are rare. One clinical study investigated the effect of carbamazepine or vigabatrin monotherapy on hippocampal atrophy over a 3-year follow-up. No difference was found between the treatments. In accord with experimental evidence, AEDs do not influence epileptogenesis in humans when studied in patients with TBI.

Conclusions

Experimental studies have revealed that neurodegeneration occurs both by necrotic and apoptotic mechanisms after BDIs leading to epilepsy. Most of the damage occurs soon after BDI but some damage can continue for weeks to months, even after the appearance of spontaneous seizures. Induced seizures in rodents have been shown to trigger neurodegeneration. Whether spontaneous seizures contribute to neuronal death is still under dispute. Studies in humans have revealed that progressive atrophy of the brain can occur in a subpopulation of patients with symptomatic epilepsies over time. Human studies are, however, compromised by the resolution of current imaging techniques, and it remains to be seen which component of the atrophy is actually caused by degeneration of neuronal somata. Also, further studies are needed to address a question whether treatment with AEDs or other compounds would have effect on atrophy and whether that would associate with more favorable outcome in patients with epilepsy.

See also: Apoptosis in Neurodegenerative Disease; Epilepsy; Epilepsy: Channelopathies; Hippocampus; Mitochondrial Dysfunction in Nervous System Injury.

Further Reading

Cory S and Adams JM (2002) The Bcl2 family: Regulators of the cellular life-or-death switch. *Nature Reviews Cancer* 2(9): 647–656.

Fujikawa DG (2005) Prolonged seizures and cellular injury: Understanding the connection. *Epilepsy and Behavior* 7(supplement 3): S3–S11.

Henshall DC and Simon RP (2005) Epilepsy and apoptosis pathways. *Journal of Cerebral Blood Flow and Metabolism* 25(12): 1557–1572.

Liu RS, Lemieux L, Bell GS, et al. (2003) Progressive neocortical damage in epilepsy. *Annals of Neurology* 53(3): 312–324.

Lukasiuk K and Pitkänen A (2004) Large scale analysis of gene expression in epilepsy research: Is synthesis already possible? *Neurochemical Research* 29(6): 1169–1178.

Morrison RS, Kinoshita Y, Johnson MD, Guo W, and Garden GA (2003) p53-dependent cell death signaling in neurons. *Neurochemical Research* 28(1): 15–27.

Pitkänen A and Kubova H (2004) Antiepileptic drugs and neuroprotection. *Expert Opinion in Experimental Therapy* 5(4): 777–798.

Pitkänen A and Sutula T (2002) Is epilepsy a progressive disease? Prospects for new therapeutic approaches in temporal lobe epilepsy. *Lancet Neurology* 1: 173–181.

Sutula T and Pitkänen A (eds.) (2002) *Do Seizures Damage the Brain? Progress in Brain Research*, vol. 135. Amsterdam: Elsevier.

Episodic Memory

M E Wheeler and E J Ploran, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Of all human capabilities, the one central to our development of a unique sense of personal identity is episodic memory. The ability to form and retrieve memories of our life experiences provides an opportunity to identify those events as our own and embed the memories in a malleable framework of related experiences. This process in turn influences our interpretation, and therefore learning, of new experiences. Indeed, people who have lost the ability to access their past episodic memories have sometimes reported feeling a loss of self identity. Over the past 50 years, significant advances in the philosophical, theoretical, and empirical study of memory have begun to shed light on the nature of memory and the intricate cognitive and neural mechanisms involved in remembering past episodes.

What Is Episodic Memory?

Episodic memories are consciously perceived memories of personal events or episodes. One defining feature of episodic memories is that they are the personal memories. Among other types of information, they include where you were, what you were doing, whom you were with, what you saw, how you felt, and what you were thinking during a particular event. A second defining feature is that episodic memories are accompanied by a conscious awareness of the temporal context in which the events occurred. In bringing to mind events from childhood, there is an awareness that the events occurred before the present time, and perhaps an awareness of precisely when they occurred.

Episodic memories represent one type of declarative memory – memories which are accessible to consciousness and which can be readily communicated or declared to others. Nondeclarative memories, such as learned skills, are memories that are not directly available to consciousness but nevertheless influence behavior. Semantic memories are memories of factual knowledge and represent a second type of declarative memory. Declarative memory is dependent upon the integrity of medial temporal lobe (MTL) structures including hippocampus, entorhinal and perirhinal cortex, and parahippocampal gyrus, and also on the fornix and anterior and mediodorsal nuclei in the thalamus. Significant damage to these structures produces amnesia and consequently an impaired ability to form and retrieve declarative memories. By some accounts, episodic and semantic memories arise from different memory, and by extension, brain systems. These accounts are supported in part by behavioral dissociations and in part by evidence that some patients are impaired on tests of episodic but not semantic memory while other patients are impaired on tests of semantic but not episodic memory. Further support for different memory systems has come from neuroimaging studies which demonstrate dissociations among brain regions associated with episodic and semantic memory tasks. By other accounts, episodic and semantic memories differ according to the presence or absence of personal spatial and temporal context, but are otherwise similar. When the initial context has been lost or is inaccessible, or the current situation does not encourage conscious recollection, the result is retrieval of semantic knowledge. By this account, the difference is due to the different processing demands of a given situation rather than differences in the underlying memory systems. These accounts are not necessarily mutually exclusive.

A distinction is often drawn between memories of personal events accompanied by subjective awareness, and memories of personal facts related to those events. For example, you may remember purchasing your first car, bringing back to mind where you found it, how it appeared, and how the seller talked. However, some of these remembered details can also be retrieved as facts. You know that you bought your car from an individual rather than a dealer. You know that it was blue. And you know that the seller was a man. Another distinction is sometimes made between episodic memories pertaining to a single event (recollections) and those pertaining to repeated similar episodes (generic personal memories). Autobiographical memory is often used synonymously with episodic memory. However, in some instances, the two terms are used to denote memories of life experiences and laboratory events, respectively.

Reconstructing Memories Using Retrieval Cues

Prevailing frameworks of episodic memory describe remembering as a dynamic process during which an active model or simulation of an event is constructed from memory traces. A memory trace, also called an engram, is a neural change reflecting the thoughts, emotions, and sensory features of the original experience. The construction account of

remembering rejects the proposition that memory traces are replayed like a song or a movie. Instead, memory cues are used to create a mental simulation of an event by drawing upon available memories and filling in the missing pieces with educated guesses. Ulric Neisser compares this process to that of assembling a dinosaur from mere bone chips. The reconstruction process begins with a retrieval cue, and subsequent retrieval is influenced by an individual's past experiences, his or her current state of mind, and the contextual demands in which remembering occurs. For example, accessing the same memory trace on two different occasions may produce qualitatively different recollections if the first occasion is immediately following a transcontinental flight and the second is after a full night of rest. The German biologist Richard Semon coined the term ecphory to describe this synergistic interaction of the original memory trace and the current retrieval cue to form the episodic memory experience. Because of an inherent dependence on the retrieval cue, memories entering consciousness need not be veridical representations of the mnemonic content of a stored engram. By this view, only the products of the interpretive interaction of current intent and past experience are available to awareness.

The effectiveness of a retrieval cue depends in large part upon the degree of similarity between the cue and the original event. The more similar the relationship, the more likely it is that the retrieval cue will lead to ecphory. Therefore, the way in which events are interpreted at encoding influences how they can later be accessed and reconstructed. Endel Tulving refers to this relationship as the encoding specificity principle. Daniel Schacter provides a memorable example of this principle by describing an experiment carried out by the artist Sophie Calle at The Museum of Modern Art. In the experiment, museum staff were asked to recollect as much detail as they could about various paintings. The level of detail provided by the staff differed widely, from just a few words to detailed sentences. Calle found that the degree and quality of detail was related to the type of work carried out by the staff member. Staff in charge of art placement provided structural descriptions of the paintings, while those in charge of procuring the art provided more detailed accounts and interpretations of the artwork. The differing experiences of the museum staff influenced how they thought about the artwork, and this in turn influenced the types of memory they were later able to retrieve. According to the encoding specificity principle, the cues that will best evoke episodic memories of the paintings will be those that most closely match the initial processing.

Components of Remembering

Episodic remembering is typically perceived as a unitary phenomenon. A memory comes to mind, sometimes easily and sometimes after some effort, but the details of the process are often unavailable to superficial introspection. There is an abundance of research demonstrating, however, that remembering depends on a variety of memory-specific and domain-general cognitive abilities, including among others, formation of search strategies, memory-based decision making, and postretrieval assessment. The most convincing evidence for separable components in human episodic memory has come from studies that produce a dissociation, either by identifying patients in whom lesions to different parts of the brain produce different types of memory deficit, or by identifying brain areas in which neural activity modulates according to variations in one component differently than the other components.

Strategic Processing

At times, memories of past events come to mind rather easily, seeming to pop into awareness after an encounter with a retrieval cue. At other times, a feeling of knowing can lead to a more strategic search through memory for the appropriate information. Strategic search involves generating retrieval cues from memory and using those cues to guide further searches. Models of strategic processing in remembering typically describe three basic stages, including initial identification of the desired information (target), a search for information and a matching process in which retrieved information is compared with the target, and an assessment or monitoring of the degree of accuracy. Neuropsychological studies of patients with lesions in the prefrontal cortex (PFC) have provided valuable insight into the role of frontal cortex in memory. Frontal patients are readily able to encode and retrieve memories. In fact, they tend to perform at levels approaching that of control subjects on simple item recognition tests that can be accomplished using nonrecollective memory (i.e., familiarity). However, patients with prefrontal damage perform significantly more poorly than control subjects on memory tests that require strategic processing, such as source memory tests of cued recall, free recall, and recognition. Source memory tests require the ability to identify the appropriate learning context, such as the time (was this item studied in the first or second study session?), place (was this item studied in the laboratory or in the classroom?), or feature (was this item presented in a male or female voice?). In these frontal patients, memory deficits are due not

to problems forming memories (as in amnesia) but instead are related to impairments of organization of new memories, strategic access to memory contexts, and the supervisory control required to maintain information in working memory and assess retrieval effectiveness. As a result, frontal patients can experience difficulty under situations that require flexible processing, such those required by source memory tasks. This difficulty is sometimes observed in the form of confabulations – the generation of false and contradictory statements.

Results from neuroimaging studies are consistent with the neuropsychological findings. Because the need for strategic search typically increases as retrieval difficulty increases, retrieval difficulty can be manipulated in order to examine the neural substrates of strategic processing in memory. For example, difficulty can be manipulated by varying encoding strength so that weakly encoded items are more difficult to retrieve than strongly encoded items, or by varying the amount of detail required so that retrieving more detail is more difficult. These studies demonstrate that frontal lobe regions near the anterior extent of the left inferior frontal gyrus (aIFG), left anterior prefrontal cortex (aPFC), and bilateral posterior middle frontal gyrus (pMFG) consistently increase activity as retrieval difficulty increases. There is also evidence for functional specialization within this set of regions. pMFG activity tends to also modulate according to the type of information retrieved, with a preference toward processing of verbal stimuli in the left hemisphere and nonverbal stimuli in the right. Other brain regions also appear to play an important role in strategic processing. Bilateral anterior insula/frontal operculum (AI/FO), the anterior cingulate cortex (ACC), and basal ganglia (bilateral caudate nucleus, putamen) also tend to increase activity as retrieval difficulty increases. While there has been relatively little attention given to the role of AI/FO and basal ganglia in these studies, the ACC has received quite a bit of attention. Current models suggest that ACC is involved in reward-based decisions, motor control, error detection, emotional self-control, problem solving, and/or conflict monitoring. The demand for all of these functions conceivably increases as retrieval difficulty increases, and therefore often it is a particular challenge to clearly identify the source of signal modulations.

Contents of Remembering

There is a host of evidence from neuropsychological, single-unit physiology, event-related potential (ERP), and neuroimaging studies supporting the view that regions of the nervous system that initially process an event (i.e., data-processing regions) are reactivated when the event is later remembered. For example, recalling your car's seller involves the coordinated recruitment of regions of the brain that initially processed the appearance of his face, the color of the car, the sound of his voice, and your thoughts about the situation. Work on human and nonhuman primates is particularly relevant to episodic remembering and mental imagery, but extends also to semantic knowledge. The findings indicate that memory traces are distributed in the nervous system as a function of the type of material encoded. For example, patients with focal lesions can exhibit impaired memory for visual objects, yet show preserved memory for spatial relationships. Neuroimaging experiments have shown that data-processing regions that are active during initial encoding of stimuli are, to some extent, reactivated during later remembering. It is yet to be determined how such a coordinated reactivation is organized and controlled, but prominent theories propose an interaction of frontoparietal attentional control processes and a medial temporal associational mechanism that provides an index of stored memory traces.

An unresolved issue concerning how memory contents are reconstructed is related to the degree to which data-processing centers are reactivated. In the visual system, for example, the most basic aspects of visual stimuli are processed in early stages, with complexity and receptive field size increasing as information is transmitted to later stages. Thus, while fundamental stimulus features such as line segments located in discrete parts of the visual field are processed in early stages of the visual processing hierarchy, more complex features that are less dependent upon location in the visual field, such as object form, are processed in later stages. Evidence that visual memories are reconstructed from the bottom up arises from studies that show involvement of early and late regions during visual retrieval. By this view, remembering the appearance of your car involves reconstructing the visual image from basic featural components up to the whole object. Other studies, however, indicate that visual memories are reconstructed by recruiting the levels most specialized to process the desired information. That is, remembering the general appearance of your car is accomplished by recruitment of higher-level object-processing regions, while remembering the color is accomplished by recruitment of color-processing regions.

Retrieval Success (Ecphory)

Episodic memory retrieval is associated with a conscious awareness of reexperiencing an event that occurred in the past. As noted previously, it has been

convincingly demonstrated that the MTL plays a critical role in retrieval success. Recent studies using ERP and neuroimaging techniques have identified a set of neo- and subcortical areas that also play a role in retrieval success. ERP studies of recognition memory were the first to note a distinct set of frontal and parietal signals related to the successful identification of previously studied items. This old/new, or retrieval-success effect, has been further studied in studies using ERP, positron emission tomography (PET), and functional magnetic resonance imaging (fMRI). The ubiquitous finding across studies is that recognizing old versus new items modulates activity in left PFC, as well as lateral and medial posterior parietal cortex. In PET and fMRI studies, frontal activations have been identified in bilateral dorsolateral PFC, medial PFC, and frontal polar cortex. Parietal activations have been consistently found in the anterior intraparietal sulcus and lateral parietal cortex near the supramarginal gyrus. While the findings are significant because they guide future research, the functional role of each of these regions has not been elucidated by the old/new manipulation.

In dual-process accounts of memory, recognition can arise from the fundamentally different memory processes of recollection and familiarity. Recollection is the ability to retrieve episodic memories, whereas familiarity is a strength-based signal that information is old. Familiarity is not supported by recollective evidence. Source memory tests, and the remember/know test developed by Endel Tulving, have been used to successfully dissociate these processes and further characterize the role of retrieval-success regions. Importantly, activity in the hippocampus has been consistently tied to memory decisions based on recollection, but not to those based on familiarity/knowing in the absence of recollection. In contrast, a related region of MTL near the rhinal cortex has been associated more with familiarity-based decisions. The neuroimaging results are consistent with an extensive animal literature that supports a role of hippocampus in recollection, and perirhinal cortex in nonrecollective recognition decisions. However, as there is also evidence that patients with lesions restricted to hippocampus proper are impaired on recognition memory, it is not currently known precisely how different MTL structures contribute to memory.

Somewhat surprisingly, ERP and fMRI studies suggest that parietal regions involved in old/new decisions are differentially related to whether those decisions were based on recollection or familiarity/knowing. ERP studies tend to indicate that parietal signals are related to recollection rather than familiarity. This general finding is supported by fMRI such that lateral parietal cortex near the supramarginal gyrus has been shown to respond preferentially to decisions based on recollection. The supramarginal regions have also been reported in studies evaluating temporal judgments and temporal orienting of attention, suggesting that they could be involved in providing temporal context during remembering, allowing us to mentally travel through time. Functional MRI studies also suggest that slightly more medial parietal cortex, near the anterior intraparietal sulcus (particularly in the left hemisphere), is similarly responsive to decisions based on recollection and familiarity. Interestingly, the literature on patients with posterior parietal lesions offers no supporting evidence for parietal involvement in long-term memory processes. Instead, parietal lesions tend to disrupt visuospatial and visuomotor abilities. There is evidence, however, that lesions affecting the angular gyrus in the left hemisphere impair short-term memory. In addition, studies of epileptic patients indicate that seizures originating in posterior parietal cortex can be associated with déjà vu or a disturbance in consciousness. This observation is broadly consistent with neuroimaging experiments that have found posterior parietal activity related to temporal attention. Thus, posterior parietal cortex may play a role in establishing temporal context during remembering.

Retrieval Mode

What would life be like if every stimulus evoked a memory of a past experience? As noted previously, ecphory occurs when there is an interaction between a retrieval cue and a stored memory trace. A retrieval cue can be any stimulus in the internal or external environment, including a word spoken by a companion, the sight of a flower, or a thought about an aching joint. Most stimuli we encounter, however, do not evoke episodic memories. In fact, because there would be little opportunity to adequately process and interpret relevant stimuli, it could be quite detrimental to basic survival if every stimulus cued an episodic memory.

At the same time, it is possible to allow stimuli to serve as cues to remember past events. One could freely associate by allowing each thought entering awareness to serve as a cue to remember a past experience. Or one could look at items in the room and think about how they were obtained. In both cases, the thoughts and objects do not typically evoke episodic memories. However, this simple suggestion could have suddenly turned an object on your desk into a cue that sparked a memory about where you purchased it. The all-or-none state in which

stimulus cues serve to evoke memories is known as retrieval mode.

Studies using neuroimaging, ERPs, and neuropsychological techniques suggest that the most anterior portions of the PFC are important for engaging in retrieval mode. A large number of neuroimaging studies using PET have consistently found right aPFC activity during episodic memory tests. aPFC was not found to modulate depending upon whether retrieval is successful or unsuccessful, indicating that it is not important for retrieving individual items from memory. Recent studies using fMRI and ERP add some insight into the role of aPFC by showing that right aPFC is active in a sustained manner, throughout the testing session and independently of the transient trial responses. aPFC could therefore be important for maintaining task-level rather than trial-level processes, playing a role in maintaining retrieval mode.

Associative Nature of Memory

Some theories of episodic memory consider the connections between the facts of the memory and the context in which they were encoded (i.e., time, place, emotional state), as well as the connections between items presented simultaneously. Since we are constantly presented with many items at once, it may not be the case that each memory trace is completely separate from the others; it is possible that each memory item can be related to many other items for organizational and efficiency reasons. Imagine if each chapter in this encyclopedia had to describe each piece of the brain it references and how certain processes work. This would add length and complexity to every chapter. Instead, the use of linking each chapter to related items allows each chapter to describe a specific piece while the reader can reference related pieces as necessary to form a whole. In this way, each related memory item can potentially serve as a recollective cue for the others in the organized group. While this can be true for the network of declarative facts alone, it can also connect declarative facts to the contextual items that allow for episodic memory. For instance, in the first car example given above, perhaps you remember that the salesperson was nervous because he had a very clammy handshake. This would be an example of the connection between a declarative fact (the seller was nervous) and a contextual fact (your disgust at his handshake) to form an episodic memory.

Several computational models attempt to use an associative system to represent the process of memory. One example is Lynne Reder's source activation confusion (SAC) model of memory, which suggests that people create different nodes of information at the time of encoding. One type of node contains the basic semantic knowledge of the item, while another type contains the contextual elements of the encoding episode. A memory within the SAC model can have many nodes and associations between these nodes, as well as common nodes that are associated with many memories (e.g., all the words studied during in an experiment might be connected to a context node encompassing the experiment room, computer, lighting, etc). A person experiences an episodic recollection when not only the item node is reinstated during recall, but also nodes containing information about the context. It is the activation of the association between the item and context nodes that makes the difference between remembering an event episodically versus simply a feeling of familiarity. It has also been suggested that a lack of association, or binding, of elements into memory can cause a lack of episodic experience at a later time. Other models that focus on the interconnected nature of memory include the search of associative memory (SAM) model and versions based on parallel distributed processing (PDP).

Neuroimaging studies have sought to investigate the neural mechanisms behind the creation of associations among items to be remembered. These studies involve presentation of multiple stimuli at one time and participants either study the items singly or as a group. The hypothesis is that studying each item singly does not prompt the creation of associations with the other items, while studying the items as a group does. These studies have shown that the hippocampus and parahippocampal gyrus (particularly on the right) are more active during associative encoding than single item encoding. Other neuroimaging studies have focused on the encoding and retrieval of contextual information about the remembered items. A variety of paradigms can be used to test recollection of contextual information, and often include a question at test regarding the temporal order of items, list membership, or color/font/spatial location at study. These studies have also found activation in the hippocampus and parahippocampal gyrus. Other areas of interest include bilateral dorsolateral PFC and parts of parietal cortex, which may be involved in the more specific contextual memories of spatial locations.

See also: Declarative Memory System: Anatomy; Hippocampus and Neural Representations; Hippocampus; Humans; Memory Representation; Memory Consolidation: Systems; Memory Disorders; Prefrontal Cortex: Structure and Anatomy.

Further Reading

Aggleton J and Brown M (1999) Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and Brain Sciences* 22: 425–489.

Brewer W (1996) What is recollective memory? In: Rubin D (ed.) *Remembering Our Past: Studies in Autobiographical Memory*, pp. 19–66. New York: Cambridge University Press.

Buckner RL (2003) Functional–anatomic correlates of control processes in memory. *Journal of Neuroscience* 23: 3999–4004.

Buckner RL and Wheeler ME (2001) The cognitive neuroscience of remembering. *Nature Reviews Neuroscience* 2: 624–634.

Burgess PW and Shallice T (1996) Confabulation and the control of recollection. *Memory* 4: 359–411.

Moscovitch M, Rosenbaum RS, Gilboa A, et al. (2005) Functional neuroanatomy of remote episodic, semantic and spatial memory: A unified account based on multiple trace theory. *Journal of Anatomy* 207: 35–66.

Schacter DL (1996) *Searching for Memory.* New York: BasicBooks.

Shimamura AP (2002) Memory retrieval and executive control processes. In: Stuss D and Knight R (eds.) *Principles of Frontal Lobe Function*, pp. 210–220. Oxford: Oxford University Press.

Squire L (1987) *Memory and Brain.* New York: Oxford University Press.

Squire L (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* 99: 195–231.

Tulving E (1983) *Elements of Episodic Memory.* Oxford: Clarendon Press.

Tulving E and Craik F (eds.) (2000) *The Oxford Handbook of Memory.* Oxford: Oxford University Press.

Velanova K, Jacoby LL, Wheeler ME, McAvoy MP, Petersen SE, and Buckner RL (2003) Functional–anatomic correlates of sustained and transient processing components engaged during controlled retrieval. *Journal of Neuroscience* 23: 8460–8470.

Wheeler MA, Stuss DT, and Tulving E (1997) Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin* 121: 331–354.

Wheeler ME, Petersen SE, and Buckner RL (2000) Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11125–11129.

Episodic Memory: Assessment in Animals

M Muñoz, University College London, London, UK
R G M Morris, University of Edinburgh, Edinburgh, UK

© 2009 Elsevier Ltd. All rights reserved.

Episodic and Episodic-like Memory

The concept of episodic memory was introduced to refer to the capacity to remember the unique events of our lives. The memory of a sequence of events collectively comprises an episode. As originally defined, an episodic memory is a mental representation of 'what' happened during an event, 'where' it happened, and 'when.' Writing in 1983, Tulving asserted that episodic memory "receives and stores information about temporally dated episodes or events, and temporal–spatial relations among these events." Thus, remembering a summer trip to France with a friend would be an instance of episodic memory. However, the concept has developed since its original definition, and it now refers to a 'memory system' whose operation has additional attributes such as the ability to travel mentally through time and a specific form of conscious awareness during the act of recall. In both its original and modern conceptions, episodic memory can be considered a form of declarative memory because the information within it is propositional. However, it is contrasted with another form of declarative memory – factual knowledge or semantic memory – which differs in not being bound to any specific temporal–spatial context. The brain mechanisms involved in these two forms of memory are likely to differ but overlap.

It has been claimed by Tulving, and separately by Suddendorf and Corballis that episodic memory is unique to humans. This 'uniqueness hypothesis' emerges from the supposition that only humans are capable of mental time travel and a self-knowing form of consciousness called 'autonoesis.' 'Mental time travel' refers to the phenomenological experience of traveling in the mind through time to that moment in the past when a specific event happened and, importantly, to travel prospectively and so imagine future events and needs. A person's experience is then one of being simultaneously in the present (the here and now) while also reexperiencing the past or the future (the memory, or imagined event) – a mental coexistence that transcends time. It is unclear whether any nonhuman animals are capable of doing this. The possession of language, another unique attribute of humans, enables the content of episodic memories be shared between persons in a way that is impossible in animals. However, it is not regarded as necessary for episodic memory and so it may be possible to develop novel, strictly behavioral tests of these additional features of episodic memory, without recourse to language, that could be tested in animals (e.g., future planning).

However, a 'comparative hypothesis' of episodic memory should also be considered. An episodic 'memory system' may be a collection of distinct neural subsystems in both humans and animals. The comparative perspective championed independently by Aggleton and Brown, Clayton and by Morris, holds that 'episodic-like' memory systems exist in animals that differ in expressing some but not all the attributes of human episodic memory. These subsystems, which may include associative mechanisms with which episodic memory is sometimes contrasted, may also be expressed in distinct ways in different species (birds, rodents, monkeys, and apes) as a function of their ecology, social structure, and so on. The reason for considering this alternative is that animals, no less than humans, also need to remember and later recall unique events (e.g., the location of a briefly seen predator). A memory system that could do this would be of adaptive significance but also likely to be distinct from the one mediating the stable body of knowledge about the world that is analogous to semantic memory in humans. Even monkeys, which are generally assumed to lack autonoesis (they fail the test of recognizing themselves in a mirror), still need a rapid, automatic system for encoding and later recalling unique events. Such a system would also be valuable for keeping track of where transiently available food or other resources are located in space and of social relationships in a troop of apes. Rodents and birds may also possess forms of memory that are episodic-like in character. The comparative hypothesis accepts that 'true' episodic memory may be unique to humans, but it argues against an inference of the uniqueness hypothesis that all forms of propositional memory in animals are necessarily semantic in character. Episodic-like and semantic-like systems would then exist in different species with a similar interdependence to that shown by the respective systems in humans.

Support for the comparative hypothesis comes from neuroanatomical, neurophysiological, and behavioral data. First, all mammals possess a number of brain structures that studies in humans have revealed to be implicated in declarative memory. Some of these may be specifically engaged during episodic-like encoding and recall, whereas others could contribute differentially to semantic-like memory. Second, these brain structures must possess specialized neurophysiological mechanisms that are essential for the online operations

of a rapid episodic-like memory system. Third, objective studies of animals in the laboratory have revealed a range of behavioral capacities that imply the possession of a memory system in vertebrates that can encode and later recall information about unique events, about their order in a sequence, and about the temporospatial context in which they happened. We now consider these three separate strands of the 'comparative' hypothesis in turn.

Neuroanatomy

Main Brain Structures Involved in Episodic-like Memory and Their Connectivity

Anatomical studies of the hippocampal region date back to the beginning of the twentieth century, but it was not until its second half that clinical reports of patients who suffered amnesia as a consequence of medial temporal lobe removals revealed its importance for episodic memory. From then onward, the neuroanatomy of the medial temporal lobe became the neuroanatomy of memory. Research on anatomical organization of the medial temporal lobe of the primate has revealed the connections between hippocampus and the parahippocampal region with the cerebral cortex and with subcortical areas also thought to be critical for episodic-like memory formation (Figure 1).

High-order processing areas of the cerebral cortex, including perirhinal and parahippocampal cortices, superior temporal gyrus, medial frontal, orbitofrontal, insular, retrosplenial, and inferior parietal cortices, send projections to the entorhinal cortex. Neurons from layer II of the entorhinal cortex originate the major afferent projection to the dentate gyrus. This projection forms the perforant pathway to the dentate gyrus' molecular layer, where individual axons terminate en passant. The en passant rather than a topological arrangement means that individual cells in the entorhinal cortex make contact with a large number of cells in the dentate; indeed, there is also a further threefold expansion in the number of dentate neurons relative to those in entorhinal cortex. The granular cells of the dentate gyrus send axons on to the CA3 field of the hippocampus (i.e., the mossy fibers) whose pattern of termination is strikingly different from that in the dentate, with individual

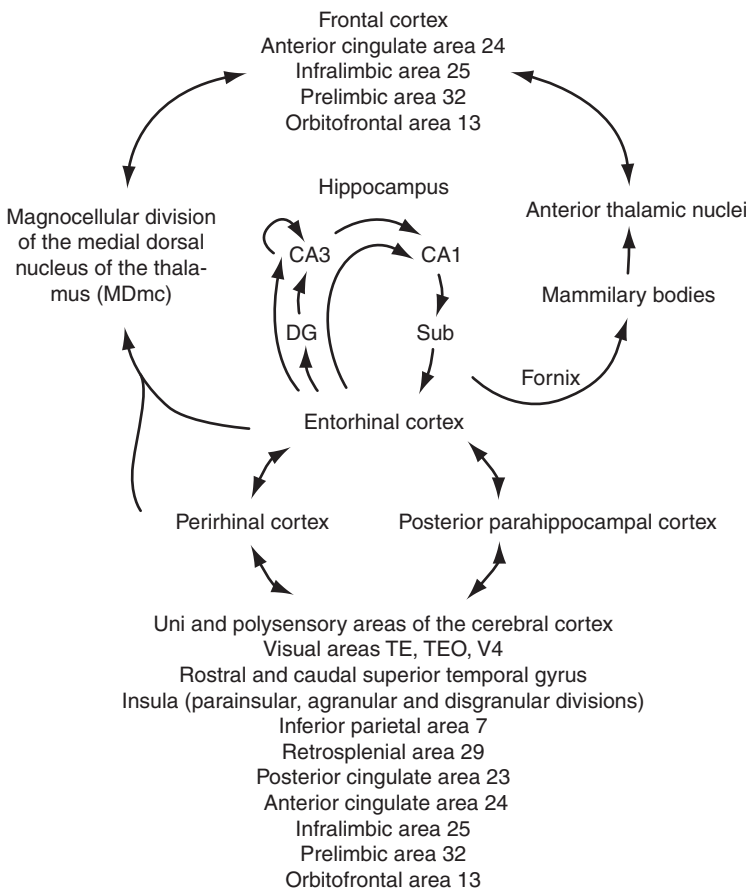

Figure 1 The neuroanatomy of episodic-like memory in nonhuman primates. The diagram summarizes the cortical and thalamic connections of the hippocampus and the entorhinal, perirhinal, and posterior parahippocampal cortices.

fibers making no more than approximately 15 axon terminals. Pyramidal cells in CA3 possess massive recurrent collaterals that can innervate other CA3 neurons along its entire anteroposterior axis, a feedback arrangement that some neural network models have suggested is ideal for the algorithms of 'pattern completion' and 'sequence learning.' Area CA3 pyramidal neurons also send collaterals, the Schaffer collaterals, to fields CA1 and CA2 of the hippocampus, where they terminate in the apical dendrites of pyramidal neurons. The neurons in area CA1 receive, in parallel, a projection from neurons of layer III of the entorhinal cortex bypassing all the processing stages of intrinsic hippocampal circuitry. The unidirectional nature of the connectivity, quite different from that in the neocortex, is striking but its functional implications are not clear. These connections reflect the potential complexity of associations that may take place within the hippocampus and they have inspired many computational models of memory formation (Figure 2).

Once the information is computed by the hippocampus, it is sent back to the cerebral cortex, where it is ultimately stored. Pyramidal cells of areas CA1/subiculum are the main origin of this projection. Their major cortical targets are neurons in the deep layers of the entorhinal cortex (V and VI), but they also send parallel projections to high-order processing areas of the cerebral cortex. Although this 'bypass' projection to the cerebral cortex is smaller in scale than that leaving the hippocampus and subicular cortex via the entorhinal cortex, it may be

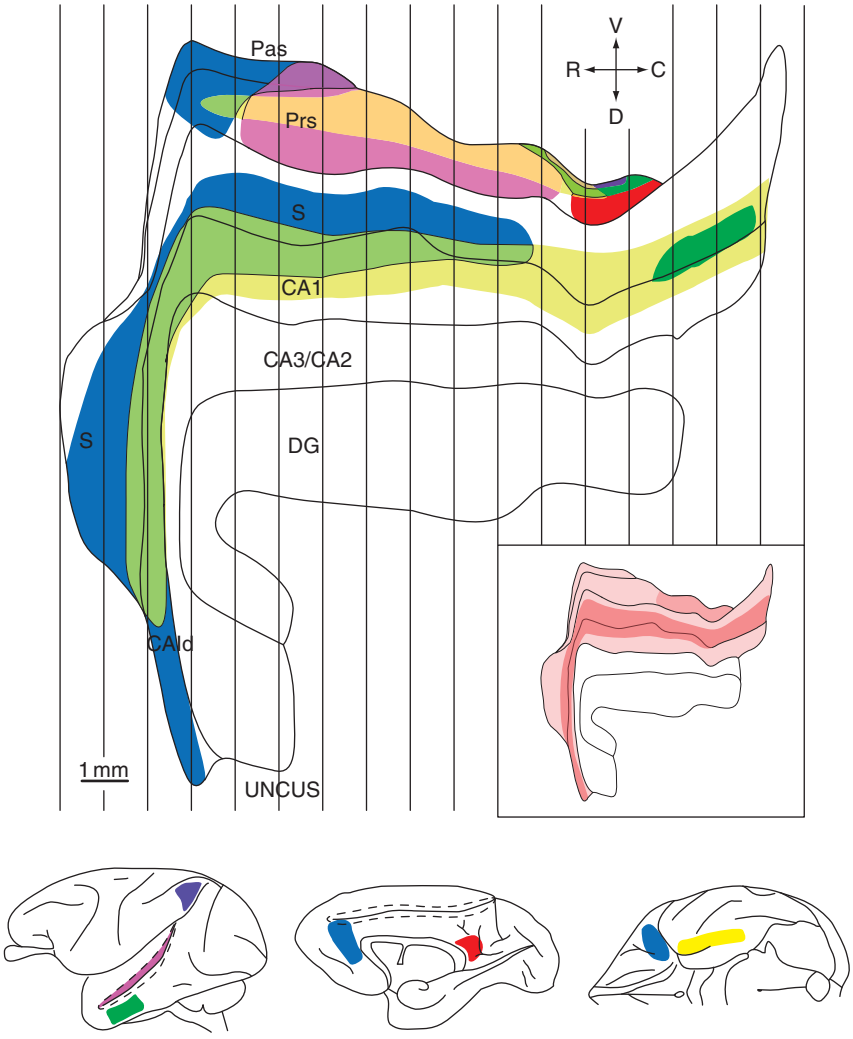

Figure 2 Direct hippocampal output to the cerebral cortex. The two-dimensional unfolded map of the hippocampus illustrates the topography of the origin of the projection to the areas of the cortex colored. A color code density map is shown in the inset to illustrate the subfields that carry the strongest projection (i.e., CA1/subiculum). Fields of the hippocampus: CA1; CA1d, dorsal field of CA1; CA2; and CA3. DG, dentate gyrus; Pas, parasubiculum; Prs, presubiculum; S, subiculum. Reproduced from Insausti R and Munoz M (2001) Cortical projections of the non-entorhinal hippocampal formation in the cynomolgus monkey (*Macaca fascicularis*). *European Journal of Neuroscience* 14(3): 435–451, with permission from Blackwell Publishing.

relevant for memory. In fact, the contribution of different portions of CA1, subiculum, presubiculum, and parasubiculum to this projection depends on their target area in the cerebral cortex. This connectional segregation likely reflects different information types processed and stored by the various cortical areas. Thus, a serial and parallel cascade arrangement applies both to the input to the hippocampus and to the outputs. One view is that the hippocampus always computes some common set of 'algorithms' that reflect features of its circuitry, and that it can do this to different kinds of information relayed from different cortical areas as a function of the species involved – mouse, monkey, or man. However, the local circuitry of hippocampus, including its myriad of distinct inhibitory interneurons, is not sufficiently well understood to be confident of this speculation.

Entorhinal projections to the cortex are reciprocal, organized topographically, originate primarily in its deep layers, and terminate in superficial and deep layers of the cortex – indicating they are of the feedback type. Like the entorhinal cortex, the perirhinal and posterior parahippocampal cortices (postrhinal in the rat) also send projections to the cerebral cortex. Their contribution is, in fact, quantitatively larger than that of the entorhinal cortex. There is little doubt that the feedback interactions from the medial temporal lobe to the cortex are critical for the formation, and possibly the strengthening, of distributed networks in the cerebral cortex, where consolidation of long-term declarative memory takes place. However, clinical reports and lesion studies in monkeys and rats have suggested that the cortical interaction of the hippocampus and that of the adjacent cortices seem to contribute differently to the two facets of declarative memory – that is, episodic and semantic memory. Connections formed by the hippocampus and subicular cortex on top of those formed by the entorhinal, perirhinal, and posterior parahippocampal cortices with the remaining cerebral cortex are thought to be critical for the storage of episodic information. Cortical connections of entorhinal, perirhinal, and posterior parahippocampal cortices are thought to be necessary for the storage of factual information (semantic memory).

The Contribution of the Thalamus and Frontal Lobe to Memory

Several anatomical pathways connect the hippocampal region with those areas in the thalamus that have been related to memory function – the magnocellular division of the mediodorsal, the anterior nuclei of the thalamus, and the mammillary bodies. There are other hippocampal–thalamic connections, such as with the nucleus reuniens, laterodorsal nucleus, and medial pulvinar, whose differential role in memory is unclear. These thalamic nuclei also connect with different areas of the frontal cortex and, altogether, temporal–thalamic–frontal connections are thought to participate in efficient encoding, ongoing memory processing, and recall of declarative information from long-term memory.

Based on clinical reports, anatomical studies, and experimental data from rat and monkey lesion studies, Aggleton and Brown proposed a model that divides the labor of the interactions between the medial temporal lobe and the thalamus and cerebral cortex in terms of episodic and semantic memory. Their model includes two parallel temporal–thalamic–frontal pathways that make qualitatively different contributions to learning and memory – one pathway critical for the 'recollective' aspects of recognition (episodic memory) and the other for item familiarity.

Specifically, their anatomical pathway critical for episodic memory includes projections from the subiculum, prosubiculum, and, to a lesser extent, the entorhinal cortex through the fornix to the medial mammillary nucleus. This circuit also includes the projections from the mammillary nucleus to the anterior thalamic nuclei and from there through the cingulum bundle back to the subicular cortex in the hippocampal region. Anterior thalamic nuclei connections with the medial and orbitofrontal cortices, which are the major recipient of hippocampal efferent connections from CA1/subiculum, also form part of this network critical for episodic memory.

The set of connections important for stimulus recognition comprises entorhinal, perirhinal, and posterior parahippocampal projections to the magnocellular division of the mediodorsal nucleus of the thalamus, which at least in the monkey are descendent, with no reciprocal connections. Like the anterior nuclei, the mediodorsal nucleus of the thalamus also has projections with medial and orbitofrontal cortices. These frontal areas are also a major target of entorhinal, temporal pole, and posterior parahippocampal cortices output.

However, there are a number of anatomical connections whose function remains unclear. Neuroanatomical tract tracing studies in monkeys have shown that thalamic input to the hippocampus is mainly directed to area CA1, whereas dentate gyrus and area CA3 do not receive direct thalamic connections. This input arises in nucleus reuniens and terminates with asymmetric synapses (i.e., indicative of excitatory input) in both CA1 principal and GABAergic interneurons. Polymophic cells in the deepest layer of the subicular and entorhinal cortices have reciprocal connections with the anterior thalamic nuclei and the laterodorsal nucleus of the thalamus.

It seems that the thalamic afferents to the hippocampus are reciprocated by descending corticothalamic projections that originate in the presubiculum and entorhinal, temporopolar, perirhinal, and parahippocampal cortices as well as from more distant regions, such as cingulate, retrosplenial, and ventral insula.

Comparative Aspects

The characteristic architectonic organization of the dentate gyrus and hippocampal region makes it easily recognizable in Nissl-stained sections in many species, from very early mammals to humans. Not all hippocampal fields and adjacent cortex, however, seem to have homogeneously expanded in primates compared to smaller mammals such as rats. Areas CA1, subiculum, and the entorhinal cortex appear disproportionately enlarged in the primate brain. Their increased size in primates parallels the expansion of association areas in the cerebral cortex, to which they are connected. As a consequence, rat and monkey hippocampo-cortical connections share the same basic organizational principle (i.e., convergence of inputs to the entorhinal cortex, unidirectional circuitry, etc.), but they also present some differences. The entorhinal cortex of the rat has connections with unimodal sensory areas and with the motor cortex in the frontal lobe. Whereas, with the exception of olfactory connections, minor connections with unimodal sensory areas and no connection with motor cortex have been reported to date in the monkey's entorhinal cortex. Its connections are primarily with polysensory areas. A particular case is the olfactory projection. Although both the rat and the monkey entorhinal cortex receive direct input from the olfactory bulb, this only reaches approximately 10% of the monkey entorhinal cortex, whereas it terminates in almost the entire extent of the rat entorhinal cortex.

In contrast to the expansion of hippocampal–cortical connections in the primate brain, hippocampal and dentate gyrus commissural connections seem to have recessed, perhaps reflecting a growing laterality of function that is especially clear in humans. In rats, extensive commissural connections originate in the polymorphic layer of the dentate gyrus to the contralateral molecular layer of the dentate gyrus and from the CA3 field of the hippocampus to contralateral CA3 and CA1 fields. In monkeys, only the most rostral portion of the dentate gyrus and hippocampus in monkeys seem to project to homotypical contralateral sites. Monkeys and rats, however, share a commissural projection that originates in the presubiculum and terminates in contralateral entorhinal layer III.

Comparative anatomy research is needed at different levels (i.e., physiological, cellular, neurochemical, etc.) and in different systems (i.e., hippocampal–thalamic–frontal systems) to determine additional commonalities and differences between species and how these map onto the development of episodic-like memory systems.

Neurophysiology

The neurophysiological correlates and possible mechanisms of episodic-like memory have been studied by means of (1) single-unit recording in several of the structures identified previously in behaving animals performing memory tasks; (2) studies of synaptic physiology both *in vivo* and *in vitro*, particularly synaptic potentiation, which is thought to be crucial for the rapid encoding of new information; and (3) experimental studies and computational models of rhythmical electrical patterns and their relationship to encoding and retrieval.

One major correlate includes the selective firing of neurons in the hippocampus when an animal occupies a particular place in the environment ('place cells') or is looking at a specific view ('view cells'). Neurons in the entorhinal cortex provide a metric for place specificity and information about the direction in which an animal is moving at the time that an event occurs ('grid cells'). These and other specialized single cells have all been identified and characterized in rodents and/or monkeys – and all would be useful components of a neural episodic-like memory system because they would provide information about the context or the scene and an animal's movement through it. Changes in context can also result in distinct forms of spatial remapping of an environment. Alterations occur in the relative rate at which place cells fire in some circumstances (rate remapping), whereas a radical reorganization of firing fields occurs under others (global remapping). What is not clear is how unique events and episodes are encoded with respect to these changes in the representation of contextual information. A difficulty in making single-unit observations of events is that obtaining reliable information about the firing correlates of a single neuron generally requires repetition of the sufficient stimulus for firing it – a condition that cannot be met for a unique event. Recording techniques that enable hundreds of cells to be monitored at one time have been developed and the use of these during episodic-like memory tasks may enable specific population vectors to be identified that correspond to a particular event. Once this breakthrough is achieved, it will be possible to ask analytic questions about the

association between events and the context or scene in which they occur.

Laboratory studies have also revealed that synaptic weight changes can be induced and made to last after very brief patterns of neural activation, as studied physiologically in the phenomenon of long-term potentiation (LTP). The most prominent form of LTP is known to require the momentary activation of *N*-methyl-D-aspartate (NMDA) receptors. LTP induction triggers changes in the availability of post-synaptic glutamate receptors responsible for fast synaptic transmission, perhaps by altering the rate at which α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors are trafficked to the synaptic membrane. In this way, the strength of connections between cells in a distributed associative memory of the kind that is processed in hippocampal circuitry can be altered in a sustained manner. There have been a number of demanding tests of the hypothesis that hippocampal synaptic plasticity is critical for memory. Studies of the role of LTP (and long-term depression) in learning and memory have used a variety of tasks of which some are good models of episodic-like memory. New studies revealing that LTP is selectively induced at hippocampal synapses during rapid learning and is necessary for learning to take place have complemented research showing how artificial alterations in synaptic weights after training alter memory recall.

If this way of thinking about the encoding of episodic-like memory is correct, the resulting changes in synaptic strength would occur in circuitry, such as in area CA1 of the hippocampal formation, where there is the potential for linking information about events (processed in hippocampal circuits) to the spatial context in which an event has occurred (projected in from the layer III cells of the entorhinal cortex). The value of this is that approaching such a location again in the future will retrieve a memory representation of the prior event that happened there (e.g., the prior presence of predator). It is less obvious that remembering 'when' an event happened is quite so valuable. The local circuitry of parts of the hippocampal formation (e.g., the dentate gyrus) is also ideally suited for separating or 'orthogonalizing' neural patterns such that the synaptic memory traces of very similar events can be stored in overlapping but sufficiently separated locations in downstream regions in order to minimize interference at the point of recall. Other parts of the circuitry (e.g., area CA3) may be useful for encoding sequences of events in the correct order and neural network models of such a process have been proposed. However, it is still unclear how changes of this kind within hippocampal circuitry give rise to lasting changes in connectivity in the cortex in the process of memory consolidation. The feedback connections described previously are likely to be critical for this.

Episodic-like memory involves both encoding and recall. Different computational models have each implicated the hippocampal theta rhythm in the switchover between these two modes. Theta is a rhythmical pattern of electroencephalogram (EEG) activity in the hippocampal formation, proceeding at between 5 and 10 Hz in rodents. It is contrasted a state in which the EEG shows large-amplitude irregular activity. Some suggest that the theta mode is a learning mode, at least in the rat, with irregular activity occurring during memory retrieval. Other models suggest that the switchover between encoding and retrieval happens extremely rapidly at different phases of the theta cycle, with the significance of the cycle being to modulate the probability that NMDA receptors can be successfully activated as a function of membrane potential. Further work linking these neurophysiological parameters to behavioral paradigms using valid episodic-like learning paradigms is required.

Behavior and Cognition

The neurobiology of learning and memory has been studied using a wide range of behavioral tasks in animals. Some are classified as 'declarative,' others as more 'procedural.' It is important to appreciate, however, that of those that are accepted as models of declarative memory, making the distinction between whether a specific task is 'episodic-like' and 'semantic-like' is often unclear. Certain widely used tasks, such as tests of recognition memory (delayed non-matching to sample in monkeys), paired-associate memory (e.g., object–place tasks in monkeys), spatial memory (the water maze in rats), and food caching (in birds), do not fall neatly into one or the other category. Moreover, the notion that one form of memory (e.g., spatial memory) is mediated solely by one brain area (e.g., the hippocampus) has fallen out of favor as neuroscientists recognize the complexity of different interacting memory processes (e.g., encoding vs. consolidation). However, a subset of behavioral tasks has attempted to determine the key distinction between episodic and semantic memory. We focus on these cases in this section.

Remembering 'What' and 'Where'

A common experience when remembering a specific event is that we also remember where that event happened. Can animals also do 'mental space travel'? Experiments in monkeys suggest they are very sensitive to the scene or context of events. In one study,

macaque monkeys were trained to reach for either of two objects (A or B), such that A was correct when the animals were facing one way in the laboratory room but B was correct when they were facing in the other direction. This task was relatively easy to learn and was impaired by bilateral damage to the fornix. Separately, the monkeys were also trained on a logically identical task in which the backdrop was the color of a tray rather than the scene in front of the animal. This task was unaffected by the lesion. When the monkey has to face in different directions, the scene before the monkey changes even though many elements of it are still there (e.g., the window). Neurons have also been identified in the hippocampus of monkeys whose firing patterns are exquisitely sensitive to what the animal is looking at (view cells), suggesting that scene processing is very important in the rapid learning about events.

Scenes can also be displayed on computer screens. In an object-in-place task, for example, the animal is trained to reach out and touch the 'object' (e.g., a typographical character) that occupies a particular position in a unique background of abstract colors and shapes on the screen (Figure 3(a)). Monkeys readily learn to do this and their success relies on the integrity of a circuit that includes the hippocampus, fornix, mammillary bodies, and anterior thalamus. The impairment is no greater when two of these structures are damaged sequentially, suggesting that these brain regions make their contribution to information processing in a serial manner – as if part of a circuit. Additional studies, using traditional training techniques that require monkeys to reach for objects on trays placed in front of them, further establish the importance of the entorhinal and perirhinal cortex in processing associations between 'where' and 'what.'

Rats can also learn event–location associations rapidly and can recall the location at which an event occurred when given an appropriate cue. For example, when trained to find and dig up distinctively

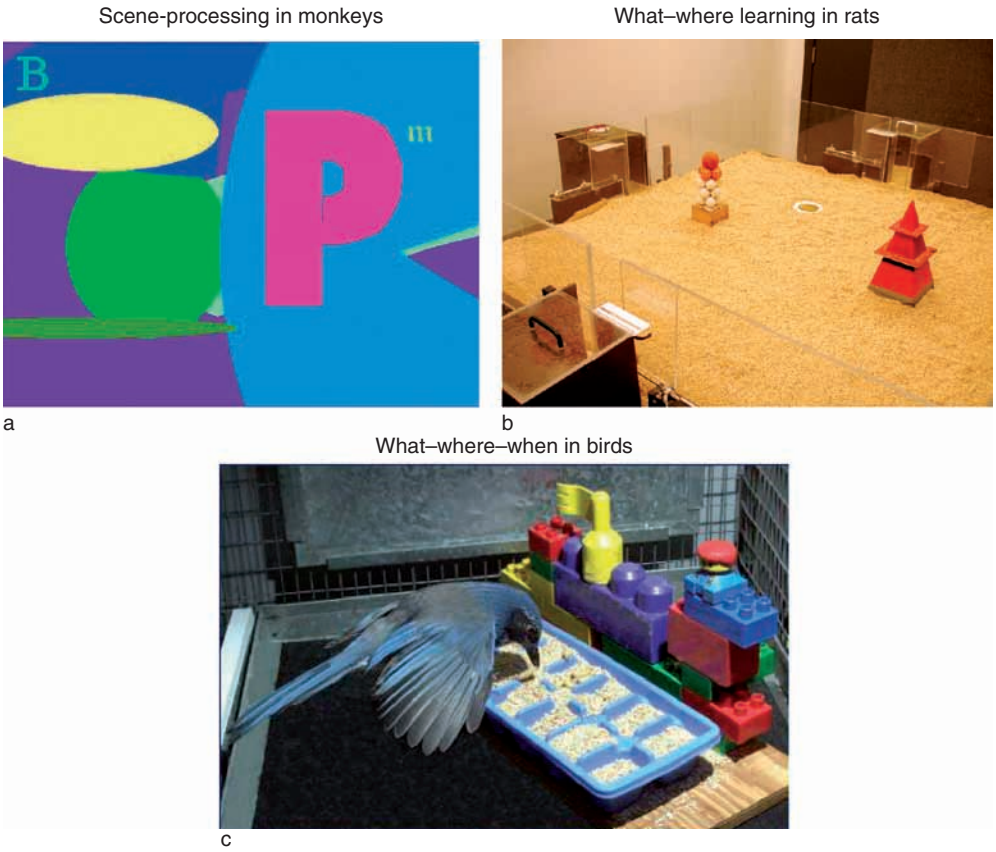

Figure 3 Contextually specific memory. (a, b) Memory for what and where. Monkeys can be trained to use 'touchscreens' and choose differentially between objects (typographical characters) as a function of the background scene against which they are presented. Rats can also be trained to find flavored food in sand wells at different locations in a large training arena. A short-lasting memory of a what–where association can be established in one trial. (c) Memory for what, where, and when in California scrub jays. A supply of different foods is made available to the food-caching bird at different times. This is cached in an ice cube tray filled with sand. The bird makes an appropriate free-recall judgment of what to search for where at different times after the caching episode. Courtesy of David Gaffan, Richard Morris, and Nicola Clayton.

flavored food at different locations in a training arena, rats can – after only a single training trial with each flavor – recall which location to visit to get more food when cued with a specific flavor (Figure 3(b)). Such memories do not last very long unless they are subject to repeated training, but this gradually increasing memory persistence may reflect a transition from episodic-like to semantic-like memory. Learning depends on the hippocampus and is impaired by a selective NMDA receptor antagonist, D-AP5, when given prior to information encoding. If only given prior to memory recall, the drug has no effect. This is directly analogous to its effect on LTP discussed previously: D-AP5 blocks the induction but not the expression of LTP.

Food Caching by Birds as a Model of 'What,' 'Where,' and 'When'

The temporal element of the 'what, where, when' trilogy is not captured in these tasks and is clearly critical. Work on birds, specifically Californian scrub jays, has started to address this critical element of episodic-like memory. Adult hand-raised scrub jays were allowed to cache food items, such as mealworms or peanuts, in sand-filled ice cube trays surrounded by distinctive visual objects (Figure 3(c)). The birds much prefer fresh mealworms to peanuts, but these perish over time and become distasteful, whereas peanuts last much longer. Two caching periods, up to 5 days apart, constituted the opportunity for the animals to encode into memory what food item was being cached, where (in the ice cube tray), and when (recently or some while ago). Recovery was permitted either immediately or several days later, with both sides of the ice cube tray uncovered and available. The main finding was that in nonrewarded probe tests in which neither food was actually available (thus preventing any olfactory or visual cues to guide search location), the birds searched preferentially for wax moths when recovery occurred 4 h after caching but switched their search to the peanut locations when recovery was delayed for 124 h. This switchover was not shown by birds in a control group in which, throughout training, fresh mealworms were placed in the correct places even after long caching–recovery delays. Thus, at recovery and using free recall, the birds could recollect what food item was stored where and, at least in some sense, how long ago.

In addition to careful control procedures, the different foods caching paradigm is a helpful innovation because it encourages the animal to be attentive and engaged when encoding the memory of where a specific type of food is located. The use of very long memory intervals (days) is beyond the capability of short-term memory, and the use of free recall at cache recovery obviates the use of a familiarity strategy that has been so problematic in studies of recognition memory. Clayton and Dickinson's claim that their initial study provides "the first conclusive behavioral evidence of 'episodic-like' memory in animals other than humans" has been vindicated by later studies in which a number of detailed issues have been taken further, such as applying information retrospectively at the time of retrieval and investigating prospective aspects of food caching. The birds have even been observed to do certain things at the time of caching that reflect some kind of 'awareness' of what food items they may prefer to secure at a point in the future. This ability would be outside the domain of semantic memory.

Similar behavioral studies have been conducted using monkeys living freely in a primate colony and these have provided evidence for the memory of 'what, where, whom.' However, neither the scrub jay nor the primate work has involved invasive neurobiological studies and so the anatomical and physiological basis of performance is unknown. This is unfortunate because, as discussed previously, including a temporal element in encoding and recall is desirable for distinguishing episodic-like memory from rapid semantic learning.

Sequences of Events

It was previously mentioned that area CA3 of the hippocampus has recurrent auto-associative circuitry. Neural network models have suggested that this may be useful for learning about sequences of events that would then collectively comprise an episode. Single-unit recording studies, such as studies of 'replay' following sleep episodes, offer some support for this idea, as do observations that the firing of place cells in a maze may be 'gated' in a retrospective or prospective way by preceding or anticipated events. Lesion studies indicate that damage to the hippocampus disrupts sequence memory. For example, a task involving successive digging through cups of sand with distinctive odors indicated that normal rats could remember trial-unique sequences of odor A before odor B, B before C, C before D, and so on, whereas the rats with hippocampal lesions could not. However, the lesioned rats were just as good as controls in remembering that these odors were familiar, although presented for only a single trial, when they were tested against completely novel odors. A follow-up study indicated that rats with hippocampal lesions could eventually acquire some sequential information

about repeated sequences with extended training but never well enough to disambiguate overlapping sequences. One theoretical view of this component of episodic-like memory in rats is that it can encode information about sequence but not necessarily the absolute time when an event occurs.

Current Dichotomies in the Neuroscience of Episodic Memory: Memory versus Perception, Remembering versus Knowing, Recollection versus Familiarity

As we have seen, early dissociations between memory systems emerged from studies of patients with large medial temporal lobe (MTL) damage. These reports were critical to the identification of different memory systems, such as declarative and non-declarative memory, and they stimulated research on the impact of restricted lesions in both animal models and in patients with respect to what areas within the MTL are critical for memory. This research has shed light on what medial temporal areas may be critical for certain processes, that is, the perirhinal cortex for recognition of objects, the posterior parahippocampal region for recognition of object-place associations, and the hippocampus for spatial memory. However, new questions have emerged. At present various dichotomies such as those listed above are a current focus of attention. So far, these dichotomies are proving difficult to relate precisely to specific brain structures, but the use of different behavioral methodologies, such as the analysis of performance using signal detection theory and, specifically, receiver operating characteristics (ROCs) are helping us to understand their validity and anatomical relevance. The use of ROCs has been championed by Yonelinas, and for animal studies by Eichenbaum, but the selectivity of this approach has also been the subject of criticism by Squire. Physiological techniques, such as event related potentials (ERP), functional MRI (fMRI), combined fMRI and ERP in patients with restricted lesions are collectively helping to shed light on theoretical controversies.

Conclusion and Outstanding Issues

Episodic memory can be defined in a way that essentially guarantees that it is unique to humans. However, a different and comparative hypothesis is both realistic and one supported by a body of data. First, the striking neuroanatomy of declarative memory in mammals points to intrinsic circuitry and patterns of connections that are important for the rapid encoding and recall of information from numerous sensory modalities. Second, context processing by neurons and rapid synaptic potentiation are essential neurophysiological features that a neural episodic-like system requires. Third, a range of new behavioral tasks is being developed to elucidate the subtle distinctions, within the declarative domain, of episodic-like and semantic-like memory.

A paradoxical feature of the 'uniqueness' claim that animals lack episodic memory and possess only semantic-like memory is that, in humans, the latter is most often acquired through communication – the very channel of learning that is not open to animals. If a paradigmatic example of semantic memory is knowing that 'Paris is the capital of France' despite the fact that the person knowing this need never have visited the city, then this is precisely the kind of information that animals could never acquire because their learning absolutely requires personal experience. Their personal experience of learning will also entail the circadian rhythm of night and day and, in most cases, of the different seasons. It is then reasonable to suppose that animals could perhaps acquire the knowledge that things change over time and that different frameworks of knowledge would be appropriate to different times of day or the year (e.g., in relation to safety by day or night or the availability of water). Yet the radical claim of those espousing the 'human uniqueness' view of episodic memory is that animals can know all these things and yet be 'stuck in time.' That is, they suppose that all animals have neither a sense of the past nor a sense of the future, despite possessing both declarative and procedural learning systems that enable their future behavior to profit from past experience. The claim may be right, but it is deeply paradoxical.

Based on the kinds of data presented here – anatomical, physiological, and behavioral – the idea that some animals may possess episodic-like memory systems seems strong. However, certain qualifications should be recognized. First, if true episodic memory is defined as consisting of certain mental experiences, such as reexperiencing the past, we may always be unable to assess whether animals also have such experiences and thus the claim is untestable. Suddendorf and Corballis have supported Tulving's 'uniqueness' hypothesis in suggesting that "the ability to travel mentally in time constitutes a discontinuity between humans and other animals." However, they also recognize that finding evidence for mental time travel in animals would be a major discovery. A second qualification is that, because episodic memory in humans encodes personal experiences, a sophisticated form of awareness that includes a sense of one's own individual identity may indeed be essential

for the full expression of true episodic memory. What is unclear is how to realize a scientifically acceptable measure of such a subjective concept as a 'sense of one's own individual identity.' The gap between the 'uniqueness' and 'comparative' hypotheses of episodic and episodic-like memory cannot be closed until additional behavioral techniques are developed and explored. This arsenal of tasks will then have to be examined in relation to the complex patterns of anatomical connections of the medial temporal lobe described previously. The imagination of researchers in behavioral neuroscience is such that techniques for addressing these issues may become available.

See also: Declarative Memory System: Anatomy; Episodic Memory; Event-Related Potentials (ERPs); Executive Function and Higher-Order Cognition: Assessment in Animals; Hippocampus and Neural Representations; Memory: Computational Models; Recognition Memory; Spatial Memory: Assessment in Animals; Synaptic Plasticity and Place Cell Formation.

Further Reading

Aggleton JP and Brown MW (1999) Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and Brain Sciences* 22: 425–489.

Amaral DG and Insausti R (2004) Hippocampal formation. In: Paxinos G and Mai JK (eds.) *The Human Nervous System*, 2nd edn. New York: Academic Press.

Fortin NJ, Wright SP, and Eichenbaum H (2004) Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature* 431: 188–191.

Gloor P (1997) *The Temporal Lobe and Limbic System*. New York: Oxford University Press.

Griffiths D, Dickinson A, and Clayton NS (1999) Episodic memory: What can animals remember about their past? *Trends in Cognitive Sciences* 3: 74–80.

Insausti R and Munoz M (2001) Cortical projections of the non-entorhinal hippocampal formation in the cynomolgus monkey (*Macaca fascicularis*). *European Journal of Neuroscience* 14(3): 435–451.

Miyashita Y (2004) Cognitive memory: Cellular and network machineries and their top-down control. *Science* 306: 435–440.

Morris RGM (2006) Elements of a neurobiological theory of the hippocampus: The role of activity dependent synaptic plasticity in episodic-like memory. *European Journal of Neuroscience* 23: 2829–2846.

Schwartz BL and Evans S (2001) Episodic memory in primates. *American Journal of Primatology* 55: 71–85.

Suddendorf T and Corballis MC (1997) Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs* 123: 133–167.

Tulving E (1983) *Elements of Episodic Memory*. New York: Oxford University Press.

Tulving E (2004) Episodic memory and autonoesis: Uniquely human? In: Terrace H and Metcalfe J (eds.) *The Missing Link in Cognition: Evolution of Self-Knowing Consciousness*. New York: Oxford University Press.

Wais PE, Wixted JT, Hopkins RO, and Squire LR (2006) The hippocampus supports both the recollection and familiarity components of recognition memory. *Neuron* 49: 459–466.

Yonelinas AP (2001) Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of Royal Society of London B: Biological Sciences* 356: 1363–1374.

Erectile Dysfunction

S H Tariq and J E Morley, St. Louis University and Veterans Affairs Medical Center, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Erectile dysfunction (ED) is defined as an inability to achieve or maintain an erection sufficient for satisfactory sexual performance. ED is estimated to occur in 10–20 million men, which increases to 30 million if partial ED is included. ED can occur in men of all ages, but it is more common in the elderly and is not a part of normal aging.

Prevalence

The prevalence of ED varies. The reported prevalence of ED is 5% in 40-year-olds, which triples to 15% at the age of 70. The occurrence of ED increases with age: it occurs in 27% of men at age 70 years, 55% at age 75 years, and 75% at age 80 years. The prevalence of ED is 50% in patients with renal disease and it increases to 75% in those who are on dialysis. ED develops in approximately 50% of patients with diabetes within 10 years of diagnosis.

ED is a major health concern, strongly associated with increasing age. Men with ED have impaired quality of life compared with unaffected individuals. Since sexual activity contributes significantly to quality of life, patients expect physicians to address such an important issue.

Physiology of Penile Erection and Changes with Aging

Sexual activity consists of libido and potency. Libido is the sexual desire/drive, thoughts and fantasies, satisfaction, and pleasure. Potency is the ability to attain and maintain an erection and to ejaculate. There are three pathways for penile erection, which are summarized in Table 1. Masters and Johnson described changes in the four stages of the sexual response cycle – excitement, plateau, orgasm, and resolution – with aging and these are outlined in Table 2.

Causes of ED

The etiology of ED varies significantly depending on the clinic population, as shown in Table 3. Etiologic causes for ED include vascular, endocrine (hypogonadism, thyroid problems, diabetes, and hyperprolactinemia), neurologic, urologic, psychogenic, medication, and alcoholism. Depression as a cause of ED is more common in those younger than 60 years of age. Similarly, anxiety and marital discord are more common in the population group younger than 60 years of age.

Medical Conditions Associated with ED

The most common medical conditions associated with ED are hypertension and coronary artery disease. Fifty-seven percent of patients who underwent coronary artery surgery had ED. Other conditions associated with ED include diabetes, smoking, and alcohol use. An increase in total cholesterol has been associated with ED.

Evaluation

A detailed history is absolutely essential to differentiate difficulties with ED from a decrease in libido. Table 4 highlights some of the important areas that are essential in evaluating ED. A history from a partner will help uncover some of the problems with their sex life. The risk factors for ED during history include vascular disease, hormonal problems (thyroid and diabetes), neurological dysfunction, medication, past surgery, or trauma; psychological and social factors, including an evaluation for depression, should be assessed. A number of medications are associated with an increase or decrease in libido, and these are shown in Table 5. Since depression can contribute to ED, a screening assessment tool such as the Beck Depression Inventory or Yesavage Geriatric Scale may be used. A number of screening tools are also available to screen patients for androgen deficiency, such as the Saint Louis University Androgen Deficiency in Aging Males (ADAM), which is shown in Table 6. A positive answer to question 1 or 7 or any other three questions is considered to indicate a high likelihood of low testosterone. The ADAM questionnaire has a sensitivity of 88% and a specificity of 60%. There is a clear improvement in the ADAM questionnaire after treatment with testosterone.

Physical examination should include assessment of secondary sexual characteristics along with a detailed examination searching for vascular and neurological problems. Local penile shaft examination is also important to rule out Peyronie's disease.

Laboratory investigations such as diabetes, renal functions, lipid profile, thyroid function tests, prolactin levels, and penile brachial pressure index will help search for the etiology of ED. The most important test for determining a decrease in sex drive or libido is to check total testosterone and bioavailable testosterone

Table 1 Physiology of penile erection

- Stimulation of the genitalia, in which sensation in the sensory receptors of the penile shaft and glans penis is carried to the spinal cord by afferent nerves. The sensory perception is then passed on to the higher centers, and efferent impulses leave the spinal cord via the sacral parasympathetic center to the pelvic plexus and the cavernous nerves into the corpus cavernosum.
- Psychogenic stimulation that occurs through visual or auditory stimuli. This is more complex and less understood.
- Penile erection occurs during the rapid eye movement stage of sleep. The physiology is not known.
- During erection the parasympathetic nervous system causes vasodilatation, increased blood supply and pressure to the cavernous spaces, and venous occlusion.
- Flaccid state is achieved by the release of catecholamines, which causes contraction of the smooth muscles around the sinusoids and arterioles, expelling the blood out of the sinusoidal spaces and decreasing blood flow, resulting in a flaccid state.
- Neurotransmitters such as nitric oxide, prostaglandins E_2 and E_1, vasoactive intestinal peptide, neuropeptide Y, and acetylcholine cause smooth muscle relaxation and penile erection. Smooth muscle contraction is caused by prostaglandin $F_{2\alpha}$, substance P, histamine, and norepinephrine.

Table 2 Changes with aging

- The excitement phase shows a delay in erection, decrease in erectile strength as well as decrease in scrotal congestion, and testicular elevation with advanced age.
- The plateau phase in prolonged along with a decrease in preejaculatory. The orgasmic phase is shorter, with a decrease in the ejaculatory force. The time for ejaculation is longer compared to that of young adults.
- In the resolution phase, there is rapid detumescence and rapid testicular descent.
- The refractory period is also prolonged with age.
- The penis becomes less sensitive to stimulation with age and it takes longer to achieve erection. The erection becomes more dependent on physical stimulation of the penis and less responsive to visual, nongenital, and psychologic stimulation.

Table 3 Causes of erectile dysfunction

	Type of clinic								
	Endocrine	*Urology*	*Sex therapist*	*Family practice*	*Medical outpatient*	*Outpatient clinic*	*Urology*[a]	*Outpatient*	*Geriatric clinic*
No. of patients	105	165	101	154	120	93	31	121	297
Mean age (years)						61	53	68	66
Etiology (%)									
Endocrine	44	6	45	24	29	35			54
Hypogonadism	26	6	39	23	19	9.6	19	2.6	43–81
Thyroid	3		4		4	2			
Hyperprolactinoma	8		2		6	4			0.3
Diabetes mellitus	7	19	2	20	9	22			27–29
Vascular		8	1	37		37.6	68	21	39–55
Neurogenic		4	2	5	7	17	26	17[b]/11[c]	3–10
Surgical/trauma		8							
Urologic		8		49	6	19	3	1.3	7–12
Alcoholism	3	0.6		7	7	20			12–14
Systemic disease	6		2	5	5				
Medication	7	2	0	12	25	67		4	58
Psychiatric	14								5–18
Psychogenic	14	51	19	60	14	15	19	9.2	3–9
Unknown	19	0.6		0.7			4		

[a]All diabetic male patients.
[b]Diabetic neuropathy.
[c]Nondiabetic neuropathy.

in older adults (non-sex hormone-bound testosterone). Cross-sectional studies have demonstrated that levels of testosterone decline with age, although the Baltimore Longitudinal Aging Study failed to find a decrease in testosterone levels in males. The decline in testosterone is 100 ng/dl per decade. A number of studies have reported a decline in bioavailable testosterone with aging. A decline in bioavailable testosterone is

Table 4 Evaluation of erectile dysfunction

History
Review of medications
Screening for depression
Screening for hypothyroidism
Screening questionnaires for ED/hypogonadism
Physical examination
Laboratory tests
Renal function
Liver function tests
Complete blood count
Prostatic serum antigen
Total or free or bioavailable testosterone

Table 5 Drugs that possibly affect sexual desire in men

Drugs that possibly decrease sexual desire
Antihistamine/barbiturates/benzodiazepines (sedative effect)
Tricyclic antidepressant
Diuretics
Spironolactone/hydrochlorothiazide/acetazolamide/methazolamide
Antihypertension
Reserpine/propranolol/clonidine/alpha-methyldopa/prazosin
Cimetidine
Hormones
Estrogen/medroxyprogesterone
Clofibrate
Digoxin
Lithium
Drugs that possibly increase sexual desire
Androgens (in androgen-deficit states)
Baclofen (possible anti-anxiety effect)
Benzodiazepines (anti-anxiety effect)
Haloperidol (indirect effect due to improved sense of well-being)

Table 6 St. Louis University ADAM Questionnaire

1. Do you have a decrease in libido (sex drive)?
2. Do you have a lack of energy?
3. Do you have a decrease in strength and/or endurance?
4. Have you lost height?
5. Have you noticed a decreased 'enjoyment of life'?
6. Are you sad and/or grumpy?
7. Are you erections less strong?
8. Have you noted a recent deterioration in your ability to play sports?
9. Are you falling asleep after dinner?
10. Has there been a recent deterioration in your work performance?

seen in 50% of men by 50 years of age. Luteinizing hormone levels fail to increase appropriately in response to a decrease in testosterone levels with age, suggesting failure of the hypothalamic–pituitary unit (i.e., secondary hypogonadism).

Treatment

Current treatments for ED include oral medications, intracavernosal injections, testosterone therapy, vacuum pumps, penile prostheses, and vascular surgery. First-line therapy includes oral drug therapy, the use of vacuum devices, and, less frequently, sex therapy. Referral to a specialist may be required upon request of the patient if initial therapy failed or a more comprehensive evaluation is needed. Patients with a suspected psychogenic etiology should be considered for sexual counseling or psychiatric referral. Involvement of the partner in the diagnostic and therapeutic plans from the beginning is crucial. If a particular drug is related to ED, then that drug should be stopped and a different class of drug should be used.

Oral Therapy

Phosphodiesterase-5 inhibitors

Currently, there are three drugs available in the United States and all inhibit phosphodiesterase-5 (PDE-5), which is responsible for the degradation of cyclic guanosine monophosphate (cGMP) in the corpus cavernosum. When the nitric oxide and cGMP pathway is activated during sexual stimulation, inhibition of cGMP by PDE-5 inhibitors results in increased levels of cGMP in the corpus cavernosum and increased relaxation of corpus cavernosal smooth muscle cells in response to sexual stimulation. This causes an increase in penile blood flow and erection. Thus, sexual stimulation is required and PDE-5 inhibitors help maintain it.

Sildenafil is well absorbed during fasting, and the plasma concentrations are maximal within 30–120 min. It is mainly excreted by the liver, and its half-life is approximately 4 h. The starting dose is 50 mg taken 1 h before sexual activity, not more than once per day. The quality of life consequences of ED include depression, anxiety, and loss of self-esteem. One study determined correlations between erectile function (EF), intercourse success, and emotional well-being measured with the Self-Esteem and Relationship (SEAR) questionnaire in men treated with sildenafil citrate for ED and stratified by age (<50 years, 50–65 years, and >65 years). For the overall population, there was significant improvement ($p < 0.0001$) in the Self-Esteem subscale and intercourse success rate. Secondary outcomes (i.e., EF domain of the International Index of Erectile Function and event log frequency of erection hard enough for sexual intercourse and of ejaculation/orgasm) also improved. All ten correlations were positive in men of all ages. The most common adverse

events were mild to moderate headache (12%), vasodilatation (7%), and rhinitis (4%). Sildenafil improved EF and there was an increased intercourse success rate, which correlated positively with improvement in SEAR measures of self-esteem and sexual relationship.

In hypogonadal patients with ED, androgen supplementation is first-line therapy. If there is no response to androgen alone or sildenafil alone, combined use may improve EF and enhance the therapeutic effect of PDE-5 inhibitors.

Tadalafil has a half-life of 17.5 h. At a dose of 20 mg, tadalafil enabled 73–80% of sexual intercourse attempts to be completed successfully with appropriate sexual stimulation between 30 min and 36 h after dosing. Tadalafil is well tolerated. Side effects are headache and dyspepsia, which are mostly mild or moderate and transient. There were no significant laboratory abnormalities or electrocardiogram changes. It is preferred over sildenafil by men in the treatment of ED, secondary to its longer duration of action, especially by those who do not have any risk factors for developing chest pain. In a small number of patients refractory to sildenafil therapy, a trial of vardenafil might be helpful.

Vardenafil is well tolerated. Side effects are headache, cutaneous flushing, dyspepsia, and rhinitis. Vardenafil successfully treated ED after nerve sparing radical retropubic prostatectomy.

All PDE-5 inhibitors are absolutely contraindicated with the use of nitrates and are hazardous in patients with positive stress test not on nitrites; patients on a complicated, multidrug, antihypertensive program; patients with congestive heart failure and borderline low-volume status; and patients on drugs which prolong the half-life of these medications.

Phentolamine

Phentolamine is an α-adrenergic blocker with poor efficacy in improving ED but may be used with other agents. A prospective, placebo-controlled, double-blind study (40 patients) showed that phentolamine had a success rate of 20.0%, 30.0%, and 36.7% at doses of 20, 40, 60 mg, respectively, in the treatment of ED compared to placebo. The mean age was 48 years for those who responded to phentolamine compared to 52.2 years for the nonresponders. An open-label, multicenter study compared the efficacy and safety of oral sildenafil and phentolamine in men with ED. Patients received sildenafil (25–100 mg, $n = 123$) or phentolamine (40 mg, $n = 119$) for 8 weeks, and efficacy was assessed. Approximately twice as many men receiving sildenafil had successful attempts at sexual intercourse, improved erections, and improved ability to have sexual intercourse compared with men receiving phentolamine. The adverse events included rhinitis, headache, tachycardia, and nausea, with a higher frequency reported in patients receiving phentolamine than sildenafil (41% vs. 33%), with the exception of headache, which was reported more frequently in sildenafil users. Overall, sildenafil was more effective and appeared to be better tolerated than phentolamine for the treatment of ED.

Yohimbine

Yohimbine is an α_2-adrenergic receptor antagonist; it acts at the serotonergic and adrenergic receptors in brain centers associated with libido and penile erection. Its exact mechanism of action is unknown. Yohimbine has been shown to be relatively effective in psychogenic ED, although other studies have shown no statistical significant difference from placebo. Yohimbine carries a risk of hypertension, palpitation, fine tremor, and anxiety. Yohimbine has had questionable effects in men with organic ED. Of a total of 18 nonsmoking men, 50% were successful in completing intercourse in more than 75% of attempts. Yohimbine could be effective therapy to treat organic ED in some men.

Apomorphine

Apomorphine is a sublingual formulation and is now undergoing clinical trials. It acts centrally on dopamine D_2 receptors of the hypothalamus and activates pathways that involve nitric oxide and oxytocin signals. It is generally well tolerated at doses of 2 or 3 mg, and it is not an opioid. An observational cohort study was conducted in England between October 2001 and December 2002. The study cohort comprised 11 185 patients, 99.3% of whom were men, with a median age of 61 years. The study concluded that the most frequently reported prescribing indication was ED and the most frequently reported reason for stopping apomorphine was that it was "not effective." However, in a small percentage of patients apomorphine has been effective. Headache was the most commonly reported adverse drug reaction and the most frequently reported clinical condition occurring in the first month.

Medicated Urethral System for Erection

The medicated urethral system for erection (MUSE) uses prostaglandin E_1, which is an endogenous unsaturated 20-carbon fatty acid derivative of arachidonic acid, and alprostadil is a more stable, synthetic form of prostaglandin E_1. Prostaglandin E_1 causes smooth muscle relaxation and vasodilatation by acting on adenylate cyclase to increase the intracellular cyclic

adenosine monophosphate concentration. The first application is usually at a dose of 50 μg in the physician's office because of the potential complications. The dose can be titrated (250–1000 μg), depending on the erectile response. Alprostadil is well tolerated and effectively restores the capacity for erections and sexual intercourse. A restrictive band can be placed at the base of the penis to increase efficacy, especially if the phenomena of pelvic steal syndrome is present. The advantages of transurethral therapy include local application, minimal systemic effects, and the rarity of drug interactions. Side effects include penile pain, local discomfort, hypotension, dizziness, and syncope. Trauma, if it occurs, consists of a superficial urethral abrasion, no urethral stricture, penile fibrosis, or priapism.

Intracavernous Injection Therapy

The most commonly used intracavernous drugs in the United States are alprostadil and a combination of papaverine, phentolamine, and alprostadil. Intracavernous injection of alprostadil is an effective therapy with tolerable side effects. It has a success rate of 60–70% in the treatment of vasculogenic and 100% in neurogenic ED. It is contraindicated in sickle cell anemia, schizophrenia or other psychiatric disease, or severe venous leakage. Pain is reported to occur in 13–33%, fibrotic complication in 1–57%, and priapism in 5–23% of patients. If priapism does occur, the patient should be informed to take oral benadryl and, if it is not effective, should then seek immediate access for medical treatment. Injection therapy should be avoided in patients who are on chronic warfarin therapy because it increases the risk of hematomas and bleeding.

Alprostadil is the only agent approved in the United States. It is given in doses of 5–20 μg. It is indicated if oral therapy fails or is not suitable; for diabetics, especially those on insulin; for spinal cord injury; and for radical prostatectomy.

Papaverine is a nonspecific PDE inhibitor. The usual dose is 15–60 mg. It is highly effective (up to 80%) in men with psychogenic and neurogenic ED, but it is less effective when ED is of vascular etiology (36–50%). Advantages include its low cost and stability at room temperature. Its major drawbacks are priapism, corporal fibrosis, and occasional increases in serum aminotransferase concentrations.

Testosterone Therapies

Testosterone replacement can be achieved in several ways: oral and injectable therapies, transscrotal and transdermal patches, subcutaneous gels, buccal preparations, and selective androgen receptor molecule (no effects on prostate). Some preparations are under development, such as sublingual and inhalation forms.

Effects of Testosterone Replacement

Coronary artery disease Studies have shown that low testosterone levels are associated with both coronary artery disease and the degree of atherosclerosis. Testosterone dilates the brachial artery and therefore increases blood flow, decreases ST depression during exercise stress test, and reduces or has no effect on myocardial infarction. Testosterone replacement therapies have not demonstrated an increased incidence of cardiovascular disease or events.

Lipid profiles Testosterone decreases cholesterol and low-density lipoprotein cholesterol levels and high-density lipoprotein (HDL) cholesterol. A meta-analysis of testosterone replacement showed that HDL levels were reduced in three studies and unchanged in 15 studies. Five studies showed a reduction in total cholesterol, two showed an increase, and 12 were unchanged. Low-density cholesterol was either unchanged or reduced in 14 of the 15 studies.

Libido and ED Epidemiologic studies have shown that decreased sexual activity and libido are associated with a decrease in free or bioavailable testosterone levels. These studies suggested that aging was a more important contributor than androgen levels in sexual behavior in healthy older men. Testosterone replacement improves libido and quality of erection in hypogonadal men. Sildenafil is reported to improve erections in older men. It is reported that certain men with low testosterone do not respond to sildenafil alone unless their testosterone is replaced. This seems to be because testosterone is essential for the synthesis of nitric oxide synthase.

Body composition and frailty Decrease in muscle mass and strength is associated with aging. Resistance training restores the muscle mass and strength. Excessive loss of muscle in a person results in sarcopenia and increases the risk of becoming frail. Studies on testosterone replacement have clearly shown that it increases the muscle mass, and this is reported even in older men who are not hypogonadal. The increased strength is reported in the upper extremities in hypogonadal men, but studies have failed to show increased strength in men who are not hypogonadal. 5α-Dihydrotestosterone has been shown to increase knee flexion strength. Testosterone replacement is also beneficial in improving the functional index measure during rehabilitation following hospitalization.

Leptin levels increase with age in males, which has been shown to be associated with decreases in testosterone. Testosterone replacement decreases fat mass, with subcutaneous adiposity being reduced to a greater extent than abdominal visceral fat.

Behavioral effects Epidemiologists have shown a strong correlation between low testosterone and cognitive decline with aging. Low testosterone at middle age is predictive of developing Alzheimer's disease. Testosterone replacement improves working memory, spatial memory, and trail making B; effects on verbal memory are controversial and in some studies testosterone replacement has failed to alter cognition.

The SAMP8 mouse has low testosterone levels. Testosterone replacement in this mouse model of Alzheimer's disease improves learning and memory. Testosterone treatment decreased amyloid precursor protein. Patients with Alzheimer's disease have low testosterone levels in brain tissue compared with controls. Testosterone replacement has been shown to produce small improvements in cognition in people with Alzheimer's disease.

Dysphoria has been related to low levels of testosterone. Testosterone replacement did improve depression in one study, but other studies have failed to show the same results.

Bone Testosterone is converted by the process of aromatization to estradiol, which acts on bone. Testosterone increases bone mineral density by acting on the osteoblasts and thus increases bone mineral density at the hip and lumbar spine. There are no studies examining the effect of testosterone on hip fractures.

Polycythemia Lower testosterone levels are associated with anemia in men. A rise in hematocrit is generally beneficial in anemia; elevation above the normal range may have grave consequences resulting in coronary, cerebrovascular, or peripheral vascular circulation. Injections increased hematocrit by 44%, transdermal nonscrotal patch by 15%, and scrotal patch by 5%. There is a direct relationship between testosterone dosage and the incidence of erythrocytosis. Erythrocytosis was reported in 2.8% of men on 5 mg of nonscrotal patches and 11% with 50 mg day^{-1} and 18% with 100 mg day^{-1} gel, respectively.

Benign prostatic hypertrophy Benign prostatic hypertrophy (BPH) is induced by androgens and it decreases with decreases in serum testosterone. However, a number of studies have failed to show that testosterone replacement exacerbates the voiding symptoms of BPH or causes increased urinary retention compared to placebo.

Prostate cancer Case reports have suggested that testosterone replacement therapy may convert occult cancer into a clinically apparent lesion. Prospective studies demonstrate a low frequency of prostate cancer in association with testosterone replacement therapy. The prevalence rate for prostate cancer is similar to that of the general population, as evident by prospective trials of testosterone replacement therapy showing prostate cancer in 1.1% of men followed for 6–36 months. There is some concern that the underlying prevalence of occult prostate cancer in men with low testosterone levels appears to be substantial. Despite decades of research, there is no compelling evidence that testosterone has a causative role in prostate cancer. Other side effects include gynecomastia, water retention, hypertension, and sleep apnea.

External Devices

A vacuum device is an external cylinder placed over the penis to allow air to be pumped out, resulting in engorgement of the penis with blood, creating the erection. A constriction band is fitted at the base of the penis to maintain the erect state. It is a noninvasive approach. Vacuum devices are suitable for a range of patients with chronic or occasional ED. They can be used after the removal of infected or defective penile prosthesis and they may prevent scar contraction. Disadvantages are discomfort, poor erection quality, and refusal of this mechanical therapy. Overall satisfaction has been reported to be between 70 and 94%. Caution should be exercised for patients using anticoagulation therapy and those with bleeding disorders. Vacuum device therapy might be a useful option in patients for whom other therapies are contraindicated, such as patients with heart disease.

Surgery

Penile prostheses are indicated after failure of other therapies for men of middle or old age. A psychiatrist or psychologist should participate in the preoperative evaluation of patients with psychogenic ED and determine the need for the prosthesis. Advantages are ease of use, reliability, and normal appearance and sensation. It requires destruction of the corpus cavernosum, which prevents pharmacological treatment later. Arterial revascularization may be indicated in young individuals with normal corporeal venous function who have arteriogenic ED secondary to trauma. Venous ligation for veno-occlusive dysfunction was not effective and therefore was abandoned.

See also: Autonomic Nervous System: Urogenital Control; Autonomic Nervous System: Central Urogenital Control; Sexual Behavior: Neuroendocrine Control; Stress Response: Sex Differences.

Further Reading

Bakhshi V, Elliott M, Gentili A, Godschalk M, and Mulligan T (2000) Testosterone improves rehabilitation outcomes in ill older men. *Journal of the American Geriatrics Society* 48: 550–553.

Barrett-Connor E, Goodman-Gruen D, and Patay B (1999) Endogenous sex hormones and cognitive function in older men. *Journal of Clinical Endocrinology & Metabolism* 84: 3681–3685.

Collins WE, Mckendry JBR, Silverman M, et al. (1983) Multidisciplinary survey of erectile impotence. *Canadian Medical Association Journal* 128: 1393–1399.

English KM, Mandour O, and Steeds RP (2000) Men with coronary artery disease have lower levels of androgens than men with normal coronary angiograms. *European Heart Journal* 21: 890–894.

Kaiser FE, Viosca SP, Morley JE, et al. (1988) Impotence and aging: Clinical and hormonal factors. *Journal of the American Geriatrics Society* 36: 511–519.

Kinsey AC, Pomeroy WB, and Martin CE (1948) *Sexual Behavior in the Human Male*. Philadelphia, PA: Saunders.

Morley JE, Charlton E, Patrick P, et al. (2000) Validation of a screening questionnaire for androgen deficiency in aging males. *Metabolism* 49: 1239–1242.

Morley JE, Perry HM 3rd, Kaiser FE, et al. (1993) Effects of testosterone replacement therapy in old hypogonadal males: A preliminary study. *Journal of the American Geriatrics Society* 41: 149–152.

NIH Consensus Development Panel on Impotence (1993) Impotence. *Journal of the American Medical Association* 270: 83–90.

Sih R, Morley JE, Kaiser FE, et al. (1997) Testosterone replacement in older hypogonadal men: A 12-month randomized controlled trial. *Journal of Clinical Endocrinology & Metabolism* 82: 1661–1667.

Slag MF, Morley JE, Elson MK, et al. (1983) Impotence in medical clinic outpatients. *Journal of the American Medical Association* 249: 1736–1740.

Tariq SH, Haleem U, Omran ML, et al. (2003) Erectile dysfunction: Etiology and treatment in young and old patients. *Clinics in Geriatric Medicine* 19: 539–551.

Tariq SH, Haren MT, Kim MJ, and Morley JE (2005) Andropause: Is the emperor wearing any clothes? *Reviews in Endocrine and Metabolic Disorders* 6: 77–84.

Wei M, Macera CA, Davis DR, et al. (1994) Total cholesterol and high density lipoprotein cholesterol as important predictors of erectile dysfunction. *American Journal of Epidemiology* 140: 930–937.

Wittert GA, Chapman I, Haren M, et al. (2003) Oral testosterone supplementation increases muscle and decreases fat mass in healthy elderly males with low-normal gonadal status. *Journals of Gerontology, Series A: Biological Sciences and Medical Sciences* 58: 618–625.

Estrus and Menstrual Cycles: Neuroendocrine Control

M S Smith, Oregon Health and Science University, Beaverton, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Estrous or menstrual cycles occur in mammals and refer to the phenomenon of cyclic ovarian function. This cyclic function is required to produce mature ova through a process called ovulation, which is necessary for fertilization to occur. Cyclic ovarian function also results in the production of estradiol and progesterone, the primary sex steroid hormones, with their important physiological effects on all tissues and the brain. Estrous cycles are named for the cyclic appearance of behavioral sexual activity (estrus) that occurs in all mammals except for higher primates. Menstrual cycles, which occur only in primates, are named for the regular appearance of menses due to the shedding of the endometrial lining of the uterus. The factors involved in regulating the cyclic nature of ovarian function are fairly similar in all mammals and primates, and although the length of the cycle varies among species, it is relatively invariant within species. Most of the information about regulation of the ovarian cycle has been gained through studies of the laboratory rat, sheep, and primate (monkey and human).

Cyclic ovarian function is under the control of three interacting components: hypothalamus, anterior pituitary, and ovary. The hypothalamus contains neurons that release the neuropeptide gonadotropin-releasing hormone (GnRH) into the hypophysial portal vasculature, which then acts on the anterior pituitary to stimulate the release of the gonadotropins, luteinizing hormone (LH) and follicle-stimulating hormone (FSH). The gonadotropins acting on the ovary provide the hormonal signals that regulate the ovarian cycle and stimulate production of estradiol and progesterone. In turn, the steroid hormones are part of exquisitely sensitive feedback loops that suppress or stimulate the release of GnRH from the hypothalamus and LH and FSH from the anterior pituitary. Therefore, normal ovarian cycles are regulated by feed-forward and feedback control systems. In addition to the internal hormonal signals that control ovarian cycles, external signals are important in some species. For example, the rat estrous cycle is critically linked to the daily circadian rhythm, whereas the sheep estrous cycle is dependent on photoperiod to stimulate periods of seasonal breeding.

The Ovarian Cycle

Overview

Figure 1 is a diagrammatic representation of the hormonal profiles observed during an estrous or menstrual cycle. The first day of the estrous or menstrual cycle is named for an externally observable event. The onset of behavioral estrus at the time of ovulation denotes the first day of each estrous cycle, whereas the onset of menses at the end of the luteal phase indicates day one of the menstrual cycle. Estrous and menstrual cycles consist of three phases that describe ovarian events. The follicular phase encompasses the period in which the follicle matures, secretes increased amounts of estradiol, and produces a mature ovum. Ovulation is the process by which the follicle ruptures and the mature ovum is released into the oviduct, where fertilization takes place. The process of ovulation induces the luteal phase, in which ovarian cellular elements reorganize into a corpus luteum, a structure which secretes large amounts of progesterone, a hormone required to maintain uterine support of implantation and pregnancy. In the absence of fertilization and pregnancy, the corpus luteum regresses, and progesterone levels fall; these are signal events that initiate a new follicular phase.

GnRH Neurons: The Central Neural Regulators

GnRH neurons are the key central regulators of reproductive function. Their cell bodies are localized in the preoptic area, anterior hypothalamic area, and medial basal hypothalamus. It should be noted that recent studies have identified two different genes in mammals that encode GnRH: GnRH-I and GnRH-II. It is likely that these different populations of GnRH neurons have overlapping and independent functions. However, since most data concerning the role of GnRH in the regulation of ovarian cyclicity involves GnRH-I, this section refers to GnRH for ease of understanding. The GnRH nerve terminals involved in regulating anterior pituitary function reside in the median eminence, where GnRH is released into the hypophysial portal vasculature. It is well established that the pattern of GnRH secretion into the portal vasculature is pulsatile and that changes in GnRH pulse frequency have profound effects on the amounts of LH and FSH that are released from the pituitary. One of the still outstanding mysteries of neuroendocrinology is the nature of the GnRH pulse generator. Currently, the most widely held view is that pulsatility is intrinsic to the GnRH neuron. However, since GnRH neurons are widely scattered, mechanisms

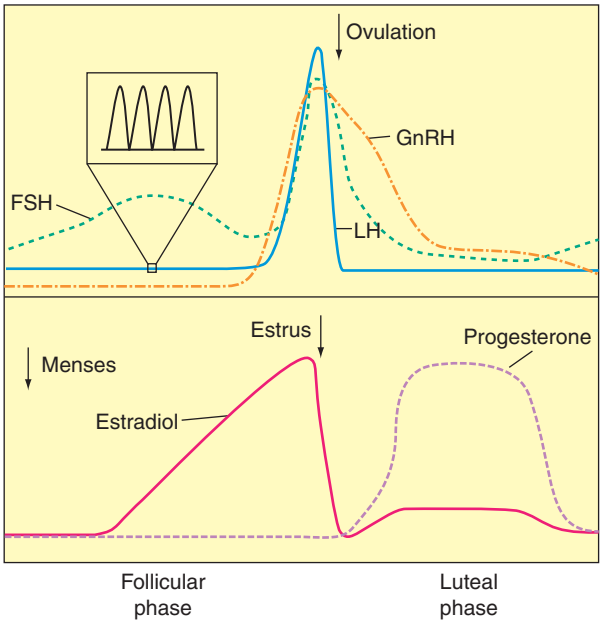

Figure 1 Diagrammatic representation of hormonal patterns and phases of the estrous or menstrual cycle. Hormones shown include hypothalamic gonadotropin-releasing hormone (GnRH), pituitary gonadotropins luteinizing hormone (LH) and follicle-stimulating hormone (FSH), and ovarian steroids (estradiol and progesterone). Estrus marks day 1 of the estrous cycle, and menses marks day 1 of the menstrual cycle.

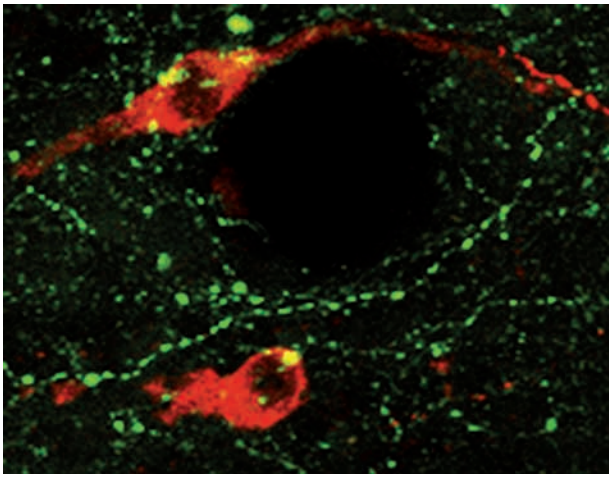

Figure 2 Photomicrograph showing gonadotropin-releasing hormone (GnRH) neurons (red) in the hypothalamus and their close association with neuropeptide Y nerve fibers (green). The image was captured using confocal microscopy, and specific neuropeptide identification was accomplished using immunocytochemistry. Neuropeptide Y fibers are shown in close contact (yellow) with GnRH neurons.

must exist to synchronize GnRH activity. Although many neurotransmitter and neuromodulatory components have been identified that act at the level of GnRH cell bodies or nerve terminals, the nature of the GnRH pulse generator remains elusive. It is generally agreed that norepinephrine, glutamate, γ-aminobutyric acid, neuropeptide Y, opioids, and GnRH itself are important modulators of GnRH neuronal activity and appear to have direct actions at GnRH cell bodies (Figure 2) or nerve terminals. Other key regulators of GnRH are estradiol and progesterone, which have both negative and positive feedback effects, although the mechanisms by which they act in this differential manner remain unclear. However, the negative and positive feedback effects of steroids on GnRH, and hence on LH and FSH, are key regulatory mechanisms that underlie the ovarian cycle (see the section titled 'Follicular phase and ovulation'). Recent studies have identified kisspeptin as a new hypothalamic neuropeptide system that regulates GnRH neuronal function. Kisspeptin neurons are located in two areas of the hypothalamus, the arcuate nucleus and the anteroventral periventricular nucleus, and are differentially regulated by ovarian steroids. It is thought that kisspeptin neurons in the arcuate nucleus are involved in the negative feedback effects of steroids to inhibit GnRH neurons, whereas kisspeptin neurons in the anteroventral periventricular nucleus are stimulated by elevated levels of estrogen and may contribute to the GnRH/LH surge at the time of ovulation.

Follicular Phase and Ovulation

LH and FSH secretion are under the control of the ovarian steroids, which primarily exert negative feedback effects at the hypothalamic level to suppress GnRH and at the pituitary level to suppress LH and FSH secretion. In most species, estradiol is the primary hormone exerting negative feedback control, although progesterone may also synergize with estradiol to enhance its effects. The ability of estradiol to exert negative feedback control on gonadotropin secretion is best illustrated by examining LH and FSH levels in the absence of estradiol, such as after ovariectomy or in postmenopausal women. LH and FSH levels increase markedly under these conditions. Under normal negative feedback conditions, LH and FSH secretion is low and pulsatile, reflecting the pulsatile release of GnRH from the GnRH-secreting neurons.

At the beginning of the follicular phase, estradiol and progesterone levels are low, reflecting the immaturity of the follicles, and the pattern of basal LH and FSH secretion is pulsatile (Figure 1). The pulse frequency varies in different species but typically averages around one pulse per hour. As the basal-levels of FSH and LH begin to stimulate follicular development during the early follicular phase, estradiol secretion from the granulosa cells increases. During the latter part of the follicular phase, as estradiol

becomes more elevated, basal levels of FSH and LH decrease somewhat because of increased negative feedback effects. The rising levels of estradiol reflect the maturation of the follicles. At some point, the estradiol levels reach a sustained threshold value such that the effect of high levels of estradiol on the hypothalamus and pituitary are dramatically different from the negative feedback effects exerted by the lower estradiol levels. Under these conditions, estradiol exerts positive feedback effects to elicit massive surges of LH and FSH (Figure 1). The surge of LH is responsible for ovulation and the release of mature ova. The positive feedback effects of estradiol are exerted at both the hypothalamus to increase GnRH release and the pituitary to enhance basal LH secretion and sensitivity to GnRH stimulation. The relative importance of the positive feedback effects of estradiol at the hypothalamus versus the pituitary in generating an LH surge most likely varies among species. Furthermore, the ability of estradiol to exert positive feedback effects is a sexually differentiated phenomenon. In most mammals, excluding higher primates, masculinization of the brain suppresses the ability of estradiol to exert positive feedback effects, precluding the ability to generate LH and FSH surges.

The length of the follicular phase varies among species and is determined by the rate at which increased amounts of estradiol are secreted and the sustained threshold value required to elicit positive feedback effects of estradiol. The combination of these factors determines when the ovulatory surge of LH will occur, signaling the end of the follicular phase. In rodents, multiple follicles develop to maturity, and the follicular phase lasts about 2 days. In sheep, one or two follicles develop during a 4-day follicular phase. In primates, the follicular phase of the menstrual cycle lasts about 12–14 days, and only one follicle typically reaches maturity.

Luteal Phase

As a direct result of the preovulatory LH surge, ova are released from mature follicles, and the follicular cells remaining in the ovary undergo a transformation from primarily estradiol-secreting cells into (primarily) progesterone-secreting luteal cells. This transition from the follicular phase to the luteal phase is characterized by rising plasma progesterone concentrations, and in some species, estradiol secretion also increases (Figure 1). Although estradiol concentrations may reach preovulatory levels during the luteal phase, LH and FSH surges do not occur because the elevated levels of progesterone block the positive feedback effects of estradiol. Progesterone's inhibitory actions are probably not exerted at the pituitary level, because pituitary responsiveness to GnRH remains normal to high during the luteal phase, but are most likely acting on the hypothalamus to greatly restrain GnRH secretion. In many species, LH pulse frequency slows considerably during the luteal phase because the elevated levels of progesterone (and possibly estradiol) cause an increase in negative feedback effects at the hypothalamus, resulting in a slowing of the GnRH pulse generator. However, the slower pulse frequency is usually associated with an increase in pulse amplitude such that mean levels of LH and FSH during the luteal phase are similar to those during the follicular phase.

During the latter part of the luteal phase, the secretory activity of the corpus luteum begins to wane. The mechanisms involved in determining the life span of the corpus luteum remain unclear but in most species they likely reside within the corpus luteum itself. However, in some species, such as ruminants, regression of the corpus luteum is dependent on the presence of the uterus, which secretes prostaglandin F2α, a potent luteolysin. The termination of luteal function is essential to initiate a new ovarian cycle. The fall in progesterone and estradiol leads to menstruation in higher primates and also causes a relaxation in negative feedback inhibition of gonadotropin secretion. Therefore, LH, and particularly FSH, levels begin to rise, initiating the growth of a new crop of follicles. The length of the luteal phase is less variable among species and averages 12–14 days. In the rat, normal luteal function lasts for only 2–3 days, but if a pseudopregnancy is induced, luteal function lasts about 12 days. In all species, fertilization of the egg and implantation of the developing embryo trigger events that prevent luteal regression. The maintenance of elevated levels of progesterone is essential to sustain pregnancy.

External Environmental Factors Influencing Ovarian Cyclicity: Energy Balance

It is evident that the hypothalamus plays the key role in regulating ovarian cyclicity by synthesizing GnRH and by acting as a site for steroid modulation. The hypothalamus is also the final common pathway whereby external factors influence reproductive activity. The influence of energy balance, coital stimulation, olfactory cues, circadian rhythms, and emotional distress, as well as the phenomenon of seasonal breeding, suggests that external environmental factors can profoundly affect the neuroendocrine reproductive axis through effects on GnRH neuronal function.

The linkage of energy balance with reproductive function is a good example of integration of the regulation of critical physiological functions at the level of the hypothalamus. Because reproduction is essential to survival of the species, and reproduction is an energy-demanding function, it is not surprising that reproduction would be tightly linked to energy balance. There are many examples that conditions of negative energy balance are linked to the suppression of cyclic ovarian function; these include fasting, anorexia nervosa, exercise, and lactation. In some cases the negative energy balance is caused by a decrease in energy intake (fasting, anorexia nervosa) and in others by an increase in energy expenditure (exercise and lactation). In all cases, the primary deficit is thought to result from the suppression of GnRH. This in turn results in very low basal levels of LH and FSH that are inadequate to stimulate ovarian follicular development. As a consequence, ovarian steroids are also very low under these conditions.

A number of signals have been identified that are likely candidates for conveying information about the status of energy balance to the hypothalamus and GnRH neurons. One category of signals is derived from peripheral tissues and conveys information about energy stores to the brain; these include leptin, insulin, glucose, fatty acids, and others. Most notable is leptin, a hormone produced by adipose tissue. Leptin levels in the blood reflect body fat stores. Therefore, under conditions of negative energy balance, leptin levels are very low. Since leptin deficiency has been shown to be associated with impaired reproductive function, leptin is thought to be an important modulator of normal reproductive function. However, leptin's effects are far-reaching, including effects on hypothalamic feeding circuitry and peripheral metabolism, so it has been difficult to determine whether leptin's role in reproductive function is direct or indirect through changes in metabolic fuels. Similar difficulties confound studies with insulin and

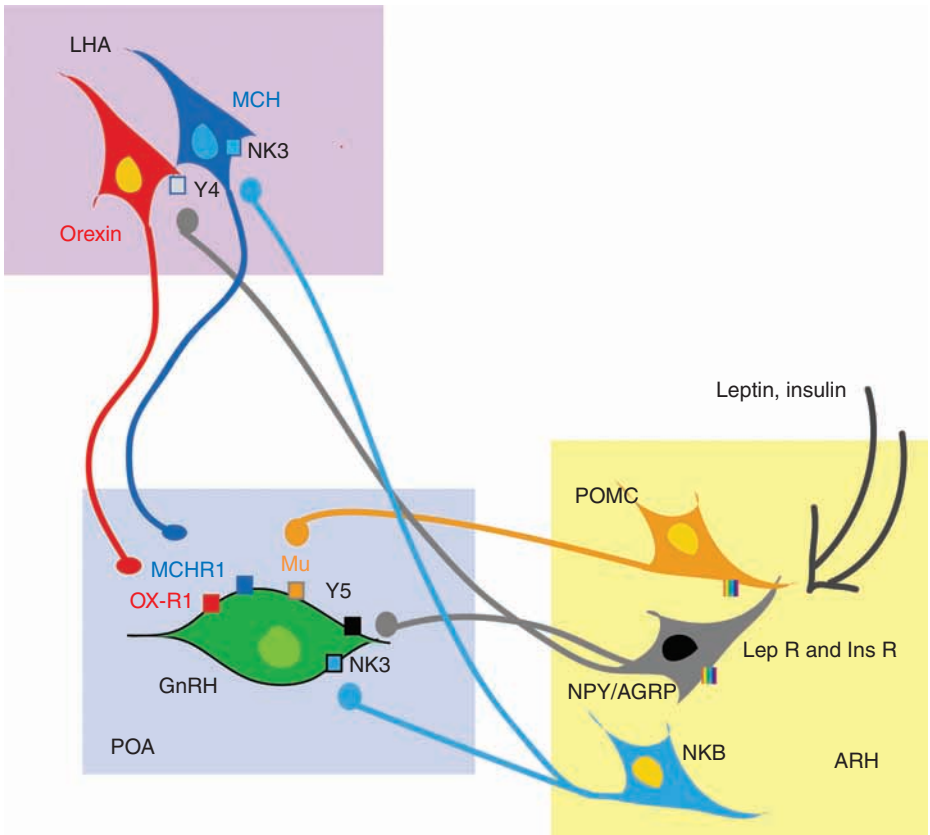

Figure 3 Hypothalamic feeding circuitry contacting gonadotropin-releasing hormone (GnRH) neurons. Arcuate nucleus (ARH) neural systems have direct input to GnRH neurons and also indirect input through connections with neural feeding circuitry in the lateral hypothalamic area (LHA), which in turn sends direct projections to GnRH neurons. Hypothalamic areas: ARH, arcuate nucleus; POA, preoptic area; LHA, lateral hypothalamic area. Feeding circuits: POMC, proopiomelanocortin (anorexigenic); NPY/AGRP, neuropeptide Y/agouti-related protein (orexigenic); NKB, neurokinin B (anorexigenic); MCH, melanin concentrating hormone (orexigenic); orexin (orexigenic). Receptor subtypes expressed on GnRH neurons: Mu (opioid), Y5 (NPY), NK (NKB), MCHR1 (MCH), and OX-R1 (orexin). Insulin, Ins; Leptin, Lep; R, receptor.

whether it plays a direct or indirect role in hypothalamic regulation of GnRH neuronal function.

Regardless of which peripheral metabolic factors may be directly involved in regulating reproductive function, it is clear they exert their effects on neural feeding circuitry within the brain, providing critical feedback information about the status of body energy stores. A key site of action for these metabolic factors is the arcuate nucleus of the hypothalamus; this nucleus contains an array of neuropeptide systems, many of which have overlapping functions in the regulation of food intake and energy balance and of GnRH activity. In addition, neuropeptide systems in the arcuate nucleus make direct contacts with additional feeding circuitry in other hypothalamic areas and can influence their activity. Therefore, the arcuate nucleus appears to be a critical integration site where metabolic signals can act to alter the activity of neural systems regulating food intake and the GnRH pulse generator. Recent studies have provided neuroanatomical evidence for the direct interaction of hypothalamic neural systems that govern food intake with GnRH neurons. Neuropeptide Y, orexin, melanin concentrating hormone, neurokinin B, cholecystokinin, and B-endorphin have all been shown to be involved in stimulating or inhibiting food intake. These same neuropeptide systems also have been shown to send fiber projections to GnRH neurons, and appropriate receptor subtypes are expressed on GnRH cell bodies (Figure 3). Numerous studies have demonstrated their ability to modulate GnRH/LH secretion. Therefore, it appears that a number of hypothalamic neuronal systems have a duality of function: regulation of food intake and regulation of reproductive function. In this manner, the control of reproductive function is tightly linked to energy availability.

See also: Gonadal Steroid Actions on Brain; Gonadotropin, Neural and Hormonal Control; Gonadotropin-Releasing Hormone: GnRH-1 System; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Neuroendocrine Control: Maternal Behavior; Neuroendocrine Peptide Processing; Neuroendocrinology; Neuroendocrinology of Affective Disorders; Neuroendocrinology of Puberty; Neuroendocrinology of Social/Affiliative Behavior; Premenstrual Dysphoric Disorder and Postpartum Major Depression: Chronobiology; Sexual Behavior: Neuroendocrine Control.

Further Reading

Dungan HM, Clifton DK, and Steiner RA (2006) Minireview: Kisspeptin neurons as central processors in the regulation of gonadotropin-releasing hormone secretion. *Endocrinology* 147: 1154–1158.

Freeman ME (2006) The neuroendocrine control of the ovarian cycle of the rat. In: Neill JD (ed.) *The Physiology of Reproduction,* 3rd edn., vol. 2, pp. 2327–2388. New York: Raven Press.

Goodman RL and Inskeep EK (2006) Neuroendocrine control of the ovarian cycle of the sheep. In: Neill JD (ed.) *The Physiology of Reproduction,* 3rd edn., vol. 2, pp. 2389–2448. New York: Raven Press.

Herbison AE (2006) Phyisology of the gonadotropin-releasing hormone neuronal network. In: Neill JD (ed.) *The Physiology of Reproduction,* 3rd edn., vol. 2, pp. 1415–1482. New York: Raven Press.

Jennes L and Conn PM (2002) Gonadotropin-releasing hormone. In: Pfaff DW, Arnold AP, Etgen AM, Fahrback SE, and Rubin RT (eds.) *Hormones, Brain and Behavior,* vol. 4, pp. 51–80. San Diego, CA: Academic Press.

Knobil E (1992) Remembrance: The discovery of the hypothalamic gonadotropin-releasing hormone pulse generator and of its physiological significance. *Endocrinology* 131: 1005–1006.

Kordon C, Gaillard RC, and Christen Y (2005) *Hormones and the Brain.* Berlin: Springer.

Krasnow SM and Steiner RA (2006) Integrating metabolism and reproduction. In: Neill JD (ed.) *The Physiology of Reproduction,* 3rd edn., vol. 2, pp. 2553–2626. New York: Raven Press.

Pfaff DW, Arnold AP, Etgen AM, Fahrback SE, and Rubin RT (eds.) (2002) *Hormones, Brain and Behavior.* San Diego, CA: Academic Press.

Smith MS and Grove KL (2002) Integration of the regulation of reproductive function and energy balance: Lactation as a model. *Frontiers in Neuroendocrinology* 23: 225–256.

Zeleznik AJ and Pohl CR (2006) Control of follicular development, corpus luteum function, recognition of pregnancy, and the neuroendocrine regulation of the menstrual cycle in higher primates. In: Neill JD (ed.) *The Physiology of Reproduction,* 3rd edn., vol. 2, pp. 2449–2510. New York: Raven Press.

Event-Related Potentials (ERPs)

M D Rugg, University of California at Irvine, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Event-Related Potentials

Event-related potentials (ERPs) are scalp-recorded fluctuations in the brain's electrical activity (the electroencephalogram or EEG) elicited by a stimulus event such as the presentation of a word. The size of these changes is usually small in comparison to the background EEG activity in which they are embedded. Therefore, to obtain ERP waveforms with satisfactory signal-to-noise ratios, EEG samples time-locked to stimulus presentation on a number of experimental trials (typically, between 20 and 30) are averaged together. The waveforms represent the average of the neural activity elicited by the stimulus events belonging to each trial.

ERPs have proved to be valuable for studying cognitive functions such as memory. Neural activity (or more accurately, the fraction of activity detectable at the scalp) associated with the processing of different classes of stimulus events can be measured with a temporal resolution that allows the neural correlates of cognitive processes to be characterized with a precision of a few tens of milliseconds. This permits quite precise upper-bound estimates to be placed on the time required by the brain to discriminate between different stimulus classes (e.g., studied vs. unstudied items in a recognition memory test). This level of temporal resolution is presently unattainable with techniques, such as functional magnetic resonance imaging (fMRI), that measure neural activity indirectly by detecting changes in blood flow or oxygenation. On the other hand, the spatial resolution of the ERP technique is in general quite poor and compares very unfavorably with the resolution attainable with fMRI. The two methods thus offer complementary perspectives on the neural correlates of human memory.

Despite their poor spatial resolution, ERPs can nonetheless be used to investigate whether different experimental conditions engage functionally dissociable cognitive processes. The rationale for this application of ERPs rests on the assumption that if two experimental conditions are associated with qualitatively different patterns of scalp electrical activity, there are grounds for assuming that the conditions engaged at least partially nonoverlapping neural, and hence cognitive, processes. By contrast, purely quantitative differences in the ERPs associated with two conditions constitute evidence for different levels of engagement of the same neurocognitive processes, at least so far as can be detected from scalp recordings. To distinguish between these scenarios in practice it is necessary to record ERP waveforms from enough electrode sites to permit the spatial configurations of the electrical fields associated with different experimental conditions to be accurately characterized. In the research described below, the analysis of the spatial distribution (the 'scalp topography') of ERP effects associated with different experimental conditions occupies a central place; when two topographies differ, this is taken as evidence that the cognitive operations engaged in the respective experimental conditions are at least partially dissociable.

Memory Encoding

Even when presented under highly uniform study conditions, stimulus events vary in how likely they are to be remembered on a later memory test. It is generally assumed that one reason for such disparities in later memory performance is variability in recruitment of processes that support the initial encoding of an event into memory. Key questions include the time course of these encoding processes, and whether the processes vary according to the class of event that is presented (e.g., words vs. pictures) or the manner in which the event is processed (e.g., whether the study task emphasizes semantic or nonsemantic processing).

ERPs have been used extensively to investigate the neural correlates of memory encoding with the 'subsequent memory procedure.' The procedure was introduced in the early 1980s, and continues in its popularity to the present time. Experimental subjects are first presented with a series of items for study while their EEGs are recorded. Sometime later they undertake a memory test, allowing each study item to be classified as 'later remembered' or 'later forgotten.' The EEG data obtained during the study phase can then be retrospectively averaged to form separate ERPs for those items that were later remembered as opposed to those that were later forgotten. This makes it possible to assess whether, and in what ways, ERPs can 'predict' whether study items will be later remembered.

An example of an ERP subsequent memory effect is illustrated in Figure 1. ERP waveforms typically begin to differ according to subsequent memory performance within 200 ms or so of the presentation of the study items, the waveforms elicited by later-remembered items demonstrating a sustained positive

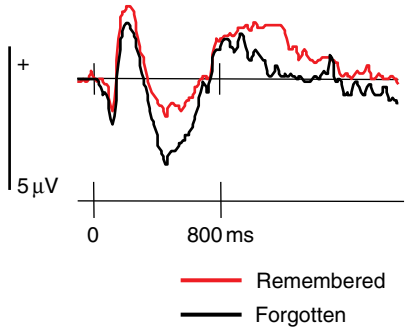

Figure 1 An ERP subsequent memory effect. Grand average (i.e., collapsed across subjects) ERP waveforms elicited by visually presented words from an electrode over the frontal scalp during an incidental study task. The ERPs have been segregated according to whether they were elicited by words that were remembered or forgotten on a recognition memory test undertaken approximately 30 min after the study phase. From approximately 250 ms onward, ERPs elicited by remembered words are more positive-going than the waveforms elicited by words that were later forgotten.

shift relative to those elicited by items that are destined to be forgotten. While the findings differ in detail across studies, these ERP differences – often referred to as 'subsequent memory' or 'Dm' (differences due to memory) effects – have been observed with a variety of different study materials and tasks, and with memory tests assessing both recognition and recall. Importantly, subsequent memory effects can be obtained in ERPs when measures of study task performance such as accuracy and response time do not differentiate items that are later remembered or forgotten. It seems unlikely, therefore, that the ERP effects merely reflect factors such as the amount of attention allocated to particular study items, or fluctuations in alertness or arousal, since such factors would be expected to impact behavior as well as brain activity.

It has proven difficult to identify all of the variables that determine the exact form that ERP subsequent memory effects will take in any given study, and in particular what dictates the scalp topography of the effects. The balance of the evidence suggests however that the cognitive processes reflected in the effects are recruited primarily in service of explicit rather than implicit memory. One study, for example, contrasted subsequent memory effects for word stem cued recall (held to depend largely on explicit memory) with the effects associated with stem completion in the absence of explicit memory (when completion is assumed to reflect implicit memory). Words that were later remembered explicitly elicited a subsequent memory effect similar in character to that illustrated in Figure 1. By contrast, the subsequent memory effects elicited by items later remembered implicitly were of opposite polarity, in that the waveforms for 'forgotten' items were more positive than those for 'remembered' items. In addition to shedding light on the functional significance of ERP subsequent memory effects, these findings support the notion that the processes supporting explicit and implicit memory encoding are qualitatively distinct.

Perhaps the most important conclusion to come out of ERP studies of encoding so far derives from the time course of subsequent memory effects. Typically, the effects onset within a few hundred milliseconds of the presentation of study items (Figure 1), during the time that the items are still being processed in service of the study task. Thus, the cascade of processes that culminate in the formation of a lasting memory is initiated early in the course of the processing of a stimulus event. A major challenge for future research is to integrate ERP findings with the results from the burgeoning number of fMRI studies that have also employed the subsequent memory procedure to investigate the neural substrates of memory encoding.

ERPs and Retrieval

Recognition Memory

As in the case of encoding, ERPs have been used to study the neural correlates of memory retrieval since the early 1980s. By far the most popular experimental approach has been to contrast waveforms elicited by different classes of item in tests of recognition memory. ERPs elicited by correctly detected old test items differ markedly from the ERPs elicited by items correctly rejected as new. These differences, often referred to as ERP 'old/new effects,' have long been held to represent neural correlates of successful retrieval. The effects extend over several hundred milliseconds, encompassing a negative-going deflection peaking around 400 ms and a later positive-going wave that peaks around 700 ms. As can be seen in Figure 2, the scalp distributions of the old/new effects centered on these two peaks are markedly different; whereas the earlier effect is maximal over the frontal scalp, the later-occurring effect shows a maximum over the left parietal scalp. The two effects are sometimes referred to as 'mid-frontal' and 'left parietal' old/new effects, respectively.

Recent years have seen an intensive effort to identify the functional significance of the old/new effects illustrated in Figure 2. Many of these studies have focused on whether the effects can be dissociated in ways compatible with 'dual-process' models of recognition memory. These models propose that recognition judgments are supported by two classes of memory process. Recognition can occur when a test

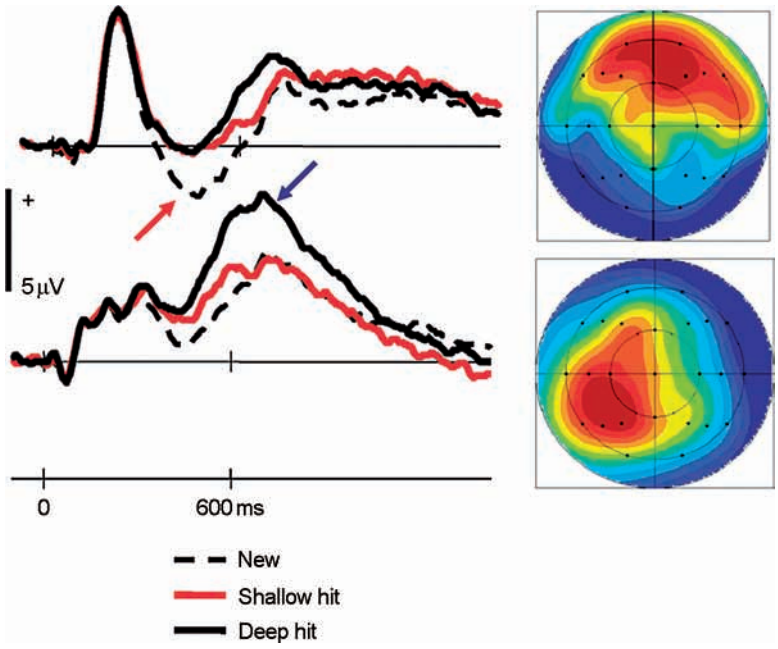

Figure 2 ERP old/new effects associated with recognition memory. Left: grand average ERP waveforms elicited by visually presented words in a recognition memory test. The waveforms were recorded from electrodes over mid-frontal (top) and left parietal (bottom) scalp. Illustrated are waveforms elicited by unstudied (new) items correctly judged to be new, along with waveforms elicited by recognized studied words that had been subjected to either a deep, semantic (deep hit) or a shallow, alphabetic (shallow hit) study judgment. Arrows indicate the mid-frontal and left parietal old/new effects (see text). Right: scalp distribution maps of old/new effects between 300 and 500 ms, and 500 and 800 ms, highlighting the distinct topographies of the mid-frontal and left parietal effects. Maps are arranged such that the nose is at the top, and the relative size of each effect over the scalp is scaled from blue (small) to red (large).

item elicits retrieval of a specific past episode involving the item, a form of memory termed 'recollection.' A test item can also be recognized on the basis of its 'familiarity' – a faster-acting, acontextual form of memory held to be dissociable from recollection on phenomenal, functional, and neurological grounds. According to dual-process models, recollection and familiarity depend upon qualitatively distinct cognitive (and neural) processes. A rival family of 'single-process' models propose that the two forms of memory depend upon the same underlying mechanisms, and differ in a quantitative rather than a qualitative fashion.

Several ERP studies have attempted to adjudicate between dual- and single-process models of recognition memory and, in parallel, to shed light on the functional significance of ERP old/new effects. The studies have employed a variety of experimental procedures originally developed by experimental psychologists to fractionate recollection and familiarity using behavioral methods. These procedures have generally involved manipulation of variables hypothesized to act selectively on one or the other recognition process, and hence to assess whether ERP old/new effects can be selectively modulated. In one study, for example, a study task manipulation ('depth' of processing) was employed that was assumed to vary the likelihood of later recollection while having little effect on familiarity. As is illustrated in Figure 2, this manipulation had a marked effect on the left parietal old/new effect. The effect was prominent when elicited by recognized items studied under conditions that promote recollection, but absent for items assumed to have been recognized largely on the basis of their familiarity. By contrast, a robust mid-frontal effect was elicited by both classes of recognized item. The mid-frontal effect was absent, however, for studied items that were misclassified as new, raising the possibility the effect reflects processes tied to explicit recognition judgments. These findings suggest that mid-frontal and left parietal old/new ERP effects are, respectively, correlates of familiarity and recollection.

Subsequent findings have added weight to this suggestion. Notably, words and pictures that, by virtue of their perceptual similarity to studied items, generate high levels familiarity-driven false alarms elicit the mid-frontal, but not the left parietal, old/new effect. In addition, the amplitude of the mid-frontal effect varies in a monotonic fashion with subjects' ratings of the familiarity of the test items. By contrast, the left-parietal effect appears to be associated exclusively with test items that elicit recollection, as indexed either by subjective report or, more objectively, in the contrast between items associated with successful

versus unsuccessful retrieval of information about their study context (i.e., items associated with accurate vs. inaccurate source memory).

The differing scalp distributions of the mid-frontal and left parietal old/new effects strongly suggest that the effects reflect activity in distinct brain regions or circuits. Coupled with the evidence that the two effects can also be fully dissociated experimentally, this gives strong support to dual-process models of recognition memory. The precise cognitive operations reflected in each effect remain to be elucidated, however.

Other ERP Old/New Effects

In addition to the mid-frontal and left parietal effects described previously, other ERP modulations have also been reported in studies employing recognition-like memory tests. A few studies, for example, have described effects that precede the mid-frontal effect. The functional significance of these effects remains uncertain. Other studies have reported old/new effects that appear to be independent of subjects' explicit memory judgments, and therefore which are possible correlates of implicit memory (priming). More work is needed however to fully characterize these effects and to assess their consistency across different experimental procedures and stimulus types.

One effect that has been consistently reported in a diverse range of studies has been termed the 'right frontal old/new effect.' An example of the effect is shown in Figure 3. Typically, it onsets around 500 ms poststimulus and takes the form of a sustained positive-going shift that can persist for well over a second. The effect was first identified in studies of source memory, where subjects are required to both identify previously studied test items and indicate the context in which the item had been studied (e.g., whether it had been heard in a male or female voice). Initially the effect was interpreted as a reflection of the additional processing engaged when recollected information must be evaluated prior to response selection. In later studies, however, it became clear that the right frontal effect can be elicited in the absence of recollection, especially if item familiarity is relatively weak. Under these circumstances, the information elicited by an old item is thought to require monitoring and evaluation to establish whether its familiarity is sufficiently high to warrant an 'old' response. Currently, the most popular interpretation of the right frontal effect is that it reflects the engagement of 'postretrieval' processes that support the evaluation of the products of a retrieval attempt.

ERP Studies of Recall

The requirement to identify an event to which EEG sampling can be time-locked precludes the use of ERPs to study free recall. A significant number of studies have however investigated the ERP correlates of cued recall, when retrieval is elicited not by 'copy cues,' as in recognition memory, but by partial cues such as word stems or word fragments (e.g., MOT or M-T-R). In one variant of the cued recall procedure, cues that can be completed with studied words are intermixed with other cues that cannot be completed with a studied item. Subjects are instructed to attempt to complete each cue with a studied item but, if this is not possible, to generate the first word that comes to mind. Importantly, subjects are further required to endorse each completion as either studied or unstudied, allowing correct completions associated with explicit memory for the study episode to be separated from completions due to chance or the influence of implicit memory. Figure 4 illustrates typical results

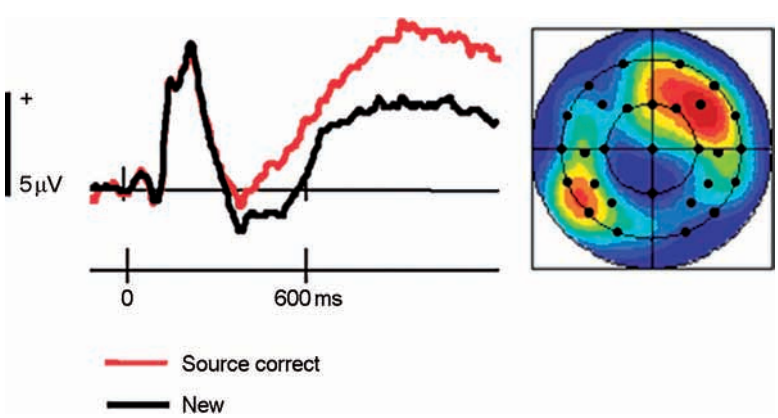

Figure 3 The right frontal ERP old/new effect. Left: Waveforms elicited by visually presented words from a right frontal electrode. Waveforms elicited by studied words accorded correct source memory judgments (source correct) are overlaid with waveforms elicited by correctly judged new items (new). Right: Scalp distribution of the old/new effect between 800 and 1100 ms. A prominent right frontal maximum is evident, along with the remains of an earlier onsetting left parietal effect.

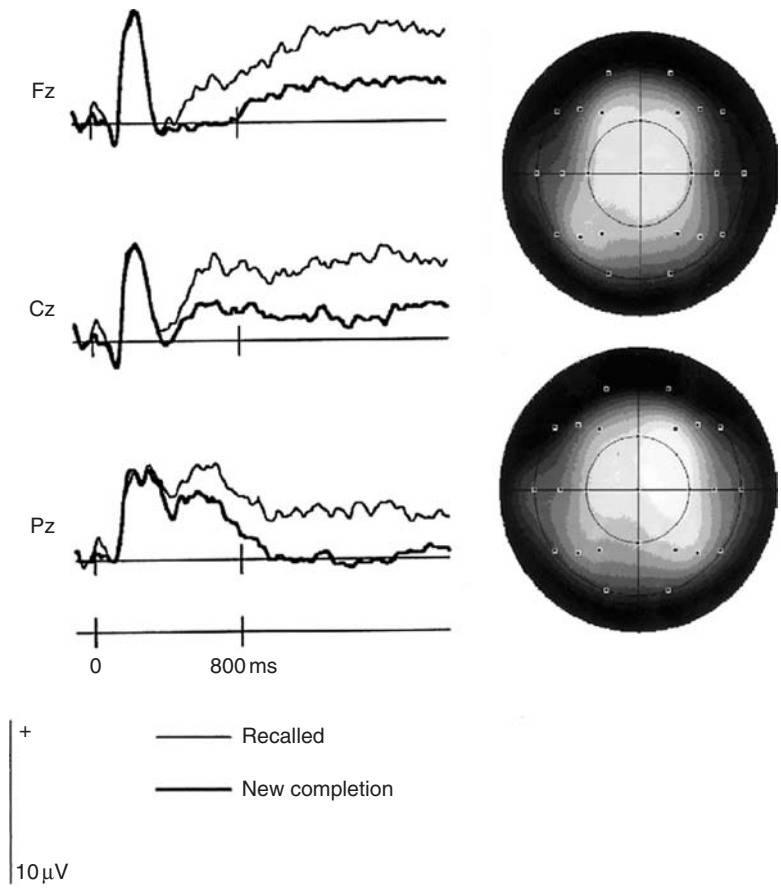

Figure 4 Old/new ERP effects associated with cued recall. Left: ERP waveforms from frontal (FZ), central (CZ), and parietal (PZ) midline lectrodes elicited by three-letter word stems (see text). ERPs elicited by stems completed by a studied word and endorsed as old are overlaid on waveforms elicited by stems completed by unstudied items and correctly endorsed as new. Right: Scalp distribution maps of the cued-recall old/new effect between 400 and 1200 ms and 1200–1900 ms, illustrating the diffuse and relatively unchanging nature of the topography of the effect (increasing amplitude scales with increasing brightness). Reproduced from Allan K and Rugg MD (1997) An event-related potential study of explicit memory on tests of cued recall and recognition. *Neuropsychologia* 35: 387–397, with permission from Elsevier.

from a study employing this procedure with word stems as the cues. Relative to the waveform elicited by stems corresponding to new items, ERPs elicited by stems completed with studied words and endorsed as old exhibit a sustained positive-going shift. This effect is not found if a 'correct' completion is incorrectly endorsed as unstudied. Nor is it evident for incorrect completions wrongly judged to be old. These findings suggest that the effect represents processing selectively associated with veridical explicit memory retrieval. The cued recall ERP effect is more diffusely distributed over the scalp than are recognition old/new effects, suggesting the involvement of neural generators additional to those active during recognition. Moreover, unlike in the case of recognition, there is little evidence that the scalp distribution of the recall effect changes with time.

Comparison of the old/new effects elicited in recognition and recall tasks leads to two principal conclusions. First, their differing scalp distributions suggest that the processes engaged during successful recognition and recall are at least partially dissociable, a conclusion that converges well with the findings from functional neuroimaging studies that have compared retrieval-related neural activity in the two kinds of test. Second, the finding that recall and recognition effects have similar onset latencies is difficult to reconcile with 'generate-recognize' models of cued recall. These models propose that cued recall involves an iterative, two-stage process, whereby candidate responses are first generated in response to the cue and then subjected to a recognition judgment. According to such models, the onset latency for ERP recall effects should be delayed relative to recognition effects, reflecting the additional generate stage that is needed for recall but not recognition. Rather, the ERP findings lend support to rival models that propose that recall cues usually gain direct access to stored memory representations.

Neural Generators of ERP Retrieval Effects

It has already been noted that the spatial resolution of scalp-recorded ERPs is poor. Indeed, there are currently no generally accepted formal methods for localizing the neural generators of an ERP effect. A popular informal approach is to seek parallels between ERP effects and the findings from functional neuroimaging studies in which the same or closely similar experimental manipulations were employed. The most striking such parallels have been observed for the left parietal and the right frontal old/new effects described previously. The relatively focal maxima of these effects are consistent with generators located in underlying cortex, that is, left-parietal and right frontal cortex, respectively.

Several functional neuroimaging studies have employed experimental procedures that allowed recognized test items to be segregated according to whether they were recollected or instead judged old solely on the basis of familiarity. Among the most consistently reported regions to demonstrate recollection-related enhancement of activity in these studies has been left lateral parietal cortex. These findings have led several investigators to propose that the left parietal old/new effect does indeed reflect, at least in part, recollection-related activity originating in this region.

An equally striking parallel exists for the right frontal effect. Functional neuroimaging studies that have investigated the neural correlates of memory monitoring have frequently implicated dorsolateral and anterior regions of right prefrontal cortex as playing a key role in this function. Thus, it has been proposed that the right frontal ERP old/new effect originates in large part from right prefrontal cortex.

Essentially nothing is known about the intracerebral loci of the generators of the other memory-related ERP effects described in the preceding sections. Even in the case of the left parietal and right-frontal effects, it should be emphasized that the hypotheses concerning their generators are based on indirect, correlative evidence, and that the validity of hypotheses has yet to be tested directly, for example, by investigating the consequences of focal cortical lesions or transcranial magnetic stimulation.

Concluding Comments

The contribution of ERPs to the cognitive neuroscience of memory has been considerable. The combination of experimental versatility, high temporal resolution, and ability to discriminate between different patterns of neural activity has allowed researchers to use ERPs to fractionate and assign distinct neural 'signatures' to several theoretically important memory functions. The challenge for the future is to go beyond the description of ERP effects in terms of their patterning over the scalp and correlations with particular experimental manipulations, and to develop methods that allow the effects to be rigorously linked to specific cognitive operations and their neural substrates. This will require close integration with complementary methods for noninvasive measurement of brain activity, most notably, fMRI and allied techniques.

See also: Aging: Brain Potential Measures and Reaction Time Studies; Electroencephalography (EEG); Electrophysiology: EEG and ERP Analysis; Event-Related Potentials (ERPs) and Cognitive Processing; Executive Function and Higher-Order Cognition: EEG Studies; Recognition Memory.

Further Reading

Allan K and Rugg MD (1997) An event-related potential study of explicit memory on tests of cued recall and recognition. *Neuropsychologia* 35: 387–397.

Curran T (2000) Brain potentials of recollection and familiarity. *Memory and Cognition* 28: 923–938.

Friedman D and Johnson JR (2000) Event-related potential (ERP) studies of memory encoding and retrieval: A selective review. *Microscopy Research and Technique* 51: 6–28.

Luck SJ (2005) *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

Paller KA (1990) Recall and stem-completion priming have different electrophysiological correlates and are modified differentially by directed forgetting. *Journal of Experimental Psychology: Learning, Memory and Cognition* 16: 1021–1032.

Paller KA, Hutson CA, Miller BB, et al. (2003) Neural manifestations of memory with and without awareness. *Neuron* 38: 507–516.

Rugg MD and Allan K (2000) Event-related potential studies of long-term memory. In: Tulving E and Craik FIM (eds.) *The Oxford Handbook of Memory*, pp. 521–538. New York: Oxford University Press.

Rugg MD, Mark RE, Walla P, et al. (1998) Dissociation of the neural correlates of implicit and explicit memory. *Nature* 392: 595–598.

Sanquist TF, Rohrbaugh JW, Syndulko K, et al. (1980) Electrocortical signs of levels of processing – perceptual analysis and recognition memory. *Psychophysiology* 17: 568–576.

Schott B, Richardson-Klavehn A, Heinze HJ, et al. (2002) Perceptual priming versus explicit memory: Dissociable neural correlates at encoding. *Journal of Cognitive Neuroscience* 14: 578–592.

Wilding EL and Rugg MD (1996) An event-related potential study of recognition memory with and without retrieval of source. *Brain* 119: 889–906.

Event-Related Potentials (ERPs) and Cognitive Processing

S A Hillyard, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Origins of Event-Related Potentials

As information is processed within the neural networks of the human brain, the electrical activity arising from millions of participating neurons is summed to form field potentials that can be recorded through the intact scalp. The brain's field potentials include the rhythmic voltage oscillations of the ongoing electroencephalogram (EEG) and the briefer evoked or event-related potentials (ERPs) that are triggered in association with sensory, motor, and cognitive events. Whereas the spontaneous EEG rhythms are sensitive monitors of general states of arousal, consciousness, and the sleep–wake cycle, the ERPs reflect more discrete patterns of neural activity that underlie specific perceptual and cognitive processes. These activity patterns are revealed in the scalp-recorded ERP with a high degree of temporal precision (on the order of milliseconds), but the method generally lacks the precision of anatomical localization that is provided by neuroimaging techniques such as positron emission tomography or functional magnetic resonance imaging.

Recent approaches in modeling the neural generators of ERPs, however, have enabled their localization to specific brain structures with improved accuracy. By mapping the scalp topography of an ERP and applying algorithms based on the physics of electrical volume conduction, the positions of the underlying neural generators can be estimated with respect to brain anatomy. In a number of cases, the anatomical regions generating surface ERPs have been verified by intracranial ERP recordings, by magnetoencephalography, or by studies of patients with circumscribed brain lesions. Combined studies of ERP recordings together with neuroimaging methods that reveal patterns of cerebral blood flow and metabolism offer the possibility of defining both the precise time course and the anatomical substrates of information processing in the human brain.

The ERP recording methodology involves the computerized analysis of the voltage–time waveforms recorded from different points on the scalp. It is generally necessary to average ERP waveforms over many stimulus presentations in order to achieve adequate signal-to-noise ratios with respect to the ongoing EEG. Newer spatiotemporal filtering procedures (e.g., wavelet filters) can also help to distinguish weak ERP signals from the background EEG.

The computer-averaged ERPs triggered by sensory signals consist of a sequence of positive (P) and negative (N) voltage deflections (components) that have characteristic time delays and wave shapes. The typical ERP elicited by a brief auditory stimulus, for example, consists of 16 or more components appearing over the interval 1.5–1000 ms after stimulus (Figure 1). The early ERP components that occur before approximately 50 ms represent evoked neural activity in the auditory pathways that is relatively insensitive to changes in the psychological state of the subject. In contrast, many of the later components are highly sensitive to the significance of a stimulus and to the cognitive demands of the task. These 'endogenous' ERPs have been studied extensively as indices of perceptual and cognitive processes. Several types of endogenous ERPs are depicted by the dotted and dashed tracings in Figure 1: the Nd wave is elicited specifically by auditory signals that are actively attended, whereas the N2, P3 (P300), and SW (slow wave) components are triggered by signals that are being actively discriminated and deliver task-relevant information.

Visual and somatosensory stimuli elicit ERPs with characteristic waveforms and scalp topographies that correspond to sensory processing in their respective modality-specific cortical areas. The visual ERP to a brief flash, for example, begins with a 'C1' component at 50–60 ms that appears to originate from the primary visual (striate) cortex, followed by a 'P1' component (at 80–120 ms) that is generated in extrastriate cortical areas of both the dorsal and ventral visual streams (Figure 2). Another type of ERP is elicited by a visual stimulus that flickers continuously at a fixed rate in the range of 6–40 Hz. This steady-state visual evoked potential (SSVEP) has a sinusoidal waveform at exactly the flicker frequency (or a harmonic thereof) and provides a continuous measure of visual–cortical activation that can be readily quantified in the frequency domain.

A major research strategy is to identify individual ERP components as markers of specific perceptual or cognitive processes. This has been achieved in a number of cases through the systematic study of the stimulus, task, and behavioral performance variables that influence the component in question. These ERP markers reveal the timing and order of successive information processing stages and can thus reveal the underlying structure of cognitive operations. In this way, the ERPs yield information about the neural bases of cognitive processes including selective attention, stimulus recognition, decision making, memory storage and retrieval,

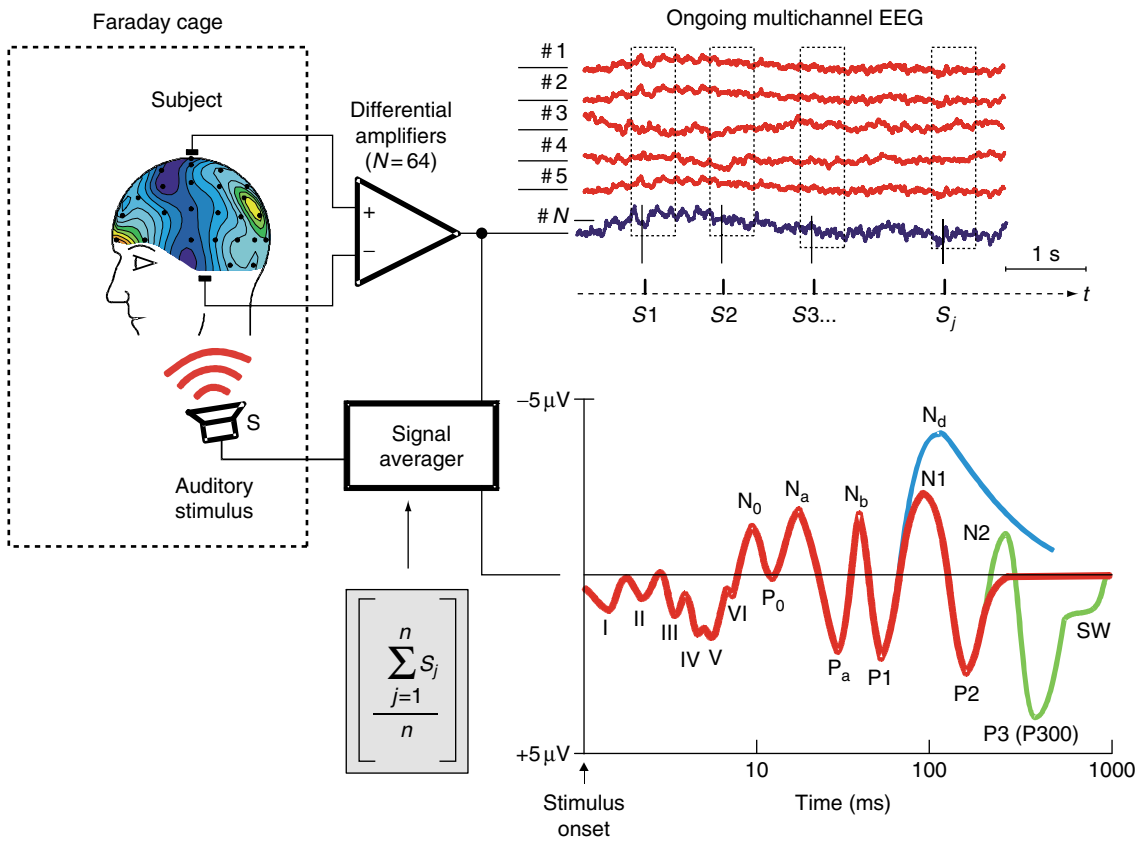

Figure 1 Waveform of the computer-averaged auditory ERP to a brief sound. The ERP is generally too small to be detected in the ongoing EEG (top) and requires computer averaging over many stimulus presentations to achieve an adequate signal-to-noise ratio. The logarithmic time display allows visualization of the early brain stem responses (waves I–VI), intermediate evoked components, and long-latency endogenous components (Nd, N2, P300, and SW, or slow wave). ERP components are labeled according to their polarity, P (positive) or N (negative), followed either by a number or letter (P3, Pa, etc.) that specifies the ordinal position of the wave in a particular sequence or by a larger number (P300, etc.) that refers to the latency of the wave in milliseconds. Courtesy of Wolfgang Teder-Salejarvi.

and language processing. As objective signs of cognitive operations, abnormal ERP waveforms have diagnostic significance for information-processing disorders associated with developmental, psychiatric, and neurological syndromes.

Selective Attention

The brain's attentional systems allow relevant sensory inputs to be processed fully while suppressing irrelevant inputs, thereby protecting higher cognitive centers from sensory overload. Considerable research has been directed at the question of whether stimulus selections occur at an early stage of processing by suppressing inputs belonging to unattended sensory channels or at a later stage of processing after stimuli have been more fully analyzed. Evidence favoring an early selection mechanism has been obtained in a number of ERP experiments in auditory, visual, and somatosensory modalities, in which discrete stimuli are delivered at rapid rates to two or more sensory channels. For example, if subjects are required to listen to a series of clicks, tones, or speech sounds in one ear and ignore a comparable sequence of sounds in the opposite ear, all stimuli in the attended ear will elicit an enlarged, negative ERP (the Nd component) starting at approximately 70 or 80 ms. Given the short-onset latency of this component and its elicitation by all stimuli belonging to the attended channel, it has been hypothesized that this ERP is a sign of an early selection process whereby stimuli are selected or rejected on the basis of whether they possess the elementary sensory feature that defines the channel (e.g., ear of entry). This early selective mechanism also appears to involve an amplitude modulation of evoked ERP components in the auditory cortex at latencies of 20–50 and 70–100 ms, as well as the longer lasting endogenous Nd wave (also called 'processing negativity') that appears to be a neural sign of the continued sensory analysis of the attended channel stimuli.

In accordance with the early selection hypothesis, stimuli in unattended channels, which lack the Nd

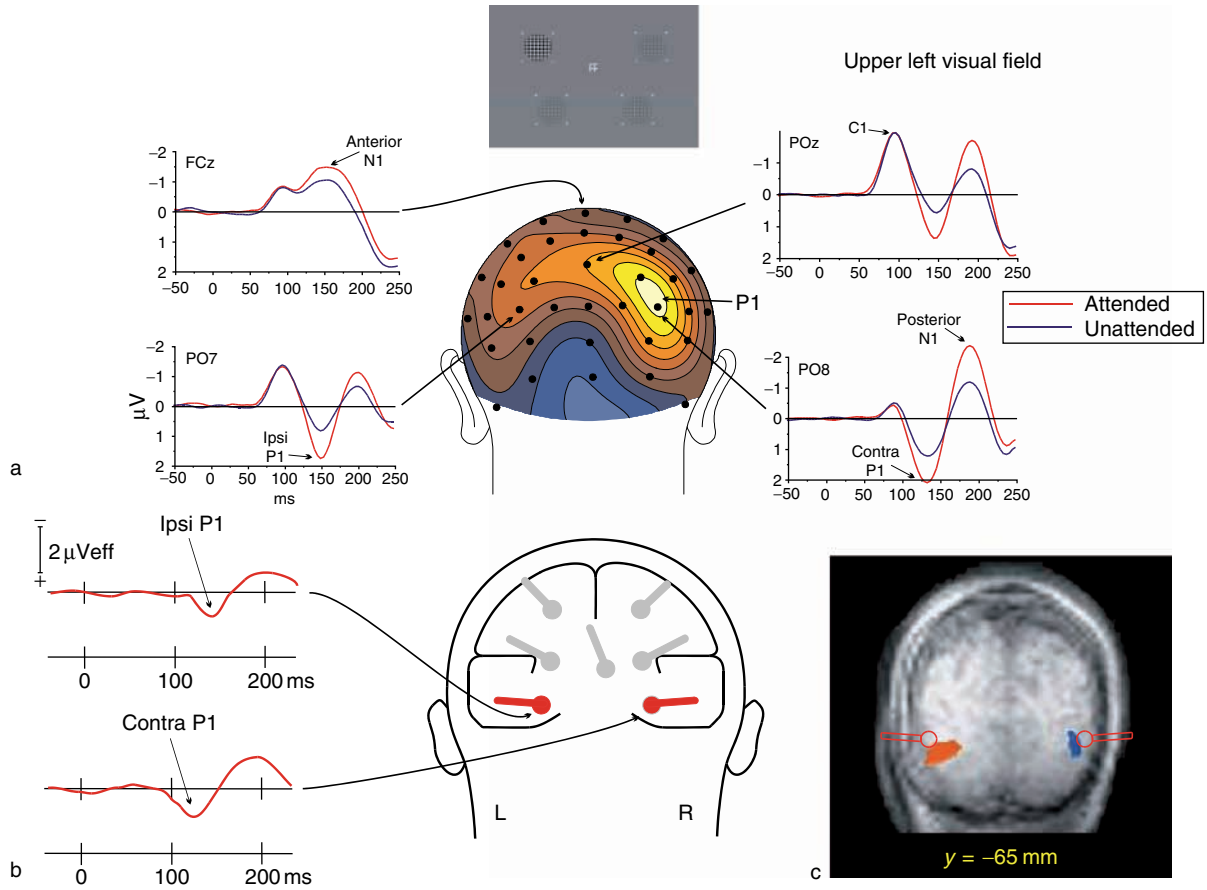

Figure 2 (a) Amplitude enhancement of visual ERP components during visual spatial attention. Circular checkerboard stimuli were flashed in random order to the four quadrants of the visual field (top) while subjects focused their attention on stimuli in only one quadrant at a time. The eyes remained fixed on a central cross. ERPs were recorded from 62 scalp sites and were averaged separately for stimuli in each quadrant under each attention condition. Waveforms shown are grand average ERPs from 4 scalp locations in response to the upper left field stimuli under conditions in which they were attended (red tracings) and when unattended (blue tracings, while attention was directed to the upper right stimulus). Note the attention-related enhancements of ERP components, including the contralateral P1 (130 ms), ipsilateral P1 (150 ms), posterior N1 (190 ms), and anterior N1 (150 ms) waves. In contrast, the earlier C1 component (90 ms), which is proposed to originate from the primary visual (striate) cortex, was not affected by attention. The head map shows the scalp voltage distribution of the late phase of the P1 component (100–130 ms) that was enhanced by attention over contralateral occipital areas. (b) Dipole model of the neural generators of the late P1 attention effect. Head model shows that dipoles accounting for the P1 attention effect are located in ventral occipital cortex. Waveforms at left represent the time course of activity in each model dipole corresponding to the enhanced neural activity in contralateral and ipsilateral occipital lobes. (c) These dipole positions correspond closely to focal regions of neural activity revealed by functional magnetic resonance imaging (fMRI) obtained in the same experiment but in a separate session. Red and blue regions represent cortical zones showing increased neural activity during attend-right and attend-left conditions, respectively. These findings indicate that spatial attention modulates the flow of visual information along the ventral–occipital pathway in the time range 100–130 ms after stimulus onset. Data from Di Russo F, Martinez A, and Hillyard SA (2003) Source analysis of event-related cortical activity during visuo-spatial attention. *Cerebral Cortex* 13: 486–499.

wave, are not processed as fully as are attended-channel stimuli. Thus, in experiments in which the attended channel is defined in terms of two features such as pitch and localization, stimuli that are rejected early on the basis of one feature do not show ERP signs of being processed for the other feature. These data and comparable findings in visual attention experiments support hierarchical, interactive models of attention. It appears that elementary stimulus features such as location and pitch are selected first, whereas more complex combinations of attributes are registered and selected at subsequent stages. The timing of different ERP components indicates that some feature analyses are performed serially, whereas others occur in parallel.

In the visual modality, the earliest ERP changes occur during spatial selective attention, which is employed, for example, when a subject attends to a flashing light in one visual field while ignoring concurrent flashes in the opposite field. Flashes in

the attended field elicit enhanced P1 and N1 components between 80 and 200 ms after stimulus onset (Figure 2). These ERP components that are modulated by visual–spatial attention appear to be generated in multiple extrastriate visual cortical areas rather than in the primary visual (striate) cortex. This ERP pattern of P1–N1 enhancement reflects an early selection mechanism that facilitates the processing of stimuli that are located within the 'spotlight' of spatially focused attention. In contrast, selective attention to other visual features, such as color, spatial frequency, and shape, is manifested by longer latency (150–300 ms) ERP configurations. Since these later attention-sensitive components differ from one another in scalp distribution as well as in wave shape, they can be used to chart the time course of separate feature analyses in different visual–cortical areas. Both ERP and neuroimaging studies suggest that visual attention operates by modulating the flow of information through multiple extrastriate areas under the influence of an attentional control network in the frontal and parietal lobes.

ERP recordings have also helped to clarify mechanisms of visual search in tasks in which subjects attempt to identify a target stimulus in a visual display. Detection of such a target triggers an ERP component known as the N2pc, which is proposed to reflect the focusing of attention on the target's location and the filtering out of interfering information from nearby distractor items. When multiple target items are present in a display, the timing of N2pc amplitude changes lends support to serial models rather than parallel models of visual search. The allocation of attention in a visual display can also be studied by means of the SSVEP, which is substantially increased in amplitude when attention is directed to the location of a flickering stimulus. SSVEP recordings have shown that the spotlight of spatial attention can be divided such that two separate regions of a visual display can be attended simultaneously.

Stimulus Discrimination and Classification

Discriminative processing of visual features such as color or shape in the ventral stream of extrastriate cortical areas is reflected by a negative ERP component in the latency range 140–200 ms. This negativity corresponds to the N1 component and is enhanced when stimulus features are actively discriminated and when a known object shape is recognized. Faces, in particular, elicit an occipital-negative ERP configuration at 170–180 ms that originates primarily from the fusiform gyrus area specialized for face processing. These ERPs provide valuable information about the timing of perceptual processes that underlie object recognition.

ERP recordings have shown that pictorial stimuli can be classified into high-level categories (e.g., animals vs. nonanimals) very rapidly, within 150 ms after stimulus onset. In general, the detection of a task-relevant stimulus class is associated with an ERP configuration that includes a negativity peaking at approximately 200 ms (N2 or N200) and a larger late positivity (P3 or P300). This P300 complex is readily elicited in tasks in which subjects detect and respond to infrequent targets that occur randomly in a sequence of nontargets. The P300 that is triggered on detecting such relevant 'oddball' stimuli is time locked to the moment of decision when the subject classifies the stimulus as a target. The P300 latency thus gives a measure of how rapidly stimuli can be recognized and classified as relevant to the assigned task.

The elicitation of the P300 complex depends critically on memory processes since incoming stimuli have to be matched against memory traces to recognize the relevant events. The P300 also depends on the subject's ability to judge the likelihood of upcoming events since its amplitude is highly sensitive to stimulus probability. The P300 also provides an indication of the subject's discriminative capabilities and, for that reason, has been useful for investigating cognitive development in children. To determine whether a child can discriminate between two classes of stimuli, for example, a random sequence of those stimuli can be presented with one being less probable than the other. If the child displays a P300 (or other late endogenous component) that is selectively triggered by the infrequent events, this would constitute unequivocal evidence that the child has made the discrimination. Thus, the presence of a P300 reveals the integrity of a number of cognitive systems, making this ERP useful in the assessment of different kinds of cognitive deficits. For example, a significant prolongation of P3 latency has been observed in cases of Alzheimer's disease and other dementias, presumably reflecting the slowing of memory comparison and recognition processes.

The timing of endogenous ERP components can also be used to study the time course and sequencing of the information processing stages that intervene between stimulus and response. Whereas the N200 and P300 waves primarily reflect perceptual discrimination and classification processes, the preparation for motor response is indexed by a 'lateralized readiness potential' (LRP) that is generated in the motor cortex contralateral to the responding musculature. Studies of these ERPs have demonstrated that stimulus information is passed along from one stage to the next in a continuous 'cascade' such that motor

processes may be activated before perceptual processing is completed. The timing of these components provides a means of tracking the time course at which different stimulus features are analyzed. The LRP has also been used to study the timing of the 'point of no return' – that is, the time at which an intended motor response cannot be stopped.

A specific ERP is elicited when a subject makes an incorrect response in a sensorimotor decision task. This 'error-related negativity' (ERN) is maximal at approximately 100 ms after an erroneous response is initiated and its amplitude grows with the perceived magnitude of the error. The neural generators of the ERN have been localized to the anterior cingulate gyrus and are influenced by the lateral prefrontal cortex. The ERN is proposed to reflect the operation of an error monitoring system that is disrupted by frontal lobe lesions and certain mood disorders.

Memory Processes

Several ERPs have been identified that bear a close relationship to specific forms of memory and retrieval processes. In the auditory modality, short-term sensory memory is reflected in a 'mismatch negativity' (MMN) component that is triggered by deviant stimuli within a repetitive sequence of sounds. The MMN is elicited (at 150–250 ms) by stimuli that deviate in pitch, intensity, timing, localization, or even in more complex acoustic features, and its occurrence signals a mismatch between the incoming stimulus and the sensory memory trace of the immediately preceding stimulus. Because the MMN can be elicited by deviant sounds that are ignored by the subject, this ERP provides evidence that multiple features of auditory inputs are automatically analyzed and stored in short-term memory whether they are actively attended or not. This automatic mismatch detection may constitute an early stage of the orienting response to novel and potentially important changes in the acoustic environment.

Given that the MMN can be elicited with minimal cooperation from the subject, it has become a valuable tool for clinical assessment of auditory processing. Indeed, the elicitation of MMN provides a sensitive index of a person's ability to make auditory discriminations. By using speech sounds as stimuli, for example, MMN experiments have shown that children with language disabilities are deficient in distinguishing critical sounds in their language. Schizophrenic patients have also been found to have reduced or prolonged MMNs, which provides evidence for a deficit in their preattentive auditory processing. In Alzheimer's disease the MMN is also reduced, most prominently at longer interstimulus intervals, which suggests an abnormally rapid decay of auditory sensory memory traces.

A late positive ERP (300–600 ms) has been found to be a reliable predictor of conscious (explicit) memory storage and retrieval. For example, if a subject is attempting to memorize a list of items for subsequent recall, the items that will later be successfully recalled (or recognized) correctly will elicit a larger positivity than words that are later forgotten. This positive ERP has been interpreted as a sign of either the initial encoding of the stimulus in memory or the updating of the memory trace due to its distinctiveness. Other long-latency ERPs are elicited when previously learned material is retrieved from memory. Memory-related ERPs corresponding to these surface-recorded components have been recorded from medial temporal lobe structures including the hippocampus in neurological patients studied with implanted electrodes. These ERPs thus provide a window into the timing of memory storage and retrieval operations in the human temporal lobe.

Language Processing

ERPs are being increasingly studied as online, nonintrusive measures of linguistic processing in the brain. Different classes of ERPs have been identified as indices of lexical, syntactic, and semantic aspects of language and of the hemispheric specialization for language functions. ERPs also reflect the division of the lexicon into open class (content) words such as nouns, verbs, and adjectives and closed class (function) words including prepositions, conjugations, and auxiliaries.

Specific ERP components have been associated with the development of semantic expectancy during reading or listening to a spoken message. The ERPs to semantically anomalous words in otherwise meaningful sentences are marked by an enhanced negativity between 300 and 600 ms after stimulus (the N400). In general, the amplitude of N400 is inversely proportional to the predictability or expectancy of a word presented in a semantic context. Thus, the N400 wave seems to provide a graded index of semantic expectancies and may reflect the amount of prior semantic priming for particular words. The fact that N400 is triggered immediately by each unexpected word provides strong support for linguistic models that postulate immediate, online sentence comprehension rather than the storage of words in a buffer for delayed analysis. The N400 shows a good deal of promise as a means of evaluating the intactness of a patient's semantic networks in a variety of neurological and psychiatric syndromes.

See also: Attentional Mechanisms in Ventral Pathway; Attentional Networks; Attentional Functions in Learning and Memory; Auditory Evoked Potentials; Cognition: An Overview of Neuroimaging Techniques; Electroencephalography (EEG); Electrophysiology: EEG and ERP Analysis; Event-Related Potentials (ERPs); Evoked Potentials: Recording Methods; Evoked Potentials: Clinical; Executive Function and Higher-Order Cognition: Neuroimaging; Executive Function and Higher-Order Cognition: EEG Studies; Language: Cortical Processes; Transcranial Magnetic Stimulation; Visual Attention.

Further Reading

Anllo-Vento L, Schoenfeld MA, and Hillyard SA (2004) Cortical mechanisms of visual attention: Electrophysiological and neuroimaging studies. In: Posner MI (ed.) *Cognitive Neuroscience of Attention*, pp. 180–193. New York: Guilford.

Di Russo F, Martinez A, and Hillyard SA (2003) Source analysis of event-related cortical activity during visuo-spatial attention. *Cerebral Cortex* 13: 486–499.

Giard MH, Fort A, Mouchetant-Rostaing Y, and Pernier J (2000) Neurophysiological mechanisms of auditory selective attention in humans. *Frontiers in Bioscience* 5: 84–94.

Hillyard SA and Kutas M (2002) Event-related potentials and magnetic fields in the human brain. In: Davis KL, Charney D, Coyle JT, and Nemeroff C (eds.) *Neuropsychopharmacology: The Fifth Generation of Progress*, pp. 427–439. Philadelphia: Williams & Wilkins.

Hopfinger JB, Luck SJ, and Hillyard SA (2004) Selective attention: Electrophysiological and neuromagnetic studies. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences III*, pp. 561–574. Boston: MIT Press.

Kutas M and Dale A (1997) Electrical and magnetic readings of mental functions. In: Rugg MD (ed.) *Cognitive Neuroscience*, pp. 197–242. Cambridge: MIT Press.

Luck SJ, Woodman GF, and Vogel EK (2000) Event-related potential studies of attention. *Trends in Cognitive Science* 4: 432–440.

Muente TF, Urbach TP, Duezel E, and Kutas M (2000) Event-related brain potentials in the study of human cognition and neuropsychology. In: Boller F, Grafman J, and Rizzolati G (eds.) *Handbook of Neuropsychology*, 2nd edn., vol. 1, part 2, pp. 1–97. Amsterdam: Elsevier.

Naatanen R (1992) *Attention and Brain Function*. Hillsdale, NJ: Erlbaum.

Rugg MD and Coles MGH (eds.) (1995) *Electrophysiology of Mind*. Oxford: Oxford University Press.

Zani A and Proverbio AM (eds.) (2003) *The Cognitive Electrophysiology of Mind and Brain*. San Diego: Academic Press.

Evoked Potentials: Clinical

B S Oken and T S Phillips, Oregon Health & Science University, Portland, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Evoked potentials (EPs) are neural responses time-locked to some stimulation (e.g., auditory or somatosensory) and can be used for clinical purposes. EPs are often low in amplitude compared to ongoing electrical activity such as electroencephalographic (EEG) and environmental noise responses, which are not time-locked to the stimulation. Thus, averaging of the response to many stimuli is necessary to obtain reliable EPs. EPs can be characterized by waveform latency and amplitude, with the peak and interpeak latencies being the prime determinant of clinically significant changes. The series of stereotypic waveforms following each modality of stimulation are named either as a sequence (e.g., wave I to wave V) or by the average latency seen in controls (P100 being the positive wave at a latency of about 100 ms). Based on lesion studies and invasive recordings in humans, the generators of the various waveforms are known with some accuracy. The waves may be dichotomized into near-field and far-field potentials: near-field potentials require a recording electrode relatively near the generator (e.g., an ear electrode can record a distal eighth nerve potential) and far-field potentials may be recorded with electrodes at some distance from the generator (e.g., a potential generated in cervicomedullary junction can be recorded with a scalp electrode).

EPs are objective physiological markers of clinically relevant abnormalities in sensory or motor systems. EPs may be most helpful when physiological responses assessed by clinical neurological exam are unreliable or not sensitive, such as in infants, comatose patients, or patients under general anesthesia. In addition, evoked potentials may be useful to objectify symptoms that are of uncertain significance (e.g., numbness) or to detect clinically unapparent lesions (e.g., evidence of optic nerve dysfunction in a person with possible multiple sclerosis).

The methodology to obtain EPs requires appropriate signal processing. This necessitates correct sampling rate (more than twice the highest frequency component of the signal, Nyquist criteria), band pass filter settings to help exclude irrelevant neurophysiologic signals or artifact, and adequate resolution of the analog-to-digital converter. The low-pass filter needs to be at least twice the sampling rate to prevent aliasing. For determining EP amplitude, it is helpful to include a prestimulus time period in the average. The number of stimuli in the average depends on the amplitude of the time-locked electrophysiologic response compared to ongoing electrophysiologic or any unwanted activity not time-locked to the stimulus. For analysis of EPs, an artifact rejection methodology, typically based on amplitude criteria, is necessary to prevent irrelevant high-amplitude signals, such as electrical artifact or movement, from entering into the average.

Each modality of EP has one or more reliable waveforms that are used to assess the abnormality of the test result. The waveforms may be categorized as short, middle, or long latency depending on their approximate latencies and generators, with short-latency potentials generated in primary sensory pathways earlier than about 40 ms, extending to long-latency waveforms generated from association cortices later than 100 ms. The waveform latencies (absolute and interpeak), amplitudes, and interside differences obtained from a patient are compared to the range of values for the clinically relevant parameters obtained from a group of healthy controls composed of people of both genders and various ages. In general, the upper limit of normal is 2.5–3.0 standard deviations above the control mean value. Left and right sides are tested separately and interside differences are one criterion for abnormality.

Visual Evoked Potentials

Visual evoked potentials (VEPs) are used to assess the visual conduction pathways from eye to visual cortex. Many types of visual stimulation may elicit a VEP, but some elicit more reliable responses. A pattern-reversal simulation (e.g., a black and white checkerboard pattern, with black and white checks synchronously alternating) is most used clinically. Grid and sinusoidal grating stimuli may also be used but are less commonly available in clinical settings. Simple luminance changes (e.g., flash stimulation) elicit a VEP, but the lack of consistent waveform morphology across individuals has made this technique less clinically useful, compared to pattern-reversal VEPs.

The most consistent waveform of the pattern-reversal VEP is the P100, generated in the visual cortex and recorded over the occipital lobe. The P100 is a positive peak with an average latency in controls of about 100 ms. There are also earlier and later negative waves (N75 and N145) (Figure 1). The recording channel commonly links the occipital electrodes to ear or

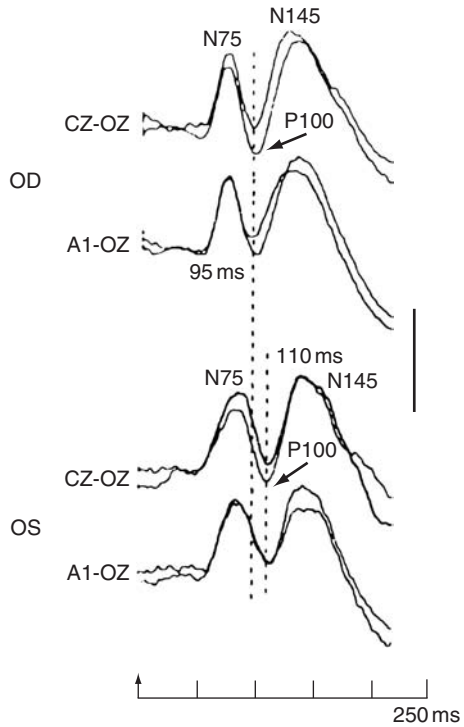

Figure 1 Pattern reversal visual evoked potentials in a patient with multiple sclerosis. The absolute latency of the P100 was normal in both eyes but the interocular latency difference is prolonged – compare right eye (OD, oculus dexter) to left eye (OS, oculus sinister); upper limit for normal was 8 ms – implying a delay in visual conduction pathways anterior to the optic chiasm on the left, although the patient had no visual symptoms. Each of the two trials was an average of 128 stimuli. A1 is left ear, OZ is midoccipital, and CZ is vertex electrode. Electrode locations for scalp electrodes are based on the International 10–20 System. Calibration mark is 10 μV. Adapted from Chiappa KH (ed.) (1997) *Evoked Potentials in Clinical Medicine*, 3rd edn. New York: Raven Press, with permission from Wolters Kluwer.

frontal electrodes. The frontal electrode may help maximize the amplitude since there is a negative waveform generated frontally at the same latency that accompanies P100.

For the standard checkerboard pattern stimulation for VEPs, the visual angle of each check is 10 min to 1 degree, and of the overall screen size for full-field stimulation it is 15–20°. Each eye is tested separately by covering the eye not being tested. To assist in detection of retrochiasmal function, hemifield stimulation of each eye can be performed because a unilateral retrochiasmal lesion is not sufficient to cause abnormalities to full-field monocular stimulation.

Nonpathological Factors Affecting VEPs

The VEP is dependent on the luminance and contrast of the black and white checks. The VEP is also sensitive to inattention and drowsiness. If an individual does not maintain gaze and fixation on the stimulus, the VEP will be attenuated. The P100 is altered by the visual field size of the individual checks and depends on the visual angle of the entire stimulus screen. If a person cannot detect stimuli of the visual angle of the check or grating, there will be no P100.

Pathologic Changes Affecting the VEP

The P100 is altered by certain pathologies along the course of the visual pathway. While not used to evaluate ocular pathologies clinically, the VEP will be altered by significant corneal opacities, dense cataracts, and retinal disease affecting the macula. Most commonly the VEP is used to assess the optic nerve function. Retrochiasmal pathology may be detected only by using hemifield stimulation. Patients with Parkinson's disease, as a group, have prolonged VEPs relative to controls, but the overlap limits clinical use of this observation.

Somatosensory Evoked Potentials

Somatosensory evoked potentials (SEPs) are elicited in response to peripheral nerve stimulation and record the afferent signal at several locations along the pathway from site of stimulation to cerebral cortex. Vibratory, light touch, and mixed-nerve stimulations have been used to investigate dorsal column integrity, and thermal or laser pain stimuli have been used to study the more anterior spinothalamic pathway. Mixed-nerve electrical stimulation is used for most clinical applications.

Upper and lower limb stimulations provide distinct benefits. The peaks following upper limb stimulation have a greater signal-to-noise ratio, allowing for more precise identification of peaks. The peaks following lower limb are generally more sensitive to delays in conduction from demyelination due to the greater length the signal must travel.

Upper Limb SEPs

Stimulating electrodes are placed on the median or ulnar nerve at the wrist. The stimulus pulse intensity is just above motor threshold at 5–10 mA and the pulse duration is 200 μs. At least three recording sites are used: ipsilateral Erb's point (Ep), posterior neck (2nd, 5th, or 7th cervical vertebrae), and contralateral parietal scalp. To identify waves, one averages 500–1000 stimulations delivered at a frequency of 3–5 Hz and the averages are performed at least twice to document waveform reliability.

Within 30 ms of stimulation, several consistent peaks are identifiable on the averaged waveforms. A negativity occurring at 9 ms (N9) after stimulation originates from the brachial plexus and is recorded from the Ep electrode. Subsequent negativity

at 10–11 ms is generated near cervical cord root entry; at 12–14 ms after stimulation, negativity and positivity (N13, P14) are generated from the dorsal column–nucleus cuneatus–medial lemniscus and are thus recorded over cervical vertebrae. A negative peak occurring between 16 and 18 ms (N18) is generated from thalamus and/or thalamocortical radiations. There is a negative peak at 20 ms (N20) and a positive peak at 22–25 ms, generated from the sensory cortex (Figure 2(a)). To calculate central conduction time, the brachial plexus latency (Ep) is subtracted from the sensory cortex negative peak latency (N20).

Lower Limb SEPs

For the lower limb, stimulating electrodes are placed over the tibial nerve at the ankle or over the common peroneal nerve at the head of the fibula. The afferent impulse can be recorded peripherally (at electrodes over the popliteal fossa following posterior tibial

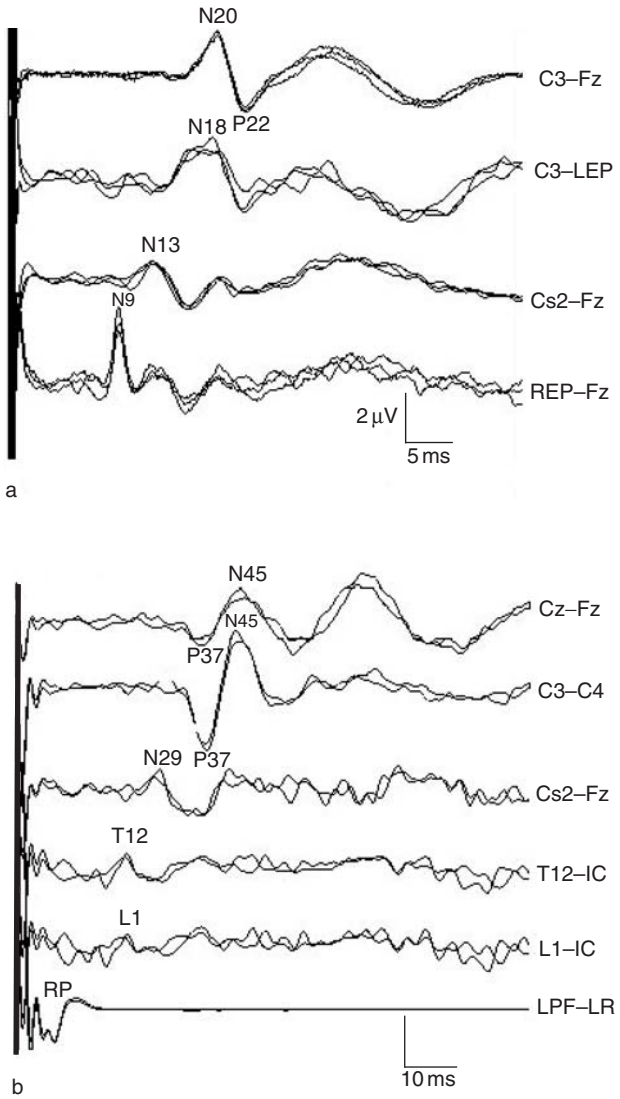

Figure 2 Normal somatosensory evoked potentials of the right median nerve and left tibial nerve. The 1000 averaged stimulations delivered at a frequency of 5.1 Hz are shown with a standard montage; several repetitions are overlaid to illustrate reliability. The vertical bars indicate stimulus onset. (a) Right median nerve stimulation: N9 is apparent at ipsilateral evoked potential and N13 is apparent at the level of cervical vertebra 2. N18, N20, and P22 can be observed at scalp electrodes. (b) Left tibial nerve stimulation: P37 and N45 can be seen maximally over central scalp electrodes Cz and C3. N29 can be seen at the level of cervical vertebra 2. In this tracing, lumbar potentials are apparent at L1 and T12 surface electrodes. Vertical calibration bar represents 0.5 μV for top tracing (Cz–Fz), 1.5 μV for bottom tracing (LPF–LR), and 1 μV for all others. C3, central right scalp; Fz, frontal midline scalp; LEP, left Erb's point; Cs2, cervical vertebra 2; REP, right Erb's point. Cz, central scalp; C4, central left scalp; T12, surface electrode over T12; L1, surface electrode over L1; IC, iliac crest; LPF, left popliteal fossa overlying tibial nerve; LR, left popliteal fossa reference.

nerve stimulation) and from the lumbar plexus or roots (LP, or lumbar potential, from electrodes over the lower back). Potentials generated from the cervicomedullary junction in the region of the nucleus gracilis are recorded from posterior cervical electrodes, and potentials generated from the sensory cortex can be recorded from central scalp electrodes. To calculate central conduction time from lower limb SEPs, latency to LP is subtracted from sensory cortex peak latency. Generally, posterior tibial nerve stimulation is used since it generates amplitude responses larger than are generated by peroneal nerve stimulation.

Following the lumbar potential elicited by posterior tibial nerve stimulation, a series of waves are generated at higher levels within the central nervous system (CNS) (Figure 2(b)). There is an N29 generated in a similar cervicomedullary region as generates the N13 that follows upper limb stimulation, and an N34 is generated in the thalamus analogous to the N18 following upper limb stimulation. About 4 ms after the initial thalamic negativity there is a positive peak (P37) generated in sensory cortex, and this is the most important waveform for assessing clinically significant changes. Following peroneal nerve stimulation at the knee, these latencies occur 5–10 ms earlier.

Nonpathological Factors Affecting SEPs

Body size will influence peak latencies and this effect is greater during lower limb testing; this can be accounted for by normalizing latencies to the height of the individual being tested. Central conduction time is calculated by subtracting latency to Ep (for upper limb) and LP (for lower limb) from the sensory cortex latency, and helps exclude peripheral pathologies as the source of SEP abnormalities. Since core temperature is usually constant, central conduction time also helps to exclude decreased limb temperature as a cause for conduction delays. Age and gender may influence some SEP characteristics, though reported findings have not been consistent. In general, age is associated with decreased conduction velocities.

Pathological Changes in SEPs

Clinical abnormalities manifest as latency prolongation (absolute or interpeak) and peak amplitude reduction. Individuals with multiple sclerosis typically exhibit increased central interpeak latencies. Peripheral demyelinating diseases may also cause increases in these so-called central interpeak latencies because peripheral myelination deficits may exist up to 12 cm from reference point (Ep or LP) to the root entry. Other disorders associated with peripheral and/or central conduction dysfunction, such as Friedrich's ataxia, produce significant SEP changes. Individuals with cortical reflex myoclonus typically have enhanced amplitude of peak components (termed 'giant SEPs'), reflecting hyperexcitability of underlying sensorimotor cortex. Attenuation or absence of cortical waveforms following hypoxic–ischemic encephalopathy or traumatic brain injury signifies a poor prognosis.

The SEP is usually less sensitive to diseases of the peripheral nervous system, compared to electromyelographic (EMG) and nerve conduction studies. Specific questions, such as whether a nerve injury is pre- or postganglionic, may be aided by SEPs. A method developed to localize nerve root pathology is dermatomal stimulation, which uses touch or vibration stimuli of a specific skin region with known sensory innervation. Considerable variability in waveforms exists in a healthy population, limiting the usefulness of this technique for identifying clinical pathologies.

Brain Stem Auditory Evoked Potentials

Brain stem auditory evoked potentials (BAEPs) are a series of waves recorded within 10 ms of an auditory stimulus and that are generated within brain stem auditory structures. The preferred stimulus is a broadband click of 90–120 dB peak sound pressure level although pure tones, or other acoustic signals may be used. All stimuli are delivered monaurally via headphones or intraaural stimulators. Click polarity is usually rarefaction, although condensation clicks are occasionally used to improve wave V definition. Stimulation intensity is 60–70 dB above the person's hearing threshold, with masking of the contralateral ear to prevent bone conduction-elicited responses from the other ear. Given the short window for averaging, stimulus repetition rate can be fairly high (usually around $10\,s^{-1}$ but up to $40\,s^{-1}$).

The waveforms used clinically are called waves I, III, and V (Figure 3). The electrocochleogram occurs prior to wave I. Wave I is thought to be generated in the distal eighth nerve, wave III within the pons in or near the superior olive, and wave V in the inferior midbrain in or near the inferior colliculus. Major criteria for clinical abnormality are delays in waves I–V or other interwave latencies, excessive interside interwave latency differences, and the wave V:wave I amplitude ratio.

Audiologic function may be assessed with BAEPs as well as electrocochleography. Wave V is the last signal to disappear when decreasing the intensity of the acoustic stimulus, and a threshold for disappearance of wave V can be obtained. In addition, wave V latency increases with decreasing stimulus intensity. The latency–intensity plot obtained by stimulating at various intensities may help define the etiology of a hearing deficit.

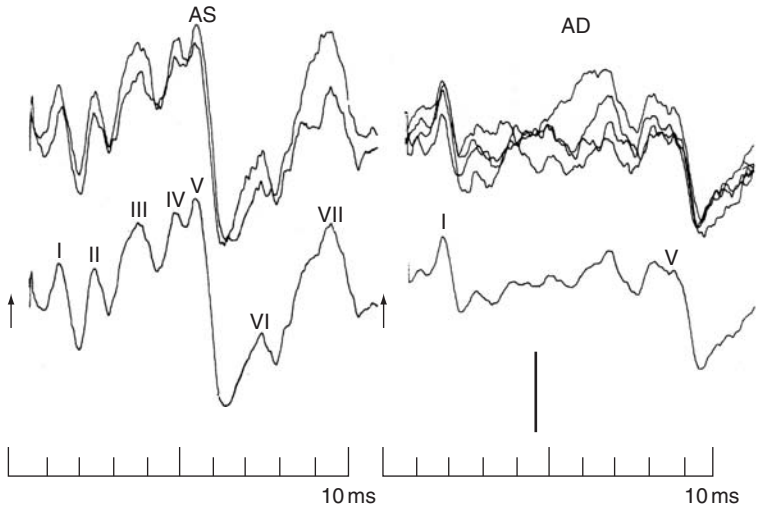

Figure 3 Brain stem auditory evoked potentials from left (AS, auris sinister) and right (AD, auris dexter) ears in a patient with multiple sclerosis, demonstrating a significant wave I–wave V interpeak latency delay on the right (6.7 ms, upper limit of normal 4.7 ms), along with attenuation of wave III, consistent with a defect in auditory conduction pathways between the distal eighth nerve and midpons. This pattern of abnormality is not specific for multiple sclerosis and can be seen with tumors or other lesions in the same region. Each trace is the average of 1024 clicks on the left and 2048 clicks on the right. The single traces at the bottom are the averages of the superimposed trials. Calibration marks are 0.25 μV. Adapted from Chiappa KH (ed.) (1997) *Evoked Potentials in Clinical Medicine*, 3rd edn. New York: Raven Press, with permission from Wolters Kluwer.

Clinically Relevant Changes in BAEPs

BAEPs are affected by age. This is caused in part by presbyacusis, since high-frequency components of the click stimulus are most important for generating the BAEPs. Any peripheral otologic condition decreasing auditory thresholds, such as wax in the ear canal or middle ear pathology, may increase the latency of all of the waveforms. The interwave latencies should remain normal if there is no central pathology.

Pathological Changes in BAEPs

BAEPs are used clinically for several reasons. In multiple sclerosis, BAEPs are less often abnormal than VEPs or SEPs are, so are not commonly performed in this setting. BAEPs may be used to help prognosticate coma following closed-head injuries and stroke. BAEPs are sensitive within brain stem structures to lesions such as acoustic neuromas or brain stem gliomas, but MRI scans are more helpful clinically. BAEPs can be used to help obtain auditory thresholds in patients who cannot easily cooperate for audiometry, such as infants and comatose patients. BAEPs may be used in screening for audiologic deficits secondary to ototoxic drugs.

Event-Related Potentials (P3)

Long-latency EPs are generated in higher order sensory processing areas or in heteromodal association areas. The component, called either the P3 (third positive component) or the P300 (positive component with a mean latency of about 300 ms in controls), is the only long-latency EP that has been used clinically.

The P3 is a symmetric positive wave, recorded maximally over the midline central and parietal regions, with a latency that varies from 250 to 600 ms. The P3 can be elicited with a stimulus of any modality. The most commonly used method to obtain the P3 is the 'oddball' paradigm. This involves the presentation of unexpected or infrequent stimuli randomly interspersed among more frequent stimuli. In most clinical studies, the unexpected stimuli differ from the more common stimuli in terms of frequency (e.g., pitch) or intensity. A method that augments the P3 and is generally used in conjunction with the oddball paradigm consists of attending to task-relevant target stimuli and ignoring nontarget stimuli. A P3 will be seen following the target but not the nontarget stimuli (Figure 4).

The generator site of the P3 is not known with certainty. Neuronal activity in multiple brain regions, including inferior parietal lobule, frontal lobe, hippocampus, medial temporal lobe along with other limbic structures, and locus coeruleus, correlate to scalp-recorded P3 activity. Patients with frontal lobe lesions have a significantly attenuated novelty P3 (P3a) while patients with temporoparietal junction lesions have markedly attenuated P3s to both novel and attended stimuli (P3a and P3b).

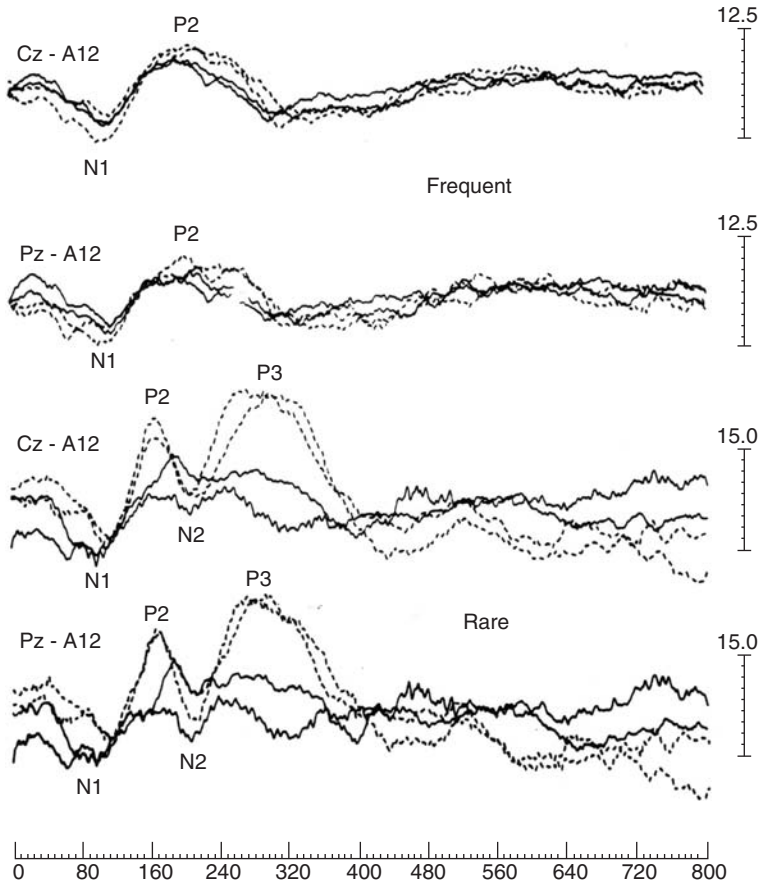

Figure 4 Auditory evoked potentials recorded from the central midline (Cz) and parietal midline (Pz) scalp following 1000-Hz tones occurring frequently (84%), interspersed with 2000 Hz tones occurring infrequently. The N1–P2 complex is relatively the same in response to frequent and infrequent tones and is not markedly affected by attending to the infrequent tones. Solid lines are in the unattended condition and dotted lines are in the attended condition, whereby participants are keeping a mental count of the rare tones. There is a marked enhancement of the N2–P3 in the attended condition with the rare tones.

At least three channels of EEG activity are used (Fz, Cz, Pz), in part to delineate the P3a and P3b components. The electrodes can be referenced to linked-ear, linked-mastoid electrodes, nose, or noncephalic electrodes. Eye movement also needs to be monitored because eye movements may be time-locked to the stimuli, especially the target (oddball) stimuli, and the field of distribution of the eye movements may include the recording electrode sites. The stimulus rate is generally about $1\,s^{-1}$. The sweep duration should be 600–1000 ms and a prestimulus period is important in assessing the baseline for amplitude measures. The P3 is a variable response, even within an individual, and, as with other EPs, it is necessary to repeat trials to ensure reproducibility.

Nonpathologic Factors Affecting the P3

The P3 is not as sensitive to physical aspects of the stimulus, such as intensity, as are shorter latency components. In fact, a very similar P3 may be elicited by stimuli of different sensory modalities. While the short-latency EPs can be recorded when the patient is under general anesthesia, the optimal P3 is recorded when the patient is alert. The two factors, stimulus infrequency or unexpectability and attention or task relevance, operate independently. In fact, they produce different P3s, with a novelty-related P3a that occurs slightly earlier and with a more frontal distribution than is seen for the parietal maximum P3b component wave, which is best elicited by attending to a task-relevant stimulus. In general, target stimulus infrequency and attention increase the amplitude of the P3, and task difficulty increases its latency.

A series of related variables affect the P3: level of attention, vigilance, and response accuracy. Drowsiness or inattention will decrease the amplitude or obliterate the P3. An assessment of behavioral response to the stimuli is important to enable estimation of the individual's vigilance or degree of attention. The P3 component does not depend on the presence of a

motor response. In fact, the P3 latency may be longer than the reaction time. The P3 latency has a positive correlation with age, with an increase in mean latency by approximately 1–1.5 ms year^{-1} after age 20 years. There is also a gradual decline of P3 latency in children, with a minimum latency being achieved between ages 15 and 25 years.

Pathologic Factors Affecting the P3

The P3 may be altered by many diseases that affect cognitive function. In Alzheimer's disease (AD), the sensitivity of a prolonged P3 latency has ranged from 10 to 80%. The sensitivity for classifying a single patient with mild AD as having an abnormal P3 latency is likely well under 50%, despite the consistently observed group difference in controls and patients with AD. Other diseases associated with cognitive change or dementia also produce changes in the P3, including vascular disease, Huntington's disease, Parkinson's disease, multiple sclerosis, human immunodeficiency virus infection, and Down's syndrome. Schizophrenia patients have a lower amplitude P3 compared to controls. Some evidence shows that patients with alcoholism or even individuals with family history of alcoholism have altered P3s when compared with groups of controls. Additionally, many CNS depressant drugs will prolong the P3 latency. Finally, the P3 following complex stimuli has been used to explore various neurobehavioral syndromes, including prosopagnosia, inattention following prefrontal lesions, visual neglect, anosognosia, blindsight, transient global amnesia, and other memory disorders.

Motor Evoked Potentials

Motor evoked potentials (MEPs) can be used to assess the integrity of the descending corticospinal tract. An electrical current is delivered or induced in the motor cortex and electromyographic recording is obtained using surface electrodes. MEPs were originally delivered using transcranial electrical stimulation (TES); however, the current attenuation by the skull is great, such that the intensity required using scalp electrodes to activate underlying neurons also activates many pain fibers and scalp muscles. The development of transcranial magnetic stimulation (TMS), which induces an electrical current orthogonal to the magnetic field of the stimulating coil, has allowed the study of MEPs without the direct application of high-intensity current.

Cortical stimulation may initially elicit a D (direct) wave, believed to represent direct activation of corticospinal axons. This is seen primarily with TES. An I (indirect) wave occurs 2 ms later, presumably through depolarization of motor cortex interneurons which synapse onto and depolarize cell bodies of the corticospinal tract. This is likely the primary descending volley from TMS, but it is also seen with TES. Both D and I waves travel down the spinal cord to synapse onto an alpha motor neuron and effectively result in a muscle twitch.

Methodology

For TMS, a magnetic stimulator is positioned over the vertex of the scalp for upper limb and lower limb MEPs, and direction of current flow through the coil can be manipulated to preferentially stimulate the right or left hemisphere. Upper or lower limb MEPs can be tested by recording EMGs of several muscles. Resting motor threshold (RMT) intensity is defined as an intensity at which 50% of trials obtain an EMG response of at least 50 μV while the muscle is relaxed. Test pulses are subsequently delivered as single pulses of electromagnetic energy at an intensity of 10% above RMT. Unlike SEPs that are on the order of 1 μV, MEPs evoke an EMG response on the order of 0.1–2 mV. As a result, trials do not need to be averaged to observe a reliable waveform, and two to three stimulations are generally sufficient to demonstrate reliability. TES is used only to elicit MEPs while someone is under general anesthesia and is used only for intraoperative neuromonitoring (see later, 'Intraoperative neurophysiologic monitoring').

Following scalp stimulation, a muscle twitch will be observed within 20–45 ms, depending on distance to recording site. For clinical purposes, response latency is the primary variable of interest, as amplitude of response will vary greatly between individuals and between sessions. Peripheral conduction time can be measured by stimulating at the root exit (upper limb, midline over cervical vertebrae; lower limb, midline over lumbar vertebrae) and measuring the latency to EMG response (Figure 5). Central conduction time can subsequently be calculated by subtracting the latency following root exit area stimulation, from the latency following scalp stimulation.

Nonpathological Changes in MEPs

As with SEPs, height will affect MEP latencies. Greater skull thickness, often associated with larger head sizes, will increase RMT. Both RMT and MEP amplitude will vary as cortical excitability of neurons underlying the stimulation site fluctuate over time from uncontrolled factors such as arousal. Pharmacological Na^+ and Ca^{2+} channel blockers will reduce RMT. MEP facilitation will be observed if the individual sustains a mild contraction of the muscle, enhancing cortical excitability (Figure 5). The latter

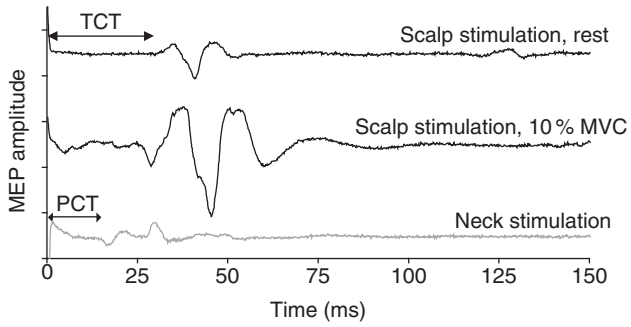

Figure 5 Single motor evoked potentials (MEPs) of extensor carpi radialis of a healthy individual. Single-pulse transcranial magnetic stimulation was applied over contralateral scalp; stimulus onset occurs at time zero. The amplitude of the MEP induced by stimulation at the scalp is greatly enhanced by a mild extensor carpi radialis contraction – 10 % of maximum voluntary contraction (MVC) – compared to rest, and the latency is slightly decreased. MEP latency after stimulation at the neck represents peripheral conduction time (PCT). Central conduction time can be calculated by subtracting PCT from total conduction time (TCT). Each tick mark on vertical axis represent 0.2 mV.

can be clinically useful in obtaining an MEP from an individual nonresponsive to high-intensity stimulation.

Pathological Changes in MEPs

MEP response amplitude and RMT are too variable to be used for clinical assessment. Response latency is the only current reliable measurement, and increased latencies are commonly observed in individuals with ischemic stroke, multiple sclerosis, upper motor neuron diseases, and spinal cord lesions. MEPs, SEPs and VEPs may have similar sensitivity for detection of dysfunction in multiple sclerosis . Individuals with epileptic syndromes have reduced RMT and larger amplitude MEPs.

Other Uses of TMS

Cortical circuitry can be indirectly studied using a paired-pulse TMS protocol. By varying prepulse stimulus intensity and interstimulus interval, intracortical inhibition or facilitation can be measured in motor disorders such as Parkinson's disease. The post-EMG silent period studied with TMS is also reduced in Parkinson's disease.

TMS has been used to probe changes in cortical excitability as well as shifts or expansion in motor representation areas for a muscle group after learning a new motor skill or practicing a movement. These plastic changes are observed as decreases in RMT or a spatial change in areas excited by a given stimulation intensity. Repetitive TMS has also been explored as a treatment for depression as well as other neurological disorders.

Premotor Potentials

Premotor potentials are evident in the averaged EEG waveform prior to a planned motor movement. For premotor potentials, averaging of trials is done by 'backaveraging' from the onset of EMG activity related to the movement. The bereitschaftspotential (BP) is one type of premotor potential, and it occurs prior to a voluntary movement. The BP is a slow negative shift composed of an early BP, beginning 0.8–2 s before the onset of EMG, and a later BP or negative slope (NS′) component, apparent 400 ms before the onset of EMG (Figure 6(a)). The BP is maximal over the contralateral central scalp for upper limb movements (Figure 6(b)) and midline central scalp for lower limb movements and reflects preparation for the upcoming motor movement. Unique premotor potentials, unlike the BP, can be recorded prior to pathologic unplanned movements such as myoclonus.

Generators

Dipole source analyses from EEG and MEG recordings, as well as recordings from subdural electrodes placed during neurosurgical procedures, have led several researchers to identify bilateral supplementary motor area (SMA), lateral premotor, and primary motor cortices as putative generators of the BP, with greater localization in contralateral M1 and premotor cortex 400 ms prior to movement onset in the late BP (NS′).

Nonpathological Changes in Premotor Potentials

The amplitude and duration of the early component of the BP are mediated by movement speed, movement precision, muscle force, intention, and movement selection, among other individual and task characteristics; thus, it is necessary to employ a standardized task and instruction set when assessing BPs in research settings. For this reason, BPs are not commonly used clinically to assess pathological abnormalities.

Pathological Changes in Premotor Potentials

Premotor potentials may be examined to help determine the cause of myoclonus. When premotor potentials are recorded in individuals with cortical myoclonus, a unique premotor potential is observed within 50 ms prior to the onset of the involuntary

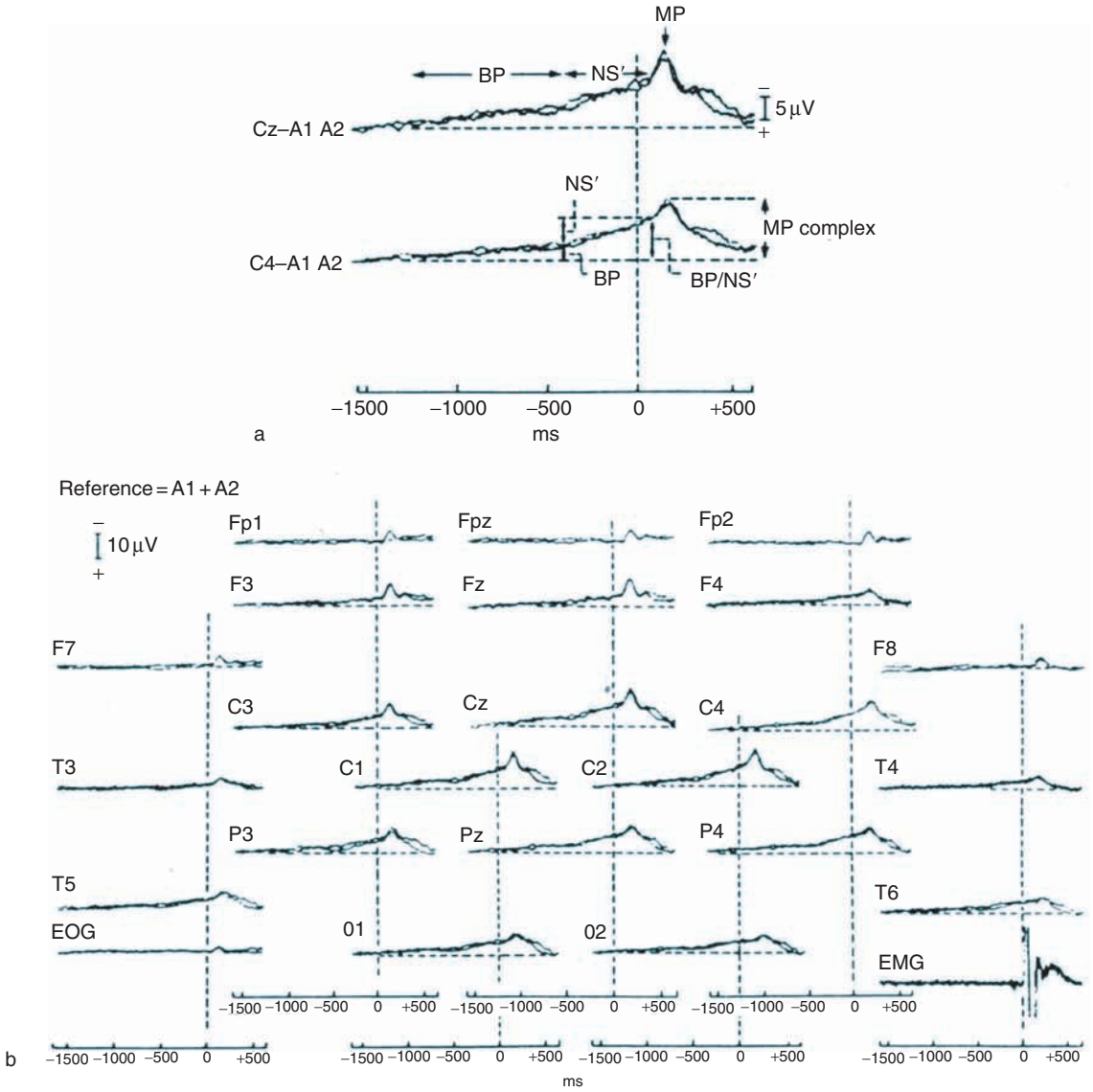

Figure 6 Terminology of premotor potential components and scalp topography. (a) Normal tracings recorded prior to self-paced middle finger extensions are shown over central midline (Cz, top tracing) and contralateral (C4, bottom tracing) scalp areas. The early bereitschaftspotential (BP) component occurs here at about 1200 ms prior to onset of movement. About 400 ms before movement, the slope of the potential steepens; this is the negative slope (NS′). Following movement, the motor potential (MP) is identified by a further increase in slope to the peak. (b) The scalp distribution of premotor potentials for the same task as in panel a. Amplitude is maximal over central midline (Cz) and contralateral (C4) sites. Two tracings of 100–250 averaged trials are shown for each recording site. F, frontal; Fp, frontopolar; z, midline placement; C, contralateral; Cz, central midline; T, temporal; P, parietal; O, occipital; EOG, electrooculographic; EMG, electromyelographic. Reproduced from Neshige R, Luders H, and Shibasaki H (1998) Recording of movement-related potentials from scalp and cortex in man. *Brain* 111: 719–736, with permission.

contraction. This potential appears over the contralateral scalp in the form of a biphasic wave at electrodes overlying sensorimotor cortex of the contracting muscle(s) and is obtained by averaging trials backward from the onset of EMG activity of the first contracting muscle (termed 'jerk-locked back-averaging'). Subcortical forms of myoclonus often lack this wave. Patients with psychogenic causes of myoclonus may demonstrate a 1–2 s duration BP, as is seen with planned movements.

Studies of patients with Parkinson's disease have reported small early BP. However, the later NS′ component is intact. Cerebellar lesions have been associated with reduced amplitude of the early BP as well as reduced NS′ amplitude over the hemisphere contralateral to the lesion. Focal dystonia has been associated with preserved early BP amplitude but a reduction in the NS′ component. Following ischemic stroke, there is a reduced late BP over the lesioned hemisphere contralateral to movement and an

enhanced early BP amplitude over the hemisphere ipsilateral to movement, suggesting a compensatory activation.

Intraoperative Neurophysiologic Monitoring

Intraoperative neurophysiologic monitoring (IONM) is used to give information in real time to surgeons operating in or near relevant neural structures. IONM is used for several general purposes. The information may consist of ongoing assessment of integrity of signal conduction. For example, acute posterior tibial-elicited cortical SEP changes during distraction in scoliosis surgery implies a posterior column dysfunction. Following acute changes in EPs secondary to CNS perfusion while under general anesthesia, the surgical team has about 10–20 min to reverse the underlying etiology until the pathology likely becomes irreversible. EPs may also be used for localization purposes – for example, localizing motor cortex by recording SEP field potentials from a subdural electrode or from direct cortical stimulation and recording EMG.

The EP modalities used for IONM include SEP (upper and lower limb), BAEP, and MEP. Due to the marked attenuation of TMS EPs by general anesthesia, intraoperative motor EPs are performed using TES. To elicit these MEPs, brief pulse trains (4–7 pulses at 200–500 Hz) are applied to central scalp regions overlying motor cortex. VEPs using goggles have not been reliable intraoperatively. In addition to EPs, an EMG from relevant muscles is often recorded if cranial nerves or nerve roots are at risk. Nerve conductions from pedicle screw stimulation or direct nerve stimulation may be used. An EEG may be used to help determine and guide the depth of anesthesia (e.g., until suppression–burst pattern is observed) or determine focal or global cerebral dysfunction during cerebrovascular or cardiovascular surgery.

EPs during surgery are sensitive to level of anesthesia, with MEPs being potentially obliterated by commonly used levels of inhalant anesthetics. Intravenous anesthetics such as propofol and narcotics have much less of an effect on EPs than do inhalants. BAEPs are relatively insensitive to anesthesia, with SEPs being in between MEPs and BAEPs. EPs are also sensitive to other physiologic parameters during surgery, including blood pressure and temperature. EP recording within the operating room requires much expertise in managing the numerous causes of artifacts potentially obliterating the signals (e.g., electrocautery).

While IONM has become standard for many neurosurgical, orthopedic, vascular, and otolaryngology surgeries, there are no randomized trials demonstrating their utility at improving outcomes. There have been several studies carefully using historical or nonmonitored controls, with results strongly suggesting their utility for scoliosis surgery, resection of small acoustic neuromas, and microvascular decompression for trigeminal neuralgia. IONM is commonly performed in other surgeries when there is significant risk to neural structures (e.g., multilevel spine surgery with preexisting myelopathy, posterior fossa surgery, and aneurysm surgery).

See also: Auditory Evoked Potentials; Electroencephalography (EEG); Electromyography (EMG) and Nerve Conduction Studies; Electrophysiology: EEG and ERP Analysis; Event-Related Potentials (ERPs) and Cognitive Processing; Event-Related Potentials (ERPs); Evoked Potentials: Recording Methods.

Further Reading

American Clinical Neurophysiology Society (2006) Guidelines on evoked potentials. *Journal of Clinical Neurophysiology* 23: 125–137.

American Clinical Neurophysiology Society (2006) Guidelines on short-latency auditory evoked potentials. *Journal of Clinical Neurophysiology* 23: 157–167.

American Clinical Neurophysiology Society (2006) Guidelines on short-latency somatosensory evoked potentials. *Journal of Clinical Neurophysiology* 23: 168–179.

American Clinical Neurophysiology Society (2006) Guidelines on visual evoked potentials. *Journal of Clinical Neurophysiology* 23: 138–156.

Chiappa KH (ed.) (1997) *Evoked Potentials in Clinical Medicine*, 3rd edn. New York: Raven Press.

Deletis V and Shils JL (2002) Neurophysiology in Neurosurgery. New York: Academic Press.

Fitzgerald PB, Fountain S, and Daskalakis ZJ (2006) A comprehensive review of the effects of TMS on motor cortical excitability and inhibition. *Clinical Neurophysiology* 117: 2584–2596.

Neshige H, Luders H, and Shibasaki H (1988) Recording of movement-related potentials from scalp and cortex in man. *Brain* 111: 719–736.

Polich J (2004) Clinical application of the P300 event-related brain potential. *Physical Medicine and Rehabilitation Clinics of North America* 15: 133–161.

Shibasaki H and Hallett M (2005) Electrophysiological studies of myoclonus. *Muscle & Nerve* 31: 157–174.

Shibasaki H and Hallett M (2006) What is the bereitschaftspotential? *Clinical Neurophysiology* 117: 2341–2356.

Sohn YH and Hallett M (2004) Motor evoked potentials. *Physical Medicine and Rehabilitation Clinics of North America* 15: 117–131.

Evoked Potentials: Recording Methods

D Regan and M P Regan, York University, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When a human receives a sensory (e.g., visual, auditory, tactile) stimulus, neurons in the corresponding sensory cortical area generate slow postsynaptic potentials, and the resulting volume currents can reach the scalp, creating potential differences (voltages) between different scalp locations. This evoked potential (EP) can be obtained by recording the electroencephalogram (EEG) from electrodes attached to the scalp and, by using prior knowledge of the EP (the signal) it can be extracted from the electrical activity of the rest of the brain (the noise). EP amplitudes range from about 0.1 to 5 μV and are buried in noise of typically 50 to 100 μV amplitude. A major limitation of the EP technique is that distant as well as nearby regions of brain tissue can contribute to the EP at any given scalp location and, because the scalp EP is affected by the conductivities of intervening tissues, sources are difficult to localize.

In addition to the electrical field, brain activity also generates a magnetic field that can be recorded outside the skull (the magnetoencephalogram, or MEG). This neuromagnetic field seems to be generated by intracellular currents rather than by the volume currents that are largely responsible for the scalp EP. Magnetic response recording has the advantage that the neuromagnetic field is unaffected by tissues between the source and the detector, because brain tissues are transparent to low-frequency magnetic fields. In addition, no reference is involved, and the evoked response near the scalp is determined by a smaller volume of brain tissue than the EP is, being little affected by distant sources. To record a neuromagnetic field, a superconducting quantum interference detector (SQUID) is placed near the head. This device converts the magnetic field into an electrical signal. Compared with imaging techniques such as functional magnetic resonance imaging (fMRI), both techniques have the advantage of millisecond resolution.

Analysis in the Time Domain

Time-Domain Averaging of Transient Evoked Responses

Signal averaging in the time domain enhances signal-to-noise ratio by exploiting prior knowledge of the instant(s) at which stimuli are applied (on the assumption that the response always occurs a fixed time after the stimulus). Figure 1 illustrates how signal averaging can enhance signal-to-noise ratio as progressively more responses are averaged. As averaging progresses, the signal-to-noise ratio increases with $\sqrt{N}$ (where N is the number of responses averaged), providing that the following requirements are satisfied: (1) the waveform to be averaged is the sum of two independent waveforms, namely the signal and the noise; (2) the signal waveform is produced by a process that is stationary from trial to trial, and its variance is negligible; (3) the noise waveform is produced by a stationary random process; and (4) the N sample of noise are uncorrelated. If these requirements are not fulfilled the signal-to-noise improvement can be much less than $\sqrt{N}$.

Finally, only if the sensory system always returns to its resting state before the next stimulus can the averaged response be regarded as the averaged transient response of the sensory system stimulated. This requirement is commonly violated.

Quantification of the Transient Evoked Response Waveform

A typical transient response consists of a series of positive and negative deflections. For example, Figure 2 shows the transient responses to the abrupt appearance and the abrupt disappearance of a checkerboard pattern. To an 'eyeball' inspection, the two responses have clearly different waveforms. But the quantification of such a difference is controversial.

A common assumption is that the waveform can be regarded as the linear sum of several component waveforms, and that each component corresponds to a functional entity in the brain. Peak analysis is a popular way of reducing the data in transient response so that the waveform can be specified in terms of a few numbers. A peak is defined as either the most positive or the most negative voltage within some specified time interval. Although this approach has intuitive appeal, peak measurements are strongly affected by noise and have dubious significance with complex waveforms that have overlapping components. An alternative approach that is less affected by noise is to measure the areas under peaks. A third possibility is principal components analysis (PCA). Whether this last approach necessarily provides the notional

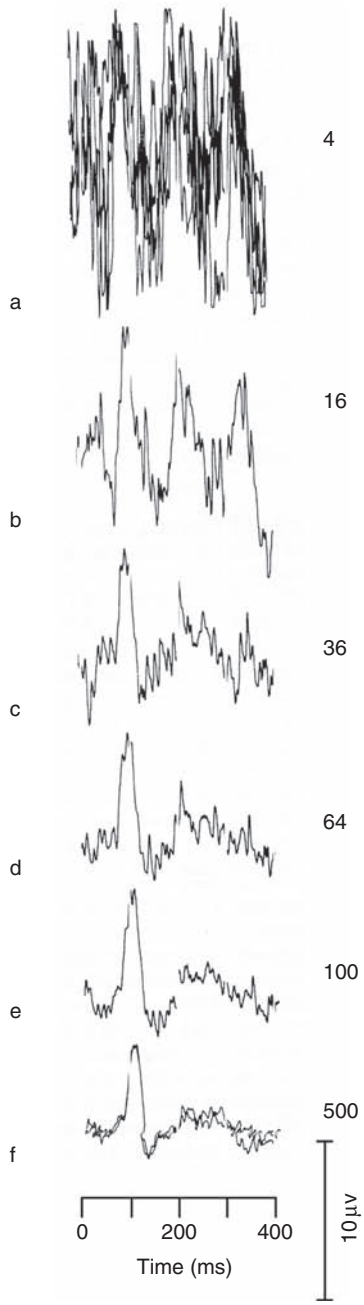

Figure 1 Signal-to-noise enhancement by time-domain signal averaging. A subject viewed a pattern of black and white checks that abruptly exchanged places (reversed) 1.88 times per second. Each reversal was a single stimulus. (a) Each trace is the average of 4 single-stimulus responses. Only noise (alpha activity in this case) is evident. (b–e) Average of successively more responses (16, 36, 64, 100). (f) Two separate averages of 500 responses superimposed to show the excellent reproducibility of the averaged response. Reproduced from Regan D (1989) *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*, p. 53. New York: Elsevier, with permission.

component waveforms that underlie the transient response waveform is controversial, though it has been stated that it does provide an estimate of the dimensionality of the data space.

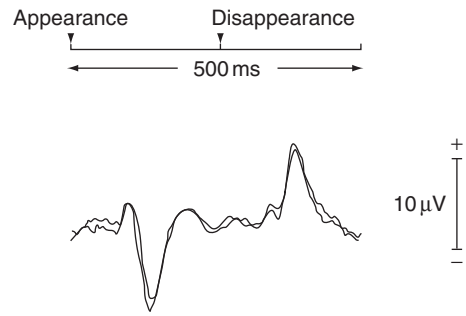

Figure 2 Transient EP to the abrupt onset and the abrupt offset of a pattern of fine checks viewed binocularly. Mean luminance was constant. Note that though the positive-upward convention is used in basic research, a negative-upward convention is common in clinical work. Reproduced from Spekreijse H, van der Tweel LH, and Regan D (1972) Interocular sustained suppression: Correlations with evoked potential amplitude and distribution. *Vision Research* 12: 521–526, with permission.

Analysis in the Frequency Domain

Fourier Analysis of the Steady-State Evoked Response

It is sometimes said that the distinction between transient and steady-state responses is merely a matter of stimulus repetition frequency. This is incorrect: in principle, steady-state responses can be evoked by a very long train of identical stimuli, whether the stimulus repetition frequency is low or high. The real distinction is that the system attains a dynamic steady state throughout the duration of the recording without ever returning to its resting state, while, as already mentioned, for an averaged waveform to be a true transient response the system must always return to its resting state before the next stimulus is applied. The dynamic steady state may not be attained immediately at the onset of stimulation, but may take several seconds. This should be checked, as should the absence of any progressive change in amplitude or phase due, for example, to sensory adaptation.

The idealized steady-state response to a very long train of identical stimuli is an electrical or magnetic waveform, the discrete frequency components of which remain constant in amplitude and phase over an infinitely long time period. Although this definition does imply that the idealized steady-state response is an infinitely long train of identical waveforms, it is more insightful to think of it in terms of discrete frequency components, because different components may have different properties. For example, the steady-state EP evoked by sinusoidal flicker with frequency F of between about 15 and 20 Hz consists predominantly of a fundamental component of frequency F Hz plus a second harmonic component of frequency $2F$ Hz; the properties of the F Hz component are similar to the

properties of parvocellular neurons in the macaque monkey lateral geniculate body, while the very different properties of the $2F$ Hz component are similar to the properties of magnocellular neurons.

Steady-state response can be more unequivocally quantified than can transient responses, because they are compared of discrete frequency components that are quite separate from each other (nonoverlapping), and each is fully described by an amplitude and a phase. In some applications, frequency domain and hybrid techniques (e.g., the sweep technique) are much faster than is time domain averaging. This is because a discrete steady-state frequency component contains considerably less information than does a typical transient response.

A method for simultaneously demonstrating the existence of a steady-state response, enhancing its signal-to-noise ratio, and quantifying the response is as follows. While a subject receives repetitive, commonly sinusoidal, stimulation, the EEG (or MEG or electroretinogram) is multiplied by a sine and by a cosine, the frequency of which is some exact multiple of the F Hz stimulus repetition frequency. Multiplication by sine/cosine pairs of frequency F, $2F$, $3F$, etc. extracts the first, second, third, etc. harmonic of the response. The products of the two multiplications are each low-pass filtered to give two running averages. Figure 3 shows how the first harmonic of the steady-state flicker EP was first demonstrated and quantified. Beyond about 15 s after the onset of stimulation in Figure 3 (this delay being largely caused by the low-pass filters used to create a running average), the running average of the sine and cosine products remained approximately constant, thus demonstrating the existence of a steady-state response. Next, the mean values of θ_1 and θ_2 were calculated and the amplitude (S) and phase (ϕ) of the first harmonic were calculated as shown.

In the case of steady-state responses, signal-to-noise ratio is enhanced by exploiting prior knowledge of the exact frequencies of the discrete response components. For example, when the stimulus is sinusoidal, the response component frequencies are multiples of the frequency of the sinusoid. Each sine/cosine multiplication acts as a narrow bandwidth

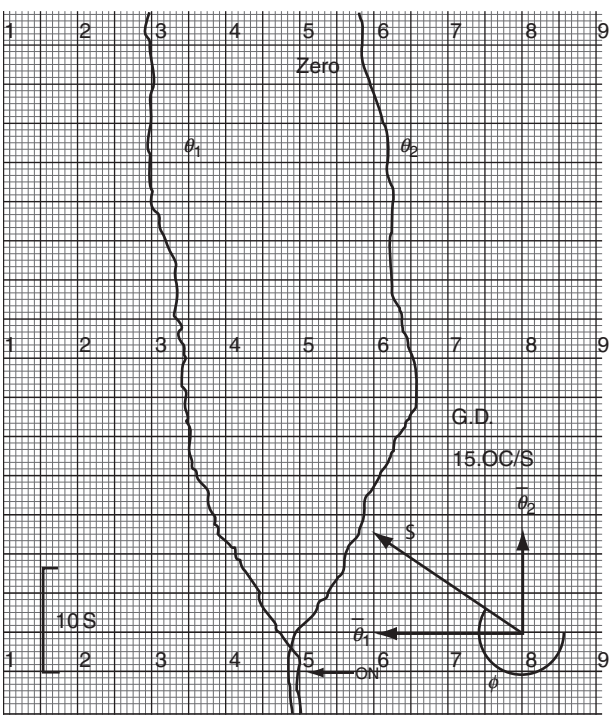

Figure 3 Steady-state EP. The two traces are the running averages of the sine and cosine multiplier outputs, the instantaneous amplitudes of which (measured horizontally from the central zero) are marked θ_1 and θ_2, respectively. Time runs vertically, and the onset of sine wave flicker stimulation is denoted by an arrow. The insert shows how the amplitude (S) and phase (ϕ) of the first harmonic frequency component of the steady-state EP were obtained using Pythagoras' theorem. From Regan D (1989) *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*, p. 83. New York: Elsevier.

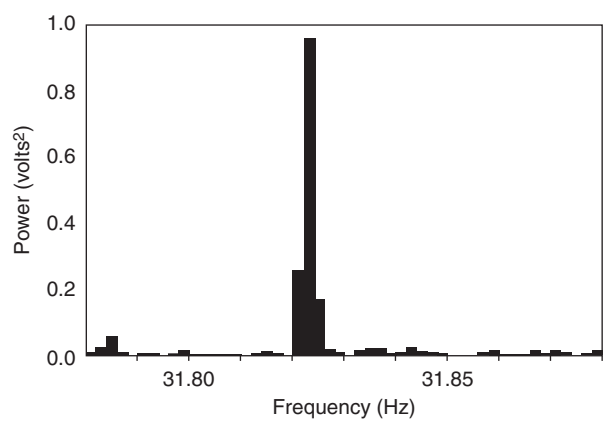

Figure 4 Two-thousandths of a hertz bandwidth. At a frequency resolution of 0.0019 Hz, approximately two-thirds of the EP signal power is contained within a single bin. Therefore the EP signal's bandwidth is no more than 0.0019 Hz. Note that even if the EP signal's bandwidth were considerably less than 0.0019 Hz, more than one bin of the analyzer would contain power when the analyzer's resolution is set at 0.0019 Hz, because the sensitivity of adjacent bins overlaps. The stimulus was a vertical sine wave grating that was counterphase-modulated at nominally $F = 8$ Hz. The $4F$ component is shown. Zoom FFT was applied to the entire length of EEG samples of 520-s duration. To allow some frequency-domain averaging (with overlap), recording duration was 640 s. From Regan D (1989) *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*, p. 95. New York: Elsevier.

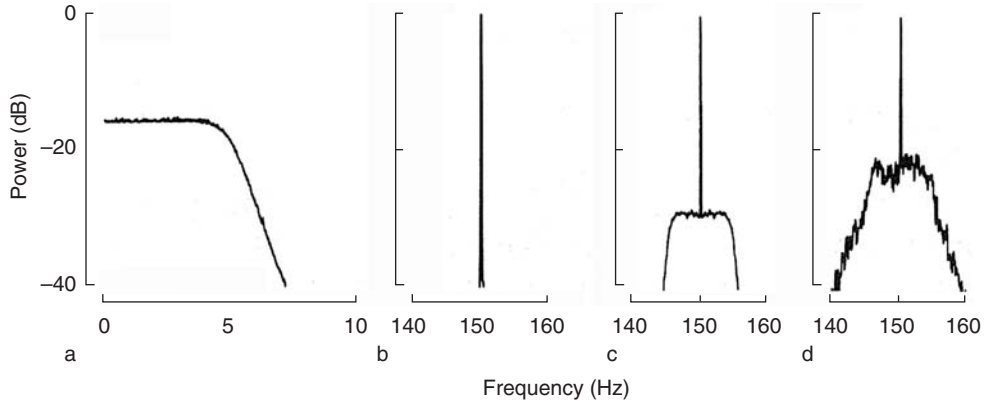

Figure 5 (a) Modulating noise waveform, direct current to 5 Hz. (b) Pure sine wave, 150 Hz. (c) 150 Hz sine wave, amplitude-modulated by noise waveform of amplitude 2.5 V RMS. (d) 150 Hz sine wave, phase-modulated by a noise waveform of amplitude 2.0 V RMS. A sine wave-modulating waveform of 1.0 V amplitude produces an amplitude modulation depth of 20% and a phase excursion of 30°. Frequency resolution was 0.0078 Hz in all cases. From Regan D (1989) *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*, p. 95. New York: Elsevier.

filter centered on the analysis frequency. If the recording duration is T seconds, the filter bandwidth is the theoretical limit of resolution – that is, T^{-1} Hz.

This last point can perhaps be more clearly understood as follows. A recording technique that uses the digital Fourier transform (fast Fourier transform (FFT) or zoom FFT) to obtain a spectrum with T^{-1} Hz resolution is in many respects more convenient than that shown in Figure 3. (Though this technique has the disadvantage that it does not immediately confirm that the amplitude and phase of a response component are stable.) In Figure 4 a subject viewed a sine wave grating that was counterphase-modulated at nominally 8 Hz. Zoom FFT was applied to the entire length of EEG samples of 520 s duration. To allow some frequency-domain averaging, the recording duration T was 640 s. Approximately two-thirds of the signal power was contained within one bin of width 0.0019 Hz. One might suppose that progressively reducing the T^{-1} Hz resolution would improve signal-to-noise ratio only up to some limit at which the broadening of a response component's bandwidth would render ineffectual any further reduction of T^{-1}. However, this is not entirely correct. Providing that the moment-to-moment variability of the amplitude and phase of the response component is not too great, the bandwidth of the physiological signal is determined entirely by the recording duration T. Figure 5 illustrates this point. Panels (b) and (c), respectively, show the spectrum of a 150 Hz sinusoid and of the same sinusoid subjected to noise (direct current to 5 Hz, panel (a)) modulation of amplitude. Panel (d) shows the spectrum of the sinusoid subjected to phase modulation. Frequency resolution was 0.0078 Hz. Panel (b) shows that the signal bandwidth

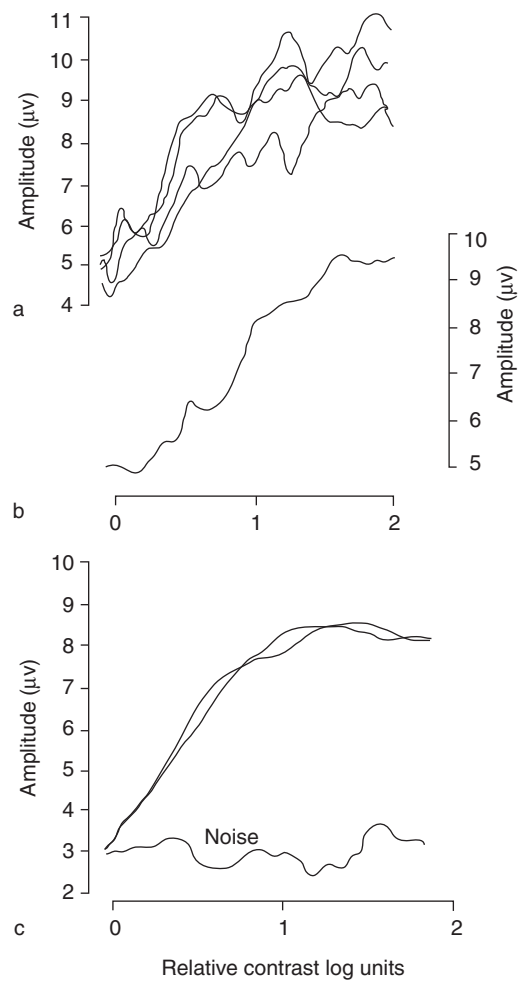

Figure 6 The sweep technique: an example. See text for details. Reproduced from Regan D (1975) Color coding of pattern responses in man investigated by evoked potential feedback and direct plot techniques. *Vision Research* 15: 175–183, with permission.

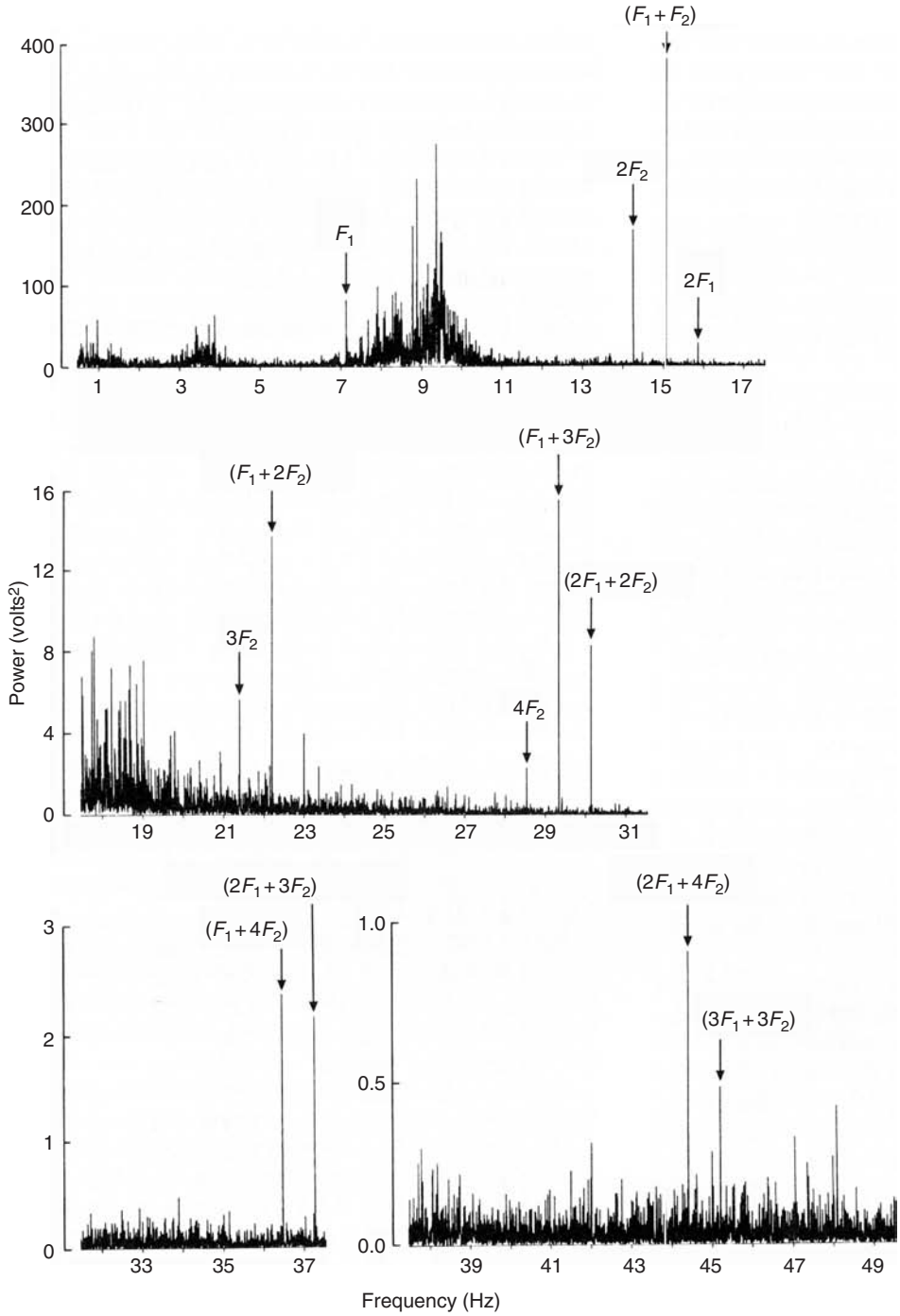

Figure 7 Ultrahigh-frequency resolution over a wide bandwidth. The EEG was analyzed by nondestructive zoom FFT at a resolution of 0.0039 Hz. See text for details. From Regan D (1989) *Human Brain Electrophysiology*: *Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*, p. 106. New York: Elsevier.

remained at 0.0078 Hz, though riding on a pedestal; the larger the noise, the higher the pedestal.

In summary, any given frequency component of the steady-state response lies within a T^{-1} bandwidth (e.g., 0.008 Hz), so that the signal-to-noise enhancement is excellent because the noise is spread along the entire EEG or MEG spectrum of, for example, 0.3 to 100 Hz.

Frequency Tagging of Multiple Stimulus Locations, Submodalities, and Modalities

The steady-state technique allows the responses evoked by stimulating different locations to be recorded simultaneously from a single electrode derivation and thus to be compared more accurately than when

recorded in succession. For example, the four quadrants of the visual field have been stimulated at four different frequencies, none differing more than 0.15 Hz. These frequency differences are negligible to the visual pathway, but can easily be separated by four parallel Fourier analyzers. Different stimulus features (e.g., orientation or spatial frequency) can also be frequency tagged. (This approach is discussed further in the section entitled 'The two-frequency method.')

The Sweep Technique

Sweep averaging is a hybrid time-domain frequency-domain technique that, because of its speed and resistance to certain kinds of noise, is useful when studying labile phenomena or when recording duration is severely limited (e.g., when recording from babies). Ordinates in Figure 6 plot the amplitude of the $2F$ Hz frequency component recorded while a subject viewed a pattern of bright and dim checks that exchanged places $2F$ times per second ($2F = 6$ in this example). Log spatial contrast of the checks was ramped linearly (abscissa). Panel (a) shows four individual sweeps; panel (b) is the average of the four traces in panel (a) and shows improved signal-to-noise ratio; each of the two traces in panel (c) is the average of 16 sweeps, while the noise trace is the average of 16 sweeps made with the stimulus occluded. A single sweep can be made within 10 s, and only one sweep may be sufficient to obtain the required information.

Evoked Response Feedback

An evoked response can exert moment-by-moment control of the stimulus that evokes it, without conscious intervention by the subject. For example, a 6 Hz steady-state response was evoked by a pattern of red checks that exchanged places six times per second. A desensitizing patch of light was superimposed on the checkerboard. A feedback loop allowed the instantaneous amplitude of the EP to control the brightness of the desensitizing light so as to hold EP amplitude constant (6 μV in this case). When the color of the desensitizing light was changed the feedback altered the brightness of the desensitizing light so as to quantify the spectral sensitivity of the response.

Analysis of Nonlinear Sensory Systems

The Two-Frequency Method

Figure 7 shows the spectrum produced by subjecting a 320 s sample of EEG to Fourier analysis while the subject viewed an unpatterned patch of light flickering at F_1 Hz superimposed on a second unpatterned patch of light flickering at F_2 Hz. In addition to the harmonic components of the responses to the F_1 Hz and the F_2 Hz flicker, Figure 7 shows multiple cross-modulation components of frequency ($nF_1 \pm mF_2$) Hz, n and m being integers. A frequency-domain approach to nonlinear systems analysis is based on the mathematical finding of MP Regan that the relative amplitudes of harmonic and cross-modulation components of the response to the sum-of-sinusoids stimulus allow candidate multistage nonlinear models to be tested against data. Note that the EP components in Figure 7 show no pedestals of the kind shown in Figure 5, indicating that moment-to-moment variations of EP amplitude and phase were small.

There is an approach that is less general than that just described: if the physiological responses to two different stimuli converge at a nonlinear stage, the properties of that stage can be investigated by applying the two stimuli simultaneously and analyzing the resulting steady-state responses. For example, in Figure 8 a subject's left and right eyes were tagged by presenting F_1 Hz flicker to the left eye (producing harmonics of F_1, including a $2F_1$ Hz response) and F_2 Hz flicker to the right eye (producing harmonics of F_2, including a $2F_2$ Hz response). The $(F_1 + F_2)$ Hz response demonstrated the presence of binocularly driven neurons.

Figure 9 demonstrates that the steady-state grating EP is generated by parallel neural mechanisms that are narrowly tuned to both spatial frequency and orientation. Figure 9(a) shows that the steady-state EP evoked by a single vertical grating has broadband spatial frequency tuning that is independent of the

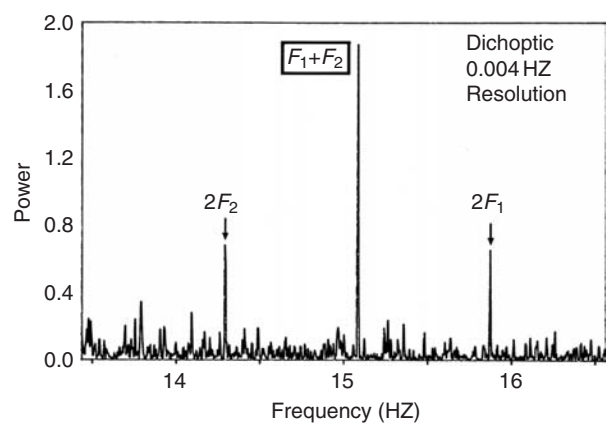

Figure 8 The 'frequency tagging' technique: an example. See text for details. Reproduced from Regan MP and Regan D (1989) Objective investigation of visual function using a nondestructive zoom-FFT technique for evoked potential analysis. *Canadian Journal of Neurological Sciences* 16: 168–179, with permission.

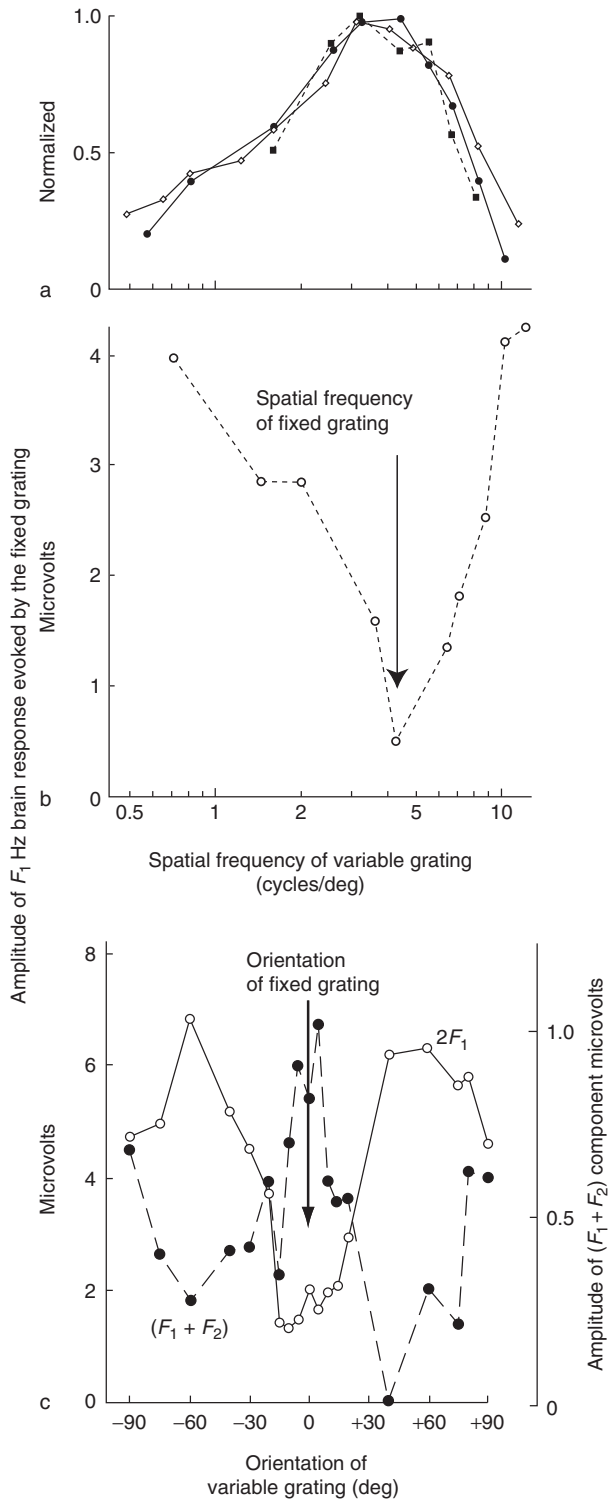

Figure 9 Selectively tuned spatial filters. (a) Normalized 20 Hz steady-state response to a single vertical grating pattern – reversing at 20 Hz for three contrasts. (b) The 20 Hz response of the fixed grating was suppressed by superimposing a second vertical grating of variable spatial frequency (abscissa), provided that the spatial frequency of the variable grating was near that of the fixed grating. (c) The 20 Hz response of the fixed grating was suppressed (open circles) by superimposing a second grating with varied orientation (abscissa). Suppression was greatest when the second grating's orientation was near vertical and near horizontal. A nonlinear $(F_1 + F_2)$ Hz component (filled circles) showed the opposite behavior. In (b) and (c) the reversal rate of the second grating was 18 Hz. (a, b) From Regan D (1983) Spatial frequency mechanisms in human vision investigated by evoked potential recording. *Vision Research* 23: 1401–1407. (c) From Regan D and Regan MP (1987) Nonlinearity in human visual responses to two-dimensional patterns and a limitation of Fourier methods. *Vision Research* 27: 2181–2183.

grating's contrast over a 15–100% range. In Figure 9(b) a second vertical grating of variable spatial frequency was superimposed on a vertical grating of fixed spatial frequency. The bright and dim bars of the gratings exchanged places $F_1 = 20$ times per second for the fixed grating and $F_2 = 18$ times second for the variable grating. The variable grating suppressed the response evoked by the first grating, but only when its spatial frequency was close to that of the fixed grating, so that the spatial filter stimulated by the fixed grating could also 'see' the variable grating. Data for only one experiment are shown in Figure 9(b), but similar results were obtained independently of the spatial frequency of the fixed grating. The bandwidth of the nonlinear suppression was 0.4 to 1.0 octaves, in good agreement with the bandwidths of spatial filters estimated psychophysically. In Figure 9(c) the spatial frequency of the variable grating was held constant while its orientation was rotated away from vertical. Evidence for an orientation-tuned mechanism is that the nonlinear suppression of the response evoked by the fixed grating (and the amplitude of the cross-modulation $(F_1 + F_2)$ Hz response) declined as the difference of the orientations of the gratings increased and reached a minimum at approximately $\pm 30°$. The suppression and the amplitude of the $(F_1 + F_2)$ component then rose to a second maximum when the two gratings were at right angles. This last finding implies that the visual processing cannot be fully understood in terms of data obtained using stimuli such as gratings or Gabors, where luminance varies along only one dimension.

A similar rationale has been used to demonstrate and locate an audio–visual convergence area in the human brain. A MEG was recorded from many sites near the scalp while a subject listened to a tone that was amplitude-modulated at F_1 Hz, while also viewing a light that flickered at F_2 Hz. Any cross-modulation response component must necessarily have been generated at some nonlinear stage where visual and auditory signals converged. The distribution of a cross-modulation component of the response allowed the site of the audio–visual convergence area to be identified.

The final example provides an indirect means of measuring the shape of the ear's hair cell transducer

function. An $f_1 = 1048.50$ Hz audio carrier tone was amplitude-modulated at $F_1 = 20.886$ Hz and fed to a subject's left ear while a second f_1 Hz carrier tone was amplitude-modulated at $F_2 = 21.778$ Hz and fed to the right ear. Clear responses at $2F_1$ Hz and $2F_2$ Hz are evident in the section of the evoked response spectrum shown in **Figure 10(A)**, but there was no $(F_1 + F_2)$ Hz component. When the two auditory stimuli were simultaneously fed to one ear, the $2F_1$ and $2F_2$ components were suppressed nonlinearly and a cross-modulation $(F_1 + F_2)$ Hz response was evident (**Figure 10(B)**). Now a small (0.1 Hz, i.e., about 0.01%) difference was introduced between the two carrier tones. **Figure 10(C)** shows that the $(F_1 + F_2)$ response component split into two parts (labeled *a* and *b*) that were separated by 0.2 Hz and, in addition, the power of the suppressed $2F_1$ Hz and $2F_2$ Hz components increased. The separation for *a* and *b* was equal to twice the frequency difference of the two carriers even when *a* and *b* moved outside the $2F_1$ and $2F_2$ components (**Figure 10(D)**, where the two carrier frequencies differed by 1.3 Hz). This strange behavior was accounted for in detail by a mathematical analysis of the processing of the sum of two amplitude-modulated auditory stimuli by the hair cell transducer function, offering a noninvasive means of detecting variations in the shape of hair cell transducer function in humans.

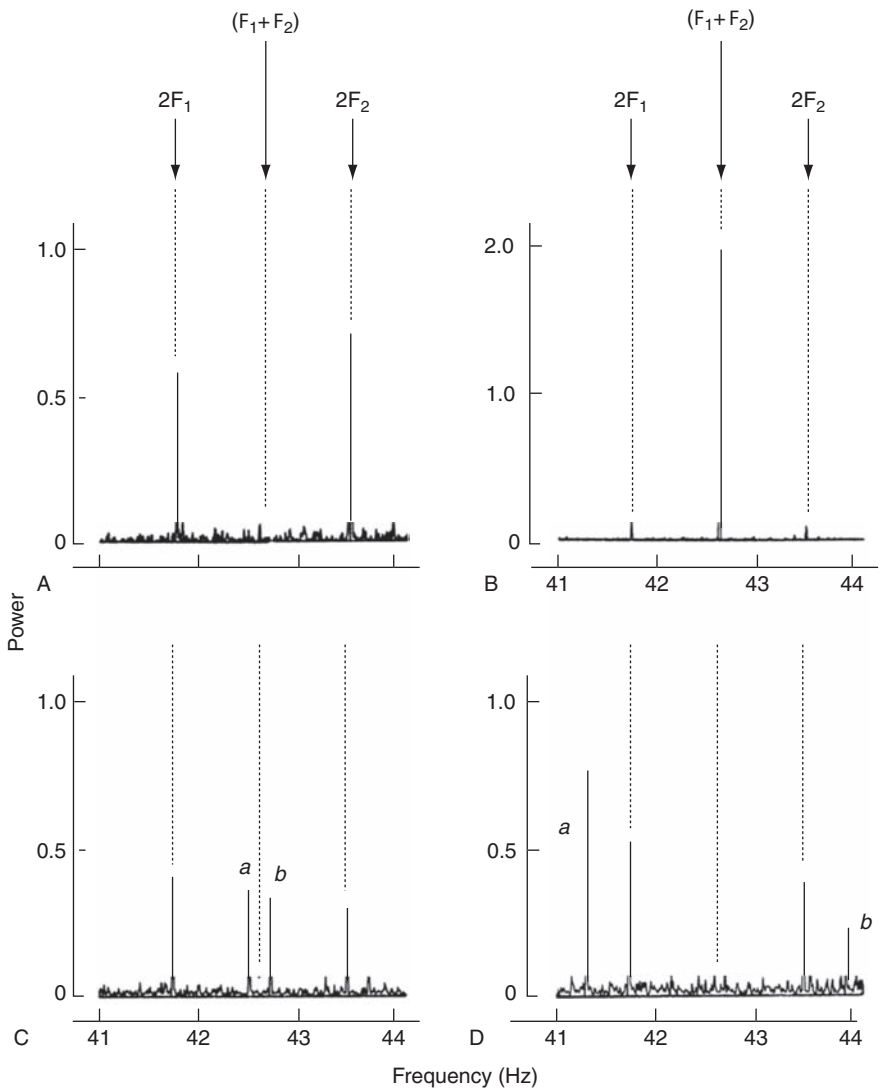

Figure 10 Steady-state responses evoked by amplitude-modulated tones. Ordinates are power with the dimensions of volts^2; 1.0 on the ordinate corresponds to 1.0 μV. Analysis frequency is plotted as abscissa. The EEG was amplified over a bandwidth of 1–200 Hz and analyzed by zoom FFT at 0.0078 Hz frequency resolution. Analyzer and stimulus generators had an hour-by-hour frequency stability better than 1 in 10^5. From Regan D and Regan MP (1988) The transducer characteristic of hair cells in the human ear: A possible objective measure. *Brain Research* 438: 363–365; and Regan MP and Regan D (1993) Nonlinear terms produced by passing amplitude-modulated sinusoids through Corey and Hudspeth's hair cell transducer function. *Biological Cybernetics* 69: 439–446.

The White Noise and Related Methods

Wiener's method of nonlinear systems analysis generates a sequence of orthogonal functionals that describe the response of the system to a Gaussian white noise input. In principle, these Wiener kernels enable the response of the system to any arbitrary input to be predicted. However, the kernels can be difficult to interpret in physiological terms. Furthermore, it is difficult to obtain higher order kernels in the presence of appreciable physiological noise, and errors can be caused if the statistical properties of the stimulus deviate from those of ideal white noise.

These last two problems are reduced in an alternative approach developed by JD Victor, BW Knight, and RM Shapley. This is to replace the stochastic white noise with a deterministic input, namely a sum of several (e.g., eight) sinusoids. The nonlinear response of the system is measured as cross-talk between the multiple input frequencies.

A second deterministic alternative to the white noise method is based on binary *m*-sequences. Sutter has applied this approach to multi-input analysis of highly nonlinear systems as, for example, the multifocal electroretinogram and visual EP.

See also: Auditory Evoked Potentials; Cognition: An Overview of Neuroimaging Techniques; Electroencephalography (EEG); Event-Related Potentials (ERPs) and Cognitive Processing; Evoked Potentials: Clinical.

Further Reading

Glaser EM and Ruchkin DS (1976) *Principles of Neurobiological Signal Analysis*. New York: Academic Press.

Marmarelis PZ and Marmarelis VZ (1978) *Analysis of Physiological Systems: The White-Noise Approach*. New York: Plenum Press.

Norcia AM and Tyler CW (1985) Visual acuity during the first year of life. *Vision Research* 25: 1399–1408.

Nunez PL (1981) *Electric Fields of the Brain: The Neurophysics of EEG*. New York: Oxford University Press.

Regan D (1973) Rapid objective refractions using evoked brain potentials. *Investigative Ophthalmology* 12: 669–679.

Regan D (1975) Color coding of pattern responses in man investigated by evoked potential feedback and direct plot techniques. *Vision Research* 15: 175–183.

Regan D (1979) Electrical responses evoked from the human brain. *Scientific American* 241(6): 134–146.

Regan D (1983) Spatial frequency mechanisms in human vision investigated by evoked potential recording. *Vision Research* 23: 1401–1407.

Regan D (1989) *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*. New York: Elsevier.

Regan MP, He P, and Regan D (1995) An audio-visual convergence area in the human brain. *Experimental Brain Research* 106: 485–487.

Regan D and Lee BB (1993) A comparison of the human 40 Hz response with the properties of macaque ganglion cells. *Visual Neuroscience* 10: 439–445.

Regan D and Regan MP (1987) Nonlinearity in human visual responses to two-dimensional patterns and a limitation of Fourier methods. *Vision Research* 27: 2181–2183.

Regan D and Regan MP (1988) The transducer characteristic of hair cells in the human ear: A possible objective measure. *Brain Research* 438: 363–365.

Regan MP and Regan D (1988) A frequency domain technique for characterizing nonlinearities in biological systems. *Journal of Theoretical Biology* 133: 293–317.

Regan MP and Regan D (1989) Objective investigation of visual function using a nondestructive zoom-FFT technique for evoked potential analysis. *Canadian Journal of Neurological Sciences* 16: 168–179.

Regan MP and Regan D (1993) Nonlinear terms produced by passing amplitude-modulated sinusoids through Corey and Hudspeth's hair cell transducer function. *Biological Cybernetics* 69: 439–446.

Slotnick SD, Klein SA, Carney T, et al. (1999) Using multi-stimulus VEP source location to obtain a retinotopic map of human primary visual cortex. *Clinical Neurophysiology* 10: 1793–1800.

Spekreijse H and Reits D (1982) Sequential analysis of the visual evoked potential system in man: Nonlinear analysis of a sandwich system. *Annals of the New York Academy of Science* 388: 72–97.

Spekreijse H, van der Tweel LH, and Regan D (1972) Interocular sustained suppression: Correlations with evoked potential amplitude and distribution. *Vision Research* 12: 521–526.

Sutter EE (2001) Imaging visual function with the multifocal *m*-sequence technique. *Vision Research* 41: 1241–1255.

Tonini G, Srinivasan R, Russell DP, et al. (1998) Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. *Proceedings of the National Academy of Science of the United States of America* 95: 3198–3203.

Victor JD (1979) Nonlinear systems analysis: Comparison of white noise and sum of sinusoids in a biological system. *Proceedings of the National Academy of Science of the United States of America* 76: 996–998.

Evolution of Sensory Receptor Specializations in the Glabrous Skin

N J Dominy, University of California, Santa Cruz, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Glabrous skin is devoid of dermal filaments such as hair or feathers. It is prevalent among amphibians and reptiles and is characteristic of the maxillary rictus, interdigital web, and combs and wattles of birds. The glabrous skin of birds is lipogenic and sometimes temporary, e.g., brood (incubation) patches, or associated with the attainment of maturity. For instance, feather loss from the head is linked to maturation in some ibises and storks. Glabrous skin is more variable among mammals. The rhinarium, lips, and genitals are glabrous, as well as parts of limb extremities in contact with the physical environment during locomotion, such as the flippers of seals and sea lions and the cutaneous pads of paws. A patch of glabrous skin also exists on the prehensile tails of atelid monkeys and certain other arboreal mammals. Such animals typically suspend their bodies by way of their tail when moving or foraging on terminal branches. More specialized examples of glabrous skin include the nasal rays of the star-nosed mole, *Condylura cristata*, and the upper bill of the platypus, *Ornithorhynchus anatinus*, which is an extension of skin that covers much of the face. At the proximal end of the lower bill is a similar extension that folds over the neck. These areas are termed upper and lower shields, respectively.

Glabrous skin is a major organ for sensing the external environment. Sensation from the glabrous skin – along with sensations from hirsute skin, muscles, tendons, joints, and bones – is termed somatic sensation. It is initiated by a variety of specialized receptors, many of which have structurally elaborate terminals that transform specific stimuli into nervous impulses. The sensory terminal may be encapsulated or linked to mesodermal or ectodermal elements. The extraneural cells are not necessarily excitable, but rather create the right circumstances for the excitation of the neuronal dendrite or modify its excitation in some way. Some receptors are remarkably uniform across vertebrate and mammalian orders, although considerable variation exists in the numbers of sensory receptors per square centimeter among the different areas of the body surface, and with advancing age the density of receptors can decrease. Not surprisingly, the adaptive function of glabrous skin and its associated receptors tends to vary according to the locomotor, foraging, and reproductive behaviors particular to a certain species.

Types of Cutaneous Receptors

The histological arrangement of a cutaneous receptor can differ greatly between hirsute and glabrous skin, as well as between animal species. This article has been written with a strong emphasis on the morphology and function of specialized receptors in the glabrous skin of mammals, which includes free nerve endings (FNEs), Merkel cells, Meissner corpuscles, Ruffini corpuscles, and Pacinian and paciniform corpuscles. Before the availability of electron microscopy, several other lamellar receptors were recognized (e.g., Golgi-Mazzoni and Dogiel corpuscles). However, they all follow the lamellar design pattern of the paciniform receptor and are considered part of this group.

When stimulation of a cutaneous receptor is maintained at a steady level, there is, as in all receptors, an initial burst – the dynamic phase – followed by gradual habituation to that steady level – the static phase. Although all receptors evince these two phases, one or the other predominates, giving a distinction between rapidly adapting (RA) endings that accurately record the rate of stimulus onset and slowly adapting (SA) endings that signal the constant amplitude of a stimulus, for example, the position sense. SA endings are further classified by the regularity of their discharge properties (type I and type II).

Free Nerve Endings

FNEs are the terminal receptor structures of sensory C-fibers or Aδ fibers. FNEs are found in all connective tissues, including the epidermis and dermis of hirsute and glabrous skin. Afferent fibers are myelinated or nonmyelinated but are always of small diameter and low conduction velocity. The terminal arborizations of myelinated afferent axons are nonmyelinated and devoid of Schwann cell investment. Electrophysiological recordings of FNEs in the dermis show that they respond to several sensory modalities. Some FNEs respond to moderate cold or heat (thermoreception), to low-threshold mechanical touch (mechanoreception), or to damaging heat, cold, or deformation (nociception). The FNEs of afferent fibers are mainly nociceptive in the nonmyelinated portion. The lips and palms of primates feature an abundance of nonmyelinated C-fibers and a paucity of finely

myelinated Aδ fibers, which together give glabrous skin a lower heat detection threshold and a higher heat pain threshold than hirsute skin. This condition is hypothesized to enhance the function of glabrous skin as an exploratory surface.

The exploratory function of glabrous skin is exemplified by the nasal rays of the star-nosed mole, *Condylura cristata*, which are richly invested with Eimer's organs and are the mole's primary sense organs. Named after the German zoologist Gustav Heinrich Theodor Eimer (1843–1898), Eimer's organs are a special form of FNE found in the rhinarium of fossorial mammals such as talpid moles. In this structure, the FNE extends through a column of thickened epidermal cells. Below the column are Merkel cell complexes and lamellated paciniform corpuscle. The presence of three types of specialized receptor raises the possibility that Eimer's organs have several important functions. The primary function appears to be the tactile detection and recognition of invertebrate prey, but the presence of up to three paciniform corpuscles at the base of each Eimer's organ suggests that it may also play a role in sensing seismic vibrations. Such vibrations are important signals of communication between moles.

The importance of FNEs for sensing prey is also evident in the infrared pit organs of snakes. In the rattlesnake, *Crotalus horridus*, the morphology of the pit organ promotes efficient thermoreception by a large number (several thousand) of intradermal myelinated axons 3–16 μm in diameter.

The Merkel Cell System

Friedrich Sigmund Merkel (1845–1919) was an outstanding German anatomist of the late nineteenth century. He described cells in the skin of amphibians, birds, and mammals and referred to them as *Tastzellen,* or 'touch cells.' The cells were given the eponym Merkel cells in 1878 by Robert Bonnet (1851–1921). Distinctive traits include dense core vesicles, cytoplasmic projections, and synaptiform contacts with enlarged nerve terminals. They measure 10–15 μm across and exist in the dermis of birds and the basal layer of the epidermis of mammals. In mammals, Merkel cells are a ubiquitous component of both hirsute and glabrous skin. The myelinated afferent fiber branches extensively in the dermis within 500 μm of the epidermis. The branches lose their myelin sheaths before they penetrate the basement membrane of the epidermis to form the flattened nerve plate (Merkel disk) lying along the lower border of the Merkel cell. The disk itself is *c.* 7 μm in diameter and 1 μm thick and contains numerous mitochondria, which occupy 50% of the terminal. Merkel disks exist only in adult mammals.

Although direct evidence is lacking, correlative electrophysiological and morphological evidence suggests that Merkel cells function as SA type I mechanoreceptors. Calcium ions enter the cell in response to maintained mechanical stimuli and trigger the release of a neurotransmitter, probably glutamate. The conserved morphology and widespread distribution of Merkel cells across such divergent animals as lizards, salamanders, and lampreys suggests an early appearance during vertebrate evolution. The adaptive significance appears to be associated with generalized somatosensory reception, particularly of prey objects. They are most sensitive to perpendicular deformation of the skin surface or to the bending of sinus hairs or vibrissae. Furthermore, the density of Merkel cells is manifested most strikingly in structures associated with foraging, such as the electroreceptive snouts of the platypus and echidna and the mystacial skin of most mammals, except humans.

An association with foraging is also evident in the beak skin of birds, where Merkel-like cells occur in dense corpuscles containing up to 50 cells. The corpuscles, termed Grandry corpuscles, are present in the glabrous skin of the bill and tongue of wading birds but absent in the domestic chicken and Japanese quail. A structural resemblance of Grandry corpuscles to Merkel cells points to analogous functions, but they are not morphologically identical, and some evidence suggests that Grandry corpuscles function as RA receptors. The functional and ecological significance of such differences awaits elucidation.

The Meissner Corpuscle

Georg Meissner (1829–1905) was a German anatomist and physiologist of the nineteenth century. He and Rudolf Wagner (1805–1864) described distinctive corpuscles in 1852, which became the subject of a monograph by Meissner in 1853. This monograph is the basis for Meissner's name being associated with the receptor. Meissner corpuscles are located in the dermal papillae of glabrous skin, where they connect to and tightly abut the basal surface of the epidermis. They are innervated by one or two myelinated fibers from the subepidermal nerve plexus. The fibers lose their myelination before entering the corpuscle, an ovoid capsule of perineural cells. Therein the nerve endings branch repeatedly, adopting a ribbonlike shape with bulbous expansions. Collagen fibers and fibrocytes bond the spiraling nerve endings to stacks of lamellar Schwann cells and attach the capsule to the epidermis. The circumscribed expansion of unmyelinated nerve endings into lamellar disks is a distinctive feature of Meissner corpuscles.

Such complex innervation has led to speculation that Meissner corpuscles have a nociceptive function, although aspects of the disciform expansion suggest a primary role in mechanoreception. Takahashi-Iwanaga and Shimoda reported that the disk margins are serrated with fine projections of lamellar Schwann cells that tightly hold the collagen trabeculae to the inner aspect of the pericorpuscular capsule. They concluded that the disks are susceptible to mechanical deformations of surrounding tissues and, furthermore, that the distortion of axonal endings during the dynamic phase of tissue deformation will generate RA receptor potentials. Castano et al. observed that "the minutest deformation of the skin [can be] immediately transmitted to the capsule and, via the cross-beam system linking it to the framework, to the transducer elements of the corpuscle" (1985: 296). The disciform expansion is thus well suited to mechanotransduction. On this basis, Meissner corpuscles are hypothesized to play a key role in the perception of elastic texture. They are classified as RA receptors that respond to transient or phasic stimuli. They are particularly sensitive to vibrations at frequencies ranging from 10 to 400 Hz with maximum sensitivity between 100 and 200 Hz.

Meissner corpuscles are reported from the glabrous skin of the lips and extremities (hands, feet, and prehensile tails) of primates and one marsupial, the Virginia opossum, *Didelphis virginiana*. Murine and sciurid corpuscles are similar in shape, but smaller and less elaborate. The high concentration of Meissner corpuscles at the tips of primate digits as well as the glabrous patch of skin present on the prehensile tail of atelid monkeys suggest that they function to sense the elastic properties of support structures, which has adaptive value in an arboreal milieu. An additional hypothesis is that the concentration of Meissner corpuscles at the digits supports the perception of food texture, particularly ripening fruits, which are often palpated by foraging primates. Hoffman et al. reported that the density of Meissner corpuscles in primate digits varied positively as a function of the overall proportion of fruit in the diet. In general, the function of Meissner corpuscles appears to be related to sensing the texture of critical objects in the environment – the paucity of Meissner corpuscles in other mammalian orders is curious.

Pacinian Corpuscle

Filippo Pacini (1812–83) was an Italian anatomist posthumously famous for isolating the cholera bacillus *Vibrio cholerae* in 1854. Pacinian corpuscles, however, were first described by Abraham Vater (1684–1751) in 1741 and were rediscovered by Pacini almost 100 years later in 1840. Pacinian corpuscles are the largest and best-known corpuscles in mammals. They are oval, spherical, or irregularly coiled and are exceptionally large, up to 2 mm in length and 100–500 μm wide. Each has a capsule formed by *c*. 30 concentrically arranged lamellae of flat cells about 0.2 μm thick. Pacinian corpuscles are located in the palmar and plantar aponeurosis or genitalia deep to the skin. They are also present in ligaments and joint capsules of mammals and ostriches and the mesentery of the cat intestine where they are visible to the unaided eye. The afferent nerve fibers are myelinated with diameters between 3 and 6 μm. Extensive studies show these endings to behave as RA mechanoreceptors responding to sudden disturbances and vibrations (40–1000 Hz), with an optimal response at the higher end of the range.

The widespread distribution of Pacinian and paciniform corpuscles throughout the body makes it difficult to assign an adaptive significance to their presence in glabrous skin. Yet attendant associations between paciniform corpuscles and specialized structures such as Eimer's organs and the push rods of monotremes suggest that paciniform corpuscles are sometimes recruited for specialized roles. The push rod is a structure unique to the glabrous snout of the echidna, *Tachyglossus aculeatus*, and the upper and lower shields of the platypus, *Ornithorhynchus anatinus*. It is composed of a column of flattened spinous cells, which are cross-linked by tight junctions. The tip of the rod emerges at the skin surface with a domelike projection. The base of the push rod resembles an Eimer's organ in that it is associated with Merkel cell complexes and lamellated paciniform corpuscles. The rod is typically 300 μm long with a diameter of 50 μm. Along part of its length the rod remains separated from adjacent regions by connective tissue papillae, which gives it a degree of independent mobility. The push rod is hypothesized to play a role in signaling contact with prey, particularly in moist or aquatic environments where electroreception is advantageous.

A functionally similar structure, termed a domed pressure receptor (DPR), has been reported on the rostrum of the American alligator, *Alligator mississippiensis*. The DPR is hypothesized to play a role in detecting the disturbance of water by prey. The types of receptors present at the base of the DPR await elucidation, but it is likely that paciniform corpuscles are present. The avian homolog of the Pacinian corpuscle is the Herbst corpuscle, which is present in the skin of the bill of wading and aquatic birds and in other species, and also in other bodily locations, including the legs and trunk. Herbst corpuscles are

particularly dense in the bill tip organ of the goose and duck, and general agreement exists that they function as RA receptors with frequency-following characteristics similar to Pacinian corpuscles.

Ruffini Corpuscle

Angelo Ruffini (1864–1929) reported cigarlike structures in 1893. Ruffini corpuscles consist of a long fusiform, encapsulated structure innervated by a myelinated large-caliber (6–12 μm) axon terminating as numerous branches intertwined among collagen bundles. Ruffini corpuscles have been reported in glabrous and hirsute skin, as well as in the tendinous insertions and joint capsules of mammals. Because of their regular discharge property in cats, they are classified as SA type II mechanoreceptors in contrast to the irregular discharge properties of SA type I fibers. Ruffini corpuscles are sensitive to skin stretch, particularly tangential forces generated in the skin during tasks like grasping.

Evolution of Cutaneous Receptors

In general, the morphological principles and mechanisms of cutaneous receptors are conserved across vertebrate orders. Each type of receptor responds to specific stimuli and contributes to the global perception of texture in the central nervous system. Specialized nerve endings of the glabrous skin tend to vary in density according to the particular locomotor or foraging behavior of an animal species. For instance, Meissner corpuscles are more complex and densely packed in the digits of primates than in other mammals, which lack the same degree of forelimb opposability. Species with very specialized receptor structures, such as the star-nosed mole and platypus, tend to also have a highly developed somatosensory cortex. The evolution of specialized receptors in the glabrous skin thus appears to result primarily from selection for the reorganization and concentration of preexisting cells. Improved somatosensory acuity favored, in turn, a greater degree of cortical investment devoted to the interpretation of those sensory inputs.

See also: Electrical Perception and Communication; Enteric Nervous System: Sensory Pathways; Seismic and Vibrational Signals in Animals; Sensory System Specializations; Tactile Texture.

Further Reading

Castano P, Ventura RG, and Maddalone M (1985) Notes on morpho-functional differences between Meissner's corpuscle of man and the green monkey (*Cercopithecus aethiops* L). *Folia Morphologica* 33: 294–298.

Catania KC and Remple FE (2004) Tactile foveation in the star-nosed mole. *Brain, Behavior, and Evolution* 63: 1–12.

Fagan BM and Cahusac PM (2001) Evidence for glutamate receptor mediated transmission at mechanoreceptors in the skin. *Neuroreport* 12: 341–347.

Granovsky Y, Matre D, Sokolik A, Lorenz J, and Casey KL (2005) Thermoreceptive innervation of human glabrous skin: A contact heat evoked potential analysis. *Pain* 115: 238–247.

Hoffman JN, Montag AG, and Dominy NJ (2004) Meissner corpuscles and somatosensory acuity: The prehensile appendages of primates and elephants. *Anatomical Record. Part A, Discoveries in Molecular, Cellular, and Evolutionary Biology* 281: 1138–1147.

Iggo A and Andres KH (1982) Morphology of cutaneous receptors. *Annual Review of Neuroscience* 5: 1–31.

Lumpkin EA and Caterina MJ (2007) Mechanisms of sensory transduction in the skin. *Nature* 445: 858–865.

Paré M, Elde R, Mazurkiiewicz JE, Smith AM, and Rice FL (2001) The Meissner corpuscle revised: A multiafferented mechanoreceptor with nociceptor immunochemical properties. *Journal of Neuroscience* 21: 7236–7246.

Pettigrew JD, Manger PR, and Fine SLB (1998) The sensory world of the platypus. *Proceedings of the Royal Society of London Series B* 353: 1199–1210.

Proske U, Gregory JE, and Iggo A (1998) Sensory receptors in monotremes. *Proceedings of the Royal Society of London Series B* 353: 1187–1198.

Soares D (2002) An ancient sensory organ in crocodilians. *Nature* 417: 241–242.

Takahashi-Iwanaga H and Shimoda H (2003) The three-dimensional microanatomy of Meissner corpuscles in monkey palmar skin. *Journal of Neurocytology* 32: 363–371.

Evolution of the Limbic System

L L Bruce, Creighton University School of Medicine, Omaha, NE, USA
M R Braford, Jr., Oberlin College, Oberlin, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The concept of the limbic system is derived from that of the limbic lobe, which is formed primarily by the parahippocampal and cingulate gyri, as well as two underlying groups, the hippocampal formation and amygdalar complex. The various components of the limbic lobe receive and integrate visceral, olfactory, and somatic information. The limbic lobe is involved in processing recent memory (hippocampal formation) and in the expression of emotion (amygdalar complex). The term 'limbic system' expands this concept to include other structures that are functionally and anatomically related to the limbic lobe, and thus includes the hypothalamus, septal nuclei, various thalamic nuclei, and olfactory systems. Today, the hippocampal formation and the amygdalar complex are usually regarded as the primary elements of the limbic system, and their evolution is the focus of this article.

The hippocampus forms along the dorsomedial edge of the telencephalon. In mammals it expands to form an arc in the caudal telencephalon, with the most prominent part in the temporal lobe. The hippocampus plays critical roles in processing learning and memory information. The amygdalar complex is composed of multiple groups that have different origins, and unique functional, connectional, and neurochemical expression characteristics. These groups play critical roles in feeding, reproduction, and flight/defense reactions, and also contribute to higher level cognitive functions, including emotional memory processing and sensory–sensory associations.

Homologs of the hippocampus and parts of amygdalar complex are usually defined by mammalian characteristics. Comparable areas have been identified in most vertebrates classes, although they show considerable variability. The following discussions provide a review of the evidence for these limbic groups in mammals, reptiles, birds, amphibians, and jawed fish, and how they may have evolved.

Mammals

Early in neurogenesis the mammalian pallium (roof) of the telencephalon is specified into four regions, the medial, dorsal, lateral, and ventral pallia. These are generally identified in the adult as hippocampus (dentate gyrus and Ammon's horn), the 'general cortex' (nonolfactory and nonhippocampal), the olfactory cortex (piriform and entorhinal), and the pallial amygdalar complex, respectively. These four regions are believed to be present in all jawed vertebrate embryos, although divergent evolution may make them difficult to recognize (**Figures 1** and **2(a)**).

Sensory Projections to the Telencephalon

Olfaction is detected in the main olfactory organ and the vomeronasal organ, although many aquatic mammals and humans lack a vomeronasal organ. The main olfactory organ detects light, airborne molecules and projects to the main olfactory bulb, which in turn projects as a superficial sheet over portions of the telencephalon. Its targets include the olfactory cortex (piriform and entorhinal), a small rostral part of the medial cortex, and several cortical amygdalar nuclei, including the anterior cortical, posterior lateral cortical, and the anteroventral part of the medial amygdala (**Figure 3(a)**). The vomeronasal organ detects heavy odor molecules and projects to the accessory olfactory bulb, which in turn projects to most of the medial amygdala and the posterior medial cortical amygdala.

Thalamic nuclei relay auditory, visual, and somatosensory information to the striatum and lateral amygdalar nucleus, as well as to the sensory cortical areas (**Figure 4**). The auditory cortical areas receive projections from the majority of the medial geniculate nucleus. The striatum and lateral amygdala receive projections from the medial subdivision of the medial geniculate, a group that is sometimes placed within the perigeniculate area because of its amygdalar projection.

Visual information from the retina reaches the telencephalon through at least three distinct pathways: from retina to thalamus (lateral geniculate), to primary visual cortex; from retina to superior colliculus, to thalamus (lateral posterior or pulvinar nuclei), to secondary visual cortical areas; and from retina to superior colliculus, to thalamus (supragenicular nucleus of the perigeniculate area), to the striatum and lateral amygdala.

Somatosensory information reaches the thalamus from the spinal cord, dorsal column nuclei, and trigeminal sensory nuclei, and a somatosensory midbrain area. There appear to be two main groups of thalamic nuclei that relay this information to the telencephalon: the ventrobasal and posterior thalamic nuclei project to the somatosensory cortex, whereas the posterior intralaminar thalamic nucleus of the perigeniculate group projects to the striatum and lateral amygdala.

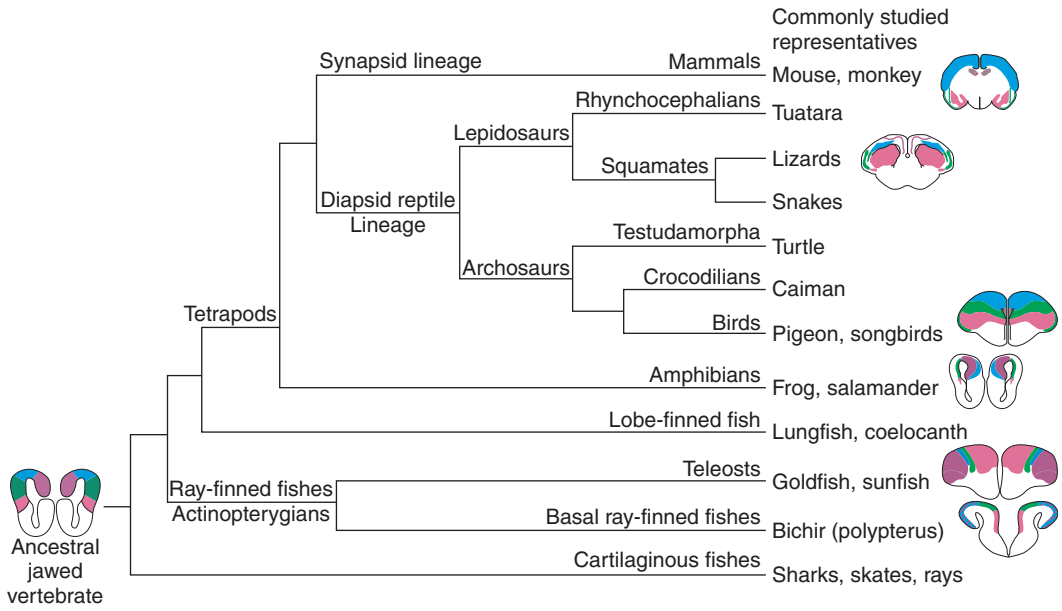

Figure 1 Relationships between the major groups of jawed vertebrates. The hypothetical pallial organization of the telencephalon of an ancestral jawed vertebrate and the variations that have evolved in six representatives are shown using the following representations: purple, medial pallium; blue, dorsal pallium; green, lateral pallium; pink, ventral pallium. The telencephala of the jawed vertebrates are derived from regions that originate from comparable parts of the embryonic ventricle. See **Figure 2** for more detailed representations of the telencephala.

In summary, auditory, visual, and somatosensory information is relayed through separate parts of the thalamus and reaches the sensory cortex and lateral amygdala through separate, parallel pathways.

Hippocampus and Cortex

The hippocampus, composed of the dentate gyrus and Ammon's horn, forms along the medial edge of the pallium. It undergoes considerable enlargement, bending caudally around the anterior commissure. The hippocampus is characterized by several massive connections. In particular, Ammon's horn receives multisensory information from the entorhinal area, projects to the subiculum, and has extensive commissural connections.

Having undergone an explosive elaboration from its ancestral tetrapod condition, the mammalian 'general cortex' is highly derived and forms numerous specific cortical subdivisions. Characteristics common to most cortical areas are: (1) descending projections to the diencephalon and midbrain, (2) commissural connections, (3) ascending projections from the thalamus, and (4) several neurochemically similar cell types.

The olfactory cortex is derived from the lateral pallium, and is distinguished by its topology, developmental origins, and connections. It is the primary target of the main olfactory bulb and reciprocates this connection. The underlying claustrum–endopiriform complex also originates mainly from the lateral pallium.

The ventral pallium consists of an amygdalar cortex, which receives olfactory and vomeronasal information, and a deep group of nuclei, the basolateral amygdalar complex, which includes the lateral, basolateral, and basomedial amygdala. Another amygdalar group, the central amygdala, originates from the subpallium. Differentiating the nuclei within the amygdalar complex requires identifying the targets of the olfactory system and the sensory projections to the telencephalon, and so these are described first.

Amygdalar Complex and Claustrum–Endopiriform Area

Five major amygdalar groups can be recognized in mammals on the basis of their connections and topological relationships (**Figure 5(a)**): (1) nuclei with long descending projections to the hypothalamus and hindbrain, (2) nuclei that receive accessory olfactory connections, (3) nuclei that receive olfactory connections and project to the hypothalamus, (4) deep amygdalar nuclei that receive ascending sensory input from the dorsal thalamus but lack hypothalamic projections, and (5) deep amygdalar nuclei that receive sensory information from the dorsal thalamus and project to the hypothalamus. In addition, the claustrum-endopiriform lies between the olfactory cortex and basolateral amygdala, and is described here because of its close association with the amygdalar complex.

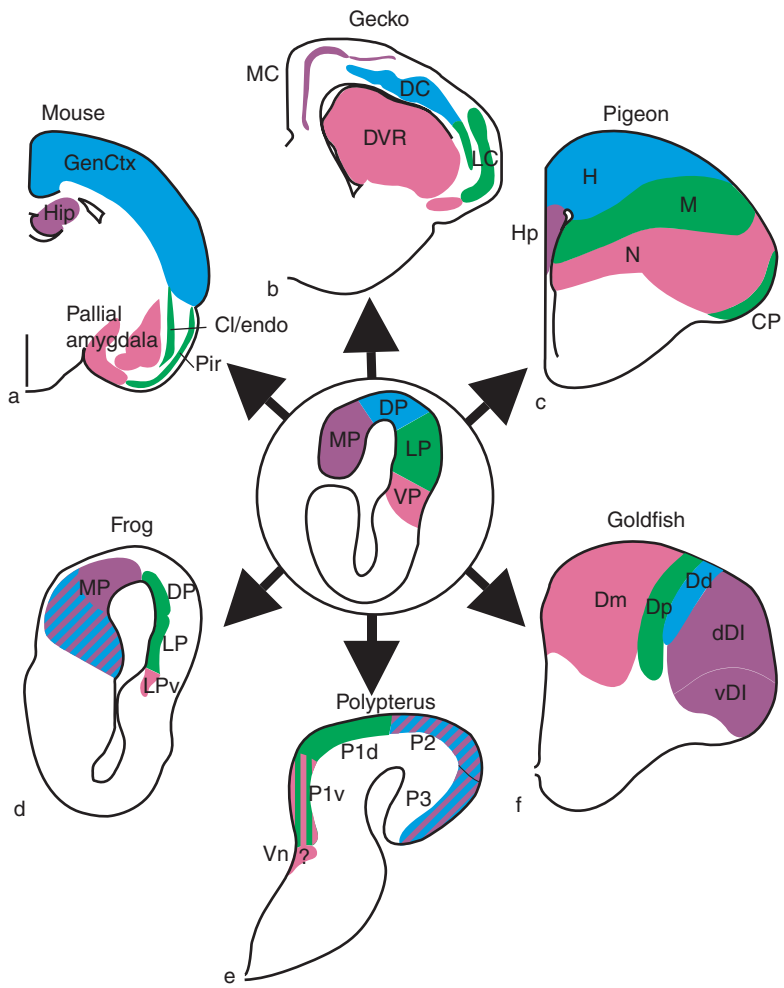

Figure 2 The telencephala of the jawed vertebrates are composed of similar regions that appear to originate from comparable parts of the embryonic ventricle (center). Observed differences among vertebrate classes are generally attributed to the flexion (eversion or inversion) of the telencephalon, or to altered rates of proliferation. (a) Mouse: Cl/Endo, claustrum–endopiriform area; GenCtx, general cortex; Hip, hippocampus; Pir, piriform cortex. (b) Gecko: DC, dorsal cortex; DVR, dorsal ventricular ridge; LC, lateral cortex; MC, medial cortex. (c) Pigeon: CP, cortex piriformis; H, hyperpallium; Hp, hippocampus; M, mesopallium; N, nidopallium. (d) Frog: DP, dorsal pallium; LP, lateral pallium; LPv, ventral part of lateral pallium; MP, medial pallium. (e) Polypterus: P1d, dorsal subdivision of pallial area 1; P1v, ventral subdivisions of pallial area 1; P2, pallial area 2; P3, pallial area 3; Vn, other nucleus of ventral pallial area. (f) Goldfish: Dd, dorsodorsal area; dDl, dorsal subdivision of dorsolateral area; Dm, dorsomedial area; Dp, dorsoposterior area; vDl, ventral subdivision of dorsolateral area.

The central amygdala is characterized by long descending projections to autonomic areas of the lateral hypothalamus and hindbrain. It plays a critical role in eliciting appropriate responses to stressful environmental conditions and in regulating behaviors associated with fear. The posterior medial cortical amygdalar nucleus and most of the medial amygdala receive input from the accessory olfactory bulb. Their main output is to the ventromedial and lateral hypothalamic areas. Functionally, the medial amygdala appears to be essential for relaying olfactory information to regions of the hypothalamus that control mating behavior. In addition to its inputs to the olfactory (piriform) cortex and the entorhinal cortex, the main olfactory bulb also projects to the cortical amygdalar nuclei that cover the surface of most of the amygdalar complex. These nuclei are the anterior cortical amygdala, posterior lateral cortical amygdala, and anteroventral division of the medial amygdala. They are all connected with the hypothalamus and basolateral complex.

Auditory, visual, and somatosensory nuclei in the caudal thalamus project to the lateral amygdala, which is a subdivision of the basolateral complex. The lateral amygdala differs from sensory cortical areas because it lacks projections to the thalamus, brain stem, and opposite hemisphere. In addition, it originates from the ventral pallium, whereas sensory cortex originates from the dorsal pallium. The lateral amygdala projects mainly to the basolateral amygdala, which

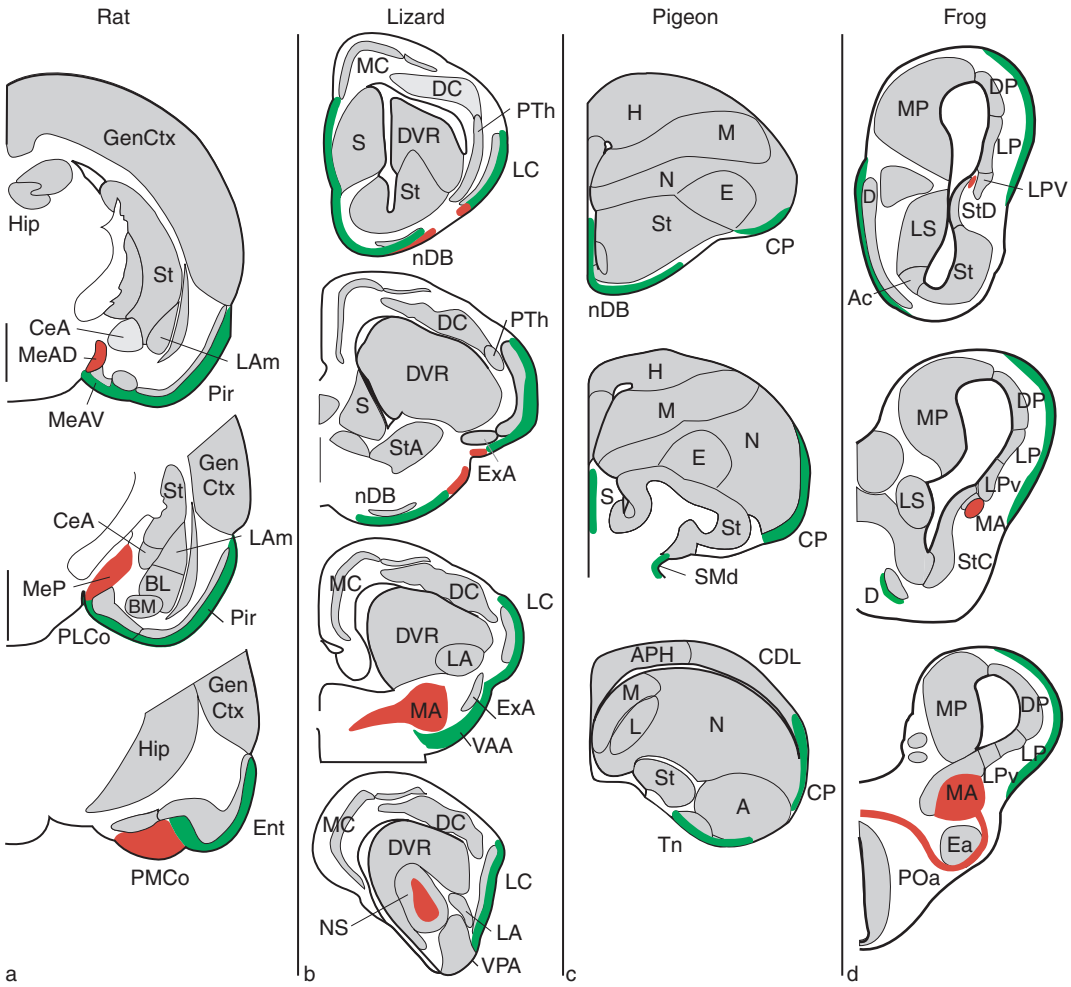

Figure 3 Comparison of the projections from the main olfactory bulb (green) and accessory olfactory bulb (red) in tetrapods. (a) Rat: BL, basolateral amygdala; BM, basomedial amygdala; CeA, central amygdala; Ent, entorhinal cortex; GenCtx, general cortex; Hip, hippocampus; LAm, lateral amygdala; MeAD, anterodorsal part of medial amygdala; MeAV, anteroventral part of medial amygdala; Pir, piriform cortex; PLCo, posterolateral cortical amygdala; PMCo, posteromedial cortical amygdala; St, striatum. (b) Lizard: DC, dorsal cortex; DVR, dorsal ventricular ridge; ExA, external amygdalar nucleus; LA, –lateral amygdalar nucleus; LC, lateral cortex; MA, medial amygdalar nucleus; MC, medial cortex; nDB, nucleus of the diagonal band; NS, nucleus sphericus; PTh, pallial thickening; S, septal nuclei; St, striatum; StA, striatoamygdalar area; VAA, ventral anterior amygdalar nucleus; PA, ventral posterior amygdalar nucleus. Pigeon: A, arcopallium; APH, perihippocampal area; CDL, dorsolateral cortical area; CP, cortex piriformis; E, endopallium; H, hyperpallium; Hp, hippocampus; L, field L; M, mesopallium; N, nidopallium; nDB, nucleus of the diagonal band; S, septum; SMd, dorsal tract of stria medullaris; St, striatum; Tn, nucleus taenia. (d) Frog: Ac, accumbens; D, diagonal band nucleus; DP, dorsal pallium; Ea, anterior entopeduncular nucleus; LP, lateral pallium; LPv, ventral part of lateral pallium; LS, lateral septum; MA, medial amygdala; MP, medial pallium; POa, anterior preoptic area; St, striatum; StC, caudal striatum; StD, dorsal striatum. Modified from Bruce LL and Neary TJ (1995) The limbic system of tetrapods: A comparative analysis of cortical and amygdalar populations. *Brain, Behavior and Evolution* 46: 224–234, with permission from S Kargel AG, Basel, with modifications from Bruce LL (2006) Evolution of the reptilian brain. In: Kaas JH (ed.) *Evolution of Nervous Systems, Vol 2: Non-Mammalian Vertebrates*, pp. 125–156. Oxford: Academic Press.

is another subdivision of the basolateral complex. The basolateral amygdala, basomedial amygdala, and the ventrolateral subdivision of the lateral amygdala all receive sensory input from the lateral amygdala and project to various parts of the hypothalamus. Some neurons within the basomedial nucleus respond to faces, and especially to faces that elicit emotional responses. The sensory input to this group appears to be critical for comprehending the emotional context of social interactions. The hypothalamic connections may contribute to responding to sensory stimuli within appropriate behavioral contexts, such as territorial signaling or courtship rituals. In addition, the claustrum and endopiriform areas develop from the lateral pallium, and can be identified by their unique gene and neurochemical expression patterns. They are extensively connected with cortical areas and the lateral amygdala.

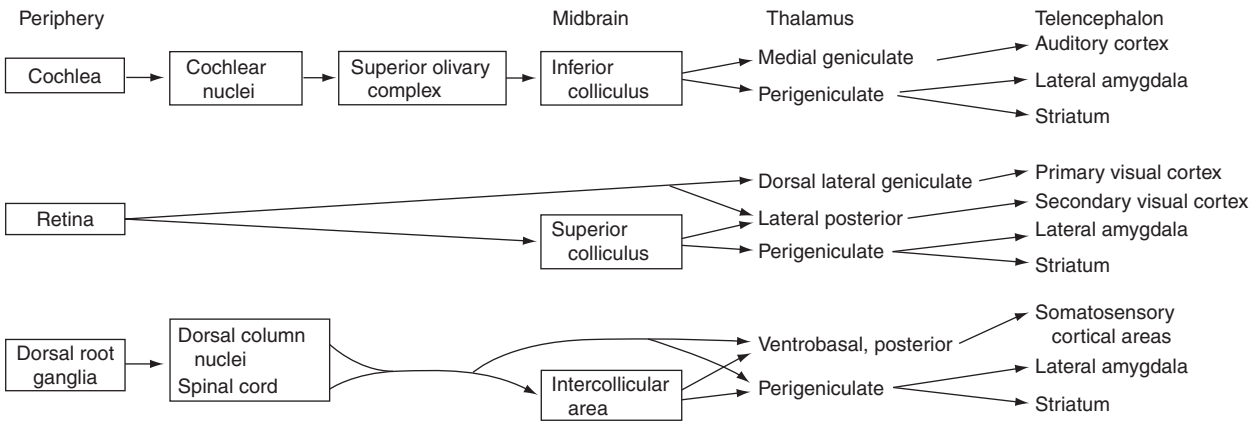

Figure 4 Comparison of the sensory pathways through the thalamus to the cortical and amygdalar targets in mammals. Note that auditory (top), visual (middle), and somatosensory (bottom) pathways reach the cortex and amygdala through separate, parallel thalamic pathways.

Reptiles

The medial, dorsal, and lateral cortices of reptiles are generally compared to the hippocampal, general (nonolfactory and nonhippocampal), and olfactory cortices of mammals, respectively (Figure 2(b)), and are derived from the medial, dorsal, and lateral pallia. The dorsal ventricular ridge is derived from the ventral pallium and appears to be homologous to part of the mammalian basolateral amygdalar complex.

Sensory Pathways to the Telencephalon

The main olfactory organ is present in all reptiles; however, the vomeronasal (accessory olfactory) system occurs in rhynchocephalians (tuatara), lizards, snakes, and turtles, but is absent in crocodilians and birds (Figure 3(b)). The main olfactory bulb projects to superficial parts of the telencephalon, including the lateral cortex, a rostral part of the medial cortex, nucleus of the diagonal band, and several cortical amygdalar nuclei. The axons decussate in the habenular commissure and innervate similar contralateral targets. The accessory olfactory bulb projections include the medial amygdala, nucleus sphericus, and the ventral posterior amygdala. The olfactory targets are very similar in reptiles and mammals, except that mammals lack contralateral main olfactory projections.

In reptiles, as in mammals, auditory, visual, and somatosensory information reaches the telencephalon through two distinct pathways. One thalamic group projects to a cortical target, and another projects to the striatum and dorsal ventricular ridge (Figure 6). The auditory midbrain is located in the medial torus semicircularis and is comparable to the mammalian inferior colliculus. It relays auditory information to two thalamic nuclei: dorsolateralis anterior and medialis (named reuniens, in turtles and crocodiles). Dorsolateralis anterior projects to the dorsal cortex, and medialis projects to the striatum and the medial part of the dorsal ventricular ridge.

Visual information reaches the telencephalon through three distinct pathways. One is a retino–thalamic–telencephalic path, and the other two are retino–tecto–thalamic–telencephalic paths. The retina projects directly to a thalamic nucleus, called intercalatus or the lateral part of dorsolateralis anterior in lizards, and called the lateral geniculate in turtles. This nucleus then projects to the lateral part of dorsal cortex. The retina also projects to the optic tectum, which in turn projects to two thalamic nuclei, dorsolateralis anterior and rotundus. Dorsolateralis anterior projects to the dorsal cortex, and rotundus projects to the striatum and dorsal ventricular ridge. Somatosensory information reaches the thalamus from the spinal cord, dorsal column nuclei, and a somatosensory midbrain area called torus semicircularis. Two somatosensory thalamic nuclei, dorsolateralis anterior and the medialis posterior complex, relay this information to the dorsal cortex, and to the striatum and dorsal ventricular ridge, respectively.

In summary, in both mammals and reptiles there are at two distinct thalamic regions that relay auditory, visual, and somatosensory information to the telencephalon. The part of the reptilian dorsal cortex that receives sensory thalamic input is clearly homologous to parts of the mammalian sensory cortex. Whether the sensory recipient part of the dorsal ventricular ridge is homologous to the mammalian sensory temporal cortex or the lateral amygdala remains a controversial issue.

Hippocampus and Cortex

The reptilian medial, dorsal, and lateral cortices are homologous to the mammalian hippocampus,

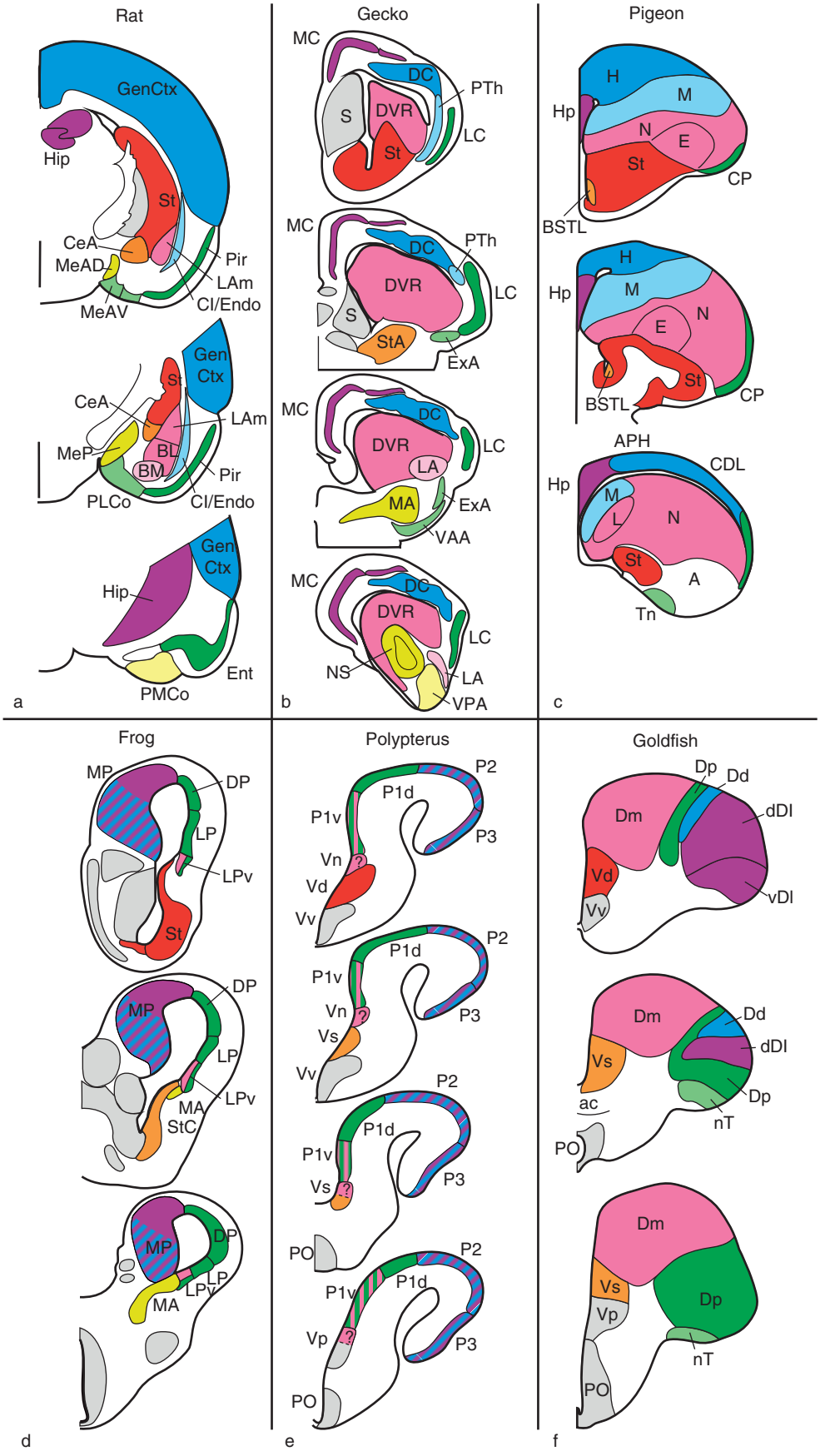

Figure 5 Continued

general cortex, and olfactory cortices, respectively (Figure 2(b)). The medial cortex can be further subdivided into two parts that closely resemble the mammalian dentate gyrus and Ammon's horn. Most cell types identified in the mammalian neocortex have been identified in the reptilian dorsal cortex, although the lamination pattern differs. A fourth pallial region, the dorsal ventricular ridge (DVR), was traditionally compared to sensory regions of the mammalian neocortex, because both receive massive sensory thalamic projections. However, a more recent hypothesis, that the DVR is homologous to the mammalian basolateral amygdalar complex, is more consistent with experimental evidence.

Amygdalar Nuclei and Pallial Thickening

Homologs of the five major amygdalar groups and the claustrum-endopiriform area present in mammals have been identified in reptiles, using similar topological, connectional, and developmental criteria and neurochemical expression patterns (Figure 5(b)): (1) the striatoamygdalar area, (2) nuclei that receive accessory olfactory bulb projections, (3) nuclei that receive olfactory connections and project to the hypothalamus, (4) pallial nuclei that receive ascending sensory input from the dorsal thalamus but lack hypothalamic projections, and (5) pallial nuclei that receive sensory information and project to the hypothalamus.

The striatoamygdalar area of reptiles is a subpallial group, located at the caudal pole of the striatum. It appears to be homologous to the mammalian central amygdalar nucleus.

The accessory olfactory bulb projects to the medial amygdalar nucleus, the ventral posterior amygdalar nucleus, and nucleus sphericus. The medial amygdalar nucleus and ventral posterior amygdala project to the ventromedial and lateral hypothalamic areas, although few neurons within nucleus sphericus project to the hypothalamus. Thus, these amygdalar groups appear to be homologous as a group to the mammalian posterior medial cortical amygdala and most of the medial amygdala.

The ventral anterior amygdala receives an olfactory bulb projection and projects to the ventromedial hypothalamus. It thus appears to be homologous to the mammalian anteroventral division of the medial amygdala and posterior lateral cortical amygdala.

Auditory, visual, and somatosensory nuclei in the caudal thalamus project to the anterior dorsal ventricular ridge. The sensory-recipient dorsal ventricular ridge is compared traditionally to parts of the mammalian neocortex, and more recently to the mammalian lateral amygdala. The anterior DVR and mammalian lateral amygdala differ from the neocortex in several significant ways. Both the DVR and lateral amygdala lack descending projections to the thalamus, brain stem, and opposite hemisphere, and are derived from the ventral pallium; the neocortex has these connections and is derived from the dorsal pallium. Thus, the best comparison for the anterior dorsal ventricular ridge is to the mammalian lateral amygdala.

On the basis of their connections and relative topology, the reptilian caudodorsal DVR, centromedial DVR, and lateral amygdalar nucleus appear to be homologous to the mammalian basolateral amygdala, ventrolateral part of the lateral amygdala, and basomedial amygdala, respectively. The sensory-recipient anterior DVR projects to the caudodorsal DVR, the centromedial DVR, and the lateral amygdalar nucleus. The caudodorsal DVR projects to the lateral hypothalamus and the other two groups project

Figure 5 Schematic overview of comparable telencephalic areas in (a) mammals, (b) reptiles, (c) birds, (d) amphibians, (e) primitive ray-finned fish, and (f) teleosts. Homologous areas are identified by the same color. Note that suggested homologies for fish are less well established than are those in tetrapods. (a) Rat: BL, basolateral amygdala; BM, basomedial amygdala; CeA, central amygdala; Ent, entorhinal cortex; GenCtx, general (nonolfactory, nonhippocampal) cortex; Hip, hippocampus; LAm, lateral amygdala; MeAD, anterodorsal division of medial amygdala; MP, medial pallium; Pir, piriform cortex; PLCo, posterior lateral cortical amygdala; PMCo, posterior medial cortical amygdala; St, striatum. (b) Gecko: DC, dorsal cortex; DVR, dorsal ventricular ridge; ExA, external amygdalar nucleus; LA, lateral amygdala; LC, lateral cortex; MA, medial amygdala; MC, medial cortex; nDB, nucleus of the diagonal band; NS, nucleus sphericus; PTh, pallial thickening; St, striatum; StA, striatoamygdalar area; VAA, ventral anterior amygdala; VPA, ventral posterior amygdala. (c) Pigeon: A, arcopallium; APH, area parahippocampalis; CP, cortex piriformis; E, entopallium; H, hyperpallium; L, field L; M, mesopallium; nDB, nucleus of the diagonal band; N, nidopallium; St, striatum; SM, stria medullaris tract; Tn, nucleus taenia. (d) Frog: Ac, accumbens; nDB, nucleus of the diagonal band; DP, dorsal pallium; LP, lateral pallium; LPv, ventral part of the lateral pallium; MA, medial amygdala; MP, medial pallium; St, striatum; StC, caudal striatum. (e) Polypterus: P1d, dorsal subdivision of pallial area 1; P1v, ventral subdivision of pallial area 1; P2, pallial area 2; P3, pallial area 3; PO, preoptic area; Vd, dorsal nucleus of ventral area; Vn, another nucleus of ventral area; Vs, supracommissural nucleus of ventral area; Vp, posterior nucleus of ventral area; Vv, ventral nucleus of ventral area. (f) Goldfish: ac, anterior commissure; Dd, dorsodorsal area; dDl, dorsal subdivision of dorsolateral area; Dm, dorsomedial area; Dp, dorsoposterior area; nT, nucleus taenia; PO, preoptic area; Vd, dorsal nucleus of ventral area; vDl, ventral subdivision of dorsolateral area; Vp, posterior nucleus of ventral area; Vs, supracommissural nucleus of ventral area; Vv, ventral nucleus of ventral area. (a–d) Reprinted with modifications from Bruce LL and Neary TJ (1995) The limbic system of tetrapods: A comparative analysis of cortical and amygdalar populations. *Brain, Behavior and Evolution* 46: 224–234, with permission from S Kager AG, Basel.

to the ventromedial hypothalamus. Stimulation of these DVR regions may elicit various motor responses, including head movements and escape responses.

The pallial thickening is situated at the rostrolateral pole of the DVR and appears to be homologous to the mammalian claustrum and endopiriform areas. The pallial thickening and claustro-endopiriform area are derived from the lateral pallium and have similar connections as well as neurochemical expression patterns.

Birds

The avian pallium includes the hippocampus, hyperpallium, cortex piriformis, and nidopallium (Figure 2(c)). These originate from the medial, dorsal, lateral, and ventral pallia, respectively.

Sensory Pathways to the Telencephalon

Birds have a main olfactory system that is remarkably similar to those in reptiles and mammals, although they lack a vomeronasal system (Figure 3(c)). Projections from the olfactory bulb include the olfactory tubercle, cortex piriformis, nucleus of the diagonal band, and nucleus taenia. Axons cross in the habenular commissure, as in reptiles but not mammals.

Auditory, visual, and somatosensory information is relayed through two separate thalamo-pallial pathways to the hyperpallium and the nidopallium (Figure 7). The auditory midbrain of birds, the torus semicircularis, projects mainly to nucleus ovoidalis in the thalamus, which in turn projects to the striatum and to field L in the nidopallium. Although an auditory thalamic nucleus that projects to the hyperpallium has not been described, acoustic responses

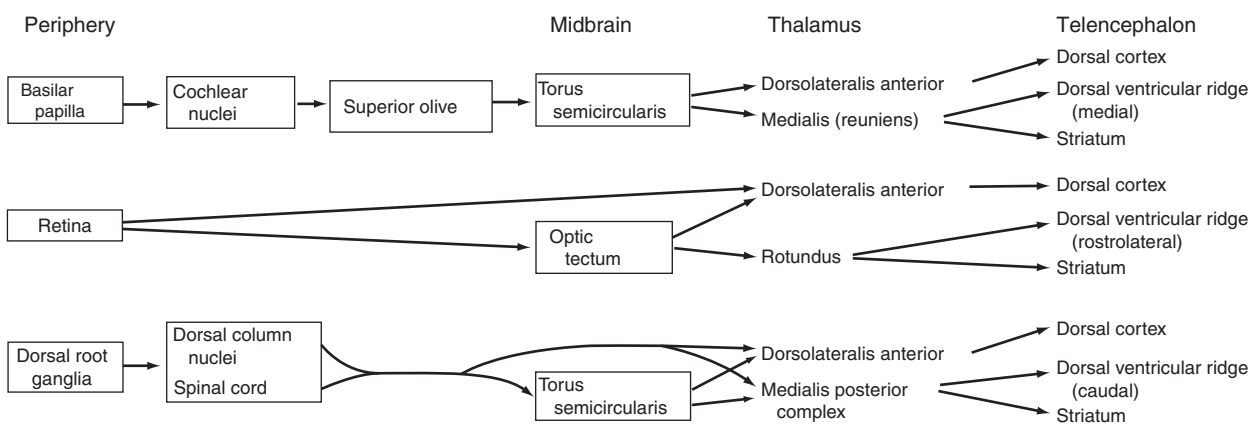

Figure 6 Comparison of the sensory pathways through the thalamus to the cortical and amygdalar targets in reptiles. Note that auditory (top), visual (middle), and somatosensory (bottom) pathways reach the cortex and amygdala through separate, parallel thalamic pathways.

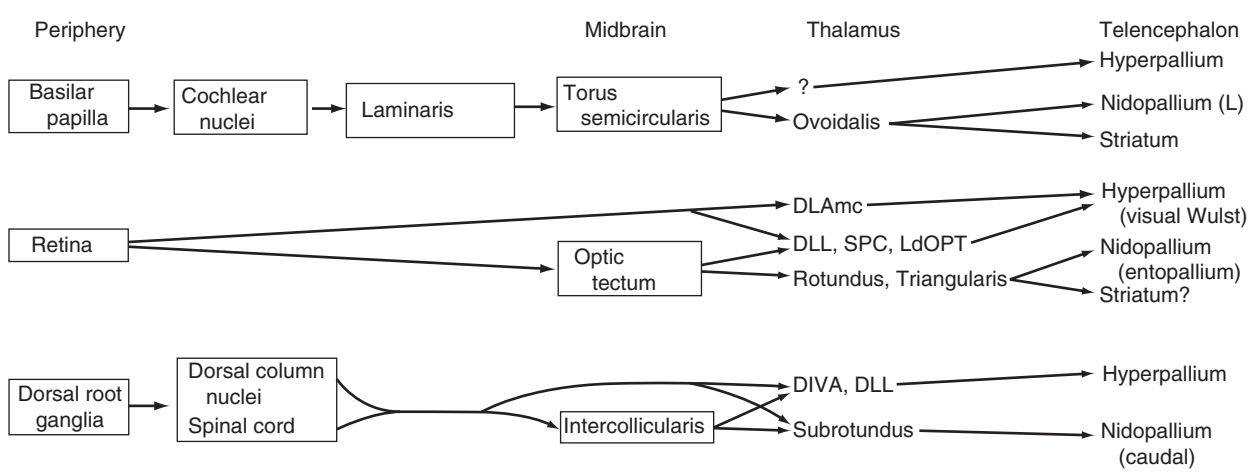

Figure 7 Comparison of the sensory pathways through the thalamus to the cortical and amygdalar targets in birds. Note that auditory (top), visual (middle), and somatosensory (bottom) pathways reach the cortex and amygdala through separate, parallel thalamic pathways. DLAmc, dorsolateralis anterior thalami, pars magnocellularis; DLL, dorsolateralis anterior thalami, pars lateralis; SPC, superficialis parvocellularis; LdOPT, lateralis dorsalis nuclei optici principalis thalami; DIVA, dorsointermedius ventralis anterior.

have been described in the medial cortex, probably corresponding to a part of the hyperpallium. Thus, there appear to be two distinct routes through which auditory information reaches the pallium.

The visual thalamic nuclei of birds may be grouped based on whether they receive only retinal projections (DLAmc, dorsolateralis anterior thalami, pars magnocellularis), projections from both the retina and optic tectum (DLL, dorsolateralis anterior thalami, pars lateralis; SPC, superficialis parvocellularis; LdOPT, lateralis dorsalis nuclei optici principalis thalami), or projections only from the optic tectum (rotundus and triangularis). Of these groups, the first two project predominantly to the hyperpallium (or visual Wulst) and the last projects to the entopallium within the nidopallium. A visual thalamic projection to the striatum has not been described in birds. Thus, there are two distinct routes through which visual information reaches the pallium.

The somatosensory region of the midbrain, nucleus intercollicularis, receives projections from the dorsal column nuclei, spinal cord, and trigeminal nuclei. Nucleus intercollicularis projects to three nuclei in the dorsal thalamus. Two of these, dorsointermedius ventralis anterior (DIVA) and dorsolateralis anterior pars lateralis (DLL), project to the hyperpallium, and the third, nucleus subrotundus, projects to the caudal nidopallium.

In conclusion, most sensory connections were conserved during the evolution from reptiles to birds, although the thalamic targets are elaborated to a greater extent in birds.

Hippocampus and Cortex

The avian hippocampus, hyperpallium, and cortex piriformis are homologous to the reptilian medial cortex, dorsal cortex, lateral cortex, respectively, and to the mammalian hippocampus, 'general cortex' and olfactory cortex, respectively (Figure 2(c)). The similarities between these domains have long been recognized. The homolog of the avian nidopallium is discussed below.

Amygdalar Nuclei and Mesopallium

Although several amygdalar regions can be recognized in birds, few studies have analyzed their projections to the hypothalamus, which are a major feature for identifying some amygdalar nuclei. Thus several regions are identified tentatively (Figure 5(c)), and a great deal more research is needed.

A homolog of the mammalian central amygdala is probably located in caudal parts of the avian striatum, although it has not been delineated at this time. Birds lack an accessory olfactory system, and thus the associated amygdalar nuclei appear to be absent. Nucleus taenia receives olfactory connections and projects to the hypothalamus. It thus appears to be comparable to the mammalian anteroventral division of the medial amygdala. Homologs of the posterior medial cortical amygdala may be present along the adjacent surface of the arcopallium, but have not been identified.

The sensory-recipient parts of the nidopallium appear to be homologous to the mammalian lateral amygdala and reptilian rostral DVR based on their similar topology, connections, and neurochemical expression patterns. Although in some laboratories it is maintained that a region with such a massive sensory input can only be homologous to sensory areas of the mammalian neocortex, many other characteristics differ. Thus this homology remains controversial.

An avian homolog of the deep amygdalar nuclei that project to the hypothalamus has not yet been identified. However, the caudolateral nidopallium and/or parts of the arcopallium seem to be likely possibilities. Topological, connectional, neurochemical expression characteristics strongly suggest that the avian mesopallium is homologous to the reptilian pallial thickening and the mammalian claustrum and endopiriform areas.

Amphibians

The amphibian pallium consists of the medial, dorsal, and lateral pallia (Figure 2(d)). Based on the topological similarities, these were traditionally compared to the mammalian hippocampus, neocortex, and olfactory cortex, respectively, but recent studies have challenged this convention, comparing the dorsal and lateral pallia to the mammalian olfactory cortex, and the medial pallium to both the hippocampus and neocortex.

Sensory Pathways to the Telencephalon

All amphibians have both a main olfactory system and most have a vomeronasal (accessory olfactory) system, although it is absent in some aquatic salamanders (Figure 3(d)). Their targets are similar to those in reptiles, mammals, and birds. The main olfactory bulb projections include the lateral and dorsal pallia, the ventral part of the lateral pallium, a rostral part of the medial amygdala, and the nucleus of the diagonal band. Axons project to the contralateral side through the habenular commissure, as in reptiles. The accessory olfactory bulb projects mainly to the medial amygdalar nucleus (formerly called the lateral amygdala). Axons project contralaterally through the anterior commissure and the habenular commissure, unlike in other tetrapods.

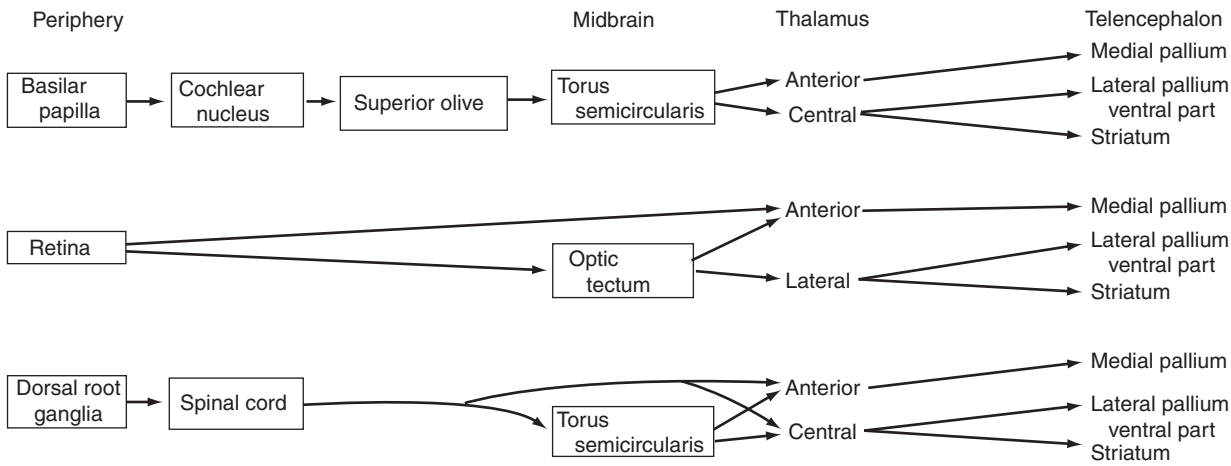

Figure 8 Comparison of the sensory pathways through the thalamus to the cortical and amygdalar targets in amphibians. Note that auditory (top), visual (middle), and somatosensory (bottom) pathways reach the cortex and amygdala through separate, parallel thalamic pathways.

Auditory information reaches the torus semicircularis in the midbrain, which projects strongly to the central thalamic nucleus and more weakly to the lateral and anterior thalamic nuclei. The auditory part of the anterior thalamic nucleus projects primarily to the medial pallium. Auditory information in the central and lateral thalamic nuclei is conveyed primarily to the striatum and sparsely to the ventral part of the lateral pallium.

Visual information reaches both the anterior and lateral thalamic nuclei. The anterior thalamic nucleus receives a projection from the optic tectum and directly from retina. In turn it projects to the medial pallium. The optic tectum also projects to the lateral thalamic nucleus, which projects to the striatum and the ventral part of the lateral pallium. Both the anterior and central thalamic nuclei receive input from the spinal cord and somatosensory part of the torus semicircularis. The anterior thalamic nucleus conveys this information mainly to the medial pallium, whereas the central thalamic nucleus conveys it mainly to the striatum and sparsely to the ventral part of the lateral pallium.

In summary, auditory, visual, and somatosensory information reaches the medial pallium and the ventral part of the lateral pallium through two distinct routes (Figure 8). Thus two thalamic–pallial pathways are present in amphibians, as in other tetrapods. These pathways appear to be phylogenetically ancient features, although the pallial targets are much more elaborated in reptiles, birds, and mammals.

Hippocampus and Cortex

Recent connectional and developmental analyses suggest that the ventral part of the lateral pallium is homologous to the reptilian dorsal ventricular ridge and the mammalian basolateral amygdalar complex (Figures 2(d) and 9). Furthermore, both the dorsal part of the lateral pallium and the dorsal pallium are homologous to the mammalian olfactory cortex and to the reptilian lateral cortices. Finally, the amphibian medial pallium is comparable both to the mammalian hippocampus and neocortex and to the reptilian medial and dorsal cortices. Further studies are needed to determine if the ancestral tetrapod had hippocampal and neocortical-like areas that were overlapping (as in amphibians) or segregated (as in reptiles, birds, and mammals). Some clues will undoubtedly come from studies of fishes.

Amygdalar Nuclei

Amygdalar areas comparable to those in other tetrapods have been identified in amphibia (Figure 5(d)). The caudal striatum contains neurons with descending projections to the hypothalamus and hindbrain, and thus appears to be homologous to the mammalian central amygdala. The amphibian medial amygdala, traditionally known as the pars lateralis of the amygdala, receives accessory olfactory projections and projects to the hypothalamus, and thus appears to be homologous to most of the mammalian medial amygdala. Both the dorsal and lateral pallia receive olfactory projections, but only the caudal, ventral part of the lateral pallium (LPv) projects to the hypothalamus. The caudal LPv thus appears, in part, to be comparable to the olfactory cortical amygdalar areas. The rostral LPv appears to be homologous to the mammalian lateral amygdala, which receives sensory thalamic input. In addition, both are derived from the ventral pallium. The caudal LPv receives sensory information from other telencephalic areas, in addition to a direct olfactory input, and it

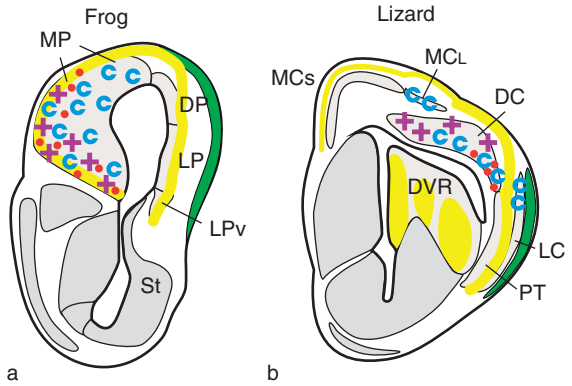

Figure 9 Pallial connections in (a) frog and (b) lizard. Schematic of a section through the right telencephalon shows the location of neurons that project to the thalamus (+), hypothalamus (red dots), and contralateral pallium (c). Terminal areas of olfactory projections (green) and thalamic nuclei (yellow) are also shown. Note that the characteristics of the lizard dorsal cortex are most similar to those of the frog medial pallium. (a) Frog: DP, dorsal pallium; LP, lateral pallium; LPv, ventral part of the lateral pallium; MP, medial pallium. (b) Lizard: DC, dorsal cortex; DVR, dorsal ventricular ridge; LC, lateral cortex; MCl, large-celled part of medial cortex; MCs, small-celled part of medial cortex; PT, pallial thickening. Reprinted from Bruce LL (2006) Evolution of the reptilian brain. In: Kaas JH (ed.) *Evolution of Nervous Systems, Vol. 2: Non-Mammalian Vertebrates*, pp. 125–156. Oxford: Academic Press, with permission from Elsevier.

is heavily interconnected with the hypothalamus. Thus, it appears to be homologous to the mammalian basolateral, basomedial, and ventrolateral part of the lateral amygdalar nuclei.

Homologs of the mammalian claustrum and endopiriform areas have not been identified in amphibians. Embryological studies suggest the claustrum and endopiriform originate from the lateral pallium. Thus it seems likely that, if present, these homologs are associated with the amphibian lateral pallium.

Fish

Among fishes, most of the data relevant to limbic system evolution come from one large group of ray-finned bony fishes, the teleosts. In the area dorsalis (presumed pallium) of adult teleosts four major areas are generally recognized and named on the basis of their topographical positions: dorsomedial (Dm), dorsodorsal (Dd), dorsolateral (Dl), and dorsoposterior (Dp). A fifth central zone (Dc) appears to contain subdivisions related to each of the other four zones. The teleost telencephalon develops by an outward bending, or eversion, of its dorsal portion, resulting in two thick, solid masses that flank a single ventricle (Figure 2(f)). The embryological details that relate this eversion to the organization of the adult teleost telencephalon are not well understood. A simple eversion would result in a mediolateral reversal of the four major pallial areas relative to their positions in the pallium of tetrapods. This would potentially place the pallial (basolateral) amygdala/ventral pallium in a dorsomedial position, followed by the lateral pallium, the dorsal pallium, and, finally, the medial pallium in a ventrolateral position. A major impediment to such a simple eversion model is that the main pallial target of the olfactory bulb in teleosts, Dp, appears to lie in this ventrolateral position where the medial pallium might be expected. Recently, several different lines of argument have suggested that Dp actually is located (or originates) between Dm and Dd, which would be topologically appropriate for the lateral pallium. If this is correct, the following working hypothesis can be suggested: Dm, Dp, Dd, and Dl are homologs of the pallial amygdala/ventral pallium, the lateral pallium, the dorsal pallium, and the medial pallium, respectively. However, more developmental and connectional data are needed to determine whether this hypothesis or an alternative – that the pallial areas of teleosts have evolved independently from those of the tetrapods – is correct. The area ventralis (presumed subpallium) is believed to contain homologs of the striatal and septal areas of tetrapods in dorsal and ventral positions, respectively, and homologs of subpallial amygdalar populations dorsally at more caudal levels.

Sensory Pathways to the Telencephalon

In all teleosts studied, the secondary olfactory projections terminate prominently in Dp, which is considered to be homologous to the lateral pallium of tetrapods (Figure 10(a)). Other prominent olfactory targets include nucleus taenia and the dorsal, supracommissural, and postcommissural portions of the area ventralis. Some or all of these latter centers have been interpreted as homologs of various portions of the amygdala of tetrapods. Among the other sensory modalities known to project to the teleost pallium, gustatory, auditory (and, to a lesser extent, lateral line), electrosensory (in some species), and visual (from tecto-recipient cells) inputs terminate in Dm. Visual (from tecto-recipient cells), auditory, and lateral line projections also terminate in Dl. These nonolfactory sensory inputs to the pallium arise largely not from what is clearly the dorsal thalamus but rather from an enigmatic group of diencephalic nuclei, the nuclei of the posterior tuberculum. It is uncertain whether these nuclei represent an hypertrophy of the posterior tuberculum in the teleosts or

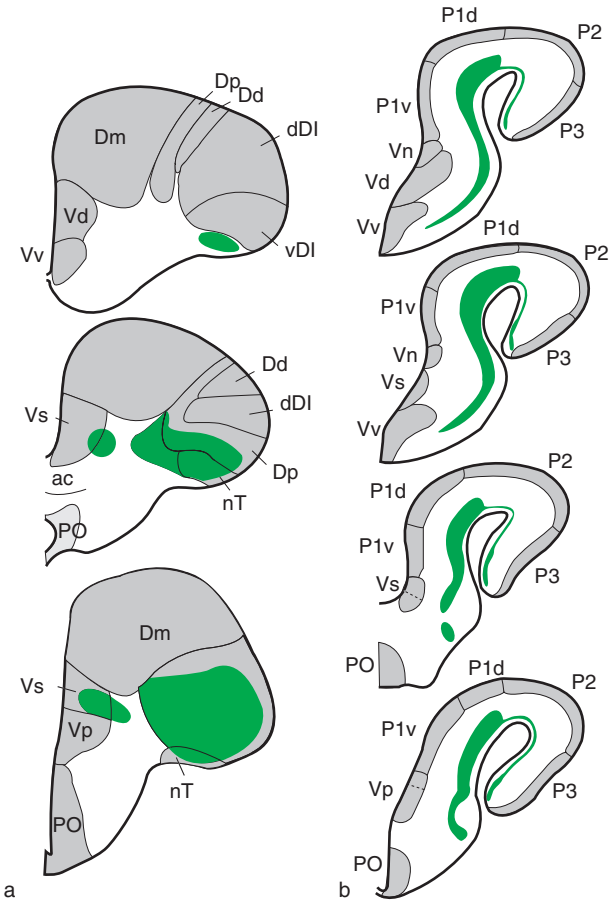

Figure 10 Comparison of projections of main olfactory bulb (green) in a teleost (the goldfish, *Carassius*) and a cladistian (the bichir, *Polypterus*). (a) Goldfish: ac, anterior commissure; Dd, dorsodorsal area; dDl, dorsal subdivision of dorsolateral area; Dm, dorsomedial area; Dp, dorsoposterior area; nT, nucleus taenia; PO, preoptic area; Vd, dorsal nucleus of ventral area; vDl, ventral subdivision of dorsolateral area; Vp, posterior nucleus of ventral area; Vs, supracommissural nucleus of ventral area; Vv, ventral nucleus of ventral area. (b) Polypterus: P1d, dorsal subdivision of pallial area 1; P1v, ventral subdivision of pallial area 1; P2, pallial area 2; P3, pallial area 3; PO, preoptic area; Vd, dorsal nucleus of ventral area; Vn, other nucleus of ventral area; Vs, supracommissural nucleus of ventral area; Vp, posterior nucleus of ventral area; Vv, ventral nucleus of ventral area.

whether some of them are homologous to parts of the dorsal thalamus of tetrapods.

Hippocampus and Cortex

The best candidate for a hippocampal (or medial pallial) homolog in teleosts appears to be Dl (Figure 2(f)). In addition to the aforementioned sensory afferents, Dl is reciprocally connected to the ventral part of the area ventralis (a presumed septal homolog) and has efferent connections to the preoptic area and the caudal hypothalamus. Behavioral studies have shown that lesions in Dl, especially its ventral division, result in deficits in spatial cognition. A small region of the pallium, Dd, that lies medially adjacent to Dl is in a position to be interpreted as the dorsal pallium (and some would lump with it the adjacent dorsal portion of Dl, dDl), but few data are known concerning its connections.

Amygdalar Areas

Although the working hypothesis suggests that all or part of Dm may constitute a pallial amygdalar/ventral pallial homolog, an alternative view that Dm (or part of it) is a dorsal cortical (isocortical) homolog has also been espoused (Figure 5(f)). If we assume that the sensory inputs to Dm from the nuclei of the posterior tuberculum are actually dorsal thalamic, we are still left with the question: are these nuclei comparable to the dorsal thalamic nuclei of tetrapods that project to the pallial amygdala/ventral pallium or to those dorsal thalamic nuclei that project to the isocortex? The topological position of Dm and its efferent connections to the hypothalamus and posterior tuberculum appear to be more consistent with the former view. In addition, lesions and electrical stimulation of Dm produce changes in aggressive and reproductive behavior, and lesions in Dm result in deficits in emotional memory. Finally, at least portions of Dm lie between the Emx-1 expression zone dorsally (characteristic of a pallial region) and the Dlx expression zone ventrally (characteristic of a subpallial region), further suggesting its similarity to the ventral pallium in tetrapods. However, the precise pattern of gene expression in this region – including that of Tbr (characteristically seen in the ventral pallium of tetrapods) – remains to be determined.

Finally, a few comments on comparative aspects of limbic telencephalic organization in another group of ray-finned fishes, the polypteriforms, are needed. The polypteriforms are the earliest derived actinopterygian lineage and thus occupy a phylogenetic position that is a key for determining the nature of the similarities found in the limbic telencephalon of the teleosts and the tetrapods (Figures 2(e) and 5(e)). As in all ray-finned fishes, the telencephalon of the polypteriforms develops by an eversion of its dorsal portion. However, this eversion is a simple outward bending in which the pallium remains thin and most neuronal populations remain in an unmigrated position. Thus there is a simple medial to lateral reversal of the pallium relative to that in tetrapods (most readily comparable to amphibians). Four pallial areas have been recognized – from medial to lateral – P1v, P1d, P2, and P3. As in the teleosts, the area ventralis is

believed to contain homologs of the striatal and septal areas of tetrapods in dorsal and ventral positions, respectively, and homologs of subpallial amygdalar populations dorsally at more caudal levels.

A simple topological comparison might nominate cell groups dorsal to the striatum as potential ventral pallial homologs. A small group, Vn (traditionally considered subpallial), and/or P1v (the most proximal portion of the traditional pallium) are thus potential candidates. This simple comparison would leave P1d as the lateral pallium, P2 as the dorsal pallium, and P3 as the medial pallium. The available data only partially support this hypothesis. The major pallial olfactory target is P1 (both portions), with a minor projection to P2 and P3 (**Figure 10(b)**). The other known sensory projection to the pallium is a visual input to P2 and P3 from tecto-recipient cells of the posterior tuberculum. Interestingly, both P2 and P3 are reciprocally connected to the ventral portion of the subpallium – the presumed septal homolog. The determination of the pallial–subpallial boundary and the further characterization of the similarities and differences of the four P divisions of the pallium in the polypteriforms are important for understanding the limbic system in this group. Moreover, they will provide strong guidance for deciding whether the limbic-like structures in telencephalon of the teleosts represent homologs of those in tetrapods or independently derived characters.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Structure and Circuitry in Rodents and Felines; Bird Brain: Evolution; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brains of Primitive Chordates.

Further Reading

Braford MR Jr. (1995) Comparative aspects of forebrain organization in the ray-finned fishes: Touchstones or not? *Brain, Behavior and Evolution* 46: 259–274.

Bruce LL (2006) Evolution of the reptilian brain. In: Kaas JH (ed.) *Evolution of Nervous Systems, Vol. 2: Non-Mammalian Vertebrates*, pp. 125–156. Oxford: Academic Press.

Bruce LL and Neary TJ (1995) The limbic system of tetrapods: A comparative analysis of cortical and amygdalar populations. *Brain, Behavior and Evolution* 46: 224–234.

Holmes PH and Northcutt RG (2003) Connections of the pallial telencephalon in the Senegal bichir, *Polypterus*. *Brain, Behavior and Evolution* 61: 113–147.

Meek J and Nieuwenhuys R (1998) Holosteans and teleosts. In: Nieuwenhuys R, ten Donkelaar HJ, and Nicholson C (eds.) *The Central Nervous System of Vertebrates*, vol. 2, pp. 759–938. Berlin: Springer.

Nikonov AA, Finger TE, and Caprio J (2005) Beyond the olfactory bulb: An odotopic map in the forebrain. *Proceedings of the National Academy of Sciences of the United States of America* 102: 18688–18693.

Northcutt RG (1995) The forebrain of gnathostomes: In search of a morphotype. *Brain, Behavior and Evolution* 46: 275–318.

Northcutt RG (2006) Connections of the lateral and medial divisions of the goldfish telencephalic pallium. *Journal of Comparative Neurology* 494: 903–943.

Price JL, Russchen FT, and Amaral DG (1987) The amygdaloid complex. In: Björklund A, Hökfelt T, and Swanson LW (eds.) *Handbook of Chemical Neuroanatomy, Vol. 5. Integrated Systems of the CNS, Part I*, pp. 279–388. Amsterdam: Elsevier Science.

Swanson LW and Petrovich GD (1998) What is the amygdala? *Trends in Neurosciences* 21: 323–331.

Turner BH and Herkenham N (1991) Thalamoamygdaloid projections in the rat: A test of the amygdala's role in sensory processing. *Journal of Comparative Neurology* 313: 295–325.

Wullimann MF and Mueller T (2004) Teleostean and mammalian forebrains contrasted: Evidence from genes to behavior. *Journal of Comparative Neurology* 475: 143–162.

Yamamoto N, Ishikawa Y, Yoshimoto M, et al. (2007) A new interpretation on the homology of the teleostean telencephalon based on hodology and a new eversion model. *Brain, Behavior and Evolution* 69: 96–104.

Evolution of Vertebrate Brains

A B Butler, George Mason University, Fairfax, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Overview

The brains of living vertebrates are a reflection of the very diverse niches occupied by the different species that comprise each major taxon (Figure 1) – agnathans (jawless vertebrates) and three radiations of jawed vertebrates: (1) the cartilaginous fishes (chimaeras and sharks, skates, and rays), (2) the ray-finned fishes (bony fishes), and (3) the sarcopterygian (fleshy-finned fish) radiation, which includes tetrapods (amphibians, mammals, reptiles, and birds). Within each of these major taxa, brain structure varies substantially, with some brains smaller relative to body size and less elaborate in terms of cytoarchitecture and others larger and more elaborate. The former can be referred to as type I brains (Figure 2) and the latter as type II (Figure 3). This type I–type II distinction is a matter of degree, and where the line is drawn is necessarily somewhat arbitrary. However, within each radiation, there are clearly some species with highly complex and enlarged brains relative to other species. Thus, this distinction is of heuristic value in appreciating the range of variation of brain evolution that has occurred within each major radiation.

That brain enlargement and elaboration has occurred four times independently presents a very different reality of how brain evolution has operated than is perceived in the widely held folk-belief of a sort of scale of nature, or Scala Naturae, that ranks all vertebrates along a simplistic scale. Instead, the picture now appreciated is a much more sophisticated and fascinating one in terms of both evolutionary history and the mechanisms by which it has proceeded. It is also important to note that the strategy of retaining a relatively simple brain in terms of cytoarchitecture, and one that is of modest size in relation to the body, is a successful one for many species, just as brain enlargement and elaboration is for other species. The variation in complexity and relative brain size that exists across all living vertebrate groups and individual species is a direct function of the available niches and the adaptations of various species that successfully occupy them.

The three major radiations of jawed vertebrates, or gnathostomes, each comprise diverse species that occupy equally diverse niches. The jawless vertebrates number fewer species and less diversity of niche but nonetheless exhibit marked diversity of brain size and degree of elaboration. Of the jawless vertebrates, or agnathans, lampreys have a relatively simple, or type I, brain (Figure 2) in terms of its number of neurons and their degree of migration away from the ventricular surface where they are generated during embryological development. In contrast, hagfishes have an enlarged and elaborated type II brain (Figure 3), characterized by considerably more neurons that are produced in and also migrate away from the ventricular surface and that form large aggregations of nuclei and/or laminated structures. Within cartilaginous fishes, chimaeras (ratfishes), and squalomorph and squantinomorph sharks have type I brains (Figure 2), while galeomorph sharks and skates and rays have type II brains (Figure 3). Within ray-finned fishes, reedfishes (Cladistia), sturgeons and paddlefishes (Chondrostei), gars (Ginglimodi), and the bowfin (Halecomorphi) have type I brains (Figure 2), while teleosts (Teleostei) have type II brains (Figure 3). Among the latter, species that exhibit complex social and territorial behaviors, such as many of the reef fish, have exceptionally complex and enlarged brains relative to other species. Within the sarcopterygian radiation that gave rise to land vertebrates, lungfishes (Dipnoi), the crossopterygian fish *Latimeria* (Actinistia), and amphibians have type I brains (Figure 2), while reptiles, birds, and mammals have type II brains (Figure 3).

The ratio of brain size to body size also varies considerably across different vertebrate groups. As shown in Figure 4, thanks to the copious data collected and analyzed by Harry Jerison, agnathans – both lampreys and hagfishes – have relatively low brain: body ratios, as do amphibians. Ray-finned fishes exhibit a wide range, with a few species overlapping the lower range of mammals. Reptiles fall within a modest range, with both mammals and birds having the largest brain:body ratios, but a number of cartilaginous fishes overlap the bird–mammal range. These data also demonstrate the wide range of variation within multiple groups of vertebrates that is the result of independent evolutionary processes.

Most of the diversity in brain complexity occurs in the dorsal part of the brain, which is derived from the alar plate during embryological development, rather than in the derivatives of the more ventrally lying basal plate. The latter gives rise to the motor nuclei for cranial nerves and reticular formation components and, while some variation occurs among these structures, they are relatively conservative in their organization and degree of development across vertebrates. The alar plate gives rise to the dorsal part of the neural tube, including the sensory nuclei of

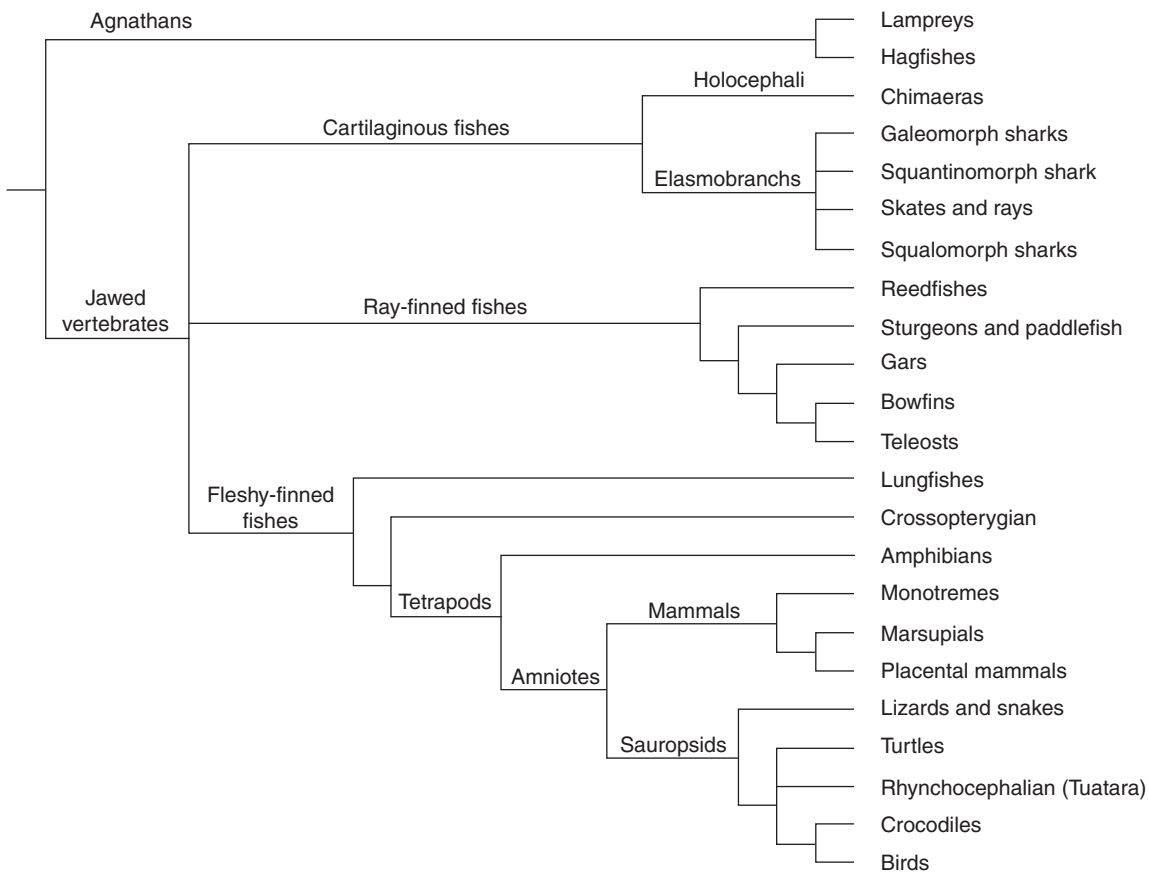

Figure 1 Cladogram showing the relationships of the four major groups of vertebrates – agnathans, cartilaginous fishes, ray-finned fishes, and the sarcopterygian radiation of fleshy-finned fishes.

cranial nerves, roof structures (cerebellum, midbrain tectum, much of the diencephalon, and all of the telencephalon), and cerebellar and basal-ganglia-related nuclei (inferior olivary nucleus, pontine nuclei, red nucleus, substantia nigra, etc.). Among the alar plate derivatives, the most variation occurs in the forebrain and particularly within the pallium, which is the dorsal part of the telencephalon. However, substantial variation also occurs in the roof of the midbrain, the tectum, and in the roof of the hindbrain, the cerebellum.

The simplistic Scala Naturae concept of evolution ranks organisms on an ascending scale that is presumed to reflect evolutionary history and that places humans in a position that is 'superior' to that of all other animals, as in fish-to-frog-to-rat-to-cat-to-monkey-to-human. While this concept is unfortunately widely and deeply embedded in the public consciousness, it is completely unsupported by the massive amount of data on evolution, not only for the brain but for all characters across the board. This concept not only errs in implying a direction to evolution, which it does not have except in terms of responses to externally driven selective pressures that vary over space and time, but it incorrectly implies that some existing taxa are identical to shared ancestral taxa, such as extant ray-finned fishes being identical to the ancestors of both themselves and land vertebrates.

To the contrary, both evolution and embryological development proceed from the general to the specific. In regard to evolution, the common ancestor of all vertebrates, for example, gave rise to the radiation of jawless vertebrates on the one hand and to that of the jawed vertebrates on the other; lampreys did not give rise to hagfishes. The common ancestor of tetrapods gave rise to amphibians on the one hand and to amniotes (reptiles, birds, and mammals) on the other; neither frogs nor newts gave rise to reptiles, birds, or mammals. Among amniotes, mammals actually appear in the fossil record before reptiles; reptiles did not give rise to mammals any more than mammals gave rise to reptiles. In regard to embryological development, it likewise generally proceeds from the general (common ancestral features) to the specific (specializations of the taxon) – for example, for a cat: from features common to vertebrates to those common to jawed vertebrates to those common to sarcopterygians (the fleshy-finned radiation) to

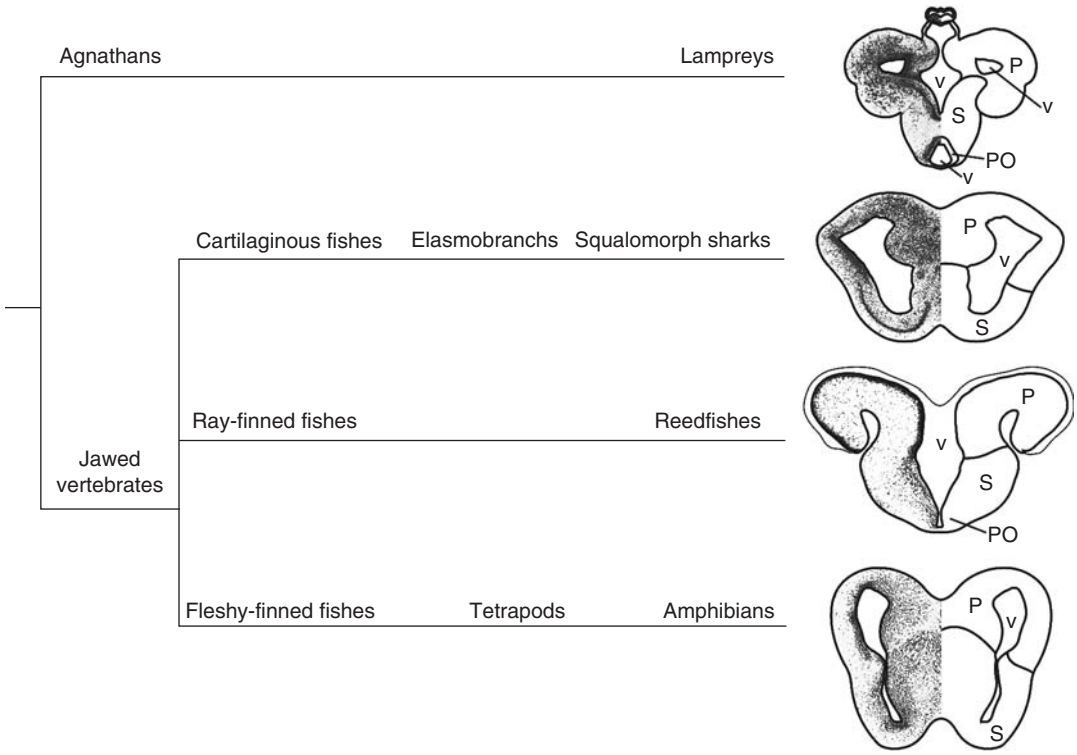

Figure 2 Simplified vertebrate cladogram showing examples of some of the taxa that have type I, cytoarchitectonically simple brains – lampreys (*Ichthyomyzon*), squalomorph sharks (the spiny dogfish shark *Squalus*), reedfishes (the bichir *Polypterus*), and amphibians (the bullfrog *Rana*). The photomicrographs of Nissl-stained hemisections with mirror-image line drawings shown at the right are through the telencephalon. D, diencephalon; H, habenula; P, pallium; Po, preoptic region; S, subpallium; v, ventricle. Most Nissl photomicrographs used in this figure were kindly provided by R Glenn Northcutt.

those common to tetrapods to those common to amniotes to those common to mammals to those specific to cats.

Understanding the basic process of evolution in these terms is essential to understanding brain evolution, since the structures present in the telencephalon of a modern reptile, for example, were not necessarily present in the common ancestor of mammals and reptiles. In order to reconstruct the evolution of a mammalian brain one must first deduce the condition of the ancestral brain, and only a comparative analysis of tetrapod brains can provide the clues necessary to do so. Features present in all tetrapod brains are those that can be hypothesized to have been present in the ancestral brain. Features present in only reptiles or only birds or only mammals are those features that can be hypothesized to be specializations of that particular taxon. Such hypotheses are not guaranteed to be correct, but they follow the principle of parsimony, which is accepted as the most conservative and probably most accurate method of dealing with the data. What is clearly established is that all taxa have their own specializations. Each taxon has a mix of primitive features – those that resemble what a distant ancestor had – and advanced features – those that are unique specializations. The challenge is to determine which features are shared across several taxa and therefore probably primitive in order to reconstruct a model of the common ancestor, then identify the specializations in each taxon, and then investigate how such specializations might have arisen.

Evolution of the Brain in Ancestral Vertebrates

Lancelets, or amphioxus, are small invertebrate chordates that are the closest living relatives of vertebrates. They have a specialization at the rostral end of the neural tube, called the cerebral vesicle, that qualifies as a brain in terms of its position and several regional features that correspond to those of vertebrate brains. From reconstructions of thin sections analyzed at the electron microscope level by Thurston Lacalli and his colleagues, it is now known that these features include (1) a single, midline group of pigment cells and associated neuronlike cells called the frontal organ, which appears to be the homolog of the paired, retinal eyes of vertebrates; (2) a more caudally lying structure called the lamellar body, homologous to the vertebrate pineal organ; (3) a ventrally situated

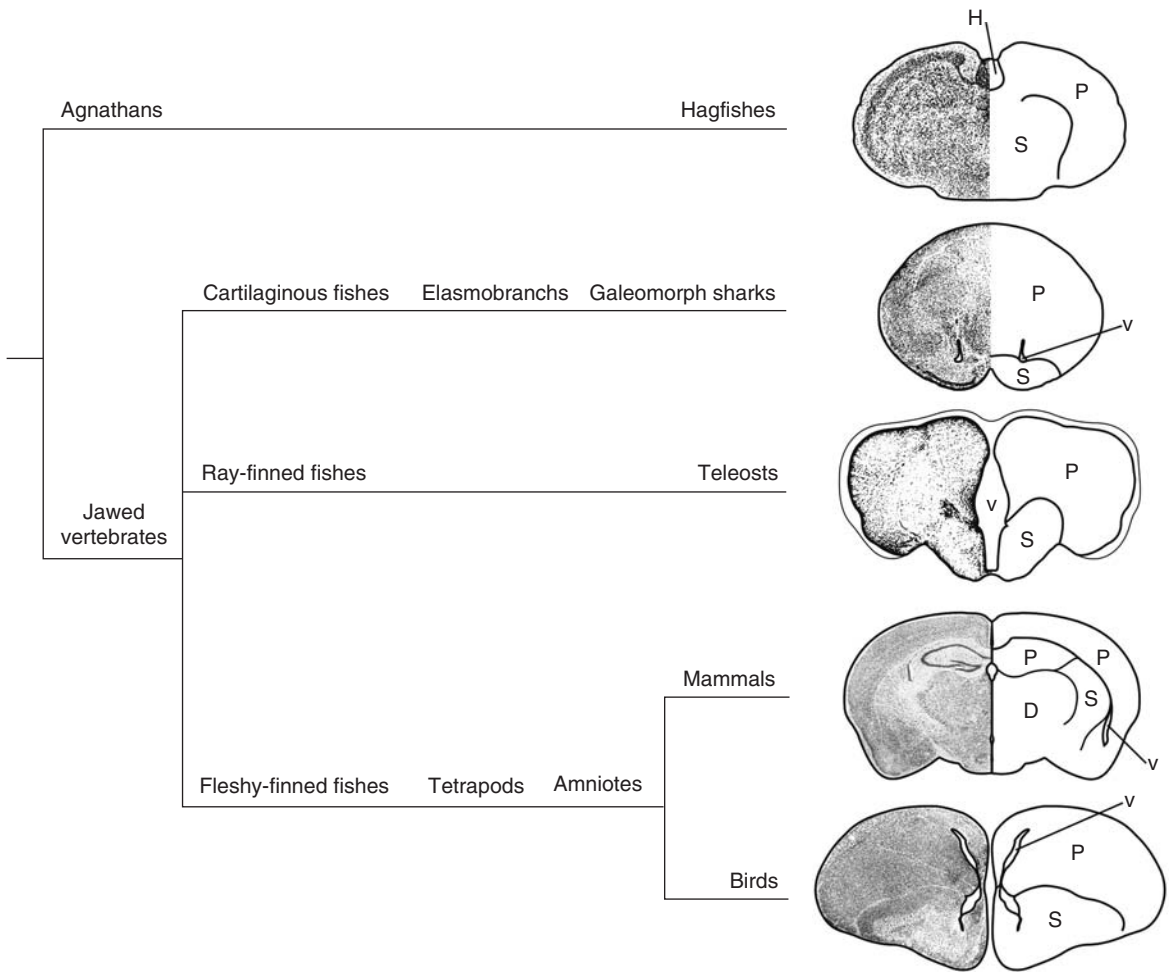

Figure 3 Simplified vertebrate cladogram showing examples of some of the taxa that have type II, cytoarchitectonically more complex and elaborated brains – hagfishes (*Eptatretus*), galeomorph sharks (the nurse shark *Ginglymostoma*), teleosts (the sunfish *Lepomis*), mammals (the mouse *Mus*), and birds (the pigeon *Columba*). The photomicrographs of Nissl-stained hemisections with mirror-image line drawings shown at the right are through the telencephalon. Most Nissl photomicrographs used in this figure were kindly provided by R Glenn Northcutt.

structure called the balance organ, homologous to at least part of the hypothalamus of mammals; and (4) associated infundibular cells, homologous to the vertebrate infundibulum (neurohypophysis). Recent findings from a cache of fossils of a small creature named *Haikouella lanceolatum*, discovered in Haikou, China and analyzed by Jon Mallatt and Jun-Yuan Chen, in addition to an array of other fossil evidence, have allowed several hypotheses on early brain evolution in the vertebrate line to be substantially clarified. *Haikouella* fossils exhibit a number of relevant features, including (1) an enlarged brain with a brain: body ratio comparable to that of lampreys, (2) brain components that consist mainly of a diencephalon and hindbrain, and (3) paired eyes. The scenario presented here may be modified in the future depending upon further fossil evidence but seems the most likely one based on the multiple lines of evidence available to date.

At the origin of vertebrates, the nervous system would have had, like amphioxus, a rostral, forebrain region that contained at least the light-receptive and hypothalamic components of the diencephalon. It lacked a telencephalon and olfactory structures. The light-receptive components comprised paired eyes, like *Haikouella*, and also a pineal organ. While still small and relatively simple in construction, the paired eyes likely had a vertebrate retinal structure. The hypothalamic component included connections with an infundibulum. As evolution proceeded, subsequent gains included a host of features associated with the neurogenic placodal tissues of head and the neural crest tissue of the body and head. Together, these tissues generate the motor neurons of the autonomic

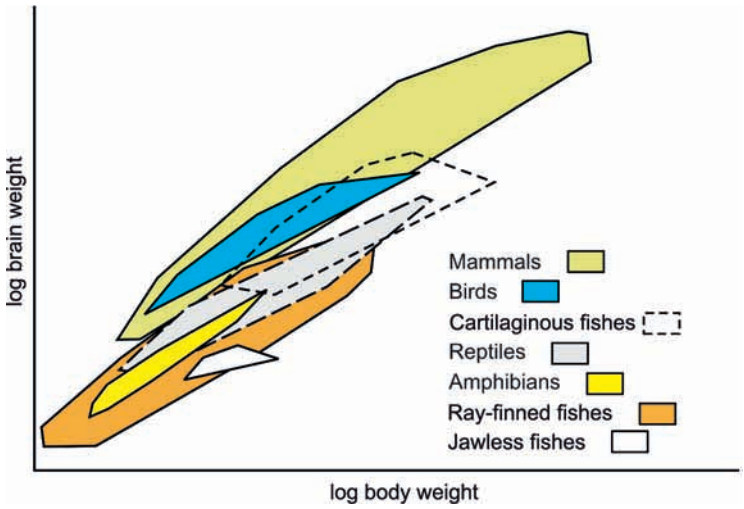

Figure 4 Convex polygons that enclose the log values for brain weight plotted against log values for body weight for a number of vertebrate taxa. Data kindly provided by Harry Jerison. Note that the polygon for birds almost completely overlaps that for mammals. The brain:body ratios for some groups of birds, including corvids (crows) and parrots, overlap the lower range of primates. Note also that the upper border of the polygon for ray-finned fishes also overlaps the lower border of the polygon for mammals and that the polygon for cartilaginous fishes substantially overlaps the polygons for both birds and mammals. These relationships demonstrate that brain elaboration and enlargement has occurred independently multiple times. Not all type II brains also show marked enlargement of brain:body ratio, however, since while hagfishes have type II brains, their brain:body ratio falls within the polygon for jawless fishes, which overlaps only the lower range for ray-finned fishes and is considerably below the polygons for cartilaginous fishes, birds, and mammals.

nervous system, the sensory neurons of cranial nerves I, V, VII, VIII, IX, and X, and the sensory neurons of the spinal nerves. Additionally, neural crest generates a host of other tissues, including the dermal bones and sensory capsules of the skull and the meninges that surround the central nervous system. Thus, with the bloom of these tissues at some point along the early vertebrate lineage, many of the structures that form the vertebrate head were gained, along with the peripheral nervous system of the head and body. Continued elaboration of the brain and spinal cord, in terms of sensory and motor nuclei and cell columns, the reticular formation, ascending and descending pathways, the telencephalon with olfactory bulbs, the dorsal and ventral thalamic components of the diencephalon, and a definitive midbrain region would likewise have been gained during this transition.

Type I Brains

As noted previously, brains that exhibit relatively little migration of neurons away from the ventricular surface and relatively simple cytoarchitecture, as compared to more complexly elaborated brains, can be classified as type I (Figure 2) for heuristic reasons. Taxa that have such brains include lampreys, some cartilaginous fishes (chimaeras, squantinomorph sharks, and squalomorph sharks), some ray-finned fishes (nonteleosts), and, among the sarcopterygian radiation, lungfishes, the crossopterygian fish *Latimeria*, and amphibians. Only a brief overview of brain organization in these taxa with selected highlights can be presented here.

Lampreys

Lampreys have most of the cranial nerve and other central nervous system components that characterize vertebrate brains in general, although there are some minor differences in the innervation pattern of the extraocular muscles. Among alar plate derivatives, a cerebellum is present but of miniscule size, consisting only of a small region of gray matter in the dorsal part of the rostral medulla. The midbrain tectum is present, with both caudal, auditory and rostral, visual and somatosensory portions. The telencephalon is of very modest size, with cell bodies migrated away from the ventricular surface to only a limited extent (Figure 2). Pallial regions homologous to those of other vertebrates have not yet been identified with certainty. A striatopallidal complex is present in the subpallium (also an alar plate derivative).

Type I Cartilaginous Fishes

Type I cartilaginous fishes comprise chimaeras (ratfishes), squantinomorph sharks, and squalomorph sharks. The latter have been most studied, particularly the spiny dogfish shark *Squalus acanthias*. In general,

the brains of these taxa exhibit many of the cranial nerve and reticular formation components of other vertebrates, a well-developed cerebellum, a moderately developed midbrain roof, an unremarkable diencephalon, and a moderately developed telencephalon. The latter in particular is less extensively populated with neurons (Figure 2) than in type II cartilaginous fishes but contains all the basic pallial and subpallial components of other vertebrates.

Type I Ray-Finned Fishes

All ray-finned fishes, both type I and type II, undergo an almost unique process of telencephalic pallial development called eversion. The only other vertebrate that shares this developmental process, and that to a lesser extent, is the crossopterygian fish *Latimeria*. Rather than undergoing an outpouching process of hemispheric expansion during development, as occurs in other vertebrate groups, eversion causes the most medial part of the telencephalic pallium to lift upward and outward, eventually arching laterally. In evagination, the pallium enlarges like a balloon expanding, with its most medial part remaining in a medial position, the dorsal part arching dorsally, and the most lateral part remaining in its lateral position. In contrast, during eversion, the medial pallium comes to lie in the most lateral position, and the originally lateral pallium remains most medially and ventrally. The geometry of eversion is akin to that of a person doing a back bend, such that the head, arms, and upper torso curve upwards and then backwards. Thus, as shown for both the reedfish telencephalon shown in Figure 2 and that of a teleost telencephalon shown in Figure 3, the ventricular surface lies dorsalmost over the dorsal aspect of the everted pallia, and the ventricular cavity extends laterally over each side to the point where the ependyma (shown as a thin line) attaches.

The brains of type I ray-finned fishes (reedfishes, sturgeons and paddlefishes, gars, and the bowfin *Amia*) are generally similar in plan and in their component parts to brains of other vertebrates, with similar cranial nerve and reticular formation components: cerebellum, midbrain roof, and forebrain components. As with other ray-finned fishes and, generally speaking cartilaginous fishes, the dorsal thalamus contains three nuclei, the more caudal two of which receive their predominant input from the midbrain roof, and the rostral one receives its predominant inputs directly from the retina, somatosensory system, and/or other sources, without relay first through the midbrain roof. The midbrain roof-recipient nuclei can be grouped as the collothalamus and the rostral dorsal thalamic component(s) can be grouped as the lemnothalamus – a basic feature of organization that applies at least to all jawed vertebrate brains.

Both ray-finned fishes and cartilaginous fishes have additional nuclei that are involved in the relay of ascending sensory information to the telencephalon, the preglomerular nuclear complex (or individual nuclei homologous to its various components), but these are better developed in type II brains than in type I brains. In reedfishes, a nucleus medianus is present that relays visual information to the pallium, and in gars and the bowfin *Amia*, a preglomerular nuclear complex is present. The latter generally relays lateral line and/or gustatory inputs to the telencephalon.

Lungfishes, Crossopterygian, and Amphibians

Less information is available for lungfish brains than for the brains of amphibians and most other major vertebrate groups, and very little information is available for the crossopterygian *Latimeria*, since this animal survives only at great ocean depths under conditions of high pressure. Nonetheless, the anatomical organization of lungfish and crossopterygian brains appears to be generally similar to that of amphibians. As already noted, the telencephalic pallium of *Latimeria* undergoes partial eversion during embryological development, whereas evagination occurs in lungfishes and amphibians.

Amphibian brains exhibit the basic structure and components of other vertebrate brains. Of note in terms of their type I development, more production and migration of neurons occurs in the medial (hippocampal) part of their pallium than in other pallial regions (Figure 2). The medial pallium is the major site to which the rostral part of their dorsal thalamus, nucleus anterior, projects. The latter nucleus constitutes the lemnothalamus, but, unlike the situation in amniotes, it does not receive direct input from either the retina or the somatosensory system (dorsal column nuclei), instead having these inputs relayed to it through the ventral thalamus. Also, the ventral thalamic projections to nucleus anterior are thought to be inhibitory, rather than excitatory as the direct inputs would be. Whether this condition represents the ancestral tetrapod condition or a secondary specialization within the amphibian lineage remains to be determined.

The collothalamus of amphibians is also different to a marked degree in its pattern of projections from the collothalamus of amniotes. These nuclei, which receive their predominant visual, somatosensory, and auditory inputs via relay through the midbrain roof, project almost exclusively to the striatum in the subpallium, with only a few fibers reaching the lateralmost aspect of the pallium. In contrast to the

medial pallium, the rest of the pallium in amphibians exhibits almost no migration of neurons away from the ventricular surface (Figure 2). Also in contrast to this pattern of projections, as will be noted subsequently, in amniotes, collothalamic projections give off collaterals to the striatum, but their heaviest targets are within the dorsolateral and lateral (lateral and ventral) portions of the pallium.

Type II Brains

As noted previously, brains that exhibit a relatively large production of neurons combined with substantial migration of most neurons away from the ventricular surface and relatively elaborate cytoarchitecture, as compared to the more simply constructed type I brains, can be classified as type II (Figure 3) for heuristic reasons. Taxa that have such brains include hagfishes, some cartilaginous fishes (galeomorph sharks, skates, and rays), many ray-finned fishes (telosts), and, among the sarcopterygian radiation, amniotes (mammals, reptiles, and birds). Only a brief overview of brain organization in these taxa, with selected highlights, can be presented here.

Hagfishes

The brain of hagfishes is largely terra incognita, due to the difficulty of obtaining and working with these animals (which are covered with a copious quantity of secreted mucus) in a laboratory setting. Also, the substantial amount of neuronal production and migration during development results in a highly complex brain in terms of its cytoarchitecture, and many features of the latter do not clearly correspond to those of other vertebrates. A telencephalon with a highly laminated pallium is present (Figure 3), and those pallial lamina are in receipt of olfactory information. A diencephalon is present with recognizable divisions of epithalamus (habenula), dorsal and (perhaps) ventral thalami, and hypothalamus. Eyes, however, are extremely reduced, and extraocular muscles are absent, with corresponding reduction of retinal projections and oculomotor nuclei. No cerebellar tissue has been identified. Instead of being similar to ancestral jawless vertebrates, a good case can be made for hagfish neural (and other) features being the consequence of evolutionary specializations.

Type II Cartilaginous Fishes

Galeomorph sharks, skates, and rays have exceptionally well-developed telencephalons, both in terms of size relative to their body size and in terms of complexity of cytoarchitecture (Figure 3). Less is known about their various neural systems and major pathways than for some other vertebrate groups – particularly ray-finned fishes and tetrapods – due to the practical difficulties of working with many of these animals in a laboratory setting. One of the major breakthroughs in comparative studies of the past century was that achieved by Sven Ebbesson, Lennart Heimer, Dolores Schroeder, and their colleagues with their work on nurse shark (*Ginglymostoma cirratum*) behavior and sensory pathways, demonstrating that (1) olfactory projections are limited to a relatively small lateral pallial area (rather than being to most or all pallial areas as previously assumed), (2) two different visual pathways to the pallium are present, one lemnothalamic (via direct projections from the retinorecipient nucleus anterior) and one collothalamic via midbrain roof relay to the dorsal thalamus and thence to the pallium, and (3) visual form discrimination requires participation of these telencephalic visual areas rather than just that of the midbrain roof. These findings, along with those of Harvey Karten and his colleagues on bird sensory system pathways and telencephalic organization (see later, the section titled 'Forebrain evolution across amniotes'), precipitated a new era of comparative insights into the evolution of vertebrate brains.

Teleosts

The range of variation in brain structure across teleosts, an exceptionally large vertebrate taxon, is extremely broad. As with cartilaginous fishes and other ray-finned fishes, teleosts exhibit many of the basic cranial nerve, reticular formation, cerebellar, midbrain, and forebrain components of the brain typical of vertebrate brains in general. As in cartilaginous fishes, olfactory projections are restricted to a small part of the pallium, leaving large areas for other sensory system inputs. Various groups and species of teleosts also exhibit a large number of specializations, mostly of alar plate derivatives but some of the basal plate as well, a sampling of which can be presented here.

A derivative of the basal plate, the oculomotor nucleus, has a very specialized structure in the stargazer, one of a number of fishes that are capable of producing an electric discharge. Part of the oculomotor nuclear complex in this fish is called the electromotor division, and it innervates specialized electric organs that are developmentally derived from the extraocular muscle anlage. Both alar and basal plate derivatives are involved in the complex circuitry that supports the sonic motor systems of a number of fishes, such as the plainfin midshipman (*Porichthys notatus*). Other cranial nerve specializations include a massive expansion of the taste-receptive nuclei in

cyprinids (which include goldfishes) and silurids (catfishes). Rather than consisting of the relatively small gustatory nucleus as in mammals and most other vertebrates, the taste-receptive cell groups in these fishes are hypertrophied into large vagal, glossopharyngeal, and facial lobes, with over a dozen cellular layers and a highly complex pattern of connections. Likewise, the lateral line system is particularly elaborated in electroreceptive teleost fishes, with mormyrids being an extreme example. These fishes have an electroreceptive cerebellum that is massively expanded relative to all of the rest of the brain and also extreme in its degree of complexity of cytoarchitecture and connections. Mormyrids use their electrosensory communication system for individual identification to conspecifics and for highly complex behavioral feats, including group cooperation in the building of underwater nests.

In addition to these cranial nerve and hindbrain specializations, teleosts display a wide range of variation in the complexity and degree of elaboration of their telencephalons (**Figure 3**). As described above for type I ray-finned fishes, the telencephalon undergoes a process of eversion during embryological development, resulting in a reversal for the adult positions of the originally medial and lateral pallial regions. Recent studies by Cosme Salas and his colleagues have demonstrated that the originally medial (hippocampal) part of the pallium in goldfish is involved in the same types of functions as the amniote hippocampus, particularly spatial mapping. In contrast, the originally lateral (amygdalar) part of the pallium in goldfish is involved in emotional, associational learning, such as the pairing of a light stimulus with an aversive, moderate electric shock. Thus, despite the process of eversion and also despite markedly different cytoarchitecture, the basic functions of pallial regions in goldfish are remarkably similar to their homologs in land vertebrates.

Amniotes

Mammals The earliest mammals appear in the fossil record slightly before the earliest reptiles. Rather than thinking of the common ancestor of all amniotes as a stem reptile, which implies reptilian structures in the brain as well as elsewhere, it is correct to think of the common ancestor as a stem amniote. Some of the most salient features of the brains of sauropsids (reptiles and birds) and those of mammals represent divergences from the common ancestral condition rather than sequential evolution of either the extant mammalian or sauropsidian condition to the other.

Cranial nerve specializations do not only occur in fishes, such as those already noted. Among mammals, a remarkable adaptation of the trigeminal nerve also occurs in monotremes – platypus and echidna. Both, but better developed in the platypus, have electroreceptors – located on the snout of the echidna and the duckbill of the platypus. These are not innervated by the lateral line system, which is present only in anamniotes, but rather by the trigeminal nerve. The platypus in particular uses these electrosensory inputs to search for prey under debris in the muddy stream beds that it inhabits. Trigeminal-associated specializations also occur in mammals such as the star-nosed mole, innervating its highly specialized Eimer's organs that cover fleshy tentacles around the oral region, and such as the barreletts-barreloids-barrels cellular formations that are present within somatosensory nuclei and cortex of rodents and some marsupials with mustacial vibrissae.

In addition to a host of features similar to those of other vertebrates, the brains of mammals vary somewhat across their taxa (monotremes vs. marsupials vs. placental mammals) and also share several salient specializations, the most prominent of which is of course the six-layered and massively developed neocortex (**Figure 3**). Likewise, the mammalian dorsal thalamus is the most highly developed and elaborated of any major taxa, in placentals containing numerous sets of multiple nuclei each. It is likely that, as in extant monotremes and to a lesser degree in extant marsupials, the earliest mammals favored a substantial expansion of the lemnothalamus (anterior, medial, rostral intralaminar, and ventral nuclear groups and the dorsal lateral geniculate nucleus) and its pallial targets. This would be in accord with the importance for early, small-sized, crepuscular mammals of depending on limbic system functions such as spatial mapping and on their lemnothalamic visual pathway to striate cortex. Subsequently, within marsupials and to a greater degree in multiple orders of placental mammals, the collothalamic (medial geniculate body and the posterior intralaminar, posterior, and LP/pulvinar nuclear groups) pathways were markedly expanded, including the collothalamic-recipient extrastriate, auditory, and secondary, etc., somatosensory cortical areas as well as the higher order association cortical areas of the temporal and parietal lobes. Likewise, a pallial part of the amygdala in mammals, its basolateral component, is in receipt of ascending collothalamic projections.

Sauropsids Just as mammals diverged from the stem amniote stock, so did the sauropsid line that gave rise first to reptiles and, subsequently, birds, which actually are a specialized group of reptiles, just as are the other extant groups of this major taxon. In fact, recent genetic analyses have demonstrated that the thecodonts (crocodiles and birds), along with

turtles and the rhynchocephalian lizardlike animal *Sphenodon*, are clustered as a monophyletic group, while the squamates (lizards and snakes) constitute the outgroup to them (Figure 1).

Among sauropsids, as in mammals and other vertebrates, cranial nerve specializations occur. In two families of snakes, boids and crotalids, sensory pits are present on the head region that can detect infrared radiation. These pits are innervated by the trigeminal nerve, and, in crotalids, the trigeminal projections terminate within a brain stem nucleus (the lateral trigeminal nucleus) unique to them. The infrared sensory information is relayed to the optic tectum (directly in boids and via another unique nucleus, the nucleus reticularis caloris, in crotalids), where it is used in conjunction with visual input to locate and accurately strike prey. Yet another specialization of the trigeminal system – the ability of magnetodetection – occurs in birds and is shared with like specializations recently found in trout, salamanders, and loggerhead sea turtle hatchlings. Naturally available magnetic stimuli are used, often in conjunction with visual clues, for navigation and orientation to the earth's magnetic field.

The telencephalic pallium of both birds (Figure 3) and reptiles contains a dorsally lying region – the Wulst of birds and the general, or dorsal, cortex/pallial thickening of reptiles – that is well established as the homolog of at least some of the more medial, lemnothalamic-recipient regions of mammalian neocortex, that is, the primary somatosensory-motor and visual cortices. Likewise, the lemnothalamus of sauropsids contains a number of individual nuclei that are homologous as either a field or one-to-one basis with various lemnothalamic nuclei and/or nuclear groups of mammals. The pallium of sauropsids also contains a structure called the dorsal ventricular ridge (DVR, also called the nidopallium and mesopallium in birds) that receives several ascending collothalamic pathways – visual, auditory, and somatosensory – and, likewise, the sauropsid collothalamus contains nuclei that are homologous to the set of collothalamic nuclei of mammals.

Within the early sauropsid line, a sequence of dorsal thalamic and pallial elaboration occurred that was opposite to that in mammals. In early sauropsids, elaboration of the collothalamus and its pallial targets was favored over that of the lemnothalamus. In some extant reptiles, such as turtles, some lizards, snakes, and the rhynchocephalian *Sphenodon*, the DVR is structured as a corticoid (layered) border of cells that surrounds a central region that hasscattered neurons but is mostly neuropil. Ascending thalamic afferents terminate on the dendrites of neurons of the corticoid band that extend into the neuropil region. In other extant reptiles – particularly crocodiles and some lizards – and birds, the DVR is a greatly expanded structure, with much more extensive and migrated populations of neurons – an exaggeration of the type II brain category. In these taxa, the DVR is composed of a number of nuclear regions, some of which receive the various ascending collothalamic projections and others of which are expanded areas of association pallium. Within the sauropsid line, birds have secondarily expanded the lemnothalamus and its pallial target area, the Wulst, although still not to as great a degree as the collothalamic-DVR systems.

Forebrain evolution across amniotes Controversy is ongoing as to the homology of the DVR with mammalian pallial components. The pioneering work of Harvey Karten and his colleagues that began in the 1960s established the pallial nature of this part of the telencephalon in birds and traced the ascending auditory and visual pathways to it, strongly indicating homology of these areas to the auditory and extrastriate visual cortices of mammals (specifically of the DVR thalamo-recipient cell populations and their efferent target cell groups to the cell populations within layer IV and the infragranular layers of neocortex, respectively). Similar pathways were then discovered in a number of reptilian taxa.

Recent embryological studies of gene expression patterns have yielded data that suggest a different correspondence – that of the sensory-recipient DVR to the basolateral amygdala of mammals. Neither of these one-to-one comparisons yet accounts for the evolution of the other structure, however; that is, the basolateral amygdala hypothesis does not yet satisfactorily explain the evolutionary derivation of the lateral neocortex, and vice versa. An alternative hypothesis that attempts to do so posits that the single DVR structure present in sauropsids is homologous, as the single derivative of a developmental field, to both of these structures in mammals. Still others argue that the pallial regions in sauropsids and mammals are so independently derived that sensible comparisons cannot be made at all. The argument also embraces the particular homologous relationships of the individual collothalamic nuclei between sauropsids and mammals and borders on becoming circular in that regard.

Whatever the outcome of the continuing research in this area in regard to phylogenetic homology of pallial areas and dorsal thalamic nuclei, it is now clearly established that the neocortex in mammals and the various pallial areas in birds, which exhibit dramatic differences in their cytoarchitecture

(Figure 3), both support highly complex cognitive abilities and may thus also both support high levels of consciousness, including working memory-related attributes such as anticipation of future events, sense of self, and theory of mind. A bounty of recent studies in birds, creatively designed to ask experimental questions of the birds in behavioral paradigms that are relevant to their response abilities, indicate extremely high levels of cognitive abilities in at least some taxa, particularly corvids (crows) and parrots. Since such abilities in humans are correlated with a high degree of consciousness, it is parsimonious to posit a similar correlation in birds. One important caveat is that high levels of consciousness also may be present but currently undetectable by us in other animal taxa, including reptiles, amphibians, and fishes, but, at least in birds, the high cognitive abilities that we can now appreciate indicate the capacity for high-level consciousness as well. A major question then becomes what role the cytoarchitecture of neocortex (six cell layers based on a pyramidal cell motif) plays in mammalian cognitive abilities and the generation of high-level consciousness and why and how a dramatically different cytoarchitecture of the DVR (nuclear) is able to support similarly high cognition and a possibly correlated high degree of consciousness as well.

See also: Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brains of Primitive Chordates; Gene Expression in the Evolution of the Human Brain.

Further Reading

Allman JM (1999) *Evolving Brains*. New York: Scientific American Library.

Bock GR and Cardew G (eds.) (2000) *Evolutionary Developmental Biology of the Cerebral Cortex, Novartis Foundation Symposium 228*. Chichester: Wiley.

Butler AB (2006) The dual elaboration hypothesis of the evolution of the dorsal thalamus. In: Krubitzer LA and Kaas JK (eds.) *Evolution of Nervous Systems, Vol. IV: The Evolution of Nervous Systems in Mammals*. Amsterdam: Elsevier.

Butler AB and Hodos W (2005) *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation*, 2nd edn. Hoboken, NJ: Wiley.

Kaas JH (1995) The evolution of isocortex. *Brain, Behavior and Evolution* 46: 187–196.

Kaas JH (ed.) (2006) *Evolution of Nervous Systems, Vols. I–V*. Amsterdam: Elsevier.

Mallatt J and Chen J-Y (2003) Fossil sister group of craniates: Predicted and found. *Journal of Morphology* 258: 1–31.

Nieuwenhuys R, ten Donkehaar HJ, and Nicholson C (1998) *The Central Nervous System of Vertebrates*. Berlin: Springer.

Northcutt RG (1995) The forebrain of gnathostomes: In search of a morphotype. *Brain, Behavior and Evolution* 46: 275–318.

Northcutt RG (2001) Changing views of brain evolution. *Brain Research Bulletin* 55: 663–674.

Pepperberg IM (1999) *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*. Cambridge, MA: Harvard University Press.

Reiner A, Yamamoto K, and Karten HJ (2005) Organization and evolution of the avian forebrain. *The Anatomical Record Part A* 287A: 1080–1102.

Roth G and Wullimann MF (eds.) (2001) *Brain Evolution and Cognition*. New York: Wiley and Spektrum Akademischer Verlag.

Striedter GF (2005) *Principles of Brain Evolution*. Sunderland, MA: Sinauer Associates.

Windhorst U, Binder M, Hirokawa N, et al. (eds.) (2007) *Encyclopedic Reference of Neuroscience*. Berlin: Springer.

Evolution of Vertebrate Respiratory Control

R J A Wilson, University of Calgary, Calgary, AB, Canada
B E Taylor and M B Harris, University of Alaska, Fairbanks, AK, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neil Armstrong's first step on the moon is more than matched by our vertebrate ancestors who transitioned to land. Like the first astronauts, the first vertebrates to invade land were well prepared. Key aspects of the respiratory controller appear to have predated tetrapods and to have been present in fish some 400 Ma.

Our understanding of central respiratory control is rudimentary. There have been relatively few modern studies in fish, frogs, and reptiles; the vast majority of data come from studies in mammals. Evolutionary theories should be grounded by data derived from diverse species; in this case, unfortunately, they are not. In effect, investigators are putting together a jigsaw puzzle of which most of the pieces are currently missing. Consequently, the evolutionary framework on which this article is based is speculative, and the homology hypothesis is very much a work in progress. It is presented here in the hope of provoking more work in this exciting area.

Evolution of Rhythm Generation

Evolution of Gill Ventilation: From Tunicate to Fish

The primordial motor patterns required for pumping water over gills can be traced back to the earliest chordates: aquatic animals that lived in water and were filter feeders (Figure 1). Living examples of primitive chordates include tunicates and amphioxus. The adult forms of these animals use rhythmic muscle contractions to pull water through their mouths, into pharyngeal-like structures, and out through secondary orifices.

In amphioxus, the passage of water through the pharynx and across the gills contributes little to gas exchange, suggesting that irrigation of the pharyngeal cavity in primordial vertebrates may have originated from the need to feed, not breathe. Nonetheless, recent molecular studies suggest that these rudimentary feeding behaviors may be the predecessor of vertebrate breathing. The motor neurons responsible for pulling water into the pharynx of adult tunicates, located in the cerebral ganglion which until recently was thought to be a highly derived structure, express transcription factors homologous to *Phox2a* and *Phox2b* (Figure 2). In fish and tetrapods, expression of *Phox2a* and *Phox2b* within the central nervous system is restricted to the brain stem, occurring in brain stem cranial motor neurons, used extensively (but not exclusively) for ventilation. In larval tunicates, when the caudate body plan is most obvious, transcription factors homologous to *Phox2a* and *Phox2b* are similarly restricted to the neck region of the animal.

By the time the jawless fishes (agnatha) evolved, the walls of the gills had likely thinned and become highly vascularized, becoming the primary site for gas exchange, as exemplified in extant agnatha such as the lamprey and hag fish.

Most fish ventilate their gills with water more or less continuously, and therefore the neuronal mechanism responsible likely involves a central rhythm-generating circuit. The presence of a central rhythm generator is supported by the fact that fish brain stems continue to exhibit rhythmic cranial nerve discharges when isolated and superfused.

Where the rhythm generator resides, however, is poorly understood. Several early studies in lamprey and modern ray-finned fishes have addressed the location and organization of the respiratory rhythm generator for gill ventilation. These studies uniformly identify the brain stem as the site of rhythmogenesis, but the rostrocaudal location of the rhythm generator within the brain stem is less clear.

In lamprey and dogfish, a rostrocaudal wave of muscle contraction ensures movement of water across the gills. The rostrocaudal wave of somatic muscle activation involved in swimming is produced by a chain of coupled neuronal oscillators within the spinal cord. While some earlier studies suggest that multiple brain stem oscillators combine to produce the rostrocaudal wave of muscle contraction involved in ventilating the gills, more-recent work suggests that the rostrocaudal progression is produced by differences in conduction delay as a result of rostral and caudal motor neurons having axons of different lengths.

In dogfish, respiratory neurons are distributed along the length of the brain stem, leading some researchers to suggest that the rhythm generator is distributed within either the motor column or the reticular system that runs the length of the brain stem. In addition, local ablation experiments in dogfish have so far failed to identify a discrete location

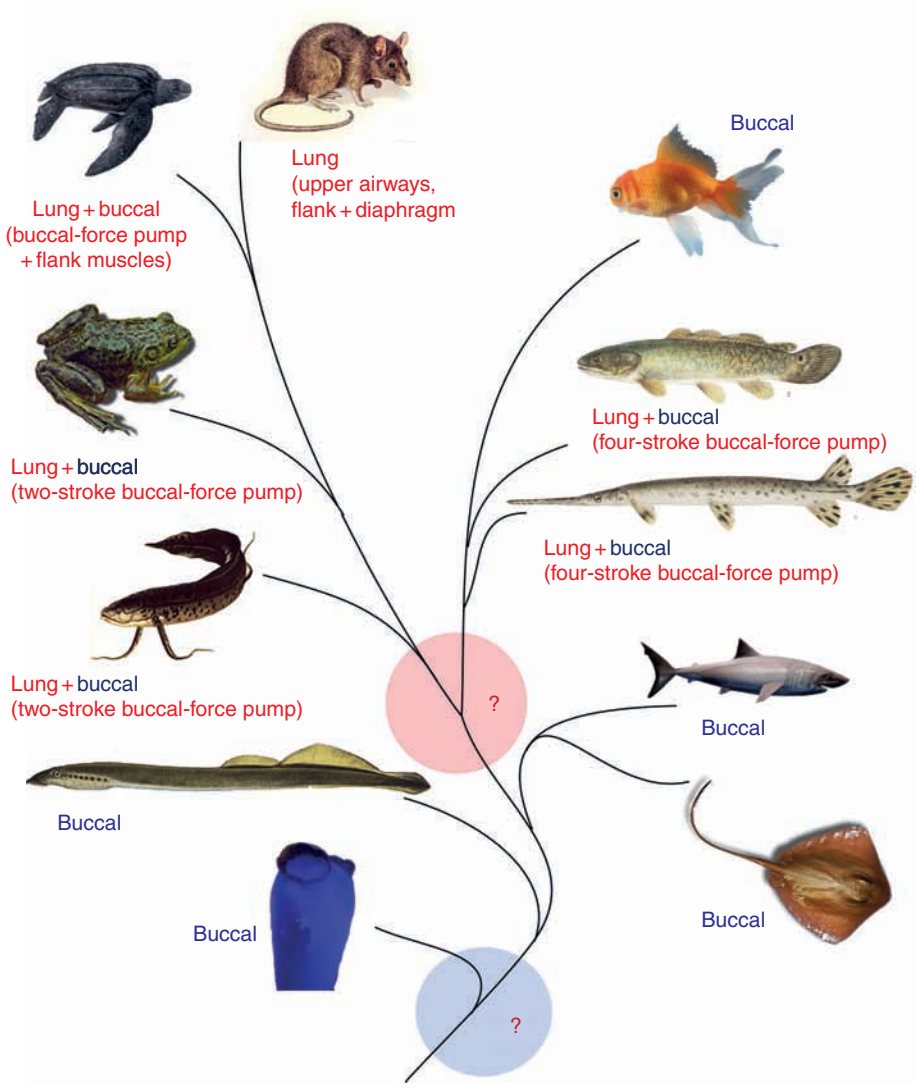

Figure 1 Phylogeny of breathing in vertebrates. Lower vertebrates use buccal and pharynx muscles to move water in and out of their buccal cavities (blue text). In the most primitive vertebrates (e.g., tunicates), this behavior may have been more important in feeding than in breathing. In bimodal ventilators that breathe air and water (e.g., gar, lungfish, and metamorphic tadpoles), the buccal musculature is also used as a 'buccal-force pump' to drive air into primitive lungs (red text). The motor pattern used to pump water appears to be retained in frogs and reptiles, co-opted into ventilating the buccal cavity with air. In reptiles, flank muscles assist the buccal-force pump to inflate the lungs. In mammals, lung inflation occurs through aspiration, the power stroke for which is provided almost entirely by the flank muscles and diaphragm. Nonetheless, even in mammals, phasic contractions of the buccal musculature retain an important respiratory role of maintaining patency of the upper airways. Blue circle, hypothesized origin of primordial neuronal circuit generating buccal contractions used to breathe water. Pink circle, hypothesized origin of air-breathing organ and primordial neuronal circuit, which inflates it. Top right, clockwise: Modern ray-finned fishes (teleosts), primitive ray-finned fishes (Amia and gar), elasmobranches (sharks and rays), tunicates, agnatha (lamprey), lobe-finned fishes (lungfish), amphibians (bullfrog), reptiles (leatherback turtle), and mammals (rat).

that stops rhythmic activity in motor nerves. However, the former evidence is correlative, and local ablation experiments may have missed a critical site. Thus, whether the rhythm generator for gill ventilation in dogfish is distributed remains an open question.

Brain stem transection studies in lamprey suggest a discrete rostral region of the brain stem, located at the level of the trigeminal nucleus, is responsible for gill ventilation. Studies of metamorphic amphibian tadpoles and juvenile frogs also suggest that a discrete region of the brain stem may be uniquely responsible for gill ventilatory rhythmogenesis, albeit at a more caudal location.

Gill Ventilation in the Amphibian Tadpole

Work to resolve the central neuronal mechanisms involved in breathing in lower vertebrates has focused mainly on the bullfrog, *Rana catesbeiana*. Based on parsimony, the paucity of data from fish to refute claims of homology, and the low likelihood that a

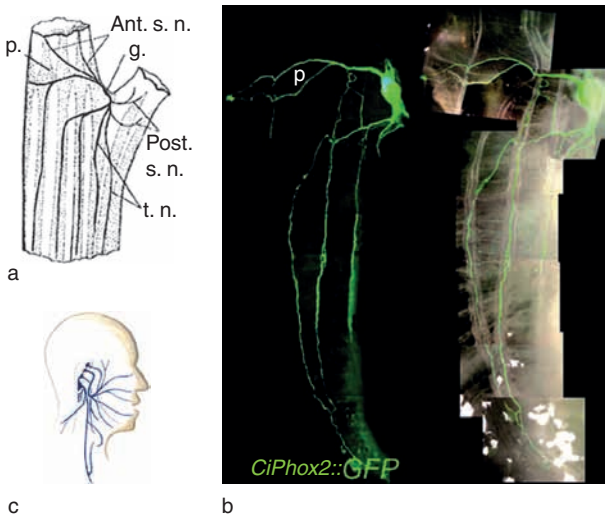

Figure 2 Motor neurons expressing Phox2 innervate the buccal cavity of both tunicates and higher vertebrates. Contractions caused by activity of Phox2 motor neurons in tunicates (a) and (b) and water-breathing vertebrates move water through the 'buccal' cavity. In higher vertebrates, including rats and humans (c), the expression pattern likely includes CO_2/H^+-chemosensitive neurons and cranial motor neurons that play an important role in air breathing. GFP, green fluorescent protein. From Dufour HD, Chettouh Z, Deyts C, et al. (2006) Precraniate origin of cranial motoneurons. *Proceedings of the National Academy of Sciences of the United States of America* 103(23): 8727–8732.

vital behavior such as gill ventilation would have been lost and re-evolved during the transition from fish to frogs, an assumption central to the field is that the neuronal mechanisms responsible for gill ventilation in fishes and bullfrog tadpoles are similar.

In tadpoles, which have both gills and lungs, a 500 μm isolated transverse brain stem slice at the level of the vagus is capable of generating a gill-like motor pattern. During metamorphosis, tadpoles lose their gills and depend more heavily on their lungs for gas exchange. However, despite the loss of gills, a facsimile to the gill ventilatory motor pattern is retained for flushing air in and out of the buccal cavity (i.e., buccal ventilation). This behavior likely serves to prime the buccal cavity for lung inflation and may also be important for olfaction.

Transection studies and drug microinjection studies have been used to locate the rhythmogenic site within the isolated superfused brain stems responsible for buccal ventilation. A small, bilaterally iterated site has been identified proximal to the vagus nerve that appears to be important for generating the buccal motor pattern (Figure 3). Injections of the non-*N*-methyl-D-aspartate glutamate receptor agonist α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid into isolated superfused brain stems at this

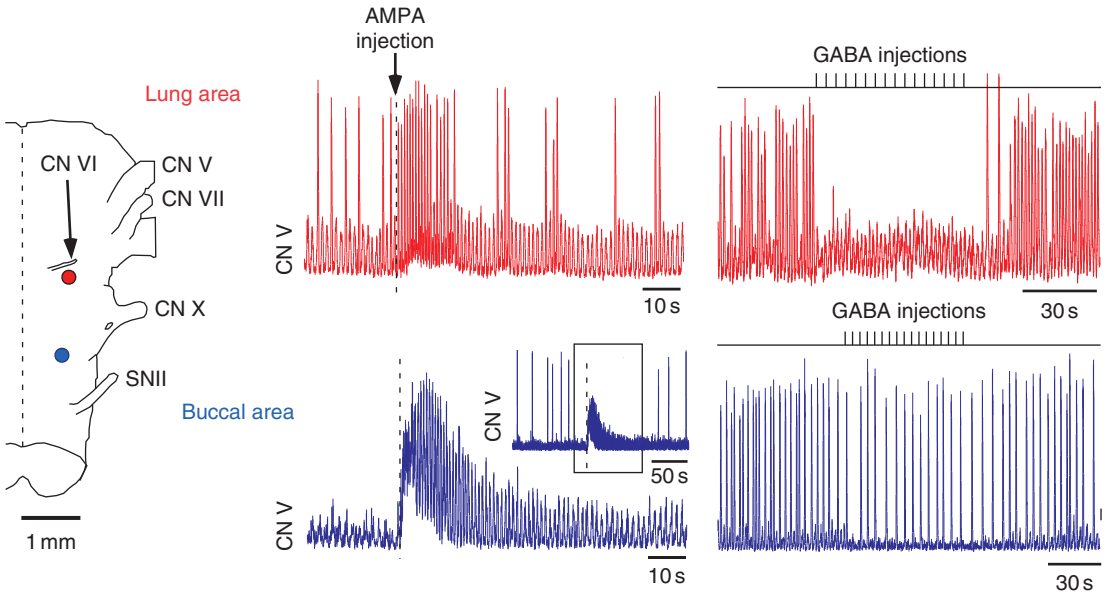

Figure 3 Two oscillators in frogs. Sufficiency (left traces) and necessity (right traces) tests suggest two areas important for generating ventilatory motor patterns, buccal (blue) and lung (red). Under control conditions the isolated juvenile frog brain stem produces buccal and lung ventilatory motor patterns in cranial nerves (CNs). These motor patterns are easily distinguished by burst amplitude (buccal small, lung large). Injecting the non-*N*-methyl-D-aspartate receptor agonist α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) into the lung area (for location see diagram on left) to produce local excitation causes an increase in lung burst frequency whereas injecting γ-aminobutyric acid (GABA) to produce local inhibition abolishes lung bursts but has no effect on buccal bursts. On the other hand, injecting AMPA into the buccal area causes an increase in buccal bursts, but abolishes lung bursts (inlay), whereas injecting GABA into the buccal area abolishes buccal bursts but has no effect on lung bursts. These data, in combination with the results of transection studies and the tight coordination of buccal and lung bursts suggest that the motor pattern is mediated by two coupled oscillators. From Wilson RJA, Vasilakos K, Harris MB, Straus C, and Remmers JE (2002) Evidence that ventilatory rhythmogenesis in the frog involves two distinct neuronal oscillators. *Journal of Physiology* 540: 557–570.

site caused an increase in the frequency of buccal bursts (sufficiency test), whereas an injection of γ-aminobutyric acid (GABA) into the same site caused cessation of buccal bursts (necessity test). These data demonstrate that the region of the brain stem at the level of the vagus is both necessary and sufficient for buccal ventilation.

The rhythmogenic mechanism responsible for generating buccal ventilation remains largely unexplored, although experiments in a number of lower vertebrates suggest that it requires postsynaptic chloride-mediated inhibition and can be abolished with strychnine and biculluline, potent agonists of GABA and glycine receptors, respectively.

Origins of Air Breathing

Ray-finned and lobe-finned fishes diverged at least 408 Ma. The ray-finned fish gave rise to modern teleosts, and the lobe-finned fish gave rise to tetrapods.

Extant but primitive ray-finned and lobe-finned fishes have saclike structures that can be opened to the atmosphere, serve a respiratory function, and have an embryonic origin similar to that of the tetrapod lung, suggesting that fish evolved lungs long before amphibians invaded land.

There are ~375 extant species of air-breathing fishes, including modern perciform teleosts such as the Siamese fighting fish, *Betta splendens*. Only a small portion of these so-called modern species use lunglike organs for gas exchange; most have secondarily derived structures that serve this purpose. Thus, while the first ray-finned fishes were capable of breathing air using a structure homologous to the tetrapod lung, air breathing was lost during the evolution of more-modern fishes, only to be rederived, in some cases, from structures unrelated to the proto-lung.

Despite the presence of lunglike organs in extant but primitive ray-finned fishes, such as the gar and amia, it has long been assumed that air breathing in this clade evolved independently from air breathing of lobe-finned fish. The following three lines of evidence, based on assumed differences in the neuronal mechanisms controlling air breathing, have been used to support this argument:

1. Compared with lobe-finned fishes and tetrapods, air breathing in ray-finned fishes is infrequent and rarely rhythmic but responds vigorously to stimulation of peripheral chemoreceptors. Therefore, some have argued that air breathing in fish is more likely to be reflex-mediated than produced by a central respiratory rhythm generator.
2. The mechanics of air breathing in ray- and lobe-finned fishes differ even though base members of both clades use buccal musculature to push air into their lungs. Specifically, ray-finned fishes use a two-cycle mechanism (one cycle of buccal dilation and constriction to expel stale air from lungs to the atmosphere, another to acquire fresh air into the buccal cavity and pump it into the lungs) whereas lobe-finned fishes and most amphibians use a one-cycle mechanism (expulsion of stale air from the lung to the atmosphere and acquisition of fresh air into the buccal cavity occur simultaneously during a single phase of buccal dilation, and the subsequent buccal constriction pumps the freshly acquired air in the buccal cavity into the lungs).
3. Air breathing in ray-finned fishes differs from that in lobe-finned fishes and tetrapods by showing little sensitivity to arterial CO_2. Therefore, unlike air breathing in tetrapods, air breathing in ray-finned fishes is not modulated by central CO_2 chemoreceptors.

Based on a study using the isolated gar brain stem, it has been proposed that the first air-breather may have predated the division of ray- and lobe-finned fish, the clades that gave rise to modern teleosts and tetrapods, respectively. Specifically, it has been shown that the isolated gar brain stem produces bursts of activity, similar to those associated with air breathing, that respond to changes in superfusate PCO_2. These controversial data suggest gar have a central CO_2-sensitive rhythm generator for air breathing, similar to lobe-finned fishes and their tetrapod descendants (**Figure 4**). If this is indeed the case, then the air-breathing ancestor of modern tetrapods likely evolved before the divergence of ray- and lobe-finned fishes some 400 Ma.

Not all evidence supports the presence of a CO_2-sensitive brain stem rhythm generator in primitive ray-finned fishes. Amia, for example, fail to increase their rate of air breathing when arterial PCO_2 is increased by systemic injection of azetazolamide. However, brain stem isolation and superfusion experiments, which are most likely to demonstrate the presence of a central rhythm generator, have yet to be performed.

The Air-Breathing Rhythm Generator in Amphibians

The presence of a central rhythm generator for air breathing in lungfish, while not investigated in any detail, is equivocal. Until data suggest otherwise, it seems reasonable to assume that the neuronal architecture is similar to that in frogs.

In frogs, a number of studies suggest that the central rhythm generator for air breathing comprises a buccal oscillator accompanied by another brain stem oscillator that is active only during lung inflation. The buccal and lung oscillators appear to be spatially distinct. Necessity, sufficiency, and transection

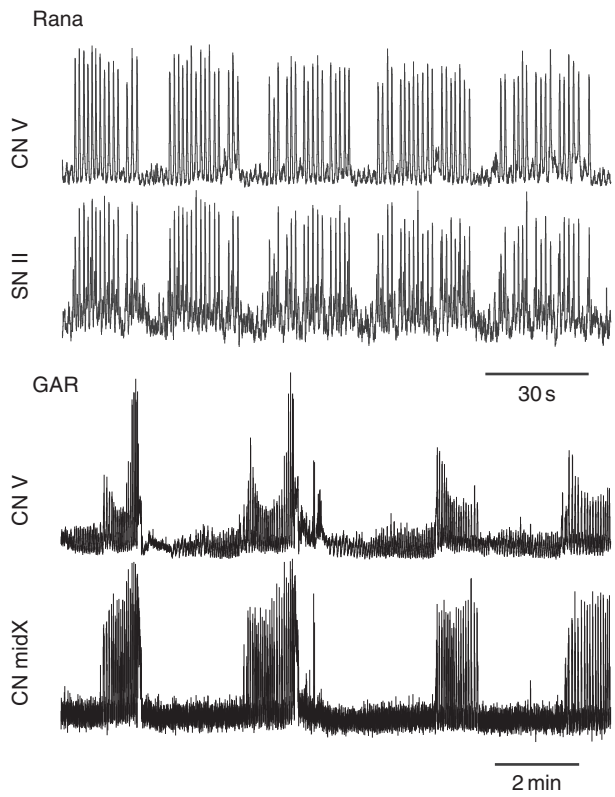

Figure 4 Comparison of motor patterns produced by frog and gar brain stems. Few studies have investigated the neuronal control of breathing in primitive air-breathing ray-finned fishes, and therefore uncertainty remains about the neuronal mechanisms that mediate breathing in this clade. However, data from a study in which the brain stem was isolated and superfused suggest that the gar brain stem is capable of producing both buccal and lung motor patterns in the absence of phasic input; that is, it contains a central respiratory rhythm generator. Furthermore, the putative lung bursts are CO_2/H^+-sensitive. It is remarkable that the motor output produced by the brain stem of the gar is similar to that produced by the juvenile frog, suggesting that brain stems of primitive air-breathing ray-finned fish may also contain two respiratory oscillators. From Wilson RJA, Harris MB, Remmers JE, and Perry SF (2000) Evolution of air-breathing and central CO_2/H^+ respiratory chemosensitivity: New insights from an old fish? *Journal of Experimental Biology* 203: 3505–3512.

studies suggest that the lung oscillator is located rostral to the buccal oscillator at a level just caudal to the facial nucleus (Figure 3). In contrast to the buccal oscillator, bursting of the lung oscillator does not depend on postsynaptic chloride-mediated inhibition, is inhibited by opioids, but is strongly stimulated by increased CO_2.

While spatially distinct, the buccal and lung oscillators interact functionally. In the intact brain stem, the buccal oscillator appears to play an important role in the timing of lung ventilation and may participate directly in lung inflation by priming the buccal cavity before the contraction phase (power stroke) mediated by the lung oscillator.

Given the strong temporal coordination between buccal and lung busts, a two coupled oscillator model has been proposed. While the possibility of additional oscillators is not excluded, simulations suggest that two coupled oscillators are capable of recapitulating the salient characteristics of the motor pattern produced by the isolated brain stem.

The Air-Breathing Rhythm Generator in Mammals

Unlike frogs and fish, which inflate their lungs using positive pressure generated by the buccal musculature, mammals inflate their lungs using negative pressure (aspiration) generated by the diaphragm. Despite the dichotomous biomechanics of the power stroke, in both frogs and mammals the power stroke is preceded by a phase of buccal expansion. In mammals this priming phase protects the upper airways from collapse during the ensuing negative pressure. In frogs, the buccal expansion draws fresh air into the mouth prior to the power stroke.

Recent data suggest that a neuronal architecture consisting of two coupled oscillators, the pre-Bötzinger complex (pre-BötC) and the preinspiratory (pre-I) oscillator, may participate in mammalian respiratory rhythmogenesis (Figure 5). The contribution of these oscillators when embedded within the intact respiratory network is currently under debate. The simplest viewpoint is that the pre-BötC generates inspiratory, power stroke activity, with the pre-I oscillator generating expiratory, pre-I, priming activity. However, on the other extreme, some have argued that within the context of an intact, possibly distributed rhythmogenic network, the contribution of each oscillator is minimal at best. Nonetheless, the presence of two putative oscillators in mammals raises the possibility that they evolved from the same ancestral air-breathing circuit that gave rise to the frog oscillators.

Figure 6 compares the properties of the priming and power stroke oscillators of frogs and mammals. Apparent from this comparison, the power stroke oscillators share a large number of similarities. The priming oscillators also share some similarity, but the resemblance is weaker. Therefore, the likelihood that the power stroke oscillators are homologous is good; the case for homology between priming oscillators is more tenuous.

Evolution of Chemoreceptors

Why Chemoreceptors?

The primary role of chemoreceptors is to maintain homeostasis. These receptors detect changes in blood and/or environmental levels of O_2, CO_2, and pH and initiate cardiorespiratory adjustments to keep gas

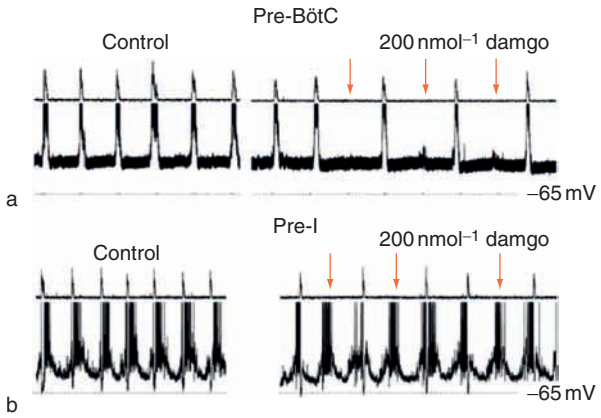

Figure 5 Recent data suggest two oscillators underlie respiratory rhythm generation in mammals. Until recently, the respiratory rhythm generator was widely considered to be composed of a single rhythmogenic kernel called the pre-Bötzinger complex (pre-BötC) embedded in a larger pattern-generating network. Substantive evidence supports the role of the pre-BötC in rhythm generation. However, recent experiments using the μ-opioid agonist DAMGO suggest that a second oscillator is also involved. The rat neonatal isolated brain stem preparation normally produces rhythmic bursts in cranial nerves and cervical nerves including the phrenic. DAMGO produces 'dropped beats' whereby occasional bursts are lost (arrows), but the underlying rhythm (illustrated with vertical bars) is not affected. Intracellular recordings from neurons within the pre-BötC demonstrate that activity of the pre-BötC is abolished during a dropped beat (a), suggesting that pre-BötC does not pace the underlying rhythm. Instead, it appears that the rhythm originates rostrally, within the parafacial respiratory group. This region is rich in preinspiratory (pre-I) neurons that fire immediately before phrenic activity and has been implicated in respiratory rhythm generation previously. It is important to note that, unlike pre-BötC neurons, rhythmic activity of neurons within this region persists during dropped beats (b). Top trace in each panel: integrated phrenic neurogram. Lower trace in (a) and (b), respectively: intracellular recording from a pre-BötC and a pre-I neuron. From Mellen NM, Janczewski WA, Bocchiaro CM, and Feldman JL (2003) Opioid-induced quantal slowing reveals dual networks for respiratory rhythm generation. *Neuron* 37(5): 821–826.

tensions and pH within set limits despite changing metabolic demands.

Water-breathing vertebrates obtain O_2 from water. The low capacity of water for O_2 necessitates a high flow across gas-exchange surfaces. In contrast, the relatively high solubility of CO_2 in water generally results in rapid CO_2 excretion. In essence, water-breathing vertebrates ventilate in order to acquire sufficient O_2 to meet metabolic demands and, in doing so, satisfy the need to excrete metabolically produced CO_2. In contrast to aquatic vertebrates, terrestrial vertebrates breathe a medium rich in O_2 but face greater limitations in the ability to excrete metabolically produced CO_2. Consequently, at rest, air-breathing terrestrial vertebrates generally ventilate to satisfy needs for CO_2 excretion and, in doing so, satisfy metabolic requirements for O_2 uptake.

The evolution from water to air thus necessitated a transition from a respiratory system dominated by the need to satisfy O_2 demand to one that is primarily preoccupied with excreting CO_2. These changing demands were met in part by the changing importance of two populations of chemoreceptors, peripheral and central, that have different sensitivities to O_2 and CO_2.

Peripheral Respiratory Chemoreceptors in Fish

The primary sites of respiratory chemoreceptors that modulate aquatic ventilation in fish are the gills. The anterior gill arches and, to a lesser extent, regions within the orobranchial pharynx contain populations of O_2-sensitive receptors oriented toward both the blood and the environment. This receptor distribution suggests that fish modulate ventilation in response to aquatic O_2 levels, the effectiveness of O_2 uptake, or the degree of metabolic O_2 consumption as indicated by blood PO_2 at the level of the gills. These locations also contain CO_2/H^+-sensitive receptors, which in most fish appear to be oriented toward the environment and, thus, likely monitor aquatic CO_2 tension.

While branches of cranial nerves (CNs) VII (facial), IX (glossopharyngeal), and X (vagus) innervate the gills, sensory information from the chemoreceptors to the brain stem is carried predominantly in CNs IX and X.

In both air- and water-breathing fishes, the peripheral chemosensory information carried by these nerves has been considered almost absolute in stimulating ventilation; in air-breathing fishes, the input was so strong that air breathing was considered an on-demand behavior, controlled entirely by a peripheral respiratory chemoreceptor-mediated reflex arch. More recent data suggest the involvement of a central rhythm generator and modulation by other sensory modalities.

Peripheral Respiratory Chemoreceptors in Tetrapods

The carotid and aortic arches house the primary peripheral respiratory chemoreceptors in tetrapods. These structures share a common embryonic origin with the gill arches of fish and, like the peripheral respiratory chemoreceptors in fish, are innervated by CNs IX and X (Figure 7). These similarities have led to the widely held view that the carotid and aortic arch chemoreceptors of tetrapods evolved from fish gill arch chemoreceptors.

In mammals, the carotid bodies dominate peripheral chemosensitivity: the ventilatory response to hypoxia decreases only slightly (<10%) when the aortic arch chemoreceptors are denervated. The carotid bodies are most strongly activated when stimulated concurrently by low levels of oxygen and high levels

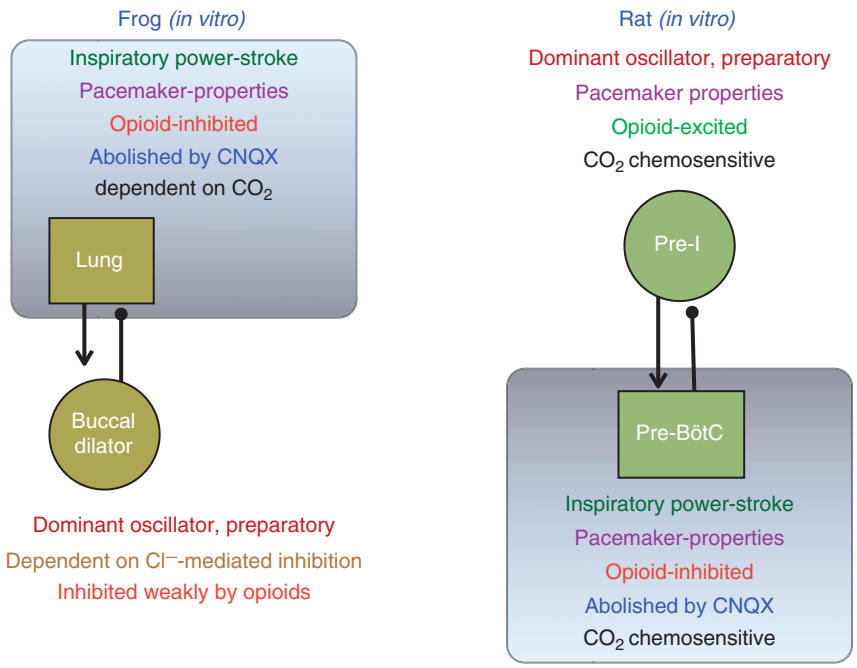

Figure 6 Functional similarities between frog and rat oscillators. There are a number of similarities between the putative pairs of oscillators in frogs and rats, as illustrated. The similarities between the frog lung oscillator and the rat pre-Bötzinger complex (pre-BötC) (squares) are strongest, and therefore these are most likely to be homologous. Arguments for homology of buccal and pre-I oscillators (circles) rely on their functional similarity: both are active immediately before inspiration, priming the upper airways for the ensuing power stroke, and appear to have a primary role in setting the underlying rhythm. Diagram illustrates the anatomical arrangement of the oscillators in frogs and rats. Note that anatomy does not support dual homology.

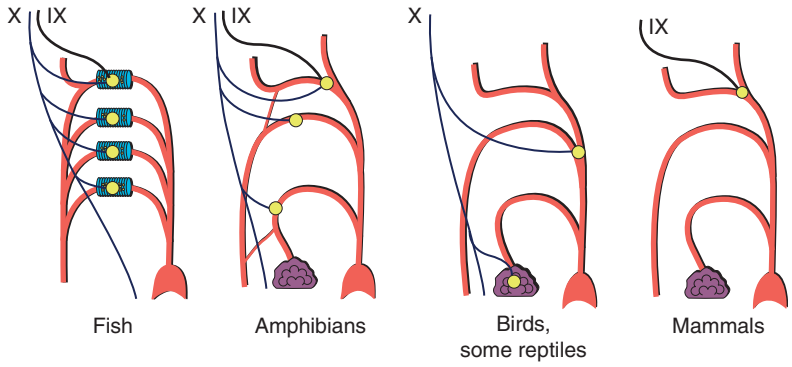

Figure 7 Comparative anatomy of peripheral chemoreceptors. This diagram depicts the distribution of peripheral arterial and pulmonary CO_2/H^+-sensitive chemoreceptors (yellow circles) in various vertebrate groups. IX and X designate the ninth (glossopharyngeal) and tenth (vagus) cranial nerves (CNs), respectively. It is important that chemoreceptors associated with the first gill arch in fish, innervated by CN IX, are homologous to the primary peripheral chemoreceptors in all vertebrates. Modified from Milsom WK (2002) Phylogeny of CO_2/H^+ chemoreception in vertebrates. *Respiratory Physiology & Neurobiology* 131(1–2): 29–41.

of CO_2. This condition arises during suffocation, leading some to describe the peripheral chemoreceptors as the primary asphyxia sensor. Under normal conditions (i.e., when blood gases are normoxic and normocapnic), the contribution of the peripheral chemoreceptors to the overall 'drive' to breathe is still debated, with the consensus around 40% and estimates ranging from 0% to 70%. The debate is complicated by the state-dependent integration of peripheral respiratory chemoreceptor afferent information, which is subject to complex modulation depending on, among other factors, the degree of lung inflation and brain stem PCO_2.

Central CO_2 Chemoreceptors in Air-Breathing Vertebrates

Although peripheral chemoreceptors contribute in important ways to ventilatory responses to transient

changes in O_2 or CO_2/H^+ in tetrapods, responses persist despite denervation of peripheral chemoreceptors. The dominant source of chemosensory drive under steady-state conditions is central chemoreceptors: CO_2/H^+-sensitive receptors located within the central nervous system. There is some evidence (discussed later) to suggest the presence of these receptors in primitive fishes, but their general presence in fish has yet to be substantiated.

Despite considerable research effort in mammals, the nature of central chemoreceptors has made their investigation difficult. In rodents and goats, activity of multiple populations of chemosensitive cells contributes to ventilatory responses invoked by central changes in CO_2/H^+. These populations are widely distributed throughout the brain stem in important respiratory areas such as the raphe, nucleus tractus solitarius, locus coeruleus, hypothalamus, and the subpopulations of the ventral lateral medulla, including the pre-BötC and retrotrapezoid nucleus. Chemostimulation of any one of these populations can elicit substantial ventilatory responses, suggesting considerable redundancy. Moreover, these different populations may utilize a diverse array of chemosensory transduction mechanisms.

The origin of these chemoreceptor populations relative to the evolution of tetrapods and the water-to-air transition is enigmatic. There is some evidence of chemosensitive cells in the brain stem of some water-breathing fishes, including one of the most primitive extant groups, the lamprey. However, the preponderance of evidence indicates that water-breathing fishes lack central ventilatory chemosensitivity. As previously noted, air breathing is a gas exchange strategy that has evolved independently in a number of fish species. Within one of these air-breathing lineages, the tetrapod lung has its origin in now-extinct, primitive air-breathing fish. An extant species of air-breathing fish (the gar, *Lepisosteus osseus*) closely related to this lineage that gave rise to tetrapods has provided a key element in the understanding of the origin of central ventilatory chemosensitivity. Brains isolated from these fish produce neural correlates to gill ventilation and air breathing.

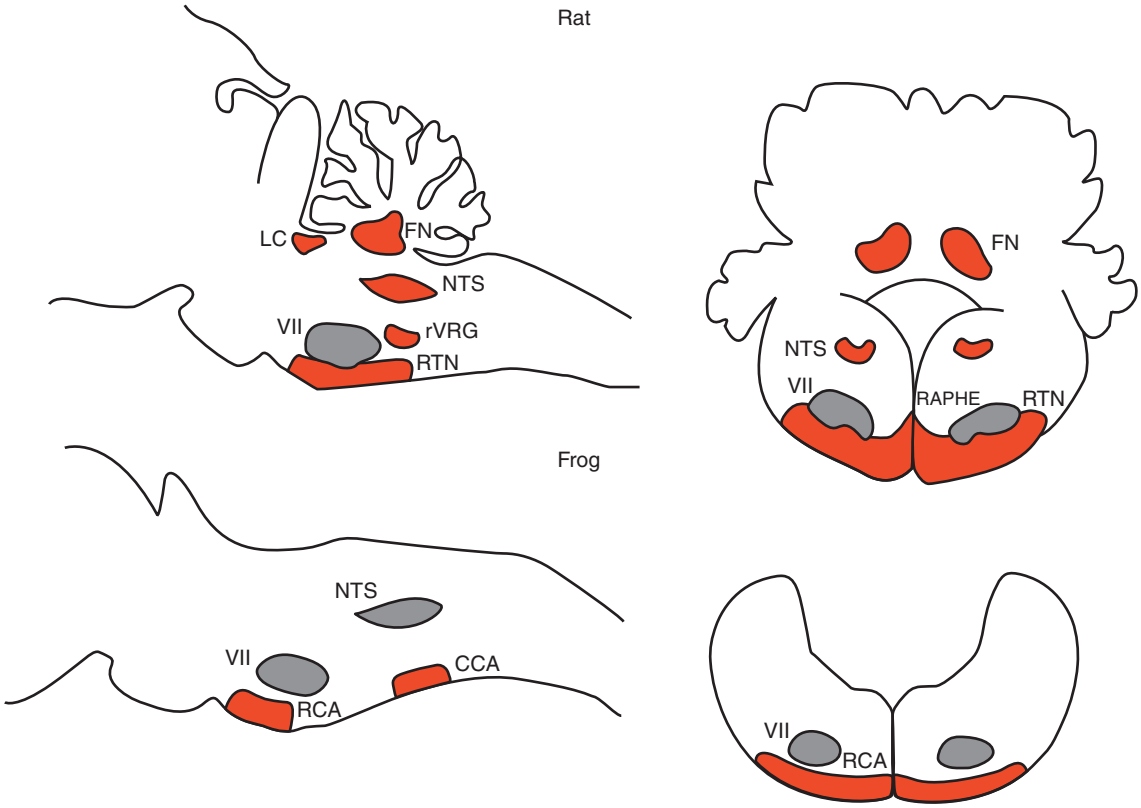

Figure 8 Central chemoreceptors in frogs and mammals. This diagram depicts areas within the central nervous system that have been shown to contain CO_2/H^+-sensitive cells. Longitudinal (left) and coronal sections (right) of a typical mammal (top) and amphibian (below) are shown. Areas with documented chemosensitivity are red, and their locations are presented relative to the facial nucleus (VII; shaded). Chemosensitive areas include the following: locus coeruleus (LC); fastigial nucleus (FN); rostral ventral respiratory group (rVRG); retrotrapezoid nucleus (RTN); nucleus tractus solitarius (NTS); and the amphibian rostral chemosensitive area (RCA) and caudal chemosensitive area (CCA).

Exposing isolated brains to elevated CO_2 prompts an increase in putative air breathing, indicating central ventilatory sensitivity to CO_2. Thus, like the tetrapod lung, CO_2/H^+-sensitive ventilatory chemoreceptors may also have their origin in a primitive air-breathing fish.

The amphibian tadpole metamorphoses from a fishlike, free-swimming larval stage that exchanges gas with water across its skin and gills into a mature semiterrestrial form that exchanges gas across its skin and lungs. Brains isolated from any of these stages produce neural correlates to gill ventilation and air breathing, which represent gill and lung ventilation in the intact animal. Mechanisms of ventilatory control are remarkably conserved among vertebrates, and these preparations have provided great insight into these mechanisms. Previous characterizations suggested that central CO_2/H^+ chemosensitivity for air breathing was absent in early metamorphic stages despite the presence of a lung and that this sensitivity developed only in the late stages, associated with loss of the gills, the adoption of a semiterrestrial lifestyle, and use of the lung ventilation as a primary gas exchange strategy. Subsequent investigations, however, have established that exposing brains isolated from tadpole of any stage to elevated CO_2 results in increases in putative air breathing. Thus, central CO_2/H^+ chemosensitivity for air breathing is present throughout tadpole development, including fishlike stages with fully functional gills.

Other Chemoreceptors: Upper Airways and Lung Epithelial Bodies

Many vertebrate species have olfactory and perhaps gustatory CO_2/H^+-sensitive receptors that can influence ventilation and are believed to be involved in general defense responses to environmental conditions. These include CO_2/H^+-sensitive receptors associated with pulmonary arteries and lung tissues present in amphibians, reptiles, birds, and mammals. The significance of these receptors is unclear, and their evolutionary origin has not been investigated.

Other Reflexes

Ventilation is modulated by a number of reflexes other than those initiated by chemoreceptors. Pulmonary stretch has a profound influence on breathing, termed the Hering-Breuer reflexes. Receptors located in the lung and/or pleural tissues and stimulated by the pulmonary stretch associated with lung inflation terminate inspiration in a reflex that prevents overinflation of the lung. Similarly, pulmonary stretch lengthens the period between subsequent breaths, and pulmonary deflation hastens the next inspiration. Such reflexes are present in amphibians as well as mammals. Air-breathing fishes such as gar and lungfish show similar alterations in ventilatory volume and frequency with pulmonary inflation and deflation. The origin of these reflexes may relate to either protection of the fragile lung or buoyancy control, but their presence in primitive air-breathing fishes suggests that Hering-Breuer pulmonary stretch reflexes are as old as the lung itself.

Another important reflex has been termed the mammalian diving reflex. This comprises a host of respiratory and cardiovascular adjustments associated with the presence of water in the pharynx or on the face. This reflex is more pronounced in neonates than in adults of most species, notably mammals, and in mammals it is exaggerated by low temperature. The ventilatory component of this reflex is airway closure and inhibition of breathing, measures that prevent water from entering the lungs. Some aspects of this reflex are observed in almost every air-breathing species, including fishes, suggesting that reflexes protecting the lung from water arose in conjunction with air breathing in aquatic species.

Summary and Conclusions

From a review of key aspects of respiratory control in vertebrates, it appears that many of the components characteristic of the neuronal controller for air breathing in mammals may have predated tetrapods. These include the oscillators that generate respiratory rhythm and the chemoreceptors that regulate it.

See also: Autonomic Nervous System: Respiratory Control; Autonomic Nervous System: Central Respiratory Control; Brainstem Respiratory Circuits; Respiration.

Further Reading

Bass AH and Baker R (1997) Phenotypic specification of hindbrain rhombomeres and the origins of rhythmic circuits in vertebrates. *Brain, Behavior and Evolution* 50(supplement 1): 3–16.

Brainerd EL (1994) The evolution of lung-gill bimodal breathing and the homology of vertebrate respiratory pumps. *American Zoologist* 34: 289–299.

Dufour HD, Chettouh Z, Deyts C, et al. (2006) Precraniate origin of cranial motoneurons. *Proceedings of the National Academy of Sciences of the United States of America* 103(23): 8727–8732.

Feldman JL and DelNegro CA (2006) Looking for inspiration: New perspectives on respiratory rhythm. *Nature Reviews Neuroscience* 7(3): 232–242.

Gans C (1970) Strategy and sequence in the evolution of the external gas exchangers of ectothermal vertebrates. *Forma et Functio* 3: 61–104.

Gilmour KM and Perry SF (2007) Branchial chemoreceptor regulation of cardiorespiratory function. In: Hara TJ and Zielinski B (eds.) *Fish Physiology, Vol. 25: Sensory Systems Neuroscience*, pp. 97–151. San Diego, CA: Academic Press.

Graham JB (1997) *Air-Breathing Fishes: Evolution, Diversity and Adaptation*. London: Academic Press.

Milsom WK (2002) Phylogeny of CO_2/H^+ chemoreception in vertebrates. *Respiratory Physiology & Neurobiology* 131(1–2): 29–41.

Smatresk N (1994) Respiratory control in the transition from water to air breathing in vertebrates. *American Zoologist* 34: 264–279.

Taylor BE, Harris MB, Leiter JC, and Gdovin MJ (2003) Ontogeny of central CO_2 chemoreception: Chemosensitivity in the ventral medulla of developing bullfrogs. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* 285: R1461–R1472.

Taylor EW, Jordan D, and Coote JH (1999) Central control of the cardiovascular and respiratory systems and their interactions in vertebrates. *Physiological Reviews* 79: 856–900.

Wilson RJA, Vasilakos K, Harris MB, Straus C, and Remmers JE (2002) Evidence that ventilatory rhythmogenesis in the frog involves two distinct neuronal oscillators. *Journal of Physiology* 540: 557–570.

Wilson RJA, Vasilakos K, and Remmers JE (2006) Phylogeny of vertebrate respiratory rhythm generators: The oscillator homology hypothesis. *Respiratory Physiology & Neurobiology* 154(1–2): 47–60.

Wilson RJA, Harris MB, Remmers JE, and Perry SF (2000) Evolution of air-breathing and central CO_2/H^+ respiratory chemosensitivity: New insights from an old fish? *Journal of Experimental Biology* 203: 3505–3512.

Excitotoxicity in Neurodegenerative Disease

S L Budd Haeberlein, AstraZeneca, Södertälje, Sweden
S A Lipton, Burnham Institute of Medical Research, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction to the Excitotoxic Theory of Neurotoxicity

Glutamate is the predominant excitatory neurotransmitter in the mammalian central nervous system (CNS) and activates several subtypes of excitatory amino acid (EAA) receptors, which fall into two functional groups – the ionotropic glutamate receptors and the metabotropic glutamate receptors (mGluRs). The ionotropic receptors are further divided into *N*-methyl-D-aspartate (NMDA), α-hydroxy-5-methylisoxazole-4-propionate (AMPA), and kainate (KA) receptors, which are named on the basis of their differential sensitivity to structural analogs. Under normal conditions, neuronal and glial uptake systems maintain the extracellular concentration of glutamate at or below low micromolar levels. However, during certain pathologic conditions, which will be discussed in the following sections, the extracellular concentrations of glutamate can be elevated or their receptors may become hyperactivated for other reasons, and can lead to neurotoxicity through a mechanism called excitotoxicity. Excitotoxicity quite literally means a toxic 'effect,' the death, even, of neurons, resulting from excessive excitatory input. Overstimulation of NMDA receptors is often a predominant factor in such situations.

It was first observed in the 1950s that the excitatory neurotransmitter glutamate could exert toxic effects on the nervous system under certain conditions. Following the systemic application of glutamate to mice, toxic effects on retinal neurons were described. Subsequent extensive studies by Olney and co-workers corroborated these toxic effects and extended this view by showing neuronal cell death following oral intake of glutamate or aspartate in brain regions devoid of the blood–brain barrier in mice and nonhuman primates. Olney hypothesized that glutamate-induced neuronal cell death was a general property of excitatory amino acids acting on central neurons, and later coined the term excitotoxicity.

With the introduction of structural analogs to the glutamate receptor agonists, local injections of the glutamate agonist kainate were shown to induce cell death with a similar pattern of damage in different brain regions, thus confirming the neurotoxic effect. Knockout mice lacking the glial glutamate transporters GLT1 and GLAST show elevated extracellular glutamate levels and the neurodegeneration characteristic of excitotoxicity. The different location of the brain damage induced by kainate alludes to the parallel discovery of the different glutamate receptor subtypes and led to further understanding of both the physiological role of glutamate as an excitatory neurotransmitter and its cytotoxic effects. Today, it is well recognized, that exogenous or endogenous agonists of excitatory amino acid receptors can induce cell death in CNS neurons through excessive ionotropic glutamate receptor activation.

Overstimulation of glutamate receptors induces the intracellular accumulation of several ions and second-messenger molecules, the most important of which may be the large influx of Ca^{2+}. Consequently the dysregulation and dysfunction in calcium signaling have been studied extensively.

Concerning the mode of cell death, glutamate-induced neuronal cell death was first judged from morphological appearance as necrotic. However, the observation of a delayed neuronal cell death in the penumbra of ischemic lesions and after EAA exposure has stimulated studies focusing on the mode of cell death. Today, there is compelling evidence that failure in extracellular glutamate homeostasis can result in different modes of cell death, with morphological and biochemical features of either apoptosis or necrosis, depending on the severity of the insult.

Clinical Relevance of Excitatory Amino Acid Neurotoxicity

The excitotoxic theory of neuronal cell death is supported by the connections between abnormalities in the glutamate system and disease states.There is evidence that excitotoxicity is involved in acute brain damage under pathophysiological conditions following status epilepticus, mechanical trauma, or ischemia. Moreover, glutamate cytotoxicity may also be partly involved in many neurodegenerative diseases, ranging from Parkinson's disease, Alzheimer's disease, Huntington's disease, and amyotrophic lateral sclerosis (ALS; also called motor neuron disease) to HIV-associated dementia. Some examples are cited in the following sections.

Hypoxia/Ischemia

Under conditions of local or global ischemia, neurons are deprived of glucose and oxygen, resulting in

energetic failure and subsequent decrease of ion gradients across the plasma membrane. During hypoxic/ischemic conditions, microdialysis measurements have shown substantial elevations in extracellular glutamate and other EAAs. Severe focal cerebral ischemia lasting from 20 min to 2 h has been shown to cause up to 20-fold increase in the extracellular concentration of glutamate, reaching concentrations of 25–80 μM. Since in central neurons glutamate exists in two pools, metabolic (cytoplasmic) and neurotransmitter (synaptic vesicle), there are two routes by which glutamate is released during ischemia. The plasma membrane depolarization which occurs subsequent to the energetic failure leads to increased synaptic release of glutamate, while the diminished Na^+ gradient is followed by attenuated Na^+-dependent glutamate uptake or even reversed glutamate transport in terms of transporter-mediated release of glutamate from both neurons and astrocytes. Moreover, osmotic cell swelling following the influx of Na^+, Cl^-, and H_2O can result in plasma membrane rupture and further release of cytoplasmic glutamate into the extracellular space. In summary, hypoxia/ischemia results in a secondary net increase in the extracellular glutamate concentration. This increase in extracellular glutamate results in excitotoxic damage. The implication of excitotoxicity in hypoxia/ischemia is corroborated by the observation that NMDA-type glutamate receptor antagonists can reduce neuronal death in animal models of cerebral ischemia. However, some negative results have been reported with glutamate receptor antagonists in models of ischemia, and the actual set of circumstances in humans appears to differ from the animal models. The metabolic and ionic changes in the ischemic region are not homogeneous; at the core, depolarization and the collapse of ion homeostasis occur rapidly and cells die by necrosis. The surrounding area, or penumbra, is a region with less severe metabolic inhibition. These cells go on to die by excitotoxic mechanisms over a delayed period. In animal models this delayed excitotoxicity is mediated by apoptosis, but whether a penumbral region exists in humans is yet to be firmly established. There are several lines of evidence, that neurons can undergo apoptosis following ischemia. Among various biochemical markers of apoptosis, DNA fragmentation and activation of caspase-3 have been observed in models for cerebral ischemia. Moreover, application of caspase-inhibitors in animal models of cerebral ischemia result in reduced infarct size and decrease cell death in cultured neurons following ischemia. In line with a downstream role, the therapeutic window for application of caspase inhibitors is longer than that of NMDA receptor antagonists such as MK-801.

Epilepsy

Histopathological studies on the brains of patients suffering from chronic epilepsy have revealed that certain brain regions show structural alterations, with severe loss of neurons and reactive gliosis. Brain pathology in epilepsy is described best for human temporal lobe epilepsy, resulting in sclerosis of the hippocampus that extends into the amygdala and the parahippocampal gyrus. Pronounced brain damage has been observed following sustained epileptiform activity with seizures lasting more than 30 min, called 'status epilepticus' (SE). Importantly, neuronal cell loss after SE is not distributed equally across all hippocampal subfields, and the extent of damage is in the order CA1 > CA4 > CA3 > CA2, thus indicating different susceptibility to SE-induced cell death.

In animal models of epilepsy using chemoconvulsant-induced or electrical stimulation-induced SE, similar patterns of brain damage are observed. Experiments aimed at the observation of ultrastructural changes have demonstrated that certain features of cell damage, such as swelling of dendrites and soma, are independent of the mechanism used to induce SE. Evidence for the implication of EAA-mediated excitotoxicity in SE-induced neuronal cell death arises from several experiments. First, the morphological appearance of cell damage following SE is similar to damage following systemic or local application of the glutamate receptor agonists L-glutamate, NMDA, or kainate. Second, the administration of ionotropic glutamate receptor antagonists can prevent cell death induced by epileptiform activity. Apoptotic cell death is primarily implicated in the pathology in epilepsy, as demonstrated by the upregulation of apoptotic gene expression and protein markers. Further, antiapoptotic treatments have been shown to reduce the extent of seizure-induced neuronal cell death.

Traumatic Brain Injury

Following head trauma, mechanical brain injury can be accompanied by secondary changes as 'metabolic' glutamate leaks uncontrollably from the neuronal cytoplasm, causing subsequent excitotoxic damage of surrounding neurons. However, compared to research on mechanisms underlying the cell damage in ischemia or epilepsy, studies on traumatic injury have been sparse and are complicated by heterogeneity in the initiating event. Studies have suggested the sudden release of EAAs from the cytoplasm into the extracellular space and subsequent energetic failure, as well as ultrastructural damage that can be diminished by NMDA antagonists, indicative of excitotoxicity. Apoptotic signaling is found in traumatic brain injury

(TBI), including activation of caspase enzymes, internucleosomal DNA fragmentation, and induction of immediate-early genes. However, in neonatal models of TBI (mortality and morbidity from head trauma is highest in children), the resulting excitotoxicity has been shown to elicit both apoptosis and necrosis. Necrosis occurs localized to the site of impact, and within 4 h of the insult, whereas secondary apoptotic damage occurs between 6 and 24 h and is found in regions surrounding the area of primary necrosis. In this model, the secondary damage was more severe than the primary damage was, suggesting a preponderance of apoptosis over necrosis, at least in the developing brain. Glutamate receptor antagonists protected against the primary damage, exacerbated the secondary damage, but have also been shown to be effective in patients with severe head injury.

Neurodegenerative Diseases

In addition to acute neurological disorders, many chronic neurodegenerative diseases may exhibit a component of glutamate-dependent neuronal damage, including apoptosis or injury to dendrites, axons, and synapses. Distinct neurodegenerative pathologies can be reproduced in animal models by, for example, kainic acid injection, and one theory is that excitotoxicity in chronic neurodegeneration is secondary to an impairment of cellular energetics. This theory hypothesizes that mitochondrial dysfunction leads to a chronic energy impairment which results in a reduction in membrane potential, and allows a physiological concentration of EAAs to become toxic. In support of this theory, the injection of inhibitors of mitochondrial energy production such as 1-methyl-1,2,3,4-tetrahydroisoquinoline (MPTP), malonate, quinolinic acid (QA), and 3-nitropropionic acid (3-NPA) produces lesions in animals similar to those in neurodegenerative disorders.

Additionally, elevated inflammatory responses in many of these diseases can also contribute to excessive glutamate release or decreased glutamate clearance.

Parkinson's disease Parkinson's disease (PD) is characterized by the selective loss of dopaminergic (DA) neurons in the nigrostriatal pathway. Administration of MPTP or 6-OH-dopamine to mice induces loss of DA neurons and a disorder that behaviorally and histologically resembles Parkinson's. Nigral DA neurons possess glutamate receptors, and they degenerate when exposed to EAAs. Studies have shown that the NMDA receptor antagonists MK-801 and 3-(2-carboxypiperazin-4-yl)-propyl-1-phosphonate (CPP), and the glutamate release inhibitors riluzole and lamotrigine, can improve parkinsonian behavioral parameters in animals lesioned with MPTP or 6-OH-dopamine.

Several antiexcitotoxins are presently used as anti-parkinsonian drugs, including amantadine, memantine, and budipine. Originally thought to be anticholinergics, these compounds are inhibitors of NMDA receptors. The drugs are used at the concentration required for NMDA receptor antagonism, though budipine may primarily enhance endogenous dopaminergic activity. These drugs provide symptomatic treatment for Parkinson's disease (reduced tremor, akinesia, and rigidity) through attenuation of the hyperactive glutaminergic mechanisms in the subthalamic nucleus (STN).

Huntington's disease Huntington's disease (HD) is a hereditary neurodegenerative disease. The disease mutation consists of an expanded CAG repeat in the gene encoding for the huntingtin protein. Huntingtin is expressed throughout the CNS; however, the striatum is the primary site affected by HD, with secondary cortical damage. At the cellular level, striatal neurons contain bodies of cleaved and ubiquitinated huntingtin. The mechanism through which mutant huntingtin mediates striatal neuronal cell death is incompletely understood, but may, at least in part, involve transcriptional dysregulation and mitochondrial dysfunction.

Injection of the NMDA-like excitotoxin, quinolinic acid (QA), into the striatum of primates elicits chorei form movements and a neuropathology that emulates that seen in HD. Moreover, NMDA receptor-mediated neuronal apoptosis is enhanced in transgenic models of HD.

Alzheimer's disease Patients with Alzheimer's disease (AD) inherit the disease as a fully penetrant, autosomal dominant trait with a young age of onset. Early-onset AD is caused by defects in any of three different genes, all related to β-amyloid pathology. In the more common sporadic form of AD, there is no clear genetic link to increased β-amyloid, and molecular genetic analyses suggest that there are likely to be many genes that influence the susceptibility to AD (presenilin, etc.). It is well established that β-amyloid induces neurotoxicity, and may do so in part through excitotoxic mechanisms. β-Amyloid has been shown to generate oxidative stress and increased intracellular Ca^{2+}, and is reported to enhance glutamate-mediated neurotoxicity *in vitro*. Additional experiments suggest that β-amyloid can increase NMDA responses and thus excitotoxicity as well as cause NMDA receptor-mediated injury to dendritic spines.

Another potential link comes from recent evidence that glutamate transporters are downregulated in Alzheimer's disease and that β-amyloid can inhibit glutamate reuptake or even enhance its release. Activated or cleaved proteins that are a result of activated apoptotic pathways have been observed in postmortem Alzheimer's disease brains, but these could be related to apoptosis induced by excitotoxic and nonexcitotoxic mechanisms such as tau hyperphosphorylation, which is present in Alzheimer's disease. Finally, memantine, an NMDA receptor antagonist, has been approved by the European Union and the US Food and Drug Administration for the treatment of moderate to severe Alzheimer's disease, and most likely acts via its antiexcitotoxic mechanism of action.

Intoxication with Exogenous Excitatory Amino Acids

Several structural analogs of endogenous EAAs naturally exist, and these exhibit similar or even higher excitatory and neurotoxic potency compared to the endogenous EAAs. In 1987, an outbreak of domoic acid poisoning following ingestion of mussels occurred in Canada. Patients suffered from acute headache, seizures, and sensory dysfunctions, and showed motor signs. In four fatal cases, neuropathological studies revealed lesions predominantly in the hippocampus and amygdala, resembling the lesion pattern after application of the exogenous excitotoxin kainate. It was reconstructed that the mussels had accumulated domoic acid, synthesized by the phytoplankton *Nitzchia pungens*. Domoic acid is a structural analog of kainic acid, and is also synthesized by seaweed, but compared to kainate, domoic acid has a higher excitatory potency. Experimental administration of domoic acid to rodents and monkeys has resulted in brain damage with ultrastructural features resembling L-glutamate excitotoxicity.

Ingestion of the chick pea *Lathyrus sativus*, which contains another EAA structural analog, results in acute spastic motor signs following the consumption. The clinical features of this motor disorder, called 'lathyrism,' were known even by the ancient Greeks, but the presumed toxic component, the amino acid β-*N*-oxalyl-L-alanine (BOAA) was identified only some decades ago. Experimental administration of BOAA is known to induce the features of lathyrism in nonhuman primates, and application of BOAA to neuronal cell cultures results in cell death with structural features of excitotoxicity, although other work has contested these results and offered other theories of lathyrism. The high excitotoxic potency of this glutamate receptor ligand is in accordance with its high excitatory potency as an agonist at AMPA receptors.

Implication of Distinct Glutamate Receptor Classes in Excitotoxicity

L-Glutamate, the most abundant excitatory transmitter in the brain, binds to different classes of receptors comprising different types of ionotropic receptors as well as metabotropic receptors. Ionotropic glutamate receptors are ligand-gated ion channels, which are named after agonists that bind preferentially to these receptor subtypes. They include AMPA, kainate, and NMDA receptors.

AMPA receptors are widely distributed in the CNS and are expressed on many different types of neurons. They control a cation channel that is permeable to Na^+ and K^+ ions. Activation of AMPA receptors results in the fast onset of an excitatory postsynaptic current, with rapid desensitization. This current shapes the fast component of glutamatergic transmission in CNS neurons. Importantly, certain AMPA receptor subtypes can exhibit substantial Ca^{2+} permeability and thereby contribute to the Ca^{2+}-dependent form of excitotoxic cell damage. Ca^{2+}-permeable channels are formed from the selective receptor subunits GluR1 or GluR3, whereas co-assembly with GluR2 results in only poor Ca^{2+} permeability.

Another type of ionotropic glutamate receptor, the kainate receptor, is less well understood because of lack of sufficiently selective agonists and antagonists, but better compounds have recently been developed. As with AMPA receptors, kainate receptors control a cation channel that desensitizes rapidly. Since the agonist kainite activates kainate receptors but also AMPA receptors, their physiological and pathophysiological role remains elusive. Recently, results based on gene-targeted rodents lacking certain kainate receptor subtypes have extended our knowledge of localization and function of kainate receptors. From these studies one can conclude that kainate receptors are not only localized to postsynaptic sites but also appear to be localized presynaptically, suggesting a role for modulation of synaptic strength.

The third type of ionotropic glutamate receptor, the NMDA receptor, which is widely distributed in different types of CNS neurons, differs fundamentally from AMPA and kainate receptors in several ways. First, the pore is in general significantly more permeable to Ca^{2+} ions, and also to K^+ and Na^+ ions. The high Ca^{2+} permeability of the NMDA receptor-associated channel probably contributes to its often predominant role in excitotoxicity. In contrast to AMPA receptors, classical NMDA receptors exhibit a high single-channel conductance and desensitize much more slowly. Second, the opening of the ligand-gated cation channel not only depends on binding of agonist but is also voltage dependent,

since the channel is blocked by Mg^{2+} ions at resting membrane potentials, and depolarization of the plasma membrane is required to relieve the Mg^{2+}-dependent block. Third, activation of NMDA receptors requires binding of a co-agonist to the so-called glycine-binding site of the NMDA receptor. In contrast to the characteristics of these so-called classical NMDA receptors, composed of NR1 and NR2A, B, C, or D subunits, the presence of the newly recognized NR3A or B subunits can alter these properties, making channels that are less permeable to Ca^{2+} and far less sensitive to Mg^{2+} blockade.

The implication of different ionotropic glutamate receptor types in induction of brain damage has been further studied by the use of selective receptor antagonists. Experimental administration of competitive or uncompetitive NMDA antagonists or application of kainate antagonists attenuates cell death in cell cultures. Unfortunately, until recently, the clinical use of these receptor antagonists has been hampered by the inhibition of physiological NMDA receptor-mediated neurotransmission, resulting in various adverse effects. However, the uncompetitive NMDA antagonist memantine (1-amino-3,5-dimethyladamantane hydrochloride), which has been used for treatment of Parkinson's disease and Alzheimer's disease, is sufficiently tolerated because of it unique mechanism of action, as first shown in the laboratory of Stuart Lipton. Memantine is an open-channel, uncompetitive blocker, but has faster kinetics, particularly 'off-rate,' than does MK-801. This results in substantial inhibitory drug action under conditions of prolonged or pathological exposure to glutamate but much less inhibition under millisecond exposure to glutamate as occurs during normal neurotransmission.

Whereas the role of ionotropic glutamate receptors in excitotoxicity has been studied extensively, the significance of metabotropic glutamate receptors in excitotoxicity is less well understood. It appears that specific mGluR subtypes can be involved in modulation of ionotropic receptor-mediated excitotoxicity. Group I mGluRs comprise the subtypes $mGluR_1$ and $mGluR_5$, which in heterologous expression systems couple to phospholipid hydrolysis through phospholipase C. Among other effects, this results in second-messenger-mediated activation of protein kinase C, and Ca^{2+} release from inositol 1,4,5-trisphosphate (IP_3)-sensitive Ca^{2+} stores. Application of recently developed selective receptor ligands suggests that activation of group I, and particularly mGluR1, receptors can exacerbate NMDA-mediated excitotoxicity.

Group II receptors comprise the subtypes $mGluR_2$ and $mGluR_3$, which are negatively coupled to the adenylate cyclase pathway. Studies using selective agonists indicate that group II receptor activation results in protection against NMDA-mediated excitotoxicity. Likewise, group III receptors ($mGluR_4$ and $mGluR_{6-8}$) are negatively coupled to adenylate cyclase, and activation of $mGluR_4$ or $mGluR_8$ group III receptors results in attenuation of NMDA-mediated neurotoxicity.

Ionic Dependence of Excitotoxic Cell Damage

Activation of different ionotropic receptor types is linked to excitotoxic cell damage through their underlying ionic currents. Depending on the predominance of either Na^+ or Ca^{2+} influx, two different components of excitotoxicity have been suggested. Na^+ ion influx mediated by activation of NMDA-type and non-NMDA-type glutamate receptors is followed by secondary influx of Cl^- and H_2O and results in swelling of neurons. This acute form of cell damage depends on the transmembrane Na^+ and Cl^- gradients and can be prevented by extracellular substitution of Na^+ and Cl^- with impermeant ions. In contrast to the acute, primarily Na^+-dependent osmotic damage, a more delayed mode of cell death has been attributed to Ca^{2+} influx. This delayed Ca^{2+}-dependent cell death can be largely attenuated by inhibition of NMDA receptors, removal of extracellular Ca^{2+} ions, or buffering of cytoplasmic Ca^{2+} by membrane-permeable chelators. Therefore, the Ca^{2+}-dependent component was traditionally believed to be induced exclusively by NMDA receptor activation.

The finding that the late component can be mimicked by calcium ionophores in presence of Ca^{2+} ions has corroborated the Ca^{2+} dependence of the delayed cell death. However, the discovery of Ca^{2+}-permeable AMPA/kainate receptors has lead to the view that neurons expressing these Ca^{2+}-permeable non-NMDA receptors can also undergo Ca^{2+}-dependent delayed cell death. In summary, different routes of Ca^{2+} influx, such as through NMDA receptors, Ca^{2+}-permeable AMPA/kainate receptors, and voltage-dependent Ca^{2+} channels, which are all involved in physiological signaling, are implicated in excitotoxic cell death.

The suggested key role of Ca^{2+} in excitotoxicity has been subsequently confirmed by the finding that the extent of excitotoxic cell death correlates with the total amount of Ca^{2+} uptake and is independent of the route of entry. The Ca^{2+} overload observed following sustained stimulation of NMDA receptors results from the inability of cellular Ca^{2+} homeostasis, such as extrusion of Ca^{2+} across the plasma membrane by the Na^+/Ca^{2+} antiporter and Ca^{2+} ATPase or Ca^{2+} sequestration by the endoplasmic reticulum and mitochondria to remove the large influx of Ca^{2+}.

Interestingly, during sustained exposure to glutamate, the intracellular calcium concentration ($[Ca^{2+}]_c$) rises rapidly to a peak value and thereafter slowly recovers to an elevated plateau level. However, the $[Ca^{2+}]_c$ can also manifest a second, delayed increase, immediately preceding cell death. This delayed $[Ca^{2+}]_c$ increase is irreversible and reflects irreversible loss of cellular Ca^{2+} homeostasis. Although the interval between the initial Ca^{2+} spike and the delayed rise varies within a cell population exposed to NMDA receptor agonists, once this rise occurs within a neuron, it invariably precedes cell death. The temporal relation between neuronal Ca^{2+} levels and delayed cell death suggests that Ca^{2+}-dependent effector mechanisms are involved, which do not need sustained high Ca^{2+} levels but are triggered by transient changes in $[Ca^{2+}]_c$. Importantly, mild excitotoxic damage often leads to apoptotic-like neuronal cell death, while more intense insults, which decrease energy production and thus disrupt ionic homeostasis, result in osmotic lysis or necrosis.

Since many enzymes are activated by transient or sustained $[Ca^{2+}]$ elevation, various effector mechanisms may be implicated in Ca^{2+}-mediated excitotoxic cell death. Consequently, a variety of Ca^{2+}-dependent hydrolytic enzymes, including lipases and proteases, have been suggested to be involved in excitotoxic neuronal damage. Activation of the Ca^{2+}-dependent phospholipase A2 has been observed following NMDA receptor activation, and the subsequent catabolism of released arachidonic acid by lipoxygenases and cyclooxygenases (COXs) is also associated with concomitant production of reactive oxygen species (ROS). In addition, activation of phospholipase A2 and subsequent release of arachidonic acid may inhibit transporter-mediated glutamate uptake from the extracellular space.

Among several Ca^{2+}-activated proteases, the activity of the Ca^{2+}-dependent cysteine protease calpain is increased following glutamate receptor-mediated Ca^{2+} loading. Calpain activation results in proteolysis of structural proteins and degradation of the neuronal cytoskeleton. Recently, Pierluigi Nicotera's group has shown that calpain-mediated cleavage of Na^+/Ca^{2+} exchanger results in a massive, secondary Ca^{2+} overload after excitotoxic insult and therefore favors necrotic over apoptotic cell death.

Ca^{2+}-mediated activation of nitric oxide synthase is also a major pathway involved in excitotoxic cell death, since neurons constitutively express the isoenzyme neuronal nitric oxide synthase (nNOS). nNOS is activated by glutamate receptor-mediated $[Ca^{2+}]_c$ increases, and there is compelling evidence that nNOS activation is linked to excitotoxic damage because inhibition of NO formation results in protection of neurons from glutamate receptor-mediated cell death. Studies on nNOS-deficient neuronal cultures have confirmed the role of NO in glutamate receptor-mediated neurotoxicity, since in nNOS-deficient cortical cultures the toxic effects elicited by administration of NMDA are markedly attenuated. The role of NO in excitotoxicity is mediated at least partly through the reaction of NO with superoxide anions (O_2^-) to form peroxynitrite ($ONOO^-$).

Mitochondrial Dysfunction

Studies on the time course of glutamate receptor-induced $[Ca^{2+}]_c$ increase have demonstrated that if the neuronal cell does not succumb to the insult, $[Ca^{2+}]_c$ slowly recovers after the termination of agonist application. This recovery is due to Ca^{2+} homeostatic mechanisms such as extrusion across the plasma membrane by the Na^+/Ca^{2+} exchanger, as mentioned previously, and the Ca^{2+}-ATPase, or sequestration of Ca^{2+} into the endoplasmic reticulum and mitochondria. In theory, the driving force for Ca^{2+} ions to enter the mitochondrial matrix in energized mitochondria is formed by the strong electrochemical gradient for Ca^{2+} across the inner mitochondrial membrane, mainly due to the mitochondrial membrane potential, $\Delta\psi_m$. Normal energized mitochondria can take up substantial amounts of calcium. Net uptake of calcium occurs in isolated mitochondria whenever $[Ca^{2+}]_c$ rises above a set point in the high nanomolar range, indicating the dynamic equilibrium of mitochondrial uptake and extrusion mechanisms. Ca^{2+} uptake into mitochondria results in a decrease in $\Delta\psi_m$ that can be monitored in isolated energized mitochondria as well as *in situ*. This depolarization is due to the balancing of the charge transfer carried by Ca^{2+} ions by reentry of protons into the mitochondrial matrix. Under normal conditions, Ca^{2+}-induced depolarization is transient and serves to activate mitochondrial dehydrogenases and the mitochondrial ATP synthase. However, studies on *in situ* mitochondrial membrane potential using cationic fluorochromes have demonstrated convincingly that the time course of $\Delta\psi_m$ depolarization can vary considerably, and depends on the time of Ca^{2+} loading. Short pulses of NMDA result in at least partial recovery of $\Delta\psi_m$, whereas prolonged NMDA exposure or higher agonist concentrations result in sustained depolarization with negligible recovery. This sustained depolarization indicates that excessive Ca^{2+} uptake triggers mitochondrial bioenergetic failure. Sustained depolarization has been ascribed to various underlying mechanisms, and ROS seem to play a key role in several of the suggested hypotheses.

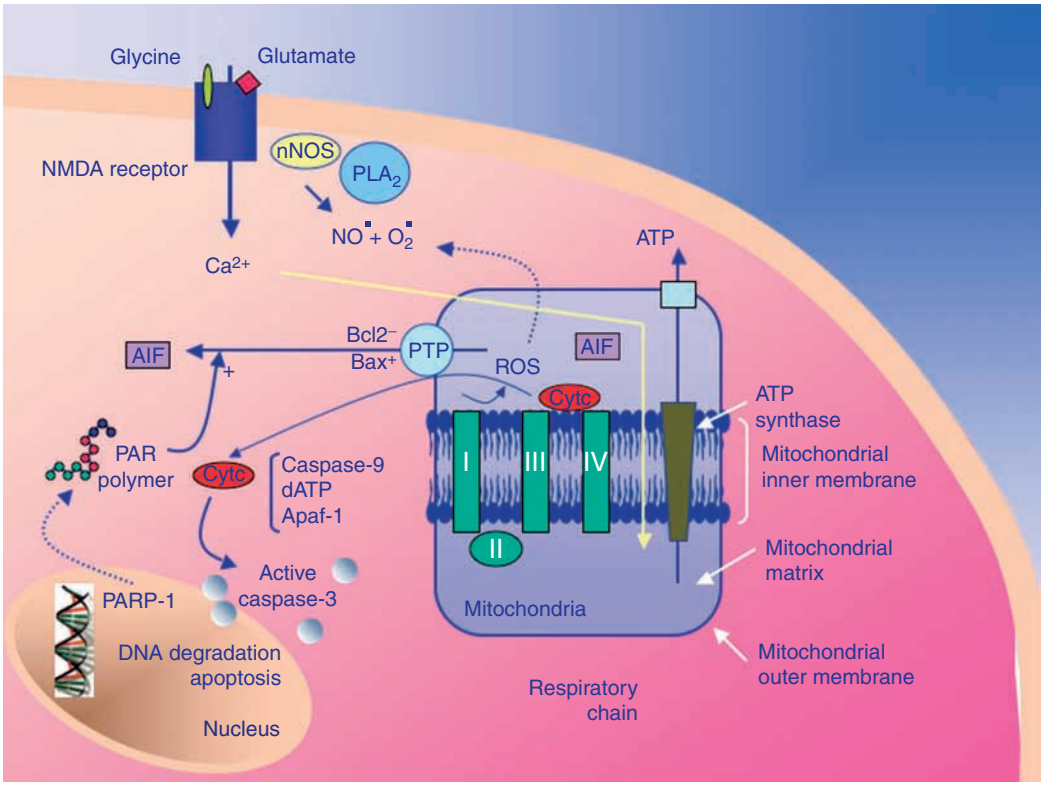

Figure 1 Schema of cell death pathways triggered by excessive *N*-methyl-D-aspartate (NMDA) receptor activity. The cascade of steps leading to apoptotic-like neuronal cell injury and death include NMDA receptor hyperactivation, leading to excessive Ca^{2+} influx, Ca^{2+} uptake into mitochondria, and subsequent mitochondrial dysfunction; toxic effects of free radicals such as nitric oxide (NO) and reactive oxygen species (ROS); and activation of apoptosis-inducing enzymes, including caspases and apoptosis-inducing factor (AIF). nNOS, nitric oxide synthase; PLA_2, phospholipase A_2; Cytc, cytochrome *c*; PAR, poly(ADP-ribose); PARP-1, PAR polymerase-1; PTP, protein tyrosine phosphatase.

In neurons, as in other tissues, mitochondria are a major source of ROS; nonetheless, mitochondria are susceptible to oxidative damage and nitrosative stress. Excessive Ca^{2+} uptake into isolated mitochondria is known to result in increased formation of reactive oxygen species, which in turn inhibit pyruvate dehydrogenase and tricarboxylic acid cycle enzymes as well as complex I of the respiratory chain. In intact neurons, ROS formation has been monitored with redox-sensitive fluorochromes, which are oxidized by ROS to form fluorescent molecules. These studies indicate that ionotropic glutamate receptor-induced Ca^{2+} loading enhances the production of O_2^-. Additionally, mitochondrial dysfunction triggered by mild excitotoxic insult can lead to both caspase-dependent and -independent apoptotic-like cell death. Several studies have shown that mild NMDA receptor-mediated insults lead to mitochondrial calcium overload, cytochrome *c* release, and caspase activation. Recently, NMDA toxicity has also been shown to be mediated in part via apoptotic-inducing factor (AIF) release from mitochondria, which is triggered by poly(ADP-ribose) (PAR) polymer generated from the activation of PAR polymerase-1 (PARP-1). Figure 1 summarizes key features of several of these excitotoxic pathways.

Role of Reactive Oxygen Species in Excitotoxicity

Evidence for the implication of ROS in excitotoxic cell damage arises from experiments showing enhanced production of O_2^- following ionotropic glutamate receptor and from studies using radical scavengers or inhibitors of the formation of certain ROS. These studies unequivocally demonstrate that removal of ROS results in attenuation of glutamate receptor-induced cell death.

Several Ca^{2+}-dependent processes that increase the endogenous production of ROS have been described, which are all assumed to be implicated in excitotoxic cell death following NMDA receptor activation. As stated previously, Ca^{2+} loading into mitochondria appears to be one important mechanism of ROS production under conditions associated with excitotoxicity. Apart from this mechanism, activation of Ca^{2+}-dependent phospholipase A2 has been observed following NMDA receptor activation, and catabolism

of released arachidonic acid by lipoxygenases and cyclooxygenases has been implicated in excitotoxicity through the concomitant production of ROS. In addition to Ca^{2+}, intracellular Zn^{2+} can be involved in excitotoxic pathways. During excitotoxic insult, peroxynitrite-induced Zn^{2+} release from intraceullular stores can trigger mitochondrial production of ROS, contributing to a p38 mitogen-acitvated protein kinase (MAPK)-dependent apoptotic pathway.

Some ROS exhibit a high reactivity with correspondingly short half-lives and can undergo many different reactions. Generally speaking, ROS can exert multiple damaging reactions to proteins, lipids, carbohydrates, and nucleic acids, thereby disrupting cellular functions. The increased production of ROS is therefore a potential threat to cellular homeostasis and neuronal survival if production is not balanced by the capacity of endogenous antioxidant mechanisms.

Apart from the inhibition of enzyme activity, the opening of a large nonselective pore in the inner mitochondrial membrane, termed 'mitochondrial permeability transition' (mPT), has been suggested as a key mechanism in neuronal excitotoxicity. The basic concept of a pore-mediated permeabilization of the inner mitochondrial membrane, rather than membrane damage, had been suggested as early as the 1970s. The early studies by Haworth and Hunter showed that various divalent cations could induce or prevent pore-mediated mitochondrial swelling. Dissipation of $\Delta\psi_m$, bioenergetic failure, and disturbance of mitochondrial Ca^{2+} homeostasis following activation of NMDA receptors all could be explained by opening of the mPT. However, conclusions regarding the involvement of mPT in neuronal excitotoxicity until recently were based on pharmacological experiments using the mPT inhibitor cyclosporine A and should be treated cautiously, since cyclosporine A also inhibits calcineurin and the multidrug-resistance channel. Our group has demonstrated that bongkrekic acid, a more specific inhibitor of mPT, could prevent NMDA-induced neuronal apoptosis, suggesting the involvement of mPT in this form of excitotoxic cell death.

Role of Nitric Oxide and Other Reactive Nitrogen Species in Excitotoxicity

Nitric oxide is a messenger molecule in the CNS that affects various cellular functions (e.g., neuronal transmitter release, synaptic plasticity, and gene expression). NO is produced by several types of cells expressing three distinct isoforms of the enzyme nitric oxide synthase (NOS; NADPH-diaphorase), which converts L-arginine into NO and citrulline: (1) neuronal NOS, (2) inducible NOS (iNOS) in microglia, and (3) endothelial NOS (eNOS) in endothelial cells of brain blood vessels. Unlike classical neurotransmitters, NO is freely diffusible and therefore cannot be stored in synaptic vesicles. Once synthesized, NO diffuses across cell membranes and thus can reach various compartments within the NO-producing cells and different types of surrounding cells. The absence of control over the activity of NO by release and reuptake mechanisms is functionally compensated by a tight regulation of its synthesis. Pharmacological experiments using NO donors, NOS inhibitors, or exogenous substrates (such as reduced hemoglobin) for competitive reactions with NO have suggested that NO is linked to neuronal cell death. Importantly, cell death under conditions assumed to induce excitotoxicity could be largely attenuated by inhibition of NO formation or blockade of NO effects. Subsequent studies on nNOS-deficient neuronal cultures confirmed the role of NO in glutamate receptor-mediated neurotoxicity, since toxic effects elicited by administration of NMDA are markedly attenuated in nNOS-deficient cortical cultures, compared with wild-type neurons. Since nNOS activity is known to increase with increasing calcium concentration, NMDA receptor-mediated calcium influx provides a mechanistic link between an excitotoxic insult and NO-mediated cell damage.

Although NO is a free radical, NO is not as reactive as are some ROS (such as hydroxyl radical), and the predominant contribution of NO to excitotoxicity has been suggested to depend on increased NO formation with concomitant production of O_2^-. Under this condition, O_2^- reacts with NO extremely fast to form peroxynitrite ($ONOO^-$). $ONOO^-$ can undergo a multitude of different chemical reactions with various substrates – for example, hydroxylation or nitration of tyrosine residues, lipid peroxidation, disulfide formation, or reaction with NO_2-like and OH-like properties; subsequently, this toxic molecule has been implicated in NO-induced cell damage. To date, several pathways of $ONOO^-$ toxicity have been suggested, but with respect to its high reactivity, it appears unlikely that these are the only mechanisms involved in NO-mediated cell death. Intense exposure to NO/O_2^- with resultant $ONOO^-$ formation has been convincingly demonstrated to result in neuronal necrosis because of energy failure, while mild insults lead to apoptosis.

Recently, S-nitrosylation (transfer of the NO group to a critical cysteine thiol, to form an SNO-protein) has been shown to modulate protein function. In some cases, neuroprotective pathways are initiated by such reaction – for example, by downregulation of excessive NMDA receptor activity or caspase activity. However, in several instances, S-nitrosylation of particular targets contributes to neurotoxicity.

Examples of such proteins that are nitrosylated to trigger neurodegenerative pathways and contribute to necrosis, apoptosis, or dendritic/synaptic damage include SNO-GAPDH, SNO-matrix metalloproteinases (MMPs), SNO-parkin, and SNO-protein disulfide isomerase (PDI).

NMDA Receptors on Glial Cells

In addition to NMDA receptors found on neurons, as already discussed, three studies have recently reported evidence for NMDA receptors on oligodendrocytes. Moreover, these receptors have been show to be pathophysiologically relevant to oligodendrocyte and myelin function in that overstimulation of the receptors could cause damage. The NMDA receptor-mediated currents of oligodendrocytes were made up not only of NR1 and various NR2 NMDA receptor subunits, but also of NR3A subunits. As we and our colleagues had previously shown, NR3 subunits can render NMDA receptors much less sensitive to blockade by Mg^{2+}, as was indeed found to the be case for the NMDA receptors on oligodendrocytes. Similarly, it was recently reported that the NMDA receptor-mediated currents found on cortical astrocytes were also highly resistant to Mg^{2+} block, raising the possibility that NR3 subunits may contribute to those functional receptors as well, and could be involved in pathological responses of glial cells.

Therapeutic Implications and Future Directions

The seminal 1984 publication of Roger Simon, Brian Meldrum, and colleagues showed that excessive NMDA receptor activity underlies ischemic damage to neurons. As alluded to in the preceding discussions, several major neurodegenerative conditions are also thought to manifest a component of NMDA receptor-mediated neurotoxicity. The newer studies showing the presence of NMDA receptors on glial cells suggest that white-matter diseases, such as periventricular leukolmalacia in infants and multiple sclerosis in adults, could have components of NMDA receptor-mediated neurotoxicity. Several reports have claimed that all NMDA receptor antagonists have failed in clinical trials. However, this is not correct, and this misinformation was then followed by the further misplaced claims that the failure of clinical trials occurred because white matter did not have NMDA receptors; hence, NMDA receptor antagonists were said not to be useful in preventing ischemic damage because white matter (axons and their surrounding myelin sheaths) and astrocytes are injured in addition to gray matter (neurons).

Actually, most NMDA receptors antagonists failed in prior clinical trials because they were poorly tolerated by humans at neuroprotective doses. Many prior studies did not use neuroprotective doses of NMDA receptor antagonists because such doses proved to be too toxic, or at least exhibited severe side effects. After all, the NMDA receptor is responsible for normal synaptic transmission contributing to long-term potentiation (LTP) and other important processes in many areas of the brain. Therefore, high-affinity, effective NMDA receptor antagonists block normal brain activity and lead to intolerable side effects, such as hallucinations, sleepiness, and even coma.

In fact, what was needed was a more gentle type of NMDA receptor antagonist, one that relatively spared normal synaptic activity while blocking excessive (pathological, probably representing extrasynaptic) activity. One such example of a clinically tolerated NMDA receptor antagonist that has passed multiple clinical trials is memantine, which, as mentioned previously, was recently approved by the Food and Drug Administration for the treatment of moderate to severe Alzheimer's disease and is also used clinically in Europe for vascular (multi-infarct or multiple stroke) dementia. While memantine may also interact with α7-nicotinic and 5-hydroxytryptamine (5-HT_3) channels, its major mode of action at clinically used concentrations is at the NMDA receptor. Our group showed that memantine is clinically tolerated because of its mode of action on the NMDA receptor – it is an uncompetitive antagonist that acts as an open-channel blocker, and, most importantly, with a relatively fast 'off-rate.' Uncompetitive action means that, at a fixed dose, the drug is relatively more effective against higher concentrations of glutamate than lower. This somewhat surprising phenomenon occurs because the action of an uncompetitive antagonist is contingent upon prior activation of the receptor by agonist, and more activation of the receptor is blocked to a greater degree. Additionally, the fast off-rate is a critical factor contributing to a low-affinity mechanism of action and thus to clinical tolerability. This tolerability in large measure occurs because memantine does not accumulate in NMDA receptor-operated channels because of the fast off-rate. Hence, memantine does not block subsequent physiological neurotransmission. This mode of drug action, employing an uncompetitive, fast off-rate (UFO), has recently led to the concept of clinically tolerated 'UFO' drugs, apropos since drug is present only when needed and then vanishes.

It has been argued that memantine has only a modest effect on rodent models of cerebral ischemia; in contrast, blocking Ca^{2+} influx occurring via acid-sensing channels (ASICs) or TMP7 channels appears

to be more effective. However, it is possible that activation of these channels is downstream of hyperstimulation of NMDA receptors, and if sufficient protection is afforded by adequately blocking NMDA receptors, the activation of these channels might not come into play. Along these lines, prior reports often used a submaximal dose of memantine, so the full neuroprotective, yet clinically tolerated, action of the drug was not achieved. In point of fact, a recent publication has shown that NMDA receptors may be very important in neuronal cell death from ischemia, because excessive NMDA receptor activity leads to phosphorylation and thus activation of ASICs. Similarly, TMP7 channels are activated only after severe hypoxic insults, which might be averted with adequate NMDA receptor antagonist treatment. Thus, all of these types of channels are important, but the most recent studies converge to the view that NMDA receptors may be even more critical in cerebral ischemia and other neurodegenerative disorders than has been appreciated in past years. Hence, as of this writing, we express a sense of renewed optimism that the recent, albeit modest, clinical success of memantine in Alzheimer's disease may in fact lead to similar treatment of other neurodegenerative diseases, cerebral ischemia, multiple sclerosis, cerebral palsy, spinal cord injury, glaucoma, and, therefore, both white-matter and gray-matter diseases with UFO-type NMDA receptor antagonists. Only further experimentation will tell if this premise holds true.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Animal Models of Alzheimer's Disease; Animal Models of Parkinson's Disease; Axonal Regeneration: Role of Growth and Guidance Cues; Cell Replacement Therapy: Parkinson's Disease; Epilepsy: Neuronal Death; Glial Responses to Injury; Glial Glutamate Transporters; Huntington's Disease: Neurodegeneration; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; NMDA Receptors and Disease + C464; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Regeneration and Functional Recovery: Strategies.

Further Reading

Andrabi SA, Kim NS, Yu SW, et al. (2006) Poly(ADP-ribose) (PAR) polymer is a death signal. *Proceedings of the National Academy of Sciences of the United States of America* 103: 18308–18313.

Ankarcrona M, Dypbukt JM, Bonfoco E, et al. (1995) Glutamate-induced neuronal death: A succession of necrosis or apoptosis depending on mitochondrial function. *Neuron* 15: 961–973.

Bano D, Young KW, Guerin CJ, et al. (2005) Cleavage of the plasma membrane Na^+/Ca^{2+} exchanger in excitotoxicity. *Cell* 120: 275–285.

Beal MF (1996) Mitochondria, free radicals, and neurodegeneration. *Current Opinion in Neurobiology* 6: 661–666.

Budd SL and Nicholls DG (1996) Mitochondria, calcium regulation, and acute glutamate excitotoxicity in cultured cerebellar granule cells. *Journal of Neurochemistry* 67: 2282–2291.

Chen H-SV and Lipton SA (2006) The chemical biology of clinically tolerated NMDA receptor antagonists. *Journal of Neurochemistry* 97: 1611–1626.

Dawson VL, Dawson TM, London ED, et al. (1991) Nitric oxide mediates glutamate neurotoxicity in primary cortical cultures. *Proceedings of the National Academy of Sciences of the United States of America* 88: 6368–6371.

Gu Z, Kaul M, Yan B, et al. (2002) S-Nitrosylation of matrix metalloproteinases: Signaling pathway to neuronal cell death. *Science* 297: 1186–1190.

Hara MR, Agrawal N, Kim SF, et al. (2005) S-Nitrosylated GAPDH initiates apoptotic cell death by nuclear translocation following Siah1 binding. *Nature Cell Biology* 7: 665–674.

Hardingham GE, Fukunaga Y, and Bading H (2002) Extrasynaptic NMDARs oppose synaptic NMDARs by triggering CREB shut-off and cell death pathways. *Nature Neuroscience* 5: 405–414.

Lipton SA (2006) Paradigm shift in neuroprotection by NMDA receptor blockade: Memantine and beyond. *Nature Reviews Drug Discovery* 5: 160–170.

Lipton SA, Choi YB, Pan ZH, et al. (1993) A redox-based mechanism for the neuroprotective and neurodestructive effects of nitric oxide and related nitroso-compounds. *Nature* 364: 626–632.

Lipton SA and Rosenberg PA (1994) Excitatory amino acids as a final common pathway for neurologic disorders. *New England Journal of Medicine* 330: 613–622.

Olney JW and Sharpe LG (1969) Brain lesions in an infant rhesus monkey treated with monosodium glutamate. *Science* 166: 386–388.

Yao D, Gu Z, Nakamura T, et al. (2004) Nitrosative stress linked to sporadic Parkinson's disease: S-nitrosylation of parkin regulates its E3 ubiquitin ligase activity. *Proceedings of the National Academy of Sciences of the United States of America* 101: 10810–10814.

Executive Function and Higher-Order Cognition: Assessment in Animals

J D Schall, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

General Issues for Animal Testing

Any use of animals in research is governed by laws interpreted by regulatory agencies and local institutions with which researchers must be conversant and compliant. Studies of cognition of necessity require measurements of behavior, and the behavior must be controlled in the context of the experiment. Verbal instructions are not effective, so animals must be coaxed to perform a task. The motivation to perform the task can be appetitive or aversive. Appetitive control occurs through pairing an event or behavior with a food or fluid reinforcement. Aversive control occurs through pairing an event or behavior with an undesirable stimulus such as electric shock or a brief loud noise. Aversive control can be used only if scientifically justified as, for example, in studies of anxiety.

Classical (Pavlovian) conditioning is accomplished by repeatedly pairing a conditioned stimulus (e.g., a tone) with an unconditioned stimulus (e.g., a puff of air) that reliably elicits an unconditioned response (e.g., an eye blink). The sensitivity of the animal to the conditioned stimulus can then be assessed by measuring the occurrence or magnitude of the unconditioned response as some characteristic of the conditioned stimulus (intensity, frequency, etc.) is varied systematically. Operant (instrumental) conditioning is accomplished by reinforcing an emitted behavior (e.g., looking at a spot of light) with a reward (e.g., a food item or fluid) or a punishment (e.g., a mild electric shock). Behaviors followed by positive reinforcements are more likely to be emitted in the future by an organism motivated to receive the reward. Behaviors that are followed by a negative reinforcement are less likely to be emitted in the future. Decades of research has detailed many principles and nuances of operant conditioning that the interested student can learn from a textbook on learning and memory. Also, detailed procedures for various species and situations can be found in the methods sections of peer-reviewed publications. Finally, we should also note the obvious specializations of behavior and associated cognition across the phylogenetic scale.

Taxonomy of Cognition

Stimuli can be presented and overt responses measured. However, to explain orderly and arbitrary relationships between responses and stimuli, it is now regarded as useful, if not necessary, to hypothesize certain covert processes mediating the encoding, selection, and categorization of stimuli and preparation of responses.

Before reviewing methods used to investigate higher order cognition in animals, we should begin by determining how to distinguish cognition from simple sensation and movement. We can also formulate at least a provisional minimal taxonomy of cognition. Reflexes are the most direct link between sensory and motor processes. It could be argued that any behavior that is not a reflex must include some cognitive component. It could also be argued that only behaviors that arise from or are concomitant with conscious mental content qualify as cognition. For the purposes of this article, we can be satisfied with an intermediate position that only holds that processes that intervene between or adaptively adjust sensation and movement constitute cognition. If so, then we can ask, how many cognitive processes can be distinguished reasonably? We will work with the following list:

- Perception, including categorization
- Attention
- Memory, including long-term and working memory
- Response selection and preparation
- Emotion
- Executive control

We are also interested in more complex activities that involve all of the processes listed previously:

- Decision making
- Problem solving
- Communication

Less space is devoted to processes that are described elsewhere in detail, and more space is devoted to less well-studied processes.

Methods to Study Perception

Discrimination and Categorization

Distinguishing the characteristics of stimuli for the purpose of guiding behavior has been investigated in many species. The testing requires designing alternative stimuli appropriate for the sensory apparatus of the particular animal and is based on the principle that two (or more) physically distinct stimuli can be rendered indistinguishable from one another by manipulation of

the magnitude of their physical differences. Thus, a natural component of an animal's behavior is placed under control of the stimulus and then the characteristics of the stimuli are varied to determine how sensitive the animal is to the variation of the stimulus. For example, monkeys can be trained to shift gaze or press bars to signal whether a visual stimulus has a particular color, shape, or pattern of motion. Similarly, rats can be trained to perform tactile discriminations with their whiskers, such as judging the width of an aperture, the texture of a surface, or the distance of a gap, or also to signal their discrimination of mixtures of odors. Animals are capable of making more complex sensory discriminations. For example, animals as diverse as monkeys and pigeons can learn rapidly to discriminate among color photographs of different categories of objects, such as trees, humans, fish, and water, and these discriminations are generalized to novel examples from the categories of objects.

Awareness

A fruitful line of inquiry in the past decade has investigated neural correlates of awareness in monkeys. This has been accomplished by presenting a single ambiguous stimulus that can be perceived as either of (usually) two alternatives to macaque monkeys trained to report which of the alternatives they perceive. One way to dissociate the presentation of a stimulus from the phenomenal awareness of that stimulus is to employ binocular rivalry. Binocular rivalry occurs when markedly different stimuli are presented to the two eyes (e.g., horizontal stripes to the left eye and vertical stripes to the right). The perception is not of some plaid combination but instead of the stimulus in one eye and then the other, alternating randomly until one or both stimuli are removed.

Another means of dissociating presentation from report is through visual masking. This is accomplished by presenting a stimulus that may be weak in some respect but not below threshold to detect and then rapidly thereafter presenting a stronger stimulus at the same or nearby location. The presence of the subsequent stronger stimulus results in a lack of awareness of the earlier stimulus; this is referred to as backward masking. Critical to these studies is some objective measure of how the sensory system registers the stimuli because animals are very likely to respond in a biased and stereotyped manner to ambiguous stimuli unless reinforcement contingencies prevent this.

Methods to Study Attention

To study the allocation of attention, the most common procedure is to present a cue that informs the animal that a response is required for a subsequent stimulus at the same location. In a fraction of trials, the test stimulus is presented at an uncued location. The difference in response time or discriminability of the test stimulus at the cued versus the uncued location is taken as a measure of the allocation of attention. Alternatively, multiple stimuli can be presented simultaneously with one of the stimuli distinguished by sensory properties, a prior cue, or a memory template to guide the response. The quality of response to the target in the presence versus the absence of distractor stimuli is taken as a measure of the allocation of attention. The more diverse the distractors' appearance and the more similar the distractors are to the target, the more attention demanding the task becomes. In addition, a stimulus with multiple attributes (e.g., color and shape) can be presented following a cue that instructs the animal to respond according to one of the attributes. The demand for attention can be manipulated by making the cued attribute more or less difficult to discriminate and by making the noncued attribute more or less potent. Note that in studies of the allocation of visual attention in space, it is essential to monitor gaze to prevent confounds introduced by movements of the eyes.

Attention has been investigated predominantly with nonhuman primates, but procedures have also been devised for rodents and even birds such as pigeons. Procedures to assess attention in a less specific sense have been formulated to study mouse models of attention deficit disorder. For example, mice can be trained to poke their nose into a hole to receive a reinforcement in response to a visual stimulus. By offering multiple alternative holes signaled by different lights turned on for short durations, the mice must attend to the stimuli to respond quickly enough to receive reinforcement. The attention capacity of the mice can then be assessed by manipulations of target light presentation rate, duration, salience, and temporal predictability.

Methods to Study Memory

Memory can be inferred by persistent and consistent changes in behavior following learning. An overview of classical and operant conditioning learning procedures was provided previously. Memory can be divided into short-term working memory and long-term procedural and declarative memory.

Short-Term Working Memory

Short-term memory has been studied using delayed-response tasks in which a stimulus or set of stimuli are presented briefly but a response cannot be produced

until after a retention interval typically lasting approximately 1–10 s. Animals can guide their responses by the location or identity of the stimuli. The task has been elaborated in various ways. Delayed alternation requires monkeys to select the object at the location that was not selected in the previous trial (delayed spatial alternation) or to select the object that was not selected in the previous trial (delayed object alternation). This task requires working memory retention across as well as within trials. Delayed match to sample exposes monkeys briefly to an object and then after an interval presents the target object plus a distractor object. Monkeys are reinforced for selecting through a gaze shift or a reaching movement the target object that they had observed previously. Delayed nonmatch to sample reverses the logic, requiring monkeys to select the object they had not observed previously. It is important to note that delayed-response tasks require not only working memory but also response inhibition to prevent premature movements and may also invoke response preparation if the monkey can anticipate the termination of the retention interval when it can respond.

Long-Term Memory

Memory persisting over days or months must be tested in a species-specific manner. For example, rats are commonly tested using spatial navigation through locomotion in a maze to locate a reinforcement. Also, the water maze was developed to assess spatial learning and memory in rats. It is commonly used with mice, but species and strain idiosyncrasies must be recognized in the interpretation of the results. A platform is placed at a particular location beneath the surface of a round pool of murky water. When placed in the pool, mice swim to locate the hidden platform, and learning can be measured by a reduction in the amount of time taken to locate the platform as the mice or rats encode orienting cues in the environment. Contrasting performance when the platform is hidden with that when it is visible permits resolving confounds of sensory or motor impairments.

Implicit in the description of all the complex tasks monkeys can be trained to perform is the fact that they have long-term memory. The nature and capacity of that memory has been probed largely through tests of categorization of stimuli and of relations among stimuli. More evidence for long-term memory is derived from learning sets (learning to learn), an improvement in the speed of acquisition of successive discriminations between arbitrary objects or of reversing responses when reinforcement changes. Learning set has been demonstrated in nonhuman primates as well as in dogs, rats, and some birds.

Methods to Study Response Selection and Preparation

All motor responses can be explained in terms of causes (the muscle contracted because the neuron fired), but some responses can be explained also in terms of reasons (the bar was pressed to obtain the food to satisfy the hunger). This section considers movements that can be explained in terms of reasons and not just causes.

Response Selection

Requiring an arbitrary response to a particular stimulus invokes a process referred to as stimulus response mapping. A wealth of studies of human subjects can inform the design of experiments with animals; however, it should be clear that different species are incapable of producing more or less complex mappings. The capacity of rodents to learn arbitrary responses through operant conditioning is more limited than that of macaque monkeys, for example. Studies with macaque monkeys have explored progressively more complicated degrees of mapping a response onto a stimulus so that we can say that monkeys are encoding instructions and following rules or even strategies.

At the lowest level, monkeys can produce an arbitrary response not directed toward a stimulus – for example, a shift of gaze to a location directly opposite a target (antisaccade). At the next level, macaque monkeys can be trained to produce arbitrary responses to specific cues on successive trials and to change the mapping if the reinforcement contingencies change (conditional motor learning). Monkeys can also be trained to produce an arbitrary sequence of a few movements guided by particular cues. Perhaps the greatest expression of this ability is seen in apes that learn sign language. In studies with humans, evidence for response selection is an increase in the latency to initiate the first movement with an increase in the number of movements in the sequence. Unfortunately, animal studies have not performed this simple test. Beyond producing sequences of movements in response to sequences of stimuli, monkeys can be trained to produce self-ordered sequences of responses in which successive responses are guided by previous responses.

At perhaps the highest level, macaque monkeys can be trained to respond according to alternative strategies that require flexibility in applying rules with no fixed stimulus–response mapping. For example, monkeys can switch between repeat-stay and change-shift strategies. To perform a task such as this, monkeys must remember the previous response, remember the stimulus or stimulus–response pair preceding that choice, use that memory to evaluate

whether the stimulus on the current trial was different from that on the previous trial, and apply the strategy to choose the same target when the stimulus repeated and to reject that target when the stimulus changed.

Response Preparation

The time of an overt response to a given stimulus is variable and unpredictable. Within that unpredictability, however, certain trends have been observed. For example, when given a warning ('ready') before an imperative trigger signal ('go'), subjects respond earlier and more reliably than when no warning is given. The interval between the warning and trigger signals is referred to as a foreperiod. Response time can also be influenced by repetition of stimuli or responses or by success in previous trials. To explain this variation, one can hypothesize a process that transpires after an instruction or warning signal and is influenced by events in preceding trials to influence the readiness to initiate a movement. Such a covert process may be called response preparation.

The influence of foreperiod on response time varies according to the statistical properties of the foreperiod. If the foreperiod is a fixed value of the appropriate interval, then the animal can learn to anticipate when it will elapse. If the foreperiod is too short (e.g., <100 ms), response time is unaffected because some time is needed for response preparation to engage. If the foreperiod is too long (e.g., >2000 ms), response time is unaffected because it is not possible to sustain response preparation indefinitely. Thus, foreperiods ranging from 100 to 1000 ms afford the largest reduction of response time. The reduction of response time occurs because the passage of time allows a sense of expectation, permitting preparation before the trigger signal so that sometimes the movement might occur before the trigger signal (jump the gun). In fact, such self-generated responses are clear evidence of a covert response preparation process.

Foreperiods can be randomized to reduce predictability. For example, the foreperiod on a given trial could be sampled with equal probability from the range 100 to 1000 ms. This does introduce some variability in response time, but response time tends to decrease as foreperiod increases, and anticipatory responses are often generated especially at the longest foreperiods. This occurs because although the foreperiods are selected randomly across trials, they have a maximum value (i.e., a deadline that the animal can learn through experience). The more time that elapses from the warning stimulus, the higher the probability that the trigger stimulus will occur given that it has not yet. This conditional probability is referred to as the hazard rate of the foreperiod. If the hazard rate increases, the foreperiod is referred to as aging. A nonaging distribution of foreperiods has a flat hazard function; in other words, the passage of time conveys no predictability about when the foreperiod will elapse. Nonaging foreperiod distributions have many short values and exponentially decreasing probability of progressively longer values. Response times obtained with nonaging foreperiods decrease little, if at all, as a function of foreperiod and anticipatory responses are produced rarely.

Further evidence for response preparation is the fact that partially prepared responses can be withheld if an imperative 'stop' signal occurs. This ability can be explained by hypothesizing another covert process that prevents movements. The stop signal or countermanding task has been introduced in the past decade into animal studies. The countermanding paradigm probes a subject's ability to control the initiation of movements by infrequently presenting an imperative stop signal in a response time task. The subjects' task is to cancel the planned movement if a stop signal is presented. Performance on this task can be accounted for by a race between a process that generates the movement (go process) and a process that cancels the movement (stop process). This race model provides an estimate of the stop signal reaction time, which is the time needed to cancel the planned movement. The stop signal reaction time can be used in tests of neural processes and manipulations.

Intention

The disposition to perform some act is a central feature of an intention, but intention cannot be identified entirely with response preparation. A statement of intention must also answer the question, 'Why was that done?' An answer about the causal path through neurons to muscles is not as satisfying as an explanation addressing the reasons for the action based on preferences, goals, and beliefs. In other words, to judge whether a movement was intended, one must refer to the agent's beliefs about which action must be performed in what circumstances to bring about the desired object of the intention.

These concepts about intention have been formulated in the domain of human interactions, but research describing communication and deception, for example, indicates that the attribution of intention to monkeys at least seems justified. A particular form of intention involves reasoning about the relation between observations of events in the environment and actions in that environment – in other words, about how actions result in consequences. In fact, some have concluded that guiding actions based on reasoning about the causes of events is uniquely

human. However, research has shown that rats can make correct inferences guiding responses for reinforcement on the basis of purely observational learning of causal relations between events. Comparable results have been obtained from nonhuman primates.

Methods to Study Emotion

Other articles should be consulted for details on emotion, and this article does not debate whether animals have emotions like humans. Most studies of animal emotion involve fear, anxiety, and depression. Being aversive, the procedures used in these studies deserve special scrutiny and oversight to ensure animal welfare in the pursuit of scientific information.

Fear in animals has been studied most effectively using classically conditioned fear responses. Fear conditioning is a form of learning in which fear is associated with a particular neutral stimulus (e.g., a tone) by pairing the neutral stimulus with an aversive stimulus (e.g., an electric shock, loud noise, or unpleasant odor). Ultimately, the neutral stimulus elicits the state of fear. Fear conditioning has been studied in numerous species, from snails to humans. In animals, conditioned fear can be measured as a period of attentive immobility (freezing), a potentiated startle reflex, as well as changes in heart rate, respiration rate, or muscle tension.

Learned helplessness occurs when an animal is exposed to inescapable aversive stimulation (e.g., electrical shock) and learns that it cannot control its environment. This is regarded as a useful model of human depression.

Anxiety has been investigated by putting mice in an open area and measuring how much time they explore in the exposed interior of the area; more anxious mice spend more time close to the walls.

Methods to Study Executive Control

When the environment is ambiguous or presents competing demands, or the mapping of stimulus onto response is complex or contrary to habit making performance prone to errors, then executive control is required to perform the task. Such executive control over the perception, selection, and response systems is seen as a central component of human cognition, and it can be investigated in animals, particularly nonhuman primates, in a variety of ways. The basic approach is to study how performance changes according to the history of trials. For example, if a particular movement is called for in response to a given stimulus on repeated trials, and then in a subsequent trial a different movement is called for in response to the same stimulus, then a cost will be observed in the accuracy and latency of that response. Such experimental paradigms are referred to as task switching. The cost for switching tasks can arise at least in part because of the conflict between producing the rehearsed response and producing the new response.

Response conflict can be investigated by training macaque monkeys extensively to respond to objects with multiple features such as color and shape with alternative responses cued by different features (e.g., red or circle means go and green or square means nogo). Once the mapping is mastered, testing proceeds by presenting congruent (e.g., red circle) stimuli that cue the same response and incongruent stimuli (e.g., red square) that cue the opposite responses. Evidence of conflict evoked by the incongruent features is slowed response times and increased error rates. Response conflict can be elicited in rats by training them to associate a response lever with a particular reinforcement (e.g., food pellet vs. sugar water). Conflict is induced by rewarding the correct responses with the reinforcement opposite the discriminative stimulus. In other words, by correctly pressing the bar associated with the food pellet stimulus, rats receive sugar water reinforcement and vice versa.

When performing in conditions that require executive control, errors are often produced. A cognitive process of error monitoring is believed to be a component of executive control. In human studies of choice responses, a sign of error monitoring is slowing of response time on the trial after errors.

Decision Making

This process has many aspects, most of which have been summarized. If we define a choice as an action performed in the context of alternatives to achieve a goal, we can appreciate that decision making occurs when the alternatives are vague or the payoffs are unclear. Accordingly, one approach to studying decision making requires animals to perform particularly challenging sensory discriminations; this can be done in any sensory modality. Another approach is to manipulate reinforcement amount or probability in a more or less predictable manner. A classic finding is that animals will produce alternative responses (e.g., bar presses) in direct proportion to the amount or probability of reinforcement for the alternative responses; this is referred to as matching law.

A popular area of inquiry in the past few years has applied concepts and procedures formulated in game theory and economics to the study of animal behavior. For example, monkeys are trained to perform tasks resembling games such as rock–paper–scissors or matching pennies. These mixed strategy games

allow investigators to explore whether theories developed to explain human decision making in games and economic settings apply to animal behavior. All of these approaches result in behavior that is more or less stochastic but amenable to analysis in terms of quantitative models of the processes hypothesized to produce the responses. This approach has shown that animals express in at least a rudimentary manner many of the biases that characterize human decision making. For example, monkeys show aversion to inequity by refusing a lower valued reinforcement if they observe another individual obtaining a higher valued reinforcement.

Problem Solving and Tool Use

Animal cognition is perhaps nowhere more evident than in the solution of problems, especially through the use of tools as highlighted by Kohler's pioneering studies of the use of tools by apes. Although several species learn to use tools in the wild as well as in the laboratory, it remains unclear to what extent animals understand the physical processes by which these tools are effective. In fact, some of the most interesting evidence to date that animals understand the causal properties of physical objects comes from birds, particularly the corvids (crows and ravens). For example, a New Caledonian crow spontaneously bent a piece of ineffective straight wire into an effective hook tool for retrieving food, and another species of corvid learns rapidly to avoid maneuvering food into a clearly visible and inaccessible trap. Chimpanzees do not fashion tools or avoid such traps without considerable training, but they do keep objects used as tools for future use.

Communication

None doubt that animals communicate in rich and diverse manners. For example, vervet monkeys in the wild exhibit a very advanced system of communication, but it is believed to be instinctive, not learned. What is more contentious is whether animals such as great apes can be said to learn language. Investigators have trained chimpanzees, bonobos, gorillas, and orangutans to communicate through symbols on a keyboard or sign language. Individual chimpanzees have acquired vocabularies of at least 200 words or signs and have demonstrated the capacity to respond appropriately and produce meaningfully different sentences.

Learning to produce words is not limited to primates. African gray parrots have gained notoriety through a phenomenal ability to learn words or even phrases, in some cases as complicated as songs. One individual bird has been trained to use English speech to label dozens of objects, features, and categories as well as use commands and requests.

See also: Anxiety Disorders; Binocular Rivalry; Cognitive Control and Development; Episodic Memory: Assessment in Animals; Game Theory and the Economics of Animal Communication; Games in Monkeys: Neurophysiology and Motor Decision-Making; Operant Conditioning of Reflexes; Pharmacology of Fear Extinction; Procedural Learning: Classical Conditioning; Reasoning and Problem Solving: Models; Referentiality and Concepts in Animal Cognition; Reward Decision-Making; Short Term and Working Memory; Spatial Memory: Assessment in Animals.

Further Reading

Blake R (1999) The behavioural analysis of animal vision. In: Carpenter RHS and Robson JG (eds.) *Vision Research: A Practical Guide to Laboratory Methods*, pp. 137–160. Oxford: Oxford University Press.

Crawley JN (2000) *What's Wrong with My Mouse? Behavioral Phenotyping of Transgenic and Knockout Mice*. New York: Wiley-Liss.

Emery NJ and Clayton NS (2004) The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science* 306: 1903–1907.

Griffin D (2001) *Animal Minds*. Chicago: University of Chicago Press.

Hanes DP and Schall JD (1995) Countermanding saccades in macaque. *Visual Neuroscience* 12: 929–937.

Karl T, Pabst R, and von Horsten S (2003) Behavioral phenotyping of mice in pharmacological and toxicological research. *Experimental Toxicology and Pathology* 55: 69–83.

Logothetis NK and Schall JD (1990) Binocular motion rivalry in macaque monkeys: Eye dominance and tracking eye movements. *Vision Research* 30: 1409–1419.

National Research Council (1996) *Guide for the Care and Use of Laboratory Animals, Institute of Laboratory Animal Resources*. Washington, DC: National Academy Press.

Panksepp J (2004) *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York: Oxford University Press.

Pearce JM (1997) *Animal Learning and Cognition: An Introduction*. Hove, UK: Psychology Press.

Terry WS (2005) *Learning and Memory: Basic Principles, Processes, and Procedures*. New York: Allyn & Bacon.

Tomasello M and Call J (1997) *Primate Cognition*. Oxford: Oxford University Press.

Wahlsten D, Rustay NR, Metten P, and Crabbe JC (2003) In search of a better mouse test. *Trends in Neuroscience* 26: 132–136.

Whishaw IQ and Kolb B (2004) *The Behavior of the Laboratory Rat: A Handbook with Tests*. Oxford: Oxford University Press.

Executive Function and Higher-Order Cognition: Computational Models

J W Brown, Indiana University, Bloomington, IN, USA
T S Braver, Washington University in St. Louis, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Executive functions include a range of processes that impact attention, working memory, goal setting, action selection and sequencing, and problem solving. Theories of higher cognitive function aim to account for how complex cognitive processes are carried out in the brain and may offer conceptual-level mechanistic explanations. Yet all too often, the mechanisms proposed to explain higher functions suffer from an implicit dependence on unspecified higher level control processes so that the actual neural mechanisms subserving the cognitive phenomena remain mysterious. This problem is called the homunculus problem. The term homunculus means 'little man' and refers to the unspecified higher level controller that is implicitly required by some theories of cognition. There has been much discussion about how to banish the homunculus – that is, to provide an explicit and mechanistic description of how control functions emerge and interact so that unspecified high-level controllers will not need to be invoked. Computational models of executive function provide critical tools in this effort. When conceptual theories are fleshed out in working computational models, accounts that claim to overcome homunculus problems are more compelling for several reasons. First, a viable computational model must work; it must actually perform a task and account for data. Second, all components of the model are explicit, which makes it easy to identify any control signal components of the model that are less well specified. Third, when a computational model does exhibit properties and behavior associated with executive function, it is easier to trace, specify, and quantify the origin and nature of the computational mechanisms that produce such cognitive properties and behavioral patterns.

Computational models of executive function include a range of formalisms from production systems to neural networks. Whereas production system models are typically more abstract, and involve symbolic processing, neural network models are more closely tied to neurobiological principles and constraints. Both types of models have focused on understanding the functions subserved by specific brain regions thought to be critical for executive control, such as the prefrontal cortex (PFC), anterior cingulate cortex (ACC), and basal ganglia (BG). Work since the late 1990s has developed a distinction between performance monitoring and control processes related to executive function, and this distinction provides a useful framework for delineating components of executive function. This article discusses computational models of distinct brain centers in executive function, and it concludes with a discussion of more abstract production system models of executive function.

Prefrontal Cortex and Control

Active Goal Maintenance

The lateral PFC is important for actively maintaining information via sustained neural firing that is robust in the face of potentially distracting information. The standard neural model of PFC has cells that reciprocally excite each other so that activity is self-sustaining: active neurons send excitation to other neurons that then send excitation back, creating a self-sustaining activity pattern that constitutes a stable attractor state. Once a particular population of PFC neurons is activated in this attractor state, it suppresses other populations via lateral inhibition. The state is maintained even when the inputs that activated the PFC are no longer present (i.e., working memory), and it is robust in the sense that activity persists despite other distracting inputs. *N*-methyl-D-aspartate (NMDA) channels may provide a cellular mechanism by which PFC cells that are more active become more easily excitable. As cells become active and depolarize, NMDA channels lose their Mg^{2+} block, thus increasing their conductivities at higher levels of depolarization. In a population of cells that reciprocally excite each other, this has the effect of enhancing self-perpetuating activity.

Neural network models of PFC are generally implemented as either rate-coded or spiking models. Rate-coded models quantify the cell activity as the instantaneous average firing rate of a local population of cells. Mathematically, a population of such rate-coded cells can exhibit bistable activity if the equilibrium output activity signal increases faster than linearly within a certain range. For example, if the input signal to a cell is doubled, the output signal from the cell would need to be greater than double at equilibrium in order for reciprocal excitatory activity to be self-sustaining. Models of PFC have been developed to account for a range of empirical data, including healthy

human and monkey performance of tasks that involve working memory, as well as executive control deficits in clinical populations such as individuals with schizophrenia and individuals with focal lesions of PFC.

Spiking models of PFC attempt to simulate the actual spikes generated by individual PFC cells. Typically, these models use a simplified integrate-and-fire scheme (which simulates action potentials as a point process) rather than a more sophisticated and computationally expensive Hodgkin–Huxley formalism (which simulates actual ionic channel kinetics and thus the full time course of the action potential). Spiking models have been developed to account for the data from single-unit recordings in awake, behaving monkeys performing a variety of tasks, including delayed match to sample, sequence learning, and various types of rule-based decision making. Other more biophysically detailed models have accounted for complex interactions among specific neurotransmitters and receptor types in mediating sustained activity, especially dopamine, NMDA, α-amino-3-hydroxy-3-methyl-4-isoxazole propionic acid (AMPA), and γ-aminobutyric acid (GABA) receptors. These models make specific predictions regarding the effects of different kinds of pharmacological manipulation on executive control.

An important computational problem involves the mechanisms that enable PFC networks to maintain information in the face of interfering input while at the same time appropriately updating stored information when relevant signals arrive. Various solutions have been proposed to solve this problem, and all of them involve an input gating mechanism. Opening the gate essentially involves making the attractor states become less stable by amplifying new inputs. The result of the open gate is that PFC is rapidly updated with the new input. In contrast, when the gate is closed, input signals do not disrupt maintained states. Some models of PFC gating depend on the neuromodulator dopamine. Cells in the ventral tegmental area (VTA) provide dopaminergic signals to the PFC. According to these models, transient increases in the level of VTA firing release more dopamine in PFC, which switches the PFC between fast updating and stable maintenance. The interactions among PFC cells and various D_1 and D_2 dopamine receptor subtypes are complex, and computational modeling has clarified how these elements work together functionally. In one model, dopamine D_1 receptor activation effectively stabilizes PFC states, whereas D_2 receptor activation leads to destabilizing effects. Others have argued that PFC stability follows an inverted U pattern, such that stability is greatest for a baseline level of dopamine but falls off rapidly for either too little or too much dopamine.

The dopamine-based gating models suffer from a limitation. The gating signal is global and nonspecific, so with these models, it is not possible to update certain areas of PFC while maintaining others. To address this issue, another class of gating mechanism has been proposed based on the strong reciprocal connections between the PFC and the BG complex. Neurons in the direct pathway of the BG drive a net disinhibition of the thalamus, which may in turn allow fast updating of selected parts of PFC. Conversely, neurons in the indirect pathway of the BG drive a net inhibition of the thalamus, which may effectively close the PFC gate and maintain stable representations against distractors in selected parts of PFC. Both the dopaminergic and BG gating mechanisms have included learning mechanisms, which allow the system to discover the appropriate signals and timing for opening the gate.

Biased Competition

Most computational models of executive control cast the PFC as serving a biasing function that regulates activity in other brain areas involved in domain-specific processing and behavior. The Stroop is the best known experimental task for illustrating the need for control biases. In the task, subjects view a color word (e.g., the words 'red' or 'blue') printed with an ink color that may be the same as (congruent) or different from (incongruent) the word (**Figure 1**). Subjects are asked to respond according to the ink color (the color-naming task) but to avoid responding based on the semantic meaning of the word (the word-reading task). For example, the word red presented in green ink (**Figure 1**) should elicit the correct response green, not red. This is a difficult task because for literate individuals, the automatic tendency to read a word is stronger than the process of indicating which color ink is used to write the word.

According to conceptual framework known as biased competition, activity in PFC representing the color naming task (the task demand units in **Figure 1**) strongly activates other areas in the brain responsible for generating specific responses based on the ink color. The PFC control signal does not directly determine the response but only enhances activity in other areas that determine the response according to the current rule. In the model of **Figure 1**, there are two task pathways, one for word reading and one for ink color naming. Each pathway can work in isolation from the other to produce task-appropriate processing and responses without executive control signals. However, control is critical when both the word-reading and color-naming pathways are simultaneously engaged. Under such conditions, there is competition during processing between the two dimensions of semantic meaning and ink color. In **Figure 1**, the thicker arrows from the word-reading units to the response

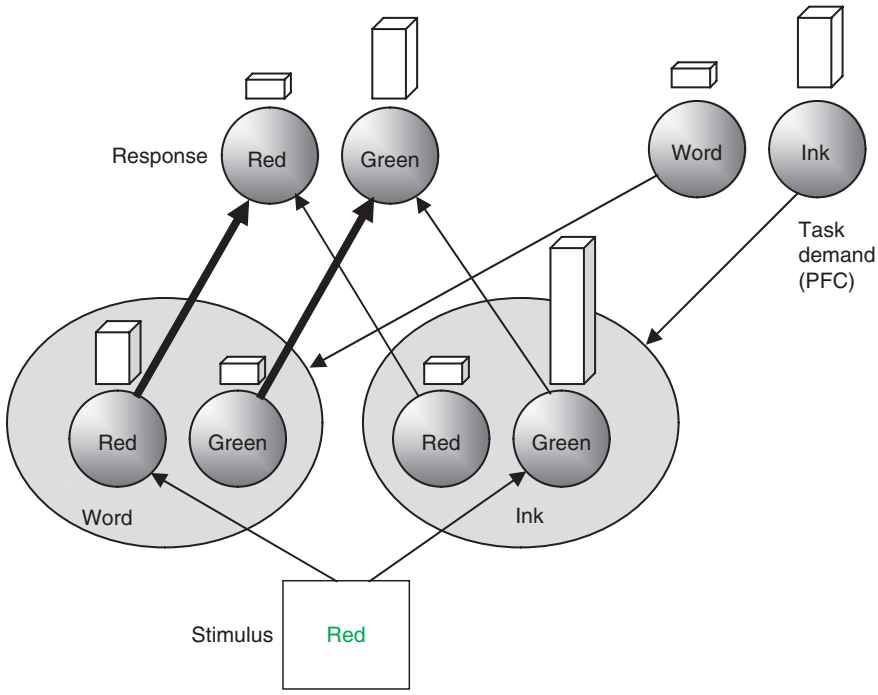

Figure 1 Executive control enables correct performance of the Stroop task.

units represent the stronger connection weights due to the effects of more learning and practice of reading words. The word-reading pathway will dominate processing competition in the absence of executive control. Thus, to be successful during execution of the color-naming task, the color pathway needs additional support to counteract the stronger weights on the word pathway. This support comes in the form of top-down biasing control, carried by a task demand unit (reflecting PFC), that represents the color-naming tasks. This task demand unit sends a signal that provides an additive source of activation to the entire input color dimension. When this control signal is present, the color pathway can successfully compete with the otherwise stronger word pathway to control responding. Conversely, without the strong biasing signal from PFC, the incorrect default response of word reading would prevail. Models of PFC that implement such control biasing signals have been found to account for human behavioral performance in a range of cognitive tasks that have Stroop-like properties of competition and interference from automatic processing tendencies.

Anterior Cingulate Cortex and Performance Monitoring

A key question in understanding goal-driven or top-down executive control is how and when it is engaged. Executive control mechanisms direct behavior, but what controls the executive? This question of what controls the controller relates to the homunculus problem discussed previously. In the case of executive control, there must be a mechanism that indicates when top-down attentional control is needed. The kinds of mechanisms that determine when control must be implemented do so by monitoring performance of the behaving cognitive system to detect the need for control.

Conflict Theory

One hypothesis that has received widespread interest is the so-called conflict monitoring account. For the purposes of executive control, conflict is defined as the coactivation of mutually incompatible response processes. This hypothesis takes as its starting point the theoretical idea that control is primarily required as a means of managing the cross-talk interference that can arise between concurrent and overlapping processing pathways in the brain. A measure of conflict provides a quantitative index of the degree of cross-talk between response pathways, and therefore the degree of executive control that must be exerted to resolve the interference. The critical addition of the conflict monitoring hypothesis is the adoption of an explicit mechanism which detects the presence and level of cross-talk interference and which conveys this information as an output signal that is used to regulate the engagement of top-down control signals. In some formulations, the conflict signal is computed

as the Hopfield energy of the response system. Essentially, this reduces to the multiplicative product of activity in two mutually incompatible response pathways. In the case of the Stroop task stimulus word red written in green ink (Figure 1), the conflict signal would be computed as the product of activity in the red and green response units. Various models have used other formalisms instead of the multiplicative model, but all compute a measure of coactivation such that two incompatible response pathways must be simultaneously coactive for a conflict signal to be generated.

Neuroanatomy

Based on an accumulating literature – primarily arising from human neuroimaging studies examining conflict-based interference effects – the area surrounding and including the ACC was identified as the neural locus of a conflict monitoring mechanism. The conflict monitoring hypothesis states explicitly that ACC implements the calculation of an online index of the conflict between competing representations, particularly at the response level. In computational models of conflict monitoring, fluctuations in ACC activity index ongoing fluctuations in conflict, and this information is then conveyed to the PFC, which in turn implements control strategies and maintains representations of the ongoing behavioral context. According to this hypothesis, more cognitive control is needed when conflict is high, but less is needed when conflict is low. The system is postulated to function as a closed loop, whereby the detection of conflict (in ACC activity) serves to engage control (in the PFC), which then acts to reduce the conflict experienced in similar future situations. Thus, the ACC has been identified as a mechanism that determines when executive control is needed and in turn activates the PFC to implement executive control. Models of the role of ACC in executive control have accounted for a variety of data from functional neuroimaging studies of cognitively demanding tasks in healthy individuals and clinical populations.

Conflict and Errors

In addition to the monitoring of response conflict, ACC also appears to be critical for situations of error commission. This brain region transiently activates both when conflict is high and when errors are actually committed. Computational modeling work reflects an ongoing controversy over the ultimate nature of performance monitoring functions in ACC. Much of the empirical data have come from studies using the event-related potential technique and focused on a specific scalp activity pattern known as the error-related negativity (ERN), which occurs at the time of error commission or at the time of explicit error feedback if the error was not predictable. There have been alternative computational models developed which postulate a specific and distinct role for the ACC in error detection rather than conflict monitoring, and these models make somewhat different claims about the function of ACC outputs in regulating behavioral performance. One class of models assumes that dopaminergic error signals drive or train the ACC to respond to errors. These models have accounted for the timing of error signals in ACC, such as whether the ERN occurs before or after receipt of feedback confirming that an error has been committed. The models have also addressed the greater magnitude of ERN signals in contexts in which errors are less frequent.

Work has also suggested that many of the phenomena originally thought to be uniquely captured by the error-detection models can also be successfully explained by the conflict monitoring account. Nevertheless, there has been a growing appreciation of the idea that the ACC may play a more general role in performance prediction and action outcome monitoring than just detecting the presence of errors or conflict. One computational model proposes that since errors are more likely when conflict is present, the ACC may generally compute the likelihood of an error based on previous learning. Despite some differences in implementation and focus, computational models of ACC primarily serve as a mechanism for evaluating the level of executive control needed in a given task situation.

Time Course of Executive Control

According to the conflict model, ACC monitors performance and then drives PFC to implement executive control. Early models suggested that control is implemented quickly, but there may be distinct timescales in which monitoring and control signals are effective. In particular, some kinds of control mechanisms may be implemented rapidly enough to detect and resolve conflict within a single trial, whereas others may use conflict detected in one trial to implement control in subsequent trials. A model of the Stroop task postulates a dual form of conflict detection: one acting in the short (i.e., subsecond) timescale and another in the long (i.e., several second/multiple trial) timescale. These different conflict measures in turn activate two separate forms of cognitive control, respectively called reactive and proactive control. Thus, a temporally extended period of high conflict will push the system toward a state of preemptive attentional control in which goal-relevant context is actively maintained across

trials to bias target-relevant aspects of attention, before stimuli are actually presented. In contrast, in situations of low conflict, this type of proactive control is not engaged. Instead, the actual detection of conflict within a trial (i.e., poststimulus presentation) rapidly engages a transient goal reactivation mechanism which can do rapid biasing of the target-relevant dimension. This computational model was able to successfully account for behavioral and brain imaging data from the Stroop task under conditions involving either a high percentage of incongruent trials or a high percentage of congruent trials. It should be noted that individuals may not be explicitly aware of changes in the frequency of conflict occurrence, and therefore the adjustments in attentional control may occur even in the absence of conscious intervention.

Multiple Kinds of Control

In the event that the need for greater control is detected, the PFC must determine what kind of control to implement. Different kinds of executive control signals have been dissected using task-switching paradigms. Task switching involves interleaved performance of two different cognitively demanding tasks. For example, the previously mentioned Stroop task might be modified so that subjects must perform the ink color-naming and the word-reading tasks on alternating trials. The most basic finding of task-switching studies is the switch cost, meaning that task performance is slower and more error-prone when a new task is performed than when an old task is repeated. The origin of the switch cost is controversial. In some models, the switch cost is due to the time needed for an executive controller to reconfigure the system to perform the new task. In other models, the switch cost is simply due to the lack of priming in the new task pathways, and therefore no executive control need be invoked to explain the switch cost.

Computational models of task switching provide insight into executive function by dissociating attentional and motor control processes. Returning to the Stroop task example, stimuli may be either congruent (e.g., the word red written in green ink) or incongruent (e.g., the word red written in red ink). Incongruent stimuli require more focused attention on the color of the ink while excluding the semantic meaning of the word from further cognitive processing. In contrast, when a task switch occurs, executive control may be required to implement the new task. Alternatively, as suggested by a more recent computational model, executive processes may not implement the task switch directly but may slow responding to enable the system to adequately process the new task prior to generating a response. Regardless of the particular form of motor control implemented, a distinction can be seen between attentional control driven by incongruency and motor control associated with the task switch.

Basal Ganglia and Action Selection

When multiple completing behavioral plans are formed and activated, an action selection mechanism allows one to be executed and suppresses or defers the remaining incompatible plans. This constitutes an executive control function. Some models posit a role of ACC in action selection, but a number of models of executive function have suggested that the BG play the main role in action selection. Of note, action selection may include control of external actions as well as internal manipulations such as deciding when and how to update the PFC state.

The BG consist of several subcortical nuclei. The striatum (caudate nucleus and putamen) receive projections from a number of frontal cortical structures. The striatum provides two main output projections. The first is known as the direct pathway, by which the striatum inhibits the internal pallidum, which in turn inhibits the motor thalamus. Computational models have suggested that the net effect of direct pathway activity is to facilitate action execution because striatal inhibition of the internal pallidum results in a net disinhibition of the thalamus, which may allow actions to be executed. In contrast, activity in the second output projection from the striatum (known as the indirect pathway) results in a net increased inhibition of the motor thalamus, which may prevent other actions from being executed. In this way, computational models suggest that the BG can actively allow some motor plans to be executed while simultaneously suppressing others. A related BG structure known as the subthalamic nucleus may further prevent multiple competing plans from being simultaneously executed. Thus, according to several computational models, these components of the BG apparently monitor ongoing activity in PFC in order to control which plans are executed.

In terms of performance monitoring, a number of BG models have focused on the substantia nigra pars compacta (SNC), which includes a dense population of dopaminergic cells that project widely to the other parts of the BG, especially the striatum. These SNC cells respond transiently to unexpected rewards and unexpected cues that predict reward, and they transiently pause their firing when an expected reward is not received. As posited by several computational models, these performance-monitoring signals train other parts of the BG to select the actions that are most likely to lead to reward.

Models of executive control by the BG have addressed a wide range of empirical data. Several

models of action selection in the BG have focused on neurophysiology. These have simulated and accounted for a variety of cell types found in the BG and connected areas of awake, behaving macaque monkeys performing cognitive tasks. Cognitive-level models generally employ a connectionist framework and have accounted for behavioral data in healthy and clinical populations, such as those with schizophrenia, Parkinson's disease, and Huntington's disease. These models have accounted for specific kinds of errors, both random and perseverative, on tasks thought to involve executive control, such as the Wisconsin Card Sort Task. These models have been applied to the domain of decision making and have been extended to include other structures such as the orbitofrontal cortex, which may serve to actively maintain (in working memory) reward prediction errors over an extended delay.

Production System Models

In contrast with neural network models, production system computational models do not explicitly simulate neural activity. Instead, they are often referred to as cognitive architectures. They rely on abstract formalisms from the field of machine intelligence, using a set of if–then conditional rules. A cognitive processor regularly evaluates the environment to determine whether a particular rule is applicable. When the conditions required for a particular rule are satisfied, then the operation specified by that rule is executed. In this way, external stimuli and internal goals lead to actions. Notably, multiple rules can be active simultaneously. Some models include specific assumptions about the time needed for a conditional rule to be activated and execute an action, and this allows production systems to make specific predictions about response time for comparison with human data, such as the psychological refractory period. Such models can also make predictions about reaction time in high-level tasks involving nested goals, such as planning tasks like the Tower of London, and the comprehension and production of complex sentences. These tasks have been traditionally difficult for neural network models to simulate. Executive control in a production system is needed to mediate among multiple rules and tasks that would otherwise interfere with each other when executed simultaneously. Such control may include lockout scheduling or interleaved scheduling. In lockout scheduling, the executive processes detect when two or more rules compete for simultaneous access to the same resource and allow only one to have access at a given time. Interleaved scheduling involves switching back and forth between the two processes. When executive control is explicitly specified, it is generally implemented as a set of rules that manipulate the internal state of the production system, such as the current goal state and working memory contents.

See also: Cognition: Basal Ganglia Role; Consciousness: Theoretical and Computational Neuroscience; Executive Function and Higher-Order Cognition: Assessment in Animals; Executive Function and Higher-Order Cognition: Neuroimaging; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex.

Further Reading

Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, and Qin Y (2004) An integrated theory of the mind. *Psychological Review* 111(4): 1036–1060.

Botvinick MM, Braver TS, Barch DM, Carter CS, and Cohen JC (2001) Conflict monitoring and cognitive control. *Psychological Review* 108: 624–652.

Braver TS and Cohen JD (2000) On the control of control: The role of dopamine in regulating prefrontal function and working memory. In: Monsell S and Driver J (eds.) *Attention and Performance XVIII*, pp. 713–738. Cambridge, MA: MIT Press.

Brown JW, Bullock D, and Grossberg S (2004) How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks* 17(4): 471–510.

Brunel N and Wang XJ (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of Computational Neuroscience* 11(1): 63–85.

Cohen JD, Dunbar K, and McClelland JL (1990) On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review* 97(3): 332–361.

Durstewitz D, Seamans JK, and Sejnowski TJ (2000) Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of Neurophysiology* 83(3): 1733–1750.

Grossberg S (1982) *Studies of Mind and Brain.* Boston: Reidel.

Holroyd CB and Coles MG (2002) The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review* 109(4): 679–709.

Meyer DE and Kieras DE (1997) A computational theory of executive cognitive processes and multiple-task performance. Part 1. Basic mechanisms. *Psychological Review* 104(1): 3–65.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 21: 167–202.

Norman DA and Shallice T (1986) Attention to action: Willed and automatic control of behavior. In: Davidson RJ, Schwartz GE, and Shapiro D (eds.) *Consciousness and Self-Regulation,* vol. 4, pp. 1–18. New York: Plenum.

O'Reilly RC and Frank MJ (2006) Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation* 18(2): 283–328.

Rougier NP, Noelle DC, Braver TS, Cohen JD, and O'Reilly RC (2005) Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America* 102(20): 7338–7343.

Yeung N, Cohen JD, and Botvinick MM (2004) The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review* 111(4): 931–959.

Executive Function and Higher-Order Cognition: Definition and Neural Substrates

E K Miller, Massachusetts Institute of Technology, Cambridge, MA, USA
J D Wallis, University of California at Berkeley, Berkeley, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Cognitive, or executive, control refers to the ability to coordinate thought and action and direct it toward obtaining goals. It is needed to overcome local considerations, plan and orchestrate complex sequences of behavior, and prioritize goals and subgoals. Simply stated, you do not need executive control to grab a beer, but you will need it to finish college.

Executive control contrasts with automatic forms of brain processing. Many of our behaviors are direct reactions to our immediate environment that do not tax executive control. If someone throws a baseball toward our face, we reflexively duck out of the way. We have not necessarily willed this behavior; it seems as if our body reacts and then our mind 'catches up' and realizes what has happened. Evolution has wired many of these reflexive, automatic processes into our nervous systems. However, others can be acquired through practice because learning mechanisms gradually and thoroughly stamp in highly familiar behaviors.

For example, consider a daily walk to work. If the route is highly familiar and if traffic is light, our mind can wander. Before we know it, we may have gone a considerable distance and negotiated street crossings and turns with little awareness of having done so. In these cases, the control of our behavior occurs in a 'bottom-up' fashion: it is determined largely by the nature of the sensory stimuli and their strong associations with certain behavioral responses. In neural terms, they are dependent on the correct sensory conditions triggering activity in well-established neural pathways.

Suppose, however, that during our walk to work something unexpected happens or we encounter difficulty. For example, we might encounter a busy street that we need to cross. Then, the executive system takes over and we need to 'take charge' of our actions. We pay attention to the people and cars around us to anticipate and accommodate their actions, or we may decide to take an alternate route. Now, straightforward stimulus–response associations are insufficient to govern our behavior. We must use knowledge of our current objective (arriving at work on time and intact) and results from previous experiences to weigh the alternatives and consequences. During this 'controlled mode,' we also engage the basic sensory, memory, and motor processes that mediate automatic behavior. Only now, the environment is not simply triggering these processes. Instead, we use our current goals to shape and control these processes in a 'top-down' fashion.

Figure 1 summarizes these ideas by illustrating a widely accepted view of the architecture of cognition. At the lower level are the automatic processes, which include sensory analysis, memories, details of motor acts, and well-learned skills. The system is in automatic mode when processing flows through the lower level, from input to output, along established pathways without any hindrance or modification. However, at any given instant the executive system can step in and modify this flow should it detect that the automatic processes are no longer sufficient to obtain our goals. In understanding the neuronal mechanisms that underlie this process, the prefrontal cortex (PFC), the brain area directly behind our forehead, seems especially critical.

The Prefrontal Cortex

PFC has dramatically expanded in size and complexity across evolution, and its development correlates with the complexity of the behavioral repertoire exhibited by an organism. It has reached its pinnacle in humans, in which it accounts for approximately 30% of the total cortical area. It consists of a collection of cortical areas that differ from one another in terms of the size, density, and distribution of their neurons. Figure 2 shows the major PFC divisions of the monkey PFC, although anatomists have subdivided these further, describing at least 18 distinct areas. The subdivisions have partly unique, but overlapping, patterns of connections with the rest of the brain, which suggests some regional differences in function. As in much of the neocortex, however, there are local connections between different PFC areas that can result in an intermixing and synthesis of the disparate information needed for cognitive control.

PFC is anatomically in a good position for a central role in executive control. Collectively, the various PFC areas have interconnections with brain areas processing external information, including all sensory systems and cortical and subcortical motor system structures, as well as internal information from limbic and midbrain structures involved in affect, memory, and reward. Indeed, neuronal activity in PFC reflects the multimodal nature of its inputs. PFC neurons

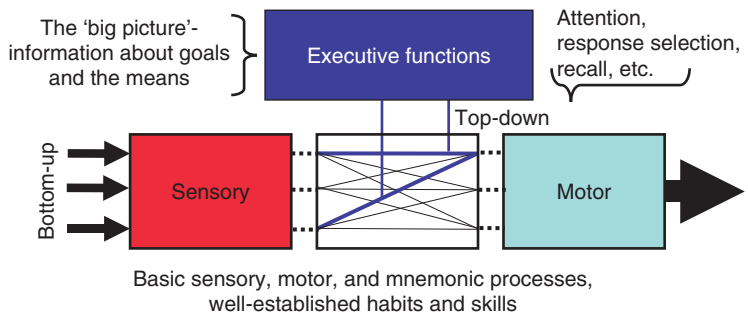

Figure 1 Two levels of cognitive processes. Specialized functions that acquire information about goals and means (top) select and coordinate among innate and well-established routines (bottom). Inputs from the environment are 'bottom-up,' whereas signals based on knowledge about goals and task demands are 'top-down.' Active processing lines are shown in blue.

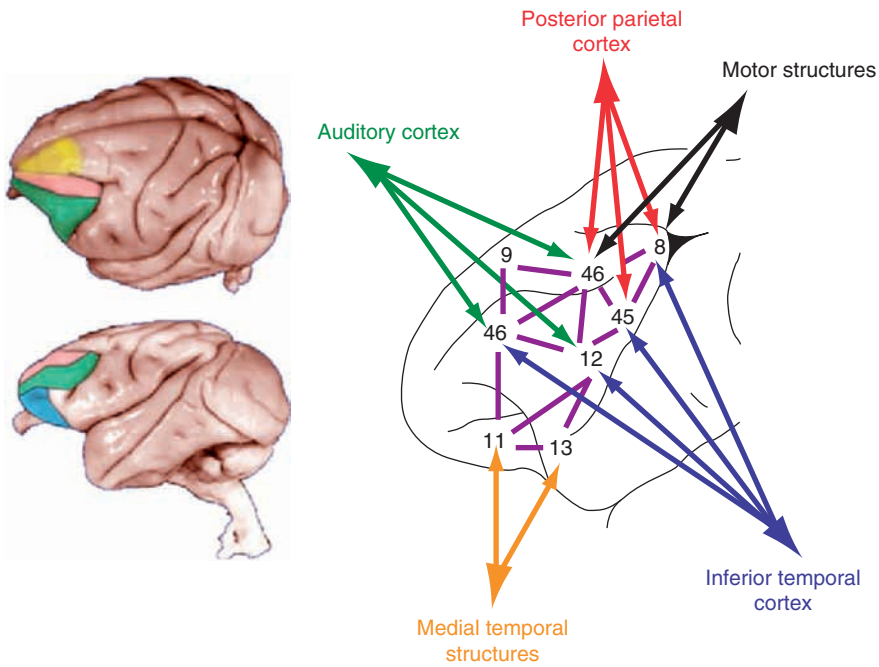

Figure 2 The monkey prefrontal cortical areas with some connections. (Left) Dorsal is yellow, dorsolateral is red, ventrolateral is green, and orbital is blue. The medial PFC is not pictured. (Right) Some extrinsic and intrinsic connections. Not all connections are illustrated; this figure is meant to convey how the PFC can synthesize and integrate diverse inputs from other brain structures.

encode visual, auditory, tactile, olfactory, and gustatory cues, as well as recalled memories and behavioral responses such as voluntary limb and eye movements. In short, the functional and morphological anatomy of PFC is consistent with its role in synthesizing diverse information about the external and internal world in order to produce goal-directed behavior.

PFC Contributions to Executive Control

PFC damage does not result in simple deficits; there is relatively little overt impairment, at least with superficial examination. Patients with PFC damage have intact sensory capabilities, they can remember events and facts, and they appear remarkably normal in casual conversation. However, despite the superficial appearance of normality, PFC damage devastates a person's life. Patients have trouble staying employed, married, or even completing simple daily errands. Furthermore, they often seem to act without any apparent consideration to future consequences. Careful testing of PFC patients has revealed a number of underlying cognitive deficits that might explain their impairments.

Inhibition

PFC damage often produces 'stimulus-bound' behavior. Patients engage in whatever habitual behaviors

happen to be triggered by cues in the immediate environment. For example, when they see a glass of water, they may drink it regardless of whose glass it is. Because other learning and memory systems are intact, they can learn to respond to specific cues in specific ways. However, they then become stuck in that behavioral rut, producing the same response to the cue even when it is no longer appropriate – a phenomenon called 'perseveration.'

Planning

After PFC damage, the basic units of behavior remain intact. What is lacking is the ability to organize them toward a goal. For example, one patient, when making coffee, first stirred and then added milk. Another example is a patient who had been an excellent cook until she had a large tumor removed from her frontal lobes. Her basic skills remained but she seemed unable to organize them. She would move haphazardly from preparing one part of the meal to another so that some parts of the meal would burn, whereas others had hardly been started.

Evaluating Consequences

Goal-directed behavior depends on our capacity to evaluate the consequences of our actions so that we can choose an optimal plan. Evidence that certain types of PFC damage impair this process comes from a gambling task. The subject has to choose cards from four different decks. Some cards win the subject some money, but others cause the subject to lose money. Unbeknownst to the subject, cards from two of the decks will occasionally win large amounts of money but are also occasionally associated with very large losses. Thus, consistently choosing from these decks will lead to a net loss. In contrast, if the subject chooses from the other two decks, he or she will win smaller amounts, but the losses are also smaller, so that overall the subject will obtain a net profit. Control subjects quickly learn to limit their choices to the profitable decks, whereas patients with PFC damage, particularly of the orbital region, continually choose from the decks associated with large rewards and larger losses until they lose all their money.

This led to the hypothesis that the orbital PFC is responsible for labeling cues or situations with an affective significance or 'somatic marker.' It associates memories of past, affect-laden events with a representation of the state of the autonomic nervous system that the event evoked. Similar events in the future can then evoke a 'gut feeling' of the appropriate course of action by recall of the somatic marker. This is particularly helpful in very complex situations in which it is difficult to evaluate rationally and deliberatively all the pros and cons of a choice.

Further evidence that PFC neurons participate in this process comes from studies showing that a large proportion of them encode information about expected rewards. For example, some neurons encode the delivery of a reward, whereas others respond when a reward does not occur as expected. After training, many PFC neurons will show responses to visual cues that signal whether a reward will or will not be forthcoming and reflect the identity, size, and preference of an expected reward.

Working Memory

The importance of a short-term memory buffer for cognition is apparent to anyone who has tried to do math in their head; intermediate answers and operations must be buffered in a way that allows them to be kept 'in mind.' The pattern of deficits following PFC damage might be sequelae of an underlying deficit in working memory. Patients may have difficulty planning or appear disinhibited because without working memory, goal-relevant information is lost over all but the briefest delays and thus their behavior takes its cue from the immediate environment.

This hypothesis stems from observations that PFC damage in monkeys impairs a spatial delayed-response task that requires them to remember the location of a stimulus or behavioral response over a brief delay of several seconds. In addition, there are many studies illustrating that PFC neurons sustain their activity over several seconds to bridge short delays imposed between a cue and a response – observations that Fuster, Niki, and colleagues first made in the 1970s. Goldman-Rakic and colleagues used a more controlled version of the task using an oculomotor response to a visual target presented in the periphery of the visual field. They demonstrated that dorsolateral PFC neurons have precise tuning for memories of particular visual field locations.

In humans, a similar task, called the 'two-back' task, is sensitive to PFC damage or dysfunction. Subjects observe a sequence of sample stimuli and must respond if a current sample stimulus matches one seen two samples ago. Because this involves matching to a specific stimulus, it may rely on working memory rather than other, more passive, forms of immediate memory, such as those that automatically detect repetition.

PFC activity sustains brief memories of a wide range of behaviorally relevant cues: objects, colors, the frequency of a vibration to the hand, and forthcoming movements. Similar activity is also evident in the sensory and motor systems. This is not

surprising; sustained activity is evident in many brain structures and must play a role in many neural processes from sensory afterimages to holding a motor state. However, specificity and robustness are what separates more 'cognitive' short-term memory processes, such as working memory processes, from such lower level processes. Working memory selectively retains task-relevant information, rather than just any stimulus, and does so over potential distractions. PFC neurons have this ability. For example, when monkeys are required to sustain the memory of a sample object across a delay period filled with visual distractors that require attention and processing, sustained activity within the PFC can still maintain a memory of the sample object. By contrast, sustained activity in visual cortical areas seems more labile; it is disrupted by the presence of distractors.

Learning and Using Rules

The PFC is important when we need top-down processing – that is, when our internal states or intentions guide behavior. It is critical in situations in which the mapping between sensory inputs, thoughts, and actions is weak relative to other existing ones or when the mapping is rapidly changing. In such circumstances, we often rely on previously constructed models of regularities in our environment and our expectancies of future events – the so-called 'rules of the game.'

Conditional learning tasks are a laboratory test of rule learning. They require learning of associative relationships that are arbitrary and extend beyond the simple one-to-one mappings that underlie reflexive reactions. Whether or not a given response is the correct response to a given cue constantly changes because it depends on additional information. For example, reaching for a beer can be rewarding, but only if one considers other information. If the beer belongs to another patron, the result could be disastrous. Monkeys with PFC damage are impaired at a variety of conditional learning tasks. When monkeys learn associations between, for example, three visual cues and three directional movements of a joystick, simply moving the joystick is not enough to produce reward. Rather, the specific movement direction is dependent on the specific cue. One means 'up,' the other 'down,' etc. Following PFC damage, monkeys have difficulty acquiring this task. Furthermore, PFC neurons encode multiple aspects of conditional tasks, including the sensory cues, the motor responses, and the specific associations between the cues and responses (Figure 2).

PFC is also essential in enabling the control of behavior by abstract, high-level rules. Consider, for example, dining in a restaurant. We are not born knowing how to act in this situation. Instead, after several experiences in restaurants we begin to abstract the common features of importance. For example, we learn the important sensory information deserving our attention (e.g., the wine list), typical events, appropriate actions, and expected consequences (e.g., paying the bill). The rules underlying ordering a meal in a restaurant are essentially networks of predictive relationships between immediate cues, internal and external context, and actions and consequences long divorced from the details of the individual restaurants in which we have dined. We can then use these abstract rules to order a meal in any subsequent restaurant, even those in which we have never dined before. Several investigators have argued that the encoding of rule information is a cardinal function of the PFC.

One task that directly tests this ability is the Wisconsin Card Sorting Test (WCST). The patient is required to sort a deck of cards on which there are a number of colored shapes. The patient has to determine by trial and error the correct property by which to sort the cards. For example, if the 'color' rule were in effect, the subject would have to sort the cards according to the color of the shapes. Unbeknownst to the subject, the experimenter can switch which is the correct rule, and the subject needs to detect this and modify his or her behavior accordingly. Patients with PFC damage, particularly lateral PFC, have difficulty doing this. Using more focal, experimentally induced lesions in monkeys, experimenters have confirmed that lateral PFC is the most important PFC region for rule or strategy implementation. Furthermore, studies show that single PFC neurons are capable of encoding abstract rules. For example, monkeys saw two pictures appear successively separated by a delay. At the start of the trial, a cue instructed the monkeys to release a lever if the two pictures were either the same or different. Thus, to solve the task the monkeys had to maintain in working memory both the picture and the currently relevant abstract rule (same or different) during the delay. Many PFC neurons, more than one-third, encoded the abstract rule.

What function does the ability to abstract a rule serve? Abstraction is a type of generalization that permits a shortcut in learning, allowing us to apply what we have learned in similar situations to novel circumstances. Without the ability to abstract the general principles behind related situations, we would have to learn the correct behavior on each new occasion by trial and error. The problem with such trial-and-error learning is that errors necessarily occur. This is a less efficient way of dealing with novelty since errors frequently lead to lost opportunities for reward. Consistent with these ideas, PFC seems to be most important when subjects must deal with novel situations.

Neurobiological Models of Executive Control

Note that the pattern of deficits following PFC damage seems to reflect a selective loss of the higher level functions from the bipartite cognitive architecture pictured in Figure 1. Well-established automatic routines on the lower level are intact and available to be triggered by the appropriate sensory cues. However, without the higher level functions specialized to acquire and represent goals and means, the system would be at the mercy of the environment, ruled by whatever sensory inputs happen to flow into the system and by whatever thoughts, emotions, and actions are strongly associated with these inputs. Thus, behavior would seem impulsive and disinhibited and reactions inappropriate because they would be emitted reflexively without any consideration of the future. Furthermore, without the influence of predicted goals to continually drive task-appropriate processes, individuals would be distractible and would easily go 'off track' when there are temporal gaps between sensory inputs or between those inputs and the individual's responses. The system might be capable of learning to react to cues, but this learning would be inflexible. Without the ability to predict goals and means, the system would be stuck in a behavioral rut, always reacting to a cue with whatever behavior it was first associated.

This architecture is also apparent in neurobiological models of executive control. Network models of Dehaene and Changeux include an executive layer of 'rule-coding' units, thought to correspond to the PFC, that controls the flow of information between the input and output layers (Figure 3(a)). Arthur Shimamura proposed a dynamic filtering model of PFC function in which patterns of information sustained by the PFC select and reroute the flow of activity in posterior association cortex (Figure 3(b)). Miller and Cohen proposed that the cardinal PFC function is to acquire and actively maintain patterns of activity that represent goals and the means to achieve them (rules) in terms of a map of the cortical pathways needed to perform the task (hence 'rulemaps') (Figure 3(c)). Under this model, activation of a PFC rulemap sets up bias signals that propagate throughout much of the rest of the cortex, affecting sensory systems as well as systems responsible for response execution, memory retrieval, emotional evaluation, etc. The aggregate effect is to guide the flow of neural activity along pathways that establish the proper mappings between inputs, internal states, and outputs to best perform the task. In short, rule information is acquired by the PFC, which provides support to related information

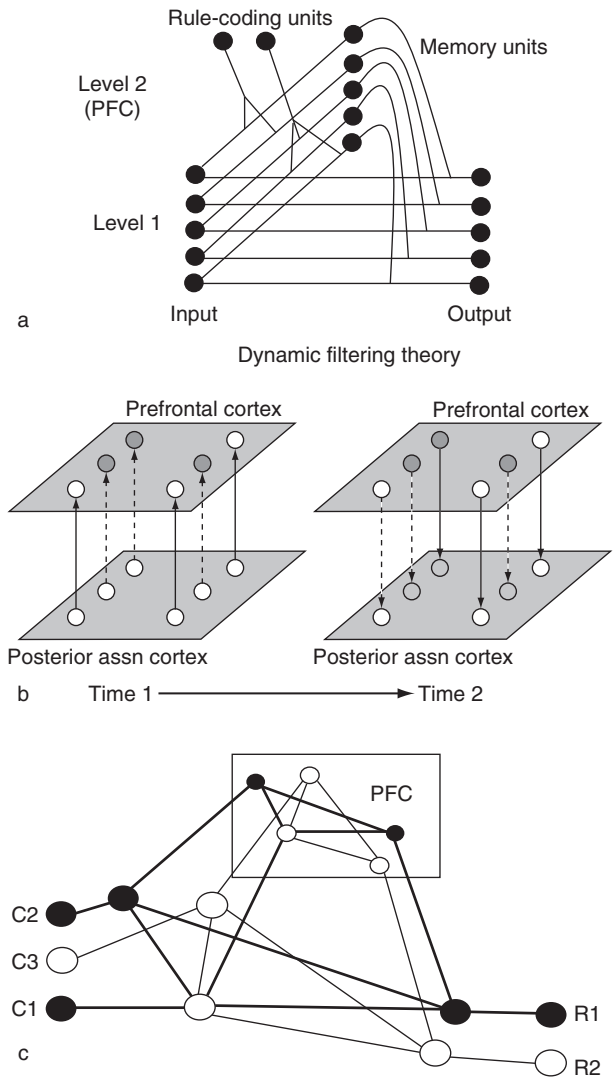

Figure 3 Models of PFC function. (a) The models of Changeux and Dehaene include an executive level (level 2, thought to correspond to the PFC) with rule-coding units and working memory units that gate the flow of activity along lower level input–output lines. (b) Shimamura's dynamic filtering model. At time 1, a region in PFC is activated by feed-forward (i.e., bottom-up) projections from posterior cortex. This activity, along with influences such as task demands, sets up a pattern of activation in PFC (light circles, activation; dark circles, inhibition). At time 2, posterior cortex is modulated by PFC via feedback (i.e., top-down) projections that enable both selection of task-relevant and inhibition of task-irrelevant neural activity. Figure provided by Arthur Shimamura. (c) In the Miller and Cohen model, a task model is formed in the PFC when reward signals link together neurons activated by the events that led to reward. A subset of cues can then activate the entire representation. Bias signals resulting from the maintenance of this pattern guide the flow of activity along neural pathways that establish the task-relevant mappings between representations of inputs (C1, C2, C3), internal states, and outputs (R1, R2) in posterior cortex. Thick lines and solid circles denote active processing lines. Reproduced from Miller EK and Cohen JD (2001) An integrative theory of prefrontal function. *Annual Review of Neuroscience* 24: 167–202.

in posterior brain systems, effectively acting as a global controller or 'traffic cop.'

A key issue is to understand how PFC acquires these high-level representations. One possibility is that they develop through the action of reinforcement signals on multimodal PFC circuitry. During learning, when a behavior meets with success, reward signals augment the unique pattern of PFC activity evoked by that situation, action, and consequence by strengthening connections between the neurons activated by those events. These representations might begin quite modestly, but with time and repeated iterations of the process, the PFC representation can 'bootstrap' into further elaboration as combinations of events and contingencies between them and the requisite actions are learned. Midbrain dopaminergic neurons have ideal properties for the reinforcement signal. Early in learning, rewards activate dopamine neurons, but later in learning the cues that predict those rewards activate dopamine neurons rather than the rewards. Furthermore, there is an inhibition of the firing rate of the dopamine neurons when an expected reward does not occur. This 'reward prediction error' signal is ideal for instructing when and what the system should learn and consequently enables the organism to acquire the representations necessary to achieve reward.

The resulting pattern of PFC activity reflects the task's contingencies and in neural terms amounts to a 'map' of the pathways between inputs, internal states, and outputs needed to perform the task successfully. The ability to construct this map arises from the intermingling of diverse information that takes place within PFC. Although the PFC may be critically involved in acquisition of this information, other parts of the brain may be responsible for its consolidation in long-term memory. The PFC, however, would retain the requisite links to retrieve the appropriate pattern and bring it online to guide behavior.

Summary

Executive control refers to the ability to take charge of one's actions and direct them toward unseen aims. Virtually all theories of cognition posit that executive control requires making predictions about available goals and what means might achieve them. This knowledge can then be used to select and coordinate among a myriad of lower level, automatic sensory, memory, and motor functions.

The PFC, a brain structure that reaches its greatest complexity in the primate brain, seems to play a central role in executive control. It has access to all major forebrain systems, as well as the means to influence them. Individuals with PFC damage seem capable only of emitting habitual or innate reactions to the immediate environment without any consideration or anticipation of future consequences and unseen goals. Neurophysiological studies indicate that PFC neurons seem to mediate the mechanisms essential for a cognitive control system: They are multimodal, they acquire and signal the formal demands of tasks, and they can sustain their activity to keep task-relevant information online and available during behavior.

Different theories of PFC function emphasize different contributions to cognitive control: maintenance and/or manipulation of information in working memory, the assignment of affective tags on events and choices, and the acquisition and online representation of task rules. These theories are not mutually exclusive; each may reflect different facets of a system that is necessarily multivariate because of its role in guiding many different behaviors and interfacing with many other brain systems. However, in the healthy individual, the different processes work in harmony and enable the organism to behave flexibly and cope efficiently with novel situations, which are hallmark features of PFC function.

See also: Cognitive Control and Development; Cognitive Deficits in Schizophrenia; Congenital Muscular Dystrophy; Executive Function and Higher-Order Cognition: Neuroimaging; Executive Function and Higher-Order Cognition: Computational Models; Frontal Lobe Syndrome; Prefrontal Cortex: Structure and Anatomy; Short Term and Working Memory.

Further Reading

Damasio AR (1994) *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Putman.

Fuster JM (1995) *Memory in the Cerebral Cortex*. Cambridge, MA: MIT Press.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal function. *Annual Review of Neuroscience* 24: 167–202.

Schultz W and Dickinson A (2000) Neuronal coding of prediction errors. *Annual Review of Neuroscience* 23: 473–500.

Executive Function and Higher-Order Cognition: EEG Studies

L Y Deouell, The Hebrew University of Jerusalem, Jerusalem, Israel
R T Knight, University of California at Berkeley, Berkeley, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Multiple goals, from survival to pleasure, coexist at every given moment, and internal needs as well as external events act in a push–pull manner to bias behavior. To maintain optimal goal-directed behavior, a control ('executive') system is needed that will dynamically prioritize the processing of information as well as the planning of actions and their execution. In addition to direct motor planning, major components of this executive system are working memory, attention, and conflict/error monitoring. Working memory would allow maintaining a goal across time, as well as information required to achieve the goal. Selective attention ('voluntary' or 'endogenous' attention) would facilitate the processing of one stream of input and suppress another while involuntary attention mechanisms allow for changes in the environment, as well as changes in internal drives, to interfere with ongoing behavior in a rapid and flexible manner. A monitoring module is needed to assess the efficiency of the executed behavior so that behavior can be adjusted in an optimal way. Consequently, the executive system needs to interact with multiple sensory regions of the brain, as well as with motor output regions involved in orientation, locomotion, and speech. Event-related brain potentials (ERPs), recorded on the scalp (and recently also intracranially) have established scalp-recorded signatures of executive functions. The effect of brain lesions on these measures of electrical brain activity provides a window into the networks supporting the executive system. This article describes ERP studies conducted specifically with patients suffering from well-circumscribed brain lesions involving mainly the lateral prefrontal cortex (LPFC), a major hub of the executive system.

Selective Attention

Patients with LPFC damage are frequently unable to suppress inappropriate reaction to objects and events in their environment (so-called environmental dependency syndrome) or to suppress prepotent responses. Whereas some of these effects may be at the level of response selection, ERP studies of these patients suggest that LPFC is essential for normal suppression of information processing at the level of unimodal sensory cortices.

Knight and coworkers delivered auditory stimuli (clicks or tone bursts), or brief electric shocks to the median nerve, to patients with damage to LPFC and to patients with comparably sized lesions in the temporoparietal junction or the lateral parietal cortex, as well as to age-matched controls. Both types of stimuli were distractors, not relevant to the task. Lesions in posterior association cortex, sparing primary sensory regions, had no effects on the amplitudes or latencies of the primary cortical evoked responses. Lesions invading either the primary auditory or somatosensory cortex reduced early latency (20–40 ms) evoked responses generated in these regions. In a distinct contrast, LPFC damage resulted in enhanced amplitude of both the primary auditory and somatosensory evoked responses generated 20–40 ms poststimulation. Spinal cord and brain stem potentials were unaffected by prefrontal damage, indicating that the amplitude enhancement of primary cortical responses was due to abnormalities in either prefrontal-thalamic connectivity or direct interactions between LPFC and sensory cortex. Selectively attending to a sensory stream and maintaining the goal of behavior requires not only enhancement of processing of some sensory stimuli or motor plans, but also suppression, or inhibition, of competitive stimuli or plans ('biased competition'). Apparently, a signal from LPFC to upstream sensory cortices provides such a biasing signal. Failure to produce such a signal may impair selective attention.

The failure to inhibit irrelevant sensory input indeed affects the patients' performance. Chao and Knight showed that patients with LPFC lesions are impaired in an auditory delayed-match-to-sample task when the delay period is filled with irrelevant tone pips. In the same task, the irrelevant tones elicited abnormally augmented primary auditory potentials (Na, Pa) in the patients, and the performance decrement directly correlated with the Pa enhancement. Similarly, in patients with right prefrontal damage, sounds presented to an unattended ear in a dichotic paradigm abnormally reduced the attentional enhancement (see the section titled 'Novelty and deviance detection and involuntary attention shift') of subsequent to-be-attended sounds. In agreement with these findings, patients with prefrontal lesions show reduced or even reversed negative priming effects, which suggests that processing of stimuli that should be outside the focus of attention is not properly suppressed.

The role of the LPFC in biased competition is not limited to inhibition. In fact, LPFC may also provide the excitatory signal. This excitation may be

implemented through three distinct mechanisms: (1) through a tonic excitatory influence on ipsilateral posterior areas, affecting attended and nonattended sensory inputs alike; (2) by enhancement of extrastriate cortex response to attended information; and (3) by a phasic excitatory influence on ipsilateral posterior areas' response to correctly perceived task-relevant stimuli (targets). This was shown in a series of ERP studies conducted by Knight and colleagues in patients with LPFC damage (centered on Brodmann's areas 9 and 46), in which the patients were asked to detect rare (odd ball) targets (inverted triangles) in a series of distractors (upright triangles). This paradigm is referred to below as the triangles task.

When the stimuli in the triangles task were presented at fixation, the patients showed a diminished extrastriate N1 component (at 170 ms) relative to controls. A similar reduction of N1 was observed in another experiment, which used visual word stimuli. Even more revealing were conditions in which targets and nontargets were lateralized and could appear in either visual field. In one of these conditions, the patients were instructed to attend to all stimuli regardless of location. The extrastriate P1, immediately preceding the N1 component, was reduced for all stimuli presented to the contralesional hemifield. Moreover, when the patients' attention was directed to one of the fields at a time, this reduction of P1 over ipsilesional extrastriate cortex was observed for attended and nonattended contralateral stimuli alike, and can thus be considered attention independent. Thus, LPC seems to exert a modulatory, tonic, attention-independent facilitation over processing at ipsilateral extrastriate cortex. A similar effect of reduced association cortex activity ipsilateral to LPFC damage has been reported for auditory-induced N1 potentials.

Directing attention to one hemifield in a target detection task normally elicits a sustained negativity (known as 'selection negativity' or 'negative difference'), starting at around 100 ms, for all stimuli presented at this hemifield. When attention was directed to the contralesional visual field of LPFC patients, this attention effect on extrastriate cortex was normal in the first 200 ms after stimulus onset but was significantly disrupted thereafter. This finding suggests that LPFC exerts attention-dependent selective facilitation starting around 200 ms after stimulus onset. Other cortical areas are probably responsible for attention-dependent regulation of extrastriate cortex in the first 200 ms. It is conceivable that inferior parietal cortex is responsible for the early reflexive component of attention whereas LPFC is responsible for more-controlled and sustained aspects of visual attention beginning later on. The effect of LPFC lesion on the selection negativity is not limited to visual areas. In an auditory selective attention task, prefrontal lesion patients generated reduced selection negativity as well. Notably, this diminished attention effect depended on the side of the lesion. Patients with left hemisphere lesions showed a mildly reduced attention effect regardless of which ear was attended. In contrast, patients with right prefrontal damage showed reduced effect mainly when the contralateral ear was to be attended. Posterior association cortex lesions in the temporoparietal junction had comparable effects on the selection negativity regardless of ear of stimulation.

Target detection in the visual oddball paradigm described above is normally associated with a prominent late response, starting at 200 ms and continuing for the next 500 ms, including the N2–P3b complex. Since this component is elicited only by detected targets, it is considered to be a manifestation of phasic top-down effects contingent on the identity of the stimulus. Following LPFC lesions, the N2, a component which likely reflects postselection processing of the target in the inferior temporal lobe, was not observable over the lesioned hemisphere in response to targets in either visual field. The P3b was reduced over the temporo-occipital electrodes but not at parietal sites, attesting to the fact that the P3b most likely reflects multiple distinct cortical processes related to target detection. The patients' performance in this study, producing more errors, was concordant with this electrophysiological evidence of impaired top-down effects following LPFC lesion. A spatially limited reduction of target P3 was recently reported also by Daffner et al., although in that group of LPFC patients, the main reduction of target P3 was reported to be in anterior electrodes. Thus, in addition to both attention-dependent and attention-independent tonic facilitation, the LPFC is critical in establishing a phasic, stimulus-dependent facilitation.

Novelty and Deviance Detection and Involuntary Attention Shift

Task requirements in an experimental setting, as well as the ongoing goals in natural settings, require that a subset of the sensory stream and a subset of motor plans be preferentially processed and, at least in humans, be available for conscious awareness and deliberation. However, such biased selection may be perilous as events outside the focus of attention might pose either dangers (e.g., an approaching predator) or opportunities (e.g., prey). Such events may be slight perturbations of an established sensory regularity, such as an intensity decrement in background noise, or grossly unexpected ('novel') events, such as the sound of screeching brakes. This need calls for a

surveillance mechanism that automatically detects deviant events.

Automatic response to slight changes in acoustic regularity outside the focus of attention bears the electrophysiological signature of the 'mismatch negativity' (MMN), a frontal scalp negative potential accompanied by lower temporal positivity, with a peak latency of 100–250 ms following the deviation. The MMN presumably reflects an 'error signal' generated automatically in the secondary auditory cortex by a neural mechanism comparing a perceived stimulus to a sensory 'memory trace' formed by the regular stimuli or the process of updating the existing model of the environment. The MMN is elicited when participants attend to a primary task and ignore the stream of sounds in which the pertinent deviations occur. Although it might be somewhat attenuated outside the focus of attention, it is completely suppressed only under very specific circumstances of competition for processing a specific feature between two auditory streams. Moreover, whereas attention-related components of the ERP are largely absent when deviation is predictable (e.g., when a deviation occurs regularly in the stream), the MMN is immune to this manipulation. Thus, the MMN is considered to represent a nonintentional, largely preattentive process. This signal may be a trigger for an ensuing shift of attention toward the deviant event. These shifts are reflected on the scalp as a positivity following the MMN, called P3a.

The P3a is a positive frontoparietal scalp potential peaking around 300 ms following the onset of a rare, task-irrelevant distractor in an oddball sequence. The novelty P3a is similar in response to either novel sounds or visual stimuli. The P3a is considered a marker of attention orienting, elicited by a distributed multimodal corticolimbic system. Both the MMN and the novelty P3a components were shown to be reduced in patients with PFC lesions.

In paradigms designed to study the MMN, patients and controls watched silent movies or were engaged with a visual reaction time paradigm and were instructed to ignore series of repetitive sounds. Infrequently, the regularity of the sound stream was broken by the occurrence of a deviant stimulus. The MMN in response to pitch and pattern changes was reduced in patients with PFC lesions, indicating an early deficit in automatic detection of deviance outside the focus of attention. Whereas temporoparietal lesions caused an MMN reduction mainly to contralesional stimuli, LPFC lesions elicited comparable deficits regardless of stimulus side in one study, and more to ipsilesional sounds in another. In addition, there was a tendency for right prefrontal lesions to be associated with larger MMN reductions than left prefrontal lesions were, but no such asymmetry was found for the temporoparietal lesions. These results suggest different contributions of the temporoparietal region and the prefrontal region to the MMN. Further research will be needed to determine whether the reduced MMN following prefrontal damage reflects a weakened memory trace of the previous regularity due to disinhibition of irrelevant information, reduced frontal facilitation of a comparator mechanism in the secondary auditory cortex (which is the main generator of scalp MMN), a failure to initiate an attention switch following the detection of the change, or damage to a postulated frontal generator of MMN. In addition, it is still unclear whether the effect of LPFC damage is independent of the dimension of the regularity which is disturbed (e.g., spatial location, pitch, or more-abstract dimensions of the sound).

Novelty P3a responses generated over prefrontal scalp sites to unexpected novel stimuli are reduced by prefrontal lesions, with reductions observed throughout the lesioned hemisphere. Comparable P3a decrements have been observed in the auditory, visual, and somatosensory modalities in humans with prefrontal damage. Reductions appear to be more severe after right prefrontal damage. In most studies of novelty P3a, the novel stimuli are completely task irrelevant, and the response is considered to reflect an 'involuntary' orienting response toward the salient event. Accordingly, galvanic skin response, a peripheral marker of the orienting response, is also reduced by damage to the prefrontal as well as posterior association cortex. LPFC damage caused a dramatic reduction in P3a amplitude even in a paradigm in which the novels where relevant to one of two tasks the patients were involved with. In the paradigm used by Daffner et al., patients had to self-pace a succession of visual stimuli while also looking for a designated (non-novel) target. LPFC patients not only showed a significantly attenuated P3a response to novel visual stimuli but also spent less time than controls did looking at novels, suggesting a more general decrement of 'novelty seeking.' Taken together, these findings converge with both clinical observations and animal experimentation supporting a critical role of prefrontal structures in the processing of novel stimuli, probably as part of a prefrontal-hippocampal network.

Monitoring

In demanding task situations, it is important to detect when actions may be erroneous and to adjust behavior to avoid additional mistakes. Evidence from electrophysiological and neuroimaging studies suggests that the anterior cingulate cortex (ACC) and the LPFC are active in situations demanding such monitoring activity. Evidence linking the ACC to action monitoring

derives from studies of error-related negativity (ERN), an ERP that occurs at the moment of an error in cognitive reaction time tasks. Dipole localization studies of ERN suggest that it is generated by a medial frontal structure, most likely the ACC. Functional magnetic resonance imaging (fMRI) studies examining error processing confirm the presence of ACC activation associated with errors or with situations prone to errors. A rare single unit recording in a human patient showed that dorsal ACC is associated with the use of rewards to guide action and further that ablation of the dorsal ACC increases errors in a situation requiring reward processing. Also, fMRI studies demonstrate that in addition to the ACC, activation may occur also in LPFC in conflict situations. Some of these studies suggest that activation in the ACC and LPFC is not related to errors *per se* but rather to the need to avoid errors or losses in more-demanding situations. Co-occurrence of prefrontal and cingulate activity related to error processing has been observed also in single-neuron recordings from nonhuman primates, as well as in intracerebral recordings from implanted depth electrodes in epileptic patients. These observations have not determined, however, whether the ACC and the LPFC interact or are independent.

The question of the interaction of the ACC and the LPFC was investigated by Gehring and Knight, who measured ERN in a group of patients with unilateral circumscribed lesions of the LPFC. A normal pattern of ERN activity in these individuals would indicate that the medial frontal regions operated independently of the LPFC in generating the ERN. In contrast, an absence or reduction of ERN activity in these individuals would indicate that the LPFC was either necessary for generation of the ERN in the ACC, or was itself a generator of the ERN. Other disruptions in the pattern of ERN activity would suggest that LPFC modulated the generation of the ERN, perhaps by supplying information or activation that was critical for ACC processing.

Six patients were studied, four with left and two with right LPFC damage. They were compared to ten age-matched controls and ten younger individuals. Participants made a squeezing response to a pair of letters, one of which was designated as a target letter. One letter appeared in red and the other in green. At 1 s before the letter pair, a precue (the word 'red' or 'green') indicated which letter was the target letter. The task was to respond with one hand if the target letter was 'H' and with the other hand if the letter was 'S'. On half of the trials, the irrelevant flanking letter was identical to the target letter; on the other half of the trials, the irrelevant flanking letter signaled the incorrect response (a manipulation that provoked erroneous responses).

The patients were slower than the control groups but did not perform with more errors than the control groups did. However, they showed less online corrective action in error trials. While controls showed reduced squeezing force during error trials, suggesting online inhibition of response, patients showed a reduction in this effect. In addition, the patients corrected their error less frequently than controls did. The pattern of the patients' electrophysiological response was also dramatically different from that of the control groups. Whereas in the young and age-matched controls, the response-locked negativity was significantly larger during error trials than during correct trials (i.e., it was an ERN), there was no difference between the error and correct trials in the patients with LPFC lesions. However, this did not result from absence of negativity. Rather, an ERN was seen for correct trials just as often as for incorrect trials, and this ERN was not different in amplitude from the response elicited by the controls during incorrect trials.

This pattern suggests an interaction or interdependence between the LPFC and the ACC in error or conflict monitoring. However, it is not consistent with a simple serial model in which the LPFC detects the need for executive control during high-conflict situations and signals the ACC, which performs the actual control function. Nor is it compatible with a serial model in which the ACC detects conflicts and transmits to the LPFC, which exerts its executive control functions. Since the most conspicuous observation was the lack of differentiation between correct and error trials, one possible account suggested by Gehring and Knight was that the damaged LPFC failed to represent or transmit the contextually appropriate stimulus–response mapping. As a consequence, the ACC may not have been able to confirm a correct response and thus elicited an ERN by default for every trial. Alternatively, the impaired stimulus–response representation in the LPFC might have created an additional conflict in each trial, detected by the ACC. In both cases, the ACC would produce an unreliable cue for errors, reflected in the disruption of corrective responses.

Conclusion

The behavior of patients with LPFC lesions suggests an interruption of executive and attentional control at multiple levels. Thus, in everyday life, patients may inadequately respond to environmental stimuli, be disproportionately distracted by irrelevant information, and be unable to focus on a task, on the one

hand, and unable to flexibly shift from one goal to another (perseveration), on the other hand. The data from electrophysiology reveal the role of LPFC in automatic processes such as maintaining an adequate level of responsiveness to all stimuli, detecting change, and processing novelty, as well as controlled processes such as inhibition of irrelevant information, enhanced processing of relevant information, phasic responses to targets, and conflict monitoring. Moreover, the electrophysiological data reveal that LPFC implements some of these functions through interaction with remote regions of the brain, including unimodal sensory cortices, the hippocampus, and the ACC. All these effects are evident early (100–300 ms) in the course of neural processing.

See also: Attention: Models; Attentional Functions in Learning and Memory; Cognitive Control and Development; Electroencephalography (EEG); Electrophysiology: EEG and ERP Analysis; Executive Function and Higher-Order Cognition: Assessment in Animals; Executive Function and Higher-Order Cognition: Definition and Neural Substrates; Working Memory: Capacity Limitations.

Further Reading

Alain C, Woods DL, and Knight RT (1998) A distributed cortical network for auditory sensory memory in humans. *Brain Research* 812: 23–37.

Barcelo F, Suwazono S, and Knight RT (2000) Prefrontal modulation of visual processing in humans. *Nature Neuroscience* 3: 399–403.

Boller F and Grafman J (eds.) (2002) *Handbook of Neuropsychology: The Frontal Lobes*. Amsterdam: Elsevier.

Chao LL and Knight RT (1998) Contribution of human prefrontal cortex to delay performance. *Journal of Cognitive Neuroscience* 10: 167–177.

Daffner KR, Scinto LFM, Weitzman AM, et al. (2003) Frontal and parietal components of a cerebral network mediating voluntary attention to novel events. *Journal of Cognitive Neuroscience* 15: 294–313.

Desimone R and Duncan J (1995) Neural mechanisms of selective visual-attention. *Annual Review of Neuroscience* 18: 193–222.

Gehring WJ and Knight RT (2000) Prefrontal-cingulate interactions in action monitoring. *Nature Neuroscience* 3: 516–520.

Knight RT, Hillyard SA, Woods DL, and Neville HJ (1981) The effects of frontal cortex lesions on event-related potentials during auditory selective attention. *Electroencephalography and Clinical Neurophysiology* 52: 571–582.

Knight RT, Scabini D, Woods DL, and Clayworth C (1988) The effects of lesions of superior temporal gyrus and inferior parietal lobe on temporal and vertex components of the human AEP. *Electroencephalography and Clinical Neurophysiology* 70: 499–509.

Lhermitte F (1986) Human autonomy and the frontal lobes. Part II: Patient behavior in complex and social situations: The "environmental dependency syndrome." *Annals of Neurology* 19: 335–343.

Metzler C and Parkin AJ (2000) Reversed negative priming following frontal lobe lesions. *Neuropsychologia* 38: 363–379.

Miller EK (2000) The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience* 1: 59.

Ramnani N and Owen AM (2004) Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience* 5: 184.

Stuss DT and Knight RT (eds.) (2002) *Principles of Frontal Lobe Function*. Oxford, UK: Oxford University Press.

Williams ZM, Bush G, Rauch SL, Cosgrove GR, and Eskandar EN (2004) Human anterior cingulate neurons and the integration of monetary reward with motor responses. *Nature Neuroscience* 7: 1370–1375.

Executive Function and Higher-Order Cognition: Neuroimaging

S A Bunge and M J Souza, University of California at Davis, Davis, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Executive Function

The terms executive function and cognitive control-refer to cognitive processes associated with the control of thought and action. Putative control functions include the ability to (1) selectively attend to relevant information while filtering out distracting information (selective attention and interference suppression), (2) work with information that is currently being held in working memory (manipulation), (3) flexibly switch between tasks (task switching), (4) inhibit inappropriate response tendencies (response inhibition), and (5) represent contextual information that determines whether a thought is relevant or whether an action is appropriate (e.g., task-set representation).

Although a number of control processes have been proposed, it is unclear which of these putative processes are distinct from one another. Thus, an important challenge in cognitive control research is to identify a set of elemental control processes and to understand the brain mechanisms underlying each of these processes. One approach that has proven useful for dissociating cognitive processes is the study of patients with various forms of neurological disease and damage. This approach, however, is not without its limitations. As such, convergent evidence from brain imaging techniques is proving increasingly useful.

Probing Executive Function with Neuroimaging Techniques

Functional magnetic resonance imaging (fMRI) has proven useful for identifying brain regions involved in cognitive control. As such, this technique allows us to test whether tasks involving putatively distinct control processes recruit separable brain networks. On the other hand, fMRI allows us to determine whether different tasks that involve a common control process recruit some of the same brain regions.

Prefrontal cortex (PFC) has long been implicated in executive function. Indeed, in 1895, the Italian physiologist Bianchi observed that experimental ablations of the frontal lobes destroyed the ability to synthesize incoming percepts together and to integrate these with outgoing motor commands. Since this early work more than a century ago, countless experiments have been conducted with a variety of techniques to better understand how PFC exerts control over other brain regions.

It is now clear that PFC should not be considered a monolithic structure; indeed, this expansive region constitutes roughly one-third of the human brain and consists of a number of different subregions. Each of these subregions is thought to provide a distinct contribution to cognition via different cellular characteristics and anatomical connectivity. As such, the use of a technique with high spatial resolution, such as fMRI, is essential for determining whether two executive tasks rely on the same or different PFC subregions.

Additionally, it is clear that if we want to understand cognitive control mechanisms, we must examine how PFC interacts with other brain regions. Some information about these interactions can be gleaned from functional connectivity analyses of fMRI data, although fMRI provides little information about the timing of these interactions. For several decades, event-related potential (ERP) data from electroencephalography (EEG) studies have contributed to our knowledge about the timing of control mechanisms. The ERP research of Robert Knight, involving patients with PFC lesions and healthy controls, demonstrated that PFC is necessary for top-down enhancement of relevant information in primary sensory regions.

Tasks Used to Study Executive Function

The most common type of executive task requires participants to override a prepotent response tendency. In the widely used Stroop task, one must override the impulse to read a word out loud and instead indicate the name of the color in which the word is printed (Figure 1a). In the go/no-go paradigm, one must press a button in response to a series of rapidly presented visual stimuli but must inhibit responding to a particular stimulus (Figure 1b). The continuous performance test known as AX-CPT is similar to go/no-go but additionally has a monitoring component because one must respond to an X followed by an A, but not to an X followed by a B. In the flanker paradigm, one must press one of two buttons in response to a central target stimulus while overriding the response specified by distracters on either side of the target (Figure 1c). In a task-switching paradigm, one must switch abruptly from one task rule to another (Figure 1d). This type of paradigm is thought to involve both suppression of the prior task rule and retrieval of the new rule. Here, several experimental paradigms are discussed in-depth.

Selective Attention

Goal-directed behavior hinges on the ability to focus on relevant information and ignore distracters, a

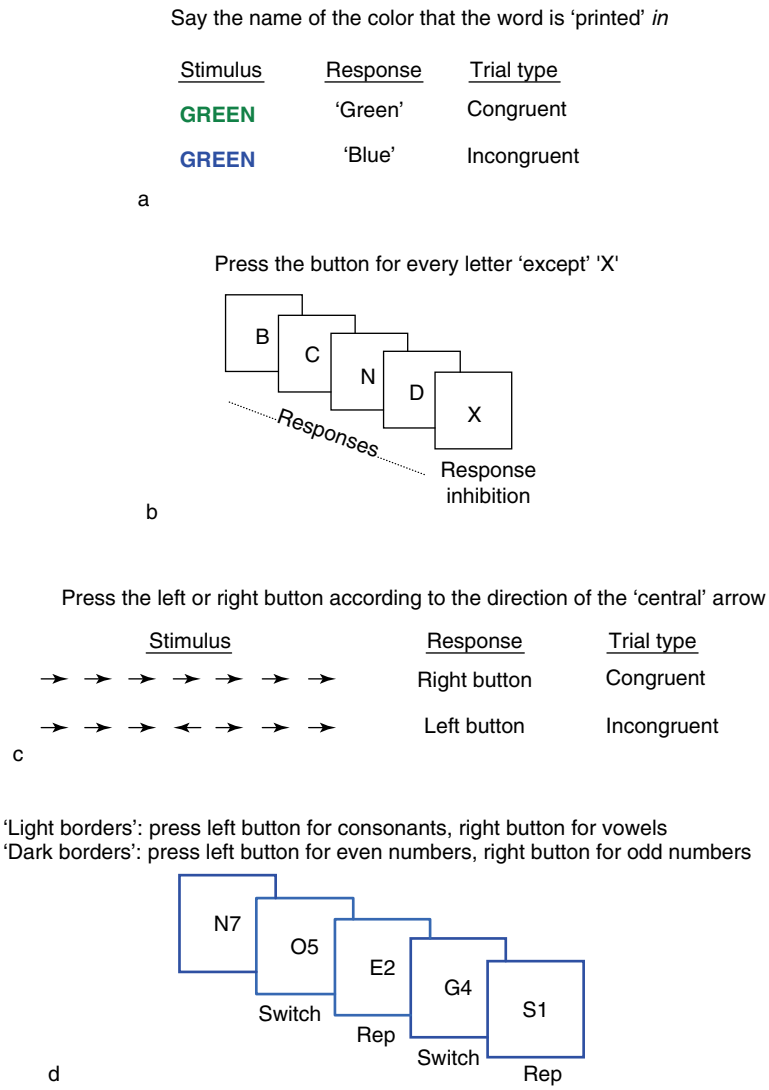

Figure 1 Standard versions of several tasks commonly used to study executive function: (a) the Stroop task, (b) the go/no-go task, (c) the Flanker task, and (d) an example of a task-switching paradigm. Variants of these basic paradigms have been used in a number of brain imaging studies.

function referred to as selective attention and/or interference suppression. Although the neural mechanisms are not yet fully understood, a widely accepted model proposed by Desimone and Duncan postulates that selective attention relies on top-down biasing mechanisms. The model proposes that long-range excitatory projections from PFC to posterior cortical regions enhance the activation of relevant representations, which in turn serve to suppress irrelevant representations through local inhibitory interactions. Herd et al. adapted this model in such a way that it can better account for performance and brain activation data from the Stroop task.

Building on earlier behavioral and fMRI work by Lavie and collaborators, Gazzaley and colleagues have conducted fMRI and ERP studies that provide compelling evidence for separable top-down enhancement and suppression effects in a selective attention paradigm. In their task, participants viewed a series of pictures of faces and scenes pseudorandomly intermixed. They were instructed to (1) remember the faces and ignore the scenes, (2) remember the scenes and ignore the faces, (3) remember both, or (4) passively view the stimuli.

The researchers found that a region involved in processing faces – the fusiform face area (FFA) – exhibited enhanced activation relative to the other conditions when participants selectively attended to faces. They further showed that FFA activation was lower relative to the passive view condition when participants ignored faces and attended to scenes instead. The opposite pattern of results was found for the parahippocampal place area (PPA), a region involved in processing scenes (Figure 2).

An ERP study involving the same paradigm also showed an enhancement of face processing-related neural activity (the face-selective N170 component)

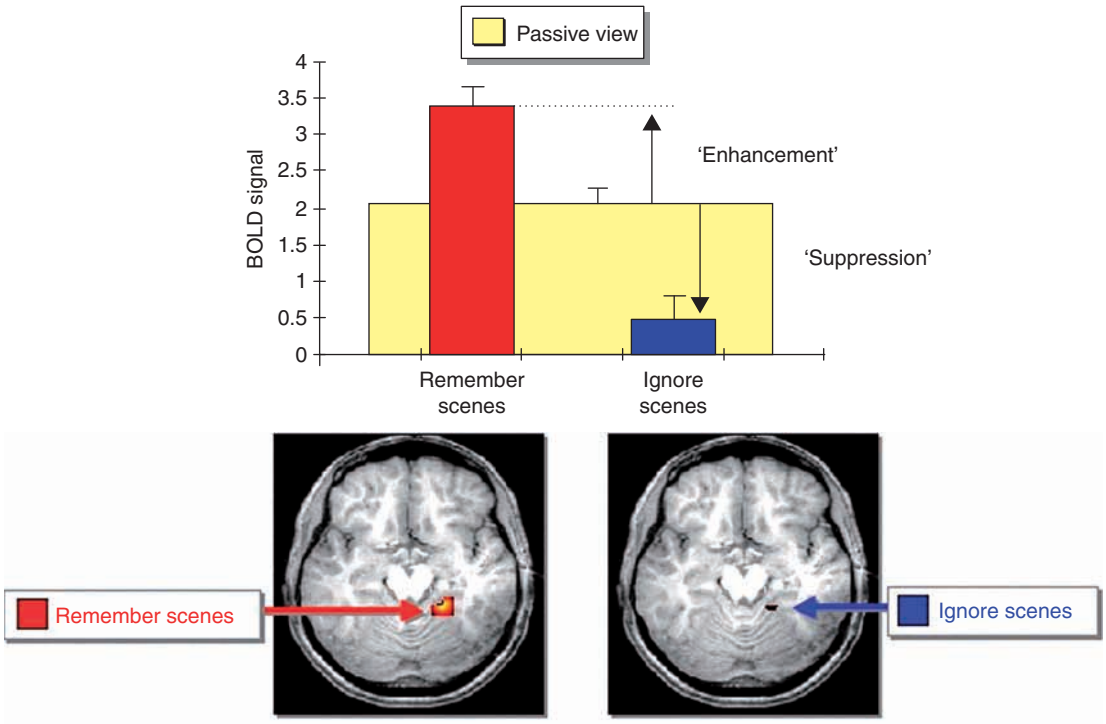

Figure 2 Data from 18 adults aged 19–30 years in an fMRI study on selective attention by Gazzaley and colleagues. These data show that the PPA was more active when participants selectively attended to scenes rather than faces than when they passively viewed both scenes and faces. Additionally, the PPA was less active when participants selectively attended to faces (ignored scenes) than in the passive viewing condition. Participants view the same number of scenes across tasks, and therefore the modulation observed in PPA must be mediated by top-down mechanisms. BOLD: Blood-oxygen level dependent. Courtesy of Adam Gazzaley.

when subjects were instructed to selectively attend to faces. The ERP data further revealed that attention to faces led to faster face processing: indeed, the peak latency of the N170 component was shifted 10 ms earlier for the 'remember faces' condition than for 'ignore faces.' Taken together, these fMRI and ERP results provide strong evidence for top-down modulation of posterior cortical representations as a function of goal relevance.

A large behavioral literature suggests that older adults have difficulty ignoring irrelevant information, but it is unclear from these data whether this deficit is rooted in a specific deficit in top-down suppression of irrelevant information, top-down enhancement of relevant information, or both. As such, Gazzaley and colleagues used the paradigm discussed previously to study selective attention in older adults. Their fMRI data revealed that in older adults, the PPA was not suppressed in the 'ignore scenes' condition relative to the 'passive view' condition (Figure 3). In contrast, the PPA did exhibit enhancement for the 'remember scenes' condition relative to 'passive view.' These data indicate that older adults showed a specific deficit in top-down suppression. In addition to providing insight into cognitive changes during aging, this finding lends credence to the idea that suppression and enhancement mechanisms are neurally separable.

Response Inhibition

The ability to inhibit a contextually inappropriate response tendency is central to goal-directed behavior. Numerous brain imaging studies have been conducted in an effort to identify the brain structures that mediate response inhibition. In the go/no-go paradigm, participants must respond to the presentation of each of a rapidly presented stream of visual stimuli, thereby building up a prepotent response tendency. Participants are also instructed to withhold their response to a particular stimulus (the 'no-go' stimulus), which will be presented without warning from time to time (Figure 1b). fMRI studies involving the go/no-go paradigm have revealed a network of largely right-hemispheric brain regions associated with successful response inhibition, including right ventrolateral PFC. The go/no-go paradigm has been used not only to study inhibitory control in healthy young adults but also to probe the neural basis of inefficient inhibitory control in children and in patient populations.

Another paradigm used to study response inhibition is the stop-signal task. In the most common variant of this paradigm, participants press a button in response to visual stimuli but must on some occasions (e.g., if they hear a tone) inhibit their response at the last moment (Figure 4a). This task is thought to tax

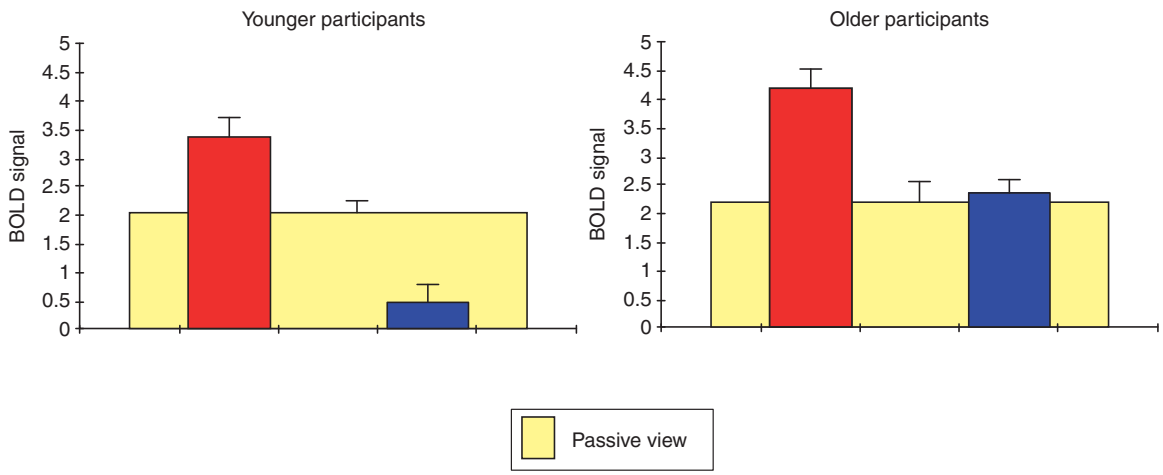

Figure 3 The data from younger adults in **Figure 2** are replotted here as a point of comparison for PPA activation in 16 older adults, aged 60–77 years. For the older adults, like the younger adults, activation in PPA was greater during selective attention to scenes than during passive viewing. In contrast, PPA activation for older adults did not differ between the ignore scenes condition and the passive viewing condition. Thus, the older participants exhibited top-down enhancement of PPA activation but failed to demonstrate top-down suppression. Blood-oxygen level dependent. Courtesy of Adam Gazzaley.

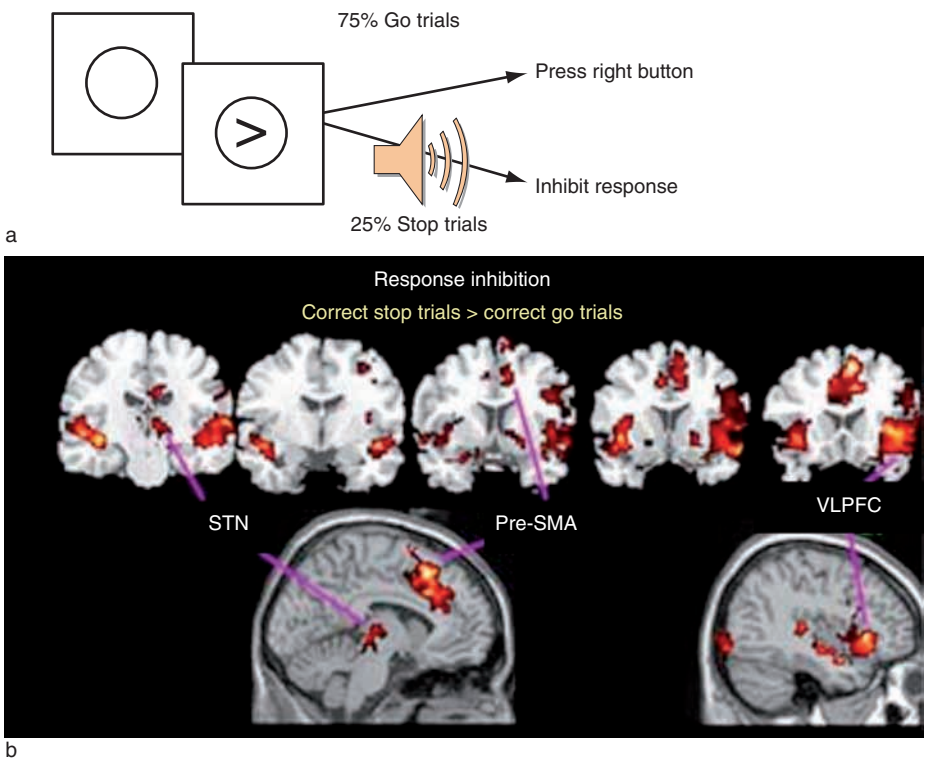

Figure 4 (a) A simplified illustration of the stop-signal paradigm used by Aron and Poldrack. (b) The neural correlates of successful inhibition of motor programs are shown in these activation maps. In particular, three regions – all in the right hemisphere – have been associated with stopping: the subthalamic nucleus (STN), pre-supplementary motor area (pre-SMA), and ventrolateral prefrontal cortex (VLPFC). This pattern of activation suggests that stopping behavior relies on the hyperdirect pathway from PFC to the STN.

response inhibition more heavily than the go/no-go paradigm because participants must override a motor response that they have already initiated. The stop-signal paradigm has an advantage over other response inhibition paradigms for use in clinical and developmental brain imaging studies: the task difficulty varies dynamically on a trial-by-trial basis to ensure a fixed level of performance across participants. As such, differences in brain activation between groups cannot simply be attributed to differences in performance.

In an fMRI study involving the stop-signal paradigm, Aron and Poldrack proposed that stopping is

achieved through excitation of the subthalamic nucleus by PFC (Figure 4b). From prior work in animals, it is known that the subthalamic nucleus is inhibitory, and that excitation of this nucleus should result in a dampening of activity in the thalamus and cortex. This fMRI study provides evidence in humans for hyperdirect inhibition, bypassing the slower, indirect loop through the basal ganglia. It is likely that some response inhibition paradigms engage the hyperdirect pathway, whereas others engage either the indirect pathway through the basal ganglia or cortico-cortical projections that result in top-down biasing of posterior representations. Further brain imaging research is needed to determine whether various types of response inhibition are indeed neurally separable.

Task-Set Representation

The majority of research on cognitive control focuses on the issue of how PFC controls thought and behavior. An equally interesting question is how PFC 'knows' which thoughts and behaviors to facilitate or suppress. Until we have a satisfactory answer for this question, we will continue to treat PFC as a homunculus – a little man inside the brain who mysteriously knows everything and can direct our thoughts and actions accordingly.

Computational modelers have attempted to decompose or decentralize the homunculus. However, more neuroanatomically detailed computational models will be needed to account for the roles of distinct PFC subregions and other parts of the brain. Such models await a larger body of research on context representations – that is, the information that is used to decide on an appropriate course of action. Context representations could include information about current goals, motivational state, the current situation, rules for behaving in this situation, etc.

The type of context representation that can perhaps be most readily operationalized and studied is the representation of a currently relevant task rule. Research in our laboratory is focused on understanding how rules for behavior are stored in long-term memory, retrieved, maintained online during task preparation, and implemented. Along the same lines, Sakai and Passingham have conducted several elegant studies examining the role of anterior prefrontal cortex (APF; lateral Brodmann area 10) in task preparation.

In these studies, Sakai and Passingham provide evidence from functional connectivity analyses that APF interacts with different brain regions during task preparation, depending on the type of task to be performed. Their first study, in 2003, showed that when participants prepared to perform either a challenging verbal or spatial working memory task, APF activation was strongly correlated with regions involved in verbal or spatial working memory, respectively. Their second study, in 2006, focused exclusively on the verbal domain, and it showed that when participants prepared to perform either a semantic or phonological task, APF was correlated with distinct regions in ventrolateral PFC that have been associated with either semantic or phonological processing. These studies show that APF coordinates task performance by interacting with brain regions that are needed to perform an upcoming task.

Current and Future Directions

During the past decade of brain imaging research, the emphasis has been on using fMRI to localize individual brain regions engaged in various tasks. These endeavors have given us a fairly good idea of the key players in cognitive control and have allowed us to test specific hypotheses regarding their contributions. Furthermore, fMRI studies focusing on adult-individual differences have shown that task performance can be tightly correlated with level of engagement of specific brain regions. Similarly, developmental and clinical fMRI studies have identified brain regions for which differences in level of activation across groups may underlie performance differences.

The next generation of brain imaging techniques will need to combine high-resolution spatial and temporal information. Additionally, further research is needed to better characterize neuropharmacological and genetic influences on brain activation. Several methods that hold promise for the next wave of studies probing executive function are discussed here.

Combined fMRI and EEG Methodology

fMRI and EEG are complimentary techniques in that the former allows for the precise spatial localization of active brain regions, and the latter provides exquisite timing information about the onset and offset of neural activity. To push our understanding of cognitive control mechanisms to the next level, it will be necessary to acquire data with high spatial and temporal resolution.

One approach is to acquire fMRI and EEG data on the same group of participants, in separate sessions. The set of brain regions identified from the fMRI data can then be used to constrain the source localization of the ERP data. Another, more technically challenging, approach is to simultaneously acquire fMRI and EEG data from the same participants. This latter approach has been used successfully in a study on performance monitoring by Debener and colleagues. Several groups are working on multimodal imaging approaches, including combined fMRI/EEG, and the further development of these methods should provide new insights into the neural mechanisms of executive function.

Event-Related Optical Signals

Gratton and Fabiani and their colleagues have developed an optical imaging technique with high spatial and temporal resolution known as EROS, or the event-related optical signal. The standard analytic approach for optical imaging data analysis is to focus on changes in optical properties of the cortical brain tissue associated with changes in blood flow, which are only indirectly related to changes in neural activity. Given the sluggish blood flow response to neuronal activity, optical imaging – like fMRI – typically has poor temporal resolution. In contrast, with EROS, the optical changes that are being analyzed are directly related to neural activity. In addition to having high temporal resolution, the EROS signal also has a high degree of spatial resolution; it can be localized to an area of less than a cubic centimeter. The main drawback of this method is that it can only be used to measure activity within 3–5 cm of the surface of the brain.

The developers of EROS have used it to study response competition in a Stroop task. They showed that both left and right motor cortices were active in response to an incongruent Stroop stimulus, even though participants were instructed to respond to the stimulus by pressing a button with only one hand. In contrast, congruent Stroop stimuli, which did not elicit response competition, elicited activation of only the contralateral motor cortex. Since this validation study was performed approximately 5 years ago, the technique has been further fine-tuned and validated. Although only a few published studies have employed EROS thus far, the use of this technique may spread as more researchers become aware of its potential.

See also: Attentional Networks; Cognition: An Overview of Neuroimaging Techniques; Cognitive Control and Development; Electroencephalography (EEG); Event-Related Potentials (ERPs) and Cognitive Processing; Executive Function and Higher-Order Cognition: Assessment in Animals; fMRI: BOLD Contrast; Neuroimaging; Optical Imaging of Intrinsic Signals; Prefrontal Cortex: Structure and Anatomy.

Further Reading

Aron AR and Poldrack RA (2006) Cortical and subcortical contributions to Stop signal response inhibition: Role of the subthalamic nucleus. *Journal of Neuroscience* 26: 2424–2433.

Braver TS and Barch DM (2002) A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews* 26: 809–817.

Bunge SA, Dudukovic NM, Thomason ME, Vaidya CJ, and Gabrieli JD (2002) Immature frontal lobe contributions to cognitive control in children: Evidence from fMRI. *Neuron* 33: 301–311.

Casey BJ, Trainor RJ, Orendi JL, et al. (1997) A developmental functional MRI study of prefrontal activation during performance of a go-no-go task. *Journal of Cognitive Neuroscience* 9: 835–847.

Collette F, Hogge M, Salmon E, and Van Der Linden M (2006) Exploration of the neural substrates of executive functioning by functional neuroimaging. *Neuroscience* 139: 209–221.

de Fockert JW, Rees G, Frith CD, and Lavie N (2001) The role of working memory in visual selective attention. *Science* 291: 1803–1806.

Debener S, Ullsperger M, Siegel M, et al. (2005) Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *Journal of Neuroscience* 25: 11730–11737.

Desimone R and Duncan J (1995) Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18: 193–222.

DeSoto MC, Fabiani M, Geary DC, and Gratton G (2001) When in doubt, do it both ways: Brain evidence of the simultaneous activation of conflicting motor responses in a spatial Stroop task. *Journal of Cognitive Neuroscience* 13: 523–536.

Durston S, Tottenham NT, Thomas KM, et al. (2003) Differential patterns of striatal activation in young children with and without ADHD. *Biological Psychiatry* 53: 871–878.

Frith CD and Friston KJ (1997) Studying brain function with neuroimaging. In: Rugg M (ed.) *Cognitive Neuroscience*, pp. 169–195. Cambridge: MIT Press.

Garavan H, Ross TJ, and Stein EA (1999) Right hemispheric dominance of inhibitory control: An event-related functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America* 96: 8301–8306.

Gazzaley A, Cooney JW, McEvoy K, Knight RT, and D'Esposito M (2005) Top-down enhancement and suppression of the magnitude and speed of neural activity. *Journal of Cognitive Neuroscience* 17: 507–517.

Gazzaley A, Cooney JW, Rissman J, and D'Esposito M (2005) Top-down suppression deficit underlies working memory impairment in normal aging. *Nature Neuroscience* 8: 1298–1300.

Goethals I, Audenaert K, Van de Wiele C, and Dierckx R (2004) The prefrontal cortex: Insights from functional neuroimaging using cognitive activation tasks. *European Journal of Nuclear Medicine and Molecular Imaging* 31(3): 408–416.

Goldberg TE and Weinberger DR (2004) Genes and the parsing of cognitive processes. *Trends in Cognitive Science* 8(7): 325–335.

Gratton G and Fabiani M (2001) Shedding light on brain function: The event-related optical signal. *Trends in Cognitive Science* 5: 357–363.

Gratton G and Fabiani M (2003) The event-related optical signal (EROS) in visual cortex: Replicability, consistency, localization, and resolution. *Psychophysiology* 40: 561–571.

Herd SA, Banich MT, and O'Rei RC (2006) Neural mechanisms of cognitive control: An integrative model of Stroop task performance and fMRI data. *Journal of Cognitive Neuroscience* 18: 22–32.

Knight RT, Grabowecky MF, and Scabini D (1995) Role of human prefrontal cortex in attention control. *Advances in Neurology* 66: 21–36.

Kutas M and Dale A (1997) Electrical and magnetic readings of mental functions. In: Rugg M (ed.) *Cognitive Neuroscience*, pp. 197–242. Cambridge: MIT Press.

O'Reilly RC, Noelle DC, Braver TS, and Cohen JD (2002) Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex* 12: 246–257.

Sakai K and Passingham R (2006) Prefrontal set activity predicts rule-specific neural processing during subsequent cognitive performance. *Journal of Neuroscience* 26: 1211–1218.

Sakai K and Passingham RE (2003) Prefrontal interactions reflect future task operations. *Nature Neuroscience* 6: 75–81.

Executive Functions: Eye Movements and Neuropsychiatric Disorders

A B Sereno, S L Babin, A J Hood, and C B Jeter, The University of Texas Health Science Center at Houston, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

What Are Executive Functions?

Executive functions are complex, higher order processes moderated primarily by the frontal lobe, specifically the prefrontal cortex. The complexity of executive function has made a universally accepted definition elusive, but attention (focusing on relevant information and ignoring distractors), working memory (maintaining information until execution), and motor planning (goal-directed motor planning and programming) are generally considered key executive functions. Executive functions influence social, emotional, intellectual, and organizational aspects of one's life. Life can be devastated when executive functions are disrupted. Unfortunately, such disruptions occur in many human disorders.

Eye Movements and Executive Functions

Eye movements are any shift of position of the eye in its orbit. There are many different kinds of eye movements, which are defined in the next section titled 'Classes of eye movements.' Eye movements determine what information reaches our retina, visual cortex, and most important, higher cortical centers. Hence, eye movements are critically important for vision, attention, and memory; they determine what we see, attend to, and remember about our surroundings. They are thus central to executive functions. Much scientific work also suggests that the brain circuitry subserving certain executive functions, such as that for spatial attention and spatial working memory, overlaps parts of the brain circuitry subserving control of eye movements. Thus, better understanding of eye movements provides a valuable window into executive functions.

Classes of Eye Movements

There are several classes of eye movements, with, for the most part, distinct neural circuitry. Eye movements serve to stabilize images on the retina and to keep objects of interest on the fovea, the retinal area that has the greatest visual acuity and the greatest representation in visual cortices. Without compensatory eye movements during self-motion (locomotion) or head movement, images of the visual world would blur and slip across the retina with each movement. Two classes of eye movements, vestibulo-ocular and optokinetic reflexes, evolved to stabilize images on the retina during such head and body perturbations. In the case of the vestibulo-ocular reflex, when the head moves to the right, 'vestibular' sensors in the inner ear are triggered and signal nuclei in the horizontal and vertical gaze centers of the brainstem to create an equal and opposite eye velocity signal. The result is that the vestibulo-ocular reflex initiates a rapid eye movement to the left (within about 15 ms) to compensate for the head movement, keeping the retina stable with respect to the world. The second mechanism, the optokinetic reflex, complements the first by using full-field 'visual' information about self-movement to compensate for the perturbation of the visual world on the retina. That is, as a result of self-motion to the right, the entire visual image will move across the retina in a leftward direction. The speed of the image drift on the retina triggers nuclei in the brainstem to create an equal leftward eye velocity signal. The result is the optokinetic reflex, an eye movement in the same direction as the retinal slip to keep the retina stable with respect to the world.

A third class of eye movements serves to rapidly shift gaze (up to $900^{\circ}s^{-1}$) to bring a new object of interest to the fovea. These saccades are often further subdivided into two general subclasses: reflexive and voluntary. Reflexive saccades are elicited in response to a sudden movement, flash, or change in the environment. The latency (onset of movement) of a reflexive saccade to a visual target is typically about 200–250 ms, but it can be as short as 70–80 ms. Voluntary saccades are more complex, involve executive functions, are thought to critically involve frontal cortex, and have latencies as long as 400 ms. In real life, for instance, the same environment can result in different voluntarily controlled saccadic eye movements, depending on what information is sought or needs to be remembered. This was elegantly demonstrated by Yarbus in 1967 when he presented the same picture to participants several times but each time asked a different question. Depending on the question, the participants scanned the picture with a strikingly different set of voluntary saccades to acquire the pertinent information. In a similar fashion, Figure 1 illustrates an example of the scanpaths of a naïve participant viewing a Vermeer painting.

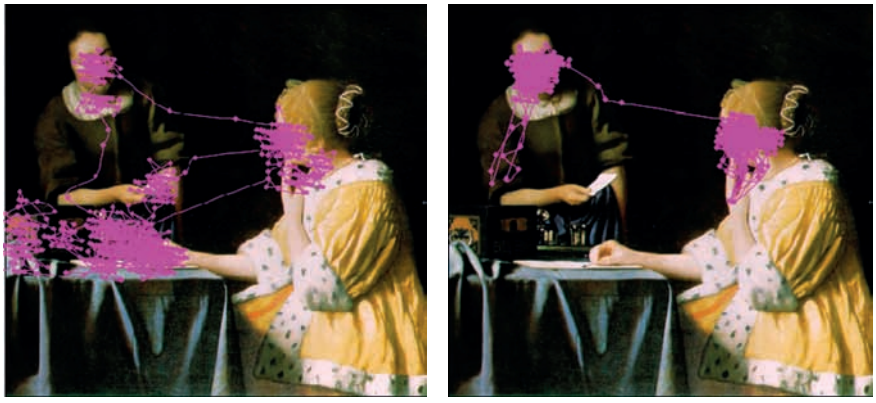

Figure 1 Scanpaths of a naïve individual viewing Vermeer's *Mistress and Maid*. This person was told that two pictures would be presented sequentially for 5 s each and that before each picture, a question would be asked. The person was told to look at the picture for the answer and reply after the viewing finished. (Left) The first question asked was, "What are they doing?" The picture was then presented, and the person's scanpath was recorded for 5 s. The individual then replied, "delivering and receiving a letter." (Right) The second question asked was, "How old are they?" The same picture was presented again and the scanpath recorded for 5 s. The person then replied "35." The scanpaths for the same picture were very different based simply on the context of what the individual was asked.

The first panel illustrates the scanpath when the picture was preceded with the question, "What are they doing?" The participant made saccades that followed the gaze of the two women in the picture and then dwelled on the hands, the table, and the objects on the table. The second panel illustrates the scanpath when the same picture was preceded with the question, "How old are they?" In this context, the participant looked almost exclusively at the women's faces.

In the research lab or clinic, a prosaccade task is frequently used to elicit reflexive saccades (see Figure 2). In this paradigm, the participant fixates a center spot, a visual target appears to the left or right of center, and the participant is instructed to look at the target. Voluntary saccades can be elicited by many different eye movement paradigms (see also Figure 2). In an antisaccade task, a peripheral target is presented (as in the prosaccade task), but the correct response is to look to the opposite side of the screen, away from the target. The antisaccade task requires a person to (1) inhibit the reflexive inclination to look at the target and (2) generate a voluntary saccade to the correct location opposite the target. Abnormal responses in the antisaccade task include increased latency and directional errors (looking to the target). A delayed saccade task also requires a voluntary saccade. This task is similar to a prosaccade task in that the person is required to make an eye movement to the target. However, the person is also instructed to wait until a go signal is given (a tone or change in fixation cue). Therefore, the person must inhibit the reflexive tendency to look at the target immediately and generate a voluntary saccade when the go signal is presented. A third commonly used voluntary saccade paradigm is a memory-guided saccade task. This task is identical to the delayed saccade task, except that the target is removed from the screen during the delay period. Now, the person must inhibit the eye movement until the go signal and also remember where the target was presented. Hence, this task also demands working memory.

Smooth pursuit eye movements (SPEMs) function to keep a small moving stimulus on the fovea and are typically much slower than saccades, with a maximum velocity of about $100\,^{\circ}\mathrm{s}^{-1}$. Smooth pursuit is a voluntary task requiring both motivation and attention. Hence, if one chooses to track a mosquito, when the mosquito takes off, one's eyes would initiate smooth pursuit 100–150 ms later, an onset of movement that is generally shorter than that for a saccade. Thus, the eyes would first begin to pursue but lag a bit behind the mosquito. After about another 100 ms, this is remedied by a 'catch-up saccade' that repositions the moving eyes directly on the mosquito. SPEMs are the result of a complex transformation from visual motion signals at the fovea to an oculomotor signal. Cortical input to the smooth eye movement generation centers in the brainstem is required to enable smooth tracking motion. At the same time, however, there must be suppression of the optokinetic reflex and, if the eye movement is accompanied by head tracking, the vestibulo-ocular reflex. SPEM paradigms are used frequently to measure executive function in clinical populations. One measure of accuracy in SPEM paradigms is pursuit gain (the ratio between eye velocity and target velocity). Gain values below 1.0 indicate that the pursuit system is not matching the velocity of the target. Errors in gain can be amended with catch-up saccades (when

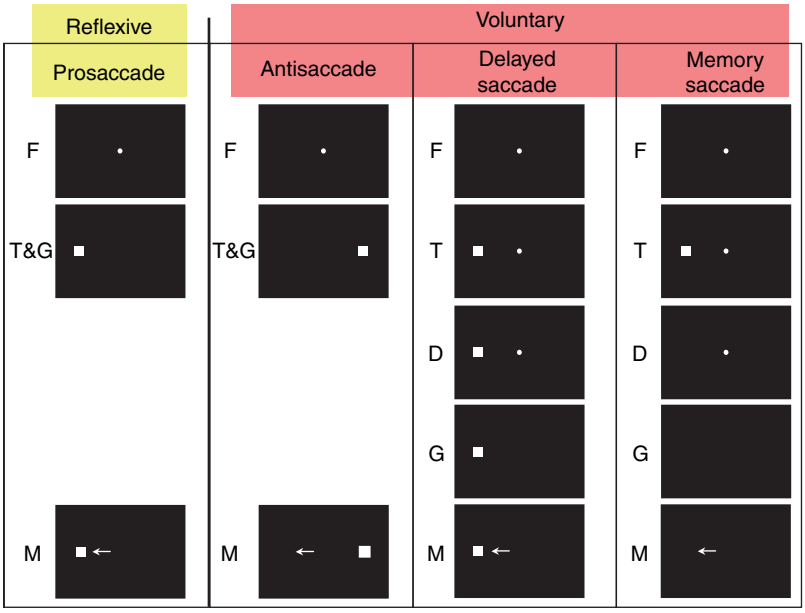

Figure 2 Examples of four saccade tasks. All tasks are presented sequentially from top to bottom, and the arrows indicate the direction of the correct eye movement. Prosaccade: This is a reflexive task in which an individual begins by fixating the center point (F). Once fixation is maintained, a target light appears, and the person makes an eye movement to the light (T&G). Antisaccade: An individual begins by fixating the center point (F). The target light then appears (T&G), and the person is to make an eye movement to the equal and opposite position of the target. Delayed Saccade: The individual begins by fixating the center point (F). Next, the target light appears (T), followed by a delay period (D), during which the person must continue fixating the center point. At some variable time, the fixation light disappears, which serves as a go signal (G) for the person to begin an eye movement to the target. Memory Saccade: The individual fixates the center point (F). The target light then flashes on the screen (T), followed by a delay period (D), during which the individual must continue fixating the center point. At some variable time, the fixation light disappears, which serves as the go signal (G) for the person to begin an eye movement to the remembered location where the target had previously flashed. F, fixation screen; T, target onset; G, go signal to begin an eye movement; D, delay period, when the participant must wait for the go signal; M, eye movement.

smooth pursuit lags behind the target) and backup saccades (when tracking gets ahead of the target). Further, there can be intrusive saccades, which interrupt smooth pursuit tracking and are indicative of inhibitory failures or inappropriate activations.

Vergence is the only disconjugate eye movement; that is, the eyes move simultaneously in opposite directions. Vergence movements occur reflexively in order to focus and reduce disparity between the locations of an image on the retina of each eye.

Eye Movements in Human Disorders

Voluntary eye movements as a general class can be used to tease out specific deficits in different aspects of executive function in clinical populations. By careful comparison of performance on voluntary eye movement tasks, a 'deficit in executive function' (as generally classified, for example, by a neuropsychiatric test) can be broken down into separable aspects of executive function, including the voluntary planning, generation, and accuracy of a response, the ability to inhibit or control impulse responding, and working memory. Voluntary eye movements have been linked to different anatomical circuitry and can be used to characterize subtypes within some patient populations.

Neurological

Tourette's syndrome and Parkinson's disease are movement disorders involving the basal ganglia. Both disorders show deficits in executive function with evidence of disrupted prefrontal–basal ganglia circuits. Voluntary eye movement deficits are consistent with these documented impaired executive functions. Both groups show increased antisaccade latency and antisaccade errors. The effect of medication on voluntary eye movements in Tourette's syndrome is not clearly known. On the other hand, dopaminergic medications, routinely given to treat Parkinson's disease, do improve voluntary eye movement performance. Further, voluntary saccade performance differentiates between subtypes of Parkinson's disease patients such that akinetic-rigid Parkinson's disease patients have more difficulty on voluntary eye movement tasks than do tremor-dominant patients.

Developmental

Autism is a developmental behavioral disorder marked by abnormal language and social communication resulting from deficits in limbic and cortical structures. Numerous studies have investigated voluntary eye movement performance in children and adults with autism as a means of understanding the etiology and social implications of this disorder. For instance, voluntary eye movements (saccades and SPEMs) have been used to show that executive function abilities are delayed in autistic patients; their voluntary abilities seem to mature to a lower level and at a slower rate. Likewise, results from voluntary saccade tasks suggest that deficits in autistic individuals critically involve the prefrontal cortex.

Psychiatric

Eye movements have been studied most extensively in schizophrenia, as is discussed in the next section. Attention-deficit/hyperactivity disorder (ADHD) is characterized by symptoms (sometimes beginning before age 7) of inattention, hyperactivity, or impulsivity that cause impairment at school, home, or work. ADHD patients have been shown to be impaired on a countermanding task (similar to a delayed saccade task, but a stop signal may be presented on some trials to indicate the response should be withheld), which suggests that their impairment mainly consists of increased impulsivity. In addition, voluntary eye movement deficits have shown subtype differences within the ADHD population. Specifically, patients classified as ADHD-Inattentive have better motor planning and are less impulsive than ADHD patients with both inattentive and impulsive characteristics. Interestingly, methylphenidate improves the voluntary eye movement performance of both ADHD subtypes. Bipolar disorder is a disturbance in mood marked by cycling periods of mania and/or depression. Bipolar patients have been shown to have an inhibition problem but to date no deficit in working memory on voluntary eye movement paradigms. Bipolar patients, as well as lithium-withdrawn manic patients, have been shown to perform more poorly than normal controls on a SPEM task.

Eye Movement Deficits in Schizophrenia

What Is Schizophrenia?

Schizophrenia is the most prevalent mental illness in the world, occurring in about 1% of the population. It is also considered the most costly and debilitating mental disorder and is among the top causes of disability in developing countries. In 2002, it cost the United States approximately $\$65 \times 10^9$, and this estimate accounted only for patient hospitalizations, loss of work, and cost of disability. Schizophrenia is a psychiatric disorder that is usually characterized by hallucinations, delusions, blunted affect, alogia (i.e., poverty of speech), and avolition. Numerous cognitive functions in areas such as executive control, working memory, and attention are disturbed. These disrupted executive functions are more persistent than psychotic symptoms. More robust measures of cognitive deficits are crucial, given that there is a strong positive correlation between improved cognitive functioning and long-term prognosis. Classically, neuropsychological tests have been used to investigate cognitive dysfunctions in schizophrenia, but recently, eye movement tasks have been shown to be a more sensitive and reliable measure of these deficits. Eye movements are also more easily mapped to specific neural structures and hence may be a more specific biological marker for differentiating subtypes. For these reasons, the study of eye movements is an important research tool that will shed light on both executive functions and debilitating human disorders such as schizophrenia.

Eye Movement Deficits

Nearly a hundred years ago (1908), using a photographic technique, Diefendorf and Dodge were the first to report that SPEMs were impaired in schizophrenia. They found that pursuit was commonly interrupted by saccades. With the advent of new techniques for measuring eye movements, as well as renewed interest, much is now known about SPEM deficits in schizophrenia.

There seem to be two basic deficits in the SPEMs of people with schizophrenia: (1) reduced gain of smooth pursuit and (2) increased saccadic events. Pursuit gain is a measure of the accuracy of the smooth eye movement, and people with schizophrenia show reduced gain, that is, their eyes do not keep up with the target, and they often fall behind. As a result, schizophrenic patients often exhibit increased saccadic activity (catch-up saccades) during smooth pursuit. These saccades are automatic and compensatory in nature and serve to quickly refoveate the target on the retina. Other increased saccadic events observed in schizophrenia are saccadic intrusions. These eye movements are not compensatory and are thus inappropriate. Common examples of saccadic intrusions are square-wave jerks and anticipatory saccades. A square-wave jerk is a small pair of saccades whose initial direction takes the eyes off the target. After a 100–250 ms intersaccadic interval, during which pursuit continues, a second saccade returns the eyes to the target. Anticipatory saccades overshoot the location of the target. These saccades have

larger amplitudes than those of square-wave jerks and longer intersaccadic intervals, and they do not maintain continuous pursuit during the intersaccadic period. Both these types of intrusive saccadic eye movements are rarely found in young controls (although they are sometimes found in older people) and indicate inappropriate activation or lack of inhibition of the saccadic system during smooth pursuit.

In addition to smoothpursuit deficits, a number of saccadic eye movement dysfunctions have been identified in schizophrenia. Numerous studies have shown deficits in various voluntary saccade tasks, including delayed, remembered, or predictive saccades. In an antisaccade task, schizophrenic participants typically are slower to generate a saccade to the correct location and make more errors (i.e., eye movements to the target) compared with normal controls. In contrast, many studies have demonstrated that schizophrenic participants perform normally or in some cases faster than controls on a prosaccade task. This pattern of saccade task performance has in some studies been related to the participant's performance on SPEMs. That is, schizophrenic patients with smooth pursuit deficits make significantly more errors on the antisaccade task than schizophrenic patients without eye tracking deficits.

Biological Basis of Eye Movement Deficits

Schizophrenic patients with SPEM deficits do not show abnormal slow phases of vestibulo-ocular and optokinetic responses. This dissociation suggests that the smooth pursuit abnormalities in schizophrenia are due to cortical deficits rather than deficits in common brainstem motor output pathways. Evidence for a prefrontal cortex dysfunction model of schizophrenia has also gained support. One key piece of evidence for such a model came from the similarity between symptoms of blunted affect, avolition, and inattention seen in schizophrenic patients and symptoms exhibited by neurological patients with a loss of frontal lobe function. A further piece of evidence in support of frontal lobe dysfunction in schizophrenia came from eye movement research. It was already well established that frontal regions, especially the frontal eye fields, play a role in saccades. Further research also demonstrated that lesions in the frontal eye fields disrupt SPEMs. Hence, frontal dysfunction could account for both pursuit deficits and saccadic deficits.

It is interesting to note that between 50% and 85% of individuals with schizophrenia exhibit SPEM dysfunction, in contrast to only about 8% of normal controls. Similarly, about 60% of individuals with schizophrenia exhibit antisaccade eye movement deficits. People with schizophrenia are heterogeneous in terms of both smooth pursuit deficits and voluntary saccade deficits. This division of impaired and nonimpaired patients (for either pursuit gain or voluntary saccade performance) may prove to be more useful than the classic clinical subtypes in characterizing homogeneous subgroups of patients.

Genetics of Eye Movement Deficits

In the late 1970s and early 1980s, a series of studies investigated the genetics of smooth pursuit deficits in monozygotic and dizygotic twins who were discordant for schizophrenia. These studies found that a larger proportion of monozygotic twins than dizygotic twins exhibited abnormal smooth pursuit eye tracking. Moreover, about half of clinically well first-degree relatives of schizophrenic patients exhibit SPEM deficits, a prevalence much greater than recurrence risk for schizophrenia in first-degree relatives, which is about 6.5%. In addition, some recent reports suggest that 25–50% of nonsymptomatic first-degree relatives of schizophrenic patients exhibit antisaccade deficits. Within the last several years, linkage has been reported for both smooth pursuit deficits and anti-saccade deficits in genetic studies of schizophrenia. Endophenotypes such as eye movement dysfunctions may be a more penetrant expression of a schizophrenia susceptibility gene than schizophrenia.

Importance of Subtypes and Genetics for Outcome

Eye movement monitoring is an important research tool because it affords the opportunity to use a quick, easy, and noninvasive technique to help identify behaviors that may indicate susceptibility for schizophrenia. Since it is possible that people with schizophrenia who exhibit eye movement deficits are genetically differentiable from those who do not, eye tracking may provide more reliable and robust criteria for the classification of different groups of people with schizophrenia. Such a biologically based classification system may allow us to investigate particular antipsychotic medication effects across different subgroups of schizophrenic patients. For example, it is possible that task-impaired patients will show a greater benefit from a specific medication. Such a finding has been reported with nicotine and schizophrenia: patients impaired on an antisaccade task showed a greater benefit from nicotine than did nonimpaired patients and controls. A similar subtype-specific medication effect has also been demonstrated in Parkinson's disease. It was found that akinetic-rigid patients are poorer than tremor-dominant patients

on voluntary eye movement tasks, such as the antisaccade task, and that only tremor-dominant patients were able to significantly improve their performance on this task after dopaminergic medications were administered. A better understanding of subtype-specific medication effects would be pivotal in assigning the most effective medication to the patients who would benefit the most.

In summary, eye movements provide invaluable information about executive functions and provide an efficient, noninvasive way to study the role of abnormal brain processes in a variety of human disorders. Their study may afford the development of more effective and efficient treatments of executive dysfunction.

See also: Attention Deficit Hyperactivity Disorder; Attention and Eye Movements; Attention: Models; Autism; Brainstem Control of Eye Movements; Cognitive Deficits in Schizophrenia; Cortical Control of Eye Movements; Eye Tracking and Mental Illness; Eye and Head Movements; Eye Movement Disorders; Frontal Eye Fields; Prefrontal Cortex: Structure and Anatomy; Psychophysics of Attention; Saccades and Visual Search; Saccadic Eye Movements; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Visual Attention.

Further Reading

Büttner U and Büttner-Ennever JA (1988) Present concepts of oculomotor organization. *Reviews of Oculomotor Research* 2: 3–32.

Calkins ME and Iacono WG (2000) Eye movement dysfunction in schizophrenia: A heritable characteristic for enhancing phenotype definition. *American Journal of Medical Genetics* 97: 72–76.

Clementz BA and Sweeney JA (1990) Is eye movement dysfunction a biological marker for schizophrenia? A methodological review. *Psychology Bulletin* 108: 77–92.

Hanisch C, Radach R, Holtkamp K, Herpertz-Dahlmann B, and Konrad K (2006) Oculomotor inhibition in children with and without attention-deficit hyperactivity disorder (ADHD). *Journal of Neural Transmission* 113(5): 671–684.

Klein C, Fischer B, and Hartnegg K (2002) Effects of methylphenidate on saccadic responses in patients with ADHD. *Experimental Brain Research* 145: 121–125.

Levin S (1984) Frontal lobe dysfunctions in schizophrenia. II. Impairments of psychological and brain functions. *Journal of Psychiatric Research* 18: 57–72.

Levy DL, Holzman PS, Matthysse S, and Mendell NR (1993) Eye tracking dysfunction and schizophrenia: A critical perspective. *Schizophrenia Bulletin* 19: 461–536.

Reuter B and Kathmann N (2004) Using saccade tasks as a tool to analyze executive dysfunctions in schizophrenia. *Acta Psychologica* 115: 255–269.

Schiess MC, Zheng H, Soukup VM, Bonnen JG, and Nauta HJ (2000) Parkinson's disease subtypes: Clinical classification and ventricular cerebrospinal fluid analysis. *Parkinsonism and Related Disorders* 6: 69–76.

Sereno AB (1996) Parsing cognitive processes: Psychopathological and neurophysiological constraints. In: Matthysse S, Levy DL, Kagan J, and Benes FM (eds.) *Psychopathology: The Evolving Science of Mental Disorder*, pp. 407–432. New York: Cambridge University Press.

Sweeney JA, Takarae Y, Macmillan C, Luna B, and Minshew NJ (2004) Eye movements in neurodevelopmental disorders. *Current Opinions in Neurology* 17: 37–42.

Exercise in Neurodegenerative Disease and Stroke

C W Cotman, University of California, Irvine, CA, USA
A D Smith, University of Pittsburgh, Pittsburgh, PA, USA
T Schallert, University of Texas at Austin, Austin, TX, USA
M J Zigmond, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

It has long been known that exercise is essential to good health. Individuals who lead a sedentary lifestyle greatly increase their risk of many diseases, including osteoporosis, cardiovascular disease, and diabetes. During the past decade, it has become apparent that physical activity also is essential to a healthy brain, particularly in aging populations. In addition to its generally beneficial effects, there is reason to believe that exercise and other forms of motor enrichment can be beneficial with regard to several neurological disorders, including Parkinson's disease (PD) and Alzheimer's disease (AD), and can even be helpful in reducing the deficits induced by stroke.

PD and AD have much in common. They involve a progressive degeneration of neurons in the brain that often goes undetected for many years. Because of this long preclinical gestation and perhaps because of a reduction in the capacity of the brain to defend itself against insults, the emergence of these conditions as clinical entities increases sharply as individuals age, with the great majority of cases emerging in people aged 60 and older. In addition, whereas their pathological and clinical manifestations are quite different at the outset, there often is an increasing overlap as PD and AD progress.

This article briefly reviews the evidence for the neuroprotective effects of exercise, discusses supportive data from animal and cellular models, and then comments on possible underlying mechanisms and implications for the causes and treatments of these neurodegenerative diseases, as well as for the treatment of stroke.

PD

The Condition

Originally identified as a clinical syndrome by James Parkinson in 1817, PD is typically characterized by its motor signs – slowness of movement (bradykinesia), stooped posture, rigidity, and tremor. In reality, however, these appear to reflect an intermediate sign of the underlying neurodegenerative process that may also affect cardiovascular function, olfactory sensitivity, speech, fatigue, and gastrointestinal function and may eventually be accompanied by cognitive dysfunction, including dementia. The prevalence of PD is about 1%. Diagnosis typically occurs at about age 60, though the disease probably has a 10- to 20-year 'preclinical' phase, during which symptoms are either not detected or not associated with PD. On the other hand, 5–10% of patients with PD are diagnosed before the age of 40. Men, particularly younger individuals, are at a slightly greater risk for the condition.

In the early 1960s, the majority of the motor deficits in PD were attributed by Oleh Hornykiewicz and his colleagues to the reduction in dopamine (DA) in the caudate and putamen (collectively referred to as the striatum). Subsequently it was shown that the DA in these regions was located in the terminals of neurons that project from the substantia nigra pars compacta (SNpc), a group of about half a million pigmented neurons per side located in the ventral mesencephalon. By the time PD is diagnosed, about half of the SNpc neurons have been lost, along with about 80% of the DA in the striatum. However, as with the symptoms of PD, the neurodegeneration underlying the condition is quite broad and eventually may include the loss of many other specific groups of neurons both caudal and rostral to the SNpc, as has been suggested most convincingly by Heiko Braak, beginning in the late 1980s. It seems likely that many of the nonmotor consequences of PD result at least in part from pathology in these extradopaminergic regions.

Clinical Evidence for the Beneficial Effects of Exercise in PD

Many drugs are used in the treatment of PD; however, virtually all of them strive directly or indirectly to replace the impact of the DA that is lost as part of the disease. Thus, current treatments target motor symptoms rather than the full complex of the disease. These medications can be highly effective for many years and usually permit patients with PD to have a normal life span. However, there is little evidence that the actual neurodegenerative process is affected by such treatments. This is also the case for surgical treatments, which have included pallidotomy and the now more common approach of deep brain stimulation of the subthalamic nucleus.

In contrast to more conventional therapies, physical exercise appears to reduce the incidence of PD and to improve the motor, cognitive, and affective deficits in patients with mild to moderate PD. A lifetime of exercise seems to reduce the incidence of PD, at least

in men. In addition, a program of aerobic activity has been reported to improve movement initiation, whereas resistance and balance training have been found to improve balance.

Despite the encouragement such studies provide, virtually all of them suffer from one of two failings. Those studies examining long-term effects of exercise derive their results from surveys. For example, a negative correlation between PD and the level of physical activity throughout life has been reported. However, whether these data are collected retrospectively or over a long period through regular questionnaires, such studies always suffer from a lack of precision. Moreover, in the case of PD, they leave open the possibility that the results reflect a tendency toward inactivity of individuals who are in the preclinical stages of the disorder.

A few studies have tried to address this problem with prospective studies, separating the patients into an exercise and a control group. However, here, too, there are serious problems. Such studies are typically very brief (weeks or at best a few months), involve relatively few participants, may not provide adequate controls for the effect of the increased attention that the exercise group receives, and fail to distinguish between symptomatic changes and an actual influence on underlying disease progression. Of course, these problems hold not only for studies of PD but also for studies of AD, described in the section titled 'AD.' Thus, although such data are encouraging, they should be viewed with caution, and for the present, studies of experimental animals form the basis for assessing the neurobiological benefits of this therapy.

Animal Studies

Although animal models of neurodegenerative diseases are of necessity imperfect, several models of PD are in use. Most rely on the fact that a major component of PD pathology is the loss of DA neurons and that these neurons can be selectively destroyed in many animals through the use of DA-specific neurotoxins. The first such toxin to be introduced was 6-hydroxydopamine (6-OHDA), which enters via DA uptake sites and, when combined with the noradrenaline uptake blocker desipramine, is selectively concentrated in DA neurons, causing toxic oxidative stress. Later, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) was introduced as a tool in the study of PD. MPTP is converted to the DA neurotoxin MPP+, which, like 6-OHDA, is concentrated in DA neurons, where it inhibits complex I of the mitochondria.

When the toxin-induced loss of DA is severe, rodents and nonhuman primates lesioned in this way show several motor deficits that are reminiscent of PD. For example, rats treated with 6-OHDA become bradykinetic or even akinetic, whereas monkeys treated with MPTP provide an even closer model of the motor symptoms of PD, exhibiting tremor, stooped posture, and rigidity.

The protective effect of constraint therapy was first used with some success in patients with stroke (see the section titled 'Stroke') and has recently been applied to 6-OHDA-treated rats as well. Adult rats receiving unilateral injections of 6-OHDA had their 'normal,' ipsilateral limb constrained by a plaster-of-paris cast for 7 days. When assessed for behavioral and neurodegenerative abnormalities several weeks later, these deficits were greatly attenuated. This was true whether the forced exercise occurred immediately before or immediately after the administration of 6-OHDA. However, if the placement of the cast on the forelimb was delayed for a week after 6-OHDA, the therapy was not effective (**Figure 1**). Moreover, if animals previously protected against 6-OHDA by limb immobilization had their 'protected' limb constrained in a cast for a 1 week period 2–3 weeks later, behavioral and neurochemical evidence of DA loss reemerged. These results indicate that there is a brief window of opportunity during which forced exercise is protective. They also suggest that continued limb use is essential to maintaining the DA neurons in a viable state.

Housing in a more complex environment or caloric restriction can also protect DA neurons. For example, the loss of DA neurons incurred from MPTP as assessed by cells immunoreactive for the rate-limiting enzyme tyrosine hydroxylase (TH) was significantly reduced by providing mice with daily access to tunnels, opportunities for climbing, running wheels, and toys for 2 months (**Figure 1**). These results have since been replicated by Dr. Richard Smeyne and colleagues by giving mice continuous access to a running wheel, suggesting that a major impact of environmental complexity is to increase exercise. Likewise, reducing caloric intake by 30–40% in young mice for 3 months prior to MPTP administration improved motor performance and caused a near-complete protection of DA cells and terminals. Similar results of caloric restriction have been obtained in a hemiparkinsonian primate MPTP model and in other brain-injury models.

The protective effects of complex environments and caloric restriction may be due at least in part to the increase in motor activity that is associated with these treatments. At least one report has indicated that the protective effects of environmental enrichment in PD models could be duplicated by housing animals with a running wheel. Moreover, running has been clearly linked to neuroprotection in studies of mouse models of AD and of stroke.

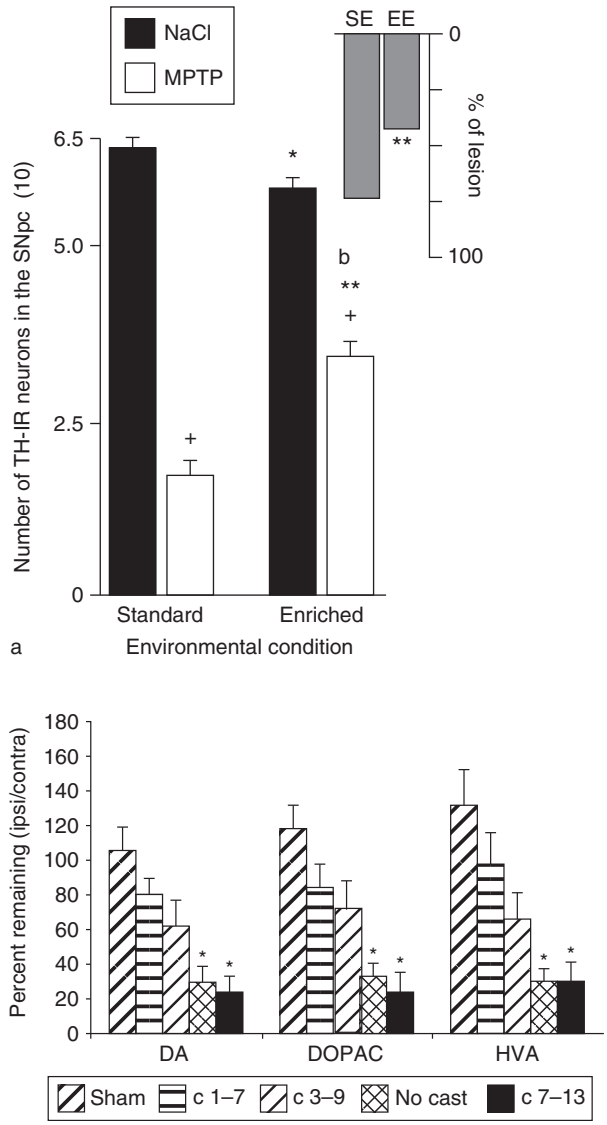

Figure 1 Effect of exercise on the response of dopamine (DA) neurons to neurotoxins in models of Parkinson's disease (PD). Upper panel: At weaning, mice were placed in either an enriched or a standard environment and examined 2 months later. Shown are the number of tyrosine hydroxylase immunoreactive (TH-IR) neurons in the substantia nigra pars compacta (SNpc) of mice from the two groups 7 days after treatment with either saline (NaCl; dark bars) or 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP; 20 mg kg^{-1} intraperitoneally; open bars). From Bezard E, Dovero S, Belin D, et al. (2003) Enriched environment confers resistance to 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine and cocaine: Involvement of dopamine transporter and trophic factors. *Journal of Neuroscience* 23: 10999–11007. Lower panel: Adult male rats were given a unilateral injection of 6-hydroxydopamine (6-OHDA; 10 μg) along the medial forebrain bundle. In some groups of rats, a unilateral cast was used to restrict the use of the ipsilateral forelimb for 7 days, thereby forcing the animals to rely on their contralateral limb for movement during this time. The cast was placed immediately after the 6-OHDA (c 1–7), 3 days later (c 3–9), or 7 days later (c 7–13). The animals were sacrificed 65–80 days later for analysis of DA, DOPAC, or HVA levels by high-performance liquid chromatography. Animals receiving early casts and sham animals did not show significant differences in DA, DOPAC, or HVA levels; animals receiving casts on days 3–9 showed intermediate DA levels, although still not significantly different from sham. Animals with late casts (days 7–13) and animals not receiving casts showed significantly lower DA levels when compared with sham-treated animals. These data parallel performance on a variety of behavioral tasks (data not shown). From Tillerson JL, Cohen AD, Philhower J, Miller GW, Zigmond MJ, and Schallert T (2001) Forced limb-use effects on the behavioral and neurochemical effects of 6-hydroxydopamine. *Journal of Neuroscience* 21: 4427–4435.

AD

The Condition

AD is the most common form of dementia. Like PD, it is a progressive, degenerative disorder that has an extended preclinical phase and becomes prominent with age. Whereas the prevalence rate of AD is only slightly higher than that of PD (~1.5%), about half of those who reach the age of 85 will be affected. Moreover, although patients with PD have a near-normal life span if properly treated, patients with AD seldom live for more than a decade beyond diagnosis. The deficits typically begin with memory loss and proceed to affect thinking and language. The condition has been associated with neuronal loss, particularly among subcortical neurons that project to the telencephalon. By the time that AD is identified, this loss is usually sufficient to have caused a gross reduction in brain size. At a cellular level, two abnormalities can be detected and are diagnostic for the condition: neuritic plaques containing amyloid-β (Aβ) and neurofibrillary tangles composed largely of tau protein. The condition and its neuropathological hallmarks were first described in 1906 by Alois Alzheimer.

Areas of the brain that are affected by AD are widespread. However, one group of neurons that often is targeted early is the cholinergic neurons of the basal nucleus of Meynert. This, in turn, has led to the use of inhibitors of acetylcholine esterase in the treatment of the condition, an approach that has limited, symptomatic benefit at best, with no evidence for any effect on the neurodegenerative process. As in the case of PD, many drugs are under development that are designed to influence the underlying pathophysiology of the disease, but they are still in early experimental stages. However, again as with PD, there is reason to believe that exercise can help hold AD at bay or slow its progression.

Clinical Findings on Efficacy of Exercise in AD

Evidence suggests that the incidence and progression of cognitive decline in later life are reduced by exercise. For example, individuals who engaged in strenuous walking three times a week showed improved executive function and increased brain activation,

as monitored by functional magnetic resonance imaging, compared with those who undertook stretching and nonaerobic activity. However, as in the case of PD, the field still awaits long-term clinical trials with enough power to permit definitive conclusions about exercise and cognition, as well as clear statements regarding the amounts and types of exercise that are most effective. Thus, we must again turn to animal studies.

Animal Studies

The first requirement for an animal study is the establishment of a model. As noted above, the classic hallmarks of AD are accumulation of Aβ and the development of neurofibrillary tangles. Transgenic mice with mutations in the amyloid precursor protein accumulate Aβ. This accumulation is correlated with cognitive impairment and a reduction in indications of neuroplasticity, including learning. Conversely, reductions in Aβ by pharmacological intervention improve cognitive function and synaptic plasticity. These data suggest that Aβ is causally linked to the cognitive difficulties present in these mutant, transgenic mice and that reducing Aβ is important for healthy brain aging.

Housing in a complex environment can decrease Aβ accumulation (Figure 2) and improve cognitive performance in mouse models of AD. However, at least one study indicates that these effects cannot always be duplicated by voluntary running alone and thus may be associated with more cognitive aspects of such enrichment. This would not be surprising given the clinical literature linking cognitive activity to a reduction in the incidence and progression of dementia.

Stroke

A stroke commonly results from either blockage (ischemic) or bursting (hemorrhagic) of an artery supplying blood to a region of the brain, with the former accounting for about 85% of all incidents. In major industrialized countries, stroke is the third leading cause of death, after heart disease and cancer, and the major cause of adult disability. Unlike the slow neurodegenerative process associated with PD and AD, hemorrhagic or ischemic injury is usually immediately severe, leading to total or near-total cell death within affected brain tissue. However, this primary injury is typically followed by slow secondary degeneration around the damage (termed the penumbra) as well as in remote regions, and this secondary degeneration can continue to develop for months. It is this component of the brain damage that is the main target for treatment.

Pharmacological treatment of stroke is limited. In the case of ischemic stroke, tissue plasminogen activator, a clot-dissolving drug, may be useful if given within the first 2–3 h of the incident. Glutamate antagonists have been examined extensively in stroke treatment, the rationale being that excitotoxicity appears to play a major role in the associated cell death. Unfortunately, although such agents appeared very promising in studies with animal models, they have not proven effective in clinical trials, in part because of their behavioral side effects.

Exercise

Physical therapy has been used in the treatment of stroke for decades. The concept of constraint therapy

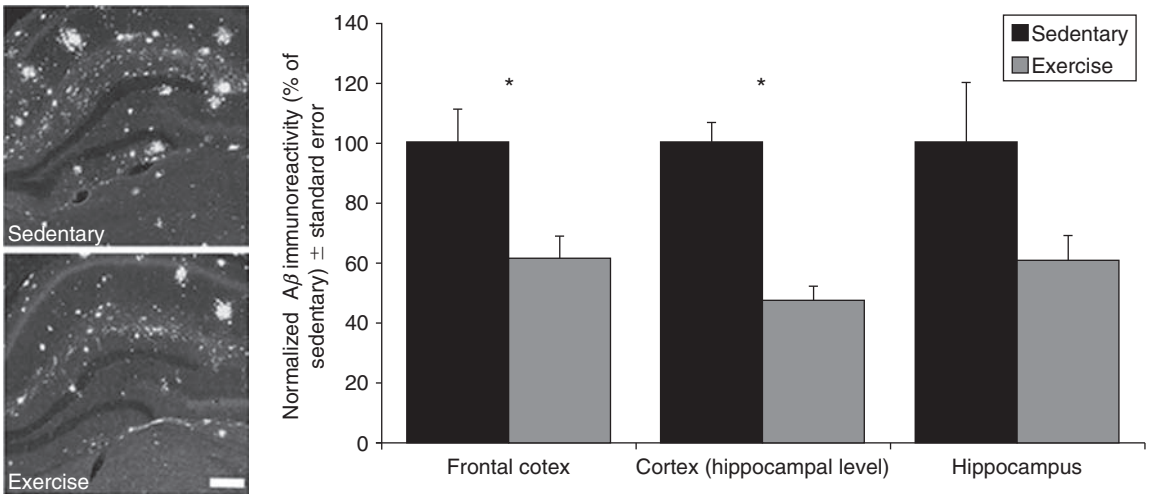

Figure 2 Effect of exercise on amyloid-β (Aβ) load in three different brain regions of transgenic mice that normally overexpress this protein (TgCRND8). Exercising animals showed a reduction in Aβ immunoreactivity in cortical regions and in hippocampus. The photomicrograph shows a representative section from both groups. Scale bar = 200 μm. From Adlard PA, Perreau VM, Pop V, and Cotman CW (2005) Voluntary exercise decreases amyloid load in a transgenic model of Alzheimer's disease. *Journal of Neuroscience* 25: 4217–4221.

in the treatment of motor deficits in stroke victims was introduced in the early 1980s, largely because of the work of Edward Taub. The concept was to increase the use of the affected limb by constraining the use of the normal limb for several hours each day for up to 2 weeks. Some restoration of function was observed, and the approach is now often used in patients with at least some capacity to move the stroke-impaired wrist or digits of the upper extremity, presumably in part because the injury to the corresponding brain region is partial. Other types of exercise have included weight-supported and unsupported treadmill and cycling. There is no evidence that the primary injury can be reduced with poststroke exercise. However, whereas the initial effects of severe ischemic injury may not be responsive to exercise, compensatory behavioral adaptations can be promoted with constraint and other forms of motor-enrichment therapies. These treatments presumably lead to attenuation of slow secondary degeneration (though see discussion below) and promote adaptive rewiring.

Exercise of the lower extremities in human stroke patients, including facilitated treadmill walking, can improve walking ability, in contrast to stretching and other control procedures. The improvement has been linked to behavior-related induction of more normal patterns of activation in cortical and subcortical motor regions. There is some evidence that poststroke training can ameliorate speech deficits, particularly when damage to speech areas in the cortex is subtotal. However, it has been shown that intense speech therapy starting immediately after a stroke can degrade function compared with a control rehabilitative strategy involving simple social contact with limited conversation.

Animal Studies

There are three major animal models of stroke: global ischemia (blockade of all major blood vessels to forebrain), focal ischemia (blockade of specific blood vessels), and hemorrhage (disruption of blood vessels). Focal ischemia, caused by occlusion of the middle cerebral artery is the most common type of stroke, can readily be modeled in animals. Two approaches are in use, brief occlusion followed by vascular reperfusion and permanent vessel occlusion. The former approach leads to more extensive and slower secondary degeneration of peri-injury and remote tissue.

Although research suggests that stroke outcome can sometimes be improved modestly by rehabilitative training, including exercise, neurodegeneration in animals can actually be worsened by either very early forced motor inactivity or overactivity. It is currently unclear what the optimal strategies might be for different degrees, locations, or types of brain injury. Slow secondary degeneration may present a more feasible target for neuroprotection than does rapid acute cell loss.

Constraint-induced movement therapy has been examined after minor focal ischemia or hemorrhage in nonhuman primates and rodents. Where examined, such forced forelimb rehabilitative experience has been found to improve function, reduce loss of cortical motor maps of the upper extremity, and expand such maps into novel regions of cortex. Housing animals in an enriched environment (**Figure 3**) or providing skill training after focal ischemic injury can enhance dendritic arborization and synaptogenesis in remaining tissue in both the peri-injury and homotopic brain regions.

The timing of the exercise is important. In some studies, exercise (usually treadmill running) has been provided before the ischemic insult in order to mimic a more active lifestyle. These studies have generally shown a 'preconditioning' effect, that is, a reduction in the size of the ischemic damage because of prior exercise. Intense rehabilitation during the first few days after injury can worsen the injury even as it improves certain functions. The exaggeration of damage, which may take many weeks to be readily detectable, does not appear to be due to a general effect on brain because forced forelimb overuse exaggerates only injury to the forelimb region of the brain but not to injury of the visual cortex. Intense forelimb use may increase the extent of delayed secondary degeneration of vulnerable peri-injury tissue via local hyperthermia, excitotoxicity, or metabolic demands. Sudden social isolation from cagemates at the time of injury can combine with the effects of limb overuse to cause damage that is even more extensive. This may be important clinically because stroke patients often have the additional stress of being confined to a relatively isolated hospital environment rather than quickly returning home to be surrounded by family and friends.

Despite exaggerating neural injury, intense exercise and motor training can improve function or delay dysfunction through learned compensatory motor behaviors in animals. Indeed, motor enrichment can overcome deficits so effectively that the adverse effects of secondary degeneration are difficult to observe with behavioral outcome measures alone. Ideally, the rehabilitation training would be applied so that the benefits accrue without tissue loss. Based on animal

models, motor enrichment might best be applied sparingly during the acute poststroke period and gradually increased in intensity during subacute periods and chronically.

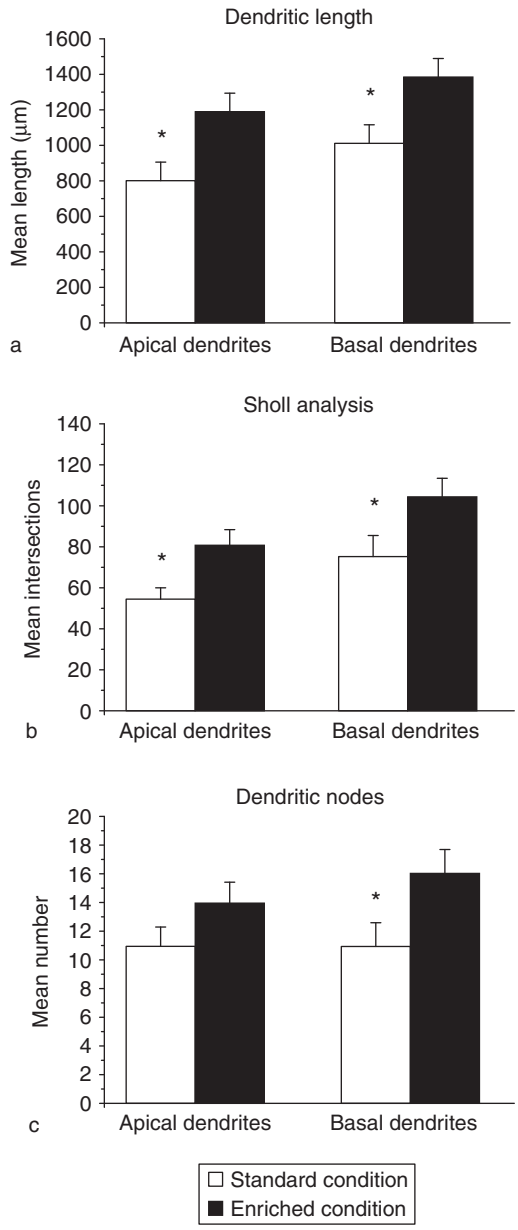

Figure 3 The effects of enriched rearing on (a) maximal dendritic length, (b) the number of intersections, and (c) branching of layer-III parietal pyramidal neurons. Rat pups were placed in an enriched environment at weaning (21 days) and examined at 3 months. As shown, environmental enrichment provoked both increased dendritic arborization and the density of dendritic spines in layer-III parietal pyramidal neurons. In addition, the rats raised in the enriched environment exhibited improved performance in a radial arm maze and a Morris water maze (data not shown). From Leggio MG, Mandolesi L, Federico F, et al. (2005) Environmental enrichment promotes improved spatial abilities and enhanced dendritic growth in the rat. *Behavioural Brain Research* 163: 78–90.

How Might Exercise Work?

It is now clear from rodent research that exercise improves learning, increases resistance to depression, and enhances synaptic plasticity such as long-term potentiation. This is accompanied in the brain by a series of biochemical events associated with enhanced plasticity (Table 1). Much of these data have come from studies of the hippocampus and cortex, but other brain regions can show plastic changes, too. In short, there are several effects of exercise on brain function, each of which could contribute to its apparent ability to reduce the incidence of neurodegeneration as well as the progression of neurodegeneration if it does occur. Thus, the question is not how exercise can be neuroprotective but which of the many changes are most critical to the neuroprotective effects of exercise and how the changes interact with each other. One of the variables in these relationships is deemed to be particularly important – trophic factors.

Table 1 Selected effects of exercise on brain[a]

Type of change	*Examples*
Transmitter release	Increased serotonin, dopamine, γ-aminobutyric acid, glutamate; decreased acetylcholinesterase
Increased levels of messenger RNA and/or protein for trophic factors or their receptors	Brain-derived neurotrophic factor, basic fibroblast growth factor, fibroblast growth factor-2, glial cell-derived neurotrophic factor, insulin-like growth factor 1, nerve growth factor, vascular endothelial growth factor, tyrosine receptor kinase B
Reduced inflammatory markers	Tumor necrosis factor-α (TNF-α), TNF-α receptors, intercellular adhesion molecule-1
Increased amounts of synaptic proteins	Synapsin, synaptophysin, NR2B, and glutamate receptor 1
Stimulation of morphological changes	Gliogenesis, neurogenesis, synaptogenesis, and vascularization (angiogenesis)
Activation of survival cascades and associated transcription factors	Ras/extracellular signal-regulated kinase and phosphoinositide-3 kinase/protein kinase B; cyclic adenosine monophosphate responsive element binding protein
Increases in antioxidant defenses	Superoxide dismutase, catalase (decreased reactive oxygen species)
Modification of brain circuitry	Increased area devoted to task; changes in anatomical circuitry, changes in activity of existing neurons and circuits

[a]Some entries are based on only one or two publications. Excluded from this list are variables that as yet have been examined only in peripheral tissue (often muscle).

Exercise Regulates Trophic Factors

Exercise is a powerful regulator of gene expression and protein levels for several trophic factors. The largest amount of data on this issue comes from studies of hippocampal brain-derived neurotrophic factor (BDNF). The neurotrophin BDNF promotes neuronal survival and resistance to insults and is essential for brain plasticity mechanisms, including certain forms of long-term potentiation. Exercise increases BDNF messenger RNA expression in hippocampal neurons. The gene upregulation is rapid (after several hours of exercise; Figure 4), occurs in both male and female rats and mice, is sustained even after several months of exercise, and is paralleled by increased amounts of BDNF protein. Once induced, protein levels remain elevated for several days after exercise has stopped, then decay back to baseline. It is interesting that an exercise regime of alternating days of exercise is as effective as daily activity for increasing hippocampal BDNF protein levels. The practical significance of these findings is that even intermittent exercise is a powerful enough stimulus to activate the molecular machinery that drives synaptic plasticity.

Can increases in trophic factors explain neuroprotection? The evidence for neuroprotective effects of trophic factors is considerable. Trophic factors such as glial cell-derived neurotrophic factor and BDNF can protect against some of the spontaneous neuronal loss that occurs *in vitro* and that is associated with normal aging. These and other trophic factors can also reduce the damage normally caused by several neurotoxins and stroke. Both of these types of neuroprotective effects may, in turn, be due in part to the ability of trophic factors to stimulate survival cascades. Further evidence for a role of trophic factors in preventing neurodegeneration comes

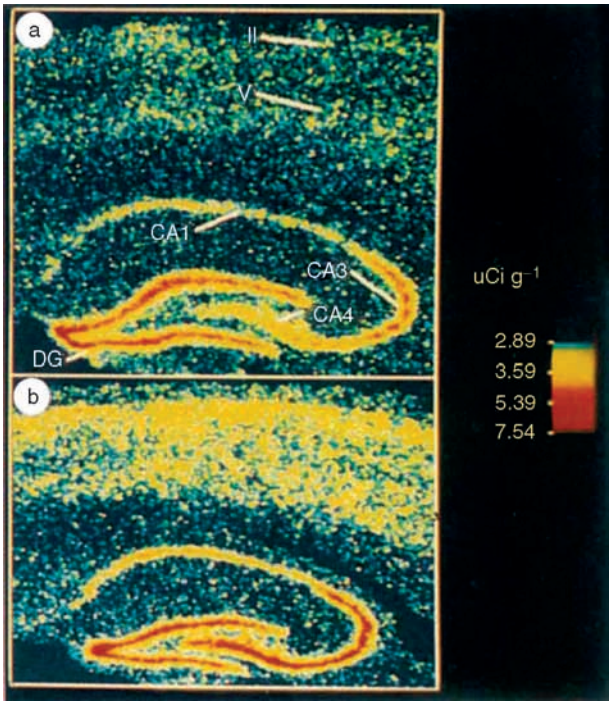

Figure 4 Brain-derived neurotrophic factor (BDNF) messenger RNA in rat brains obtained under (a) basal, sedentary conditions and (b) after 7 days of exercise. Neocortical layers II and V, as well as hippocampal areas CA1, CA3, and CA4 and dentate gyrus (DG) are shown. This *in situ* analysis of exercise-induced increases in BDNF message may have been the first observation of an effect of exercise on a neurotrophic factor in brain. From Neeper SA, Gómez-Pinilla F, Choi J, and Cotman C (1995) Exercise and brain neurotrophins. *Nature* 373: 109.

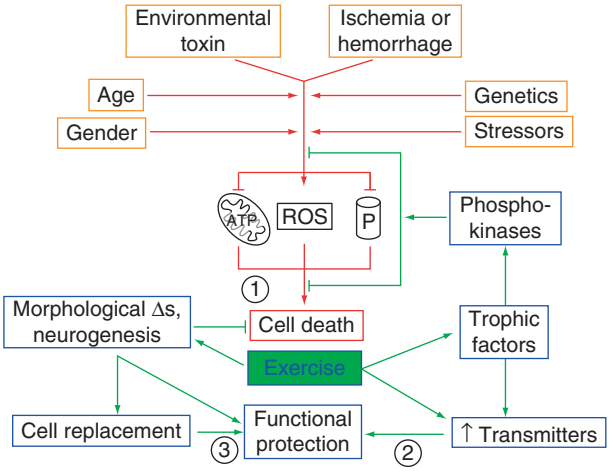

Figure 5 Schematic indicating multiple ways by which exercise may alter brain. Shown (orange) are several of the factors that can influence the vulnerability of neurons, which might include age, gender, environmental toxins, or other insults (e.g., head injury), hemorrhage, ischemia, genetic predisposition, or stressors. These factors can combine to cause cell death through one or more mechanisms, including inhibition of mitochondrial function (leading, in part, to a reduction in adenosine triphosphate (ATP) production), an increase in oxidative stress, and disruption of the processes involved in handling misfolded proteins (e.g., via inhibition of the proteasome $\mathcal{P}$). Exercise may counteract these events, blocking neuronal death and/or reducing its rate of progression (①). This could result from an increase in the expression of trophic factors (e.g., glial cell-derived neurotrophic factor and brain-derived neurotrophic factor) leading to an increase in prosurvival cascades, such as those involving the kinases, extracellular signal-regulated kinase and protein kinase B. Other factors include morphological changes, such as synaptogenesis, neurogenesis (and associated cell replacement), and angiogenesis (and associated increases in oxygen and nutrients) (②). Among the other changes that could ameliorate cell death are neurochemical changes that could lead to an exercise-induced increase in the activity of residual neurons within a damaged system, and changes in neuronal circuitry leading to a transfer of function from damaged neurons to healthy neurons (③). ROS, reactive oxygen species.

from observations of reductions in trophic factors in brains of patients with PD and AD.

Summary and Conclusions

Stroke and neurodegenerative diseases such as PD and AD are a major health problem throughout the world. Moreover, since the principal risk factor for these conditions is age, the burden of these conditions will only increase in years to come unless neuroprotective treatments can be found. Although a great deal of research is ongoing to develop pharmacotherapies, there is not yet convincing evidence that such drugs are now in hand. Moreover, in the condition in which drug development and testing have been most intense, the results have been extremely disappointing. This work is continuing, as it should. In the meantime, however, the evidence for neuroprotective effects of exercise, at least in animal models, is considerable and mounting. Indeed, it is now fair to say that in addition to its many other beneficial effects, exercise also is a key to brain health.

Although this discussion enumerated many possible factors (Table 1) and then focused on a role for trophic factors, this focus should suggest that changes in trophic factors exist apart from the other changes enumerated. On the contrary, changes in the availability of trophic factors may be a consequence of some of the effects listed (e.g., gliogenesis) and in turn may be the stimulus for other effects (e.g., synaptogenesis and increased antioxidant defenses; Figure 5).

In the years ahead, it will be essential to learn more about the mechanism of these effects, to move carefully with research-based treatment strategies from animal models to humans, and to determine how best to optimize an exercise intervention to promote healthy aging as well as to reduce the progression of disease.

See also: Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: Neurodegeneration; Animal Models of Alzheimer's Disease; Animal Models of Parkinson's Disease; Animal Models of Stroke; Enteric Nervous System: Neurotrophic Factors; Exercise: Optimizing Function and Survival at the Cellular Level; Neurotoxins; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Stroke.

Further Reading

Adlard PA, Perreau VM, Pop V, and Cotman CW (2005) Voluntary exercise decreases amyloid load in a transgenic model of Alzheimer's disease. *Journal of Neuroscience* 25: 4217–4221.

Berchtold NC, Chinn G, Chou M, Kesslak JP, and Cotman CW (2005) Exercise primes a molecular memory for brain-derived neurotrophic factor protein induction in the rat hippocampus. *Neuroscience* 133: 853–861.

Bergen JL, Toole T, Elliott RG III, Wallace B, Robinson K, and Maitland CG (2002) Aerobic exercise intervention improves aerobic capacity and movement initiation in Parkinson's disease patients. *NeuroRehabilitation* 17: 161–168.

Bezard E, Dovero S, Belin D, et al. (2003) Enriched environment confers resistance to 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine and cocaine: Involvement of dopamine transporter and trophic factors. *Journal of Neuroscience* 23: 10999–11007.

Chen H, Zhang SM, Schwarzschild MA, Hernan MA, and Ascherio A (2005) Physical activity and the risk of Parkinson disease. *Neurology* 64: 664–669.

Cotman CW and Berchtold NC (2002) Exercise: A behavioral intervention to enhance brain health and plasticity. *Trends in Neurosciences* 25: 295–301.

Crizzle AM and Newhouse IJ (2006) Is physical exercise beneficial for persons with Parkinson's disease? *Clinical Journal of Sport Medicine* 16: 422–425.

Debow SB, Davies ML, Clarke HL, and Colbourne F (2003) Constraint-induced movement therapy and rehabilitation exercises lessen motor deficits and volume of brain injury after striatal hemorrhagic stroke in rats. *Stroke* 34: 1021–1026.

Ding Y-H, Young CN, Luan X, et al. (2005) Exercise preconditioning ameliorates inflammatory injury in ischemic rats during reperfusion. *Acta Neuropathologica* 109: 237–246.

Dishman RK, Berthoud HR, Booth FW, et al. (2006) Neurobiology of exercise. *Obesity* 14: 345–356.

Faherty CJ, Raviie Shepherd K, Herasimtschuk A, and Smeyne RJ (2005) Environmental enrichment in adulthood eliminates neuronal death in experimental Parkinsonism. *Brain Research Molecular Brain Research* 134: 170–179.

Hirsch MA, Toole T, Maitland CG, and Rider RA (2003) The effects of balance training and high-intensity resistance training on persons with idiopathic Parkinson's disease. *Archives of Physical Medicine and Rehabilitation* 84: 1109–1117.

Ivanco TL and Greenough WT (2000) Physiological consequences of morphologically detectable synaptic plasticity: Potential uses for examining recovery following damage. *Neuropharmacology* 39: 765–776.

Jankowsky JL, Slunt HH, Gonzales V, et al. (2005) Persistent amyloidosis following suppression of Abeta production in a transgenic model of Alzheimer disease. *PLoS Medicine* 2: e355.

Kleim JA, Jones TA, and Schallert T (2003) Motor enrichment and the induction of plasticity before and after brain injury. *Neurochemical Research* 28: 1757–1769.

Kramer AF, Bherer L, Colcombe SJ, Dong W, and Greenough WT (2004) Environmental influences on cognitive and brain plasticity during aging. *Journals of Gerontology, Series A: Biological Sciences and Medical Sciences* 59: M940–M957.

Lazarov O, Robinson J, Tang YP, et al. (2005) Environmental enrichment reduces A-beta levels and amyloid deposition in transgenic mice. *Cell* 120: 701–713.

Mattson MP (2000) Neuroprotective signaling and the aging brain: Take away my food and let me run. *Brain Research* 886: 47–53.

Schallert T and Whishaw IQ (1978) Two types of aphagia and two types of sensorimotor impairment after lateral hypothalamic lesions: Observations in normal weight, dieted and fattened

rats. *Journal of Comparative and Physiological Psychology* 92: 720–741.

Smith AD and Zigmond MJ (2003) Can the brain be protected through exercise? Lessons from an animal model of parkinsonism. *Experimental Neurology* 184: 31–39.

Tillerson JL, Cohen AD, Philhower J, Miller GW, Zigmond MJ, and Schallert T (2001) Forced limb-use effects on the behavioral and neurochemical effects of 6-hydroxydopamine. *Journal of Neuroscience* 21: 4427–4435.

Woodlee MT and Schallert T (2006) The impact of motor activity and inactivity on the brain: Implications for the prevention and treatment of nervous system disorders. *Current Directions in Psychological Science* 15: 203–206.

Exercise: Optimizing Function and Survival at the Cellular Level

A Russo-Neustadt, California State University, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

For many years, consistent physical activity has been associated with improved mental health as well as physical health and longevity. Through recorded history, several nonscientific assertions have been made that 'an active body makes for a healthy mind.' In the past 30 years, several reports from the sports medicine field have documented acute alterations in mood and cognitive functioning immediately following a bout of vigorous physical activity. Both aerobic and weight-bearing forms of activity have been associated with decreased anxiety, improved mood, and a subjective or actual experience of sharper mental functioning. In more recent years, formal, long-term exercise studies have been conducted, reporting improved cognitive functioning, heightened coping abilities, and decreased incidence of mental disorders such as dementia and depression. Importantly, many of these studies have focused on the aging population.

Intervention studies of the effects of exercise on mental functioning in older adults have revealed prevention of memory decline, greater attentional abilities, and substantial improvements in executive functions such as the ability to plan activities, focus on task-relevant information, and perform multiple skills concurrently. Recent prospective cohort studies indicate that regular exercise is associated with reduced risk and delayed onset of age-associated cognitive impairments and dementia. A direct imaging study of older adults has revealed decreased brain tissue volume in several brain areas with age, and a striking reduction in brain tissue loss with regular activity.

Exercise has not only shown promise as an intervention to preserve cognitive functioning, but also has been demonstrated to improve emotional functioning and mood. For example, long-term exercise has been linked with improved morale in normal elderly populations. Exercise has also been introduced as an adjunct for antidepressant medication treatment or as a single intervention in controlled clinical trials for depression, and significant improvements in symptoms have been reported. Studies have shown that a long-term (16 weeks) exercise treatment program was equally as effective as medication treatment for major depression. Regular activity has also been associated with a reduced risk for developing depression in elderly populations.

It is thought that natural selection of human ancestors, after many generations of survival as hunter-gatherers, led to the evolution of humans for whom regular vigorous activity improves and prolongs survival and function. It is likely that normal human cellular function has evolved in such a way that a threshold of physical activity is required for optimal physiologic gene expression in many tissues throughout the body, including the brain. As discussed in the following sections, exercise maintains the activity of a number of important central neurotransmitter systems, and this activation state sets into motion key signal transduction pathways promoting survival and functioning of neurons at the cellular level. It appears that the activation state of neurons (as well as the entire organism) enhances processes associated with growth and plasticity.

Exercise and Central Nervous System Activation: Neurotransmitter Function

A wealth of evidence from both animal and human studies indicates that general aerobic activity leads to enhanced secretion of a number of important neurotransmitters and neuromodulators in the central nervous system (CNS). Such findings serve as neurochemical evidence for increased activation of the CNS (both global excitation of the brain and stimulation within specific brain regions). Enhanced concentrations of neurotransmitters and neurotransmitter metabolites have been measured in blood and cerebrospinal fluid (CSF) of humans. In animal models, microdialysis studies have allowed investigators to observe dynamic neurotransmitter release within specific brain areas during (and following) exercise. These specific forms of neurotransmitter activation have important implications for vulnerability to stress and depression, and impact survival and plasticity at the cellular level.

The monoamines, most notably serotonin, norephinephrine, and dopamine, are essential for multiple CNS functions, including arousal, cognition, and the regulation of mood. The activity of these systems is compromised with chronic stress and depression, and most antidepressant treatments in current practice are designed to enhance monoaminergic activity. In animal studies, exercise has been observed to enhance norephinephrine (NE) release in the brain stem, hypothalamus, and limbic brain regions. In humans, increased NE release into plasma has been observed following acute

NRSC: 00144

exercise. These changes prevent NE depletion during stress, and thus have been demonstrated to modulate physiological vulnerability to stressors. Some evidence also exists for changes in serotonin (5-hydroxytryptamine; 5-HT) release and metabolism during exercise. Clinical studies have indicated enhanced 5-HT availability and/or release in physically trained men, or after acute exercise, and another study has demonstrated enhanced release of dopamine and tryptophan (a metabolic precursor of serotonin). It has been proposed that enhanced dopamine neurotransmission occurring during exercise may contribute to enhanced central as well as peripheral function. Adequate monoaminergic (particularly serotonergic) function optimizes behavioral neuromodulation during various stressors.

The excitatory neurotransmitter, glutamate, plays a very important role in functional neuronal plasticity, signaling changes in synaptic function and morphology occurring during learning and other forms of neuronal activation. Studies have revealed that long-term exercise increases the levels or activated forms of several postsynaptic glutamatergic receptors in the cortex and hippocampus of rodents. Exercise-enhanced neurogenesis (discussed later) has been demonstrated to be dependent upon glutamatergic (*N*-methyl-D-aspartate; NMDA) receptor function. Neuropeptides such as vasopressin and galanin have also been studied in exercise research. Investigators have demonstrated changes in affinity of vasopressin, along with noradrenergic, receptors following exercise training. Both voluntary wheel-running and treadmill exercise training increase galanin gene expression in the rat locus coeruleus. Increased serum galanin secretion has also been detected after acute exercise in humans. Galanin may have an important neuromodulatory role in the NE system adaptations occurring during exercise.

In addition to an acute impact on intracellular activity and metabolism, cellular neurotransmission, as stimulated by exercise, also directs gene expression via the activation of signal transduction pathways. A key example is the activation of protein kinase A (PKA) and cyclic adenosine monophosphate (AMP) response element-binding protein (CREB) by monoamines via G-protein-coupled receptors, subsequently directing the expression of neurotrophins (discussed later) and other molecules promoting neuronal growth and survival.

Exercise and Growth Factor Expression

The neurotrophins are a family of molecules, including nerve growth factor (NGF), brain-derived neurotrophic factor (BDNF), and neurotrophins (NT-3 and NT-4/5). These compounds promote survival, growth, and differentiation of neurons during development and throughout the life span. Recent evidence has indicated that BDNF may play an important role in supporting normal cognitive and emotional function into old age and senescence, and that its expression is influenced by neuronal activity.

It has become evident that physical activity increases the expression of BDNF and other growth factors in several areas of the CNS. In animal models, both voluntary wheel-running and forced treadmill training (under conditions controlling stress) rapidly and robustly enhance BDNF mRNA and protein expression in the hippocampus. It is known that environmental enrichment, or the addition of visual, motor, tactile, and social stimulation to the rodents' cage environment, enhances BDNF expression in the hippocampus and neocortex, and that activity has been suggested to be the single most important component of this model for the observed increases in neurotrophin expression and CNS function. Increased hippocampal BDNF mRNA occurs in as little as 6 h and is sustained with long-term (several weeks) activity. Increases in BDNF protein are evident within 1–2 weeks and are sustained for at least 3 months. Increases in both BDNF mRNA and BDNF protein are highly correlated with running distance. Increased peripheral motor activity has been demonstrated to be required for these observed changes, which are thought to mediate many of the prosurvival effects of exercise. Confocal microscopic analysis revealed that short-term (7 days) activity caused a redistribution of BDNF from the cell body to the dendrites in neurons of the lumbar spinal cord, whereas longer term (28 days) exercise also led to increased perikaryonal BDNF.

The phenomenon of increased hippocampal BDNF with exercise is very much intact throughout the life span of rodents. Even with relatively small amounts of activity, aged rats demonstrate an ability to significantly enhance BDNF mRNA and protein to an extent no less than that achieved with antidepressant medications. Voluntary wheel running has also greatly enhanced neurogenesis and hippocampal (learning) function in aged mice without prior exercise experience. Recent evidence also indicates that swimming exercise of a mother rat during pregnancy enhances hippocampal BDNF mRNA expression and neurogenesis of the offspring. Exercise can therefore represent a powerful, nonpharmacological means for maintaining neurotrophin levels and neuronal health throughout development, and throughout the life span.

It is becoming increasingly evident that exercise, and the resulting enhancement in neurotrophin expression, can improve neuronal function. Exercise has been

demonstrated to increase the levels of proteins involved in vesicle formation and neurotransmitter release, such as synaptophysin and synapsin I. Enhanced synaptic function, such as that occurring during development or learning, can thus result from exercise. One cellular phenomenon that is highly dependent upon BDNF, and that in turn stimulates additional BDNF expression, is long-term potentiation (LTP). LTP is activated in the hippocampus during the process of spatial learning. An increase in hippocampal LTP has been observed in rats undergoing voluntary wheel-running activity, which appears to reflect an alteration in the induction threshold for synaptic plasticity occurring during exercise. Voluntary wheel-running activity for 3 weeks has been shown to enhance performance of rats in the Morris water maze, a widely studied spatial learning paradigm. The acquisition of another challenging spatial learning task, the radial arm maze, has been facilitated by long-term voluntary exercise. In aged (23 months) rats, long-term (7 weeks) mild, forced treadmill exercise improved performance in the Morris water maze, demonstrating both a shorter path length and significantly decreased latency to find the hidden platform.

Evidence exists that the BDNF-increasing effects of exercise in the hippocampus follow, and are dependent upon, noradrenergic activation. In addition, some cognitive benefits of exercise, such as the augmentation of contextual fear conditioning, are attenuated by blockade of β-adrenoreceptors. Stimulation of the PKA pathway following adrenergic G-protein-coupled receptor activation mediates increased BDNF expression via the transcription factor CREB. Abundant evidence exists for increased CREB phosphorylation (activation) and expression following exercise. The enhanced neurotrophin expression brought about by exercise subsequently activates tyrosine kinase (Trk) receptors and key intracellular signaling pathways mediating survival and growth. One important transcriptional target of these pathways is, once again, CREB. CREB, along with other transcription factors (including nuclear factor-kappa B, which has been shown to be increased with exercise), enhances the expression of BDNF along with several other prosurvival genes. In this way, BDNF expression can be increased via a positive feedback mechanism. This type of mechanism may explain the very rapid and robust increase in BDNF expression observed with exercise. Researchers in several laboratories have found evidence for neurotrophin expression being enhanced via a type of positive feedback loop. One of their reports cited direct evidence for changes in neurotrophin expression being driven by neurotrophin pulses in neuronal culture. This mechanism is discussed further in the following section.

In addition to the striking changes in hippocampal BDNF, evident with exercise, increases in BDNF and/or NGF and NT-3 have been demonstrated in the hippocampus, neocortex, basal forebrain, striatum, and spinal cord. The spinal cord has been a focus area for study of the effects of activity on neurotrophin expression, survival, and recovery from injury. Motor activity (treadmill walking or stepping) has been shown to enhance the expression of several neurotrophins in the spinal cord, and to enhance neuronal outgrowth and recovery in injured areas. In intact rats, locomotor exercise increases levels of BDNF and its receptor, TrkB, in the lumbar spinal cord, and also increases levels of NT-3 and its receptor, TrkC. In models of spinal cord injury, exercise also compensated for reductions in BDNF occurring in the lumbar spinal cord after hemisection to the midthoracic region, and increased BDNF was associated with reduction in sensory symptoms such as allodynia. In the latter study, increasing BDNF protein levels were highly correlated with running distance. Exercise was also shown to enhance regrowth of axons after a nerve-crush injury; and sensory (dorsal root ganglion) neurons from animals undergoing exercise also showed increased neurite outgrowth. Exercise interventions, therefore, show much promise to promote the recovery of motor function following spinal cord injury.

Another highly important growth factor for the mediation of exercise-induced changes within the brain is insulin-like growth factor (IGF). Investigators have shown that the uptake of peripheral IGF into the CNS occurs during voluntary exercise, and that this action is essential for both increased CNS expression of neurotrophins and functional recovery from stress. Recent evidence indicates that IGF from central as well as peripheral sources participates in exercise-induced increases in hippocampal BDNF expression as well as augmentation of recall in the Morris water maze, and also increases progenitor cell proliferation in the dentate gyrus. Much evidence is accumulating that IGF plays a very important role in antiaging and anticellular stress phenomena occurring with exercise, environmental enrichment, and dietary restriction.

It is becoming increasingly evident that interaction between multiple growth factors may be playing a role in exercise-associated neurochemical changes. Glial maturation factor (GMF), production of which is increased with exercise, appears to also be essential for exercise induction of BDNF expression. Vascular endothelial growth factor (VEGF) has

NRSC: 00144

been demonstrated to be particularly important for exercise-induced adult hippocampal neurogenesis.

Exercise and Cellular Survival Signaling

The survival of neurons is dependent upon the activity of a number of finely tuned cellular signaling cascades. Two primary signaling pathways known to promote neuronal growth and survival are the phosphatidylinositol 3-kinase (PI3K)/Akt and mitogen-activated protein kinase (MAPK) pathways, which influence gene expression through activation of transcription factors, such as CREB. The MAPK and PI3K pathways are regulated by neurotrophin/Trk receptor signaling, and therefore represent two of the major signaling pathways mediating BDNF-induced TrkB activation. In cultured cortical neurons subjected to apoptotic conditions, BDNF has been shown to promote survival and neuroprotection by activating the MAPK and PI3K pathways. Consistently, BDNF stimulates the phosphorylation of Akt, which plays a critical role in controlling the balance between survival and apoptosis. Akt is activated via the PI3K pathway, following stimulation by insulin and several other growth/survival factors, and promotes cell survival by inhibiting apoptosis through its ability to phosphorylate and inactivate several substrates, including BAD, the forkhead transcription factors, caspase-9, and GSK-3β. In addition to increasing cellular survival, the PI3K and MAPK pathways play principal roles in promoting neurite growth, synaptic strength, and plasticity. Via these pathways, excitatory neural stimulation via exercise may therefore be an important means to enhance neuronal growth or recovery in an activity-dependent manner.

Microarray analyses of brain tissue from rodents undergoing physical activity have provided an excellent preliminary view of the genes and intracellular molecules activated by exercise. The 'exercise profile' includes a number of molecules along the PI3K survival pathway, and several other genes promoting cellular growth and survival. Some recent studies have also directly demonstrated increased concentrations of proteins along the PI3K and MAPK pathways in the brains of rodents undergoing chronic or subchronic voluntary activity.

Exercise enhances by two- to threefold the generation and survival of new neurons derived from progenitor cells within the adult hippocampus. Several of the key humoral factors outlined in the previous section, such as IGF-1 and BDNF, are thought to mediate exercise-stimulated neurogenesis. Evidence indicates that the MAPK signaling pathway is required for IGF-1-stimulated proliferation, whereas the PI3K/Akt pathway plays a major role in maintaining survival and inhibiting apoptosis. Recent evidence also indicates that neurogranin, a specific substrate of protein kinase C, is required for exercise-induced neurogenesis and resulting cognitive enhancements. Regarding the birth of new neurons, exercise appears to have a more profound effect upon proliferation than on differentiation or survival. Anatomical examination has revealed that the proportion of cells exhibiting an immature morphology is increased within the hippocampal dentate gyrus following exercise. Subsequent differentiation of these progenitors into mature neurons, and integration into the neuronal circuitry, may be stimulated by environmental enrichment and learning. Exercise may therefore benefit hippocampal function by increasing the potential for neurogenesis. Indeed, human evidence exists that cognitive enrichment complements physical fitness training in promoting life span and mental health.

The effectiveness of antidepressant medications for improving mood and emotional function has been attributed to increased neurotrophin expression, stable signaling changes, and increased neurogenesis within the hippocampus, all of which occur after 2–3 weeks of treatment (the time frame often required for symptom relief in humans). The relatively rapid effectiveness of exercise for improving neurotrophin expression (within hours to days) may be attributed to the simultaneous activation of several survival- and plasticity-promoting signaling cascades, including the PI3K and MAPK pathways, pathway cross-talk, and positive feedback. Several clinical studies have indicated that exercise is an effective antidepressant or adjunctive agent. Electroconvulsive therapy (ECT), a treatment of choice for medication-resistant patients with major depression, has been demonstrated to activate the expression of multiple growth-promoting genes, including the 'exercise profile.'

Investigations into the intracellular mechanisms of survival-promoting interventions such as exercise have also been pursued in cell culture models. It has been revealed that primary neurons harvested from exercising animals retain an activated or stimulated state. In one study, dorsal root ganglion neurons showed an increase in neurite outgrowth when cultured from animals that had undergone 3 or 7 days of exercise. Since much evidence has pointed to the possibility that NE activation is an essential component of the cascade of events leading to exercise-induced increases in neurotrophin expression, we studied cellular signaling within a model system of NE applied to primary hippocampal neurons. NE application to hippocampal neurons in culture significantly increased BDNF levels in a time- and dose-dependent manner, and the phenomenon appeared to be dependent upon Trk receptor activation as

well as on activity of the PI3K and MAPK pathways. The NE-induced activation of the MAPK cascade via Trk activation has also been demonstrated in primary cortical neurons.

Neurodegeneration: How Exercise Can Decrease Risk and Contribute to CNS Repair Mechanisms

The aging process is associated with metabolic dysfunction and an impaired ability of the organism to cope with adversity. With normal aging, the efficiency of metabolic processes decreases along with the efficacy of systems designed to remove molecules associated with oxidative stress. Oxidative stress can lead to neurodegeneration in many areas of the brain. The hippocampus is also particularly vulnerable to the effects of psychosocial stress and the secretion of stress hormones (glucocorticoids), since it is an area with very high glucocorticoid receptor numbers. Severe, prolonged stress is thought to play an important role in the pathophysiology of several mental disorders such as major depression. Age-associated neurological disorders such as Alzheimer's disease and Parkinson's disease involve a great pronouncement of neurodegeneration beyond the normal aging process. Exercise has shown increasing promise as a preventive or reparative intervention for these various forms of neurodegeneration.

Although exercise imposes increased metabolic demands on the organism, evidence exists that it improves metabolic efficiency and decreases cellular vulnerability. As noted earlier, exercise is capable of increasing the expression of hippocampal BDNF throughout the life span of rodents. Via neurotransmitter activation and enhanced neurotrophin production, exercise can protect neurons by stimulating the production of proteins that suppress oxyradical production, stabilize cellular calcium homeostasis, and inhibit apoptotic biochemical cascades. Decreased oxidative damage and improvement in cognitive function have been demonstrated in the rat brain following exercise. Exercise also modifies the age-related decrease of hippocampal growth-associated protein, GAP-43, and synaptophysin, and the age-related decrease in antioxidant enzyme activities. It has also been shown that exercise potentiates the expression of the heat shock protein, Hsp72, and that this could contribute to neuroprotection from stress. In addition, although exercise is associated with increased glucocorticoid secretion, voluntary activity can still decrease glucocorticoid levels and hippocampal damage associated with ongoing physical/psychological stress. Prior exercise was also demonstrated to be protective against reductions in hippocampal BDNF protein caused by acute immobilization stress, even though increased corticosterone levels were sustained. Exercise has also been demonstrated to reverse several detrimental effects of a high-fat diet, including increased levels of reactive oxygen species, decreased neurotrophin (BDNF) expression, decreased expression of downstream effectors of synaptic function, and behavioral (spatial learning) deficits.

Specific evidence is accumulating that exercise can provide benefits in the face of neurodegenerative disease. Investigators have shown that exercise almost completely protects the brain against experimentally induced parkinsonism. This finding may be associated with increased glial-derived neurotrophic factor levels. Moderate exercise has been demonstrated to induce neurotrophin production in humans (evident in serum levels), and this may underlie the benefits of exercise for multiple sclerosis. Exercise has also shown much promise as a therapeutic strategy against amyotrophic lateral sclerosis. Even brief exposures to voluntary exercise improved motor function and survival in a mouse model of this disorder. Evidence also exists that increased physical activity reduces stroke damage. Preischemic treadmill activity reduced neurologic deficits and brain infarct volume in rats; increased neurotrophin levels and microvessel density were associated findings. Microvessel integrity within the brain may be a key factor for the prevention of many types of neurodegenerative disorders. Exercise stimulates brain vessel density in experimental animals in a manner that is dependent upon IGF-1.

Aging is associated with reductions in both hippocampal neurogenesis and IGF-1 levels, and administration of IGF-1 to old rats reverses age-related changes in central energy metabolism/function, increases neurogenesis, and reverses cognitive impairments. A reduction in the activity of the growth hormone/IGF-1 axis with age is well documented in humans. Analogs of IGF-1, along with the molecules of the PI3K pathway, are present in a wide variety of organisms from nematodes to humans, and evidence is strong that their function is closely tied to stress resistance and survival. Insulin-like signaling, along with the actions of neurotransmitters and neurotrophins, therefore link energy metabolism to cellular survival in an evolutionarily conserved manner.

Conclusion

In mammals, it is becoming increasingly evident that optimal cellular function within the nervous system and throughout the body is favored by heightened

NRSC: 00144

activity. The evidence outlined in this article has shown us that various forms of physical activity, such as voluntary aerobic exercise, activate highly specific intracellular systems that optimize function and survival. These systems are put into motion by molecules such as monoaminergic neurotransmitters, IGF-1, and the neurotrophin BDNF. This brings us, at the cellular level, to a full-circle connection to our earlier discussion of evolution of humans as active hunter-gatherers, and how cellular machinery likely similarly evolved in such a way that optimal function would be favored by increased activity. The aforementioned studies have suggested that, at the molecular level, monoamines, BDNF, and IGFs can stimulate the production of proteins involved in cellular stress adaptation, survival, growth/repair, and neurogenesis. All of these actions appear to be dependent upon synaptic activity, which is enhanced during exercise. These mechanisms may also compensate for changes taking place with aging and situations of severe stress. There is much yet to be learned about the specific mechanisms involved in exercise-induced cellular adaptations, but we have much evidence that exercise is an essential intervention to maintain brain health throughout life.

See also: Cholinergic System; Exercise in Neurodegenerative Disease and Stroke; Glutamatergic and Gabaergic Systems; Monoamines; Neurotransmitters and Growth Factors: Overview; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Neurotrophins: Physiology and Pharmacology; Retrograde Neurotrophic Signaling.

Further Reading

Aberg ND, Brywe KG, and Isgaard J (2006) Aspects of growth hormone and insulin-like growth factor-I related to neuroprotection, regeneration, and functional plasticity in the adult brain. *Scientific World Journal* 6: 53–80.

Barbour KA and Blumenthal JA (2005) Exercise training and depression in older adults. *Neurobiology of Aging* 26(supplement 1): 119–123.

Colcombe SJ, Kramer AF, McAuley E, et al. (2004) Neurocognitive aging and cardiovascular fitness: Recent findings and future directions. *Journal of Molecular Neuroscience* 24(1): 9–14.

Mattson MP (2002) Brain evolution and lifespan regulation: Conservation of signal transduction pathways that regulate energy metabolism. *Mechanisms of Ageing and Development* 123: 947–953.

Mattson MP, Maudsley S, and Martin B (2004) A neural signaling triumvirate that influences ageing and age-related disease: Insulin/IGF-1, BDNF and serotonin. *Ageing Research Reviews* 3(4): 445–464.

McEwen BS (2002) The neurobiology and neuroendocrinology of stress. Implications for post-traumatic stress disorder from a basic science perspective. *Psychiatric Clinics of North America* 25(2): 469–494, ix.

Meeusen R and De Meirleir K (1995) Exercise and brain neurotransmission. *Sports Medicine* 20(3): 160–188.

Pereira AC, Huddleston DE, Brickman AM, et al. (2007) An *in vivo* correlate of exercise-induced neurogenesis in the adult dentate gyrus. *Proceedings of the National Academy of Sciences of the United States of America* 104(13): 5638–5643.

Popoli M, Gennarelli M, and Racagni G (2002) Modulation of synaptic plasticity by stress and antidepressants. *Bipolar Disorder* 4(3): 166–182.

Russo-Neustadt A (2003) Brain-derived neurotrophic factor, behavior, and new directions for the treatment of mental disorders. *Seminars Clinical Neuropsychiatry* 8(2): 109–118.

Sapolsky RM (2000) The possibility of neurotoxicity in the hippocampus in major depression: A primer on neuron death. *Biological Psychiatry* 48(8): 755–765.

Vaidya VA and Duman RS (2001) Depression – Emerging insights from neurobiology. *British Medical Bulletin* 57: 61–79.

Exocytosis: Ca^{2+}-Sensitivity

R Schneggenburger, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Transmission at chemical synapses is mediated by exocytosis of synaptic vesicles with the neuronal plasma membrane, a process which takes place at the active zones of nerve terminals. When an action potential arrives at a presynaptic active zone, voltage-gated Ca^{2+} channels open, causing a short inflow of Ca^{2+} ions into the cytoplasm of the presynaptic neuron (Figure 1). The subsequent increase in intracellular Ca^{2+} concentration ($[Ca^{2+}]_i$) triggers the membrane fusion of neurotransmitter-loaded synaptic vesicles, probably initiated by the binding of Ca^{2+} ions to synaptotagmin molecules, which are located on the surface of synaptic vesicles. By this process of exocytosis, neurotransmitter is released from a presynaptic neuron and can activate receptors on the surface of a postsynaptic cell.

The Microdomain Ca^{2+} Signal

Although it has been known since the pioneering studies of B Katz and colleagues in the 1960s that Ca^{2+} influx through voltage-gated Ca^{2+} channel triggers vesicle fusion at synapses, it has been less clear how intracellular Ca^{2+} regulates vesicle fusion and which intracellular Ca^{2+} concentration must be reached at the sites of vesicle fusion to efficiently trigger the release of neurotransmitter. This is because action potentials in neurons are brief ($\sim$1 ms) and consequently, Ca^{2+} channels admit only a brief inflow of Ca^{2+}. Computer simulations have shown that as long as a Ca^{2+} channel stays open, a 'spike' of enhanced $[Ca^{2+}]_i$ builds up around the open Ca^{2+} channel (Figure 1). The exact Ca^{2+} concentration reached at a neighboring docked and readily releasable vesicle critically depends on the distance between the Ca^{2+} channel and the vesicle, with Ca^{2+} concentration dropping off steeply with distance to the Ca^{2+} channel. The $[Ca^{2+}]_i$ reached in this single channel 'microdomain' also depends on a number of other parameters, such as the single channel conductance of the Ca^{2+} channel and the Ca^{2+} buffering strength of the cytoplasm. When the Ca^{2+} channel closes, the $[Ca^{2+}]_i$ in the microdomain collapses rapidly, within microseconds, to much lower values. To further complicate matters, the $[Ca^{2+}]_i$ signal at any given docked and readily releasable vesicle is, in many synapses, influenced by several Ca^{2+} channels located in close vicinity. We refer to the $[Ca^{2+}]_i$ signal experienced by a docked and readily releasable vesicle as the 'local' $[Ca^{2+}]_i$ signal. Although tremendous advances have been made in microscopic imaging techniques in recent years, it is still not possible to directly measure both the amplitude and the time course of this local $[Ca^{2+}]_i$ signal.

The Ca^{2+} Uncaging Experiment

In order to learn more about how intracellular Ca^{2+} regulates vesicle fusion and transmitter release, Ca^{2+} uncaging in the cytoplasm of single cells has been used to trigger vesicle fusion. During Ca^{2+} uncaging, a Ca^{2+}-loaded photosensitive Ca^{2+} chelator in the cytoplasm of a cell is photolyzed by a brief UV light pulse from a flash lamp or a UV laser coupled with an epifluorescence microscope. With this method of $[Ca^{2+}]_i$ elevation, spatial gradients that occur during the opening of voltage-gated Ca^{2+} channels are avoided. A fluorescent Ca^{2+} indicator dye can be used to measure $[Ca^{2+}]_i$. Since the $[Ca^{2+}]_i$ signal is spatially homogeneous, this technique ensures that the Ca^{2+} indicator dye and the readily releasable vesicles docked at the membrane 'see' the same $[Ca^{2+}]_i$. Thus, the jump in $[Ca^{2+}]_i$ that triggers vesicle fusion in these experiments can be directly measured by the fluorescent Ca^{2+} indicator dye, and a quantitative relationship between the rate of vesicle fusion and $[Ca^{2+}]_i$ can be derived. Such experiments are usually combined with whole cell patch clamp recording of a single secretory cell or a nerve terminal. A solution containing the Ca^{2+}-loaded photosensitive Ca^{2+} chelator, the Ca^{2+} indicator dye, and additional constituents is perfused inside the cell via the recording patch pipette.

Ca^{2+}-Dependent Rates of Exocytosis in Secretory Cells and Nerve Terminals

Ca^{2+} uncaging experiments to measure the intracellular Ca^{2+} sensitivity of vesicle fusion were first done in neuroendocrine secretory cells such as chromaffin cells, which are easily amenable to patch clamp studies. In these cells, hormones and transmitters are released via membrane fusion of large dense core vesicles. The rate of exocytosis, as measured by membrane capacitance, depended steeply on the amplitude of the $[Ca^{2+}]_i$ step produced by Ca^{2+} uncaging, and with $[Ca^{2+}]_i$ steps to approximately 10 μM, the fastest component of membrane capacitance rise had a time constant (τ) of approximately 50 ms. This means that the fastest component of the readily releasable pool is released at a rate of $1/\tau$, or 20 s^{-1} at

10 μM $[Ca^{2+}]_i$ (Table 1). Thus, in chromaffin cells and other neuroendocrine cells, secretion takes place at low micromolar Ca^{2+} concentrations, but the kinetics of vesicle fusion at this $[Ca^{2+}]_i$ are relatively slow. This might reflect the fact that for their physiological functions, neuroendocrine cells probably do not need very fast secretory responses in the low milliseconds range.

The intracellular Ca^{2+} sensitivity of vesicle fusion has also been measured in terminals of bipolar cells of the retina and in inner hair cells of the cochlea. These cells release the neurotransmitter glutamate from small synaptic vesicles. Glutamate release from these cells occurs at specialized release sites called 'ribbon' synapses, which have a protein-containing, electron-dense organelle – the ribbon – close to their active zones. At ribbon synapses, neurotransmitter can be released both rapidly and in a sustained fashion. Ca^{2+} uncaging has shown that fast release rates of approximately 1000 s^{-1} (corresponding to a secretion time constant of 1 ms) are reached at approximately 100 μM $[Ca^{2+}]_i$ or higher in bipolar cell terminals and in inner hair cells (Table 1). Thus, the results of these experiments agreed with earlier estimates from theoretical work, which had predicted that an intracellular Ca^{2+} concentration of approximately 100 μM must occur close to docked vesicles in order to ensure fast rates of vesicle fusion.

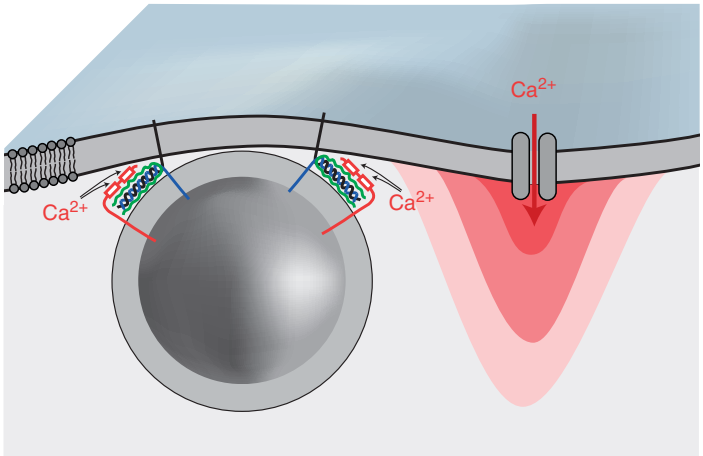

Figure 1 Scheme of a membrane-docked, readily releasable vesicle with a neighboring Ca^{2+} channel. The intracellular Ca^{2+} signal that builds up when the Ca^{2+} channel opens is shown in red. The SNARE proteins synaptobrevin (blue), SNAP-25 (green), and syntaxin (black), which define the readily releasable state of a vesicle and contribute to tethering the vesicle to the plasma membrane, are also shown. The vesicle protein synaptotagmin 1/2 with its two Ca^{2+} binding domains (C2 domains), which is probably the Ca^{2+} sensor for vesicle fusion, is shown in red. Note that in many synapses, the microdomain Ca^{2+} signal experienced by readily releasable vesicles is created by several neighboring Ca^{2+} channels.

Table 1 Ca^{2+}-dependent rates of vesicle fusion estimated by Ca^{2+} uncaging in different secretory cells

	Rate constant (at μM $[Ca^{2+}]_i$)[a]	*Ca^{2+} cooperativity*[b]	*References*[c]
Chromaffin cells	20 s^{-1}, at 10 μM	3	1, 2
Bipolar nerve terminals	1000 s^{-1}, at 150 μM	4	3
Inner hair cells	1000 s^{-1}, at 100 μM	5	4
Calyx of Held	1000 s^{-1}, at 10 μM	5	5
	1000 s^{-1}, at 30 μM	5	6

[a]Rate constant (= inverse time constant) of pool depletion for the fastest component of exocytosis.
[b]Number of Ca^{2+} ion binding steps in kinetic models of Ca^{2+} binding and vesicle fusion.
[c]References: 1, Heinemann C, Chow RH, Neher E, and Zucker RS (1994) Kinetics of the secretory response in bovine chromaffin cells following flash photolysis of caged Ca^{2+}. *Biophysical Journal* 67: 2546–2557; 2, Voets T (2000) Dissection of three Ca^{2+}-dependent steps leading to secretion in chromaffin cells from mouse adrenal slices. *Neuron* 28: 537–545; 3, Heidelberger R, Heinemann C, Neher E, and Matthews G (1994) Calcium dependence of the rate of exocytosis in a synaptic terminal. *Nature* 371: 513–515; 4, Beutner D, Voets T, Neher E, and Moser T (2001) Calcium dependence of exocytosis and endocytosis at the cochlear inner hair cell afferent synapse. *Neuron* 29: 681–690; 5, Bollmann J, Sakmann B, and Borst J (2000) Calcium sensitivity of glutamate release in a calyx-type terminal. *Science* 289: 953–957; 6, Wölfel M and Schneggenburger R (2003) Presynaptic capacitance measurements and Ca^{2+} uncaging reveal sub-millisecond exocytosis kinetics and characterize the Ca^{2+} sensitivity of vesicle pool depletion at a fast CNS synapse. *Journal of Neuroscience* 23: 7059–7068.

The Ca^{2+} Sensitivity of Transmitter Release at the Calyx of Held

The general expectation that fast rates of neurotransmitter release need high transients of local $[Ca^{2+}]_i$ (~100 μM) has not been found to hold true at a large, fast-releasing central nervous system (CNS) synapse, the calyx of Held. This axosomatic synapse between a large nerve terminal and a principal cell of the medial nucleus of the trapezoid body contains several hundred individual active zones. At this synaptic connection, an action potential in the presynaptic nerve cell produces a fast-rising and decaying excitatory postsynaptic current (EPSC) with an amplitude of several nanoamperes, which is mediated by the release of approximately 100–200 vesicles. The calyx of Held nerve terminal is unusually large (~10–15 μm in circumference), which has allowed researchers to make patch clamp recordings directly from this nerve terminal. This has facilitated the use of Ca^{2+} uncaging at this CNS nerve terminal (Figure 2). The calyx of Held synapse is the only nerve terminal in the mammalian CNS at which the intracellular Ca^{2+} sensitivity of vesicle fusion has been studied systematically.

Figure 2 illustrates a typical Ca^{2+} uncaging experiment performed at the calyx of Held. Two whole cell patch clamp recordings were made simultaneously, one from the soma of the postsynaptic neuron and the other from the presynaptic element, the calyx of Held nerve terminal. The presynaptic patch pipette contained 1 mM of the photolyzable Ca^{2+} chelator DM-nitrophen as well as 0.1 mM of the Ca^{2+} indicator fura-2FF, which stained the presynaptic calyceal nerve terminal (Figure 2(a)). Flashes of various intensities were given every 30–40 s to elevate the intracellular Ca^{2+} concentration in the nerve terminal (Figure 2(b)). In response to these presynaptic $[Ca^{2+}]_i$ elevations, EPSCs were recorded, which reflect the neurotransmitter release activity from the nerve terminal. As can be seen in Figure 2, Ca^{2+} uncaging stimuli of increasing amplitude evoked EPSCs of increasing amplitudes. The relationship between EPSC amplitude and presynaptic $[Ca^{2+}]_i$ was apparently strongly supralinear (Figures 2(b)–2(d)).

In order to derive the rates of transmitter release from the EPSCs recorded in the postsynaptic cell,

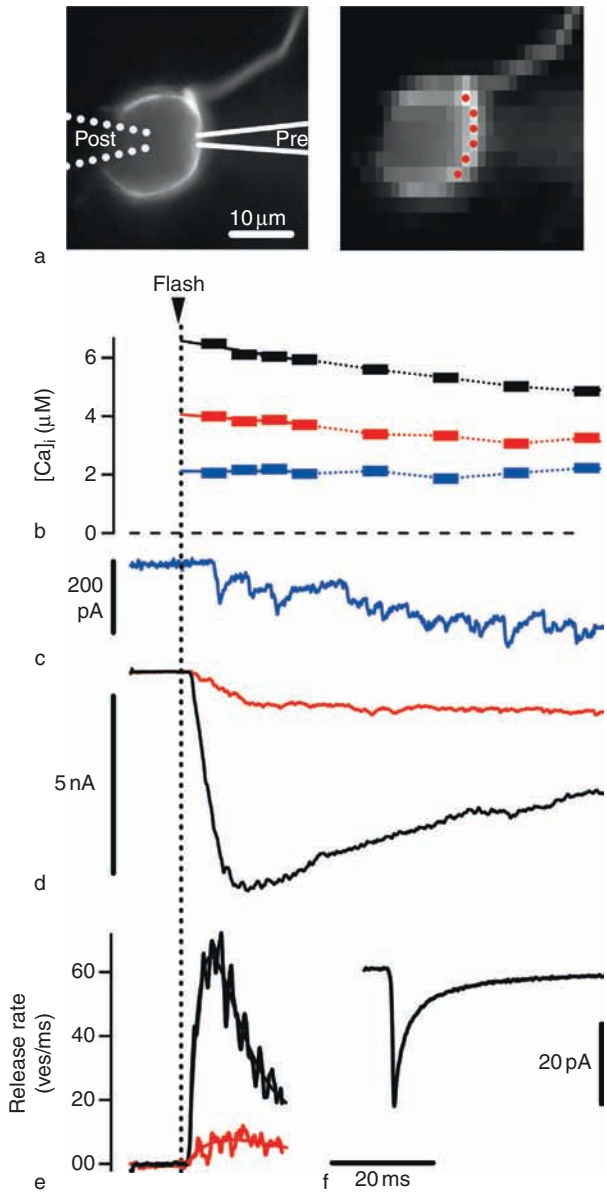

Figure 2 Presynaptic Ca^{2+} uncaging at the calyx of Held. (a) High-resolution image (left) and binned image (right) of a calyx of Held filled via a recording whole cell patch clamp pipette with the Ca^{2+}-sensitive fluorescent indicator, fura-2FF, and with the photolyzable Ca^{2+} chelator, DM-nitrophen. Pixel binning (see right image) was used during the fluorometric measurements of the intracellular Ca^{2+} concentration ($[Ca^{2+}]_i$; see (b)) to enable a high time resolution. (b) Steplike presynaptic $[Ca^{2+}]_i$ elevations in response to three UV light flashes with different intensities. (c, d) Excitatory postsynaptic currents (EPSCs) measured simultaneously by a second whole cell patch clamp recording of the postsynaptic cell. The traces that correspond to the presynaptic $[Ca^{2+}]_i$ elevations shown in (b) are identified by identical colors. Note that with presynaptic $[Ca^{2+}]_i$ steps of increasing amplitude, the EPSCs increase in amplitude and their rise times decrease. (e) Transmitter release rates as analyzed by deconvolution of the two EPSC traces shown in (d). Note that the two $[Ca^{2+}]_i$ steps to 4 and 6.5 μM amplitude (see (b), red and black trace) produced peak transmitter release rates with a sixfold difference, showing the highly supralinear action of Ca^{2+} in triggering synaptic vesicle fusion. (f) Averaged miniature EPSC trace from 436 single events. The kinetic parameters of this mEPSCs were used for the deconvolution of the evoked EPSC responses (see (d) and (e)). Reproduced from Schneggenburger R and Neher E (2000) Intracellular calcium dependence of transmitter release rates at a fast central synapse. *Nature* 406: 889–893.

the EPSCs were deconvolved with the average waveform of the miniature EPSC (Figure 2(f)). Miniature EPSCs (mEPSCs) likely represent the spontaneous fusion of a single presynaptic vesicle. The transmitter release rate traces obtained from the deconvolution showed peak release rates of 10 and 60 ves/ms (Figure 2(e)), which were evoked by presynaptic $[Ca^{2+}]_i$ steps to 4 and 6.5 μM, respectively (Figure 2(e), red and black traces). This analysis thus establishes the relationship between transmitter release rate and the intracellular Ca^{2+} concentration in a presynaptic nerve terminal of the CNS.

Figure 3(a) shows a plot of peak release rates, as analyzed in Figure 2(e), as a function of the $[Ca^{2+}]_i$ step in the nerve terminal, measured in Ca^{2+} uncaging experiments such as that illustrated in Figure 2. Figure 3(b) shows a similar plot of the Ca^{2+} sensitivity of transmitter release rate at the calyx of Held, which has been established independently by another group. In the plot of Figure 3(b), the transmitter release rates were divided by the estimated number of readily releasable vesicles (~1000), thus obtaining release rates per readily releasable vesicle. The measurements shown in Figures 3(a) and 3(b) clearly establish that fast vesicle release rates of approximately 1000 s^{-1} can be reached at 10–25 μM $[Ca^{2+}]_i$.

The local $[Ca^{2+}]_i$ signal experienced by an average readily releasable vesicle at the calyx of Held during the physiological stimulus (a presynaptic action potential) has been estimated to be a brief (~0.5 ms) rise of $[Ca^{2+}]_i$ to 10–25 μM based on kinetic modeling. This was done by first fitting the relationship between

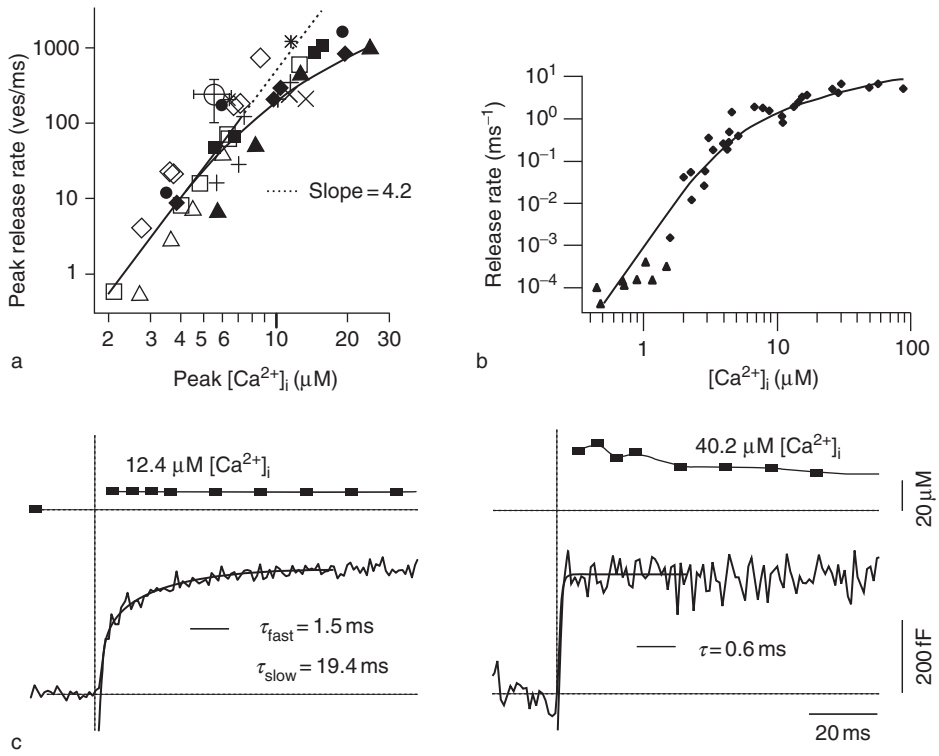

Figure 3 Intracellular Ca^{2+} dependency of the rate of transmitter release at the calyx of Held. (a) Plot of the peak transmitter release rate as a function of the peak $[Ca^{2+}]_i$ step as produced by Ca^{2+} uncaging in experiments similar to those shown in Figure 2. (b) A similar relationship obtained independently from Ca^{2+} uncaging experiments at the calyx of Held (taken from Bollmann et al., 2000, *Science* 289, 953–957). Here, the release rates were normalized to the estimated number of readily releasable vesicles (~1000). Note that high rates of transmitter release of approximately 1000 s^{-1} per readily releasable vesicle are already reached at 10 μM $[Ca^{2+}]_i$. In both (a) and (b), the relationship between transmitter release and $[Ca^{2+}]_i$ was fitted with a kinetic model of Ca^{2+} binding and vesicle fusion (lines). (c) Membrane capacitance measurements at the calyx of Held after presynaptic Ca^{2+} uncaging. During the early times after a stimulus shown here, membrane capacitance measurements report cumulative vesicle fusion. Note that with $[Ca^{2+}]_i$ steps above 30 μM, the rise of the membrane capacitance signal is faster than 1 ms, as estimated from exponential fits. This confirms that fast rates of vesicle fusion (~1000 s^{-1}, equivalent to a time constant of 1 ms) are reached at moderate $[Ca^{2+}]_i$ at the calyx of Held. (a) Reproduced from Schneggenburger R and Neher E (2000) Intracellular calcium dependence of transmitter release rates at a fast central synapse. *Nature* 406: 889–893. (b) Reproduced from Bollmann J, Sakmann B, and Borst J (2000) Calcium sensitivity of glutamate release in a calyx-type terminal. *Science* 289: 953–957. (c) Reproduced from Wölfel M and Schneggenburger R (2003) Presynaptic capacitance measurements and Ca^{2+} uncaging reveal submillisecond exocytosis kinetics and characterize the Ca^{2+} sensitivity of vesicle pool depletion at a fast CNS synapse. *Journal of Neuroscience* 23: 7059–7068.

transmitter release rates and $[Ca^{2+}]_i$ with a minimal kinetic model of Ca^{2+} binding and vesicle fusion (Figures 3(a) and 3(b), lines). Because of the steep relationship between transmitter release rates and $[Ca^{2+}]_i$, with a slope of approximately 4 or 5 in double logarithmic coordinates (Figures 3(a) and 3(b)), five Ca^{2+} binding sites needed to be postulated for the kinetic models. After fixing the parameters of the model, the local $[Ca^{2+}]_i$ compatible with the measured Ca^{2+} sensitivity of vesicle fusion as measured in the Ca^{2+} uncaging experiment was then calculated. It was found that an intracellular Ca^{2+} signal of 10–25 μM amplitude with brief half-duration (~0.5 ms) experienced by an average readily releasable vesicle could explain the physiological waveform of transmitter release at the calyx of Held.

The relatively high rates of vesicle fusion at the calyx of Held reached at surprisingly moderate $[Ca^{2+}]_i$ have been confirmed by combined presynaptic Ca^{2+} uncaging and membrane capacitance measurements (Figure 3(c)). This approach is similar to the one used in neuroendocrine cells, bipolar cells, and inner hair cells and thus allows direct comparison of the rates of exocytosis measured at the calyx of Held to that estimated in the other cell types (Table 1). The time constant of the fastest component in the membrane capacitance rise was approximately 2 ms at 15 μM $[Ca^{2+}]_i$, and it was less than 1 ms at >30 μM $[Ca^{2+}]_i$. These measurements therefore indicate secretory rates of $>1000\,s^{-1}$ that are reached at approximately 30 μM $[Ca^{2+}]_i$ and thus confirm the estimates based on measurements of EPSCs at the calyx of Held (Figures 3(a) and 3(b)). Therefore, at the calyx of Held nerve terminal, fast rates of exocytosis ($\sim 1000\,s^{-1}$) are reached at significantly lower Ca^{2+} concentrations (~10–30 μM $[Ca^{2+}]_i$) than at the ribbon-type synapses of retinal bipolar cells and inner hair cells, where comparable rates are only reached at approximately 100 μM $[Ca^{2+}]_i$ (Table 1).

Molecular Determinants for Different Ca^{2+} Sensitivities of Vesicle Fusion

The molecular differences responsible for the differences in Ca^{2+} sensitivity between different secretory cells and nerve terminals are not clearly understood. In neurons, the synaptotagmin isoform 1 or 2 is thought to be the most important Ca^{2+} sensor which initiates vesicle fusion. However, these genes are not expressed in inner hair cells of the cochlea, where other proteins with C2 domain Ca^{2+} binding sites might take over the role of the Ca^{2+} sensor for vesicle fusion. These molecular differences may explain the different intracellular Ca^{2+} sensitivities of vesicle fusion in inner hair cells and in the calyx of Held (Table 1). It is also not known whether other CNS nerve terminals have a Ca^{2+} sensitivity for vesicle fusion comparable to that of the calyx of Held, or whether they have a lower Ca^{2+} sensitivity comparable to that observed in the ribbon-type synapses (Table 1). It is possible that synapses with different forms of short-term plasticity (facilitation as opposed to depression) have different Ca^{2+} sensitivities of vesicle fusion. At neuromuscular junctions of crustaceans, a difference in the intracellular Ca^{2+} sensitivity of transmitter release has been found between two types of synapses that show different short-term plasticities (facilitation as opposed to depression).

Ca^{2+} Cooperativity in Vesicle Fusion

An important property of the action of Ca^{2+} in triggering vesicle fusion is its high cooperativity. Studies that have used Ca^{2+} uncaging to measure the intracellular Ca^{2+} sensitivity of vesicle fusion have all shown a highly supralinear relationship between the rate of vesicle fusion and the intracellular Ca^{2+} concentration which drives vesicle fusion (Table 1). When release rates were plotted as a function of $[Ca^{2+}]_i$ in double-logarithmic coordinates (Figures 3(a) and 3(b)), a region of extreme steepness (~4 or 5) became apparent in the range of 2–8 μM $[Ca^{2+}]_i$. This indicates that the Ca^{2+} sensor(s) of a given readily releaseable vesicle has to bind at least four or five Ca^{2+} ions before the vesicle can fuse with the plasma membrane. In kinetic models of Ca^{2+} binding and vesicle fusion, five Ca^{2+} ions have been found necessary to adequately describe the steep Ca^{2+} dependency of transmitter release from the calyx of Held and from inner hair cells, which both release glutamate from small synaptic vesicles. However, in chromaffin cells, which release catecholamines from larger storage vesicles, the intrinsic Ca^{2+} cooperativity is somewhat smaller (Table 1).

It has been shown at the calyx of Held synapse that below 2 μM $[Ca^{2+}]_i$, the steepness in the relationship between transmitter release and $[Ca^{2+}]_i$ is gradually reduced until a slope value of approximately 1 is reached close to the resting $[Ca^{2+}]_i$ of the nerve terminal. This indicates that at low $[Ca^{2+}]_i$, vesicles can fuse at low rates even when less than five Ca^{2+} ions are bound to the Ca^{2+} sensor for vesicle fusion. The low cooperativity of Ca^{2+} in inducing vesicle fusion close to resting $[Ca^{2+}]_i$ is probably of physiological relevance, because it prevents a large rise in asynchronous transmitter release, which would occur in response to a weak rise of basal $[Ca^{2+}]_i$ if a Ca^{2+}-sensing mechanisms with high cooperativity operated close to resting $[Ca^{2+}]_i$. Thus, the low cooperativity of Ca^{2+} in inducing vesicle fusion close to resting

$[Ca^{2+}]_i$ probably protects the nerve terminal from excessive loss of transmitter during small variations of the resting $[Ca^{2+}]_i$.

What might be the advantage of the high cooperativity of Ca^{2+} in inducing vesicle fusion at synapses? The answer might lie in an optimal signal-to-background 'noise' ratio for synaptic transmission. At the calyx of Held, it has been found that during a presynaptic action potential, the peak transmitter release rate is 10^5–10^6 fold higher than the spontaneous release rate as measured by the frequency of mEPSC events. Thus, the relative increase in transmitter release rate during an action potential is much larger than the relative increase in presynaptic $[Ca^{2+}]_i$ by which it is triggered (from a basal value of ~50 nM to ~10–25 μM local $[Ca^{2+}]_i$). The strong, 10^5–10^6 fold increase in peak transmitter release rate during an action potential can only be achieved with a highly nonlinear transduction mechanism. Thus, the nonlinear, cooperative action of Ca^{2+} at the molecular machinery that triggers vesicle fusion might have evolved to allow a high dynamic range of peak transmitter release rate compared to background release at the synapse.

The high cooperativity of Ca^{2+} in inducing vesicle fusion is also an important factor during short-term plasticity of synaptic transmission and for the modulation of transmitter release by presynaptic metabotropic receptors. Small changes in the amount of presynaptic Ca^{2+} influx that might occur during short-term plasticity or after metabotropic receptor activation will be translated into relatively large changes in transmitter output because of the highly supralinear relationship between Ca^{2+} and transmitter release.

See also: Active Zone; Calcium Channels; Calcium Channel Subtypes Involved in Neurotransmitter Release; Glutamate; Large Dense Core Vesicles (LDCVs); Neurotransmitter Release: Synchronous and Asynchronous; Synaptic Plasticity: Short-Term Mechanisms; Synaptic Vesicles; Synaptotagmins; Ultrastructural Organization of Release Sites in the Calyx of Held; Vesicle Pools; Voltage-Gated Calcium Channels.

Further Reading

Beutner D, Voets T, Neher E, and Moser T (2001) Calcium dependence of exocytosis and endocytosis at the cochlear inner hair cell afferent synapse. *Neuron* 29: 681–690.

Bollmann J, Sakmann B, and Borst J (2000) Calcium sensitivity of glutamate release in a calyx-type terminal. *Science* 289: 953–957.

Borst JGG and Sakmann B (1996) Calcium influx and transmitter release in a fast CNS synapse. *Nature* 383: 431–434.

Heidelberger R, Heinemann C, Neher E, and Matthews G (1994) Calcium dependence of the rate of exocytosis in a synaptic terminal. *Nature* 371: 513–515.

Heinemann C, Chow RH, Neher E, and Zucker RS (1994) Kinetics of the secretory response in bovine chromaffin cells following flash photolysis of caged Ca^{2+}. *Biophysical Journal* 67: 2546–2557.

Katz B (1969) The Release of Neural Transmitter Substances. Liverpool, UK: Liverpool University Press.

Lou X, Scheuss V, and Schneggenburger R (2005) Allosteric modulation of the presynaptic Ca^{2+} sensor for vesicle fusion. *Nature* 435: 497–501.

Millar AG, Zucker RS, Ellis-Davies GCR, Charlton MP, and Atwood HL (2005) Calcium sensitivity of neurotransmitter release differs at phasic and tonic synapses. *Journal of Neuroscience* 25: 3113–3125.

Neher E (1998) Vesicle pools and Ca^{2+} microdomains: New tools for understanding their roles in neurotransmitter release. *Neuron* 20: 389–399.

Roberts WM (1994) Localization of calcium signals by a mobile calcium buffer in frog saccular hair cells. *Journal of Neuroscience* 14: 3246–3262.

Schneggenburger R and Neher E (2000) Intracellular calcium dependence of transmitter release rates at a fast central synapse. *Nature* 406: 889–893.

Simon SM and Llinás RR (1985) Compartmentalization of the submembrane calcium activity during calcium influx and its significance in transmitter release. *Biophysical Journal* 48: 485–498.

Thomas P, Wong JG, Lee AK, and Almers W (1993) A low affinity Ca^{2+} receptor controls the final steps in peptide secretion from pituitary melanotrophs. *Neuron* 11: 93–104.

Voets T (2000) Dissection of three Ca^{2+}-dependent steps leading to secretion in chromaffin cells from mouse adrenal slices. *Neuron* 28: 537–545.

Wölfel M and Schneggenburger R (2003) Presynaptic capacitance measurements and Ca^{2+} uncaging reveal submillisecond exocytosis kinetics and characterize the Ca^{2+} sensitivity of vesicle pool depletion at a fast CNS synapse. *Journal of Neuroscience* 23: 7059–7068.

Extinction: Anatomy

K A Corcoran and G J Quirk, University of Puerto Rico School of Medicine, San Juan, Puerto Rico

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Extinction occurs when a conditioned stimulus (CS) is repeatedly presented in the absence of reinforcement; it is measured as a decline in the frequency and amplitude of conditioned responses. Current theories of extinction suggest that this decline reflects the development of an inhibitory memory that competes with the conditioning memory to drive behavior. Like other forms of memory, extinction consists of three phases: acquisition, consolidation, and retrieval (Figure 1). Acquisition of extinction is evident as decreased responding to the CS across an extinction session. Consolidation begins some time during the extinction session and continues for several hours afterward to form a more stable, long-term memory trace. Retrieval of extinction is triggered when the CS is once again presented after the extinction session. Research on the mechanisms of extinction has pinpointed many of the molecular mechanisms involved in these processes and has identified a critical neural circuit, which includes the amygdala, medial prefrontal cortex (mPFC), and hippocampus.

Amygdala

Acquisition of Pavlovian fear conditioning depends on cellular and molecular processes in the basolateral complex of the amygdala (BLA), which includes the basal and lateral amygdalar nuclei. After pairing a tone CS with a footshock unconditioned stimulus (US), for example, neurons in the lateral amygdala (LA) demonstrate increased firing in response to the CS. This increased activity reflects *N*-methyl-D-aspartate (NMDA) receptor- and protein synthesis-dependent synaptic plasticity. Lesions in or inactivation of the amygdala prevent the acquisition of conditioned fear as well as the expression of previously acquired fear memories. Thus, the amygdala is considered an essential site for the acquisition, consolidation, and retrieval of conditioned fear memories.

Given the critical role for the BLA in fear learning, it is not surprising that it is also involved in the inhibitory learning underlying extinction. Accumulating evidence suggests that the BLA mediates the acquisition of fear extinction, similar to its role in the acquisition of fear conditioning. Extinction of conditioned fear is similar to fear acquisition in that it requires the activation of glutamate receptors in the BLA. Within-session extinction is prevented by intra-BLA blockade of NMDA or metabotropic glutamate receptors. Extinction training engages several intracellular signaling cascades in the BLA, as indicated by increased levels of phosphorylated mitogen-activated protein kinase (MAPK) and phosphatidylinositol (PI)-3 kinase. Disruption of MAPK or PI-3 kinase activity in the BLA shortly after extinction training prevents the subsequent retrieval of extinction memory. These signaling pathways lead to the activation of immediate early genes such as *c-fos* and *zif*; failure to activate these genes is correlated with poor extinction learning. Within-session extinction of appetitive instrumental responses is blocked by inactivation of the BLA, suggesting that the BLA is similarly involved in the acquisition of extinction for other forms of affective learning.

Consolidation of extinction is also reliant on synaptic plasticity in the BLA. Protein synthesis inhibition in the BLA prevents extinction recall in conditioned fear and taste aversion paradigms. Shortly after extinction training, increased levels of mRNA for the neurotropic factor brain-derived neurotropic factor (BDNF) are observed in the BLA, indicating that structural changes help stabilize the extinction memory. Interestingly, some synaptic changes underlying extinction appear to directly oppose some of the plastic changes that underlie conditioning. For example, fear conditioning decreases levels of the γ-aminobutyric acid (GABA)-clustering protein gephyrin in the BLA, whereas extinction training is correlated with an upregulation of gephyrin and an increase in the surface expression of GABA receptors. Thus, extinction memories may be represented as increases in inhibitory neurotransmission within the amygdala. Similarly, extinction can reverse conditioning-induced phosphorylation of the transcription factor cyclic adenosine monophosphate (cAMP) responsive element-binding protein (CREB). Extinction training increases the activity of the phosphatase calcineurin, which dephosphorylates CREB. This reversal suggests that, to a certain extent, extinction might cause an erasure of the conditioning memory stored in the BLA.

Medial Prefrontal Cortex

The mPFC is a higher-order processing center involved in many cognitive functions, such as working memory and decision making. The mPFC consists of several interrelated subregions; the two that have been studied most thoroughly with regard to fear and extinction are the infralimbic (IL) and prelimbic

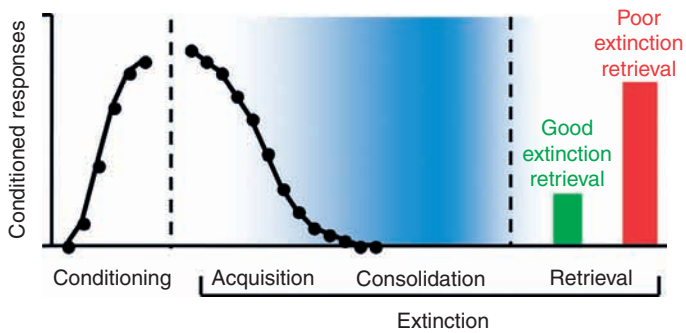

Figure 1 Extinction is a learning process that occurs in three phases. During extinction training, when conditioned stimuli (CSs) are presented without any reinforcement, the frequency and magnitude of conditioned responses decline. Consolidation of the extinction memory begins some time during extinction training and continues for several hours after the end of the extinction session. When the CS is again presented sometime later, extinction is retrieved, as evidenced by low levels of conditioned responding. A failure to retrieve extinction (high levels of responding) occurs when CS is presented outside the extinction context.

(PL) subregions. Activity in these areas is intimately linked to amygdala function, fear expression, and extinction.

Rats with lesions of the IL subregion of mPFC are able to learn extinction within a single session, suggesting that the IL is not involved in the within-session acquisition of extinction. However, they are unable to retrieve the extinction memory the following day, demonstrating high levels of fear. Consistent with this, inactivation of the IL prior to extinction training also prevents extinction retrieval. Thus, activity in the IL during or after an extinction session is important for later extinction retrieval. Consistent with a role for the IL in consolidation of the extinction memory, local infusion of protein synthesis inhibitors or NMDA antagonists in the IL disrupts extinction retrieval the following day.

Although manipulations of the mPFC disrupt the consolidation of extinction, there is evidence that the extinction memory itself is not completely lost. When rats are reextinguished during testing via additional CS presentations, extinction proceeds at a faster rate than during the original extinction session. This faster rate of reextinction is evidence of savings of the original extinction memory. Thus, the IL is not necessary for the consolidation of the extinction memory itself. Rather, extinction-related plasticity in the IL is necessary for successfully retrieving extinction at a later time. In other words, the IL stores a memory that facilitates the retrieval of the extinction memory stored elsewhere.

One mechanism by which the IL encodes this extinction retrieval memory is by increased bursting of IL neurons. Bursting measured shortly after extinction training is correlated with the degree of extinction recall the following day. Bursting in the IL is dependent on NMDA receptors, suggesting that it may contribute to the potentiation of IL activity following extinction. During retrieval of extinction, the IL shows increases in unit activity, evoked potentials, metabolic activity, and c-Fos expression. In each case, the degree of potentiation is correlated with the amount of extinction recalled. Thus, it is believed that IL neurons specifically signal extinction retrieval. A role for the IL in extinction retrieval is also evident in studies of instrumental appetitive responses. Lesions of the IL enhance the renewal of bar-press responses outside the extinction context and the spontaneous recovery of bar-press responses after extinction.

The IL projects to both the amygdala and the amygdala's targets, and it is therefore well situated to modulate fear responses. Consistent with this, IL activity has an inhibitory influence on the amygdala. Neurons in the IL project to the intercalated GABAergic neurons (ITCs) situated between the BLA and the central nucleus of the amygdala. Stimulation of the IL activates ITCs, which subsequently inhibit neurons projecting from the central nucleus to other brain regions involved in the generation of fear responses. Stimulation of the IL also reduces the expression of conditioned fear responses, supporting a role for this pathway in fear inhibition.

Contrary to the inhibitory effects of the IL, the PL augments fear expression. Lesions or inactivation of the PL diminishes the expression of conditioned fear responses. In contrast, microstimulation of the PL increases freezing in response to CSs. Like the IL, the PL is intimately connected with the amygdala, but its targets are somewhat different. The PL projects to the BLA, and activity in the PL drives BLA neurons. The BLA projects to output neurons of the central nucleus of the amygdala. The PL could therefore be involved in the renewal or spontaneous recovery of fear after extinction. Thus, because of its ability to bidirectionally modulate amygdala activity, the mPFC can regulate both fear conditioning and extinction.

Hippocampus

Unlike the retrieval of fear memory, the retrieval of extinction memory is strongly regulated by context. When a CS is presented outside the context in which extinction training occurred, extinction memory is not retrieved and fear responses are renewed. The hippocampus plays an important role in many forms of contextual learning and in using contextual cues to aid memory retrieval. As a result, it has been the focus of several studies exploring how contexts modulate extinction retrieval.

Hippocampal inactivation does not prevent the acquisition of extinction. The next day, however, extinction retrieval in these rats is deficient, and they demonstrate higher levels of freezing than controls. The increase in freezing among these rats is equivalent to levels of renewed freezing among rats that extinguished drug-free and were tested outside the extinction context. Thus, during extinction training, the dorsal hippocampus acquires a memory of the extinction context that is used for subsequent retrieval of extinction. In contrast to preextinction inactivation of the hippocampus, inactivation prior to testing prevents fear renewal. Although the dorsal or ventral hippocampus is inactivated, rats show little freezing regardless of the context. These results suggest that, following extinction, the hippocampus uses contextual cues to determine whether fear or extinction memory is retrieved.

The pattern of effects seen with hippocampal inactivation mirrors the effects seen with mPFC inactivation, suggesting that these two structures work together to mediate extinction retrieval. The hippocampus sends dense projections to the mPFC, and hippocampal stimulation gates the response of mPFC neurons to other inputs. This pathway could trigger neural plasticity essential for extinction recall. In support of this, hippocampo-prefrontal connections are potentiated by extinction training. Thus, when a CS is presented in the extinction context, contextual information processed by the hippocampus permits the inhibition of fear expression via its connections with the mPFC (Figure 2). Activity of neurons in the BLA during extinction retrieval is less when the CS is presented inside the extinction context compared to outside. Inactivation prevents this contextual modulation of LA activity. Thus, the hippocampus is capable of modulating LA responsiveness to CSs, either directly or indirectly.

The hippocampus is necessary for context processing in a variety of learning and memory paradigms. In each of these, the hippocampus is also necessary for extinction. Acquisition of extinction of context fear requires NMDA receptors in the hippocampus, as well as various protein kinase signaling pathways, and the rearrangement of actin, a cytoskeletal protein. Another form of context learning is inhibitory avoidance, in which a rat learns to avoid stepping onto an electrified floor. Extinction of this behavior similarly requires protein kinase activation and protein synthesis in the hippocampus. Thus, the role of the hippocampus in the acquisition and consolidation of extinction for contextual memories is very much like the role of the amygdala in the extinction of cued fear conditioning.

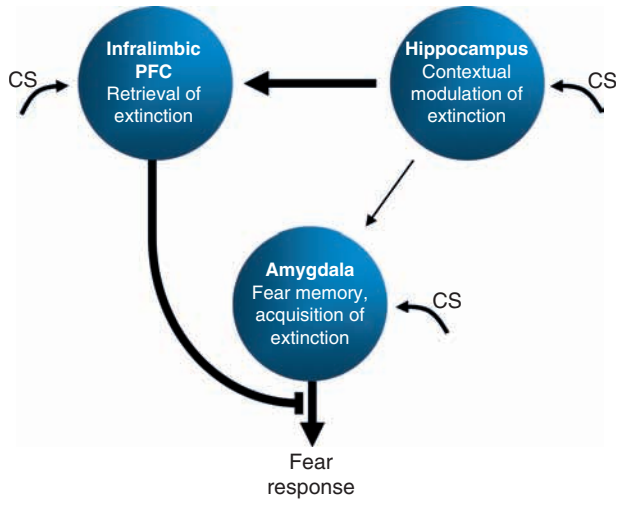

Figure 2 Retrieval of fear extinction involves the amygdala, medial prefrontal cortex, and hippocampus. The amygdala acquires, stores, and expresses fear memories. Plasticity in the amygdala is also necessary for acquisition of extinction memory. The infralimbic subregion of medial prefrontal cortex is necessary for the retrieval of extinction memory and the inhibition of amygdala output. The hippocampus processes contextual information surrounding CS presentation and uses this information to help determine whether fear or extinction memory will be retrieved. CS, conditioned stimulus; PFC, prefrontal cortex.

Periaqueductal Gray and Error Signaling

Predominant theories of how conditioning memories are acquired hold that learning is driven by a mismatch between what an organism expects in response to a CS and what the organism actually experiences. For example, early in fear conditioning, a rat does not expect a tone to be followed by footshock. The arrival of the footshock after the tone violates this expectancy, however, and the rat quickly learns the tone–shock association. Similarly, early in extinction training, the rat expects the tone to be followed by a shock that never comes. It is thought that the error between expectancy and actual outcome during both fear and extinction learning is signaled by opioids in the midbrain periaqueductal gray (PAG).

Opioid antagonists infused directly into the PAG facilitate the acquisition of fear conditioning, but disrupt the acquisition of extinction. Likewise, increasing opioids in PAG disrupts fear acquisition and enhances extinction acquisition. Thus, in fear conditioning, opioids signal the degree to which a tone is expected to be followed by a shock. Early in extinction training, opioid release in response to the CS is high, but the expected shock never comes. This discrepancy between opioid-mediated expectancy of a shock and the actual nonreinforcement provides a mechanism to initiate the inhibitory learning that occurs as a result of extinction training. After opiod receptor blockade in PAG, the error signal that drives extinction learning is missing and within-session extinction is prevented.

Anatomy of Extinction in Humans

Neuroimaging studies of extinction learning in humans indicate that the network of structures involved in extinction learning is conserved across species.

In humans, both the amygdala and mPFC are active during extinction training. Consistent with rat findings, both the hippocampus and mPFC are activated during extinction retrieval, and this activity is correlated with the amount of extinction retrieved. Moreover, this activity during extinction retrieval is modulated by contextual cues.

Extinction learning in humans falls into a broader category of emotional regulation processes. In the cognitive sense, emotional regulation can take the form of willfully suppressing emotional responses to stimuli or reappraising the meaning of the stimuli themselves to either reduce or enhance emotional reactions. Intentional reduction of behavioral responses to emotional stimuli can facilitate cognitive reappraisals of those stimuli. Neuroimaging studies show that reappraisal of negative stimuli often decreases amygdala activity and increases activity in higher cognitive centers, such as medial and lateral prefrontal regions.

The role of the mPFC and hippocampus in extinction retrieval suggests that these structures are deficient in posttraumatic stress disorder (PTSD), which is characterized by inappropriate fear responses. In fact, PTSD patients show reduced volume and reduced activity in the mPFC and hippocampus when reexposed to traumatic reminders. As a result, PTSD subjects show exaggerated fear responses, which may reflect an inability to retrieve extinction memory. Although this remains to be confirmed experimentally, it is likely that pharmacological agents designed to strengthen extinction retrieval will help control PTSD and other disorders stemming from reduced inhibition.

See also: Amygdala: Structure and Circuitry in Primates; Amygdala: Structure and Circuitry in Rodents and Felines; Amygdala: Contributions to Fear; Aversive Emotions: Molecular Basis of Unconditioned Fear; Fear Conditioning and Synaptic Plasticity; Genetics of Human Anxiety and Its Disorders; Hippocampus; Pharmacology of Fear Extinction; Posttraumatic Stress Disorder: Neurobiology; Posttraumatic Stress Disorder as an Emotional Disorder; Prefrontal Cortex: Structure and Anatomy.

Further Reading

Burgos-Robles A, Vidal-Gonzalez I, Santini E, and Quirk GJ (2007) Consolidation of fear extinction requires NMDA receptor-dependent bursting in the ventromedial prefrontal cortex. *Neuron* 53: 871–880.

Chhatwal JP, Stanek-Rattiner L, Davis M, and Ressler KJ (2006) Amygdala BDNF signaling is required for consolidation but not encoding of extinction. *Nature Neuroscience* 9: 870–872.

Delamater AR (2004) Experimental extinction in Pavlovian conditioning: Behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology B* 57: 97–132.

Maren S and Quirk GJ (2004) Neuronal signalling of fear memory. *Nature Reviews Neuroscience* 5: 844–852.

Milad MR, Wright CI, Orr SP, Pitman RK, Quirk GJ, and Rauch SL (2007) Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological Psychiatry* 62(5): 446–454.

Myers KM and Davis M (2007) Mechanisms of fear extinction. *Molecular Psychiatry* 12: 120–150.

Paré D, Quirk GJ, and LeDoux JE (2004) New vistas on amygdala networks in conditioned fear. *Journal of Neurophysiology* 92: 1–9.

Phelps EA and LeDoux JE (2005) Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* 48: 175–187.

Quirk GJ (ed.) (2006) *Special Issue: Extinction Biological Psychiatry* 60(4).

Quirk GJ, Garcia R, and Gonzalez-Lima F (2006) Prefrontal mechanisms in extinction of conditioned fear. *Biological Psychiatry* 60: 337–343.

Quirk GJ and Mueller D (2008) Neural mechanisms of extinction learning and retrieval. *Neuropsychopharmacology Reviews* 33: 56–72.

Extracellular Matrix Molecules: Synaptic Plasticity and Learning

A Dityatev, A Irintchev, F Morellini, and M Schachner, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The importance of the extracellular matrix (ECM) in synaptic organization has been known for decades from structural and molecular studies of the neuromuscular junction. A striking feature of this synapse is its prominent ECM basal lamina that is interposed in its synaptic cleft. It contains several ECM molecules, of which the heparan sulfate proteoglycan agrin and laminin proved to be required for different aspects of post- and presynaptic synaptic differentiation of the neuromuscular junction.

Conspicuous aggregates of the ECM molecules in the central nervous system appear as so-called perineuronal nets, which surround axon initial segments and cell bodies and proximal dendrites of some neurons in a meshlike structure that interdigitates with synaptic contacts and astrocytic processes. Perineuronal nets are heterogeneous in structure and composition, incorporating molecules secreted from both neurons and astrocytes. Apart from these prominent ECM structures, immunoreactivity for many ECM molecules appears as diffuse staining, documenting the widespread occurrence of these molecules in the extrasynaptic space, the structural and molecular organization of which remains to be elucidated. Studies indicate that ECM molecules such as thrombospondins, reelin, tenascin-R (TN-R), chondroitin sulfate proteoglycan versican, neuronal pentraxins NP1 and NP2, and integrin ligands may shape pre- and postsynaptic differentiation of excitatory and inhibitory synapses in the central nervous system.

Not only formation and maturation of synapses but also synaptic plasticity, learning, and memory in adults are dependent on the ECM. These functions of several ECM molecules are the focus of this article.

Integrins and Synaptic Plasticity

Integrins are transmembrane heterodimeric receptors to ECM molecules, including fibronectin, laminins, and thrombospondins. There are 16 different α and eight different β mammalian integrin subunits, which associate to form $\alpha\beta$ heterodimers; many of these are expressed in the developing and adult hippocampus. Initial evidence for a potential role of integrins in hippocampal synaptic plasticity was presented in 1990 by reducing the stability of hippocampal long-term potentiation (LTP), a widely used model of synaptic plasticity, using broad-spectrum peptide ligands of integrins. Further support for integrin function in LTP was gained by experiments with injection of snake toxins (disintegrins), which preferentially inhibit the binding of ECM ligands to β_1- or β_3-containing integrins, and blocking antibodies to the α_3 integrin subunit. These treatments were efficient in blocking the stabilization of LTP and facilitation of depotentiation (i.e., reversal of LTP by low-frequency stimulation). Since laminin also modulates stabilization of LTP and the α_3 subunit was found to be associated with the β_1 subunit in synaptosomal preparations, it is plausible to assume that interactions between $\alpha_3\beta_1$ integrin receptors and laminins are important for stabilization of LTP.

The critical role of integrins in synaptic plasticity is also supported by analysis of mouse mutants deficient in integrins. Heterozygous mice with genetically reduced expression of the α_3 subunit fail to maintain CA1 LTP. Mice with reduced expression of the α_3 and α_5 subunits also exhibit a defect in paired-pulse facilitation. Furthermore, conditional mutants lacking the β_1 subunit in excitatory neurons in the postnatal forebrain have impaired synaptic transmission through α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors and diminished *N*-methyl-D-aspartate (NMDA) receptor-dependent LTP. These results demonstrate that multiple integrins are involved in synaptic plasticity.

The potential mechanisms by which integrins may affect induction of synaptic plasticity appear to involve phosphorylation of glutamate receptors. Microperfusion of a peptide integrin ligand caused an approximately twofold increase in the amplitude and duration of NMDA receptor-gated synaptic currents. Biochemical analysis demonstrated that the peptide and the native ligand fibronectin trigger β_1-dependent rapid and robust increases in tyrosine phosphorylation of focal adhesion kinase, proline-rich tyrosine kinase 2, and Src family kinases. Pharmacological experiments also support the view that integrin-triggered Src activity underlies tyrosine phosphorylation of NMDA receptor subunits NR2A and NR2B (Figure 1).

To maintain postsynaptic changes induced by synaptic stimulation, the mechanisms regulating activity, number, or composition of synaptic AMPA receptors have to come into play. In this context, it is noteworthy that infusion of the peptide integrin ligand into

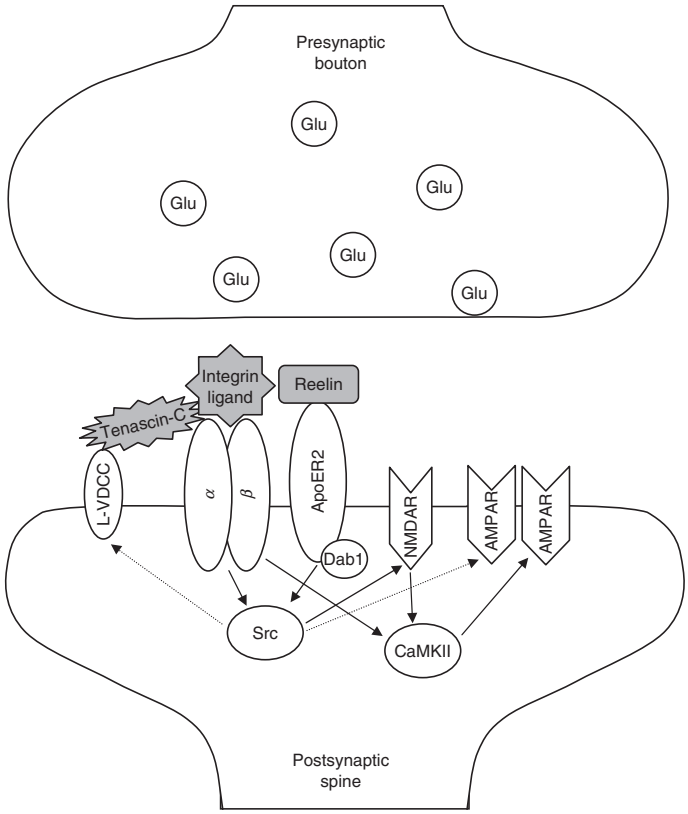

Figure 1 A summary of mechanisms by which ECM molecules may regulate synaptic plasticity. Reelin and integrin ligands may bind to their cell surface receptors (ApoER2 and $\alpha\beta$ integrins) to trigger intracellular cascades via interaction with Dab1 and activation of tyrosine kinases of the Src family, which result in tyrosine phosphorylation of effectors, including the NMDA receptor (NMDARs), AMPA receptors (AMPARs), and the L-type voltage-dependent Ca^{2+} channels (L-VDCC). Additionally, integrin ligands may activate Ca^{2+}/calmodulin-dependent kinase II, which regulates the number and activity of AMPA receptors. Tenascin-C affects L-VDCC via an unknown mechanism that may involve a direct interaction with channels or signaling via integrins.

rat hippocampal slices reversibly increased the slope and amplitude of excitatory postsynaptic potentials mediated by AMPA receptors. A mixture of function-blocking antibodies to integrin subunits α_3, α_5, and α_V reduced the ligand effects on synaptic responses. The effects were also blocked by inhibitors of Src tyrosine kinase, antagonists of NMDA receptors, and inhibitors of calcium calmodulin-dependent protein kinase II. Furthermore, changes in AMPA receptor-mediated responses were accompanied by phosphorylation of both the Thr286 site on calmodulin-dependent protein kinase II and the Ser831 site on the GluR1 subunit of the AMPA receptor. Thus, integrins may affect AMPA receptors via two signaling pathways, involving Src tyrosine kinases and calmodulin-dependent protein kinase II (Figure 1).

Integrins and Learning and Memory

The first evidence that integrins are involved in memory processes came from the Volado mutant of *Drosophila*, which mutated gene encodes two isoforms of α integrin: these mutants were not able to show short-term olfactory memory, an impairment that could be rescued by conditional expression of a Volado transgene in adulthood. It is noteworthy that, as observed in *Drosophila*, integrins regulate short-term memory also in mice. Although mice heterozygous for three α integrins (α_3, α_5, and α_8) and mice deficient for the β_1 integrins showed normal learning and long-term memory in the water maze and contextual fear conditioning tasks, these mice exhibited specific memory impairments soon after training. In fact, mice heterozygous for the three α integrins did not show any preference for the platform in a probe trial performed 1 h after training in the water maze test, whereas working memory in the nonmatching-to-place task was reduced in mice deficient for the β_1 integrins when they had to retain the information of the visited arm for 10 s. Thus, it seems that integrins modulate memory formation through short-term changes in synaptic transmission independent of protein synthesis.

Reelin and Induction of NMDA Receptor-Dependent LTP

Reelin is an ECM glycoprotein with a long and exciting history of research: it is named after the spontaneously occurring *reeler* mouse mutant, which shows abnormal migration of neurons during embryonic development. These deficits result in an outside–in instead of inside–out positioning of migrating neuronal cell bodies in the cerebral cortex. Strikingly, mice deficient in receptors for reelin, the very low-density lipoprotein (VLDL) receptor and ApoE receptor 2 (ApoER2), and cytoplasmic adaptor protein mDab1, which mediates signaling through VLDL and ApoER2 receptors, also display the *reeler* phenotype.

Several studies suggest that reelin together with its receptor, ApoER2, are involved in synaptic plasticity. Application of reelin to wild-type hippocampal slices produces a significant increase in LTP. The heterozygote reeler mouse shows impaired LTP and long-term depression (LTD). Application of receptor-associated protein RAP, a specific and broad-spectrum inhibitor of ligand binding to all low-density lipoprotein receptor family members, almost completely blocks LTP. A modest decrease in short-term potentiation is observed in VLDL receptor-deficient mice, although LTP in these mutants is nearly identical to that of wild-type mice. However, ApoER2-deficient mice are strongly impaired in LTP. Importantly, when reelin was applied to slices from VLDL receptor-deficient or ApoER2-deficient mice, the LTP-enhancing effect of reelin was abolished, supporting the view that reelin signaling via ApoER2 is critical for LTP.

Two studies uncovered some of the underlying mechanisms: ApoER2 is present in the postsynaptic densities of excitatory synapses, where it forms a functional complex with NMDA receptors. Reelin signaling through ApoER2 markedly enhances LTP through a mechanism that requires the presence of a stretch of amino acid residues in the intracellular domain of ApoER2 encoded by a single exon. This exon is alternatively spliced in an activity-dependent manner and is required for reelin-induced tyrosine phosphorylation of NMDA receptor subunits (Figure 1). Furthermore, a mouse mutant in which ApoER2 is mutated to disrupt its interaction with Dab1 lacks LTP, suggesting that signaling of ApoER2 via Dab1 is necessary for LTP. Since exogenously added reelin could significantly although only slightly increase the levels of LTP in this mutant, whereas reelin does not enhance LTP in ApoER2-deficient mice, it is plausible to assume that some LTP-promoting signaling can also occur independently of ApoER2–Dab1 binding. It may involve binding of other adaptor proteins such as JIP1/2 and PSD95 to exon 19 of reelin.

Reelin and Learning and Memory

Genetic analysis of mice deficient in either ApoER2 or VLDL receptor provided evidence that both reelin receptors regulate cognitive processes, assessed as contextual long-term memory in the fear conditioning task. Similarly, contextual but not cued memory in the fear conditioning task was reduced in mice heterozygous for the *reeler* mutation. The normal spatial learning and memory in the water maze task and minor deficit in olfactory discrimination learning of heterozygote reeler mice suggest that reelin regulates some but not all hippocampus-dependent cognitive functions. Since previous studies revealed that mice deficient in reelin, ApoER2 and VLDL receptor, and Dab1 have abnormal migration patterns, the same fundamental signaling pathway appears to regulate neuronal migration during brain development and also synaptic plasticity and contextual memory in the adult brain.

Tenascin-C and Induction of Synaptic Plasticity via Ca^{2+} Channels

Induction of some forms of synaptic plasticity depends on activity of the L-type voltage-dependent Ca^{2+} channels (L-VDCC), and there is evidence that this activity is under the influence of ECM glycoprotein tenascin-C (TN-C). In TN-C-deficient mice, theta-burst stimulation of CA3–CA1 synapses resulted in impaired LTP, whereas CA1 LTD was completely abolished. Recordings of LTP in the presence of nifedipine, an antagonist of L-VDCCs, did not show an effect on TN-C-deficient mice but reduced LTP in wild-type mice to the levels seen in the mutant, implying a link between L-VDCCs and TN-C in the regulation of synaptic plasticity. Furthermore, chemical induction of an L-VDCC-dependent LTP in the CA1 region by application of the K^+ channel blocker tetraethylammonium resulted in impaired LTP in TN-C-deficient mice. Because NMDA receptor-mediated responses and three forms of L-VDCC-independent LTP are normal in TN-C-deficient mice, the data point to a specific role for TN-C in L-VDCC-mediated signaling. These results are supported by data showing a strong reduction of LTP in the CA1 hippocampal area after injection of the TN-C fragment comprising fibronectin-like repeats 6–8. This fragment is known to bind to another ECM molecule, fibronectin, and thus could act via competition with endogenous TN-C, disrupting the targeting of TN-C to the fibronectin-containing ECM.

The hypothesis that TN-C affects signaling via L-VDCCs was verified by optical measurements of intracellular Ca^{2+} concentrations in CA1 pyramidal cells. The difference between Ca^{2+} signals triggered by

theta-burst stimulation under normal physiological solution and in the presence of nifedipine was used as a measure of Ca^{2+} influx via L-VDCCs. These measurements revealed a reduction in Ca^{2+} influx in TN-C-deficient mice. The underlying mechanisms may involve direct interaction between TN-C and Ca^{2+} channels or interaction of TN-C with integrins which are known to regulate activity of L-VDCCs (Figure 1).

TN-C and Learning and Memory

Possible roles of TN-C in learning and memory have been investigated in mice deficient for TN-C and by intrahippocampal injection of TN-C fragments containing the fibronectin type III repeats in mice tested in a cognitive task. Injection of a fragment containing the fibronectin type III repeats 6–8 into the hippocampus of mice inhibited memory retention in the one-trial step-down passive avoidance task. The observation that application of this fragment also reduced LTP in the CA1 region of hippocampal slices suggested that TN-C is involved in synaptic plasticity and required for memory consolidation in hippocampus-dependent cognitive tasks. Similarly, TN-C-deficient mice had an impaired performance in the step-down avoidance task. On the other hand, they performed as well as wild-type mice in a variety of tasks designed to measure hippocampus-dependent learning and memory, such as the water maze, contextual fear conditioning, step-through passive avoidance, and spontaneous alternation tasks. A detailed analysis of the behavior of TN-C-deficient mice strongly suggested that the higher novelty- and stress-induced active responses of these mutant mice account for their poorer performance in passive avoidance tasks, whereas learning is unaltered. This conclusion is in agreement with the fact that the reduced LTP observed in the CA1 region of TN-C-deficient slices has been interpreted as a consequence of lower activity of L-VDCCs, since pharmacological studies have shown that these channels are not required for acquisition or expression of conditioned fear in mice. On the other hand, L-VDCCs are involved in extinction of fear memory, which is the decline in the expression of a conditioned response following the withdrawal of reinforcement. We found that TN-C-deficient mice exhibit reduced short- and long-term extinction of fear memories, suggesting that TN-C regulates this cognitive process by modulating the activity of L-VDCCs.

TN-R, GABAergic Transmission, and Metaplasticity

Another example of how ECM molecules affect induction of LTP is provided by the glycoprotein TN-R. TN-R is enriched at the nodes of Ranvier in the central nervous system and in perineuronal nets and is essential for assembly of these nets since distribution of net-associated ECM molecules is altered in TN-R-deficient mice. Also, TN-R-deficient mutant mice show impaired CA1 LTP, elevated levels of excitatory synaptic transmission, elevated $GABA_B$ receptor-activated K^+ currents, and reduced levels of perisomatic inhibitory currents mediated by $GABA_A$ receptors. Several studies suggest that these abnormalities are related to the deficiency in the HNK-1 carbohydrate (a structure containing a 3′-sulfated glucuronic acid, which was first discovered on human natural killer cells – hence the name), which is prominently expressed by TN-R. Mice deficient in the glucuronyltransferase and HNK-1 sulfotransferase, which are the final pathway enzymes in the synthesis of the HNK-1 carbohydrate, showed a similar reduction in CA1 LTP as was found in TN-R-deficient mice. The HNK-1 antibody produced a strong and specific reduction in perisomatic inhibitory currents when applied to hippocampal slices via a relief of a constitutive block of postsynaptic $GABA_B$ receptors exerted by endogenous HNK-1 carbohydrate in perisomatic synapses. The functional link between the HNK-1 carbohydrate and TN-R as its cognate carrier molecule was demonstrated by experiments in which the antagonistic effect of the HNK-1 antibody on inhibitory currents was abolished in TN-R mutant mice.

Thus, TN-R/HNK-1 deficiency leads to reduction in GABAergic inhibition and impairment in LTP. This appears to be due to a 10 mV shift in the threshold for induction of LTP at CA3–CA1 synapses, which develops between the second and third weeks of postnatal development (i.e., during maturation of the GABAergic system). To evaluate whether changes in LTP are related to reduced $GABA_A$ and/or increased $GABA_B$ receptor-mediated currents in TN-R-deficient mice, we examined whether changes of these currents would lead to a recovery of LTP. Pretreatment of hippocampal slices with either a $GABA_B$ receptor antagonist or $GABA_A$ receptor agonists restored theta-burst-induced LTP in TN-R-deficient mice and normalized the threshold for pairing-induced LTP. These observations provide the first evidence that reduced levels of inhibition in TN-R-deficient mice may cause metaplastic changes in LTP.

Interestingly, elevation of $GABA_A$ receptor-mediated inhibition could also lead to impairment in LTP. For instance, there is an increase in inhibitory currents and impaired LTP in heparin-binding growth-associated molecule (HB-GAM) overexpressing mice, which can be rescued by an inhibitor of $GABA_A$ receptors. Thus, the ECM and its associated molecules may regulate

synaptic plasticity by bidirectionally setting the levels of GABAergic inhibition.

TN-R and Learning and Memory

In the water maze task, mice deficient for TN-R showed an altered searching strategy, swimming more in the proximity of the wall of the maze and less in the proximity of the platform compared to wild-type littermates, although it was suggested that exploratory behavior rather than cognitive abilities was altered in these mice. Because TN-R carries the carbohydrate HNK-1, it is extremely interesting that mice deficient for HNK-1 sulfotransferase showed the same subtle but significant alterations as those observed in TN-R-deficient mice in the water maze task. Moreover, HNK-1 sulfotransferase-deficient mice were not capable of finding the platform when tested after a long time interval and did not learn a new platform position when trained with a mass protocol that required fast re-learning rate. It would therefore be interesting to test long-term memory and fast re-learning rate of TN-R-deficient mice to determine whether HNK-1 and TN-R deficiencies lead to the same phenotype. In summary, the similar behavioral alteration of TN-R and HNK-1 sulfotransferase-deficient mice in a hippocampus-dependent cognitive task correlates with reduced LTP and enhanced basal excitatory synaptic transmission in the CA1 region.

Tenascins and Neural Circuitries

Beyond modulation of synaptic efficacy, ECM molecules may influence the formation, maintenance, and functional properties of neuronal and glial cell populations and neural circuitries, and thus regulate functions of cell ensembles underlying cognition. However, little experimental evidence supporting this notion is available. Studies on mice deficient in the expression of TN-R and TN-C provide some preliminary insights into how primary effects of synaptic functions or aberrant formation of cell populations may be linked to systemic dysfunction.

Monitoring of spontaneous and event-related electrical activity in awake TN-R-deficient mice has revealed enhanced hippocampal gamma oscillations, increased neocortical electroencephalogram (EEG) activity over a wide frequency range, and enhanced responses to auditory stimuli in the hippocampus and cerebral cortex compared to wild-type mice. Stereological estimates of major neuronal cell types in the cortex and hippocampus have provided no evidence for altered ratios between inhibitory and excitatory neurons. Thus, the enhanced excitability of neuronal ensembles in the TN-R-deficient mouse might be directly related to the aforementioned deficient inhibitory input to excitatory neurons. Abnormal operation of inhibitory networks may also underlie the alterations of gamma-frequency oscillations in the hippocampus. Since these oscillations provide a temporal framework for various forms of cognitive processing, a direct link between synaptic and behavioral abnormalities is plausible. Findings of reduced extracellular space and tortuosity in the neocortex and hippocampus of TN-R-deficient mice suggest another possibility for the direct influence of an ECM molecule on synaptic efficacy and extrasynaptic transmission – that is, by regulation of extracellular diffusion of neuroactive substances.

Similar to TN-R, ablation of TN-C causes enhancement of cortical event-related responses. Synaptic dysfunction that could readily explain this phenomenon, as opposed to TN-R deficiency, has not been identified. Quantitative morphological analysis has revealed abnormally high numbers of excitatory neurons and a reduced population of parvalbumin interneurons in the motor and sensory cortices of TN-C-deficient mice. Thus, abnormal physiological responses in the TN-C-deficient mouse may be related to an altered ratio of excitatory to inhibitory neurons which, as evidenced by synaptic coverage analysis, leads to reduced inhibitory input to pyramidal cell bodies. The origin of these abnormalities apparently in part lies in early neurogenesis in which TN-C, expressed in radial glia cells, is involved. However, the failure to compensate supernumerary neuronal numbers during later development and adulthood suggests that TN-C has other, yet unknown, influences on the shaping of neuronal populations and connectivity. In accord with this speculation is the observation that posttraumatic plasticity of cortical whisker representation, requiring rewiring of circuitries and changes in synaptic efficacy, is reduced in the adult TN-C-deficient mouse. Last, it is worth mentioning that numbers of astrocytes, which normally express TN-C, are strongly increased in the TN-C-deficient neocortex and hippocampus. Since there is no evidence for neurodegenerative processes, an increase in astrocytes numbers may be a compensatory response related to synaptic functions.

Plastic and Cognitive Functions of Chondroitin Sulfate Proteoglycans

Other ECM molecules to be discussed in the context of synaptic functioning are chondroitin sulfate proteoglycans of the lectican family. This family comprises four abundant components of brain ECM: aggrecan, versican, neurocan, and brevican. Lecticans together with link proteins, hyaluronic acid, and

glycoproteins such as the tenascins form huge carbohydrate–protein aggregates of perineuronal nets.

Mice deficient in neurocan show reduced LTP of CA3–CA1 synapses. Interestingly, the deficit appears only >2 h after induction of LTP but not earlier. Since late LTP phases are characterized by synaptic remodeling, it is plausible to assume a role for neurocan in such processes. Brevican is the closest relative of neurocan and one of the most abundant chondroitin sulfate proteoglycans in mature rodent central nervous system. Application of anti-brevican antibodies to naïve hippocampal slices leads to a drastic reduction of CA1 LTP 30 min after tetanization of the Schaffer collaterals. This finding is corroborated by LTP recordings in brevican-deficient mice. Although basic properties of excitatory and inhibitory synaptic transmission in these mutants are normal, LTP maintenance is dramatically impaired within the same time frame as after application of brevican antibody. The observed LTP deficit in brevican-deficient mice, however, seems not to be accompanied by changes in learning and memory, as measured in the water maze and active avoidance tasks. However, it is noteworthy that available results were obtained with a relatively small number of mice, and that in both tests mutant mice tended to show a poorer performance (less preference for the target during the probe trial of the water maze task and lower avoidance in the active avoidance task) compared to wild-type mice, suggesting normal learning but reduced memory consolidation under brevican deficiency. If we take into consideration the subtle but significant alteration of TN-R-deficient mice or that impairments of HNK-1 sulfotransferase-deficient mice were observed only when specific cognitive abilities (e.g., long-term memory or fast re-learning rate) were challenged, it is plausible to assume that more detailed behavioral analyses could reveal a role of brevican in specific cognitive functions.

Another interesting chondroitin sulfate proteoglycan is phosphacan/receptor phosphotyrosine phosphatase beta (RPTPβ). It is expressed in close vicinity to synapses, where it is well in place to bridge and stabilize perisynaptic glial processes near synapses via an interaction with the contactin–Caspr complex. The involvement of the RPTPβ family in synaptic plasticity is suggested by the analysis of RPTPβ knockout mice. These mice exhibited enhanced hippocampal LTP in an age-dependent manner, which could be overruled by pharmacological inhibition of the Rho-associated kinase (ROCK), the major downstream effector of a small GTPase Rho. Interestingly, this finding was accompanied by impairments in the Morris in the water maze task, suggesting that enhanced LTP is not necessarily beneficial for learning.

Improvement of cognitive functions was reported for biglycan, a leucine-rich chondroitin sulfate proteoglycan with neurotrophic effects. When injected into the nucleus basalis magnocellularis of the ventral pallidum or into the ventricles, biglycan ameliorated the cognitive deficits of old rats. This effect was mediated by attached chondroitin sulfate C and not by the proteoglycan *per se*, suggesting that other chondroitin sulfate proteoglycans may have the same effects as biglycan. Similarly, intrapallidal injection of heparin, another glycosaminoglycan, facilitated memory consolidation in the step-through passive avoidance task in rats. In the case of both biglycan and heparin, the beneficial effects on cognitive abilities appeared to be mediated by increasing cholinergic activity in the frontal cortex. It should be mentioned that the improved performance in a passive avoidance task could also result from enhanced anxiety in the animals: in fact, injection of chondroitin sulfate C into the dorsal periaqueductal gray enhanced anxiety and risk assessment as tested in the elevated-plus maze. At least some plasticity-related activity of chondroitin sulfate proteoglycans is mediated via the attached chondroitin sulfate side chains because treatment of hippocampal slices with the enzyme chondroitinase ABC, which removes chondroitin sulfates from chondroitin sulfate proteoglycans, leads to a reduction in LTP and LTD.

The mechanisms of how lecticans might influence synaptic transmission are unknown, but many interesting possibilities exist. Lectican-enriched ECM may insulate synaptic contact sites to prevent spillover of released neurotransmitters to neighboring synapses or extrasynaptic neurotransmitter receptors. Proteoglycans may also shape synaptic processes via trapping and concentrating neurotrophins at synapses and presenting them to their cognate receptors. Finally, chondroitin sulfate proteoglycans may bind to not yet characterized receptors, which might actively transduce signals from the ECM to postsynaptic neurons, thus directly influencing synaptic transmission properties. A candidate receptor could be neural cell adhesion molecule (NCAM), a cell adhesion molecule implicated in synaptic plasticity, which is also known to interact with neurocan and phosphacan.

Activity-Dependent Remodeling of the ECM

Although one of ECM functions is to provide mechanical stability to a tissue, this is a dynamic structure. Some ECM molecules in the brain are secreted in an activity-dependent manner. These include neuronal pentraxin NP2, reelin, aggrecan, TN-C, and TN-R. The ECM molecules released may trigger signaling

via cognate receptors (i.e., integrins) and thus affect synaptic plasticity. Furthermore, extracellular proteases mediating proteolytic cleavage of ECM molecules are also released in an activity-dependent manner. Obviously, their activity may lead to structural changes in the ECM and inhibition of ECM-mediated signaling due to digestion of some ECM molecules. However, their activity may also lead to a relief of cryptic ECM receptor binding sites and thus trigger signaling events.

More than 20 metalloproteinases (MMPs) and four of their endogenous tissue inhibitors act together to control tightly temporally restricted, focal proteolysis of ECM. Since the cascade of proteases is initiated by plasminogen activators, these activators have the potential to be the primary enzymes in the control of events such as learning and LTP, which presumably involve matrix remodeling. Indeed, mice genetically deficient in the plasminogen activator tissue plasminogen activator (tPA) showed a decrease in late phase of LTP in the CA1 field and impaired memory in the contextual fear conditioning and two-way active avoidance tasks. Conversely, transgenic mice overexpressing tPA in postnatal neurons exhibited exaggerated LTP in CA1 and displayed better learning and re-learning in the water maze task compared to wild-type mice. There is also evidence that tPA is involved in motor learning because tPA was elevated in cerebellar Purkinje cells of rats trained to rapidly traverse an irregular pegged straight runaway.

Studies have demonstrated that MMP-3 and MMP-9 increase transiently during water maze acquisition in a NMDA receptor-dependent manner. Interestingly, whereas MMP-9 transiently increased only at the beginning of learning, MMP-3 levels were enhanced throughout the training in the water maze over several days, suggesting that the two MMPs subserve distinct functions during memory consolidation. Inhibition of MMP activity with MMP-3 and MMP-9 antisense oligonucleotides and/or MMP inhibitor FN-439 reduced LTP at CA3–CA1 synapses. Similarly, in slices from MMP-9-deficient mice, hippocampal LTP, but not LTD, was impaired in magnitude and duration, whereas adding recombinant active MMP-9 to slices derived from mutants restored the magnitude and duration of LTP to wild-type levels. Activated MMP-9 localizes in part to synapses and modulates hippocampal synaptic physiology through integrin receptors because integrin function-blocking reagents prevent an MMP-9-mediated potentiation of synaptic strength. The learning-dependent MMP alterations modify the stability of the actin-binding protein cortactin, which plays an essential role in regulating the dendritic cytoskeleton and synaptic efficiency. The important role of MMP-9 in regulating synaptic plasticity is indicated by the reduced contextual long-term memory displayed by MMP-9-deficient mice in the fear conditioning task. Tone memory (which, contrary to contextual memory, is not supposed to require the hippocampus) was normal in these mice, but it was only tested after a short interval after training, so it is not possible to conclude whether the memory impairment of MMP-9-deficient mice is specific for hippocampal function, long-term memories, or both. In summary, these results indicate that changes in MMP function and remodeling of ECM are critical to synaptic plasticity and hippocampal-dependent learning.

Conclusions and Future Directions

Several ECM molecules have been identified as crucial players in hippocampal synaptic plasticity and learning and memory (Table 1). The signaling events underlying the contributions of these and probably other ECM molecules to synaptic functions remain to be elucidated, particularly at the presynaptic level. Yet we believe that it can be concluded that the ECM in the nervous system is much more than just a passive microenvironment repellant or conducive for cells and neurites. Tight functional links to synaptic plasticity and learning, together with data showing remarkable activity and learning-induced changes in expression of ECM molecules and enzymes involved in remodeling of the ECM in the adult nervous system, suggest that the ECM is a dynamic, functionally indispensable structure between glial cells and their pre- and postsynaptic neuronal partners.

Table 1 ECM molecules, synaptic plasticity, and learning and memory

Functions	*ECM molecules*
Synaptic plasticity	
Activity of NMDARs	Reelin, integrin ligand(s)
Activity of L-type Ca^{2+} channels	Tenascin-C
GABAergic inhibition	Tenascin-R, HB-GAM
Activity/recycling of AMPARs	Integrin ligand(s)
Activity of Rho/ROCK	RPTPβ
Unknown mechanisms	Brevican, neurocan
Learning and memory	
Water maze	Tenascin-R, HNK-1, integrin ligand(s), RPTPβ, brevican[a]
Contextual fear conditioning	Reelin, tenascin-C (on extinction)
One-trial passive avoidance	Biglycan, heparin
Working memory	Integrin ligand(s)

[a]Only tendencies were observed in a small number of mice.

See also: Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Axonal Pathfinding: Extracellular Matrix Role; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Plasticity: Neuronal Sprouting.

Further Reading

Beffert U, Durudas A, Weeber EJ, et al. (2006) Functional dissection of Reelin signaling by site-directed disruption of Disabled-1 adaptor binding to apolipoprotein E receptor 2: Distinct roles in development and synaptic plasticity. *Journal of Neuroscience* 26: 2041–2052.

Beffert U, Weeber EJ, Durudas A, et al. (2005) Modulation of synaptic plasticity and memory by Reelin involves differential splicing of the lipoprotein receptor Apoer2. *Neuron* 47: 567–579.

Bernard-Trifilo JA, Kramar EA, Torp R, et al. (2005) Integrin signaling cascades are operational in adult hippocampal synapses and modulate NMDA receptor physiology. *Journal of Neurochemistry* 93: 834–849.

Chan CS, Weeber EJ, Kurup S, Sweatt JD, and Davis RL (2003) Integrin requirement for hippocampal synaptic plasticity and spatial memory. *Journal of Neuroscience* 23: 7107–7116.

Chan CS, Weeber EJ, Zong L, Fuchs E, Sweatt JD, and Davis RL (2006) Beta1-integrins are required for hippocampal AMPA receptor-dependent synaptic transmission, synaptic plasticity, and working memory. *Journal of Neuroscience* 26: 223–232.

Dityatev A and Schachner M (2003) Extracellular matrix molecules and synaptic plasticity. *Nature Reviews Neuroscience* 4: 456–468.

Evers MR, Salmen B, Bukalo O, et al. (2002) Impairment of L-type Ca^{2+} channel-dependent forms of hippocampal synaptic plasticity in mice deficient in the extracellular matrix glycoprotein tenascin-C. *Journal of Neuroscience* 22: 7177–7194.

Kramar EA, Bernard JA, Gall CM, and Lynch G (2003) Integrins modulate fast excitatory transmission at hippocampal synapses. *Journal of Biological Chemistry* 278: 10722–10730.

Meighan SE, Meighan PC, Choudhury P, et al. (2006) Effects of extracellular matrix-degrading proteases matrix metalloproteinases 3 and 9 on spatial learning and synaptic plasticity. *Journal of Neurochemistry* 96: 1227–1241.

Morellini F and Schachner M (2006) Enhanced novelty-induced activity, reduced anxiety, delayed resynchronization to daylight reversal and weaker muscle strength in tenascin-C-deficient mice. *European Journal of Neuroscience* 23: 1255–1268.

Nagy V, Bozdagi O, Matynia A, et al. (2006) Matrix metalloproteinase-9 is required for hippocampal late-phase long-term potentiation and memory. *Journal of Neuroscience* 26: 1923–1934.

Nakagami Y, Abe K, Nishiyama N, and Matsuki N (2000) Laminin degradation by plasmin regulates long-term potentiation. *Journal of Neuroscience* 20: 2003–2010.

Qiu S, Korwek KM, Pratt-Davis AR, Peters M, Bergman MY, and Weeber EJ (2006) Cognitive disruption and altered hippocampus synaptic function in Reelin haploinsufficient mice. *Neurobiology of Learning and Memory* 85: 228–242.

Saghatelyan AK, Dityatev A, Schmidt S, Schuster T, Bartsch U, and Schachner M (2001) Reduced perisomatic inhibition, increased excitatory transmission, and impaired long-term potentiation in mice deficient for the extracellular matrix glycoprotein tenascin-R. *Molecular and Cellular Neuroscience* 17: 226–240.

Saghatelyan AK, Snapyan M, Gorissen S, et al. (2003) Recognition molecule associated carbohydrate inhibits postsynaptic GABA (B) receptors: A mechanism for homeostatic regulation of GABA release in perisomatic synapses. *Molecular and Cellular Neuroscience* 24: 271–282.

Sanes JR and Lichtman JW (2001) Induction, assembly, maturation and maintenance of a postsynaptic apparatus. *Nature Reviews Neuroscience* 2: 791–805.

Senn C, Kutsche M, Saghatelyan A, et al. (2002) Mice deficient for the HNK-1 sulfotransferase show alterations in synaptic efficacy and spatial learning and memory. *Molecular and Cellular Neuroscience* 20: 712–729.

Strekalova T, Sun M, Sibbe M, et al. (2002) Fibronectin domains of extracellular matrix molecule tenascin-C modulate hippocampal learning and synaptic plasticity. *Molecular and Cellular Neuroscience* 21: 173–187.

Weeber EJ, Beffert U, Jones C, et al. (2002) Reelin and ApoE receptors cooperate to enhance hippocampal synaptic plasticity and learning. *Journal of Biological Chemistry* 277: 39944–39952.

Relevant Website

http://www.biochemweb.org/ – BioChemWeb.org – Cell adhesion and extracellular matrix.

Eye and Head Movements

K E Cullen, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Interactions between the Vestibulo-Ocular Reflex and Voluntary Eye Movements

The traditional approach of studying the oculomotor subsystems that generate stabilizing reflexes (i.e., the vestibulo-ocular reflex (VOR) and optokinetic system) and voluntary eye movements (i.e., saccades, smooth pursuit, and vergence) in isolation has led to an excellent understanding of the brain stem circuitry underlying each of these types of eye movements. However, during most natural voluntary orienting and tracking behaviors, two or more subsystems work together to control gaze. When we shift our axis of gaze in space by making a combined eye–head gaze shift, the VOR and saccadic subsystems must work together to ensure that the axis of gaze (where gaze = eye-in-orbit + head-in-space) is placed precisely on the target of interest. Similarly, to track a slowly moving target, we move our head as well as eyes in a manner that is dependent on the target's velocity and eccentricity. In this case, the VOR and smooth pursuit subsystems must work together to ensure that the axis of gaze is aligned with target movement. The manner in which the brain coordinates the interactions between these eye movement subsystems and head premotor circuitry to accurately redirect gaze is an area of ongoing research.

The Control of Eye–Head Gaze Shifts

Behavioral Investigations

The VOR functions to produce a compensatory eye movement of equal and opposite amplitude to the motion of the head. This reflex rapidly (within 5–7 ms) and effectively stabilizes the visual world on the retina over the wide range of head motions that are produced during our daily activities. However, primates also make voluntary head movements to redirect their visual axis to novel targets in their environment. Under natural conditions, where the head is not restrained, coordinated eye and head movements (i.e., gaze shifts) are commonly used to rapidly redirect the visual axis to a new point in space. The eye movement command produced by the VOR during gaze shifts would be counterproductive because it would code an eye movement in the direction opposite to that of the intended shift of the axis of gaze.

Bizzi and colleagues originally proposed that the VOR remains completely functional during gaze shifts, such that it effectively eliminates the head's contribution to the change in gaze. In this schema, termed the linear summation hypothesis, the actual eye movement generated at the level of the extraocular motor neurons reflects the summation of two signals: an intact VOR from vestibular pathways (i.e., a compensatory eye movement signal (Figure 1(a)) and a separate oculocentric saccadic eye movement signal from the brain stem saccadic burst generator (Figure 1(b)), which counters the VOR and redirects the axis of gaze. The conceptual simplicity of this proposal has made it an attractive model of gaze shift generation. Figure 2(a) shows a schematic of how these two commands sum to drive to the extraocular motor neurons during a rightward eye–head gaze shift. The saccadic pathway commands a rightward eye movement, whereas the VOR pathway, in response to the rightward head movement that is produced, commands a compensatory (i.e., leftward) eye movement.

Substantial evidence, however, has accumulated over the past three decades to indicate that the linear summation hypothesis is generally not valid during gaze shifts. The status of the VOR during gaze shifts has been probed by perturbing head movement during gaze shifts using electromagnetic clutches to brake the head and torque motors to apply whole-body and head-on-body rotations. There is considerable agreement across studies that the VOR is completely turned off during large gaze shifts and is significantly suppressed during smaller gaze shifts. The attenuation of the VOR is illustrated in Figure 2 by the box labeled 'behaviorally dependent modulation.' For gaze shifts less than 50°, VOR gain (defined as the change in eye velocity/change in head velocity caused by a perturbation) decreases as a function of increasing gaze shift amplitude. Taken together, these findings indicate that the head movements generated during gaze shifts contribute to changing the axis of gaze in space because the contribution of the head movement to gaze is not completely cancelled out (i.e., VOR gain $\neq 1$). As a result, several models have been proposed to explain how eye and head movements are coordinated to produce accurate gaze shifts.

The gain of the VOR is not only dependent on the amplitude of an ongoing gaze shift, but also varies temporally during the execution of a single gaze shift such that there is a gradual restoration of the VOR efficacy from its off state. The traces in the top right corner of Figure 2(b) are examples of transient head perturbations that have been used to explicitly quantify VOR gain at specific intervals during large gaze

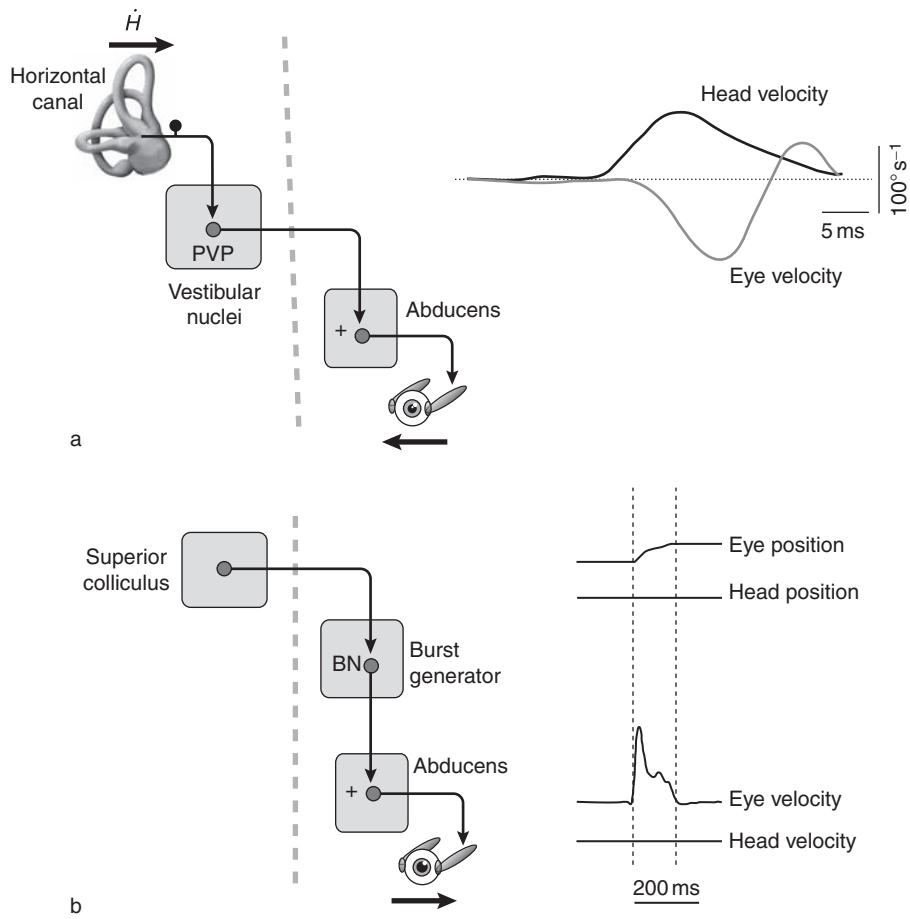

Figure 1 The primary inputs to the extraocular motor neurons in the abducens nucleus: (a) the VOR pathway; (b) the saccadic pathway. The panels on the right side of (a) and (b) show examples of eye- and head-velocity responses during the VOR and saccadic eye movements, respectively. BN, premotor saccadic burst neuron; PVP, position-vestibular-pause neuron; VOR, vestibulo-ocular reflex.

shifts. The graph in Figure 2(b) shows the time course of the VOR gain before, during, and after a gaze shift. VOR suppression is maximal early in the gaze shift and progressively recovers to reach normal values near gaze shift end. The level of VOR suppression, however, is not a constant value at gaze shift onset but, instead, shows considerable variation across subjects.

Neurophysiology: Premotor Control of the Eye during Gaze Shifts

Brain stem saccadic pathway During gaze shifts, the superior colliculus (SC) generates a gaze displacement command that is used by downstream premotor neurons to control the movement of the eye and head. The output neurons of the SC send direct projections to two regions in the brain stem reticular formation which control saccadic eye movements: the paramedian pontine reticular formation (PPRF) and the rostral interstitial nucleus of the medial longitudinal fasciculus (riMLF) of the midbrain. In the head-restrained condition, neurons in the PPRF and riMLF encode temporally modulated command signals that provide the primary input to the extraocular motor neurons to drive horizontal and vertical saccadic eye movements, respectively.

During head-restrained saccades, the responses of premotor burst neurons are well described by a simple model ($\mathrm{FR} = b + r\dot{E}$; where FR is firing rate, $\dot{E}$ is eye velocity, and b and r are constants), which is based on eye movement dynamics. However, the same eye-based dynamic model that describes a neuron's response during head-restrained saccades cannot be used to predict its response during gaze shifts (Figure 3(a), left inset, eye-based model, heavy line). This is because these neurons encode head as well as eye velocity during gaze shifts. In general, premotor saccadic burst neurons encode head velocity with only about half as much response sensitivity as eye velocity, whereas collicular neurons encode gaze-related information (where gaze = eye + head). This progressive loss of head-movement-related response in the saccadic pathway is consistent with downstream interactions between the saccadic and VOR circuits.

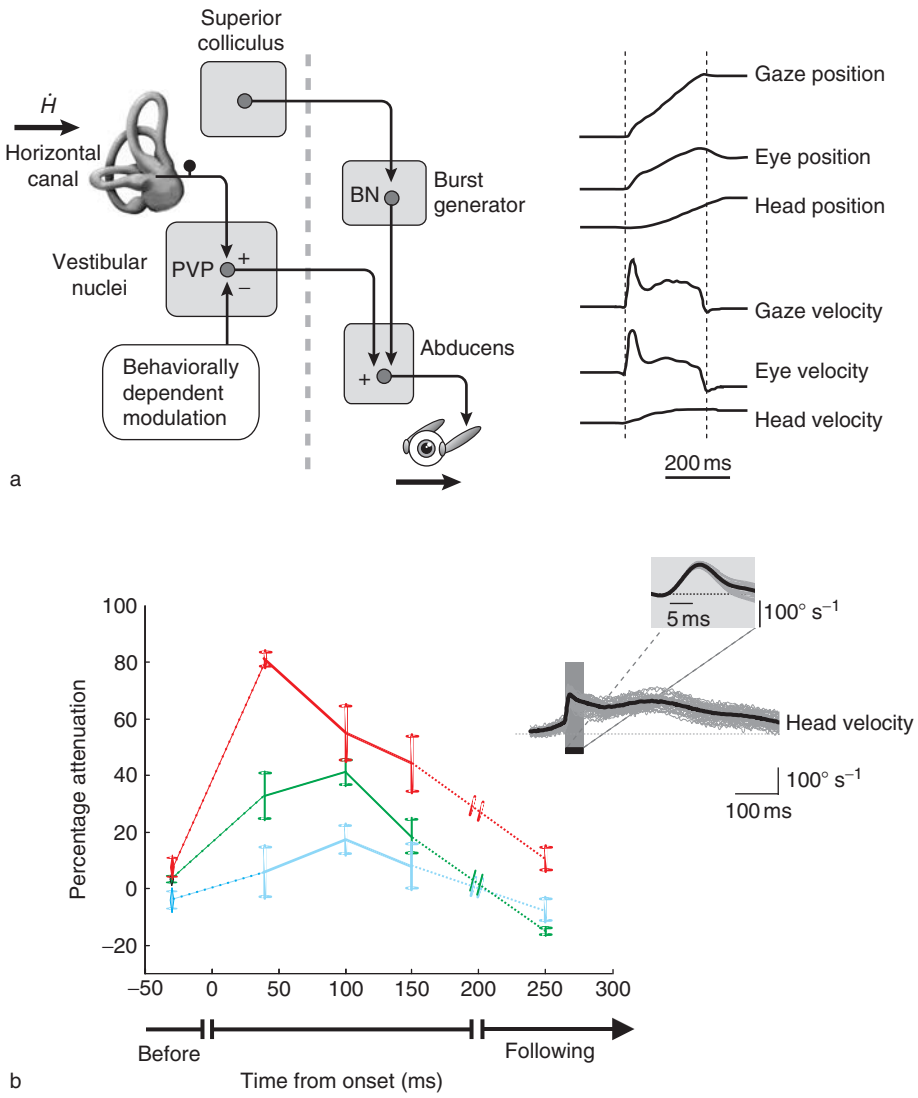

Figure 2 (a) Interactions between the VOR and saccadic pathways during a rightward gaze shift; (b) VOR gain in response increases over time from gaze shift onset, ultimately reaching unity at gaze shift end. Error bars represent standard errors. Each color shows data from a different subject. To the right in (b) are example head perturbations (vertical shaded gray bar and inset) that were applied to quantify the VOR during gaze shifts. BN, premotor saccadic burst neuron; PVP, position-vestibular-pause neuron; VOR, vestibulo-ocular reflex. (b) Data from Cullen KE, Huterer M, Braidwood DA, and Sylvestre PA (2004) Time course of vestibuloocular reflex suppression during gaze shifts. *Journal of Neurophysiology* 92: 3408–3422.

As a result, the gaze-related drive that originates in the SC ultimately becomes completely oculomotor at the level of the extraocular motor neurons (see the section titled 'Extraocular motor neurons').

Vestibulo-ocular reflex pathways A simple three neuron arc (vestibular afferents to interneurons in the vestibular nuclei to extraocular motor neurons) mediates the most direct pathway of the VOR. Investigations in head-restrained monkeys have revealed that there are several classes of vestibular nuclei neurons that discharge in relation to horizontal passive whole-body rotation and/or eye movements. One class, called position-vestibular-pause (PVP) neurons, constitutes most of the intermediate leg of the direct VOR pathway. These neurons receive a strong monosynaptic connection from the ipsilateral vestibular afferents and, in turn, project directly to contralateral extraocular motor neurons (Figure 1(b)).

Type I PVP neurons derive their name from the signals they carry during head-restrained oculomotor and vestibular paradigms: (1) their firing rates increase when the eyes move to more contralaterally directed positions; (2) during slow phase vestibular nystagmus, these neurons are sensitive to ipsilateral head rotations (i.e., a type I response); and (3) their discharges completely cease (pause) for ipsilaterally directed saccades and vestibular quick phases. Although these

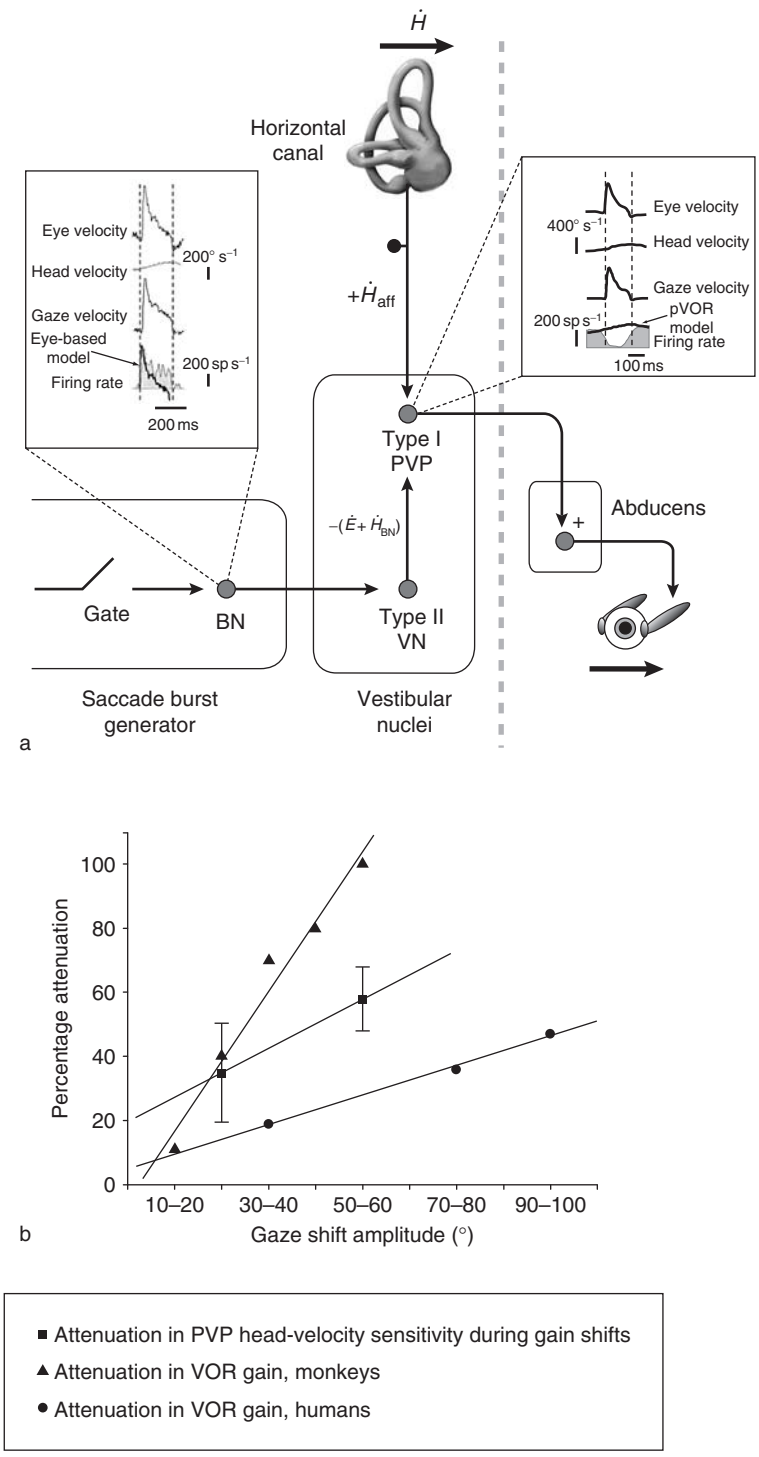

Figure 3 (a) Brain stem mechanism that underlies the reduction in VOR gain during gaze shifts; (b) comparison of the percentage attenuation of PVP head-velocity sensitivity during gaze shifts and the percentage attenuation of the VOR gain, as measured in two behavioral experiments (monkeys and humans). In (a), the insets show the responses of BN and PVP neurons during gaze shifts. BN, burst neuron; $\dot{E} + \dot{H}_{BN}$, eye and head velocity encoded by saccadic burst neurons; $\dot{H}_{aff}$, head velocity encoded by horizontal semicircular canal afferents; PVP, position-vestibular-pause neuron; VOR, vestibulo-ocular reflex. Data from Cullen KE and Guitton D (1997) Analysis of primate IBN spike trains using system identification techniques, II: Relationship to gaze, eye, and head movement dynamics during head-free gaze shifts. *Journal of Neurophysiology* 78: 3283–3306; Roy JE and Cullen KE (1998) A neural correlate for vestibulo-ocular reflex suppression during voluntary eye-head gaze shifts. *Nature Neuroscience* 1: 404–410; Tabak S, Smeets JBJ, and Collewijn H (1996) Modulation of the human vestibuloocular reflex during saccades: Probing by high-frequency oscillation and torque pulses of the head. *Journal of Neurophysiology* 76: 3249–3263; Tomlinson RD (1990) Combined eye-head gaze shifts in the primate, III: Contributions to the accuracy of gaze saccades. *Journal of Neurophysiology* 64: 1873–1891.

neurons are robustly modulated in response to passive head rotations, their responses to head velocity are dramatically attenuated during gaze shifts. The response of a typical PVP neuron during gaze shifts is shown in Figure 3(a) (right inset). The neuron's activity is interrupted during the gaze shift interval, which is denoted by the two vertical dashed lines. The response predicted by the same neuron's response during passive whole-body rotation is superimposed for comparison (pVOR model, heavy line). Moreover, the head-velocity signal encoded by PVP neurons is increasingly attenuated as a function of gaze-shift amplitude. This neurophysiological trend (Figure 3(b)) is consistent with the results of the behavioral studies of the VOR which we already discussed. The results from two of these prior studies have been superimposed on PVP neuron data for comparison. Thus the amplitude-dependent reduction of the responses of VOR interneurons (i.e., PVP neurons) underlies the amplitude-dependent decrease in behavioral VOR gain that occurs during gaze shifts.

The attenuation of PVP responses during gaze shifts can be accounted for by a known mechanism, namely projections from the brain stem saccade generator to the vestibular nuclei (Figure 3(a)). During saccades, vestibular quick phases and gaze shifts, brain stem burst neurons (BNs) become active, as denoted by the gate. As a result, BNs (Figure 3(a), left inset) provide an inhibitory input related to eye velocity (E) and head velocity ($\dot{H}_{\mathrm{BN}}$) to type I PVP neurons during gaze shifts via their projections to type II neurons in the vestibular nuclei. (Note that type II vestibular nuclei neurons are activated in response to contralateral head rotations.) In contrast, immediately following a gaze shift (when the gaze is stable but the head is still moving), the head-velocity sensitivity of PVP neurons completely recovers to reach that observed during the VOR (Figure 3(a), right inset). This observation is also consistent with the mechanism illustrated because the brain stem saccade generator is not active during this interval.

Extraocular motor neurons Because the responses of PVP neurons are attenuated (but not completely abolished) during gaze shifts, some head-movement related information is directly conveyed to the extraocular motor neurons (Figure 3(a), right inset). Moreover, saccadic BNs carry head- as well as eye-movement-related signals to the extraocular motor neurons (i.e., Figure 3(a), left inset). The questions thus arise: What signals are encoded by the extraocular motor neurons during gaze shifts? Do individual motor neurons encode only eye movement, or is the relationship between firing rates and eye movement influenced by the additional premotor signals that are present during gaze shifts?

A first-order eye-movement-based model ($\mathrm{FR} = b + kE + r\dot{E}$; where FR is firing rate; E and $\dot{E}$ are the eye position and velocity, respectively; and b, k, and r are constants) provides a good representation of the firing rate of extraocular motor neurons during saccadic eye movements made in the head-restrained condition. The addition of an acceleration term ($u\ddot{E}$) and an exponentially decaying term (or slide term, $c\dot{\mathrm{F}}\mathrm{R}$) further improves our ability to mathematically describe the saccade-related activity in the head-restrained condition. During combined eye–head gaze shifts, the parametric relationship between an individual motor neuron's firing rate and eye motion does not change because the head is moving; the activity of each extraocular motor neuron is related to eye motion by the same dynamic relationship during head-unrestrained gaze shifts and head-restrained saccades. This finding is logically consistent with the projections of these neurons because they drive the extraocular muscles of the eyes alone.

The finding that premotor VOR neurons (PVPs) and saccadic BNs encode head-movement-related information but that extraocular motor neurons do not can be easily reconciled with what is known about signals encoded by these premotor pathways during gaze shifts. These two premotor pathways respond to contralaterally and ipsilaterally directed head-movement-related information, respectively. As a result, the residual head-velocity information carried by premotor VOR (i.e., PVP) neurons is appropriate to offset the incoming head-velocity-related drive from saccadic premotor BNs, such that the responses of individual extraocular motor neurons are solely related to eye motion during gaze shifts.

Neurophysiology: Premotor Control of the Head during Gaze Shifts

The SC has been identified as a primary center that controls both eye and head movements during gaze shifts. The output neurons in the intermediate layers of this structure not only send descending projections to the premotor saccadic circuitry, but also project to the neck motor neurons via two major pathways.

First a tectoreticulospinal pathway, which accesses the reticular formation of the pons and medulla, controls orienting movements in the horizontal direction. In particular, the reticulospinal neurons of the nucleus reticularis gigantocellularis (NRG) and nucleus reticularis pontis caudalis (NRPc) receive monosynaptic input from the SC. Electrical stimulation of the NRG, which is located just caudal to the abducens nucleus, evokes ipsilateral deviations of the head in the horizontal plane. The amplitudes and peak velocities of the evoked head movement depend on the frequency

and duration of stimulation, consistent with the idea that these nuclei play a critical role in transforming SC signals into horizontal head-movement commands. In cats, combined lesions of both the NRG and NRPc impair the ability to make horizontal orienting head movements. In contrast, lesions restricted to either area only minimally influence orienting head movements, suggesting that the NRG and NRPc can partially compensate for one another's loss of function.

Second, a tectoreticulospinal pathway via the reticular formation of the mesodiencephalic region controls orienting head movements in the vertical direction. The SC sends projections to an area known as the Forel's field H (FFH) in cats, which corresponds to the riMLF in monkeys. This area, in turn, projects, directly to ipsilateral neck motor neurons. During orienting head movements, the responses of FFH neurons precede the onset of head movements and are temporally related to the dynamics of the head movement. Stimulation of the FFH induces upward and torsional head movements consistent with the idea that the vertical component of orienting head movements is encoded by activation of the FFH.

The tectoreticulospinal pathways to the pons/medualla and mesodiencephalic regions of the reticular formation transform the topographic representation of gaze displacement that is encoded by the SC into the temporally modulated commands signals that are required to drive the neck musculature in the horizontal and vertical directions, respectively. These pathways function in concert with other descending pathways that also converge at the level of neck motor neurons, including the vestibulo-spinal, rubrospinal, and interstitiospinal tracts. Vestibulo-spinal pathways play a particularly important role in the regulation of gait and posture. Neurons in the vestibular cerebellum project to neurons in the vestibular nuclei, which in turn project to neck motor neurons. The modulation of this pathway is attenuated but not completely suppressed during active orienting head movements. As a result, the drive to neck motor neurons is determined by the inputs from the vestibulo-spinal pathway as well as those from the tectospinal pathways. Accordingly, temporary lesions of areas of the vestibular cerebellum, such as the rostral fastigial nucleus, not only result in postural deficits but can also alter the metrics of gaze and head orienting movements.

Overall, it is important to note that our understanding of the premotor pathways that project to the neck motor neurons is largely based on experimental studies done in cats. In this species, neurons in the reticular formation just rostral to the NRG and caudal to the abducens nucleus have been shown to project both to neck and eye motor neurons. Neurons in the corresponding location of primates are known to project to the extraocular motor neurons, but to date there has been no explicit demonstration that these same neurons also project to the cervical spinal cord.

Integrated Feedback Control of Eye–Head Gaze Shifts and Implications for Upstream Control

That the VOR is attenuated during gaze shifts has important implications regarding the premotor control of gaze control. As previously explained, if the VOR is not fully operational during gaze shift, it then follows that the head as well as the eyes contribute to shifting the axis of gaze relative to space. Thus, a fundamental question is: how are the motor commands to the eye and head musculature coordinated to ensure accurate gaze shifts? Two general classes of models have emerged that can account for the control of eye–head gaze shifts. In the first model, a single controller is used to minimize gaze error rather than enforce a specific eye or head trajectory, making it conceptually analogous to optimal feedback control. Alternatively, it has been proposed that a gaze displacement command is decomposed early on to control separate eye and head comparators. In this schema, it is the individual displacements of the eye and head, rather than the overall accuracy of the movement *per se*, that are under feedback control. To date, the available evidence has proven equivocal with respect to these two model classes. The finding that movement accuracy is preserved when head trajectories are experimentally altered during gaze shifts is readily accounted for by gaze comparator models. However, models based on separate comparators can also simulate this finding with the incorporation of neck reflexes. Likewise, the responses of saccadic brain stem neurons during gaze shifts have been reconciled with both single and separate comparator models.

Firm evidence that gaze shift accuracy is ensured by a single comparator has been provided by more recent experiments. The responses of neurons in the premotor circuit that controls saccadic eye movement (saccade related burst neurons and omnipause neurons) are instantaneously updated following experimentally applied head perturbations (**Figure 4(a)**). Rapid (>3 ms) online feedback of head-movement updates the durations and dynamics of neurons in the saccadic premotor network in order to preserve global movement accuracy during gaze shifts. This result is consistent with a single controller model because it predicts that the drive to both eye as well as head premotor circuits should be updated in response to a head perturbation in order to preserve gaze accuracy. In contrast, models in which gaze-shift accuracy is

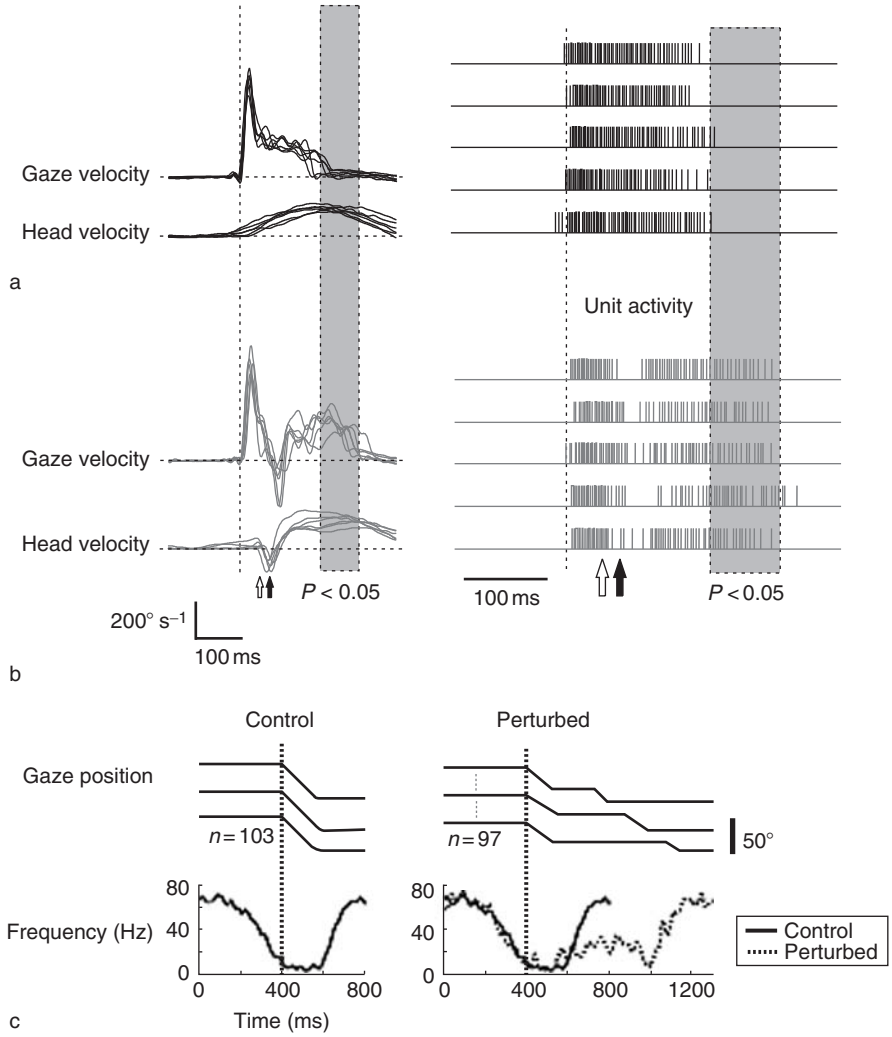

Figure 4 Gaze shifts and example burst neuron activity recorded during the same trials: (a) control; (b) perturbed; and (c) discharges of superior colliculus neurons during control (left) and perturbed trials (right), aligned on gaze shift onset. In (a) and (b), the vertical dashed line indicates gaze onset; the shaded box indicates average increase in perturbed trials duration; and the open and filled arrows indicate average perturbation onset and time at which the resulting head perturbation reached maximum velocity, respectively. Traces are aligned on gaze shift onset. In (c), the number of control and perturbed gaze shifts used to calculate average neuronal response is provided. Data from Sylvestre PA and Cullen KE (2006) Premotor correlates of integrated feedback control for eye-head gaze shifts. *Journal of Neuroscience* 26(18): 4922–4929; Choi WY and Guitton D (2006) Responses of collicular fixation neurons to gaze shift perturbations in head-unrestrained monkey reveal gaze feedback control. *Neuron* 50(3): 491–505.

ensured by separate controllers, specifying *a priori* the amplitude of eye and head, are difficult to reconcile with the present results. These models do not predict that the neurons in the brain stem saccadic pathway should be altered to reflect the changes in gaze shift dynamics that result from experimentally applied head-movement perturbations.

Prior reports have provided evidence that an integrated controller that functions to coordinate the movements of the eyes and head during gaze shifts might be centered on the SC. The output neurons of the SC have direct access to both the saccadic and neck premotor circuitry, thereby providing a physiological substrate by which eye and head movements can be driven in a coordinated fashion during gaze shifts (see sections titled 'Behavioral investigations' and 'Neurophysiology: Premotor control of the eye during gaze shifts' before this). Firm evidence that neurons in the SC control coordinated eye–head gaze saccades has been obtained from electrical stimulation and single-unit recording experiments. However, whether the SC is within an integrated feedback loop or operates open loop has remained a point of contention. In recent experiments, recordings were made from neurons in the SC during gaze shifts that were transiently interrupted by braking head movements

(Figure 4(b)). Neurons continued to fire steadily during interrupted gaze movements in a manner consistent with the proposal that the SC is within an integrated feedback loop. However, because the mechanical perturbations employed in these previous studies interrupted gaze movements (i.e., rendered gaze stable for hundreds of milliseconds), it is difficult conclude if the perturbations' effects represent dynamic corrections to the ongoing gaze shift or, rather, the programming of a second corrective movement.

Electrical stimulation of the supplementary eye fields (SEFs), like that of the SC, can evoke coordinated eye–head gaze shifts that are kinematically indistinguishable from natural gaze shifts. Similarly, stimulation of the frontal eye field (FEF) can evoke combined movements of the eye and the head that appear qualitatively similar to natural gaze shifts. These findings have been taken as evidence that cortical areas such as these in frontal cortex are not concerned with planning only the eye but also the head, components of gaze shifts. The latencies of eye and head movements evoked by stimulation in the FEF and SEF are longer than those evoked by stimulation of the SC consistent with projections from the two cortical areas to the SC. Electrical stimulation experiments show that the SEF encodes gaze direction in a reference frame that varies across a continuum from eye- to head- to space/body-centered coding schemes. How these multiple coding schemes are reflected in the processing that takes place in the SEF before the gaze command is released toward the FEF, the SC, and brain stem structures remains to be determined.

The Control of Eye–Head Pursuit

Behavioral Investigations

Humans and other primates follow a moving visual target of interest by generating smooth pursuit eye movements (Figure 5(a), left). These slow continuous eye movements function to keep a small moving target on the fovea so that it is viewed with the greatest accuracy. When the head is unrestrained, a combination of both head and eye movements are commonly used to track target motion (Figure 5(a), right). During combined eye–head pursuit movements, also called gaze pursuit, the VOR would be counterproductive for the same reason that it is counterproductive during gaze shifts. The eye-movement commands generated would reflect the summation of two opposing signals: a VOR, which cancels out the contribution of head motion to gaze, and an oculocentric pursuit signal of cerebellar origin. As a result, a fully functional VOR generates an eye-movement command in the direction opposite to that of the ongoing tracking during eye–head pursuit. Figure 5(b) shows a schematic of how these two commands sum to drive to the extraocular motor neurons during rightward eye–head pursuit. The pursuit pathway commands a rightward eye movement, while the VOR pathway, in response to the rightward head movement that is produced, commands a compensatory (i.e., leftward) eye movement.

The results of behavioral investigations have established that the VOR is partially suppressed during combined eye–head tracking. Unambiguous evidence for non-pursuit-mediated adjustment of VOR gain during eye–head tracking was found by transiently perturbing the head while subjects voluntarily tracked a head-stationary target. In this condition, the VOR would need to be completely suppressed in order to maintain foveation of a target. Indeed, in both humans and monkeys, the gain of the evoked VOR was rapidly suppressed (i.e., within 30 ms), indicating that the latency of VOR suppression is too fast to be explained by pursuit eye movements (>90 ms). The question then arises: What is the advantage of having a non-pursuit-related mechanism to negate the VOR during eye–head tracking? The general consensus is that parametric suppression of the VOR operates to extend the range of eye–head tracking movements beyond what would be possible by the pursuit system alone. Reports that subjects can suppress their VOR gain by imaging a head-stationary target and that the ability to generate smooth pursuit eye movements to cancel the VOR are not always equally affected by central nervous system (CNS) lesions are consistent with the existence of a nonvisual VOR cancellation mechanism.

Neurophysiology: Premotor Control of the Eye during Eye–Head Pursuit

Brain stem pursuit pathways The implication of the results of the behavioral investigations just described is that interactions between premotor vestibular and oculomotor pathways must be coordinated to ensure the accurate coordination of eye and head movements during pursuit in the head-unrestrained condition. Eye–head (EH) cells receive a pursuit-related command from the cerebellar flocculus and a head-velocity-related command from the vestibular afferents and project directly to the extraocular motor neurons. EH cells encode eye velocity and provide the primary drive to the extraocular motor neurons for pursuit in the head-restrained condition (Figure 5(b)). During combined eye–head pursuit, EH neuron responses are well predicted by the sum of their gaze-movement sensitivity during smooth pursuit and their head-movement

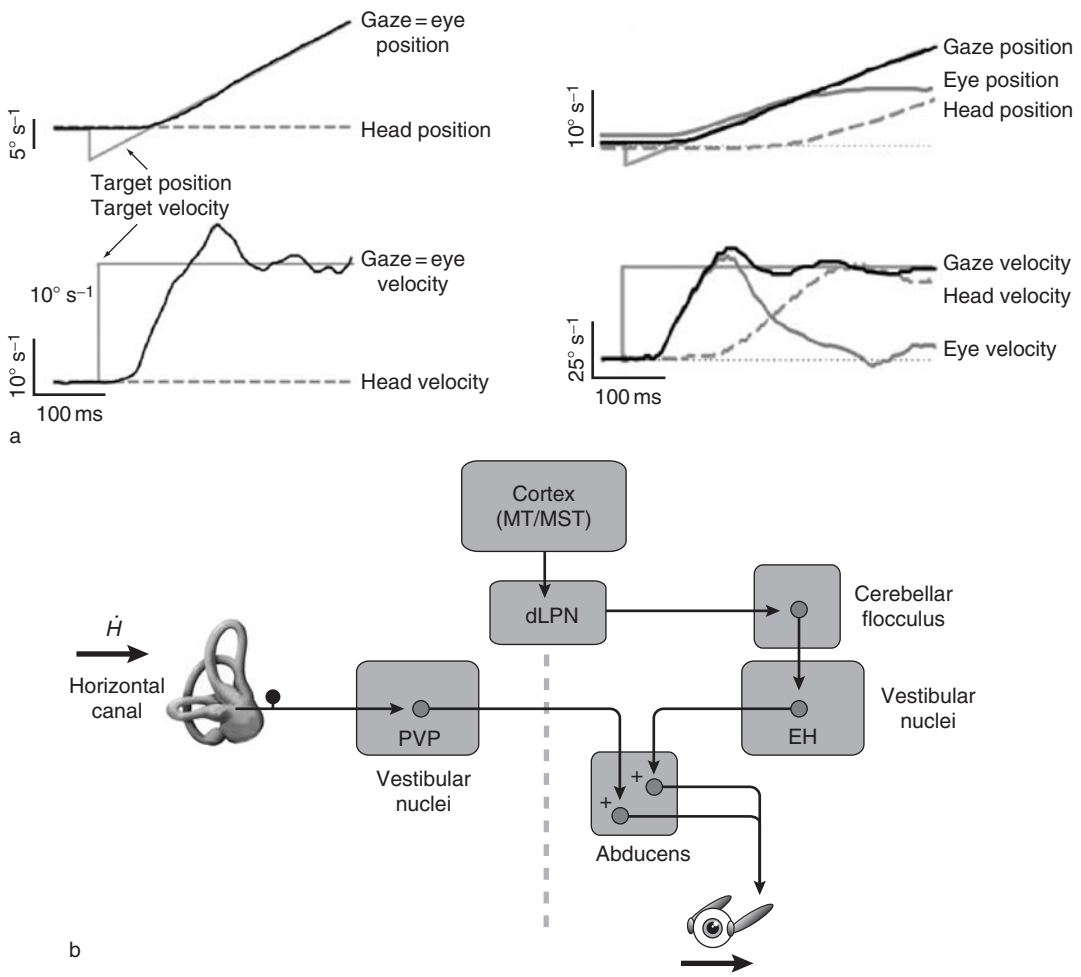

Figure 5 (a) Pursuit response to a step-ramp target trajectory made with the head restrained (left; smooth pursuit) and head unrestrained (right; gaze pursuit); (b) interactions between VOR and smooth pursuit pathways during combined eye–head tracking. dLPN, dorsal lateral pontine nucleus; EH, eye–head cells; MT/MST, middle temporal and middle superior temporal areas; PVP, position-vestibular-pause neuron; VOR, vestibulo-ocular reflex.

sensitivity during VOR in the dark. This suggests that their response reflects the integration of head-movement-related inputs from the vestibular afferents and gaze-movement-related information from the floccular lobe.

Vestibulo-ocular reflex pathways The efficacy of the VOR pathway during combined eye–head pursuit has been directly measured. The head-velocity-related responses of PVP neurons are suppressed by 30% during eye–head tracking relative to their responses during the VOR. During cancellation of the VOR (by fixation of a head-stationary target), the velocity modulation of the direct VOR pathways is likewise attenuated by 30%. Thus, the efficacy of transmission in the direct VOR pathways is similarly suppressed during cancellation of the VOR and during the active head movements made during coordinated eye–head pursuit. The responses of PVP neurons are more attenuated during gaze shifts than during eye–head pursuit (i.e., ~70 vs. 30% attenuation). This difference in the relative levels of suppression occurs because the inhibition that is mediated by the premotor circuitry that generates smooth pursuit is less powerful than that which generates saccades. Less is currently known about the connectivity of the VOR circuitry with the brain stem premotor pathway that mediates smooth pursuit than that which mediates saccades. Accordingly, fully elucidating the mechanism responsible for the attenuation in PVP neuron responses during slow gaze redirection requires further study.

Premotor inputs to the extraocular motor neurons As previously described, PVP neurons, which mediate the VOR, remain sensitive to head movements during eye–head pursuit; their head-velocity-related responses are attenuated by only ~30% relative to VOR in the

dark. This modulation has the appropriate polarity to offset the oppositely directed head-related signals carried by premotor smooth pursuit pathways (i.e., EH cells) at the level of the extraocular motor neurons during eye–head pursuit functions. Indeed, as is the case for gaze shifts, individual motor neurons encode only eye-movement-related signals during eye–head pursuit.

Neurophysiology: Premotor Control of the Head during Eye–Head Pursuit

Much less is known about the premotor control of voluntary head movements during pursuit compared to the head movements that are made during gaze shifts. This is largely because most work on voluntary head movements has been done in cat and its pursuit system is far less developed than that of monkeys. Whether neurons in the brain stem, which are distinct from the tecto-recipient neurons that control gaze shifts, drive head movements during pursuit remains to be determined. As is the case during gaze shifts, head movements during pursuit are most likely shaped by multiple spinal projecting pathways that converge at the level of the neck motor neurons.

Upstream Control

Quantification of initial pursuit dynamics has demonstrated that: (1) eye and head movement onsets are tightly coupled and (2) initial eye and head movement dynamics both similarly vary as a function of target velocity. In addition, the primate pursuit system has the ability to use and store information about the motion of a visual stimulus in order to drive the initiation and the maintenance of both the eyes and the head during gaze pursuit. Thus, the coordination of eye and head movements during pursuit is likely to be achieved, at least in part, by a common drive mechanism, whereby a shared-controller drives both the eye and neck motor systems.

Although the neural substrate of the shared controller for pursuit is not known, one possible candidate is the cerebellar flocculus and/or ventral paraflocculus (herein referred to as the floccular lobe). Patients with cerebellar disease generally show comparable deficits during combined eye–head and ocular only pursuit of pseudorandom targets. Furthermore, electrical stimulation of the floccular lobe elicits smooth eye movements in the head-restrained primate and combined smooth eye–head movements in the head-unrestrained rabbit, suggesting that the structure may be involved in eye–head gaze control. Conversely, studies in monkeys suggest that the output of the floccular lobe is not adequate to perform this role. Instead, it has been proposed that the floccular lobe primarily controls eye movement (in the orbit) as a component of the oculomotor neural integrator, whereas the dorsal vermis of the cerebellum may be more involved in the control of gaze movement.

Another possibility, although not a mutually exclusive one, is that the coordination of the eyes and the head during pursuit emerges more centrally. For example, neurons in the SEFs and FEFs respond to smooth pursuit eye movements as well as saccadic eye movements. Moreover, as previously reviewed, stimulation of both the SEFs or FEFs evoke head as well as eye movements. Although prior studies have largely focused on the kinematics of the head movements that accompany gaze shifts, stimulation can also evoke eye and head movements that are slow and continuous. Given that (1) the SEF and FEF are reciprocally connected and (2) the FEF efferents project directly to the brain stem as well as to the dorsal lateral pontine nuclei, which in turn project to the floccular lobe of the cerebellum, these structures are well situated within a neural network that could function to coordinate combined eye–head gaze pursuit.

Conclusion

In conclusion, the premotor pathways that mediate the VOR are modulated in a manner that depends strictly on the current gaze strategy of the monkey. The head-velocity sensitivities of premotor neurons in this pathway are attenuated for the active head movements that are made during gaze shifts and gaze pursuit. This attenuation of the direct VOR pathways is mediated via inputs from the premotor circuitries that are known to generate saccades and smooth pursuit eye movements. Because the VOR is attenuated during eye–head gaze shifts and pursuit, it follows that the head and the eyes contribute to shifting the axis of gaze relative to space. How the upstream structures that are involved in the planning of these coordinated eye and head movements ensure accurate gaze is an important question that requires future investigation.

See also: Brainstem Control of Eye Movements; Cerebellum and Oculomotor Control; Eye Movement Disorders; Frontal Eye Fields; Oculomotor Control: Anatomical Pathways; Oculomotor System: Models; Pursuit Eye Movements; Saccadic Eye Movements; Supplementary Eye Fields; Vestibular System; Vestibulo-Ocular Reflex; Vestibulospinal System and Eye–Head/Neck Movement; Visual–Vestibular Interactions.

Further Reading

Choi WY and Guitton D (2006) Responses of collicular fixation neurons to gaze shift perturbations in head-unrestrained monkey reveal gaze feedback control. *Neuron* 50(3): 491–505.

Cullen KE, Galiana HL, and Sylvestre PA (2000) Comparing extraocular motoneuron discharges during head-restrained saccades and head-unrestrained gaze shifts. *Journal of Neurophysiology* 83(1): 630–637.

Cullen KE and Guitton D (1997) Analysis of primate IBN spike trains using system identification techniques, II: Relationship to gaze, eye, and head movement dynamics during head-free gaze shifts. *Journal of Neurophysiology* 78: 3823–3306.

Cullen KE, Huterer M, Braidwood DA, and Sylvestre PA (2004) Time course of vestibuloocular reflex suppression during gaze shifts. *Journal of Neurophysiology* 92: 3408–3422.

Cullen KE and Roy JE (2004) Signal processing in the vestibular system during active versus passive head movements. *Journal of Neurophysiology* 91: 1919–1933 (review).

Dubrovsky AS and Cullen KE (2002) Gaze-, eye-, and head-movement dynamics during closed- and open-loop gaze pursuit. *Journal of Neurophysiology* 87(2): 859–875.

Fukushima K (2003) Roles of the cerebellum in pursuit-vestibular interactions. *Cerebellum* 2(3): 223–232.

Grantyn A, Moschovakis AK, and Kitama T (2004) Control of orienting movements: Role of multiple tectal projections to the lower brainstem. *Progress in Brain Research* 143: 423–438 (review).

Guitton D, Bergeron A, Choi WY, and Matsuo S (2003) On the feedback control of orienting gaze shifts made with eye and head movements. *Progress in Brain Research* 142: 55–68.

Huterer M and Cullen KE (2002) Vestibuloocular reflex dynamics during high-frequency and high-acceleration rotations of the head on body in rhesus monkey. *Journal of Neurophysiology* 88: 13–28.

Isa T and Sasaki S (2002) Brainstem control of head movements during orienting: Organization of the premotor circuits. *Progress in Neurobiology* 66(4): 205–241.

Lanman J, Bizzi E, and Allum J (1978) The coordination of eye and head movement during smooth pursuit. *Brain Research* 153: 39–53.

Martinez-Trujillo JC, Medendorp WP, Wang H, and Crawford JD (2004) Frames of reference for eye-head gaze commands in primate supplementary eye fields. *Neuron* 44(6): 1057–1066.

Roy JE and Cullen KE (1998) A neural correlate for vestibulo-ocular reflex suppression during voluntary eye-head gaze shifts. *Nature Neuroscience* 1: 404–410.

Sparks DL (1999) Conceptual issues related to the role of the superior colliculus in the control of gaze. *Current Opinion in Neurobiology* 9(6): 698–707.

Sylvestre PA and Cullen KE (2006) Premotor correlates of integrated feedback control for eye-head gaze shifts. *Journal of Neuroscience* 26(18): 4922–4929.

Tabak S, Smeets JBJ, and Collewijn H (1996) Modulation of the human vestibuloocular reflex during saccades: Probing by high-frequency oscillation and torque pulses of the head. *Journal of Neurophysiology* 76: 3249–3263.

Tomlinson RD (1990) Combined eye-head gaze shifts in the primate, III: Contributions to the accuracy of gaze saccades. *Journal of Neurophysiology* 64: 1873–1891.

Eye Movement Disorders

R J Leigh and M Gross, Case Western Reserve University, Cleveland, OH, USA

Published by Elsevier Ltd.

Introduction

Neuroscientists have traditionally been skeptical of studies of abnormal eye movements in patients with neurological disorders because of doubts about the discreteness of lesions, the quality of experimental design or measurement of eye movements, and the potentially confounding effects of adaptive mechanisms. During the past decade, these concerns have been addressed by development of sensitive neuroimaging techniques, application of sophisticated test paradigms, use of reliable methods for measuring eye movements, and standardization of procedures to control mental set. This means that studies of eye movements now provide an important opportunity to test current hypotheses, forged by basic scientists, about how the brain controls several aspects of eye movements. Most diseases affecting the brain have some effect on eye movements.

In this article, several examples are presented of how study of eye movement disorders has provided insights, as well as instances in which basic research has translated into improved techniques for diagnosis and treatment of patients with neurological diseases. A bottom-up approach is used, proceeding from the extraocular muscles to cerebral cortex. In the final section, pathological forms of nystagmus and saccadic intrusions are summarized. Several functional classes of eye movements are summarized in Table 1.

Disorders of the Ocular Motor Periphery

Movements of each eye are controlled by six extraocular muscles, which originate at the back or side of the orbit and travel to fibromuscular pulleys that are formed by the fascia of orbit. The outer orbital part of each muscle inserts partially on the pulley, and the inner global part passes through the pulley and inserts onto the eyeball. The direction of ocular movements in three dimensions depends upon the geometry of the muscle origins, pulleys, and insertions. During saccades and pursuit movements, the eyes rotate around axes that lie in Listing's plane, which approximates the frontal (coronal) plane of the head. Palsies of the abducens, trochlear, or oculomotor nerves, which supply the extraocular muscles, acutely, but not chronically, cause the eyes to rotate out of Listing's plane. Congenital misplacement of the fibromuscular pulleys may cause some forms of childhood strabismus (misalignment of the eyes).

The unique properties of extraocular muscles result in some unusual responses to systemic disease. Thus, Duchenne muscular dystrophy, an X-linked disorder leading to profound weakness of skeletal muscles and death of affected males in their early twenties, largely spares the extraocular muscles. This insight was gained from the clinical observation that patients with advanced weakness were able to make normal-velocity saccades. It has been postulated that extraocular muscle may be protected either by endogenous upregulation of an analog of the abnormal protein (dystrophin) called utrophin, or by the unique properties of the basement membranes of the extrasynaptic sarcolemma of extraocular muscle. Improved understanding of the reason why extraocular muscles are spared in Duchenne dystrophy may provide insights into possible therapies for this disorder.

In contrast, disorders of the neuromuscular junction have a propensity for affecting eye movements. Botulinum poisoning causes slow saccades with limited range of movement. Myasthenia gravis, the commonest neuromuscular disorder in humans, which is probably due to an immune attack on nicotinic acetylcholine receptors, affects eye movements in up to 80% of affected individuals. This propensity has been attributed to the reduced 'safety factor' that relates neurotransmitter release to receptors available. Other factors, such as reduced expression of decay-accelerating factor, an inhibitor of complement-mediated responses by extraocular muscles, may play a role. Ocular myasthenia is characterized by fatigue (e.g., impaired ability to sustain the eye in eccentric gaze), variability (e.g., saccades of similar size show increased variance of peak velocity), and, ultimately, restricted range of eye movement. Administration of anticholinesterase drugs, such as edrophonium, temporally increases acetylcholine and improves neuromuscular transmission, causing improvement of eye movements. Sometimes edrophonium causes saccadic overshooting (hypermetria), reflecting central adaptive responses to ocular motor weakness.

Disorders Affecting the Brain Stem Contributions to Gaze Control – the Pons

The pons and rostral medulla are important in the control of horizontal gaze. The abducens nucleus is the horizontal conjugate gaze center; it contains motor neurons that innervate the lateral rectus muscle

Table 1 Functional classes of human eye movements

Class of eye movement	*Main function*
Gaze Holding	
Vestibular	Holds images of the seen world steady on the retina during brief head rotations or translations
Visual fixation	Holds the image of a stationary object on the fovea by minimizing ocular drifts
Optokinetic	Holds images of the seen world steady on the retina during sustained head rotation
Gaze Shifting	
Smooth pursuit	Holds the image of a small moving target on the fovea; or holds the image of a small near target on the retina during linear self-motion; with optokinetic responses, aids gaze stabilization during sustained head rotation
Nystagmus quick phases	Resets the eyes during prolonged rotation and directs gaze toward the oncoming visual scene
Saccades	Brings images of objects of interest onto the fovea
Vergence	Moves the eyes in opposite directions so that images of a single object are placed or held simultaneously on the fovea of each eye

Adapted from Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University Press.

and internuclear neurons that project across the midline, via the medial longitudinal fasciculus (MLF), to the contralateral medial rectus motor neurons (Figure 1). Lesions of the abducens nucleus (site 1, Figure 1) cause loss of all conjugate movements toward the side of the lesion, but vergence movements are still possible, because they depend on independent projections to the medial rectus motor neurons. Interruption of the MLF (site 2, Figure 1) causes internuclear ophthalmoplegia (INO), with restriction and slowing of the adducting eye ipsilateral to the lesion. Multiple sclerosis is the most common etiology for an INO in young persons, especially when it is present bilaterally.

Inputs to the abducens nucleus contributing to each functional class of eye movement may be independently affected by disease. Thus, the vestibuloocular reflex (VOR) for horizontal head rotations depends on vestibular afferents from the lateral semicircular canals, which relay their signal to the contralateral abducens nucleus via the medial vestibular nucleus (MVN) (Figure 1). Wernicke's encephalopathy, a disorder in alcoholics due to thiamine deficiency, involves the vestibular nuclei (site 3, Figure 1) and thus impairs the horizontal VOR. The horizontal gaze center also receives command signals for horizontal saccades, signals which project to the abducens nucleus from the adjacent paramedian pontine reticular formation (PPRF); lesions here (site 4, Figure 1) cause slow or absent horizontal saccades.

The nucleus prepositus hypoglossi (NPH), the adjacent MVN, and the cerebellum play an important role in holding the eyes in an eccentric position (e.g., far right gaze) against the elastic pull of the orbital tissues. This function depends on mathematical integration of premotor (visual, vestibular, saccadic) signals by the NPH/MVN-cerebellar network. Impaired function of this network (leaky integration) due, for example, to intoxication with alcohol causes the eyes to drift back to the center, leading to gaze-evoked nystagmus. It has also been postulated that the ocular motor neural integrator network may become unstable, causing either increasing-velocity drifts away from center position or quasi-sinusoidal eye rotations (acquired pendular nystagmus; discussed further elsewhere in this encyclopedia.

Disorders Affecting the Brain Stem Contributions to Gaze Control – the Midbrain

The midbrain plays an important role in the control of vertical conjugate gaze and vergence eye movements. The oculomotor and trochlear nuclei (Figure 2) house the motor neurons that innervate extraocular muscles that mainly rotate the eyes vertically (superior and inferior rectus muscles) or torsionally (around the line of sight; the superior and inferior oblique muscles). These motor neurons receive their saccadic input from the rostral interstitial nucleus of the medial longitudinal fasciculus (riMLF), which lies in the prerubral fields of the rostral midbrain. Bilateral lesions involving the riMLF (site 1, Figure 2) cause slow or absent vertical saccades (such as in progressive supranuclear palsy; PSP). Unilateral lesions of riMLF, such as in rostral brain stem strokes, cause loss of torsional quick phases that rotate the top pole of the eye toward the side of the lesion.

The signals for vertical vestibular and pursuit eye movements ascend from the medulla and pons to the midbrain in the MLF and other pathways. Thus, bilateral MLF lesions (site 2, Figure 2), which occur in multiple sclerosis, cause impaired vertical pursuit and vestibular responses (as well as bilateral adduction failure during horizontal saccades). The interstitial

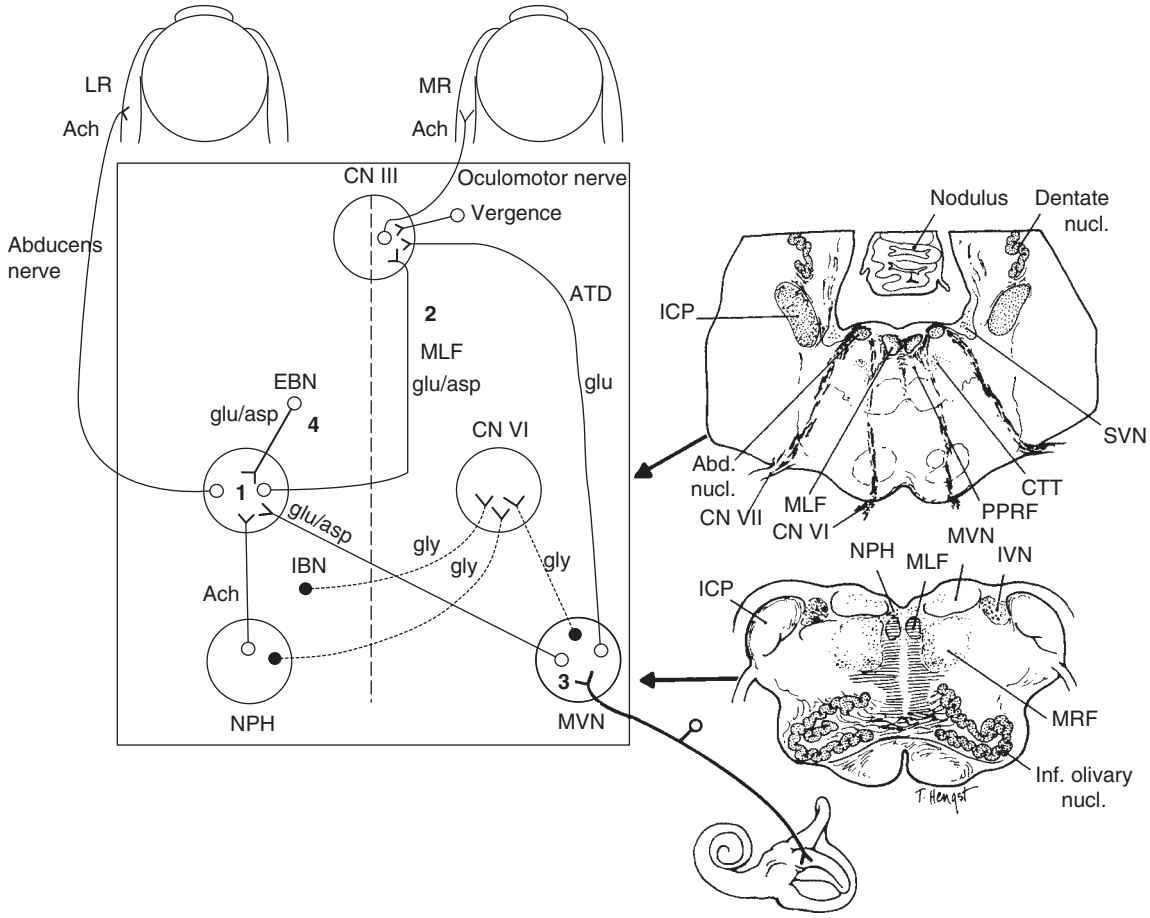

Figure 1 Anatomic scheme for the synthesis of signals for horizontal eye movements. The abducens nucleus (CN VI) contains abducens motor neurons that innervate the ipsilateral lateral rectus muscle (LR), and abducens internuclear neurons that send an ascending projection in the contralateral medial longitudinal fasciculus (MLF) to contact medial rectus (MR) motor neurons in the contralateral third nerve nucleus (CN III). From the horizontal semicircular canal, primary afferents on the vestibular nerve project mainly to the medial vestibular nucleus (MVN), where they synapse and then send an excitatory connection to the contralateral abducens nucleus and an inhibitory projection to the ipsilateral abducens nucleus. Saccadic inputs reach the abducens nucleus from ipsilateral excitatory burst neurons (EBN) and contralateral inhibitory burst neurons (IBN). Eye position information (the output of the neural integrator) reaches the abducens nucleus from neurons within the nucleus prepositus hypoglossi (NPH) and adjacent MVN. The medial rectus motor neurons in CN III also receive a command for vergence eye movements. Putative neurotransmitters for each pathway are shown: Ach, acetylcholine; asp, aspartate; glu, glutamate; gly, glycine. The anatomic sections on the right correspond to the level of the arrowheads on the schematic on the left. Other components: Abd. nucl., abducens nucleus; CN VI, abducens nerve; CN VII, facial nerve; CTT, central tegmental tract; ICP, inferior cerebellar peduncle; IVN, inferior vestibular nucleus; Inf. olivary nucl., inferior olivary nucleus; MRF, medullary reticular formation; PPRF, paramedian pontine reticular formation; SVN, superior vestibular nucleus. Numbers indicate lesion sites that are discussed in the text. Reproduced from Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University Press, with permission.

nucleus of Cajal plays an important role in holding steady vertical eccentric gaze (e.g., far upward gaze); lesions here (site 3, Figure 2) cause gaze-evoked nystagmus on upward or downward gaze (due to a leaky ocular motor integrator). The superior colliculus is a midbrain tectal structure that receives inputs from the cortical eye fields, and is important for triggering both horizontal and vertical saccades. Extensive studies in the macaque monkey have demonstrated the important role of the superior colliculus in target selection and triggering of eye movements (saccades and pursuit). Functional imaging studies in humans have demonstrated activation during generation of short-latency (express) saccades. Lesions restricted to the superior colliculus in humans are rare, but more diffuse disorders (such as PSP) involve this structure.

Disorders Affecting the Cerebellar Contributions to Gaze Control

Three regions of the cerebellum contribute to the control of eye movements. The flocculus and paraflocculus

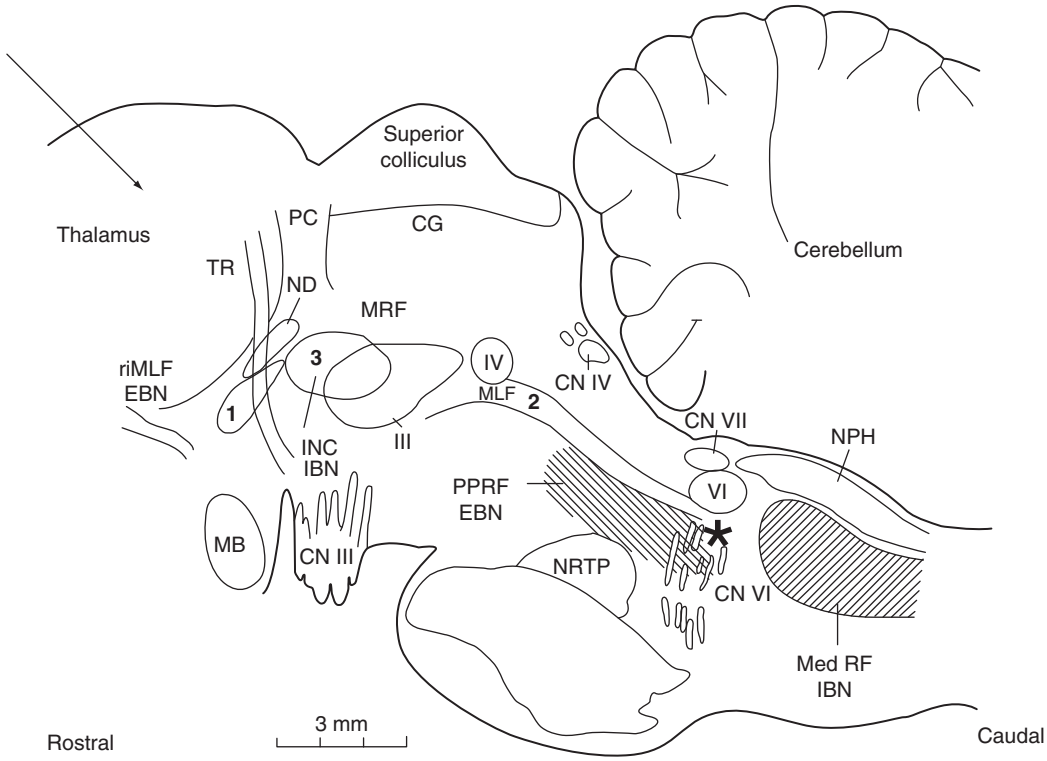

Figure 2 A sagittal section of the monkey brain stem showing the locations of premotor burst neurons: excitatory burst neurons for horizontal saccades lie in the paramedian pontine reticular formation (PPRF) and, for vertical and torsional saccades, lie in the rostral interstitial nucleus of the medial longitudinal fasciculus (riMLF). Burst neurons project to ocular motor neurons lying in the abducens nucleus (VI), the trochlear nucleus (IV), and the oculomotor nucleus (III). Omnipause neurons (indicated by an asterisk) lie in the midline raphe of the pons between the rootlets of the abducens nerve (CN VI) and gate the activity of burst neurons. Other components: CG, central gray; EBN, excitatory burst neuron; IBN, inhibitory burst neuron; MB, mammillary body; CN III, rootlets of the oculomotor nerve; CN IV, trochlear nerve; MRF, medullary reticular formation; ND, nucleus of Darkschewitsch; NPH, nucleus prepositus hypoglossi; NRTP, nucleus reticularis tegmenti pontis; PC, posterior commissure; TR, tractus retroflexus. The arrow refers to the Horsley–Clarke plane of section. Numbers indicate lesion sites that are discussed in the text. Adapted courtesy of Dr. Jean Büttner-Ennever.

are important for smooth pursuit (eyes alone or eye–head tracking), eccentric gaze holding, and adjustment of the VOR so that it is optimized to guarantee clear vision during head perturbations. Lesions of the flocculus and paraflocculus impair each of these functions and also cause downbeat nystagmus. This syndrome is commonly encountered in patients with the Chiari malformation (an abnormality of brain stem and cerebellar development) and with hereditary spinocerebellar ataxias including episodic ataxia type 2, which is due to abnormal P/Q calcium channels and which can be treated with the potassium channel blocker, 4-aminopyridine.

The midline nodulus and adjacent ventral uvula regulate a property of the VOR called velocity storage, which causes the raw vestibular input to sustain vestibular nystagmus after the head has stopped turning. Lesions of the nodulus and ventral uvula cause sustained vestibular responses, which are subject to adaptive repair mechanisms resulting in periodic alternating nystagmus, a form of horizontal nystagmus that reverses direction every 2 min (see later, section titled 'Pathological nystagmus and saccadic intrusions').

The dorsal vermis and the fastigial nucleus to which it projects play an important role in making saccades accurate, perhaps by feedback control of brain stem saccade-generating burst neurons. Dorsal vermis lesions cause saccadic hypometria (undershoots). Fastigial nucleus lesions cause hypermetria (overshoots). Axons from each fastigial nucleus cross in its counterpart, thus, clinical lesions cause bilateral saccadic hypermetria. Wallenberg's syndrome (infarction of the lateral medulla) causes lateropulsion of eye movements (overshooting to the side of the lesion; undershooting to the opposite side). A similar ipsipulsion occurs after experimental inactivation of the fastigial nucleus in macaques, suggesting that lateropulsion in Wallenberg's syndrome is due to interruption of climbing fiber inputs (inferior cerebellar peduncle infarction) to Purkinje cells, which then inhibit fastigial nucleus neurons.

Disorders of the Basal Ganglia That Cause Disordered Gaze Control

The basal ganglia (caudate, substantia nigra pars reticulata, and subthalamic nucleus) are important relay sites for pathways from the cortical eye fields to the superior colliculus and brain stem nuclei. In general, to elicit disorders of eye movements with basal ganglia diseases requires special test paradigms, such as predictive saccades, saccades to remembered targets, and antisaccades (looking in the direction opposite to a visual stimulus). In Parkinson's disease, which affects dopaminergic neurons in the substantia nigra, patients are able to make relatively accurate saccades between two stationary targets, provided they are cued to do so, but saccades become progressively hypometric when they are self-generated, so that gaze shifts become a 'staircase' of saccades to the target. Procedures to treat Parkinson's disease, such as deep brain stimulation of the subthalamic nucleus, are reported to improve the ability to make saccades to remembered targets. Pallidotomy (a stereotactically guided lesion of the globus pallidus) may increase small saccadic intrusions (square-wave jerks), possibly because of interruption of inhibitory projections from the substantia nigra pars reticulata to the superior colliculus.

In Huntington's disease, a diffuse neurodegenerative disorder which prominently affects the caudate nucleus, patients show difficulties initiating saccades but, paradoxically, also have difficulties in suppressing saccades during fixation. These findings can be related to the parallel pathways that control the various types of saccadic responses (Figure 3). Thus, on the one hand, diseases affecting either the frontal lobes or the caudate nucleus, which inhibits the nondopaminergic substantia nigra pars reticulata (SNpr), may lead to difficulties in initiating voluntary saccades. On the other hand, the SNpr also inhibits the superior colliculus in order to suppress reflexive saccades to visual stimuli and, therefore, affected patients exhibit excessive distractibility during attempted fixation. Recent animal studies have emphasized how the basal ganglia pathway is concerned with behavior that is rewarded, but this role has yet to be explored in patients with basal ganglionic disease.

Disorders of the Cerebral Cortex That Affect Gaze Control

The cerebral cortex contains several areas that are important for eye movements (Figure 4). Primary visual cortex (V1) is the gateway for vision; without it, humans cannot make visually guided eye movements. Cortical areas concerned with motion vision lie in cortex at the junction of areas 19, 37, and 39, close to the intersection of the ascending limb of the inferior temporal sulcus and the lateral occipital sulcus. These areas correspond to the middle temporal visual area (MT, or V5) and the medial superior visual temporal area (MST) in the macaque. Patients with lesions affecting the human homolog MT show a defect of eye movements (saccades and pursuit) to targets moving in the contralateral visual hemifield; saccades to stationary targets may be normal. Patients with lesions involving the homolog of MST show a defect of smooth pursuit for targets moving ipsilateral to the side of the lesion. Acute parietal lobe lesions, especially affecting the right hemisphere, produce a classic syndrome of contralateral hemineglect that contributes to an ipsilateral gaze preference. The accuracy of saccades to contralateral targets may also be impaired, but a more impressive dysmetria occurs when patients are required to respond to a double-step stimulus, in which the target jumps twice before a response can be initiated. If the target jumps first into the contralateral hemifield and then into the ipsilateral field, patients cannot make accurate saccades to the final target position, even though it lies in the 'intact' hemifield. This finding has been taken as evidence that the parietal lobe plays a pivotal role in computing target position from both visual stimuli and an internal (efference) copy of eye movements (in this case, the change in eye position due to the first saccade).

The frontal eye field (FEF) is important for voluntary saccades and suppression of saccades during steady fixation. Acute lesions that involve the FEF may produce an ipsilateral gaze deviation. With more chronic FEF lesions, saccades made in response to the overlap paradigm, during which the fixation light remains on when the target light is presented, show increased reaction time, both ipsilaterally and contralaterally. The supplementary eye fields (SEFs), and adjacent presupplementary motor cortex, guide saccades during complex tasks, such as sequences of movements and responses when the instructional set changes. Patients with SEF lesions cannot make a sequence of saccades to an array of visible targets in the order that they were turned on. Another effect of SEF lesions is difficulty in changing the direction of saccades as part of a reversal of a previously established pattern of response. The dorsolateral prefrontal cortex (DLPC) is important for memory-guided saccades, and lesions in this region impair both this function and antisaccades. There is also some evidence that lesions involving the anterior cingulate cortex impair memory-guided saccades and antisaccades.

Pathological Nystagmus and Saccadic Intrusions

Abnormal eye movements that disrupt steady fixation are of two main types: pathological nystagmus and saccadic intrusions. The essential difference between

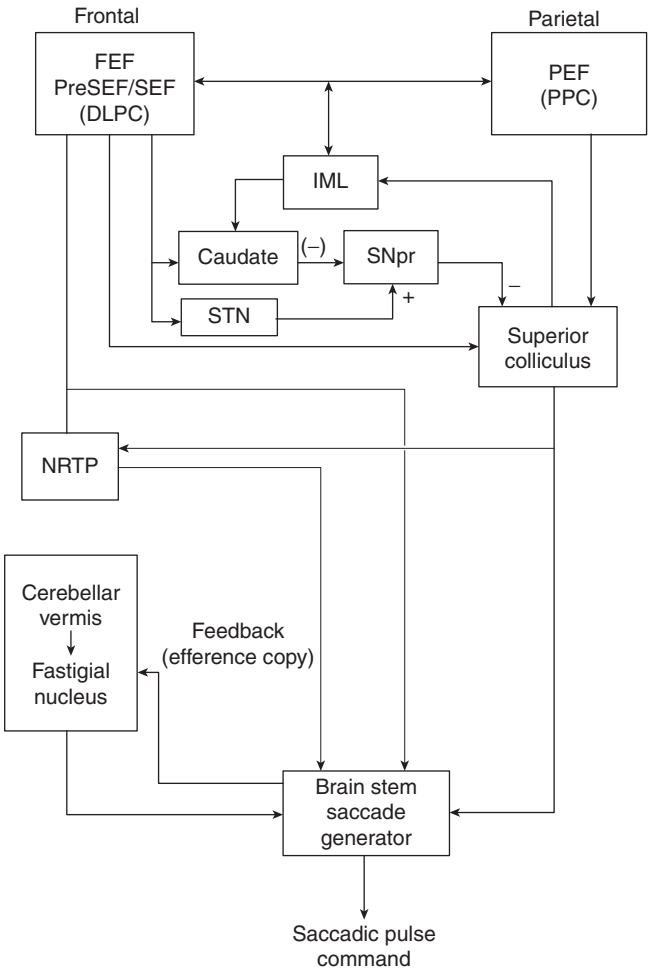

Figure 3 Block diagram of the major structures that project to the brain stem saccade generator (premotor burst neurons in pontine reticular formation and rostral interstitial nucleus of the medial longitudinal fasciculus). Also shown are projections from cortical eye fields to superior colliculus. DLPC, dorsolateral prefrontal cortex; FEF, frontal eye fields; IML, intramedullary lamina of thalamus; NRTP, nucleus reticularis tegmenti pontis; PEF, parietal eye fields; PPC, posterior parietal cortex; SEF, supplementary eye fields; SNpr, substantia nigra pars reticulata; STN, subthalamic nucleus. Not shown are the pulvinar, which has connections with the superior colliculus and both the frontal and parietal lobes, projections from the caudate nucleus to the subthalamic nucleus via globus pallidus, and the pathway that conveys efference copy from brain stem and cerebellum, via thalamus, to cerebral cortex. Reproduced from Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University Press, with permission.

nystagmus and saccadic intrusions lies in the initial eye movement that takes the line of sight away from the visual target. Thus, for nystagmus, the initial movement is a slow drift (or 'slow-phase' – see Figure 5), as opposed to an initial inappropriate quick movement that interrupts steady fixation in saccadic intrusions (Figure 6). After the initial slow movement of nystagmus, corrective or abnormal eye movements may follow. Thus, a definition of nystagmus is repetitive to-and-fro movements of the eyes that are initiated by slow phases. Nystagmus may consist mainly of sinusoidal slow-phase oscillations (pendular nystagmus – Figure 5(d)) or, more commonly, of an alternation of slow drift and corrective quick phase (jerk nystagmus – Figures 5(a)–5(c)). Although nystagmus is often described by the direction of its quick phases (e.g., downbeat nystagmus), it is the slow phase that indicates the underlying disorder. Not all nystagmus is pathological. Thus, when we turn our heads, the slow phases of physiological nystagmus – vestibular and optokinetic – act to preserve clear vision by preventing excessive slip of images on the retina; quick phases reset the eyes into their working range. In contrast, saccadic intrusions are fast movements that take the eye away from the target (Figure 6), and comprise a spectrum ranging from single saccades to sustained saccadic oscillations.

Under normal circumstances, three mechanisms collaborate to prevent the line of sight drifting away from the object of regard. Disorders of each of these

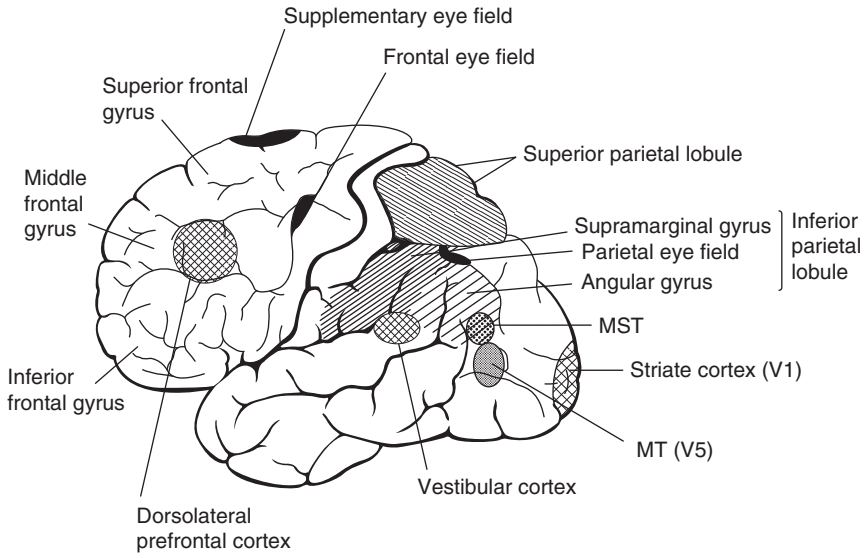

Figure 4 Probable location of cortical areas important for eye movements in human brain. The medial superior temporal visual area (MST) and middle temporal visual area (MT) may form a contiguous cortical area. Reproduced from Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University Press, with permission.

mechanisms disrupt steady gaze and lead to nystagmus. The first mechanism is the VOR, by which eye movements compensate for head perturbations at short latency and thereby maintain clear vision during natural activities, especially locomotion. The second mechanism (previously referred to as the neural integrator for eye movements) concerns the brain's ability to hold the eye at an eccentric position in the orbit against the elastic pull of the eyeball's suspensory ligaments and muscles. The third mechanism is visual fixation, which has three components. The first component is visual suppression of unwanted saccades during focused tasks, such as threading a needle. The second component is the ability to detect retinal image drift and rapidly program both slow and saccadic corrective eye movements. The third component is a more long-term effect of visual inputs that monitor the visual consequences of eye movements and continuously tunes them to optimize gaze stability.

Nystagmus Due to Peripheral or Central Imbalance

An important characteristic of the vestibular system is its resting discharge rate ('vestibular tone'). In response to motion stimuli, such as rotation, inputs from one end organ (the vestibular labyrinth) will increase above this background rate while the contralateral one will decrease. Thus, vestibular tone bestows increased sensitivity to motion stimuli due to this push–pull mechanism. However, the consequence of vestibular tone is that, if disease affects one labyrinth, the discharge rate may fall, and the brain will misinterpret the difference between the two end organs as self-rotation (known clinically as vertigo) and generate eye movements to compensate for the illusory motion (vestibular nystagmus). Such nystagmus (Figure 5(a)) is often mixed horizontal–torsional in direction (reflecting the sum of inputs from the three semicircular canals that sense motion in each ear) and can be suppressed by visual fixation, since central mechanisms for gaze stabilization are intact.

Imbalance may also arise within the central vestibular system. For example, the flocculus of the cerebellum governs dynamic properties of the VOR and, for vertical head rotations, achieves this through an asymmetric inhibition. Consequently, lesions of the flocculus tend to cause upward drifts of the eyes with downward quick phases (downbeat nystagmus). As noted previously, the nodulus and ventral uvula govern 'velocity storage' – perseveration of the rotational vestibular signal – so that the nystagmus response is prolonged beyond the duration of the head rotation. Lesions of the nodulus and ventral uvula further prolong velocity storage, then adaptive mechanisms attempt to correct the prolonged responses, so that the nystagmus periodically reverses, resulting in periodic alternating nystagmus, which changes direction approximately every 2 min. Periodic alternating nystagmus can be successfully abolished by the γ-aminobutyric acid ($GABA_B$)ergic drug baclofen.

Nystagmus Due to Disturbance of the Neural Integrator for Eye Movements

As already noted, if the gaze-holding mechanism is deficient, the eyes cannot be held steadily in an eccentric orbital position, but drift back to the midline with

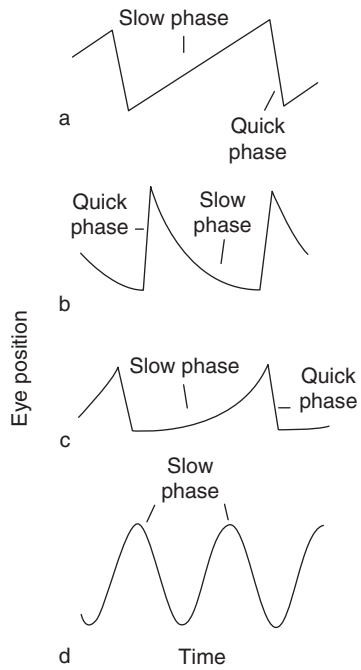

Figure 5 Four common slow-phase waveforms of nystagmus. (a) Constant velocity drift of the eyes. This occurs in nystagmus caused by peripheral or central vestibular disease and also with lesions of the cerebral hemisphere. The added quick phases give a 'sawtooth' appearance. (b) Drift of the eyes back from an eccentric orbital position toward the midline (gaze-evoked nystagmus). The drift shows a negative exponential time course, with decreasing velocity. This waveform reflects an unsustained eye position signal caused by an impaired neural integrator. (c) Drift of the eyes away from the central position, with a positive exponential time course (increasing velocity). This waveform suggests an unstable neural integrator and is encountered in the horizontal plane in congenital nystagmus and in the vertical plane in cerebellar disease. (d) Pendular nystagmus, which is encountered as a type of congenital nystagmus and with acquired disease. Reproduced from Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University Press, with permission.

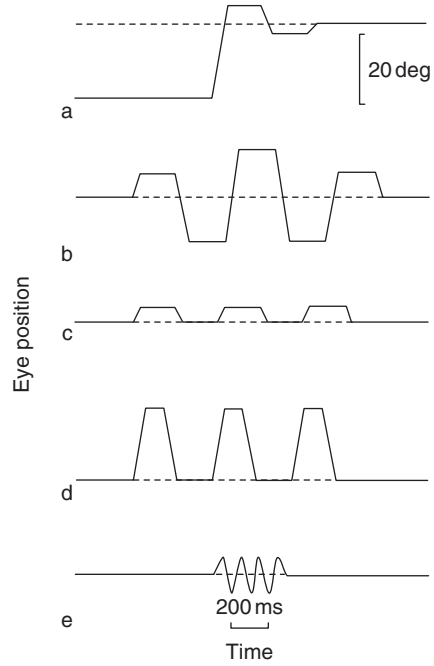

Figure 6 Schematic of saccadic intrusions and oscillations. (a) Dysmetria: inaccurate saccades. (b) Macrosaccadic oscillations: hypermetric saccades about the position of the target. (c) Square-wave jerks: small, uncalled-for saccades away from and back to the position of the target. (d) Macrosquare-wave jerks or macrosaccadic pulses: large, uncalled-for saccades away from and back to the position of the target. (e) Ocular flutter: to-and-fro, back-to-back saccades without an intersaccadic interval. Reproduced from Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University Press, with permission.

a decreasing-velocity waveform (Figure 5(b)); this is gaze-evoked nystagmus. Because the gaze-holding mechanism depends, in part, upon the vestibular nuclei, nystagmus due to brain stem lesions often manifests the properties of both vestibular imbalance and disturbed gaze holding. A wide range of disorders affecting the brain stem and cerebellum can cause gaze-evoked nystagmus, along with a range of sedative medicines and alcohol. Instability in the gaze-holding mechanism may lead to a slow-phase drift that increases in velocity (Figure 5(c)); this is described in patients with cerebellar disorders. It is also possible that the ocular motor neural integrator network may become unstable, causing either increasing-velocity drifts away from center position or quasi-sinusoidal eye rotations (acquired pendular nystagmus – Figure 5(d)). Determination of the neurotransmitters involved in the ocular motor neural integrator (GABA and glutamate) using the technique of pharmacological inactivation in monkeys has led to development of treatments for acquired pendular nystagmus in patients (e.g., the noncompetitive glutamate antagonist, memantine).

Nystagmus Occurring with Disorders of the Visual Pathways

Individuals who are born blind have continuous nystagmus which is constantly changing in direction (shifting null). Such nystagmus reflects an uncalibrated ocular motor system. Nystagmus also occurs with a range of less severe congenital visual disorders, such as in albinism, in which anomalies of the visual pathways are documented. Congenital nystagmus, or infantile nystagmus syndrome, may also occur in individuals who have normal visual systems, and thus may reflect a different, poorly understood mechanism. In adulthood, loss of vision due, for example, to optic nerve demyelination in multiple sclerosis may be associated with acquired pendular nystagmus.

Although it seemed possible that delays in the visual system could lead to these quasi-sinusoidal oscillations, experimental manipulation of visual delay using electronic feedback did not change the frequency of oscillations. A more plausible hypothesis, noted in the prior section, is that acquired pendular nystagmus arises from instability of the neural integrator mechanism. Other forms of nystagmus exist for which there is no good explanation; these are discussed further in the text on eye movement disorders, *The Neurology of Eye Movements*, by Leigh and Zee.

Saccadic Intrusions and Oscillations

The range of saccadic intrusions and oscillations are summarized schematically in Figure 6. Square-wave jerks are small, uncalled-for saccades that take the eye away from the fixation point and return it within 200–400 ms. Normal subjects show square-wave jerks and, in some individuals they are frequent. However, square-wave jerks are most prominent with certain disease states, especially PSP and Friedreich's ataxia. Macrosaccadic oscillations are larger saccades that oscillate the eye about the point of fixation. They usually occur in patients with midline cerebellar lesions involving the fastigial nucleus, which also cause saccadic hypermetria. Ocular flutter consists of horizontal high-frequency, back-to-back saccades without an intersaccadic interval. When such oscillations have horizontal, vertical, and torsional components they are referred to as opsoclonus. Current evidence suggests that flutter and opsoclonus, as well as brief, low-amplitude, high-frequency oscillations in normal subjects, may be generated by a brain stem network of burst neurons, although disturbed cerebellar feedback of saccades might also contribute. With improved knowledge of the pharmacology of gaze-holding and saccadic systems, new drug treatments for these disorders are becoming available.

Conclusions

This overview of abnormal eye movements briefly describes only a few examples from a broad range of disorders that are now fairly well understood. These and other examples may provide insights into current models of normal eye movements. If the goal is to develop neuromimetic models for eye movements, then abnormal eye movements are well suited to identify biologically plausible hypotheses. Indeed, certain disorders, such as slow saccades, have played a pivotal role in development of current models. Another important potential outcome of such a translational approach is development of new therapies – for example, the identification of drug treatmen ts for acquired pendular nystagmus based on the pharmacology of the neural integrator for eye movements.

See also: Brainstem Control of Eye Movements; Cerebellum and Oculomotor Control; Cortical Control of Eye Movements; Eye and Head Movements; Nystagmus; Saccade–Pursuit Interactions; Saccades and Visual Search; Saccades and Visual Search; Vestibulo-Ocular Reflex; Visual–Vestibular Interactions.

Further Reading

Ali FR, Michell AW, Barker RA, et al. (2006) The use of quantitative oculometry in the assessment of Huntington's disease. *Experimental Brain Research* 169(2): 237–245.

Chan F, Armstrong IT, Pari G, et al. (2005) Deficits in saccadic eye-movement control in Parkinson's disease. *Neuropsychologia* 43: 784–796.

Das VE, Oruganti P, Kramer PD, et al. (2000) Experimental tests of a neural-network model for ocular oscillations caused by disease of central myelin. *Experimental Brain Research* 133: 189–197.

Demer JL (2004) Pivotal role of orbital connective tissues in binocular alignment and strabismus. *Investigative Ophthalmology and Vision Science* 45: 729–738.

Hughes BW, Moro De Casillas ML, and Kaminski HJ (2005) Pathophysiology of myasthenia gravis. *Seminars in Neurology* 24: 21–30.

Leigh RJ and Kennard C (2004) Using saccades as a research tool in the clinical neurosciences. *Brain* 127: 460–477.

Leigh RJ and Tomsak R (2003) Drug treatments for eye movement disorders. *Journal of Neurology, Neurosurgery and Psychiatry* 74: 1–4.

Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. New York: Oxford University Press.

Ramat S, Leigh RJ, Zee DS, et al. (2005) Ocular oscillations generated by coupling of brainstem excitatory and inhibitory saccadic burst neurons. *Experimental Brain Research* 160: 89–106.

Strupp M, Kalla R, Dichgans M, et al. (2004) Treatment of episodic ataxia type 2 with the potassium channel blocker 4-aminopyridine. *Neurology* 62: 1623–1625.

Waespe W and Baumgartner R (1992) Enduring dysmetria and impaired gain adaptivity of saccadic eye movements in Wallenberg's lateral medullary syndrome. *Brain* 115: 1125–1146.

Eye Tracking and Mental Illness

D L Levy, McLean Hospital, Belmont, MA, USA
J A Sweeney, University of Illinois, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

The study of eye tracking in psychiatric populations began in 1908, when Diefendorf and Dodge first observed an association between abnormal visual pursuit tracking and the illness that we now refer to as schizophrenia. Included among their study populations were hospitalized individuals with diagnoses of dementia praecox (this term was later changed to schizophrenia), manic-depressive psychosis (now termed bipolar disorder), and organic brain disorders such as epilepsy and neurosyphilis. Subjects were asked to follow a pendulum as it moved back and forth while an ingeniously designed camera system recorded their eye movements. The key finding was that impaired smooth pursuit eye movements were characteristic of patients with a diagnosis of dementia praecox but were unusual in patients with mood and organic disorders. Because of the high degree of specificity in distinguishing dementia praecox from the other conditions, the authors labeled the phenomenon 'praecox pursuit'.

This intriguing finding went largely unnoticed for the next 65 years. By 1938, only two additional studies of oculomotor function in psychiatric populations had been published. Another 35 years would pass before Philip S Holzman and his colleagues rediscovered the association between eye-tracking dysfunction (ETD) and schizophrenia. In the context of studying vestibular function, they serendipitously found that ETD was present in 52–86% of schizophrenic patients. Chronically hospitalized patients showed a higher rate of ETD than patients who had been hospitalized for less than 6 months. This frequency of ETD was much higher than that found in patients with nonschizophrenic psychotic conditions (22%), nonpsychotic disorders (21%), or normal controls (8%). Scores of studies have since replicated the finding that ETD, or impaired smooth pursuit eye movements, is highly correlated with the diagnosis of schizophrenia. ETD is found during acute episodes as well as in remitted states, consistent with it being a trait rather than a secondary effect of an acute psychotic state. Moreover, ETD occurs in first-episode never-medicated patients, indicating that it is present at the onset of the illness and is thus not a function of chronicity, a degenerative process, or exposure to antipsychotic medication. Of particular importance ETD occurs at a much higher rate than schizophrenia in first-degree biological relatives of schizophrenics.

Quantitative Characterization of Eye-Tracking Dysfunction

During the first 15 years of modern studies of ETD in schizophrenia, performance was evaluated using one of two methods: a qualitative rating of normal or abnormal pursuit based on visual inspection of the degree to which the eye movement record matched the target waveform (see Figure 1) or global quantitative scores, such as the ln S/N or root mean square error. These measures yielded global indices of performance, but could not help to identify the nature of the abnormality, which, in turn, is relevant for localizing which brain systems are involved. More recent work has adopted the quantitative measures that are now routinely used to assess oculomotor function in clinical and basic science settings.

The smooth pursuit and saccadic eye movement systems work interactively during an eye-tracking task. Typical targets move predictably at a constant velocity (usually 10–30°s^{-1}) or sinusoidally (0.4 Hz). Smooth pursuit is initiated when a target moves across the visual field. The pursuit system attempts to stabilize the image of the target on the fovea by matching eye velocity to target velocity. The better the match between eye velocity and target velocity, the closer the ratio of these two values, or closed loop gain, is to 1.0. The more the eyes are moving too slowly and lag behind the target, the lower the gain. Two types of fast eye movements, or saccades, occur during pursuit. Compensatory saccades correct for pursuit that falls behind (catch-up saccades) or gets ahead (back-up saccades) of the target. Thus, low gain is typically associated with an increase in the number (and sometimes the size) of catch-up saccades. Intrusive saccades, on the other hand, are not corrective and are ordinarily inhibited during pursuit. The two most common types of intrusive saccades found in psychiatric populations are square wave jerks and anticipatory saccades. Both kinds of saccades take the eyes off the target. Anticipatory saccades are generally larger than 5° in amplitude and move the eyes ahead of the target, where the eyes either wait for it, or make a back-up saccade, before resuming pursuit. Square wave jerks are pairs of small amplitude (up to 5°) saccades separated by 130–450 ms, during which pursuit continues; the first intrusive saccade takes the eyes off the target and the second saccade repositions the eyes on the target.

Numerous studies have evaluated eye tracking using these quantitative measures of performance. ETD in schizophrenia seems to reflect two processes. The predominant pattern is reduced gain and increased

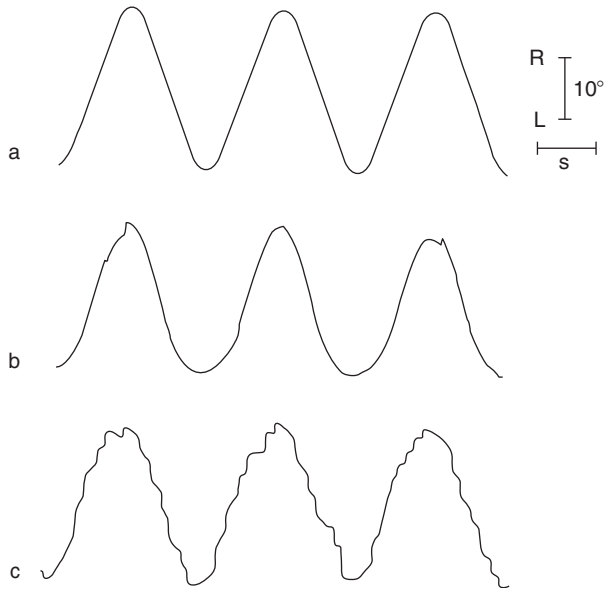

Figure 1 Eye tracking patterns in normal and abnormal pursuit: (a) 0.4 Hz sinusoidal target, (b) qualitatively normal eye tracking of target in (a, c) qualitatively abnormal eye tracking of target in (a).

rates of catch-up saccades (see Figure 1(c)), implicating a disturbance in the smooth pursuit system involving slowed eye velocity accompanied by increased compensatory catch-up saccades. Some evidence suggests that schizophrenics may also have an increased tolerance for position error. This results in sluggishness to initiate catch-up saccades, which increases their amplitude. An increased frequency of square wave jerks and anticipatory saccades also contribute to ETD, but to a lesser extent, and implicate disinhibition of the saccadic system.

Specificity

Numerous studies have reported that ETD is specific to schizophrenia and is not characteristic of nonschizophrenic psychiatric disorders. ETD appears to be a trait in schizophrenia. ETD is stable, and it is independent of medication effects or changes in clinical state. ETD is also present from the onset of the illness. Consistent with the data supporting the specificity of ETD for schizophrenia, some studies have reported that ETD is not characteristic of first-episode mood disorder patients. Moreover, ETD occurs in a significant proportion of clinically unaffected first-degree biological relatives of schizophrenic patients (~45%) but is atypical in relatives of nonschizophrenic patients. ETD is also found in childhood-onset schizophrenics, and a similar pattern of impaired pursuit (low gain and increased catch-up saccades) has been reported in their biological relatives, supporting the notion that childhood- and adult-onset forms of schizophrenia exist along a continuum.

Other studies, however, have found ETD in bipolar disorder (manic-depressive illness) and major depressive disorder as well as in nonpsychotic disorders such as obsessive–compulsive disorder and autism. For example, pursuit deficits of equivalent severity to those seen in schizophrenic patients have been reported in some samples of never-medicated first-episode affective disorder patients. The state–trait issue has not been examined extensively in relation to nonschizophrenic disorders. Some evidence indicates that ETD in bipolar disorder and major depressive disorder may be state-related. For example, certain pharmacological treatments can induce ETD in patients with affective disorders. Lithium carbonate has been shown to induce ETD in bipolar patients, and the monoamine oxidase inhibitor medication, phenelzine, causes pursuit abnormalities in patients with major depressive disorder. Conversely, ETD has been reported to improve following electroconvulsive therapy. Although ETD is clearly not treatment-induced when it occurs in first-episode never-medicated patients, it remains unclear whether it is stable in nonschizophrenic patients or normalizes as a function of some treatments. Whether such changes reflect the primary effects of the treatment intervention or secondary effects of clinical improvement remains unclear. Longitudinal studies are necessary to determine whether the nonspecificity of ETD in some samples is restricted to the acute phase of the illness or is a more enduring trait. As with many biological findings in psychiatry, abnormalities that are relatively specific to one disorder are often seen in more modest form and less frequently in other disorders. This seems to be the case with ETD as well, and may provide important clues to issues of heterogeneity and the boundaries between mental disorders.

Psychotropic Medications

Most studies addressing the issue of medication effects were carried out when first-generation antipsychotic drugs (e.g., haloperidol and fluphenazine) were the primary pharmacological treatment for schizophrenia. These antipsychotic medications are potent blockers of D2 dopamine receptors in the basal ganglia, resulting in extrapyramidal side effects resembling the motor disturbances seen in Parkinson's disease. The presence of impaired pursuit in individuals with idiopathic Parkinson's disease raised the possibility that ETD in schizophrenic patients could in part be a secondary effect of treatment with these medications.

Several lines of evidence indicate that ETD is not caused by treatment with antipsychotic medications. First, ETD (as well as low gain and increased catch-up

saccades) is found in schizophrenic patients who have never been exposed to antipsychotic medication. Second, performance is stable in never-before-medicated first-episode patients who are followed longitudinally while receiving drug treatment. Third, normal pursuit occurs in medicated schizophrenics, indicating that drug treatment does not sufficiently account for ETD. Fourth, indices of performance do not differ in first-episode patients who have never received antipsychotic drug treatment, unmedicated chronic schizophrenics with prior drug exposure, and currently medicated patients. Fifth, the dose of antipsychotic medication is generally not correlated with indices of performance. Sixth, eye tracking tends to be stable, despite changes in clinical state and medication status. Seventh, ETD is found at an elevated prevalence in nonschizophrenic individuals with schizotypal traits (e.g., perceptual aberrations and magical thinking) who are not treated with antipsychotic medication. Of particular importance, ETD occurs in a substantial proportion of the clinically well relatives of schizophrenics who have never been exposed to antipsychotic drugs. The latter two findings provide especially strong support for the conclusion that ETD occurs independently of a schizophrenic psychosis or its treatment. Thus, antipsychotic medications do not seem to affect ETD in a robust consistent manner.

Although current or past treatment with antipsychotic medication is not a necessary condition for the presence of ETD, several caveats warrant recognition. Drug-withdrawal periods in studies of chronic patients tend to be relatively short compared with the length of prior exposure. Most of the information about medication effects is based on first-generation antipsychotics, and much less data are available on patients treated with the second generation, or atypical, antipsychotics. In some studies, but not others, currently unmedicated schizophrenics with prior exposure to antipsychotics have worse eye tracking than never-medicated patients. It is unclear whether this difference reflects an effect of chronic prior drug exposure on the smooth pursuit system *per se* or heightened sensitivity to the effects of medication withdrawal on attentional functioning in patients with a protracted history of drug treatment. Of particular relevance in this regard, antipsychotic drug treatment improves certain aspects of performance, perhaps by improving attentional engagement in the task. This effect is most evident in reduced anticipatory saccades, but can also be manifested in decreased catch-up saccade amplitude, indicating more efficient correction of position error. Thus, the data are most robust in excluding acute rather than chronic effects of typical antipsychotic medications as a cause of ETD.

Genetic Studies – Familial Aggregation, Twin Studies, and Linkage Studies

ETD occurs in first-degree biological relatives of schizophrenics at a rate (~45%) that is much higher than that of the general population (~8%); this finding has been replicated numerous times. The increased prevalence of ETD in relatives of schizophrenics suggests that ETD may be genetically transmitted. Studies of psychiatrically healthy twins as well as of twins discordant for schizophrenia are consistent with genetic control of eye-tracking performance. In two samples of twins discordant for schizophrenia, eye-tracking concordance was 0.77 in monozygotic twins and 0.39 in dizygotic twins. These figures are approximately 80% of those expected for traits under genetic control. The primary features of ETD in relatives are the same as those seen in patients – decreased gain and increased catch-up saccades. In contrast, relatives of patients with nonschizophrenic disorders do not show elevated rates of ETD. Thus, although patients with some other psychiatric disorders may have ETD, ETD does not seem to be familial in nonschizophrenic disorders.

ETD in a schizophrenic patient usually coincides with ETD in at least one first-degree relative. However, there are exceptions. In some instances, normal eye tracking is found in both the patient and in family members. In other cases, patients have normal eye tracking but family member(s) have ETD. The Mendelian latent trait model was developed in order to account for three observations: (1) ETD and schizophrenia frequently occur together in patients, (2) the two traits are occasionally dissociated in patients and are usually dissociated in relatives, and (3) ETD occurs in relatives of schizophrenics at a significantly higher prevalence than does schizophrenia. The empirical distribution of schizophrenia and ETD (in schizophrenia families) is consistent with the proposition that schizophrenia and ETD are independent manifestations of a latent trait that is transmitted by a single autosomal dominant gene. The latent trait has a much higher probability of being expressed as ETD than as schizophrenia. Thus, even though the genetic data on schizophrenia are not consistent with single-gene transmission, a single autosomal dominant gene can account for the distribution of ETD, schizophrenia, or both in schizophrenia families.

ETD has a much higher recurrence risk than schizophrenia (~6.5%) among first-degree relatives of schizophrenics. This observation raises the possibility that ETD is a more penetrant expression of a schizophrenia susceptibility gene than the clinical disorder. In that case, a gene would have a higher probability of being detected through its expression as ETD than

through its expression as schizophrenia – a locus that is one of several for schizophrenia (i.e., oligogenic) can also be a major gene for ETD. The advantages of incorporating highly penetrant disease-related traits, or endophenotypes, into linkage studies has led to the successful identification of genes for genetically complex nonpsychiatric disorders (e.g., long QT cardiac syndrome and juvenile myoclonic epilepsy). The endophenotype strategy has also yielded promising results in linkage studies of schizophrenia. ETD has been linked to loci on chromosome 6p in two independent studies. Another endophenotype, P50 inhibition deficits, has been linked to the α-7 nicotinic receptor gene on chromosome 15q14.

Pathophysiology

The importance of the finding of ETD in schizophrenia is strengthened by the fact that abnormal smooth pursuit can result from a wide range of pathology affecting diverse regions of the central nervous system. This includes the middle temporal and medial superior temporal areas, the frontal eye fields, posterior parietal cortex, parts of the cerebellum, the pontine nuclei, basal ganglia, and brain stem regions, all of which play a role in generating smooth pursuit eye movements. Much of what is known about the cortical processing of moving targets comes from single-cell recording and lesion studies of nonhuman primates and humans with neurological damage in these brain regions. Different regions optimize specific parameters of pursuit (and saccades). Therefore, the pattern of oculomotor deficits and/or normal function provides localizing information about the involvement of specific brain regions. In lesions of striate and extrastriate areas, for example, the accuracy of visual motion information is compromised, resulting in effects on both the pursuit and saccadic systems. Pursuit initiation is slowed and saccades to moving targets are hypometric. Effects on steady-state gain are more modest, presumably because the frontal eye fields and posterior parietal cortex use perceptual feedback and extraretinal signals to compensate for faulty sensory input. Damage to the frontal eye fields, on the other hand, affects pursuit initiation and steady-state gain, but has little effect on the accuracy of saccades to moving targets, indicating that processing of early sensory motion information is intact.

In order to localize the brain regions involved in ETD in schizophrenia, three complementary strategies have been used. The first uses functional brain imaging (positron emission tomography and functional magnetic resonance imaging) to identify different patterns of regional brain activation in schizophrenics and controls while performing an eye-tracking task. The second capitalizes on the availability of experimental paradigms that provide localizing information about the role of specific brain regions in subserving the different components of pursuit and saccadic responses. The third employs psychophysical methods to parse ETD into simpler components.

Brain Imaging

The results of brain imaging studies of schizophrenic patients consistently implicate a disturbance in a frontotemporal network in individuals with ETD. Smooth pursuit deficits have been associated with hyperactivation of the posterior hippocampus and right fusiform gyrus as well as lower activation of the frontal eye fields, cingulate gyrus, and occipital cortex. Consistent with this finding, ETD is associated with reduced resting frontostriatal glucose metabolism and reduced blood flow in the frontal cortex. Moreover, the magnitude of activation in the right dorsal frontal eye fields is significantly correlated with pursuit gain in clinically unaffected relatives of schizophrenics. Schizophrenics also show reduced activation in the frontal eye fields and supplementary eye fields, medial superior temporal cortex, and anterior cingulate and increased, possibly compensatory, activation in medial occipitotemporal cortex. Neuronal activity in motion sensitive area V5 is also abnormal in schizophrenics during pursuit. The involvement of frontal regions in pursuit deficits is consistent with the finding that ETD is correlated with performance on tasks (e.g., spatial working memory) that are thought to depend on the functional integrity of more rostral frontal lobe regions. Thus, widely distributed brain abnormalities appear to contribute to the manifestation of ETD.

Oculomotor Paradigms

Most studies of ETD have used targets that moved sinusoidally or at constant velocities. Such targets can establish that an eye-tracking abnormality is present. However, because target movement is highly predictable, the relative contributions of early sensorimotor (e.g., visual motion information mediated in V5 and the human homologs of the middle temporal and medial superior temporal areas in the monkey) and later higher cortical areas (e.g., frontal eye fields and posterior parietal cortex) to pursuit dysfunction cannot be distinguished. Several recent studies of schizophrenic patients have employed oculomotor paradigms that are less dependent on predictive inputs and that can better separate the early sensory processing of motion signals from later higher-order cortical involvement. The results showed

that sensory input is not effectively used by the pursuit system, as seen in reduced open-loop gain and delayed initiation of pursuit. Closed-loop gain was also reduced but not as much as open-loop gain, suggesting that extraretinal information and perceptual feedback can partially compensate for the faulty use of motion input by the pursuit system. In contrast, saccade latency and accuracy to moving targets were normal, indicating that the saccadic system receives accurate sensory information. This dissociation has been interpreted to suggest that the processing of motion signals is intact at the early stage but that higher-order cortical areas (e.g., frontal eye fields and posterior parietal cortex) are impaired in making optimal use of retinal and extraretinal input for generating accurate pursuit eye movements. Related work suggests that schizophrenic patients and their relatives use retinal error signals excessively to compensate for a deficit in processing extraretinal signals.

Psychophysical Methods

The possibility that a deficit in motion processing contributes to ETD was suggested by three findings. First, lesions of the motion-sensitive areas of the brain (the middle temporal and medial superior temporal areas of the parietal lobe) result in smooth pursuit dysfunction, deficits in detection of coherent motion, reduced open-loop acceleration and impaired ability to use motion cues. Second, open-loop acceleration during pursuit initiation, a response based solely on visual motion signals, is significantly reduced in schizophrenics. Third, reduced steady-state pursuit gain in schizophrenics suggests that impaired velocity detection may be a factor in ETD. In light of these considerations, several studies have examined various aspects of motion processing. Schizophrenics (and their relatives) have increased velocity sensitivity thresholds, but perform normally on visual tasks that do not involve the detection of motion (e.g., contrast sensitivity and orientation). This diminished velocity discrimination is significantly correlated with difficulty generating smooth pursuit as reflected in decreased open-loop acceleration. Moreover, schizophrenics use nonmotion cues to compensate for diminished velocity discrimination and have elevated thresholds for detecting the direction of coherent motion, implicating a deficit in processing global motion. The results of these studies implicate a disturbance in early posterior cortical sensory processing in ETD, whereas the findings of studies that used oculomotor paradigms came to the opposite conclusion. One potential way to resolve this inconsistency would be to examine the same subjects on both kinds of tasks. It is possible, for example, that pursuit initiation deficits may be related to a visual motion processing deficit but that the impairment in steady-state pursuit may be caused by other deficits in the generation or use of extraretinal signals.

In summary, ETD appears to reflect dysfunction in multiple brain regions, although recent research suggests that disturbances in specific brain systems, and the cognitive processes they subserve, may contribute in separate ways to visual tracking deficits. Future research is needed to better localize the neural substrate of ETD in patients and unaffected relatives using functional imaging and to link those regional disturbances to the genotypes related to ETD. In so doing, we can further extend the use of ETD as a window into brain abnormalities in schizophrenia and the genetics of the disorder.

See also: Attention and Eye Movements; Cognitive Deficits in Schizophrenia; Cognitive Dysfunction in Psychiatric Disorders; Cortical Control of Eye Movements; Eye Movement Disorders; Neuroleptics; Pursuit Eye Movements; Saccade–Pursuit Interactions; Saccades and Visual Search; Saccadic Eye Movements; Spatial Orientation: Our Whole-Body Motion and Orientation Sense.

Further Reading

Abel L and Ziegler A (1988) Smooth pursuit eye movements in schizophrenics – what constitutes quantitative assessment? *Biological Psychiatry* 24: 747–761.

Arolt V, Lencer R, Purmann S, et al. (1999) Testing for linkage of eye tracking dysfunction to markers on chromosomes 6, 8, 9, 20, 22 in families multiply affected with schizophrenia. *American Journal of Medical Genetics* 88: 603–606.

Chen Y, Nakayama K, Levy DL, Matthysse S, and Holzman PS (1999) Psychophysical isolation of a motion processing deficit in schizophrenics and their relatives and its association with impaired smooth pursuit. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4724–4729.

Clementz BA and McDowell JE (1994) Smooth-pursuit in schizophrenia: Abnormalities of open-loop and closed-loop responses. *Psychophysiology* 31: 79–86.

Gooding DC, Iacono WG, and Beiser M (1994) Temporal stability of smooth-pursuit eye tracking in first-episode psychosis. *Psychophysiology* 31: 62–67.

Holzman PS, Kringlen E, Matthysse S, et al. (1988) A single dominant gene can account for eye tracking dysfunctions and schizophrenia in offspring of discordant twins. *Archives of General Psychiatry* 45: 641–647.

Holzman PS, Proctor LR, Levy DL, et al. (1974) Eye tracking dysfunctions in schizophrenic patients and their relatives. *Archives of General Psychiatry* 31: 143–151.

Iacono WG, Moreau M, Beiser M, Fleming J, and Lin T (1992) Smooth-pursuit eye tracking in first-episode psychotic patients and their relatives. *Journal of Abnormal Psychology* 101: 104–116.

Levy DL, Holzman PS, Matthysse S, and Mendell NR (1993) Eye tracking dysfunction and schizophrenia: A critical perspective. *Schizophrenia Bulletin* 19: 461–536.

Levy DL, Lajonchere CM, Dorogusker B, et al. (2000) Quantitative characterization of eye tracking dysfunction in schizophrenia. *Schizophrenia Research* 42: 171–185.

Matthysse S, Holzman PS, Gusella J, et al. (2004) Linkage of eye movement dysfunction to chromosome 6p in schizophrenia: A confirmation. *American Journal of Human Genetics (Neuropsychiatric Genetics)* 128B: 30–36.

O'Driscoll GA, Benkelfat C, Florencio PS, et al. (1999) Neural correlates of eye tracking deficits in first-degree relatives of schizophrenic patients: A positron emission tomography study. *Archives of General Psychiatry* 56: 1127–1134.

Ross RG, Olincy A, Mikulich SK, et al. (2002) Admixture analysis of smooth pursuit eye movements in probands with schizophrenia and their relatives suggests gain and leading saccades are potential endophenotypes. *Psychophysiology* 39: 809–819.

Sweeney JA, Luna B, Srinivasagam NM, et al. (1998) Eye tracking abnormalities in schizophrenia: Evidence for dysfunction in the frontal eye fields. *Biological Psychiatry* 44: 698–708.

Thaker GK, Ross DE, Buchanan RW, Adami HM, and Medoff DR (1999) Smooth pursuit eye movements to extraretinal motion signals: Deficits in patients with schizophrenia. *Archives of General Psychiatry* 88: 209–219.

Eyeblink Conditioning

K M Christian, National Institute of Mental Health, Bethesda, MD, USA
R F Thompson, University of Southern California, Los Angeles, CA, USA

Published by Elsevier Ltd.

Classical Conditioning of the Eyeblink Response

Eyeblink conditioning is the most widely studied form of basic associative learning in mammals. In standard delay conditioning, a tone or a light conditioned stimulus (CS) precedes and coterminates with a reflex-eliciting unconditioned stimulus (US), such as periorbital shock or a puff of air to the cornea. Classical conditioning of the eyeblink response is a form of procedural learning and results in the formation of associative memory. Extensive investigation into the neural substrates of this associative memory has resulted in perhaps the most complete description of mammalian memory formation to date.

The Nature of the Eyeblink Unconditioned and Conditioned Responses

Isidore Gormezano showed some years ago in studies of the rabbit that in response to corneal air puff or periorbital shock, the reflex eyeblink response includes eyeball retraction, nictitating membrane (NM) extension, and external eyelid closure, all with essentially identical acquisition functions (NM extension is a largely passive response to eyeball retraction). Simultaneous recording of NM extension and external eyelid closure (electromyogram (EMG) recordings from orbicularis oculi) during acquisition and extinction showed that they were, in essence, perfectly correlated, both within trials and during training. Substantial learning-induced increases in neuronal unit activity that correlate very closely with the conditioned NM extension response have been reported in several motor nuclei: oculomotor, trochlear, motor trigeminal, abducens, accessory abducens, and facial. Coordinated activity in these nuclei and associated muscles contribute to the same global conditioned response (CR), of which the NM extension response is but one component. The suggestion that different motor nuclei might somehow exhibit different CRs in the eyeblink-conditioning paradigm is not supported by evidence.

The CR and the unconditioned response (UR) are similar in eyeblink conditioning in the sense that to a large extent the same muscles and motor nuclei are engaged. However, the CR and the UR differ fundamentally in a number of respects. The minimum onset latency of the CR to a tone CS in well-trained rabbits, measured as NM extension, is about 90–100 ms; the minimum onset latency of the NM extension UR to a 3 psi corneal air puff US in the rabbit is about 25–40 ms. Perhaps most important, the variables that determine the topographies of the UR and CR are quite different. The topography of the UR is under the control of the properties of the US: for example, stimulus intensity, rise time, and duration. In marked contrast, the topography of the CR is substantially independent of the properties of the US and is determined primarily by the interstimulus interval (the CS–US onset interval), with the CR peaking at about the onset of the US over a wide range of effective CS–US onset intervals. This key property of the CR cannot be derived from the properties of the US or the UR. The CR and the UR also differ in that certain components of the UR can be elicited separately by appropriate peripheral stimuli, but the CR always occurs as a global coordinated response. Another important difference is that the CR exhibits much greater plasticity in recovery from lesions of the motor nuclei that impair performance of the UR than does the UR itself. Lesions of the interpositus successful in abolishing the CR have at most only a transient depressive effect on UR amplitude. Numerous parameters of the UR, including amplitude, rise time, frequency, and latency, have been investigated, and there are no significant lasting effects of interpositus lesions in any of these properties.

In sum, the conditioned eyeblink response involves highly coordinated activity in a number of motor nuclei and muscles; it is one global defensive response that is conditioned to a neutral stimulus as a result of associative training. Electrical stimulation of the small critical region of the cerebellar interpositus nucleus (see below) elicits this full complement of coordinated behaviors; it is a 'higher motor program.' The very small lesion of interpositus nucleus which is effective in completely and permanently abolishing the conditioned NM extension response also completely and permanently abolishes all other components of the CR that have been studied – eyeball retraction, external eyelid closure, and orbicularis oculi EMG – without producing any impairment in performance of the reflex response.

Brain Systems Involved in Eyeblink Conditioning

Electrophysiological recordings of neural activity in eyeblink conditioning indicate that, in addition to the motor nuclei described above, a number of brain areas and systems become engaged, most prominently the hippocampus and the cerebellum. In all these

systems, neuronal activity in the CS period increases over the course of learning and precedes the onset of the behavioral CR within trials. Indeed, this increased pattern of action potential discharges in the hippocampus and cerebellum predicts the occurrence and the actual temporal form of the CR (but not the reflex response). In standard delay conditioning, when the CS and the US overlap, lesion studies show that the cerebellar system is critical but the hippocampal system is not. This article focuses on basic delay conditioning and the cerebellum. Within the cerebellar system, the following regions develop the predictive neuronal response: relevant motor nuclei, a region of the fifth nucleus, various reticular brain stem regions, the cerebellar cortex (ansiform and anterior lobes), the cerebellar interpositus nucleus, the pontine nuclei, and the red nucleus.

Cerebellar System: The Pathways

A large literature is in general agreement in identifying the essential circuitry for classical conditioning of eyeblink and other discrete responses (see Figure 1). This circuitry follows closely the well-established anatomy of the cerebellar system and is in general accord with classical theories of cerebellar learning. A brief summary of this work follows.

The UR Pathways

In terms of the reflex pathways, there are direct projections from neurons in regions of the trigeminal nucleus to the accessory abducens (and abducens) nuclei and to the facial nucleus, as well as indirect projections relaying via the brain stem reticular formation, at least to the facial nucleus.

The CR Pathway

Neurons in a localized region of the interpositus nucleus ipsilateral to the trained eye develop a neuronal model of the learned behavioral CR; lesions of this region selectively abolish the CR with no effect in the UR; electrical stimulation of this region evokes the eyeblink response before training (as noted above).

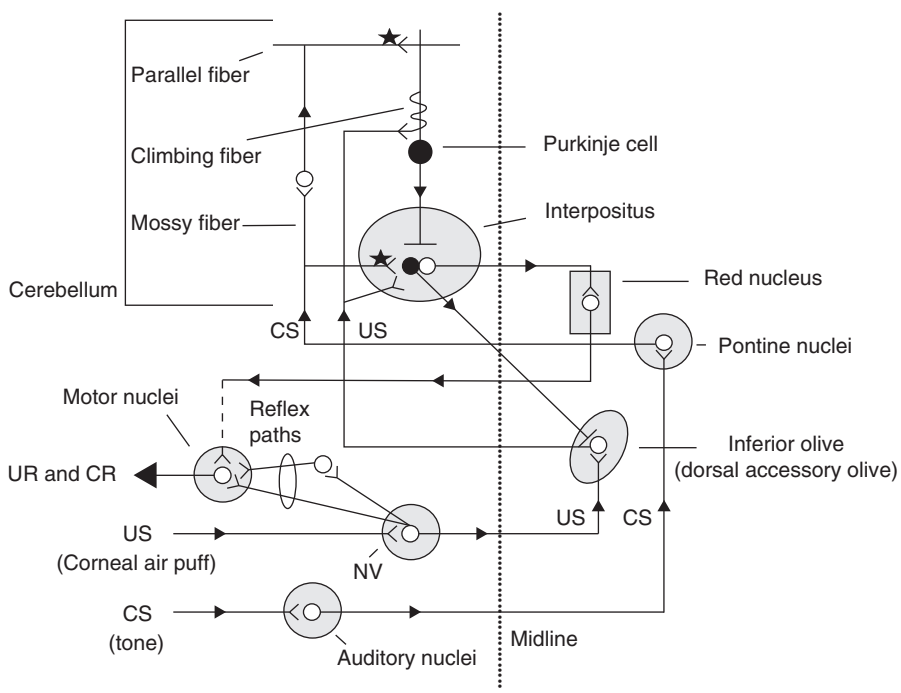

Figure 1 Simplified schematic (most interneurons omitted) of the putative essential circuitry for delay classical conditioning of eyeblink (and other discrete responses) learned with an aversive unconditioned stimulus (US). The reflex US–unconditioned response (UR) pathway involves direct and indirect projections from the trigeminal nucleus to the motor nuclei (for the eyeblink UR and conditioned response (CR) primarily accessory 6 and 7). The tone conditioned stimulus (CS) pathway projects from auditory nuclei to the pontine nuclei and to the cerebellum as mossy fibers. The US pathway includes projections from the trigeminal to the inferior olive and to the cerebellum as climbing fibers. The CR pathway projects from the interpositus to the red nucleus and on to premotor and motor nuclei. There is also a direct γ-aminobutyricacidergic inhibitory projection from the interpositus to the inferior olive. Omitted from this diagram are feedback projections from the cerebellum to several afferent structures. These projections have generally been described physiologically, not anatomically, and may well be involved in some aspect of CR modulation. See article for discussion. Solid cell bodies and bar terminals indicate inhibitory neurons, open cell bodies and fork terminals indicate excitatory neurons. Stars indicate sites of plasticity based on current evidence. NV, trigeminal nucleus. Adapted from Thompson RF (1986) The neurobiology of learning and memory. *Science* 233: 941–947.

This region of the interpositus projects via the superior cerebellar peduncle to a region of the contralateral magnocellular red nucleus. Lesions of the peduncle abolish the CR, as do lesions of the key region of the red nucleus, where neurons also show a model of the learned response. Stimulation of this rubral region also elicits the eyeblink response. The descending rubral pathways project contralaterally to premotor and motor nuclei, seventh for external eyelid closure, accessory sixth and sixth for NM extension. Overwhelming evidence identifies this circuit as the efferent CR pathway for the conditioned eyeblink response. A similar pathway projecting to the spinal cord subserves classical conditioning of the limb flexion response in cat.

The CS Pathway

The pontine nuclei send axons as mossy fibers directly to the cerebellar cortex and interpositus nucleus, mostly contralaterally. The pontine nuclei in turn receive projections from auditory, visual, somatosensory, and association systems, both cortical and subcortical. A wide range of evidence identifies the pontine nuclei–mossy fibers as the essential CS pathway.

The US Pathway

Neurons in the inferior olive (IO) send climbing fiber projections contralaterally directly to cerebellar cortex and interpositus nucleus. The critical region of the IO for eyeblink conditioning is the dorsal accessory olive (DAO), which receives somatosensory and nociceptive input from the spinal cord and appropriate cranial nuclei. Lesions of the critical region of the IO, the face representation in the DAO, completely prevent learning if made before training and result in extinction of the CR if made after training. Neurons in this critical DAO region do not respond to auditory stimuli (CS), respond only to US onset, and show no learning-related activity, and the US-evoked response decreases as animals learn. Electrical microstimulation of this region serves as a very effective US. All these data argue that the DAO–climbing fibers system is the essential US reinforcing pathway for the learning of discrete responses.

The interpositus nucleus sends direct γ-aminobutyricacidergic (GABAergic) projections to the DAO. Hence, as learning-induced increases in interpositus neuron activity develop, inhibition of the DAO neurons will increase. This accounts for the fact that US-evoked activity in the DAO decreases as learning develops, consistent with the classical conditioning model proposed by Robert Rescorla and Allan Wagner. This also appears to serve as a part of the neural circuit essential for the behavioral learning phenomenon of 'blocking,' in which prior training to one CS, such as tone, prevents subsequent learning to a light CS when it is then presented together with the tone in paired compound stimulus training. Infusion of picrotoxin in the DAO to block GABA inhibition from the interpositus during compound stimulus training completely prevents the development of behavioral blocking.

Conjoint Activation of CS and US Pathways

If the above hypotheses concerning the identities of the CS and US pathways are correct, it should be possible to train behavioral CRs by conjoint stimulation of these pathways, which is indeed the case. Stimulation of the pontine nuclei mossy fibers as the CS (below movement threshold) and DAO-climbing fibers as a US results in effective conditioning. Stimulus (US)-elicited movements including eyeblink, head turn, and limb flexion were learned to the mossy fiber stimulation CS just as was the case for peripheral CSs. In an even more reduced preparation, successful conditioning was achieved when parallel fibers were stimulated through a cortical surface electrode as the CS and stimulation of the white matter directly beneath the surface electrode served as the US.

The Cerebellar System: Interpositus Nucleus

Lesions

The initial discovery of the key role of the cerebellum in eyeblink conditioning involved both large aspiration lesions including cerebellar cortex and nuclei and electrolytic lesions of the dentate-interpositus nuclear region. The key nuclear region is the anterior lateral interpositus nucleus ipsilateral to the trained eye. If this region is lesioned before training, learning is completely prevented. These lesions have no effect at all on performance of the UR. Selective abolition of the CR with no effect on the UR following interpositus lesion or inactivation has now been reported in more than 20 studies on several species of mammals. It is important to note that lesion of the interpositus 1 month after training also completely abolishes the eyeblink CR – the interpositus nucleus appears to be the site of long-term storage for the memory.

Kainic acid lesions of the interpositus as small as about 1 mm^3 in the anterior lateral region of the nucleus abolish the CR, indicating extreme localization of the critical region and ruling out the possibility that fibers of passage are involved. Finally, lesions of the output of the cerebellar nuclei, the superior

cerebellar peduncle (scp), abolish the behavioral CR with no effect on the UR. It is important to emphasize that the cerebellar interpositus nucleus is essential for the learning of all discrete movements trained with an aversive US: eyeblink, limb flexion, head turn, and so forth. To the extent tested, the memory traces for all these learned movements are stored (at separate loci) in the interpositus nucleus. Eyeblink conditioning has simply been the most widely studied, although there is also considerable evidence for limb flexion conditioning.

Results of clinical studies of eyeblink conditioning in humans with brain damage are strikingly parallel to the infrahuman animal literature. Appropriate cerebellar lesions markedly impair or completely prevent learning of the standard delay eyeblink-conditioning task; if the cerebellar damage is unilateral, only the ipsilateral eye is affected. In general, effective lesions are large, including damage to cerebellar cortical regions and nuclei, although exact extent of damage is difficult to determine from the brain scans.

Recordings

In a number of studies to date in several species of mammals, recordings of multiple and single neuronal unit activity from the interpositus nucleus during eyeblink conditioning revealed populations of cells in the critical region of the nucleus that, as a result of training, discharged prior to the execution of the learned eyeblink response and fired in a pattern of increased frequency of response that predicted the temporal form of the behavioral CR (the 'neuronal model' of the CR). It is compelling that multiple and single unit activity recorded from the critical interpositus region is appropriately timed to effect the downstream motor pathway and culminate in the well-timed CR. A tight causal chain has been demonstrated from the interpositus to the conditioned eyeblink response (see Figure 2).

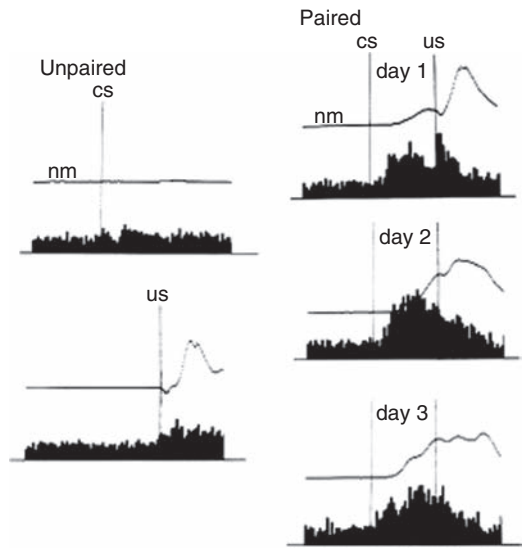

Figure 2 Neuronal model of learned eyeblink response recorded from the cerebellar interpositus nucleus. Each graph shows nictitating membrane (nm) movement (top trace) and a histogram of multiple unit activity (bottom trace). Animals that received explicitly unpaired presentations of the conditioning stimuli (cs) do not develop altered patterns of neuronal activity (left column). Trained animals develop a neuronal model of the learned behavior, as shown by an increase in unit activity prior to the US onset (right column). Total trace duration, 750 ms. us, unconditioned stimulus. From McCormick DA and Thompson RF (1984) Cerebellum: Essential involvement in the classically conditioned eyelid response. *Science* 223(4633): 296–299.

Cerebellar System: Reversible Inactivation

Although the evidence cited above is consistent with the interpositus nucleus as the locus of the conditioned eyeblink memory trace, it does not prove this hypothesis. Predictive neuronal models of the CR develop in the interpositus nucleus, the red nucleus, and the motor nuclei, as noted. Further, lesions of the interpositus, of the scp (containing all the output from the interpositus), and of the red nucleus all abolish the CR with no effect on the UR. Lesions of the motor nuclei of course abolish both the CR and the UR. The technique of reversible inactivation was used to determine where within the circuit the memory trace was formed and stored. Muscimol, a $GABA_A$ agonist, was used to inactivate neuron somas, as was cooling, and tetradotoxin was used to inactivate axons.

Inactivation of the scp, the red nucleus, and the motor nuclei during training all prevent expression of the CR (and the UR with motor nuclei inactivation). However, in postinactivation testing/training, the CR has been learned to asymptote; therefore none of these areas contains the memory trace. However, inactivation of the anterior interpositus nucleus, which also prevents expression of the CR, completely prevents learning. Hence the memory trace is established in the interpositus nucleus. If any portion of the trace were established efferent from the interpositus, learning would be incomplete following inactivation of the scp, but it is not. If any portion of the trace were established afferent to the interpositus, then interpositus inactivation would not completely prevent learning, but it does.

Cerebellar System: Cerebellar Cortex

The morphology of cortical neurons is remarkably distinct, and the cortex itself is structurally homogeneous. It is important to note, however, that the

cortex does not share the discrete somatotopy of the deep nuclei, in which damage to a single cubic millimeter of the anterior lateral interpositus will abolish a specific learned behavior, that is, the conditioned eyeblink. Although the spatial distribution of the mossy fiber and climbing fiber afferents, pathways carrying CS and US information, respectively, is not understood fully, data from anatomical and microstimulation studies suggest a strict topographical relationship between the cortex, deep nuclei, and IO. Although this relationship appears to be preserved among these structures, multiple cortical regions are associated with the conditioned eyeblink circuitry. To date, three regions have received the bulk of attention because of reported anatomical connections with the critical region of the interpositus: Larsell's hemispheric lobule VI (HVI), the anterior lobe, and the C1 and C3 zones – parasagittal regions encompassing portions of both HVI and anterior lobe.

Lesions

In the first study to address the role of the cortex in eyeblink conditioning explicitly, aspiration lesions of ansiform or ansiform and paramedian lobules in well-trained rabbits resulted in the expression of short-latency, small amplitude CRs. The most significant aspect of this early finding was the implication that the CR consists of separable elements related to frequency, magnitude, and timing. While there is currently much support for this hypothesis, there remains considerable disagreement as to the extent of cortical involvement and the anatomical loci that may contribute to the acquisition and maintenance of each of these elements.

Early studies suggested that selective lesions of HVI or HVI and the ansiform lobule completely block expression of the CR in well-trained animals. Subsequent attempts to replicate this result have proven unsuccessful, however. Instead, these latter studies have shown consistently that unilateral lesions of HVI and adjacent regions in well-trained animals result in a transient decrease in conditioned responding but that reacquisition to prelesion levels of CR frequency and amplitude can be achieved with additional training. It is important to note that animals showed significant savings during reacquisition, requiring fewer trials to reach criterion than during initial acquisition, which suggests that the cortical lesions did not eliminate the memory trace. Pretraining lesions of HVI and HVIIa lobules and significant portions of the anterior lobe were shown to significantly impair the rate of learning. However, animals were able to achieve significant levels of conditioned responding in the absence of HVI. Response latencies following targeted disruptions of the HVI region have either not been reported or were not statistically different from prelesion latencies.

Selective modification of CR timing, frequency, and magnitude have been observed following aspiration lesions in which frequency of conditioned responding correlated with lesion size whereas latency and amplitude of the CR were only affected by the most extensive lesions that included the anterior lobe. There are now several reports indicating that the anterior lobe may play a critical role in normal development of well-timed CRs. Although it has been claimed that this region is also necessary for the acquisition and extinction of CRs, acquisition in naïve animals with anterior lobe lesions has yet to be tested directly.

Inactivation

Temporary inactivation of the ventral HVI lobule via infusions of the $GABA_A$ receptor agonist muscimol has no effect on the rate of acquisition or the quality of the CRs. Extensive cortical inactivation encompassing both HVI and portions of the anterior lobe significantly impairs but does not prevent acquisition and has no effect on the CR performance once the animal reaches asymptotic levels of conditioned responding. Similarly, reversible inactivation of the HVI region achieved with a cortical cooling probe significantly decreased the learning rate but did not disrupt CR expression in well-trained animals. If there are multiple and noncontiguous regions within the cerebellar cortex that receive appropriate sensory information to influence the conditioned eyeblink response, then selective lesions and inactivation may not affect all critical cortical regions.

Complete Cortical Elimination

Circumventing these complications, several attempts have been made to investigate the effects of functionally inactivating the entire cerebellar cortex. Mutant mice (pcd mice), in which Purkinje cells degenerate approximately 2 weeks postnatally, show significant deficits in both the rate of acquisition and the asymptotic level of conditioning. Nevertheless, pcd mice express significant levels of learning and extinction in the absence of cortical input to the interpositus. Peak latencies of the CRs in the mutants were significantly decreased compared with wild-type mice. A later study confirmed that the interpositus was responsible for the residual learning by demonstrating that bilateral lesions of the interpositus in the pcd mice completely block acquisition. Recently, it was shown that widespread ablation of Purkinje cells via intraventricular infusions of the immunotoxin

OX7-saporin disrupts retention and expression of previously acquired CRs. Reacquisition of the CR following this treatment occurred at a slower rate than did original acquisition.

Functional and reversible lesions of the entire cerebellar cortex can also be achieved through a targeted disruption of the afferent inhibitory input of Purkinje cells at the level of the interpositus through sequential infusions of muscimol and picrotoxin. Muscimol first blocks the synaptic transmission from the cortex, and the baseline level of excitability in the interpositus neurons is restored through application of the GABA receptor antagonist picrotoxin. With this procedure, onset and peak latencies were decreased in well-trained animals, CR amplitudes were increased, and no effects (i.e., no CRs) were observed in naïve animals.

In contrast to the clear and compelling evidence confirming the essential role of the interpositus in the acquisition and expression of the conditioned eyeblink response, reports of cerebellar cortical involvement have proven more difficult to interpret. The absence of savings following inactivation of the interpositus during acquisition argues against the formation of an interpositus-independent memory trace established in the cerebellar cortex. This result does not, however, eliminate the possibility of interpositus-dependent memory trace formation in the cortex or interpositus-independent cortical modulation of the conditioned response in intact animals.

Recordings

Any direct cortical modulation of the CR must be mediated through the Purkinje cell, which serves as the sole efferent pathway innervating the interpositus. Purkinje neurons are inhibitory and have high spontaneous firing rates (~25–60 Hz), and thus it may be likely that the primary influence on the interpositus results from a well-timed disinhibition. If this hypothesis is correct, then an appropriately timed decrease in Purkinje cell activity should develop over the course of learning. Several studies have shown that Purkinje cells in HVI, the adjacent paramedian lobule, the C3 microzone, and the anterior lobe do exhibit learning-related decreases in firing. However, there is also evidence of learning-related increases in Purkinje cell firing during the CS period. The functional significance of the increase in simple spike activity has not been determined, but it has been suggested that an increased inhibition of a subset of interpositus neurons may suppress non-eyeblink-related movements. Multiple-unit activity recorded from the cerebellar cortex exhibits the same temporal model of the behavioral response observed in other brain structures (interpositus, red nucleus, hippocampus, etc.) and is likewise dependent on the interpositus. It is interesting that if the interpositus is lesioned in trained animals, the timing of learning-related activity in single Purkinje cells is also altered. These results suggest that the interpositus may play a critical role in the timing of the cortical plasticity and that the reciprocal connections between the two structures are involved in the development of the well-timed response.

Mechanisms of Memory Storage in the Interpositus

Several studies report that GABA antagonists prevent both learning and expression of the CR, but blocking the *N*-methyl-D-aspartate receptors in the interpositus prevents learning but not expression of the CR. Hence both GABAergic and glutamate transmitter systems are involved. Protein synthesis inhibitors infused in the interpositus prevent learning but not expression of the CR. Recently, subsets of genes have been identified that are differentially regulated during conditioning – an important step in understanding the molecular basis of the learning-induced changes in the interpositus. All these results argue for the formation of neuronal/synaptic plasticity, a memory trace in the interpositus nucleus.

Extremely important direct evidence for a strengthening of the mossy fiber–interpositus neuron synapses comes from the observation of a highly significant increase in the number of excitatory synapses in the interpositus nucleus (likely from mossy fibers) but no change in inhibitory synapses (from Purkinje neurons) following eyeblink conditioning. This increase appears to be a result of associative learning as this synaptic modification did not occur in unpaired stimulation control animals.

In sum, the evidence is now very strong from behavioral, physiological, pharmacological, anatomical, and reversible inactivation studies that the basic associative memory trace in eyeblink conditioning is established in the interpositus nucleus. The next step is to elucidate the causal chain from behavioral training to increased synaptic efficacy and synapse formation.

Mechanisms of Memory Storage in the Cerebellar Cortex

Although the nature of cortical involvement in classical conditioning is not yet fully understood, it is

clear from lesion, inactivation, and genetic ablation studies that the cerebellar cortex does contribute substantially to the development of the CR under normal conditions. Whereas the interpositus was empirically demonstrated to be essential for the acquisition of the conditioned eyeblink prior to an understanding of the underlying cellular mechanisms, empirical studies of the cortex have largely been designed and interpreted with respect to existing cerebellar learning theory. Specifically, it was predicted that parallel fibers encode contextual information associated with the CS and that the climbing fibers encode US information that serves to signal movement error. The discovery of long-term depression at the parallel fiber–Purkinje cell synapse induced by conjunctive activation of parallel and climbing fibers was interpreted as strong support for these key predictions of cerebellar motor skill learning based on cortical organization. Evidence suggesting a functional role for cortical LTD in eyeblink conditioning has emerged from behavioral assays of mutant mice with deficits in LTD expression. Activation of glutamate receptors on the Purkinje cell dendrites is necessary for the induction of LTD, and mutant mice in which these receptors or downstream signaling pathways are disrupted have shown deficient LTD and impaired eyeblink conditioning. The behavioral impairment seen in these mutants is similar to that observed in pcd mice, in which both the acquisition rate and the asymptotic levels of responding are compromised, with no significant effect on either the timing or the extinction of the CR. Additional evidence for a functional role of cortical LTD in the generation of learned behavior comes from the observation that the temporal and spatial parameters of afferent activity eliciting this plasticity mirror the parameters governing stimulus presentation in eyeblink conditioning. It has been shown that LTD induced by sparse stimulation is most significant if parallel fiber activation precedes climbing fiber activation by 50 to 200 ms, a temporal range close to the optimal interstimulus interval for eyeblink conditioning. Furthermore, synaptic changes in Purkinje cell morphology consistent with the induction of LTD have been observed in trained animals. A significant decrease in α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) binding, consistent with LTD-mediated endocytosis of the AMPA receptors, was reported in the ipsilateral cortex of trained animals compared with controls receiving unpaired presentations of the stimuli. Insofar as cerebellar cortex is concerned, most evidence supports LTD as a mechanism of synaptic plasticity underlying eyeblink conditioning. However, the evidence is indirect, in marked contrast to the more direct evidence of synapse formation in the interpositus nucleus.

Conclusion

It is now established beyond reasonable doubt that the cerebellum and its associated circuitry are the essential (necessary and sufficient) brain system for delay classical conditioning of eyeblink and other discrete responses learned with effective USs. Furthermore, it is clear that the interpositus nucleus is necessary for acquisition and expression of this associative memory. While much work remains to be done to elucidate the molecular and cellular events that contribute to this learned behavior, investigation of the classically conditioned eyeblink response has resulted in perhaps the most striking and complete example of a localized memory trace in the mammalian brain.

See also: Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Cerebellar Microcircuitry; Cerebellar Deep Nuclei; Cerebellum: Models; Conditioned Reflex; Olivocerebellar System; Procedural Learning: Classical Conditioning; Procedural Learning: Cerebellum Models.

Further Reading

Christian KM and Thompson RF (2003) Neural substrates of eyeblink conditioning: Acquisition and retention. *Learning & Memory* 10(6): 427–455.

Christian KM and Thompson RF (2005) Long-term storage of an associative memory trace in the cerebellum. *Behavioral Neuroscience* 119(2): 526–537.

Green JT and Steinmetz JE (2005) Purkinje cell activity in the cerebellar anterior lobe after rabbit eyeblink conditioning. *Learning & Memory* 12(3): 260–269.

Hesslow G and Yeo C (1998) Cerebellum and learning: A complex problem. *Science* 280(5371): 1817–1819.

King DA, Krupa DJ, Foy MR, and Thompson RF (2001) Mechanisms of neuronal conditioning. *International Review of Neurobiology* 45: 313–337.

Kleim JA, Freeman JH Jr., Bruneau R, et al. (2002) Synapse formation is associated with memory storage in the cerebellum. *Proceedings of the National Academy of Sciences of the United States of America* 99(20): 13228–13231.

Krupa DJ, Thompson JK, and Thompson RF (1993) Localization of a memory trace in the mammalian brain. *Science* 260(5110): 989–991.

McCormick DA, Lavond DG, Clark GA, Kettner RE, and Thompson RF (1981) The engram found? Role of the cerebellum in classical conditioning of nictitating membrane and eyelid repsonses. *Bulletin of the Psychonomic Society* 28: 769–775.

McCormick DA and Thompson RF (1984) Cerebellum: Essential involvement in the classically conditioned eyelid response. *Science* 223(4633): 296–299.

Nolan BC and Freeman JH Jr. (2005) Purkinje cell loss by OX7-saporin impairs excitatory and inhibitory eyeblink conditioning. *Behavioral Neuroscience* 119(1): 190–201.

Ohyama T, Nores WL, Murphy M, and Mauk MD (2003) What the cerebellum computes. *Trends in Neurosciences* 26(4): 222–227.

Robleto K, Poulos AM, and Thompson RF (2004) Brain mechanisms of extinction of the classically conditioned eyeblink response. *Learning & Memory* 11(5): 517–524.

Thompson RF (1986) The neurobiology of learning and memory. *Science* 233: 941–947.

Thompson RF (2005) In search of memory traces. *Annual Review of Psychology* 56: 1–23.

Voneida TJ (2000) The effect of brachium conjunctivum transection on a conditioned limb response in the cat. *Behavioural Brain Research* 109(2): 167–175.

Woodruff-Pak DS and Steinmetz JE (eds.) (2000) *Eyeblink Classical Conditioning*, 2 vols. Boston, MA: Kluwer Academic.

Facial Expression in Primate Communication

L A Parr, Yerkes National Primate Research Center, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Faces Are Salient Social Stimuli

Faces are one of the most important classes of stimuli involved in social communication. They provide viewers with a wealth of information, including individual identity, age, sex, and even emotion. Much of the important information from faces is registered quickly by perceivers, without any need for prolonged exposure or scrutiny. A quick glance at a photograph can tell you whether the face is of a man or woman, someone you know or a stranger, as well as the person's general age and race. Important information is also provided by short-term, dynamic changes in expression, head orientation, and gaze that are essential components of ongoing social interactions. For example, studies have shown distinct coding of gaze and emotion from facial expressions in the monkey amygdala using functional magnetic resonance imaging, highlighting the important and dissociable qualities of these two sources of information from the face. Finally, some of the information in faces changes over a much longer timescale and would not be apparent if a person was only encountered once. These life history characteristics include appearance changes due to normal aging and general information about health status.

Faces are also highly memorable. Although general information such as identity, age, and sex are processed without conscious effort, many additional details, such as general affective state, attractiveness, or physical details about a person, can be easily drawn from memory. The phrase 'one never forgets a familiar face' appropriately highlights the salience of faces. Even a brief encounter with an unfamiliar individual might, on the surface, seem quite unremarkable, but when asked later to provide more detail about the person (i.e., if the person seemed in good health and whether he or she was happy or not), these questions could be answered with ease. Therefore, faces carry a rich array of information that not only helps to structure an ongoing interaction but also can provide more details about the encounter at a later time. This article reviews some of the important information that is transmitted from the face. This is divided into two main themes – information transmitted from static faces, such as individual identity, and information transmitted from dynamic faces, such as expression. These sources of information are not mutually exclusive, and although there is a wealth of information about the specialized nature of these processes in humans, this article focuses primarily on the evolution of these skills by drawing on comparative data from nonhuman primates, hereafter referred to as primates.

Static Faces

Individual Identity

Human faces are recognized primarily using the configurational arrangement of facial features. The configuration refers to the spatial arrangement of features on the face (i.e., their gestalt). Two types of configural information are present in faces. First-order relational properties refer to the relative arrangement of facial features which is similar in every face; that is, the eyes are laterally displaced above the nose, which is above the mouth, etc. This type of configural information provides the means for recognizing faces from other categories of visual stimuli, and attraction to these basic configural patters appears to be present from birth. Second-order relational properties refer to the relative size and spatial arrangement of inner facial features with regard to one another; that is, the spacing between the eyes, nose, and mouth, which is more or less unique to every face and enables us to remember specific individuals from facial cues alone. Thus, whereas first-order relational properties enable the identification of faces at a basic categorical level, second-order relational properties provide the information necessary to form subordinate-level classifications of faces (i.e., the ability to distinguish among different individuals such as Mary vs. Jane).

One of the main tests for configural processing is the inversion effect, the finding that faces are difficult to recognize when turned upside down because the configural arrangement of facial features is orientation dependent. The face inversion effect has become one of the main tests for configural processing in both humans and nonhuman primates and provides a simple, but elegant, method for comparing face recognition processes across species. However, studies of the face inversion effect in nonhuman primates have revealed very inconsistent results. There are some reports confirming evidence of the inversion effect in monkeys and apes, including cotton-top tamarins, pigtail macaques, chimpanzees, Japanese

macaques, and rhesus monkeys, whereas many studies have failed to find differences for discriminations between upright and inverted faces in longtail macaques, rhesus monkeys, and cotton-top tamarins. One potential explanation for these differences is that a number of different methodological procedures for testing face recognition skills have been used, ranging from the use of viewing preferences to discriminations of a pair of stimuli based on whether they are the same or different. These different tasks may tap different cognitive and perceptual processes, revealing inconsistent results. Moreover, the stimuli used in these studies are primarily images of conspecifics' faces; however, the quality and presentation of these images differ considerably across studies, including line drawing/schematics of primate faces, a heterogeneous assortment of primate pictures, real photographs versus computer images, and restricted sets of human photos. Therefore, although the inconsistencies are numerous, the methodologies are also quite diverse.

Face Expertise

Much of the research on face recognition in primates has focused on whether some kinds of faces are recognized better than others. For example, studies have demonstrated that monkeys prefer to view pictures of their own species compared to pictures of other species. These findings have been used to support the idea that monkeys are specialized for viewing their own species' faces and preferentially attend to these images because they provide the most salient social information. Among humans, this phenomenon is functionally similar to what is known as the 'other-race' effect and suggests that people develop an expertise for processing faces of their own race, compared to other races, which leads to encoding biases for same-race faces. This preference for own-race faces develops quite early, having been observed in infants as young as 3 month of age. Moreover, it has been suggested that the other-race encoding bias stems from a greater reliance on configural cues when processing same-race faces, as evidenced by greater inversion effects found for same-race versus other-race faces. Moreover, these biases are dependent on the level of expertise with the race of the faces presented. One test of face expertise is the parts-to-wholes test, a measure of configural processing that predicts better recognition of facial features when they are embedded within a face (holistic processing) compared to when they are isolated. Using this test, Caucasian subjects show greater holistic processing for own-race faces compared to Asian faces, but no differences in the parts-to-wholes test were found for Asian subjects who had been living in the United States for some time, suggesting an equal reliance on holistic information for Asian and Caucasian faces.

The issue of expertise and configural processing in face recognition is extremely important for several reasons. First, face expertise enables humans to remember and differentiate between many different individuals over a lifetime. Therefore, individual recognition skills are integrally associated with expertise in face processing. Second, developmentally, face expertise takes many years to acquire – some suggest as long as 10 years – and this expertise has been linked directly to the ability to detect subtle differences in the spacing of facial features, or sensitivity to second-order relational features. Moreover, sensitivity to second-order relational features and individuation of faces activates a region of the fusiform gyrus referred to as the fusiform face area (FFA). The FFA is believed to be a brain region specialized for processing faces, or at the very least it is a brain region that is sensitive to the type of learned subordinate-level classifications that are inherently necessary for human face processing. Interestingly, discriminations involving the second-order relational properties of faces, the spacing between internal features compared to discriminations of face identity, selectively activate a region just adjacent to the FFA. Again, this suggests a causal relationship between face expertise and the ability to detect second-order configural cues.

Studies of Face Expertise in Primates

Several studies have demonstrated greater reliance on configural cues in chimpanzees when discriminating faces of species for which they have considerable expertise (human and conspecific faces) but not the faces of unfamiliar species (capuchin monkeys) or nonface objects such as clip art. Chimpanzees, for example, show the inversion effect for the expert face categories but not the nonexpert or nonface categories. Although this was not specifically a test of the other-race effect, it was the first demonstration of face expertise for processing an 'other-race' category in a nonhuman primate and, moreover, the expertise was dependent on configural processing. Findings on the inversion effect in rhesus monkeys have revealed a quite different picture. Monkeys tested using the same procedures have shown significant inversion effects for discriminations of conspecifics' faces but also unfamiliar face categories, including capuchin monkey faces, and nonface categories, including automobiles; however, no inversion effects were observed for the familiar category of human faces or clip art objects. These findings have been replicated in

subsequent groups of rhesus monkeys in which again there was a failure to find any effect of expertise during tasks requiring monkeys to discriminate upright and inverted faces (i.e., conspecifics' faces, chimpanzee faces, and human faces) compared to houses and clip art objects. Monkeys in this study showed a general inversion effect for all face categories, regardless of their familiarity. This finding is similar to those of other studies that have failed to find evidence for species expertise in Japanese monkeys viewing upright and inverted pictures of rhesus monkeys and conspecifics' faces, despite finding an overall effect of orientation. Therefore, prominent species differences exist between the face processing skills of chimpanzees and rhesus monkeys using an identical methodology and stimulus set, leading to a general conclusion that rhesus monkeys lack the kind of face expertise that is characterized by chimpanzee and human face processing.

Studies on chimpanzees and rhesus monkeys have expanded the question of configural face processing and expertise to examine the role of first- and second-order relational features. Three types of configural manipulations were performed (Figure 1). The inner part of the face was extracted, preserving both first- and second-order relational features. The features were split apart, or fractured, disrupting only second-order relational features. Finally, the faces were fractured and some features were rearranged, disrupting both first- and second-order relational features. The chimpanzees were only significantly impaired when discriminating the fractured and rearranged faces – those that contained manipulations of both first- and second-order relational features. Using the same configural manipulations seen in Figure 1 but applied to conspecific faces, rhesus monkeys were impaired for all of the discriminations, with no preferential sensitivity to second-order relational cues. Therefore, comparative studies of monkeys and apes suggest differences in the use of configural cues when processing the faces of familiar and unfamiliar species, suggesting a lack of face expertise in the monkeys, whereas chimpanzees showed behavioral evidence of face expertise on a level that is comparable to adult humans.

Faces as Dynamic Signals

Ritualization and Signal Form

Almost any behavior is capable of conveying information and thus functioning as a signal. Even if referential communication is not the primary function, an observant animal may derive information from behaviors performed by other individuals; providing this information is relatively easy to identify (perceptual consistency) and corresponds to some predicted outcome (functional consistency). Behavior patterns that convey information in this 'incidental' fashion are referred to as unritualized signals and are contrasted with behaviors that have been specialized, or ritualized, through evolution to convey particular information in specific ways. Ritualization refers to the process by which a formerly adaptive, unspecialized behavior becomes divorced from this original, adaptive function to take on a different meaning. An action of tongue protrusion to repel noxious food from the mouth, for example, may have led to the display of disgust. As a result of this ritualization, communicative signals become highly stereotypical and conspicuous so that they will be easily recognized by receivers and their messages/meanings will have little ambiguity. Another example is the silent pout, a facial expression made by chimpanzees, particularly infants and juveniles. When infant chimpanzees suckle, they pucker their lips around their mother's nipple, forming a pout face. When used outside the context of suckling, the pout is seem most often when an individual is denied access to something, is socially ignored or rejected, during

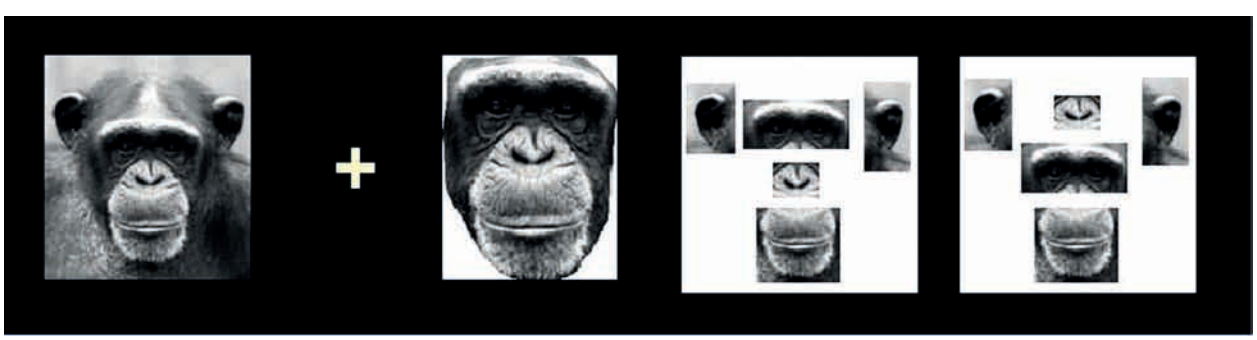

Figure 1 Manipulations of first- and second-order relational features. (Left to right) An unaltered face, inner features, split/fractured features, and fractured/rearranged features. Images were paired in a matching to sample format so that the manipulated image was the sample, the unaltered image was the correct choice, and an unaltered image of a second individual was the nonmatch. Adapted from Parr LA, Heintz M, and Akamagwuna U (2006) Three studies of configural face processing by chimpanzees. *Brain and Cognition* 62: 30–42.

begging, or when startled. Thus, its unritualized function was for feeding, but its ritualized function is to express want or dejection.

The chimpanzee has a set of prototypical, highly ritualized facial expressions, including the bared-teeth display, scream face, pout, stretch-pout whimper, pant-hoot, play face, and bulging lip face (Figure 2). However, with the exception of excellent studies on the structure and function of primate vocalizations, very little is known about how primates discriminate among these important nonverbal signals, and even less is understood about how these signals function within different social contexts, their associated emotional meaning, and, potentially, their altered meaning when used in combination with other signals. A few studies have been undertaken by the author requiring chimpanzees to discriminate pairs of photographs showing different individuals making the same facial expression, from either a neutral face or photographs of other facial expressions. In these studies, chimpanzees perform very well, categorizing the most prototypical example of their facial expressions. Therefore, specific features unique to particular expression categories, such as the upper and lower teeth showing in the bared-teeth display, or unique combinations/configurations of features were likely used by subjects when performing these categorization tasks. There is considerable debate in the human literature as to how facial expressions are processed. Some studies support the role of configural cues in expression categorization, where similar types of tasks have been used as those described for face identity, including the inversion effect. However, other studies have clearly shown that altering a single feature, such as the wrinkling around the eye corners, can change the interpretation of a facial expression such as smiling. Smiling faces that did not contain the crow's feet eye wrinkles were judged less happy than faces that did contain the wrinkles. Thus, although configural cues are important for facial expression categorization in humans, altering individual component features can have a significant impact on the emotional judgment of the entire configuration.

The Social Meaning of Graded and Blended Signals

The facial expressions of nonhuman primates, particularly the great apes, are best described as graded signals because rarely are they observed in their peak-intensity prototypical form, as shown in Figure 2. Similar issues are known for human facial expressions, and it should be noted that most of the work on the meaning of human facial expressions has been done in a

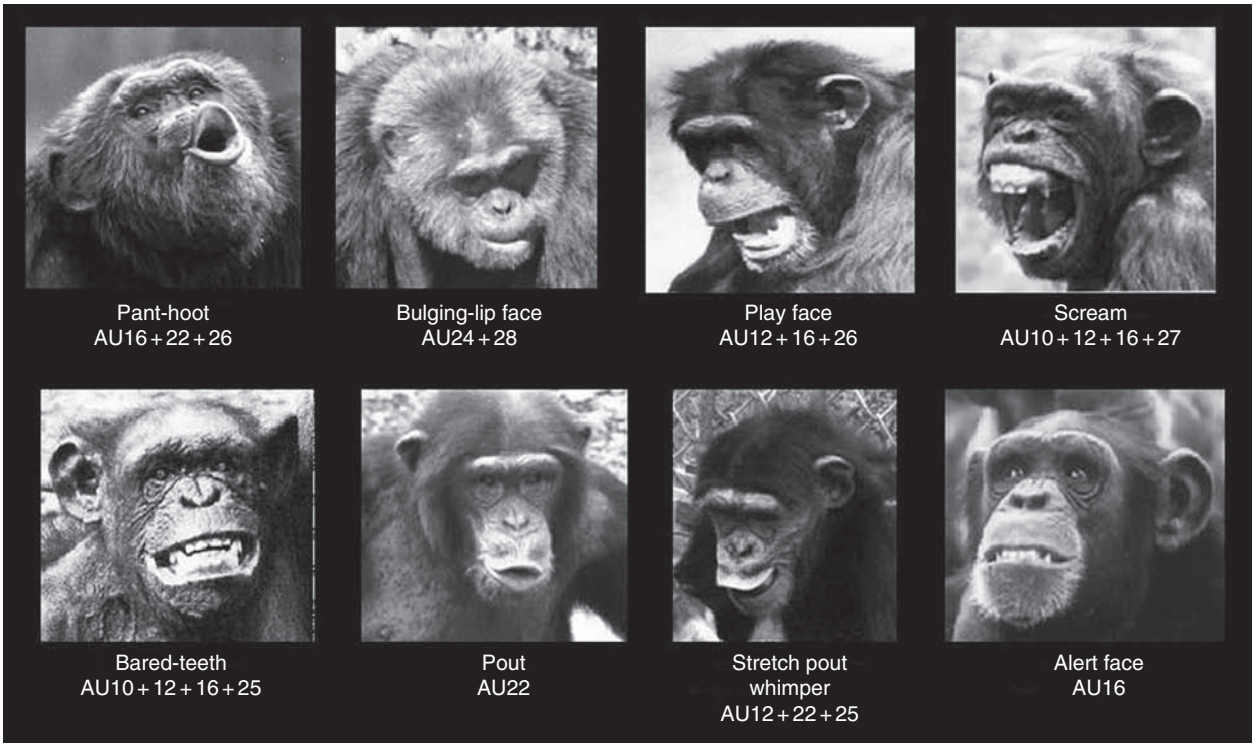

Figure 2 Prototypical chimpanzee facial expressions, verified according to component movements using ChimpFACS. The unique action unit configurations are listed beneath the expression names. These categories are abbreviated from a more complete ethogram. Reproduced from Vick SJ, Waller BM, Parr LA, Pasqualini-Smith M, and Bard KA (2007) A cross species comparison of facial morphology and movement in humans and chimpanzees using FACS. *Journal of Nonverbal Behavior* 31: 1–20.

laboratory setting using either highly standardized, posed expressions or facial expressions represented only as peak-intensity displays. Their graded nature makes it extremely difficult to characterize facial expressions during ongoing social interaction in both chimpanzees and humans and challenges our ability to understand their form, preferred context of use, and social meaning or function. Moreover, there are two dimensions on which signals may be graded: They may vary quantitatively in terms of intensity level and/or qualitatively by forming intermediates with other signals. In an effort to understand the relationship between chimpanzee facial expressions and social behavior, observational studies have recorded in what social contexts specific expressions are made. Prototypical facial expressions were associated most often with a single social context; for example, the relaxed open-mouth face (or play face) was observed to occur almost exclusively within the social context of play. However, there was also considerable overlap in the contexts that gave rise to many of the other facial expressions present in the chimpanzee repertoire. Some facial expressions, for example, appear to share features in common with two or more prototypical categories; thus, they seem to be blended between two parent groups. Figure 3 illustrates one example, the stretch-pout whimper, which is an expression blend because it shares prominent features in common with both the bared-teeth display and the pout, and it was as likely to be expressed in the contexts associated with both of the parent expressions. Thus, it was both a morphological/structural blend and a functional blend. In most cases, however, the expression blends were only blended in visual appearance, not in terms of their context of use, because they were used in a manner similar to only one of the parent expressions that they were blended between.

Some variation in facial expression can also occur as a result of the accompanying vocalization. Although not all facial expressions have an accompanying vocalization, it is obvious that all vocalizations will have some changes in facial appearance. Whether these changes provide information in the visual domain, as opposed to being exclusively auditory signals, is not clear. A multimodal information-processing experiment was conducted in which videos of various facial expressions were presented with either their congruent vocalization (i.e., scream faces with scream vocalizations) or an incongruent vocalization (i.e., a scream face paired with a vocalization of another facial expression category, such as a pant-hoot). The chimpanzees were presented with the choice of matching the incongruent videos by selecting a facial expression photograph that depicted either the auditory cue (i.e., the pant-hoot) or the visual cue (i.e., the scream face). Results showed clear preferences for auditory modality for some expressions (pant-hoots and play faces (laughter)) but the visual modality for others (i.e., the scream). Other authors have shown that rhesus monkeys will orient to the visual display of one of two vocalizations, either a coo or a threat call, when they hear that corresponding vocalization, suggesting an important convergence between the auditory and visual components of multimodal facial signals.

Thus, the communication system among some primates, such as humans, appears to be highly varied in terms of signal appearance (graded and blended facial expressions), the ability to integrate multimodal cues, and in the varied use of facial expressions across a range of social contexts. This presents unique cognitive challenges for primates, particularly chimpanzees, that must know/learn how to use these varied signals as well as interpret their meaning in multiple contexts. The rules for this are unclear, but they may be simple in that specific facial expressions garner different meaning depending on their context of use, as has been suggested in some monkey species, although it was not studied how subtle differences in the form and/or temporal patterning of these signals may also have contributed to their differential interpretation. This type of redundancy may minimize the need to develop a completely new signal to express the same message in different contexts, but it raises the awareness that social context plays

Figure 3 Graded and blended facial expressions in chimpanzees. The leftmost photograph shows a silent pout and the rightmost photograph shows a silent bared-teeth display. The middle photograph is a blended expression, the stretch-pout whimper, and the remaining photographs are graded variations between these phototypes. Reproduced from Parr LA, Cohen M, and de Waal FBM (2005) The influence of social context on the use of blended and graded facial displays in chimpanzees (*Pan troglodytes*). *International Journal of Primatology* 26: 73–103.

as important a role in signal meaning as signal type. This is also quite similar to how human facial expressions are interpreted, since both context and the identity of the performer help to interpret the meaning of expression.

Alternatively, the rules may be quite complex because the range of potential meanings that are possible when different forms of a signal are used in different social contexts, and perhaps even by different age/sex class individuals, grows exponentially large. In this latter case, the meaning conveyed by a specific signal may be a simple heuristic, as opposed to having a specific/referential meaning. Consider the bared-teeth face. This expression in the chimpanzee has been shown to be structurally homologous to the smile in humans: These expressions look very similar and probably share a similar evolutionary history. The bared-teeth expression is also used by chimpanzees in a wide range of social contexts and by all individuals in a social group. In contrast, this signal is highly ritualized in macaques in both form and context, being expressed unidirectionally from subordinate to dominant. Among rhesus monkeys, this has led to its label as a formal signal of dominance because it can be used to reliably predict dominance within a social group. It has also received a functional label drawing on its proposed emotional meaning, referred to by some as the 'fear grin' or 'grimace.' However, the relationship between this signal and the emotion 'fear' is unclear, and often this interpretation does not accurately describe the situation because the bared teeth may be used in unthreatening or nonconflict situations in which there is little risk to the sender. Among chimpanzees, the bared-teeth display appears to mean something much more general, and it is used across a wide variety of social contexts, including during play, with no predicted directionality between sender and receiver. It has been proposed that the bared-teeth display functions to indicate a general state of benign intent (i.e., "I mean you no harm"), which enables individuals to reduce proximity and potentially socially interact in a wide variety of situations. Its use in these situations can result in more grooming between individuals and improved quality of their social relationship. Further studies are needed before researchers will more fully understand how these facial expressions function during chimpanzee social interactions. The final section describes a new comparative research tool, ChimpFACS, which was developed from the human Facial Action Coding System (FACS) to aid in identifying homologous facial movements between chimpanzees and humans and to aid in the classification of graded signals.

Deconstructing Facial Expressions into Component Movements

ChimpFACS is an anatomically based coding system for measuring and describing facial movement according to the action of the underlying facial muscles. It was developed, in part, to provide a standardized (i.e., certification is required) observational coding scheme for describing facial expressions in chimpanzees that would be comparable across users and also to make possible direct comparisons between the facial expressions of chimpanzees and humans by identifying the homologous muscular bases of facial movement. Similar to the human FACS, each component facial movement is based on the action of an individual muscle or, in some cases, a combination of muscles. Therefore, prior to the development of ChimpFACS, detailed dissections of chimpanzee facial muscles were undertaken from cadavers and all of the 23 mimetic facial muscles described in humans were confirmed in the chimpanzee. These component movements were then given numerical codes called action units (AUs), similar to the human FACS. ChimpFACS identifies the muscular basis, appearance changes, and minimum coding criteria necessary for describing a total of 43 AUs (Table 1).

With this tool in hand, several initial studies were undertaken to determine whether this bottom-up approach could validate the previous classification of chimpanzee facial expressions (i.e., those shown in Figure 2, among others) or to provide any new information about the categorization of these highly graded signals through the identification of common underlying movements. Using a discriminant functions analysis (DFA), we validated most of the existing categories of chimpanzee facial expressions by reporting high agreement between the AU configurations coded by ChimpFACS experts and the *a priori* classifications given to facial expressions by me based on experience. Therefore, ChimpFACS was accurately and objectively able to classify prototypical categories of facial expressions according to their component movements. The DFA also enabled researchers to identify the most prototypical combination, or configuration, of facial movements for each expression. The expressions shown in Figure 2 illustrate these prototypes with their unique AU configurations. These configurations will be very useful in future studies on expression categorization because they can confirm the categories used and provide a measure to standardize the best examples. Moreover, this approach will enable more detailed studies to examine whether configuration

Table 1 Action units (AUs) describing facial movement in chimpanzees

Upper-face chimp AUs		*Lower-face chimp AUs*	
1+2	Brow raise	9	Nose wrinkler
5	Upper lid raise	10	Upper lip raiser
6	Cheek raise	12	Lip corner puller
43	Eye closure	15	Lip corner depressor
45	Blink	16	Lower lip depress
46	Wink	17	Chin raiser
Head positions		22	Lip funneler
51	Turn left	24	Lip pressor
52	Turn right	28	Lips suck
53	Head up	Miscellaneous AUs	
54	Head down	19	Tongue show
55	Tilt left	21	Neck tightener
56	Tilt right	29	Jaw thrust
57	Forward	30	Jaw sideways
58	Back	31	Jaw clencher
Eye positions		32	Bite
61	Eyes left	33	Blow
62	Eyes right	34	Puff
63	Eyes up	35	Cheek suck
64	Eyes down	36	Tongue bulge
65	Walleye	37	Lip wipe
66	Cross eye		
Lip parting and jaw opening			
25	Lips part		
26	Jaw drop		
27	Mouth stretch		

Adapted from Vick SJ, Waller BM, Parr LA, et al. (2007) A cross species comparison of facial morphology and movement in humans and chimpanzees using FACS. *Journal of Nonverbal Behavior* 31: 1–20.

or component features are most important for chimpanzee expression categorization and compare the result to recent findings in humans.

Conclusion

The primate face provides a rich source of information pertaining to individual characteristics and social communication. Static information in faces, such as facial configuration, provides the details necessary for individual recognition, but the face can also convey information regarding an individual's age, sex, health status, attractiveness, and race. Humans and chimpanzees have developed an expertise for processing faces of familiar species, particularly own-race faces, which enables the individuation of faces using unique configural features. The current data do not support such an expertise in rhesus monkeys. The face also provides a rich source of socioemotional information via dynamic movement, or facial expression. Although researchers are in the early stages of understanding the comparative nature of expressive communication across primates, comparative tools such as the ChimpFACS will be invaluable in understanding the homologous basis of expressive communication.

See also: Primate Communication: Evolution; Referentiality and Concepts in Animal Cognition; Signal Transmission in Natural Environments; Signal Design Rules in Animal Communication; Social Interaction; Visual Signaling in Animals.

Further Reading

Diamond R and Carey S (1986) Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology* 115: 107–117.

Ekman P and Friesen WV (1978) *Facial Action Coding System*. Mountain View, CA: Consulting Psychologists Press.

Gothard KM, Battaglai FP, Erickson CA, Spitler KM, and Amaral DG (2007) Neural responses to facial expression and face identity in the monkey amygdala. *Journal of Neurophysiology* 7: 1671–1683.

Hoffman KL, Gothard KM, Scmid MC, and Logothetis NK (2007) Facial expression and gaze-selective responses in the monkey amygdala. *Current Biology* 17: 766–772.

Marler P and Tenaza R (1976) Signaling behavior of apes with special reference to vocalization. In: Sebeok T (ed.) *How Animals Communicate*, pp. 965–1033. Bloomington, IN: Indiana University Press.

Maurer D, Le Grand R, and Mondloch CJ (2002) The many faces of configural processing. *Trends in Cognitive Science* 6: 255–260.

McKone E and Kanwisher N (2005) Does the human brain process objects of expertise like faces? A review of the evidence. In: Dehaenne S, Duhamel JR, Hauser M, and Rizzolatti G (eds.) *From Monkey Brain to Human Brain*, pp. 339–356. Cambridge, MA: MIT Press.

Mondloch CJ, Maurer D, and Ahola S (2006) Becoming a face expert. *Psychological Science* 17: 930–934.

Parr LA, Cohen M, and de Waal FBM (2005) The influence of social context on the use of blended and graded facial displays in chimpanzees (*Pan troglodytes*). *International Journal of Primatology* 26: 73–103.

Parr LA, Heintz M, and Akamagwuna U (2006) Three studies of configural face processing by chimpanzees. *Brain and Cognition* 62: 30–42.

Preuschoft S and van Hooff JARAM (1995) Homologizing primate facial displays: A critical review of methods. *Folia Primatologica* 65: 121–137.

Rhodes G, Hayward WG, and Winkler C (2006) Expert face coding: Configural and component coding of own-race and other-race faces. *Psychonomic Bulletin and Review* 13: 499–505.

Seyfarth RM and Cheney DL (2002) Signallers and receivers in animal communication. *Annual Review of Psychology* 54: 145–173.

van Hooff JARAM (1973) A structural analysis of the social behavior of a semi-captive groups of chimpanzees. In: von Cranach M and Vine I (eds.) *Social Communication and Movement*, pp. 75–162. Academic Press: London.

Vick SJ, Waller BM, Parr LA, Pasqualini-Smith M, and Bard KA (2007) A cross species comparison of facial morphology and

movement in humans and chimpanzees using FACS. *Journal of Nonverbal Behavior* 31: 1–20.

Waller B and Dunbar RIM (2005) Differential behavioural effects of silent bared teeth display and relaxed open mouth display in chimpanzees (*Pan troglodytes*). *Ethology* 111: 129–142.

Relevant Website

http://www.chimpfacs.com – ChimpFACS.

Fatigue

R M Enoka, University of Colorado, Boulder, CO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In common usage, the term fatigue is used to denote a variety of conditions that range from an exercise-induced impairment of motor performance through to the sensations of tiredness and weakness that accompany some neurological disorders. Although convenient for casual conversation, such broad usage precludes the systematic study of the underlying mechanisms. Accordingly, many investigators who study the physiology of fatigue prefer a more restricted definition that considers fatigue to be an exercise-induced reduction in the ability of muscle to produce force or power. With this definition, the amount of fatigue caused by a bout of exercise can be quantified as either the reduction in the peak force achieved during a maximal voluntary contraction or the peak power produced during a single contraction or a series of contractions.

Sites of Impairment

Since the seminal work of Mosso, the typical strategy used by experimentalists in the study of fatigue has been to determine whether the mechanism responsible for fatigue was located in the exercising muscle or in the nervous system. The strategy has been to distinguish between an impairment of contractile protein function and a decrease in the magnitude of the activation signal. This work has demonstrated that the function of the contractile proteins and the capacity of the myofibrils to produce force can be impaired by a decline in the efficacy of excitation–contraction coupling, metabolic changes that influence the supply of adenosine triphosphate and impede the function of the cross-bridges, and an inadequate perfusion of the active muscles. In contrast, activation of the contractile proteins can be reduced by a decline in neuromuscular propagation and the inward spread of action potentials down the transverse tubules, a reduction in the motor output from the spinal cord, a decline in the afferent support of motor neuron discharge, an increase in afferent feedback that reduces motor neuron excitability, a decrease in the activation of the motor unit population by descending inputs, and an insufficient drive to the neurons in the motor cortex (Figure 1).

The experimental models that have been used to identify the sites of impairment range from *in vitro* measurements on isolated muscle fibers through the evaluation of a motor performance by a human participant. Each model has its strengths and weaknesses. To identify the metabolic factors that impair cross-bridge function, for example, it is necessary to use a reduced preparation that permits the assessment of contractile protein function. To examine the changes that occur in the descending drive to the motor neurons during a voluntary contraction, however, it is necessary to measure evoked responses in humans as they perform a prescribed task.

A common technique used to distinguish impairments of contractile function and the activation signal in humans is to superimpose electrical stimulation on maximal voluntary contractions performed during a fatigue protocol. This method is known as the twitch interpolation technique and typically involves superimposing from two to ten stimuli at a high frequency to determine whether additional force can be evoked from a test muscle when an individual attempts to activate it maximally. Any increment in force is normalized to the force that can be elicited by the same stimulus when the muscle is at rest, and the ratio is expressed as an index of voluntary activation. A number of studies have found that voluntary activation is often less than 100% and that the deficit can increase during a fatiguing contraction.

One approach to identify the locus of the deficit in voluntary activation is to evoke muscle contractions with stimuli delivered at various locations along the pathway from the motor cortex to the muscle. For example, studies on the fatigue experienced by the elbow flexor muscles can involve a comparison of responses evoked by transcranial stimulation of the motor cortex, stimulation of the corticospinal tract at the level of the cervicomedullary junction (transmastoid processes), stimulation of the brachial plexus at the supraclavicular fossa, and stimulation of the intramuscular branches of the motor nerve (Figure 2). Stimulation of the motor cortex, corticospinal tract, and brachial plexus evoke both electromyographic and force responses that can be measured, whereas stimulation applied over the muscle elicits only a force response. Because transcranial magnetic stimulation generates action potentials in the axons of cortical neurons that impinge on the output neurons of the motor cortex, a difference in the responses evoked by cortical and corticospinal stimulation provides information on changes in the motor cortical neurons. Similarly, a difference in the responses between the corticospinal and brachial plexus stimulation informs the investigator about changes in the excitability of the motor neurons. These comparisons have indicated that

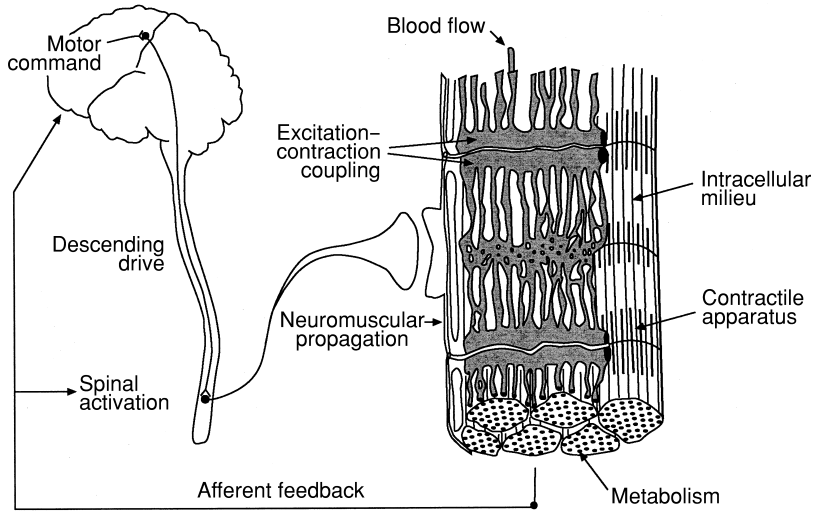

Figure 1 The decline in force or power that occurs during a fatiguing contraction can be caused by an impairment of one or many of the physiological processes associated with transforming the motor command into a contractile force. From Enoka RM (2008) *Neuromechanics of Human Movement.* Champaign, IL: Human Kinetics.

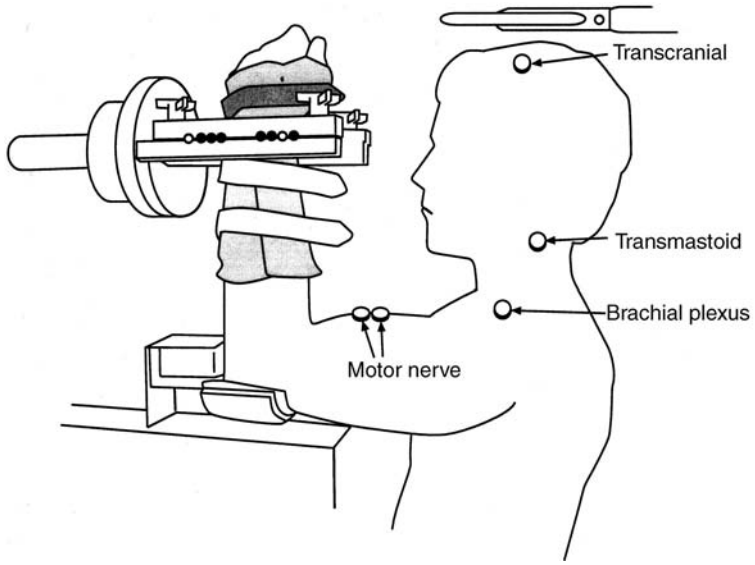

Figure 2 Changes that occur in the pathway from the motor cortex to the muscle fibers of the elbow flexor muscles can be examined by comparing responses evoked by stimulation of the motor cortex (transcranial stimulation), the corticospinal tract (transmastoid stimulation), the peripheral nerve (brachial plexus), and the intramuscular branches of the motor axons.

group III and IV afferents sensitive to ischemia do not mediate depression of the responses to motor cortical or corticospinal stimulation after fatiguing exercise and that the depression of voluntary activation when exercise is performed in hyperthermic conditions is not related to a decrease in the response evoked by motor cortical stimulation.

Another approach to examining the changes that occur in the pathways that provide activation of the muscle fibers is to measure the adjustments in reflex responses, such as changes in the electromyogram evoked by a stimulus that activates afferent fibers: for example, the measurement of changes in short- and long-latency reflexes evoked by applying either a brief electrical stimulation to the group Ia fibers that arise from muscle spindles or a quick stretch of the muscle to activate the muscle spindle afferents. The short-latency component largely comprises monosynaptic input to the spinal motor neurons, whereas the long-latency component involves transmission of the elicited activity traveling to supraspinal centers before being received by the motor

neurons. Comparison of short- and long-latency components can distinguish the involvement of spinal and supraspinal pathways in the fatiguing contraction, particularly in the upper limbs. Such studies have indicated that adjustments during fatiguing contractions typically involve a change in the balance of excitation and inhibition delivered from spinal and supraspinal sources to the motor neuron pool.

A few studies that have examined changes in the responses evoked by transcranial magnetic stimulation have concluded that some of the deficit in voluntary activation can be attributed to reduction in the net excitation of the neurons that contribute to the descending pathways. Although the origin of this limitation is uncertain, some exercise conditions have adverse effects on cerebral function. For example, the hypoglycemia that is caused by a failure of hepatic glucose production to satisfy the increased demands for glucose during exercise can impair the ability of the brain to provide an adequate descending drive to the spinal motor neurons during a fatiguing contraction. Nybo found that when participants sustained a maximal contraction for 2 min, the decline in force exerted by the knee extensor muscles was greater when the task was preceded by 3 h of exercise without glucose supplementation (placebo) than when the same task was performed with glucose supplementation. Furthermore, the greater rate of force decline for the placebo condition was accompanied by a greater impairment of voluntary activation (Figure 3). These findings, along with others, suggest that the fatigue associated with hypoglycemia is largely related to the decrease in energy turnover in the brain. This effect, which is similar to one produced by hyperthermia, depends on the intensity and mode of exercise.

Although a few pioneers in this field did appreciate that fatigue could result from the failure to sustain an adequate activation, the classical approach to the study of muscle fatigue has been to associate the decline in force capacity with the intramuscular accumulation of various substances (e.g., lactate, phosphate, and ammonia), a disturbance of the Na^+–K^+ balance, changes in intracellular pH, an accumulation of free radical species, alterations in intracellular Ca^{2+} handling or sensitivity, or impaired energy metabolism due to the depletion of substrates such as glucose and the intermediates for the tricarboxylic acid cycle. For example, the accumulation of inorganic phosphate due to the breakdown of creatine phosphate has been shown to evoke at least two major effects that could contribute to fatigue: (1) reduction of the maximal cross-bridge force, either by impeding the transition of the cross-bridge from the low-force state to the high-force state or by decreasing the sensitivity of the myofibrils to calcium; and (2) interference with the handling of calcium by the sarcoplasmic reticulum, such as by increasing the probability of calcium channels' opening, inhibiting the uptake of calcium, and augmenting the precipitation of inorganic phosphate and calcium within the sarcoplasmic reticulum. Similarly, disturbances in the Na^+ and K^+ gradients across the muscle fiber membrane due to an inability of Na^+–K^+pumps to restore ionic gradients during intense exercise can depress the excitability of the membrane and contribute to a decline in muscle fiber force. These studies, which are performed on skinned muscle fibers, have also shown that acidosis has a negligible direct effect on muscle function at physiological temperatures.

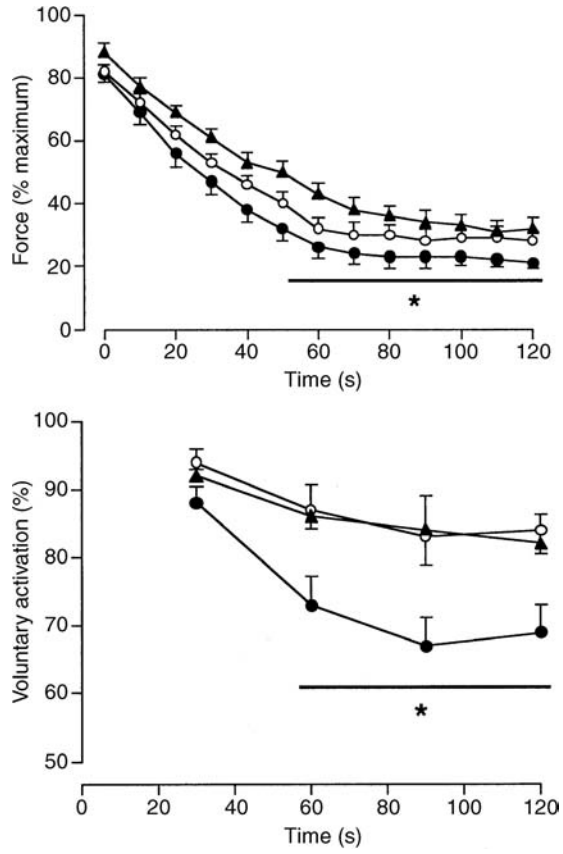

Figure 3 The force exerted by the knee extensor muscles (top) and the associated levels of voluntary activation (bottom) during maximal contractions that were sustained for 120 s. Three conditions were examined: baseline (filled triangles), placebo trial (filled circles), and glucose trial (open circles). The baseline condition represents the decreases in force and voluntary activation when the task was performed after a 15 min warm-up. The placebo and glucose trials were performed after 3 h of cycling, either with (glucose trial) or without (placebo trial) glucose supplementation. The force data (top) are reported as a percentage of the peak values recorded during a maximal voluntary contraction. $^*p < 0.05$ compared with the baseline and glucose trial. Mean ± SE. Data from Nybo L (2003) CNS fatigue and prolonged exercise: Effect of glucose supplementation. *Medicine & Science in Sports & Exercise* 35: 589–594.

In intact muscles, the release of calcium from the sarcoplasmic reticulum in response to the propagation of action potentials down the transverse tubules can be impaired during fatiguing contractions. The effect of the impairment is a reduction in the quantity of calcium released in response to a single action potential. As a consequence, the intracellular calcium concentration required to achieve maximal activation of the contractile proteins can be achieved only with high rates of activation. Because this deficit in excitation–contraction coupling is apparent only at low frequencies of activation, it is referred to as low-frequency fatigue. Low-frequency fatigue is more common after voluntary contractions sustained at low forces for long duration or contraction evoked by low frequencies of stimulation. Recovery from low-frequency fatigue can take up to 24 h. Because the force depression is exhibited by stimulation rates that encompass the typical discharge rates of motor units, it is presumed that low-frequency fatigue represents a functionally significant phenomenon.

Despite the observations derived from experiments on single muscle fibers that the decline in force capacity can be caused by an accumulation of some metabolites, other studies have found that such effects can occur even in the absence of an increase in the intracellular concentrations of these metabolites. For example, some of the properties of single fibers that were chemically skinned after performing a fatigue protocol differed from those of resting fibers. Because the process of chemical skinning replaces the sarcoplasm of the fiber with control levels of the intracellular milieu, the difference in myofibrillar sensitivity to calcium and rate of calcium uptake were attributed to changes intrinsic to the myofibrils and sarcoplasmic reticulum.

Taken together, the single-fiber experiments provide evidence that changes within the muscle fiber can contribute to the decline in force associated with fatigue. However, the relative significance of these mechanisms *in vivo* remains uncertain. Nonetheless, it does seem likely the force produced by single muscle fibers does decrease during a fatiguing contraction. The increase in motor unit recruitment during a sustained submaximal contraction, for example, presumably represents an attempt to maintain the desired force in the presence of a declining contribution from the previously active muscle fibers. Despite the many candidate mechanisms that could cause the decline in muscle fiber force, there is no consensus on the primary mechanism.

Task Dependency

One principal conclusion from a century of work on the sites of impairment is that fatigue can be caused by a decrease in the efficacy of multiple physiological processes. It is important to note that the dominant mechanism depends on the characteristics of the task being performed. The task details that influence the contributing mechanisms include the type and intensity of exercise, the muscle groups involved in the activity, and the physical environment in which the task is performed. These findings have led to a principle known as the task dependency of muscle fatigue: There is no single cause of muscle fatigue, and the dominant mechanism depends on the details of the task being performed. Furthermore, some of these observations on task-dependent effects also indicate that the decline in force during a fatiguing contraction often involves multiple mechanisms.

Variation in the dominant mechanism across tasks has confused the literature on muscle fatigue. For example, compare the results of two studies that attempted to determine if old adults are more fatigable than young adults. Baudry and colleagues had young and old adults perform a series of maximal shortening and lengthening contractions with the dorsiflexor muscles. The decline in peak torque during both types of contractions was greater for the old adults, which indicated that they were more fatigable than the young adults. In contrast, Hunter and colleagues found that old men could sustain an isometric contraction with the elbow flexor muscles at a force that was 20% of maximum for a longer duration than young men could. At the conclusion of the submaximal contraction, the two groups of men exhibited a similar amount of fatigue, as indicated by equivalent declines in the maximal force that they could exert. Hence, the old men were less fatigable when they performed this task. This comparison indicates that the influence of age on muscle fatigue depends on the task that is being performed.

Task Failure

Given the task dependency of muscle fatigue, an alternative to searching for an omnipresent fatigue mechanism is to identify the mechanisms that limit the ability of individuals to perform specific tasks. Rather than simply quantifying the amount of fatigue as indicated by the decline in peak force or power, however, recent work has focused on the mechanisms that limit the duration the task can be sustained. Such an approach emphasizes the functional significance of the changes that occur during fatiguing contractions. The relative role of the adjustments that occur during fatiguing contractions can be assessed by identifying those changes that are associated with differences in the time to failure of two performances. Comparisons can include, for example, one group of participants

performing two similar tasks, two groups of participants sustaining the same task, and one group of participants performing the same task before and after an intervention.

One approach has been to compare how varying the type of load influences the duration that a submaximal contraction can be sustained. For example, one could compare two tasks that both involved sustained isometric contractions and required the same submaximal net muscle torque and joint angle yet differed in the type of feedback provided to the participant and the type of load supported by the muscles. In one such task, referred to as the force task, the limb was attached to a restraint, and the participant was required to maintain a constant force for as long as possible while viewing force feedback on a monitor. In the other task, referred to as the position task, the participant supported an inertial load that was equivalent to the force exerted during the force task and was required to maintain a constant joint angle for as long as possible while viewing position feedback on a monitor. The time to failure of the position task was briefer when the two tasks were performed by a hand muscle in the position shown in Figure 4(a). The difference in the time to failure was caused by a more rapid recruitment of the motor unit pool during the position task. Because the net muscle torque exerted by each participant was similar for the two tasks, the load on the muscle fibers did not differ, at least initially. As the fatiguing contractions progressed, however, the more rapid recruitment of motor units during the position task should have reduced the load experienced by the active muscle fibers. Consequently, the more rapid recruitment of motor units during the position task was not caused by a more substantial decrease in the force produced by the active muscle fibers, which suggests that the difference in recruitment rate was a consequence of the control strategy used by the central nervous system. To underscore this interpretation, there was no difference in the time to failure or the rate of motor unit recruitment for the force and position tasks when the target force exceeded the upper limit of motor unit recruitment for the muscle.

When the force and position tasks are performed with the elbow flexor muscles in the position shown in Figure 4(b), the difference in the time to task failure is greater than that for the hand muscle. This suggests that another factor besides the control strategy influences the duration that the position task can be sustained by the elbow flexor muscles. When the two tasks were performed with the elbow flexor muscles in the position shown in Figure 4(c), the difference in the time to failure between the two tasks was much less

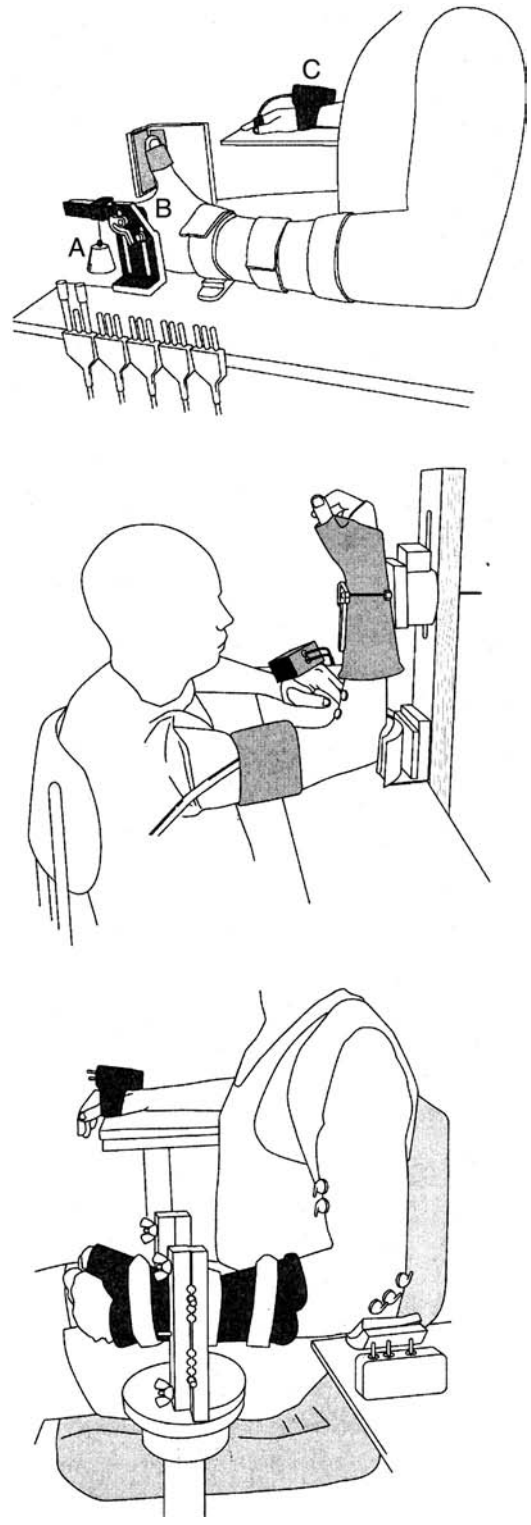

Figure 4 Selected models that have been used to compare the time to failure of the force and position tasks. Top, the position task with the first dorsal interosseus muscle. Middle, the force task with the elbow flexor muscles and with the upper arm vertical and the forearm in a neutral position. Bottom, the force task with the elbow flexor muscles and with the upper arm horizontal and the forearm supinated. From Maluf KS and Enoka RM (2005) Task failure during fatiguing contractions performed by humans. *Journal of Applied Physiology* 99: 389–396.

than when the forearm was horizontal (Figure 4(b)). Hence, some factor related to the posture of the arm also contributed to the briefer time to failure of the position task when the forearm was horizontal. The additional factor was the demand on the muscles around the shoulder joint (rotator cuff and deltoid muscles) to maintain the position of the arm. The rate of increase in the electromyographic activity of the shoulder muscles was more rapid during the position task when the forearm was horizontal compared with the force task with the arm in the same position and with both tasks when the forearm was vertical. This observation indicates that the demands experienced by the postural muscles can limit the duration a task can be sustained, which has implications for the exercises that are prescribed during rehabilitation and for the ergonomic design of work stations.

These initial studies with the task failure approach suggest that it provides a strategy to identify the functional significance of the physiological adjustments that occur during fatiguing contractions. It is important to note that the approach should provide a foundation for the systematic evaluation of the task dependency of muscle fatigue and the extent to which the limiting factors generalize across muscles, postures, types of contractions, and individuals in health and disease.

See also: Computational Approaches to Motor Control; Corticomotoneuronal System; Corticospinal Development; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): Presynaptic Stretch Effects on Neuromuscular Transmission.

Further Reading

Allen DG (2004) Skeletal muscle function: Role of ionic changes in fatigue, damage and disease. *Clinical and Experimental Pharmacology & Physiology* 31: 485–493.

Baudry S, Klass M, Pasquet B, and Duchateau J (2007) Age-related fatigability of the ankle dorsiflexor muscles during concentric and eccentric contractions. *European Journal of Applied Physiology* 100: 515–525.

Danieli-Betto D, Germinario E, Esposito A, Biral D, and Betto R (2000) Effects of fatigue on sarcoplasmic reticulum and myofibrillar properties of rat single muscle fibers. *Journal of Applied Physiology* 89: 891–898.

Enoka RM (2008) *Neuromechanics of Human Movement*. Champaign, IL: Human Kinetics.

Gandevia SC (2001) Spinal and supraspinal factors in human muscle fatigue. *Physiological Reviews* 81: 1725–1789.

Hunter SK, Critchlow A, and Enoka RM (2005) Muscle endurance is greater for old men compared with strength-matched young men. *Journal of Applied Physiology* 99: 890–897.

Hunter SK, Duchateau J, and Enoka RM (2004) Muscle fatigue and the mechanisms of task failure. *Exercise and Sport Sciences Reviews* 32: 4449.

Maluf KS and Enoka RM (2005) Task failure during fatiguing contractions performed by humans. *Journal of Applied Physiology* 99: 389–396.

Nybo L (2003) CNS fatigue and prolonged exercise: Effect of glucose supplementation. *Medicine & Science in Sports & Exercise* 35: 589–594.

Nybo L and Secher NH (2004) Cerebral perturbations provoked by prolonged exercise. *Progress in Neurobiology* 72: 223–261.

Westerblad H, Allen DG, and Lannergren J (2002) Muscle fatigue: Lactic acid or inorganic phosphate the major cause. *News in Physiological Sciences* 17: 17–21.

Fear Conditioning and Synaptic Plasticity

J M Levenson, K P Mangan, P S Mukhopadhyay, and K J O'Riordan, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Changes in synaptic efficacy potentially represent a major mechanism whereby information is stored in the nervous system. Many studies have since characterized the mechanisms that underlie induction and expression of synaptic plasticity. However, surprisingly little direct evidence exists that directly implicates synaptic plasticity in memory formation in the nervous system.

Pavlovian fear conditioning is one of the most widely used behavioral paradigms for the study of the cellular and molecular bases of long-term memory formation. Fear conditioning is an attractive model system for the study of long-term memory formation, as the training session is brief, the behavioral expression of fear is robust and easy to quantify, and the circuitry that underlies the acquisition, storage, and expression of conditioned fear memory has been well characterized both anatomically and functionally. Moreover, the advent of modern molecular biology and molecular genetics has provided a plethora of information regarding the molecular and biochemical bases for conditioned fear memory. These properties make fear conditioning an ideal behavioral and neuroanatomical paradigm to investigate the hypothesis that synaptic plasticity is an important cellular mechanism for information storage in the nervous system. In the following sections we review the evidence that synaptic plasticity plays a role in storage of conditioned fear memories *in vivo*.

Pavlovian Fear Conditioning

Fear conditioning refers to training procedures whereby organisms learn to associate previously neutral stimuli with noxious or unpleasant events. Fear conditioning is based on the pioneering work of Ivan Pavlov, who demonstrated that neutral stimuli could acquire affective properties in general. During fear conditioning, stimuli that normally elicit no behavioral response are transformed to elicit robust fear responses. Fear in rodents is most commonly measured as either freezing behavior or potentiated startle reflexes. Fear conditioning begins by presenting an animal to a new environment and/or a neutral stimulus such as a tone or light. These neutral stimuli are referred to as the conditioned stimulus (CS). A noxious stimulus (shock, malaise-inducing chemical), the unconditioned stimulus (US), is then delivered coincident with the end of presentation of the CS. Close temporal association of the CS and US facilitates formation of an association between these two stimuli and a conditioned fear memory. Quantification of conditioned fear memory is performed by exposure of animals to the CS and measurement of stereotypic fear behaviors.

Anatomical Basis of Conditioned Fear

Conditioned fear has been used as a paradigm to study the cellular and molecular bases of memory formation for over 30 years. During this time, several classic experiments involving lesions, inactivation, track-tracing, and functional imaging have allowed for a thorough characterization of the neural circuitry that underlies acquisition and storage of conditioned fear memories. The amygdala and hippocampus are the primary regions of the brain involved in processing and storage of conditioned fear memory (Figure 1). The hippocampus receives information regarding the environmental context of conditioning and transmits this information into the basal (B) and accessory basal (AB) nuclei of the amygdala. The lateral nucleus of the amygdala (LA) receives input from the auditory cortex and auditory thalamus regarding auditory-based CS. The LA nucleus receives context information from inputs via the visual cortex. US information is transmitted to the LA, central (CE), and AB nuclei via distinct pathways. Behavioral and physiologic responses to fear are relayed from the amygdala to various target nuclei via the CE nucleus. Thus, the amygdala is critical for integration of US and CS, and the subsequent effect on behavior. The hippocampus is necessary for formation of context representations, but does not directly elicit fear-motivated responses. Collectively, these observations indicate that cellular and molecular events associated with formation of conditioned fear memory should occur in both the amygdala and hippocampus.

Molecular Mechanisms of Conditioned Fear

Molecular Fingerprint of Synaptic Plasticity

Synaptic plasticity refers generally to any change in the efficacy of a given synapse. That basic description covers a number of different types of plasticity, each with different durations and distinct mechanisms. A number of advances in our understanding of the molecular basis of synaptic plasticity have been made

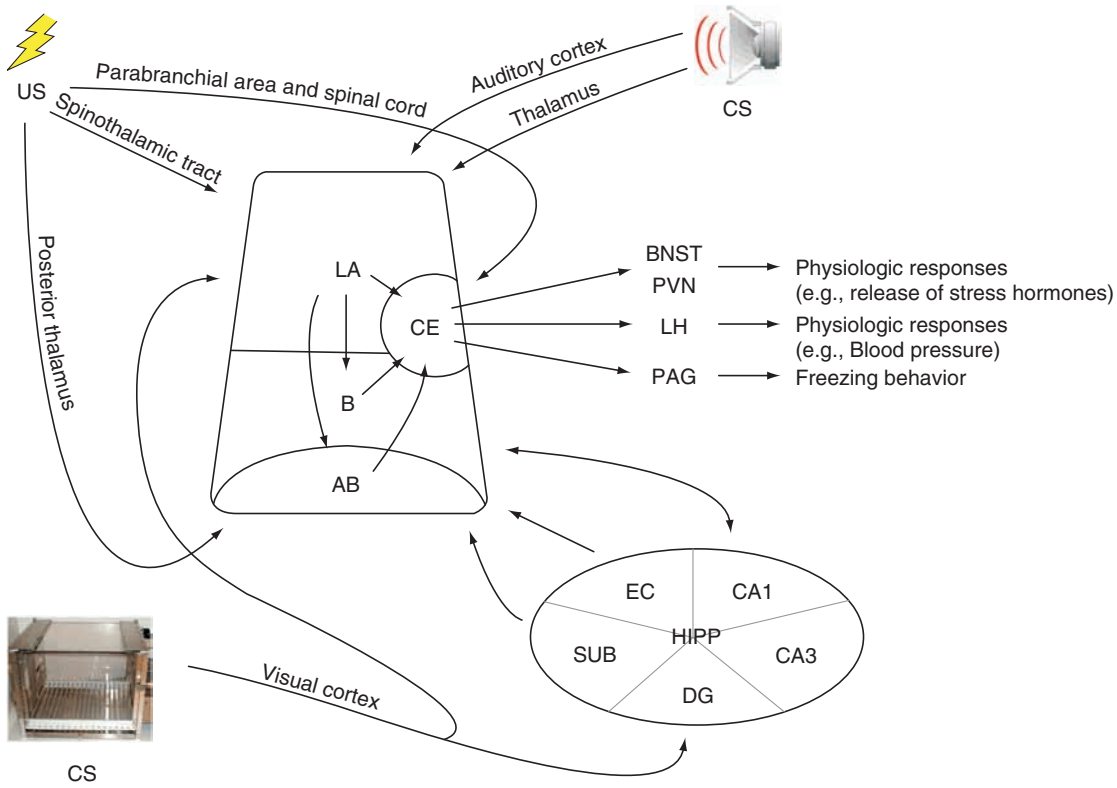

Figure 1 Anatomical basis of fear conditioning. Formation of conditioned fear memory requires the function of the amygdala. The hippocampus is required only for association of context with fearful stimuli. The amygdala serves to receive and integrate information regarding the conditioned stimulus (CS) and unconditioned stimulus (US), and then mediates the behavioral and physiologic response through outputs from the central (CE) nucleus. The hippocampus receives information regarding environmental context through the visual cortex and then transmits this information to the amygdala through several pathways into the basal (B) and accessory basal (AB) nuclei. Although not explicitly shown, many of the pathways contain reciprocal connections. LA, lateral nuclei of the amygdala; subfields of hippocampus – CA1, area CA1; CA3, area CA3; DG, dentate gyrus; SUB, subiculum; EC, entorhinal cortex; regions receiving CE afferents – BNST, bed nucleus of the stria terminalis; PVN, paraventricular nucleus of the hypothalamus; LH, lateral hypothalamus; PAG, periaqueductal gray of the brain stem; HIPP, hippocampus.

in the past 20 years. Studies of the molecular mechanisms that underlie synaptic plasticity have served two important purposes. First, they have elucidated novel mechanisms involved in all aspects of neuronal function. Second, they have provided a molecular signature that can be used to probe whether synaptic plasticity has occurred *in vivo*. For brevity, a schematic of many of the signaling pathways and molecules engaged during induction of synaptic plasticity has been included (Figure 2).

If synaptic plasticity mediates storage of conditioned fear memories, then the mechanisms responsible for induction and expression of plasticity should also be engaged in the amygdala and hippocampus during storage of conditioned fear memory. In this section we trace the molecular pathways implicated in synaptic plasticity (Figure 2) and review what is known regarding conditioned fear memory. When we refer to the hippocampus in this section, it is implied that we are specifically referring to processes involved in contextual fear conditioning.

Glutamate Receptors

Glutamate is the predominant excitatory neurotransmitter in the mammalian brain. Therefore, induction and expression of many forms of synaptic plasticity in the amygdala and hippocampus begin with activation of ionotropic glutamate receptors. The α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor (AMPA-R) is responsible for a majority of the basal excitatory synaptic transmission that occurs in the amygdala and hippocampus. Activation of the AMPA-R is the first step in the induction of synaptic plasticity, and subsequent regulation of AMPA-R expression and function represents key processes involved in expression plasticity. Pharmacologic enhancement of AMPA-R function increases the rate of conditioned fear learning but not the capacity, indicating that AMPA-Rs play a critical role in acquisition of memory. Interestingly, direct application of AMPA-R antagonists prior to testing blocks expression of conditioned fear memory, suggesting that

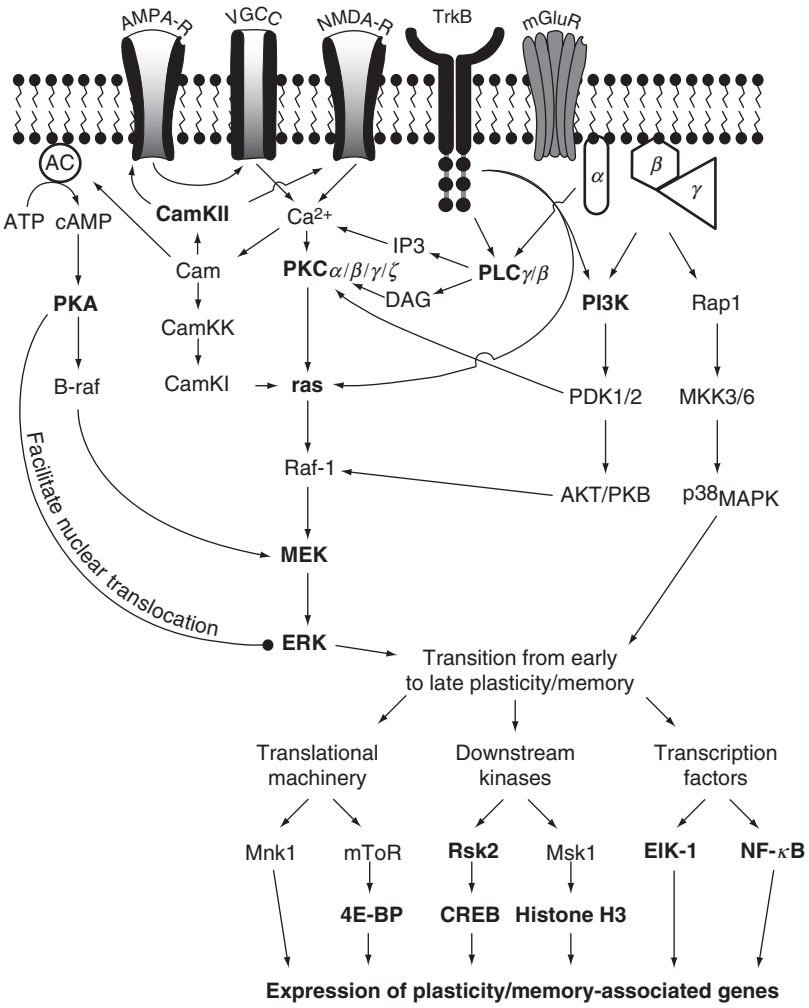

Figure 2 Molecular mechanisms implicated in both synaptic plasticity and conditioned fear memory. A schematic indicating the signaling pathways and molecular processes implicated in induction of synaptic plasticity and formation of long-term memory. The process begins at the postsynaptic membrane where *N*-methyl-D-aspartate receptor (NMDA-R) activation results in Ca^{2+} influx. The rise in intracellular Ca^{2+} results in activation of several parallel signaling pathways that ultimately converge on the mitogen-activated protein extracellular kinase–extracellular signal-regulated kinase (MEK–ERK) signaling pathway. Activation of ERK engages a complex cascade of events, including protein synthesis, kinases, and transcription factors, which results in a coordinated pattern of gene expression necessary for maintenance of long-term changes in synaptic efficacy. In parallel to NMDA-R-mediated events, activation of mGluRs results in recruitment of phosphatidylinositol 3-kinase (PI3K) and the p38–mitogen-activated protein kinase (MAPK) pathway. All molecules in bold have been directly implicated in conditioned fear memory. See text for discussion of other molecules.

AMPA-Rs also play a role in expression of memory. In support of a role for AMPA-Rs in expression of memory, work in several laboratories has documented an increase in surface expression of AMPA-Rs in the amygdala and hippocampus in response to conditioned fear training. Moreover, electrophysiologic recordings of tissue slices containing the medial geniculate nucleus–lateral nucleus of the amygdala (MGN–LA) synapse *in vitro* indicate that fear conditioning results in an increase in AMPA-R-mediated synaptic transmission. Collectively, these results indicate that AMPA-R-mediated function is necessary for both acquisition and expression of conditioned fear.

A critical step in induction of many forms of activity-dependent synaptic plasticity is activation of a unique class of ionotropic glutamate receptors, the *N*-methyl-D-aspartate receptor (NMDA-R). Formation of conditioned fear memory requires NMDA-R function both in the amygdala and the hippocampus. Pharmacologic inhibition of NMDA-R function in the amygdala prior to training completely blocks acquisition of conditioned fear. Blockade of NMDA-Rs in the hippocampus prevents acquisition of contextual, but not cued, conditioned fear. Pharmacologic inhibition of NMDA-Rs after training has no effect on expression or maintenance of conditioned fear,

indicating that NMDA-Rs specifically function in acquisition of memory. Recent experiments using genetically modified mice demonstrated that overexpression of NMDA-Rs in the forebrain enhances formation of conditioned fear memory. Collectively, these results indicate that NMDA-Rs are necessary for formation of conditioned fear memory and that these receptors are strictly involved in acquisition, but not expression, of conditioned fear.

Metabotropic glutamate receptors (mGluRs) are a final class of glutamate receptor that functions in synaptic plasticity. Activation of mGluRs has been implicated in induction of numerous forms of synaptic plasticity. Interestingly, both pharmacologic inhibition and activation of mGluRs in the amygdala and hippocampus interfere with acquisition of conditioned fear memory. Genetically modified animals that contain either gain or loss of function in specific isoforms of mGluRs exhibit similar phenotypes when acquisition of conditioned fear memory is significantly impaired. These results suggest that mGluRs function *in vivo* in acquisition of conditioned fear memory.

Brain-Derived Neurotrophic Factor

Growth factors play a critical role in several aspects of brain function, and several studies indicate that proper expression and function of growth factors and their associated receptors is necessary for transcription-dependent, long-term forms of synaptic plasticity. Brain-derived neurotrophic factor (BDNF) appears to play a unique role in activity-dependent synaptic plasticity. BDNF mRNA and protein, and the associated receptor, the tyrosine kinase receptor B (TrkB), are expressed in relatively high levels in the hippocampus and basolateral amygdaloid nucleus (BLA). BDNF mRNA is significantly increased in the hippocampus and BLA after fear conditioning. Overexpression of TrkB facilitates acquisition of conditioned fear memory. Pharmacologic and genetic disruptions of TrkB expression and function significantly attenuate conditioned fear learning. Extinction studies suggest that the BDNF/TrkB system is essential for consolidation, but not acquisition, of long-term conditioned fear memory. These studies indicate that the BDNF/TrkB system is engaged during consolidation of conditioned fear memory and represents another molecular correlate suggesting that synaptic plasticity occurs in the hippocampus and amygdala during consolidation of conditioned fear memory.

Signaling Pathways

The role of any membrane receptor is to transduce extracellular stimuli into intracellular signaling events. A great deal is known regarding the signaling pathways engaged during induction of synaptic plasticity. These signaling pathways provide another molecular signature that can be used to determine whether synaptic plasticity has occurred *in vivo* during acquisition and consolidation of conditioned fear memory.

The primary trigger of plasticity-associated signaling is an influx of Ca^{2+} through the NMDA-R (**Figure 2**). As reviewed previously, many studies indicate that NMDA-R function is critical for acquisition of conditioned fear memory. In addition, Ca^{2+} flux can also occur through voltage-gated Ca^{2+} channels (VGCCs). VGCCs have been implicated in both expression and consolidation of conditioned fear memory. After intracellular Ca^{2+} levels are increased, several signaling cascades are engaged to mediate expression and consolidation of conditioned fear memory. The activities of several signaling molecules increase in response to fear conditioning in both the amygdala and the hippocampus, including Ca^{2+}/calmodulin-dependent protein kinase II (CamKII), protein kinase A (PKA), protein kinase C (PKC), and extracellular signal-regulated kinase (ERK). Pharmacologic disruptions of CamKII, PKA, PKC, and mitogen-activated protein extracellular kinase (MEK) in either the amygdala or hippocampus inhibit acquisition of conditioned fear. Genetic disruptions of function, either through knockout or expression of a dominant negative enzyme, of CamKII, PKA, PKC, ras, and MEK all diminish or inhibit formation of conditioned fear memory. Enhancement of cAMP–PKA signaling pharmacologically or genetically increases acquisition of conditioned fear memory. Thus, many of the Ca^{2+}-dependent signaling pathways that are engaged during induction of synaptic plasticity are also activated in response to acquisition of conditioned fear memory.

Plasticity-associated signaling also occurs through non-Ca^{2+}-dependent pathways. For example, signaling through TrkB and mGluRs is necessary for induction of synaptic plasticity. Two major targets of TrkB and mGluR signaling are phospholipase C (PLC) and phosphatidylinositol 3-kinase (PI3K). The β and γ isoforms of PLC are activated by fear conditioning and can be found complexed with ERK-2. Overexpression of TrkB receptors results in enhanced PLC-γ signaling and memory formation. Infusion of a PI3K inhibitor into the hippocampus inhibits formation of contextual fear conditioning. Infusion into the hippocampus of phosphorylated growth factor receptor fragments that activate PI3K facilitates formation of contextual fear conditioning. Indirect evidence suggests that Akt, a kinase that is activated in response to PI3K signaling, is activated in the amygdala by fear conditioning. Considered together with the preceding observations, these studies indicate that

a significant overlap exists between the signaling pathways activated during induction of synaptic plasticity and formation of conditioned fear memory.

Transcription

One of the consequences of engagement of the plasticity-associated signaling pathways is regulation of gene expression. Early studies of long-term memory formation observed that direct injection of RNA synthesis inhibitors into the brain resulted in selective loss of long-term memory, indicating that gene expression was a critical component of long-term memory consolidation. These studies also indicated that mechanistically, memory formation was similar to synaptic plasticity in that short-term memories did not require synthesis of new macromolecules, but long-term memory required *de novo* protein and RNA synthesis. Using a candidate-based approach, initial studies of gene expression during memory consolidation identified several transcripts, the expression of which changes in the amygdala and hippocampus after fear conditioning, including the immediate-early genes c-*fos* and *egr-1*. Contemporary studies using a combination of molecular genetics and high-density oligonucleotide arrays have confirmed these early studies and extended them to include several additional genes, including those for BDNF, tissue plasminogen activator (tPA), activity-regulated cytoskeletal-associated protein (Arc), postsynaptic density-95 protein (PSD-95), and homer. Comprehensive characterization of gene expression during memory formation has revealed the involvement of several transcription factor families including cAMP response element-binding protein (CREB), CCAAT/enhancer-binding protein (C/EBP), and nuclear factor-kappa B (NF-κB). Notably, mouse models that contain either loss or gain of function for these transcription factors have been made and indicate that all of their genes are important for formation of conditioned fear memory. Expression of any one gene can be influenced by literally hundreds of different transcription factors. Thus, the significant degree of overlap between the patterns of gene expression observed after induction of synaptic plasticity and consolidation of conditioned fear memory strongly suggests that synaptic plasticity occurs *in vivo* during consolidation of conditioned fear memory.

Translation

One endpoint of gene expression is protein synthesis. Classic studies have demonstrated that infusion of protein synthesis inhibitors into the brain inhibits consolidation of conditioned fear memory. Several recent studies have discovered some of the molecules that are involved in regulation of protein synthesis during induction of synaptic plasticity. Fragile X mental retardation is an X-linked cognitive disease that is due to loss of function in one of several fragile X mental retardation (FMR) genes. All FMR genes are mRNA-binding proteins, act to transport mRNA to dendritic spines, and have been implicated in the transcription-dependent, late-phase of long-term potentiation (L-LTP). Loss of *FMR1* and *FMR2* results in a diminution of conditioned fear. Moreover, loss of the translational repressors 4E-BP2 and GCN2, which results in loss of L-LTP, inhibits consolidation of conditioned fear memory. Thus, several molecules implicated in protein synthesis and synaptic plasticity are necessary for formation of conditioned fear.

Direct Measures of Synaptic Plasticity during Conditioned Fear

Experiments into the molecular and cellular bases of formation of conditioned fear memory are suggestive, but demonstration of a change in synaptic efficacy *in vivo* would provide the most compelling evidence that synaptic plasticity mediates or at least contributes to storage of memory. At its core, synaptic plasticity is a change in the efficiency of communication at a particular synapse. One of the major mechanisms that underlies both LTP and long-term depression (LTD) is a redistribution in the complement of AMPA-Rs in the postsynaptic density (PSD). Induction of LTP results in an increase in AMPA-Rs, while LTD results in a decrease. Therefore, the relative level of AMPA-Rs in the PSD is a molecular correlate of synaptic efficacy. Several studies have demonstrated that levels of AMPA-Rs in the PSD increase after conditioned fear training in both the LA nucleus and area CA1 of the hippocampus, suggesting that plasticity occurs in these areas of the brain in response to formation of conditioned fear memories.

Induction of synaptic plasticity results in persistent physiologic changes in neurons. Input/output curves in area CA1 of the hippocampus were increased up to 7 days after fear conditioning, suggesting either synaptic transmission and/or neuronal excitability were increased. Efferents from the amygdala appear to be necessary for at least some of the conditioned fear-induced modulation of neuronal excitability in the hippocampus. Interestingly, several studies indicate that either conditioned fear modulates induction of synaptic plasticity, or that induction of synaptic plasticity modulates sensory-evoked potentials. In the hippocampus, conditioned fear occludes induction of LTP in slices *in vitro*. Slices isolated immediately after training exhibit impaired early-phase LTP (E-LTP), while slices isolated 24 h after training exhibit normal E-LTP but impaired L-LTP. Occlusion of LTP suggests that fear conditioning results in

induction of LTP in the hippocampus that prevents further potentiation of slices *in vitro*. In the amygdala, induction of LTP in the MGN–LA pathway results in enhancement of auditory-evoked potentials in this pathway, indicating that the endogenous systems for information processing share at least some mechanisms with synaptic plasticity.

Perhaps the most compelling evidence for synaptic plasticity in conditioned fear memory formation comes from a series of studies measuring changes in electrophysiology of the amygdala and hippocampus. Fear conditioning results in increases in synaptic efficacy in the MGN–LA pathway of the amygdala and the Schaffer-collateral pathway of the hippocampus. Implantation of electrodes *in vivo* has revealed that auditory fear conditioning enhances synaptically evoked potentials in the LA in a manner similar to what occurs when LTP is induced in this same pathway. Moreover, a hippocampus-dependent passive avoidance task increases evoked potentials in the Schaffer-collateral pathway of the hippocampus. In both the amygdala and the hippocampus, fear-induced increases in synaptic potentials occlude the effect of tetanus-induced LTP, suggesting that fear conditioning results in LTP-like changes in synaptic efficacy. Moreover, E-LTP induced in the dentate gyrus *in vivo* can be transformed to L-LTP by exposure of individuals to either an emotionally motivated experience or a hippocampus-dependent learning task, providing direct evidence of significant overlap in the molecular mechanisms used to consolidate long-term memory and long-term synaptic plasticity.

Summary and Conclusions

A significant portion of research into the cellular basis of information storage in the nervous system has focused on the synapse. A great deal has been learned about the mechanisms governing synaptic function and the processes that underlie synaptic plasticity. The convergence of cellular, molecular, and physiologic research on the mechanisms underlying synaptic plasticity have provided a comprehensive description of the processes involved in this phenomenon. This research has provided the necessary tools to test the hypothesis that synaptic plasticity occurs *in vivo* during memory formation. Indeed, several lines of compelling evidence exist that strongly suggest that synaptic plasticity occurs in response to formation of conditioned fear memory. However, all of the research presented thus far is correlative in nature; changes in synaptic responses mimic changes induced by artificial induction of LTP. No study to date has demonstrated that specific patterns of synaptic activity that mimic LTP or LTD induction occur *in vivo* during conditioned fear training and that these firing patterns are followed by long-term changes in synaptic efficacy. Moreover, no study has been able to artificially induce a fear memory through induction of synaptic plasticity in either the amygdala or the hippocampus. Thus, while there is still no direct evidence that synaptic plasticity is involved in memory formation, all indications are that synaptic plasticity is an important part of conditioned fear memory.

See also: Aggression: Neurochemical and Molecular Mechanisms; Amygdala: Contributions to Fear; Aversive Emotions: Molecular Basis of Unconditioned Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Genetics of Human Anxiety and Its Disorders; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Adams JP and Sweatt JD (2002) Molecular psychology: Roles for the ERK MAP kinase cascade in memory. *Annual Review of Pharmacology and Toxicology* 42: 135–163.

LeDoux JE (2000) Emotion circuits in the brain. *Annual Review of Neuroscience* 23: 155–184.

Maren S (2001) Neurobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience* 24: 897–931.

Maren S (2005) Synaptic mechanisms of associative memory in the amygdala. *Neuron* 47: 783–786.

Matynia A, Kushner SA, and Silva AJ (2002) Genetic approaches to molecular and cellular cognition: A focus on LTP and learning and memory. *Annual Review of Genetics* 36: 687–720.

Rodrigues SM, Schafe GE, and LeDoux JE (2004) Molecular mechanisms underlying emotional learning and memory in the lateral amygdala. *Neuron* 44: 75–91.

Rogan MT, Staubli UV, and LeDoux JE (1997) Fear conditioning induces associative long-term potentiation in the amygdala. *Nature* 390: 604–607.

Sigurdsson T, Doyere V, Cain CK, et al. (2007) Long-term potentiation in the amygdala: A cellular mechanism of fear learning and memory. *Neuropharmacology* 52(1): 215–227.

Uzakov S, Frey JU, and Korz V (2005) Reinforcement of rat hippocampal LTP by holeboard training. *Learning & Memory* 12: 165–171.

Whitlock JR, Heynen AJ, Shuler MG, et al. (2006) Learning induces long-term potentiation in the hippocampus. *Science* 313: 1093–1097.

Fetal Alcohol Syndrome

E P Riley, S N Mattson, and J D Thomas, San Diego State University, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Fetal alcohol syndrome (FAS) was first described by Kenneth Lyons Jones and David Smith in 1973 as a constellation of abnormalities resulting from heavy *in utero* alcohol exposure involving growth retardation, central nervous system (CNS) dysfunction, and distinct facial characteristics. Research during the three decades since that report has demonstrated that FAS is only one outcome on a continuum of effects resulting from prenatal exposure to alcohol. Considering the full range of possible outcomes resulting from prenatal alcohol exposure, FAS and its associated conditions represent a major public health problem and one that is entirely preventable. Herein, we review FAS, the effects of prenatal alcohol exposure on brain and behavior, possible mechanisms of damage, and recent attempts to mitigate these outcomes.

Fetal Alcohol Syndrome

FAS is the most physically apparent consequence of heavy gestational alcohol exposure, and while there are differences between the various diagnostic schemas, in general, anomalies are required in three distinct areas. Centers for Disease Control and Prevention (CDC) guidelines for the diagnosis of FAS require evidence of facial dysmorphology, growth problems, and evidence of CNS abnormalities. The facial dysmorphology consists of the presence of a smooth philtrum, a thin vermilion border, and small palpebral fissures ($\leq$10th centile) (Figure 1). Prenatal or postnatal growth retardation ($\leq$10th centile after adjustment for age, sex, gestational age, and race or ethnicity) is also required. The CNS abnormalities can be structural (significant anomalies as seen on imaging or a head circumference $\leq$10th centile), neurological, or functional. Functional anomalies are defined as global cognitive or intellectual deficits in multiple domains below the 3rd centile or deficits on three or more specific domains (e.g., executive or motor functioning, attention and hyperactivity, social skills, language, developmental deficits). As mentioned, other diagnostic schemas differ slightly. For example, a clarification of the Institute of Medicine (IOM) criteria for a diagnosis of FAS requires only two of the three facial features, but is more restrictive in terms of the CNS abnormality. In addition, this clarification suggests criteria for the diagnosis of partial FAS, alcohol-related birth defects (ARBDs), and alcohol-related neurodevelopmental disorders (ARNDs). Both the CDC and revised IOM guidelines allow the diagnosis of FAS to be made in the absence of information on maternal alcohol consumption. It is important to note that individuals with other syndromes (e.g., velocardiofacial, Williams, and de Lange syndromes) can display some of the characteristic facial features of FAS and therefore partial phenocopies should be ruled out prior to diagnosing FAS.

Although estimates of the prevalence of FAS vary depending upon the population under study and the type of ascertainment used, estimates generally run between 0.5 and 2 cases per 1000 births. However, in some areas, such as the Western Cape of South Africa, rates as high as 70/1000 are reported. Furthermore, FAS is only one outcome on a continuum of effects resulting from prenatal alcohol exposure, and the designation 'fetal alcohol spectrum disorders' (FASDs) is now used as an umbrella term for all possible outcomes resulting from prenatal alcohol exposure. Importantly, as suggested by the use of the term FASDs, both ARBDs and ARNDs can occur in the absence of FAS. Considering the full range of FASDs, physical and behavioral effects resulting from prenatal alcohol exposure may occur in as many as 1 in 100 births.

Animal models of FASDs have been developed that show a great deal of concordance with human findings. These models have helped to demonstrate the behavioral teratogenecity of alcohol, as confounding factors that may be present in clinical populations – such as polydrug use, nutritional factors, and poor prenatal care – can be controlled. Moreover, these models allow mechanisms to be assessed and potential interventions tested. For example, virtually all of the facial characteristics seen in FAS have been produced in mouse embryos by inducing apoptotic cell death in select cell populations in the anterior neural ridge, by exposing the embryos to alcohol during the equivalent of the third week of human gestation.

Effects of Alcohol on the Brain

The most deleterious consequence resulting from gestational alcohol exposure is the effect of alcohol on brain development and the impact that the brain changes have on behavior and cognition. Overwhelming evidence from both human and animal studies indicates that prenatal alcohol exposure can produce permanent changes in brain structure and function.

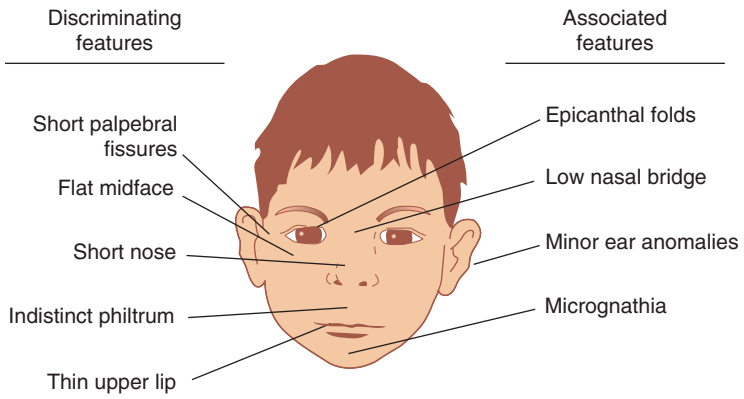

Figure 1 Features of fetal alcohol syndrome. Reproduced from Streissguth AP (1994) A long-term perspective of FAS. *Alcohol Health and Research World* 18(1): 74–81.

Early autopsy studies indicated a wide range of damage, including microcephaly, neuroglial heterotopias, agenesis of the corpus callosum, and cerebellar hypoplasia. Although one of the most consistent findings is that alcohol produces a reduction in brain volume, the brain is not uniformly vulnerable to a developmental alcohol insult.

One brain region particularly sensitive to developmental alcohol exposure is the cerebellum. Animal studies indicate that perinatal alcohol exposure reduces the number of cerebellar Purkinje and granule cells and alters cerebellar glial cells. These studies demonstrate that the cerebellum is particularly vulnerable during the equivalent of the human early third trimester, with lobules I–V and IX–X of the vermis being most sensitive. Consistent with these reports, magnetic resonance imaging (MRI) studies find about a 15% reduction in cerebellar volume in children with FAS and smaller volume reductions in those without FAS but with histories of heavy prenatal alcohol exposure. Reductions are more apparent in the anterior (lobules I–V) compared to the posterior cerebellar vermis. Thus, within the cerebellum there may be regional differences in sensitivity to alcohol-induced cell loss. Besides these size reductions, the anterior vermis and posterior-inferior vermis are spatially displaced in alcohol-exposed individuals. The degree of displacement of the anterior vermis is 1–3 mm and is most prominent on the anterior and superior edges. Interestingly, the degree of anterior vermis displacement is related to verbal learning and memory deficits.

Studies in children with FASDs suggest regional sensitivity to alcohol in the cortex. Analysis of the overall shape of the brain indicates narrowing in the inferior parietal/perisylvian region, as well as reduced growth in the inferior frontal region. Density mapping of gray and white matter in brain also supports abnormalities in the temporal–parietal regions, where increased gray matter and decreased white matter are apparent, particularly in the left hemisphere. Two small single-photon computed tomography (SPECT) studies support alcohol's effects in this brain region: one demonstrated mild hypoperfusion of the left hemisphere in the parietal–occipital regions and the other noted blood flow reductions in the temporal region.

Human MRI and animal studies indicate that the basal ganglia and, in particular, the caudate nucleus are disproportionately reduced in volume following heavy prenatal alcohol exposure. These results are supported by brain metabolism studies in both humans and animals showing reductions in activity in the thalamus and basal ganglia. Thus, both volumetric and metabolic changes support regional vulnerability in the teratogenic effects of alcohol. These effects on the caudate may be related to executive function and other cognitive deficits in FASDs.

The hippocampus is another subcortical structure affected by developmental alcohol exposure. Animal studies report that such exposure leads to cell loss in the hippocampus and lower dendritic spine density on pyramidal neurons, as well as aberrant mossy fiber projections. This exposure also alters synaptic activity in hippocampal brain slices. Human MRI data on a limited number of individuals indicate that while the hippocampus is smaller than in controls, it is unclear if these reductions are simply proportionate to overall brain size. Hippocampal changes could influence learning and memory in individuals with FASDs, given the significant role of the hippocampus in these cognitive domains.

Finally, abnormalities in the corpus callosum are reported in autopsy, imaging, and animal studies. In fact, it has been suggested that prenatal alcohol exposure is the most common cause of agenesis of the

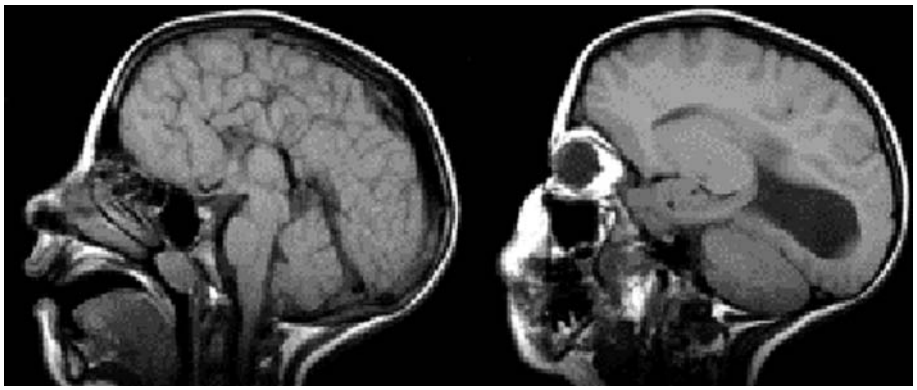

Figure 2 Magnetic resonance images of a 9-year-old girl with fetal alcohol syndrome. These sagittal images illustrate agenesis of the corpus callosum (left) and colpocephaly (right) present in this child. Reproduced from Mattson SN and Riley EP (1995) Prenatal exposure to alcohol: What the images reveal. *Alcohol Health and Research World* 19(4): 273–278.

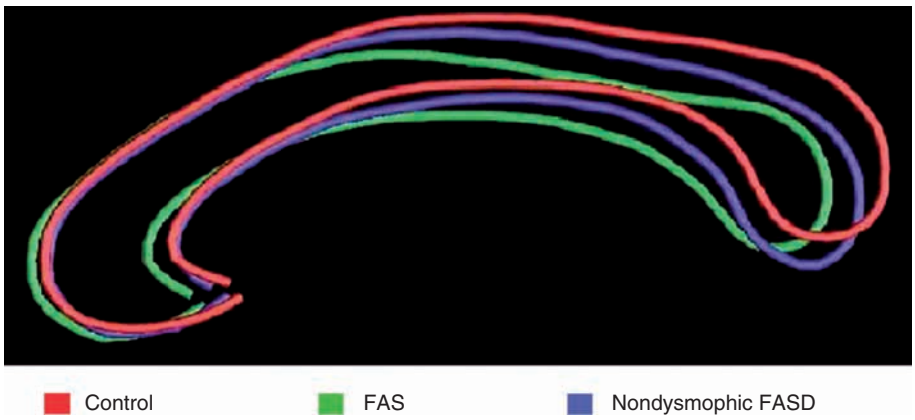

Figure 3 Average outlines of the corpus callosum (the major connection between the two hemispheres) in controls, children with fetal alcohol syndrome (FAS), and nondysmorphic children exposed to alcohol prenatally. The children in this last group were exposed to large amounts of alcohol prenatally, but they do not have the facial characteristics required for the FAS diagnosis. Notice that the corpus callosum of the two alcohol groups is displaced, with the greatest displacement occurring in the isthmus and splenium at the back of the corpus callosum. Also note that the nondysmorphic group is somewhat between the FAS and control groups. The corpus callosum is oriented so that the face would be on the left and the back of the head would be on the right. Reproduced from Sowell ER, Mattson SN, Thompson PM, Jernigan TL, Riley EP, and Toga AW (2001) Mapping callosal morphology and cognitive correlates: Effects of heavy prenatal alcohol exposure. *Neurology* 57(2): 235–244.

corpus callosum (Figure 2). Morphometric studies indicate that it is reduced in size and is spatially displaced. Specifically, the posterior corpus callosum in alcohol-exposed individuals is located more anterior and inferior than in controls (Figure 3). Verbal, but not nonverbal, learning/memory deficits are related to the degree of callosal displacement. In addition, new research suggests that assessment of callosal shape, via neonatal cranial ultrasound, may assist in the early identification of FAS. The integrity of the corpus callosum was recently investigated using diffusion tensor imaging in a sample of adults with FAS. Differences were noted in the diffusion fractional anisotropy and the apparent diffusion coefficient between alcohol-exposed and nonexposed young adults, supporting changes in white-matter integrity. Animal studies also report reductions in corpus callosum size, again most prominent in the posterior region. These reductions may be due to alcohol's effect on the development and function of oligodendrocytes, leading to a reduction or delay in myelin production, or to fewer neurons crossing within the corpus callosum.

It has been suggested that high levels of prenatal alcohol exposure may result in the mild end of holoprosencephaly spectrum malformation. Altered craniofacial development, ocular anomalies, and forebrain deficiencies, including hypoplasia of the corpus callosum, basal ganglia, and hippocampus, are consistent with this suggestion. Exposure to alcohol during a

discrete period of development may lead to deficiencies in the development of midline brain structures.

In addition to these structural changes, prenatal alcohol exposure alters the development and function of several neurotransmitter systems. For example, animal studies show that developmental alcohol exposure retards migration, reduces fiber density, and reduces the number of serotonergic neurons in various areas, including the midline raphe nuclei. Prenatal alcohol exposure also reduces glutamate–*N*-methyl-D-aspartate (NMDA) receptor number and activity in the hippocampus and cortex. Given the role glutamate plays in neuronal plasticity, long-lasting changes in glutamatergic signaling could contribute to learning deficits associated with developmental alcohol exposure. Such deficits could also be influenced by disruptions in cholinergic functioning. Alcohol exposure during development reduces the number of cholinergic neurons, reduces levels of choline acetyltransferase (ChAT), and alters the number of muscarinic receptors in regions such as the cortex and hippocampus. Reduced behavioral responsivity to cholinergic drug challenges provides further support for cholinergic hypofunctioning.

Dopaminergic activity is reduced by prenatal alcohol exposure, although the specific effects depend on timing and dose of alcohol. Alterations in dopaminergic functioning may contribute to the attention deficits observed in children with FASDs. Finally, although less well studied, prenatal alcohol also leads to alterations in noradrenergic, histaminergic, and γ-aminobutyric acid (GABA)ergic systems.

Neuroendocrine function is also affected by prenatal alcohol exposure, either indirectly by influencing maternal–fetal hormonal interactions or directly by altering the hormonal balance of the developing fetus. Animal studies demonstrate that prenatal alcohol exposure alters functioning of the hypothalamic–pituitary–adrenal (HPA) axis, the hypothalamic–pituitary–thyroid axis, and hypothalamic–pituitary–gonadal axis, which may influence stress and metabolic and reproductive function. Animals exposed to alcohol prenatally exhibit augmented responses to stress and take longer to return to baseline levels, although these effects may interact with sex. Hyperactivity of the HPA axis may also contribute to immune dysfunction, and limited data suggest that immune function is compromised following prenatal alcohol exposure.

Behavioral and Cognitive Effects in FASDs

Behavioral and cognitive changes related to prenatal alcohol exposure include deficits in sensory and motor function, attention, and learning as well as developmental delays, hyperactivity, and altered social behaviors. Recent research has focused on defining a profile of relative strengths and weaknesses in this population.

IQ Deficits

Probably the most devastating consequence of prenatal alcohol exposure is intellectual impairment. The mean IQ in FAS is in the low 70s, but the range is rather broad and includes functioning in the average range. In individuals with histories of heavy prenatal alcohol exposure, but without the facial characteristics of FAS, the average IQ is in the 80s, but again the range is rather broad. While frequently noted as one of the most common causes of mental retardation, the majority of individuals with FAS are not mentally retarded.

Motor Dysfunction

Given the cerebellar alterations, it is not surprising that prenatal alcohol exposure has been linked with problems in balance, motor coordination, and gait. The ability to maintain balance can be affected in children and adults with FASDs and, importantly, as the mechanisms involved in balance (e.g., visual, vestibular, and somatosensory cues) are challenged, deficits become more obvious. Animal studies have also demonstrated effects on gait, balance, and coordination. Moreover, balance deficits and motor incoordination following perinatal alcohol are correlated with the severity of cerebellar Purkinje cell loss.

Learning and Memory Impairments

Multiple studies have suggested that perinatal alcohol exposure leads to deficits in learning and memory. In humans, these include verbal learning deficits and nonverbal learning and memory deficits. The pervasiveness of the nonverbal deficits may be related to visuospatial dysfunction. For example, recent studies utilizing a virtual Morris maze show deficits in place but not cued navigation in children with FASDs. Animal studies also demonstrate deficits in Morris maze spatial learning that are likely related to alcohol-induced hippocampal or frontal lobe dysfunction. Additional learning deficits have been noted in human and animal studies using a classically conditioned eyeblink response, a task that engages neural circuitry in the cerebellum and brain stem. In animals, these delay conditioning deficits have been related to alterations in interpositus nucleus activity in the cerebellum and have been correlated with the loss of cerebellar deep nuclear cells, which are reduced by about 50% in alcohol-exposed animals. In one human study, children exposed to alcohol prenatally were compared to children with

either dyslexia or attention-deficit/hyperactivity disorder (ADHD), but no history of alcohol exposure. Both alcohol-exposed and dyslexic children failed to learn the conditioned eyeblink response.

Executive Function Deficits

Another pervasive characterization in this population is one of executive function deficits. Weaknesses in planning, response inhibition, verbal and nonverbal fluency, set-shifting, and reversal learning have all been documented in individuals with FASDs. Executive function deficits have been related to the frontal-subcortical system, which is affected by prenatal alcohol exposure. Although preliminary, a recent functional MRI (fMRI) study in individuals with FASDs reported increased functional activation in the inferior and middle frontal cortices on a fairly simple spatial working memory task, another domain of executive functioning. Among findings in animal models of FASDs are deficits in response inhibition and reversal learning. For example, rodents perinatally exposed to alcohol typically perform worse than do controls on tasks requiring successive reversals. In one study, impaired reversal learning was associated with cell loss in the medial prefrontal cortex. Executive functioning deficits may cause profound and disabling difficulties in everyday life and may account for the problems in social, legal, alcohol use, and mental health domains in individuals with histories of heavy prenatal alcohol exposure.

Activity Levels

Deficits in response inhibition may also be expressed as changes in activity level. Since the development of animal models of FASDs in the late 1970s, increased activity and exploration behaviors in the open field are among the most common findings. Although empirical data from clinical populations regarding activity level is limited, children with FASDs have been described as hyperactive and fidgety, characteristics that may also be related to deficits in attention.

Attention Deficits

Children with prenatal alcohol exposure are commonly diagnosed with ADHD. A case-control study of psychiatric inpatients revealed that cases with ADHD were 2.5 times more likely to have been exposed to alcohol compared to control cases, and this was independent of other important risk factors for ADHD, such as smoking or familial factors. Although prenatal alcohol exposure as a risk factor for ADHD has not been universally supported, examination of attention in FASDs suggests that this domain represents a core deficit in this population. However, these attention deficits are not global in nature; visual attention deficits are pervasive whereas auditory deficits are dependent on task demands. Research comparing children with heavy prenatal alcohol exposure to children with ADHD but no history of alcohol exposure is limited, but there is some suggestion that the two groups can be distinguished from each other based upon differences in performance on tests of vigilance and visual spatial reasoning, and in the eyeblink conditioning study mentioned previously.

Psychiatric Comorbidity

In addition to ADHD, other psychiatric disorders are frequently mentioned in conjunction with FASDs. Rates of depression between 27% and 44% have been noted in individuals with histories of prenatal alcohol exposure, and 25% of a sample of adults with FASDs had attempted suicide. Importantly, there is evidence that the level of depression in children with prenatal exposure histories cannot be explained by maternal depression or level of current drinking in the home. Higher rates of disruptive, anxiety, and substance use disorders have also been reported.

Risk Factors

One obvious risk factor for FASDs is the level of alcohol exposure. Alcohol's adverse effects on development are dose dependent, and heavier maternal consumption is associated with more severe outcomes. However, there are data suggesting that even low levels of alcohol may damage the developing fetus. In addition to alcohol dose, pattern of consumption is an important factor. For example, animal models have shown that binge drinking results in more severe neuropathology and behavioral alterations than does more chronic exposure. Clinical studies also suggest that maternal binge drinking places the fetus at higher risk for FASDs. Developmental timing of exposure is another factor. For example, using a mouse model, it was demonstrated that facial features associated with FAS are related to alcohol exposure during a discrete period early in gestation, and that a difference in the day of exposure (gestation day 7 vs. 8) could cause different outcomes in terms of facial characteristics.

Maternal and/or fetal genetic factors also interact with alcohol exposure and affect the relative risk for FASDs. For example, different alleles for the *ADH1B* gene may play a role in the etiology of FAS. Maternal *ADH1B*2* and possibly *ADH1B*3* alleles, which are associated with more rapid alcohol metabolism, reduce the risk of FAS. Using crosses of selectively

bred lines, it has been shown that maternal genotype related to functional alcohol sensitivity influences the outcome in the offspring, regardless of the offspring's genotype. Prenatal care, nutritional factors, maternal age, parity, and use of other drugs of abuse may also interact with prenatal alcohol exposure in determining outcomes.

Mechanisms of Damage and Potential Interventions

Alcohol disrupts brain development via a number of mechanisms and affects many stages of neuronal and glial development, including proliferation, migration, differentiation, and cell survival. Moreover, the mechanisms of damage depend on the developmental timing of alcohol exposure and characteristics of various cellular populations. For example, alcohol alters the cell cycle during proliferation, alters radial glial cells and neuronal responses to cell adhesion molecules, alters neurite outgrowth and dendritic branching, and leads to both apoptotic and necrotic cell death. Several putative mechanisms have been investigated using animal model systems.

Based upon similarities in neuropathology between children with FAS and those with mutations in the L1 cell adhesion molecule, it is hypothesized that alcohol might interfere with normal cell adhesion. *In vitro* studies have shown that ethanol impairs neuronal responses to L1 cell adhesion molecules, which are important for neurite extension and neuronal migration (**Figure 4**). Likewise, coadministration of compounds that antagonize ethanol's inhibition of the L1 molecule should protect against alcohol's teratogenicity. For example, octanol inhibits ethanol's effects on L1, and when coadministered with ethanol inhibits toxicity to the embryo (**Figure 5**).

Another possible teratogenic mechanism is oxidative stress, which leads to cellular damage and death. Perinatal alcohol exposure increases the production of destructive reactive oxygen species and disrupts intracellular antioxidant capacity. Consistent with this mechanism, administration of antioxidants, such as vitamin E, during ethanol exposure can reduce oxidative stress and protect against ethanol-related hippocampal and cerebellar cell death, although the functional significance of this has yet to be established. Administration of a synthetic superoxide dismutase and catalase mimetic antioxidant, EUK-134, reduces ethanol-induced apoptotic cell death and limb malformations in whole mouse embryos. The success of antioxidant treatment may depend on the particular antioxidant agent used and the cell population studied, as well as timing and dosing of alcohol exposure.

The neuroactive small peptide fragments NAPVSIPQ (NAP) and SALLRSIPA (SAL) have antioxidant properties as well as other neuroprotective actions. Treatment with these peptides prior to or shortly after prenatal ethanol exposure reduces alcohol-induced fetal death, growth retardation, and neuropathology. Whether these peptides are effective because of their antioxidant properties, ability to block ethanol's effects on L1 cell adhesion molecules, or another protective mechanism is unknown. Recently, it was reported that nicotinamide administration shortly after perinatal alcohol exposure in the mouse significantly attenuated activation of caspase-3, apoptotic cell death, and behavioral deficits. Nicotinamide is a novel neuroprotective agent that can protect against various insults, such as oxidative stress.

Ethanol also disrupts the production or response to a variety of growth factors, disruptions that may reduce cell proliferation as well as survival. Furthermore, given that some neurotransmitters exert trophic effects during development, alcohol-related reductions in neurotransmitter levels can augment ethanol's teratogenic effects. For example, serotonin exhibits neurotrophic effects during development primarily by activating both neuronal and glial serotonin ($5\text{-}HT_{1A}$) receptors, and alcohol-related reductions in serotonergic activity could alter the development of a number of neuronal populations, including forebrain structures. Concurrent administration of serotonin $5\text{-}HT_{1A}$ agonists, buspirone and ispaspirone, can prevent alcohol's damaging effects on developing serotonergic and other rhombencephalic neurons.

Ethanol may also disrupt development by acting on receptors and ion channels. For example, perinatal alcohol exposure leads to widespread apoptotic cell death and this may be related, in part, to ethanol-induced blockade of NMDA receptors and stimulation of $GABA_A$ receptors. In addition, blockade of NMDA receptors with the antagonists MK-801 or eliprodil during perinatal ethanol withdrawal protects against some ethanol-related behavioral alterations. These results suggest that neurocompensatory overactivity of NMDA receptors may lead to excitotoxicity. Thus, both the acute action and withdrawal from alcohol may be damaging to the developing brain.

Indirectly, prenatal ethanol exposure can also disrupt nutrient absorption and maternal–fetal blood flow. Moreover, acetaldehyde, a metabolite of ethanol, is also teratogenic. Other factors that may contribute to alcohol-related damage include retinoic acid deficiencies, altered regulation of gene expression, and endocrine-related mechanisms. Thus, alcohol disrupts brain development via numerous and complex mechanisms. Elucidation of these mechanisms may lead to the identification of effective treatments.

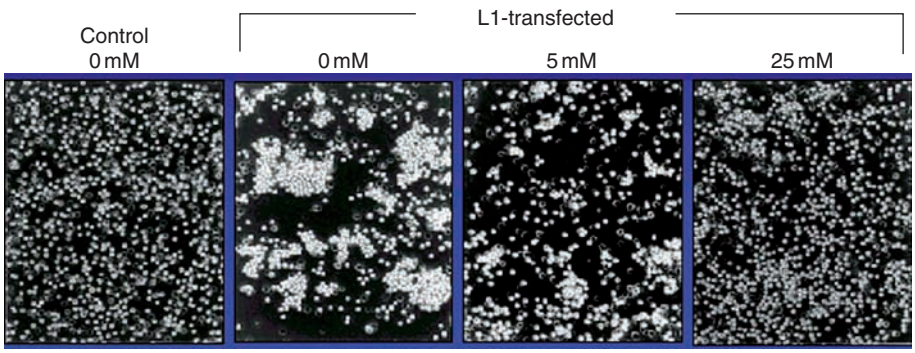

Figure 4 Ethanol seems to alter cell responses to molecules that regulate neuronal proliferation, migration, and differentiation. One interesting mechanism involves L1 and other cell adhesion molecules. These are essential for normal human nervous system development. L1 and other cell adhesion molecules guide neuronal migration and tract formation during nervous system development, and mutations in the gene for L1 result in brain malformations. Scientists recently recognized that children born with L1 mutations exhibit mental retardation, spastic gait, and a variety of brain malformations, including enlarged ventricles (hydrocephalus), and agenesis of the corpus callosum. Noting that retardation, hydrocephalus, and agenesis of the corpus callosum also occur in fetal alcohol syndrome, Michael Charness and colleagues asked whether alcohol inhibits the adhesiveness of cells bearing the L1 molecule. L cells (mouse fibroblasts) were engineered to express the human gene for the cell adhesion molecule L1 (see the three slides labeled 'L1-transfected'). The cells were allowed to aggregate by gentle shaking for 30 min. Control cells (left slide) form few clusters of adherent cells. Cells transfected with human L1 cell adhesion molecules aggregate much more than do control cells. Of the L1-transfected cells, the first slide (0 mM) shows this aggregation (white clumps) when alcohol is not added to the system. The slides showing the two other L1-transfected cells (5 and 25 mM) show a dose-dependent decrease in aggregation when ethanol is added to the medium. This increase in cell adhesion is inhibited by ethanol in a dose-dependent manner. Ethanol seems to alter cell responses to molecules that regulate neuronal proliferation, migration, and differentiation. Reproduced from Ramanathan R, Wilkemeyer MF, Mittal B, et al. (1996) Alcohol exhibits cell–cell adhesion mediated by human L1. *Journal of Cell Biology* 133(2): 381–390.

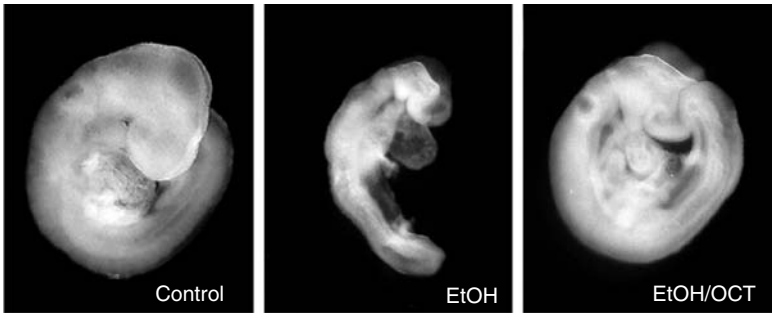

Figure 5 Representative morphology of cultured control and alcohol-exposed C57BL/6J mouse embryos at gestational day 8. The embryos in the middle and right panels were exposed to 100 mM alcohol (EtOH) and 100 mM EtOH + 3 μM 1-octanol (OCT), respectively; administration of octanol significantly reduced the severity of alcohol-related teratogenicity. Reproduced from Chen S-Y, Wilkemeycer MF, Sulik KK, and Charness ME (2001) Octanol antagonism of ethanol teratogenesis. *FASEB Journal* 15(9): 1649–1651.

In hopes of intervening in affected individuals with FASDs, factors known to positively influence brain structure and function have been applied to animal models of FASDs. For example, rearing in an enriched environment, motor learning rehabilitation, and exercise improve outcomes in animals exposed to alcohol perinatally. Motor rehabilitation initiated in adulthood increases synaptic connections in the cerebellum and improves motor coordination in subjects exposed to alcohol during development. More recent evidence suggests that even exercise alone can increase hippocampal plasticity and increase neurogenesis in adult animals exposed to alcohol during development. Finally, dietary treatments may also be effective. For example, choline supplementation can reduce the severity of overactivity and learning deficits associated with developmental alcohol exposure even when administered after ethanol treatment has ceased. Treatments that are relatively safe to administer may hold promise for translation into the clinical population.

Summary

FASDs are a devastating, yet preventable, public health concern that span the physical, neurological, behavioral, cognitive, and social domains. Following

heavy prenatal alcohol exposure, both structural and functional changes in the brain are documented. These changes are permanent and are seen in children with FAS and in those that have fewer or none of the facial characteristics of the syndrome. Because they occur in both dysmorphic and nondysmorphic individuals, identification of the full spectrum can be challenging. Efforts to define a behavioral and cognitive profile may assist in more accurate identification of children affected by prenatal alcohol exposure and thus provide the basis for interventions that are desperately needed in this population. Research using animal model systems has both supported and provided impetus for clinical studies. Alcohol-related changes in brain structure, neurochemistry, and resulting behavior have all been documented. Importantly, research using animal model systems has recently identified several promising interventions that may lead to either prevention or treatment of FASDs. Given that women continue to drink alcohol during pregnancy, identification of safe and effective treatments is among the most important challenges.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Alcoholism; Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine; Brain Damage: Functional Reorganization; Gene Therapy and Protection from Stress-Induced Brain Damage; Single Photon Emission Computed Tomography (SPECT); Stress and Vulnerability to Brain Damage; Substance Abuse and Dependence.

Further Reading

Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities, National Task Force on Fetal Alcohol Syndrome and Fetal Alcohol Effects (2004) Fetal alcohol syndrome: Guidelines for referral and diagnosis. Available at http://www.cdc.gov/ncbddd/fas/.

Chen S-Y, Wilkemeyer MF, Sulik KK, et al. (2001) Octanol antagonism of ethanol teratogenesis. *FASEB Journal* 15: 1649–1651.

Chudley AE, Conry J, Cook JL, et al. (2005) Fetal alcohol spectrum disorder: Canadian guidelines for diagnosis. *Canadian Medical Association Journal* 172(supplement 5): S1–S21.

Cohen-Kerem R and Koren G (2003) Antioxidants and fetal protection against ethanol teratogenicity. I. Review of the experimental data and implications to humans. *Neurotoxicology and Teratology* 25: 1–9.

Cudd TA (2005) Animal model systems for the study of alcohol teratology. *Experimental Biology and Medicine* 230: 389–393.

Floyd RL, O'Connor MJ, Sokol RJ, et al. (2005) Recognition and prevention of fetal alcohol syndrome. *Obstetrics and Gynecology* 106: 1059–1064.

Goodlett CR, Horn KH, and Zhou FC (2005) Alcohol teratogenesis: Mechanisms of damage and strategies for intervention. *Experimental Biology and Medicine* 230: 394–406.

Hoyme HE, May PA, Kalberg WO, et al. (2005) A practical clinical approach to diagnosis of fetal alcohol spectrum disorders: Clarification of the 1996 Institute of Medicine criteria. *Pediatrics* 115: 39–47.

Jones KL, Smith DW, Ulleland CN, et al. (1973) Pattern of malformation in offspring of chronic alcoholic mothers. *Lancet* 1(7815): 1267–1271.

Mattson SN and Riley EP (1995) Prenatal exposure to alcohol: What the image reveal. *Alcohol Health and Research World* 19(4): 273–278.

McGee CL, Fryer SL, and Mattson SN (2006) Lessons learned: Fetal alcohol spectrum disorders. In: Bellinger DC (ed.) *Clinical and Developmental Neurotoxicology*, pp. 169–191. New York: Taylor & Francis.

Olney JW (2004) Fetal alcohol syndrome at the cellular level. *Addiction Biology* 9: 137–149.

Ramanathan R, Wilkemeyer MF, Mitlal B, et al. (1996) Alcohol exhibits cell–cell adhesion mediated by human L1. *Journal of Cell Biology* 133(2): 381–390.

Riley EP and McGee CL (2005) Fetal alcohol spectrum disorders: An overview with emphasis on changes in brain and behavior. *Experimental Biology and Medicine* 230: 357–365.

Sowell ER, Mattson SN, Thompson PM, Jernigan TL, Riley EP, and Toga AW (2001) Mapping callosal morphology and cognitive correlates: Effects of heavy prenatal alcohol exposure. *Neurology* 57(2): 235–244.

Stratton K, Howe C, and Battaglia F (1996) *Fetal Alcohol Syndrome: Diagnosis, Epidemiology, Prevention, and Treatment*. Washington, D.C: National Academy Press.

Streissguth AP (1994) A long-term perspective of FAS. *Alcohol Health and Research World* 18(1): 74–81.

Sulik KK (2005) Genesis of alcohol-induced craniofacial dysmorphism. *Experimental Biology and Medicine* 230: 366–375.

Zhang X, Sliwowska JH, and Weinberg J (2005) Prenatal alcohol exposure and fetal programming: Effects on neuroendocrine and immune function. *Experimental Biology and Medicine* 230: 376–388.

Relevant Websites

http://www.med.unc.edu – Bowles Center for Alcohol Studies, University of North Carolina at Chapel Hill.

http://fascenter.samhsa.gov – Center for Substance Abuse Prevention's Fetal Alcohol Syndrome (FAS) Center for Excellence.

http://www.cdc.gov – Centers for Disease Control and Prevention: National Center on Birth Defects and Developmental Disabilities.

http://depts.washington.edu – Fetal alcohol and drug unit, University of Washington.

http://www.niaaa.nih.gov – National Institute on Alcohol Abuse and Alcoholism.

http://www.nofas.org – National Organization on Fetal Alcohol Syndrome.

http://www.phac-aspc.gc.ca – Public Health Agency of Canada.

http://www.med.unc.edu – University of North Carolina FAS slide show.

http://www.rsoa.org – Research Society on Alcoholism.

Finger Movements: Control

A M Smith, Université de Montréal, Montréal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

A Brief History

Since at least the latter half of the nineteenth century, the motor cortex has been considered the most important brain region controlling the hand. The clinical link between small lesions of the motor cortex and contralateral paresis, particularly those restricted to the hand, was taken by Hughlings-Jackson, Ferrier, and Charcot as persuasive evidence in favor of cortical localization. For these early pioneers of neurology, the motor cortex also included what today is known as the primary somatosensory cortex. However, research during the first half of the twentieth century using better anatomical and physiological methods demonstrated that although both the motor and somatosensory cortical areas contribute corticospinal axons to the pyramidal tract, the axons of the primary somatosensory cortex terminate in the dorsal horn and do not seem to affect movement directly. The electrical stimulation studies of Penfield and Woolsey demonstrated a clearer picture of the somatotopography of the motor cortex and the disproportionately large motor representation of the regions controlling the face and hand. These and other studies showed that several regions in the frontal lobe, collectively called premotor areas, also had somatotopically organized body representations.

Using staining techniques for visualizing degenerating axonal terminals, Hans Kuypers was the first to compare spinal terminations of the pyramidal tract with those of the brain stem motor pathways. In order to produce the necessary corticospinal degeneration, Kuypers sectioned one side of the medullary pyramid and found that approximately 80% of the fibers crossed the midline to descend the cord in the lateral funiculus. The uncrossed remaining 20% of fibers descended in the anterior funiculus. Knowing myotopic organization of the ventral horn and intermediate zone, Kuypers inferred that the uncrossed fibers act on proximal and axial muscles, whereas the crossed fibers act on the more distal limb muscles. Kuypers further demonstrated that in Old World monkeys and apes, a select number of corticospinal axons originating from the motor cortex terminate directly on the dorsolateral motor neurons of lower cervical cord innervating the muscles of the hand.

In subsequent behavioral studies, Kuypers and his colleagues, Lawrence and Hopkins, studied the effects of bilateral bulbar pyramidotomy. They found that the monkeys recovered remarkably well, in contrast to the near permanent hemiplegia neurologists have long associated with infarctions of the motor cortex and internal capsule. However, the investigators made the interesting observation that destruction of the corticospinal tract, whether in the infant or the adult monkey, led to a loss of manual dexterity, ostensibly due to a loss of relatively independent finger movements. That is, the monkeys recovered whole-hand grasping and the ability to use the hand for climbing but were unable to use a single finger to winkle out pieces of food from narrow food wells. Nevertheless, Weisendanger and Hepp-Reymond found that monkeys could still perform a thumb–forefinger precision grip after bilateral pyramidotomy although both their reaction times and their force-building times were significantly increased.

At about the same time, Charles Phillips's electrophysiological studies demonstrated strong monosynaptic connections between the motor cortex and the motor neurons of both intrinsic and extrinsic wrist and finger muscles in the baboon. Also about this time, Edward Evarts took the bold step of recording from single identified pyramidal tract cells in awake monkeys trained to perform specific wrist movements. These studies confirmed that many pyramidal tract neurons discharged prior to voluntary hand movements, a finding consistent with a causal relationship. Evarts went on to show that the discharge frequency of at least some pyramidal tract neurons was directly related to wrist muscle force. Movements produced by external forces such as gravity were not associated with pyramidal tract activity, thereby demonstrating that neither movement direction nor final position was directly encoded by the discharge frequency of motor cortical neurons. Subsequent experiments by Georgopoulos and collaborators suggested that the activity of each neuron might be thought of as having a 'preferred direction' – and that populations of coactive motor cortical cells produce a 'population vector' – that determines the resultant movement direction of a limb.

These general observations remain both valid and relevant today, although a number of more recent discoveries have added considerably to our understanding of how the motor cortex controls finger movements. Brodmann's area 4 is no longer seen as the singular final common path to the motor areas of the spinal cord, and motor areas of the cortex must include the dorsal premotor area (PMd), the ventral premotor (PMv) area, and the supplementary motor area (SMA) and three cingulate motor

areas on the medial wall of the hemisphere. The digit representation in Brodmann's area 4, which is commonly accepted as the primary motor cortex (M1), receives direct input from six separate cortical areas. Dum and Strick showed that each of these premotor areas also has substantial direct projections to the spinal cord that terminate in the motor regions of the intermediate zone. Additional anatomical studies found that the PMv and the PMd have individual digit representations, identified by movements evoked by microstimulation, and that both regions are densely interconnected with M1. It was unexpected that the projections from the SMA were stronger to the PMd and PMv than those with M1.

In general, the current anatomical data do not support the belief that the motor functions of the cerebral cortex are strictly hierarchically organized. Single cell recording studies in the behaving monkey have substantiated that each of these areas is active during voluntary hand movements, although the distinct contribution of each of these areas to hand function is still unknown.

Corticomotor Neuronal Connections

As one might expect, the direct corticomotor (CM) neuronal connections have received the most attention with respect to hand function because of the initial association with independent finger movements. However, as the Weisendanger and Hepp-Reymond studies have shown, this is probably not an exclusive or unique function of the CM neurons. Initially the CM–neuron projections were demonstrated with intracellular recordings of monosynaptic excitatory postsynaptic potentials (EPSPs) in cervical motor neurons in anesthetized monkeys. However, another method, called spike-triggered averaging, has been used to identify these neurons in awake, behaving monkeys. This method, initially implemented by Fetz and Cheney, is based on the supposition that individual EPSPs generated by CM cells in target motor neurons should increase the probability of discharge in motor units at a fixed latency. By averaging the electromyographic (EMG) activity immediately following each action potential from a putative CM cell, the appearance of time-locked postspike facilitation can be demonstrated after averaging hundreds or thousands of times. Another version of the method uses stimulus-triggered averaging, in which an intracortical microelectrode applies single pulses to activate a small region of CM cells. Studies employing these methods have shown that single CM neurons act on the motor neurons of distal muscles, that is, both the extrinsic and intrinsic muscles of the wrist and fingers. However, recent evidence suggests that shoulder and elbow muscles may also have CM connections and the same site may cofacilitate muscles acting at contiguous joints synergistically. Furthermore, CM neurons generally establish synaptic contacts with the motor neurons of more than one muscle, and a single cortical CM cell can excite the motor neurons of as many as four or five muscles. In general, these synaptic contacts are found on muscles that tend to be coactive as opposed to mechanically opposing antagonists. In fact, at a slightly longer latency, postspike suppression of the electromyogram has also been reported in some antagonist muscles.

A recent study by Rathelot and Strick has introduced a new anatomical method for studying CM connections, and this has revealed a number of new and important findings. Rathelot and Strick were able to inject a small group of intrinsic and extrinsic hand muscles in monkeys with rabies virus and use the transsynaptic retrograde transport of the virus to trace the origin of CM cells in the motor cortex. They injected three thumb and finger muscles and showed that the origin of the CM cells projecting to these muscles was widespread, extending medially well beyond the finger representation area into a region that evoked shoulder movements when stimulated. The CM cells originated mainly from area 4, deep within the rostral bank of the central sulcus, and also from area 3a, that part of what was traditionally considered the somatosensory cortex receiving input from primary muscle spindles. Very few CM cells were found on the more rostral precentral convexity. In general, there was no evidence for separate anatomically segregated representations of each digit or each digit muscle in the motor cortex. Instead the hand muscles appear to be represented by multiple colonies spread over broad territories within M1, as suggested by Phillips many years earlier.

Sensory Afferents to the Motor Cortex Hand Area

It has long been known that the motor cortex, like the somatosensory cortex, receives afferents from skin and deeper tissues such as muscles, tendons, and joints. Overall, the locations of afferent sensory receptive fields match the target of motor output. However, the anatomical route of the sensory input to the motor cortex is still uncertain. In addition, the topography of this input–output relationship is not quite the 'finely graded mosaic' of input–output relations that Asanuma proposed in 1981. The discovery of CM neurons with output connections to hand muscle motor neurons in the shoulder regions of the motor cortex increases the likelihood that synergies of the shoulder, elbow, and hand muscles

may be mediated by single cortical neurons. Also hand-related neurons in motor cortex with receptive fields split between the hand and mouth are not uncommon. Consequently, finger muscle activity may be influenced by other selected regions of the body, such as the mouth or shoulder. Whether this receptive field organization is a reflection of a genetically determined inherent connectivity or whether it is the product of use-dependent plasticity is at present unknown. Several studies indicate that both the receptive fields and the hand movements evoked by electrical stimulation can be modified by experience. These use-dependent plastic changes are mediated in part by local inhibitory circuits within the cortex as opposed to the growth of new connections within the spinal cord.

Strick and Preston demonstrated that in the squirrel monkey, the motor cortex contained two distinct representations of the wrist and fingers, one rostral and the other caudal bordering on area 3a. On the output side, the two representations appeared to be identical in terms of the movements and EMG activity evoked by microstimulation, but the two zones differed markedly on the afferent side. Neurons receiving input from deep tissue were located in the rostral zone, whereas neurons receiving input mainly from glabrous skin on the volar surface of the hand were located in the caudal region bordering on area 3a. A similar organization was reported by Picard and Smith in unanesthetized fascicularis monkeys. These findings, taken together with the location of CM cells deep within the rostral bank of the central sulcus, imply that the majority of CM neurons receive cutaneous afferents from the glabrous skin of the hand. Picard and Smith also found that although some motor cortex neurons with cutaneous receptive fields increased activity with the strength of the grip force, a substantial number of neurons increased their discharge rate in response to abrupt shear forces and slip of a handheld object.

There has been a long-standing debate as to whether somatosensory afferents are transmitted to motor cortex through a separate thalamic relay channel or whether they arrive over corticocortical pathways from the postcentral parietal cortex and the second somatosensory cortex (SII) or both. Recent evidence using transsynaptic virus tracers indicates that thalamic neurons receiving input from the spinothalamic tract do not project to motor cortex. Nevertheless, there is still the possibility that some lemniscal afferents project to motor cortex through the rostral ventral posterior lateral thalamic nucleus. However, it is known that the hand area of the motor cortex of monkeys receives abundant projections from the corresponding regions of areas 1, 2, and 5. These, in addition to projections from SII, are perhaps the most likely source of cutaneous afferents to motor cortex. Moreover, temporary inactivation of the somatosensory cortex by direct injections of the γ-aminobutyric acid agonist muscimol produces clear deficits in hand dexterity that are strikingly similar to the hand dysfunction seen after local anesthesia of the fingers or numbness produced by extreme cold. The cutaneous afferents from the fingertips are particularly important for sensing the direction of pressure vectors required for the dexterous handling of small objects.

Although reports of responses to visual stimuli in the monkey motor cortex have been relatively rare, Wannier and collaborators described such responses unrelated to EMG activity throughout the hand area of area 4. In general, responses to moving visual stimuli are more firmly established neurons in the cat than in the monkey. In contrast, visual motion-related responses have been frequently reported for the hand representation region of the ventral premotor area (PMv). Similarly, auditory evoked responses have been recorded in the cat but not in the hand representation area of the primate.

Force Control and Object Manipulation

The discharge frequency of many motor cortical neurons, including CM cells recorded during static holding, is related to pinch force, or the rate of force increment. However, some neurons actually decrease activity with higher grip force, and these are generally neurons with cutaneous receptive fields. This is not surprising in that the manipulation of a delicate object requires the application of a sufficient grip force to prevent the object from slipping but not so excessive as to either crush the object or expend energy unnecessarily. In addition, the discharge of cutaneous finger-related motor cortical cells increases with the friction or surface roughness of the object against the fingers as well as dynamic changes in the shear force or slip of hand-held objects. During object holding a shear force increase or slip will trigger an increase in motor cortex activity in less than 50 ms on average. As yet the role of motor cortex during tactile exploration has not been examined, but it is likely that the cutaneous motor cortex plays an important role in deploying the contact and sliding forces that optimize the mechanical sensitivity of the fingertip skin as well.

Motor Cortex and Relatively Independent Finger Movements

Although the motor cortex has long been acknowledged to play an important role in the execution of

voluntary independent finger movements, there has been some controversy over the degree of independence in finger movements and whether each finger has its own segregated representation territory within the cortex. This question has been most intensively studied by Scheiber and his colleagues, who trained monkeys to execute flexion and extension movements with each of the five digits. Although thumb flexion and extension was performed without movement of the other digits, the movements of the other four digits were not nearly as autonomous. Single motor cortical neurons were active with movements of each of the different fingers. However, the neurons that discharged during movements of different fingers overlapped extensively, suggesting that finger movement is achieved by a population of neurons distributed rather widely in the hand area rather than by a somatotopically segregated population. This arrangement agrees well with Rathelot and Strick's wide distribution of CM cells within the cortex. Subsequent studies of individual finger movements in patients with well-defined vascular infarcts of the motor cortex or its efferents have demonstrated that movements of the thumb were the least impaired, whereas digits 3, 4, and 5 lost their ability to move separately from one another.

Bimanual Coordination

Many, if not most, skilled hand movements involve the coordinated use of both hands together. Wiesendanger and collaborators have studied a number of bimanual tasks in an attempt to elucidate the particular brain mechanisms responsible for two-handed dexterity. Surprisingly, the primary motor cortices of each hemisphere have few callosal connections between the hand representations although the proximal arm regions are well interconnected. In contrast, the hand areas of the two SMAs are reciprocally connected, and neurons with bimanual-related activity have been found there. Although neurons with activity related to the movement of both arms have been clearly demonstrated, unilateral lesions of the primary motor cortex apparently do not disrupt bimanual coordination as such. Specific deficits in bimanual coordination are more frequently seen in the various forms of apraxia associated with posterior parietal lesions of the nondominant hemisphere.

Cortical Synchronous Activity, Hand Muscles, and Prehension

A number of studies by Lemon and collaborators, in both monkey and humans, have demonstrated synchronous oscillatory activity in motor cortex and in hand muscles during prehensile activity. This oscillatory coherence within a 15–30-Hz band is not confined to the motor cortex and has been reported in both somatosensory cortex and cerebellum. In motor cortex the synchronous activity involves CM cells as well as other corticospinal neurons. The coherence is greater during the steady state holding in a precision grip task than during dynamic grasping or release. What is most intriguing is that the degree of coherence covaries with the degree of compliance of the grasped object and appears to be independent of the grip force or the rate of force change. Coherence was considerably reduced when an isometric grip was exerted on a solid, noncompliant object. At present, the physiological significance of the synchrony is unknown. Nevertheless, maintaining a steady force on a compliant object requires constant monitoring of cutaneous and proprioceptive afferents. Given the spatial segregation of these afferents in the motor cortex mentioned above, the synchronous activity may serve to coordinate and unite their actions on spinal neurons and the muscles of the hand.

Mental Imagery, Cortical Excitability, and Hand Orientation

There seems little doubt that mental imagery, or imagining a hand movement, can increase the excitability of corticospinal neurons. Transcranial magnetic stimulation of the motor cortex has been used to evoke motor-evoked potentials (MEPs) in muscles related to opening or closing the hand as participants were imagining these gestures. The MEPs were larger when an imagined movement involved the muscle as an agonist than when the imagined movement involved the same muscle as an antagonist. Numerous studies have also indicated that the observation of actions done by others (called the mirror neuron system by Rizzolatti and Craighero) activates a number of cortical areas including the motor cortex. How these changes in corticospinal excitability result from observing complex actions of the hand is difficult to assess. However, it appears that observation increases neuronal discharge in the posterior parietal regions and in the PMv that ultimately contributes to increases in corticospinal excitability, as shown by transcranial magnetic stimulation.

It is somewhat easier to understand that the sight of a familiar object evokes a mental image of how it is to be grasped even though the neurons activated in this context are not technically 'mirror neurons.' That is, an object's shape and size dictate the degree of wrist rotation and the shaping of the fingers in

preparation for grasping. Current evidence suggests that the visual preparation for grasping is organized in the posterior parietal area and that the reach-to-grasp commands are relayed through PMv to the motor cortex.

See also: Corticomotoneuronal System; Neural Coding in Primary Motor Cortex; Posterior Parietal Cortex and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape; Premotor Cortex in Primates: Dorsal and Ventral; Premotor Areas: Medial; Reaching and Grasping; Sensorimotor Control of Manipulation.

Further Reading

Baker SN, Spinks R, Jackson A, and Lemon RN (2001) Synchronization in monkey motor cortex during a precision grip task. I. Task-dependent modulation in single-unit synchrony. *Journal of Neurophysiology* 85: 869–885.

Donchin O, De Oliveira SC, and Vaadia E (1999) Who tells one hand what the other is doing: The neurophysiology of bimanual movements. *Neuron* 23: 15–18.

Dum RP and Strick PL (1991) The origin of corticospinal projections from the premotor areas in the frontal lobe. *Journal of Neuroscience* 11: 667–689.

Evarts EV (1968) Relation of pyramidal tract activity to force exerted during voluntary movement. *Journal of Neurophysiology* 31: 14–27.

Fetz EE and Cheney PD (1980) Postspike facilitation of forelimb muscle activity by primate corticomotoneuronal cells. *Journal of Neurophysiology* 44: 751–772.

Georgopoulos AP, Schwartz AB, and Kettner RE (1986) Neuronal population coding of movement direction. *Science* 233: 1416–1419.

Hepp-Reymond M-C, Trouche E, and Wiesendanger M (1974) Effects of unilateral and bilateral pyramidotomy on a conditioned rapid precision grip in monkeys (Macaca fascicularis). *Experimental Brain Research* 21: 519–527.

Kilner JM, Alonso-Alonso M, Fisher R, and Lemon RN (2002) Modulation of synchrony between single motor units during precision grip tasks in humans. *Journal of Physiology* 541: 937–948.

Lawrence DG and Kuypers HGJM (1968) The functional organization of the motor system in the monkey. I. The effects of bilateral pyramidal lesions. *Brain* 91: 1–14.

Lawrence DG and Hopkins DA (1976) The development of motor control in the rhesus monkey: Evidence concerning the role of corticomotoneuronal connections. *Brain* 99: 235–254.

Maier MA, Bennett KMB, Hepp-Reymond M-C, and Lemon RN (1993) Contribution of the monkey corticomotoneuronal system to the control of force in precision grip. *Journal of Neurophysiology* 69: 772–785.

Park MC, Belhaj-Saif A, and Cheney PD (2004) Properties of primary motor cortex output to forelimb muscles in rhesus macaques. *Journal of Neurophysiology* 92: 2968–2984.

Phillips CG (1969) The motor apparatus of the baboon's hand. *Proceedings of the Royal Society of London, Series B. Containing Papers of a Biological Character* 173: 141–174.

Picard N and Smith AM (1992) Primary motor cortical activity related to the weight and texture of grasped objects in the monkey. *Journal of Neurophysiology* 68: 1867–1881.

Poliakov Av and Schieber MH (1999) Limited functional grouping of neurons in the motor cortex hand area during individuated finger movements: A cluster analysis. *Journal of Neurophysiology* 82: 3488–3505.

Rathelot JA and Strick PL (2006) Muscle representation in the macaque motor cortex: An anatomical perspective. *Proceedings of the National Academy of Sciences of the United States of America* 103: 8257–8262.

Rizzolatti G and Craighero L (2004) The mirror–neuron system. *Annual Review of Neuroscience* 27: 169–192.

Strick PL and Preston JB (1982) Two representations of the hand in area 4 of a primate. II. Somatosensory input organization. *Journal of Neurophysiology* 48: 150–159.

Wannier TMJ, Maier MA, and Hepp-Reymond M-C (1989) Responses of motor cortex neurons to visual stimulation in the alert monkey. *Neuroscience Letters* 98: 63–68.

Wiesendanger M and Serrien DJ (2001) Neurological problems affecting hand dexterity. *Brain Research Reviews* 36: 161–168.

Flavor Physiology

E T Rolls, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Taste Processing in the Primate Brain

Pathways

This article describes how taste, olfactory, and somatosensory inputs are brought together in the cortex to represent food flavor, and how these representations are influenced by visual and even cognitive inputs. To make the results relevant to understanding these processes and the control of food intake in humans, complementary evidence is provided by neurophysiological studies in nonhuman primates, and by functional neuroimaging studies in humans.

The taste and related olfactory, somatosensory, and visual pathways in primates are shown in Figure 1. Of particular interest is that in primates there is a direct projection from the rostral part of the nucleus of the solitary tract (NTS) to the taste thalamus and thus to the primary taste cortex in the frontal operculum and adjoining insula, with no pontine taste area and associated subcortical projections as in rodents. This emphasis on cortical processing of taste in primates may be related to the great development of the cerebral cortex in primates, with sophisticated processing of each sensory modality before unimodal representations are brought together in multimodal areas to form multimodal representations of objects, such as food.

The Primary Taste Cortex

The primary taste cortex in the primate anterior insula and adjoining frontal operculum contains not only taste neurons tuned to sweet, salt, bitter, sour, and umami as exemplified by monosodium glutamate, but also other neurons that encode oral somatosensory stimuli, including viscosity, fat texture, temperature, and capsaicin. Some neurons in the primary taste cortex respond to particular combinations of taste and oral texture stimuli, but do not respond to olfactory stimuli or visual stimuli such as the sight of food. Neurons in the primary taste cortex do not represent the reward value of taste, that is, the appetite for a food, in that their firing is not decreased to zero by feeding the taste to satiety. Corresponding areas are activated by taste and oral texture in humans.

The Secondary Taste Cortex

A secondary cortical taste area in primates was discovered in the orbitofrontal cortex (OFC), extending several millimeters in front of the primary taste cortex. In addition to neurons responding to the same taste and oral somatosensory stimuli, including viscosity, fat texture, temperature, and capsaicin, as in the primary taste cortex, some OFC neurons respond to water in the mouth, and others to the somatosensory stimuli astringency as exemplified by tannic acid, and to capsaicin. In the OFC, it has been shown in monkeys that responses of the neurons to the taste of glucose decreased to zero while the monkey ate glucose to satiety, during the course of which the feeding behavior of the monkey changed from avid acceptance to active rejection. These OFC neurons thus have activity related to the reward value of food.

Representation of Flavor: Convergence of Olfactory and Taste Inputs

At some stage in taste processing, it is likely that taste representations are brought together with inputs from different modalities, for example, with olfactory inputs to form a representation of flavor (see Figure 1). In the OFC, some neurons are unimodal for taste, olfactory, or visual inputs, but others respond to combinations of these inputs. Some of the multimodal single neurons had corresponding sensitivities in the two modalities, in that they responded best to sweet tastes (e.g., 1 M glucose) and also responded more in a visual discrimination task to the visual stimulus which signified sweet fruit juice than to that which signified saline, or responded to sweet taste and in an olfactory discrimination task to fruit odor.

Olfactory inputs become associated with taste inputs to represent flavor by olfactory-to-taste association learning. This has been shown by reversing the taste with which an odor is associated, and this has been found to produce slow changes of the odor to which a neuron responds, taking typically 40 trials in primates, as illustrated in Figure 2. Thus flavor, represented by taste–olfactory convergence, is relatively stable. Thus for some neurons in the OFC the neuronal response to the odor depends on the taste with which it is associated and does not depend primarily on the physicochemical structure of the odor. These findings demonstrate directly a coding principle in primate olfaction whereby the responses of some OFC olfactory neurons are modified by, and depend upon, the taste with which the odor is associated. This convergence can be detected in human functional magnetic resonance imaging (fMRI) studies by supralinear responses to a flavor compound (such as sweet taste and strawberry odor), compared to the sum of the components presented separately. Similarly, Small and colleagues also found supradditive interactions between

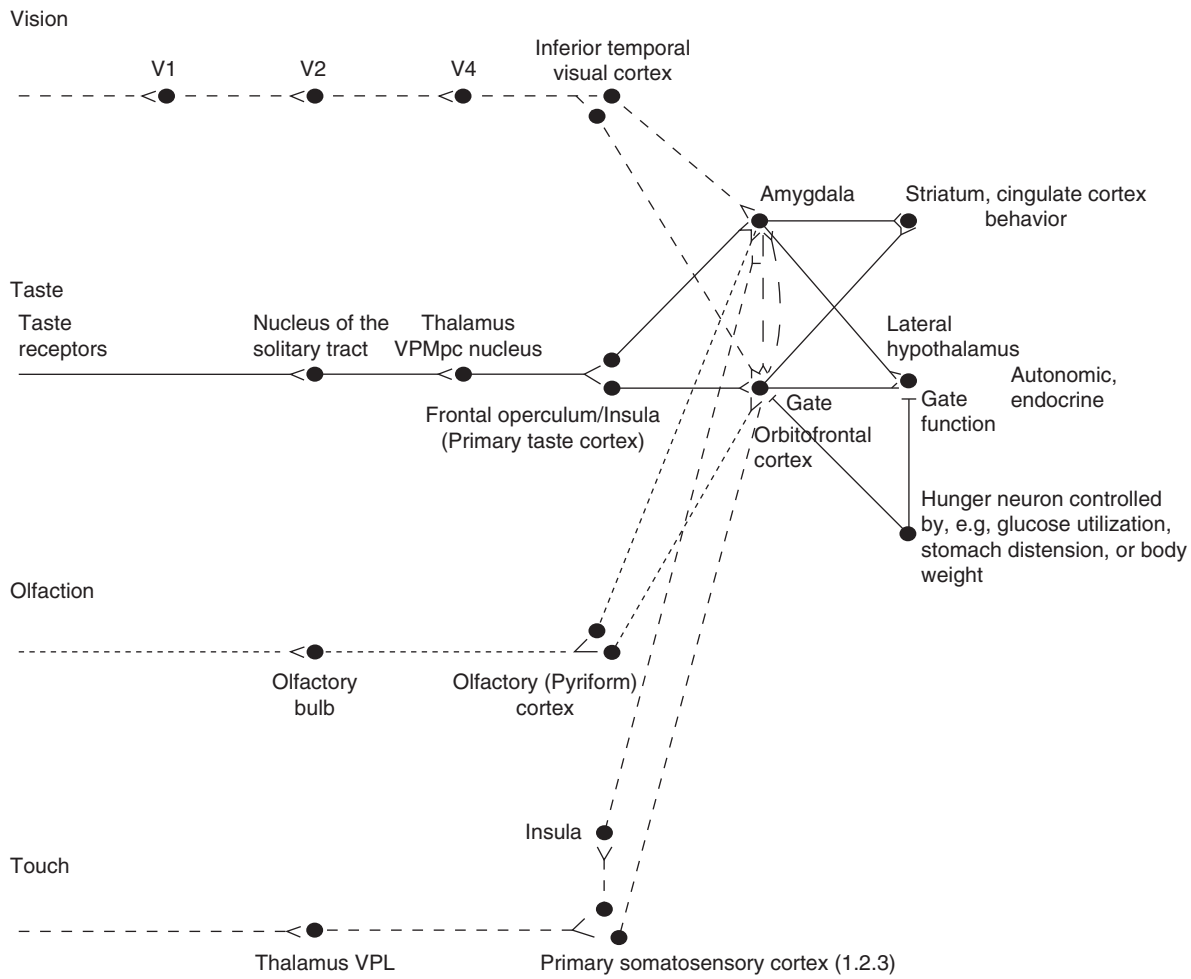

Figure 1 Schematic diagram of the taste and olfactory pathways in primates, showing how they converge with each other and with visual pathways. The gate functions shown refer to the finding that the responses of taste neurons in the orbitofrontal cortex and the lateral hypothalamus are modulated by hunger. VPMpc, ventral posteromedial thalamic nucleus; VPL, ventral posterolateral thalamic nucleus; V1, V2, and V4, visual cortical areas.

congruent taste and smell stimuli in areas including the caudal OFC and anterior cingulate cortex (ACC).

In humans, in addition to activation of the pyriform (olfactory) cortex, there is strong and consistent activation of the OFC by olfactory stimuli. In an investigation of where the pleasantness of olfactory stimuli might be represented in humans, O'Doherty and colleagues showed that the activation of an area of the OFC by banana odor was decreased (relative to a control vanilla odor) after bananas were eaten to satiety. Thus activity in a part of the human OFC olfactory area is related to sensory-specific satiety, and this is one brain region where the pleasantness of odor is represented.

An important issue is whether there are separate regions of the brain, discriminable with fMRI, that represent pleasant and unpleasant odors. To investigate this, we measured the brain activations produced by three pleasant and three unpleasant odors. The pleasant odors chosen were linalyl acetate (floral, sweet), geranyl acetate (floral), and α-ionone (woody, slightly food-related). (Chiral substances were used as racemates.) The unpleasant odors chosen were hexanoic acid, octanol, and isovaleric acid. We found that they activated dissociable parts of the human brain. Pleasant but not unpleasant odors were found to activate a medial region of the rostral OFC. Further, there was a correlation between the subjective pleasantness ratings of the six odors and activation of a medial region of the rostral OFC. In contrast, a correlation between the subjective unpleasantness ratings of the six odors was found in regions of the left and more lateral OFC.

Representation of Flavor: Convergence of Visual Inputs with Taste, Olfactory, and Oral Texture Inputs

Visual stimuli such as the sight of food can produce responses in some OFC neurons that have taste or

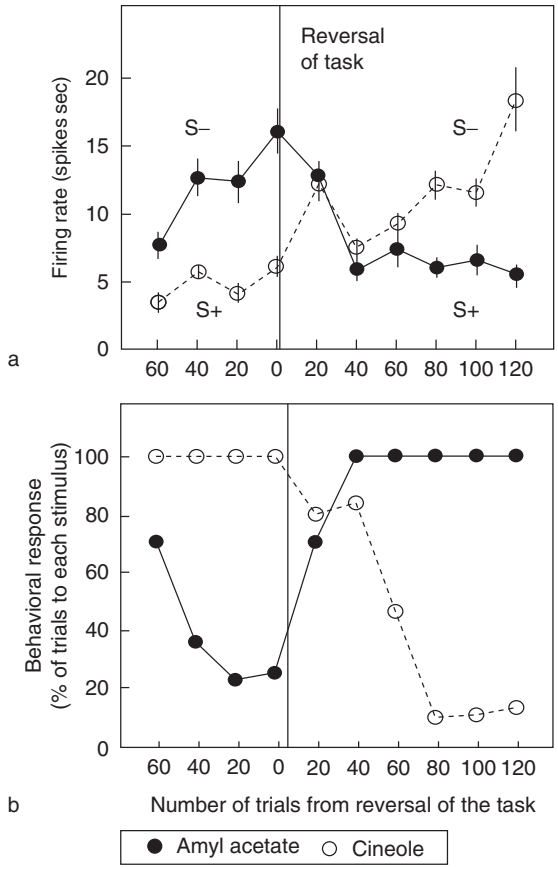

Figure 2 Orbitofrontal cortex: olfactory to taste association reversal. (a) The activity of a single orbitofrontal olfactory neuron during the performance of a two-odor olfactory discrimination task and its reversal is shown. Each point represents the mean poststimulus activity of the neuron in a 500 ms period on approximately 10 trials of the different odorants. The standard errors of these responses are shown. The odorants were amyl acetate (closed circle) (initially S–, that is, associated with the taste of saline) and cineole (open circle) (initially S+, that is, associated with the taste of glucose). After 80 trials of the task, the reward associations of the stimuli were reversed. This neuron reversed its responses to the odorants following the task reversal. (b) The behavioral responses of the monkey during the performance of the olfactory discrimination task. The number of lick responses to each odorant is plotted as a percentage of the number of trials for that odorant in a block of 20 trials of the task. Rolls ET, Critchley H, Mason R, and Wakeman EA (1996) Orbitofrontal cortex neurons: Role in olfactory and visual association learning. *Journal of Neurophysiology* 75: 1970–1981.

olfactory inputs. The visual inputs are associated by learning with the taste inputs to a neuron, and this learning can be very fast, in as little as one trial. This enables prediction of the taste associated with ingestion of what is seen, and thus in the visual selection of foods. This representation also provides a way for the sight of a food to influence the flavor of a food. Oral temperature, which can also affect the palatability of a food, also influences the responses of OFC neurons, and indeed some of these neurons are tuned to oral temperature.

Representation of the Pleasantness and Reward Value of Flavor in the Brain

The responses of OFC neurons to the taste, olfactory, visual, and/or oral texture components of food flavor all show a reduction that is correlated with the decrease in the reward value of food. This reflects probably an influence of satiety signals, including gastric distension, on the responses of these neurons. An additional effect is a sensory-specific reduction in the responses of these neurons after feeding to satiety with a food, and this is likely to be computed in the orbitofrontal cortex, and to be important in which foods are eaten in a meal and in the control of food intake.

To assess how satiety influences the brain activations to a whole food which produces taste, olfactory, and texture stimulation, we measured brain activation by whole foods before and after the food is eaten to satiety. The aim was to show, using a food that has olfactory, taste, and texture components, the extent of the region that shows decreases when the food becomes less pleasant, in order to identify the different brain areas where the pleasantness of the odor, taste, and texture of food is represented. The foods eaten to satiety were either chocolate milk or tomato juice. A decrease in activation by the food eaten to satiety relative to the other food was found in the OFC but not in the primary taste cortex. This study provided evidence that the pleasantness of the flavor of food, and sensory-specific satiety, are represented in the OFC.

Cognitive Effects on Representations of Food

Brie can smell pleasant. However, the same odor taken out of the context of cheese might be unpleasant. There is evidence that the sight (including color) of a food or wine can influence its flavor. However, what about a more cognitive influence, such as a word? Can this influence the perception and hedonics of food-related stimuli, and if so, how far back down into the sensory system does the cognitive influence reach? To address this, we performed an fMRI investigation in which the delivery of a standard test odor (isovaleric acid combined with cheddar cheese flavor, presented orthonasally using an olfactometer) was paired with a descriptor word on a screen, which on different trials was 'cheddar cheese' or 'body odor.' The subjects rated the pleasantness and the intensity of the odor on every trial. α-Ionone (pleasant, labeled 'flowers') and octanol (unpleasant, labeled 'burned plastic') were used as reference pleasant and unpleasant stimuli for the psychophysics and neuroimaging.

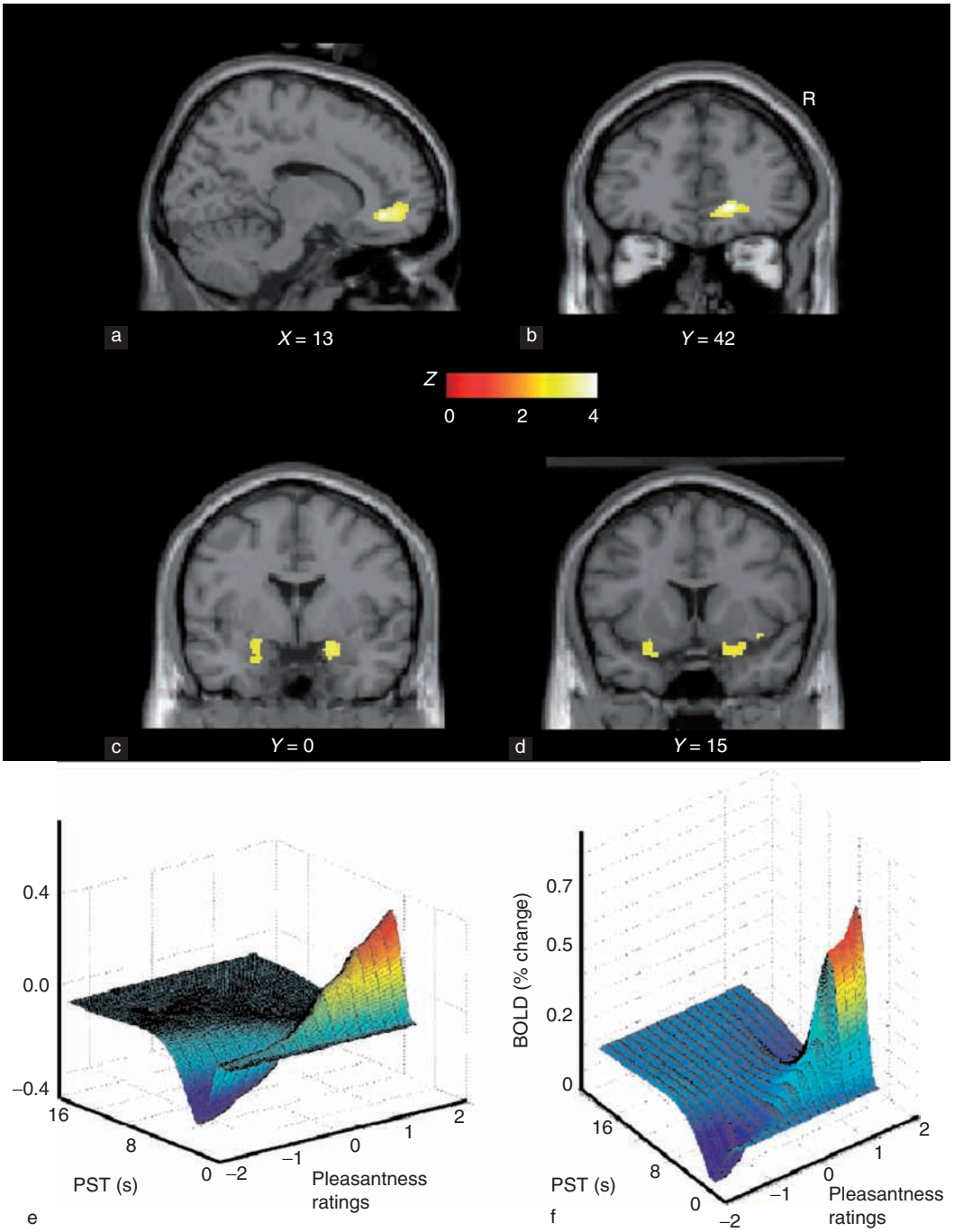

Figure 3 Cognitive influences on olfactory representations in the human brain. Group (random) effects analysis showing the brain regions where the BOLD signal was correlated with pleasantness ratings given to the test odor. The pleasantness ratings were being modulated by the word labels. (a) Activations in the rostral anterior cingulate cortex, in the region adjoining the medial OFC, shown in a sagittal slice. (b) The same activation shown coronally. (c) Bilateral activations in the amygdala. (d) These activations extended anteriorly to the primary olfactory cortex. The image was thresholded at $p < 0.0001$ uncorrected in order to show the extent of the activation. (e) Parametric plots of the data averaged across all subjects, showing that the percentage BOLD change (fitted) correlates with the pleasantness ratings in the region shown in (a) and (b). The parametric plots were very similar for the primary olfactory region shown in (d). PST, poststimulus time (in seconds). (f) Parametric plots for the amygdala region shown in (c). Adapted from de Araujo IET, Rolls ET, Velazco MI, et al. (2005) Cognitive modulation of olfactory processing. *Neuron* 46: 671–679.

Subjects rated the affective value of the test odor as significantly more unpleasant when labeled 'body odor' than when labeled 'cheddar cheese'. We found that the medial OFC/rostral ACC was significantly more activated by the test stimulus labeled 'cheddar cheese' than when labeled 'body odor,' and that these activations were correlated with the pleasantness ratings (see Figure 3). This cognitive modulation was also found in the medial amygdala olfactory area, and this extended toward the olfactory tubercle. Thus cognitive modulation extends into the olfactory system as far down as the secondary olfactory cortex, in the orbitofrontal cortex, and may even influence some parts of the primary olfactory areas, such as the olfactory tubercle. The implication is that cognitive factors can have profound effects on our responses to the hedonic and sensory properties of food, in that these effects are manifest quite far back into sensory processing, so that at least hedonic representations of odors are affected, and even perceptual representations may be modulated.

The Amygdala

These oral sensory properties of food, and also the sight and smell of food, are also represented in the primate amygdala. Interestingly, the responses of these amygdala neurons do not correlate well with the preferences of the macaques for the oral stimuli, and feeding to satiety does not produce the large reduction in the responses of amygdala neurons to food that is typical of OFC neurons. O'Doherty and co-workers showed that the human amygdala was as much activated by the affectively pleasant taste of glucose as by the affectively negative taste of NaCl, and thus provided evidence that the human amygdala is not especially involved in processing aversive as compared to rewarding stimuli.

Conclusions

Food flavors, and the subjective flavor complements, including the consonance of food taste and olfactory components, are represented in primates, including humans, only after several stages of analysis. First the representation of the taste of the food (its identity and intensity) is made explicit in the primary taste cortex. In the pyriform (primary) olfactory cortex, the intensity of an odor, and not its pleasantness or taste, is represented. The OFC is where multimodal representations of food are built, which include taste, texture, olfactory, oral temperature, and visual components.

Although representations of the taste and texture of food are found in the primate (including human) amygdala, the primate OFC is more closely related to the changing affective value of food than the amygdala is, in that the OFC shows responses that decrease to zero as the reward decreases to zero with satiety, and in that the OFC tracks (and probably computes) the changing reward value of stimuli as they are altered by stimulus–reinforcer association learning and reversal.

It is thus becoming possible to start to understand how smell, taste, and other stimuli combine to represent flavor in the brain. Although functional neuroimaging with its spatial resolution of several millimeters shows that some of these representations are close together in the human medial orbitofrontal cortex, the primate neurophysiology provides clear evidence for the exquisitely rich and separate representations of different types of sensory input, because each neuron in this region is tuned to one or to a unique combination of the range of food-related stimuli described here. The information relevant to the control of food intake concerning the sensory properties of foods is thus made explicit in the firing-rate profiles of single neurons and groups of single neurons to the sensory stimuli. It is at the neuron level that brain computations are performed and that the information being transmitted between the computing elements of the brain, the neurons, can be measured by the spiking activity. By making models of the computations taking place in networks of neurons at the spiking neuron level, it is now possible, by integration over the energy required for the activated neurons, to predict signals at a much more global level of functional neuroimaging in humans. By combining information from the neuron and functional imaging level, we are beginning to understand the details of the neuronal operations that underlie the representation of food in the brain, and what makes food pleasant when we are hungry.

The outputs of the OFC reach brain regions such as the striatum, cingulate cortex, and dorsolateral prefrontal cortex, where behavioral responses to food may be elicited because these structures produce behavior that makes the OFC reward neurons fire, as they represent a goal for behavior. At the same time, outputs from the OFC and amygdala, in part via the hypothalamus, may provide for appropriate autonomic and endocrine responses to food to be produced, including the release of hormones such as insulin.

See also: Appetitive Systems: Amygdala and Striatum; Cortical Processing of the Reward Value of Food; Olfactory Receptors; Olfactory Cortex Physiology; Orbitofrontal Cortex: Visual Functions; Sensory Aging: Chemical Senses; Taste: Vertebrate Taste Bud Physiology.

Further Reading

Critchley HD and Rolls ET (1996) Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. *Journal of Neurophysiology* 75: 1673–1686.

Critchley HD and Rolls ET (1996) Olfactory neuronal responses in the primate orbitofrontal cortex: Analysis in an olfactory discrimination task. *Journal of Neurophysiology* 75: 1659–1672.

de Araujo IET, Kringelbach ML, Rolls ET, et al. (2003) The representation of umami taste in the human brain. *Journal of Neurophysiology* 90: 313–319.

de Araujo IET and Rolls ET (2004) The representation in the human brain of food texture and oral fat. *Journal of Neuroscience* 24: 3086–3093.

de Araujo IET, Rolls ET, Kringgelbach ML, et al. (2003) Taste olfactory convergence, and the representation of the pleasantness of flavour, in the human brain. *European Journal of Neuroscience* 18: 2374–2390.

de Araujo IET, Rolls ET, Velazco MI, et al. (2005) Cognitive modulation of olfactory processing. *Neuron* 46: 671–679.

Deco G and Rolls ET (2005) Attention, short-term memory, and action selection: A unifying theory. *Progress in Neurobiology* 76: 236–256.

Kadohisa M, Rolls ET, and Verhagen JV (2004) Orbitofrontal cortex neuronal representation of temperature and capsaicin in the mouth. *Neuroscience* 127: 207–221.

Kadohisa M, Rolls ET, and Verhagen JV (2005) Neuronal representations of stimuli in the mouth: The primate insular taste cortex, orbitofrontal cortex, and amygdala. *Chemical Senses* 30: 401–419.

Kadohisa M, Rolls ET, and Verhagen JV (2005) The primate amygdala: Neuronal representations of the viscosity, fat texture, grittiness and taste of foods. *Neuroscience* 132: 33–48.

Kringelbach ML, O'Doherty J, Rolls ET, et al. (2003) Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. *Cerebral Cortex* 13: 1064–1071.

Norgren R (1984) Central neural mechanisms of taste. In: Darien-Smith I (ed.) *Handbook of Physiology – The Nervous System III. Sensory Processes 1*, pp. 1087–1128. Washington, DC: American Physiological Society.

O'Doherty J, Rolls ET, Francis S, et al. (2001) The representation of pleasant and aversive taste in the human brain. *Journal of Neurophysiology* 85: 1315–1321.

Rolls ET (2005) *Emotion Explained*. Oxford: Oxford University Press.

Rolls ET and Baylis LL (1994) Gustatory, olfactory, and visual convergence within the primate orbitofrontal cortex. *Journal of Neuroscience* 14: 5437–5452.

Rolls ET, Critchley H, Mason R, and Wakeman EA (1996) Orbitofrontal cortex neurons: Role in olfactory and visual association learning. *Journal of Neurophysiology* 75: 1970–1981.

Rolls ET, Critchley H, Wakeman EA, et al. (1996) Responses of neurons in the primate taste cortex to the glutamate ion and to inosine 5′-monophosphate. *Physiology and Behaviour* 59: 991–1000.

Rolls ET, Kringelbach ML, and de Araujo IET (2003) Different representations of pleasant and unpleasant odors in the human brain. *European Journal of Neuroscience* 18: 695–703.

Rolls ET and Scott TR (2003) Central taste anatomy and neurophysiology. In: Doty RL (ed.) *Handbook of Olfaction and Gustation,* 2nd edn., pp. 679–705. New York: Dekker.

Rolls ET, Sienkiewicz ZJ, and Yaxley S (1989) Hunger modulates the responses to gustatory stimuli of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. *European Journal of Neuroscience* 1: 53–60.

Rolls ET, Yaxley S, and Sienkiewicz ZJ (1990) Gustatory responses of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. *Journal of Neurophysiology* 64: 1055–1066.

Small DM and Prescott J (2005) Odor/taste integration and the perception of flavor. *Experimental Brain Research* 166: 345–357.

Verhagen JV, Kadohisa M, and Rolls ET (2004) The primate insular/opercular taste cortex: Neuronal representations of the viscosity, fat texture, grittiness and taste of foods in the mouth. *Journal of Neurophysiology* 92: 1685–1699.

Floor Plate Patterning of Ventral Cell Types: Ventral Patterning

E Dessaud, J Briscoe, and K Hill, National Institute for Medical Research, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Interest in understanding the patterning of the vertebrate neural tube stems from its complex yet highly organized structure and its central role in the life and behavior of animals. Here we concentrate on the strategies and molecular mechanisms by which neuronal subtype identity is assigned in ventral regions of the spinal cord – the region of the central nervous system (CNS) that contains the neurons which control and coordinate motor output. At early stages of embryonic development, five distinct subtypes of neurons can be molecularly identified in the ventral spinal cord; among these are four types of interneurons (V0–V3) and somatic motor neurons (MNs). Each neuronal population arises from blocks of proliferating progenitors that occupy stereotyped locations along the dorsal/ventral axis of the neural tube (Figure 1). These domains of progenitor cells can be distinguished by their expression of distinct molecular markers, including a set of transcription factors, notably homeodomain-containing (HD) proteins and basic helix–loop–helix (bHLH) proteins. The combinatorial expression profile of these proteins delineates five populations of progenitor cells, each of which generates one of the identified neuronal subtypes; consequently, the early spinal cord develops as a dorsal/ventral array of neuronal subtypes. We review the mechanisms involved in the establishment of this pattern of neurogenesis; in particular, we describe the function of the Sonic hedgehog (Shh) signaling pathway, which appears to act as a graded morphogen in the ventral neural tube to control pattern formation.

Evidence That Notochord-Secreted Signals Pattern the Dorsoventral Axis of the Neural Tube

Studies initiated in the 1920s suggested that, rather than containing intrinsic self-organizing capacity, the major determinant of dorsoventral (D/V) polarity is the mesoderm surrounding and underlying the neural tube. From these studies, the notochord was identified as one of the mesodermal structures from which instructive signal(s) emanate that direct D/V patterning of the developing neural tube (Figure 2). In the absence of the notochord, floor plate cells – the characteristic wedge-shaped cells that are generated at the ventral midline of the neural tube – were missing, whereas in embryos containing two notochords a duplication of floor-plate-like structures was evident.

The same principles were reinforced in subsequent experiments using chick embryos, from the 1950s onward. Moreover, these experiments indicated that contact between the notochord and neural tube was essential for floor plate formation. Ensuing experiments in which extra notochords were grafted into chick embryos confirmed the inductive effect on the floor plate, and added an extra dimension; neuroblasts (presumptive motor neurons) were also induced by the notochord but at a distance from the position of the graft (Figure 2(b)). In addition, the floor plate was found to display similar inductive properties to the notochord, indicating that the floor plate also has capacity to promote the generation of both more floor plate cells and MNs. These experiments led to the conclusion that the notochord and floor plate are the source of a signal(s) that induced different cell types and that the distance from the source of this signal(s) determines the cell type generated.

Shh Patterning of the Neural Tube

We now know a considerable amount about the molecules involved in D/V patterning and the understanding of their action relies to a significant degree on the foundation provided by these early embryological experiments. In the search for the signal providing instructive cues to the ventral neural tube, a secreted protein, Sonic hedgehog, was identified. This protein is an ortholog of the *Drosophila* segment polarity gene product, Hedgehog (Hh), a signaling molecule involved in the development of several embryonic tissues. Consistent with this, Shh is expressed in many developing tissues, including the notochord and floor plate. The period when Shh is expressed in these structures corresponds to the time when neural tube polarity is being established. Since the discovery of the Hh family of signaling molecules, much effort has been employed in identifying the components of the signaling pathway, and an understanding of the mechanism of signal transduction is beginning to emerge. Transduction depends on two transmembrane proteins: Patched 1 (Ptc1), the receptor which binds Hh proteins, and Smoothened (Smo), which is responsible for transmitting Hh signals intracellularly (see later).

A range of loss-of-function experiments support the idea that Shh corresponds to the signal that

provides instructive cues to control the generation and arrangement of distinct neuronal populations in the neural tube. In mouse embryos harboring a targeted null allele of Shh, there is a loss of MNs and other ventral interneuron subtypes; concomitantly molecular markers normally restricted to dorsal aspects of the spinal cord expand toward the ventral midline. Furthermore, antibodies raised against Shh that block its ability to bind Ptc1 also disrupt patterning of the ventral neural tube and inhibit the ability of notochord to induce ventral neuronal subtypes. Together, these studies suggest that production of Shh is necessary for correct formation of ventral neuronal subtypes *in vivo*.

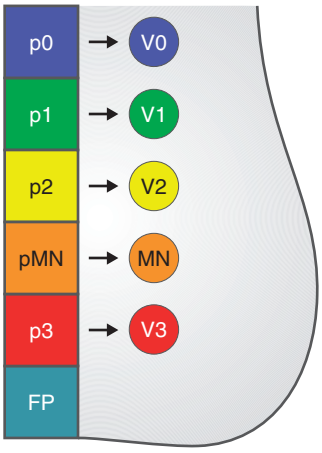

Figure 1 Organization of the ventral neural tube: progenitor domains and ventral neuronal cell types and their fates in the spinal cord (schematic of the ventral half of the neural tube). Five progenitor domains are found in stereotyped positions within the ventricular zone of the ventral neural tube, lying dorsal to the floor plate (FP). These progenitor domains – p0, p1, p2, p3, and pMN – will generate the corresponding neuronal cell types – V0 to V3 interneurons and motor neurons (MN) – involved in motor output control.

Experiments carried out *in vitro* using explants of naive neural tissue provided further understanding of the mechanism of Shh activity. These studies indicated that exposure to varying concentrations of Shh protein induced the populations of neurons characteristic of the ventral neural tube. Incremental two- to threefold changes in Shh concentration led to the generation of different neuronal subtypes; moreover, a correlation between the *in vitro* concentration of Shh necessary to induce each neuronal class and their position of generation *in vivo* was apparent (Figure 3). Thus, neurons generated in progressively more ventral regions of the neural tube required correspondingly higher Shh concentrations for their induction. These data indicated that in addition to being necessary for their development, Shh is sufficient to induce ventral interneurons and motor neurons in neural tissue, suggesting that Shh functions in a graded manner to impart positional identity. Thus, cells in the ventral neural tube are exposed to a ventral^{HIGH}–dorsal^{LOW} concentration gradient of Shh, and the concentration of Shh to which a progenitor cells is exposed determines the neuronal subtype generated (Figure 3).

A key prediction of this model is that Shh should operate as a long-range signal, acting directly on cells at a distance from its source. Several lines of evidence indicate that this is indeed the case. First, Shh protein is detectable several cell diameters away from the producing cells of the floor plate (Figure 3). Second, a dominant active version of Smo, a component of the signaling pathway, is sufficient to induce ectopic expression of ventral markers in the neural tube. Third, blockade of Shh signaling, by either cell autonomous removal of Smo or ectopic expression of a mutated version of Ptc1 (the Shh receptor) that acts as a dominant inhibitor of signaling, blocks the generation of ventral neurons, which instead develop

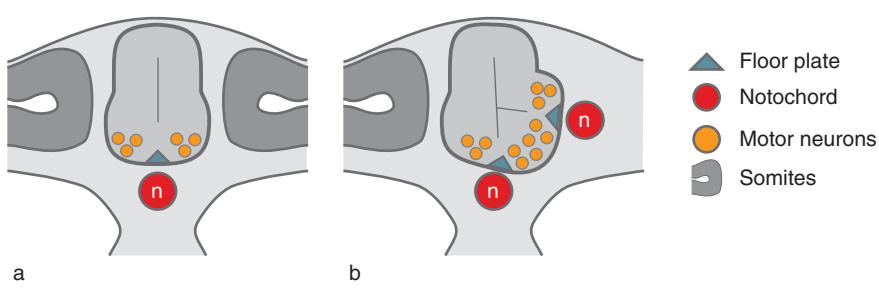

Figure 2 Notochord-secreted signal(s) influence patterning of the neural tube. (a) The notochord (n) is an axial mesodermal structure that underlies the caudal neural tube. The floor plate is formed in the ventral neural tube, close to the notochord, and motor neurons are generated in the ventral horns of the spinal cord, at a distance from the notochord. (b) Grafting an ectopic notochord next to the neural tube induces an ectopic floor plate at a short distance, whereas at a greater distance excess motor neurons are generated. These experiments suggest the existence of an instructive secreted signal(s) from the notochord that can control the patterning of the ventral neural tube.

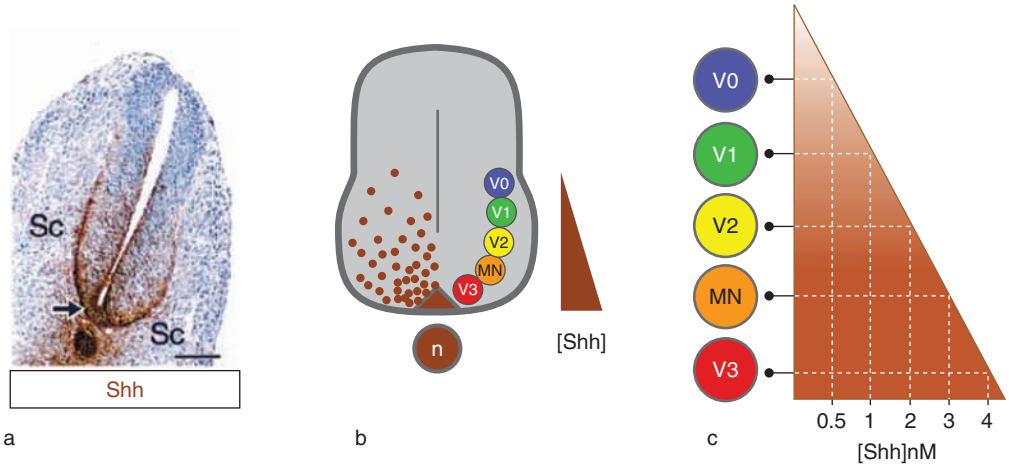

Figure 3 Sonic hedgehog (Shh) diffuses in the ventral neural tube and controls neuronal cell fate in a concentration-dependent manner. (a) Staining for Shh in E10.5 mouse embryo. Shh protein can diffuse through the ventral half of the neural tube to establish a ventral-to-dorsal gradient (high levels of protein are detected in the Shh-secreting floor plate and lower levels in the intermediate neural tube). (b) Schematic of Shh distribution, illustrating the graded distribution of the protein in the developing neural tube, together with the relative position of newborn ventral neuronal cell types (interneurons, V0–V3; motor neurons, MN). (c) Naive neural plate chick explants can generate the five ventral neuronal cell types, depending on exposure to Shh. Two- to threefold increments in Shh concentration allow the induction of ventral interneurons and motor neurons, suggesting that Shh protein is acting as a morphogen, patterning the ventral neural tube. (a) Reproduced with permission from Gritli-Linde A, Lewis P, McMahon AP, et al. (2001) The whereabouts of a morphogen: Direct evidence for short- and long-range activity of hedgehog signaling peptides. *Developmental Biology* 236: 364–386.

with characteristics of neurons located in the dorsal neural tube. Together, these results suggest that Shh acts directly on cells at a distance through its receptors, rather than indirectly through a relay signal.

A Transcription Factor Code and Neural Cell Fate

The ability of graded Shh signaling to directly control D/V patterning in the ventral neural tube raised the question of how positional identity is imposed on progenitor cells and how this determines neuronal subtype identity. Progenitors respond to Shh signaling by regulating the expression of a series of transcription factors, which, with the exception of the bHLH protein Olig2, are members of the homeodomain protein family, including Pax7, Pax3, Dbx1, Dbx2, Pax6, Nkx2.2, Nkx2.9, Nkx6.1, Nkx6.2, and Irx3. These transcription factors play a crucial part in patterning the ventral neural tube and can be split into two groups according to their response to Shh: class I, the genes of which are repressed by Shh signaling, and class II, the genes of which are activated by Shh (Figure 4(a)). Furthermore, within the same class, genes respond in a concentration-dependent manner to Shh, such that the repression of class I genes with more ventral limits of expression requires higher levels of Shh signaling compared to more dorsally restricted class I genes. Conversely, class II proteins that have a broad domain of expression in the ventral neural tube are induced by lower concentrations of Shh than are class II proteins that have a more ventral limit of expression. In addition, class I and class II genes can be grouped in pairs based on two prerequisites: (1) the ventral limit of expression of a class I protein coincides with the dorsal limit of the paired class II protein and (2) the pair of class I and class II factors display mutual cross-repression.

Graded Shh Controls Class I and Class II Protein Expression

Evidence that graded Shh signaling establishes the transcription factor code emerged from studies of the expression patterns of Nkx2.2 and Pax6. In the developing ventral neural tube, the class I protein Pax6 and the class II protein Nkx2.2 have a mutually exclusive expression pattern, with the most ventral progenitors expressing Nkx2.2, while Pax6 is expressed in progenitors situated dorsal to the boundary of Nkx2.2 (Figure 4(a)). *In vitro* experiments demonstrated that the expression of both of these proteins is regulated by Shh signaling in a concentration-dependent manner. In the absence of Shh, explants express Pax6; exposure to concentrations of Shh above a given threshold results in the repression of Pax6 and the concomitant induction of Nkx2.2. Importantly, the mutually exclusive relationship between Pax6 and Nkx2.2 is

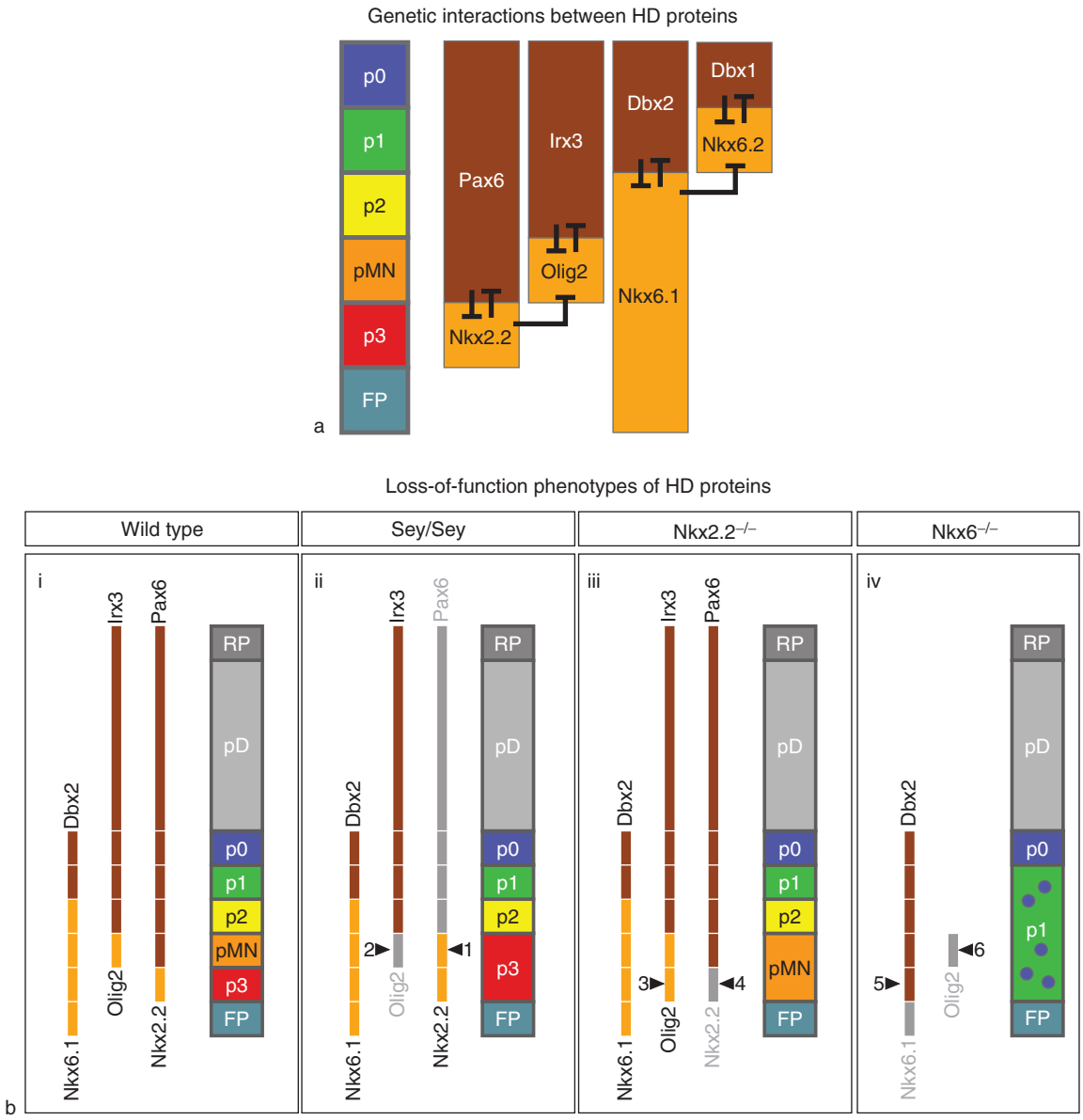

Figure 4 A transcription factor code and patterning of the ventral neural tube. (a) Combinatorial expression of homeodomain-containing (HD) proteins Pax, Irx3, Nkx, and Dbx, together with the basic helix–loop–helix transcription factor Olig2, distinguishes five ventral progenitor domains in the caudal neural tube. HD proteins can be subdivided into two groups according to their response to Sonic hedgehog (Shh) signaling, class I (brown) and class II (orange) proteins, which are inhibited or activated by Shh, respectively. Cross-repression between selected pairs of proteins controls the formation of boundaries between progenitor domains. (b) Cross-repression between pairs of HD proteins controls ventral neural tube patterning. Compared with wild-type embryos (i), in the absence of Pax6 (Sey/Sey; ii), there is a dorsal expansion in Nkx2.2 expression (1) that leads to the inhibition of Olig2 expression (2) and results in the absence of the motor neuron progenitor (pMN) domain and a concomitant expansion of the p3 domain. Conversely, in the absence of Nkx2.2 (iii), there is a ventral expansion of Olig2 (3), but not of Pax6 (4; see text for more details), leading to an expansion of pMN at the expense of p3. Another example of a pair of HD proteins is illustrated with Nkx6.1 and Dbx2 (iv). In double mutants lacking Nkx6.1 and Nkx6.2 (Nkx6$^{-/-}$), there is a ventral expansion of Dbx2 (5) and a downregulation of Olig2 (6). As a consequence, p2, pMN, and p3 domains are lost and p1 progenitors occupy most of the ventral spinal cord. There is also a sporadic expression of Dbx1 (blue circles) in this ectopic p1 domain due to the loss of inhibition by Nkx6.2 (RP, roof plate; pD, dorsal spinal cord domain). Expression patterns of Irx3, Pax6, and Nkx2.2 in these mutants have not been determined.

maintained in these experiments in a cell autonomous manner. Furthermore, addition of anti-Shh antibodies promotes Pax6 expression and inhibits Nkx2.2 expression, even in the presence of Shh. Similar findings have been documented for other class I and class II proteins. Additional signaling molecules have also been implicated in modulating the expression of the progenitor-expressed

transcription factors (see later) and these may act by either influencing the Shh signal transduction pathway or by directly regulating target gene expression.

Cross-Repression between Class I and Class II Proteins Establishes the Mutually Exclusive Expression Profiles

The regulation of class I and class II protein expression by graded Shh signaling is not alone sufficient to explain how the discrete, sharp boundaries of gene expression arise in the neural tube. Gain- and loss-of-function studies indicate that selective cross-repressive interactions between pairs of class I and class II proteins expressed in adjacent domains are important for establishing the all-or-none changes in gene expression that characterize progenitor domain boundaries. This first became apparent for Nkx2.2 and Pax6. Gain-of-function experiments indicate a cross-repressive interaction between Pax6 and Nkx2.2. Moreover, expression of Nkx2.2 expands dorsally in mutant mice lacking Pax6 (Figure 4(b)). This results in a decrease in MN production coupled with an expansion in Nkx2.2-produced V3 neurons. Pax6 is therefore required for accurate positioning of the boundary between MN and V3 progenitors, and for correct production of the neurons from these progenitors. The absence of Pax6 also results in defects in the generation of other ventral neuronal subtypes, in particular V1 and V2 neurons, suggesting that Pax6 has further roles in the control of neuronal subtype identity in the neural tube. Additional studies indicated that in embryos lacking Nkx2.2 there was a loss of V3 neuronal subtype identity, and a ventral shift in the MN progenitor marker Olig2 and subsequently in the generation of MNs (Figure 4(b)). Pax6 expression remained unchanged, however, possibly due to the presence of Nkx2.9, another Nkx family member, which is expressed in a similar domain to Nkx2.2.

Evidence from a second set of class I and class II proteins lends support to the idea of reciprocal cross-repressive interactions. The class II proteins Nkx6.1 and Nkx6.2 adjoin the ventral boundaries of the class I proteins Dbx2 and Dbx1, respectively (Figure 4(a)). In double-knockout embryos lacking both Nkx6 proteins, a ventral expansion of Dbx2 expression was evident (Figure 4(b)). Furthermore, forced expression of Nkx6.1 cell autonomously repressed Dbx2. Conversely misexpression of Dbx2 resulted in the down-regulation of Nkx6.1. A similar relationship between a third pair of proteins, Olig2 and Irx3, is also apparent (Figure 4(a)). With the exception of Pax6, the class I and class II proteins act as transcriptional repressors in neural progenitors, raising the possibility of direct interactions between the proteins and promoters of class I and class II genes.

Control of Neuronal Identity by Graded Shh Signaling

The regulation of class I and class II proteins ultimately leads to the establishment of the progenitor domains p0, p1, p2, pMN, and p3, defined by the expression of distinct combinations of transcription factors, which generate the five neuronal subtypes (Figure 4(a)). Gain- and loss-of-function experiments support the hypothesis that the progenitor transcription factor expression profile specifies the identity of neurons generated from each progenitor domain. The forced expression of a class I or class II protein in the neural tube and targeted inactivation of individual class I or class II proteins change the fate and position of generation of individual neuronal subtypes in a manner predicted by the normal profile of transcription factor expression.

Overall, these studies suggest a model to explain ventral patterning of the neural tube. Crucially, the initial activation or repression of the class I and class II proteins by graded Shh signaling imparts positional identity to progenitor cells. The reciprocal repression between pairs of class I and class II proteins provides a mechanism to convert the gradient of Shh signaling into discrete all-or-none changes in gene expression that may account for the formation of sharp boundaries between adjacent progenitor domains. Each domain then generates a specific neuronal subtype. The repressive interactions between progenitor expressed transcription factors could also serve to consolidate progenitor domain identity, relieving a requirement for a prolonged period of Shh signaling to maintain gene expression domains. Significantly, *Drosophila* orthologs of several of the class I and class II proteins have been demonstrated to play a similar role in allocating regional identity in ventral nerve cord of the developing embryo, raising the possibility that this strategy represents a fundamental feature of nervous system patterning of many extant animals.

Additional Signals Involved in Dorsoventral Patterning of the Neural Tube

Although the evidence confirms the importance of Shh signaling in controlling spatial patterning within the neural tube, other extracellular signals also contribute to correct D/V patterning. During neural induction, progenitors are exposed to fibroblast growth factors (FGFs), originating from the presomitic mesoderm, the regressing node, and the neural plate. FGF signaling acts as an inhibitor of neural differentiation, ensuring cells remain progenitors. FGFs also inhibit the expression of many of the

progenitor transcription factors and it is not until cells have emerged from the influence of FGFs that they mature and begin to express class I and class II proteins. Conversely, retinoic acid (RA) is expressed anteriorly to FGF, in paraxial mesoderm adjacent to the neural tube, and RA is necessary for neural differentiation and progenitor transcription factor expression. RA promotes class I protein expression, therefore counteracting the ventralizing effects of Shh. A third signaling pathway is also implicated in D/V patterning – bone morphogenetic proteins (BMPs). Several members of the BMP family are expressed in the dorsal pole of the neural tube, where they play a role in specifying dorsal neuronal fates. In addition, BMP proteins appear to oppose Shh-mediated ventralization of the neural tube, limiting the dorsal extent of the ventral neural tube. Hence, precise coordination of ventral neural patterning depends on the complex interplay between several signaling pathways, the details of which are only now beginning to be understood.

Formation and Maintenance of Shh Gradient

The central role of graded Shh signaling in the control of D/V patterning in the neural tube has focused attention on how the production and distribution of Shh are regulated. For this, knowledge of the concentration and shape of the gradient is required; parameters such as the rate at which Shh is released from its source, the speed of its spread, and its stability are necessary but difficult to obtain in embryos *in vivo*. Additional complexity to this problem is introduced by the attachment of lipid moieties to Shh and the presence of extracellular proteins that bind to Shh.

Posttranslational Modification and Release of Hedgehog Proteins

The Hh precursor protein is posttranslationally processed to a mature Hh protein by an autocatalytic cleavage that removes the carboxyl (C)-terminus. As part of this reaction, a cholesterol moiety is covalently attached to the C-terminal end of the processed protein. Shh is also acylated on its amino (N)-terminus, for example, by addition of palmitate; the end result is a Shh protein linked to two hydrophobic groups. Although these posttranslational modifications are not absolutely required for Hh activity, as illustrated by the induction of several Hh-target genes with recombinant forms of Shh that lack lipid moieties, the lipid modifications appear to affect the spread and/or activity of Shh in neural tissue. The secretion of cholesterol modified Hh appears to require the transmembrane protein, Dispatched-1 (Disp1), as mutations in this gene result in alterations in neural tube patterning consistent with a lack of Shh released. Palmitoylation of Shh also appears essential in vertebrate neural tissue. Prevention of palmitate addition to Shh protein by mutation of the target cysteine (Shh^{C25S}) leads to a severe deficiency in ventral neural tube patterning, reminiscent of embryos lacking Shh protein. Similar effects are also observed in mice deficient in Skn, the murine ortholog of the product of the *Drosophila ski* gene, which catalyzes Hh palmitoylation. Thus, the hydrophobic modifications of Shh have important and complex effects on Shh function *in vivo*.

Hedgehog-Binding Proteins and Feedback Loops

In addition to lipidation, Hh-binding proteins also modify Hh activity. One important regulatory strategy used by many signaling systems is negative feedback, where signal transduction leads to upregulation of inhibitors of the pathway to attenuate signaling. Hh uses such feedback regulation and inhibition is mediated by several Hh-binding proteins. Among these are Hh receptor Ptc1 and an additional protein, Hh-interacting protein 1 (Hhip1). Both Ptc1 and Hhip1 are upregulated by Shh signaling in the ventral neural tube (**Figures 5(a) and 5(b)**). Evidence that upregulation of Ptc may be involved in limiting the spread of Hh proteins initially came to light in experiments in *Drosophila* which demonstrated an extended range of Hh in the absence of Ptc. Similar observations were later made in the vertebrate neural tube, where a dorsal-to-ventral change in neural fate was noted in cells dorsal to clusters of cells unable to respond to Hh. Elegant studies in mouse subsequently documented the role Ptc1- and Hhip1-mediated Shh sequestration plays in patterning the ventral neural tube (**Figure 5(c)**). In these experiments, a transgenic mouse strain expressing a low basal level of Ptc1 from a heterologous promoter (MtPtc1) was used. This was sufficient to suppress unregulated Smo activity; however, in the absence of endogenous Ptc1 and Hhip1 genes, MtPtc1 failed to restrain the range of Shh signaling, resulting in a ventralization of the neural tube (**Figure 5(c)**). Thus, the upregulation of Ptc1 and Hhip1 in response to Shh signaling normally acts to limit the range over which the ligand can diffuse (**Figure 5(b)**). The appearance of intermediate phenotypes in compound heterozygote mutants for both Ptc1 and Hhip1 indicates that this mechanism utilizes both Ptc1 and Hhip1 in a dose-dependent manner.

Two further transmembrane Hh-binding proteins, Cdo and Boc, have been identified as both targets and regulators of Shh signaling. Both are expressed in the dorsal spinal cord – a domain with low level of Ptc1

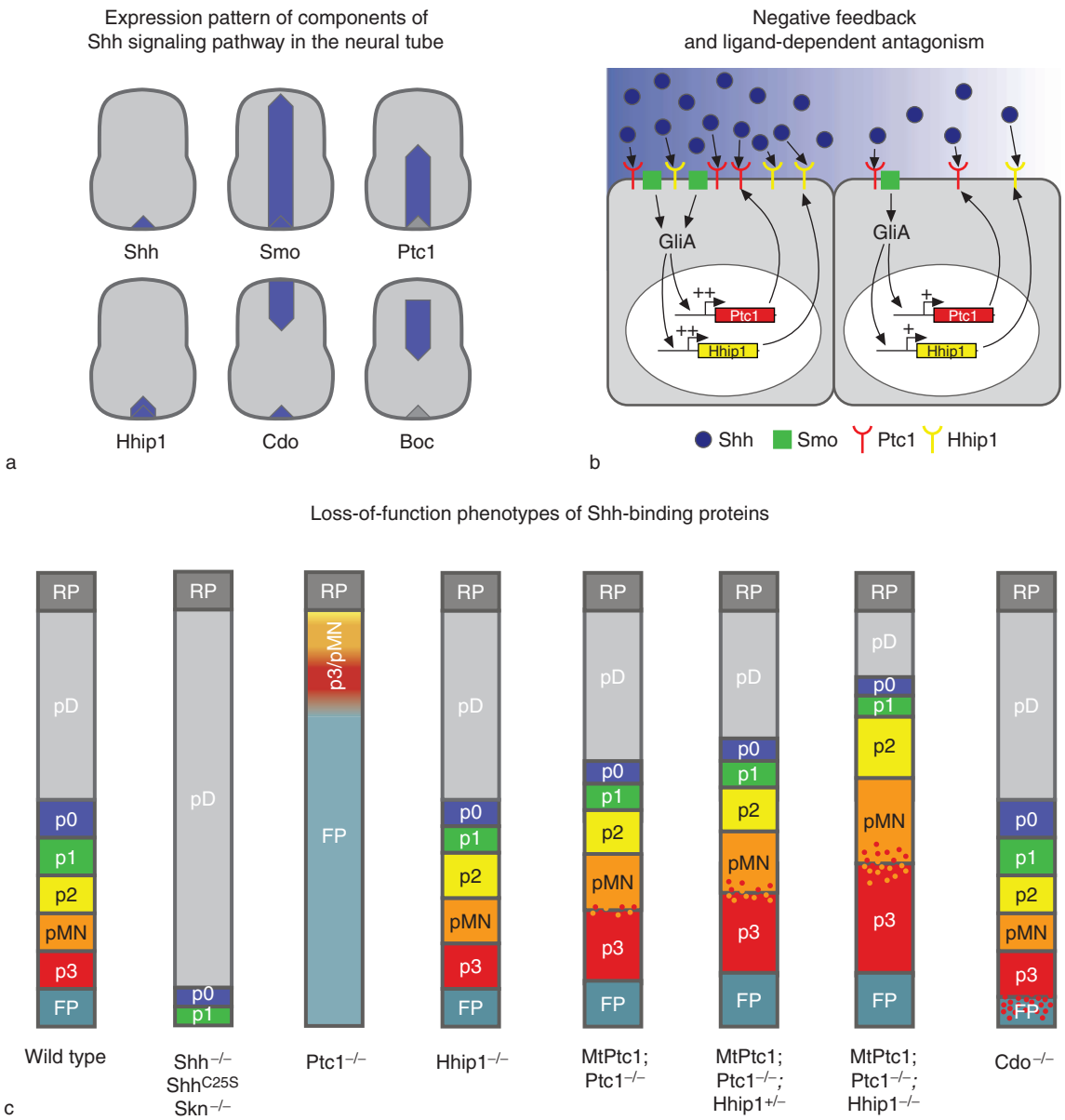

Figure 5 Regulation of the Sonic hedgehog (Shh) gradient is important for dorsoventral patterning of the neural tube. (a) Shh is expressed by the notochord (not shown) and the floor plate. Shh-binding proteins are found along the dorsoventral axis of the neural tube; Smoothened (Smo) and Patched1 (Ptc1) are expressed in a ventral-to-dorsal gradient in the spinal cord, together with Hhip1, found in the ventralmost part of the spinal cord. Cdo and Boc are expressed in the dorsal spinal cord; in addition, Cdo is also expressed in the floor plate. (b) Schematic of the control of the Shh gradient by the negative feedback loop triggered by Shh-binding proteins Ptc1 and Hhip1. Shh signaling enhances the expression of its receptors Ptc1 and Hhip1 through Gli transcriptional activators (GliA). The more Shh signaling received by a cell, the more it will increase the expression of Ptc1 and Hhip1. As a result, cells located near the source of Shh express higher levels of Ptc1 and Hhip1 (and also GliA) compared to cells located far away from the Shh source. The two Shh-binding proteins sequester Shh, thus limiting the range of Shh spread. (c) Abrogating the production and posttranslational modifications of Shh results in a dramatic dorsalization of the neural tube as observed in Shh$^{-/-}$ mice or mutants lacking Shh palmitoylation (Shh^{C25S} and Skn$^{-/-}$). In the absence of Ptc1, there is high Shh activity along the dorsoventral axis due to the lack of Smo inhibition. This is not observed in Hhip1$^{-/-}$ embryos. When Ptc1 is ubiquitously expressed at low level under the control of the metallothionein promoter (MtPtc1) and Ptc1/Hhip1 ligand-dependent antagonism is abolished, the distance of Shh spread is increased and consequently ventralization of the spinal cord results. This phenotype is dependent on the quantity of Hhip1, as illustrated by the comparison between MtPtc1;Ptc1$^{-/-}$ with MtPtc1;Ptc1$^{-/-}$;Hhip$^{+/-}$ and MtPtc1;Ptc1$^{-/-}$;Hhip$^{-/-}$. In parallel, this perturbation in Shh gradient decreases the precision in ventral patterning, as the boundary between p3 and pMN is less precise. There is also a lack of precision in the boundary between the floor plate (FP) and the p3 domain in the absence of Cdo, together with a reduction of the FP and a ventral expansion of the p3 domain. RP, roof plate; pD, dorsal spinal cord domain; pMN, motor neuron progenitor.

and Hhip1 – and Cdo is also expressed in the floor plate (Figure 5(a)). Gain- and loss-of-function experiments indicate that Cdo and Boc cell autonomously enhance Hh signaling. Given their expression profiles, it is possible that Cdo increases Shh signaling where the highest Shh signaling levels are required (e.g., in the floor plate). Conversely, at the fringes of the Hh target field (e.g., in intermediate and dorsal regions of the neural tube), where Ptc1 and Hhip1 levels are expected to be low, Cdo and Boc may sensitize cells to even low levels of Shh protein. Ectopic expression of Cdo and Boc led to a non-cell autonomous expansion of dorsal genes in cells lying dorsal to expressing cells, consistent with the idea that these proteins bind Shh, impeding its ability to spread further through tissue. Together, these data begin to indicate that the extracellular spread of Shh is the subject of finely controlled regulation which establishes the appropriate distribution of Shh that ensures the accurate supply of positional information to progenitors residing in the ventral neural tube.

Intracellular Mediators of Shh-Dependent Neural Tube Patterning: A Major Role of Gli Proteins

In receiving cells, intracellular signal transmission depends on two transmembrane proteins: Patched 1 (Ptc1), as already mentioned, the receptor which binds Hh proteins, and Smoothened (Smo), which is responsible for transducing Hh signals intracellularly (Figure 6(a)). In the absence of Shh, Ptc1 inhibits Smo activity, and binding of Shh to Ptc1 releases this inhibition, allowing intracellular signal transduction. Although the exact mechanism of signal transmission downstream of Smo is still the subject of study, several key components of the pathway have been identified. In vertebrates, Suppressor of Fused (Sufu) has a central role in Shh signaling, as gene disruption of Sufu in mice leads to potent ligand-independent activation of the Hh pathway, mimicking the defect in neural tube patterning observed in Ptc1 mutant embryos. Several other factors, including Rab23 (a vesicle transport family protein) and a number of factors involved in the assembly and/or function of a cell's primary cilium, have also been implicated in Shh signal transduction.

Gli Proteins and Dorsoventral Patterning

Although the mechanism of signal transduction remains to be clarified, the evidence suggests that the signal culminates in the regulation of a family of zinc finger-containing transcriptional effectors known, in vertebrates, as the Gli transcriptional regulators (Gli1, 2, and 3). These control target gene expression. All three Gli genes are expressed in the neural tube and several studies have begun to examine the regulation of Gli gene activity and the roles they play in neural tube patterning. Gli1 is expressed in the ventralmost neural tube, and its expression is dependent on Shh signaling. Gli1 appears to act solely as a transcriptional activator. In contrast, Gli2 and Gli3 are expressed in neural tissue prior to neural tube closure; later on, the expression pattern of Gli3 becomes confined to intermediate and dorsal spinal cord regions (Figure 6(b)). Analogous to the regulation of the *Drosophila* Gli ortholog, Cubitus interruptus (Ci), Shh signaling appears to control the transcriptional activity of Gli2 and Gli3. The C-terminally deleted versions of Gli2 and Gli3 proteins are transcriptional repressors, and in the absence of Shh signaling Gli3 is processed to this transcriptional repressor form. On Shh binding to Ptc1, the formation of Gli3 repressor is blocked and the transcriptional activation potential of Gli3 is revealed. In the case of Gli2, exactly how Shh controls transcriptional activity remains unclear, but may involve regulation of Gli2 stability or transcriptional potency by posttranslational modification.

In the mouse, targeted deletion of Gli2 leads to a failure in the generation of floor plate and the adjacent p3 domain and a concomitant ventral expansion in the production of MNs (Figure 6(c)). Cell types located dorsal to the pMN are, however, unaffected. This suggests that Gli2 is required for specifying the cell types that require the highest levels of Shh signaling; in support of this, analysis of mutants for Gli2 crossed with Ptc1 mutants indicates that expansion of the floor plate observed in the absence of Ptc1 is abolished when Gli2 is also absent. Surprisingly, although ectopic Gli1 expression is able to induce ventral cell types, no neural defects have been detected in Gli1 mutant mice. However, a compound mutant lacking both Gli1 and Gli2 has more severe defects than do $Gli2^{-/-}$ mutants. Moreover, the neural tube defects seen in Gli2 mutant embryos can be rescued by replacing Gli2 with Gli1.

On the other hand, Gli3 has been proposed to function primarily as an inhibitor of Shh signaling in the neural tube. Supporting this idea, in the absence of Gli3, progenitor domains located in the intermediate region of the neural tube expand dorsally, concomitant with a switch in the identity of the neurons generated in this region. This phenotype can be corrected by a truncated form of Gli3, corresponding to the processed Gli3, suggesting that the repressor form of Gli3 is responsible for patterning in the intermediate neural tube. However, Gli3 can act as a transcriptional activator, as illustrated by its ability to partially

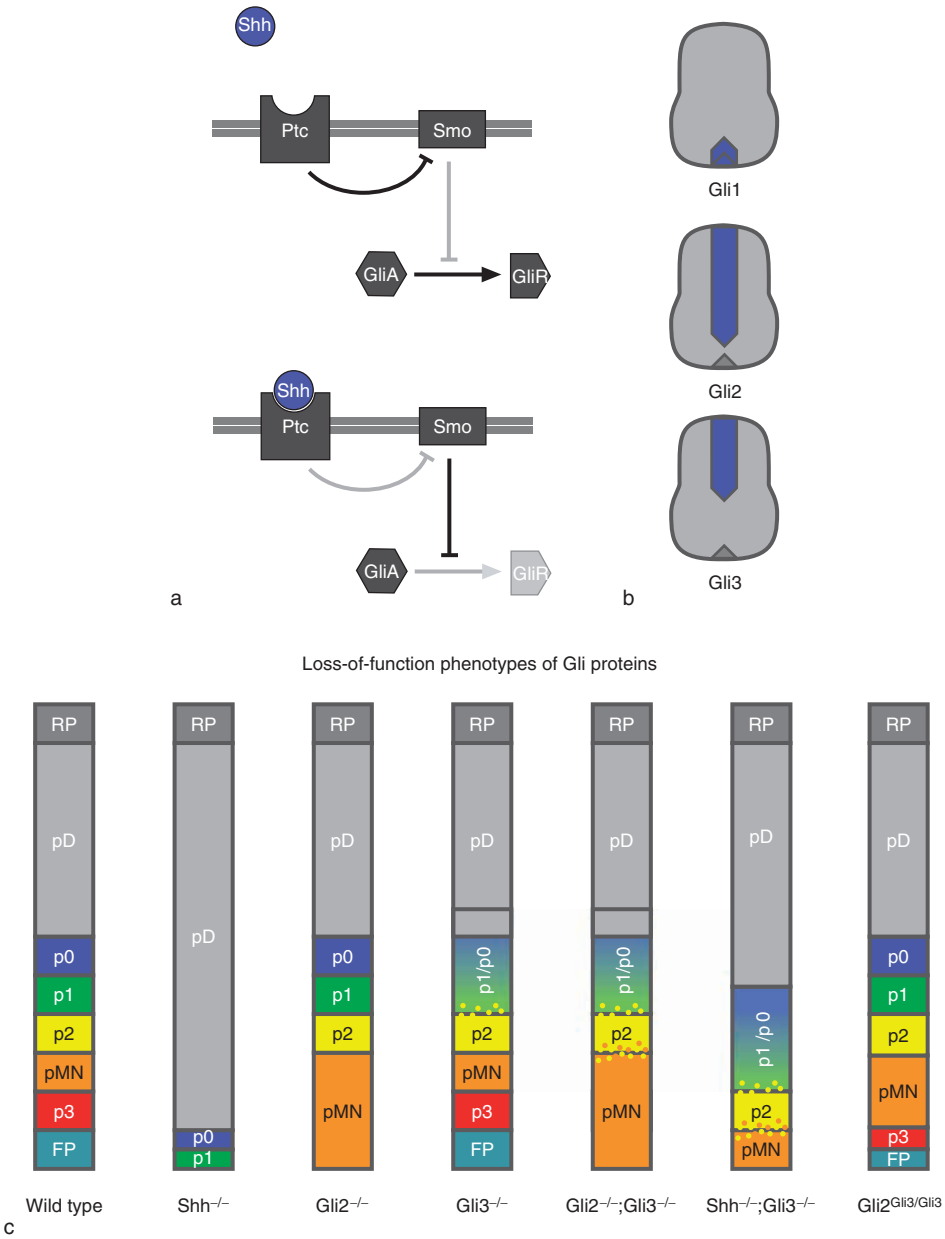

Figure 6 Gli proteins are involved in dorsoventral patterning of the neural tube. (a) In the absence of Sonic hedgehog (Shh), Patched (Ptc) inhibits the transmembrane protein Smoothened (Smo). Downstream transcriptional effectors, the Gli proteins, are thought to be processed from an activator form (GliA) to a repressor form (GliR). Shh binding to Ptc releases the inhibition of Smo, which inhibits production of inhibitory Gli proteins and promotes their transcriptional activity. (b) Schematic representing the expression of Gli proteins in the developing neural tube. (c) Abolishing Gli1 expression does not affect dorsoventral patterning. Loss of Gli2 (Gli2$^{-/-}$) results in an expansion of the ventral boundary of pMN domain together with a loss of the floor plate (FP) and the p3 domain. These results show that Gli2 is required for the patterning of the ventralmost region of the neural tube. Conversely, null mutants for Gli3 have defects in the patterning of the intermediate region of the neural tube accompanied with a mixing of the cells at the p2/p1 boundary. Double mutants for both Gli2 and Gli3 have a composite phenotype of the single mutants. Whereas Gli2 is acting as a transcriptional activator, Gli3 seems to mainly act as a repressor, as illustrated by the partial rescue of Shh$^{-/-}$ phenotype in the Shh$^{-/-}$;Gli3$^{-/-}$ mutant neural tube. However, Gli3 can also act as a transcriptional activator because it partially complements Gli2 when Gli3 is used to replace Gli2 in transgenic embryos. RP, roof plate; pD, dorsal spinal cord domain; pMN, motor neuron progenitor.

rescue the Gli2 mutant phenotype when expressed from this locus.

Importantly, abolishing Gli3 function in Shh$^{-/-}$ embryos partially rescues the severe defects in ventral neural tube patterning observed in Shh mutant embryos (Figure 6(c)). This confirms the inhibitory role for Gli3 in Shh signaling. It also indicates that although the spatial arrangement of neurogenesis

depends on Shh signaling, the induction of most ventral cell types can take place in the absence of Shh signaling, as long as the repressive activity of Gli3 is removed. Consistent with this, embryos lacking both Gli2 and Gli3 display a D/V pattern similar to that of Shh/Gli3 double mutants.

In zebra fish embryos, although the exact role of individual Gli genes appears to differ from that in the mouse, loss-of-function experiments indicate that the Gli proteins are also required for the reception of Hh signals. Together, these results demonstrate that Gli proteins act downstream of Shh to control ventral patterning through transcriptional regulation of target genes (**Figure 6(c)**). Moreover, the data can be accommodated by a model in which graded Shh signaling establishes a ventral-to-dorsal gradient of Gli activity that reflects the gradient of Shh activity. This Gli gradient corresponds to the sum of the transcriptional activator and repressor forms of all Gli proteins present in a responding cell. In support of this model, gain-of-function experiments in the chick suggest a tight correlation between progenitor cell fate and the level of Gli activity within a progenitor cell; furthermore, different levels of Gli activity are sufficient to emulate different levels of Shh signaling *in vivo*.

Perspective

Numerous studies show that Shh, emanating from the notochord and floor plate, patterns the ventral neural tube by controlling cell fate in a concentration-dependent manner. Several major issues remain to be resolved, including understanding how the Shh signaling pathway regulates Gli activity and how this controls the differential expression of class I and class II protein expression. Moreover, how other extracellular signals modulate and cross-talk with Shh signaling requires further investigation. In addition, the integration of D/V patterning with the anterior–posterior differences and temporal changes in neural cell fate remains to be explained. Finally, the evidence suggests that Shh signaling also controls the proliferation, adhesive properties, and survival of cells in the neural tube. The downstream targets of Shh signaling involved in these processes are not well defined, nor how these cell behaviors are coordinated with and contribute to the control of D/V patterning. Nevertheless, the progress that has been made provides a framework to understand fundamental aspects of the mechanism by which positional cues control the arrangement of distinct neuronal subtypes in the ventral neural tube.

See also: Helix–Loop–Helix (bHLH) Proteins: Proneural; Helix–Loop–Helix (bHLH) Proteins: Hes Family; Motor Neuron Specification in Vertebrates; Oligodendrocyte Specification; Sonic Hedgehog and Neural Patterning; Wnt Pathway and Neural Patterning.

Further Reading

Bai CB, Stephen D, and Joyner AL (2004) All mouse ventral spinal cord patterning by Hedgehog is Gli dependent and involves an activator function of Gli3. *Developmental Cell* 6: 103–115.

Briscoe J, Pierani A, Jessell TM, et al. (2000) A homeodomain protein code specifies progenitor cell identity and neuronal fate in the ventral neural tube. *Cell* 101: 435–445.

Diez del Corral R and Storey KG (2004) Opposing FGF and retinoid pathways: A signalling switch that controls differentiation and patterning onset in the extending vertebrate body axis. *BioEssays* 26: 857–869.

Ericson J, Rashbass P, Schedl A, et al. (1997) Pax6 controls progenitor cell identity and neuronal fate in response to graded Shh signalling. *Cell* 90: 169–180.

Gritli-Linde A, Lewis P, McMahon AP, et al. (2001) The whereabouts of a morphogen: Direct evidence for short- and long-range activity of hedgehog signaling peptides. *Developmental Biology* 236: 364–386.

Ingham PW and McMahon AP (2001) Hedgehog signaling in animal development: Paradigms and principles. *Genes and Development* 15: 3059–3087.

Jeong J and McMahon AP (2005) Growth and pattern of mammalian neural tube are govern by partially overlapping feedback activities of the hedgehog antagonists patched 1 and Hhip 1. *Development* 132: 143–154.

Jessell TM (2000) Neuronal specification in the spinal cord: Inductive signals and transcriptional codes. *Nature Reviews Genetics* 1: 20–29.

Lum L and Beachy PA (2004) The hedgehog response network: sensors, switches and routers. *Science* 304: 1755–1759.

Matise MP, Epstein DJ, Park HL, et al. (1998) Gli2 is required for induction of floor plate and adjacent cells, but not most ventral neurons in the mouse central nervous system. *Development* 125: 2759–2770.

Roelink H, Augsburger A, Heemskerk J, et al. (1994) Floor plate and motor neuron induction by vhh-1, a vertebrate homolog of hedgehog expressed by the notochord. *Cell* 76: 761–775.

Ruiz i Altaba A, Nguyen V, and Palma V (2003) The emergent design of the neural tube: Prepattern, SHH morphogen and GLI code. *Current Opinion in Genetics & Development* 13: 513–521.

Stamataki D, Ulloa F, Tsoni SV, et al. (2005) A gradient of Gli activity mediates graded Sonic Hedgehog signaling in the neural tube. *Genes & Development* 19: 626–641.

Relevant Website

http://hedgehog.sfsu.edu – Hedgehog Signaling Pathway Database, San Francisco State University (SFSU).

Fluid and Electrolyte Homeostasis: Clinical Disease

J G Verbalis, Georgetown University School of Medicine, Washington, DC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Disorders of body fluids are among the most commonly encountered problems in clinical medicine. This is because many different disease states can potentially disrupt the finely balanced mechanisms that control the intake and output of water and solutes. Although solute and water homeostasis mechanisms are closely linked, clinical disorders of body fluids are generally divided into disorders of water homeostasis and disorders of solute homeostasis. Since body water is the primary determinant of the osmolality of the extracellular fluid (ECF), disorders of water homeostasis can be broadly divided into hypoosmolar disorders, in which there is an excess of body water relative to body solute, and hyperosmolar disorders, in which there is a deficiency of body water relative to body solute. Because sodium is the main constituent of plasma osmolality, these disorders are typically characterized by hyponatremia and hypernatremia, respectively. After disorders of water metabolism, abnormalities of sodium balance are the next most common cause of body fluid disorders in clinical practice. Since body sodium is the primary determinant of the volume of the ECF, disorders of sodium metabolism can be broadly divided into hypovolemic disorders, in which there is a deficiency of ECF and plasma volume, and hypervolemic disorders, in which there is an excess of ECF and plasma volume. Depending on the relative balance between body sodium and water, these disorders can also be associated with either hyponatremia or hypernatremia.

Disorders of Hypoosmolality

Hypoosmolality indicates excess water relative to solute in the ECF; because water moves freely between the ECF and the intracellular fluid (ICF), this also indicates an excess of total body water relative to total body solute. Imbalances between body water and solute can be generated either by depletion of body solute greater than depletion of body water, or by dilution of body solute from increases in body water greater than increases in than body solute (Table 1). This is an oversimplification of complex physiology, and most hypoosmolar states include components of both solute depletion and water retention. Nonetheless, this general concept has proved to be useful because it provides a simple framework for understanding the basic etiologies of hypoosmolar disorders.

Definitive identification of the etiology of hypoosmolality is not always possible at the time of presentation, but categorization according to the patient's ECF volume status represents the first step in ascertaining the underlying cause of the disorder.

Decreased ECF Volume (Hypovolemia)

Clinically detectable hypovolemia indicates some degree of solute depletion. Even isotonic or hypotonic fluid losses can cause hypoosmolality if water or hypotonic fluids are subsequently ingested or infused. A low urine sodium concentration (U_{Na}) suggests a nonrenal cause of solute depletion, whereas a high U_{Na} suggests renal causes of solute depletion. Diuretic use is the most common cause of hypovolemic hypoosmolality. Many such patients do not manifest clinical evidence of hypovolemia, in part because ingested water has been retained in response to nonosmotically stimulated arginine vasopressin (AVP) secretion, which occurs in all disorders of solute depletion as an attempt to maintain volume homeostasis. Most etiologies of solute losses causing hypovolemic hypoosmolality will be clinically apparent, although some salt-wasting nephropathies and mineralocorticoid deficiency may be difficult to diagnose during early phases of these diseases.

Normal ECF Volume (Euvolemia)

Virtually any disorder causing hypoosmolality can present with a volume status that appears normal by standard methods of clinical evaluation. Because clinical assessment of volume status is not very sensitive, normal or low blood urea nitrogen and uric acid concentrations are helpful laboratory correlates of relatively normal ECF volume. In these cases, a low U_{Na} ($<30\,\mathrm{mmol\,l^{-1}}$) suggests depletional hypoosmolality secondary to ECF losses, with subsequent volume replacement by water or other hypotonic fluids; as discussed earlier, such patients may appear euvolemic by the usual clinical parameters used to assess hydration status. Hypoosmolar disorders caused primarily by dilution (Table 1) are less likely with a low U_{Na}, although this can occur in hypothyroidism or in the syndrome of inappropriate antidiuretic hormone secretion (SIADH) with superimposed hypovolemia. A high U_{Na} ($>30\,\mathrm{mmol\,l^{-1}}$) generally indicates a dilutional hypoosmolality such as SIADH, the most common cause of euvolemic hypoosmolality. The clinical criteria necessary to diagnose SIADH remain as initially defined by Bartter and Schwartz in 1967. Many different

Table 1 Pathogenesis of hypoosmolar disorders

Solute depletion (primary decreases in total body solute + secondary water retention)[a]
Renal solute loss
Diuretic use
Solute diuresis (glucose, mannitol)
Salt-wasting nephropathy
Mineralocorticoid deficiency
Nonrenal solute loss
Gastrointestinal (diarrhea, vomiting, pancreatitis, bowel obstruction)
Cutaneous (sweating, burns)
Blood loss
Solute dilution (primary increases in total body water + secondary solute depletion)[a]
Impaired renal free water excretion
Increased proximal nephron reabsorption
Congestive heart failure
Cirrhosis
Nephrotic syndrome
Hypothyroidism
Impaired distal nephron dilution
Syndrome of inappropriate antidiuretic hormone secretion (SIADH)
Glucocorticoid deficiency
Excess water intake
Primary polydipsia

[a]Virtually all disorders of solute depletion are accompanied by some degree of secondary retention of water by the kidneys in response to the resulting hypovolemia; this mechanism can lead to hypoosmolality even when the solute depletion occurs via hypotonic or isotonic body fluid losses. Disorders of water retention can cause hypoosmolality in the absence of any solute losses, but often some secondary solute losses occur in response to the resulting hypervolemia, and this can then further aggravate the dilutional hypoosmolality.

disorders are associated with SIADH, which can be divided into four major etiologic groups: tumors, central nervous system disorders, drug effects, and pulmonary diseases.

Increased ECF Volume (Hypervolemia)

Clinically detectable hypervolemia indicates whole-body sodium excess, and hypoosmolality in such cases suggests a relatively decreased intravascular volume and/or pressure leading to water retention as a result of elevated plasma AVP levels and decreased distal delivery of glomerular filtrate to the kidneys. Patients with these findings usually have a low U_{Na} because of secondary hyperaldosteronism, but under certain conditions the U_{Na} may be elevated (e.g., diuretic therapy). Hyponatremia generally does not occur until relatively advanced stages of congestive heart failure, cirrhosis, or the nephrotic syndrome, by which time diagnosis is usually not difficult. Renal failure can also cause retention of both sodium and water, but in this case the factor limiting excretion of excess body fluid is not decreased effective circulating volume, but rather decreased renal glomerular filtration.

Clinical Manifestations of Hyponatremia

The clinical manifestations of hyponatremia are largely neurological and primarily reflect brain edema resulting from osmotic water shifts into the brain. These range from nonspecific symptoms such as headache and confusion, to more severe manifestations such as decreased sensorium, coma, seizures, and death. Significant central nervous system (CNS) symptoms generally do not occur until the serum sodium concentration ($[Na^+]$) falls below $125\,mmol\,l^{-1}$, and the severity of symptoms can be roughly correlated with the degree of hypoosmolality. Individual variability is marked, and for any patient the level of serum $[Na^+]$ at which symptoms will appear cannot be accurately predicted. Several factors other than the severity of the hypoosmolality also affect the degree of neurological dysfunction. The most important is the time course over which hypoosmolality develops. Rapid development of severe hypoosmolality frequently causes marked neurological symptoms, whereas gradual development over several days or weeks is often associated with relatively mild symptomatology despite profound degrees of hypoosmolality. This is because the brain counteracts osmotic swelling by extruding extracellular and intracellular solutes, including potassium and a variety of small organic molecules (amino acids, polyols, and methylamines) called organic osmolytes. Since this is a time-dependent process, rapid development of hypoosmolality can result in brain edema before this adaptation occurs, but, with slower development of the same degree of hypoosmolality, brain cells can lose solute sufficiently rapidly to prevent cell swelling, brain edema, and

neurological dysfunction. Underlying neurological disease also affects the level of hypoosmolality at which CNS symptoms appear; moderate hypoosmolality is of little concern in an otherwise healthy patient, but can cause morbidity in a patient with an underlying seizure disorder. Nonneurological metabolic disorders (hypoxia, hypercapnia, acidosis, hypercalcemia, etc.) similarly can affect the level of plasma osmolality at which CNS symptoms occur. Recent clinical studies have suggested that menstruating females and young children may be particularly susceptible to the development of neurological morbidity and mortality during hyponatremia, especially in the acute postoperative setting.

Disorders of Hyperosmolality

Hyperosmolality indicates a deficiency of water relative to solute in the ECF. Because water moves freely between the ICF and ECF, this also indicates a deficiency of total body water relative to total body solute. Although hypernatremia can be caused by an excess of body sodium, the vast majority of cases are due to losses of body water in excess of body solutes, caused by either insufficient water intake or excessive water excretion. Consequently, most of the disorders causing hyperosmolality are those associated with inadequate water intake and/or deficient pituitary AVP secretion (Table 2). Although hyperosmolality from inadequate water intake is seen frequently in clinical practice, this is usually not due to an underlying defect in thirst but rather results from a generalized incapacity to obtain and/or ingest fluids, often stemming from a depressed sensorium.

Evaluation of the patient's ECF volume status is important as a guide to fluid replacement therapy, but is not as useful for differential diagnosis since most hyperosmolar patients will manifest some degree of hypovolemia. Rather, assessment of urinary concentrating ability provides the most useful data with regard to the type of disorder present. Using this approach, disorders of hyperosmolality can be categorized as those in which renal water conservation mechanisms are intact but are unable to compensate for inadequately replaced losses of hypotonic fluids from other sources, or those in which renal concentrating defects are a contributing factor to the deficiency of body water. Most cases of CNS-associated hyperosmolality are due to deficits of pituitary AVP secretion.

Diabetes Insipidus

Diabetes insipidus (DI) can result from either inadequate AVP secretion (central or neurogenic DI) or inadequate renal response to AVP (nephrogenic DI).

Table 2 Pathogenesis of hyperosmolar disorders

Water depletion (decreases in total body water in excess of body solute)
Insufficient water intake
Unavailability of water
Hypodipsia (osmoreceptor dysfunction, age)
Neurological deficits (cognitive dysfunction, motor impairments)
Hypotonic fluid loss[a]
Renal: diabetes insipidus (DI)
Insufficient AVP secretion (central DI, osmoreceptor dysfunction)
Insufficient AVP effect (nephrogenic DI)
Renal: other fluid loss
Osmotic diuresis (hyperglycemia, mannitol)
Diuretic drugs (furosemide, ethacrynic acid, thiazides)
Postobstructive diuresis
Diuretic phase of acute tubular necrosis
Nonrenal fluid loss
Gastrointestinal (vomiting, diarrhea, nasogastric suction)
Cutaneous (sweating, burns)
Pulmonary (hyperventilation)
Peritoneal dialysis
Solute excess (increases in total body solute in excess of body water)
Sodium
Excess Na^+ administration (NaCl, $NaHCO_3$)
Sea water drowning
Other
Hyperalimentation (intravenous, parenteral)

[a]Most hypotonic fluid losses will not produce hyperosmolality unless insufficient free water is ingested or infused to replace the ongoing losses, so these disorders also usually involve some component of insufficient water intake.

Central DI is caused by a variety of acquired or congenital anatomic lesions that disrupt the neurohypophysis, including pituitary surgery, tumors, trauma,hemorrhage, thrombosis, infarction, or granulomatous disease. Severe nephrogenic DI is most commonly congenital due to defects in the gene for the AVP V_2 receptor (X-linked recessive pattern of inheritance) or in the gene for the aquaporin-2 water channel (autosomal recessive pattern of inheritance), but relief of chronic urinary obstruction or therapy with drugs such as lithium can cause an acquired form sufficient to warrant treatment. Short-lived nephrogenic DI can result from hypokalemia or hypercalcemia, but the mild concentrating defect generally does not by itself cause hypertonicity, and responds to correction of the underlying disorder. Regardless of the etiology of the DI, the end result is a water diuresis due to an inability to concentrate urine appropriately.

Because patients with DI do not have impaired urine Na^+ conservation, the ECF volume is generally not markedly decreased and regulatory mechanisms for maintenance of osmotic homeostasis are primarily activated: stimulation of thirst and pituitary AVP secretion (to whatever degree the neurohypophysis is still able to secrete AVP). In cases in which AVP secretion is totally absent (complete DI), patients are dependent entirely on water intake for maintenance of water balance. Although untreated DI can lead to both hyperosmolality and volume depletion, until the water losses become severe, volume depletion is minimized by osmotic shifts of water from the ICF into the more osmotically concentrated ECF.

Osmoreceptor Dysfunction

There is an extensive literature on animal studies indicating that the primary osmoreceptors that control AVP secretion and thirst are located in the anterior hypothalamus. Lesions of this region in animals cause hyperosmolality through a combination of impairments of thirst and of osmotically stimulated AVP secretion. Initial reports in humans described this syndrome as 'essential hypernatremia,' and subsequent studies used the term 'adipsic hypernatremia' in recognition of the profound thirst deficits found in most of the patients. All of these syndromes are now grouped together as disorders of osmoreceptor function. Most of the cases reported to date have represented various degrees of osmoreceptor destruction associated with different brain lesions. In contrast to lesions causing central DI, these lesions usually occur more rostrally in the hypothalamus, consistent with the anterior hypothalamic location of the primary osmoreceptor cells. For all cases of osmoreceptor dysfunction it is important to remember that afferent pathways from the brain stem to the hypothalamus generally remain intact; therefore, the patients will usually have normal AVP and renal concentrating responses to baroreceptor-mediated stimuli such as hypovolemia and hypotension.

Clinical Manifestations of Hyperosmolality

The clinical manifestations of hyperosmolality can be divided into the signs and symptoms produced by dehydration, which are largely cardiovascular, those caused by the hyperosmolality itself, which are predominantly neurological and reflect brain dehydration as a result of osmotic water shifts out of the central nervous system, and those secondary to excessive renal water losses in patients with DI. Cardiovascular manifestations of hypertonic dehydration include hypotension, kidney failure secondary to decreased renal perfusion, acute tubular necrosis or rhabdomyolysis, and, in severe cases, hypotensive shock. Neurological manifestations range from nonspecific symptoms such as irritability and decreased sensorium to more severe manifestations such as chorea, seizures, coma, focal neurological deficits, and cerebral infarction. The severity of symptoms can be roughly correlated with the degree of hyperosmolality, but individual variability is marked, and for any single patient the level of serum $[Na^+]$ at which symptoms will appear cannot be predicted. Similar to hypoosmolar syndromes, the length of time over which hyperosmolality develops can markedly affect clinical symptomatology. Rapid development of severe hyperosmolality is frequently associated with marked neurological symptoms, whereas gradual development over several days or weeks generally causes milder symptoms. In this case the brain counteracts osmotic shrinkage by increasing intracellular content of solutes. These include electrolytes, such as potassium, and organic osmolytes, which previously had been called 'idiogenic osmoles' (for the most part, these are the same organic osmolytes that are lost from the brain during adaptation to hypoosmolality, discussed previously). The net effect of this process is to protect the brain against excessive shrinkage during sustained hypertonicity. However, once the brain has adapted by increasing its solute content, rapid correction of the hyperosmolality can cause brain edema, since it takes a finite time (24–48 h in animal studies) to dissipate the accumulated solutes, and until this process has been completed the brain will accumulate excess water as the plasma osmolality is normalized. This effect is most often seen in dehydrated pediatric patients, who can develop seizures as a result of rapid rehydration.

Disorders of Hypovolemia

It is important to distinguish dehydration, which indicates a relative deficiency of body water, from hypovolemia, which indicates insufficient extracellular fluid to maintain blood volume within normal ranges and almost always requires a deficiency of body sodium. Since dehydration often occurs in association with hypovolemia, these terms are frequently used interchangeably but inappropriately. Although water loss or inadequate intake also can produce hypovolemia, to a large degree the ECF volume deficits will be reduced by osmotic water shifts out of the ICF space. The degrees of hyperosmolality that can be achieved in dehydrated patients with relatively small effects on circulating blood volume and pressure are striking. Clinically significant hypovolemia develops much more quickly, and to a much more pathological degree, when water losses occur in conjunction with solute losses. In this case, sodium balance is critically important for maintenance of body fluid homeostasis. Because of a relative absence of regulated sodium intake in humans, renal sodium conservation represents the main defense for maintenance of volume homeostasis during hypovolemic conditions.

Hypovolemic states are generally classified into those in which solute is lost through extrarenal causes but renal function is unimpaired, and those in which an impairment of renal sodium conservation mechanisms either causes or exacerbates the hypovolemia (Table 3).

Table 3 Pathogenesis of hypovolemic disorders

Extrarenal solute and water depletion
Blood loss
Hemorrhage
Severe anemia
Gastrointestinal fluid losses
Diarrhea
Vomiting
Nasogastric suctioning
Pancreatitis
Bowel obstruction
Cutaneous fluid losses
Sweating
Burns
Renal solute and water depletion
Mineralocorticoid deficiency
Addison's disease
Salt-wasting nephropathy
Chronic interstitial nephropathy
Polycystic kidney disease
Medullary cystic disease
Obstructive uropathy
Bartter's syndrome
Drugs
Diuretic agents
Solute diuresis
Glucose
Mannitol

Solute Losses by Extrarenal Pathways

The presence of a low U_{Na} in a patient with clinical manifestations of hypovolemia generally indicates an extrarenal cause. This can occur via a variety of etiologies. In most cases the source of the sodium and water losses (sweating, diarrhea, hemorrhage, etc.) will be readily apparent. In some patients the loss of water is disproportionately greater than the loss of sodium. When primarily water is lost, as in DI, then intracellular dehydration will be more marked than hypovolemia, as described earlier, and osmotic homeostatic mechanisms will be activated to a greater degree than are the renal mechanisms that act to preserve volume homeostasis. In contrast, isotonic hypovolemia, as produced by large amounts of blood loss from hemorrhage, represents a situation in which the regulatory mechanisms for water and sodium balance that are activated will be primarily those that protect volume homeostasis, without much contribution from osmotic regulatory systems. Most body fluid losses, whether from excessive sweating, gastrointestinal excretions, or diuretic use, result in losses of both water and sodium, but the fluid lost in such situations is generally in the range of 75 mM, or half-isotonic. As a result, both hypovolemia and hyperosmolality occur, and volume homeostatic mechanisms will be activated more than in cases of pure water losses, such as diabetes insipidus. Consequently, renal sodium conservation, as reflected by a low U_{Na}, occurs prominently in hypovolemic patients with normal kidney function.

Renal Solute Losses

The presence of an elevated U_{Na} in a patient with clinical manifestations of hypovolemia generally indicates a renal cause of hypovolemia. This can occur via several different etiologies. Adrenal insufficiency (Addison's disease) is caused by bilateral destruction of the adrenal glands. Because the adrenal glands are the site of synthesis of aldosterone in addition to glucocorticoids, patients with Addison's disease have impaired renal sodium conservation. This leads to renal sodium wasting in excess of the sodium losses observed with other forms of hypovolemia, resulting in both hypoosmolality and hypovolemia. The net effect on sodium balance will be to activate mechanisms for both sodium intake and renal sodium conservation. However, because aldosterone acts on the more distal part of the nephron, proximal renal sodium-conserving mechanisms cannot adequately compensate for the sodium wasting in this case. Analogous to

primary adrenal insufficiency, other primary disorders of renal sodium wasting also predispose patients to the development of clinical hypovolemia. These include chronic interstitial nephropathy, polycystic kidney disease, medullary cystic disease, obstructive uropathy, Bartter's syndrome, and drugs such as cisplatinum. However, the sodium losses in these diseases are usually relatively mild and only rarely cause clinically significant hypovolemia, since dietary sodium intakes are generally more than sufficient to replace daily urinary sodium losses even when these are fairly large.

Therapy with diuretic drugs is one of the most common causes of hyponatremia. Although the resulting sodium losses often induce some degree of hypovolemia as well, the magnitude of ECF volume reduction is generally mild. In contrast, obligatory solute diureses, particularly those associated with glucosuria, can cause substantially greater urinary sodium losses and severe hypovolemia relatively quickly, primarily because of the ongoing natriuresis and diuresis. One manifestation of this is the circulatory collapse that represents a major cause of early mortality in patients with uncontrolled diabetic ketoacidosis.

Clinical Manifestations of Hypovolemia

The hallmarks of hypovolemia are clinical signs of inadequate ECF and plasma volume, most of which reflect increased activity of the sympathetic nervous system to maintain mean arterial pressure via release of norepinephrine from nerve terminals and epinephrine from the adrenal medulla. Associated symptoms are those of orthostatic dizziness, dry mucus membranes, and thirst. The last symptom is a crucial part of compensatory mechanisms that maintain ECF volume. However, the absence of increased sodium appetite in response to hypovolemia limits the effectiveness of thirst to restore ECF volume and contributes to the production of hypoosmolality as a result of ingestion of hypotonic fluids. In addition, age-related changes appear to impair some of the compensatory homeostatic responses to hypovolemia. Elderly patients are known to be particularly prone to hypovolemia, regardless of the cause. This suggests the presence of defects in the mechanisms that regulate osmotic and volume homeostasis. Studies in elderly humans have consistently demonstrated intact or enhanced AVP secretion in response to osmotic stimuli. Although elderly patients generally have a decreased renal response to AVP, given the nature of the urine volume–urine osmolality relationship shown in Figure 1, a small decrease in urinary concentrating ability will not cause markedly increased urine volume losses under most conditions. The most significant defect in the elderly appears to be a decreased sensation of thirst. After 24 h of fluid deprivation, elderly individuals report less subjective thirst than do younger individuals, which can lead to less spontaneous drinking in response to periods of water deprivation.

Disorders of Hypervolemia

Hypervolemia results from excess ECF and plasma volume. Because water distributes across all body fluid compartments, excess sodium is much more effective at producing ECF and plasma volume expansion because of its more restricted extracellular distribution. However, hyperosmolality from an excess of other solutes does not invariably lead to volume expansion. Patients with renal failure have high levels of blood urea that often increase total plasma osmolality by as much as 20–30 mOsm kg^{-1} H_2O, yet this causes neither intracellular dehydration nor ECF and plasma volume expansion. This is because urea is a nonpolar solute that freely permeates cell membranes. Only increases in concentrations of impermeable solutes are accompanied by ECF volume expansion, and such clinical disorders are caused almost exclusively by sodium retention.

Most clinical disorders characterized by increased total body sodium and fluid are the result of a perceived deficit in the 'effective' ECF volume. Thus, in diseases such as congestive heart failure (CHF) and cirrhosis, the kidney avidly retains sodium and water despite a total body surfeit of both of these substances. Although many questions about specific disease processes remain to be answered, evidence over the past decade has supported the concept that the critical component of the ECF volume that is the primary determinant of renal sodium retention is the effective arterial blood volume (EABV), which represents only approximately 15% of the blood volume, or about 0.5–0.7 l in a 70 kg human.

Blood volume is sensed in the arterial and venous circulations by baroreceptors that alter their electrical output according to the degree of stretch of the blood vessel wall. In arteries this is predominantly a function of the blood pressure (mean arterial pressure as well as dynamic changes in pulse pressure) and of the intrinsic tensile properties of the arterial wall, whereas in veins the stretch is mainly a function of the volume of blood within the vessel. Consequently, under normal conditions the major determinant of arterial baroreceptor activity is the pressure in the large vessels of the arterial circulation, primarily the aorta and carotid arteries, and the major determinant of venous baroreceptor activity is the circulating blood volume. The arterial pressure is primarily determined by two factors, cardiac output and peripheral arterial resistance. Similarly, renal perfusion pressure

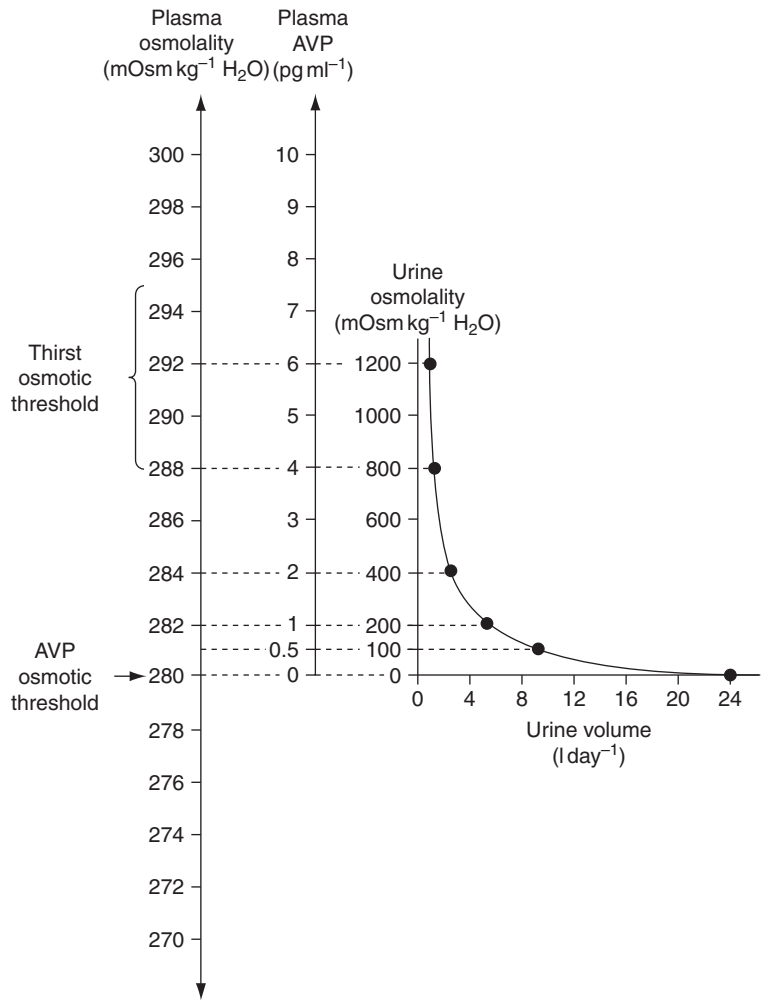

Figure 1 Schematic representation of normal physiological relationships between plasma osmolality, plasma arginine vasopressin (AVP) concentrations, urine osmolality, and urine volume in humans. Note particularly the inverse nature of the relation between urine osmolality and urine volume, resulting in disproportionate effects of small changes in plasma AVP concentrations on urine volume at lower AVP levels. Adapted from Robinson AG (1985) Disorders of antidiuretic hormone secretion. *Journal of Clinical Endocrinology & Metabolism* 4: 55–88.

determines the activity of the baroreceptors of the renal afferent arterioles that control renin secretion from the juxtaglomerular apparatus. Acute decreases in either variable are interpreted as a decreased EABV: in CHF this occurs as a result of the decrease in cardiac output (and possibly pulse pressure) as myocardial function is compromised, and in cirrhosis as a result of the decrease in peripheral arterial resistance due to arteriovenous shunting and circulating vasodilators. In response to either of these factors, the activity of arterial and renal afferent arteriolar baroreceptors is increased and a cascade of neural and humoral responses is activated to increase arterial pressure back to normal ranges. These include (1) activation of the sympathetic nervous system (both norepinephrine release from sympathetic nerve terminals and epinephrine release from the adrenal medulla), which increases peripheral arteriolar resistance and stimulates renin production by the kidney; (2) suppression of brain stem vagal efferent output to the heart, which causes a reflex tachycardia; (3) activation of the renin–angiotensin system, which promotes renal sodium conservation via angiotensin-stimulated aldosterone secretion; and (4) stimulation of pituitary AVP secretion, which promotes renal water retention. If these compensatory responses are sufficient to allow the EABV to return to normal, then cardiac output and mean arterial pressure will return toward normal too, but this is only because they are now being supported by the secondary effects set in motion as a result of the initial decrease in EABV. Thus, a new steady state is achieved in which an increase in total body sodium and water supports the maintenance of EABV at acceptable levels.

Table 4 Pathogenesis of hypervolemic disorders

Primary renal sodium retention
Renal failure
Primary hyperaldosteronism
Adrenal adenoma
Adrenal hyperplasia
Exogenous mineralocorticoid therapy
Secondary renal sodium retention
Congestive heart failure
Cirrhosis
Nephrotic syndrome

Hypervolemic states can be broadly divided into those in which pathological sodium and water retention occurs as a result of primary renal or adrenal disease and independently of the ECF volume status, and those in which renal sodium and water retention occurs secondarily to stimuli that signal an insufficiency of EABV (Table 4).

Primary Renal Sodium Retention (Baroreceptor Suppression)

Diseases of primary renal sodium retention produce an expansion of all components of the ECF, including the circulating arterial volume. As a result, these disorders are characterized by suppression of most of the neural and humoral mechanisms that act to maintain EABV. Acute and chronic renal failure represent the quintessential examples of primary renal sodium retention. As the glomerular filtration rate deceases, the kidney eventually becomes unable to excrete sufficient sodium and water to equal ongoing fluid and sodium intake. At what stage of renal failure clinically significant fluid retention and edema occur depends on the patient's salt and water intakes as well as their degree of renal impairment. Because all ECF compartments are expanded, the renin–angiotensin–aldosterone system and the sympathetic nervous system are suppressed via the effects of baroreceptor activation. Hyperaldosteronism is another example of primary sodium retention with subsequent ECF volume expansion. Although the kidneys are normal and the disease is secondary to elevated mineralocorticoid activity, because this process occurs independently of the ECF volume status it falls into the category of primary renal sodium retention. High levels of plasma mineralocorticoid activity can result from an adrenal tumor or hyperplasia (primary hyperaldosteronism) or from pharmacological therapy with synthetic mineralocorticoids. Regardless of the etiology, the net effect is sodium retention.

Secondary Renal Sodium Retention (Baroreceptor Activation)

In diseases of secondary renal sodium retention, high plasma aldosterone levels are produced by angiotensin II stimulation of adrenal aldosterone secretion. Normally, activation of the renin–angiotensin system occurs under physiological conditions of hypovolemia to promote renal sodium conservation; however, when this cascade is activated in response to a perceived decrease in EABV, then pathological sodium and fluid retention results. Many different pathophysiological states can cause sodium retention and its sequelae. Congestive heart failure is the best known cause of hypervolemia. Although the pathophysiology seems straightforward, because the disease evolves slowly over long periods of time, many of the clinical manifestations with regard to sodium and fluid balance depend on the point in the disease process at which the patient is evaluated. As cardiac output continues to decrease, more defense mechanisms will be called into play to support EABV.

The pathophysiology underlying the fluid and electrolyte abnormalities of cirrhosis remains unsettled. The sodium and fluid retention are clearly not the result of a primary renal defect, but rather are secondary to a perceived decreased EABV, since all of the regulatory systems for maintenance of EABV are activated as they are in CHF. However, the cause of the decreased EABV is less certain. It has been proposed that a key factor in the pathophysiology is decreased peripheral vascular resistance, as a result of both the arteriovenous shunting in the gastrointestinal and other vascular beds in this disease, as well as from the decreased metabolism and clearance of a variety of vasoactive agents that accumulate during hepatic insufficiency. However, other pathogenetic factors are likely involved in this very complex disease as well.

Clinical Manifestations of Hypervolemia

The clinical manifestations of hypervolemia depend on the underlying disease process. Patients with primary aldosteronism generally have mild hypernatremia and hypokalemia, and manifest varying degrees of hypertension but less frequently subcutaneous edema. When the serum $[K^+]$ is $<3\,\mathrm{mmol\,l^{-1}}$, symptoms of hypokalemia, such as muscle weakness, can predominate. Patients with hyperaldosteronism secondary to perceived reduced EABV, such as congestive heart failure and cirrhosis, generally do not have hypertension but tend to have greater degrees of edema, reflecting more sodium and water retention,

because they cannot escape the effects of elevated aldosterone levels due to their decreased renal perfusion. This produces greater amounts of subcutaneous edema and, in later stages of the diseases, accumulations of excess retained fluid in the peritoneal cavity (ascites) and lungs (pulmonary edema) and whole body swelling (anasarca).

See also: Angiotensin II; Atrial Natriuretic Peptide: Fluid/Mineral Balance; Neurohypophyseal System; Osmoregulation; Salt Appetite; Thirst.

Further Reading

Adrogue HJ and Madias NE (2000) Hyponatremia. *New England Journal of Medicine* 342: 1581–1589.

Bartter FC and Schwartz WB (1967) The syndrome of inappropriate secretion of antidiuretic hormone. *American Journal of Medicine* 42: 790–806.

Knepper MA (1997) Molecular physiology of urinary concentrating mechanism: Regulation of aquaporin water channels by vasopressin. *American Journal of Physiology* 272: F3–F12.

Phillips PA, Rolls BJ, Ledingham JG, et al. (1984) Reduced thirst after water deprivation in healthy elderly men. *New England Journal of Medicine* 311: 753–759.

Robertson GL (1983) Thirst and vasopressin function in normal and disordered states of water balance. *Journal of Laboratory and Clinical Medicine* 101: 351–371.

Robinson AG (1985) Disorders of antidiuretic hormone secretion. *Journal of Clinical Endocrinology & Metabolism* 4: 55–88.

Robinson AG and Verbalis JG (2003) The posterior pituitary. In: Larsen PR, Kronenberg HM, Melmed S, et al. (eds.) *Williams Textbook of Endocrinology,* 10th edn., pp. 281–329. Philadelphia: W.B. Saunders.

Schrier RW (1988) Pathogenesis of sodium and water retention in high-output and low-output cardiac failure, nephrotic syndrome, cirrhosis, and pregnancy. *New England Journal of Medicine* 319: 1065–1072, 1127–1134.

Stricker EM and Verbalis JG (2003) Water intake and body fluids. In: Squire L, Bloom F, McConnell S, et al. (eds.) *Fundamental Neuroscience*, pp. 1011–1029. New York: Academic Press.

Verbalis JG (2001) The syndrome of inappropriate antidiuretic hormone secretion and other hypoosmolar disorders. In: Schrier RW (ed.) *Diseases of the Kidney and Urinary Tract*, pp. 2511–2548. Philadelphia: Lippincott, Williams & Wilkins.

Verbalis JG (2006) Control of brain volume during hypoosmolality and hyperosmolality. *Advances in Experimental Medicine and Biology* 576: 113–129.

Fluorescence Microscopy in the Neurosciences

F G Wouterlood and A J Boekel, Vrije Universiteit Medical Center, Amsterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Photoluminescence Phenomena

Fluorescence and phosphorescence belong to the class of physical events called photoluminescence; many chemical substances emit light of a particular wavelength (perceived by the human eye as a particular color) very shortly after being hit by light of another wavelength. Based on the delay between the absorption and emission of light, photoluminescence events are called fluorescence (very short delay; around 10^{-6}–10^{-8} s) or phosphorescence (delay 10^{-6} s and longer). Substances that fluoresce are called fluorochromes.

High-energy electromagnetic radiation has a wavelength shorter than that of visible light. With decreasing energy contents this wavelength becomes increasingly longer. For instance, the wavelength of X-rays is measured in angstrom units while the wavelength of visible light is suitably expressed in nanometers (1 nm = 10 Å units). The relationship between energy and wavelength is understandable if one is familiar with Max Planck's mathematical equation:

$$E = h\nu$$

where E is energy, h is Planck's constant, and ν is frequency.

In the color spectrum, the wavelength of the emitted light is always longer than that of the absorbed light (the so-called Stokes shift). Thus, ultraviolet (UV) light causes a suitable fluorochrome to fluoresce blue light, blue light causes a suitable fluorochrome to fluoresce green, green light causes red fluorescence, and red light causes infrared fluorescence. Fluorescence and the wavelength shift between excitation light and emission light were first scientifically described by the English physicist George Gabriel Stokes in 1852.

What Is Light?

Light can be considered as an electromagnetic waveform or as a stream of quanta of energy, photons. The quantum theory explanation of fluorescence is as follows. Fluorochrome molecules in their ground state of energy absorb photons. The energy contained by the photons is transferred to the fluorochrome molecule, which enters a higher energy state (called the excited state). A split-second later, the excited fluorochrome molecule returns to the ground state while emitting the excess energy as a photon in the visible light band of the electromagnetic spectrum. The quantum physics underlying excitation and emission is known as the Franck–Condon principle. This principle also explains the Stokes wavelength shift.

Design of the Fluorescence Microscope

The construction most often applied in modern fluorescence microscopes follows the principle of epifluorescence (Figure 1). In short, high-intensity light is projected via a mirror placed in the optic path, an excitation filter, and the objective lens, onto the section or other object of study. The excitation filter is necessary to select light with the appropriate excitation wavelength (see section on Filters). Fluorescence emitted by fluorochrome molecules inside or attached to the specimen passes through the objective lens and also hits the mirror. This mirror is the construction's centerpiece. It is covered with a suitable coating that reflects the high-intensity excitation light from the light source toward the objective lens and the specimen, while at the same time it reflects away from the observer's eye the excess excitation light reflected back from the specimen into the microscope via the objective lens. Most important is that the mirror does not reflect the fluorescence emitted by the specimen. As a consequence, the light that has emerged from the object as fluorescence passes the mirror unreflected, passes an extra filter (the emission filter), and reaches the observer's eye, the camera, or some detector. The characteristic that this mirror separates the excitation light from the fluorescence light is basic to its name: dichroic mirror. Sometimes this mirror is referred to as the 'beam splitter.' In some microscope designs the excitation light first passes a so-called barrier filter to limit the wavelength band and then passes through the excitation filter.

Thus, the epifluorescence microscope is designed in such a way that only fluorescence light reaches (via the ocular lens the eye of the observer) the camera or other imaging device, while the excitation light does not. Every major manufacturer of optical instruments carries a line of fluorescence microscopes, both upright and inverted models and nearly always based on the epifluorescence principle.

Fluorochromes: Absorption and Emission Spectra

The effectiveness of a specific fluorochrome for a particular staining purpose depends on several characteristics. In this respect the associated absorption

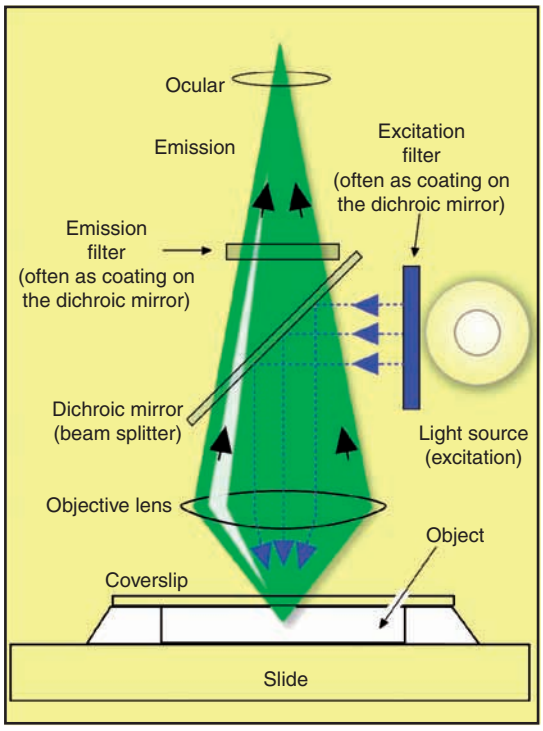

Figure 1 Principle of epifluorescence microscopy involving a dichroic mirror. Excitation light is reflected, emitted light (fluorescence) is not.

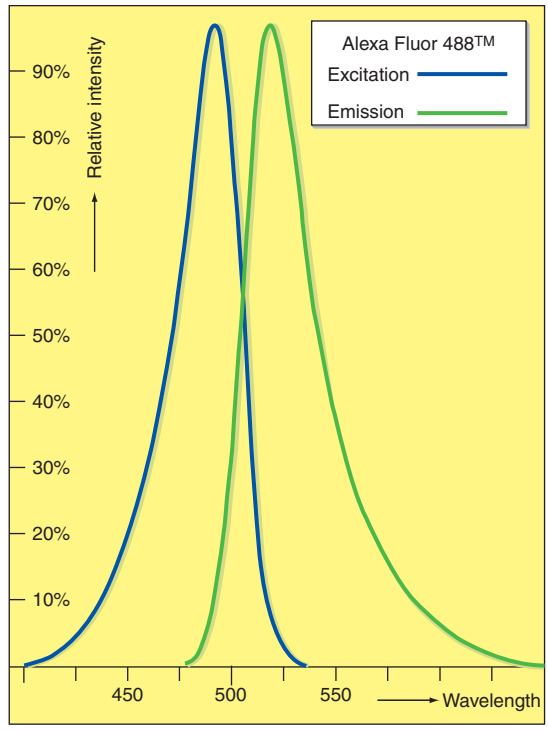

Figure 2 Typical excitation and emission spectra for Alexa Fluor 488. Note the shift toward longer wavelengths of the emission spectrum compared with the excitation spectrum (Stokes shift). The excitation spectrum of Alexa Fluor 488 has a peak at 492 nm; the emission spectrum has a maximum at 520 nm.

and emission spectra take a high priority on the list. The absorption spectrum (also called excitation spectrum) is a graph in which the intensity of the emission is plotted against the wavelength of the light used to excite the fluorochrome (Figure 2). From this graph the user decides which 'color' light (which excitation wavelength) should preferably be used to get the maximum fluorescence out of the fluorochrome.

The emission spectrum is a graph that shows the intensity of the emitted fluorescent light plotted against the wavelength of the emitted fluorescence (Figure 2). This graph informs the user with which filter or detector settings the highest yield or best signal of the fluorescence output can be obtained. Note that because of the Stokes wavelength shift, the emission spectrum of any fluorochrome always has a peak at a longer wavelength, compared to the peak that occurs in the excitation spectrum.

Polychromatic and Monochromatic Light

The conventional mercury and xenon light sources applied in fluorescence microscopes provide white light – that is, a mixture of light of all colors of the rainbow. This type of light is called polychromatic. Also, the novel metal halide illumination systems produce white light. The intensity of the light from these light sources can be plotted against the wavelength. In xenon and metal halide illumination systems this plot shows a more or less linear or flat graph. In mercury illumination the intensity–wavelength plot shows typical peaks in intensity. These peaks are called 'mercury lines.' Fluorescence detection systems have traditionally been designed around such 'lines' – for example, the 365 nm or the 436 nm lines provided by an HBO lightbulb (more details later). As a consequence of the polychromatism of these sources of excitation light there is the necessity for appropriate filtering to select a narrow excitation wavelength band. Only a laser provides monochromatic light with a single color (i.e., an extremely narrow wavelength band).

High Quantum Efficiency with a Laser Excitation System

Because fluorescence-based laser scanning image acquisition systems have become widely available, we will, for the sake of simplicity, further deal with monochromatic light. An example is when an excitation spectrum shows a peak at, say, 488 nm: in this case, 488 nm light (blue light) is the best wavelength

to excite this fluorochrome. Excitation with light of this wavelength (blue) will get the maximum amount of fluorescence out of the fluorochrome (green light). Excitation of this fluorochrome with light of a different wavelength – for instance, red light (600 nm) – will not bring about much fluorescence. Thus, in order to make maximal use of a particular fluorochrome it is necessary to be familiar with its excitation spectrum. In general, the sharper and narrower the peak of its excitation spectrum, the better a fluorochrome fits its purpose. This especially holds for multicolor fluorescence applications and in the extreme for laser scanning fluorescence microscopy. In the latter application, excitation light is per definition monochromatic and because of this the laser wavelength should ideally match the peak in the absorption spectrum of the fluorochrome. If the laser light wavelength does not match this peak, most fluorochrome molecules will simply not be excited and weak or no fluorescence will appear. In this context the term 'quantum efficiency' is used. A quantum efficiency of 100% means 'one excitation photon in, one fluorescent photon out,' or, in other words, the maximal fluorescence output obtainable with an excitation light input with a given wavelength is close to 100% quantum efficiency.

Importance of Absorption and Emission Spectra in Double-Fluorescence Studies

Anyone interested in the simultaneous visualization of two fluorochromes – for instance, in a double-immunofluorescence procedure designed to test whether a given substance A is present in cells also containing substance B (so-called co-localization) – should be concerned about the spectral characteristics of the two selected fluorochromes to be associated with A and B. The excitation spectrum of fluorochrome A should not overlap that of fluorochrome B. If it does, then fluorochrome B may be excited and produce a signal in response to the light meant to excite fluorochrome A. The effect in this case is false positivity (cells with substance B erroneously observed as containing A). Similarly, the emission spectra of fluorochromes A and B should not overlap. When they do, A and B will appear with the same color and it will be impossible to discriminate between cells expressing A, cells expressing B, or cells containing both A and B.

The occurrence of observable fluorescence of fluorochrome B with the illumination and the filter set designed to highlight A, and, vice versa, of observable fluorescence of fluorochrome A with the illumination and the filter set designed to highlight B, is called cross-talk, bleeding through, or shining through. A notorious example of bleeding through occurs in many routine fluorescence microscopes when some fluorescence emitted by tetramethylrhodamine isothiocyanate (TRITC; red fluorescence) bleeds through the filter set used to observe fluorescein isothiocyanate (FITC; green fluorescence). One reason, apart from poorly designed filter sets, is that the excitation spectrum of FITC is quite wide. When it is absolutely necessary to avoid bleeding through, FITC could be replaced by a fluorochrome with a better (i.e., narrower) excitation spectrum. Alternatively or simultaneously, TRITC could be replaced by a fluorochrome with a peak in its excitation spectrum at a longer wavelength. Usually, FITC is replaced by Cy2 or Alexa Fluor 488. TRITC can be replaced by Texas Red or Alexa Fluor 594. An alternative way to reduce bleeding through is to 'play' with the filter sets or to use a special microscope (i.e., the confocal laser scanning microscope) that is equipped with one or more monochromatic light sources.

Shortpass, Bandpass, and Longpass Filters

With the exception of the confocal laser scanning microscope with its monochromatic lasers, the illumination light source in most fluorescence microscopes is a mercury lamp. This light source produces light ranging from UV to infrared (the lamp becomes extremely hot!). The emission spectrum of this light has several wavelength peaks that are referred to as 'excitation lines' (see earlier): 280, 313, 334, 365, 405, 435, 546, 577, and 579 nm. Some of these lines are exploited by filter manufacturers to design filter sets that optimally match the excitation spectrum of a particular fluorochrome. For instance, filter sets that employ the 546 nm line are used to excite TRITC, and filter sets that employ the 577 and 578 nm lines are used to excite Texas Red or Alexa Fluor 594. Note here that if one considers the purchase of a particular filter set, one should inform the manufacturer which lamp is used as the illumination source. Xenon metal halide and other metal halide lamps have emission spectra that area quite different than the mercury lamp emission spectra, and for this reason might yield disappointing results with a filter set designed around a specific mercury excitation line.

Filters may be used to filter away unwanted excitation wavelengths from the mercury lamp's light (such a filter is also called a 'barrier filter'), or they may be used to pass a selected portion of the emitted light to the eyes of the observer or to his camera. In everyday practice an excitation filter is always complemented with an associated emission filter. Filters that allow

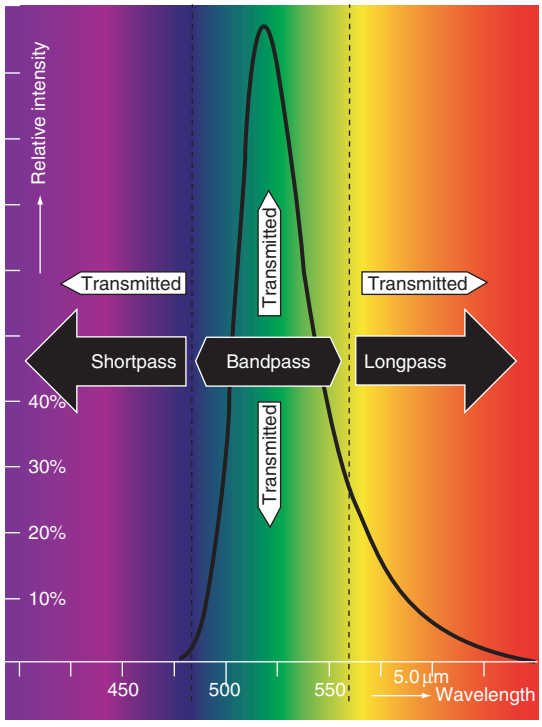

Figure 3 Depiction of shortpass, bandpass, and longpass filters. This fluorochrome (curve) reveals most of its fluorescence in the wavelength area covered by the bandpass filter.

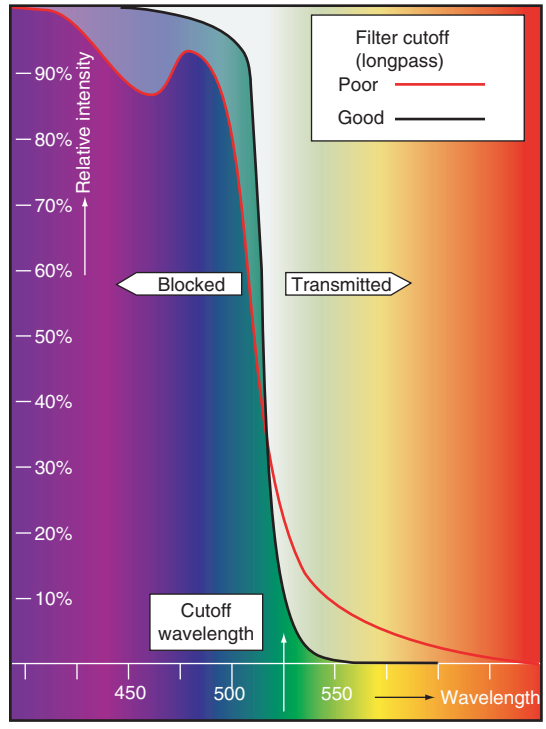

Figure 4 Cutoff wavelengths of a fluorescence filter, in this case a longpass filter. The black curve shows the cutoff of a nearly ideal filter, whereas the red curve belongs to a filter with poor cutoff characteristics. The latter filter allows passage of traces of fluorescence of wavelengths shorter than the official cutoff wavelength.

the light of the short-wavelength portion of an excitation (or emission) spectrum to pass, but block light of longer wavelengths, are called shortpass filters (Figure 3). Conversely, filters that block light of the short-wavelength portion of an excitation (or emission) spectrum and that allow light of the remaining, longer wavelengths to pass are called longpass filters (Figure 3). Filters that allow only a specific part of the light in a given spectrum to pass, blocking light with wavelengths shorter than a particular short wavelength and light with wavelengths longer than a particular long wavelength, are called bandpass filters (Figure 3). A typical and appropriate wavelength referred to in bandpass filters is the center wavelength, which defines the center of the wavelength band of light allowed to pass the filter. Another important parameter in bandpass filters is the bandwidth. Filter manufacturers provide tables that inform researchers about the wavelengths that matter for a particular filter set that is made for use with a particular fluorochrome (for instance, the wavelength at which a longpass filter begins transmitting light). Note that this wavelength, referred to as the cutoff wavelength, is in fact a short-wavelength trajectory, and the better (and often more expensive) the filter (Figure 4), the narrower this trajectory and the steeper the transmission curve in the cutoff trajectory.

There are several principles according to which filters can be manufactured, but this is outside the scope of this article.

Multiple-Bandpass Filters

A relatively new development is the appearance on the market of filter sets designed such that they allow the simultaneous observation of the fluorescence emitted by two or three different fluorochromes: these double- and triple-bandpass filters are also called combination filters. These filter sets can be very handy to screen quickly whether and where co-localization of fluorochromes occurs in tissues or cells. The classical method to study co-localization of fluorescent markers is to inspect an object switching back and forth between two different filter sets (each set specific for one of the fluorochromes). The disadvantage of the combination filter sets is that their design is always a compromise between specific demands set by the excitation/emission spectra of each of the individual fluorochromes. In fact, these are filters that allow a certain amount of cross-talk! Also, if for some reason fluorochrome A shines brighter than fluorochrome B, the fluorescence emitted by B might become overwhelmed by that of fluorochrome A.

Suitable Fluorochromes

In the past 30 years ever better fluorochromes have been developed, and each year new fluorochromes reach the market. The most recent catalog published by the leading company (see section on 'Relevant Websites'), currently in its tenth version, contains more than 3000 products, the majority of which are fluorescent substances. The basic demands underlying these developments are the quest for stable and specific fluorochromes that have a high photon efficiency, yet can be combined with bioactive molecules. Fluorochromes should comply with the following demands: chemical stability, photostability, sensitivity, high quantum yield at a particular excitation wavelength, a narrow excitation spectrum, and nontoxicity. They should possess the ability to be combined with antibodies, nucleic acids, agonists, or other reporter molecules used in fluorescence detection systems in biomedical applications. Some of the currently available fluorochromes are provided in Table 1. A detailed description of nearly all commercially available fluorochromes for cell biological purposes can be found in the catalog published on the Internet or in print by Invitrogen/Molecular Probes. The first fluorochrome to be applied in immunofluorescence, FITC, has long been surpassed by fluorescing substances with much better optical and chemical characteristics. FITC is so entrenched, however, notably in some pathology applications, that it is still being used.

Table 1 Excitation/emission peak wavelength of widely used fluorochromes

Traditional dyes		
Tetramethylrhodamine isothiocyanate (TRITC)[1]	550	573
Fluorescein isothiocyanate (FITC)[1]	494	520
Texas Red[1]	595	615
Alexa dyes		
Alexa Fluor 350[1]	342	441
Alexa Fluor 405[1]	401	422
Alexa Fluor 430[1]	431	540
Alexa Fluor 488[1]	492	520
Alexa Fluor 532[1]	532	553
Alexa Fluor 546[1]	557	572
Alexa Fluor 555[1]	553	568
Alexa Fluor 568[1]	577	603
Alexa Fluor 594[1]	594	618
Alexa Fluor 633[1]	632	649
Alexa Fluor 647[1]	650	668
Alexa Fluor 660[1]	664	689
Alexa Fluor 700[1]	696	719
Alexa Fluor 750[1]	752	779
Carbocyanine dyes		
Allophycocyanin[1]	651	661
Cy2[1]	489	506
Cy3[1]	554	566
Cy 3.5[1]	581	596
Cy5[1]	649	666
Cy5.5[1]	675	694
Cy7[1]	743	476
Green fluorescent protein dyes		
Cyan Fluorescent Protein[1] (CFP)	430	476
Enhanced Cyan Fluorescent Protein[1] (ECFP)	430	476
Yellow Fluorescent Protein[1] (YFP)	520	532
Enhanced Yellow Fluorescent Protein[1] (EYFP)	520	532
Traditional dyes		
wtGFP[1]	397	515
Other		
Acridine Orange	500	520
AMCA[1]	347	444
Amino-coumarin	346	442
Bodipy 630[1]	625	642
Bodipy 650[1]	646	660
Calcium Crimson[1]	589	615
Calcium Green[1]	506	531
Calcium Orange[1]	549	575
Cascade Blue[1]	398	420
Cascade Yellow[1]	399	552
Coumarin[1]	360	449
DiI[1]	551	565
DiO[1]	487	501
DiR[1]	750	779
Eosin[1]	525	546
Ethidium Bromide[1]	523	605
FluoroRuby	555	582
Fluoro-Emerald	495	524
Fura Red (low calcium)[1]	472	669
Fura Red (high calcium)[1]	436	657
Hydroxystilbamine (FluoroGold)	385	536
Indo-1 (Ca-free)[1]	346	478
Indo-1 (Ca-saturated)[1]	330	404
Lissamine rhodamine[1]	572	591
Lucifer Yellow CH[1]	428	536
Neuro Trace 500/525 green[1]	497	524
Oregon Green 488[1]	488	526
Oregon Green 514[1]	513	533
Phycoerythrin[1]	567	576
Propidium Iodide[1]	305	617
Rhodamine-123	505	533
Rhodamine Green[1]	504	532
Rhodamine Phallacidin[1]	558	575

Sources: [1]Obtained from Bio-Rad Laboratories; other fluorochromes obtained from Molecular Probes, Inc. Trademarks: Alexa Fluor, trademark of Molecular Probes, Inc., Eugene, OR; Cy, trademark of Biological Detection Systems, Pittsburgh, PA.

Quenching and Antiquenching Measures

Although fluorescence is a cyclical process, fluorochromes have a limited lifetime due to, for instance, chemical interaction of the fluorochrome with molecules in the immediate vicinity, exposure to light, or both. Anyone working with fluorescence is familiar with deterioration of fluorescence in the tissue while the material is being examined in the fluorescence

microscope. This phenomenon is referred to as quenching, photobleaching, or fading. Quenching may occur through exposure to daylight and can be accelerated by an increase in the temperature of the environment. Because of this, most researchers at least store their fluorescence material in the dark. The more prudent researcher stores the precious material in a refrigerator or a freezer. In our experience, the fluorescence in brain sections immunostained and detected with secondary antibodies or avidin conjugated to Cy-fluorochromes or Alexa Fluor fluorochromes, dehydrated, coverslipped, and stored in a freezer at –20 °C has remained remarkably stable over at least 5 years.

Manufacturers of fluorescence microscopes equip their top models with electronic shutters to limit the exposure of the specimen to high-intensity excitation light. Quenching can be abated by the addition to the mounting medium of antiquenching agents, such as *p*-phenylenediamine, or of a commercial antifading product.

Background and Autofluorescence

Unwanted, nonspecific fluorescence surrounding fluorescence-stained structures is usually referred to as background fluorescence. Causes of background staining might be, for instance, the unspecific binding of the fluorochrome or its vehicle to tissue components. A particular source of background is autofluorescence, which is the characteristic of biological tissue to fluoresce 'spontaneously' at wavelengths in the green visible light range (up to 500 nm excitation). Notorious autofluorescence is produced by pigment particles (e.g., lipofuscin). Autofluorescence can be aggravated by fixatives containing glutaraldehyde. Remedies can be to use fluorochromes that can be operated at higher wavelengths (e.g., Texas Red and Carbocyanine 5), or through the use of more specific secondary antibodies, improvements of the signal-to-noise ratio by signal enhancement, checking the pH of the mounting medium (the intensity of fluorescence is pH dependent), the omission of glutaraldehyde, and, finally, the use of the thinnest sections manageable. Lipofuscin autofluorescence can be reduced by specific measures.

First Success

The first great success of a fluorescence-based method in neuroscience occurred after the discovery that catecholamines in dried sections of brain tissue, when exposed to formaldehyde vapor, propagate the deposition of a fluorescing, cyclic product in cells and fibers. This treatment of brain tissue, known as the formaldehyde-induced fluorescence technique (FIF), named the Falck–Hillarp technique after its discoverers, offered neuroscientists for the first time in history an instrument to study brain neuronal systems belonging to a particular chemical family, the catecholaminergic neuron. The discovery that neurons in the substantia nigra are responsible for the dopaminergic innervation of the striatum is largely based on observations on FIF-treated brain sections. Thus, the FIF technique marks for the neuroanatomical field the beginning of the transition from purely descriptive neuroanatomy to the current era of function-related chemical and molecular neuroscience. Improvements further refined this method into a fluorescence tool that dominated neuroscience research until the advent of more sensitive methods based on immunofluorescence. Because of its limited sensitivity and its inability to distinguish among various monoamines, the FIF technique no longer plays a major role in neuroscience research.

Genetic Engineering

If it were possible to have available tissue in which cells of interest are intrinsically labeled with a fluorescent marker, while the remaining cells are not, then all attempts to stain with fluorochromes would be unnecessary or could be aimed at other interesting but non-autofluorescent cells. Such tissue can indeed be created by genetic engineering. Organisms (or cell cultures) exist into which a gene that codes for green fluorescent protein (GFP; a photostable autofluorescent protein from the jellyfish, *Aequorea victoria*) has been transfected into the genome of particular cells. As a consequence, these cells metabolize and accumulate GFP that fluoresces 'spontaneously' when the cells are illuminated with light of the excitation wavelength.

The excitation spectrum of GFP has an excitation peak at 498 nm. GFP is reportedly extremely photostable. Since the introduction of GFP, a whole range of GFP–look-alikes has been introduced, each with its own specific excitation/emission characteristics: there is yellow-GFP, cyan-GFP, and wild-type (wt)-GFP, and so forth (see Table 1). Specific types of neurons exclusively expressing GFP can be of great help to the anatomist for the unraveling of neuronal networks (Figure 5).

Applications

Applications of fluorescence in neuroscience are numerous, and the reader is encouraged to consult a vast literature on this subject. We mention here only some broad areas of application. In fixed tissue, immunofluorescence allows detection of neuroactive substances, neurofilaments, cell proteins, and gene

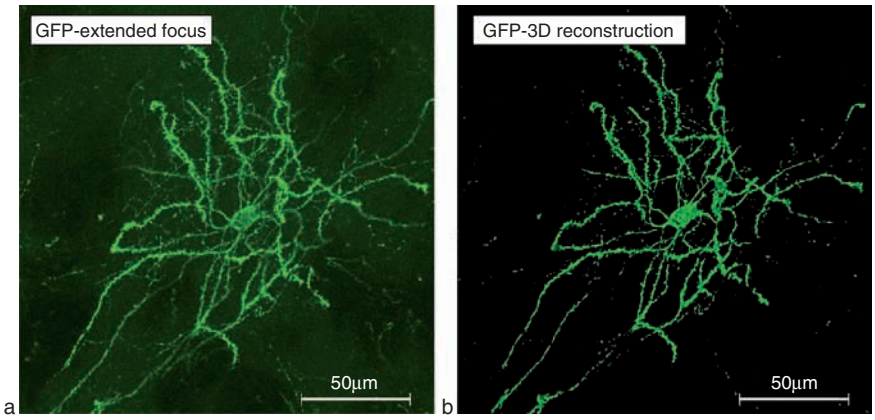

Figure 5 GFP expressing a medium-size spiny neuron in the striatum of a genetically engineered mouse. (a) Fluorescence image in a 125-µm-thick slice. Note that all cell components and processes contain GFP. Visualization with 488 nm excitation in a confocal laser scanning microscope. (b) A three-dimensional computer reconstruction of the same cell. The presence of such cells as isolated neurons in mouse strains has greatly facilitated anatomical neuronal network analysis. Mouse material was a gift by Christiaan Levelt (University of Amsterdam).

products (by a technique named fluorescence *in situ* hybridization (FISH)). Detection of tracers incorporated in tissue before the tissue was fixed (i.e., by the living organism, via introduced markers and transfected fluorescent markers) is another approach. In living tissue and cells, voltage-sensitive dyes, calcium imaging, uptake and transport phenomena, and photobleaching experiments (fluorescence resonance energy transfer (FRET)) may be used.

See also: Fluorescent Biomarkers in Neurons; Glial Ion Homeostasis: A Fluorescence Microscopy Approach.

Further Reading

Coons AH and Kaplan MH (1950) Localization of antigens in tissue cells. II: Improvement in a method in the detection of antigens by means of fluorescent antibody. *Journal of Experimental Medicine* 91: 1–14.

Falck B, Hillarp N-A, Thieme G, et al. (1962) Fluorescence of catecholamines and related compounds condensed with formaldehyde. *Journal of Histochemistry and Cytochemistry* 10: 348–354.

Haugland RP, Gregory J, Spence MTZ, et al. (2003) *Handbook of Fluorescent Probes and Research Products* (commercial product catalog). Eugene, OR: Molecular Probes, Inc.

Kiernan JA (1997) Making and using aqueous mounting media. *Microscopy Today* 10: 16–17.

Longin A, Souchier C, Ffrench M, et al. (1993) Comparison of antifading agents used in fluorescence microscopy: Image analysis and laser confocal microscopy study. *Journal of Histochemistry and Cytochemistry* 41: 1833–1840.

Moore RY (1992) Fluorescence histochemical methods: Neurotransmitter histochemistry. In: Heimer L and RoBards M (eds.) *Neuroanatomical Tract-Tracing Methods*, pp. 441–482. New York: Plenum.

Pawley JB (1995) *Handbook of Biological Confocal Microscopy*, 2nd edn. New York: Plenum.

Ploem JS (1967) The use of a vertical illuminator with interchangeable dichroic mirrors for fluorescence microscopy with incidental light. *Zeitschrift fuer Wissenschaftliche Mikroskopische* 68: 129–142.

Ploem JS (1971) A study of filters and light sources in immunofluorescence microscopy. In: Beatner EH (ed.) *Defined Immunofluorescent Staining*, pp. 414–429. New York: New York Academy of Sciences.

Reichman J (2000) *Handbook of Optical Filters for Fluorescence Microscopy*. Brattleboro, VT: Chroma Technology Corporation.

Schnell EA, Staines WA, and Wessendorf MW (1999) Reduction of lipofuscin-like autofluorescence in fluorescently labeled tissue. *Journal of Histochemistry and Cytochemistry* 47: 719–730.

Standard JM and Clark BK (1999) The Franck-Condon principle and condon parabolas in a physical chemistry or quantum physics course. *Journal of Chemical Education* 76: 1363–1366.

Sullivan KF and Kay SA (1998) *Green Fluorescent Proteins. Methods in Cell Biology*. New York: Academic Press.

Taylor DL and Salmon ED (1989) Basic fluorescence microscopy. In: Wang Y-L and Taylor DL (eds.) *Methods in Cell Biology, Volume 29, Fluorescence Microscopy of Living Cells in Culture, Part A*, pp. 207–237. London: Academic Press.

Wessendorf MW (1990) Characterization and use of multi-color fluorescence microscopic techniques. In: Björklund A, Hökfelt T, Wouterlood FG, et al. (eds.) *Analysis of Neuronal Microcircuits and Synaptic Interactions. Handbook of Chemical Neuroanatomy*, vol. 8, pp. 1–45. Amsterdam: Elsevier Science Publishers.

Wouterlood FG and Vinkenoog M (2003) Confocal laser scanning microscopy. In: Adelman G and Smith B (eds.) *Encyclopedia of Neuroscience*. Amsterdam: Elsevier.

Relevant Websites

http://www.chroma.com – Chroma Technology Corporation (filters).

http://micro.magnet.fsu.edu – Florida State University (FSU), National High Magnetic Field Laboratory (microscopy).

http://probes.invitrogen.com – Invitrogen Corporation (fluorochromes).

http://home.ncifcrf.gov – National Cancer Institute Experimental (NCI) Transplantation and Immunology Branch Flow

Cytometry Core Laboratory (fluorochrome excitation/emission spectra).

http://www.nikon-instruments.jp – Nikon Corporation Instruments Company (filters).

http://www.olympusfluoview.com – Olympus Corporation (filters).

http://www.semrock.com – Semrock Inc. (filters).

http://fluorescence.nexus-solutions.net – The Bio-Rad Fluorescence Database (fluorochrome excitation/emission spectra).

Fluorescent Biomarkers in Neurons

K N Fish, University of Pittsburgh, Pittsburgh, PA, USA
M W Davidson, Florida State University, Tallahassee, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

The Fluorescence Microscope

Long before fluorescence imaging become an essential tool in neurobiology, shaping the way we think about neurons and the brain, Camillo Golgi, Santiago Ramon y Cajal, and others used light-absorbing tissue stains coupled with bright-field microscopy to visualize the brain's architecture. Their ability to observe the elements of the nervous system led to the early development of modern neuroscience. The field of neuroscience is still shaped by our ability to visualize the structures and cells of the nervous system. Modern microscopes far exceed the design specifications of those made prior to the mid-1900s. Modern glass formulations are vastly improved, enabling greater correction for optical aberration than ever before, and synthetic antiglare lens coatings are now very advanced. Integrated circuit technology has allowed manufacturers to produce smart microscopes that incorporate microprocessors into the microscope stand. Today's microscopes incorporate multiple illuminators, analyzers and polarizers, differential interference contrast (DIC) prisms, fluorescence attachments, and phase-contrast capabilities. When combined with the existing array of fluorochromes, a fluorescence microscope can be used to identify cells and submicroscopic cellular components with a high degree of specificity amid nonfluorescing material (Figure 1). In fact, at its highest resolution, the fluorescence microscope is capable of revealing the presence of a single molecule. Through the use of multiple fluorescence labeling, different probes can simultaneously identify several target molecules. Although the fluorescence microscope cannot provide a spatial resolution below the diffraction limit of specific specimen features, the detection of fluorescing molecules below such limits is readily achieved. In this section, we briefly described the microscopy equipment regularly found in the laboratory of a neuroscientist.

Fundamentals of Excitation and Emission

Epifluorescence illumination is the overwhelming choice of techniques in modern microscopy. The basic function of a fluorescence microscope is to irradiate the specimen with a desired and specific band of wavelengths and then to separate the much weaker-emitted fluorescence from the excitation light. In a properly configured microscope, only the emission light should reach the eye or detector so that the resulting fluorescent structures are superimposed with high contrast against a very dark background. The limits of detection are generally governed by the darkness of the background, and the excitation light is typically several hundred thousand to a million times brighter than the emitted fluorescence.

The effective separation and detection of excitation and emission wavelengths are achieved in fluorescence microscopy through the proper selection of filters to block or pass specific wavelength bands in the ultraviolet, visible, and near-infrared spectral regions. Epifluorescence microscopes feature a vertical illuminator interposed between the observation viewing tubes and the nosepiece housing the objectives and have the light source positioned at one end and the filter cube turret at the other. The illuminator is designed to direct light on to the specimen by first passing the excitation light through the microscope objective on the way to the specimen and then using that same objective to capture the emitted fluorescence. For fluorescence microscopy, the light of a specific wavelength (or defined band of wavelengths), often in the ultraviolet, blue, or green regions of the visible spectrum, is produced by passing multispectral light from an arc-discharge lamp (usually a mercury or xenon burner) or other source through a wavelength-selective excitation filter. Wavelengths passed by the excitation filter reflect from the surface of a dichromatic mirror, through the microscope objective to bathe the specimen with intense light. The light that emits from the fluorescing specimen radiates spherically in all directions, regardless of the excitation light source direction. Some of the emitted light is gathered by the objective and is sent back to the dichromatic mirror. Because the emitted light consists of longer wavelengths than the excitation illumination, it is able to pass through the dichromatic mirror and up toward the observation tubes or electronic detector. Most of the scattered excitation light reaching the dichromatic mirror is reflected back toward the light source. Before the emitted fluorescence can reach the eyepiece or detector, it first passes through the emission filter. This filter blocks any residual excitation light and passes the desired longer emission wavelengths. The excitation and emission filters can either be placed in front of the light source and capture device (or eyepiece), respectively, or they can be incorporated into the filter cube that houses the dichromatic mirror.

As already discussed, the fluorophore emission intensity peak is usually lower in frequency and

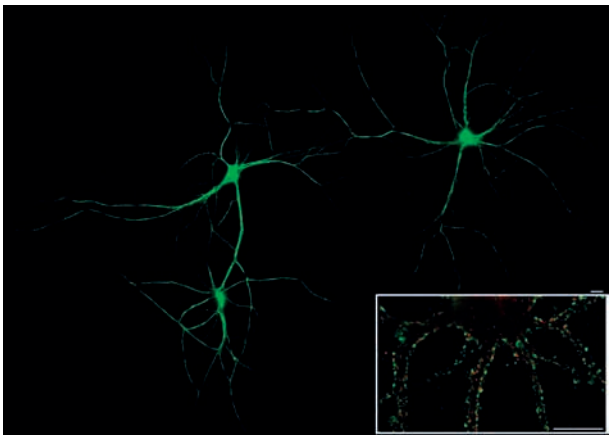

Figure 1 Using fluorescent microscopy to visualize cultured neurons. The dendrites of maturing primary rat hippocampal neurons in culture were imaged using a mouse anti-microtubule-associated protein (MAP)2 antibody and a goat antimouse secondary antibody conjugated to Alexa 488. The inset image is a higher magnification of synaptic proteins (green is synaptophysin; and red is postsynaptic density protein (PSD) 95) in mature cultured hippocampal neurons. Scale bar = 10 μm.

magnitude than that exhibited by the excitation peak. In addition, the emission spectral profile is often a mirror image (or nearly so) of the excitation curve but shifted to longer wavelengths. In order to achieve maximum fluorescence intensity, a fluorophore (often termed a dye) is usually excited at wavelengths near or at the peak of the excitation curve, and the widest possible range of emission wavelengths that include the emission peak are selected for detection. Perhaps the most important criteria, in view of relatively low fluorescence emission intensities, are that (1) the light source used for excitation be of sufficient brightness so that the weak emission light can be maximized and (2) the fluorochromes possess adequate absorption properties and emission quantum yields.

Fading: Photobleaching and Quenching

A wide spectrum of conditions often comes into play that ultimately affects the reradiation of fluorescence emission and thus reduces the intensity. The general term for a reduction of fluorescence emission intensity is fading, a catch-all category that is usually further subdivided into quenching and photobleaching phenomena. The excited state relaxation process of quenching results in reduced fluorescence intensity through a variety of mechanisms involving nonradiative energy loss and frequently occurs as a result of oxidizing agents or the presence of salts or heavy metals or halogen compounds. Photobleaching, on the other hand, is the irreversible decomposition of the fluorescent molecules in the excited state because of their interaction with molecular oxygen before emission.

Fluorescence Light Sources

An unfortunate consequence of low emission levels in most fluorescence microscopy applications is that the number of photons that reach the eye or camera detector is also very low. In most cases, the collection efficiency of optical microscopes is less than 30% and the concentration of many fluorophores in the optical path is in the micromolar or nanomolar range. In order to generate sufficient excitation light intensity to produce detectable emission, powerful compact light sources, such as high-energy short arc-discharge lamps, are necessary. The most common lamps are mercury burners, ranging in wattage from 50 to 200 W, and xenon burners, ranging from 75 to 150 W. These light sources are usually powered by an external direct current supply, furnishing enough startup power to ignite the burner through the ionization of the gaseous vapor and to keep it burning with a minimum of flicker for approximately 200–300 h. The mercury burners do not provide even intensity across the spectrum from ultraviolet to infrared, and much of the intensity of the lamp is expended in the near ultraviolet. Prominent peaks of intensity occur at 313, 334, 365, 406, 435, 546, and 578 nm. At other wavelengths in the visible light region, the intensity is steady, although not nearly so bright (but still usable in most applications). In considering illumination efficiency, mere lamp wattage is not the prime consideration. Instead, the mean luminance must be considered, which takes into account the source brightness, arc geometry, and the angular spread of emission.

In the past few years, optical microscopy has experienced an increase in the application of laser light sources, particularly the argon ion, argon–krypton ion, and a variety of helium–neon lasers. These lasers have the virtues of small source size, low divergence, near-monochromicity, and high mean luminance. Lasers have become essential for the development of certain techniques, such as scanning confocal microscopy, and a great deal of effort has been expended on developing new solid-state devices to replace the older gas lasers.

Objectives

Microscope objectives are perhaps the most important components of an optical microscope because they are responsible for primary image formation and play a central role in determining the quality of images that the microscope is capable of producing. Objectives are also instrumental in determining the

magnification of a particular specimen and the resolution under which fine specimen detail can be observed in the microscope. Modern objectives, composed of numerous internal glass lens elements, have reached a high state of quality and performance, with the extent of correction for aberrations and flatness of field determining the usefulness and cost of an objective. Construction techniques and materials used to manufacture objectives have greatly improved over the course of the past 100 years. Today, objectives are designed with the assistance of computer-aided-design systems, using advanced rare-element glass formulations of uniform composition and quality and having highly specific refractive indices. The enhanced performance that is demonstrated using these advanced techniques has allowed manufacturers to produce objectives that are very low in dispersion and corrected for most of the common optical artifacts such as coma, astigmatism, geometrical distortion, field curvature, and spherical and chromatic aberration. Not only are microscope objectives now corrected for more aberrations over wider fields, but image flare has been dramatically reduced with a substantial increase in light transmission, yielding images that are remarkably bright, sharp, and crisp.

Microscope Stage

Modern microscope stages are precisely manufactured and very durable. They are cast from iron, composite ceramics, or aluminum and machine finished to accurate tolerances. Translation mechanisms are usually rack-and-pinion styles, machined from aluminum, brass, or synthetic polymers. There are a wide variety of microscope stages that are designed for specific purposes. These include stages equipped with auxiliary equipment for manipulating samples during observation, measuring systems that allow precise measurements to be made over very small distances, universal stages that allow measurements over many specimen angles, and other specialized stages that perform a myriad of unique functions.

Microscope stages are classified according to design and functionality. In the simplest case, the plain stage consists of a rectangular or square design containing several clips to hold the specimen slide. Although manual movement of a specimen is very useful for quick examination of specimens, it is very difficult to use with higher-power objectives (above 20×). At high magnification, small translations of the specimen slide will usually throw the features of interest completely out of the view field, and attempting to relocate them can lead to frustration. Therefore, most stages can be fitted with the added feature of a specimen-slide translational control device, commonly referred to as a mechanical stage. Mechanical stages contain controls that allow the movement of the specimen slide in both the *X* (right and left) and *Y* (back and forth) directions in minute distances, making it easier for the microscopist to find specific areas on the slide.

Our ability to electronically control mechanical stages has improved tremendously over the last 15 years. These advances have been crucial in the development of the high-magnification measuring systems used in quantitative microscopy. The stages on these systems are equipped with electronically controlled micrometer-style translational devices that allow for very accurate and controlled movements of the stage and specimen, give accurate digital readouts of the stage position, and can be programmed for precise stage movements in a sequence of steps. There are a wide variety of designs for these specialized stages and many after-market manufacturers are capable of supplying custom-configured stages built to the customer's specifications.

Data Collection

Almost all microscopists, at some point, have the need or desire to record the images seen through the microscope. The use of photography to capture images in a microscope dates back to the invention of the photographic process. Early photomicrographs were remarkable for their quality, but the techniques were laborious and burdened with long exposures and a difficult process for developing emulsion plates and films. The primary medium for photomicrography was film until the past decade, when improvements in electronic camera and computer technology made digital imaging cheaper and easier to use than conventional photography. The imaging device is one of the most critical components in fluorescence microscopy because it determines at what level specimen fluorescence may be detected, the relevant structures resolved, and/or the dynamics of a process visualized and recorded. The problems of photodestruction of fluorochromes in the presence of oxygen (photobleaching) and limitations on the numbers of fluorochromes that can be involved in a single region make it essential that a sensitive electronic detector is used in fluorescence microscopy. In addition, only approximately 5–10% of the emitted light from an excited fluorochrome is collected and transferred to the sensor in a typical epifluorescence microscope. There are two approaches to capturing as much of this limited light flux as possible: integration by a slow-scan charge-coupled device (CCD) and image intensification and capture on a video-rate or progressive-scan CCD camera. The general finding is

that a cooled, slow-scan CCD camera always produces a higher signal-to-noise ratio than an intensified CCD, provided sufficient integration time is available. The CCD is presently the most widely used detector for fluorescence microscopy. Emerging electron-multiplying CCD technology is opening the door to even more sensitive detectors and may become a primary component in the microscopist's toolbox in the next few years.

Fluorescent Probes

Many fluorescent probes are constructed around synthetic aromatic organic chemicals designed to bind with a biological macromolecule (e.g., a protein or nucleic acid) or to localize within a specific structural region, such as the cytoskeleton, mitochondria, Golgi apparatus, endoplasmic reticulum, or nucleus. Other probes, both synthetic and those derived from naturally occurring proteins, are employed to monitor dynamic processes and localized environmental variables, including concentrations of inorganic metallic ions, pH, reactive oxygen species, and membrane potential. Fluorescent dyes are also useful in monitoring cellular integrity (live vs. dead and apoptosis), endocytosis, exocytosis, membrane fluidity, protein trafficking, signal transduction, and enzymatic activity. As such, the investigation of living and fixed neurons has benefited greatly from the evolution of fluorescent biomarkers. Although many of the available fluorescent reagents are used in the study of neurons, we describe here only the major groups.

Classic Synthetic Fluorescent Probes

Fluorochromes have been of great value in the study of both neurons in culture and tissue. These stains (also termed fluorophores) are excited by specific wavelengths of irradiating light and emit light of defined and useful intensity. Importantly, modern fluorochromes have a significant quantum yield (the ratio of photon absorption to emission). Fluorochromes were introduced to microscopy in the early twentieth century as vital stains for bacteria, protozoa, and trypanosomes, but did not see widespread use until the 1920s when fluorescence microscopy was first used to study dye binding in fixed tissues and living cells. Although these reagents proved to be useful, it was not until the early 1940s that Albert Coons developed a technique for labeling antibodies with fluorescent dyes, which gave rise to the field of immunofluorescence. Over the past 60 years, advances in immunology and molecular biology have produced a wide spectrum of secondary antibodies and provided insights into the molecular design of fluorescent probes targeted at specific regions within macromolecular complexes. Surprisingly, the advances made in fluorescent dye synthesis during the past few decades have provided very little insight into molecular design rules for developing new fluorochromes.

Many of the classic fluorescent probes, including fluorescein isothiocyanate, Lissamine rhodamine, and Texas red, have been successfully used to investigate neurons for many years in wide-field fluorescence. Although some of these fluorochromes (especially fluorescein isothiocynate) have been replaced with newly developed synthetic analogs, others, such as tetramethyl rhodamine and the isothiocyanate derivative, are still frequently employed in multiple-labeling investigations in microscopy due to their efficient excitation by the 546 nm spectral line from mercury arc-discharge lamps and 543 nm gas lasers.

The Birth of a New Generation of Synthetic Fluorescent Probes

The dramatic advances in modern fluorophore technology are exemplified by the Alexa Fluor dyes introduced by Molecular Probes. These sulfonated rhodamine derivatives exhibit higher quantum yields for more intense fluorescence emission than spectrally similar probes, and they have several additional improved features, including enhanced photostability, absorption spectra matched to common laser lines, pH insensitivity, and a high degree of water solubility.

The popular family of cyanine dyes, Cy2, Cy3, Cy5, Cy7, and their derivatives, exhibit fluorescence excitation and emission profiles that are similar to many of the traditional dyes, such as fluorescein and tetramethylrhodamine, but with enhanced water solubility, photostability, and higher quantum yields. Most of the cyanine dyes are more environmentally stable than their traditional counterparts, rendering their fluorescence emission intensity less sensitive to pH and organic mounting media. In a manner similar to the Alexa Fluors, the excitation wavelengths of the Cy series of synthetic dyes are tuned specifically for use with common laser and arc-discharge sources, and the fluorescence emission can be detected with traditional filter combinations.

In addition to the new generation of fluorochromes just described, nanometer-size crystals of purified semiconductors, principally cadmium selenide, known as quantum dots, are emerging as a potentially useful fluorescent labeling agent for living and fixed neurons in both traditional wide-field and laser scanning confocal fluorescence microscopy. Recently introduced techniques enable the purified tiny semiconductor crystals to be coated with a hydrophilic polymer shell and conjugated to antibodies or

other biologically active peptides and carbohydrates for application in many of the classic immunocytochemistry protocols, as well as tracking agents in live-cell imaging. These probes have significant benefits over organic dyes and fluorescent proteins, including long-term photostability, high fluorescence-intensity levels, and multiple colors with single-wavelength excitation for all emission profiles. In addition, the fluorescence emission intensity is confined to a symmetrical peak with a maximum wavelength that is dependent on the dot size but independent of the excitation wavelength. As a result, the same emission profile is observed regardless of whether the quantum dot is excited at 300, 400, 500, or 600 nm, but the fluorescence intensity increases dramatically at shorter excitation wavelengths.

Directly Targeted Fluorophores

DNA and chromatin in neurons can be stained with dyes that bind externally to the double helix. The most popular fluorochromes in this category are 4′,6-diamidino-2-phenylindole and the bisbenzimide Hoechst dyes that are designated by the numbers 33 258, 33 342, and 34 580. These probes are quite water-soluble and bind externally to AT-rich base pair clusters in the minor groove of double-stranded DNA with a dramatic increase in fluorescence intensity.

Fluorophores targeted at specific intracellular organelles, such as the mitochondria, lysosomes, Golgi apparatus, and endoplasmic reticulum, are useful for monitoring a variety of biological processes in living and fixed neurons. In general, organelle probes consist of a fluorochrome nucleus attached to a target-specific moiety that assists in localizing the fluorophore through covalent, electrostatic, hydrophobic, or similar types of bonds. Many of the fluorescent probes designed for selecting organelles are able to permeate or sequester within the cell membrane (and, therefore, are useful in living cells), whereas others must be installed using monoclonal antibodies with traditional immunocytochemistry techniques.

Fluorophores Used to Sense the Intracellular Neuronal Environment

Fluorophores designed to probe the internal environment of living neurons have been widely examined by a number of investigators, and many hundreds have been developed to monitor effects such as localized concentrations of alkali and alkaline earth metals, heavy metals (employed biochemically as enzyme cofactors), inorganic ions, thiols and sulfides, and nitrite; pH, solvent polarity, and membrane potential. Ca^{2+} is a metabolically important ion in neurons that plays a vital role in cellular response to many forms of external stimuli. Because transient fluctuations in Ca^{2+} ion concentration are typically involved when cells undergo a response, fluorophores must be designed not only to measure localized concentrations within segregated compartments but also to produce quantitative changes when flux density waves progress throughout the entire cytoplasm. Many of the synthetic molecules designed to measure Ca^{2+} levels are based on the nonfluorescent chelation agents ethylene glycol tetraacetic acid (EGTA) and 1,2-bis(*o*-aminophenoxy)ethane-*N*,*N*,*N*′,*N*′-tetraacetic acid (BAPTA), which have been used for years to sequester Ca^{2+} ions in buffer solutions. Several of the most common Ca^{2+} probes are the ratiometric indicators fura-2, fluo-3, and indo-1. Important fluorescence monitors for intracellular pH include a pyrene derivative known as 8-hydroxypyrene-1,3,6-trisulfonic acid, trisodium salt (HPTS) or pyranine; the fluorescein derivative (2-carboxy ethyl)-5-(and 6)-carboxy fluorescein (BCECF); and another substituted xanthene termed carboxy seminaphtorhodafluor (SNARF)-1.

Fluorescent Proteins

Fluorescent probe technology and cell biology (including neurobiology) have been dramatically altered over the past few years by the discovery and development of naturally occurring fluorescent proteins (Table 1). Mutated derivatives of these proteins have rapidly advanced to center stage in the investigation of a wide range of intracellular processes in living organisms. These biological probes have provided scientists with the ability to visualize, monitor, and track individual molecules with high spatial and temporal resolution in both steady-state and kinetic experiments. A variety of marine organisms have been the source of more than 100 fluorescent proteins and their analogs, which arm the investigator with a balanced palette of noninvasive biological probes for single-, dual-, and multispectral fluorescence analysis. Among the advantages of fluorescent proteins over the traditional organic and new semiconductor probes already described is their response to a wider variety of biological events and signals. Coupled with their ability to specifically target fluorescent probes to subcellular compartments, their extremely low or absent photodynamic toxicity, and their widespread compatibility with tissues and intact organisms, these biological macromolecules offer an exciting new frontier in live-cell imaging.

The first member of this series to be discovered, green fluorescent protein (GFP), was isolated from the North Atlantic jellyfish, *Aequorea victoria*, and found to exhibit a high degree of fluorescence without

Table 1 A compilation of properties displayed by several of the most popular and useful fluorescent protein variants

Protein (acronym)	*Ex (nm)*	*Em (nm)*	*EC* $\times 10^{-3}$	*QY*	*In vivo structure*	*Relative brightness (% of EGFP)*
Blue fluorescent proteins						
Azurite	384	450	26.2	0.55	Monomer*	43
EPFP2	383	448	32.0	0.56	Monomer*	53
Cyan fluorescent proteins						
ECFP	439	476	32.5	0.40	Monomer*	39
mECFP	433	475	32.5	0.40	Monomer	39
mCerulean	433	475	43.0	0.62	Monomer*	79
CyPet	435	477	35.0	0.51	Monomer*	53
AmCyan1	458	489	44.0	0.24	Tetramer	31
Midoriishi Cyan	472	495	27.3	0.90	Dimer	73
mTFP1	462	492	64	0.85	Monomer	162
Green fluorescent proteins						
GFP	484	507	56.0	0.60	Monomer*	100
Emerald	487	509	57.5	0.68	Monomer*	116
Azami Green	492	505	55.0	0.74	Monomer	121
ZsGreen	493	505	43.0	0.91	Tetramer	117
Superfolder GFP	485	510	83.3	0.65	Monomer*	160
T-Sapphire	399	511	44.0	0.60	Monomer*	79
Yellow fluorescent proteins						
EYFP	514	527	83.4	0.61	Monomer*	151
Topaz	514	527	94.5	0.60	Monomer*	169
Venus	515	528	92.2	0.57	Monomer*	156
mCitrine	516	529	77.0	0.76	Monomer	174
YPet	517	530	104	0.77	Monomer*	238
ZsYellow1	529	539	20.2	0.42	Tetramer	25
mBanana	540	553	6.00	0.70	Monomer	13
Orange and red fluorescent proteins						
Kusabira Orange	548	559	51.6	0.60	Monomer	92
mOrange	548	562	71.0	0.69	Monomer	146
dTomato	554	581	69.0	0.69	Dimer	142
dTomato-Tandem	554	581	138	0.69	Pseudomonomer	283
DsRed	558	583	75.0	0.79	Tetramer	176
DsRed2	563	582	43.8	0.55	Tetramer	72
DsRed-Express (T1)	555	584	38.0	0.51	Tetramer	58
DsRed-Monomer	556	586	35.0	0.10	Monomer	10
TagRFP	555	584	100.0	0.48	Monomer	142
mStrawberry	574	596	90.0	0.29	Monomer	78
AsRed2	576	592	56.2	0.05	Tetramer	8
mRFP1	584	607	50.0	0.25	Monomer	37
JRed	584	610	44.0	0.20	Dimer	26
mCherry	587	610	72.0	0.22	Monomer	47
HcRed1	588	618	20.0	0.015	Dimer	1
mRaspberry	598	625	86.0	0.15	Monomer	38
mKate	588	635	45.0	0.33	Monomer	44
HcRed-Tandem	590	637	160	0.04	Pseudomonomer	19
mPlum	590	649	41.0	0.10	Monomer	12

Along with the common name and/or acronym for each fluorescent protein, the peak excitation (Ex) and emission (Em) wavelengths, molar extinction coefficient (EC), quantum yield (QY), relative brightness, and *in vivo* structural associations are listed (*signifies a weak dimer). The computed brightness values were derived from the product of the molar extinction coefficient and quantum yield, divided by the value for EGFP. This listing was created from scientific and commercial literature resources and is not intended to be comprehensive, but instead represents fluorescent protein derivatives that have received considerable attention in the literature and may prove valuable in research efforts. The excitation and emission peak values listed may vary in published reports due to the broad spectral profiles. In actual fluorescence microscopy investigations, spectral profiles and wavelength maxima are influenced by environmental effects, such as pH, ionic concentration, and solvent polarity, as well as fluctuations in localized probe concentration. Therefore, the listed extinction coefficients and quantum yields may differ from those actually observed under experimental conditions.

the aid of additional substrates or coenzymes. In native GFP, the fluorescent moiety is a cyclized tripeptide derivative of serine, tyrosine, and glycine that requires molecular oxygen for activation, but no additional cofactors or enzymes. Subsequent investigations revealed that the GFP gene could be expressed in other organisms, including mammals, to yield fully functional analogs that display no adverse biological effects. In fact, fluorescent proteins can be fused to virtually any protein in living cells using recombinant complementary DNA cloning technology, and the resulting fusion protein gene product can be expressed in cultured neurons and transgenic animals. Lack of a need for cell-specific activation cofactors renders the fluorescent proteins much more useful as generalized probes than other biological macromolecules, such as the phycobiliproteins, which require the insertion of accessory pigments in order to produce fluorescence.

Mutagenesis experiments with GFP have produced a large number of variants with improved folding and expression characteristics, which have eliminated wild-type dimerization artifacts and fine-tuned the absorption and fluorescence properties. These studies have uncovered GFP variants that exhibit a variety of absorption and emission characteristics across the entire visible spectral region, which has enabled researchers to develop probe combinations for simultaneous observation of two or more distinct fluorescent proteins in a single organism. Several of these derivatives, known as biosensors, are useful in resonance energy-transfer experiments and other investigations that rely on multicolor fluorescence imaging.

Attempts to shift the absorption and emission spectra of *Aequorea victoria* fluorescent proteins to wavelengths in the orange and red regions of the spectrum have met with little success. However, fluorescent proteins from other marine species have enabled investigators to extend the available spectral regions to well within the red and near-infrared wavelength range. Newly developed photoactivation fluorescent proteins, including photoactivatable GFP, Kaede, and Eos, exhibit dramatic emission wavelength shifts on irradiation with ultraviolet light. These probes should prove useful in live-cell neuron imaging studies involving selective irradiation of specific target regions and the subsequent kinetic analysis of diffusional mobility and compartmental residency time of fusion proteins.

An emerging set of tools for the neuroscientist's fluorescent probe palette is hybrid systems designed around small molecules that can be covalently targeted to genetically specified proteins within or on the surface of living cells. The most promising system is a tetracysteine-biarsenical system that requires modification of the target protein using a 12-residue peptide sequence containing four cysteine residues. When membrane-permeable biarsenical molecules, known as FlAsH and ReAsH, interact with the modified target, fluorescence can be detected at picomolar levels.

Advanced Microscopy Techniques

Neurons are polarized cells that have three morphologically and biochemically distinct compartments: the soma, axon, and dendrites. It is the existence of these subregions and their unique characteristics that make studying protein synthesis, assembly, trafficking, and function in neurons using some of the more traditional approaches difficult. Therefore, neuroscientists have taken advantage of advanced microscopy techniques to perform biochemical-like studies in subneuronal compartments. In this section, we discuss the techniques that make this possible.

Fluorescence Recovery after Photobleaching

The phenomenon of photobleaching is exploited in a technique known as fluorescence recovery after photobleaching (FRAP), a very useful mechanism for investigating the diffusion and motion of biological macromolecules. The method is based on photobleaching a fluorescently tagged molecule in a precisely defined region of the specimen by an intense burst of laser light, accompanied by the subsequent observation of the rates and pattern of fluorescence recovery in the photobleached area. A related technique, known as fluorescence loss in photobleaching (FLIP), is employed to monitor the decrease of fluorescence in a defined region lying adjacent to a photobleached area. Similar to FRAP, this technique is useful in the investigation of molecular mobility and dynamics in living cells.

Fluorescence Resonance Energy Transfer

In some cases, fluorescence quenching results from the transfer of energy to another molecule (termed the acceptor; e.g., EYFP), which resides physically close to the excited fluorophore (the donor; e.g., ECFP), a phenomenon known as fluorescence resonance energy transfer (FRET). This particular mechanism has become the basis for a useful technique involving the study of molecular interactions and associations at distances far below the lateral resolution of the optical microscope. FRET is useful for obtaining quantitative temporal and spatial information about the binding and interactions of proteins, lipids, enzymes, and nucleic acids in living cells. The technique is performed using either steady-state or time-resolved

methodology, but time-resolved FRET imaging has the advantage of more accurately mapping the donor–acceptor distance.

Optical Highlighter Fluorescent Proteins

Protein chromophores that can be activated to initiate fluorescence emission from a quiescent state (a process known as photoactivation) or that are capable of being optically converted from one fluorescence emission bandwidth to another (photoconversion) represent perhaps the most promising approach to the *in vivo* investigation of protein lifetimes, transport, and turnover rates in neurons. Appropriately termed molecular or optical highlighters, photoactivated fluorescent proteins generally display little or no initial fluorescence under excitation at the imaging wavelength, but dramatically increase their fluorescence intensity after activation by irradiation at a different (usually lower) wavelength. Photoconversion optical highlighters, on the other hand, undergo a change in the fluorescence emission bandwidth profile on optically induced changes to the chromophore. These effects result in the direct and controlled highlighting of distinct molecular pools within the cell.

The ability to selectively initiate or alter fluorescence emission profiles in photoconversion optical highlighter proteins renders these probes excellent tools for exploring protein behavior in living cells. Because the fluorescence intensity (or color spectrum) of highlighters occurs only after photon-mediated conversion, newly synthesized non-photo-activated protein pools remain unobserved and do not complicate experimental results. This signal independence from new protein synthesis could potentially enable the study of protein degradation kinetics in tagged molecules by techniques such as optical pulse labeling and monitoring of the fluorescence over time. Additional quantitative techniques, including fluorescence-correlation spectroscopy, should prove useful in measuring the mobility of photoactivated optical highlighters in small numbers, even down to the single-molecule level.

Photoactivation and photoconversion of optical highlighters are generally rapid and produce stable fluorescence signals. Therefore, these probes can be used to examine a variety of kinetic properties of tagged proteins, such as their diffusion coefficients, fractional mobility, compartmental residency time, and exchange rates. In many cases, photoconversion offers greater flexibility than the traditional photobleaching techniques of FRAP and FLIP. In addition, cell lineage or translocation in developing organisms can be easily monitored by imaging the fluorescence dispersion after photoactivation of a single cell or subpopulation of cells. In these regards, optical highlighters demonstrate remarkable promise to complement and extend the range of present fluorescent protein-imaging applications.

Fluorescence Deconvolution Microscopy

An intrinsic problem exists in wide-field fluorescence microscopy – blur arising from fluorescently stained structures removed from the point of focus hinders the visualization of information in the focal plane. Deconvolution is the computational process that reduces out-of-focus blur. Deconvolution microscopy has become the preferred method for imaging weak signals, particularly in live imaging of events in culture cells, because CCDs have much higher quantum efficiency than the photomultipliers used in confocal microscopy. The reason why this technique is usually limited to imaging culture cells and relatively thin tissue specimens is because all the algorithms used to perform deconvolution have the potential to amplify input signal noise. However, when used correctly, the application of deconvolution algorithms to image stacks made up of several consecutive *Z*-axis image planes results in images that have substantially improved contrast and resolution.

Confocal Microscopy

Laser scanning confocal microscopy represents one of the most significant advances in optical microscopy ever developed, primarily because the technique enables visualization deep within both living and fixed cells and tissues and because it affords the ability to collect sharply defined optical sections from which three-dimensional renderings can be created (Figure 2). Confocal microscopy offers several advantages over conventional wide-field optical microscopy, including the ability to control depth of field, the elimination or reduction of background information away from the focal plane (that leads to image degradation), and the capability of collecting serial optical sections from thick specimens. The basic key to the confocal approach is the use of spatial filtering techniques to eliminate out-of-focus light or glare in specimens whose thickness exceeds the immediate plane of focus. There has been a tremendous explosion in the popularity of confocal microscopy in recent years, due in part to the relative ease with which extremely high-quality images can be obtained from specimens prepared for conventional fluorescence microscopy and to the growing number of applications in cell biology that rely on imaging both fixed and living cells and tissues. In fact, confocal technology is proving to be one of the most important advances ever achieved in optical microscopy.

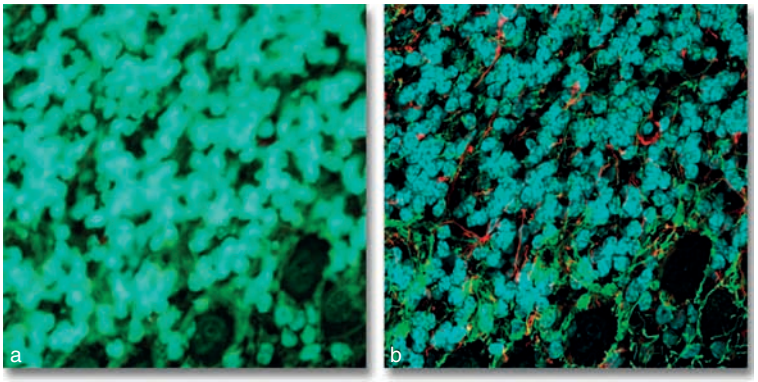

Figure 2 Rat brain hippocampus section (30 μm thick): (a) wide-field fluorescence image; (b) laser scanning confocal image. The tissue was stained with triple immunofluorescence using primary antibodies to histones (cyan), heavy neurofilaments (green), and glial fibrillary acidic protein (GFAP; red).

In modern confocal microscopes, two fundamentally different techniques for beam scanning have been developed. Single-beam scanning, one of the more popular methods employed in a majority of the commercial laser scanning microscopes, uses a pair of computer-controlled galvanometer mirrors to scan the specimen in a raster pattern at a rate of approximately one frame per second. Faster scanning rates (to near video speed) can be achieved using acousto-optic devices or oscillating mirrors. In contrast, multiple-beam scanning confocal microscopes are equipped with a spinning disk containing an array of slits or pinholes coupled to microlenses. These instruments often use arc-discharge lamps for illumination, instead of lasers, to reduce specimen damage and enhance the detection of low fluorescence levels during real-time image collection. Another important feature of the multiple-beam microscopes is their ability to readily capture images with an array detector, such as a CCD camera system.

The broad range of applications available to laser scanning and spinning-disk confocal microscopy includes a wide variety of studies in neuroanatomy and neurophysiology, as well as morphological studies of a wide spectrum of cells and tissues. In addition, the growing use of new fluorescent proteins is rapidly expanding the number of original research reports coupling these useful tools to modern microscopic investigations. Other applications include resonance energy transfer, stem-cell research, photobleaching studies, lifetime imaging, multiphoton microscopy, total internal reflection, DNA hybridization, membrane and ion probes, bioluminescent proteins, and epitope tagging.

Multiphoton Fluorescence Microscopy

Multiphoton fluorescence microscopy is a powerful research tool that combines the advanced optical techniques of laser scanning microscopy with long-wavelength nonlinear multiphoton fluorescence excitation to capture high-resolution, three-dimensional images of specimens tagged with highly specific fluorophores.

The methodology is particularly useful to neurobiologists who endeavor to study dynamic processes in living brain tissues without inflicting significant, and often lethal, damage to the specimen. Excitation in multiphoton microscopy occurs only at the focal point of a diffraction-limited microscope, providing the ability to optically section thick biological specimens to obtain three-dimensional resolution. Because the position of the focal point can be accurately determined and controlled, multiphoton fluorescence is useful for probing selected regions beneath the specimen surface. The highly localized excitation energy serves to minimize the photobleaching of fluorophores attached to the specimen and to reduce photodamage, which increases cell viability and the subsequent duration of experiments that investigate the properties of living cells. In addition, the application of near-infrared excitation wavelengths permit deeper penetration into biological materials and reduces the high degree of light scattering that is observed at shorter wavelengths. These advantages enable researchers to conduct experiments on thick living-tissue samples, such as brain slices and developing embryos, that would be difficult, if not impossible, to image with other microscopy techniques.

A major impediment to research in multiphoton microscopy is the high cost of the equipment, in particular the necessary mode-locked pulsed laser systems that are required for two-photon and three-photon excitation. Typically, Ti:sapphire lasers, which have a wavelength tunability of 700–1100 nm, are used for multiphoton imaging of fluorescent probes. As newer, cheaper, and user-friendly lasers are introduced into this competitive market, more commercial multiphoton

microscope systems will be introduced at a cheaper price. This availability, coupled to the fact that there are no physical limitations on the implementation of multiphoton fluorescence, will encourage the widespread application of this technique throughout the biological sciences.

Detecting Single Molecules Using Total Internal Reflection Microscopy

Under ideal conditions, it is often possible to detect the fluorescence emission from a single molecule, provided that the optical background and detector noise are sufficiently low. A single fluorescein molecule emits as many as 300 000 photons before it is destroyed by photobleaching. Assuming a 20% collection and detection efficiency, approximately 60 000 photons would be detected. Using avalanche photodiode or electron-multiplying CCD detectors for these experiments, investigators have been able to monitor the behavior of single molecules for many seconds and even minutes. The major problem is adequate suppression of the optical background noise. Fluorescence microscopy techniques using total internal reflection provide the desired combination of low background and high excitation light flux.

Total internal reflection fluorescence microscopy (TIRFM) takes advantage of the evanescent wave that is developed when light is totally internally reflected at the interface between two media having dissimilar refractive indices. In this technique, a beam of light (usually an expanded laser beam) is directed through a prism of high refractive index, such as glass or sapphire, which abuts a lower refractive index medium of glass or aqueous solution. If the light is directed into the prism at a higher than the critical angle, the beam will be totally internally reflected at the interface. The reflection phenomenon develops an evanescent wave at the interface by the generation of an electromagnetic field that permeates approximately 200 nm into the lower refractive index space. The light intensity in the evanescent wave is sufficiently high to excite the fluorophores within it, but because of its shallow depth, the volume excited is very small. The result is an extremely low-level background because so little of the specimen is exposed to the excitation light.

TIRFM can also be conducted through a modification of the epi-illumination approach used in wide-field techniques. This method requires a very high numerical aperture objective (at least 1.4, but preferably 1.45–1.6) and partial illumination of the microscope field from one side by a small sport or more uniform illumination by a thin annulus. High-refractive-index lens immersion medium and microscope cover glass are required to achieve the illumination angle resulting in total internal reflection. Light rays exiting the objective front lens element at an angle less than the critical angle are transmitted away from the microscope. When the angle is increased to or beyond the critical angle, total internal reflection results.

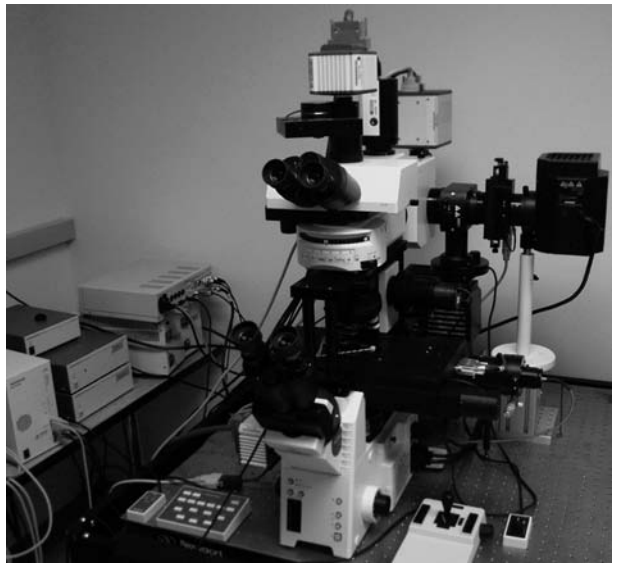

Figure 3 The newly developed up/down microscope. This advanced microscope is capable of simultaneously performing both total internal reflection fluorescence microscopy and epifluorescence or confocal microscopy. The upper part of the system is equipped with an Olympus digital server unit confocal, a Hamamatsu ORCA II ERG camera, a QICAM-IR, and a Hamamatsu electron-multiplying charge-coupled device camera. The bottom half of the system is a multi-laser-line total internal reflection fluorescence microscope equipped with a Hamamatsu ORCA II ERG camera. Because two different microscope heads are used, equipped with independent Z-axis motors, simultaneous imaging can be performed on the exact same focal plane or on different planes: distinctive *Z*-axis, *XY*-plane, or both positions.

Other popular advanced fluorescence techniques, such as FRET and FRAP, are often combined with TIRFM to achieve additional information. In addition, microscopes that are capable of simultaneously performing both TIRFM and epifluorescence or confocal microscopy are now being developed (Figure 3). The result is a very powerful tool for the study of individual fluorophores and fluorescently labeled molecules. The advantages resulting from the study of the properties of single molecules are only beginning to be appreciated.

Conclusion

The modern fluorescence microscope combines the power of high-performance optical components with computerized control of the instrument and digital image acquisition to achieve a level of sophistication

that far exceeds that of simple observation by the human eye. Microscopy now depends heavily on electronic imaging to rapidly acquire information at low light levels or at visually undetectable wavelengths. These technical improvements are not mere window dressing but, rather, are essential components of the light microscope as a system.

The era when optical microscopy was purely a descriptive instrument or an intellectual toy is past. At present, optical image formation is only the first step toward data analysis. The microscope accomplishes this first step in conjunction with electronic detectors, image processors, and display devices that can be viewed as extensions of the imaging system. Computerized control of focus, stage position, optical components, shutters, filters, and detectors is in widespread use and enables experimental manipulations that were not humanly possible with mechanical microscopes. The increasing application of electro-optics in fluorescence microscopy has led to the development of optical tweezers capable of manipulating subcellular structures or particles, the imaging of single molecules, and a wide range of sophisticated spectroscopic applications. In conclusion, the current range of optical microscopy now extends from the single molecule to the entire animal.

See also: Cajal's Place in the History of Neuroscience; Fluorescence Microscopy in the Neurosciences; Glial Ion Homeostasis: A Fluorescence Microscopy Approach; History of Neuroscience: Early Neuroscience; Imaging Studies Using Reporter-Gene Transgenic Rats.

Further Reading

Arndt-Jovin DJ and Jovin TM (1989) Fluorescence labeling and microscopy of DNA. *Methods in Cell Biology* 30: 417–448.

Axelrod D, Thompson NL, and Burghardt T (1983) Total internal inflection fluorescent microscopy. *Journal of Microscopy* 129 (part 1): 19–28.

Chalfie M and Kalin SR (eds.) (2006) *Green Fluorescent Protein: Properties, Applications, and Protocols*. New York: Wiley-Interscience.

Giepmans BN, Adams SR, Ellisman MH, and Tsien RY (2006) The fluorescent toolbox for assessing protein location and function. *Science* 312(5771): 217–224.

Haugland RP (2005) *The Handbook: A Guide to Fluorescent Probes and Labeling Technologies*. Chicago: Invitrogen Molecular Probes.

Hibbs AR (2004) *Confocal Microscopy for Biologists*. New York: Kluwer Academic.

Kasten FH (1999) Introduction to fluorescent probes: Properties, history, and applications. In: Mason WT (ed.) *Fluorescent and Luminescent Probes for Biological Activity: A Practical Guide to Technology for Quantitative Real-Time Analysis*, 2nd edn., pp. 17–39. San Diego, CA: Academic Press.

Lippincott-Schwartz J, Altan-Bonnet N, and Patterson G (2003) Photobleaching and photoactivation: Following protein dynamics in living cells. *Nature Cell Biology* 5(supplement): S7–S14.

Lippincott-Schwartz J and Patterson GH (2003) Development and use of fluorescent protein markers in living cells. *Science* 300 (5616): 87–91.

Miyawaki A, Sawano A, and Kogure T (2003) Lighting up cells: Labeling proteins with fluorophores. *Nature Cell Biology* 5 (supplement): S1–S7.

Panchuk-Voloshina N, Haugland RP, Bishop-Stewart J, et al. (1999) Alexa dyes, a series of new fluorescent dyes that yield exceptionally bright, photostable conjugates. *Journal of Histochemistry and Cytochemistry* 47(9): 1179–1188.

Pathak S, Cao E, Davidson MC, Jin S, and Silva GA (2006) Quantum dot applications to neuroscience: New tools for probing neurons and glia. *Journal of Neuroscience* 26(7): 1893–1895.

Shaner NC, Steinbach PA, and Tsien RY (2005) A guide to choosing fluorescent proteins. *Nature Methods* 2(12): 905–909.

Tsien RY (1998) The green fluorescent protein. *Annual Review of Biochemistry* 67: 509–544.

Wiedenmann J, Ivanchenko S, Oswald F, et al. (2004) EosFP, a fluorescent marker protein with UV-inducible green-to-red fluorescence conversion. *Proceedings of the National Academy of Sciences of the United States of America* 101(45): 15905–15910.

Zimmer M (2002) Green fluorescent protein (GFP): Applications, structure, and related photophysical behavior. *Chemical Reviews* 102(3): 759–781.

fMRI: BOLD Contrast

S A Huettel, Duke University Medical Center, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Many neuroscience techniques provide information about human brain function. Perhaps the most influential in recent years has been functional magnetic resonance imaging (fMRI). The success of fMRI can be traced to several factors. Whereas its predecessor positron emission tomography requires injection of some radioactively labeled metabolite, fMRI measures naturally occurring blood oxygenation level-dependent (BOLD) contrast. Data can be acquired using standard clinical magnetic resonance imaging (MRI) scanners (Figure 1), of which there are more than 10 000 in North America alone. The broad availability of these machines has made fMRI accessible to thousands of researchers, many originally from fields outside of neuroscience. Also, the measurements obtained using fMRI allow study of many cognitive processes. The spatial resolution of fMRI is on the order of a few millimeters, allowing researchers to identify areas of distinct function. The temporal resolution of fMRI is on the order of 1 s, which is much slower than measurements of brain electrophysiology but fast enough for many types of experiments.

The advantages of fMRI over other neuroimaging techniques have resulted in remarkable growth. No longer restricted to a handful of institutions, fMRI has become the central research technique in hundreds of laboratories. During the year prior to the publication of this edition of the encyclopedia, more than 1500 peer-reviewed scientific articles (by a conservative count) cited fMRI in their abstracts or titles. Yet, with growth should come caution. Like any other neuroscience technique, fMRI can be readily used to conduct confusing, meaningless, and uninterpretable experiments. Although a complete discussion of fMRI is beyond its scope, this article provides a broad overview to the basic principles of fMRI, from physics and neurophysiology to experimental design and analysis.

Physical Principles of MRI

The process of MRI scanning can be partitioned into three components (Figure 2), corresponding roughly to the three parts of that abbreviation: magnetization, resonance, and image formation. First, a large superconducting solenoid coil generates a very strong magnetic field that induces magnetization in tissue protons. Second, an excitation coil sends radiowaves into the tissue; by calibrating this coil to a particular, resonant frequency of the magnetized protons, energy from the radio waves will be absorbed by the protons and later reemitted as the MR signal. Third, a set of gradient coils are used to strengthen or weaken the magnetic field very slightly across space, which provides spatial variation in the signal that is received. Depending on the particular pattern (or pulse sequence) of radio waves and magnetic gradients, images can be acquired that are sensitive to brain structure or function.

Magnetization

Atomic nuclei whose characteristics can be measured by MRI are often referred to as 'spins,' reflecting that under normal conditions, thermal energy causes the nucleus to spin about itself. The most common MRI-visible nucleus in the human body is that of hydrogen within water molecules, and the processes described in the remainder of this article involve creating images of the properties of hydrogen protons. One can think of these protons as analogous to spinning bar magnets (like tops or gyroscopes), each described by the frequency and the orientation of its spinning.

In the absence of any strong magnetic field, the hydrogen protons are oriented randomly and thus their magnetic moments tend to cancel each other out. However, when the spins enter the MRI scanner, their axes of spinning tend to become aligned either parallel or antiparallel to its strong magnetic field (Figure 3). Protons in the parallel state have a lower energy level, whereas protons in the antiparallel state have a higher energy level. At room temperature in a standard 1.5 T MRI scanner, roughly equal numbers of protons are in the parallel (approximately 50.002%) and antiparallel (approximately 49.998%) energy states. This difference, although proportionally very small, results in a net magnetization. We can think of the net magnetization as a vector with two components: (1) a longitudinal component that is either parallel or antiparallel to the magnetic field and (2) a transverse component that is perpendicular to the magnetic field.

Because of their angular momentum, protons in a magnetic field initiate a gyroscopic motion known as precession. The precession frequency is proportional to the strength of the field (i.e., approximately 42.58 MHz/T). Within even the smallest quantity of tissue, there are an enormous number of protons, each precessing with different phase.

Figure 1 A modern MRI scanner. Machines such as this 1.5 T scanner can be used for both structural MRI measurements and fMRI studies. The subject lies down on the table at the front of the scanner and places his or her head inside the volume coil at the center of the image. Surrounding the central bore of the scanner are the coils that generate the strong static magnetic field and weaker gradient magnetic fields; these two sets of coils are hidden behind the plastic cowling in this image. Reproduced from Huettel SA, Song AW, and McCarthy G (2004) *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer, with permission.

Thus, the transverse components of their magnetization will tend to cancel out under normal conditions, leaving only a net magnetization that is precessing in parallel with the main magnetic field.

Resonance

Spins in the low-energy state (parallel) can jump to the high-energy state (antiparallel) by absorbing electromagnetic energy (i.e., photons). Yet, because the net magnetization rotates around the magnetic field axis many millions of times per second, a single pulse of energy would have little effect. Instead, protons absorb energy when it is applied at a particular frequency for an extended period of time. As an analogy, imagine a backyard swing set. By continually applying energy at the swing's natural (or resonant) frequency, a series of very small pushes contributes energy to the system and increases the velocity of the swinging. Similarly, MRI scanners use specialized coils to transmit electromagnetic energy at the precession frequency of the hydrogen protons within the magnetic field, whereupon some spins in the low-energy state absorb this energy and change to the high-energy state. This process is known as excitation (Figure 4(a) and 4(b)).

After excitation, some protons will move back to the low-energy state, emitting energy in the process. During this period, the net magnetization gains a significant transverse component – oscillating at the resonant frequency of the precessing spins – that can be measured in detector coils as the MR signal (Figure 4(c)). This process is known as reception. Since the frequencies of excitation and reception are identical, the same radiofrequency coil can be used for both. The MR signal, which is the coherent part of the transverse magnetization, decays quickly over time because of spin–spin interactions, generally within a few tens of milliseconds – a process known as transverse relaxation. The time constant that characterizes this decay process is T2. Simultaneously over a period of several seconds, some of the individual spins move back to the low-energy state in a process known as longitudinal recovery (of the net magnetization). The time constant that characterizes this recovery process is T1.

An additional, extrinsic cause of MR signal decay is the presence of inhomogeneity in the external magnetic field. Because each spin precesses at a frequency proportional to the magnetic field it experiences, spatial variations in field strength cause variation in precession frequency, also leading to the loss of coherent MR signal. Because field inhomogeneity has an additive effect on top of T2 spin interactions, the combined effects of these two factors are characterized by a third time constant, T2*. This final time constant forms the basis of the BOLD contrast used in nearly all fMRI studies.

Image Formation

In the context of MRI, an image is a map that depicts the spatial distribution of some spin property – for example, how many spins there are, how those spins move, or the spins' relaxation constants such as T1 or T2. The fundamental concept underlying image formation in MRI is that of the magnetic gradient, or spatially varying magnetic field. By superimposing a second magnetic field that varies linearly across space, the scanner causes spins at different locations to precess at different frequencies. In most forms of MRI, including fMRI, several gradients are used in succession. Following an initial spatial gradient along one direction (e.g., from the top to bottom of the brain), an electromagnetic (or excitation) pulse is applied at a frequency designed to excite spins within a single two-dimensional slice of tissue. Then, magnetic gradients in the other two directions are rapidly cycled on and off in prescribed patterns, depending on how the image is to be collected. The particular combination of radiofrequency excitation pulse(s) and patterned changes in magnetic gradients used for collection of a particular image is known as a pulse sequence.

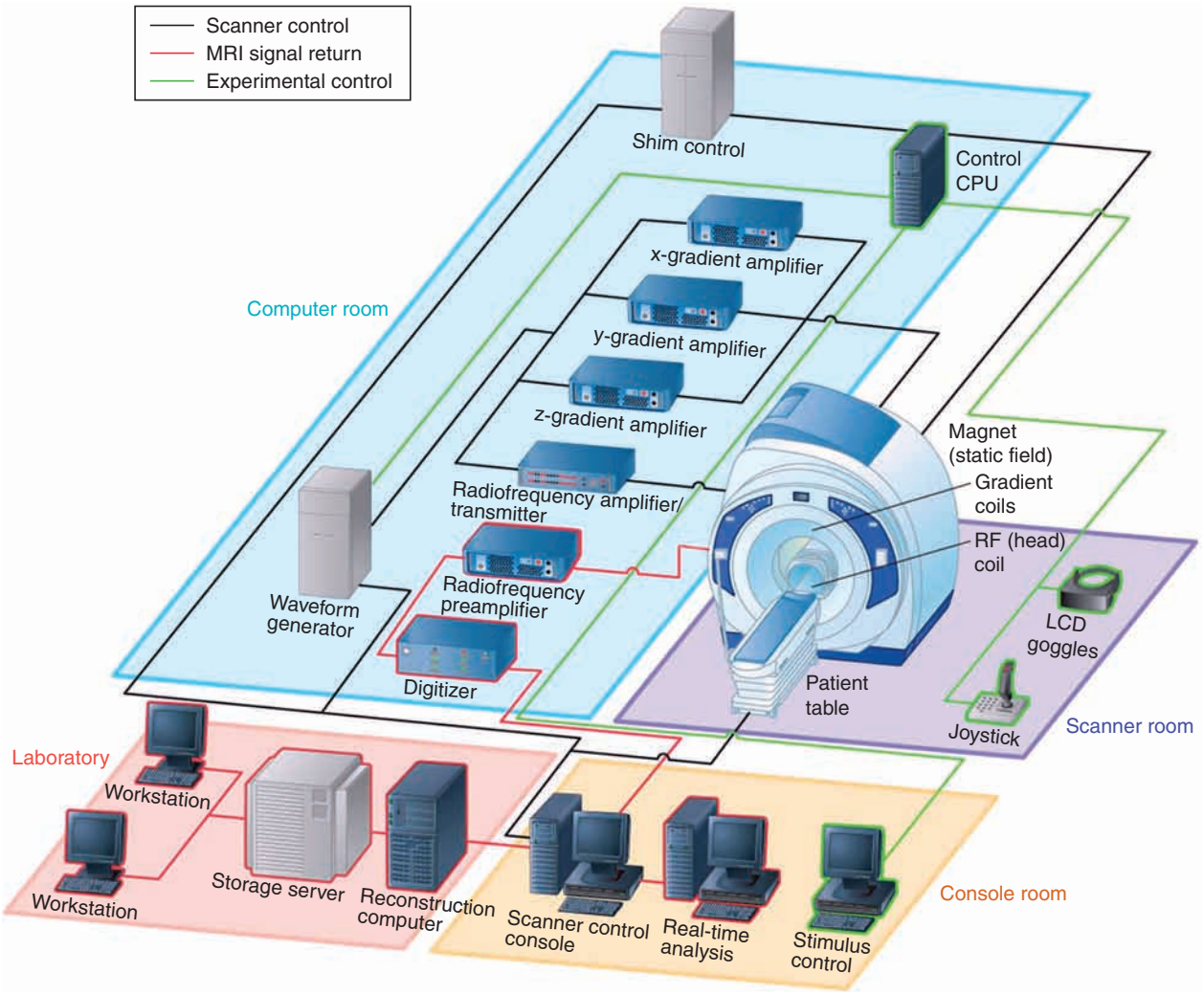

Figure 2 A schematic overview of the hardware required for fMRI research. The scanner is controlled by electronics that generate the pulse sequences used in fMRI experiments. Key components include amplifiers that create spatial gradients in three directions and a radiofrequency amplifier which controls the output of the head coil. The MR signal is measured in that same head coil and then transmitted back to the scanner console room for subsequent reconstruction. For fMRI experiments, additional systems are necessary for displaying stimuli, recording behavior, and storing and analyzing the data. Reproduced from Huettel SA, Song AW, and McCarthy G (2004) *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer, with permission.

How changing magnetic gradients can be used to create an image may be the most complex aspect of MRI. To understand the underlying concepts, it is important to recognize that any image can be mathematically represented as either a map of intensity values (normal space) or a map of spatial frequency values (*k*-space). For MRI, each point in *k*-space reflects the accumulated changes in phase based on the spatial gradients that have been presented following excitation. For example, if positive horizontal and positive vertical gradients were together applied for a few milliseconds, any signal then recorded would be considered to come from the upper right quadrant of *k*-space. At the center of *k*-space are low spatial frequency components, reflecting only small net exposures to gradients. These correspond to the overall intensity and coarse features of the normal space image. The periphery of *k*-space contains high spatial frequency components; these provide information about fine details of the sample.

Most pulse sequences manipulate spatial gradients to collect data from all parts of *k*-space following one excitation pulse. In echo-planar imaging, which is the most commonly used approach to fMRI, horizontal and vertical gradients are alternated to fill *k*-space in a back-and-forth pattern. Spiral imaging sequences use a different approach: simultaneously using both gradients to move from the center of *k*-space outward (or from the periphery inward) in a whirlpool pattern. Regardless of the pulse sequence, two parameters are

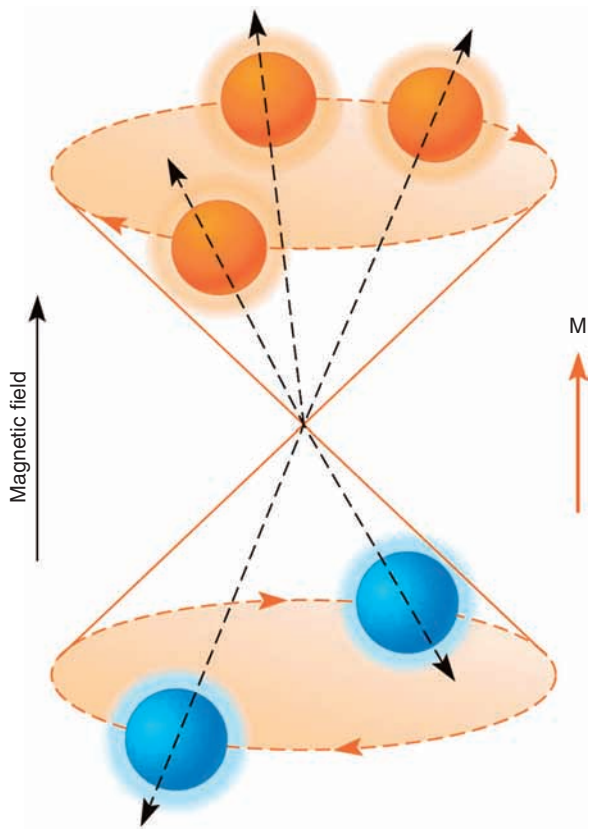

Figure 3 Net magnetization. In the strong magnetic field of the MRI scanner, slightly more than half of the protons align their axis of precession parallel to the magnetic field, whereas slightly less than half are antiparallel. This difference results in a net magnetization, labeled M, along the direction of the magnetic field. Reproduced from Huettel SA, Song AW, and McCarthy G (2004) *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer, with permission.

important. The echo time (or TE) is the time between the excitation pulse and the recording of data from the center of *k*-space; to give maximal T2* contrast for fMRI, TE values are typically a few tens of milliseconds. The time between successive excitations, known as repetition time (TR), depends on how many slices are to be collected and the characteristics of the pulse sequence, including TE. In most fMRI studies that sample the entire brain, often using 20 or more individual slices, TR is between 1 and 3 s.

In summary, standard MRI scanners can collect, every few seconds, an image of the brain that is sensitive to magnetic field inhomogeneities. As described in the following section, one consequence of neuronal activity is changes in blood oxygenation that cause exactly that – spatially specific inhomogeneities in the local magnetic field.

Physiological Principles of fMRI

Although constituting only approximately 2% or 3% of the body's mass, the human brain uses approximately 20% of the body's energy. This enormous energy budget supports the integrative (dendritic) and signaling (axonal) activity of neurons. Both forms of neuronal activity do not require external sources of energy because the associated changes in electrical potential are caused by movements of ions along concentration gradients. Energy must be supplied, however, for the restoration of membrane potentials following neuronal activity. It has been estimated that for humans and higher primates, the restoration of

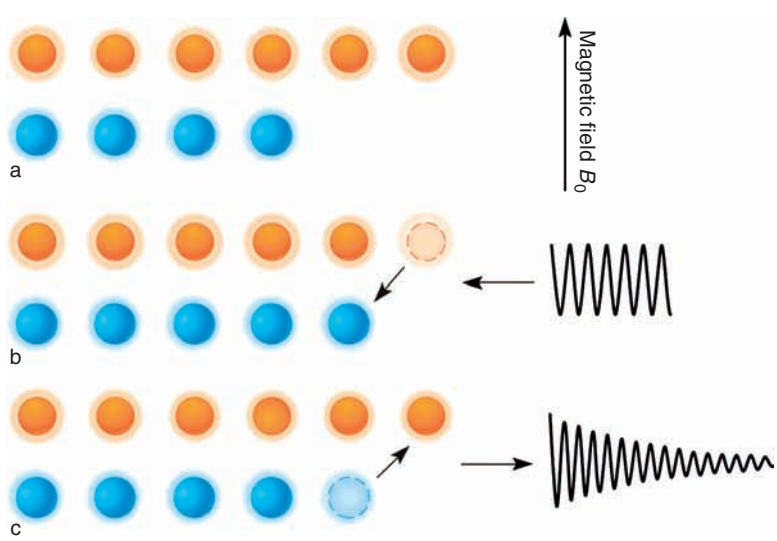

Figure 4 A schematic representation of excitation and reception. (a) When protons are placed in an external magnetic field, often labeled B_0, more will be at the low-energy state (orange) than at the high-energy state (blue). (b) If an excitation pulse is applied at the resonant frequency of the protons, some will absorb that energy and jump to the high-energy state. (c) However, after the excitation pulse is turned off, some of the protons in the high-energy state will return to the low-energy state, releasing the energy they earlier absorbed. Reproduced from Huettel SA, Song AW, and McCarthy G (2004) *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer, with permission.

these potentials requires the lion's share of energy expenditure, perhaps 80% or more of all ATP consumption in the brain.

To meet this energy demand, the vascular system supplies two important metabolites: glucose and oxygen. It has long been speculated that regional blood flow – and thus delivery of these metabolites – depends on the local activity of neurons. In a famous passage, the nineteenth-century British physiologists Roy and Sherrington argued that blood flow might depend on function: "We are of opinion that an automatic mechanism . . . is well fitted to provide for a local variation of the blood supply in accordance with local variations of the functional activity" (1890, p. 105).

Recent studies have demonstrated that neuronal stimulation leads to increases in blood flow in arterioles near the area of stimulation, consistent with Roy and Sherrington's predictions. However, the increased blood flow is imperfectly coupled to the increased metabolic demands of neurons, both in spatial specificity and in quantity. Even if only a small population of neurons are active, there are increases in arteriole diameter several millimeters or more from the active neurons. The arterial system supplies more oxygen than can be extracted within the capillary bed. Thus, following neuronal activity, there is an increase in the amount of oxygenated hemoglobin and a decrease in the amount of deoxygenated hemoglobin in venous blood. These changes in blood flow and especially blood oxygenation suggest potential markers for brain function.

BOLD fMRI

In the early 1980s, Thulborn and colleagues found that the relative oxygenation of blood (*in vitro*) influenced its T2* decay constant. Later that decade, Seiji Ogawa and colleagues examined whether manipulating the proportion of blood oxygen would affect the visibility of neural blood vessels on T2*-weighted images. They found that when their rodent subjects breathed pure oxygen, MR images revealed structural differences but not blood vessels. However, when the rodents breathed normal air, blood vessels became visible on the images as thin, dark lines. Ogawa and colleagues concluded that these thin lines represented magnetic susceptibility effects caused by the presence of paramagnetic deoxygenated hemoglobin. A summary of the BOLD effect is presented in Figure 5.

The effect discovered by Ogawa came to be labeled blood oxygenation level-dependent contrast, or often simply BOLD contrast. Soon after his initial reports, several groups began to explore whether this endogenous contrast could be used to create images of brain function. (Note that, at the same time, several fMRI studies used exogenous contrast agents, typically based on the injection of compounds containing the element gadolinium. The use of contrast agents remains important for many clinical uses of fMRI, but it has become superseded by BOLD contrast because of its invasive nature.) The first papers using BOLD fMRI, which were published in 1992, provided proof that this new technique could provide images of activation in primary motor and visual cortices. Throughout subsequent years, researchers extended the reach of BOLD fMRI to more cognitive applications, using it to map the neural substrates supporting attention, memory, decision making, and other complex functions.

Only recently has there been significant progress toward understanding the neurobiological origins of the BOLD signal. Recall that the energy demands of the brain are primarily attributable to neuronal activity, specifically the restoration of membrane potentials. However, it was long unclear whether BOLD changes tracked axonal action potentials, dendritic postsynaptic potentials, both, or neither. Simultaneous measurements of fMRI BOLD signal and electrophysiological activity – originally in monkeys and recently in humans – have indicated that BOLD signal contrast arises primarily from postsynaptic potentials, not action potentials. In particular, synchronized dendritic activity across populations of neurons seems to be a key contributor to BOLD amplitude.

fMRI Experimental Design and Analysis

The physical and physiological properties described in the previous sections establish important constraints on the design and analysis of fMRI experiments. Underlying most fMRI experiments is the linear systems framework, which suggests that the measured MR signal arises from the linear convolution of a set of impulses representing neuronal activity with a standardized hemodynamic response function. This framework implies that one can determine whether a brain region exhibited a hypothesized pattern of neuronal activity by statistical evaluation of its BOLD time course. In general, fMRI designs (and subsequent analyses) are created to answer two types of questions: detection, or "What brain regions showed task-related neuronal activity?" and estimation, or "How did the BOLD response change over time in a particular brain region?"

Experimental Design

Considered broadly, fMRI experimental designs take blocked or event-related forms. The former tends to have superior detection power, whereas the latter

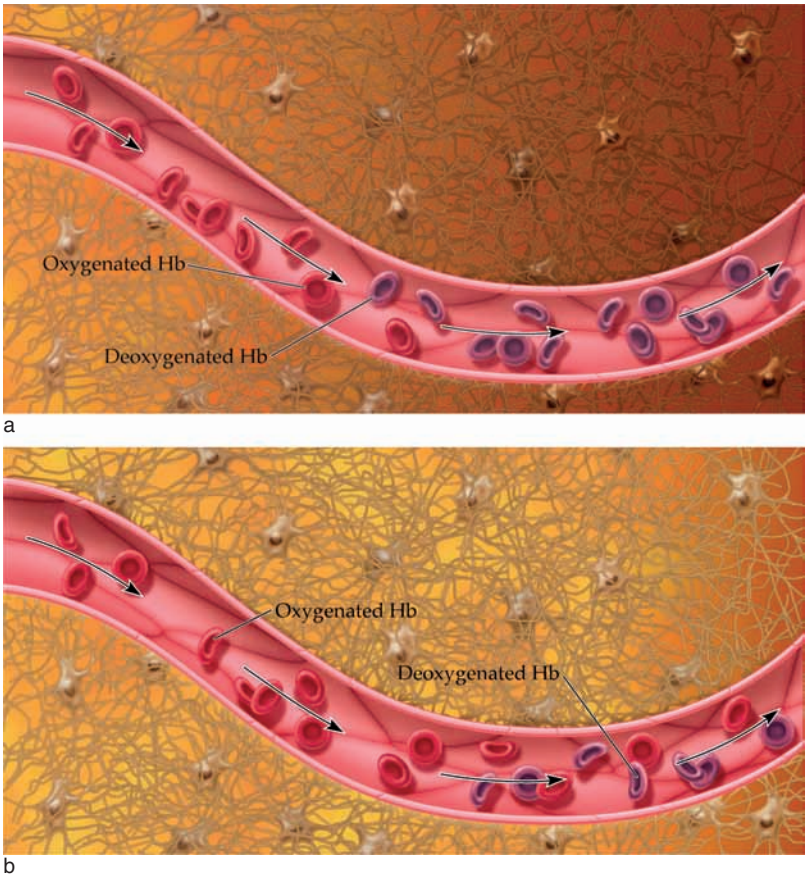

Figure 5 Blood oxygenation level-dependent (BOLD) contrast. Under normal conditions, oxygenated hemoglobin is converted to deoxygenated hemoglobin within the capillary bed, resulting in a darkening of T2* images (a). However, when neurons become active, there is an increase in the supply of oxygenated hemoglobin. This causes a decrease in the amount of deoxygenated hemoglobin, which in turn results in brightening on T2* images (b). Reproduced from Huettel SA, Song AW, and McCarthy G (2004) *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer, with permission.

tends to have better estimation power. In recent years, optimized event-related designs that provide both excellent detection and good estimation have become increasingly common. Mixed approaches which combine features of both types of design are also used to examine both transient and sustained processes. The choice of design depends not only on the experimental goals but also on the constraints of the tasks. For example, some types of experiments (e.g., detecting an infrequent target) can only be presented in an event-related form, whereas others (e.g., performing a sustained attention task) are more reasonably conducted in blocked form.

Blocked designs present stimuli or tasks for prolonged intervals (or blocks), with the goal of evoking an extended period of steady-state neuronal activity. The central challenge for blocked designs is matching the different blocks on all factors except the main independent variable. Blocks that involve no experimental task (i.e., resting) are not absent of mental processes. Instead, they contain particular types of processes associated with reflection, daydreaming, self-assessment, bodily attention, and emotion. Thus, certain brain regions repeatedly show up as deactivations (i.e., exhibiting greater activation on control than task blocks) across a wide variety of tasks. Although well-controlled blocked designs have maximal detection power, they provide almost no information about the timing of neuronal activity or about the shape of the evoked BOLD response.

Event-related designs present stimuli or tasks in discrete, punctate fashion. In most such designs, different conditions of the independent variables are associated with different event types. Unlike blocked designs, the different events are usually presented in either a random or a temporally jittered order to maximize the ability to distinguish among their BOLD contributions. Without such steps, the BOLD signal would saturate to some maximal value with repeated events. The greatest advantage of event-related designs is their flexibility: any property could be used to define different classes of events. In memory

studies, for example, a common approach is to present a series of words to be encoded into memory. These words are sorted into event categories not based on what the subject does within the scanner but, rather, based on whether or not a word is remembered at a subsequent testing session hours (or even days) later.

Finally, mixed designs combine the basic elements of the previous two types. They present stimuli in discrete and regular blocks, each containing multiple types of events. The task blocks are assumed to evoke sustained changes in task-strategy, attention, or other cognitive processes, and blocked analyses measure state-related processes. The individual events are assumed to evoke transient activation associated with the particular task demands. Event-related analyses would therefore measure item-related processes. Mixed designs are preferable over either of the two basic design types when both sustained and transient aspects of cognition are of interest.

Preprocessing

Once collected, fMRI data are usually subjected to a series of steps known as preprocessing to measure and remove unwanted variability. These steps are usually undertaken after the *k*-space data have been reconstructed into image space but before engaging in any task-related statistical analyses. Initial temporal and spatial preprocessing steps correct for variability both in the timing of slice acquisition and in the spatial position of voxels. The most insidious cause of spatial error is head motion, which is often more easily prevented by physical restrictions than corrected during preprocessing. Although a variety of corrective algorithms are used, the most common are rigid-body transformations that attempt to minimize the error between each volume and a reference volume (e.g., the first in the data set).

To facilitate comparisons across subjects, fMRI data are often normalized into a standard stereotaxic space, usually through a combination of linear and nonlinear transformations. High-resolution anatomical images for each subject are often used as an intermediate step in this normalization process. Depending on the experimental goals, temporal and spatial filtering may also be introduced. Temporal filtering can minimize the contributions of several forms of noise, such as physiological variability (e.g., respiration) or slow, temperature-related changes in the strength of the magnetic field. Spatial filtering (i.e., smoothing) can reduce apparent noise and increase the validity of comparisons across subjects. Finally, quality assurance should be an integral part of any preprocessing. Without rigorous quality assurance testing, unnoticed problems can propagate into the final results of an experiment with potentially disastrous results.

Data Analysis

As for all of science, analyses in fMRI can be either data driven or hypothesis driven. In data-driven analyses, which are only found in a minority of fMRI studies, the researcher identifies regularities present in the experimental data. Common approaches include independent components analysis, partial least squares, and intersubject correlations. Data-driven analyses are largely exploratory in that they are intended to discover systematic and regular changes in BOLD signal. Note that data-driven techniques are often used as a first step before setting up subsequent hypotheses. This combined approach can be extraordinarily powerful, particularly when used to examine interrelations among brain regions.

In contrast, hypothesis-driven analyses are intended to test specific predictions about the relations between an experimental task and evoked BOLD activation. Most hypothesis-driven analyses use a variant of the general linear model to evaluate whether and how different explanatory variables contribute to the observed pattern of activation. The explanatory variables are usually considered to represent particular cognitive processes (or sets of processes), such as encoding an item into memory or deciding between two options. A central challenge of the data analysis, therefore, is identifying when these processes occurred. Then, impulses defined by the timing and estimated duration of the processes are convolved with some representation of the fMRI hemodynamic response; these convolutions generate one or more regressors that predict how BOLD signal would change were that process to occur. All such regressors are entered into a model – often accompanied by nuisance regressors such as head motion parameters – to form a design matrix. A statistical program then calculates what combination of regression weights best predicts the experimental data. Data from different subjects are combined using fixed- or random-effects approaches; for most experiments, random-effects approaches are preferable because they allow extension of the results to the population from which the subject sample is drawn.

Correction for Multiple Comparisons

A typical fMRI data set can contain a few tens of thousands of voxels. Thus, using a conventional threshold (e.g., alpha of 0.05) will result in an enormous number of type I errors (false positives). Conversely, the use of very strict correction factors (e.g., Bonferroni comparison) risks missing many regions

with true activity (e.g., type II errors). A common compromise between these extremes is to adjust the correction factor based on the data's spatial autocorrelation, which can result from many sources: the spatial extent of activity, hemodynamic spread, smoothing during preprocessing, and variability across subjects. For fMRI data, the true number of independent comparisons can be orders of magnitude smaller than the number of voxels. Another, nonexclusive approach is to require clusters of activation to reach a particular size (e.g., ten or more active voxels) to be counted as significant. Finally, other correction metrics can be used, such as the false discovery rate, which controls the type I error rate among voxels labeled as significant.

Displaying and Interpreting fMRI Data

Nearly all neuroscientists (and most of the lay public) have been exposed to images created by fMRI analyses. Images of a high-resolution brain with colored patches can be extraordinarily striking (Figure 6), sometimes leading to the idea that scientists have captured a photograph of the brain in action. The apparent detail of the background MRIs – or three-dimensional renderings – suggests that the research has exquisite spatial resolution. Also, when researchers show an area of activation, it seems like they have discovered the brain area that processes faces, makes decisions, or evaluates numbers. These ideas, though all compelling, are false. The images arising from fMRI experiments are not photographs: they are statistical maps. The resolution of fMRI maps is relatively coarse (i.e., acquired at 3- to 5-mm resolution and then smoothed further during processing) compared to the potentially submillimeter resolution of structural MRI. In addition, the identification of a single active region does not rule out the possibilities that many other regions may contribute or that the experiment may not have sufficient power to identify other effects.

Described plainly, the colors of fMRI activation maps reflect the statistical confidence that a particular brain region expresses a particular function. Bright colors usually indicate regions where there is strong evidence that the experimental manipulation affected the BOLD signal, whereas uncolored regions lacked any significant statistical effect. Like other experiments, fMRI studies are not designed to identify null effects, and a region that is uncolored on a fMRI activation map may still contribute. Furthermore, the meaning of an image depends on the quality of the underlying experimental design, the adherence of the subjects to the expected task, the analyses conducted, and even the threshold chosen. When evaluating fMRI data, it is thus extraordinarily important to look beyond the images to evaluate the underlying methods.

Future Directions

In many ways, fMRI has indeed become a mature technique. Researchers and students can download publicly available (and anonymized) data sets and freely available analysis programs. Dozens of fMRI

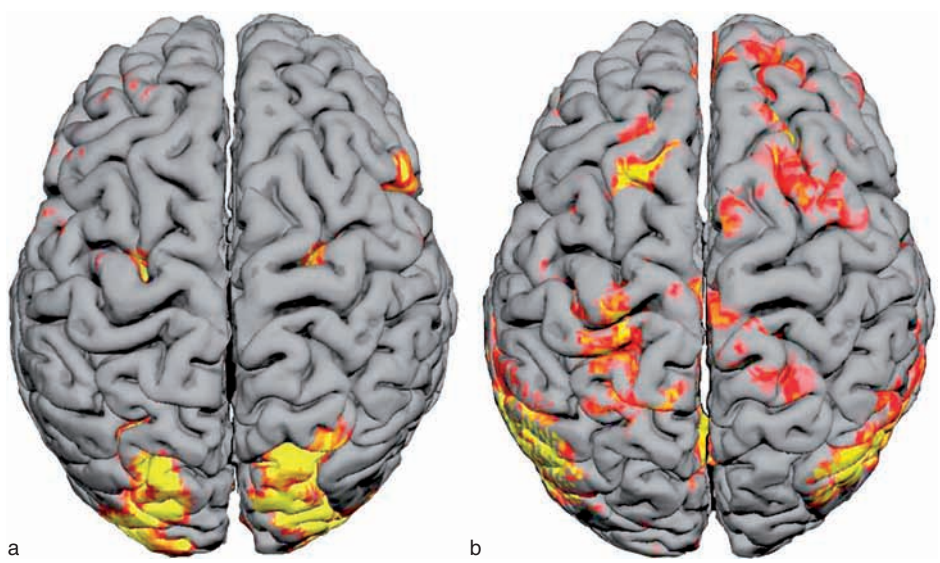

Figure 6 fMRI data displayed on a high-resolution rendered brain. Shown are statistical maps derived from a visual search task, displayed on a high-resolution brain that has been rendered into a three-dimensional surface. The colors represent areas of statistically significant activation, with yellow indicating greatest significance. Shown in (a) are areas with positive activation during visual search, whereas in (b) areas with reduced activation during visual search are indicated.

courses are offered throughout the world yearly. Even the collection of fMRI data has become democratized: new MRI scanners, even those for clinical use, come equipped with hardware, pulse sequences, and online analysis for fMRI experiments.

Given this maturity, from where will the next advances in fMRI arise? One possibility is increased field strength. In the early years of fMRI, there was a continual push toward high-field (i.e., >1.5 T) scanning. Now, 3 T and higher field machines are common, both in research and in clinical settings, and further improvements in field strength seem likely to be incremental at best. Another possibility is improved spatial and temporal resolution. Although some new techniques have arisen (e.g., the growth of spiral imaging), current fMRI data acquisition is – to a first approximation – largely similar to that of a decade ago. The stability of fMRI technology does not mean that no future advances are likely. In particular, simultaneous multiple-coil recording promises to allow researchers to reduce the time required to collect fMRI images and, more important, increase image quality. Yet, even this will not revolutionize BOLD fMRI experimentation.

Instead, the main advances in fMRI will come from improved experimental designs, new forms of data analysis, and novel combinations with other techniques. Most early fMRI studies tried to localize particular brain functions to particular locations. As argued by critics, this approach overemphasizes functional specificity but fails to account for how different regions interact to support complex behavior. Although this criticism is still regularly repeated, its validity diminishes with each passing year. Researchers are increasingly likely to relate fMRI activity to specific behavioral measures, such as memory accuracy, decision outcome, perceptual reports, or even individual traits. Likewise, increasingly more studies report interactions between regions, not just a list of activation foci. Combining fMRI with electrophysiological techniques will remain a key area of methodological research because of their posited complementarity of spatial and temporal properties. However, during the next few years, there may be more value in adding transcranial magnetic stimulation to fMRI to better address questions about causality. Regardless of whether some (or all) of these new developments take shape over the coming years, it seems clear that fMRI will continue to be a central technique in the study of the human brain.

See also: Alzheimer's Disease: MRI Studies; Cognition: An Overview of Neuroimaging Techniques; Cognitive Neuroscience: An Overview; Event-Related Potentials (ERPs) and Cognitive Processing; Executive Function and Higher-Order Cognition: Neuroimaging; Neuroanatomy Methods in Humans and Animals; Neuroimaging; Perfusion MRI; Positron Emission Tomography (PET); Statistical Tests and Inferences.

Further Reading

Blamire AM, Ogawa S, Ugurbil K, et al. (1992) Dynamic mapping of the human visual cortex by high-speed magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America* 89: 11069–11073.

Boynton GM, Engel SA, Glover GH, and Heeger DJ (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience* 16: 4207–4221.

Buxton RB (2002) *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*. New York: Cambridge University Press.

Gusnard DA and Raichle ME (2001) Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience* 2: 685–694.

Huettel SA, Song AW, and McCarthy G (2004) *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer.

Kwong KK, Belliveau JW, Chesler DA, et al. (1992) Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences of the United States of America* 89: 5675–5679.

Logothetis NK, Pauls J, Augath M, Trinath T, and Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412: 150–157.

Ogawa S, Lee TM, Kay AR, and Tank DW (1990) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America* 87: 9868–9872.

Racine E, Bar-Ilan O, and Illes J (2005) fMRI in the public eye. *Nature Reviews Neuroscience* 6: 159–164.

Roy CS and Sherrington CS (1890) On the regulation of the blood-supply of the brain. *Journal of Physiology* 11: 85–108.

Thulborn KR, Waterton JC, Matthews PM, and Radda GK (1982) Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field. *Biochimica et Biophysica Acta* 714: 265–270.

Folate Deficiency States

V T Ramaekers, University Hospital Aachen, Aachen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Since the discovery of folic acid and folate deficiency states, the knowledge from rare inborn errors of folate metabolism and folate transport disorders across placental, intestinal, and blood–central nervous system (CNS) barriers has contributed to our understanding of the origin of neural tube defects, homocysteinemia, and vascular disease, and a range of neuropsychiatric disorders. Herein, we review the most important aspects of prenatal, systemic, and cerebral folate deficiency (CFD), their underlying mechanisms including the role of folate-receptor autoantibodies, the current treatment strategies for folate deficiency states, as well as the effect of the folate food fortification program in North America.

Physiology of Folate Metabolism

Folic acid is a water-soluble vitamin, of which only traces are found in nature. Folic acid is the parent structure of folates, which represent a family of more than 25 structurally related and interconvertible compounds. The basic structure of folic acid consists of a 2-amino-4-hydroxypteridine molecule linked through a methylene carbon to a para-aminobenzoyl group, which is attached to a monoglutamate residue or polyglutamate chain (Figure 1). Reduction of the pteridine ring at the 5,6,7,8 carbon positions results in the reduced folate forms which function as active cofactors in cellular metabolism. These reduced folate forms may carry one-carbon units at three different oxidation levels in the form of methyl, methenyl, or formyl groups, attached to the N-5 and/or N-10 position of tetrahydrofolate, which can enzymatically be interconverted.

The naturally occurring folates in foods are reduced polyglutamyl derivates of mostly L-5-methyltetrahydrofolate (5MTHF), present in leafy green vegetables, citrus fruits, liver, and yeast extracts. After food ingestion, polyglutamyl folates are hydrolyzed to monoglutamyl derivatives by a zinc-dependent pteroylglutamate hydrolase at the brush-border membrane of the jejunum. This is followed by a carrier-mediated transport (reduced folate carrier 1; RFC1) into the intestinal cell where the various monoglutamyl folates are converted to 5MTHF, which exits the cell into the portal circulation for further transport to the liver and peripheral organs. Food-derived folic acid can be reduced to dihydrofolate and L-tetrahydrofolate by the enzyme dihydrofolate reductase (DHFR) and further converted to 5MTHF before leaving the intestinal cell. However, part of the folic acid can also enter directly into the circulation in case of diminished activities of the reducing and interconverting enzymes. Another source of folate is derived from the colonic folate-producing bacterial flora.

The physiological function of the various reduced folate forms subserves the *de novo* synthesis of purines and thymidine, methylation of DNA, conversion of homocysteine to methionine, and the formation of the activated methyl-group donor *S*-adenosylmethionine (SAM).

5MTHF is the physiologically most active and important circulating folate form in blood which passes cellular membranes, placental, and blood–brain barriers. 5MTHF is the precursor of the activated methyl-group donor SAM (Figure 2). SAM is used in more than 100 chemical reactions among which are the methylation of DNA, fatty acids, phospholipids, polysaccharides, and proteins. For example, the methylation of arginine at position 107 within myelin-basic protein is necessary to maintain stability of CNS myelin. In addition, the universal methyl-group donor SAM acts in the catabolism of biogenic monoamines.

An adequate folate pool and metabolism is not only necessary for DNA synthesis and cell growth, but also for normal embryonic development, maturation, and function of the nervous system.

Three different transport mechanisms promote cellular folate uptake and transport across cellular membranes, that is, folate-receptor mediated endocytosis by folate receptor 1 (FR1) and folate receptor 2 (FR2), the RFC1, and an ATP-dependent folate exporter system (Figure 3(a)). FR1 is mainly distributed at epithelial cells such as choroid plexus, lung, thyroid, and renal tubular cells, while FR2 is mainly located within mesenchymal derived cells, like red blood cells. Both FR1 and FR2 are high-affinity proteins which function at the physiological nanomolar range of extracellular folate concentrations. RFC1 is ubiquitously distributed and represents a low-affinity folate-transporting system with bidirectional transport across cellular membranes. The ATP-dependent folate exporter regulates intracellular folate monoglutamate concentrations.

The transport mechanisms across the placental, intestinal, and blood–CNS barriers deliver sufficient folate stores within organs and the CNS (Figure 3(b)).

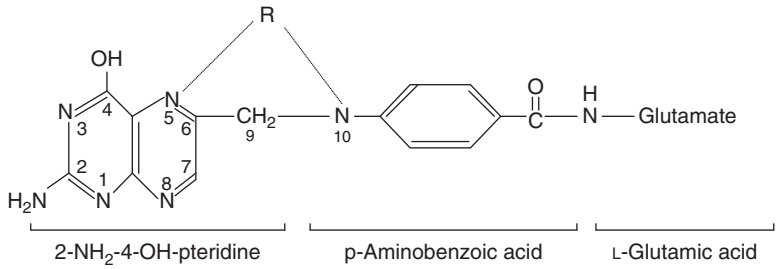

Figure 1 Structural formula of folic acid. This is composed of an unreduced 2-amino-4-hydroxypteridine molecule linked through methylene (C9-position) to p-aminobenzoylmono-glutamate, while R represents the carried one-carbon group at different oxidation states in the form of methyl, methenyl, or formyl, attached to the N-5 and/or N-10 positions as shown in the drawing. The metabolically active form tetrahydrofolate is reduced at the 5,6,7,8 positions of the pteridine group.

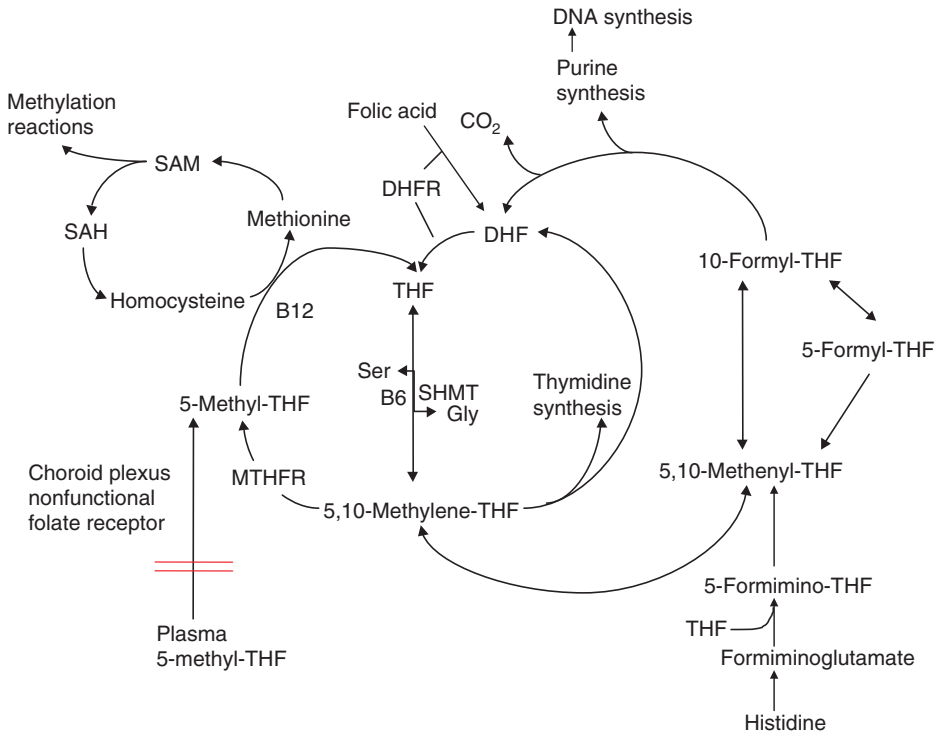

Figure 2 Metabolic pathways showing the interconversion of different folate forms in the brain with major transport of 5-methyl-THF (5MTHF) across the choroid plexus where, in CFD, a nonfunctional FR1 protein is thought to be responsible for reduced folate transport to the CNS. At the upper left the conversion of homocysteine to methionine is shown by the enzyme methionine synthase requiring 5MTHF and vitamin B12. Methionine is converted to the activated methyl-donor SAM, necessary for methylation reactions. The other folate interconversions and metabolic pathways are also shown. SAM, *S*-adenosylmethionine; SAH, *S*-adenosylhomocysteine; 5-methyl-THF, 5-methyl-tetrahydrofolate; Ser, serine; Gly, glycine; THF, tetrahydrofolate; DHF, dihydrofolate; DHFR, dihydrofolate reductase; SHMT, serine hydroxymethyltransferase requiring cofactor B6; MTHFR, methylenetetrahydrofolate reductase; 5-formyl-THF is synonymous with folinic acid or leucovorin.

For passage across the blood–cerebrospinal fluid (CSF) barrier, plasma 5MTHF is bound by the FR1 protein, anchored to choroid epithelial cell membranes by a glycosylphosphatidylinositol (GPI) moiety. This is followed by receptor-mediated endocytosis, storage, and subsequent export. The choroid plexus is the main site of active folate transport to the CNS. Folate transport by FR1 across the choroid plexus is an active ATP-dependent transport leading to an at least twofold higher concentration in CSF compared to plasma. The FR1 is mainly localized at the basolateral surface of choroid epithelial cell membranes, facing the plasma side, where it functions through receptor-mediated endocytosis (Figure 3(b)). After endocytosis and release of folate from its receptor into the cytosol, part of the endocytotic membrane-attached FR1 will recycle to the basolateral choroid cell surface (Figure 4). For normal expression of FR1 on the cell

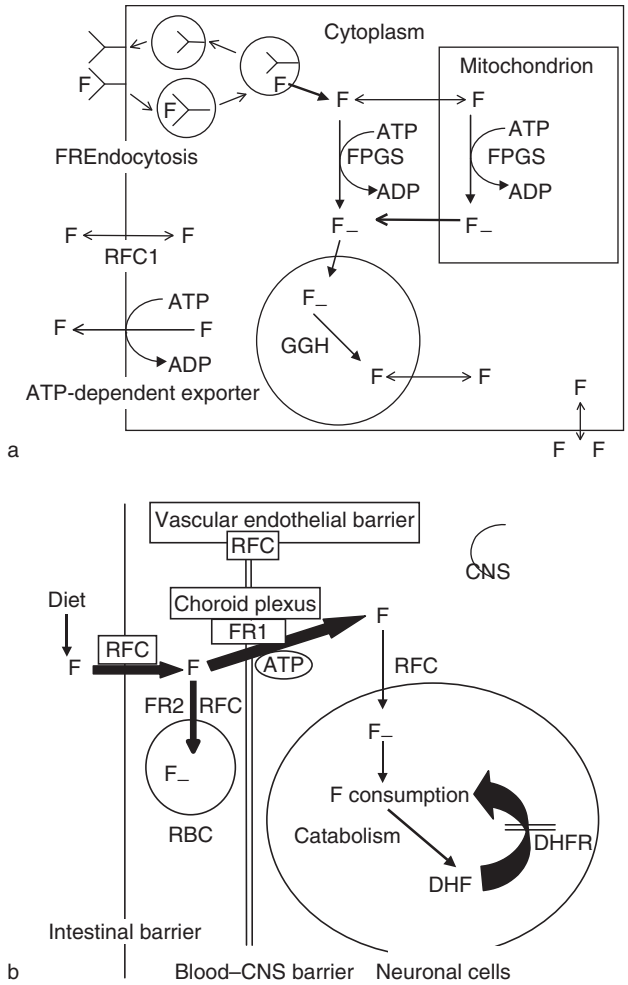

Figure 3 (a) The three different folate transport mechanisms to and from the cell interior, that is, FR1-mediated endocytosis, RFC1 and the ATP-dependent folate exporter. Folate monoglutamates can also enter mitochondria and will be converted to the folylpolyglutamate form. Folylpolyglutamates can be recoverted to their monoglutamate form by lysosomal gamma-glutamyl-hydrolase (GGH). (b) After intestinal absorption across the intestinal barrier with the use of RFC1, folates enter the circulation and can be transported into mesenchymal tissues, for example, red blood cells (RBC) by the folate receptor 2 (FR2). After entry into the cell, folate monoglutamate can be retained within cells after polyglutamylation by folylpoly-gamma-glutamate synthetase (FPGS), the major part of which will be tightly bound to proteins (F– stands for folylpolyglutamate). After passage across choroid epithelial cells using FR1 receptor-mediated endocytosis, folates will enter the CSF and enter neuronal cells by the RFC1, located at axons and dendrites. Within the neurons, part of the folate pool will be catabolized by oxidation to dihydrofolates (DHF) and folic acid, which can be reconverted to reduced tetrahydrofolate by dihydrofolate reductase (DHFR). However, in the brain of man there is no or virtually absent DHFR activity. *In vitro* experiments indicate that part can be oxidized to form 6-formyltetrahydropterin and para-aminobenzoylglutamate.

membrane surface and proper functioning of this folate-receptor-mediated endocytosis, many cooperating genetic and nongenetic factors have been identified as important determinants, such as intracellular homocysteine concentration, the extent of N-glycosylation of FR1, and the sphingolipid and cholesterol content of membranes. At the apical surface of choroid epithelial cells, RFC1 facilitates the folate export. RFC1 is also located at vessel membranes of the blood–brain barrier, where its contribution to the main bulk of folate transport is of minor importance because it has low folate affinity and will only bind folate concentrations within the micromolar range. At the cellular level, folate concentrations are influenced by folate availability, cellular folate transport efficiency, folate polyglutamylation, and folate turnover which can be influenced by different degrees of folate catabolism. The different folate forms are also known to be highly susceptible to oxidative degradation with the exception of 5-formyltetrahydrofolate (synonyms: folinic acid, calcium folinate), which represents a stable form of folate.

Oxidative degradation of the physiologically active reduced folate compounds within the brain will deplete the active folate pool and its dependent

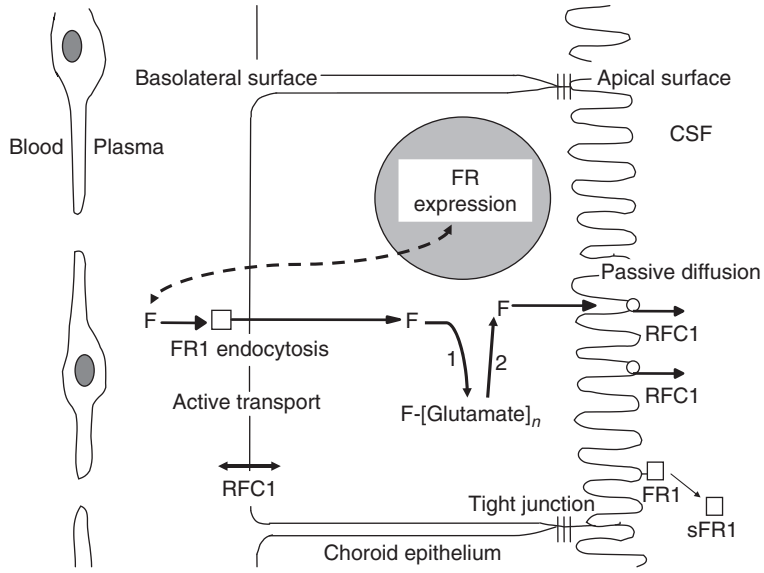

Figure 4 Active folate transport to the CNS occurs mainly at the choroid plexus epithelial cells, where the folate-binding proteins (FR1), anchored at the plasma side, bind and incorporate 5MTHF through an endocytotic process, after which conversion to folylpolyglutamate takes place by the action of folylpoly-gamma-glutamate synthetase (1). The stored folylpolyglutamate can be converted again to its monoglutamate form by gamma-glutamylhydrolase (2) after which the monoglutamate form of folate can leave the choroid plexus to the spinal fluid compartment. Part of the anchored FR1 will be cleaved from its GPI–peptide anchor and released to the spinal fluid as soluble FR1 (sFR1) without losing its physico-chemical properties. At the apical surface RFC1 facilitates the export of folate monoglutamate. The expression of FR1 is thought to correlate inversely with extracellular folate concentrations.

pathways because of absent activity of the reducing enzyme DHFR in the nervous system of man.

Clinical Features of Folate Deficiency States

Since 1965 folate-deficient states were recognized to be associated with megaloblastic anemia, neural tube defects, and cardiovascular disease. Systemic folate deficiency was suspected to play a role in the pathogenesis of various neuropsychiatric disorders and predispose to cancer. However, in many points, the pathogenesis remained still obscure. In systemic folate deficiency megaloblastic anemia is not always a consistent feature. Laboratory investigations include measurement of serum and red blood cell folate, homocysteine, and amino acids. Because clinical features sometimes mimic vitamin B12 deficiency states, vitamin B12 measurements should be included as well.

Prenatal Folate Deficiency

During pregnancy the daily folate requirements increase. Intrauterine folate deficiency during embryogenesis increases the risk of a pregnancy complicated by neural tube defects, conotruncal heart defects, and craniofacial malformations (Figure 5). Several drugs with an antifolate effect should be avoided during pregnancy (Table 1). Low socioeconomic status was found to be associated with multivitamin deficiencies during pregnancy, thereby increasing the likelihood of a neural tube defect. For this reason women of childbearing age are advised to take extra folate supplements before conception. The most important reason for the folate food fortification program introduced in North America since 1998 was to reduce the overall occurrence of neural tube defects in the population, and it resulted in a significant reduction.

Rare genetic, environmental, and nongenetic predisposing factors affecting folate metabolism and transfer across the placental barrier have been identified to play a role in the origin of neural tube defects. Among the nongenetic factors, the presence of autoantibodies directed against the folate receptor has been identified in the serum of mothers whose pregnancy was complicated by a neural tube defect. These FR autoantibodies in serum of the mothers block folate transfer across the placental barrier from the mother to the fetus, which predisposes to intrauterine folate deficiency at the time of neural tube closure in the embryo.

Systemic Folate Deficiency

Systemic folate deficiency is often the consequence of inadequate dietary folate intake and malabsorption, frequently occurring in patients with alcohol abuse, chronic debilitating diseases, regional intestinal diseases, and celiac disease. A hereditary disorder of congenital folate malabsorption and several rare

- Well-known:
  - megaloblastic anemia
- Assumed:
  - increased rate of neural tube defects (NTD) and other birth deformities (inborn heart failure, deformities of urogenital tract, cleft lip and palate)
  - increased risk of arteriosclerosis (Homocysteine↑)
  - .....
- Risk groups:
  - limited alimentation (alcoholics)
  - increased need (pregnancy)
  - decreased uptake (celiac disease)

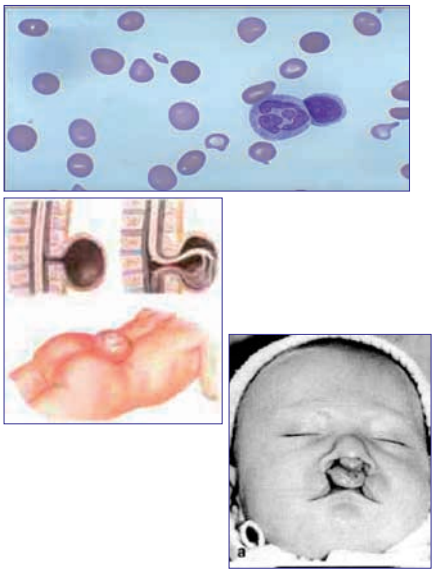

Figure 5 Megaloblastic anemia is well known in folate deficiency and should be differentiated from vitamin B12 deficiency. Increased rates of congenital malformations associated with intrauterine folate deficiency are known, while folate deficiency in adults at risk will predispose to homocysteinemia and cardiovascular diseases.

Table 1 Overview of systemic folate depletion and conditions of CFD associated with isolated 5MTHF loss in the CNS

Classification of condition	*Underlying mechanism*
Systemic depletion	
Malnutrition	Folate deficient in the diet or food deprivation
Malabsorption	Reduced folate absorption in the jejenum
Coeliac disease	Gluten enteropathy
Chron and jejunal diseases	Affects folate absorption
Antifolate agents	
Chemotherapy (MTX)	Blocks RFC1 carrier; inhibits dihydrofolate reductase
Isoniazide	Tuberculostatic drug
Sulfonamides	Analogs of para-aminobenzoic acid, interfering with THF synthesis in sensitive bacteria
Anticonvulsant drugs	Interfering with cellular folate uptake
Carbidopa	Inhibits aromatic amino acid decarboxylase with consequent SAM and 5MTHF overconsumption
Congenital folate malabsorption	Hereditary factor involving the RFC1 gene?
Inborn errors of metabolism	
MTHFR deficiency	Depletion of enzymatic product 5MTHF
Formiminotransferase deficiency	Defective histidine derived one-carbon transfer to THF
Cerebral folate deficiency	
Infantile-onset CFD	Serum FR autoantibodies of the blocking type
Low-IQ autism with neurological deficits	Serum FR autoantibodies of the blocking type
Late-onset spastic-ataxic with mental deficits	Serum FR autoantibodies of the blocking type
One-carbon pool deficiencies in the CNS	
3-Phosphoglycerate dehydrogenase def.	Defective serine synthesis affecting one-carbon pool
Inborn errors of metabolism	
Dihydropteridine reductase deficiency	Unproven mechanism
Aromatic L-amino acid decarboxylase def.	Overconsumption of 5MTHF and SAM
Kearns–Sayre syndrome	Disturbed active folate transport at choroid plexus
Mitochondrial encephalopathies	Disturbed active folate transport at choroid plexus
Rett syndrome	FR pseudogene expression due to MECP2 defect?
Aicardi–Goutières syndrome	Variable CFD of unknown origin

inborn errors of folate metabolism are accompanied by megaloblastic anemia and severe neurological features and will be outlined further. An important aspect to be considered in systemic folate deficiency is the antifolate effect of many drugs like isoniazide, sulfonamides, salicylazosulfapyridine, pyrimethamin (Daraprim), methotrexate, carbidopa, and the anticonvulsants phenytoin, primidone, phenobarbital, carbamazepin, and valproate.

Homocysteinemia with an increased risk of thrombosis and thromboembolism, stroke, and cardiovascular disease has been associated with commonly occurring polymorphisms of several genes, like for example homozygosity for the 677C→T mutation of the methylenetetrahydrofolate reductase (MTHFR) gene, that results in the change of an alanine to valine in the catalytic domain of the enzyme leading to a 50–60% decrease in the specific activity of MTHFR. The T allele frequency is around 30% in Europeans and Japanese but is only around 11% in African Americans. Studies have clarified that this polymorphism is only associated with elevated homocysteine levels in patients with low folate intake. The reduced enzyme activity will affect the interconversion of various folate forms and change the folate form distribution. MTHFR polymorphism has been found to play a role in the pathogenesis of homocysteinemia which together with other genetic and environmental factors further increases the risk of thrombosis and thromboembolism, stroke, and cardiovascular disease.

The first reported neuropsychiatric manifestations of systemic folate deficiency were the single occurrence or combinations of polyneuropathy, ataxia, and spastic–ataxic syndromes due to subacute combined degeneration of the spinal cord, already known to occur in cobalamin deficiency. In addition, systemic folate deficiency was reported to be associated with mental retardation, dementia, parkinsonian features, athetosis, seizures, myopathy, and an increased incidence of affective disorders, in particular depression, organic brain syndrome, and schizophrenia-like syndromes. Arguments in support of the etiologic role of folate deficiency were megaloblastic anemia with low serum and red blood cell folate, elevated plasma homocysteine, and clinical improvement after folate administration.

The pathophysiology of CNS dysfunction due to folate deficiency in psychiatric disorders remains obscure but is probably related to derangements of DNA synthesis and gene expression, disturbed methylation processes, and serotonin turnover. Folate-deficient patients with neuropsychiatric phenotypes were found to have low levels of the serotonin end-metabolite 5-hydroxyindole acetic acid (5HIAA) in their CSF, returning to normal after folate administration. Folate-deficient patients without neuropsychiatric abnormalities had normal CSF 5HIAA levels. There appeared to exist a close relationship between folate-responsive neuropsychiatric symptoms and changes of CNS serotonin metabolism.

Neuropathological correlates of folate deficiency described vessel wall lesions and demyelinating lesions of white matter around blood vessels in the brain and spinal cord. The histopathology of demyelinating lesions in the spinal cord showed vacuolization of myelin sheaths of the posterior and lateral columns and often the anterior columns, which has been called subacute combined degeneration. Subacute combined degeneration of the spinal cord has earlier been associated with vitamin B12 deficiency in patients with pernicious anemia (**Figure 6**).

The advent of improved neuroimaging of the nervous system with MRI has revealed white matter demyelinating lesions and leukoencephalopathy in patients with systemic folate deficiency or children suffering from a severely reduced enzyme activity of methylenetetrahydrofolate reductase.

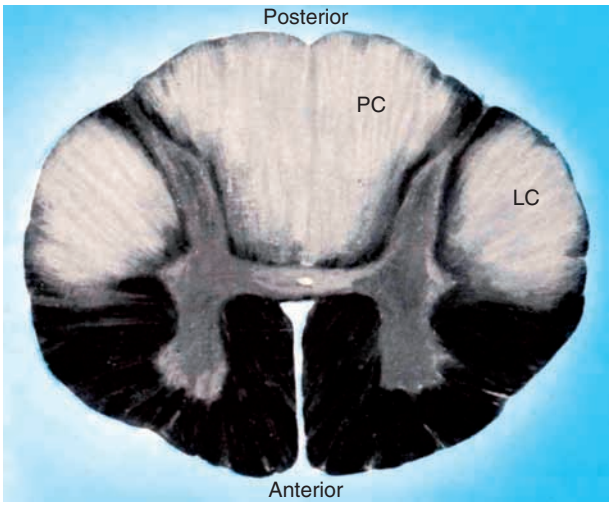

Figure 6 Subacute combined degeneration of the spinal cord is a progressive, degenerative disease involving the posterior (PC) and lateral columns (LC) of the spinal cord. It was first described in pernicious anemia due to vitamin B12 deficiency. In the mid-thoracic region of the spinal cord, there is axis cylinder and medullary sheath degeneration, involving the posterior columns, direct and indirect pyramidal tracts, and direct cerebellar tracts. Clinical signs in adults present insidiously with generalized weakness and vague gastrointestinal discomfort. Most characteristic is a sensation of 'pins and needles' in toes and fingers, with muscular atrophy in arms and legs. Touch, vibration, position, deep pain, and muscle sensibility are steadily impaired, and the Romberg test becomes positive. Pain is uncommon. The gait becomes spastic and ataxic with hyperactive deep reflexes and pathologic toe signs. The sphincters are seldom affected. In the terminal stages flaccid paraplegia, loss of reflexes, and mental changes appear. Spinal fluid studies show normal cells, glucose, and protein content.

Cerebral Folate Deficiency

CFD can be defined as a recognizable neurological syndrome associated with low CSF 5-methyltetrahydrofolate (5MTHF), the active folate metabolite, in the presence of normal folate metabolism outside the nervous system. CFD results from either disturbed folate transport or from increased folate turnover within the CNS.

An infantile-onset CFD syndrome is characterized by the presence of three or more typical features manifesting from the age of 4 to 6 months, starting with marked unrest, irritability, and sleep disturbances followed by psychomotor retardation, cerebellar ataxia, and hypotonia, pyramidal signs in the lower limbs which may ascend to spastic tetraplegia with age, if left untreated. In about one-third of the cases, dyskinesias (choreoathetosis, ballism) and epileptic seizures develop. Most children show deceleration of head growth from the age of 4 to 6 months. Visual disturbances are not invariably present but may start to develop around the age of 3 years. Progressive sensorineural hearing loss may begin to develop from the age of 6 years (Figure 7). Neuroimaging can be normal but in about 50% of the children, atrophy of frontotemporal regions and periventricular demyelination can be noted, and in a minority slowly progressive supra- and infratentorial atrophy can be found.

CSF analysis of the secreted membrane-derived FR protein could identify this protein as a nonfunctional FR protein with normal immunoreactivity but diminished binding to radioactively labeled folate. Because active folate transport to the CNS occurs mainly at the choroid plexus through folate-receptor-mediated endocytosis, the FR1 and FR 2 genes were sequenced and found to be normal. Further investigations have now clarified that the most important mechanism underlying infantile-onset CFD is the presence in serum of circulating autoantibodies, directed against the membrane-attached FR to choroid epithelial cells (Figure 8). These FR autoantibodies were characterized by a high affinity to the FR and were shown to block the binding of folate, thereby preventing its passage into the CSF and nervous tissues. Serum FR autoantibodies were absent among the asymptomatic mothers of the children affected by infantile CFD.

A particular subgroup of infantile-onset CFD patients manifest low-IQ autism, which is mostly associated with part of the neurological deficits described for infantile-onset CFD. In a high number of these autistic children serum FR autoantibodies have been identified.

Later onset spastic–ataxic syndromes followed by stagnation of mental development have been identified to be associated with CFD and also found to be associated with FR autoantibodies. The patients have normal development during the first 12–18 months and show no or only mild cognitive deficits.

The discovery of FR autoimmunity as a major cause of CFD has opened a new window on those neuropsychiatric conditions suspected to be linked to folate deficiency, although folate levels in serum and red blood cells, as well as homocysteine levels remain within the normal range. It may well be that a particular subgroup of patients with depression, schizophrenia-like syndromes, and dementias suffer from CFD. Clinicians should be alerted especially if

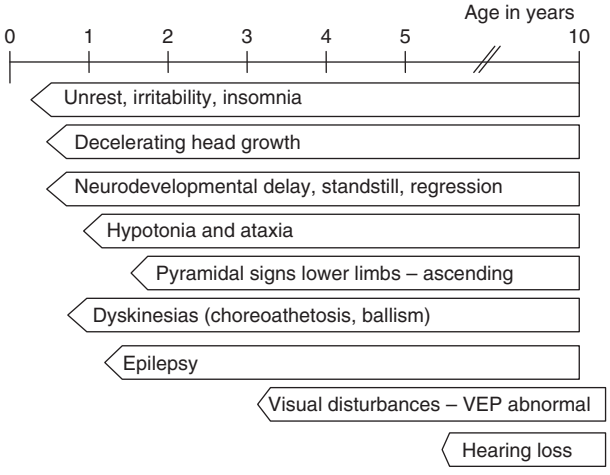

Figure 7 The figure displays the age of onset of the main clinical features. Visual and hearing loss are not consistent features.

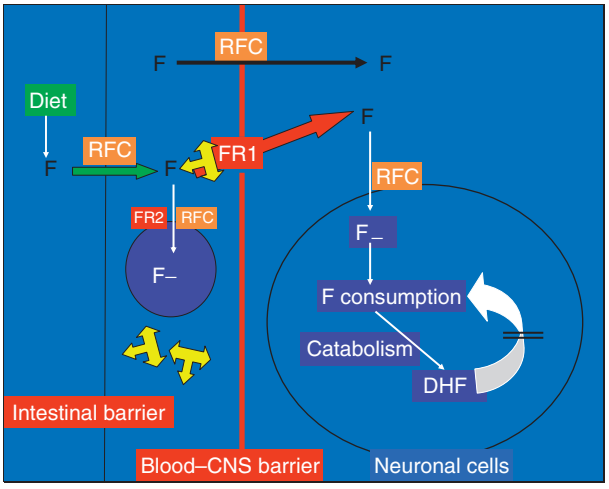

Figure 8 The FR autoantibodies circulating in serum will block the FRs attached at the basolateral surface of choroid epithelial cells and impair folate transfer to the CSF and from there to nervous tissues. Supplementation with high doses of reduced folates will increase the 5MTHF serum values by more than 100 times the physiological range, so that 5MTHF can pass the blood–CNS barriers through the low-affinity RFC1. The autoantibody against the FR is indicated as a yellow T. The arrow on the right represents the virtually absent dihydrofolate reductase activity in neuronal cells. DHF, sihydrofolate; RFC, reduced folate carrier; FR, folate receptor.

these patients do not respond to the conventional treatment modes and consider the possibility of CFD and FR autoimmunity. However, the prevalence of CFD and serum autoantibodies in these patients awaits further studies.

Genetic Studies and Inborn Errors of Folate Metabolism

Because systemic folate deficiency and the use of antifolate drugs in pregnant mothers was known to lead to congenital malformations, including neural tube defects, gene knockout experiments have been performed in mice to study the specific effects after knockout of the FR1, FR2, and RFC1 genes with respect to prenatal development.

To study whether the FR genes are involved in maternal-to-fetal folate transport, the FR genes in mice were inactivated. Nullizygous FR2−/− embryos developed normally, but FR1−/− embryos had severe morphogenetic abnormalities and died *in utero* by embryonic day 10. Supplementing pregnant FR1+/− dams with folinic acid reversed this phenotype in nullizygous pups. Decreasing the folate dose administered to FR1+/− dams resulted in a wide range of malformations of FR1−/− pups, involving the neural tube, craniofacies, eyes, and abdominal wall. These experiments in mice support the earlier findings that periconceptional folate supplementation reduces the occurrence of several human congenital malformations, including craniofacial, heart, and neural tube defects.

RFC1 knockout experiments also lead to intrauterine lethality in nullizygous mice. Following supplementation of RFC1+/− dams with a daily subcutaneous administration of 1 mg folic acid, near-normal development could be sustained in RFC−/− embryos, of whom 10% went on to live after birth but died within 12 days. These RFC−/− mice showed a marked absence of erythropoiesis in bone marrow, spleen, and liver with lymphoid depletion in the splenic white pulp and thymus. These data indicated that in the absence of RFC1 function, neonatal animals die due to failure of hematopoietic organs.

A hereditary congenital folate malabsorption is a very rare condition with preponderance in females. The patients suffer from an isolated defect in intestinal absorption of folic acid and a defect of its transport across the blood–brain barrier. The clinical manifestations are severe megaloblastic anemia, increased susceptibility to infections, mental retardation, recurrent seizures, movement disorders with ataxia or athetosis, and basal ganglia calcifications of the brain. Laboratory investigations confirm low folate levels in serum, red blood cells, and CSF. The genetic defect has not yet been identified. A number of rare inborn errors of folate metabolism are known to have a genetic basis (see also Table 2).

The most common inborn error of folate metabolism is MTHFR deficiency, a riboflavin-dependent enzyme which reduces the substrate 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, which is used for the remethylation of homocysteine to methionine. Deficient enzyme activity leads to homocysteine elevation and low methionine levels in plasma. There is no megaloblastic anemia because 5,10-methylenetetrahydrofolate, present in normal amounts, can still be utilized by thymidylate synthase for the synthesis of thymidine. Clinical severity varies and the age of onset ranges from the neonatal period to adulthood. The patients show features of neurodevelopmental delay, muscular weakness, seizures, gait abnormalities, incoordination, paresthesias, and psychiatric disturbances. There is a high risk of venous thrombosis and thromboembolism with stroke. Neuroimaging shows signs of cerebral atrophy and leukoencephalopathy. This autosomal recessive condition can be caused by different mutations of the MTHFR gene on chromosome 1p36.3.

Other rare inborn errors of folate metabolism are formiminoglutamate transferase deficiency and dihydrofolate reductase deficiency, of which only a few cases have been reported.

Diagnosis and Differential Diagnosis

Several rare conditions are known which are associated with secondary systemic or cerebral folate deficiency and should be considered in the differential diagnosis (Tables 1 and 2).

Autosomal recessive forms of severe methylenetetrahydrofolate reductase deficiency (MTHFR) should be considered, and leucocyte or fibroblast MTHFR activity should be measured in these patients.

The amino acids serine, glycine, and histidine act as the most important one-carbon donors to replenish the pool of unsubstituted tetrahydrofolates. Therefore, hereditary metabolic conditions associated with lowered levels of these amino acids can be expected to reduce secondarily the 5MTHF concentrations and thus the SAM-mediated methylation reactions. A *de novo* disorder of serine synthesis due to 3-phosphoglycerate dehydrogenase deficiency leads to a recognizable clinical entity with microcephaly, intractable seizures, feeding disturbances, and severe neurological disability. The presence of low serine in CSF and the brain will consequently reduce the substrate for the enzyme serine hydroxymethyltransferase in this disorder which will reduce the production of 5,10-methylenetetrahydrofolate, the precursor for 5MTHF and SAM.

Table 2 Congenital folate malabsorption and inborn errors of folate metabolism

Essential clinical features	*Diagnostic investigations*
Congenital folate malabsorption	
Megaloblastic anemia, mental retardation, ataxia, athetosis, seizures, susceptibility to infections, NMR: Basal ganglia calcifications	Intestinal folate malabsorption Megaloblastic anemia Folate deficiency (P, RBC, CSF) Low methionine (P) Correction of anemia after intramuscular folinic acid
Methylenetetrahydrofolate reductase deficiency	
Microcephaly, retardation, muscle weakness and neuropathy, seizures, thromboembolic events	Homocysteinemia and homocystinuria Low to normal plasma methionine Low MTHFR activity in lymphocytes or fibroblasts
Formimino transferase deficiency	
Retardation, convulsions, hypotonia and megaloblastic anaemia	Formimino glutamate excretion ↑(U) Hydantoin-5-propionic acid ↑ (U)
Dihydropteridine reductase (DHPR) deficiency	
Progressive neurological illness with microcephaly, myoclonic seizures, temperature instability, tone fluctuations, choreoathetosis, NMR: Basal ganglia and subcortical calcifications	Phenylketonuria unresponsive to diet CSF HVA↓, 5HIAA↓, 5MTHF↓ Neopterin normal or ↑, biopterin ↑↑ (U,CSF) RBC DHPR activity ↓ (Guthrie card)
3-Phosphoglycerate dehydrogenase deficiency	
Microcephaly, retardation, intractable seizures, feeding difficulties, NMR: Dysmyelination	CSF serine ↓↓, glycine ↓, 5MTHF ↓ Low enzyme activity in fibroblasts
Aromatic amino acid decarboxylase deficiency	
Low blood pressure, psychomotor retardation, oculogyric crises, eyelid ptosis, miosis, hypotonia, sweating, temperature instability, chorea, dystonia	Plasma epinephrine and serotonin ↓ CSF HVA ↓, 5HIAA ↓, L-dopa ↑, 3-*O*-methyl-dopa ↑ and 5-OH-tryptophane ↑ Vanillactic acid ↑ (U) L-Dopa decarboxylase activity ↓↓ (P)

P, plasma; RBCs, red blood cells, U, urine; HVA, homovanillic acid; 5HIAA, 5-hydroxyindoleacetic acid.

Other inborn errors of metabolism associated with secondary CFD include dihydropteridine reductase (DHPR) and aromatic L-amino acid decarboxylase deficiency. Diagnosis depends on a high index of clinical suspicion and typical findings from plasma amino acid levels and CSF amino acid levels, biogenic monoamine, and pterin metabolites.

DHPR deficiency (OMIM # 261630) is an autosomal recessively inherited disorder involving failure of reduction of quinonoid-dihydrobiopterin to tetrahydrobiopterin, the common cofactor of the aromatic amino acid hydroxylases phenylalanine-, tyrosine-, and tryptophane hydroxylases. The resulting enzyme dysfunction leads to elevated serum phenylalanine with reduced synthesis of the neurotransmitters dopamine and serotonin. Serum amino acid analysis will reveal hyperphenylalaninemia, and CSF analysis the reduced end-metabolites of dopamine and serotonin, that is, homovanillic acid and 5-hydroxyindole acetic acid. Children manifest with progressive mental and physical retardation, microcephaly, myoclonic seizures, hypotonia/hypertonia, swallowing difficulties, hypersalivation, choreoathetosis, and temperature instability. MRI and CT brain scans reveal cortical atrophy with subcortical and basal ganglia calcifications. Secondary CFD develops due to a suspected interaction of the accumulated quinonoid-dihydrobiopterin metabolites with folate metabolism resulting in lowered CSF 5MTHF.

Aromatic L-amino acid decarboxylase deficiency (OMIM # 608643) is another autosomal recessive disorder with secondary CFD as a result of increased utilization of 5MTHF and SAM which are required to convert the accumulated L-dopa to 3-ortho-methyldopa. The CSF pattern shows accumulation of the substrates 5-hydroxytryptophan, L-dopa and its converted metabolite 3-ortho-methyldopa together with lowered metabolites of the enzymatic reaction products serotonin and dopamine. The clinical picture manifests soon after birth and is associated with severe signs related to depletion of dopamine, epinephrine, and serotonin.

Recent reports have described secondary CFD to occur in a number of girls with Rett syndrome and patients with a variant or classical Aicardi–Goutières syndrome. Multicenter studies from Europe and Israel now have shown low 5MTHF to be present in CSF among about 50% of all Rett patients, independent of the presence of an MECP2 genotype. Rett patients with low CSF 5MTHF seem to suffer more frequently from epilepsy that is resistant to antiepileptic drugs. Further CSF studies of children with Rett

syndrome have also been performed in North America and were unable to detect low CSF folates in a large number of patients. The contrast between the North American and European studies on Rett syndrome is an interesting finding which might be linked to the folate fortification food program. The exact link between the MECP2 genotype and reduced folate transfer to the CNS remains to be established.

Kearns–Sayre syndrome is associated with clonal expansions of large-scale mitochondrial DNA rearrangements (i.e., partial Δ-mtDNAs and/or dup-mtDNAs) arising in the maternal oocyte or in early fetal life. It has long been known that CSF folate levels are reduced with or without concomitant reduction of plasma folate. The choroid plexus epithelial cells which harbor these mtDNA rearrangements will fail to produce sufficient ATP to transport 5MTHF actively across the blood–CSF barrier. Other mitochondrial encephalopathies are now known to also lead to reductions of active folate transport with resulting CFD, and are amenable to treatment with folinic acid (see **Figure 9**).

Treatment Guidelines and Outcome

Treatment should be adopted according to the origin and causes of the folate-deficiency state. For women of child-bearing age, no consensus exists about the optimal preconceptional daily dosage of folic acid, for which advice ranges from 400 μg up to 5 mg folic acid. The folate food fortification program in North America has achieved a dramatic reduction of the incidence of neural tube defects, whereas European efforts to prescribe folic acid to women on an individual basis have failed due to noncompliance.

In patients with systemic folate-deficiency states, the underlying cause should be clarified first by appropriate investigations, and promptly followed by the administration of currently available and registered folate compounds. The choice of the folate compound depends on the underlying condition. Folinic acid represents a stable form of a metabolically active reduced folate compound in contrast to folic acid, which is an oxidized and metabolically inactive folate form. Paradoxically, treatment with folic acid may further enhance CNS depletion of the metabolically active pool of reduced folates because folic acid has a higher FR1-binding affinity compared to 5MTHF and consequently competes at the FR1 receptor with the transport and storage of reduced folate species. Due to absent DHFR activity in the brain, the folic acid cannot be reduced to the active form of reduced tetrahydrofolate. Hunter et al. have previously described in a study on healthy volunteers that daily administration of extremely high doses of 15 mg folic acid for more than 1 month caused mental changes, sleep disturbances, and irritability. Some became hyperactive, whereas others became depressed and confused and had increasing difficulties with concentration. In addition, Hommes and Obbens reported that intravenous folate at extremely high doses (45–120 mg) caused generalized seizures in rats.

In infantile-onset CFD with or without concomitant FR autoantibodies, treatment is started with a daily oral administration of 0.5–1 mg kg^{-1} body weight folinic acid (calcium folinate), divided into two or three doses. Most children showed excellent tolerance. While treating children with CFD, caution should be exercised to avoid overdosage and neurotoxicity with folinic acid. Abrupt increases of the folinic acid dose by more than 0.5 mg kg^{-1} day^{-1} over a period less than one month must be avoided for safety reasons. Side-effects such as gastrointestinal upset, an increased risk of seizures, transient tics, and agitation have been observed in a minority of children treated with folinic acid; hence, careful clinical and EEG follow-up studies should be performed after 1, 3, and 6 months.

These extremely high folinic acid doses for children with CFD increase the serum 5MTHF values by a factor of more than 100 times the normal physiologic concentration. At these high 5MTHF concentrations in serum, 5MTHF can be bound by RFC1 at the blood–CNS barrier for further transport into the nervous system, and thus bypass the FRs blocked by autoantibodies. Although the group of youngest children with early diagnosis and treatment showed a favorable response with marked neurological recovery and cessation of seizures, the group of older children tended to show a more delayed response with incomplete neurological recovery. Mental deficits, in particular, persist if diagnosis and treatment are started at a later age (**Figure 10**), although folinic acid was still able to prevent further deterioration. In children who did not show any clinical effect or a doubtful effect after treatment during an initial period of 4–6 months at a folinic acid dose of 0.5–1 mg kg^{-1} day^{-1}, a lumbar puncture has been repeated in order to monitor the change of CSF 5MTHF levels. The reason for repeating the spinal fluid analysis after 4–6 months was to detect whether 5MTHF levels in CSF had normalized, remained low, or had become too high with the risk of clinical toxicity. Based upon the initial CSF levels together with the monitored levels during treatment, dose readjustments for folinic acid have been calculated.

Severe MTHFR deficiency is difficult to treat with poor outcome; high betaine doses can be tried to improve methylation and reduce homocysteine levels. Folinic acid treatment has been tried in patients with

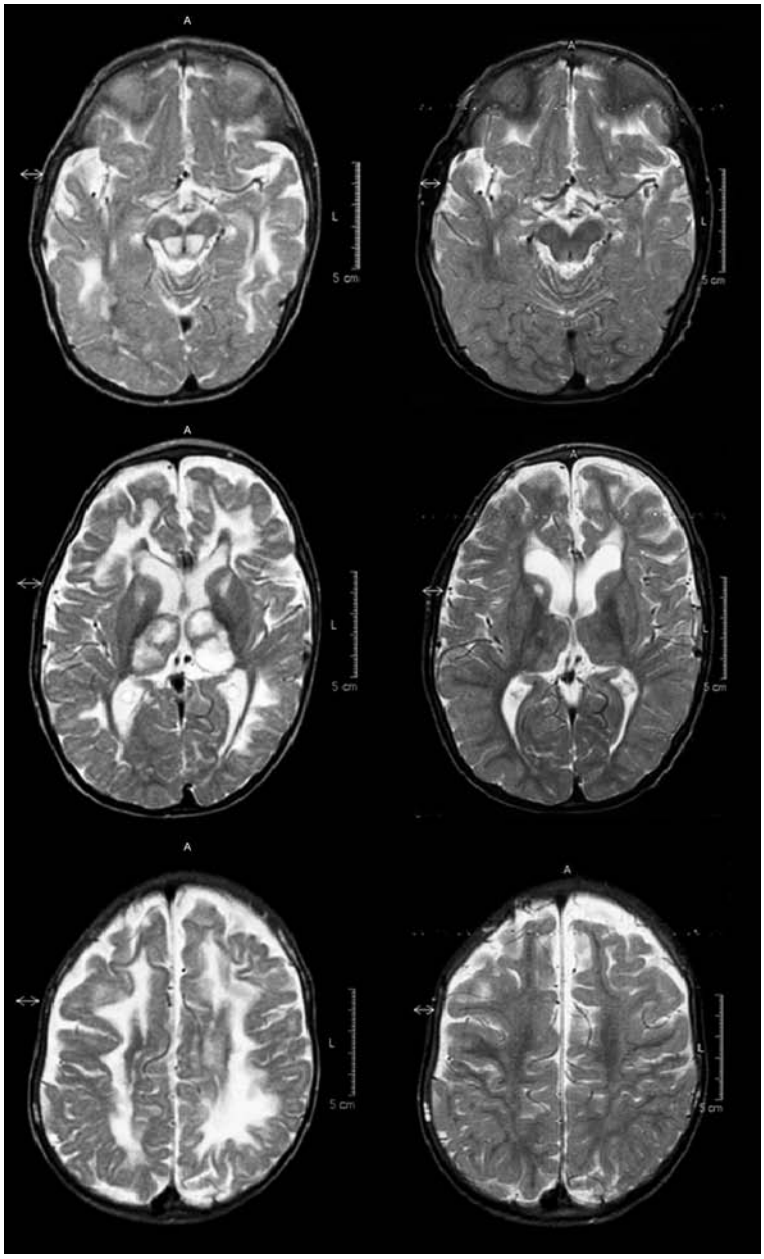

Figure 9 From top to bottom on the left, the brain MR imaging (T2-weighted) of a child affected by mitochondrial complex I deficiency with a history of congenital myopathy, psychomotor retardation, and resistant epilepsy, showing a severe leukoencephalopathy at the age of 18 months, involving white matter, brain stem, basal ganglia, and thalamus. Further CSF studies confirmed CFD. Treatment consisted of supplementation with the complex I cofactor Q10, the radical scavengers vitamin C and E, and high doses of folinic acid. After treatment at the age of 4 years, significant reversal of the leukoencephalopathy can be noted on the comparable axial images at the right side.

secondary CFD due to deficient ATP production and failure of active folate transport to the CNS. Kearns–Sayre syndrome has been treated with positive clinical effects and reversal of the leukoencephalopathy. In one patient with a myopathy, seizures, and severe leukoencephalopathy due to reduced activity of mitochondrial complex I, folinic acid therapy was combined with high doses of Q10, riboflavine, and the radical scavengers vitamin C and E (Figure 9). There was mild clinical recovery and a dramatic reversal of the leukoencephalopathy.

The food folate fortification in North America has not only reduced the incidence of neural tube defects, but in Ontario the occurrence of neuroblastoma has now been found to have declined. Future statistical analyses will reveal whether other conditions like

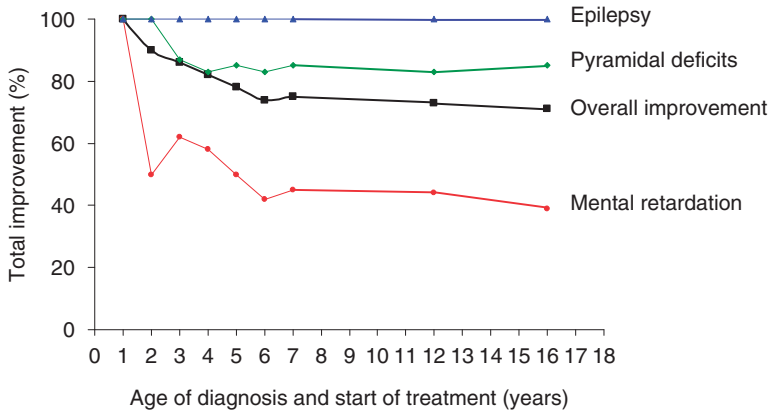

Figure 10 The overall improvement of infantile-onset CFD becomes better the earlier it is diagnosed and treatment started. The figure shows overall improvement in function of the age before which the child is diagnosed, expressed as the percentage of ameliorated signs and symptoms after folinic acid therapy during at least one year. Children treated before the age of 6 years show better results with even better prognosis at earlier ages. The red curve represents the percentage of recovery for mental deficits, which can be noticed to decline sharply between the age of 1 and 2 years. In contrast, control of epilepsy and pyramidal deficits showed more favorable responses after treatment. The critical time window to detect and treat infantile-onset CFD is very short and thus early treatment should be started to prevent persistent mental deficits.

cardiovascular disease, strokes, cancer, major depression, and psychiatric disease have declined since introduction of the folate fortification in food.

See also: Autism; Dementia; Depression and the Brain; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Schizophrenia: Genetics.

Further Reading

Beckman DR, Hoganson C, Berlow S, and Gilbert EF (1987) Pathological findings in 5,10-methylenetetrahydrofolate deficiency. *Birth Defects* 23: 47–64.

Blau N, Bonafe L, Krägeloh-Mann I, et al. (2003) Cerebrospinal fluid pterins and folates in Aicardi–Goutieres syndrome: A new phenotype. *Neurology* 61: 642–647.

Botez MI, Young SN, Bachevalier J, and Guthier S (1982) Effect of folic acid and vitamin B12 deficiencies on 5-hydroxyindoleacetic acid in human cerebrospinal fluid. *Annals of Neurology* 12: 479–484.

Corbeel L, Van den Berghe G, Jaeken J, et al. (1985) Congenital folate malabsorption. *European Journal of Pediatrics* 143: 284–290.

French AE, Grant R, Weitzman S, et al. (2003) Folic acid food fortification is associated with a decline in neuroblastoma. *Clinical Pharmacology and Therapeutics* 74(3): 288–294.

Ghosh SK, Syed SK, Jung S, et al. (1990) Substrate specificity for myelin basic protein-specific protein methylase I. *Biochimica et Biophysica Acta* 1039: 142–148.

Gospe SM, Gietzen DW, Summers PJ, et al. (1995) Behavioral and neurochemical changes in folate-deficient mice. *Physiology and Behavior* 58: 935–941.

Hansen FJ and Blau N (2005) Cerebral folate deficiency: Life-changing supplementation with folinic acid. *Molecular Genetics and Metabolism* 84: 371–373.

Holm J, Hansen SI, Hoier-Madsen M, and Bostad L (1991) High-affinity folate binding in human choroid plexus. *The Biochemical Journal* 280: 267–271.

Hommes OR and Obbens EA (1972) The epileptogenic action of Na-folate in the rat. *Journal of Neurological Sciences* 16: 271–281.

Hunter R, Barnes J, Oakeley HF, et al. (1970) Toxicity of folic acid given in pharmacological doses to healthy volunteers. *Lancet* 1: 61–63.

Piedrahita JA, Oetama B, Bennett G, et al. (1999) Mice lacking the folic acid-binding protein Folbp1 are defective in early embryonic development. *Nature Genetics* 23: 228–232.

Pineda M, Ormazabal A, Lopez-Gallardo E, et al. (2006) Cerebral folate deficiency and leukoencephalopathy caused by a mitochondrial DNA deletion. *Annals of Neurology* 59: 394–398.

Ramaekers V, Hansen SI, Holm J, et al. (2003) Reduced folate transport to the CNS in female Rett patients. *Neurology* 61: 506–514.

Ramaekers V, Häusler M, Opladen T, et al. (2002) Psychomotor retardation, spastic paraplegia, cerebellar ataxia, and dyskinesia associated with low 5-methyltetrahydrofolate in cerebrospinal fluid: A novel neurometabolic condition responding to folinic acid substitution. *Neuropediatrics* 33: 301–308.

Ramaekers VT and Blau N (2004) Cerebral folate deficiency. *Developmental Medicine and Child Neurology* 46: 843–851.

Ramaekers VT, Rothenberg SP, Sequeira J, et al. (2005) Autoantibodies against folate receptors of the choroid plexus block folate uptake and cause cerebral folate deficiency in childhood. *New England Journal of Medicine* 352: 1985–1991.

Ray JG, Meier C, Vermeulen MJ, et al. (2002) Association of neural tube defects and folic acid food fortification in Canada. *Lancet* 360: 2047–2048.

Reynolds EH, Preece JM, Bailey J, and Coppen A (1970) Folate deficiency in depressive illness. *British Journal of Psychiatry* 117: 287–292.

Rosenblatt D (1995) Inherited disorders of folate transport and metabolism. In: Scriver CR, Beaudet AL, Sly WS, and Valle D (eds.) *The Metabolic and Molecular Basis of Inherited Disease*, 6th edn., pp. 3111–3128. New York: McGraw-Hill.

Rothenberg SP, Da Costa MP, Sequeira JM, et al. (2004) Autoantibodies against folate receptors in women with a pregnancy complicated by a neural tube defect. *New England Journal of Medicine* 59(6): 410–411.

Spiegelstein O, Mitchell LE, Merriweather MY, et al. (2004) Embryonic development of folate binding protein-1 (Folbp1) knock-out mice: Effects of the chemical form, dose, and timing of maternal folate supplementation. *Developmental Dynamics* 231: 221–231.

Suh JR, Herbig AK, and Stover PJ (2001) New perspectives on folate catabolism. *Annual Review of Nutrition* 21: 255–282.

Surtees R (1993) Biochemical pathogenesis of subacute combined degeneration of the spinal cord and brain. *Journal of Inherited Metabolic Diseases* 16: 762–770.

Surtees R (2001) Cobalamin and folate responsive disorders. In: Baxter P (ed.) *Vitamin responsive Conditions in Paediatric Neurology, International Review of Child Neurology Series*, pp. 96–108. London: Mac Keith Press.

Wang Y, Zhao R, Russell RG, and Goldman ID (2001) Localization of the murine reduced folate carrier as assessed by immunohistochemical analysis. *Biochimica et Biophysica Acta* 1513: 49–54.

Zhao R, Russell RG, Wang Y, et al. (2001) Rescue of embryonic lethality in reduced folate carrier-deficient mice by maternal folic acid supplementation reveals early neonatal failure of hematopoietic organs. *The Journal of Biological Chemistry* 276(13): 10224–10228.

Relevant Website

http://www.folates.com – Cerebral Folates Deficiency.

Food and Water Intake: Regulation

E M Stricker, University of Pittsburgh, Pittsburgh, PA, USA
S C Woods, University of Cincinnati, Cincinnati, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Water intake and food intake are often discussed in the same chapter in textbooks and in related reviews of ingestive behavior or motivated behavior more generally. This practice reflects obvious similarities in the manifestation and control of the two behaviors, and it implies that analogies and/or overlap in their regulatory circuits are sufficient to consider them together rather than separately. It also reflects the fact that the two behaviors are interdependent, with changes in one often associated with changes in the other. As technology in neuroscience has become increasingly more sophisticated, it is now also recognized that control of the two behaviors involves parallel levels of sensory signals. This article reviews those similarities.

In considering the biological basis of water and food intake, the following general approach has long been taken. Proceeding from the assumption that a critical amount of each nutrient is needed, it was assumed (and sometimes demonstrated) that this amount is regulated by physiological processes, that a deficiency of each nutrient creates a specific excitatory stimulus to ingest the needed nutrient and thereby correct its deficiency, and that removal of the deficiency also removes the excitatory stimulus for the motivation to consume it. This simple negative feedback model reflecting the homeostatic control of ingestive behavior is, in fact, too simple; thirst and the control of water intake are much more complex than the schema suggests, whereas hunger and the control of food intake are not adequately represented by the schema.

Hunger and Food Intake

Because of limited storage capacity, most living cells cannot survive long intervals without an influx of new calories. If the animal is to survive, the sensation of hunger must therefore occur long before there is a generalized lack of calories throughout the body so as to ensure that energy will be available when needed. Thus, a strict depletion–repletion model cannot provide a suitable perspective from which to consider the control of food intake, at least at the organismal level, since hunger and eating generally occur when animals have ample stored nutrients to last over prolonged intervals of fasting.

Three quite different strategies have been proposed throughout the years to account for hunger and eating in animals that already have abundant stored calories. The first posits that critical sensors exist in the brain and elsewhere that are sensitive to local fluxes of nutrients, which herald impending demands on storage depots. The most popular of these models was the glucostatic hypothesis proposed by Jean Mayer 50 years ago, in which sensory cells in the hypothalamus were thought to be sensitive to local glucose. Decreases of glucose signaling in the hypothalamus were hypothesized to act as a bellwether for the organism as a whole and to elicit eating, whereas increases of glucose signaling were posited to stop eating. Consistent with this hypothesis, eating can be initiated by experimentally induced reductions of glucose signaling or glucose metabolism, whether systemically or locally in the hypothalamus, and meals generally end when the use of glucose as a fuel by the body is high. Analogous hypotheses have been proposed based on fluxes of lipid metabolism, of correlates of protein metabolism, and of available intracellular energy in the form of ATP. In every instance, specialized receptor cells were hypothesized to be exquisitely sensitive to small fluctuations and to inform brain areas controlling food intake.

The second strategy posits that the sensation of hunger does not need to be stimulated but is always present, and that animals are constantly stimulated to eat unless they have just eaten, in which case the inhibitory effects of satiation generated from the previous meal are still operative. Such an arrangement presumably evolved because, in the wild, the appearance of food is unpredictable, and animals can take advantage of opportunities to eat whenever they occur. Consistent with this perspective are laboratory findings that (1) it is far more difficult to precipitate eating in satiated rats than to stop eating in hungry rats; (2) the size of spontaneous meals eaten by laboratory rats is unpredictable, but meal sizes do predict how long they wait until they eat again; and (3) food deprivation does not cause a shortage of caloric nutrients in the circulation (in contrast to evidence of water deficiency during water deprivation). Hunger in this model is thought to be initiated by the gradual disappearance of inhibitory satiety stimuli that were generated by the previous meal, in contrast to a gradual increase in an excitatory hunger stimulus.

The third strategy is to dissociate meal initiation from fluxes of nutrients or satiety signals altogether, unless nutrients drop so low that emergency intake is

necessary for survival. Hunger and meal initiation are posited to occur based on learning, habits, and various social and other situations. This arrangement allows animals to take maximal advantage of their environment without having to respond to urgent signals when engaged in other activities. Any biological influence related to nutrient levels or needs is therefore exerted on how many calories are consumed once a meal begins and the fate of the ingested calories.

There are experimental data supporting each of these positions, and they are discussed as appropriate later. Although there is no clear consensus as to the signals that initiate hunger and/or eating, it is generally acknowledged that signals arising during meals contribute to a sense of fullness or satiation. Several types of signals are generated as food passes from the mouth to the stomach to the intestine, and their cumulative effect is to elicit satiation. Signals from the oral cavity (taste, texture, and swallowing stimuli) are not critical in this regard since satiation is not produced by allowing animals to eat (i.e., to taste, chew, and swallow food) but having the ingested food drain from the stomach through an open gastric fistula. This kind of sham eating persists for long intervals in fasted rats with no sign of slowing over several hours. Hence, satiation signals must arise distally to swallowing. One obvious candidate is gastric distension, detected by stretch receptors within the walls of the stomach. A general rule is that the greater the distension, the greater the inhibition of eating. This afferent signal projects to the caudal brain stem via the gastric branch of the vagus nerve, and animals eat larger than normal meals when these vagal fibers are acutely disabled.

Whereas signals from the stomach indicate the volume of ingested material, signals from the small intestine indicate the nutrient content of ingested food. Because the digestion of different macronutrients (carbohydrates, lipids, and proteins) requires different enzymes and cofactors, sensory cells lining the intestinal lumen have the ability to analyze what has been eaten and to initiate the secretion of the appropriate blend of digestive juices from the pancreas, liver, and intestinal cells. These sensory cells secrete peptidergic signals that act locally (paracrine action) or hormonally to orchestrate digestion. Some of these same peptidergic signals convey messages to the brain coding the macronutrient and caloric content of the ingested food, and these are collectively called satiation signals.

The best known satiation signal is cholecystokinin (CCK), a peptide secreted from duodenal I cells in response to lipids in the ingested food. Administration of exogenous CCK reduces meal size and elicits feelings of fullness in humans. Administration of a selective CCK receptor antagonist increases meal size, implying that endogenous CCK normally contributes to limiting meal size. CCK secreted from the duodenum in response to food acts locally on CCK receptors on sensory vagal neurons, and CCK is no longer able to reduce meal size when these neurons are severed or their destination point in the hindbrain is destroyed. Importantly, the same vagal axons that convey the CCK signal have branches that convey the signal indicating stretch from the stomach, and there is evidence that a potentiated signal to the brain is provided by the synergistic interaction of the two signals – one indicating the quantity of food ingested and the other indicating an aspect of the quality. CCK is but one of many satiation peptides secreted from the intestine in response to the contents of a meal. Although specifics of the signaling system vary slightly among them, all have in common the ability to reduce meal size and to converge in the hindbrain where they are integrated and relayed to appropriate circuits that control meal continuation.

Calories can also be detected in the liver, and there is evidence that hepatic signals are relayed to the hindbrain via the vagus nerves. Therefore, as food is eaten and processed there is a continuous wave of information informing the brain as to how much and which nutrients have been consumed. Early signals from the stomach convey mainly volume, and these become integrated with duodenal and other intestinal signals as the meal progresses. At the same time, insulin is secreted from the pancreas into the hepatic portal vein, and there is evidence that it potentiates the hepatic signals. The important point is that the brain receives continuous information and integrates it with other factors to decide when to stop an ongoing meal.

In addition to compensating for short-term fluctuations in metabolism and energy expenditure, food intake also interacts directly with the organism's energy stores. The energy equation states that body weight will change over time if energy intake does not equal energy expenditure. It has been recognized for more than 50 years that body weight is regulated, and that circulating signals tonically inform the brain about how much fat is present in the body and whether more or less food needs be eaten to balance input with expenditure. Adiposity signals include the hormones insulin and leptin. Both insulin, from the pancreatic islets, and leptin, from white adipose tissue, are secreted and circulate in the blood in direct proportion to body fat. Both are transported through the blood–brain barrier directly into the hypothalamus and other brain areas where they interact with specific receptors on neurons. When either insulin or leptin is increased locally in or near the ventral

hypothalamus, animals act as if they have suddenly gained weight; that is, they eat less and lose weight. Conversely, when the insulin or leptin signal is reduced in the hypothalamus, animals eat more and gain weight.

Because insulin levels are more closely related to the amount of visceral fat, and leptin is more closely related to the amount of subcutaneous fat, the brain receives an integrated signal from the two peptide hormones. This combined adiposity signal is then integrated with information from the hindbrain regarding the progress of meals. When the insulin or leptin signal is elevated (indicating more stored energy), animals are more sensitive to satiation signals such as CCK and they eat smaller meals. In contrast, when the insulin or leptin signal is reduced, CCK is less potent and larger meals are consumed. Hence, adiposity signals combine with satiation signals to help maintain fat stores over time.

Satiation signals and adiposity signals act against a background of other factors to determine food intake. These include psychological factors such as learning, habits, the social situation, and stress, as well as physical factors. Among the latter are dehydration and nausea, both of which greatly reduce intake, and phase of the ovarian cycle in females.

In summary, food intake is a complex behavior. Most animals eat in discrete bouts or meals, and there is considerable evidence that the factors that control when meals occur (i.e., meal onset) are quite different than those that control meal size. The amount of food eaten in a given meal is determined by the quantity and quality of the food (via phasic satiation signals), by fluctuations in body fat stores (via tonic adiposity signals), and by myriad other factors. Although the impact of any of these factors may be small in a particular meal, collectively they ensure that energy stores are maintained over time when ample food is available.

Thirst and Water Intake

Water deprivation elicits the sensation of thirst when certain brain cells become dehydrated by as little as 1% or 2% by the osmotic loss of water. Those cells are located in the basal forebrain, just rostral to the third cerebral ventricle. They exist outside the blood–brain barrier and consequently they respond readily to changes in the particulate concentration, or osmolality, of plasma (pOsm). Although they are not uniquely sensitive to dehydration, such as retinal cells that detect only photons, they do have unique synaptic connections to other brain neurons that stimulate thirst. When the 'osmoreceptor cells' are impaired, either by experimental lesions or by brain disease, animals become blind to increases in pOsm and do not experience thirst normally.

This arrangement whereby osmoreceptors cells detect small increases in pOsm and initiate thirst and water intake, which consequently lower pOsm and thereby eliminate the excitatory signal for behavior, allows the possibility that osmoregulation is controlled by a single-loop negative feedback system. However, a major obstacle to this possibility is that ingested water becomes equilibrated with body fluids only after it is emptied from the stomach, and therefore it cannot provide rapid rehydration and negative feedback in the control of water intake. Although this delay would be expected to lead to continued consumption of large volumes of water in excess of amounts needed for rehydration, ample evidence indicates that such overconsumption of water by dehydrated animals does not occur, and that many species, including humans, replace water deficits by drinking large volumes of water very rapidly and then stopping abruptly. Thus, some early signal must inhibit thirst. Actually, an appropriate signal has been identified, which originates in the oropharynx and is derived from the rapid swallowing of liquids. In support of this conclusion are findings that (1) dehydrated dogs reduce thirst well before decreases in systemic pOsm can be observed; (2) the rapid cessation of thirst also occurs both when ingested water drains through an open gastric fistula before reaching the small intestine and when dehydrated dogs drink isotonic saline; and (3) no effect on thirst is found when water is infused directly into the stomach, bypassing the oropharynx, until the administered fluid decreases systemic pOsm. These striking observations collectively provide strong support for the proposal that some component of the act of drinking, such as rapid swallowing, generates an early stimulus that inhibits water intake in dogs.

These seminal findings have been replicated and extended to other species, including monkeys, sheep, and humans. The repeated confirmations implied that volume-dependent oropharyngeal signals were a general feature in the inhibition of thirst in animals. However, most investigations of the central controls of fluid homeostasis now use laboratory rats as experimental subjects, and whereas dehydrated rats also use presystemic signals to modulate thirst, the signals in rats are not oropharyngeal. In addition, in the rodent system early signals can provide both stimulation and inhibition of thirst, and the early signals are associated with both the volume and concentration of ingested fluid.

However, even with the inclusion of presystemic signals, our consideration of thirst is not complete; there are multiple signals of thirst, not just those resulting from osmoregulatory needs. Animals deprived of drinking water lose water from plasma in addition

to water from cells, and the loss of plasma volume (hypovolemia) is itself a stimulus of thirst. Indeed, thirst can be elicited even when no increase in pOsm occurs, such as following hemorrhage. Animals detect deficits in blood volume by stretch receptors embedded in the distensible walls of the inferior vena cava (which delivers much of the venous return to the heart) and of the right atrium. The stretch of the vessels is proportional to the volume contained therein, so when blood volumes are low the sensory neurons send an afferent signal of hypovolemia to the caudal brain stem, which then relays that message to the forebrain for stimulation of thirst.

One problem with this arrangement, however, is that ingested water, when absorbed, does not repair the plasma volume deficits that stimulated thirst. Instead, approximately two-thirds of the water moves by osmosis into cells. That outcome is desirable when thirst is associated with increased pOsm and cellular dehydration but not when plasma volume is diminished and pOsm is not elevated; then, water consumption only causes osmotic dilution without much correction of hypovolemia. Therefore, it should not be surprising that merely 3–5% osmotic dilution provides a potent stimulus for inhibiting thirst even in the presence of marked hypovolemia. This inhibition of hypovolemic thirst by osmotic dilution may be contrasted with the satiation of osmoregulatory thirst that occurs when appropriate amounts of water are consumed.

The inhibition of water intake despite hypovolemia usefully prevents osmotic dilution from becoming severe but it does not repair the plasma volume deficits that stimulated thirst in the first place. To restore those volumes, animals must ingest plasma or an equivalently dilute NaCl solution. Having first drunk water due to thirst, they must then consume salt. In fact, rats have been shown to drink water and concentrated NaCl solution in appropriate amounts, moving back and forth between the two solutions, to create the isotonic NaCl mixture that is ideal for plasma volume restoration. Space does not allow description of the central mechanisms for this control of thirst and salt appetite during hypovolemia, although much of this information is now known.

It is important to note that thirst in response to plasma volume deficits is not eliminated after destruction of the sites in the caudal brain stem that receive neuronal projections from the cardiovascular stretch receptors that detect hypovolemia. This finding indicates that another stimulus of thirst exists during hypovolemia. This signal likely is provided by angiotensin, a peptide hormone formed in the blood after secretion of the enzyme renin from the kidneys. Angiotensin also stimulates salt appetite as well as the hormones that enable the retention of water and sodium in urine, and it is a very potent vasoconstrictor agent as well (thus helping to support blood pressure during hypovolemia). By having so many functionally related actions, angiotensin ensures that diverse behavioral and physiological responses to hypovolemia occur simultaneously. Note that a stimulus of thirst also results from acute decreases in arterial blood pressure, and that mediation of this thirst by angiotensin, once controversial, has been established.

Circulating angiotensin acts in the brain at the subfornical organ, which is located in the dorsal portion of the third cerebral ventricle. Local angiotensin receptors can respond to elevations in systemic levels of the hormone because the subfornical organ lacks a blood–brain barrier, thus allowing angiotensin to diffuse into that region of the brain. Surgical destruction of this brain structure eliminates thirst and salt appetite stimulated by angiotensin. However, it does not abolish these dual effects of hypovolemia, which indicates that redundant mechanisms can be used to detect the loss of plasma volume and activate appropriate behavioral responses. Such redundancy should not be surprising given the great significance of adequate blood volume to life.

Conclusions

Despite the superficial similarity of food and water consumption, the mechanisms by which the two ingestive behaviors are controlled differ greatly from one another. This difference is in accord with other fundamental features of caloric and water homeostasis, which also differ greatly. For example, water deprivation produces cell dehydration and hypovolemia, whereas food deprivation does not compromise cellular metabolism. Water consumed in excess is excreted readily in urine, whereas food consumed in excess is stored in liver and adipose tissue. Evidently, calories are a more precious commodity than water.

Food intake is not likely to be controlled according to a depletion–repletion scheme whereby hunger is generated by caloric deficiencies. As mentioned previously, the autonomic recruitment of stored calories prevents such deficiencies, and normally there are no apparent need-based excitatory signals for hunger. Instead, multiple signals for inhibiting food intake have been identified. Some of these satiation signals are specific to food ingestion (e.g., CCK, postgastric delivery of calories, and leptin), whereas others are nonspecific and disrupt all ingestive behavior (e.g., gastric distension and nausea). Thus, brain mechanisms that control food intake generally respond to inhibitory signals, both neural and blood-borne, some of which are related to recently ingested and stored calories and some that are not.

In contrast, water intake is controlled according to a modified depletion–repletion scheme whereby bodily water loss generates signals that stimulate thirst and water consumption, which ultimately repairs dehydration. Two neural signals of thirst have been identified – one associated with increased plasma osmolality and cellular dehydration in brain osmoreceptor cells and one associated with decreased plasma volume as detected by cardiovascular receptors. A third, humoral signal of thirst, provided by the peptide hormone angiotensin II, arises from the periphery during hypovolemia and arterial hypotension and acts in the brain. Water intake satiates osmoregulatory thirst but inhibits thirst due to hypovolemia or angiotensin II, and salt consumption (also stimulated by hypovolemia and angiotensin) is required to repair plasma volume deficits and thereby satiate thirst. Thus, brain mechanisms respond to excitatory and inhibitory signals, both neural and blood-borne, that integrate thirst and salt appetite.

During the past four decades, numerous studies have sought to define the biological bases of hunger, food intake, and satiation. Two general perspectives on the problem have been recognized. One views food intake in relation to body weight maintenance and the influence of leptin and insulin, whereas the other views food intake in relation to the control of meals and the influence of gastric distension and intestinal nutrients. In other words, one perspective involves systemic signals associated with body fat stores and the other involves presystemic signals that occur during a meal but before ingested nutrients reach the systemic circulation. The presystemic signals allow animals to prepare for storage and utilization of the nutrients by providing information about what has been consumed and in what quantity, and thereby allow the animals to regulate the delivery of nutrients from the stomach to the small intestine (i.e., gastric emptying) and to vary it according to the concentration of gastric nutrients and bodily needs. Analogously, the control of drinking bouts is influenced by systemic signals associated with body fluid stores and derived from the volume, osmolality, and pressure of circulating blood. The same benefits result when thirsty animals are informed of the volume and content of ingested fluid before it reaches the circulation. The animals then modulate fluid consumption and do not drink in excess, and they modulate gastric emptying of ingested fluid; they also modulate secretion of vasopressin in anticipation of changes in blood osmolality and volume that occur once ingested fluid reaches the small intestine, equilibrates osmotically with body fluids, and is absorbed.

Both food and water intake are complemented by secretion of a peptide hormone – insulin and vasopressin, respectively – that conserves glucose and water, respectively. Eating stimulates insulin secretion, as does the ingested food; hence the coexistence of the behavioral and physiological responses. In contrast, many of the same signals that stimulate thirst also stimulate vasopressin secretion; hence the coexistence of the behavioral and physiological responses. In both cases, the absence of the hormone leads to pronounced urinary nutrient loss and compensatory increases in intake (i.e., diabetes insipidus and diabetes mellitus, respectively). Food consumed in excess does not eliminate renal glucose conservation, and in consequence fat stores increase (although large fat stores may diminish further storage). In contrast, water consumed in excess rapidly dilutes body fluids, thereby inhibiting the hormone responsible for renal water conservation and allowing excretion of the excess fluid in urine.

Multiple signals, both excitatory and inhibitory, control food and water intake. In each case, the signals involve various peptide hormones, afferent neural messages, and substrate levels. Hunger can be inhibited by CCK, insulin, and leptin, along with neural signals from the stomach (in response to gastric distension) and liver (in response to the hepatic delivery of utilizable nutrients), whereas hunger can be stimulated by ghrelin and the absence of satiety signals. Conversely, thirst can be stimulated by angiotensin II along with neural signals from cerebral and visceral osmoreceptors (in response to osmotic loads) and from cardiac baroreceptors (in response to hypovolemia), whereas thirst can be inhibited by atrial natriuretic peptide (which blocks angiotensin II receptors) and by neural signals from cerebral and visceral osmoreceptors (in response to osmotic loads), from arterial baroreceptors (in response to acute hypertension), and from presystemic receptors (in response to volume-dependent signals of rapid swallowing (e.g., in dogs and humans) or of gastric and small intestine distension (e.g., in rats).

Integrated central nervous system circuitry has been identified for a dual-control system that responds to excitatory and inhibitory signals in the control of food intake. One feature of such a system is that eating onset depends on the relative magnitude of the collective signals. Thus, a small excitatory signal can be effective in eliciting eating in the presence of even smaller inhibition, whereas great inhibition precludes eating onset. In contrast, the control of water intake is influenced largely by the appearance and disappearance of excitatory signals, with early cessation provided by acute and potent inhibitory signals that are stimulated by drinking. The relevant receptors that generate these signals have been identified, as have most of their anatomical interconnections.

The control of food intake is complex, at least in humans, and involves many factors unrelated to

caloric homeostasis and hunger, such as social pressures, food palatability, food appearance, and mood. Thus, despite the absence of hunger, individuals tend to eat when in a group that is eating, when presented with food that is especially tasty or attractive (e.g., dessert), or when stressed (i.e., the calming effect of eating) or bored (i.e., the stimulating effect of food). The control of water intake is much less complex because it involves few factors other than those related to fluid homeostasis. Nonetheless, in the absence of true thirst, fluid can be consumed for reasons of palatability (e.g., flavored drinks) or to obtain a desired chemical in the beverage (e.g., alcohol and caffeine). It seems likely that much of contemporary human food and fluid ingestion in affluent Western culture is unrelated to homeostatic needs and actually creates imbalances caused by nutrient excess.

See also: Angiotensin II; Energy Homeostasis: Adiposity Signals; Energy Homeostasis: Hypothalamic Development; Gastrointestinal Signals: Stimulation; Gastrointestinal Signals: Satiety; Gastrointestinal Tract role in Neural Control of Metabolism, Food Intake and Body Weight: A Summary; Hypothalamic Structure–Function Relationships; Neuropeptides: Food Intake; Salt Appetite; Thirst.

Further Reading

Fitzsimons JT (1998) Angiotensin, thirst, and sodium appetite. *Physiological Reviews* 78: 583–686.

Seeley RJ and Woods SC (2003) Glucostatic versus lipostatic regulation of energy balance by the CNS. *Nature Reviews Neuroscience* 4: 902–909.

Strader AD and Woods SC (2005) Gastrointestinal hormones and food intake. *Gastroenterology* 128: 175–191.

Stricker EM (ed.) (1990) *Handbook of Behavioral Neurobiology,* vol. 10. New York: Plenum.

Stricker EM (2004) Thirst. *Handbook of Behavioral Neurobiology* 14: 505–543.

Stricker EM and Verbalis JG (2002) Fluid intake and homeostasis. In: Squire LR, Bloom FE, Roberts JL, Zigmond MJ, McConnell SK, and Spitzer NC (eds.) *Fundamental Neuroscience,* 2nd edn., pp. 991–1009. San Diego: Academic Press.

Stricker EM and Woods SC (eds.) (2004) *Handbook of Behavioral Neurobiology* vol. 14. New York: Kluwer Academic Plenum.

Woods SC and Stricker EM (2002) Central control of food intake. In: Squire LR, Bloom FE, Roberts JL, Zigmond MJ, McConnell SK, and Spitzer NC (eds.) *Fundamental Neuroscience,* 2nd edn., pp. 1011–1029. San Diego: Academic Press.

Forebrain Development: Holoprosencephaly (HPE)

E S Monuki, University of California at Irvine, Irvine, CA, USA

Published by Elsevier Ltd.

Introduction and Scope

Holoprosencephaly (HPE) has long fascinated clinicians, basic scientists, and the public alike. This fascination results largely from its striking phenotype, which includes features such as a single midline eye (cyclopia) and a single forebrain holosphere rather than two hemispheres (Figure 1(a)). HPE also represents the most common congenital birth defect of the human forebrain (1 in 10 000 live births, 1 in 250 conceptions). Although previously considered to result from defective cleavage (i.e., splitting of the eye and forebrain fields), it is now well appreciated that the forebrain malformation results from primary defects in midline induction. Correspondingly, when factors needed to adopt midline fates are deficient, HPE is commonly observed. HPE, therefore, serves as a central paradigm for understanding how forebrain midline fates are specified.

The focus of this article is to describe forebrain midline development as revealed by studies on the cells and molecules implicated in HPE. After a brief description of HPE, we describe the following as they relate to forebrain midline development: (1) the organizers (signaling centers), (2) the genes and signaling pathways, and (3) the interactions among signaling pathways. Among the highlights are a tidy categorization of causal HPE genes into four well-known signaling pathways and the recent delineation of interactions among these pathways that help to explain longstanding conundrums about HPE phenotypes.

Definition of Holoprosencephaly

Primary midline failure in HPE has myriad secondary effects on forebrain development, which often lead to confusion and error regarding the HPE designation. Morphologically, the brunt of HPE neuropathology occurs at the midline rather than in more lateral structures. This includes marked reduction to total absence of midline tissues and, correspondingly, the failed separation of adjacent forebrain structures that are normally bilateral (Figure 1(b)). (Although often referred to as fusion, the term failed separation is more accurate.) At the molecular level, these morphological defects equate to absence of midline marker expression and aberrant continuity of lateral markers across the midline. This definition helps to exclude disorders such as microencephaly (small brain size), absence of the corpus callosum, hydrocephalus and absence of the septum pellucidum (which give the appearance of fused ventricles), and aprosencephaly/atelencephaly (absence of forebrain/telencephalon).

HPE can be divided into two categories with qualitatively different midline phenotypes: (1) classic and (2) middle interhemispheric (MIH) (Figure 1(a)). Classic HPE is further divided into alobar, semilobar, and lobar subtypes based on the degree of severity, with alobar being the most severe. The feature common to classic HPE subtypes is ventral predominance – that is, the neuropathology is most severe ventrally and then extends variably and in graded fashion to the rostral, dorsal, and posterior domains. In contrast, the MIH form (also known as syntelencephaly) exhibits midline failure that is restricted to the dorsal forebrain, generally affecting the posterior frontal and parietal lobes in humans (Figure 1(a)). Notably, MIH HPE lacks significant pathology away from the dorsal region – that is, the rostral and ventral domains, including the eyes and face, are relatively unaffected.

Organizers in Forebrain Midline Induction

Like other tissues, the development of the forebrain midline is governed by morphogenetic movements and organizers (signaling centers), which are localized groups of cells that produce morphogens. Morphogens are molecules that promote different cell fates at different concentrations. The key signals identified in forebrain midline induction are secreted proteins that are well-known morphogens in development.

Conceptually, induction of the forebrain midline occurs after neural induction, anterior–posterior (AP) patterning, and forebrain specification. Despite this useful conceptual distinction, however, there is significant overlap among these processes in terms of timing and the factors that govern them. For example, many of the signaling centers that govern neural induction, AP patterning, and forebrain specification are located at the embryonic midline, and these same signaling centers regulate forebrain midline induction. Thus, to provide an appropriate and comprehensive framework, we first summarize the earliest stages of forebrain development.

Nonneural Organizers of the Neural Plate

In the early embryo, forebrain development is regulated by organizers that lie outside the developing nervous system (i.e., in nonneural tissues). These

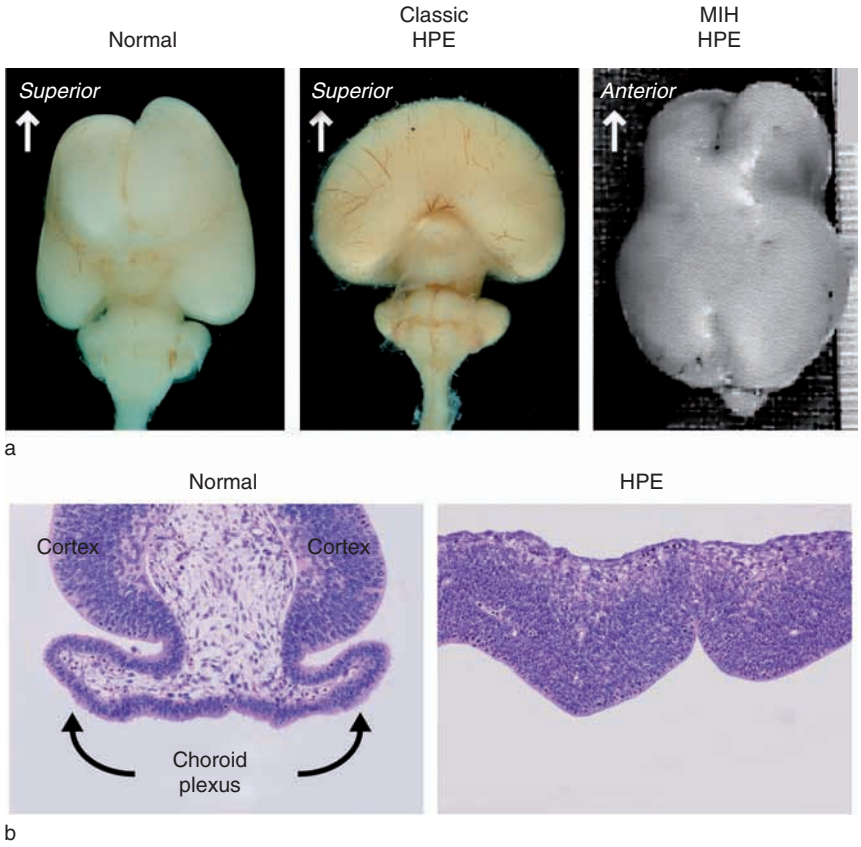

Figure 1 Holoprosencephaly (HPE): (a) normal forebrain compared to classic and middle interhemispheric (MIH) HPE in human fetuses; (b) HPE histology. In (a), the normal forebrain is composed of two hemispheres separated by an interhemispheric fissure. In severe (alobar) classic HPE, hemispheric separation is undetectable throughout the forebrain. Classic HPE neuropathology is always most severe ventrally, with variable extension to the rostral and dorsal domains. In MIH HPE, the lack of hemispheric separation is restricted to the dorsal regions, with no apparent extension to rostral or ventral domains. In (b), central to all types of HPE is failed midline induction, which results in absent midline structures and the continuity of lateral tissues across the midline. In this example of MIH HPE (following roof plate ablation), mutant mice lack choroid plexus and have a continuous cortex across the dorsal midline. Ventral tissues in (b) removed for visual clarity. (a, right panel) Reproduced from Marcorelles P, Loget P, Fallet-Bianco C, et al. (2002) Unusual varient of holoprosencephaly in monosomy 13q. In: *Pediatric and Developmental Pathology*, vol. 5, figure 2a, p. 174, with permission from Springer Science and Business Media. (a (left, middle), b) Reproduced with permission from the Journal of Neuropathology and Experimental Neurology.

nonneural organizers act before and during gastrulation to induce neural fate and AP patterning and then induce organizers within the forebrain tissue itself. Induction of these local neural organizers is often homeogenetic in nature – the nonneural cells instruct their neural neighbors to adopt a fate similar to its own – with the homeogenetic signals being the morphogens themselves. Important nonneural organizers of the forebrain include the node, anterior visceral endoderm (AVE), epidermal ectoderm, and prechordal plate.

The Node and Anterior Visceral Endoderm: Morphogen Antagonists and Forebrain Specification

Neural induction The node and AVE have primary importance in neural induction in mammals (Figure 2). Despite lingering questions and apparent differences among species, antagonism of bone morphogenetic protein (Bmp) signaling consistently takes center stage in the process of neural induction, and both the node and AVE are rich sources of Bmp antagonists, such as Noggin and Chordin. In some species, Bmp antagonism suffices for neural induction, whereas additional competence signal(s) – most likely fibroblast growth factors (Fgfs) – seem to be required in others.

Forebrain specification The AVE, which eventually underlies the future forebrain after cell movements and rotation (Figure 2(a)), has been shown in mice and chicks to regulate AP patterning of the neural plate. In addition, the AVE is necessary and sufficient for forebrain specification. Multiple AVE-derived

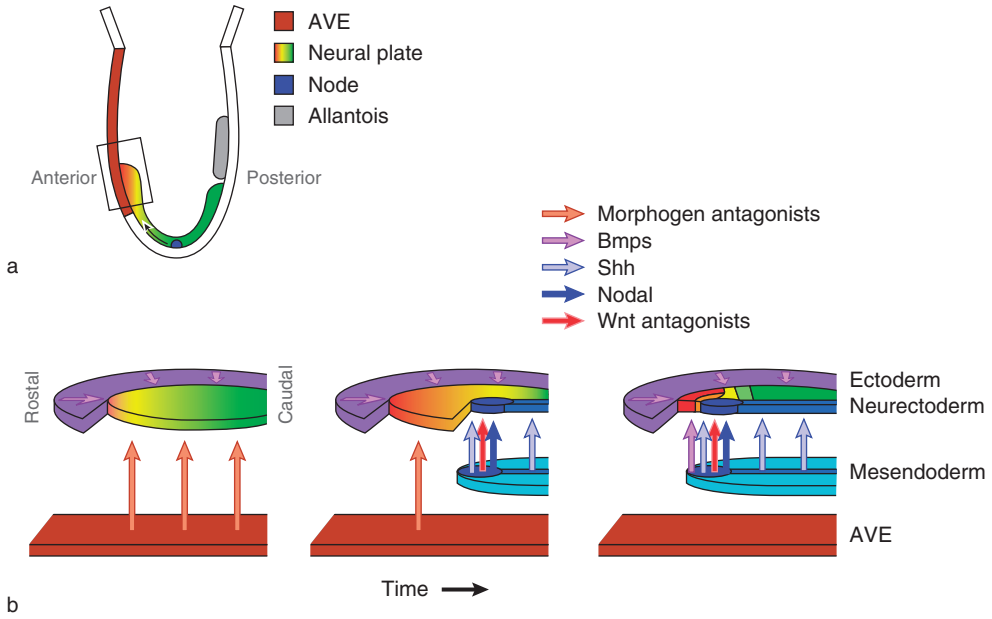

Figure 2 Nonneural organizers of the forebrain: (a) schematic cross section of early embryo; (b) patterning of the anterior neural plate by nonneural organizers (tissue layers separated for visual clarity). In (a), the extraembryonic anterior visceral endoderm (AVE) has rotated to underlie the future forebrain while cells in and around the embryonic node intercalate and migrate rostrally between the ectoderm and endoderm to form the prechordal plate and notochord. In (b), before prechordal plate arrival, the forebrain anlage is adjacent to the AVE (brown) and nonneural ectoderm (purple). AVE-derived morphogen antagonists are involved in AP patterning and forebrain specification. Antagonism of Bmp signaling is central to the earlier process of neural induction, but the roles of ectoderm-derived Bmps at the neural plate stage remain largely undefined. The prechordal plate and associated mesendoderm (blue) intercalate between the AVE and neural plate (neurectoderm). The prechordal plate produces multiple signals (Nodal, Shh, Bmp7) that influence rostral morphogenesis and induction of midline neural tissue. AP, anterior–posterior; Bmp, bone morphogenetic protein; Shh, Sonic hedgehog; Wnt, Wingless-int. (b) Adapted from *Developmental Cell*, vol. 6, Wilson SW and Houart C, Early steps in the development of the forebrain, p. 171, 2004, with permission from Elsevier.

signals mediate these processes, including cerebrus and dickkopf, which antagonize Bmp, Wingless-int (Wnt), and Nodal signaling. Antagonism of multiple morphogens, many of which caudalize neural tissue, is central to AVE function. Thus, AVE-dependent forebrain specification is not instructive *per se* but, instead, involves protecting the forebrain from factors that impart caudal character, such as Wnts, Fgfs, and retinoic acid. This has contributed to the so-called default model in which forebrain fate represents the default state of neural tissue unexposed to patterning signals. Eventually, AVE signaling to neural tissue is altered by the intercalation of mesendoderm during gastrulation (Figure 2(b)).

Role in anterior neural border specification The AVE has been implicated in specifying the anterior neural border (ANB), the local neural organizer at the edge of the rostal neural plate. Removal of the AVE or inactivation of its morphogen antagonists results in the absence of the ANB and telencephalon. In addition to cerebrus and dickkopf, the AVE produces secreted Frizzled-related proteins (sFRPs), which are (Figure 3), Wnt antagonists, and Wnt antagonism is known to be required for ANB formation.

Epidermal Ectoderm: Bmps and Roof Plate Induction

Bmp antagonism in neural induction Continuous with the neural plate is nonneural epidermal ectoderm, the future skin. Prior to and during neural induction, the ectoderm possesses high Bmp signaling levels, and antagonism of this high-level signaling is central to neural induction. Ultimately, high Bmp signaling confers epidermal fate. Epidermal ectoderm-derived Bmps are likely to mediate multiple processes at the rostral and lateral margins of forebrain neural plate, although these functions remain undefined.

Role in roof plate induction In the chick and mouse spinal cord, Bmps from epidermal ectoderm induce the formation of the roof plate (RP), the dorsal midline organizer of the neural tube (Figure 3(a)). (Note that the lateral margins of neural plate become the dorsal midline of neural tube after neurulation.) Because the RP continues to produce Bmps, epidermal ectoderm-dependent RP formation serves as a classic example of homeogenetic induction, and it appears likely that homeogenetic RP induction occurs at all levels of the developing neural tube.

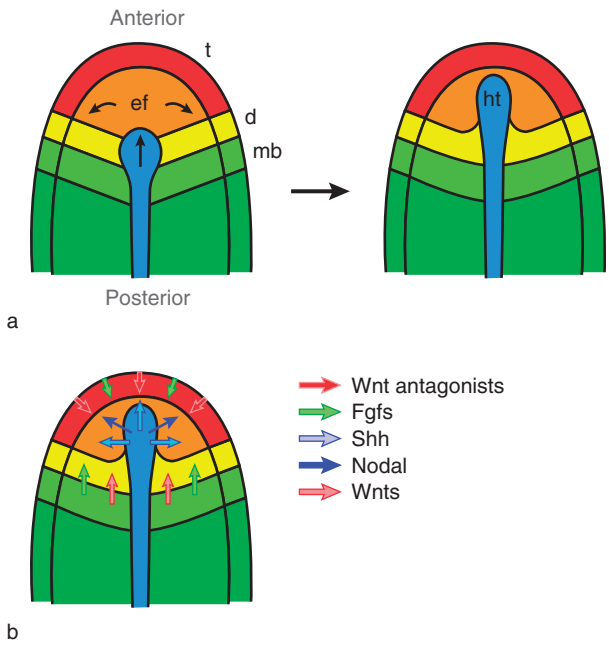

Figure 3 Local neural organizers of the neural plate: (a) rostral morphogenesis of the hypothalamic anlage; (b) signals produced by local neural organizers. In (a), cells of the hypothalamus (ht) originate at or near the node and then move rostrally under the influence of the prechordal plate. This morphogenesis results in the lateral displacement of cells in more rostral fields, particularly the eye field (ef). In (b), Wnt antagonists and Fgfs (most notably Fgf8) are produced at the rostral margin of the neural plate (the anterior neural border). The hypothalamic anlage produces Nodal and Shh, whereas Fgfs and Wnts are produced caudally. d, diencephalon; Fgf, fibroblast growth factor; mb, midbrain; t, telencephalon; Adapted from Wilson SW and Houart C (2004) Early steps in the development of the forebrain. *Developmental Cell* 6: 168, figure 1, with permission from Elsevier.

Prechordal Plate: Nodal, Shh, Bmp, Ventral Midline Induction, and Role in Holoprosencephaly

Prechordal plate formation During gastrulation, mesendodermal cells from the node migrate rostrally along the midline between the ectoderm and endoderm layers. These cells form the tight notochord, which underlies the neural tube up to the midbrain–forebrain boundary and the prechordal plate, a loose cell collection that underlies the emerging forebrain (Figure 2(b)). The prechordal plate shares certain similarities with the notochord, in particular the production of the Sonic hedgehog (Shh) morphogen. Unlike the notochord, however, the prechordal plate also produces several other signaling proteins, including Nodal, Bmp7, and Wnt antagonists.

Regulation of hypothalamic morphogenesis and induction of ventral midline organizers The prechordal plate regulates three different aspects of hypothalamic (HT) development: (1) the rostral migration of future HT cells, (2) HT induction, and (3) HT patterning. The hypothalamus then serves as a local neural organizer caudomedial to the eye field. Prechordal plate signals also induce local organizer function in the ventral telencephalic midline (VTM; preoptic and entopeduncular regions, and then later in the medial ganglionic eminence). Nodal and Shh signals account for these prechordal plate-dependent functions. The prechordal plate has also been implicated in ANB development because Shh is needed to maintain Fgf8 expression in the ANB.

Role in classic holoprosencephaly Among nonneural organizers, only the prechordal plate has been directly implicated in HPE induction, and prechordal plate defects generate classic HPE phenotypes. Manual prechordal plate removal results in a single eye field, failed HT induction, and molecular defects characteristic of classic HPE. In addition, genetic analyses of the Nodal, Shh, and Bmp signaling pathways point to the primary involvement of the prechordal plate in classic HPE induction.

Local Neural Organizers at the Forebrain Midline

Under direction of nonneural organizers, local organizers are established within forebrain tissue itself (Figure 4). This occurs during the period spanning the end of gastrulation and the beginning of neurulation (neural tube formation). Local midline organizers include the hypothalamus and ventral telencephalic midline (VTM) ventrally, the ANB rostrally, and the RP dorsally. Induction of ventral and rostral organizers occurs before that of the RP, which is completed after neurulation. Among these, the RP has been directly implicated in MIH HPE, whereas molecular genetic and transplant studies also suggest direct ANB involvement in classical HPE.

Hypothalamus and Ventral Telencephalic Midline: Nodal, Shh, and Ventral Patterning

Hypothalamic morphogenesis and induction In fish and chicks, HT cells originate near the node, where they are intermingled with floor plate cells. HT cells then migrate rostrally toward the telencephalon, which results in a lateral displacement of forebrain tissue, most notably the eye field (Figure 3(a)). This morphogenesis lags slightly behind that of the prechordal plate (Figure 2(b)), and the prechordal plate regulates HT morphogenesis via Nodal signals. Although less well appreciated than the signaling functions, defects in rostral HT morphogenesis are easy to envision as the culprits of cyclopia and HPE. Prechordal plate signaling via Nodal and Shh signals is then responsible for inducing HT and VTM fates and signaling functions.

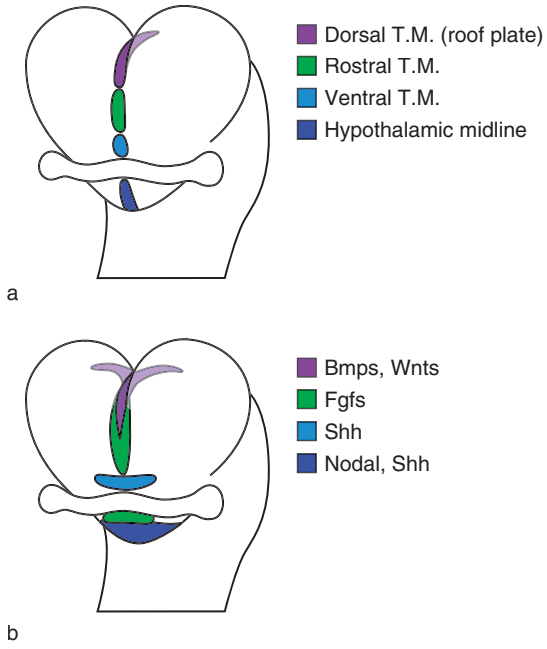

Figure 4 Local neural organizers of the rostral neural tube: (a) midline organizers; (b) morphogen expression. In (a), specialized signaling centers are aligned at the midline following neurulation, including at the closed rostral end of the neural tube. In (b), the midline organizers serve as central sites of morphogen production, which then extends beyond the midline to involve additional sites. Dorsally, Bmps and Wnts are produced at the midline and then secondarily at the dorsomedial edges of telencephalon (choroid plexus and cortical hem). Rostral Fgf8 expression extends to involve a domain caudal to the eye field. Conversely, Shh production extends rostrally to involve the ventral telencephalic midline rostral to the eye field. TM, telencephalic midline.

Nodal and Shh expression Following induction, the HT continues to produce Nodal and Shh (Figure 3(b)). Nodal and Shh signals from the prechordal plate and/or produced locally then regulate HT patterning. The VTM is induced to express Shh, but not Nodal, and Shh expression from the VTM eventually extends into the medial ganglionic eminence (MGE) (Figure 4). These ventral Shh sources are probably responsible for inducing and patterning the septum, MGE, and lateral ganglionic eminence (LGE) in the ventral telencephalon.

Roles in patterning and holoprosencephaly pathogenesis Distinguishing the roles of ventral neural organizers from those of the prechordal plate has been difficult due to their common expression of morphogens. Nonetheless, both the hypothalamus and VTM seem certain to be involved in HPE pathogenesis, if not its induction. Based on proximity and timing, the VTM may be directly responsible for the basal ganglia defects (e.g., basal ganglionic fusion) that are frequently seen in classic HPE.

Anterior Neural Border: Wnt Antagonists, Fibroblast Growth Factors, and Telencephalic Specification and Patterning

Anterior neural border induction In addition to the AVE, the prechordal plate and epidermal ectoderm are well positioned to participate in ANB specification. Shh signaling, most likely from prechordal plate, is needed to maintain Fgf8 expression in the ANB, although initial Fgf8 induction can occur in the absence of Shh.

Expression of Wnt antagonists and Fibroblast growth factors Following induction, the ANB expresses two different classes of secreted products: (1) Tlc, a Wnt antagonist of the sFRP family and (2) a number of Fgfs, which vary among species (e.g., Fgf3/8 in fish and Fgf8/15/17/18 in mice) (Figure 3(b)). Fgf8 is particularly important for ANB function. Fgf8 expression does not remain restricted to the ANB, however; it extends caudally into the optic region and then splits. Low-level Fgf8 expression is also detectable in dorsal telencephalic and diencephalic midline regions after neurulation, which may represent continued Fgf8 expression that originated at the lateral margins of neural plate.

Roles in telencephalic specification and patterning Ablation and transplant studies suggest a central role for the ANB in telencephalic specification and patterning. These two functions are mediated by the two different classes of ANB products. Tlc mediates ANB-dependent telencephalic specification by antagonizing caudal Wnt sources. The Fgfs (particularly Fgf8) mediate ANB-dependent telencephalic patterning rather than its initial specification, promoting rostral telencephalic fates over caudal ones.

Potential roles in midline patterning and holoprosencephaly induction ANB tissue rescues midline eye-field fates in Nodal mutants, implicating the ANB in the eye defect of classic HPE. Although Fgf8 expression extends beyond the ANB, genetic analyses suggest a prominent role for the ANB in Fgf8 mutant HPE phenotypes.

Roof Plate: Bmps, Wnts, Dorsal Midline Induction, and Role in Middle Interhemispheric Holoprosencephaly

Roof-plate induction Based on analogy to the spinal cord, nonneural (epidermal) ectoderm probably has a central role in inducing forebrain RP, although this remains unproven.

Expression of Bmps and Wnts Following its formation and induction during neurulation, the RP expresses several Bmps (2/4/5/6/7/12), as well as multiple Wnts.

Bmp4, which is the best-studied Bmp in forebrain development, is expressed in the RP and abuts the Fgf8 domain in the rostral midline (Figure 4(b)).

Roles in dorsal midline induction In addition to minor lineage contributions, the RP governs induction of dorsal fates near the telencephalic midline, most notably the telencephalic choroid plexus epithelium (CPe) and cortical hem (the small CPe–cortex junctional region). The RP also patterns the adjacent dorsal cortical primordium. Based on rescue and gain of function studies, many of these RP-dependent functions are mediated by Bmps. These Bmp-dependent RP functions at the dorsal telencephalic midline parallel those in the dorsal spinal cord and cerebellum.

Role in middle interhemispheric holoprosencephaly In addition to failed dorsal midline induction, genetic RP ablation in mice results in MIH HPE. The ablation phenotype also accurately predicts CPe and hippocampal defects in MIH HPE patients, thus supporting a causal role for the RP in human MIH HPE.

Molecules and Pathways in Forebrain Midline Induction and Holoprosencephaly

With a cellular framework now established, we turn our attention to the individual genes and signaling proteins involved. Although selective molecular expression allows for functional assignment to specific organizers in some cases, individual signals and genes are most often expressed beyond the confines of single organizers (e.g., after homeogenetic induction; Figure 4). However, whether they inform organizer function or not, molecular analyses have illuminated a great deal about midline patterning and HPE. We start by grouping causal HPE genes into defined signaling pathways, then discuss each pathway in turn.

Genes Implicated in Holoprosencephaly

Several HPE genes have been identified from studies in fish, mice, and humans. In humans, nine genes have been identified, with a similar number of genes awaiting identification. Most of the identified genes have known functions in specific morphogen pathways. Some encode transcription factors (e.g., Six3, Zic2, and Tgif) with less certain relationships to individual pathways or organizers, although recent studies have begun to clarify their roles as well (Table 1).

Table 1 Holoprosencephaly genes and signaling pathways[a]

NODAL	*SHH*	*FGF8*	*BMP*
Nodal	SHH	Fgf8	Chordin
Cyclops	PTCH	SIX3	Noggin
Squint	DHCR7		Twsg1
One-eyed pinhead	GLI2		ZIC2
TDGF1/CRIPTO	Smo		
FAST1	Cdo		
	DispA		
	SIX3		

[a]Genes from humans and experimental organisms linked to definitive holoprosencephaly (HPE) phenotypes, grouped by signaling pathway. Human genes are shown in upper case. TGIF may be involved in both Nodal, Bmp, or retinoic acid signaling in the forebrain, although this remains undefined. Megalin may be involved in either Shh or Bmp signaling. SIX3 and ZIC2 may be involved in additional signaling pathways.

Signaling Pathways in Forebrain Midline Induction and Holoprosencephaly

Nodal signaling at the ventral midline

Genetic associations with holoprosencephaly. Defects in several Nodal signaling components lead to HPE phenotypes in fish, mice, and humans. These include Nodal ligands, a receptor subunit (ActRIIA), extracellular cofactors (EGF-CFC family), a transcriptional transducer (Smad2), and a transcriptional cofactor (Tgif). Genetic loss of function yields predictable phenotypes based on the known positive or negative regulation of Nodal signaling by these genes, with the exception of Tgif. (HPE is associated with reduced Nodal signaling and Tgif loss of function, but, as a negative regulator, Tgif loss of function should lead to increased Nodal signaling.)

Nodal sites and functions: Confined expression to ventral tissues. Among its many developmental functions, Nodal signaling is required for the initial specification of the prechordal plate, where Nodal ligands are expressed. Prechordal plate-derived Nodal signals then regulate HT morphogenesis and induction (Figure 2(b)). Nodal signaling is the only pathway known to regulate rostral HT morphogenesis. Later, Nodal is expressed in the hypothalamus itself (Figure 3(b)), but does not extend beyond the ventral midline during this developmental period.

Nodal mutant holoprosencephaly phenotypes: Severe classic holoprosencephaly extending to the dorsal midline. Despite the restriction of Nodal expression to ventral tissues, Nodal mutants have a particularly severe HPE phenotype that extends to involve the rostral and dorsal midlines. This severe classic HPE phenotype is reminiscent of those associated with defects in Shh signaling.

Shh signaling at the ventral midline

Genetic associations with holoprosencephaly. Like Nodal, studies from different organisms implicate several Shh signaling components in ventral midline induction of the forebrain. These include the ligand, its cholesterol-based posttranslational modifications,

Shh release (DispA), receptors (Ptch, Smo), and the coreceptor or trafficking molecule (megalin). By and large, genetic analyses of individual components have led to midline and HPE phenotypes consistent with their known positive or negative effects on Shh signaling.

Shh sites and functions: Confined expression to ventral tissues. Shh is co-expressed in prechordal plate with Nodals (Figure 2(b)). Prechordal plate Shh induces the overlying hypothalamus and VTM (preoptic and anterior entopeduncular regions), which leads to activated Shh expression in these tissues (Figures 3(b) and 4(b)). Although the relative contributions of Shh from prechordal plate, the hypothalamus, and the VTM remain difficult to dissociate, ventral Shh signaling is clearly essential for midline eye-field specification and for excluding laterally expressed genes from the midline. The midline eye-field defects in Shh mutants do not result from aberrant HT morphogenesis, unlike Nodal mutants, and are therefore more clearly due to failed inductive signaling. At later stages, Shh production extends into more lateral regions of the ventral telencephalon (medial ganglionic eminence) (Figure 4(b)). Like Nodal, all Shh sources remain confined to ventral tissues during the HPE inductive period.

Shh mutant holoprosencephaly phenotypes: Severe classic holoprosencephaly extending to the dorsal midline. Like Nodal, Shh signaling defects in mice and humans lead to ventral predominant classical HPE, which typically extend to involve the rostral and dorsal midlines. For example, human HPE phenotypes due to Shh mutations include absence of the interhemispheric fissure, absence of corpus callosum, and failed cortical separation in rostral and dorsal telencephalic domains. At the dorsal midline, Shh null mice exhibit no telencephalic CPe and aberrant expression of lateral cortical markers across the midline. Notably, the absence of CPe and abnormal midline expression of lateral cortical markers mimics the dorsal midline phenotype induced by RP ablation.

Fgf8 signaling at the rostral midline

Genetic association with holoprosencephaly. Although not yet implicated in humans, Fgf signaling from the ANB has been implicated in midline patterning defects in fish (Fgf8 and Fgf3) and mice (Fgf8). Significantly, Fgf8 mutations have recently been shown to result in definitive HPE phenotypes in mice. Other dedicated Fgf signaling components have yet to be implicated in HPE, although Six3 has a critical role in Fgf8 induction.

Fgf8 sites and functions: Anterior neural border with ventral and dorsal extensions. Fgfs are not expressed in nonneural tissues near the forebrain but are, instead, limited to local neural organizers. In the forebrain neural plate, the expression of Fgf8 and other Fgfs (which differ among species) initiates in the ANB at the margin of the telencephalic field (Figure 3(b)). Fgf8 expression eventually extends caudally into the midline eye field and then splits bilaterally (Figure 4 (b)). After neurulation, high-level Fgf8 expression continues at the rostral telencephalic midline, although lower levels also become detectable in the dorsal telencephalic and diencephalic midline regions. Caudally, Fgfs are also expressed in the isthmus organizer, the local neural organizer at the midbrain–hindbrain boundary.

Fgf8 studies and mutant. holoprosencephaly phenotypes: Variable, extending to the dorsal midline Fgf8 mutant phenotypes are variable but include those corresponding to classic HPE. Defects in HT and VTM induction correlate with Fgf8 activity level, based on analysis of hypomorphic Fgf8 mouse mutants. Moreover, ectopic Fgf8 studies in chicks indicate the positive regulation of midline fate (including induction of an ectopic sulcus reminiscent of the interhemispheric fissure). Notably, midline induction failure in Fgf8-dependent HPE can also extend to the dorsal midline. Unlike its consistently positive regulation of ventral midline fates, however, Fgf8 effects on dorsal midline development are nonlinear. Recent studies in the chick midbrain provide a plausible explanation – namely that Fgf8 both promotes RP competence (positive regulation) and inhibits RP differentiation (negative regulation).

Bmp signaling at the ventral and dorsal midlines

Genetic association with holoprosencephaly. Genetic studies on the Bmp antagonists Noggin and Chordin in mice have directly implicated Bmp signaling in HPE. Other dedicated Bmp signaling components have yet to be associated, although a major site of Zic2 action appears to be upstream of the Bmp-producing RP. Tgif (a Smad corepressor) may also act in the Bmp pathway, and Bmp signaling has been implicated in many of the midline defects associated with RP ablation in mice.

Bmp sites and functions: Ventral and dorsal midline tissues. Prior to gastrulation, multiple Bmps are expressed in ectoderm, where the antagonism of Bmp signaling induces neural fate. Following neural induction, Bmps continue to be expressed at high levels in nonneural (epidermal) ectoderm contiguous with the neural plate. Noggin and Chordin are expressed at highest levels in the prechordal plate and notochord from early stages, whereas Bmp7 expression in the prechordal plate begins later (Figure 2(b)). Following neurulation, the RP becomes an epicenter for Bmp expression (Figure 4), although many Bmps maintain expression domains in the overlying ectoderm,

mesenchyme, and elsewhere. Bmp expression is maintained at later stages in two dorsal telencephalic midline tissues derived or induced by the RP: CPe and cortical hem, which is the CPe–cerebral cortex junctional region that serves as a Wnt production epicenter.

Bmp mutant holoprosencephaly phenotypes: Classic and middle interhemispheric forms. The Noggin/Chordin mutant HPE phenotype falls into the ventral-predominant classic HPE category. This correlates with their major site of expression in the forebrain region, the prechordal plate. Bmp signaling has been shown to suppress Shh expression, which probably accounts for the need to antagonize Bmp signaling in the prechordal plate (Figure 2(b)). Bmp-antagonist studies provide additional evidence for the primary role of prechordal plate in classic HPE induction. On the other hand, RP ablation induces MIH HPE, including the absence of CPe and aberrant midline expression of cortical markers (Figure 1(b)). Bmp signaling has been directly implicated in CPe induction and suppression of cortical markers at the midline, and multiple midline defects caused by RP ablation can be rescued by exogenous Bmp4. Thus, Bmp signaling appears to be centrally involved in RP-dependent midline and HPE induction.

Interactions among the Signaling Pathways

Several insights into midline patterning and HPE have come from recent *in vivo* and *in vitro* studies that have determined how the various signaling pathways interact. In addition to these interactions, we discuss how the transcription factors implicated in HPE impact these signaling pathways (Figure 5).

Nodal → Shh Pathway in the Ventral Midline

As previously discussed, Nodal and Shh share not only common expression sites but also similarity in mutant phenotypes. Genetic epistasis experiments provide significant evidence for Nodal being upstream of Shh in the pathway leading to HT induction. For example, Nodal null mutants lack prechordal plate and Shh expression, and exogenous Shh rescues HT and ventral telencephalic Nodal defects. Shh appears sufficient to mediate many, but not all, Nodal-dependent functions (e.g., diencephalic defects in Nodal mutants are not rescued by Shh, suggesting a separate Shh-independent Nodal pathway). Subsequent HT patterning involves parallel pathways for Nodal and Shh (as well as Bmp7), with the two morphogens biasing toward opposing HT fates.

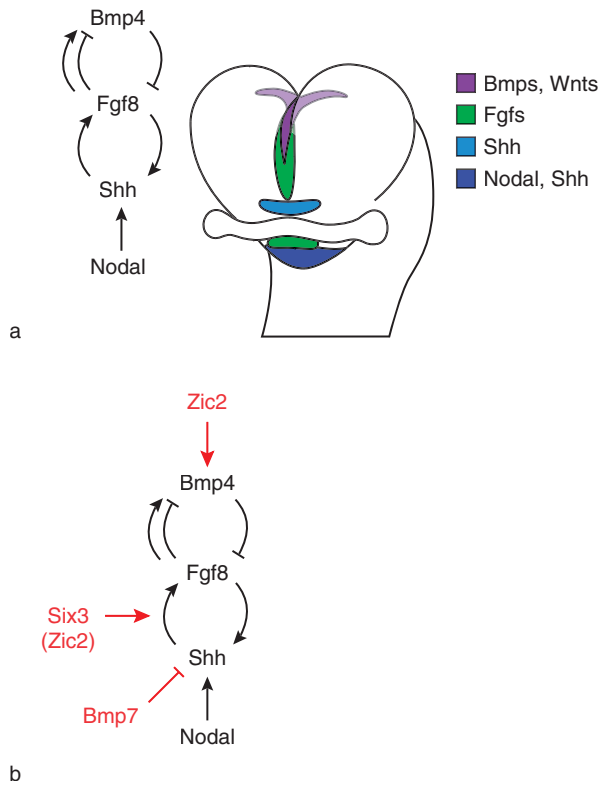

Figure 5 Genetic network in forebrain midline induction: (a) the core morphogen interaction network; (b) additional network components. In (a), The genetic network starts with ventral Nodal, which positively regulates ventral Shh production. Shh then engages in a positive feedback circuit with Fgf8 in the ventral and rostral domains. Fgf8 has both positive and negative influences on the dorsal midline, serving initially as a competence factor for RP induction. Later, Fgf8 acts as an inhibitor of RP maturation and participates in a negative feedback loop with dorsal Bmp4 and other Bmp signals. This core network explains how defects in morphogens expressed exclusively in ventral domains (Nodal and Shh) can spread to involve the rostral and dorsal midlines, which is characteristic of classical HPE. Conversely, this network also explains why primary RP defects that cause MIH HPE do not spread rostrally or ventrally. In (b), Bmp7 produced by the prechordal plate antagonizes Shh production, which accounts for the classic HPE phenotype due to Noggin and Chordin loss. Six3 mutations also cause classic HPE, which is consistent with Six3 involvement in multiple circuits that all serve to promote the Shh → Fgf8 positive feedback loop. Although the mechanism by which Zic2 mutations lead to classic HPE remains uncertain, Zic2 mutation can also result in MIH HPE due to its role in RP induction. HPE, holoprosencephaly; MIH, middle interhemispheric; RP, roof plate.

Shh → Fgf8 Positive Feedback Loop in the Ventral and Rostral Midlines

Evidence for a positive feedback loop Fgf8 expression is almost entirely lost in Shh mutant mice, but Fgf8 expression is dramatically upregulated in mouse mutants with increased Shh signaling (Gli3 mutants), indicating that Shh signaling positively regulates Fgf8 expression. The critical Shh sources at the ventral

midline (whether the prechordal plate, hypothalamus, or VTM) remain to be determined. Interestingly, Shh is not required for Fgf8 induction (based on rescued Fgf8 expression in Shh/Gli3 double null mutants) but, rather, for its maintenance once induced. Conversely, Fgf8 loss in fish and mice leads to decreased Shh expression in the hypothalamus and VTM. Consistent with a positive feedback loop, Fgf8 and Shh expression are consistently co-regulated whenever Bmp signaling is altered, either genetically or with exogenous proteins in chicks and mice. The positive feedback loop between Shh and Fgf8 is similar to that described in limb development.

Convergence of Shh and Fgf8 signaling in the eye field In addition to the positive feedback loop, Shh and Fgf8 signaling cooperatively induce optic stalk fate in the eye field. The eye field is well placed for such convergence, where nearby Shh and Fgf8 expression domains are closely apposed (Figure 4(b)). In cyclopic Shh mutant eyes, exogenous Fgfs or ANB transplants are sufficient to rescue optic stalk fate. In addition, the absence of ventral telencephalic fates in Shh −/− mice is suppressed by Gli3 inactivation, implying the presence of another ventralizing signal. This most likely includes Fgf8, which is dramatically upregulated in Shh/Gli3 compound null mice.

Positive Regulation of the Shh → Fgf8 Loop by Six3

Although Six3 does not fit neatly into a single signaling pathway, it is embedded in two defined positive feedback loops that serve to promote the Shh → Fgf8 loop. Six3 is expressed in the anterior neural plate and ANB, where it is required for forebrain and eye specification. These Six3 functions involve a feedback loop between Six3 and Wnt repression. Six3 is also required for ANB and Fgf8 induction. Following Fgf8 induction, Six3 serves as a competence factor for Fgf8-dependent telencephalic gene expression, resulting in a Six3 → Fgf8 positive feedback loop in the telencephalon. Six3 is also strongly expressed in the eye field and hypothalamus, where it confers competence for Shh-dependent HT fate. Thus, all defined Six3-dependent loops and pathways act to promote the Shh → Fgf8 positive feedback loop.

Negative Regulation of the Shh → Fgf8 Loop by Bmps

As already discussed, several experimental manipulations that alter Bmp signaling (e.g., Bmp and Noggin bead placements in chicks; RP ablations, Chordin/Noggin double knockouts, Noggin electroporations, and exogenous Bmp applications in mice) lead to the same conclusion – namely that Bmp signaling negatively regulates both Fgf8 and Shh expression. As discussed earlier, Bmps in the prechordal plate (Figure 2(b)) probably account for the HPE phenotype seen in Chordin/Noggin compound null mice. Bmps produced in the dorsal midline region may also negatively regulate Fgf8 and Shh because Fgf8 and Shh expression are well maintained (and Fgf8 possibly upregulated) following genetic RP ablation in mice.

Fgf8 → Roof Plate Pathway: Positive and Negative Regulation by Fgf8

Studies in fish and mice indicate that reduced Fgf8 expression can lead to dorsal midline defects and HPE. This indicates that Fgf8 can positively regulate dorsal midline induction. However, until recently, most gain-of-function studies have provided evidence for negative Fgf8 regulation of dorsal midline development. Significantly, analyses of hypomorphic Fgf8 mouse mutants have provided evidence for nonlinear Fgf8 effects (i.e., reduced Fgf8 signaling can have both positive and negative effects on the dorsal midline), and recent studies in chick midbrain provide an explanation for this nonlinearity. These studies identified two different roles for Fgf8 in RP development. Fgf8 first acts as a competence factor for RP induction (e.g., Fgf8 beads can induce an ectopic RP *in vivo*). At later stages, Fgf8 inhibits RP maturation, probably by antagonizing transforming growth factor (TGF)β signals, including Bmps, which positively regulate RP differentiation. Earlier studies in the chick forebrain suggest similar Fgf8 functions in the forebrain, although these remain unproven. The ability of Fgf8 to confer RP competence provides, for the first time, a mechanism accounting for failed dorsal midline development following Fgf8 reduction.

Zic2 → Roof Plate Pathway in the Dorsal Midline

Studies in mice demonstrate Zic2 expression prior to neurulation, followed by preferential Zic2 expression in the dorsal forebrain. Zic2 knockdown in mice leads to defective neurulation kinetics and absence of RP induction, which leads to a dorsal HPE phenotype. Notably, among known HPE genes, only *ZIC2* has been specifically linked to human MIH HPE (single case predicting a hypomorphic 12-amino-acid in-frame deletion), although most human *ZIC2* mutations lead to classic HPE. The classic phenotype could be related to Zic2 expression in ventral midline tissues. Nonetheless, the mouse knockdown studies and hypomorphic human allele suggest that dorsal midline development – and RP induction in particular – are particularly sensitive to reduced Zic2 level. Zic2 expression is well maintained following RP ablation in

mice and is insensitive to exogenous Bmps in explants and dissociated cells, which argues against positive feedback from the RP back to Zic2.

Sufficiency of Morphogen Interactions to Account for Classic and Middle Interhemispheric Holoprosencephaly Phenotypes

As mentioned earlier, how pure ventral signaling defects caused by Shh and Nodal mutations can spread to involve the dorsal midline has been a conundrum that lacked explanation. Studies of the developing spinal cord, which so often serve as the precedent for the developing forebrain, do not help in this regard. For example, Shh loss in the spinal cord leads to ventral expansion of Bmp-regulated genes (as occurs in the forebrain), but it does not lead to RP failure or other dorsal midline defects.

Recent studies, particularly those involving Fgf8 and the RP, provide for a coherent genetic regulatory network that resolves this apparent conundrum and is sufficient to account for both classic and MIH HPE phenotypes (Figure 5). In this network, Fgf8 serves as the key intermediary between ventral Nodal/Shh and dorsal Bmp signaling (via the ANB and rostral midline, which distinguish the anterior neural tube from more caudal regions). The role of Fgf8 as an intermediary results from both its involvement in positive feedback with Shh and its positive regulation of dorsal midline development by providing competence for RP (and dorsal Bmp) induction. Via Fgf8, defects in Nodal and Shh signaling then lead plausibly to failed dorsal midline development, as seen commonly in classic HPE.

Conversely, the RP ablation and Bmp genetic studies help to explain why MIH HPE defects do not spread ventrally. Based on a relatively large number of studies, Bmp signaling acts uniformly to negatively regulate both Fgf8 and Shh expression (Figure 5). Reduction of dorsal Bmp signaling (e.g., due to RP failure), therefore, serves to maintain or upregulate the pathways responsible for patterning the rostral and ventral midlines, thus preserving these domains in MIH HPE. The unidirectionality of spread in HPE (i.e., ventral to dorsal in classic HPE but not vice versa in MIH HPE) is also consistent with the temporal order of differentation, with the induction of ventral and rostral midline signaling centers occurring prior to neurulation and RP induction.

Interestingly, in virtually all cases examined, positive and negative regulation of the involved signaling pathways occurs at the level of individual morphogen mRNAs. Conceivably, regulation of signaling pathways could occur at myriad levels (e.g., extracellular antagonists, extracellular diffusivity, cell surface receptors, intracellular transducers, or degradation), and regulation at multiple levels is likely at the forebrain midline, as it is elsewhere. Thus, it is perhaps remarkable that steady-state mRNA levels of the morphogens alone are sufficient to account for the spectrum of midline patterning defects seen in HPE. Although unproven, we might then predict that regulation at other levels of signaling should occur predominantly in the same positive or negative direction as those for the individual morphogen mRNAs.

Future Directions

Several unanswered questions remain about the pathways involved in forebrain midline induction and HPE. The extent to which the simple morphogen interaction network models the HPE phenotypes requires further testing. In addition, notably absent from most of this article is a detailed discussion of Wnt signaling. Wnts have well-defined roles in AP patterning, including telencephalic specification, but their precise roles in midline patterning remain undefined. Because Six3 and Wnt suppression make up a positive feedback loop central to telencephalic induction, Wnt suppression is at least indirectly involved in promoting competence (via Six3) for Shh- and Fgf8-induced ventral midline fates. Rostral Fgf8 signaling has also been implicated as a negative regulator of dorsal Wnt signaling, although the role of this pathway in HPE pathogenesis is unclear. Another conceivable participant is retinoic acid (RA). RA has been implicated as one of the caudalizing signals whose antagonism is required for forebrain fate specification. Excessive RA signaling can lead to an HPE-like phenotype in mice, which is consistent with its ability to downregulate Shh expression.

The role of Tgif also remains somewhat mysterious. A significant number of human *TGIF* mutations (mostly predicting loss of function) have now been associated with HPE. As a repressor of Nodal and RA signaling, Tgif does not yet fit well into the Nodal–HPE pathway (Tgif loss-of-function mutations should lead to increased rather than decreased Nodal signaling), but it fits with RA (the same mutations mimic excessive RA). However, total Tgif loss in mice does not lead to midline induction failure or HPE. In addition, Tgif expression after neurulation appears preferentially in dorsal domains and is higher in the cortex than the dorsal midline. Maintained Tgif expression is dependent on an intact RP and is positively regulated by exogenous Bmps in culture, which is consistent with a role for Tgif in HPE pathogenesis but not in induction *per se*. Additional basic studies on Tgif function in early forebrain development are needed to clarify its placement in the genetic regulatory network that underlies forebrain midline and HPE induction.

See also: Bone Morphogenetic Protein (BMP) Signaling in the Neuroectoderm; Forebrain Development: Prosomere Model; Forebrain: Early Development; Holoprosencephaly; Midbrain Patterning; Sonic Hedgehog and Neural Patterning.

Further Reading

Alexandre P, Bachy I, Marcou M, and Wassef M (2006) Positive and negative regulations by FGF8 contribute to midbrain roof plate developmental plasticity. *Development* 133: 2905–2913.

Bachiller D, Klingensmith J, Kemp C, et al. (2000) The organizer factors Chordin and Noggin are required for mouse forebrain development. *Nature* 403: 658–661.

Cheng X, Hsu CM, Currle DS, et al. (2006) Central roles of the roof plate in telencephalic development and holoprosencephaly. *Journal of Neuroscience* 26: 7640–7649.

Golden JA (1998) Holoprosencephaly: A defect in brain patterning. *Journal of Neuropathology and Experimental Neurology* 57: 991–999.

Hayhurst M and McConnell SK (2003) Mouse models of holoprosencephaly. *Current Opinion in Neurology* 16: 135–141.

Marcorelles P, Loget P, Fallet-Bianco C, et al. (2002) Unusual variant of holoprosencephaly in monosomy 13q. *Pediatric and Developmental Pathology* 5: 170–178.

Monuki ES and Walsh CA (2001) Mechanisms of cerebral cortical patterning in mice and humans. *Nature Neuroscience* 4(supplement): 1199–1206.

Nagai T, Aruga J, Minowa O, et al. (2000) Zic2 regulates the kinetics of neurulation. *Proceedings of the National Academy of Sciences of the United States of America* 97: 1618–1623.

Simon EM, Hevner RF, Pinter JD, et al. (2002) The middle interhemispheric variant of holoprosencephaly. *American Journal of Neuroradiology* 23: 151–156.

Storm EE, Garel S, Borello U, et al. (2006) Dose-dependent functions of Fgf8 in regulating telencephalic patterning centers. *Development* 133: 1831–1844.

Wallis D and Muenke M (2000) Mutations in holoprosencephaly. *Human Mutation* 16: 99–108.

Wilson SW and Houart C (2004) Early steps in the development of the forebrain. *Developmental Cell* 6: 167–181.

Forebrain Development: Prosomere Model

L Puelles, University of Murcia, Murcia, Spain

© 2009 Elsevier Ltd. All rights reserved.

Definition

The prosomeric model is a segmental structural model of the brain of vertebrates that explicitly holds that the brain is formed by an uninterrupted series of transverse subunits of the neural tube, generally called neuromeres. Among such subunits is a large rostral forebrain unit – the secondary prosencephalon – that encompasses hypothalamus, eyes, and telencephalon, followed by three caudal forebrain or diencephalic neuromeres (i.e., prosomeres), which are regarded as being serially homologous with more-caudal neuromeres, namely, a single midbrain mesomere, 11 hindbrain rhombomeres, and the spinal myelomeres. It is important to note that for descriptive purposes, the model postulates in all vertebrates a morphogenetic bending of the longitudinal axis of the tubular neural primordium, most marked at the cephalic flexure, whose incurvation causes wedge-shaped deformation of the topologically transverse cylindrical neuromeric sectors of the neural tube. These units share a set of fundamental longitudinal zones (due to common dorsoventral (DV) patterning processes) and therefore represent segments, that is, metameric developmental units (anteroposterior (AP) patterning). The common causal background of the longitudinal zones establishes the property of metamery (i.e., serial homology) across all neuromeres, irrespective of their differential molecular identities and individual prospective adult fates and of the variable border properties of the cells found at the interneuromeric boundaries. Therefore, the prosomeric model visualizes all vertebrate brains as segmented structures constructed along the same Bauplan (same set of DV and AP developmental units). Orthogonal intersection of DV and AP boundaries in the neural tube wall defines a checkerboard pattern of domains (histogenetic areas) in which specific properties and finer regionalization phenomena appear (shared or not among vertebrates). This makes the model useful for systematic descriptive neuroembryology, comparative neuroanatomy, and causal analysis of conserved or variant brain morphogenesis.

Characteristics

Why We Use Models

Structural neurobiology studies the form and functional inner structure of brains. This needs a conceptual model in which all sorts of detailed data on form and structure down to cellular aspects can be systematically accumulated, organized, compared, and differentiated one from another in their mutual relationships. The basic model of the vertebrate brain is the concept of the closed neural tube, from which adult brains emerge via differential morpho- and histogenesis. Since we cannot know all from the beginning, morphological models essentially are reasonable and useful conjectures about how many parts there are and how they are patched together. Such models are periodically perfected over time, becoming in the case of brains increasingly complex operational scaffolds based on accumulated data and a number of assumptions. There is always the possibility of constructing better (or worse) models.

A good model in essence should be parsimonious; that is, it should identify a minimal set of characteristic parts or landmarks in the modeled system, which can be generally recognized and seem to encompass, or be able to explain, most if not all available structural data. A good model also may delimit various 'unfilled' conceptual domains, where new data should fit in (as the periodic table of chemical elements did when it was first formulated). Such predictive aspects of models are highly useful because they implicitly indicate which new questions might be meaningful or how best to pose and answer them in practice. Simultaneously, models are instrumental in providing possible significance to any new, unexpected observation. Scientists in principle believe in and use a particular model as long as it seems to accommodate established knowledge, inspire significant research, and allow satisfactory incorporation of emerging sets of new data. Historical periods in which technological improvements produce radically novel sorts of data are particularly critical for the survival of a model.

Models widely shared among a scientific community represent a scientific paradigm. In contrast to hypotheses and theories, paradigms are not meant to be tested, since one must believe in one of them and use it as if it represented the truth, in the very process of testing a hypothesis experimentally. A paradigm comes dangerously close to becoming a dogma, a belief that wholly escapes criticism or doubt and is considered ascientific. Several models may coexist historically, sometimes because each one is perceived to have different advantages, but usually due to lack of awareness that one of them is distinctly better than the others, compounded with the human tendency to persist irrationally in long-held beliefs. Nevertheless, models and paradigms eventually may be perceived as

obsolete and be discarded by newer, less committed generations of scientists, particularly when they are manifestly unable to account for some data and efforts to apply them lead to highly unparsimonious, complicated lines of thought. Continued use of an obsolete model, or mixed-up joint use of elements of different models, tends to obstruct the progress of science.

Some neuroscientists wrongly think that they do not use a neural model. This means they simply are unaware of the model they are using. Frequently, some aspects they mistakenly regard as facts actually are conjectures. Dogmatic conscious or unconscious belief in models is a condition that is prone to poor thinking and poor science. Due to the great complexity of the studied organ, brain science is a field where such interpretive malfunctioning is not uncommon.

Neuromeric Models

The earliest morphological models of the brain were based on the adult form of the human and animal brains. This approach provided over time a rich set of neuroanatomical terms and conjectural meanings, many of which are now obsolete, although some old terms and concepts still persist in textbooks. During the late nineteenth century, and as a result of various important advances such as microscopy, evolutionary theory, and cell theory, comparative anatomical and embryological knowledge of brains advanced enough to allow the initial formulation of developmental brain models generally valid for all vertebrates. Developmental models also appeared for the entire body. The first generally accepted developmental structural paradigm for brains was a segmental model of the neural tube, which appeared hand in hand with a segmental model of the body and head of vertebrates. The axial skeleton was conceived as being segmented into metameric vertebrae (with a number of units fused together in the sacrum and in the cranial basis). The branchial apparatus also seemed segmented into serial branchial arches and slits. The brain and the set of spinal and cranial nerves were postulated to consist of a number of segmental units correlated one to one with the vertebrae and/or branchial arches.

The term 'neuromere' that was soon applied to these transverse neural units was coined by the American scientist Orr, who very ably characterized histologically in lizard embryos the relevant hindbrain, midbrain, and forebrain neuromeric units. He also provided a clear-headed morphological analysis of longitudinal zonation and axial bending of the brain, largely consistent with the present-day prosomeric model (Figure 1). Orr's study is the historic root of the prosomeric model, though previous and subsequent writings by von Kupffer, Hill, His, Neal, Palmgren, Rendahl, Tello, and Vaage, among others, contain less explicit antecedents. A large-scale review of shared neuromeric structural data collected for all vertebrate lineages from agnatha to mammals was published by von Kupffer at the turn of the twentieth century.

Wilhelm His produced an alternative very influential neural model, though he certainly must have known the neuromeric views of von Kupffer and other contemporaries well. His defined the floor, basal, alar, and roof plates, the alar-basal boundary (sulcus limitans of His), the concept of isthmus, and the idea of neural tube morphogenetic deformation due to axial bending. This model was very influential because it underpinned the first Nomina Anatomica in 1895, whose committee was presided over by His. The model was not explicitly neuromeric, though His' concepts of axial bending and longitudinal zonation and most of his transverse boundaries, including those of the isthmus, clearly were consistent with neuromeric models (Figure 2).

Columnar Models

At the height of the prestige of neuromeric brain models, an important unrelated breakthrough resulted from the analysis of functional components in the cranial and spinal nerves. It was discovered that each nerve component (motor or sensory fibers) either originates from or projects on a distinct columnar domain of the hindbrain or spinal cord. Separate columns could be assigned to visceral and somatic nerve components. Afferent fibers usually bifurcate into ascending and descending branches that distribute widely within the corresponding column. In so doing, they do not respect the neuromeric boundaries. These data were widely perceived as important, and they threw doubt on the neuromeric models, at least for application to advanced embryos and adults, since the basic functional organization of the hindbrain and spinal cord seemed to be columnar and not segmental, irrespective of the separate, more or less periodic nerve roots, and the peripheral dermatomes and myotomes. It was increasingly thought that maybe neuromeres were transient early embryonic phenomena without impact in the mature brain, in which a columnar arrangement of functions emerges. While Europe immersed itself in World Wars I and II, a new school of US neuroanatomists bloomed, led by JB Johnston and CJ Herrick, and its members proceeded to explore these new columnar ideas. Already in 1910, Herrick postulated columnar subdivisions in the diencephalon, which he initially thought might be continuous caudally with brain stem columns and extend rostrally into telencephalic ones. This work

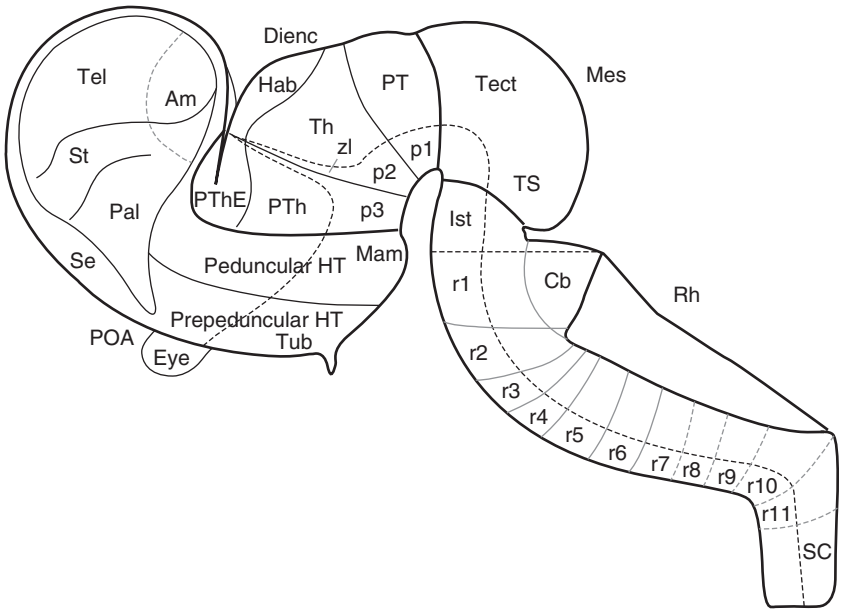

Figure 1 Schema of the prosomeric model of Puelles and Rubenstein (2003). The forebrain lies to the left. Note axial bending at cephalic flexure. The longitudinal alarbasal boundary is present throughout the lateral wall of the neural tube, symbolizing all longitudinal components (floor and roof plates not represented); a singularity known as zona limitans (zl) is a transversal spike-like deviation of the general alar-basal boundary. The secondary prosencephalon (Sec.Pros.) is the rostralmost and most complex unit, consisting of telencephalon (Tel), eye and hypothalamus (HT: divided in two parts). Septum (Se), striatum (St), pallidum (Pal), preoptic area (POA), and amygdala (Am) regions are identified within the telencephalon; the pallium lies under the label Tel. Tuberal (Tub) and mammillary (Mam) subregions of the hypothalamus are marked. The caudal forebrain or diencephalon consists of three prosomeres (p1–p3), whose alar regions include the pretectum (PT), the thalamus and habenula (Th–Hab), and the prethalamus and prethalamic eminence (PTh, PThE); a specific tegmental domain corresponds to each of them (under p1–p3 labels). A simplified view of the large mesencephalic alar plate (Mes) divides it into superior colliculus or tectum (Tect) and inferior colliculus or torus semicircularis (TS); 'colliculi' are mammalian terms. The hindbrain or rhombencephalon (Rh) contains 12 neuromeric units, from the isthmus (Ist) and ehombomere 1 (rl) down to rhombomere 11 (r11), which limits with the spinal cord (SC). Note the cerebellum (Cb) forms mainly across isthmus and r1.

originated the prevalent present-day dogma of the structural division of the diencephalon into epithalamus, dorsal thalamus, ventral thalamus, and hypothalamus, considered to be longitudinal columns of the forebrain. This emphasis was accompanied by negation of the cephalic flexure (or simply, not attaching morphological meaning to it).

Soon afterward it was recognized that the postulated diencephalic columns are not continuous with the brain stem and telencephalic ones. As a consequence, each of these sets had to be conceived of as forming independent partial models of the respective brain parts. This left in between three vaguely defined, unmodeled, and badly understood transition areas: isthmus, pretectum, and telencephalic stalk. Efforts to extrapolate to diencephalon and telencephalon the columnar 'functions' of the brain stem and spinal cord (i.e., visceral-somatic sensory and motor functional correlations) were also unproductive. Paradoxically, though the promise of the columnar model stumbled on the forebrain, this did not lead to any doubts about its potency as a paradigm or usefulness as a forebrain model, because by that time it had become a neuroanatomical dogma. For a long time, dogmatic transmission of the columnar Herrick model in research and classroom pushed the alternative neuromeric models nearly to oblivion. Most neuroscientists to this day have been made to believe that the supposedly dorsoventral columnar series of epithalamus–dorsal thalamus–ventral thalamus–hypothalamus is a fact, not a risky conjecture of a hundred years ago.

The fundamental failure of the columnar forebrain model was that it redefined the observable forebrain axis, negating its observable curvature and substituting an arbitrary ideal straight axis which is not supported by any specific data. The columnar straight brain axis crosses from the pontine brain stem into the the 'caudal' hypothalamus, then traverses the hypothalamus and preoptic area 'longitudinally,' to enter the telencephalon and end in the olfactory bulb (this last part is obviously inconsistent with the paired paramedian nature of the olfactory bulbs and telencephalic hemispheres). Herrick curtly explained such pragmatic axial redefinition as "controversial . . . but convenient."

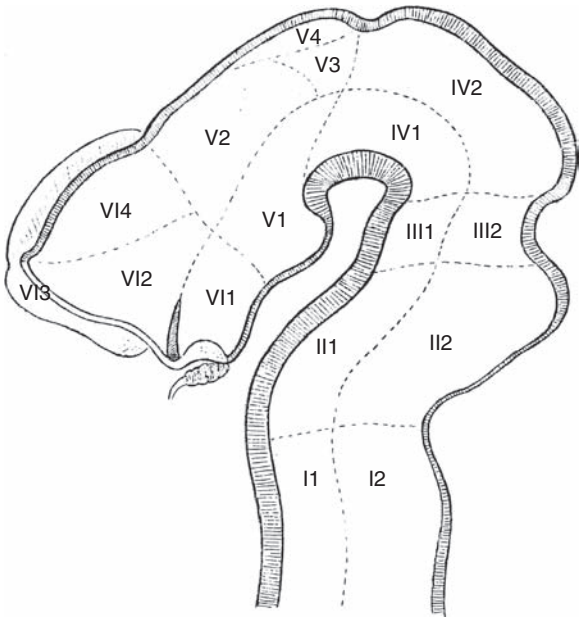

Figure 2 Model of W His (1895). Six transversal units (I–VI) are recognized along the bent neural tube. The alar-basal boundary appears through the entire lateral wall, parallel to the floor and roof plates. Domains I and II correspond to the myelencephalon and metencephalon, respectively. Domain III is the isthmus. Domain IV is the midbrain, and domain V is the 'diencephalon proper.' Domain VI contains part of hypothalamus, the eye, and the telencephalon. Basal plate domains are identified as I (II–VII) and alar plate domains are marked as 2 (I2–VI2). An additional domain VI3 represents the olfactory bulb, and the V3 and V4 domains refer to the metathalamus and habenula, respectively.

In the subsequent era of spectacular experimental neuroanatomical advances (axonal degeneration, axonal transport, electron microscopy, chemical anatomy), which extended up to the recent 1980s, Herrick's columnar model seemed to encompass without problems the accumulating hodological and chemoarchitectonic data on the forebrain. Stereotaxic topographic references for lesions and tracer injections worked well with the idea of a straight axis of the entire brain, which could be naively thought to be reproduced by the length axis of the stereotaxic apparatus. Only isolated embryologists (and then only those who looked at whole mounts and sagittal sections, procedures that curiously fell into disuse) insisted now and then that the brain axis is always curved and therefore topologic transversal and longitudinal dimensions had to be defined in accordance with the specific part of the neural tube considered.

The Rebirth of Neuromeric Models

A fully new set of neuroanatomical developmental data started to accrue during the 1980s and 1990s. These data included observations on the expression domains of neural developmental genes, possible thanks to the new *in situ* hybridization protocol for transcribed messenger RNA (and other correlative molecular biology and genomic advances). This procedure renders visible the cells that are in the process of reading out piecewise the information coded in the genome. Since many of these genes are causally determinant of the structural and histogenetic patterning of the neural tube wall, their expression patterns and the boundaries defined by them are highly relevant for brain models. It was soon discovered that some genes show longitudinal patterns of expression and others show transversal patterns (actually, both aspects usually appear in combination). All efforts to encompass these patterns within the forebrain columnar model have failed or have led to unparsimonious and highly convoluted *ad hoc* interpretations. On the other hand, the hindbrain and spinal cord columnar model does agree significantly with longitudinal gene patterns but highlights at the same time that observed transversal patterns relate specifically to the old neuromeric models. It turns out that neuromeric and columnar patterns coexist in the hindbrain and spinal cord, as predicted long ago by defenders of the segmental approach.

Similar analysis of forebrain gene expression data in the context of a forebrain neuromeric model (using the original bent axis) showed the capacity of this model to encompass and give morphologic significance (developmental function) to the new set of molecular causal data. The apparent rebirth of a neuromeric paradigm in the forebrain and hindbrain (where neuromeres are best visible) pointed the way to the possibility of conceiving a general segmental model of the entire central nervous system, in which longitudinal zones (columns, but different ones in the forebrain than those postulated by Herrick and his followers) and transverse neuromeres combine to interpret and predict the nature of causal phenomena operating in the construction of the brain. This sort of ultimate or synthetic brain model was called the prosomeric model, as developed in several reports and reviews by Puelles and Rubenstein.

See also: Brains of Primitive Chordates; Evolution of Vertebrate Brains; Forebrain Development: Holoprosencephaly (HPE); Forebrain: Early Development.

Further Reading

Herrick CJ (1910) The morphology of the forebrain in amphibia and reptilia. *Journal of Comparative Neurology* 20: 413–547.

Herrick CJ (1933) Morphogenesis of the brain. *Journal of Morphology* 54: 233–258.

Herrick CJ (1948) *The Brain of the Tiger Salamander, Amblystoma tigrinum*. Chicago: The University of Chicago Press.

His W (1892) Zur allgemeinen Morphologie des Gehirns. *Archiv für Anatomie und Entwicklungsgeschichte* 346–383.

His W (1893) Über das frontale Ende des Gehirnrohres. *Archiv für Anatomie und Physiologie/Anatomie Abteilung* 3/4: 157–171.

His W (1893) Vorschläge zur Eintheilung des Gehirns. *Archiv für Anatomie und Physiologie/Anatomie Abteilung* 3/4: 173–179.

His W (1904) *Die Entwicklung des menschlichen Gehirns während der ersten Monate*. Leipzig, Germany: Hirzel.

Kuhlenbeck H (1973) *The Central Nervous System of Vertebrates, Vol. 3, Part II: Overall Morphological Pattern*. Basel, Switzerland: Karger.

Orr H (1887) Contribution to the embryology of the lizard. *Journal of Morphology* 1: 311–363 (plates XII–XVI).

Puelles L (1995) A segmental morphological paradigm for understanding vertebrate forebrains. *Brain, Behavior and Evolution* 46: 319–337.

Puelles L (2001) Brain segmentation and forebrain development in amniotes. *Brain Research Bulletin 55*: 695–710.

Puelles L and Rubenstein JLR (1993) Expression patterns of homeobox and other putative regulatory genes in the embryonic mouse forebrain suggest a neuromeric organization. *Trends in Neuroscience* 16: 472–479.

Puelles L and Rubenstein JLR (2003) Forebrain gene expression domains and the evolving prosomeric model. *Trends in Neuroscience* 26: 469–476.

Von Kupffer K (1906) Die Morphogenie des Centralnervensystems. In: Hertwig O (ed.) *Handbuch der Vergleichende und Experimentelle Entwicklungslehre der Wirbelthiere*, vol. 2, pp. 1–119. Jena, Germany: Fischer.

Forebrain: Early Development

F Cavodeassi, M Kapsimali, S W Wilson, and R M Young, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Key Features of Early Forebrain Development

The forebrain is specified within the anterior neural plate during gastrulation, and by the end of somitogenesis, this territory has established ventrally the hypothalamus and floor plate and dorsally the telencephalon, eye, and diencephalon, which itself is further subdivided in prethalamus, zona limitans intrathalamica (ZLI), thalamus, and pretectum (Figure 1). All forebrain structures arise from a simple sheet of neuroepithelial cells, which undergoes regionalization and dramatic morphogenesis during early developmental stages. The morphology and function of various forebrain structures may be different among vertebrates, but the general organization and early patterning steps are likely conserved.

The initial formation of the forebrain is linked to the specification of neural identity in the embryonic ectoderm. Neural character is induced by gastrula stages through a combination of fibroblast growth factor (FGF) activity and extracellular antagonists of bone morphogenetic proteins (BMPs). The induced neural plate initially has anterior character, as evidenced by expression of genes that later become restricted to prospective forebrain and midbrain regions. Subsequently, the activity of various signaling molecules (including FGFs, Wnts, and retinoic acid) emanating from posterior regions of the embryo leads to caudalization of the neural tissue. These two events, neural induction and caudalization, eventually result in the generation of a nascent central nervous system (CNS) with graded anteroposterior (AP) character.

To acquire and maintain anterior character, the rostral neural plate must avoid exposure to caudalizing factors. Several mechanisms ensure this happens. These vary between species, but there are some common themes. For instance, gastrulation movements shift the anterior neural plate away from the sources of caudalizing signals. In addition, locally secreted signals from a variety of rostral tissues antagonize the activity of the caudalizing factors, protecting the anterior neural plate and ensuring it retains anterior character.

The Wnt/β-catenin pathway is perhaps the most crucial of the signaling cascades critical for early regional AP patterning of the rostral neural plate. Embryos carrying mutations that lead to enhanced Wnt activity in the anterior neural plate lack the telencephalon, rostral hypothalamus, eyes, and prethalamic territories. Thus, specification of the rostral forebrain requires low levels of Wnt/β-catenin activity. Conversely, local suppression of Wnt signaling can lead to the expansion of the rostral at the expense of more caudal forebrain domains.

The gradual regionalization of the neural plate as a consequence of the activity of Wnts and other signals leads to the subdivision of the anterior neural plate (or prospective forebrain) into discrete territories defined by the restricted expression of various transcription factors of the Emx, Irx, Pax, Six, and other families. Some of these transcription factors are likely to be direct effectors of the signals that regionalize the neural plate. For instance, Six family genes are repressed by Wnt signals and consequently expressed only in rostral regions. In turn, Six family proteins suppress Wnt activity, thus establishing a feedback loop that ensures tightly regulated levels of Wnt activity in the prospective forebrain.

One consequence of the initial regionalization of the neural plate is the establishment of groups of cells that function as local signaling centers, orchestrating subsequent neural development at different positions of the neuroepithelium (also called secondary organizers). These organizers influence cellular fate and growth of the adjacent tissues, further refining the initial pattern of the neural plate.

Along the rostrocaudal axis of the forming rostral CNS, cell populations with organizer properties are found at the anterior neural border or anterior neural ridge (ANB or ANR), at the ZLI, and at the midbrain–hindbrain boundary (MHB). The ANB/ANR is a source of signals including Wnt antagonists and FGFs that promote telencephalic fate; the ZLI is established later than the ANB/ANR and the MHB and is a source of sonic hedgehog (Shh) which influences the patterning and growth of surrounding prethalamic and thalamic structures; and the MHB is a source of FGF and Wnt signals that pattern the midbrain and rostral hindbrain and contribute to defining the caudal boundary of the forebrain.

Along the dorsoventral axis, the most prominent cells with signaling properties are ventral midline cells of the floor plate and hypothalamus. These cells produce Hh, Nodal, and other signals and are important for regionalization and neuronal patterning of the ventral forebrain and for splitting of the initially coherent eye field into left and right eyes. Cells along the dorsal midline of the forebrain, including the derivatives of the ANB/ANR and more caudal roof plate tissue, are also the source of various secreted signals, including FGFs, Wnts, and BMPs. The roles for such signals are

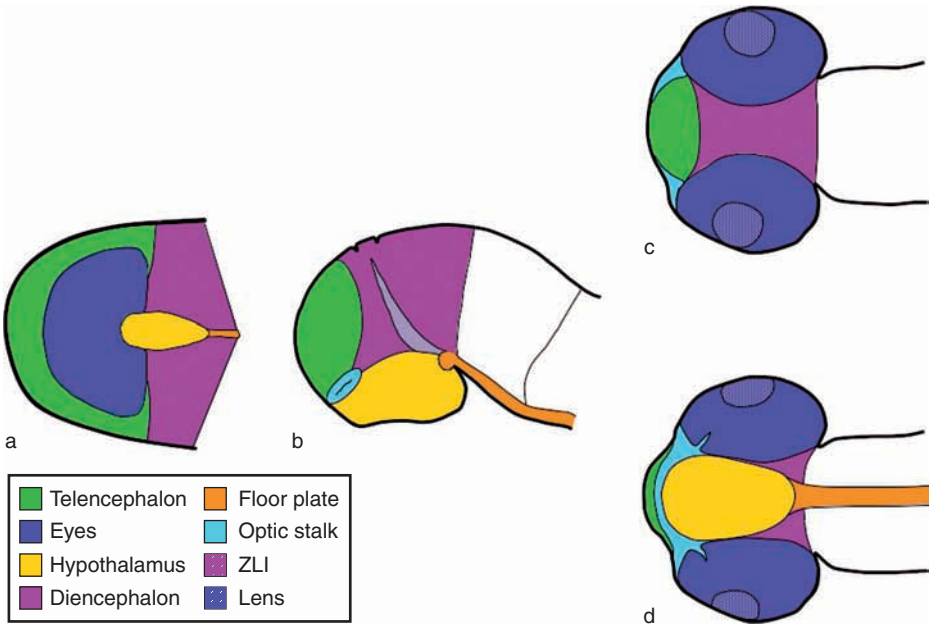

Figure 1 (a) Organization of the forebrain territories at neural plate stage. The groups of cells giving rise to each territory are shown in different colors. (b–d) The same color code is used to show the relative positions of the different territories in embryonic brains at a later stage of development, from lateral (b), dorsal (c), and ventral (d) views. ZLI, zona limitans intrathalamica.

perhaps best understood with respect to the regionalization of the mammalian dorsal telencephalon.

The morphogenesis that occurs during transition from the simple sheet of cells that constitutes the anterior neural plate to the complex structures of the mature forebrain and eyes (see Figure 1) is the least understood aspect of early forebrain development. This area is likely to see considerable research activity in the next few years.

Specification and Early Development of Various Forebrain Domains

Hypothalamus

The hypothalamus is the master regulator of neuroendocrine functions in the brain. It is a derivative of the most ventral/rostral regions of the prospective forebrain, and medial (ventral) regions of the hypothalamus are in continuity with floor plate tissue that runs the entire length of the ventral CNS.

During gastrula stages, prospective midline neural plate cells that will form floor plate and hypothalamus are located near the organizer from which they will extend rostrally and caudally along the CNS as convergence and extension movements shape the AP axis of the embryo. Initially, all prospective neural midline may be specified with floor plate character. However, as rostral midline neural tissue progressively extends within the rostral neural plate, it escapes the caudalizing influence of signals from caudal mesendodermal and neural tissues and begins to acquire rostral, hypothalamic identity. Thus by the end of gastrulation, ventral midline tissue of the neural plate is overtly divided into rostral, hypothalamic, and more caudal floor plate domains (Figure 1). The prospective hypothalamic domain is subsequently further subdivided along the AP axis and differentiates as specific territories with different identities and different signaling properties.

The Nodal and Hedgehog (Hh) signaling pathways are implicated in the induction of both hypothalamus and floor plate although their relative contributions to this process vary between classes of vertebrates. In fish, Nodal signaling is essential for induction of medial regions of the floor plate and caudomedial regions of the hypothalamus. Hh signaling is not essential for formation of these domains but is critical for the induction and patterning of ventral tissues adjacent to the most medial floor plate and lateral and rostral to the caudal hypothalamus. Similar distinctions between medial and lateral floor plate and hypothalamic domains probably exist in all other vertebrate classes, but at least in mammals, the Hh pathway plays a more prominent role in the formation of all midline structures.

As with more dorsal regions of the CNS, the level of Wnt/β-catenin signaling influences the rostrocaudal character of the ventral midline and adjacent tissue. Hypothalamic fate specification requires low

levels of Wnt/β-catenin activity, and if Wnt signaling is enhanced, cells with more caudal floor plate character extend further rostrally into the territory normally occupied by hypothalamus. Within the hypothalamus itself, raised Wnt signaling can promote caudal at the expense of rostral hypothalamic identity, suggesting that a low level of Wnt signaling may cooperate with Nodal signals to confer posterior hypothalamic identity. In addition to Nodals and Wnts, BMP and Hh signals are also important in conferring regional hypothalamic character to neurons generated in this portion of the ventral forebrain. For instance, Shh together with BMP7 is sufficient to confer hypothalamic dopaminergic identity to neural progenitor cells *in vitro*.

The various signals that influence hypothalamic development are thought to induce transcription factors that further refine signaling activity and confer regional character and specific neuronal fates on ventral CNS cells. For instance, *six3* expression is suppressed by Wnt signaling and in turn represses Wnt activity whereas, conversely, *irx3* expression is induced by Wnt signaling and itself promotes Wnt signaling. Mutual repressive interactions between Six and Irx family homeodomain proteins therefore contribute to the division of rostral neural tissue into rostral *six3*-expressing and caudal *irx3*-expressing domains. Several of the transcription factors that function downstream of these early regional patterning events to promote the generation of hypothalamic neurons have been characterized and include homeodomain proteins of the Nk2 family and neurogenin subfamily of basic helix–loop–helix proteins.

Telencephalon

The adult telencephalon is the most structurally diverse region of the vertebrate forebrain, but at early stages of development, patterning mechanisms are likely to be conserved across all vertebrates. At these early stages, the telencephalon has two major subdivisions: a dorsal pallial region that gives rise to the neocortex in mammals and analogous structures in other vertebrate classes and a ventral, or subpallial, region that gives rise to basal ganglia or striatopallidal structures.

At the neurula stage, the prospective telencephalon constitutes a band of cells located in the most anterior neural plate, rostral and lateral to the eye field (Figure 1). For the telencephalon to form, Wnt/β-catenin signaling must be suppressed, and several genetic mutations in fish which lead to enhanced Wnt activity in the anterior neural plate lead to transformation of the telencephalon (and eyes) to more caudal forebrain fates. Wnt signaling is suppressed by a variety of antagonists secreted from cells within the prospective telencephalon itself and from adjacent tissues, including underlying mesendoderm. For instance, in fish, cells located in the ANB/ANR express a Wnt antagonist of the secreted frizzled-related protein family, which when overexpressed expands telencephalic territories. Secreted frizzled-related proteins are believed to function by binding Wnt proteins and thus preventing Wnt ligands from interacting with their receptors. Although high levels of Wnt activity at early stages of development can suppress all telencephalic development, graded levels of Wnt activity can also influence regional patterning within the nascent telencephalon, promoting dorsal, pallial fates and suppressing ventral, subpallial identity.

In addition to Wnt antagonists, the ANB/ANR expresses FGF family members, and these signaling proteins play a variety of crucial roles in telencephalic development. It is unclear whether FGF signaling is required for induction of the telencephalon, but subsequent to induction, FGFs regulate regional patterning. Together with Shh, FGF signaling is required for induction of ventral, subpallial telencephalic character. Studies in mouse have revealed that within the dorsal, pallial telencephalon, FGF signaling is a critical regulator of arealization of the cortex. This polarization is mediated by transcription factors that appear to have graded activity across the pallium. For instance, antagonistic interplay between Pax6 and Emx2 influences the caudomedial and rostrolateral character of pallial territories. Thus, Emx2 mutant mice display expanded rostrolateral pallial character at the expense of caudomedial territories, whereas Pax6 mutants display the opposite phenotype.

Eye

The neural retina, pigment epithelium, and optic stalk or nerve are all derivatives of the anterior CNS. Within the neural plate, precursors for these eye structures initially form a coherent field of cells, termed the eye field, which spans the midline. Morphogenetic movements in combination with signals derived from midline tissues subsequently result in cells from the eye field moving laterally and evaginating to form the left and right optic vesicles. Distal cells within the vesicles form the neural and pigmented retina while more proximal and medial cells contribute to the optic stalks or nerves.

During gastrulation, the eye field is defined by the combinatorial expression of various transcription factors including Pax6, Six3, and Rx. Mutations in these genes result in eye malformations in all organisms analyzed, from humans to mice and fish. All these genes, generically known as eye specification genes, are part of a complex cross-regulatory network essential for eye development, and their functions during this process are highly conserved throughout the animal kingdom. It is unclear how the expression of these genes is first

induced, but expression is at least modulated by the same signals that regionalize other domains of the anterior neural plate. Thus, for example, Pax6, Rx3, and Six3 are restricted to anterior regions of the zebra fish neural plate by the activity of the canonical Wnt/β-catenin pathway, and excess Wnt activity can suppress both their expression and subsequent eye formation. The eye field is therefore induced in anterior regions of the prospective forebrain, where high levels of Wnt/β-catenin activity are absent. At least in fish, the eye field is further protected from canonical Wnt signaling by the activation, within the prospective eye field, of a noncanonical Wnt pathway that antagonizes canonical Wnt signaling, thereby promoting the expression of *rx3* and *six3* in this domain.

Both telencephalon and eye field are induced in regions of low or absent Wnt signaling. Although the mechanisms that distinguish prospective telencephalon from eye field are unclear, it is likely that graded BMP signaling is involved in the segregation of these two territories. At least in fish, high levels of BMP signaling encroaching from the margins of the neural plate suppress expression of eye field genes, limiting the lateral extent of eye field specification. Thus, the activation of the eye specification genes is promoted in the region sandwiched between prospective telencephalic and dorsal diencephalic domains that receive high levels of BMP signaling and more caudal diencephalic domains that receive high levels of Wnt signaling. The Rx3 transcription factor is likely to be an effector of the fate choice between telencephalon and eye (in addition to a role in eye morphogenesis), as telencephalic markers are ectopically expressed in the eye field of zebra fish rx3 mutants. It is unclear whether the eye specification genes are under the direct transcriptional control of these signaling pathways; to date, only *six3* has been shown to be transcriptionally regulated by the Wnt signaling pathway.

The eye primordium is specified as a single domain straddling the midline, but two optic vesicles, evaginating from the lateral walls of the neural tube, are established from this single domain. This is accomplished by morphogenetic processes that split the territory in two and lead to lateral evagination of the optic primordia. Do the eye specification genes instruct a specific program of morphogenesis for the eyes, or does this process depend on other, yet to be described, mechanisms? The answer is probably a combination of both. For example, *rx3* mutants in zebra fish and medaka show defects in the evagination of the optic vesicles, but early markers of eye field formation are initially expressed normally, suggesting that *rx3* regulates morphogenesis of the forming optic vesicles. However, the genes and proteins that are the direct effectors of the morphogenetic movements are currently unknown.

Nodal and Hh signals emanating from ventral midline tissues are also involved in the separation of the retinae. Disruption of either of these signals leads to phenotypes in which the retinae remain fused across the midline of the brain and the optic stalks are absent. In these conditions, the hypothalamus also fails to form, and it has been proposed that the absence of this structure, which normally would move anteriorly between the eyes, contributes to the failure of eye field separation. Other mutant conditions affecting the rostral movement of the hypothalamus, such as *silberblick/wnt11* mutant in zebra fish, similarly show varying degrees of cyclopia. Ongoing studies using advanced imaging and transgenic approaches are allowing much better documentation of the morphogenetic processes that accompany eye formation in normal development and in conditions that lead to cyclopia, and new insights into these processes will soon be forthcoming.

Diencephalon

According to the prosomeric model of forebrain organization, its most caudal part is subdivided into prethalamus (previously ventral thalamus), thalamus (previously dorsal thalamus), and pretectum, with the ZLI being located between prethalamus and thalamus (**Figure 1(b)**). Of all the major subdivisions of the forebrain, these domains are the most poorly understood in terms of their early development. For instance, it remains uncertain whether these three subdivisions are already specified at neural plate stages of development or whether they resolve at later stages. The ZLI is considered to be the boundary between prechordal and chordal domains of the brain, and one would expect this position to be defined very early, but as yet no sufficiently detailed fate maps exist to indicate where within the neural plate the cells that constitute the ZLI originate.

Despite this uncertainty, it is generally thought that the ZLI will form at a position defined on the neural plate by the boundary of expression between *six3* in the rostral forebrain and *irx3* in the posterior diencephalon–midbrain at neural plate stages. As discussed in the section titled 'Hypothalamus,' the expression of these genes is regulated by Wnt/β-catenin activity, which activates *irx3* and suppresses *six3*. This implies that a specific level of Wnt/β-catenin activity defines the position at which the ZLI forms and thus the relative size of prospective prethalamus and thalamus.

Later in development, when the embryo has gone through neurulation, the ZLI forms a compartment between the prethalamus and the thalamus. The ZLI expresses Shh and is therefore potentially a secondary organizer in the forebrain. In support of this, studies in mice, chick, and zebra fish suggest that Hh proteins secreted from the ZLI regulate regional patterning and growth of the prethalamus and the thalamus. It is likely that further roles will be found for the ZLI in the near future.

The boundary between the caudalmost forebrain and the midbrain corresponds to the interface between *pax6* expression in the forebrain and *pax2* and *engrailed* expression in the midbrain. Similar to the case with Six3 and Irx3, corepressive interactions between these genes are thought to define the position of the forebrain–midbrain boundary. Midbrain development is promoted by FGF8, and overexpression of *fgf8* is able to expand *en2*, inhibit *pax6* expression, and therefore shift the forebrain–midbrain boundary rostrally.

The dorsal part of the diencephalon is unusual in exhibiting marked asymmetries between left and right sides of the brain. These asymmetries are likely present in all vertebrates but are more prominent in some species than others. Studies in zebra fish are unraveling the mechanisms that establish this asymmetry and determine the directionality and laterality of the asymmetric nuclei.

See also: Brain Asymmetry: Evolution; Brain Development: The Generation of Large Brains; Evolution of Vertebrate Brains; Forebrain Development: Prosomere Model; Forebrain Development: Holoprosencephaly (HPE); Retinoic Acid Signaling and Neural Patterning; Wnt Pathway and Neural Patterning.

Further Reading

Campbell K (2003) Dorsal-ventral patterning in the mammalian telencephalon. *Current Opinion in Neurobiology* 13: 50–56.

Cavodeassi F, Carreira-Barbosa F, Young RM, et al. (2005) Early stages of zebrafish eye formation require the coordinated activity of Wnt11, Fz5, and the Wnt/β-catenin pathway. *Neuron* 47: 43–56.

Chuang JC and Raymond P (2002) Embryonic origin of the eyes in teleost fish. *Bioessays* 24: 519–529.

Donner AL and Maas RL (2004) Conservation and non-conservation of genetic pathways in eye specification. *International Journal of Developmental Biology* 48: 743–753.

Foley AC and Stern CD (2001) Evolution of vertebrate forebrain development: How many different mechanisms? *Journal of Anatomy* 199(1–2): 35–52.

Kapsimali M, Caneparo L, Houart C, and Wilson SW (2004) Inhibition of Wnt/Axin/beta-catenin pathway activity promotes ventral CNS ventral midline tissue to adopt hypothalamic rather than floor plate identity. *Development* 131: 5923–5933.

Kiecker C and Lumsden A (2005) Compartments and their boundaries in vertebrate brain development. *Nature Reviews Neuroscience* 6: 553–564.

Ohyama K, Ellis P, Kimura S, and Placzek M (2005) Directed differentiation of neural cells to hypothalamic dopaminergic neurons. *Development* 132: 5185–5197.

Puelles L and Rubenstein JL (2003) Forebrain gene expression domains and the evolving prosomeric model. *Trends in Neurosciences* 26: 469–476.

Rhinn M, Picker A, and Brand M (2006) Global and local mechanisms of forebrain and midbrain patterning. *Current Opinion in Neurobiology* 16: 5–12.

Stigloher C, Ninkovic J, Laplante M, et al. (2006) Segregation of telencephalic and eye field identities inside the zebrafish forebrain territory is controlled by Rx3. *Development* 133(15): 2925–2935.

Sur M and Rubenstein JL (2005) Patterning and plasticity of the cerebral cortex. *Science* 310: 805–810.

Wilson SW and Houart C (2004) Early steps in the development of the forebrain. *Developmental Cell* 6: 167–181.

Fovea: Primate

A Hendrickson, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Anatomy of the Primate Fovea

The 'fovea centralis' is a morphological specialization for high acuity which is found in the central retina of *Tarsius*, Old World monkeys, New World monkeys except *Aotes*, apes, and humans. The human fovea lies on the visual axis of the eye 4 mm temporal and 0.8 mm inferior to the center of the optic disc. Helmholz showed in the early twentieth century that the fovea is the site of maximal visual acuity. It has four anatomical characteristics which distinguish it from the surrounding retina and are modifications to support high acuity:

1. Retinal layers from the outer plexiform to the ganglion cell layer (GCL) are displaced laterally, forming a pit. *Fovea* is Latin for small pit, and it was identified as a characteristic of human retina by the late eighteenth century.
2. The foveal center or 'foveola' contains the highest density of cone photoreceptors in the retina. Cone photoreceptors function in bright light and support high acuity and color vision. Foveolar cone densities average 200 000 mm^{-2} in most primates compared to peripheral cone densities <20 000 mm^{-2}.
3. The most important foveal synaptic circuit is the 'private line' midget circuit which runs from cone to bipolar cell (BP) to GC. This preserves a faithful representation of the high-pixel foveal cone mosaic for transmission to central visual centers.
4. The fovea is free of blood vessels throughout life which has implications for foveal development. Each of these modifications is discussed in more detail below.

The fovea centralis is divided into four concentric circular zones which are shown for a *Macaca* monkey fovea at low magnification in Figure 1. The central retina overlapping the four foveal rings contains a concentration of yellow pigment, so this region is also called the macula lutea or 'yellow spot.' Details of retinal layers and their neurons in a 13-year-old human foveola, fovea, and parafovea are shown at higher magnification in Figure 2. A drawing of the primate fovea is shown in Figure 3 which contains a diagram of foveal cone synaptic circuits.

The 'foveola' is the central 250–300 μm, and represents the central 1°20′ of the visual field. It is slightly more than 100 μm thick and is formed only by cone outer and inner segments, a tightly packed 10–12 deep layer of cone cell bodies (Figures 2(a) and 2(c)), and surrounding Müller glial cell processes. All other layers are displaced peripherally by the pit. The foveola is the region of highest cone density so that foveolar cones are very thin and long and have the longest inner and outer segments in the retina (Figures 2 and 3). Rod photoreceptors are absent. The next ring is the 'fovea,' which extends 750 μm around the foveola. The fovea covers 5.5° of the central visual field. The foveal region contains the pit which is round, about 1000 μm at its widest, and 200–240 μm deep. The walls of the pit form the foveal slope which is characterized histologically by having a GCL which goes from no cells near the foveola to a GCL up to eight cells deep on the edge of the fovea. Cone density is lower than in the foveola, and cones are thicker. Rods are present at low density in the fovea (Figure 2(b), R). The outer plexiform layer appears in the fovea. It contains a thick layer of long cone axons called the fibers of Henle and the synaptic pedicles of the foveolar cones (Figure 2(b), P) which are displaced peripherally during foveal development. A foveal cone midget circuit is drawn in color in Figure 3. Capillaries appear in the inner retina and encircle the foveal slope, where they form a foveal avascular zone (FAZ) about 400–500 μm wide (Figure 3). The third ring is the 'parafovea,' which is ~500 μm wide, and the outermost ring is the 'perifovea,' which is ~1.5 mm wide and extends almost to the optic disc (Figure 1). These zones do not have sharp morphological boundaries. Rods are more numerous in the parafovea (Figure 2(c), R) and cones are thicker and less numerous. In the perifovea rods are the dominant photoreceptor and the GCL decreases to one to two cells thick near the optic disc.

Foveal Pit

Although numerous reasons have been put forth for having a foveal pit, its role in vision is not clear. Blood vessels encircle the foveal slope (Figure 3), forming an FAZ about 500 μm in diameter in humans and monkeys. Snodderly et al. show that vascular coverage or 'screening of light' of the photoreceptor layer by blood vessels is 0 in the FAZ, but then rises rapidly to 30% at 1 mm and 45% at 2 mm. One functional

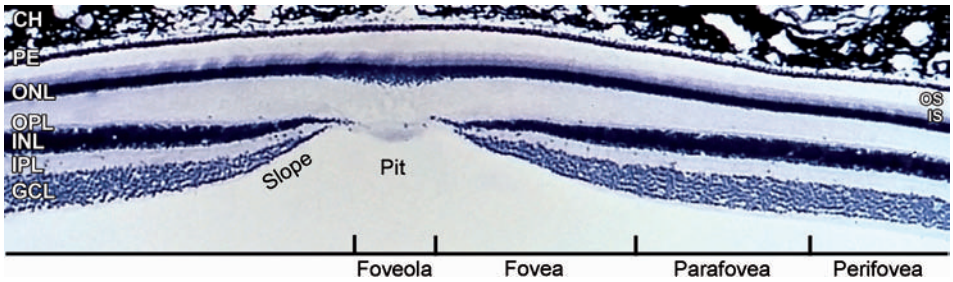

Figure 1 Photomicrograph of a section through the central retina of an adult *Macaca* monkey. The layers and their abbreviations are used throughout all of the figures. CH, choroid; PE, pigment epithelium; OS, outer segment; IS, inner segment; ONL, outer nuclear or photoreceptor layer; OPL, outer plexiform layer; INL, inner nuclear layer; IPL, inner plexiform layer; GCL, ganglion cell layer. The subdivisions of the foveal region are indicated.

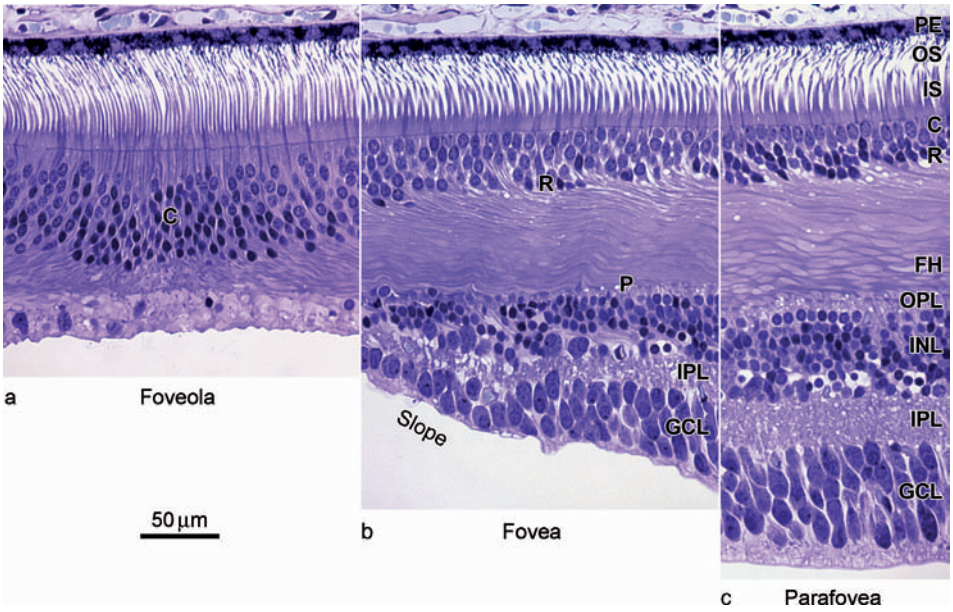

Figure 2 Photomicrographs of a 2 µm plastic section through a 13-year-old human retina. The three retinal regions and their labels are aligned on the outer limiting membrane. In the foveola (a) cone (C), cell bodies form a tightly packed multilayer. These cones have very long and thin OS and IS. Cone pedicles are absent. In the fovea (b), inner retinal layers and the foveal slope or walls of the pit appear. Cone ISs become thicker and a few rods (R) appear under the cone cell bodies. The outer plexiform layer contains many fibers of Henle (FH) ending in cone pedicles (P). These are the pedicles of the foveolar cones which were displaced peripherally during pit formation (see Figure 3). In the parafovea (c), rods are the predominant type of photoreceptor while cones are reduced to a single layer. Inner retinal layers are at their full thickness, although the GCL will become thinner in more peripheral retina.

reason for the FAZ and foveal pit may be to provide an optically clear and stable path to the foveolar cones. A second function of the pit suggested by Springer and Hendrickson may be to facilitate cone packing during foveal development (see below).

Foveal Photoreceptor Distribution

Cones and rods can be identified by their size difference using Nomarski optics as shown in the top row of photographs in Figure 4, or by riboprobes or antisera that recognize cone versus rod opsins. Cone subtypes can be differentiated by antisera that recognize opsin protein in short wavelength-selective or blue cones versus long and medium wavelength-selective or red and green cones. To date, no riboprobes or antisera have been devised that can morphologically differentiate red versus green cone opsin. Cone-selective wavelength absorption combined with adaptive optics for the first time enabled Williams and Roorda to visualize the blue, red, and green cone ratios and mosaics in living human central retina.

Photoreceptor distribution in human retina is shown in Figure 4. The mean number of photoreceptors in a young adult human retina is 4.6 million cones and 92 million rods based on the Curcio et al.

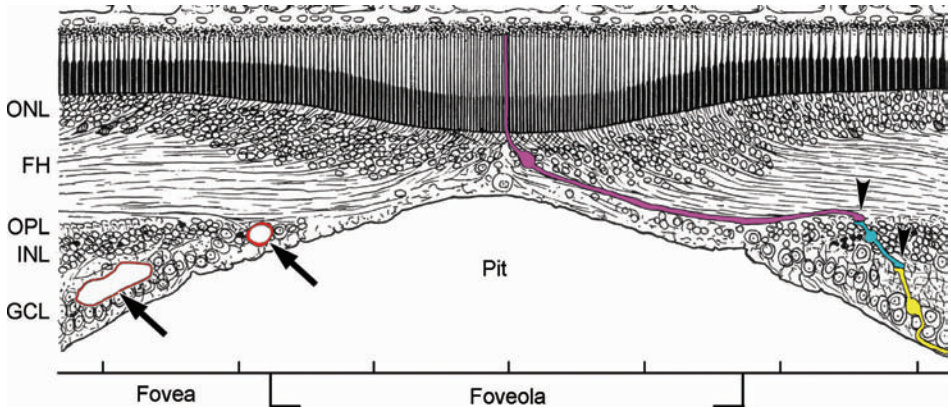

Figure 3 A drawing modified from Polyak of a monkey foveola and fovea. The ticks on the lower bar indicate 100 μm. A midget circuit is drawn on the right side with the foveolar cone in magenta, the midget bipolar cell in cyan, and the midget ganglion cell in yellow. Arrowheads indicate synaptic contacts. On the left, the arrows indicate the centralmost blood vessels on the foveal slope which outline the pit and delineate the foveal avascular zone.

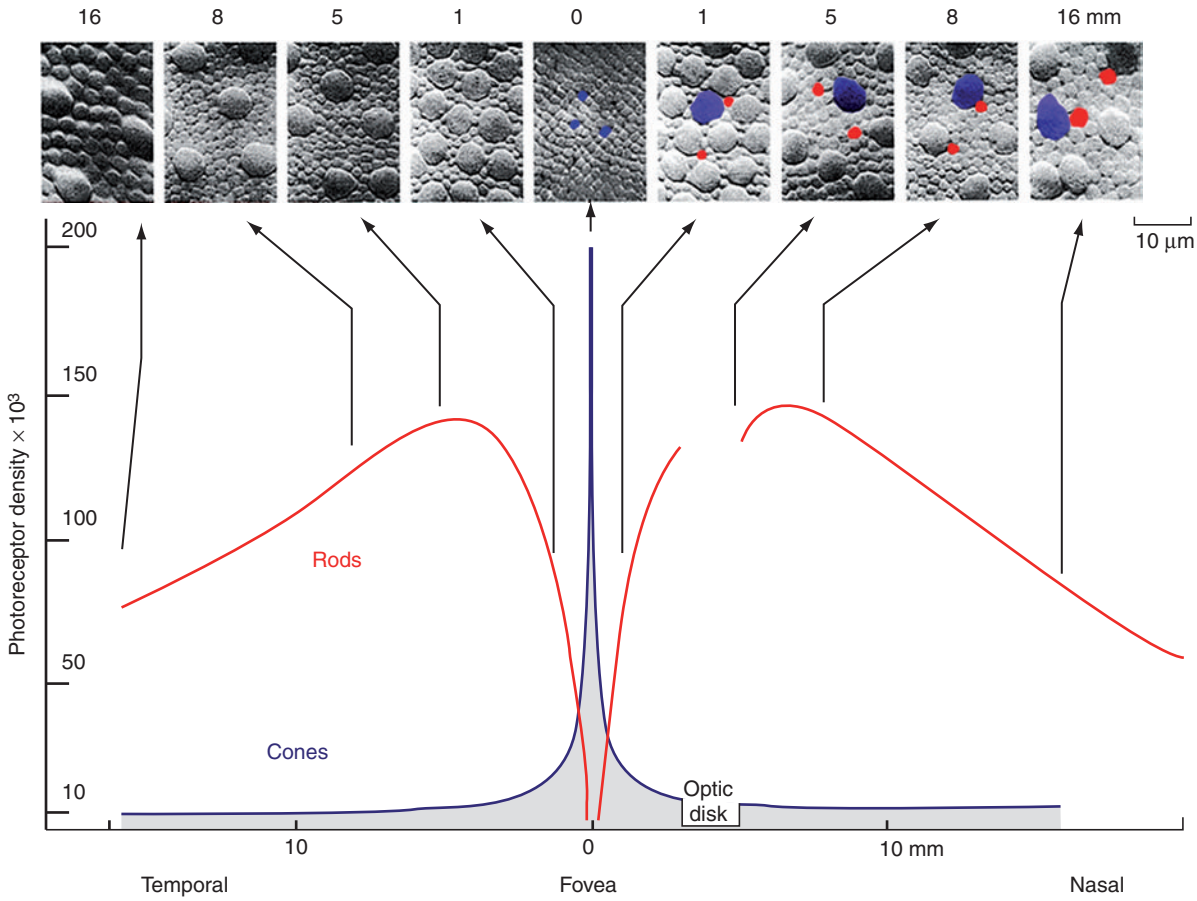

Figure 4 The distribution and density per square millimeter of primate photoreceptors. Cones are in blue and rods in red. The upper band of Nomarski images shows the change in size and ratio of human rods (red) and cones (blue) out to 16 mm from the foveal center. Upper scale bar is for these photographs. Adapted from Curcio CA, Sloan KA, Kalina RE, and Hendrickson AE (1990) Human photoreceptor topography. *Journal of Comparative Neurology* 292: 497–523 and Dennis Dacey.

sample. Using the same counting methods, Packer et al. find that the smaller *Macaca* monkey retina averages 3.1 million cones and 60 million rods. In both primates, rods are absent in the fovea, then their density rises sharply to a peak (Figure 4, red line) of $\sim 150\,000\,\mathrm{mm}^{-2}$ at the eccentricity of the optic disc, and from there rod density decreases slowly across peripheral retina. The rod:cone ratio is

20:1 across much of the human retina. Cone density (Figure 4, blue line) peaks sharply in the foveola and then drops rapidly into the periphery, with a 1:1 rod: cone ratio reaching 500 μm eccentricity. Cone density remains higher along the nasal horizontal meridian so that at the same eccentricity nasal retina has 2–3 times as many cones as temporal retina. The same pattern is found for GC in both human and monkey retinas. In Old World monkeys and humans, red and green cones are around 90% of the cone population with blue cones constituting 8–12%. In humans, blue cones are typically separated from one another by other cones forming a six-sided but random mosaic while some New World monkeys have two to four blue cones in a cluster. Adaptive optical imaging finds that central human red and green cones form little clusters with no regular overall pattern. A surprising finding from these studies is that, although most humans have a ratio averaging 2red:1green, individuals can range from 1red:8green to 9red:1green, yet have normal color vision. By contrast, *Macaca* and baboon Old World monkeys typically have a cone ratio of 1green:1red across the retina.

The fovea is the center of photoreceptor topography and has a circular arrangement in which photoreceptor types systematically drop out. Near the optic disc (Figure 4, 5 mm), cone density is low and rod density is at its peak. Central to this eccentricity, cone density rapidly rises and rod density drops until in the fovea (Figure 4, 1 mm), the density is 40 000 mm^{-2} for each population. In humans, the last significant number of rods is found at 300 μm (Figure 5, black dots) but in *Macaca* monkeys a few rods continue into the foveola. In humans, blue cones (Figure 5, blue dots) are absent in the central 100 μm of the foveola, where only red and green cones (Figure 5, red and green dots) are present. By contrast, in *Macaca* and marmoset monkeys, blue cones continue across the foveola at decreased density. Thus, the primate fovea has subtle differences between humans and monkeys for blue cone and rod distribution.

Foveal cone diameter changes dramatically over the central 5 mm, which is illustrated in Figures 2 and 4 (blue circles). Rods also increase in diameter across the periphery (Figure 4, red circles), although not as dramatically as cones. Foveolar cones are the longest and thinnest in the retina, as well as the most tightly packed. Foveolar and foveal cones have a very long axon or fiber of Henle which ends in a synaptic pedicle which is displaced peripherally (Figure 3, diagram). The overall total length of foveal cones is 500–700 μm. Adult foveal cones measured in plastic sections have outer segments 1.2 μm wide and 40–50 μm long while the slightly tapered inner segments are 3 μm at the outer limiting membrane and 2 μm at their tip and 30–35 μm long. Cone length decreases and cone diameter increases rapidly with increasing distance from the foveal center (Figure 4, top row). For instance, 400 μm outside the foveola center, inner segments are 6 μm wide and 24 μm long. This increase in size is correlated with a drop in cone density from 200 000 mm^{-2} at the foveola to 100 000 mm^{-2} just 150 μm away to 15 000 mm^{-2} at 1.5 mm (Figure 4). Despite the apparent high density, the actual number of cones is only 7000–10 000 in the foveola and 90 000 in the central 2 mm of human retina. Therefore, our highest visual acuity depends on the functional integrity of less than 100 000 cones out of 5 million, emphasizing how critical optimal health of the fovea is for good vision throughout life.

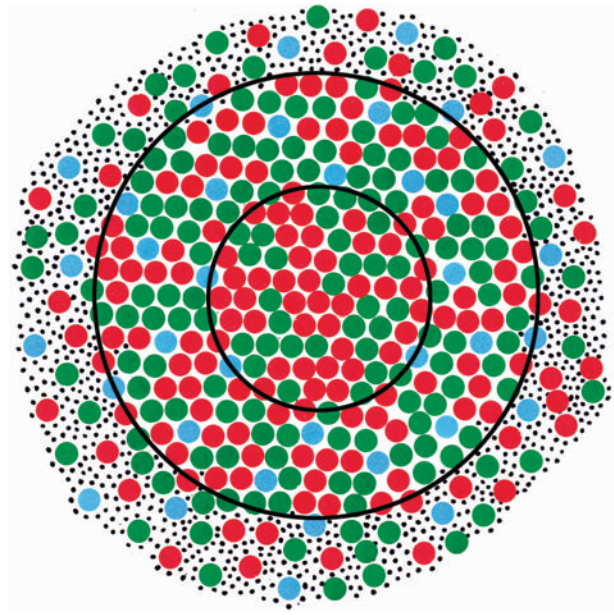

Figure 5 A schematic showing the arrangement of red, green, and blue cones and rods in the central human fovea. The innermost circle is 100 μm wide and depicts the foveolar center which has only red and green cones. The middle circle is 200 μm wide, and here blue cones are added to the red and green cones. Rods (small black dots) appear at 300 μm, and these four photoreceptor types will continue to the retinal edge with rods predominating over peripheral retina.

Foveal Synaptic Circuitry

Stefan Polyak, Brian Boycott, John Dowling, and Helga Kolb used Golgi impregnation for the light and electron microscopic analysis which first revealed the intricate synaptic relationships of primate foveal neurons. These studies concluded that there was a 'private line' synaptic circuit which was called the midget circuit because of the small dendritic trees, cell bodies, and axon terminals found in it. The midget circuit begins with one foveal cone synapsing

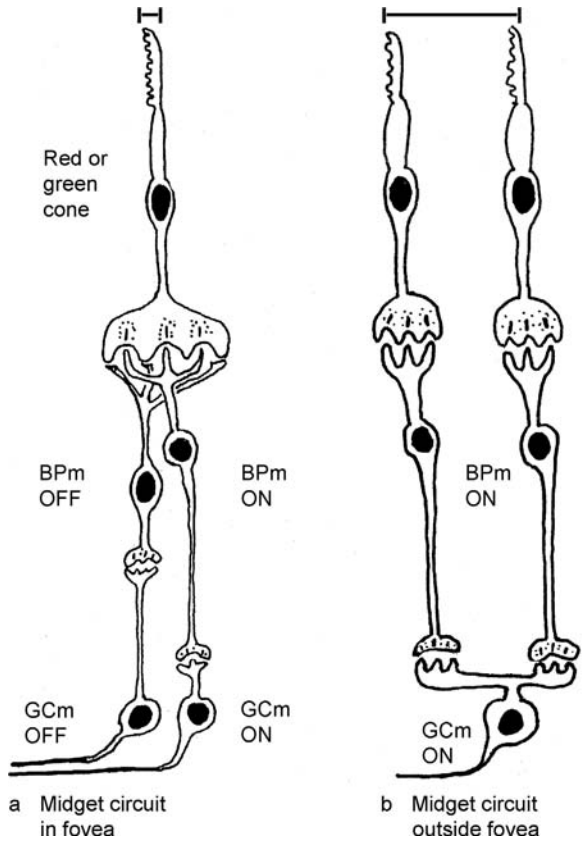

Figure 6 Drawing of the neurons involved in the midget circuit. (a) In the fovea, one red or green cone is presynaptic to two midget bipolar cells (BPm). The OFF is excited when the light goes down and the ON is excited when the light goes up. Each BPm receives only from this single cone. In turn, each BPm provides all excitatory input to a midget ganglion cell (GCm), so a pair of GCm signal all of the activity in the visual space covered by a single cone (line above cone). Because only a single cone drives this circuit, it also will transmit either red or green color information. (b) Outside the fovea, cones are thicker and more widely spaced. One red or green cone still is presynaptic to two BPm (only the ON is shown for clarity here); however, both BPm synapse on the same GCm. This means that the GCm receives visual input from the larger area covered by both cones (line above cones), and visual acuity is decreased. If by chance, both cones are red or green, color coding will be retained. If one cone is green and the other red, this will provide mixed input to the GCm and degrade color vision.

onto two midget bipolar cells (BPm) and is diagrammatically shown in magenta and cyan in Figure 3 and in Figure 6(a). Each BPm receives input only from that single cone. The ON BPm is stimulated when light increases and the OFF BPm when light decreases. In turn, each BPm provides all of the synaptic input to the dendritic tree of a midget ganglion cell (GCm; cyan to yellow (Figure 3); Figure 6(a)), so these GCm respond to the same ON or OFF stimulus. Thus, a single cone signals via two BPm to a pair of GCm which transmit centrally all ON and OFF information about the point in visual space covered by this cone (Figure 6(a), horizontal bar). Because cones are the initial receptor, and foveolar cones are packed at the highest density, high-pixel transmission requires a large number of GCm. About half of the 1.1 million GC in human retina are found within 4.5 mm of the foveal center. In turn, each of the GCm needs a set amount of visual cortex to process its information, reflected in that >40% of the primary visual cortex is devoted to processing information arising within the fovea.

Only red or green cones drive the midget circuit. However, it is not necessary to have both red and green color or trichromatic vision to have high acuity and a fovea. New World male marmoset monkeys have only a single red or green functional gene and are dichromatic, yet they have a normal fovea and acuity equal to female marmosets with two functional genes. The same is true of colorblind human males. This indicates that the fovea evolved for high acuity before the separation of New World and Old World primates 30–40 million years ago. Blue cones have a separate retinal circuit in which two to five blue cones synapse onto a single blue BP and several blue BPs synapse onto a blue GC, so even in the human fovea the blue cone circuit is not designed for high acuity. In trichromatic monkey and human retinas, it appears that red and green color processing 'piggybacks' on the midget circuit. Because each foveal cone contains only red or green opsin, this means that the GCm will carry color as well as high acuity information.

Outside of the fovea, GCm receive input from more than one BPm (Figure 6(b)), and this ratio increases into the periphery. This means that a single peripheral GCm is receiving from multiple cones, and therefore the visual space transmitted centrally by this peripheral GCm is the size of the area covered by all of the cones feeding into that GCm (Figure 6(b)). As you move away from the fovea, this causes visual acuity to decrease roughly proportional to the decrease in cone packing and the increase in bipolar:GC ratio. Color vision also declines into the periphery, presumably because of increased mixing of BPm inputs from red and green cones.

Foveal Development

Adult foveal dimensions, peak cone density, and relative position in the temporal retina remain constant across primates, despite an eightfold range of eye and retinal sizes from the tiny marmoset monkey eye to large ape and human eyes. This strongly suggests that the molecular factors and mechanical forces that create a fovea arose early in primate evolution and are localized to a small region. Given the large

amount of visual thalamus and cortex that already is required for processing foveal visual information, having larger foveas may not be possible without enlarging the brain beyond its cranial capacity.

Foveal development is shown in Figure 7 with colored diagrams of a midget circuit overlaid on actual photographs of fetal and neonatal *Macaca* monkey retina. At midgestation in monkey or human retina, the incipient foveal region lies ~3 mm temporal to the optic disc, but lacks many of its adult characteristics. This is a region >2 mm in diameter which has continuous retinal layers superficially like the adult parafovea (Figure 7(a)). A marked difference is that in the central 1.5 mm, the outer nuclear layer consists of a single layer of photoreceptors which are almost entirely cones (pure-cone region). As early as cones and rods can be identified morphologically or with immunomarkers, in humans only red and green cones are present in the center of the pure-cone region; this is the future foveola. Blue cones appear more peripheral and rods are not present until the edge of the pure-cone region. By contrast, in monkey retina, blue cones are found across the wide pure-cone region, but rods are excluded. Thus, in the fetal primate retina, the same three concentric circles of photoreceptor types found in the adult fovea are present, but each fetal circle is much wider. The molecule(s) determining this pattern are not known.

Foveal neurons undergo remarkable remodeling during development. A key point is that synapses form between cones, BP, and GC at an early stage in primate retina. In *Macaca* fovea, cones have quite mature ribbon synapses by Fd70 and BP/GC ribbon synapses at Fd60 which is more than a month before pit formation begins. In humans, synaptic markers are present by fetal week 11 which is 4–5 weeks before opsin is expressed and 3 months before pit formation begins. This early midget synaptic circuit is illustrated by the vertical column of connected neurons in Figure 7(a).

The first evidence of pit formation is seen around Fd100 in *Macaca* monkeys, Fd135 in New World

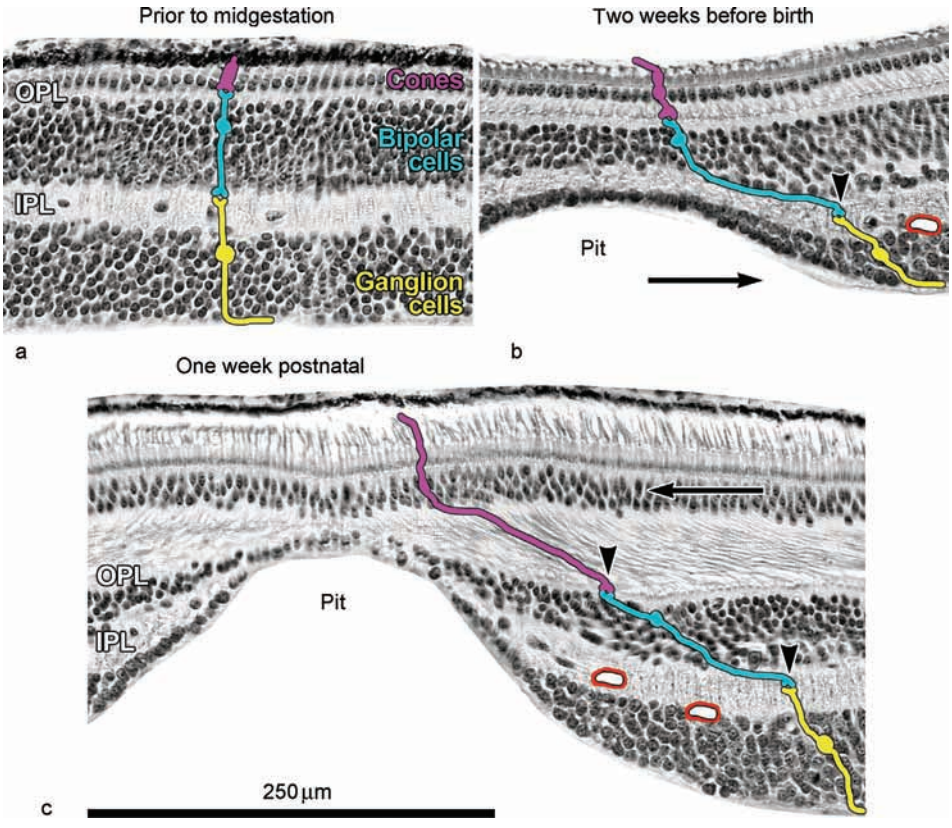

Figure 7 Sections of fetal monkey fovea at Fd 73 (a), Fd 160 (b), and 7 days after birth (c). The midget circuit is drawn on each photograph with the cone in magenta, BP in cyan, and GC in yellow. (a) At midgestation, the circuit is aligned vertically because all layers are present. (b) In later gestation, the pit forms by displacing the GCL first (arrow); note that the GCL is one cell thick over the pit center while the inner nuclear layer is only slightly thinned. As the GC are displaced, each BP has to elongate its axon to keep synaptic contact which is now more peripheral (arrowhead). A blood vessel is drawn in red to indicate the position of the foveal avascular zone. (c) As the pit becomes deeper around birth, cone cell bodies begin to pack into the foveal center (arrow). This requires the cone to elongate its axon to maintain the cone–BP synapse (upper arrowhead), which has been displaced onto the foveal slope during the later stages of pit formation. The adult midget circuit is shown in Figure 3.

marmoset monkeys, and fetal week 24–26 in humans (Figure 7(b)). Pit formation begins with a slight indentation of the GCL in the center of the pure-cone region, and then over time first the GCL and inner plexiform layer, and then the inner nuclear layer neurons, are displaced peripherally (Figure 7(b), arrow). Pit formation takes a considerable time; in humans, an adult-like pit is not present until 11–15 months postnatal, and in monkeys until 2–4 months postnatal. Because foveal neurons have formed synaptic contacts before pit formation, it is necessary to retain these contacts throughout the ensuing complex displacements. Morphological data suggest that in the early stages of pit formation, BPs elongate their axons to retain contact with GCs, and GCs elongate their dendritic stalks to retain contact with BPs as the GCs move peripherally. This produces the offset of inner retinal neurons shown diagrammatically in Figure 7(b) with the BP to GC synapse marked by an arrowhead.

Significantly, at the same age as pit formation begins, blood vessels and their accompanying astrocytes have vascularized central retina, but are excluded from entering the central 400–500 μm of the foveal GCL. This forms a fetal FAZ indicated by the red blood vessel in Figure 7(b). A two-step mechanical model for pit formation has been proposed by Springer and Hendrickson based on a combination of finite element modeling and quantitative morphology. This model requires interaction between the FAZ and two known mechanical factors within the eye, intraocular pressure and retinal stretch caused by eye growth. The first step is that intraocular pressure acts on the more elastic tissue within the FAZ and slowly indents it. This slowly pushes peripherally the neurons, synapses, and glia within the FAZ, first in the GCL, then in the inner plexiform layer, and finally in the inner nuclear layer. Intraocular pressure creates the narrow relatively deep pit seen in all primates around birth (Figures 7(b) and 7(c)). The second step occurs postnatally, and involves a small amount of stretch within the fovea due to postnatal eye growth. This stretch slowly remodels the fovea into the flatter, wider adult pit (Figures 1 and 3).

Well after pit formation is underway, cones and rods across central retina begin to pack centrally into the fovea (Figure 7(c), arrow). Cone packing is marked by (1) the pure-cone region slowly decreasing in diameter so that over time rods lie within 300 μm of the foveola; (2) foveal cones becoming much longer and thinner, enabling tighter packing of their inner segments and cell bodies; (3) a marked elongation of fibers of Henle on cone and central rods; and (4) a steady rise in foveal cone density. For instance, in humans, cone density rises from 19 000 mm^{-2} at birth to 41 000 mm^{-2} around 1 year to 113 000 mm^{-2} at 4 years, which is in the lower range of adults. Because no new cones are produced during this period, and the number of the cones in the pure-cone region does not change, this density increase is due to cone remodeling which allows packing of a pre-existing population into a smaller space. Elongated thin cones with long fibers of Henle appear as packing commences (Figure 7(c)). The fiber of Henle is formed because foveal cones must retain their synaptic connections with BP cells, and the timing of its appearance suggests that Henle fiber elongation is mainly due to foveolar, foveal, and parafoveal cone cell bodies slowly being displaced centrally. Additional elongation is stimulated by cone synaptic pedicles being displaced more peripherally to where the BP now reside (Figure 7(c), upper arrowhead). The mechanism of cone packing is debatable, but one factor may be the same retinal stretch which flattens the foveal pit. Finite element analysis modeling shows that stretching the pit creates centralward lift forces around the pit which could help pack cones. This is supported by the observation that temporally both the morphological changes in cones and increased cone density strongly correlate with pit flattening in both New and Old World monkeys.

See also: Cone Photopigment Evolution; Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Pharmacology: Inner Retinal Layers; Retinal Color Mechanisms; Retinal Development: An Overview.

Further Reading

Bumsted K, Jasoni CL, Szél A, and Hendrickson A (1997) Spatial and temporal expression of cone opsins during monkey retinal development. *The Journal of Comparative Neurology* 378: 117–134.

Curcio CA and Hendrickson AE (1991) Organization and development of the primate photoreceptor mosaic. *Progress in Retinal Research* 10: 89–120.

Curcio CA, Sloan KA, Kalina RE, and Hendrickson AE (1990) Human photoreceptor topography. *The Journal of Comparative Neurology* 292: 497–523.

Hendrickson A (1992) A morphological comparison of foveal development in man and monkey. *Eye* 6: 136–144.

Hendrickson A (1996) Synaptic development in Macaque monkey retina and its implications for other developmental sequences. *Perspectives on Developmental Neurobiology* 3: 195–203.

Hendrickson A (2005) Organization of the adult primate fovea. In: Penfold PL and Provis JM (eds.) *Macular Degeneration*. Heidelberg: Springer.

Hogan MJ, Alvarado JA, and Weddell JE (1971) *Histology of the Human Eye*. Philadelphia: W.B. Saunders.

Packer O, Hendrickson AE, and Curcio CA (1989) Photoreceptor topography of the retina in the adult pigtail macaque (*Macaca nemestrina*). *The Journal of Comparative Neurology* 288: 165–183.

Provis JM, Diaz CM, and Dreher B (1998) Ontogeny of the primate fovea: A central issue in retinal development. *Progress in Neurobiology* 54: 549–580.

Provis JM, Sandercoe T, and Hendrickson AE (2000) Astrocytes and blood vessels define the foveal rim during primate retinal development. *Investigative Ophthalmology & Visual Science* 41: 2827–2836.

Roorda A and Williams DR (1999) The arrangement of the three cone classes in the living human eye. *Nature* 397: 520–522.

Snodderly DM, Weinhaus RS, and Choi JC (1992) Neural–vascular relationships in central retina of macaque monkeys (*Macaca fascicularis*). *The Journal of Neuroscience* 12: 1169–1193.

Springer A and Hendrickson A (2005) Development of the primate area of high acuity. 3: Temporal relationships between pit formation, retinal elongation and cone packing. *Visual Neuroscience* 22: 171–185.

Wässle H and Boycott BB (1991) Functional architecture of the mammalian retina. *Physiological Reviews* 71: 447–480.

Wässle H, Grünert U, Röhrenbeck J, and Boycott BB (1990) Retinal ganglion cell density and cortical magnification factor in the primate. *Vision Research* 30: 1897–1911.

Yuodelis C and Hendrickson A (1986) A qualitative and quantitative analysis of the human fovea during development. *Vision Research* 26: 847–855.

Fragile X Associated Tremor Ataxia Syndrome

H E Bronsky, S Coffey, and R J Hagerman,
University of California at Davis Medical Center,
Sacramento, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Fragile X-associated tremor/ataxia syndrome (FXTAS) is a newly discovered neurodegenerative disorder that typically affects men in their fifth or sixth decade. It is characterized by intention tremor, unstable gait, cognitive decline, neuropathy and psychiatric features, along with brain atrophy, white matter disease, and intranuclear inclusions (abnormal protein clusters) in neurons and astrocytes throughout the brain. FXTAS is caused by the premutation of the fragile X mental retardation (*FMR1*) gene and, along with fragile X syndrome (FXS) and premature ovarian failure (POF), is one of three separate disorders that are now associated with this gene. POF, a disorder characterized by the cessation of menses in women younger than 40 years of age, is also caused by a premutation of the *FMR1* gene. FXS is the window by which FXTAS and fragile X-related POF were discovered. FXS is a childhood-onset neurodevelopmental disorder caused by a full mutation of the *FMR1* gene. These disorders are all associated with a mutation on the *FMR1* gene, which is located on the bottom part of the X chromosome. The mutation is characterized by a CGG (trinucleotide repeat) expansion in the front part of the gene that is not translated into protein.

Normal CGG repeat size in the *FMR1* gene is between 5 and 44. The mutation falls into three categories: gray zone, premutation, and full mutation. Repeat size between 45 and 54 CGG repeats is known as the 'gray zone' or 'intermediate' size range. There are no known clinical associations with individuals that carry the gray zone allele; however, this CGG repeat size may be unstable when it passes to the next generation and this instability is difficult to predict based on repeat size alone. The premutation allele is defined as a repeat size of 55–200. The premutation allele is associated with POF and FXTAS, as well as various medical, neurological, psychological, and cognitive deficits including autistic spectrum disorders (ASD) and attention deficit hyperactivity disorders (ADHD). The full mutation allele, characterized as having a repeat number of greater than 200, is associated with FXS. FXS has a broad array of social, cognitive, physical, and psychological abnormalities that present in the affected individual with a wide degree of variability. In order to achieve a better understanding of FXTAS, a discussion of the associated disorder, FXS, as well as clinical implication of carrying the premutation allele, is necessary.

Fragile X Syndrome

FXS is the most common form of inherited mental retardation and one of the known causes of autism and autistic-like characteristics. The prevalence of the full mutation is 1 in 3600 in males and 1 in 6000 in females. It is thus named because in the event of a full mutation, the bottom part of the long arm of the X chromosome where the *FMR1* gene is located (Xq27.3) appears fragile as if it is falling off the chromosome. In the full mutation, the gene becomes methylated at the CpG island (the first part of the CGG repeat) causing transcriptional silencing of the gene and the fragile X mental retardation protein (FMRP) is not produced. FMRP is an RNA-binding protein that is thought to regulate translation at the dendrite in response to glutamatergic neural activation. FMRP is therefore a modulator of synaptic plasticity and dendritic morphology. The absence or deficiency of FMRP results in FXS.

From the first identification of FXS, developmental disabilities and mental impairment have been the hallmarks of the syndrome in both affected males and females. Males with the full mutation may exhibit distinctive physical characteristics such as large and/or protruding ears, high arched palate with rare cleft palate, prominent forehead, macrocephaly (where head circumference is larger than average), and mandibular prognathism (where the lower jaw is more protruded than the upper jaw) (**Figure 1**). Connective tissue abnormalities are common in FXS and include hyperflexible joints, soft velvety skin, flat feet, and mitral valve prolapse (a heart valve abnormality). Genital abnormalities consist of macroorchidism (abnormally enlarged testes). The cognitive effects of the full mutation include developmental delays, learning disabilities, and mental retardation ranging from mild to severe. Behavioral effects include ADHD, hand biting, anxiety, sensory integration disorders, and language delays. Autism occurs in approximately 30% of males and in approximately 10% of females with FXS.

Females with the full mutation typically have milder deficits due to the second X chromosome, which presumably makes FMRP. FMRP production results from only the normal X chromosome, and each cell randomly selects which *FMR1* gene to silence. Individuals with a greater proportion of silenced mutated genes

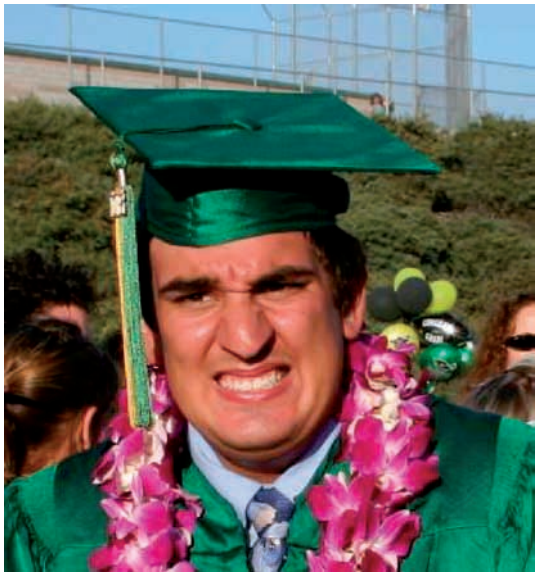

Figure 1 An adult male with FXS.

will have less of the characteristics of FXS and females with a greater percentage of their mutated *FMR1* alleles remaining active will exhibit more of the characteristics of FXS. Females with a full mutation and a greater proportion of silenced mutated genes will have a normal IQ, whereas females with an unfavorable activation ratio (below 0.5) will have learning disabilities or mental retardation because of an absence or deficiency of FMRP, similar to males with FXS. Typically, intellectual impairment is milder in females than affected males. It is more common for women to have normal IQs; however, most females with a normal IQ have learning disabilities, particularly math difficulties, or emotional problems such as depression, obsessive thinking, social anxiety, impulsivity, or poor eye contact. Approximately 30% of females with the full mutation exhibit the distinctive physical features of prominent ears and hyperextensible finger joints.

Premutation Carriers

The prevalence of the premutation in the general population is approximately 1 in 259 females and 1 in 813 males. Before the discovery of the *FMR1* gene, premutation carriers were thought to be unaffected. For several years after the discovery of the *FMR1* gene, the clinical involvement in the premutation carrier was unknown. POF was identified as a specific phenotype associated with the premutation after discussions at a 1987 FXS conference between families, patients, and clinicians about early menopause. Research studies have estimated that POF affects approximately 15–20% of female premutation carriers (compared to 1.7% of noncarriers). POF is not seen in females with the full mutation, suggesting that specific molecular abnormalities in the premutation are not present in the full mutation.

Within the past few years, there has been a significant change in understanding the clinical and molecular consequences of being a carrier of the premutation. Prior to this, behavioral and cognitive testing were not routinely done. Carriers, in fact, may have mild physical features of FXS, as well as cognitive and behavioral abnormalities including social deficits, attention problems, and ASD. When seen clinically for behavior problems, 73% of males with the premutation who were probands were diagnosed with either ASD or ADHD. Nonproband males with the premutation who were identified by cascade testing (routine testing among family members after a diagnosis is made) are still at a significantly greater risk for ASD (8%) or social deficits but not nearly as high as that for the probands. This rate is also higher than normal for brothers without the premutation. The premutation allele is also associated with an increased risk of obsessive–compulsive behavior, particularly in males, and with an increased prevalence of obsessive–compulsive behavior. The clinical involvement now apparent in carriers of the premutation has clear molecular underpinnings.

It was the discovery of elevated *FMR1* mRNA in males and females with the premutation that confirmed a unique molecular phenotype in carriers. POF and FXTAS, the disorders of premutation carriers, are not associated with FMRP deficit as in FXS but instead with an increased level of *FMR1* mRNA. Individuals with the full mutation do not have elevated levels of mRNA since the gene is silenced, and they do not develop POF or FXTAS later in life. It has been found that *FMR1* mRNA is increased two- to eightfold above normal when compared to that in controls or individuals with the normal *FMR1* gene in peripheral blood leukocytes (white blood cells) in premutation carriers. This is thought to result in mRNA toxicity, whereby proteins bind to the expanded CGG repeat in the mRNA and are sequestered into inclusions seen in the nuclei of neurons and astrocytes. It has been found that as repeat size increases, *FMR1* mRNA levels increase despite mildly decreased levels of FMRP.

Fragile X-Associated Tremor/Ataxia Syndrome

As discussed previously, FXTAS and FXS are mechanistically distinct syndromes. The connection between FXS and its associated disorder, FXTAS, was not

made until 2001. Mother's of children with FXS provided the missing link between FXS, a childhood neurodevelopmental syndrome of carriers of the full mutation, and FXTAS, an adult-onset disorder of carriers of the premutation. Often, mothers would express concern for their fathers, who were carriers of the premutation, whereas their children were being seen for FXS-related problems. They would report that their fathers were experiencing hand tremor, resulting in loss of daily living skills, as well as gait problems, often causing serious falls. These males were then seen clinically. It is now known that approximately one-third of male carriers will develop FXTAS, with varying degrees of symptom progression. Approximately 1 in 800 males carry the premutation in the general population, which means that approximately 1 in 3000 men run the risk of developing FXTAS.

FXTAS is occasionally seen in female carriers but less frequently than in the male carriers. Rare females have been documented to have the clinical and radiological features of FXTAS. The first five cases of FXTAS in females demonstrated clinical involvement similar to males with FXTAS, including intention tremor, ataxia, and peripheral neuropathy. However, none of these patients exhibited the cognitive decline seen in approximately 20% or more of males with FXTAS.

The diagnostic criteria for FXTAS include major and minor clinical and radiological findings, in addition to a CGG repeat in the range of 55–200. Based on the patient's findings, he or she may be diagnosed with 'definite,' 'probable,' or 'possible' FXTAS. Because FXTAS is a progressive neurodegenerative disease, six stages further describe how the patient's activities of daily living are impacted. Stage 1 is a stage of subtle or questionable tremor or ataxia, whereas the patient later relies on a cane or walker in stage 4, uses a wheelchair on a daily basis in stage 5, and finally becomes bedridden in stage 6. A description of the clinical, radiological, and pathological features of FXTAS follows.

Clinical Features

In the early stages of the disease, a patient may present with one clinical symptom (usually intention tremor in the dominant hand). Then, as the disease progresses, further clinical symptoms typical of FXTAS, plus a radiological sign, can determine a definitive diagnosis. Clinical symptoms, and major clinical diagnostic criteria, include gait ataxia (balance problems when walking), intention tremor (tremor when trying to do something as opposed to the resting tremor associated with Parkinson's disease), parkinsonism (symptoms associated with Parkinson's disease – that is masked facies, resting tremor, etc.), peripheral neuropathy (loss of vibration sense and reflexes particularly in the lower extremities), muscle weakness, and cognitive decline that usually begins with memory and executive function deficits. Patients may also have bowel/bladder incontinence, impotence, hypertension, and orthostatic hypotension. Patients with FXTAS have been misdiagnosed with several different disorders that touch on the clinical involvements (i.e., atypical Parkinson's, peripheral neuropathy, essential tremor, multiple system atrophy, and cerebellar ataxia). If a patient presents with symptoms resembling those of FXTAS, it would be helpful for the examining physician to ask targeted questions about children or grandchildren with mental retardation, autism, or behavioral/learning disorders; daughters or female relatives with infertility or POF; and family members with ataxia, neurological problems, and psychiatric problems.

Patients typically develop a frontal–subcortical dementia in addition to the motor problems of tremor and ataxia. Typically, patients with FXTAS will present with memory loss, executive function deficits, and some anxiety and/or depression. Executive function deficits include disinhibition and working memory difficulties, whereas verbal abilities are relatively well preserved. The most striking deficits are impairment in the initiation of purposeful activity and the disinhibition of inappropriate or irrelevant behavior. Investigation of the cognitive decline in FXTAS is newly described and warrants further study due to the considerable variability in its progression and nature.

Radiological Features

Males with the premutation have been found to have a decrease in total brain volume, including the cerebrum, cerebellum, and cerebral cortex, on magnetic resonance imaging (MRI). The total brain and cerebrum volume is significantly related to CGG repeat size; that is, the higher the CGG repeat number, the lower the brain volume.

There is white matter disease, shown by increased T2 signaling on MRI in males with FXTAS (Figure 2). There is also spongiosis (increased intercellular fluid in the brain) in the middle cerebellar peduncles (MCP) and an increase in ventricular size, resulting from brain atrophy. The MCP sign is composed of white matter tracts coursing from the pons to the cerebellum, and this area of the brain has demonstrated white matter disease in a characteristic way in approximately 60% of individuals with FXTAS. For this reason, the MCP sign is a major radiological diagnostic criterion.

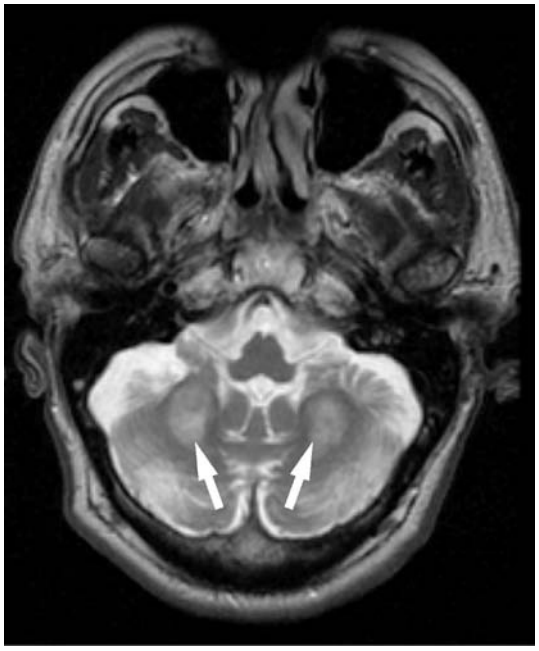

Figure 2 An MRI image of the neurodegeneraton in FXTAS. Note the increased T2 signaling in the MCP (arrows).

Pathological Features

Purkinje cell loss is seen in the cerebellum of patients with FXTAS. Purkinje cells control the output from the cerebellum and are among the largest neurons in the brain.

In addition, the presence of ubiquitin-positive eosinophilic intranuclear inclusions is part of the characteristic neuropathology of FXTAS. The inclusions are found in the astrocytes (a subtype of glial cell) of cerebral white matter, cerebral cortex, cerebellum, and in large numbers in the hippocampus. Inclusions are also found in the autonomic neurons of the spinal chord, which may account for the autonomic dysfunction that often occurs. There is a significant correlation between CGG repeat number, number of inclusions in cortical gray matter, and CA1 pyramidal neurons (the major excitatory neurons of the cortex) in the hippocampus. There is also a correlation between inclusion-bearing astrocytes in cortical white matter and CGG repeat number. *FMR1* mRNA has been identified within the inclusions, supporting the RNA toxic gain-of-function mechanism. This is also supported by the tendency for symptom severity to increase as mRNA level increases. A variety of proteins are sequestered inside the inclusions, including myelin basic protein, lamin A/C, and HnRNPA2. The dysregulation of these proteins is likely the cause of neurodegeneration found in FXTAS.

The mRNA gain-of-function theory in FXTAS is based on the RNA gain-of-function model for myotonic dystrophy. Either expansion of the CUG repeat on the dystrophia myotonica-protein kinase (DMPK) mRNA or the expansion of the CCUG repeat on the *ZNF9* gene lead to inclusions with C(C)UG-binding proteins. Placing an expanded CTG repeat in transgenic mice leads to myotonia. 'Knockin' mice with a CGG repeat of approximately 100 placed in the *FMR1* gene develop ubiquitin-positive intranuclear inclusions in neurons that are similar to what is seen in FXTAS. These mice with the premutation also develop neurological disease with aging.

Conclusions

FXTAS is a newly identified neurodegenerative disorder. Patients present with tremor and ataxia, but neuropathy, autonomic dysfunction, psychiatric problems, and cognitive deficits are common. Because this is a new diagnosis that requires DNA study to document the premutation expansion of the *FMR1* gene, many individuals with FXTAS are misdiagnosed with another disorder. Given the prevalence of FMR1 premutations in the general population, screening of all individuals with tremor and/or ataxia older than 50 years with the FMR1 DNA test is recommended. A positive family history of POF or ovarian dysfunction, mental retardation, or autism will increase the likelihood of a positive test. When the premutation is identified, referral to a genetic counselor or geneticist is recommended for further workup of the family pedigree to identify family members with the premutation or full mutation or family members at risk of having children with FXS. Supportive treatment can also be provided with medications currently available to treat some of the symptoms associated with FXTAS.

FXTAS is one of the most common progressive disorders associated with a single gene defect. It is currently underdiagnosed because it has only recently been recognized. Research toward understanding FXTAS is of far-reaching utility in understanding other inclusion-associated neurodegenerative disorders.

See also: Autism; Dementia; Developmental Disability and Fragile X Syndrome: Clinical Overview; Friedreich Ataxia; Triplicate Repeats: Huntington's disease.

Further Reading

Berry-Kravis E, Lewin F, Wuu J, et al. (2003) Tremor and ataxia in fragile X premutation carriers: Blinded videotape study. *Annals of Neurology* 53: 616–623.

Brunberg JA, Jacquemont S, Hagerman RJ, et al. (2002) Fragile X premutation carriers: Characteristic MR imaging findings in adult males with progressive cerebellar and cognitive dysfunction. *American Journal of Neuroradiology* 23: 1757–1766.

Cornish KM, Kogan C, Turk J, et al. (2005) The emerging fragile X premutation phenotype: Evidence from the domain of social cognition. *Brain and Cognition* 57: 53–60.

Greco CM, Berman RF, Martin RM, et al. (2006) Neuropathology of fragile X-associated tremor/ataxia syndrome (FXTAS). *Brain* 129: 243–255.

Grigsby J, Brega AG, Jacquemont S, et al. (2006) Impairment in the cognitive functioning of men with fragile X-associated tremor/ataxia syndrome (FXTAS). *Journal of Neurological Sciences* 248: 227–233.

Hagerman PJ and Hagerman RJ (2004) Fragile X-associated tremor/ataxia syndrome (FXTAS). *Mental Retardation and Developmental Disabilities Research Reviews* 10: 25–30.

Hagerman RJ (2006) Lessons from fragile X regarding neurobiology, autism, and neurodegeneration. *Journal of Developmental and Behavioral Pediatrics* 27: 63–74.

Hagerman RJ and Hagerman PJ (2002) *Fragile X Syndrome: Diagnosis, Treatment, and Research,* 3rd edn. Baltimore, MD: Johns Hopkins University Press.

Hall DA, Berry-Kravis E, Jacquemont S, et al. (2005) Initial diagnoses given to persons with the fragile X associated tremor/ataxia syndrome (FXTAS). *Neurology* 65: 299–301.

Hessl D, Tassone F, Loesch DZ, et al. (2005) Abnormal elevation of FMR1 mRNA is associated with psychological symptoms in individuals with the fragile X premutation. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* 139: 115–121.

Iwahashi CK, Yasui DH, An HJ, et al. (2006) Protein composition of the intranuclear inclusions of FXTAS. *Brain* 129: 256–271.

Jacquemont S, Hagerman RJ, Leehey M, et al. (2003) Fragile X premutation tremor/ataxia syndrome: Molecular, clinical, and neuroimaging correlates. *American Journal of Human Genetics* 72: 869–878.

Leehey MA, Hagerman RJ, Landau WM, et al. (2002) Tremor/ataxia syndrome in fragile X carrier males. *Movement Disorders* 17: 744–745.

Van Esch H, Dom R, Bex D, et al. (2005) Screening for FMR-1 premutations in 122 older Flemish males presenting with ataxia. *European Journal of Human Genetics* 13: 121–123.

Relevant Website

http://www.fragilex.org – National Fragile X Foundation.

Fragile X Syndrome

A W Grossman and W T Greenough, University of Illinois at Urbana–Champaign, Urbana, IL, USA
E M Berry-Kravis, Rush University Medical Center, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Fragile X syndrome (FXS) represents the most common form of inherited mental retardation, and is associated with a highly variable cognitive and behavioral phenotype as well as a characteristic profile of physical features. FXS is an X-linked disorder that affects approximately 1 in 4000 males and females. Previously termed Martin–Bell syndrome, this disorder was named fragile X syndrome because of an apparent fragile site observed at Xq27.3 on the long arm of the X chromosome when cells from affected individuals were cultured in folate-deficient media. It is now known that FXS is caused by a trinucleotide repeat expansion in the *FMR1* gene, which is located at and accounts for the fragile site. This mutation leads to hypermethylation of the promoter region and transcriptional silencing of the gene. Typically, no functional *FMR1* mRNA is produced, and the fragile X mental retardation protein (FMRP) is also absent. In the past decade, significant research advances have been made toward delineating the role of FMRP in normal brain function and the effects of its absence on neuroanatomy and physiology.

Symptoms of FXS

Individuals with FXS exhibit, to variable degree, a characteristic physical profile including a long, narrow face with large, protruding ears and macroorchidism (enlarged testes). Other physical features may include macrocephaly, hyperextensible joints, loose skin over the hands, a high arched palate, a vertical plantar crease, and flat feet.

The predominant and most disabling problem in FXS is mental retardation, with the average IQ in males around 40 and the average IQ in females with the full mutation around 80. IQ varies among individuals, however, ranging from severe functional impairment to normal intelligence, and appears to correlate positively with expression levels of FMRP. Nearly all males with intelligence in the borderline or normal range are genetic mosaics in which methylation of the *FMR1* gene is incomplete and FMRP is expressed, albeit at reduced levels. In females, the distribution is shifted toward higher levels of functioning, compared to males, and intellectual functioning has been correlated with X inactivation ratios (i.e., women and girls with higher percentages of cells with the normal X chromosome active tend to be higher functioning).

When compared with IQ-matched children with other forms of developmental delay (e.g., Down syndrome), children with FXS demonstrate a specific profile of cognitive strengths and weaknesses. Though the available data are not entirely consistent, many of these tests reveal relative strengths in receptive language skills, visual memory, acquisition of factual information, imitation skills, and gestalt processing. Relative weaknesses are often observed in expressive language skills as well as spatial and sequential processing, quantitative and mathematical skills, and executive function. Individuals with FXS often exhibit perseverative speech including imitation of expressions, swearing, or automatic verbalizations; these patterns can be exacerbated by social anxiety. These cognitive profiles are more easily observed in higher-functioning, more testable individuals.

Behavioral and social dysfunction in children with FXS tends to be more severe than would be expected based on cognitive levels. Approximately 90% of boys exhibit hyperactivity, and virtually all have attentional difficulties. Girls frequently have attention problems and may occasionally manifest hyperactivity. Males and females with FXS demonstrate mood lability, compulsive and perseverative behaviors, as well as stereotyped movements such as hand flapping and hand biting. Aggressive outbursts are observed in about 30% of boys with FXS and continue through young adulthood. Many of these behavioral problems appear to be related to an elevated level of anxiety, which also results in social shyness and poor eye contact. Studies of physiologic arousal suggest that sympathetic activity is increased, parasympathetic activity is decreased, and that these systems are poorly coordinated in FXS. Panic attacks are common in individuals with FXS and are often triggered by novelty (e.g., novel social situations), undesired or excessive sensory stimuli, or when the individual is placed 'on the spot' and made the center of attention. A related phenotype in FXS may be an exaggerated hormonal response to behavioral stress. Individuals with FXS also display hyperresponsivity to a variety of sensory stimuli, including auditory stimuli (expressed as an elevated startle response), olfactory stimuli (perceiving neutral scents as noxious), and somatosensory stimuli (expressed as tactile defensiveness).

Autistic symptoms are common in males with FXS (but relatively rare in females with FXS). These

symptoms include poor eye contact, difficulties with peer interactions, unusual nonverbal interactions, and stereotyped behaviors such as hand flapping. Indeed, approximately 30% of these individuals meet the diagnostic criteria for autism spectrum disorders, although they have a pattern of more severe deficit in communication and stereotyped behavioral domains than in reciprocal social interactions. Social anxiety as opposed to social disinterest seems to drive autistic behavioral patterns in FXS individuals. Males with FXS and autism tend to exhibit more severe developmental delays than do those without autism.

Beyond the cognitive and behavioral phenotype, individuals with FXS often exhibit a number of medical problems that may require monitoring and management. Perhaps the most problematic for families of children with FXS is seizures, which has been estimated to occur in 15–20% of male children with FXS and a smaller percentage of females. In most individuals, seizures are limited to childhood, although in up to 25% of cases, seizures can continue into adulthood. Other medical problems seen more often in FXS than in the general population include gastroesophageal reflux and failure to thrive in early infancy, hypotonia with subsequent poor coordination, recurrent otitis media and sinusitis, vision problems including strabismus and farsightedness, cardiac valve prolapse (especially the mitral valve), sleep disorders such as obstructive sleep apnea, orthopedic issues such as joint laxity, and dental malocclusions. Many of these problems have been described in boys with FXS and are thought to be due to mild dysfunction of connective tissue. Girls with FXS may also exhibit these medical problems, although their presence is much more variable.

Current Treatment Strategies for FXS

Currently available treatments for FXS are supportive and are aimed at management of specific symptoms rather than the underlying disorder. Medical problems outlined in the preceding section should be managed aggressively, with placement of myringotomy tubes for chronic ear infections, audiological monitoring, vision monitoring and use of glasses and strabismus surgery as appropriate, tonsilectomy/adenoidectomy for sleep apnea or recurrent infections, and prophylactic antibiotics for dental procedures and surgery if mitral valve prolapse is present. If seizures occur, these should be treated (after appearance of a second seizure episode) with anticonvulsants that do not aggravate behavioral and social dysfunction – carbamazepine, oxcarbazine, and valproic acid can give good control. Phenobarbital should be avoided because of the tendency to aggravate hyperactive and aggressive behavior. Phenytoin also has undesired side effects such as gingival hyperplasia, and should be avoided.

Medications for behavior may be quite helpful for symptomatic management of behavior problems in FXS. Stimulants (methylphenidate and amphetamine salt preparations) have been shown to be useful, both in clinic settings and in a small placebo-controlled trial, for as many as 70% of boys with FXS for management of distractibility and hyperactivity. Clonidine can be helpful for hyperactivity and hypersensitivity, especially in young boys. Selective serotonin reuptake inhibitors (SSRIs; e.g., fluoxetine, sertraline, citalopram) are effective in managing anxiety and mood problems in FXS, and aggression, outbursts, irritable behavior, and severe oppositionality often respond well to atypical antipsychotics (such as risperidone and aripiprazole). Aripiprazole has a unique pharmacological profile in the central nervous system (CNS) which includes partial dopamine agonist activity and thus seems to be particularly useful in FXS because it targets multiple categories of behavior, including hyperactivity/distractibility as well as irritability/mood, with high clinical response rates. Good clinical success with aripiprazole in FXS has prompted a formal clinical trial, currently in progress.

Treatment of individuals with FXS also includes supportive strategies to assist with development of adaptive skills (early intervention, physical therapy, orthotics for foot positioning and gait training, occupational therapy, and speech therapy). Educational programming should be maximized with a curriculum and environment tailored to the cognitive profile in FXS. Psychological support and social skills training should be available for the individual with FXS. Family therapy can be helpful for coping and comprehensive genetic counseling is crucial for families of individuals with FXS.

Neuropathology in FXS

Although postmortem studies of gross brain morphology have not revealed striking abnormalities in FXS brain, neuroimaging methods have provided researchers excellent tools to investigate brain abnormalities associated with FXS. Structural magnetic resonance imaging (MRI) has revealed a reduction in the size of the posterior cerebellar vermis that may result in the observed enlargement of the fourth ventricle in FXS males. The volume of hippocampus and caudate nucleus appears to be increased in a cohort of males with FXS, and lateral ventricles appear to be slightly enlarged. More recently, diffusion tensor imaging has been used to identify white matter tract alterations in frontostriatal and parietal sensorimotor pathways.

The neuroanatomical phenotype of FXS has also been described on the microscopic level. Early studies of Golgi-stained tissue revealed an abundance of long, thin, tortuous spines (postsynaptic elements of excitatory synapses) along pyramidal cell dendrites in the cerebral cortex of postmortem tissue from adult FXS males. Later quantitative analyses of temporal and occipital cortex revealed that spine length and spine density were elevated in males with FXS compared with controls, and that the adults with FXS exhibited a profile of spine shapes (morphologies) comparable to the young or developing controls. Findings of abnormal spine morphology have been reported in other disorders associated with mental retardation and are not unique to FXS. Yet, the nature of the spine 'dysgenesis' observed in the cerebral cortex of patients with FXS is distinct from other disorders associated with mental retardation, in that adult FXS patients exhibit an elevation in spine density along with longer spines and a morphology profile that resembles an immature state. In the context of changes that typically occur during synaptic development (e.g., synapse overproduction and pruning during development), it has been hypothesized that the spine phenotype in FXS reflects a deficit of synaptic plasticity.

Molecular Genetics of FXS

In 1991, researchers discovered that the genetic basis for FXS is the expansion during early development of an unstable CGG repeat in the 5′ untranslated region of the *FMR1* gene, found at locus q27.3 on the X chromosome. Most repeat lengths of fewer than 45 triplets are stable and considered normal. A repeat length of between 55 and 200 triplets represents the FXS carrier state or 'premutation.' Expansion of this repeat beyond approximately 200 triplets occurs only when the gene is passed on through a female and leads to hypermethylation of the promoter region, and silencing transcription of *FMR1* mRNA, with resultant absence of FMRP and the development of FXS (see **Figure 1**).

Knowledge of the basis of repeat expansion in the transmission of fragile X remains incomplete. Borderline repeat lengths of 45–54 triplets may be stable or unstable, and if unstable, can be passed on to future generations as larger alleles. These alleles have been called 'gray zone' alleles; they are not known to cause disease and do not expand directly to a fragile X full mutation or result in FXS in a single generation. The smallest premutation allele known to expand directly to a full mutation in a single generation contains 59 repeats. Repeat sizes between 150 and 250 are often only partially methylated and result in varying levels of FMRP deficiency and milder FXS phenotypes. Unmethylated alleles in the upper premutation and full mutation length range are transcribed but result in reduced FMRP due to impaired translation in the presence of the long CGG repeat sequences. These mutations typically result in milder FXS phenotypes with symptoms such as attention-deficit/hyperactivity disorder (ADHD), anxiety, social dysfunction, and learning disability rather than overt mental retardation.

Most premutation carriers do not exhibit hypermethylation of the *FMR1* gene and make near normal amounts of FMRP and therefore do not exhibit FXS. The mechanism that underlies hypermethylation and promoter inactivation in alleles with repeats greater than 200 triplets in length remains unknown but is an area of active investigation with therapeutic implications. In males, variability in *FMR1* mRNA and FMRP levels may be due in part to mosaicism of repeat length and methylation (e.g., subsets of cells in the brain or elsewhere in the body with the premutation and other subsets with the full mutation) or to methylation mosaicism alone (as described previously). In females, variability is likely due to differential X chromosome inactivation. These sources of genetic variability reflect the complexity of FXS inheritance and contribute to the remarkably high variability in the phenotype of FXS.

FMRP is a 69 kDa protein found in most adult and fetal tissues (with other splice variants at slightly different molecular masses), but it is present in particularly high levels in the brain and testes. In neurons, FMRP is produced in the soma and also near synapses in response to specific neurotransmitter receptor activation. FMRP expression in oligodendrocytes (glial cells that give rise to myelin) appears to be highest early in development, and FMRP expression in astrocytes (glial cells that serve molecular needs of neurons and synapses) appears to be minimal. FMRP expression in the brain seems to be activity or experience dependent, as protein levels are increased in activated brain regions in response to behavioral demands such as exposure to complex environment or motor skill training, whisker stimulation, and light exposure in dark-reared animals.

Although the specific cellular functions of FMRP are unknown, its role as a putative mRNA-binding protein has been most thoroughly investigated. As shown in **Figure 1**, the protein sequence of FMRP contains several mRNA-binding motifs (two KH domains, an RGG box, and an amino terminus with a strong affinity for RNA), a nuclear localization signal (NLS), and a nuclear export signal (NES). FMRP has been shown to bind to multiple mRNAs and associates with actively translating ribosomes in

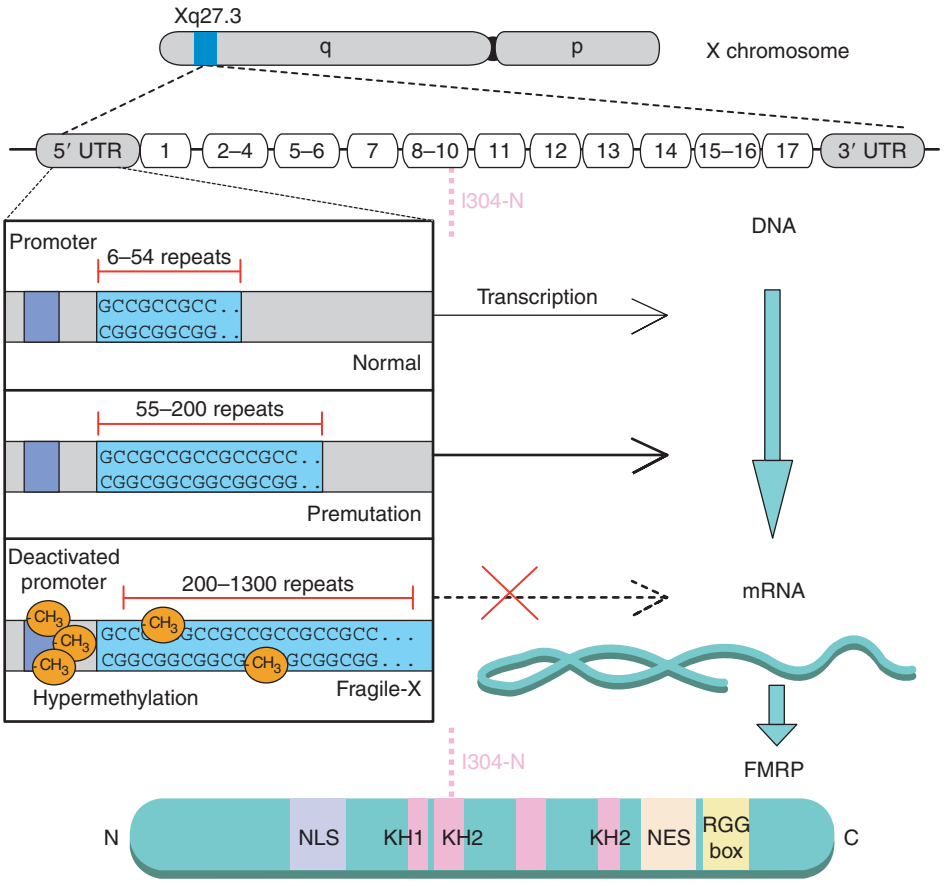

Figure 1 The gene *FMR1* is located at Xq27.3 on the long arm of the X chromosome. A CGG repeat sequence in the 5′ untranslated region (UTR) is normally less than 55 repeats in length. Transcribed from this gene is *FMR1* mRNA, which codes for FMRP. Expansion of this sequence to 55–200 repeats represents the unstable 'premutation' and levels of *FMR1* mRNA are elevated. Beyond 200 repeats, the sequence becomes hypermethylated, inactivating the promoter and silencing transcription. The resultant absence of FMRP characterizes the fragile X syndrome. FMRP contains nuclear localization and nuclear export sequences (NLS, NES) as well as several RNA-binding domains (KH1, KH2, RGG box). A missense mutation (I304N) in one of these domains (KH2) is associated with the full fragile X phenotype, indicating the importance of binding RNA in the function of FMRP. Figure created by Georgina Aldridge.

neurons, suggesting not only that FMRP assists in mRNA transport and processing, but also that it is involved in the translation of those mRNAs. Binding of mRNA is critical to FMRP function as evidenced by the identification of a point mutation (I304N) in one of the mRNA-binding motifs (KH2) that leads to loss of mRNA binding and a severe FXS phenotype in an individual with otherwise normal levels of FMRP. Several groups have published lists of the mRNAs bound by FMRP; some of these mRNAs exhibit a secondary structure that has a G-quartet motif shown to bind to the RGG box of FMRP, and others form a 'kissing complex' with FMRP's KH2 domain.

One common theory is that FMRP is involved in nucleocytoplasmic shuttling of a subset of mRNA, and that it helps regulate the synthesis of these messages into proteins that are important for synaptic plasticity and learning and memory (see Figure 2). It is likely that the phenotype of FXS reflects the mRNA cargoes of FMRP and proteins whose synthesis is regulated by FMRP. An aspect of this theory that is under active investigation is whether FMRP functions as an initiator or as a repressor of translation. Several groups have described FMRP as an inhibitor of protein synthesis and have shown that, in the absence of FMRP, baseline levels of protein synthesis are elevated in some brain regions and that the levels of several specific proteins are elevated in the brain. By contrast, FMRP also appears to be required for the initiation of synaptic protein synthesis following metabotropic glutamate receptor activation. Protein translation at synapses in response to neurotransmitter activation is diminished in mice lacking an intact FMR1 gene. These seemingly inconsistent findings may reflect a dual role for FMRP: in a 'resting state,' FMRP may help keep protein translation of its bound mRNAs at low baseline levels, whereas, when activated via cellular signaling pathways,

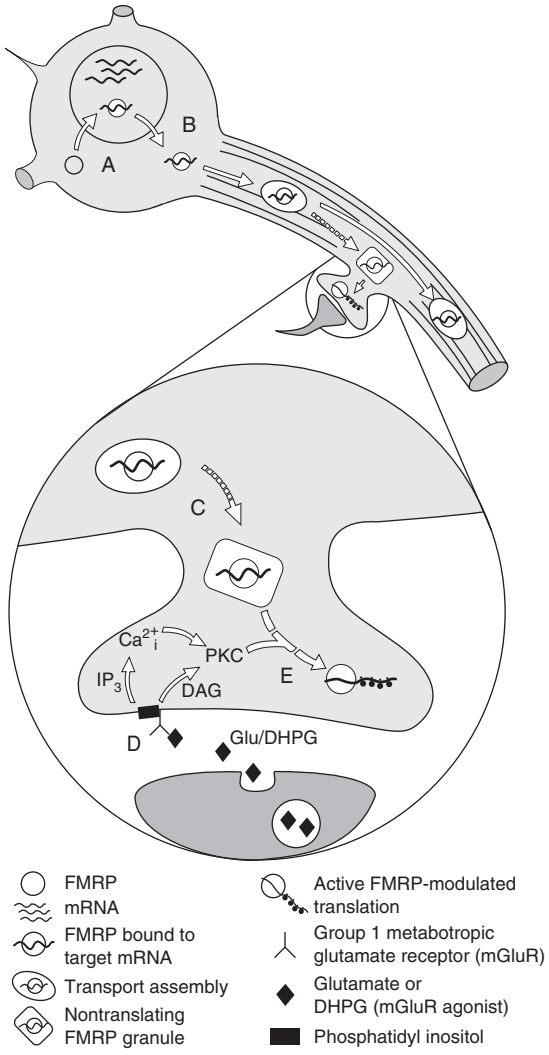

Figure 2 A putative role for fragile X mental retardation protein (FMRP) in neurons. FMRP binds to its target mRNAs in the nucleus (A) and helps export them to somatic cytoplasm. FMRP and its target mRNA are packaged into transport assemblies (B) and travel, likely via microtubules, down dendrites toward synapses. FMRP–mRNA transport assemblies may repress translation (C) by taking the form of nontranslating granules in the cytoplasm and near synapses, where they await some synaptic signal. Activation of group 1 metabotropic glutamate (Glu) receptors (D) triggers an enzyme cascade that, via many intermediates (broken arrow), signals nontranslating granules to release FMRP-bound target mRNA for translation (E). PKC, protein kinase C; DAG, diacyl glycerol; IP_3, inositol triphosphate; DHPG, (*S*)3,5-dihydroxyphenylglycine. Adapted from Weiler IJ, Spangler CC, Klintsova AY, et al. (2004) Fragile X mental retardation protein is necessary for neurotransmitter-activated protein translation at synapses. *Proceedings of the National Academy of Sciences of the United States of America* 101: 17504–17509, with permission.

FMRP may participate in the initiation of protein synthesis in response to synaptic activity. Very recent reports have suggested that FMRP is transported throughout neurons in granules, and that its dual role in protein synthesis regulation may involve shuttling between these nontranslating granules and actively translating polyribosomes. It therefore is reasonable to hypothesize that many among the broad spectrum of symptoms that stem from this single gene disorder may be associated with the subset of mRNAs whose transport and translation is regulated by FMRP. One putative FMRP cargo mRNA is the glucocorticoid receptor, which provides negative feedback based on the amount of glucocorticoid stress hormones in the circulation. It has been shown that fragile X patients have elevated glucocorticoid levels and protracted glucocorticoid responses to stress. Elevated glucocorticoids are associated with heightened anxiety, a behavioral characteristic of many FXS patients.

Fragile X Knockout Mice

In 1994 the Dutch-Belgian Fragile X Consortium developed a mouse model of FXS that has opened many new doors to investigators studying FXS and the function of FMRP. In this mouse model, the basis for FMRP's absence is not an expanded CGG repeat, but instead a neomycin cassette inserted into exon 5 of the murine *fmr1* gene. This insert disrupts transcription of *fmr1* mRNA and hence translation of functional FMRP. Macroorchidism has been observed consistently in these animals, but compared with the cognitive and behavioral deficits observed in patients with FXS, the behavioral phenotype in *fmr1* knockout mice is quite subtle and, by most accounts, these animals appear relatively normal.

Nevertheless, after over a decade of experimentation in these animals, a profile of behavioral phenotypes, many of which parallel FXS symptoms, is beginning to emerge. The first behavioral phenotype described in *fmr1* knockout mice was impairment on the reversal trial of a spatial memory task, the Morris water maze. This finding has been replicated by others, though animals consistently performed well on other aspects of this task. Impairments in radial maze learning and plus-shaped water maze learning have been reported, and *fmr1* knockout mice also exhibit deficits in object recognition memory, including a failure to habituate to objects. Another behavioral phenotype that has received much attention is behavior suggestive of anxiety, paralleling the social anxiety seen in human patients. In a passive avoidance task, for example, *fmr1* knockout mice demonstrate decreased latency (compared with wild-type mice) to enter the dark area where these animals had received a shock 24 h earlier. With regard to social interaction, *fmr1* knockout mice exhibited increased anxiety-like responses (in the mirrored chamber task), and diminished social dominance.

The behavioral phenotype in *fmr1* knockout mice that is described most consistently involves sensory hyperresponsiveness, particularly to auditory stimuli. Remarkably, a very similar set of symptoms has been reported in FXS patients. The *fmr1* knockout mice bred in various background strains exhibit greater startle responses than do wild-type mice. Similar findings have been reported in a prepulse tone inhibition paradigm: in knockout mice, the first tone elicited a greater inhibition of response to the second tone than in wild-type mice. Loud tones induce audiogenic seizures (leading often to respiratory arrest and death) in *fmr1* knockout mice with much higher frequency than in wild-type mice.

Throughout the investigation for a behavioral phenotype, results on some tasks have varied by laboratory and depending on which strain of mice was used. For example, several groups have measured activity of *fmr1* knockout mice in an open field and results have varied, with some groups reporting a higher proportion of line crosses in the periphery (suggestive of anxiety), and others reporting increased entry to center of the open field. The consistent finding in studies of knockout mouse behavior in the open field, however, has been locomotor hyperactivity. In conditioned fear paradigms, some authors suggest that knockout mice spend significantly less time freezing than do wild-type mice, whereas others have observed no differences. In other tasks, including a complex visual and auditory discrimination task, *fmr1* knockout mice appear to perform better than do wild-type animals. The fact that of some aspects of the *fmr1* knockout phenotype depend on mouse strain or genetic background may model the contribution of genetic background to the variability in expression of symptoms among individuals with FXS. It is particularly interesting in this regard that FMRP regulates the translation of the mRNA products of multiple genes.

Despite their relatively normal behavior, *fmr1* knockout mice exhibit a neuropathologic phenotype that is remarkably similar to that of FXS patients. In the visual and somatosensory cortex, the density of spines along dendrites of pyramidal neurons (an estimate of synapse number) is greater in adult *fmr1* knockout mice than in wild-type mice. Furthermore, knockout mice have, on average, longer dendritic spines in these brain regions than do wild types, as well as spines whose morphology might be considered more characteristic of developing than of mature spines. The manner in which this spine phenotype emerges in *fmr1* knockout mice (and correspondingly in FXS patients) is still an area of active investigation. In the neocortex, *fmr1* knockout mice may fail to undergo normal processes of developmental synapse elimination. Instead, some synapses that might normally mature or undergo 'synaptosis' appear to persist in an immature state into adulthood.

The hypothesis that FMRP is important for developmental pruning has been tested with regard to dendrites in the somatosensory cortex and olfactory bulb of *fmr1* knockout mice. During normal development, excess dendrites in both of these brain regions have been shown to be pruned in an experience-dependent process. Yet in adult *fmr1* knockout mice, many of these excess dendrites remained (compared with wild-type controls), suggesting a deficit in the pruning process. A growing body of evidence indicates that these deficits are part of more global state of brain disorganization which, in human FXS patients, also appears to involve subcortical brain regions, such as the lateral geniculate nucleus.

Further investigation of neural plasticity in *fmr1* knockout mice has revealed important deficits in these animals, particularly in the ability to exhibit long-term potentiation (LTP) and long-term depression (LTD). Although studies in several laboratories have examined LTP in the hippocampus, no consistent differences between wild-type and *fmr1* knockout mice in the extent of short- or long-term potentiation have been reported. By contrast, several groups have reported significant reductions in the magnitude of LTP in the neocortex of *fmr1* knockout mice. Furthermore, one form of LTD appears to be enhanced in *fmr1* knockout mice (metabotropic glutamate receptor- and protein synthesis-dependent LTD). These groups have suggested that FMRP normally limits LTD by suppressing the synthesis of proteins required for this form of synaptic plasticity, and that in the absence of FMRP, translation of these important proteins is increased, such that LTD is enhanced and that *de novo* protein synthesis may no longer be required for induction of LTD in *fmr1* knockout mice. From the finding of enhanced LTD in the absence of FMRP, the metabotropic glutamate receptor, or 'mGluR,' theory of FXS has emerged. This theory hypothesizes that FMRP normally inhibits aspects of postsynaptic signaling that follow stimulation of group 1 metabotropic glutamate receptors. Proponents of the mGluR theory further propose that many of the symptoms of FXS may stem from unbridled transduction of mGluR signals and that antagonism of mGluRs may attenuate some of these symptoms. Early tests using mGluR antagonists have shown some promise, rescuing to near wild-type-levels some behavioral phenotypes of *fmr1* knockout mice. The mRNAs of several proteins in mGluR-driven signaling pathways bind selectively to FMRP, and it is possible that the activity of these pathways is affected in FXS.

Two genes homologous to *FMR1* have been identified in vertebrates, *FXR1* (fragile X related gene 1) and *FXR2*. These genes have relatively high homology to *FMR1* at the protein structural level, bind mRNA, and bind to FMRP. The *FXR1* protein, FXR1P, appears to be expressed in the nucleus of fetal tissue and particularly in developing myocytes, whereas FXR2P is expressed in cell cytoplasm. In the brain, FXR1P and FXR2P are expressed primarily in cell bodies but also in dendrites at synapses. The two FMRP homologs are thought to exhibit different patterns of shuttling between nucleus and cytoplasm. It is thought, therefore, that the mRNA targets of these two proteins are distinct from one another. Both an *fxr1* and an *fxr2* knockout mouse have been produced. The FXR1P homozygous knockout dies within 24 h of birth, and a heterozygous version of this mouse exhibits abnormal limb musculature. The FXR2P knockout has a normal life span and exhibits a behavioral profile of learning deficits similar to that of *fmr1* knockout mice (though some differences exist) as well as circadian rhythm deficits.

A homolog to FMRP has been identified in the fruit fly *Drosophila melanogaster*. Mutant flies that lack the dFMR1 (also called dFXR) protein exhibit overextension of neurites during development of the mushroom bodies (a brain region in *Drosophila* that is linked to memory) and a behavioral phenotype that includes circadian rhythm abnormalities as well as altered courtship behavior. Administration of mGluR antagonists has been shown to restore memory impairments in courtship behavior and structural defects in the mushroom bodies.

Future Treatment Strategies Aimed at the Underlying Mechanism in FXS

Based on information related to synaptic function obtained from the *fmr1* knockout mouse, new treatment strategies have been proposed aimed at correcting synaptic deficits in the underlying disorder in FXS. The finding of reduced LTP and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor number and activity in the *fmr1* knockout prompted a phase II clinical trial of CX516, an AMPA receptor activator, in FXS. The drug proved safe, although low potency precluded definitive results, and use of more potent AMPA activators remains a potential treatment strategy for the future. Lithium has been shown to reduce inositol phosphate turnover, and thus acts on the signaling cascade required for mGluR activation of protein synthesis. It has been proposed that lithium might thus reduce overactive mGluR-mediated protein synthesis in FXS and administration of lithium does rescue the phenotype in *dfxr* fly mutants. There are some anecdotal data suggesting lithium may be quite helpful for behavior in FXS, and a formal clinical trial examining the effects of lithium on cognition and behavior in individuals with FXS is in progress. Negative modulators of mGluR5 receptor activity, similar to those utilized to rescue phenotypes in the mouse and fly models, have been shown to have potent anxiolytic activity in preclinical studies and are in development for use in humans with FXS. In the future, deficits in synaptic plasticity in FXS may treated with one or a combination of these agents or with yet-to-be-discovered pharmaceuticals that act on the CNS mechanism of disability in the absence of FMRP.

Premutation Phenotype and Fragile X-Associated Tremor/Ataxia Syndrome

In the past decade, several phenotypes associated with the FXS premutation have emerged. Some of these phenotypic features represent milder versions of features seen in FXS. Such mild phenotypic features had been difficult to demonstrate because they tend to overlap with the normal population and because there appears to be no evidence of impairment in overall cognitive functioning in premutation carriers. Behavioral features (e.g., anxiety, social phobia, and neurotic tendencies) are increased in premutation carriers relative to control cohorts. In addition, mild executive function deficits and learning disability can be seen in premutation carriers, predominantly those at the higher end of the repeat range (>150 repeats). Physical features of FXS (e.g., prominent ears) are also increased in premutation carriers relative to control groups.

A second set of premutation-associated phenotypes is specific to the premutation and these phenotypes do not occur in individuals with a full mutation. Premature ovarian failure (POF) is one such phenotype, occurring in approximately 20% of premutation carriers. Indeed, roughly 2% of women with sporadic POF and 14% of women with familial POF carry the fragile X premutation, and genetic screening should be offered to these women.

Another phenotype, occurring predominantly in male premutation carriers, is fragile X-associated tremor/ataxia syndrome (FXTAS). These men present with complex tremor, gait disorder, parkinsonian symptoms, and in some cases executive function deficits progressing to a dementia with subcortical features. Autonomic dysfunction and neuropathy may also be seen. The disorder is variably progressive, with many cases progressing over quite prolonged time frames. Traditionally, these individuals have been misdiagnosed with essential tremor, parkinsonism, dementia, stroke, or even multiple sclerosis. The radiological hallmark of FXTAS is increased signal in

the middle cerebellar peduncle (MCP sign) on T2-weighted and fast fluid-attenuated inversion recovery (FLAIR) MR images. Increased signal in subcortical white matter and generalized cerebral and cerebellar atrophy are also often seen. Neuropathological findings associated with FXTAS include eosinophilic intranuclear inclusions in neurons and glia and spongiform change in the white matter with mild axonal and myelin loss. Infrequently, females do seem to develop symptoms of FXTAS, although the symptoms are typically milder, occur later, and are less progressive than in males. This is likely due to the protective effect of the normal *FMR1* gene, and symptom predisposition in females is thought to relate to X-inactivation ratios.

In contrast to patients with the full mutation, premutation carriers have elevated levels of *FMR1* mRNA. Premutation-specific phenotypes are thought to result from cellular toxicity related to elevated levels of the CGG-repeat-containing *FMR1* mRNA. Indeed, the *FMR1* mRNA has been identified within the inclusions in brain samples from FXTAS patients, along with a list of proteins including nuclear regulatory protein lamin A/C. In support of the RNA toxicity mechanism for FXTAS, mouse and *Drosophila* models expressing an expanded CGG repeat have been generated and also develop neurodegenerative signs and neuronal inclusions.

See also: Autism; Developmental Disability and Fragile X Syndrome: Clinical Overview; Fragile X Associated Tremor Ataxia Syndrome.

Further Reading

Bagni C and Greenough WT (2005) From mRNP trafficking to spine dysmorphogenesis: The roots of fragile X syndrome. *Nature Reviews Neuroscience* 6: 376–387.

Bear MF, Huber KM, and Warren ST (2004) The mGluR theory of fragile X mental retardation. *Trends in Neuroscience* 27: 370–377.

Beckel-Mitchener A and Greenough WT (2004) Correlates across the structural, functional, and molecular phenotypes of fragile X syndrome. *Mental Retardation and Developmental Disability Research Reviews* 10: 53–59.

Berry-Kravis E, Grossman AW, Crnic LS, et al. (2002) Understanding fragile X syndrome. *Current Paediatrics* 12: 316–324.

Berry-Kravis E and Potanos K (2004) Psychopharmacology in fragile X syndrome – present and future. *Mental Retardation and Developmental Disability Research Reviews* 10: 42–48.

Darnell JC, Fraser CE, Mostovetsky O, et al. (2005) Kissing complex RNAs mediate interaction between the fragile-X mental retardation protein KH2 domain and brain polyribosomes. *Genes & Development* 19: 903–918.

Hagerman PJ and Hagerman RJ (2004) Fragile X-associated tremor/ataxia syndrome (FXTAS). *Mental Retardation and Developmental Disability Research Reviews* 10: 25–30.

Hagerman PJ and Hagerman RJ (2004) The fragile-X premutation: A maturing perspective. *American Journal of Human Genetics* 74: 805–816.

Hall DA, Berry-Kravis E, Jacquemont S, et al. (2005) Initial diagnoses given to persons with the fragile X associated tremor/ataxia syndrome (FXTAS). *Neurology* 65: 299–301.

Hessl D, Rivera SM, and Reiss AL (2004) The neuroanatomy and neuroendocrinology of fragile X syndrome. *Mental Retardation and Developmental Disability Research Reviews* 10: 17–24.

Hessl D, Tassone F, Loesch DZ, et al. (2005) Abnormal elevation of FMR1 mRNA is associated with psychological symptoms in individuals with the fragile X premutation. *American Journal of Medical Genetics B (Neuropsychiatric Genetics)* 139: 15–121.

Jin P, Zarnescu DC, Zhang F, et al. (2003) RNA-mediated neurodegeneration caused by the fragile X premutation rCGG repeats in Drosophila. *Neuron* 39: 739–747.

Maddalena A, Richards CS, McGinniss MJ, et al. (2001) Technical standards and guidelines for fragile X: The first of a series of disease-specific supplements to the Standards and Guidelines for Clinical Genetics Laboratories of the American College of Medical Genetics. Quality Assurance Subcommittee of the Laboratory Practice Committee. *Genetics in Medicine* 3: 200–205.

McBride SM, Choi CH, Wang Y, et al. (2005) Pharmacological rescue of synaptic plasticity, courtship behavior, and mushroom body defects in a *Drosophila* model of fragile X syndrome. *Neuron* 45: 753–764.

O'Donnell WT and Warren ST (2002) A decade of molecular studies of fragile X syndrome. *Annual Review of Neuroscience* 25: 315–338.

Tassone F, Iwahashi CK, and Hagerman PJ (2004) FMR1 RNA within the intranuclear inclusions of fragile X-associated tremor/ataxia syndrome (FXTAS). *RNA Biology* 1: 103–105.

Van Dam D, Errijgers V, Kooy RF, et al. (2005) Cognitive decline, neuromotor and behavioural disturbances in a mouse model for fragile-X-associated tremor/ataxia syndrome (FXTAS). *Behavioural Brain Research* 162: 233–239.

Vanderklish PW and Edelman GM (2005) Differential translation and fragile X syndrome. *Genes, Brain and Behavior* 4: 360–384.

Weiler IJ, Spangler CC, Klintsova AY, et al. (2004) Fragile X mental retardation protein is necessary for neurotransmitter-activated protein translation at synapses. *Proceedings of the National Academy of Sciences of the United States of America* 101: 17504–17509.

Free Radicals in Autonomic Functions

J Lincoln, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Free radicals are molecules that have an unpaired electron. As a consequence, they are highly reactive and they pass freely through membranes. Such properties mean that they do not conform well to standard 'biological rules' where molecular shape and cellular localization govern function. Thus, free radicals have generally been viewed as harmful to cells. However, in 1987, endothelium-derived relaxing factor, known to contribute to normal relaxation of vascular smooth muscle, was identified as the free radical nitric oxide (NO). Subsequent studies established that NO is also synthesized by neurons within the autonomic nervous system and have altered significantly our understanding of the autonomic control of a variety of physiological processes and introduced new approaches for clinical therapy. However, cells still need to handle free radicals with care, and there are a variety of intracellular mechanisms designed to restrict the production and/or reactions of free radicals. When such mechanisms are impaired or the intracellular environment of cells is altered, this can result in toxic effects that contribute to autonomic dysfunction.

Free Radicals, Reactive Oxygen Species, and Reactive Nitrogen Species

Because of their molecular form, free radicals readily undergo oxidation and reduction reactions resulting in a variety of products. In the case of radicals derived from oxygen, these are referred to as reactive oxygen species (ROS), and those derived from nitrogen are referred to as reactive nitrogen species (RNS). Overproduction of ROS or RNS results in the intracellular conditions of oxidative stress and nitrosative stress, respectively. The major free radical derived from oxygen produced by autonomic neurons is the superoxide anion ($O_2^{-\bullet}$) that is readily converted to hydrogen peroxide (H_2O_2) and subsequently to water. In the presence of ferrous or cuprous ions, H_2O_2 can form the hydroxyl radical (OH•) that has the potential to damage all classes of biological molecules. NO can act as an antioxidant preventing lipid peroxidation. However, under conditions in which relatively high levels of both NO and $O_2^{-\bullet}$ are produced, they react together to produce peroxynitrite ($ONOO^-$). Peroxynitrite is a pro-oxidant that has toxic effects. Thus, the intracellular environment can influence what species are produced and whether the result is harmful.

Synthesis of ROS

Neurons have a high level of metabolic activity and contain relatively high numbers of mitochondria. However, there is a price to pay for the production of energy by aerobic metabolism in mammalian cells, and this is the formation of ROS as a by-product. The major site for the production of ROS in cells is mitochondria, and this occurs normally during respiration. Electrons produced by the oxidation of NADH are passed along the electron transport chain in the mitochondrial inner membrane. During this process, oxygen is reduced to water and redox energy is conserved by pumping protons across the inner membrane to build up a proton electrochemical gradient. The proton electrochemical gradient drives the synthesis of adenosine 5′-triphosphate (ATP), during which there is inward flow of protons through ATP synthase. Protons can also re-enter via nonspecific proton leak and by translocation via uncoupling proteins. However, not all of the molecular oxygen is fully reduced to water, and a fraction is incompletely reduced to $O_2^{-\bullet}$. The formation of $O_2^{-\bullet}$ is particularly favored under a condition known as state 4 respiration. This occurs when mitochondria are respiring but there are high levels of ATP. Electrons accumulate on the electron transport chain because the proton electrochemical gradient is not dissipated by ATP synthesis. Reaction of the accumulated electrons with oxygen produces increased levels of ROS.

Synthesis of NO

NO is synthesized by the enzyme nitric oxide synthase (NOS). Three isoforms have been identified. Inducible NOS (iNOS or type 2 NOS) is not present under normal conditions but is induced as part of an immune reaction and subsequently synthesizes large quantities of NO that are toxic to invading organisms. Two isoforms are present constitutively. Endothelial NOS (eNOS or type 3 NOS) was first identified in endothelial cells, and neuronal NOS (nNOS or type 1 NOS) was first identified in neurons. nNOS is a very complex molecule and the synthesis of NO requires the presence of three cosubstrates (L-arginine, molecular oxygen, and NADPH) and five cofactors or prosthetic groups (flavin adenine dinucleotide, flavin mononucleotide, tetrahydrobiopterin, calmodulin, and heme).

Active nNOS consists of two enzymes in one. The C-terminal end of nNOS resembles cytochrome P450

reductase (CPR) and is responsible for the production of electrons from NADPH. The flavin cofactors direct the electrons toward the N-terminal portion. In the N-terminal region, in the presence of electrons, heme, and molecular oxygen, L-arginine is converted to citrulline producing NO in the process. The role of tetrahydrobiopterin has not been fully established, but it is thought to contribute to the conformation of the active nNOS and possibly to the redox reactions required for synthesizing NO. The calcium-binding protein, calmodulin, plays the crucial role of making nNOS activity dependent on the intracellular concentration of Ca^{2+}. Normally, nNOS only synthesizes low concentrations of NO that do not reach a toxic level. However, under certain conditions, nNOS has the capacity for synthesizing alternative products. When L-arginine levels are low, nNOS can actually synthesize $O_2^{-\bullet}$. Furthermore, there is evidence that when tetrahydrobiopterin binding is not optimal, nNOS can synthesize both NO and $O_2^{-\bullet}$. Since NO and $O_2^{-\bullet}$ can react together to produce peroxynitrite, this can be potentially harmful. A spliced variant of nNOS has been identified that is localized in mitochondria (mtNOS). The function of NO synthesized by mtNOS has not been fully established, but NO has the capacity to interact with metal groups present on enzymes within the electron transport chain, particularly cytochrome oxidase. In this way, it has been suggested that it can modify respiration and de-energize mitochondria in a transient and reversible manner.

Intracellular Defense Mechanisms

In view of their toxic nature, autonomic neurons have a variety of mechanisms to defend themselves against ROS and RNS. In the case of NO, the major protective mechanisms act to limit the levels produced. ROS are removed from neurons by a number of different antioxidant processes that act in different cellular localizations. Within mitochondria, the two main protective mechanisms are manganese superoxide dismutase (MnSOD) and glutathione peroxidase (GPx4). MnSOD converts $O_2^{-\bullet}$ to H_2O_2, whereas GPx4 converts H_2O_2 to water, oxidizing glutathione in the process. Glutathione is a major intracellular antioxidant and to maintain its protective effects it has to be continually recycled from its oxidized form by the action of the NADPH-dependent enzyme glutathione reductase. In the cytoplasm, alternative forms of the antioxidant enzymes (Cu/ZnSOD and GPx1) are present and catalyze similar reactions. Additionally, catalase converts H_2O_2 to water. Vitamin C, vitamin E, and taurine also have antioxidant properties. Vitamin C exists in the aqueous environment of the cytoplasm, whereas vitamin E is localized in the hydrophobic environment of membranes. Taurine not only is an antioxidant but also is released from cells under conditions of osmotic stress to maintain the internal osmolarity of the cell.

NO as a Neurotransmitter in the Autonomic Nervous System

Nitrergic Transmission

Despite the fact that NO does not conform to the classical criteria for the definition of a neurotransmitter, it is now accepted that it plays an important role in autonomic neuromuscular transmission. The presence of nNOS in autonomic neurons can readily be demonstrated using specific antibodies raised against nNOS and immunohistochemical techniques. Unlike classical neurotransmitters, in which the arrival of an action potential causes the calcium-dependent release of synaptic vesicles, NO cannot be stored in vesicles within nerves for release. However, the synthesis of NO by nNOS is calcium dependent, and it is this property that enables NO to mediate neurotransmission. A schematic representation of the events of nitrergic neurotransmission is given in **Figure 1**. Under resting conditions, the levels of Ca^{2+} within a nerve are not sufficient to cause the change in conformation of calmodulin required to activate nNOS. During an action potential, the intracellular concentrations of Ca^{2+} rise sufficiently to initiate the activation of nNOS and NO synthesis. NO is only synthesized for as long as the Ca^{2+} levels within the nerve are raised. Thus, only small amounts of NO are synthesized for short periods of time, and they do not reach toxic levels. Both calcium chelators and calmodulin antagonists can inhibit nitrergic transmission. NO is not as reactive as many free radicals and, under normal intracellular conditions, the most likely reaction it will undergo is with transition metals. NO can react with the heme group on nNOS inhibiting activity, providing a mechanism for feedback inhibition that further controls the levels of NO that can be produced.

NO does not produce a response by interacting with extracellular receptors on the target like classical neurotransmitters. Instead, NO reaches and enters the target by diffusion and reacts with the heme group of soluble guanylate cyclase (sGC) within the target. NO binds to heme, forming a five-coordinated nitrosyl–heme complex that results in up to 200-fold activation of sGC. Thus, NO increases synthesis of cGMP, and this causes relaxation of smooth muscle. Evidence that NO can act as an inhibitory neurotransmitter in the autonomic nervous system via the

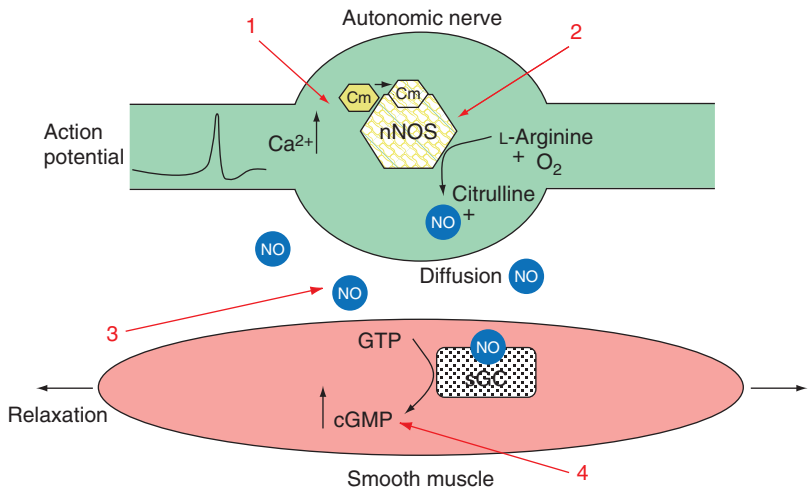

Figure 1 Schematic representation of nitrergic transmission in the autonomic nervous system. The arrival of an action potential causes a rise in intracellular calcium that binds to calmodulin (Cm) and activates neuronal nitric oxide synthase (nNOS). Nitric oxide (NO) is synthesized by nNOS during the conversion of L-arginine to citrulline. NO is released from autonomic nerves by diffusion and enters the target smooth muscle, where it interacts with the heme group on guanylate cyclase (sGC) activating the synthesis of cGMP and causing relaxation of the smooth muscle. Nitrergic transmission can be modified at the sites indicated: 1, calcium chelators and calmodulin antagonists inhibit the activation of nNOS; 2, L-arginine analogues are specific inhibitors of nNOS and are used to identify whether inhibitory transmission is specifically nitrergic; 3, free radical scavengers and superoxides can react with NO, inhibiting transmission; 4, agents that inhibit sGC can inhibit both the response to exogenous NO and the response to nerve stimulation, whereas drugs that inhibit the breakdown of cGMP by phosphodiesterase have the opposite effect.

production of cGMP has been obtained by demonstrating that a relaxation response following electrical stimulation of autonomic nerves can be prevented by NOS inhibitors and drugs that decrease the stimulation of sGC by NO (Figure 1). L-Arginine analogues are an important group of NOS inhibitors that compete with L-arginine at the active site of NOS, preventing NO synthesis and nitrergic transmission. The effect of these competitive NOS inhibitors can be reversed by high concentrations of L-but not D-arginine. Methylene blue is a nonspecific inhibitor that interacts with heme on both NOS and sGC. ODQ (1H-[1,2,4] oxadiazolo[4,3-α]-quinoxalin-1-one) is a selective inhibitor of sGC that competes with NO for binding and inhibits NO-stimulated cGMP synthesis while leaving basal synthesis unchanged. Some doubt was raised about whether NO was released from autonomic nerves as free NO or a related molecule formed by NO, such as a nitrosothiol, when some free radical scavengers were shown to prevent muscle relaxation in response to application of exogenous NO but had little effect on the response to nerve stimulation. However, it has been proposed that endogenous antioxidant mechanisms within tissue are able to protect the endogenous free NO that is released from the nerve. The susceptibility of NO to reactions with ROS emphasizes the importance of the antioxidant status of an organ in determining the effectiveness of nitrergic transmission.

Physiological Roles of NO in the Autonomic Nervous System

This section focuses on examples of autonomic transmission in which NO has been clearly demonstrated to play the major role and in which a deficit has been implicated in pathological conditions. However, it should be noted that the presence of nNOS-containing autonomic nerves has been demonstrated throughout the gastrointestinal, cardiovascular, and urogenital systems together with nerve-mediated relaxation responses that can be inhibited by NOS inhibitors. As with all autonomic neurotransmitters in which co-localization and cotransmission is a characteristic phenomenon, NO does not occur on its own but, rather, in combination with other transmitters. nNOS is most often present in postganglionic parasympathetic nerves, in which it is commonly co-localized with acetylcholine and/or vasoactive intestinal polypeptide. It is also very well represented in the enteric nervous system. However, there is less evidence for the presence of nNOS in postganglionic sympathetic nerves.

Penile erection Historically, studies of the erectile process have played a pivotal role in establishing NO as an autonomic neurotransmitter. Penile erection is controlled by parasympathetic nerves, in which stimulation results in the relaxation of the

smooth muscle of erectile tissue. However, it has long been known that blockade of the actions of acetylcholine by atropine fails to block the erectile response. The release of an additional neurotransmitter from parasympathetic nerves had been proposed, but its identity was unknown until it was demonstrated that NO could be synthesized by mammalian cells. Autonomic neurons containing nNOS have been demonstrated in pelvic ganglia that supply the penis, and nNOS-containing nerve fibers are present in the erectile tissue of all species including humans. The development of L-arginine analogues to inhibit NOS activity provided the tools to demonstrate that NO is the major neurotransmitter involved in the erectile response.

A decrease in nNOS-containing nerve fibers occurs in erectile tissue from patients with erectile dysfunction. Furthermore, a deficit in neurogenic NO-mediated relaxation occurs in erectile tissue from animal models of diabetes and human diabetic patients and is considered to be a major factor in the increased incidence of impotence associated with diabetes. Understanding of the NO–cGMP pathway in the control of penile erection has led to the development of sildenafil citrate (Viagra) in the clinical treatment of impotence. Viagra is a specific phosphodiesterase inhibitor that acts by decreasing the breakdown of cGMP, thus maintaining the relaxation of erectile smooth muscle induced by NO.

The gastrointestinal tract The enteric nervous system supplies the entire gastrointestinal tract and can act independently of any input from the parasympathetic and sympathetic nervous systems. Enteric neurons are localized in two main plexuses: the myenteric plexus, located between the longitudinal and circular smooth muscle layers, and the submucosal plexus, located between the circular smooth muscle layer and the mucosa. Neurons containing nNOS can readily be demonstrated in both the myenteric (Figure 2) and the submucosal plexuses, sometimes co-localized with vasoactive intestinal polypeptide. Nerve-mediated relaxation of smooth muscle that can be inhibited by NOS inhibitors has also been demonstrated throughout the gastrointestinal tract from the esophagus to the anal sphincter. The first indication of the importance of NO in the autonomic control of the gastrointestinal tract was provided when knockout mice in which the gene for nNOS had been deleted were generated. The major gross change observed in these mice was a markedly distended stomach, indicating that relaxation of the pyloric sphincter had been impaired. Since that time, evidence has accumulated that NO is the major neurotransmitter involved in the relaxation of sphincters throughout the gastrointestinal tract. Similar studies of the urinary tract have suggested that NO is also an inhibitory neurotransmitter in the urethra and bladder neck.

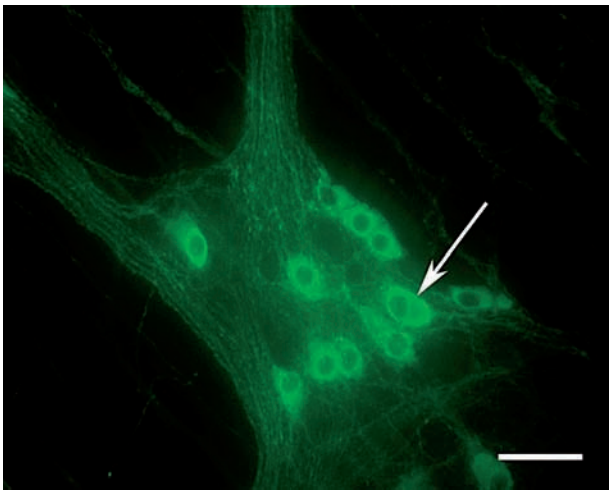

Figure 2 Localization of nNOS in intrinsic neurons of the myenteric plexus. Immunohistochemical labeling of nNOS (green fluorescence) in a whole mount preparation of the myenteric plexus of the rat proximal colon enables the plexus to be observed through the longitudinal muscle layer. The white arrow indicates one of the neurons within the myenteric ganglion positively labeled for nNOS with a negative nucleus. Note that some of the neurons within the same ganglion do not contain nNOS. Scale bar = 55 μm.

Impaired nitrergic transmission has been implicated in a variety of disorders of the gastrointestinal tract. A loss of nNOS-containing nerves has been demonstrated in agangliosis of the gut (Hirschsprung's disease), infantile pyloric stenosis, and achalasia of the gastroesophageal junction and internal anal sphincter, all of which result in impaired motility. In the case of infantile pyloric stenosis, the condition strongly resembles that observed in the stomachs of nNOS knockout mice.

Cerebral blood vessels The cerebral vessels provide an example of cotransmission of NO with the classical neurotransmitter, acetylcholine, in which the NO-mediated response is modulated by its co-transmitter. Co-localization of nNOS and choline acetyltransferase has been demonstrated in the parasympathetic neurons of the sphenopalatine ganglion that supply the cerebral blood vessels. Nerve-mediated relaxation of cerebral blood vessels is blocked by NOS and sGC inhibitors but not by atropine, indicating that NO is the main neurotransmitter mediating cerebral neurogenic vasodilation. Indeed, atropine enhances the relaxation response to nerve stimulation providing evidence that acetylcholine can modulate nitrergic transmission. It has been demonstrated that acetylcholine acts on presynaptic muscarinic

receptors, reducing Ca^{2+} influx via voltage-sensitive N-type Ca^{2+} channels. As a consequence, acetylcholine inhibits the synthesis of NO and nitrergic transmission in parasympathetic nerves supplying cerebral vessels. A loss of nNOS-containing nerves supplying cerebral vessels has been observed in experimental diabetes and implicated in the increased incidence of stroke in diabetic patients.

Parallels between NO and carbon monoxide Once NO had been demonstrated as a neurotransmitter, studies were undertaken to investigate whether other similar molecules could also mediate transmission. These revealed that carbon monoxide (CO) acts in a similar way. Although CO is not a free radical, it has been included here because there are some parallels between the synthesis of CO and NO and their effects. Electrons provided by the activity of CPR are used by the enzyme heme oxygenase (HO) to synthesize CO during the conversion of heme to biliverdin. HO2 is a constitutive enzyme that has been localized in autonomic neurons. CO can also activate sGC by interacting with the heme group causing increased synthesis of cGMP. However, CO forms a different type of heme complex with sGC from that formed by NO, and this is considerably less effective in activating sGC (sixfold compared to 200-fold). One problem in establishing CO as a neurotransmitter within the autonomic nervous system is that it has been difficult to demonstrate that its synthesis is related to nerve stimulation.

HO2 has been co-localized in some nNOS-containing neurons in the myenteric plexus of the gastrointestinal tract, and there appears to be cooperation between the actions of NO and CO in mediating relaxation of gastrointestinal smooth muscle. There is evidence that low concentrations of CO can stimulate the synthesis of NO by nNOS. In studies of nNOS, HO2, and double nNOS/HO2 knockout mice, depolarization of intestinal smooth muscle occurs in both single knockouts and is additive in the double knockout. In HO2 knockout mice, inhibitory responses to electrical stimulation of nerves are nearly abolished despite the fact that nNOS is still expressed. However, exogenous CO can restore normal responses in a manner that is prevented by an NOS inhibitor. Thus, CO is able to regulate nitrergic transmission in the gastrointestinal tract.

Pathological Conditions of the Autonomic Nervous System Involving ROS

Diabetes Mellitus and Oxidative Stress

Diabetes is the single major cause of autonomic dysfunction in humans. All parts of the autonomic nervous system can be affected in diabetes, resulting in a range of symptoms including postural hypotension, impaired control of heart rate, disordered gut motility, impotence, and bladder dysfunction. The presence of autonomic neuropathy in diabetic patients is associated with considerable morbidity and increased mortality and, in view of the dramatic increase in the number of diabetic patients, is a considerable clinical problem. Diabetic neuropathy is often characterized by an initial accumulation of neurotransmitter in the neuronal cell body and loss in the nerve terminal. Ultimately, degeneration of the peripheral nerve fibers occurs as a dying back type of axonopathy. The major factor associated with neuropathy in diabetic patients is poor glucose control, and metabolic changes associated with hyperglycemia have received considerable attention in research on the mechanisms underlying the development of diabetic neuropathy. A common feature of many of the diabetes-induced changes in metabolism in peripheral nerves is oxidative stress. The current major focus for research is now on the role of increased production of ROS in diabetic neuropathy and the potential use of antioxidants in its prevention and treatment. A schematic representation of how oxidative stress may occur in diabetes is shown in Figure 3. The evidence for the pathways described has been derived from a combination of *in vivo* studies of animal models of diabetes and *in vitro* studies of sensory and autonomic neurons exposed to high levels of glucose. It should also be emphasized that the metabolic changes given are not exclusive to neurons. Indeed, there is considerable evidence that similar changes occur in endothelial cells, leading to vascular dysfunction and potentially to decreased nerve blood flow and subsequent nerve deficits.

Nerves do not require insulin for the uptake of glucose; therefore, hyperglycemia in diabetes results in increased intracellular concentrations of glucose. This results in the activation of aldose reductase that converts glucose to sorbitol. Sorbitol dehydrogenase then converts sorbitol to fructose. Aldose reductase is an NADPH-dependent enzyme and it has been proposed that increased aldose reductase activity causes depletion of NADPH. One of the consequences of NADPH depletion is impaired recycling of the antioxidant glutathione by glutathione reductase, which is also an NADPH-dependent enzyme. Increased intracellular glucose also causes osmotic stress that results in depletion of the antioxidant taurine. Glucose can undergo autoxidation, forming reactive species. Under normal conditions, glycation of macromolecules only occurs by an enzymatic process. However, at high concentrations both glucose and fructose can glycate macromolecules via a nonenzymatic process. Following a series of chemical modifications, this results in

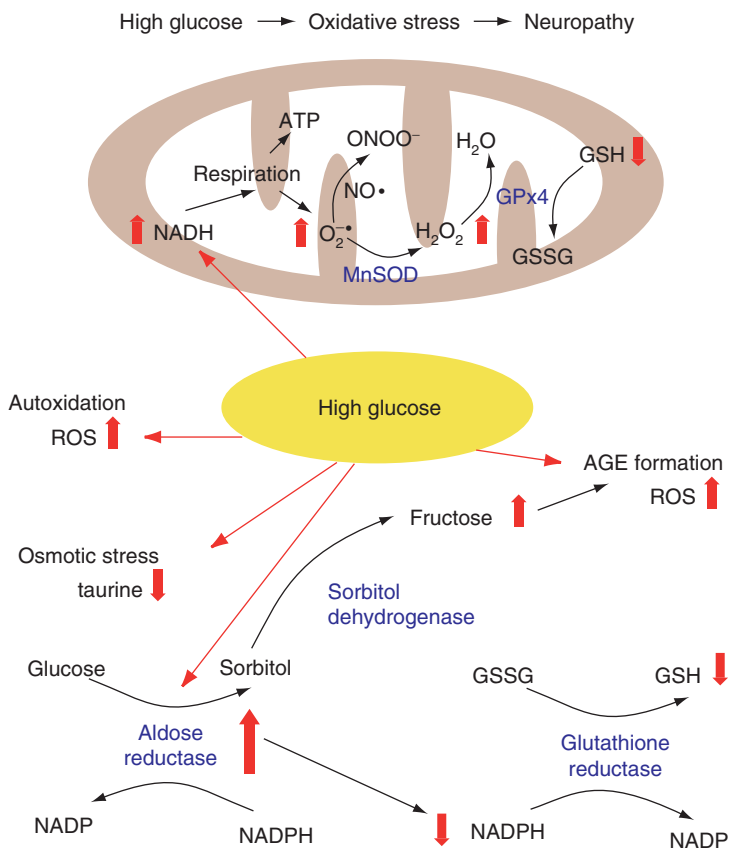

Figure 3 Schematic representation of the pathways by which high glucose levels in diabetes may lead to oxidative stress. Hyperglycemia results in high intracellular levels of glucose in nerves that activates aldose reductase, resulting in the formation of fructose and depletion of NADPH. Glucose can undergo autoxidation forming reactive oxygen species (ROS), and high levels of glucose and fructose can also lead to the formation of advanced glycation end products (AGEs) that are highly reactive. Antioxidant defense is decreased as osmotic stress induced by glucose results in the depletion of the antioxidant taurine and the reduction in NADPH levels reduces the recycling of the antioxidant glutathione (GSH) from its oxidized form (GSSG). Glucose also affects mitochondrial function, resulting in increased production of ROS. Glucose results in increased levels of NADH entering the electron transport chain, increasing the production of superoxide ($O_2^{-\bullet}$), hydrogen peroxide (H_2O_2), and also peroxynitrite ($ONOO^-$) by interaction with NO. Depletion of GSH reduces the activity of the mitochondrial form of the antioxidant enzyme, glutathione peroxidase (GPx4).

the formation of advanced glycation end products (AGEs), which are not only highly reactive but also can act on the AGE receptor inducing oxidative stress. Oxidative stress can lead to mitochondrial dysfunction, resulting in a vicious cycle in which increased mitochondrial production of ROS is induced following the production of ROS in the cytoplasm. It has been proposed that high glucose can, via the tricarboxylic acid cycle, produce excessive levels of NADH that enter the electron transport chain and lead directly to increased production of ROS by mitochondria. There is also preliminary evidence that diabetes can result in reduced expression of uncoupling protein-3, thus impairing the ability of mitochondria to dissipate the proton electrochemical gradient and favoring the production of ROS.

The exact sequence of events that occur in diabetic neuropathy has yet to be established unequivocally. One current area of debate is whether diabetes-induced oxidative stress leads to neuronal cell death *in vivo* rather than just leading to axonopathy. There are conflicting reports on the effect of experimental diabetes on sensory dorsal root ganglion neuron number. Studies of the sympathetic superior mesenteric ganglia have demonstrated neuroaxonal dystrophy in diabetes in the absence of any neuronal cell loss. Apoptosis has been observed in nNOS-containing but not tyrosine hydroxylase-containing neurons in the autonomic pelvic ganglia from diabetic rats. The presence of apoptosis was correlated with the diabetes-induced formation of AGEs, and it has been suggested that the interaction of NO with reactive AGEs to produce peroxynitrite may make nNOS-containing neurons particularly vulnerable in diabetes. Whether neuronal cell death occurs in diabetes or not, oxidative stress is regarded as a key

factor in the development of neuropathy. As a consequence, considerable attention is being paid to the use of antioxidants in the prevention and/or treatment of diabetic neuropathy. α-Lipoic acid is an antioxidant that acts in both an aqueous and a lipid environment, and there is evidence from clinical trials that it has beneficial effects in the treatment of cardiac autonomic neuropathy in diabetic patients.

Aging

It is recognized that autonomic dysfunction can occur as a consequence of aging. It has been proposed for some time that oxidative stress contributes to altered function of central neurons in aging and neurodegenerative disease, but there has been less direct investigation of this in autonomic neurons. However, the similarities between the effect of diabetes and the effect of aging on sympathetic neurons suggest that oxidative stress may also be significant in the aging autonomic nervous system. The same features of neuroaxonal dystrophy have been observed in the paravertebral celiac/superior mesenteric sympathetic ganglion in both aging and diabetes in animals and humans. Furthermore, sympathetic neurons of the prevertebral sympathetic superior cervical ganglion are relatively resistant to a degenerative process in both conditions. A loss of intrinsic neurons in the gut in aging has been demonstrated in both rodents and humans. There is preliminary evidence for increased ROS levels and a greater susceptibility to ROS-induced apoptotic cell death in myenteric neurons from aged rats. In addition, caloric restriction of aging animals may exert a protective effect on myenteric neurons against ROS. This is of interest because caloric restriction has been shown to decrease the production of ROS by mitochondria and has been studied extensively in relation to delayed aging and increased longevity in rodents. However, further study is required in order to demonstrate a direct relationship between oxidative stress and autonomic dysfunction in aging.

Conclusions

The transitory nature of free radicals and the complexity of the reactions that they can undergo make it inherently difficult to study their role in autonomic function and in pathological conditions affecting the autonomic nervous system. The discovery of NO as a neurotransmitter in the autonomic nervous system has had a dramatic effect on our understanding of the autonomic control of visceral organs under normal conditions and emphasized that free radicals cannot simply be regarded as dangerous to autonomic neurons. However, studies of the mechanisms underlying the development of neuropathy in diabetes have also highlighted the importance of ROS in pathological conditions and are having a significant impact on current approaches to the treatment of autonomic neuropathy. Research on the autonomic nervous system has lagged behind similar studies on the central nervous system. In particular, there is still a great deal to learn on the role of free radicals in autonomic dysfunction in old age.

See also: Autonomic Disorders; Autonomic Failure; Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Diabetic Neuropathy; Nitric Oxide; Role of NO in Neurodegeneration.

Further Reading

Cellek S, Foxwell NA, and Moncada S (2003) Two phases of nitrergic neuropathy in streptozotocin-induced diabetic rats. *Diabetes* 52: 2353–2362.

Ghafourifar P and Cadenas E (2005) Mitochondrial nitric oxide synthase. *Trends in Pharmacological Sciences* 26: 190–195.

Gibbons SJ and Farrugia G (2004) The role of carbon monoxide in the gastrointestinal tract. *Journal of Physiology* 556: 325–336.

Green K, Brand MD, and Murphy MP (2004) Prevention of mitochondrial oxidative damage as a therapeutic strategy in diabetes. *Diabetes* 53(supplement 1): S110–S118.

Lee TJF, Liu J, and Evans MS (2001) Cholinergic–nitrergic transmitter mechanisms in the cerebral circulation. *Microscopy Research and Technique* 53: 119–128.

Lincoln J, Hoyle CHV, and Burnstock G (1997) *Nitric Oxide in Health and Disease*. Cambridge, UK: Cambridge University Press.

Merry BJ (2003) Oxidative stress and mitochondrial function with aging – The effects of caloric restriction. *Aging Cell* 3: 7–12.

Russwurm M and Koesling D (2004) Guanylyl cyclase: NO hits its target. *Biochemical Society Symposia* 71: 51–63.

Schmidt RE (2002) Neuropathology and pathogenesis of diabetic autonomic neuropathy. *International Review of Neurobiology* 50: 257–292.

Stuehr DJ, Wei C-C, Santolini J, et al. (2004) Radical reactions of nitric oxide synthase. *Biochemical Society Symposia* 71: 39–49.

Toda N, Ayajiki K, and Okamura T (2005) Nitric oxide and penile erectile function. *Pharmacology & Therapeutics* 106: 233–266.

Toda N and Herman AG (2005) Gastrointestinal function regulation by nitrergic efferent nerves. *Pharmacological Reviews* 57: 315–338.

Toda N and Okamura T (2003) The pharmacology of nitric oxide in the peripheral nervous system of blood vessels. *Pharmacological Reviews* 56: 271–324.

Vincent AM, Russell JW, Low PA, and Feldman EL (2004) Oxidative stress in the pathogenesis of diabetic neuropathy. *Endocrine Reviews* 25: 612–628.

Wade PR and Cowen T (2004) Neurodegeneration: A key factor in the ageing gut. *Neurogastroenterology & Motility* 16(supplement 1): 19–23.

Friedreich Ataxia

M Pandolfo, Brussels Free University (ULB), Erasme Hospital, Brussels, Belgium

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Friedreich ataxia (FA), first described in 1863 by Nicholaus Friedreich, professor of medicine in Heidelberg, is the most common of the early-onset hereditary ataxias in Caucasians. It is an autosomal recessive hereditary disease characterized by progressive neurological disability, hypertrophic cardiopathy, and increased risk of diabetes mellitus. Skeletal abnormalities such as kyphoscoliosis and pes cavus are common. On average, after 10–15 years of disease, the progressive loss of coordination and balance results in the use of a wheelchair and in the need of help for all daily living activities. Some patients have a very severe cardiomyopathy that can cause premature death for cardiac insufficiency or arrhythmia. The complications of diabetes may further contribute to disease burden and disability. As it is the case for almost all neurodegenerative diseases, there is currently no treatment to stop the progression of FA. Furthermore, there are no effective symptomatic treatments comparable to levodopa for Parkinson disease, so the patient's condition relentlessly deteriorates as the neurodegenerative process continues its progression. Here, after providing some information on the disease, the most significant advances in our understanding of the molecular and cellular pathogenesis of FA 10 years after the discovery of the causative gene mutation, and how these may lead to the development of effective therapies, are reviewed.

FA: Clinicopathological Features

In Caucasians, FA accounts for half of the overall hereditary ataxia cases and for three-quarters of those with onset before age 25 years, with a prevalence of around 2×10^{-5}. The disease may not exist in non-Caucasian populations. Friedreich's papers reported the essential clinical and pathological features of the disease, a "degenerative atrophy of the posterior columns of the spinal cord" leading to progressive ataxia, sensory loss, and muscle weakness, often associated with scoliosis, foot deformity, and heart disease. However, rigorous diagnostic criteria were established for research purposes only in the late 1970s and early 1980s. The phenotype has been redefined and extended after the discovery in 1996 of the disease gene, FRDA, encoding the mitochondrial protein FXN, and of its most common mutation, an intronic GAA triplet repeat expansion. The typical age of onset is around puberty, but it may vary substantially even within a sibship. Late-onset cases exist, even in the sixth decade and later, a phenomenon now in part explained by the dynamic nature of the underlying mutation. Gait instability or generalized clumsiness are typical presenting symptoms. Scoliosis may already be present when neurological symptoms appear. In rare cases, hypertrophic cardiomyopathy is diagnosed before the onset of ataxia. Ataxia is of mixed cerebellar and sensory type, affecting the trunk, with swaying, imbalance, and falls, and the limbs, with increasing difficulty in activities such as writing, dressing, handling utensils. Ataxia is progressive and unremitting, though periods of stability are frequent at the beginning of the illness. With progression, gait becomes broad-based, with frequent losses of balance requiring intermittent, then constant support. Fine motor skills deteriorate, dysmetria and intention tremor become evident. Mild to moderate limb weakness of central origin is a constant feature of advanced FA. Some amyotrophy, particularly in the hands, is also very common. On average, 10–15 years after onset, patients lose the ability to walk, stand, and sit without support. Evolution is variable, however, with mild cases who are still ambulatory decades after onset and severe cases that are wheelchair bound in a few years. Dysarthria consisting of slow, jerky speech with sudden utterances is a relatively early feature and progresses until speech become almost inintelligible. Dysphagia, particularly for liquids, appears with advancing disease, requiring modified foods and eventually a nasogastric tube or gastrostomy feedings. Cognitive functions are generally well preserved. However, because of substantial physical disability, FA may have a substantial impact on academic, professional, and personal development.

Findings at the neurological examination reflect the underlying pathology. A sensory neuropathy in dorsal root ganglia, accompanied by loss of sensory fibers in peripheral nerves and degeneration of the posterior columns in the spinal cord, is a hallmark of the disease. The larger neurons carrying proprioceptive information are mostly affected, resulting in loss of position and vibration sense and abolished reflexes. Perception of light touch, pain, and temperature only decrease with advancing disease. A progressive degeneration of pyramidal tracts usually determines extensor plantar responses and progressive muscular weakness. Some patients with more limited sensory neuronopathy have retained reflexes.

In these cases, pyramidal involvement may lead to hyperreflexia and spasticity, to the point of mimicking a spastic paraplegia. The relative impact of sensory neuropathy and of pyramidal tract degeneration clearly varies from patient to patient, resulting most often in the typical picture of areflexia associated with extensor plantar responses, but sometimes one component prevails. Such partial pictures are usually observed in milder cases of the disease. Fixation instability with square-wave jerks is the typical oculomotor abnormailty in FA, while gaze-evoked nystagmus is uncommon and ophthalmoparesis does not occur. About 30% of the patients develop optic atrophy, with or without visual impairment, and about 20% develop sensorineural hearing loss.

Neurophysiological studies of the peripheral nervous system (PNS) reveal a severe reduction or complete loss of sensory nerve action potentials (SNAPs), with motor and sensory (when measurable) nerve conduction velocities (NCVs) within or just below the normal range. These neurophysiological abnormalities do not appear to progress with time. In the central nervous system (CNS), degeneration of both peripheral and central sensory fibers also results in dispersion and delay of somatosensory evoked potentials (SEPs) and brain stem auditory evoked potentials (BAEPs). Visual evoked potentials (VEPs) are commonly reduced in amplitude, but not delayed. Central motor conduction velocity, reflecting pyramidal tract function, determined by cortical magnetic stimulation, is slower than normal, and, contrary to the sensory involvement, this slowing becomes more severe with increasing disease duration.

Structural neuroimaging, by computed tomography (CT) or magnetic resonance imaging (MRI), typically shows an atrophic cervical spinal cord; cerebellar atrophy is mild and commonly seen only in advanced cases. Newer MRI approaches, including tractography, may have the potential to reveal degeneration of specific fiber systems, but they have not yet been systematically evaluated in this disease. Preliminary studies with magnetic resonance specroscopy (MRS) are compatible with mitochondrial dysfunction, as were older metabolic imaging studies with fluordeoxyglucose (FDG) positron emission tomography (PET) that showed a diffusely increased FDG uptake in the brain in early FA, suggesting inadequate utilization of energy sources.

Heart disease is most commonly asymptomatic, but it may contribute to disability and cause premature death in a significant minority of patients, particularly those with earlier age of onset. Shortness of breath and palpitations are the usual initial symptoms. The electrocardiogram shows inverted T-waves in essentially all patients, ventricular hypertrophy in most, conduction disturbances in about 10%, occasionally supraventricular ectopic beats and atrial fibrillation. Echocardiography demonstrates concentric hypertrophy of the ventricles or asymmetric septal hypertrophy.

About 10% of FA patients develop diabetes mellitus, 20% have carbohydrate intolerance. The mechanisms are complex, with beta-cell dysfunction as well as peripheral insulin resistance, resembling other mitochondrial disorders.

Kyphoscoliosis may cause pain and cardiorespiratory problems. Pes cavus and pes equinovarus may further affect ambulation. Autonomic disturbances, most commonly cold and cyanotic legs and feet, become increasingly frequent as the disease advances. Sphincteric problems are rare.

The FA Gene (FRDA) and Its Mutations

The FA gene (FRDA) is localized in the proximal long arm of chromosome 9. The main mRNA has a size of 1.3 kb and corresponds to five exons, numbered 1–5a. The encoded protein, predicted to contain 210 amino acids, was called frataxin (FXN) and shown to localize in mitochondria.

The gene is expressed in all cells, but at variable levels in different tissues and during development that can only in part be accounted for by differences in mitochondrial content. In adult humans, FXN mRNA is most abundant in the heart and spinal cord, followed by liver, skeletal muscle, and pancreas. In mouse embryos, expression starts in the neuroepithelium at embryonic day 10.5 (E10.5), it reaches its highest level at E14.5 and into the postnatal period. In developing mice, the highest levels of FXN mRNA are found in the spinal cord, particularly at the thoracolumbar level, and in the dorsal root ganglia. The developing brain is also very rich in FXN mRNA, which is abundant in the proliferating neural cells in the periventricular zone, in the cortical plates, and in the ganglionic eminence. Expression is also high in the heart, liver, and brown fat.

FA is the consequence of FXN deficiency. Most patients are homozygous for the expansion of a GAA triplet repeat sequence (TRS) in the first intron, 5% are heterozygous for a GAA expansion and a point mutation in the FXN gene. Repeats in normal chromosomes contain up to ~40 triplets, 90 to >1000 triplets in FA chromosomes. As in other TRS expansion diseases, pathological alleles show meiotic and mitotic instability.

Two major hypotheses have been proposed to explain the reduced FXN expression in the presence of lengths of the GAA TRS that correspond to pathological FRDA alleles. The first hypothesis is based on

the observation that such repeats can adopt a triple helical structure in physiological conditions. Two triplexes may associate to form a novel DNA structure called 'sticky DNA', a phenomenon that may occur intramolecularly when the TRS is sufficiently long. This structure has been shown to inhibit transcription in *in vitro* and *in vivo* models, providing a mechanism to explain FXN deficiency in FA. The second hypothesis proposes that the GAA repeat, possibly by recruiting specific binding proteins, triggers modifications in DNA packaging proteins that eventually result in chromatin condensation and gene silencing. Chromatin condensation is a well-known mechanism to regulate gene expression. It is largely determined by posttranslational modification of histones, including acetylation, methylation, and ubiquitination. The GAA TRS resembles known heterochromatin-associated DNA sequences like satellite III in *Drosophila*, whose repeating sequence is GAGAA, and it may trigger such modifications by recruiting specific DNA-binding proteins. Experimental evidence supporting this hypothesis mostly comes from studies with transgenic mice. When a reporter gene is associated with GAA TRS in a construct used to create transgenic mice, it will be silenced in a proportion of cells, a phenomenon called position effect variegation that is observed in genes localized close to heterochromatic regions. Recent preliminary data (J Gottesfeld, personal communication) suggest that changes in histone acetylation and methylation indeed occur in the vicinity of the expanded GAA TRS in lymphocyte DNA from FA patients.

As expected by the experimental finding that smaller expansions allow a higher residual gene expression, expansion sizes have an influence on the severity of the phenotype. A direct correlation has been firmly established between the size of GAA repeats and earlier age of onset, earlier age when confined in wheelchair, more rapid rate of disease progression, and presence of nonobligatory disease manifestations indicative of more widespread degeneration. However, differences in GAA expansions account for only about 50% of the variability in age of onset, indicating that other factors influence the phenotype. These may include somatic mosaicism for expansion sizes, possibly variation in the TRS sequence, modifier genes, and environmental factors. In addition, about 2% of the FA chromosomes carry GAA repeat of normal length, but have a missense, nonsense, or splice site mutations ultimately affecting the FXN coding sequence. Most of these individuals have typical FA, but a few missense mutations are associated with atypical, slowly evolving disease, suggesting that the corresponding mutated proteins have some preserved function.

The Function of FXN and Pathogenesis

FXN does not resemble any protein of known function. Its amino acid sequence predicts a small soluble protein with no transmembrane domain. It is highly conserved during evolution, with homologs in mammals, invertebrates, yeast, and plants. FXN is imported into the mitochondria, where an N-terminal peptide of about 40 amino acids is removed generating the mature, functional protein.

Much information on FXN function comes from yeast studies. The yeast frataxin homolog (yfh1) is thought to share function with human FXN, not only because of high sequence similarity and corresponding mitochondrial localization, but also because the latter can fully complement a yfh1 deletion. Most yfh1 knockout (KO) yeast strains, called Δ*yfh1*, lose the ability to carry out oxidative phosphorylation and form petite colonies with defects or loss of mitochondrial DNA that cannot grow on nonfermentable substrates. Most remarkably, iron accumulates in Δyfh1 mitochondria, averaging a more then tenfold excess compared to the corresponding wild-type yeast. Loss of respiratory competence requires the presence of iron in the culture medium, and occurs more rapidly as iron concentration in the medium is increased, suggesting that permanent mitochondrial damage is the consequence of iron toxicity. Ferrous iron reacts with reactive oxygen species (ROS) and can form the highly toxic hydroxyl radical ($OH^{\cdot}$) through the Fenton reaction with H_2O_2 , a process likely to occur in Δ*yfh1* yeast, as suggested by its enhanced sensitivity to H_2O_2. In mitochondria, H_2O_2 is produced by the mitochondrial form of the enzyme superoxide dysmutase (SOD) from superoxide (O_2^-), in turn generated by the reaction of oxygen with electrons prematurely leaking from the respiratory chain. FXN deficiency likely establishes a vicious cycle involving damage to the respiratory chain, increased ROS production, and iron toxicity. Accordingly, a role of oxidative stress in the pathogenesis of FA has been proposed.

Excess iron in Δ*yfh1* yeast has been related to a marked induction (10–50-fold) of the high-affinity iron transport system on the cell membrane, normally not expressed in yeast cells that are iron replete. This induction has been recently identified as a general response to a deficit in the mitochondrial synthesis of iron–sulfur clusters (ISCs). ISCs are complexes of iron and sulfur atoms that act as prosthetic groups in several enzymes of different function and cellular localization. They are synthesized in mitochondria; cytosolic synthesis has been proposed but is somewhat controversial. ISC-containing enzymes as respiratory chain complexes I, II, and III and aconitase (AC) are impaired in

Δ*yfh1* yeast. FXN is thought to be involved in an early step of ISC synthesis, through its interaction with the scaffold protein IscU, where the first ISC assembly takes place, probably facilitating iron incorporation. Iron enters mitochondria through specific trasporters, called Mrs3 and -4 in yeast. In this process, sulfur is provided by the cysteine desulfurase Nsf1, that also appears to interact with FXN. FXN may be a mitochondrial iron chaperone with a more general function to protect this metal from ROS and making it bioavailable. *In vitro*, FXN binds Fe^{2+} with low affinity, as it would be expected for a chaperone, and has the capacity to oxidize it to the much less reactive Fe^{3+} form when its concentration increases. A mitochondrial form of ferritin (m. ferr) may also be involved in this detoxification process. Some data suggest that FXN also acts as an iron chaperone in heme synthesis, providing iron to ferrochelatase (FCH), the enzyme that inserts iron into protoporphyrin IX to form heme (Figure 1). A much higher affinity of FXN for the heme synthesis enzyme FCH than for IscU may explain why heme synthesis is resistant to low FXN level and essentially unaffected in FA patients. FXN has also been implicated in the modulation of AC activity through assembly and disassembly of its ISC.

A mouse model of FA has been difficult to generate because complete loss of FXN, as in FXN KO mice, causes early embryonic lethality. Viable mouse models have been so far obtained only through a conditional gene targeting approach. The first two models utilized Cre transgenes under the control of the muscle creatine kinase (MCK) and the neuron-specific enolase (NSE) promoters to induce striated muscle- and neuron-restricted exon deletion, respectively. NSE mutants have a low birth weight and develop a progressive neurological phenotype with the average onset of ataxia at 12 days, hunched stance, and loss of proprioception. MCK mutants show cardiac hypertrophy with thickening of the walls of the left ventricle, and show myocardial degeneration with cytoplasmic vacuolization in the myocytes, evidence of necrosis, and postnecrotic fibrosis. Loss of activity of ISC-containing enzymes is an early finding in these models. The MCK mutants accumulate iron in heart mitochondria at later stages. Using a similar conditional KO approach, but with a tamoxifen-inducible Cre recombinase under the control of a neuron-specific prion protein promoter, two different lines which exhibit a progressive neurological phenotype with slow evolution that recreates the neurological features of the human disease were developed. An autophagic process was detected in the DRG, leading to removal of mitochondrial debris and apparition of lipofuscin deposits.

Altered iron metabolism, free radical damage, and mitochondrial dysfunction all occur in FA patients, suggesting that the information derived from the investigations on FXN function and from the yeast and animal models is relevant for the pathogenesis of the human disease. Oxidative stress is revealed by increased plasma levels of malondialdheyde, a lipid peroxidation product, increased urinary 8-hydroxy-2′-deoxyguanosine, a marker of oxidative DNA damage, decreased plasma free glutathione, and increased plasma glutathione-S-transferase activity. Increased

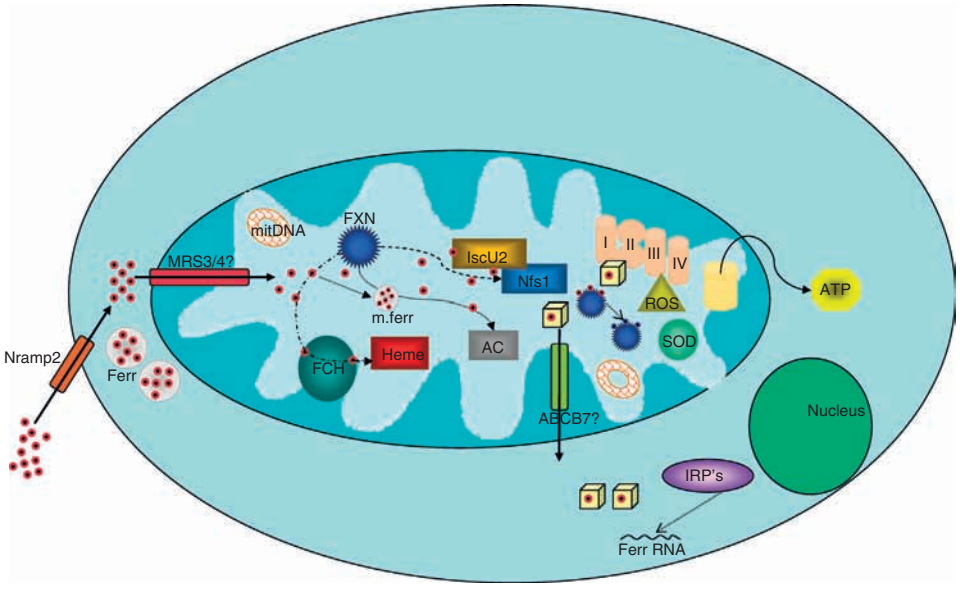

Figure 1 Potential functions of frataxin (FXN). General iron chaperone, making Fe^{2+} (●) available to FCH for heme biosynthesis, to IscU2 and Nfs1 for mitochondrial ISC (▣) biosynthesis, to mitochondrial AC. Prevent oxidative stress, protect mitochondrial proteins and mitochondrial DNA (mitDNA) from free Fe^{2+}; by converting Fe^{2+} to Fe^{3+} (●).

free radical production could be directly demonstrated in cultured cells engineered to produce reduced levels of FXN. In addition, patients' fibroblasts are sensitive to low doses of H_2O_2, which induce apoptosis at lower doses than in control fibroblasts, suggesting that even nonaffected cells are in an 'at-risk' status for oxidative stress as a consequence of the primary genetic defect. FA fibroblasts also show abnormal antioxidant responses, in particular a blunted increase in mitochondrial SOD triggered by iron and by oxidants in control cells. Intriguingly, no evidence of oxidative stress was obtained in studies on conditional KO mouse models. However, interfering factors may have affected these results, including the admixture of cells with normal FXN with progressively disappearing cells with no FXN, and the almost complete shutdown of the respiratory chain in the latter.

Mitochondrial dysfunction has been proven to occur *in vivo* in FA patients. Phosphorus-MRS analysis of skeletal muscle and heart shows a reduced rate of ATP synthesis. Finally and most importantly, the same multiple ISC-containing enzyme dysfunctions found in Δ*yfh1* yeast and in mouse models are also found in affected tissues from FA patient.

Activation of stress pathways, triggered by mitochondrial dysfunction causing energy deficit and oxidative stress, occurs in FA and likely plays an important role in cell atrophy and death. Studies on cultured PC12 cells, rat pheochromocytoma cells that can be differentiated into neurons by adding nerve growth factor (NGF), showed in particular an increased expression and activity of the MKK4-JNK kinase pathway, which may be at first a protective response but also eventually trigger apoptosis. Different vulnerable cell types may activate different pathways, as suggested by the observation of the specific occurrence of autophagic vacuoles only in primary sensory neurons in the inducible conditional KO mouse model.

All FA patients carry at least one expanded GAA repeat and therefore make an insufficient amount of otherwise normal FXN. If it were possible to increase their FXN production even to levels that are similar to those of healthy carriers, one could possibly stop the course of the disease and maybe even induce some improvement. An increased FXN production could be obtained: (1) through gene replacement therapy, that is, by introducing a FXN gene without the GAA expansion into the patient cells; (2) by giving FXN directly (the protein should, however, be modified in such a way to be able to reach the nerve cells affected by the disease and the mitochondria inside these cells); or (3) by intervening on the GAA expansion with molecules that can destabilize the triple helical structure and shift the equilibrium toward the physiological double helix that allows FXN expression. Though still in their infancy, all these appoaches are under study. Recently, encouraging results have been obtained for gene replacement therapy, with partial correction of the oxidative stress hypersensitivity of FA fibroblasts by FXN-encoding adeno-associated virus and lentivirus vectors, and for some approaches aimed to reactivate FXN transcription by using histode deacetylase inhibitors (J Gottesfeld, personal communication).

Further possibilities to treat the disease may come from the studies on the function of FXN. Based on these findings, therapeutic approaches aimed to free radical control and respiratory chain activation may be proposed. Concerning antioxidant molecules and respiratory chain stimulants, some coenzyme Q derivatives (idebenone, CoQ-10) have already given promising results, not only in experimental models, but also in clinical trials, at least on the FA cardiomyopathy. An intriguing possibility would be the identification of small molecules capable to effectively replace FXN by binding mitochondrial iron and increasing its bioavailability. A further possibility to treat FA may come from cellular therapies, in particular from the use of stem cells. However, the widespread nature of neurodegeneration in FA is a major obstacle to this approach, because it would require to diffusely deliver cells in the CNS of the patients.

The most positive note is the remarkable progress that was made in understanding the pathogenesis of FA since the responsible gene was discovered in 1996. In addition, investigating the pathogenesis of FA has stimulated research on numerous basic areas of biology, from DNA structure and biochemistry to iron metabolism. But most exciting is the now realistic perspective to develop a treatment for this sofar incurable neurodegenerative disease.

See also: Apoptosis in Neurodegenerative Disease; Fragile X Associated Tremor Ataxia Syndrome; Sporadic Degenerative Ataxias and the Dominantly Inherited Spinocerebellar Ataxias; Transgenic Models of Neurodegenerative Disease.

Further Reading

Babcock M, de Silva D, Oaks R, et al. (1997) Regulation of mitochondrial iron accumulation by Yfh1, a putative homolog of frataxin. *Science* 276: 1709–1712.

Campuzano V, Montermini L, Moltó MD, et al. (1996) Friedreich ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 271: 1423–1427.

Cossée M, Puccio H, Gansmuller A, et al. (2000) Inactivation of the Friedreich ataxia mouse gene leads to early embryonic lethality without iron accumulation. *Human Molecular Genetics* 9: 1219–1226.

Jiralerspong S, Liu Y, Montermini L, Stifani S, and Pandolfo M (1997) Frataxin shows developmentally regulated tissue-specific

expression in the mouse embryo. *Neurobiology of Disease* 4: 103–113.

Lodi R, Cooper JM, Bradley JL, et al. (1999) Deficit of *in vivo* mitochondrial ATP production in patients with Friedreich ataxia. *Proceedings of the National Academy of Sciences of the United States of America* 96: 11492–11495.

Montermini L, Richter A, Morgan K, et al. (1997) Phenotypic variability in Friedreich ataxia: Role of the associated GAA triplet repeat expansion. *Annals of Neurology* 41: 675–682.

Puccio H, Simon D, Cossée M, et al. (2001) Mouse models for Friedreich ataxia exhibit cardiomyopathy, sensory nerve defect and Fe–S enzyme deficiency followed by intramitochondrial iron deposits. *Nature Genetics* 27: 181–618.

Rötig A, deLonlay P, Chretien D, et al. (1997) Frataxin gene expansion causes aconitase and mitochondrial iron–sulfur protein deficiency in Friedreich ataxia. *Nature Genetics* 17: 215–217.

Yoon T and Cowan JA (2003) Iron–sulfur cluster biosynthesis. Characterization of frataxin as an iron donor for assembly of [2Fe–2S] clusters in ISU-type proteins. *Journal of the American Chemical Society* 125: 6078–6084.

Yoon T and Cowan JA (2004) Frataxin-mediated iron delivery to ferrochelatase in the final step of heme biosynthesis. *Journal of Biological Chemistry* 279: 25943–25946.

Frontal Cortex Evolution in Primates

E C Bush, Harvey Mudd College, Claremont, CA, USA
J M Allman, California Institute of Technology, Pasadena, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Primate-Specific Adaptations

The frontal cortex can be defined as the neocortex anterior to the motor somatosensory–cortex border. This is a large region in primates, containing areas involved directly or indirectly in the control of almost every behavior. It has long been thought that the frontal cortex played an important role in primate evolution. Modern evidence supports this view. Indeed, given the diversity of functions in the region and the variety of unique behaviors exhibited by primates, it would be surprising if the frontal cortex had not evolved unique adaptations.

The common ancestor of eutherian mammals probably had a small body and a small brain. Comparative work in modern mammals suggests it had a basic complement of cortical areas including primary visual, auditory, and somatosensory areas. It also probably had a primary motor area (M1). This means that we could define a frontal cortex in that mammal and that the region of the frontal cortex in modern eutherians (taken as a whole) can be thought of as being homologous.

It is likely however that this broad homology obscures substantial differences in frontal cortex structure between primates and nonprimates. One piece of evidence to this effect is that the frontal cortex scales differently in primates and nonprimates. In primates, the frontal cortex hyperscales with the brain size. This can be seen in the reconstructions in Figure 1 which show the brain of a small primate, the galago, and a larger primate, the macaque. A primate with a larger brain tends to have a disproportionately large frontal cortex. In contrast, in a nonprimate order, carnivores, the frontal cortex size does not vary systematically with brain size. This suggests that the structure and development of the frontal cortex differs substantially in the two orders.

When we focus on specific cortical areas and regions, we again find that the primate frontal cortex differs in important ways from that found in other orders. A variety of evidence suggests that the two main branches of primates, strepsirrhines and anthropoids, share up to ten motor areas in the frontal cortex. Of these, only one or two have clear homologs outside primates. Most eutherians have an agranular M1, with large layer 5 pyramidal cells, somatotopy, and relatively low thresholds of stimulation. Rostral and medial to M1 is a supplementary motor area (SMA, also called M2) in primates. It is also somatotopically organized, and has slightly smaller layer 5 pyramidals than M1. This area may have homologs in other mammals such as rats.

The remaining motor areas which are shared among primates appear not to have one-to-one homologs in other orders. Among these areas are several cingulate motor areas and a number of premotor areas involved in higher-level coordination of movement. Several of these are particularly interesting. The frontal eye field (FEF) is connected with both the sensory visual areas and the prefrontal cortex. Microstimulation here produces saccadic and smooth pursuit eye movements. The FEF may be involved in the voluntary control of eye movements, an attribute of great importance for highly visual primates. Strepsirrhines and haplorhines also share a ventral premotor area (PMV), which appears to have representations of the upper body. One possibility is that PMV is involved in the visually guided grasping of objects, something that would be especially important to a small visual predator. Early primates are thought to have been small arboreal visual predators, and it is tempting to view the evolution of primate motor areas in this light. These areas could be seen as motor adaptations of the earliest primates for a very demanding arboreal niche.

A frontal region of particular interest in primate evolution is the prefrontal cortex. This region has a wide variety of connections with other brain regions, and patients with prefrontal damage show deficits in the ability to plan and organize behavior. This suggests that it is involved in aspects of behavior which have been important in primate evolution.

The question of whether the prefrontal cortex is unique to primates has been controversial. Brodmann originally argued that it was, identifying prefrontal cortex with the well-developed granular layer 4 which is seen in primates. This was later disputed based on connectivity data. Today, there is general agreement that neither of these two types of data alone provides an adequate answer. However, there is less consensus on whether the prefrontal cortex is unique to primates. What is clear, however, is that even if a broadly homologous prefrontal cortex exists in other mammals, it is substantially different from that found in primates. There is no evidence that rats possess anything like the diversity of prefrontal areas found in primates, and many of the details of the primate prefrontal cortex are likely to be unique to the order.

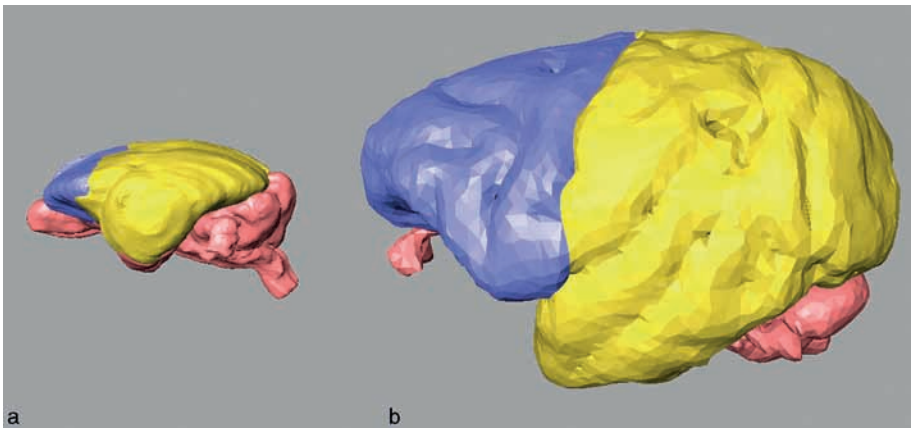

Figure 1 Reconstructions of the brains: (a) *Galago senegalensis*; (b) *Maccaca mulatta*. Frontal cortex is indicated in blue. The position of the motor-somatosensory cortex border was used to delineate frontal cortex. The two brains are shown here in their proper proportions relative to one another.

Specializations within Primates

Anthropoids versus Strepsirrhines

There is also important variation in frontal cortical structure between the two large divisions of primates, anthropoids and strepsirrhines (represented by the macaque and galago, respectively). Among the premotor areas we have discussed, area PMV appears to have developed interesting adaptations in anthropoids. In monkeys, it seems to have an increasing representation of distal forelimb movements. At the same time, it has strong connections to parts of somatosensory cortex which represent information from the cutaneous receptors of the hands. Such information is particularly useful for reaching and grasping motions. These differences have led to the suggestion that anthropoid PMV evolved mechanisms for the improved guidance of reaching and grasping.

In terms of the broader arrangement of areas, however, the premotor cortices in the two groups have many similarities. As already mentioned, electrophysiological data suggest that the motor and premotor regions of anthropoids and strepsirrhines share up to ten areas. Architectonic evidence supports this view. This broad similarity in the arrangement of areas can be found in several other frontal regions. Comparisons in cytoarchitecture and myeloarchitecture suggest that anthropoids and strepsirrhines have a similar arrangement of areas in the orbital and medial regions of the frontal cortex.

In contrast, it seems that there are important differences between the two groups in the prefrontal cortex. Macaques appear to have a much larger number of areas in this region than galagos do. This is suggested by patterns of connectivity between the prefrontal cortex and other regions. Labeling experiments show that the prefrontal cortex in both species has numerous connections with other parts of the cortex. But the pattern of connectivity suggests that galago has fewer prefrontal areas. For example, tracer injections in regions of the parietal cortex produce fewer discrete zones of labeling in the galago prefrontal cortex than in the macaque. A similar result has been found in comparisons of architectonic parcelations in the two species.

Current architectonic and connectional data suggest that most prefrontal areas in galagos can be identified with a homologous area in the macaque. For example, the posterior-most prefrontal areas in galago appear to have homologs in the macaque arcuate cortex region. These conclusions reflect similarities in cell density and size, fiber-staining patterns, and connections to the parietal, superior temporal, and inferotemporal cortices. Similar conclusions have been reached about other areas in the galago prefrontal cortex (e.g., areas in the superior, polar, and ventral parts of the prefrontal cortex).

However, there are a number of areas in the macaque which have no clear homolog in galagos. These areas are concentrated especially around the principal sulcus in macaques, and they may represent anthropoid-specific specializations. This region of the cortex, often referred to as the dorsolateral prefrontal cortex, is thought to be involved in working memory. Classic lesion studies showed that monkeys with lesions in this region have deficits on delayed-response tasks. Such tasks may involve spatial working memory, for example, requiring the monkey to remember a physical position where food was last given. They can also involve nonspatial problems which require the monkey to remember an object's identity over a delay. Working memory of this type is thought to

play an important role in cognition, and in light of this, it is especially interesting that the dorsolateral prefrontal cortex appears to have anthropoid-specific specializations. Also, it is interesting to note that area PMV has developed connections to the dorsolateral prefrontal cortex which are especially prominent in Old World monkeys.

Great Apes and Humans

The fronto-polar cortex (Brodmann's area 10) expands in apes and especially in humans both absolutely and relative to total cortical size. Functional imaging studies indicate that area 10 is activated in the retrieval of episodic memory, in the receipt of monetary rewards, in weighing cost versus benefit, in the formulation of auction bids, in deciding how much to spend to punish cheaters, and in moral decision making. The retrieval of specific past episodes contributes to the complex deliberative socioeconomic decision making that this structure participates in. The slow deliberative nature of this form of cognition stands in contrast to rapid intuition, which another hominoid specialization, the von Economo neurons (VENs) may participate in.

The VENs are large bipolar cells located in anterior cingulate (aCC) and fronto-insular (FI) cortex. They are distinguished from pyramidal cells because they have only a single, large, basal dendrite, whereas pyramidal cells have an array of smaller basal dendrites extending from the cell body. The VENs are present only in humans and great apes and are far more abundant in humans than in apes.

The apical dendrites of VENs are very similar to those of the apical dendrites of neighboring pyramidal cells. The radial orientation and narrow width of the dendritic arborization indicate that the VENs sample a sharply circumscribed cylinder of cortex, possibly corresponding to a minicolumn. They may thus constitute a fast-fire output from minicolumns that provides a rapid relay to other parts of the brain. VEN functions are revealed by immunocytochemical staining with antibodies to neurotransmitter receptors. The VENs are strongly labeled with antibodies to the dopamine D3 receptor, which may signal the expectation of reward under uncertainty. The activation of the FI and aCC increases with the degree of uncertainty. FI and aCC activity is coupled to situations in which the subject sustains a gambling loss (punishment) and then switches to a different behavioral strategy, implying that in normal subjects these areas are involved in adaptive decision making and cognitive flexibility.

The VENs may participate in intuition, a form of cognition in which many variables are rapidly evaluated to yield a fast decision. Typically we are unaware of the logical steps or assumptions underlying the process although intuition is based on experience-based probabilistic models. We experience the intuitive process at a visceral level. Intuitive decision making enables us to react quickly in situations that involve a high degree of uncertainty, which commonly involve social interactions. Frequently we do not have the luxury of sufficient time to perform a deliberative cost-benefit analysis to determine the most appropriate course of action but, instead, must rely on rapid intuitive judgments. The aCC and FI are active when subjects make decisions under a high degree of uncertainty. These areas are also active when subjects experience guilt, experience embarrassment, and engage in deception. The aCC and FI are also active in humor, trust, empathy, and the discrimination of the mental states of others. All these social emotions are influenced by the degree of uncertainty involved. Their large size suggests that the VENs may relay a fast intuitive assessment of complex social situations to allow the rapid adjustment of behavior in quickly changing social situations. For example, humor, which activates the FI and aCC in proportion to subjective ratings of funniness, may serve as a way to recalibrate intuitive judgments in changing social situations, thus resolving uncertainty and relieving tension. The VENs can thus be seen as an adaptation supporting the increased complexity of hominoid and especially human social networks.

See also: Brodmann's Areas; Frontal Eye Fields; Humans; Orbitofrontal Cortex: Visual Functions; Prefrontal Cortex: Structure and Anatomy; Prefrontal Contributions to Reward Encoding; Prefrontal Cortex; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Allman JM, Hakeem A, and Watson K (2002) Two phylogenetic specializations in the human brain. *Neuroscientist* 8: 335–346.

Allman JM, Watson K, Tetreault N, and Hakeem A (2005) Intuition and autism: A possible role for von Economo neurons. *Trends in Cognitive Sciences* 9: 367–373.

Bush EC and Allman JM (2004) The scaling of frontal cortex in primates and carnivores. *Proceedings of the National Academy of Sciences of the United States of America* 101: 3962–3966.

Kaas JH and Preuss TM (2003) Human brain evolution. In: Squire L, Bloom F, Landis SR, and Zigmond M (eds.) *Fundamental Neuroscience,* 2nd edn., pp. 1147–1166. San Diego, CA: Academic Press.

Nimchinsky EA, Gilissen E, Allman JM, Perl DP, Erwin JM, and Hof PR (1999) A neuronal morphologic type unique to humans and great apes. *Proceedings of the National Academy of Sciences of the United States of America* 96: 5268–5273.

Preuss TM and Goldman-Rakic PS (1991) Ipsilateral cortical connections of granular frontal-cortex in the strepsirhine primate galago, with comparative comments on anthropoid primates. *Journal of Comparative Neurology* 310: 507–549.

Preuss TM and Goldman-Rakic PS (1991) Myeloarchitecture and cytoarchitecture of the granular frontal-cortex and surrounding regions in the strepsirhine primate Galago and the anthropoid primate Macaca. *Journal of Comparative Neurology* 310: 429–474.

Semendeferi K, Armstrong E, Schleichter A, Zilles K, and Van Hoesen G (2001) Prefrontal cortex in humans and apes: A comparative study of area 10. *American Journal of Physical Anthropology* 114: 224–241.

Stepniewska I, Preuss TM, and Kaas JH (2006) Ipsilateral cortical connections of dorsal and ventral premotor areas in new world owl monkeys. *Journal of Comparative Neurology* 495: 691–708.

Watson K, Matthews B, and Allman J (2007) Brain activation during sight-gags and language-dependent humor. *Cerebral Cortex* 17(2): 314–324.

Wu CW, Bichot NP, and Kaas JH (2000) Converging evidence from microstimulation, architecture, and connections for multiple motor areas in the frontal and cingulate cortex of prosimian primates. *Journal of Comparative Neurology* 423: 140–177.

Frontal Eye Fields

J D Schall, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Location and Connections

The frontal eye field (FEF) was described first in the late 1800s by David Ferrier, who discovered that electrical stimulation of a portion of the frontal lobe of macaque monkeys elicited movements of the eyes. Today we know that the FEF exists in all primates, including humans. A homolog of the FEF has been described in cats and owls too. Thus, the FEF is present in animals that employ visually guided conjugate movements of the eyes or head to locate objects to approach or avoid.

In humans, the FEF is located in the rostral bank of a portion of the precentral sulcus at the caudal end of the middle frontal gyrus (Figure 1(b)). In macaque monkeys, the animal of choice for investigations of this part of the brain, the FEF occupies the rostral bank of the arcuate sulcus (Figure 1(a)). It occupies the caudal end of the granular prefrontal cortex, with the agranular premotor cortex making up its caudal boundary in the fundus of the arcuate sulcus.

The distinctive cytoarchitecture of the FEF distinguishes it from the surrounding cortex. It contains a high concentration of large pyramidal cells in layer V. This cortex is also characterized by a thinner granular layer 4 compared to the more rostral cortex. The ventrolateral portion of this zone, designated area 45, is distinguished by the presence of large pyramidal cells in layer 3 as well as layer 5. The dorsomedial zone, referred to as area 8A, can be further parcellated based on cytoarchitectonic differences. The cortex within the medial portion of the rostral bank of the arcuate sulcus contains fewer large pyramidal cells in layer 3 and a loosely organized granular layer; it is referred to as area 8Ac. Rostral to area 8Ac medially and area 45 laterally, occupying the convexity of the prearcuate gyrus is a transitional zone, designated area 8Ar, with fewer large pyramidal cells and a thicker, more clearly delineated layer 4. Area 8Ar forms the caudal boundary of area 46 and should probably be considered functionally distinct from the FEF. Curiously, in humans the FEF has been described as being in Brodmann's area 6. However, recent analyses have indicated that this difference between species is more about labels than cortical architecture.

The FEF influences saccade production through four major pathways: (1) a projection to the ipsilateral superior colliculus concentrated in the intermediate layers, (2) a pathway through the basal ganglia via the ipsilateral striatum, (3) a projection to the cerebellum via the pontine nuclei, and (4) a weaker projection to mesencephalic and pontine nuclei that make up the saccade generator circuit (Figure 2). The FEF is innervated by nuclei in the thalamus bordering the internal medullary lamina, mainly the lateral part of the mediodorsal nucleus and the medial part of the ventroanterior nucleus. The thalamic zones most heavily connected with the FEF are themselves innervated by oculomotor afferents from the intermediate layer of the superior colliculus, the substantia nigra pars reticulata, and the dentate nucleus of the cerebellum.

The FEF is connected with diverse cortical areas. Within the frontal lobe, the FEF is interconnected with the supplementary eye field, with prefrontal areas 46 and 12, and weakly with anterior cingulate area 24 and the postarcuate premotor cortical areas.

The FEF also receives abundant inputs from a multitude of extrastriate visual cortical areas in both the dorsal and ventral streams (Figure 3). In fact, the FEF is unique in the extent of its connectivity with extrastriate visual cortex. Moreover, the FEF provides reciprocal connections to the same diverse areas. Thus, the FEF can influence the activation of neurons in the extrastriate visual cortex. Recent evidence and current models identify the FEF as a source of top-down control on visual processing.

The connectivity of the FEF with visual areas caudal to the central sulcus is topographically organized. The more ventrolateral portion of the FEF, which is responsible for generating shorter saccades, is interconnected with the perifoveal representation in retinotopically organized areas, from areas that represent central vision in the inferotemporal cortex and from other areas having no retinotopic order. In contrast, the mediodorsal FEF, which is responsible for generating longer saccades, is interconnected with the peripheral visual field representation of retinotopically organized areas, from areas that emphasize peripheral vision or are multimodal and from other areas that have no retinotopic order.

Role in Producing Movements of the Eyes

The FEF is a key part of the ocular motor system. Most evidence has been obtained through invasive studies with nonhuman primates, but the human FEF has been located through brain imaging, transdural recording and stimulation, and transcranial magnetic stimulation. The FEF contributes to the generation of rapid saccadic gaze shifts as well as

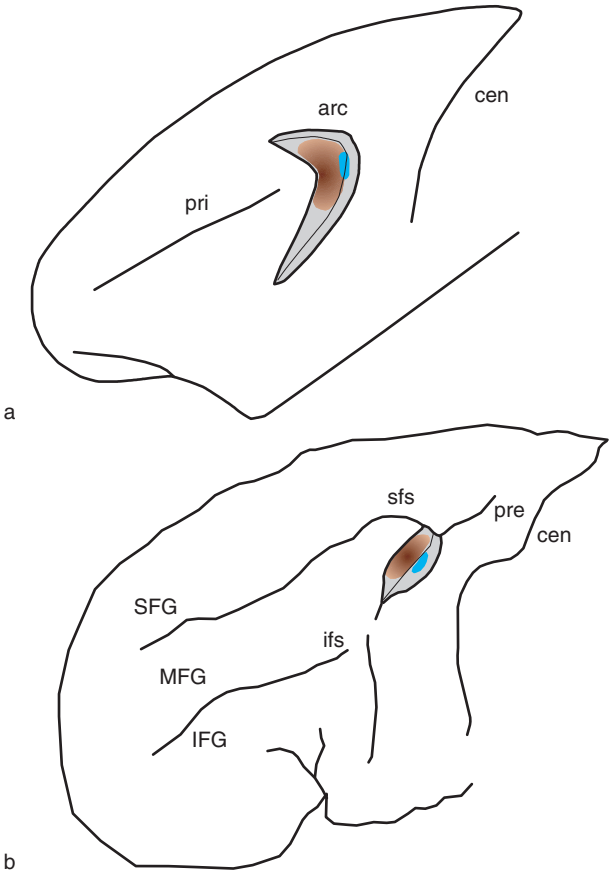

Figure 1 Stylized lateral view of the frontal cortex: (a) of a macaque monkey; (b) of a human. The macaque arcuate sulcus is opened to expose the frontal eye field (brown) on the rostral bank and the frontal pursuit zone (blue) on the fundus. The human precentral sulcus just ventral to the superior frontal sulcus is opened to expose the frontal eye field (red) on the caudal end of the middle frontal gyrus with the pursuit zone (blue) in the fundus. arc, arcuate sulcus; cen, central sulcus; IFG, inferior frontal gyrus; ifs, inferior frontal sulcus; MFG, middle frontal gyrus; pre, precentral sulcus; pri, principal sulcus; SFG, superior frontal gyrus; sfs, superior frontal sulcus.

slow pursuit eye movements and probably also vergence eye movements. In macaques, the zone contributing to the production of smooth pursuit eye movements is located at the fundus of the sulcus, apparently somewhat separate from the area of the FEF that is involved in saccade production (Figure 1(a)).

Electrical Stimulation

Low-intensity electrical stimulation of the FEF in the bank of the arcuate sulcus in monkeys elicits saccadic eye movements directed contralateral to the stimulated hemisphere. Stimulation of the frontal pursuit zone in the fundus elicits pursuit eye movements directed ipsilateral to the stimulated hemisphere. Currents less than 50 μA are sufficient to elicit eye movements, and the movements begin ~30 ms following the stimulation. The saccades elicited by stimulation of a particular site in the FEF have a particular direction and amplitude that does not vary with the orientation of the eyes in the head. The direction and amplitude of evoked saccades varies gradually as stimulating electrodes are moved through the FEF. The amplitude of the evoked saccade increases laterally to medially along the arcuate sulcus (Figure 4(a)). The FEF also contributes to orienting movements of the pinna in macaques. Recent research has also shown that stimulation of the FEF in monkeys with unrestrained heads can evoke coordinated eye and head gaze shifts.

Inactivation and Ablation

Reversible inactivation of parts of the map in the FEF impairs the production of both visually guided and memory-guided saccades. Monkeys initiate fewer saccades to the retinotopic representation of the inactivated FEF site than to others location in the visual field, and the saccades that are produced are initiated with longer latencies, have lower velocities, and have larger targeting errors than normal. These effects are more pronounced for memory-guided saccades. Also, when tested with memory-guided saccades, monkeys produce more premature saccades to targets in the hemifield ipsilateral to the inactivation. The results from inactivation studies show that the FEF contributes to generating contraversive saccades, especially to locations of remembered targets, maintaining fixation and suppressing inappropriate ipsiversive saccades.

These results complement earlier observations that ablation of the FEF causes an initially severe impairment in saccade production that recovers in some but not all respects over time. If the fundus of the arcuate sulcus is ablated, deficits in pursuit occur; the deficits are most evident in producing pursuit that anticipates the movement of predictable stimuli. Clinical signs of FEF damage in humans include deviation of gaze toward the side of the damage and impaired ability to produce saccades that are more voluntary in character, such as a gaze shift opposite a stimulus (antisaccade) or a saccade to a remembered location.

Functional Brain Imaging

In humans as well as macaque monkeys, the FEFs exhibit increased blood flow and oxygen use when saccadic and pursuit eye movements are produced. The functional magnetic resonance imaging (fMRI) studies locate the human FEF in the precentral sulcus. High-resolution scanning indicates that the upper

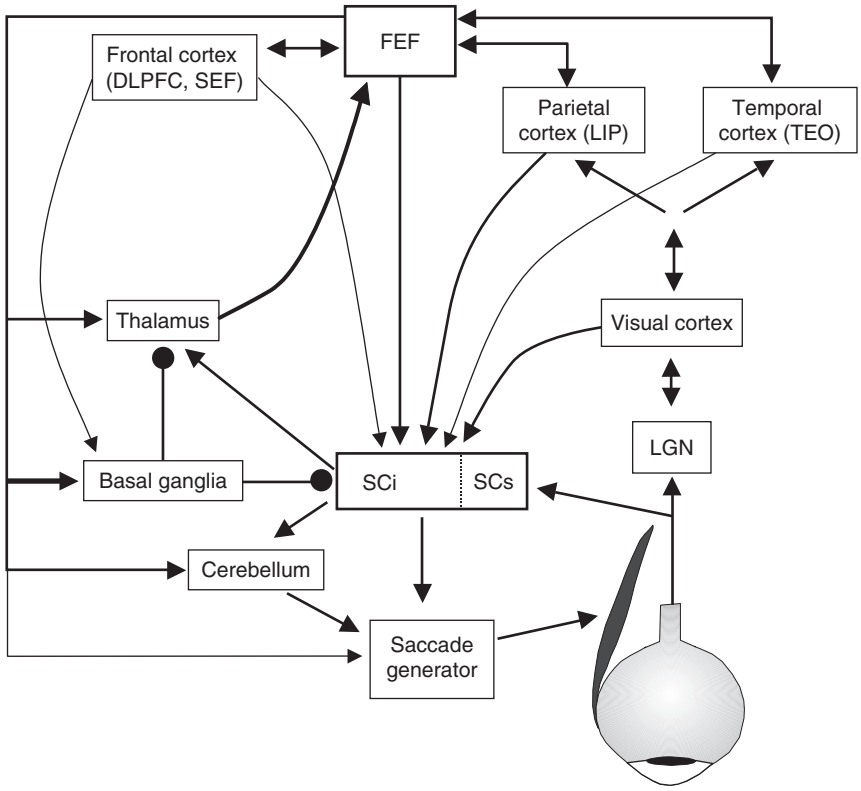

Figure 2 Schematic of brain structures and some of the anatomical pathways involved in saccade generation. Visual input from the retina in the eye is transmitted to the lateral geniculate nucleus (LGN) and the superficial layers of the superior colliculus (SCs). The LGN transmits visual signals to primary visual cortex, where processing proceeds through areas in the temporal lobe (such as area TEO) and the parietal lobe (such as area LIP). Extrastriate visual areas are reciprocally connected with the frontal eye field. The frontal eye field is also connected with areas in the frontal lobe such as dorsolateral prefrontal cortex (DLPFC) and the supplementary eye field (SEF). The frontal eye field influences saccade production through projections to the intermediate layers of the superior colliculus (SCi), the basal ganglia, the cerebellum, and the mesencephalic and pontine brain stem saccade nuclei. FEF, frontal eye field; LIP, lateral intraparietal area; TEO, area TEO.

portion of the anterior wall of the precentral sulcus contributes to saccade production and that a deeper region along the anterior wall extending in some individuals to the fundus or even the deep posterior bank of the sulcus contributes to pursuit production. This localization shows a high degree of homology in the organization of the FEFs in humans and macaque monkeys.

Neuronal Recordings

The influence of the FEF on saccade production is mediated by neurons that are activated specifically before and during saccades. Two kinds of neurons that control gaze have been distinguished. In general, movement neurons contribute to gaze shifting and fixation neurons contribute to gaze holding (Figure 5(b)). In other words, an increase in the discharge rate of movement neurons increases the probability that a saccade will be produced. In contrast, elevated discharge rates of fixation neurons occur when gaze is fixed on a stimulus of interest and reduced discharge rate increases the probability of saccade initiation. Current models of saccade production conceive of movement and fixation neurons as producing signals that are in a reciprocally inhibitory relationship.

Neurons in the FEF that generate movement-related or fixation-related activity are located in layer 5 and innervate the superior colliculus and parts of the neural circuit in the brain stem that generate saccades. Indirect evidence suggests that fixation neurons are smaller than movement neurons. Physiological recordings indicate that these neurons, in concert with a network that includes the superior colliculus, produce signals necessary to produce saccadic eye movements.

Fixation neurons are modulated during saccades in all directions. Movement neurons, on the other hand, are active before saccades of a particular range of directions and amplitudes, referred to as the movement field. The movement fields of neurons in the lateral aspect of the FEF are less eccentric and smaller,

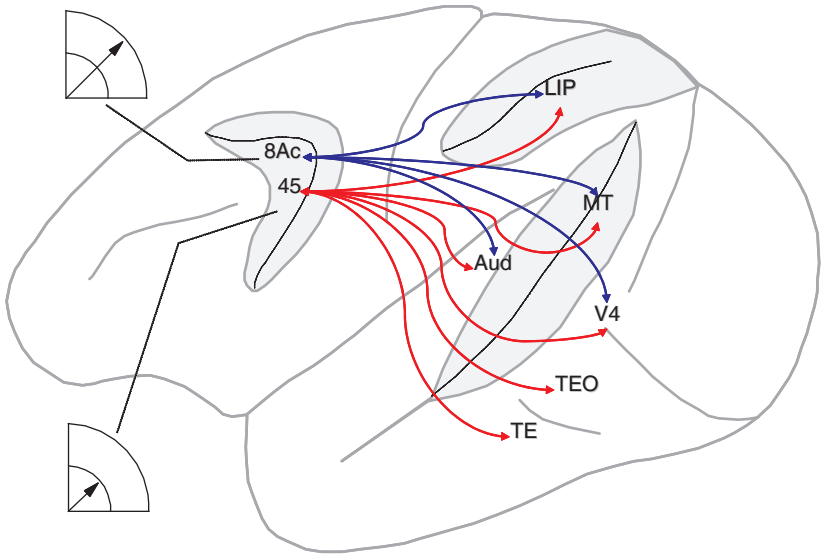

Figure 3 Selective summary of connections of the frontal eye field with extrastriate visual cortical areas. Red lines indicate connections of the ventrolateral frontal eye field in area 45, and blue lines mark connections of the dorsomedial frontal eye field in area 8Ac. Also indicated schematically are connections with belt and parabelt auditory areas (Aud). LIP, lateral intraparietal area; MT, middle temporal area; TE, area TE; TEO, area TEO, V4, area V4.

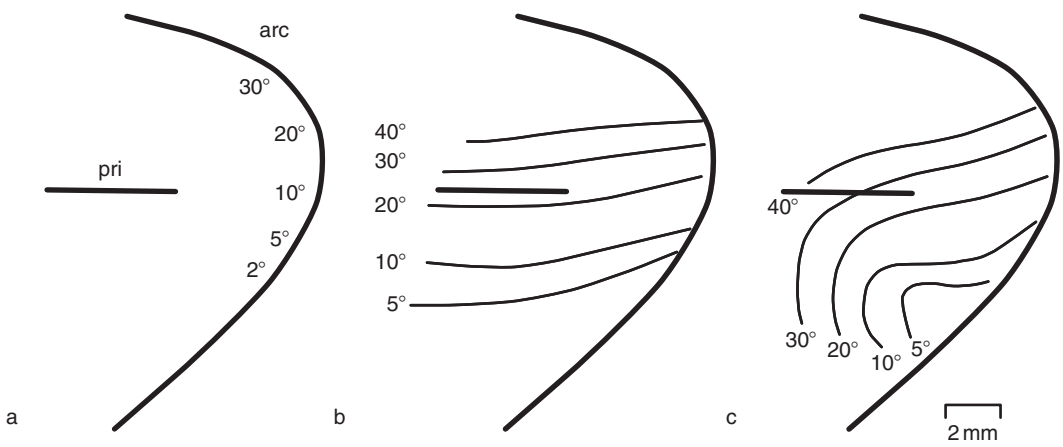

Figure 4 Maps of the frontal eye field and rostrally adjacent cortex: (a) saccade amplitude; (b) visual receptive field eccentricity; (c) receptive field size. arc, arcuate sulcus; pri, principal sulcus.

and movement field eccentricity and size increase medially in the FEF. In direction relative to current fixation, the movement fields are Gaussian shaped. In amplitude, they exhibit a log-Gaussian shape; this means that the outer border of the receptive field extends further than the inner border from the point of peak activity. The more eccentric the movement field, the further the outer border extends. This arrangement corresponds to the mapping of the visual field in the visual pathways and the mapping of saccade direction and amplitude in the superior colliculus.

Saccades are initiated when the activity of fixation neurons decreases and the activity of movement neurons reaches a threshold (Figure 5(b)). Variability in saccade latency can be accounted for by the time taken to reach this threshold. In general, the form of the activation of these neurons in the FEF corresponds to diffusion processes employed in models of response time. If an interrupting stimulus occurs, saccade preparation is canceled if and only if the activity of these neurons is reduced and prevented from reaching the threshold.

Each movement of the eyes causes a shift of the retinal image. A long-standing problem has been understanding how the brain registers whether a movement in the retinal image is due to a movement in the world or a movement of the eyes. The FEF may contribute to maintaining an up-to-date representation

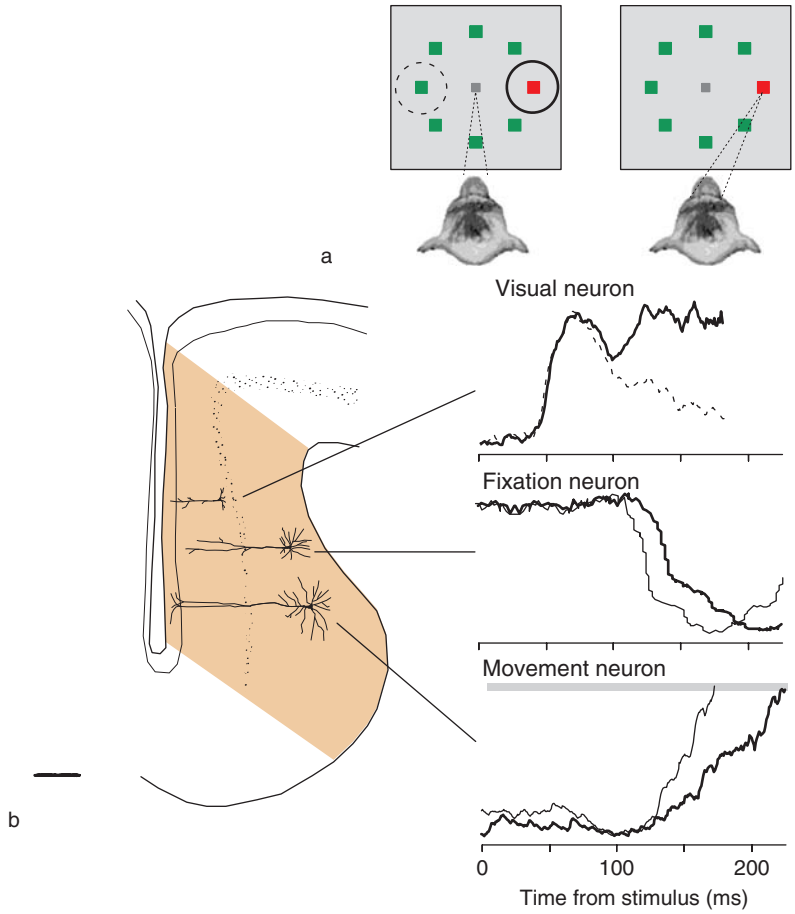

Figure 5 Target selection and saccade preparation by neurons in the frontal eye field: (a) performance of a visual performance task; (b) location and some neural constituents of the frontal eye field and graphs of modulating discharge rates of the main types of neurons found in the frontal eye field. In the visual search task in (a), the gaze must be shifted directly to the uniquely colored target among distractors. Targets for saccades are selected by the pattern of activity of visually responsive neurons in the frontal eye field as part of a distributed network. In (b), the frontal eye field is shown to be distinguished from the surrounding cortex by the large pyramidal cells in layers 3 and 5. In layer 5, pyramidal cells vary in size from smaller to larger. It is likely that each morphological cell type corresponds to a different functional cell type, as indicated in the graphs. The top neural activity graph illustrates the response to the target in the receptive field (thicker solid line) and the response to a distractor (thinner dotted line), superimposed for comparison. After around 100 ms, a selection process transpires that results in an accurate representation of the location of the target that can be used to guide saccade generation. Saccades are produced when the activity of gaze-holding fixation neurons in the frontal eye field, as part of a network, decreases while the activity of gaze-shifting movement-related neurons in the frontal eye field, as part of a network, increases to reach a fixed threshold. Variability in the time of initiation of the saccade originates in the variable time taken for the activity to increase to the threshold. The relation of activity in the frontal eye field to saccade latency is illustrated by portraying activity on trials with shorter (thin) and longer (thick) saccade latencies.

of eye orientation in the head. This is necessary to direct the gaze correctly when confronted with unexpected perturbations resulting in a mismatch between the retinal locus of a target and the angle of the saccade needed to fixate that target.

Role in Selecting Targets and Allocating Attention

Since the discovery of the FEF, research has focused on the contributions of the FEF to the generation of movements of the eyes. In the last 10 years, a wealth of data has demonstrated that the FEF contributes, as well, to the allocation of covert attention.

Neuronal Recordings

By far, most of the neurons sampled by extracellular recordings in the FEF have visual responses. The neurons are not selective for stimulus features except for some particular exceptions (noted later). The receptive fields of FEF neurons are restricted primarily to the contralateral hemifield and scale in size with

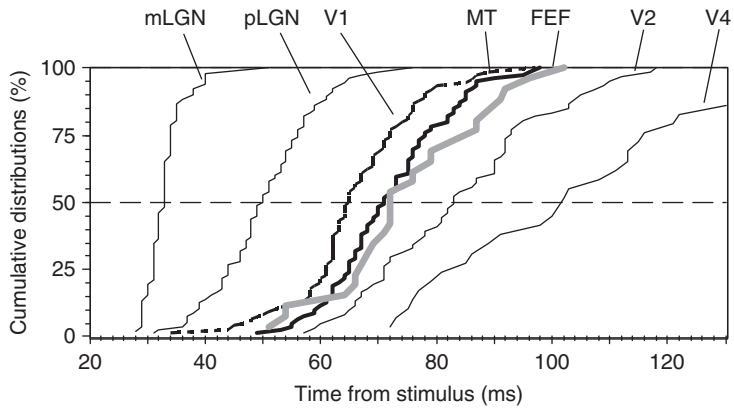

Figure 6 Timing across the visual pathway. Cumulative distributions of visual response latencies are compared for neurons in the magnocellular layers of the dorsal lateral geniculate nucleus (mLGN), neurons in the parvocellular layers of the dorsal lateral geniculate nucleus (pLGN), primary visual cortex (V1), middle temporal visual area (MT), frontal eye field (FEF), visual area 2 (V2), and visual area 4 (V4).

eccentricity. Less eccentric receptive fields are smaller than more peripheral receptive fields. The FEF has a rough map of visual field eccentricity with the central visual field represented laterally and progressively more eccentric visual field represented more medially (Figure 4(b)). This topography arises from the pattern of inputs from the extrastriate visual cortex.

The visual responses occur with latencies comparable to those observed in most extrastriate visual areas, averaging around 70 ms (Figure 6). Furthermore, the neural signals occurring in the FEF coincide with identical signals occurring in a network of interconnected structures, including the superior colliculus and posterior parietal cortex. Recent evidence indicates that FEF neurons may exhibit some selectivity for disparity and direction of motion. Some neurons in the FEF also respond to acoustic stimuli located preferentially in the contralateral hemifield. These responses contribute to orienting to sound sources.

The FEF contributes to selecting the target and shifting attention before gaze shifts, both saccadic and pursuit. In macaque monkeys trained to shift gaze to the odd-ball target in visual search arrays, most visually responsive cells in the FEF responded initially indiscriminately to the target or the distractor of the search array in their receptive field because neurons in the FEF are not intrisinically feature selective (Figure 5(a)). However, before the eyes move, a selection process transpires by which most visually responsive cells in the FEF signal the location of the target stimulus through suppression of the response to nontarget stimuli, leaving only the neurons responding to the target active.

This target selection process occurs if no saccade is made and the target location is signaled by a manual response. The selection process also occurs even if the saccade is directed to a nontarget stimulus. The selection process is influenced by the similarity of the target and nontarget stimuli; if the target is more difficult to distinguish from distractors, then the selection process takes more time and accounts for more of the variability of saccade latency.

The target selection process is influenced by knowledge of target properties acquired over preceding trials or sessions. At one extreme, if monkeys are trained exclusively on one target–distractor pairing, then many neurons in the FEF can appear to acquire immediate selectivity for the overpracticed target. At the other extreme, if monkeys perform a search for one target–distractor pair (e.g., red among green) and then switch to the opposite pair (green among red), they are slower and more error-prone on the first few trials after the switch. Concomitantly, the target selection process in the FEF is later and less certain on the first few trials after the switch.

To investigate more directly the relationship between the allocation of attention and preparation of a saccade, it is advantageous to dissociate at least momentarily the focus of attention from the endpoint of a saccade. This can be accomplished by training macaque monkeys to perform a visual search for an attention-capturing pop-out color target in a search array and then shifting gaze either toward (prosaccade) or opposite (antisaccade) the pop-out target, according to its orientation. Most visually responsive neurons in the FEF initially select the pop-out target while attention is allocated to distinguish its shape, but then these neurons display a strong modulation of the discharge rate whereby they come to select the ultimate endpoint of the saccade. This modulation can be identified with the shift of attention from the pop-out target to the endpoint of the saccade preceding the gaze shift.

These observations motivate the hypothesis that attention is allocated when and to the extent that these neurons in the FEF (and elsewhere in a coordinated network) are active for an object at one (as opposed to any other) location in the visual field. Mechanistically, it can be surmised that these neurons in the FEF (and elsewhere in a coordinated network) make up a salience map guiding covert and overt orienting.

Functional Brain Imaging

Numerous fMRI studies have demonstrated that the FEF and surrounding cortex are involved in the allocation of spatial attention. In fact, current thinking holds that the FEF and areas in posterior parietal cortex make up a network that mediates covert and overt orienting. However, the frontal and parietal areas have distinct roles with the FEF, concerned more with encoding the target of the saccade as such and preparing to execute the saccade, and the parietal cortex, concerned more with representing the visual stimuli.

Several studies have demonstrated the involvement of the FEF in both top-down and bottom-up control of attention. These studies present a cue (such as a small arrow) that instructs observers about some aspect of the forthcoming stimulus (such as the location or visual feature of a target stimulus). While attention is allocated in preparation for the stimulus, activation is observed in the FEF, predominantly in the hemisphere contralateral to the attended location. Studies have also demonstrated fMRI activation in the FEF of participants searching for salient stimuli. Activation is also observed in the FEF when participants maintain a working memory representation. The area of activation overlaps substantially with that activated when saccades are produced.

Inactivation and Ablation

A role for the FEF in the cognitive operations of working memory and allocation of attention is also shown by the effects of ablation or inactivation. Ablation of the FEF in monkeys results in severe deficits in locating a target presented among distractors or maintaining a representation of target location during a delay period. Reversible inactivation of the FEF impaired their detecting a visual target presented among distractors when no eye movements were permitted.

It should be noted that the effects of inactivation of the FEF on search performance are notably different from those seen following inactivation of the lateral intraparietal area in the posterior parietal cortex. Whereas inactivation of the lateral intraparietal area produced deficits that varied in magnitude with task difficulty, inactivation of the FEF produced severe deficits in visual search for single targets that were easy or hard to distinguish from the distractors as well as in search for a target defined by a conjunction of features.

Electrical Stimulation

The contribution of the FEF and surrounding cortex to the allocation of spatial attention has been shown very convincingly using intracortical electrical stimulation in monkeys and transcranial magnetic stimulation in humans. In monkeys trained to locate a single target among conspicuous distractors, stimulation of the FEF increased the probability of locating the target if it was located within the movement field of the site of stimulation. Other work has indicated that this influence may be mediated through a direct influence on the activation of neurons in visual area 4 (V4). In humans, transcranial magnetic stimulation over the FEF can affect attention allocation and awareness of stimuli. Curiously, these effects in humans tend to be specific to the right hemisphere.

These results have been cited in support of the premotor theory of attention, which states that shifting visual spatial attention corresponds to preparing a saccade. The FEF is an ideal region in which to test this theory, being a central node in the network that selects targets and produces saccades. As previously described, the focus of attention can be dissociated momentarily from the endpoint of a saccade when monkeys search for an attention-capturing pop-out color target in a search array and then shift gaze either toward (prosaccade) or opposite (antisaccade) the pop-out target according to its orientation. An experiment used this task and probed the evolution of saccade preparation by measuring the direction of saccades evoked by intracortical microstimulation of the FEF at different times following the search array. The saccades evoked on prosaccade trials deviated progressively toward the target that was the endpoint of the saccade. However, the saccades evoked on antisaccade trials deviated not at all toward the pop-out target but only toward the ultimate endpoint of the saccade opposite the singleton. Recall that, on antisaccade trials, the most visually responsive neurons in the FEF initially select the pop-out target while attention is allocated to distinguish its shape. In contrast, movement neurons are activated but do not produce a directional signal until immediately before the saccade is made. Thus, the FEF can covertly orient attention without preparing a saccade to the locus of attention. These results indicate that the premotor theory of attention must be revised.

In summary, the FEF was discovered over a century ago when it was found that electrical stimulation elicited movements of the eyes. Today, electrical

stimulation of the FEF confirms that it contributes significantly to overt orienting through gaze shift and also to covert orienting through attention shifts.

See also: Attention and Eye Movements; Cortical Control of Eye Movements; Eye and Head Movements; Eye Movement Disorders; Neural Coding of Spatial Representations; Oculomotor Control: Anatomical Pathways; Oculomotor System: Models; Parietal Cortex and Spatial Attention; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Pursuit Eye Movements; Saccade–Pursuit Interactions; Saccades and Visual Search; Saccadic Eye Movements; Sensorimotor Integration: Attention and the Premotor Theory; Superior Colliculus; Supplementary Eye Fields; Target Selection for Pursuit and Saccades; Visual System: Multiple Visual Areas in Monkeys; Visual Cortex in Humans.

Further Reading

Bruce CJ and Goldberg ME (1985) Primate frontal eye fields. I: Single neurons discharging before saccades. *Journal of Neurophysiology* 53: 603–635.

Corbetta M and Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* 3: 201–215.

Fukushima K (2003) Frontal cortical control of smooth-pursuit. *Current Opinion in Neurobiology* 13: 647–654.

Juan CH, Shorter-Jacobi SM, and Schall JD (2004) Dissociation of spatial attention and saccade preparation. *Proceedings of the National Academy of Sciences of the United States of America* 101: 15541–15544.

Moore T and Fallah M (2004) Microstimulation of the frontal eye field and its effects on covert spatial attention. *Journal of Neurophysiology* 91: 152–162.

Muggleton NG, Juan C-H, Cowey A, and Walsh V (2003) Human frontal eye fields and visual search. *Journal of Neurophysiology* 91: 3340–3343.

Munoz DP and Schall JD (2003) Concurrent distributed control of saccades. In: Hall WC and Moschovakis AK (eds.) *The Oculomotor System: New Approaches for Studying Sensorimotor Integration*, pp. 55–82. Boca Raton, FL: CRC Press.

Pierrot-Deseilligny C, Milea D, and Muri RM (2004) Eye movement control by the cerebral cortex. *Current Opinion in Neurobiology* 17: 17–25.

Schall JD (1997) Visuomotor areas of the frontal lobe. In: Rockland K, Peters A, and Kaas J (eds.) *Extrastriate Cortex of Primates, Vol. 12: Cerebral Cortex*, pp. 527–638. New York: Plenum.

Schall JD (2002) The neural selection and control of saccades by frontal eye field. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1073–1082.

Schall JD, Stuphorn V, and Brown JW (2002) Monitoring and control of action by the frontal lobes. *Neuron* 36: 309–322.

Schiller PH and Tehovnik EJ (2001) Look and see: How the brain moves your eyes about. *Progress in Brain Research* 134: 127–142.

Thompson KG and Bichot NP (2005) A visual salience map in the primate frontal eye field. *Progress in Brain Research* 147: 251–262.

Thompson KG, Biscoe KL, and Sato TR (2005) Neuronal basis of covert spatial attention in the frontal eye field. *Journal of Neuroscience* 25: 9479–9487.

Frontal Lobe Syndrome

D T Stuss, Rotman Research Institute, Baycrest, Toronto, ON, Canada; and University of Toronto, Toronto, ON, Canada
M P Alexander, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA, USA and Rotman Research Institute, Baycrest, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

This article describes the functions associated with the frontal lobes, particularly as demonstrated by studies of patients with focal brain lesions. Modern neuroscience is overwhelmingly dominated by functional imaging studies in normal subjects, but functional imaging can only demonstrate, or in some cases only imply, a role for a region in a particular, often very narrowly defined cognitive task. Lesion studies are an essential counterbalance to functional imaging as the only mechanism to identify areas that are necessary for specific functions. Lesion studies have, however, often been hampered by sweeping up an amorphous and varied group of deficits, including cognitive, affective, comportment, and personality changes, under the term 'frontal lobe syndrome' as though there were a single frontal lobe syndrome that only varied in severity. The notion of a unitary frontal lobe syndrome, based solely on the coarsest anatomical distinctions, evolved at a time when lesion location was not easily specified and often in patients who were examined at a later stage of disease. The unitary syndrome notion was not strongly influenced by the observations that it could include diametrically opposite properties: patients could be impulsive or apathetic, distractible or perseverative, restless or bradykinetic, and so on. In recent years, the clinical research focus has changed from a concentration on the anatomical basis of the disordered cognition or behavior to an emphasis on the functional nature of those impairments under the supraordinate label of 'dysexecutive' disorder, but attribution of the dysexecutive syndrome to lesions of the frontal lobes has always been clear, although usually only implied and not demonstrated. This switch to a functional definition has its own problems. It, too, was often seen as a unitary disorder, and many, possibly most, patients with executive impairments did not have lesions restricted to frontal structures. In fact, many had disorders such as depression, sleep apnea, concussions, chronic pain, and other problems not easily tied to a structural injury of the frontal lobes. Although clinical neurology and neuropsychology are still quite tied to the concepts of frontal lobe syndrome and dysexecutive disorder, clinical research has increasingly demonstrated that specific, highly differentiated cognitive, motivational, behavioral, and attentional deficits are associated with discrete regional frontal injuries.

Our analysis suggests that frontal lobe functions fall into four distinct categories, each associated with a different frontal region. Only one of these categories is truly 'executive.' The other three categories cannot be subsumed under the common understanding of executive functions. Executive functions are best understood within a model of higher level attentional abilities. The executive functions are associated with discrete regions within the frontal lobes. We conclude that there is no single frontal system and not even a single frontal executive system and, thus, no undifferentiated frontal lobe or dysexecutive syndrome. There are domain general executive functions, each with a specific regional frontal basis, that are recruited over time in response to contextual demands.

There are four sections to this article. The first summarizes the anatomy and pathophysiology of the frontal lobes to provide a context for understanding research in frontal lobe functions. The second defines the proposed four categories of frontal lobe functions, with an emphasis on the one category most consonant with the term executive. The third section describes the interactive relationship between the executive functions and between executive functions and the other categories of frontal functions, emphasizing the organization of networks of processes and the unfolding of processes in time. We conclude with a brief description of the potential treatment of patients with frontal lobe damage.

Anatomy and Pathophysiology

The frontal lobes are the regions anterior to the Rolandic fissure and superior to the Sylvian fissure. They constitute one-fourth to one-third of the human brain. This article focuses on the prefrontal cortex, the area rostral to primary motor (Brodmann area 4, generation of movement) and premotor (Brodmann area 6, programming of sequential movements) cortex. The prefrontal cortex can be roughly subdivided into regions in which adjacent Brodmann area regions have similar connectivity patterns: lateral (dorsolateral: 9, 46, and perhaps 8; ventrolateral: 44, 45, and lateral 47/12), orbitofrontal (11–14; and orbital 47/12), polar (area 10), inferior medial (ventral anterior cingulate and paracingulate 24, 25, and 32), and superior medial (8, 9, dorsal anterior cingulate 24, 32, and perhaps including supplementary motor

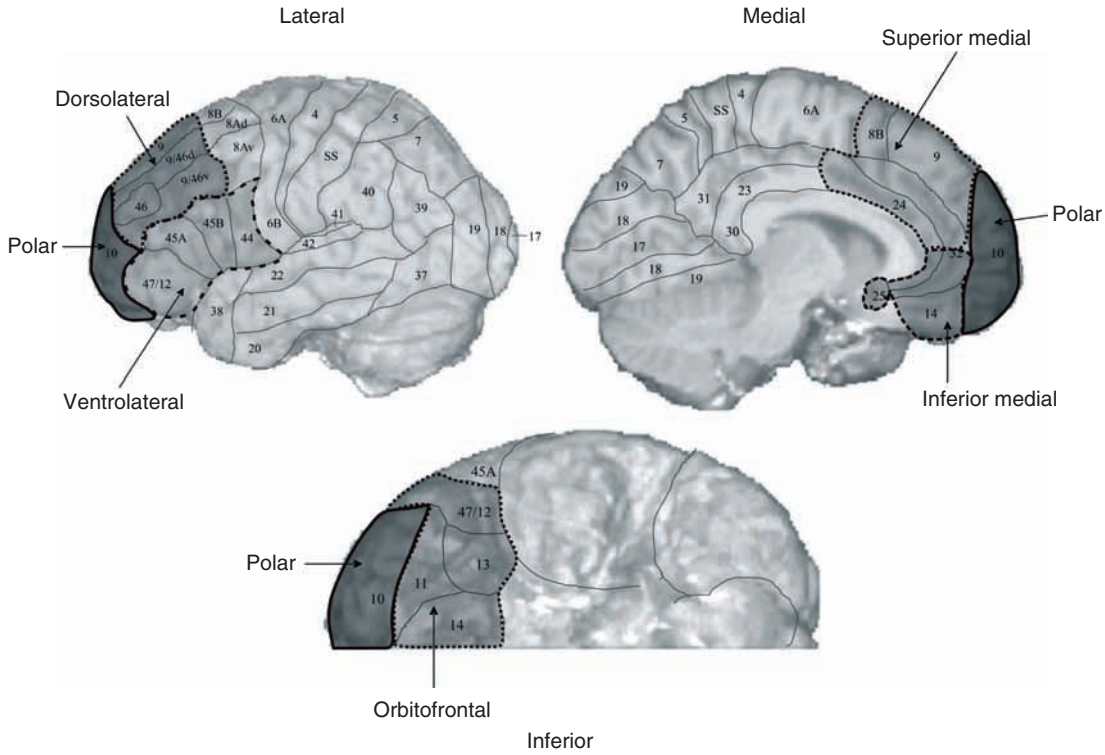

Figure 1 Depiction of the different regions within the frontal lobes.

area 6) (Figure 1). Only the lateral regions show strong functional hemispheric asymmetry.

The prefrontal cortex has strong reciprocal connections with virtually all other cortical regions. The specific patterns of connectivity determine the roles of different prefrontal regions in integrating and influencing all types of mental operations, cognitive and affective. In addition to functionally relevant cortico-cortical connections, there are parallel but independent frontal–subcortical circuits. The circuits are functionally distinct, and damage to any part of a circuit may produce similar discrete functional impairments. These may be more pervasive than suggested by cortical lesion alone and may be significant even in the absence of cortical lesions.

With the possible exception of early frontal–temporal lobe degeneration, there is no natural clinical condition in which pathology only involves the frontal lobes, but almost any neurological disease may preferentially affect the frontal lobes. Each has specific advantages for illuminating frontal functions.

Infarction is most common. Because infarcts are almost always unilateral and lesion boundaries are sharp, infarcts have always been the 'preferred' pathology for clinical–anatomical studies. Anterior cerebral artery (ACA) territory infarction can involve the orbital cortex, the frontal pole, the anterior cingulate, and/or the supplementary motor area. Bilateral lesions may be more frequent with ACA infarction since both ACAs may originate from a common internal carotid artery. Involvement of the anterior corpus callosum aggravates deficits caused by damage to the medial frontal cortex. Rupture of an ACA aneurysm is a common cause of ACA infarction. The perforator branches of that vessel supply the septal nuclei, and infarctions may cause significant limbic dysfunction including memory impairment. The vascular territory of the recurrent artery of Heubner includes the 'limbic caudate' and the inferior anterior limb of the internal capsule, which contains thalamofrontal connections. Middle cerebral artery (MCA) territory infarcts can involve significant regions of the dorsolateral and ventrolateral frontal convexity. Since MCA infarcts are not often restricted to the frontal lobes, care must be taken to demonstrate lesions restricted to frontal regions when studying the effects of MCA infarcts. If the lenticulostriate branches of the MCA are affected, the dorsolateral caudate, dorsolateral anterior limb of the internal capsule, and the putamen may also be damaged. Deep frontal white matter lesions can also affect the interconnection between prefrontal cortex (undamaged) and connected regions (e.g., posterior cortex and fibers of passage from ipsilateral medial frontal lobe and the contralateral frontal lobe).

Intracerebral hemorrhages do not necessarily follow vascular territories, potentially resulting in more anatomical variability than infarctions. Hemorrhages may also complicate behavior–lesion studies because of mass effects, secondary herniation damage, and hydrocephalus. Nevertheless, the potential involvement of combinations of structures not involved in infarctions indicates that hemorrhage cases can be of importance for brain–behavior studies.

Tumors can unilaterally infiltrate the frontal lobe, and as with hemorrhages they may produce a much wider range of combinations of regional injury than infarcts. Not all types of tumors are suitable for brain–behavior correlation study (e.g., rapidly growing intracerebral tumors (gliomas)) because of mass effects and uncertain margins of lesion. Relatively benign primary tumors (ependymomas, oligodendrogliomas, cystic astrocytomas, etc.) that have been excised may provide useful correlations. Extracerebral tumors (meningiomas) may cause considerable cortical impairment by compression and vascular compromise. Both of these tumor types are most informatively studied after resection, and only postoperative neuroimaging can define the areas of damaged brain.

Traumatic brain injury results in two broad categories of brain damage: diffuse axonal injury (DAI) and focal cortical contusions (FCCs). Although much DAI pathology is located in the deep frontal white matter, the widespread nature of DAI hinders demonstration of specific frontal–behavioral relationships. FCCs, on the other hand, may be appropriate for analysis of specific frontal–behavioral relationships provided that there is strong evidence that DAI is relatively modest (e.g., short duration of posttraumatic amnesia). Indeed, inclusion of such patients may be the only way to assess involvement of basal/polar frontal regions. Gunshot or other penetrating head wounds almost routinely produce unusual patterns of brain damage. Diffuse effects may be difficult to measure.

Many other disorders may affect the frontal lobes directly or disrupt the subcortical connections of frontal systems. Herpes simplex encephalitis and frontotemporal dementia commonly involve discrete frontal regions, but usually with extensive temporal injury as well. Large vessel and small vessel vascular diseases, hydrocephalus, and multiple sclerosis produce considerable white matter damage but rarely restricted to the frontal lobes. There are degenerative diseases with a predilection for subcortical structures that are critical components of frontal systems – Parkinson's disease, Huntington's disease, and progressive supranuclear palsy – but pathology is not restricted to these structures and is difficult to specify in life. Since many of these disorders involve multiple brain regions, they are not optimal for analysis of specific functions of the frontal lobes.

For much of the history of neuropsychology, strategies for lesion localization were fairly primitive, often analyzing case material only by frontal/posterior or right–left–bilateral frontal distinctions. In recent years, a variety of different procedures have evolved that sharpen frontal functional/anatomical correlations. Dividing the patients into those with and those without impairment in comparison to the control group and analyzing lesion site differences, either by simple overlaps or by lesion size within site correlations, have been fruitful. The classification and regression tree is a technique that can be used to divide patients based on their performance on a dependent variable into anatomically diverse groups. The performance groups are then analyzed for lesion region differences. 'Hotspotting' requires mapping the lesion of each patient onto an architectonic map. Then, for a defined dependent measurement (e.g., performance on a specific task), all individuals who have damage in a specific architectonic region are compared to all those who do not have damage in that area. This effectively identifies which region (or regions) results in most impairment of that particular process. This procedure requires experimental isolation of the specific process being studied and evaluation of a relatively large number of patients with focal lesions. The end result of fractionation of frontal processes is not phrenology but the identification of specific regions involved in simpler processes as a first step in understanding the unfolding of these abilities in more complex tasks and in real time.

Functional Domains within the Frontal Lobes

The existence of distinct functional domains within the frontal lobes is supported by evidence from developmental, anatomical connectivity, and lesion effects. The human cortex develops from two discrete sources. The lateral prefrontal cortical cortex (LPFC) has developed from hippocampal, archicortical sources. It determines spatial and conceptual reasoning processes. The ventral (medial) prefrontal cortex (VPFC) has evolved from paleocortical sources. It is associated with emotional processing, including the acquisition and reversal of stimulus–reward associations. These two broad developmental divisions form the two major (executive cognitive and behavioral/emotional self-regulatory) functional divisions within the frontal lobes and map onto two of the three proposed frontal–subcortical circuits involved in cognitive and/or emotional processing. Another frontal system is determined by the unique convergence of sensory and

limbic input combined with a distributed frontal output. This characterizes the medial frontal structures that are uniquely placed for regulating energization of all cognitive activities. The final category of frontal functions (metacognitive) is suggested by research on higher order integrative functions of the frontal polar area 10. Some authors use the term 'executive' for all four of these functional domains, but, as described later, we consider this inappropriate.

Regulating Energization

Energization of cognitive functions is important for adequate performance of any function. (This is commonly called activation, but this term has come to have multiple meanings – psychological, vascular, imaging, etc. – and we prefer energization.) Extensive damage to more superior medial (anterior cingulate and superior) frontal pathology results in abulia or severe apathy or, at its worst, akinetic mutism. This lack of energization of activity, defined by a diminished capacity to generate and maintain actions of mental processes, can be demonstrated even in much less clinically obvious cases. Patients with damage in this region are slow in generating lists of words (particularly in the first 15 s); have notably slower reaction time (RT), particularly if tasks are more demanding; are deficient in maintaining over time the benefit of a warning stimulus; and have problems maintaining a selected target such as in the Stroop interference test. RT speed is correlated with functional activation in the same superior medial region. Verbal fluency and the Stroop task are commonly used in the clinical setting to assess energization, but they do not isolate this deficit because they are not specific and require other cognitive (often executive) abilities. Perhaps the best measures to evaluate impaired activation are demanding reaction time tests.

Executive Cognitive Functions

Executive cognitive functions are defined as, and the definition is limited to, those high-level cognitive functions that require control and direction (e.g., planning, monitoring, activating, switching, and inhibiting) of lower level, more automatic functions. Much of the research on executive functions in neuropsychological studies has demonstrated deficits in patients with primarily LPFC damage. Tests commonly used by many clinicians as measures of frontal lobe executive functioning include the Wisconsin Card Sorting Test, Trail Making Test Part B, the Stroop Interference subtest, and specific measures within verbal fluency tasks. These tests are indeed more sensitive in general to focal LPFC (and also not generally to orbitofrontal/ventral medial) pathology and fall within the classification of tests of 'executive cognitive' processes. However, the tests are complex and multifactorial, and individuals can fail for many reasons.

Significant efforts have been made to develop and evaluate tests that more precisely isolate specific executive processes. Structured RT tasks that isolate discrete performance measures have identified specific dissociations of executive cognitive functions. Two major dissociations of executive cognitive functions within the dorsolateral prefrontal regions have been demonstrated. Impaired task setting has been related to left lateral frontal lobe damage, particularly ventrolateral. This has been shown as impaired bias in a feature integration task (errors are primarily false positives) and as significantly increased errors in the first block of 100 trials in a concentrating task requiring rapid stimulus–response matching. This deficient task setting is domain general; that is, deficits can be seen in any task, not just in RT tasks. For example, in a word list learning task, patients with left frontal damage had increased false positive errors in memory recognition.

The right lateral frontal area, on the other hand, is important in monitoring. Damage in this region results in several types of deficits that are attributable to deficient monitoring: as increased errors of all types in a complex feature integration task, indicating a problem in monitoring the difference between targets and nontargets (i.e., impaired sensitivity); as an abnormal foreperiod effect – that is, an increase in RT with longer interstimulus intervals, rather than the normal decrease in RT, suggesting a deficiency in monitoring the occurrence of stimuli over time; and as significantly increased variability of timing performance in a tapping task.

The mechanisms behind this top-down within-subject (intra-individual) variability of performance both within a testing session and across sessions appear to be the same attentional mechanisms evidenced by other RT measures – that is, task setting, monitoring, and energization. As would be expected of impairment in domain general abilities, such intra-individual variability can be seen in other tasks such as memory.

There are efforts to bridge the gap between laboratory tasks of executive cognitive functions and everyday functioning with the investigation of naturalistic actions under controlled conditions. Since 'real-life' measures are more complex and require the involvement of multiple functional systems, these measures overlap with behavioral/emotional self-regulatory measures.

Behavioral/Emotional Self-Regulatory Functions

The VPFC region is involved in emotional processing. An important, and natural, role for the VPFC, because of its involvement in reward processing, is behavioral self-regulation required in situations in which cognitive analysis, habit, or environmental cues are not sufficient to determine the most adaptive response. Patients with damage to the inferior medial frontal cortex have difficulty in understanding the emotional consequences of their behavior, despite intact performance on commonly used neuropsychological tests of executive functioning. Assessments of these abilities tend to be more experimental in nature, and they include gambling tasks and naturalistic multiple subgoal tasks.

Because of the role of the ventral prefrontal cortex in emotional processing (basic drives and rewards that inform and direct high-level decision making), tests assessing the acquisition and reversal of stimulus–reward associations can be used. This reversal learning (interpreted as affective) is dissociable from the impairment in attentional (extradimensional) set shifting found after LPFC lesions, reinforcing the distinction between executive attentional and affective/emotional behavioral measures.

The VPFC is also involved in higher level decision-making tasks involving reward processing in unstructured situations, such as the gambling task that has been suggested as both sensitive and specific to VPFC lesions. However, as with many neuropsychological tests, further evaluation suggests that these tests may also be multifactorial in nature, and the true behavior/anatomy specificity remains to be determined. Decision-making performance on the gambling task has been dissociated from deficits in working memory and inhibition.

A broad term to encompass the type of deficits observed on the gambling task is 'self-regulatory disorder,' defined as the inability to regulate behavior according to internal goals and constraints particularly in less structured situations. Qualitative descriptions with naturalistic multiple subgoal tasks, and paper-and-pencil laboratory versions, have been developed. There are also efforts to develop various scales of behavioral dyscontrol or assessment of frontal lobe personality characteristics.

Metacognitive Processes

The fourth category of frontal function is related to the frontal polar region, Brodmann area 10, although there is still debate as to how much this functional category is associated with damage to a more general area including the anterior medial regions. This frontal region appears to be maximally involved in the metacognitive aspects of human nature: integrative aspects of personality, social cognition, autonoetic consciousness, theory of mind, humor appreciation, and self-awareness. Self-awareness implies a meta cognitive representation of one's own mental states, beliefs, attitudes, and experiences. This self-reflection is key to understand the relationship of one's own thoughts to external events and, from this self-knowledge, to understand the mental states of others. This ability to make inferences about the world, and to empathize with others, allows correct interpretation of events and correct social judgments. Family reports often precisely describe the changes in behavior that have occurred: lack of empathy, unconcern, and inability to appreciate humor that requires self-reflection (appreciation of slapstick humor may be intact).

Because area 10 is among the most recently evolved of human brain regions, it may be uniquely positioned to integrate the higher level executive cognitive functions and emotional or drive-related inputs. This bridging of behavioral self-regulatory and executive cognitive functions may be the neural basis for their role in metacognitive functions. There is evidence, however, that there are potential dissociations within this domain, and that metacognitive processes are not reducible to these other functions.

The neuropsychological assessments in this category of frontal lobe functions are generally experimental. They include reactions to verbal and cartoon humor, visual perspective-taking tasks, and comparison of performance on remember–know memory tasks. For some individuals, however, these tasks can be solved on the basis of factual knowledge, not inference. Moreover, the relationship of process to anatomy is not certain, and it may depend on the differential demand of different tests within the overall 'metacognitive' rubric.

The general localization of the four separable frontal lobe processes is depicted in Figure 2.

Networks and Integration in Time

The evidence for fractionation of the functions of the frontal lobes does not imply that frontal lobes are simply a series of independent processes. Depending on task demands, there is the fluid recruitment of different processes anywhere in the brain into different networks. Lesion research demonstrates the effect of impairment in the networks; functional imaging demonstrates the recruitment of processes. For some task demands, one or more anatomically distinct frontal processes may be recruited. In tasks with simple demands, the more automatic nonfrontal processes function independently; as task demands increase or

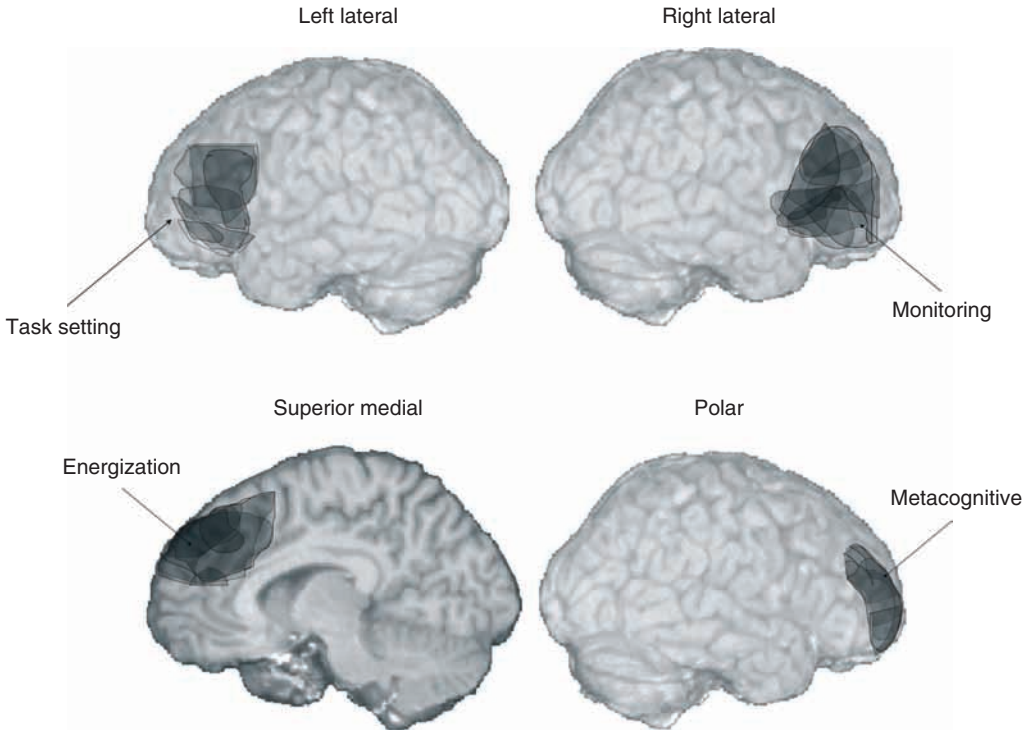

Figure 2 Illustration of three of the four major anatomical regions within the prefrontal cortex related to the different functional categories within this brain region. The four categories are energization (superior medial), executive (lateral), behavioral/self-regulatory (inferior medial; not illustrated), and metacognition (polar). Two executive processes are illustrated, one related with the left lateral region (task setting) and the other with the right (monitoring).

alter, there may be increased involvement of different frontal (more 'strategic') regions, even to the point where it appears all frontal regions are involved. In other demands the network may function 'top-down.'

These networks function in real time. Language provides one example. To request an object, answer a simple question, or tell a familiar story require essentially no control or monitoring and these presumably occur with no executive processes required. To give a complicated response to a difficult question will require more control. To tell a complex narrative – deciding what to include, what the listener may already know of the story, what is socially appropriate to include for a particular audience, and to keep track of the goal of the story while adding a critical digression – will require substantial executive control. Setting the task goal, monitoring the content and the listener, and sustaining and maintaining intention are essentially dependent on the several functions described previously. They are recruited at different levels, at different times across the task, and they in turn recruit a variety of posterior cognitive operations, each at a specific point in the narrative. Multiple executive processes, each primarily dependent on a specific frontal region, must be knitted together in real time.

Rehabilitation

There are no class I randomized controlled trials of executive rehabilitation in focal frontal patients, although there are a few studies on patients with traumatic brain injury, primarily DAI. Clinical treatments have been targeted at the disability level of rehabilitation, not at specified frontal lobe impairment. Three types of treatments have been utilized, individually or alone, with some success:

Drugs. Stimulants may help patients with low arousal states or who are easily distracted. Dopaminergic agents may help patients with poor activation (energization).

Cognitive. A few theoretically sound programs addressing 'executive dysfunction' have demonstrated some possible efficacy. These techniques often rely on training patients to stop before they begin a task (inhibit impulsive responses), then explicitly define goals (set task requirements), check off performance as they progress (monitor subgoals), and confirm appropriate completion (compare goal to outcome). Each step in this treatment would appear to be directed at a specific executive or self-regulatory factor of frontal function, but no study has yet assessed the specificity of treatment effectiveness.

Behavioral self-regulation/self-awareness. There have been successful efforts to improve self-awareness. It is uncertain how much of this is dependent on the environmental support provided to the patient.

Compensatory. Commonly, environmental compensation has focused on a limiting disability or handicap, resulting in greater functional independence. These programs make, appropriately, no claims for directly treating executive dysfunction.

Conclusion

Prior to the availability of computed tomography scans, patients with frontal lobe lesions were often not identified until clinical signs were well advanced, and by the time of identification, many patients had lesion extension into brain regions outside the frontal lobes or had extremely large frontal lesions or both. Patients' deficits were often characterized only with behavioral descriptions or with multidimensional tests that have high sensitivity but low specificity for frontal injury. The historical notions of the frontal lobe syndrome come from observations of such patients. In the past 25 years, early diagnosis, precise lesion definition, and a broad range of specific and sensitive tests of regulation of attention, discrete executive functions, self-regulatory behaviors, and empathy and social awareness have combined to provide a much more precise understanding of the many consequences of specific frontal injuries and have made it clear that there is little evidence of a single frontal lobe or dysexecutive syndrome.

See also: Brain Trauma; Executive Function and Higher-Order Cognition: Neuroimaging; Executive Function and Higher-Order Cognition: Definition and Neural Substrates; Humans; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex.

Further Reading

Bechara A, Tranel D, and Damasio H (2000) Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain* 123: 2189–2202.

Cicerone KD, Dahlberg C, Malec JF, et al. (2005) Evidence-based cognitive rehabilitation: Updated review of the literature from 1998 through 2002. *Archives of Physical Medicine and Rehabilitation* 86: 1681–1692.

Corbetta M and Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* 3: 201–215.

Cummings JL (1993) Frontal–subcortical circuits and human behavior. *Archives of Neurology* 50: 873–880.

Grafman J, Holyoak KJ, and Boller F (eds.) (1995) *Structure and Functions of the Human Prefrontal Cortex*. New York: New York Academy of Sciences.

Lhermitte F (1986) Human anatomy and the frontal lobes. 2. Patient behavior in complex and social situations: The environmental dependency syndrome. *Annals of Neurology* 19: 335–343.

Miller EK (2000) The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience* 1: 59–65.

Norman DA and Shallice T (1986) Attention to action: Willed and automatic control of behavior. In: Davidson RJ, Shwartz GE, and Shapiro D (eds.) *Consciousness and Self-Regulation: Advances in Research and Theory*, vol. 4, pp. 1–18. New York: Plenum.

Pandya DN and Yeterian EH (1996) Comparison of prefrontal architecture and connections. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 351: 1423–1432.

Passingham R (1995) *The Frontal Lobes and Voluntary Action*. New York: Clarendon.

Petrides M (2005) Lateral prefrontal cortex: Architectonic and functional organization. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences* 360: 781–795.

Rolls ET (2004) The functions of the orbitofrontal cortex. *Brain and Cognition* 55: 11–29.

Siegal M and Varley R (2002) Neural systems involved in 'theory of mind.' *Nature Reviews Neuroscience* 3: 463–471.

Stuss DT and Knight RT (eds.) (2002) *Principles of Frontal Lobe Function*. New York: Oxford University Press.

Stuss DT and Levine B (2002) Adult clinical neuropsychology: Lessons from studies of the frontal lobes. *Annual Review of Psychology* 53: 401–433.

Functional Amnesia

M Kritchevsky, San Diego VA Healthcare System, San Diego, CA, USA

Published by Elsevier Ltd.

Introduction

Functional amnesia is an uncommon condition in which patients develop severe retrograde amnesia in the absence of significant anterograde amnesia and without any known brain injury or disorder. This nonneurological syndrome is variously referred to as psychogenic amnesia, hysterical amnesia, dissociative amnesia, and functional retrograde amnesia. Functional amnesia is easily distinguished from the neurological amnesic syndrome that is commonly seen with dysfunction of bilateral medial temporal lobe or bilateral medial diencephalic structures important for memory, in which patients have severe anterograde amnesia with variable retrograde amnesia. In contrast, functional amnesia may be indistinguishable from malingered amnesia. Functional amnesia is the disorder of memory that is most commonly popularized in literature and film, and it appears to be the condition that the majority of lay persons regard as amnesia.

Early descriptions of functional amnesia included large numbers of cases that were characterized primarily by a loss of personal identity. Memory for previously known facts and personal events sometimes was noted to be impaired as well. In 1982 the first case of functional amnesia that had been formally studied with neuropsychological testing was published. Subsequently many additional case studies have been described. In 2004 the first series of patients with functional amnesia who had been studied clinically and neuropsychologically in a systematic fashion was reported. The increasing understanding of functional amnesia has facilitated diagnosis of this condition, but treatment of many patients with functional amnesia has remained difficult.

Clinical Characteristics of the Patient with Functional Amnesia

The Clinical Picture of the Patient

The typical patient with functional amnesia presents with the sudden onset of retrograde amnesia without clinically significant anterograde amnesia. Most commonly, the retrograde amnesia affects all of the memories that occurred prior to the onset of amnesia, including memory for personal identity. Thus, the patient reports that he or she cannot remember their past. They do not know their name, where they were born, or where they were raised. They can not recall where they went to school, when or to whom they were married, where they live, or by whom they are employed. Once their identity is discovered, they do not recognize their name. When they meet their friends and family, they do not recognize them. In general, the loss of past memories affects public as well as personal material. For example, the patient will not recall or recognize previously known public figures such as sports figures, movie stars, and political leaders. They will not recall or recognize previously known public events such as wars or national disasters. Because the person with functional amnesia has no significant anterograde amnesia, they are able to relearn past memories. However, these newly learned memories typically have no sense of familiarity to them.

Uncommonly, the patient with functional amnesia has retrograde amnesia that affects only a period of time before the onset of memory loss. For example, if the period of amnesia is 2 years, then the patient will not recall or recognize memories that occurred during the 2-year period before the onset of amnesia. The patient will know their name and will have normal recollection of memories that occurred more than 2 years ago. However, he or she will give the year as 2 years earlier than the correct year, and will report their marital status, place of residence, and workplace as they were 2 years previously despite any significant changes that may have occurred since. The patient can relearn past memories, but the new memories will not have a sense of familiarity.

The patient with functional amnesia most often will report a complete loss of memories from the period of retrograde amnesia. However, on occasion a patient will be aware of one or more vague and incomplete memories from the period of amnesia.

At about the time of onset of functional amnesia, many patients have one or more abnormalities, in addition to the amnesia, that also appear to be psychogenic. These have included inability to name objects, inability to name numbers, difficulty with calculations, loss of ability to read and write, inability to move the eyes voluntarily during formal testing, bilateral leg weakness with inability to walk independently, weakness on one side of the body, and an inability to carry out previously familiar activities such as using a telephone or driving a car with a manual transmission. One patient has reported that he had to relearn the English language by reading a dictionary during the first 2 days of his amnesia. Another has reported that she could

no longer remember the Spanish language, though she had been bilingual prior to the onset of her amnesia. With regard to anterograde amnesia, patients characteristically have good memory for events that occurred after the onset of amnesia. One patient has reported difficulty with new learning, and on neurological examination was poor at learning new verbal, but not nonverbal, material. However, like other patients with functional amnesia, he did not appear to be forgetful.

Occasionally, a patient with functional amnesia will present immediately after a period of sudden and unexpected travel during which they appear to have exhibited normal behavior. The period of travel is termed a fugue state, and whatever occurred during the fugue state will be included in the period of retrograde amnesia.

Risk and Precipitating Factors

Many patients with functional amnesia have significant premorbid psychiatric histories, including one or more of the following: alcohol abuse, other substance abuse, previous conversion symptoms, anxiety disorder, posttraumatic stress disorder, schizophrenia, depression, history of suicide attempt, and histrionic or borderline personality disorder. Many patients with functional amnesia have one or more of the following possible precipitating factors at the onset of amnesia: intoxication with alcohol, mild closed head injury, active depression, psychological stress, and involvement or alleged involvement in illegal activity.

Clinical Course

The clinical course of patients with functional amnesia is highly variable. Some patients appear to have a full recovery of their retrograde amnesia. In these cases the recovery most often occurs gradually over days to weeks. The recovery most often appears to consist of the gradual return of memories from throughout the period of amnesia. Rarely, the amnesia appears to resolve suddenly. Many patients have only a partial recovery of their lost memories. Sometimes this involves partial but incomplete recovery of memories from throughout the period of memory loss. In contrast, at times the initially extensive retrograde amnesia appears to shrink over a period of months to a shorter period of retrograde amnesia for a period of time (e.g., 6 months) immediately preceding the onset of amnesia. Some patients with functional amnesia report that they experience little or no recovery of their lost memories over an interval as long as 42 months. Patients who experience no significant recovery of memory occasionally appear to have established new personalities. It appears to be very uncommon for functional amnesia to suddenly and completely resolve in association with a mild head injury or a psychological catharsis.

The recovery of lost memories sometimes appears to have been facilitated by reacquainting the patient with their past, or by addressing and initiating treatment for ongoing psychological stressors. Also, sometimes hypnosis or an amobarbital interview appears to facilitate the recovery of lost memories. Memories initially remembered with these therapies sometimes are again inaccessible immediately following the session, but then rapidly return in the following days.

Recovery from retrograde amnesia should be distinguished from relearning lost memories. The patient with functional amnesia usually will report with confidence that a particular memory was relearned. They sometimes will recall the event of relearning and usually will report that the memory feels new and is unassociated with the familiar feel of an old memory.

Additional psychogenic cognitive deficits that are associated with the retrograde amnesia often resolve within hours to days of onset. General neurological deficits, particularly weakness, of psychogenic origin may persist throughout the course of the amnesia.

Laboratory Tests

In patients with functional amnesia, structural neuroimaging tests such as computed tomography (CT) and magnetic resonance imaging (MRI) brain scans are usually normal, though they occasionally show abnormalities unrelated to the functional amnesia. Electroencephalography also is usually normal, but may show abnormalities that are unrelated to the amnesia. Functional neuroimaging tests such as single-photon computed tomography (SPECT), positron emission tomography (PET), or functional MRI brain scans have been reported to show abnormal findings related to the retrograde amnesia in some patients. The findings have not been consistent in the different patients studied. Although the demonstrated functional abnormalities may be related to the patients' retrograde amnesia, the presence of such abnormalities does not imply that functional amnesia is due to or associated with brain injury or disorder.

Differential Diagnosis

Malingered (intentionally feigned) amnesia also can present with the sudden onset of retrograde amnesia without clinically significant anterograde amnesia. In fact, it may be that a patient with malingered amnesia only can be distinguished from a patient with functional amnesia with certainty if the malingerer confesses. One clue that a patient has malingered amnesia may be the absence of a significant past psychiatric history. The precipitating factors for malingered amnesia may be similar to those for

functional amnesia. One malingerer hit himself in the head with a bottle so there would be evidence of an appropriate precipitating factor for his amnesia.

Rare neurological patients have been reported to have significant retrograde amnesia with a lesser degree of anterograde amnesia. This may be associated with bilateral damage to the anterior or inferior temporal lobes. Personal identity probably is preserved, the retrograde amnesia covers a period of several years to several decades, and there is some degree of anterograde amnesia. Neuroimaging typically will reveal the responsible brain injury.

The neurological amnesic syndrome that is commonly seen with dysfunction of bilateral medial temporal lobe or bilateral medial diencephalic structures important for memory is easily distinguished from functional amnesia. Patients with the neurological amnesic syndrome have significant anterograde amnesia with variable retrograde amnesia.

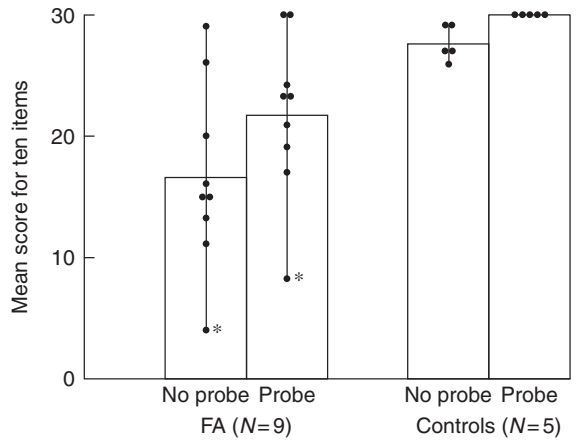

Figure 1 Patients with functional amnesia (FA) and controls were asked to recollect ten autobiographical episodes. Responses were scored (0–3) before (no probe) and after (probe) encouragement by the examiner to elicit as specific a recollection as possible. Each filled circle indicates the score for one subject. The asterisk indicates the score for the patient who later confessed that he had malingered his amnesia. Kritchevsky M, Chang J, and Squire LR (2004) Functional amnesia: Clinical description and neuropsychological profile of 10 cases. *Learning and Memory* 11: 213–226.

Neuropsychological Findings in Patients with Functional Amnesia

In 2004, a series of ten patients who had initially been diagnosed with functional amnesia was published. The patients had been studied clinically and neuropsychologically in a systematic fashion. Nine of the patients presented with loss of past memories that extended into childhood and included loss of personal identity. One had retrograde amnesia that was limited to 2–3 years. Eight of the ten patients were thought to have definite functional amnesia. One of the patients (whose performance is noted by an asterisk in the figures) subsequently confessed that he had malingered his amnesia. The final patient was considered to be a possible malingerer. The patients ranged in age from 28 to 54 years (mean, 37 years). For the eight patients for whom the data were known, education ranged from 9 to 16 years (mean, 13 years). Eight patients were men and two women. Seven were tested neuropsychologically 1–14 days after the onset of amnesia. Three were tested 10 weeks to 9 months after the onset of amnesia. Five patients were tested between 1982 and 1985. The other five were tested between 1992 and 1995. Four men and one woman, matched to the patients with respect to age and education, served as controls. They were tested in 1996. The neuropsychological data presented in this section are from these patients and controls.

Retrograde Amnesia

Figure 1 shows the performance of nine of the ten patients on a test of past autobiographical memory. This test was not administered to the patient who had retrograde amnesia of only 2–3 years. Ten common words (e.g., bird, clock) were presented one at a time with the instruction to recall a specific personal event from any time in the past that involved the stimulus word. Subjects were asked to describe the memory in as much detail as possible and then to date the memory. When recall was not clearly specific to time and place, the examiner probed to elicit the most specific memory possible. Responses were scored on a 0–3 scale, with 3 representing a well-formed episodic memory. The patients performed more poorly than the controls, both without (t [12] = 3.2, $p < 0.01$) and with probing (t [12] = 2.7, $p < 0.05$) by the examiner. The two patients who obtained the best scores were tested 3 and 9 months after the onset of the amnesia. Their good performances resulted from their ability to recall memories from the period after the onset of their amnesia. On neurological examination, both of these patients exhibited significant loss of autobiographical memories for events that had occurred before the onset of amnesia.

Figure 2 shows the percentage of well-formed episodic memories that were recalled from different past time periods (0 days to > 20 years) on the test of past autobiographical memory. One patient, the malingerer, was unable to produce any well-formed recollections and was not included in this analysis. The remaining patients each recalled from three to ten memories. The patients differed markedly from the controls in that they recalled most of their memories from the most recent time period (patients, 78%; controls = 24%; t [11] = 3.1, $p < 0.01$) and very few

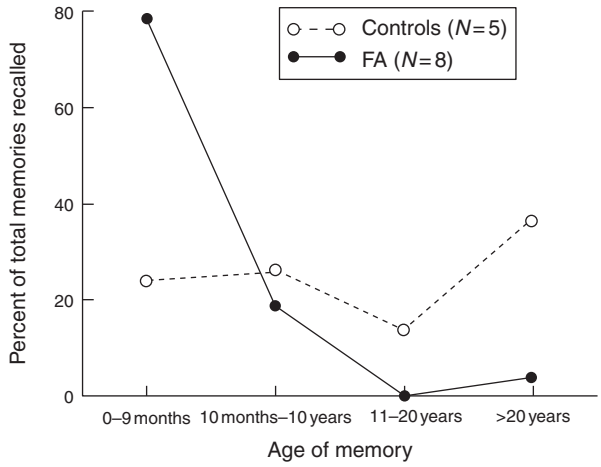

Figure 2 Percentage of memories that were recalled from the indicated time periods for patients with functional amnesia (FA) and controls. The data are based only on those recollections given a maximum score of 3 on the ten-item test of past autobiographical memory (**Figure 1**). The patient who later confessed that he had malingered his amnesia had no recollections that were given a three-point score. For the eight patients, all the memories from the 0–9 months time period were taken from the period after the onset of amnesia. Kritchevsky M, Chang J, and Squire LR (2004) Functional amnesia: Clinical description and neuropsychological profile of 10 cases. *Learning and Memory* 11: 213–226.

of their memories from the most remote time period (patients, 3%; controls, 36%; t [11] = 3.2, $p < 0.01$). It was notable that all of the memories recalled by the patients from the most recent time period were drawn from the time after the onset of amnesia. Thus, compared with the controls, the patients had a striking tendency not to recall remote memories, but rather to draw their memories from the period after the onset of the amnesia.

Figures 3(a) and 3(b) show the performance of four of the patients on a test of recall and recognition of public events. These patients were tested between 1992 and 1994. Twenty-four questions about public events that had occurred from 1970 to 1985 were asked. This test first was administered in a recall format (e.g., Who killed John Lennon?) and then, for each item that was not recalled or was incorrectly recalled, in a four-alternative, multiple-choice format (John Hinkley, Sara Jane Moore, David Roth, and Mark Chapman). The recognition score was based on the items that were recalled correctly plus the items that were recognized correctly. The four patients were impaired both at recall (16.7% correct for patients; 39.2% correct for controls; t [7] = 4.2, $p < 0.01$) and at recognition (44.8% correct for patients; 76.6% correct for controls; t[7] = 4.2, $p < 0.01$). One of these patients, who had retrograde amnesia of only 2–3 years, performed better than the other three. However, he still performed at the low end of the control range on recall (29.2%) and poorer than the controls on recognition (62.5%). The six other patients took a similar test consisting of 17–26 questions about the two most recent decades before the onset of their amnesia. They also scored poorly at recall (16.3% correct; one patient did not take the recall test) and at recognition (48.3% correct). One of these patients, the possible malingerer, had a score of 0% on the recall test and 11% on the recognition test (chance = 25%).

Figures 3(c) and 3(d) show the performance of four of the patients on a test of recall and recognition of famous faces. These patients were tested between 1992 and 1994. Nineteen photographs of famous people who came into the news from 1970 to 1989 were shown. The test first was administered in a recall format and then, for each item that was not recalled or was incorrectly recalled, in a recognition format. For recognition, half the items were yes/no questions (e.g., Is this person's name Liza Minelli?) and half were three-alternative, multiple-choice questions (e.g., Liza Minelli, Ann Margaret, Helen Reddy). The recognition score was based on the items that were recalled correctly plus the items that were recognized correctly. As a group, the four patients were impaired both at recall (40.8% correct for patients; 64.2% correct for controls; t [7] = 5.7, $p < 0.01$) and at recognition (80.2% correct for patients; 94.7% for controls; t [7] = 4.2, $p < 0.01$). The six other patients took a similar test consisting of 20–27 faces from the two most recent decades before the onset of their amnesia. They also scored poorly as a group at recall (39.1% correct) and at recognition (77.8% correct; one patient did not take the recognition test). There was marked variability in the individual scores of the ten patients. Thus, both the recall and recognition scores of two patients (one of these was the malingerer) were as good as or better than the control scores. Also, the patient who had retrograde amnesia of only 2–3 years performed just below the controls at recall and obtained a perfect score at recognition. A fourth patient scored well enough on the recognition test to reach the low end of the control range. A fifth patient scored 0% correct at recall but nearly reached the control range at recognition. Finally, a sixth patient (the possible malingerer) scored 0% correct at recall and 18.5% correct at recognition (chance = 41.7%). His overall score on the public events and famous faces recognition tests was 2.75 standard deviations below the score that would have been expected from guessing.

Figure 4 shows the performance of nine of the ten patients on tests of identification of US and southern California cities. First, each subject was read the names of 16 real and 16 plausible US cities and associated

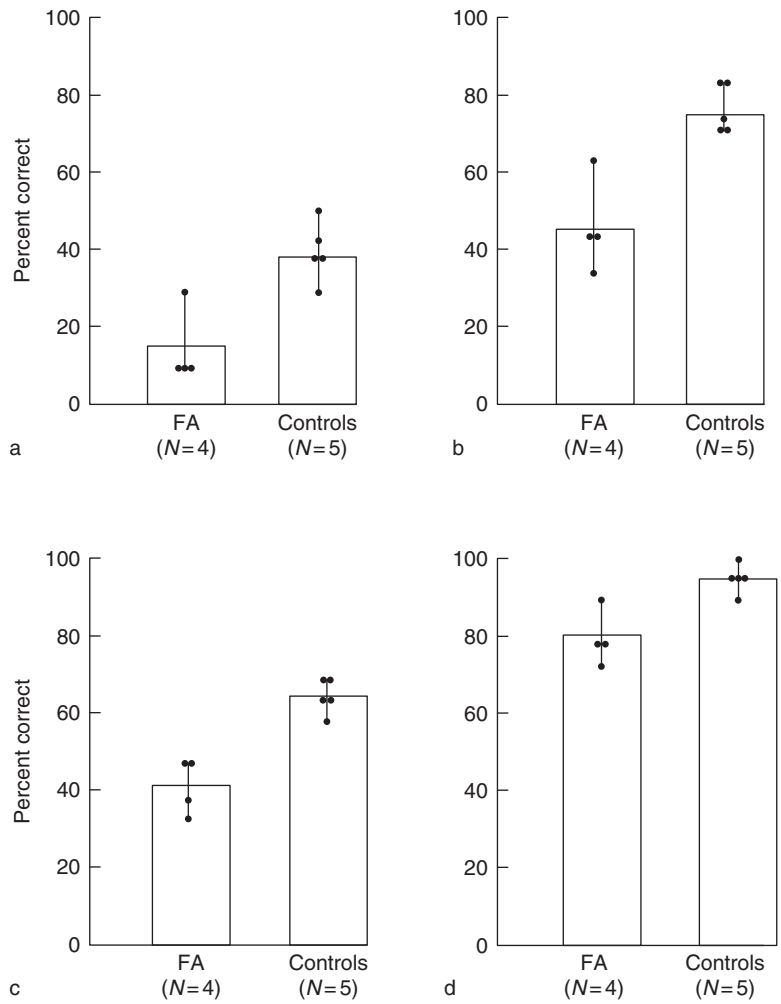

Figure 3 Performance on tests of remote memory for public events and famous faces. Patients with functional amnesia (FA) and controls were asked 24 questions about public events that had occurred during the two decades prior to testing (a) and then took a four-alternative, multiple-choice test about the same events (b). Subjects also were asked to identify 19 photographs of famous people who came into the news during the two decades before testing (c) and then to recognize the names that they could not recall (d). The recognition scores were based on the number of items recalled correctly plus the items that were recognized correctly. Chance = 25% for public events, 41.7% for famous faces. Each filled circle indicates the score for one subject. Kritchevsky M, Chang J, and Squire LR (2004) Functional amnesia: Clinical description and neuropsychological profile of 10 cases. *Learning and Memory* 11: 213–226.

states (e.g., Phoenix, Arizona, or Nelander, Michigan) and asked to identify whether each city name was real or fabricated. Then each subject was read a list of ten real and ten plausible southern California cities (e.g., Oceanside or Palmville) and again asked whether each city name was real or fabricated. The score on each test was the number of cities correctly identified. Overall, the patients performed similarly to the controls (US cities; $t[12] = 1.8$, $p = 0.09$; southern California cities, $p > 0.10$). However, two patients scored more than two standard deviations below the control mean on both tests. One of these, the possible malingerer, scored 47% correct on the US cities test and 30% correct on the southern California cities test (chance = 50% for each test).

In summary, all patients were impaired in their ability to recall autobiographical memories of specific events from their past. In contrast, there was marked variability in their performances on tests of their ability to recall and recognize public events and famous faces from the two decades before the onset of their amnesia. Thus, the possible malingerer scored significantly below chance on the recognition tests, and it is possible that he intentionally chose the wrong answers. Also, three patients (one was the malingerer) performed particularly well on the famous faces tests despite performing poorly on the public events tests. Another patient was unable to recall a single item on the famous faces test, but then did well on recognition of the faces. There was less variability in the patients'

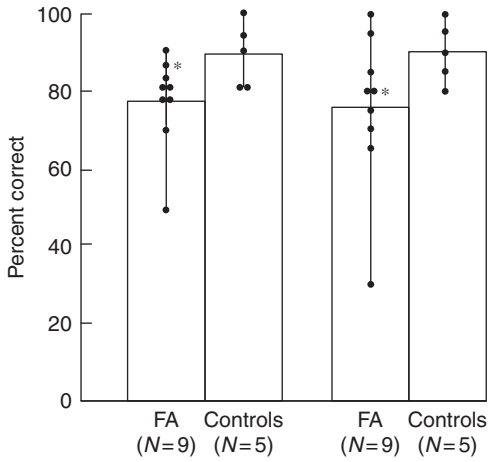

Figure 4 Patients with functional amnesia (FA) and controls were asked to identify the names of cities in the United States (left two bars) and the names of cities in southern California (right two bars) from lists of real and fictitious city names. Chance = 50%. Each filled circle indicates the score for one subject. The asterisk indicates the scores for the patient who later confessed that he had malingered his amnesia. Kritchevsky M, Chang J, and Squire LR (2004) Functional amnesia: Clinical description and neuropsychological profile of 10 cases. *Learning and Memory* 11: 213–226.

ability to distinguish names of real cities from the names of fictitious cities. Two patients, one the possible malingerer, performed abnormally on these tests.

Anterograde Amnesia

Figure 5(a) shows the performance of the ten patients with functional amnesia on a test of story recall. A short prose passage was read to the patient. Recall was tested immediately and again after a 10–20 min delay. The score was the number of story segments correctly recalled. Overall, the patients were marginally impaired at immediate recall (t [13] = 2.1, $p < 0.06$) but similar to controls at delayed recall ($p > 0.10$). Nonetheless, there was considerable variability in the performance of individual patients. Six of the patients (including the malingerer and the possible malingerer) had immediate recall scores that were more than two standard deviations below the control mean. Only the malingerer had a delayed recall score that was more than two standard deviations below the control mean.

Figure 5(b) shows the performance of the ten patients on a test of paired associate learning. A series of ten noun–noun word pairs was presented verbally and visually on each of three study trials. After each study trial, the patient was shown the first word of each pair and asked to recall the second word. The score was the number of words correctly recalled on each trial. Patients and controls performed similarly overall (F [1,13] = 2.7, $p > 0.10$) and improved at a similar rate across the three trials (F [2,26] = 0.5, $p > 0.10$). There was some variability in the scores,

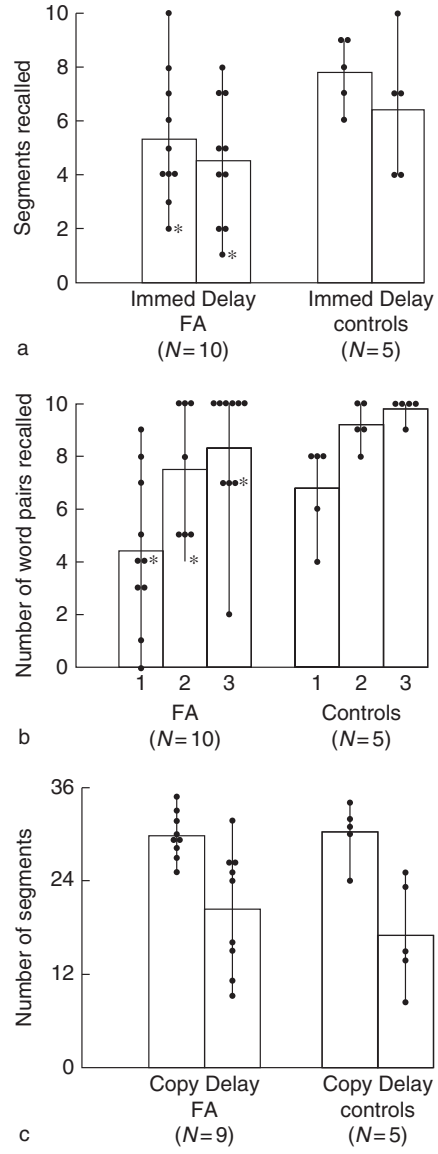

Figure 5 Performance of patients with functional amnesia (FA) and controls on tests of anterograde amnesia. (a) Story recall. Recall was tested immediately (Immed) and again after a delay of 10–20 min (Delay). (b) Paired-associate learning. Ten word pairs were presented on each of three study trials. (c) Copy of a complex diagram and reconstruction of the diagram from memory after 10–20 min (Delay). Maximum score = 36. Each filled circle indicates the score for one subject. The asterisk indicates the scores for the patient who later confessed that he had malingered his amnesia. He did not receive the diagram recall test. Kritchevsky M, Chang J, and Squire LR (2004) Functional amnesia: Clinical description and neuropsychological profile of 10 cases. *Learning and Memory* 11: 213–226.

with four of the patients (including the malingerer) having total number of words recalled across the three trials more than two standard deviations below the overall control mean.

Figure 5(c) shows the performance of nine of the patients on a test of recall of nonverbal material.

This test was not administered to the malingerer. Each patient copied a complex diagram. After a 10–20 min delay and without forewarning, they were asked to reproduce the diagram from memory. The score was the number of parts of the diagram correctly recalled. The patients scored similarly to the controls, both when they copied the diagram and when they reconstructed it from memory ($p > 0.10$). No patient obtained a score on this test that was below the lowest score obtained by the controls.

In summary, the patients as a group performed well on tests of anterograde amnesia for verbal and nonverbal material. However, there was some variability among patients. Some appeared to have abnormal performances on immediate verbal memory for a prose passage or on paired associate learning. Despite these abnormal performances, no patient exhibited any difficulty learning and remembering day-to-day events after the onset of amnesia.

Etiology of Functional Amnesia

Functional amnesia is a nonneurological, conversion symptom-like disorder. The model for the amnesia is the amnesic syndrome that is popularized in literature and film, namely a loss of past memories that usually includes a loss of personal identity. Each patient has their own individual conception of what past memories should be lost, which leads to the striking variability in the clinical and neuropsychological findings of patients with functional amnesia. In functional amnesia, as in a conversion disorder, the patient is not intentionally producing the deficits. It has been proposed that posthypnotic amnesia might serve as a laboratory model of functional amnesia, and self-hypnosis is one possible mechanism of conversion disorder and of functional amnesia.

Evaluation and Treatment of the Patient with Functional Amnesia

The evaluation of the patient should begin with neurological and psychiatric histories. The patient may be able to provide some of this information, but early in the course of evaluation most of the history will come from family members, friends, and healthcare or legal professionals. Neurological and psychiatric examinations should be performed. Laboratory investigations usually are not required, but may be obtained in selected patients.

The patient should be treated in the same manner as a patient with a conversion disorder. They should be reassured that there is no evidence of serious or significant brain problem. Significant psychological stressors should be identified and treated. They should be informed that their memories may begin returning within days of the onset of the amnesia. Contact with friends and family members should be encouraged unless there is a contraindication. If there has been little or no return of past memories within about a week, then treatment with hypnosis or an amobarbital interview should be considered if the treating psychiatrist agrees. If psychologically painful memories are remembered during hypnosis or an amobarbital interview, then the patient will require careful observation to be sure that these memories do not lead to undue distress or even to a suicide attempt. Despite the best possible therapy, some patients have only partial recoveries and some patients do not appear to have any recovery from their functional amnesia.

See also: Cognition: An Overview of Neuroimaging Techniques; Cognitive Dysfunction in Psychiatric Disorders; Declarative Memory System: Anatomy; Episodic Memory; Humans; Memory Consolidation: Systems; Memory Disorders.

Further Reading

Abeles M and Schilder P (1935) Psychogenic loss of personal identity. *Archives of Neurology and Psychiatry* 34: 587–604.

American Psychiatric Association (2000) Dissociative amnesia. In: *Diagnostic and Statistical Manual of Mental Disorders,* 4th edn., pp. 520–523. Arlington, VA: American Psychiatric Association.

Barnier A (2002) Posthypnotic amnesia for autobiographical episodes: A laboratory model of functional amnesia? *Psychological Science* 13: 232–237.

Kanzer M (1939) Amnesia: A statistical study. *American Journal of Psychiatry* 96: 711–716.

Kapur N (1993) Focal retrograde amnesia in neurological disease: A critical review. *Cortex* 29: 217–234.

Kritchevsky M, Chang J, and Squire LR (2004) Functional amnesia: Clinical description and neuropsychological profile of 10 cases. *Learning and Memory* 11: 213–226.

Kritchevsky M, Zouzounis J, and Squire LR (1997) Transient global amnesia and functional retrograde amnesia: Contrasting examples of episodic memory loss. *Philosophical Transactions of the Royal Society of London B* 352: 1747–1754.

Schacter DL, Wang PL, Tulving E, and Freedman M (1982) Functional retrograde amnesia: A quantitative case study. *Neuropsychologia* 20: 523–532.

Yang JC, Jeong GW, Lee MS, et al. (2005) Functional MR imaging of psychogenic amnesia: A case report. *Korean Journal of Radiology* 6: 196–199.

Functional Connectivity

K E Stephan and K J Friston, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Electrophysiological methods, such as electroencephalography (EEG) and magnetoencephalography (MEG), and functional imaging techniques, such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), are used widely in cognitive neuroscience. A common aim is to understand brain function along two dimensions: functional specialization and functional integration. Functional specialization assumes a local specialization for certain aspects of information processing but allows for the possibility that this specialization is anatomically segregated across different cortical areas. Most current functional neuroimaging experiments have adopted this view and interpret the areas that are activated by a certain task component as the elements of a distributed system. However, this characterization does not address how the locally specialized areas are bound together by context-dependent interactions among these areas – that is, the functional integration within the system. This integration can be quantified in two ways – functional connectivity and effective connectivity. Whereas functional connectivity describes statistical dependencies between remote neuronal data, effective connectivity rests on a mechanistic model of how the data were caused. This article reviews established techniques for characterizing functional integration on the basis of electrophysiological and functional neuroimaging data.

Functional Connectivity

Functional connectivity is operationally defined as the temporal correlation between the time series of different brain regions. (A more general definition of functional connectivity is the presence of statistical dependencies between two sets of neurophysiological data. If one assumes that these data conform to Gaussian assumptions, one only needs to investigate second-order dependencies (i.e., covariances or correlations). For non-Gaussian processes, higher order dependencies need to be taken into account. This can be done, for example, with independent component analysis.) In fMRI and PET, such time series are available for each volume element (voxel) of the brain. For example, given two voxels with time series x and y, the functional connectivity simply corresponds to the Pearson correlation coefficient r of the two time series

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{s_x \cdot s_y} \qquad [1]$$

where s_x, s_y are the standard deviations, and cov_{xy} is the covariance of the two time series. Note that functional connectivity suffers from the general problem of interpreting correlations: are the two time series correlated because (1) x influences y, (2) y influences x, (3) both influence each other, or (4) both are influenced by a third variable? Disambiguating these possibilities requires a model of the causal influences (i.e., effective connectivity).

The previous definition of functional connectivity can be applied practically in different ways. For example, for PET or fMRI data, one can compute functional connectivity maps with reference to a particular seed region or, alternatively, use a variety of multivariate techniques that find sets of voxels whose time series represent distinct (orthogonal or independent) components of the covariance structure of the data. Because of their central role in analyses of functional connectivity, the following section summarizes approaches that are based on eigenimage decomposition of spatiotemporal data.

Eigendecomposition of Spatiotemporal Imaging Data

Given a matrix A, a vector $\mathbf{x}$ is called the eigenvector of A if there is a scalar eigenvalue λ such that

$$A\mathbf{x} = \lambda\mathbf{x} \qquad [2]$$

In other words, eigenvectors are those vectors which are only scaled by A; their direction remains unchanged. Eigenvectors thus represent the principal axes of the linear function encoded by A.

Let us now consider a PET or fMRI data set of m scans obtained for n voxels that is represented by a mean-corrected $m \times n$ matrix Y. The temporal covariance matrix of this data set is given by YY^T and the spatial covariance matrix is given by Y^TY (where T denotes the transpose operator). The challenge is to characterize those patterns in time and space, respectively, that contribute most to the covariance structure of the data. Focusing on temporal covariance, one could perform a principal components analysis (PCA) that reexpresses the data Y in terms of a new orthonormal basis set U comprising the eigenvectors of the temporal covariance matrix YY^T:

$$Y^* = UY \qquad [3]$$

Here, Y^* are the data after projection onto the space defined by U. In this space, the axes account for extreme and orthogonal components of the temporal covariance structure.

A more general and powerful procedure for eigendecomposition of spatiotemporal data is singular decomposition (SVD). SVD decomposes the data matrix Y into three matrices:

$$\mathrm{SVD}(Y) = USV^T \qquad [4]$$

This is a so-called eigendecomposition because the column vectors of U and V represent the eigenvectors of the temporal and spatial covariance matrices, respectively. Each eigenvector in V contains one value per voxel; therefore, it can be displayed as an image and is also called an eigenimage (or spatial mode). The first eigenimage is the spatial pattern that accounts for the greatest amount of the covariance structure of the data. The second eigenimage is orthogonal to the first and explains the greatest amount of the remaining covariance and so on. The amount of covariance explained by each eigenimage is given by S, which is a diagonal matrix of decreasing singular values (the square root of the eigenvalues). (Note that the eigenvalues of the temporal and spatial covariance matrices are identical and, given a data matrix Y with rank r, S has r singular values which occupy the leading diagonal in decreasing order.) The temporal eigenvectors, or eigenvariates, contained by the columns of U, reflect the degree to which an eigenimage is expressed over time and can help to interpret the function of the distributed neural system defined by the associated eigenimage (e.g., by comparing the temporal evolution of the eigenvariate to the temporal sequence of task conditions).

Eigenimage analysis based on SVD has been elaborated in several ways. Notable among these is canonical variate analysis (CVA), which uses a generalized eigenvector solution to maximize the variance that can be explained by some explanatory variables relative to variance that cannot. CVA can be considered as an extension of eigenimage analysis that refers explicitly to some explanatory variables and allows for statistical inference. A technique closely related to SVD is partial least squares, which has been applied in the context of neuroimaging using behavioral data or time series from a reference voxel as explanatory variables.

As explained previously, analyses of functional connectivity only refer to statistical dependencies among measured data and do not incorporate any knowledge or assumptions about the structure and mechanisms of the neural system of interest. This is appropriate if the system of interest is largely unknown and needs to be approached in an exploratory fashion. The results from functional connectivity analyses can then be used to generate hypotheses about the system. However, given a specific mechanistic hypothesis of how system function results from system structure, models of effective connectivity are generally more appropriate and powerful.

Effective Connectivity

Effective connectivity describes the causal influences that the elements of a neural system exert over another. Determining effective connectivity requires a specific model of the neural system of interest that explains how the observed dynamics of the system result from the interactions among its elements. Such a model accounts for the intrinsic structure of the system (i.e., the structural presence or absence of connections) and the external inputs that perturb the system. Before discussing specific models for functional neuroimaging, it is helpful to describe how system models can be formalized in general.

Although systems can be described mathematically in different ways, the underlying concept is quite universal: a system is defined by a set of elements with n time-variant properties that interact with each other. Here, we choose deterministic differential equations with time-invariant parameters as a mathematical framework. Each time-variant property x_i ($1 \leq i \leq n$) is called a state variable, and the n-vector $\mathbf{x}(t)$ of all state variables in the system is called the state vector or simply state) of the system at time t:

$$\mathbf{x}(t) = \begin{bmatrix} \mathbf{x}_1(t) \\ \vdots \\ \mathbf{x}_n(t) \end{bmatrix} \qquad [5]$$

For example, the system elements could correspond to individual neurons, each of which is represented by one or more state variables. These state variables could refer to neurophysiological properties (e.g., postsynaptic potentials and status of ion channels). The crucial point is that the state variables interact with each other; that is, the evolution of each state variable depends on at least one other state variable. For example, the postsynaptic membrane potential depends on which and how many ion channels are open, and the probability of voltage-dependent ion channels opening depends on the membrane potential. Such functional dependencies among the state variables of the system can be expressed quite naturally by a set of ordinary differential equations that operate on the state vector:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_n(x_1, \ldots, x_n) \end{bmatrix} = F(x) \qquad [6]$$

However, this description is not yet sufficient. First, the functional dependencies f_i must be specified. This requires a set of parameters θ which represent the form and strength of causal influences between the state variables. In neural systems, these parameters usually correspond to time constants or synaptic connection strengths. Second, we need to consider exogenous inputs into the system (e.g., sensory information entering the brain). Here, we represent the set of all m known inputs by the m-vector function $u(t)$. Extending eqn [6] accordingly leads to a general state equation for nonautonomous deterministic systems:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = F(x, u, \theta) \qquad [7]$$

A model of this form provides a causal description of how system dynamics results from system structure because it describes (1) when and where external inputs enter the system and (2) how the state changes induced by these inputs evolve in time depending on the system structure. Given a particular temporal sequence of inputs $u(t)$, integration of eqn [7] leads to a complete description of how the dynamics of the system (i.e., the trajectory of its state in time) results from its structure and initial state:

$$x(\tau) = x(0) + \int_0^\tau F(x, u, \theta)\mathrm{d}t \qquad [8]$$

Equation [7] therefore provides a general form for models of effective connectivity in neural systems. (We have made two main assumptions to simplify the exposition. First, it is assumed that all processes in the system are deterministic and occur instantaneously. Random components (noise) and delays could be accounted for by using stochastic differential equations and delay differential equations, respectively. Second, we assume that we know the inputs that enter the system. This is a tenable assumption in neuroimaging because the inputs are experimentally controlled variables, such as changes in sensory stimulation or task set.) In the next sections, some specific models of effective connectivity are discussed.

Structural Equation Modeling

Structural equation modeling (SEM) is a multivariate, hypothesis-driven technique that is based on a structural model representing a hypothesis about the causal relations among several variables. In the context of fMRI, for example, these variables are the measured blood oxygen level-dependent (BOLD) time series $y_1, \ldots, y_n$ of n brain regions and the hypothetical causal relations are based on anatomically plausible connections between the regions. The strength of each connection $y_i \rightarrow y_j$ is specified by a so-called path coefficient which, by analogy to a partial regression coefficient, indicates how the variance of y_i depends on the variance of y_j if all other influences on y_j are held constant.

The statistical model of standard SEM can be summarized by the equation

$$y = Ay + u \qquad [9]$$

where y is an $n \times s$ matrix of n area-specific time series with s scans each, A is an $n \times n$ matrix of path coefficients (with zeros for absent connections), and u is an $n \times s$ matrix of zero mean Gaussian error terms, which are driving the modeled system ('innovations'; see eqn [10]). Parameter estimation is achieved by minimization of the difference between the observed and the modeled covariance matrix Σ. For any given set of parameters, Σ can be computed by transforming eqn [9]:

$$\begin{aligned} y &= (I - A)^{-1}u \\ \Sigma &= yy^T \\ &= (I - A)^{-1}uu^T(I - A)^{-1^T} \end{aligned} \qquad [10]$$

where I is the identity matrix. The first line of eqn [10] can be understood as a generative model of how system function results from the system's connectional structure: the measured time series y results by applying a function of the interregional connectivity matrix – that is, $(I-A)^{-1}$ – to the Gaussian innovations u.

In the special case of fMRI, the path coefficients (i.e., the parameters in A) describe the effective connectivity of the system across the entire experimental session. What one would often prefer to know, however, is how the coupling between certain regions changes as a function of experimentally controlled context (e.g., differences in coupling between two different tasks). Notably, SEM does not account for temporal order: if all regional time series were permuted in the same fashion, the estimated parameters would not change. In case of blocked designs, this makes it possible to proceed as if one were dealing with PET data – that is, to partition the time series into condition-specific subseries and fit separate SEMs to them. These SEMs can then be compared statistically to test for condition-specific differences in effective connectivity. An alternative approach is to augment the model with bilinear terms (cf. eqn [9])

which represent the modulation of a given connection by an experimental condition. In this case, only a single SEM is fitted to the entire time series.

A problem with SEM is that one is obliged to use models of relatively low complexity since models with reciprocal connections and loops are often non-identifiable. As a simple example, consider a fully connected model: in this case, the number of free parameters exceeds the number of observed covariances, which means there are no degrees of freedom. Although some heuristics for dealing with complex models have been proposed, this problem, together with the neglect of temporal order, is a critical limitation in the application of SEM to neural systems.

Dynamic Causal Modeling

An important limitation of all approaches discussed so far is that they operate at the level of the measured signals. Taking the example of fMRI, the model parameters are fitted to BOLD series which result from a hemodynamic convolution of the underlying neural activity. Any inference about interregional connectivity obtained by SEM is only an indirect one because these models do not include the forward model linking neuronal activity to the measured hemodynamic response. In the case of EEG, this forward model means there is a major difference between signals measured at each electrode and the underlying neuronal activity: changes in neural activity in different brain regions lead to changes in electric potentials that superimpose linearly. The scalp electrodes therefore record a mixture, with unknown weightings, of potentials generated by a number of different sources.

The causal architecture of the system that we would like to identify is expressed at the level of neuronal dynamics. Therefore, to enable inferences about connectivity between neural units we need models that combine (1) a parsimonious but neurobiologically plausible model of neural population dynamics and (2) a biophysically plausible forward model that describes the transduction of neural activity into measured signals. Such models make it possible to fit jointly the parameters of the neural and the forward model such that the predicted time series are optimally similar to the observed time series. Dynamic causal modeling (DCM) is an approach in which the combination of models of neural dynamics and biophysical forward models is mandatory. DCMs currently exist for both fMRI data and event-related potentials (ERPs) as measured by EEG/MEG. These modality-specific implementations are summarized briefly in the remainder of this section.

DCM for fMRI

DCM for fMRI uses a simple model of neural dynamics in a system of n interacting brain regions. It models the change of a neural state vector $\mathbf{x}$ in time, with each region in the system being represented by a single state variable, using the following bilinear differential equation:

$$\begin{aligned}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= F\left(\mathbf{x}, u, \theta^{(n)}\right)\\ &= \left(A + \sum_{j=1}^{m} u_j B^{(j)}\right)\mathbf{x} + Cu\end{aligned} \qquad [11]$$

Note that this neural state equation follows the general form for deterministic system models introduced by eqn [7] – that is, the modeled state changes are a function of the system state, the inputs u, and some parameters $\theta^{(n)}$ that define the functional architecture and interactions among brain regions at a neuronal level. The neural state variables represent a summary index of neural population dynamics in the respective regions. The neural dynamics are driven by experimentally controlled external inputs that can enter the model in two different ways (Figure 1): they can elicit responses through direct influences on specific regions (e.g., evoked responses in early sensory cortices; the C matrix) or they can modulate the coupling among regions (e.g., during learning or attention; the B matrices). Given this bilinear state equation, the neural parameters $\theta^{(n)} = \{A, B, C\}$ can be expressed as partial derivatives of F:

$$\begin{aligned}A &= \left.\frac{\partial F}{\partial x}\right|_{u=0}\\ B^{(j)} &= \frac{\partial^2 F}{\partial x \partial u_j}\\ C &= \left.\frac{\partial F}{\partial u}\right|_{x=0}\end{aligned} \qquad [12]$$

The matrix A represents the effective connectivity among the regions in the absence of input, the matrices $B^{(j)}$ encode the change in effective connectivity induced by the jth input u_j, and C embodies the strength of direct influences of inputs on neuronal activity. See Figure 1 for an example.

DCM for fMRI combines this model of neural dynamics with a hemodynamic model that describes the transformation of neuronal activity into a BOLD response. This 'balloon model' consists of a set of differential equations with parameters $\theta^{(b)}$ that describe how changes in neural activity elicit a vasodilatory signal that leads to increases in blood flow and subsequently to changes in blood volume and deoxyhemoglobin content. The predicted BOLD signal is a

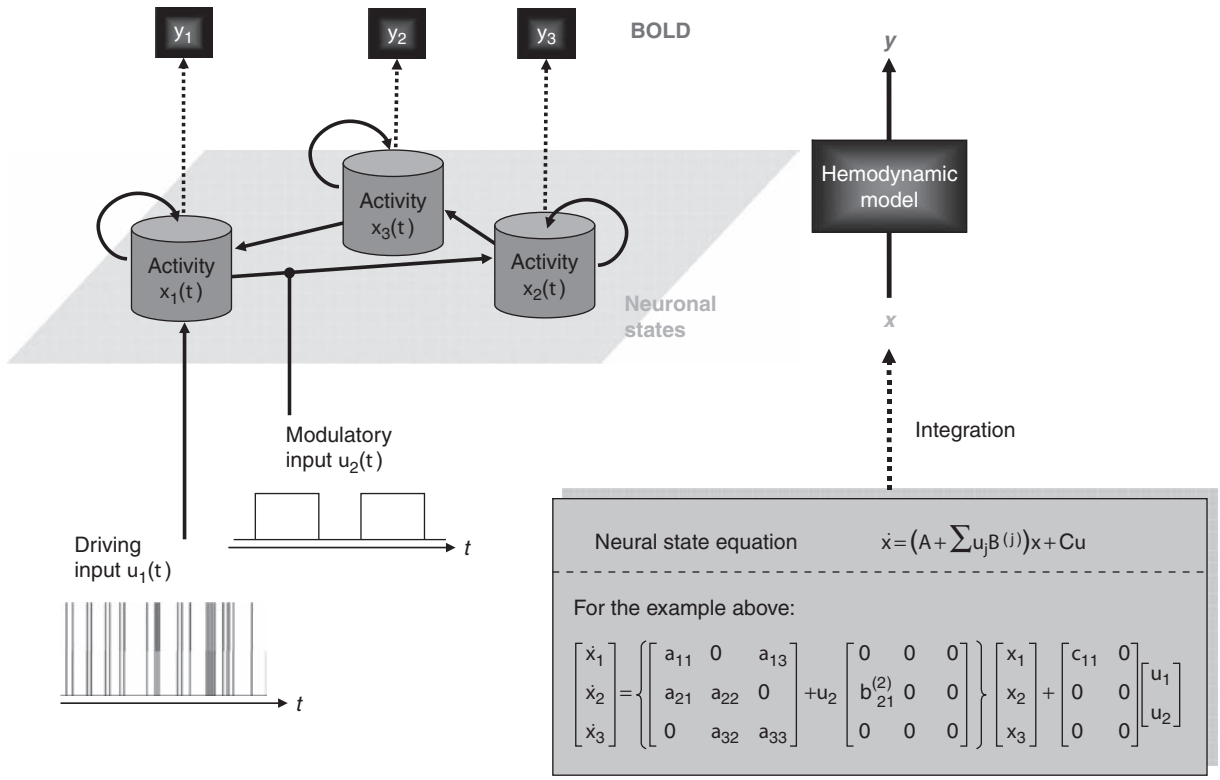

Figure 1 Schematic summary of DCM for fMRI. The dynamics in a system of interacting neuronal populations (top left quadrant), which are not directly observable by fMRI, are modeled using a bilinear state equation. The gray box in the bottom right quadrant shows the general state equation of DCM and its specific form for the network shown in this figure. Integrating the state equation gives predicted neural dynamics (x) which are transformed into predicted BOLD responses (y) by means of a hemodynamic forward model. Neural and hemodynamic parameters are adjusted jointly such that the differences between predicted and measured BOLD series are minimized. The neural dynamics are determined by experimental manipulations that enter the model in the form of external inputs. Driving inputs (u_1; e.g., sensory stimuli) elicit local responses which are propagated through the system according to the intrinsic connections. The strengths of these connections can be changed by modulatory inputs (u_2; e.g., changes in task, attention, or due to learning). Note that in this figure the structure of the system and the scaling of the inputs have been chosen arbitrarily.

nonlinear function of blood volume and deoxyhemoglobin content. The combined neural and hemodynamic parameter $\theta = \{\theta^{(n)}, \theta^{(h)}\}$ set is estimated from the measured BOLD data using a fully Bayesian approach with empirical priors for the hemodynamic parameters $\theta^{(h)}$ and conservative shrinkage priors for the coupling parameters $\theta^{(n)}$. The parameter estimation scheme rests on expectation maximization and uses a Laplace (i.e., Gaussian) approximation to the true posterior. The posterior distributions of the parameter estimates can be used to test hypotheses about connection strengths. Usually, these hypotheses concern context-dependent changes in coupling. If there is uncertainty about the connectional structure of the modeled system, or if one would like to compare competing hypotheses (represented by different DCMs), a Bayesian model selection procedure can be used to find the DCM that exhibits an optimal balance between model fit and model complexity.

DCM for ERPs

ERPs as measured with EEG or MEG have been used for decades to study electrophysiological correlates of cognitive operations. DCM for ERPs was developed as a biologically plausible model to understand how event-related responses result from the dynamics in coupled neural ensembles. It rests on a neural mass model which uses established connectivity rules in hierarchical sensory systems to assemble a network of coupled cortical sources. These rules characterize connections with respect to their laminar patterns of origin and termination and distinguish between (1) forward (or bottom-up) connections originating in agranular layers and terminating in layer 4, (2) backward (or top-down) connections originating and terminating in agranular layers, and (3) lateral connections originating in agranular layers and targeting all layers. These interareal cortico-cortical connections are excitatory and arise from pyramidal cells.

Each region or source is modeled as a microcircuit in which three neuronal subpopulations are combined and assigned to granular and supra-/infragranular layers. A population of excitatory pyramidal (output) cells receives inputs from inhibitory and excitatory populations of interneurons via intrinsic (intra-areal) connections. Within this model, excitatory interneurons can be regarded as spiny stellate cells found predominantly in layer 4 and in receipt of forward connections. Excitatory pyramidal cells and inhibitory interneurons are considered to occupy infra- and supragranular layers and receive backward and lateral inputs.

The neural state equations are summarized in Figure 2. To perturb the system and model event-related responses, the network receives inputs via input connections. These connections are exactly the same as forward connections and deliver input u to the spiny stellate cells in layer 4. Input u represents afferent activity relayed by subcortical structures and is modeled as two parameterized components, a gamma density function (representing an event-related burst of input that is delayed and dispersed by subcortical synapses and axonal conduction) and a discrete cosine set (representing fluctuations in input over peristimulus time). The influence of this input on each source is controlled by a parameter vector **C**. Overall, the DCM is specified in terms of the state equations shown in Figure 2 and a linear output equation

$$\frac{dx}{dt} = f(x, u, \theta)$$
$$y = Lx_0 + \varepsilon \qquad [13]$$

where x_0 represents the transmembrane potential of pyramidal cells, and L is a lead field matrix coupling electrical sources to the EEG channels. In comparison to DCM for fMRI, the forward model is a simple linearity as opposed to the nonlinear hemodynamic model in DCM for fMRI. In contrast, the state equations of DCM for ERPs are much more complex and realistic (compare Figures 1 and 2). As an example, the state equation for the inhibitory subpopulation is

$$\dot{x}_7 = x_s$$
$$\dot{x}_8 = \frac{H_e}{\tau_e}((A^B + A^L + \gamma_3 I)S(x_0)) - \frac{2x_8}{\tau_e} - \frac{x_7}{\tau_e^2} \qquad [14]$$

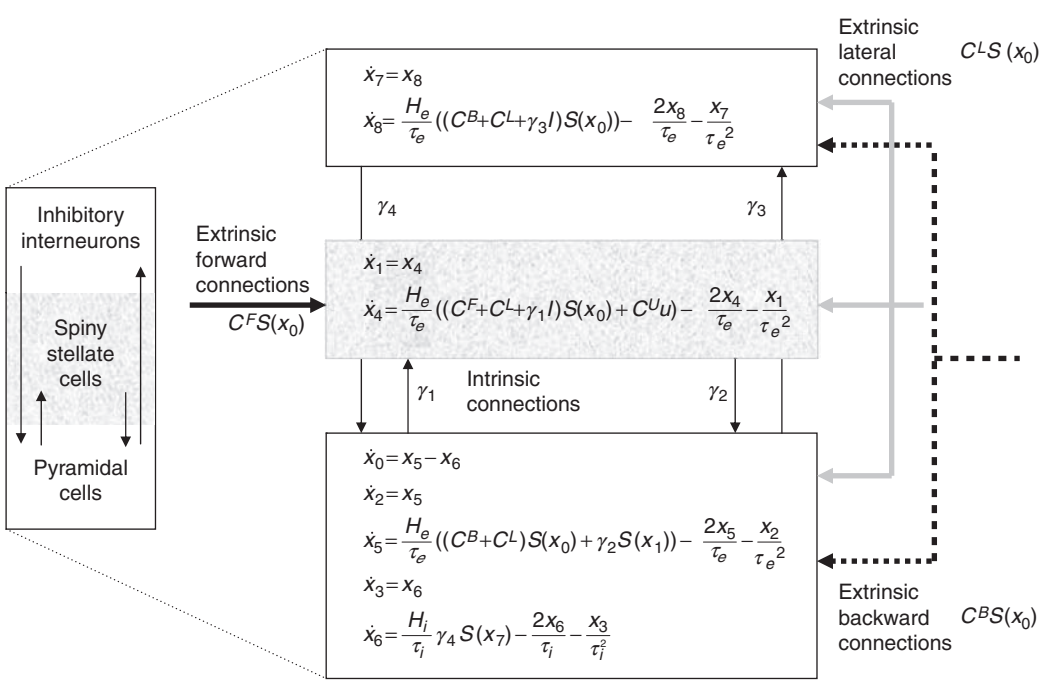

Figure 2 Schematic of the neural model in DCM for ERPs. This schema shows the state equations describing the dynamics of a microcircuit representing an individual region (source). Each region contains three subpopulations (pyramidal, spiny stellate, and inhibitory interneurons) that are linked by intrinsic connections and have been assigned to supragranular, granular, and infragranular cortical layers. Different regions are coupled through extrinsic (long-range) excitatory connections that follow the laminar patterns of forward, backward, and lateral connections, respectively. Reproduced from David O, Kiebel SJ, Harrison LM, Mattout J, Kilner JM, and Friston KJ (2006) Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage* 30: 1255–1272, with permission by Elsevier Ltd.

The parameter matrices A^F, A^B, A^L encode forward, backward, and lateral connections, respectively. Within each subpopulation, the dynamics of neural states are determined by two operators. The first transforms the average density of presynaptic inputs into the average postsynaptic membrane potential. This is modeled by a linear transformation with excitatory (e) and inhibitory (i) kernels parameterized by $H_{e,i}$ and $\tau_{e,i}$. $H_{e,i}$ control the maximum postsynaptic potential, and $\tau_{e,i}$ represent lumped rate constants (i.e., lumped across dendritic spines and the dendritic tree). The second operator S transforms the average potential of each subpopulation into an average firing rate. This is assumed to be instantaneous and is a sigmoid function. Intra-areal interactions among the subpopulations depend on constants $\gamma_{1\ldots4}$ which control the strength of intrinsic connections and reflect the total number of synapses expressed by each subpopulation. In eqn [14], the top line expresses the rate of change of voltage as a function of current. The second line specifies how current changes as a function of voltage, current, and presynaptic input from extrinsic and intrinsic sources. For simplification, our description has omitted the fact that in DCM for ERPs all intra- and interareal connections have conduction delays. This is implemented by delay differential equations.

Summary and Outlook

This article has reviewed different approaches to characterizing functional integration in neural systems on the basis of neuroimaging data. Starting with exploratory analyses of functional connectivity, we discussed various models of effective connectivity that allow for increasingly sophisticated modeling of neurobiological mechanisms. DCM is currently the most advanced framework for system models that is applicable to functional neuroimaging data. Various extensions of DCM are in development, such as application to other data modalities than fMRI or ERPs and enhanced state equations for existing models. The hope is that during the next few years, these developments will contribute to a more mechanistic understanding of brain function. Of particular interest will be the use of neural system models such as DCM to understand the mechanisms of drugs and to develop models that can serve as diagnostic tools for diseases linked to abnormalities of connectivity and synaptic plasticity (e.g., schizophrenia).

See also: Cognition: An Overview of Neuroimaging Techniques; Electroencephalography (EEG); Event-Related Potentials (ERPs) and Cognitive Processing; Executive Function and Higher-Order Cognition: Computational Models; fMRI: BOLD Contrast; Neuroimaging; Perfusion MRI; Positron Emission Tomography (PET).

Further Reading

Büchel C and Friston KJ (1997) Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex* 7: 768–778.

David O, Kiebel SJ, Harrison LM, Mattout J, Kilner JM, and Friston KJ (2006) Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage* 30: 1255–1272.

Friston KJ (1994) Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping* 2: 56–78.

Friston KJ (2002) Beyond phrenology: What can neuroimaging tell us about distributed circuitry? *Annual Reviews in Neuroscience* 25: 221–250.

Friston KJ and Büchel C (2004) Functional connectivity: Eigenimages and multivariate analyses. In: Frackowiack R, et al. (eds.) *Human Brain Function*, 2nd edn., pp. 999–1018. Elsevier: New York.

Friston KJ, Büchel C, Fink GR, Morris J, Rolls E, and Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage* 6: 218–229.

Friston KJ, Harrison L, and Penny W (2003) Dynamic causal modelling. *NeuroImage* 19: 1273–1302.

Gerstein GL and Perkel DH (1969) Simultaneously recorded trains of action potentials: Analysis and functional interpretation. *Science* 164: 828–830.

Horwitz B, Duara R, and Rapoport SI (1984) Intercorrelations of glucose metabolic rates between brain regions: Application to healthy males in a state of reduced sensory input. *Journal of Cerebral Blood Flow and Metabolism* 4: 484–499.

Horwitz B, Tagamets MA, and McIntosh AR (1999) Neural modeling, functional brain imaging, and cognition. *Trends in Cognitive Sciences* 3: 91–98.

McIntosh AR and Gonzalez-Lima F (1994) Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping* 2: 2–22.

McIntosh AR and Lobaugh NJ (2004) Partial least squares analysis of neuroimaging data: Applications and advances. *NeuroImage* 23: S250–S263.

Penny WD, Stephan KE, Mechelli A, and Friston KJ (2004) Comparing dynamic causal models. *NeuroImage* 22: 1157–1172.

Stephan KE (2004) On the role of general system theory for functional neuroimaging. *Journal of Anatomy* 205: 443–470.

Relevant Websites

http://psych.utoronto.ca/~mcintosh – Partial least squares software for neuroimaging data.

http://www.fil.ion.ucl.ac.uk/spm – Statistical parametric mapping software (including DCM).

Functional Electrical Stimulation

R B Stein, University of Alberta, Edmonton, AB, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Many diseases or injuries interrupt the normal conduction of electrical signals such as action potentials. Therefore, people have long tried to replace the natural generation of action potentials by electrical stimulation below the level of the interrupting lesion. Indeed, reports go back over 2000 years to the use of electric eels for stimulation in ancient Greece. Physical therapists and others involved in rehabilitation commonly use electrical stimulation for therapy, to strengthen weakened muscles or increase fatigue resistance; this practice is often referred to as therapeutic electrical stimulation. We and others have shown that 1–2 h of neuromuscular stimulation per day can build muscle strength and restore the fatigue resistance of disused muscles. Rather than a generalized strengthening of muscles, functional electrical stimulation (FES) uses patterned electrical stimulation to assist or replace a particular motor function.

Classes of FES and Related Devices

FES is a general term that can be applied to a variety of functions; some of the devices employed are listed in Table 1 and examples are shown in Figure 1. By far the most successful device has been the heart pacemaker, which is now implanted in more than 400 000 people worldwide every year. Pacemakers replace the normal pattern of heartbeats when the natural cardiac pacemaker is diseased or damaged. The lives of people receiving these implants have been prolonged for years or even decades. Although implanted devices tend to be expensive and have some risk of infection, the life-saving benefits clearly outweigh the costs and risks for most people.

Cochlear stimulators (Figure 1(b)) restore the sense of hearing in completely deaf individuals by patterned stimulation of cochlear neurons in the inner ear. The patterns are determined by an auditory receiver and speech processor (Figure 1(b1)) and are transmitted across the skin to stimulate cochlear neurons using electrodes implanted in the inner ear (Figure 1(b2)). The best stimulators use more than ten channels that are tonotopically organized to match the layout of cochlear neurons and can produce a remarkable degree of recovery in auditory function. Some completely deaf people can understand speech well enough that they can engage in telephone conversations. Deaf children, implanted with a cochlear prosthesis, are often able to attain similar educational achievement as children with some hearing ability. Again, for tens of thousands of individuals (Table 1) the benefits outweigh the costs and risks of implanting a stimulator in the inner ear.

The two cochlear applications differ from the other devices in Table 1 in important ways. Heart pacemakers stimulate heart muscle directly, while cochlear neurons stimulate sensory neurons. Other techniques, such as transcutaneous electric nerve stimulation (TENS), spinal cord stimulators, and acupuncture, also act by stimulating sensory fibers. Their effects, in alleviating pain, for example, could be classified as therapeutic, although the term 'neuromodulation' is often used. This term derives from the fact that many peptide transmitters, such as are released by pain fibers, act on a much longer time scale than classic neurotransmitters do and hence modulate the state of a system rather than providing detailed sensory or motor information. Similarly, deep brain stimulators for Parkinson's disease and other conditions act as neuromodulators in reducing tremor and enabling movements, rather than specifying the pattern of movement to be executed. Therapeutic electrical stimulation and neuromodulation are not considered further here, and the focus will instead be on the replacement of function.

The rest of the FES devices in Table 1 act by stimulating nerve fibers, either in a nerve trunk or at the neuromuscular junction. As a result, the term 'functional neuromuscular stimulation' (FNS) is sometimes used for these devices. Even when the electrodes are placed within a muscle, the lowest threshold is found along the nerve or near the neuromuscular junction. The reason is the tremendous amplification of the neuromuscular junction. The amplification occurs since a small current generated in a nerve can release large numbers of molecules of the transmitter acetylcholine (ACh), each of which can open channels in the much larger muscle membrane and create much larger currents. In addition, a motor nerve fiber often branches profusely to innervate hundreds of muscle fibers, thus providing further amplification.

Finally, the capacitance of the muscle membrane has long been known to be much greater than the capacitance of the nerve membrane, because of the additional membranous systems, such as the transverse tubules in muscle fibers. Thus, the capacity of

Table 1 Estimated number of users of various FES devices

Device	*Number of users*
Heart pacemaker	>400 000 yr^{-1}
Cochlear stimulator	>80 000 total
Bladder stimulator	>2500 total
Phrenic nerve stimulator	1500 total
All FES limb stimulators	<10 000 total

the muscle membrane to store charge is much greater than that of a nerve membrane and much more current must be supplied to depolarize muscle cells directly. Stimulating denervated muscles with large, long-lasting currents has been claimed to prevent or reverse atrophy to some extent. However, the stimulation has not been shown to be functional for particular motor tasks, so further discussion here will be restricted to stimulating nerves for FES.

An example of an FES device for motor tasks that has benefited thousands of people is the bladder stimulator (Figure 1(c); Table 1). With this device, the push of a button produces urination. Previously, people with complete spinal cord injury needed intermittent catheterization or other methods. However, urination has the complication that the bladder muscle must be contracted while the urinary sphincter must be relaxed. In practice, the anterior roots to bladder and sphincter are stimulated at the same time. However, the sphincter is fast skeletal muscle and the bladder is slow smooth muscle, so the bladder continues to contract after the sphincter has relaxed. Clinically, the method requires that sacral dorsal roots be cut, since stimulating these afferents would cause a longer reflex contraction of the sphincter that would block the flow of urine. Various methods are being considered to overcome this problem, but they are outside the scope of this short review.

Similarly, phrenic nerve stimulators (Figure 1(d)) can generate breathing movements in people who have high spinal cord injuries, freeing these people from use of respirators. The late Christopher Reeve benefited from a phrenic nerve stimulator in his final years. Recently, the use of laparoscopy to insert electrodes into the diaphragm has been tested; this procedure could eliminate the need for a thoracotomy and its associated risks. One remaining problem with current devices is that speech must be phrased appropriately with the expiration phase of the stimulator, rather than the pattern of breathing adjusting to the pattern of speech, as occurs in the respiration of able-bodied people.

All of the devices described here are surgically implanted and have been successful clinically to a greater or lesser extent, depending on the relative benefits and costs (both monetary and in terms of risks or drawbacks of the procedures). FES devices to replace limb movements have been less successful, and it is estimated that probably less than 10 000 devices have been fitted worldwide. The vast majority of these have been surface stimulators, so the difference between surface and implanted stimulators must be considered. Actually, limb stimulators for functional movements fall into three categories, based on whether they use implanted, percutaneous, or surface electrodes, and these will be considered in turn.

Electrodes

Implanted Electrodes

Implanted electrodes generally require open surgery. An exception is the BION ('bionic neuron'), developed by Loeb and colleagues, which is small enough that it can be implanted through the barrel of a hypodermic needle near a nerve or neuromuscular junction of interest. An implanted stimulator may contain (1) a battery that will eventually run down and need replacement, such as in a heart pacemaker, (2) a rechargeable battery that is recharged periodically by sending power as a radio wave (telemetry), which is received and processed (e.g., rectified) by implanted circuitry, or (3) no battery, so power must be sent continuously by telemetry whenever operation of the device is desired. Which of the three options is best depends on the total power requirements, the space available for batteries in the device, and the consequences of power loss (e.g., is it life-threatening or not?). In addition to power, the implanted devices must have control circuitry. This can be very simple, such as a pulse generator to generate a pulse about once per second, as in a heart pacemaker, or a series of pulses every few seconds, as in a phrenic nerve stimulator. Where the processing is quite complex, such as for a cochlear stimulator, much of the processing may be external and only a series of commands that encode the stimulation parameters are sent to the internal device periodically. For the bladder stimulator a switch is external, but the circuitry generating the pulses is internal.

Percutaneous Electrodes

Percutaneous electrodes are injected, as for the BION, but the insulated wire from the electrode comes through the skin. This eliminates the need for open surgery, but provides a path for infection to enter through the skin. In addition, stress points occur

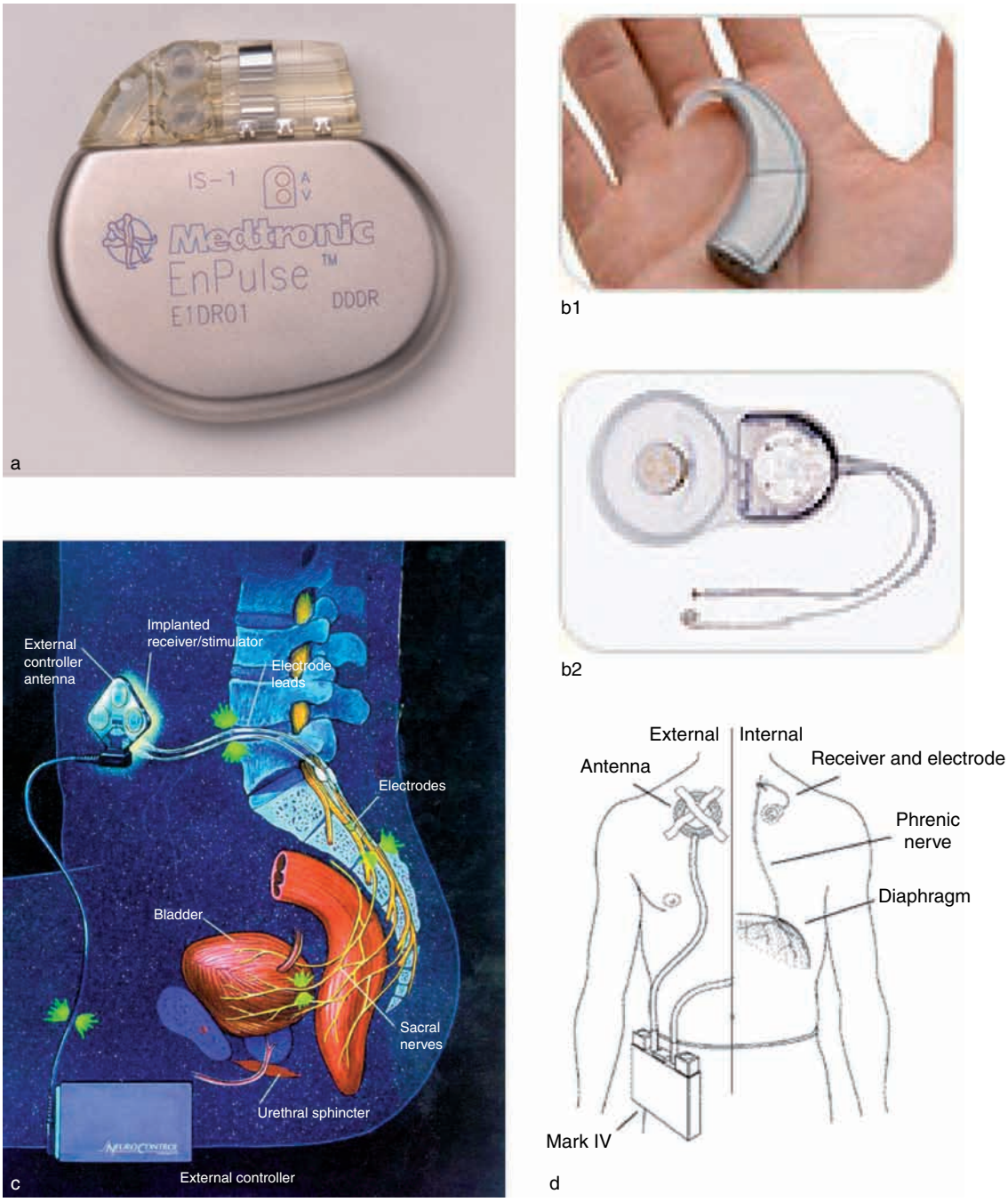

Figure 1 (a) A cardiac pacemaker (EnPulse, Medtronic, Inc.). Leads go from the top ports, to stimulate the heart muscle to contract and expel blood from the heart with each heartbeat. (b) A cochlear stimulator (Nucleus Freedom, Cochlear Ltd.). The microphone and speech processor are external (1), but the one shown is small enough to wear behind the ear. Coded signals are then transmitted across the skin to the receiver and stimulating electrodes (2) that are implanted in the canal of the inner ear. Different auditory frequencies stimulate electrodes that activate different cochlear neurons, according to the normal tonotopic organization. (c) A bladder stimulator (Brindley Bladder System, FineTech Medical Ltd.). To urinate, a command from the external controller is transmitted to the receiver, which activates the implanted electrodes on the sacral roots. The bladder and urethral sphincter contract, but the sphincter relaxes first so that urine is expelled. (d) Schematic diagram of a phrenic pacemaker (the Mark IV version developed by Glenn and colleagues). An external pacemaker transmits signals from the antenna across the skin to the implanted electronics. Stimulation of the phrenic nerve contracts the diaphragm and expels air. As the diaphragm relaxes, air enters the lungs. External components are shown on the left side of the picture; internal components are implanted (right side).

where the wires cross the skin and fascial planes between muscles. Frequent bending at these points can eventually cause the wire to break. Percutaneous wires have been used to develop both arm and leg systems, and the properties of metals used in the electrodes have been improved to reduce breakage. However, percutaneous electrodes are now rarely used for long-term systems.

Surface Electrodes

Surface electrodes are applied to the outside of the skin over a nerve or motor endplate and depend on a conductive path through the skin to the underlying target. The conductive path is most commonly a hydrogel, but can be as simple as water, which has a low but usable electrical conductance (thousands of stimulators that used water-filled electrodes were fitted in the former Yugoslavia). Hydrogels are generally sticky enough to adhere to the skin and provide a higher conductance path, particularly if the skin and surface of the electrode are moistened somewhat with water or there is some sweating through the skin under the electrodes. A continuing problem with surface electrodes is that even hypoallergic electrodes can produce allergic reactions or become a site for skin infections (dermatitis) to develop. Passage of electricity for many hours a day can also produce pinhole lesions in the skin, depending on the intensity and shape of the stimulus current (see later). Skin problems are of particular importance in people who have diabetes or other conditions that compromise the health of the skin, and can be more severe in obese individuals, since subcutaneous fat is a poor conductor of electricity. Increasing the voltage to pass more current through the high resistance will increase discomfort by stimulation of pain fibers near the surface of the skin and will increase the chance of lesions. Some disorders produce conditions in which even normal sensations become noxious (allodynia). This may preclude the use of surface and even implanted electrodes. Finally, surface electrodes can only be used to excite nerves reliably that lie close to the surface of the skin. Despite these problems, the vast majority of limb stimulators have used surface electrodes, as will be described later. However, before describing the types of limb stimulators, we need to consider stimulus waveforms briefly.

Stimulus Waveforms

A stimulator can control either the voltage or current applied to the electrodes, irrespective of whether implanted, percutaneous, or surface electrodes are used. A constant voltage is simpler to control (e.g., a battery), although constant-current stimulators are often considered to be superior. The reason is that the current, flowing through a tissue and across the resistance of a nerve membrane, depolarizes the nerve to the threshold for generating an action potential. If the resistance of the skin increases due to the electrode drying out or decreases due to sweating, a constant-current stimulator will adjust automatically. This is typically done by sensing the current flowing through the electrode on each pulse and adjusting the voltage up or down to compensate for any changes, thereby maintaining the current at a desired level. However, constant-current stimulators may be dangerous if an electrode pulls away from the skin and loses contact with all but a small area of skin. Then, a constant-current stimulator will maintain the current, but the current density flowing through the small remaining area of contact can be increased enormously and may lead to skin damage.

For implanted or percutaneous electrodes, the difference between constant-current and constant-voltage stimulators is minimal in the chronic state. The tissue resistance, which links these two variables through Ohm's law, remains constant from day to day, although there is typically a slow increase in resistance of about 30% in the first few weeks after implantation, as connective tissue grows around the electrode and replaces lower resistance body fluids.

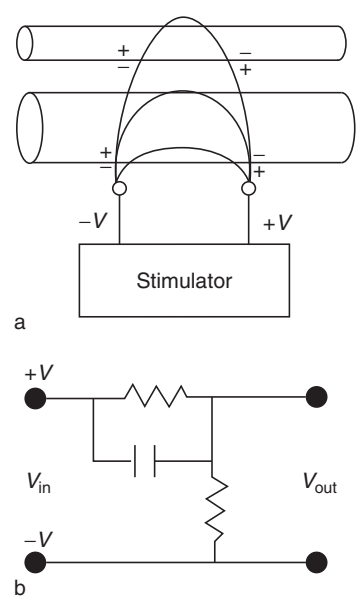

Figure 2 (a) Positive and negative charges from the stimulator will bind to the outside of nerve cell membranes and attract negative and positive charges from inside the nerve membrane. If the inside of the membrane near the negative electrode (cathode) becomes sufficiently depolarized, action potentials will be generated. This occurs preferentially in large axons, as explained in the text. (b) The electrodes have a capacitance as well as a resistance, so the input voltage to the electrodes will differ from the output that is actually applied to the tissue (see **Figure 3**).

All stimulators contain at least two electrodes and the current will flow from the positive electrode to the negative electrode. Figure 2 illustrates the path of current flow through the tissue. The region outside the nerve fiber will be more negative near the negative electrode. Since the membrane has a capacity to store charge (capacitance), and unlike charges attract, positive charge will be attracted to the inside of the membrane and can depolarize the membrane to threshold for firing an action potential. Thus, action potentials are generated near the negative electrode (cathode). The flow of current to depolarize the membrane will be much easier in larger fibers, since a fiber twice as large will have 4 times the cross-sectional area and thus one-quarter the resistance to currents flowing along the fiber. The membrane area will be twice as large, but supplied by 4 times the current, so the larger fibers will be more readily depolarized to threshold. This principle is important in enabling large motor nerve fibers to be stimulated without activating small pain fibers. However, other factors, such as proximity, play a role, and fibers nearer to the electrode will be stimulated before more distant ones are. Proximity is particularly important with surface electrodes, since pain fibers will lie in the subcutaneous tissue immediately below the skin and stimulating deep-lying muscles before stimulating pain fibers will be difficult.

Electrodes are complex interfaces that include capacitors as well as resistors. The circuit diagram in Figure 2(b) is a simplification that ignores the spatial properties of the current flow, described in Figure 2(a), as well as the full complexity of the electrode–solution interface, but will illustrate the needed points. Even though the applied voltage (or current) is constant, these variables will not be constant in the tissue, but will change as the electrode's capacitor charges up. The tissue contains a capacitance from the membranes of nerve and other cells, but this has also been ignored in the simplified diagram of Figure 2(b). The voltage in the tissue will actually reverse in sign as the electrode's capacitance discharges (Figure 3(b)), even though the applied voltage does not reverse (Figure 3(a)). Depending on the voltage built up across the capacitor, metal

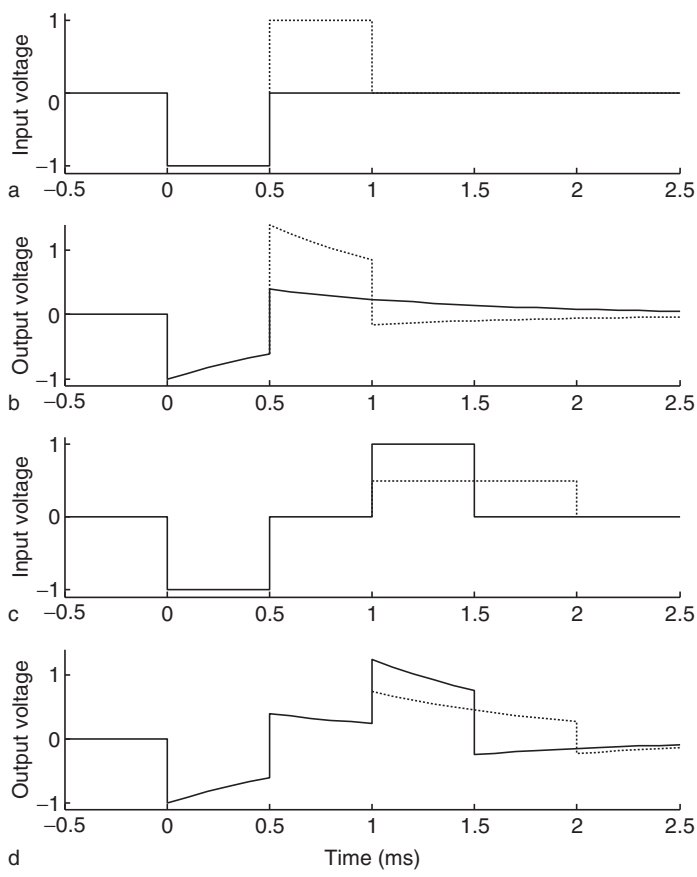

Figure 3 If a negative-going voltage pulse (solid line in (a)) is applied to electrodes with the circuit diagram of Figure 2(b), the resulting waveform in the tissue will have negative and positive components (solid line in (b)). If the applied pulse is balanced by an equal and opposite waveform (dotted line in (a)), the positive-going voltage in the tissue will be greater in magnitude than the initial negative-going pulse (dotted line in (b)), which can cause excitation at the second electrode. This problem can be alleviated to some extent by delaying the opposite waveform (solid lines in (c) and (d)) and/or changing its time course (dotted lines in (c) and (d)).

particles from the electrode will be lost, including the surface coating. The loss of metal particles will lead to erosion and eventual breakage of the electrode, so this is an important variable in the design of implanted systems. Most stimulators produce biphasic pulses, with partial or complete charge balance. Charge is the integral of current over time, so for complete charge balance the area under the curve of the applied currents as a function of time should be the same for both phases.

Charge balancing is an attempt to recover the material from the electrode that was lost during the stimulation pulse, but many of the metal and other ions will diffuse away. They cannot be recovered, but are replaced by ions from the tissue. For this reason, full charge balance may not be needed. If the second phase is equal and opposite to the first phase, a higher absolute voltage will be reached in the tissue during the second phase (**Figure 3(b)**). The positive electrode (anode) in the first phase will become the cathode in the second phase and may excite nerve fibers. To minimize this secondary excitation, the second phase can be delayed (**Figure 3(c)**) or have a smaller amplitude over a longer period of time, so that the area under the curve is still the same (**Figure 3(d)**). However, the delays allow more opportunity for metal ions to diffuse away. Different implanted FES systems use some of the options illustrated in **Figure 3** for charge balance. Charge balance is also important for surface systems, since injecting materials from the electrode through the skin is thought to lead to some of the unwanted tissue reactions described earlier.

Foot-Drop Stimulators

Liberson and colleagues proposed the first FES system for limb function. They considered the condition of foot drop that occurs commonly after stroke or other central nervous system disorders. In this condition dorsiflexion of the ankle is prominently affected and the foot drops and may drag on the ground during the swing phase of the step cycle. Liberson et al. suggested use of a heel sensor to detect when the heel comes on and off the ground. They stimulated the common peroneal nerve, when the heel came off the ground, to flex the ankle and thereby lift the foot until the heel contacted the ground again. Yet, FES is not the current solution for most patients, but rather use of a plastic brace (ankle–foot orthosis, or AFO) physically prevents the foot from dropping. In fact, hundreds of thousands of AFOs are fitted every year, while probably less than 10 000 FES devices have been fitted for all limb problems since Liberson's suggestion 45 years ago. When given the choice, people overwhelmingly prefer an FES system. Until recently, no US Food and Drug Administration (FDA)-approved FES systems were available for sale in the United States, but three systems have recently come on the market (WalkAide; Odstock Dropped Foot Stimulator; and Ness L300) and the market will determine whether these systems are widely accepted. These systems all use surface stimulation, but implanted stimulators using open surgery or BIONs are being developed. The cost of a sophisticated surface stimulator will inevitably be higher than that of a plastic brace, but the advantages of actively dorsiflexing the ankle, rather than passively fixing it at about 90°, may be sufficient to outweigh the costs. These advantages include (1) flexing the ankle to clear the ground without 'hiking' the hip or circumducting the leg around the body, (2) strengthening the dorsiflexor muscles and any residual descending connections, and (3) being able to wear the same size shoe on both feet or even to go barefoot. This is possible with the WalkAide (see **Figure 4(a)**), which has a tilt sensor built in, rather than using a heel switch.

The second point, the advantage of strengthening muscles and connections, is interesting, since several studies have reported a carryover effect of using the stimulator, both in the short and long term. In other words, after using the stimulator for a period of time, some but not all persons report being able to walk better even without stimulation. This is confirmed by measuring walking speeds and the effort required to walk a given distance. Typically, about half the improvement seen with the stimulator is still present when the stimulator is turned off. This improvement could be due to strengthening leg muscles and spinal circuits by more frequent use. However, recent studies with transcranial magnetic stimulation (TMS) show that the motor evoked potentials (MEPs) recorded in the tibialis anterior muscle, a prime ankle dorsiflexor, are increased by stimulating the common peroneal nerve for 10–30 min. This increase is observed in study controls, as well as in persons requiring a foot-drop stimulator, but the increased MEPs return to control levels within an hour or so. Preliminary evidence from my lab suggests, that by using FES on a daily basis, these transient increases are maintained (**Figure 5**). However, increases in corticospinal connections are not found in all persons, and the reasons for the differences between persons are not known.

Walking Systems after Complete Paraplegia

New training systems that enhance walking abilities in people with incomplete spinal cord injury have been developed, including treadmills, to support

Figure 4 (a) Stimulator for correcting foot drop (WalkAide, Innovative Neurotronics, Inc.). The stimulator contains an inbuilt tilt sensor for control and attaches to a cuff that contains the surface electrodes and fits on the leg below the knee. (b) Stimulator for a walking system for paraplegics with complete spinal cord injury (Parastep, Sigmedics, Inc.). The connectors shown go to surface electrodes on each leg and to a hand switch on a walker for control of the stimulation. (c) System consisting of brace with electrodes built in for stimulating hand grasp and release (H200, Bioness, Inc.) (d) Second-generation implanted hand system developed by Peckham and colleagues, containing ten electrodes for stimulating intrinsic muscles of the hand and muscles above and below the elbow. Also shown is an implanted, Hall-effect, joint angle sensor which can be used to control the external control unit in persons who have some voluntary movement of the wrist.

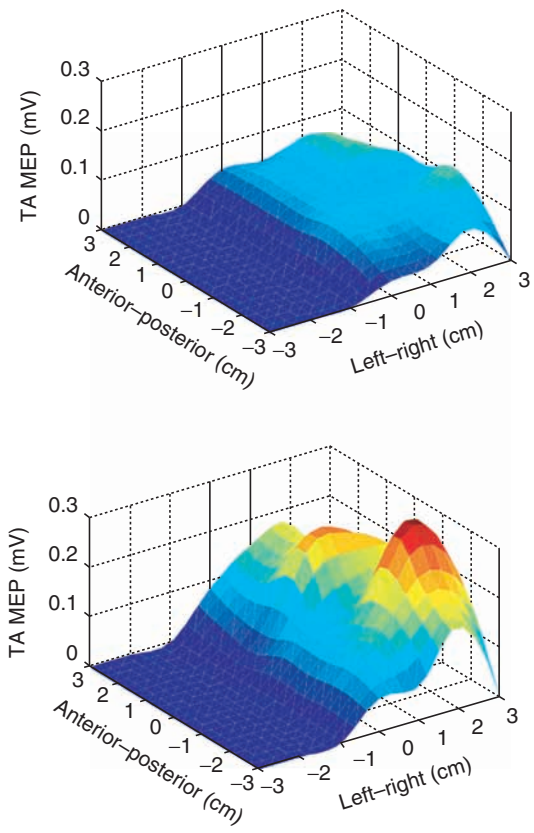

Figure 5 Effect of 6 months of FES in a head injury patient who began to use WalkAide (see **Figure 4(a)**) 19 years after his injury, showing the motor evoked potentials (MEP) in the tibialis anterior (TA) when the stimulator was applied at various positions anterior, posterior, left, and right from the vertex of the head. (Top) Before using WalkAide. (Bottom) After using WalkAide2 for 6 months. The valley in the center is the normal leg area of the motor cortex. The head injury destroyed this area, although surround areas retained some residual connections. These residual connections were strengthened substantially after using FES, as can be seen by comparing before-use and after-use effects. Adapted from Stein RB, Chong S, Everaret DG, et al. (2006) A multicenter trial of footdrop stimulator controlled by a tilt sensor. *Neurorehabilitation and Neural Repair* 20: 371–379.

part of the body's weight, and robotic manipulators. These devices probably act on the spinal pattern generator for walking and residual descending pathways. Many individuals with incomplete injuries are able to regain functional walking with only crutches or walkers for balance. Others still have problems such as foot drop, which can be treated with the foot-drop stimulators. However, these techniques have proved ineffective in people with complete spinal cord injuries. Systems for restoring walking have been developed for people with a complete paraplegia at the thoracic level. The simplest system involves stimulating the quadriceps muscles, to lock the knees in stance, and the common peroneal nerve, to produce a flexion reflex to enable the leg to swing forward. Others add additional channels of stimulation, such as the gluteal or paraspinal muscles, to enhance upright stance. One such system, Parastep, has received FDA approval and is shown in **Figure 4(b)**. Fatigue remains a problem because of the high level of stimulation required of the quadriceps to maintain the standing posture. The walking is slow and is typically limited to the order of 10 m because of the very high energy consumption.

Popovic and colleagues suggested the concept of a hybrid system which used bracing to maintain the upright posture and FES for propulsion, and a number of these systems have been developed. Some are based on existing braces, such as the reciprocal gait orthosis (RGO). RGOs brace the ankle, knee, and hip. They allow a stiff-legged gait without FES in many paraplegics, and the walking can be enhanced with FES. Speed and endurance are better and the energy cost is less, but the system is cumbersome and difficult to don and doff except from a bed. Sitting and standing up after putting the device on can be difficult. New types of braces (stance control) that allow the knee joint to be locked and unlocked have been developed and control only the joints from the knee down (knee–ankle–foot orthosis; KAFO). A stance-control KAFO can be donned and doffed while sitting in a wheelchair, and my group has added FES to allow easy standing and propulsion for walking. Several groups have developed systems using implanted electrodes to allow stimulation of more muscles and better control of the stepping movements. So far, none of these systems has been commercialized and published data are based on a few intensively trained individuals. No studies directly compare the best implanted systems and the best surface systems, so the relative costs and benefits of surgery are still unknown.

Systems for Hand Grasp

Several FES systems have also been developed for allowing opening of the hand to grasp objects. One system, Ness H200 (**Figure 4(c)**), has received FDA approval and is available for sale in the United States. The H200 is intended for people with stroke or other central nervous system disorders that produce flexor spasticity. In other words, the hand remains in a flexed posture and stimulating the extensors allows the hand to be opened to grasp objects. The stimulation can be used therapeutically to reduce spasticity, as well as in functional tasks. A more ambitious, eight-channel, implanted system was developed by Peckham and colleagues at Case Western University, and successfully tested on about 150 spinal cord-injured individuals. However, the market did not grow fast enough for the company that had commercialized it (Neural Control, Inc.) and the device was withdrawn from the market. Work continues at Case

Western and elsewhere; hopefully, this proven system will again be available, probably in an updated form.

Multijoint Control of the Arm

Clearly, the higher the level of a spinal cord injury and the more joints that must be controlled, the more difficult the problem. In addition, the spinal lesion will denervate some muscles of the arm, so tendon transfers or other surgical procedures may be needed to restore missing functions. Despite the challenges, surface, percutaneous, and implanted FES systems have been developed to control arm movements. To simplify the control problem, synergies involving more than one joint are used, but the movements remain slow and tedious to an external observer. However, for the severely affected individuals, these systems represent progress toward regaining control of their environment. Volunteers are eager to participate in testing these systems, hoping to speed the progress of the product from the laboratory to the community. The higher the level of the injury, the fewer biological control signals are available. Among those that have been used are the contralateral shoulder, electromyogram control from shoulder muscles, and tooth clicks. The ultimate challenge is to record signals from the motor cortex of individuals who have no voluntary arm movements and use these signals about intended movements to control all of the muscles of the arm with FES. Current implants in the motor cortex of humans by Donoghue's group and in nonhuman primates by several groups offer a proof of principle that such a venture is feasible, although many technical problems remain. In conclusion, although the field of FES for limb control has been slow to emerge, several commercial systems are now available. If successful, these systems can form the basis for advancing toward more sophisticated systems that help people with more severe disabilities to participate in a range of activities and to live a fuller and more rewarding life.

See also: Cochlear Prosthesis; Deep Brain Stimulation; Motor Skill Learning; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation.

Further Reading

Chae J and Yu D (1999) Neuromuscular stimulation for motor relearning in hemiplegia. *Critical Reviews in Physical and Rehabilitation Medicine* 11: 279–297.

Creasey GH, Ho CH, Triolo RJ, et al. (2004) Clinical applications of electrical stimulation after spinal cord injury. *Journal of Spinal Cord Medicine* 27: 365–375.

DiMarco AF (2005) Restoration of respiratory muscle function following spinal cord injury. Review of electrical and magnetic stimulation techniques. *Respiratory Physiology & Neurobiology* 147: 273–287.

Gaunt RA and Prochazka A (2006) Control of urinary bladder function with devices: Successes and failures. *Progress in Brain Research* 152: 163–194.

Hochberg LR, Serruya MD, Friehs GM, et al. (2006) Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* 442: 164–171.

Horch KW and Dhillon GS (2004) *Neuroprosthetics: Theory and Practice*. Hackensack, NJ: World Scientific Publishers.

Kralj A and Bajd T (1989) *Functional Electrical Stimulation, Standing and Walking after Spinal Cord Injury*. Boca Raton, FL: CRC Press.

Liberson WT, Holmquest HJ, Scott D, et al. (1961) Functional electrotherapy, stimulation of the peroneal nerve synchronized with the swing phase of the gait of hemiplegic patients. *Archives of Physical Medicine and Rehabilitation* 42: 101–105.

Loeb GE, Peck RA, Moore WH, et al. (2001) BION system for distributed neural prosthetic interfaces. *Medical Engineering & Physics* 23: 9–18.

McNeal DR (1977) 2000 years of electrical stimulation. In: Hambrecht FT and Reswick JP (eds.) *Functional Electrical Stimulation: Applications to Neural Prostheses*, pp. 3–35. New York: Marcel Dekker.

Nathan RH (1993) Control strategies in FNS systems for the upper extremities. *Critical Reviews in Biomedical Engineering* 21: 485–568.

Peckham PH, Keith MW, Kilgore KL, et al. (2001) Efficacy of an implanted neuroprosthesis for restoring hand grasp in tetraplegia: A multicenter study. *Archives of Physical Medicine and Rehabilitation* 82: 1380–1388.

Popovic D, Tomovic R, and Schwirtlich L (1989) Hybrid assistive system: Neuroprosthesis for motion. *IEEE Transactions on Biomedical Engineering* BME37: 729–738.

Solomonow M, Aguilar E, Reisin E, et al. (1997) Reciprocating gait orthosis powered with electrical muscle stimulation (RGO II). Part I: Performance evaluation of 70 paraplegic patients. *Orthopedics* 20: 315–324.

Stein RB, Chong S, Everaert DG, et al. (2006) A multicenter trial of a footdrop stimulator controlled by a tilt sensor. *Neurorehabilitation and Neural Repair* 20: 371–379.

Stein RB and Mushahwar V (2005) Reanimating limbs after injury or disease. *Trends in Neuroscience* 28: 518–524.

Relevant Websites

http://www.bionessinc.com – Bioness, Inc. (Ness L300, Ness H200).
http://www.cochlear.com – Cochlear Ltd. (Nucleus Freedom).
http://www.finetech-medical.co.uk – FineTech Medical Ltd. (Brindley Bladder System).
http://www.walkaide.com – Innovative Neurotronics, Inc. (WalkAide).
http://www.medtronic.com – Medtronic, Inc. (EnPulse).
http://www.odfs.com – NDI Medical (Odstock Dropped Foot Stimulator).
http://www.sigmedics.com – Sigmedics, Inc. (Parastep).

Functional Imaging of the Motor System

S Grafton, University of California at Santa Barbara, Santa Barbara, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuroimaging is used to generate maps of cortical and subcortical regions involved in the control of movement. Historically, human neuroimaging measured regional brain activity with positron emission tomography imaging that utilized radioactive isotopes to quantify cerebral glucose metabolism or blood flow. These techniques have been largely replaced by functional magnetic resonance imaging (fMRI) of blood oxygen level-dependent (BOLD) signal changes. The BOLD signal correlates with changes of blood flow, local field potentials, and to a lesser degree intrinsic neural activity. Changes of the BOLD signal are not specific to inhibitory or excitatory neurons and reflect a complex combination of inputs, intrinsic neurons, and output neurons within a given brain area. Spatial resolution can be as good as 1 mm, and absolute temporal resolution is limited by the latency of the BOLD response to 3–4 s. However, relative differences in the time of activation between brain areas can be measured with a higher degree of precision.

The most common goal in brain mapping is to localize specific functional attributes within brain regions. This is done with 'subtraction' paradigms and univariate statistics. In subtraction paradigms, experimental conditions are manipulated so that participants are scanned during at least two predefined behavioral states. For example, a participant might perform a visuomotor tracking task versus rest. Differences in brain activity between these two conditions will identify brain areas associated with movement execution (Figure 1). By the use of multiple regression, it is possible to employ more-complex behavioral paradigms, with which multiple trial types can be directly compared or brain activity can be correlated with performance variables such as speed or reaction time. A less common approach is to characterize interactions between sets of brain regions and to determine whether these network interactions are task specific or change as a function of learning or intervention. This approach can be used with multivariate statistics, including partial least squares or independent component analysis. The rich temporal information of neural processing measured with fMRI also provides the opportunity to use dynamic causal modeling or coherence mapping to examine interactions between brain regions. For example, with coherence measurements it is possible to identify cortical areas showing strong task-specific interactions between the supplementary motor area and the motor cortex during voluntary movement (Figure 2). Similarly, it is possible to show that the onset of activity in one brain area with respect to another, that is, their relative phasic relationship, can change as a function of task (Figure 3).

Execution-Related Areas

Imaging studies of voluntary movement compared with rest reveal a large-scale circuit of both cortical and subcortical brain regions (Figure 1). The motor execution circuit involves the sensorimotor cortex as well as premotor areas located on both the mesial and the lateral convexity of the hemisphere. There is typically involvement of the intraparietal sulcus and adjacent superior and inferior parietal lobules and multiple subcortical projection areas, including striatum, internal segment of the globus pallidus (GPi), ventrolateral thalamus, and cerebellar cortex and associated output nuclei. Within the sensory and motor cortex, there is evidence for a crude homunculus with activity related to leg movements in the most superior position, mouth movements in the most inferior aspect, and limb movements in a mid position along the cortical convexity. Within-limb somatotopy of the arm is observed in the motor cortex when measured at a very low spatial resolution. At high resolution there is evidence for a complex, fractured representation, with multiple representations of each joint of the hand and arm distributed throughout the motor cortex. In right-handed people, there is strong evidence for cortical lateralization to the left hemisphere for right-handed movements, particularly in premotor and motor cortex. In contrast, left-handed movements are associated with a more symmetric recruitment of cortical areas. This provides some evidence that the left hemisphere motor areas are normally dominant for right-handed people.

The BOLD signal change in motor areas is very sensitive to the proportion of scanning time consisting of active movement rather than rest. For example, slight increases of movement duration or number of movements per unit of time will lead to large increases in the magnitude of motor cortex activity. Subsets of motor areas demonstrate activity that correlates with kinematic measurements. Increasing movement velocity, force, or amplitude increases activity in

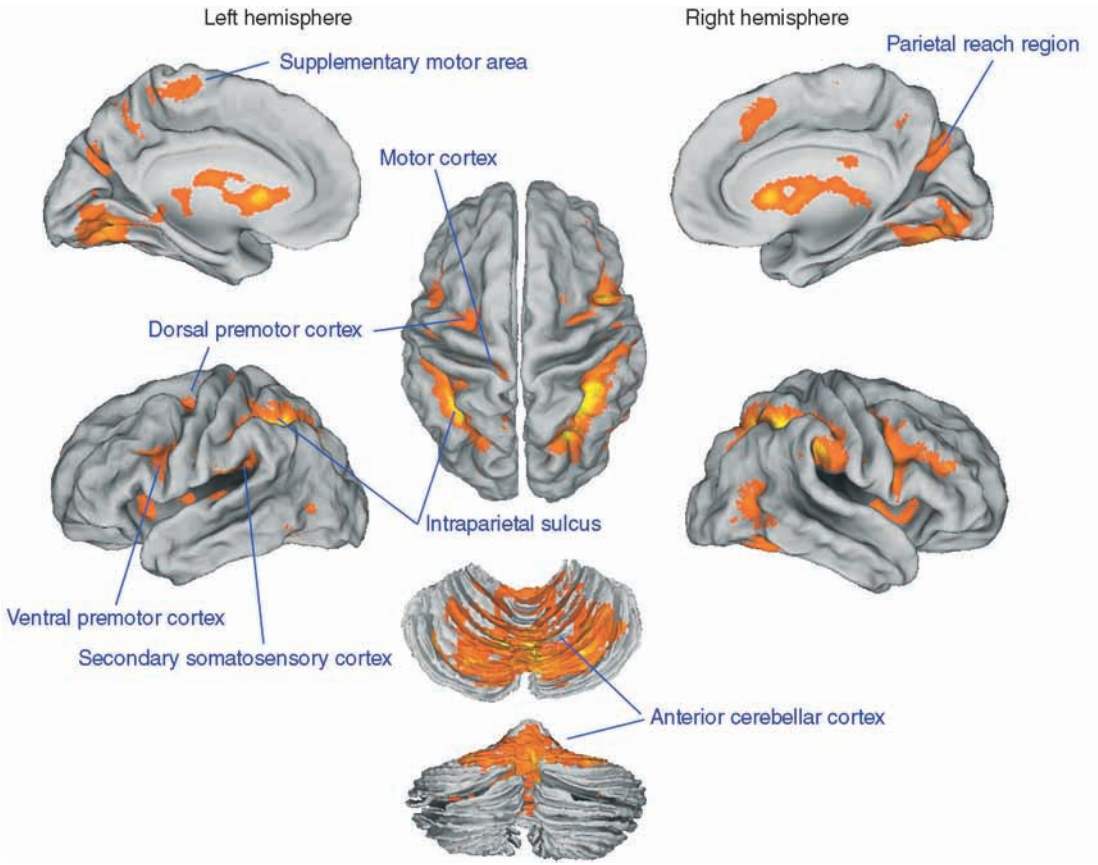

Figure 1 Cortical regions activated during visually guided tracking with the right hand compared with a rest condition are shown in orange. The results are a composite of 15 participants, superimposed on a population-based MRI atlas (PALS atlas, Caret software). Even the simplest of motor tasks engages a broad set of cortical areas to accomplish a motor goal.

contralateral motor cortex and GPi, as well as in ipsilateral cerebellar cortex. From this evidence, it has been proposed that basal ganglia activity may be critical for scaling motor output to a given context and to given task requirements.

Within the network of areas that are recruited during motor execution, there is specialization for the type of action that is being performed. For example, grasping objects relative to pointing recruits the most anterior portion of the intraparietal sulcus in an area considered a homolog of the nonhuman primate anterior intraparietal (AIP) area (Figure 4). Neurons in this area code visual features of an object and are also active during grasping without vision of the hand. In contrast, pointing studies are more likely to recruit the posterior superior parietal lobule and adjacent posterior intraparietal sulcus (IPS) and mesial parietal cortex. This cortex has been proposed as the human homolog of the nonhuman primate area called the parietal reach region. These parietal regions have strong interconnections with the dorsal premotor cortex. Together, they demonstrate that there are multiple specialized pathways between the posterior parietal cortex and premotor cortex for action execution.

Motor Preparation

Motor behavior requires the selection of appropriate motor commands to achieve a particular goal. Using paradigms in which a delay period occurs between task instruction and movement imperative or paradigms in which the instruction is followed by a go or no-go instruction, task planning can be distinguished from motor execution. Tasks requiring selection of a learned response typically recruit the posterior parietal cortex, particularly the superior parietal lobule and adjacent intraparietal sulcus, and the dorsal premotor cortex. This is true for choosing among many types of responses, including movements with different digits, pointing to different locations, or grasping with power or precision grip. These selection areas are particularly important when the relationship between the stimulus and movement is arbitrary,

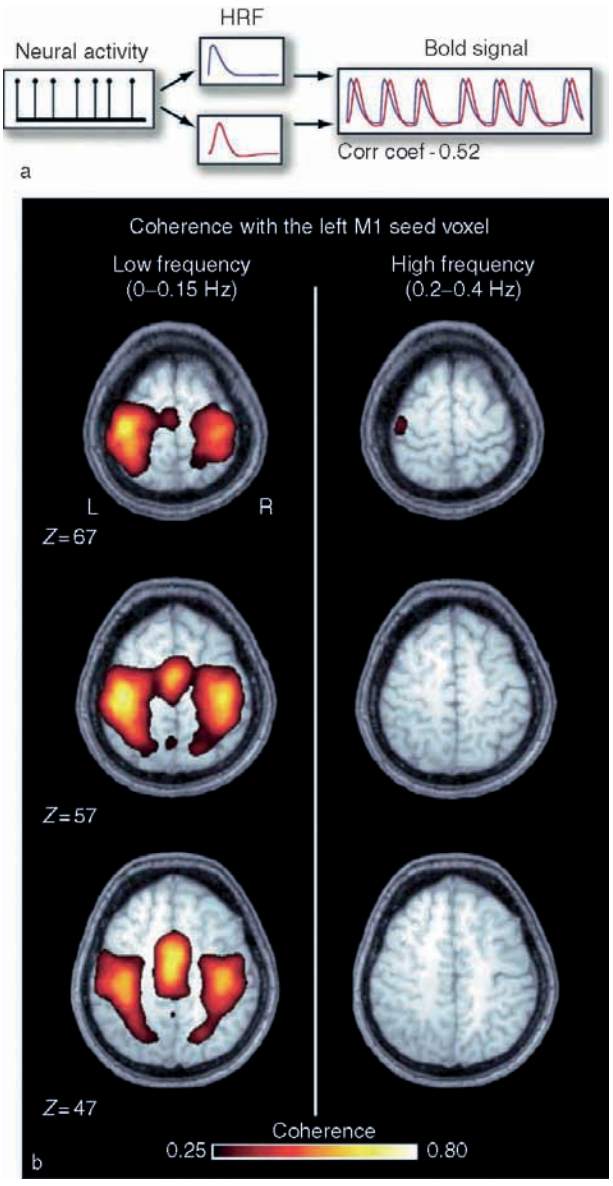

Figure 2 (a) Neural activation (left) leads to hemodynamic responses (HRF) in the brain (middle) that are delayed and dispersed in time. Coherence mapping is a method to determine whether two areas demonstrate common frequencies of responses irrespective of phasic differences (right). (b) Demonstration of brain areas demonstrating strong coherence with motor cortex at low (left column) and high (right column) frequencies in a movement task. Corr Coef, correlation coefficient; *Z*, distance in millimeters above the anterior commissure. Adapted from Sun FT, Miller LM, and D'Esposito M (2004) Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. *Neuroimage* 21: 647–658.

such as tasks where nonspatial cues are used to indicate where someone should point in space. This dorsal parietal-premotor circuit forms a substrate for learning many types of mappings between stimuli and responses, such as learning that a red light means stop and a green light, go. The left parietal cortex is also involved in implicit action selection. It has long been observed that covert motor attention can be directed to an upcoming movement. Neuroimaging studies show that the left IPL is recruited in experiments when participants covertly prepare movements.

Successful motor behavior requires the ability to switch between mappings as a function of context. Do we continue with the task at hand, such as reading and drinking coffee, or is it time to shift to an alternate behavior such as checking email? The ability to switch or stay with the current task, depending on the context, requires involvement of the medial frontal cortex including the presupplementary motor area (pre-SMA) and adjacent cingulate cortex. These areas are transiently activated at the time that participants switch from one way of responding during a task to another.

Sequential Movement

Complex behavior requires the organization of multiple movements across time. An extensive neuroimaging literature demonstrates that retrieval of already learned sequences is associated with activity in the supplementary motor area (SMA) and pre-SMA regions of the medial bank of the superior frontal gyrus (Figure 5). This is found with human and nonhuman primate lesion studies. Damage to this area of the mesial frontal cortex leads to impairment in organizing sequential movements. Single unit recording studies in nonhuman primate demonstrate individual neurons that represent sequential information in SMA and pre-SMA. There is emerging evidence for involvement of the inferior parietal lobule during sequence learning and retrieval as well. The sequence related activity in IPL overlaps with functional studies localizing brain areas associated with tool use and complex motor behaviors (see below), and is consistent with clinical observations that lesions to the inferior parietal lobule (IPL) in humans can lead to impairments in motor sequencing for complex tasks.

When motor sequences are learned under explicit conditions, such as learning to tap the digits of the right hand in a particular pattern, humans can also draw on working memory and higher-level cognitive circuits to enhance performance. This is reflected in recruitment of dorsal prefrontal cortex when explicit sequences are performed. It has been proposed that the basal ganglia represent sequential information as tasks become overlearned and automated. However, additional evidence is needed to establish a storage role for the basal ganglia or specific cortical regions for motor memories.

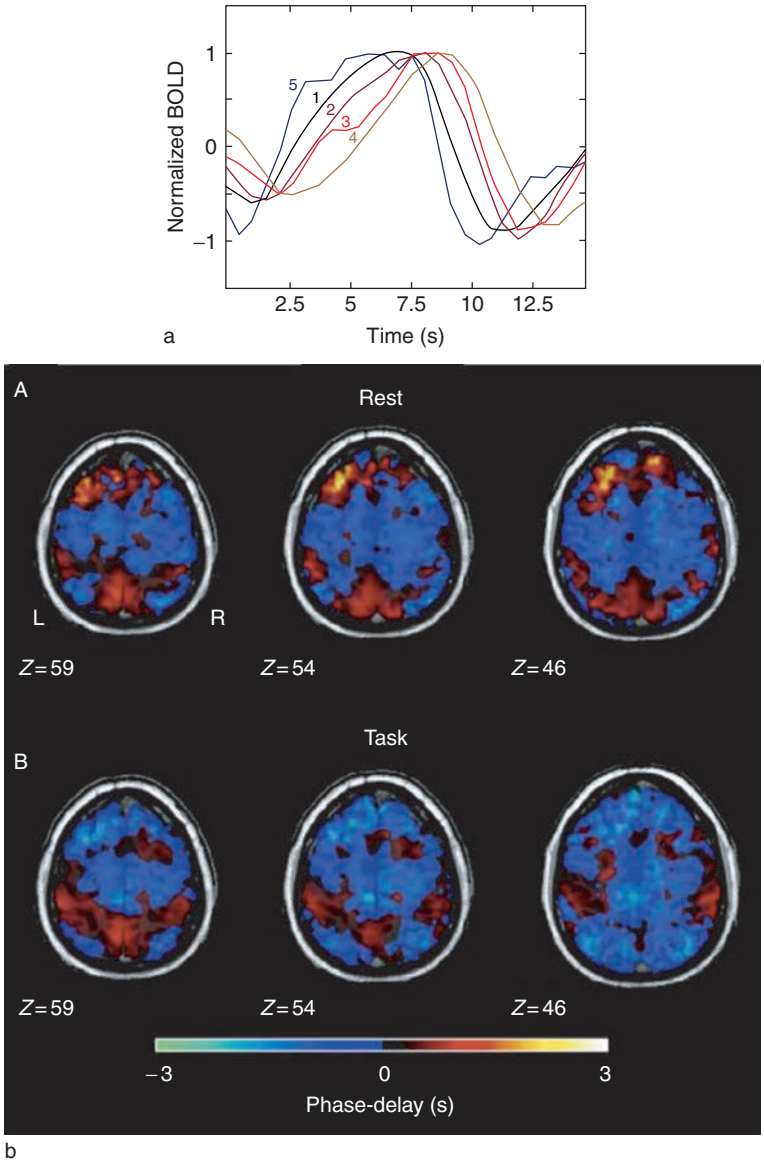

Figure 3 The hemodynamic response in the brain limits the time resolution for measuring when an event occurs. However, relative latencies between a set of areas can be distinguished at a subsecond resolution, as shown in (a). The plots show the relative timing of responses in supplementary motor area (1), left premotor cortex (2), left motor cortex (3), left parietal lobe (4), and right ventral premotor cortex (5). Relative differences of the timing when the BOLD response occurs during a motor task are shown in (b). Note the differential timing of response in motor cortex compared to other areas as a function of movement (lower row) or rest (upper row). *Z*, height in millimeters above the anterior commissure. Adapted from Sun FT, Miller LM, and D'Esposito M (2005) Measuring temporal dynamics of functional networks using phase spectrum of fMRI data. *Neuroimage* 28: 227–237.

Online Control

A precise movement, such as threading a needle, requires the ability to adjust an ongoing movement in the face of inaccurate preplanning, a changing environment, or an imperfect musculoskeletal apparatus. All these introduce added noise to any planned action. Studies of precision reaching demonstrate robust recruitment of the anterior cerebellar cortex, ipsilateral to the reaching arm during task execution. If external errors are introduced by shifting the target location near the time of movement onset or by introducing a perturbing force to the reaching arm, then increased activity is also seen along the postcentral sulcus, particularly near the anterior intraparietal sulcus contralateral to the reaching arm (Figure 6), suggesting that a common parietal–cerebellar circuit is involved in correcting random errors induced by a number of different causes.

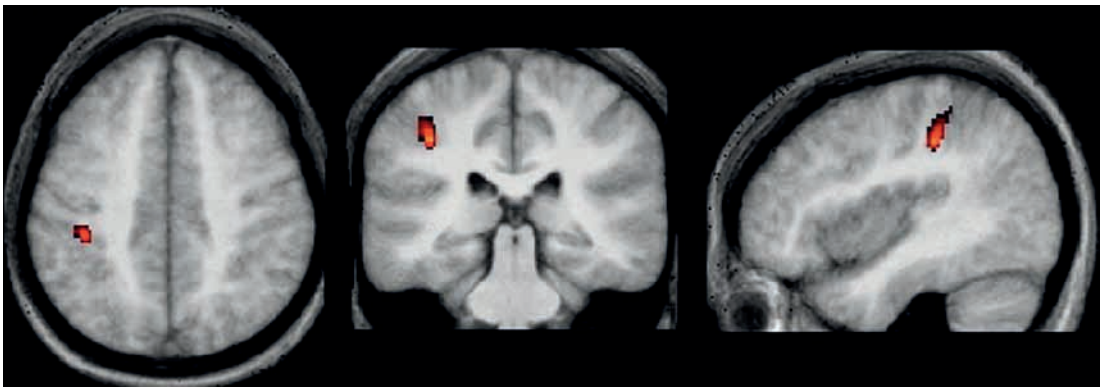

Figure 4 Localization of grasp-related activity in the anterior intraparietal sulcus. This area is considered to be a homolog of the anterior intraparietal (AIP) area in nonhuman primates. Adapted from Frey SH, Vinton D, Norlund R, and Grafton ST (2005) Cortical topography of human anterior intraparietal cortex active during visually-guided grasping. *Cognitive Brain Research* 23: 397–405.

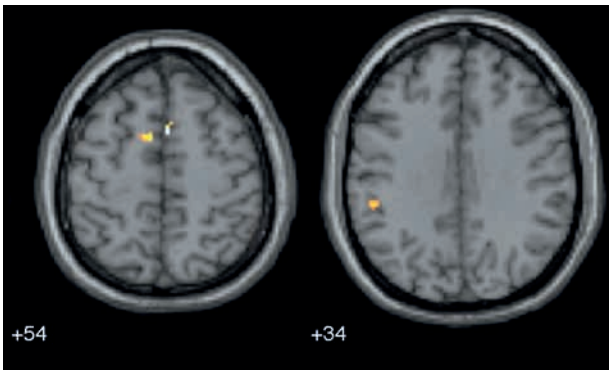

Figure 5 Recruitment of pre-SMA (left image) and inferior parietal lobule (right image) when participants retrieve a previously learned sequence of finger movements. +54, +34 indicate height in millimeters above the anterior commissure. Adapted from Bischoff-Grethe A, Goedert KM, Willingham DT, and Grafton ST (2004) Neural substrates of response-based sequence learning using fMRI. *Journal of Cognitive Neuroscience* 16: 127–138.

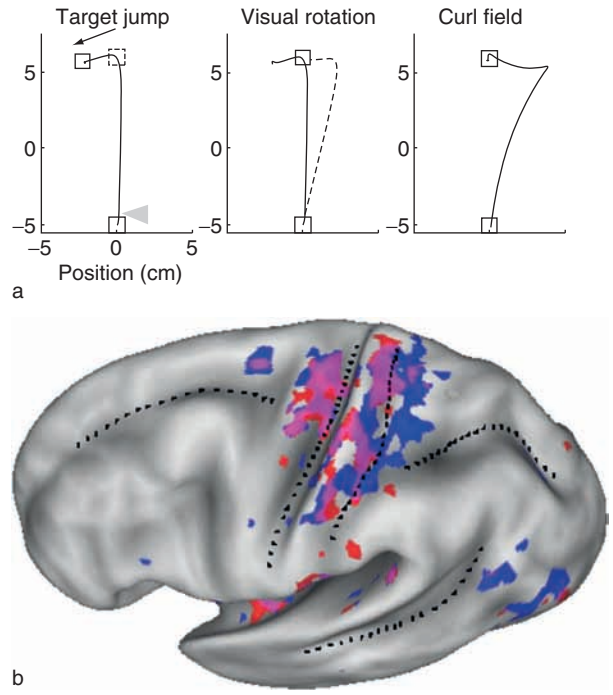

Figure 6 (a) Online control of movement can be assessed by randomly perturbing a target location at movement onset (left panel), by inducing a visuomotor rotation between hand and cursor direction (middle panel), or by making a joystick movement through a curl force field (right panel). In all three cases, the net trajectory is similar, but the source of error is very different. (b) Brain areas of the left hemisphere involved in visual-rotation (blue) and curl-field (red) conditions relative to normal movements for the left hemisphere. The hemisphere has been 'inflated' and is viewed from a superior oblique position. Purple areas indicate overlap. The dotted lines, starting left and in clockwise order, denote the following sulci: superior frontal, central, postcentral, intraparietal, and superior temporal. Adapted from Diedrichsen J, Hashambhoy Y, Rane T, and Shadmehr R (2005) Neural correlates of reach errors. *Journal of Neuroscience* 25: 9919–9931.

Organization of Complex Behaviors

Humans maintain a broad repertoire of physical skills for solving real-world problems, combined with a seamless ability to work with tools. It has long been known from lesion studies that the explicit retrieval of action knowledge depends critically on the integrity of the left parietal lobule and posterior middle frontal gyrus. Lesions to either of these regions lead to ideomotor apraxia. This localization has been closely replicated in functional neuroimaging studies of participants as they mentally retrieve knowledge about how to use tools. For example, in one study, participants were asked how to use a hammer and other tools. By the use of a go, no-go paradigm, the mental retrieval process was separated from the actual motor execution used to pantomime the action needed for each of the tools (Figure 7). Functional neuroimaging established that during action retrieval, the left supramarginal gyrus was the most consistently recruited sector of the IPL across individual participants, irrespective of whether the action would be pantomimed by the right or left hand.

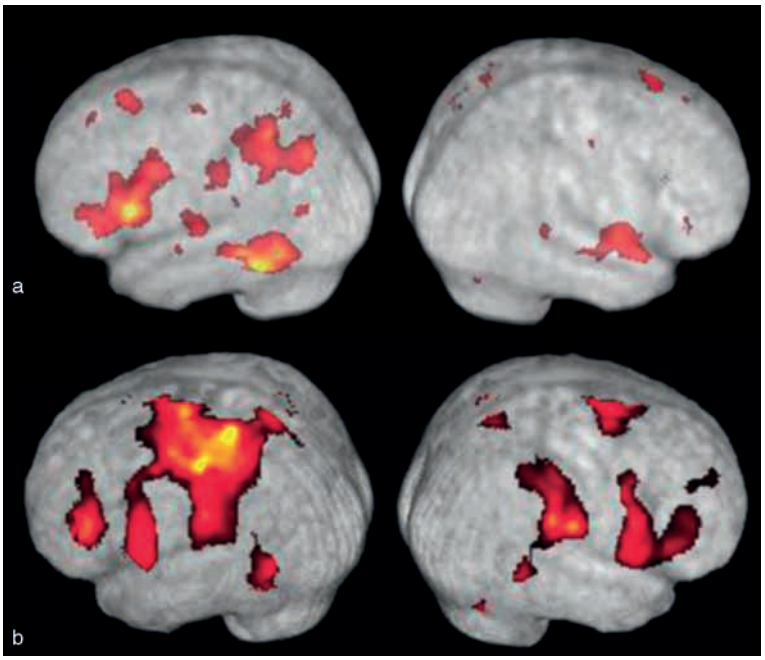

Figure 7 Cortical areas involved in retrieving (a) and executing (b) pantomimed gestures of common tool use by the right hand. Left and right columns are the left and right hemispheres, respectively. Retrieval of tool-based actions requires the left supramarginal and inferior frontal gyri. Similar retrieval patterns are found irrespective of the hand used to make the gesture. Adapted from Johnson-Frey SH, Newman-Norlund R, and Grafton ST (2005) A distributed left hemisphere network active during planning of everyday tool use skills. *Cerebral Cortex* 15: 681–695.

The neural substrates for action representation can also be identified in studies of action observation. In nonhuman primates there are strong interconnections of AIP, the grasp-related area, with the inferior frontal cortex. These sites contain mirror neurons, individual cells that fire in response to observation of grasping actions by another agent. These neurons are also active during actual grasping. A homologous set of regions, including the ventral premotor cortex and adjacent inferior frontal gyrus (pars opercularis) as well as inferior parietal lobule in the human are activated with action observation as well as execution and imitation, as shown in Figure 8. This network in humans has been labeled the mirror neuron system (MNS) and may be a critical substrate for action organization as well as in understanding actions of others. Activity within some parts of the MNS is experience-dependent. For example, studies have shown that the more physical expertise dancers have in performing a particular movement, the greater the activity in the MNS as they observe others performing the same action.

Complex actions are organized with respect to desired outcomes. From this, intermediate solutions are used to generate actions that lead to those outcomes. For example, on the elevator someone can push the button for the desired floor with the hand, unless it is occupied with another task such as holding an umbrella. In this case one of several alternate solutions is implemented, such as pushing the button with the umbrella, the other hand, or even the foot. There is emerging evidence from studies of action observation that dissociable neural substrates are recruited for these different levels of action organization. When an action is observed, such as someone picking up a cup, the intention of the action will differ if the surrounding dishes are dirty (clearing the table) or filled with food (taking a drink). Activity within the inferior frontal gyrus is modulated when this level of intention is inferred. In contrast, for intermediate levels of action organization, such as the detection of what particular object is being grasped rather than how it is grasped, the anterior IPS demonstrates sensitivity.

Motor Skill Learning

Motor learning is the ability to perform a novel movement to accomplish a behavioral goal, such as playing a musical instrument. In skill acquisition of this type, someone may take years to reach proficiency, and it is difficult to use brain imaging to study the accompanying long-term changes within an individual's nervous system. Instead, neuroimaging studies have focused on other aspects of learning that occur along

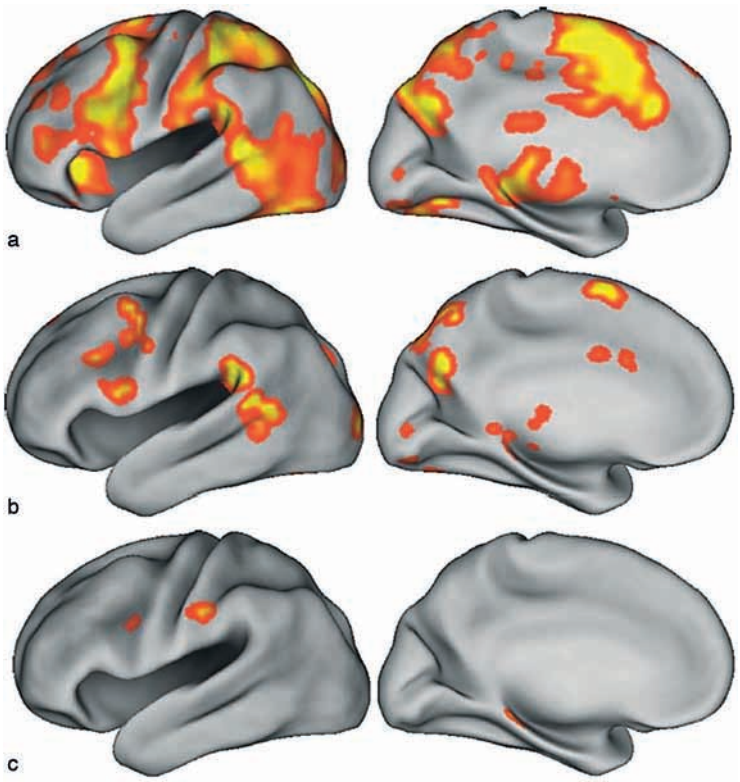

Figure 8 Example of the interaction between brain areas that respond to the observation of action and the role of physical experience for those same actions. (a) Participants watched video clips of a teacher performing complex modern dance movements. Note the widespread recruitment of premotor, temporal, and parietal cortex of the lateral left hemisphere (left column), as well as supplementary motor area on the medial surface (right column). (b) A comparison of trials after observers have learned to perform complex dance movements with those in which they have not identifies a more specific set of brain regions in the action observation network. (c) If a measurement of self rated competency is correlated with brain activity acquired while subjects observe of the same movement, there is a strong correlation in inferior parietal and frontal gyri. This demonstrates that some, but not all action observation areas have experience-dependent responses. Adapted from Cross ES, Hamilton AF, and Grafton ST (2006) Building a motor simulation *de novo*: Observation of dance by dancers. *Neuroimage* 31: 1257–1267.

faster timescales. For example, adaptation paradigms can be used to examine learning of task dynamics, analogous to playing a musical instrument with a heavy coat on. In this case the motor skill is already learned, and the performer is adjusting motor output to address a novel environment. In this case, adaptation can occur quickly. Adaptation-related activity is associated with recruitment of parietal and cerebellar cortex, suggesting a close connection between online error correction and adaptive control. Another strategy for examining learning is to examine repeated performance of relatively simple tasks, such as visuomotor tracking. In these cases, activity increases rapidly and dramatically in sensory and motor cortices contralateral to the moving hand. These early increases might be related to temporary changes of cortical excitability that occur with repeated use and reflect reduced cortical inhibition rather than formation of a specific skill. With longer-term practice over hours to days, activity changes have been observed in numerous motor planning areas, including the supplementary motor area, dorsal premotor cortex, and inferior parietal lobule. There is also a close correspondence between activity in the cerebellum and coordination between the eyes and the hand.

Comparisons of expert and novice performers provide an important avenue for defining learning-dependent changes in the brain that emerge over years of practice. Musicians such as professional violinists and pianists demonstrate greater gray matter volume in the motor cortex than their novice counterparts. In contrast, functional imaging experiments of these people performing relevant skills such as keyboard performance show that the experts recruit far less cortex than the novices do in areas including motor cortex, SMA, and dorsal premotor cortex. This is true even when the performance of the two groups is matched in speed and difficulty. This finding supports a model whereby long-term training in

complex skills leads to a strengthening within a subset of neural circuits that can be recruited for a given task, reflected in reduced recruitment of cortical area.

See also: Basal Ganglia: Motor Functions; Computational Approaches to Motor Control; fMRI: BOLD Contrast; Motor Skill Learning; Motor Sequences; Motor Timing; Neuroimaging; Premotor Cortex in Primates: Dorsal and Ventral; Premotor Areas: Medial.

Further Reading

Bischoff-Grethe A, Goedert KM, Willingham DT, and Grafton ST (2004) Neural substrates of response-based sequence learning using fMRI. *Journal of Cognitive Neuroscience* 16: 127–138.

Cross ES, Hamilton AF, and Grafton ST (2006) Building a motor simulation *de novo*: Observation of dance by dancers. *Neuroimage* 31: 1257–1267.

Culham JC and Valyear KF (2006) Human parietal cortex in action. *Current Opinion in Neurobiology* 16: 205–212.

Diedrichsen J, Hashambhoy Y, Rane T, and Shadmehr R (2005) Neural correlates of reach errors. *Journal of Neuroscience* 25: 9919–9931.

Floyer-Lea A and Matthews PM (2005) Distinguishable brain activation networks for short- and long-term motor skill learning. *Journal of Neurophysiology* 94: 512–518.

Frey SH, Vinton D, Norlund R, and Grafton ST (2005) Cortical topography of human anterior intraparietal cortex active during visually-guided grasping. *Cognitive Brain Research* 23: 397–405.

Hamilton AF and Grafton ST (2006) Goal representation in human anterior intraparietal sulcus. *Journal of Neuroscience* 26: 1133–1137.

Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, Mazziotta JC, and Rizzolatti G (2005) Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology* 3: e79.

Johnson-Frey SH, Newman-Norlund R, and Grafton ST (2005) A distributed left hemisphere network active during planning of everyday tool use skills. *Cerebral Cortex* 15: 681–695.

Logothetis NK (2002) The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1003–1037.

Miall RC and Jenkinson EW (2005) Functional imaging of changes in cerebellar activity related to learning during a novel eye-hand tracking task. *Experimental Brain Research* 166(2): 170–183.

Rizzolatti G and Craighero L (2004) The mirror-neuron system. *Annual Review of Neuroscience* 27: 169–192.

Sanes JN and Schieber MH (2001) Orderly somatotopy in primary motor cortex: Does it exist? *Neuroimage* 13: 968–974.

Sun FT, Miller LM, and D'Esposito M (2004) Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. *Neuroimage* 21: 647–658.

Sun FT, Miller LM, and D'Esposito M (2005) Measuring temporal dynamics of functional networks using phase spectrum of fMRI data. *Neuroimage* 28: 227–237.

Relevant Website

http:/brainmap.wustl.edu – Van Essen Lab; See CARET and Suyace-Based Atlases for PALS Atlas, Caret Software.

Functional Neuroimaging Studies of Aging

C L Grady, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Age differences in cognitive function have been studied for many years, and of all the cognitive domains, memory has probably been studied the most extensively. Compared to young adults, older adults have particular difficulty with episodic memory, defined as the conscious recollection of events that have occurred in a person's experience. In the laboratory, these age differences in episodic memory are seen in a reduced ability to learn and retrieve lists of stimuli, such as words or pictures of common objects. Older adults have particular difficulty in retrieving accompanying information about stimuli learned during the course of an experiment, so-called source memory. For example, when learning a list of words presented verbally by either a male or a female speaker, young adults are readily able to recall both the words and whether they were spoken by a man or woman, whereas older adults will have more difficulty remembering the speaker's gender than the words. Older adults also recall fewer details about real-life, autobiographical memories, and substantial age-related declines are seen on working memory tasks (i.e., tasks that require holding information on line for brief periods of time). In addition to these differences in memory, older adults also show reductions in inhibitory function, reflected in an increased vulnerability to distracting information, as well as lower performance on tests tapping other types of executive function. In contrast, semantic memory, or the accumulation of knowledge about the world, is maintained or even increased in older adults, compared to younger adults. Some aspects of emotional function also are preserved with age, although identifying and remembering negative emotional material are sometimes reduced in older adults. Thus, not all aspects of cognitive function are affected adversely by older age.

In recent years, functional neuroimaging has been used to explore brain activity accompanying cognitive tasks in young and old adults, and to understand the neural mechanisms underlying behavioral differences. Memory has received the most attention in this work, similar to the predominance of memory studies in the behavioral literature. Nevertheless, other cognitive domains have been studied, and some of the fundamental processes contributing to cognitive aging are becoming clear.

Methodological Issues in Studying Aging

Prior to reviewing the results of neuroimaging studies of cognitive aging, it is useful to consider some of the issues that one may encounter when using this methodology to study older adults. The typical techniques used are either positron emission tomography (PET) or, more commonly today, functional magnetic resonance imaging (fMRI). Both of these techniques are based on task-dependent modulation of hemodynamic function, such as blood flow or oxygenation. Therefore, age-related changes in the vasculature might influence the coupling between neural activity and hemodynamic functional measures obtained in the elderly. Although the shape of the hemodynamic response function, as measured with fMRI, is similar in young and old adults, there is reduced signal-to-noise in the elderly. In addition, older adults undergo brain shrinkage, or atrophy, and some loss of white-matter integrity, both of which could impact functional measures independently of any changes in neural activity. Despite these difficulties, several robust and replicable age-related differences in brain activity have been noted, suggesting that these potential artifacts may not pose too severe a problem.

When brain activity in young and older adults is compared on a task, there are at least three possible outcomes in any given brain area: (1) young and old groups could have equivalent brain activity, (2) older adults could show less activity, or (3) old adults could show greater activity. Reduced activity in the elderly can reasonably be assumed to reflect a reduced level of functioning, particularly when accompanied by poorer performance on the task. Equivalent activity is generally considered evidence for spared function in the elderly. The major challenge has been to interpret increased recruitment of task-related brain areas, or activation of unique brain regions in the elderly.

Several analytic approaches have been applied to the study of brain mechanisms underlying cognitive aging. The most common approach has been a straightforward contrast of brain activity between young and old groups of participants and interpretation of the observed differences in activity in the context of behavioral performance on the tasks. Much of this work has focused on the frontal lobes, which are thought to be particularly vulnerable to the effects of aging. However, this type of univariate approach is often plagued by issues related to multiple comparisons and statistical power, which are compounded when more than one group of individuals is studied. Other approaches that utilize multivariate analyses to identify whole-brain patterns of

functional changes, and hence avoid problems with multiple comparisons, have been used with some success in the aging field. Functional connectivity is another type of multivariate approach that focuses on cognitive networks (i.e., groups of regions working in an integrated fashion during cognitive tasks). Connectivity involves assessing how activity in a given region correlates with activity in other areas of the brain during a task. This type of correlation can be done on a pair-wise region basis or across the whole brain. Thus, connectivity approaches emphasize the functional interactions among brain areas and the ways in which these interactions mediate cognitive processing, rather than focusing on the activity in any individual brain region. This 'whole-brain' type of approach has revealed that a number of brain areas, including the frontal lobes, appear to be critical sites of age-related change in function.

Reduced Brain Activity in Older Adults

Reduced brain activity in older adults typically has been noted during encoding, or learning new material. Younger adults have increased activity in left prefrontal cortex when learning new stimuli, and a common finding is that older adults show less activity in this area during encoding compared to younger adults, across a variety of stimulus types (e.g., words, faces, objects; see Figure 1). The age-related reduction of activity in left prefrontal cortex can be ameliorated to some extent by using encoding tasks that promote deep, or semantic, processing of the to-be-remembered information. Both young and old adults show more left prefrontal activity during a semantic encoding task, compared to a task that does not require a focus on the semantic aspects of the stimulus, and this increase in older adults can be

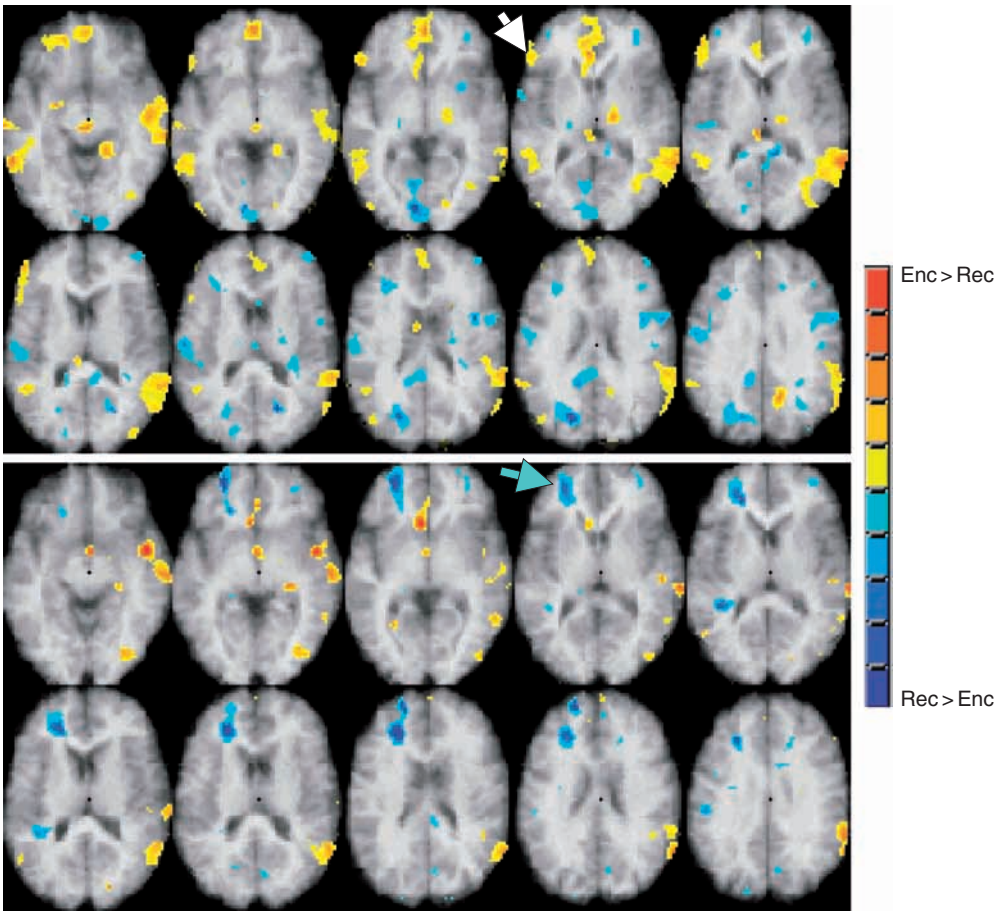

Figure 1 The images depict brain areas with changes in blood flow (measured with positron emission tomography) during encoding (Enc) and recognition (Rec) of faces in a group of young adults (top panel, $n = 12$; mean age 23 years) and a group of older adults (bottom panel, $n = 11$; mean age 70 years). Results are shown on images from a standard structural magnetic resonance image. Yellow/red areas are those with more activity during encoding, compared to recognition, and blue areas indicate more activity during recognition. Young adults show increased activity during encoding in left inferior prefrontal cortex (white arrow), but older adults do not. In contrast, old adults have more activity during recognition, compared to young adults, in widespread areas of left prefrontal cortex (cyan arrow).

large enough to remove any age difference. This pattern of results is consistent with behavioral data suggesting that older adults may not spontaneously engage effective encoding strategies – that is, those involving semantic processing by left inferior frontal cortex – but can use such strategies when required to do so.

In contrast to the left frontal activity seen during encoding, during memory retrieval, when people are attempting to recall or recognize previously learned information, frontal activity often is found predominately in the right hemisphere. Although this right frontal activity is often of the same magnitude regardless of age, older adults do show reduced activity under some conditions. When the retrieval task places a heavy demand on memory search, without providing cues for retrieval, older adults show a reduction in right prefrontal activation. Retrieval of source memory, such as whether a word was spoken by a man or a woman, also places significant demand on memory search and typically activates prefrontal regions in young adults. Consistent with their reduced source memory, older adults have less activation of prefrontal cortex during source retrieval compared to their younger counterparts.

Another area of memory that has been examined fairly extensively in the aging field is that of working memory. A number of working memory tasks have been used, including delayed match-to-sample and *n*-back tasks. Regardless of task, these studies have generally found less prefrontal activity in older adults. When the different phases of the working memory task have been examined separately (memory set presentation, delay, and probe), age differences have been seen primarily during the presentation of the probe – that is, when the memory decision is being made. This suggests that some aspects of working memory, such as rehearsal, might not be affected by age. Some recent work has shown that older adults are able to recruit prefrontal cortex at lower levels of working memory task load, but that this declines at higher task loads, whereas younger adults continue to increase prefrontal recruitment with increasing load. This finding is similar to that noted earlier for episodic retrieval, as both would suggest that tasks with high processing demands are likely to be associated with reduced brain activity in older adults, and potentially poorer performance as well.

In addition to prefrontal regions, less activity during encoding and/or recognition memory in older adults also has been found in other areas important for memory, including the hippocampus and the parahippocampal gyrus, and in visual areas of cortex (Figure 2). Interestingly, when brain activity has been compared across different kinds of tasks, such as working memory and episodic retrieval, similar age reductions in brain activity have been found in visual cortex and hippocampus. Such a result is consistent with the idea that there is a 'common cause' of cognitive aging, and indicates that reductions in both

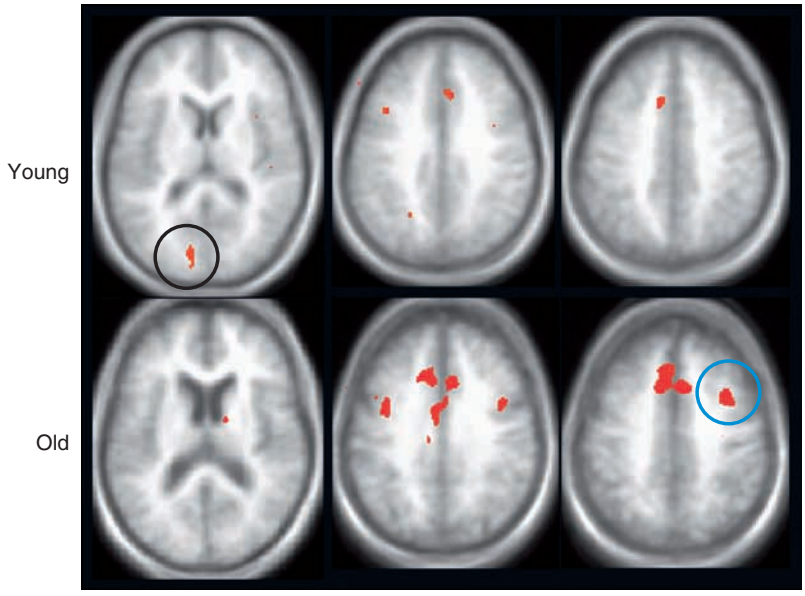

Figure 2 The images depict brain areas with changes in blood oxygenation (measured with functional magnetic resonance imaging) during recognition of auditorily presented words in a group of young adults ($n = 12$; mean age 26 years) and a group of older adults ($n = 12$; mean age 71 years). Results are shown in red on mean structural magnetic resonance images for the young adults (top) and the older adults (bottom). Young adults had greater activity in a region of left extrastriate cortex (black circle), whereas older adults had more activity in prefrontal regions (one of which is indicated by the blue circle).

sensory processing and memory encoding could be involved in such a common cause. One thing to keep in mind, however, particularly when considering age differences in sensory cortex, is that less activity during a particular cognitive task may mean that older adults have more activity in these areas during the baseline condition. Thus, when activity during the baseline is subtracted from task activity, older adults will have a smaller task-related increase, but this does not mean that the region cannot be activated. Rather, the older adults are starting out from a higher level of activity. Such a phenomenon has been observed in visual cortex.

Functional neuroimaging has also been used to study executive functions, thought to be mediated primarily by frontal cortex, in older adults. One such function, inhibition, has been studied using tasks that require inhibiting a prepotent response or suppressing information irrelevant to the task. Consistent with their poorer performance, older adults have less activity in frontal regions thought to mediate inhibitory processes on such tasks. Activity in frontal regions also is reduced with age during performance of the Wisconsin Card Sorting Task, a standard instrument measuring concept formation and cognitive flexibility. Other brain regions also show age differences during tasks requiring inhibitory processing. Young adults either enhance or suppress activity in regions of occipital cortex that respond selectively to different types of objects depending on whether the objects are attended or ignored. In contrast, older adults show enhanced activity in these areas, but fail to reduce activity in object-specific regions when viewing objects that should be ignored because they are irrelevant to the task at hand. Thus, age reductions in the ability to ignore irrelevant and distracting information may be the result of both top-down (i.e., frontal) and bottom-up (occipital) mechanisms.

Emotional processing and emotional enhancement of memory have been studied for a number of years using animal models, but only recently has investigation of this area of cognition been explored with functional neuroimaging. Consistent with animal work showing critical involvement of the amygdala in emotion, young adults show increased activity in the amygdala when emotional stimuli, particularly negative items, are viewed. This activity also is related to later memory for this material, even when memory is tested months after initial exposure to the stimuli. Older adults show less amygdala activity compared to younger adults during viewing of negative faces and scenes, consistent with their reduced ability to both identify and remember negatively valenced stimuli.

Finally, there has been recent interest in a group of brain areas that have more activity when people are not carrying out a specific task, but are merely monitoring their internal and external milieus, the so-called 'default mode' regions. Young adults show marked reductions of activity in these default areas when carrying out a variety of tasks, whereas older adults have much less change in these areas during tasks. In fact, there appears to be linear reduction in this default region 'suppression' over the adult life span, suggesting a gradual, age-related reduction in the ability to suspend non-task-related, or default, activity and to engage areas for carrying out memory tasks. Thus, it may be just as important to look at age-related differences in activity reductions as it is to explore differences in activity increases.

Increased Brain Activity in Older Adults

It is perhaps surprising that, in addition to reduced brain activity in some areas, older adults often show increased brain activity, compared to young adults. Additional recruitment of brain areas by older adults typically is found in the frontal lobes, either in areas of frontal cortex not active in young adults, or in homologous regions in the opposite hemisphere from regions active in the young. During episodic memory retrieval older adults frequently show more extensive activation of prefrontal regions compared to younger adults, sometimes in those left hemisphere frontal areas that are less active during encoding in older adults. Greater prefrontal activation during memory retrieval has been found in older adults for a variety of stimuli, both verbal and nonverbal (**Figures 1** and **2**). Another way of looking at the brain activity that supports retrieval is to see how activity during encoding of items that are later correctly remembered differs from that for encoding of items that are forgotten, known as the 'subsequent memory effect.' That is, what is the brain activity, during learning, that predicts whether a stimulus will be successfully remembered later? Using the subsequent memory effect, a number of investigators have found that activity in left inferior prefrontal cortex and the left medial temporal region in young adults is greater during encoding of subsequently remembered stimuli. A similar pattern has been found in old adults; however, older adults also have an association between more activity in right prefrontal cortex and subsequent recognition. Thus, older adults can show a more extensive pattern of frontal activity during encoding, as they do during retrieval, but only when that encoding is sufficient to support later retrieval. A similar pattern has been found during some working memory tasks, where frontal activation often is unilateral in young adults, but in both hemispheres in older adults.

There is some evidence for greater brain activity in older adults during memory retrieval tasks in areas outside of the frontal lobes, such as medial temporal cortex. In particular, studies of recognition memory, such as recognition for words presented earlier in the experiment, have shown more activation of parahippocampal regions in older adults, compared to younger adults. Indeed, parahippocampal cortex in young adults often shows less activity when previously seen stimuli are presented during a recognition task, and this reduction in activity is thought to represent a signal of familiarity with the stimulus. Increased activity in older adults in the parahippocampal gyrus may still indicate familiarity, but the nature of the signal may undergo a change with age. Interestingly, although hippocampal activity during memory retrieval is often reduced with aging, bilateral activation of the hippocampus has been reported in older adults during an autobiographical memory retrieval task, whereas young adults had only left hippocampal activation. This bilateral activation is similar to the bilateral prefrontal activity seen in older adults during other memory tasks, and may support those aspects of autobiographical retrieval which are maintained in the elderly.

Increased prefrontal activation in older adults also has been found in other cognitive domains, although less information is available than for memory. For example, there are reports of more prefrontal activity in older adults when carrying out what one might consider relatively simple motor tasks. This may indicate that even sensorimotor tasks require more cognitive control on the part of older adults.

So, what could be the impact of this increased brain activity on behavioral performance in older adults? There have been suggestions that additional recruitment of task-related brain areas, or engagement of unique areas not active in young adults, compensates for age-related changes in brain structure and function. Compensation has been invoked to explain greater prefrontal activity in older adults when they are able to perform a particular task, as well as younger adults. Additional support for this idea has been found in studies showing that older adults who perform better on memory tasks recruit prefrontal cortex to a greater degree than do those performing less well. Older adults also may have more prefrontal activity than do young adults when brain activity is examined just for those task trials where participants respond correctly. Such findings suggest that correct performance may be supported by this overrecruitment of frontal regions. Another approach has been to determine how brain activity is related to task performance by correlating these two measures in groups of young and old adults. Increased prefrontal activity is often correlated with better performance on memory tasks in older adults, but rarely in young adults. In fact, in young adults prefrontal activity is usually correlated with poorer task performance, indicating that with age, the impact of brain activity on memory performance is altered.

On the other hand, some of this increased prefrontal activity in the elderly may reflect greater need or use of frontally mediated executive functions at lower levels of task demand than would be necessary for activation of this area in young adults. In this case the increased activity in prefrontal cortex might not be related to performance on the particular task, but would reflect a type of nonselective recruitment in older adults, perhaps related to task difficulty. Examples of this kind of scenario have been observed, and include increased activity in widely distributed areas of both lateral and medial portions of prefrontal cortex in older adults during tasks on which they perform worse than young adults (Figure 2). These findings suggest that while additional recruitment of prefrontal cortex in older adults may often be compensatory, it is not necessarily so.

Different Cognitive Networks in Older Adults

A relatively new approach to studying brain function is the focus on functional connectivity, or how activity in a given brain area is correlated with activity elsewhere in the brain, to identify networks of interacting regions. Studies using this approach have provided intriguing evidence that young and older adults recruit different networks, even when the measured behavior is the same. For example, when pictures are used as stimuli in memory tasks, they are recognized well by both young and old adults. Encoding of these pictures also is associated with increased activity in medial temporal regions, including the hippocampus, in both young and old groups. However, the areas that are functionally connected to the hippocampus during encoding vary with age. In young adults, during picture encoding, hippocampal activity is correlated with activity in ventral prefrontal and visual regions, whereas older adults show correlations between hippocampal activity and dorsolateral prefrontal and parietal regions (Figure 3). Of particular interest is that these age-specific networks are associated with better memory performance. That is, those individuals in each age group who engage the specific network to a greater degree are the ones who remember the greatest number of pictures.

Different functional connectivity of medial temporal regions in older adults, compared to young adults, also has been found during recognition tasks.

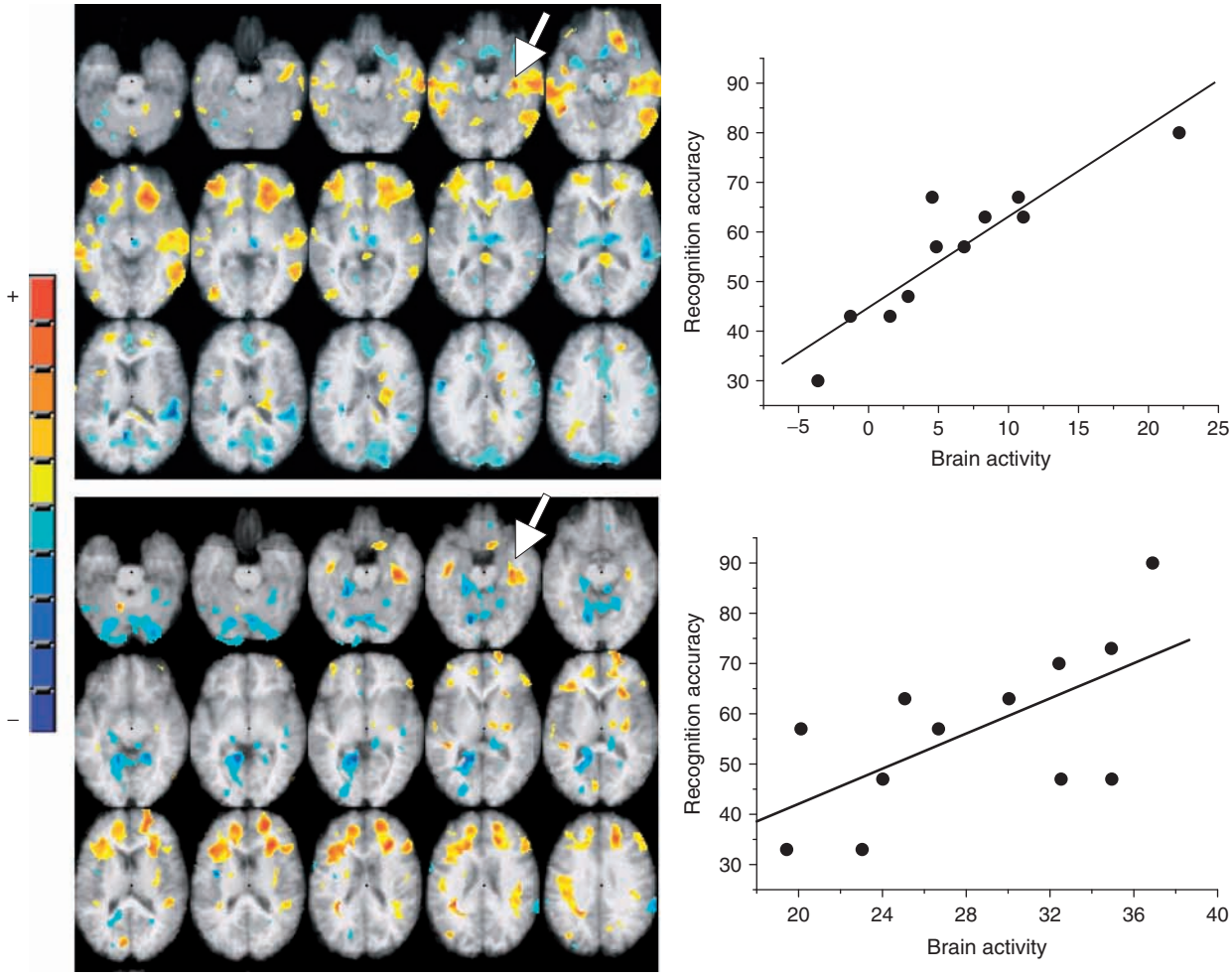

Figure 3 The images depict brain areas where blood flow (measured with positron emission tomography) during picture encoding was significantly correlated with flow in right hippocampus (white arrows). Results are from a group of young adults (top, $n = 12$; mean age 23 years) and a group of older adults (bottom, $n = 12$; mean age 66 years) and are shown on a standard structural magnetic resonance image. Yellow/red areas are those with positive correlations and blue areas are those with negative correlations. In young adults hippocampal activity was positively correlated with activity in ventral prefrontal regions, whereas in old adults hippocampal activity was positively correlated with activity in dorsolateral frontal cortex. The graphs show the scatterplots of activity in these age-specific networks, and recognition accuracy. As activity increased during encoding in the hippocampus, and the areas positively correlated with it, subsequent recognition accuracy increased in both groups.

A recent study found that during recognition, the hippocampus showed stronger functional connections with cortical regions in young adults, whereas the entorhinal cortex showed stronger connections in older adults. This finding was consistent with evidence that these two medial temporal areas mediate recollection and familiarity, respectively, two aspects of memory that tap different qualitative experiences, and with the many studies showing age differences in recollection but not familiarity. Thus, age differences in recollective processes mediated by the hippocampus may be offset by more reliance on familiarity mediated by medial temporal cortex adjacent to the hippocampus. Similar evidence for age differences in cognitive networks involving the hippocampus and its interactions with the cortex has been found in experiments tapping working memory as well. All of these studies that focus on network approaches to image analysis have provided evidence that aging results in the modification of large-scale network operations, in addition to affecting activity in specific brain regions, which in turn has an impact on memory ability.

Genetic Factors in Brain Aging

Recently there has been interest in using functional neuroimaging to study the genetic basis for cognitive behaviors, including memory. Most of this work has been done in young adults, by dividing healthy

samples of people according to different genetic alleles that bear some relation to risk for psychiatric diseases, such as schizophrenia. This work has shown that activity in prefrontal cortex and medial temporal areas differs across genotypes, and that these differences are associated with differences in personality traits, such as risk taking, and in memory function. Little work in this area has been applied to the study of cognitive aging, but a few studies have looked at the effect of *APOE* genotype on brain function and memory in older adults. *APOE* is one of the genes that confers risk for Alzheimer's disease when the *APOE* $\varepsilon 4$ allele is present. Several studies have shown that middle-aged and older individuals who have the $\varepsilon 4$ allele, but are nevertheless asymptomatic, have reduced resting metabolic activity in parieto-temporal cortex compared to those without the $\varepsilon 4$ allele. Follow-up assessments showed that cognitive decline after 2 years was greatest in those *APOE* $\varepsilon 4$ carriers with the lowest metabolism in parietal and temporal regions at baseline. In addition, older *APOE* $\varepsilon 4$ carriers have faster cognitive decline than do noncarriers, and show a relation between subjective memory complaints and decline in temporal lobe metabolism over time. Interestingly, it was found with fMRI that activation of prefrontal and medial temporal areas was increased in *APOE* $\varepsilon 4$ carriers, compared to noncarriers, but only during memory tasks. This finding suggests that the cognitive and neural effort necessary for memory function varies according to one's genetic makeup, and that this effort may also reflect the likelihood of cognitive decline in later years. All of these studies indicate that cognitive changes noted in older adults, even in those individuals who appear to be cognitively normal for their age, may be influenced by genetic factors. In addition, it is possible that genetic differences underlie some of the variability that has been found across neuroimaging studies of older adults.

Conclusions

It is clear that functional neuroimaging has opened a window into cognitive aging, shedding light on the roles of brain areas, such as prefrontal cortex and the hippocampus, in age-related differences in cognitive performance. However, a number of questions remain. For example, the extent to which alterations of brain activity or brain networks seen in older adults are due to changes in the function of the brain regions, or to changes in the white-matter tracts that connect the regions, is unknown. Communication between cortical regions could be reduced or altered because the neurons in the regions have altered function or because the fibers that connect them are damaged or less efficient. Changes with age in white matter, and the impact of these on cognition, have been documented, but how these changes affect functional connectivity of gray-matter regions during cognitive tasks is not well understood. Recent evidence indicates that age differences in white matter are more prominent in frontal than in posterior brain areas, and that white-matter integrity of the frontal lobes is related to activation in prefrontal gray-matter regions in older adults. These structural changes are consistent with age changes in frontal activity that suggest nonselectivity, but not with evidence that frontal cortex is most often recruited for compensatory purposes. Clearly we have much to learn about the complex interplay between brain structure and function, and how this interaction is affected by age.

There are other factors that could influence brain activity in older and younger adults and these are just now beginning to receive attention. An example is education, which has been shown to have a protective effect against decline in cognitive function in older adults. There is some evidence that education, or other indices of intellectual achievement, are related differently to brain activity in young versus older adults, such that education is positively correlated with frontal activity only in older adults. This finding is consistent with the evidence of greater frontal recruitment by older adults and suggests that frontal cortex is engaged particularly by those older adults who are highly educated. Thus, it may be that compensatory brain activity in older adults is most likely to be seen in those with higher education, but more work needs to be done in this area. Other variables, such as personality traits or mood, have yet to be examined, but as age differences have been found on these measures, these factors also could be influencing brain activity in older adults.

Finally, it is important to note that it is possible to identify a difference in functional connectivity, or network interactions, involving a particular brain area even when no change in mean activation is seen in this area. This indicates that a failure to find differences in brain activity between younger and older adults in a given region on a particular task does not mean that the functional communication of that region with other area is the same as well. Thus, one cannot always assume that the function of a brain area is unchanged with age merely by assessing its mean activity level. A more detailed examination of the dynamic interactions among brain regions, in conjunction with underlying structural changes and assessment of the impact of network activity on behavioral performance, will likely be necessary to ultimately understand the impact of aging on cognitive function.

See also: Aging of the Brain; Aging and Memory in Humans; Cognition in Aging and Age-Related Disease; Cognition: An Overview of Neuroimaging Techniques; Declarative Memory System: Anatomy; Emotional Learning in Humans; Episodic Memory; Humans; Memory Disorders; Neuroimaging.

Further Reading

Anderson ND and Grady CL (2004) Functional imaging in healthy aging and Alzheimer's disease. In: Cronin-Golomb A, and Hof P (eds.) *Vision in Alzheimer's Disease*, pp. 62–95. Basel: Karger.

Bookheimer SY, Strojwas MH, Cohen MS, et al. (2000) Patterns of brain activation in people at risk for Alzheimer's disease. *New England Journal of Medicine* 343: 450–456.

Buckner RL (2004) Memory and executive function in aging and AD: Multiple factors that cause decline and reserve factors that compensate. *Neuron* 44: 195–208.

Cabeza R, Nyberg L, and Park D (eds.) (2005) *Cognitive Neuroscience of Aging*. Oxford: Oxford University Press.

Ercoli L, Siddarth P, Huang SC, et al. (2006) Perceived loss of memory ability and cerebral metabolic decline in persons with the apolipoprotein E-IV genetic risk for Alzheimer disease. *Archives of General Psychiatry* 63: 442–448.

Gazzaley A, Cooney JW, Rissman J, et al. (2005) Top-down suppression deficit underlies working memory impairment in normal aging. *Nature Neuroscience* 8: 1298–1300.

Grady CL and Craik FIM (2000) Changes in memory processing with age. *Current Opinion in Neurobiology* 10: 224–231.

Grady CL, McIntosh AR, and Craik FI (2003) Age-related differences in the functional connectivity of the hippocampus during memory encoding. *Hippocampus* 13: 572–586.

Gutchess AH, Welsh RC, Hedden T, et al. (2005) Aging and the neural correlates of successful picture encoding: Frontal activations compensate for decreased medial temporal activity. *Journal of Cognitive Neuroscience* 17: 84–96.

Morcom AM, Good CD, Frackowiak RS, et al. (2003) Age effects on the neural correlates of successful memory encoding. *Brain* 126: 213–229.

Nordahl CW, Ranganath C, Yonelinas AP, et al. (2006) White matter changes compromise prefrontal cortex function in healthy elderly individuals. *Journal of Cognitive Neuroscience* 18: 418–429.

Rajah MN and D'Esposito M (2005) Region-specific changes in prefrontal function with age: A review of PET and fMRI studies on working and episodic memory. *Brain* 128: 1964–1983.

Small GW, Ercoli LM, Silverman DH, et al. (2000) Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* 97: 6037–6042.

Stern Y, Habeck C, Moeller J, et al. (2005) Brain networks associated with cognitive reserve in healthy young and old adults. *Cerebral Cortex* 15: 394–402.

Fusion Pore

J Zimmerberg, L-G Wu, and T S Reese, National Institutes of Health, Bethesda, MD, USA

© 2009 Published by Elsevier Ltd.

When an action potential in an axon depolarizes the membrane at a synapse, voltage-dependent calcium channels open to admit calcium, which then induces vesicles laden with neurotransmitters or neuropeptide to fuse with the presynaptic membrane and secrete their contents. This process of membrane fusion is not restricted to the nervous system but is a general feature of membranes in living cells. Fusion is used for exocytotic secretion at synapses and elsewhere, as well as viral entry, protein trafficking, fertilization, organogenesis, and wound healing. Fusion is not instantaneous but, rather, begins with a set of intermediate states establishing a direct aqueous pathway – a fusion pore – between the interior of the fusing vesicle or interior of a fusing cell and the extracellular space or the interior of the target cell (Figure 1). A minute pore connects the inside of the granule with the extracellular space, and then the pore expands rapidly to expel the contents of the granule.

Techniques

Detailed structural views of exocytosis came from electron microscope examination of ultrathin sections of stimulated secretory cells fixed, stained, and embedded in plastic. The description of omega-shaped fusion openings originated with the work of Douglass on the exocytosis of secretory granules in chromaffin cells. It was expected from the beginning that earlier stages consisting of much smaller vesicle openings or fusion pores must exist.

Early fusion pores were captured by freezing of cells just as they were stimulated to secrete. The idea was to cool the cells so rapidly – within a few milliseconds – that short-lived fusion pore intermediates would be trapped in vitreous ice. Electron microscopy of this material was done after freeze substitution to preserve the original conformations of membranes for ultrathin sectioning. In this way, fusion pores in the 10 nm range were found during granular exocytosis in both mast cells and the amebocyte of the horseshoe crab, *Limulus* (Figure 2). A search for pores in membrane fractures through the pore in both types of secretory cell, as well as at neuromuscular synapses, revealed an actual pore connecting the vesicle with the outside.

Because the early fusion pores were almost 10 times smaller than the thickness of a thin section, detecting them in *Limulus* required viewing stereo electron micrographs. The dimensions of the smallest pore that can be detected by stereo electron microscopy is in the range of 10 nm, but this range could potentially be extended to smaller pores by modern electron microscope tomography. Small fusion pores (10 nm range) were also detected in the fusion of the influenza virus to liposomes that contain the viral receptor ganglioside by freeze-fracture technique. The paucity of fusion pores in many systems is interpreted to reflect the short lifetime of this intermediate, estimated to be less than a few milliseconds at the presynaptic terminal.

One problem in the interpretation of electron microscopic images of fusion pores is that fusion pores associated with exocytosis are not easily differentiated in static images from endocytic fission pores, which are structural intermediates that arise immediately before the final pinching off of an endocytosing vesicle. Time-resolved admittance measurement is a method of detecting fusion that is useful for both differentiating between fusion and fission pores and detecting pores of smaller dimensions. This technique relies on the fact that fusion adds membrane to a cell, and the addition of membrane results in an increase in the capacitance of the whole cell (Figure 3). Although the fusion pore is small, the voltage drop across the vesicle membrane is submaximal because part of the voltage falls across the resistance of the fusion pore. Thus, the fusion pore conductance (permeability to ions) can be calculated at millisecond time resolution from time-resolved measurements of capacitance. More generally, two-phase admittance can be measured, extending the range of fusion pore detection. In addition, the frequency of the inquiring alternating voltage used to measure admittance or capacitance can be altered to probe different ranges of fusion pore conductance. In this way, pores were first detected in the exocytosis of a mast cell from a mutant (beige) mouse, an animal model for Chediak–Higashi syndrome of humans (Figure 3). The pore conductances in exocytosis measured by this technique range from 0.1 to 100 nS. To determine the dimensions of fusion pores that fit this measured conductance, one must make certain assumptions about the geometry of the pore, mainly the relationship between its length and width. Assuming a fusion pore length of 15 nm to account for twice the bilayer thickness and 5 nm of cytoplasm, and assuming a cylindrical geometry, we estimate the internal width of the smallest initial fusion pores measured by admittance to be approximately 1.5 nm. Since an α-helix of a protein is 1 nm in diameter, the measured fusion pores are thus of molecular dimensions.

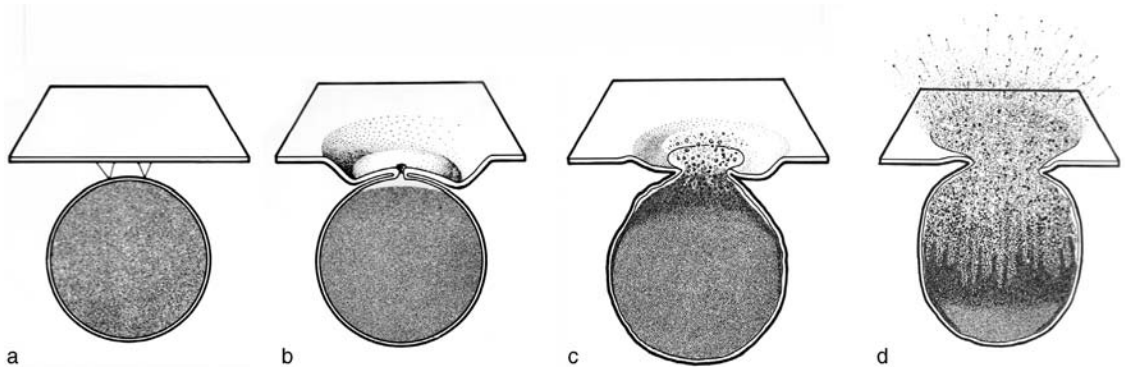

Figure 1 Stages of pore formation and enlargement based on observation with the electron microscope of thin sections through horseshoe crab blood cells (amebocytes). The amebocytes were stimulated to secrete clotting proteins from their granules. The granule membrane is pulled against the cell membrane and a minute pore forms (b, top). The pore rapidly expands (c, d) to release the clotting proteins. Expansion of the pore progresses from b to c in a fraction of a second, but it proceeds more slowly from c to d. These granules are approximately 1 μm in diameter. Reproduced from Ornberg RL and Reese TS (1981) Beginning of exocytosis captured by rapid-freezing of *Limulus* amebocytes. *Journal of Cell Biology* 90: 40–54.

Almost immediately upon application of the time-resolved admittance measurement technique to cells with large granules, where fusion pore openings could be separated from each other in time, it was appreciated that the fusion pore can close once open. This phenomenon, termed capacitance 'flicker,' is widely found in exocytosis and viral fusion. Even very large pores, up to 8 nm in diameter, can close relatively soon after opening.

Another method for detecting initial fusion pores relies on the discharge of charge stored within vesicles (Figure 4). The resting voltage across the cell membrane can be set to a value different from that of the transmembrane potential. Upon the opening of the fusion pore, the difference in potential is equilibrated, resulting in an exponential flux of charge across the fusion pore. By measuring the exponential decay in this charge dissipation, one can calculate the conductance of the initial fusion pore and confirm its uniformity of size over the time constant of the vesicle capacitance–fusion pore resistance product.

A final method for fusion pore detection, amperometry, relies on detecting directly the discharge of vesicular contents (Figure 5). Certain vesicular contents oxidize or reduce at the surface of an electrochemical electrode (usually a carbon fiber). By placing a carbon fiber near the cell membrane, the momentary burst of released contents can be detected in a time-resolved manner. Occasionally, a slow release can be seen prior to the burst of release, and from this slow flux of contents the fusion pore conductance can be estimated. For vesicles containing insufficient amounts of electrochemically active compounds, often appropriate chemicals can be incubated with the secreting cells to load the vesicles with these exogenous compounds, thus providing a stronger signal for electrochemical measurements than the endogenous contents.

The methods for detecting fusion pores have been refined to allow detection of not only changes to the whole cell membrane but also individual events in both on-cell patches of membrane and excised patches (Figure 6). In addition, hybrid pipettes have been created that allow for simultaneous measurements of capacitance and amperometry.

Fusion Pore Dimensions and Dynamics

Since the smallest fusion pores detected have conductances comparable to those of single ionic channels, the measurement of fusion pores reflects the molecular architecture of the fusion apparatus. Remarkably, both the initial fusion pore and the dynamics of its widening are the same in almost every exocytotic system in which it has been measured, supporting the notion that there is a common molecular mechanism to exocytic fusion.

The initial fusion pore has a very broad distribution of conductance and, hence, size (Figure 7(a)). This fact has been used to reflect a fluid component to the architecture of the fusion pore and thus to the mechanism of fusion (Figure 7(b)). It is not known whether the lumen of the fusion pore is lined with proteins or lipids. There is no way to determine this directly, and the indirect evidence is contradictory. Phenylalanine mutations of the transmembrane domain of syntaxin alter amperometric 'foot' conductance, which suggests an effect of the transmembrane domain on pore conductance. Phospholipid composition alters the fusion pore opening, suggesting that it begins as a lipidic pore within a proteinaceous scaffold. The simplest way to accommodate the data is to imagine that the protein contributing to the initial fusion pore is intimately associated with its structure, either within lipid that arrives later to wet the

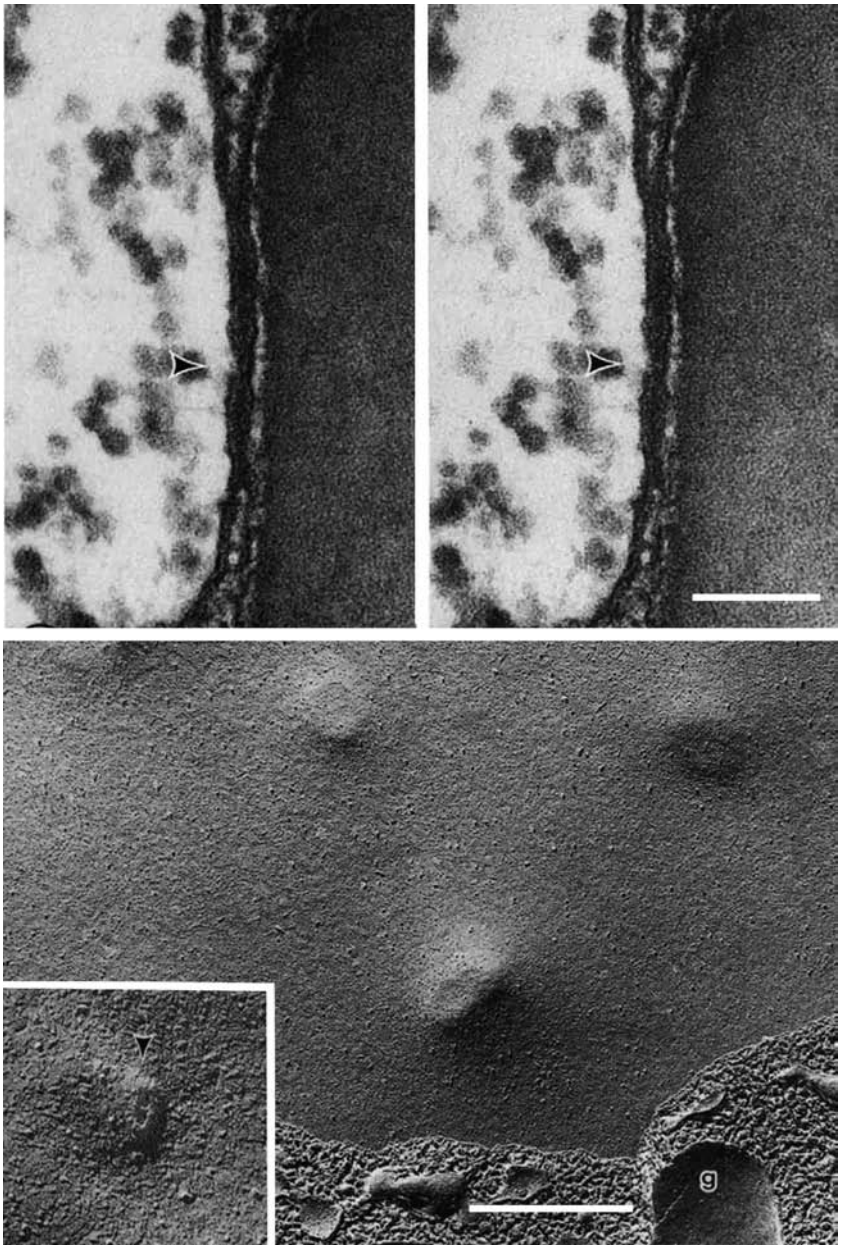

Figure 2 Electron micrographs of early fusion pores in horseshoe crab amebocytes (see **Figure 1b**) as seen by freeze substitution (top) and freeze fracture (bottom). The two freeze-substitution images (top) are a stereo pair of a contact between a granule (on the right in both views) and the outside of the amebocyte (on the left in both views). A pore approximately 30 nm in diameter (arrowhead) connects the contents of a granule to the outside of the cell. The freeze-fracture views (below) show the same stage in pore formation. The outer membrane of the amebocyte is pulled up to make a cup containing the granule (granule is fractured away). Another cup (inset) shows a pore approximately 30 nm in diameter that has formed on top of a cup. The small pores are very rare, reflecting their transient, millisecond lifetimes. Scale bars = 0.1 μm (top) and 0.5 μm (bottom). Reproduced from Ornberg RL and Reese TS (1981) Beginning of exocytosis captured by rapid-freezing of *Limulus* amebocytes. *Journal of Cell Biology* 90: 40–54.

connection or from outside as a scaffold. Note that the ionic conductance of the initial pore is in series with its associated access resistance, so the effect of protein or lipid on pore conductance can be due to changes to the outer vestibules of the pore rather than the pore center. In summary, the initial fusion pore has both lipidic and proteinaceous aspects; the structure of the fusion pore is dependent on both proteins and lipids.

There is good evidence for a stage prior to fusion pore opening and widening that depends on the spontaneous curvature of the membrane. Lysolipids

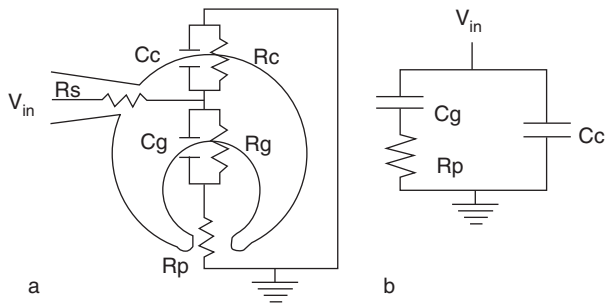

Figure 3 (a) Sketch of the exocytotic pore that has resulted from fusion. Superimposed on the figure is the equivalent circuit. The phase of the lock-in amplifier was set as described by Debus K and Lindau M (2000), thereby compensating the phase shift induced by the series resistance. Rc and Cc, resistance and capacitance of the cell membrane, respectively; Rg and Cg, resistance and capacitance of the granule membrane, respectively; Rp, resistance of the exocytotic pore; Rs, series resistance of the patch electrode. (b) The simplified equivalent circuit. Rc and Rg can be neglected for a sine wave of 806 Hz because Rc,g $<<$ $1/2\pi f$Cc,g. The change in admittance as a result of fusion is given in complex notation by $\Delta Y(\omega) = \mathrm{Rp}(\omega\mathrm{Cg})^2/((\omega\mathrm{CgRp})^2+1) + j(\omega\mathrm{Cg})/((\omega\mathrm{CgRp})^2+1)$. Note that as a result of fusion, Rp varies from infinity before fusion to a small value after fusion. Fusion does not cause any net change in the real part of the admittance. However, during the time that the capacitance increases, Rp decreases, and the real part of the admittance increases and then decreases. The maximum possible change (due to varying Rp) in the real component of the admittance equals ωCg/2, which is one-half the final change of the imaginary component, ωCg. The experimental conductance transients (real component of the admittance) are in quantitative agreement with this theoretical prediction, lending support to this model. For different events, we recorded times $t1/2$ between the initiation of the capacitance growth and the half-maximal capacitance value over the range of 2–59 ms. From the measured capacitance jumps, we calculated that Rp was between 3.1×10^8 and 2.5×10^9 ohms (Gp between 0.4 and 3.2 nS) at times $= t1/2$. Assuming the pore to have the conductivity of bulk solution and a length of 100 A, we estimate that at the time of half-maximal capacitance the pore diameter is between 18 and 52 A. In contrast, in *Limulus* amebocytes the smallest exocytotic pore seen in freeze-substituted sections by electron microscopy was -100 A wide and 150 A long. Adapted from Zimmerberg J, Curran M, Cohen FS, and Brodwick M (1987) Simultaneous electrical and optical measurements show that membrane fusion precedes secretory granule swelling during exocytosis of beige mouse mast cells. *Proceedings of the National Academy of Sciences of the United States of America* 84: 1585–1589.

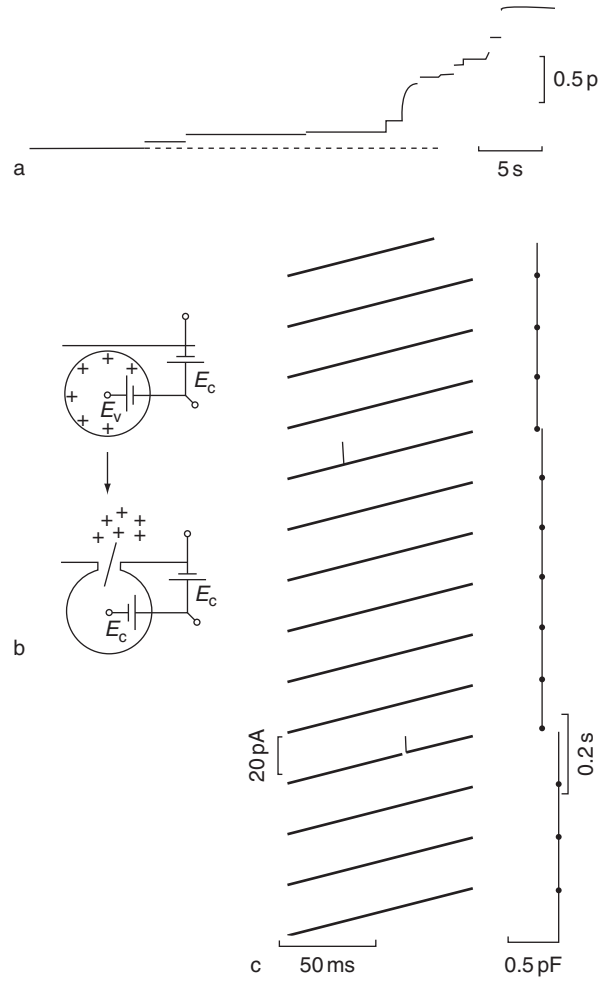

Figure 4 (a) Membrane capacitance plotted against time in a degranulating mast cell voltage-clamped at 0 mV. Each time a vesicle fuses with the cell membrane, there is a step increase in C_m. (b) Potentials across cell membrane (E_c) and vesicle membrane (E_v) defined as cytosol minus outside and cytosol minus lumen, respectively. (c) Membrane current (thick traces) during a short portion of an experiment as in a except that the sinusoid used to measure C_m was applied only once every 112.5 ms for 12.5 ms at a time. Holding potential, 10 mV. Vertical trace, mean C_m values during the successive 12.5 ms periods; dots, time of measurement. Often, the increase in C_m due to exocytosis of a single vesicle continued over two or three 125 ms episodes; in such cases, the final amplitude of the C_m step was used in the analysis. Reproduced from Breckenridge LJ and Almers W (1987) Currents through the fusion pore that forms during exocytosis of a secretory vesicle. *Nature* 328: 814–817.

that confer positive spontaneous curvature to the monolayer leaflet to which they are added universally inhibit membrane fusion when added to the contacting leaflets before stimulation to fuse. These and other effects of lipids are consistent with the stalk–pore model of membrane fusion. In this hypothesis, the pore develops within a high-energy intermediate, the tension-stressed stalk, whose free energy exceeds that of a small, bent, closed pre-pore. The transmembrane segment of syntaxin, the juxtamembrane segment of SNAREs, and the C2 domains of synaptotagmin have all been implicated in the formation of the fusion pore at this stage.

The Lifetime of the Initial Fusion Pore Varies Widely

After the initial fusion pore widens, a much longer phase of fusion pore fluctuation occurs during which

there is a slow superimposed growth to the fusion pore conductance. Since the widening of the fusion pore requires less motion of bulk material than does the lengthening, we assume that the fast fluctuations detected reflect widening of the pore as the pore slowly lengthens. At some point, there is an abrupt increase in pore conductance and usually recording is terminated at this time. However, there appear to be another phase of fusion pore widening that correlates with a swelling of the vesicular gel components.

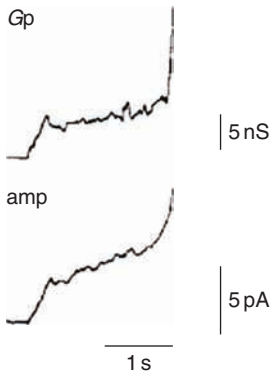

Figure 5 Release of serotonin during slow expansion of a fusion pore before completing irreversible fusion. The complex admittance is used to calculate the conductance of the fusion pore (G_p) during these early moments of expansion. The fusion pore conductance is shown to be well correlated with the amperometric current (amp). Reproduced from Alvarez De Toledo G, Fernandez-Chacon R, and Fernandez JM (1993) Release of secretory products during transient vesicle fusion. *Nature* 363: 554–558.

Fusion Pores in Synaptic Transmission

The techniques been used in endocrine cells to measure the fusion pore have encountered two problems when applied to synapses. First, synaptic vesicles are much smaller than the large dense-core vesicles in endocrine cells, making it much more difficult to detect individual vesicle fusion and the fusion pore conductance. Second, for detection of small single vesicle fusion, fusion must occur at the patched membrane. However, the majority of fusion, if not all, at the nerve terminal occurs at release sites apposing the postsynaptic neuron. The postsynaptic neuron may thus prevent the electrode from accessing the release site. These two hurdles have recently been overcome.

In pituitary nerve terminals that do not form synapses, individual vesicle fusion and the fusion pore conductance have been resolved for microvesicles with a size similar to that of synaptic vesicles. Like the much larger granules in secretory cells, the smallest pores appear to be reversible. A pore conductance of approximately 19 pS was found in approximately 5% of fusion events, indicating reversibility. The reversibility led to the hypothesis that some vesicle contents might normally be released through the small fusion pore, which could then close and allow the vesicle to be reloaded, and the hypothesis that has come to be known as kiss-and-run in the context of release of transmitter from synaptic vesicles. Note that no images from the electron microscope show kiss-and-run openings since, as noted previously, static images cannot distinguish between exo- and endocytosis.

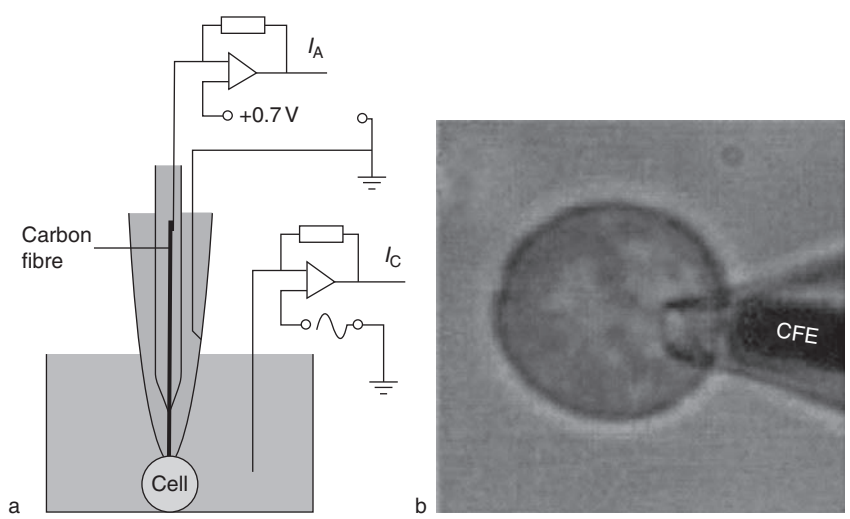

Figure 6 (a) Arrangement of a CFE inside a patch pipette. I_A, amperometric current; I_C, sine wave current used to measure capacitance changes. (b) Chromaffin cell with attached patch pipette containing CFE at the beginning of the experiment. Reproduced from Abillos et al.

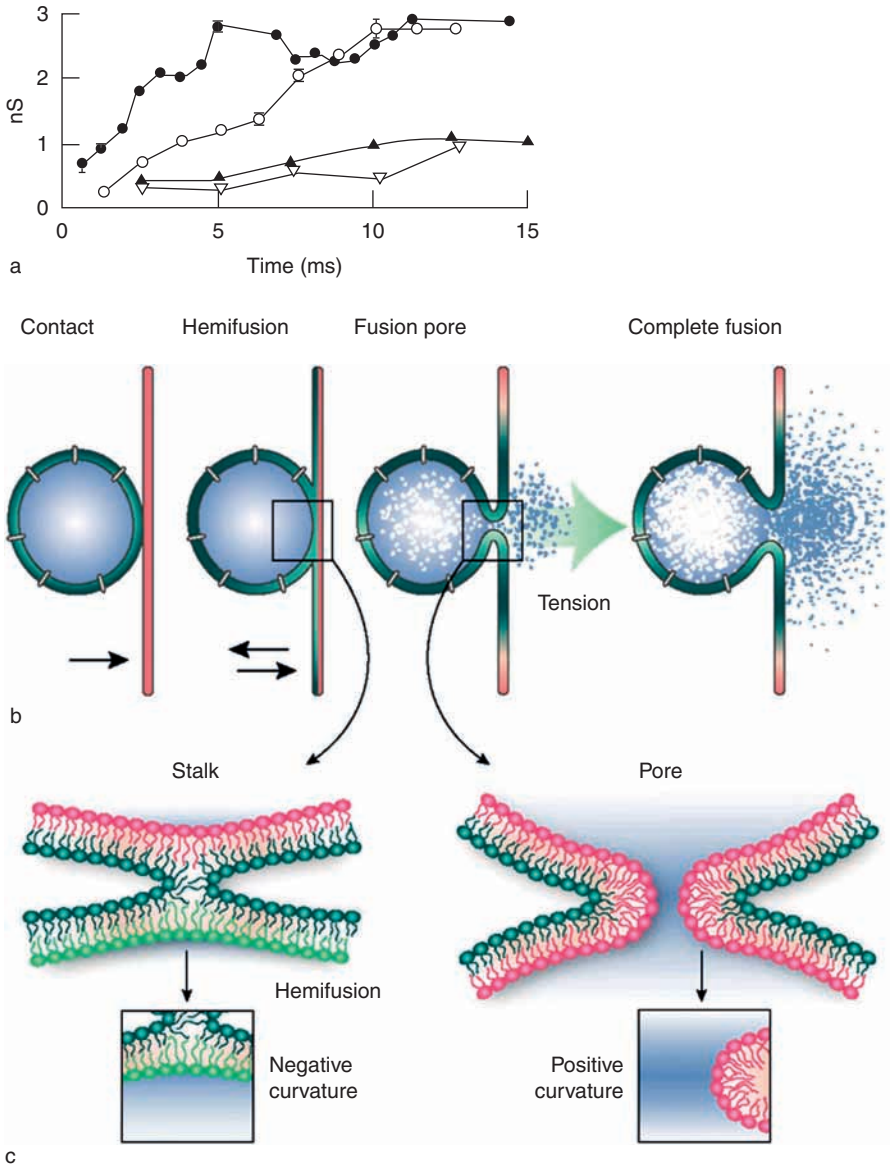

Figure 7 The first detectable fusion pore is polymorphic in its behavior. It can open to just about any size, without any restricted regions. (b) Biological membrane fusion is likely to proceed through hemifusion to prevent hydration of lipid tails during this transformation of topology. The initial hemifusion intermediate (the stalk) is a structure of overall negative spontaneous monolayer curvature and should be favored in membranes whose composition is negative overall in spontaneous monolayer curvature, such as membranes rich in fatty acids. Likewise, fusion should be inhibited in membranes whose composition is positive overall in spontaneous monolayer curvature, such as membranes rich in lysophospholipids; this prediction is borne out experimentally in all systems tested. (c) Energetically, the major problem in fusion is to convert the stalk to a pore, increasing the radius of the neck of the stalk. For phospholipid membrane fusion, tension is needed to widen the fusion pore. (a) Reproduced from Curran M, Cohen FS, Chandler DE, Munson PJ, and Zimmerberg J (1993) Exocytotic fusion pores exhibit semi-stable states. *Journal of Membrane Biology* 133: 61–75, with permission. (b, c) Reproduced from Zimmerberg J and Gawrisch K (2006) The physical chemistry of biological membranes. *Nature Chemical Biology* 2: 564–567.

The capacitance recording methods developed to resolve small vesicle fusion in this study made it possible to apply the same method to synaptic nerve terminals. To date, the capacitance technique has been applied to only the calyx of Held synapse. To expose the release site of the synaptic nerve terminal, the calyx of Held, a large pipette was used to pull out the post-synaptic neuron (Figure 8(a)). At the exposed release sites, cell-attached recordings revealed individual capacitance up-steps reflecting single synaptic vesicle fusion (Figure 8(b)). For full collapse fusion, the initial fusion pore conductance is usually >375 pS and increased rapidly to ≥299 pS/ms. In a small number of full collapse events, the initial fusion pore conductance of 260 pS was detected (Figure 8(c)). Support for the kiss-and-run type or release is reflected as a brief

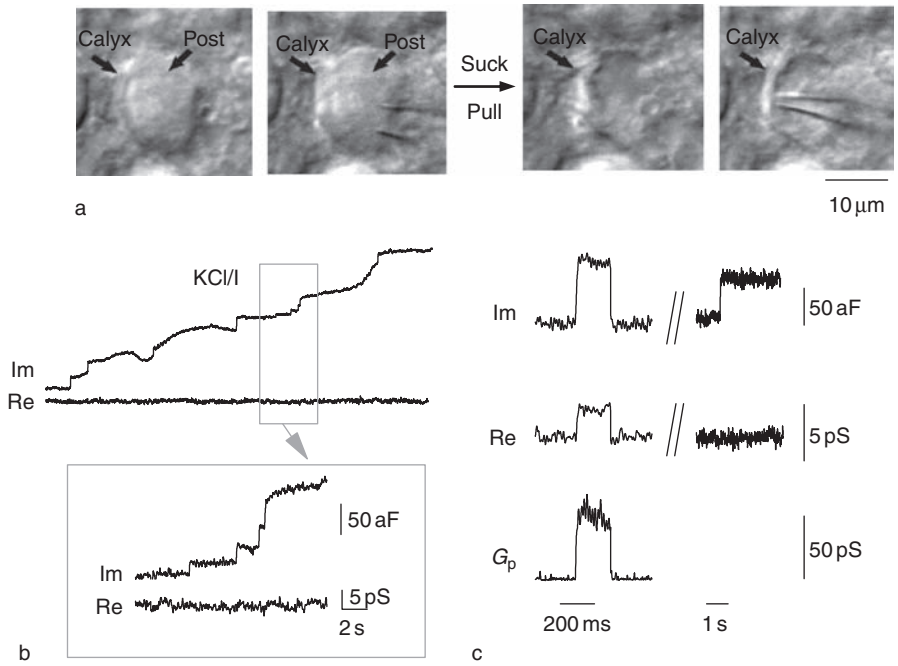

Figure 8 Full collapse and kiss-and-run fusion at a calyx-type synapse. (a) The procedure to perform cell-attached recordings at the release face of a calyx. A calyx associated with a postsynaptic neuron was first identified (left). The postsynaptic neuron was sucked and pulled away by a pipette (middle two panels). Another pipette was positioned at the release face of the calyx membrane (right) for cell-attached recordings, from which individual vesicle fusions, such as those shown in b and c were obtained. (b) Im (imaginary component of the admittance, reflecting capacitance) and Re (real component of the admittance, reflecting conductance) traces from a cell-attached recording during application of 25 mM KCl. The inset shows discrete up-steps reflecting single vesicle fusion. (c) Im, Re, and the fusion pore conductance (G_p) during a capacitance flicker. The capacitance flicker with a detectable fusion pore conductance indicates kiss-and-run fusion. A nonflicker up-step (right) occurring 10 s later was not accompanied by detectable Re changes, suggesting a full collapse fusion with the fusion pore that is too fast and/or too large to resolve. Adapted from He L, Wu XS, Mohan R, and Wu LG (2006) Two modes of fusion pore opening revealed by cell-attached recordings at a synapse. *Nature* 444: 102–105.

(<2 s) capacitance flicker with $G_p > 288$ pS for most flickers, but within 15–288 pS for the remaining flickers. As in secretory cells, most fusion events end in full collapse under our conditions.

Comparison of Exocytic and Cell–Cell Fusion Pores

Although all exocytic fusion pores studied to date have very similar initial parameters, the initial fusion pores catalyzed by viral fusion proteins are more varied with regard to conductance, lifetime, and voltage dependence. For example, the initial fusion pore mediated by the baculovirus envelope protein gp64 is approximately 1 nS in conductance and does not show any flickering behavior. Since these proteins are typically expressed in genetically engineered cells, and their fusion pores are measured in cell–cell fusion systems, there are many differences between both the proteins and the membranes in which they are expressed, as well as differences in the local density of the crucial proteins. The relative contribution of these various factors remains to be determined.

See also: Action Potential Initiation and Conduction in Axons; Active Zone; Endocytosis: Kiss and Run; Ion Channel Localization in Axons; Large Dense Core Vesicles (LDCVs); Synaptic Vesicles.

Further Reading

Alvarez De Toledo G, Fernandez-Chacon R, and Fernandez JM (1993) Release of secretory products during transient vesicle fusion. *Nature* 363: 554–558.

Breckenridge LJ and Almers W (1987) Currents through the fusion pore that forms during exocytosis of a secretory vesicle. *Nature* 328: 814–817.

Ceccarelli B, Hurlbut WP, and Mauro A (1973) Turnover of transmitter and synaptic vesicles at the frog neuromuscular junction. *Journal of Cell Biology* 57(2): 499–524.

Chandler DE and Heuser JE (1980) Arrest of membrane fusion events in mast cells by quick-freezing. *Journal of Cell Biology* 86: 666–674.

Chanturiya A, Chernomordik LV, and Zimmerberg J (1997) Flickering fusion pores comparable with initial exocytotic pores occur in protein-free phospholipid bilayers. *Proceedings of the National Academy of Sciences of the United States of America* 94: 14423–14428.

Chernomordik L, Frolov V, Leikina E, Bronk P, and Zimmerberg J (1998) The pathway of membrane fusion catalyzed by influenza

hemagglutinin: Restriction of lipids, hemifusion, and lipidic fusion pore formation. *Journal of Cell Biology* 140: 1369–1382.

Chizmadzhev YA, Cohen FS, Shcherbakov A, and Zimmerberg J (1995) Membrane mechanics can account for fusion pore dilation in stages. *Biophysical Journal* 69: 2489–2500.

Curran M, Cohen FS, Chandler DE, Munson PJ, and Zimmerberg J (1993) Exocytotic fusion pores exhibit semi-stable states. *Journal of Membrane Biology* 133: 61–75.

Debus K and Lindau M (2000) Resolution of patch capacitance recordings and of fusion pore conductances in small vesicles. *Biophysical Journal* 78: 2983–2997.

Fisher RJ, Pevsner J, and Burgoyne RD (2001) Control of fusion pore dynamics during exocytosis by Munc18. *Science* 291: 875–878.

Frolov V, Cho M, Bronk P, Reese T, and Zimmerberg J (2000) Multiple local contact sites are induced by GPI-linked influenza hemagglutinin during hemifusion and flickering pore formation. *Traffic* 1: 622–630.

Graham ME, Fisher RJ, and Burgoyne RD (2000) Measurement of exocytosis by amperometry in adrenal chromaffin cells: Effects of clostridial neurotoxins and activation of protein kinase C on fusion pore kinetics. *Biochimie* 82: 469–479.

Hafez I, Kisler K, Berberian K, et al. (2005) Electrochemical imaging of fusion pore openings by electrochemical detector arrays. *Proceedings of the National Academy of Sciences of the United States of America* 102: 13879–13884.

He L, Wu XS, Mohan R, and Wu LG (2006) Two modes of fusion pore opening revealed by cell-attached recordings at a synapse. *Nature* 444: 102–105.

Heuser JE and Reese TS (1973) Evidence for recycling of synaptic vesicle membrane during transmitter release at the frog neuromuscular junction. *Journal of Cell Biology* 57(2): 315–344.

Heuser JE, Reese TS, Dennis MJ, Jan Y, Jan L, and Evans L (1979) Synaptic vesicle exocytosis captured by quick freezing and correlated with quantal transmitter release. *Journal of Cell Biology* 81: 275–300.

Kanaseki T, Kawasaki K, Murata M, Ikeuchi Y, and Ohnishi S (1997) Structural features of membrane fusion between influenza virus and liposome as revealed by quick-freezing electron microscopy. *Journal of Cell Biology* 137: 1041–1056.

Klyachko VA and Jackson MB (2002) Capacitance steps and fusion pores of small and large dense-core vesicles in nerve terminals. *Nature* 418: 89–92.

Lindau M and Alvarez de Toledo G (2003) The fusion pore. *Biochimica et Biophysica Acta* 164: 167–173.

Lollike K, Borregaard N, and Lindau M (1995) The exocytotic fusion pore of small granules has a conductance similar to an ion channel. *Journal of Cell Biology* 129: 99–104.

Melikyan GB, Markosyan RM, Brener SA, Rozenberg Y, and Cohen FS (2000) Role of the cytoplasmic tail of ecotropic moloney murine leukemia virus Env protein in fusion pore formation. *Journal of Virology* 74: 447–455.

Melikyan GB, Niles WD, and Cohen FS (1995) The fusion kinetics of influenza hemagglutinin expressing cells to planar bilayer membranes is affected by HA density and host cell surface. *Journal of General Physiology* 106: 783–802.

Melikyan GB, Niles WD, Peeples ME, and Cohen FS (1993) Influenza hemagglutinin-mediated fusion pores connecting cells to planar membranes: Flickering to final expansion. *Journal of General Physiology* 102: 1131–1149.

Monck JR, Alvarez de Toledo G, and Fernandez JM (1991) Tension in secretory granule membranes causes extensive membrane transfer through the exocytotic fusion pore. *Proceedings of the National Academy of Sciences of the United States of America* 87: 7804–7808. [Erratum in *Proceedings of the National Academy of Sciences of the United States of America* 88: 2035].

Ornberg RL and Reese TS (1981) Beginning of exocytosis captured by rapid-freezing of *Limulus* amebocytes. *Journal of Cell Biology* 90: 40–54.

Plonsky I and Zimmerberg J (1996) The initial fusion pore induced by baculovirus GP64 is large and forms quickly. *Journal of Cell Biology* 135: 1831–1839.

Ratinov V, Plonsky I, and Zimmerberg J (1998) Fusion pore conductance: Experimental approaches and theoretical algorithms. *Biophysical Journal* 74: 2374–2387.

Scepek S, Coorssen JR, and Lindau M (1998) Fusion pore expansion in horse eosinophils is modulated by Ca^{2+} and protein kinase C via distinct mechanisms. *EMBO Journal* 17: 4340–4345.

Spruce AE, Breckenridge LJ, Lee AK, and Almers W (1990) Properties of the fusion pore that forms during exocytosis of a mast cell secretory vesicle. *Neuron* 4: 643–654.

Wang CT, Grishanin R, Earles CA, et al. (2001) Synaptotagmin modulation of fusion pore kinetics in regulated exocytosis of dense-core vesicles. *Science* 294: 1111–1115.

Zimmerberg J, Blumenthal R, Curran M, Sarkar DP, and Morris SJ (1994) Restricted movement of lipid and aqueous dyes through pores formed by influenza hemagglutinin during cell fusion. *Journal of Cell Biology* 127: 1885–1894.

Zimmerberg J, Blumenthal R, Sarkar DP, Curran M, and Morris SJ (1994) Restricted movement of lipid and aqueous dyes through pores formed by influenza hemagglutinin during cell fusion. *Journal of Cell Biology* 127: 1885–1894.

Zimmerberg J, Curran M, Cohen FS, and Brodwick M (1987) Simultaneous electrical and optical measurements show that membrane fusion precedes secretory granule swelling during exocytosis of beige mouse mast cells. *Proceedings of the National Academy of Sciences of the United States of America* 84: 1585–1589.

Zimmerberg J and Gawrisch K (2006) The physical chemistry of biological membranes. *Nature Chemical Biology* 2: 564–567.

GABA Synthesis and Metabolism

K L Behar, Yale University School of Medicine, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

γ-Aminobutyric acid (GABA) is the major inhibitory neurotransmitter in mammalian brain and is found widely throughout the central and peripheral nervous systems. In the neocortex GABAergic neurons are plentiful, constituting 15–30% of all neurons. GABA serves both metabolic and trophic functions, in addition to its role as a neurotransmitter, influencing the migration of neurons and astroglia to their target locations in the cortex. During early brain development GABA elicits excitatory (depolarizing) rather than inhibitory (hyperpolarizing) postsynaptic responses. Later in development GABA influences the synaptic organization and fine-tuning of local circuits. Because of GABA's many important roles – as neurotransmitter, neuromodulator, trophic factor, and cellular metabolite – GABA metabolism may impact many aspects of brain function.

Brain tissue is highly heterogeneous, and compartmentation of the enzymes of the GABA shunt between neurons and astroglia plays an important role in the function of GABA as a neurotransmitter. The discovery of metabolic compartmentation in the early 1960s to the late 1970s laid the conceptual foundation for our current understanding of brain glutamate and GABA metabolism. Studies using ^{14}C- and ^{15}N-labeled substrates revealed that brain glutamate metabolism was not homogeneous and could be separated kinetically into distinct pools – a large pool labeled from [^{14}C]glucose and a small, rapidly turning over pool labeled from [^{14}C]acetate, [^{14}C]butyrate, and [^{14}C]bicarbonate. The subsequent discoveries that glutamine synthetase (GS; glutamate–ammonia ligase, EC 6.3.1.2) and pyruvate carboxylase (PC; EC 6.4.1.1) were astroglial enzymes provided key information that the large and small pools of glutamate corresponded to metabolism in neurons and astroglia, respectively. Glutamatergic and GABAergic neurons are included within the definition of the 'large pool,' although glutamate levels in the respective neurons differ greatly. Most of the 8–12 μmol g^{-1} of glutamate measured in brain tissue is present in glutamatergic neurons (~80%), with GABA neurons (~2–10%) and astroglia (~10%) containing much less. Astroglia play a critical role in glutamatergic and GABAergic function by maintaining low synaptic and interstitial levels of glutamate and GABA through active uptake and metabolism. Astroglia also provide glutamine, and potentially other precursors, necessary for the resynthesis of these neurotransmitters.

The study of brain glutamate and GABA metabolism has been advanced significantly in recent years through the use of nuclear magnetic resonance (NMR) spectroscopy in conjunction with ^{13}C- and ^{15}N-labeled substrates. The use of isotopically labeled substrates in these studies allows metabolic pathways of GABA and glutamate synthesis and neuron/glial substrate trafficking to be examined noninvasively and in unprecedented detail, revealing new insights into these pathways and their functional interactions.

GABA Synthesis and Degradation in the GABA Shunt

GABA synthesis and degradation occur at the level of the tricarboxylic acid (TCA) cycle. GABA, a product of glucose metabolism, is synthesized almost exclusively from glutamate, a product of the TCA cycle intermediate, α-ketoglutarate, in an α-decarboxylation catalyzed by the pyridoxal-5′-phosphate-dependent enzyme, glutamate decarboxylase (GAD; EC 4.1.1.15). GABA is degraded to succinate by the concerted actions of two enzymes: pyridoxal-5′-phosphate-dependent GABA transaminase (GABA-T; 4-aminobutyrate transaminase, EC 2.6.1.19) and nicotinamide adenine dinucleotide (NAD)-dependent succinate-semialdehyde dehydrogenase (SSADH; EC 1.2.1.24). In the first step of catabolism, GABA nitrogen is transferred to α-ketoglutarate by transamination, forming succinic-semialdehyde and glutamate. In the second step succinic-semialdehyde is oxidized to succinate with reduction of NAD^+ to $NADH_2$; succinic-semialdehyde is a short-lived intermediate and is present at low cellular concentrations. The pathway of GABA metabolism from α-ketoglutarate to succinate forms a metabolic shunt, by passing two enzymes of the TCA cycle, α-ketoglutarate dehydrogenase and succinyl-CoA synthetase (Figure 1). Glucose oxidation through the GABA shunt is slightly (~3%) less energy efficient than through the complete TCA cycle, resulting in one less molecule of guanosine triphosphate (GTP) for each molecule of α-ketoglutarate traversing the shunt.

GABA synthesis occurs only in neurons, whereas GABA catabolism occurs both in neurons and glia. Thus, the resynthesis of glutamate from α-ketoglutarate during transamination with GABA does not occur totally in the same cells in which GABA was formed. Because GABAergic and other neurons are incapable of *de novo* synthesis of glutamate precursors, due to

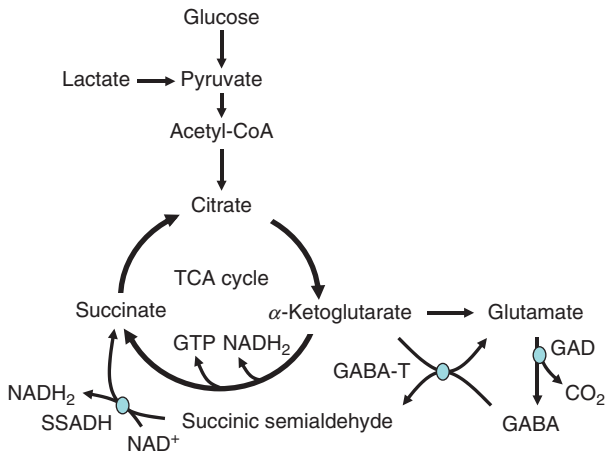

Figure 1 Metabolic pathway depicting the synthesis and catabolism of γ-aminobutyric acid (GABA) in the GABA shunt and its relationship to glucose metabolism and the tricarboxylic acid (TCA) cycle. The GABA shunt bypasses two enzymatic steps of the TCA cycle, resulting in one less molecule of guanosine triphosphate (GTP) formed for each molecule of α-ketoglutarate metabolized to succinate through the shunt. GABA synthesis occurs in GABAergic neurons but GABA catabolism occurs in neurons and astroglia. Astroglial glutamate precursors (e.g., glutamine) are required to replenish GABA removed and metabolized by astroglia. GAD, glutamic acid decarboxylase; GABA-T, GABA transaminase; SSADH, succinic semialdehyde dehydrogenase; NAD, nicotinamide dinucleotide; CoA, coenzyme A.

the absence of the necessary anaplerotic enzymes, glutamate precursors must be supplied by astroglia to maintain TCA cycle intermediates (and GABA) at constant levels.

The concentration of GABA in brain tissue ranges from ~1 to 5 μmol g^{-1}, with the variation in concentration reflecting the density of GABA neurons and their terminals. GABA levels are very low in the extracellular fluid and in cells which do not express GAD. The GABA concentration within GABAergic terminals may reach 50–100 mM, although much of this is contained in synaptic vesicles and is inaccessible to GABA-T. The concentration of GABA in the cytoplasm and the mitochondrial intermembrane space where GABA-T is located is not well known, but levels are likely to be high and saturating for GABA-T, which has a relatively low K_m (~1 mM). Thus, in neurons GABA levels are likely to be regulated by changes in synthesis and GAD activity rather than catabolism. In contrast to neurons, in astroglia where GABA levels are low and nonsaturating for the glial transaminase, GABA level is likely to regulate the rate of catabolism.

The GABA shunt also generates precursors for synthesis of a limited number of other compounds, the most studied of which are the neuromodulator/neurotransmitters, homocarnosine (γ-aminobutyryl histidine) and γ-hydroxybutyrate (GHB). Homocarnosine is synthesized from GABA and histidine by homocarnosine synthetase, and hydrolysis of this dipeptide by a serum carnosinase regenerates GABA and histidine. GHB is synthesized from succinic-semialdehyde by a specific NADPH-dependent succinic-semialdehyde reductase, and oxidation back to succinic-semialdehyde is catalyzed by GHB dehydrogenase. Tissue levels of GABA, homocarnosine, and GHB can be elevated substantially by pharmacologic inhibitors or gene deficiencies in GABA-T and SSADH.

GABA Transporters Are Expressed in GABAergic Neurons and Astroglia

GABA is transported across the plasma membranes of neurons and astroglia by electrogenic GABA transporters (GATs), which facilitate the reversible symport of GABA with two Na^+ ions and one Cl^- ion. These transporters, like the electrogenic glutamate transporters, require the energy of ATP to maintain the Na^+ ion gradient that provides the driving force and direction of GABA transport. GATs operate at much slower rates than do glutamate transporters, so that efficient clearance of GABA from the synapse requires relatively high transporter densities. The rate of GABA transport is determined by the number of functional GATs residing on the cell surface which are recruited from the cytoplasm. The distribution of GATs between cell surface and cytoplasm is regulated by extracellular GABA levels, protein kinase C, and ligands of G-protein-coupled receptors.

Pathways of GABA Clearance Following Release: Neuronal Reuptake and Glial Uptake

GABA clearance following release into the synaptic cleft occurs both by reuptake into presynaptic terminals and by transport into surrounding astroglia. Isotopic studies which follow the labeling of GABA from different precursors cannot distinguish between the different uptake pathways because GABA metabolism in neurons and glia leads to the same labeling patterns. Thus, the relative quantitative importance of these two pathways to total GABA clearance *in vivo* is unknown and this issue remains a vexing problem. Studies using cell cultures have shown that GABA transport capacity and uptake are greater in neurons than in glia, supporting a predominant role for reuptake in GABA clearance from the synapse. Although reuptake has been suggested to be more energetically efficient because new GABA

synthesis is not required, the activity of GABAergic neurons and vesicular loading of GABA has been shown to involve new GABA synthesis, thus weakening a rationale for reuptake.

GATs are reversible and can function in GABA release as well as uptake. GATs in interneurons operate closer to their equilibrium potentials than do glutamate transporters, providing more favorable conditions for calcium-independent, nonvesicular transport of GABA in either direction. Under depolarizing conditions GABA transport reversal can lead to loss of GABA from the cell, but it has not been shown whether the recovery of GABA levels during repolarization involves reuptake or new synthesis from glial precursors.

GABAergic synapses are closely surrounded by glial end-processes possessing high densities of GATs, and astroglia possess all the enzymes needed to degrade GABA. Studies of isolated glial fractions, astrocyte cultures, and intracellular recordings of glial GABA transporter currents in tissue slices show that glia have a substantial capacity to transport GABA. *In situ* levels of extracellular GABA and the occupancy of postsynaptic GABA receptors are regulated by astroglial transporters. Inhibitors of astroglial transporters show greater antiepileptic efficacy than do neuronal transport inhibitors of GABA, suggesting that glial clearance is quantitatively significant. These observations are consistent with the high rate of GABA synthesis from glutamine (~23% of total glutamate plus GABA cycling) measured in rat cortex *in vivo* using ^{13}C NMR. Whether glial GABA transport is more important in paracrine signaling or synaptic transmission is unclear at present, but such differences may be related to the functions of the two isoforms of GAD, as discussed in a later section.

Precursors of GABA Synthesis

The Glutamate/GABA–Glutamine Cycle

Glutamate precursors for GABA synthesis can be produced directly from the metabolism of glucose in neurons. However, glucose metabolism cannot replace glutamate and GABA carbon lost from neurons during activity due to the absence of the necessary anaplerotic enzymes (e.g., PC), which are expressed only in the astroglia. GABAergic neurons import the needed precursors from astroglia mainly in the form of glutamine. The astroglia which surround GABAergic (and glutamatergic) neuronal processes express high densities of the N-type electroneutral glutamine transporter (SN1), while GABAergic neurons express both of the A-type electrogenic glutamine transporters (SA1 and SA2). Glutamine is hydrolyzed to glutamate and ammonia by phosphate-activated glutaminase (PAG; EC 3.5.1.2), a highly regulated enzyme, present on the outer surface of the inner mitochondrial membrane, which provides glutamate for GABA synthesis to recharge cytoplasmic and vesicular pools. Release of GABA from neurons followed by uptake and metabolism to succinate in the astroglia completes the GABA shunt while regenerating glutamate (by transamination of α-ketoglutarate with GABA). Thus, *de novo* synthesis of α-ketoglutarate by anaplerosis is not required to replenish the GABA catabolized in astroglia. However, GABA catabolism requires an equivalent flow of acetyl-CoA from glucose-derived pyruvate or other substrates (e.g., acetate) to maintain the availability of α-ketoglutarate for transamination with GABA. The glutamate generated by transamination (GABA-T) can be converted to glutamine by GS and is released for uptake and replenishment of glutamate and GABA in neurons. In contrast to GABA, which requires oxidative metabolism in the TCA cycle, much of the glutamate removed by glia (~70–80%) following release from glutamatergic neurons is converted directly to glutamine. A schematic of GABA and glutamate neurotransmitter cycling between GABAergic and glutamatergic neurons and astroglia is shown in Figure 2.

Although glutamine has been shown to be a major glial precursor of GABA *in vivo* and *in vitro*, studies of isolated nerve terminal preparations indicate that glial TCA cycle intermediates (e.g., α-ketoglutarate or malate) might also contribute to GABA synthesis. In addition, the role of phosphate-activated glutaminase in GABA synthesis has been challenged; immunohistochemical findings *in situ* have shown that many neocortical interneurons which co-express certain peptides do not stain positively for PAG, suggesting that a substrate other than glutamine (e.g., α-ketoglutarate) might be utilized for GABA synthesis in these neurons.

GABA Synthesis from Extracellular Glutamate

In addition to glial precursors, extracellular glutamate can enter some GABA neurons through the EAAC1 glutamate transporter, where it can be used in GABA synthesis. This pathway has been observed in the hippocampus and thalamus but not in the cortex. Regional differences in GABA precursors might be related to the apparent heterogeneity seen in PAG immunostaining for different populations of GABAergic neurons. For interneurons using extracellular glutamate in GABA synthesis, glial precursors must ultimately replenish this glutamate in an extended loop, whereby glutamine is redirected first through the glutamatergic neuron and released as glutamate before uptake and conversion to GABA.

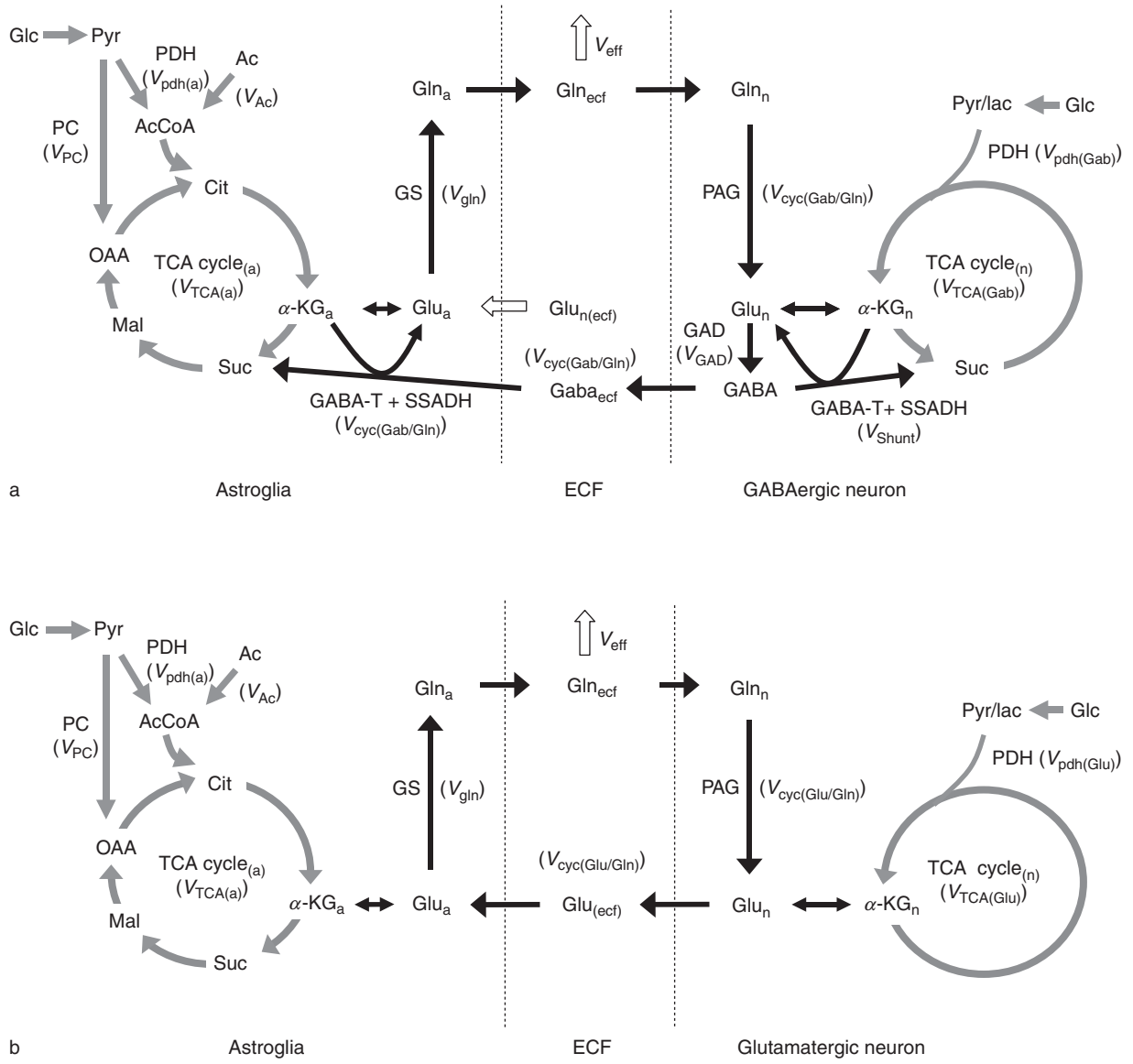

Figure 2 Depiction of GABA and glutamate cycling with glutamine between GABAergic and glutamatergic neurons with astroglia via extracellular fluid (ECF). (a) GABAergic neuron and astroglial trafficking in the glutamate/GABA–glutamine cycle. (b) Glutamatergic neuron and astroglial trafficking in the glutamate–glutamine cycle. Enzymes and their catalyzed fluxes: Ac, acetate; AcCoA, acetyl-coenzyme A; αKG, α-ketoglutarate; Cit, citrate; GAD, glutamic acid decarboxylase; Glc, glucose; Gln, glutamine; Glu, glutamic acid; GS, glutamine synthetase; lac, lactate; Mal, malate; OAA, oxaloacetic acid; PAG, phosphate-activated glutaminase; PC, pyruvate carboxylase; PDH, pyruvate dehydrogenase; Pyr, pyruvate; Suc, succinate; V_{GAD}, rate of GABA synthesis; V_{shunt}, rate of GABA shunt in GABAergic neurons; $V_{cyc(Gab/Gln)}$, rate of glutamate/GABA–glutamine cycling; V_{gln}, rate of glutamine synthesis; $V_{eff,}$ rate of glutamine efflux; $V_{pdh(Gab)}$ and $V_{pdh(a)}$, rate of acetyl-CoA formation from pyruvate in GABAergic neurons and astroglia, respectively; V_{Ac}, rate of acetyl-CoA formation from acetate.

GABA Synthesis from Polyamines

Alternative but quantitatively minor pathways of GABA synthesis from putrescine and other polyamines have been described, but these pathways account for only a small fraction (≤1%) of synthesis in the mature brain. Polyamines may play important but largely unknown roles in GABA metabolism in the developing brain and retina. Putrescine and other polyamines increase in response to stress, but a potential link to GABA synthesis has not been reported.

Nitrogen Homeostasis in Brain Glutamate and GABA Metabolism

The operation of the glutamate/GABA–glutamine cycle requires the efficient transfer of nitrogen as well as

carbon between neurons and astroglia. Nitrogen fixed as glutamine and released from astroglia for the synthesis of glutamate and GABA in neurons must ultimately return to the glia to permit continuous operation of the neurotransmitter cycles. Ammonia is freely diffusible through cell membranes, and the ammonia generated by PAG in the neurons could provide, in principle, the amide-N needed for glutamine synthesis from glutamate, part of which is acquired through uptake of extracellular glutamate or is produced by transamination with GABA. However, astroglia also undergo significant *de novo* synthesis of glutamate (anaplerosis) to replace glutamate and GABA lost by oxidation and diffusion, which may approach ~18–30% of total glutamine synthesis, thus requiring a source of amine-N. Glutamate dehydrogenase (GDH; EC 1.4.1.2) is highly expressed in astroglia but is not believed to be sufficiently active in the direction of glutamate formation under normal physiological conditions. Recent studies suggest instead that amine-N for *de novo* glutamate (and glutamine) synthesis from α-ketoglutarate occurs by transamination with alanine or a branched-chain amino acid (BCAA), such as leucine. Two different nitrogen carrier cycles have been proposed – one involving an exchange of alanine for glutamine in the glia with the product of transamination, pyruvate, returning to the neuron as lactate (alanine/lactate shuttle), and the other involving an exchange of a BCAA (leucine) for glutamine with the product of transamination, branched-chain keto acid (α-ketoisocaproic), returning to the neuron. Alanine and leucine are regenerated in the neurons by transamination with glutamate, the latter formed by GDH and the ammonia from the PAG reaction. Both neurons and astroglia express the necessary transaminases needed to support the operation of these cycles, alanine aminotransaminase (AAT; EC 2.6.1.2) and branched-chain-amino-acid transaminase (BCAT; EC 2.6.1.42). Distinct cytosolic (neuronal) and mitochondrial (astroglial) isoforms of BCAT have been found. Studies of the BCAAs and their transaminases, together with findings of significant ^{15}N labeling of brain glutamate in rodents during intravenous infusion of [^{15}N]leucine *in vivo,* indicate that as much as 25% of brain glutamate nitrogen may be derived from leucine. While studies of nitrogen shuttling between neurons and astroglia have focused mainly on glutamatergic neurons, leucine is likely to be a source of glutamate nitrogen in some GABAergic neurons. GABAergic Purkinje neurons in the cerebellum and basket neurons in the hippocampus express BCATc in their cell bodies. In Purkinje neurons, PAG expression is low, suggesting that GABA synthesis involves substrates other than glutamine. In these neurons and others potentially not expressing PAG, leucine transamination of α-ketoglutarate derived from astroglia during GABA catabolism or *de novo* synthesis could supply glutamate needed for GABA synthesis. Whether the GABA synthetic pathways operating in the many different classes of interneurons (e.g., chandelier, double bouquet, Martinotti neurons) differ from one another is largely unknown, but such differences could play an important role in their function.

Relationship of Cytoplasmic and Vesicular GABA to Different Release Pathways

Studies of cultured neocortical neurons enriched in the GABAergic phenotype demonstrate the presence of cytoplasmic and vesicular GABA pools; one is released by agonists of the glutamate *N*-methyl-D-aspartate (NMDA) receptor and is calcium independent (cytoplasmic GABA), and the other is released by 55 mM potassium and is calcium dependent (vesicular GABA). The vesicular and cytoplasmic GABA produced from glutamine can be distinguished metabolically when neurons are depolarized in a manner related to the pathway of GABA release. Vesicular GABA produced from glutamine involves metabolic processing in the TCA cycle and equilibration of the glutamate carbon skeleton with endogenous mitochondrial glutamate. In contrast, glutamate produced directly from glutamine appears to be a better precursor of GABA released through reversal of membrane transporters (i.e., cytoplasmic GABA). Neuronal firing patterns may also play an important role in the mode of GABA release and thus the functional expression of the GAD isoforms, as discussed in the following section.

Relationship of the GAD Isoforms to Cytoplasmic and Vesicular GABA Synthesis

GAD is expressed as two major isoforms, of 65 and 67 kDa, which are the products of two different genes. The two isoforms of GAD are distributed differently within GABAergic neurons: GAD_{67} is present throughout the cytoplasm, with more in the cell bodies, whereas GAD_{65} is highly enriched in axon terminals and is associated with synaptic vesicles. GAD_{67} appears to play the major role in basal GABA synthesis, and knockout of the gene for this isoform results in a major loss of tissue GABA. GAD_{67} expression is regulated both by transcriptional and by posttranscriptional mechanisms, whereas GAD_{65} expression is regulated at the transcriptional level and by short-term (kinetic) mechanisms. GAD_{65} knockout mice display electrophysiological features consistent with reduced vesicular release of GABA. GAD_{65} binding to synaptic vesicles was shown recently to involve a multiprotein

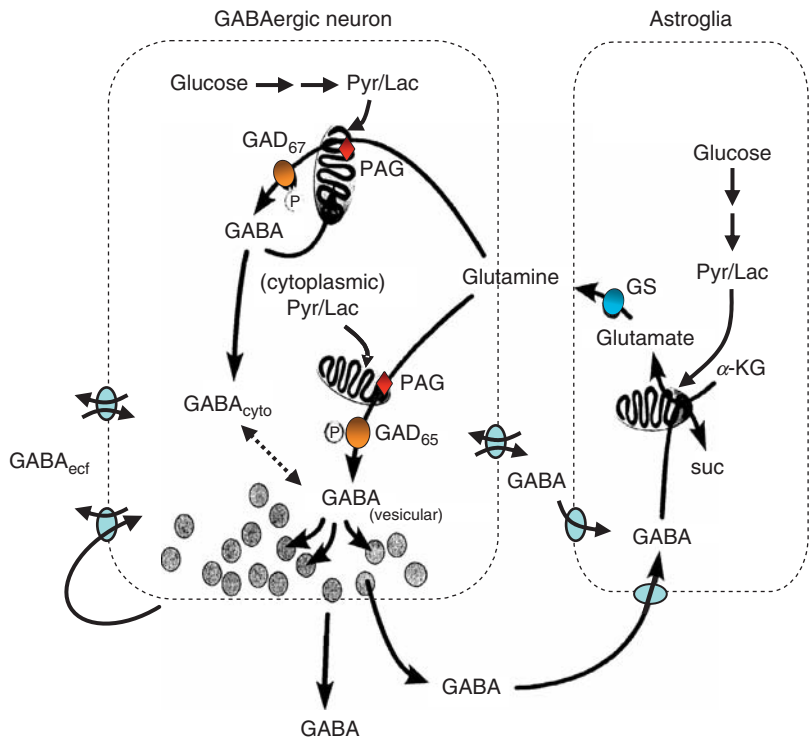

Figure 3 An idealized depiction of the relationships between cytoplasmic and vesicular γ-aminobutyric acid (GABA) synthesis, GABA release pathways, and glutamic acid decarboxylase (GAD) isoforms. Cytoplasmic GABA ($GABA_{cyto}$) synthesis is provided mainly by GAD_{67}, and this pool/isoform is associated with the majority of neuronal GABA shunt flux. Forward and reverse movements of GABA through GATs lead to uptake or release of GABA to the extracellular fluid ($GABA_{ecf}$). Vesicular GABA synthesis is catalyzed by GAD_{65}, and GABA not returned to the terminals by reuptake is metabolized in the astroglia to succinate (suc) and oxidized in the mitochondrial TCA. Transamination of GABA in the glia by GABA-T generates glutamate from α-ketoglutarate (α-KG), which is used in the synthesis of glutamine by glutamine synthetase (GS). Phosphate-activated glutaminase (PAG) provides glutamate precursors for GABA synthesis. In some GABA neurons not expressing PAG, α-ketoglutarate could serve as a GABA precursor following its uptake and transamination to glutamate in the terminals, possibly by leucine or alanine.

complex of vesicle-associated proteins, the vesicular $GABA/H^+$ antiporter and H^+-ATPase, and Ca^{2+}/calmodulin protein kinase II, which together facilitate vesicular loading of GABA driven by the hydrogen ion gradient. An idealized depiction of the relationship between the two GAD isoforms and the synthesis of the vesicular and cytoplasmic GABA pools is shown in Figure 3.

GAD_{65} is more strongly regulated by the cofactor, pyridoxal phosphate, than is GAD_{67}. The majority of GAD in brain is present as inactive apoenzyme (GAD without bound cofactor), which serves as a reservoir of inactive GAD that can be activated when demand for GABA increases. GAD_{65} binds pyridoxal phosphate less tightly than does GAD_{67} and accounts for the majority of the apoGAD in rat brain. In contrast, pyridoxal phosphate is more tightly bound to GAD_{67}, which explains the importance of this isoform in basal synthesis. The interconversion between pyridoxal phosphate bound (active) and free (inactive) GAD involves a complex cycle of reactions catalyzed by GAD. This cycle is strongly regulated by nucleoside triphosphates (ATP) and inorganic phosphate, which could serve as important regulators to increase GABA synthesis in response to increased demand. The activities of both GAD isoforms also appear to be regulated in opposing directions by protein phosphorylation.

Inherited Disorders of GABA Metabolism

Rare inherited disorders of GABA metabolism have been described involving deficiencies in enzymes in the pathways of GABA catabolism (GABA-T, SSADH, and serum carnosinase) and a cofactor (pyridoxine) required in GABA synthesis. High levels of GABA accumulate in brain and cerebrospinal fluid of individuals deficient in GABA-T and SSADH. In SSADH deficiency, the buildup of the product of GABA-T, succinic semialdehyde, promotes increased synthesis and accumulation of GHB. Homocarnosine levels are elevated in brain and cerebrospinal fluid of individuals with homocarnosinosis, a condition resulting from deficiency of serum carnosinase, a dipeptidase which hydrolyzes homocarnosine and carnosine. Pyridoxine-dependent epilepsy, a rare autosomal recessive inherited

trait characterized by intractable seizures, is treated by administration of pyridoxine (vitamin B6) at pharmacologic doses. Although pyridoxine is a required cofactor of GABA synthesis, and reduced pyridoxine interaction with GAD is suspected in this disorder, a definitive mechanism has not been established.

See also: GABA-A Receptors: Developmental Roles; GABA-A Receptor Synaptic Functions; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-A Receptors and Disease; GABA-B Receptors: Molecular Biology and Pharmacology; GABA-B Receptor Function; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism.

Further Reading

Asada H, Kawamura Y, Maruyama K, et al. (1996) Mice lacking the 65 kDa isoform of glutamic acid decarboxylase (GAD_{65}) maintain normal levels of GAD_{67} and GABA in their brains but are susceptible to seizures. *Biochemical and Biophysical Research Communications* 229: 891–895.

Asada H, Kawamura Y, Maruyama K, et al. (1997) Cleft palate and decreased brain gamma-aminobutyric acid in mice lacking the 67-kDa isoform of glutamic acid decarboxylase. *Proceedings of the National Academy of Sciences of the United States of America* 94: 6496–6499.

Bak LK, Schousboe A, and Waagepetersen HS (2006) The glutamate/GABA-glutamine cycle: Aspects of transport, neurotransmitter homeostasis and ammonia transfer. *Journal of Neurochemistry* 98: 641–653.

Battaglioli G, Liu H, and Martin DL (2003) Kinetic differences between the isoforms of glutamate decarboxylase: Implications for the regulation of GABA synthesis. *Journal of Neurochemistry* 86: 879–887.

Beckman ML, Bernstein EM, and Quick MW (1999) Multiple G protein-coupled receptors initiate protein kinase C redistribution of GABA transporters in hippocampal neurons. *Journal of Neuroscience* 19: RC9 1–6.

Belhage B, Hansen GH, and Schousboe A (1993) Depolarization by K^+ and glutamate activates different neurotransmitter release mechanisms in GABAergic neurons: Vesicular versus non-vesicular release of GABA. *Neuroscience* 54: 1019–1034.

Berl S and Clarke DD (1969) Compartmentation of amino acid metabolism. In: Lajtha A (ed.) *Handbook of Neurochemistry,* vol. 2, pp. 447–472. New York: Plenum.

Chaudhry FA, Schmitz D, Reimer RJ, et al. (2002) Glutamine uptake by neurons: Interaction of protons with system a transporters. *Journal of Neuroscience* 22: 62–72.

Conti F, Minelli A, and Melone M (2004) GABA transporters in the mammalian cerebral cortex: Localization, development and pathological implications. *Brain Research: Brain Research Reviews* 45: 196–212.

Fonnum F and Walberg F (1973) An estimation of the concentration of γ-aminobutyric acid and glutamate decarboxylase in the inhibitory Purkinje axon terminals in the cat. *Brain Research* 54: 115–127.

Hutson SM, Berkich D, Drown P, et al. (1998) Role of branched-chain aminotransferase isoenzymes and gabapentin in neurotransmitter metabolism. *Journal of Neurochemistry* 71: 863–874.

Jin H, Wu H, Osterhaus G, et al. (2003) Demonstration of functional coupling between gamma-aminobutyric acid (GABA) synthesis and vesicular GABA transport into synaptic vesicles. *Proceedings of the National Academy of Sciences of the United States America* 100: 4293–4298.

Kanamori K, Ross BD, and Kondrat RW (1998) Rate of glutamate synthesis from leucine in rat brain measured *in vivo* by ^{15}N NMR. *Journal of Neurochemistry* 70: 1304–1315.

Kaufman DL, Houser CR, and Tobin AJ (1991) Two forms of the gamma-aminobutyric acid synthetic enzyme glutamate decarboxylase have distinct intraneuronal distributions and cofactor interactions. *Journal of Neurochemistry* 56: 720–723.

Lieth E, LaNoue KF, Berkich DA, et al. (2001) Nitrogen shuttling between neurons and glial cells during glutamate synthesis. *Journal of Neurochemistry* 76: 1712–1723.

Martin DL and Tobin AJ (2000) Mechanisms controlling GABA synthesis and degradation in the brain. In: Martin DL and Olsen RW (eds.) *GABA in the Nervous System: The View at Fifty Years*, pp. 25–41. Philadelphia: Lippincott Williams & Wilkins.

Patel AB, de Graaf RA, Martin DL, et al. (2005) Evidence that GAD_{65} mediates increased GABA synthesis during intense neuronal activity *in vivo*. *Journal of Neurochemistry* 97: 385–396.

Patel AB, de Graaf RA, Mason GF, et al. (2005) The contribution of GABA to glutamate/glutamine cycling and energy metabolism in the rat cortex *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* 102: 5588–5593.

Pearl PL and Gibson KM (2004) Clinical aspects of the disorders of GABA metabolism in children. *Current Opinion in Neurology* 17: 107–113.

Richerson GB and Wu Y (2003) Dynamic equilibrium of neurotransmitter transporters: Not just for reuptake anymore. *Journal of Neurophysiology* 90: 1363–1374.

Seiler N (1980) On the role of GABA in vertebrate polyamine metabolism. *Physiological Chemistry and Physics* 12: 411–429.

Sepkuty JP, Cohen AS, Eccles C, et al. (2002) A neuronal glutamate transporter contributes to neurotransmitter GABA synthesis and epilepsy. *Journal of Neuroscience* 22: 6372–6379.

Shank RP, Bennett GS, Freytag SO, et al. (1985) Pyruvate carboxylase: An astrocyte-specific enzyme implicated in the replenishment of amino acid neurotransmitter pools. *Brain Research* 329: 364–367.

Soghomonian JJ and Martin DL (1998) Two isoforms of glutamate decarboxylase: Why? *Trends in Pharmacological Science* 19: 500–505.

Waagepetersen HS, Sonnewald U, Gegelashvili G, et al. (2001) Metabolic distinction between vesicular and cytosolic GABA in cultured GABAergic neurons using ^{13}C magnetic resonance spectroscopy. *Journal of Neuroscience Research* 63: 347–355.

Wei J, Davis KM, Wu H, et al. (2004) Protein phosphorylation of human brain glutamic acid decarboxylase GAD_{65} and GAD_{67} and its physiological implications. *Biochemistry* 43: 6182–6189.

GABA$_A$ Receptor Synaptic Functions

I Mody, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The amino acid neurotransmitter γ-aminobutyric acid (GABA) is responsible for mediating most of chemical inhibition in the central nervous system (CNS). The ionotropic GABA receptors (GABA$_A$Rs) are members of the cysteine-loop ligand-gated ion channel family and form a Cl^-- and HCO_3^--permeable ion pore assembled from five (heteropentameric) subunits selected from the following subunits: α1–6, β1–3, γ1–3, δ, ϵ, θ1–3, π, and ρ1–3. Hundreds of thousands of different combinations are possible, yet no more than a few dozen receptor combinations are likely to exist in the mammalian brain. This is thought to result in part from precise rules of subunit assembly (e.g., two α and two β subunits assemble with either one γ or one δ subunit) and in part from certain rules of specific subunit partnerships that also appear to define their subcellular localization inside and outside of synapses. Loss-of-function mutations in some of these subunits are associated with various rare genetic forms of epilepsy in humans, indicating their importance in subduing neuronal hyperactivity. Over the years GABA$_A$Rs have become synonymous with inhibition, but we now know that depending on a variety of conditions, their activation may excite neurons or may synchronize them, resulting in physiological oscillations or pathological synchrony leading to seizures. In 1963, on behalf of the Nobel Committee upon awarding the Prize to Sir John C Eccles, Ragnar Granit stated: "If the arriving impulse is connected to excitatory synapses the response of the cell is yes, i.e., excitability increases; vice versa the inhibitory synapses make the cell respond with a no, a diminution of excitability." Forty-three years later it is clear that in response to GABAergic 'inhibition' many cells have expanded their vocabulary to 'maybe' and even to 'yes.'

GABA$_A$Rs at Synapses

In the mammalian brain GABA synapses constitute a minority. Only about every sixth synaptic terminal is GABAergic and only every fifth neuron utilizes GABA as a transmitter. Excitation clearly outnumbers inhibition. Yet, in spite of this overpowering excitatory drive, most of the time the GABAergic system manages to keep excitability in check. It does this through a spatially and temporally selective activation of the GABAergic system. The pioneering Golgi impregnation studies of Ramón y Cajal, Lorente de No, and János Szentágothai have distinguished a marked feature of cortical interneurons: they have a highly selective axonal arborization, innervating distinct spatial domains of the principal cells. GABA synapses can be found on cell bodies, dendritic shafts and branch points, spine necks, and axon initial segments, in far more diverse places than excitatory synapses. The strategic positioning of GABA synapses on the target cells can annihilate a specific input or the entire output of the cell. A good example of such selective control of a given input is the hippocampal oriens-lacunosum moleculare (OLM) cell that precisely innervates the most distal dendrites of pyramidal cells. This is precisely the region where the entorhinal cortical excitatory input lands on the principal cells. Since OLM cells are activated by massive intrahippocampal excitatory events, this spatial arrangement of their inhibitory output ensures that the entorhinal input does not have a say during this time. In contrast, the cortical chandelier or axo-axonic cells do not synapse onto a cellular region receiving a specific input. Instead, they control the output of pyramidal cells, by innervating the axon initial segment, where the ultimate action potential output of the target neuron is generated. According to the latest findings however, some of these axo-axonic inhibitory interneurons may elicit firing at the axon initial segment (see later), in stark contradiction to the function implied by their 'inhibitory' name.

The location specificity of inhibitory synapses is complemented by a highly specific assembly of the GABA$_A$Rs that depends on the type of inhibitory interneuron that forms the synapse. There is little known about the mechanisms that sustain this specificity at a given synapse over the lifetime of the neuron. Of the thousands of possible heteropentameric GABA$_A$R combinations that can be formed through assembly from nearly 20 subunits, only a few dozen combinations occur in the brain. These combinations do not appear to be randomly inserted at synapses. The types of subunits present at a given synapse tend to be specified by the GABA neuron that provides the presynaptic bouton of the synapse. The axo-axonic chandelier cells, and the pyramidal cell soma innervating cholecystokinin (CCK)-positive basket cells, are known to form synapses on predominantly α2 subunit-containing receptors. The parvalbumin-containing basket cells synapse onto α1-containing

$GABA_A$Rs, while other combinations are found at synapses made by other interneuron types. The functional significance of this highly specific arrangement is only beginning to emerge. Different GABA receptor subunit assemblies have very different kinetic and desensitization properties when studied in expression systems. Accordingly, one would expect the various synaptic currents generated by diverse $GABA_A$Rs also to be highly variable. However, paired recordings from two interneurons and their target cells usually show similar inhibitory postsynaptic current (IPSC) kinetics, in spite of known differences of $GABA_A$R assemblies at the two synapses. Even if the kinetic differences observed in expression systems do not hold at $GABA_A$Rs assembled at synapses in the brain, the different compositions and locations of $GABA_A$Rs have undoubtedly evolved to participate in widely different control functions of neuronal excitability. Insight into these functions has come from cleverly engineered mice expressing various α subunits insensitive to benzodiazepines (BZs), a class of potent allosteric $GABA_A$R modulators.

Two α, two β, and a single γ subunit assemble to form the most commonly found $GABA_A$Rs in the brain. The function of these receptors is enhanced by BZs as long as the α subunits are either $\alpha1$, $\alpha2$, $\alpha3$, or $\alpha5$, but not $\alpha4$ or $\alpha6$. There is a single amino acid in a critical position (His) that differs between the BZ-sensitive and BZ-insensitive α subunits (Arg in the latter). Mice have been genetically engineered to have BZ-insensitive $\alpha1$, $\alpha2$, $\alpha3$, or $\alpha5$ subunit-containing $GABA_A$Rs. Studies in these mice conclusively demonstrated that the $\alpha1$ subunit-containing $GABA_A$Rs are involved in the sedative–hypnotic actions of BZs, while the receptors containing $\alpha2$ subunits mediate anxiolytic effects of BZs. Does this mean that the function of the synapses predominantly containing $\alpha2$ subunits, such as those found at the axon initial segment or those apposed to the terminals of CCK-containing basket cells, are mainly related to the control of anxiety? At this time it is still too early to say, as there might be specific brain regions rich in $\alpha2$ subunits that could play a significant role in the control of anxiety. Nevertheless, it is clear that specific $GABA_A$R subunit assemblies innervated by specific interneurons play highly specialized roles in the development of neuronal circuitry and in mediating specific higher nervous system functions.

Activation of $GABA_A$Rs at Synapses – Phasic Inhibition

The $GABA_A$Rs usually found at synapses are most likely to contain $\gamma2$ subunits. GABA synapses are 'symmetrical' synapses; called such because on electron micrographs they are devoid of a wide electron-dense postsynaptic density (PSD) which would make the appearance of the synapse 'asymmetric.' Therefore, compared to the PSDs of excitatory glutamatergic synapses containing a stable assembly of over 70 different receptor, scaffold, and signaling proteins, there is a lackluster showing of postsynaptic proteins at the poorly developed PSDs of GABA synapses. Glutamate receptors at synapses associate with the subsynaptic cytoskeleton. $GABA_A$Rs do not. There are, however, some proteins that seem to play a role in the anchoring, insertion, and removal of $GABA_A$Rs at synapses. Gephyrin links up with the subsynaptic microtubule and actin cytoskeleton. It interacts with several microfilament-regulating proteins, including collybistin, profilins, Mena/VASP proteins, and dynein light chain (Dlc)/myosin-Va. However, the role of these protein–protein interactions for the shaping of GABA synapses is largely a mystery. Some GABA synapses express the dystrophin–glycoprotein complex, better known as a connecting element between the extracellular and cytoskeletal matrices of muscle cells. Other proteins are implicated in the trafficking of $GABA_A$Rs: AP2, BIG2, GABARAP, GODZ, Plic-1, and PRIP1/2. To make things more complicated, trafficking of $GABA_A$R subunits seems to depend on the phosphorylation state of the subunit. Various kinases and phosphatases are at hand, as they bind directly to the subunits or are linked to specific $GABA_A$Rs through highly specific adaptor proteins.

As molecular biologists, biochemists, and anatomists continue to pursue the means of $GABA_A$R anchoring and targeting to synapses, physiologists and biophysicist have focused on the means of activation of $GABA_A$Rs at synaptic junctions. The kinetic profile of synaptic $GABA_A$Rs consists of a moderate to low affinity for GABA, some desensitization, and a rapid activation/deactivation. The time course of GABA in the synaptic cleft has been estimated to be less than 500 μs, and the peak concentration reaching the receptors is probably in the low millimolar range. This means that at many synapses where usually no more than tens of $GABA_A$Rs are present, the number of GABA molecules is in excess of the available binding sites, thus providing for a high occupancy of the available receptors. However, this may not be the case at all GABA synapses, as in cerebellar basket/stellate cells, where synapses with hundreds of receptors and synapses with tens of receptors coexist, only the latter appear to be saturated by the released transmitter. In general, the relatively low number of postsynaptic GABA receptors compared to the amount of GABA released into the cleft renders the process regulating the number of receptors an important aspect of $GABA_A$R plasticity. The kinetic properties of the

subunits assembled at the synapses should also affect the properties of inhibition, but thus far there are consistent reports only on the speedy decay of α1 subunit-containing synaptic receptors, while the variety of the biophysical receptor properties seen in expression systems has yet to be demonstrated at synapses. Desensitization is also an important mechanism to consider for regulating the decay of IPSCs, but prolonging the synaptic GABA transient by using uptake blockers does not always prolong the duration of synaptic events as it would be expected from a significant role of desensitization. It appears, however that the duration of the GABA transient in the cleft may itself regulate the decay of synaptic events, which in turn might make it difficult to discern differences that are simply due to underlying channel kinetics. The various brain-state-altering effects of BZs that act to prolong the duration of IPSCs predominantly affecting γ2 subunit-containing synaptic receptors show the importance of regulating the decay phase of IPSCs.

Like at any central synapse, presynaptic mechanisms can also effectively determine how GABA synapses are activated. Again, the most interesting presynaptic aspect of GABA synapses is their specificity with regard to the cell type supplying the presynaptic terminal. For example, there are clear differences in the way GABA is released from parvalbumin- and CCK-containing basket cells, and this may have important functional consequences for the operation of neuronal networks. Other GABA synapses are depressed most of the time by ambient levels of endocannabinoids, and do not become active unless the presynaptic cell fires at high frequencies. Different presynaptic Ca^{2+} channels are responsible for activating GABA release in different interneurons. Moreover, there is also accumulating evidence that the intricate vesicular release machinery may be different between glutamate- and GABA-containing terminals.

Activation of Peri- and Extrasynaptic GABA Receptors: Tonic Inhibition

As is the case for many ligand-gated ion channels, synapses are not the only place $GABA_A$Rs are likely to show up on a cell's surface. But perhaps for more than any other ligand-gated channel, $GABA_A$Rs found outside the synapses are quite special. Although many different combinations of $GABA_A$Rs are scattered outside of synapses, there are only two types of combinations that have been shown to actually respond to the ambient GABA concentrations of less than a few micromoles per liter, as found in the interstitial space of neurons. These two combinations are the δ subunit-containing and the α5 subunit-containing $GABA_A$Rs. The δ subunits combine with α4, α6, and most likely other α subunits to form, together with β2 or β3 subunits, a $GABA_A$R combination that is well suited to be activated by low levels of ambient GABA. These receptors have a high affinity for GABA, they do not desensitize, and are preferentially found extra- or perisynaptically. Their modulation mainly occurs by increasing the efficacy of GABA rather than its potency, as the natural agonist is not very efficacious at these receptors. The other $GABA_A$R combination shown to be activated by ambient GABA is the α5 subunit-containing receptors. The α5 subunits are not exclusively extrasynaptic; they probably combine with other α subunits in synaptic receptors, but in the cells where they are present they readily mediate tonic inhibition. The tonic inhibition, when present, is a significant force in controlling the excitability of the neurons. Charge movement through tonically active GABA channels outweighs 3:1 to 5:1 the charge going through the synaptic receptors, even when the frequency of the synaptic activity is quite high (40–60 Hz). Such a tonically active conductance globally is well suited to regulate the input–output function of the cells.

The δ subunit-containing $GABA_A$Rs mediating tonic inhibition are also an important site of action for various endogenous and exogenous allosteric modulators. Neurosteroids, brain-synthesized metabolites of various sex and stress steroid hormones, specifically act upon these receptors to enhance their activity. Ethanol, at concentrations relevant to impairing sobriety, also appears to target the δ subunit-containing $GABA_A$Rs, which may justify them lately being referred to as the 'ethanol receptor.' The receptors mediating tonic inhibition show a remarkable plasticity, both during physiological (ovarian cycle) and pathological (epilepsy) alterations in neuronal excitability. Selectively eliminating tonic inhibition in hippocampal CA1 pyramidal cells facilitates certain forms of learning, presumably by increasing the sensitivity of pyramidal cells to incoming stimuli. These recent developments in the physiology and pharmacology of extrasynaptic tonically active $GABA_A$Rs will lead to exciting insights into their complex roles in the control of neuronal excitability.

Excitation and Synchrony through 'Inhibitory' Synapses

There is nothing mysterious about 'inhibitory' synapses extending their vocabulary to 'maybe' and 'yes.' The relationship between the time-averaged membrane potential of the neuron and the GABA

equilibrium potential, which is a function of the equilibrium potentials of the two permeant ions, Cl^- and HCO_3, determines whether activation of $GABA_A$Rs will depolarize or hyperpolarize a cell. In fact, 'yes' is the first word uttered by an activated $GABA_A$R. Early on during brain development, when cells are loaded with Cl^-, the activation of $GABA_A$R channels leads to efflux of negatively charged Cl^- ions, thus depolarizing the cell. This makes GABA one of the first depolarizing signaling molecules during development. The depolarization helps bring the membrane potential to the threshold of activation of voltage-gated Ca^{2+} entry, a cation that, once inside the cells, is critical for the activation of a myriad of important developmental signaling cascades. Eventually, later during development, Cl^- is extruded from the cells by various transport mechanisms. Inhibition can thus start to learn to say 'no.' Yet, depolarizing $GABA_A$R-mediated events continue to play an important role in the adult nervous system. The survival of newly generated neurons in the mammalian brain depends largely on the activation of depolarizing, tonically active $GABA_A$Rs. Local peaks in the GABA reversal potential can trigger depolarizing inhibitory postsynaptic potentials (IPSPs) that can passively propagate along the neuron to elevate excitability, or can outright trigger action potentials at the axon initial segment, as we have recently learned from the potentially 'excitatory' function of axo-axonic cells in the neocortex.

Regulated by several exchangers and transporters, the GABA reversal potential can itself become a dynamic force in the control of neuronal excitability. One of the key players is a K^+/Cl^- cotransporter called KCC2. This molecule is critical in extruding Cl^- during development, but reducing its function during spike timing-dependent plasticity of GABA synapses, high-frequency firing of the cells, or neuronal injury leads to an intracellular accumulation of Cl^- and thus the conversion of a hyperpolarizing event into one that will depolarize the affected neuron. An altered KCC2 function is most likely at work to cause the pathological alteration of the GABA reversal potential in subicular pyramidal neurons recorded in tissue samples obtained from temporal lobe epilepsy patients, and causes the conversion to a GABA-induced depolarization of spinal cord neurons during neuropathic pain. Even if depolarizing GABA-mediated events do not push the membrane potential all the way past the threshold of activation of voltage-dependent Na^+ or Ca^{2+} conductances, such events can passively propagate through the cell, thus increasing excitability and promoting neuronal synchrony.

Depolarizing neurons is not the only means by which activation of inhibitory synapses can promote synchrony. If the neurons are equipped with a hyperpolarization-activated excitatory conductance (e.g., I_h) or a low-threshold (T-type) Ca^{2+} current, as most neurons are, then a hyperpolarization induced in several neurons at the same time can lead to a rebound excitation following the cessation of the hyperpolarizing event. Hundreds or thousands of principal cells innervated by a single inhibitory interneuron, or by a few interneurons activated in unison, can thus synchronize their firing without requiring any synchronous excitatory drive. This type of synchronization may specifically spring into action when diffusely acting neurotransmitters, such as acetylcholine (ACh) or glutamate, enhance GABA release by acting at once on the terminals of several inhibitory cells through actions at nicotinic ACh or *N*-methyl-D-aspartate (NMDA) receptors.

In summary, phasic and tonic GABAergic inhibitions exert a substantial control over neuronal excitability. The incredibly diverse GABAergic cells of the brain are the real puppet masters of the principal cells, by controlling their activity both at the level of different single neuronal compartments and at the level of intricately connected neuronal networks. For certain, the diversity and specificity of $GABA_A$Rs, GABAergic neurons, and their functions will keep neuroscientists busy for a while. It may have taken some time, but now we know that it is much easier for the 'inhibitory' GABAergic system to say 'maybe' and 'yes' than it is for excitatory glutamate synapses to learn to say 'no.'

See also: GABA Synthesis and Metabolism; GABA-A Receptors: Developmental Roles; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-A Receptors and Disease; GABA-B Receptors: Molecular Biology and Pharmacology; GABA-B Receptor Function; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism.

Further Reading

Ben-Ari Y (2002) Excitatory actions of GABA during development: The nature of the nurture. *Nature Reviews in Neuroscience* 3: 728–739.

Buzsáki G (2006) *Rhythms of the Brain*. New York: Oxford University Press.

Buzsáki G, Geisler C, Henze DA, et al. (2004) Interneuron diversity series: Circuit complexity and axon wiring economy of cortical interneurons. *Trends in Neuroscience* 27: 186–193.

Chen ZW and Olsen RW (2007) GABA receptor associated proteins: A key factor regulating GABA receptor function. *Journal of Neurochemistry* 100(2): 279–294.

Cobb SR, Buhl EH, Halasy K, et al. (1995) Synchronization of neuronal activity in hippocampus by individual GABAergic interneurons. *Nature* 378: 75–78.

Cohen I, Navarro V, Clemenceau S, et al. (2002) On the origin of interictal activity in human temporal lobe epilepsy *in vitro*. *Science* 298: 1418–1421.

Coull JA, Beggs S, Boudreau D, et al. (2005) BDNF from microglia causes the shift in neuronal anion gradient underlying neuropathic pain. *Nature* 438: 1017–1021.

Farrant M and Nusser Z (2005) Variations on an inhibitory theme: Phasic and tonic activation of $GABA_A$ receptors. *Nature Reviews Neuroscience* 6: 215–229.

Fiumelli H, Cancedda L, and Poo MM (2005) Modulation of GABAergic transmission by activity via postsynaptic Ca^{2+}-dependent regulation of KCC2 function. *Neuron* 48: 773–786.

Freund TF and Buzsáki G (1996) Interneurons of the hippocampus. *Hippocampus* 6: 347–470.

Gaiarsa JL, Caillard O, and Ben-Ari Y (2002) Long-term plasticity at GABAergic and glycinergic synapses: Mechanisms and functional significance. *Trends in Neuroscience* 25: 564–570.

Hanchar HJ, Wallner M, and Olsen RW (2004) Alcohol effects on gamma-aminobutyric acid type A receptors: Are extrasynaptic receptors the answer? *Life Sciences* 76: 1–8.

Hefft S and Jonas P (2005) Asynchronous GABA release generates long-lasting inhibition at a hippocampal interneuron–principal neuron synapse. *Nature Neuroscience* 8: 1319–1328.

Hevers W and Lüddens H (1998) The diversity of $GABA_A$ receptors. Pharmacological and electrophysiological properties of $GABA_A$ channel subtypes. *Molecular Neurobiology* 18: 35–86.

Lien CC, Mu Y, Vargas-Caballero M, et al. (2006) Visual stimuli-induced LTD of GABAergic synapses mediated by presynaptic NMDA receptors. *Nature Neuroscience* 9: 372–380.

Lüscher B and Fritschy JM (2001) Subcellular localization and regulation of $GABA_A$ receptors and associated proteins. *International Review of Neurobiology* 48: 31–64.

Lüscher B and Keller CA (2004) Regulation of $GABA_A$ receptor trafficking, channel activity, and functional plasticity of inhibitory synapses. *Pharmacology and Therapeutics* 102: 195–221.

Macdonald RL, Gallagher MJ, Feng HJ, et al. (2004) $GABA_A$ receptor epilepsy mutations. *Biochemical Pharmacology* 68: 1497–1506.

Maguire JL, Stell BM, Rafizadeh M, et al. (2005) Ovarian cycle-linked changes in $GABA_A$ receptors mediating tonic inhibition alter seizure susceptibility and anxiety. *Nature Neuroscience* 8: 797–804.

Mody I and Pearce RA (2004) Diversity of inhibitory neurotransmission through GABA(A) receptors. *Trends in Neuroscience* 27(9): 569–575.

Mozrzymas JW (2004) Dynamism of $GABA_A$ receptor activation shapes the 'personality' of inhibitory synapses. *Neuropharmacology* 47: 945–960.

Rivera C, Voipio J, Thomas-Crusells J, et al. (2004) Mechanism of activity-dependent downregulation of the neuron-specific K–Cl cotransporter KCC2. *Journal of Neuroscience* 24: 4683–4691.

Rudolph U and Möhler H (2004) Analysis of $GABA_A$ receptor function and dissection of the pharmacology of benzodiazepines and general anesthetics through mouse genetics. *Annual Reviews of Pharmacology and Toxicology* 44: 475–498.

Sieghart W and Sperk G (2002) Subunit composition, distribution and function of $GABA_A$ receptor subtypes. *Current Topics in Medicinal Chemistry* 2: 795–816.

Somogyi P and Klausberger T (2005) Defined types of cortical interneurone structure space and spike timing in the hippocampus. *Journal of Physiology* 562(1): 9–26.

Somogyi P, Tamas G, Lujan R, et al. (1998) Salient features of synaptic organisation in the cerebral cortex. *Brain Research: Brain Research Reviews* 26: 113–135.

Soltesz I (2005) *Diversity in the Neuronal Machine: Order and Variability in Interneuronal Microcircuits*. New York: Oxford University Press.

Stell BM, Brickley SG, Tang CY, et al. (2003) Neuroactive steroids reduce neuronal excitability by selectively enhancing tonic inhibition mediated by delta subunit-containing $GABA_A$ receptors. *Proceeding of the National Academy of Sciences United States of America* 100: 14439–14444.

Szabadics J, Varga C, Molnar G, et al. (2006) Excitatory effect of GABAergic axo-axonic cells in cortical microcircuits. *Science* 311: 233–235.

Watts J and Thomson AM (2005) Excitatory and inhibitory connections show selectivity in the neocortex. *Journal of Physiology* 562: 89–97.

Whiting PJ (2003) The $GABA_A$ receptor gene family: New opportunities for drug development. *Current Opinion in Drug Discovery and Development* 6: 648–657.

$GABA_A$ Receptors and Disease

K Staley, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

γ-Aminobutyric acid (GABA) is the principal inhibitory neurotransmitter in the mature nervous system. It is produced by decarboxylation of glutamate, a reaction for which vitamin B_6 is a necessary cofactor. GABA is released primarily by calcium-dependent fusion of neurotransmitter vesicles with the cytoplasmic membrane of the presynaptic terminal. Binding of GABA to postsynaptic $GABA_A$ receptors opens anion-permeable channels. The main permeant anions are chloride, which flows into the neuron, and bicarbonate, which flows out. Chloride is more plentiful and more permeant, so under most circumstances the net effect of $GABA_A$ receptor activation is hyperpolarization due to intracellular accumulation of negatively charged chloride ions. $GABA_B$ receptors are G-protein-linked receptors that are located both pre- and postsynaptically. $GABA_B$ receptors inhibit calcium channels and activate G-protein-linked inwardly rectifying potassium channels whose activation kinetics are one to two orders of magnitude slower than the $GABA_A$ receptor-operated ion channels. Because GABA inhibits neurons, the principal disorders associated with GABA dysfunction are related to pathological excitation. The prototypical disorder of neuronal overexcitation is epilepsy. Recent research has clearly linked GABA dysfunction to psychiatric, sleep, pain, and movement disorders as well.

Epilepsy

Epilepsy is the propensity for episodes of pathologically synchronous neural network activation called seizures. The most commonly accepted etiology for epilepsy is an imbalance of excitation and inhibition. Although the imbalance idea is an oversimplification, it was conceived on the basis of a large body of experimental data indicating that reductions in inhibition could induce acute epileptiform activity in otherwise normal preparations. The convulsants penicillin and pentylene tetrazol, which are both antagonists of the $GABA_A$ receptor, have been used as experimental convulsants for many decades. More recently, links between dysfunctions of specific aspects of the $GABA_A$ system and epilepsy have been established. These dysfunctions are grouped into presynaptic, postsynaptic, and ion flux problems, and therapies have been developed for each category.

Presynaptic Disorders

Reduced availability of GABA is thought to form the basis for a form of epilepsy manifested in pyridoxine-dependent seizures. In this rare disorder, newborn infants present with recurrent seizures that respond poorly to anticonvulsants but resolve rapidly with the administration of pyridoxine (vitamin B_6) or pyridoxal phosphate, the active form of the vitamin. A similar syndrome occurs in anyone exposed to high concentrations of the antitubercular drug isoniazid, which is also a pyridoxine antagonist. As mentioned above, pyridoxine is a necessary coenzyme for glutamic acid decarboxylase, which catalyzes the conversion of glutamate, the principal excitatory amino acid, into GABA, the principal inhibitory amino acid. This shift between glutamate and GABA is perhaps the clearest embodiment of the concept of an imbalance between excitation and inhibition as the etiology of epilepsy. Glutamate and GABA levels in the cerebrospinal fluid correspond to seizure control, although it may be possible in the future to assay these levels more directly (and less painfully) with magnetic resonance spectroscopy of GABA and glutamate/glutamine levels in the brain.

Therapeutically, several anticonvulsant drugs have been designed as GABA analogs or to increase the availability of GABA. Tiagabine blocks the GABA transporter GAT-1 that is responsible for much of the reuptake of GABA after synaptic release. Tiagabine should therefore increase the concentration of extracellular GABA, an important determinant of the activity of the extrasynaptic $GABA_A$ receptors that mediate tonic inhibition. Two drugs, vigabatrin and pregabalin, were designed as GABA analogs; recent studies have indicated that their anticonvulsant effects are more likely to be mediated via effects on voltage-gated calcium channels.

Postsynaptic Disorders

Subunit mutations A variety of genetically determined epilepsies have been discovered to be due to mutations in genes encoding subunits of the $GABA_A$ receptor. $GABA_A$ receptors are pentameric polypeptides that are typically composed of two α, two β, and either a γ or a δ subunit. Each of these Greek letters represents a family of related subunits, the members of which are designated by numerical subscripts, such as γ_2.

Generalized epilepsy with febrile seizures plus (GEFS+) is a vague name for an epilepsy syndrome, but that vagueness accurately reflects the clinical heterogeneity of the syndrome itself. GEFS+ is an

autosomal dominant disorder in which affected family members may have only febrile seizures. Febrile seizures are triggered in young children when the body temperature rises at the beginning of a fever. Other affected family members may manifest absence seizures, in which brief behavioral arrests are associated with generalized 3 Hz spike-wave discharges of several seconds' duration. However, affected family members may have more-severe seizure disorders, the worst of which is severe myoconic epilepsy of infancy (SMEI). In SMEI seizures also begin as febrile seizures but progress to include afebrile myoclonic and generalized tonic-clonic seizures that are associated with ataxia and developmental delays that are usually profound. These diverse epilepsy syndromes share the same mutations within a family. In keeping with the concept of epilepsy's arising from a change in the balance of excitation and inhibition, the most frequently isolated mutations are in either sodium channel subunits or the γ_2 and δ subunits of the $GABA_A$ receptor. Three laboratories have investigated the γ_2 mutation, R43Q, and uncovered a variety of possible functional implications, including more-rapid desensitization, a reduction in benzodiazepine modulation, and a reduction in the peak current amplitude. The variety of functional defects, while somewhat disconcerting, is common in these types of investigation and arises from a combination of differences in expression systems and the tremendous impact of posttranslational modification on the functional characteristics of ion channels. This issue is also reflected in the heterogeneity of epilepsy phenotypes for a given mutation and demonstrates the strong interaction between mutations in a gene for a GABA receptor subunit, the functional implications of that mutation after posttranslational processing, and the expression of that function in the setting of all the other characteristics of the brain that may either compensate for or exacerbate the underlying functional deficit in GABA signaling. Identification of the factors that modify the functional expression of mutations in GABA signaling proteins may make possible more-accurate predictions of the severity of disease in affected individuals.

Juvenile myoclonic epilepsy (JME) is manifest in adolescence and adulthood by early morning isolated jerks of the limbs and trunk (myoclonus) as well as tonic-clonic convulsions. The electroencephalogram demonstrates bilateral 6 Hz spikes and slow wave discharges. JME was recently demonstrated to be due to a mutation in the $GABA_A$ receptor. Identified mutations include the A322D mutation in the α_1 subunit of the $GABA_A$ receptor, which strongly reduces $GABA_A$ receptor-mediated currents via reduced expression of the receptor on the cytoplasmic membrane in the human embryonic kidney cell 293T expression system.

Subunit expression Extrasynaptic $GABA_A$ receptors mediate the tonic GABA conductance, in which ambient extracellular GABA concentrations activate high-affinity, slowly desensitizing $GABA_A$ receptors that usually contain a delta subunit. These $GABA_A$ receptors are uniquely sensitive to alcohol and neurosteroids. Recently, it has been discovered that changing levels of endogenous neurosteroids occurring during the estrous cycle may contribute to cyclic fluctuations in seizure susceptibility. In humans, the premenstrual reduction in seizure threshold is referred to as catamenial epilepsy. Rather than a direct neurosteroid modulation of $GABA_A$ receptors, the estrous cycle effects were shown to be due at least in part to alterations in the expression of delta subunit-containing $GABA_A$ receptors. The changing level of delta subunit-containing $GABA_A$ receptors alters the level of tonic inhibition, which contributes to the resting conductance of neurons and strongly modulates neuronal excitability and seizure threshold.

Therapeutically, anticonvulsants that increase the postsynaptic actions of GABA include the barbiturates and the benzodiazepines. There is evidence that the oldest of the anticonvulsants, bromide, also works by enhancing the postsynaptic effect of GABA. All these agents prolong the open probability of the GABA receptor. Benzodiazepines do this by increasing the frequency of channel openings, while barbiturates do so by increasing the probability of long channel openings. The net effect of the prolongation of the $GABA_A$ receptor open time is an increase in GABA-gated anionic conductance that hyperpolarizes the neuronal membrane and shunts or short-circuits excitatory conductances. These effects shift the balance of excitation and inhibition back toward inhibition. However, nonspecific prolongation of GABA currents will have a variety of impacts on other brain functions. For example, electroencephalographic rhythms in the gamma frequency band (30–80 Hz) are considered to be important in binding neural activity in different areas of the brain. The precise frequency of the gamma oscillations is determined by the duration of $GABA_A$ receptor-mediated postsynaptic conductances. Prolongation of these conductances by barbiturates and benzodiazepines alters the frequency of the gamma rhythm and is likely to contribute to reduced levels of alertness and cognitive capacity in many patients. Similarly, prolongation and enhancement of $GABA_A$ conductances will limit the voltage-dependent activation of *N*-methyl-D-aspartate (NMDA) receptors and are likely to inhibit NMDA receptor-dependent synaptic alterations thought to be important in learning and memory.

Receptor trafficking As described for the mutation associated with JME, reductions in the surface

expression of $GABA_A$ receptors can be just as problematic as mutations that affect the function of the receptor. This is because shortly after vesicle fusion at GABAergic synapses, GABA receptors are essentially saturated so that receptor number rather than the concentration of extracellular GABA is rate limiting. Under these conditions, a reduction in the density of membrane-bound receptors leads to a reduction in the $GABA_A$ receptor-mediated conductance. Recent studies have demonstrated endocytosis of $GABA_A$ receptors during prolonged seizures, an effect that is thought to be due to prolonged exposure to increased levels of GABA released during the course of the seizure. Similarly, some mutations in $GABA_A$ receptor subunits appear to reduce synaptic $GABA_A$ receptor-mediated conductances by altering receptor trafficking rather than by changing channel function.

Therapeutically, $GABA_A$ receptor trafficking has important implications for the treatment of status epilepticus. Benzodiazepines, because of their limited effect on respiratory drive, are usually the first-line agents in the treatment of prolonged seizures. Reduction in membrane-bound $GABA_A$ receptors correlates with reduced responsiveness to benzodiazepines so that there is a gradual reduction in benzodiazepine anticonvulsant efficacy during prolonged seizures. Barbiturates at high doses remain effective in this circumstance because of the wide variety of channels that interact with higher concentrations of this anticonvulsant. Presumably, anticonvulsants that work separately from the GABA system would also not be affected.

Charge-Carrying Ions

$GABA_A$ receptor-mediated signaling is unique in that it can be modulated not only by manipulation of ligand and receptor interactions but also by the flux of anions through the receptor-operated channel. This is because the driving force for chloride flux through the $GABA_A$ receptor is small and can be reversed by accumulation of a few millimoles of chloride within the cytoplasm. For example, chloride accumulation can be effected by the activity of chloride-accumulating transporters such as Na-K-2Cl cotransporter (NKCC1), as occurs in immature neurons in which GABA receptor activation is excitatory. In adult neurons, the chloride exporter KCC2 is the principal chloride transmembrane transporter and serves to maintain low cytoplasmic chloride concentrations that are conducive to hyperpolarizing influx of chloride during openings of $GABA_A$ receptor-operated channels. However, during prolonged, intense $GABA_A$ receptor activity, chloride accumulation within the cytoplasm eventually abolishes the driving force for further chloride entry. In contrast, bicarbonate efflux can continue indefinitely as a consequence of CO_2 diffusion across the membrane in concert with extracellular hydration and intracellular dehydration of CO_2, as long as the protons that accumulate intracellularly during bicarbonate dehydration can be buffered. The differential duration of the oppositely directed chloride and bicarbonate currents leads to a late depolarizing shift during prolonged $GABA_A$ receptor activation that may contribute to seizures. The importance of chloride homeostasis to $GABA_A$ receptor function was recently demonstrated by the association of generalized epilepsy syndromes with loss-of-function mutations in chloride channel (ClC) 2. ClC-2 is a hyperpolarization-activated, noninactivating ClC that serves as an emergency valve that reduces chloride accumulation within neurons and thereby maintains the $GABA_A$ reversal potential at values negative to action potential threshold.

Therapeutically, depolarizing responses to GABA can be blocked by diuretics. Acetazolamide inhibits carbonic anhydrase, the enzyme that catalyzes the hydration and dehydration of CO_2. This limits the bicarbonate efflux through the $GABA_A$ receptor-operated channel. Acetazolamide is a useful anticonvulsant although, curiously, its efficacy wears off after a few days, perhaps due to increased expression of carbonic anhydrase isoforms that are more resistant to acetazolamide. This limited duration of action is perfect for catamenial epilepsy, and acetazolamide is one of the most frequently used anticonvulsants in this clinical situation. NKCC1, the transporter that accumulates chloride in immature neurons, can be blocked by the diuretic bumetanide, and this has been shown to be an effective anticonvulsant strategy in newborn rats. Human neonatal seizures respond poorly to benzodiazepines and barbiturates, and human neocortical neurons also express NKCC1 in the neonatal period, so it is likely that bumetanide would be an effective anticonvulsant in human neonates as well. Clinical trials to test this hypothesis are currently under way in human neonates.

Alcohol

Extrasynaptic $GABA_A$ receptors that contain delta receptors have a high sensitivity to GABA and a low rate of desensitization, making them well suited for tonic inhibition that contributes to the resting membrane conductance of neurons. As discussed in the section titled 'Epilepsy,' these receptors are sensitive not only to neurosteroids but also to ethanol, and recent research suggests that enhanced tonic inhibition contributes to many of the effects of ethanol intoxication. Chronic ethanol administration is associated with reduced expression of the ethanol-sensitive delta

receptors. Although the delta subunit is an attractive locus for the effect of acute and chronic ethanol exposure on the central nervous system, the characteristics of $GABA_A$ receptors change, as do many other membrane proteins, depending on the cells in which they are expressed, perhaps reflecting the substantial functional impact of cell-specific phosphorylation cascades on receptor function. For example, much less ethanol sensitivity is noted when delta subunit-containing GABA receptors are expressed in oocytes. Ethanol effects on the GABA system may also be mediated by increases in neurosteroids and GABA release. This is an active area of research, driven in part by the hypothesis that familial polymorphisms in GABA receptor subunits may underlie genetic predispositions for ethanol abuse.

Psychiatry

Anxiety and Sleep

There is a long history of use of modulators of GABA receptors for relief of anxiety and sleep, including the use of ethanol, bromides, barbiturates, and benzodiazepines. These agents all have less than optimal side effect profiles, and for all of them, sedation is tightly linked to anxiety relief. Recently, additional insights have been developed using benzodiazepines that are selective for particular subunits and mice in which specific $GABA_A$ receptor subunits have been knocked out. Zolpidem is selective for α1-containing $GABA_A$ receptors and has much more substantial sedative-hypnotic effects than anxiolytic effects. Conversely, α1 knockout mice have anxiolytic but reduced sedative-hypnotic effects in response to benzodiazepine administration. On the other hand, α2 knockout mice show sedative but no anxiolytic effects with benzodiazepine administration. The goal of a pure anxiolytic benzodiazepine is being approached through the development of α2-selective benzodiazepines.

Psychosis

A number of lines of evidence point to altered GABA function in schizophrenia and related disorders, although, with these complex diseases, it is not clear what changes are causal. Schizophrenic patients have reduced cortical levels of an isoform of the enzyme glutamic acid decarboxylase (GAD), GAD67. GAD67 produces GABA from glutamate, and treatment with antipsychotic drugs increases GAD67 levels. Because early cortical injury also alters GAD67 expression, one possible pathophysiology links early injury or dysgenesis to altered cortical circuitry, GAD expression, and psychosis. However, there are many other possible links between GAD67 and psychotic illnesses. For example, single nucleotide polymorphisms in GAD1, the gene encoding GAD67, have been linked to childhood onset schizophrenia.

Movement Disorders and Tremor

Many nuclei of the basal ganglia have GABAergic outputs, as recently described in an excellent group of reviews in *Progress in Brain Research*. Alterations in GABAergic synaptic transmission have long been postulated to be involved in the expression of many movement disorders. Alteration of GABAergic outputs does not mean that these outputs are involved in the pathophysiology. For example, in Parkinson's disease the deficit involves loss of dopaminergic neurons in the substantia nigra. The output of the substantia nigra pars reticulata is predominantly GABAergic, and this output is increased in Parkinson's disease. However, GABA is not directly involved in the pathogenesis of Parkinson's disease, and treatment is directed at replacing the lost dopaminergic drive, rather than antagonizing GABAergic synaptic transmission, because the latter would be nonspecific and would be associated with unacceptable side effects, such as epilepsy.

Some forms of tremor may be related to specific $GABA_A$ receptor polymorphisms. For example, the α1 knockout mice discussed in the section titled 'Anxiety and sleep' also display a prominent tremor that is relieved by alcohol. This is very similar to a tremor in human patients that is often called benign essential tremor or familial tremor. In humans this tremor is inherited in an autosomal dominant fashion. It becomes manifest during skilled movements and is therefore called an intention tremor. It can be disabling and can lead to substance abuse by virtue of tremor suppression by alcohol and other sedative-hypnotic enhancers of $GABA_A$ receptor function. A recent study in humans did not find mutations in the DNA encoding the α1 receptor in a substantial cohort of patients with essential tremor. However, experiences with $GABA_A$ receptor mutations in epilepsy suggest that it is quite possible that receptor trafficking rather than receptor function is altered in essential tremor so that mutations in proteins associated with $GABA_A$ receptor trafficking might represent the next best target for mutational analyses. Alternatively, the knockout mouse may mimic features of human essential tremor; the human disease arises when other subunits are abnormally expressed, and this is mimicked in the α1 knockout mouse.

Spinal Cord Disorders

Spasticity

Spasticity is excessive resting muscle tone that results from inadequate inhibition of lower motor neurons

by descending inputs from the corticospinal tract. Generally, this occurs because of damage to the corticospinal tract, and so spasticity is highly associated with deficits in the skilled use of the affected limb. Because the net effect of activation of these corticospinal inputs is inhibition of lower motor neurons, GABAergic agonists have frequently been used to alleviate spasticity. As mentioned in the introduction, presynaptic $GABA_B$ receptors inhibit transmitter release via activation of potassium channels and inactivation of calcium channels. $GABA_B$ agonists reduce glutamate release that excites lower motor neurons and thereby reduce activation of muscle fibers to decrease spasticity. Systemic administration of $GABA_B$ agonists can produce both unacceptable sedation and, if GABA release is inhibited more than glutamate, seizures. To improve specificity, the restriction of the problem to the spinal cord has been exploited to develop more-specific therapeutic effects via local administration of agonists. Intrathecal delivery of baclofen, a $GABA_B$ receptor agonist, has been demonstrated to be effective in the treatment of both adult and pediatric spasticity, and this method of delivery is associated with minimal central nervous system side effects.

Pain

Phantom or neuropathic pain is a difficult clinical problem in which damage or loss of peripheral nerves is followed by intractable pain experienced in the original somatosensory distribution of the damaged nerve. Many mechanisms have been postulated for this phenomenon, but recent data suggest that an important pathophysiology involves an alteration in the reversal potential for $GABA_A$ receptor-mediated currents in neurons within lamina 1 of the spinal cord. The neurons in lamina 1 are postsynaptic to the neurons in the dorsal root ganglia whose peripheral processes were damaged or severed. There is evidence that this effect on E_{GABA} in the uninjured, postsynaptic neurons in lamina 1 occurs as a consequence of brain-derived neurotrophic factor (BDNF) secreted by microglia. BDNF has been shown to reduce the expression of KCC2, the cation-chloride cotransporter responsible for maintaining low chloride levels in mature neurons. This downregulation reduces the inhibitory effect of GABA by removing the hyperpolarizing component of inhibition. Shunting inhibition remains intact, although modeling studies suggest that at very high intensities of excitatory input, shunting inhibition may fail because the associated reduction in the membrane time constant permits a higher frequency of action potential transmission. Failure of shunting has not been a factor in neocortical studies of epilepsy, but the differences could be due to different rates of synaptic activity in the spinal cord as well as different dendritic geometries that may predispose to activity-dependent shifts in E_{GABA} due to chloride accumulation. Because the shift in E_{GABA} in the spinal cord is due to downregulation of KCC2 instead of upregulation of NKCC1, bumetanide cannot be used to ameliorate neuropathic pain. The involvement of BDNF has provided an important clue for understanding the regulation of transporters, and intervening in the downregulation of KCC2 is an important focus of current research.

See also: Epilepsy; Epilepsy: Channelopathies; GABA Synthesis and Metabolism; GABA-A Receptors: Developmental Roles; GABA-A Receptor Synaptic Functions; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-B Receptors: Molecular Biology and Pharmacology; GABA-B Receptor Function; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism.

Further Reading

Akbarian S and Huang HS (2006) Molecular and cellular mechanisms of altered GAD1/GAD67 expression in schizophrenia and related disorders. *Brain Research Reviews* 52(2): 293–304.

Audenaert D, Van Broeckhoven C, and De Jonghe P (2006) Genes and loci involved in febrile seizures and related epilepsy syndromes. *Human Mutation* 27(5): 391–401.

Ben-Ari Y and Holmes GL (2005) The multiple facets of gamma-aminobutyric acid dysfunction in epilepsy. *Current Opinion in Neurology* 18(2): 141–145.

Benarroch EE (2007) $GABA_A$ receptor heterogeneity, function, and implications for epilepsy. *Neurology* 68(8): 612–614.

Chen JW, Naylor DE, and Wasterlain CG (2007) Advances in the pathophysiology of status epilepticus. *Acta Neurologica Scandinavica* 115(4): 7–15.

De Koninck Y (2007) Altered chloride homeostasis in neurological disorders: A new target. *Current Opinion in Pharmacology* 7(1): 93–99.

Deng H, Xie WJ, Le WD, Huang MS, and Jankovic J (2006) Genetic analysis of the GABRA1 gene in patients with essential tremor. *Neuroscience Letters* 401(1–2): 16–19.

Dzhala VI, Talos DM, Sdrulla DA, et al. (2005) NKCC1 transporter facilitates seizures in the developing brain. *Nature Medicine* 11(11): 1205–1213.

Galvan A and Wichmann T (2007) GABAergic circuits in the basal ganglia and movement disorders. *Progress in Brain Research* 160: 287–312.

Krystal JH, Staley J, Mason G, et al. (2006) Gamma-aminobutyric acid type A receptors and alcoholism: Intoxication, dependence, vulnerability, and treatment. *Archives of General Psychiatry* 63(9): 957–968.

Macdonald RL, Kang JQ, Gallagher MJ, and Feng HJ (2006) GABA(A) receptor mutations associated with generalized epilepsies. *Advances in Pharmacology (San Diego, CA)* 54: 147–169.

Maguire JL, Stell BM, Rafizadeh M, and Mody I (2005) Ovarian cycle-linked changes in GABA(A) receptors mediating tonic

inhibition alter seizure susceptibility and anxiety. *Nature Neuroscience* 8(6): 797–804.

Michels G and Moss T (2007) GABAA receptors: Properties and trafficking. *Critical Reviews in Biochemistry and Molecular Biology* 42: 3–14.

Ong J and Kerr DI (2005) Clinical potential of GABAB receptor modulators. *Cns Drug Reviews* 11(3): 317–334.

White HS, Smith MD, and Wilcox KS (2007) Mechanisms of action of antiepileptic drugs. *International Review of Neurobiology* 81: 85–110.

GABA-A Receptors: Developmental Roles

Y Ben-Ari, Université de la Méditerranée, Marseille, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In adult neurons, GABAergic synapses provide most of the transmitter-gated inhibitory drive, whereas glutamatergic synapses are responsible for the ongoing excitatory drive that impinges on neurons. $GABA_A$ receptor-mediated signaling also plays a key role in the control of neuronal excitability and the generation of behaviorally relevant patterns and oscillations. Activation of $GABA_A$ receptors synchronizes the activity of principal neurons, leading to the generation of gamma and other network-driven oscillations. Blockade of GABAergic receptors generates seizures and other insults, reflecting the importance of an equilibrated activity of GABAergic and glutamatergic signaling. In cortical structures, GABAergic interneurons represent around 10–15% of the total neuronal population and constitute a very heterogeneous population of neurons with unique features that facilitate the control of all aspects of the activity of principal neurons. One key question in developmental neurobiology is how GABA signaling is installed during maturation and how the balance between excitation and inhibition is preserved throughout. This is an important issue considering the fact that any imbalance would lead to either excitotoxic consequences, if the excitatory glutamatergic drive prevails at any time, or inadequate construction of cortical networks and migration disorders, if inhibitory GABA signaling prevails. An equally important question is the maturation of network-driven patterns that shift from silence to adult networks that generate behaviorally relevant patterns. How does the developmental sequence of intrinsic voltage- and transmitter-gated currents provide a substrate for the development of networks? This is of course central in our understanding of how activity modulates genetic programs to entrain the construction of cortical networks.

Studies initiated primarily in Paris almost 20 years ago have unraveled a developmental program in the establishment of GABAergic signaling that appears to have been conserved throughout evolution. This cascade of events includes several key rules that appear also to provide a progressive shift of developing neurons that will first generate a universal primitive pattern that runs on the engine before the network has a sufficient density of synapses and is capable of generating adult patterns and the integrative functions associated with them. These events and how they shed light on the maturation of neurons and networks are discussed in the following sections.

The Excitatory/Inhibitory Shift of GABA Actions during Development

The activation of GABA receptors leads to the opening of a channel primarily permeable to chloride. In adult neurons, the intracellular concentration of chloride is quite low (a few millimoles per liter) and GABA will induce an influx of chloride and a hyperpolarization associated with a shunt of membrane potential that prevents the generation of action potentials. In developing neurons, studies using a wide range of recording techniques, including single GABA channels and perforated patch recordings that are impermeable to chloride, show that the intracellular chloride concentration ($[Cl^-]_i$) is higher (in the range of tens of millimoles per liter) than in adult neurons (Figure 1). As a consequence, the activation of $GABA_A$ receptors will lead to quite different actions, including a depolarization, the activation of sodium spikes, and voltage-dependent calcium currents leading to an excitation and an increase of $[Ca^{2+}]_i$. Perhaps even more strikingly, the activation of $GABA_A$ receptors leads to a removal of the voltage-dependent Mg^{2+} block of *N*-methyl-D-aspartate (NMDA) channels and thus to an additional increase of $[Ca^{2+}]_i$. This 'synergistic' action of GABA receptors and NMDA receptors that is quite opposite to the actions of GABA in adult neurons has important consequences in terms, notably, of plasticity. Several studies indicate that coactivations of GABA and NMDA receptors or calcium channels lead to alterations of synaptic efficacy of GABAergic synapses, including a long-term depression or potentiation of GABAergic synapses according to the source of calcium influx activated. They also modulate the formation of novel synapses in neurons that are quiescent. Thus, repetitive electrical stimulation of neurons that are silent – with no functional synapses – triggers the expression of GABAergic synapses, as reflected by the presence of spontaneous GABAergic postsynaptic currents (PSCs).

The developmentally regulated reduction of $[Cl^-]_i$ and the excitatory (E) to inhibitory (I) shift of the actions of GABA (E to I) have now been confirmed in all developing brain structures and animal species studied, suggesting that they have been preserved throughout evolution. It is suggested that excitatory actions of GABA provide a source for calcium influx

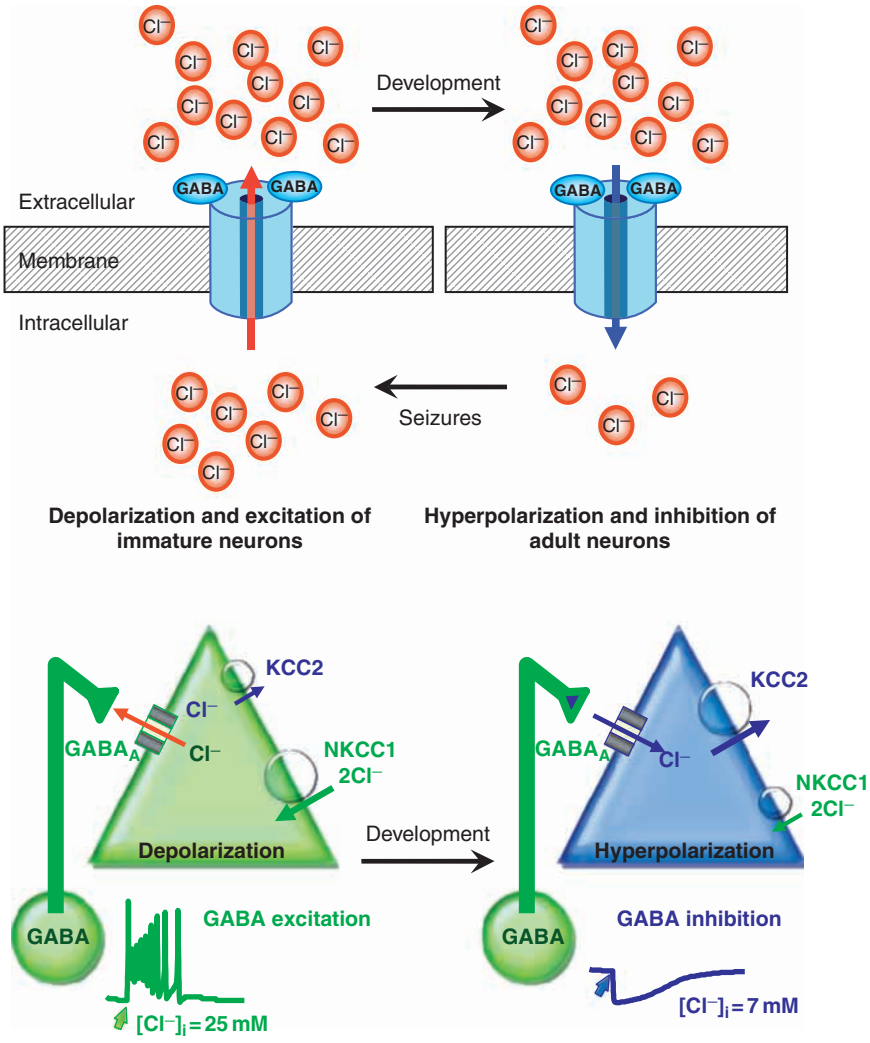

Figure 1 The developmental shift of γ-aminobutyric acid (GABA) actions. In immature neurons, there is a high intracellular concentration of chloride (Cl^-). This is reduced during maturation, leading to an excitatory-to-inhibitory shift. The mechanisms include an early expression of a chloride importer (NKCC1) and a late expression of a chloride transport extruder (KCC2). From Ben-Ari Y (2002) Excitatory actions of GABA during development: The nature of the nurture. *Nature Reviews Neuroscience* 3(9): 728–739.

needed for growth while avoiding the possible excitotoxic actions of unbalanced glutamatergic drive, because of the small magnitude of GABA excitation and the shunting actions intrinsic to its operation. In this model, the E-to-I shift occurs when the density of glutamatergic synapses reaches a level that requires an efficient compensatory inhibition.

How Does the E-to-I Shift Occur?

The concentration of $[Cl^-]_i$ is regulated by several cotransporters, and two of them appear to play an important role (Figure 1):

1. KCC2 is the principal transporter for Cl^- extrusion from neurons. KCC2 extrudes K^+ and Cl^- using an electrochemical gradient for K^+. Developmental studies suggest that KCC2 is exclusively expressed in mature neurons: the KCC2 that extrudes chloride follows a developmental expression curve that corresponds well with the higher $[Cl^-]_i$ at an early stage. An early expression by gene transfection of the transporter leads to the expected E-to-I shift. Interestingly, in neuronal cultures, an early expression of the transporter at a time when immature neurons have high $[Cl^-]_i$ leads, in addition to the expected shift, to a selective increase of the density of GABAergic – but not glutamatergic – synapses in the transfected neuron. Therefore, the intracellular concentration of chloride plays an important role in synapse formation and by a yet unknown mechanism may regulate the establishment of GABAergic synapses.

2. The NKCC1 is a membrane transport protein that mediates chloride uptake across the plasma membrane, internalizing one Na^+, one K^+, and two Cl^- ions in an electroneutrally coupled fashion. NKCC1 also plays a housekeeping role in cell volume homeostasis. The early expression of this chloride importer plays an important role in maintaining high intracellular $[Cl^-]_i$. Studies using selective antagonists of this transporter suggest that it also plays an important role in early epileptogenesis.

The mechanisms responsible for the reduction of $[Cl^-]_i$ and the excitatory-to-inhibitory shift of GABA actions are not fully understood. Some observations suggest that GABA itself may do the job – that is, it is the activation of immature GABAergic synapses providing the signal that activates KCC2 and the removal of chloride. This is based on neurons in cultures that do not exhibit the shift when $GABA_A$ receptors are blocked. Other studies, however, cannot be reconciled with such a mechanism and suggest that the shift is programmed and occurs even when GABA signaling is blocked.

Therefore, the E-to-I shift adheres clearly to a basic mechanism that includes the expression of mechanisms that remove excess chloride. However, it remains to determine quantitatively the alterations of $[Cl^-]_i$ that occur dynamically during ongoing activity in developing neurons. The issue is to understand how chloride is dynamically regulated when an accumulation occurs and how that process is modified by activity. Other observations suggest that in a variety of pathological conditions, KCC2 is downregulated, leading to a chronic accumulation of chloride occurring on a persistent basis. This suggests that pathological insults in adults recapitulate the developing brain situation, but also that $[Cl^-]_i$ is an important signal in neurological disorders.

When Does the Shift Occur?

The E-to-I shift occurs in a species- and structure-dependent manner. In the rodent hippocampus, the shift occurs during the second postnatal week, although there are important differences between neurons according to their developmental stage: the shift will occur earlier in interneurons or CA3 pyramidal neurons that mature before CA1 pyramidal neurons. In other structures, the E-to-I shift will occur earlier or later according to developmental programs. This is particularly important in humans and subhuman primates, which have an extended developmental period, because recently born cortical neurons will have excitatory GABA whereas adjacent neurons that belong to the same neuronal population but that became postmitotic at an earlier stage will already operate with thousands of inhibitory-operated GABAergic synapses. This heterogeneity must be taken into account in models of network development. In addition, in a clinical perspective, drugs that augment the efficacy of GABA actions (e.g., benzodiazepines) will affect GABA signaling differently in the mother's brain and that of the embryo at a given developmental stage.

GABAergic Synapses Are Formed before Glutamatergic Synapses

In adult neurons, applications of GABA receptor antagonists generate seizures because of the removal of a tonic and phasic inhibitory drive, whereas similar applications of glutamate receptor antagonists produce a full blockade of ongoing excitatory activity. Studies performed in immature slices indicate, in contrast, that similar applications of GABA receptor antagonists block ongoing activity instead of generating seizures, suggesting that the excitatory drive is provided by GABA. However, the effects of antagonists can be misleading, particularly in developing networks, because of their heterogeneous features and expression of different receptor subunits. Direct demonstration that GABA signaling does in fact mature before glutamate signaling matures came from studies performed in primate hippocampal neurons *in utero* and rodent embryos and early postnatal neurons; in these studies neurons were first recorded, their physiological properties were determined, and morphological *post hoc* reconstruction of neurons was performed to determine their developmental stage (**Figure 2**). This showed that silent neurons with no active PSCs have a small axon and little or no dendrites, neurons with a small apical dendrite have only GABAergic PSCs, and neurons with both GABA and glutamate PSCs have longer and more developed apical and basal dendrites. Neurons with glutamate but no GABAergic PSCs were not encountered, suggesting that GABAergic synapses are formed first on apical dendrites and glutamatergic synapses are formed only once the postsynaptic neuron has reached a certain degree of maturity (**Figure 2**). Since afferent glutamatergic axons are already present at an early stage, this sequence is not due to a late arrival of inputs but rather to different requirements for the formation of GABA and glutamatergic synapses: glutamatergic synapses require a more developed postsynaptic target to form synapses than GABAergic axons. Therefore, once neurons have a primitive apical dendrite, they will form GABAergic but not

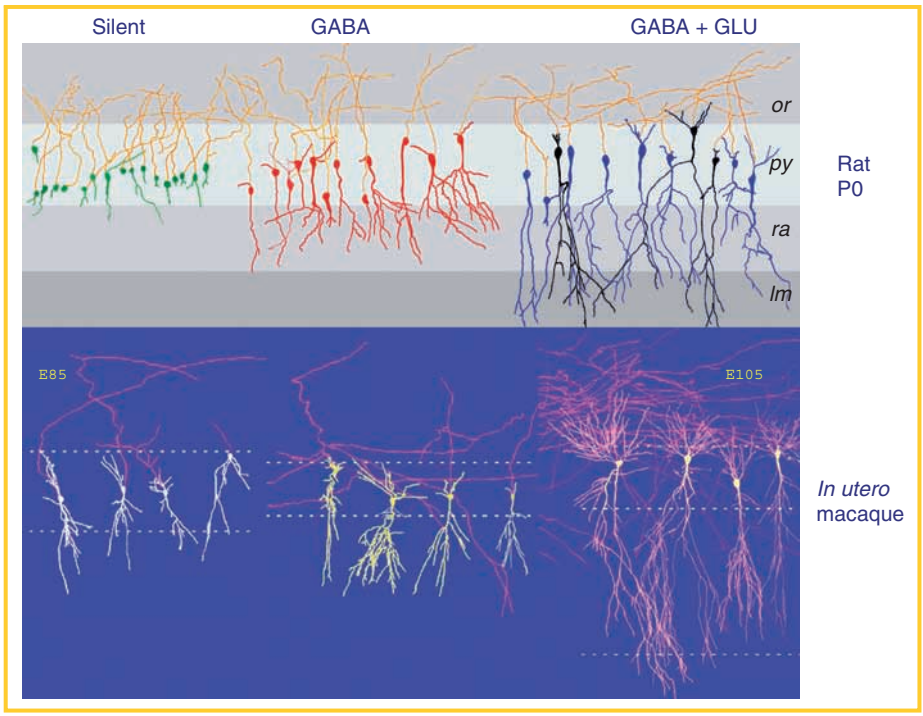

Figure 2 Early formation of γ-aminobutyric acid (GABA) synapses and the GABA–glutamate (GLU) sequence. CA1 pyramidal neurons from rodent slices at birth (top) and from primate hippocampus at various gestational stages (E85 to E105; bottom) were recorded in slices and reconstructed after identification of their synaptic responses. In both cases, silent neurons with no functional postsynaptic currents are small, with an axon but no developed dendrites; neurons with GABA but no GLU postsynaptic currents have a small apical dendrite, and neurons with both GABA and GLU postsynaptic currents have an elaborated apical and basal dendrite. (Top) Data from Tyzio R, Represa A, Ben-Ari Y, et al. (1999) The establishment of GABAergic and glutamatergic synapses on CA1 pyramidal neurons is sequential and correlates with the development of the apical dendrite. *Journal of Neuroscience* 19(23): 10372–10382. (Bottom) Data from Khazipov R, Esclapez M, Caillard O, et al. (2001) Early development of neuronal activity in the primate hippocampus *in utero*. *Journal of Neuroscience* 21(24): 9770–9781.

glutamatergic synapses, and only once they have well-developed apical and basal dendrites will they also have functional glutamatergic synapses. A similar developmental gradient is observed in other structures. Interestingly, studies in proliferating neurons in the adult fascia dentata show a similar gradient: newly proliferating neurons have initially only excitatory GABAergic synapses and only when they have a more mature dendritic arbor and morphology do they also have functional glutamatergic synapses. Therefore, this sequence is somehow obligatory; neurons must express the E-to-I shift during their maturation. One consequence of this second developmental rule is that at an early stage, GABA provides most if not all ongoing activity.

If GABAergic synapses mature before there is maturation of glutamatergic synapses, then the neurons that synthesize and release GABA must mature before glutamatergic neurons do. This indeed was found to be the case. Patch clamp recordings from immature GABAergic interneurons and reconstruction revealed that they follow the same sequence as the principal neurons do, but earlier: *in utero*, in rodents, interneurons first have GABAergic synapses, and when they are more developed also have glutamatergic synapses, at a stage when most pyramidal neurons are silent. Thus, at E18, in rodents, most pyramidal neurons have no functional synapse at a time when most interneurons already operate with functional GABA and glutamate synapses. In fact, the earliest patterns recorded *in utero* are generated by a network of interneurons, indicating that these neurons, which control the generation of many fundamental adult patterns, also drive early patterns. Therefore, in spite of their long journey along the tangential migration pathway, GABAergic neurons, at least in the hippocampus, become postmitotic first, then establish synapses and extend a dendritic and axonal arbor to organize the first network pattern. The first functional synapses in the developing hippocampus are formed between GABAergic interneurons.

Giant Depolarizing Potentials: A Primitive Network Oscillation, Present in All Developing Systems, That Disappears When the E-to-I Shift Is Completed

Studies performed initially in the developing hippocampus *in vivo* and *in vitro* have unraveled a unique primitive pattern that disappears around the second postnatal week when the E-to-I shift has been completed. This pattern was named 'giant depolarizing potentials' (GDPs; Figure 3), since it was identified in intracellular recordings and was characterized by its long-lasting duration and high amplitude at resting membrane potential. GDPs are characterized by the presence of synchronized synaptic currents mediated by excitatory GABA and glutamatergic PSCs and, particularly, by NMDA receptor-mediated PSCs, suggesting that the GABA and NMDA synergistic actions play a central role in GDP generation. GDPs disappear in the hippocampus by the second week of postnatal life, when more elaborated behaviorally relevant patterns, including gamma oscillations, are first recorded. This pattern is universal in that similar oscillations have been observed in a wide range of developing structures and preparations, including the spinal cord and neocortex. The key signal is the completion of the E-to-I shift that is associated with the loss of GDPs, suggesting that the excitatory actions of GABA are instrumental in the generation of GDPs. GDPs can be generated in small slices in all subregions of the hippocampus, suggesting that a small network of GABAergic interneurons and principal cells can generate them.

To study the propagation of GDPs, an interesting preparation was developed: the intact hippocampi from both hemispheres, along with their dissected connecting commissure, are placed in a triple chamber with three independent compartments, allowing a perfusion of the three components with different liquids. With this preparation, GDPs were found to propagate from one hemisphere to the other, and within the same hippocampus to propagate following developmental gradients: higher-frequency GDPs generated in more developed subregions propagate to less developed regions that generate less frequent GDPs. The entire hippocampus behaves like a syncytium, with all regions generating GDPs but more developed regions having a higher density of glutamatergic synapses generating GDPs with a higher probability of propagating to less developed regions. GDPs have been recorded *in vivo* in rodents, including in their neocortex. These patterns may correspond to the 'delta brush' pattern observed in preterm babies.

GDPs thus provide a signal that entrains large networks together at an early developmental stage by means of the excitatory actions of GABA and the activation of NMDA receptor-driven PSCs that are very long-lasting in immature neurons because of the expression of subunits with longer kinetics. The long-lasting kinetics of GDPs are highly suitable

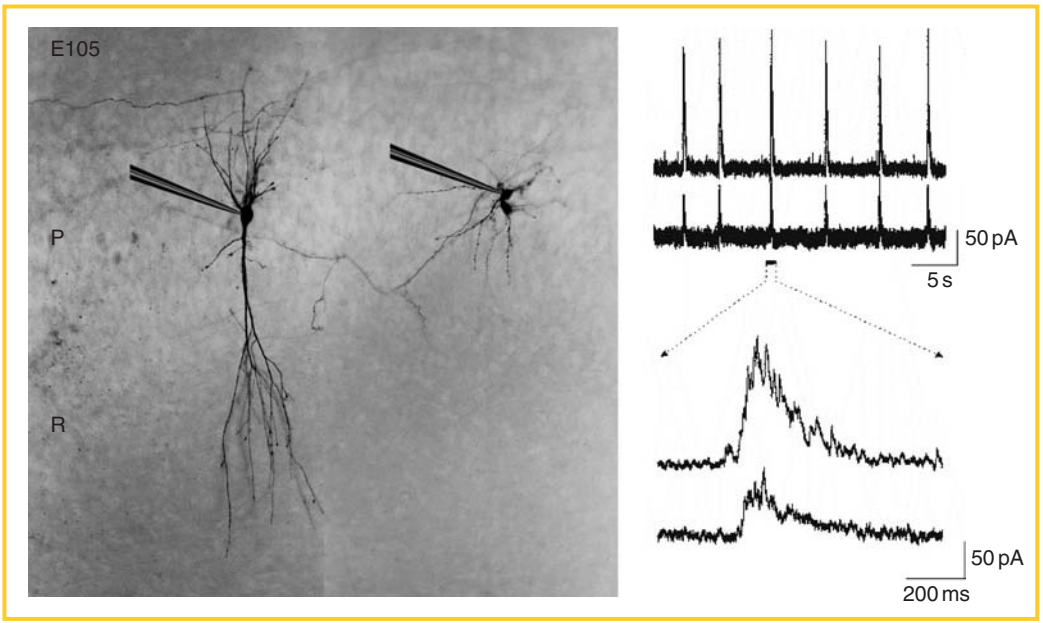

Figure 3 Giant depolarizing potentials in primate hippocampal neurons *in utero* (E105; P, pyramidal layer; R, stratum radiatum). Two neurons were patch clamp recorded and reconstructed *post hoc*: one was a pyramidal neuron and the other was an interneuron. Both had synchronized giant depolarizing potentials. From Khazipov R, Esclapez M, Caillard O, et al. (2001) Early development of neuronal activity in the primate hippocampus *in utero*. *Journal of Neuroscience* 21(24): 9770–9781.

to activate a very heterogeneous population of neurons, some of which have very few active synapses and others of which already have large numbers of synapses. The long time constant of GABaergic PSCs, in comparison with AMPA receptor-driven PSCs, is useful: with a short kinetic event it would have been quite difficult to activate neurons endowed with a few active synapses. This is an instrumental property that will enable neurons that fire together to 'wire' together, and most likely facilitates the formation of functional units. It remains to determine the mechanisms that eliminate GDPs and the relation between the density of active synapses and this shift. Also, the role of activity in that elimination will have to be better understood.

Developmental Curve of Primate Neurons

Are the principal elements of this cascade present in primate neurons *in utero*? This is quite important, considering the relevance to humans, but also because of the ontogenetic implications of better determining the developmental curves of neuronal maturation in rodents and primates. A pioneer study recorded central primate neurons during fetal development: a cesarean operation was performed in pregnant macaques, the fetal brain was removed, and the entire developmental sequence was determined using a morphofunctional characterization of the properties of neurons in relation to their developmental stage (Figures 2 and 4). At midgestation, pyramidal neurons of CA1 are immature with little or no dendrites and no functional synapses (i.e., no PSCs). Next, neurons have a small apical dendrite and only GABAergic synapses, and only later do they also form glutamatergic synapses. A quantitative analysis revealed that axons develop before dendrites do; also, whereas at midgestation neurons have no spines, a few weeks before full term they have as many as 7000 spines; these data allow quantitation of the speed of formation of synapses in primate central neurons. GABA excites immature neurons, as reflected by the fact that a GABA receptor antagonist has no effect at midgestation but generates seizures a few weeks later. This also reflects the importance of seizures *in utero*, in that neurons are able to generate synchronized patterns *in utero* and are highly excitable (see later).

The curve depicting this maturational sequence in primates, which in fact is more completed than that in rodents, is shown in Figure 4. Note that GDPs are present until a few weeks before delivery; this suggests that the loss of GDP signals is an important maturational stage, when networks can generate behaviorally relevant patterns. Again, the E-to-I

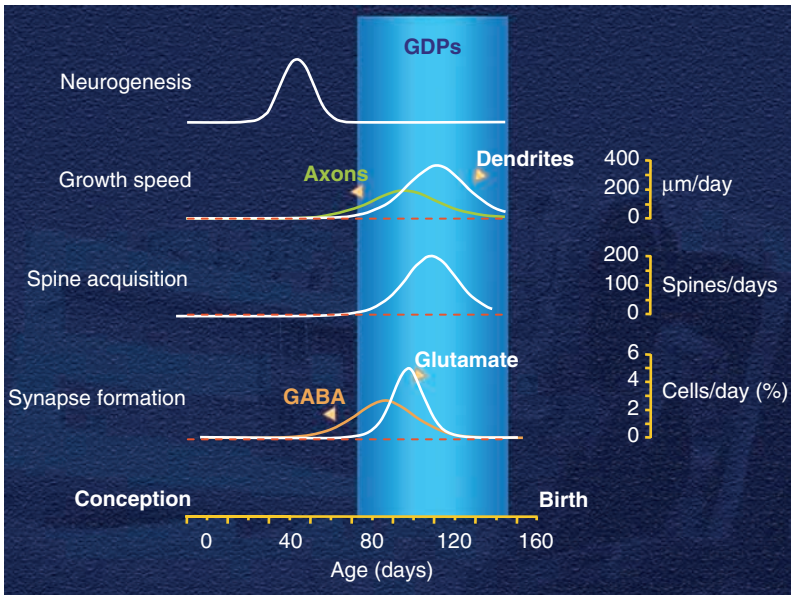

Figure 4 Boltzmann equation used to depict various functions during maturation of primate neurons *in utero*, from neurogenesis in the hippocampus to speed of growth of axons, spine acquisition, and synapse formation. All values are derived from intracellular recordings and reconstruction of primate neurons in slices *in utero* (removed by cesarean surgery), with values calculated. Note the early formation of axons (prior to dendrite formation), the early maturation of GABA signals, and the rapid formation of spines, between midgestation and delivery, shifting from 0 to over 7000 synapses a few weeks before delivery. Note also the domination of giant depolarizing potentials (GDPs) until a few weeks before delivery, indicating that GDPs in primates are an *in utero* pattern, in contrast to that in rodents. From Khazipov R, Esclapez M, Caillard O, et al. (2001) Early development of neuronal activity in the primate hippocampus *in utero*. *Journal of Neuroscience* 21(24): 9770–9781.

shift is an important signal, since the disappearance of GDPs and the completion of the E-to-I shift indicate that at this stage there is a sufficient density of glutamatergic synapses and a network capable of integrative activity. This curve is therefore an important source of information on network development in primates.

Paracrinic Action of GABA Prior to Synapse Formation

In all developing peripheral (i.e., the neuromuscular junction) or central cells, receptors are established before synapses are. Does this contribute to the process that established cortical networks? Do the excitatory actions of GABA also play a role at that earlier stage? Studies performed in immature hippocampal neurons confirm the presence of functional receptors at an early stage and indicate that stimulation of immature neurons that bear no functional synapse generates a long-lasting current present only at an early stage. This early current is mediated primarily by GABAergic receptors, as the current is almost entirely blocked by GABA receptor antagonists. The remaining current is NMDA, but not AMPA, receptor driven. Therefore, these early expressed receptors somehow sense a release of GABA and generate long-lasting currents in neurons with no functional synapses.

How are the transmitters released? Blocking both sodium and calcium currents and vesicular release with toxins fails to block that response. This response persists even in recordings made in a knockout in which vesicular release had been obliterated by inactivating Munc18, a protein required for vesicular release. Therefore, a nonvesicular release of GABA occurs before the conventional vesicular release, and this diffuses to distal neurons to generate long-lasting currents. Again, GABA appears to operate earlier than glutamate does and to exert an important control on early activity.

Are these early actions of GABA – and to a lesser extent, glutamate acting on NMDA receptors – functional? Studies using slice cultures indicate that blocking GABA receptors at a time when neurons have no synapses retards considerably neuronal migration. Therefore, a nonvesicular release of GABA acts on neurons that have receptors, but no functional synapses, to modulate neuronal migration.

The diffusion of GABA to distal sites depends on the efficacy of transporters. Using GABA and glutamate transporter blockers, it was found that glutamate, but not GABA, transporters operate at an early stage. Therefore, GABA will diffuse to distal sites and act in a paracrinic manner whereas the concentration of extracellular glutamate is tightly controlled from the start to prevent toxic actions. In keeping with this finding, *in vivo* and *in vitro* injections of a glutamate transport blocker generate long-lasting oscillations and seizures with a clinical and electrographic pattern reminiscent of encephalopathies.

Trophic Actions of Excitatory GABA

Early studies suggest that GABA exerts a plethora of trophic actions in several systems. GABA stimulates neuronal growth and dendritic arborization, synapse formation, and neuronal differentiation. GABA also exerts an important action on neuronal proliferation by means of action on DNA in neuroblasts. Therefore, at all early stages, GABA can modulate important functions by means of its excitatory action.

I-to-E Shift after Insults: Pathogenesis Recapitulates Ontogenesis

It has long been known that high-frequency electrical activity induces an accumulation of chloride – most likely because the capacity to remove chloride is limited. This is most conspicuously shown in seizures and other insults associated with high-frequency electrical activity. Such activity leads to excitatory actions of GABA in a transient or even quasi-permanent manner; excitatory actions of GABA have been observed in human epileptic tissue months after the last seizure, and various insults, including trauma and lesions, also lead to a long-lasting I-to-E shift.

Recent studies using a convenient novel *in vitro* preparation have enabled determination of the mechanism of induction and expression of this shift and of the electrographic events required for its occurrence. A triple chamber was constructed to accommodate the two intact hippocampi and the commissure connection in three independent chambers that can be perfused with different liquids. One hippocampus was perfused with a convulsive agent (kainate) and the other with an artificial cerebrospinal fluid (ACSF). Seizures generated in one hippocampus propagated to the other hippocampus and led to the formation of a chronic mirror focus, since after a few seizures, the naive hippocampus (the one that did not receive kainate but was subjected to recurrent propagated seizures) became epileptic and generated seizures when disconnected from the other hippocampus. This situation persisted for up to 2–3 days (i.e., as long as the preparation could be kept alive *in vitro*).

This preparation was then used to determine the conditions required for the shift. Applications of an NMDA or GABA receptor antagonist to the naive side – while kainate was applied only to

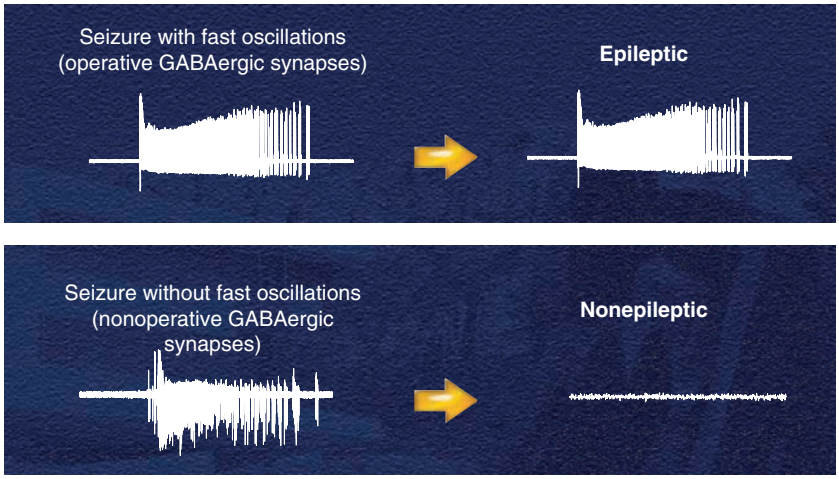

Figure 5 Seizures beget seizures in the triple chamber *in vitro*. When a seizure is generated in one intact hippocampus, it travels to the other hippocampus and leads to the formation of a secondary epileptogenic mirror focus, capable of spontaneously generating seizures. However, this transformation of a naive network to one that seizes requires high-frequency oscillations in the propagated paroxysmal events. If the electrographic event does not include high-frequency oscillations, it will not lead to long-lasting consequences. Since high-frequency oscillations require functional γ-aminobutyric acid (GABA; and *N*-methyl-D-aspartate) signaling for their generation, blocking GABA receptors does not block the seizures but rather prevents their pathogenic consequences: GABA receptor antagonists are anticonvulsive but not antiepileptogenic. From Khalilov I, Holmes GL, Ben-Ari Y (2003) *In vitro* formation of a secondary epileptogenic mirror focus by interhippocampal propagation of seizures. *Nature Neuroscience* 6(10): 1079–1085.

the stimulated hippocampus – prevented the formation of a mirror focus, although this did not block the propagation of seizures. Therefore, functional GABA and NMDA receptors are required for seizures to beget seizures, although they are not required for the propagation of seizures. What are the elements in the seizure needed for this transformation to take place? When time frequency curves were constructed to evaluate the components of the seizure, it was found that high-frequency oscillations (60–90 Hz) are present in the ictal events that led to the formation of a mirror focus: seizures that included lower frequencies failed to generate a mirror focus, and blocking GABA or NMDA receptors also blocked the high-frequency components of the seizures (Figure 5). Therefore, the GABA–NMDA synergy is needed for seizures to activate a persistent transformation of a network from that of control to an epileptic one. Finally, using this preparation, GABA was found to excite neurons in the chronically epileptic tissue, suggesting that epileptogenesis has induced a shift to an immature state: epileptogenesis recapitulates ontogenesis. This illustrates the dynamic properties of the shift and its contribution to neurological disorders.

Implications of These Basic Rules and a General Model of the Establishment of Activity in Developing Cortical Networks

These observations suggest that the unique features of GABA receptors that are not shared by other transmitters (notably, glutamate) – that is, the capacity to shift from excitation to inhibition in an activity-dependent manner – play an important role in early brain development. The high $[Cl^-]_i$ that appears to be a basic feature of developing neurons, most likely an evolutionarily conserved feature, is used by developing neurons to provide a small excitatory drive and is needed to augment intracellular calcium concentrations with little danger, when compared to the potentially toxic actions of glutamate. The long duration of GABA PSCs – when compared to AMPA receptor-driven PSCs – will also facilitate the generation of synchronized events in neurons that have a few functional synapses. The synergistic actions of GABA and NMDA receptors will further stimulate these effects, in particular when the long-lasting actions of the NMDA receptor subunits, present at an early developmental stage, are considered. The synergistic actions of GABA and NMDA will generate the first network-driven pattern capable of synchronizing a wide range of neurons at very different developmental stages. This primitive pattern, which has little informative content, is present to activate neurons synchronistically; it will disappear when a sufficient density of glutamatergic synapses is present and when the E-to-I shift has taken place in the vast majority of GABA synapses. Only at that stage is the network capable of generating more complex, behaviorally relevant patterns. This suggests that this sequence provides a solution to the issue of how to develop excitatory and inhibitory drives while preventing a

transient imbalance between excitation and inhibition that would have potentially excitotoxic actions.

See also: GABA Synthesis and Metabolism; GABA-A Receptor Synaptic Functions; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-A Receptors and Disease; GABA-B Receptors: Molecular Biology and Pharmacology; GABA-B Receptor Function; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism.

Further Reading

Ben-Ari Y (2003) Neuroscience from a to z. *Nature Neuroscience* 6(9): 903.

Ben-Ari Y, Cherubini E, Corradetti R, et al. (1989) Giant synaptic potentials in immature rat CA3 hippocampal neurones. *Journal of Physiology (London)* 416: 303–325.

Ben-Ari Y, Khazipov R, Caillard O, et al. (1997) GABAA, NMDA and AMPA receptors: A developmentally regulated "ménage à trois." *Trends in Neurosciences* 20(11): 499–547.

Ben-Ari Y, Khazipov R, and Gaiarsa JL (2001) Developing networks play the same melody. *Trends in Neuroscience* 24(6): 353–360.

Khalilov I, Holmes GL, and Ben-Ari Y (2003) *In vitro* formation of a secondary epileptogenic mirror focus by interhippocampal propagation of seizures. *Nature Neuroscience* 6(10): 1079–1085.

Khazipov R, Esclapez M, Caillard O, et al. (2001) Early development of neuronal activity in the primate hippocampus *in utero*. *Journal of Neuroscience* 21(24): 9770–9781.

Rivera C, Voipio J, Payne JA, et al. (1999) The K^+/Cl^- cotransporter KCC2 renders GABA hyperpolarizing during neuronal maturation. *Nature* 397(6716): 251–255.

GABA$_A$ Receptors: Molecular Biology, Cell Biology, and Pharmacology

T Goetz, P Wulff, and W Wisden, University of Aberdeen, Aberdeen, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

If nervous systems possessed only excitatory neurons, there would be no brake on any signal, excitation would generate further excitation, and a chain of neurons would produce amplifying cascades of excitation, making the system incapable of doing anything useful. By contrast, networks made from both excitatory and inhibitory neurons can self-organize and generate complex properties. Inhibition is essential for every operation performed by any neuronal circuit in any brain region. In vertebrate brains, the inhibitory agent used most often by neurons is γ-aminobutyric acid (GABA), which after release from the presynaptic terminals diffuses across the synaptic cleft to bind to receptor molecules – the so-called GABA$_A$ receptors. As for all other ligand-gated channels, GABA$_A$ receptors convert chemical messages into electrical signals. In less than a millisecond, the binding of two (tiny) molecules of GABA induces a conformational change in the (giant) receptor oligomer that opens the central chloride ion channel. This remarkable process is called 'gating.' Cl^- diffuses through the receptor pore, down the electrochemical gradient, entering the cell and hyperpolarizing it. This results in fast synaptic inhibition (on the millisecond timescale) of a domain of the postsynaptic cell (Figure 1).

Strictly, GABA$_A$ receptors are actually GABA-gated anion-permeable channels, with a HCO_3^-/Cl^- permeability ratio of approximately 0.2–0.4. In adult cells, Cl^- ions usually move into the cell to produce strong inhibitory hyperpolarization, as the reversal potential for Cl^- is 15–20 mV more negative than the resting membrane potential. The Cl^- gradient is maintained by K-Cl cotransporters. HCO_3^- ions move out of the cells through the GABA$_A$ receptor channel (Figure 1); the HCO_3^- efflux is mildly depolarizing (HCO_3^- has a reversal potential of -12 mV), but this is normally offset by the Cl^- hyperpolarization. Depending on the local internal Cl^- concentration, Cl^- ions can also move out after GABA$_A$ receptor activation and depolarize the cell; this happens especially during embryonic and postnatal neuronal development.

Molecular Biology of GABA$_A$ Receptors

In mammals, GABA$_A$ receptors form as heteropentameric assemblies from a family of 19 subunits encoded by distinct genes (α1–α6, β1–β3, γ1–γ3, δ, ε, θ, π, and ρ1–ρ3). Depending on the subunit composition, GABA$_A$ receptors differ in their biophysical properties and affinity for GABA, their pharmacology, and their location on the cell. GABA$_A$ receptors were cloned by the combined efforts of the Eric Barnard and Peter Seeburg groups in 1987 by the classical method: Peptide sequences obtained from purified (bovine brain) receptors were used to construct synthetic DNA probes to screen brain complementary DNA (cDNA) libraries. This was the starting point. By 1990, this now historical technique of screening cDNA libraries had revealed most of the gene family, all the α1–6, β1–3, γ1–3 subunits and one δ subunit; over the remaining decade, a few more subunits, such as ε, θ, and π, were characterized. With the completion of the human genome database, an *in silico* hybridization method was used to screen for further mammalian GABA$_A$ receptor genes, but it found none. Most of the subunit gene family members are in clusters, suggesting gene and then cluster duplication during the evolutionary origin of vertebrates: β2, α6, α1, γ2 form a cluster in that order on human chromosome 5q34; the β3, α5, γ3 genes cluster in that order on human chromosome 15q13; the γ1, α2, α4, β1 genes cluster in that order on chromosome 4p12; the ε, α3, θ genes in that order on X q28; the ρ1 and ρ2 genes, 40 kbp apart on 6 q15; and the π, ρ3, and δ subunit genes are isolated on human chromosomes 5q35.1, 3q12.1, and 1p36.3, respectively.

As determined by both *in situ* hybridization (mRNA localization) with gene-specific probes (Figure 2) and immunocytochemistry (protein localization) with subunit-specific antibodies, the expression of the individual subunit genes is age and region specific. Does the clustering of the GABA$_A$ receptor subunit genes imply they are coregulated? The α1 and β2 genes do indeed share identical transcription patterns; thus, these two genes may share regulatory elements. All the other subunit genes have sometimes common, sometimes divergent expression patterns, with no correlation with which gene is in which gene cluster.

GABA$_A$ Receptor Structure

The GABA$_A$ receptor belongs to a superfamily of ligand-gated ion channels ('Cys-loop receptors') that

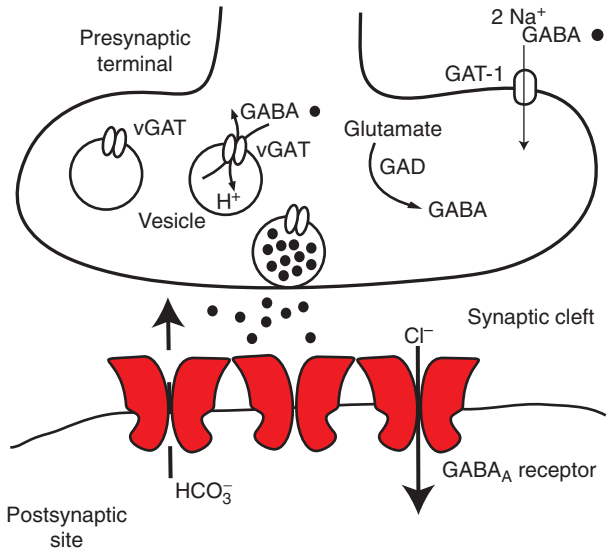

Figure 1 The GABAergic synapse in the mammalian brain. GABA is synthesized from glutamate by glutamic acid decarboxylase (GAD) and transported into vesicles by the vesicular GABA transporter (vGAT). When the presynaptic terminal is depolarized sufficiently, GABA vesicles are released into the cleft and diffuse rapidly to activate the postsynaptic GABA$_A$ receptors. GABA is salvaged from the synaptic cleft by, for example, the GAT-1 transporter on presynaptic terminals. The GABA$_A$ receptor is a GABA-gated Cl^- and HCO_3^- channel.

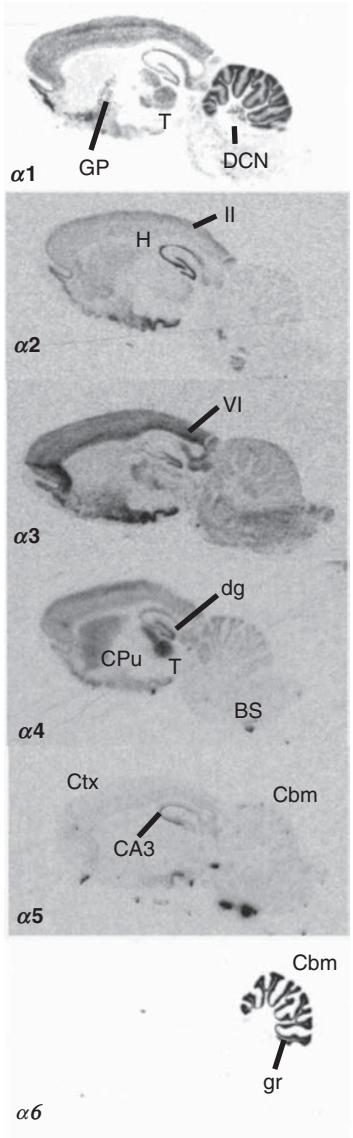

Figure 2 Expression of the GABA$_A$ receptor α subunit genes in adult rat brain (sagittal sections) as detected by *in situ* hybridization. BS, brain stem; CA3, subdomain of hippocampus; Cbm, cerebellum; CPu, caudate-putamen; Ctx, neocortex; DCN, deep cerebellar nuclei; dg, dentate granule cells; GP, globus pallidus; gr, cerebellar granule cells; H, hippocampus; T, thalamus. Roman numerals mark neocortical layers.

in vertebrates includes the nicotinic acetylcholine receptors, the 5-hydroxytryptamine type 3 (5-HT3) receptors, the zinc-activated ion channel, and the glycine receptors. No direct structural information is available for GABA$_A$ receptors. So in imagining how the GABA$_A$ receptor must look, we can do no better than quote Unwin for his empirical observations on the *Torpedo* nicotinic acetylcholine receptor:

> The receptor (a large 290 kDa glycoprotein) is composed of elongated subunits, which associate with their long axes approximately normal to the membrane, creating a continuous wall around the central ion-conducting path. The whole assembly presents a rounded, nearly 5-fold symmetric assembly when viewed from the synaptic cleft, but is wedge-shaped when viewed parallel with the membrane plane. The subunits of the receptor all have a similar size 30 A × 40 A × 160 A and the same three-dimensional fold. Each subunit is a three-domain protein and so portions the channel naturally into its ligand-binding, membrane-spanning and intracellular parts.

In GABA$_A$ receptors, the arrangement of subunits around the channel is probably $\gamma\beta\alpha\beta\alpha$ or $\delta\beta\alpha\beta\alpha$ counterclockwise when viewed from the extracellular space (Figure 3(a)). For those cells in which they are expressed, ε and π subunits probably replace the γ and δ subunit within the pentamer, whereas the θ subunit might replace a β subunit. As for all members of the nicotinic receptor superfamily, all GABA$_A$ receptor subunits contain a large extracellular N-terminal domain of approximately 200 amino acids shaped by a cysteine disulfide bridge (the so-called Cys-loop, Figure 3(b)). For GABA$_A$ subunits, the amino acid consensus sequence of the Cys-loop is C*****F/YP*D***C*****S (where * is a degenerate residue).

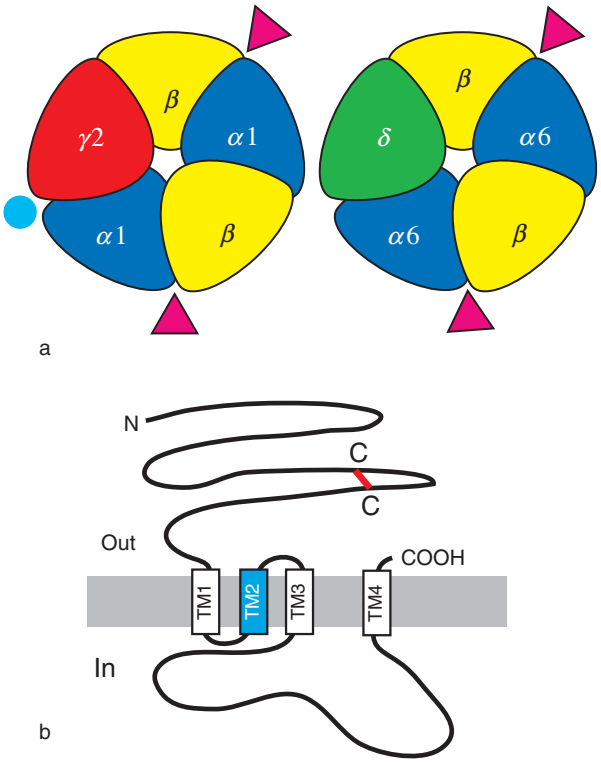

Figure 3 (a) Subunit arrangement in γ2-containing (synaptic) and δ-containing (extrasynaptic) GABA$_A$ receptors; (b) schematic of an individual subunit topology. C, cysteine; N, N-terminus of the protein TM, transmembrane domain. The pore lining domain, TM2, is shown in blue. Pink triangles represent GABA; blue circle represents a BZ ligand.

Each subunit contains four predicted transmembrane spanning domains (TM1–TM4) of about 20 amino acids and a large intracellular loop between TM3 and TM4 (TM3–TM4 loop) (Figure 3(b)). Many GABA$_A$ receptor subunits have the amino acid sequence (TTVLTMTT) in the TM2 domain. Five of these eight amino acids probably line the ion channel. TM1, TM3, and TM4 segregate TM2 from membrane lipid. The selectivity filter and gate lies at the intracellular end of the TM2 domains and includes part of the TM1–TM2 loop. In the 5-HT3 and nicotinic receptors, residues in the TM3–TM4 loop region influence single channel conductance. The TM3–TM4 loop contributes key sites for attaching anchor and regulatory proteins involved in locating the receptor at synapses and in governing the activity of GABA$_A$ receptor.

In the extracellular domain of a typical GABA$_A$ receptor consisting of two α, two β, and one γ2 subunit, the binding pocket for GABA forms at the interface between the α and the β subunit, and the binding pocket for benzodiazepines (BZs) lies at the interface of the α and the γ subunit (Figure 3).

Subunit Assembly Rules for GABA$_A$ Receptors

GABA$_A$ receptor subunit combinations are partly governed by which cell types express which genes and partly by preferential partnering of subunits within a given cell; for example, the α4 and α6 subunits assemble preferentially with the δ subunit (Figure 4(a)). The majority of mammalian brain GABA$_A$ receptors are probably $\alpha\beta\gamma$2 combinations. The subunit ratio is probably $2\alpha/2\beta/1\gamma$. Some receptors also contain different α and β subunits, such as $\alpha1\alpha2\beta2\gamma2$. The $\alpha1\beta\gamma2$ combination is the most abundant GABA$_A$ receptor subtype in the brain (60% of total; Figure 4(b)).

Synaptic GABA$_A$ Receptors: $\alpha\beta\gamma$ Subunit Combinations and Anchoring Role of the γ2 Subunit and Gephyrin

Placing GABA$_A$ receptors at synapses requires specific proteins that interact directly or indirectly with the γ subunits. For example, targeting some GABA$_A$ receptor subtypes to GABAergic terminals involves the widely expressed microtubule-binding protein gephyrin. Gephyrin either helps convey some GABA$_A$ receptor subtypes to the synapse or anchors them there – this requires the γ2 subunit. Without the γ2 subunit, no GABA$_A$ receptors are found in synapses in the developing or adult hippocampus, and without gephyrin, much-reduced numbers of some synaptic GABA$_A$ receptor subtypes, especially those containing α2, are found; some receptor clusters, especially those containing the α1 subunit, persist in hippocampal gephyrin knockout neurons. Other γ subunits can replace synaptic targeting function of γ2; in γ2 knockout mice, GABA$_A$ receptors can be restored to hippocampal synapses by expressing the γ3 subunit by transgenic rescue. Some conserved sequence identity in the large TM3–TM4 intracellular loops of the γ subunits may indicate binding sites for parts of the synapse anchoring mechanism. Distributed cysteine residues are conserved in the γ subunit large intracellular loops but are absent from the β and δ subunits. Palmitoylation of these cysteine residues via a thioester bond plays some role in targeting γ subunit-containing receptors to the synapse. A surprising finding is that the TM4 region of the γ2 subunit is also involved in synaptic targeting, possibly by interacting with lipid rafts occurring in the synapse or by other membrane proteins. The γ2 TM4 is necessary and sufficient for postsynaptic clustering of GABA$_A$ receptors, whereas the cytoplasmic γ2 subunit domains are dispensable. In contrast, both the TM3–TM4 loop and the TM4 domain of the γ2 subunit contribute to efficient recruitment of

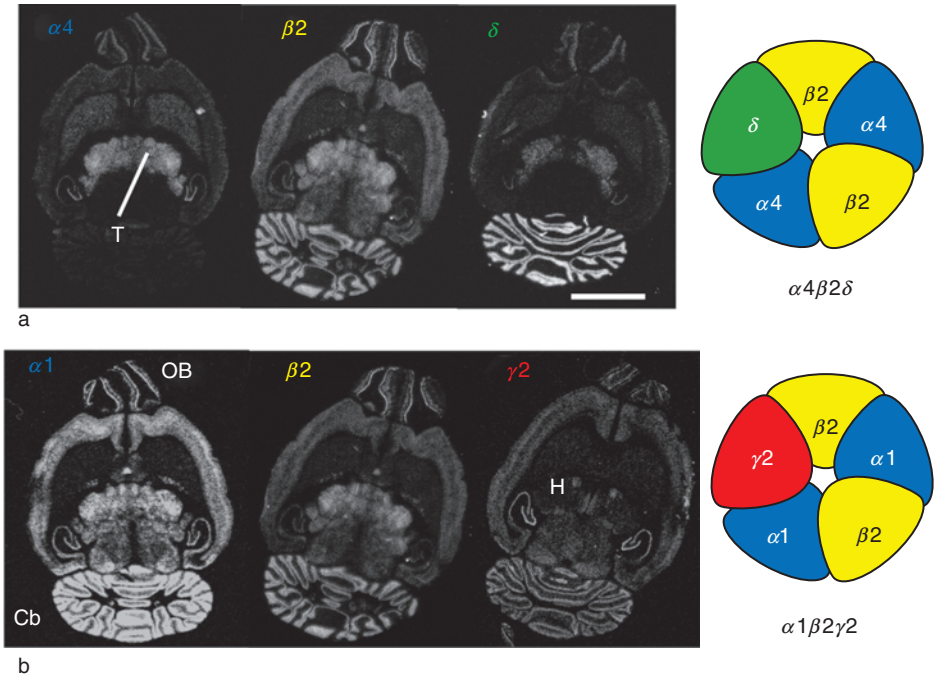

Figure 4 Some GABA$_A$ receptor subunit combinations found in adult rat brain: (a) $\alpha4\beta2\delta$; (b) $\alpha1\beta2\gamma2$. The gene expression is visualized by *in situ* hybridization (adult rat brain, horizontal sections). Cb, cerebellum; H, hippocampus; OB, olfactory bulb; T, thalamus. The $\alpha4\beta2\delta$ receptors are particularly prominent in the thalamus; the $\alpha1\beta2\gamma2$ subtype is widespread.

gephyrin to postsynaptic receptor clusters. Thus, the $\gamma2$ subunit TM3–TM4 cytoplasmic loop might be needed for inserting receptors into the plasma membrane but is dispensable for delivery of receptors to subsynaptic dendritic sites. Gephyrin does not bind the $\gamma2$ receptor subunit directly. The identity of the missing link(s) between gephyrin and GABA$_A$ receptor subunits is unknown.

GABA$_A$ Receptor Occupancy at Synapses Is Dynamic

GABA$_A$ receptor expression on the surface of neurons is dynamic; receptors rapidly recycle and leave from or insert into the synapse by lateral diffusion and/or endo/exocytosis. GABA$_A$ receptors diffuse into a synaptic zone and are transiently 'captured' by the anchoring complex. However, for some inhibitory hippocampal synapses, a direct relationship exists between the number of synaptic GABA$_A$ receptors and the strength of the synapse, but it is not clear what mechanisms maintain fixed numbers of GABA$_A$ receptors long-term at specific synapses. As for glutamate receptors at excitatory synapses, neurons probably recycle GABA$_A$ receptors as a strategy for setting their degree of excitability. GABA$_A$ receptors constitutively internalize by clathrin-dependent endocytosis; this requires interactions between the β and $\gamma2$ subunits and the AP2 adaptin complex.

Different Synaptic GABA$_A$ Receptor Subtypes Can Be Enriched on Different Domains of the Same Neuron

As mentioned above, considerable GABA$_A$ receptor complexity arises from differential subunit gene expression and specific subunit assembly rules. In addition, there are undefined determinants specifying that particular $\alpha\beta\gamma2$ subunit combinations are enriched at different postsynaptic locations on the same cell. An example of GABA$_A$ receptor complexity on a single cell type is provided by a hippocampal pyramidal cell (Figure 5). A pyramidal cell is covered with GABAergic terminals; a typical rat CA1 pyramidal cell receives around 1700 GABAergic synapses, with the highest density on the perisomatic region. Inhibition on different domains affects different aspects of pyramidal cell function. GABA$_A$ receptor-mediated inhibitory postsynaptic currents can be generated along the whole somatodendritic domain and on the axon-initial segment (AIS) of CA1 pyramidal cells. The apical dendritic trunk has a high density of GABAergic terminals relative to the rest of the dendrite; strong GABAergic stimulation onto this apical trunk region will isolate the dendritic compartment from the cell body. The inhibition on the more distal dendrites controls Ca^{2+} spike propagation. Inhibiting the AIS will powerfully clamp down the pyramidal cell's activity as this is where action potentials initiate – a typical rat CA1 pyramidal cell has more

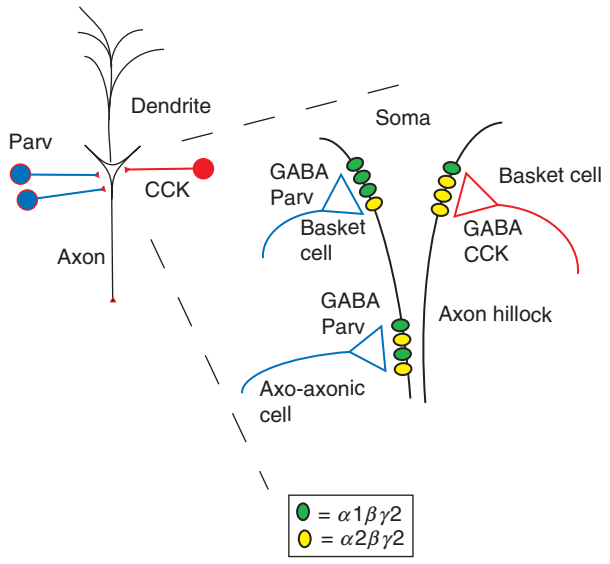

Figure 5 Enrichment of α1- and α2-subunit-containing αβγ2 GABA$_A$ receptors opposite different classes of GABAergic terminal (cholecystokinin (CCK)- or parvalbumin (Parv)-containing) on a hippocampal pyramidal cell. Note that for clarity, only a small subset of GABAergic interneurons that project onto any given pyramidal cell are shown.

than 25 GABAergic terminals per 50 μm of the AIS. Thus, GABA$_A$ receptors in the AIS deliver inhibition which controls the overall level of output activity of pyramidal cells (Figure 5).

GABA$_A$ receptors show differential subunit distribution at different synapses of the same pyramidal cell. On hippocampal pyramidal cells, the α2 subunit is enriched in the AIS but is present at only a minority of cell body synapses and synapses on dendrites. The AIS synapse also contains the α1 subunit and possibly the α4 and α5 subunits. GABA$_A$ receptors at the AIS are positioned to exert strong inhibitory influence over action potential generation.

Synapses formed by two types of presynaptic interneuron cells (GABAergic basket cells) on the soma of CA1 pyramidal cells contain distinct GABA$_A$ receptor subtypes: Only those synapses coming from parvalbumin-negative/CCK-positive basket cell interneurons contain the α2 subunit; synapses from parvalbumin-positive basket cells are often α2-immunonegative. This is a preferential targeting or enrichment of the α2 in a particular type of synapse – not an absolute selectivity. Both synapse types contain the same level of immunoreactivity for the β2/3 subunits; both types of synapse made by basket cells onto pyramidal cell somata also contain the α1 and the γ2 subunits. However, the α1 subunit is less abundant in the parvalbumin-negative pyramidal cell synapses than in the parvalbumin-positive ones (Figure 5).

Extrasynaptic GABA$_A$ Receptors: α4βδ and α6βδ Subtypes

Besides mediating precisely timed synaptic point-to-point inhibition (phasic inhibition) via γ2 subunit-containing receptors, GABA$_A$ receptors can convey less-time-locked signals. Low GABA concentrations in the extracellular space, resulting from GABA diffusing from the synapse, can tonically activate extrasynaptic GABA$_A$ receptors. This 'tonic inhibition' is temporally uncoupled from the fast synaptic events, causing a continually present background inhibitory conductance. Such conductances alter the input resistance of the cell and thus influence synaptic efficacy and integration; tonic extrasynaptic conductances, by increasing the electrical leakiness of the dendritic membrane, substantially and indiscriminately diminish the size of excitatory signals in dendrites (Figure 6).

Receptors with the δ subunit, α4βδ in forebrain and α6βδ in cerebellar granule cells, are extrasynaptic; δ subunits are perisynaptic (annular), localized around the edge of synapses in hippocampal dentate granule cells, and totally extrasynaptic on cerebellar granule cells. In all regions so far tested (cerebellar granule cells, hippocampal dentate granule cells, thalamic relay nuclei), δ subunits contribute to GABA$_A$ receptors that provide an extrasynaptic tonic conductance. For GABA$_A$ receptors containing α4βδ and α6βδ subunits, their key properties are high affinity for neurotransmitter and limited desensitization, enabling them to contribute to tonic background conductances (Figure 6).

GABA$_A$ Receptor Agonists, Antagonists, and Allosteric Modulators

GABA$_A$ receptors display a rich pharmacology. Generic GABA$_A$ receptors are selectively activated by the GABA agonist muscimol and blocked competitively by the GABA antagonists bicuculline and SR95531 (receptors assembled with ρ subunits are bicuculline and barbiturate insensitive, having their own unique pharmacology). Picrotoxin blocks GABA$_A$ receptors noncompetitively, probably by binding to a site in the channel. Many drugs bind at sites on the GABA$_A$ receptor distinct from the GABA binding site; these drugs change the shape of the receptor oligomer so that the efficacy of GABA at opening the channel is either increased (positive allosteric agonists, e.g., BZs such as diazepam) or decreased (negative allosteric agonists, e.g., β-carbolines). A few allosteric modulators occur naturally in the brain (e.g., Zn^{2+}, neurosteroids). Generally, positive allosteric agonists are used widely in medicine (e.g., for the induction and maintenance of general anesthesia or to treat anxiety disorders, states of agitation, epilepsy, or

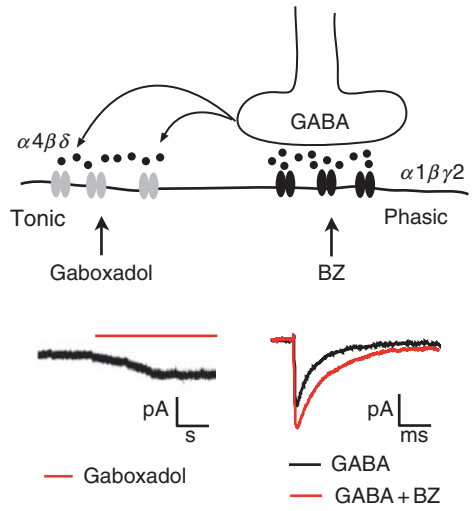

Figure 6 Synaptic and extrasynaptic GABA$_A$ receptors conferring tonic ($\alpha4\beta\delta$ or $\alpha4\beta\delta$) and phasic ($\gamma2$) inhibition, respectively. Synaptic ($\gamma2$-containing) GABA$_A$ receptors can be activated by BZs and BZ-like ligands (e.g., zolpidem). BZs act to increase the peak amplitude and slow the rate of decay of the Cl^- current (red trace). Synaptic transmission is on the millisecond timescale. Extrasynaptic (δ-containing) receptors can be selectively activated by Gaboxadol (red). Note the longer timescale.

sleep disorders), and there is scope to develop these drugs further to produce receptor subtype-selective drugs with fewer side effects; however, negative allosteric agonists also have potential clinical applications; for example, the drug L-655 708 works selectively at $\alpha5\beta\gamma2$ receptors (a subtype mainly expressed in the hippocampus), and by decreasing GABA's action there, it acts as a cognition enhancer. A feature of all allosteric modulators is that they usually work only when GABA is at submaximal activating concentrations (below 1 mmol l^{-1}), and they do not work in the absence of GABA (with the exception of some intravenous anesthetics). Nevertheless, some modulators (e.g., BZs) also strongly influence the deactivation rate of the receptors, even at peak synaptic GABA concentrations, and this may be how some of their *in vivo* effects originate. In the following sections, the drugs that act on particular GABA$_A$ receptor subunit combinations are considered in more detail.

GABA$_A$ Receptors: Allosteric Modulation by BZs and Related Ligands

The main effects of BZs are sedation, anxiolysis, suppression of seizures, and muscle relaxation. These drugs require $\alpha\beta\gamma2$-type receptors ($\alpha1\beta2\gamma2$, $\alpha2\beta2\gamma2$, $\alpha3\beta3\gamma2$, and $\alpha5\beta\gamma2$) with the drug-binding site located between the α and $\gamma2$ subunits (Figure 3(a)). Note that $\alpha4\beta\gamma2$- and $\alpha6\beta\gamma2$-type receptors are insensitive to most BZ drugs, as are any receptors that contain the δ subunit. The substances that act at the BZ binding site include the classical BZs like diazepam or flunitrazepam, as well as chemically different substances, such as the imidazopyridine zolpidem (relatively selective for $\alpha1\beta\gamma2$-type receptors), which is used as a sleeping pill. The BZ site is situated at the interface between the α and the γ subunit. In the $\alpha1$, $\alpha2$, $\alpha3$, and $\alpha5$ subunits, a mutation from histidine to arginine at position 101 abolishes binding of classic agonists like diazepam. The diazepam-insensitive $\alpha4$ (or $\alpha6$) subunits naturally contain an arginine residue at the homologous position, and so $\alpha4\beta\gamma2$ or $\alpha6\beta\gamma2$ receptors are insensitive to most BZ ligands.

In a beautiful series of studies, mice with specific mutations in the key H101 coding position affecting BZ sensitivity were generated in the $\alpha1$, $\alpha2$, $\alpha3$, and $\alpha5$ subunit genes, and the behavioral effects of diazepam were tested. These mice have normal GABA$_A$ receptors, but in α1H101R mice, for example, only the $\alpha2\beta\gamma2$-, $\alpha3\beta\gamma2$-, and $\alpha5\beta\gamma2$-type GABA$_A$ receptors are diazepam sensitive. Thus by a process of subtraction, it can be deduced how different $\alpha1\beta\gamma2$, $\alpha2\beta\gamma2$, $\alpha3\beta\gamma2$, and $\alpha5\beta\gamma2$ subtypes contribute to the diverse *in vivo* pharmacological effects of diazepam and other ligands requiring the H101 site. Thus, α1H101R mice no longer become sleepy when given diazepam, and so the $\alpha1\beta\gamma2$ receptors are required for the sedative effects of diazepam, whereas the $\alpha2$ (and $\alpha3$) mediates diazepam's anxiolytic effects (under the influence of diazepam, α2H101R mice do not venture more into threatening areas, whereas their wild-type littermates do). A different set of studies using $\alpha3$ selective agonists and inverse agonists also showed a significant contribution of the $\alpha3$ subunit in anxiogenesis and anxiolysis. The muscle relaxant activity of diazepam is mediated by the $\alpha2$ and $\alpha3$ subunits, probably because these subunits are expressed in spinal motor neurons.

New Subtype-Selective Drugs for GABA$_A$ Receptors That Work at the BZ Site

GABA$_A$ receptors have always been fertile ground for drug companies. BZs, although for many years the mainstay of clinical treatments for anxiety disorders, fell out of favor to selective serotonin reuptake inhibitors (SSRIs) because of side effects like sedation, cognitive impairment, and abuse liability. But SSRIs are too slow acting for some situations, requiring several weeks to work. Thus there is a medical need for fast-acting anxiolytics with few or no side effects. The knock-in mouse studies mentioned above suggest that $\alpha2$-containing GABA$_A$ receptors would be a prominent target for developing drugs which would be specifically anxiolytic.

$GABA_A$ Receptors: Allosteric Modulation by Intravenous Anesthetics

At clinically relevant concentrations, general anesthetics modulate the activity of various ion channels. Whereas volatile anesthetics (e.g., halothane, enflurane, or isoflurane) are positive modulators of recombinant $GABA_A$ receptors, the intravenous anesthetics (e.g., barbiturates, steroidal anesthetics, propofol, and etomidate) can modulate GABA's action at the receptor but can also activate the receptor directly in the absence of GABA at higher concentrations. Based on the analysis of knock-in mouse lines with propofol- and etomidate-insensitive β subunits (see below), propofol and etomidate exert nearly all their anesthetic actions entirely through $GABA_A$ receptors, whereas volatile anesthetics produce their actions through many diverse ion channel targets.

The action of etomidate and propofol requires residues in TM2 and TM3 in the $\beta2$ or $\beta3$ subunits. A mutation of asparagine to methionine at position 265 (N265M) in the second transmembrane domain of the $\beta3$ subunit abolishes the modulatory and direct effects of etomidate and propofol in recombinant receptors. A mutation of aspargine at the same position in the $\beta2$ subunit also abolishes the action of etomidate on the $GABA_A$ receptor. In $\beta3$(N265M) mice, propofol and etomidate do not suppress noxious-evoked movements and show a strongly decreased duration of the loss of righting reflex, two different endpoints of anesthesia. These results suggest that propofol and etomidate act mainly via the $GABA_A$ receptor and the $\beta3$ subunit in particular to induce deep anesthesia. The remaining effects of propofol and etomidate could be mediated by $\beta2$ subunit-containing receptors. Studies on $\beta2$(N265S) mice have suggested that the $\beta2$ subunit mediates the sedative effects of etomidate, whereas the $\beta3$ subunit is required for etomidate to induce a loss of consciousness. A highly interesting issue is the location in the brain where etomidate and propofol exert their anesthetic effects. Is the modulation of $GABA_A$ receptors in specific nuclei required to induce anesthesia or do these drugs produce global effects at many $GABA_A$ receptors in all brain circuits?

$GABA_A$ Receptors: Allosteric Modulation by Neurosteroids

Neuroactive steroids modulate $GABA_A$ receptor function in many brain regions. Naturally occurring steroid metabolites form locally in the brain: 5α-reductase transforms progesterone to 5α-DPH, which in turn is reduced by 3α-hydroxysteroid oxidoreductase to allopregnanalone. Allopregnanalone potently activates $GABA_A$ receptors. No absolute specificity of neurosteroids for particular $GABA_A$ receptor subunit combinations exists. Many $GABA_A$ receptors are sensitive to the steroid THDOC, but receptors with the δ subunit are particularly sensitive. Thus, endogenous allopregnanolone may act on extrasynaptic $\alpha4\beta\delta$ and $\alpha6\beta\delta$ $GABA_A$ receptors to increase basal levels of inhibition.

New Sleep-Promoting Ligands Acting as GABA Mimetics on Extrasynaptic $\alpha4\beta\delta$ Receptors

Modulating extrasynaptic $GABA_A$ receptors can produce profound effects on the nervous system (Figure 6). Many people suffer from insomnia. This is often treated with ligands, such as zolpidem, with a preferential affinity for the BZ site on synaptic $\alpha1\beta\gamma2$ receptors. By contrast, Gaboxadol is a selective extrasynaptic $GABA_A$ receptor agonist that has its greatest efficacy at $\alpha4\beta\delta$ and $\alpha6\beta\delta$ $GABA_A$ receptors, that is, BZ-insensitive receptors that contribute to tonic inhibitory conductances rather than synaptic inhibitory postsynaptic currents. It is important to emphasize that Gaboxadol, unlike BZs or related ligands, is an agonist at the GABA binding site itself – that is, it is a GABA mimetic, opening the receptor directly. Drugs which promote sleep acting through the BZ biding site of $\alpha1\beta\gamma2$-type receptors can cause various side effects such as disturbances in memory, rebound insomnia on drug withdrawal, and drug dependence. Gaboxadol acts on $\alpha4\beta\delta$ occurring exclusively extrasynaptically and enriched in particular neuronal pathways (e.g., thalamus – Figures 4(a) and 6), and thus has a different side effect profile. Sleep induced by increasing specifically the activity of extrasynaptic $GABA_A$ receptors may be of higher quality.

Conclusions

Many subtypes of $GABA_A$ receptor are used by the nervous system to impart inhibition. These $GABA_A$ receptors differ in their affinity for neurotransmitter and allosteric modulators, activation rate, desensitization rate, channel conductance, and location on the cell. But exactly what selective advantages are conferred by so many receptor subtypes to brain function is unclear and remains a fascinating research issue. In particular, the large number of $\alpha\beta\gamma$ subunit-containing synaptic receptor subtypes, often co-expressed on a single cell type such as a cortical pyramidal cell (Figures 2 and 5), might enhance the computational potential of the cell. Receptors at different synapses might have properties (e.g., deactivation rates) fine-tuned for individual inputs. In principle, evolutionary processes selected these subunit combinations

to be optimal, or at least to provide a reasonable compromise, for the inhibitory function required at each type of location. In some cases, however, multiple genes may exist because of the need to have complex transcriptional regulation that would have been difficult to organize from one gene promoter, rather than from the need to have different receptor properties. If one could engineer it, the hippocampus, or even the entire brain, might work perfectly well with just one type of $GABA_A$ receptor, such as an $\alpha1\beta2\gamma2$ combination. In practice, the $GABA_A$ receptor system remains an important target for new therapeutic approaches.

See also: GABA Synthesis and Metabolism; GABA-A Receptors: Developmental Roles; GABA-A Receptor Synaptic Functions; GABA-A Receptors and Disease; GABA-B Receptors: Molecular Biology and Pharmacology; GABA-B Receptor Function; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism.

Further Reading

Belelli D and Lambert JJ (2005) Neurosteroids: Endogenous regulators of the $GABA_A$ receptor. *Nature Reviews Neuroscience* 6: 565–575.

Buzsaki G (2006) Cycle 3: Diversity of cortical functions is provided by inhibition. In: Buzaki G (ed.) *Rhythms of the Brain*, pp. 61–79. New York: Oxford University Press.

Darlison MG, Pahal I, and Thode C (2005) Consequences of the evolution of the $GABA_A$ receptor gene family. *Cellular and Molecular Neurobiology* 25: 607–624.

Ernst M, Bruckner S, Boresch S, and Sieghart W (2005) Comparative models of $GABA_A$ receptor extracellular and transmembrane domains: Important insights in pharmacology and function. *Molecular Pharmacology* 68: 1291–1300.

Fang C, Dent L, Keller CA, et al. (2006) GODZ-mediated palmitoylation of $GABA_A$ receptors is required for normal assembly and function of GABAergic inhibitory synapses. *Journal of Neuroscience* 26: 12758–12768.

Farrant M and Nusser Z (2005) Variations on an inhibitory theme: Phasic and tonic activation of $GABA_A$ receptors. *Nature Reviews Neuroscience* 6: 215–229.

Fritschy JM and Brunig I (2003) Formation and plasticity of GABAergic synapses: Physiological mechanisms and pathophysiological implications. *Pharmacology & Therapeutics* 98: 299–323.

Hammond C (2001) The Ionotropic $GABA_A$ receptor. In: Hammond C (ed.) *Cellular and Molecular Neurobiology*, 2nd edn., pp. 227–250. San Diego, CA: Academic Press.

Korpi ER and Sinkkonen ST (2006) $GABA_A$ receptor subtypes as targets for neuropsychiatric drug development. *Pharmacology & Therapeutics* 109: 12–32.

Olsen RW and Betz H (2006) GABA and glycine. In: Siegel GJ, Albers RW, Brady ST, and Price DL (eds.) *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, 7th edn., pp. 291–301. Amsterdam: Academic Press.

Rudloph U and Mohler H (2004) Analysis of $GABA_A$ receptor function and dissection of the pharmacology of benzodiazepines and general anesthetics through mouse genetics. *Annual Review of Pharmacology and Toxicology* 44: 475–498.

Schofield PR, Darlison MG, Fujita N, et al. (1987) Sequence and functional expression of the $GABA_A$ receptor shows a ligand-gated receptor super-family. *Nature* 328: 221–227.

Seeburg PH, Wisden W, Verdoorn TA, et al. (1990) The $GABA_A$ receptor family: Molecular and functional diversity. *Cold Spring Harbor Symposia on Quantitative Biology* 55: 29–40.

Siegel E, Baur R, Boulineau N, and Minier F (2006) Impact of subunit positioning on $GABA_A$ receptor function. *Biochemical Society Transactions* 34: 868–871.

Simon J, Wakimoto H, Fujita N, Lalande M, and Barnard EA (2004) Analysis of the set of $GABA_A$ receptor genes in the human genome. *Journal of Biological Chemistry* 279: 41422–41435.

Somogyi P and Klausberger T (2005) Defined types of cortical interneurone structure space and spike timing in the hippocampus. *Journal of Physiology* 562: 9–26.

Unwin N (2005) Refined structure of the nicotinic acetylcholine receptor at 4A resolution. *Journal of Molecular Biology* 346: 967–989.

Wafford KA and Ebert B (2006) Gaboxadol: A new awakening in sleep. *Current Opinion in Pharmacology* 6: 30–36.

Whiting PJ (2006) $GABA_A$ receptors: A viable target for novel anxiolytics? *Current Opinion in Pharmacology* 6: 24–29.

Relevant Websites

http://www.ionchannels.org – Ionchannels.org.
http://www.ebi.ac.uk – Ligand-Gated Ion Channel Database.
http://www.nature.com – *Nature*: Supplement: Insight: Ion Channels.
http://en.wikipedia.org – Wikipedia articles on benzodiazepine and GABA.

$GABA_B$ Receptor Function

A J Doherty, G L Collingridge, and S M Fitzjohn, University of Bristol, Bristol, UK

© 2009 Elsevier Ltd. All rights reserved.

Postsynaptic γ-Aminobutyric Acid B-Mediated Inhibitory Postsynaptic Potential

Activation of postsynaptic γ-aminobutyric acid B ($GABA_B$) receptors produces a hyperpolarization of neuronal membrane potential, accompanied by a decrease in cell input resistance. Synaptic activation of GABA receptors elicits a biphasic inhibitory postsynaptic potential (IPSP), of which the early component is sensitive to $GABA_A$ receptor blockade, and inhibitors of $GABA_B$ receptors block the slower, late component of the IPSP (Figure 1). This late component typically has a time to peak of 50–250 ms and decay time of 100–500 ms. The difference in time course of these two components is due to the different nature of the two GABA receptors. Thus, $GABA_A$ receptors consist of a ligand-gated Cl^- channel with fast kinetics, whereas the $GABA_B$ component is via G-protein activation of a K^+ conductance.

The $GABA_B$ IPSP is mediated by $\beta\gamma$ G-protein subunits activating inwardly rectifying K^+ (Kir) channels, with efflux of K^+ causing membrane hyperpolarization (Figure 1). In the hippocampus, $GABA_B$ receptors couple to activation of the Kir3.2 subunit, and the $GABA_B$ receptor-mediated hyperpolarization of membrane potential is absent in Kir3.2 knockout mice. However, evidence suggests that a small component of the slow IPSP may, in some cases, also be mediated by activation of other K^+ channels, such as small-conductance Ca^{2+}-activated K^+ (SK) channels.

In hippocampal pyramidal neurons, there is a strong co-localization of $GABA_B$ receptors and the Kir3.2 subunit in dendritic spines, with both proteins being enriched around glutamate synapses. In the main dendritic shaft, however, the two proteins show a much more segregated distribution. Kir channels are coupled to G-protein-coupled receptors by membrane diffusion of the $\beta\gamma$ subunit, with effective coupling occurring if the two proteins are located within a 500 nm distance. Thus, it is likely that $GABA_B$-mediated activation of Kir and the associated membrane hyperpolarization are functionally much more important in dendritic spines than in the main dendritic shaft. GABAergic presynaptic fibers generally synapse onto the main dendritic shaft of hippocampal pyramidal neurons, suggesting that the activation of Kir by postsynaptic $GABA_B$ receptors is initiated by GABA spillover from nearby GABA terminals.

The slow time course of the $GABA_B$ IPSP and the localization of $GABA_B$ receptors in close vicinity to glutamate synapses have important implications for $GABA_B$ receptor function. Being closely located near the excitatory input to neurons, the $GABA_B$-mediated response limits postsynaptic excitation. One function of the $GABA_B$ receptor-mediated increase in K^+ conductance and membrane hyperpolarization is to reduce activity of the *N*-methyl-D-aspartate (NMDA) subtype of ionotropic glutamate receptor (NMDAR). The NMDAR has slow kinetics and also displays a voltage-dependent block of the receptor channel by Mg^{2+} ions at membrane potentials more negative than −35 mV. As GABA is released from feed-forward interneurons, which receive their synaptic drive from the same glutamatergic fibers that form excitatory synapses onto hippocampal pyramidal neurons, the dual $GABA_A$/$GABA_B$ IPSP is perfectly timed to prevent NMDAR activation during basal low-frequency activity. Activation of NMDARs is a key initiator of multiple forms of synaptic plasticity, which means that the $GABA_B$ receptor is an important factor in controlling the induction of synaptic plasticity.

Presynaptic Regulation of Neurotransmitter Release

At presynaptic sites, $GABA_B$ receptors inhibit neurotransmitter release at both GABA synapses (acting as autoreceptors) and other neurotransmitter synapses (acting as heteroreceptors, e.g., at glutamate and noradrenaline synapses). This regulation is predominantly caused by inhibition of N- and P/Q-type voltage-gated Ca^{2+} channels. The inhibition is mediated by the $\beta\gamma$ subunits of $GABA_B$ receptor-coupled G-proteins. Inhibition of transmitter release brought about by activation of presynaptic K^+ channels may also occur. Presynaptic $GABA_B$ receptors produce a powerful effect on transmitter release, being capable of inhibiting release by more than 90%.

Autoreceptors

$GABA_B$ receptors located on presynaptic GABA terminals act to decrease GABA release when two or more stimuli are delivered to presynaptic axons, leading to paired-pulse depression (PPD) of postsynaptic $GABA_A$ receptor-mediated responses (Figure 2). PPD occurs when pairs of stimuli are delivered in the frequency range of 0.1–50 Hz, with maximum PPD

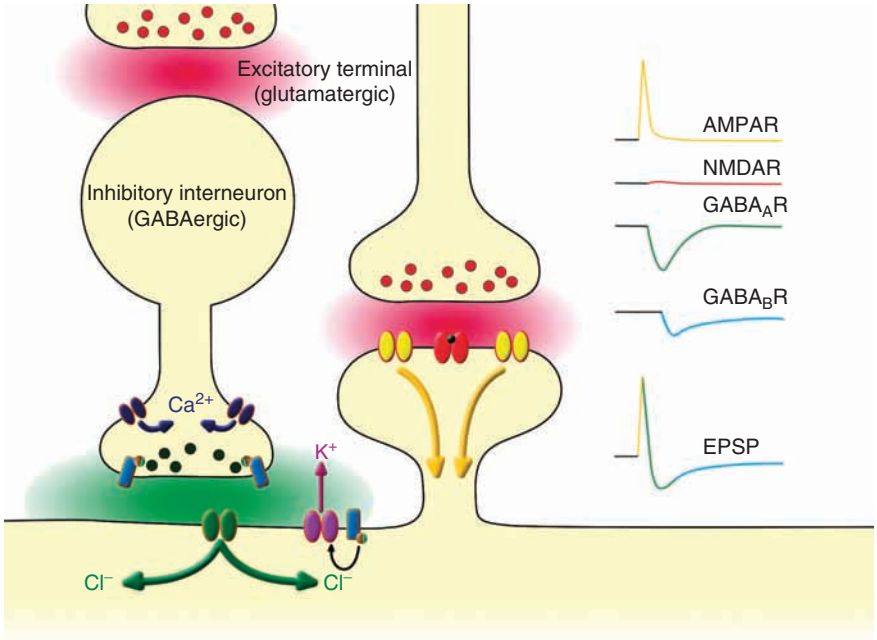

Figure 1 Postsynaptic γ-aminobutyric acid B (GABA$_B$) receptors produce a slow postsynaptic membrane hyperpolarization. Schematic diagram showing membrane potential changes produced by synaptic stimulation. Release of glutamate from excitatory synapses activates postsynaptic α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs), producing a fast excitatory postsynaptic potential (EPSP). Activation of GABAergic interneurons releases GABA that activates postsynaptic GABA$_A$ receptors, leading to Cl^- efflux and a fast inhibitory postsynaptic potential (IPSP). Spillover of GABA from the synapse activates GABA$_B$ receptors, where $\beta\gamma$ G-protein subunits activate nearby inwardly reactivating K^+ channels, producing a slow IPSP. This combined IPSP curtails the AMPAR-mediated EPSP and prevents activation of *N*-methyl-D-aspartate receptors (NMDARs). The traces show the change in membrane potential produced by activation of each receptor type. The color of the lines in these traces corresponds to the color of the receptor in the main diagram. The composite EPSP recorded in the postsynaptic cell is also shown.

seen at around 5–10 Hz. Depression of postsynaptic GABA$_A$ responses also occurs during trains of stimuli due to a depression of GABA release, although the effect of presynaptic GABA$_B$ inhibition decreases as the number of stimuli in the train increases or the frequency of the train increases (beyond around 10 Hz for a train of ten stimuli). With sufficient numbers of stimuli, the postsynaptic GABA$_A$ response undergoes activity-dependent facilitation and converts into a depolarizing response.

Role of GABA$_B$ Autoreceptors in the Induction of NMDAR-Dependent Long-Term Potentiation

The effect of decreasing postsynaptic GABA-mediated inhibition during short trains of activity is important for the induction of long-term potentiation (LTP) of excitatory glutamatergic synaptic transmission, which has been particularly well studied in the CA1 region of the hippocampus. LTP is a well-studied form of synaptic plasticity that occurs in many brain regions and is considered a synaptic model of learning. Many, but not all, forms of LTP rely on the activation of NMDARs to act as a trigger for LTP induction. As discussed above, the postsynaptic GABA$_B$ receptor-mediated IPSP acts to inhibit the induction of LTP during basal, low-frequency synaptic activity by preventing activation of the NMDAR. However, during short, high-frequency trains of activity, GABA release from presynaptic terminals is reduced by feedback of GABA onto presynaptic autoreceptors. The result is a decrease in postsynaptic GABA inhibition which, coupled to release from glutamate terminals, leads to a more persistent depolarization of the postsynaptic cell, activation of NMDARs, calcium influx, and the subsequent induction of LTP (Figure 2).

Multiple protocols exist for inducing LTP in experimental systems, and the role of GABA$_B$ receptors in LTP induction varies depending on the protocol used. A typical induction protocol is the delivery of brief high-frequency bursts (e.g., two or more trains of four shocks at 100 Hz) with an interburst frequency of around 5 Hz (i.e., an interburst interval of 200 ms). This is termed a theta burst stimulus train as the 5 Hz pattern of activity mimics the physiological activity that occurs during the theta rhythm, a frequency of activation observed in the EEG when animals are exploring their environment. Such theta burst trains are thus considered potential physiological induction protocols for LTP induction. This frequency corresponds to the near-maximal effect of presynaptic GABA$_B$ autoreceptors, and activity at GABA$_B$ receptors is essential for induction of LTP. Thus,

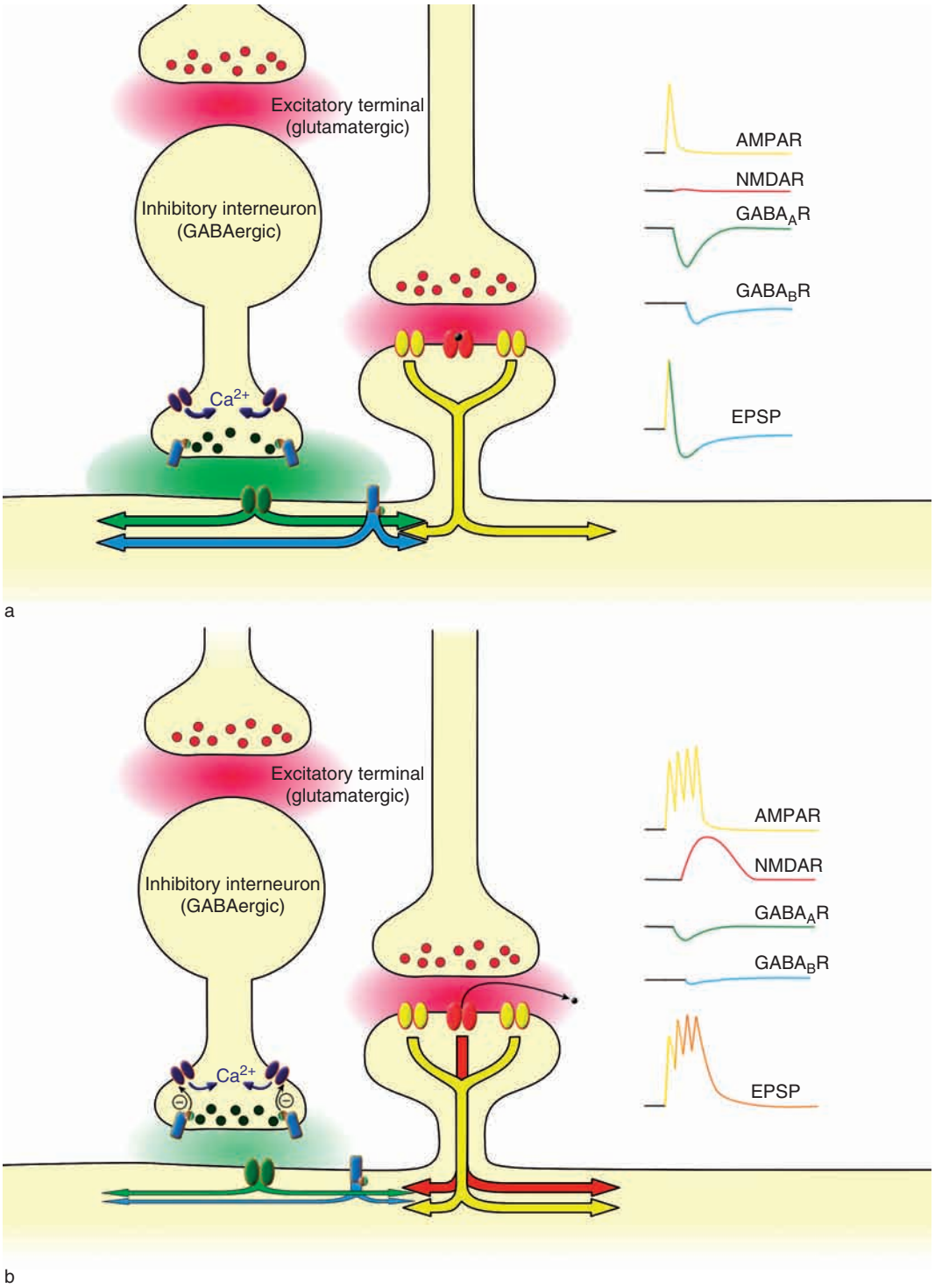

Figure 2 Activation of presynaptic γ-aminobutyric acid B (GABA$_B$) autoreceptors contributes to the induction of *N*-methyl-D-aspartate (NMDA) receptor-dependent LTP. (a) In response to a single stimulus, activation of postsynaptic GABA receptors produces a membrane hyperpolarization that curtails the excitatory postsynaptic potential (EPSP) and prevents NMDA receptor (NMDAR) activation, which remain blocked by Mg^{2+} ions (.). (b) In response to a train of stimuli, GABA activates presynaptic GABA$_B$ autoreceptors which decrease Ca^{2+} entry into the presynaptic bouton and thus decrease GABA release. This reduces the postsynaptic GABA-mediated inhibitory postsynaptic potential, which prolongs the AMPA receptor (AMPAR)-mediated membrane depolarization leading to expulsion of Mg^{2+} ions from the NMDAR and subsequent Ca^{2+} entry into the postsynaptic neuron. The inset traces show the membrane potential change produced by each constituent receptor and the composite EPSP recorded form the postsynaptic neuron.

$GABA_B$ receptor antagonists block the induction of LTP by theta burst stimuli. Likewise, LTP induced by the delivery of a single priming stimulus followed approximately 200 ms later by a single burst (which could comprise as few as two stimuli, though typically four are used) is blocked by $GABA_B$ receptor antagonists. In this case the priming stimulus allows activation of the presynaptic $GABA_B$ autoreceptors such that during the subsequent burst of stimuli, GABA release is depressed sufficiently to allow NMDA receptor activation.

LTP may also be induced by stronger stimulus trains, typically 100 stimuli delivered at a frequency of 100 Hz. Under these conditions, LTP induction is independent of $GABA_B$ receptor activity due to the conversion of the postsynaptic $GABA_A$ response into a depolarizing response.

Depression of Vesicle Recruitment

At the excitatory calyx of Held synapse in the brain stem, activation of presynaptic $GABA_B$ receptors by GABA spillover from nearby inhibitory synapses elicits heterosynaptic depression of glutamate by depression of Ca^{2+} influx, as seen at other synapses in the brain. However, presynaptic $GABA_B$ receptor activation has an additional effect at this synapse in that it slows vesicle recruitment after strong presynaptic stimulation. This effect is mediated by G-protein inhibition of adenylyl cyclase and thus a reduction in cyclic adenosine monophosphate (cAMP) production. Vesicle priming at this synapse is enhanced by cAMP and Ca^{2+}/calmodulin, and thus a reduction in cAMP levels slows the recruitment of vesicles ready for release.

LTP of Postsynaptic $GABA_B$ Responses

In addition to being involved in the induction of LTP of excitatory synaptic transmission, postsynaptic $GABA_B$ receptors may also undergo LTP. Thus in the hippocampus, NMDAR activation can lead to an enhancement of the slow, $GABA_B$ receptor-mediated IPSC in CA1 hippocampal pyramidal neurons. This form of LTP requires influx of Ca^{2+} through the NMDA receptor and the subsequent activation of Ca^{2+}/calmodulin dependent protein kinase II, although the precise mechanism of the enhancement of $GABA_B$ receptor function is unclear. It remains to be seen whether other forms of plasticity exist to modulate $GABA_B$ receptor function.

Coupling of $GABA_B$ and Metabotropic Glutamate Receptors

In the cerebellar cortex, $GABA_B$ receptors are expressed in Purkinje cells (PCs) at the extrasynaptic site of excitatory input from parallel fibers. This extrasynaptic localization is also seen for the metabotropic glutamate receptor subtype 1 (mGluR1), stimulation of which produces a slow excitatory postsynaptic current and mobilization of Ca^{2+} from intracellular stores. Activation of postsynaptic $GABA_B$ receptors by agonists or synaptic GABA release produces an enhancement of mGluR1-mediated synaptic currents and Ca^{2+} mobilization, an effect that is dependent on $G_{i/o}$ G-protein signaling, activation of phospholipase C, and release of Ca^{2+} from intracellular stores.

It is interesting that facilitation of mGluR1 in PCs by $GABA_B$ receptors can also occur independently of classical receptor activation. In the absence of $GABA_B$ receptor agonists, increases in extracellular Ca^{2+} elicit a constitutive increase in the sensitivity of mGluR1 to glutamate, facilitating mGluR1 effects. This effect of extracellular Ca^{2+} is independent of $G_{i/o}$ activity and thus differs from the enhancement described above. $GABA_B$ and mGluR1 receptors appear to exist in a protein complex in PCs, and thus this enhancement may be due to changes in protein complex structure rather than to G-protein signaling.

Therapeutic Potential of $GABA_B$ Receptor Ligands

Genetic deletion of either of the two $GABA_B$ receptor subunits or pharmacological blockade of native $GABA_B$ receptors produces spontaneous epileptic seizures, demonstrating the important role that $GABA_B$ receptors have in preventing overexcitation of neuronal networks and maintaining normal brain function. As such, it is possible that dysfunction of $GABA_B$ receptor activity may be implicated in human epilepsies, and polymorphisms of the $GABA_{B1}$ receptor subunit are linked to increased susceptibility to temporal lobe epilepsy. In contrast, antagonists of $GABA_B$ receptors suppress absence seizures seen in animal models, whereas agonists promote seizures in these animals, suggesting that antagonists may be of potential therapeutic use in human absence seizures. However, in these models, high doses of $GABA_B$ receptor antagonists are prone to induce convulsions, highlighting the difficulty in targeting the $GABA_B$ receptor to treat this disease, in particular because of the widespread distribution of these receptors and the lack of receptor subtypes.

The $GABA_B$ receptor agonist baclofen is currently used to treat spasticity and muscle rigidity in patients with multiple sclerosis and spinal cord injury. The effect of baclofen is thought to be due to activation of presynaptic heteroreceptors in the spinal cord, reducing the release of excitatory transmitters. Reduction of transmitter release in the spinal cord (particularly glutamate and substance P) probably also contributes to the antinociceptive effect of baclofen administration, although additional actions in the

brain, such as in the rostral agranular insular cortex, likely also contribute to it. $GABA_B$ receptor subunit knockout animals display heightened sensitivity to painful stimuli such as heat, confirming the role of $GABA_B$ receptors in modulating pain pathways.

A potential use of $GABA_B$ receptor agonists is in the treatment of drug addiction. Thus, agonists such as baclofen can reduce craving seen with drugs such as alcohol, cocaine, morphine, and nicotine. In addition, genetic studies suggest that there is a link between $GABA_B$ receptor variants and potential for addiction. This action of $GABA_B$ receptor activation in reducing addiction probably relates to activity in the mesolimbic dopamine system, which is part of the reward and reinforcing pathways of the brain. A further postulated use of activating $GABA_B$ receptors is to treat anxiety as $GABA_B$ receptor subunit knockout mice show increased anxiety in behavioral tests and baclofen has anxiolytic effects.

See also: GABA Synthesis and Metabolism; GABA-A Receptors: Developmental Roles; GABA-A Receptor Synaptic Functions; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-B Receptors: Molecular Biology and Pharmacology; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism; Long-Term Potentiation (LTP): NMDA Receptor Role; NMDA Receptor Function and Physiological Modulation.

Further Reading

Bettler B, Kaupmann K, Mosbacher J, and Gasmann M (2004) Molecular structure and physiological functions of $GABA_B$ receptors. *Physiological Reviews* 84: 835–867.

Bowery NG, Hill DR, Hudson AL, et al. (1980) (−)Baclofen decreases neurotransmitter release in the mammalian CNS by an action at a novel GABA receptor. *Nature* 283: 92–94.

Davies CH, Davies SN, and Collingridge GL (1990) Paired-pulse depression of monosynaptic GABA-mediated inhibitory postsynaptic responses in rat hippocampus. *Journal of Physiology* 424: 513–531.

Davies CH, Starkey SJ, Pozza MF, and Collingridge GL (1991) GABA autoreceptors regulate the induction of LTP. *Nature* 349: 609–611.

Dutar P and Nicoll RA (1988) A physiological role for $GABA_B$ receptors in the central nervous system. *Nature* 332: 156–158.

Hirono M, Yoshioka T, and Konishi S (2001) GABA(B) receptor activation enhances mGluR-mediated responses at cerebellar excitatory synapses. *Nature Neuroscience* 4: 1207–1216.

Huang CS, Shi SH, Ule J, et al. (2005) Common molecular pathways mediate long-term potentiation of synaptic excitation and slow synaptic inhibition. *Cell* 123: 105–118.

Kornau HC (2006) GABA(B) receptors and synaptic modulation. *Cell Tissue Research* 326: 517–533.

Kulik A, Vida I, Fukazawa Y, et al. (2006) Compartment-dependent colocalization of Kir3.2-containing K^+ channels and $GABA_B$ receptors in hippocampal pyramidal cells. *Journal of Neuroscience* 26: 4289–4297.

Newberry NR and Nicoll RA (1984) Direct hyperpolarizing action of baclofen on hippocampal pyramidal cells. *Nature* 308: 450–452.

Sakaba T and Neher E (2003) Direct modulation of synaptic vesicle priming by GABA(B) receptor activation at a glutamatergic synapse. *Nature* 424: 775–778.

Soltesz I, Haby M, Leresche N, and Crunelli V (1988) The $GABA_B$ antagonist phaclofen inhibits the late K^+-dependent IPSP in cat and rat thalamic and hippocampal neurones. *Brain Research* 448: 351–354.

Tabata T, Araishi K, Hashimoto K, et al. (2004) Ca^{2+} activity at $GABA_B$ receptors constitutively promotes metabotropic glutamate signaling in the absence of GABA. *Proceedings of the National Academy of Sciences of the United States of America* 101: 16952–16957.

$GABA_B$ Receptors: Molecular Biology and Pharmacology

N G Bowery, GlaxoSmithKline, Verona, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

γ-Aminobutyric acid (GABA) is the single most important inhibitory neurotransmitter within the mammalian brain, in which it is estimated that 40% of all inhibitory synaptic activity is mediated through this neutral amino acid. GABA can activate both ionotropic ($GABA_A$) and metabotropic ($GABA_B$) receptors, and these are present inside and outside the brain of mammals and also evident in lower species. This article focuses on the metabotropic receptor, $GABA_B$.

The receptor was originally defined in the 1980s by pharmacological characterization in isolated tissue preparations. The structure of the receptor was identified approximately 18 years later.

Unlike $GABA_A$ receptors, $GABA_B$ sites do not affect Cl^- membrane conductance in neurons but instead their activation increases K^+ conductance and decreases Ca^{2+} conductance. These effects tend, although not exclusively, to be site directed such that the decrease in Ca^{2+} is more associated with presynaptic receptors, whereas K^+ effects are predominantly postsynaptic. As a consequence of these effects, receptor activation can produce neuronal hyperpolarization (K^+ effect) or a decrease in evoked neurotransmitter release (Ca^{2+} effect). It is this latter effect that was the basis of the first observed characteristic of $GABA_B$ receptors. GABA dose dependently reduced the electrically evoked release of tritiated noradrenaline from rat isolated atria. This effect was neither blocked by recognized GABA antagonists nor mimicked by established GABA agonists. Conversely, baclofen (β-chlorophenyl GABA), which is inactive at chloride-dependent GABA receptors, was able to mimic the action of GABA in the atrial preparation and in other $GABA_B$ receptor systems. The subsequent development of a membrane receptor binding assay firmly established the existence of the metabotropic GABA receptor in mammalian brain tissue.

Receptor Structure

$GABA_B$ receptors exist naturally as heterodimers and comprise subunits designated $GABA_{B1}$ and $GABA_{B2}$ (Figure 1). Although the subunits exhibit 35% homology, they have quite distinct characteristics. Whereas $GABA_{B1}$ contains the binding domain for GABA in its extracellular N-terminal, the $GABA_{B2}$ subunit appears to be responsible for engaging and activating the G-protein-coupled signaling system. $GABA_{B1}$ and $GABA_{B2}$ must remain linked as a dimer after insertion into the cell membrane to fully maintain receptor function. $GABA_{B1}$ subunits remain associated with the endoplasmic reticulum, and only when they link with $GABA_{B2}$ is the dimer transported to the cell membrane. Each subunit has a seven transmembrane spanning domain and they are linked via their C-termini through coiled-coil structures. A number of splice variants of the $GABA_{B1}$ subunit have been identified, namely $GABA_{B1a–1g}$ (Figure 1, inset), but not all have been shown to have a physiological function. Only $GABA_{B1a}$ and $GABA_{B1b}$ have been implicated in the functional receptor, and although $GABA_{B1c}$ has been reported to be functional *in vitro*, evidence for an *in vivo* role is equivocal. The 1a and 1b proteins differ significantly in their amino acid sequences within the N-terminal; 1a has a unique sequence of 162 amino acids, whereas 1b has a shorter unique sequence of 47 amino acids. The distribution patterns of these two subunits within the rat brain indicate that they have different roles, although these have yet to be established. For example, within the spinal cord, the level of $GABA_{B1a}$ is reported to be tenfold higher than that of $GABA_{B1b}$ within the primary afferent terminals, and the $GABA_{B1b}$ subunit appears to predominate at postsynaptic sites in the cerebellum. However, elsewhere in the brain, $GABA_{B1b}$ appears to predominate in presynaptic terminals, whereas $GABA_{B1a}$ is associated with postsynaptic sites, for example, within the thalamocortical system.

There is no unequivocal evidence to support the existence of variants of the $GABA_{B2}$ subunit, although two variants of the human $GABA_{B2}$ subunit have been proposed. However, these were probably artifacts of the isolation process.

The expression of $GABA_{B1}$ and that of $GABA_{B2}$ appear to be independent of each, even though there is a 1:1 stoichiometric relation between the two proteins. Invariably, $GABA_{B2}$ is expressed more abundantly than $GABA_{B1}$ under experimental pathological conditions.

Protein–Protein Interactions

The possibility of a mismatch between $GABA_{B1}$ and $GABA_{B2}$ levels raises the possibility that other proteins might form a dimer with the $GABA_B$ subunit to form a functional receptor. Numerous proteins have been described which interact with either $GABA_{B1}$ or $GABA_{B2}$, and some of these have such high affinity that they are able to compete with one $GABA_B$

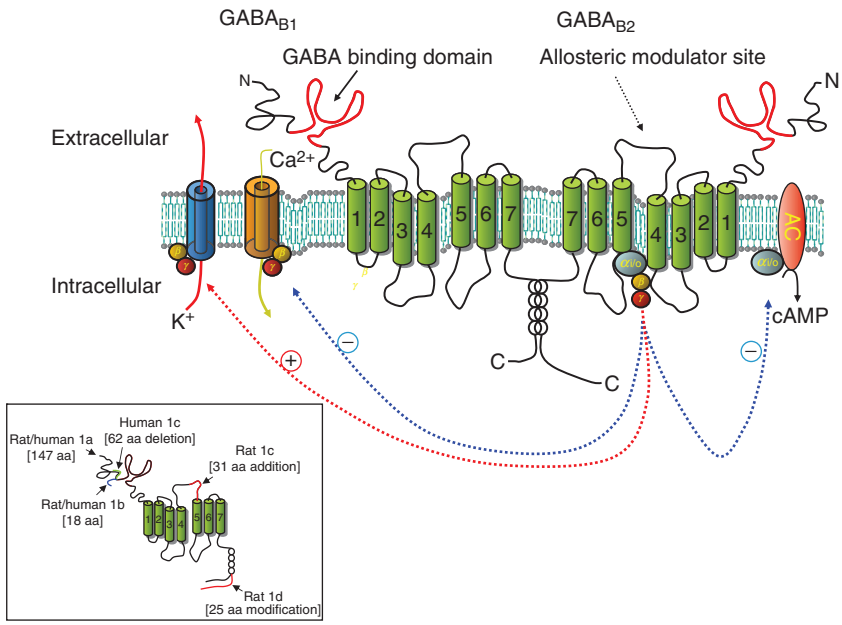

Figure 1 Diagram of the heterodimer that forms the GABA$_B$ receptor. The receptor comprises two subunits, GABA$_{B1}$ and GABA$_{B2}$, which are inserted into the plasma membrane after having been coupled intracellularly. GABA$_{B1}$ contains the GABA binding domain in its extracellular N-terminal chain, whereas GABA$_{B2}$ does not bind GABA but has a modulatory site within the transmembrane region which appears to be specific to this subunit. This subunit also seems to be responsible for coupling to the second messenger G-protein system. Receptor activation manifests as an increase in K^+ conductance, a decrease in Ca^{2+} conductance, and/or an inhibition of adenylate cyclase (AC). The inset indicates the primary isoforms of GABA$_{B1}$. Courtesy of Stefano Tacconi and Rachel Ginham.

subunit for binding to the other subunit. However, no evidence of the formation of a functional receptor from any protein–protein interaction has been reported. Thus, if the interaction is not producing a receptor, what purpose(s) does this protein coupling serve? It might be to provide a directional mechanism to enable the correct site insertion of the dimer into the plasma membrane or merely to act as a scaffold or anchoring device to support the dimer when inserted. Alternatively, because of the high affinity of these proteins, they might regulate the formation of the dimer, limiting the level of functional receptor, or by interacting intracellularly with the formed dimer, they may reduce its cell surface expression. The exact role of these interacting proteins, including CREB2/ATF4, 14-3-3, fibulin-2, CHOP, and Marlin-1, has yet to be established.

Receptor Subtypes

Despite the presence of structural variants of the GABA$_{B1}$ subunits and the possibility of functional modifications of the receptor heterodimer by interacting proteins, there is still no unequivocal evidence for the existence of functional GABA$_B$ receptor subtypes. Pharmacological differences between autoreceptors and heteroreceptors as well as between the dual actions of GABA$_B$ agonists on adenylyl cyclase activity have been suggested by individual research groups. However, both a lack of reproducibility and sufficient distinction between the effects and the failure to emulate observations made by other groups have cast doubt on the existence of functional receptor subtypes. In general, the available receptor agonists and antagonists do not distinguish between potential subtypes of the GABA$_B$ receptor. Nevertheless, it has been suggested that gabapentin is a GABA$_B$ agonist which selectively activates the form of the receptor comprising a combination of GABA$_{B1a}$ and GABA$_{B2}$ and expressed in isolated *Xenopus* oocytes. However, this could not be confirmed by others, especially in *in vivo* systems. Thus, although this was an exciting and potentially important observation, the failure of others to reproduce the original findings has tempered enthusiasm not only for understanding the mechanism of action of this anticonvulsant/analgesic agent but also for the possibility of relating receptor structure to functional diversity. A further lack of support for receptor subtyping derives from studies performed on developmental knockout mice in which either GABA$_{B1}$ or GABA$_{B2}$ subunits have been deleted. No supporting evidence for any receptor subtyping was observed in these animals.

Receptor Distribution

GABA$_B$ receptors are distributed throughout the mammalian system and are not confined to nervous tissue.

Although their function is probably of greatest physiological importance within the central nervous system, their presence in peripheral organs is well established. In fact, it was their presence in the heart atrium of the rat that led to their discovery. They are present on autonomic nerve terminals, and when activated by GABA, this decreases the evoked release of transmitter from the terminal. Of course, this is only of physiological importance when the receptor agonist is present, and in the periphery this is unlikely to occur at most sites. However, in the enteric nervous system, in which GABA-releasing neurons are present and impinge on autonomic cholinergic nerve fibers, $GABA_B$ receptors probably contribute to the control of intestinal movement and sphincters. Individual $GABA_{B1}$ and $GABA_{B2}$ subunits have been detected in peripheral tissue, but their distribution does not always appear to be coincident, for example, in the uterus and spleen. In general, there appears to be a paucity of $GABA_{B2}$, suggesting that another protein may form a dimer with $GABA_{B1}$ if the receptor is to function as described elsewhere. There is also a differential distribution of the $GABA_{B1}$ isoforms such that $GABA_{B1a}$ is in, for example, the adrenals and prostate of the rat, whereas $GABA_{B1b}$ is the form present in the kidney.

The distribution of $GABA_B$ sites in the mammalian brain is quite wide, although there are regional variations. Receptor autoradiography has shown that the highest densities of the receptor are in the thalamic nuclei, cerebellum (molecular layer), cerebral cortex, interpeduncular nucleus, and dorsal horn of the spinal cord. Moderate to low levels are present elsewhere, but this does not necessarily reflect the physiological importance of the receptor in those regions. For example, in the hippocampus the significance of $GABA_B$ receptors in neural transmission may not be matched by the overall density of receptors. The distribution patterns of the individual receptor subunits, as determined by immunocytochemistry, generally match each other, supporting the concept of heterodimer receptors. However, in the caudate putamen $GABA_{B2}$ is not detectable even though $GABA_{B1}$ and the native receptor are present, which could support the notion of another, as yet undefined, protein subunit.

Physiological Significance

The physiological relevance of the $GABA_B$ receptor was first established within the hippocampus, where it was shown to be responsible for generating the late hyperpolarization associated with orthodromic synaptic transmission in Ammon's horn. Subsequent studies provided evidence for the same pattern of activity within other brain regions, such as the lateral geniculate nucleus.

The mechanism underlying this neuronal hyperpolarization is an increase in membrane K^+ conductance mediated by activation of postsynaptic $GABA_B$ receptors. However, this is not the only mechanism through which $GABA_B$ activation occurs. The predominant location of the receptors appears to be presynaptic, where they function as autoreceptors and heteroreceptors to limit the release of a variety of neurotransmitters. A reduction in membrane Ca^{2+} conductance is produced by activation of $GABA_B$ sites, which reduces the cytosolic concentration of Ca^{2+} to inhibit transmitter release. Whereas the physiological significance of autoreceptors seems straightforward because of the local availability of neurotransmitter, the role of heteroreceptors on non-GABAergic terminals is not so obvious because evidence for axo-axonic contacts at nerve terminals is lacking. The only substantiated evidence is at primary afferent terminals in the spinal cord, where GABA interneurons innervate the terminal regions of the primary afferent fibers. However, despite the apparent lack of innervation, there is strong evidence for GABA acting in a 'paracrine' manner to activate $GABA_B$ heteroreceptors. GABA released from GABAergic terminals 'washes over' onto adjacent terminals where $GABA_B$ sites are present. The estimated concentration of GABA in the cleft, following synaptic release, is approximately 0.01 M, and because the affinity of GABA for heteroreceptors is approximately 0.000 000 01 M, sufficient GABA should be available to interact with sites in close proximity. Heteroreceptors are therefore of physiological as well as pharmacological importance.

Effector Mechanisms

As indicated previously, more than one signaling system mediates the response to $GABA_B$ receptor activation. Either an increase in K^+ or a reduction in Ca^{2+} membrane conductance provides the neuronal response, and both of these events are mediated by G-proteins that are members of the pertussis toxin-sensitive family $G_{i\alpha}/G_{o\alpha}$. However, presynaptically mediated events, which are generally associated with reduced Ca^{2+} conductance, appear to be less sensitive to pertussis toxin. The predominant calcium channel linked to $GABA_B$ sites appears to be the 'N' type, although 'P' and 'Q' type channels are also implicated. Multiple K^+ channel types seem to be associated with postsynaptic $GABA_B$ receptors.

The third signaling system associated with $GABA_B$ sites is adenylyl cyclase, which is normally inhibited by receptor activation but if the enzyme has been activated by G_s-coupled receptor agonists such as isoprenaline, the β-adrenoceptor ligand, $GABA_B$

receptor activation enhances the formation of cAMP above the level achieved with isoprenaline alone. This observation prompted a search for receptor subtypes by comparing the abilities of different GABA$_B$ receptor agonists to inhibit or enhance cAMP production. However, no clear separation has been established.

GABA$_B$ Receptor Ligands

Agonists

One of the original criteria for establishing the presence of GABA$_B$ receptors was to show that β-[4-chlorophenyl] GABA (baclofen) was a stereospecific agonist. Subsequently, other agonists, such as 3-aminopropyl-phosphinic acid (3-APPA) and 3-aminopropyl-methyl-phosphinic acid (3-APMPA), emerged which were approximately tenfold more potent than R-(−)-baclofen, the active isomer, at GABA$_B$ receptor binding sites (Figure 2). *In vivo* studies have also indicated that 3-APMPA has greater brain penetration than either baclofen or 3-APPA. The agonist CGP44532, which has an affinity comparable to that of (−) baclofen for GABA$_B$ receptors (Figure 2), is reported to be selective for GABA$_B$ autoreceptors.

Antagonists

The first selective GABA$_B$ antagonists to be described were phaclofen, saclofen, and 2-hydroxy saclofen (Figure 2). These have only low affinity for the receptor, with p*K*i values of 4–5. However, phaclofen was used to great effect in obtaining the first evidence in 1988 for the physiological role of GABA$_B$ receptors in synaptic transmission within the rat hippocampus. Subsequent studies led to the introduction of a compound that was able to cross the blood–brain barrier (CGP 35348) and the first orally active agent (CGP 36742). However, both of these compounds have low affinities, as does the chemically distinct antagonist SCH50911 (Figure 2). A major breakthrough was obtained when 3,4-dichlorobenzyl or 3-carboxybenzyl substituents were attached to the existing molecules to produce a variety of compounds with affinities in the low nanomolar range, such as CGP 55845 and many others, which in all cases contain a phosphinic acid moiety. Ultimately, such antagonists, when radiolabeled with ^{125}I, provided photoaffinity ligands suitable for the successful elucidation of the structure of the GABA$_B$ receptor.

Effects of GABA$_B$ Receptor Activation

Receptor activation within the mammalian brain can produce a variety of effects as a consequence of

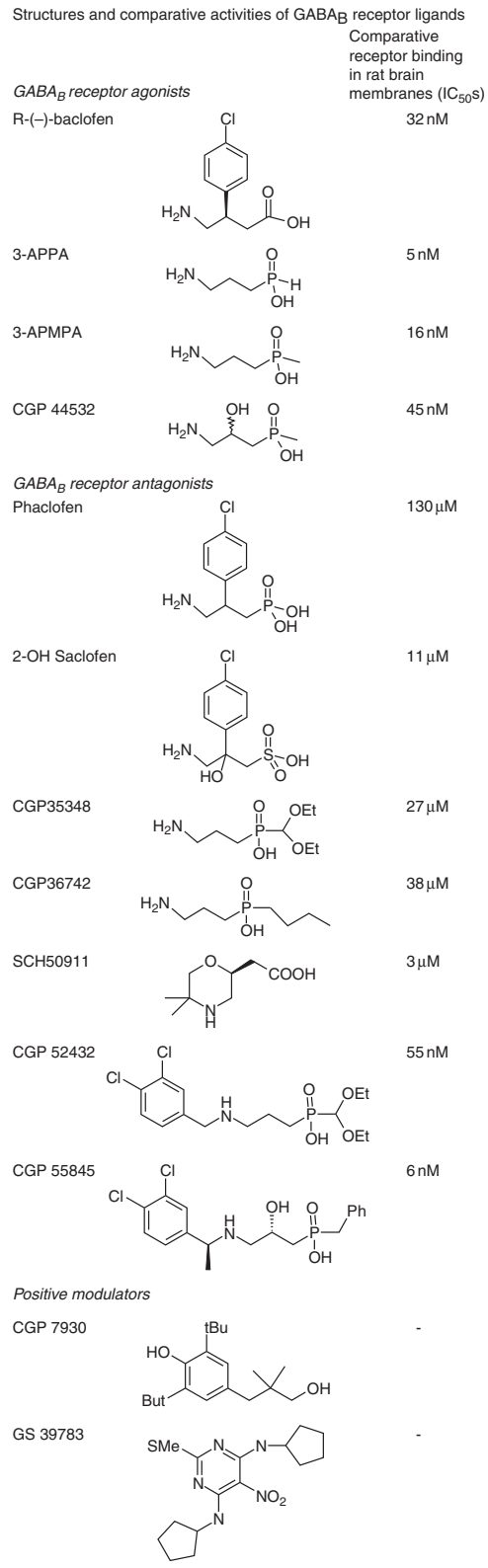

Figure 2 Selection of GABA$_B$ receptor ligands and their comparative potencies at the native receptor in rat brain cerebral cortex membranes. Note the radical increase in potency of the examples of the dichlorobenzyl-substituted antagonists, CGP52432 and CGP55845.

Table 1 $GABA_B$ receptor activation

Effect	*Site of action*
Antinociception	Spinal cord, thalamus
Antitussive action	Cough center in medulla
Drug addiction suppression	CNS–mesolimbic system
Enhanced feeding	Higher centers
Fat intake reduction	Higher centers
Gastrin/gastric acid secretion altered	Vagal center
Generation/exacerbation of absence epilepsy	Thalamus/somatosensory cortex
Inhibition of neurotransmitter release	CNS and peripheral nerve terminals
Inhibition of cognitive function	Higher centers
Insulin/glucagon release	Pancreas
Modulation of long-term potentiation	CNS, hippocampus
Muscle relaxation	Spinal cord
Smooth muscle relaxation	Lung, bladder, intestine
Smooth muscle contraction	Uterus, oviduct, gall bladder
Neuronal hyperpolarization	CNS (numerous locations)
Neutrophil chemotaxis enhanced	Leukocytes
Respiratory depression	Brain stem
Suppression of CRH/MSH release	Pituitary
Suppression of panic behavior	Dorsal periaqueductal gray
Vasopressor action	Nucleus tractus solitarius

inhibition of transmitter release and/or postsynaptic neuronal hyperpolarization. Information derived from knockout mice, which exhibit hyperalgesia, seizures, hyperlocomotion, impaired learning, loss of responses to baclofen, and lack of $GABA_B$ binding sites throughout the brain, together with the actions of $GABA_B$ agonists and antagonists have provided the basis for defining the physiological role of the receptor as well as indicating the potential therapeutic benefits of receptor ligands. Some of the actions of baclofen (*in vitro* and *in vivo*) are shown in Table 1, and predominant among the *in vivo* effects are the muscle relaxant, antinociceptive, anti-drug-craving effects and reduction in cognitive behavior.

Potential Therapeutic Uses of $GABA_B$ Receptor Ligands

Skeletal Muscle Relaxation

The centrally mediated muscle relaxant properties of baclofen make it the drug of choice for spasticity associated with cerebral palsy, multiple sclerosis, stiff-man syndrome, and tetanus. However, the side effects produced by this drug, which include seizures, nausea, drowsiness, dizziness, hypotension, muscle weakness, hallucinations, and mental confusion, are often poorly tolerated by patients. These are, in part, due to the need for high doses to be administered to allow for poor brain penetration. This problem has been addressed by the introduction of intrathecal administration using an indwelling pump inserted into the peritoneal cavity. Administration in this way, directly to the site of action within the spinal cord, requires only local concentrations that are too low to appear in the systemic circulation, thus avoiding the production of adverse effects. The reduced extracellular level of the agonist also decreases the possibility of receptor desensitization. The response to systemic administration of baclofen is reduced after chronic treatment, but this does not seem to be the case after chronic intrathecal infusion.

Nociception

Baclofen and other $GABA_B$ receptor agonists have antinociceptive activity in acute pain models, such as the tail flick and hot plate tests in rodents. This occurs at doses below the threshold for muscle relaxation, enabling impairment of locomotor activity to be excluded as a confounding reason for the effect. This antinociceptive action stems, in part, from a reduction in the release of nociceptive transmitter from primary afferent fibers within the dorsal horn of the spinal cord. In addition, action within higher centers, particularly the thalamus, also contributes. In spinal cord slices from control rats, the application of $GABA_B$ receptor antagonists produces little or no increase in the evoked release of transmitter, but if the antagonist is applied to a cord slice from a monoarthritic rat, an evoked release of primary afferent transmitter occurs. If the antagonist is administered *in vivo* to rats with the same lesion, significant hyperalgesia occurs. This contrasts markedly with a lack of effect in control rats. These results suggest that an increase in $GABA_B$ innervation to primary afferent terminals occurs during chronic inflammation, and this acts as a pathological antinociceptive process to decrease the enhanced sensory input. The use of baclofen as an analgesic in humans has been very limited, presumably due, in part, to rapid tolerance and adverse effects following systemic administration.

As mentioned previously, the spinal cord is not the only site where $GABA_B$ agonists exert their antinociceptive effect. Focal injections into the ventrobasal complex within the thalamus can suppress nociceptive processing in chronic inflammation. It has also been observed that the antinociceptive effect of the GABA uptake inhibitor tiagabine in rodents, which can be attributed to $GABA_B$ receptor activation, is associated with an increase in the extracellular concentration of GABA within the thalamus.

In contrast to models of inflammatory pain, the induction of neuropathic pain in rodents does not

produce an increase in GABA levels within the spinal cord. However, baclofen can produce an antinociceptive effect in models of chronic neuropathy. Support for a role for GABA$_B$ receptors in pain mechanisms also derives from developmental knockout studies in mice. In these mice, functional GABA$_B$ receptors are not formed because the mice are deficient in either of the individual GABA$_{B1}$ or GABA$_{B2}$ subunits. In both forms of null mutant mice, hyperalgesia was exhibited in acute nociceptive tests, suggesting that functional heteromeric GABA$_B$ receptors are required to maintain pain thresholds.

Cognitive Function

GABA$_B$ receptor agonists suppress cognitive behavior in animals, and this action is reversed by GABA$_B$ antagonists. Moreover, basal cognitive activity can even be enhanced by GABA$_B$ antagonists without prior suppression by an agonist. This raises the possibility that GABA$_B$ antagonists might provide a novel opportunity to treat cognitive impairment in humans, and this hypothesis is being tested in clinical trials. Evidence suggests that the site of action of GABA$_B$ antagonists in relation to cognition may be the hippocampus, where an increase in long-term potentiation (LTP) has been implicated but the nature of this modification appears to depend on the frequency of stimulation employed to produce LTP.

Depression and Anxiety

A role for GABA$_B$ receptors in functional depression was first proposed more than 20 years ago. An upregulation in GABA$_B$ binding sites occurs in rat frontal cortex after chronic administration of a variety of antidepressant drugs, and although these findings were disputed, there is little doubt that GABA$_B$ mechanisms can be associated with depression. Antagonism of GABA$_B$ receptors produces a reversal of depressant-like behavior in animal models such as the rodent forced swim test and learned helplessness models. It has been observed that mice lacking GABA$_{B1}$ or GABA$_{B2}$ receptor subunits exhibit antidepressant-like behavior but they appear to be more anxious. Therefore, GABA$_B$ receptor activation appears to produce anxiolytic activity, whereas a loss or blockade of GABA$_B$ receptor function produces antidepressant-like effects. In support of this distinction, GABA$_B$ receptor activation in the dorsal periaqueductal gray of rats impairs one-way escape in the elevated T-maze test, which is consistent with an anxiolytic effect.

Drug Addiction

Therapeutic treatment of dependence on drugs of abuse is still inadequate and is therefore a major clinical goal. A number of receptor systems appear to provide potential targets, including dopamine and glutamate receptors, but GABA$_B$ receptor activation may provide an additional approach. Baclofen was initially shown to reduce the reinforcing effects of cocaine in rats, but it soon became clear that other drugs of addiction, including nicotine, morphine-related agents, and ethanol, were also sensitive to GABA$_B$ agonists, whereas food reinforcement was unaffected.

The finding that baclofen reduces craving for a variety of unrelated addictive substances, including heroin, alcohol, and nicotine, suggests that there may be an underlying common mechanism for the GABA$_B$ agonist in all cases. The reward center within the mesolimbic system, possibly the ventral tegmental area, would provide the focus for this action where control of the release/action of dopamine is implicated. Raising endogenous GABA levels within the mesolimbic system should have the same effect as administering a GABA$_B$ agonist. Thus, if vigabatrin, an inhibitor of GABA metabolism, or the GABA uptake inhibitor NO-711 are administered centrally in rats, they both attenuate heroin and cocaine self-administration and prevent cocaine-induced increases in dopamine in this brain region.

Clinical data indicate that baclofen is also effective against cocaine and alcohol craving in humans. However, the administration of baclofen can produce adverse effects as well as muscle relaxation, and these effects may detract from any potential benefits.

Therefore, what might provide an alternative approach? The presence of a positive allosteric modulatory site on the GABA$_{B2}$ subunit could provide the answer.

Allosteric Modulation of GABA$_B$ Receptors

The location of the modulator allosteric site appears to be the heptahelical domain of the GABA$_{B2}$ subunit. Although the GABA$_{B1}$ subunit has the agonist binding domain, it does not have an allosteric site, whereas the converse appears to be true for GABA$_{B2}$. Two compounds, CGP7930 and GS39783 (Figure 2), were originally reported to be positive modulators. Neither has any direct agonist activity, but both accentuate the effects of GABA and baclofen. These compounds have been examined in rat models of addiction and both modulators reduced the self-administration of cocaine.

Seizure Generation in Absence Epilepsy

Absence seizures have a characteristic EEG waveform of a 3-Hz spike and wave which stem from discharges

in the thalamic nuclei. It is believed that the thalamus is the site from which these discharges originate, and although an intact thalamocortical network is necessary for generating spike and wave discharges, the origin of these discharges appears to lie outside the thalamus. The site of origin appears to be within the perioral region of the somatosensory cortex. This discharge spreads rapidly across the cortex and initiates a corticothalamic cascade. Injection of a $GABA_B$ agonist into the ventrobasal thalamus or reticular nucleus of a rat exacerbates this activity. By contrast, injection of a $GABA_B$ antagonist into the same regions suppresses the spike and wave discharges. If the $GABA_B$ antagonist is administered systemically, the same effect occurs. This might indicate that interference with the GABAergic innervation from the reticular nucleus disrupts the thalamocortical loop which generates the spike and wave activity. However, microinjection of the same selective antagonist into the somatosensory cortex can also suppress the seizure activity, indicating the involvement of $GABA_B$ergic mechanisms also at this level.

In patients with existing absence seizures, $GABA_B$ receptor agonists would likely enhance seizure activity, and any increase in GABA concentration in the vicinity of the thalamus would be expected to enhance seizure activity. Thus, for example, vigabatrin and tiagabine would be, and are, contraindicated in such individuals. The mechanism(s) underlying seizure exacerbation by $GABA_B$ agonists is unclear, but involvement of transient Ca^{2+} T currents seems possible.

It has been suggested that γ-hydroxybutyrate (GHB), which can produce absence-like seizures, is a weak $GABA_B$ receptor agonist. If GHB mimics the effect of the endogenous agonist, GABA, at $GABA_B$ receptors, then $GABA_B$ receptor antagonists would be expected to block the spike and wave discharges produced by GHB, which they do.

Conclusions

The $GABA_B$ receptor system is of predominant, but not exclusive, importance within the central nervous system, and it may provide an important drug target in a variety of central nervous system disorders. It is a member of group 3 G-protein-coupled receptors and exists as a heterodimer with various isoforms comprising the $GABA_{B1}$ subunit. However, lack of demonstrable functional subtyping of the receptor has limited exploitation of any pharmacological specificity. The only receptor agonist in current clinical use is baclofen, which has many limitations, including poor brain penetration. To date, little has emerged which is an improvement over baclofen. However, the discovery of allosteric modulators of the $GABA_B$ system may provide the possibility of producing compounds that will readily gain access to the central nervous system while facilitating receptor function.

See also: GABA Synthesis and Metabolism; GABA-A Receptors: Developmental Roles; GABA-A Receptor Synaptic Functions; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-A Receptors and Disease; GABA-B Receptor Function; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism.

Further Reading

Bettler B, Kaupman K, Mosbacher J, and Gassmann M (2004) Molecular structure and physiological functions of $GABA_B$ receptors. *Physiological Reviews* 84: 835–867.

Binet V, Brajon C, Le Corre L, Acher F, and Pin J-P (2004) The heptahelical domain of $GABA_{B2}$ is activated directly by CGP7930, a positive allosteric modulator of the $GABA_B$ receptor. *Journal of Biological Chemistry* 279: 29085–29091.

Bowery NG, Bettler B, Froestl W, et al. (2002) International Union of Pharmacology. XXXIII: Mammalian γ-aminobutyric acid$_B$ receptors: Structure and function. *Pharmacological Reviews* 54(2): 247–264.

Cryan JF and Kaupmann K (2005) Don't worry 'B' happy! A role for $GABA_B$ receptors in anxiety and depression. *Trends in Pharmacological Sciences* 26(1): 36–43.

Froestl W and Mickel SW (1997) Chemistry of $GABA_B$ modulators. In: Enna SJ and Bowery NG (eds.) *The GABA Receptors*, pp. 271–296. Totowa, NJ: Humana Press.

Gassmann M, Shaban H, Vigot R, et al. (2004) Redistribution of $GABA_{B(1)}$ protein and atypical $GABA_B$ responses in $GABA_{B(2)}$-deficient mice. *Journal of Neuroscience* 24: 6086–6097.

Ipponi A, Lamberti C, Medica A, Bartolini A, and Malmberg-Aiello P (1999) Tiagabine antinociception in rodents depends on $GABA_B$ receptor activation: Parallel antinociception testing and medial thalamus GABA microdialysis. *European Journal of Pharmacology* 368: 205–211.

Kaupmann K, Malitschek B, Schuler V, et al. (1998) $GABA_B$ receptor subtypes assemble into functional heteromeric complexes. *Nature* 396: 683–687.

Lehmann A, Antonsson M, Bremner-Danielsen M, Flärdh M, Hansson-Branden L, and Kärrberg L (1999) Activation of the $GABA_B$ receptor inhibits transient lower esophageal sphincter relaxations in dogs. *Gastroenterology* 117: 1147–1154.

Lloyd KG, Thuret F, and Pilc A (1985) Upregulation of gamma-aminobutyric acid ($GABA_B$) binding sites in rat frontal cortex: A common action of repeated administration of different classes of antidepressant and electroshock. *Journal of Pharmacology and Experimental Therapeutics* 235: 191–199.

Malcangio M and Bowery NG (1994) Spinal cord SP release and hyperalgesia in monoarthritic rats: Involvement of the $GABA_B$ receptor system. *British Journal of Pharmacology* 113: 1561–1566.

Malcangio M and Bowery NG (1996) GABA and its receptors in the spinal cord. *Trends in Pharmacological Sciences* 17: 457–462.

Manning J-P, Richards DA, and Bowery NG (2003) Pharmacology of absence epilepsy. *Trends in Pharmacological Sciences* 24: 542–549.

Prosser HM, Gill CH, Hirst WD, et al. (2001) Epileptogenesis and enhanced prepulse inhibition in $GABA_{B1}$-deficient mice. *Molecular and Cellular Neuroscience* 17: 1059–1070.

Robbins MJ, Calver AR, Fillipov AK, Couve A, Moss SJ, and Pangalos MN (2001) The $GABA_{B2}$ subunit is essential for G protein coupling of the $GABA_B$ receptor heterodimer. *Journal of Neuroscience* 21: 8043–8052.

Schuler V, Luscher C, Blanchet C, et al. (2001) Epilepsy, hyperalgesia, impaired memory, and loss of pre- and post-synaptic $GABA_B$ responses in mice lacking $GABA_{B1}$. *Neuron* 31: 47–58.

Slattery DA, Markou A, Froestl W, and Cryan JF (2005) The $GABA_B$ receptor-positive modulator GS39783 and the $GABA_B$ receptor agonist baclofen attenuate the reward-facilitating effects of cocaine: Intracranial self-stimulation studies in the rat. *Neuropsychopharmacology* 30: 2065–2072.

Thuault SJ, Brown JT, Sheardown SA, et al. (2004) The $GABA_{B2}$ subunit is critical for the trafficking and function of native $GABA_B$ receptors. *Biochemical Pharmacology* 68(8): 1655–1666.

Urwyler S, Mosbacher J, Lingenhoehl K, et al. (2001) Positive allosteric modulation of native and recombinant γ-aminobutyric $acid_B$ receptors by 2,6-di-*tert*-butyl-4-(3-hydroxy-2,2-dimethyl-propyl)-phenol (CGP7930) and its aldehyde analog CGP13501. *Molecular Pharmacology* 60: 963–971.

Urwyler S, Pozza MF, Lingenhoehl K, et al. (2003) GS39783 (*N*,*N*′-dicyclopentyl-2-methylsulfanyl-5-nitro-pyrimidine-4,6-diamine) and structurally related compounds: Novel allosteric enhancers of γ-aminobutyric $acid_B$ receptor function. *Journal of Pharmacology and Experimental Therapeutics* 307: 322–330.

White JH, Wise A, Main M, et al. (1998) Heterodimerization is required for the formation of a functional $GABA_B$ receptor. *Nature* 396: 679–682.

Gain Modulation

E Salinas, Wake Forest University School of Medicine, Winston-Salem, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Origin and Key Issues

The term 'gain field' was coined to describe an interaction between visually evoked responses and eye position first documented by Andersen and Mountcastle in 1983. In their experiment, as well as in those that followed, neurons from parietal cortex were recorded in awake monkeys and the responses were systematically mapped by varying both eye position and the location of a spot of light. Surprisingly, although the shape of the response curves as functions of stimulus location did not change with eye position, their amplitude, or gain, did. Figure 1(e) shows an idealized example. In analogy with the receptive field, the gain field corresponded to the full map showing how the response changed as a function of eye position. Initial work described gain fields that depended on eye position, but an additional dependence on head position was demonstrated later, such that the combination of head and eye angles – the gaze angle – seems to be the relevant quantity. These interactions are well approximated by a multiplication of two factors, such that the firing rate r of a visual parietal neuron can be expressed as $r = f(x)g(y)$, where x is the location of a visual stimulus and y the gaze angle. The functions f and g may have a variety of forms, but what is important computationally is that there are two separate terms that are combined nonlinearly (multiplicatively in this case).

In the last 25 years or so, abundant experimental evidence has accumulated showing that, in the central nervous system, gain modulation is a widespread mechanism for combining information from multiple sources, not just vision and gaze angle. At the same time, complementary modeling studies have addressed two issues: (1) the kinds of computational problems that can be solved or substantially simplified by using gain modulation and (2), the biophysical mechanisms that may give rise to nonlinear interactions between two types of inputs converging on a neuron, as observed in experiments reporting changes in gain. Some of the highlights of this work are described here.

Gain Modulation and Coordinate Transformations

Zipser and Andersen were the first to demonstrate the power of gain modulation in a specific biological computation. They trained an artificial neural network to transform the location of a stimulus from an initial eye-centered representation to one in head-centered coordinates. Their network essentially learned to produce multiple neural responses like that in Figure 1(f) given an input that encoded gaze angle and a set of standard eye-centered (or retino-centered) responses like the one in Figure 1(d). What they found, after their network learned to perform the transformation, were intermediate model units that had developed gaze-dependent gain fields much like those in parietal cortex (see Figure 1(e)). This result suggested that gain fields were in some sense optimal for performing a coordinate transformation, which was an extraordinary breakthrough. However, it was not entirely clear why gain modulation was optimal. Subsequent theoretical studies showed that a network of neurons that respond to stumulus location x and are gain-modulated by gaze angle y is useful because it acts as a flexible set of building blocks from which any other, arbitrary functions of x and y can be constructed downstream. Mathematically, the gain-modulated neurons are said to serve as a basis set, meaning that by adding their responses in various proportions one can construct any desired function of x and y. Furthermore, with gain-modulated cells many important downstream functions can be generated using simple synaptic modification rules (Hebbian, or correlation-based, learning rules).

The implication is that if area A has neurons that respond to stimulus x and are gain modulated by a quantity y, neurons in a downstream area B that are driven by A will probably respond in more complicated ways as functions of x and y; perhaps they will respond to $x+y$ or $x-y$, in which case their tuning curves as functions of x will shift with different values of y (as in Figure 1(f)), or perhaps the dependencies will be nonlinear. Tuning curve shifts associated with a variety of coordinate transformations have been documented in several cortical and subcortical areas, including the lateral parietal area (LIP), the ventral parietal area (VIP), premotor cortex, and the superior colliculus.

Other Reference Frames

Although the variables x and y introduced here initially represented target location and eye position, the mechanism for generating transformed representations is extremely general, so the results are valid for any other encoded variables. Indeed, neurophysiological experiments have documented modulation by a variety of internal or proprioceptive signals. Examples include gaze direction, eye or head velocity, arm position, and attentional location. Typically, these signals seem to be directly combined with spatially tuned

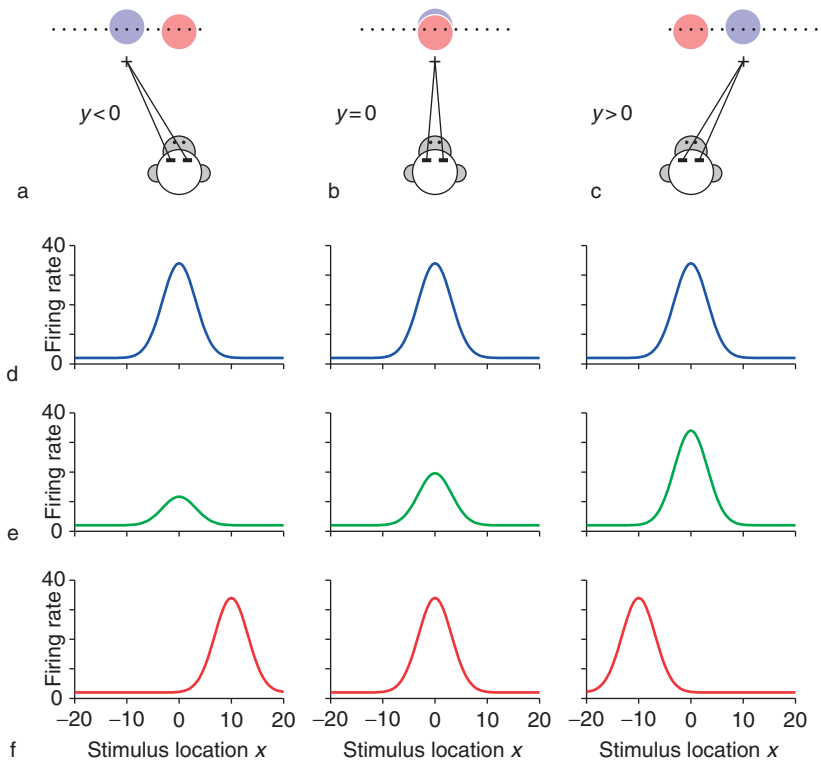

Figure 1 Transforming stimulus location from retino-centered to head-centered coordinates. The top panels schematize experiments in which visual receptive fields are mapped at three eye positions. The cross corresponds to the fixation point; small dots above it indicate locations tested with flashed stimuli. Blue and red circles indicate receptive fields of neurons in (d) and (f) respectively. Gaze angle is *y*. (a) Gaze is directed to the left. (b) Gaze is directed straight ahead. (c) Gaze is directed to the right. (d) Responses of a hypothetical early visual neuron in the three conditions schematized above. Firing rate is plotted as a function of stimulus location *x*, which is measured relative to the fixation point (i.e., in retinal coordinates). The tuning curve is independent of gaze angle. (e) Responses of a hypothetical parietal neuron that is gain modulated by gaze. Its receptive field location stays invariant relative to the fixation point, but its amplitude or gain depends on gaze direction. (f) Responses of a hypothetical neuron that encodes stimulus location in head-centered coordinates. Its receptive field stays in the same location relative to the head, so its tuning curve shifts as a function of *x*.

sensory responses to produce a change in reference frame, as in the case of eye position.

The generality of this framework – the idea that arrays of gain-modulated neurons constitute basis sets for constructing more elaborate representations downstream – also extends the concept of coordinate transformation to perceptual processes, such as object recognition. For instance, directing attention to different locations in space modifies the gain of many visual neurons that react selectively to specific features of an image. According to work by Salinas and Abbott, such attentional modulation may give rise to visual responses that are, to a good degree, independent of where they appear in the visual field. Some neurons in inferior-temporal cortex (IT) seem to have this property, also known as translation invariance: they respond depending only on the type of image presented, regardless of its location. According to the theory, they encode visual information in attention-centered coordinates. These cells are considered the neural correlate of our ability to recognize objects independently of where they appear in the visual field.

Partial Shifts in Recurrent Networks

Recurrent connections are ubiquitous in cortical networks, so it is important to understand how computations mediated by gain modulation change in networks with feedback interactions. Studies by Xing and Andersen and by Pouget and collaborators have investigated this in networks trained to perform coordinate transformations to and from diverse reference frames, but where feedback between neurons is allowed. With such feedback, the sensory responses that are optimal for implementing the transformations – the basis functions – still show nonlinear dependencies on the relevant proprioceptive parameters (eye position, head position, etc.). But in addition to changes in gain, their tuning curves show partial shifts as functions of stimulus location, particularly when more than two reference frames are involved. In other words, with feedback and multiple coordinate systems, the most efficient neuronal responses turn out to be intermediate between those shown in Figures 1(e) and 1(f), so individual neurons should display changes

in gain together with tuning curve shifts that are smaller than those expected for a full transformation. Indeed, this mixture is typical of data from various experimental preparations, suggesting that models with feedback capture more of the complexity of real neuronal circuits.

Arbitrary Visuomotor Remapping

Most theoretical studies on gain modulation have focused on the idea that proprioceptive signals combined with spatial sensory information lead to changes in reference frame. However, gain interactions may serve to integrate sensory information with other types of events, not necessarily proprioceptive, and in that case the mechanism works for establishing arbitrary associations between sensory stimuli and motor actions. This refers to tasks or behaviors in which a given stimulus is arbitrarily associated with two or more motor responses depending on separate cues, which provide a context. A simple example is illustrated in Figure 2, where the color of the fixation point indicates whether the correct movement should be a saccade (eye movement) toward a spot of light or an antisaccade away from it. Importantly, the stimulus may appear at many different locations. Gain modulation can be used to construct network models for tasks like this, in which multiple maps between sensory stimuli and motor actions are possible, but only one map, depending on the context, is implemented at any given time. This is illustrated in Figures 3 and 4.

Switching Sensory–Motor Maps

Figure 3 depicts four hypothetical tuning curves from neurons that respond in the task of Figure 2. Each unit has a preferred stimulus location and a preferred context, with red and green corresponding to saccade and antisaccade conditions, respectively. Modulation strength varies across cells and is strongest for the unit in Figure 3(a), which is fully suppressed in the nonpreferred context. Such a 'switching' neuron is entirely turned on and off by context. It is straightforward to see that an array of switching neurons can be easily configured to generate saccades in one context and antisaccades in the other: the neurons that are turned on in context 1 can establish their own set of synaptic connections regardless of the connections established by the other neurons, which are active only in context 2, and vice versa. However, to switch sensory–motor maps it is not necessary that individual neurons turn on and off with context. Figure 4 illustrates this point.

The results in Figure 4 were obtained with a simple network without feedback in which 60 gain-modulated sensory neurons drive a set of 25 output or motor neurons through synaptic connections, and the output firing rates (black dots) are determined by weighted sums of the sensory firing rates (colored dots), such that

$$R_i(x,y) = \sum_j w_{ij} r_j(x,y) \qquad [1]$$

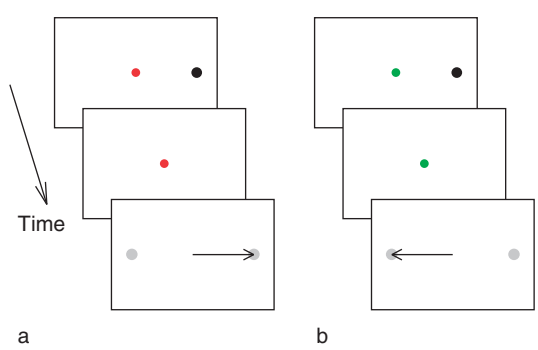

Figure 2 An arbitrary sensory–motor mapping task. In each trial, a stimulus (black dot) is presented at a distance x from the fixation point (colored dot); the stimulus disappears; two targets appear (gray dots) and the subject responds by making an eye movement (arrow) to one of them. The color of the fixation spot acts as a contextual cue indicating whether the movement should be a saccade or an antisaccade. (a) In context 1 the fixation spot is red and the movement is to the target at x. (b) In context 2 the fixation spot is green and the movement is to the opposite target, at $-x$. Adapted from Salinas E (2004) Context-dependent selection of visuomotor maps. *BioMed Central Neuroscience* 5: 47.

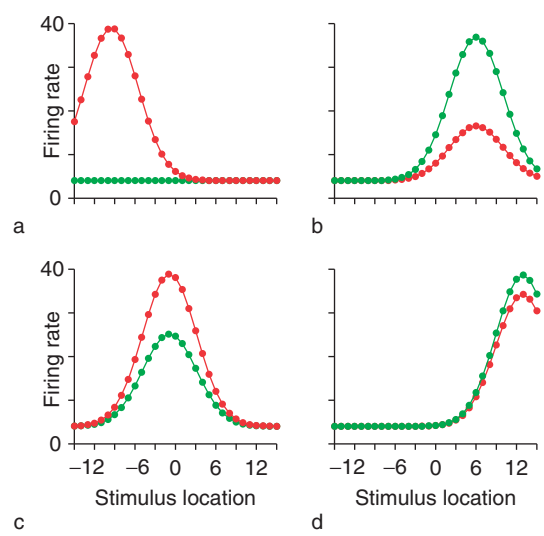

Figure 3 Responses of four hypothetical gain-modulated neurons in the remapping task of Figure 2. Each graph plots the mean firing rate of a model neuron as a function of stimulus location. Red and green traces correspond to sensory responses evoked during contexts 1 and 2, respectively. (a) A unit that prefers context 1 and is 100% suppressed in the nonpreferred condition. (b) A unit that prefers context 2 and is 62% suppressed in the nonpreferred condition. (c) A unit that prefers context 1 and is 39% suppressed in context 2. (d) A unit that prefers context 2 and is 13% suppressed in context 1. Adapted from Salinas E (2004) Context-dependent selection of visuomotor maps. *BioMed Central Neuroscience* 5: 47.

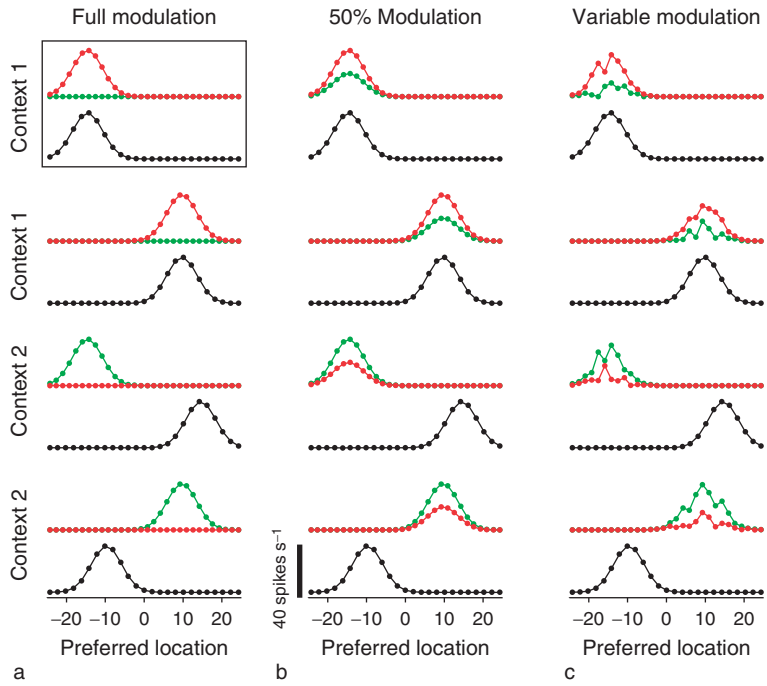

Figure 4 A model network that switches between saccades and antisaccades. The model consists of 60 sensory neurons driving 25 motor or output neurons through synaptic connections. The box marks a single trial. Colored traces are the firing rates of the 60 gain-modulated neurons in the network; 30 of them (red) prefer the direct saccade condition and 30 (green) prefer the antisaccade condition. Black dots are the 25 motor responses driven by the gain-modulated neurons. The *x*-axis is preferred stimulus location for the sensory cells and preferred movement location for the motor cells. In context 1 (direct saccades), the peak of motor activity should be aligned with the peak of sensory activity; in context 2 (antisaccades), the motor activity should peak at the mirror-symmetric point from the sensory activity. (a) Firing rates of all model cells when the sensory units are fully modulated by context (switching neurons). The context in each of four trials is indicated on the left. Trials with $x = -15$ and $x = 10$ alternate. The profile of output activity always peaks at the location of the correct saccade. (b) As in (a), but when all gain-modulated neurons are partially modulated by the same amount. (c) As in (a), but when all the gain factors are chosen randomly. Sets of neurons with diverse gain modulation factors can, with appropriate synaptic weights, perform the same input-output transformations as neurons that are fully modulated. Adapted from Salinas E (2004) Context-dependent selection of visuomotor maps. *BioMed Central Neuroscience* 5: 47.

Here, $R_i(x, y)$ is the firing rate of output unit i, w_{ij} is the synaptic connection from sensory neuron j to output unit i, and $r_j(x, y)$ is the firing rate of sensory neuron j when the context is y (1 or 2) and the stimulus is located at x. In Figure 4(a) the model neurons are fully modulated by context, so only one subnetwork is active in context 1 (red dots) and only the other in context 2 (green dots). The two subnetworks are never active simultaneously, so they can establish independent connection matrices and generate separate sensory–motor maps for contexts 1 and 2. In Figure 4(b) the model works similarly, except that the sensory neurons are only partially suppressed in their nonpreferred context. Although the synaptic connections that allow the model to perform the two versions of the task (saccades and antisaccades) are different than in Figure 4(a), with the appropriate connections w_{ij}, performance is again accurate. This is true even when each sensory cell is modulated by a different, random amount, as in Figure 4(c).

More generally, the key result is that, if certain mild conditions on the modulation strengths are satisfied, then any downstream function $R_i(x, y)$ that can be constructed with switching neurons can also be constructed with neurons that are only partially suppressed. In practice, this means that anything that can be done with subpopulations of switching neurons can also be done with neurons that modulate their responses only moderately. This result offers a useful intuition as to why gain modulation is so powerful: modulating the activity of a neuronal population is equivalent to flipping a switch that turns on or off various sets of neurons.

Evidence of Context-Dependent Modulation

Although relatively little is known about how sensory-triggered activity is modified by arbitrary contextual cues, there are several experimental reports revealing (1) neuronal responses that are sensitive to more than one variable or stimulus feature and (2) that the interactions between those variables are generally nonlinear, and thus consistent with changes in gain. For example, the activity evoked in prefrontal neurons by moving dots varies substantially depending on

whether their color or their direction of motion is relevant to the task at hand. Similarly, visual responses may change depending on the probability of obtaining a reward in the current trial, or on the rule that is used to link them to a motor behavior. Such effects could be used for generating abstract transformations in which the contextual or modulatory signal itself lacks any spatial information.

Mechanisms Underlying Changes in Gain

What biophysical mechanisms might allow a neuron to combine two input signals in a multiplicative way? The problem is this. Suppose there are two nearby synapses, a and b, that generate depolarizations ΔV_a and ΔV_b, respectively, when activated individually. To a first approximation, the depolarization observed when activated together should be $\Delta V_a + \Delta V_b$, and indeed this assumption is valid under various conditions. The question is how to make the combined response of synapses a and b nonlinear, and in particular, how to make it so that the neuron's firing rate behaves approximately as a product of a- and b-dependent terms, as observed in extracellular recordings.

Proposed Solutions

In recent years this issue has attracted considerable attention, so several candidate mechanisms with quite different properties have been proposed. One of the earliest suggestions was that of Mel, who showed using biophysically detailed computer simulations, that dendritic conductances with nonlinear voltage dependencies could boost the responses of nearby synapses that are activated simultaneously. That is, due to depolarizing currents in the dendrites, nearby synapses a and b that are active together can generate a depolarization that is much larger than $\Delta V_a + \Delta V_b$. There is experimental support for this effect: on some pyramidal neurons, nearby synapses on the same dendritic branch appear to sum sigmoidally, whereas synapses that are either far apart or on different branches sum linearly.

A second possibility is based on the observation that circuits with positive feedback tend to amplify their inputs. In this scheme, strong recurrent activity between neurons with similar gain fields is the basis for generating a multiplicative effect without the need for synapses to interact. Another mechanism that does not place any restrictions on the synapses themselves has been proposed by Chance and collaborators. They showed that the gain of a neuronal response to excitatory drive can be modulated by varying the overall level of synaptic input. Simultaneously increasing both excitatory and inhibitory background firing rates in a balanced manner results in a divisive effect that modulates the responsiveness to excitatory drive. This, however, requires coordinated changes in the background activity; that is, although each input neuron needs to modify its firing rate by only a small amount, similar changes must occur across a large number of input neurons for the mechanism to work. In addition to these proposals, shunting inhibition, inhibitory synchrony, and input–output functions described by a power law have also been discussed in theoretical studies as being possibly implicated in regulating the gain of neurons.

Two more mechanisms have been suggested recently on the basis of experimental observations. The first one is based on results by Larkum and colleagues showing that, in layer 5 pyramidal cells, there is a nonlinear interaction between inputs that arrive at the soma and at the apical dendrites. The key is that back-propagating action potentials help activate high-threshold dendritic calcium conductances, such that bursts of spikes are produced only when somatic and dendritic inputs coincide. Thus, a weak, asynchronous current at the distal dendrites can generate extra spikes, because of the evoked bursts, and this increases the gain of the response to somatic input. Finally, Mehaffey and colleagues suggested that a neuron's gain may be controlled through inhibition that is applied at the dendrites, and specifically not at the soma. Due to the back-propagation of action potentials and to voltage-dependent currents at the dendrites, after a neuron fires a spike, a depolarizing afterpotential is typically generated. The crucial observation is that the size of this afterpotential regulates the gain of the cell, and that dendritic inhibition – but not somatic inhibition – reduces it. Therefore, dendritic inhibition decreases the gain.

This last idea is interesting because the biophysical requirements for this form of divisive gain control are relatively simple. However, it is likely that neurons use a variety of mechanisms to regulate their gain over different timescales, so the possibilities mentioned earlier (and future ones) should not be considered mutually exclusive.

See also: Motor Psychophysics; Neural Coding of Spatial Representations; Parietal Cortex and Spatial Attention; Saccadic Eye Movements; Spatial Transformations for Eye–Hand Coordination.

Further Reading

Andersen RA, Essick GK, and Siegel RM (1985) Encoding of spatial location by posterior parietal neurons. *Science* 230: 450–458.

Andersen RA and Mountcastle VB (1983) The influence of the angle of gaze upon the excitability of light-sensitive neurons

of the posterior parietal cortex. *Journal of Neuroscience* 3: 532–548.

Avillac M, Deneve S, Olivier E, Pouget A, and Duhamel JR (2005) Reference frames for representing visual and tactile locations in parietal cortex. *Nature Neuroscience* 8: 941–949.

Brotchie PR, Andersen RA, Snyder LH, and Goodman SJ (1995) Head position signals used by parietal neurons to encode locations of visual stimuli. *Nature* 375: 232–235.

Chance FS, Abbott LF, and Reyes AD (2002) Gain modulation from background synaptic input. *Neuron* 35: 773–782.

Deneve S, Latham PE, and Pouget A (2001) Efficient computation and cue integration with noisy population codes. *Nature Neuroscience* 4: 826–831.

Graziano MSA, Hu TX, and Gross CG (1997) Visuospatial properties of ventral premotor cortex. *Journal of Neurophysiology* 77: 2268–2292.

Larkum ME, Senn W, and Lüscher H-R (2004) Top-down dendritic input increases the gain of layer 5 pyramidal neurons. *Cerebral Cortex* 14: 1059–1070.

Lauwereyns J, Sakagami M, Tsutsui K, Kobayashi S, Koizumi M, and Hikosaka O (2001) Responses to task-irrelevant visual features by primate prefrontal neurons. *Journal of Neurophysiology* 86: 2001–2010.

Mehaffey WH, Doiron B, Maler L, and Turner RW (2005) Deterministic multiplicative gain control with active dendrites. *Journal of Neuroscience* 25: 9968–9977.

Poggio T (1990) A theory of how the brain might work. *Cold Spring Harbor Symposium on Quantitative Biology* 5: 899–910.

Polsky A, Mel BW, and Schiller J (2004) Computational subunits in thin dendrites of pyramidal cells. *Nature Neuroscience* 7: 621–627.

Pouget A and Sejnowski TJ (1997) Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience* 9: 222–237.

Salinas E (2004) Context-dependent selection of visuomotor maps. *BioMed Central Neuroscience* 5: 47.

Salinas E and Abbott LF (1995) Transfer of coded information from sensory to motor networks. *Journal of Neuroscience* 15: 6461–6474.

Salinas E and Abbott LF (1997) Invariant visual responses from attentional gain fields. *Journal of Neurophysiology* 77: 3267–3272.

Salinas E and Their P (2000) Gain modulation: A major computational principle of the central nervous system. *Neuron* 27: 15–21.

Wallis JD and Miller EK (2003) From rule to response: Neuronal processes in the premotor and prefrontal cortex. *Journal of Neurophysiology* 90: 1790–1806.

Xing J and Andersen RA (2000) Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames. *Journal of Cognitive Neuroscience* 12: 601–614.

Zipser D and Andersen RA (1988) A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature* 331: 679–684.

Galanin and Receptors

K R Bailey and J N Crawley, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Discovery and Distribution of Galanin

Kazuhiko Tatemoto and Viktor Mutt at the Karolinska Institutet in Stockholm discovered galanin in extracts from porcine intestine. It was first reported in 1983; the name galanin was taken from the first and last amino acids of the porcine sequence, glycine and alanine. The amino acid sequence of galanin is highly conserved across species, consisting of 29 amino acids with an amidated N-terminal. The exception is human galanin, composed of the non-amidated 30 amino acid sequence Gly-Trp-Thr-Leu-Asn-Ser-Ala-Gly-Tyr-Leu-Leu-Gly-Pro-His-Ala-Val-Gly-Asn-His-Arg-Ser-Phe-Ser-Asp-Lys-Asn-Gly-Leu-Thr-Ser. Synthesis is by enzymatic cleavage from a larger preprogalanin protein. Preprogalanin contains a second biologically active sequence, galanin-like peptide, composed of 60 amino acids. Galanin and galanin-like peptide share 13 homologous amino acids but appear to mediate different biological functions. The chromosomal location of the human preprogalanin gene is 11q13.3–q13.5.

In the mammalian central nervous system, galanin is discretely distributed and coexists with several other neurotransmitters. High concentrations of galanin-like immunoreactivity are seen in the rat hypothalamus, locus coeruleus, and dorsal horn of the spinal cord. Galanin and galanin mRNA are present in the amygdala, basal forebrain, cerebral cortex, dorsal raphe, hippocampus, medulla, striatum, and many other brain regions. Coexistences vary by species. In the rat, galanin-like immunoreactivity is found within approximately 90% of tyrosine hydroxylase-immunoreactive neurons of the locus coeruleus, 30% of the serotonergic dorsal raphe neurons, and 30% of the cholinergic neurons of the diagonal band and medial septum. Smaller percentages are seen in the coexistences of galanin with dopamine in the median eminence, with gonadotropin-releasing hormone in the preoptic area and with vasopressin in the bed nucleus of the stria terminalis. In the peripheral nervous system, galanin is prominent in the dorsal root ganglia, sympathetic nerves, heart, lung, pancreas, liver, gastrointestinal tract, and pituitary gland. Galanin expression in the pituitary is strongly upregulated by estrogen.

Discovery and Distribution of Galanin Receptors

Galanin binds selectively with high affinity to three receptors, presently termed GalR1, GalR2, and GalR3 (Figure 1). Cloned sequences in rodents and humans were published in 1994–98 by several prominent galanin research teams. The three known galanin receptor subtypes are typical members of the G-protein-coupled superfamily of receptors. Chromosomal locations in the human genome are 18q23 for GalR1, 17q25.3 for GalR2, and 22q12–q13.3 for GalR3. The three subtypes share only approximately 50% sequence homology with one another, but their gene sequences are highly conserved across species. The three subtypes activate somewhat different postsynaptic signaling events. GalR1 and GalR3 activate $G_{i/o}$ and inhibit cyclic adenosine monophosphate (cAMP) synthesis, whereas GalR2 couples to $G_{q/11}$, activates phosphatidyl inositol hydrolysis, stimulates Ca^{2+} release, opens Ca^{2+}-dependent channels and, subsequently, activates $G_{i/o}$.

In the mammalian central and peripheral nervous systems, high-affinity galanin binding is widely distributed. The anatomical distribution of mRNAs for the three galanin receptor subtypes is discretely localized. GalR1 and GalR2 are expressed at moderate to high levels in the hypothalamus, hippocampus, amygdala, cortex, basal forebrain, dorsal root ganglia, spinal cord, pituitary, pancreas, stomach, and large intestines. Some differences between GalR1 and GalR2 localizations are notable; for example, only GalR1 mRNA is seen in the substantia nigra. GalR3 is primarily expressed in the hypothalamus, locus coeruleus, dorsal raphe, central gray, dorsal root ganglia, heart, lung, adrenal gland, pancreas, spleen, kidney, colon, and testis.

Peptidergic ligands with moderate selectivity for each galanin receptor subtype have been synthesized (Figure 1). Historically, the lack of nonpeptide ligands with high subtype selectivity and bioavailability slowed the galanin research field.

Central Nervous Systems Actions of Galanin

Corroborating results from biochemical assays, tissue-slice preparations, and *in vivo* microdialysis indicate an overall dampening action of galanin on neural circuitry. Presynaptic actions of galanin include the inhibition of the release of glutamate, acetylcholine, serotonin, and norepinephrine. Postsynaptic actions

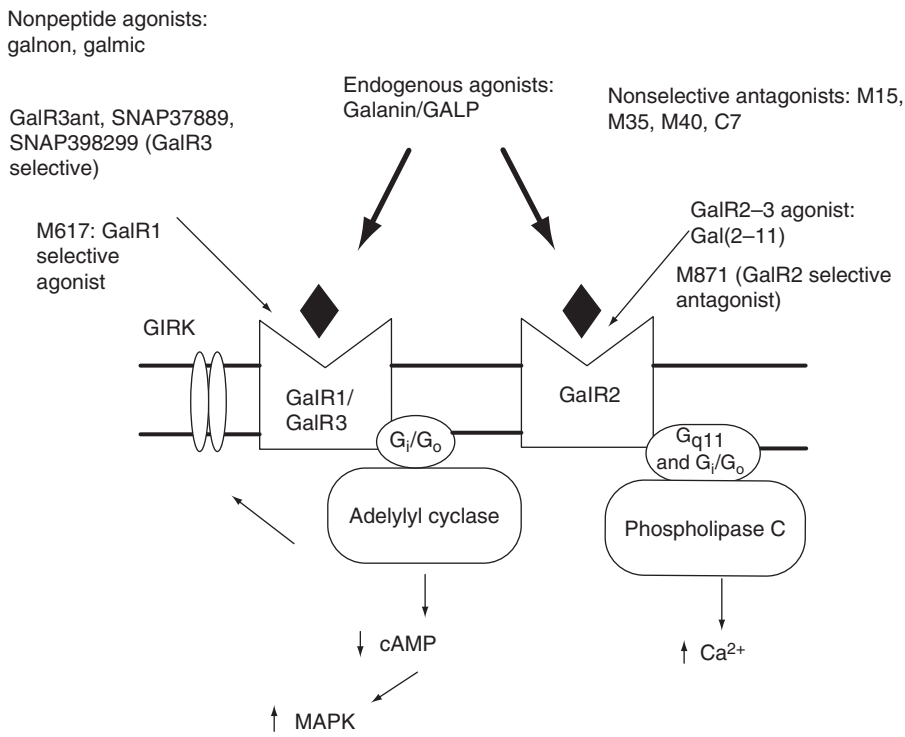

Figure 1 Three known galanin receptor subtypes, their signaling mechanisms, and their selective receptor ligands. Both peptidergic and nonpeptide small-molecule ligands are included. cAMP, cyclic adenosine monophosphate; GALP, galanin-like peptide; GalR, galanin receptor; GIRK, G-protein-activated inwardly rectifying potassium; MAPK, mitogen-activated protein kinase; SNAP, synaptosomal-associated protein. From Robinson JK, Bartfai T, and Langel Ü (2006) Galanin/GALP receptors and CNS homeostatic processes. *CNS & Neurological Disorders – Drug Targets* 5: 327–334.

of galanin include inhibition of phosphatidyl inositol hydrolysis, adenylate cyclase production of cAMP, and cAMP responsive element-binding protein (CREB) signaling.

Galanin Stimulates Feeding

Central microinjection of galanin into the lateral ventricles, third ventricle, paraventricular nucleus of the hypothalamus, and nucleus tractus solitarius evokes feeding in satiated rats (Table 1). Galanin-induced stimulation of consumption was highest at the beginning of the dark phase of the circadian cycle and greatest for foods with high-fat and -carbohydrate contents. Peptidergic galanin antagonists decreased feeding in some model systems (e.g., in rats selected for high levels of fat consumption) but not consistently across model systems and laboratories. Expression levels of galanin mRNA in hypothalamic nuclei varied with level of fasting and refeeding. However, mice with a conventional lifetime targeted mutation in the gene for galanin, the GalR1 receptor, or the GalR2 receptor maintained normal body weights. The actions of galanin at multiple receptors and the interactions of galanin with other neurotransmitters and hormones, including norepinephrine, serotonin, neuropeptide Y, leptin, and insulin, may be responsible for the modulatory actions of galanin on food consumption.

Ethanol and Drugs of Abuse

A single-nucleotide polymorphism haplotype for the galanin gene was found to be associated with alcoholism in Finnish and Plains Indians populations. This genetic finding in the clinical literature is supported by reports that paraventricular hypothalamic administration of galanin increases ethanol consumption in rats (Table 1).

Galanin treatment reduced morphine-withdrawal behaviors and blocked place preference to morphine in normal mice. Expression of galanin mRNA was highly elevated in the locus coeruleus of mice and rats after morphine withdrawal. Galanin-overexpressing transgenic mice displayed fewer morphine withdrawal signs, whereas galanin knockout mice displayed exacerbated morphine-withdrawal signs. Attenuation of cAMP production and CREB phosphorylation by galanin appears to be the mechanism underlying the reduction in morphine-withdrawal responses (Table 1).

Table 1 Behavioral and physiological functions of galanin on central and peripheral nervous systems[a]

Nervous system actions	*Test*	*Species*	*Treatment/transgenic mutation*	*Biological effect*
Central nervous system				
Feeding		Rat	Exogenous galanin	↑ Feeding
		Rat	Galanin receptor agonists	↑ Feeding
		Mouse	GalR1 −/−	Normal body weight
		Mouse	GalR2 −/−	Normal body weight
Reinforcement/withdrawal		Rat	Exogenous galanin	↑ EtOH intake
		Rat	Galanin antagonist M40	Attenuated galanin ↓ EtOH intake
		Mouse	Exogenous galanin	↓ Morphine withdrawal symptoms
		Mouse	Galanin −/−	↑ Withdrawal symptoms
		Mouse	Galanin transgenic (Dβ H promoter)	↓ Withdrawal symptoms
Learning and memory				
Spatial memory	Acquisition	Rat	Exogenous galanin	Impaired
		Mouse	Galanin transgenic (Dβ H promoter)	Normal
		Mouse	Galanin transgenic (PDGF promoter)	Impaired at older ages
		Mouse	GalR1 −/−	Normal
		Mouse	GalR2 −/−	Normal
	Retention	Rat	Exogenous galanin	Impaired
		Mouse	Galanin transgenic (Dβ H promoter)	Impaired
		Mouse	GalR1 −/−	Normal
		Mouse	GalR2 −/−	Normal
		Mouse	Galanin transgenic (PDFG promoter)	Normal
	Working memory	Rat	Exogenous galanin	Impaired
Emotional memory	Trace fear conditioning to context	Rat	Exogenous galanin	Normal
		Mouse	Galanin transgenic (Dβ H promoter)	Normal
		Mouse	GalR1 −/−	Normal
		Mouse	GalR2 −/−	Normal
	Trace auditory cue fear conditioning	Rat	Exogenous galanin	Impaired
		Mouse	Galanin transgenic (Dβ H promoter)	Impaired
		Mouse	GalR1 −/−	Impaired
		Mouse	GalR2 −/−	Normal
	Passive avoidance	Mouse	Exogenous galanin	Impaired
Olfactory memory	Social transmission of food preference	Mouse	Galanin transgenic (Dβ H promoter)	Impaired
		Mouse	GalR1 −/−	Impaired
Seizure susceptibility	Seizure activity	Rat	Exogenous galanin	Suppressed
		Rat	Galanin receptor agonist	Suppressed
		Rat	Galanin receptor antagonist	Increased
		Mouse	Galanin −/−	Increased
		Mouse	GalR1 −/−	Increased
		Mouse	Galanin transgenic (Dβ H promoter)	Suppressed
		Mouse	Galanin transgenic (PDGF promoter)	Suppressed
	Hippocampal excitability	Mouse	Gal −/−	Increased
		Mouse	GalR1 −/−	Increased
Anxiety-related/depression-related	Approach–avoidance conflict tests			
	Elevated-plus maze	Rat	Exogenous galanin	No effect

Continued

Table 1 Continued

Nervous system actions	*Test*	*Species*	*Treatment/transgenic mutation*	*Biological effect*
		Mouse	Galanin transgenic (DβH promoter)	No effect
		Mouse	Galanin transgenic (PDGF promoter)	No effect
		Mouse	GalR1 −/−	Anxiogenic-like
		Mouse	GalR2 −/−	Anxiogenic-like
	Vogel test	Rat	Exogenous galanin	Site-specific effects
		Rat	GalR3 antagonist	Anxiolytic-like
	Light↔dark exploration	Mouse	Exogenous galanin, Gal-OE	No effect
		Mouse	Galanin transgenic (DβH promoter)	
		Mouse	Galanin transgenic (DβH promoter)	Blocked anxiogenic-like effect of yohimbine
	Open field exploration	Rat	Exogenous galanin	No effect
		Mouse	Galanin transgenic (DβH promoter)	No effect
		Mouse	Galanin transgenic (PDGF promoter)	No effect
		Mouse	Gal −/−	No effect
		Mouse	GalR1 −/−	No effect
		Mouse	GalR2 −/−	No effect
	Behavioral despair tests			
	Forced swim	Rat	Exogenous galanin	Site-specific effects
		Rat	Galanin agonist	Antidepressant-like
		Rat	Galanin antagonist	Antidepressant-like
		Mouse	Galanin transgenic (PDGF promoter)	↑ Depression-like behavior
		Mouse	GalR3 antagonist	↓ Depression-like behavior
	Tail suspension test	Mouse	Exogenous galanin	No effect
		Mouse	Galanin transgenic (DβH promoter)	No effect
		Mouse	GalR1 −/−	No effect
		Mouse	GalR2 −/−	No effect
Nociception: spinal cord[b]		Rat	Exogenous galanin	Antinociceptive
		Mouse	Gal −/−	Antinociceptive
		Rat	Galanin + morphine	Potentiated analgesia
		Rat	Galanin antagonist + morphine	Blocked analgesia
		Mouse	GalR1 −/−	↑ Hyperalgesia
Peripheral nervous system				
Reproductive/sexual function		Rat	Exogenous galanin (MPA)	↑ Male copulatory behavior
		Rat	Exogenous galanin	↓ Male copulatory behavior
		Rat	Exogenous galanin antagonist	Blocked galanin inhibitory effects
Smooth muscle		Rat	Galanin	↓ Smooth muscle contractions
Cardiovascular		Mouse	Intravenous galanin	↓ Pulse rate interval
		Mouse	Intravenous galanin	↑ Blood pressure
		Cat	Galanin antagonist	Blocked galanin ↓ pulse rate interval
		Cat	Galanin antagonist	No effect on galanin ↓ in blood pressure
Endocrine function		Rat	Exogenous galanin	↑ Growth hormone release

Continued

Table 1 Continued

Nervous system actions	*Test*	*Species*	*Treatment/transgenic mutation*	*Biological effect*
		Rat	Exogenous galanin	↑ Luteinizing hormone release
		Rat	Exogenous galanin	↑ LHRH release
		Rat	Exogenous galanin	↑ Prolactin levels
		Rat	Exogenous galanin	↓ Glucose-induced insulin release
		Rat	Exogenous galanin	↓ Oxytocin levels

[a]Exogenous administration by an intraventricular or site-specific microinjection route into brain regions, or of the galanin gene by adeno-associated viral vector delivery; administration by intrathecal routes for nociception studies and intravenous routes for peripheral studies. Phenotypes of mice with a targeted mutation in the gene for galanin or a galanin receptor subtype are included. ↑, increased; ↓, decreased, DβH, dopamine β-hydroxylase; EtOH, ethanol; Gal, galanin; Gal-OE, galanin overexpressing; GalR, galanin receptor; LHRH, luteinizing hormone releasing hormone; MPA, medial preoptic area; PDGF, platelet-derived growth factor.
[b]Galanin-null mutant mice were hyporesponsive in the formalin test. Possible loss of small unmyelinated sensory fibers, throughout development, may explain this apparent contradiction with the actions of exogenous galanin.

Overexpression in Alzheimer's Disease and in Learning and Memory Impairments

Galanin is overexpressed in the basal forebrain in Alzheimer's disease (Figures 2(g) and 2(h)). Dense galanin-immunoreactive fibers course through the region of the nucleus basalis of Meynert, and galanin-immunoreactive terminals appear to decorate and synapse on the remaining cholinergic nucleus basalis of Meynert cell bodies. Postmortem concentrations of galanin in basal-forebrain samples from Alzheimer's patients are approximately twice as high as in age-matched controls. In the same samples, cholinergic markers are lower in Alzheimer's versus age-matched controls, and levels of other neuropeptides localized in the basal forebrain do not differ between groups. Increased galanin in the basal forebrain is apparent at middle to late stages of the disease progression.

Administered centrally to rats, galanin impairs performance on a wide variety of learning and memory tasks (Table 1). Transgenic mice overexpressing galanin selectively in adrenergic neurons displayed deficits on more difficult memory tasks. The release of glutamate and acetylcholine was reduced in the hippocampus of galanin-overexpressing transgenic mice, indicating a mechanism through which galanin impairs performance on challenging learning and memory tasks. In hippocampal slices, galanin reduced long-term potentiation when administered exogenously and when overexpressed in galanin transgenic mice. Peptidergic galanin receptor antagonists have been reported to improve cognitive performance in some rat models. Galanin receptor antagonists have been proposed as a treatment for Alzheimer's disease to remove the putative inhibitory actions of excess galanin within the neuropharmacological circuitry required for learning and memory. Targeting the most relevant galanin receptor(s) could improve memory in Alzheimer's patients directly or could increase the efficacy of existing therapeutics by removing inhibitory modulation.

Anticonvulsant

Central microinjection of galanin inhibits picrotoxin-kindled seizures in rats (Table 1). Resistance to seizures induced by electrical stimulation of the perforant path was seen in galanin-overexpressing transgenic mice. Galanin-null mutant mice displayed higher seizure susceptibility. Adeno-associated viral vector delivery of the galanin gene into the inferior colliculus of rats increased the threshold for focal seizures after electrical stimulation of the collicular cortex and protected hippocampal hilar neurons from death after kainic acid administration. Inhibition of glutamate release by galanin, demonstrated in potassium-stimulated hippocampal slices taken from galanin-overexpressing transgenic mice, appears to be one mechanism underlying the antiseizure actions of galanin. GalR1 and GalR2 peptidergic agonists mimicked the actions of galanin (1–29), indicating the potential clinical efficacy of ligands at both receptor subtypes as novel treatments for epilepsy.

Anxiety and Depression

Centrally administered galanin produces anxiolytic-like responses in some, but not all, model systems (Table 1). In rats, galanin decreased punished responding when administered intraventricularly before the Vogel conflict test. Following immobilization stress or treatment with the noradrenergic autoreceptor antagonist yohimbine, anxiety-related behaviors on the elevated-plus maze and social interaction tests were attenuated by galanin treatment. However, in normal C57BL/6J mice, the intraventricular administration

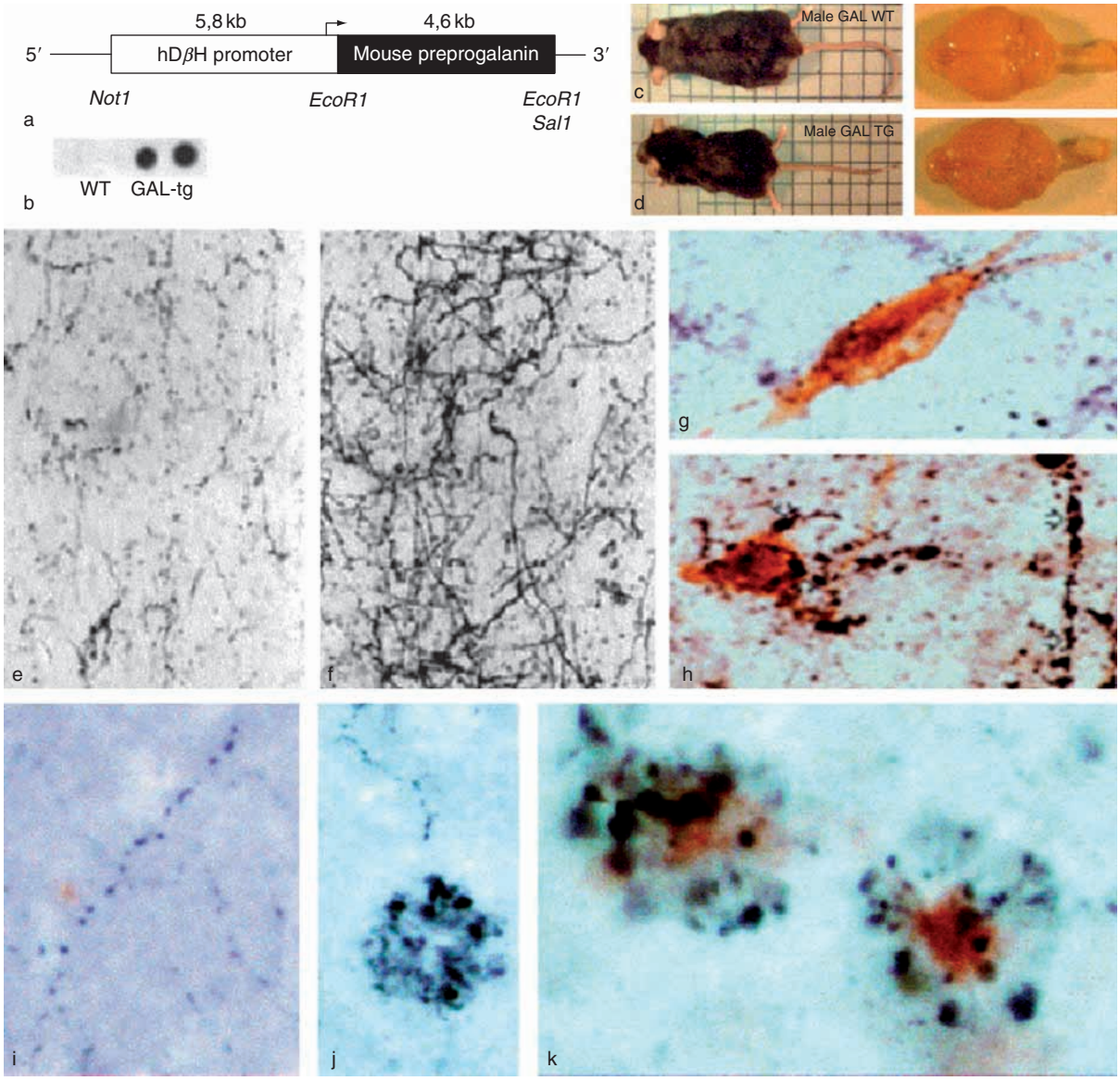

Figure 2 Similarities between galanin overexpression in the basal forebrain in Alzheimer's disease and galanin overexpression in transgenic mice with the galanin transgene generated on a dopamine β-hydroxylase (DβH) promoter: (a) schematic of the insertion of hDβH promoter into the mouse preprogalanin gene; (b) dot blot showing the increased expression of galanin in the forebrain of this mutant mouse; (c) photograph of a wild-type mouse (left) and its brain (right); (d) photograph of a transgenic mouse showing similar body phenotype (left) and its brain (right); (e) minor galanin-immunoreactive fiber staining in the wild-type mouse septal area; (f) overexpression of galanin-immunoreactive fiber staining in the transgenic mouse septal area; (g) cholinergic neuron (brown) with minor galanin innervation (black) in a human age-matched control case; (h) galanin hyperinnervation surrounding a cholinergic neuron and fiber in Alzheimer's disease; (i) galanin-positive fiber in the hippocampus of an APP/PS1Detla E9 wt mouse; (j) galanin-positive plaques and neuritis in the hippocampus of an APP/PS1Delta E9 transgenic mouse; (k) hippocampal plaque immunoreactive for amyloid (yellow) and galanin-immunoreactive neurites (black) in an APP/PS1Detla E9 transgenic mouse. In (c) and (d), the apparent difference in size of the mice is a photographic artifact, and the photographs of the brains show the lack of gross morphological difference between the genotypes. GAL-tg, galanin transgenic; WT, wild type. Reproduced from Mufson EJ, Counts SE, Perez SE, and Binder L (2005). Galanin plasticity in the cholinergic basal forebrain in Alzheimer's disease and transgenic mice. *Neuropeptides* 39: 233–237, with permission from Elsevier.

of galanin did not affect baseline performance on the elevated-plus maze or light↔dark exploration. Galanin-overexpressing transgenic mice displayed normal baseline performance on anxiety-related tasks but displayed less anxiety-like behavior in response to the anxiogenic-like actions of yohimbine. GalR1 and GalR2 knockout mice displayed a significant anxiolytic-like phenotype on the elevated-plus maze task but not on other anxiety-related tasks.

Centrally administered galanin increases immobility on the forced-swim test in rats, a finding replicated by several laboratories. Galanin-overexpressing

transgenics generated with a platelet-derived growth factor B promoter displayed increased immobility on the forced-swim test, particularly at older ages. However, galanin transgenics generated on a dopamine β-hydroxylase promoter and GalR1 knockout mice were not significantly different than wild-type littermate controls on the tail suspension test. The inhibitory actions of galanin on the release of serotonin, norepinephrine, and dopamine, as well as the effects of galanin on serotonin receptors, may mediate its depression-like effects. Antidepressant-like and anxiolytic-like actions of a GalR3 antagonist in rats, mice, and guinea pigs have led to clinical trials, presently in progress.

Analgesia

Intrathecally administered galanin produces analgesic actions on the C-fiber flexor reflex in rats. Systemically administered galanin increased latencies on the hot-plate and tail flick tests in rats and mice, and potentiated morphine-induced antinociception. Galanin mRNA was upregulated in neurons of the rat dorsal root ganglia after sciatic nerve transaction and constriction, and after persistent inflammation. Galanin-overexpressing transgenic mice displayed longer latencies to respond to noxious heat stimulation and showed attenuated responses to repetitive C-fiber stimulation. GalR1-knockout mice showed a somewhat longer duration of mechanical hypersensitivity after sciatic nerve injury in a neuropathic pain model (Table 1). Agonists of galanin receptors are under consideration for the clinical treatment of chronic pain.

Neuroprotection

Upregulation of galanin mRNA is highly significant in dorsal root ganglion and some brain nuclei after nerve damage, including sciatic nerve ligation, colchicine treatment, and site-specific excitotoxin lesions of brain structures. Knockout mice deficient in galanin showed slower rates of nerve regeneration and reduced neurite outgrowth of cultured dorsal root ganglia neurons after sciatic nerve crush. In GalR1 knockout mice, neurite outgrowth was normal in cultured dorsal root ganglia neurons after sciatic nerve damage. However, GalR2 knockout mice had fewer surviving dorsal root ganglion neurons than controls after sciatic nerve transaction, suggesting a role for the GalR2 receptor subtype in neuroprotection. Higher levels of galanin promoted the survival of hippocampal neurons after kainic acid lesions, possibly through the inhibition of glutamate release. Long-term neuroprotective actions of endogenously administered galanin in the adult central nervous system remain controversial.

Peripheral Nervous Systems Actions of Galanin

Galanin and its receptors are localized in neurons of the sympathetic nervous system and peripheral sensory nerves. Galanin immunoreactivity is found in the epithelial cells of the colon, myenteric plexus of the gastrointestinal tract, stomach, liver, pancreatic islet cells, vas deferens, heart, carotid artery, dorsal root ganglia, trigeminal ganglia, skin, and pituitary gland. As in the central nervous system, the biological actions of galanin are generally inhibitory. Release of insulin from pancreatic islet cells was inhibited by galanin in some species, implicating galanin in metabolism, liver function, diabetes, and obesity. Actions of galanin on gastrointestinal motility included inhibiting contraction of circular smooth muscle in the ileum, increasing basal tone in the lower esophageal sphincter, and causing smooth muscle relaxation in the internal anal sphincter. Experimentally induced inflammation was found to upregulate galanin GalR1 expression in intestinal smooth muscle cells. Colonic fluid secretion in response to inflammation was reduced in GalR1 knockout mice, implicating galanin in colitis. In some species, galanin inhibited cholinergic innervation of the heart, thereby attenuating parasympathetic slowing of heart rate (Table 1).

Progenitor Cells

Galanin was identified by serial analysis of gene expression (SAGE) as being highly expressed in undifferentiated mouse pluripotent stem cells, compared to differentiated embryonic tissues. Galanin was the third most abundant transcript in these embryonic stem cells compared to other SAGE catalogs. Galanin expression was transiently induced in oligodendrocyte progenitor cells in the rat cortex. Several classes of tumors and cancer cell lines expressed galanin-like immunoreactivity and galanin mRNA, including neuroblastomas, gliomas, pheochromocytomas, melanomas, and testicular germ cell tumors. Galanin expression in the anterior pituitary was dramatically upregulated by estrogen treatment. Galanin was secreted from pituitary tumor cultures, and galanin levels were high in estrogen-induced prolactinomas, pituitary adenomas, and adrenocorticotropin-secreting tumors that cause Cushing's syndrome. A cocktail of galanin, serotonin, and octreotide was reported to be efficacious in treating colorectal cancer.

Although the neurotransmitter actions of galanin in the central and peripheral nervous systems have received the greatest attention, the actions of galanin on endocrine, proliferative, and neurodegenerative processes are increasingly intriguing.

Acknowledgment

Supported by the National Institute of Mental Health Intramural Research Program.

See also: Herbal Products and GABA Receptors; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Discovery; Neuropeptides: Food Intake; Neuropeptides: Mental Disease; Neuropeptides: Pain.

Further Reading

Belfer I, Hipp H, McKnight C, et al. (2006) Association of galanin haplotypes with alcoholism and anxiety in two ethnically distinct populations. *Molecular Psychiatry* 11: 301–311.

Berger A, Santic R, Almer D, et al. (2003) Galanin and galanin receptors in human gliomas. *Acta Neuropathologica (Berlin)* 105: 555–660.

Hökfelt T (2005) Galanin and its receptors. In: Hökfelt T, and Crawley JN (eds.) *Special Issue: Third International Symposium on Galanin and its Receptors Neuropeptides,* 39, pp. 125–142.

Hökfelt T, Bartfai T, and Crawley JN (eds.) (1998) *Special Issue: Galanin: Basic Research Discoveries and Therapeutic Implications Annals of the New York Academy of Sciences* 863.

Hökfelt T, Bartfai T, Jacobowitz D, and Ottoson D (1991) *Wenner-Gren International Symposium Series, Vol. 58: Galanin: A New Multifunctional Peptide in the Neuro-endocrine System.* Cambridge, UK: Macmillan.

Hökfelt T, and Crawley JN (eds.) (2005) *Special Issue: Third International Symposium on Galanin and its Receptors. Neuropeptides,* 39(3).

Holmes A and Picciotto MR (2006) Galanin: A novel therapeutic target for depression, anxiety disorders and drug addiction? *CNS & Neurological Disorders – Drug Targets* 5: 225–232.

Leibowitz SF (2005) Regulation and effects of hypothalamic galanin: Relation to dietary fat, alcohol ingestion, circulating lipids and energy homeostasis. *Neuropeptides* 39: 327–332.

Mazarati AM (2004) Galanin and galanin receptors in epilepsy. *Neuropeptides* 38: 331–343.

Mufson EJ, Counts SE, Perez SE, and Binder L (2005) Galanin plasticity in the cholinergic basal forebrain in Alzheimer's disease and transgenic mice. *Neuropeptides* 39: 233–237.

Robinson JK, Bartfai T, and Langel Ü (2006) Galanin/GALP receptors and CNS homeostatic processes. *CNS & Neurological Disorders – Drug Targets* 5: 327–334.

Rustay NR, Wrenn CC, Kinney JW, et al. (2004) Galanin impairs performance on learning and memory tasks: Findings from galanin transgenic and GAL-R1 knockout mice. *Neuropeptides* 39: 239–243.

Swanson CJ, Blackburn TP, Zhang X, et al. (2005) Anxiolytic- and antidepressant-like profiles of the galanin-3 receptor (Gal3) antagonists SNAP 37889 and SNAP 398299. *Proceedings of the National Academy of Sciences of the United States of America* 102: 17489–17494.

Tarasov KV, Tarasova YS, Crider DG, Anisimov SV, Wobus AM, and Boheler KR (2002) Galanin and galanin receptors in embryonic stem cells: Accidental or essential? *Neuropeptides* 36: 239–245.

Xu XJ, Hökfelt T, Bartfai T, and Wiesenfeld-Hallin Z (2000) Galanin and spinal nociceptive mechanisms: Recent advances and therapeutic implications. *Neuropeptides* 34: 137–147.

Zachariou V, Brunzell DH, Hawes J, et al. (2003) The neuropeptide galanin modulates behavioral and neurochemical signs of opiate withdrawal. *Proceedings of the National Academy of Sciences of the United States of America* 100: 9028–9033.

Game Theory and the Economics of Animal Communication

R A Johnstone, University of Cambridge, Cambridge, UK

© 2009 Published by Elsevier Ltd.

Introduction

One may contrast signals with cues – traits that influence the behavior of others but have not been favored by selection for this reason. The rustling of a mouse running through long grass may, for instance, draw the attention of a predatory owl; the sound here serves as a cue to the location of the mouse, but it does not represent a signal (nor is this an instance of communication), because it is not adapted for this purpose. Selection has not favored mice who attract the notice of predators in this way; if anything, quieter individuals who minimize the noise they make are more likely to survive. By contrast, a display trait such as the song of a nightingale, the roar of a red deer, or the train of a peacock has been favored by selection precisely because of the responses it evokes, and thus it constitutes a true signal.

Selection of Signals – Efficacy versus Strategy

The first requirement for an effective signal, in any context, is that it should be readily detectable and identifiable by the appropriate recipient. Yet communication may often take place over long distances or in a noisy environment, so by the time they reach the receiver, signals are attenuated, degraded, and overlaid by irrelevant stimuli. Selection will therefore tend to favor conspicuous and distinctive signals that are easier to detect and identify, even in the presence of noise. At the same time, such displays may be energetically costly or time consuming to produce and may draw the unwanted attention of 'eavesdroppers' such as predators and parasites. The displays we see in nature should be those that strike the best balance between the opposing selective pressures for greater efficacy and lower cost.

Many studies have looked at the design of animal signals from this perspective, focusing on the features that enhance signal efficacy. There is evidence, for instance, that the details of visual and vocal signal design are often adapted to the particular physical environment in which communication occurs, and to the sensory capacities and 'psychology' of receivers, so as to facilitate efficient detection. But the emphasis in this article is on a different aspect of signal design. It is not enough that a display should be readily detectable. If it is to be favored by natural selection, it must also, once detected, tend to elicit responses that are beneficial for the signaler. This 'strategic' perspective on signal evolution emphasizes those properties of a display that render it effective in eliciting a beneficial response rather than those that facilitate detection.

To understand the strategic aspect of signal design, we must consider selection from the perspective of both the signaler and the receiver. Selection favors signalers that elicit favorable responses from receivers; it also favors receivers that respond appropriately to signalers. Game theory, with its emphasis on strategic interaction, is thus a powerful tool for modeling this aspect of signal evolution.

Conflicts of Interest and the Problem of Honesty

In some 'cooperative' contexts, the recipient of a signal also gains by responding in a manner that is beneficial for the signaler. Under these circumstances, the evolutionary interests of the two parties are aligned. Consider, for example, the 'dance language' of honeybees, by means of which a worker returning to the hive can convey to her sisters the location of a food source. The worker benefits by recruiting others to the food, just as they stand to gain by responding to the information she provides. Selection on both parties thus favors efficient communication.

In most cases, however, there is an evolutionary conflict of interest between signaler and receiver because the latter suffers by responding in a manner beneficial to the former. Bolas spiders, for instance, attract male insect prey by means of a pheromonal lure that mimics the sex pheromones of females of the prey species. Responding to such a chemical 'lure,' while it may benefit the spider, is clearly detrimental to the prey.

The conflict in the above case is extreme, but even in situations where there is a degree of common interest between signaler and receiver, there will usually be an element of conflict too. For instance, both males and females share a common interest in finding and identifying potential mates of their own species. However, males are typically under stronger selection to acquire many mates, and females to acquire better mates. This gives rise to a conflict of interest. Females stand to gain by accurately assessing the mating benefits (material or genetic) that a potential partner has to offer, in order to reject suitors of low value. Conversely, males stand to gain by misleading females as to their own value, in order to gain additional matings. In a similar way, conflict arises between

parents and offspring, even though both have a genetic stake in each other's survival. Offspring are selected to demand more food than parents are selected to provide. As a result, parents stand to gain by accurately assessing offspring need, but offspring stand to gain by misleading parents as to their own condition in order to extract additional resources from them.

The first theoretical game models of animal communication suggested that because of the underlying conflict of interest between signaler and receiver, informative or honest signaling cannot persist. Consider, for instance, a population in which signalers reliably advertise their aggressive intentions by means of a threat display. Maynard Smith, using game theory to model conflict behavior, argued that threat displays that provided receivers with accurate information about an animal's intentions would not be evolutionarily stable. If a display provides honest information about an opponent's propensity to fight, receivers should attend to it and should be more likely to back down when they confront an opponent who signals strongly. Consequently, it seems that a 'deceitful' mutant that always announced a high level of aggression would prosper because its opponents would more often retreat. The spread of such mutants would, however, devalue the signal. Ultimately, it would no longer pay to attend to the display, because it would no longer provide reliable information.

The same argument seems to apply to any signaling system in which there is a conflict of interest between signaler and receiver. Any correlation between the production or intensity of a display and the underlying state of the signaler can be disrupted by the spread of a 'deceitful' mutant that gains by adopting a misrepresentative signal. Consequently, the models seemed to suggest that communication should be viewed, not as a cooperative exchange of information, but as the focus of an arms race between signalers as 'manipulators' and receivers as 'mind-readers,' the former standing to gain by effective exploitation, the latter by effective extraction of information. Honesty, from this perspective, merely represented a temporary phase of the arms race in which receivers have 'drawn ahead' of signalers.

Constraints and Costs – The Handicap Principle

Despite the potential benefits that signalers stand to gain by misleading receivers, the majority of signals seem to convey at least some reliable information. In other words, stable honesty is the usual outcome of the arms race between signaler and receiver. How can such honesty persist in the face of the evolutionary conflict between the two?

Honesty may endure if it is physically impossible for a signaler to produce a misleading signal. In some cases there is a material link between the signal and some underlying aspect of the signaler's state, such that the former is constrained to provide information about the latter. For instance, many animals display bright coloration attributable (in part at least) to carotenoid pigments. Since most cannot synthesize carotenoids themselves but must acquire them as part of their diet, the intensity of carotenoid coloration is inescapably linked to (past) foraging success. An individual that fails to acquire sufficient carotenoids through foraging simply cannot manufacture the pigments needed for vivid color. Studies of the house finch (*Carpodacus mexicanus*), for instance, a species in which males display patches of carotenoid pigment on the crown, breast, and rump, have shown that females prefer redder males and that vivid coloration reflects nutritional state at the time of molt. Since molting precedes mate choice by some time, redder males may not necessarily be in better condition when they are actually chosen as partners, but plumage coloration was correlated with nest attentiveness and overwinter survival, suggesting that brighter males do make, on average, better fathers.

While a few signals may be 'unfakable' because of such physical constraints, most are not. An individual's choice of one threat display or another, for instance, is not physically constrained by its fighting ability or motivation, nor is the length of a bird's tail inescapably tied to its value as a mate. To account for widespread honesty, therefore, an explanation is required that is applicable to situations in which the relationship between the signal and the underlying state of the signaler is open to behavioral or evolutionary change.

Amotz Zahavi was the first biologist to propose such a general explanation for honesty. He suggested that a signal could provide reliable or 'honest' information about the quality of the signaler if it were costly to produce. Consider, for instance, a population in which males reliably advertise their value as mates by means of some sexual display. The correlation between signal and mate value would seem vulnerable to disruption by the spread of a deceitful mutant that adopts the display typical of the most desirable males, regardless of its own quality. Suppose, however, that the signal is costly to produce but that superior males can more easily bear these costs than can inferior males. Under these circumstances, it might pay an inferior male to refrain from signaling, even though it would be physically able to do so, because the costs involved would outweigh the benefits to be gained. At the same time, for superior males who can more easily bear the costs of display,

production of the signal would yield a net benefit. Under these circumstances, honesty can persist.

While Zahavi initially focused chiefly on sexual displays, he has subsequently argued that the handicap principle is applicable to many and perhaps all forms of communication. Costly signals could serve to advertise fighting ability, for instance, and thereby to deter rivals, or could deter predators by advertising ability to escape or fend off attack. Costly contributions to group activities could, in a similar way, advertise social status, provided that only dominant, high-quality individuals can afford to invest time and energy in such pursuits. In short, receivers in most contexts can obtain reliable information about signaler quality by attending to displays that are costly to produce.

A Simple Model

Zahavi's handicap principle suggests a possible solution to the problem of honesty. At first, however, it was widely dismissed as unworkable. In part, this rejection of the idea was due to Zahavi's verbal formulation, which was open to different interpretations, leaving some doubt as to precisely what was meant. Acceptance had to wait until the development of formal models, in many cases based on game theoretical techniques. Indeed, the problem of honesty has been, and remains, the chief focus of game theoretical models of animal communication, and such models have played a key role in establishing the plausibility of honest, costly advertisement (it must also be noted that economic game theory had reached similar conclusions rather earlier).

To illustrate this approach, this article presents a very simple (perhaps the simplest possible) game theoretical model of honest signaling. Suppose that one individual, the signaler, is informed of the value of an uncertain parameter q (possibly representing some aspect of the external environment, but more likely some aspect of its own type or state). The signaler than chooses an action s, referred to as a signal. Another individual, the receiver, who is unaware of the value of the parameter q, observes the signal s and decides, on this basis, to adopt some response r. In our simple case, q, s, and r all take values of either 0 or 1. Each player then receives a payoff that may depend on q, s, and r.

The problem of honesty arises when receivers gain by responding differently to signalers of different type, but at least some signalers stands to gain by eliciting a response that is, from the perspective of the receiver, inappropriate for their type. Suppose, for instance, that the receiver obtains a payoff of 1 if its response matches the signaler's type (i.e., if $r = q = 0$ or $r = q = 1$), but a payoff of 0 otherwise (if $r \neq q$). All signalers, by contrast, obtain a payoff equal to r and therefore prefer a response of 1 to a response of 0. How is a separating equilibrium, at which signalers of different types adopt different signals, possible under these conditions? Any such equilibrium appears vulnerable to invasion by a rare class of mutant signaler that, regardless of its true type, adopts the signal typical of individuals of type $q = 1$, thereby eliciting the favorable response $r = 1$.

Suppose, however, that the signal adopted by individuals of type 1 is costly to produce. The handicap principle suggest that this can serve to maintain honesty. For simplicity, let us restrict our attention to the potential separating equilibrium at which individuals of type $q = 1$ adopt the signal $s = 1$ (eliciting the response $r = 1$), while individuals of type $q = 0$ adopt the signal $s = 0$ (eliciting the response $r = 0$). Now suppose that production of the signal $s = 1$ entails an additive cost of magnitude c_i (≥ 0) for signalers of type i, whereas the signal $s = 0$ is cost-free. Provided that $c_0 > 1 > c_1$, the separating equilibrium is stable against invasion by the cheating mutant suggested above because the cost of adopting signal $s = 1$ outweighs the benefits to be gained thereby for individuals of type 0 but not for those of type 1. Cheating in this case may be possible, in the sense that a mutant individual can arise that adopts signal $s = 1$ when of type $q = 0$, but it proves unprofitable.

In the above case, signalers differ in the cost they pay to produce the 'favored' signal $s = 1$, but all stand to gain the same benefits by eliciting the 'favorable' response $r = 1$. Suppose, by contrast, that all signalers pay the same cost, of magnitude 1, to produce the favored signal, but that the benefit they stand to gain by eliciting the favorable response $r = 1$, denoted b_i for signalers of type i, differs according to their type. Provided that $b_1 > 0 > b_0$, the separating equilibrium is once again stable against invasion by the cheating mutant because the signal $s = 1$ is unprofitable for individuals of type $q = 0$. In this case, however, honesty is maintained not by the differential cost of the signal but by differential benefit. The former case is often described as a signal of quality, the latter as a signal of need.

In reality, of course, signalers may vary both in the cost of signaling and in the benefits they stand to gain, and some more-realistic models have incorporated both factors (though surprisingly many analyses continue to focus exclusively on one or the other).

Evidence for Honesty

Even the very simple model presented above suffices to demonstrate that signal cost can, in principle at least, serve to maintain honesty. But is there any evidence that the handicap principle plays such a role in nature? This article offers two illustrative

examples of the kind of studies that have been undertaken to confirm its importance.

Perhaps the best evidence for honest, costly advertisement of quality comes from work by Mappes and colleagues on sexual display in the wolf spider *Hygrolycosa rubrofasciata*. Males of this ground-dwelling species attract mates by drumming on dry leaves on the substrate with their abdomen. Experiments have shown that drumming entails energetic costs – males kept under standardized conditions lose weight faster and experience greater mortality when induced to drum more often by exposure to females. At the same time, those males who were observed to drum more frequently and vigorously under natural conditions suffer less when subsequently induced to drum in the lab, indicating that the costs of the display are indeed lower for those males that signal more strongly, as the handicap principle predicts.

Turning to advertisement of need, some of the best evidence for honest, costly signaling comes from work by Kilner on the begging displays of captive canary (*Serinus canaria*) nestlings. When an adult canary returns to the nest with food, the offspring stretch and posture, gape and call as the adult empties its crop and distributes the regurgitated food among the brood. Chicks that attain positions closer to the parent are likely to receive more food, but when offspring proximity to the parent was experimentally controlled and hunger manipulated, parents were observed to preferentially feed hungrier chicks, suggesting that they respond to cues indicative of offspring need. What form might these cues take? The mouths of canary nestlings exhibit a rapid change in color intensity, the 'flush,' following the onset of begging, and subsequent experiments revealed that this change in color accurately reflects a nestling's state of need and that parents preferentially feed chicks with artificially reddened mouths. Gape color thus appears to provide parents with an honest signal of chick hunger (although parents may also respond to chick posturing and calling as well). Finally, when pairs of canary siblings were induced to beg for markedly different durations, greater begging was found to retard growth, both immediately and in the long term. This growth cost is likely to entail a reduction in fitness, because daily mass gain was strongly correlated with survival of chicks to independence. It thus appears that chick begging in this species constitutes a costly, as well as an honest, signal of hunger.

Occasional Deceit

The handicap principle seems to offer a simple and widely applicable solution to the problem of honesty. But if receivers can obtain reliable information about signaler quality or condition merely by attending to costly displays, how can one account for instances of deceit?

Selection will favor receivers who respond to a signal if, on average, they obtain a net benefit by doing so. But a benefit 'on average' is compatible with occasional exploitation. Suppose, for instance, that a rare class of signaler finds display cheaper than other individuals do, for the same level of quality, or stands to gain more by eliciting a favorable response, for the same level of condition. Such individuals do best to produce a 'misleading' signal, typical (in most cases) of signalers of higher quality or in greater need than themselves. This will lead receivers to respond more favorably to them than to other signalers. We might expect receivers to adjust their behavior to take account of the presence of these deceptive individuals – the risk of encountering one will favor lower responsiveness to intense displays. Nevertheless, provided that the frequency of such individuals is low enough, and the cost of responding to them inappropriately is not too great, they need not disrupt the signaling system altogether. Evidence of occasional deceit can thus support the handicap principle if it can be established that deceitful signalers do differ from others in the net benefit they stand to gain from display (and if there are constraints on the frequency of deceit).

A good example is the use of deceptive begging signals by avian brood parasites such as cuckoos. As discussed above, there is evidence that chick begging constitutes (in at least some cases) an honest signal of hunger. Models of begging predict that chicks who are less closely related to their nestmates and to the adults raising them should beg more intensely because they have less to lose by depriving their competitors of food and/or forcing the adults to work harder. A brood parasite chick is an extreme instance of this: Since it is entirely unrelated to the other chicks in the nest or to its host parents, it has no genetic interest in their survival. Accordingly, Kilner and colleagues have shown that chicks of the common cuckoo exploit the signaling system of reed warbler hosts by begging at a greater rate than do host chicks. Intriguingly, parents appear to respond to both the vocal and the visual signals of their offspring. Cuckoo chicks, which eject all host young from the nest, display a reduced gape area compared with a host brood yet apparently compensate by more rapid calling.

Exploitation of host parents by specialized brood parasites may appear an extreme and unusual instance of deceit. However, more subtle forms of intraspecific deception may be widespread. A nice illustration is the occurrence of bluffing in the marine stomatopod crustacean *Gonodactylus bredini*, studied by Adams and

Caldwell. Males of this species defend cavities in coral rubble and ward off intruders with a meral spread threat display. Following molt, an individual's soft condition means that it is unable to fight effectively, yet newly molted owners threaten more often than those that are capable of fighting, thus reducing the chances that intruders will attack and discover their opponents' vulnerability. That this deceit does not destabilize the signaling system entirely presumably reflects the relativey low probability of encountering a newly molted opponent.

Cost-Free Honesty?

The evidence discussed earlier shows that signal cost can indeed, as the handicap principle suggests, favor quality-dependent or condition-dependent expression of display traits. Cost, in other words, is sufficient to maintain honesty. But is it necessary? The more-enthusiastic proponents of the handicap principle have argued that honest signals must be costly, otherwise what is to prevent cheating? But more-recent analyses suggest that cost-free honesty is possible. Two rather different mechanisms have been proposed that might maintain signal reliability even in the absence of cost.

The first mechanism was suggested in the context of signaling of need by offspring to their parents. As pointed out above, offspring are selected to demand more resources than parents are selected to supply. Consequently, an honest signal of need is potentially subject to devaluation by deceitful mutants that misrepresent their need, thereby attaining additional resources. However, since offspring have a genetic stake in the well-being of their parents, selection will not favor unlimited selfishness on the part of young. Offspring stand to gain by simulating a level of need that is slightly greater than their true need, but not so great that they might elicit excessive investment that would be overly costly to the parent. Selection, in other words, will act against offspring that are too deceptive.

As a result of selection against excessive deception, a cost-free signaling equilibrium is possible at which offspring divide into two or more discrete categories, according to their need, in each of which all individuals adopt the same begging behavior (sometimes referred to as a 'pooling' equilibrium). In the simplest case, for instance, those whose need falls below some threshold refrain from begging while the remainder beg. Under these circumstances, the parent cannot distinguish among nonbegging young and must attribute to them all the same 'average nonbegging' level of need; similarly, the parent cannot distinguish among begging young and must attribute to them all the same 'average begging' level of need (greater than the average for nonbegging young). A nonbegging offspring may thus stand to gain by exaggerating its need slightly, but it cannot do so – by choosing to beg, it will cause the parent to greatly overestimate its level of need, to an extent that can prove disadvantageous, even if begging entails no direct costs.

Formal game-theoretical models have confirmed that a cost-free equilibrium of the above kind is possible in principle. Whether actual chick begging represents a cost-free or a costly signal remains, however, controversial. As described earlier, there is evidence that begging entails fitness costs in at least some cases. But in other instances, the costs are less obvious – most attempts to measure the energetic cost of begging, for instance, suggest that this is very small. Similarly, while experiments have shown that playback of begging calls can increase the risk of nestling predation, this result holds true only for offspring in open nests and perhaps overlooks the fact that parents may normally be able to warn their offspring to remain silent when predators are nearby, so reducing the effective cost of begging. Despite the paucity of evidence for substantial costs, however, offspring typically do not exhibit 'pooling' but instead show continuous variation in begging intensity, which better fits the predictions of costly signaling models. Perhaps, in reality, begging may involve both costly and cost-free components – it is certainly a composite signal, often combining posturing, calling, and gaping.

In any case, 'pooling' models of cost-free begging are based on the fact that signalers (offspring) have a genetic stake in the well-being of receivers (their parents). Because this is not the case in many other signaling systems, such models are of restricted applicability. A second mechanism for the maintenance of cost-free honesty, however, is of much broader scope. Looking back at the very simple model we introduced earlier, we showed that in the case of a signal of quality, stable honesty is possible provided that $c_1 > 1 > c_2$, that is, provided that the cost of signaling exceeds the benefit to be gained for individuals of class 1, but not for individuals of class 2. This condition allows for the possibility that $c_2 = 0$, that is, that the signal is entirely cost-free for individuals of class 2. Because it is only these individuals that employ the signal at equilibrium, the outcome is then one in which no cost is paid. What maintains honesty (in the face of a conflict of interest between signaler and receiver) is the fact that low-quality individuals 'would' incur a cost 'if they were to adopt' the display behavior typical of a high-quality signaler. This does not mean, however, that such display behavior must entail a cost for the high-quality individuals that actually employ it.

The above argument suggests that cost-free honesty is possible in any context, provided that cost is incurred only by those signalers that misrepresent their state. It is difficult to envisage a physical mechanism that could give rise to such conditional costs. When costs are socially imposed, however, it does seem plausible that receivers might punish only those signalers that are subsequently discovered to have employed misrepresentative or deceptive signals. For example, rhesus monkeys living on the island of Cayo Santiago, Puerto Rico, often (though not always) produce distinctive vocalizations when they discover or eat food, particularly high-quality food such as coconut. Following the discovery of food by one individual, other group members were in some cases observed to detect and approach the food source and its original discoverer. Some discoverers were then attacked and injured, whereas others were allowed to keep the food they had collected. It appears that individuals who call when they discover food are more likely to be approached by other group members, and as a result suffer from increased feeding competition. However, discoverers who call incur less aggression when subsequently detected than do those discoverers who remain silent. In this case, then, the threat of aggression directed at 'deceptive' individuals who refrain from calling may serve to maintain honest advertisement of food discoveries.

Further Questions

This brief review makes clear, the important role that game-theoretical models have played, and continue to play, in the study of animal communication. Because most forms of communication involve an evolutionary conflict of interest between signaler and receiver, they are potentially vulnerable to disruption by 'dishonest' signalers. Game-theoretical models have served to verify the logic of Zahavi's handicap principle and to confirm that signal cost may indeed support honest signaling in the face of a conflict of interest, a suggestion for which there is now some empirical support. At the same time, such models have also suggested that honesty need not always entail signal cost. This last issue remains the subject of much controversy, and the study of putative cases of cost-free honesty continues to attract much interest. There are also many further questions that fall within the scope of game theory, but which have not been touched on here. Three topics that are of particular interest, for instance, are the evolution and use of multiple signals, the dynamics of extended communicatory interchanges, and the evolution of signaling within a communication network in which multiple receivers may eavesdrop on a signal (and multiple signalers may compete for the attention of a receiver). Given the many questions that remain to be answered, it seems likely that game theory still has much to contribute to the study of animal communication.

See also: Animal Communication: Honesty and Deception; Communication Networks and Eavesdropping in Animals; Electrical Perception and Communication; Endocrinology of Animal Communication: Behavioral; Kinship Signals in Animals; Pheromones and other Chemical Communication in Animals; Referentiality and Concepts in Animal Cognition; Seismic and Vibrational Signals in Animals; Sexual Selection and the Evolution of Animal Signals; Signal Design Rules in Animal Communication; Visual Signaling in Animals.

Further Reading

Adams ES and Caldwell RL (1990) Deceptive communication in asymmetric fights of the stomatopod crustacean *Gonodactylus bredini*. *Animal Behavior* 39: 706–716.

Bergstrom CT and Lachmann M (1998) Signalling among relatives. III. Talk is cheap. *Proceedings of the National Academy of Sciences of the United States of America* 95: 5100–5105.

Grafen A (1990) Biological signals as handicaps. *Journal of Theoretical Biology* 144: 517–546.

Hauser MD and Marler P (1993) Food-associated calls in rhesus macacques (*Macaca mulatta*). II. Costs and benefits of call production and suppression. *Behavioral Ecology* 4: 206–212.

Hill GE, Inouye CY, and Montgomerie R (2002) Dietary carotenoids predict plumage colouration in wild house finches. *Proceedings of the Royal Society of London, Series B* 269: 1119–1124.

Johnstone RA and Grafen A (1993) Dishonesty and the handicap principle. *Animal Behaviour* 46: 759–764.

Kilner RM (2001) A growth cost of begging in captive canary chicks. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11394–11398.

Kilner RM, Noble DG, and Davies NB (1999) Signals of need in parent-offspring communication and their exploitation by the common cuckoo. *Nature* 397: 667–672.

Lachmann M, Szamado S, and Bergstrom CT (2001) Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences of the United States of America* 98: 13189–13194.

Mappes J, Alatalo RV, Kotiaho J, and Parri S (1996) Viability costs of condition-dependent sexual male display in a drumming wolf spider. *Proceedings of the Royal Society of London, Series B* 263: 785–789.

Maynard Smith J (1974) The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology* 57: 239–242.

Maynard Smith J (1991) Honest signalling: The Philip Sidney game. *Animal Behaviour* 47: 1115–1120.

Maynard Smith J and Harper D (2003) *Animal signals*. Oxford, UK: Oxford University Press.

Searcy WA and Nowicki S (2005) *The Evolution of Animal Communication*. Princeton, NJ: Princeton University Press.

Zahavi A (1975) Mate selection: A selection for a handicap. *Journal of Theoretical Biology* 53: 205–214.

Games in Monkeys: Neurophysiology and Motor Decision-Making

D Lee, Yale University School of Medicine, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Stages of Decision Making

Decision making can be defined as choosing a particular option from multiple alternatives, and it is often carried out in order to maximize certain desirable quantity, such as reward or utility. Animals rely heavily on the sensory stimuli they receive from their environment to choose appropriate behavioral responses (Figure 1). A precise pattern of sensory stimuli impinging on the animal's nervous system is determined by the positions and orientations of the animal's sensory organs as well as the current state of the animal's environment. These sensory inputs are inevitably corrupted to a variable degree by the noise in the animal's nervous system. Based on this noisy representation, individual objects in the environment are recognized and additional properties of these objects that the animals have learned from their previous experience can be retrieved from their memory storage. This is the process of perceptual decision making. Based on the outcome of perceptual decision making, the animal chooses what to do, and this is referred to as cognitive decision making. For example, should the animal eat the apple it just discovered from a tree, or save it for a rainy day? Even when all the objects in the animal's environment are correctly identified, the animal's behavioral response may change depending on the animal's metabolic needs, and the animal's cognitive limitations may prevent the animal from selecting a behavioral response that produces the most desirable outcome. Finally, the number of various movement trajectories that can be chosen to accomplish a particular behavioral objective is often infinitely large, and the selection of a particular motor trajectory can be referred to as motor decision making. For example, a light switch can be turned on by the right or left hand or even by one's head if both hands are occupied or injured. This article focuses on the neural mechanisms of cognitive and motor decision making in monkeys with a particular emphasis on cognitive decision making in the context of social interaction.

Utility Maximization and Reinforcement Learning

Standard economic analysis of choice behavior, such as the expected utility theory, begins with the assumption that decision makers assign a numerical score to each available option according to the subjective desirability of its outcome, commonly referred to as utility function. It is also assumed that decision makers make their choices so as to maximize the expected value of utility function. Numerous studies of human choice behavior have shown that these assumptions are often violated. Many of these violations are accounted for by alternative theoretical proposals, such as the prospective theory of Kahneman and Tversky and the regret theory of Bell, Loomes, and Sugden. Nevertheless, the expected utility theory still remains as a core theory in microeconomics, and accounts for a wide range of choice behavior in both humans and animals.

To understand the neural basis of decision making within the framework of the expected utility theory, it is important to distinguish two different types of utility which Kahneman referred to as decision utility and experienced utility. Decision utility corresponds to the amount of subjective pleasure expected by the decision maker during the process of decision making. In contrast, experienced utility is the amount of actual pleasure or satisfaction derived from chosen actions or from the consumption of chosen objects. If the decision maker makes correct predictions regarding the outcomes of chosen actions during the time of decision making, experienced utility would be equal to decision utility. In reality, however, the decision maker's knowledge of his or her environment is seldom complete, and there is often a discrepancy between these two types of utilities. It is, therefore, possible that decision utility and experienced utility might be represented and processed in separate brain areas. The results from previous neuroimaging studies in human subjects, however, show that decision utility and experienced utility influence the level of metabolic activity in similar brain areas, including the dorsolateral prefrontal cortex, the orbitofrontal cortex, the anterior and posterior cingulate cortex, the posterior parietal cortex, and the basal ganglia. Activity of individual neurons in many of these brain areas has been examined in nonhuman primates, and the results are largely consistent with the findings from the human neuroimaging studies. Thus, signals related to decision and experienced utilities are encoded in a broadly distributed network of cortical and subcortical areas.

If the animal's environment is stationary, the probability that a given action would produce a particular outcome would be fixed, and, therefore, optimal decision-making strategies can be hardwired in the form of stimulus–response mapping. In reality, the environment is almost always dynamic, and the actual outcomes frequently differ from the outcomes

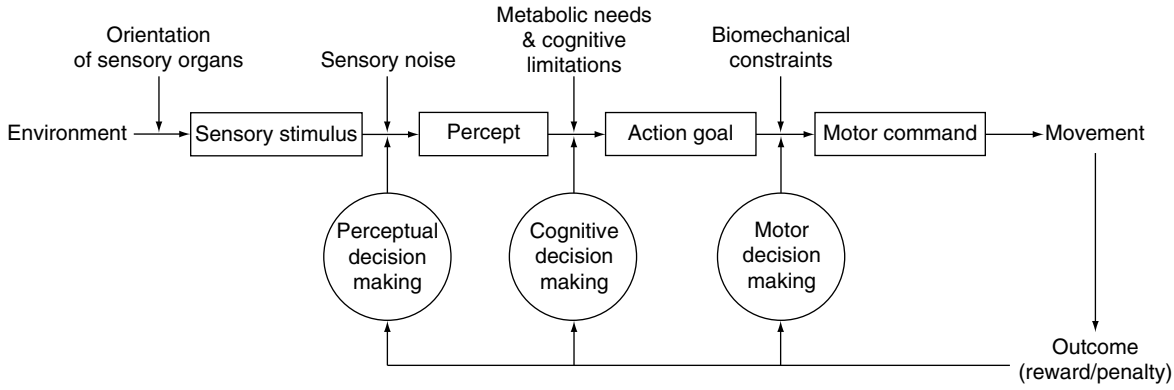

Figure 1 Different stages of decision making.

anticipated during the process of decision making, so decision utilities may deviate from the experienced utilities. To make choices optimally, therefore, decision makers need to adjust decision utilities according to their experience. In reinforcement learning theory, decision utilities are referred to as value functions and the difference between decision and experienced utilities is referred to as reward prediction error. The work of Schultz and his colleagues showed that reward prediction error is encoded by the activity of dopamine neurons in the primate ventral tegmental area and substantia nigra pars compacta. These dopamine neurons project densely to the striatum, and neuroimaging studies in human subjects found that the activity in the striatum also reflects reward prediction error. Therefore, dopamine neurons and their target neurons in the striatum might play a key role in updating and encoding the value functions during dynamic decision making.

Game Theory and Adaptive Social Decision Making

To make choices optimally, decision makers must be able to predict the outcomes of their actions correctly. For animals in a social context, the outcomes of their choices often depend on the behaviors of other animals, so they must be able to predict the choices of other animals. The mathematical problem of finding an optimal decision-making strategy in a socially interactive setting is addressed by the game theory introduced by von Neumann and Morgenstern.

In game theory, a game is characterized by a set of players, a set of available actions for each player, and a payoff function that assigns utility to each player according to the choices of all players. As in the expected utility theory, game theory assumes that players in a game seek to maximize their utility functions individually, and then provides normative predictions as to the choices that rational players should make. One such prediction is provided by Nash equilibrium, which is defined as a set of strategies for all players from which no players can deviate individually to increase their payoffs. Although game theory provides such solutions to many different types of games, many studies in behavioral economics have found that people often display systematic deviations from such theoretical predictions.

Matching Pennies Game

Barraclough and his colleagues investigated whether the choice behavior of monkeys during a simple zero-sum game, known as matching pennies (Figure 2(a)), follows the prediction of Nash equilibrium. In each trial, a monkey and a computer opponent each chose either the left-hand or the right-hand target, and the monkey was rewarded only when it chose the same target as the computer. The computer was programmed to simulate a utility-maximizing player capable of analyzing and exploiting the animal's behavior. In the matching pennies game illustrated in Figure 2(a), choosing one of the options exclusively is not a part of the Nash equilibrium. For example, if the monkey and the computer always choose the left-hand and right-hand target, respectively, their payoffs would be both zero. This is not a Nash equilibrium, since the monkey can increase its payoff by increasing the probability of choosing the right-hand target. The Nash equilibrium for this matching pennies game is for each player to choose the two targets with equal probabilities. Barraclough and his colleagues found that the overall probability that the animal would choose each of the two targets was quite close to this equilibrium prediction. However, there was still a bias for the animal to choose the same target chosen by the computer opponent in the previous trial, which was exploited by the computer to reduce the probability of reward for the animal.

a

		Computer: Left	Computer: Right
Monkey	Left	(1, –1)	(0, 0)
	Right	(0, 0)	(1, –1)

b

		Computer: Rock	Computer: Paper	Computer: Scissors
Monkey	Rock	(1, –1)	(0, 0)	(2, –2)
	Paper	(2, –2)	(1, –1)	(0, 0)
	Scissors	(0, 0)	(2, –2)	(1, –1)

c

		Computer: Inspect	Computer: Not inspect
Monkey	Work	(0.5, 2–*I*)	(0.5, 2)
	Shirk	(0, 1–*I*)	(1, 0)

Figure 2 Payoff matrix for matching pennies game (a), rock–paper–scissors game (b), and inspection game (c). Each game is played by a monkey and a computer opponent. For each payoff matrix, rows correspond to the alternative choices for the monkey, and columns correspond to the choices of the computer opponent. Two numbers inside the parentheses correspond to the payoffs or utilities for the monkey and the computer opponent, respectively. In the matching pennies game, the optimal strategy, known as Nash equilibrium, is for each player to choose the left-hand and right-hand targets randomly with equal probabilities. The equilibrium strategy in the rock–paper–scissors game is to choose each option with a 1/3 probability. In the inspection game, I indicates the cost of inspection ($0 < I < 1$), and the optimal strategy is for the computer opponent to inspect with the probability of 0.5 and for the monkey to shirk with the probability of I.

The fact that monkeys displayed small but systematic biases in their choice behavior during the matching pennies game might reflect a specific learning algorithm used by the animal to approximate the optimal decision-making strategy. Therefore, the choice behavior of monkeys during the matching pennies game was analyzed further with a reinforcement learning model. Such a model formalizes Thorndike's law of effect and generalizes the so-called win–stay–lose–switch strategy. In this model, the probability of choosing each target is determined by the logistic function of the difference between the value functions for the two targets. In addition, value functions are adjusted according to the outcome of the animal's choice in each trial, and the effect of the animal's choice in a given trial decays exponentially during the following trials. Accordingly, the value function of a target increases gradually if its choice is rewarded frequently. The animal's choice behavior during the matching pennies game was indeed consistent with the predictions from this simple reinforcement learning model.

During the matching pennies game, neurons in the dorsolateral prefrontal cortex display several different types of signals that might reflect the process of reinforcement learning. First, many neurons modulate their activity according to the animal's choice in the previous trial. In many decision-making problems, the outcome of a particular choice is often revealed after some delay. In such cases, it is necessary to maintain information about the action taken by the animal until its outcome is revealed, so that its value function can be adjusted appropriately. In reinforcement learning, such memory of a previous action is referred to as eligibility trace. Therefore, activity in the dorsolateral prefrontal cortex related to the animal's choice in the previous trial might provide the neural substrate of eligibility trace. Second, activity of neurons in the dorsolateral prefrontal cortex often reflects whether the animal was rewarded in the previous trial or not. This suggests that the dorsolateral prefrontal cortex may be involved in adjusting the value functions for the recently chosen actions, since the outcome of a given choice determines how its value function would be updated in reinforcement learning. Finally, some neurons in the dorsolateral prefrontal cortex displayed modulations in their activity according to the combination of the animal's choice in the previous trial and its outcome (Figure 3). Such signals can provide information about the choice of the computer opponent, and therefore can also contribute to the process of updating value functions, since during the matching pennies game, the choice of the opponent determines the animal's choice that would be rewarded.

Rock–Paper–Scissors Game: Reinforcement Learning versus Belief Learning

Although the choice behavior of monkeys during the matching pennies game was relatively well described by a simple reinforcement learning model, the simplicity of this game makes it difficult to test whether the animal might rely on an alternative learning algorithm. For example, in belief learning, the decision makers update their beliefs about the choices of other players, and choose the best action that maximizes the expected payoffs based on such beliefs. For the matching pennies game, reinforcement learning and belief learning models cannot be distinguished. To understand this, imagine that the animal selects the left-hand and right-hand target in two successive trials while the computer selects the right-hand targets in both trials. The animal would then be rewarded only in the second trial. In a reinforcement learning model, the value function for the left-hand target would decrease after the first trial and the value

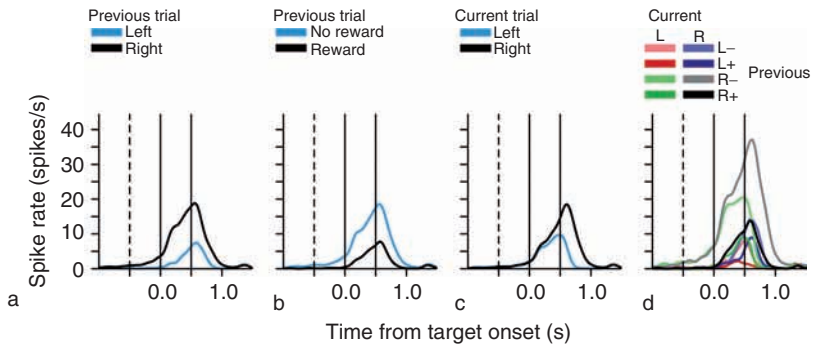

Figure 3 Temporal changes in the spike rate of a single neuron in the dorsolateral prefrontal cortex of a monkey during a matching pennies game. The activity of this neuron was modulated by the animal's choice in the previous trial (a), its outcome (b), and the animal's choice in the current trial (c). In addition, this neuron increased its activity significantly when the animal chose the right-hand target after the same choice was not rewarded in the previous trial (d). Adapted from Barraclough DJ, Conroy ML, and Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience* 7: 404–410.

function for the right-hand target would increase after the second trial. Accordingly, after two such trials, the animal would be more likely to choose the right-hand target. In a belief learning model, the animal would strengthen its belief that the computer opponent would choose the right-hand target, and therefore the value function for the right target would increase after each trial. As in reinforcement learning, this would also increase the probability that the animal would choose the right-hand target. These two scenarios would become indistinguishable if the value functions are related to the probability of choosing each action through a logistic function that determines the animal's choice as a function of the difference between the two value functions.

The difference between reinforcement learning and belief learning can be revealed if a player has to choose from more than two alternatives. In such cases, belief learning might be a more efficient way to improve the strategy of a decision maker compared to reinforcement learning. This can be illustrated using the example of a rock–paper–scissors game (**Figure 2(b)**). In reinforcement learning, if a player chooses rock and is defeated by the opponent who chooses paper, the value function for rock and the probability of choosing it would decrease, but the value functions for paper and scissors would not be affected. In belief learning, the player would strengthen his or her belief that the opponent would choose paper again, so the value function for scissors and the probability of choosing scissors would increase more than for paper. It should be noted that if the opponent plays according to the Nash equilibrium strategy, the average payoff to the other player would be fixed regardless of his or her strategy and regardless of the learning algorithm used to update the strategy. However, if the opponent has a bias toward a particular choice, then belief learning may lead more efficiently to the strategy that can exploit such a bias.

The experienced-weighted attraction (EWA) model proposed by Camerer and Ho incorporates the features of both reinforcement learning and belief learning model according to a weight parameter that specifies the relative contribution of each learning algorithm. They found that the EWA model was more consistent with the results obtained from experimental games in human subjects compared to either reinforcement learning or belief learning models. Consistent with these findings, the choice behavior of monkeys during a computerized version of a rock–paper–scissors was also better accounted for by a model that incorporated the features of both reinforcement learning and belief learning.

Inspection Game

When multiple decision makers interact in a dynamic environment, their payoff matrix and therefore the optimal decision-making strategies might change unpredictably. Dorris and Glimcher investigated the abilities of monkeys to adapt to such unpredictable changes in the payoff matrix during an inspection game. Similar to matching pennies game, each of the two players in this game has two alternatives. For the monkey, these two targets correspond to work and shirk, respectively (**Figure 2(c)**). If the animal chooses the 'work' target, it receives a fixed amount of reward. Otherwise, it receives a larger or smaller amount of reward, depending on whether the computer decides to inspect or not. The Nash equilibrium for the computer is to inspect with the probability of 0.5, whereas the optimal strategy for the monkey is to shirk with the probability equal to the inspection cost (I). Although the probability that the animal would choose each of the two targets roughly followed the

predictions of the Nash equilibrium, there was still a systematic discrepancy between the theoretical prediction and the animal's choice behavior. Specifically, the animal chose to shirk more often than the equilibrium prediction when the inspection cost for the computer opponent was small. It should be noted that during this inspection game, the animal was always rewarded by the same amount of reward for choosing to work, and therefore did not receive any information as to whether the computer opponent chose to inspect or not. Therefore, choosing to shirk might have additional advantage in that it provides some information about the behavior of the opponent, and this may explain why the animal chose to shirk more frequently than the Nash equilibrium solution, especially after an unpredictable change in the payoff matrix.

The lateral intraparietal area (LIP) is a region in the primate posterior parietal cortex closely connected with other oculomotor centers, such as the superior colliculus and the frontal eye field. Neurons in the LIP often modulate their activity according to the direction of eye movement during its preparation and execution, and might be actively involved in the process of decision making. For example, activity of LIP neurons is systematically related to the probability and magnitude of reward expected from certain eye movements. The possibility that the activity of LIP neurons reflects the subjective value of the expected outcome from an eye movement toward its receptive field was tested using the inspection game paradigm. Although the decision to shirk leads to a variable outcome depending on the choice of the opponent, the expected utilities for shirking and working must be equal at the Nash equilibrium. This prediction was tested by requiring an animal to indicate its choice to shirk by making an eye movement toward a visual target in the receptive field of an LIP neuron. Although the probability of making such eye movements varied according to the inspection cost and the corresponding Nash equilibrium, the activity of LIP neurons remained relatively constant. In addition, LIP activity was significantly correlated with the trial-by-trial variation in the utility of choosing the target in the neuron's receptive field. These results are consistent with the hypothesis that LIP neurons encode subjective utilities of eye movements.

Utility and Cost Functions in Motor Control

In combination with appropriately chosen utility functions, the principle of utility maximization can account for a wide range of choice behaviors in both humans and animals. Similarly, changes in the choice behavior during social interactions and other behavioral changes in a dynamic environment can be modeled with reinforcement learning algorithms. In order to produce the physical movement in the animal's body, however, the brain has to translate its chosen behavioral goal into a specific pattern of muscle activations.

Due to a large degree of redundancy in the motor system, there are practically an infinite number of muscle activation patterns that can accomplish more or less the same desired final configuration of the animal's body. For example, when a person or a monkey reaches for an object, either of the two hands can be used, and often a variety of grasp postures can be chosen. When monkeys are allowed to choose freely their hands to acquire a food item, they often display hand preferences. However, such hand preference is not complete, and the animals tend to switch their hands in successive trials. This indicates that utilities for reaching and grasping movements incorporate the relationship between successive movements. Similar to the cortical circuitry involved in the selection of a behavioral goal, the neural system responsible for motor decision making is also distributed. For example, neurons encoding information about the movement target and the hand chosen to acquire it are found in multiple cortical areas, including the dorsal premotor cortex, the dorsolateral prefrontal cortex, and the supplementary motor area.

Even the trajectories of simple movements, such as reaching or saccadic eye movement, have a large degree of freedom. Nevertheless, the trajectories of such movements tend to be highly stereotyped and typically bell shaped. This led to the assumption that specific movement trajectories are chosen as a result of optimization. For example, it has been proposed that the motor system chooses a particular trajectory to minimize the rate of change in acceleration (jerk) or torque. The function to be minimized during optimization, such as the total jerk, is referred to as the cost or loss function, and therefore can be considered as negative utility function. In general, the cost function computed from the error at the end of the movement does not make specific predictions about the movement trajectory, unless the effect of neural noise is considered. Trains of action potentials recorded from the neurons throughout the motor system, including the motor neurons in the spinal cord, display highly irregular patterns, and the amount of such neuronal noise tends to increase with the average activity. This is referred to as signal-dependent noise. The variability at the end point of a given movement, such as reaching, would increase if the hand is suddenly accelerated, requiring rapid muscle contraction and higher activity of motor neurons. In fact, the optimal trajectory that minimizes the variability in the end point in the presence of signal-dependent noise resembles

those with minimum jerk. Motor errors are encoded by climbing fibers in the cerebellum, and such signals can be used to reduce the positional error by modifying the corresponding movement trajectory.

See also: Cognition: Basal Ganglia Role; Decision-Making and Vision; Dopamine Neurons: Reward and Uncertainty; Game Theory and the Economics of Animal Communication; Learning, Action, Inference and Neuromodulation; Posterior Parietal Cortex and Arm Movement; Reward Decision-Making; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex; Social Brain: Evolution.

Further Reading

Barraclough DJ, Conroy ML, and Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience* 7: 404–410.

Camerer CF (2003) *Behavioral Game Theory.* Princeton: Princeton University Press.

Camerer CF and Ho T-H (1999) Experience-weighted attraction learning in normal form games. *Econometrica* 67: 827–874.

Dorris MC and Glimcher PW (2004) Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* 44: 365–378.

Harris CM and Wolpert DM (1998) Signal-dependent noise determines motor planning. *Nature* 394: 780–784.

Körding KP and Wolpert DM (2004) The loss function of sensorimotor learning. *Proceedings of the National Academy of Sciences of the United States of America* 101: 9839–9842.

Lee D (2006) Neural basis of quasi-rational decision making. *Current Opinion in Neurobiology* 16: 191–198.

Lee D, McGreevy BP, and Barraclough DJ (2005) Learning and decision making in monkeys during a rock-paper-scissors game. *Cognitive Brain Research* 25: 416–430.

Lee D and Schieber MH (2006) Serial correlation in lateralized choices of hand and target. *Experimental Brain Research* 174: 499–509.

Montague PR, King-Casas B, and Cohen JD (2006) Imaging valuation models in human choice. *Annual Review of Neuroscience* 29: 417–448.

Platt ML and Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400: 233–238.

Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology* 57: 87–115.

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

von Neumann J and Morgenstern O (1944) *Theory of Games and Economic Behavior.* Princeton: Princeton University Press.

Wolpert DM and Ghahramani Z (2000) Computational principles of movement neuroscience. *Nature Neuroscience Supplement* 3: 1212–1217.

Relevant Website

http://www.bme.jhu.edu – Reza Shadmehr Home Page, The Whitaker Biomedical Engineering Institute at Johns Hopkins.

Gamma-Aminobutryic Acid (GABA)

A Schousboe and H S Waagepetersen, University of Copenhagen, Copenhagen, Denmark

© 2009 Elsevier Ltd. All rights reserved.

GABA – The Inhibitory Neurotransmitter

γ-Aminobutyric acid (GABA) is the most abundant inhibitory neurotransmitter in the central nervous system (CNS). It was discovered in the CNS around 1950 by Eugene Roberts, but it was not until the 1960s that GABA was suggested to be an inhibitory neurotransmitter. GABA is also present in organs other than the brain. For example, it also is present in pancreatic β cells, where it appears to act as a signaling molecule between cells within the pancreatic islets. Moreover, the entities of the GABAergic neurotransmission (biosynthetic/metabolic enzymes, receptors, and transporters) are expressed in the reproductive organs (testes and ovaries), the adrenal gland, and the gastrointestinal tract. With regard to the reproductive system, GABA has an effect on release of estradiol and progesterone in the ovary and it also appears that GABA may modulate sperm function, perhaps *via* a stimulatory effect on testosterone production. In the adrenal gland, steroidogenesis is influenced by GABA in a paracrine or autocrine manner. This could modulate release of catecholamines from the adrenal glands. In the gastrointestinal tract, there is ample evidence to suggest a neurotransmitter or neuromodulatory role for GABA. Hence, GABAergic activity seems to modulate contractile responses *via* an action on acetylcholine release.

GABA Metabolism

GABA is synthesized from glutamate, catalyzed by the enzyme L-glutamate decarboxylase (EC 4.1.1.15; GAD). It has been suggested that during early postnatal development, GABA might be synthesized through an alternative pathway involving ornithine. Classically, GAD has been considered to be a marker for GABAergic neurons in the CNS. However, recent studies have indicated that GAD might be present in at least a subpopulation of excitatory neurons as well. This localization in excitatory neurons might reflect the possible role of GABA as a neurodifferentiative agent during early postnatal development in the CNS.

GABA is metabolized by GABA transaminase (EC 2.6.1.19; GABA-T) in the GABA shunt (Figure 1). In contrast to GAD, GABA-T, which is localized in the mitochondrial matrix, is an enzyme that is ubiquitously expressed in not only the brain but also other organs. GABA-T converts GABA to succinic semialdehyde, using α-ketoglutarate as the amino group acceptor forming glutamate. The enzyme requires pyridoxal phosphate (PLP) as a coenzyme, and hence, carboxyl trapping agents inhibit GABA-T. Succinic semialdehyde is subsequently oxidized to succinate, which enters the tricarboxylic acid (TCA) cycle (Figure 1). Large differences exist in the reported fluxes through the GABA shunt compared with TCA cycle activity. They range from 2% to 17%, with the highest value obtained from analysis of mouse brain metabolism.

GAD_{65} and GAD_{67}

GAD exists in two isoforms, known as GAD_{65} and GAD_{67}, according to their molecular masses (i.e., approximately 65 and 67 kDa), and they are encoded by two independent genes. GAD_{67} seems to be a cytosolic enzyme distributed throughout the GABAergic neurons in the cell body as well as in their processes. In contrast, GAD_{65} is predominantly found to be located in nerve terminals and it might be anchored in the membrane of the vesicles containing the neurotransmitter. The regulation of GAD is complex and has not been fully elucidated. PLP is the cofactor for GAD, and the association and dissociation of this cofactor play a major role in the short-term regulation of GAD. Fifty percent of GAD is present in the brain as the apoenzyme, thereby providing a large 'dormant' catalytic capacity, which may be readily recruited whenever necessary. GAD_{65} is the predominant contributor to this large pool of apoenzyme, whereas GAD_{67} primarily exists as a holoenzyme. The interconversion between apo- and holoGAD is regulated by physiological concentrations of adenosine triphosphate (ATP) and inorganic phosphate. It has been shown that soluble GAD (presumably GAD_{67}) is inhibited by protein kinase A-mediated protein phosphorylation, whereas membrane-associated GAD (presumably GAD_{65}) recently has been shown to be activated by ATP-mediated phosphorylation. These findings illustrate the pronounced complexity in the regulation of GABA synthesis.

The rationale for the existence of two enzymes, GAD_{65} and GAD_{67}, that are expressed by independently regulated genes and that are distinct with regard to regulation and localization is at present not understood. It has been speculated how this might be connected to the fact that GABA exists in at least two separate compartments, named the

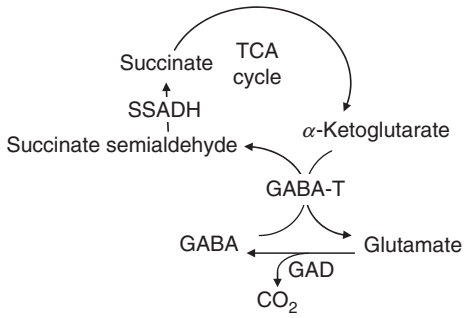

Figure 1 The γ-aminobutyric acid (GABA) shunt reactions are responsible for the synthesis and degradation of GABA. TCA, tricarboxylic acid; SSADH, succinic semialdehyde dehydrogenase; GABA-T, GABA transaminase; GAD, glutamate decarboxylase.

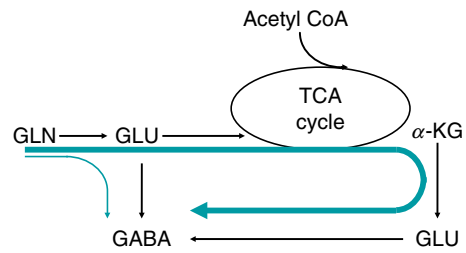

Figure 2 Synthesis of γ-aminobutyric acid (GABA) from glutamine via the TCA. An illustration of two pathways for GABA synthesis from glutamine. The bold green arrow illustrates the predominant pathway, where the carbon skeleton of glutamine (GLN) is metabolized via the tricarboxcylic acid (TCA) cycle prior to synthesis of GABA. The thin green arrow illustrates the direct synthesis of GABA from glutamine without involvement of the TCA cycle. GLU, glutamate; α-KG, α-ketoglutarate.

cytoplasmic and the vesicular pools, which might have distinct functions. The neurotransmitter function from the vesicular pool seems obvious. However, the function of the metabolic pool is more uncertain. However, in addition to a possible modulatory role during neurotransmission, it may be of importance for the action of GABA as a neurodifferentiative agent during development. As further evidence in support for multiple pools of GABA and their functional significance, antiepileptic drugs acting through an interference with GABA metabolism have been found both *in vitro* and *in vivo* to affect distinct pools of GABA. Because GAD is dependent on PLP, carbonyl trapping agents inhibit GAD and act as anticonvulsants.

Glutamine as GABA Precursor

In GABAergic neurotransmission, a net flow of GABA from the neuronal to the astrocytic compartment occurs. This net flow needs to be compensated by a flow of a GABA precursor in the opposite direction due to the lack of a neuronal anaplerotic pathway. It has been convincingly demonstrated in different brain tissue preparations that glutamine, which is exclusively synthesized in astrocytes, can function as the precursor for the synthesis of GABA. Glutamine is to a large extent metabolized through the TCA cycle before conversion to glutamate and subsequently to GABA (Figure 2). This mitochondrial involvement in GABA biosynthesis might be an important regulatory point. The extent of TCA cycle metabolism before GABA synthesis from glutamine appears to be dependent on the pool – that is, the vesicular or metabolic pool from which GABA is released.

GABA Release

GABAergic neurons contain specialized vesicles equipped with a GABA transporter that is capable of concentrating GABA approximately 10- to 20- fold over that present in the cytoplasm. After depolarization of the neuronal plasma membrane, these vesicles may fuse with the membrane in a process involving a number of specific proteins (synaptin, neurexins, synaptotagmin, syntaxin, and synaptophysin) that are activated and reorganized on an elevation of the intracellular Ca^{2+} concentration brought about by opening a number of pharmacologically distinct voltage-gated Ca^{2+} channels. Ultimately, this leads to the release of the vesicular GABA, which spreads in the narrow synaptic cleft, allowing activation of a set of postsynaptic and presynaptic GABA receptors.

In addition to the vesicular or quantal release mechanism that leads to a spatially highly localized release, GABA may also be released by reversal of the plasma membrane GABA transporters, which are electrogenic and therefore sensitive to depolarization. When this mechanism is operating, GABA release will be spatially different from the vesicular release, which may have important functional consequences. Another important difference is that the transporter-mediated release of cytoplasmic GABA is not Ca^{2+} dependent.

GABA Receptors

Following its release from GABAergic nerve endings, GABA is bound to a variety of receptors (Table 1). These are separated in two major classes – that is, the ionotropic ($GABA_A$ and $GABA_C$) and the metabotropic ($GABA_B$) – in accordance with either formation of an ion channel (A and C) or association with G-proteins, which link receptor activation to either K^+ or Ca^{2+} channels or to adenylate cyclase.

Ionotropic Receptors

The $GABA_A$ and $GABA_C$ receptors form chloride channels, which, upon activation by GABA, open

Table 1 GABA receptor classification

A^a	C^a	B^a
Subunits		
α_1 β_1 γ_1 δ ε θ	ρ_1	B1a
π		
α_2 β_2 γ_2	ρ_2	B1b
α_3 β_3 γ_3	ρ_3	B2
α_4		
α_5		
α_6		

[a]$GABA_A$ and $GABA_C$ receptors form Cl^- channels in pentameric homo (C) or heteromeric subunit complexes. The Greek letters denote subunit families, each of which shows about 70% homology in the amino acid sequence. The corresponding homology across the families is about 30–40%. $GABA_B$ receptors are G-protein-coupled 7-TM receptors and consist of heterodimeric complexes of B1a/B2 or B1b/B2.

and allow Cl^- to be fluxed into the cell, resulting in a hyperpolarization. In this context, it may be mentioned that owing to differences in the Cl^- distribution intracellularly versus extracellularly, in neurons during early developmental stages compared with mature neurons, opposite Cl^- fluxes occur in immature neurons. Therefore, depolarizing effects of GABA are often observed in the embryonic nervous system, a property of GABAergic neurotransmission that very likely is important for the neurotropic action of GABA during early development.

The distinction between $GABA_A$ and $GABA_C$ receptors rests essentially only on pharmacological grounds (Table 2). $GABA_C$ receptors are characterized by an activation by *cis*-4-aminobut-2-enoate (CACA), whereas classic $GABA_A$ receptor agonists such as isoguvacine have no effect. Because the $GABA_C$ receptor is a homomeric complex of ρ-subunits sharing a considerable amino acid sequence homology with the remaining $GABA_A$ receptor subunits, it may be concluded that this receptor subclass should be termed a pharmacologically distinct $GABA_A$ receptor.

In addition to the GABA-binding site normally thought to be located at the interphase between the α and β subunits, which are found in essentially all $GABA_A$ receptors, the receptor complex is endowed with a multitude of pharmacologically highly important recognition sites for drugs such as benzodiazepines (γ subunit), barbiturates, ethanol, steroids, and picrotoxinin. This is the basis for $GABA_A$ receptors being a highly important target for drugs acting as tranquilizers, anesthetics, anticonvulsants, and convulsants.

The ionotropic GABA receptors are regulated in terms of activity in numerous ways. Essentially all subunits contain phosphorylation sites for protein kinases A, C, and G, as well as calmodulin kinase and tyrosine kinase, leading to regulation of activity of the ion channel. Another important regulation is associated with desensitization, the characteristics of which exhibit subunit specificity. Also, channel opening kinetics are characterized by subunit diversity. Recently, a GABA receptor anchoring protein has been identified and cloned. This protein plays an important role in regulating trafficking, targeting, and assembling $GABA_A$ receptors in the plasma membrane. Attention has also been focused on the subset of $GABA_A$ receptors containing $\alpha4$, $\alpha6$, and δ subunits and which are mainly located extrasynaptically. It appears that this subclass of receptors has interesting pharmacological properties. The classical $GABA_A$ specific agonist 4,5,6,7-tetrahydroisoxazole[5,4-*c*] pyridin-3-ol (THIP) has been shown to act particularly effectively on these receptors, and the mechanism of action of some anticonvulsants acting via inhibition of GABA transporters (see later) also appears to involve extrasynaptic GABA receptors.

Metabotropic Receptors

The metabotropic GABA receptors, originally termed $GABA_B$ because they are pharmacologically distinct from $GABA_A$ (being activated by baclofen and insensitive to the $GABA_A$ receptor antagonist bicuculline), belong to the G-protein-coupled seven-transmembrane (7TM) receptor family (Table 1). Recent cloning strategies have resulted in the identification of two subtypes, $GABA_{B1}$, and $GABA_{B2}$, which have to form a heterodimeric complex in order to be functionally active. As mentioned earlier, baclofen is the prototype agonist, whereas 2-OH-saclofen is the most commonly used antagonist (Table 2). $GABA_B$ receptors expressed at postsynaptic sites normally lead to hyperpolarization if activated due to coupling to K^+ channels that increase the membrane K^+ conductance. Presynaptically, $GABA_B$ receptors are coupled to Ca^{2+} channels, and activation leads to a decrease in Ca^{2+} conductance, resulting in inhibition of transmitter release. These basic properties of $GABA_B$ receptors lie behind their therapeutic relevance because agonists produce muscle relaxation and antagonists act as antiepileptics, particularly in absence seizures.

GABA Inactivation

GABA is inactivated by uptake either into the presynaptic GABAergic nerve ending, allowing recycling of the neurotransmitter, or into surrounding astrocytes, leading to degradation via transamination (GABA-T) and oxidation to succinate. So far, three GABA transporters and a betaine transporter also

Table 2 Basic pharmacology of GABA receptors

$GABA_A$		*$GABA_C$*		*$GABA_B$*	
Agonist	*Antagonist*	*Agonist*	*Antagonist*[a]	*Agonist*	*Antagonist*
GABA	Bicuculline	GABA	3-APMPA	GABA	Phaclofen
THIP	Picrotoxinin	CACA	3-APPA	Baclofen	5-OH-Saclofen
Isoguvacine		CAMP			
Isonipecotic acid					
Muscimol					

[a]Nonselective ($GABA_A$–$GABA_C$).
THIP, 4,5,6,7-tetrahydroisoxazole[5,4-*c*]pyridin-3-ol; CACA, *cis*-4-aminocrotonic acid (*cis*-4-amino-2-butenoic acid); CAMP, *cis*-2-(aminomethyl)cyclopropane-1-carboxylic acid; 3-APMPA, 3-aminopropyl(methyl)phosphinic acid; 3-APPA, 3-aminopropylphosphonic acid.

capable of GABA transport have been cloned. In the mouse, these transporters are termed GAT1–GAT4, where GAT2 is synonymous with the betaine transporter. Although GAT1 is preferentially located on neurons but also to some extent on astroglia, the other transporters and particularly GAT4 are mainly nonneuronal. These differences in expression among neurons and astroglial cells are of interest in relation to pharmacological differences among the cloned transporters and in regard to the possibility to use this as a basis to explain some of the pharmacological differences noted between neuronal and astroglial GABA uptake. Close inspection of the pharmacological properties of these GABA transport entities reveals, however, that it is not possible to explain the distinct differences between neuronal and astroglial GABA uptake on the basis of known pharmacological properties of the cloned transporters (Table 3). This has led to the proposal that GABA transporters may exhibit cooperativity when expressed heterologously.

GABA transporters may be regulated by phosphorylation – for example, through protein kinases A and C – and, moreover, membrane trafficking and anchoring appear to be activity dependent. Furthermore, neuronally produced proteins have been shown to be of importance for the regulation of expression of GABA transporters in astrocytes. Hence, the synaptic machinery for GABA recognition and inactivation appears to be highly plastic and regulated in an activity-dependent manner. The development of tiagabine, which specifically inhibits the GAT1 transporter (Table 3), as an antiepileptic drug has provided proof of principle that GABA transporters constitute important drug targets. This notion has been extended to include other than GAT1 transporters by the demonstration that GAT2 (the betaine transporter) inhibitors may have novel anticonvulsant properties involving a mechanism implicating activation of extrasynaptic $GABA_A$ receptors (see earlier).

Functional Implications of Inhibition

Immunohistochemical visualization of GABAergic neurotransmission through the GABA-synthesizing enzyme GAD, the neurotransmitter (GABA) itself, and its receptors has clearly and indisputably shown GABAergic transmission to be highly ubiquitous in

Table 3 Key substrates and inhibitors of GABA transporters

Substrate/inhibitor	*Transporter*					
	Neuron	*Glia*	*GAT1*	*GAT2*	*GAT3*	*GAT4*
K_m, K_i, or IC_{50} values (μM)						
GABA	10	30	20	50	15	20
Nipecotic acid	10	15	25	>1000	100	150
Guvacine	30	30	40	>1000	200	400
ACHC	200	700	150	1000	>1000	>10 000
THPO[a]	500	250	1000	3000	800	5000
Tiagabine[a]	0.5	0.2	0.1	>100	>100	100
EF 1502	2	2	7	26	>300	>300

[a]These inhibitors do not act as substrates.
ACHC, *cis*-3-aminocyclohexane carboxylic acid; THPO, 4,5,6,7-tetrahydroisoxazolo[4,5-*c*]pyridin-3-ol; GAT1–GAT4, these represent the GABA transporters cloned from mouse brain; EF 1502, *N*-[4,4-bis(3-methyl-2-thienyl)-3-butenyl]-4-(methylamino)-4,5,6,7-tetrahydrobenzo[d]isoxazol-3-ol.

the CNS. This confirms the original notion by Eugene Roberts that inhibition is the governing principle of function in the brain. This means that disinhibition leading to excitation may be as important as, if not more important than, excitation itself. Therefore, it is not surprising that aberrations in GABAergic function at the level of enzyme defects, receptor dysfunction, or transporter failure are involved in a vast number of neurologic and psychiatric disorders, including epilepsy, anxiety, insomnia, Alzheimer's disease, and schizophrenia, just mentioning a few examples. Therefore, the different entities of the GABA system have been among the major targets for development of drugs that are therapeutically active against these diseases. Among these drugs, the benzodiazepines may constitute one of the most important drug discoveries of modern pharmacology.

See also: GABA Synthesis and Metabolism; GABA-A Receptors: Developmental Roles; GABA-A Receptor Synaptic Functions; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-B Receptors: Molecular Biology and Pharmacology; GABA-B Receptor Function; Muscarinic Receptors: Autonomic Neurons.

Further Reading

Bowery NG (2002) Pharmacology of $GABA_B$ receptors. In: Egebjerg J, Schousboe A, and Krogsgaard-Larsen P (eds.) *Glutamate and GABA Receptors and Transporters; Structure, Function and Pharmacology*, pp. 287–304. Taylor and Francis: London.

Clausen RP, Madsen K, Larsson OM, et al. (2006) Structure–activity relationship and pharmacology of γ-aminobutyric acid (GABA) transport inhibitors. *Advances in Pharmacology* 54: 265–284.

Froestl W and Mickel SJ (1997) Chemistry of $GABA_B$ modulators. In: Enna SJ and Bowery NG (eds.) *The GABA Receptors*, pp. 271–296. New Jersey: Humana Press Inc.

Gladkevich A, Korf J, Hakobyan VP, et al. (2006) The peripheral GABAergic system as a target in endocrine disorders. *Autonomic Neuroscience* 124: 1–8.

Jessen KR, Mirsky R, and Hills JM (1987) GABA as an autonomic neurotransmitter in the myenteric plexus of the gut. *Trends in Neurosciences* 10: 255–261.

Johnston GAR (1997) Molecular biology, pharmacology and physiology of $GABA_C$ receptors. In: Enna SJ and Bowery NG (eds.) *The GABA Receptors*, pp. 297–323. New Jersey: Humana Press Inc.

Klix N and Bettler B (2002) Molecular structure of the $GABA_B$ receptors. In: Egebjerg J, Schousboe A, and Krogsgaard-Larsen P (eds.) *Glutamate and GABA Receptors and Transporters; Structure, Function and Pharmacology*, pp. 277–286. London: Taylor and Francis.

Krogsgaard-Larsen P, Frølund B, Kristiansen U, et al. (2002) Ligands for the $GABA_A$ receptor complex. In: Egebjerg J, Schousboe A, and Krogsgaard-Larsen P (eds.) *Glutamate and GABA Receptors and Transporters; Structure, Function and Pharmacology*, pp. 236–274. London: Taylor and Francis.

Krogsgaard-Larsen P, Frølund B, Liljefors T, et al. (2004) $GABA_A$ agonists and partial agonists: THIP (Gaboxadol) as a non-opioid analgesic and a novel type of hypnotic. *Biochemical Pharmacology* 68: 1573–1580.

Martin DL and Rimvall K (1993) Regulation of gamma-aminobutyric acid synthesis in the brain. *Journal of Neurochemistry* 60: 395–407.

Olsen RW and Macdonald RL (2002) $GABA_A$ receptor complex structure and function. In: Egebjerg J, Schousboe A, and Krogsgaard-Larsen P (eds.) *Glutamate and GABA Receptors and Transporters; Structure, Function and Pharmacology*, pp. 202–235. London: Taylor and Francis.

Otsuka M (1996) Establishment of GABA as an inhibitory neurotransmitter at crustacean neuromuscular junction and in the mammalian central nervous system. In: Egebjerg J, Schousboe A, and Krogsgaard-Larsen P (eds.) *Glutamate and GABA Receptors and Transporters; Structure, Function and Pharmacology*, pp. 1–6. London: Taylor and Francis.

Roberts E (1974) Disinhibition as an organizing principle in the nervous system. The role of gamma-aminobutyric acid. *Advances in Neurology* 5: 127–143.

Waagepetersen HS, Sonnewald U, and Schousboe A (1999) The GABA paradox: Multiple roles as metabolite, neurotransmitter and neurodifferentiative agent. *Journal of Neurochemistry* 73: 1335–1342.

White HS, Watson WP, Hansen S, et al. (2005) First demonstration of a functional role for CNS betaine/GABA transporter (mGAT2) based on synergistic anticonvulsant action among inhibitors of mGAT1 and mGAT2. *Journal of Pharmacology and Experimental Therapeutics* 312: 866–874.

Gap Junction Abnormalities and Disorders of the Nervous System

C C Naus, The University of British Columbia, Vancouver, BC, Canada
C Giaume, Collège de France, Paris, France
P L Carlen, University Health Network, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Gap Junctions, Connexins, and Intercellular Communication in the Nervous System

Gap junctions are intercellular channels which directly connect the cytoplasm between neighboring cells. As such, they form an extensive pathway of intercellular communication in the nervous system. The structural unit of the gap junction is the 'connexon,' a proteinaceous cylinder with a hydrophilic channel, which spans the plasma membranes of closely apposed cells to align end-to-end, forming intercellular channels which provide for the exchange of small molecules (less than 1200 Da). This channel underlies the phenomenon of gap junctional intercellular communication (GJC). The gap junction proteins, called connexins (Cx), are encoded by a multigene family consisting of at least 20 members in rodents (21 in humans), many of which are found in the nervous system, where they are involved in coupling various kinds of cells. In addition, other gap junctional candidates which should also be considered are the pannexins and the mammalian counterpart of the invertebrate gap junction proteins called innexins. The subset and the expression of connexins vary depending on the type of cell and the stage of development. Neuronal gap junctions are involved in electrical coupling, and may play a critical role in neural development. Astrocytes are generally highly coupled by gap junctions, leading to an extensive cellular network which is critical to maintaining ionic and metabolic homeostasis. The myelinating cells of the nervous system, that is, the oligodendrocytes (in the central nervous system (CNS)) and Schwann cells (in the peripheral nervous system (PNS)), also express a number of connexins. In Schwann cells, these are present between the layers of myelin in order to allow for efficient intracellular metabolic coupling. Of increasing importance is the unraveling of the function of gap junctions and connexins in the neural cell network, involving neurons, astrocytes, oligodendrocytes, and microglia.

The role of connexins, gap junctions, and hemichannels in the pathology of diseases in the nervous system is coming to light through a combination of cellular and animal models and the ongoing discovery of human genetic disorders associated with connexin mutations. Since connexins associate in both homomeric or heteromeric forms, and can form homotypic or heterotypic cellular associations, mutations in individual connexin genes can impact the entire intercellular communication network within a tissue, organ, or organism. In addition, modulation of connexin gene expression by growth factors or cytokines, and alterations in phosphorylation can profoundly affect gap junction channel gating, and thus function. In this review, we address the possible roles of gap junctions and connexins in neural development, as well as in neurological disorders such as stroke and epilepsy in light of insights gained from genetically engineered mice and characterized human genetic mutations.

Connexins and Neural Development – Building the Gap

During development, there is extensive regulation of the spatial and temporal expression of gap junctions. This regulation is being implicated in the control of differentiation and growth. In general, there is high intercellular coupling during development, with a loss or attenuation of coupling during terminal differentiation. GJC and cellular expression of connexins have been partially characterized in the developing nervous system. The functional significance of connexin expression in astrocytes, oligodendrocytes, and neurons is only beginning to be resolved. The changes in the expression of gap junction genes during postnatal brain development coincides with various major differentiative events, including cellular proliferation and migration, neurite outgrowth and synaptogenesis, myelination, neurotransmitter differentiation, and neuronal death. For the most part, it remains to be determined if gap junctions play any direct role in such developmental events.

Electrotonic coupling between mammalian neurons has been shown in many areas of the CNS, and has been implicated in neuronal synchronization. For example, during neocortical development, there is a high degree of intercellular coupling between neurons. Studies in the rat indicate that neuroblasts are coupled with approximately 30–60 others in columns within the ventricular zone of the developing cerebral cortex. The zone is comprised of mitotically active epithelial cells lining the ventricles. These cells have been shown to express Cx43 and Cx26, and the

coupling appears to involve both neural precursors and radial glia. This extensive coupling is maintained throughout all neocortical layers during the first 2 weeks after birth, decreasing significantly thereafter. The specific connexins responsible for this transient neuronal coupling have not been definitively determined. Although Cx43 has been reported in neurons of the cortical plate, with some Cx32 appearing as neurons mature, the pattern of expression of Cx26 and Cx36 in the developing brain is consistent with the transient GJC observed in the neocortex.

It has been suggested that this early electrotonic coupling via gap junctions may function in determining the phenotype and positional fate of neurons in the cortex, providing a mechanism for laminar organization. Thus, GJC may establish cortical domains in the developing neocortex that underlie the adult pattern of functional cytoarchitecture. In addition, the establishment of an early system of electrotonic synapses may be the precursor of the more differentiated chemical synapses which follow. When embryonal carcinoma cells differentiate into neurons in response to retinoic acid, the expression of Cx43, and the level of GJC, progressively disappear, similar to what is seen *in vivo*. Interestingly, blocking of GJC disrupts retinoic acid-induced neuronal differentiation in this same system. It therefore appears that the temporal pattern of connexin expression and GJC during neuronal differentiation *in vitro* is critical.

Gap junctions have also been implicated in the migration of neurons during development. A null mutation of Cx43 perturbs the migration of neurons during neocortical development. This suggests that GJC or connexin expression defects may underlie some of the migrational disorders in nervous system development, and thus may also contribute to epilepsy (see below).

Connexin Mutations – Breaking the Gap

Following the identification and cloning of the connexin genes, functional studies were rapidly pursued utilizing knockout and transgenic technologies. Coincidental with the characterization of these genetically engineered mice, mutations of some of the human connexin genes were identified. This allowed for a phenotypic comparison between mice and men. These are summarized in Tables 1 and 2.

Connexin Knockouts and Transgenics

While targeted gene knockout through homologous recombination to eliminate expression of connexin genes in transgenic mice has been pursued since 1995, this strategy has provided little insight into the role of GJC in neural development. The homozygous Cx43 null mutation in mice is lethal at birth. Thus, any neural phenotype of these knockout mice is confounded by cardiac malformation and neonatal death. Knockout of another neuronal candidate, Cx26, is embryo lethal, and thus additional strategies must be followed to investigate its function in neuronal development. Most recently, Cx36 knockout mice have been shown to be viable, and although visual deficits are apparent, no major CNS phenotype has been reported. These results highlight several problems inherent in this genetic approach.

1. In the brain, except for Cx36 which is neuronal, most connexins are expressed in several cell types.
2. All CNS cell types express multiple connexins, so possible compensation for loss of function of one connexin by another may occur.
3. Many connexins are expressed during development, so embryonic lethal phenotypes are not uncommon in knockout studies.
4. If the knockout is not lethal, we need to interpret the phenotype in light of possible disruption of developmental processes which rely on specific connexins. To circumvent many of these problems, more recent approaches have employed conditional knockout strategies.

Table 1 Some observed nervous system abnormalities in connexin knockout mice

Knockout mouse	*Abnormalities*
Cx26	Embryonal lethal
($Cx26^{fl/fl}$, Otog-cre)	Hearing impairment
Cx29	Hearing impairment
	Demyelination
Cx30	Severe hearing impairment
	Reduced exploratory activity
Cx32	Demyelination in PNS
	Hyperexcitability in CNS
	Enhanced neuronal injury in ischemia
Cx36	Visual deficit
	Synchronization defect in hippocampus
	Disturbed synchronous inhibitory activity
	Disrupted gamma frequency oscillations in cortex
Cx43	Neonatal lethal
	Subtle neural changes in embryo
	Neural crest migration defect
	Disturbance of neural migration in embryonic neocortex
(Cx43+/−)	Enhanced neuronal injury in ischemia
($Cx43^{fl/fl}$,GFAP-cre)	Increased spreading depression in hippocampus
	Enhanced neuronal injury in ischemia
Cx47	Vacuolation in nerve fibers
(Cx47/32 double)	Severe demyelination in CNS

Table 2 Human connexin mutations

Gene/protein	*Abnormalities*
GJB2/Cx26	Dominant deafness; skin disorders
GJB6/Cx30	Dominant deafness; skin disorders
GJB4/Cx30.3	Dominant deafness; skin disorders
GJB3/Cx31	Dominant deafness; skin disorder
GJB1/Cx32	CMTX demyelinating disorder; deafness
GJA9/Cx36	Juvenile myoclonic epilepsy
GJA1/Cx43	Oculodentodigital dysplasia; deafness
GJA3/Cx46	Cataracts
GJA8/Cx50	Cataracts

Connexin Mutations and Genetic Disorders

With the identification of connexin genes, several genetic disorders have been linked to specific mutations. While some of these result in a nervous system phenotype, others manifest in diverse systems. Several of these are briefly summarized here.

Cx32 mutation and CMTX The first human genetic disorder attributed to a connexin mutation was X-linked Charcot-Marie-Tooth disease (CMTX), a dominant demyelinating peripheral neuropathy. Cx32 is expressed in the Schwann cells that myelinate peripheral nerves. To date, over 270 mutations have been identified in Cx32, with most of these being point mutations that alter connexin protein trafficking and gating properties. Interestingly, when Cx32 knockout mice were generated, they also exhibited a demyelinating peripheral nerve phenotype, and in addition display neuronal hyperexcitability in the CNS. Furthermore, mice with a double knockout of Cx32 and Cx47 die by postnatal week 6 from profound abnormalities in CNS myelin indicating that these two connexins are critical for normal central myelination.

Cx43 mutation and ODDD Oculodentodigital dysplasia (ODDD) is a congenital autosomal dominant disorder with face, eye, limb, and tooth abnormalities associated with mutation of Cx43. There are approximately 30 Cx43 mutations that have been identified in ODDD. On the one hand, the broad spectrum of systems involved is not surprising given the extensive expression of Cx43 in the body. On the other hand, this being the case, it is surprising that the phenotypes are not more severe.

Connexins mutated in deafness Four different connexin genes have been found with mutations associated with deafness (i.e., nonsyndromic sensorineural hearing loss) – Cx26, Cx30, Cx30.3, and Cx31. For Cx26 alone, over 100 mutations have been identified which lead to deafness, with some of these also linked to skin disorders (see below). These connexins are expressed in the cochlea and appear to be critical for potassium buffering and recycling. In mice, targeted ablation or dominant negative mutation of Cx26 in the Organ of Corti causes degeneration of the sensory hairs cells resulting in deafness.

Connexins mutated in skin disorders Interestingly, the same connexin genes mutated in deafness are also associated with hyperproliferative skin disorders (i.e., erythrokeratodermia). At least nine different connexin genes are co-expressed in the epidermis and there is evidence indicating that they play a critical role in regulating keratinocyte growth and differentiation. The loss of connexins and GJC has long been correlated with disruption in cell growth control and cancer (see below).

Connexins mutated in cataracts In the lens three connexins are expressed: Cx43, Cx46, and Cx50. Point mutations in these genes have been identified in patients with inherited zonular pulverulent cataracts. In the mouse, Cx46 knockout is characterized by severe cataracts without effects on lens development and growth. In contrast, Cx43/Cx50 double knockout exhibited normal lens demonstrating that these two connexins are not required for lens development.

Connexins and Neuroprotection – Minding the Gap

Given the extent of GJC among astrocytes in the CNS and the proposed role of these glial cells in spatial buffering of potassium and glutamate, a protective role of gap junctions has been proposed. In contrast, this extensive network of intercellular communication has been implicated in the propagation of cellular injury at least between astrocytes. These contrasting opinions are likely based, in part, on experimental brain ischemia models (see Table 3). The use of these models is based on the clinical impact of such injuries. Cerebrovascular diseases rank as the third leading cause of death in the United States and the fourth in Canada, most commonly manifesting as ischemic brain stroke.

Several lines of evidence support a neuroprotective role of gap junctions. (1) Blocking astrocytic gap junctions enhances neuronal vulnerability either to glutamate cytotoxicity or metabolic stress in neuron/glial co-cultures and under experimental ischemia using oxygen and glucose depletion in hippocampal slice cultures. (2) *In vivo*, heterozygote Cx43 knockout mice showed a significantly increased stroke volume compared to wild-type mice following middle cerebral artery occlusion (MCAO), accompanied by

Table 3 Connexins, gap junctions, and neuroprotection

Injury	*Cx*	*Main observation*
Focal ischemia		Octanol reduces infarct volume
Metabolic stress in glial cultures	Cx43	GJC amplifies cell injury
Focal brain ischemia		Halothane reduces injury
Global cerebral ischemia		Octanol decreases neuronal death
Hypoxia-hypoglycemia to organotypic slices	Cx32, Cx26, Cx43	Reduction of GJC decreases neuronal vulnerability
Traumatic impact injury to organotypic slices	Cx32, Cx26, Cx43	Reduction (enhancement) of GJC decreases (increases) neuronal vulnerability
Oxygen glucose deprivation		Increased GJC and apoptosis which was reduced by blockers
Excitotoxic injury to organotypic and dissociated cultures		Blocking GJC exacerbates injury
Focal brain ischemia	Cx43	Cx43 heterozygotes present larger infarct volume
Excitotoxicity to dissociated cultures		Blockade of GJC increased cell death
Global brain ischemia	Cx32	Cx32 knockouts show enhanced neuronal vulnerability
Focal brain ischemia	GFAP-Cre/Cx43$^{fl/fl}$	Cx43 deletion resulted in increased infarct volume
	GFAP-Cre/Cx43$^{fl/fl}$	Increased hippocampal spreading depression

enhanced apoptosis. (3) More recent use of targeted deletion of connexins, specifically in astrocytes, provides further evidence that astrocytic gap junctions play a neuroprotective role in ischemic insults. (4) Cx32 knockout mice have been reported to exhibit an enhanced vulnerability of hippocampal neurons against global brain ischemia compared to wild-type mice, suggesting that gap junctions of hippocampal interneurons play a neuroprotective role in ischemic insults.

On the other hand, there is evidence that gap junctions enhance neuronal injury. (1) Neuronal death caused by oxygen and glucose depletion was decreased when Cx43 was blocked by specific antisense oligodeoxynucleotides in hippocampal slice cultures. (2) Similarly, the stroke volume following MCAO in the rat model was reduced by blocking gap junctions using octanol. (3) It has been suggested that the spreading depression caused by ischemic insult propagates via astrocytic gap junctions which remain open during the ischemic condition, resulting in the expansion of the stroke volume.

The role of astrocytic connexins could also be mediated by hemichannel opening. Recently, it has been reported that astrocytic hemichannels which are closed under normal conditions remained open under experimental ischemia induced by glucose and oxygen depletion. It is possible that open hemichannels allow for the release of glutamate causing loss of membrane potential, or hemichannels may contribute to the distribution of antiapoptotic factors in the lesion area.

A number of factors may contribute to the above apparent contradictions. (1) Studies vary considerably with regard to the system employed, including the ischemia model and the gap junction blockers used. (2) In ischemic injury, there are critical stages with specific time parameters. Thus, GJC may be beneficial at one point in the injury cascade but deleterious at another point. (3) The regulation of connexin expression can be mediated by various types of cells other than neurons and astrocytes following ischemic stress. For example, expression of connexins can be affected by macrophages and microglia.

Connexins and Epilepsy – Wiring the Gap

Epilepsy affects about 1% of the world's population and presents a greater burden of disease than stroke, in part because it can be a lifetime disease. About 25% of epileptics have seizures that do not respond to available pharmacotherapy. Recently, gap junctions have become an attractive target for the development of novel therapeutics, although at present there are no available clinically used drugs which knowingly affect GJC. Gap junctions are logical targets for anticonvulsants since they facilitate neuronal synchronization, a hallmark of epilepsy.

Blocking GJC potently depresses seizure activity in many diverse *in vitro* and *in vivo* animal models. *In vitro* studies include the use of low magnesium, 4-aminopyridine, bicuculline, kainate, and repeated tetanizations. *In vivo* studies showed efficacy of gap junctional blockade in animals made epileptic by focal cortical application of 4-aminopyridine, atypical absence seizures induced by the cholesterol synthesis inhibitor AY9944, and in a genetic model of absence epilepsy (GAERS rats). The problem with these studies is that all gap junction blocking agents used to date are nonspecific and have other

effects not related to gap junctions, including effects on neuronal excitability. Hence, epileptogenesis in genetically modified preparations is the obvious strategy, but very little work in this regard is published. The next question is whether connexin expression and function are altered in seizure models. There are reports of increased GJC in seizure models, using fluorescent recovery after photobleaching, which does not differentiate between glial and neuronal GJC. The gold standard for measuring GJC between cells is dual electrophysiological recordings wherein one measures the passage of electrical current from one cell to the other, an approach not yet published in seizure models. As for gap junctional protein expression, there are several conflicting reports showing both up- and downregulation of connexins (usually mRNA measures) in seizure models. The cellular locations and functionality of these epilepsy-related gap junctional proteins are unclear. Furthermore, there are no published reports concerning epilepsy, involving the recently discovered gap junctional proteins in the mammalian CNS, called pannexins, which localize primarily to neurons. Finally, connexon hemichannels control neuronal excitability and can release metabolites (e.g., ATP) and glutamate, but their role in epileptogenesis is not yet studied.

We tend to assume that neuronal gap junctions are the most important for epileptogenesis and anticonvulsant action. However, gap junctions are far more prevalent in glia, which are more abundant than neurons and are now being recognized as playing a key role in neuronal function and in epileptogenesis. Furthermore, when the various gap junctional blockers are administered, one does not know whether it is a neuronal and/or a glial action that is the basis of their very powerful anticonvulsant effects. Glia form a cellular network coupled by gap junctions through which ions such as calcium and potassium flow. They are considered to play a major role in spatial buffering of these ions, both of which have profound effects on neuronal excitability and synaptic transmission. For example, blocked glial gap junctions could focally raise extracellular potassium which could enhance epileptogenesis at lower concentrations and diminish seizures at higher concentrations. The role of glial gap junctions in epilepsy is surely critical, but as yet remains poorly understood.

Studies of brain gap junctions in humans and their role in seizures are sparse indeed. There are conflicting reports of increases and decreases of connexins (mainly mRNA and from glia) in tissue removed from epileptic or brain tumor patients. Furthermore, the expression of proteins does not tell us about their function. A recent report demonstrated an association of the connexin36 gene with juvenile myoclonic epilepsy, suggesting that mutation of this gene is related to the pathogenesis of this common type of human epilepsy. Further work integrating animal studies, genetic studies, and the development of novel pharmacotherapeutics is required so that we can better understand clearly the powerful role that GJC plays in epilepsy, and thereby develop better treatments.

Connexins and Brain Tumors – Enhancing the Gap

Glioma is the most common brain tumor. The severity of gliomas is divided into four categories, from most malignant tumor to benign tumor. In general, glioma is a space occupying mass in the brain and causes high intracranial pressure, vessel occlusion, and brain edema. In some preclinical and clinical studies, higher-grade gliomas express lower levels of Cx43.

Gap junctions have long been implicated in the control of cell proliferation and suppression of tumor progression. Downregulation of connexins, gap junctions, and GJC has been observed in a variety of tumor cell lines and cancer tissues. The expression of connexins and the presence of gap junction channels can be readily altered by exogenous factors. In the context of carcinogenesis, many tumor-promoting agents have the ability to downregulate connexin expression and GJC, while many antineoplastic agents are capable of reducing tumor cell growth in association with enhanced connexin expression and GJC.

The most direct mechanism to increase connexin protein resides at the level of gene expression, which has been shown to be increased by numerous endogenous factors, including growth factors. The effect of gap junction proteins on proliferation has been more directly studied by analyzing both the inhibition and the forced expression of connexins. The decrease in connexin expression and loss of gap junctions in many cases correlates with an increase in growth rate *in vitro* and tumor formation *in vivo*. In glioma studies, both Cx43 and Cx30, the main astrocytic connexins, act as tumor suppressors. Many transformed cells transfected with connexin cDNAs exhibit suppression of growth and/or tumorigenicity. However, some other connexins fail to exhibit this growth suppressive phenotype. Reasons for selectivity remain elusive, but may depend on connexin-interacting proteins and associated signaling pathways.

Upregulated GJC reduces the *in vitro* proliferation and *in vivo* tumor formation of a number of cancer cells, including glioma cells. On the other hand, gap

junctions appear to facilitate glioma cell aggregation and invasion. Astrocytes co-cultured with C6 glioma cells overexpressing Cx43 underwent phenotypic transformation, suggesting that tumor cells may affect phenotypical alteration of surrounding tissue through gap junctions.

Gene transduction is a potential strategy for glioma treatment. Gap junctions between glioma cells may play a critical role in the mechanism of this therapy. Glioma cells infected with the herpes simplex virus thymidine kinase gene can be killed upon exposure to ganciclovir. Neighboring glioma cells that are coupled by gap junctions to the targeted cell subsequently die due to a 'bystander effect.' The lethal bystander effect can be enhanced when the glioma cells are forced to express Cx43.

Recently, several reports have shown that Cx43 plays a role in suppressing tumor growth independent of gap junction formation. Many intracellular proteins have been identified to interact with the connexins, including Cx43, providing a context for signaling events which may influence cell growth without the formation of GJC.

Conclusions

In a simple sense, gap junctions can be viewed as forming an intercellular communication pathway that links cells together, allowing them to share a pool of small molecules and ions. However, it is clear that there are several levels of complexity associated with the regulation of the connexin genes and proteins (i.e., transcription, translation, posttranslational modifications, oligomerization, intracellular trafficking, and turnover), the gating of gap junction channels, the selectivity of GJC, and the specificity of cell types in which they are expressed. This is particularly evident in the nervous system, where these channels play an integral role in the functions of neurons and glial cells. The impact of gap junctions in normal brain function is realized when one considers the consequences of perturbed connexin expression and gap junction formation.

See also: Epilepsy; Epilepsy: Channelopathies; Gap Junctions and Hemichannels in Glia; Gap Junctions and Neuronal Oscillations; Gap Junctions and Electrical Synapses; Glioma; Tumors of the Brain and the Spinal Cord.

Further Reading

Carlen PL, Skinner F, Zhang L, et al. (2000) The role of gap junctions in seizures. *Brain Research Reviews* 32: 235–241.

Connors BW and Long MA (2004) Electrical synapses in the mammalian brain. *Annual Review of Neuroscience* 27: 393–418.

Contreras JE, Sanchez HA, Veliz LP, et al. (2004) Role of connexin-based gap junction channels and hemichannels in ischemia-induced cell death in nervous tissue. *Brain Research Reviews* 47: 290–303.

Kirchhoff F, Dringen R, and Giaume C (2001) Pathways of neuron–astrocyte interactions and their possible role in neuroprotection. *European Archives of Psychiatry and Clinical Neuroscience* 251: 159–169.

Laird DW (2006) Life cycle of connexins in health and disease. *Biochemical Journal* 394: 527–543.

Perez VJL, Frantseva MV, and Naus CC (2003) Gap junctions and neuronal injury: Protectants or executioners? *Neuroscientist* 9: 5–9.

Rouach N, Avignone E, Meme W, et al. (2002) Gap junctions and connexin expression in the normal and pathological central nervous system. *Biologie Cellulaire* 94: 457–475.

Rozental R, Giaume C, and Spray DC (2000) Gap junctions in the nervous system. *Brain Research Reviews* 32: 11–15.

Sohl G, Maxeiner S, and Willecke K (2005) Expression and functions of neuronal gap junctions. *Nature Reviews Neuroscience* 6: 191–200.

Sohl G, Odermatt B, Maxeiner S, Degen J, and Willecke K (2004) New insights into the expression and function of neural connexins with transgenic mouse mutants. *Brain Research Reviews* 47: 245–259.

Sutor B and Hagerty T (2005) Involvement of gap junctions in the development of the neocortex. *Biochimica et Biophysica Acta* 1719: 59–68.

Theis M, Sohl G, Eiberger J, and Willecke K (2005) Emerging complexities in identity and function of glial connexins. *Trends in Neuroscience* 28: 188–195.

Gap Junction Communication

C Giaume, Collège de France, Paris, France
C C Naus, University of British Columbia, Vancouver, BC, Canada

© 2009 Elsevier Ltd. All rights reserved.

For a long time, it has been well established that gap junctions provide a direct pathway for cell-to-cell communication. This mode of intercellular communication is widely present in all vertebrate species and also in most tissues and organs including the central nervous system (CNS). Compared to other types of cell interaction, gap junctional communication (GJC) mediates direct cytoplasmic exchanges from one cell to one or several neighbors. These exchanges are characterized by a weak ionic selectivity (electrical coupling) and a permeability for small molecules up to 1.8 kDa (metabolic or biochemical coupling). Such properties indicate that intercellular communication through gap junctions is likely relevant for excitable cells as neurons where they constitute the morphological substrate of electrical synapses. However, GJC also plays an important role in nonexcitable cells such as glial cells since they are permeable to signaling molecules.

Connexins (Cxs) are the molecular constituents of gap junction plaques consisting of closely packed hemichannels formed by a hexameric ring of Cxs which align head-to-head between neighboring cells to form functional and regulated intercellular channels. Invertebrates have a different, but functionally equivalent, family of proteins making up gap junctions, the innexins. Moreover, another gene family with functional characteristics similar to the Cxs and innexins has been discovered and named the pannexins, at least two of which are expressed in the mammalian CNS, whereas their function is not well defined. Interestingly, evidence is accumulating that Cxs interact by their cytoplasmic domains with other proteins that may be important for different aspects of their biology. These Cx-interacting partners, which are signaling and/or scaffolding proteins, have been proposed to be components of a dynamic multiprotein complex involved in the intracellular transport of Cxs, hemichannel assembly, full channel formation, junctional plaque assembly and stability, and regulation of channel permeability.

In addition to the well-known intercellular function of full gap junction channels, it has been reported that Cx channels can also operate as hemichannels, allowing exchanges of ions and small molecules between the cytoplasm and the extracellular space. Thus, in addition to providing a direct cell-to-cell pathway, Cxs are also involved in extracellular pathways mediating autocrine and paracrine processes.

Brain Connexins: Where and When

Between Neurons

Although coupling persists in many brain regions in the adult, during development there is a progressive decrease in Cx expression and GJC in CNS neurons. The correlation between changes in Cx expression and neuronal differentiation suggests that early GJC permits signaling molecules to diffuse between neurons and contributes to their differentiation. Eight different Cxs have been detected in defined neuronal cell populations (Cx26, Cx32, Cx36, Cx37, Cx40, Cx43, Cx45, and Cx47), mainly in the hypothalamus, striatum, inferior olive, hippocampus, olfactory bulb, retina, cerebral cortex, and cerebellum. Interestingly, in the CNS Cx36 represents the first Cx isoform expressed exclusively in neurons and exhibits typical functional properties compared to other Cxs. In the hippocampus of the adult rat, this Cx predominates and is mainly expressed in interneurons.

Between Glial Cells

In the CNS, glia represents the major population of cells coupled through gap junctions, and their connections establish compartments of communicating cells. This expression persists throughout adulthood, with qualitative and quantitative differences between glial classes: (1) each glial cell type expresses a set of specific Cxs, none being specific for glia; (2) the strength of coupling and the level of connexin expression depend on the glial cell type; and (3) not all classes of glial cells are coupled together, which contradicts the initial concept of a glial syncitium. Indeed, astrocytes are organized as a network composed of only subpopulations of glia that are under the control of biological signals, including neurotransmitters.

This network organization can be studied in acute brain slices after injecting low-molecular-weight tracers into a single astrocyte that results in the staining of large groups of 10–100 coupled cells. Such large coupling areas have also been described *in vivo* in the cortex, the hippocampus, and the cerebellum (Bergmann glial cells). Besides minor detection of Cx26, Cx40, and Cx45, two main junctional proteins are expressed in astrocytes: Cx43, which is widely present at both adult and embryonic stages, and Cx30, which is detected in mature gray matter astrocytes. In addition, the relative levels of Cx43 and

Cx30 expression vary according to the developmental stage and region studied.

Oligodendrocytes, the myelin-forming cells of the CNS, are endowed with gap junctions found in almost all areas of their plasma membranes; however, they are not detected between successive layers of myelin. Expression of Cx32 and Cx47 has been shown in oligodendrocytes studied *in vitro* as well as *in situ*, and indirect evidence suggests that Cx29 is also expressed in oligodendrocytes.

Finally, the presence of Cx43 has been described in microglial cells. In normal tissues, <5% of microglial cells in adult rat cerebral cortex are Cx43 immunoreactive, and in culture only low levels of Cx43 have been detected. However, *in vivo* after a stab wound or in culture after treatment with cytokines, Cx43 is observed at the interfaces between activated microglia cells that become dye coupled.

The Language of Connexins

The molecular complexity of Cx channels bears important consequences on their function. Cxs speak an articulate language and have evolved diverse forms of intercellular communication among cells. First, channels composed of different Cxs are endowed with distinct biophysical and regulatory properties. Although Cxs do not form ion-selective channels in the classical sense, they exhibit a certain degree of selectivity with a preference for cations versus anions. Because gap junctions are permeable to cyclic nucleotides and inositol 1,4,5-trisphosphate (IP3), which are negatively charged at normal intracellular pH, ionic selectivity may allow the discrimination and modulation of the kinetics of intercellular propagation of known second messengers. Size selectivity is also a Cx-specific feature. It had long been thought that connexins were permeable to molecules up to 1800 Da. The use of different fluorescent tracers demonstrated that some connexins can discriminate between molecules of 400 and 600 Da, thus favoring the passage of a distinct set of signals. A second form of intercellular language is based on compatibility – that is, the ability of Cxs to functionally interact with different family members. On the basis of Cx topology relative to the plasma membrane, the docking of two facing hemichannels implicates the recognition between their two extracellular domains. Experimental data have shown that the ability to establish direct communication between adjacent cells is dependent, at least in part, on the ability of Cxs to discriminate between different partners. A consequence of the restricted ability of Cxs for heterotypic interactions is that expression of different Cxs in a group of cells may provide a powerful means to limit or segregate intercellular communication. Finally, the functional status of intercellular channels can be regulated by both covalent and noncovalent modifications of Cxs which affect the channel structure, a third form of intercellular language called 'gating.' Despite the overall degree of homology, Cxs exhibit in their two major cytoplasmic domains stretches of unique sequences that probably adapt channel properties to the specific tissue requirements. The behavior of Cx channels has been studied extensively in various *in vitro* expression systems, and these biophysical studies have established that individual Cxs are endowed with unique gating properties, including voltage dependency, pH sensitivity, subconductance states, and residual conductance.

What Goes through Connexin Full and Hemichannels?

Intercellular Communication through Gap Junctions in Neurons and Glia

The initial evidence of gap junction-mediated exchanges of ions was obtained by electrophysiological recordings of electrical coupling between neurons or between glial cells. These pioneering works were then confirmed by a number of studies that established the contribution of gap junctions to electrical coupling between CNS cells and their regulation by several factors, including neurotransmitters, voltage, and certain brain pathologies or lesions. This was further investigated by expressing neuronal or glial Cxs in cell lines lacking endogenous Cxs and by using the double patch clamp technique to establish selectivity profiles for cations versus anions and the number of charges. The development of dual patch clamp recordings from neurons in brain slices permitted functional studies of electrical synapses.

Of great interest is the permeability of Cx channels for small molecules since this intercellular exchange supports a biochemical and metabolic coupling that is likely important for signaling between nonexcitable cells. Thus, an important issue concerning the role of gap junctions between glia is the identification of signaling molecules exchanged through this intercellular pathway. Since the permeability profile of Cx channels depends on their molecular constituents, it is expected that the amount and the nature of permeating ions or signals is directly related to their identity. The diversity of Cx expression in CNS cells suggests that intercellular exchanges are likely dependent on several types of gap junction channels made by these Cxs. Although essential, this issue has been very difficult to address *in situ* due to technical limitations and the expression of several Cxs per cell. For this reason, the situation is much more favorable *in vitro* for

cultured astrocytes since most gap junction channels are made up of Cx43, as demonstrated by quantitative analysis showing that 95% of the electrical coupling is lacking in astrocytes cultured from Cx43 knockout mouse brain.

Although Cx43 channels are poorly selective for ions, double patch clamp recordings have established that they favor the passage of cations versus anions. In addition, there is converging evidence indicating that intercellular exchange of signaling molecules occurs in cultured astrocytes and in glioma cell lines transfected with Cx43. The permeability of astrocyte gap junctions for glucose and its metabolites, including lactate, was initially characterized by using radiolabeled compounds and by investigating their intercellular diffusion by adapting the scrape-loading dye-transfer technique. Moreover, the use of several unrelated gap junction inhibitors has shown that this diffusion occurs through gap junction channels. This has been confirmed by use of a fluorescent glucose molecule, 2-(*N*-(7-nitrobenz-2-oxa-1,3-diazol-4-Y1) amino)-2-deoxyglucose. Interestingly, gap junctions were shown to participate in the propagation of intercellular calcium waves. In this context, the demonstration that astrocyte gap junctions are permeable to IP3 has contributed to establish the key role of this second messenger in the propagation process. In addition, transfection of C6 cells with Cx43 allowed the demonstration that several nucleotides, particularly ATP and ADP, can be transferred through Cx43 channels. Finally, radiolabeled glutamate and glutamine have also been shown to permeate Cx43 channels in astrocytes.

Whereas in some systems electrical coupling between neurons is strong enough to allow transmission of action potentials, such as in early development, in many cases coupling is weak. In addition, gap junction channels act as a low-pass filter, and gap junctions are generally located primarily on dendrites so that fast action potentials are attenuated. Taking these considerations into account, it is unlikely that coordination of spike propagation through gap junctions is the primary function of coupling between neurons. Although this statement argues for a role of biochemical coupling between coupled neurons, there is very little information about what can cross neuronal gap junctions. Indeed, besides the demonstration that glycine can be exchanged through gap junctions between neurons in the retina, the other evidence for the passage of signaling molecule concerns IP3. In this case, the permeation of IP3 was reported to contribute to the propagation of calcium waves in neuronal domains, whereas an increase in Ca^{2+} alone was not sufficient.

Connexin Hemichannels: An Alternative Release Pathway

Classical pathways for neurotransmitter release that have been extensively described in neurons and in astrocytes include calcium-dependent vesicular release and the purinergic ionotropic receptor (P2X7). In addition, Cx hemichannels have also been proposed to operate in certain retinal neurons and in astrocytes. The unitary conductance of Cx hemichannels has been recorded in cultured astrocytes and is twofold higher than that recorded for the Cx43 full channel. The activity of these Cx hemichannels is very low in resting conditions but can be increased either after ischemic treatment or metabolic inhibition or in low divalent solution. There is evidence that ATP and glutamate can be released from astrocyte Cx hemichannels under these nonphysiological conditions.

Functional Consequences of Cx-Mediated Communication

In addition to its contribution to synaptic transmission, the permeability of gap junction channels for ions may serve to maintain ionic homeostasis between cells, particularly between glial cells, which have for a long time been proposed to play a role in ionic buffering. This is particularly relevant for K^+ since astrocytes mainly exhibit voltage-dependent and passive conductance for this ion. Indeed, several studies have also demonstrated that in cultured astrocytes, GJC is also involved in the intercellular homeostasis of Na^+ and Ca^{2+}. This is particularly relevant when the pharmacological heterogeneity of astrocytes is considered, since not all astrocytes will respond primarily to the stimulation of a membrane receptor agonist. However, the permeability of gap junction channels for calcium ions and IP3 will enable neighboring coupled cells to develop a secondary calcium response. Accordingly, the regulation of GJC may contribute to unmask a pharmacological heterogeneity of astrocytes. In addition, the permeability of Cx43 channels for IP3 is proposed to contribute to the propagation process of intercellular calcium waves, establishing an important functional role for gap junctions in astrocyte physiology and neuroglial interactions.

The possibility for astrocytes to exchange energetic compounds provides an intercellular pathway that may participate in metabolic trafficking between the source of energetic substrates and the main site of energy consumption – respectively, the blood circulation and neurons. In this way, gap junctions may contribute to the distribution of cellular metabolites and the maintenance of the energy status over a glial

cell population. This role for gap junctions is further strengthened by the enriched expression of Cx43 and Cx30 antibodies at contacts between astrocytic endfeet enwrapping blood vessels. This may create local zones of intercellular communication that favor glucose uptake and its spreading into several astrocytes through a metabolic network.

The observation that astrocyte gap junctions are permeable to glutamate and glutamine is of particular interest since after release from glutamatergic neurons, the neurotransmitters are rapidly inactivated by cellular uptake, predominantly into astrocytes. Glutamate is then amidated to glutamine by the ATP-consuming reaction catalyzed by glutamine synthetase, an enzyme present in astrocytes but not in neurons. Glutamine is then released by astrocytes and taken up by neurons, where its hydrolysis provides glutamate. Consequently, the regeneration of this neurotransmitter should not solely be considered as the result of the contribution of isolated and independent astrocytes but, rather, as a result of groups of communicating astrocytes that can exchange glutamate and glutamine. Thus, the permeability for this excitatory amino acid may play an important role in the interaction between astrocytes and neurons that are in part mediated by glial release of glutamate. In certain conditions, this release could also be mediated by Cx hemichannels. A similar situation may occur for the release of ATP that was also shown to operate through Cx43 hemichannels in astrocytes. In these situations, Cx hemichannel-mediated communication has been proposed to contribute to the propagation of calcium waves between astrocytes as well as to the modification in neuronal activity. Finally, Cx hemichannels in horizontal cells of the retina exert a feedback-mediated response that affects the activity of calcium channels and subsequently glutamate release controlling the output of the cones.

Conclusions and Perspectives

It is clear that the analysis of the functional state of gap junctions in the brain is still limited for at least several reasons. First, it is still difficult to study intercellular coupling in brain slices, especially at adult stages, and only a few studies have been performed in the mature brain. Second, dye-coupling experiments may not be sufficiently informative to identify the role of gap junctions in nonexcitable cells. Indeed, this approach only allows one to follow the intercellular diffusion of small molecules (Lucifer yellow, biocytin, 6-carboxyfluorosceine, sulforhodamine B, etc.) that have no biological relevance and solely diffuse along their concentration gradient. Also, not all Cxs have the same permeability for these intercellular tracers. Moreover, the morphology, the location of coupling sites, and the electrical properties of astrocytes studied *in situ* and in cultures are not suitable for optimizing the information obtained by dual patch clamp recordings. Consequently, in the future it will be beneficial to develop and use new molecules or probes that will allow the tracking of biological signals which are relevant to astrocytic gap junctions. Also, many of the experiments performed to demonstrate that gap junctions in the CNS are permeable to signaling molecules used cultured cells. This limits the relevance of these observations, and in future this should be assessed in more integrated systems to weigh the physiological significance of this mode of intercellular communication. Finally, the role and the real contribution of Cx hemichannels is still unclear and needs further investigation to evaluate the conditions under which they are functional and which physiologically relevant molecules cross them.

See also: Calcium Waves in Glia; Gap Junctions and Hemichannels in Glia; Gap Junctions and Neuronal Oscillations; Gap Junctions and Electrical Synapses; Gap Junction Abnormalities and Disorders of the Nervous System; Gap Junctions: Metabolic Exchanges; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging.

Further Reading

Bennett MV, Contreras JE, Bukauskas FF, and Saez JC (2003) New roles for astrocytes: Gap junction hemichannels have something to communicate. *Trends in Neurosciences* 26: 610–617.

Bukauskas FF and Verselis VK (2004) Gap junction channel gating. *Biochimica Biophysica Acta* 1662: 42–60.

Giaume C and McCarthy KD (1996) Control of gap-junctional communication in astrocytic networks. *Trends in Neurosciences* 19: 319–325.

Giaume C, Tabernero A, Medina JM, Swanson RA, and Magistretti PJ (1997) Metabolic trafficking through astrocytic gap junctions. *Glia* 21: 114–123.

Hormuzdi SG, Filippov MA, Mitropoulou G, Monyer H, and Bruzzone R (2004) Electrical synapses: A dynamic signaling system that shapes the activity of neuronal networks. *Biochimica Biophysica Acta* 1662: 113–137.

Nagy JI and Rash JE (2000) Connexins and gap junctions of astrocytes and oligodendrocytes in the CNS. *Brain Research Reviews* 32: 29–44.

Nakase T and Naus CC (2004) Gap junctions and neurological disorders of the central nervous system. *Biochimica Biophysica Acta* 1662: 149–158.

Pereda AE, Rash JE, Nagy JI, and Bennett MVL (2004) Dynamics of electrical transmission at club endings on the Mauthner cells. *Brain Research Reviews* 47: 227–244.

Rouach N, Avignone E, Même W, et al. (2002) Gap junctions and connexin expression in the normal and pathological central nervous system. *Biology of the Cell* 94: 457–475.

Scemes E (2000) Components of astrocytic intercellular calcium signaling. *Molecular Neurobiology* 22: 167–179.

Theis M, Söhl G, Eiberger E, and Willecke K (2005) Emerging complexities in identity and function of glial connexins. *Trends in Neuroscience* 28: 188–195.

Volterra A and Meldolesi J (2005) Astrocytes, from brain glue to communication elements: The revolution continues. *Nature Reviews Neuroscience* 6: 626–640.

Gap Junctions and Electrical Synapses

M V L Bennett, Albert Einstein College of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Synapses are specialized areas of contact between neurons, between sensory cells and neurons, or between neurons and effector cells. Synapses mediate communication between cells; synaptic transmission from pre- to postsynaptic cell can be mediated by electrical current flowing from the presynaptic cell or by a neurotransmitter molecule secreted by the presynaptic cell. (Neurons are not known to communicate by mechanical means, but some muscles are mechanosensitive, and junctions between them can mediate mechanical transmission.) A key word in the first sentence is 'specialized'; there may be disagreement about what it means. Electrical and chemical interactions can occur over greater separations than at the canonical synapses. 'Ephapse' and 'ephaptic transmission' have been used to describe electrical interaction at a fortuitous close apposition of two cells; 'field effects' denotes interactions across greater distances and may involve many cells. Analogs for chemical transmission include transmitter 'spillover,' in which a high frequency of presynaptic secretory activity causes transmitter to diffuse beyond where its usual receptors are located, and paracrine action, in which transmitter is released and acts at some distance without an obviously specialized relation between secreting and responding cells. Endocrine actions by hormones involve widely separated secretors and targets. The same molecule can be a transmitter at synapses, have paracrine actions, or be a hormone, such as epinephrine.

The most common and best characterized electrical synapses are gap junctions between neurons. A gap junction is a cluster of small channels connecting the interiors of adjoining cells (Figure 1(b)). There are also electrical inhibitory synapses, originally described at the Mauthner cell of teleosts and subsequently in the mammalian cerebellum. This type of synapse is uncommon but will be described briefly in the section titled 'Electrical Inhibitory Synapses'.

Some appositions between neuronal cell bodies are specialized in that intervening glia are absent (e.g., hippocampal pyramidal cells, some hypothalamic neurons), but there do not appear to be gap junctions. Electrical interactions are more likely to be found between these cells than between neurons separated by glial ensheathment. If there are interactions, these relatively unspecialized appositions might well be considered electrical synapses.

Since gap junctions provide ultramicroscopic cytoplasmic continuity between cells, their presence could be viewed as contradicting Cajal's neuron doctrine that each neuron is a distinct cell (although he wrote in his wonderful autobiography, "Neuronal discontinuity . . . could sustain some exceptions"). As will be seen, the properties of gap junctions are reasonably subtle in spite of their relative simplicity; a functional syncytium, as coupled cells have been described, does not necessarily mean that all cells do the same thing.

Properties of Gap Junctions

Molecular Composition of Gap Junctions: Connexins and Pannexins

Known gap junctions in vertebrates are composed of connexins, encoded by a gene family in mammals of ~20 members. Connexins are commonly named by their predicted molecular mass to the nearest kilodalton with a prefix for species or an additional significant figure where necessary to avoid ambiguity, as in human Cx30, Cx30.2, and Cx30.3, encoded by three distinct genes. Molecular weights range from 25 to 57 kDa. Most connexin genes have the entire coding region in a single exon; the Cx36 gene and several others have the coding region in two exons. In some connexins, alternative splicing in the 5′ noncoding region permits use of different tissue-specific promoters.

The connexins are tetraspan integral membrane proteins, that is, they cross the membrane four times. The N- and C-termini (NT, CT) are cytoplasmic, and there are four transmembrane domains (TM1–TM4) connected by two highly conserved extracellular loops (E1 and E2) and a cytoplasmic loop (CL). This topology was confirmed for several connexins by antibody labeling and identification of peptides following protease treatment that cleaved at exposed sites.

The degree of conservation of sequence differs along the molecule. The N-terminus, membrane-spanning regions and extracellular loops are well conserved; the cytoplasmic loop and C-terminal domain are much more variable, in length as well as sequence, and length of the CT domain accounts for most of the differences in molecular weight. Conservation of the extracellular loops may reflect the shared function of docking with another hemichannel. Several regions in the CT contain binding sites for protein–protein interaction, including with the cytoskeletal protein ZO-1.

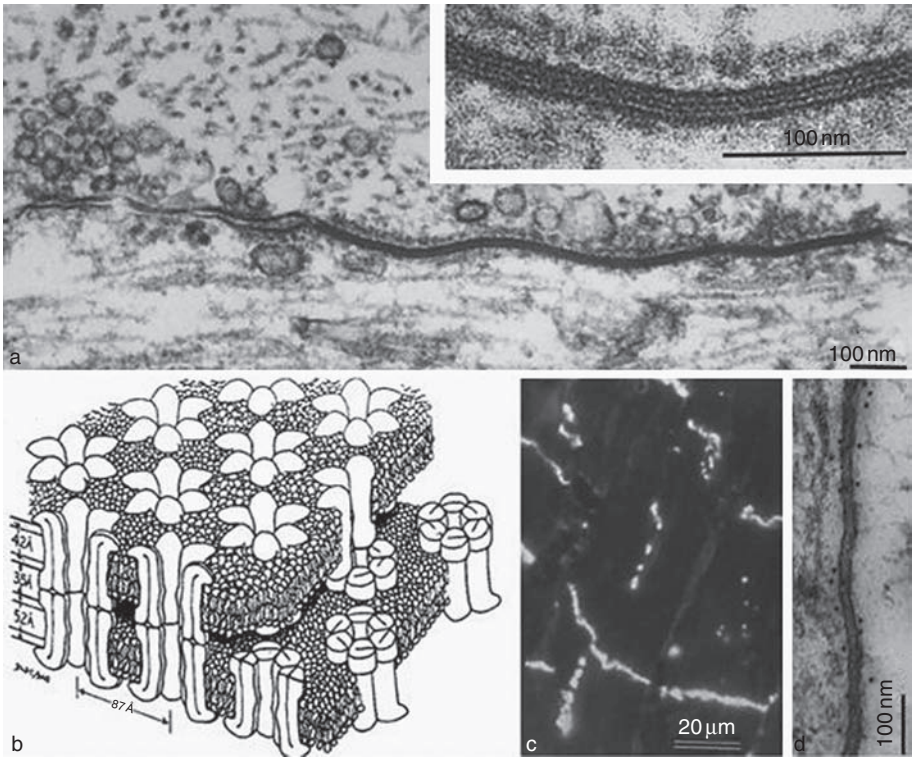

Figure 1 Gap junction morphology. (a) A gap junction in thin section, an electrical synapse on a Mauthner cell dendrite. A long gap junction is seen between the axon (upper) and dendrite (lower). The axon also contains presynaptic vesicles. In an enlarged view (inset) the gap junction exhibits the typical seven-layer structure. (b) Diagram of a gap junction showing how hexamers of the subunit protein, connexin, in each membrane form a channel crossing the extracellular gap. (c) Immunofluorescence stain of ventricular myocytes with an anticonnexin 43 antibody. Intercalated discs and some lateral appositions are labeled. (d) Immunogold labeling of a cardiac gap junction with the same antibody, electron micrograph. (a, c) From Bennett MVL, Barrio LC, Bargiello TA, Spray DC, Hertzberg E, and Saez JC (1991) Gap junctions: New tools, new answers, new questions. *Neuron* 6: 305–320. (b) From Makowski L, Caspar DL, Phillips WC, and Goodenaugh DA (1977) Gap junction structures. II. Analysis of the x-ray diffraction data. *Journal of Cell Biology* 74: 629–645.

Given that the human and mouse genomes have been sequenced, few additional mammalian connexins are likely to be found. Different connexins assemble to form junctions that differ in single-channel conductance, gating, permeability (depending on both size and charge), and temporal and spatial patterns of expression. At a gap junction, each cell membrane provides hemichannels, or connexons, that dock one to one with hemichannels in the other membrane. Hemichannels are hexamers, homomeric if they are comprised of one kind of connexin and heteromeric if they are comprised of more than one kind.

Many connexins are expressed in more than one kind of cell, and co-expression of multiple connexins is common in a given cell type. Heteromeric hemichannels do occur, although their prevalence and stoichiometry are poorly known. (Factors controlling oligomerization are not well understood, but assembly does not appear to be random.) Hemichannels of the same kind form homotypic junctions; hemichannels of different kinds form heterotypic junctions. It is also useful to distinguish homo- and heterocellular junctions, that is, junctions between the same or between different types of cells, and these junctions may be homo- or heterotypic. (The terms 'homologous' and 'heterologous' are sometimes used for junctions between the same and between different kinds of cells, respectively.) 'Compatibility' between connexins, the property that allows the formation of heterotypic junctions, has been determined for only a small fraction of the 190 possible pairwise combinations of homomeric hemichannels (for 20 different connexins), but it is not universal and may be uncommon. For example, Cx32, Cx36, and Cx43 do not form junctions in any of the three pairwise heterotypic combinations. Compatibility in this sense is only one of the factors determining where gap junctions form; cell adhesion molecules and almost certainly other as yet undetermined mechanisms are also operative.

The connexins found in the nervous system include the following: the most common neuronal connexin is Cx36 (which is also expressed in pancreatic islets). An additional neuronal connexin is Cx45. Retinal horizontal cells are coupled by Cx57 junctions. Other connexins may be expressed by neurons transiently during development. Astrocytes express predominantly Cx43, but also Cx26 and Cx30. Oligodendrocytes express Cx29, Cx32, and Cx47. Distributions may vary. Astrocyte Cx26 and Cx30 can mediate coupling to oligodendrocyte Cx32.

Identification of the connexin genes permits gene knockout, antisense or small interfering RNA(siRNA) knockdown, replacement of the coding sequence with a reporter gene, and exogenous expression in cell lines or other tissues. Knowledge of sequence also permits *in situ* hybridization, polymerase chain reaction, and the raising of antibodies against specific (or common) sequences to be applied to identification and localization of connexin expression. Site-directed mutagenesis can be used to identify residues critical for particular junctional properties. These approaches have led to an explosion of knowledge of function and distribution.

Although most mammalian connexins have orthologs in other vertebrates and known vertebrate gap junctions are connexin based, gap junctions in nematodes, mollusks, and arthropods (the protostome line) are formed of a family of proteins known variously as innexins (which they were named when they were thought to be expressed only in invertebrates) or pannexins (when it was demonstrated that the same gene family is found in both major branches of the animal kingdom, protostomes and deuterostomes, which include vertebrates). Cnideria have pannexins, and connexins may have evolved in the deuterostome lineage. *C. elegans* has 25 innexins/pannexins; *Drosophila* has eight. Mouse and human each have three pannexins. Two of the mammalian pannexins are expressed in neocortex, hippocampus, and cerebellum, apparently in association with the postsynaptic density. The mammalian pannexins can form gap junctions when exogenously expressed in *Xenopus* oocytes but may only form hemichannels in mammals. The many convergences and few differences between connexin gap junctions in deuterostomes and pannexin gap junctions in protostomes will be noted in the section titled 'Evolution.' Otherwise, gap junction will refer to connexin-based junctions.

Structure

In an early electron microscopic thin section study of gap junctions, the extracellular space between the adjoining cells appeared reduced to a narrow gap (Figure 1(a)), and the channels crossing the gap were not well resolved. Thus, the term 'gap junction' derives from a structural characteristic not obviously related to function. The gap distinguishes these junctions from 'tight junctions' or 'zonulae occludentes' (singular: zonula occludens), in which the intercellular space between adjoining cells appears completely occluded. The gap could be penetrated by small extracellular markers, such as colloidal $La(OH)_3$. Freeze-fracture replicas in the plane of the junction show junctional particles on the protoplasmic face and corresponding pits on the external face (Figures 2 and 3). These structures are often hexagonally packed. With good resolution, the particles show a central depression that presumably is the site of the aqueous channel (Figure 2, inset). Gap junctions can be isolated by several techniques that

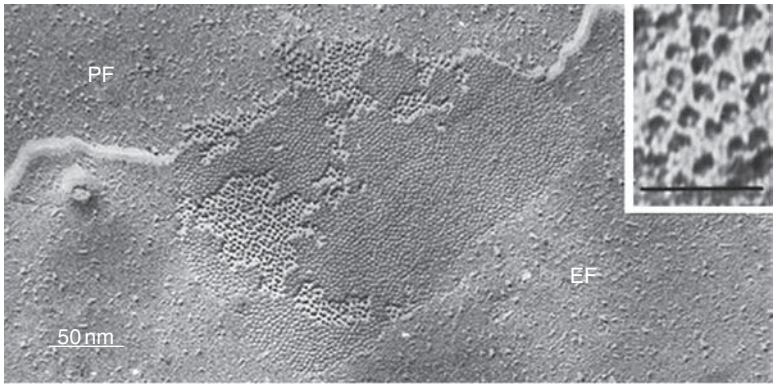

Figure 2 Gap junction between blastomeres of a fish shown by freeze fracture electron microscopy. The protoplasmic leaflet of the plasma membrane (P face, or PF) lies to the upper left; the external leaflet (E face, or EF) lies to the lower left. The gap junction is characterized by an aggregate of P-face particles and complementary E-face pits where the particles were pulled out in the fracture process. The inset shows the particles at higher magnification. Many of them show a small central pit that corresponds to the channel. Reprinted from Ne'eman Z, Spira ME, and Bennett MV (1980) Formation of gap and tight junctions between reaggregated blastomeres of the killfish, Fundulus. *American Journal of Anatomy* 158: 251–262, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc. Copyright (1980).

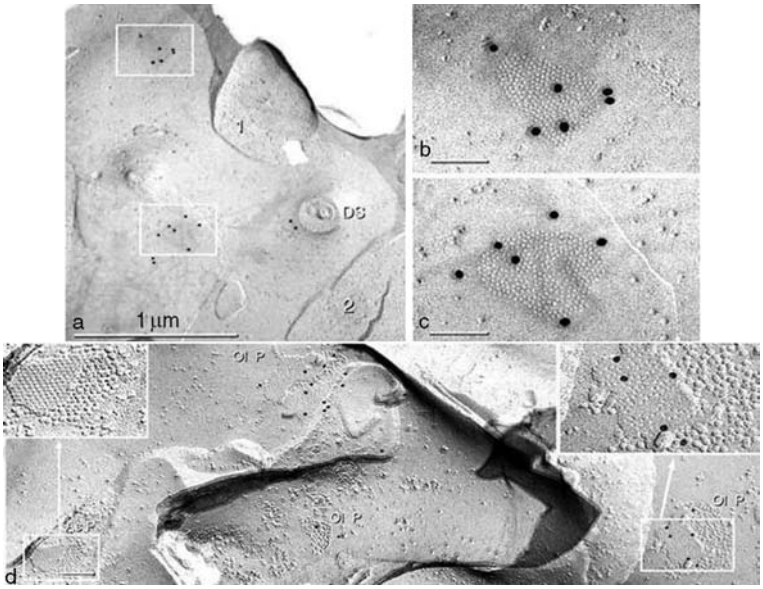

Figure 3 Freeze fracture replica immunolabeling (FRIL) of gap junctions between cells of the nervous system. Gap junctions, characterized by aggregates of external leaflet face (E-face) pits or plasma membrane face (P-face) particles, are clearly labeled by specific antibodies followed by secondary antibodies with attached gold beads of predetermined diameter. The technique labels only a small fraction of the replicated channels, and the length of the antibodies permits the gold beads to be some distance from the edge of the junctions. (a) A dendrite in the inferior olive of the rat shows two gap junctions as arrays of E-face pits. 1 and 2 are endings on the dendrite; DS is a cross-fractured dendritic spine. The gap junctions are shown enlarged in (b, c), and each is immunolabeled with six 20 nm gold beads. The primary antibody is against Cx36, a neuron-specific connexin. (d) Gap junctions on an inferior olive oligodendrocyte show both P-face particles (Ol P) and E-face pits; the junctions are labeled with an anti-Cx32 primary antibody and 20 nm gold beads; the rightmost junction is enlarged. A gap junction on an astrocytic process in the same replica (lower left, identified by aquaporin square arrays of particles not included in the micrograph) is negative for Cx32, but these junctions can be labeled using Cx43 antibodies. Scale bar = 100 nm (a–c). Reproduced from Rash JE, Staines WA, Yasumura T, et al. (2000) Immunogold evidence that neuronal gap junctions in adult rat brain and spinal cord contain connexin-36 but not connexin-32 or connexin-43. *Proceedings of the National Academy of Sciences of the United States of America* 97: 7573–7578, with permission from National Academy of Sciences, USA. Copyright (2000).

dissolve away surrounding membrane. Isolated junctions, when negatively stained, usually show domains of hexagonal symmetry (Figure 4). There is a central dot at the presumed site of the channel lumen and a hexagonal lattice surrounding the dots that results from stain in the intercellular gap, which is interrupted by the bridging structures (cf. Figure 1(b)). Atomic force microscopy of separated junctional membranes confirms the general features of the junctions and suggests structural changes associated with gating (Figure 5). Low-dose electron microscopy of isolated junctions reveals four alpha helical regions crossing the membrane, but matching to the amino acid sequences is not yet possible (Figure 6). Scanning cysteine accessibility mutagenesis is revealing channel lining residues, although the data are still somewhat controversial.

When hemichannels come together to form a cell–cell channel, one hemichannel is rotated by 30° relative to the other (Figure 7). Antibodies against the junctional protein or connexin can be used to label gap junctions at the light and electron microscopic levels; the intercalated discs of cardiac ventricle are illustrated in Figures 1(c) and 1(d)). Freeze fracture followed by immunolabeling of residual protein on the replica allows precise localization of particular connexins in identifiable gap junctions (Figure 3).

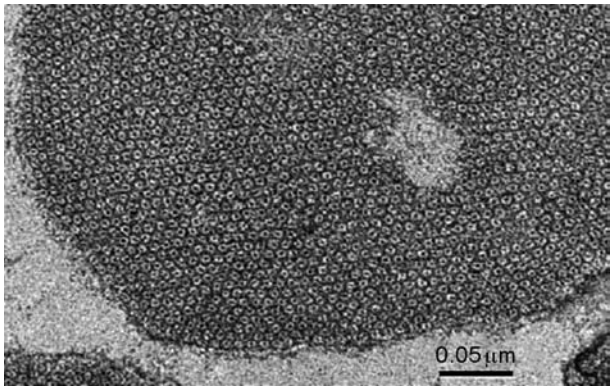

Figure 4 Isolated liver gap junction from rat, negatively stained with uranyl formate. The junctional channels are hexagonally arrayed. The hexagonal lattice represents stain in the extracellular gap; the central dots mark the location of the channel. From Hertzberg EL and Gilula NB (1979) Isolation and characterization of gap junctions from rat liver. *Journal of Biological Chemistry* 254: 2138–2147.

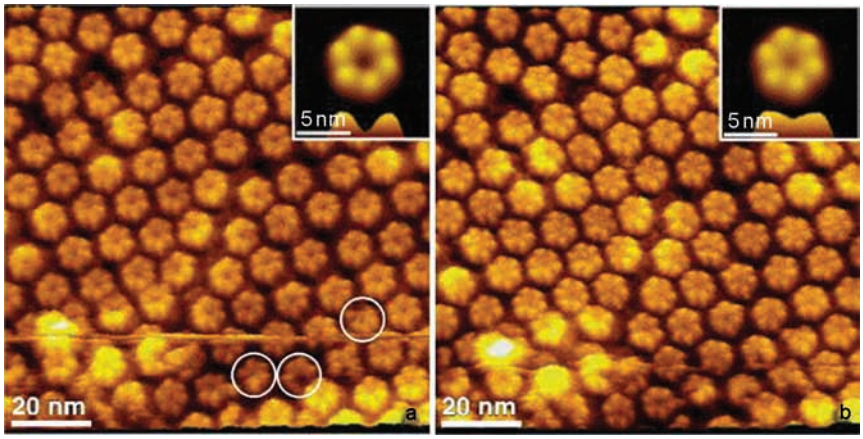

Figure 5 Atomic force microscopy of external face of an isolated gap junction that has been split open mechanically. The same junction in zero Ca^{2+} (left) shows larger openings than after addition of 0.5 mmol l^{-1} Ca (right). The insets are averages. The (hemi)channel has six subunits. The circles indicate channels that presumably have been damaged. The apparent openings in the channels are wider in zero Ca^{2+} (left) than in 0.5 mmol l^{-1} Ca^{2+}, a concentration that closes gap junction channels when in the cytoplasm. From Muller DJ, Hand GM, Engel A, and Sosinsky GE (2002) *EMBO Journal* 21: 3598–3607.

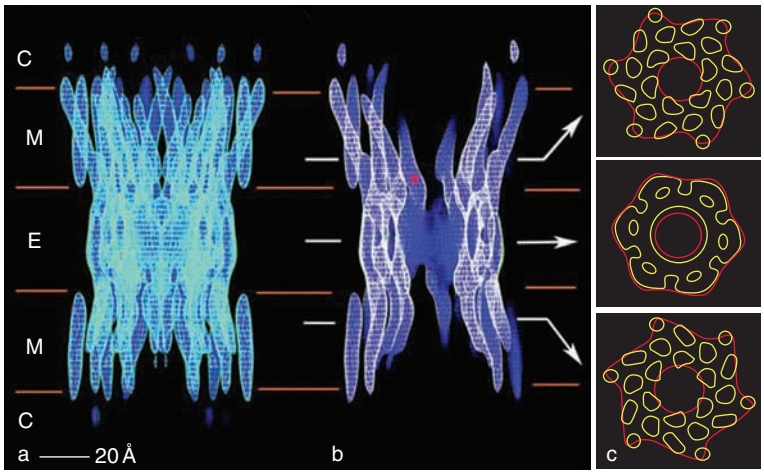

Figure 6 Molecular organization of a gap junction channel. (a) Side view. (b) Side view showing the channel interior. The approximate boundaries for the membrane bilayers (M), extracellular gap (E), and the cytoplasmic space (C) are indicated to the left. (c) Cross-sections parallel to the membrane at the levels indicated by the arrows in (b). The red contours in (c) define the boundary of the channel. The yellow contours are at increased density and are consistent with four transmembrane alpha helices for each of the six connexins, as predicted from the amino acid sequence. The red asterisk in (b) marks the narrowest part of the channel, where the channel lumen is ~1.5 nm in diameter, excluding the contribution of amino acid side chains that were not resolved. Comparison of upper and lower cross sections in (c) shows the 30° rotation in docking the two hemichannels. Note that both are viewed from the same side of the channel and are therefore mirror images. From **Figure 2** in Unger VM, Kumar NM, Gilula NB, and Yeager M (1999) Three-dimensional structure of a recombinant gap junction membrane channel. *Science* 283: 1176–1180.

Distribution

Gap junctions connect many kinds of nonneuronal excitable cells where they mediate propagation of action potentials, such as in heart, smooth muscle, and pancreatic islets. Gap junctions are also found widely in inexcitable tissues, such as liver, epidermis, exocrine pancreas, salivary glands, neuroglia, and lining of the gut from mouth to anus. Although gap junctions are very common in these tissues, they are absent in many contacts between specific kinds of cells. They are not found between adult vertebrate skeletal muscle fibers or erythrocytes and are rare or absent between neurons and glia and between many kinds of neurons that come into close contact.

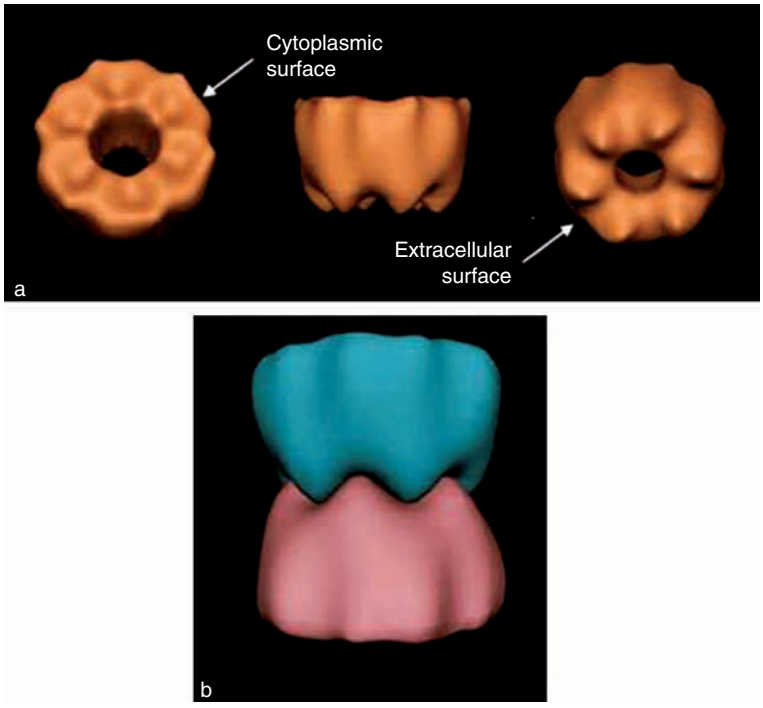

Figure 7 Fitting of the extracellular surfaces of two hemichannels to form a cell–cell channel. (a) Surface representation of a single hemichannel (connexon) viewed from the cytoplasm, from the membrane side, and from the extracellular side. The structure of the external and cytoplasmic surfaces are quite different. (b) Two hemichannels fit together at their external surfaces by rotating one 30°. From Perkins GA, Goodenough DA, and Sosinsky GE (1998) Formation of the gap junction intercellular channel requires a 30 degree rotation for interdigitating two apposing connexons. *Journal of Molecular Biology* 277: 171–177.

Electrical Coupling and Circuit Parameters

When a current is passed in one cell to change its potential, a smaller potential change is recorded in cells joined to it by gap junctions (Figure 8). Current flows from cell to cell through the junctional channels just as in electrotonic spread along a core conductor (hence the terms 'electrotonic junction' and 'electrotonic synapse'); the gap junctions act as a conductor, and both hyperpolarizing and depolarizing potentials are recorded in the follower cell. This current is carried largely by K^+ ions, the primary mobile ion in the cytoplasm. Before concluding that the cells are coupled by gap junctions, cytoplasmic continuity must be excluded, which is commonly done with gap junction blockers (see the section titled 'Pharmacological block of gap junctions') or by injecting a visualizable molecule that is too large to permeate gap junctions, such as fluorescently labeled dextran (see the section titled 'Dye coupling and permeability'). Also, recording outside the cells may be necessary to demonstrate that the intracellular potential in the coupled cell is not a result of a voltage drop in the extracellular space. Both electrical and dye (see the section titled 'Voltage gating of gap junctions') measurements indicate that there is little leakage from the channels as they cross the gap and that they constitute a private pathway between the cell interiors that is not accessible from the extracellular space.

For steady state potentials, the ratio of pre- to postjunctional potential is termed the coupling coefficient or coupling ratio. In the equivalent circuit of two isopotential cells connected by a gap junction, each cell is represented by its conductance and capacitance in parallel, and the cells are connected by the junctional conductance, g_j (Figure 9). (Isopotential means that the cytoplasmic resistance is small compared with the membrane resistance and that all points within the cell are at virtually the same potential. In neuronal processes, i.e., the axon and dendrites, cytoplasmic resistance is no longer negligible, and the potentials in these processes can differ from that in the soma.) For small potentials, junctional and nonjunctional conductances can be treated as constant. The junctional membrane has little capacitance compared with nonjunctional membrane because of its small area and the presence of the gap, which is likely to be at a potential close to that of the surrounding extracellular space. The parameters of steady state transmission (which excludes the effect of cell capacitances) are easily described using the circuit. The

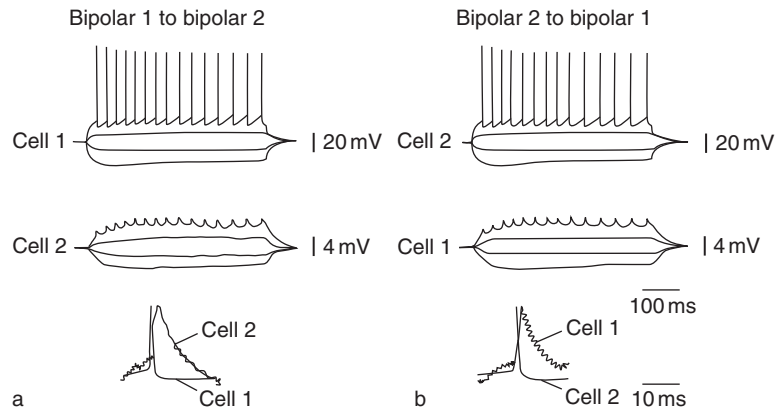

Figure 8 Electrical coupling mediated by gap junctions between bipolar (inhibitory) interneurons in the visual cortex of the rat. Four superimposed records with current applied in cell 1 in the upper set in (a) and in cell 2 in the upper set in (b); recordings from the follower cell at higher gain in the middle sets of traces. Both hyperpolarization and subthreshold depolarization spread between cells. Larger polarizations in the polarized cell result in larger polarizations in the follower cell, and the relation is essentially linear under these conditions. Impulses, which are evoked by the strongest current, are more attenuated than steady polarizations, and the peak of the 'spikelet' in the postsynaptic, nonstimulated cell is delayed compared to the presynaptic spike; superimposed presynaptic spikes and postsynaptic responses are shown in the bottom records at greater time resolution. Pre- and postsynaptic are defined here as the stimulated and follower cells; the synapse is symmetric, and transmission is bidirectional. Reproduced from Venance L, Rozov A, Blatow M, et al. (2000) Connexin expression in electrically coupled postnatal rat brain neurons. *Proceedings of the National Academy of Sciences of the United States of America* 97: 10260–10265, with permission from National Academy of Sciences, USA. Copyright (2000).

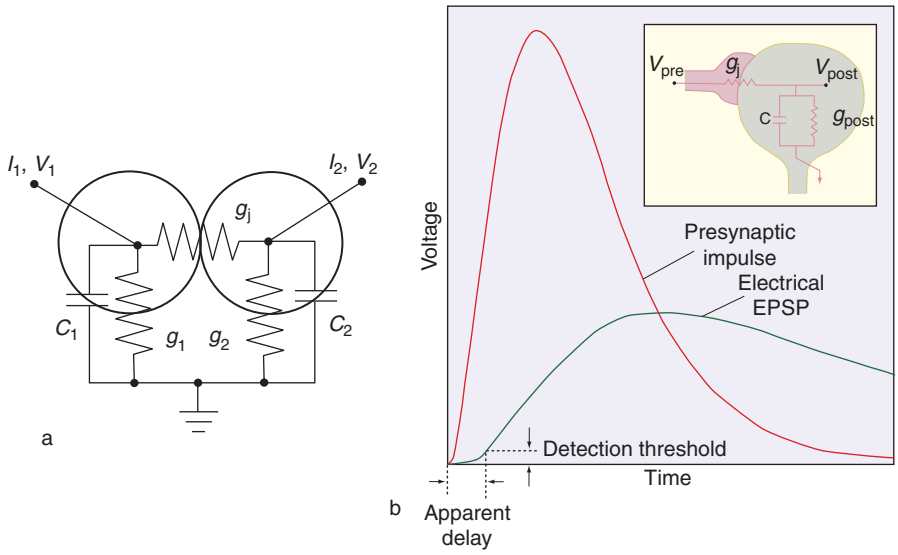

Figure 9 Equivalent circuit of a coupled pair of cells. (a) Isopotential cells: Each of the cell has a nonjunctional conductance (g_1, g_2) and a membrane capacitance (C_1, C_2). The cells are joined by a junctional conductance, g_j. Current is applied and voltage measured in each of the cells (I_1, I_2, V_1, V_2). Capacitance of the junctional membrane is negligible. (b) An axosomatic synapse. The postsynaptic response (electrical excitatory postsynaptic potential, EPSP, V_{pre}) is attenuated and delayed with respect to the presynaptic impulse, V_{pre}. The delay introduces a synaptic delay that can be comparable to delays at chemical synapses. g_j and g_{post}, junctional and postsynaptic conductances; *c*, postsynaptic capacitance. (b) From Bennett MVL (2000) Seeing is relieving: Electrical synapses between visualized neurons. *Nature Neuroscience* 3: 7–9.

coupling coefficient for a current applied in cell 1 is the ratio of the potential generated in cell 2 to that in cell 1: $k_{12} = V_2/V_1$. If the nonjunctional conductances of the cells are g_1 and g_2 and the junctional conductance is g_j, $k_{12} = g_j/(g_2 + g_j)$, and for currents applied in cell 2, $k_{21} = g_j/(g_1 + g_j)$. Note that k_{12} is independent of g_1, and k_{21} is independent of g_2. If k_{12} is small and $g_j < g_2$, $g_2 \approx I_2/V_2$ and $g_j \approx g_2 k_{12}$, and similarly for k_{21}.

Where the circuit of Figure 9(a) is valid, one can apply current first in one cell and then in the other and from the resulting voltages calculate the cell and junctional conductances precisely, but the equations are somewhat messy. A simpler method to determine g_j is

to voltage clamp the two cells to a common holding potential by means of a voltage clamp and then step the voltage in one cell to a new value. The current applied in the other cell to keep its voltage constant is equal and opposite to the current flowing through the junction at the given transjunctional voltage, $V_j = V_1 - V_2$ and $g_j = -I_2/V_j$ the minus sign is because the current into cell 2 is opposite in sign from the current flowing through g_j); g_1 is given by $(I_1 - I_2)/V_1$. Stepping cell 2 while holding V_1 constant measures g_j and g_2. (If both cells are stepped to the same potential, $V_j = 0$, $g_1 = I_1/V_1$, and $g_2 = I_2/V_2$.) Equality of g_j for V_j of either sign is an important validation of the equivalent circuit.

If the cells are not isopotential and even if they are part of a large coupled network, one can still calculate effective values for g_1, g_j, and g_2 for any two recording sites. However, the values will not correspond in any simple way to the conductances of the gap junctions coupling the cells, and coupling conductance, or g_c, is a more accurate term than g_j.

Voltage Gating of Gap Junctions

Gap junctions formed of most connexins exhibit varying degrees of sensitivity to V_j. In dual voltage clamp, a large step in cell 1 causes a current in cell 2 that decreases to a new steady state more or less exponentially, and this new value is smaller for larger V_j (Figure 10). (A virtue of dual voltage clamp is that the change in g_j is immediately apparent in the I_2 record.) In homotypic junctions, the decrease is in most cases symmetric for transjunctional voltages of either sign generated by hyperpolarization or depolarization in either cell. This finding indicates that the junctions are insensitive to the potential between the interior of the cells and the external medium. A few connexin-based junctions show a minor degree of dependence on the voltage between the cell interior and the extracellular medium ($V_{i\text{-}o}$ or V_m). The eponymous gap at the junction is accessible to the extracellular medium, so there should be little or no potential change along the gap, and the voltage across the channel walls should be the voltage between the channel lumen and the extracellular space at some distance from the junction. If the voltage differs in the two cells, the voltage is not uniform along the channels.

The steady state macroscopic currents for each polarity of V_j can be reasonably described by a Boltzmann relation:

$$g_j = [g_{jmax} - g_{jmin}]/(1 + \exp[A(V_j - V_0)]) + g_{jmin}$$

where g_{jmax} is the maximum value of g_j, which is approached at large values of V_j; A is a constant expressing voltage sensitivity; and V_0 is the voltage at which the conductance is halfway between g_{max} and g_{min}. The Boltzmann relation describes the equilibrium between two states, in this case high and low conductance, where the energy difference between them is a linear function of $V_j - V_0$ and $A = nz/(kT)$, where n is the equivalent number of electron charges, z, moving through the full voltage V_j, and kT has its usual meaning. The time course of changes in macroscopic currents in response to voltage steps

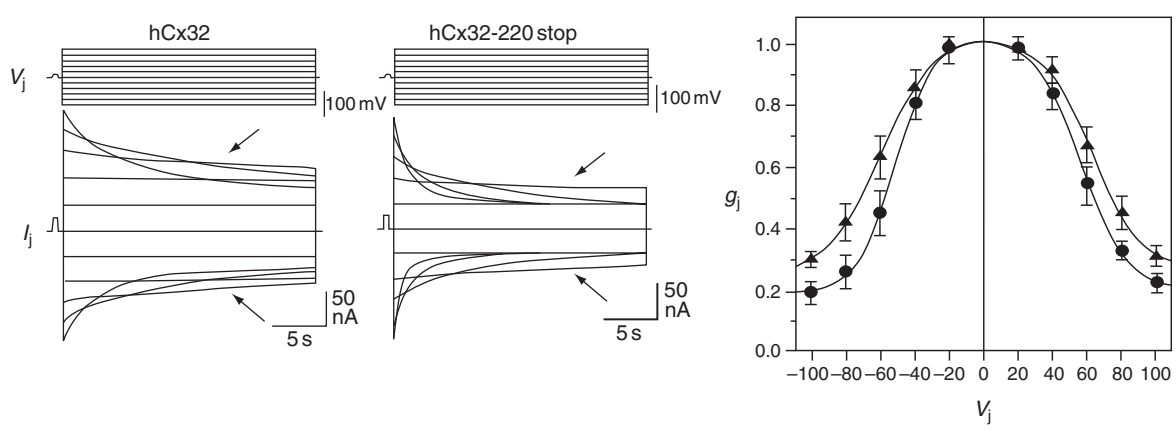

Figure 10 Voltage-dependent gating of gap junctions formed by human Cx32 wild-type and a mutant, Cx32-220 stop, one of many mutations that cause Charcot–Marie–Tooth disease, a peripheral neuropathy. The cRNAs were expressed in *Xenopus laevis* oocytes that were then paired and allowed to form junctions. The upper sets of traces show voltages applied in one cell (V_j) while the other cell was clamped at the holding potential. The lower sets of traces are the currents in the cell clamped at the holding potential and are equal and apposite to the transjunctional currents (I_j) associated with the voltage steps. The graph shows the steady state junctional conductance (g_j) as a function of transjunctional voltage (V_j). g_j decreases quite symmetrically for increasing V_j of either sign. The decrease approaches a plateau value accounted for by the residual conductance seen in voltage gating at the single-channel level (see Figure 11). The decline is steeper for the mutant (filled circles) than for wild-type (triangles). From Revilla A, Castro C, Barrio LC (1999) Molecular dissection of transjuctional voltage dependence in the connexins-32 and connexins-43 junction. *Biophysical Journal* 77: 1374–1383.

is approximately exponential, consistent with a Boltzmann process.

Single Channels and Gating

The conductances of single gap junction channels of known composition can be measured in connexin-deficient, high-resistance cells after transfection with specific connexins. In well-coupled cells, the number of open channels can be reduced with various blocking agents (see the section titled 'Pharmacological block of gap junctions') until single-channel activity is observable (Figure 11). Alternatively, separated cells can be apposed and the formation of junctions observed at the level of the first channels to open. The clear transitions between closed and quite stable open states, and often substates, allow identification of single-channel conductances. These conductances range between ~15 pS for Cx36 and ~200 pS for Cx37. Two modes of voltage gating, fast and slow, can be identified from single-channel recordings. Fast gating occurs on a millisecond timescale between the fully open state and a residual state with a conductance in the range of 10–30% of the maximal value. Slow gating requires 5 ms or more and occurs between the fully open state and the substate or the fully closed state. It appears to result from poorly resolved transitions between multiple unstable substates. For most connexins studied, both kinds of gate have a high open probability at $V_j = 0$, and open probability is decreased by increasing V_j. The more rapid rate constants of the fast gating process and its greater voltage sensitivity cause it to dominate the current–voltage relation. In this case g_j for either sign of V_j can be reasonably described by a Boltzmann relation for each of the oppositely oriented series hemichannels; one sign of V_j closes the gate in one hemichannel and has little effect on the gate in the other hemichannel. In this description, probability of being in the fully open state is the dependent variable; macroscopic g_j is given by the open probability P_o multiplied by the number of channels, N, and single-channel conductance, γ.

$$g_j = P_o N \gamma$$

The oppositely oriented gates in the two halves of a cell–cell channel are closed by opposite signs of V_j, which accounts for the symmetry of the I/V relation. Symmetry, however, precludes inference of the sign voltage sensitivity of the individual hemichannels from these data. Comparison of heterotypic and homotypic junctions or experimental induction of asymmetry by, for example, unilateral acidification, does allow determination of gating polarity, and it has been shown that different connexins can form hemichannels with opposite signs of gating polarity. Moreover, charged amino acids that constitute part of the voltage sensor have been identified, and charge substitutions in the N-terminal residues can reverse the gating polarity.

While only an approximation, and definitely not mechanistically justified for slow gating, Boltzmann relations are a convenient descriptor of the degree of voltage sensitivity of fast gating, as well as of slow gating when fast gating is absent. Voltage sensitivity in terms of charge can approach that of the channels

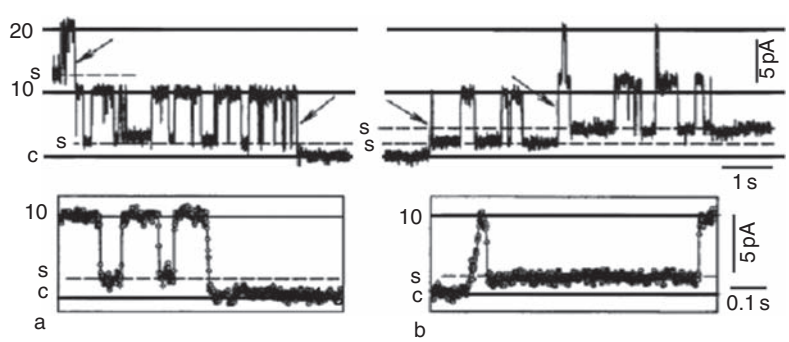

Figure 11 Single-channel currents showing fast gating between the fully open state and a substate and slow gating between the closed and fully open states. A well-coupled pair of fibroblasts expressing Cx43 were exposed to CO_2 equilibrated saline, which blocked the junctional conductance. (a) The upper record shows the last two channels closing in response to cytoplasmic acidification (arrows). The solid lines indicate the conductance level of the two open channels (10, 20). The dotted lines (s) indicate the substate conductance level, and c indicates the closed state. The boxed record below shows at higher gain and faster timescale two sojourns to the substate via fast transitions before the last slow transition to the closed state. The small circles on the records occur at 1 ms intervals. (b) The upper record shows the first two channels to open again after ~10 min washing in normal saline (arrows). The second channel opens fully while the first channel is in the substate (right arrow); subsequent fast transitions between open state and substate cannot be unambiguously assigned to one or the other channel. The boxed record below shows the initial slow transition from closed to open state, followed by a fast transition to the substate and a fast transition from substate to open state. From Bukauskas FF and Peracchia C (1997) Two distinct gating mechanisms in gap junction channels: CO_2-sensitive and voltage-sensitive. *Biophysical Journal* 72: 2137–2142.

of impulse-generating membrane, but the transition rate constants are long, in many cases hundreds of milliseconds to seconds. For a given connexin, the slow gating mechanism may have voltage sensitivity opposite that of the fast gating mechanism.

Heterotypic Channels, Gating Asymmetry, and Rectification

Hemichannels formed of different connexins can form heterotypic gap junctions that are quite symmetric or very asymmetric. Asymmetry of voltage gating results in junctional conductance that depends on polarity of transjunctional voltage, but the characteristic rates of change in conductance are slow, as described in the preceding section. At one extreme, the hemichannels have opposite polarity of voltage sensitivity referred to the voltage at their cytoplasmic ends, and both hemichannels are closed by the same sign of V_j. The macroscopic current–voltage relation of these junctions is very asymmetric.

Another factor in gating asymmetry is single hemichannel conductance. If a low-conductance hemichannel is in series with a high-conductance hemichannel, most of an applied voltage will be developed across the low-conductance (high-resistance) hemichannel. Its sensitivity to an applied V_j will be increased relative to its sensitivity in a homotypic junction. Similarly, the sensitivity of the high-conductance hemichannel will be reduced compared to that in its homotypic junction. In heterotypic junctions, small shifts in resting potential difference between two cells can cause large changes in junctional conductance and may be important in controlling chemical communication.

Asymmetric junctions can also show fast rectification (important in some rectifying synapses) in which the open channel rectifies, presumably on a microsecond timescale, and passes more or less current for a given transjunctional voltage, dependent on its sign. Rectifying electrical synapses are morphologically asymmetrical in that they occur between cells of different types and have a preferred direction of transmission. The polarity of rectification determines direction of impulse propagation and which elements are the pre- and postsynaptic ones. The constituent connexins have not yet been identified and may be encoded by different genes, or the same connexin on both sides may be differentially modified posttranslationally. The rectification is ascribed to differences in fixed charge in the constituent hemichannels. Fast rectification in inwardly rectifying K channels is due to channel block by internal polyamines. Similarly, differences in the composition of cytoplasm on the two sides of a gap junction could lead to rectification.

For most gap junctions, V_j dependence of g_j appears too weak and slow to have much physiological significance and appears unsuitable for processing of neuronal impulse activity. For some tissues, such as heart, with its long-duration action potentials, voltage insensitivity could be an advantage.

Dye Coupling and Permeability

A cell may be injected with (or induced to take up) a probe molecule such as a dye or a radioactive tracer, and contacting cells can then be assayed for appearance of the probe (**Figure 12**). Measurements of this kind indicate that the channel diameter is ~1.4 nm, and molecules of molecular weight of ~1 kDa can permeate (as can extended molecules of somewhat larger molecular weight, such as short interfering RNAs). The channel can be traversed by small ions, polypeptides, and nucleotides, including for some connexins cyclic adenosine monophosphate (cAMP), adenosine triphosphate (ATP), and other molecules of intermediary metabolism. This chemical communication may mediate metabolic cooperation in which one cell provides a molecule that another cannot synthesize or where random metabolic differences between cells are compensated for. Some but not all connexins form junctions with significant charge selectivity, and maximum size for permeation varies with connexin (**Figure 12**). Gap junctions are impermeable to large polypeptides, long nucleic acids, and, of course, subcellular organelles.

Dye coupling can be quantified in terms of junctional permeability, which is conveniently normalized to junctional conductance. For probes to which the surface membrane is impermeable, the dye coupling coefficient, defined as steady state ratio of concentration in a recipient cell to that in the source cell, is unity. However, in large networks of cells, the time to reach equilibrium may be longer than the usual measurement time, and postjunctional concentration may not reach its theoretical final value. Gap junctions do not rectify in respect to dye flux; in the absence of an asymmetric driving force, permeation is not polarized. However, dye coupling between a large cell and a small cell can be asymmetric at short times; the same dye flux in one or the other direction would produce a higher concentration in the smaller cell. Differences in nonjunctional membrane permeability could also lead to apparent asymmetry in dye coupling. A further issue in asymmetric systems is degree of binding, which can alter fluorescence of the bound molecules and concentration of the free dye in the cytoplasm.

A crude but convenient technique to measure dye coupling is scrape loading. As commonly used,

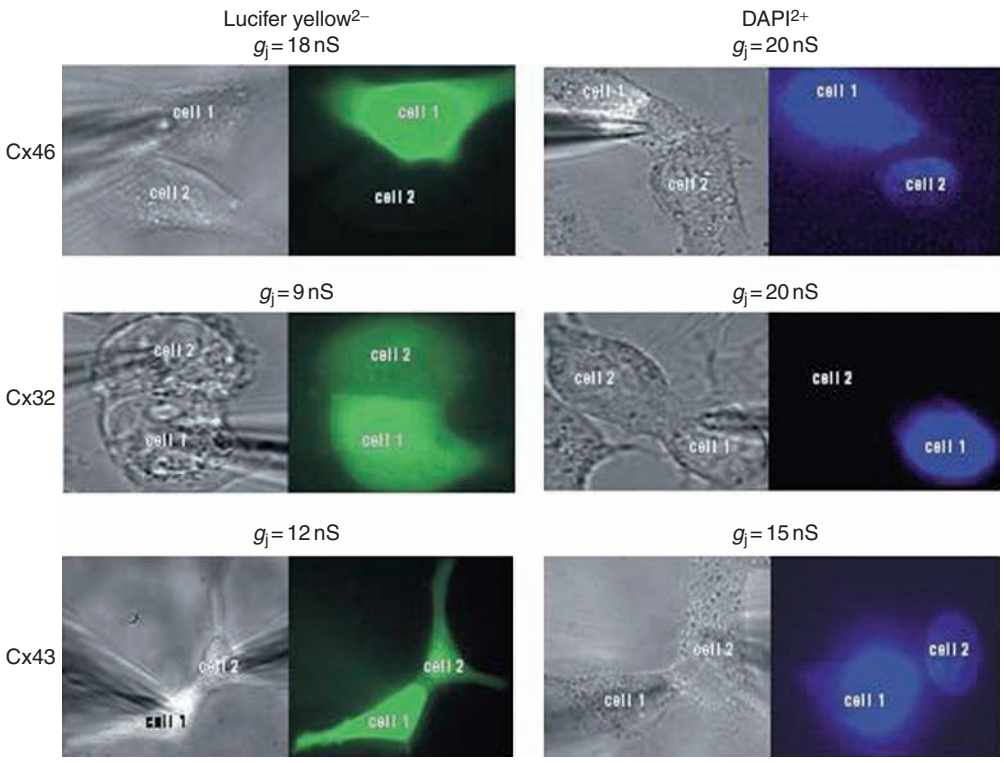

Figure 12 Cells joined by gap junctions show spread of dyes from one to another, that is, dye coupling. Dye coupling depends on the connexin, some of which are permeable to larger molecules than others are and some of which are charge selective. The figure shows Neuro-2a cells transfected with Cx46 (upper row), Cx32 (middle row), or Cx43 (lower row), phase micrographs on the left and fluorescence on the right in each pair of images. In each case, cell 1 was injected with the fluorescent dye Lucifer yellow (ion molecular mass 443 Da, two negative charges, left column) and DAPI (4′,6-diamidino-2-phenylindole, ion molecular mass 279, two positive charges, right column). Cx46 junctions are cation selective and are impermeant to Lucifer yellow but show some permeation by DAPI, which fluoresces when bound to nuclei. Cx32 junctions are anion selective and are permeable to Lucifer yellow but not to DAPI. Cx43 junctions readily transfer both Lucifer yellow and DAPI. Junctional conductance, g_j, is shown for each cell pair, and low level of coupling did not account for selectivity. From Trexler EB, Bukauskas FF, Kronengold J, Bargiello TA, Verselis VK (2000) The first extracellular 100p domain is a major determinant of charge selectivity in connexin 46 channels. *Biophysical Journal* 79: 3036–3051.

a confluent monolayer is cut with a razor blade in the presence of dye. Many of the cells can heal over, trapping dye in their cytoplasm. The dye bathing solution is then washed away, permitting observation of dye coupling to cells at some distance from the cut. Simultaneous application of a larger, junction-impermeant molecule with a different fluorescence allows identification of cells that were cut open, which have both markers, and the recipient cells, which have only the smaller, permeant species.

Formation and turnover In addition to the rapid gating mechanisms, specific treatments can induce cells to form gap junctions or to remove them. Most types of connexin hemichannel are assembled in membrane of the endoplasmic reticulum (ER) or post-ER compartments. Cx26 may be an exception and can insert into membranes posttranslationally. Vesicles with hemichannels are transported to the cell surface and inserted into the external membrane (Figure 13). Hemichannels then diffuse in the plane of the membrane until finding a hemichannel in an apposed membrane. Unapposed hemichannels in the surface in general are not open. The final steps are apposition or docking of intact hemichannels in the two membranes (Figure 7) and interaction with the extracellular loops to form the complete channel in a kind of ligand–receptor interaction in which each is a molecule of the same kind. When the first channel opens, the current through the nonjunctional membrane is unchanged (Figure 14). Thus, the hemichannels forming the first channel are not open to the extracellular space before they open to connect the two cells. (In some conditions, physiological or pathological, hemichannels open without interacting with a hemichannel in an apposed membrane (see the section titled 'Hemichannels that open').

Gap junctions can grow by accretion of new channels at their periphery because membranes are closely apposed in this region rather than exhibiting the wider gap of nonjunctional appositions. Targeted

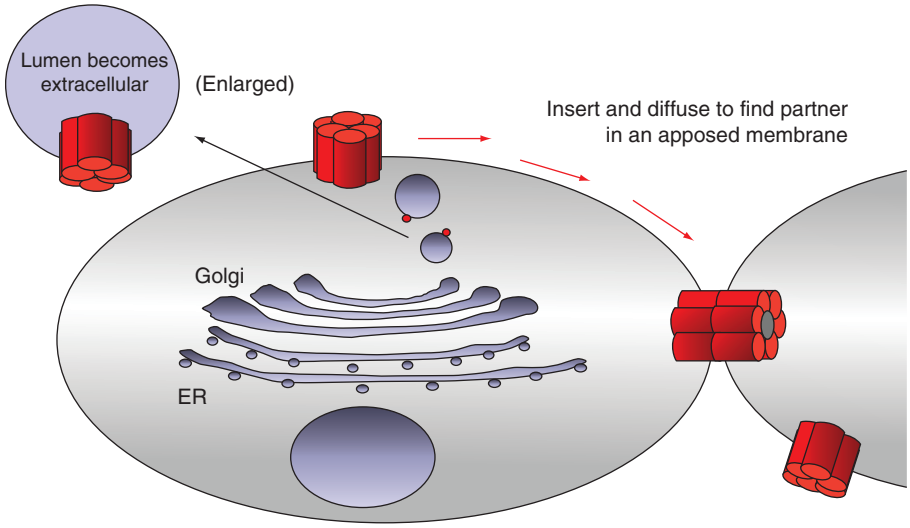

Figure 13 Connexins are synthesized in the endoplasmic reticulum (ER) and assembled into hemichannels in the ER or later compartment. They are then transported to the surface as integral membrane proteins in vesicles, which are inserted. The hemichannels diffuse in the surface to find a hemichannel in an apposed membrane and then join to form a cell–cell channel; each cell contributes a hemichannel to a cell–cell channel. Unpublished diagram by Jorge E. Contreras.

delivery of vesicles with hemichannels to junctional regions has been reported, but membrane insertion can also occur without directional transport. The mechanism by which the separated membranes are pushed close enough together for initial junction formation is unknown.

There is evidence that a large number of hemichannels must be clustered before the first cell–cell channel opens. Quantitation of junctional area and number of open channels determined from conductance indicates that only a small fraction of cell–cell channels are open, ~10%, a fraction that is likely to be connexin dependent. It is not clear whether a subpopulation of channels opens with a larger probability or whether probability is uniform and small. The atomic force data in Figure 5 suggest that all the channels are changing conformation during gating.

Normal connexin turnover time is a few hours in several cell types in culture and in rodent liver and heart. In marked contrast, gap junctions in the lens do not turn over in the lifetime of the organism after maturation of the lens fibers, in which their messenger RNA (mRNA) is lost. In turnover, junction removal apparently proceeds through internalization of the full thickness of the junction by one or the other cell (Figure 15) and may involve an entire junction or just a small region from the center. The hole in each cell that this mechanism requires presumably seals over with negligible leak. Internalization leads to the production of 'annular' gap junctions that can be seen in the subsurface cytoplasm; these structures are the internalized junctions with a little perijunctional membrane, and they contain a small amount of cytoplasm from the 'loser' cell. The transfer of cytoplasm from one to the other cell could be important in antigen transfer and presentation.

Eventually the internalized junctions are transported to lysosomes for degradation; there is no indication of recycling of hemichannels or connexins that have formed cell–cell channels. However, unapposed hemichannels in the surface membrane can enter the constitutive membrane recycling pathway and be internalized and reinserted.

Activation of kinases may accelerate junction formation or removal, depending on the tissue. In some systems, precursor connexin appear to be present, and junction formation can occur spontaneously even when ATP levels are reduced; in other systems, protein must be synthesized. Formation requires cell adhesion, and block of adhesion can block junction formation. Also, specific antibodies to the extracellular loops or peptides corresponding to the extracellular loops can prevent junction formation. If junction turnover is sufficiently rapid, these agents potentially provide a highly specific method of blocking junctional communication.

Degree of coupling can be regulated by controls of transcription and mRNA degradation as well as by those of formation and internalization. The cell biology of regulation of coupling is a field of active exploration.

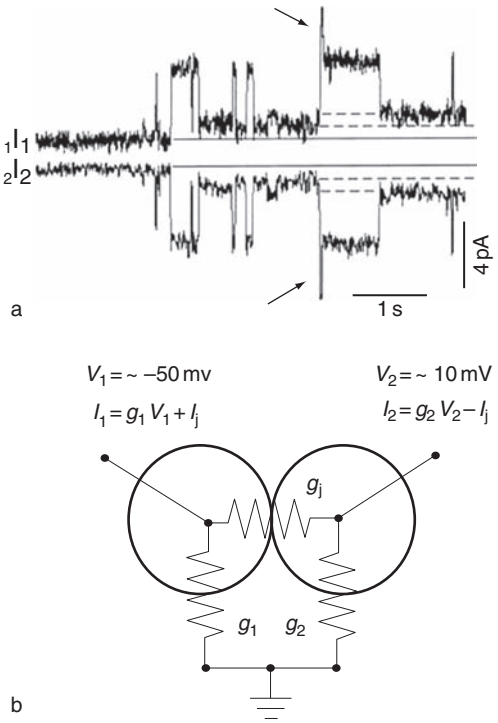

Figure 14 Hemichannels in cell–cell channels are closed before opening to connect the interiors of the apposed cells. (a) Opening of the first channels after washout of a agent that blocked gap junctions connecting two cardiac myocytes. (b) Equivalent circuit of the cells, where cell 1 and cell 2 each has a voltage, V_1 and V_2, a current, I_1 and I_2, and a conductance g_1 and g_2, and I_j is the current through g_j, the junctional conductance. The first cell–cell channel to open caused equal and opposite currents in the two cells. Thus, the nonjunctional currents in the two cells were unchanged, indicating that the hemichannels were not open to the extracellular medium prior to opening to couple the cells. At the arrow, a second channel opened. The dotted lines indicate the substate conductances of the two channels. Similar results are obtained with the first opening of newly formed channels when dissociated cells are pushed together while recording. (a) From Valiunas V, Bukauskas FF, and Weingart R (1997) Conductances and selective permeability of connexin43 gap junction channels examined in neonatal rat heart cells. *Circulation Research* 80: 708–719.

Pharmacological block of gap junctions An important tool in functional analysis is the use of specific blockers. Although numerous blockers are available, such as heptanol, octanol, halothane, carbenoxolone, glycyrrhetinic acid, flufenamic acid, anandamide, and oleamide, they are not very specific. Octanol and heptanol are widely used but affect many membrane channels. Carbenoxolone and glycyrrhetinic acid are more specific but may also affect excitability properties. Mefloquine is quite selective for Cx36 junctions, common between neurons. Where measured, conductance is reduced to zero by the slow gating mechanism. Low cytoplasmic pH (Figure 11) and high Ca^{2+} block most if not all gap junctions. The sensitivity to pH varies with the connexin and possibly the tissue. Domains involved in pH sensitivity are being identified. In some instances, conductance is close to maximum at resting pH, and small degrees of acidification markedly depress it. Here intracellular acidification is a plausible mechanism for cellular control of junctional conductance. In the few cases analyzed quantitatively, the sensitivity to Ca^{2+} appears to be less than that to H^+. Calcium ions would, however, act to uncouple from its neighbors a cell whose surface membrane became leaky. 'Healing over' of an injury to cardiac muscle is thought to involve a decrease in junctional conductance due to an effect of this kind, although scrape loading and retention of a gap junction-impermeant molecule indicates that ruptured cells can indeed heal over.

One can pretty reliably infer that if several gap junction blockers do not block a particular response, then the response is not mediated by gap junctions. At this time, it is much less reliable to infer that gap junctions mediate a response because gap junction blockers reduce or prevent it. So-called connexin- (or pannexin-) mimetic peptides with sequences corresponding to regions of the extracellular loops, as well as molecular approaches using siRNAs, should be quite specific in preventing junction formation and will be useful where gap junctions turn over rapidly enough for an effect of loss of function to be observed (see the section titled 'Formation and turnover'). Connexin knockout by recombinant techniques is and has been extremely useful but suffers from poor time resolution.

Phosphorylation Probably most connexins are phosphoproteins. There are consensus sites for phosphorylation in the cytoplasmic C-terminal domain that vary among connexins. Cx43 has at least 14. Phosphorylation may result in an increase or decrease in macroscopic conductance associated with altered single-channel conductance or open time. For example, gap junctions between hepatocytes are increased in conductance by cAMP-increasing agents; conversely, conductance of the junctions between horizontal cells in the retina is decreased by the same treatments. Nitric oxide mediates synaptic modulation of gap junction synapses at several sites. Because these changes are relatively rapid and are likely to involve changes in single-channel properties rather than formation or removal, they may be considered a form of gating. Phosphorylation can also affect insertion and removal.

What Do Gap Junctions Do for Neurons?

Transmission between neurons In the equivalent circuit of Figure 9, transmission of small potentials is governed by the steady-state relations described

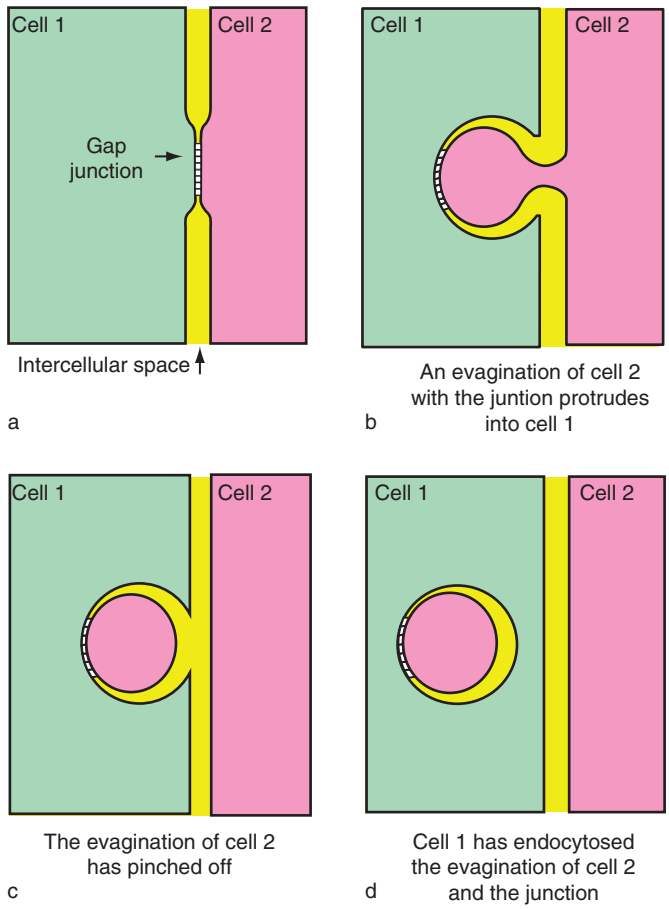

Figure 15 Removal of gap junction channels involves internalization of both junction membranes by one cell and transfer of a small volume of cytoplasm from one cell to the other. Successive stages of internalization of an entire junction are indicated, but small areas from the center of a junction can also be internalized. Between (b) and (c), the protrusion from cell 2 into cell 1 has pinched off from cell 2, and its membranes have resealed. Between (c) and (d), cell 1 has endocytosed the protrusion from cell 2 and resealed its membranes to include a small volume of extracellular material as well as the content of the protrusion.

above, and g_j values can be determined. Generally, voltages in neurons are too small or brief to engage voltage dependence of g_j, and the principle neuronal connexin, Cx36, is relatively insensitive to voltage. Because the membrane capacity of the nonjunctional membrane is much greater than that of the junctions, gap junction synapses act like a low pass filter (Figure 9). The postsynaptic potential is slowed because some of the junctional current goes to change the charge on the postsynaptic capacitance. Steady state coupling does not depend on capacity, but coupling is less for faster presynaptic potentials; peak amplitude is reduced, and the peak is delayed (Figures 8 and 9). One important consequence is that the measured synaptic delay can be as great at an electrical synapse as at a chemical synapse; although the electrical synapse starts to transmit without delay, time is required for the postsynaptic potential to reach a level at which it can be detected. The delays tend to be more nearly alike in warm-blooded animals because the delay at chemical synapses is less at higher temperatures whereas the delay at electrical synapses is not very temperature sensitive.

Slowing of the postsynaptic potential is generally an obvious feature of neural signaling across gap junctions. Neuronal action potentials are usually considerably briefer than the input time constant and so are relatively attenuated compared with slower potential changes such as synaptic potentials or applied hyperpolarization (Figures 8 and 9). In coupled cells, presynaptic impulses can produce postsynaptic responses that have been termed spikelets or fast prepotentials. Spikelets are briefer than synaptic potentials, and one hypothesis has been that they result from spikes generated in a dendrite that fail to invade the soma. While this may be true in some cases, dendritic spikes are often longer lasting than somatic spikes and spikelets. A factor that may contribute to the relatively short time course of an electrical postsynaptic potential (PSP)/spikelet is that the

presynaptic action potential may have an afterhyperpolarization that reverses the current flow across the electrical synapse and shortens the postsynaptic response.

In Cajal's neuron, transmission is polarized. At a gap junction synapse, if $g_1 \gg g_2$, then $k_{12} \gg k_{21}$ and coupling is much stronger in one direction that the other. Although inefficient, electrical coupling could be made arbitrarily unidirectional by adjusting the cell conductances. Unidirectional impulse propagation can also be achieved where the junctional membrane rectifies (see the section titled 'Heterotypic channels, gating asymmetry, and rectification,' above). At electrical axosomatic or axodendritic synapses, either electrical or chemical, impulses need not propagate from axon to postsynaptic cell on a one-to-one basis. Transmission is bidirectional, but coupling is weak, and the postsynaptic cell is excited only when there are enough presynaptic impulses to depolarize the cell to threshold. Impulses in the postsynaptic cell may or may not invade the electrically coupled presynaptic axons.

In coupled groups of neurons, a significant part of the input conductance of a single cell can be due to the conductances into the junctions it makes with other cells, and there generally are multiple pathways connecting any two cells. Significant voltage drops can be developed along dendrites connecting the cells. Furthermore, active conductances in the coupling pathways can influence the measured coupling and calculated value of g_c.

Synchronization Dendrodendritic gap junctions are the most common type of electrical synapse in mammals; these synapses generally synchronize electrical activity of cells of the same kind. In synchronization, current flows from a more positive cell to depolarize a less positive cell, and the first cell is itself made less depolarized by that current flow; thus, the synchronizing function is both excitatory and inhibitory, and pre- and postsynaptic do not have an obvious, fixed meaning where electrical signals can go in either direction. Examples of these coupled systems include numerous types of inhibitory interneurons, such as fast spiking cells, low threshold spiking cells in the neocortex, comparable cells in the hippocampus and striatum, stellate cells in cerebellum, thalamic reticular neurons, and various subclasses of interneurons expressing somatostatin, parvalbumin, or calbindin. Like neurons tend to be coupled, and a coupled network can extend for quite long distances, although electrical spread is measurable only between close neighbors. In these systems, synchronization of the inhibitory neurons is not very precise, but the circuitry is made simpler by allowing reciprocal connections that transmit not only impulse activity but subthreshold potentials of either sign. Moreover, the cells can excite each other through their gap junctions and inhibit downstream cells as well as each other by chemical synapses, a problematic arrangement for chemical transmission.

Synchronization can be increased by mutual inhibition mediated chemically. In a rhythmically discharging pool of neurons that are electrically coupled, firing of one cell tends to fire its neighbors, but depolarization is immediately followed by inhibition, restricting the time at which those cells can fire and also preventing delayed firing so that they are more likely to fire the next time the first cell fires. Generally in these pools of neurons, the influence of one cell on a coupled neighbor is small, but if many cells are active simultaneously, the excitatory effect on coupled cells can be very strong.

A number of classes of synchronously active coupled cells have axons that project out of the brain region in which they are coupled. Mitral cells of the olfactory bulb project to olfactory cortex and are coupled in their dendritic regions in the olfactory glomeruli. Olfactory neurons expressing the same receptor project to the same glomerulus, and the innervating mitral cells tend to fire synchronously. Cells of the inferior olive, which give rise to climbing fibers exciting cerebellar Purkinje cells, are coupled. The olive is a relatively large structure and tends to be subdivided into regions that are synchronously firing; the subdivision may be under control of inhibitory inputs to the dendrites that decrease electrical current spread along the dendrites and decouple groups of cells from each other. Neurons of the striatum are coupled inhibitory neurons that project to other brain regions. Neurons of the hypothalamus that release neuropeptides in the median eminence can be coupled with like neurons to facilitate hormone release. Neurons of the locus coeruleus are coupled, fire more or less synchronously, and project to much of the cortex. The suprachiasmatic neurons are coupled and project to the paraventricular nucleus.

Many neurons in the retina are coupled, including type II amacrine cells and horizontal cells, where coupling can be both within and between classes. Coupling mediates lateral spread of excitation (rather than synchronization) that then causes recurrent inhibition. Electrical synapses may also involve other types of retinal cell. Rod bipolar cells excite type II amacrine cells that are electrically coupled to each other and to cone bipolar cells that then excite ganglion cells.

Pools of neurons connected by gap junctions differ markedly in the degree of coupling. At one extreme, an impulse initiated in one cell propagates to all the

other cells (not true of mammalian inhibitory interneurons). In less well coupled groups of cells, the synchrony of firing may be rather poor. However, even weak coupling can promote precise synchronization if other inputs are driving the cells to fire at about the same frequency.

Synchronization mediated by electrical coupling has been characterized in a number of lower vertebrate and invertebrate species. Examples include neurons controlling electric organ discharge in fishes, which in most species is highly synchronous; motor systems in which speed and synchronization are important, such as those mediating saccadic eye movements, and sonic muscles, in which the fundamental frequency of the emitted sound is the frequency of contraction. Other systems are less well synchronized; examples include neurons controlling cardiac ganglia in arthropods and supramedullary neurons, a group of autonomic effectors found in teleosts.

In the highly synchronized systems, the short delay possible at electrical synapses is required for the observed degree of synchronization. In other cases, synchronization could be mediated by drive from a higher center with fewer cells or by feedback mediated by chemical synapses.

When a pool of neurons has fired synchronously, additional adaptations are necessary to assure that the effector cells are active synchronously, if distances from the 'command center' to effectors are sufficiently different to introduce asynchrony if conduction is 'as the crow flies' and of fixed velocity. One solution is to have the pathway to the more distant effectors conduct more rapidly. Others are to have more devious pathways to the nearer effectors or to introduce compensatory delays along the pathways.

Escape! Most animals have a fast escape response to escape from predators and other dangers. An example that has been important in the study of synaptic transmission is the Mauthner cell system found in teleost fishes and some urodele amphibians. There are two Mauthner cells, one on either side of the medulla. Their axons decussate and descend the spinal cord to excite the axial motor neurons on the side contralateral to the Mauthner cell body, and firing of one Mauthner cell mediates a tail flip escape response away from the stimulated side. Large myelinated fibers from the sacculus, a hearing organ, form 'club endings' on the lateral dendrite of the Mauthner cells that transmit both electrically and chemically; the electrical component is of shorter latency and presumably reduces the latency of the escape response. Excitation of the axial motor neurons by the Mauthner axons is chemically mediated, as is transmission at the peripheral neuromuscular junction. The Mauthner cells also inhibit each other through interneurons that act electrically (see the section titled 'Electrical inhibitory synapse') and then chemically so that firing of one cell rapidly inhibits the other; clearly, it would be counterproductive for swimming if muscles on both sides contracted simultaneously. (Each Mauthner axon also inhibits the contralateral motor neurons through crossed inhibitory interneurons on its side of the cord. Excitation of the inhibitory neurons is electrical so that in a given motor neuron, the chemical excitation from one Mauthner axon and the crossed chemical inhibition from the other have about the same latency, preventing simultaneous contraction of muscles on the two sides, should the Mauthner cells fire at the same time. In contrast, there is a bilateral component of the Mauthner cell escape reaction which closes the mouth and adducts the pectoral fins. Interneurons mediating these actions are excited by both Mauthner cells at chemical synapses and then transmit electrically to the motor neurons.)

The lateral giant axons controlling the tail flip of many crustaceans, including lobsters and crayfish, are made up of multiple segments, one neuron per segment coupled together end to end to make in effect a single giant axon. (The gap junctions connecting the segments are presumably pannexin based.) The lateral giant axons conduct impulses at a comparable speed to the medial giant axons, which are true single axons arising from a single pair of cell bodies in the supraesophageal ganglia. One can appeal to developmental factors to account for the structural difference. The medial giants are excited by synapses in the head region, and the lateral giants can be excited in many of the abdominal ganglia. The two systems give somewhat different escape responses. The giant axons of many annelids are segmental and mediate the fast withdrawal. The septa between segments contain gap junctions. The gap junctions in all the invertebrates are presumably pannexin based; the pannexin mediating the escape jump reflex in *Drosophila* is known.

How important is electrical transmission in these systems? Synaptic delays can be less at electrical synapses, particularly at lower temperatures. In an escape response, shorter latency is better, which may account for the high incidence of electrical transmission in escape systems of poikilotherms. There remains an obligatory chemical synapse to the muscle fibers, and one appeals to as yet undefined developmental constraints to account for the lack of electrical neuromuscular junctions. The time for muscle contraction is greater than the delay at a chemical synapse, but a small percentage increase in escape probability could provide powerful selection pressure. However, as an organism gets bigger, conduction times get

longer, the time required for movement becomes greater, and the saving at an electrical synapse a smaller fraction of the response time.

Electrical synapses are is found in the vestibular nuclei and mesencephalic nucleus of the trigeminal nerve of rodents, where it may reduce response latency. In mammals, where transmission in motor systems has been characterized, there are few instances of electrical transmission. However, anatomical studies of Cx36 distribution indicate that it may be much more common that realized hitherto.

Spatial and temporal summation If many axons form gap junction synapses on a postsynaptic cell, the postsynaptic response is graded with the number of presynaptic axons activated, and firing of the postsynaptic cell depends on the summation of inputs from the presynaptic axons, just as at a cell receiving many excitatory chemical synapses. In the classic sense, the system shows spatial summation of the excitatory synaptic potentials. A second classical form of short-term interaction is temporal summation. Electrical as well as chemical PSPs can change the charge on the postsynaptic capacity, which requires time to decay back to the resting value. Electrical synapses can show temporal summation, when the postsynaptic time constant is comparable to or greater than the interval between presynaptic impulses. Furthermore, presynaptic action potentials can be prolonged or decreased in amplitude as a result of repetitive firing, both of which can modulate both electrical and chemical postsynaptic potentials.

Other forms of plasticity Coupling between neurons can be regulated in a number of ways. Inhibitory synapses along a dendrodendritic path coupling cell bodies can short-circuit coupling between them and allow them to fire independently in response to excitatory inputs closer to their cell bodies. This form of regulation has been proposed for the mammalian inferior olive and for buccal motor neurons in a mollusk.

Gap junctions may be regulated by second messengers, as in retinal horizontal cells and amacrine cells. Transmitters, such as dopamine, acting through G-protein-coupled receptors, lead to increase in cytoplasmic cAMP, which is followed by decrease in junctional conductance. It remains to be determined whether phosphorylation sites on the connexins are involved. Long-term potentiation and depression of both electrical and chemical components have been demonstrated at the club endings on the Mauthner cell. As more electrical synapses are examined, more instances of this class of regulation are likely to appear.

Coupling of cortical neurons is very widespread and promiscuous in early development but then decreases to the less frequent and more precisely targeted level seen in the adult. The roles of this early coupling are obscure. It may mediate synchronous firing that helps guide axons to their appropriate targets, or it may mediate chemical communication for differentiation. Mice with targeted deletion of any one of a number of connexins expressed in brain still exhibit relatively normal central nervous system (CNS) structure and function, suggesting either redundancy or relative unimportance.

Electrical Inhibitory Synapses

As noted, gap junction synapses mediating synchronization are both excitatory and inhibitory. Furthermore the postsynaptic potential generated by a presynaptic action potential with an afterhyperpolarization is biphasic, depolarization followed by hyperpolarization. Depending on the cell, the mean potential can be hyperpolarizing, and the result of a train of action potentials is an initial excitation followed by a prolonged inhibition with small superimposed depolarizations associated with each presynaptic spike.

A completely different kind of electrical inhibitory synapse occurs at the initial segment of the Mauthner cell. Interneuron axons converge on a structure termed the axon cap, a specialized region around the initial segment, which is where the impulse arises. Presynaptic action potentials in the inhibitory neurons generate a positivity in the extracellular space of the axon cap, the time course of which is like that of an action potential (Figure 16). It is likely that the axonal membrane acts simply as a fixed battery with a series resistance. The axons may go on to form chemical inhibitory synapses on the nearby soma of the Mauthner cell. The positivity, or 'extrinsic hyperpolarizing potential,' causes current to flow inward through the initial segment membrane (hyperpolarizing it) into the cell soma (and axon), along the cell cytoplasm and out through more-distant membrane (causing depolarization of this membrane). Since the conductance of this latter pathway is large compared with that of the initial segment membrane, most of the external potential is developed across the initial segment membrane, and impulse initiation is inhibited.

When the Mauthner cell is active, it generates a large negativity in the axon cap region. This current tends to hyperpolarize the excitable membrane in the inhibitory interneurons and inhibit them. The current flows and equivalent circuit during inhibitory neuron and Mauthner cell activity are diagrammed in Figure 16.

The electrical inhibition is a specialization for speed. Electrical inhibition reduces the latency of onset of inhibition by one chemical synaptic delay. The electrical component is brief, but it is followed by long-lasting chemical inhibition. As noted (see the

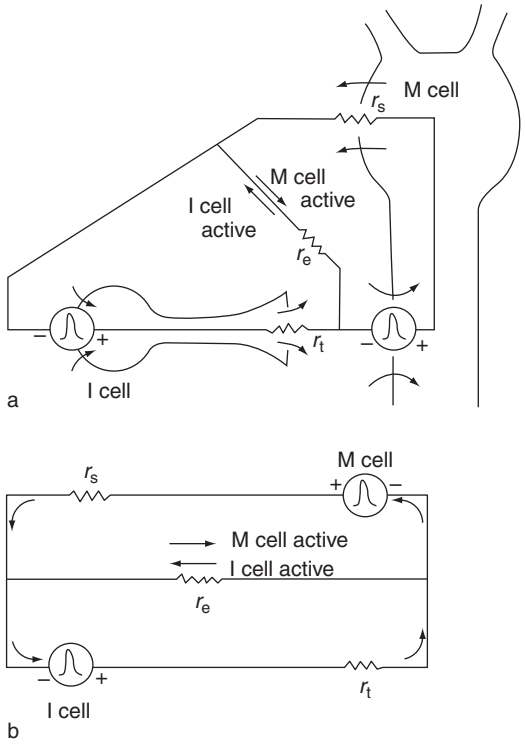

Figure 16 Diagram of an electrical inhibitory synapse at the initial segment of the Mauthner (M) cell. (a) Impulse activity of the inhibitory interneurons (I cell) produces outward current through their axon terminals (r_t) that divides to (1) flow across the external resistance (r_e) of the axon cap or (2) enter the M cell through its impulse initiation site, hyperpolarizing it and inhibiting M-cell firing. The current entering the M cell leaves through the resistance of dendritic (r_s) and axonal membrane, but the resulting depolarization does not significantly affect impulse initiation. The circles with impulses drawn in them represent the impulse-generating membranes. Conversely, impulse activity of the M cell causes current that hyperpolarizes and inhibits the impulse-generating membrane of the I cell. Reciprocity of the inhibitory interaction is made clearer by the redrawn equivalent circuit in (b). From Bennett MVL (1977) Electrical transmission: A functional analysis and comparison with chemical transmission. In: Kandel ER (ed.) *Cellular Biology of Neurons*, pp. 357–416. Baltimore: Williams and Wilkins.

section titled 'Escape!') a requirement for an effective escape response is that the Mauthner cells do not fire simultaneously and each inhibits the other. There is evidence for similar electrical inhibition at the axon hillock of mammalian cerebellar Purkinje cells.

Gap Junctions and Genetic Disease

Given the size of the connexin gene family (and Murphy's law that anything that can go wrong, will), one would expect there to be diseases caused by mutations in these genes. The first discovery was that mutations or deletions of Cx32 cause X-linked Charcot–Marie–Tooth disease, a peripheral neuropathy. Only after positional cloning demonstrated the involvement of Cx32 was it shown that this connexin is expressed by Schwann cells that form myelin. The effects of the mutations on Cx32 junctions may be subtle, such as minor changes in voltage sensitivity or permeability, or marked, leading to perturbed trafficking or complete loss of the protein. Light microscopic immunocytochemistry shows Cx32 at perinodes and Schmidt-Lanterman incisures, but their arrangement has not been clearly resolved in thin section or freeze-fracture electron microscopy. The gap junctions are apparently reflexive, that is, between different parts of the same cell, and are presumed to facilitate the interchange of small molecules between the periaxonal cytoplasm and the outermost part of the Schwann cell. The question remains as to what is communicated from outside to inside or vice versa. And it is unclear why loss of this communication, which has no apparent effect on myelin formation during growth and development, leads eventually to demyelination and degeneration of nerve fibers. Furthermore, there is no obvious pathology in other tissues expressing Cx32, either because Cx32-mediated communication is unimportant in these tissues or because another connexin takes over the function of Cx32. Some minor effects are seen in central nerve conduction and in the Cx32 knockout mouse in the liver and pancreas.

Mutations in Cx26 are the most common cause of nonsyndromic deafness (deafness without other defects). Cx26 is expressed in the cochlear epithelium, not including the hair cells, and the pathogenesis may involve impaired communication between supporting cells. In any case, degeneration of the hair cells occurs early in the development of hearing but after the major structures of the inner ear have developed. As for Cx32, few other symptoms are reported, although a minority of patients exhibit defects in skin. Mutations of Cx31, which is expressed in skin, cause erythrokeratodermia, a thickening and reddening of the skin. Cx30 mutations cause deafness or skin disease; differential effects of the mutations at different points in the life cycle of the connexin may account for these differences. Mutations in the lens connexins, Cx46 and Cx50, cause congenital cataract. Mutations in Cx43 cause oculodentodigital dysplasia, a disease involving multiple defects, including in the nervous system. Cx36 is implicated in juvenile myoclonic epilepsy, but mechanisms are obscure.

Targeted gene disruption (knockout) in mice reveals other potential genetic diseases involving mutations in connexins. Knockout of Cx32 is quite benign, but the double knockout of Cx32 and Cx47 causes loss of central myelin and early death. Knockout of Cx43 is lethal at birth because of cardiac malformation, and there are subtle abnormalities in cortical layering.

Knockout of Cx26 is embryonic lethal because of defects in the trophoblast, but conditional knockouts using various strategies to permit knockout of a gene in a specific tissue are becoming important tools for analysis of gene function that will be relevant to human disease.

Evolution

Vertebrate connexins have been grouped into alpha and beta (or group II and group I, respectively). Assignment into further divisions, gamma and delta, may be justifiable, but deep branches of the evolutionary tree are less reliable at this time. The zebra fish, *Danio rerio*, has 37 connexins, almost twice as many as mammals have, probably because teleosts have undergone a whole genome duplication. Connexins are also found in ascidians. The number of connexins in the common ancestor of ascidians and vertebrates may be no more than one. Most vertebrate connexins have a single coding exon, and alternative splicing is limited to noncoding regions. Cx36 is an exception and has two coding exons. Ascidian connexins have several coding exons, as do pannexins/innexins.

The ancestral connexin is uncertain. Occurrence in hemichordates has not been investigated. The sea urchin genome has been sequenced and is reported to lack both connexins and pannexins. The anatomical basis of coupling in this group, reported in several studies, is thus unknown. Although there were early inferences of connexins in Cnideria based on antibody labeling and Western blotting, genomic sequencing has yet to demonstrate a connexin. Their gap junctions, coupling, and electrical synapses are likely to be pannexin/innexins based.

What is the survival value of having multiple gap junction-forming proteins? Functional differences between connexins in gating, permeability, and trafficking have been demonstrated and provide one explanation for the observed diversity. There are the obvious functional differences in permeability, gating, and posttranscriptional regulation of formation and degradation. Differences in transcriptional and translational control may be more important than the functional differences in the connexins themselves.

Convergence between Connexins and Pannexins

Although differences in the gene sequences indicate that connexin-based and pannexin-based gap junctions are separate evolutionary adaptations, there is a remarkable degree of functional convergence, including block by many of the same pharmacological agents, by low cytoplasmic pH, and by high cytoplasmic Ca^{2+}. Fast V_j gating to a substate and slow gating to the completely closed state are also found in both classes, although dependence on membrane potential ($V_{i\text{-}o}$) is more pronounced in at least one pannexin junction. In pannexin junctions, the channel diameter is a little bigger, the gap is a little wider, and the number of channels per unit area is a little lower. Some pannexin junctions under specific uncoupling conditions may split, with each membrane taking its own hemichannels, which can then be reused. Pannexins do mediate coupling when expressed in *Xenopus* oocytes, but it is not yet clear whether they do so in mammals. Like connexins, pannexins form hemichannels that can open in the surface membrane without contact with a hemichannel in an apposed membrane.

Hemichannels That Open

The life history of a connexin molecule includes some time in the membrane as a hemichannel without a partner (**Figure 13**). Because of the predicted high conductance and lack of permselectivity, it was thought that these hemichannels should not open because they would stress the cells by letting in excess Ca^{2+} and Na^{+} and permitting loss of K^{+} and metabolites. Furthermore, measurements during junction formation indicated that channels were not open before connecting with hemichannels in an apposed membrane (see **Figure 14**). This conventional wisdom was confirmed when expression of Cx46, which is expressed only in lens and forms hemichannels when expressed in *Xenopus* oocytes, killed the oocytes unless Ca^{2+} in the bathing solution was kept at relatively high levels to keep the hemichannels closed. Now it is clear that hemichannels formed of a number of connexins can open to the extracellular space and can mediate both physiological and pathological responses. The open probability is generally very low, and cells can release significant amounts of signaling molecules, such as ATP, without impact on viability. Conversely, under pathological conditions hemichannels can speed death of the cells expressing them. Also, the molecules they release, such as glutamate from astrocytes, can have a toxic action on neighboring cells, in particular neurons.

A number of criteria indicate that connexins can form hemichannels. Different treatments, such as bathing in low Ca^{2+} solution, cause uptake of dyes to which gap junctions are permeable but normal cell membranes are not, and this uptake requires expression of connexins. The uptake is also blocked by gap junction blockers. Furthermore, single-channel activity can be recorded that has the predicted conductance of about twice that of the cell–cell channel formed by the same connexin. Moreover, their gating properties correspond to those of the cell–cell channels.

Pannexins in mammals may form only hemichannels, and one mode of opening is through activation of purinergic receptors. The linkage between purinergic activation and pannexin hemichannel opening is unclear. As with connexins, pannexin hemichannel opening may result in release of signaling molecules, such as ATP.

See also: Gap Junctions and Neuronal Oscillations; Gap Junction Abnormalities and Disorders of the Nervous System; Gap Junction Communication; Gap Junctions: Metabolic Exchanges.

Further Reading

Bennett MVL (1977) Electrical transmission: A functional analysis and comparison with chemical transmission. In: Kandel ER (ed.) *Cellular Biology of Neurons*, pp. 357–416. Baltimore: Williams and Wilkins.

Bennett MVL (1997) Gap junctions as electrical synapses. *Journal of Neurocytology* 26: 349–366.

Bennett MVL, Barrio LC, Bargiello TA, Spray DC, Hertzberg E, and Saez JC (1991) Gap junctions: New tools, new answers, new questions. *Neuron* 6: 305–320.

Bennett MVL (2000) Seeing is relieving: Electrical synapses between visualized neurons. *Nature Neuroscience* 3: 7–9.

Bennett MVL, Contreras JE, Bukauskas FF, and Saez JC (2003) New roles for astrocytes: Gap junction hemichannels have something to communicate. *Trends in Neuroscience* 26: 610–617.

Bennett MVL and Zukin RS (2004) Electrical coupling and neuronal synchronization in the mammalian brain. *Neuron* 41: 495–511.

Bukauskas FF and Peracchia C (1997) Two distinct gating mechanisms in gap junction channels: CO_2-sensitive and voltage-sensitive. *Biophysical Journal* 72: 2137–2142.

Connors BW and Long MA (2004) Electrical synapses in the mammalian brain. *Annual Review of Neuroscience* 27: 393–418.

Cruciani V and Mikalsen SO (2007) Evolutionary selection pressure and family relationships among connexin genes. *Biological Chemistry* 388: 253–264.

Eastman SD, Chen TH, Falk MM, Mendelson TC, and Iovine MK (2006) Phylogenetic analysis of three complete gap junction gene families reveals lineage-specific duplications and highly supported gene classes. *Genomics* 87: 265–274.

Gaietta G, Deerinck TJ, Adams SR, et al. (2002) Multicolor and electron microscopic imaging of connexin trafficking. *Science* 296: 503–507.

Harris AL (2002) Emerging issues of connexin channels: Biophysics fills the gap. *Quarterly Reviews of Biophysics* 34: 325–472. Erratum in: *Quarterly Reviews of Biophysics* (2002) 35:109.

Harris AL (2007) Connexin channel permeability to cytoplasmic molecules. *Progress in Biophysics and Molecular Biology* 94: 120–143.

Meier C and Dermietzel R (2006) Electrical synapses: Gap junctions in the brain. *Results and Problems in Cell Differentiation* 43: 99–128.

Perkins GA, Goodenough DA, and Sosinsky GE (1998) Formation of the gap junction intercellular channel requires a 30 degree rotation for interdigitating two apposing connexons. *Journal of Molecular Biology* 277: 171–177.

Revilla A, Castro C, and Barrio LC (1999) Molecular dissection of transjuctional voltage dependence in the connexins-32 and connexins-43 junction. *Biophysical Journal* 77: 1374–1383.

Saez JC, Retamal MA, Basilio D, Bukauskas FF, and Bennett MVL (2005) Connexin-based gap junction hemichannels: Gating mechanisms. *Biochimica et Biophysica Acta* 1711: 215–224.

Sohl G, Maxeiner S, and Willecke K (2005) Expression and functions of neuronal gap junctions. *Nature Reviews Neuroscience* 6: 191–200.

Theis M, Sohl G, Eiberger J, and Willecke K (2005) Emerging complexities in identity and function of glial connexins. *Trends in Neuroscience* 28: 188–195.

Trexler EB, Bukauskas FF, Kronengold J, Bargiello TA, and Verselis VK (2000) The first extracellular 100p domain is a major determinant of charge selectivity in connexin 46 channels. *Biophysical Journal* 79: 3036–3051.

Valiunas V, Bukauskas FF, and Weingart R (1997) Conductances and selective permeability of connexin43 gap junction channels examined in neonatal rat heart cells. *Circulation Research* 80: 708–719.

Venance L, Rozov A, Blatow M, et al. (2000) Connexin expression in electrically coupled postnatal rat brain neurons. *Proceedings of the National Academy of Sciences of the United States of America* 97: 10260–10265.

Yen MR and Saier MH Jr (2007) Gap junctional proteins of animals: The innexin/pannexin superfamily. *Progress in Biophysics and Molecular Biology* 94: 5–14.

Relevant Website

http://www.scholarpedia.org – Eugene M. Izhikevich, Scholarpedia.

Gap Junctions and Hemichannels in Glia

Z Ye and B R Ransom, University of Washington School of Medicine, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Gap junctions (GJ) are unique large channels that connect the cytoplasm of two adjacent cells. Half GJs, called connexons or hemichannels (HC), expressed by individual cells may also be functionally active, and when open would connect a cell's cytoplasm to extracellular space. Gap junctions mediate a primitive and important form of intercellular communication that is seen across phyla. Discovered in 1952 in heart muscle cells, GJs are expressed in most mammalian cells; mature skeletal muscle cells, spermatozoa and erythrocytes are the only exceptions. Gap junction mediated coupling between glial cells was first noted in the leech central nervous system (CNS) by Stephen Kuffler and colleagues in 1964. In the mammalian CNS, connexins, the proteins that form GJs and HCs, are most abundantly expressed by glial cells. All subtypes of glia express connexins and GJs, with the possible exception of the newly defined NG2 cell. The characteristics and possible functions of GJs and HCs in glial cells are the focus of this review. First, however, some important general principles about these channels and the proteins that make them will be discussed.

Connexins Form Gap Junctions

Gap junctions (GJs) are so named based on their appearance in electron micrographs. At GJs, the width of the extracellular space is narrowed to a small 'gap' of ~2.5 nm (normal width ~20 nm; Figure 1). In mammals, GJs are formed from connexins, a protein family with ancient origins and containing >20 members (Figures 1 and 2). The molecular mass of cloned connexins ranges from 25 to 62 kDa, and they are named on this basis (e.g., the 43 kDa connexin is called Cx43; Figure 2). Most cells express at least two connexin subtypes, although one is likely to predominate. Invertebrates form abundant GJs but do not express connexins. Instead, their GJs are composed of proteins called innexins that are analogous, but not strongly homologous, to connexins. Curiously, innexin-like proteins called pannexins have been detected in mammals and may mediate coupling between hippocampal pyramidal neurons. Pannexins, in contrast to connexins, however, are more likely to form hemichannels (HCs) than GJs.

Each GJ is composed of two HCs, one embedded in each of the closely opposed membranes (Figures 1 and 3). Each HC consists of six connexin molecules, and these can be aggregates of six identical connexins or a mixture of different connexins. Functional GJs can form between HCs composed of different connexins. For example, astrocytes and oligodendrocytes form GJs composed of Cx43 and Cx47, respectively. Connexins and related proteins have four transmembrane domains, and the N-terminal (NT) and C-terminal (CT) segments are both intracellular, resulting in two extracellular loops (ELs) and one cytoplasmic loop (CL). When two HCs 'dock' to make a GJ, the extracellular loops mesh to form a structural 'tight seal,' guaranteeing that the channel will not leak to the extracellular space. Covalent bonds are not involved in the formation of this tight seal. Curiously, connexin proteins have a very high turnover rate; the average half-life of these proteins is several hours.

Macroscopic GJs are aggregates of many, usually hundreds, of tightly packed GJ channels. In the mammalian central nervous system (CNS), GJs are most numerous between glial cells, especially between astrocytes. They are also seen between certain populations of neurons, including cortical interneurons and certain brain stem neurons. More detail about the expression of connexins and GJs in glial cells is given in the following sections.

Biophysical Properties of GJs

The biophysical properties of GJs can be partly understood based on their structure. Open GJs have a very large limiting pore diameter of 1.5 nm and are permeable to a much wider range of ions and molecules than are typical ion channels (typical pore diameter, ~3–5 nm). Consequently, GJ pores can admit molecules up to a molecular mass of about 1 kDa, largely independent of charge. Coupled cells, therefore, are able to exchange a wide range of biologically important molecules, including cyclic nucleotides, small molecules of intermediate metabolism, vitamins, and inorganic ions. Obviously, strongly coupled cells will have a tendency to have similar concentrations of those cytoplasmic constituents that easily pass through GJs. This principle has been conclusively demonstrated for the concentration of intracellular Na^+ in astrocytes. GJs composed of different connexins, however, vary considerably in what can pass through them. For example, GJ channels composed of Cx32 and Cx43 show marked differences in permeability to adenosine, which is 12-fold more permeable through Cx32 versus Cx43 channels.

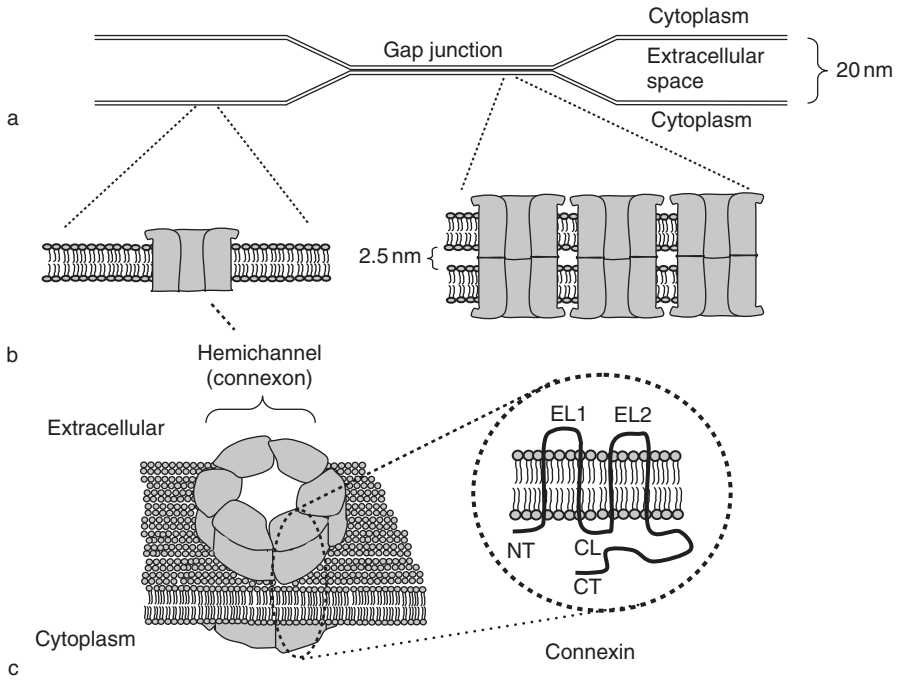

Figure 1 Gap junction and hemichannel structure. (a) Schematic drawing of a gap junction as visualized by electron microscopy. The distance between neighboring cells narrows to about 2 nm in the junction. (b) Gap junctions are formed by two aligned hemichannels or connexons, one in each adjacent cell. Hemichannels are unopposed connexons. This is the only possible configuration when a connexon is inserted into the membrane outside the junction area. (c) Schematic drawing of a hemichannel. Each hemichannel is composed of six connexin molecules. An exploded view of a connexin molecule is shown on the right. Each connexin has four-transmembrane-spanning domains. Both the C-terminal (CT) and N-terminal (NT) portions of the connexin are located within the cell. This creates a single cytoplasmic loop (CL) and two extracellular loops (EL1 and EL2). Modified from Ransom BR and Ye ZC (2005) Gap junctions and hemichannels. In: Kettenmann H and Ransom BR (eds.) *Neuroglia*, 2nd edn., pp. 177–189. New York: Oxford University Press.

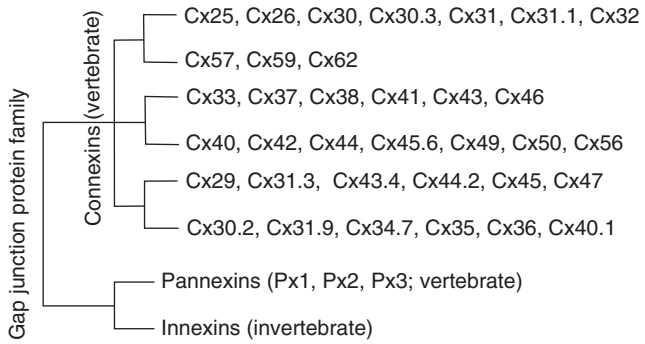

Figure 2 Phylogenetic 'tree' of known gap junction proteins in vertebrate (connexin and pannexin) and invertebrate (innexin) animal cells. More than 20 distinctive gap junction proteins have been identified (note that some of the 39 proteins listed here are only slightly different from one another and therefore do not count as distinctive proteins).

Electrophysiological analysis of single GJ channels indicates that, like other ion channels, they abruptly open and close. The permeability of GJ channels is influenced by physiological variables, including membrane potential and the intracellular concentrations of H^+ and Ca^{2+}. As a general principle, the permeability of GJs is greatest when there is no voltage difference across the junction; this corresponds to the situation when the two coupled cells have the same resting potential. Increases in intracellular concentrations of either H^+ or Ca^{2+} rapidly decrease junctional conductance (i.e., increase electrical resistance) and permeability to large molecules. Coupling conductance is sensitive to intracellular pH changes that are within the physiological range of this variable (i.e., pH 6.5 ± 0.5). Because neuronal activity can cause fluctuations of intracellular pH in astrocytes, this variable may be an important modulator of GJ function. Gating by pH depends on the connexin CT. Truncation of the CT eliminates pH sensitivity. It is believed that pH changes induce the CT to interact with a portion of the CL to close the channel, in a

fashion analogous to a 'ball and chain.' The increases in intracellular $[Ca^{2+}]$ ($[Ca^{2+}]_i$) that decrease junctional conductance are high (e.g., $\geq 10\,\mu M$), and this variable is not likely to affect coupling under normal physiological conditions; cell injury causes large increases in $[Ca^{2+}]_i$ that could lead to uncoupling.

GJ coupling is decreased by a wide range of drugs, including lipophilic agents such as the higher-order alcohols heptanol and octanol, and the anesthetic halothane. The mechanism of blockade by these agents is not clearly understood.

Hemichannels

HCs have not traditionally been thought of as 'stand-alone' ion channels. The presumption was that these channels existed only transiently in membranes before being incorporated into GJs. Moreover, to the extent that unpaired HCs existed, they would predominantly be in the closed state because of high extracellular $[Ca^{2+}]$ ($[Ca^{2+}]_o$) (Figure 3). A structural explanation for this was provided by the discovery of a site sensitive to near-millimolar $[Ca^{2+}]$ that is located on the extracellular portion of the HC pore. This site would not affect the permeability of GJs, however, because it would be inaccessible due to the tight meshing of the extracellular loops necessary to form a GJ. Because of their large pore diameter and presumed high permeability and low level of selectivity, open HCs could dissipate ion and metabolite gradients would seemingly threaten cellular integrity. This reasoning, and the fact that HCs have proved difficult to study for technical reasons, may explain why relatively little attention has been paid to HCs. Their involvement in brain function and pathology remains controversial. Because HCs connect cell cytoplasm to the extracellular space, the roles of HCs and traditional GJs are fundamentally different.

Unlike gap junctions, HCs are directly exposed to extracellular ions and thus can be effectively gated by extracellular, as well as intracellular, ion changes. A wide range of factors influence HC gating. For example, Ca^{2+}-free solution rapidly opens these channels in astrocytes and other cells. Unexpectedly, and in contradistinction to GJs, some HCs may also open in response to modest increases in $[Ca^{2+}]_i$. Open HCs are detected by uptake of appropriate marker dyes and by the release of molecules such as glutamate and ATP.

Most factors that regulate gap junctions, but probably not all, have similar effects on HCs. The ELs of an unopposed HC, for example, are far more exposed than are the same loops of two aligned HCs forming a GJ, making these ELs a prime candidate site for differential modulation (see earlier). For now, however, caution is needed when interpreting results obtained with agents that block gap junctions, because these would probably also block HCs. HCs may also be confused with other membrane ion channels that have very large pores, such as the P2X7 purinergic receptor channel.

Gap Junctions in Glial Cells

Connexins are variably expressed by all the main types of mammalian glial cells, with the possible exception of NG2 cells. Even microglial cells, felt at one time to be devoid of connexins, can express Cx43 and form gap junctions. Astrocytes are the most robustly coupled glial cells. Astrocytes of various types (e.g., protoplasmic, fibrous, Bergmann glia) form homotypic gap junctions with other astrocytes and heterotypic junctions with oligodendrocytes and ependymal cells. Astrocytes, as well as other glial cells, express several different connexins (Table 1) that may be regulated during development and affected by injury or cellular environment. Physiological variables

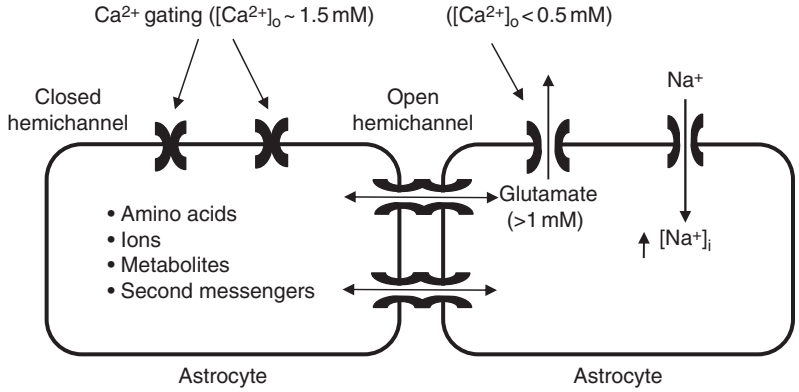

Figure 3 Schematic diagram of gap junctions and hemichannels in astrocytes. Gap junctions are typically open because they are exposed to low intracellular $[Ca^{2+}]$; they mediate diffusion of molecules up to 1000 Da. Hemichannels (i.e., unopposed connexons) are gated closed under normal conditions by high extracellular $[Ca^{2+}]$ (see text). When opened, hemichannels allow intracellular contents to equilibrate with extracellular fluid (e.g., glutamate leaves and Na^+ enters).

Table 1 Gap junction coupling in mammalian glial cells

Glial cell pair	*Relative coupling strength*[a]	*Connexin pairings*[b]
Astrocytes	++++	Cx43–Cx43, CX43–Cx30, Cx30–Cx30, Cx30–Cx26, Cx26–Cx26
Astrocyte–oligodendrocyte	++	Cx43–Cx47, Cx30–Cx32, CX26–Cx32
Oligodendrocytes	++	Cx32–Cx32, Cx29–Cx29
Schwann cells	++	Cx32–Cx32, Cx29–Cx29
Microglia	+	Cx43–Cx43
Müller cells	++	Cx43–Cx43, Cx43–Cx45, Cx45–Cx45
Müller cell–astrocyte	++	Cx43–Cx43, Cx43–Cx45, Cx45–Cx45
Ependymal cells	+++	Cx43–Cx43

[a]Relative coupling strength is shown as a semiquantitative score between + and ++++, with + being the lowest strength of coupling.
[b]Connexin pairings are shown in order of decreasing frequency. For example, Cx43–Cx43 is the most frequent connexin pairing for gap junctions between astrocytes.

such as glutamate application or depolarization can increase, while inflammatory cytokines decrease, astrocyte GJ coupling. The predominant connexin expressed by these cells is Cx43, but its expression varies by brain region. Not surprisingly, Cx43 is the most abundantly expressed connexin in the CNS. There are intriguing reports indicating that astrocytes form gap junctions with neurons. However, these junctions may be limited to immature cells and occur only during development. Mammalian astrocytes *in situ* typically show widespread dye coupling with each other. When the GJ-permeable dye Lucifer Yellow is injected into individual cortical astrocytes, as many as 100 adjacent cells are stained.

Oligodendrocytes primarily form gap junctions with astrocytes. The coupling between oligodendrocytes and astrocytes is mainly mediated by GJs composed of Cx43 (astrocytes) and Cx47 (oligodendrocytes) (Table 1). Although functional coupling between oligodendrocytes is readily detected *in vitro*, actual GJs are only rarely seen using electron microscopy on freshly prepared whole brain. Based on the scarcity of morphological GJs, it has been suggested that communication between oligodendrocytes might be mediated by connections through the astrocytic syncytium. 'Coupling strength' measured electrically and by dye passage is much greater between astrocytes than between astrocytes and oligodendrocytes.

In the peripheral nervous system, proliferating Schwann cells during development, or after peripheral nerve crush injury, primarily express Cx46 and are coupled to one another. When Schwann cells begin myelinating axons, they stop expressing Cx46 but increase expression of Cx32, which is confined to paranodal regions and Schmidt–Lanterman incisures. These so-called reflexic gap junctions connect the paranodal folds, perhaps facilitating ion and small-molecule movements to and from the tight, periaxonal space. The X-linked form of the hereditary neuropathy known as Charcot–Marie–Tooth disease is associated with mutations in the Cx32 gene.

Other specialized glial cells, such as ependymal cells, Müller cells of the retina, and Bergmann glia of the cerebellum, express connexins and GJs (Table 1). Activated, but not resting, microglia express Cx43 and show dye coupling *in vitro*. Interestingly, activated microglia can downregulate Cx43 expression and coupling in nearby astrocytes, probably via cytokine secretion, suggesting that astrocyte communication may be modulated by CNS inflammation mediated by microglia.

Functions of GJs and HCs

The functions of connexins, and the channels that they form (both GJs and HCs), are still debated and undoubtedly vary by cell type. This uncertainty about function is especially true for GJs and HCs in glial cells. The following comments about function are directed mainly at astrocytes. In general, GJs may serve to coordinate the electrical and metabolic activities of cell populations, act to amplify the consequences of signal transduction, control intrinsic proliferative capacity, and help to orchestrate the complex events of embryonic morphogenesis. Roles of GJs in astrocytes include equilibrating intracellular ions among coupled cells (Figures 3 and 4(a)), regulating cell proliferation, participation in signal dispersion (Figure 4(b)), participation in Ca^{2+} wave propagation (Figure 4(c)), and involvement in brain injury. Because astrocytes are strongly coupled, they can experience GJ-mediated signal dispersion (or amplification). Consider a cell stimulated to produce an intracellular second-messenger molecule (e.g., cAMP, inositol 1,4,5-trisphosphate, Ca^{2+}) by binding a specific ligand (Figure 4(b)). If the cell is not connected to its neighbors by gap junctions it will act in isolation. If the cell is widely coupled to other cells, however, the diffusible junction-permeable second messenger has the potential of affecting many cells, effectively amplifying the initial signal.

Astrocyte Ca^{2+} waves are propagated waves of elevated intracellular Ca^{2+} that may be elicited by

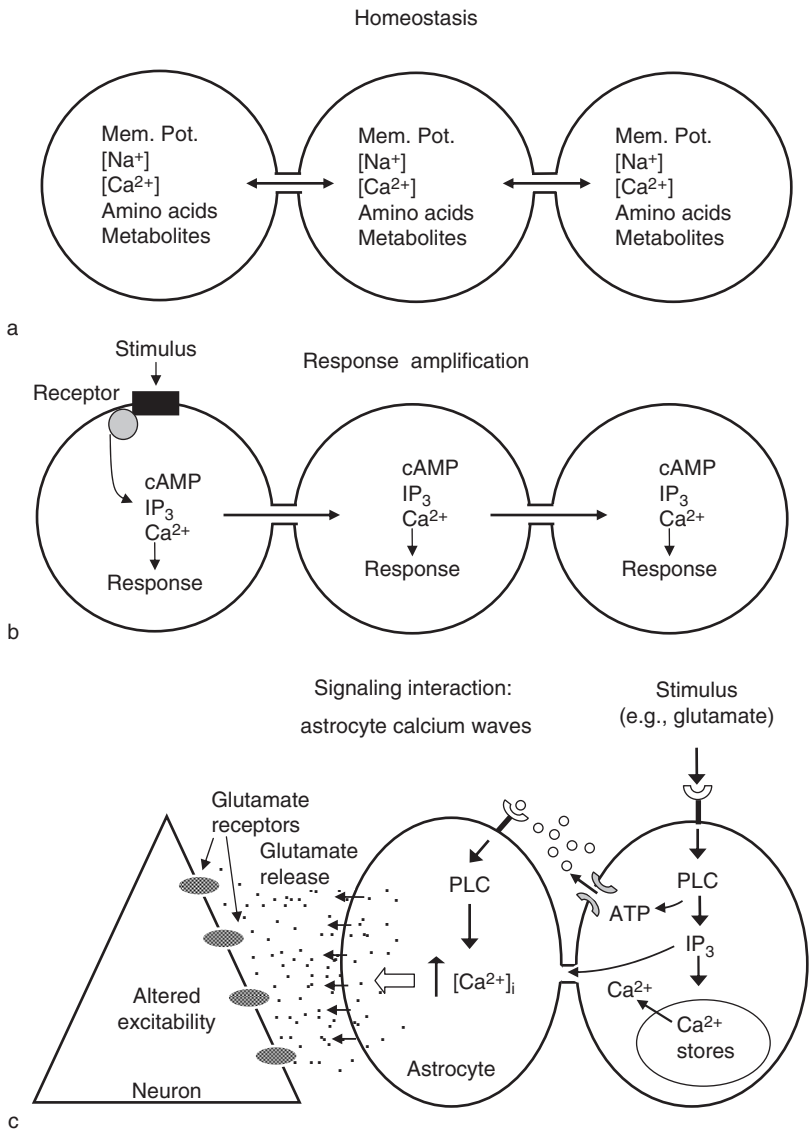

Figure 4 Illustration of possible functions of gap junctions and hemichannels. (a) In strongly coupled cell aggregates, the intracellular concentrations of ions and molecules less than ~1000 Da in size (including amino acids, sugars, and second-messenger molecules) will tend to be similar due to free exchange across gap junctions. (b) Glial cells express many receptors for signaling molecules (e.g., neurotransmitters, cytokines). Coupling can amplify the physiological consequences of a chemical signal acting upon a single cell. Such stimuli often evoke production of second-messenger molecules, such as Ca^{2+}, inositol 1,4,5-trisphosphate (IP_3), and cAMP, that can diffuse into adjacent cells via gap junctions; the coupled cells are then recruited to respond. (c) Gap junctions and hemichannels participate in astrocyte Ca^{2+} waves. Astrocyte Ca^{2+} waves are provoked in many ways, such as by the application of glutamate. Effective stimuli activate intracellular signaling pathways, such as those involving phospholipase C (PLC) and IP_3, which cause Ca^{2+} release from Ca^{2+} stores. IP_3 and Ca^{2+} diffuse to adjacent cells through gap junctions, creating a 'Ca^{2+} wave'. Astrocytes also release ATP upon activation, probably through hemichannels. In turn, ATP activates purinergic receptors on nearby astrocytes, the predominant mode by which Ca^{2+} waves propagate. Increase in astrocyte $[Ca^{2+}]_i$ causes glutamate release, which modulates the excitability of neighboring neurons. Reproduced from Ransom BR and Ye ZC (2005) Gap junctions and hemichannels. In: Kettenmann H and Ransom BR (eds.) *Neuroglia*, 2nd edn., pp. 177–189. New York: Oxford University Press, with permission.

glutamate application, mechanical perturbation, or ischemia (Figure 4(c)). As the wave passes through astrocytes it can elicit glutamate release and alter the behavior of adjacent neurons; in essence, this event is a slow form of signaling (slow compared to action potentials and synaptic transmission in neurons). Ca^{2+} wave propagation is dependent on both gap junctions and extracellular signaling pathways, including hemichannel-mediated ATP release; the relative contribution of these two mechanisms may vary in different brain regions.

Gap junctions tend to restrict cell proliferation. During neoplastic transformation, gap junctions are usually lost and most malignant cells are not coupled. Most primary brain tumors derive from glial cells, and loss of GJs could plausibly contribute to

CNS neoplasia. How GJs suppress mitotic activity is not clear.

Genetic manipulation is one method of assessing the function of connexins. Animals null for Cx43 die at birth because of a cardiac defect. Conditional mutants have been created in which Cx43 is suppressed only in astrocytes. These animals live to adulthood and exhibit subtle neurological changes, including increased exploratory behavior, an anxiolytic-like state, and impaired motor capacities. In humans, mutations in Cx43 cause an autosomal-dominant syndrome called oculodentodigital dysplasia (ODDD). At least half of these patients have neurological abnormalities, including spastic paraplegia and neurodegeneration (i.e., marked changes in cortex and white matter on magnetic resonance imaging), in addition to craniofacial and limb dysmorphisms. The exact pathophysiology of this condition remains uncertain and some mutations may cause a toxic gain of function or disturb the normal interactions known to occur between connexin and nonconnexin proteins.

The roles of glial HCs are mainly a matter of speculation at this time. Possible roles include (1) release of signaling molecules such as ATP or glutamate, (2) modulation of membrane potential or ion gradients in a manner similar to conventional ion channels, and (3) mediation of cell injury under pathological conditions such as ischemia. HCs can open under pathological conditions, perhaps due to oxidative stress, and may release glutamate, which could contribute to excitotoxic injury.

Conclusions

Connexins and the channels they form have been the focus of intense investigation for greater than a half century since their discovery. As more information about these fascinating channels has emerged, more questions have appeared. This is especially so in the brain, where these channels are widely expressed in glial cells. We lack conclusive insight about how they participate in brain function and pathological conditions. We are at a point, however, where critical hypotheses about their function can be formulated and tested using improved research techniques, including regional and cell-type-specific genetic manipulation of connexin expression.

See also: Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Gap Junctions and Neuronal Oscillations; Gap Junctions and Electrical Synapses; Gap Junction Abnormalities and Disorders of the Nervous System; Gap Junctions: Metabolic Exchanges; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Oligodendrocyte Morphology.

Further Reading

Bennett MV, Barrio LC, Bargiello TA, et al. (1991) Gap junctions: New tools, new answers, new questions. *Neuron* 6: 305–320.

Bruzzone R, Hormuzdi SG, Barbe MT, et al. (2003) Pannexins, a family of gap junction proteins expressed in brain. *Proceedings of the National Academy of Sciences of the United States of America* 100: 13644–13649.

Contreras JE, Sanchez HA, Eugenin EA, et al. (2002) Metabolic inhibition induces opening of unapposed connexin 43 gap junction hemichannels and reduces gap junctional communication in cortical astrocytes in culture. *Proceedings of the National Academy of Sciences of the United States of America* 99: 495–500.

Cotrina ML, Lin JH, Alves-Rodrigues A, et al. (1998) Connexins regulate calcium signaling by controlling ATP release. *Proceedings of the National Academy of Sciences of the United States of America* 95: 15735–15740.

Dermietzel R, Gao Y, Scemes E, et al. (2000) Connexin43 null mice reveal that astrocytes express multiple connexins. *Brain Research – Brain Research Reviews* 32: 45–56.

Froes MM, Correia AH, Garcia-Abreu J, et al. (1999) Gap-junctional coupling between neurons and astrocytes in primary central nervous system cultures. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7541–7546.

Gomez-Hernandez JM, de Miguel M, Larrosa B, et al. (2003) Molecular basis of calcium regulation in connexin-32 hemichannels. *Proceedings of the National Academy of Sciences of the United States of America* 100: 16030–16035.

Goodenough DA, Goliger JA, and Paul DL (1996) Connexins, connexons, and intercellular communication. *Annual Review of Biochemistry* 65: 475–502.

Harris AL (2001) Emerging issues of connexin channels: Biophysics fills the gap. *Quarterly Review of Biophysics* 34: 325–472.

Kettenmann H and Ransom BR (1988) Electrical coupling between astrocytes and between oligodendrocytes studied in mammalian cell cultures. *Glia* 1: 64–73.

Levin M (2002) Isolation and community: A review of the role of gap-junctional communication in embryonic patterning. *Journal of Membrane Biology* 185: 177–192.

Phelan P and Starich TA (2001) Innexins get into the gap. *BioEssays* 23: 388–396.

Ransom BR and Ye ZC (2005) Gap junctions and hemichannels. In: Kettenmann H and Ransom BR (eds.) *Neuroglia,* 2nd edn., pp. 177–189. New York: Oxford University Press.

Rouach N, Glowinski J, and Giaume C (2000) Activity-dependent neuronal control of gap-junctional communication in astrocytes. *Journal of Cell Biology* 149: 1513–1526.

Saez JC, Berthoud VM, Branes MC, et al. (2003) Plasma membrane channels formed by connexins: Their regulation and functions. *Physiological Reviews* 83: 1359–1400.

Saez JC, Contreras JE, Bukauskas FF, et al. (2003) Gap junction hemichannels in astrocytes of the CNS. *Acta Physiologica Scandinavika* 179: 9–22.

Sontheimer H, Minturn JE, Ransom BR, et al. (1991) Cell coupling is restricted to subpopulations of astrocytes cultured from rat hippocampus and optic nerve. *Annals of the New York Academy of Sciences* 633: 592–596.

Spray DC, Ye ZC, and Ransom BR (2006) Functional connexin hemichannels: A critical appraisal. *Glia* 54(7): 758–773.

Unger VM, Kumar NM, Gilula NB, et al. (1999) Three-dimensional structure of a recombinant gap junction membrane channel. *Science* 283: 1176–1180.

Ye ZC, Wyeth MS, Baltan-Tekkok S, et al. (2003) Functional hemichannels in astrocytes: A novel mechanism of glutamate release. *Journal of Neuroscience* 23: 3588–3596.

Gap Junctions and Neuronal Oscillations

M O Cunningham and F E N LeBeau, Newcastle University, Newcastle upon Tyne, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In addition to neuronal chemical neurotransmission, electrical transmission via gap junctions, which are protein channels that directly link the cytoplasm of adjacent cells, is now known to constitute an important type of cell-to-cell communication. Although initially thought to be important mainly in development, it is now clear that electrical signaling via gap junctions plays an important role in the adult central nervous system (CNS).

Considerable advances have been made in understanding the contribution of electrical signaling, via gap junctions, to neuronal signaling, most notably the role of synchronizing oscillatory or rhythmic activity. Oscillatory neuronal activity underlies the production of the signals recorded with an electroencephalogram (EEG) and reflects the coordinated activity of large populations of neurons. An EEG of the brain reveals rhythmic oscillatory activity in distinct frequency bands that occur during different behavioral states, and can be altered in a number of neurological and psychological diseases. Work from many research groups over the past decade has highlighted the role of electrical signaling via gap junctions in the generation, and/or spread, of synchronous oscillatory network activity. Electrical coupling via gap junctions can occur in both neurons and glial cells throughout the CNS. However, in this article, we focus on electrical coupling in excitatory principal cells and inhibitory interneurons, mainly in the neocortex and hippocampus. We provide an overview of the role of electrical signaling via gap junctions in the generation of oscillatory network activity and discuss briefly the importance of gap junction function in neurological conditions such as epilepsy.

Historical Background

The idea that neurons can form an interconnected network via direct cell-to-cell connections is not a novel concept. Proposed by Camillo Golgi in the late nineteenth century, the 'reticular theory' postulated that the nervous system could form a syncytium consisting of nerve fibers forming an intricate connected network and that the nerve impulse could propagate along such a diffuse network. However, at the same time, Ramón y Cajal's 'neuron doctrine' hypothesized that the nervous system was composed of anatomically and functionally independent units (neurons) with chemical synapses connecting them. Today we understand that in the adult mammalian brain, each of these theories reflects an important aspect of the communication in neuronal networks.

Structure and Properties of Gap Junctions in the CNS

Gap junctions form a continuous link between two separate neuronal structures (Figure 1) that allows the direct transmission of electrical signals between neurons. Neurons can be coupled together via gap junctions connecting soma-to-soma, dendrite-to-dendrite, or soma-to-dendrite. More recently, evidence has appeared of gap junctions linking neurons via axon-to-axon connections. The major proteins that form gap junctions are termed connexins (Cxs). Different Cx subtypes are defined by their molecular mass (in kilodaltons). A number of different Cx subtypes have now been identified in the vertebrate CNS, including Cx26, Cx32, Cx36, Cx43, and Cx45, but of these only Cx36 forms have been shown to be specific for neurons. A functional channel connecting two different neurons is formed by the apposition of two connexons, one in the plasma membrane of each neuron. Each connexon contains six Cx proteins. A functional gap junction is thus composed of 12 Cx subunits. Diversity in the composition of connexons in a gap junction can arise, as hemichannels can either consist of the same Cx subtype (homomeric) or different Cx subtypes (heteromeric).

Recently, a new family of gap junction proteins, the pannexins, has been identified in the CNS. Related to innexins, an invertebrate family of gap junction proteins, the so-called pannexins (Px) have been demonstrated to form functional gap junctions in recordings from paired *Xenopus* oocytes. Three Px subtypes have been described in the rat and human genome, with Px1 and Px2 expressed solely in the CNS. There are strong expression patterns for Px1 and Px2 in the cortex and hippocampus, but their functional role in network activity is unknown.

Gap junctions are not simply inert pores between neurons, as they can be dynamically modulated in a number of ways. Evidence suggests that the number of gap junctions, the pattern of expression of Cx proteins, and conductance properties of the individual channels can all be modified by neurotransmitters, second messengers, calcium ion concentrations, intra- and extracellular pH, and transmembrane voltage. This sort of modification occurs during development

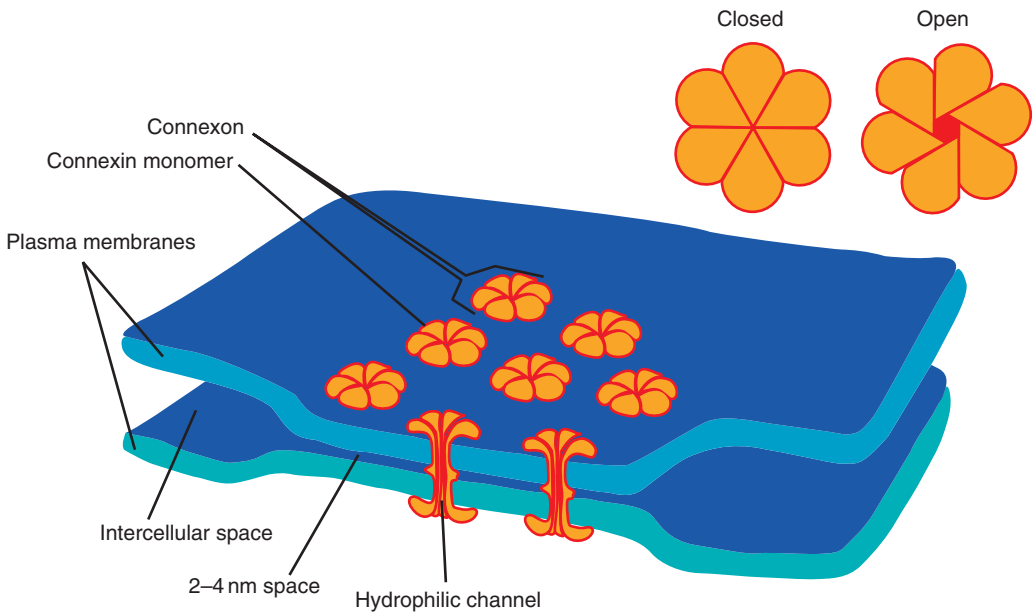

Figure 1 Structure of a gap junction, showing a cluster of gap junctions (yellow) spanning two opposing plasma membranes (blue sheets). The intercellular space narrows at the point of the gap junction, bringing the two membranes into close contact. A gap junction consists of two hemichannels (one in each plasma membrane) that form a continuous channel between the two cells. Each hemichannel contains six connexin proteins, which, as a unit, are called a connexon. Insets show that a conformational change in the connexin proteins allows the gap junctions to open and close in response to modulators such as pH, calcium, and cAMP.

and in response to a variety of factors in the adult. Modulation allows for the complex, connexin-specific regulation of electrical signaling via gap junctions, although to date little is known about the mechanism.

Approaches to Studying Gap Junctions

Evidence for electrical coupling mediated via gap junctions in the CNS comes from using a multidisciplinary approach. The CNS consists of excitatory principal cells and inhibitory interneurons, and there is evidence for electrical signaling in both neuronal populations. Electron microscopy and freeze-fracture studies have anatomically demonstrated gap junctions between both interneurons and pyramidal cells in the cortex.

Early electrophysiological evidence for electrical signaling via gap junctions was obtained by recording from principal cells in the cortex and hippocampus. These studies also showed that certain dyes (e.g., Lucifer yellow) could pass from one principal neuron to another via gap junctions. In dye coupling experiments a specific dye is injected into one cell, and if that cell is coupled to other cells, dye transfer may be detected in the connected cells. However, dye coupling studies are complicated because different connexins may have different permeabilities to dyes, and dye transfer depends on the time for filling and the locations of the gap junctions. For this reason a lack of dye coupling does not completely exclude the presence of functional gap junctions.

Interneurons represent only 10–20% of the neuronal population and are thus difficult to record 'blind' with intracellular sharp electrode techniques. In order to surmount this issue experimentalists have utilized infrared differential interface contrast (IR-DIC) microscopy and whole-cell patch recording techniques. This has made it possible to identify interneurons visually from their basic morphology and thus conduct simultaneous whole-cell patch recordings from two or more interneurons. Using this technique electrical coupling between pairs of interneurons has been demonstrated. When current is injected into one cell to change the membrane potential, a potential change is also seen in the coupled cell (Figure 2). The ratio of the potential change in cell 1/cell 2 is termed the coupling coefficient or coupling ratio. As with the dye transfer studies, the absence of electrical coupling potentials cannot totally exclude the possibility that gap junctions may be present at sites too distant for current transfer to be detected with somatic recordings.

Electrically coupled interneurons have now been shown to be widespread in many areas of the vertebrate CNS, including the hippocampus and neocortex. However, elucidating the functional significance of electrical signaling via gap junctions has been made difficult by the lack of specific agents to block gap

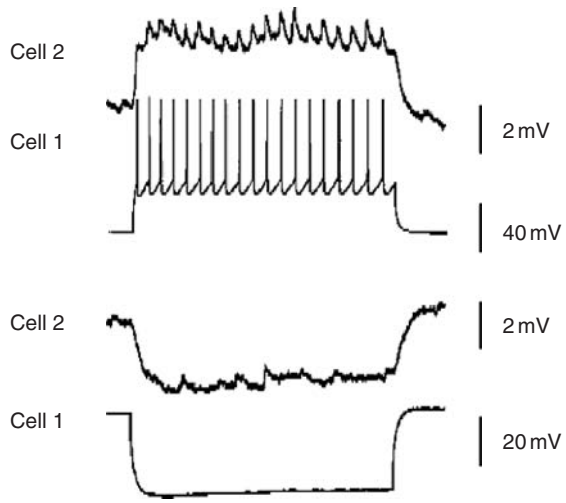

Figure 2 Electrical coupling potentials, showing electrical coupling between fast-spiking interneurons. Paired whole-cell recordings of two fast-spiking interneurons from the hippocampal dentate gyrus region in the mouse are depicted. Traces show the voltage responses in cell 1 following depolarizing (upper) and hyperpolarizing (lower) current injections; voltage responses were also reflected in cell 2 but were smaller in amplitude.

junctions. Uncoupling agents that are currently used include heptanol, octanol, halothane, carbenoxolone, and 18α-glycyrretinic acid. However, these agents will block all gap junctions (including glial gap junctions), making it impossible to clearly identify the importance of any specific gap junctions. Recently, blockers for Cx36 have been described, including quinine and the more specific mefloquine. However, all of these compounds have been reported to have a range of nonspecific effects that could also affect neuronal activity. New strategies have been developed to study the role of electrical signaling via gap junctions using genetically engineered mice that lack specific connexin proteins (see later).

Evidence for Gap Junctions between Principal Neurons

Dye coupling studies show that when one principal cell is injected with dye, often one or two other cells also fill with dye, suggesting direct coupling between the cells. Procedures that increase gap junction conductance (e.g., reduced extracellular calcium) tend to increase the number of dye-coupled neurons. Conversely, the degree of dye coupling is diminished when gap junction conductance is reduced by the application of gap junction blockers. In these and earlier studies, potentials resembling small action potentials (<10 mV) were observed in recordings from the soma of principal cells. Termed fast prepotentials (FPPs), and first observed *in vivo*, these events were initially believed to arise from electrogenesis in dendrites. Subsequent experimental studies, however, suggested that FPPs or spikelets (presently the more commonly used term) represent action potentials in electrically coupled neurons.

The location of the gap junctions that mediate coupling between principal cells has yet to be conclusively demonstrated. Paired whole-cell patch recording techniques have failed to demonstrate significant electrical coupling between the somas of principal cells. However, recent intracellular recordings with sharp electrodes show high levels of coupling between pyramidal cells in the hippocampus. There is evidence to suggest that electrical signaling via gap junctions could be axo-axonic in principal cells. Computational studies first proposed that the fast kinetics of the spikelets seen in pyramidal cells could only be explained if the gap junctions were located between axonal compartments. If the gap junctions were located at remote sites on the dendrites of pyramidal cells, then the electrical-filtering properties of the dendrites would result in slower kinetic properties of the spikelets. Experimental evidence was subsequently provided when one study, using paired recordings from the soma and axon hillock/axon initial segment of the same cell, showed that following antidromic stimulation of the relevant fiber tract, the evoked spikelet appeared first in the axon hillock/initial segment and then in the soma. Subsequent visualization experiments revealed that the path taken by dye loaded into one cell could be followed in real time, as it was observed first filling in the axon and then subsequently in the soma of a second neuron. One recent electron microscopy study has found ultrastructural evidence for gap junctions between pairs of mossy fiber axons in the CA3 region of the hippocampus. However, the identities of the proteins that compose the putative axo-axonic gap junctions are not fully known, although the pannexins and Cx36 have both been proposed as possible candidates.

Evidence for Gap Junctions between Interneurons

Perhaps the greatest advance in our understanding of the role of electrical signaling via gap junctions has come from studies of interneurons. Interneurons in the hippocampus and neocortex consist of a number of distinct subtypes that can be identified on the basis of their morphology, physiological properties, and their expression of specific chemical markers such as peptides or calcium-binding proteins. Early anatomical studies clearly demonstrated evidence for the existence of dendrodendritic or dendrosomatic gap junctions between interneurons in the cortex.

Recent physiological studies have also shown that interneurons can be electrically connected. Electrical coupling between interneuron pairs has been shown to be bidirectional, with current usually flowing equally in both directions. In addition, gap junctions function as low-pass filters in that they preferentially transmit low-frequency signals. Fast signals, such as action potentials, are not transmitted as well and appear in the coupled cell as a truncated spikelet. Nonetheless, spikelets can still facilitate the generation of action potentials in the coupled cell and so enhance the spread of activity. The traditional view of electrical signaling via gap junctions as important for mediating fast excitation is an oversimplification. Gap junctions can transmit depolarizing and hyperpolarizing current equally well. For example, the afterhyperpolarization that follows an action potential is also transmitted to the coupled cell, so a spikelet can be followed by a significant hyperpolarization. Thus electrical signaling can be both excitatory and inhibitory.

A key feature of gap junction coupling observed between cortical interneurons is that it is mainly between interneurons of the same class (e.g., fast-spiking basket cells, multipolar bursting cells, neurogliaform cells, and low-threshold spiking cells). As it has long been postulated that different classes of interneurons are involved in different functions, this specificity of connectivity leads to the formation of distinct networks of electrically coupled interneurons, each of which could be involved in a separate function. In the neocortex, at least five different interneuron networks of electrically coupled cells have now been identified (Figure 3). Interneurons belonging to the same morphological class exhibit gap junction coupling, often with a high incidence (e.g., 50% of basket cells coupled). However, only a few cases of electrical coupling between different interneuron classes have been reported. Paired whole-cell patch recordings reveal that electrical coupling between all inhibitory interneurons so far studied is mediated by the same connexin (Cx36). In mice (though not in rats), Cx36 is expressed only in interneurons, and mice that lack Cx36 exhibit no functional electrical coupling between pairs of interneurons. As yet we know nothing about the mechanisms that regulate the specificity of interneuron coupling so as to ensure that connections only form between interneurons of the same type. This remains an important issue to address in the future.

There is strong evidence for the developmental refinement of the cell-specific distribution of gap junctions. The degree of electrical coupling in both rodent and human brain declines with maturation. This decline in the extent of electrical coupling was one reason that electrical signaling via gap junctions

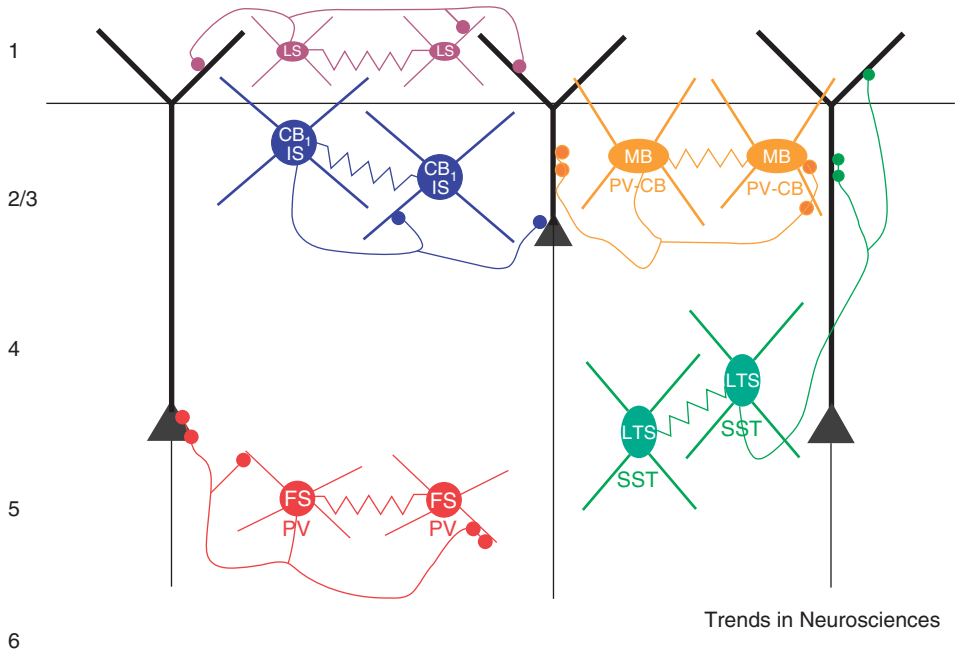

Figure 3 Five networks of distinct electrically coupled γ-aminobutyric acid (GABAergic) interneurons in the cerebral cortex. Numbers indicate cortical laminae. Pyramidal cells are shown in black. Different colored circles represent different interneuron classes that are electrically coupled (jagged lines). Some interneuron classes are also chemically synaptically coupled, and all form synapses onto the pyramidal cells. CB, calbindin; CB_1 IS, cannabinoid receptor-1-expressing irregular-spiking cell; FS, fast-spiking cell; LS, late-spiking cell; LTS, low-threshold spiking cell; MB, multipolar bursting cell; PV, parvalbumin; SST, somatostatin.

was thought to be important mainly in development. However, a number of studies have demonstrated that electrical coupling persists in the adult brain, and computational studies predict that low levels of coupling (e.g., one to two gap junctions per cell) are in fact sufficient to enhance network synchronization.

Gap Junctions and Oscillations of Cognitive Relevance

Oscillatory, or synchronized, network activity occurs in the EEG in different frequency bands during specific behavioral states in mammals, including humans. Oscillations can occur at from less than 1 Hz to several hundred hertz. This broad range of activity is tightly linked to higher cognitive function such as memory formation, sensory perception, and integration. Two types of oscillatory activity in which the role of electrical signaling via gap junctions has been particularly well studied are the gamma frequency (20–80 Hz) and ultrafast (>80 Hz) or very fast oscillations (VFOs). Gamma frequency activity can be recorded in many brain areas but is particularly prominent in the neocortex and hippocampus. It is associated with a number of functions such as sensory perception, attention, learning, and memory. Ultrafast oscillations in the hippocampus are also associated with memory functions, particularly the consolidation of declarative memories.

Role of Electrical Signaling via Gap Junctions in Generating Fast Network Oscillations

A major methodological advance in the study of gamma frequency oscillations has been the ability to study this activity *in vitro*. Persistent gamma oscillations can be induced and sustained in brain slices of hippocampus and cortex by the bath application of drugs that can depolarize both pyramidal cells and interneurons such as kainate, muscarinic, or glutamate metabotropic receptor agonists. Gamma frequency activity cannot be generated in the absence of inhibitory neurotransmission, and several studies have identified specific interneuron classes, particularly fast-spiking basket cells, that fire at gamma frequency during the oscillation and so contribute to the generation of this activity. However, in addition to its role in GABAergic inhibition, electrical signaling via gap junctions also appears to be important for the generation of gamma frequency oscillations. Recent work suggests that electrical coupling between both interneurons and pyramidal cells is important for the generation of gamma frequency activity.

It had been known for some time that application of gap junction blockers (halothane, octanol, and carbenoxolone) can abolish gamma frequency oscillations, but these studies could not distinguish the contribution of electrical signaling via gap junctions in interneurons versus pyramidal cells. However, it is possible *in vitro* to generate a gamma frequency oscillation in which pyramidal cell activity has been blocked – the so-called interneuronal network gamma (ING). Computational modeling studies initially suggested that, along with chemical inhibitory transmission, electrical signaling via dendritic gap junctions between interneurons was important for synchronizing gamma frequency activity. Using the *in vitro* ING model it was found that blockade of gap junctions markedly reduced the synchrony of the oscillations, although some gamma frequency activity still remained.

The contribution of electrical coupling between interneurons to fast network oscillations was tested further using Cx36 knockout mice that specifically lack functional electrical coupling between interneurons. Using the Cx36 mice, in which excitation is intact, it was found that the ability of hippocampal neuronal networks to generate gamma frequency activity was significantly reduced. Reductions in gamma frequency activity were seen both *in vitro* and *in vivo*. At a cellular level, recordings from principal cells and interneurons *in vitro* showed a disruption in the amplitude, rhythmicity, and coherence of inhibitory inputs. Under normal conditions, spikelets in interneurons increase the likelihood of full action potentials being generated in the coupled cells, thus enabling a network of interneurons to fire synchronously. These synchronous inhibitory outputs impinge on the postsynaptic principal cells and control the timing of the spike output, as the principal cells can only fire once the inhibition has decayed. The inability of interneurons in the Cx36 knockout mice to synchronize properly their firing via gap junctions leads to a reduction in the precision of their action potential timing and thus their rhythmic inhibitory output. Behavioral studies have shown that mice lacking Cx36 exhibit learning and memory impairments, an observation consistent with the idea that normal gamma frequency activity is important for these cognitive tasks.

Importantly, the loss of Cx36 did not lead to the complete abolition of gamma frequency activity. A residual gamma oscillation was still present both *in vivo* and *in vitro*. *In vitro* it was shown that this remaining gamma frequency oscillation was blocked by the putative gap junction blocking agent, carbenoxolone. This suggested that some other form of

gap junction-mediated signaling is also required for the generation of the gamma frequency activity, and this is now thought to arise from the pyramidal cell population.

Electrical signaling via gap junctions between pyramidal cells was initially proposed to underlie the generation of certain types of ultrafast or 'ripple' activity. Using an *in vitro* brain slice model, high-frequency oscillatory activity, or 'ripples,' occurring at ~80–150 Hz can be generated in the hippocampus in the absence of chemical synaptic transmission when using a nominally calcium-free bath medium. The lack of chemical synaptic transmission in these experiments supports the hypothesis that this activity is dependent on direct electrical connections between neurons. The 'ripple' activity is abolished in the presence of pharmacological blockers of gap junctions, such as halothane or carbenoxolone. At the level of individual neurons, intracellular recording from principal cells revealed that, concomitant with the extracellular field potential 'ripple,' the cells discharged either a full-action potential or a spikelet. The kinetic properties of these spikelets (as discussed earlier) were too fast to be due to dendrodendritic gap junctions, and so axo-axonal connections were proposed. In addition, in the Cx36 knockout mice, this ultrafast oscillatory activity was normal both *in vivo* and *in vitro,* suggesting that interneuronal electrical coupling did not underlie this activity.

More recent work has now demonstrated that the ultrafast oscillations coexist with the gamma frequency activity. Ultrafast oscillations can be generated in the axon plexus of pyramidal cells in the hippocampus. It is currently thought that these high-frequency oscillations occur as a result of random ectopic action potentials (action potentials arising from regions other than the axon hillock) in the axon plexus. This random activity spreads throughout the axon plexus, via axo-axonic gap junctions, resulting in a barrage of fast excitatory synaptic potentials that can be recorded in the interneurons, and provides the necessary excitatory drive for the interneurons. The combination of chemical and electrical connections between the interneurons then allows a synchronized output back to the pyramidal cells at gamma frequency. Gap junctions are therefore required between the pyramidal cells to generate the drive to the interneurons, and the gap junctions between the interneurons are needed to temporally coordinate their activity.

Gap Junctions and Epilepsy

The EEG is an important test in the diagnostic process of treating epilepsy. Traditionally, this approach has been noninvasive, using small, metal electrodes attached to the scalp. This approach is satisfactory for the medical evaluation of the seizure activity. However, there are limitations to this approach. The advent of digital recordings and invasive human studies using depth electrode recordings or subdural electrode mats now allows the measurement of electrical signals with frequencies greater than those recordable with conventional EEG equipment. Using these techniques, ultrafast oscillatory activity has been observed in close proximity to seizure onset zones in both animal models and humans. Indeed, it has been suggested that this type of activity is a reliable electrophysiological marker of epileptogenicity. A common feature of focal epileptic seizure activity, in both humans and experimental models, is the presence of ultrafast oscillations. Thus fast ripples associated with interictal spikes currently provide information on localization and epileptogenicity of the tissue involved in their generation. Evidence from developmental studies further supports a role of gap junctions in epilepsy. Thus genetic linkage mapping demonstrates that Cx36 is a high-ranking positional and functional candidate gene for juvenile myoclonic epilepsy.

A number of molecular biology studies have demonstrated that in neuronal tissue obtained from animal models and human cases of epilepsy there are obvious alterations in the abundance of connexins. Most of this work concerns the expression of connexins, which is limited to astrocytes. Several studies have examined protein and or mRNA expression, but the results from these studies yield conflicting results, with some reporting increases in connexin expression and others reporting decreases. Differences in a number of the experimental parameters applied could account for these discrepancies. It is obvious that more work is required to determine whether Cxs are modified in epileptic tissue, and if so in what directions. Moreover, if putative axo-axonic gap junctions are composed of pannexins, then attempts should be made to examine the expression and localization of such proteins in both human epileptic tissue and tissue from animal experimental models of epilepsy.

Conclusions

Recent advances over the past decade have demonstrated that both chemical signaling and electrical signaling play a central role in cortical communication. It is generally widely accepted that chemical signaling allows for highly complex, specific, and modifiable neuronal communication. However, it is now apparent that electrical signaling mediated via gap junctions can also demonstrate a high degree of functional specificity. Thus, the combination of

multiple coexistent synaptic and electrical networks serves to greatly enrich the processing capabilities on neuronal networks.

See also: Epilepsy; Epilepsy: Channelopathies; Gap Junctions and Electrical Synapses; Gap Junction Abnormalities and Disorders of the Nervous System; Neural Synchrony and Feature Binding; Neural Oscillators and Dymanmical Systems Models; Neurotransmitter Release: Synchronous and Asynchronous; Presynaptic Receptor Signaling.

Further Reading

Bennet MLV and Zukin RS (2004) Electrical coupling and neuronal synchronization in the mammalian brain. *Neuron* 41: 495–511.

Buzsaki G and Draguhn A (2004) Neuronal oscillations in cortical networks. *Science* 304: 1926–1929.

Connors BW and Long MA (2004) Electrical synapses in the mammalian brain. *Annual Review of Neuroscience* 27: 393–418.

Galarreta M and Hestrin S (2001) Electrical synapses between GABA-releasing interneurons. *Nature Reviews Neuroscience* 2: 425–433.

Hestrin S and Galerreta M (2005) Electrical synapses define networks of neocortical GABAergic neurons. *Trends in Neuroscience* 28: 304–309.

Schmitz D, Schuchmann S, Fisahn A, et al. (2001) Axo-axonal coupling: A novel mechanism for ultrafast neuronal communication. *Neuron* 31: 831–840.

Whittington MA and Traub RD (2003) Interneuron diversity series: Inhibitory interneurons and network oscillations *in vitro*. *Trends in Neuroscience* 26: 676–682.

Gap Junctions: Metabolic Exchanges

J M Medina and A Tabernero, Universidad de Salamanca, Salamanca, Spain

© 2009 Elsevier Ltd. All rights reserved.

Introduction

One of the first roles assigned to astrocytes was the supply of energetic compounds to neurons, since astrocytes form the first cellular barrier encountered by glucose as it enters the brain. In fact, astrocytic end-feet surround intraparenchymal blood vessels and astrocytic processes are in close contact with neurons. Thus, astrocytes are physically located between the capillaries and neuronal elements. In addition, astrocytes are ideally suited for such a nutrient role: they store glycogen, to the point where they can be identified at the ultrastructural level by the presence of glycogen granules. In addition, astrocytes process glucose glycolytically and release lactate, which serves as a metabolic fuel for neurons. Thus, due to their typical morphology, their intimate ensheathing of blood capillaries and synaptic contacts, and their specific metabolic activity, astrocytes can sense synaptic activity, evaluate the neuronal need for energetic metabolites, and dispatch metabolites for the maintenance of energy balance in neurons.

This metabolic role of astrocytes may be played by single and independent astrocytes, but a characteristic feature of astrocytes is communication through intercellular gap junctions. Since these intercellular channels allow diffusion of molecules between adjacent cells, it seems more likely that functionally coupled astrocytes are involved in metabolic trafficking. According to this hypothesis, two or more astrocytes can form a 'shuttle system' between blood vessels and neurons, and therefore the participation of gap junctions in metabolic trafficking should be considered. The aim of this article is to put together the evidence obtained so far, regarding diffusion of metabolic substrates through astrocytic gap junctions that can support this hypothesis. In addition, we focus on the effects of disrupting gap junctional communication on glucose uptake and metabolism and its relation with proliferation.

Permeability of Astrocytic Gap Junctions for Metabolic Compounds

Classically, most studies performed to demonstrate the occurrence of gap junction-mediated intercellular communication have been carried out using either electrophysiological techniques or intracellular loading of low molecular mass (up to 1000–1200 Da) tracers or fluorescent dyes. Although initial studies led to the conclusion that the junctional channel is an aqueous pore of large diameter and low selectivity, several authors reported that channels composed of different connexins (Cx) have different permeability properties. In fact, the rate of permeation of metabolites through gap junctions clearly differs according to the Cx composition. Therefore, the diffusion of metabolic compounds through astrocytic gap junctions needs to be addressed by using astrocytes, or at least cells that communicate through gap junctions composed of Cx43, since this is the main Cx expressed in astrocytes *in vivo* and *in vitro*.

The diffusion of various energetic metabolites through astrocytic gap junctions has been evaluated by adapting the scrape-loading dye transfer technique, using several radioactive metabolic compounds and following their intercellular spreading by autoradiography. This method, which must be performed with a confluent cell population, results in the loading of a large number of cells after they have been transiently opened to an extracellular medium devoid of calcium ions and containing a high concentration of the probe used to test gap junction permeability. After washing, if the junctional channels are functional, an intercellular diffusion of the dye is observed, provided the probe used does not leak through the plasma membrane to the outside. Such a technique is well suited for a membrane-impermeant dye such as Lucifer yellow, but has to be modified when metabolic compounds are used. Thus, to prevent the diffusion of metabolites through the plasma membrane, the experiments must be carried out at low temperature (4 °C) in the presence of a saturating concentration of the same nonradioactive metabolite. Moreover, after a given time (1–2 min) allowed for intercellular diffusion, preparations are fixed by drying the cells on a hot plate (100 °C). Prints of the distribution of the labeled compounds are finally obtained by exposing the dried coverslips to an X-ray-sensitive film, which is then photomicrographed and scanned for quantification (Figure 1).

Passage of Glutamine and Glutamate through Astrocytic Gap Junctions

Using the aforementioned technique, passage of glutamine and glutamate through astrocytic gap junctions has been achieved. In addition, passage of glutamate between Cx43-transfected C6 glioma cells has been documented. Diffusion of these metabolites

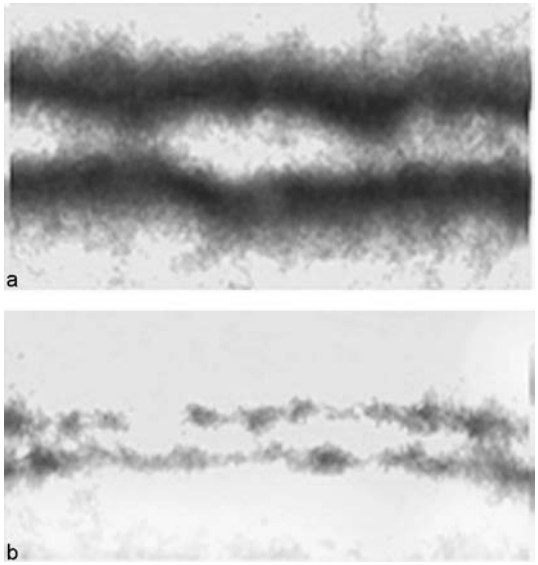

Figure 1 Adaptation of the scrape-loading dye transfer technique to measure the diffusion of radiolabeled substrates. (a) Passage of radiolabeled glucose-6-phosphate through astrocyte gap junctions. Astrocytes in primary culture growing on coverslips were scrape-loaded in the presence of 50 mM glucose and 20 μCi ml^{-1} of [^{14}C]U-glucose-6-phosphate. (b) Gap junctional communication was inhibited by the addition 100 μM endothelin-1 to the solutions used in the experiment. Dried coverslips were autoradiographed and the images were digitalized for quantification.

is inhibited by octanol or carbenoxolone, well-known inhibitors of gap junction permeability, suggesting that glutamine and glutamate can be distributed throughout the brain by passage through astrocytic gap junctions. In fact, glutamine is an excellent energetic substrate for both neurons and astrocytes in primary culture. Glutamine crosses the blood–brain barrier and its carbons are readily decarboxylated in the tricarboxylic acid cycle, consistent with the idea that glutamine is a very important energetic substrate. Although glutamine is a poor substrate for lipid synthesis in both neurons and astrocytes, its carbons can also be used to replenish tricarboxylic acid cycle intermediates in neurons. The passage of glutamine and glutamate through astrocytic gap junctions may also play an important role in the operation of a glutamine–glutamate cycle between neurons and astrocytes (Figure 2).

Passage of Glucose and Glucose-6-Phosphate through Astrocytic Gap Junctions

In agreement with the idea that astrocytic gap junctions are freely permeable to glucose, both radiolabeled glucose and radiolabeled 3-*O*-methylglucose, which cannot be metabolized, diffuse through gap junctions, indicating that it is glucose *per se* which

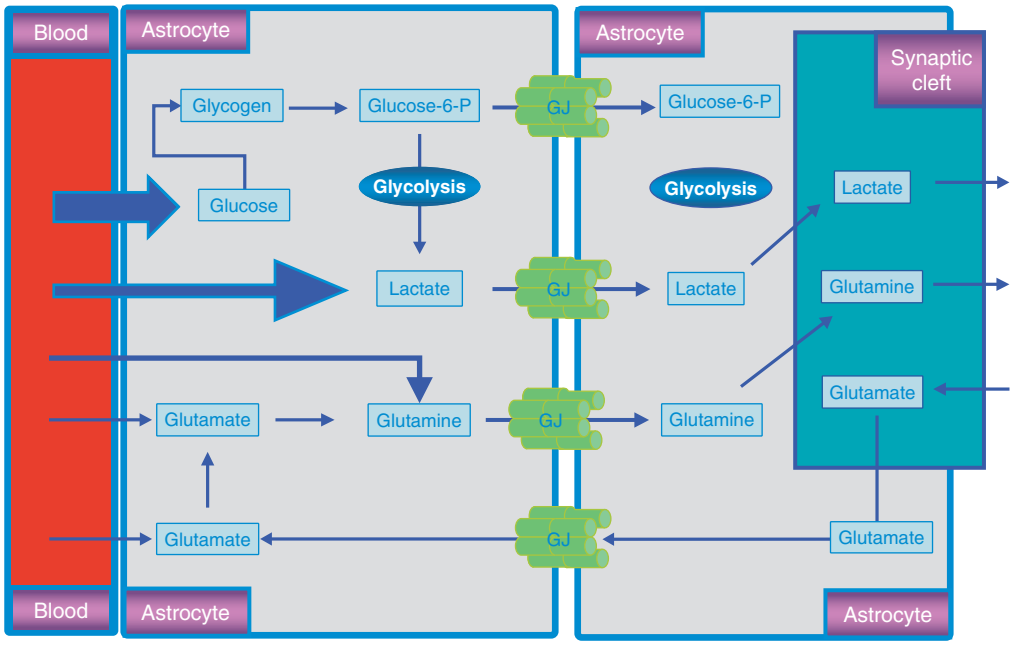

Figure 2 Syncytial metabolic operation of astrocytes coupled by gap junctions. Brain blood capillaries are surrounded by astrocyte processes through which energy substrates are taken up. To reach brain areas distant from the blood supply, energy substrates must cross astrocytic gap junctions (GJ) using an on/off mechanism that may be regulated by ATP concentrations through the ATP-sensitive potassium channel. Indeed, astrocytic gap junctions are permeable to glucose, glucose-6-phosphate (glucose-6-P), lactate, glutamate, and glutamine. Owing to the scarce presence of glucose-6-P in astrocytes, glucose-6-P from glycogen breakdown is converted into lactate, which can be used by neurons both as an energy substrate and as a lipid precursor. Astrocyte gap junctions may also participate in the glutamate/glutamine cycle that takes place between neuron and astrocyte.

passes through gap junctions. In addition, Goldberg and colleagues have reported radiolabeled-glucose diffusion through gap junctions in C6 glioma cells transfected with Cx43. More recently, Blomstrand and Giaume have shown that fluorescence-labeled glucose (2-[*N*-(7-nitrobenz-2-oxa-1,3-diazol-4-yl) amino]2-deoxy-D-glucose) can diffuse through astrocytic gap junctions. This molecule allows online visualization of glucose diffusion in living astrocytes. All of these results serve to demonstrate the passage of glucose through astrocytic gap junctions.

The first and compulsory step in glucose metabolism is its phosphorylation, forming glucose-6-phosphate. Interestingly, the phosphorylated form of glucose also crosses astrocytic gap junctions. Since the plasma membrane is not permeable to glucose-6-phosphate, gap junctions may be the only way to transfer glucose-6-phosphate between neighboring astrocytes. In conclusion, astrocytes that communicate through gap junctions can operate as a metabolic syncytium, since gap junctions are permeable to low molecular weight phosphorylated compounds that otherwise cannot pass through the plasma membrane (Figure 2).

Passage of Lactate through Astrocytic Gap Junctions

The main product from glucose and probably from glycogen metabolism in astrocytes is lactate, which is synthesized in both normal and pathological situations. Although both neurons and astrocytes can use lactate as a substrate to provide energy and carbon skeletons, an attractive hypothesis for this substrate has been proposed in which lactate is supplied by astrocytes to neurons for metabolic purposes, depending on synaptic activity. Astrocytic gap junctions are permeable to lactate, which may contribute to its diffusion through brain structures (Figure 2). This helps to explain why lactate is able to support neuronal activity.

Metabolic Exchange between Astrocytic Gap Junctions: Relevance for Distribution of Glycogen Breakdown Products

The passage of glucose-6-phosphate and lactate through gap junctions may play a role in the trafficking of glucose carbons between astrocytes. Although brain glycogen is mainly confined to astrocytes, these cells lack glucose-6-phosphatase activity. Therefore, glucose-6-phosphate must be confined within astrocytes. Since the plasma membrane is not permeable to phosphorylated compounds, glucose-6-phosphate can be channeled to adjacent astrocytes through gap junctions, or it can be transformed into lactate through the glycolytic pathway for release to neurons. Indeed, it has been reported that the main product of glycogen breakdown in cultured astrocytes is lactate, which is released into the medium, presumably for use by neurons. It should be noted that glycogen has been proposed as a source of energy for memory processes. If so, the aforementioned mechanism for trafficking glycogen products through brain structures should be considered very relevant (Figure 2).

Regulation of Astrocytic Gap Junctional Communication: Contribution to Regulation of Metabolic Syncytial Operation

Astrocytes have the equipment to receive, integrate, and transmit signals, since they express a broad variety of membrane ionic channels, transporters, and receptors linked to most known signal-transduction pathways. One consequence of the activation of several astrocytic membrane receptor subtypes is the up- or downregulation of gap junction permeability, which indicates that junctional communication in astrocytes may exhibit some degree of plasticity and remodeling. For instance, endothelin-1, a peptide produced by several brain cell types, including endothelial cells, astrocytes, and certain neurons, is a potent inhibitor of astrocytic gap junctions. The regulation of gap junctional communication between astrocytes promoted by these molecules can affect the passage of metabolites through gap junctions, because endothelin or its second messenger, arachidonic acid, inhibit gap junction permeability for glucose and its derivatives. This suggests that the passage of metabolites through astrocytic gap junctions can be regulated through a paracrine mechanism. In fact, gap junction permeability is controlled by the activity of the K-ATP channel, which is composed of two subunits, KIR (inwardly rectifying K^+ channel) and SUR (sulfonylurea receptor), the latter conferring the channel sensitivity to ATP, sulfonylureas, and to an endogenous ligand, endosulfine. The existence of the K-ATP channel has been reported in the central nervous system (CNS), and this channel is sensitive to sulfonylureas and to endosulfine. Actually, sulfonylureas significantly decrease the inhibitory effect of Ca^{2+} on gap junctional permeability in astrocytes, suggesting that the K-ATP channel may control permeability through gap junctions in astrocytes. Sulfonylureas still decrease the effect of Ca^{2+} in the presence of the Ca^{2+} ionophore A-23187, suggesting that the effect of the K-ATP channel on gap junction permeability is not associated with changes in cytosolic free-Ca^{2+} concentrations. Instead, membrane depolarization caused by the inhibition of the K-ATP channel would prevent the inhibition of gap junction permeability caused by Ca^{2+}, suggesting that the K-ATP channel may regulate gap junction permeability by controlling membrane potential through a

mechanism that is independent of, but mandatory in, the Ca^{2+} effect. This hypothesis is in agreement with the finding that the increase in external K^+ concentration, which leads to depolarization, increases gap junction permeability in astrocytes.

Because sulfonylureas mimic the effects of ATP on the K-ATP channel, it is proposed that intracellular ATP concentrations might regulate gap junction permeability controlling the K-ATP channel. Therefore, the energy status of the astrocyte would be able to regulate intercellular communication through gap junctions. When the energy status of the astrocyte is sufficient to sustain its own metabolic machinery, gap junction permeability would increase, to allow the passage of metabolites to adjacent cells. However, if the energy reserves of the astrocyte were compromised, ATP concentrations would decrease, resulting in a drop in gap junction permeability. Therefore, astrocytic gap junctions may play an important role in 'pipelining' metabolic substrates through the CNS, by acting like waterway locks that allow establishment of a cell-to-cell metabolite gradient, which then fuels the transport of substrates to targeted neurons. Accordingly, ATP concentrations may regulate the open/closed state of the 'locks' by controlling the activity of the K-ATP channel. Once the incoming substrates have fulfilled the energy requirements of the cells, gap junction permeability would increase, to allow spared metabolic substrates to pass to adjacent cells, thus distributing energy and carbon skeletons through the CNS (Figure 2). Therefore, it is reasonable to suggest that the K-ATP channel plays an important role in controlling the transport metabolites and signals through the CNS, thus regulating the crucial collaboration between astrocytes and neurons.

Besides these metabolic substrates, important endogenous metabolites have been reported to be able to diffuse through gap junctions. In particular, ATP and ADP can be transferred through Cx43 channels, indicating that gap junctional communication may contribute to the maintenance of the energy status of the astrocytic network.

Inhibition of Gap Junctional Communication Increases Uptake of Glucose in Astrocytes

Several lines of evidence suggest that the inhibition of gap junctional communication increases the uptake of glucose in astrocytes (Figure 3). Thus, structurally unrelated inhibitors of astrocytic gap junctional communication, such as endothelin-1 (ET-1), arachidonic acid, oleic acid, carbenoxolone (CBX), or α-glycyrrhetinic acid (AGA), promote an

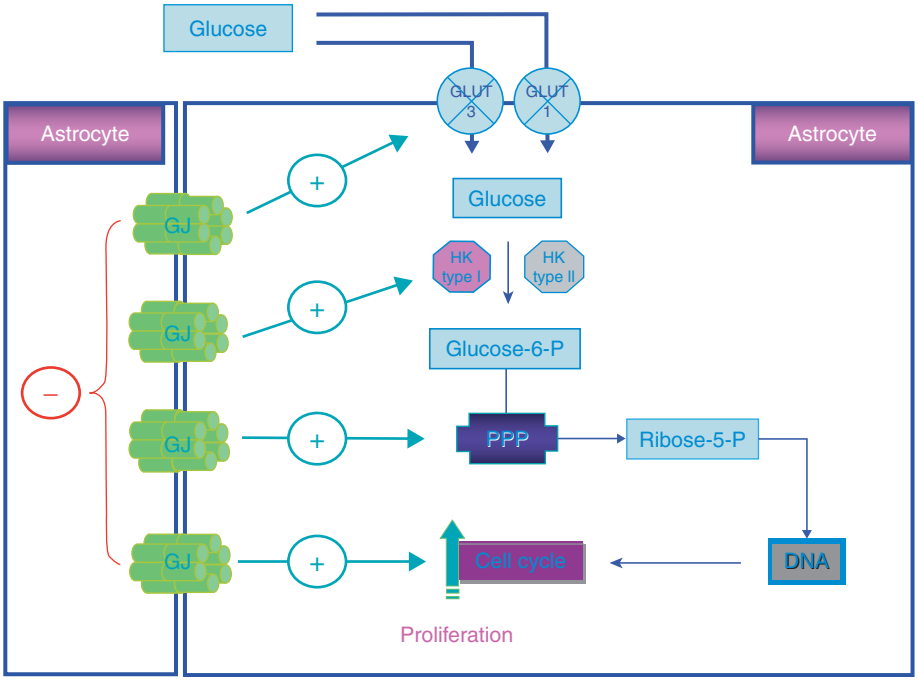

Figure 3 Gap junction (GJ) inhibition increases the rate of astrocyte proliferation. Gap junction inhibition increases glucose uptake by increasing the activity of glucose transporter 1 (GLUT-1) and type I hexokinase (HK). The increase in glucose uptake is maintained by the upregulation of GLUT-1 and type I HK and the appearance of different isoforms of glucose transporter (GLUT-3) and HK (type II). Glucose-6-phosphate (glucose-6-P) is then transformed to ribose-5-P through the pentose phosphate pathway (PPP), which is used for DNA synthesis to account for the increase in the rate of astrocyte proliferation.

increase in the rate of glucose uptake. The effect of these compounds on glucose uptake seems to be related to their ability to inhibit gap junctional communication. For instance, changes in their structure affecting their role in gap junctional communication, such as the esterification of oleic acid to form methyl-oleyl ester, abolish their role in glucose uptake. Other fatty acids, such as stearic acid, which does not affect gap junctional communication, do not increase glucose uptake either. The intensive coupling of astrocytes through gap junctions is required for the effect of the inhibitors of gap junctional communication on glucose uptake to be exerted, because their effect depends on the level of Cx43 expression and on the extent of gap junctional communication. In fact, in dissociated astrocytes, the gap junction channels of which are not functional, the inhibitors of gap junctional communication do not affect glucose uptake. Moreover, these compounds also fail to affect glucose uptake in C6 glioma cells, which are known to be poorly coupled through gap junctions. Finally, when the inhibition of gap junctional communication is prevented by tolbutamide, the inhibitors of gap junctional communication are not able to increase glucose uptake. Altogether, these results strongly suggest that the inhibition of gap junctional communication promotes an increase in the rate of glucose uptake in astrocytes. It can be speculated that the purpose of increasing the uptake of glucose is to compensate the decrease in the entry of glucose and other metabolites through astrocytic gap junctions.

Effect on Glucose Transporters

Glucose transport by astrocytes is mainly brought about through the energy-independent facilitative glucose transporter GLUT-1. The inhibition of gap junctional communication promotes a rapid translocation of GLUT-1 from an intracellular pool to the plasma membrane, which may be responsible for the increase in the V_{max} for glucose uptake found when gap junctional communication is inhibited. Similar changes in GLUT-1 distribution have been reported in response to some growth factors, indicating that this system may be designed to rapidly increase glucose uptake to meet cellular metabolic requirements.

In addition to this short-term regulation, the level of GLUT-1 is increased after 24 h of treatment with inhibitors of gap junctional communication, indicating that the inhibition of gap junctional communication also triggers a long-term regulation of glucose transport. In fact, GLUT-3, which is not normally present in astrocytes, although its expression can be induced under several circumstances, is induced after the inhibition of gap junctional communication. The upregulation of GLUT-3 has been associated with an increase in glucose uptake because this isoform exhibits a higher capacity for glucose transport than does GLUT-1. In fact, the K_m for glucose uptake is significantly reduced after treatment with the inhibitors of gap junctional communication.

Effect on Glucose Phosphorylation

The method of estimating glucose uptake with 2-deoxy-[U-^{14}C]glucose includes glucose entry and phosphorylation to 2-deoxy-[U-^{14}C]glucose-6-phosphate, which is confined inside the cell, where it is not further metabolized. Therefore, glucose uptake includes not only glucose transport but also glucose phosphorylation, which in the brain is mainly catalyzed by type I hexokinase. This enzyme can undergo rapid and reversible changes in subcellular compartmentation between mitochondria and the cytosol. In astrocytes, most of the type I hexokinase is associated with mitochondria. Interestingly, the inhibition of gap junctional communication promotes a rapid detachment of a significant proportion of type I hexokinase from mitochondria. This is rather unexpected, since it is generally accepted that the activity of hexokinase bound to mitochondria is greater than that of cytosolic hexokinase. However, the inhibition of gap junctional communication increases the translocation of type I hexokinase toward a specific subcellular location (i.e., to the microtubules). In this context, it has been reported that the binding of hexokinase to microtubules increases the activity of the enzyme. Consequently, the inhibition of gap junctional communication promotes a rapid translocation of type I hexokinase from mitochondria to microtubules, which could be responsible for the rapid increase in the activity of the enzyme. Thus, both glucose entry and phosphorylation are rapidly increased, probably to compensate the closure of an important route of metabolite supply (i.e., gap junctions). However, in addition to this rapid activation of the enzyme, longer exposure to inhibitors of gap junctional communication (24 h) induces a significant increase in the amount of type I hexokinase and the induction of type II hexokinase. It should be mentioned that type II hexokinase is not normally present in astrocytes but it is specifically induced in tumoral cells, contributing to their high rate of glycolytic activity.

Fate of Glucose after Inhibition of Gap Junctional Communication

As previously mentioned, prolonged inhibition of gap junctional communication in astrocytes induces the expression of isoforms characteristic of highly proliferative cells, such as GLUT-3 and type II hexokinase, suggesting that these phenotypic changes are designed to sustain a higher rate of cell proliferation.

In fact, analysis of the metabolic fate of glucose reveals that the inhibition of gap junctional communication specifically increases the flux of glucose through the pentose phosphate pathway (PPP). This pathway provides the ribose-5-phosphate needed for nucleic acid synthesis. In agreement with this, the inhibition of gap junctional communication increases the synthesis of nucleic acids from glucose, because glucose incorporation into both DNA and RNA increases (Figure 3).

Prolonged Inhibition of Gap Junctional Communication Induces Astrocyte Proliferation

For decades, gap junctions have been proposed to be required for an adequate rate of cell proliferation. In fact, the inhibition of gap junctional communication increases astrocyte proliferation (Figure 3). Thus, the presence of several inhibitors of gap junctional communication almost doubles the rate of 5-bromo-2-deoxyuridine (BrdU) incorporation in astrocytes. This effect is abolished by preventing the inhibition of gap junctional communication with tolbutamide, or when astrocytes are poorly communicated through gap junctions (i.e., only after 10 days in culture or in astrocytes after 21 days in culture, replated at half the density to avoid the establishment of gap junctional communication).

The percentage of astrocytes found in the S, G2, and M phases increases significantly after the inhibition of gap junctional communication, with a concomitant decrease in the G0 and G1 phases. In addition, the levels of Ki-67, a protein present during all active phases of the cell cycle but absent from resting cells, increases after the inhibition of gap junctional communication, indicating that the inhibition of gap junctional communication promotes the entry of astrocytes into the cell cycle.

In G0 and early G1, retinoblastoma protein (pRb) exists in a hypophosphorylated state but is sequentially phosphorylated on several residues; as a consequence, the cell passes through G1 and enters the S phase. The inhibition of gap junctional communication promotes the phosphorylation of pRb on Ser780. It should be mentioned that Ser780 has been identified as belonging to the sites that are initially phosphorylated in the G1 phase. Once pRb has been phosphorylated on Ser780 it loses its ability to bind and inhibit E2F, a family of transcription factors that activate the genes required for progression from the G1 to the S phase. pRb is phosphorylated on Ser780 by Cdk4/6, which is modulated positively by D-type cyclins. As a consequence of the inhibition of gap junctional communication, the levels of cyclins D1 and D3 are increased, suggesting that these cyclins may be responsible for activating the phosphorylation of pRb on Ser780 and hence for the transition of astrocytes from the G0/G1 to the S phase observed after the inhibition of gap junctional communication. Together, these results provide evidence for the involvement of cyclins D1 and D3 as sensors of the inhibition of gap junctional communication in astrocytes and suggest that they could be mediators in the mitogenic effect observed in these circumstances. These findings also suggest that D-type cyclins, apart from their classical role as sensors of growth factors, can respond to changes in the cellular context, such as the interruption of cell-to-cell communication.

Increase in Gap Junctional Communication Is Associated with a Decrease in Both Cell Proliferation and Glucose Uptake in Glioma Cells

Gliomas are the most common brain tumors and are generally highly malignant, with a poor prognosis. The neoplastic phenotype of gliomas and several other tumors includes the absence of gap junctional communication. Thus, the expression of Cx43, the main protein forming gap junction channels in astrocytes, correlates inversely with the degree of glioma malignancy. In addition, when Cx43 is overexpressed in glioma cells, the rate of cell growth decreases, indicating that Cx43 behaves as a tumor suppressor protein for gliomas. In this sense, tolbutamide increases gap junctional permeability and Cx43 expression in poorly coupled C6 glioma cells. This effect was similar to that found with dbcAMP, a well-known activator of gap junctional communication in tumor cells. Furthermore, the increase in communication promoted by tolbutamide or dbcAMP is concurrent with the inhibition of the proliferation of C6 glioma cells (Figure 4). Thus, tolbutamide significantly decreases the rate of C6 glioma cell proliferation, estimated by BrdU incorporation, and this effect is additive with that of dbcAMP. In addition, the percentage of cells in different cell cycle phases is affected by these two agents. In fact, tolbutamide and dbcAMP decrease the level of Ki-67, a protein expressed when cells are proliferating. These effects are accompanied by a reduction in the phosphorylation of pRb, mainly on Ser795, a residue critical for the control of cell proliferation. The decrease in the phosphorylation of pRb is probably due to the upregulation of the Cdk inhibitors p21 and p27. When Cx43 is silenced by siRNA, neither tolbutamide nor dbcAMP is able to upregulate p21 and consequently to reduce glioma cell proliferation, as judged by Ki-67 expression. In

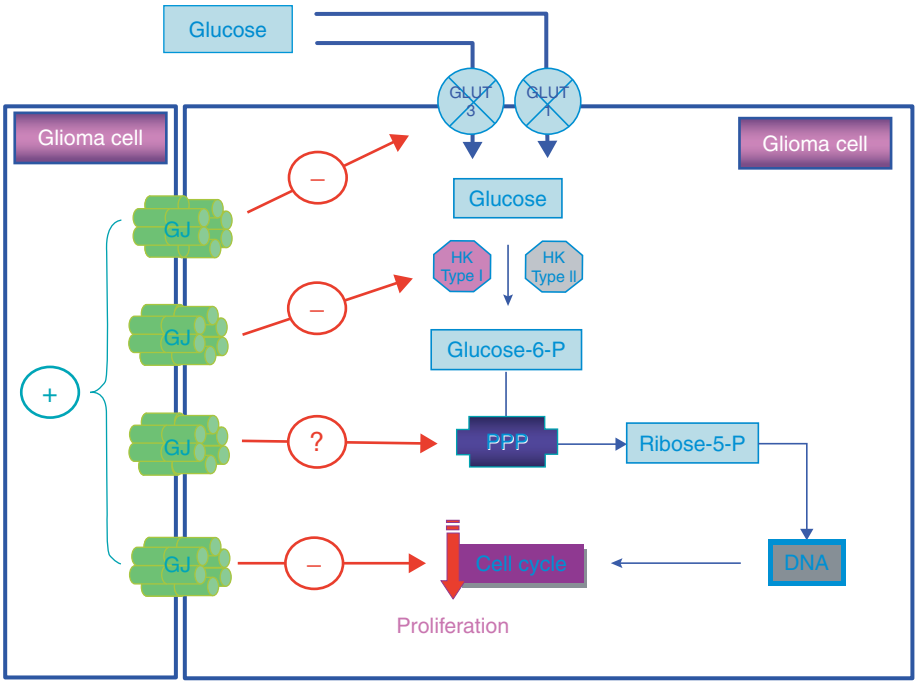

Figure 4 The enhancement of gap junctional (GJ) communication decreases the rate of C6 glioma cell proliferation. The enhancement of gap junctional communication decreases glucose uptake by decreasing hexokinase (HK) activity. Concomitantly, the rate of C6 glioma cell proliferation declines (PPP, pentose phosphate pathway).

conclusion, tolbutamide and dbcAMP inhibit C6 glioma cell proliferation by increasing Cx43, which correlates with a reduction in pRb phosphorylation due to the upregulation of the Cdk inhibitors p21 and p27.

The ability to take up glucose at high rates is one of the most characteristic biochemical features of neoplastic cells and transformed phenotypes. The high glycolytic rate and increased glucose uptake by tumors were first reported by Warburg. Increased glucose uptake by tumor cells is exploited in cancer imaging using the positron-labeled glucose analog [^{18}F]fluoro-2-deoxy-D-glucose (FDG), which is transported into cells and then phosphorylated, but undergoes little further metabolism. The FDG accumulated can then be detected using positron emission tomography (PET). Several reports have indicated that FDG PET assesses the degree of tumor malignancy, since high-grade gliomas exhibit a higher glucose uptake than do low-grade tumors. In agreement with this, tolbutamide and dbcAMP, in addition to decreasing glioma cell proliferation, trigger a rapid decrease in the rate of glucose uptake. In addition, the effects of tolbutamide and dbcAMP on the rate of glucose uptake are additive, like their effects on gap junctional communication. Consequently, the increase in gap junctional communication is associated with a decreased rate of glucose uptake.

In C6 glioma cells, glucose phosphorylation is mainly catalyzed by type I and type II hexokinase. This enzyme can undergo rapid and reversible changes in subcellular compartmentation between the mitochondria and the cytosol. In C6 glioma cells, most type I and type II hexokinase is found associated with mitochondria. Interestingly, the presence of tolbutamide or dbcAMP promotes the rapid detachment of a significant proportion of type I and type II hexokinase from the mitochondria to the cytosol, together with a decrease in hexokinase activity. This is in agreement with several reports that the activity of hexokinase is higher when the enzyme is bound to mitochondria than when it is soluble in the cytosol. This seems to be the mechanism through which tolbutamide and dbcAMP decrease glucose uptake, since the activity of glucose transporters is not modified. In agreement with this, one of the mechanisms involved in regulating hexokinase activity in tumor cells is the shift of hexokinase between the cytosol and mitochondria. In fact, a strong correlation between growth rate and mitochondrial hexokinase activity has been reported for several types of tumors. In tumor cells, the release of hexokinase from mitochondria is associated with the inhibition of cell proliferation and is therefore a target for anticancer therapy. Taken together, these data show that the

increase in gap junctional communication in glioma cells reduces the exacerbated rate of glucose uptake characteristic of tumor cells by releasing hexokinase from the mitochondria. This is concurrent with a decrease of the rate of proliferation.

Summary

It is reasonable to suggest that astrocytic gap junctional communication is an important element for maintaining brain metabolic homeostasis. In fact, the inhibition of gap junctional communication induces the increase in the uptake of glucose through specific transporters in the plasma membrane, probably to compensate the lack of the diffusion of glucose and other substrates through gap junctions. However, prolonged inhibition of gap junctional communication induces an alteration of the astrocytic phenotype toward the phenotype of a highly proliferative cell. In fact, gliomas are an example of a pathology in which glial cells lose their ability to communicate through gap junctions. The upregulation of Cx43 in glioma cells decreases both cell proliferation and the exacerbated rate of glucose uptake, indicating that these events are closely linked. Knowledge of cellular and molecular events that associate them should contribute to a better understanding of the role of astrocytes in brain dysfunction and should lead to the identification of new targets for reversing pathological situations.

See also: Astrocyte: Calcium Signaling; Astrocyte: Neurotransmitter and Hormone Receptors; Astrocyte: Identification Methods; Gap Junctions and Hemichannels in Glia; Gap Junctions and Neuronal Oscillations; Gap Junctions and Electrical Synapses; Gap Junction Abnormalities and Disorders of the Nervous System; Gap Junction Communication; Neurotransmitter Release from Astrocytes.

Further Reading

Dringen R, Gebhardt R, and Hamprecht B (1993) Glycogen in astrocytes, possible function as lactate supply for neighboring cells. *Brain Research* 62: 208–214.

Giaume C and McCarthy KD (1996) Control of gap-junctional communication in astrocytic networks. *Trends in Neuroscience* 19: 319–325.

Giaume C, Tabernero A, and Medina JM (1997) Metabolic trafficking through astrocytic gap junctions. *Glia* 21: 114–123.

Goldberg GS, Valiunas V, and Brink PR (2004) Selective permeability of gap junction channels. *Biochimica et Biophysica Acta* 23: 96–101.

Huang R-P, Hossain M, Sehgal A, et al. (1999) Reduced connexin43 expression in high-grade human brain glioma cells. *Journal of Surgical Oncology* 70: 21–24.

Kettenmann H and Ransom BR (1995) *Neuroglia*. Oxford: Oxford University Press.

Loewenstein W (1981) Junctional intercellular communication: The cell-to-cell membrane channel. *Physiological Reviews* 61: 829–913.

Medina JM and Tabernero A (2005) Lactate utilization by brain cells and its role in CNS development. *Journal of Neuroscience Research* 79: 2–10.

Mugnaini E (1986) Cell junctions of astrocytes, ependyma and related cells in the mammalian central nervous system, with emphasis on the hypothesis of a generalized functional syncytium of supporting cells. In: Fedoroff S and Vernadakis A (eds.) *Astrocytes*, 1, pp. 329–371. New York: Academic Press.

Pedersen PL, Mathupala S, Rempel A, et al. (2002) Mitochondrial bound type II hexokinase: A key player in the growth and survival of many cancers and an ideal prospect for therapeutic intervention. *Biochimica et Biophysica Acta* 1555: 14–20.

Pellerin L and Magistretti PJ (2004) Neuroenergetics: Calling upon astrocytes to satisfy hungry neurons. *Neuroscientist* 10: 53–62.

Sánchez-Alvarez R, Paino T, Herrero-González S, Medina JM, and Tabernero A (2006) Tolbutamide reduces glioma cell proliferation by increasing conexin 43, which promotes the up-regulation of p21 and p27 and subsequent changes in retinoblastoma phosphorilation. *Glia* 54: 125–134.

Spray DC (1994) Physiological and pharmacological regulation of gap junction channels. In: Citi S (ed.) *Molecular Mechanism of Epithelial Cell Junctions: From Development to Disease*, pp. 195–215. Austin, TX: Landes Company.

Tabernero A, Giaume C, and Medina JM (1996) Endothelin-1 regulates glucose utilization in cultured rat astrocytes by controlling intercellular communication through gap junctions. *Glia* 16: 187–195.

Tabernero A, Sanchez-Alvarez R, and Medina JM (2006) Increased levels of cyclins D1 and D3 after inhibition of gap junctional communication in astrocytes. *Journal of Neurochemistry* 96: 973–982.

Yamasaki H and Naus C (1996) Role of connexin genes in growth control. *Carcinogenesis* 17: 1199–1213.

Gastrointestinal Signals: Satiety

T H Moran, The Johns Hopkins University School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In many mammalian species, food intake occurs in distinct bouts or meals. Although meal initiation is determined by a variety of factors, especially food availability, meal termination is primarily the result of feedback signaling from the gastrointestinal (GI) tract. During a meal, ingested nutrients stimulate the activation or release of a variety of afferent neural and peptide signals. Many of these signals engage mechanisms that facilitate nutrient GI transit and nutrient digestion and absorption. A subset of these provide information about the quality and quantity of ingested nutrients and serve as feedback mediators controlling food intake within and between meals. Meal size can be highly variable and alterations in meal size appear to be a major determinant of overall food intake.

The clearest evidence for GI feedback controls on food intake comes form experiments involving sham feeding. Various models have been used, but the most common involves a gastric fistula that can be opened or closed (Figure 1). When closed, ingested food collects in the stomach and is gradually emptied into the upper intestine. Intake stops at a point where feedback inhibition exceeds any signals that are driving intake. When the fistula is opened, ingested nutrients drain from the stomach at the same rate that they are consumed. With such sham feeding, food does not accumulate in the stomach and does not pass into the upper intestine. Under these circumstances, feeding continues for much longer time intervals and much larger 'meals' are consumed. The fact that sham feeding does eventually stop suggests pregastric contributions to satiety. The amount sham fed depends upon the degree of experience with the sham-feeding paradigm. Although the first time rats sham feed, their intakes double, continued experience with sham feeding significantly increases intake. These data demonstrate the presence of a conditioned inhibition on food intake that is due to an association between oral stimulation and postingestion negative feedback. Only with continued experience is this conditioned inhibition on intake overcome. With experience, sham-fed meals can be 6–10 times the size of meals consumed with the fistula closed.

In normal ingestion, consumed nutrients contact mechano- and chemosensitive receptors in the stomach and the intestine. The potential range of feedback mechanism that could contribute to meal termination is dependent upon the distribution of ingested nutrients during the meal. Studies of gastric emptying suggest that this distribution during a meal is extensive. In both rats and monkeys, filling the stomach at rates mimicking normal ingestion results in rates of gastric emptying that are much more rapid than those that occur following the filling period. Such data demonstrate that a significant portion of ingested nutrients (in the rat as much as 30%) enters the duodenum, contacts duodenal receptors, and is available for absorption during the meal. Thus, the stomach and a significant proportion of the upper intestine are potential sites for within-meal feedback signal generation.

Neural Afferent Signaling

The vagus (X cranial) nerve is the major neuroanatomical link between the GI tract and brain. Vagal afferent fibers with cell bodies in the nodose ganglion innervate the stomach and various intestinal segments. In the rat, afferent fibers constitute the majority of vagal fibers, accounting for greater that 75% of subdiaphragmatic vagal nerve fibers. The subdiaphragmatic vagus has two major trunks that travel along the esophagus. In the rat, the ventral trunk bifurcates into the left gastric, the hepatic, and the accessory celiac branches. The dorsal trunk bifurcates into the right gastric and the celiac branches. The nomenclature implies a specificity of innervation. However, branches often innervate additional organs. For example, the hepatic branch provides the major innervation to the proximal duodenum and some duodenal innervation also derives from the gastric branches. Although there are also significant spinal gut neural connections, the response properties of this system have not been well characterized.

The response properties of vagal afferents depend in part upon the target organ from which they arise. Vagal afferents innervating the stomach have terminations in the gastric musculature. These afferents are primarily mechanosensitive – that is, they increase their firing rates in response to increasing gastric load volume. There are multiple mechanoreceptive response patterns. Slowly adapting mechanoreceptive fibers increase their response rate with increasing gastric volume. The fibers remain active while load volume

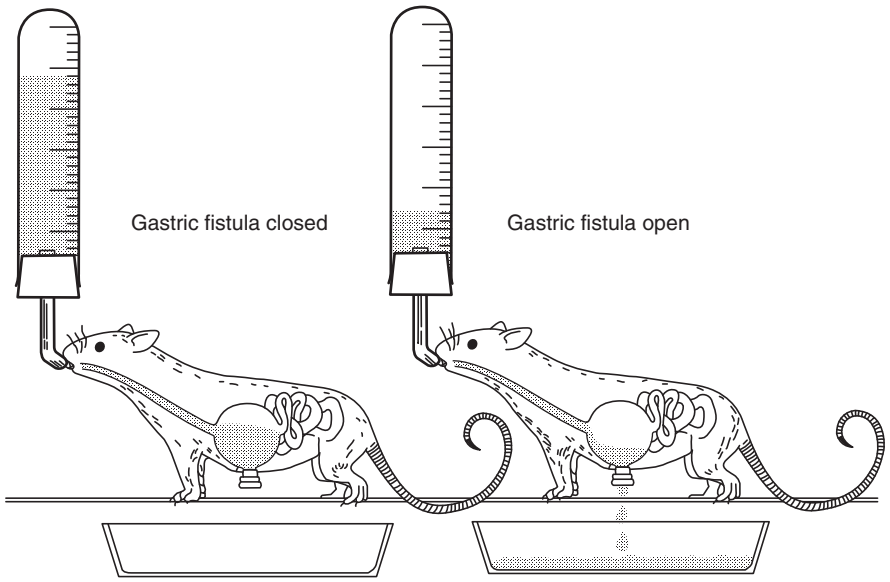

Figure 1 Increased food intake in sham feeding. Reproduced from Smith GP, Gibbs J, and Young RC (1974) Cholecystokinin and intestinal satiety in the rat. *Federation Proceedings* 33: 1146–1149, with permission.

is retained and show an off response in which activity briefly drops below baseline levels when the load volume is removed. Individual afferents are differentially tuned such that there are differences in their dynamic range. Some afferents reach their maximal activity at small intragastric volumes while others do not begin to respond until a significant gastric load is present. Another class of mechanosensitive fibers increase their firing rate during a volume increase but then return to baseline rates when the new load is achieved. Gastric mechanoreceptive vagal afferents do not respond directly to the chemical character of the gastric load. Response rates are similarly increased by nutrient and nonnutrient load volumes. Although gastric vagal activity is not directly responsive to intragastric nutrient character, gastric afferent responsivity can be altered by duodenal nutrient. Thus, the degree to which a gastric load activates gastric vagal afferent activity is modulated in the presence of duodenal nutrients.

Duodenal vagal afferents are activated by both intraluminal load volume and nutrient character. Duodenal slowly adapting mechanoreceptive fibers have been identified. Similar to gastric mechanoreceptive fibers, activity increases with increases in load volume. Other duodenal vagal afferents are directly responsive to nutrient character. For example, both intestinal casein and lipid infusions have been shown to result in increases in vagal afferent activity. Vagal afferent terminations have been identified within the duodenal mucosa. Although fine afferent fibers have been localized within the intestinal villa, the terminations do not appear to penetrate into the duodenal lumen, leading to the view that any direct nutrient stimulation would be postabsorptive rather than preabsorptive. Alternatively, nutrient absorption could stimulate local chemical release and activate vagal afferents secondary to such release.

Vagal afferent fibers project to the nucleus of the solitary tract (NTS) in the dorsal hindbrain following a rough viscerotopic representation of the alimentary canal. Activation of vagal afferent fibers results in a modulation of activity of NTS neurons. This can be shown electrophysiologically or through the detection of immediate activation of early genes such as c-*fos*. Just as the presence of intestinal nutrients can modulate the mechanoreceptive response to gastric load volume, meal-induced activation within the NTS reflects response integration. Studies that have provided stimulation to individual or multiple GI compartments have demonstrated that nutrient stimulation of one GI compartment can either enhance or reduce the activation produced by stimulation of another compartment.

Elimination of aspects of vagal afferent feedback can result in significant alterations in the way that rats pattern their food intake. Surgical vagal deafferentation results in alterations in meal patterns in rats maintained on a liquid diet, in that such rats consume larger meals than sham-operated controls do. Meal frequency is reduced in response to these meal size increases such that overall food intake is unchanged. Similar alterations in meal size have been reported in response to capsaicin-induced chemical vagal deafferentation. Although these data demonstrate a role

for vagal afferent feedback in the controls of meal size, the particular signal(s) that is being blocked by the vagotomy procedures has yet to be identified.

Within-Meal Peptide Satiety Signaling

As well as resulting in the stimulation of vagal afferent fibers, nutrients in the GI tract stimulate the release of a number of gut peptides that play various roles in food intake. Different peptides have different patterns of release during and following food intake and the patterns of release are consistent with their proposed roles in controlling the size of a current or in influencing subsequent meals.

A number of peptides are released into circulation at the onset of eating or shortly thereafter. A few of these have been demonstrated to play feedback roles in the control of the size of the meal that stimulates their release. Among these are the intestinal peptide, cholecystokinin (CCK), and the pancreatic peptides, glucagon and amylin (Figure 2). These peptides share a number of properties, including relative brief durations of action, the ability to produce reductions in meal size following their exogenous administration, documented roles for the endogenous peptide in limiting meal size, and signal transduction involving either the vagus nerve and/or the dorsal hindbrain.

The best characterized of these peptides is CCK. Exogenously administered CCK was originally demonstrated to decrease food intake over 30 years ago. The feeding inhibitory actions of CCK and CCK agonists have been demonstrated in a variety of species, including nonhuman primates and humans. Exogenously administered CCK reduces meal size and results in an earlier appearance of a behavioral satiety sequence. Consistent with its rapid release from the proximal intestine and short duration of effective bioactivity, the actions of CCK are specific to terminating an individual meal. Repeated, meal-contingent CCK administration consistently reduces meal size without producing a significant change in overall daily food intake.

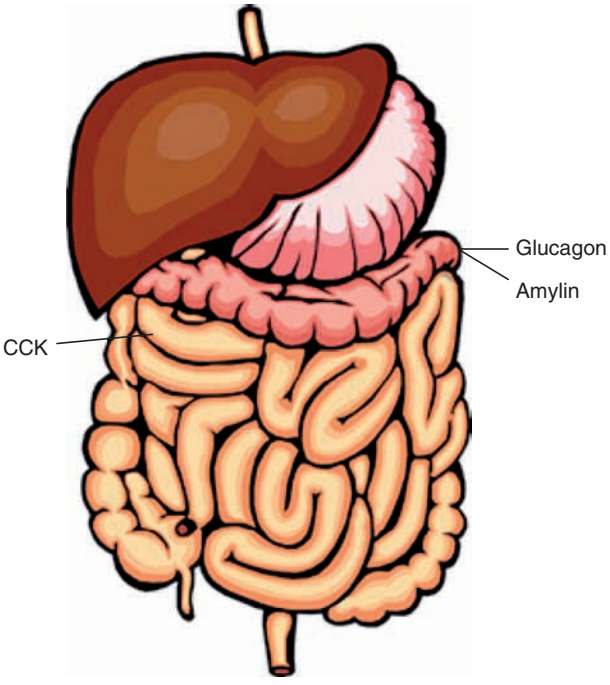

Figure 2 Within-meal gut peptide feedback (CCK).

The satiety actions of CCK depend upon interactions with multiple receptor sites. Local arterial CCK administration aimed at the proximal gut results in increases in vagal gastric mechanoreceptive afferent activity similar to those produced by intragastric load. Combinations of gastric load and CCK produce greater degrees of activity compared to either load or CCK alone. CCK also modifies responses to subsequent intragastric load such that load volume results in greater degrees of activity following CCK administration than prior to it, even at times when the initial response has disappeared. Duodenal vagal afferents also are activated by CCK, and combinations of load and CCK combine to produce greater duodenal vagal afferent activity compared to either load or CCK alone, and CCK affects the response to subsequent load volumes. CCK also plays a role in the response of duodenal afferents to nutrients. Administration of a CCK antagonist blocks the increase in vagal afferent activity produced by intraduodenal casein. CCK-induced changes in gastric vagal afferent activity appear to result from a direct action of the peptide on the vagal afferent fibers. Vagal afferents contain CCK receptors, and CCK induces decreases in intragastric pressure that by themselves decrease rather than increase vagal afferent activity.

Surgical or chemical disruption of subdiaphragmatic vagal afferent innervation significantly affects the ability of CCK to inhibit food intake. The nature of this disruption is a rightward shift in the CCK satiety dose-response curve. Low doses of CCK that inhibit food intake in intact rats are ineffective following vagal afferent lesions. Higher doses inhibit intake, but to a smaller degree. CCK's feeding inhibitory actions also depend upon interactions with pyloric CCK receptors, and the effect appears to be additive to that relayed through vagal CCK receptors. In contrast to the effect of vagotomy, surgical removal of the pyloric sphincter does not affect the ability of low doses of CCK to inhibit intake, but instead truncates the dose-effect curve such that the additional suppression that normally accompanies higher CCK doses is eliminated. Results such as these have led to the proposal that the satiety actions of CCK have

multiple mediations and are, in part, secondary to its local GI effects.

A role for endogenous CCK in satiety is supported by data demonstrating that administration of CCK antagonists with specificity for the CCK1 receptor results in increases food intake. In the primate, the effects have been demonstrated to be dose related, with a maximum increase of around 40% in daily food intake in monkeys with 4-h daily food access. This increase is almost completely accounted for by an increase in the size of their first meal. Alterations in meal patterns are also evident in rats lacking CCK1 receptors. Otsuka Long–Evans Tokushima Fatty (OLETF) rats, which have a ~6-kb deletion in the CCK1 receptor gene spanning the promoter region and the first and second exons, lack functional CCK1 receptors. Consistent with this deficit, OLETF rats consume significantly larger meals and do so when maintained on either solid or liquid nutrients.

Satiety actions for the pancreatic peptides glucagon and amylin have also been demonstrated. Eating rapidly elicits an increase in pancreatic glucagon secretion. Because an increase in plasma glucagon is also stimulated by sham feeding, this appears to be a cephalic-phase response. Glucagon is rapidly cleared from the circulation by the liver and the liver appears to be the site of glucagon's satiety action. Hepatic portal infusions of glucagon at meal onset elicit dose-related reductions in meal size. Consistent with a hepatic portal site of action, glucagon-induced feeding inhibition is prevented by transection of the vagal afferent fibers that innervate the liver. In contrast to CCK, glucagon's satiety action requires the presence of other forms of ingestional consequences. CCK inhibits sham feeding but glucagon does not. A role for endogenous glucagon in the control of meal size is supported by data demonstrating the ability of hepatic portal infusions of a glucagon antibody to increase meal size.

Amylin inhibits feeding in a dose-dependent and behaviorally specific manner following either peripheral or central administration. Amylin is obligatorily cosecreted with insulin by pancreatic β cells. Thus, amylin levels rise rapidly with meal onset and remain elevated for a significant period of time during and following meals. Similar to CCK and glucagon, meal-contingent amylin administration reduces meal size and meal duration without affecting intake in subsequent meals. A physiological role for endogenous amylin in feeding controls is supported by experiments demonstrating increases in food intake in response to administration of amylin antagonists. Although overall food intake of amylin knockout mice does not differ from controls, whether meal patterns are altered in these mice has not been reported. However, the anorectic effect of CCK was abolished in amylin knockout mice and this effect was reversed by the administration of doses of amylin that by themselves did not affect food intake. These data have been interpreted to suggest a role for endogenous amylin in the satiety actions of CCK.

In contrast to the satiety actions of CCK and glucagons, amylin feeding inhibitory effects are not vagally dependent. Amylin's feeding inhibitory actions are mediated by the area postrema (AP), a circumventricular organ with a porous blood–brain barrier that lies adjacent to the NTS, the site of vagal afferent terminations. The AP contains amylin receptors, AP lesions block the ability of amylin to inhibit eating and amylin administration activates cells in both the AP and underlying NTS. Mediation through the NTS is a shared property of all three peptides. NTS lesions block the ability of both CCK and glucagon to inhibit food intake. This similar hindbrain mediation suggests that the three peptides may be interacting with a common substrate at this level of the neural axis.

Across-Meal Peptide Signaling

Recent work has demonstrated the ability of a number of additional gut peptides to inhibit food intake following their peripheral administration, and the dynamics of their nutrient-stimulated release and feeding effects are consistent with actions that may extend beyond the individual meal (Figure 3). Thus, peptides such as glucagon-like peptide-1 (GLP-1),

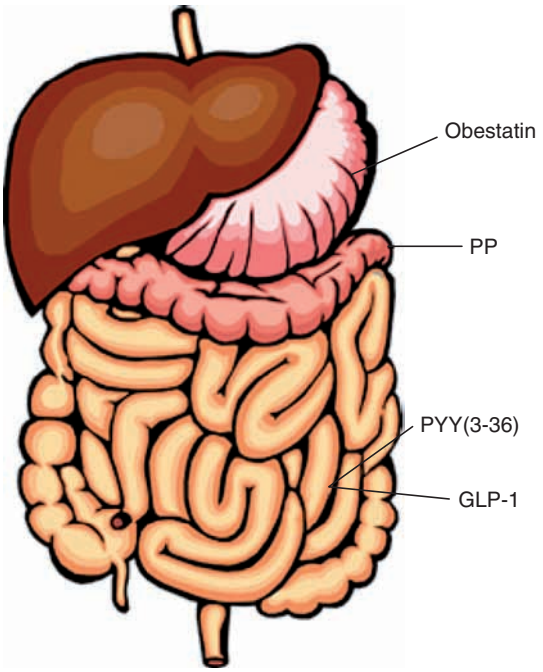

Figure 3 Across-meal gut peptide feedback (PP, pancreatic polypeptide; PYY(3-36), peptide YY(3-36); GLP-1, glucagon-like peptide-1).

pancreatic polypeptide (PP), and peptide YY (PYY) (3-36) have actions in feeding that do not appear to solely depend upon terminating the individual meal that stimulated their release.

GLP-1 is a posttranslational product of proglucagon that is released from L cells in the distal intestine upon ingestion of nutrients. Circulating levels of GLP-1 rise within minutes of food ingestion and remain elevated for over 3 h. The rapid rise in plasma GLP-1 following meal initiation makes it unlikely that this is a local response to nutrients in the ileum. Duodenal nutrients likely stimulate endocrine and neural mediators that activate GLP-1 secretion. Consistent with this view, pharmacological or surgical vagotomy significantly attenuates meal-stimulated increases in GLP-1 secretion. GLP-1 functions to help regulate blood glucose. Exogenous administration of GLP-1 reduces blood glucose in both normal and diabetic individuals by stimulating insulin synthesis and secretion and by suppressing glucagon secretion. GLP-1 is also involved in the ileal brake. Gastric inhibitory and intestinal motility actions of ileal nutrients are in part mediated through GLP-1 release. In humans, peripheral GLP-1 infusions have been reported to reduce food intake in lean, obese, and type 2 diabetic patients. However, peripheral bolus injections of GLP-1 into the rat have little or no effect on food intake. The short half-life of peripheral GLP-1 provides a potential explanation for the inability of exogenous GLP-1 to elicit an effect on food intake in the rat. Prolonged GLP-1 infusions prior to and during meal periods have been shown to reduce amounts consumed. The availability of a peptide GLP-1 agonist with a longer duration of action has allowed a more complete assessment of the effects of increased GLP-1 signaling on food intake. Peripheral bolus administration of such a compound potently suppresses food intake and does so for extended time intervals. The role of endogenous GLP-1 in feeding control has yet to be completely assessed. Although central administration of a GLP-1 antagonist has been shown to increase food intake, similar findings for peripheral administration have not been reported.

The mechanisms underlying the feeding inhibitory actions of intestinally derived GLP-1 are beginning to be understood. Vagal afferent denervation abolishes the actions of central and peripheral GLP-1 on gastric emptying, and vagotomy has recently been demonstrated to also prevent the feeding inhibitory actions of GLP-1. A second possible mode of action for peripheral GLP-1 to exert its effects is via specific carriers or endothelial leaks that allow peripheral GLP-1 access to central binding sites. Evidence supporting this idea comes from data demonstrating that when radiolabeled GLP-1 is injected peripherally, it can be identified in the AP and other circumventricular organs. These data suggest that peripheral GLP-1 gains access to dorsal hindbrain sites to inhibit food intake. Consistent with this view, administration of GLP-1 to the hindbrain does inhibit food intake.

PP is a member of the neuropeptide Y (NPY) and PYY family of peptides that interacts primarily the Y4 receptor subtype. PP is produced in endocrine cells in the pancreas and released into the circulation in response to food intake. The release of PP has multiple components with an early cephalic peak followed by a prolonged plateau. PP exerts a variety of regulatory actions in the GI tract, including inhibition of pancreatic exocrine secretion, gall bladder contraction, modulation of gastric acid secretion, and stimulation of GI motility. A role for PP in food intake was first suggested in studies in mice demonstrating that twice daily peripheral PP injections into obese mice reduced food intake and body weight. Recent work has suggested a role for PP in both food intake and energy expenditure and identified PP-induced decreases in the expression of NPY, orexin, and ghrelin as potential mechanisms for these actions. A vagal mediation has been suggested based on experiments demonstrating PP-induced increases in vagal afferent activity and the blockade of the feeding inhibitory actions by vagotomy. Exogenous PP administration has also been shown to result in reductions in food intake in human subjects. A role for endogenous PP in feeding control has not yet been demonstrated. Paradoxically, combined Y2/Y4 receptor knockout protects against high-fat-diet-induced obesity.

PYY and PYY(3-36) are also members of the NPY family of peptides; like GLP-1, they are secreted from I cells in the distal intestine. PYY and PYY(3-36) are synthesized through differential posttranslational processing of proPYY. The peptides are found throughout the GI tract, with the highest concentrations found at distal small and large intestinal sites. PYY(3-36) has been shown to be released into the plasma in response to a meal. Although little work has been done on the dynamics of PYY(3-36) release, its pattern of release has been inferred from data that measured postprandial plasma levels of PYY. The temporal pattern of PYY release is different from that of other GI peptides with suggested roles in feeding control. Plasma levels of PYY gradually rise following meal initiation, reaching a peak at about 60 min, and high levels are maintained for a number of hours following a meal. Such a pattern of release suggests the potential for a role in the control of food intake across meals rather than in meal termination. PYY has long been identified as a peptide involved in the ileal brake. PYY inhibits gastric emptying and mediates the effects of ileal fat on gastric motility.

Batterham and colleagues first reported significant suppression of food intake in rodents in response to peripheral bolus PYY(3-36) administration and significant decreases in food intake in humans in response to PYY(3-36) infusions. Originally, peripheral PYY(3-36) was proposed to affect feeding by directly inhibiting hypothalamic arcuate nucleus NPY-containing neurons by activating presynaptic Y2 receptors. Consistent with this proposal, peripheral PYY(3-36) was shown to decrease arcuate NPY mRNA expression. However, recent work supports the view that rather than directly accessing brain sites, PYY(3-36) may inhibit feeding through alterations in vagal signaling. The vagus contains Y2 receptors, and PYY(3-36) administration activates gastric vagal afferent fibers. Furthermore, vagotomy blocks the ability of peripheral administration of PYY(3-36) to inhibit eating and to activate NPY neurons.

It is not yet clear whether feeding inhibition is a physiological action of meal-induced PYY(3-36). There has been some controversy over the reproducibility of PYY(3-36)-induced feeding inhibition in rodents. In addition, doses of PYY that inhibit food intake in mice also produce a conditioned taste aversion. Finally, infusions that reduce food intake in primates produce plasma levels greatly in excess of those found following a meal. However, there is a recent suggestion that combining PYY(3-36) and GLP-1 may allow feeding inhibitions at plasma levels of each closer to those found following a meal.

An additional peptide that may play a role in controlling food intake is the recently identified novel peptide obestatin. Obestatin is derived from the prepropeptide that codes for the gastric orexigenic peptide ghrelin. In opposition to the actions of ghrelin, obestatin suppresses food intake and decreases body weight gain. Obestatin has been shown to act at a separate receptor site from ghrelin, suggesting a novel role for a gastric peptide in inhibiting food intake.

Summary

Gut vagal afferent neurons and peptide signaling modulate eating, providing feedback inhibition within individual or across multiple meals. The peptides involved have different sites of release along the GI tract and different patterns of release that are consistent with their actions. Vagal afferent neurons and/or hindbrain signaling appear to mediate the feeding actions of many of these peptides.

See also: Autonomic Nervous System: Gastrointestinal Control; Food and Water Intake: Regulation; Gastrointestinal Signals: Stimulation; Neuropeptides: Food Intake.

Further Reading

Badman MK and Flier JS (2005) The gut and energy balance: Visceral allies in the obesity wars. *Science* 307: 1909–1914.

Bloom S, Wynne K, and Chaudhri O (2005) Gut feelings – the secret of satiety? *Clinical Medicine* 5: 147–152.

Dhillo WS and Bloom SR (2004) Gastrointestinal hormones and regulation of food intake. *Hormone and Metabolic Research* 36: 846–851.

Kaplan JM and Moran TH (2004) Gastrointestinal signaling in the control of food intake. In: Stricker E and Woods S (eds.) *Handbook of Behavioral Neurobiology: 14 Neurobiology of Food and Fluid Intake*, pp. 275–306. Kluwer Academic/Plenum Publishers: New York.

Lutz TA (2005) Pancreatic amylin as a centrally acting satiating hormone. *Current Drug Targets* 6: 181–189.

Moran TH (2004) Gut peptides in the control of food intake: 30 years of ideas. *Physiology and Behavior* 82: 175–180.

Moran TH and Kinzig KP (2004) Gastrointestinal satiety signals II. Cholecystokinin. *American Journal of Physiology, Gastrointestinal and Liver Physiology* 286: G183–G188.

Smith GP (1998) *Satiation: From Gut to Brain*. New York: Oxford University Press.

Smith GP, Gibbs J, and Young RC (1974) Cholecystokinin and intestinal satiety in the rat. *Federation Proceedings* 33: 1146–1149.

Strader AD and Woods SC (2005) Gastrointestinal hormones and food intake. *Gastroenterology* 128: 175–191.

Gastrointestinal Signals: Stimulation

R Nogueiras, University of Cincinnati, Cincinnati, OH, USA
P Wiedmer, German Institute for Human Nutrition, Potsdam-Rehbrucke, Germany
D Perez-Tilve, University of Cincinnati, Cincinnati, OH, USA
S J Caton, Ludwig-Maximilians-University, Munich, Germany
M H Tschöp, German Institute for Human Nutrition, Potsdam-Rehbrucke, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The role of the gut in the digestion and absorption of nutrients is well established. The gut is not only involved in the assimilation and storage of nutrients, but is also an important endocrine organ, secreting a number of factors informing the brain about the nutritional status of the organism. Thus, the gut plays an essential role in the regulation of food intake and energy balance. These gastrointestinal (GI) signals are released from specialized enteroendocrine cells that are localized along the GI tract. For this action the gut uses neural and endocrine pathways to communicate with controllers of energy balance in the hypothalamus and hindbrain. Interestingly, many of the signals secreted by the GI are also found in the central nervous system, supporting a role for these signals not only peripherally but also at the central level. Most of these factors possess anorexigenic actions; however, some of them increase food intake.

The first hormone to be discovered in the GI tract was secretin, isolated from duodenal mucosa. Some years later, two additional GI hormones were described, gastrin and cholecystokinin (CCK). Over the past few decades a number of other gut peptides have been discovered. Several of these peptides, glucagon-like peptide-1 (GLP-1), pancreatic polypeptide (PP), and peptide YY (PYY), to name a few, have been reported to decrease food intake. However, in this article we focus on the few GI tract-secreted peptides which have been demonstrated to stimulate food intake, and discuss their regulation under different physiological conditions.

β-Casomorphin

Enterostatin is a pentapeptide produced by trypsin cleavage of pancreatic procolipase in the intestine during the absorption of fat. Enterostatin has been demonstrated to decrease food intake regardless of whether it is administered peripherally or centrally, and several studies have proposed that it interacts with opioid receptors in this regard. The opioid receptors involved in feeding are subdivided into the μ and κ subtypes. Opioid receptor agonists stimulate food intake, and one such example is β-casomorphin (β-CM), an agonist for the μ receptor. This peptide is a cleavage product of milk casein and has only seven amino acids. There is a high homology between the sequences of β-CM and enterostation, which consist of five amino acids, each peptide having proline residues in positions two and four. However, these two agents have opposite effects. β-CM stimulates food intake when given peripherally, and this effect is lost when the three C-terminal amino acids (Gly-His-Ile) are removed.

One study investigated the effect of several isoforms of β-CMs (β-CM1–β-CM7, β-CM1–β-CM5, and β-CM1–β-CM4) on food intake. The intracerebroventricular (icv) injection of isoforms 1–7 stimulated intake of a high-fat (HF) diet in animals, but isoforms 1–5 and 1–4 failed to promote intake of this diet. Peripheral injection of isoforms 1–7 also resulted in a dose-dependent increase in consumption of a HF diet, but reduced the intake of a high-carbohydrate (HC) diet in satiated rats with a similar dose-dependent relationship. Enterostatin (icv) inhibited the stimulation of HF intake by isoforms 1–7, as did the general opioid antagonist naloxone. These data suggest that at high doses, enterostatin and isoforms 1–7 may interact with the same low-affinity receptor to modulate intake of dietary fat. Consistent with this, it was observed that when rats with a gastric cannula were given a choice of a HF and a low-fat (LF) diet, enterostatin caused a dose-dependent reduction in the intake of the HF diet for the first 2 h after infusion, but had no effect on the LF intake. β-CM isoforms 1–7 stimulated intake of the HF diet, but had no effect on the LF diet, and blocked the inhibitory effect of enterostatin on HF intake in fasted rats.

To clarify the relationship between enterostatin and β-CM at different doses, K Berger and collaborators investigated the effect of intravenous (iv) administration of enterostatin in the presence and absence of β-CM, on HF intake in rodents. At low doses, enterostatin decreased HF intake in rats, whereas high doses stimulated food intake during the first hour after injection, compared to control rats. Contrary to this finding, iv β-CM increased HF intake at all concentrations. When enterostatin and β-CM were administered together at a low dose, there was no effect on food intake; however, at a higher dose an increase in food intake was detected, compared to

controls. These findings demonstrate that enterostatin and β-CM act in an antagonistic way at certain doses, and that β-CM is capable of abolishing the anorexigenic effect of enterostatin.

Motilin

Based on structural and effect-related similarities, motilin, a 22-amino-acid GI peptide, and ghrelin, the natural ligand of the growth hormone secretagogue receptor (GHS-R), are considered to represent a novel GI hormone family. Motilin is secreted from the endocrine cells of the mucosa of the upper part of the small intestine. This peptide stimulates GI motor activity especially in the antrum and upper duodenum, and it is widely acknowledged that motilin is involved in the regulation of interdigestive motility.

The production of motilin in the brain has been debated. Early studies suggest that motilin immunoreactivity is present in the Purkinje cells of the cerebellum in mice, rats, pigs, monkeys, and humans, as well as in the cerebrum, pituitary body, hypothalamus, and epiphysis in mice, rats, dogs, pigs, cows, monkeys, and humans. Motilin is also reported to excite neurons in the cerebral cortex and spinal cord of rats. Using longitudinal muscle/myenteric plexus (LMMP) preparations and cultures of myenteric neurons of the guinea pig ileum, it was observed that most of the motilin-immunoreactive and motilin receptor-immunoreactive neurons were also positive for neuronal nitric oxide synthase, with only a few being positive for choline acetyltransferase. In contrast, ghrelin was mainly co-localized with choline acetyltransferase. Motilin, ghrelin, and ghrelin receptor mRNA transcripts were found in LMMP preparations and in cultured myenteric neurons, suggesting that both peptides may play a role in the activation of the enteric nervous system and in the regulation of GI motility.

Motilin has been reported to possess orexigenic actions. A significant increase in food intake was observed after acute icv administration of motilin in both mice and rats. This effect was attenuated markedly by the motilin receptor antagonist GM-109. These results are consistent with the possibility that motilin receptors are present in the brain and may have a role in the regulation of food intake. Furthermore, peripheral administration of motilin has also been found to stimulate feeding in rats. In agreement with the effect of motilin on food intake it was reported that circulating motilin levels rise progressively during fasting and fall to a nadir within an hour of refeeding.

In addition to its actions on food intake, motilin is also considered to be an important factor in the control of the interdigestive migrating motor complex in the upper GI tract. In the fasting state, exogenous administration of motilin has been reported to induce a contraction pattern in the gastric antrum of both dogs and humans, and motilin also contracts the porcine gastric fundus *in vitro*. Moreover, administration of motilin causes a contraction of the proximal stomach in humans and increases meal-induced satiety in clinical studies.

Ghrelin

In 1999, the endogenous ligand for the GHS-R was discovered and named ghrelin. Its highest expression was found in the stomach, and more recently substantially lower amounts have been found in the bowel, pancreas, pituitary, kidney, and placenta. Ghrelin takes its name from the proto-Indo-European word *ghre*, which means 'grow,' and *relin*, relating to its role in the release of growth hormone (GH). Ghrelin results from the cleavage of a precursor called preproghrelin, which is composed of 117 amino acids. Preproghrelin contains a 23-amino-acid signal peptide and the 94-amino-acid proghrelin, which includes the 28-amino-acid mature ghrelin and a 66-amino-acid tail.

Ghrelin has potent growth hormone secretagogue activity and also stimulates food intake in humans and rodents. One peculiarity of this peptide is the *n*-octanoylation in the serine-3, because this modification of a peptide by esterification with a fatty acid is unique in mammalian biology; yet this modification is essential for ghrelin's biological functions. The attached acyl group enables the binding of ghrelin to the GHS-R and is responsible for the observed effects on appetite and energy intake. However, the main circulating form of ghrelin is not the active form, but the desacylated ghrelin, which constitutes about 80–90% of total ghrelin. Desacyl ghrelin cannot activate GHS-R1a, but it was recently reported that this isoform has cardioprotective, antiproliferative, and adipogenic properties, and that it can antagonize octanoyl-ghrelin-induced effects on insulin secretion and blood glucose levels in humans. However, the receptor that mediates those actions is not known. Desacyl ghrelin is not known to be directly involved in the regulation of appetite or food intake.

Another isoform of ghrelin that misses the Gln residue at position 14 of the molecule was also purified in rat and was named des-Gln14-ghrelin. This 27-amino-acid des-Gln14-ghrelin is produced by alternative splicing of the ghrelin gene. It also has an acylated and a nonacylated form, the last one being the most abundant. The acylation appears to be essential for the release of GH in both natural forms of ghrelin, although desacyl ghrelin may also be

biologically active in the cardiovascular system, in the promotion of adipogenesis, or in the regulation of insulin sensitivity. Other minor forms of ghrelin are present in human plasma and stomach, measured as the total immunoreactivity by conventional radioimmunoassay (RIA) based on the C-terminal fragment of ghrelin.

Regulation of Ghrelin by Nutritional Status and Obesity

Gastric ghrelin production is regulated by nutritional status and various hormonal factors. Ghrelin is an important factor in meal initiation; levels rise prior to the onset of meal ingestion and then decline following consumption. The postprandial suppression of ghrelin may require a postgastric factor, since the intragastric administration of saline or glucose did not modify ghrelin levels when gastric emptying was prevented. This profile is also independent of any meal-related cues. When all external cues indicating the onset of a meal were removed, this 'typical' profile of ghrelin release was still observed. In addition, ghrelin levels were positively associated with subjective hunger ratings. Intravenous infusion of ghrelin significantly increases food intake, and ghrelin levels were highly correlated with self-reported hunger. Preprandial increases in ghrelin release have also been reported to have a learned component, related to habitual meal pattern in sheep trained to consume one, two, or three meals per day. Accordingly, in each group ghrelin increased preprandially in contrast to the lower ghrelin levels observed in the *ad libitum*-fed control animals.

The meal-related suppression of plasma ghrelin requires postgastric stimulation. The energy content of the ingested meal is also an important factor in postprandial ghrelin suppression, since it has been reported to decline in proportion to the ingested caloric load. The macronutrient composition of the ingested meal has also been reported to influence postprandial ghrelin secretion. Meals rich in carbohydrate have a more profound effect. The postprandial suppression of ghrelin correlates with a rise in insulin, and this may partially explain the greater suppression of ghrelin by HC meals. Circulating ghrelin levels are low in obese patients and higher after fasting. However, obesity may not be characterized by ghrelin resistance, since obese individuals have been reported to respond to peripheral administration of ghrelin by increasing food intake. Obese individuals also have been reported to have an attenuated postprandial suppression of ghrelin regardless of the energy content of the meal. The low ghrelin levels in obese individuals are reversed by diet-induced weight loss.

Patients with Prader-Willi syndrome (PWS), a disorder characterized by cognitive deficits, GH deficiency, extreme hyperphagia, and obesity, do not have low ghrelin levels. On the contrary, ghrelin levels are elevated in these patients and thus may contribute to the insatiable levels of hunger observed in these patients. However, when somatostatin, a potent suppressor of ghrelin secretion, was administered to patients with PWS, plasma ghrelin was suppressed but no changes of food intake were observed, thus questioning the role of hyperghrelinemia in hyperphagia observed in PWS patients. It is clear that ghrelin plays a complex role in appetite regulation and overall energy balance, yet the precise mechanism for the orexigenic effect of ghrelin is not completely understood.

Pharmacological Effects of Ghrelin on Energy Balance

Central (icv) administration of ghrelin induces c-fos expression in regions of primary importance in the regulation of feeding, primarily in the hypothalamus, including arcuate nucleus (ARC), paraventricular nucleus (PVN), and dorsomedial and ventromedial hypothalamic nuclei. Depending of the exact site of administration, c-fos was stimulated in several nuclei. Ghrelin-containing neurons are localized in the ARC and these neurons project efferent fibers onto neuropeptide Y (NPY)- and agouti-related protein (AgRP)-expressing neurons to stimulate the release of these orexigenic peptides, and onto proopiomelanocortin (POMC) neurons to suppress the release of anorexigenic factors from these cells. Most NPY/AgRP-containing neurons and POMC/cocaine- and amphetamine-regulated transcript (CART)-containing neurons also express leptin receptors, and both types of neurons are regulated by leptin in a reciprocal manner. Leptin inhibits ghrelin-induced feeding, and ghrelin substantially reverses the anorexic effect of leptin. However, the hypothalamus is not the only site of action for ghrelin, as brain stem areas have also been implicated in the stimulation of food intake by ghrelin. Administration of ghrelin into the fourth ventricle or directly into the dorsal vagal complex resulted in a hyperphagic response with a magnitude similar to that obtained with administration in the third ventricle. Moreover, GHS-R is expressed in neurons of the nucleus of the solitary tract, the dorsal motor nucleus of the vagus, the rostral ventrolateral medulla, and the caudal ventrolateral medulla. It has been reported that injection of ghrelin into the hippocampus or dorsal raphe nucleus also evokes an increase in food intake, suggesting that, in addition to areas of ghrelin control in the hypothalamus, ghrelin exhibits

orexigenic activity in various zones of the central nervous system, especially in the caudal brain stem. Ghrelin also directly acts at GHS-Rs expressed in ventral tegmentum neurons of the midbrain. Since such circuitry is known to be important for reward and addiction, it has now been proposed that similar to leptin, ghrelin directly targets the hedonic as well as the homeostatic drive for food intake.

The role of the vagus nerve in the orexigenic effects of ghrelin still remains to be fully elucidated. The detection of ghrelin receptors on vagal afferent neurons in the rat nodose ganglion suggests that ghrelin signals from the stomach are transmitted to the brain via the vagus nerve. However, it was recently suggested that the preprandial increase in ghrelin occurs independently of vagal afferent signaling, suggesting that circulating ghrelin likely acts directly at target neurons of the mediobasal hypothalamus, the ventral tegmentum, and the brain stem.

In addition to its effects on food intake and body weight, intravenous ghrelin dose-dependently increases gastric acid secretion and stimulates gastric motility and emptying. Intracerebroventricular ghrelin also increases gastric acid secretion in a dose-dependent manner and induces c-fos expression in the nucleus of the solitary tract and the dorsomotor nucleus of the vagus nerve.

Physiological Functions of Ghrelin: Lessons from Targeted Mouse Mutagenesis

Early investigations examining ghrelin and GHS-R knockout mice failed to identify effects on food intake, bodyweight, or body composition, although mice lacking ghrelin signaling were found to be smaller. Notably though, it was reported that these mice had a reduced respiratory quotient when fed a HF diet, suggesting that endogenous ghrelin may be an essential regulator of substrate utilization. Later investigations have revealed that ghrelin-deficient as well as GHS-R-deficient mice are resistant to diet-induced obesity, indicating that endogenous ghrelin does have an essential function in the control of body fat mass. More recent reports indicate that endogenous ghrelin may be of crucial importance for insulin sensitivity and maintenance of glucose homeostasis. This function appears to become most apparent in the absence of both leptin and ghrelin, where ghrelin gene deletion seems to partially rescue the diabetic state of leptin-deficient mice.

In summary, ghrelin or ghrelin analogs are orexigenic and adipogenic agents with the largely untested potential to treat human cachexia or anorexia. On the other hand, GHS-R antagonists may be useful candidates for the pharmacological treatment of obesity or type 2 diabetes. The multifaceted character and pleiotropic action profile of ghrelin will continue to represent a major challenge for the development of specific and safe therapeutics until the molecular underpinnings of ghrelin action are better understood.

See also: Autonomic Nervous System: Gastrointestinal Control; Autonomic Nervous System: Central Control of the Gastrointestinal Tract; Cortical Processing of the Reward Value of Food; Energy Homeostasis: Adiposity Signals; Food and Water Intake: Regulation; Gastrointestinal Signals: Satiety; Nutrient Sensing: Cellular Metabolism; Nutrient Sensing: Carbohydrates.

Further Reading

Abizaid A, Liu ZW, Andrews ZB, et al. (2006) Ghrelin modulates the activity and synaptic input organization of midbrain dopamine neurons while promoting appetite. *Journal of Clinical Investigation* 116: 3229–3239.

Arnold M, Mura A, Langhans W, et al. (2006) Gut vagal afferents are not necessary for the eating-stimulatory effect of intraperitoneally injected ghrelin in the rat. *Journal of Neuroscience* 26: 11052–11060.

Berger K, Winzell MS, Mei J, et al. (2004) Enterostatin and its target mechanisms during regulation of fat intake. *Physiology & Behavior* 83: 623–630.

Casaueva FF and Dieguez C (2002) Ghrelin: The link connecting growth with metabolism and energy homeostasis. *Reviews in Endocrine & Metabolic Disorders* 3: 325–338.

Cowley MA, Smith RG, Diano S, et al. (2003) The distribution and mechanism of action of ghrelin in the CNS demonstrates a novel hypothalamic circuit regulating energy homeostasis. *Neuron* 37: 649–661.

Cummings DE, Weigle DS, Frayo RS, et al. (2002) Plasma ghrelin levels after diet-induced weight loss or gastric bypass surgery. *New England Journal of Medicine* 346: 1623–1630.

Date Y, Murakami N, Toshinai K, et al. (2002) The role of the gastric afferent vagal nerve in ghrelin-induced feeding and growth hormone secretion in rats. *Gastroenterology* 123: 1120–1128.

Faulconbridge LF, Cummings DE, Kaplan JM, et al. (2003) Hyperphagic effects of brainstem ghrelin administration. *Diabetes* 52: 2260–2265.

Itoh Z (1997) Motilin and clinical application. *Peptides* 18: 593–608.

Kojima M, Hosoda H, Date Y, et al. (1999) Ghrelin is a growth-hormone-releasing acylated peptide from stomach. *Nature* 402: 656–660.

Sun Y, Ahmed S, and Smith RG (2003) Deletion of ghrelin impairs neither growth nor appetite. *Molecular Cellular Biology* 23: 7973–7981.

Sun Y, Asnicar M, Saha PK, et al. (2006) Ablation of ghrelin improves the diabetic but not obese phenotype of ob/ob mice. *Cell Metabolism* 3: 379–386.

Sun Y, Wang P, Zheng H, et al. (2004) Ghrelin stimulation of growth hormone release and appetite is mediated through the growth hormone secretagogue receptor. *Proceedings of the National Academy of Sciences of the United States of America* 101: 4679–4684.

Tschöp M, Smiley DL, and Heiman ML (2000) Ghrelin induces adiposity in rodents. *Nature* 407: 908–913.

Tschöp M, Weyer C, Tataranni PA, et al. (2001) Circulating ghrelin levels are decreased in human obesity. *Diabetes* 50: 707–709.

van der Lely AJ, Tschöp M, Heiman ML, et al. (2004) Biological, physiological, pathophysiological, and pharmacological aspects of ghrelin. *Endocrine Review* 25: 426–457.

Wortley KE, Anderson KD, Garcia K, et al. (2004) Genetic deletion of ghrelin does not decrease food intake but influences metabolic fuel preference. *Proceedings of the National Academy of Sciences of the United States of America* 101: 8227–8232.

Wortley KE, del Rincon JP, Murray JD, et al. (2005) Absence of ghrelin protects against early-onset obesity. *Journal of Clinical Investigation* 115: 3573–3578.

Wren AM, Seal LJ, Cohen MA, et al. (2001) Ghrelin enhances appetite and increases food intake in humans. *Journal of Clinical Endocrinology and Metabolism* 86: 5992–5996.

Zigman JM, Nakano Y, Coppari R, et al. (2005) Mice lacking ghrelin receptors resist the development of diet-induced obesity. *Journal of Clinical Investigation* 115: 3564–3572.

Gastrointestinal Tract Role in Neural Control of Metabolism, Food Intake and Body Weight: A Summary

W Jänig, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neural and endocrine regulation of metabolism and food intake by the brain is the main integrative function in which the gastrointestinal tract (GIT) is involved. Closely related to this regulation is the regulation of hunger, satiety, and body weight. The neural machineries that are important in this regulation are situated in the lower brain stem (in particular in the dorsal vagal complex (DVC) consisting of the nucleus tractus solitarii (NTS), the dorsal motor nucleus of the vagus nerve (DMNX) and the area postrema (AP)) and in the hypothalamus. The afferent neural signals from the GIT that are important in this regulation project through the vagus nerves to the nucleus tractus solitarii. The endocrine signals derive from endocrine cells in the mucosa of the GIT (most of them in the anterior part), pancreas, and adipose tissue, and they act on the brain mainly via the area postrema in the medulla oblongata and/or via the arcuate nucleus (ARC) in the hypothalamus.

A Model of Regulation of Metabolism and Food Intake

Regulation of metabolism and food intake is homeostatic and consists of a long-term regulation and a short-term regulation, both of which are closely integrated (Figure 1):

- The adipose tissue is the main energy store of the body. A normal human male of 75 kg body weight stores approximately $5–6\times10^5$ kJ in his fat tissue (15% of his body mass). Neural regulation of this energy store is a long-term regulation. The size of the body fat mass is proportional to the body weight. Regulation of the body fat mass (and therefore also body weight) is quantitatively precise and normally stable to <1% over months and years. The control centers, located in the hypothalamus and in the lower brain stem (mainly the DVC), receive two signals that are proportional in their blood concentration to the size of the body fat mass – leptin and insulin, also called adiposity-related (or adipositas) signals. Leptin is synthesized and released by adipocytes, and the hormone insulin is released by B cells of the endocrine pancreas. Increase of both hormones in the blood (e.g., during ample food availability and fat stores) leads to decreased energy (food) intake and of endogenous glucose production by the liver, increased energy expenditure, and a mobilization of fat stores (catabolic state of the organism). Decrease of both hormones in the blood (e.g., during starvation) leads to increased energy intake, decreased energy expenditure, elevated fat stores, and increased glucose production by the liver (anabolic state of the organism). Genetically engineered changes of these adipositas feedback signals, their receptors, or their postreceptor pathways result in changes predicted by this hypothesis.
- Neural regulation of food (energy) intake (regulation of meal size and meal frequency) is a short-term regulation. The regulation centers are situated in the DVC of the medulla oblongata and in the hypothalamus. This regulation is fast but quantitatively not precise. It is dependent on distinct neural, hormonal, and nutritive afferent signals to the brain. The DVC receives signals in vagal afferents encoding in their activity mechanical and/or intraluminal chemical events of the GIT, signals in taste afferents, and hormonal signals from the GIT via the area postrema (cholecystokinin (CCK), glucagon-like peptide 1, pancreatic polypeptide, and peptide YY (PYY)). However, it probably also receives hormonal signals that are important in the regulation of metabolism, namely leptin and insulin, as well as nutritive signals (glucose and lipids in the blood). Neurons in the ARC of the hypothalamus receive hormonal signals and nutritive signals. The hormonal signals are ghrelin from the oxyntic cells of the mucosa of the stomach that initiates food intake and PYY from the mucosa cells of the hindgut and ileum. With the exception of ghrelin, all afferent signals (neural, hormonal, or nutritive) are satiety signals; that is, they terminate, delay, or reduce food intake.

The homeostatic regulation of metabolism and food intake interact with other homeostatic regulations represented in the hypothalamus (body core temperature, reproduction, growth, fluid matrix, etc.). Furthermore, this regulation is under the control of the cerebral systems. This control is nonhomeostatic and is related to satiety and hunger as well as to reward and hedonic control of food intake.

Neural Components of Regulation of Metabolism and Food Intake

The central neural components that are involved in the regulation of metabolism and food intake are summarized in Figure 2.

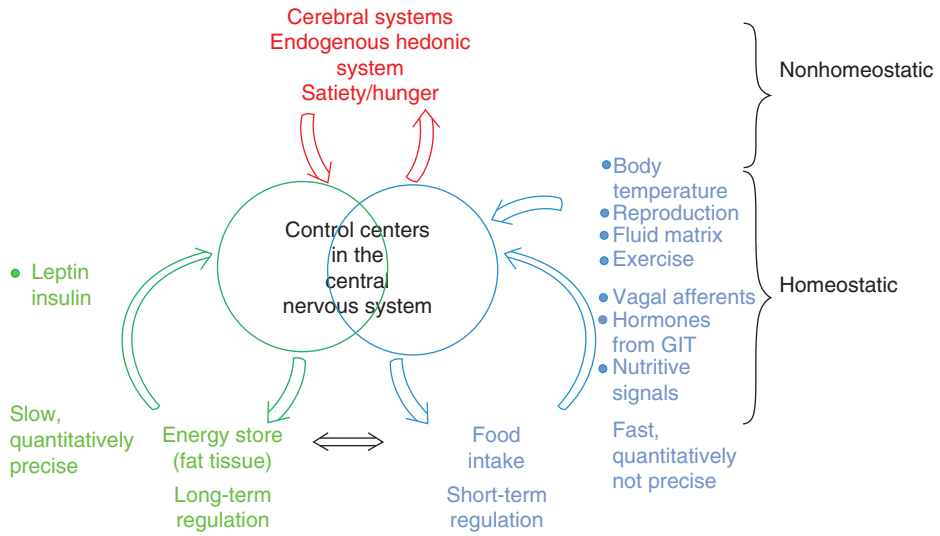

Figure 1 Homeostatic long-term regulation of energy stores and short-term regulation of food intake by the brain. Homeostatic control centers are located in the hypothalamus and the lower brain stem. Long-term regulation of the main energy store (fat tissue) is slow but quantitatively precise, leading to a precision of the regulation of fat stores and body weight in the range of $\leq$1%. The main feedback signals are the peptides leptin (from the fat stores) and insulin (from the endocrine pancreas) to the hypothalamus. Short-term regulation of food intake (meal size and frequency of food intake) is fast and quantitatively relatively inaccurate. The main feedback signals are activity in vagal afferents from the GIT, hormones from the GIT, and nutritive signals (level of glucose and free fatty acids (FFA) in the blood) to the lower brain stem and/or the hypothalamus. The control centers involved in this regulation are integrated with other homeostatic regulations represented in the hypothalamus (regulation of body temperature, reproduction, fluid matrix, etc.). The homeostatic control centers in hypothalamus and brain stem are under the control of the cerebral systems responsible for the general sensations of hunger and satiety and represent the endogenous hedonic system. These cerebral controls are not homeostatic. GIT, gastrointestinal tract.

As already mentioned, the control centers of regulation of metabolism and food intake are represented in neural networks of the hypothalamus and the lower brain stem. The functioning of these centers and their interaction cannot be explained by a hierarchical model in which the hypothalamic centers take the lead, via command signals in descending projecting neurons, over the lower brain stem (mainly the DVC). In fact, the sensorimotor circuits in the lower brain stem can generate the full patterns of ingestive behavior involving the GIT and somatomotor systems. The way afferent neural, hormonal, and nutritive signals are translated by these circuits into coordinated behavior involving autonomic systems to the GIT, other autonomic systems, and somatomotor systems is poorly understood. Finally, the communication between hypothalamic and lower brain stem centers is reciprocal. Details of this communication are just beginning to be worked out, and the description given here is a somewhat simplified view.

1. Leptin and insulin in the blood activate, via leptin and insulin receptors, neurons in the ARC of the hypothalamus, which contain proopiomelanocortin (POMC), the precursor of α-melanocyte-stimulating hormone used by these neurons as transmitter. These neurons activate the parvocellular neurons in the paraventricular nucleus of the hypothalamus (PVH) via the melanocortin-3 and -4 receptors (Mc3R and Mc4R). The parvocellular neurons contain either oxytocin (Oxt), corticotropin-releasing hormone (CRH), or thyrotropin-releasing hormone (TRH). Activation of these neurons generates a catabolic state of the organism, with weight loss and decrease in food intake.

2. Decrease of food intake and the accompanying changes of the GIT (decrease of motility, endocrine and exocrine secretion; inhibition of hepatic glucose production) are generated, at least in part, by the activation of oxytocinergic neurons in the PVH that project to the DVC. This activation enhances the effect of neural and hormonal satiety signals from the GIT to the DVC. The changes in the GIT, including the inhibition of hepatic glucose production, are mediated by vagal preganglionic neurons in the DVC.

3. Interruption of the specific signaling chain leptin–insulin–POMC neurons–PVN neurons (e.g., by deletion of the leptin or insulin receptor in the neurons of the ARC, in transgenic mice lacking the functioning peptide leptin or by blockade of the Mc4 receptor in the PVH nucleus) results in an anabolic state of the organism with an increase of body weight and food intake. This view does not

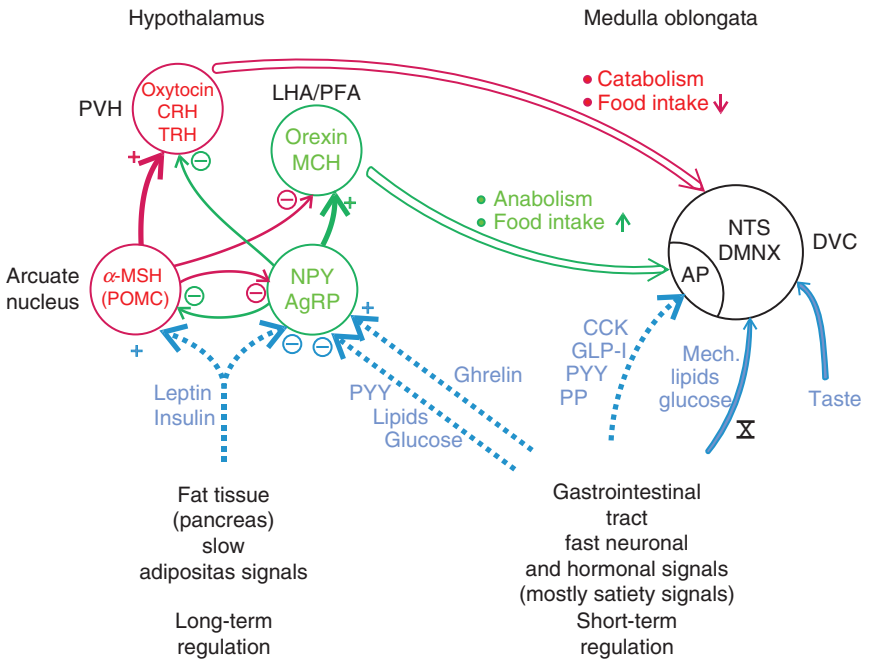

Figure 2 Centers in the hypothalamus and lower brain stem involved in the regulation of metabolism (anabolism and catabolism) and food intake. Long-term regulation of fat tissue and short-term regulation of food (energy) intake are integrated. Adipositas signals (leptin from fat tissue and insulin from the endocrine pancreas) activate neurons containing the peptides proopiomelanocortin (POMC) and α-melanocyte stimulating hormone (α-MSH) and inhibit neurons containing neuropeptide Y (NPY) and agouti-related peptide (AgRP) in the arcuate nucleus (ARC) of the hypothalamus. Stimulation of POMC neurons in the ARC of the hypothalamus activates parvocellular neurons in the paraventricular nucleus of the hypothalamus (PVH, containing either oxytocin or CRH or TRH) via melanocortin receptors (Mc3R and Mc4R). This activation generates a catabolic state of the organism. Stimulation of NPY/AgRP neurons in the ARC (e.g., by decrease of concentration of leptin and insulin in the blood and the subsequent decrease of inhibition of NPY/AgRP neurons) activates neurons in the lateral hypothalamic area (LHA) and parafornical area (PFA), which contain orexin or melatonin-concentrating hormone (MCH), via the Y1 and Y5 receptors. This activation generates an anabolic state of the organism. Both catabolic (anorexigenic) circuits (red) and anabolic (orexigenic) circuits (green) inhibit each other reciprocally. This inhibition is exerted by GABAergic neurons or by AgRP acting as an endogenous antagonist at the Mc4 receptor of the neurons in the PVH. Fast (neural and hormonal) signals from the GIT and nutritional signals in the blood (glucose and lipids) act either on neurons in the dorsal vagal complex (DVC; in the nucleus tractus solitarii (NTS) or via the area postrema (AP)) or on the NPY/AgRP neurons in the ARC. The hormone ghrelin stimulates food intake. With the exception of ghrelin, the signals are involved in short-term regulation and satiety signals. They lead to a reduction of food intake and further catabolism by the inhibition of NPY/AgRP neurons. +, excitation; −, inhibition; CCK, cholecystokinin; CRH, corticotropin-releasing hormone; DMNX, dorsal motor nucleus of the vagus nerve; GABA, γ-aminobutyric acid; GIT, gastrointestinal tract; GLP-1, glucagon-like peptide 1; mech., mechanical; PP, pancreatic polypeptide; PYY, peptide YY; TRH, thyrotropin-releasing hormone; Ⅹ, vagus nerve. Data from Batterham RL, Cowley MA, Small CJ, et al. (2002) Gut hormone PYY(3–36) physiologically inhibits food intake. *Nature* 418: 650–654; Inui A, Asakawa A, Bowers CY, et al. (2004) Ghrelin, appetite, and gastric motility: The emerging role of the stomach as an endocrine organ. *FASEB Journal* 18: 439–456; Lam TK, Pocai A, Gutierrez-Juarez R, et al. (2005) Hypothalamic sensing of circulating fatty acids is required for glucose homeostasis. *Nature Medicine* 11: 320–327; Morton GJ, Blevins JE, Williams DL, et al. (2005) Leptin action in the forebrain regulates the hindbrain response to satiety signals. *Journal of Clinical Investigations* 115: 703–710; Schwartz MW and Morton GJ (2002) Obesity: Keeping hunger at bay. *Nature* 418: 595–597; Yamamoto H, Kishi T, Lee CE, et al. (2003) Glucagon-like peptide-1-responsive catecholamine neurons in the area postrema link peripheral glucagon-like peptide-1 with central autonomic control sites. *Journal of Neuroscience* 23: 2939–2946. Adapted from Schwartz MW, Woods SC, Porte D Jr., Seeley, RJ, and Baskin DG (2000) Central nervous system control of food intake. *Nature* 404: 661–671.

exclude leptin and insulin also acting on neurons in the DVC, POMC neurons in the ARC projecting to the DVC, and similar reaction patterns also being generated independently of the PVH from the DVC.

4. Activation of a second group of neurons in the ARC of the hypothalamus generates an anabolic state of the organism with an increase of food intake and gain of body weight. These neurons contain neuropeptide Y (NPY) and the peptide agouti-related peptide (AgRP). Their excitation activates neurons in the lateral hypothalamic area (LHA) and the parafornical area of the hypothalamus (PFA) via the Y1 and Y5 receptors. Neurons in the LHA and PFA contain the peptides orexin or melanin-concentrating hormone (MCH). NPY and AgRP neurons in the

ARC are inhibited by leptin, insulin, and PYY and are excited by the peptide ghrelin.

5. Activation of the orexin and MCH neurons attenuates the response of the neurons in the DVC to the neural and hormonal satiety signals from the GIT. Injection of orexin or MCH into the third ventricle or overexpression of MCH in transgenic mice generates an anabolic state in the animals. Targeted deletion of MCH or its receptor leads to weight loss and a hypermetabolic (catabolic) state of the animals. Nutritive signals (glucose and lipids in the blood) also inhibit NPY and AgRP neurons in the ARC through specific intracellular pathways.

6. Neurons whose activation generates a catabolic state of the organism (red in **Figure 2**) and neurons whose activation generates an anabolic state of the organism (green in **Figure 1**) inhibit one another reciprocally. This inhibition is mediated by the transmitter γ-aminobutyric acid (GABA) in the ARC. In the PVH nucleus, AgRP acts as an endogenous antagonist at the Mc4 receptor, thus blocking the signal transmission from the POMC neurons to the Oxt, CRH, and TRH neurons. In the lateral hypothalamus, the inhibition of orexin and MCH neurons is probably mediated by GABA too. This reciprocal inhibition between neurons of the catabolic and neurons of the anabolic system means that activation of a catabolic state is also produced by the disinhibition of POMC neurons and of Oxt, CRH, and TRH neurons generated by the inhibition of NPY and AgRP neurons by leptin, insulin, PYY, and/or nutritive signals.

Future experimental investigations will elaborate the details of this complex neural regulation system, in the center of which is the GIT. These investigations will work out the relationship of this homeostatic regulation with cortical systems, such as the insular cortex, the anterior and middle cingulated cortex, and the prefrontal cortex, that are responsible for reward, hedonic aspects of food intake, and the sensations of hunger and satiety. Unraveling these nonhomeostatic cerebral control mechanisms may give us better insight into the mechanisms of the failure of regulation of food intake and body weight in higher vertebrates, in particular humans.

See also: Autonomic Nervous System: Gastrointestinal Control; Autonomic Nervous System: Central Control of the Gastrointestinal Tract; Food and Water Intake: Regulation; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Hypothalamic Structure–Function Relationships; Neuroendocrine Control of Energy Balance (Central Circuits/Mechanisms); Neuropeptides: Food Intake.

Further Reading

Batterham RL, Cowley MA, Small CJ, et al. (2002) Gut hormone PYY(3–36) physiologically inhibits food intake. *Nature* 418: 650–654.

Blevins JE, Schwartz MW, and Baskin DG (2004) Evidence that paraventricular nucleus oxytocin neurons link hypothalamic leptin action to caudal brain stem nuclei controlling meal size. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 287: R87–R96.

Gao Q and Horvath TL (2007) Neurobiology of feeding and energy expenditure. *Annual Review of Neuroscience* 30: 367–399.

Grill HJ and Kaplan JM (2002) The neuroanatomical axis for control of energy balance. *Frontiers in Neuroendocrinology* 23: 2–40.

Inui A, Asakawa A, Bowers CY, et al. (2004) Ghrelin, appetite, and gastric motility: The emerging role of the stomach as an endocrine organ. *FASEB Journal* 18: 439–456.

Jänig W (2006) *The Integrative Action of the Autonomic Nervous System: Neurobiology of Homeostasis*. Cambridge, UK: Cambridge University Press.

Lam TK, Pocai A, Gutierrez-Juarez R, et al. (2005) Hypothalamic sensing of circulating fatty acids is required for glucose homeostasis. *Nature Medicine* 11: 320–327.

Morton GJ, Blevins JE, Williams DL, et al. (2005) Leptin action in the forebrain regulates the hindbrain response to satiety signals. *Journal of Clinical Investigations* 115: 703–710.

Peters A, Pellerin L, Dallman MF, et al. (2007) Causes of obesity: Looking beyond the hypothalamus. *Progress in Neurobiology* 81: 61–68.

Pocai A, Lam TK, Gutierrez-Juarez R, et al. (2005) Hypothalamic K(ATP) channels control hepatic glucose production. *Nature* 434: 1026–1031.

Pocai A, Obici S, Schwartz GJ, and Rossetti L (2005) A brain–liver circuit regulates glucose homeostasis. *Cell Metabolism* 1: 53–61.

Porte D Jr., Baskin DG, and Schwartz MW (2005) Insulin signaling in the central nervous system: A critical role in metabolic homeostasis and disease from *C. elegans* to humans. *Diabetes* 54: 1264–1276.

Schwartz MW (2006) Central nervous system regulation of food intake. *Obesity* 14: 1S–8S.

Schwartz MW and Porte D Jr. (2005) Diabetes, obesity, and the brain. *Science* 307: 375–379.

Schwartz MW, Woods SC, Porte D Jr., Seeley RJ, and Baskin DG (2000) Central nervous system control of food intake. *Nature* 404: 661–671.

Stanley S, Wynne K, McGowan B, and Bloom S (2005) Hormonal regulation of food intake. *Physiological Reviews* 85: 1131–1158.

van der Lely AJ, Tschop M, Heiman ML, and Ghigo E (2004) Biological, physiological, pathophysiological, and pharmacological aspects of ghrelin. *Endocrine Review* 25: 426–457.

Yamamoto H, Kishi T, Lee CE, et al. (2003) Glucagon-like peptide-1-responsive catecholamine neurons in the area postrema link peripheral glucagon-like peptide-1 with central autonomic control sites. *Journal of Neuroscience* 23: 2939–2946.

Gene Expression Dysregulation in CNS Pathophysiology

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

Pathophysiological changes in the brain include the deterioration or degeneration of specific brain regions, alterations in cell morphology, changes in the relative size of different brain regions, and the reorganization of neuronal connections. These changes in physiology often reflect more subtle changes in brain chemistry, signaling processes, and, underlying all else, changes in the levels of gene expression. While signaling processes in the developing and adult brain may have direct effects on cell behavior through influences on the cytoskeleton or on cell adhesion, most signals culminate in the nucleus and result in the activation or repression of specific sets of genes. Any neurological disorder that has a significant genetic component must in some way reflect the abnormal activity of such genes, and it is becoming increasingly clear, therefore, that pathophysiological changes are intricately linked to changes in the level of gene expression. In some cases, such changes are a direct reflection of mutations in specific genes, while in others they reflect the overactivity or underactivity of signaling pathways or regulators caused by particular external stimuli, such as stress or drug use. Some changes in gene expression are short-lived and easily reversible, while others lead to long-lasting effects which must be addressed with long-term therapies. The regulation and dysregulation of gene expression can mediate both the transient and the stable maladaptations in the brain that lead to pathophysiological changes.

Causes of Gene Dysregulation

Mutations with *cis*-Acting Effects on Single Genes

Conceptually the simplest way for gene expression to become dysregulated is for a mutation to occur in the regulatory elements of a particular gene. Devastating single mutations are rare, but in a number of cases the predisposition to a particular disease has been shown to reflect the presence of mutations in the promoter or occasionally in the sites that regulate splicing, resulting in quantitative rather than qualitative differences in the level of mRNA. In some cases, it is not the total level of mRNA that matters, but rather the balance of alternative transcripts where a primary transcript undergoes alternative splicing.

Examples of *cis*-acting mutations that affect individual genes with pathological consequences include promoter mutations upstream of the *APP* gene, increasing the risk of Alzheimer's disease; a mutation in the acetylcholine receptor ε-subunit promoter, causing congenital myasthenic syndrome; and single-nucleotide polymorphisms in the apolipoprotein E promoter, which affect the phenotype of primary open-angle glaucoma.

A more complex example which involves the dysregulation of splicing and therefore modulates the ratio of alternative transcripts has been described for the *NRG1* gene, encoding neuregulin 1 (NRG1). At least six different isoforms of NRG1 have been described, with a diverse range of functions in the nervous system and links to several diseases, including schizophrenia. A recent association study using the Icelandic deCODE database yielded a high-risk haplotype at the 5′ end of the *NRG1* gene (SNP8NRG243177) that enhances the expression of type IV NRG1 in the brains of schizophrenic patients.

The transcript levels of key genes can also be modified by segmental deletions and duplications, as occurs in Charcot–Marie–Tooth disease when the dosage of the *PMP22* gene is increased from two to three copies, and in Alagille syndrome, which results from haploinsufficiency for the gene *JAG1*.

Mutations with *trans*-Acting Effects on Multiple Genes

The function of certain genes is to regulate other genes, and, depending on the number of downstream targets, mutations affecting such genes may cause significant dysregulation of expression. Examples of genes in this category include those encoding signaling proteins, transcription factors, or proteins that control RNA processing and protein synthesis. Importantly, there is scope here for wide-ranging changes because the affected genes might also encode regulatory proteins, which go on to have further knock-on effects on the nervous system.

A good example of this phenomenon is mutations affecting the activity of PAX6. The *PAX6* gene is expressed in the developing retina, iris, lens, and surface ectoderm, as well as in discrete regions of the forebrain, hindbrain, neural tube, pituitary gland, and nasal epithelium. It encodes a transcription factor that plays a critical early role in the developing nervous system and the effects of which are dosage critical. Mutations in *PAX6* are responsible for a range of neuroectodermal defects, including aniridia. The expression of *PAX6* is complicated by the presence of

two promoters and six enhancers, and the fact that alternative splicing yields three isoforms that recognize different DNA sequences. Genomic sequencing and microarray analysis have identified a number of downstream targets that become dysregulated following mutation of the *PAX6* gene. Interestingly, these include many further regulatory factors, such as transcription factors involved in neurogenesis (AP-2γ, Bhlhb5, NeuroD6, Nfia, Ngn2, Satb2, and Tbr2) and the retinoic acid signaling molecule Rlbp1. These have additional downstream targets that would also become dysregulated. From this example, it is clear that a single mutation in a critical master gene can impact on the expression of potentially hundreds of other genes by affecting the availability of other transcription factors and signaling proteins.

Epigenetic Changes Affecting Many Genes

Mutations by definition achieve dysregulation by changing the DNA sequence of a gene, and these effects are described as genetic. In contrast, epigenetic changes modify gene expression without modifying the gene sequence. This is achieved by altering the way DNA is packaged in the cell, principally by altering the structure of chromatin. Chromatin remodeling is brought about by a combined process of histone modification and DNA methylation.

Histones are the proteins that package DNA into chromatin in the nucleus, and they are subject to various forms of posttranslational modification, some of which are clearly linked to transcriptional accessibility. In particular, histone acetylation usually corresponds to active chromatin whereas the loss of acetylation correlates with inactive, silenced chromatin. Histone modification may also result in the specific recruitment of certain transcriptional activators and repressors, or proteins that sequester DNA into higher-order structures. Histone modification is mediated by specific enzymes, and those controlling acetylation are the best understood. Histone acetyltransferases (HATs) catalyze the acetylation of lysine and/or arginine residues, whereas histone deacetylases (HDACs) catalyze the removal of acetyl groups and associate with transcriptional repressors. The actual state of chromatin in any region of the genome depends to a large extent on the balance of these two opposing enzyme activities.

Transcriptional repression is also associated with the methylation of DNA at the 5′ position of cytidine residues in the context of the dinucleotide sequence CpG. The methylation of DNA is catalyzed by DNA methyltransferases (DNMTs), which transfer the methyl group from an *S*-adenosylmethionine (SAM) donor. As might be expected, patterns of DNA methylation correspond to patterns of histone modification that establish inactive chromatin. This is achieved through the presence of methyl-CpG-binding proteins (MeCPs) that associate with protein complexes containing HDACs and histone methyltransferases. DNA methylation, histone methylation, and histone deacetylation are therefore intrinsically linked in order to facilitate transcriptional silencing, establish epigenetic markers, and allow the silenced state to persist through DNA replication and cell division. Since chromatin remodeling is a general rather than gene-specific phenomenon, it is easy to see how altering the activity of histone-modifying enzymes, DNA methyltransferases, or other chromatin components could have a profound impact on the expression of many genes. It is becoming clear that a number of complex neurological disorders, including psychiatric disorders, may reflect gene dysregulation caused by the perturbation of chromatin remodeling, rather than specific defects in individual genes.

To exemplify the role of epigenetic modifications in pathological gene dysregulation, there is now a lot of evidence that epigenetic mechanisms underlie disorders such as schizophrenia. Microarray analysis of postmortem brains suggests that significant numbers of genes are misexpressed. Histone remodeling has been implicated in schizophrenia because antipsychotic drugs, such as the dopamine receptor antagonist haloperidol, induce rapid phosphoacetylation of histone H3 in the c-*fos* promoter in mouse striatal neurons. Similarly, an anticonvulsant and mood stabilizer, valproic acid, is able to increase expression of the *reelin* gene by demethylating the *reelin* promoter and inhibiting the activity of HDACs. Reelin is a glycoprotein with a main function to control neuronal migration during brain development, although it is also expressed in γ-aminobutyric acid (GABA-)ergic neurons in the adult brain. Reelin is significantly underexpressed in schizophrenic brains, and treatment with agents that either promote or inhibit methylation of the *reelin* promoter can repress or elevate its expression, respectively. These data suggest that disturbances in the levels of DNMTs, and histone-modifying enzymes could contribute to schizophrenia by altering the degree of methylation in the promoters of genes such as *reelin*, causing a generalized disruption in genes involved in the GABA and glutamate systems.

Detecting Gene Dysregulation

Traditional Methods

Traditional methods for studying gene regulation are difficult to apply on a large scale and therefore tend to be used to look at specific candidate genes. Analysis

at the RNA level invariably relies on the specificity of nucleic acid hybridization, with the target mRNA recognized either directly using a complementary labeled probe, or indirectly through amplification of cDNA using the polymerase chain reaction (PCR) and specific primers. The methods also differ considerably in their sensitivity. The easiest methods, the Northern blot and RNA dot blot, simply involve the transfer of RNA from a complex source to a membrane either without prior fractionation (dot blot) or after fractionation by electrophoresis (Northern blot), followed by probing with a complementary nucleic acid and detection of the signal. This allows rough quantitation and, in the case of the Northern blot, size determination. However, it is difficult to detect rare transcripts because of the lack of sensitivity.

More sensitive techniques include nuclease protection and reverse transcriptase PCR (RT-PCR). In nuclease protection, a labeled antisense RNA probe is hybridized in solution with a complex RNA population. If its target is present, it forms a duplex and is thus protected from degradation with a nuclease specific for single-stranded RNA and can be visualized on a sequencing gel. In the RT-PCR method, a population of mRNA molecules is reverse-transcribed to generate an equivalent population of cDNAs, which are then amplified by the PCR. Through the incorporation of suitable labels, this reaction can be rendered semiquantitative or fully quantitative, and moderate multiplexing is possible.

High-Throughput Methods

The aforementioned traditional techniques are not suitable for large-scale analysis because of the materials and labor involved. Two new methods have been devised which allow the simultaneous monitoring of all genes – one based on the high-throughput sampling of cDNA sequences and one based on miniature devices called microarrays, which allow mRNA levels to be measured by hybridization. Sequence sampling is based on the principle that a strongly expressed gene will be abundantly represented in the mRNA population of a particular cell or tissue, and will be similarly represented in an unbiased cDNA library or sequence database derived from that population. Therefore, random sampling will tend to pull out clones or sequences of strongly expressed genes much more frequently than those of poorly expressed genes. Relative expression levels can be determined simply by counting the score for each sequence pulled out during sampling, and because these data are in the form of real numbers, they are digital. A method called serial analysis of gene expression (SAGE) accelerates the sampling process by reducing each cDNA to a short tag 9–15 bp in length, and allowing up to 200 such tags to be concatemerized prior to sequencing.

Microarrays are miniature devices comprising a collection of immobilized DNA sequences arranged in a grid. Each address on the grid corresponds to a unique gene, or at least to a unique mRNA if the gene undergoes alternative splicing. These arrayed targets are hybridized (interrogated) with a complex mRNA or cDNA probe, that is, a probe comprising many different sequences representing the entire population of mRNAs from the cell or tissue under investigation. In this case, quantitative data are obtained not by counting, but rather by measuring the strength of the hybridization signal at each address. If the source mRNA/cDNA probe is universally labeled, the most abundant mRNA/cDNA molecules in the population will generate quantitatively stronger signals on the array, as long as hybridization does not reach saturation. Both sequence sampling and microarrays can be used for comparative expression profiling, in one case by comparing the representation of different sequences in two different tissues (e.g., prefrontal cortex from normal brain and schizophrenic brain) and in the other case by using two complex probes, labeled with different fluorophores, and comparing the signal intensity of each fluorophore for each gene.

See also: Alternative Splicing in the Nervous System; Gene Expression Regulation: Activity-Dependent; Gene Expression Regulation: Chromatin Modification in the CNS; Genetic Influence on CNS Gene Expression: Impact on Behavior; Genomic Disorder and Gene Expression in the Developing CNS.

Further Reading

Brouwers N, Sleegers K, Engelborghs S, et al. (2006) Genetic risk and transcriptional variability of amyloid precursor protein in Alzheimer's disease. *Brain* 129: 2984–2991.

Feng J, Fouse S, and Fan G (2007) Epigenetic regulation of neural gene expression and neuronal function. *Pediatric Research* 61: 58R–63R.

Guyant-Marechal L, Rovelet-Lecrux A, Goumidi L, et al. (2007) Variations in the *APP* gene promoter region and risk of Alzheimer disease. *Neurology* 68: 684–687.

Holm PC, Mader MT, Haubst N, et al. (2007) Loss- and gain-of-function analyses reveal targets of Pax6 in the developing mouse telencephalon. *Molecular and Cellular Neuroscience* 34: 99–119.

Hsieh J and Gage FH (2005) Chromatin remodeling in neural development and plasticity. *Current Opinion in Cell Biology* 17: 664–671.

Mirnics K, Levitt P, and Lewis DA (2006) Critical appraisal of DNA microarrays in psychiatric genomics. *Biological Psychiatry* 60: 163–176.

Morris CM and Wilson KE (2004) High throughput approaches in neuroscience. *International Journal of Developmental Neuroscience* 22: 515–522.

Nichols P, Croxen R, Vincent A, et al. (1999) Mutation of the acetylcholine receptor epsilon-subunit promoter in congenital myasthenic syndrome. *Annals of Neurology* 45: 439–443.

Perini G and Tupler R (2006) Altered gene silencing and human diseases. *Clinical Genetics* 69: 1–7.

Santos M, Coelho PA, and Maciel P (2006) Chromatin remodeling and neuronal function: Exciting links. *Genes, Brain and Behavior* 5: 80–91.

Tsankova N, Renthal W, Kumar A, et al. (2007) Epigenetic regulation in psychiatric disorders. *Nature Reviews Neuroscience* 8: 355–367.

Gene Expression in Normal Aging Brain

T E Morgan, University of Southern California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Genes, lifestyle, and life circumstances all contribute to aging. Impairments in cognitive and motor skills commonly occur with increasing age. These progressive changes in cognition are subtle but become apparent later in life. Aging is a major risk factor for numerous neurological disorders. Cognitive and motor impairments resulting from neurodegenerative diseases are distinct from age-related changes and are influenced by peripheral diseases, such as diabetes and cardiovascular disease. In fact, there are strong associations between cardiovascular disease, diabetes, and Alzheimer's disease (AD).

The complex process of brain aging involves multiple systems, cells types, and cellular pathways. As with aging in general, numerous mechanisms have been proposed to underlie brain aging, including instability of nuclear and mitochondrial genomes, neuroendocrine dysfunction, production of and damage from reactive oxygen species, altered calcium metabolism, and inflammation-mediated neuronal damage. The common thread woven through all of these mechanisms is changes in gene expression.

Minimal Neuron Loss with Age

Prior to advances in stereological procedures for estimating neuron number it was widely believed that brain aging was accompanied by significant neuron death. Today, as verified by electron microscopy, stereologic counts, and gene expression data, brain aging is not due to loss of neurons. Instead, loss of synapses, morphologic changes in dendritic arbors and spines, and changes in myelination contribute to age-related cognitive and motor impairment. The causes of these changes are not known. Molecular studies with transgenic mouse models indicate that specific genetic manipulation can lead to defects in long-term potentiation (LTP) and learning in the absence of degenerative changes (e.g., *N*-methyl-D-aspartate receptor subunits and calcium- and calmodulin-dependent kinase II). These studies support the important role that specific genes have in brain aging.

Gene Alleles: Single Nucleotide Polymorphisms

Brain aging is not only influenced by changes in gene expression but is also affected by the particular gene allele one inherits. Specific sequence variations (single-nucleotide polymorphism, or SNP), which make people more or less susceptible to certain diseases, have also been shown to influence age-related cognitive changes. An SNP in the coding or regulatory region can affect transcription, translation, and ultimately function. For example, in humans, there are three unique alleles for the apolipoprotein E (*APOE*) gene (designated *APOE* $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$). *APOE* $\varepsilon 3$ allele is most prevalent, followed by *APOE* $\varepsilon 4$, with rare occurrence of *APOE* $\varepsilon 2$. Genetic linkage studies indicate that the inheritance of *APOE* $\varepsilon 2$ is associated with increased longevity. The neurotrophic activities of *APOE* are allele specific, with $\varepsilon 2 > \varepsilon 3 > \varepsilon 4$. Furthermore, inheritance of *APOE* $\varepsilon 4$ increases the risk for AD. Another example is brain-derived neurotrophic factor (BDNF), a neurotrophin that is essential for neuronal growth, survival, and function. In addition, BDNF is a critical factor in learning and memory. An SNP identified in the coding region of BDNF (Val66Met), which affects the pro-region of the protein, is associated with age-related cognitive deficits.

Gene Expression

Most studies that examine age-related gene expression compare mRNA levels in brain tissue of different age groups. A number of important caveats must be acknowledged when interpreting these studies. First, gene expression is highly selective to different brain regions and specific cell types. Second, alterations of mRNA levels may represent changes in gene expression or mRNA stability or turnover. Third, changes in mRNA levels may not ultimately reflect translational or functional changes. Fourth, aging changes can be quite subtle, which requires confirmation of each potential age-regulated gene by multiple methods and labs, and even in a variety of species. Unfortunately, most studies do not look at the 'middle years' of a species' life span, and results are generally reported from only two age groups, young and old. Careful examination at multiple ages throughout the life span is necessary for identifying potential intervention strategies.

Age-related pathways that may contribute to aging are difficult to ascertain in humans. Since gene expression in the human brain can only be examined at autopsy, the changes measured may be causing aging or may be a consequence of aging. Due to this difficulty, animal models provide a reasonable approach to studying brain aging. In addition, specific pathways can be modeled *in vitro* but these studies are complicated by the fact that neurons are postmitotic.

Some of the first studies documenting changes in gene expression with age showed declines in specific pre- and postsynaptic markers which are often correlated with age-related decline of cognition and motor skills. Synaptophysin, a presynaptic vesicular protein that is often used as a marker for synaptic plasticity and integrity, decreases with increasing age in hippocampal and cortical regions. Specific subunits of *N*-methyl-D-aspartate (NMDA) receptors (e.g., NR1, NR2), which have critical roles in learning and memory, decrease with age. Also, dopamine receptor (D2R) levels decrease with increasing age. The decline in expression of these synaptic components effectively reduces synaptic transmission, which contributes to age-related cognitive and motor changes.

In addition to these neuronal changes, age-related gene expression also occurs in glia. Glial fibrillary acidic protein (GFAP) is an astrocyte-specific intermediate filament that consistently shows elevation in aging brain. Beginning at middle age, the increased expression of GFAP ranges from 1.5 to 4 times normal, depending on brain region. Widely accepted as an indicator of astrocyte activation, the consequence of the progressive increase during aging is not known. Recent data suggest that the age-related increase of GFAP may contribute to the loss of synaptic plasticity during aging. Astrocytes originating from aged rat brains possess elevated GFAP and are less able to support neurite outgrowth. Experimental manipulation of GFAP levels (increase with complementary DNA; decrease with small interfering RNA) demonstrates a direct correlation between GFAP levels and neurite outgrowth. This exciting line of discovery implicates astrocytes as a driver of age-related cognitive changes.

Myelin-associated oligodendrocyte basic protein (MOBP) is increased with age in oligodendrocytes. Although MOBP is the third most abundant protein in myelin, its exact physiological function is not known. Based on its localization to compact myelin, it is proposed to have a role in stabilization of the myelin. White-matter integrity declines with increasing age. In addition, a decline in the volume of brain white matter had been shown with magnetic resonance imaging (MRI). The impact of these age-related changes in the ultrastructure of the myelin sheath may have a profound effect on neuronal function.

Another glial gene that is increased with age is apolipoprotein D (ApoD). As with many of the age-regulated genes, the exact function of ApoD is not known. It is widely expressed by both astrocytes and oligodendrocytes. The ApoD protein can transport small hydrophobic ligands, including cholesterol, pregnenolone, and sterols, but it is not known if these are its physiological ligands. Importantly, the *Drosophila* homolog has been shown to increase stress resistance and extend life span. In studies of transgenic flies that overexpress the homolog of human ApoD, the protein named glial Lazarillo (GLaz), the flies were less susceptible to hyperoxic conditions, had improved motor and behavioral activities when subjected to sublethal oxygen exposure, and lived 29% longer. Therefore, the age-associated increase in ApoD expression may be a compensatory response to the age-related progressive increase in oxidative stress.

Genotypic changes in microglia are believed to underlie their activated phenotype which appears with increasing age. One indication of this is the increased expression of CD68 (the mouse homolog is macrosialin). CD68 is a lysosomal/endosomal-associated membrane protein with characteristics of scavenger receptors. Its function is not known but is believed to play a role in phagocytosis. Under normal conditions CD68 is predominantly found in microglia residing in white matter. Since oxidized lipids induce CD68 expression, it may be playing a role in age-related myelin changes.

Studies examining the expression of individual genes that change with age are vital to advancing our knowledge of the molecular mechanisms underlying aging; however, such studies do not address the polygenetic component of brain aging. This need is met by large-scale microarray analysis, which allows simultaneous examination of thousands of genes in a single sample. This type of analysis has been performed on brains from many mammalian species, including humans, chimpanzees, and rodents. There is significant change with age in a relatively small percentage (<5%) of the >20 000 genes examined. Although specific genes across all species may not be identical, classes of genes emerge (Table 1). The greatest percentage of genes that are upregulated consists of those involved in inflammatory processes and the stress response. Downregulated genes concentrate in synaptic function and plasticity. A comparison of gene classes in different species suggests that inflammatory and oxidative pathways are at the center of age-related cognitive and motor deficits.

An important group of inflammation-related genes that are increased in the aging brain includes components of the complement cascade. Resident brain cells

Table 1 Functional gene classes showing consistent changes with brain aging

Age change	*Functional class*	*Gene family*
Increase	Inflammatory and stress response	Complement subunits, heat shock proteins, cathepsins, inflammatory mediators
Decrease	Synaptic function and plasticity	Presynaptic and postsynaptic proteins, neurotransmitter receptors

produce each member of the complement cascade. Activation of the complement cascade initiates innate immune responses such as phagocytosis, chemotaxis, and cytolysis. The initial components, C1q, C3, and C4, are commonly shown to be increased in the aging brain, suggesting that activation of the complement cascade occurs during aging.

Brain Aging across Species

A new informatics tool to compare microarray results from different specifies and/or different laboratories is the web database and analysis platform, Gene Aging Nexus (GAN). GAN is freely accessible online (see 'Relevant website' section), where new datasets can be uploaded, queries can be run on existing datasets, and extensive integrative analysis can be performed to identify differentially or co-expressed genes with functional annotation. The database currently has data for the aging brain in humans, mice, and rats. Seven genes are identified to be increased with age across these three species (Table 2).

This cross-species comparison identifies known age-changers such as *GFAP* and *APOD*, but also identifies age-regulated genes that have not been documented prior to the microarray analysis. For example, annexin A3 (gene *ANXA3*) increases in human, mouse, and rat brains. Annexin A3 is a member of the annexin family of proteins that have calcium-dependent phospholipid-binding activities. The function of annexin A3 is not known; it is expressed in activated microglia in response to neuron injury. Carbonic anhydrase IV (gene *CA4*) expression is decreased in the aged brain. Reduced levels of carbonic anhydrase IV may contribute to age-related declines in synaptic transmission, since this protein is responsible for buffering the alkaline pH shifts associated with neuronal activity.

It should be noted that some of the individual genes identified as age changers were not shown to be changed during aging in the microarray studies. Detection of individual genes is influenced by the specific set of oligonucleotide probes used on a particular microarray chip, which may explain these differences. Also, differences could arise due to the complex statistical considerations microarray analysis requires.

Brain Aging and Insulin-Like Growth Factor

A discussion of aging is not complete without including insulin-like growth factor-1 (IGF-1). The importance of the insulin/IGF-1 pathway in aging and longevity is apparent from the vast experimental literature documenting a central role for IGF-1 pathways in regulating the life span of diverse organisms, including yeast, worms, flies, and mice. IGF-1 pathways orchestrate cell metabolism, growth, and survival. Systemic IGF-1 decreases with increasing age. IGF-1 crosses the blood–brain barrier and resident brain cells express IGF-1. Correlational and experimental observations suggest that IGF-1 has a generally positive influence on cognitive performance. At the cellular level, IGF-1 is neurotrophic and neuroprotective while participating in the stress response. The influence IGF-1 has on cellular activities predicts a critical role for IGF-1 pathways in brain aging.

Table 2 Age-regulated genes common to human, rat, and mouse brain: fold change (*p*-value)

Human gene nomenclature/protein	*Human*	*Mouse*	*Rat*
ANXA3/annexin A3	+2.2 (0.006)	+1.4 (0.04)	+1.3 (0.002)
APOD/apolipoprotein D	+2.0 (0.000 05)	+2.6 (0.000 03)	+1.4 (0.003)
CA4/carbonic anhydrase IV	−1.5 (0.01)	−1.9 (0.009)	−1.3 (0.009)
EGR2/early growth response 2	−1.6(0.04)	+1.6(0.04)	+2.8(0.000 9)
GFAP/glial fibrillary acidic protein	+1.5(0.01)	+2.1(0.000 1)	+1.7(0.05)
MOBP/myelin-associated oligodendrocyte basic protein	+2.1(0.03)	+1.8(0.007)	+1.6(0.02)
QDPR/quinoid dihydropteridine reductase	+1.4(0.04)	+1.3(0.04)	+1.6(0.000 003)

Age-regulated genes identified from three brain aging datasets of gene expression microarray analysis in Gene Aging Nexus.
Human data from B Yankner, Harvard Medical School; mouse data from T Prolla/R Weindruch, University of Wisconsin-Madison; rat data from P Landfield, University of Kentucky.

Regulation of Gene Expression

Gene expression is regulated through transcription, translation, and posttranslational means. Transcriptional factors, which are at the center of gene regulation during development, may also have a major role during aging. Generally, transcription factors exist in an inactive form and become active through phosphorylation, ligand binding, or removal of inhibitory subunits. Likely candidates involved in regulating age changes include activator protein-1 (AP-1), glucocorticoid receptor (GR), cyclic AMP response element-binding protein (CREB), peroxisome proliferator-activated receptor (PPAR), and nuclear factor-kappa B (NF-κB). Each transcriptionally activates unique signal transduction pathways that are implicated in learning and memory. The promoter region of the age-regulated gene *GFAP* contains putative transcriptional binding sites for all of these transcription factors. Although transcriptional regulation of *GFAP* through these sites has been extensively documented, it is still unresolved which ones are responsible for increasing *GFAP* expression with age. Additional levels of regulation have recently been identified. These include the silencing RNAs, microRNAs, and chromatin acetylation pathways. The contribution each of these has in aging is not known.

Attenuation of Brain Aging

Dietary restriction (DR; also called caloric restriction, or CR), which is achieved through restricted caloric intake, is the most powerful and general manipulation of aging processes in laboratory animals. In addition to extending life span, DR attenuates many of the age-related changes that occur in the brain. Depending on the brain region, 30–75% of the age-associated gene expression changes are either completely or partially prevented by DR. In the neocortex, DR prevents the age-related increased expression of 65% of those genes involved in the inflammatory response.

In addition to DR, specific dietary nutrients that possess anti-inflammatory and antioxidant properties have beneficial effects on brain aging. For example, supplementing rodent diets with strawberry, blueberry, or spinach extracts attenuates age-related motor and cognitive declines. Polyphenolics, such as the flavonoids found in these dietary supplements, are believed to underlie this beneficial effect. The omega-3 unsaturated fatty acids, such as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), improve cognitive performance. The phytoalexin, resveratrol, also has anti-aging properties. Since resveratrol is concentrated in the skin of red grapes, one has an additional reason to savor that glass of burgundy.

Brain Aging and AD

Since aging is the primary risk factor for AD it should not be a surprise that age-regulated genes are also found in AD. In fact, inflammatory and oxidative pathways are implicated in AD progression. Activated astrocytes and microglia are integral components of AD-related pathologies. Furthermore, using transgenic mouse models of AD, experimental reduction in disease progression occurs with many of the methods shown to be beneficial for brain aging (e.g., DR, DHA, blueberry extract, and resveratrol).

Technologic Advances

New technologies will soon advance our understanding of brain aging significantly. Microarray analysis is now available for screening microRNAs and SNPs. Laser capture microdissection (LCM) will allow researchers to perform molecular analysis of pure cells from specific microscopic regions of the brain; in addition, the synergism of molecular and computational biology disciplines is poised to make major contributions to the complex, multifaceted process of brain aging.

Open Questions

Some of the critical issues that remain to be addressed include the unique roles that transcription, translation, and posttranslational modification (which currently is a major target of drug development) have in orchestrating age-related changes in motor and cognitive skills. The delineation between normal age-related changes and disease-related changes, such as those that occur in AD and Parkinson's disease, will identify unique molecular mechanisms underlying normal brain aging. Also, careful consideration must be given to the specific contributions made by vascular aging in the aging brain.

Summary

Underlying normal age-related cognitive and motor deficits are gene expression changes that define the aging brain. The specific changes in gene expression that occur in each type of brain cell contribute to subtle dysfunctions in brain activities. A comparison of gene classes in the different species supports the belief that the aging brain experiences progressive increases of inflammatory and oxidative stress pathways, and that this results in the accumulation of altered or damaged proteins that interrupt normal synaptic functions. The current status of age-related gene expression supports an important role for glia in orchestrating these age changes in the brain.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: Transgenic Mouse Models; Alzheimer's Disease: MRI Studies; Animal Models of Alzheimer's Disease; Genomics of Brain Aging: Nuclear and Mitochondrial Genomes; Genomics of Brain Aging: Apolipoprotein E; Genomics of Brain Aging: Twin Studies.

Further Reading

Blalock EM, Geddes JW, Chen KC, et al. (2004) Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America* 101: 2173–2178.

Lee CK, Weindruch R, and Prolla TA (2000) Gene-expression profile of the ageing brain in mice. *Nature Genetics* 25: 294–297.

Lu T, Pan Y, Kao SY, et al. (2004) Gene regulation and DNA damage in the ageing human brain. *Nature* 429: 883–891.

Morgan TE, Wong AM, and Finch CE (2006) Anti-inflammatory mechanisms of dietary restriction in slowing aging processes. In: Chanson P (ed.) *Endocrine Aspects of Successful Aging*, pp. 191–206. Berlin: Springer.

Pan F, Chiu CH, Pulapura S, et al. (2007) Gene Aging Nexus: A web database and data mining platform for microarray data on aging. *Nucleic Acids Research* 35: D756–D759.

Relevant Website

http://gan.usc.edu – Gene Aging Nexus (online database of aging-related data developed at USC Davis School of Gerontology).

Gene Expression in the Evolution of the Human Brain

M C Oldham and D H Geschwind, UCLA School of Medicine, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Humans and chimpanzees diverged from a common ancestor within the past 5–7 My. During this relatively brief evolutionary time period, humans have acquired a number of singular characteristics, including a bipedal gait, the widespread use of tools, and a vastly expanded neocortex. Above all else, it is our unique cognitive abilities that separate us from other primates, endowing us with the specific epithet *sapiens*: the 'wise' or 'intelligent' human. What changes precipitated the remodeling of the human brain? Although we may never know with certainty the selective forces that shaped the recent evolution of our species, the telltale signature of natural selection must exist in the finite number of DNA sequence differences that distinguish our genome from those of our closest living relatives.

The first detailed molecular comparisons between humans and chimpanzees were performed in the 1960s and 1970s and produced a remarkable conclusion that holds true today: the overwhelming majority of macromolecules are nearly identical between the species. The degree of similarity between human and chimpanzee proteins suggested that, despite our overt phenotypic differences, we are as similar to one another molecularly as sibling species. Furthermore, these observations suggested that the evolution of many human-specific traits may have been driven by changes affecting the regulation of gene expression, in addition to changes in protein-coding sequences. The recent completion of the human and chimpanzee genome sequences and the advent of high-throughput technologies for surveying gene expression have begun to extend and refine these ideas.

Comparison of the human and chimpanzee genomes has revealed that they are approximately 96% identical, with about 1% of the differences resulting from single base substitutions and about 3% of the differences resulting from numerous insertions, deletions, duplications, inversions, and rearrangements of DNA sequences of various size. The majority of these differences exist outside of protein-coding regions, which is not surprising, since exonic sequence comprises only a small fraction of primate genomes. At present, it is difficult to discern on the basis of sequence data alone which of these changes have arisen through positive selection and which are the result of neutral evolution (i.e., genetic drift). For example, one commonly used metric assesses the ratio of single nucleotide changes that result in nonconservative amino acid differences and single nucleotide changes that do not affect coding sequence. However, this metric ignores many types of evolutionary change that may have significant functional consequences, including changes in introns that may alter splicing, changes in 3′-untranslated regions that may alter messenger RNA (mRNA) stability or trafficking, and changes in promoter regions that may alter the level of gene expression. Because many functionally significant changes in noncoding regions of DNA exert their effects by altering the regulation of gene expression, a new approach to the study of human brain evolution involves large-scale comparisons of gene-expression levels between human and chimpanzee brains. Here, we discuss what we have learned about brain evolution from these studies.

Comparing the Transcriptomes of Human and Chimpanzee Brains

The transcriptome is broadly defined as the set of all mRNA molecules that are transcribed in a given organism in a given tissue (or cell type) at a given time. In reality, no technique currently exists to survey all transcripts in any multicellular organism, as no complete set of transcripts, expressed across all developmental time points, is definitively known. However, existing techniques can measure the expression levels for upward of 20 000 transcripts in a single experiment. The most popular of these techniques involves the use of microarrays. A microarray consists of thousands of molecular probes arranged in a small grid, each designed to hybridize to a specific mRNA, thereby providing a quantitative measure of the abundance of that mRNA in a sample (i.e., the degree of gene expression). Recently, investigators have begun to use microarrays to compare expression levels for thousands of genes in adult human and chimpanzee brains, as well as other tissues. Below, we review the results of these initial studies and discuss some of the challenges that must be addressed to enable proper interpretation of the data produced by this promising experimental approach.

Initial Findings

As with humans, it is considered unethical to sacrifice a great ape for the sole purpose of obtaining tissue samples. As a result, microarray experiments

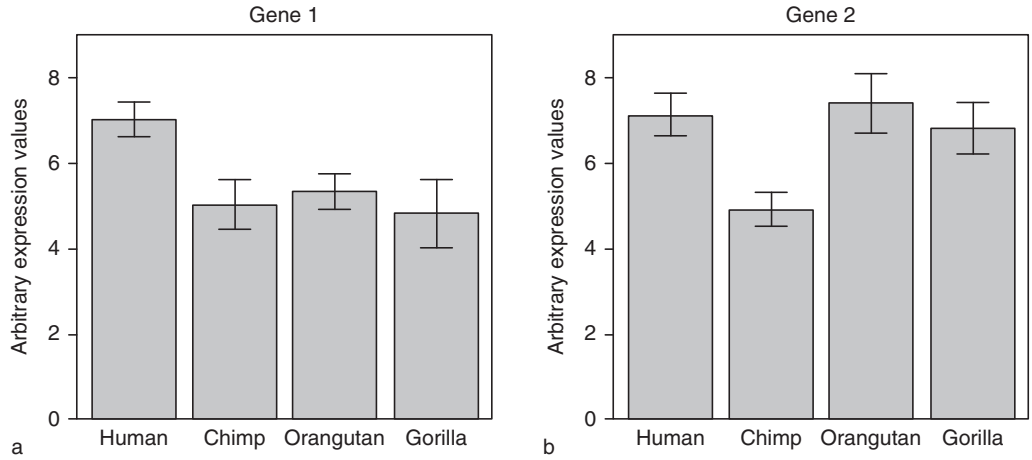

Figure 1 Outgroup data are required to infer the direction of evolutionary changes in gene-expression levels. (a) Mean expression levels are shown for a hypothetical gene in human, chimpanzee, orangutan, and gorilla. Although the gene is expressed higher in humans relative to chimpanzees, without outgroup data it is unclear whether expression levels have increased on the human lineage or decreased on the chimpanzee lineage. Comparison with orangutan and gorilla suggests an increase in expression on the human lineage. (b) Mean expression levels for another hypothetical gene are shown. The difference between humans and chimpanzees is comparable to that in (a). However, comparison with orangutan and gorilla suggests a decrease in expression on the chimpanzee lineage.

comparing gene expression between humans and chimpanzees have been carried out using mRNA extracted from postmortem tissue samples from individuals who have died of natural causes. Certainly, the cause of death or the process of death itself may influence gene expression, which is a potential confound in these studies. However, it is reassuring that gene-expression levels from postmortem tissue that has been rapidly frozen correlate strongly with gene-expression levels derived from tissue extracted during surgical biopsies from living human subjects, lending confidence to the integrity of this experimental approach. Microarray studies comparing gene expression in human and chimpanzee brains have focused primarily on cerebral cortex, although comparisons have also been performed for cerebellar vermis, caudate nucleus, and some nonneural tissues. An initial and perhaps somewhat counterintuitive observation is that humans and chimpanzees appear to have fewer gene-expression differences in prefrontal cortex than they do in nonneural tissues. This observation was first made in comparison to liver and subsequently extended to heart, kidney, and testis. These data have been interpreted to suggest that the brain may impose greater functional and evolutionary constraints on gene expression than other tissues. However, there are technical obstacles and other explanations that challenge this interpretation, which are discussed below.

An important question is whether gene-expression differences between humans and chimpanzees reflect evolutionary changes that have occurred on the human lineage, the chimpanzee lineage, or both. To answer this question, it is necessary to compare human and chimpanzee gene-expression levels with those of another closely related species (or outgroup), such as gorilla or orangutan (Figure 1). Unfortunately, tissue samples from these species are extremely scarce, often necessitating the use of more distant relatives for which captive colonies exist in large numbers, such as rhesus macaque. Comparisons of human/chimpanzee gene-expression differences with outgroup data have yielded another interesting observation: in the brain, but not in other tissues, a greater proportion of gene-expression differences between humans and chimpanzees seems to have occurred on the human lineage. Thus, although the brain appears to show fewer gene-expression differences overall between humans and chimpanzees compared with other tissues, relatively more of those changes are human specific.

Is there any pattern to the gene-expression differences between human and chimpanzee brains? *A priori*, one might expect that approximately equal numbers of genes would show higher expression in human and chimpanzee brains, respectively. However, there is some evidence to suggest that the distribution of gene-expression differences might be skewed, with more differentially expressed genes showing higher expression in human cerebral cortex relative to chimpanzee. Interestingly, among genes with higher expression levels in the human brain, several functional categories are enriched, including cell growth and maintenance, protein folding, and energy metabolism. The last category is especially noteworthy, as multiple lines of evidence are beginning to

converge upon the idea that genes involved in energy metabolism, including several members of the electron transport chain, have experienced accelerated evolution in anthropoid primates. In light of the massive expansion of the neocortex that has occurred on the human lineage within the past 5–7 MY, it is reasonable to expect that solutions to newly imposed energetic demands must have emerged concurrently.

Another important question is whether gene-expression differences between human and chimpanzee brains are ubiquitous or whether they are confined to a particular brain region or cell type. A comparison of four cortical regions (Broca's area, anterior cingulate cortex, prefrontal cortex, and primary visual cortex), caudate nucleus, and cerebellar vermis by Khaitovich and colleagues is the only study that has attempted to address this question directly. These authors observed very little variation in gene expression between cortical regions in individuals of either species (although there is some evidence to suggest that primary sensory cortex is distinguishable from areas typically termed 'association' cortex). In contrast, clear differences in gene expression exist between cortex, caudate nucleus, and cerebellum within individuals from both species. Analysis of genes that are differentially expressed between humans and chimpanzees reveals that the vast majority of gene-expression differences between the species are found in all of the brain regions surveyed. Among those genes that show expression patterns specific to a single brain region (cortex, caudate nucleus, or cerebellum), there appears to be a slightly greater percentage of human-specific cortical genes than there are human-specific caudate nucleus or cerebellar genes. However, the absolute number of these genes is quite low. Furthermore, direct comparison between humans and chimpanzees in Broca's area, a region of cortex known to be critically involved in a human-specific phenotype (language), did not identify any major differences in gene expression between the species.

Challenges to Data Interpretation

An important caveat is that, owing to the scarcity of human and chimpanzee tissue samples, all of the findings described above have been derived from comparisons of 3–6 individuals from each species. Because cortical gene expression can vary significantly between individuals (indeed, differences between individuals are typically greater than differences between cortical areas from the same individual), 'outlier' samples have the potential to significantly alter data interpretation. For example, it has been suggested that the apparent acceleration of gene-expression changes in the brain on the human lineage might result from greater variability in gene expression among the analyzed human samples relative to the chimpanzee samples, which could reflect true individual differences or differences in postmortem mRNA quality. In addition, many of the studies described above used samples from related chimpanzee individuals, depressing their degree of genetic diversity compared to the unrelated human individuals. Dissociating individual variability from technical variability will require many more samples representing different regional populations, genders, and ages. Thus, caution is required when attempting to generalize these findings. The issue of sample size and composition also holds true for the use of outgroup species. When outgroups have been used to infer the directionality of changes in gene expression between humans and chimpanzees, they have consisted of a handful of individuals (from one to four orangutans, gorillas, or rhesus macaques).

The use of outgroup species highlights another challenge to data interpretation – the effect of interspecies sequence differences on hybridization efficiencies between microarray probes and their mRNA target sequences. Because the probes have been designed from human mRNA sequences, they will not match perfectly with their mRNA targets in closely related primate species (including chimpanzees). Conservation between human and chimpanzee coding regions is ~99%, implying that one out of every four 25-mer probes (the probe length on Affymetrix microarrays, which have been used in all of the studies discussed above) will have a mismatch. A sequence mismatch can reduce the hybridization efficiency between a probe and its mRNA target, resulting in expression levels that appear 'decreased' in nonhuman primate species or 'increased' in humans. This effect clouds the observation that genes differentially expressed between human and chimpanzee brains are skewed toward higher expression in humans (although skewing of gene-expression differences was not evident in nonneural tissues such as heart or liver). However, this problem can be mitigated by using available sequence data to mask probes with sequence differences between species before calculating gene-expression values. Because each transcript on an oligonucleotide microarray is queried by multiple probes, eliminating ~1/4 of the probes (as might be done for chimpanzee samples) does not preclude data collection. Indeed, application of this approach to chimpanzee gene-expression data appears to preserve the skewing of gene-expression differences between the species. Unfortunately, this approach is untenable for more evolutionarily distant outgroup species, for which most probes are expected to possess sequence differences with their intended

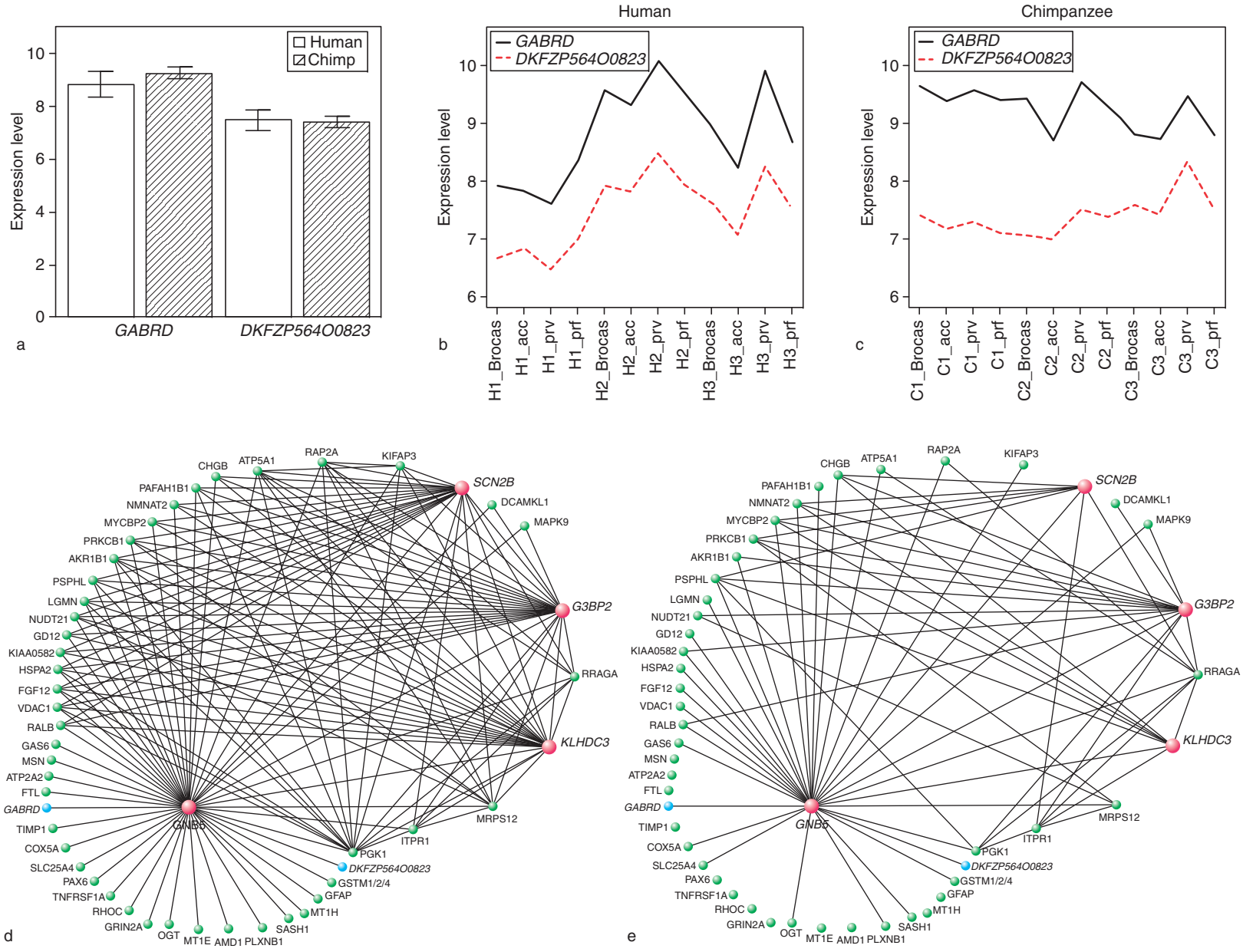
Human
Chimp
Expression level
GABRD
DKFZP564O0823
a
Human
Chimpanzee
GABRD
DKFZP564O0823
Expression level
H1_Brocas
H1_acc
H1_prv
H1_prf
H2_Brocas
H2_acc
H2_prv
H2_prf
H3_Brocas
H3_acc
H3_prv
H3_prf
b
C1_Brocas
C1_acc
C1_prv
C1_prf
C2_Brocas
C2_acc
C2_prv
C2_prf
C3_Brocas
C3_acc
C3_prv
C3_prf
c
SCN2B
G3BP2
RRAGA
KLHDC3
GNB5
CHGB
ATP5A1
RAP2A
KIFAP3
PAFAH1B1
NMNAT2
MYCBP2
PRKCB1
AKR1B1
PSPHL
LGMN
NUDT21
GD12
KIAA0582
HSPA2
FGF12
VDAC1
RALB
GAS6
MSN
ATP2A2
FTL
GABRD
TIMP1
COX5A
SLC25A4
PAX6
TNFRSF1A
RHOC
GRIN2A
OGT
MT1E
AMD1
PLXNB1
SASH1
MT1H
GFAP
GSTM1/2/4
DKFZP564O0823
PGK1
ITPR1
MRPS12
DCAMKL1
MAPK9
d
e

targets. Ultimately, the solution to this problem may depend upon the development of species-specific microarrays with probes designed to match orthologous exons.

Microarray studies comparing human and chimpanzee brains pose several additional challenges. First, all studies to date have surveyed gene expression in tissue samples from adult individuals, although many of the most important evolutionary adaptations must manifest during fetal development. It is unclear to what extent such adaptations will persist into adulthood at the level of gene expression. Furthermore, it is unlikely that fetal chimpanzee tissue samples will become available anytime soon; therefore, additional strategies aimed at elucidating the unique aspects of the transcriptional program in the developing human brain will be required. Second, interpretation of microarray data derived from brain samples is complicated by the issue of cellular heterogeneity. Within any given sample of brain tissue there are anywhere from dozens to hundreds of different cell types; each of these cell types is expected to vary in terms of the number and degree of expressed genes. In effect, the level of expression for a given gene measured by microarray in a sample of brain tissue represents the breadth of expression (across cell types) multiplied by the representation of each cell type in the sample. Seen in this light, it may seem less surprising that gene expression appears to show less divergence between humans and chimpanzees in the brain than it does in other, more homogeneous tissues such as liver, heart, or testis. The genes that are detected by microarray in samples of brain tissue are more likely to be genes that are broadly expressed across a variety of cell types; expression in a wide variety of cell types, in turn, is expected to impose greater functional and evolutionary constraints than expression in a few cell types, as would be the case in a more homogenous tissue. Clarification of this issue will require cellular resolution (i.e., comparisons of gene-expression levels for specific cell types). Finally, most current analyses of microarray data focus on identifying genes that show differential expression between two conditions. Since gene-expression levels in human and chimpanzee brains represent population averages across cell types and individuals, these comparisons offer a rather severe method of data reduction that do not extract all of the available information present in microarray data. It is likely that gene-by-gene assessments of differential expression will be supplanted by more comprehensive analytic techniques such as network-based approaches, which we discuss below. These approaches study gene expression in the context of the whole biological system. As such, they enable understanding of the relationships between genes and the relative importance of a gene within the system being studied.

Application of Network Theory to Gene-Expression Data

Recently there has been a great deal of interest in the application of network theory to biological systems. Many biological systems exhibit properties of complex networks, such as scale-free topology, in which the probability that a given node (a gene or its product) in the network has a certain number of connections follows an inverse power law (this property predicts that certain 'hub' nodes will have many connections and serve as convergence points in the network, whereas most nodes will have very few connections). Network analysis of gene-expression

Figure 2 Analysis of gene co-expression relationships provides insights that comparison of average gene-expression levels cannot identify. (a) Comparison of average gene-expression levels for *GABRD* and *DKFZP564O0823* suggests that neither gene is differentially expressed between human and chimpanzee cerebral cortex. On the basis of these comparisons, there is no reason to suspect that these genes have anything in common. (b, c) Network analysis reveals that *GABRD* and *DKFZP564O0823* are highly co-expressed in human cerebral cortex (b) ($r = 0.99$, $p = 6.45e{-}10$), but not in chimpanzee cerebral cortex (c) ($r = 0.11$, $p = 0.73$). *GABRD* encodes the delta subunit of the $GABA_A$ receptor, while the function of *DKFZP564O0823*, which also encodes a membrane protein, is unknown. The strength of co-expression in human cerebral cortex suggests that these genes may be co-regulated and participate in similar biological processes, while the absence of co-expression in chimpanzees suggests that coregulation of these genes may have disappeared during recent chimpanzee evolution or appeared during recent human evolution. H1, Human 1, C1, Chimp 1, etc.; Brocas, Broca's area; acc, anterior cingulate cortex; prv, primary visual cortex; prf, prefrontal cortex. Raw data are from Khaitovich P, Muetzel B, She X, et al. (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Research* 14: 1462–1473. (d) By searching for other genes that are highly co-expressed in human cerebral cortex with both *GABRD* and *DKFZP564O0823*, a module of gene co-expression relationships is revealed. The hub genes in this module are highlighted in pink: *GNB5*, *SCN2B*, *G3BP2*, and *KLHDC3*. *GABRD* and *DKFZP564O0823* (highlighted in turquoise) are seen to be highly correlated with one another by virtue of their mutual association with the hub gene *GNB5*. (e) By comparing the strength of the connections depicted in (d) between humans and chimpanzees, it becomes possible to apply a filter such that only the most human-specific connections are retained as is depicted here. So, the network module shown in (e) represents human-specific connections. This pattern suggests that this module is poorly conserved in chimpanzee cerebral cortex. The lack of conservation is largely due to the differential connectivity of the hub genes. Such detail and knowledge of which genes are the focal points of evolution cannot be obtained from simple analysis of differential gene expression.

data proceeds by treating each probe set on the microarray as a node and modeling the connections between nodes as a function of the correlations between gene-expression levels. By searching for clusters of nodes that are densely interconnected, one can identify groups of genes (modules) that are highly co-expressed. Modules of co-expressed genes may be expressed in the same cell type(s) or be involved in similar biological processes, providing an excellent method for generating 'guilt-by-association' functional hypotheses about genes for which little is known. In addition, the position of each gene in the network can be described by its connectivity (k), which is a quantitative summary of its connection strengths to all other genes in the network. Genes with high k (hub genes) are critical to the integrity and proper functioning of the network. Comparisons of k between species can thus serve to place changes in gene expression or protein sequence in an appropriate functional context and may aid in discerning between changes that are adaptive and changes that are evolutionarily neutral. Similarly, comparisons of gene co-expression relationships between human and chimpanzee brains have the potential to identify genes with vastly different patterns of connectivity. Importantly, a gene may differ dramatically in its degree of co-expression with other genes, yet have the same average expression level in both species (Figure 2). By considering genes only in terms of differential expression, a world of available information goes unnoticed. Network-based approaches can thus provide a more sensitive method for comparing transcriptomes and are likely to become increasingly prevalent in large-scale analyses of microarray data sets in the future.

In Search of a Theoretical Framework

What forces govern the evolution of gene-expression differences between primate species? The answer to this question is not yet clear, and probably will not be for some time. One view holds that most genes are under stabilizing selection and expression levels should therefore be largely conserved across species; where differences exist, they are thought to be adaptive. An alternative theory posits that most gene-expression differences between species are evolutionarily neutral or nearly neutral. This theory predicts that gene-expression differences between species should accumulate approximately linearly with time. Although limited evidence has been produced in favor of both arguments, it is clear that much more data will be required to answer this question with any certainty. It is also evident that the search for a theoretical framework governing the evolution of gene expression must take place within appropriate species and tissue contexts. Can one assume that the rules governing the evolution of gene expression are the same in human appendix as they are in human brain, or the same in fly brain as they are in human brain? In addition, the question of selective pressure cannot be disentangled from the technical hurdles described above. The gene-expression 'levels' typically used to promote various theoretical frameworks represent population averages across cell types and individuals. Consequently, individual differences and differences in cellular composition have the potential to significantly impact data interpretation, especially when considering so few samples. Moving forward, the challenge for researchers in the field will be to assign differences in gene expression levels and gene co-expression relationships to specific cell types in the human brain. It is only at this level that a clear picture of the evolutionary forces that have shaped gene-expression changes in the human brain will begin to emerge.

Summary

Over 30 years have passed since it was first proposed that changes in the regulation of gene expression may underlie many human-specific traits. Today, with the availability of the human genome sequence, draft sequences for chimpanzee and rhesus macaque, and the impending availability of genome sequences for orangutan and gorilla, we are closer than ever to zeroing in on the parts of our DNA that make us unique as a species. While this goal is within our reach, it will in many ways mark a new beginning, as researchers attempt to determine which of these sequence differences have contributed to the emergence of uniquely human traits. High-throughput techniques such as microarrays that enable transcriptomic comparisons between species offer great promise to identify functionally significant molecular phenotypes that may underlie species-specific adaptations. Application of these techniques to larger data sets consisting of increasingly refined brain regions or specific populations of cells, combined with new analytic approaches, have the potential to discern between evolutionarily neutral and functionally significant differences in gene expression between humans and other species. Understanding these differences may enable the cultivation of a theoretical framework governing the evolution of gene expression in the human brain and simultaneously reveal the critical evolutionary changes that have contributed to the emergence of our singular cognitive abilities.

See also: Brain Development: The Generation of Large Brains; Broca's Area: Evolution; Cerebellum: Evolution

and Comparative Anatomy; Comparative Neurobiology: History; Frontal Cortex Evolution in Primates; Microarray use for the Analysis of the CNS; Social Brain: Evolution.

Further Reading

Barabási AL and Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* 5: 101–113.

Cáceres M, Lachuer J, Zapala MA, et al. (2003) Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences of the United States of America* 100: 13030–13035.

Cheng Z, Ventura M, She X, et al. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437: 88–93.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69–87.

Enard W, Khaitovich P, Klose J, et al. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science* 296: 340–343.

Gilad Y, Oshlack A, Smyth GK, Speed TP, and White KP (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440: 242–245.

Gu J and Gu X (2003) Induced gene expression in human brain after the split from chimpanzee. *Trends in Genetics* 19: 63–65.

Hsieh WP, Chu TM, Wolfinger RD, and Gibson G (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* 165: 747–757.

Khaitovich P, Enard W, Lachmann M, and Pääbo S (2006) Evolution of primate gene expression. *Nature Reviews Genetics* 7: 693–702.

Khaitovich P, Hellmann I, Enard W, et al. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309: 1850–1854.

Khaitovich P, Muetzel B, She X, et al. (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Research* 14: 1462–1473.

Khaitovich P, Weiss G, Lachmann M, et al. (2004) A neutral model of transcriptome evolution. *PLoS Biology* 2: E132.

King MC and Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* 188: 107–116.

Preuss TM, Cáceres M, Oldham MC, and Geschwind DH (2004) Human brain evolution: Insights from microarrays. *Nature Reviews Genetics* 5: 850–860.

Gene Expression Regulation: Activity-Dependent

E D Nelson and L M Monteggia, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Regulation of gene expression changes is thought to be the basis for long-lasting forms of synaptic plasticity in the brain. Paradoxically, synaptic activity can regulate the transcription of some of the genes responsible for the changes in synaptic plasticity. Synaptic activity can effect gene expression through the direct control of transcription factors, both activators and repressors, as well as through the regulation of epigenetic mechanisms important for controlling gene transcription, such as DNA methylation and histone acetylation. The common starting point by which synaptic input can regulate gene expression is through the enhancement of postsynaptic Ca^{2+} influx through *N*-methyl-D-aspartate (NMDA)-type glutamate receptors and/or voltage-gated Ca^{2+} channels (particularly of the L type). This rise in intracellular Ca^{2+} levels leads to the activation of signaling pathways important for regulating certain transcription factors (Figure 1). Most often, Ca^{2+} will bind either directly to kinases that can phosphorylate substrates important for controlling gene expression, or to proteins that can then activate downstream kinases in a Ca^{2+}-dependent manner. Activation of particular kinases depends upon the temporal and spatial properties of the Ca^{2+} signals in response to synaptic inputs.

Regulation of Transcriptional Activation by Synaptic Activity

One of the best-understood examples for how activity-driven signaling pathways can induce the activation of gene expression is the phosphorylation of the transcriptional activator cyclic AMP response element (CRE)-binding protein (CREB). Upon synaptic stimulation by excitatory neurotransmitters, increased intracellular Ca^{2+} binds to the protein calmodulin, which activates a number of different stimulus-dependent protein kinases that can all lead to the phosphorylation of CREB at serine 133. Immediately after receiving synaptic stimulation, Ca^{2+}-activated calmodulin can translocate from the cytoplasm to the nucleus, where it can then activate Ca^{2+}/calmodulin-dependent kinase IV (CaMKIV), which in turn phosphorylates CREB. Ca^{2+}-bound calmodulin can also activate additional CaM kinases within the cytoplasm or other signaling cascades, such as the mitogen-activated protein (MAP) kinase and protein kinase A (PKA) pathways. After phosphorylation of CREB by one of these kinases, it can bind to CRE sequences in the promoters of certain genes and activate their transcription (Figure 2). However, phosphorylation of CREB alone is not sufficient to activate gene transcription. Increased nuclear Ca^{2+} must also activate the transcriptional coactivator CREB-binding protein (CBP), which then associates with phosphorylated CREB and leads to the expression of target genes. Many of the targets of activity regulated transcription factors are gene products known to have effects on neuronal activity. The transcriptional activator CREB controls the expression of genes such as brain-derived neurotrophic factor (BDNF) and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type glutamate receptor 1 (GluR1). BDNF, a growth factor important for neuronal survival, can mediate changes in long-term potentiation (LTP), a form of synaptic plasticity. GluR1 is an excitatory neurotransmitter receptor subunit required for LTP to occur at synapses.

Ca^{2+}-dependent shuttling of transcription factors between the cytoplasm and the nucleus is another mechanism by which activity can drive changes in gene expression. Two different transcriptional activators, nuclear factor of activated T cells (NFAT) and nuclear factor-κB (NF-κB), are relocalized from the cytoplasm to the nucleus in response to signaling cascades activated by a variety of stimuli. The activation of the Ca^{2+}-dependent phosphatase, calcineurin, leads to the dephosphorylation of NFAT, allowing NFAT to translocate to the nucleus where it can then turn on the expression of specific target genes. In contrast, NF-κB is kept inactive in the cytoplasm by inhibitor proteins, IκBs, until they are phosphorylated and degraded in response to increased intracellular Ca^{2+}. NF-κB is then free to relocate into the nucleus and activate gene transcription. The nuclear translocation of these proteins in response to synaptic activity suggests that their specific localization within the cytoplasm may influence the transcriptional response to cellular stimuli, thus making them possible regulators of gene expression-dependent localized processes such as synapse-specific modulation and/or axon guidance. Indeed, it is known that NFAT controls the expression of the inositol 1,4,5-trisphosphate receptor type 1 (IP_3R1) gene that encodes an endoplasmic reticulum Ca^{2+} channel important for releasing Ca^{2+} from intracellular stores in response to synaptic activity.

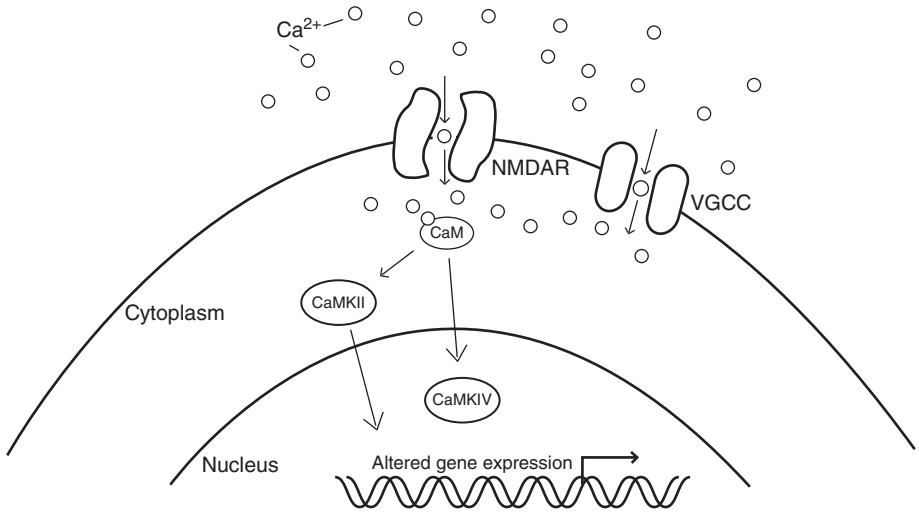

Figure 1 Example of a postsynaptic mechanism by which activity can regulate gene expression. In response to neurotransmitter release from presynaptic neurons, ligand- and voltage-gated ion channels open up and allow the influx of Ca^{2+} into the postsynaptic compartment. Two channels that are critically important for mediating this rise in intracellular Ca^{2+} are *N*-methyl-D-aspartate receptors (NMDAR) and voltage-gated Ca^{2+} channels (VGCC). The rise in Ca^{2+} concentrations can be sensed by the binding of Ca^{2+} to the protein calmodulin (CaM), which then becomes ready to activate further downstream the proteins important for the regulation of gene expression. Two such proteins are Ca^{2+}/CaM-dependent kinases II and IV (CaMKII and CaMKIV). These kinases then go on to phosphorylate transcriptional activators or repressors and thereby have an effect on gene transcription.

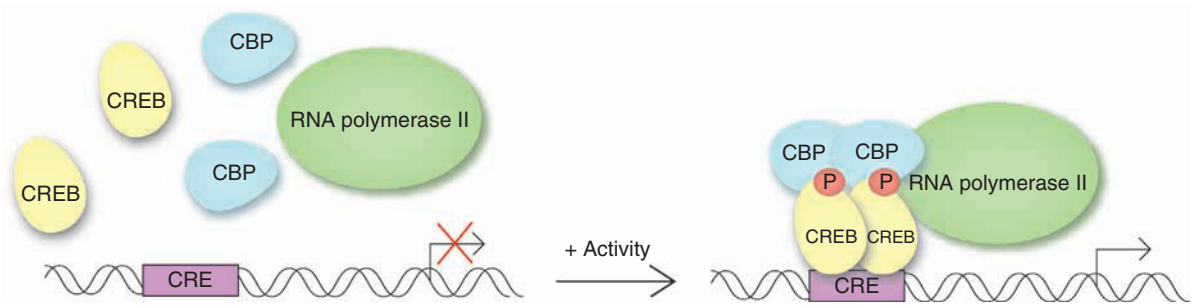

Figure 2 Gene transcription is activated upon synaptic stimulation via the phosphorylation of cAMP response element (CRE)-binding protein (CREB). The rise in intracellular Ca^{2+} after synaptic activity leads to the activation of various kinases that can phosphorylate the transcriptional activator CREB. Upon phosphorylation, CREB dimerizes and binds to CRE sequences in the upstream promoter regions of target genes. The coactivator, CREB-binding protein (CBP), must also bind to CREB before the recruitment of proteins important for gene transcription, such as RNA polymerase II, can turn on the expression of these genes.

Regulation of Transcriptional Repression by Synaptic Activity

Though much work has been done demonstrating the activation of gene expression through synaptic activity, there is accumulating evidence that activity can regulate transcription repression as well. One example is through the direct binding of intracellular Ca^{2+} to the downstream regulatory element antagonist modulator (DREAM). DREAM represses transcription by binding to downstream response element (DRE) sequences found in the untranslated regions of target genes. Upon synaptic stimulation, Ca^{2+}-binding domains within the DREAM protein bind Ca^{2+}, causing a conformational change and the release of DREAM from DRE-containing genes, thereby relieving the repression of transcription. DREAM has been shown to repress the transcription of a specific Na^{2+}/Ca^{2+} exchanger, NCX3, important for controlling intracellular Ca^{2+} homeostasis.

Another example of the activity-dependent regulation of transcriptional repression is through the Ca^{2+}-induced phosphorylation of the methyl-CpG-binding protein 2 (MeCP2). MeCP2 is a unique transcriptional repressor protein that links two different epigenetic mechanisms for controlling gene expression, DNA methylation and histone deacetylation. Under unstimulated conditions, MeCP2 associates with methylated CpG dinucleotides in the promoters

of target genes and recruits histone deacetylases (HDACs) 1 and 2. In response to neuronal activity, CaMKII phosphorylates MeCP2 at serine 421, causing its release from methylated promoters and relieving its repression of target genes (Figure 3). Recent work has demonstrated that MeCP2 controls the expression of BDNF in neurons by repressing its transcription in the absence of synaptic inputs.

Activity-Dependent Regulation of Epigenetic Mechanisms

The mention of MeCP2 introduces a manner by which changes in the interpretation of epigenetic markers can be regulated through alterations in neuronal activity. There is recent evidence that epigenetic mechanisms can be regulated by synaptic activity. The previously mentioned CREB-associated protein, CBP, can also be phosphorylated by the specific kinase Ca^{2+}/calmodulin-dependent kinase IV and likely increases transcription through its histone acetyltransferase (HAT) activity. CBP's activation by intracellular calcium is only one mechanism by which synaptic stimulation can cause chromatin remodeling that can then lead to changes in gene expression.

Activity-dependent increases in gene expression can also be regulated by the activation of another transcriptional activator, myocyte-enhancer factor 2 (MEF2), and the subcellular localization of HDACs. In quiescent cells, phosphorylated MEF2 is bound to the promoters of target genes and interacts with class II HDACs (4, 5, 7, and 9). HDAC activity represses the expression of MEF2-regulated genes until the cell receives a large enough stimulation to induce Ca^{2+} influx. The activation of CaMKs then phosphorylates the class II HDACs, causing them to be shuttled from the nucleus to the cytoplasm, therefore relieving their repression of gene transcription. Activity can then induce the dephosphorylation of MEF2 through the Ca^{2+}-dependent regulation of calcineurin, which results in active transcription of MEF2 target genes. Activity-dependent activation of the transcription factor MEF2 leads to an increase in the expression levels of activity-regulated cytoskeletal-associated protein (arc) and synaptic Ras guanosine triphosphatase activating protein (synGAP). The products of the *ARC* and *SYNGAP1* genes control synapse number and therefore have impacts on neuronal excitability.

There is recent evidence that DNA methylation, an epigenetic mechanism for controlling transcriptional repression, can also be regulated by neuronal activity. A commonly used learning and memory paradigm, contextual fear conditioning, increases the levels of DNA methyltransferase (DNMT) enzymes in the brain and alters promoter methylation patterns in two genes associated with synaptic plasticity. The molecular signaling mechanisms that drive these changes in DNA methylation are still not understood, but they are likely to involve similar increases in postsynaptic calcium levels.

In summary, there are many ways in which changes in synaptic activity result in the regulation of gene expression. Interestingly, there are known neurodevelopmental disorders associated with deficits in the activities of some of these transcriptional regulators, making their understanding of great clinical relevance. In humans, mutations in the transcriptional activating protein, CBP, cause Rubinstein–Taybi syndrome, a disorder associated with both growth and mental retardation. Mutations in the gene *MECP2* have been shown to lead to another mental retardation disorder, Rett syndrome. Future studies of these transcription factors will hopefully lead to a better understanding of how neuronal activity can alter gene expression and may give some mechanistic insights into how their malfunctions can lead to neurodevelopmental defects.

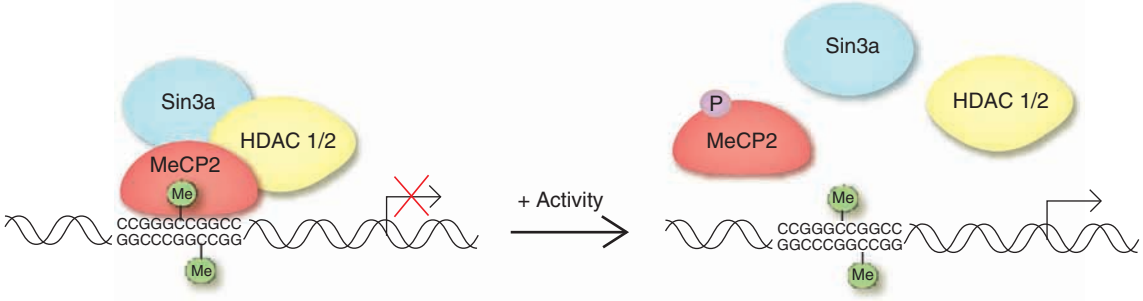

Figure 3 Transcriptional repression by methyl-CpG-binding protein 2 (MeCP2) is relieved following synaptic activity. In the absence of activity, MeCP2 is bound to methylated CpG sequences in the promoters of certain genes. The expression of these genes is repressed due to the inability of transcriptional activators to bind to promoter regions as well as to the deacetylation of histones by histone deacetylases 1 and 2 (HDAC1/2), which are recruited to chromatin by MeCP2 and the corepressor Sin3a. Upon stimulation, Ca^{2+}/calmodulin-dependent kinase II phosphorylates MeCP2, causing release of the repressor complex from CpG islands, allowing subsequent binding of transcriptional activators that can enhance gene expression.

From the Nucleus to Synapse: Transcriptional Regulation of Synaptic Function

In addition to synaptic activity-dependent regulation of gene expression, alterations in gene expression can in turn have significant impact on neurotransmission. Acute or chronic changes in the levels of activity-dependent genes can alter the degree of synaptic plasticity as well as baseline neurotransmission. The most salient example of this type of regulation is the activity-dependent alteration in levels of the brain-derived neurotrophic factor, a target gene for CREB-dependent transcription and MeCP2-dependent transcriptional repression. BDNF, via its binding to TrkB receptors, can exert several effects on neurotransmission by acting on postsynaptic as well as presynaptic mechanisms. On the postsynaptic site, BDNF action can increase the activity of the NMDA subtype of glutamate receptors and it can be a potent cofactor facilitating long-term synaptic plasticity. BDNF can also augment neuronal excitability by directly acting on voltage-gated ion channels. In the case of presynaptic nerve terminals, however, BDNF action has been proposed to target neurotransmitter release machinery directly in addition to its more global effect on presynaptic excitability. Due to its intimate association with multiple forms of synaptic plasticity and robust activity-dependent regulation through gene expression, BDNF has been a focus of several studies on the transcriptional regulation of neurotransmission.

Regulation of neuronal gene transcription via CREB or MeCP2 can have multiple, albeit specific, effects on neuronal excitability and synaptic transmission, which may not be solely attributable to alterations in BDNF levels. For instance, expression of constitutively active CREB or CaMKIV in the CA1 region of hippocampus causes a concomitant increase in NMDA receptor-mediated synaptic responses and in the extent of long-term potentiation. This increase in synaptic plasticity is attributable to a parallel increase in the number of 'silent synapses' that possess functional NMDA responses but no detectable AMPA receptor-mediated currents. As silent synapses are considered to be excellent substrates for synaptic plasticity, an increase in their number can substantially augment the dynamic range for ensuing plasticity. In addition, *in vivo* expression of active CREB in the nucleus accumbens medium spiny neurons causes a substantial increase in the excitability of these neurons and modifies cocaine-induced neuroadaptations.

Studies on loss-of-function phenotypes of MeCP2 have revealed several changes in baseline synaptic transmission as well as synaptic plasticity. Loss of MeCP2 causes impairments in long-term potentiation as well as long-term depression. This phenotype becomes evident in aged MeCP2-deficient mice at a time when the mice begin to manifest behavioral phenotypes akin to those seen in human patients with Rett syndrome. Synaptic connections formed between dissociated hippocampal neurons extracted from MeCP2-deficient mice also display several abnormalities, suggesting that MeCP2 is a *bona fide* cell-autonomous regulator of synaptic function. In this reduced system, loss of MeCP2 causes a specific decrease in the frequency of unitary excitatory neurotransmission coupled to a small increase in evoked neurotransmitter release from excitatory nerve terminals. Interestingly, treatment of wild-type control cultures with a drug that impairs histone deacetylation induces a similar decrease in unitary excitatory neurotransmission, but this decrease is occluded in MeCP2-deficient neurons. Furthermore, neither the loss of MeCP2 nor the drug treatment results in changes in inhibitory neurotransmission. Taken together, these findings suggest a role for MeCP2 in control of excitatory presynaptic function through regulation of gene expression.

The studies on CREB- and MeCP2-dependent alterations in synaptic function indicate a strong interaction between the regulatory mechanisms that tightly control gene expression in neuronal nuclei and the postsynaptic and presynaptic signaling elements. Therefore, activity-dependent regulation of gene expression forms a closed-loop system, which underlies potent feedback regulation of synaptic plasticity and neuronal homeostasis. Strong reciprocal feedback regulation of gene expression and synaptic function is clearly a key factor in sustained stability of synaptic networks, despite their vast capacity of information storage and processing.

See also: Alternative Splicing in the Nervous System; Gene Expression Regulation: Chromatin Modification in the CNS; Gene Expression Dysregulation in CNS Pathophysiology; Genetic Influence on CNS Gene Expression: Impact on Behavior; Genomic Disorder and Gene Expression in the Developing CNS; Plasticity and Activity-Dependent Regulation of Gene Expression.

Further Reading

Asaka Y, Jugloff DGN, Zhang L, et al. (2006) Hippocampal synaptic plasticity is impaired in the Mecp2-null mouse model of Rett syndrome. *Neurobiology of Disease* 21: 217–227.

Deisseroth K, Mermelstein PG, Xia H, et al. (2003) Signaling from synapse to nucleus: The logic behind the mechanisms. *Current Opinion in Neurobiology* 13: 354–365.

Dong Y, Green T, Saal D, et al. (2006) CREB modulates excitability of nucleus accumbens neurons. *Nature Neuroscience* 9: 475–477.

Klose R and Bird A (2003) Molecular biology. MeCP2 repression goes nonglobal. *Science* 302(5646): 793–795.

Marie H, Morishita W, Yu X, et al. (2005) Generation of silent synapses by acute *in vivo* expression of CaMKIV and CREB. *Neuron* 45: 741–752.

Nelson ED, Kavalali ET, and Monteggia LM (2006) MeCP2-dependent transcriptional repression regulates excitatory neurotransmission. *Current Biology* 16: 710–716.

West AE, Chen WG, Dalva MB, et al. (2001) Calcium regulation of neuronal gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11024–11031.

Zoghby HY (2003) Postnatal neurodevelopmental disorders: Meeting at the synapse? *Science* 302(5646): 826–830.

Gene Expression Regulation: Chromatin Modification in the CNS

J Hsieh, A Tsai, and K Ure, UT Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Epigenetic gene mechanisms are widely studied in the developmental and cancer biology fields but are relatively new to neurobiology. Nevertheless, many emerging discoveries suggest the importance of epigenetics in the central nervous system (CNS), which is the subject of this article. Historically studied in simple genetic model systems such as plants and invertebrates, and once considered in the realm of strange and unexplained gene expression phenomenon, it is now generally regarded as a fundamental mechanism of gene regulation comparable to genetic mechanisms. Although there are several definitions of epigenetics, the most commonly used one applies to meiotically and mitotically heritable changes in gene expression that is not encoded in the DNA sequence. Using biochemical modifications of chromatin, this allows genotypically identical cells to behave phenotypically different. Two well-documented epigenetic mechanisms are chromatin packaging and DNA methylation. From biochemical approaches and the use of simple model organisms, much of the basic cellular chromatin and transcriptional machinery has been worked out; however, a full understanding of the mechanisms that regulate gene expression in complex organs such as the mammalian brain is still in its infancy. This article discusses current ideas regarding epigenetic mechanisms of neural development and plasticity in both normal and disease states.

Chromatin Structure and Histone Modifications

Chromatin is a complex of nucleic acid and protein made up of nucleosome repeats of 147 bp of DNA wrapped around two copies each of histones H2A, H2B, H3, and H4. One of the most exciting breakthroughs in chromatin biology in the past decade is that the amino (N)-terminal tails of core histones are subject to many types of covalent modifications including acetylation and methylation that control the activation or repression of gene expression on the associated DNA (Figure 1). Some histone modifications, such as acetylation of histone H3 on lysine 9 and 14, dimethylated or trimethylated at lysine 4, and phosphorylated at serine 10, are associated with active gene expression. Other marks, such as dimethylated or trimethylated histone H3 on lysine 9, are associated with inactive chromatin. These histone modifications are regulated by specific enzymes; histone acetyltransferases (HATs), histone deacetylases (HDACs), and histone methyltranferases (HMTs) are just some of the chromatin-modifying enzymes that have emerged to play central roles in cell growth and differentiation.

Among the growing list of chromatin-modifying enzymes, the HATs and HDACs have been extensively studied. HATs induce histone acetylation which usually results in relaxation of the nucleosomes and increased transcription, whereas HDACs catalyze the reverse reaction; in the deacetylated state, nucleosomes are more condensed, preventing access of transcriptional activators to their target sites, which results in transcriptional repression. HDAC proteins fall into three classes. Class I and class II are based on their homology with yeast HDACs Rpd3p and Hda1p, respectively. Class III HDACs represent a distinct class and are similar to yeast Sir2. In terms of the brain, global changes in histone modifications have been observed in different regions, particularly after traumatic brain injury, electroconvulsive seizures, or drugs of abuse. Although these findings are intriguing, the role of cell type-specific histone-modifying enzymes in neural development and function remains to be determined. Even though one class of tissue-specific HDACs (class II) have well-documented roles in cardiac hypertrophy and other tissue types, despite the fact that they are highly enriched in the developing and adult brain, little is known regarding their specific function in neurobiology. One important clue regarding the role of class I and class II HDACs comes from small molecule HDAC inhibitors such as trichostatin A or valproic acid. HDAC inhibitors can induce neuronal differentiation in embryonic and adult hippocampal neural stem cells. Moreover, HDAC inhibitors are becoming one promising class of potential therapeutics with apparent neuroprotective effects in Huntington's disease, amyotrophic lateral sclerosis, Friedreich's ataxia, spinal muscular atrophy, ischemic brain injury, seizure disorders, and even brain cancer cells. The future application of HDAC inhibitors for treating CNS disorders remains an exciting possibility. Our own studies suggest that one of the most commonly prescribed antiepileptics, valproic acid, appears to block seizure-induced aberrant neurogenesis and hippocampal-related cognitive deficits through a mechanism involving HDAC inhibition. Although these global studies of histone modifications and HDAC inhibitors in neurological disorders and brain function remain intriguing, detailed studies are needed to understand how specific genes are regulated

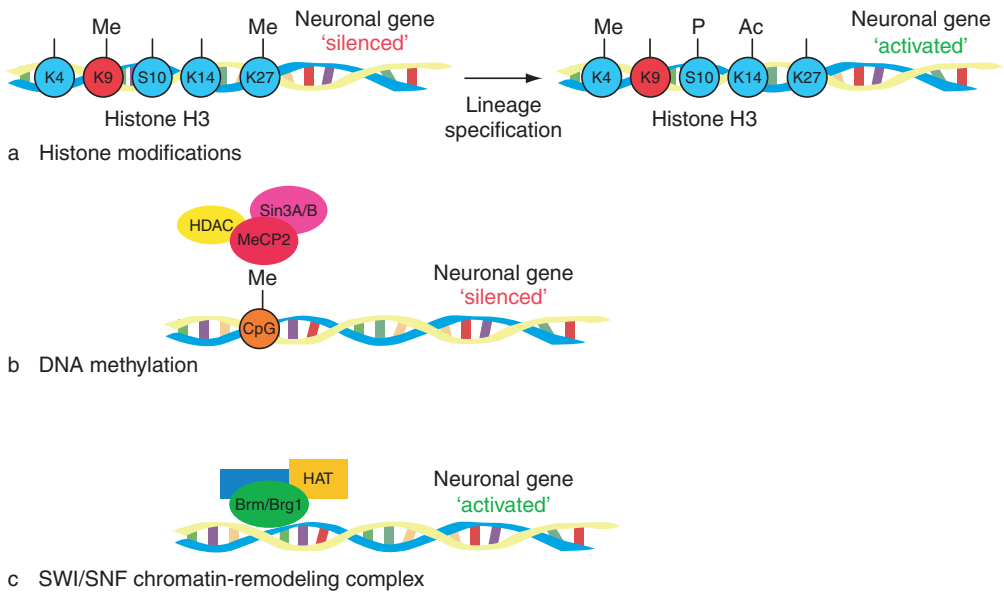

Figure 1 Epigenetic mechanisms involved in neural development. (a) Example of histone H3 acetylated (Ac), methylated (Me), and/or phosphorylated (P) in the promoter of a neuronal gene before ('silenced') and after lineage specification ('activated'). (b) Methyl-CpG-binding domain protein (MeCP2) binds to a methylated CpG site in DNA in the promoter of a neuronal gene and recruits repressor complexes containing HDACs and Sin3A/3B to mediate gene silencing. (c) An SWI/SNF chromatin-remodeling complex containing a Brm/Brg1 subunit and HAT activity to activate neuronal gene expression.

to control neural cell growth and plasticity. The mechanism for regulating HDAC activity in the cell can involve export from the nucleus and de-repression of gene expression or, more commonly, the recruitment of HDACs to gene-specific promoters by association with DNA-binding proteins, a process that is discussed in the next section.

Chromatin Complexes in the Brain

During cortical development, neural stem cells first undergo limited expansion through rounds of symmetric divisions and then undergo neurogenesis, mainly through asymmetric division. Toward the end of neurogenesis, cortical progenitors switch back to symmetric divisions and give rise to astrocytes and oligodendrocytes. Genes that specify one lineage must be repressed when a stem cell generates a different lineage. Thus, neuronal genes must be silenced until the onset of neurogenesis, and gliogenic genes must be inhibited during neurogenesis. This is coordinated through both transcriptional and epigenetic regulatory mechanisms that determine the potential of the cell to respond to extrinsic signals. Many elegant studies collectively describe the existence of transcriptional codes involving the spatial and temporal expression of cell type-specific transcriptional activators and repressors that control neuronal subtype specification and patterning in the developing CNS. Many of these factors have been studied in terms of their regulation of individual target genes during neural specification; however, a unifying mechanism that controls the process of neural cell fate specification through maturation has not been found.

A candidate master regulator of neurogenesis and gliogenesis is the transcriptional regulator NRSF (neuron-restrictive silencing factor, also called REST) (Figure 2). Initially described as a repressor of neuronal genes in nonneuronal cells, NRSF is a zinc finger protein that binds to a conserved 21–23 bp motif known as NRSE (neuron-restrictive silencing element, also called RE1). The NRSE sequence is found throughout the genome; many are found in the regulatory regions of neuron-specific genes, including neuronal growth factors, ion channels, neurotransmitter receptors, and guidance/migration molecules. The mechanism by which NRSF mediates transcriptional regulation of its target genes is becoming more complicated and appears to be context dependent. Emerging evidence suggests that NRSF functions by recruiting other corepressor complexes, usually through interaction of its N- or C-terminal domains, such as the mSin3A/B, N-CoR, CtBP, or CoREST complexes, which in turn may recruit class I HDACs. Nakagawa and colleagues showed that NRSF also interacts with class II HDACs in ventricular myocytes, although it remains to be determined whether the interaction is direct or indirect. Bioinformatic analyses

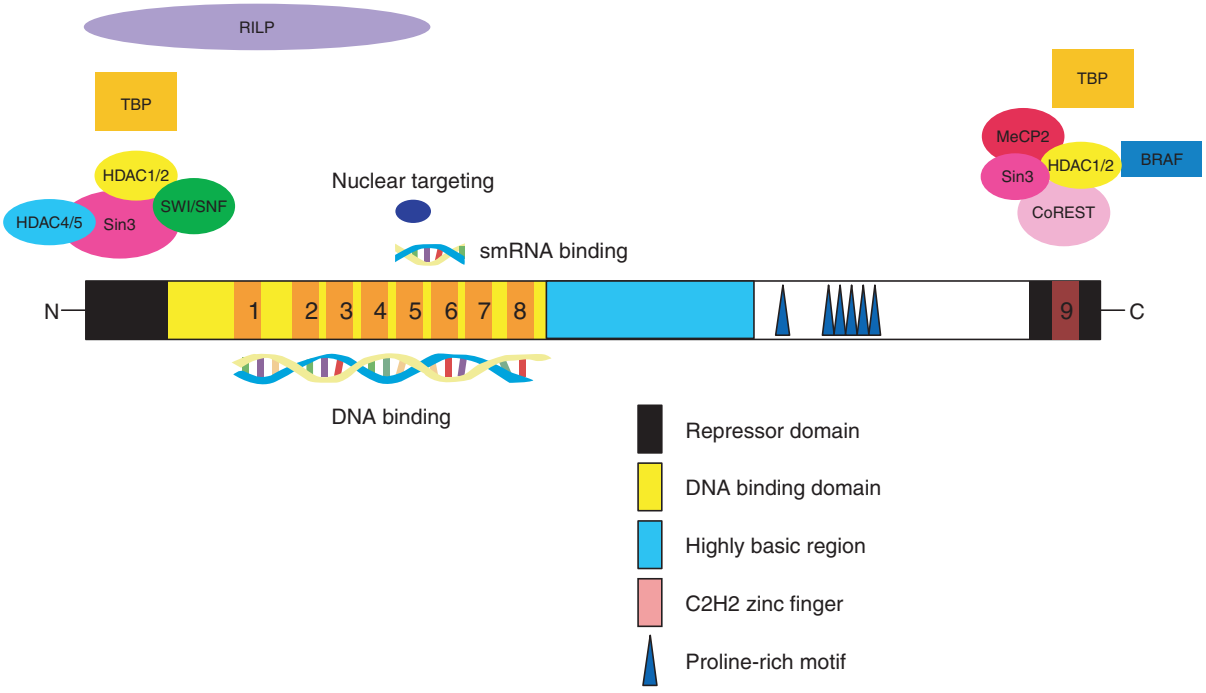

Figure 2 NRSF/REST is a context-dependent regulator of gene expression. Originally discovered as a repressor of neuronal gene expression in nonneuronal cells by both the Anderson (called NRSF) and Mandel (called REST) labs, cumulative data over many years have arrived at the consensus that the activity of NRSF/REST is dependent on cellular context and the nature of the specific target gene and its many possible interacting proteins. There is accumulating evidence to suggest that NRSF/REST splice variants, small modulatory RNAs, protein dimerization, cellular localization, and posttranslational modification may affect the function of the transcription factor, giving it the ability to spatially and temporally regulate its hundreds to thousands of putative target genes in both neuronal and nonneuronal cells. BRAF, B-raf; RILP, REST/NRSF-interacting LIM domain protein; SWI/SNF, switching/sucrose nonfermenting; TBP, TATA-binding protein.

have identified hundreds to thousands of NRSE sites throughout the mammalian genome, supporting the idea that NRSF functions as a master regulator of neuronal gene expression in neural stem/progenitor cells as well as in mature neurons. To add to the intrigue, NRSF function has also been implicated in various pathological states in brain as well as outside the brain, such as in different kinds of neural and neuroendocrine cancer cells, Huntington's disease, ischemic insults, epilepsy, and cardiovascular disease.

In addition to neuronal gene regulation by NRSF and recruitment of histone-modifying enzymes, evidence suggests that members of the switching/sucrose nonfermenting (SWI/SNF) and polycomb family of chromatin remodeling complexes have important roles in neural development and cancer and stem cell biology. SWI/SNF complexes use ATP hydrolysis to disrupt histone–DNA associations and contain different combinations of proteins that include both conserved and nonconserved components. They also interact with HATs or HDACs and/or sequence-specific transcription factors to activate or repress target genes. Typically, SWI/SNF complexes contain either ATPase Brahma (Brm) or Brahma-related protein 1 (Brg1). Chromodomain proteins, which possess the ATPase domain found in SWI/SNF proteins, are also recruited by polycomb repressors and are involved in homeotic gene silencing to control tissue patterning. Collectively, Brm, Brg1, and the polycomb repressor protein Bmi-1 have all been implicated in some aspect of neural development and function. Regulation of the sex-determining region Y (SRY)-related, high-mobility group (HMG)-box-containing protein (Sox-2), a key transcription factor in neural stem cells, is associated with the recruitment of Brm and appears linked to its role in stem cell maintenance. Brg1 is also required for *Xenopus* neurogenesis; loss of Brg1 correlated with increased neural progenitor proliferation and an expansion of Sox2-positive cells in embryos at later stages of neurogenesis. Brg1 function appears to be linked to the regulation of two proneural bHLH proteins, Neurogenin (Ngn)-related 1 and NeuroD. Moreover, the polycomb group proteins are emerging to play fundamental roles in stem cell maintenance and cancer. Bmi-1 is required for the self-renewal of stem cells in the CNS and peripheral nervous system; mice lacking Bmi-1 experience neurological defects, presumably due to a postnatal

depletion of their neural stem cells. Together, these data indicate the important role of chromatin remodeling factors in brain development and function.

The Role of DNA Methylation in the Brain

One classic epigenetic modification is DNA methylation, where mammalian DNA can be covalently modified through methylation of the carbon at the fifth position on the pyrimidine ring of the cytosine residue. This typically occurs at symmetrical CpG dinucleotides. DNA methylation is mainly involved in the establishment of parental-specific imprints during gametogenesis and gene silencing of the inactivated X chromosome and retrotransposons. *De novo* methylation of DNA is catalyzed by DNA methyltransferases, Dnmt3a and Dnmt3b, which add methyl groups onto unmethylated DNA. When cells divide, the methyltransferase Dnmt1 preferentially recognizes hemimethylated DNA and maintains DNA methylation. Interestingly, Dnmt3a and Dnmt3b have both been shown to associate with the regulatory regions of genes during the initial specification of neurons and glia, as well as during later stages of neuronal maturation and function. Dnmt1 is also very important in the brain and is involved in Janus kinase-single transducer and activator of transcription (JAK-STAT) signaling to control the timing of when precursor cells switch from neurogenesis to gliogenesis during development. Site-specific methylation of DNA has been shown to be important for astrocyte differentiation and activation of the glial fibrillary acidic protein (GFAP) promoter. Activation of the GFAP promoter requires binding of the STAT3 in response to leukemia inhibitory factor (LIF) signaling. Early cortical progenitors are refractory to astrocyte fates, even in the presence of LIF and STAT3 activation, likely due to methylation of the STAT3 binding site. At later stages, this site becomes demethylated and gliogenesis can proceed. A similar case was observed in the STAT3 binding site of the S100β promoter, a calcium-binding protein expressed in astrocytes.

In addition to DNA methylation, histone methylation has also been found to be important for the regulation of the GFAP gene in cortical progenitors. Using chromatin immunoprecipitation assays, Song and Ghosh showed there was increased histone H3 lysine 9 methylation, a mark associated with gene silencing, at the STAT3 binding site within GFAP at progenitor stages, when GFAP was repressed. Stimulation with basic fibroblast growth factor (FGF-2) potentiated the ability of CNTF (ciliary neurotrophic factor) to induce astrocyte differentiation. Strikingly, FGF-2 induction caused lysine 9 to become demethylated but lysine 4, the active mark, to become hypermethylated, resulting in STAT3 binding and activation of GFAP. By contrast, mature neurons had low levels of both lysine 9 and lysine 4 methylation. This leaves open the question of how GFAP is repressed in neurons. These studies highlight the diverse transcriptional and epigenetic mechanisms that control lineage-specific gene expression for the generation of different neural cell types.

Epigenetics and Neurodevelopmental Disorders

As described previously, DNA methylation-mediated silencing can be mediated by inhibition of transcription factor binding to target genes by methylation at CpG sites. Another mechanism by which DNA methylation leads to transcriptional gene silencing is the binding of methyl-CpGs by methyl-CpG-binding proteins (MBDs), which further recruit HDAC repressor complexes, resulting in a repression of chromatin. MBDs comprise a family of proteins, two of which (methyl-CpG-binding domain protein (MeCP2) and MBD1) play key roles in the nervous system. MeCP2 is found highly enriched in postmitotic neurons, and mutations in MeCP2 have been linked to the neurological disorder Rett syndrome. Rett syndrome is characterized by normal development until 1 year of age, most commonly seen in females, and is followed by a rapid deterioration, involving loss of speech and motor skills, microcephaly, seizures, autism, ataxia, and stereotypical hand wringing. The exact role of MeCp2 in neurons is unclear. There is evidence that MeCP2 is a key regulator of the essential CNS protein, brain-derived neurotrophic factor (BDNF). Interestingly, membrane depolarization of neurons by exposure to elevated levels of extracellular potassium resulted in the phosphorylation of MeCP2 and transcriptional de-repression of the BDNF promoter. Zhou and colleagues showed that phosphorylation of MeCP2 at a specific amino acid residue, S421, occurs selectively in the nervous system and is important for regulating dendritic growth and spine maturation, explaining, at least in part, the neural-specific pathology of Rett syndrome. Other MBD proteins may also play roles in the brain; MBD1 knockout mice have diminished hippocampal neurogenesis, defects in spatial learning, and a reduction in long-term potentiation in the dentate gyrus of the hippocampus.

Fragile X is another commonly inherited mental retardation. Patients have excessive expansion of the CGG sequence from the 5′ untranslated region (UTR) of the *Fmr1* gene, resulting in transcriptional silencing of the gene. When the CGG repeats exceed 200, the region is hypermethylated, preventing the binding of transcriptional machinery. Treatments with DNA demethylating drugs result in increased histone

H3 methylation at the lysine 9 residue, increased H3 and H4 acetylation, and decreased methylation at H3 lysine 4, resulting in the activation of *Fmr1*. This finding strongly suggests a role for chromatin remodeling in the silencing of *Fmr1*. Even more supportive is the observation that MeCP2 and Brm associate and jointly silence the *Fmr1* gene. RNA interference knockdown of MeCP2 and Brm resulted in increased mRNA transcription of *Fmr1*. Taken together, these results present a compelling case for epigenetic factors underlying neurological disorders such as Rett syndrome and fragile X mental retardation.

Disorders in genomic imprinting have also been shown to have effects on the integrity of the CNS. Genomic imprinting has been defined as an epigenetic effect on a parental chromosome in the gamete or zygote that causes differential gene expression in the offspring's somatic cells. The main epigenetic effect in genomic imprinting is differential allele-specific DNA methylation, whereas the effects of DNA methylation were described previously. Two well-described imprinting diseases related to the CNS are Prader–Willi syndrome (PWS), which is characterized by hyperphagia, obesity during childhood, mental retardation, and behavioral abnormalities; and Angelman syndrome (AS), which is characterized by mental retardation, speech impairments, and behavioral problems. Both molecular defects seem to lie in an ~2 Mb imprinted domain at 15q11–q13 which contains both maternally and paternally expressed genes in humans. In PWS, no single gene defect has been shown to be responsible for the disease, whereas in AS the loss of a single gene is the defect. The maternally expressed UBE3A (ubiquitin protein ligase E3A) is only imprinted in the brain and is involved in the ubiquitin–proteasome degradation pathway; loss of function of this gene results in AS. Therefore, another body of evidence suggests that epigenetic mechanisms play an essential role in the CNS.

Epigenetic Control of Neural Plasticity, Learning, and Memory

In addition to regulating gene expression in mature neuronal or glial cells or their progenitors during lineage specification, epigenetic mechanisms have also been suggested to underlie certain forms of cellular and organismal memory. Histone modifications and chromatin remodeling factors have been observed in animal models of drug addiction as well as certain types of learning and memory. Since drug addiction has a long-term sustained effect, evidence is leading to alterations of chromatin structure in the brain as a possible underlying mechanism to mediate long-lasting changes in gene expression associated with the addiction state. c-*fos* and ΔFosB, members of the Fos family of transcription factors, are implicated in behavioral responses and synaptic plasticity induced by abused drugs. Interestingly, cocaine administration can induce H4 acetylation and H3 phosphoacetylation at the c-*fos* gene promoter in the striatum after acute administration but is desensitized after repeated exposure to cocaine. Regarding the FosB promoter, H3 acetylation is still observed under chronic cocaine treatment. ΔFosB, a product of the *FosB* gene, also known to accumulate under chronic cocaine exposure, increases the binding of Brg1 (a component of the SWI/SNF complex mentioned previously) with cyclin-dependent kinase 5, a gene implicated in neurogenesis, synaptic plasticity, learning, and memory. This evidence suggests that ΔFosB can recruit chromatin-remodeling factors such as HATs and Brg1-containing complexes to its downstream targets, thus regulating the expression of key genes involved in sustaining the long-term effects of drugs of abuse.

Epigenetic modifications have also been linked to changes in neural plasticity and long-term memory. Two different mouse models for Rubinstein–Taybi syndrome (RTS), which in human patients has well-defined multiple congenital anomalies – mental retardation syndrome characterized by postnatal growth deficiency, microcephaly, specific facial characteristics, broad thumbs, and big toes – were used to demonstrate that histone acetylation (since the syndrome is linked to mutations in protein (CBP), which has HAT activity) is required for long-term potentiation, learning, and memory. CBP heterozygotes exhibited normal levels of short-term memory; however, their long-term memory, as determined in contextual and cued fear conditioning and novel object recognition tasks, was abnormal. Interestingly, when the HDAC inhibitor suberoylanilide hydroxamic acid was delivered intraventricularly into CBP heterozygotes, the mice showed improvement in contextual fear conditioning. In a second mouse model bearing a hippocampal CA1- and dentate gyrus-specific inducible form of mutant CBP lacking HAT activity, there were long-term memory deficits, although short-term memory appeared intact. Similarly, when an HDAC inhibitor trichostatin A was used, or when the transgene was shut off, the memory deficits could be rescued, suggesting that pharmacological manipulation of HAT activity might be a potential therapeutic approach to treat RTS. Perhaps more important, the findings that genes such as CBP and MeCP2, which have global effects on gene expression, have specific phenotypic consequences in the brain suggest that epigenetic mechanisms may represent an untapped area for understanding and treating cognitive disorders. Many more examples of

neurodevelopmental and cognitive disorders linked to transcriptional regulators have been described and discussed elsewhere.

Transcriptional Silencing in the Brain and Emerging Role of Small RNAs

Small noncoding regulatory RNAs have emerged to be a central player in all of biology as well as in the mammalian CNS. Their small size and sequence complementarity allow extreme versatility to target mRNAs for repression or activation of gene expression or modify chromatin structures of targeted genes. Double-stranded RNA was originally found to silence genes by targeting their mRNA intermediates for degradation, but during the past several years, more mechanisms of action of small RNAs have been discovered and potential new classes of such RNAs have been revealed. Small regulatory RNAs fall into at least five major classes. Small interfering (si)RNAs and micro (mi)RNAs both silence gene expression; usually, siRNAs target mRNAs for degradation by sequence complementarity to a single site, whereas miRNAs can recognize numerous partially complementary binding sites that act synergistically within target mRNAs found to be abundantly distributed in the postnatal brain, suggesting they play important roles in neuronal function. Out of the ~250 miRNAs in mammals, it appears that at least 125 miRNAs are enriched or expressed in the brain. Those enriched in brain include miR9,-29,-124,-125a,-125b,-127,-128,-132,-137,-138, and -139. A CREB-induced miRNA, miR-132, was found to regulate neuronal morphogenesis in cultured cortical neurons. Another miRNA implicated in neuronal function is miR-134, which is dendritically localized and involved in the control of spine size. Besides siRNAs and miRNAs, newly discovered classes of short RNAs are emerging that are species specific (tiny 20–22 nucleotide noncoding RNAs in *Caenorhabditis elegans*), tissue specific (small 29 or 30 nucleotide noncoding RNAs that interact with the germ cell protein PIWI in germline), and activators, not silencers, of gene expression (short, double-stranded RNAs called NRSE (described later)).

The first demonstrated intersection of small RNAs and chromatin remodeling factors in the mammalian CNS was when a small dsRNA matching the binding sequence of NRSF, called NRSE small modulatory RNA (or just smRNA), was discovered in adult hippocampal neural stem cells. This smRNA appeared to mediate its function in an entirely novel way, unlike siRNAs or miRNAs, by interacting with the NRSF protein and inducing neuron-specific gene expression. Clearly, there will be many more studies regarding brain-enriched small RNAs in both physiological states and disease states. Interestingly, the protein product of *Fmr1*, FMRP, has also been shown to interact with miRNAs and Argonaute 2, although the exact function of this association in the pathogenesis of fragile X mental retardation has not been fully described. However, much more research is required to fully elucidate the effects of these and other small RNAs in the CNS.

Conclusions

What is amazing and still one of the life's major mysteries is how the brain balances a specific set of instructions to carry out cellular and tissue patterning but still retains plasticity to respond to the ever-changing environment. The brain must have a way to receive input with the outside world and relay environmental information down to the cellular and molecular level to affect changes in gene expression. Epigenetic and chromatin modifications of target genes in response to variations in environmental conditions might serve as a major source of variation in gene expression and function, and ultimately as a mechanism mediating neuronal plasticity and individual differences in behavior. We propose that chromatin remodeling conveys dynamic environmental experiences on a static genome, resulting in heritable and reversible changes in the CNS during development and throughout adulthood.

See also: Alternative Splicing in the Nervous System; Gene Expression Regulation: Activity-Dependent; Gene Expression Dysregulation in CNS Pathophysiology; Genetic Influence on CNS Gene Expression: Impact on Behavior; Genomic Disorder and Gene Expression in the Developing CNS; Transcriptional Silencing.

Further Reading

Alarcon JM, Malleret G, Touzani K, et al. (2004) Chromatin acetylation, memory, and LTP are impaired in CBP+/−mice: A model for the cognitive deficit in Rubinstein–Taybi syndrome and its amelioration. *Neuron* 42(6): 947–959.

Cao X, Yeo G, Muotri AR, et al. (2006) Noncoding RNAs in the mammalian central nervous system. *Annual Review of Neuroscience* 29: 77–103.

Chen WG, Chang Q, Lin Y, et al. (2003) Derepression of BDNF transcription involves calcium-dependent phosphorylation of Mecp2. *Science* 302(5646): 885–889.

Hong EJ, West AE, and Greenberg ME (2005) Transcriptional control of cognitive development. *Current Opinion in Neurobiology* 15(1): 21–28.

Hsieh J and Gage FH (2004) Epigenetic control of neural stem cell fate. *Current Opinion in Genetics & Development* 14(5): 461–469.

Hsieh J and Gage FH (2005) Chromatin remodeling in neural development and plasticity. *Current Opinion in Cell Biology* 17(6): 664–671.

Hsieh J, Nakashima K, Kuwabara T, et al. (2004) Histone deacetylase inhibition-mediated neuronal differentiation of multipotent adult neural progenitor cells. *Proceedings of the National Academy of Sciences of the United States of America* 101(47): 16659–16664.

Jenuwein T and Allis CD (2001) Translating the histone code. *Science* 293(5532): 1074–1080.

Korzus E, Rosenfeld MG, and Mayford M (2004) CBP histone acetyltransferase activity is a critical component of memory consolidation. *Neuron* 42(6): 961–972.

Kumar A, Choi KH, Renthal W, et al. (2005) Chromatin remodeling is a key mechanism underlying cocaine-induced plasticity in striatum. *Neuron* 48(2): 303–314.

Kuwabara T, Hsieh J, Nakashima K, et al. (2004) A small modulatory dsRNA specifies the fate of adult neural stem cells. *Cell* 116 (6): 779–793.

Martinowich K, Hattori D, Wu H, et al. (2003) DNA methylation-related chromatin remodeling in activity-dependent *Bdnf* gene regulation. *Science* 302(5646): 890–893.

Novina CD and Sharp PA (2004) The RNAi revolution. *Nature* 430 (6996): 161–164.

Petrij F, Giles RH, Dauwerse HG, et al. (1995) Rubinstein–Taybi syndrome caused by mutations in the transcriptional coactivator CBP. *Nature* 376(6538): 348–351.

Schratt GM, Tuebing F, Nigh EA, et al. (2006) A brain-specific microRNA regulates dendritic spine development. *Nature* 439 (7074): 283–289.

Takizawa T, Nakashima K, Namihira M, et al. (2001) DNA methylation is a critical cell-intrinsic determinant of astrocyte differentiation in the fetal brain. *Developmental Cell* 1(6): 749–758.

Zhao X, Ueba T, Christie BR, et al. (2003) Mice lacking methyl-Cpg binding protein 1 have deficits in adult neurogenesis and hippocampal function. *Proceedings of the National Academy of Sciences of the United States of America* 100(11): 6777–6782.

Zoghbi HY (2003) Postnatal neurodevelopmental disorders: Meeting at the synapse? *Science* 302(5646): 826–830.

Gene Expression Regulation: Steroid Hormone Effects

M Warner, S Andersson, and J-A Gustafsson, Karolinska Institutet, Stockholm, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The most extensive information regarding the distribution of estrogen receptor (ER) α and ERβ messenger RNA (mRNA) expression patterns in the central nervous system (CNS) has come from studies of the rat. In adult rats, cells expressing ERβ mRNA are more widely distributed throughout the brain than are those expressing ERα mRNA, although ERα-positive neurons are more numerous than those expressing ERβ. Within the hypothalamus, ERα is the primary ER in the arcuate and ventromedial nuclei, whereas in the paraventricular nucleus, ERβ is the predominant ER. Within the midbrain, pons, and brain stem, a predominance of ERβ has been observed in several of the major monoaminergic cell populations. These include the dopaminergic substantia nigra and ventral tegmental area, the serotonergic raphe nuclei, and the noradrenergic locus coeruleus. In the cerebellum, only ERβ mRNA was exclusively detected, both in the Purkinje cell layer and scattered within granule cells. Despite many differences in the distribution of ERα and ERβ mRNAs, considerable overlap was seen in the preoptic area and the bed nucleus of stria terminalis and throughout the lower brain stem. Studies also suggest considerable overlap in expression of ERα and ERβ mRNA in various regions of the forebrain, such as the cerebral cortex and hippocampus.

Comparison of ER mRNA expression patterns in rat, mouse, monkey, and human indicates that although the general distribution is similar, data cannot be extrapolated from one species to another. For example, in the mouse brain, ERβ mRNA is more abundant in the suprachiasmatic nucleus, raphe nuclei, and entorhinal cortex, whereas in these areas the ERβ mRNA expression is low in the rat brain. In contrast, levels of ERβ mRNA in the mouse supraoptic nucleus and preoptic area are low. The rat paraventricular nucleus has been reported to express only ERβ, but the mouse paraventricular nucleus contains cells that also express moderate levels of ERα. It is interesting that the human paraventricular nucleus has been reported to express mRNA for both ERα and ERβ; in fact, ERα appears to be the predominant subtype.

The distribution of ERα and ERβ protein in the CNS seems to coordinate relatively well with the mRNA distribution. However, some exceptions have been reported: In the suprachiasmatic nucleus of the mouse brain, only weak ERβ immunoreactivity has been detected, and in the nuclei of Purkinje cells and granular cells of the mouse cerebellum, ERβ was not detectable even though these regions are reported to abundantly express ERβ mRNA.

The differential distribution of ERs in the brain may provide a mechanism for selective modulation of physiological parameters involved with reproduction, as well as the regulation of learning and memory. ER expression in discrete brain regions depends on gender, reproductive status, age, and exposure to estradiol (E2). ERα expression is acutely reduced shortly after exposure to E2, whereas ERβ expression decreases after chronic E2 treatment (1–2 weeks) but not after short-term (2 h) E2 treatment of ovariectomized rats. A recent study indicates altered expression of ERβ in specific brain regions in the male ERα knockout ($ER\alpha^{-/-}$) mouse.

Estrogen Action in the CNS

Estrogens influence a wide variety of brain functions in mammals. In the adult CNS, these include the feedback regulation of hypothalamic and pituitary hormone secretion and the facilitation of feminine reproductive behavior via regulation of brain progesterone receptor levels. Furthermore, estrogen has an effect on the modulation of motor behavior. There is also an effect on mood and mental state by the control of central aminergic and peptidergic neurotransmission. Various experimental studies and clinical observations have shown estrogen to be effective in the preservation of cognitive brain function, suggesting that it could exert a neuroprotective effect in some neurological disorders such as Alzheimer's disease, Parkinson's disease, schizophrenia, and ischemic stroke. In schizophrenia, psychotic symptoms are exacerbated when estrogen levels are low, and estrogen treatment produces beneficial effects. Furthermore, abrupt changes in hormone levels such as those occurring at menopause or at parturition are associated with depression.

Multiple epidemiological reports have suggested that estrogen replacement therapy may delay the onset of and slow the cognitive decline associated with Alzheimer's disease. Accumulation of amyloid β is thought to be causal in the development of cerebral plaques, a characteristic in the neuropathology of Alzheimer's disease. A number of studies have reported that E2 reduces the generation of amyloid β peptides and protects against amyloid β-induced

cell degeneration in primary cell cultures. E2 also attenuates impairment of synaptic Na^+, K^+ adenosine triphosphatase activity and glucose and glutamate transport induced by amyloid β.

Whether estrogen replacement therapy can improve cognition and memory functions in neurological diseases remains an actively debated issue. It seems that there is a critical period around menopause when hormone replacement delays the onset of and slows cognitive decline. However, estrogen replacement at a later stage does not seem to improve cognitive function. It has been suggested that estrogen replacement may protect younger women from Alzheimer's disease or reduce the risk of early onset forms of Alzheimer's disease.

Estrogen as a Trophic Factor for Neurons

More than 20 years ago, it was found that estrogen enhances the growth and arborization of axons and dendrites in hypothalamic neurons grown in organotypic cultures. Initial studies concerning synaptic remodeling in developing and adult brain were performed on the arcuate nucleus, where neonatal estrogen treatment increases synapse formation during postnatal development in rats. Estrogen treatment was also shown to restore deafferentation-induced loss of axodendritic synapses in adult rats. Later on it was shown that synapses in the arcuate nucleus exhibit a phasic remodeling, which could be linked to the fluctuation in hormone levels during the estrous cycle. Growth-promoting effects of estrogen have now also been described in the midbrain, cortex, hippocampus, spinal cord, and pituitary.

One of the processes regulated by ovarian hormones is the cyclic formation and breakdown of excitatory synapses in the hippocampus. Estrogens do not act alone; in fact, ongoing excitatory neurotransmission is required for synapse induction, and activation of *N*-methyl-D-aspartate (NMDA) receptors by glutamate is an essential factor in the development of new excitatory synapses. Estrogen treatment has been shown to increase the density of NMDA receptors in the CA1 region of the hippocampus. Exactly how estrogen induces the formation of new synaptic contacts is not clear, but estrogen treatment has been shown to induce both pre- and postsynaptic markers in the spine synapses. In light of the critical role of dendritic spine synapses in hippocampal function, it is not surprising that it has been suggested that Alzheimer's disease, schizophrenia, and stress are all associated with a substantial pathological loss of dendritic spine synapses in the hippocampus.

It has been suggested that the effect of estrogens on neuronal growth may be mediated indirectly via regulation of other endogenous neurotrophic substances. Expression of several growth factors and their receptors has been shown to be estrogen regulated. The most studied are nerve growth factor (NGF) and its receptors in cholinergic neurons transforming growth factor in the hypothalamus, brain-derived neurotrophic factor in the cortex, NGF receptors in sensory neurons, and the insulin-like growth factor and its receptor in the hypothalamus. The effects of estrogen on growth factor signaling are reciprocated, and growth factors can influence estrogen signaling. Thus, NGF has been shown to regulate ER action in the forebrain, possibly via phosphorylation events. However, estrogen can also affect neurite elongation and differentiation directly through its regulation of cytoskeletal genes that are required for neurite growth.

Estrogen and the Sexual Differentiation of the CNS

The criteria that have been used to measure sexual differentiation in the brain are volumes of brain nuclei, cell numbers/densities, and synaptic morphology. The first morphological evidence of sexual dimorphism in the brain was the observation that three vocal control areas of songbirds are larger in males than in females. Size differences have since been found in the preoptic region in rat and human brains. Sex differences have also been reported in higher cognitive centers and in sensory and autonomic ganglia as well as structures of the limbic system.

Virtually all morphological differences between male and female brains and all sexually differentiated behaviors are thought to be influenced by embryonic exposure to gonadal steroids, a phenomenon called imprinting. The embryonic testes secrete testosterone that enters the brain and is converted to its metabolites 5α-dihydrotestosterone (DHT) and E2. There is a large body of evidence showing that E2 is the hormone which masculinizes the brain permanently. Although it has been the focus of much research effort, even today, little is known about how E2 exposure imprints the fetal brain. Besides acting as a growth-stimulating agent, estrogen has been suggested to influence neurogenesis and apoptotic cell death. There is still no detailed study of the profile of ERα and ERβ in the fetal brain, so it is not possible to pinpoint the cells which are the targets of E2 during this period.

Estrogen and the Cholinergic System

The basal forebrain contains cholinergic neurons that project to cerebral cortex and hippocampus, where they play an important role in cognitive function.

Experiments with ovariectomy and estrogen replacement therapy have revealed an induction of choline acetyltransferase (ChAT), the rate-limiting enzyme for acetylcholine formation, in basal forebrain of female rats. One possible candidate as a regulator of the cholinergic system of the basal forebrain is NGF, which is produced by the hippocampus and transported retrogradely to basal forebrain neurons to produce trophic effects. Although the effects of estrogen treatment on NGF levels in hippocampus remain to be investigated, ER has been reported to co-localize with low-affinity NGF receptors in cholinergic neurons of the basal forebrain of the newborn rat. Moreover, estrogen replacement in both young and aged female rats increases both NGF receptor mRNA and ChAT mRNA expression in basal forebrain.

Estrogen and the Serotonergic System

Serotonin neurons of the midbrain–brain stem raphe nuclei are among the earliest neuronal phenotypes to become differentiated during CNS development, and serotonin is believed to act as a regulatory-developmental agent. The serotonergic system is involved in the regulation of diverse functions such as reproduction, mood, sleep, and cognition. Serotonergic activity seems to be regulated by the ovarian steroids, but the mechanism by which such regulation occurs is not fully understood. It has been demonstrated that ovarian hormones increase the expression of tryptophan hydroxylase, the key enzyme in serotonin biosynthesis, and suppress expression of the serotonin transporter in the macaque raphe nuclei. Estrogen treatment has also been reported to affect the expression of certain subtypes (5HT2A, 2C, and 1A) of serotonin receptors in rat brain.

Estrogen's affecting serotonergic function is relevant to the effect of estrogen on mood and cognition. High doses of estrogen have been reported to have antidepressant effects in humans, and estrogen treatment influences the response to antidepressant drugs in animal models and in clinical studies.

Estrogen and Catecholaminergic Systems

Noradrenergic System

Brain stem catecholaminergic neurons contain ERs, and estrogen treatment after gonadectomy exerts complex, time-dependent effects on the level of tyrosine hydroxylase mRNA. In the rat locus coeruleus, galanin is co-expressed in many noradrenergic neurons, and estrogen treatment increases the expression of galanin mRNA. It is thought that estrogen treatment might reduce noradrenergic tone by enhancing the secretion of galanin, which in turn reduces noradrenaline release.

Dopaminergic System

Estrogen has been shown to modulate the binding of dopamine to its receptor. Estrogen also potentiates K^+-stimulated dopamine release from rat nucleus accumbens, and estrogen pretreatment increases the firing rate of neostriatal neurons in response to dopamine. Furthermore, estrogen has been shown to facilitate amphetamine- and apomorphine-stimulated dopamine release and locomotor activity in lesioned rats, and this activity is responsive to natural fluctuations in E2 levels and is generally increased during late proestrus and early estrus. Intrastriatal application of E2 rapidly causes rotational behaviors and enhances sensorimotor performance. One explanation of this phenomenon is reduced L-type Ca^{2+}-channel activity in striatal neurons via a G-protein-coupled receptor. Relatively high levels of estrogen, such as provided through oral contraceptives and estrogen replacement therapy, have been shown to exacerbate symptoms of Parkinson's disease, pointing to antidopaminergic actions that are the opposite of the actions of physiological levels of E2.

An Alternative Estrogen

More than 30 years ago, it was noted that 5α-androstane-3β,17β-diol (3βAdiol) is an estrogenic metabolite of DHT which has no affinity for the androgen receptor. The formation of 3βAdiol is catalyzed by several steroid metabolizing enzymes – 17β-hydroxysteroid dehydrogenase and 3α- and 3β-hydroxysteroid dehydrogenase – and like DHT, 3βAdiol is a local (autocrine/paracrine) hormone in the cells where it is synthesized.

The immature ovaries in prepubertal female rats do not synthesize E2, but substantial quantities of 3βAdiol and 5α-androstane-3α,17β-diol (3αAdiol) seem to be synthesized instead, and it has been suggested that these steroids may have a role in the onset of puberty. The immature testes also produce 3βAdiol and 3αAdiol in large amounts. It has been reported that until postnatal day 20, the main metabolite of DHT is 3αAdiol; thereafter 3βAdiol is the predominant testicular metabolite of DHT. In adult male rats, the plasma concentration of 3βAdiol is about 20 times higher than the E2 level. Treatment with 3βAdiol in male rats does not seem to affect the serum levels of gonadotropins, although both DHT and 3αAdiol depress the gonadotropin levels in serum. A recent report has stated that 3βAdiol treatment, as well as DHT treatment, depresses the corticosterone response to stress.

The enzymes responsible for the conversion of DHT to 3βAdiol are expressed in many regions of the rat brain throughout development, including the cortex, hypothalamus, thalamus, septum, hippocampus, and cerebellum. In the brain, the exact function of 3βAdiol is unclear, but in a neuronal cell type, HT-22, 3βAdiol has been shown to activate ERβ-induced transcription mediated by an estrogen response element with the same efficacy as E2. Also, the transcriptional selectivity for ERβ over ERα has been demonstrated to be much greater than its binding selectivity in these neuronal cells.

Moreover, 3βAdiol is further metabolized via hydroxylation to nonestrogenic androstanetriols, which are excreted in the urine. Oxidative conversions and metabolism of cholesterol and its steroid derivatives are performed by cytochrome P450s (CYPs), a family of heme-containing mono-oxygenases. Tissue levels and estrogenicity of 3βAdiol are regulated by CYP7B1, which was cloned from rat and mouse hippocampus in 1995. Before the complementary DNA of this enzyme was cloned, the tissue distribution of '3βAdiol hydroxylase' was assessed by its catalytic activity. The tissues with the highest activity were the ovary and testis, as well as the prostate, brain, and pituitary. With the cloning of CYP7B1, it was identified as a steroid/sterol B-ring hydroxylase with selectivity for hydroxylation at the 7α, 6α, and 7β positions. It accepts dehydroepiandrosterone (DHEA), pregnenolone, 25- and 27-hydroxycholesterol, and 3βAdiol, but 3αAdiol is not a substrate, and neither is E2 or testosterone.

Analysis of mRNA from various tissues shows CYP7B1expression in several tissues, such as brain, prostate, liver, and kidney. It is interesting that CYP7B1 mRNA expression and enzyme activity have been reported to be very high in the neuroepithelium of most brain areas during early to midgestation but to decrease in later stages of gestation, and in adults it has been detected only in a few specific regions of the brain, most notably the dentate gyrus. No expression of CYP7B1 in liver was reported at any point in gestation. In liver, CYP7B1 activity has initially been detected at 3 weeks of age, after which it increases gradually, then remains constant after 8 weeks of age. Sterol B-ring hydroxylation in the liver has been shown to be important for the metabolic elimination of oxysterol via conversion to bile acids. This is promoted by cholesterol 7α-hydroxylase (CYP7A1) and CYP7B1 operating in parallel with a liver-specific 24(S)-hydroxycholesterol 7α-hydroxylase. It has been suggested that in the brain, CYP7B1 activity may be important for neurosteroid synthesis and metabolism. Baulieu and colleagues originally used the term 'neurosteroid' to describe the steroids that were produced *de novo* within the brain. Nowadays the term is also used for steroids that are synthesized in the CNS from precursor steroid hormones secreted from peripheral endocrine organs. The understanding of neurosteroid synthesis and where neurosteroids act is still limited. It is interesting that a reduced level of CYP7B1 mRNA has been reported in the dentate gyrus and the CA1 hippocampal area of patients with Alzheimer's disease.

Knockout Models

In 1998, ERβ knockout (ER$\beta^{-/-}$) mice were produced, and the phenotype of ER$\beta^{-/-}$ mice is at present being investigated and compared with that of ER$\alpha^{-/-}$ mice. A part of the phenotypes present in the two knockouts relates to reproductive function and to other tissues, including the CNS, the skeletal system, and the cardiovascular system.

Phenotype of ER$\beta^{-/-}$ Mice

In contrast to ER$\alpha^{-/-}$ mice, the lack of functional ERβ does not impair adult sexual behavior. However, transient behavioral differences between wild-type (WT) and ER$\beta^{-/-}$ males during puberty have been described. At the time of puberty, but not in adult males, aggression is higher in ER$\beta^{-/-}$ than in WT males. Furthermore, ERβ has been shown to be involved in defeminization or the ability of males to display female-typical behaviors, such as lordosis (receptive mating posture).

Increased anxiety has been reported in ER$\beta^{-/-}$ females compared with WT mouse behavior. Local increase of the serotonin receptor 5-HT1A expression in medial amygdala of the brain has been suggested to contribute to these changes as well as to lower serotonin content in several brain regions including the bed nucleus of the stria terminalis, preoptic area, and hippocampus.

In the brain, estrogen has been suggested to influence spatial learning in females via ERβ. The results are puzzling, however: WT females were able to learn a spatial task with or without E2 replacement, but when given E2 replacement, ER$\beta^{-/-}$ females showed an impaired spatial learning ability in the Morris water maze. Without E2 treatment, the spatial learning was normal in ER$\beta^{-/-}$ females compared with WT females. Furthermore, ERβ has been shown to be involved in the estrogenic regulation of oxytocin and vasopressin mRNA levels in the paraventricular nucleus.

Phenotype of CYP7B1$^{-/-}$ Mice

CYP7B1 knockout (CYP7B1$^{-/-}$) mice have been generated in two independent laboratories, and both

lines show that CYP7B1 is not essential for viability of mice. CYP7B1$^{-/-}$ mice have been reported to accumulate 25- and 27-hydroxycholesterol in their plasma and tissues, while the levels of brain-derived 24(S)-hydroxycholesterol are normal. In addition to the abnormalities in certain plasma sterol levels, brain extracts from CYP7B1$^{-/-}$ mice fail to catalyze hydroxylation of DHEA and 3βAdiol.

In prostate, CYP7B1 has been suggested to play a role in epithelial cell proliferation by regulating the tissue level of 3βAdiol, which could be the quantitatively major estrogenic steroid and the most likely ligand for ERβ in this tissue. The ventral prostate, but not the other lobes of prostate, has a decreased number of proliferating cells in CYP7B1$^{-/-}$ mice before puberty and are smaller than those of their WT littermates at adulthood. Because CYP7B1 represents the major pathway for inactivating 3βAdiol in the prostate, it regulates, via ERβ, androgen receptor expression and epithelial cell proliferation.

It has been reported that the uteri and mammary glands of young CYP7B1$^{-/-}$ female mice have characteristics of tissues chronically exposed to estrogen. There is a precocious onset of puberty in female CYP7B1$^{-/-}$ mice, but at 15 months of age, the estrous cycling has stopped. These data suggest premature exhaustion of ovarian function in CYP7B1$^{-/-}$ females.

Why Study the Function of ERβ in the CNS?

Evidence suggests multiple roles for estrogen in the brain. It influences development, plasticity, and survival of neurons, with consequences for age-related neuronal degenerative diseases. Even though we have an anatomical framework of the distribution of the two ER subtypes in the CNS, the functional relevance of the two ERs in different cells remains a question. Distinguishing the effects of the different ER subtypes *in vivo* would give insight into the kind of ER-selective ligands that would exert positive effects on brain and bone but spare the uterus and breast. The estrogen replacement therapy of today is associated with serious unwanted side effects, including development of breast cancer and endometrial cancer. ERβ is highly expressed in the brain but very poorly in the adult uterus, and it is possible that ERβ-selective ligands would have the desired tissue specificity.

ERβ is important for neuronal survival, during both brain development and the following neuronal life span. In the brains of ER$\beta^{-/-}$ mice, a regional neuronal hypocellularity with a severe neuronal deficit is seen in the somatosensory cortex, especially layers II, III, IV, and V. Neuronal loss was also evident in certain regions of the hypothalamus, as well as in the amygdala and ventral tegmental area, but not in the cerebellum or hippocampus. Proliferation of astroglia was also detected in a regional manner; it was evident in the limbic system but not in the cortex. Regarding the hyperproliferation of astroglia in the limbic system, it is possible that in ER$\beta^{-/-}$ mice, it occurs in response to the neuronal loss as proliferation of astroglia usually is a reactive response to neurodegeneration or injury.

These morphological abnormalities in the brain of ER$\beta^{-/-}$ mice are evident as early as 2 months of age. As ER$\beta^{-/-}$ mice age, the neuronal deficit becomes more pronounced, and by 2 years of age, there is degeneration of neuronal cell bodies throughout the brain. At this stage, there is a gross atrophy of the brain, especially pronounced in the cortex. Neuronal shrinkage is particularly evident in the substantia nigra. However, no astrogliosis is evident in 2-year-old ER$\beta^{-/-}$ mice. One possible explanation is that ER$\beta^{-/-}$ mice lose both neurons and astrocytes as they age. It appears that ERβ is one critical factor for neuronal life span in CNS. No morphological changes have been reported in the brains of ER$\alpha^{-/-}$ mice, but some *in vitro* studies have suggested that ERα has a neuroprotective effect, whereas ERβ mediates apoptosis in neuronal cells. Further studies are needed regarding the actual role of the different ERs in the neuroprotective action of estrogen, but ERβ may well be a protective factor for neurons and could influence the development of degenerative diseases of the CNS, such as Alzheimer's disease and Parkinson's disease, as well as those resulting from trauma or stroke in the brain.

Because the morphological changes appear quite early in ER$\beta^{-/-}$ mouse brains, it is possible that the phenotype has resulted from defective embryonic brain development. During embryonic development, brain size is determined by proliferation of neural progenitor cells and by differentiation, survival, and migration of postmitotic neurons. A reduction in the number of neurons in the cortex could therefore occur via several possible mechanisms: decreased proliferation, increased apoptotic cell death, or abnormal migration and differentiation. In embryonic brains from ER$\beta^{-/-}$ mice, there are no differences in size or cellularity at embryonic day (E) 14.5; however, at E18.5, ER$\beta^{-/-}$ mouse brains are smaller than those of their WT littermates. Nissl staining of brain sections shows a thinner cortex, but no morphological differences are evident in the hippocampus, thalamus, or hypothalamus. Because the neuronal loss in the adult ER$\beta^{-/-}$ mouse cortex is more pronounced in superficial layers than in deeper layers, it may be that neurons in the deep layer would be affected little by

the lack of ERβ. When proliferating cells are labeled with 5′-bromodeoxyuridine (BrdUrd) on E12.5, when cortical neurogenesis in mice begins, and examined on E14.5, there is no difference between WT and ER$\beta^{-/-}$ mice in the number of labeled cells in the cortex. However, when BrdUrd is administered between E14.5 and E16.5, fewer labeled cells can be detected in the superficial cortical layers by E18.5 and postnatal day 14 in mice lacking ERβ, suggesting that neuronal migration could be defective. When one stains for radial glia, which are essential for guiding the migrating neurons, their processes appear to be truncated. The cytoarchitecture of ER$\beta^{-/-}$ mouse cortex also seems disorganized, and cells are not clustered in a clear radial pattern, as seen in WT littermates. The radial glia also have more branched end-feet at the marginal zone, and this might result in abnormal contact with Cajal–Retzius cells. These cells produce reelin, shown to be important for migration and positioning of neurons in the cerebral cortex. Although the number and morphology of Cajal–Retzius cells in ER$\beta^{-/-}$ mice appear normal, their functional abilities in the absence of ERβ need further investigation.

At E18.5, more apoptotic cells are in the ventricular zones in ER$\beta^{-/-}$ mice, indicating that the lack of ERβ promotes apoptosis. Regarding the regional astrogliosis of adult ER$\beta^{-/-}$ mice, few astroglia cells have been seen in either ER$\beta^{-/-}$ or WT mouse brains during the late gestational stage. Thus, a lack of ERβ does not hasten the maturation of glial cells into astroglia, and the increase in the number of astroglia in the adult ER$\beta^{-/-}$ mouse brain seems to occur postnatally. It has been reported that ERβ mRNA is detected from E10.5 in brain of mouse embryos, whereas the earliest detection of ERα mRNA is on E16.5. Although the exact distribution pattern of ERβ in the CNS during development is still unclear, data suggest that by influencing migration and neuronal survival, ERβ has an important role in the late embryonic development of the brain. However, it is unclear whether ERβ is ligand activated in the embryonic brain because of the limited exposure of the brain at this stage to E2. The placenta efficiently inactivates E2 by catalyzing its conjugation with sulphate. In addition, α-fetoprotein in the fetal circulation binds E2 with high affinity and prevents its entry into the tissues. One alternative ligand for ERβ could be 3βAdiol, an estrogenic metabolite of DHT. Although 3βAdiol activates both ERα and ERβ, it has no affinity for the androgen receptor. Recent studies report that the human placenta is capable of synthesis and metabolism of androgen during early gestation and is thereby able to synthesize 3βAdiol. To explore the function of 3βAdiol *in vivo*, CYP7B1$^{-/-}$ mice, in which the pathway for degradation of 3βAdiol is inactivated, have been examined, and their early postnatal brain is larger than that of WT littermates of the same body weight. Nissl staining of brain sections has shown all regions of the forebrain to be affected. The increased sizes of the cortex, caudate putamen, and preoptic area are all directly proportional to the increased overall brain size in CYP7B1$^{-/-}$ mice at 14 days of age. There is no change in cell density of either neurons or astroglia, and no apparent distortion in cytoarchitecture or cortical layering has been detected. However, the increased brain mass results in compression of the lateral ventricles in CYP7B1$^{-/-}$ mice.

When BrdUrd was administered on E12.5 and embryos examined on E13.5, no difference was detected between WT and CYP7B1$^{-/-}$ littermates in the number or pattern of S-phase-labeled neurons in the brain. The pattern of late-born neurons, generated and labeled with BrdUrd at E17.5 and observed at E18.5, was also normal. No difference in the number of M-phase-labeled progenitor cells lining the ventricles has been seen, and no abnormalities in neuronal guidance by radial glia or neuronal differentiation have been detected. However, a profound decrease in apoptosis is apparent in the ventricular/subventricular zone between E13.5 and E15.5 in CYP7B1$^{-/-}$ mice. At later gestational stage, no changes in cell death have been detected, indicating that the increase in brain size in CYP7B1$^{-/-}$ mice is most probably due to increased cell survival during early brain development. However, by the time mice were 3 months old, the brains of CYP7B1$^{-/-}$ mice and their WT littermates were indistinguishable on the basis of size and shape. Thus at some time in the late postnatal period, the excess cells must be removed. Competition for the limited access to target-derived neurotrophic substances may be involved in this later neuronal loss and needs to be investigated further.

These data are supportive evidence for a role for the ERβ-3βAdiol-CYP7B1 pathway in regulation of neuronal survival in the embryonic brain. By metabolizing steroids in the CNS, CYP7B1 could be important for the availability of steroids to their receptors, either by preventing or by potentiating receptor interactions. If 3βAdiol acts as an estrogen in the fetal brain, ERs are active earlier in fetal life than has previously been thought. It is now essential to have a thorough understanding of the expression and cellular localization of ERs in the developing brain.

Summary and Conclusions

A central finding is that ERβ is important for neuronal survival, during both brain development and the

following neuronal life span. Data of the phenotype in the brains of ER$\beta^{-/-}$ mice have strengthened the view that ERβ is a protective factor for neurons and could influence the development of degenerative diseases of the CNS, such as Alzheimer's disease and Parkinson's disease, as well as those resulting from trauma or stroke in the brain. The data suggest that by influencing migration and neuronal survival, ERβ has an important role in the late embryonic development of the brain. However, it is unclear, because of the limited exposure of the brain at this stage to E2, whether ERβ is ligand activated in embryonic brain. One alternative ligand for ERβ could be 3βAdiol, an estrogenic metabolite of DHT. The data of the phenotype in the brains of CYP7B1$^{-/-}$ mice are supportive evidence for a role for the ERβ-3βAdiol-CYP7B1 pathway in regulation of neuronal survival in the embryonic brain. By metabolizing steroids in the CNS, CYP7B1 could be important for the availability of steroids to their receptors, either by preventing or by potentiating receptor interactions. If 3βAdiol acts as an estrogen in the fetal brain, then ERs are active earlier in fetal life than has been thought previously. It is now essential to have a thorough understanding of the expression and cellular localization of ERs in the developing brain.

See also: Emotional Hormones and Memory Modulation; Hormones and Memory; Hormones and Behavior; Thyroid Hormone and Transcriptional Regulation in the CNS.

Further Reading

Gorski RA, Gordon JH, Shryne JE, and Southam AM (1978) Evidence for a morphological sex difference within the medial preoptic area of the rat brain. *Brain Research* 148: 333–346.

Greco B, Allegretto EA, Tetel MJ, and Blaustein JD (2001) Coexpression of ER beta with ER alpha and progestin receptor proteins in the female rat forebrain: Effects of estradiol treatment. *Endocrinology* 142: 5172–5181.

Imwalle DB, Gustafsson J-A, and Rissman EF (2005) Lack of functional estrogen receptor beta influences anxiety behavior and serotonin content in female mice. *Physiology & Behavior* 84: 157–163.

Kuiper GG, Enmark E, Pelto-Huikko M, Nilsson S, and Gustafsson J-A (1996) Cloning of a novel receptor expressed in rat prostate and ovary. *Proceedings of the National Academy of Sciences of the United States of America* 93: 5925–5930.

Laflamme N, Nappi RE, Drolet G, Labrie C, and Rivest S (1998) Expression and neuropeptidergic characterization of estrogen receptors (ERalpha and ERbeta) throughout the rat brain: Anatomical evidence of distinct roles of each subtype. *Journal of Neurobiology* 36: 357–378.

Li X, Schwartz PE, and Rissman EF (1997) Distribution of estrogen receptor-beta-like immunoreactivity in rat forebrain. *Neuroendocrinology* 66: 63–67.

Merchenthaler I, Lane MV, Numan S, and Dellovade TL (2004) Distribution of estrogen receptor alpha and beta in the mouse central nervous system: *In vivo* autoradiographic and immunocytochemical analyses. *Journal of Comparative Neurology* 473: 270–291.

Mitra SW, Hoskin E, Yudkovitz J, et al. (2003) Immunolocalization of estrogen receptor beta in the mouse brain: Comparison with estrogen receptor alpha. *Endocrinology* 144: 2055–2067.

Nomura M, Durbak L, Chan J, et al. (2002) Genotype/age interactions on aggressive behavior in gonadally intact estrogen receptor beta knockout (betaERKO) male mice. *Hormones and Behavior* 41: 288–296.

Osterlund M, Kuiper GG, Gustafsson J-A, and Hurd YL (1998) Differential distribution and regulation of estrogen receptor-alpha and -beta mRNA within the female rat brain. *Brain Research. Molecular Brain Research* 54: 175–180.

Osterlund MK, Gustafsson J-A, Keller E, and Hurd YL (2000) Estrogen receptor beta (ERbeta) messenger ribonucleic acid (mRNA) expression within the human forebrain: Distinct distribution pattern to ERalpha mRNA. *Journal of Clinical Endocrinology & Metabolism* 85: 3840–3846.

Price RH, Jr, Lorenzon N, and Handa RJ (2000) Differential expression of estrogen receptor beta splice variants in rat brain: Identification and characterization of a novel variant missing exon 4. *Brain Research. Molecular Brain Research* 80: 260–268.

Rissman EF, Heck AL, Leonard JE, Shupnik MA, and Gustafsson J-A (2002) Disruption of estrogen receptor beta gene impairs spatial learning in female mice. *Proceedings of the National Academy of Sciences of the United States of America* 99: 3996–4001.

Toran-Allerand CD (1976) Sex steroids and the development of the newborn mouse hypothalamus and preoptic area *in vitro*: Implications for sexual differentiation. *Brain Research* 106: 407–412.

Wang L, Andersson S, Warner M, and Gustafsson J-A (2001) Morphological abnormalities in the brains of estrogen receptor beta knockout mice. *Proceedings of the National Academy of Sciences of the United States of America* 98: 2792–2796.

Wang L, Andersson S, Warner M, and Gustafsson J-A (2003) Estrogen receptor (ER) beta knockout mice reveal a role for ERbeta in migration of corticol neurons in the developing brain. *Proceedings of the National Academy of Sciences of the United States of America* 100: 703–708.

Gene Therapy and Protection from Stress-Induced Brain Damage

R Sapolsky, Stanford University, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

There are few challenges to health that can be more devastating than a major neurological injury. This is because of the ease with which such injury can cause extensive damage, the frequent irreparability of injury, our limited therapeutic skills, and because such an injury can compromise the essence of whom we are.

Amid this depressing picture, nonetheless, recent decades have brought tremendous insight into the mechanisms underlying neuron death following acute necrotic neurological insults such as seizure, stroke, cardiac arrest, severe hypoglycemia, and concussive head trauma. Central to such death is the 'excitotoxic' cascade, initiated by pathologically elevated concentrations of the highly excitatory neurotransmitter glutamate. Initially, this excess is due to depolarization-induced release of glutamate, but as the insult progresses, the bulk of the excess becomes due to failure of costly high-affinity uptake of synaptic glutamate by glia and the presynaptic neuron. This excess triggers a vast increase in free cytosolic calcium concentrations in the postsynaptic neuron; this is due to influx, via both the NMDA receptor and the opening of voltage-gated calcium channels, the release of calcium from the endoplasmic reticulum, and the failure of costly extrusion and sequestration. The calcium excess, in turn, triggers an array of unregulated and damaging consequences via calcium-dependent enzymes. This includes cytoskeletal degradation, damage to DNA, and, most important, the generation of reactive oxygen species (ROS). Finally, in a subset of neurons, activation of this cascade can activate apoptotic programs that, given sufficient energy, can trigger controlled, noninflammatory programmed cell death rather than the default state of chaotic and proinflammatory necrotic cell death.

Adverse Neurological Effects of Glucocorticoids

With this knowledge has come an understanding of factors extrinsic to the nervous system, which can modulate this pathway. A number of such modulators are endocrine. As one example that will become relevant later in this article, estrogen can strongly protect against necrotic insults, both by decreasing the likelihood of neuron death following an insult and by aiding recovery and compensation.

The focus of this article is on another case of endocrine modulation of necrotic neuronal injury, namely stress, and stress-induced secretion of glucocorticoids (GCs). These adrenal steroid hormones (e.g., cortisol in primates, corticosterone in rats, and synthetic versions such as prednisone and dexamethasone) are secreted in response to virtually all stressors, both physical and psychological. Their short-term actions typify the benefits of the stress response in the face of the acute insults typical of most mammals (e.g., evading a predator or pursuing a prey); they mobilize energy to exercise muscle; synergize with catecholamines in enhancing cardiovascular tone and thus energy delivery; enhance immunity; suppress unessentials such as growth, tissue repair, and reproduction; and enhance aspects of cognition and sensory processing (as well as the synaptic plasticity underlying such processing). However, when secreted in excess, typically in response to psychosocial stressors, GCs can increase the risk of adult diabetes, hypertension, obesity, immunosuppression, and reproductive suppression.

It is in this context that one can appreciate the array of adverse effects of GCs in the nervous system; these are most dramatic in the hippocampus, which contains high concentrations of receptors for GCs. Beginning hours after the first elevation of GC concentrations, hippocampal-dependent cognition and synaptic plasticity is impaired. Within a similar time span, there is inhibition of neurogenesis in the structure. Over longer time periods, the hormone will cause atrophy of the dendritic tree. Finally, over the course of days, the hormone will compromise the ability of neurons to survive necrotic neurological insults and, over the course of months to years, will cause overt neuron death in the hippocampus.

The final case of overt neurotoxicity is demonstrable in both rodents and nonhuman primates; however, it remains unsettled whether this is only a pharmacological effect, requiring elevations of GC levels above those seen physiologically, even in the case of major stressors. Far more reliable and studied is the 'endangerment' of neurons by GCs (i.e., the compromising of neuronal ability to survive insults).

One key step in such endangerment is the ability of GCs to decrease glucose uptake in the hippocampus. It has long been known that GCs inhibit glucose uptake throughout the body, with the exception of exercising muscle, constituting a means of diverting energy to such muscles during an acute physical stressor. This inhibition is protein synthesis dependent and involves the withdrawal of glucose transporter molecules from the cell surface. Inhibition of glucose transport occurs in hippocampal neurons and

Table 1 Adverse effects in the hippocampus of prolonged exposure to stress and/or elevated concentrations of glucocorticoids

Effect	*Comment*
Impaired cognition	Impairment of subsequent acquisition and retrieval of information; inconsistent effects on consolidation; effects are reversible.
Impaired synaptic plasticity	Disruption of long-term potentiation and primed burst potentiation; enhancement of long-term depression; effects are reversible.
Inhibition of adult neurogenesis	Mechanisms and consequences relatively unknown; effects are reversible.
Dendritic atrophy and reorganization of neural networks	Likely to result in decreased connectiveness of dentritic spines; effects are reversible.
Neurotoxicity	Rare and of questionable physiologic relevance; effects are essentially irreversible.
Impaired capacity of neurons to survive necrotic disruption of energetics, of the control of glutamate	Demonstrable with an array of insults; involves insults and calcium trafficking, and impaired antioxidant defenses; effects are reversible.

glia as well (albeit to a lesser extent than in, for example, adipose tissue). Although insufficient to cause a change in energy profiles in the absence of insult, it leaves the hippocampus energetically vulnerable, such that during necrotic insults, the declines in ATP concentrations and mitochondrial potential are exacerbated.

Containing the excitotoxic cascade of neuron death is highly energy dependent; as a result of the inhibition of glucose uptake by GCs, the energy-sensitive steps in the cascade are worsened. The GC-induced worsening of the energy profile biases the presynaptic neuron toward depolarization and thus increased glutamate release. Moreover, glutamate reuptake is impaired, as is uptake by glia, because of the expense of the process. The net result of these presynaptic effects is a greater mobilization of calcium in the postsynaptic neuron. This is exacerbated by the energy problems occurring at that neuron; depolarization will be worsened by GCs, enhancing calcium influx via the NMDA receptor and voltage-gated calcium channel, and there will be less energy for costly extrusion. The worsening of the insult-induced calcium mobilization by GCs will worsen the downstream consequences. This has been shown to involve generation of more ROS, more cytoskeletal damage (i.e., more spectrin proteolysis), and more microtubule abnormalities (i.e., increased levels of the abnormally phosphorylated tau microtubule protein).

A small number of studies have examined whether the consequence of these GC actions is an increase in the fraction of neurons that die necrotically and/or in the fraction that die apoptotically. Despite the well-established fact that GCs can trigger apoptosis in immune cells, there are roughly equal amounts of evidence for and against physiologically elevated levels of GCs worsening the apoptotic profile of neuron death.

Neurons are not passively pummeled by necrotic insults. Instead, there are a variety of defenses that are mobilized. The uptake of synaptic glutamate and the extrusion of calcium qualify as such, but they are rarely thought of this way. More recognized defenses include responses such as expression of heat shock proteins, upregulation of antioxidant genes, and release of inhibitory retrograde neurotransmitters. Although the subject has not been explored in great detail, there have been some reports of GCs impairing the upregulation of such antioxidant genes or curtailing the activity of antioxidant proteins. One such case involves glutathione peroxidase, in which its activity was decreased by GCs decreasing NADH levels (secondary to the hormone's energetic effects), thereby reducing the available pool of reduced glutathione.

These adverse effects of GCs are summarized in Table 1.

Potential Clinical Implications

The previously discussed findings raise a number of worrisome issues in the clinical realm, related to the possible damaging effects of excessive GCs and/or stress. To date, few data are relevant to the ability of GCs to worsen the outcome of neurological insults in humans, beyond some correlative findings (e.g., that GC levels predict the extent of cognitive impairment post-stroke or following electroconvulsive therapy). Instead, the human literature is more focused on instances of GC excess, independent of coincident neurological disease.

Administration of Synthetic GCs

Experimental work with healthy volunteers demonstrates that a few days of exposure to only moderately high doses of a synthetic GCs such as prednisone causes subtle neuropsychological impairments. Synthetic GCs are prescribed in enormous quantities for prolonged periods for a number of disorders, include autoimmune diseases, inflammatory disorders, and to prevent organ transplant rejection. This raises the issue of whether such prolonged GC treatment has

adverse cognitive effects. This has indeed been long recognized clinically, and recent case-controlled studies suggest that the GC treatment impairs hippocampal-dependent memory. As is the case with all of these human syndromes, there is currently no insight as to the cellular basis of these GC effects.

Cushing's syndrome arises from any of a number of tumors, which, as their commonality, result in secretion of supraphysiological GC concentrations. This has long been recognized to result in cognitive impairments, traditionally referred to as 'Cushingoid dementia.' Recent work reports selective hippocampal volume loss in Cushingoid patients, with the magnitude of atrophy correlating with the severity of GC hypersecretion and cognitive impairments. Whatever the underlying cellular basis of this volume loss, it reverses slowly with the correction of the GC excess.

Posttraumatic Stress Disorder

An array of studies report that posttraumatic stress disorder (PTSD) arising from combat trauma or childhood abuse is associated with hippocampal volume loss in adulthood, the magnitude of which correlates with the severity of cognitive deficits and of the trauma. This suggests a model in which the volume loss arises as a result of the GC excess during the trauma and/or the posttraumatic period. However, this view is controversial for two reasons.

First, does the volume loss emerge posttrauma or, for reasons unrelated to stress, does it precede and predispose toward PTSD (i.e., a vulnerability model, reflecting the fact that trauma results in PTSD in only a subset of victims)? There is evidence for both ideas. As particularly strong support for the predisposition model, Vietnam veterans with PTSD secondary to combat trauma demonstrated hippocampal volume loss. As an important feature of this unique study, these subjects were members of pairs of identical twins, in which the sibling had not gone to Vietnam nor suffered from PTSD for other reasons; however, the same degree of volume loss was observed in these siblings as well. As particularly strong evidence for the model in which atrophy emerges as a result of the trauma rather than precedes it, children with PTSD due to abuse do now show hippocampal volume loss; it only emerges later on, gradually over subsequent months to years.

Second, it remains unclear whether GC levels are, in fact, elevated in PTSD. One literature reports such elevations, whereas another reports lower than normal levels, accompanied by enhanced sensitivity to negative feedback and downregulation of GC receptors in lymphocytes. Confounding this is a great deal of uncertainty (due to minimal data) as to GC levels during the relevant traumas.

Regardless of the resolution of these controversies, there is little evidence that the volume loss is reversible with time.

Major Depression

Stress and GCs play a central role in understanding the onset of the first few episodes of major depression, the neurochemistry and symptomatology of the disorder, and its endocrine concomitants. Importantly, approximately half of individuals with major depression have elevated GC levels. Commensurate with this, long-term depression is associated with hippocampal volume loss, the extent of which is correlated with the duration of depression and severity of cognitive impairments. This literature is weakened by the lack of demonstration that the volume loss is only found among the depressives with the elevated GC levels. As with the other syndromes, the cellular basis of the hippocampal volume loss is not known. One controversial theory is that it reflects GC/stress-induced inhibition of hippocampal neurogenesis; currently, there are no definitive data for or against this idea. There is also only minimal evidence that the volume loss precedes and predisposes toward depression. Finally, whatever the underlying cellular mechanism, the volume loss is demonstrable long after depressions have been pharmacologically contained.

Occupational Stress

Currently, the notion that everyday stressors can lead to hippocampal atrophy is almost entirely speculative. One provocative study supporting this idea concerned the fact that airlines differ as to the length of time that flight crews are allowed to recover after flights that involve major circadian shifts. After controlling for length of time in the profession and the total number of such shifts, individuals from the airline that allowed the shortest recovery time had significantly smaller temporal lobes and worse hippocampal-dependent cognition relative to crews from the airline allowing the longest recovery time.

Adverse Effects of Stress in the Amygdala

As should be clear, the literature regarding the deleterious effects of stress and GCs in the brain has been heavily skewed toward the hippocampus. A recent literature, however, shows there to be different GC/stress effects in the amygdala, ones that can be argued to be deleterious as well.

An extensive literature implicates the amygdala in fear and anxiety-related behaviors, findings derived from both animal and human studies (in which, for example, brain imaging shows amygdaloid activation

in response to frightening faces, or hyperactivity of the amygdala in individuals with PTSD). Thus, whereas the pathologies of the hippocampus involve decreased function (impaired hippocampal-dependent cognition, synaptic plasticity, and dendritic atrophy), those of the amygdala involve the opposite for each of those endpoints. Recent work shows that stress and GCs augment this picture of the amygdala as 'hyperexcitable.' As such, both will enhance fear conditioning and the startle response, phenomena that reflect amygdaloid-dependent learning, and will enhance the underlying increases in synaptic plasticity. Most remarkably, stress augments the dendritic tree in amygdaloid neurons, which is a plausible contribution to long-term increase in excitability.

On the level of cell biology, these are instances in which stress and GCs improve the function of neurons, and it remains a fascinating mechanistic question as to how the same external stimulus (GCs) can cause diametrically opposite cellular effects in the hippocampus and amygdala. On the more integrative level, insofar as enhanced excitability of the amygdala helps mediate an array of maladaptive behaviors (e.g., PTSD), the effects of stress and GCs can be viewed as being adverse in the amygdala.

Strategies to Protect against Adverse GC Effects

The range of stress-related pathologies throughout the body has long emphasized the importance of minimizing activation of the stress response in psychosocial circumstances; the growing literature regarding the adverse neurobiological effects of stress add to this conclusion. Although interventions meant to minimize the stress response are predominately psychological in nature, more biological approaches have been tried as well (Table 2). For example, some investigators have attempted to decrease the impact of the elevated GC levels frequent in major depression (because one important consequence of the elevated levels is to increase the severity of the anhedonia in the disease). Such approaches have taken the form of treating individuals with adrenal steroidogenesis inhibitors or, more recently, antagonists to one type of GC receptor.

Some first tentative attempts at developing 'anti-stress' gene therapy approaches have been made as well. Gene therapy in the nervous system is a burgeoning field. The approach most typically involves the use of neurotropic viruses (i.e., viruses that can infect postmitotic cells such as neurons), including herpes simplex virus or adenovirus. Potentially protective genes are inserted into crippled (i.e., safe) versions of the virus, and the resulting vector is introduced into the nervous system, where the transgene is expressed. Nearly all such gene therapy studies have targeted the basic biology of neuron death or neuronal recovery in the context of neurodegenerative diseases and acute necrotic neurological insults. Considerable preclinical success has been reported with these approaches, and the first few clinical trials derived from those are in progress.

Making use of this technology, some gene therapy approaches have been designed to block the disruptive effects of GCs and stress. Two components of these vectors have been developed. The first is to use vectors that will only express in response to elevated GC levels. This has been accomplished with the introduction of glucocorticoid responsive elements upstream of the transgene to be expressed. The result is a vector that can be introduced into neurons in advance of a stressor and that will then be activated out of latency by a stressor.

The second element is to identify genes that, if overexpressed, will blunt GC action in some manner. To date, three have been tried. In the first two cases, the transgene expressed was either a renal enzyme that unidirectionally degrades GCs or a mutant transdominant version of the glucocorticoid receptor; both decrease the hormone's access to its receptors. The third represents an advance in gene therapy technology and takes advantage of the fact that estrogen, as discussed, is neuroprotective in many realms. This

Table 2 Approaches to block the deleterious effects of glucocorticoids in the nervous system

Intervention	*Rationale*
Metyrapone, aminoglutethimide	Blocking adrenal steroidogenesis
RU486 (Mifepristone)	Blocking glucocorticoid receptors
Gene therapeutic overexpression of	
• 11-β-Hydoxyrosteroid dehydrogenase	Degradation of glucocorticoids
• Transdominant glucocorticoid receptor	Outcompeting wild-type glucocorticoid receptor for ligand
• ER/GR chimera containing the hormone-binding domain of the glucocorticoid receptor and the DNA-binding domain of the estrogen receptor	Converting deleterious glucocorticoid effects into protective estrogenic effects

vector exploits the modular structure of steroid receptors by expressing a chimeric gene that is half a GC and half an estrogenic receptor; the hormone-binding domain is derived from the GC receptor, and the DNA-binding domain is derived from the estrogen receptor. The result is that GCs, when bound to this receptor, trigger its translocation to estrogen responsive elements in the DNA, triggering the expression of estrogen-regulated genes; as such, this vector transduces the endangering effects of GCs into the protective effects of estrogen. Recent work suggests that these 'anti-GC' vectors can not only protect against the endangering effects of GCs and stress in the context of neurological insult but also are protective against their disruptive effects on cognition and their enhancement of amygdaloid function.

Conclusions

We are now nearly three-quarters of a century into recognizing the potentially pathogenic effects of sustained stress, often via the excessive secretion of GCs. The past few decades have shown that this pathogenicity applies to the nervous system as well, and that this act may be relevant to understanding the maladaptive responses to trauma, the side effects of heavy pharmacological use of synthetic GCs, the extent of neurotoxicity following the most common acute neurological insults, and why some brains will age more successfully than others. However, the knowledge of the cell biology of how these adverse actions emerge is also paving the way for the development of some potential therapies against the neurological effects of stress.

See also: Adrenal Steroids: Biphasic Effects on Neurons; Posttraumatic Stress Disorder: Overview; Stress and Vulnerability to Brain Damage; Stress Response: Genetic Consequences; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History.

Further Reading

Bremner J, Vythilingam M, Vermetten E, Anderson G, Newcomer J, and Charney D (2004) Effects of glucocorticoids on declarative memory function in major depression. *Biological Psychiatry* 55: 811–819.

Bremner JD (2005) Effects of traumatic stress on brain structure and function: Relevance to early responses to trauma. *Journal of Trauma and Dissociation* 6: 51–68.

Cahill L (2003) Sex- and hemisphere-related influences on the neurobiology of emotionally influenced memory. *Progress in Neuropsychopharmacology and Biological Psychiatry* 27: 1235–1241.

Cho K (2001) Chronic "jet lag" produces temporal lobe atrophy and spatial cognitive deficits. *Nature Neuroscience* 4: 567–570.

Flores BH, Kenna H, Keller J, Solvason HB, and Schatzberg AF (2006) Clinical and biological effects of mifepristone treatment for psychotic depression. *Neuropsychopharmacology* 31: 628–636.

Gilbertson MW, Shenton ME, Ciszewski A, et al. (2002) Smaller hippocampal volume predicts pathologic vulnerability to psychological trauma. *Nature Neuroscience* 5: 1242–127.

Gould E and Gross C (2002) Neurogenesis in adult mammals: Some progress and problems. *Journal of Neuroscience* 22: 619–623.

Kaufer D, Ogle W, Pincus Z, et al. (2004) Restructuring the neuronal stress response with anti-glucocorticoid gene delivery. *Nature Neuroscience* 7: 947–954.

Keenan PA, Jacobson MW, Soleymani RM, Mayes MD, Stress ME, and Yaldoo DT (1996) The effect on memory of chronic prednisone treatment in patients with systemic disease. *Neurology* 47: 1396–1402.

Kim JJ and Diamond DM (2002) The stressed hippocampus, synaptic plasticity and lost memories. *Nature Reviews Neuroscience* 3: 453–462.

Kristian T and Siesjo B (1998) Calcium in ischemic cell death. *Stroke* 29: 705–718.

Maren S and Quirk G (2004) Neuronal signaling of fear memory. *Nature Reviews Neuroscience* 5: 844–850.

McEwen BS (2005) Glucocorticoids, depression, and mood disorders: Structural remodeling in the brain. *Metabolism* 54: 20–23.

Nemeroff CB, Bremner JD, Foa EB, Mayberg HS, North CS, and Stein MB (2006) Posttraumatic stress disorder: A state-of-the-science review. *Journal of Psychiatric Research* 40: 1–21.

Newcomer J, Craft S, Hershey T, Askins K, and Bardgett M (1994) Glucocorticoid-induced impairment in declarative memory performance in adult humans. *Journal of Neuroscience* l4: 2047–2053.

Nicholas A, Munhoz C, Ferguson D, Campbell L, and Sapolsky R (2006) Enhancing cognition after stress with gene therapy. *Journal of Neuroscience* 26: 11637–11643.

Roy M and Sapolsky R (2003) The exacerbation of hippocampal excitotoxicity by glucocorticoids is not mediated by apoptosis. *Neuroendocrinology* 77: 24–31.

Sapolsky R (1999) Stress, glucocorticoids and their adverse neurological effects: Relevance to aging. *Experimental Gerontology* 34: 721–732.

Sapolsky R (2000) Glucocorticoids and hippocampal atrophy in neuropsychiatric disorders. *Archives of General Psychiatry* 57: 925–931.

Sapolsky R (2003) Neuroprotective gene therapy. *Nature Reviews Neuroscience* 4: 61–68.

Sapolsky R, Romero M, and Munck A (2000) How do glucocorticoids influence the stress-response? Integrating permissive, suppressive, stimulatory, and preparative actions. *Endocrine Reviews* 21: 55–81.

Sheline YI, Mittler BL, and Mintun MA (2002) The hippocampus and depression. *European Psychiatry* 17: 300–305.

Shors T (2006) Stressful experience and learning across the lifespan. *Annual Review of Psychology* 57: 55–85.

Starkman M, Gebarski S, Berent S, and Schteingart D (1992) Hippocampal formation volume, memory dysfunction, and cortisol levels in patients with Cushing's syndrome. *Biological Psychiatry* 32: 756–763.

Vyas A, Mitra R, Shankaranarayana Rao BS, and Chattarji S (2002) Chronic stress induces contrasting patterns of dendritic remodeling in hippocampal and amygdaloid neurons. *Journal of Neuroscience* 22: 6810–6818.

Wolkowitz O, Reus V, Chan T, et al. (1999) Antiglucocorticoid treatment of depression: Double-blind ketoconazole. *Biological Psychiatry* 45: 1070–1079.

Won S, Kim D, and Gwag B (2002) Review cellular and molecular pathways of ischemic neuronal death. *Journal of Biochemistry and Molecular Biology* 35: 67–86.

Yang SH, Liu R, Wu SS, and Simpkins JW (2003) The use of estrogens and related compounds in the treatment of damage from cerebral ischemia. *Annals of the New York Academy of Sciences* 1007: 101–107.

Gene Therapy: Direct Viral Delivery

B K Kaspar, The Ohio State University, Columbus, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

As populations age, the burden of health-related diseases due to nervous system disorders rises. Given the challenge to deliver drugs to the complex structure and diverse cell types of the nervous system, new therapeutic approaches are being investigated. Recent progress in molecular biology and virology has allowed tremendous advancements in delivery of recombinant DNA to the brain. This article provides an update on the status of direct viral gene delivery and examines the challenges to advance this approach as a therapeutic modality.

Direct viral injections focused on treating or curing acquired and inherited disorders can be utilized through a variety of approaches, including (1) delivery of a gene in excess (such as a trophic factor), (2) suppression of a gene (via antisense or small interfering RNA), (3) swapping an abnormal gene with a normal gene through homologous recombination, (4) controlling the regulation of endogenous gene expression (via transcription factor or zinc finger approaches), or (5) modulating immune defense mechanisms by either activating or suppressing immune function. Furthermore, direct viral injection can be used to generate animal models of mutant or toxic gene expression or, using *loxP* technology, conditional gene deletions.

There have been tremendous advances in the development of recombinant viral vectors that transduce postmitotic cells of the central nervous system with no associated toxicity. Among the two most commonly used are recombinant adeno-associated virus (AAV) vectors and recombinant lentiviral vectors. Initial studies in rodents have shown that neurons are the predominant cells that are transduced, with little transduction of other central nervous system (CNS) cell types. Both adeno-associated virus and lentiviral vectors have also been utilized in larger species, including in the monkey CNS, demonstrating a neural tropism. Over the years, it has been shown that these vectors are very nontoxic, with little to no activation of astrocytes or immune reactions against the vector. Additionally, long-term gene expression has been achieved, lasting well over 1 year post-infection with no evidence of cell death.

Utilizing precise maps of the brain and spinal cord, direct injection provides an ability to specifically target regions of the brain. The dynamics of vector spread have been studied, and, given the structure of the target and well as the type of virus; vector spread may be confined or may spread several millimeters. It has become apparent in many studies that virus may be transported in a retrograde and anterograde fashion, thereby targeting multiple areas of the CNS. Therefore, in many situations a direct injection will target multiple projection pathways.

Delivery Platforms

Numerous viral-mediated gene delivery platforms exist, each having particular advantages and disadvantages. Two of the most common platforms utilized to target nondividing cells, such as neurons, are adeno-associated virus and lentiviral vectors. These two vectors are attractive since they deliver genes into postmitotic cells and have demonstrated long-term gene transduction with limited or no immune response to the vector. Retroviral vectors are used for direct injection into the brain to target dividing cells, such as astrocytes or neural progenitor cells. It has recently been shown that retroviral vectors could transduce neural progenitor cells in different regions of the brain, with long-term gene expression. In addition, adenoviral and herpes simplex viruses are advantageous for transducing all cell types in the CNS very efficiently. Both of these vectors have the potential to induce an immune response, but new modified, or 'gutless,' versions of the viruses have bypassed many of the immune responses. Adenoviral and herpes simplex vectors are commonly used in many brain-related cancer studies.

The following sections provide a brief description of the various viruses that are used most commonly for direct injections into the brain.

Adeno-Associated Viral Vectors

The AAV is a member of the genus *Dependovirus* of the family Parvoviridae, among the smallest of DNA animal viruses, consisting of a virion particle approximately 25 nm in diameter and composed of an icosohedron-structured protein capsid outer coat. The viral genome is linear, consisting of a single-stranded DNA molecule of approximately 4.7 kb which contains two open reading frames (ORFs) flanked by 145 bp inverted terminal repeats (ITRs). AAV requires adenovirus or herpes virus as helper virus to initiate and complete its viral life cycle of rescue, replication, and packaging. Recombinant AAV-based vectors have the viral genome deleted, other than the ITRs, allowing for insertion of vector

construct to a maximum of approximately 4.9 kb. In recombinant form, most of the AAV vectors persist as concatemers, forming linear and circular forms, which lead to long-term persistence of the viral genome in transduced cells.

Specifically, to make these vectors recombinant, the rep- and cap-coding regions are removed and a promoter with a respective cDNA and polyadenlyation signal is cloned in between the two ITRs. For viral production, this recombinant AAV plasmid, along with *a trans* rep and cap plasmid and helper adenovirus or adenoviral plasmid, is transfected into cells such as human embryonic kidney (HEK293) carcinoma cells. Viral production consists of lysing cells followed by gradient ultracentrifugation and/or column chromatography. Vector preparations of titers between 10^{12} and 10^{13} are now routinely produced. Viral packaging efficiency and purity are important for optimal transduction *in vivo*. Of importance, AAV vectors can target postmitotic cells such as neurons. Most serotypes are neurotropic; however, some have achieved astrocyte gene transfer. Most regions of the brain have been transduced by AAV, with expression lasting long term in rodents as well as nonhuman primates. Currently, the most widely utilized serotype of AAV has been based on AAV-2, which utilizes the heparan sulfate proteoglycan as the primary receptor as well as $\alpha v\beta 5$ integrin and fibroblast growth factor receptor 1 (FGF-1R) as coreceptors. This allows a wide host range of tissue and cellular types. Other serotypes have shown great promise for expression in ependymal cells as well as various regions of the brain.

In sum, recombinant AAV vectors offer high promise for direct *in vivo* injection. This is due to the high stability of the virus, including an ability to inject highly concentrated stocks. AAV vectors have shown the ability to infect a wide spectrum of cells, including postmitotic cells, leading to long-term expression. Furthermore, AAV vectors are nonpathogenic, induce no etiological factors, and have been safely demonstrated in a number of clinical trials. However, AAV has a smaller packaging capacity compared to other vectors and requires several weeks to reach the maximal level of gene expression. Additionally, the scale of virus production and the purification processes remain a challenge, though significant progress is being made. New approaches using self-complementary vectors may offer some benefits for improving the timing and level of expression in the brain.

Retroviral

Historically, retrovirus-mediated gene transfer methods have been the most commonly used in gene transfer techniques, based on the Moloney murine leukemia virus (MoMuLV) as the backbone. Retroviruses target dividing cells and involve integration of the proviral DNA into the infected cell, allowing stable transmission to subsequent generations of the infected cells. Retroviruses have several advantages, including the ease of manipulation, the ability to clone relatively large genes (up to 7 kb), and the ability to infect a variety of cell types from various species. The majority of viral genes can be removed from the vectors, requiring regions in and around the long terminal repeats (LTRs) for packaging the RNA genome into the viral coat and allowing reverse transcription and integration. Since the proteins necessary for the viral replication can be provided *in trans*, replication-defective viruses can be generated easily. Production of retroviral vectors typically utilizes packaging and producer cell lines in which the viral gag, pol, and env proteins are expressed.

Given the ability to stably infect dividing cells, current direct injections of retroviral vectors to the brain have focused on tumor cell injections as well as direct injections into regions of neurogenesis to target neural progenitor cells. This approach to label neural progenitor cells has provided a very useful tool to investigate the neurobiology and histology of neural progenitor cells *in vivo*. Furthermore, this approach may be very useful to target these cells for therapeutic options in a number of diseases. There are several problems with retroviruses, such as stability (very short half-life of the virus) as well as the difficulty in making concentrated high-titer stocks. Furthermore, promoter shutdown has been an issue; however, advances in vector and promoter design have proved successful.

Lentiviral

Since conventional retroviruses only infect dividing cells, considerable attention has been directed to vectors based on the human immunodeficiency virus (HIV) or equine infectious anemia virus (EIAV), species of the genus *Lentivirus*, which infect postmitotic cells. A generation of HIV-based retroviral vectors that are not limited to infection of replicating cells has gained considerable use as a means to deliver genes to the CNS. Unlike other retroviruses that rely on the breakdown of the nuclear membrane and passive diffusion into the nucleus during cell division to transport their genetic information, HIV can be actively transported to the host cell nucleus by the cellular nuclear transport mechanism. This HIV-based expression system can efficiently transfer genes into nondividing cells. The HIV-based vector has been developed through several generations, reducing the potential for recombination. Extensive studies selectively removing the HIV virulence genes (*tat*, *vif*, *vpu*, *vpr*, and *nef*) have

demonstrated that these new generations are helper-virus free. Lentiviruses have an advantage in that larger transgenes may be cloned into the vector, versus AAV vectors. Packaging of lentiviruses is very similar to that of retroviruses, in which the *gal–pol* gene is used, along with the *rev* gene in some packaging systems. In addition, the packaging system includes the pseudotyping of the envelopes with other viruses, such as the vesicular stomatitis virus glycoprotein (VSVG). One drawback in lentiviruses is that production in large quantities for *in vivo* studies remains cumbersome and also that purification techniques have not been fully developed for highly purified vectors. Recent advances in column chromatography may aid in the enhancing lentiviral purification. For human studies, it may be useful to develop vectors with nonhuman lentiviruses, such as EIAV, bovine immunodeficiency virus, or simian immunodeficiency virus.

Adenoviral

Adenoviral vectors have been one of the most widely used viruses in gene delivery. They are nonenveloped, double-stranded (ds) DNA viral vectors with a packaging capacity of approximately 35 kb. Over 50 different adenoviral serotypes exist, grouped into six species. Recombinant adenoviral vectors are commonly derived from types 2 and 5. Adenoviruses of these serotypes have not been associated with human malignancies. The majority of these vectors are *E1* gene-deleted to render the virus replication deficient, along with *E3* deletions to allow greater packaging space. Most of the serotypes are able to enter most cells via the cellular coxsackie/adenovirus receptor (CAR), for attachment, and utilize αv integrins for internalization. This ability is indeed a major strength of adenoviruses, enabling a broad tissue tropism and infection of many cell/tissue types, with a very fast expression profile. In addition, large genes can be inserted relatively easy and adenoviral vectors are produced with relative ease at high titers and are easily purified. Virus is produced typically in HEK293 cells that stably express the *E1A* and *E1B* genes, allowing for replication and packaging. Transgene expression from first-generation vectors is very strong, but is reduced and abolished 2–3 weeks following direct injection. These recombinant adenoviral approaches have been utilized for many direct brain injections for tumor cell killing. The adenoviruses are recognized by the immune system and therefore have problems with long-term gene expression. Second-generation vectors have been used in an attempt to decrease immune responses to the vector, by deleting early regions involved in DNA replication, such as the *E2A*, *E2B*, and *E4* regions.

In sum, adenoviral vectors can be produced in high titers that hold a large amount of foreign DNA with the ability to infect postmitotic and dividing cells. The major disadvantage is that adenoviral proteins are highly immunogenic and many people have circulating antibodies and cellular immune responses toward the virus. Other drawbacks are that transgene expression is relatively unstable because the virus remains episomal, and that the virus may be toxic at high doses. The recent development of new generations of adenoviral vectors might overcome the toxicity and immunogenicity issues experienced with the previously used adenoviral vectors in the nervous system. Helper-dependent ('gutless') adenoviral vectors have shown reduced toxicity and larger cloning capacity compared to the standard adenoviruses. However, these gutless adenoviral systems are difficult to produce in large quantities.

Herpes Simplex Virus

Recombinant herpes simplex virus (HSV-1) is a double-stranded virus that has been produced as a recombinant vector system based on a natural neurotropic virus. The two types that are being exploited for *in vivo* injection are the herpes virus amplicon and the recombinant virus. The virus has a large cloning capacity and has been shown to remain in a latent state within neurons. Recombinant HSV-1 vectors are created by deleting the nonessential viral genes and replacing them with a transgene or a construct of interest (up to approximately 50 kb). HSV-1 has a wide ability to infect different cells and tissues. The virus has the capacity to retrogradely transport from remote sites, affording HSV-1 the ability to target CNS sensory and motor neurons in a highly efficient manner. There is also evidence to support transsynaptic passage from neuron to neuron. Most neurons are efficiently infected by HSV-1 vectors; however, HSV-1 is not solely a neurotropic virus, and in fact it will infect astrocytes and oligodendrocytes. HSV-1 vectors have been predominantly utilized experimentally for tumor cell killing. There is, however, potential to develop an immune response to the vector, and at present, sustained transgene expression has been difficult to achieve.

Control of Vector Systems

There are numerous methods to gain control of vector systems. One method is to engineer cell type-specific promoters into the recombinant vector. In this case, known cell type promoters are cloned in front of the recombinant transgene, which leads to specific cellular expression. Numerous promoters have been utilized, including promoters specific for neurons and

for specific subtypes of neurons (such as motor neurons), as well as astrocytes. In addition to cell type-specific expression, numerous advances have allowed for regulated transgene expression. In this case, transgene expression can be turned on and off as well as regulated as to the level of expression. Technology of regulatable systems is growing, and new developments are under way that are currently being tested in a variety of vector systems. Virologists have also utilized another method to gain control of vector systems: each viral particle is coated or encapsidated, allowing researchers to modify the outer coat or capsid. Many viruses have numerous serotypes and are species specific. It has been determined that these different serotypes display different transduction profiles. Some serotypes work better than others in neurons. In addition, numerous viral vectors can be pseudotyped with various proteins that allow the virus to infect cells more efficiently. Also, specific engineering can be employed on the capsid by inserting peptides or mutating the capsid regions in specific or nonspecific regions to alter viral tropism and transduction properties.

The control of vector systems in the CNS is important for a number of reasons. First, any therapeutic approach will necessitate the therapy being delivered to the proper cell type at the right dosage. Furthermore, while the brain is normally immune privileged due to the blood–brain barrier, there is considerable evidence that immune responses to viral vectors need to be considered in the design of direct viral injections. In the case of AAV, approximately 80% of the human population is seropositive for antibodies to wild-type AAV 2 capsid protein. The presence of neutralizing antibodies may prevent effective infection of this serotype. While it was originally thought that the blood–brain barrier would prevent peripheral immune responses, it has been shown that direct viral delivery locally disrupts the blood–brain barrier. Vector redosing is not possible with the same serotype.

With the demonstration that viral vectors are highly efficient for delivering genes, with long-term gene expression in rodents exceeding 18 months and over 3.5 years in primate brain, it is apparent that gene delivery may be a useful modality for the treatment of progressive neurodegenerative disorders. The following sections describe recent preclinical research of direct viral delivery in a number of degenerative disorders.

Diseases of the CNS

While not exclusive to all disorders, direct viral injection to the nervous system has been studied in numerous diseases and injuries, including Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis, lysosomal storage disorders, and Huntington's disease (HD).

Alzheimer's Disease

Affecting approximately 20 million people worldwide, AD is the most prevalent neurodegenerative disease. The disease is characterized by progressive cell death in cortical structures, with progressive memory loss, impaired judgment, confusion, and personality changes. The disease is characterized by a neuropathological component of accumulated amyloid beta (Aβ) plaques and neurofibrillary tangles throughout the brain. Since AD affects multiple regions of the brain, including widespread structures such as the cortex, direct viral delivery has been difficult. To date, there are no effective drugs for AD, although several inhibitors of acetylcholinesterase are prescribed. New strategies have focused on the use of trophic factors and antiapoptotic factors to target the cholinergic neurons of the basal forebrain. Nerve growth factor (NGF) is the most potent neurotrophic factor for cholinergic neurons and has been tested by direct viral delivery of NGF to spare cholinergic neurons of the basal forebrain in animal models. Furthermore, NGF has been shown to increase acetylcholine levels, which leads to improved memory, and is therefore a major candidate being pursued in the clinic. New viral approaches have focused on using natural peptidases that hydrolyze Aβ. Using a number of different viral vectors, neprilysin has shown promise in decreasing Aβ levels in AD mice. Another approach at reducing amyloid plaques has focused on the use of developing antibodies to clear the plaques. The use of vectors that express Aβ has induced autoantibodies that have been shown to vaccinate against and thereby reduce the levels of Aβ. Still another method using viral vectors has focused on the Aβ proteolytic pathway to prevent the toxic form of Aβ. Using a small interfering RNA (siRNA) approach against a protease (BACE-1) in the cleavage of amyloid precursor protein, cleavage has also shown significant promise as a therapy. Delivering the therapy in AD to the right cells in a widespread manner and utilizing the most efficacious genes are current challenges in AD research.

Parkinson's Disease

PD is one of the most studied neurodegenerative diseases and is the second most common disorder following AD, with symptoms manifesting at approximately 60 years of age. This neurological disorder is due to a gradual degeneration of dopamine-containing neurons in the substantia nigra, leading to the loss of dopamine transmission in the striatum. In fact, when

symptoms of Parkinson's appear, approximately 80% of striatal dopamine is lost. The loss of striatal dopamine leads to a resting tremor, bradykinesia, and akinesia. The etiology of the idiopathic form of the disease is largely unknown, although there are certain forms of the disorder with genetic components. Currently the only pharmaceutical therapy is L-dopa (L-3,4-dihydroxyphenylalanine), a precursor of the neurotransmitter dopamine. As the disease progresses over 3–5 years, this treatment becomes ineffective. Numerous studies have focused on direct gene delivery to deliver factors to prevent the neurons from dying, or to serve in dopamine replacement strategies. Several studies have utilized antiapoptotic factors or neurotrophic factors, including glial-derived neurotrophic factor for potent neuroprotection and regeneration of dopaminergic neurons. Additionally, there have been studies for striatal L-dopa for dopamine delivery via intrastriatal gene delivery. Numerous genes have been employed, including those encoding tyrosine hydroxylase (TH), GTP-cyclohydrolase 1 (GTPCH-1; for tetrahydrobiopterin availability), and aromatic amino acid decarboxylase (AADC; to produce dopamine). Still another approach has targeted the subthalamic nucleus to modulate basal ganglia during the nigrostriatal dopamine degeneration. In this therapy, direct injection into the subthalamic nucleus of glutamic acid decarboxylase (GAD) has been tested to increase the inhibitory neurotransmitter γ-aminobutyric acid (GABA). All of these approaches are showing promise in rodent and nonhuman primate models of PD using direct viral injection into the CNS.

Clinical trials in PD A neurotrophic strategy is currently in clinical trials to treat PD. This approach has utilized a neurotrophic factor, neurturin, to restore lost motor function and protect against further losses of dopaminergic neurons. Using an AAV encoding this gene, the virus is delivered into dopamine neuron-rich regions of the brain. Phase I safety trials in 12 patients with advanced PD were injected with the virus into the putamen at two different doses. Patients that received the low dose of virus demonstrated approximately a 36% improvement in the Unified Parkinson's Disease Rating Score (UPDRS) while off their current drug therapies, which was detected at 9 months post-gene delivery. The group of patients receiving a fourfold higher dosage experienced similar improvements at 6 months postinjection. These patients had persistent improvements throughout the 1-year analysis. No untoward or serious adverse events were noted in the study, demonstrating that gene delivery was safe and well tolerated over long periods of time. Self-reports from patients demonstrated a 50% reduction in the time when normal medication was ineffective and a doubling of time without dyskinesias while on medication. Currently, a phase II trial of this treatment is underway, enrolling 51 patients with advanced PD, and will be studied for 12 months for safety and efficacy.

A second trial using AAV virus is currently underway to evaluate the safety and efficacy of a therapy designed to restore and extend the therapeutic effectiveness of L-dopa, the precursor to dopamine. This vector encodes aromatic amino acid decarboxylase and is currently administered bilaterally to the putamen of patients. As of 2006, four patients had received treatment. Positron emission tomography (PET) anlaysis showed an increase in the ratio of L-dopa to dopamine, demonstrating an increase in the AADC tracer fluoro-L-*m*-tyrosine (FMT) uptake at 1 month in all four patients following gene delivery, with slightly less uptake at 6 months. To date, there has been no report on changes in the average UPDRS motor score. Further recruitment is currently under way to evaluate this clinical study.

Another trial that has demonstrated safety and function of AAV has involved 11 men and one woman with PD, a disease noted for substantial reduction in the activity and amount of GABA in the brain; GABA is a major inhibitory brain neurotransmitter that suppresses excessive neuronal firing. In this trial approach, using an adeno-associated viral vector, the glutamic acid decarboxylase (GAD) gene was injected into a single site of the brain, the subthalamic nucleus, that is a key regulatory center within the motor circuit. In this procedure, only one side of the brain was injected with virus. Using the UPDRS and positron emission tomography scans, the investigators monitored patients for safety and the potential efficacy. They found that the injection was safe and well tolerated, and no adverse events were reported during the trial over 1 year. Interestingly, there was a 25–30% improvement in the UPDRS scores in patients taken off the standard drug therapies for a period of time, to evaluate effects of the gene therapy treatment alone. Similar improvements were found even while patients were on their normal medications. These improvements persisted over the course of the trial. Trends noted on reduced medication linked dyskinesias and improvements in daily living, though the data did not reach statistical significance. Furthermore, PET scans showed that the treated side of the brain had more normal levels of activity than did the nontreated side. Larger, efficacy-driven trials are planned to evaluate this therapy.

Amyotrophic Lateral Sclerosis

Also known as Lou Gehrig's disease, ALS is a degenerative disease of the motor system affecting upper

and lower motor neurons of the motor cortex and spinal cord. The disease is characterized by loss of skeletal muscle control, ultimately resulting in death due to loss of muscles involved in respiration. Currently there are no effective treatments for this disease, but significant progress and promise have been demonstrated by gene therapy approaches. Numerous strategies have evaluated viral-mediated gene delivery of neurotrophic factors – glial-derived neurotrophic factor (GDNF), insulin-like growth factor-1 (IGF-1), and vascular endothelial growth factor (VEGF) – as well as antiapoptotic genes using numerous viral vectors, including adenoviruses, AAV, and lentiviruses. In addition, there are developments to deliver antisense or siRNAs directed against one of the known genetic mutations accounting for the familial form of the disease in which patients have mutations within the superoxide dismutase 1 (*SOD1*) gene. Several different approaches have been utilized to deliver a therapeutic gene to the spinal cord. These approaches have utilized the concept of remote delivery by injecting the therapeutic vector intramuscularly; the virus is taken up by the axon terminal and transported in a retrograde fashion to the motor neuron cell body. This retrograde transport approach of the viral vector was accomplished using adenoviral and adeno-associated viral vectors as well as rabies glycoprotein pseudotyping of lentiviral vectors. Other approaches have investigated direct viral delivery by intraparenchymal delivery to the spinal cord. These studies have all showed the promise to significantly delay disease onset and/or progression in mouse models of the disease. Several of the approaches are being developed for clinical translation. Ultimately, as further genetic analysis of patients leads to new targets and mechanisms of the disease, the efforts gained in targeting spinal cord motor neurons and the surrounding cellular environment will lead to more efficacious therapies.

Lysosomal Storage Disorders

Over 40 known lysosomal disorders result from deficient enzymes, leading to accumulation of cellular proteins in the lysosomes. A deficiency of lysosomal acid hydrolases results in several lysosomal disorders, including mucopolysaccharidosis type VII (MPS VII), Niemann–Pick A, MPS IIIB, Krabbe disease, and Batten disease. Current therapies include enzyme replacement strategies. Human therapy for prenatally determined lysosomal disorders requires administration of the missing enzyme postnatally. Approaches to gene therapy have tested numerous enzyme replacement strategies utilizing viral vectors, and new approaches are investigating minimally invasive, yet widespread gene delivery to the CNS in addition to delivering the gene to peripheral organs. The CNS is targeted to treat the cognitive neural problems associated with the disease. A clinical trial for Batten disease is currently underway to transfer a correct version of the ceroid lipofuscinosis neuronal 2 (*CLN2*) gene into infants.

Huntington's Disease

HD is a fatal neurodegenerative disease classified by trinucleotide repeats of CAG in the first exon of the HD gene. These polyglutamine expansions require a specific length to trigger neurodegeneration, in which the mutant protein forms aggregate, leading to loss of GABA spiny neurons in the caudate putamen. The disease leads to cognitive defects and motor abnormalities. Trophic factor support utilizing a number of viral vectors is currently being tested to protect against the GABAergic cell loss. Furthermore, new approaches utilizing short-hairpin RNAs (shRNAs) to target the mutant huntington gene has shown significant promise for a therapeutic modality in this disease. Given that a number of diseases of the CNS are due to the expansion of polyglutamine repeats, this directed approach for silencing the mutant protein may be further applicable to other diseases. Proper delivery to target the affected cells in the disease and the optimal timing of delivery are challenges that researchers in the field are currently developing.

Conclusion

In addition to use in therapeutic approaches to the diseases discussed in this article, gene therapy by direct viral injection has shown significant promise for the treatment of brain disorders such as Canavan's disease, and in treating pain, epilepsy, stroke, and spinal cord injury; besides, there is interest in direct viral injection in a wide spectrum of studies on treating cancers of the brain, including neuroblastoma and glioblastoma.

Utilization of gene therapy by direct viral injection for nervous system disorders has the potential to revolutionize medicine in the near future. Indeed, there have been striking preclinical results demonstrating that viral vectors may be the only feasible therapeutic modality to target a specific region of the brain, or to deliver a therapy that can get past the blood–brain barrier. Many of these preclinical studies warrant clinical development. As additional genetic understanding for brain diseases develops, the ability to utilize the optimal vectors with the right gene, by the optimal injection parameter, will ultimately result in therapeutic benefit.

Truly one can see the challenges and struggles in the field of gene delivery as similar to those faced in the dramatic development of open heart surgery, when the early studies resulted in scientific challenges that were far beyond those ever imagined, including the loss of the earliest treated patients. The field of gene therapy, in particular as it relates to neuroscience, has recruited a diversity of specialists, ranging from molecular biologists to virologists, anatomists, physicians, and biotechnologists, to work together to revolutionize healthcare and ultimately to unburden patients from the effects of disease and injury of the nervous system. As clinical development of gene therapy ensues, the field will take the lessons learned and apply them to evolve genetic medicines to use in standard practice.

See also: Cell Replacement Therapy: Parkinson's Disease; Cell Replacement Therapy: Mechanisms of Functional Recovery; Cell Replacement Therapy for Huntington's Disease; Enteric Nervous System: Neurotrophic Factors; Gene Therapy: Genetically Modified Cells; GFL Neurotrophic Factors: Physiology and Pharmacology; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Viral Vectors in the CNS.

Further Reading

Bankiewicz KS, Leff SE, Nagy D, et al. (1997) Practical aspects of the development of *ex vivo* and *in vivo* gene therapy for Parkinson's disease. *Experimental Neurology* 144: 147–156.

Bohn MC (1999) A commentary on glial cell line-derived neurotrophic factor (GDNF). From a glial secreted molecule to gene therapy. *Biochemical Pharmacology* 57: 135–142.

Bruijn LI and Cudkowicz M (2006) Therapeutic targets for amyotrophic lateral sclerosis: Current treatments and prospects for more effective therapies. *Expert Review of Neurotherapeutics* 6: 417–428.

Burger C, Gorbatyuk OS, Velardo MJ, et al. (2004) Recombinant AAV viral vectors pseudotyped with viral capsids from serotypes 1, 2, and 5 display differential efficiency and cell tropism after delivery to different regions of the central nervous system. *Molecular Therapy* 10: 302–317.

Carlsson T, Bjorklund T, and Kirik D (2007) Restoration of the striatal dopamine synthesis for Parkinson's disease: Viral vector-mediated enzyme replacement strategy. *Current Gene Therapy* 7: 109–120.

Denovan-Wright EM and Davidson BL (2006) RNAi: A potential therapy for the dominantly inherited nucleotide repeat diseases. *Gene Therapy* 13: 525–531.

Federici T and Boulis NM (2006) Gene-based treatment of motor neuron diseases. *Muscle & Nerve* 33: 302–323.

Jakobsson J and Lundberg C (2006) Lentiviral vectors for use in the central nervous system. *Molecular Therapy* 13: 484–493.

Kaplitt MG, Feigin A, Tang C, et al. (2007) Safety and tolerability of gene therapy with an adeno-associated virus (AAV) borne GAD gene for Parkinson's disease: An open label, phase I trial. *Lancet* 369: 2097–2105.

Kaspar BK, Llado J, Sherkat N, et al. (2003) Retrograde viral delivery of IGF-1 prolongs survival in a mouse ALS model. *Science* 301: 839–842.

Naldini L, Blomer U, Gallay P, et al. (1996) *In vivo* gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* 272: 263–267.

Ryan DA and Federoff HJ (2007) Translational considerations for CNS gene therapy. *Expert Opinion on Biological Therapy* 7: 305–318.

Samulski RJ (2003) AAV vectors, the future workhorse of human gene therapy. *Ernst Schering Research Foundation Workshop* 43: 25–40.

Sands MS and Davidson BL (2006) Gene therapy for lysosomal storage diseases. *Molecular Therapy* 13: 839–849.

St George JA (2003) Gene therapy progress and prospects: Adenoviral vectors. *Gene Therapy* 10: 1135–1141.

Suhr ST and Gage FH (1993) Gene therapy for neurologic disease. *Archives of Neurology* 50: 1252–1268.

Suhr ST and Gage FH (1999) Gene therapy in the central nervous system: The use of recombinant retroviruses. *Archives of Neurology* 56: 287–292.

Tashiro A, Zhao C, and Gage FH (2006) Retrovirus-mediated single-cell gene knockout technique in adult newborn neurons *in vivo*. *Nature Protocols* 1: 3049–3055.

Tyler CM, Wuertzer CA, Bowers WJ, et al. (2006) HSV amplicons: Neuro applications. *Current Gene Therapy* 6: 337–350.

Wu Z, Asokan A, and Samulski RJ (2006) Adeno-associated virus serotypes: Vector toolkit for human gene therapy. *Molecular Therapy* 14: 316–327.

Gene Therapy: Genetically Modified Cells

U J Kang, The University of Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Gene therapy has been introduced as a potential therapeutic modality that alters the genetic machinery of the cells and tissues to a desired end. The host or implantable cells are modified with externally provided genetic information so that they produce therapeutic molecules by utilizing endogenous mechanisms to transcribe the message and translate into protein products. Since the introduction of this concept for central nervous system (CNS) disorders in the late 1980s, significant progress has been made in understanding its potential and in the development of new techniques for genetic manipulation of various cell types in *ex vivo* and *in vivo* settings. This article reviews the methodological issues and examples of CNS application that have been studied.

Gene Therapy Methods to Produce Genetically Modified Cells

Two general approaches have been taken to introduce genes of interest and artificially produce a desired product in a therapeutically relevant part of the CNS: *ex vivo* and *in vivo* gene therapy. *Ex vivo* gene therapy involves genetically modifying cells *in vitro* to express relevant gene products in the cells and then delivering the modified cells to target areas *in vivo*. The cells are sometimes taken from the host and implanted back to the host after genetic modification. The *ex vivo* approach was developed first since early gene therapy vehicles such as retrovirus vectors required actively dividing cells for successful and stable integration of the genetic materials into the chromosomes and they were not efficient in transducing cells *in vivo*. *In vivo* gene therapy is discussed in a separate article; this article focuses on *ex vivo* methods utilizing genetically modified cells.

Cell Types Used for *Ex Vivo* Gene Therapy

The cell types that are genetically modified play a critical role in *ex vivo* gene therapy. Properties of the cells that are implanted can determine a significant part of the therapeutic effect of genetically modified cells. To avoid the detrimental effects of cell delivery, the potential of the genetically modified cells to form tumors or produce immunological reactions by the host must first be considered. Placing the cells within encapsulation polymers can also provide a physical barrier restricting further growth and limiting immunological reactions. Often, primary cell types such as fibroblasts, myoblasts, astrocytes, Schwann cells, and olfactory ensheathing glia have been used as cellular vehicles for these reasons. Particularly attractive cell types for genetic modification are various stem cells of both embryonic and adult origin, which may also provide a substrate for neurite outgrowth and have a tendency to migrate into the pathological areas such as tumor cells or sites of injury or degeneration.

Vectors Used for *Ex Vivo* Gene Therapy

Viral vectors are currently the most preferred method of gene transduction for gene therapy because of their high transfection efficiency for most cell types used for *ex vivo* gene therapy. Retroviruses derived from the Moloney murine leukemia virus transduce cells with high efficiency by integrating into the cellular genome of dividing cells. However, integration is apparently random and could disrupt important normal cellular function by insertional mutagenesis including oncogenesis. Lentiviruses are retroviruses derived from human immunodeficiency viruses and have the ability to transduce nondividing neurons *in vivo* with great efficiency. They have also been successfully used for transducing cells for *ex vivo* gene therapy. Adenovirus is a double-stranded DNA virus associated with human respiratory infections, conjunctivitis, and gastroenteritis. Adenovirus has been shown to infect both neurons and glia after direct intracerebral injection and is very efficient in transducing genes into primary neurons in culture.

The main issue in using viral vectors is whether expression levels can be sustained after many passages and after implantation of the genetically modified cells into the host animals. The long-term expression depends on the type of viral vectors, promoters, cells, and transgenes. Viral promoters have initial high levels of expression but tend to turn off their expression as the cells become quiescent in culture and after grafting. Constitutive promoters have been used and they show sustained expression, but at a very low level compared to viral promoters. For some transgenes such as tyrosine hydroxylase (TH), short-lived *in vivo* expression has become a major limitation of *ex vivo* gene therapy. However, other genes, such as choline acetyltransferase (ChAT) and nerve growth factor (NGF), have been shown to maintain their expression for as long as 1 year. Another issue is the ability to regulate the expression of the transgenes so that appropriate quantities of products

are delivered to exert intended biological effects. Various novel promoter constructs that can be regulated by external agents are being developed and tested.

Summary of *Ex Vivo* Gene Therapy Methods

The advantage of *ex vivo* gene therapy is the ability to characterize the genetically modified cells extensively and select them for desired phenotypes before using them for therapy. The property of the cells may also be chosen to enhance the therapeutic potential. They may produce trophic factors endogenously, provide substrates that are conducive for neurite growth, and be capable of migrating within the target brain area. Confining the genetic modification within the implanted cells may provide a measure of safety against unexpected detrimental effects that cannot be controlled in the host cells after viral vector-mediated direct *in vivo* gene therapy. Incorporation of a suicidal gene that could be controlled in genetically modified cells can be envisioned.

The disadvantages of *ex vivo* therapy include the extraneous graft mass of many cell types such as fibroblasts and astrocytes in the CNS that could be disruptive of normal host function. Another potential limitation of *ex vivo* therapy is that the genetic modification is most appropriate for secreted molecules that can act extracellularly and not for therapeutic molecules that need to gain entry into the target cells to intervene in the intracellular processes, which is better suited for *in vivo* gene therapy.

Application of Genetically Modified Cells

Genetically modified cells can serve as a biological minipump that delivers therapeutic molecules such as neurotransmitters or neurotrophic factors locally and in a continuous manner. Neurotransmitter replacement is a mainstay of neurological therapy in disorders such as Parkinson's disease. However, selective delivery into the target site in a physiological manner is not possible by pharmacological means. Therefore, a biological pump that can provide therapeutic products in a steady fashion appropriately targeted to selected brain areas has the potential for increased therapeutic efficacy with reduced side effects.

Neurotrophic factors are critical in the growth and survival of developing neurons. In addition, they can enhance certain neuronal phenotypes and promote the survival and regeneration of damaged adult neurons. For these reasons, neurotrophic factors have tremendous therapeutic potential for a wide range of neurological diseases. However, delivering appropriate concentrations of neurotrophic factors to target cells is not easily achieved by systemic administration. Neurotrophic factors generally have short half-lives, do not cross the blood–brain barrier, and have a wide range of effects on various targets. Intraventricular delivery could overcome the blood–brain barrier but lacks anatomic and cellular specificity. Intraparenchymal delivery is limited by short diffusion distance from the delivery site. Therefore, gene therapy represents one of the most promising means of providing continuous delivery of neurotrophic factors to specific anatomic and cellular targets.

The neurotrophin family, including NGF, brain-derived neurotrophic factor (BDNF), neurotrophin (NT)-3, and NT-4/5, has been extensively studied for its therapeutic potential, including use in gene therapy. Glial cell line-derived neurotrophic factor (GDNF) is a distant member of the transforming growth factor-β superfamily of trophic factors that also includes neurturin, artemin (neublastin), and persephin (Table 1).

Parkinson's Disease

Parkinson's disease (PD) is the second most common neurodegenerative disorder. PD is well suited for gene therapy to deliver neurotransmitters or neurotrophic factors because it is biochemically and pathologically well defined. The degeneration of the nigrostriatal dopaminergic system and ensuing dopamine loss underlie the majority of motor disabilities of PD. The clinical efficacy of dopamine replacement therapy in improving major aspects of motor

Table 1 Target genes delivered by genetically modified cells in animal models of neurological disorders

Disorder	*Neurotransmitter replacement*	*Neurotrophic factor delivery*
Parkinson's disease	TH, GTPCH1, AADC, VMAT	BDNF, GDNF, neurturin, persephin
Alzheimer's disease	ChAT	NGF
Huntington's disease		NGF, BDNF, NT-3, NT-4, GDNF, neurturin, CNTF
Amyotrophic lateral sclerosis		GDNF
Spinal cord injury		GDNF, NGF, BDNF, NT-3, NT-4/5

AADC, aromatic L-amino acid decarboxylase; BDNF, brain-derived neurotrophic factor; ChAT, choline acetyltransferase; CNTF, ciliary neurotrophic factor; GDNF, glial cell line-derived neurotrophic factor; GTPCH1, GTP cyclohydrolase 1; NGF, nerve growth factor; NT, neurotrophin; TH, tyrosine hydroxylase; VMAT, vesicular monoamine transporter.

disability has been well established. On the other hand, pharmacological therapy with dopamine precursor, L-3,4-dihydroxyphenylalanine (L-DOPA), or synthetic dopamine agonists is limited by development of motor response fluctuation and other side effects that develop over time. The emergence of motor response complications has been attributed to the supersensitivity of the dopamine signaling pathway and further maladaptive plasticity that result from nonphysiological chronic drug therapy. It has been hypothesized that steady levels of L-DOPA or dopamine may be more physiological than fluctuating levels that are typically achieved by intermittent oral administration of dopaminergic agents. Experimental studies have shown that maintaining steady levels of L-DOPA achieved by either intravenous or intraintestinal delivery reduces the likelihood of motor response complications. In addition, it is hypothesized that localized delivery would avoid cognitive side effects such as hallucinations that are attributed to excessive stimulation of dopamine system outside the pathogenic nigrostriatal system by systemic delivery of dopaminergic agents. Therefore, methods to deliver L-DOPA or dopamine continuously and locally have been sought. Transplantation of various cells that provide L-DOPA or dopamine has been attempted in patients and experimental animals. Fetal dopamine cells have been considered the gold standard for such vehicles. Stem cells and gene therapy have been proposed as alternative means to provide well-defined and readily available sources of localized and sustained dopamine.

Neurotransmitter replacement To deliver dopamine by genetically modified cells, one has to consider many genes depending on the cell types to be used as a vehicle. Some cell types, such as stem cells, can be coaxed into expressing some such phenotypes, but for most implantable cell types, artificial expression of transgenes is necessary or can supplement the phenotype. Dopamine synthesis is a two-step enzymatic process. The essential amino acid tyrosine is first converted into L-DOPA by the rate-limiting enzyme TH. TH requires cofactors such as iron, oxygen, and tetrahydrobiopterin (BH_4) for its activity. Although most of the initial studies employed only TH, the most efficient production of L-DOPA is achieved by adding the GTP cyclohydrolase 1 gene along with the TH gene. GTP cyclohydrolase 1 is the first and rate-limiting step of BH_4 synthesis. Both *ex vivo* gene therapies utilizing various cell types and *in vivo* gene therapies introducing genes directly into striatal cells have shown the necessity and effectiveness of using both TH and GTP cyclohydrolase 1 genes for significant production of L-DOPA and subsequent behavioral restoration in PD models of rats, mice, and nonhuman primates. The critical role of GTP cyclohydrolase 1 in dopamine synthesis is also demonstrated by hereditary progressive dystonia and parkinsonism (also termed DOPA-responsive dystonia), which results from the deficiency in GTP cyclohydrolase 1.

The second step in dopamine synthesis involves decarboxylation of L-DOPA to dopamine by aromatic L-amino acid decarboxylase (AADC). *Ex vivo* gene therapy experiments in which AADC is added to TH and GTPCH1 have provided interesting insight about dopamine regulation. When fibroblast grafts expressing three genes, TH and GTPCH1 to produce L-DOPA and AADC to convert L-DOPA to dopamine, were compared to those expressing only TH and AADC, both L-DOPA and dopamine levels were higher in those with two genes than those with three genes. The detrimental effect of AADC on synthesis of L-DOPA and dopamine in genetically modified fibroblast cells both in culture and *in vivo* was attributed to the end product – dopamine inhibition of TH activity. End-product inhibition is a well-established regulatory mechanism of TH, and dopaminergic neurons sequester dopamine into synaptic vesicles and thus accumulate high levels of dopamine in the cytoplasm that would inhibit TH. Therefore, AADC is marginally helpful given its endogenous expression in nondopaminergic cells and may even be detrimental depending on the parameters of its delivery.

Engineering cells with the vesicular monoamine transporter (VMAT) gene to increase storage capacity circumvents the feedback inhibition. Although typical synaptic vesicles are not present in most cell types used for *ex vivo* gene therapy, experiments on fibroblast cells have shown that even nonneuronal cells can incorporate VMAT into tubulovesicular structures and store dopamine at higher levels and release them over longer periods of time than cells without VMAT expression. *In vivo* gene therapy study also showed that engineering all four genes – TH, GTPCH1, AADC, and VMAT – produced higher levels of dopamine and better behavioral recovery than three gene constructs consisting of TH, GTPCH1, and AADC in a rat model of unilateral PD.

An alternative approach to providing either L-DOPA or dopamine by gene therapy is to graft cells that express AADC without TH or GTPCH1 to enhance the efficiency of converting systemically administered L-DOPA to dopamine. If one could achieve the same level of striatal dopamine using a lower dose of L-DOPA compared to those without grafts, this would decrease the diffusion of L-DOPA to other parts of the brain, notably the limbic system, and minimize side effects such as hallucinations and

confusion. Both *ex vivo* and *in vivo* gene transfers of AADC genes in the striatum have been employed to demonstrate enhanced dopamine production in the striatum in rodents and nonhuman primates. This approach also has the advantage of being able to regulate dopamine production by varying the amount of precursor, L-DOPA. A phase I clinical trial using adeno-associated virus (AAV)-mediated AADC gene therapy in PD patients is in progress. To further enhance the efficacy of dopamine conversion from L-DOPA, genetically modified fibroblasts containing both AADC and VMAT genes have been used for gene therapy. The addition of VMAT not only enhances dopamine levels but also prolongs the duration of dopamine release compared to AADC alone.

A major technical issue with the dopamine gene approach is the safety of regulating the level of dopamine replacement. The most feasible approach is to use regulatory promoters to control either TH or GTPCH1 expression or use a combination of precursor, L-DOPA regulation, and gene therapy with-AADC to enhance the production with or without VMAT to store dopamine. In the meantime, the rationale for central and continuous delivery of dopamine is being re-examined. First, limitations of treating only dopamine-associated symptoms are recognized as more widespread system degeneration in PD is being appreciated. Second, current gene therapy or cell replacement therapies may function mainly as a tool for tonic replacement of dopamine, whereas phasic activity of dopamine seems to be critical in many physiological functions. Lastly, the use of continuous dopamine stimulation to reverse fundamental mechanisms of motor response complication is largely untested beyond its pharmacokinetic effect. Of note is that most dopamine-related genes seem to be relatively unstable, achieving relatively short-term expression by genetically modified cells, and most long-term expression has been achieved by direct *in vivo* viral delivery.

Neurotrophic factor delivery A complementary approach to treating PD is neurorestorative therapy using genes that could prevent degeneration or promote dopaminergic phenotypes and survival. To do so effectively would require a good understanding of the etiology of PD, which remains largely unknown. Nonetheless, mitochondrial dysfunction, oxidative stress, and abnormal protein processing have been implicated in the pathogenesis. In addition, several approaches seem to be restorative to dopaminergic neurons, although these approaches may not be based on the precise etiology or pathogenesis of PD. For example, although it is not well understood whether neurotrophic factor deficiency plays a major role in the pathogenesis of PD, neurotrophic factors can still augment the function and retard degeneration of dopaminergic neurons in PD models.

Although many factors have been shown to promote the survival of dopaminergic neurons in experimental culture models, only a few have been effective *in vivo*. BDNF has been delivered by a variety of *ex vivo* gene therapy paradigms. Implants of BDNF-producing fibroblast cells protect against the toxicities of 6-hydroxydopamine (6-OHDA) and 1-methyl-4-phenylpyridinium. In addition, implantation of BDNF-expressing fibroblasts into the substantia nigra induced sprouting of dopaminergic fibers into grafts.

GDNF delivered to the substantia nigra by *ex vivo* gene therapy using neural stem cells and encapsulated cells has been found to protect dopaminergic neurons and terminals from 6-OHDA toxicity in rodents and 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine in primates. Fibroblasts expressing neurturin and neuronal stem cells expressing persephin protected against 6-OHDA neurotoxicity. The timing and the location of GDNF gene therapy are important variables in experimental models of PD usually produced by dopaminergic toxins. Therefore, one would need to extrapolate the efficacy in these acute toxic lesion studies to the chronic neurodegenerative process of PD. The effect of *in vivo* viral gene therapy with GDNF in boosting the dopaminergic phenotype in aged monkeys demonstrates the feasibility of this approach in chronic human diseases such as PD.

Intraventricular and intraparenchymal deliveries of GDNF have been attempted in PD patients without significant benefit, most likely due to inefficient delivery to target areas. To overcome the delivery issue, *in vivo* AAV-mediated gene therapy of neurturin is in phase I clinical trials in PD patients.

Alzheimer's Disease

Alzheimer's disease (AD) is characterized by generalized cortical atrophy with characteristic pathology of neuritic plaques and neurofibrillary tangles. Although the degeneration is widespread, reductions in acetylcholine are most profound and associated with the loss of cholinergic neurons in the basal forebrain. Cholinergic deficits are thought to be responsible for memory loss and attention problems, but neurotransmitter replacement therapy using centrally acting acetylcholinesterase inhibitors has been only modestly effective in improving cognitive dysfunction.

Neurotransmitter replacement Acetylcholine is synthesized by ChAT from acetyl coenzyme A and choline. Once released from the nerve terminals, acetylcholine is rapidly metabolized by acetylcholinesterase and choline is taken backup to the terminal

by the choline transporter. Gene therapy targeted to replace ChAT by genetically modified fibroblasts cells has been tested in two different rat models of AD: the ibotenic acid lesion of the nucleus basalis of Meynert in the basal forebrain and the fimbria-fornix lesion that axotomizes the basal forebrain neurons. The cells were shown to express the transgenes at least for 1 year. Increased ChAT activity and acetylcholine levels were noted in these genetically modified cells grafted in the frontal and parietal cortex, and this increase was associated with improvement in their performance in the water maze task. The behavioral effects of the genetically modified cells were comparable to those of pharmacological treatment with acetylcholinesterase inhibitors or hippocampal grafts that contain cholinergic neurons. In addition, aging rats with cognitive deficits improved their water maze spatial memory performance when grafted with acetylcholine-producing fibroblast cells in cortex and hippocampus.

Neurotrophic factor delivery The effect of NGF on basal forebrain cholinergic neurons has been well demonstrated in rodent and primate animal models. Rat fibroblast cells that were genetically modified to express NGF protected against cholinergic degeneration following fimbria-fornix lesions in rats. The protective effects of NGF have been replicated in many *ex vivo* and *in vivo* gene therapy experiments in rodents and primates of AD models produced by lesions of basal forebrain as well as against age-related atrophy in rats and monkeys. The cells that have been modified include immortalized embryonic skeletal muscle cells, NT2N cells, and encapsulated Schwannoma cells. Intraventricular delivery of NGF was attempted in several AD patients but was limited by side effects such as weight loss and back pain. A phase I gene therapy study was conducted in which NGF-expressing fibroblasts were implanted near the nucleus basalis of Meynert in the brains of a small number of patients with early stage AD. After a mean follow-up of 22 months, no adverse effects were noted. Clinical assessment showed attenuation in the rate of cognitive decline, positron emission tomography scans showed an increase in cortical glucose uptake, and one autopsy showed robust sprouting of cholinergic axons into the NGF delivery site. Further studies including control groups will be critical for proper assessment of this approach, but these initial findings are promising and consistent with experimental data in animals.

Other Disease Models

For other diseases, neurotransmitter replacement strategy by genetically modified cells has not been explored much due to the fact that there is little clinical evidence for the efficacy of such an approach. Genetically modified cells that produce neurotrophic factors have been tested in animal models of several other neurodegenerative disorders.

Huntington's disease (HD) is an autosomal dominant disease characterized by abnormal involuntary movements, cognitive deficits, and psychiatric manifestations. The expansion of trinucleotide CAG repeat sequences results in abnormally long polyglutamine tracts in the huntingtin protein and degeneration of medium spiny neurons in the striatum and cortical neurons. BDNF levels are decreased by 45% in the frontoparietal cortex of HD brains. BDNF delivered by *ex vivo* gene therapy using fibroblasts has been shown to protect striatal projection neurons against quinolinic acid lesions. GDNF-expressing fibroblast cell lines protected GABA/substance P-positive neurons in the direct pathway but did not protect striatal projection neurons in the indirect pathway following quinolinic acid lesions. Interestingly, neurturin-expressing cell lines have the opposite effect, protecting striatal neurons in the indirect but not direct pathway. Polymer-encapsulated baby hamster kidney fibroblast cells expressing CNTF protected striatal neurons in rats against the toxicity of intrastriatal quinolinic acid injections and against the toxicity of 3-nitropropionic acid in monkeys. Few clinical trials using neurotrophic factors have been performed in HD.

Amyotrophic lateral sclerosis is a devastating disease of unknown etiology involving the progressive degeneration of motor neurons. Most applications of gene therapy using neurotrophic factors have utilized direct viral-mediated *in vivo* gene therapy.

Spinal cord injury is another major therapeutic target of genetically modified cells that express neurotrophic factors. A variety of growth factors have been studied in animal models of spinal cord injury, including GDNF, NT-3, NT-4/5, BDNF, and NGF, each with its distinct range of target neuronal populations. For instance, NGF-expressing fibroblasts induced growth of sensory axons and noradrenergic axons into the graft, whereas fibroblasts genetically modified to express NT-3 resulted in growth of corticospinal axons through host gray matter and distal to lesion sites following acute dorsal spinal cord hemisection and produced partial functional recovery. BDNF-expressing Schwann cells induced growth of coerulospinal and raphespinal axons, and BDNF-expressing fibroblasts into a partial cervical hemisection model resulted in partial regeneration of the rubrospinal tract axons and functional recovery and partially rescued axotomized red nucleus neurons from loss and atrophy. Fibroblasts expressing GDNF also promoted significant regeneration of dorsal

column sensory, propriospinal, and local motor axons when grafted into transected spinal rat spinal cord. Schwann cells expressing GDNF also promoted propriospinal axonal regeneration and myelination. Despite some success in the initiation of axonal growth, stimulating axons to grow beyond neurotrophic factor-secreting grafts has proved to be much more difficult since they are trapped at the graft with a high concentration of neurotrophic factors. It may be necessary to employ novel strategies such as turning off growth factor expression at the injury site and initiating expression at more distal sites or producing a gradient of neurotrophic factors.

Summary

Gene therapy using genetically modified cells provides an opportunity to understand the biology of disease by examining the effect of one gene at a time on restoration of functions of the diseased system. Genetically modified cells have the potential to effectively deliver neurotransmitters, neurotrophic factors, or a combination of factors in a wide range of neurological diseases. Some of these experiments have led to early phase clinical trials, such as *in vivo* viral gene therapy of aromatic L-amino acid decarboxylase, glutamic acid decarboxylase, or neurturin for PD and *ex vivo* therapy of NGF for AD. One must consider the specific advantages offered by gene therapy and balance them against the limits of current therapeutic modalities.

See also: Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Cell Replacement Therapy: Parkinson's Disease; Cell Replacement Therapy: Mechanisms of Functional Recovery; Cell Replacement Therapy for Huntington's Disease; Gene Therapy: Direct Viral Delivery; Neurotrophic Factor Therapy: GDNF and CNTF.

Further Reading

Akerud P, Canals JM, Snyder EY, and Arenas E (2001) Neuroprotection through delivery of glial cell line-derived neurotrophic factor by neural stem cells in a mouse model of Parkinson's disease. *Journal of Neuroscience* 21: 8108–8118.

Bencsics C, Wachtel SR, Milstien S, Hatakeyama K, Becker JB, and Kang UJ (1996) Double transduction with GTP cyclohydrolase I and tyrosine hydroxylase is necessary for spontaneous synthesis of L-DOPA by primary fibroblasts. *Journal of Neuroscience* 16: 4449–4456.

Blesch A, Paul Lu P, and Tuszynski MH (2002) Neurotrophic factors, gene therapy, and neural stem cells for spinal cord repair. *Brain Research Bulletin* 57: 833–838.

Kang UJ, Lee WY, and Chang JW (2001) Gene therapy for Parkinson's disease: Determining the genes necessary for optimal dopamine replacement in rat models. *Human Cell* 14: 39–48.

Nakamura K and Kang UJ (2005) Neurotrophic factor delivery by gene therapy. In: Selzer ME, Cohen L, Gage FH, Clarke S, and Duncan PW (eds.) *Textbook of Neural Repair and Neurorehabilitation*, pp. 532–547. Cambridge, UK: Cambridge University Press.

Rosenberg MB, Friedmann T, Robertson RC, et al. (1988) Grafting genetically modified cells to the damaged brain: Restorative effects of NGF expression. *Science* 242: 1575–1578.

Tuszynski MH, Thal L, Pay M, et al. (2005) A phase 1 clinical trial of nerve growth factor gene therapy for Alzheimer disease. *Nature Medicine* 11: 551–555.

Winkler J, Suhr ST, Gage FH, Thal LJ, and Fisher LJ (1995) Essential role of neocortical acetylcholine in spatial memory. *Nature* 375(6531): 484–487.

Genetic Influence on CNS Gene Expression: Impact on Behavior

I Hovatta, National Public Health Institute, Helsinki, Finland; Institut National de la Santé et de la Recherche Médicale, Créteil, France; Biomedicum Helsinki, and University of Helsinki, Helsinki, Finland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Advanced genomics technologies allow neuroscientists to tackle more complex questions than ever before. Most of the new genomewide analysis methods concentrate on evaluating global gene expression or protein abundance differences. However, RNA-based methods are currently more popular because of their relative ease and the methodological difficulties associated with proteomic approaches. DNA microarrays enable the detection of gene expression profiles from small quantities of brain tissue or even from single cells. This ability is essential for most neurobiological applications because the brain is extremely complex tissue in which neighboring cells might have very different gene expression patterns.

Microarrays offer a tool for unbiased genomewide gene expression analysis. They work by hybridization of labeled RNA or DNA to DNA molecules attached at specific locations on a surface (Figure 1). The concept of DNA microarrays is more than 15 years old, and several different techniques emerged in the mid-1990s. Current arrays offer high sensitivity, reproducibility, and a large number of target genes. A wide variety of different arrays is commercially available. Therefore, the initial problems related to microarray technology have been largely overcome, although reliable identification of small differences in expression level between two samples (<20%) remains challenging. With the advancement of the technology, the need to independently replicate array-based results is disappearing as microarrays themselves are in many cases more sensitive than methods used to validate microarray results, such as quantitative real-time PCR or Northern blots. Several excellent reviews and book chapters discuss the technical aspects of gene expression profiling using microarrays.

Multifactorial or complex diseases are likely to be caused by some combination of multiple genetic and environmental factors. While a great number of genes that cause monogenic disorders are known, success in mapping susceptibility genes for multifactorial diseases has been limited because of several confounding factors, such as genetic heterogeneity, unknown mode of inheritance, classification of the phenotype, phenocopies, reduced penetrance, and high frequency of disease-predisposing alleles. The initial disappointments in gene mapping of complex diseases have led geneticists to use new methods, such as microarrays, in order to find genes predisposing to complex diseases. Moreover, it has been hoped that microarrays would provide information about pathogenetic mechanisms of both Mendelian and complex diseases. In the past 5 years, a large number of studies have been published using human brain samples and microarrays, and some of this work is discussed below. There is also an ongoing search for relevant animal models for complex behavioral traits. The major advantage of working with animals is that the environment can be entirely controlled, and one can concentrate on phenotypic differences resulting from genetic differences. Gene expression profiling has provided valuable information about disease mechanisms in various animal models, such as transgenic mouse lines or pharmacological models.

This article reviews some recent research aimed at the identification of genes with altered expression in human behavioral disorders. It also emphasizes the use of animal models in conjunction with new genomic methodology in trying to understand complex neurobiological diseases and behavior.

Gene Expression Profiling of Human Brain Tissue

Gene expression profiling has contributed to our understanding of the pathophysiology of many diseases, such as cancer. The progress in brain-related diseases has not been as rapid. In part, this is due to the poor availability of good-quality human brain samples. Although fresh tissue is always preferred, it is rarely available, with the exception of central nervous system tumors. They have been classified according to their cell type of origin as it is possible to obtain good-quality brain tissue from patients who undergo surgical treatment for brain tumors. However, for many other brain-related diseases, it is very difficult, and this has slowed the use of microarray-based technologies in the study of neurological and mental illnesses. Recently established human brain collections, such as the Stanley Foundation Brain Collection, which have collected and matched brains especially for gene expression profiling experiments, have already proven valuable. It seems that agonal status and the freezing of the tissue impact mRNA integrity more than postmortem interval does, and the mRNA in most regions of the brain appears to be surprisingly well preserved, at least up to 48 h postmortem. An essential requirement in

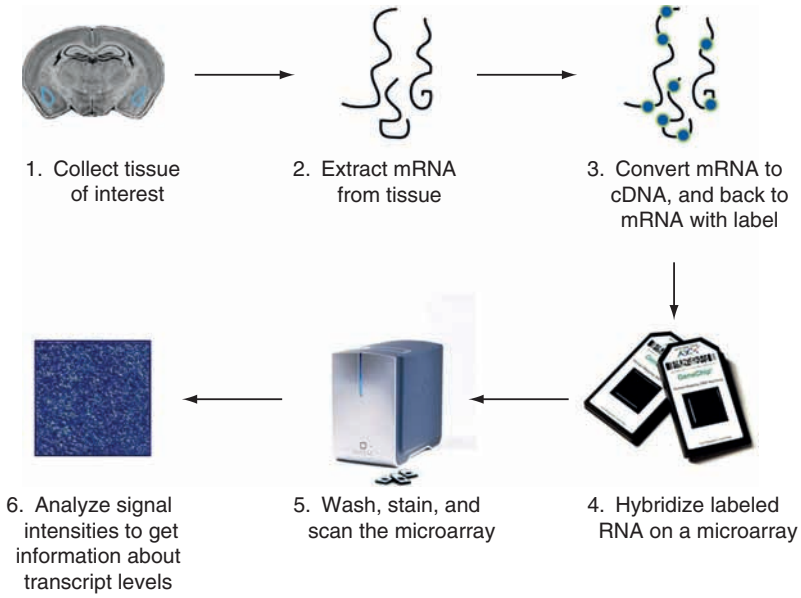

Figure 1 An overview of sample processing for microarrays. (1) First, a brain region or cells of interest are dissected. (2) Messenger RNA (mRNA) is extracted from the tissue/cells. (3) mRNA is converted to complementary DNA (cDNA) by a reverse transcriptase, and the resulting cDNA is converted back to RNA and labeled simultaneously. (4) Labeled RNA is then hybridized on a microarray, usually overnight. (5) Unbound probe is washed, and further staining is performed, depending on the type of microarray used. The signal intensities of bound probe are then read in a scanner. (6) The signal intensities of bound probe are then read in a scanner. A higher signal intensity generally means a larger number of the corresponding transcript in the original sample. Careful quality control assessment and statistical analysis are needed in order to determine the expression level of transcripts on the microarray. Pictures in 4–6 from www.affymetrix.com

studies using human brain samples is that a sufficient number of experiments be performed across multiple individuals and multiple tissue samples to account for individual variation due to age, sex, disease state, medication, and possible tissue inhomogeneity. This has been the major problem of most microarray studies of human brain tissue, and therefore interpretation of the results between different studies has been difficult. As studies of cancer have shown, it might be possible to use microarray profiling and cluster analysis to better define the subtypes of a given disease, identify target genes and pathways, predict optimal therapeutic interventions, and assess pharmacodynamic response. In the past few years, a growing number of studies have shed light on gene expression perturbations in various neurobiological diseases. The next sections provide an overview of some studies aiming to resolve the biology behind these complex brain diseases.

Recent Advances in Gene Expression Profiling of Behavioral Diseases

Alcoholism

Alcohol dependence is a complex disorder with an estimated lifetime prevalence of 10–15%. Genetic factors involved in alcoholism are largely unknown, but linkage and association studies have suggested the involvement of several genes in the etiology of alcoholism, such as alcohol dehydrogenases; cytochrome P-450, family 2, subfamily E, polypeptide 1; γ-aminobutyric acid receptor subunits; dopamine receptors; and serotonin receptors.

Several recent studies have investigated gene expression differences in various brain regions between individuals with alcoholism and controls. The most studied tissue has been the prefrontal cortex due to the most severe brain damage in this region in patients with alcoholism. Other profiled tissues include brain regions involved in reward and dependence, such as other mesolimbic brain structures. These studies have consistently found altered expression of genes involved in myelination, trafficking, ubiquitination, mitochondrial function, cell adhesion, and apoptosis. One recent study of 14 people who were alcoholic and 13 controls suggested that relatively small expression changes in the frontal cortex of brains of alcoholics are quite consistent across individuals and likely represent important alcohol-responsive genes. It is interesting that some of the differentially expressed genes in alcoholism studies are the same as those identified in Alzheimer's disease (AD), and others are the same as those identified in schizophrenia, suggesting that

these genes perhaps are involved in more-general brain malfunction.

Alzheimer's Disease

AD accounts for the majority of dementia in the older population. By means of linkage analysis in pedigrees with several affected individuals, genes predisposing to early-onset familial AD have been described: amyloid-β precursor protein (APP), presenilin 1, and presenilin 2. Also, one risk factor, the ApoE4-ε4 allele of the apolipoprotein E gene, has been described in late-onset AD. However, variation in these four genes accounts for only a fraction of the total genetic predisposition to AD.

In recent years, several studies have used microarrays to examine gene expression changes in brain aging and AD. Results from these studies, which have used both animal models and postmortem human brain tissue, vary considerably, which is not surprising given the variability in microarray platforms, statistical approaches, and tissue studied. However, these studies have identified some common pathways that might regulate brain aging and related pathological processes in both rodents and humans. There is an encouraging overlap between findings of microarray studies and pathological processes identified with more conventional neurobiological methods. Most microarray studies of brain aging have been performed in rodent models. Essentially all have found altered expression of genes related to inflammatory responses, oxidative stress, mitochondrial dysfunction, and glial processes.

The microarray studies of human postmortem AD brain tissue have compared either AD-affected brain structures across AD patients and matched controls or affected and unaffected tissue across AD patients and controls. Some studies have looked at early stages of disease, and others have concentrated on advanced stages. The results suggest downregulation of transcripts whose protein products are associated with neuronal function and upregulation of messages functioning as transcription factors and tumor suppressors.

Transgenic animal models for genes that are associated with AD show many characteristics similar to those seen in human AD patients. Microarray studies of these models (presenilin 1 and 2 and APP) have suggested changes in the expression of genes related to inflammatory response, lysosomal function, transcription machinery, synaptic or intracellular signaling, cell growth, extracellular matrix/cytoskeletal elements, energy transduction, and protein degradation, among others. It has been postulated that many of these changes relate to disorganization of the synaptic structure of extant neurons or loss of neurons. This dysfunction in the neuronal compartment might be accompanied by glial activation and inflammatory processes. Overall, microarray studies have provided new systems-level information about normal and pathological aging of the brain and have provided testable hypotheses for future functional studies.

Autism

Autism is a severe neurodevelopmental disorder with a prevalence of 10–20 per 10 000. Autism is the most severe form of a broad group of disorders known as autism spectrum disorders, which include also milder conditions such as Asperger syndrome, disintegrative disorder, and atypical forms of autism. Twin and family studies have repeatedly and unequivocally indicated a strong genetic component in autism spectrum disorders, but the underlying genetic mechanisms are still largely unknown.

Three studies investigating gene expression profiles of autism have been published to date. One of them studied mRNA extracted from cerebellum of ten patients with autism and 23 controls. Of the 30 genes differentially expressed between the patients with autism and the controls, the glutamate receptor system was targeted for further analysis. It was shown that the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid-type glutamate receptor density was decreased in the cerebellum of individuals with autism. The two other studies investigated gene expression profiles from lymphoblast cell lines. There is emerging evidence that the presence of chromosomal abnormalities may play a significant role in the etiology of autism, and chromosome 15 duplication in particular accounts for $\sim$1–3% of autism cases. Gene expression profiling in ten cell lines with the duplication and nine control cell lines revealed 112 genes that were differentially expressed between these two groups, including genes in several functional categories associated with macromolecular catabolic processes, especially the ubiquitin-dependent protein catabolism category. Another recent study of lymphoblast cell lines looked at gene expression differences between five pairs of monozygotic twins discordant or concordant for autism. Global analysis of the significantly differentially expressed genes based on functional classes identified genes involved in nervous system development and function as significantly enriched across the five gene expression data sets. It is interesting that *in silico* mapping of the most differentially expressed genes across as well as within the twin sets demonstrated that many of these genes are located in or close to chromosomal regions previously identified as autism susceptibility loci by genetic analyses. Even though the

sample sizes of these studies have been low, these initial experiments show that microarray analyses can identify interesting new candidate genes that might help investigators understand the pathogenetic mechanisms of autism.

Bipolar Disorder

Bipolar disorder (BPD) is a common psychiatric disorder of adulthood affecting about 1% of the population. It is characterized by episodes of abnormal mood (mania or hypomania and depression). Family and twin studies have demonstrated the importance of genetic factors in the etiology of BPD but have also suggested multifactorial inheritance. Several genomewide screens have been performed in families with multiple affected family members. Encouragingly, some chromosomal regions have been implicated in several studies, and specific susceptibility genes are being identified.

A growing number of microarray studies have compared postmortem brain tissue of patients with BPD and controls. The profiled regions include hippocampus, dorsolateral prefrontal cortex, and orbitofrontal cortex, all of which have been implicated in the pathogenesis of BPD. Lymphoblast cells from patients with BPD have also been profiled. Many of the studies have used samples from the Stanley Foundation Brain Collection. Comparison of results from different studies has been difficult because of slightly different diagnostic classification, statistical analysis methods with varying covariates, and different microarray platforms. However, several interesting processes that might be involved in the pathogenesis of BPD have been identified, such as mitochondrial dysfunction, regulation of oligodendrocytes and myelination, cell growth and development of the nervous system, apoptosis, the ubiquitin–proteasome pathway, regulation of synaptic function, G-protein-coupled receptor signaling, and immune response. As in the case of other diseases discussed in this article, it remains to be determined how these various processes coordinately contribute to the pathogenesis of BPD, which gene expression changes are causal, and which are reactive.

Major Depression

Major depression (MD) is one of the most common mental disorders, with a lifetime prevalence of at least 10%. Heritability based on twin studies is 40–50%, suggesting that it is a complex disorder with genetic and environmental risk factors. The potent environmental risk factors include childhood abuse and neglect and stressful life events. As in the case of other psychiatric disorders, genetic linkage and association studies have identified some susceptibility genes for MD, but replications of the results have been rare. However, one of the recent meta-analyses has shown that serotonin transporter gene polymorphism may be associated with MD. Other suggested susceptibility genes include serotonin 2A receptor, tyrosine hydroxylase, tryptophan hydroxylases 1 and 2, catechol-o-methyltransferase (COMT), and brain-derived neurotrophic factor.

A recent study investigated gene expression profiles of three brain regions, the orbital cortex, the dorsolateral prefrontal cortex, and the motor cortex in French Canadian individuals. The study included 24 people who had committed suicide, of whom 16 had a diagnosis of MD and eight did not. In addition, 12 controls were included. Twenty-six genes were differentially expressed between patients and controls, and one of these, spermine/spermidine N^1-acetyltransferase (*SSAT*), was targeted for further studies. The expression difference has been confirmed with other methods. It is interesting that an association analysis of genetic markers in the gene coding for *SSAT* revealed evidence of association in an independent cohort of males who had committed suicide. Therefore, this study suggests that SSAT, the rate-limiting enzyme in the catabolism of polyamines, might have a role in the predisposition to suicide and depression. Another recent study looked at gene expression profiles of eight patients with MD, six with BPD, and seven control individuals. The brain regions studied included anterior cingulate cortex and dorsolateral prefrontal cortex. Altered expression of a specific subset of genes encoding the glial high-affinity glutamate transporters, glutamate-ammonia ligase, and various subunits of glutamate receptors and γ-aminobutyric acid receptors was found in patients with MD compared with controls. Altered expression of fibroblast growth factor system-related genes in MD patients was also observed in this data set.

Schizophrenia

Schizophrenia is a severe and debilitating mental disorder characterized by so-called positive (psychotic) and negative (low levels of emotional arousal, mental activity, and social drive) symptoms. In general, the lifetime prevalence of schizophrenia is thought to be between 0.5% and 1.0%, but some geographical variation occurs. The evidence for a genetic predisposition to schizophrenia has been obtained from family, adoption, and twin studies. However, despite the effort of numerous groups, only a few susceptibility genes have been found by linkage analysis in families with several affected individuals. Genes that have been independently identified in several studies

include disrupted in schizophrenia 1, dystrobrevin binding protein 1, neuregulin 1, and D-amino acid oxidase activator, among others. In addition, multiple promising chromosomal loci have been found that are likely to harbor genes predisposing to schizophrenia.

With regard to gene expression profiling, schizophrenia is clearly the most extensively studied mental disorder. In the past few years, more than 30 reports have been published on gene expression profiles studied in either postmortem brain tissue or blood cells. As discussed above, there are many confounding factors that make the comparison of the different studies difficult, and the number of replicated findings is small. However, several well-replicated findings have emerged: dysregulation of genes involved in the function of mitochondria and energy metabolism, altered expression of genes that control oligodendrocyte function and myelination, and decrease of presynaptic markers. It is interesting that several genes in these pathways, not just single genes, show altered expression in schizophrenia patients. On the other hand, these changes do not seem to be specific for schizophrenia as patients with BPD or MD show similar alterations in brain gene expression patterns. It is encouraging that some genes that have been found differentially expressed between schizophrenia patients and controls, such as PDZ and LIM domain 5 and regulator of G-protein signaling 4, have also revealed evidence of genetic association.

Gene Expression Profiling of Mouse Brain Tissue

Progress in understanding the neurobiology of behavior has increased rapidly in recent years as result of several major advances, most notably advances in mouse genetics. The rapid pace at which new genes are being identified, the availability of the DNA sequence for several mouse strains, the extensive analysis of inbred strains of mice, and the use of induced mutations have created new tools for investigating the molecular and genetic basis of behavioral phenotypes. Highly parallel gene expression approaches allow one to look at the global interactions of genes and modifiers and their effects and will greatly enhance our ability to define the role of developmental alterations, mutations, and compensatory mechanisms in causing or modifying particular behaviors. Behavioral testing combined with microarray profiling has proven to be a powerful approach to identifying genes regulating mouse behaviors. Since the basic neuronal mechanisms are shared across mammalian species, the same set of genes likely regulates critical aspects of behavior in humans and mice. The advantage of studies in mouse is that when strict procedures are applied during animal and tissue handling, sample preparation, and array hybridization, overall assay variability can be minimized, and the false-positive rate can be kept extremely low. The drawback of using animals as models for human behavioral phenotypes is that it requires a well-characterized animal model for a given disease, and such models do not exist for many psychiatric diseases, such as schizophrenia. However, individual components of complex diseases can be studied separately, and in the case of schizophrenia, this has been done by studying prepulse inhibition (PPI). PPI is the suppression of the normal response to a startling stimulus when that stimulus is immediately preceded by a weak prestimulus or prepulse. PPI, measured using acoustic startle, differs between patients with schizophrenia and controls as well as among different inbred mouse strains. For some behavioral phenotypes, such as anxiety and substance abuse, animal models have been well characterized and widely used.

The advantage of working with mice is that the animals can be handled and the tissue dissections performed systematically and consistently. This consistency is extremely important in order to minimize sample-to-sample variation. Brain dissections can be performed by hand, by means of a sophisticated dissection microscope and very fine dissection tools, or by laser capture microdissection. It is important to dissect the exact anatomical structure of interest. Inclusion of too much additional material might 'dilute' a weak expression difference seen between two regions, and failure to include all small subnuclei might cause important differences to be missed. As in the case of human RNA, it is important to use stringent quality control measures at all steps of the procedure, from RNA extraction to the analysis of hybridized arrays.

Mouse Gene Expression Studies of Behavioral Phenotypes

Numerous transgenic mouse models show differences in behavioral phenotypes. Gene expression profiling of the brain tissue from these models has helped reveal molecular pathways affected by the transgene expression. However, this article concentrates on the genetic variation across a number of inbred mouse strains because it provides an excellent model to correlate various complex phenotypes with brain transcript levels in a genomewide unbiased manner. The involvement of the identified genes in behavior can then be studied further by means of various methods (Figure 2).

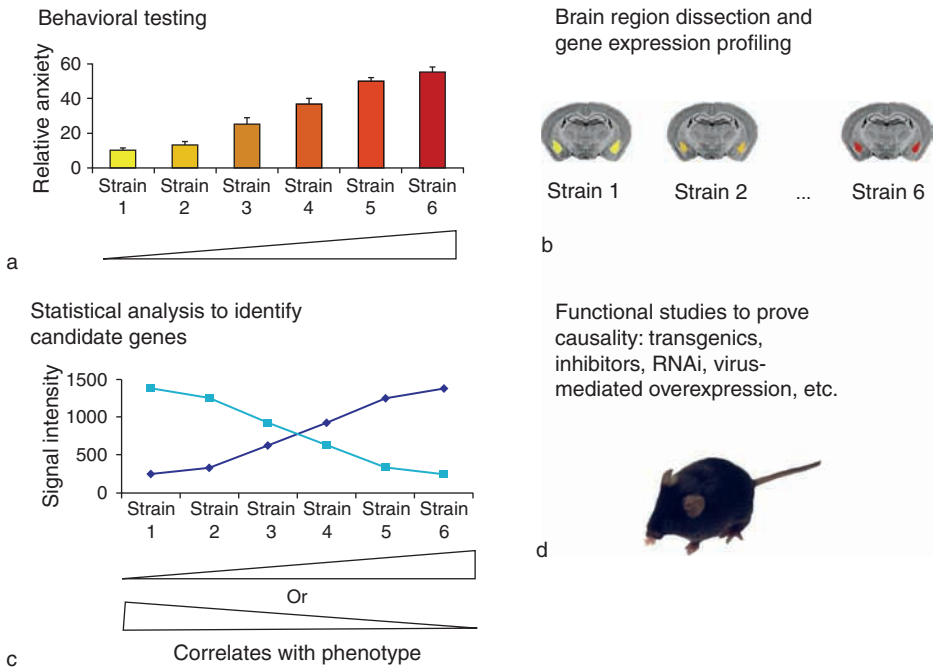

Figure 2 A gene expression-based approach to identifying genes involved in the regulation of behavior. (a) First, behavioral testing is carried out across inbred mouse strains to determine their phenotype (e.g., anxiety level). (b) Brain regions that regulate the behavior are dissected from the same strains. Messenger RNA extracted from the tissue is hybridized on a microarray to determine expression levels of all genes. (c) Statistical analyses are carried out to identify genes with an expression pattern that correlates with the behavioral phenotype across the strains. (d) The involvement of the candidate genes in behavior can be further studied by manipulating the gene expression and/or protein level(s) in mouse. (a–c) Reproduced from Hovatta I, Tennant RS, Helton R, et al. (2005) Glyoxalase 1 and glutathione reductase regulate anxiety in mice. *Nature* 438: 662–666.

Inbred mouse strains vary with regard to many behavioral phenotypes, such as anxiety, learning and memory, activity, alcohol preference, aggression, and PPI. Since animals within one inbred strain are genetically identical, interstrain differences result from genetic differences between the strains. Investigators have proposed combining behavioral testing in inbred mouse strains with gene expression profiling to find candidate genes for behavioral phenotypes. Gene expression differences that correlate with a given phenotype might be causal (i.e., directly regulate the phenotype in question) or reactive (i.e., are secondary changes due to the primary genetic defect). Causal genes can be directly tested as susceptibility genes for a given phenotype, and reactive changes may help illuminate the pathogenesis of the disease. Several groups are currently developing statistical methods to distinguish between these possibilities by combining single-nucleotide polymorphism (SNP) genotyping and gene expression profiling in populations that segregate the phenotype of interest.

In one recent study, gene expression profiles from the hippocampus of eight inbred mouse strains were compared with aggression phenotype from the Mouse Phenome Database, where information regarding many different phenotypes across several inbred mouse strains has been collected. It is interesting that the gene with an expression pattern across the strains that correlated most significantly with the aggression behavior was COMT. In humans, it has been shown that individuals with lower levels and activity of COMT are generally more aggressive than people who have an allele associated with higher activity.

Gene expression profiles of six inbred mouse strains have also been correlated with anxiety-related behavior measures from two behavioral tests, and follow-on functional studies have proven causality of a subset of identified genes. First, anxiety-related behavior of six inbred mouse strains was assessed through two pharmacologically validated behavioral tests. Concurrently, seven brain regions (amygdala, bed nucleus of the stria terminalis, cingulate cortex, hippocampus, hypothalamus, periaqueductal gray, and pituitary gland) that are known to regulate some aspects of anxiety were dissected from the same strains. The mRNA extracted from these tissues was hybridized on microarrays, and 18 genes with an expression profile that correlated significantly with the behavioral phenotype were identified. Further efforts concentrated on two of these genes: glyoxalase 1 and glutathione reductase 1, which are expressed at

a higher level in anxious mouse strains than in nonanxious ones. Both of these genes code for enzymes that are involved in the regulation of oxidative stress. Functional analyses using lentivirus-mediated gene transfer to overexpress or silence these genes in the cingulate cortex of anxious 129S6/SvEvTac mice or nonanxious C57BL/6J mice have shown that both these genes regulate anxiety-related behavior in mice. This study shows that inbred mouse strains differ markedly in behavioral phenotypes and that these differences are partially due to gene expression differences of genes regulating the phenotype. Moreover, these expression changes can be detected from appropriate brain regions by gene expression profiling.

Conclusions and Future Directions

In recent years, gene expression profiling has become one of the strongest methods for revealing the molecular basis of many neurobiological diseases. In studies with postmortem brain tissue, several new pathways have already been implicated in the pathogenesis of neurobiological diseases. It is interesting that many gene expression changes seem to be common to a group of brain diseases, while some changes are specific to individual conditions, perhaps as would be expected as some of these diseases share a subset of clinical symptoms. Many groups have also investigated gene expression profiles of blood cells from patients with various behavioral disorders. Even though only a small fraction of genes that are differentially expressed in the relevant brain region for the disease are expected to be differentially expressed in the blood, several groups have started to investigate whether blood-based gene expression profiling could be used as a diagnostic test for neurobiological disorders. It is assumed that some of the pathological changes in the brain are reflected in the blood as well and that this information could be used to classify patients into subgroups, much as in the case of various cancers.

However, simply measuring gene expression may not be enough to unravel the molecular underpinnings of a complex behavior or a disease state. Taking advantage of an increasing amount of DNA sequence information in the context of data sets on gene expression enables investigators to combine expression studies with sequence analysis and to move from phenomenology to causation. Since the sequences of the human and mouse genomes are known, it is relatively simple to look at the SNPs in the coding regions of the differentially expressed genes and identify SNPs that may either directly or indirectly lead to the expression differences. However, little is known about *cis*-acting elements outside the coding regions of the gene or *trans*-acting elements that regulate gene expression levels, and identification of such SNPs is time consuming. New approaches to solve these methodological shortcomings have been developed recently and should help in identification of susceptibility genes for complex diseases.

With large amounts of gene expression data emerging from various neurobiological studies, it is important to consider how to store and share the data. Clearly, a large database that would enable efficient sharing of data between laboratories would be desirable, but is difficult to realize in practice, as different studies have used widely varying sample collection procedures and microarray platforms, as discussed in this article. Moreover, there will be a need to incorporate data from other sources, such as SNP genotyping or proteomic or epigenetic experiments, in the same database with gene expression data. Combined analysis of genomewide data should lead to testable hypotheses about pathogenetic mechanisms of behavioral traits.

See also: Alcoholism; Alternative Splicing in the Nervous System; Alzheimer's Disease: Molecular Genetics; Anxiety Disorders; Autism; Aversive Emotions: Genetic Mechanisms of Serotonin; Bipolar Disorder; Gene Expression Regulation: Activity-Dependent; Gene Expression Regulation: Chromatin Modification in the CNS; Gene Expression Dysregulation in CNS Pathophysiology; Genomic Disorder and Gene Expression in the Developing CNS; Molecular Anatomy of the Mammalian Brain; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Schizophrenia: Genetics; Stress Response: Genetic Consequences.

Further Reading

Bahn S, Augood SJ, Ryan M, et al. (2001) Gene expression profiling in the post-mortem human brain: No cause for dismay. *Journal of Chemical Neuroanatomy* 22: 79–94.

Barlow C and Lockhart DJ (2002) DNA arrays and neurobiology: What's new and what's next? *Current Opinion in Neurobiology* 12: 554–561.

Baron CA, Tepper CG, Liu SY, et al. (2006) Genomic and functional profiling of duplicated chromosome 15 cell lines reveal regulatory alterations in UBE3A-associated ubiquitin-proteasome pathway processes. *Human Molecular Genetics* 15: 853–869.

Blalock EM, Chen KC, Stromberg AJ, et al. (2005) Harnessing the power of gene microarrays for the study of brain aging and Alzheimer's disease: Statistical reliability and functional correlation. *Ageing Research Reviews* 4: 481–512.

Fernandes C, Paya-Cano JL, Sluyter F, et al. (2004) Hippocampal gene expression profiling across eight mouse inbred strains: Towards understanding the molecular basis for behaviour. *European Journal of Neuroscience* 19: 2576–2582.

Flatscher-Bader T, van der Brug MP, Landis N, et al. (2006) Comparative gene expression in brain regions of human alcoholics. *Genes, Brain and Behavior* 5(supplement 1): 78–84.

Higgs BW, Elashoff M, Richman S, and Barci B (2006) An online database for brain disease research. *BMC Genomics* 7: 70.

Hovatta I and Barlow C (2005) Genome to disease. In: Koslow SH, and Subramaniam S (eds.) *Databasing the Brain: From Data to Knowledge,* 1st edn., pp. 169–190. New York: Wiley.

Hovatta I, Tennant RS, Helton R, et al. (2005) Glyoxalase 1 and glutathione reductase regulate anxiety in mice. *Nature* 438: 662–666.

Hu VW, Frank BC, Heine S, Lee NH, and Quackenbush J (2006) Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. *BMC Genomics* 7: 118.

Iwamoto K and Kato T (2006) Gene expression profiling in schizophrenia and related mental disorders. *Neuroscientist* 12: 349–361.

Liu J, Lewohl JM, Harris RA, et al. (2006) Patterns of gene expression in the frontal cortex discriminate alcoholic from nonalcoholic individuals. *Neuropsychopharmacology* 31: 1574–1582.

Lockhart DJ and Barlow C (2001) DNA arrays and gene expression analysis in the brain. In: Chin HR, and Moldin SO (eds.) *Methods in Genomic Neuroscience,* 1st edn., pp. 143–169. Boca Raton, FL: CRC Press.

Lockhart DJ and Barlow C (2001) Expressing what's on your mind: DNA arrays and the brain. *Nature Reviews Neuroscience* 2: 63–68.

Mirnics K, Levitt P, and Lewis DA (2006) Critical appraisal of DNA microarrays in psychiatric genomics. *Biological Psychiatry* 60: 163–176.

Purcell AE, Jeon OH, Zimmerman AW, Blue ME, and Pevsner J (2001) Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* 57: 1618–1628.

Relevant Websites

http://www.ebi.ac.uk – European Bioinformatics Institute's ArrayExpress.

http://www.jax.org – Mouse Phenome Database.

http://www.ncbi.nlm.nih.gov – National Center for Biotechnology Information's Gene Expression Omnibus.

http://www.stanleyresearch.org – The Stanley Brain Collection.

Genetic Regulation of Circadian Rhythms in *Drosophila*

R Allada, Northwestern University, Evanston, IL, USA
P Emery, University of Massachusetts Medical School, Worcester, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Drosophila Circadian Rhythms and the *period* Gene

The fruit fly, *Drosophila melanogaster*, is one of the preeminent genetic model organisms for the study of circadian rhythms. Because of its short generation time, the ease of screening large populations of mutant animals, its century long history of development of genetic and molecular tools and ultimately its conservation with higher organisms, especially humans, the fly has played an important role in both the discovery of new circadian rhythm genes and more importantly, the revelation of underlying principles in clock function. In the premolecular age, *Drosophila* had been used as a model for defining fundamental properties of circadian rhythms. Indeed, the genus name *Drosophila* comes from the Latin meaning 'dew loving' derived from the rhythm of eclosion (emergence of the adult from the pupal case) in the morning. The molecular era can be traced to behavioral genetic screens initiated by Seymour Benzer during the late 1960s and early 1970s. Benzer believed that the same genetic approaches used to elucidate the basic principles of cell and molecular biology could be used to analyze animal behavior. Among many behavioral screens, his laboratory screened for mutants that eclosed at different times of day. Three mutants were identified: one with short-period rhythms, one with long-period rhythms, and one with no discernable period. These mutants also displayed altered locomotor activity rhythms, the assay most commonly used today to monitor circadian behavior. All three mutants mapped to a single locus they termed *period* (*per*). Unfortunately the cloning of the *per* gene during the 1980s failed to immediately reveal underlying clock mechanisms. Ultimately, studies of the *per* transcript demonstrated daily oscillations in levels, and mutations in the *per* coding sequence that alter behavioral period length alter the period length of these molecular oscillations in a parallel manner. This led to the notion of a transcriptional feedback loop in which PER protein must feed back and regulate its own transcription to generate circadian oscillations that can persist even in the absence of any environmental cue. Studies in many organisms, including mammals, have identified similar transcriptional feedback loops at their core, suggesting that such loops are a conserved aspect of circadian clock function.

In addition to oscillatory transcription, PER was also found to undergo rhythms in protein levels, phosphorylation, and nuclear localization. As for transcript oscillations, circadian oscillations at the protein level have been defined subsequently in the core clocks of many other organisms. PER is not only found in the central nervous system (CNS) but in many other neurons and non-neuronal tissues, helping establish the notion of peripheral clocks (those outside the CNS). Indeed, most organs in mammals are known to harbor circadian clocks. Thus, studies of the *per* gene in *Drosophila* have elucidated many of the operating principles that have been subsequently found among circadian clocks throughout nature. Indeed, circadian gene discovery in the fly has often led to the discovery of new components and underlying mechanisms in mammalian, even human, circadian clock function.

The Molecular Circadian Pacemaker of *Drosophila*

Transcriptional Feedback

The transcriptional activators: CLK and CYC Early molecular studies demonstrated that the *per* gene is a central element of a transcriptional feedback loop that generates circadian rhythms in *Drosophila*. The positive elements of this loop are the transcription factors CLOCK (CLK) and CYCLE (CYC; Figure 1). These two transcription factors are members of the basic helix-loop-helix PAS family of transcription factors. Of note, PER is one of the three founding members of this family, with aryl hydrocarbon nuclear translocator (ARNT) and single-minded (SIM). CLK and CYC form a heterodimer that binds tightly to a DNA motif called the E-box (CACGTG) to activate transcription. An E-box is present in the *per* gene and is responsible for the circadian control of its transcription. Flies carrying inactive or null alleles of *Clk* and *cyc* express very little *per* and *tim* transcripts and are behaviorally arrhythmic. Thus, CLK and CYC are obligate heterodimers, at least for circadian rhythms.

The transcriptional repressors: PER and TIM The molecules responsible for blocking the activity of the CLK/CYC dimer are PER and its heterodimeric partner TIM (Figure 1). As for *per*, flies with a null mutation in the *tim* gene are arrhythmic. *tim* transcription

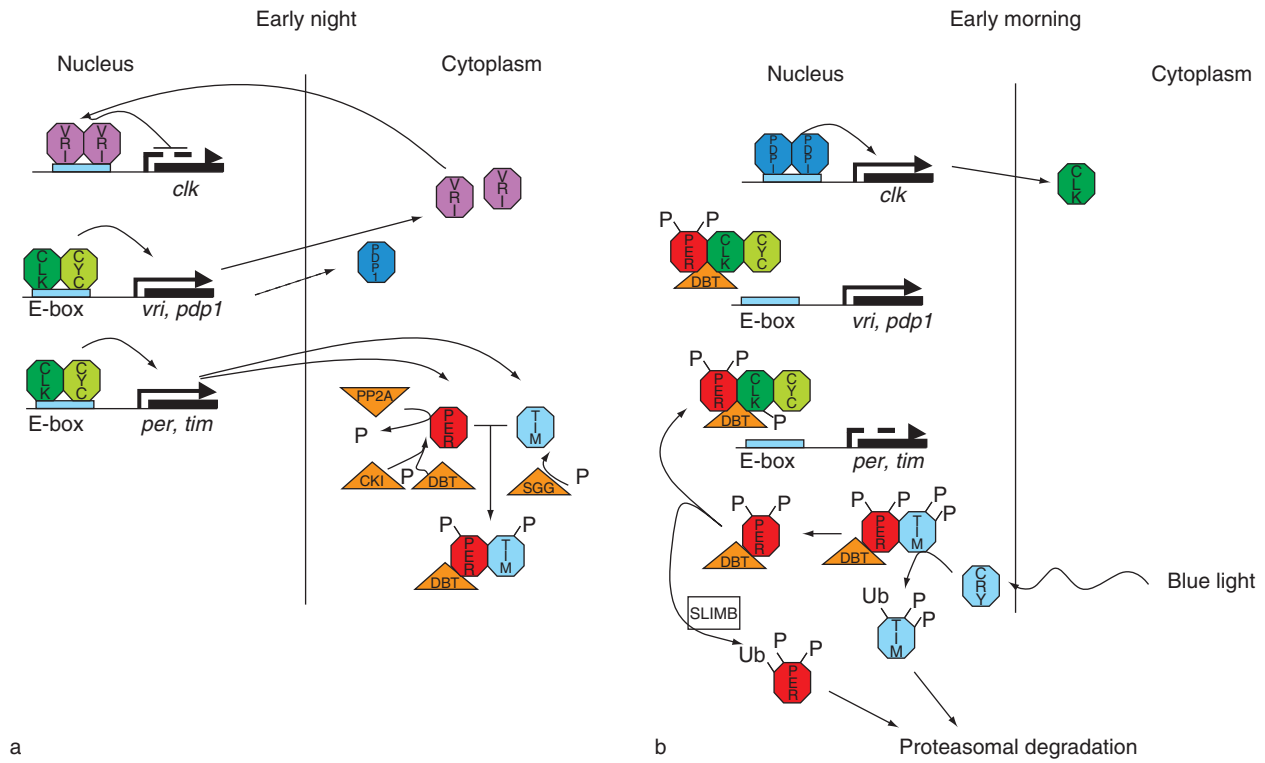

Figure 1 A model for the molecular circadian pacemaker in *Drosophila*. In the early night (a), the *per* and *tim* genes are actively transcribed under the positive influence of the transcription factors CLK and CYC. PER and TIM dimerize and begin to be phosphorylated in the cytoplasm. SGG phosphorylates TIM, CK2 and DBT phosphorylate PER, while PP2A counteracts the effect of these two kinases. In addition, DBT also associates durably with PER. In the early morning (b), PER and TIM are mostly nuclear, and highly phosphorylated. PER binds to the CLK/CYC dimer and reduces its affinity for the E-box, possibly with the help of DBT. This results in the repression of PER and TIM expression and the closing of a first transcriptional feedback loop. To restart the cycle, PER and TIM must be eliminated. PER and TIM degradation begins during the second half of the night. Hyperphosphorylated PER is targeted for proteasomal degradation by the E3 ubiquitin ligase SLIMB. TIM is also degraded, but the mechanism of its degradation is not clear. When the flies are exposed to light (b), the blue-light photoreceptor CRY is activated by light, binds to TIM, and greatly accelerates its proteasomal degradation. PER degradation is also accelerated since PER is no longer protected by TIM. Importantly, light is not required for this feedback loop to generate and maintain its circadian oscillations, but it is important to reset properly the phase of these oscillations every day. In addition, a second transcription loop controls CLK expression. In the early night, the *clk* gene repressor VRI, whose circadian expression is controlled by CLK, is highly expressed, while the circadian expression of the *clk* activator PDP1, also regulated by CLK/CYC, has not peaked yet. As a result, *clk* expression is low. However, at the end of the night and the early morning, PDP1 dominates and CLK expression is high. The role of this second transcriptional loop is still mysterious.

is also circadianly regulated by CLK/CYC, and TIM protein levels oscillate robustly during the day and night. TIM is thus an integral part of the transcriptional feedback loop that generates circadian oscillations.

While PER and TIM are heterodimeric partners, it is likely that PER is playing the dominant role in repressing CLK/CYC-activated transcription. PER and TIM are believed to repress transcription by reducing the affinity of CLK/CYC for the E-box. In the pacemaker neurons controlling circadian behavior (the LNv, see below), PER enters the nucleus before TIM. Moreover, PER does not need TIM to translocate into the nucleus when expressed at a high level. Finally, it appears that PER can inhibit transcription independently of TIM in a cell culture assay. The principal role of TIM appears to be protecting PER from proteasomal degradation, rather than a biochemically direct role in repressing transcription.

The precise mechanism by which PER represses CLK/CYC remains unclear. PER forms complexes with CLK/CYC and the formation of this complex coincides with a reduction in the binding of CLK/CYC to E-box DNA *in vivo*. Although PER and CLK/CYC share a PAS dimerization domain, this PAS domain does not appear to be required for PER repression of CLK. It is noteworthy that over 30 years after the isolation of the original short- and long-period alleles of *per*, we still do not have a clear explanation of how PER works and the molecular basis of these short- and long-period phenotypes.

In summary, the current model for the transcriptional loop involves four transcription factors: two

transactivators, CLK and CYC; and two repressors, PER and TIM (Figure 1). This 24 h feedback loop has turned out to be remarkably conserved in the animal kingdom. Homologs of CLK, CYC, and PER are also involved in the transcriptional feedback loop that generates circadian rhythms in mammals, but TIM's function in the mammalian clock is not yet clear.

In mammals, mammalian cryptochromes (*crys*) rather than *per*s are the primary transcriptional repressors of the CLK/CYC orthologs mClock and Bmal1. In flies, CRY is a circadian photoreceptor (see below) that is not required for circadian oscillations. Nevertheless, fly CRY does support robust rhythms in constant conditions in the antennae (the main olfactory organ) and the Malpighian tubules (insect kidneys). Moreover, in the eyes, CRY can repress CLK/CYC activation, although it is not required for the circadian oscillations of PER and TIM proteins. Interestingly, several non-drosophilid insects have two cryptochromes: one homologous to mammalian CRYs that represses CLK/CYC-mediated transcription and the other that resembles the photoreceptive *Drosophila* CRY. It therefore appears that drosophilids lost the repressive CRY during evolution, and that the photoreceptive CRY adopted some repressor activity.

The *Clock* Loop

In addition to *per* and *tim*, the CLK/CYC dimer activates multiple genes including the transcriptional activator PDP1 and the repressor VRILLE (VRI; Figure 1). These transcription factors, in turn, bind to the same DNA motif in the *Clk* promoter. Interestingly, the concentration of both factors oscillates but with distinct phases, with VRI peaking a few hours earlier. The differential oscillation in activator and repressor results in coordinated circadian oscillations in transcription of *Clk*. The function of this second transcriptional feedback loop remains uncertain. It is clear that this loop is not necessary for generating circadian rhythms, since flies in which *Clk* is overexpressed under the control of the *per* promoter display essentially normal circadian rhythms. Since very similar feedback loops are also found in mammals and fungi, it suggests that interlocked feedback loops are important for the circadian systems.

Posttranslational Regulation

Rhythmic protein phosphorylation appears to be a ubiquitous feature of core circadian clocks. There are accumulating data suggesting that rhythmic protein metabolism may not require rhythmic transcription and indeed may even drive rhythmic transcription. Rhythmic PER and TIM oscillations still occur even when they are driven by constitutive promoters. In cyanobacteria, rhythmic protein levels and phosphorylation persist in the absence of rhythmic transcription. A 24 h phosphorylation cycle can even be reconstituted *in vitro* using purified cyanobacterial clock proteins and ATP. These data suggest that transcriptional feedback may be dispensable and thus, may not be a central feature of the clockwork. On the other hand, these studies highlight a potential dominant role for rhythmic phosphorylation in driving circadian timing. Consistent with this hypothesis, the strongest period-altering alleles are of *per*, *tim*, and their associated kinases.

At least three protein kinases have been defined as important players in terms of period length: DOUBLETIME (DBT), the fly ortholog of CK1ϵ; Shaggy (SGG), the fly ortholog of glycogen synthase kinase 3 (GSK3); and casein kinase 2 (CK2; Figure 1). Taken together, these kinases modify the activity, subcellular localization, and stability of core clock components. In the case of PER and TIM, rhythmic phosphorylation has been assessed by Western blotting and observation of mobility shifts that can be reversed by phosphatase treatment. In both cases, disappearance of the proteins coincides with peak levels of phosphorylation, suggesting that phosphorylation serves as a signal for protein degradation.

DBT phosphorylation of PER is a key element in circadian timekeeping. *Dbt* orthologs (CK1δ and ε) also function in mammalian, including human, clocks. The most potent dominant period-altering alleles have been identified in *Dbt*, suggesting an important role in setting period length. Surprisingly, both short- and long-period alleles appear to reduce the DBT catalytic activity. In *Dbt* null mutant larvae, PER accumulates to very high levels in a hypophosphorylated state, while *per* mRNA is not strongly affected, suggesting that DBT destabilizes PER. Moreover, DBT can induce PER phosphorylation and degradation in cultured cells and DBT co-immunoprecipitates with PER *in vivo*, suggesting that PER is an *in vivo* DBT substrate.

DBT-induced phosphorylation of PER likely triggers its degradation by the ubiquitin proteasome system. Accumulation of hyperphosphorylated PER and TIM is typically followed by their degradation. In mutants of the E3 ubiquitin ligase, *supernumerary limbs* (*slimb*), these forms of PER and TIM accumulate. Moreover, SLIMB directly interacts with PER and DBT. Cell culture experiments suggest that SLIMB may specifically interact with DBT-phosphorylated PER, and thus signal its ubiquitination and subsequent proteasomal degradation.

DBT may also mediate negative feedback of PER onto CLK/CYC. Knockdown of DBT in tissue culture cells reduces PER repression of CLK activation. Indeed, DBT may also phosphorylate and destabilize CLK. Moreover, this phosphorylation may depend on PER. Thus, PER repression of CLK activation may be mediated through DBT phosphorylation of CLK. This is analogous to fungal clocks where changes in phosphorylation may mediate rhythmic repression of transcriptional activators.

CK2 and SGG are two kinases that phosphorylate PER and TIM, respectively and promote their translocation to the nucleus (Figure 1). The timing of TIM and PER nuclear translocation is delayed in long-period *sgg* loss-of-function mutants. Of note, *sgg* overexpression can increase TIM phosphorylation even in per^{01} mutant flies and can phosphorylate TIM *in vitro*, suggesting direct effects of SGG on TIM. Interestingly, GSK3 is also a target of lithium, a therapeutic agent for bipolar depression. Evidence suggests that defects in the circadian clock may contribute to bipolar disorder.

The CK2 holoenzyme is a heterotetramer consisting of two catalytic α subunits and two regulatory β subunits. Mutations that disrupt α- or β-function result in long-period behavioral rhythms. Both subunits are cytoplasmic in pacemaker neurons and the transition of PER from cytoplasm to nucleus is delayed in CK2 mutants. CK2 increases PER phosphorylation *in vivo* and can specifically phosphorylate PER *in vitro*. *In vivo* mutation of CK2 phosphorylation sites in PER recapitulates many aspects of CK2 mutants. These data implicate CK2 in phosphorylating PER in the cytoplasm and signaling its entry into the nucleus.

Both SGG and DBT are kinases that recognize and phosphorylate previously phosphorylated recognition sequences. The imposition of a requirement of prior phosphorylation for subsequent phosphorylation may introduce the necessary time delays to produce a 24 h phosphorylation rhythm. In this model, the phosphorylation of target substrates such as PER and TIM would be greatly facilitated by other kinases that prime their phosphorylation. The identity of the priming kinase is unknown.

A phosphatase that may mediate rhythms in phosphorylation state is PP2A (Figure 1). PP2A consists of three subunits: structural, regulatory, and catalytic. Two genes encoding regulatory subunits, *wiederborst* (*wdb*) and *twins* (*tws*) both appear to play a role in rhythms. Genetic manipulation of *wdb*, *tws,* and the catalytic subunit, *microtubule star* (*mts*), indicates that PP2A both stabilizes and promotes nuclear localization of PER. PP2A efficiently dephosphorylates DBT-induced PER phosphorylation, suggesting that PP2A may specifically antagonize this kinase. One of the transcripts of the regulatory subunit *tws* is clock-regulated. As none of the kinases thus far exhibit daily rhythms in activity or expression, rhythmic *tws* expression and consequent rhythms in PP2A activity may drive daily rhythms of PER phosphorylation. In this model, the 24 h posttranslational rhythm of PER phosphorylation and degradation is driven by the delicate balance between kinase and phosphatase action.

Synchronization of the *Drosophila* Circadian Clock

The period of circadian clocks only approximates 24 h. Therefore, circadian clocks need to be reset every day to remain synchronized with the environment. In *Drosophila*, three inputs are known to synchronize the clock: light, temperature, and social cues.

The Dedicated Circadian Light Input Pathway

The circadian clock responds quickly to light exposure. Light pulses as short as 1 min can shift the phase of circadian locomotor activity by 4 h. If the pulse is applied early during the night, the result will be a phase delay, while if it is applied late it will be a phase advance. The molecular correlate of this behavioral response to short light pulses is the proteasomal degradation of TIM, which will subsequently result in the reduction of PER levels, since TIM protects PER from degradation.

The photoreceptor responsible for TIM degradation is CRYPTOCHROME (CRY), a member of a family of blue-light sensitive proteins (Figure 1). Light pulses during the early night (when TIM levels are rising) delay TIM accumulation, resulting in delayed rhythms, while light pulses during the late night (while TIM levels are falling) accelerate TIM disappearance, advancing rhythms. While CRY binds directly to TIM, it is unknown how this event triggers TIM degradation.

CRY is expressed in almost all neurons and non-neuronal peripheral tissues that contain circadian clocks. Remarkably, these isolated peripheral organs can be synchronized to external light–dark cycles. In addition, flies that express CRY only in the pacemaker neurons controlling circadian behavior can respond to short light pulses behaviorally, but their peripheral clocks (with few exceptions, see below) remain blind, suggesting that CRY operates cell-autonomously. Of note, mammalian peripheral clocks do not appear to be able to synchronize to external light–dark cycles but rather respond to rhythmic humoral factors. Mammals are too large and opaque for light to reach internal organs, and thus the cell-autonomy of the light input pathway was lost during their evolution.

Other Light Input Pathways

CRY is not the only photoreceptor that can synchronize the circadian clock. Visual photoreceptive organs using rhodopsin as a photopigment (the eyes, the ocelli, and the Hofbauer–Buchner eyelets) can synchronize pacemaker neurons responsible for rhythmic behaviors in the absence of CRY. Nevertheless, in the absence of CRY, the phase of circadian behavior is not shifted by short light pulses and TIM is not acutely degraded, suggesting that CRY and rhodopsin synchronize the pacemaker by distinct mechanisms. However, the anatomic and molecular pathways by which visual input reaches the circadian pacemaker neurons are not clear. CRY-independent synchronization is also evident in some peripheral organs. In flies without functional CRY, the isolated antennae and the Malpighian tubules can still synchronize to external light–dark cycles. The nature of the photoreceptor, a rhodopsin or an as-yet-unidentified photoreceptor, remains unclear.

The utility of having different and partially redundant photoreceptor systems may be related to the fact that the spectrum of ambient light significantly changes during the sunrise and sunset. The combination of CRY and other photoreceptive organs may help the fly to monitor these changes in light spectrum, and thus more efficiently synchronize its circadian clock.

Nonphotic Environmental Cues That Synchronize the Circadian Clock

Temperature is an important synchronizing cue in animals, especially those that cannot maintain their own body temperature. The isolation of a mutant (*nocte*) with normal circadian light responses but insensitivity to temperature changes reveals a dedicated temperature input pathway in *Drosophila*. This pathway is likely cell-autonomous as dissected antennae or wings can be synchronized to temperature cycles. Nonetheless, the pace of the circadian pacemaker is protected from changes in ambient temperature, a phenomenon termed temperature compensation. A clock whose period changes with the weather would be useless. Since, the kinetics of biochemical reactions increase with temperature – usually doubling every 10 °C – there are compensatory mechanisms within the circadian pacemaker that prevent acceleration of the clock with increasing temperature. The mechanisms that underlie temperature compensation are unknown. While period is relatively insensitive to temperature, the phase of the circadian clock is temperature sensitive, perhaps mediating seasonal adaptations. A temperature-sensitive splicing in the 3′ untranslated region of *per* determines the phase of *per* mRNA and protein accumulation. Delayed PER accumulation at higher temperatures results in a later onset of behavioral activity in the evening, thus helping flies to avoid heat-mediated dessication during the middle of the day.

Olfactory social cues also influence circadian phase, suggesting the existence of a circadian pheromone. Surprisingly, when a small population of flies with a specific circadian phase is mixed with a larger population with a different phase, the larger population will shift its phase toward the phase of the smaller population, not vice versa.

Pacemaker Neural Network

Under 12 h light:12 h dark (LD) conditions, flies exhibit two peaks of activity, a morning peak around the time of lights-on and an evening peak around the time of lights-off. Flies anticipate each transition by increasing their activity, a process mediated by the circadian clock. Upon release into constant darkness (DD), flies continue to exhibit morning and evening peaks that gradually morph into a single broad peak of activity occurring approximately every 24 h.

Considerable progress has been made in understanding how these molecular oscillators function within cellular networks that control these rhythmic behaviors. There are roughly 200 neurons in the central brain that express core clock components. These can be subdivided into six main clusters: small ventral lateral neurons (sLNv), large ventral lateral neurons (lLNv), dorsal lateral neurons (LNd); and three groups of dorsal neurons (DN1, DN2, and DN3; Figure 2).

While molecular oscillators within these neurons largely keep time, network communication is essential to maintain synchrony between these intracellular clocks. The neuropeptide pigment dispersing factor (PDF) is a major synchronizing agent (Figure 2). PDF is selectively expressed in the LNv. The sLNv send dorsal projections toward the DNs. PDF is rhythmically expressed in these dorsal projections, presumably reflecting clock-controlled release. In addition, the lLNv send projections to their contralateral counterparts. Loss of *pdf* and its receptor *pdfr* result in changes in the core molecular oscillator of both PDF and non-PDF neurons. In the absence of PDF/PDFR, flies exhibit advanced timing of evening anticipation in LD and short (albeit weak) periods in DD. These behavioral changes are accompanied by changes in molecular oscillations within PDF and non-PDF neurons including phase advances, desynchrony, and reduced amplitude. Selective acceleration of the clock in the LNv speeds up the molecular oscillator in the LNd, DN1 and DN3 groups, strongly

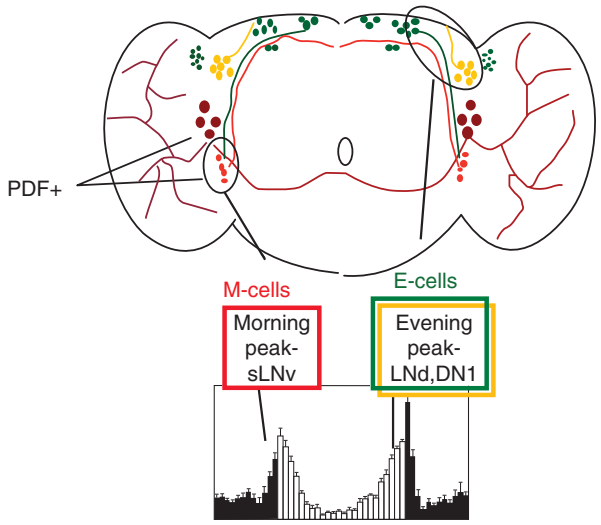

Figure 2 The *Drosophila* pacemaker neural network. Top panel shows a schematic of the fly brain with small ventral lateral neurons (sLNv) indicated in orange, large ventral lateral neurons (lLNv) indicated in maroon, dorsal lateral neurons (LNd) indicated in yellow, and three groups of dorsal neurons (DN) indicated in green. The sLNv send projections dorsally and lLNv send contralateral projection. The ventral lateral neurons express the neuropeptide PDF. The bottom panel shows the locomotor activity behavior of a population of wild-type fruit flies. Location of morning and evening peaks is indicated. Cells that underlie morning and evening behavior are circled and indicated as M- and E-cells.

suggesting a role for PDF in resetting and synchronizing these other oscillators to the period set in the LNv. Of note, speeding up the oscillators of the non-PDF neurons does not drive faster behavioral or LNv rhythms, suggesting that the PDF neurons are at the top of a network hierarchy that controls circadian period.

Selective pacemaker neurons mediate morning and evening behavior. PDF is likely communicating to output neural circuits that ultimately drive morning behavior. *pdf*/*pdfr* mutants display little or no LD morning anticipation or DD morning peak. Of note, ablation of the PDF neurons have a similar phenotype suggesting that PDF is the principal LNv neurotransmitter, indicating that these cells are responsible for morning anticipation, acquiring the name 'morning cells.' Yet these large changes in morning behavior are not accompanied by similarly large changes in the molecular oscillators. These data suggest that PDF is playing a role downstream of the core oscillator to regulate morning behavior. In contrast to the role of the PDF neurons, the non-PDF neurons, particularly the LNd/DN1, are essential for evening anticipation and have been termed the 'evening cells.' In addition, they set the phase (but not period) of the evening peak relative to the morning peak under DD conditions.

In understanding the molecular mechanisms by which different circadian oscillators are synchronized requires elucidating how the core oscillator regulates neural outputs such as PDF and in turn, how PDF receptor signaling feeds back and resets target oscillators. Microarray studies have identified hundreds of rhythmically expressed transcripts, suggesting that clock-mediated regulation of transcription mediates clock outputs. One presumed neural target of clocks is membrane excitability. In a variety of systems, the function or activity of ion channels is under clock control. In flies, there is some evidence that clocks control resting membrane potential. Moreover, a novel voltage-gated ion channel, NARROW ABDOMEN, appears to play an important role in regulating pacemaker neuron output. Silencing of pacemaker neurons can abolish free-running rhythms suggesting an important role for membrane excitability in regulating core clock function.

Unlike most peripheral oscillators in *Drosophila*, molecular oscillators in the core pacemaker neurons exhibit long-term persistent rhythmicity in DD. Network connections between these neurons may serve to sustain rhythms in the absence of external cues. As loss of PDF results in altered neuronal oscillators, it suggests that PDF receptor signaling may play a role in synchronizing and sustaining core oscillators. Studies in heterologous systems demonstrate that the PDF receptor is coupled to $G\alpha_s$, assisted by the GTPase-activating protein, neurofibromatosis-1 (NF1), and its activation results in increases in cAMP-dependent gene transcription. In addition, calcium-dependent transcription is also promoted. However, the PDF-R signaling pathway still needs to be elucidated *in vivo*. In the mammalian circadian pacemaker, the suprachiasmatic nucleus (SCN), the neuropeptide, vasoactive intestinal peptide (VIP) plays an analogous role to PDF in synchronizing circadian neurons. Thus, the similarity between the mechanisms generating circadian rhythms in *Drosophila* and mammals is not limited to the molecular pacemaker, but also includes the neural network controlling circadian rhythms.

See also: Circadian Organization; Circadian Organization in Non-Mammalian Vertebrates; Circadian Regulation in Invertebrates; Circadian Rhythms: Influence of Light in Humans; Mammalian Sleep and Circadian Rhythms: Flies; Photoreceptors: Physiology; Sleep and Waking in Drosophila.

Further Reading

Chang DC (2006) Neural circuits underlying circadian behavior in *Drosophila melanogaster*. *Behavioural Processes* 71: 211–225.

Dunlap JC, Loros JJ, and DeCoursey PJ (eds.) (2004) *Chronobiology: Biological Timekeeping*. Sunderland, MA: Sinauer Associates.

Hardin PE (2005) The circadian timekeeping system of *Drosophila*. *Current Biology* 15: R714–R722.

Helfrich-Forster C (2005) Neurobiology of the fruit fly's circadian clock. *Genes, Brain, and Behavior* 4: 65–76.

Hendricks JC and Sehgal A (2004) Why a fly? Using *Drosophila* to understand the genetics of circadian rhythms and sleep. *Sleep* 27: 334–342.

Sehgal A (ed.) (2004) *Molecular Genetics of Circadian Rhythms*. Hoboken, NJ: Wiley.

Weiner J (1999) *Time, Love, and Memory*. New York: Alfred A. Knopf.

Yu W, Zheng H, Houl JH, Dauwalder B, and Hardin PE (2006) PER-dependent rhythms in CLK phosphorylation and E-box binding regulate circadian transcription. *Genes and Development* 20: 723–733.

Genetics of Circadian Disorders in Humans

A-M Chang, Harvard Medical School, Boston, MA, USA
P C Zee, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Circadian Rhythms in Mammals

As terrestrial beings living on a planet with an axial rotation of about 24 h, giving rise to rhythms in light and darkness, almost all living organisms display circadian rhythms in their behavior. It is this near-24 h, period that gives the name 'circadian' or 'about a day' to the system. In mammals the circadian timing system, or biological clock, is located in the suprachiasmatic nuclei (SCN) of the hypothalamus. This circadian pacemaker synchronizes many physiologic processes to cyclic changes in the external environment and also maintains the temporal organization of these processes to each other. The daily light and dark cycles entrain the circadian system which signals the appropriate time for activity and rest. The circadian activity/rest rhythm as well as other physiological, biochemical, and molecular processes are synchronized to the external environment in response to changes across the 24 h day. From an evolutionary viewpoint, this kind of circadian organization creates specific temporal groups of animals sharing times of the day in which to do certain activities, for example, diurnal or nocturnal. This may have developed in organisms in response to several needs: food, search for prey, protection from predators, and reduced competition for resources in a physical environment.

Although humans are diurnal animals, preference for certain times of the day, or chronotype, varies among individuals. Some people prefer activity in the early part of the day; others prefer night-time activity, while still others have no preference. Early types have been described as larks and late types as owls. Subjective measures of diurnal preference by questionnaire have been used to assess the morningness or eveningness of individuals for several decades, driven in large part by the desire to identify workers' abilities to adjust to changing work schedules such as shift work. A number of studies have shown that diurnal preference has, in part, a genetic basis yet there are certainly environmental, social, and behavioral influences on the preference for sleep/wake schedules. Different subjective chronotypes have been associated with objective measures of circadian rhythms, including body temperature and hormonal rhythms. Diurnal preference has also been correlated to a sleep/wake pattern and performance measures.

The expression of circadian behavior is also, in part, genetically determined. The first circadian mutant mammal identified, the *tau* mutant hamster, has a shortened endogenous period compared to the wild-type hamster. This altered circadian period of locomotor activity was seen in the offspring of mutant animals demonstrating the heritability of circadian behavior and was eventually associated with a specific genetic mechanism. The identification of other mammalian circadian mutants that show heritability of their altered circadian rhythms has led to the identification of circadian genes that comprise the molecular mechanism for the generation of circadian rhythms. Homologs of these circadian genes have been identified in numerous species, including humans, showing a remarkable conservation of the molecular mechanisms of the circadian pacemaker. Evidence in different model systems shows that circadian timing is regulated transcriptionally, translationally, and post-translationally by genes that make up the clock mechanism. They form autoregulatory feedback loops that drive the essential cycling patterns of daily rhythms and mutations in these genes result in alterations or abolition of properties of circadian rhythms.

Circadian Rhythm Sleep Disorders

Sleep is a physiological and behavioral state that is necessary for life and highly conserved in numerous species. There are several theories about the function of sleep and it appears likely that it fulfills multiple functions. It is also possible that sleep served one primordial function necessary for survival and that multiple purposes evolved. Moreover, the lower threshold for arousal that occurs during sleep, putting an organism at great risk, reflects its essential function. Sleep loss impacts general health, mood, and cognitive function. Results from acute and chronic sleep deprivation studies have shown that the consequences of sleep loss on health, determined by a number of physiological and endocrine measures, and cognitive function, assessed on several performance tasks, are severe and detrimental. In fact, the result of near-total sleep deprivation in rat models is death in about 3 weeks.

For optimal sleep, the desired sleep time should coincide with the timing of the circadian rhythm of sleep propensity. A misalignment between sleep and the 24 h social and physical environment results in circadian sleep/wake cycle disruption and circadian

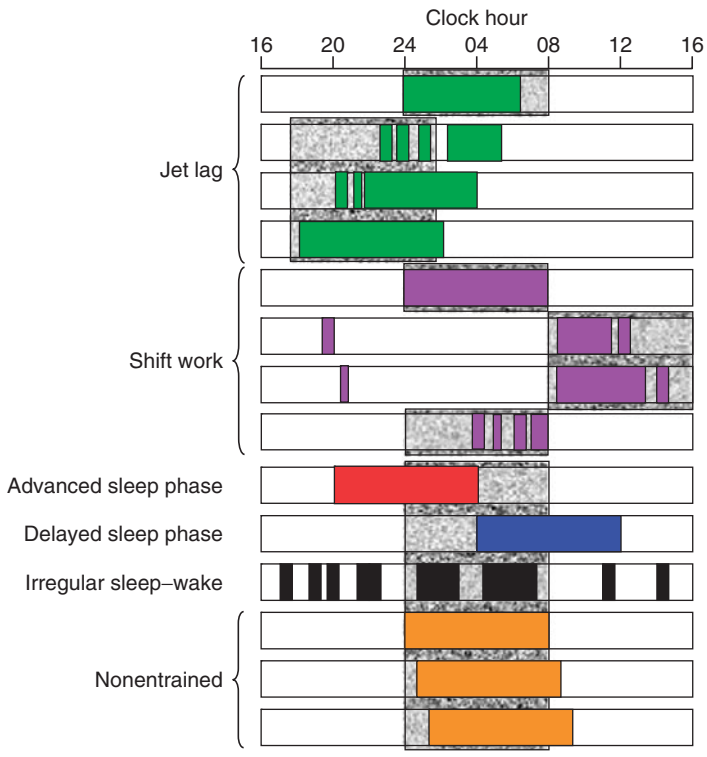

Figure 1 Circadian rhythm sleep disorders. The colored bars represent the sleep pattern for each type and the gray areas represent typical sleep times. For the first two types the environmental schedule is shifted, relative to the internal circadian clock. Examples are shown for jet lag (green) where the typical sleep time is advanced by 7 h (trans-Atlantic flight) and for shift work (purple) where a day shift (sleep at night) is followed by two consecutive night shifts (sleep during the day) followed again by a day shift (rotating shift schedule). For the last four types the circadian clock is altered relative to the environmental schedule.

rhythm sleep disorders, shown in Figure 1. Two types of circadian rhythm sleep disorders, jet lag and shift work, occur when the physical environment is altered relative to the internal circadian clock. Four types of circadian rhythm sleep disorders may result when the endogenous circadian rhythm is altered relative to the external environment: advanced sleep phase (ASPS), delayed sleep phase (DSPS), irregular sleep/wake, and nonentrained types. These types of disorders are thought to occur because of chronic alterations in the circadian pacemaker that result in the inability of the individual to achieve or maintain a conventional phase relationship with the external world. Of these four types, the most common are DSPS and ASPS, in which the timing of the major consolidated sleep period is altered, delayed, or advanced, in relation to local clock time. Despite being apparent disorders of circadian timing, little is known about how the circadian system and the diverse rhythms under its control are affected. As increasing numbers of patients with ASPS and DSPS are diagnosed and studied, more can be discovered about the pathophysiology of these disorders and rational strategies for treatments developed. In this article, the focus is on the genetic basis for the circadian rhythm sleep phase disorders, ASPS and DSPS, because little is known regarding the genetic contributions to jet lag, shift work sleep disorder, irregular sleep/wake cycle, or nonentrained sleep/wake disorder.

Diagnosis

The diagnoses of both DSPS and ASPS are based on published clinical criteria. Their diagnosis requires a complaint of inability to fall asleep or to remain asleep at the desired bedtime. Diagnostic criteria also include an advanced or delayed sleep pattern that is stable, and normal sleep quality and duration when patients are allowed to choose their preferred sleep schedule. A stable advance or delay of the habitual sleep period must be demonstrated by either sleep log or actigraphic recording in combination with sleep log for a minimum of 1 week. These criteria exclude the existence of any other sleep disorder, medical or neurological disorder, mental disorder, medication use, or substance use that can better explain the altered sleep pattern.

DSPS, first described by Weitzman and colleagues, is characterized by bedtimes and wake times that are

delayed 3–6 h, relative to desired sleep/wake times. The patient typically finds it difficult to fall asleep before 2–6 a.m. and wake up earlier than 10 a.m.–1 p.m. Classic symptoms include chronic sleep onset insomnia and difficulty waking in the morning. Patients also report daytime sleepiness, primarily during the earlier part of the day. Patients with DSPS often report feeling most alert and active in the late evening and most sleepy in the morning. They typically score as 'evening' types on the Horne and Östberg questionnaire of diurnal preference and are described as night people, or owls. The delay of the sleep period, particularly of the wake time, results in the patient's complaint of difficulty maintaining early morning work times, school schedule, or social obligations. Enforced conventional wake times frequently result in chronic sleep deprivation as the patient forces him- or herself to awaken early but still cannot fall asleep until later than desired. Often sleep is extended on days when the individual has no obligations that require an early wake time, such as on weekends, days off, and vacations.

In contrast to DSPS, ASPS is characterized by habitual and involuntary sleep (6–9 p.m.) and wake (2–5 a.m.) times that are several hours earlier than desired. ASPS patients complain of early morning awakening insomnia or sleep maintenance insomnia and also of sleepiness in the late afternoon or early evening. They typically consider themselves larks and score as morning types on subjective questionnaires. They report feeling best and most alert in the morning; however, some patients who delay their bedtime due to familial or social pressures still experience early morning awakenings and as a result are chronically sleep-deprived. An advance in the phase of the sleep/wake cycle is commonly seen with aging; however, sleep is more fragmented in this population as well, and alterations of the sleep homeostat could account for this advanced phase. In the few reported cases of young ASPS patients, sleep duration and architecture appeared normal without the fragmentation or sleep disturbances commonly seen in aging.

Prevalence

The prevalence of DSPS and ASPS is estimated to be low in the general population. The prevalence of DSPS is about 0.13–0.17%. The number appears to be much higher among adolescents and young adults, with a reported prevalence of 7–16%. This is supported by a number of studies showing that the sleep/wake rhythm is delayed in this group. The prevalence of ASPS is estimated to be much lower than DSPS. One reason for this disparity could be due to underreporting of ASPS, since an early schedule may have advantages in fitting in with societal norms and schedules and therefore 'morning people' do not have complaints about their sleep habits or feel a need to seek treatment. Another possible explanation for the low prevalence of ASPS is the fact that many of the symptoms are commonly seen with aging; therefore, complaints of early morning awakening may be attributed to normal senescence by patients and clinicians alike.

Pathophysiology

Based on the fundamental properties of the circadian timing system and sleep, several mechanisms could account for a persistently altered phase relationship between the sleep/wake cycle and the environment. One possible explanation is that the endogenous period of the circadian pacemaker is unusually short in individuals with ASPS and long in individuals with DSPS. With respect to DSPS, there is evidence showing a greater frequency of longer periods in adolescents, the same demographic group to show a higher prevalence of the DSPS. A long period that lies outside the range of entrainment or is too long to make the necessary adjustments to maintain a 24 h day may cause non-24 h sleep/wake syndrome. A long endogenous cycle that is still entrained to a 24 h day may have an altered phase angle of entrainment, delaying the sleep period in relation to the light/dark cycle. Likewise, in the case of ASPS, a short period could have an effect on the phase angle of entrainment causing a subsequent advance in the sleep and wake times relative to the light/dark cycle. The report that one member with familial ASPS had a short circadian period supports this explanation and suggests that an alteration in the period of the pacemaker may be involved in the development of ASPS.

Another possible explanation underlying the phase advance or delay seen in ASPS and DSPS is the masking of the phase advance or delay regions of the phase response curve (PRC) to light due to the sleep schedule (Figure 2). ASPS patients who are consistently awakening early are likely to be exposed to light at the region of their PRC causing phase advances. If they are also going to bed earlier in the evening, they are less likely to be receiving light at the delay region, thereby reinforcing the advanced schedule. Similarly, DSPS patients who awaken later in the morning are presumably sleeping through the phase advancing portion of their PRC and are more likely to be receiving phase delaying light in the late evening hours. A longer interval between body temperature nadir and sleep offset has been described in a few patients with DSPS, suggesting that the sleep period extended into the morning may be masking the advance

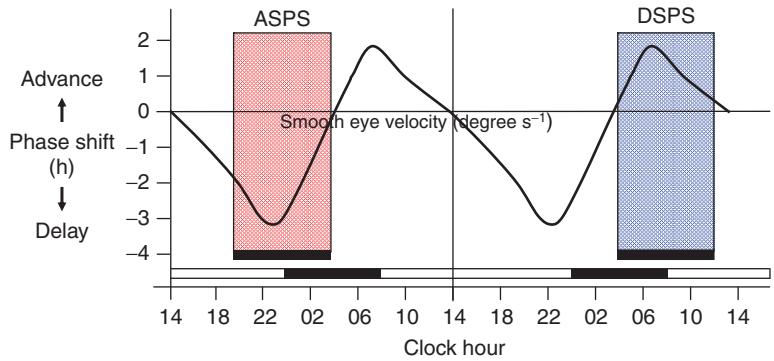

Figure 2 Representative human phase response curve to light, double-plotted. The light dark bar shows a 'typical' sleep/wake schedule (midnight–8 a.m.). The black bars at the bottom of the colored areas, represent the sleep times for ASPS and DSPS subjects, respectively. The colored bars show the proposed 'masking' of the PRC by the sleep schedule.

portion of the PRC in DSPS patients and therefore preventing them from phase advancing their circadian rhythms. This behavior acts as a vicious cycle and perpetuates the already advanced or delayed schedule in ASPS and DSPS, respectively.

An alteration in the entraining effects of light is yet another explanation for the altered phase relationship between the sleep/wake rhythm and the light/dark cycle. This has been described as a weak or low-amplitude advanced or delayed portion of the light PRC. For example, DSPS patients may have abnormally small advance portions of the light PRC explaining the inability to advance the phase of their sleep. Decreased sensitivity to the photic cues may also have an effect on the ability to synchronize an individual's sleep/wake rhythm to the environment. An increased sensitivity to night-time light has been shown in DSPS patients and a reduced response to morning light was demonstrated in subjects with a long free-running period, suggesting the possibility of an altered responsiveness to light cues in these conditions.

Finally, an alteration in the phase relationship between the circadian clock and the sleep/wake rhythm may play a role in these disorders. There have been reports of changes in the phase relationship between sleep times and measures of the circadian clock: specifically, an altered phase relationship between the melatonin rhythm and that of the sleep/wake cycle. The most consistent finding was a prolonged interval between melatonin acrophase and sleep offset. A change in the phase relationship of melatonin rhythm and core body temperature with an individual's diurnal preference has also been described with an earlier circadian phase of melatonin and core temperature in morning types and a later phase in evening types. Furthermore, the interval between circadian phase and habitual wake time was longer in morning types than in evening types. Because DSPS and ASPS patients segregate into the two diurnal groups, evening types or morning types, this supports the explanation that an alteration in the phase relationship between the clock and the sleep/wake rhythm may affect the phase of the sleep episode. Interestingly, when the morning types were divided by age, there was a difference in this phase relationship. Older morning types who also had an earlier circadian phase showed a shorter interval between core body temperature nadir and sleep offset, opposite to that seen in the young morning types. This suggests that different mechanisms may be responsible for the earlier phase seen in association with aging and ASPS. Therefore, it is likely that not only changes in the free-running period but also entrainment mechanisms of the circadian system contribute to the alteration in the timing of sleep in DSPS and ASPS.

Though it is commonly accepted that both DSPS and ASPS are predominantly a result of alterations of circadian timing, there is recent evidence that alterations in the homeostatic regulation of sleep may play an important role in the pathophysiology of circadian rhythm sleep disorders and in diurnal preference. Polysomnographic recordings of sleep in DSPS patients showed that sleep architecture was not disrupted after the initiation of sleep when the subject was allowed to sleep until their desired wake time. However, following 24 h of sleep deprivation, DSPS patients, when compared to controls, showed a decreased ability to compensate for sleep loss during the subjective day and the first hours of subjective night. In addition, the phase angle between melatonin onset and sleep propensity was larger in DSPS patients. It is possible that poor sleep recovery perpetuates the inability of patients with DSPS to reset their sleep phase. In a recent study of morning and evening types, those with similar intermediate circadian phases showed significant differences in dissipation of sleep pressure. In contrast, the morning and evening types with extremely early and late circadian phases did

not show this difference, suggesting that both the circadian system and the sleep homeostat contribute to diurnal preference in addition to sleep regulation and timing.

Genetic Basis of Human Circadian Rhythm Sleep Disorders

Our current understanding of circadian rhythm sleep disorders indicates that alterations in the interaction between circadian and homeostatic processes that regulate sleep and wakefulness play an essential role in the pathophysiology of DSPS and ASPS. Animal work examining the genetic regulation of both circadian and homeostatic mechanisms has revealed a complicated interaction between these two processes in determining sleep duration and quality. Sleep studies of *Clock* mutant mice showed that they not only sleep less but also appear to need less sleep, suggesting that the *Clock* mutation alters the homeostatic drive in these mutants. Similarly, sleep studies of *Cry1* and *2* double knockouts showed an increased sleep drive in the mutants and suggest that the *Cry* genes play an important role in both circadian behavior and the homeostatic regulation of sleep. In contrast, sleep analysis of m*Per1* and m*Per2* mutants showed no significant difference in homeostatic sleep regulation relative to wild-type mice. The only difference between the genotypes of each mutation was seen in the distribution of sleep throughout the day expressing the circadian regulation of sleep timing with the m*Per2* mutants in particular, showing an advance of the activity rhythm. This negative result suggests an independence of the circadian and homeostatic processes regulating sleep. However, it has been proposed that changes in the sleep homeostasis have been demonstrated in mutants having no functional circadian rhythms, such as arrhythmicity in constant darkness (e.g., *Clock*, *Bmal1*, *Cry* double-knockouts, and m*Per1*and m*Per2* double knockouts). The evidence in rodents that mutations in circadian genes can alter both circadian rhythmicity and sleep homeostasis suggests the possibility that a similar genetic predisposition could be involved in the etiology of the circadian and sleep homeostatic changes of sleep regulation seen in circadian rhythm sleep disorders.

In the last few years, a genetic basis for ASPS and DSPS has been established. Familial cases of ASPS and one published family with DSPS demonstrate the heritability of these disorders. The ASPS and DSPS phenotypes appear to segregate with an autosomal dominant mode of transmission. Furthermore, a family history was reported in 11 out of 53 patients with DSPS from a clinic in Japan. Genetic analysis of a family with ASPS led to the identification of a mutation in the human *PER2* gene causing the disorder. However, other published cases of fASPS had not been found to carry this mutation and showed heterogeneity for fASPS. More recently, a mutation in the *CKIδ* gene was linked to another fASPS kindred.

Association studies in idiopathic cases of circadian rhythm sleep disorders have also identified mutations in several circadian genes associated with altered circadian properties. Structural polymorphisms in the human *PER3* gene have been associated with the affected status in DSPS patients and different polymorphisms in the same *PER3* gene were also identified and associated with DSPS and eveningness. A missense mutation in the *CKIε* gene showed an inverse association with circadian rhythm sleep disorders DSPS and non-24 h sleep/wake syndrome. Genetic analysis of diurnal preference to identify polymorphisms in circadian genes has yielded similar results. A polymorphism in the *CLOCK* gene was correlated with evening type although two other groups did not find this association with eveningness in their study populations. More recently, association studies of diurnal preference have identified polymorphisms in the human *PER2* and *PER3* genes associated with extreme morningness and eveningness, respectively.

Summary

The study of disorders of the circadian timing of the sleep/wake rhythm in humans provides valuable insight into the interaction and regulation of various neurobiological processes. Further phenotypic characterization of aspects to the circadian regulation of sleep in people with an altered circadian phase of sleep (i.e., ASPS and DSPS) may serve to elucidate important pathways and functions of photic regulation and entrainment. Thorough evaluation of the sleep in these disorders may also yield information about the underlying processes affecting and causing the expressed changes of sleep and sleep timing. Genetic analysis of ASPS and DSPS familial and idiopathic cases is likely to reveal the molecular and cellular processes generating and modifying the phenotype and behavior. Together, these endeavors will undoubtedly provide greater understanding of the mechanisms underlying circadian rhythm and sleep regulation.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake

Cycle; Circadian Function and Therapeutic Potential of Melatonin in Humans; Circadian Rhythms: Influence of Light in Humans; Circadian Rhythms in Sleepiness, Alertness, and Performance; Psychiatric Disorders Associated with Disturbed Sleep and Circadian Rhythms; Shift Work and Circadian Rhythms; Sleep and Circadian Rhythm Disorders in Human Aging and Dementia.

Further Reading

American Sleep Disorders Association, Diagnostic Classification Steering Committee (2005) *International Classification of Sleep Disorders: Diagnostic and Coding Manual,* ICSD-2, 2nd edn., Westchester, IL: American Academy of Sleep Medicine.

Aoki H, Ozeki Y, and Yamada N (2001) Hypersensitivity of melatonin suppression in response to light in patients with delayed sleep phase syndrome. *Chronobiology International* 18: 263–271.

Archer SN, Robilliard DL, Skene, et al. (2003) A length polymorphism in the circadian clock gene Per3 is linked to delayed sleep phase syndrome and extreme diurnal preference. *Sleep* 26: 413–415.

Ebisawa T, Uchiyama M, Kajimura N, et al. (2001) Association of structural polymorphisms in the human *Period3* gene with delayed sleep phase syndrome. *EMBO Reports* 2: 342–346.

Horne JA and Östberg O (1976) A self-assessment questionnaire to determine morningness–eveningness in human circadian rhythms. *International Journal of Chronobiology* 4: 97–110.

Jones CR, Campbell SS, Zone SE, et al. (1999) Familial advanced sleep-phase syndrome: A short-period circadian rhythm variant in humans. *Nature Medicine* 5: 1062–1065.

Ozaki S, Uchiyama M, Shirakawa S, and Okawa M (1996) Prolonged interval from body temperature nadir to sleep offset in patients with delayed sleep phase syndrome. *Sleep* 19: 36–40.

Regestein QR and Monk TH (1995) Delayed sleep phase syndrome: A review of its clinical aspects. *American Journal of Psychiatry* 152: 602–608.

Reid KJ, Chang AM, Dubocovich ML, et al. (2001) Familial advanced sleep phase syndrome. *Archives of Neurology* 58: 1089–1094.

Schrader H, Bovim G, and Sand T (1993) The prevalence of delayed and advanced sleep phase syndromes. *Journal of Sleep Research* 2: 51–55.

Toh KL, Jones CR, He Y, et al. (2001) An h*Per2* phosphorylation site mutation in familial advanced sleep phase syndrome. *Science* 291: 1040–1043.

Uchiyama M, Okawa M, Shibui K, et al. (1999) Poor recovery sleep after sleep deprivation in delayed sleep phase syndrome. *Psychiatry and Clinical Neurosciences* 53: 195–197.

Weitzman ED, Czeisler CA, Coleman RM, et al. (1981) Delayed sleep phase syndrome. A chronobiological disorder with sleep-onset insomnia. *Archives of General Psychiatry* 38: 737–746.

Xu Y, Padiath QS, Shapiro RE, et al. (2005) Functional consequences of a CKIdelta mutation causing familial advanced sleep phase syndrome. *Nature* 434: 640–644.

Genetics of Human Anxiety and Its Disorders

A R Hariri, University of Pittsburgh, Pittsburgh, PA, USA
D R Weinberger, National Institutes of Health, Bethesda, MD, USA

© 2009 Published by Elsevier Ltd.

Introduction

Recent advances in both molecular genetics and human neuroimaging have begun to provide the tools necessary to explore how individual differences in complex cognitive and emotional behaviors emerge and how such differences may confer vulnerability to psychopathology. With the publication of the reference human genome sequence, a major effort is underway to exhaustively identify common variations in this sequence that impact on gene function (i.e., functional polymorphisms) and subsequently to understand how such functional variations alter human biology. Because approximately 70% of all genes are expressed in the brain, many of these functional polymorphisms will influence how the brain processes information and, as a consequence, regulates both cognitive and affective behaviors. Human neuroimaging (e.g., magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), and positron emissions tomography (PET)), because of its capacity to assay detailed brain structure and function within individuals, has unique potential as a tool for characterizing functional genetics in neural circuitry. This article (1) describes the conceptual basis for imaging genetics and (2) outlines some major findings in imaging genetics related to aspects of human emotion and temperament. As a first step, we consider behavioral association studies relevant to the genetics of emotion regulation in healthy and clinical populations. These studies, with their challenges and limitations, provide a foundation for the imaging genetics approach, which aims to examine mechanisms at the neural-systems level that lead to individual variability in emotion regulation that is more proximal to genetic factors.

Behavioral Genetics

Why Study Genes?

Genes represent the 'Go' square on the monopoly board of life. They are the biological toolbox with which we negotiate the environment. Although most human behaviors, including emotion regulation, cannot be explained by genes alone, and certainly much of the variance in aspects of brain information processing is not genetically determined, variations in genetic sequence contributes some variance to these more complex phenomena. This conclusion is implicit in the results of studies of twins, which have revealed heritabilities of 40–70% for various aspects of cognition, temperament, and personality.

A powerful approach to understanding the nature of individual differences in emotion regulation revolves around genes because genes have an unparalleled impact on all levels of biology. In the context of disease states, particularly behavioral disorders, genes not only transcend phenomenological diagnosis, they represent mechanisms of disease. Moreover, genes offer the potential to identify at-risk individuals and biological pathways for the development of new treatments. In the case of psychiatric illness, genes appear to be the only consistent risk factors that have been identified across populations, and the lion's share of susceptibility to major psychiatric disorders is accounted for by inheritance. Although the strategy for finding susceptibility genes for complex disorders by traditional linkage and association methods may seem relatively straightforward (albeit not easily achieved), understanding the mechanisms by which such genes increase biological risk is a much more daunting challenge. How many genes contribute to a particular complex behavior or complex disease state? What genetic overlap exists across behaviors and diseases? How large are the effects of candidate genes on particular brain functions? And, perhaps most important, how does a gene affect brain information processing to increase the risk for a behavioral disorder?

Traditional Association Studies

The candidate-gene association approach has been a particularly popular strategy for attempting to answer these questions. Genetic association is a test of a relationship between a particular phenotype and a specific allele of a gene. This approach usually begins with selecting a biological aspect of a particular condition or disease, then identifying variants in genes thought to impact the candidate biological process, and next searching for evidence that the frequency of a particular variant (allele) is increased in populations having the disease or condition. A significant increase in allele frequency in the selected population is evidence of association. When a particular allele is significantly associated with a particular phenotype, it is potentially a causative factor in determining that phenotype. (There are caveats to the design and interpretation of genetic association

studies, such as linkage disequilibrium with other loci and ancestral stratification, that are beyond the scope of this review and that have been discussed at length elsewhere.)

A starting point for personality or psychiatric genetics is the findings of genetic influence through traditional behavioral genetics studies such as twin studies. By comparing the similarity of monozygotic and dizygotic twins on individual differences variables such as extraversion, such studies examine the extent to which variability is due to heritable (i.e., genetic) factors or to environmental factors. These studies indicate that genetic influences shape personality traits and do so across the life span.

Evidence that personality traits are heritable suggests that genes can be identified that account for their heritability. This is the foundation for genetic association studies of behavioral phenotypes. These studies, which take a molecular genetics approach, identify candidate genes of interest and examine their relation to a phenotype. The candidate genes included in these studies are typically those related to the patterns of brain function that putatively underlie personality style.

Association Studies of Genes Involved in Emotion Regulation

Behavioral and molecular genetics approaches have been applied to two topics that are relevant to stable emotional regulatory style: personality and affective disorders. Personality refers to stable normal individual differences, many of which pertain to emotional experience and expression. Affective disorders, although more in the realm of abnormal emotional experience, can be considered examples of pathological emotion dysregulation. These disorders – which include intense and long-duration depressed, manic, or anxious emotional states – involve reduced emotional flexibility. Presumably, difficulty with modulating the frequency, intensity, or duration of affective states underlies these disorders. For example, depression is characterized by sustained sadness and unusually low-frequency, low-intensity positive affect. The genes that predispose people to experience depression therefore may constitute genetic influences on healthy emotion regulation.

Molecular genetics approaches to emotion regulation have tended to focus on polymorphisms leading to variability in neurotransmitter availability or neurotransmitter receptor function. In general, this approach has been somewhat arbitrary and probably superficial. For example, the personality dimension of extraversion, which is characterized by measures of social dominance, novelty-seeking, and reward sensitivity, has been theorized to be driven by variability in function of the dopamine (DA) system. This has encouraged studies of association of variants in DA system genes and measures of extraversion. However, it is probable that many neurotransmitter systems influence aspects of brain function that impact on the extraversion personality style.

In spite of the oversimplification, two particular systems appear to be especially relevant to questions of emotion regulation: the serotonin (5-HT) system and the DA system. 5-HT has been implicated in the generation and regulation of emotional behavior, and the manipulation of 5-HT activity has effects on behaviors such as impulsivity and aggression. The DA system plays a critical role in reward processing and has been linked to normal individual differences in reward traits as well as to disorders involving enhanced reward-seeking such as addiction.

We address both neurotransmitter systems in our review of association studies here, and we focus specifically on the 5-HT system in our discussion of imaging genetics in the remainder of the article. In addition, although we address genetic factors in both normal and abnormal individual differences in the review of association studies, we emphasize normal individual differences in our treatment of imaging genetics. This is because the conceptual foundation for imaging genetics lends itself best to first examining normal variability in neural function.

Personality

There are a number of models characterizing stable and enduring aspects of personality, each involving different dimensions of character style. We believe that a five-factor model has the most established validity and is reflective of alternative styles of emotion regulation.

Extraversion, for example, is likely to involve an approach toward goals despite setbacks and assertion that serves to modify a current situation. Consequently, studies of the genetic underpinnings of extraversion have focused on polymorphisms related to DA function because of its critical role in shaping goal-oriented appetitive behaviors. Specifically, genetic variants of DA receptor subtypes, such as the D2 and D4 receptors, which mediate the myriad neuromodulatory effects of DA, as well as the DA transporter, which facilitates the active reuptake of DA from the extracellular space, have been examined in relation to the broad trait of extraversion and to one of its facets, novelty seeking. More recently, studies have begun to examine other genes that influence broader DA and other catecholamine availability, including catechol-*O*-methyltransferase (COMT) and monoamine oxidase A (MAO-A). As recent reviews and

meta-analyses have noted, the associations between specific DA polymorphisms and complex measures of personality have been inconsistent across studies, with null findings being relatively common.

Another significant line of related research from the field of personality genetics is the examination of 5-HT subsystem polymorphisms on emotional behaviors linked to the general experience of negative affect, such as neuroticism, impulsivity, and aggression. A variant of particular interest has been a common sequence-length polymorphism in the promoter or regulatory region of the 5-HT transporter (5-HTT) gene. Numerous studies have indicated that the short (S) variant of this gene, resulting in altered 5-HT signaling, is associated with higher levels of temperamental anxiety. Other investigators have established links between variation in the 5-HT genes controlling biosynthesis, receptor function, and metabolic degradation and additional dimensional measures of negative emotionality such as impulsive aggression and suicidality. Despite some replication, these lines of investigation have also been marked by frequent null findings, with several reports, including meta-analyses, emphasizing that the ability to detect associations depends on the personality instruments used, with broad-bandwidth personality measures (e.g., extraversion) typically representing constructs that are too heterogeneous to map meaningfully on to biological systems.

Affective Disorders

The leap from studies of genetic influences on dimensional indices of normal variability in personality and temperament to studies of genetic influences on affective disorders, such as depression and anxiety, is understandable given the correlation of these indices with the symptoms of these disorders and the genetic influences on such correlations. For example, depression and the personality trait neuroticism appear to share a genetic influence, and in addition, the correlation between depression and neuroticism appears to be influenced by genetic factors. Such attempts to link polymorphisms directly with clinical syndromes have been fueled by the suggestion that genes might have a more detectable influence at the extreme, pathological ends of the emotional trait distribution. A single gene, however, is not likely to account for a complex disorder such as major depression, and major gene effects on the expression of psychiatric disorders have not been found.

Studies of genetic influences on depression and anxiety in humans have emphasized the role of genes related to 5-HT and hypothalamic–pituitary–adrenal (HPA) axis function. Both the 5-HT and HPA systems play a critical role in emotional reactivity and regulation and are thus prime candidates for studies of these mood disorders. Many candidate polymorphisms in these systems have been linked to increased risk for mood disorders. Moreover, the existence of an association has been demonstrated to be moderated by the environment. In particular, social stress, such as maltreatment during childhood or divorce in adulthood, appears to unmask genetic vulnerability for depression and anxiety.

Limitations of Behavioral Association Studies

All the findings from traditional behavioral association studies based on clinical ratings or clinical diagnosis have been inconsistent, with an impressive amount of null findings for each gene studied. In many ways, this underscores the argument that, in the context of behavior and psychiatric illness, there are only susceptibility genes and not disease genes that clearly and specifically determine affective disorders and that the genetic origins of such disorders are heterogeneous and polygenic, involving gene–gene and gene–environment interactions. Association studies have important limitations, not least of which is the long chain of events from gene function to personality or psychiatric disorder. Additional limitations include the specificity of findings to particular personality instruments, the reliance on self-report rather than observed behavior, the failure to account for developmental effects, and the difficulties of defining and examining gene-by-environment interactions.

Conceptual Basis of Imaging Genetics

The failure of the genetic association strategy to link genotype to behavioral phenotype is ultimately not surprising given the fact that genes do not encode for complex behaviors, personality style, depression, or anxiety. Gene effects on behavior are mediated by their molecular and cellular effects on information processing in brain. Thus, examining gene effects on the brain represents a critical step in understanding their ultimate contribution to variability in behavior and is a biological readout with expected greater genetic penetrance.

Because genes are directly involved in the development and function of brain regions subserving specific cognitive and emotional processes, functional polymorphisms in genes would be expected to be more strongly related to the function of these specific neural systems than to the behavioral correlates of the function of these systems. This is the underlying assumption of our investigations examining the relation between genes and neural systems assayed with

neuroimaging techniques, what we initially called imaging genomics and more recently describe as imaging genetics because this approach is used to explore variation in specific genes and not in the genome broadly. Neuroimaging techniques, especially those that are noninvasive such as MRI and EEG/MEG, typically require no more than a few minutes of subject participation to acquire substantial data sets, reflecting the acquisition of many hundreds of repeated measures of brain structure or function within a single subject. The efficiency and sensitivity of these techniques makes it possible to investigate gene effects on multiple functional systems (e.g., prefrontal, striatal, and limbic) in a single subject in one experimental session. This capacity to rapidly assay differences in brain structure and information processing with enhanced power and sensitivity places neuroimaging at the forefront of available tools for the *in vivo* study of functional genetic variation.

The protocol for imaging genetics typically involves first identifying a meaningful variation in the DNA sequence within a candidate gene. For the variant to be meaningful, it should have an impact at the molecular and cellular level in gene or protein function (i.e., be a functional variant) and the distribution of such effects at the level of brain systems involved in specific forms of information processing should be predictable. For example, a polymorphism in the gene for the 5-HT transporter that impacts the availability of synaptic 5-HT would be expected to affect amygdala function because 5-HT is important in amygdala physiology. Alternatively, recent imaging genetics studies have taken the lead in exploring the functionality of candidate variants by first describing the *in vivo* effects at the level of brain systems. As such, imaging genetics can provide the initial impetus for the further characterization of molecular and functional effects of specific candidate genes in brain systems thought to be involved in behavior. In this manner, the contributions of abnormalities in these systems to complex behaviors and emergent behavioral phenomena, possibly including psychiatric syndromes, can then be understood from the perspective of their biological origins.

Selected Overview of Imaging Genetics Findings Related to Emotional Regulation

Converging evidence from animal and human studies indicates that 5-HT is a critical neuromodulator in the generation and regulation of emotional behavior. Serotonergic neurotransmission has also been an efficacious target for the pharmacological treatment of mood disorders, including depression, obsessive–compulsive disorder, anxiety, and panic. Moreover, genetic variation in several key 5-HT subsystems, presumably resulting in altered central serotonergic tone and neurotransmission, has been associated with various aspects of personality and temperament as well as susceptibility to affective illness. However, enthusiasm for the potential of such genetic variation to affect behaviors and especially disease liability has been tempered by weak, inconsistent, and failed attempts at replication of specific associations with psychiatric syndromes.

The inability to substantiate such relationships through consistent replication in independent cohorts may simply reflect methodological issues such as inadequate control for population stratification, insufficient power, and/or inconsistency in the methods applied. Alternatively, and perhaps more important, such inconsistency may reflect the underlying biological nature of the relationship between allelic variants in 5-HT genes, each of presumably small effect, and observable behaviors in the domain of mood and emotion that typically reflect complex functional interactions and emergent phenomena. Given that the biological impact of variation in a gene traverses an increasingly divergent path from cells to neural systems to behavior, the response of brain regions subserving emotional processes in humans (e.g., the amygdala, hippocampus, prefrontal cortex, and anterior cingulate gyrus) represents a critical first step in their impact on behavior. Thus, functional polymorphisms in 5-HT genes may be strongly related to the integrity of these underlying neural systems and mediate/moderate their ultimate effect on behavior.

Serotonin Transporter

The 5-HTT plays an important role in serotonergic neurotransmission by mediating the reuptake of 5-HT from the synaptic cleft. In 1996, a relatively common polymorphism was identified in the human 5-HTT gene located on chromosome 17q11.1–q12. The polymorphism is a variable repeat sequence in the promoter region (5-HTTLPR), resulting in two common alleles: a short (S) variant comprising 14 copies of a 20- to 23-bp tandem repeat unit and the long (L) variant comprised of 16 copies. In populations of European ancestry, the frequency of the S allele is approximately 0.40, and the genotype frequencies are in Hardy–Weinberg equilibrium (L/L = 0.36, L/S = 0.48, and S/S = 0.16). These relative allele frequencies, however, can vary substantially across populations.

Following the identification of this polymorphism, Lesch and colleagues demonstrated *in vitro* that the 5-HTTLPR alters both gene transcription and the

level of 5-HTT function. Cultured human lymphoblast cell lines homozygous for the long allele have higher concentrations of 5-HTT mRNA and express nearly twofold greater 5-HT re-uptake compared to cells possessing either one or two copies of the short allele. Subsequently, both *in vivo* imaging measures of radioligand binding to 5-HTT and postmortem calculation of 5-HTT density in humans reported nearly identical reductions in 5-HTT binding levels associated with the short allele as observed *in vitro* (but there were also negative reports).

In their initial study, Lesch and colleagues also demonstrated that individuals carrying the short allele are slightly more likely to display abnormal levels of anxiety compared to L/L homozygotes. Since their original report, others have confirmed the association between the 5-HTTLPR short (S) allele and heightened anxiety and have also demonstrated that individuals possessing the short allele more readily acquire conditioned fear responses and develop affective illness than do those homozygous for the long allele. Recent studies using pharmacological challenge paradigms of the 5-HT system suggested that these differences in affect, mood, and temperament may reflect 5-HTTLPR-driven variation in 5-HTT expression and subsequent changes in synaptic concentrations of 5-HT. Furthermore, reduced 5-HTT availability, as putatively indexed by the 5-HTTLPR S allele, has been associated with mood disturbances, including major depression and the severity of depression and anxiety in various psychiatric disorders. Intriguingly, it appears that exposure to stressful life events moderates the impact of the 5-HTTLPR on the development of depression.

The amygdala is a central brain structure in the generation of both normal and pathological emotional behavior. Furthermore, the amygdala is densely innervated by serotonergic neurons, and 5-HT receptors are abundant throughout amygdala subnuclei. Thus, the activity of the amygdala may be uniquely sensitive to alterations in serotonergic neurotransmission and any resulting variability in amygdala excitability is likely to contribute to individual differences in emergent phenomena such as mood and temperament. The response of the amygdala, however, occurs in the context of a distributed and interconnected network of cortical and subcortical brain regions for the generation, integration, and modulation of emotional behavior. Results from a series of imaging studies suggest that the dynamic interactions of the amygdala and prefrontal cortex may be critical in regulating emotional behavior.

Although the potential influence of genetic variation in 5-HTT function on human mood and temperament was bolstered by subsequent studies demonstrating increased anxiety-like behavior and abnormal fear conditioning in 5-HTT knockout mice, the underlying neurobiological correlates of this functional relationship remain unknown. Because the physiological response of the amygdala during the processing of fearful or threatening stimuli temporally precedes the subjective experience of emotionality, the 5-HTTLPR may have a more obvious impact at the level of amygdala biology.

In 2002, our research group at the National Institute for Mental Health (NIMH) used an imaging genetics strategy with fMRI to directly explore the neural basis of the apparent relationship between the 5-HTTLPR and emotional behavior. Specifically, we hypothesized that 5-HTTLPR S-allele carriers, who presumably have relatively lower 5-HTT function and higher synaptic concentrations of 5-HT (analogous to the 5-HTT knockout mice) and have been reported to be more anxious and fearful, would exhibit greater amygdala activity in response to fearful or threatening stimuli than those homozygous for the long allele, who presumably have lower levels of synaptic 5-HT and have been reported to be less anxious and fearful (analogous to the contrasting wild-type mice).

In our initial study, normal subjects from two independent cohorts ($n = 14$ in each) were divided into equal groups based on their 5-HTTLPR genotype, with the groups matched for age, sex, IQ, and task performance. During scanning, the subjects performed a simple perceptual processing task involving the matching of fearful and angry human facial expressions. Significantly, this task has been effective at consistently engaging the amygdala across multiple subject populations and experimental paradigms. Consistent with our hypothesis, we found that subjects carrying the less efficient 5-HTTLPR S allele exhibited significantly increased amygdala activity compared to subjects homozygous for the L allele. In fact, the difference in amygdala activity between 5-HTTLPR genotype groups in this study was nearly fivefold, accounting for 20% of the total variance in the amygdala response. This initial finding suggested that the increased anxiety and fearfulness associated with individuals possessing the 5-HTTLPR S allele may reflect the relatively increased reactivity of their amygdala to relevant environmental stimuli.

Subsequently, multiple independent functional imaging studies reported identical increased 5-HTTLPR S-allele-driven amygdala reactivity in cohorts of healthy adult volunteers as well as patients with social phobia, panic disorder and depression. Moreover, we also replicated our initial finding of 5-HTTLPR S-allele effects on amygdala reactivity in a large independent cohort of adult volunteers ($n = 92$) who were

carefully screened to exclude individuals with a past history of psychiatric illness or treatment. We again observed that 5-HTTLPR S-allele carriers exhibited significantly increased right amygdala activation in response to our fMRI challenge paradigm. In addition, our latest data revealed that 5-HTTLPR S-allele-driven amygdala hyperresponsivity is equally pronounced in both sexes and independent of S-allele load. The equivalent effect of one or two S alleles on amygdala function is consistent with the original observations of Lesch and colleagues on the influence of the 5-HTTLPR on *in vitro* gene-transcription efficiency and subsequent 5-HTT availability. The absence of sex differences suggests that the increased prevalence of mood disorders in females may be related to factors other than the direct risk effect of the 5-HTTLPR S allele.

A third study from the large NIMH cohort ($N = 114$) further captured the dynamic effects of the 5-HTTLPR on genes, brain, and behavior by examining effects on brain structure and corticolimbic functional connectivity. We found that, compared to the L/L genotype subjects, S-allele carriers showed a significantly reduced gray matter volume of the perigenual anterior cingulate cortex (pACC) and amygdala. Moreover, there was a positive correlation (structural covariation) between amygdala and pACC volume, and S-allele carriers showed significantly lower structural covariation between the amygdala and subgenual anterior cingulate than did L/L individuals. This suggests that the pACC and amygdala represent a functional anatomical circuit whose development is modulated by genetic variation in the serotonergic system.

We next explored the impact of these observed structural effects on the functional interactions of the amygdala and pACC in our fMRI data set. Independent of 5-HTTLPR genotype status, we found that the amygdala and pACC were significantly functionally connected. Two distinct regions of functional connectivity were identified within the pACC: (1) a positive coupling between the amygdala and the subgenual cingulate and (2) a negative coupling between the amygdala and the supragenual cingulate. This pattern of functional connectivity is consistent with anatomical tracing studies in nonhuman primates, which have defined a feedback circuit from the amygdala to the rostral cingulate and then from the dorsal cingulate back to the amygdala. These intrinsic cingulate regions also showed strong positive connectivity with one another, suggesting that the corticolimbic feedback loop is closed via local processing within the cingulate cortex. This intrinsic cingulate connection also is consistent with anatomical studies in nonhuman primates. Remarkably, 5-HTTLPR S-allele carriers showed a significant reduction of amygdala–pACC functional connectivity compared to L/L homozygotes. This difference was most pronounced in the coupling of the amygdala and subgenual ACC. These findings suggest that a disruption of this amygdala–pACC feedback circuitry could underlie the earlier observation of increased amygdala reactivity in S-allele carriers during the processing of biologically salient stimuli. More specifically, the data suggest that the increased activity of the amygdala associated with the 5-HTTLPR S allele may reflect more a relative failure of regulation of the amygdala response than an abnormal primary response *per se*. Taken together these data show that the 5-HTTLPR genotype affects the structure and putative wiring of a core region within the limbic system thought to be crucial for anxiety-related temperamental traits and depression.

The collective results of these imaging genetics studies reveal that the 5-HTTLPR S allele has a robust effect on human amygdala structure and function as well as the functional interactions of corticolimbic circuitry implicated in both normal and pathological mood states. Significantly, the absence of group differences in age, sex, IQ, and ethnicity in each of these studies indicates that the observed effects are not likely to be due to a bias resulting from population stratification. Rather, the data suggest that heritable variation in 5-HT signaling associated with the 5-HTTLPR results in structural alterations of the amygdala and pACC, accompanied by biased amygdala reactivity and functional coupling with the pACC in response to salient environmental cues. Furthermore, the emergence of these effects in samples of ethnically matched volunteers carefully screened to exclude any lifetime history of psychiatric illness or treatment argues that they represent genetically determined biological traits that are not altered by the presence of a psychiatric illness.

In contrast to the striking imaging genetics findings of 5-HTTLPR S-allele effects on amygdala reactivity and limbic circuitry dynamics, initial attempts to link these effects on brain function with measures of emergent behavioral phenomena, namely the personality trait of harm avoidance (HA), failed to detect any significant direct relationships. Specifically, in both our initial and replication studies we did not find any significant 5-HTTLPR genotype association with subjective behavioral measures of anxiety-like or fear-related traits as indexed by the HA component of the Tridimensional Personality Questionnaire, a putative personality measure related to trait anxiety and 5-HT function. This failure to find a behavioral association is not surprising given the relatively small sample sizes of each study, that the participants were screened not to have extreme scores on psychiatric inventories, thus

limiting the power to detect probably small (e.g., 1–5%) genetically mediated differences in behavior, and the theses of this paper – namely that genes do not directly predict behavior and their effect on behavior is mediated/moderated by their effects on distinct brain circuitry.

A convergence of evidence from animal and human studies clearly demonstrates that emotional behaviors, especially those as complex as HA, are probably influenced by a densely interconnected and distributed cortical and subcortical circuitry of which the amygdala is only one component. Thus, we were compelled to examine the relationship between HA and the observed 5-HTTLPR effects on the functional connectivity of the amygdala and pACC. We reasoned that if functional uncoupling of the amygdala–pACC affective circuit underlies reported associations of the 5-HTTLPR with emotional phenotypes, functional connectivity indices between these regions should predict normal variation in temperamental trait measures related to anxiety and depression such as HA. These analyses revealed a striking result; nearly 30% of the variance in HA scores was predicted by our measure of amygdala–pACC functional connectivity. This is particularly remarkable because approximately 50% of the individual differences in HA are estimated to be due to genetic effects, thus suggesting that the 5-HTTLPR effects on the amygdala–pACC coupling may account for as much as 60% of the genetically driven variance in HA. Consistent with our previous studies, functional (or structural) measures of single brain regions (i.e., the amygdala or pACC) were of no predictive value. Thus, 5-HTTLPR-mediated corticolimbic functional connectivity alterations are manifested in anxiety-related temperamental traits, possibly reflecting inadequate regulation and integration of amygdala-mediated arousal, leading to an increased vulnerability for persistent negative affect and eventually depression in the context of accumulating environmental adversity. Although investigations of localized structural and functional abnormalities have provided insights about depression, our data underscore the importance of studying the genetic mechanisms of complex brain disorders at the level of dynamically interacting neural systems. We suggest that such relationships capture more proximally the functional consequences of neurodevelopmental processes altering circuitry function implicated in human temperament and psychiatric disorders.

Monoamine Oxidase A

To the extent that the effects of the 5-HTTLPR variant on corticolimbic development and function related to emotion processing are 5-HT-mediated, it would be expected that other genes related to 5-HT function would show similar effects on the function of this circuitry. 5-HT neurotransmission is also regulated through intracellular degradation via the metabolic enzyme, MAO-A. A common polymorphism in the MAO-A gene, resulting in a relatively low-activity enzyme, has been associated with increased risk for violent or antisocial behavior, as well as for depression and anxiety. A recent fMRI study reported that the low-activity MAO-A allele is associated with relatively exaggerated amygdala reactivity and diminished prefrontal regulation of the amygdala. The magnitude of functional coupling between these regions predicted levels of temperamental anxiety, suggesting that the genetic association between the MAO-A low-activity variant and abnormal behavior may be mediated through this circuit. Interestingly, both the 5-HTTLPR S allele and MAO-A low-activity alleles presumably result in relatively increased 5-HT signaling and exaggerated amygdala reactivity. Because the directionality of these effects are consistent with animal studies documenting the anxiogenic effects of 5-HT, the imaging genetics data provide important insight regarding the neurobiological and behavioral effects of 5-HT.

Tryptophan Hydroxylase 2

Tryptophan hydroxylase 2 (TPH2) is the rate-limiting enzyme in the synthesis of neuronal 5-HT and thus plays a key role in regulating 5-HT neurotransmission. A recent study found that a single nucleotide polymorphism (SNP) in the regulatory region of the human TPH2 gene affects amygdala function. Specifically, the T allele of the relatively frequent G(-844)T polymorphism was associated with relatively exaggerated amygdala reactivity. In contrast to previous studies of genetic effects on brain function, in which the molecular and cellular effects of the candidate variants had been previously demonstrated (e.g., 5-HTTLPR, MAO-A, COMT, and brain-derived neurotrophic factor (BDNF)), these fMRI data provide the first evidence for potential functionality of a novel candidate polymorphism. In this way, the initial identification of a systems-level effect of a specific polymorphism provides impetus for the subsequent characterization of its functional effects at the molecular and cellular level. Building on this initial imaging genetics finding (and a subsequent replication), a recent molecular study demonstrated that the G(-844)T is in strong linkage with another promoter SNP that impacts the transcriptional regulation of TPH2 and may affect enzyme availability and 5-HT biosynthesis. Such scientific reciprocity between imaging and molecular genetics illustrates how the contributions of abnormalities in candidate neural systems to complex behaviors and emergent phenomena, possibly

including psychiatric illnesses, can be understood from the perspective of their neurobiological origins.

Brain-Derived Neurotrophic Factor

BDNF is a critical peptide neurotrophic factor involved in neuronal survival, differentiation, and synaptic plasticity. BDNF plays an important role in the expression of long-term potentiation (LTP) in the hippocampus and in learning and memory in animal models. BDNF is also regulated by serotonergic signaling and is involved in the developmental sculpting of serotonergic innervation patterns. A popular theory about the efficacy of selective serotonin reuptake inhibitor (SSRI) antidepressant/antianxiety drugs is that they act via stimulation of BDNF expression. The BNDF gene contains a common functional polymorphism, a valine (Val) to methionine (Met) substitution at codon 66 (Val66Met), which is associated with abnormal intracellular trafficking and the regulated secretion of pro-BDNF, the precursor to functional BDNF. This SNP also has been associated with mood disorders and with anxious temperament. Because of its abundant expression in the hippocampus and its role in learning and memory, BDNF has been studied using neuroimaging primarily in terms of hippocampal anatomy and memory processing. Consistent with the molecular effects of the Met66 allele, studies have documented relatively impaired episodic memory in human subjects carrying this allele. Converging evidence from several imaging genetics studies using different modalities suggests that the Met66 effects on memory are mediated in part by its impact on hippocampal structure and function. Most recently, a structural imaging genetics study revealed that the BDNF Val66Met can modulate the impact of the 5-HTTLPR S allele on corticolimbic circuitry. Specifically, reduced amygdala gray matter volume typically associated with the 5-HTTLPR S allele was absent in subjects also possessing the BDNF Met66 allele. The data suggest that the Met66 allele effectively blocks the relatively increased 5-HT stimulation of BDNF-mediated changes in brain morphology associated with the 5-HTTLPR S allele. As such, this functional genetic epistasis may partially account for the reduced risk of mood disorders associated with the BDNF Met66 allele.

Imaging Genetics, Development, and Developmental Psychopathology

It is important to emphasize that the 5-HTTLPR S-allele effect on amygdala structure, reactivity, and connectivity in our studies, as well as those of others, exists in samples of healthy adult volunteers with no history of affective or other psychiatric disorders. On the one hand, this is consistent with a recent fMRI study reporting that, although amygdala hyperexcitability reflects a stable, heritable trait associated with inhibited behavior, it does not by itself predict the development of affective disorders. On the other hand, more and more evidence is accumulating that the majority of psychopathology is rooted early in life, first emerging during childhood and adolescence. Thus, it is possible that the relevance of gene effects on corticolimbic brain circuitry will be more manifest during the development of individuals predisposed to psychopathology. Moreover, it is likely that exposure to environmental stressors impacts this gene–brain pathway, which in turn increases an individual's risk of developing psychopathology. The hallmark study of Caspi and colleagues and subsequent replication studies suggest that the existence of significant stressors in the environment of individuals carrying the 5-HTTLPR S allele is necessary to further tip the balance toward the development of psychopathology. Similarly, abnormal social behavior and 5-HT metabolism have been reported in rhesus macaques with the 5-HTTLPR S-allele homolog but only in peer-reared, and thus environmentally stressed, individuals. Emerging data from studies of the 5-HTT knockout mouse implicate similar early developmental phenomena interacting with genetically driven variation in 5-HT in shaping the neurobiological landscape contributing to emotional behaviors. It is pertinent to note that in many of these examples the genetic vulnerability has manifested as a consequence of environmental stressors that occurred early in development.

This shift from normal to pathological behaviors and when during the life span this shift occurs may reflect the effects of cumulative environmental stress on brain regions, most notably the prefrontal cortex, critical in the regulation of amygdala activity. For example, repeated exposure to environmental insults before the maturation of relatively late-developing prefrontal regulatory circuits may result in further biased amygdala drive in S-allele carriers. Such relative hyperamygdala and hypoprefrontal activity has been documented in affective disorders and thus, may represent a critical pathway or predictive biological marker for the future development of psychopathology.

Recent imaging studies demonstrating altered functional coupling of the amygdala and medial prefrontal cortex during affective processing in healthy adult S-allele carriers underscored that complex dynamic interactions of the amygdala and prefrontal cortex may be critical for normal behavioral responses in

individuals possessing this risk allele. These results suggest that individual differences in indices of complex emergent behaviors, such as HA, reflect the effects of genetic variation on a distributed brain system involved in not only mediating physiological and behavioral arousal (e.g., the amygdala) but also regulating and integrating this arousal in the service of adaptive responses to environmental challenges (e.g., the prefrontal cortex).

Summary

This article has reviewed efforts to parse human variation in anxious temperament and risk for affective disorders based on polymorphisms in genes implicated in development and function of neural circuitries mediating emotional arousal and regulation. We have outlined an experimental strategy by which genetic effects on brain function can be explored using neuroimaging, namely imaging genetics. This work is just in its infancy; the number of genes explored so far is very few and the strategies for looking at gene effects in brain are relatively simplistic. The studies cited provide compelling evidence that gene effects at the level of emotional neuroscience are much more robust (i.e., penetrant) than at the level of manifest behavior. This is consistent with the conclusion that genes related to emotional disorders and temperament do not encode for behavior but, rather, impact the development and function of neural systems that mediate emotional experience and behavior. Emotional responses are complex and not the result of variation in any single gene. Future studies will emphasize the interactions of genes and the interactions of genes with the environment. This is likely to add complexity but also a finer-grained resolution to the analyses.

See also: Aggression: Hormonal Basis; Amygdala: Contributions to Fear; Anxiety Disorders; Aversive Emotions: Molecular Basis of Unconditioned Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Emotion Systems and the Brain; Emotion: Neuroimaging.

Further Reading

Bertolino A, Arciero G, Rubino V, et al. (2005) Variation of human amygdala response during threatening stimuli as a function of 5′HTTLPR genotype and personality style. *Biological Psychiatry* 57(12): 1517–1525.

Brown SM, Peet E, Manuck SB, et al. (2005) A regulatory variant of the human tryptophan hydroxylase-2 gene biases amygdala reactivity. *Molecular Psychiatry* 10(9): 805.

Caspi A, Sugden K, Moffitt TE, et al. (2003) Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* 301(5631): 386–389.

Egan MF, Kojima M, Callicott JH, et al. (2003) The BDNF val66-met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell* 112(2): 257–269.

Hariri AR, Drabant EM, and Weinberger DR (2006) Imaging genetics: Perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. *Biological Psychiatry* 59(10): 888–897.

Hariri AR, Mattay VS, Tessitore A, et al. (2002) Serotonin transporter genetic variation and the response of the human amygdala. *Science* 297(5580): 400–403.

Hariri AR and Weinberger DR (2003) Imaging genomics. *British Medical Bulletin* 65: 259–270.

Heinz A, Braus DF, Smolka MN, et al. (2005) Amygdala-prefrontal coupling depends on a genetic variation of the serotonin transporter. *Nature Neuroscience* 8(1): 20–21.

LeDoux JE (2000) Emotion circuits in the brain. *Annual Review of Neuroscience* 23: 155–184.

Lesch KP, Bengel D, Heils A, et al. (1996) Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science* 274(5292): 1527–1531.

Meyer-Lindenberg A, Buckholtz JW, Kolachana B, et al. (2006) Neural mechanisms of genetic risk for impulsivity and violence in humans. *Proceedings of the National Academy of Sciences of the United States of America* 103(16): 6269–6274.

Neumeister A, Konstantinidis A, Stastny J, et al. (2002) Association between serotonin transporter gene promoter polymorphism (5HTTLPR) and behavioral responses to tryptophan depletion in healthy women with and without family history of depression. *Archives of General Psychiatry* 59(7): 613–620.

Pezawas L, Meyer-Lindenberg A, Drabant EM, et al. (2005) 5-HTTLPR polymorphism impacts human cingulate-amygdala interactions: A genetic susceptibility mechanism for depression. *Nature Neuroscience* 8(6): 828–834.

Plomin R, Owen MJ, and McGuffin P (1994) The genetic basis of complex human behaviors. *Science* 264(5166): 1733–1739.

Viding E, Williamson DE, and Hariri AR (2006) Developmental imaging genetics: Challenges and promises for translational research. *Developmental Psychopathology* 18(3): 877–892.

Genomic Disorder and Gene Expression in the Developing CNS

C M Paczkowski and J Chun, The Scripps Research Institute, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

The central nervous system (CNS) is composed of billions of neurons with trillions of synaptic connections specifically designed for large amounts of information processing through parallel representational neural networks. The level of diversity required for higher order human cognition is, in large part, genetically regulated and, for this reason, highly susceptible to genomic instability resulting in neurological dysfunction. Cellular diversity is regulated by molecular mechanisms at multiple levels: DNA, RNA, and protein. This article focuses on the level of DNA – specifically, how genomic changes influence gene expression in the normal brain and in human genomic disorders.

Genomic Regulation of CNS Development

Development of the CNS requires an array of cellular processes to produce a functioning brain, including proliferation, programmed cell death, differentiation, migration, axonal/dendritic outgrowth, and synaptogenesis. A large number of genes are involved in the transition from neural stem cell to progenitor cell to postmitotic neuron, and in the age of microarrays and genome sequencing, new genes are constantly being discovered. Genomic and environmental regulation of gene expression determines the ultimate fate of individual neurons and their integration into functioning networks. Within the highly interconnected brain, each neuron has the potential to exert far-reaching effects on numerous other cells. Therefore, altered gene expression within even a single brain cell can impact nervous system development and function.

Neural stem cells lining the surface of the embryonic neural tube proliferate and give rise to lineage-committed neural progenitor cells, specifying neural patterning (see Figure 1). A large number of proneural transcription factors are involved in directing the fate of neural progenitor cells. In addition, epigenetic mechanisms (discussed below) are important in selectively silencing genes, committing cells to neuronal or nonneuronal lineages. As stem and progenitor cells divide, genomic instability can produce daughter cells with distinct genotypes.

Upon exiting the cell cycle, neurons and glia migrate away from their proliferative zone, guided by soluble and cell surface molecular cues. A host of factors can influence the subsequent maturation of neural populations, including those intrinsically encoded within the genome as well as an orchestrated milieu of activity-dependent intra- and intercellular signaling pathways.

Cell death is a prominent aspect of brain development, with substantial overproduction of neurons that are subsequently culled. Two waves of cell death occur over the course of nervous system maturation, the first during neurogenesis and the second during synaptogenesis, which together shape the final organization of the mature brain. Neuroproliferative cell death is likely a means of eliminating new cells with undesirable genotypes, resulting from a host of potential variables. Later in development, postmitotic cell death occurs as neurons are forging synaptic contacts. Competition, largely for limited neurotrophic factors, permits the survival of only a subset of neurons. The high rate of cell death in the developing CNS suggests that there is a selection process whereby neurons with more advantageous phenotypes (based on a combination of physiological environment and genotype) survive to populate the mature brain.

Genomic Changes That Alter Gene Expression

Multiple changes at the level of the genome have been shown to alter gene expression in neural cells. Genomic changes that occur during neurogenesis give rise to populations of cells with distinct genotypes. Altered genomes are often associated with brain malfunction; however, accumulating evidence points to genomic mosaicism as a feature of normal vertebrate brains.

The most extreme form of genomic change is aneuploidy – the loss and/or gain of whole chromosomes. Aneuploidy can occur as a result of missegregation during mitosis, through events such as lagging chromosomes, supernumary centrosomes, and nondisjunction (Figure 2(a)). With several hundred genes on each chromosome, loss or gain of a single chromosome leads to widespread changes in cellular gene expression. Changes in gene expression that result from aneuploidy include gene dosage effects and allelic variation. Increases or decreases in gene dosage for genes expressed on the aberrant chromosome (*cis* effects) and on other chromosomes (*trans* effects) have been reported in aneuploid neurons. Allelic variation results in the gain or loss of one parental allele, altering the balance of allele-specific gene expression. Allelic variation is due to genetic imprinting, expression of recessive mutated alleles (as seen in loss of tumor

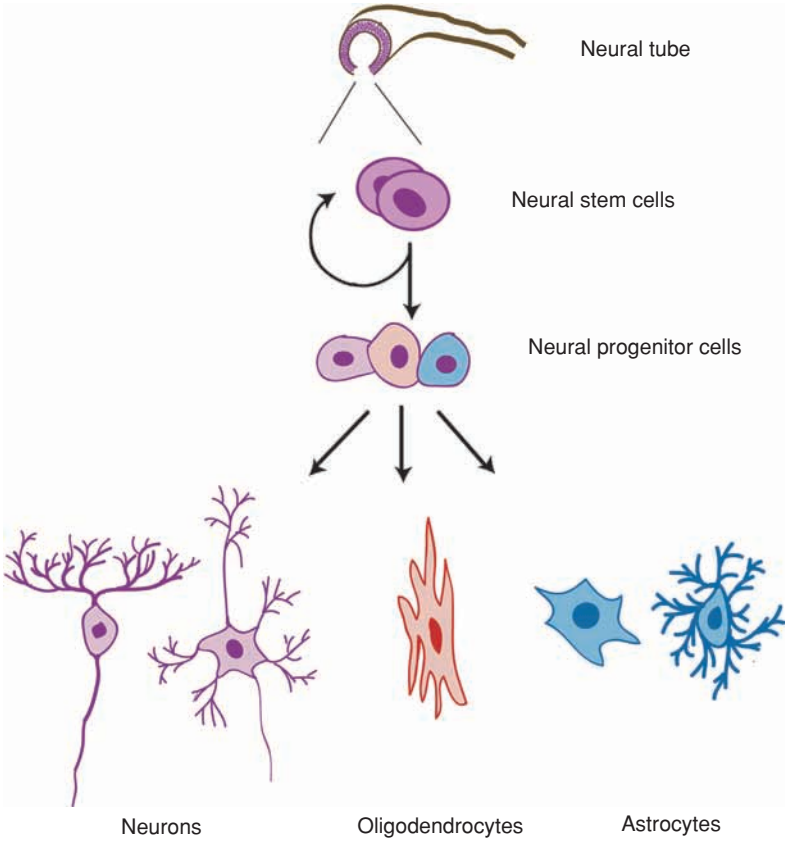

Figure 1 Process of neurogenesis. Neural stem cells dividing within the neuroepithelium of the neural tube give rise to lineage-committed neural progenitor cells. Neural progenitors differentiate into multiple diverse cell types that constitute the mature nervous system.

suppressor genes in aneuploid cancer cells), and polymorphisms. This degree of genetic diversity will affect cellular function through changes in any number of signaling pathways and, in turn, influence nervous system development. Given the range of possible functional changes produced by aneuploidy, the implications of this recent finding extend to normal neuronal diversity, behavioral diversity (as seen in monozygotic twins with exactly the same genome yet individual personalities), and neuropsychiatric diseases.

Retrotransposons, primarily those within the long interspersed nuclear element 1, or LINE1 (L1), can cause genomic mutations by disrupting gene coding regions or noncoding promoters/regulatory elements (Figure 3). L1 protein binds to untranslated L1 mRNA and escorts it back to the nucleus where the L1 endonuclease nicks a single strand of DNA and re-inserts itself through reverse transcription. Both germ line and somatic transposition events have been noted. L1 retrotransposons have been recently reported as a mechanism for genomic alteration in neurons.

Genetic imprinting is a mechanism that can regulate gene expression without altering the DNA sequence, yet the specific modifications are heritable. A classic example that affects CNS function is Prader-Willi syndrome, in which partial deletion of chromosome 15 results in mono-allelic expression of a subset of genes, causing mental retardation. Specific genomic imprinting modifications include methylation of CpG sites within a gene or promoter region and histone modifications which are encoded during DNA replication and thus passed on in proliferating cells. Genetic imprinting likely contributes to other brain functions, as there are estimates of 100–200 known imprinted genes in humans. This mechanism plays an important role in determining maternal or paternal monoallelic expression of many genes involved in brain development.

Genomic Disorders Affecting CNS Development

Down Syndrome

Down syndrome (DS), or trisomy 21, is perhaps the best-known genomic disorder affecting the CNS. It is characterized by stereotypical craniofacial abnormalities, moderate to severe mental retardation, and an

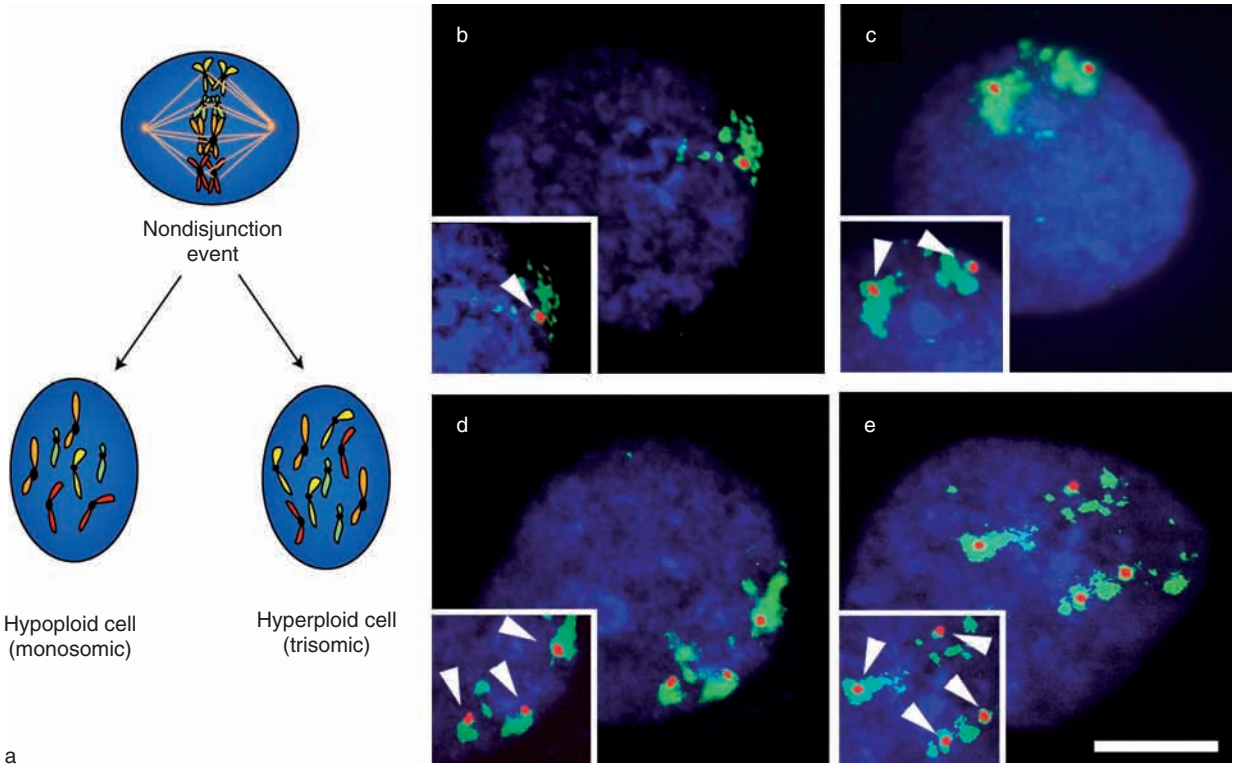

Figure 2 Aneuploidy is one mechanism for altering the genome of individual neurons. (a) Missegregation events, such as nondisjunction, can result in daughter cells that have lost (monosomy) or gained (trisomy) chromosomes (yellow chromosome). Here, genetically distinct mature brain cells from an adult human cortex are shown to have different copy numbers of chromosome 21 (b–e). Nuclei are visualized with DAPI and chromosome 21 is detected with a whole chromosome-specific paint (green) and a locus-specific point probe. Arrowheads denote distinct chromosomes. Error bar = 10 μm.

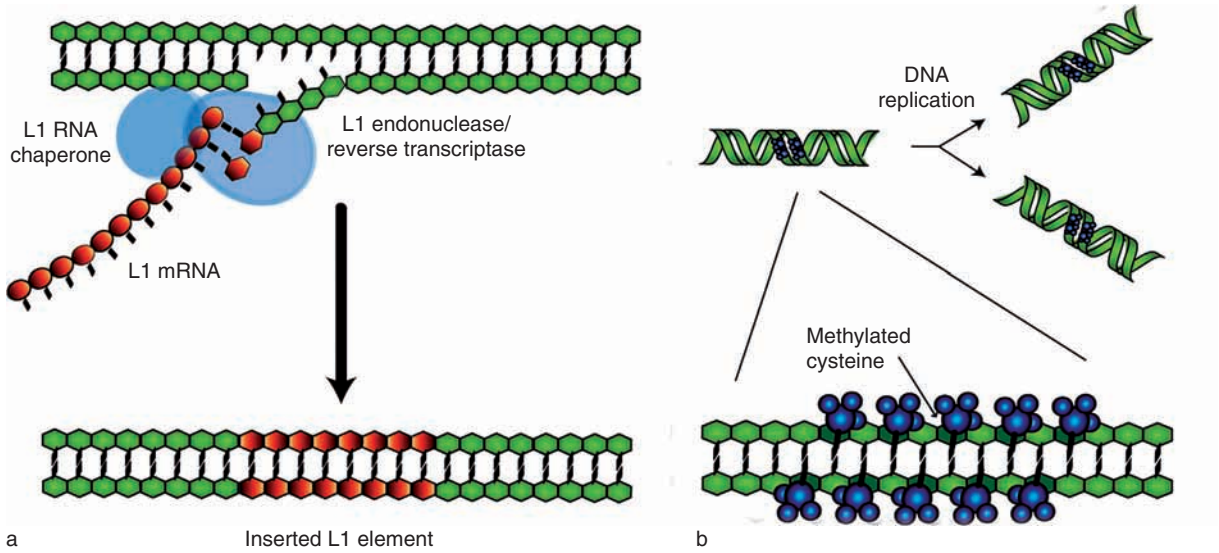

Figure 3 Intrachromosomal mechanisms can also alter the genome of individual neurons. (a) LINE1 (L1) retrotransposons randomly insert their sequence into the genome. L1 proteins escort L1 mRNA into the nucleus, nick a single strand of DNA, and reverse transcribe the L1 sequence. The cell's replication and repair machinery incorporate the new sequence into the genome. (b) Methylation of cysteine nucleotides, particularly in CG repeat sequences, silences gene transcription. This epigenetic event is reproduced during DNA replication; thus, it is permanently imprinted on the genome.

increased risk for leukemia and cardiovascular disease. Patients with DS often live into the forth or fifth decade of life, during which they prematurely develop Alzheimer's disease (AD)-like amyloid plaques and neurofibrillary tangles. Edwards (trisomy 18) and Patau (trisomy 13) syndromes are other major trisomic disorders compatible with postnatal life, and they similarly show pronounced mental retardation and physical dysmorphism.

Much research has focused on defining the genotype-phenotype relationship in DS. Chromosome 21 contains roughly 360 known genes, and the main effects of trisomy are thought to be related to the 1.5-fold increase in gene dosage for genes on that chromosome. Gene expression studies on trisomic cells from DS individuals find that most genes on chromosome 21 are overexpressed, although the pattern of altered gene expression depends on tissue type and developmental stage. It has been noted that none of the features of DS are unique to DS, suggesting a shared etiology in disorders with overlapping symptoms. An example is the possible relationship that has been reported between DS and AD in which elevated gene dosage for the *APP* (amyloid precursor protein) gene on chromosome 21 has been implicated in amyloid plaque formation, although data linking trisomy with AD is lacking. Most phenotypic features do not appear to follow this single-gene/single-phenotype pattern, suggesting that *trans*-regulation of genes on other chromosomes could result in the disruption of many common signaling pathways. Such loss of genetic homeostasis, produced by aneuploidy in brain cells, could produce general neuronal dysfunction and result in mental retardation, common to all genomic brain disorders.

Chromosomal Deletion Disorders

A long list of chromosomal disorders in which partial deletions of various chromosomes exist: Miller-Dieker syndrome (del 17p13), Angelman and Prader-Willi syndromes (del chr 15q11–12; see previous discussion), DiGeorge syndrome (del 22q11), Cri-du-chat syndrome (partial del 5p-), and Wolf-Hirshhorn syndrome (partial del 4p-), among others. Invariably, these chromosome deletion syndromes are associated with brain malformation and/or mental retardation, supporting the idea that systemic genomic abnormalities disrupt general nerve cell function and brain development.

Turner syndrome (X0) is an intriguing disorder that results in females with normal verbal intelligence but impaired numerical and spatial cognition. In addition, social skills and emotion recognition are poorly developed in these patients, possibly related to structural abnormalities of the amygdalae. Since normal XX cells are known to inactivate one X chromosome early in development, much research has focused on gene expression in Turner syndrome. The phenotype is attributed to haploinsufficiency of genes with biallelic expression on chromosome X, representing up to 20% of the genes on that chromosome. It is not known how these biallelic genes are regulated in normal females and males, and to what degree biallelic expression is required in various cell types.

Fragile X Syndrome

Fragile sites are unstable chromosomal regions that are prone to double-strand breaks and rearrangements. The most common genomic disorder associated with a fragile site is the inherited disorder fragile X syndrome, in which a CGG trinucleotide repeat causes methylation-mediated silencing of the *frm1* gene on chromosome X. The gene product, FMRP, is a mRNA-binding protein that appears to traffic specialized mRNAs to the dendrites of neurons necessary for synaptic plasticity. CGG repeat length expands during germ line transmission and the full phenotype of fragile X syndrome does not manifest until there are more than 200 repeats. Mental retardation (IQs ranging from 20 to 60) is the main clinical phenotype in fragile X syndrome, although other symptoms can include autistic behaviors, anxiety disorder, attention deficit-hyperactivity disorder, and seizures. In addition to the predominant germ line mutations, repeat length instability has been reported in somatic tissues, including the developing brain, indicating that the phenotype may partially depend on neural genotypes distinct from the inherited genotype.

Ataxia-Telangiectasia

Ataxia-telangiectasia (A-T) is the best-known disorder resulting from disruption in a DNA repair pathway. The disorder is characterized by cancer predisposition and radiation sensitivity as well as progressive neurodegeneration. A mutation in the *atm* gene (A-T mutated) has been linked to the disorder. ATM phosphorylates multiple substrates involved in cell cycle regulation and genomic stability, playing an important role in genome surveillance. Loss of ATM function has been proposed to result in an accumulation of neurons with genomic abnormalities, leading to the progressive loss of neurons. Interestingly, mice lacking *atm* have an increase in aneuploid brain cells, suggesting that aneuploidy is one form of genomic instability associated with loss of genome surveillance mechanisms. It is notable that both defects in DNA repair and aneuploidy are also associated with medulloblastoma formation. Other

DNA repair disorders with neurological phenotypes include Nijmegen-breakage syndrome, Cockayne's syndrome, and xeroderma pigmentosa.

Mosaic Variegated Aneuploidy

Mosaic variegated aneuploidy (MVA) is a rare disorder characterized by mosaic aneuploidy in multiple somatic tissues, with the requirement of at least three distinct aneuploidies for diagnosis. Up to 65% of cases of MVA show premature centromere division, leading to a high frequency of different trisomies. In other cases, chromosome missegregation and mitotic checkpoint failure are suggested as mechanisms of aneuploidy. The gene *BUB1B*, which encodes the spindle checkpoint protein BUBR1, has been implicated in some forms of this disease. Rates of aneuploidy in blood lymphocytes and fibroblasts have been reported ranging from less than 10% to more than 80%, but no information is known about the effect of this high rate of aneuploidy on gene expression or cell function. Neurologically, the disorder is associated with microcephaly, mental retardation, and myoclonic seizures, presumably related to high levels of aneuploid mosaicism in the brain, similar to other chromosomal disorders.

Neuropsychiatric Disorders

It has been hypothesized that sporadic neuropsychiatric disorders may be the result of as-yet-undetected genomic changes. Evidence for this includes Tourette's syndrome, which has been linked to mutations in microRNAs that regulate gene expression; epilepsy, which has an increased risk associated with most chromosomal disorders, including DS; idiopathic mental retardation, in which *de novo* microdeletions have been identified in several cases; and neurodegenerative diseases, in which single gene mutations account for only a small minority of cases. Since many neuropsychiatric disorders present with clinical symptoms that overlap with known genomic disorders, it is not unreasonable to postulate that undetected genomic abnormalities may be present in these patients.

Recently, fragile sites have been implicated in some neuropsychiatric disorders. FRA13A is a fragile site that maps to chromosome 13q12–14, a region flanked by tumor suppressor genes and containing the neurobeachin gene *NBEA*. In humans, haploinsufficiency for NBEA has been linked to sporadic autism, suggesting that instability of FRA13A may be a contributing factor in some cases of autism. Interestingly, the unstable nature of the FRA13A site may be related to a high number of LINE1 elements within the region. Another known fragile site, FRA6E, maps to chromosome 6q26, which includes the *PARK2* locus. Mutations in *PARK2* have been associated with autosomal recessive early parkinsonism, and it is compelling to speculate that genomic instability of FRA6E could be a factor in sporadic forms of Parkinson's disease.

Genotype-Phenotype Variability within Genomic Disorders

The clinical phenotype of most genomic disorders is highly variable, especially regarding cognitive ability; however, the basis for this clinical presentation is unclear. Genomic mosaicism in the CNS is a compelling mechanism for explaining at least some degree of the phenotypic variability seen in neuropsychiatric disorders. Genomic mosaicism produced by multiple mechanisms (e.g., aneuploidy, genetic imprinting, and retrotransposons) exists within all human brains. The precise function of this mosaicism in normal and diseased brain is an area of active research, and will likely have repercussions for our understanding of many of the genomic diseases discussed in this section. Future research in this area will provide further insight into the link between genomic mosaicism and neurological phenotype.

Acknowledgment

The authors would like to acknowledge NIH grant number MH076145.

See also: Alternative Splicing in the Nervous System; Gene Expression Regulation: Activity-Dependent; Gene Expression Regulation: Chromatin Modification in the CNS; Gene Expression Dysregulation in CNS Pathophysiology; Genetic Influence on CNS Gene Expression: Impact on Behavior; Plasticity and Activity-Dependent Regulation of Gene Expression.

Further Reading

Barlow C and Treuner K (2005) DNA instability in the brain: Survival of the 'fittest' *Nature Medicine* 11: 474–475.

Blaschke AJ, Weiner JA, and Chun J (1998) Programmed cell death is a universal feature of embryonic and postnatal neuroproliferative regions throughout the central nervous system. *Journal of Comparative Neurology* 396: 39–50.

Callier P, Faivre L, Cusin V, et al. (2005) Microcephaly is not mandatory for the diagnosis of mosaic variegated aneuploidy syndrome. *American Journal of Medical Genetics A* 137: 204–207.

Kingsbury MA, Yung YC, Peterson SE, Westra JW, and Chun J (2006) Aneuploidy in the normal and diseased brain. *Cellular and Molecular Life Sciences* 63: 2626–2641.

McKinnon PJ (2004) ATM and ataxia telangiectasia. *EMBO Report* 5: 772–776.

Muotri AR and Gage FH (2006) Generation of neuronal variability and complexity. *Nature* 441: 1087–1093.

O'Doherty A, Ruf S, Mulligan C, et al. (2005) An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes. *Science* 309: 2033–2037.

Pearson CE, Nichol Edamura K, and Cleary JD (2005) Repeat instability: Mechanisms of dynamic mutations. *Nature Review Genetics* 6: 729–742.

Rehen SK, Yung YC, McCreight MP, et al. (2005) Constitutional aneuploidy in the normal human brain. *Journal of Neuroscience* 25: 2176–2180.

Roper RJ and Reeves RH (2006) Understanding the basis for Down syndrome phenotypes. *PLoS Genetics* 2: e50.

Schwab M (2006) Fragilome: Determining the role of fragile sites in human disease. *Cancer Letters* 232: 1–3.

Sebat J, Lakshmi B, Troge J, et al. (2004) Large-scale copy number polymorphism in the human genome. *Science* 305: 525–528.

Spilianakis CG, Lalioti MD, Town T, Lee GR, and Flavell RA (2005) Interchromosomal associations between alternatively expressed loci. *Nature* 435: 637–645.

Temple S (2001) The development of neural stem cells. *Nature* 414: 112–117.

Genomics of Brain Aging: Apolipoprotein E

D K Lahiri and B Maloney, Indiana University School of Medicine, Indianapolis, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Dementia is rapidly becoming a common diagnosis in the current aging population, and the numbers are expected to rise exponentially in coming years. Alzheimer's disease (AD) alone now affects over 4.5 million people in the United States. In addition, millions more are currently affected by vascular dementia, Lewy body disease, and frontotemporal dementia (FTD). Therefore it is critical to understand potentially modifiable risk factors and early disease markers and the application of new disease-specific diagnostic tools and treatment modalities for these age-related dementias. In the present review, we focus on the current state of knowledge and directions of research on the role of apolipoprotein E (ApoE), encoded by the *APOE* gene, on the major dementias, particularly AD.

ApoE has a major role in cholesterol transport and plasma lipoprotein metabolism. It is also an important cholesterol transporter in the brain, being involved in neuronal repair, dendritic growth, maintenance of synaptic plasticity, and anti-inflammatory activities. Cholesterol regulates amyloid production and its deposition. ApoE is associated with both plaques and tangles, which are the two major pathological hallmarks of AD. Importantly, plaque formation seems to require the presence of ApoE. It is essential for amyloid deposition in the β-amyloid protein precursor (APP) Val717 $\rightarrow$ Phe (V717F) transgenic model of AD, suggesting that the role of ApoE in AD is to increase the yield and/or toxicity of amyloid plaques. *APOE* is known to have three alleles, $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$. The risk of AD increases in accordance with genotype such that $\varepsilon 4 > \varepsilon 3 > \varepsilon 2$. The $\varepsilon 4$ allele has also been associated with anomic aphasia in FTD, while $\varepsilon 2$ may be a risk factor for FTD in general, with the $\varepsilon 2/\varepsilon 4$ genotype conferring greatest risk for primary progressive aphasia.

However, multiple factors in addition to ApoE play an important role in neurological pathogenesis, making ApoE only one of several important contributors. In transgenic mice, for example, *APOE* $\varepsilon 4$ expression causes neuropathology and behavioral deficits, but it is unclear whether the mechanism involves modulation of amyloid levels. ApoE is clearly an important and robust genetic risk factor for AD and other aging-related dementias. Perhaps the most important question in the area of age-related disorders concerns how ApoE accelerates dementia pathogenesis. In this regard, AD would serve as a model for ApoE's overall role, as this is currently the most extensively characterized relationship.

ApoE Expression in Different Tissues and Cell Types and during Aging

ApoE mRNA is most abundant in the liver, followed by the brain, where it is synthesized and secreted primarily by astrocytes, oligodendrocytes, activated microglia, and ependymal layer cells. The synthesis and secretion of ApoE also occur in some human neurons, in which ApoE is induced by brain injury. Central nervous system (CNS) neurons express ApoE at lower levels than do astrocytes. APOE protein and mRNA are detected in cortical and hippocampal neurons in humans and in transgenic mice expressing human ApoE under the control of the human *APOE* gene promoter. ApoE is expressed in primary cultured human CNS neurons and in several human neuronal cell lines. Cerebrospinal fluid (CSF) ApoE is a component of CSF high-density lipoproteins (HDLs), the major lipoprotein type in the CSF.

APOE expression is upregulated in astrocytes during aging. Macrophages are the major source of *APOE* among various cell types studied. Two distal enhancers that specify *APOE* gene expression in macrophages were reported in transgenic mice with specific constructs of the human *APOE/APOC1/APOC1′/APOC4/APOC2* gene cluster. A multienhancer-1 (ME-l) element consists of a 620-bp sequence located 3.3 kb downstream of the *APOE* gene and the multienhancer-2 (ME-2) element comprises a 619-bp region located 15.9 kb downstream of the *APOE* gene and 5.9 kb downstream of the *APOC1* gene. Both ME-1 and ME-2 elements drive *APOE* gene expression in astrocytes of the olfactory bulb, cerebellum, and hippocampus in transgenic mice.

ApoE and AD

Alzheimer's disease, the most common form of dementia among elderly, is a progressive, degenerative disorder of the central nervous system. The consequence of the disease is loss of memory; this and other cognitive and behavioral symptoms progressively impair function in the activities of daily living. Neurochemically, AD is characterized by selective neuronal cell death, the presence of amyloid deposits

in the core of neuritic plaques, and the formation of neurofibrillary tangles in the brain of afflicted individuals. Plaque and tangles are deposited extraneuronally and intraneuronally, respectively. Notably, ApoE protein has been reported to co-localize with both of these neuropathological lesions of AD.

An increase in ApoE mRNA was reported in the hippocampus in AD and in rats after entorhinal cortex lesioning. In both AD and the lesion model, investigators observed a shift in the location of astrocytes containing prevalent ApoE mRNA, from the neuropil to regions with densely packed neurons. The increased abundance of ApoE mRNA in astrocytes close to neuron cell bodies could be indicative of lipid uptake in regions where neurons are degenerating or where synaptic remodeling is taking place.

Protein Isoforms of ApoE and Their Role in AD

ApoE is a polymorphic 299-amino-acid protein with a molecular mass of 34 200 Da. The human *APOE* gene consists of at least three alleles ($\varepsilon 2$, $\varepsilon 3$, $\varepsilon 4$) encoding three protein isoforms. The three common ApoE isoforms, $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$, are all products of the same gene, with three alleles at a single gene locus. These isoforms differ from one another at residues 112 and 158. ApoE $\varepsilon 3$ has Cys112 and Arg158, ApoE $\varepsilon 4$ has arginine at both positions, and ApoE $\varepsilon 2$ has cysteine. In almost all populations, the $\varepsilon 3$ allele accounts for the vast majority of the *APOE* gene pool (typically ~70–80%) and the $\varepsilon 4$ and $\varepsilon 2$ alleles account for 10–15% and 5–10% of the population, respectively.

Among these isoforms, the $\varepsilon 4$ allele of *APOE* constitutes a major susceptibility factor for the development of the familial and sporadic forms of late-onset AD. Individuals with one copy of *APOE* $\varepsilon 4$ show a higher risk compared to those without the allele, and two copies of the *APOE* $\varepsilon 4$ further increase the risk of AD. At the protein-binding level, ApoE binds β-amyloid (Aβ) peptide with high avidity, and the three isoforms differ in their binding properties, which may contribute to the isoform-specific effects of ApoE on Aβ deposition. Animal studies suggest that ApoE participates in β-amyloid deposition in an isoform-specific manner. At drug treatment level, AD patients with ApoE $\varepsilon 4$ responded poorly to treatment with tacrine compared to patients with ApoE $\varepsilon 2$ and ApoE $\varepsilon 3$.

ApoE Isoforms and Familial AD-Linked Mutations

Familial AD (FAD) is caused by mutations in several genes, including *APP*, *PSEN1*, and *PSEN2*, which encode β-amyloid protein precursor, presenilin 1 (PS1), and presenilin 2 (PS2), respectively. Mutations in *APP*, *PSEN1*, and *PSEN2* account for only a small percentage of AD patients, even among those who have early age of onset and a family history of the disease. The role of ApoE isoforms in the pathogenesis of AD is visible by their interaction with FAD mutations, as well as in Down syndrome. In those cases when FAD is associated with a mutation in APP or increased load of APP, the presence of an *APOE* $\varepsilon 4$ allele, in comparison to *APOE* $\varepsilon 3$, significantly reduces the age of onset of AD. In contrast, most mutations in PS1 and PS2, which together account for the bulk of the known FAD families, are only weakly affected by the *APOE* $\varepsilon 4$ allele. This finding cannot be explained by assuming a ceiling effect on the age of onset. It would appear that the *APOE* $\varepsilon 4$ allele affects a particular pathway that is either removed or saturated by the PS1/2 mutations. It is then most likely that ApoE influences the pathogenesis of AD by altering the pathology, although cell culture studies indicate that there are no changes in Aβ production because of co-expression of any of the three *APOE* alleles.

Neurobiological Mechanisms: ApoE Isoform Status Influences Risk of AD by a Relatively Selective Effect on Episodic Memory

The neurobiological mechanisms linking the $\varepsilon 4$ allele risk of AD are beginning to be understood. Previous histopathological studies report that the $\varepsilon 4$ allele is associated with the number of neuritic plaques and neurofibrillary tangles, suggesting that the effect of the $\varepsilon 4$ allele on disease may be mediated by an increase in the rate that AD pathology accumulates, but not all studies have found this association, and others suggest that the $\varepsilon 4$ allele is related to other mechanisms, such as neural repair or survival. Data from the Religious Orders Study, an ongoing longitudinal clinical–pathologic study of aging and AD, were used to examine the pathologic mechanisms linking the $\varepsilon 4$ allele to the development of AD. These studies suggest that *APOE* allele status influences risk of AD by a relatively selective effect on episodic memory. Other studies suggest that the presence of a single *APOE* $\varepsilon 4$ allele is associated with an increased rate of hippocampal volume loss in healthy women in their sixth decade of life, but is not related to any detectable memory changes.

Neuropathological Association and Role of *APOE* $\epsilon 4$ Allele in the Clinical Manifestation of AD

Bennett and colleagues have tested the hypothesis that the $\varepsilon 4$ allele is associated with the clinical manifestations of AD through an association with the pathologic hallmarks of disease among 128 deceased participants in the Religious Orders Study. The effect

was present when measures of neuritic plaques, diffuse plaques, and neurofibrillary tangles were analyzed separately and when measures of five different cognitive systems (episodic memory, semantic memory, working memory, perceptual speed, and visuospatial ability) were analyzed independently. The ε4 allele is, therefore, associated with the clinical manifestations of AD through an association with the pathologic hallmarks of AD, rather than some other mechanism (e.g., direct effect on neuronal survival).

The relationships underlying neuropathology also exist in normal individuals. A study of Aβ42 and Aβ40 concentration in individuals without AD revealed a significant association between the *APOE* ε4 allele and greater decline in the amount of Aβ42 in cerebrospinal fluid than in non-*APOE* ε4 individuals. This sharper decrease in Aβ42 in CSF is likely due to greater precipitation of Aβ42 as amyloid plaque in brain parenchyma.

ApoE Genotype/Age Interaction Affects Cerebral Glucose Metabolism

Positron emission tomography (PET) studies in healthy elders showed that the *APOE* ε4 allele is disruptive to the regional cerebral metabolic rate of glucose metabolism (rCMRGlu), perhaps via an interaction with the aging process. Another study assessed whether this interaction occurs in AD patients. ApoE groups were comparable for age, gender, age at onset, and disease duration. Correlations between age and rCMRGlu confirmed a negative relationship for both groups. For example, lower rCMRGlu was found within the frontal and cingulate areas for *APOE* ε4 carriers as compared with the noncarriers. Additionally, a significant ApoE/age interaction was detected in the frontal and anterior cingulate cortex, with the *APOE* ε4 carriers having a steeper regression slope with respect to the noncarriers. These results indicate that age-related regional rCMRGlu decreases within the frontal and anterior cingulate areas may be more severe in AD patients carrying the *APOE* ε4 allele.

Increased ApoE mRNA Level in AD Subjects Carrying the *APOE* ε4 Allele

In addition to functional alterations associated with the ε4 isoform, gene expression of *APOE* has been assessed in the brains of AD subjects according to *APOE* genotype. Increased ApoE mRNA was reported in AD subjects carrying the ε4 allele. ApoE mRNA level was significantly higher in the AD group than in the control group ($p < 0.05$). ApoE mRNA level in the AD group with ε4 was also significantly higher than that in the control group with ε4 ($p < 0.05$). However, this is controversial. For example, in another study, levels of ApoD, but not ApoE, were reported to increase in the CSF and hippocampus of AD patients as a function of inheritance of the APOE ε4 allele.

APOE ε4 Still May Not Be a Major Susceptibility Factor for All Racial and Ethnic Groups

Even considering all of the aforementioned evidence, the relationship of the ApoE isoform to AD risk is not necessarily clear-cut. To understand the role of the ε4 allele in different racial groups, Bennett and colleagues presented data from a geographically defined biracial population on the south side of Chicago. The effect of the ε4 allele on the risk of AD differed strongly between African-Americans and Caucasians. For example, among Caucasians the presence of the ε4 allele was associated with a 2.73-fold increase in risk, while among African-Americans there was no significant increase in risk. Other studies also report a weaker (or no) association of the ε4 allele among racial and ethnic minorities, although this remains controversial. Other genetic and environmental risk factors may be important for the development of AD among racial and ethnic minorities in particular.

The Regulatory Region of the *APOE* Gene

In addition to the differential risk related to ε2/ε3/ε4 alleles, additional risk for AD may be related to transcriptional activity of the *APOE* gene. ApoE mRNA has been shown to be significantly more abundant in AD patients than in a control group, regardless of ApoE isoform genotype. In addition, reports have been made of correlation related to the presence of promoter single-nucleotide polymorphisms (SNPs), ApoE expression levels, and risk of AD.

The gene encoding ApoE in humans and in the most common model system (mice) resides in a well-conserved cluster of genes. The human *APOE* gene is on chromosome 19, while the mouse homolog, *Apoe*, is on mouse chromosome 7 (Figure 1). The genes for ApoE of mouse and human both consist of a 5′-untranslated region (5′-UTR) divided between two exons, three coding-sequence exons, and a terminator (Figure 2). All introns in the human gene are longer than those in the mouse.

Structure of the 5′-Flanking Region of the Gene for ApoE

The proximal region of the ApoE gene promoter is highly conserved in the mouse, rat, and human; the relative position of the 'TATA box' and the two copies of the 'GC box' are identical. Such characterization of the ApoE promoter is important for understanding the mechanism of ApoE gene expression and its regulation.

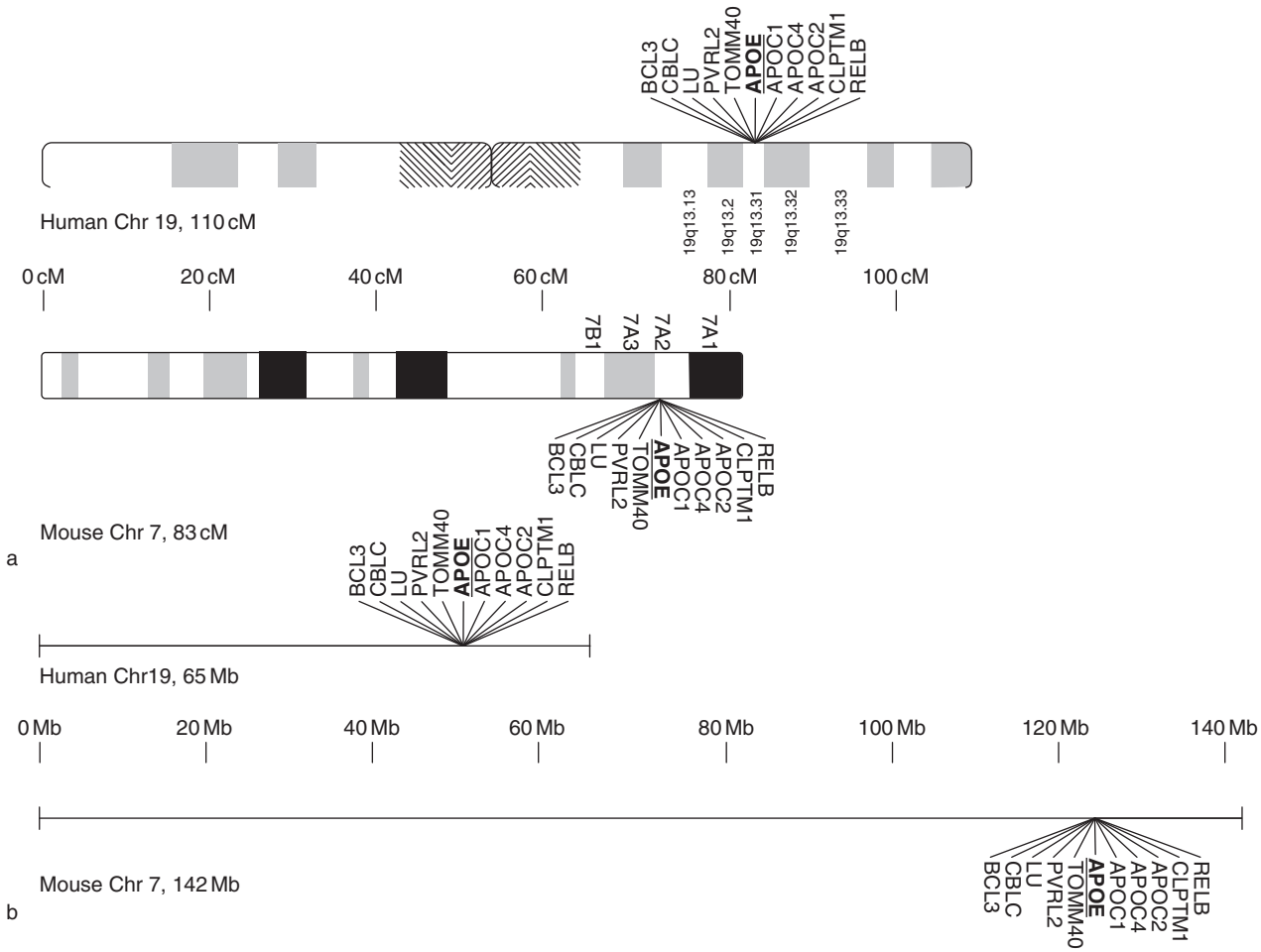

Figure 1 Human and mouse chromosome regions near their respective (*APOE*, *Apoe*) genes encoding apolipoprotein E. Databases for human and mouse genomes were searched for genes in the region of each species' apolipoprotein E gene on the respective chromosomes. Homologous genes were mapped according to genetic (cM, centimorgan; 1 cM ~ 1 Mb) and physical (Mb, megabase pair) distances. Mouse maps are presented in 'reversed' orientation for clarity of comparison. (a) Genetic maps of human chromosome 19 and mouse chromosome 7. Maps are to the same centimorgan scale. (b) Physical maps of human chromosome 19 and mouse chromosome 7. Maps are to the same megabase pair scale.

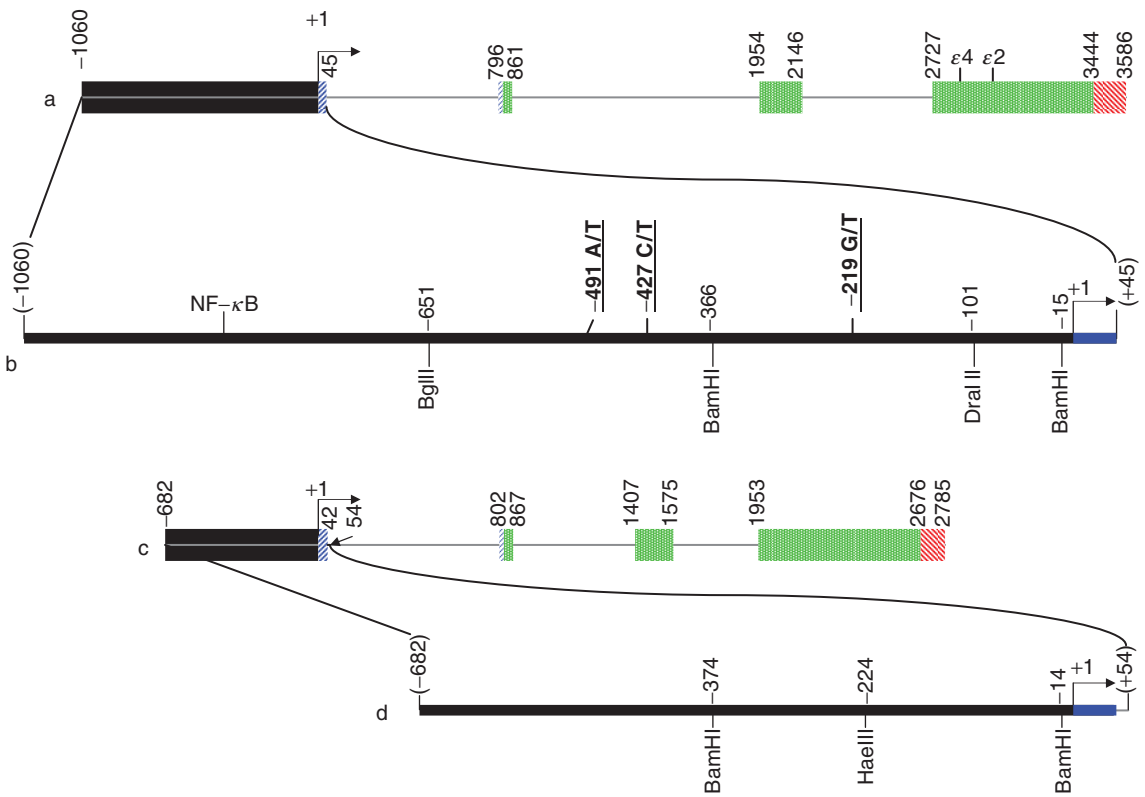

Figure 2 Human (a, b) and mouse (c, d) genes that encode apolipoprotein E, and 5′-flanking region structures. (a) Map of the human *APOE* gene, including introns, exons, terminator region, and 5′-flanking region. The 5′-untranslated region is in blue, coding sequence exons are indicated in green, and the 3′-untranslated region/terminators are in red. Locations of the sites of the ε2 and ε4 alleles are indicated. (b) Map of the human *APOE* 5′-flanking region, including locations of the polymorphisms at −491, −427, and −219. (c) Map of the mouse *Apoe* gene, including introns, exons, terminator region, and 5′-flanking region. The 5′-untranslated region is in blue, coding sequence exons are indicated in green, and the 3′-untranslated region/terminators are in red. (d) Map of the mouse *Apoe* 5′-flanking region. Dashed vertical line indicates the 5′-untranslated region exon/intron border. Maps (a) and (c) are drawn on the same scale, and maps (b) and (d) are drawn on the same scale. See text for discussion of other sites. Adapted from Maloney B, Ge Y-W, Alley GM, and Lahiri DK (2007) Important differences between human and mouse APOE gene promoters with implications for Alzheimer's disease. *Journal of Neurochemistry* 103: 1237–1257.

Transcription control of the mouse *Apoe* (m-*Apoe*) gene has been studied. Promoter activity of a 729-bp 5′-flanking region, which is located 729-bp upstream from the translation initiation codon (ATG), has been thoroughly examined. The 729-bp region was cloned into the reporter chloramphenicol acetyl transferase (*CAT*) gene in a promoterless vector (pBLCAT3). Transfection studies of the m-*Apoe* promoter in rat glial C6 and neuronal PC12 cell lines suggest that the 729-bp *Apoe* region (from position −675 to +54) is functionally active in these cell types. The serial deletion analysis indicates that the 278-bp promoter region (from −224 to +54) has the highest, and the 68-bp region (from −14 to +54) the lowest, activity on the reporter gene in neuronal and astrocytic cell lines. These results indicate that the 151-bp region (from −374 to −224) has a negative regulatory effect on the reporter gene. DNA–protein interaction studies reveal that the 68-bp region binds a specific transcription factor(s) in PC12 nuclear extracts. These functional studies suggest that mouse *Apoe* gene can also be expressed in neuronal cells in addition to the astrocytic cells.

Promoter Polymorphisms of the *APOE* Gene

In addition to the highly characterized isoform ε2/ε3/ε4 polymorphisms of *APOE*, three promoter SNPs have been reported with some level of associated AD risk. These polymorphisms reside at −491 (A/T), −427 (C/T), and −219 (G/T). Of the three, −491A/T has the most agreement in the literature as representing a risk factor (as −491A) for AD independent of *APOE* ε4, although some reports only indicate that it increases the effect of the ε4 allele and it has been reported that this polymorphism has no effect in some ethnic groups. In addition, it has been suggested that −491A is associated with increases in subjective memory loss. The −491T variant has been found to correlate with improved response to the drugs atorvastatin and benzafibrate. The −427C/T polymorphism has usually

been reported to offer no significant association with AD. The −219G/T polymorphism has been shown to have no significant effect on age-related neurodegeneration, to be an independent risk factor (as −219T) for AD, or to only be a risk factor in association with *APOE* ε4. Interestingly, the −219T polymorphism has been suggested to have an association with coronary artery disease and open-angle glaucoma. The incompleteness of effect for these polymorphisms strongly indicates that environmental interaction plays an important role in any effects ApoE might have on aging and age-related pathology.

Interaction of Environmental Factor(s) with the *APOE* Gene Promoter

The *APOE* promoter region contains important transcription factor sites through which the promoter interacts with various cellular factors. There are some functionally tested DNA sites in the *APOE* promoter for possible binding of interesting transcriptional factors. Nuclear factor-kappa B (NF-κB) and activator protein-2 (AP-2) sites present within the *APOE* promoter are mainly responsible for environmental factor-mediated *APOE* gene expression. DNA sequence analysis predicts the presence of putative sites for activator protein-1 (AP-1) and AP-2, and a functional site for the NF-κB transcription factor (TF) in the human *APOE* promoter. Gel electrophoretic mobility analysis has confirmed the presence of one predominant NF-κB site within the human *APOE* promoter. The significance of the presence of these important TF sites in the promoter is evident from a finding that cytokine/Aβ-induced glial activation, NF-κB, and ApoE play critical roles in neuroinflammatory process seen in AD. Notably, some of the aforementioned *APOE* TF sites are also present in the *APP* promoter region, suggesting a common regulatory pathway for both of these important genes in the pathogenesis of AD. Thus, these sites can potentially be effective drug targets to control the expression of the *APOE* gene.

In addition to acute environmental–gene promoter interactions, the possibility may exist that *APOE* gene expression may be subject to alteration in a LEARn (latent early-life associated regulation) fashion. LEARn begins by transient exposure of an organism to a stimulus such as lead (Pb), which alters DNA methylation and/or oxidation patterns. However, unlike conventional Pb-toxicity, there is no acute effect of the environmental exposure. Instead, the potential of gene expression is latently altered, only to be triggered later in life by a second event, such as the widespread changes in levels of transcription and inflammatory factors seen in the transition from maturity to old age. Mechanistically, a gene's primary DNA sequence would not be altered either by the original stimulus nor by the later trigger. Instead, methylation (or oxidation) would produce a latent 'somatic epitype' that would ultimately determine gene expression levels. The LEARn model provides a mechanism for explanation of the difficulty in tracking environment–gene interactions.

Interaction of the Aβ Peptide with the *APOE* Promoter

Treating cultured cells with the Aβ peptide increased *APOE* promoter activity in a dose-dependent manner. Interestingly, Aβ-mediated *APOE* promoter activity is regulated by the NF-κB site, as evident from other functional studies, and this interaction can be altered by salicylate. Thus, the NF-κB site could potentially be a therapeutic target to control the increased *APOE* promoter activity seen in AD (Figure 3).

Effects of Statins on *APOE* Gene Expression: Role in Regulating Inflammation and Cell Proliferation/Differentiation

The effects of statins have been tested on the activation of transcription factors, which are known to regulate inflammation and cell proliferation/differentiation. Simvastatin, atorvastatin, and lovastatin inhibited the binding of nuclear proteins to both the NF-κB and AP-1 DNA consensus oligonucleotides in human endothelial and vascular smooth muscle cells. Hydroxymethyl glutaryl coenzyme A (HMG-CoA) reductase inhibitors upregulate the inhibitory IκB-α protein levels in endothelial cells and decrease c-Jun mRNA expression in smooth muscle cells. Thus, HMG-CoA reductase inhibitors downregulate the activation of transcription factors NF-κB, AP-1, and hypoxia-inducible factor-1α. These results corroborate the concept that statins have anti-inflammatory and antiproliferative effects that are relevant in the treatment of atherosclerotic diseases. Current studies are directed to examine the role of those transcriptional factors' sites within the *APOE* promoter that would be affected by the HMG-CoA reductase inhibitors, which otherwise regulate inflammatory transcription factors that are involved in age-related disorders such as AD.

Interaction of ApoE with Other Risk Factors: Oxidative Stress, Inflammation, and Cholesterol Transport

Besides genetic etiological effects on AD, ApoE interacts with other risk factors, such as oxidative and inflammatory processes, nutrition, and environment. ApoE has been shown to possess anti-inflammatory

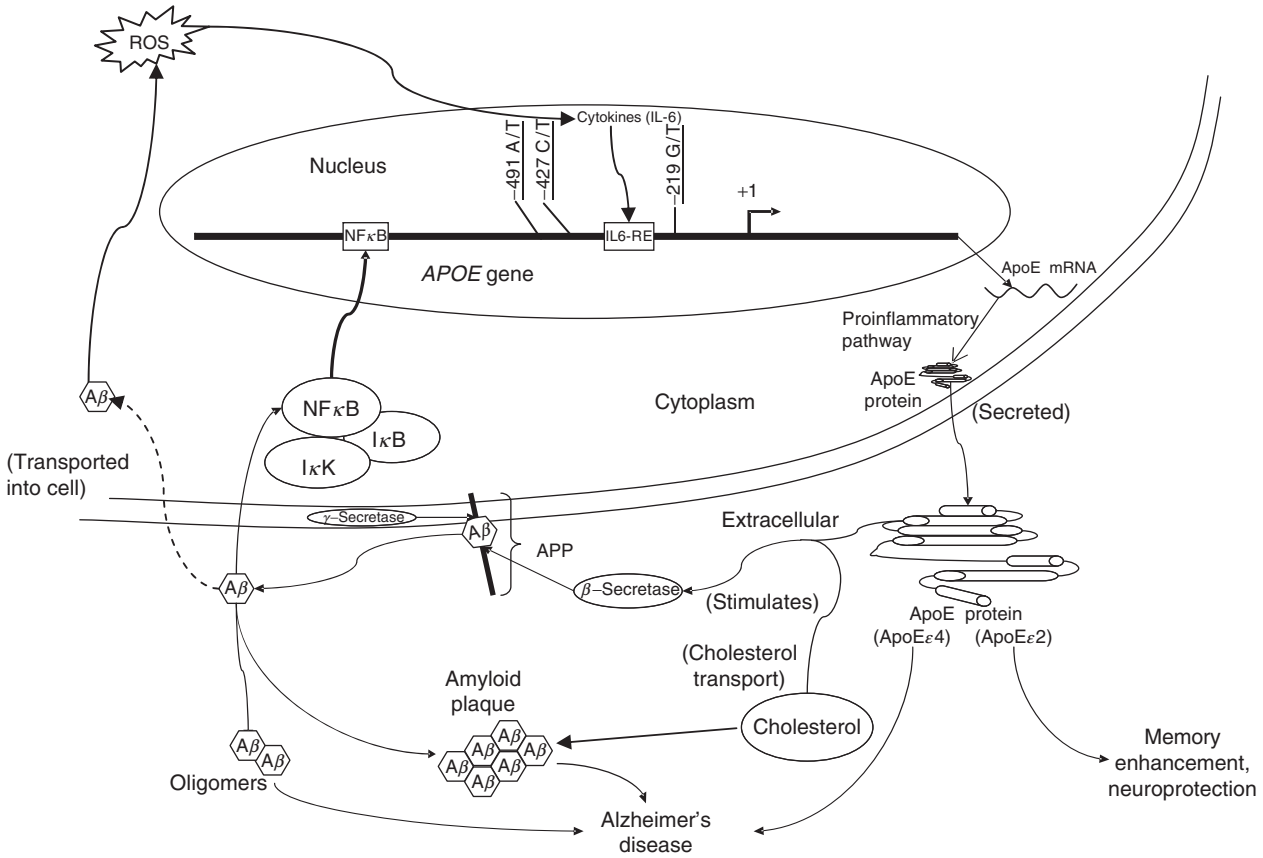

Figure 3 Schematic of interactions of the apolipoprotein E (ApoE)-encoding *APOE* gene with β-amyloid (Aβ) and with environmentally driven risk factors that lead to Alzheimer's disease pathogenesis. The *APOE* regulatory region contains binding sites for several transcription factors, such as nuclear factor-kappa B (NF-κB), and inflammatory factors, such as interleukins (IL-6 RE; interleukin-6 response element), which modulate expression of the *APOE* gene. In addition, promoter single-nucleotide polymorphisms influence ApoE levels and risk for age-related neurodegenerative disorders such as Alzheimer's disease. Modulation of this regulatory region alters *APOE* transcription levels, thereby changing ApoE protein synthesis. ApoE acts as a cholesterol transporter protein and carries cholesterol into the brain. Cholesterol modulates β-amyloid protein precursor (APP) processing, as it increases the activity of one of the key secretase enzymes, β-secretase, which results in an increased Aβ production. Aβ plays several roles in the neuropathology: Aβ transported into the cell contributes to reactive oxidative species (ROS), inducing cytokine response. Oligomeric Aβ causes neurodegeneration, and excessive Aβ results in amyloid deposits, which also contain cholesterol and ApoE. Furthermore, Aβ activates NF-κB, which in turn goes to the nucleus and activates the *APOE* gene. In addition to this pathway, the *APOE* ε2 allele has a positive effect on memory and protective effect against Alzheimer's disease. However, allele ε4 may have a role in neurotoxicity and promotes fibrillogenesis, as deduced from cell culture and animal studies. All of these events lead to neurodegeneration and accelerate Aβ deposition. Thus, a vicious loop of interacting Aβ, *APOE*, cholesterol, and APP forms a cascade of events, culminating in Alzheimer's disease pathogenesis. From Lahiri DK, Sambamurti K, and Bennett DA (2004) Apolipoprotein gene and its interaction with the environmentally driven risk factors: Molecular, genetic and epidomiological studies with Alzheimer's disease. *Neurobiology of Aging* 25: 651–660.

properties in a variety of models. During normal aging, the brain undergoes several alterations, including reduced neuronal functioning and alterations in glia homeostasis. An increase in inflammatory signaling has also been reported in some studies of the aging brain, with inflammation potentially mediating age-related changes in the brain. Age-related alterations in ApoE expression have been shown to correlate with age-related changes in the cytokine interleukin-1β (IL-1β) in the brain of aging rats. In another study, ApoE expression was not altered during normal brain aging, but a relationship may exist between ApoE and IL-1β transcription in the cerebral cortex.

The role of cholesterol and oxidative stress is important in the pathogenesis of AD. Our current understanding of the role of oxidative stress in AD has resulted in the potentially beneficial use of vitamin E or selegiline, in delaying the clinical progression of functional loss and the time of nursing home placement. Cerebral oxidative damage is a feature of aging and increases in neurodegenerative diseases such as AD. In aged male and female mice, it has been shown that lack of ApoE is associated with

oxidative damage in the brain, with results more pronounced for male mice, especially those fed a tocopherol-deficient diet, and that supplementation with tocopherol is not effective in reversing arachidonic acid oxidation in APOE(−/−) mice.

Summary

Apolipoprotein E is encoded by a gene located on the long arm of human chromosome 19 (19q13). The protein is a vital part of cholesterol transport and functions in neurite outgrowth. ApoE mRNA and protein are found predominantly in astrocytes within the CNS. ApoE appears in three isoforms, ε2/ε3/ε4. The *APOE* ε4 allele is associated with the development of late-onset familial AD as well as late-onset sporadic AD. *APOE* ε4 has a gene dose effect on the risk and age of onset of AD. At the gene expression level, there is a report of a high expression of *APOE* mRNA in the brains with sporadic AD. Two promoter polymorphisms, −491A/T and −219G/C, have been repeatedly shown to influence AD risk. At the protein level, ApoE has been found to co-localize with one of the typical neuropathological lesions of AD, specifically the extracellular amyloid deposits.

Herein we have reviewed *APOE* as a risk factor for AD (as the most common form of aging-related neurodegeneration), both from the standpoint of the ε4 allele and potentially influential promoter polymorphisms, in general populations, among different racial and ethnic groups. The effect of the ε4 allele on annual decline in episodic memory is significantly stronger than is its effect on decline in other cognitive systems and at baseline, but there is an equal and opposite effect of the ε2 allele. Thus, the *APOE* allele status may influence risk of AD by a relatively selective effect on episodic memory. In addition, allele status alters cerebral glucose metabolism and influences age-related hippocampal volume loss.

APOE operates in age-related neurodegeneration, such as in AD, in a complex set of feedback loops that are influenced by both intrinsic and environmental factors (Figure 3). Briefly, the *APOE* gene regulatory region is a target for transcription factors such as NF-κB and inflammatory factors such as interleukins. Single-nucleotide polymorphisms in the promoter influence ApoE levels. Modulation of this regulatory region by differences in transcription factor/cytokine binding would alter *APOE* transcription levels, changing ApoE protein synthesis levels. Different responses to the same levels of transcription factors according to promoter polymorphism would also alter *APOE* mRNA levels. ApoE acts as a cholesterol transporter protein and carries cholesterol into the brain. Cholesterol modulates APP processing, since it increases the activity of one of the key secretase enzymes, β-secretase, which results in increased Aβ production. Aβ transported back into the cell would contribute to reactive oxidative species (ROS), inducing cytokine response. Oligomeric Aβ would cause gross neurodegeneration, and excess extra cellular Aβ accumulation results in amyloid deposits, which also contain cholesterol and ApoE. Furthermore, Aβ activates NF-κB, which in turn activates the *APOE* gene. Thus a vicious loop of interacting Aβ, *APOE*, cholesterol, and APP forms a cascade of events culminating in AD pathogenesis.

The proximal region of the *APOE* gene is highly conserved in the mouse, rat, and human; the relative positions of the 'TATA box' and the two copies of 'GC box' are identical. The human *APOE* promoter DNA contains putative sites for AP-1, AP-2, and NF-κB transcription factors. The NF-κB site interacts with Aβ, and this interaction can be blocked by salicylate. The promoter region also contains inflammatory response transcription factor sites such as IL-6 response element binding protein. Moreover, some of these transcription factor sites should provide suitable targets for drug development in treating devastating aging-related diseases such as AD.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: Transgenic Mouse Models; Alzheimer's Disease: MRI Studies; Amyloid: Vascular and Parenchymal; Animal Models of Alzheimer's Disease; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Gene Expression in Normal Aging Brain; Genomics of Brain Aging: Twin Studies.

Further Reading

Artiga MJ, Bullido MJ, Frank A, et al. (1998) Risk for Alzheimer's disease correlates with transcriptional activity of the APOE gene. *Human Molecular Genetics* 7: 1887–9182.

Bales KR, Du Y, Holtzman D, et al. (2000) Neuroinflammation and Alzheimer's disease: Critical roles for cytokine/Abeta-induced glial activation, NF-kappaB, and apolipoprotein E. *Neurobiology of Aging* 21: 427–432.

Biere AL, Ostaszewski B, Zhao H, et al. (1995) Co-expression of beta-amyloid precursor protein (betaAPP) and apolipoprotein E in cell culture: Analysis of betaAPP processing. *Neurobiology of Disease* 2: 177–187.

Corder EH, Saunders AM, Strittmatter WJ, et al. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261: 921–923.

Evans DA, Bennett DA, Wilson RS, et al. (2003) Incidence of Alzheimer disease in a biracial urban community: Relation to apolipoprotein E allele status. *Archives of Neurology* 60: 185–189.

Graff–Radford NR, Green RC, Go RC, et al. (2002) Association between apolipoprotein E genotype and Alzheimer disease in African American subjects. *Archives of Neurology* 59: 594–600.

Harwood DG, Barker WW, Ownby RL, et al. (2002) Apolipoprotein E polymorphism and cognitive impairment in a bi-ethnic community-dwelling elderly sample. *Alzheimer Disease & Associated Disorders* 16: 8–14.

Hebert LE, Scherr PA, Bienias JL, et al. (2003) Alzheimer disease in the US population: Prevalence estimates using the 2000 census. *Archives of Neurology* 60: 1119–1122.

Lahiri DK (2004) Apolipoprotein E as a target for developing new therapeutics for Alzheimer's disease: Characterization of the regulatory regions of the apolipoprotein gene. *Journal of Molecular Neuroscience* 23: 225–233.

Lahiri DK, Alley GM, Ge YW, et al. (2002) Functional characterization of the 5′-regulatory region of the murine apolipoprotein gene. *Annals of the New York Academy of Sciences* 973: 340–344.

Lahiri DK and Maloney B (2006) Genes are not our density: The somatic epitype bridges between the genotype and the phenotype. *Nature Reviews Neuroscience* 7: doi:10.1038/nrn2022-C1.

Lahiri DK, Maloney B, Basha MR, Ge Y-W, and Zawia NH (2007) How and when environmental agents and dietary factors affect the course of Alzheimer's disease: The 'LEARn' model (Latent Early Associated Regulation) may explain the triggering of AD. *Current Alzheimer Research* 4: 219–228.

Maloney B, Ge Y-W, Alley GM, and Lahiri DK (2007) Important differences between human and mouse APOE gene promoters with implications for Alzheimer's disease. *Journal of Neurochemistry* 103: 1237–1257.

Mayeux R, Small SA, Tang M, et al. (2001) Memory performance in healthy elderly without Alzheimer's disease: Effects of time and apolipoprotein-E. *Neurobiology of Aging* 22: 683–689.

Namba Y, Tomonoga M, Kawasaki E, et al. (1991) Apolipoprotein E immunoreactivity in cerebral amyloid deposits and neurofibrillary tangles in Alzheimer's disease and kuru plaque amyloid in Creutzfeldt–Jakob disease. *Brain Research* 541: 163–166.

Pitas RE, Ji ZS, Weisgraber KH, et al. (1998) Role of apolipoprotein E in modulating neurite outgrowth: Potential effect of intracellular apolipoprotein E. *Biochemical Society Transactions* 26: 257–262.

Refolo LM, Malester B, LaFrancois J, et al. (2000) Hypercholesterolemia accelerates the Alzheimer's amyloid pathology in a transgenic mouse model. *Neurobiology of Disease* 7: 321–331.

Saunders AM, Trowers MK, Shimkets RA, et al. (2000) The role of apolipoprotein E in Alzheimer's disease: Pharmacogenomic target selection. *Biochimica et Biophysica Acta* 1502: 85–94.

Slooter AJ, Houwing-Duistermaat JJ, van Harskamp F, et al. (1999) Apolipoprotein E genotype and progression of Alzheimer's disease: The Rotterdam Study. *Journal of Neurology* 246: 304–308.

Wang JC, Kwon MD, Shah P, et al. (2000) Effects of *APOE* genotype and promoter polymorphisms on risk of Alzheimer's disease. *Neurology* 55: 1644–1649.

Wilson RS, Schneider JA, Barnes LL, et al. (2002) Effect of apolipoprotein E $\varepsilon 4$ allele status on decline in different cognitive systems and development of Alzheimer's disease over a 6-year period. *Archives of Neurology* 59: 1154–1160.

Genomics of Brain Aging: Nuclear and Mitochondrial Genomes

A Prigione and G Cortopassi, University of California at Davis, Davis, CA, USA
C Ferrarese, University of Milan-Bicocca, Monza, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Over the past few years, significant advances in the study of brain genomics have been made due to the advent of DNA microarray technology, which has become the selected method for large-scale gene expression assessment. Gene microarrays provide a powerful new approach for addressing the complexity of brain function and aging-related processes, as they allow expression monitoring of thousand of genes simultaneously, permitting a more global analysis of genomic data.

The technique makes use of DNA sequences from a library of synthetic oligonucleotides; these are arrayed on a chip at a very high density. After extraction and labeling, nucleic acids from experimental samples are hybridized on the array chip and the amount of emitted fluorescence can be monitored, allowing a whole-genome expression analysis. The observations are not limited to single molecules, as the results can be integrated into biochemical pathways or classes of related genes using biostatistical tools. In fact, the significantly modified genes are automatically assigned to specific functional categories, such as molecular bioprocesses, chromosomes, cell compartments, and relevant pathways based on databased gene product functions (Figure 1).

Several recent endeavors have applied microarray technology to the study of the transcriptional modifications associated with the physiology and pathology of brain aging in both mammalian models and humans. Additionally, microarrays have been used to define the effects of mutations in the mitochondrial DNA (mtDNA) on nuclear gene expression. Mutations in the mitochondrial genome accumulate over time and mitochondria appear to play a central role in the process of brain aging. Herein we review the trends in gene regulation in the aging brain, the changes in mitochondrial mutations with age, and the direct consequences of mitochondrial mutations on the transcriptional profile of the nuclear genome.

Gene Expression Profiles of the Aging Brain

The gift of long life is counterbalanced by the price of aging. Aging is a progressive and inevitable process that involves all the organs, and mental decline is a broadly experienced status consequent to physiological brain aging. The same changes experienced by brain cells during normal aging are exacerbated in vulnerable populations of neurons in neurodegenerative disorders. Neurodegenerative diseases, most frequently manifested in the form of Alzheimer's disease (AD) and Parkinson's disease (PD), affect an increasing proportion of the aging population and are considered as the pathological side effects of normal aging. Genetic, biological, and environmental factors all contribute to physiological and pathological brain aging, and thus understanding the transcriptional response in both cases might help in creating therapeutic strategies to fight pathology and in the definition of interventions to slow down physiological cognitive decline.

Several hypotheses have been proposed regarding the nature of neural death during aging, and the three main mechanisms involve oxidative stress, calcium dysregulation, and apoptosis. Oxidative stress occurs when the cellular redox homeostasis is imbalanced by an increase in pro-oxidant species, known as reactive oxygen species (ROS), without a corresponding increase in the protective antioxidant counterparts. ROS include a variety of diverse chemical species and, although partially generated by cytosolic enzyme systems such as NADPH oxidase or cytosolic reactions as dopamine catabolism, a vast majority of intracellular ROS are derived from the mitochondria. Mitochondrial ROS (superoxide anion, hydrogen peroxide, and hydroxyl radicals) are generated as by-products of mitochondrial oxidative phosphorylation (OXPHOS), which provides ATP via the reaction of hydrocarbons with oxygen. The free-radical theory of aging, formulated in the mid-1950s, hypothesizes that life span is limited by the accumulation of cellular damage from reaction with endogenous oxygen radicals. This endogenous formation of oxidative species may be spontaneous or triggered by specific external stimuli and can increase as humans age, eventually leading to irreversible and deleterious cellular modifications. Indeed, once ROS are generated, they can react with DNA, proteins, and lipids, and damage them. All organs and proteins can potentially be modified by oxidative stress, but certain tissues such as neurons and muscle may be particularly sensitive, as they are highly energy dependent. Energy-dependent tissues are in fact very sensitive to any mitochondrial alterations, the mitochondrion being the main cellular powerhouse and also the cellular compartment that has the greatest exposure to oxidative insults.

The 'calcium dyshomeostasis' hypothesis of brain aging proposes that early in the aging process, Ca^{2+}

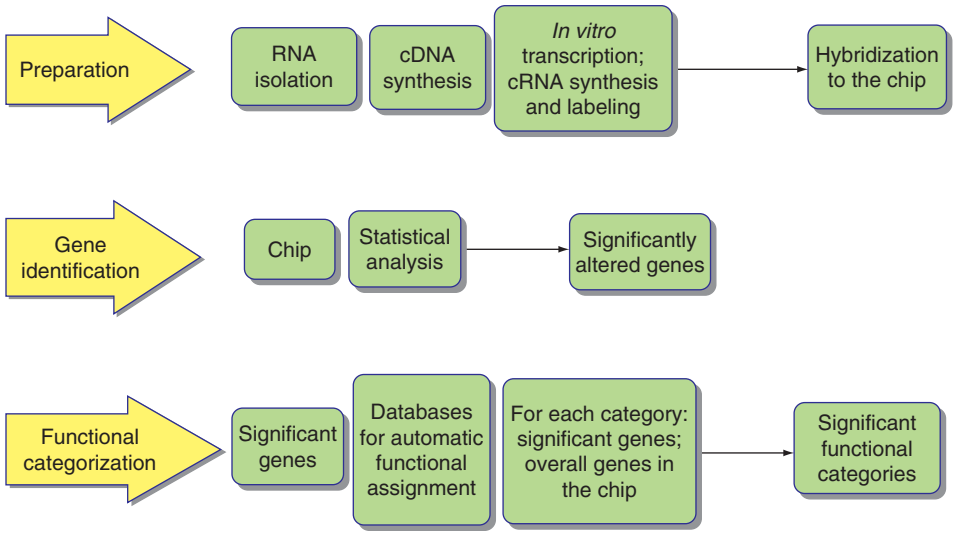

Figure 1 A summary of basic steps in microarray technology.

balance in brain cells begins to be dysregulated, influencing multiple signaling pathways modulating cellular physiology, molecular functions, and cell structures of the normal brain. The same subtle age-dependent changes in Ca^{2+} homeostasis may increase vulnerability to neurodegenerative diseases such as AD. Changes in free Ca^{2+} concentrations in various subcellular compartments act as a universal signaling system and can be modified by uptake into the endoplasmic reticulum (ER) or into mitochondria, which can function as Ca^{2+} reservoirs. Unbalanced increase of Ca^{2+} influx into neurons can also be caused by excessive activation of glutamate receptors, a process known as glutamate excitotoxicity, which may result in dendritic damage and cell death. Glutamate is the main excitatory neurotransmitter in the central nervous system (CNS) and most neurons commonly receive synaptic glutamatergic input. Several molecular and cellular changes occurring during physiological aging and disease-specific abnormalities are known to increase neural vulnerability to excitotoxicity, and glutamatergic defects are thought to underlie the development of memory loss associated with aging.

Apoptosis is the most widely studied type of programmed cell death in the CNS. It can be triggered by several upstream factors, suggesting that, even if specific gene mutations give rise to different neurodegenerative disorders, there may be a convergence of disease- and age-related pathways on the same well-known cell death cascades. Apoptosis is regulated by cysteine proteases called caspases, which trigger cell death by degrading various structural proteins. Two main pathways are responsible for the activation of the caspases that execute the apoptotic death program. The extrinsic pathway involves cell-surface signaling complexes that activate caspase 8 and downstream caspases. The intrinsic, or mitochondrial, pathway controls activation of caspase 9 by regulating the release of cytochrome *c* from mitochondria. The latter appears to be predominant in the brain and, independently from the origin of the upstream activators, mitochondria, together with the ER, regulate and modulate the entire process, playing a role in integrating cellular stress signals.

Nonetheless, whether the mechanisms of cellular death involve ROS generation, Ca^{2+} dysregulation, or activation of apoptotic factors, the study of nuclear profiling is of enormous help in understanding the common mechanisms upstream of neuronal death and the complex process of brain aging.

Trends in Animal Models

Most genomic studies of brain aging have been performed in rodent models. Aging results in a gene expression profile indicative of oxidative stress, altered Ca^{2+} regulation, apoptosis, and reduced mitochondrial function, suggesting that the three hypotheses regarding the nature of neural death may all be part of the same process, sharing mitochondrial dysfunction as a central common element. In addition, other relevant transcriptional modifications are related to inflammatory response, altered synaptic signaling, decreased protein turnover, and decreased levels of growth and trophic factors. Overall, upregulated genes can be assigned to the following categories: stress response, Ca^{2+}-related signal transduction, vesicle trafficking, myelin and lipid synthesis, inflammatory response, and glial function (Table 1).

The stress response category comprises genes involved in oxidative stress, heat shock response,

lysosomal proteases, and apoptosis. These include ROS-activated elements, such as glutathione-*S*-transferase, several heat shock factors, components of the lysosomal proteolytic system, and injury or hypoxic stress-induced elements. The stress category thus appears to play an essential role in the aging process, although most likely a secondary one consequent to cellular insults, rather than being a primary step. The putative triggers could be calcium pathway dysfunctions as well as mitochondrial modifications. In fact, upregulation of transcripts encoding Ca^{2+}-regulating pathways has been detected, including modulators of Ca^{2+}-induced release and the voltage-dependent anion channel (VDAC) that regulates Ca^{2+} flux into the mitochondria and plays a central part in apoptosis. Elevated intracellular Ca^{2+} concentration can reduce neural activity through the activation of inhibitory Ca^{2+}-dependent conductances; amplified Ca^{2+} release and influx in aged neurons may eventually lead to decreased neural responsiveness and synaptic plasticity.

Vesicle trafficking genes include those encoding SNARE (*N*-ethylmaleimide-sensitive factor attachment protein receptor) proteins that are involved in docking and fusion of synaptic vesicles to the active zone, as well as in the calcium-triggering step itself, as part of the exocytotic process at the presynaptic plasma membrane. Their increase has been proposed to be linked to the upregulation of myelin-related proteins, myelin-associated oligodendrocytic basic protein, and myelin-associated glycoprotein. The production and the secretion of myelin in oligondendrocytes may be increased in response to light-activated demyelinating processes occurring during aging. Demyelination and, most importantly, oxidative stress can then trigger a widespread neuroinflammatory response.

Neuroinflammation is mediated by microglial activation and cytokine release and can exacerbate the initial demyelination and redox status disequilibrium, creating a positive feedback. Activated transcripts included the microglia and macrophage migration factor, receptors that mediate the microglial activation pathway, inflammation markers, complement cascade genes, and major histocompatibility complex (MHC) class I antigen-presenting molecules. Both the widespread inflammatory response and the myelin production involve glial cells, and, as further confirmation, glial function genes appear to be upregulated in rodent brain aging.

Downregulated genes are mainly involved in mitochondrial function and energy metabolism, protein synthesis and turnover, synaptic plasticity, extracellular matrix (ECM) production, and neurotrophic factors. Downregulated mitochondrial transcripts include genes responsible for mitochondrial function, as well as electron transport chain genes. The overall inhibition of mitochondrial-related transcripts may be interpreted either as a primary event related to aging or, more likely, as the nuclear response to mitochondrial alterations which cause or are caused by oxidative stress. Mitochondrial dysfunction may be responsible, together with calcium dyshomeostasis, for evoking the cellular stress response; the downregulation of mitochondrial genes could then represent a nuclear attempt to shut down a process that is not working properly. The consequences would be deleterious for energy metabolism and would lead to metabolic and biosynthetic involutions. Accordingly, a decrease in protein biosynthesis and turnover is also observed.

Table 1 Global view of nuclear transcriptional changes induced by brain aging (in mammalian models and in humans) and by human mtDNA mutations

Brain aging	*mtDNA mutations*
Upregulated transcripts	Upregulated transcripts
Stress response	Stress response
Ca^{2+}-related signal transduction	Ca^{2+}-related signal transduction
Vesicle trafficking	Vesicle trafficking
Myelin and lipid synthesis	Cell cycle and cell proliferation
Inflammatory response	Glycolysis
Glial functions	
Downregulated transcripts	Downregulated transcripts
Mitochondrial function and energy metabolism	Mitochondrial function and energy metabolism
Protein biosynthesis	Protein biosynthesis
Protein turnover	Protein turnover (mtDNA deletion specific)
Synaptic plasticity	Synaptic transmission
Extracellular matrix production	Cell migration
Growth and neurotrophic factors	Oligodendrosis

Transcripts encoding the main cellular pathway responsible for cytoplasmic protein degradation, the ubiquitin–proteasome system (UPS), are in fact mainly downregulated. Interestingly, UPS inhibition is known to cause accumulation of altered or misfolded proteins in some neurodegenerative diseases (e.g., PD and AD).

Downregulated activity-dependent synaptic/neurite plasticity genes comprise genes which have been previously linked to synaptogenesis, neurite remodeling, plasticity, or memory. Reduced synaptic signaling and plasticity in neurons have been linked to disruption of memory consolidation, which is a common feature of both physiological and pathological cognitive impairment. Moreover, some of these genes are also significant components of the ECM, and their downregulation may suggest a genomically mediated ECM erosion, which has been hypothesized to be related to the demyelination process. As a further confirmation, other ECM elements are also downregulated.

Inhibition of several genes associated with either trophic or growth functions has been detected. Lack of neurotrophic factors, which are involved in neural development and rescue from excitotoxic or metabolic insults, can thus exacerbate the whole process, leading aged brain cells to death. Transcripts comprise fibroblast growth factor, glial neurotrophic factor, and factors involved in neural migration.

Similar genomic transcriptional profiles have been identified in rodent models for AD, as multiple studies have demonstrated increased expression of inflammation-associated genes, apoptosis-related transcripts, and altered mitochondrial energy metabolism genes. Thus, common central elements for physiological and pathological aging appear to be calcium homeostasis, mitochondrial dysregulation, and stress- and inflammation-mediated neural damage. So far, the main mechanism that has proved effective in retarding the aging process in rodents appears to be caloric restriction (CR). CR may also extend the maximum life span of several nonmammalian models. Explanations involve decrease in ROS production, decreased body temperature associated with a hypometabolic state, hormone signaling changes, and alterations in gene expression and protein degradation. Interestingly, CR is found to selectively attenuate the age-associated induction of genes encoding for inflammatory and stress responses.

Trends in Humans

The patterns of changes in gene expression observed in microarray experiments performed on aged human neurons have mainly reproduced the ones identified in rodents, helping to define a potential fingerprint of brain aging. The human transcriptional profiles are characterized by the induction of genes mediating stress response and repair functions, inflammatory and immune responses, myelination, and lipid metabolism. Transcripts reduced with aging include genes involved in signal transduction, synaptic function and plasticity, learning and memory processes, calcium homeostasis, protein turnover, and mitochondrial functions (Table 1).

Genomic studies of human pathological aging mainly have focused on AD. Two basic approaches are used: the comparison of AD-affected and AD-resistant brain tissues from the same individuals or the comparison of AD-affected brain structures of AD patients and age-matched controls. Although transcriptional changes of important elements of AD neuropathology have been identified, alterations shared with normal aging are also observed. Common trends are represented by the activation of apoptotic and neuroinflammatory signaling, stress response, and lipid metabolism, and the inhibition of the gene categories involved in synaptic plasticity and neurotrophic factors.

Taken together, common features of brain aging appear to be the repression of genes implicated in mitochondrial oxidative respiration, the decline in physiological neural activities, and inflammatory and glial reactions. In order to establish a cascade of events, some authors hypothesize that the primary dysfunctions occurring in aged neurons may be either calcium or mitochondrial dysfunctions, leading to cellular stress and genome damage, which in turn may compromise biosynthetic and metabolic pathways as well as systems that subserve synaptic functions and neural survival. Eventually, neural alterations might translate into widespread reactions involving glial responses, inflammation, and demyelination.

Mitochondrial Mutations and Aging

Mitochondria, in addition to their well-known role in respiration and OXPHOS, are important sites for diverse cellular functions such as fatty acid oxidation and heme, urea, amino acid, and phospholipid biosyntheses. However, as explained previously, their relevance for the aging process consists in the central role they play in ROS generation, Ca^{2+} signaling, and apoptosis.

Mitochondria, unlike other cellular compartments, possess an individual genome. mtDNA is a small 16.6 kb double-stranded circular DNA that encodes 13 respiratory chain polypeptides and 24 nucleic acids that are needed for intramitochondrial protein synthesis. mtDNA genes cannot be translated by use of the universal genetic code, hence the need for

mtDNA to encode its own ribosomal RNA and transfer RNA genes. Replication is independent of the cell cycle and can occur even in nondividing cells. The rest of the respiratory chain elements and mitochondrial proteins (approximately 900 gene products) are all encoded by the nuclear genome. Nuclear DNA is also relevant for it maintenance, as it encodes the only DNA polymerase known to be targeted to mitochondria: DNA polymerase-γ (POLG). POLG is responsible for all aspects of mtDNA synthesis, replication, and repair of mtDNA damage. It corrects mistakes made during replication by copying and proofreading the genome and it can participate in the resynthesis of DNA during the DNA repair processes. Overall, the cross-talk between nuclear and mitochondrial genomes plays an essential function in regulating mitochondrial activities.

mtDNA damage is a very likely event, given the proximity of mitochondria to an environment rich with reactive species, the lack of genomic introns and protective histones, and the presence of a single DNA polymerase. Consequently, mtDNA accumulates mutations more than 10 times faster than does the nuclear genome. Because multiple copies of mtDNA reside within each mitochondrion, and several hundred mitochondria are present within each cell, when a mutation occurs, the result is a dual population of mtDNAs – mutant and wild-type – coexisting within the same cell, a phenomenon called heteroplasmy. Moreover, mtDNA randomly segregates during mitosis, leading to different levels of heteroplasmy in different tissues. Eventually, when the amount of mutated mtDNA in a particular tissue reaches a threshold, pathological manifestations can become evident.

In the general population, mitochondrial genes in several nonreproductive (somatic) tissues accumulate mutations over time, and this event has been proposed to underlie the deleterious cellular modifications associated with aging. Somatic mtDNA mutations, in fact, may lead to energy-generation defects, increased number of harmful free radicals, or calcium dyshomeostasis, resulting in lowered cell and tissue functions characteristic of advanced age. Point mutations (single base-pair changes), deletions, and wider rearrangements of mtDNA can all occur and increase as humans age. These mutations are unevenly distributed and can accumulate clonally in certain areas, causing a mosaic pattern of respiratory chain deficiencies in tissues such as skeletal muscle, heart, and brain. A frequently observed mutation is a deletion of mtDNA of about 5 kb in length, the so-called common deletion, which preferentially occurs at much higher levels in nervous and muscle tissues. Deletions, accumulated exponentially with age, reach much lower concentrations in most tissues of healthy individuals than in patients affected by mitochondrial disorders such as Kearns–Sayre syndrome (KSS) or chronic progressive external opthalmoplegia (CPEO). However, deletions levels may come close to 5% in muscle tissue and to 50% in substantia nigra. These deletions occur as spontaneous errors in individual cells as a consequence of age, and rise in concentration over time, with each cell containing a different predominant deletion species. Age-dependent mtDNA deletions have also been linked with respiratory chain defects in muscle and substantia nigra.

More recently, it has been demonstrated that a defective POLG is correlated with an increased level of mtDNA mutations, comprising deletions and point mutations. Eliminating the proofreading activity, POLG becomes error prone, giving rise to increased somatic mtDNA mutations. Accumulation of mutations is associated with reduced life span and premature onset of aging-related phenotypes in mice models carrying defective POLG, demonstrating a causative link between mitochondrial mutations and aging in mammals. In POLG mutant mice, the increase in mutations in the mitochondrial genome is not associated with oxidative stress, but has been correlated with the induction of apoptotic markers. Accelerated development of aging through mtDNA mutations may thus occur even without redox imbalances. In the apoptotic pathway, mitochondrial dysfunction can lead to mitochondrial outer membrane permeabilization and subsequent release of cytochrome *c* in the cytoplasm, which in turn activates the key effector, protease caspase 3. However, the hypothesis of a contribution of oxidative DNA damage to aging is supported by other studies in humans and mice. In fact, the event leading to an error-prone POLG might be secondary to protein damage caused by ROS. However, the disruption of physiological mitochondrial activities appears to be a major element in the chain of events leading to aging.

Interestingly, CR, the only nutritional intervention that retards aging in animal models, has been demonstrated to reduce the mitochondrial-mediated apoptosis pathway and to delay the accumulation of mtDNA mutations in rodents. Indeed, CR reduces fiber loss and mitochondrial abnormalities in aged muscle and attenuates the increase in cytosolic cytochrome *c* and caspase 2 activity with age. As a therapeutic strategy, the possible application of CR to the human population would face ethical and practical problems, but the study of the specific mechanisms responsible for the antiaging influences of CR could lead to the development of pharmacological agents. Moreover, the demonstration that CR increases life span by decreasing the accumulation of mtDNA

mutations and reducing mitochondrial apoptosis strengthens the concept that losses of mtDNA integrity and mitochondrial function are central factors in brain aging.

Genomic Effects of Mitochondrial Mutations

Mutations in gene products expressed in the mitochondrion cause a nuclear transcriptional response that leads to neurological disease, such as Leber's hereditary optic neuropathy (LHON), neurogenic ataxia and retinitis pigmentosa (NARP), mitochondrial myopathy/encephalopathy/lactacidosis/stroke (MELAS), myoclonic epilepsy with ragged red fibers (MERRF), KSS, and CPEO. As indicated previously, the same mutations may play a causative role in the aging process as well. In order to assess mitochondrial disease nuclear genomic profiles, microarray analyses have been performed in different cell types. The results have demonstrated that a similar transcriptional profile is shared among different mitochondrial diseases, as if each mitochondrial disease and mitochondrial mutation could generate a common nuclear response (Table 1).

Upregulated transcripts encode for genes in the categories of stress response, calcium signaling regulation, cell cycle and cell proliferation, vesicle-mediated transport, and glycolysis. In the stress response category, ER- and unfolded protein response (UPR)-related transcripts are detected. This suggests that mitochondrial inhibition might be sufficient for activation of the UPR, which occurs when ER function is perturbed. Apoptosis is known to be stimulated in the presence of prolonged ER stress. Thus, mitochondrial mutations or other factors inducing a cellular stress response may eventually give rise to the mitochondrial apoptotic pathway. UPR induction may occur consequently to ER stress, which can be generated as a result of mitochondrial mutations, either by increased ROS production or by Ca^{2+} signaling perturbations caused by mitochondrial bioenergetic defects. Upregulation of transcripts related to Ca^{2+} modulation is in fact detected, including VDAC, which is thought to reside at the mitochondrial–ER interface and to control Ca^{2+} transients. Moreover, it has been shown that depletion of mitochondrial membrane potential induces both the UPR genes and cell cycle genes. Accordingly, cell cycle and cell proliferation genes are induced. As observed for the brain aging genomic profiles, vesicle trafficking genes, protein transport genes, and ER-to-Golgi transport genes appear to be induced also as a consequence of mitochondrial mutations. However, the reason for this activation is still unclear. Glycolysis transcripts, such as aldolase C, are also upregulated, and their increase may be interpreted as the nuclear attempt to restore energetic metabolism functions.

Functional clusters inhibited by mitochondrial diseases include mitochondrial ATP synthesis, protein biosynthesis, synaptic transmission and vesicular secretion, cell migration, and oligodendrogenesis. Transcripts coding for mitochondrial proteins and specifically for ATP synthesis are downregulated. The data suggest that the nucleus senses the irreversible depletion of mtDNA-encoded mitochondrial subunits and transfer RNAs, and responds by downregulating the interacting subunits that would normally form a functional complex. The downregulation of nuclear-encoded mitochondrial ribosomal subunits and OXPHOS transcripts thus appears to reinforce the mitochondrial defect initiated by the mutations. As observed in aged brain transcriptomes, deleterious modifications of energy metabolism may lead to metabolic and biosynthetic involutions, as the functional category of protein biosynthesis appears to be inhibited. Moreover, members of the vesicular secretion pathway, or proteins that are transported by the vesicular secretion pathway, are also decreased. In addition to their function in vesicular secretion, some gene products of this category have also been shown to be involved in cellular migration, as cellular migration is dependent upon the process of vesicular secretion, both for secretion of cell migratory signals into the ECM as well as for the interpretation of them. Additionally, UPR and ER stress induction are known to inhibit protein synthesis and vesicular secretion. Oligodendrocyte-specific transcripts are found downregulated in mitochondrial diseases. Oligodendrocytes are especially dependent on lipid synthesis, cell-to-cell contact, synaptogenesis, and migration. In addition, the dihydroceramide desaturase transcript (DEGS1), which produces ceramide, the precursor of sphingomyelin, a major component of the myelin sheath, is also inhibited. These findings suggest that demyelination could be an important pathophysiologic feature of mitochondrial mutations, especially because myelin is transported by the vesicular secretion pathway.

Overall, microarray data show that there is a significantly shared transcriptomal imprint of mitochondrial mutations, which implicates specific pathophysiological pathways. The data support the hypothesis that mitochondrial mutations may preferentially causes ER stress, which in turn triggers the UPR and decreases vesicular secretion, resulting in dysfunctions of secreted proteins involved in synaptogenesis, neural cell migration, and oligodendrocyte function. Interestingly, most of these transcriptional modifications are similar to the ones relative to brain aging, as

if mtDNA mutations play a primary causative event in the aging process.

Recently published microarray data from our laboratory, conducted on several cell types harboring mtDNA deletions, show a prominent inhibition of transcripts encoding ubiquitin-mediated proteasome activity. Thus, the downregulation of the ubiquitin–proteasomal system appears to be a unique consequence of mtDNA deletions. As already described, mtDNA deletions are the most common type of mitochondrial mutations known to accumulate exponentially during aging and to preferentially occur at higher levels within the CNS, and in the substantia nigra in particular. Thus, the genomic signal of UPS inhibition characteristic of brain aging might be secondary to an increase in mtDNA deletions, and this may even be more dramatic for neurodegenerative diseases with UPS defects, such as PD, which is characterized by specific death of dopaminergic neurons within the substantia nigra.

Conclusion

The field of aging research has been completely transformed in the past decade. Genomic profiles associated with brain aging became available due to the advent of microarray technology, and it is now widely accepted that the aging process is subject to regulation, like most biological processes. Mitochondrial dysfunctions were already known to play a relevant role because of the centrality of the organelle in apoptosis, Ca^{2+} regulation, and oxidative stress, and because mtDNA mutations increase with age. In addition, the nuclear transcriptional profiles consequent to mtDNA mutations are showing multiple similarities with the ones relative to brain aging. Similar trends in activation and inhibition of functional categories can be observed. Overall, it is becoming more evident that mutations in the mitochondrial genome in aged neurons play a causative role rather than merely being correlative, and the concepts connected to the mitochondrial theory of aging are thus undergoing change.

See also: Alzheimer's Disease: Neurodegeneration; Calcium and Signal Transduction; Gene Expression in Normal Aging Brain; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Glutamate and GABA Metabolism; Lipids and Membranes in Brain Aging; Microarray use for the Analysis of the CNS; Rodent Aging; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Plasticity: Neuronal Sprouting.

Further Reading

Beal MF (2005) Mitochondria take center stage in aging and neurodegeneration. *Annals of Neurology* 58: 495–505.

Blalock EM, Chen KC, Stromberg AJ, et al. (2005) Harnessing the power of gene microarrays for the study of brain aging and Alzheimer's disease: Statistical reliability and functional correlation. *Ageing Research Reviews* 4: 481–512.

Blalock EM, Geddes JW, Chen KC, et al. (2004) Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America* 101: 2173–2178.

Cortopassi G, Danielson S, Alemi M, et al. (2006) Mitochondrial disease activates transcripts of the unfolded protein response and cell cycle and inhibits vesicular secretion and oligodendrocyte-specific transcripts. *Mitochondrion* 6(4): 161–175.

Cortopassi GA, Shibata D, Soong NW, et al. (1992) A pattern of accumulation of a somatic deletion of mitochondrial DNA in aging human tissues. *Proceedings of the National Academy of Sciences of the United States of America* 89: 7370–7374.

Dickey CA, Loring JF, Montgomery J, et al. (2003) Selectively reduced expression of synaptic plasticity-related genes in amyloid precursor protein + presenilin-1 transgenic mice. *Journal of Neuroscience* 23: 5219–5226.

DiMauro S and Schon EA (2003) Mitochondrial respiratory-chain diseases. *New England Journal of Medicine* 348: 2656–2668.

Finkel T and Holbrook NJ (2000) Oxidants, oxidative stress and the biology of ageing. *Nature* 408: 239–247.

Lee CK, Weindruch R, and Prolla TA (2000) Gene-expression profile of the ageing brain in mice. *Nature Genetics* 25: 294–297.

Loeb LA, Wallace DC, and Martin GM (2005) The mitochondrial theory of aging and its relationship to reactive oxygen species damage and somatic mtDNA mutations. *Proceedings of the National Academy of Sciences of the United States of America* 102: 18769–18770.

Lu T, Pan Y, Kao SY, et al. (2004) Gene regulation and DNA damage in the ageing human brain. *Nature* 429: 883–891.

Mattson MP and Magnus T (2006) Ageing and neuronal vulnerability. *Nature Reviews Neuroscience* 7: 278–294.

Trifunovic A, Wredenberg A, Falkenberg M, et al. (2004) Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* 429: 417–423.

Relevant Websites

http://www.affymetrix.com – Affymetrix Corporation product/research Website.

http://www.alz.org – Alzheimer's Association (Chicago, IL, USA).

http://www.nia.nih.gov – National Institute on Aging (U.S. National Institutes of Health).

http://www.parkinson.org – National Parkinson Foundation, Inc. (Miami, FL, USA).

http://www.umdf.org – United Mitochondrial Disease Foundation (Pittsburgh, PA, USA).

Genomics of Brain Aging: Twin Studies

C A Reynolds, University of California at Riverside, Riverside, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

The Twin Method

The study of adult twin siblings offers a distinctive perspective on brain aging. Monozygotic (MZ; identical) twins share 100% of their genes in common, while dizygotic (DZ; fraternal) twins share 50% of their segregating genes in common. This fortunate experiment of nature affords multiple avenues to consider the etiology of brain aging. For example, comparing the relative similarity of MZ twins to DZ twins allows one to consider the extent of genetic and environmental influences on individual variation in cognitive performance or dementia status. If MZ twins are more similar than are DZ twins in recall memory performance, for example, then genetic influences are implicated. For this same trait, if DZ twins are as similar to each other as are MZ twins, then shared environmental effects, such as family environments and school environments, are implied. To the extent that MZ twins are dissimilar (e.g., a correlation less than 1.0), then nonshared environmental influences are implied.

Quantitative Traits

For quantitative traits, such as recall memory performance scores, covariances between twin pair scores are calculated for MZ pairs (C_{mz}) and for DZ pairs (C_{dz}). Next, a formal biometric model is fit to the covariances. The biometrical model quantifies understanding of all possible, though anonymous, genetic and environmental causes and shows how these factors contribute to sibling similarity. For example, the biometrical model could state that individual differences in recall memory performance (V_{mem}) are a function of genetic factors that act additively (V_a), plus rearing environment factors (V_r), plus unique person-specific (nonshared) experiences (V_e):

$$V_{mem} = V_a + V_r + V_e \quad [1]$$

Next, the model specifies that two related individuals have similar recall memory scores due to the amount of genes shared, and the environment shared. For MZ twins reared together, their recall memory scores are similar, or co-vary, because the twins are genetically identical and they share a rearing environment:

$$C_{mz} = V_a + V_r \quad [2]$$

For DZ twins reared together, their recall memory scores are similar because the twins share half of their segregating genes in common and share a rearing environment:

$$C_{dz} = 1/2V_a + V_r \quad [3]$$

Heritability is often reported as a summary statistic and is a numerical ratio of genetic factors (V_a) over the total amount of individual differences in recall memory (V_{mem}). Thus, heritability (h) indicates the proportion of individual differences in recall memory scores that is due to genetic differences:

$$h^2 = V_a/V_{mem} \quad [4]$$

Environmentality estimates are computed in similar fashion (e.g., shared environmentality is the ratio of V_r/V_{mem}). In longitudinal studies, there is increasing emphasis on unscaled genetic and environmental variance (i.e., V_a, V_r, V_e). For example, one can better judge whether both genetic and environmental influences are increasing or decreasing with age, a process that otherwise would be masked with ratio-based heritability and environmentality statistics. This knowledge is important in theorizing about potential gene action, specific environmental factors to consider, or longevity-related mechanisms whereby individuals with particular genotypes may die at earlier (or later) ages.

Disease Traits

Disease traits are typically measured in terms of absence or presence (i.e., dichotomously), or otherwise as ordinal categories. Concordance rates are commonly reported for dichotomous traits and are compared between MZ and DZ twin pairs. Probandwise concordance is the simple proportion of co-twins of affected proband twins that exhibit the disease trait (e.g., Alzheimer's disease (AD)). In much the same way as for quantitative measures of pair similarity, the extent of differences between MZ and DZ concordance rates gives an indication of genetic influence. Furthermore, the extent to which the MZ concordance rate is less than 100% suggests the degree to which nonshared environmental influences, including random factors, are important.

Other measures of pair similarity may be reported for dichotomous or ordinal categorical traits, including tetrachoric and polychoric correlations, respectively. Herein, the assumptions are that there is an underlying continuum of liability or risk of the disease and that the observed proportion of those with the disease constitutes a threshold point on the normal curve where the disorder is manifested.

The interpretation of the correlations computed for twin pairs is that they reflect the magnitude of pair similarity for the underlying disease liability continuum. Prevalence rates increase concordance rates even when the underlying pair similarity for the disease liability is identical. Thus, tetrachoric correlations are preferred over concordance rates when the prevalence rates of the disease differ between groups included in the analysis, as can be the case for males versus females (e.g., AD). The biometrical model described in eqns [1]–[3] is applied to the tetrachoric or polychoric correlations in similar fashion.

Assumptions of the Twin Method

Analytical requirements include the assumption of equal environments for twins-reared-together, as well as no assortative mating. The equal environments assumption (EEA) states that the shared environment (e.g., family environment, school environment) contributes equally to a trait in both MZ and DZ pairs. The vast majority of analyses which have tested for EEA violation across multiple psychological and psychiatric traits have indicated little support for EEA violation, thus providing sufficient confidence in the findings from studies of twins-reared-together, though it must be noted that some controversy remains over the adequacy of tests for violations of the EEA.

A second assumption is the absence of assortative mating – that is, spouses or mates do not select one another on the basis of the trait in question. There is ample evidence that spouses are moderately similar in terms of cognitive abilities, extant from the time of marriage. For aging-related cognitive traits (i.e., cognitive aging, AD), there are few data reported on spouse similarity. For now, it appears that the similarity between spouses for dementia is no greater than the similarity between any two people of the same age and sex.

Co-Twin Control

In studies of discordant twins, the co-twin control method compares affected twins to their unaffected partners on a host of environmental factors. This design is advantageous to comparable case-control designs, as matching the discordant twins to their co-twin accounts for any confounding genetic of familial factors that might bear upon the disease trait. Indeed, discordant MZ twin pairs who share 100% of their genes in common are most valuable for full control of genetic and familial factors.

Candidate Genes

Twin siblings who differ in terms of their recall memory scores, for example, and who differ in genotype at a particular gene locus (e.g., *APOE* ε3/ε4 vs. *APOE* ε3/ε3), give potential information about the association of the candidate gene with recall memory performance. True gene effects contribute to within-twin-pair differences, in the case of DZ twins, as well as to between-twin-pair differences for both MZ and DZ twins. Thus, formal tests of association in twins compare whether, according to genotype, within-pair differences and between-pair differences in trait scores are of the same magnitude. Twin designs have a distinct advantage over standard case-control designs, in which genetic association may be spurious if cases and controls originate from separate populations with differing gene frequencies (i.e., population stratification).

Cognitive Aging and Dementia via Twin Models

Twin Studies

Several research programs include adult twins as a focus in studies of cognitive abilities, most situated in the US (e.g., MTSADA, Minnesota Twin Study of Adult Development and Aging; NHLBI, National Heart, Lung, and Blood Institute Twin Study; New York Twin Study) or Scandinavia (LSADT, Longitudinal Study of Aging Danish Twins; Norwegian Twin Register; OCTO-Twin, Origins of Variance in the Oldest Old; SATSA, Swedish Adoption/Twin Study of Aging; the Gender Study of Opposite-Sex Twins), though others are located in Germany (German Observational Study of Adult Twins), Italy (Italian Twin Registry), Japan (Osaka/Kinki University Twin Study; OUATR, Osaka University Aged Twin Registry), and The Netherlands (Netherlands Twin Register). Among these research programs, the longitudinal studies, LSADT, NHLBI, OCTO-Twin, and SATSA have provided much of our knowledge about the etiology of cognitive decline. Summaries of findings from these studies are presented in the following sections.

Twin studies of cognitive impairment and dementia are fewer in number than are those addressing cognitive aging, and include two large-scale programs in the USA (NAS-NRC Registry, National Academy of Sciences – National Research Council Registry of World War II veteran twins; NHLBI), and several in Scandinavia (LSADT; Norwegian Twin Register; HARMONY, Study of Dementia in Swedish Twins).

Contributions of Genes and Environments

General cognitive abilities Cross-sectional findings of adult twins indicate a strong heritability for general cognitive ability, or *g*, with a median value of 0.69 (see Figure 1). Therefore, 69% of the individual

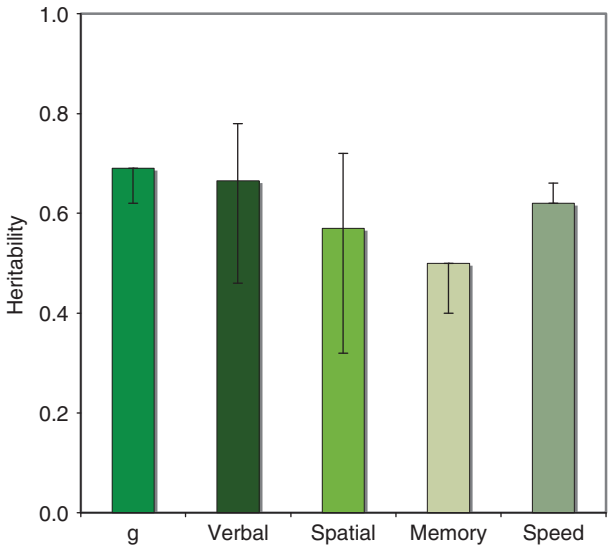

Figure 1 Heritability estimates of cognitive abilities; *g* represents general cognitive ability. Note: Tops of solid colored bars indicate median heritability estimates from all adult twins studies that included older adult twins. Lower ends of 'error bars' indicate median old-old estimates and upper ends indicate young-old median.

differences in general cognitive ability results from genetic differences. Furthermore, the described studies find no evidence for shared environmental factors (e.g., rearing environment) affecting general cognitive abilities in adults, thus the remaining proportion of individual difference, 0.31, or 31%, is due to person-specific or nonshared environmental influences. Studies that report separate estimates for young-old adults versus old-old adults consistently suggest a waning of heritable influences for old-old adults, and concomitantly an increase in nonshared environmentality influences. The median heritability estimate for young-old twins is 0.69 versus 0.62 in the old-old (see error bars in Figure 1). Swedish twin studies, in particular SATSA and OCTO-Twin, suggest even more marked difference in heritability: as high as 0.80 in young-old twins and 0.62 in old-old twins. Thus, cross-sectional findings imply waning genetic influences in old-old age.

Specific cognitive abilities Specific ability domains, such as verbal, spatial, memory, and perceptual speed abilities, show varying patterns of the extent of genetic influences. The median heritability ranges from 0.50 for memory to 0.66 for verbal abilities (see Figure 1). Virtually all cross-sectional studies suggest that the remaining variation is largely due to nonshared environmental influences, with few hints of shared environmental effects on select cognitive measures within a domain. Furthermore, no convincing evidence for sex differences in heritability or for sex-limited genetic or environmental factors has been presented to date.

The median estimate of 0.66 for verbal ability across all studies is flanked by a median heritability of 0.78 in the young-old and a heritability of 0.46 in the old-old, again suggesting a similar pattern of waning genetic influences in old-old age. The most disparate pattern between young-old and old-old heritability estimates is observed for spatial ability (median of 0.72 in young-old and 0.32 in old-old), and the least is observed for perceptual speed (median of 0.66 in young-old and 0.62 in old-old). The estimates for the memory domain indicate a median heritability of 0.50 across all studies, and a median heritability of 0.40 in old-old versus 0.50 in young-old adult twins.

The memory domain is made up of multiple distinctive processes, including working memory, semantic memory, episodic memory, and long-term memory. Twin studies suggest higher heritability estimates for working memory measures than for episodic memory measures. Furthermore, the SATSA study indicates a lower heritability for working memory measures in young-old versus old-old (0.36 vs. 0.63) but a higher heritability for episodic memory measures in young-old versus old-old (0.50 vs. 0.33), as has been noted for the other cognitive domains.

Across domain, the differences in heritability and environmentality estimates among young-old and old-old suggest that new environmental factors may emerge in old-old adults, or perhaps that cohort differences are important. Longitudinal studies of twins best address these possible explanations and most directly assess the heritability of cognitive decline.

Cognitive change Longitudinal studies of twins have added to our understanding of the etiology of cognitive abilities and particularly of cognitive decline with age. Figure 2(a) illustrates the nature of the data analyzed in longitudinal studies of twins. Twins have their own individual cognitive test score trajectory that can be aligned to age of measurement. Furthermore, the trajectories within and between pairs are examined to establish the extent of environmental and genetic influences. The illustrative trajectories in Figure 2(a) suggest that cognitive performance (e.g., episodic memory ability) declines with age for all four twins. Furthermore, the extent of decline differs across individuals and within pairs, particularly for the DZ twin pair versus the MZ pair. Analyses of longitudinal twin data consist of fitting characteristic curves (e.g., linear, quadratic) to individual cognitive data and comparing twin similarity for curve parameters. By and large, the curves of MZ twin pairs are more similar than are those of DZ twin pairs, but

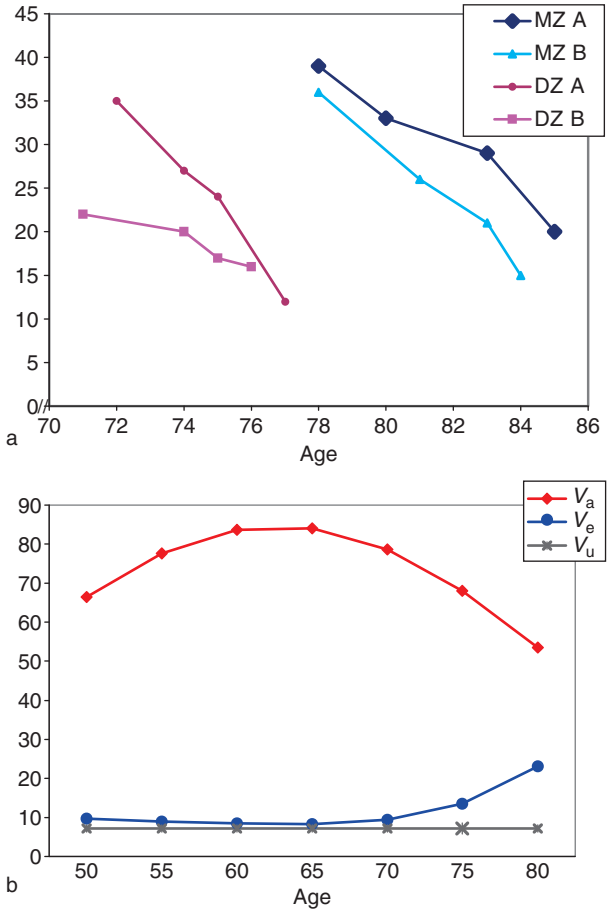

Figure 2 Application of longitudinal twin models to longitudinal cognitive data: (a) illustrative data for monozygotic (MZ) and dizygotic (DZ) pairs of twins, ages 70–86 years; (b) actual results from the Swedish Adoption/Twin Study of Aging for general cognitive ability, *g*, for twins aged 50–80 years. V_a represents additive genetic variance contributing to systematic longitudinal growth; V_e represents person-specific (nonshared) environmental variance contributing to systematic longitudinal growth; V_u represents occasion-specific variance (nongrowth-related variance). (b) Data from Reynolds CA, Finkel D, McArdle JJ, Gatz M, Berg S, and Pedersen NL (2005) Quantitative genetic analysis of latent growth curve models of cognitive abilities in adulthood. *Developmental Psychology* 41: 3–16; and adapted from Reynolds CA (in press) Genetic and environmental influences on cognitive change. In: Hofer SM and Alwin D (eds.) *The Handbook of Coginitive Aging. Interdisciplinary Perspective*. London: SAGE, with permission.

pairs become less similar to one another with age, particularly after age 65 years. Genetic factors influence average cognitive performance levels, as well as the shape of curve trajectories, though to a lesser degree. Furthermore, the course of genetic and environmental influences with age is consistent with cross-sectional studies, with the exception of memory; that is, there are waning genetic influences after age 65 years and increasing nonshared environmental influences. Figure 2(b) shows findings from the SATSA study for general cognitive ability, *g*, where only additive genetic factors (V_a) and nonshared environmental factors (V_e) contribute to individual differences in average performance and change. Genetic factors contribute the most to individual differences in *g* with age, but a decreasing trend is evident after age 65 years, when nonshared environmental factors assume greater importance. Thus, unique person-specific environmental factors may become amplified in late adulthood, or, alternatively, new environmental factors emerge. Likewise, the increase in nonshared environmental effects may suggest the possibility of gene–environment (G × E) interactions, as G × E effects become part of the nonshared environment estimate if present but not modeled directly.

The pattern of increasing unique person-specific environmental factors and decreasing role for genetic factors has been observed nearly uniformly across cognitive domains, with the exception of cognitive measures that tap memory. For example, for both working memory and episodic memory tasks, both genetic and unique person-specific environmental factors increase in importance with age, suggesting the amplification or emergence of new genetic and environmental factors in late adulthood. Thus, the search for measured genes and measured environments for cognitive change is of interest.

Endophenotypes of cognitive aging Locating relevant and heritable intermediary neural traits (i.e., endophenotypes) which relate to cognitive aging is important to understanding mechanisms underlying cognitive change. Studies of adult twins suggest that processing speed, brain volume, and electroencephalogram (EEG) measures may be reasonable endophenotypes in the search for measured genes, given their significant though typically modest to moderate relationships with measures of intellectual ability. Recently, regional brain volume metrics have been associated with cognitive performance in Dutch twins, as well as in aging male twins from the NHLBI study, in which a significant percentage of the relationship between lateral ventricular volumes and executive functioning performance was due to common genetic influences; other lobar brain volume–cognition relationships were associated through shared and nonshared environmental pathways. Furthermore, higher white-matter hyperintensities (WMHs) predict greater rates of decline in perceptual speed performance, and this association may be due largely to common genetic effects influencing both cognitive function and WMH.

Cognitive functioning The heritability for mental status is 0.49 across studies. However, mental status performance as assessed with telephone screening

methods tends to yield lower heritability estimates than when assessed with in-person instruments (e.g., Mini-Mental Status Exam), which may in part reflect selection effects on health for those well enough to participate in person or by telephone. Furthermore, compared to cognitive performance measures or risk of AD, heritability of cognitive dysfunction is typically lower and may be the result of the diverse range of factors that lead to cognitive impairment, including health conditions and poor lifestyle habits, some of which show little evidence of genetic influences, such as cardiovascular disease.

Dementia Late-onset AD is due to multiple factors and is highly heritable, with MZ twins showing consistently higher concordance (and tetrachoric correlations) compared to DZ twins. Figure 3 shows the median heritability for AD of 0.52 across all studies to date, with minimum and maximum values indicated by the error bars. The HARMONY study, at present the biggest population-based twin study of AD, estimates heritability to be as large as 79% for prevalent AD, while nonshared environmental effects account for the remaining variation in risk. Heritability of AD appears to be of similar magnitude for males and females. Vascular dementia, on the other hand, appears to be influenced completely by shared and nonshared environmental factors, given essentially equal concordance rates for MZ and DZ twins. Late-onset Parkinson's disease (PD), of which dementia sometimes features, appears to be mediated primarily by environmental factors, with nearly identical but low concordance rates in MZ and DZ twin pairs.

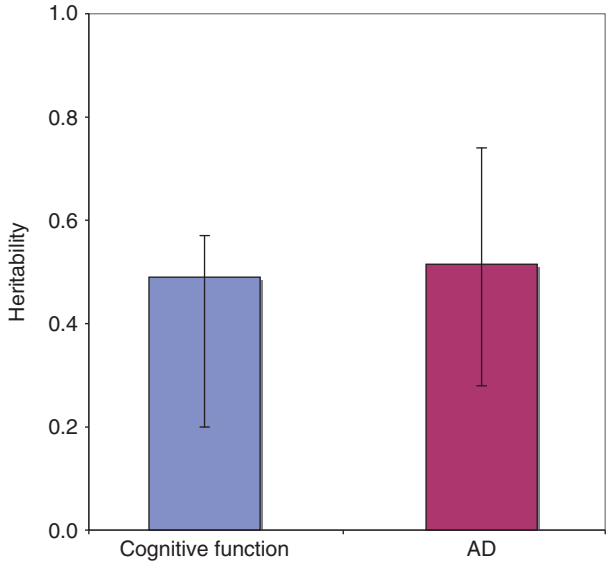

Figure 3 Heritability estimates of cognitive functioning and Alzheimer's disease (AD). Note: Tops of solid colored bars indicate median heritability estimates from all adult twins studies. Lower ends of 'error bars' indicate lowest estimate and upper ends indicate highest estimate.

Measured Genes

Twin sibling designs are valuable to address the presence of true gene-trait association. However, such studies have not been used to examine candidate association with cognitive aging traits with much frequency, or the pair aspect of the twins has not been used to full advantage. However, reports from twin studies are beginning to appear.

Cognitive aging and decline The *APOE* gene coding for the cholesterol transporter apolipoprotein E is considered a major genetic risk factor for late-onset AD. The weight of evidence for *APOE* and cognitive change in nondemented adults has been less consistently observed, in comparison. Twin association studies suggest an effect on memory-related tasks: the *APOE* $\varepsilon 4$ polymorphism has been associated with poorer recall memory, poorer working memory performance, and a steeper rate of change in working memory over age.

Apart from *APOE*, additional reports using twin models of association with cognitive aging or related traits are beginning to appear in the literature. For example, a recent SATSA study of twins suggests an association of the −1438 G/A [dbSNP: rs6311] variant of the *HTR2A* gene that codes for 5-hydroxytryptamine 2A (5-HT_{2A}; serotonin) receptors with differences in memory decline over time in late adulthood. Those with the common allele (G) exhibited higher performance on an episodic figure recognition task over time than did those with the rarer allele (A). These results extend a prior case-control study and are consistent with molecular and imaging studies that indicate decreases in serotonin 2A receptors with age in hippocampus and prefrontal cortex regions. Additional promising gene candidates have been reported in association studies for general intelligence in child and adult twins, including cholinergic muscarinic receptor 2 (*CHRM2*) gene polymorphisms. Finally, there are candidates for cognitive aging and decline that ought to be considered from a twin perspective, including the catechol *O*-methyltransferase (*COMT*) gene, which has been associated with performance and/or change in executive functioning and spatial performance in older adults.

Cognitive functioning *APOE* $\varepsilon 4$ status is related to an increased risk of mild cognitive impairment (MCI) according to analyses of twins as individuals from the

NHLBI study. Aging Danish DZ twins from the LSADT study were examined for four polymorphisms in the Werner syndrome *WRN* gene. In an analysis of all twins, without consideration of pair status, a positive association was found between the *Wrn-i1* gene variant located in a noncoding region of the *WRN* gene and cognitive functioning, a composite measure consisting of word fluency, word learning, and span memory performance, as well as self-rated health. Among DZ pairs who differed in genotype for the *WRN* gene variant, findings for the cognitive composite were consistent with the overall findings, but too few pairs led to lack of significance. Overall, the findings implicate *WRN* in cognitive functioning and health, and a further search of variants in regulatory regions is warranted.

Dementia *APOE* ε4 status predicts concordance for AD among twin pairs, as well as increased risk of AD among first-degree relatives. Several additional gene association studies have included twins in analyses of AD risk but have not considered AD risk association from a twin perspective *per se* to confirm associations. Beyond *APOE*, the most well-known and replicated genetic risk factors for late-onset AD, genes for insulin-degrading enzyme *(IDE)* and ATP-binding cassette A1 transporter (*ABCA1*), have been associated with increased AD risk or with earlier age of onset. Thus, a broader role of cholesterol metabolism on AD risk may be implied (e.g., *APOE*, *ABCA1*). Replication via within- and between-twin pair comparison is needed to confirm these and other associations suggested in the epidemiological literature.

Measured Environmental Factors

Both positive and negative aspects of the environment appear to play an increasingly vital part in cognitive aging and risk of dementia. Epidemiological research points to low education, poor early-life socioeconomic factors, and head injury as risk factors, while protective factors may include physical and cognitive activity, antioxidant vitamin supplements, and nonsteroidal anti-inflammatory drugs (NSAIDs). Twin studies of measured environmental influences on cognitive aging or dementia have been conducted generally using two methods: by addressing the etiology of associations between complex traits to determine if their common variance is due to genetic or environmental pathways (e.g., education and cognitive functioning), or by comparing MZ twins discordant for AD or cognitive decline, for example, and examining differences in potential risk or protective variables (e.g., NSAID use).

Cognitive aging and decline Low education is a risk for compromised mental functioning. Theories have weighed in with biological (e.g., genetic factors influence both educational attainment and intelligence) or environmental explanations (e.g., cognitive reserve, or 'use-it-or-lose-it'). Twin studies suggest that education, cognitive ability level, and mental status performance are primarily associated due to genetic influences in common to all three, supportive of biological theories. However, as is discussed later, some weight should be accorded to environmental explanations, given the predictive value of pair differences in education and dementia risk in discordant twins.

Lobar brain-volume relationships with cognitive performance appear to be associated through environmental pathways in an analysis conducted on data from elderly male twins participating in the NHLBI twin study. Additional analyses further suggest that lower total brain volume is associated with a greater rate of decline in memory performance, perhaps influenced by common environmental factors.

Dementia Recent co-twin control studies of dementia-discordant twin pairs in the HARMONY study suggest that environmental factors may mediate dementia or AD risk, including occupational complexity, physical exercise, participation in leisure activities decades earlier, and tooth loss prior to age 35. Unaffected twins had a greater lifetime occupational complexity score, greater involvement in leisure activities 20 years prior to assessment, and were less likely to report tooth loss prior to age 35 years than was their AD affected twin partner. The reported tooth loss in earlier adulthood may reflect inflammation processes that result from poor health or nutrition factors earlier in life. Additional reports from the NAS–NRC Twin Registry and OCTO-Twin studies suggest that use of anti-inflammatory medications (NSAIDs, aspirin) may reduce risk of AD, replicating prior epidemiological studies.

In addition, co-twin-control reports from HARMONY indicate significant environmental factors important to diseases other than AD, or that are nonspecific to dementia type. Deficient physical exercise was more often observed for dementia-affected twins than for their co-twins, exclusive of AD. Risk of dementia of any type was increased in the case of lower educational attainment, and was evident to greater degree in dementia-affected twins than in their co-twins. The findings for education are supportive of environmental explanations, such as the cognitive reserve hypothesis, which posits that higher educational attainment may afford protection via greater synaptic density or cognitive compensation, despite neuropathologies, and thereby lower risk of, or delay clinically, meaningful cognitive impairments. Indeed, a host of potential neuroprotective or compensatory

factors may coalesce in twins with higher education. In education-discordant twins, the twin with higher education reported greater participation in leisure activities and higher occupational status, for example.

The evidence from co-twin-control studies indicates intriguing possibilities that particular environmental factors may modify dementia or AD risk, though some factors appear to confer advantages or disadvantages in early or mid-adulthood, a time when individuals may not reflect on their future health.

Gene–Environment Interplay

Environmentally mediated risk factors, as described earlier, may influence cognitive decline or risk of AD, and recent epidemiological literature suggests often this is dependent on genotype. For example, mid-adulthood hypertension or the occurrence of stroke or head injury, when coupled with *APOE* ε4 carrier status, may lead to poorer cognitive performance or poorer cognitive recuperation, respectively. Viewing gene–environment interplay from a twin perspective could replicate and build upon existing epidemiologic research, but this is not often done, or the twin aspect is ignored. One twin approach to establish $G \times E$, while controlling for genetic similarity, is to compare MZ twins (e.g., on measures of cognitive decline), stratified by genotype, to examine whether pair differences are smaller or greater for pairs with particular genotypes. This approach would indicate whether individuals with certain genotypes are more or less 'sensitive' or responsive to environmental factors. The advantage of the twin $G \times E$ model is that background genetic and familial factors would be controlled.

Conclusion

Twin studies provide a unique window on the etiology of brain aging. Identical twins are more similar or concordant than are fraternal twins for cognitive aging, cognitive dysfunction, and certain forms of dementia (i.e., AD). Other dementia- and aging-related traits show little evidence of genetic influences, including vascular dementia and PD, in which identical and fraternal twins are equally similar. Longitudinal twins studies of cognitive aging indicate that genetic influences decrease with age while person-specific environmental influences increase in importance. Co-twin-control studies suggest that individuals may be able to modify their risk of AD or dementia through environmental means, but that this may require effort even before midlife, before individuals may be cognizant of dementia risk. Epidemiological studies hint that the degree to which risk may be modifiable could be dependent on genotype. Twin approaches to gene–environment interplay should confirm and build upon epidemiological findings, given the strength of the twin design.

See also: Aging of the Brain and Alzheimer's Disease; Aging: Extracellular Space; Aging: Brain Potential Measures and Reaction Time Studies; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: MRI Studies; Amyloid: Vascular and Parenchymal; Animal Models of Alzheimer's Disease; Brain Volume: Age-Related Changes; Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Memory: Genetic Approaches; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Bendixen MH, Nexo BA, Bohr VA, et al. (2004) A polymorphic marker in the first intron of the Werner gene associates with cognitive function in aged Danish twins. *Experimental Gerontology* 39: 1101–1107.

Boomsma D, Busjahn A, and Peltonen L (2002) Classical twin studies and beyond. *Nature Reviews Genetics* 3: 872–882.

Carmelli D, Reed T, and DeCarli C (2002) A bivariate genetic analysis of cerebral white matter hyperintensities and cognitive performance in elderly male twins. *Neurobiology of Aging* 23: 413–420.

Finkel D and Reynolds CA (in press) Behavioral genetic investigations of cognitive aging. In: Kim Y-K (ed.) *Handbook of Behavior Genetics*. New York: Springer.

Gatz M (2007) Genetics, dementia, and the elderly. *Current Directions in Psychological Science* 16(3): 123–127.

Gatz M, Prescott CA, and Pedersen NL (2006) Lifestyle risk and delaying factors. *Alzheimer Disease and Associated Disorders* 20: S84–S88.

Gatz M, Mortimer J, Fratiglioni L, et al. (2006) Potentially modifiable risk factors for dementia in identical twins. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 2: 110–117.

McClearn GE, Johansson B, Berg S, et al. (1997) Substantial genetic influence on cognitive abilities in twins 80 or more years old. *Science* 276: 1560–1563.

Posthuma D and de Geus EJC (2006) Progress in the molecular-genetic study of intelligence. *Current Directions in Psychological Science* 15: 151–155.

Reynolds CA, Finkel D, McArdle JJ, et al. (2005) Quantitative genetic analysis of latent growth curve models of cognitive abilities in adulthood. *Developmental Psychology* 41: 3–16.

Special Issue on Aging. Behavior Genetics 33(2): 79–203.

Special Issue on TwinRegisters, Across the Globe. Twin Research 5 (5): v505.

Vogler GP (2006) Behavior genetics and aging. In: Birren JE and Schaire KW (eds.) *Handbook of the Psychology of Aging,* 6th edn., pp. 41–55. Amsterdam: Elsevier.

Relevant Websites

http://www.sri.com – Genetic Epidemiology and Aging Research (NHLBI Twin Study), Stanford Research Institute.

http://www.meb.ki.se – Karolinska Institutet, Department of Medical Epidemiology and Biostatistics (Swedish Twin Registry, Including the HARMONY and SATSA Studies).

http://www.psych.umn.edu – Minnesota Twin Family Study (Study of Adult Development), University of Minnesota.

http://www.fds.duke.edu – Program on Population, Policy, and Aging, Duke University (Longitudinal Study of Aging Danish Twins).

http://www.iom.edu – Study of U.S. Veteran Twins, Institute of Medicine of the National Academies (NAS–NRC Twin Registry).

GFL Neurotrophic Factors: Physiology and Pharmacology

M Saarma, University of Helsinki, Helsinki, Finland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Growth factors are secretory proteins that bind to their cognate receptors at the plasma membrane of the cells and activate receptors that in turn trigger the activation of cellular biochemical signaling processes and lead to the cell proliferation, differentiation, migration, morphological changes, or induce the death of cells. Currently, about 100 growth factors that affect neurons have been described, but only those that mainly regulate the development and physiology of neurons are called neurotrophic factors. Neurotrophic factors are growth factors that regulate the number of neurons in a given population, neurite branching and synaptogenesis, adult synaptic plasticity, and maturation of neuronal phenotype. Neurotrophic factors include, for example, neurotrophins (NGF, BDNF, NT-3, NT-4), neurokines (CNTF, LIF, IL-6, CT-1, etc.), and glial cell line-derived neurotrophic factor (GDNF) family ligands (GFLs).

GFLs: Structure, Secretion, and Processing

GDNF and the related factors artemin (ARTN), neurturin (NRTN), and persephin (PSPN) form the GDNF family of neurotrophic factors (Figure 1(a)). GFLs are synthesized by the cells as preproGFL proteins that are in the endoplasmic reticulum (ER) processed to proGFLs and secreted into the extracelluar space as mature GFLs or as proGFLs. The 211 amino-acid-long preproGDNF protein is finally processed into 134 amino acid mature GDNF monomer (Figure 1(b)) that is inactive and must homodimerize to generate a biologically active factor. All GFLs have seven conserved cysteines with similar relative spacing. Six of these cysteines are involved in the intramolecular S–S bonding whereas the seventh cysteine forms an intermolecular S–S bond stabilizing the homodimers of GFLs. Based on the conserved location of the seven cysteines, GFLs form a distant subfamily in the large transforming growth factor-beta (TGF-β) superfamily of growth factors. Orthologs of all four GFLs are found in all mammals, and also in teleost fishes. Interestingly, chicken genome lacks PSPN gene, but has a functional PSPN coreceptor GDNF family receptor α4 (GFRα4). Clawed frog *Xenopus tropicalis* genome surprisingly lacks the ortholog of NRTN. Even more surprisingly insects have a receptor tyrosine kinase (RTK) rearranged during transfection (RET), a putative GFL coreceptor homolog but no obvious GFL homologs. The atomic structure of the major part of GDNF has been determined and GDNF together with all other TGF-β superfamily growth factors, including all GFLs, belong to the cystine-knot family of proteins. The main structural elements of the GDNF molecule are helical regions and two fingers. Fingers 1 and 2 are involved in the binding of GDNF to its receptor GFRα1 (Figure 1(c)). In the GDNF structure the first 38 residues are disordered. Sequence analysis and experimental data indicate that this region binds heparin stabilizing the structure of GDNF and mediating GFL interactions with extracellular matrix (ECM). Recently, two laboratories have solved the crystal structure of ARTN. Although the overall fold of the ARTN covalent homodimer is similar to GDNF, the shape and the flexibility of elongated homodimer differs. ARTN also differs from GDNF in its overall charge and electrostatic distribution. Arginines in ARTN at positions 48, 49, and 51, as well as its amino terminus play a role in heparin binding. Although the three-dimensional (3-D) structures of NRTN and PSPN are not known, it is likely that their general spatial structure is very similar to GDNF and ARTN (Figure 1(c)). There are also heparin-binding regions in NRTN, but not in PSPN.

The detailed mechanism of the secretion, processing, and activation of GFLs is poorly studied. Several cells and tissues can also secrete proGDNF in addition to mature GDNF (Figure 1(b)), suggesting that processing of proGDNF (and possibly all GFLs) can occur also in the ECM. Whether proGFLs can bind GFL receptors and trigger cellular signaling is currently not known.

GFL Receptors

GFL receptor system is unique in several ways. First, GFLs can signal through three different receptor systems. Second, all three signaling receptors, RET, neural cell adhesion molecule (NCAM), and syndecans, are shared by four ligands, with the exception of PSPN that is not binding to syndecans. Third, the main GFL signaling receptor is the RTK RET, which is the only known RTK that does not bind the ligands directly but via coreceptors of GFRα family. Finally, RET activity is strictly regulated by Ca^{2+} ions, and in the intracellular part of RET in addition to tyrosine residues serines are also phosphorylated.

The RET gene was identified in 1985 as a novel oncogene. Normal RET is the transmembrane RTK.

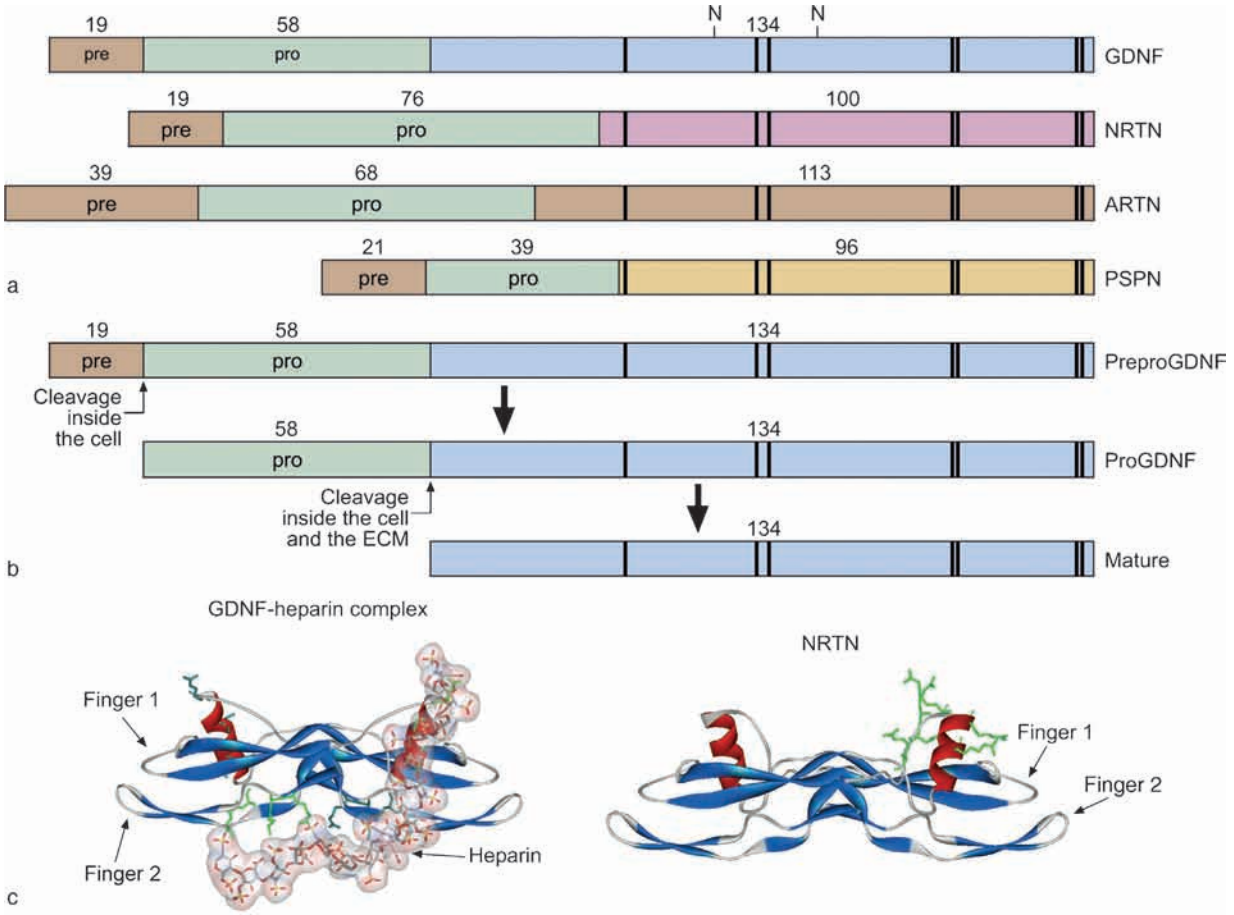

Figure 1 (a) Schematic structure of GFL showing the relative lengths (and number of amino acids) of the mature GDNF (blue), NRTN (pink), ARTN (brown), and PSPN (yellow) molecules and their pre-(brown) and pro-(light green) domains, as well as relative positions of the seven conserved cysteine residues (black vertical lines). GDNF has two putative N-glycosylation sites (N). (b) The preregion of GFLs is cleaved off inside the cells during secretion. Cells can secrete mature GDNF and proGDNF, indicating that the processing of proGDNF can also occur in the ECM. (c) The 3-D structure of dimeric GDNF shows that it belongs to the cystine-knot family of proteins. The main structural elements of the GDNF molecule are helical regions and fingers 1 and 2 that are involved in its binding to GFRα1 coreceptor. The 3-D structure of ARTN is similar to that of GDNF. The 3-D structures of NRTN and PSPN are not known, but the modeling of NRTN shows that its spatial structure is very similar to GDNF. The binding of heparan sulfate to GDNF is shown (left). NRTN can also bind heparan sulfate. Putative binding site is shown (green).

In human cancer, especially in papillary thyroid carcinoma, the RET tyrosine kinase domain is fused with N-terminal regions of the variety of unrelated genes. In addition to these somatic rearrangements germ line mutations in RET are responsible for two inherited disorders: activating mutations cause multiple endocrine neoplasia 2 (MEN2) and inactivating mutations Hirschsprung's disease (HSCR) or congenital central hypoventilation syndrome. The alternative splicing of the human RET gene gives rise to at least three splice isoforms. The larger of these encodes a protein of 1114 amino acids that contains 51 amino acids at the C-terminus (RET51) that are replaced by nine amino acids in the RET9 isoforms. RET51 has two additional tyrosines: Tyr1090 and 1096. Orthologs of RET have been identified in higher and lower vertebrates, as well as in *Drosophila* (D-RET). The tyrosine kinase domain of D-RET is conserved as it is functional *in vitro* and activates similar intracellular signaling pathways as mammalian RET. The extracellular domain of D-RET is less conserved and fails to interact with mammalian GDNF/GFRα1 complex. Interestingly, the *Drosophila* genome does not encode putative GFLs, but has a GFRα-like gene, called munin.

RET is the only RTK among the receptors of TGF-β superfamily factors, all others being the receptor serine–threonine kinases. Although RET is similar to other RTKs, it also has several special features. First, the extracellular domain of RET consists of four cadherin-like domains (CLD1–4) and a single cysteine-rich domain (CRD). RET binds one molecule of Ca^{2+} that is required for its correct folding and function as the GFL receptor. Second, RET is the only

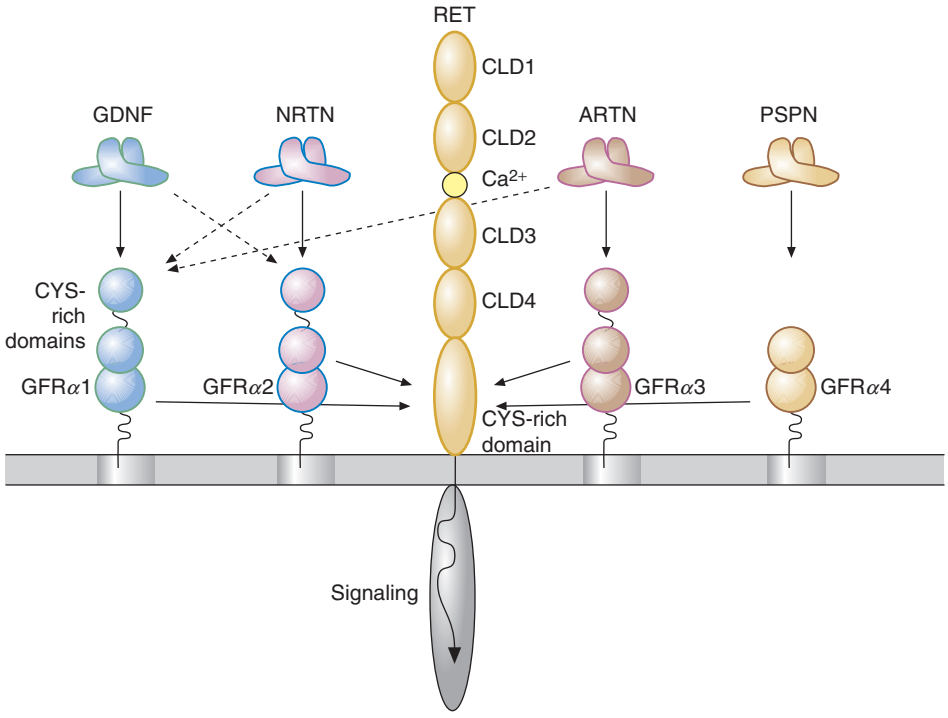

Figure 2 GDNF-family ligands and receptor interactions. Homodimeric GFLs activate RET RTK by first binding to their cognate GFRα coreceptors. GDNF binds preferentially to GFRα1, NRTN to GFRα2, ARTN to GFRα3, and PSPN to GFRα4. Solid arrows indicate the preferred functional ligand–coreceptor interactions, whereas dotted arrows indicate putative cross-talk. GFRα proteins are attached to the plasma membrane through a GPI-anchor and consist of three (GFRα4 has only two) globular cysteine-rich domains joined together by adapter sequences. The extracellular domain of RET, that consists of CLD1–4 and one CRD interacts with all four GFL–GFRα complexes. Binding of Ca^{2+} ions to one CLD of RET is required for its activation by GFLs. RET intracellular part contains tyrosine kinase domain and tyrosine residues that become phosphorylated upon RET activation.

RTK that does not directly bind the ligands, that is, GFLs (Figure 2). Instead, RET is activated by binding to a complex formed by a GFL bound to its cognate glycosylphosphatidylinositol (GPI)-anchored coreceptor GFRα. Thus, GFLs interact with RET, but only in the presence of specific GFRα coreceptors. Four different GFL-GFRα pairs exist in mammals: GDNF-GFRα1, NRTN-GFRα2, ARTN-GFRα3, and PSPN-GFRα4 (Figure 2). With significantly lower-affinity GDNF can interact with GFRα2 and GFRα3, and NRTN and ARTN with GFRα1, but the physiological significance of these interactions is unclear. The GFL and GFRα homodimeric complex brings two RET molecules together, triggering transphosphorylation of specific tyrosine residues in their tyrosine kinase domains and activating intracellular signaling (Figure 3).

GFRα Coreceptors

The GFL binding specificity to RET is determined by GFRα proteins that have unique binding affinities for each GFL. GFRαs are cysteine-rich GPI-anchored proteins. Mammalian GFRα1–3 and most of the vertebrate GFRα consist of three CRDs, whereas mammalian PSPN receptor GFRα4 lacks CRD1 and thus has only two CRDs (Figure 2). The structure of GFRα1 domain 3 revealed a novel protein fold: an all-α five-helix bundle stabilized by five disulfide bridges. The model for domain 2 was constructed, using its homology to domain 3. Recently, the crystal structure of ARTN in the complex with domains 2 and 3 of the GFRα3 was reported. These data reveal crucial amino acids in domain 2 of the GFRα3 receptor involved in ARTN binding and indicate that domains 2 and 3 define the potential RET binding site. Structural data on the ARTN-GFRα3 complex, on GFRα1 domain 3, and the conserved location of cysteine residues in GFRα1–4 CRDs predict a very similar general structural fold for all GFRαs. Phylogenetic analysis indicates that all vertebrate classes from teleost fishes to mammals have orthologs for all four GFRα receptors.

For several neuronal populations the full biological activity of GDNF is exerted only in the presence of TGF-β. It appears that TGF-β is required for the action of GDNF via GFRα1, but not for the action of NRTN, as TGF-β is increasing the surface location of GFRα1 but not of GFRα2.

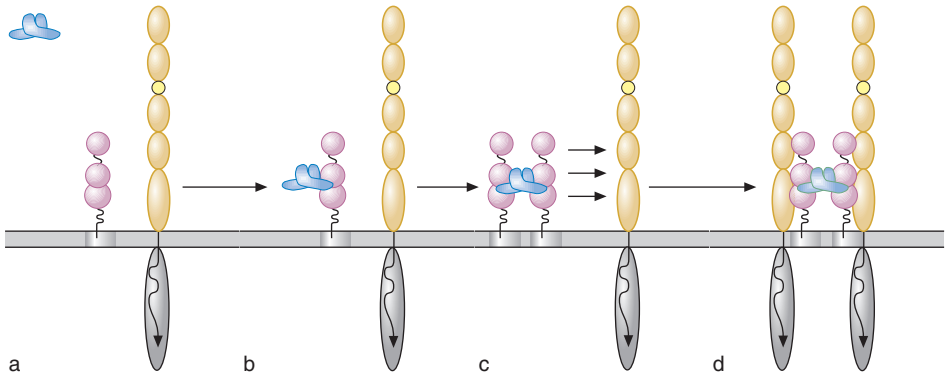

Figure 3 GDNF interacts with coreceptor GFRα1 and activates RET. (a) Biologically active GDNF is a homodimer. In the absence of GDNF, GFRα1 and RET interaction is very weak. (b) A dimer of GDNF binds to the second CRD of GFRα1. (c) GDNF brings together two molecules of GFRα1. (d) GDNF-GFRα1 complex induces dimerization of two molecules of RET leading to transphosphorylation and activation of their tyrosine kinase domains.

Two proteins that are structurally similar to GFRα proteins – GAS1 and GFRAL – are described, but their relatedness to the GFL system is not yet clear.

NCAM

NCAM plays an important role in neural development as the homophilic cell adhesion molecule. $P140^{NCAM}$ can also function as the alternative signaling receptor for all four GDNF family ligands. GFLs can directly bind to NCAM with low affinity, but do not trigger intracellular signaling. In the presence of cognate GFRα coreceptors GFLs can bind to $p140^{NCAM}$ with higher affinity and activate Src-like kinase Fyn and focal adhesion kinase FAK (Figure 4(a)). GFRα1 interaction with NCAM even in the absence of GDNF can inhibit NCAM-mediated cell adhesion. Thus GDNF–GFRα1–NCAM and GFRα1–NCAM interactions may have different biological readouts and physiological consequences. Current experiments indicate that GDNF by binding to GFRα1 and activating NCAM stimulates Schwann cell migration, hippocampal neurite outgrowth, and development of olfactory neurons in RET-independent manner. Further experiments are needed to unravel the *in vivo* roles of GDNF–GFRα1–NCAM and GFRα1–NCAM signaling and possible cross-talk of RET and NCAM in GFL signaling.

Syndecans

Although the vast majority of GFL signaling utilizes GFRα/RET or GFRα/NCAM receptor systems, novel data indicate that additional GFL receptors exist. GDNF was originally purified by heparin-affinity chromatography, and later its interaction with heparan sulfates was documented. Recent experimental data indicate that heparan sulfates are required for GDNF signaling via GFRα1 and RET. GDNF, ARTN, and NRTN, but not PSPN, can with high affinity bind to heparan sulfate proteoglycan (HSPG) syndecans (Figure 4(b)). Syndecans that bind GDNF can function as coreceptors and deliver GDNF to GFRα1/RET complex or activate directly Src signaling pathway triggering cell spreading and inducing neurite outgrowth of hippocampal neurons. GDNF interaction with syndecan also stimulates the migration of cortical neurons. Thus, GDNF can signal in two different ways – as a soluble protein and as a matrix-bound protein. In a soluble form it binds to GFRα1 and activates intracellular signaling via RET or NCAM. When bound to ECM (both transmembrane and soluble HSPGs), GDNF triggers qualitatively different signaling that is currently poorly defined.

Signaling Pathways

The interactions of the intracellular part of RET with various signaling molecules are mainly triggered by the cognate ligands (GFLs), but in some cases also GFL-independently. GFL–GFRα homodimeric complex brings two molecules of RET together, triggering transphosphorylation of specific tyrosine residues of RET, followed by the activation of intracellular signaling cascades (Figures 3 and 5). Several cascades can be activated, which regulate cell survival, differentiation, proliferation, migration, chemotaxis, branching morphogenesis, neurite outgrowth, synaptic plasticity, etc. The mitogen activated protein (MAP) kinase pathway is involved in neurite outgrowth of the neurons. The phosphoinositide 3-kinase (PI3K)-Akt pathway is responsible for both neuronal survival and neurite outgrowth. GFL signaling also activates Src-family kinases, which elicit, for example, neurite outgrowth, neuronal survival, and kidney ureteric branching. In most cases tyrosine residues

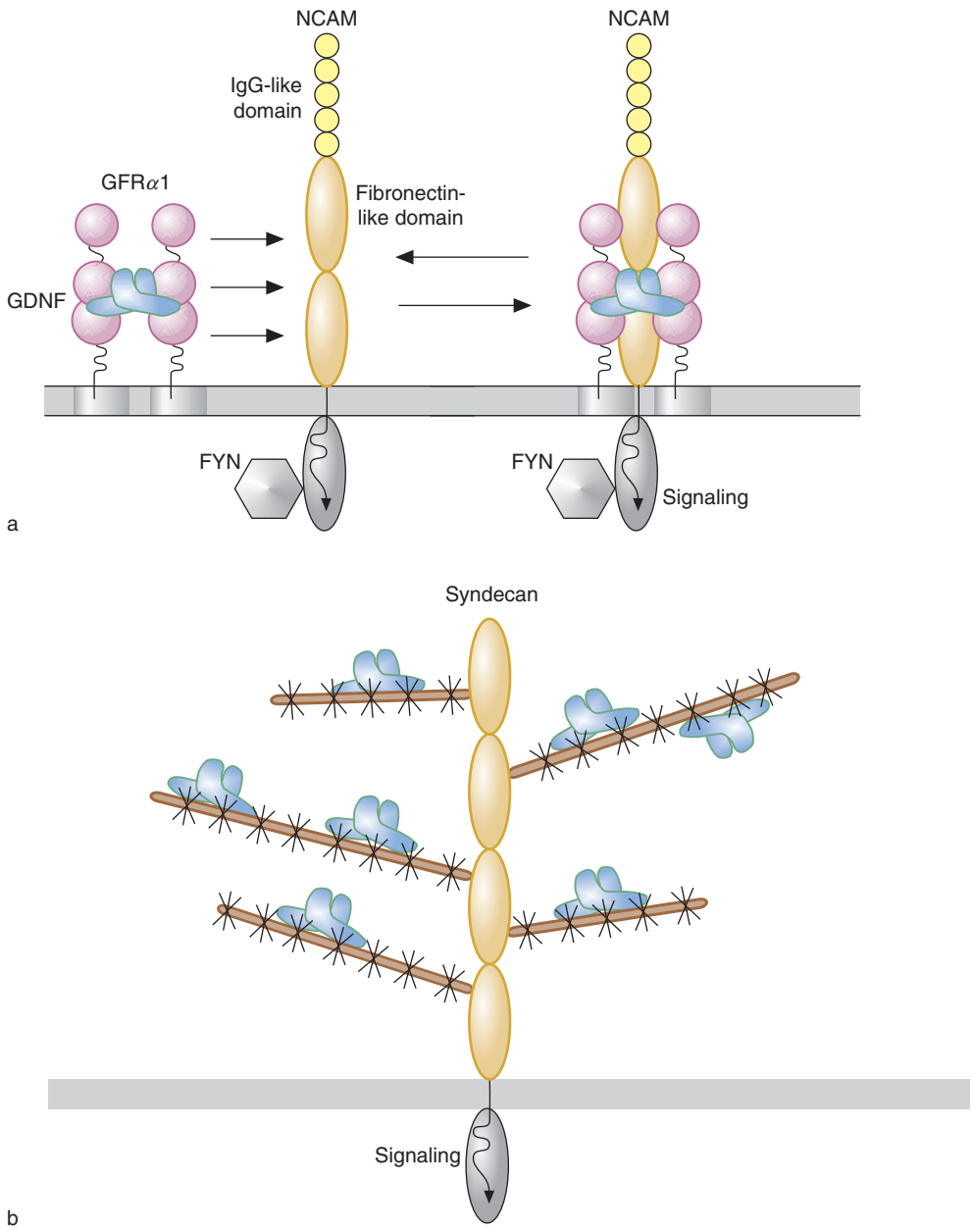

Figure 4 RET-independent signaling by GFLs. (a) p140^{NCAM} is an alternative signaling receptor for GFLs. It interacts with a GDNF–GFRα1 dimer leading to the activation of Fyn, a Src-like kinase. It is unclear, whether GDNF triggers NCAM dimerization. NCAM has five IgG-like domains and two fibronectin-like domains. (b) HSPG syndecans carry heparan sulfate side chains that bind with high affinity GDNF, ARTN, and NRTN. Unlike GFRα1, syndecans can bind many GFL molecules simultaneously. GDNF binding to syndecans can activate Src-family kinases or modulate signaling through GFRα-RET.

Tyr905, Tyr981, Tyr1015, Tyr1062, and Tyr1096 of RET are phosphorylated and docking proteins and intracellular proteins are bound and activated through interaction with these and possibly other RET phosphotyrosines (Figure 5). Interestingly, RET activation affects different downstream targets inside and outside lipid rafts that are the dynamic assemblies of cholesterol and sphingolipids scattered within the disordered phase of the lipid bilayer. Differences in GDNF signaling through RET within and outside the rafts could lead to significantly different cellular responses.

RET is also activated by several GFL-independent pathways. First, increase in the intracellular concentration of cyclic AMP level triggers protein kinase A (PKA)-dependent Ser696 phosphorylation of RET independently of GFL that regulates the migration of enteric neural crest cells in the developing gut. Second, in the mature sympathetic neurons, RET51

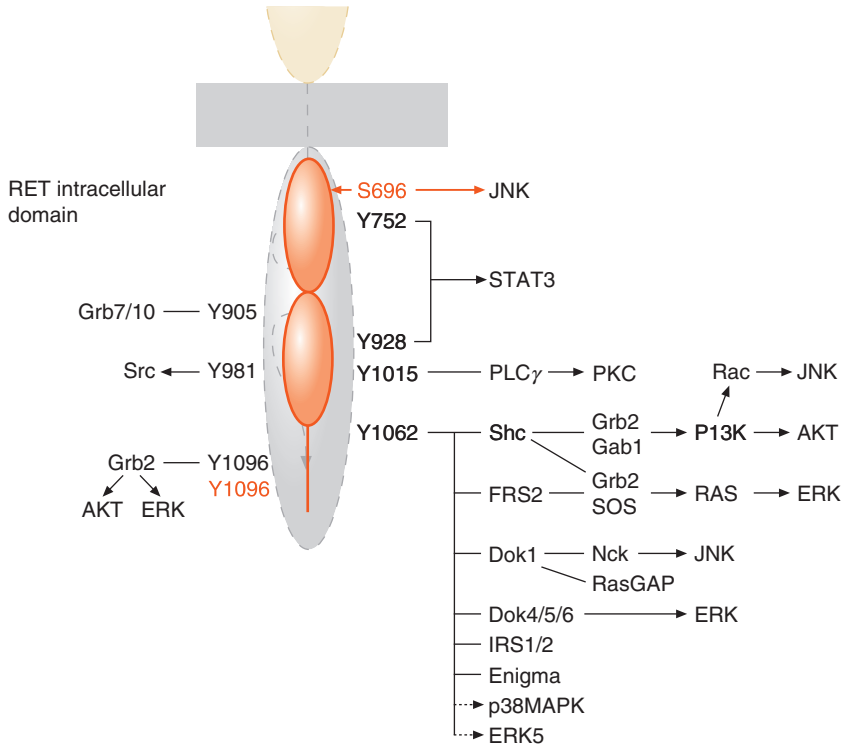

Figure 5 GDNF-induced signaling pathways. The model shows the intracellular domain of RET and highlights the phosphorylated tyrosines and different intracellular RET-binding proteins and activated downstream signaling pathways. The phosphorylated tyrosine residues (Y752, Y905, Y928, Y981, Y1015, Y1062, Y1096) activate multiple signaling pathways. Serine 696 (orange) is phosphorylated GFL-independently by PKA, and Y1096 (orange) can be phosphorylated GFL-independently by NGF.

isoform becomes phosphorylated and functionally activated by nerve growth factor that signals via unrelated receptor TrkA. The physiological importance of this signaling is not yet known.

Third, there is *in vitro* evidence that RET is also a dependence receptor that triggers apoptotic cell death when not bound by the ligands. However, the *in vivo* relevance of this RET-triggered apoptosis is completely unclear.

Biology and Physiology

Mice that lack GDNF, GFRα1, or RET die soon after birth, whereas mice lacking other GFLs or GFRα coreceptors or alternative receptors NCAM or syndecans are viable and fertile. The neuronal phenotypes of the different GFLs and their receptor-knockout mice are summarized in Table 1. The strongly overlapping phenotypes of ligand and coreceptor knockouts demonstrate, under physiological conditions, a specific pairing of each GFL and corresponding GFRαs. The vast majority of cells and tissues that are affected in GFL- or GFRα-knockout mice also express RET, indicating that *in vivo* this is the main signaling receptor for GFLs. GDNF-, GFRα1-, or RET-deficient mice share a phenotype of kidney agenesis, and an absence of many parasympathetic and enteric neurons. Mice that lack NRTN or GFRα2 have similar deficits in enteric and parasympathetic innervation. Characterization of ARTN- and GFRα3-deficient mice revealed similar abnormalities in the migration and axonal projection pattern of the entire sympathetic nervous system. Ablation of GFRα4 or its ligand PSPN impairs thyroid calcitonin production in young mice. Both GDNF- and NCAM-deficient mice have impaired migration of the rostral migratory stream-derived neuronal precursors. Likewise, GDNF- and syndecan-deficient mice have an impaired migration of GABAergic cortical neurons. In some cases the ligand- and its receptor-deficient mice phenotypes differ.

Mice lacking the long RET51 isoform seem to be normal, whereas mice lacking the short RET9 isoform were similar to mice lacking all RET isoforms. Only the short isoform can rescue the phenotype of the RET-null mutation in the kidney and enteric nervous system. In another study, however, homozygous RET9 and RET51 mice were viable and show normally developed kidneys.

Knock-in mice with mutated Tyr1062 of RET had defects in the enteric nervous system similar to those

Table 1 Phenotypes of mice lacking GFLs and their receptors

Gene knockout	*RET*	*GDNF/GFRα1*	*NRTN/GFRα2*	*ARTN/GFRα3*	*PSPN/GFRα4*
Gross phenotype	P0 lethal	P0 lethal	Viable, fertile. Pseudoptosis growth retardation[b]	Viable, fertile ptosis	Viable, fertile *Gfra4*−/−
PNS					ND
Sensory: PG	Breathing defect	<40%, breathing defect Soma reduced			
DRG			Soma size reduced, loss of heat sensitivity	NS	
TG		Loss in whisker follicles			
Autonomic					
Sympathetic	SCG; migration defect	SCG; <35% in *Gdnf*−/−mice NS in *Gfra1*−/− mice	NS	SCG and other ganglia; defect in migration and axon growth	
Sympathetic cholinergic			Soma size, target innervation reduced		
Parasympathetic					
Cranial: SPG	Lack of ganglia	Lack of ganglia	NS		
OG	Lack of ganglia	Lack of ganglia	40% reduced no. of neurons		
SMG	Reduced number and soma size	Reduced number and soma size	42% reduced no. of neurons	NS	
Enteric neurons	No neurons in bowel below stomach	No neurons in bowel below stomach	Moderate loss of fibers in small intestine	NS	
CNS					
Spinal motorneurons	Loss in various nuclei	22–31% loss of neurons	No gross defects	No gross defects	No gross defects
Brain	Substantia nigra dopamine neurons; NS	Impaired learning, reduced loco-motor activity[a]	Impaired behavioral flexibility and memory[b]	No gross defects	*Pspn*−/−: hypersensitive to cerebral ischemia
Other tissues	No kidneys, moderate thyroid C-cell loss	No kidneys, testis degeneration in adult *Gdnf*+/− mice	ND	ND	*Gfra4*−/− mice; calcitonin levels reduced in young mice

[a]*Gdfn*+/−.
[b]*Gfra2*−/− mice.
NS, not significantly different from wild-type; ND, not determined; PG, petrosal ganglion; DRG, dorsal root ganglion; TG, trigeminal ganglion; SCG, superior cervical ganglion; SPG, sphenopalatine ganglion; OG, otic ganglion; SMG, submandibular ganglion. Adapted from Airaksinen MS and Saarma M (2002) The GDNF family: Signaling, biological functions and therapeutic value. *Nature Reviews Neuroscience* 3: 383–394.

observed in RET-deficient mice but with only slight histological changes in the kidneys.

Dopamine neurons. GDNF and NRTN are potent survival factors for midbrain dopamine (DA) neurons and they signal through the GFRα1–RET receptor complex, which is expressed in developing and mature DA neurons. GDNF, but not NRTN, can also promote axon growth and sprouting of DA neurons. Analysis of GDNF-, NRTN-, GFRα1-, and RET-deficient mice reveals that GDNF signaling through GFRα1/RET is not required for the embryonic development of DA neurons. However, DA neurons of the substantia nigra undergo programmed cell death (PCD) during postnatal development. This is a biphasic process with an initial major peak just after birth and a second minor peak 2 weeks after birth. Intrastriatal injection of GDNF diminishes the PCD event and GDNF-neutralizing antibodies augment it. Thus, GDNF seems to serve as a physiological target-derived neurotrophic factor for DA neurons during

the first phase of PCD. This is also supported by the observation that GDNF is expressed prominently in the striatum during the first few postnatal weeks, when DA neuron target innervation is taking place. Results from mice, where RET or GFRα1 will be deleted from DA neurons, should demonstrate whether RET signaling is required for postnatal survival and maintenance of DA neurons.

Motor neurons. *In vitro*, GDNF is the most potent survival factor for motor neurons (MNs). In GDNF- and GFRα1-deficient mouse embryos, there is an increased (20–40%) loss of spinal and cranial MNs. RET-deficient mice also show significant losses in all motor neuron populations examined. Motor neuron survival is usually supported by synergistically acting trophic factors that are produced by muscle and glial cells. During development, GDNF is produced primarily by Schwann cells and by astrocytes throughout the nervous system. Muscle-specific expression of GDNF transgene or GDNF treatment *in utero* enhances survival of the MNs. In mice overexpressing GDNF under the control of an astrocyte-specific GFAP promoter, the survival of MNs in brachial, lumbar, thoracic spinal cord, and the abducens nucleus was increased. GDNF also prevented MN PCD in several cranial motor nuclei. Furthermore, GDNF could support complete and relatively long-term survival of MNs following neonatal facial-nerve axotomy. Although exogenous NRTN supports motor neuron survival, this is mostly mediated by GFRα1, as no clear loss of MNs is found in NRTN- and GFRα2-deficient mice. Continued subcutaneous injections of GDNF postnatally, or transgenic overexpression of GDNF in skeletal muscle, produce hyperinnervation of neuromuscular junctions. GDNF seems to induce this extra innervation by promoting terminal branching. Therefore, the main role of GDNF in postnatal MNs is to promote terminal axon branching and synapse formation.

Parasympathetic neurons. In the cranial region, postganglionic parasympathetic neurons are located in small ganglia: ciliary, sphenopalatine, submandibular, and otic ganglia. In GDNF-, RET-, and GFRα1-knockout mice, the otic and sphenopalatine ganglia are absent, demonstrating that GDNF signaling is critically required for the development of these parasympathetic neurons. In addition, GDNF signaling through GFRα1/RET is required for the migration and proliferation of parasympathetic neuronal precursors during early embryonic development. Postnatally, NRTN is upregulated in target tissues and parasympathetic neurons continue to express RET and GFRα2. In line with that, NRTN- or GFRα2-deficient mice have deficits in parasympathetic target innervation in many regions. Taken together, there is a shift from GDNF to NRTN signaling in developing parasympathetic neurons, where both growth factors have distinct roles.

Enteric neurons. In mice that lack RET, GDNF, or GFRα1 enteric neurons and glial cells that are derived from the vagal and sacral neural crest are missing below the stomach. GDNF promotes the migration, proliferation, survival, and differentiation of multipotent enteric precursor cells that express cognate receptors. NRTN or GFRα2 mutant mice show a significant loss of the cholinergic, substance P-containing fine myenteric nerve plexus in the small intestine, and very little loss of enteric neurons. As the release of substance P *in vitro* from the NRTN-deficient mouse colon is reduced significantly, NRTN signaling could regulate neurotransmitter release. The function of GDNF and NRTN in the adult enteric nervous system is unknown.

Sympathetic neuron development requires chemoattractant ARTN signaling through GFRα3/RET for migration and initial axon outgrowth. The role of GFLs in the postnatal sympathetic system is poorly studied.

GDNF, NRTN, and ARTN *in vitro* support subpopulations of primary sensory neurons. Although GFLs can be survival-promoting factors for some sensory neurons, their specific roles for these neurons are poorly understood.

Kidney Development and Spermatogenesis

Although GDNF is mainly known as the neurotrophic factor, it also plays a crucial role in kidney development and spermatogenesis. In mice in which GDNF, RET, or GFRα1 is knocked out, the kidneys show aplasia or severe hypodysplasia. GDNF is the major mesenchyme-derived factor that activates GFRα1/RET complex in the ureteric bud, thereby regulating ureteric budding and branching during nephrogenesis. Mutation of Tyr1015 in both RET51 and RET9 isoforms results in severe renal anomalies. In addition, loss of RET9 (Y1062)-mediated Akt/MAPK activation resulted in renal agenesis or kidney rudiments, whereas mutation of this residue in RET51 had no obvious effect on AKT/MAPK activity and renal development. However, the precise role of GDNF/RET signaling in renal branching morphogenesis and the specific responses of ureteric bud cells to GDNF remain unclear. The localized expression of GDNF in the metanephric mesenchyme, together with several types of negative regulation, is important to elicit and correctly position the initial budding event from the Wolffian duct. GDNF also promotes the continued branching of the ureteric

bud. However, it seems not to provide the positional information required to specify the pattern of ureteric bud growth and branching, as its site of synthesis can be drastically altered with minimal effects on kidney development.

In the testis, GDNF is expressed by the Sertoli cells that regulate spermatogenesis in a paracrine manner, whereas RET/GFRα1 are found in the undifferentiated spermatogonia, including spermatogenic stem cells. Mice lacking one GDNF allele show partial depletion of spermatogenic stem cells, whereas mice overexpressing GDNF show clusters of undifferentiated spermatogonia that develop seminoma-like germ line tumors in adulthood. GDNF is thus a crucial paracrine regulator of the self-renewal and differentiation of spermatogonial stem cells. Actually, GDNF is the first factor known to determine the fate of spermatogenic stem cells. GDNF signaling is therefore an attractive target for the development of male contraceptives. However, such efforts are severely shadowed by the carcinogenic consequences of deregulated RET activation. NRTN, GFRα2, and NCAM are also expressed in the developing kidney and testis, but they have no renal or testis phenotype when knocked out.

The function of GDNF and its receptors in kidney and testis should be very seriously considered when RET/GFRα1 agonists or antagonists to treat cancer or neurodegenerative diseases are developed.

Clinical Potential

GFLs have received attention as potential therapeutic agents for the treatment of certain neurological diseases, in particular the Parkinson's disease (PD). All current therapies for PD are symptomatic and do not slow down or reverse the underlying degeneration of midbrain DA neurons. GDNF has been an attractive treatment option since its discovery, because it is a naturally occurring protein that *in vivo* can protect and rescue nigrostriatal DA neurons that degenerate in PD. In various animal models of PD, GDNF can prevent the neurotoxin-induced death of DA neurons and can promote functional recovery. In the first clinical trial GDNF that was delivered into the lateral ventricles of patients with PD was without therapeutic efficacy and caused side effects, including weight loss. Later trials tested delivery of the GDNF into the site of the major DA loss in PD, namely in the caudate putamen. In a small open-label study with five patients clear positive clinical effects were found. In contrast, a more recent double-blind, placebo-controlled trial on 34 patients has shown no such effects. In this study patients with moderately advanced PD were randomly assigned intraputamenal GDNF or placebo. In addition, anti-GDNF antibodies were reported in three patients out of 34. After several failures a question arises whether there is a future for GDNF in the treatment of PD. The ability to promote endogenous repair through the use of target-derived growth factor is still very attractive. This is further supported by recent improvements in GDNF delivery in animal models of PD that include: heparin coadministration, which increases GFL spread; novel viral vectors; and engrafted- or encapsulated-GDNF-producing neuronal cells, which maintain high levels of human GDNF production. Combined with safety measures, including means to control gene expression, these new vehicles for GDNF delivery look promising for PD therapy.

GDNF is a survival factor for MNs and therefore it may have clinical importance for the treatment of amyotropic lateral sclerosis. In addition, GDNF can be useful in the treatment of various neurological traumas, as it supports the survival and axon regeneration of both MNs and sensory neurons. The combination of different GFLs might be the most efficient to promote functional regeneration of the sensory axons. Peripheral nerve injury often leads to chronic neuropathic pain. ARTN and GDNF have shown their effect in animal experiments and can be considered as drug targets for pain treatment.

In experimental models of ischemia, exogenous GDNF administered before or just after anoxia can reduce ischemic brain injury. PSPN-deficient mice show normal development and behavior, but are hypersensitive to cerebral ischemia. More recent results highlight the importance of GDNF as a new target for drug addiction. In addition to GDNF, NRTN also has promise for PD therapy and for epilepsy, and GDNF and NRTN might modulate seizure susceptibility.

Small Molecules

All known growth factors that activate RTKs bind directly to the receptor and trigger RTK dimerization. The only exception is GFLs that first bind to the ligand-specific GFRα-coreceptor that together with the ligand trigger RTK RET dimerization and activation. Unique mechanism of RET activation by GFLs also offers unique opportunities for rational drug design. Small number of GFRα receptors compared to large number of homologous RTKs predict less stringent specificity requirements for the designed drugs. Using high-throughput screening system XIB4035, a novel non-peptidyl small molecule agonist for GFRα1 was described. This compound at 12 μM concentration induced RET phosphorylation and neurite outgrowth in Neuro 2A cells. Although the *in vivo* effects of XIB4035 are still unknown, the fact that a nonpeptidyl small molecule with high affinity binds to GFRα1 and

activates RET is remarkable and very encouraging. New information on the structure of GFL receptors and their complexes with GFLs is helpful to improve the specificity and efficacy of small molecules that will be beneficial for the treatment of PD.

Summary

There are four members of the GDNF ligand family (GFLs): GDNF, NRTN, ARTN, and PSPN. These small secreted dimeric proteins, distantly related to TGF-β, promote the survival of central and peripheral neurons. GFL binds and activates the transmembrane RTK RET via specific GDNF family coreceptors α (GFRα). GDNF binds preferentially to GFRα1, NRTN to GFRα2, ARTN to GFRα3, and PSPN to GFRα4. Upon activation, RET dimerizes, transphosphorylates its tyrosine residues, and triggers activation of signaling cascades leading to survival, growth, or differentiation of cells. GDNF protects DA and MNs *in vivo*, making it a candidate for therapeutic use in several neurodegenerative diseases. Knockout mice revealed the function of GFLs and their receptors. GDNF −/− +/− mice lack enteric neurons, as in human HSCR. GDNF- and its receptor-deficient mice also lack kidneys, showing that GDNF is important in early kidney development. GDNF is also important for the regulation of stem cell differentiation in the testis. GDNF and NRTN regulate the development of enteric and parasympathetic neurons, ARTN controls the migration and initial growth of sympathetic neurons, and PSPN may protect the brain from ischemic insult and regulates calcitonin secretion.

See also: Enteric Nervous System: Neurotrophic Factors; Glial Growth Factors; Growth Factors: Neuronal Atrophy; Nerve Growth Factor; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3.

Further Reading

Airaksinen MS and Saarma M (2002) The GDNF family: Signalling, biological functions and therapeutic value. *Nature Reviews Neuroscience* 3: 383–394.

Baloh RH, Enomoto H, Johnson J, and Milbrandt J (2000) The GDNF family ligands and receptors – Implications for neural development. *Current Opinion in Neurobiology* 10: 103–110.

Cacalano G, Farinas I, Wang LC, et al. (1998) GFRalpha1 is an essential receptor component for GDNF in the developing nervous system and kidney. *Neuron* 21: 53–62.

Eigenbrot C and Gerber N (1997) X-ray structure of glial cell-derived neurotrophic factor at 1.9 Å resolution and implications for receptor binding. *Nature Structural Biology* 4: 435–438.

Enomoto H, Araki T, Jackman A, et al. (1998) GFR alpha1-deficient mice have deficits in the enteric nervous system and kidneys. *Neuron* 21: 317–324.

Gill SS, Patel NK, Hotton GR, et al. (2003) Direct brain infusion of glial cell line-derived neurotrophic factor in Parkinson disease. *Nature Medicine* 9: 589–595.

Jain S, Encinas M, Johnson EM Jr., and Milbrandt J (2006) Critical and distinct roles for key RET tyrosine docking sites in renal development. *Genes and Development* 20: 321–333.

Jing S, Wen D, Yu Y, et al. (1996) GDNF-induced activation of the RET protein tyrosine kinase is mediated by GDNFR-[alpha], a novel receptor for GDNF. *Cell* 85: 1113–1124.

Kotzbauer PT, Lampe PA, Heuckeroth RO, et al. (1996) Neurturin, a relative of glial-cell-line-derived neurotrophic factor. *Nature* 384: 467–470.

Leppänen VM, Bespalov MM, Runeberg-Roos P, et al. (2004) The structure of GFRalpha1 domain 3 reveals new insights into GDNF binding and RET activation. *EMBO Journal* 23: 1452–1462.

Lin LF, Doherty DH, Lile JD, Bektesh S, and Collins F (1993) GDNF: A glial cell line-derived neurotrophic factor for midbrain dopaminergic neurons. *Science* 260: 1130–1132.

Manie S, Santoro M, Fusco A, and Billaud M (2001) The RET receptor: Function in development and dysfunction in congenital malformation. *Trends in Genetics* 17: 580–589.

Meng X, Lindahl M, Hyvönen ME, et al. (2000) Regulation of cell fate decision of undifferentiated spermatogonia by GDNF. *Science* 287: 1489–1493.

Oppenheim RW, Houenou LJ, Johnson JE, et al. (1995) Developing motor neurons rescued from programmed and axotomy-induced cell death by GDNF. *Nature* 373: 344–346.

Rossi J, Luukko K, Poteryaev D, et al. (1999) Retarded growth and deficits in the enteric and parasympathetic nervous system in mice lacking GFR alpha2, a functional neurturin receptor. *Neuron* 22: 243–252.

Schuchardt A, D'Agati V, Larsson-Blomberg L, Costantini F, and Pachnis V (1994) Defects in the kidney and enteric nervous system of mice lacking the tyrosine kinase receptor RET. *Nature* 367: 380–383.

Takahashi M, Ritz J, and Cooper GM (1985) Activation of a novel human transforming gene, RET, by DNA rearrangement. *Cell* 42: 581–588.

Tomac A, Lindqvist E, Lin L-F, et al. (1995) Protection and repair of the nigrostriatal dopaminergic system by GDNF *in vivo*. *Nature* 373: 335–339.

Trupp M, Arenas E, Fainzilber M, et al. (1996) Functional receptor for GDNF encoded by the c-RET proto-oncogene. *Nature* 381: 785–789.

Wang X, Baloh RH, Milbrandt J, and Garcia KC (2006) Structure of artemin complexed with its receptor GFRα3: Convergent recognition of glial cell line-derived neurotrophic factors. *Structure* 14: 1083–1092.

Glia and Stroke

G Trendelenburg and U Dirnagl, Humboldt University of Berlin, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Cerebral Ischemia

Ischemic stroke is the result of a reversible or permanent focal perfusion deficit in the territory of a brain artery. Although the early work on stroke pathophysiology focused on blood flow and metabolism, a shift has occurred toward biochemical and cellular mechanisms, and only recently have molecular and genetic aspects been added. It is becoming evident that circulatory, metabolic, cellular, and molecular mechanisms have to be considered if we are to understand, treat, and possibly prevent the consequences of focal cerebral ischemia. With this shift in focus, the contribution of nonneuronal and particularly glial cells to stroke pathophysiology is increasingly appreciated.

Recent work demonstrates that under physiological, as well as pathophysiological conditions, highly dynamic and complex glial–neuronal interactions secure brain function and survival. During the past decade, a number of molecular mechanisms by which brain cells die, or survive, after occlusion of a cerebral artery have been unraveled. In the core of ischemic lesions brain tissue undergoes necrosis, a rather passive disintegration of all cellular elements in which energy supply has been critically impeded for some time. In brain areas surrounding the densely ischemic core ('penumbra'), however, biochemical and molecular processes determine the ultimate fate of the tissue. Pathophysiological entities such as excitotoxicity, spreading depression, inflammation, and apoptosis are involved in postischemic cell death. After minor insults or in the penumbra, cellular stress can be so subtle that only certain types of neurons die with a considerable delay, whereas other neurons and nonneuronal cells survive. The fate of brain tissue after focal cerebral ischemia is thus determined by the degree and duration of ischemia and the ability of the threatened brain to mount a competent defense on the cellular level. Glial cells play important roles in the destructive, as well as protective cascades evolving in the focally ischemic or reperfused brain.

It has been demonstrated that there is a complex network of communication between brain endothelial cells, neurons, and astrocytes. This network has been termed the 'neurovascular unit,' and it is believed to play a key role in regulating cerebral blood flow and in maintaining the integrity of the blood–brain barrier (BBB). The position of the astrocyte between the neuron and the blood vessel as part of the neurovascular unit emphasizes its potential to influence and coordinate the activities of both. Glial cells are sensors of changes in the brain environment, to which they immediately react on the genomic (e.g., production of trophic factors), as well as nongenomic level (e.g., uptake of ions). Glial cells, like astrocytes, not only support and protect themselves and all the other cellular elements (including neurons) with which they are in intimate contact but also actively participate in the demise of brain tissue after ischemia.

In such a complex organ as the brain, and with such a complex stimulus as ischemia, it is very difficult to experimentally dissect protective from destructive mechanisms because they coexist in time and space. To facilitate the exploration of protective mechanisms, researchers have turned to an experimental paradigm called 'ischemic preconditioning' or 'ischemic tolerance.' The preconditioning paradigm may also help to explain why the same mechanism or mediator may be protective or destructive in cerebral ischemia: whether a signaling pathway leads to survival or cell death depends on stimulus strength and context.

Ischemic Tolerance

In many, if not all, tissues and phyla, noxious stimuli, given below the threshold of damage, are able to induce protection (tolerance) against a subsequent deleterious stimulus of the same or even another modality (cross-tolerance). Although many specific cellular and molecular elements of the protective cascade have been identified, most researchers agree that there is not one cell type or one mechanism of neuroprotection in ischemic preconditioning. Rather, a concerted action of different pathways is postulated, which affects almost all pathophysiological mechanisms of ischemic cell damage. Preconditioning reduces the deleterious inflammatory response, adapts energy metabolism, as well as enhances neuroprotective mechanisms on various other levels. Specifically, it was shown that ischemic tolerance reprograms the response to ischemia, leading to a protected state. Because of the very close interaction between neuronal and nonneuronal elements in brain physiology and pathophysiology in general, we herein examine the principal mechanisms of ischemic damage and neuroprotection and discuss what is known about the specific contribution of glial cells.

Pathophysiology of Focal Cerebral Ischemia

Early Mechanisms: Excitotoxicity and Peri-Infarct Depolarizations

Especially within the core of the ischemic territory, blood flow deficits, low ATP levels and energy stores, ionic disruption, and metabolic failure are severe. Energy depletion due to lack of substrate (oxygen and glucose) leads to the depolarization of neurons but also of astroglia, resulting in the release of glutamate into the extracellular space. At the same time, the energy-dependent reuptake of glutamate by astroglial cells is impeded, further increasing its extracellular concentration. Under excitotoxic conditions, glia is thought to be neuroprotective due to its ability to clear extracellular glutamate. However, despite protecting neurons against glutamate or *N*-methyl-D-aspartate (NMDA) toxicity, the presence of astrocytes exacerbates α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) neurotoxicity *in vitro*. As a result of receptor activation (in particular of the neuronal NMDA and AMPA receptor subtypes), intracellular calcium rises. It was shown that in the penumbra, the neurotransmitter pool of glutamate coming from astrocytic glutamine was preserved while neuronal metabolism progressively declined, implying that astrocytic glutamine contributed significantly to glutamate excitotoxicity. This is somewhat paradoxical: due to the expression of the high-affinity glutamate transporters EAAT1/GLAST and EAAT2/GLT-1, astrocytes are thought to represent the most important cells for limiting the excitotoxic potential of glutamate, and in agreement, downregulation of both produces excitotoxic neurodegeneration and exacerbates ischemic injury, whereas ischemic preconditioning was shown *in vitro* to increase cellular glutamate uptake and expression of astrocytic EAAT2/GLT-1 glutamate transporters.

In summary, astrocytes have the capacity to limit excitotoxicity by clearing glutamate, but under certain conditions they may contribute to damage by providing precursors for excitotoxin synthesis. It is believed by many researchers that the reversed uptake of glutamate by neuronal and/or astrocytic glutamate transporters is the primary cause of ischemia-induced neuronal death. Ischemic stress to the brain *in vivo* or cultured brain cells induces a biphasic increase in extracellular concentration of glutamate, the first phase of which is dependent on Ca^{2+} and caused by the release of glutamate from nerve endings, whereas the second phase is caused by reversed uptake of glutamate through neuronal and/or astrocytic glutamate transporters. This is in agreement with previous studies which suggested that the activation of neuronal NMDA receptors is essential to preconditioning-induced ischemic tolerance.

The activation of metabotropic receptors and phospholipase C and IP3 signaling mobilizes calcium from intracellular stores, further contributing to the calcium overload. The increase in the universal second messenger, calcium, overstimulates numerous intracellular pathways and enzymes. Cell membranes are degraded (e.g., via activation of cyclo-oxygenase) and free radical species are produced in high concentrations, overwhelming endogenous scavenging mechanisms, which are mainly provided by astroglial cells. Neurons, in contrast to astroglia, depend almost exclusively on mitochondrial respiration. Neuronal mitochondrial function is impaired, and particularly the appearance of the so-called mitochondrial permeability transition pore in the mitochondrial membrane leads to mitochondrial swelling, cessation of ATP production, and possibly an oxygen free radical burst. In addition, the entry of sodium and water leads to swelling (edema) and even osmotic lysis of the cells. Due to the depolarization-induced release of neuronal and glial potassium and glutamate, repetitive depolarizations propagate and the so-called peri-infarct depolarizations may occur. The depolarization-induced release of excessive amounts of glutamate from neurons is a key event in triggering downstream mechanisms of tissue damage and in regions of intermediate perfusion and remaining energy equivalents (the so-called 'penumbra'), and the glutamate buffering capacity of astrocytes may be critical in limiting damage to neurons by excitotoxicity.

Astrocytes are coupled by small pores, so-called gap junction channels. Thereby a cellular network is formed which integrates and modulates homeostasis in larger brain regions. Astrocytic gap junctions remain open during ischemic conditions, and they may link ischemic astrocytes in an evolving infarct with the surroundings. Gap junction coupling may contribute to ischemic cell death because proapoptotic substances can diffuse from dying cells to otherwise healthy cells; ATP is channeled from healthy to dying cells, critically depriving the periphery of the infarct from energy-rich phosphates; and gap junction coupling may contribute to the propagation of spreading depressions. Since peri-infarct depolarizations contribute to lesion growth, inhibition of such depolarization waves may be at least partially responsible for the neuroprotection observed after astrocytic gap junction inhibition in experimental models of stroke.

However, the results of these studies conflict with evidence for a protective role of gap junctions. Gap

junctions are responsible for the special buffering capacity of astrocytes, allowing elevated extracellular K^+ during increased neuronal activity to be dispersed. Mice lacking one allele of connexin-43, the major constituent of astrocytic gap junctions, display impaired gap junction coupling, attenuated intercellular calcium signaling, enhanced neuronal apoptosis, and increased infarct volumes after experimental stroke. Furthermore, a role for connexin-43 in mediating ischemic preconditioning has been established.

Besides their metabolic role, astrocytes are thought to play a strategic role with regard to development of brain swelling (edema) and water imbalance following cerebral ischemia. Specific channel proteins, the aquaporins (AQPs), have been identified which are abundantly expressed in astrocyte pericapillary endfoot processes and which play an important role in astrocyte swelling and cerebral edema formation. Although conflicting data exist with regard to AQP4 regulation in cerebral ischemia, downregulation of AQP4 protects against brain edema and AQP4 knockout mice display a reduced neurological deficit after cerebral ischemia.

Astrocyte swelling – which may result from water flux across AQP channels or high levels of glutamate and K^+, lactic acid, oxygen free radicals, and arachidonic acid, among others – is one of the earliest morphological changes detectable in focally ischemic brain tissue. Local cell swelling may also be responsible for further reducing tissue perfusion by vascular compression, and neuronal swelling is an early consequence of the breakdown of energy metabolism, whereas the extent of brain edema is attenuated in a model of induced tolerance.

Astrocytes are essential elements of the BBB. Astrocytes lacking glial fibrillary acidic protein (GFAP) have been shown to have defects in process formation, induction of the BBB, and volume regulation. In this context, it is interesting that GFAP null mice are highly susceptible to damage by focal cerebral ischemia. However, astrocytes can release trophic factors, which regenerate and protect the BBB: ischemic and hypoxic preconditioning induces astrocyte-derived factors such as vascular endothelial growth factor (VEGF) and adrenomedullin, which are thought to regulate BBB properties. Secretion of VEGF by activated astrocytes has been shown to be required for ischemic preconditioning.

Astroglia is the principal repository of glycogen in the brain. Studies have reported an increase in basal glycogen within the brain after hypoxic preconditioning. As intrinsic anaerobes, astrocytes may be able to transiently support anaerobic metabolism and support the structural integrity of neurons in the penumbra.

Another important aspect of the astrocytic contribution to early mechanisms of stroke pathophysiology is their role in defense against oxidants. Following cerebral ischemia and reperfusion, astrocytes are less vulnerable to reactive oxygen species (ROS) than neurons because they contain high levels of antioxidants and can initially maintain ATP levels via glycolysis. Astrocytes provide neurons with substrates for antioxidants, such as glutathione.

Moreover, metallothioneins (MTs) are expressed by astrocytes and act as scavengers of ROS, as well as of neurotoxic metals such as zinc. They are highly inducible by various stimuli including cerebral ischemia, they are neuroprotective *in vitro*, as well as *in vivo*, and they have been shown to be involved in ischemic tolerance. Thioredoxin (TRX), which scavenges ROS and regulates various intracellular molecules by modifying their redox status, is also induced in glial cells in cerebral ischemia *in vivo* and released upon exposure to H_2O_2 *in vitro*, and overexpression of TRX has been shown to attenuate focal ischemic brain damage.

Very little is known about the role of oligodendrocytes in this early phase of cerebral ischemia. Most of our knowledge is derived from work on white matter rather than gray matter, and there may be relevant differences in the pathophysiology of ischemia and the role of oligodendrocytes in both regions. Interestingly, and in contrast to previous concepts, oligodendrocytes appear to be highly vulnerable to cerebral ischemia. Early, non-Wallerian damage to oligodendrocytes seems to be mediated via AMPA receptors, which, in contrast to NMDA receptors, are abundant on oligodendrocytes. These findings may be relevant from a clinical standpoint because they could at least in part explain the failure of NMDA receptor antagonists in human stroke trials. NMDA receptor antagonists significantly reduce infarct sizes in many stroke models. However, in contrast to AMPA receptor antagonists, they do not protect oligodendrocytes. Thus, although neurons may be rescued by NMDA receptor blockade, they may remain nonfunctional due to the axonal damage resulting from oligodendrocyte death.

Inflammation

Focal cerebral ischemia is a potent inducer of inflammation. Astrocytes, microglia, as well as invading leukocytes are traditionally regarded as the key cellular elements of postischemic inflammation. However, although glial cells are the focus of this article, it should be kept in mind that endothelial cells are crucial in initiating and modulating inflammation by attracting inflammatory cells and regulating their

activity. The induction of intercellular adhesion molecule (ICAM)-1 on endothelium parallels the mobilization of CD11b/CD18 (Mac-1) on circulating leukocytes. Consequently, administration of monoclonal antibodies against CD18, CD11b, or ICAM-1 was found to significantly reduce leukocyte accumulation and stroke severity in experimental models of cerebral ischemia.

Disturbed central nervous system (CNS) homeostasis and disrupture of the BBB is associated with invading immune cells and production of a plethora of different cytokines by activated glia. As in other tissues, inflammation in the brain has two contrasting aspects: although the deleterious role of inflammation after stroke is acknowledged by most researchers, protective and particularly regenerative aspects of ischemia-induced inflammation have been suggested. Indirect evidence exists that certain types of inflammatory reactions are neuroprotective and neuroregenerative. Activated glial cells produce a host of trophic cytokines, and macrophages or T cells exposed to certain CNS antigens *ex vivo* partake in tissue repair and recovery after nerve transection and spinal cord injury. Moreover, lipopolysaccharide (LPS), the prototypic stimulus for expression of proinflammatory cytokines, can precondition the brain against subsequent ischemia.

Astrocytes appear to play a key role in brain inflammation after stroke. This is hardly surprising because astrocytes initiate and regulate immune responses through the release of numerous proinflammatory cytokines, chemokines, and components of the complement system.

Microglia and macrophages are thought to be the key effector cells of the postischemic inflammatory response. Inhibition of microglial activation by tetracylines or the deficiency of tissue plasminogen activator have been shown to protect the brain against focal cerebral ischemia in rodents. Microglia become activated within hours after focal cerebral ischemia. At this time after transient ischemia, neutrophils accumulate in the brain parenchyma and contribute to neuronal damage. With the expression of macrophage inflammatory protein-1 in microglia and monocyte chemoattractant protein-1 in astrocytes, a switch from neutrophil to macrophage chemoattraction occurs 1 or 2 days after the onset of ischemia. Although the immune function of microglia is evident, the ability of astrocytes to function as nonprofessional antigen-presenting cells in the diseased brain remains controversial. Nevertheless, astrocytes contribute to the production of proinflammatory cytokines and neuroprotective factors, such as transforming growth factor (TGF)-β and metallothionein-II, after cerebral ischemia.

In addition to the acute detrimental effects of the postischemic inflammatory response, inflammation is also linked with repair and plasticity. Activated macrophages and microglia may release neurotrophic factors, and interleukin (IL)-1β has been shown to stimulate the synthesis of nerve growth factor. Mice deficient in tumor necrosis factor (TNF) receptors had significantly larger infarcts than wild-type mice after transient ischemia and reperfusion. Moreover, TNF-α has been shown to promote the remyelination of damaged axons by promoting the proliferation of oligodendrocyte progenitors during inflammation.

IL-1β and TNF-α Under ischemic conditions, a variety of inflammatory mediators, such as IL-1, IL-6, IL-8, TNF-α, and monocyte chemoattractant protein-1 (MCP-1), are generated by diverse cell types including astrocytes. A dual role for neuronal outcome has been demonstrated for some of them. The chemokine MCP-1, for example, which is expressed in astrocytes and induced after cerebral ischemia, has been shown to cause neuronal death via direct apoptotic mechanisms through specific cell surface chemokine receptors and possibly, in the long term, neuronal survival through the induction of basic fibroblast growth factor production by astrocytes.

The two key proinflammatory cytokines in the setting of cerebral ischemia are IL-1β and TNF-α, both of which are primarily produced by activated microglia and macrophages. The biological effects of IL-1β and TNF-α are diverse and include the conversion of endothelium to a prothrombotic surface, the induction of leukocyte adhesion molecules, an increase in BBB permeability, and the activation of macrophages. Inhibition of either IL-1β or TNF-α activity by monoclonal antibodies or recombinant receptor antagonists after focal cerebral ischemia significantly reduced infarct sizes in rats. Moreover, mice expressing a dominant negative mutant of IL-1β converting enzyme developed smaller strokes.

Interleukins such as the lesion-induced interleukins of the IL-6 family are thought to exhibit a dichotomous type of action. They contribute at least partially to tissue damage, as well as induce protective mechanisms (e.g., induction of protective genes such as nerve growth factor and metallothionein). Similarly, IL-1, in contrast to the IL-1-like cytokine IL-18, mainly contributes to acute brain damage but also has been identified as a mediator of ischemic tolerance and is thought to be necessary for astroglial proliferation.

TNF-α is induced in both glial and neuronal cells after ischemic and excitotoxic brain injury. TNF-α dose dependently has neuroprotective properties despite its role as a prototypic proinflammatory

cytokine. TNF-α-mediated ischemic tolerance has been ascribed to the downstream messenger ceramide, the neuronal glutamate transporter EAAT3, and a p300-involving mechanism of selective suppression of TNF-α-induced harmful genes in preconditioned cells.

Inducible nitric oxide synthase Studies suggest that inducible nitric oxide synthase (iNOS) produced by infiltrating inflammatory cells and astrocytes also contributes to postischemic damage, possibly by the generation of toxic peroxynitrite and ROS. Pharmacological blockade of iNOS blocked the development of tolerance to focal cerebral ischemia induced by volatile anesthetics. Thus, a role for iNOS in mediating ischemic preconditioning appears likely. NO produced by iNOS also activates matrix metalloproteinase-9 (MMP-9), thereby linking MMP signaling with the well-recognized NO pathway in stroke. 'Activated' astrocytes express MMPs such as MMP-2 and MMP-9 and their inhibitors, the family of tissue inhibitors of metalloproteinases (TIMPs). The deleterious effect of MMP-9, which is tightly regulated and induced in ischemic brain tissue, is demonstrated by the protective effect of MMP-9 deficiency.

TGF-β and IL-10 The overwhelming proinflammatory signaling following ischemic brain damage is counterbalanced by the production of cytokines with homeostatic and potentially neurotrophic action, such as TGF-β1 and IL-10. TGF-β1 suppresses the translation of cytokine mRNAs and deactivates macrophages. IL-10 is secreted from peripheral blood mononuclear cells after ischemic stroke. Both TGF-β1 and IL-10 have been shown to protect rodents from ischemic injury, and it has been suggested that bacterial LPS-induced ischemic tolerance in the brain is mediated via TGF-β. It has been shown that murine heat shock protein-32 (Hsp32) mediates the anti-inflammatory effect of IL-10. Heat shock proteins are stress-induced protective genes thought to act as molecular chaperones, but numerous other functions have been described, including direct antiapoptotic effects. Whereas induction of Hsp72 has been observed mainly in neurons, astrocytes express Hsp32 (heme oxygenase-1). Induced Hsp32 protected neurons against ischemic damage *in vivo* and hypoxic preconditioning induces Hsp32 in combination with sustained astrocytic activation.

Of all the gene transcripts upregulated in the ischemic hemisphere, metallothionein-II has been identified as the most significantly induced. This small, cysteine-rich molecule is synthesized by reactive astrocytes and has been shown to exert neuroprotective effects after cerebral ischemia. Purines are released on a massive scale after ischemic cell injury, and extracellular ATP has been found to transcriptionally activate microglia and astrocytes *in vitro*. Purines may contribute to the regeneration of damaged neuronal axons by enhancing the release and activity of neurotrophic factors.

Apoptosis

Inflammation and apoptosis in cerebral ischemia appear to have a similar temporal profile. In general, apoptosis is particularly prominent in stroke areas resulting from focal ischemia of short duration (30–60 min) and in the penumbral border zone, where selective neuronal degeneration may occur. In addition, days and weeks after infarction, neurons in areas projecting to the infarct zone may also undergo apoptosis, while microglial cells in these areas become activated, as a sign of either neuronal deafferentation or neuronophagia.

Glial cells may be involved in stroke-induced apoptosis in several ways: by producing signals that induce apoptosis in neurons, by dying apoptotically triggered by the same signals and in a manner similar to neurons, by self-elimination via apoptosis, and by producing factors that protect neurons against apoptosis.

All of these processes are closely associated with inflammation. Inflammation produces toxins (e.g., oxygen free radicals) that induce mitochondrial and DNA damage as important upstream mediators of apoptosis in neurons and oligodendrocytes. In addition, apoptosis is a critical mechanism in limiting and terminating inflammation, although it is controversial whether inflammatory cells are eliminated by apoptosis. Macrophages express Bax protein, and nuclear DNA fragmentation has been found in the infarcted tissue. It is very likely that astrocytes play a major role in limiting the growth of ischemic lesions, and that antiapoptotic mechanisms are a key element of astrocyte-mediated neuroprotection after stroke. Astrocytes are a major source of growth factors which protect neurons against ischemic damage. Reactive astrocytes produce nerve growth factor, basic fibroblast growth factor, brain-derived nerve growth factor, ciliary neurotrophic factor, insulin-like growth factor-1, and TGF-β, among others.

Nerve Growth Factor, Brain-Derived Neurotrophic Factor, and Glial Cell Line-Derived Neurotrophic Factor

Neuroprotective effects of nerve growth factor (NGF) and brain-derived neurotrophic factor (BDNF) in cerebral ischemia have been demonstrated *in vitro* and *in vivo*. Although neutrophins such as NGF or

BDNF are expressed in neurons, there is *in vitro* evidence for NGF production and NGF-like activity in astrocytes. Both BNDF and NGF were shown to be induced after ischemic preconditioning, and ischemic tolerance was mimicked by the application of BDNF *in vivo*.

Glial cell line-derived neurotrophic factor (GDNF) is another potent neuroprotective factor produced by astrocytes. GDNF is regulated after cerebral ischemia, astrocytes appear to be a major source of GDNF, and GDNF has a protective effect after ischemic brain injury.

Erythropoietin

The trophic cytokine, erythropoietin (EPO), has been shown to protect neurons from ischemic damage *in vivo* and *in vitro*. EPO is induced after cerebral ischemia, as well as after hypoxic preconditioning *in vivo*. Astrocytes are the main cellular source of the glycoprotein hormone EPO in the brain and low oxygen tension stimulates EPO mRNA expression in astrocytes via hypoxia-inducible factor-1. A synergistic effect of EPO and insulin-like growth factor-1 has been shown. EPO can be viewed as the prototypical paracrine protein by which astrocytes provide protection to neurons.

Regeneration and Repair

In most stroke survivors, functional deficits are most severe in the initial phase after the event and neurological function recovers at least to some degree within days, weeks, or months. Although this improvement has been attributed to the reorganization of brain function toward brain areas that have survived the insult, it is increasingly appreciated that regeneration and repair contribute to recovery after stroke. Glial cells, particularly astrocytes and microglia, may play an important role in this process.

The neuropathological hallmark of ischemic stroke is brain infarction, which is defined as pannecrosis, the necrotic disintegration of all cellular elements. The infarct may be surrounded by a zone of neuronal degeneration, historically termed selective neuronal necrosis, which is probably mechanistically more related to apoptosis. Depending on the severity of the ischemic insult and the type of stroke, the ischemic infarct and selective neuronal degeneration develop over a period of hours to days. Tissue breakdown and phagocytosis continue, and cystic spaces filled with fluid replace the parenchyma. At the border of the lesion, which is well demarcated at that time, swollen astrocytes are abundant and astrocytes proliferate. Trophic factors induced and released by reactive astrocytes include NGF, basic fibroblast growth factor, TGF-β, platelet-derived growth factor, BDNF, ciliary neurotrophic factor, neuropilin-1, and VEGF. Moreover, reactive astrocytes exhibit elongated processes in areas surrounding the ischemic infarction, which suggests that astrocytes also respond to injury by adopting a more immature phenotype. As the developing brain is conditioned to support cell genesis, growth, and synaptogenesis, a possible return to an immature state and the release of associated growth factors may provide a local environment conductive to regeneration and neurogenesis. As discussed previously, almost all pathophysiological mechanisms appear to be 'double edged,' and this seems to also be true for astrocytes and regeneration. Within a few days after the insult, near ischemic lesions in the brain a glial scar forms around the necrotic brain tissue of the infarct. This scar is mainly generated by reactive astrocytes, and it is composed primarily of astrocytes and the extracellular matrix molecules of proteoglycans. Although there is no doubt that the glial scar can limit regeneration in the adult mammalian by inhibiting axonal sprouting, the glial scar may be important for containing the injury site and preventing its extension.

Microglia at the infarct border and microglia/macrophages in the infarct core may produce astrocyte chemoattractants, such as osteopontin, and thereby organize the formation of the astroglial scar after stroke. Microglia may also contribute to tissue reorganization by secreting MMPs (particularly MMP-2), which lead to extracellular proteolysis and may mediate parenchymal and angiogenic recovery. Reactive astrocytes upregulate adhesion molecules and produce extracellular matrix proteins, such as neural cell adhesion molecule (NCAM), laminin, tenascin, chondroitin sulfate proteoglycans, and fibronectin. Some of these products inhibit neurite outgrowth (e.g., tenascin), whereas others (e.g., NCAM) promote neurite outgrowth. Both astrocytes and microglia/macrophages produce TGF-β1 and VEGF after stroke, which are important signals for the angiogenic response. In addition, they may provide signals leading to the proliferation of oligodendrocytes which can be observed around the lesion on day 5 after focal cerebral ischemia.

From the previous discussion, it is apparent that the contributions of glial cells to stroke pathophysiology, regeneration, and repair have only recently come into focus and that our knowledge is patchy at best. It is undisputed that glial cells are major contributors to tissue damage and repair after focal cerebral ischemia. The involvement of glial cells in stroke is considered to have contrasting aspects: whereas ischemia induces the production of cytotoxins in these cells, at the same time potent defense mechanisms are

turned on. Since the transcription and translation of the effector proteins of inflammation and apoptosis are delayed, they represent interesting new therapeutic targets given that only a minority of stroke patients can be recruited immediately after symptom onset.

Whereas the previous discussion focused on the contribution of glial cells to stroke pathophysiology along the temporal evolution of damage and repair, in the following sections we synthesize and summarize our current knowledge stratified for astrocytes, microglia, and oligodendrocytes.

Astrocytes

Astrocytes are comparatively resistant to substrate deprivation and are activated very early in and around a focally ischemic brain region. Reactive astrocytes are hyperthrophic and hyperplastic, and they can be identified by their increased expression of GFAP. At the molecular level, reactive astrocytes increase the production of numerous mediators, receptors, and channels which are responsible for the altered functional state of these cells in cerebral ischemia.

Closely related to their role in inflammation, astrocytes are sources of a host of trophic factors, which affect neuronal survival in a paracrine manner. *In vitro* studies have reported that the astroglial production of growth factors and cytokines is regulated via complex autocrine and paracrine loops. Trophic support from astrocytes not only provides protection in the acute phase after cerebral ischemia but also promotes neurogenesis, angiogenesis, and regeneration in the chronic phase.

Astroglial activation is thought to be protective via destruction of pathogens, removal of debris, and promotion of tissue repair. However, beyond a certain threshold, activation might spread tissue damage. Activated astrocytes can kill neurons in co-culture, and this may occur *in vivo*. The mechanisms by which activated astrocytes kill neurons include the release of high levels of NO, peroxynitrite ($ONOO^-$), glutamate, TNF, and IL-1β. Thus, astrocytes play a key role in the induction of the inflammatory reaction to cerebral ischemia. As infarction and selective neuronal degeneration progress, astrocytes in and around the lesion (and even in remote projection areas) remain activated and even proliferate. Astrocytes are the main cellular constituent of the scar that forms around the lesion within approximately 1 week after the event. Astrocytes, particularly reactive astrocytes, have been identified as a source and potential target of endothelins, VEGF, and angiopoietin I – three factors that are prominently involved in glia-mediated vasculogenesis, neurogenesis, directed migration of neural precursors, neuritogenesis, and glial plasticity.

Microglia/Macrophages

As for astrocytes, the involvement of microglia and macrophages in ischemic injury has contrasting aspects. Resident microglia are rapidly activated around the ischemic lesion. They transform from a resting, ramified morphology into rod cell-shaped and amoeboid phenotypes. Activated microglia proliferate and migrate to the site of injury. Their response is rather stereotypic, and the transformation of resident microglia into phagocytic cells occurs in the presence of neuronal cell death. At the same time, circulating monocytes and macrophages are recruited into the ischemic brain tissue. Activated microglia and macrophages produce proinflammatory cytokines, chemoattractants, and MMPs that sustain the inflammatory response. While clearing the areas of cell death, they are a source of potentially cytotoxic substances, such as NO, ROS, proteases, and excitotoxins. Microglia and macrophages participate in the immunological reactions to cerebral ischemia by upregulating major histocompatibility and costimulatory molecules and by presenting antigens to invading lymphocytes. They also secrete factors inducing apoptosis in neurons and oligodendrocytes. On the other hand, microglia and peripheral blood mononuclear cells produce neuroprotective cytokines and trophic factors after cerebral ischemia. As has been outlined previously, they also participate actively in tissue remodeling and angiogenesis.

Oligodendrocytes

Oligodendrocytes are exquisitely sensitive to hypoxia and ischemia, possibly even more so than neurons. This may have clinical consequences since some neuroprotective strategies might fail to protect oligodendrocytes. As a consequence, neuroprotectants tested in experimental models of stroke (e.g., NMDA receptor antagonists) may have saved neurons but not the oligodendrocytes responsible for maintaining axonal function of these neurons, which may at least partially explain the failure of these substances in clinical trials. Surprisingly little is known about the role of oligodendrocytes in regeneration and repair after stroke, although it is obvious that they must be involved in maintaining myelination or helping to remyelinate axons.

Conclusions

Given the tremendous complexity of cellular and molecular interactions in health and disease in general, and in the brain in particular, it is not surprising

that most of the issues discussed in this article remain tentative, and many questions remain unanswered. We should further take into account that any medically relevant biological signaling pathway is highly complex and consists of functionally independent sets of proteins which act in a coordinated way to produce phenotypic responses.

What are the pressing issues concerning the functional roles of glial cells in cerebral ischemia? With regard to the neuroprotective roles, we need to know how the same cell type or biochemical or molecular mechanism can promote either cell death or survival. Is it possible to devise therapeutic strategies to separate beneficial from deleterious mechanisms? Can we, by pharmacological or physical means, induce beneficial mechanisms selectively? This implies that there are pathways that are purely 'prosurvival.' Apparently this is the case for the astrocytic protein erythropoietin, for which only neuroprotective, but no detrimental, actions on the CNS have been reported.

A plethora of experimental evidence suggests that astrocytes protect neurons, as well as other elements of the CNS after cerebral ischemia. Astrocytes are effective sensors of the brain microenvironment. They are capable of detecting substrate deprivation, against which they acutely and chronically activate a host of countermeasures. They modulate inflammation, participate in the clearing of cell debris, and produce trophic factors to neurons and vascular cells. Increasing evidence indicates that there are molecular parallels between neural/neuronal and vascular development and maintenance, in which astroglia might have important orchestrating functions. However, under certain conditions, activated astrocytes are perpetrators and actively contribute to the demise of brain tissue after cerebral ischemia. By the repulsion of regenerating axons, reactive astrocytes and their proteoglycan products may actively impede brain reorganization. It is likely that the same stimuli, and possibly even the same mediators and effectors, which are involved in astrocytic protection are also part of astrocyte-mediated cell death. Further research is needed to determine whether and how astrocyte protection can be therapeutically utilized after cerebral ischemia. This could be done either by boosting astrocytic neuroprotection or by delivering therapeutics that mimic astrocyte-derived neuroprotectants.

In summary, glial cells are major contributors to damage, as well as to endogenous protection and repair after stroke. The fact that the same cell types partake in damaging, as well as in protective signaling precludes simple therapeutic approaches aimed at blocking or inducing the activities of glial cells. Nevertheless, the appreciation of the high susceptibility of oligodendrocytes to AMPA-mediated cell death and the successful pilot trial on the use of erythropoietin in stroke in humans are examples of the outstanding clinical relevance of glial mechanisms in stroke. Future studies will elucidate the interplay between glia, neurons, and other nonneuronal elements in stroke in greater detail.

See also: Animal Models of Stroke; Astrocyte: Response to Injury; Gap Junctions and Hemichannels in Glia; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Inflammation in Neurodegenerative Disease and Injury; Microglial Response to Injury; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neuroprotection: Endogenous Mechanisms; Stroke; Stroke: Injury Mechanisms.

Further Reading

Anderson MF, Blomstrand F, Blomstrand C, Eriksson PS, and Nilsson M (2003) Astrocytes and stroke: Networking for survival? *Neurochemical Research* 28: 293–305.

Carmichael ST (2006) Cellular and molecular mechanisms of neural repair after stroke: Making waves. *Annals of Neurology* 59: 735–742.

Chen Y and Swanson RA (2003) Astrocytes and brain injury. *Journal of Cerebral Blood Flow & Metabolism* 23: 137–149.

Chung RS and West AK (2004) A role for extracellular metallothioneins in CNS injury and repair. *Neuroscience* 123: 595–599.

Cotrina ML, Kang J, Lin JH, et al. (1998) Astrocytic gap junctions remain open during ischemic conditions. *Journal of Neuroscience* 18: 2520–2537.

D'Ambrosio AL, Pinsky DJ, and Connolly ES (2001) The role of the complement cascade in ischemia/reperfusion injury: Implications for neuroprotection. *Molecular Medicine* 7: 367–382.

Dirnagl U, Iadecola C, and Moskowitz MA (1999) Pathobiology of ischaemic stroke: An integrated view. *Trends in Neurosciences* 22: 391–397.

Dirnagl U, Simon RP, and Hallenbeck JM (2003) Ischemic tolerance and endogenous neuroprotection. *Trends in Neurosciences* 26: 248–254.

Farahani R, Pina-Benabou MH, Kyrozis A, et al. (2005) Alterations in metabolism and gap junction expression may determine the role of astrocytes as 'good samaritans' or executioners. *Glia* 50: 351–361.

Kirino T (2002) Ischemic tolerance. *Journal of Cerebral Blood Flow & Metabolism* 22: 1283–1896.

Liang D, Dawson TM, and Dawson VL (2004) What have genetically engineered mice taught us about ischemic injury? *Current Molecular Medicine* 4: 207–225.

Lin JH, Weigel H, Cotrina ML, et al. (1998) Gap-junction-mediated propagation and amplification of cell injury. *Nature Neuroscience* 1: 494–500.

Lo EH, Dalkara T, and Moskowitz MA (2003) Mechanisms, challenges and opportunities in stroke. *Nature Reviews Neuroscience* 4: 399–415.

Mabuchi T, Kitagawa K, Ohtsuki T, et al. (2000) Contribution of microglia/macrophages to expansion of infarction and response of oligodendrocytes after focal cerebral ischemia in rats. *Stroke* 31: 1735–1743.

Nedergaard M and Dirnagl U (2005) Role of glial cells in cerebral ischemia. *Glia* 50: 281–286.

Seifert G, Schilling K, and Steinhäuser C (2006) Astrocyte dysfunction in neurological disorders: A molecular perspective. *Nature Reviews Neuroscience* 7: 194–206.

Silver J and Miller JH (2004) Regeneration beyond the glial scar. *Nature Reviews Neuroscience* 5: 146–156.

Stenzel-Poore MP, Stevens SL, Xiong Z, et al. (2003) Effect of ischemic preconditioning on genomic response to cerebral ischemia: Similarity to neuroprotective strategies in hibernation and hypoxia-tolerant states. *Lancet* 362: 1028..

Swanson RA, Ying W, and Kauppinen TM (2004) Astrocyte influences on ischemic neuronal death. *Current Molecular Medicine* 4: 193–205.

Trendelenburg G and Dirnagl U (2005) Neuroprotective role of astrocytes in cerebral ischemia: Focus on ischemic preconditioning. *Glia* 50: 307–320.

Glia and Synapse Formation: An Overview

N J Allen and B A Barres, Stanford University School of Medicine, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The majority of mature synapses are not just composed of a pre- and postsynaptic neuronal element, but also an astrocytic process that envelops the synapse (Figure 1). This close spatial relationship has led to the term 'tripartite synapse' to acknowledge the glial contribution. Their synaptic localization means that glial cells are ideally placed to monitor and respond to synaptic activity. Indeed, one glial cell can contact and ensheathe thousands of synapses formed between many different neurons. This morphological specialization suggests important roles for glia in synaptic development and function, which are reviewed here.

Glial Induction of Synaptogenesis

Glial cells are in the right place at the right time to play an active role in neuronal synaptogenesis. For example, there is a temporal correlation between when retinal ganglion cell (RGC) axons reach their target structure, the superior colliculus, and when they form synapses. There is a delay of 1 week between target innervation and synaptogenesis, during which time glia are generated, which suggests that neurons require glial-derived signals to enable them to form synapses. The mechanism by which glial signals act is still unclear. Are they permissive in allowing synapses to form, with the location of synapses determined by neurons, or do they instruct neurons where to form synapses?

The first demonstration of a role for glial cells in neuronal synaptogenesis was by Pfrieger and Barres in 1997. They showed that when RGCs were cultured in the presence of astrocytes they possessed significantly more synaptic activity than when cultured alone in serum-free media. This was later shown to be due to astrocytes inducing the formation of new synapses in addition to enhancing the efficacy of existing synapses. This effect was not due to astrocytes enhancing the survival of RGCs, as the same effect was seen when astrocytes were added after most of the neuronal death had occurred and RGC survival factors were present in the media. Contact between astrocytes and neurons was not required, as the same effect was seen if the astrocytes were grown in contact with the neurons or placed in a feeding layer above them, thus demonstrating that a soluble signal released from astrocytes increased the number of synapses. Interestingly, astrocytes in culture did not require a neuronal signal to stimulate secretion of synaptogenic factors – they were constitutively released. Addition of media conditioned by astrocytes was equally effective in inducing synapses as a feeding layer of astrocytes.

There are a number of lines of evidence demonstrating that the synapses induced by astrocytes are functional (Figure 2). First, electrophysiological recordings from RGCs showed an increase in the frequency and amplitude of spontaneous excitatory postsynaptic currents (sEPSCs) following exposure to astrocyte signals. Second, immunostaining for pre- and postsynaptic markers showed a sevenfold increase in co-localized pre- and postsynaptic puncta, defined as a synapse, following astrocyte addition. Third, FM dye-uptake studies showed that presynaptic vesicle recycling was enhanced by astrocytes. Fourth, electron microscopy analysis of the synapses induced between RGCs by astrocyte exposure showed them to be ultrastructurally normal, exhibiting presynaptic vesicles and an electron-dense postsynaptic density. Taken together these observations show that the synapses induced to form between neurons by glial cells are fully functional.

Glia Induce Synapses in Multiple Neuron Classes

Since this initial study, numerous researchers have demonstrated a role for glial cells in inducing synaptogenesis in multiple classes of neurons. In addition to glutamatergic RGCs, spinal motor neurons, γ-aminobutyric acid (GABA)ergic hippocampal neurons, and glycinergic neurons all show enhanced synapse formation in the presence of glial cells. Also, synaptogenic effects are not restricted to astrocytes. Oligodendrocytes and Schwann cells, whose primary function is to myelinate and ensheathe axons, have also been shown to induce neuronal synaptogenesis.

Perisynaptic Schwann cells, specialized Schwann cells that are present at the neuromuscular junction synapse, are involved in guiding nerve terminals to the muscle during development. These specialized Schwann cells have been shown to enhance synaptogenesis between cultured motor neurons from rodents and between motor neurons and target muscle in *Xenopus*. Perisynaptic Schwann cell processes make contact with the muscle first and the nerve terminal uses the processes as a guide to locate and innervate the muscle. Selective ablation of perisynaptic Schwann

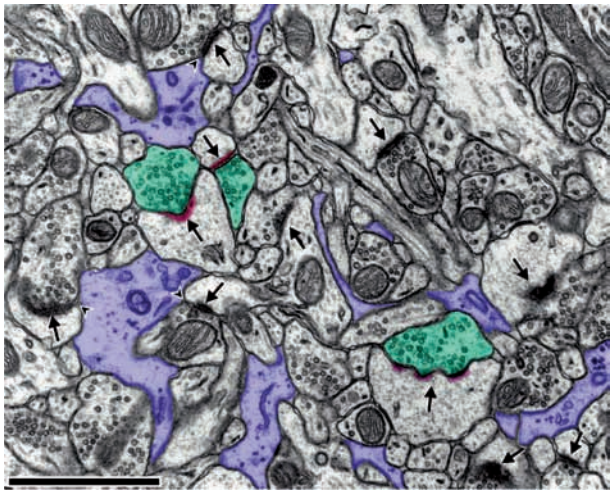

Figure 1 Electron micrograph illustrating the close relationship of astrocytic processes to synapses in the hippocampus. Astrocytic profiles are illustrated (blue) in the vicinity of 11 synapses (arrows). On this section, three synapses have astrocytic profiles at their perimeters (arrowheads). Scale bar = 1 μm. Reproduced from Ventura R and Harris KM (1999) Three dimensional relationships between hippocampal synapses and antrocytes. *Journal of Neuroscience* 19(16): 6897–6906, with permission.

cells in *Xenopus* tadpoles resulted in a decrease in the number of synapses that formed, the retraction of nerve terminals that had already formed synapses, and the loss of postsynaptic acetylcholine receptors.

GABAergic hippocampal neurons undergo enhanced synaptogenesis in the presence of astrocytes, and as for RGCs, the astrocytic effect is via a soluble factor. Astrocytes induce an increase in pre- and postsynaptic clusters, increased surface levels of $GABA_A$ receptors, and increased frequency of miniature inhibitory postsynaptic currents (mIPSCs). One of the downstream effects of astrocyte signals is to regulate brain-derived neurotropic factor (BDNF) and tyrosine receptor kinase B (TrkB) signaling between neurons to enhance synaptogenesis, specifically the maturation of the postsynaptic site and insertion of $GABA_A$ receptors, but the identity of the astrocytic signal is currently unknown.

Glia Induce Synaptogenesis via Secreted Factors

Much work has focused on identifying the soluble signals secreted by astrocytes that induce formation of functional synapses. To date two signals have been identified, thrombospondin (TSP) and cholesterol, but there are likely to be many more.

TSPs are a family of large extracellular matrix proteins that can mediate both cell–cell and cell–matrix interactions. Addition of TSP to cultured RGCs increased the number of structural synapses that formed, to the same extent as that induced by astrocytes. TSP-1 and -2 are expressed in the developing brain during the peak period of synaptogenesis but decrease by adulthood. Mice in which both TSP-1 and -2 had been knocked out formed 30% fewer synapses in their brains. These observations have led to the hypothesis that immature astrocytes provide a developmental window during which synaptogenesis can occur, by producing a permissive environment via the secretion of TSP.

Interestingly, the synapses induced by TSP are postsynaptically silent – they lack the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid subtype of glutamate receptor (AMPAR) but contain extrasynaptic *N*-methyl-D-aspartate receptors (NMDARs). This suggests that a second astrocyte-derived signal is required to trigger insertion of AMPARs into the postsynaptic membrane and make the synapse fully functional.

Another astrocyte-secreted factor suggested to be involved in synaptogenesis is cholesterol bound to apolipoprotein (ApoE). Cholesterol is required for normal neuronal synaptogenesis and can be provided by glia when neuronal cholesterol is lacking. Cholesterol enhances presynaptic function and transmitter release, and contributes to dendrite development, although these effects are strongest in clonal-density cultures where neuronal cholesterol may be lacking. Indeed, it has been demonstrated that if cholesterol is depleted from neurons in culture, then surface AMPARs are reduced due to disruption of lipid rafts.

Therefore, multiple astrocyte-derived signals promote both synapse formation and function. There are many signals yet to be identified, and it remains to be seen whether different signals induce the formation of different synapse types or between different classes of neurons.

Contact-Mediated Synaptogenesis

Glia can also enhance synaptogenesis by contact-mediated signaling. It has been shown that integrin-mediated astrocyte–neuron contact was required to initiate synaptogenesis between embryonic hippocampal neurons in culture. In these experiments neurons were encircled by noncontacting astrocytes to provide trophic support, but very few synapses were formed even in the presence of soluble astrocyte factors. However, when an astrocyte was added directly to a neuron and physically contacted it, multiple synapses were formed. In contrast, direct contact by astrocytes was not needed to induce postnatal RGCs to form synapses. This raises the question of whether astrocytic contact during embryonic development can render neurons receptive to secreted factors that can induce synaptogenesis.

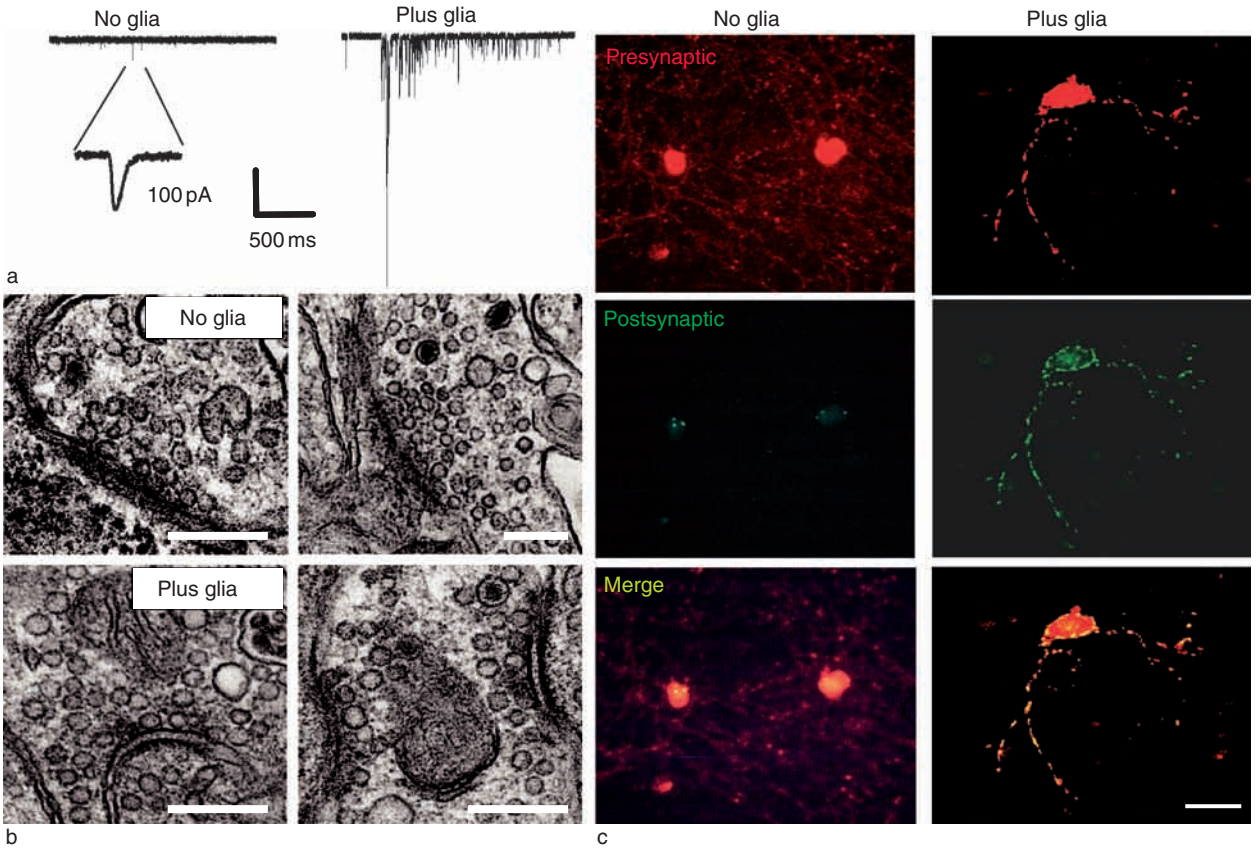

Figure 2 Glia induce functional synapses to form between retinal ganglion cells in culture. (a) Electrophysiological recordings from cultured retinal ganglion cells. In the absence of glia, only a low frequency of small spontaneous synaptic currents is observed, whereas in the presence of glia large spontaneous synaptic currents are frequently observed. (b) Electron micrographs of synapses between retinal ganglion cell neurons cultured in the absence of glia and presence of glia. There is no apparent difference in synaptic ultrastructure between the two conditions. Scale bar = 200 nm. (c) Glia increase the clustering of pre- and postsynaptic proteins and their co-localization. Staining of retinal ganglion cells with an antibody against the presynaptic marker synaptotagmin shows that in the absence of glia presynaptic markers are diffuse (left), whereas in the presence of glia presynaptic markers are discretely clustered (right). Staining of retinal ganglion cells with an antibody to the postsynaptic marker PSD-95 reveals that in the absence of glia there are relatively few PSD-95 puncta, whereas in the presence of glia numerous puncta are apparent. In the absence of glia there is little overlap between pre- and postsynaptic markers, whereas in the presence of glia there is a high degree of overlap. Scale bar = 50 μm. Reproduced from Ullian EM, Sapperstein SK, Christophernon KS, and Barres BA (2001) Control of synapse number by glia. *Science* 291(5504): 657–661, with permission.

Glia Influence Synapse Elimination

During development an excess of synapses are formed, but there is a reduction to the final adult number by a process of synapse elimination. It is hypothesized that elimination can be due to retraction of presynaptic terminals and axons or by degeneration of axons and subsequent phagocytosis of the debris by surrounding glial cells. It has now been suggested that glia are not just scavenging axonal debris, but are actively involved in the pruning process. In the *Drosophila* nervous system, glial cells were shown to engulf axonal varicosities prior to their degeneration. When the ability of glia to phagocytose was perturbed, axon pruning was significantly inhibited. A separate electron microscopy (EM) study showed peroxidase from labeled neurons inside neighboring glial cells in lysosomal compartments, suggesting phagocytosis. These findings are not direct evidence, but strongly suggest that glia actively contribute to pruning in *Drosophila*.

A similar process of axon removal by Schwann cells, termed axosome shedding, occurs during synapse elimination at the mammalian neuromuscular junction. This raises the possibility that glia actively contribute to synapse elimination in mammals. A combination of time-lapse imaging and serial EM showed that as axons disappeared they shed small membrane-bound particles termed axosomes, and that these axosomes appeared inside neighboring Schwann cells. It was suggested that this is a novel

means of transferring information between the cytoplasmic compartments of different cell types.

Glia Modulate Synaptic Strength

In addition to dictating the formation of synapses during development, glial cells are also actively involved in modulating the strength of mature synaptic connections. They do this by releasing factors that alter the number of neurotransmitter receptors present at synapses, and via factors that modulate the release of neurotransmitter and act as cofactors for receptor activation.

Glia, particularly microglia, release cytokines, including tumor necrosis factor-α (TNF-α), which has been shown to modulate synaptic strength via effects on the trafficking of AMPA and $GABA_A$ receptors. Addition of TNF-α to cultured hippocampal neurons increased the surface levels of AMPA receptors while simultaneously decreasing surface $GABA_A$ receptors, thus leading to an overall strengthening of the synapse, a phenomenon known as homeostatic synaptic scaling. Blockade of the actions of endogenous TNF-α led to a decrease in surface AMPA receptors. A glial source for TNF-α was shown by experiments in which TNF-α knockout (KO) glia were cultured with wild-type (WT) neurons and synaptic scaling due to activity blockade was no longer present. There were no deficits in long-term potentiation (LTP) in hippocampal slices from TNF-α KO mice, showing that glial TNF-α is involved in controlling overall synaptic strength but has no role in plasticity.

Glial Release of Transmitters and Neuromodulators

What are the effects of astrocytic transmitter release on synaptic transmission? Astrocytes have been shown to release ATP, which acts to inhibit overall neuronal excitability, and glutamate receptor agonists, which enhance overall activity levels.

Astrocytic ATP release has been implicated in heterosynaptic depression, a process whereby synapses neighboring those that have undergone LTP are depressed. ATP can exert its effects either as itself or by being converted to adenosine in the extracellular space by ectonucleotidases. In the hippocampus glial-derived adenosine has been shown to act on presynaptic A1 receptors, leading to inhibition of calcium channels, a decreased probability of vesicular release, and reduced frequency of sEPSCs. These effects were long-lasting and occurred over a large area – many synapses were inhibited. It is not clear if this inhibition was due to the diffusion of adenosine through the extracellular space or to the release of ATP from multiple regions on the astrocyte or multiple astrocytes within a network.

It has been suggested that glutamate can be released from astrocytes in an activity-dependent manner. Astrocytic glutamate has been shown to activate extrasynaptic NMDAR on neighboring neurons, which could contribute to excitation of the cell. As there is little evidence that glutamate can be accumulated in glia, it remains uncertain whether these effects are attributable to glial glutamate release or release of some other neuroactive substance.

In addition to the release of transmitters that directly activate neurotransmitter receptors, a role for astrocytes in releasing receptor co-agonists has been proposed. D-Serine, which has been suggested to be the endogenous ligand of the glycine binding site of the NMDAR, is expressed in glial cells but not neurons, and is necessary for NMDAR to be activated. There is circumstantial evidence that D-serine can be released from astrocytes in response to neuronal activity and that this release is necessary for hippocampal LTP. When astrocytes were cultured in contact with hippocampal neurons, the neurons were capable of undergoing LTP, whereas neurons grown in media conditioned by astrocytes (but without contact) were not. The addition of exogenous D-serine restored the ability of neurons grown in astrocyte-conditioned medium to undergo LTP. These data, and further studies carried out in hippocampal slices, suggest that astrocytes are actively involved in synaptic plasticity by regulating the availability of D-serine.

Glia Influence Synaptic Structure, Stability and Location

Studies on cultured neurons have shown that if synapses are induced by astrocytes and then the astrocytes are removed, these initial synaptic connections are lost. While the identity of this maintenance factor is unknown, this demonstrates that a constant astrocytic signal is needed in order to maintain synaptic connections.

Studies have shown transient physical coupling between astrocytic processes and synapses via ephrin/EphR interactions in the hippocampus. Dendritic spines express EphA4, which interacts with ephrin-A3 on astrocytic processes. Activation of EphA4 via ephrin-A3 caused a reduction in the length of dendritic spines by 30% and the collapse of 20% of spines, leading to an overall reduction in spine density. Conversely, inhibition of EphA4 caused an increase in spine length and a more disorganized pattern of spines appeared. EphA4 expression levels decrease during

development, and it is present in an inactive form in the adult brain. This raises the question of whether developmental contact between astrocytic processes and dendritic spines plays a role in the elimination and localization of synapses.

Live confocal imaging studies have produced new information about the motility of astrocytic processes at synapses. Imaging of green fluorescent protein (GFP)-labeled astrocytes in brain slices revealed that astrocytes frequently extend and retract fillipodial processes, on a timescale of minutes. Interestingly, this motility only occurred at synaptic sites, not at regions where astrocytes were contacting neuronal somata or blood vessels, and occurred in numerous brain regions. These observations add to the evidence that astrocytes are actively monitoring the synaptic environment.

In the hypothalamo-neurohypophysial system (HNS) astrocytes are capable of fully retracting from synapses, in a reversible manner, under the control of hormonal signals. When oxytocin is released it causes astrocytes to retract from neighboring neurons, which then receive more synaptic inputs. Astrocytes were previously blocking these sites and preventing neuronal innervation. In this way astrocytes can control the location and number of synaptic inputs that a neuron receives, thus influencing the overall activity and output of the neuron. This is a normal physiological process, and a similar phenomenon has been demonstrated in the cerebellum when astrocytic transmitter receptors were altered. Climbing fibers release ectopic vesicles directly onto Bergmann glia (BG) in the cerebellum. Released glutamate activates AMPA receptors on the BG; glutamate spilling over from neighboring synapses would not be at a high enough concentration to do this. BG express calcium-permeable AMPA receptors lacking the GluR2 subunit, which usually confers calcium impermeability to the channel. Expression of GluR2 in BG led to retraction of glial processes from the Purkinje cells (PCs) that they are normally in close association with and caused the PCs to be aberrantly innervated by multiple climbing fibers. Therefore, ectopic vesicular release of glutamate directly onto BG AMPA receptors signals the Bergmann glial cell in a calcium-dependent manner to remain in close apposition with the PC it is surrounding.

In the central nervous system it has been hard to interpret the effects of removal of astrocytes on synaptic function in mature animals. Experiments in which astrocytes have been ablated have had effects mainly in the cerebellum, with the loss of BG. This led to widespread death of granule neurons, presumably due to excitotoxicity from glutamate that accumulated, as the BG were no longer removing it from the extracellular space. The more global effects of astrocyte ablation on neuronal survival make it hard to infer anything about effects on synaptic function.

More subtle perturbations in astrocyte function do lead to alterations in synaptic physiology. Mice in which the astrocyte-specific intermediate filament protein glial fibrillary acidic protein (GFAP) has been deleted show altered synaptic plasticity, with enhanced levels of LTP. Deletion of GFAP will probably affect the structure of astrocytic processes and presumably their synaptic apposition, although whether this is responsible for the enhanced LTP remains to be determined.

An elegant study using the frog neuromuscular junction as a model system investigated the effects of removing synaptic glia from a mature synaptic contact. Perisynaptic Schwann cells in an intact adult frog were selectively labeled with a monoclonal antibody, exposed to complement, and lysed via the complement cascade. This approach left the presynaptic motor neuron terminal and the postsynaptic muscle cell intact, but removed their synaptic partner, the perisynaptic Schwann cell. Interestingly, this had no effect on the structure or function of the synapse until 1 week after the ablation. At this time presynaptic function decreased by half and there was a tenfold increase in the retraction of presynaptic terminals from the muscle. Therefore, glial cells play a role in stabilizing mature synaptic contacts, both in the central and peripheral nervous systems.

The studies discussed here demonstrate that glia are constantly monitoring the synaptic environment and can alter their structure and synaptic association in response to neuronal activity. This is yet another way in which they can control synaptic activity. Glial cells are ideally placed to respond to alterations in neuronal activity and to integrate information from many sources.

Conclusions

Glial cells are intimately associated with synapses at all stages of development and adult life, both in the central and peripheral nervous system. Glia induce the formation of synapses via the secretion of synaptogenic substances, and secrete additional signals that regulate both pre- and postsynaptic function. Glia contribute to the maintenance of synaptic structure and arrangement, ensuring that neurons receive the correct pattern of innervation.

See also: Glial Responses to Injury; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Glutamate Transporters: Electrophysiology; Neuromuscular Junction (NMJ): Presynaptic Schwann

Cells and Modulation of Neuromuscular Transmission; Neuron–Glia Coupling in Glutathione Metabolism; Retinal Glia; Schwann Cells and Plasticity of the Neuromuscular Junction.

Further Reading

Allen NJ and Barres BA (2005) Signaling between glia and neurons: Focus on synaptic plasticity. *Current Opinion in Neurobiology* 15: 542–548.

Auld DS and Robitaille R (2003) Glial cells and neurotransmission: An inclusive view of synaptic function. *Neuron* 40: 389–400.

Fellin T and Carmignoto G (2004) Neurone-to-astrocyte signalling in the brain represents a distinct multifunctional unit. *Journal of Physiology* 559: 3–15.

Feng Z, Koirala S, and Ko C (2005) Synapse–glia interactions at the vertebrate neuromuscular junction. *The Neuroscientist* 11: 503–513.

Glia – Special issue (2004) Glial control of synaptic function. *Glia* 47(3): 207–298.

Kettenmann H and Ranson BR (eds.) (2005) *Neuroglia*, 2nd edn. Oxford: Oxford University Press.

Luo L and O'Leary DD (2005) Axon retraction and degeneration in development and disease. *Annual Review of Neuroscience* 28: 127–156.

Pfrieger FW and Barres BA (1997) Synaptic efficacy enhanced by glial cells *in vitro*. *Science* 277: 1684–1687.

Ullian EM, Sapperstein SK, Christophernon KS, and Barres BA (2001) Control of synapse number by glia. *Science* 291(5504): 657–661.

Venture R and Harris KM (1999) Three-dimensional relationships between hippocampal synapses and artrocytes. *Journal of Neuroscience* 277: 6897–6900.

Glia Control of Blood Flow

G R J Gordon and B A MacVicar, University of British Columbia, Vancouver, BC, Canada
S J Mulligan, University of Saskatchewan, Saskatoon, SK, Canada

© 2009 Published by Elsevier Ltd.

Functional Hyperemia and the Astrocyte Link

More than 100 years have passed since Roy and Sherrington first appreciated that the brain possesses intrinsic factors that participate in the control of cerebral blood flow (CBF). Functional hyperemia, which is the active process of broadening vessel width in response to enhanced synaptic transmission and neuronal spiking, is a fundamental component of the entirety of CBF control, but despite its long history, there is much knowledge to be gained concerning the precise physiology responsible for the regulation of vessel diameter. Although the purpose of functional hyperemia is to augment the delivery of oxygen-rich hemoglobin and other nutrients to the working cells, while simultaneously clearing metabolic products such as CO_2, there persists an implicit assumption in the use of imaging to map brain function that changes in CBF are a simple function of neuronal activation. This has formed the basis for several imaging strategies so that specific volumes of activated brain tissue can be identified during both normal and pathological cerebral functioning. However, there is an incomplete understanding of the cellular mechanisms involved in transforming changes in neuronal activity to changes in CBF. Furthermore, there is a growing appreciation that the link between neural activity and vasomotor responses is more complicated than originally anticipated. Many factors have been identified that have direct consequences on cerebral vessel tone – including, but not limited to, prostaglandin E_2 (PGE_2), epoxyeicosatrienoic acid (EET), 20-hydroxyeicosatetraenoic acid (20-HETE), adenosine, nitric oxide (NO), and K^+. The number of cell types participating in neurovascular coupling is growing as new experimentation is performed and the cellular source of these vasoactive signaling molecules in some cases is still a matter of debate.

However, contributing to our understanding of this important physiological process, astrocytes have been examined as the missing element coupling changes in neuronal activity to alterations in the diameter of brain vessels. This article focuses on the role of astrocytes, particularly that of astrocyte endfeet Ca^{2+} signaling, in the control of brain blood vessel girth. The different experimental models and approaches used to test astrocyte involvement, the possible signaling molecules participating, and the potential mechanisms of astrocyte action are highlighted.

Astrocytes Are Ideal Regulators of Cerebral Blood Flow

Processes from individual astrocytes can form a physical bridge between numerous synapses, adjacent astrocytes – forming a glial syncytium – and the microvessels of the brain. The latter is accomplished by modified astrocytic processes termed perivascular 'endfeet.' Endfeet are enlarged astrocytic compartments that are specialized for direct interaction with vessels. It is a defining feature of the central nervous system that all cerebral arterioles in the brain are circumscribed by endfeet, suggesting a prominent role in vasculature physiology. Astrocytes also possess an equally impressive ability for cell-to-cell communication. Astrocytes can sense transmitter molecules such as glutamate, γ-aminobutyric acid (GABA), adenosine triphosphate (ATP), and norepinephrine and may respond to such signals by releasing gliotransmitters. These include ATP, adenosine via ATP breakdown, NO, PGE_2, prostacyclins, and EETs, all of which are vasoactive compounds. Furthermore, astrocytes can propagate a signal over long distances via interastrocytic Ca^{2+} waves, which rely on the release of glial-derived ATP. It is these properties that have led to the concept of the 'neurovascular unit,' in which astrocytes are well enough endowed, both anatomically and physiologically, to transmit relevant information about the extracellular environment from neurons to vessels.

The endfeet are also endowed with many features thought to be important for cerebrovasculature control. Endfeet express soluble nitric oxide synthase (NOS), noradrenergic receptors, and gap junction proteins, whereas Ca^{2+}-activated K^+ channels, aquaporin 4 water channels, and purinergic receptors are present in greater amounts here than in other astrocytic compartments. In particular, phospholipase A_2 (PLA_2) activation can lead to the downstream production of vasoactive arachidonic acid (AA), PGE_2, EET, and 20-HETE. These characteristics suggest that the endfeet are capable of producing diffusible molecules to communicate with arteriole smooth muscle cells (SMCs) and thus contribute to the physiology of blood vessel control.

Astrocytes in the Control of Cerebrovascular Diameter

Astrocytes respond to glutamate with an increase in intracellular Ca^{2+} through the activation of the mGluR subtype, and the first evidence that mGluR

activation of astrocytes is a component of neurovascular coupling was demonstrated by the Carmignoto laboratory. Activating astrocytes indirectly by disrupting membranes with patch electrodes or by applying mGluR agonists triggered the release of diffusible factors that then acted on SMCs to cause dilation of arterioles. Furthermore, an elevation of extracellular glutamate from enhanced synaptic activation evoked Ca^{2+} increases in astrocyte endfeet (via mGluR) and caused vasodilation. The dilation was reduced by blocking astrocyte Ca^{2+} signals with the cyclooxygenase (COX) inhibitor aspirin, suggesting that products from cyclooxygenase activity such as PGE_2 and/or prostacyclin were involved. The complexity of this model for functional hyperemia, in which astrocytes simply sense the level of extracellular glutamate, was soon expanded. Work by the Nelson laboratory demonstrated that in the presence of a thromboxane analog to increase vascular tone, electrical stimulation of the tissue increased endfoot Ca^{2+} and reduced spontaneous Ca^{2+} signals in SMCs, but neither of these effects were completely blocked by mGluR antagonists. These data indicate that other, nonglutamatergic inputs or transmitters are likely involved in relaying information about the state of the cellular environment to vessels through astrocytes.

The MacVicar laboratory has examined the effects of elevations in astrocyte Ca^{2+} alone, without incorporating the involvement of membrane-bound receptors, by uncaging Ca^{2+} using two-photon photolysis. This technique allowed them to increase Ca^{2+} within the discrete volume of astrocytes without provoking other cell types. Astrocytes were identified in transgenic mice that expressed enhanced green fluorescent protein driven by the glial fibrillary acidic protein promoter and these cells were loaded with the acetoxymethyl (AM) ester form of Rhod-2, the Ca^{2+}-sensitive dye. Under typical brain slice recording conditions in the hippocampus and cortex, Ca^{2+} uncaging in astrocytes induced a Ca^{2+} wave that propagated throughout the astrocyte syncytium and invaded endfeet, and the endfeet Ca^{2+} signal was immediately followed by a constriction of adjacent arterioles. Mulligan and MacVicar found a strong positive relationship between the extent of the constriction and the extent of the Ca^{2+} signal within and among the endfeet. Astrocyte-mediated vasoconstrictions were blocked by inhibiting the Ca^{2+}-dependent enzyme PLA_2 and preventing the release of the diffusible factor AA. Arteriole constrictions were found to be caused by the generation of 20-HETE within SMCs from the astrocyte-derived AA by a CYP4504a enzyme.

A study from the Newman laboratory conducted in retinal explants examined the effect of Ca^{2+} elevations within retinal glia by using UV photolysis of caged Ca^{2+} and caged IP_3. Notably, in response to Ca^{2+} uncaging, constrictions as well as dilations were observed, whereas increases in free IP_3 produced mostly dilation. Consistent with results from the MacVicar laboratory, constrictions were dependent on the generation of astrocyte AA and its conversion to 20-HETE. Dilations were also caused by astrocyte AA, but instead of its conversion to the potent constrictor molecule, AA was converted to EET by another CYP450 enzyme. These data indicate an intriguing complexity in the dialog between glial cells and arterioles with respect to the potential factors determining the polarity of the response and the different outcomes observed depending on the method of Ca^{2+} liberation within glia.

An *in vivo* study conducted by the Nedergaard laboratory described the effects of uncaging Ca^{2+} in astrocyte endfeet on vasomotor responses in arteries of the somatosensory cortex. Astrocytes were loaded with the Ca^{2+} indicator dye Rhod2-AM and caged Ca^{2+} DM-nitrophen, similar to previous strategies. Uncaging Ca^{2+} within astrocyte endfeet triggered an intracellular Ca^{2+} rise that was followed by the dilation of the adjacent penetrating arteriole. Using this protocol very few constrictions were observed. Dilations were found to be blocked by inhibitors of COX-1 but not COX-2 enzymes, suggesting that the AA conversion to PGE_2 underlies the dilation, similar to the findings of the Carmignoto laboratory (see **Figure 1** for the possible astrocyte endfoot-mediated vasoconstriction and vasodilation mechanisms).

Nitric Oxide: An Important Modulator of Astrocyte-Mediated Cerebral Vessel Control

As stated previously, the first study implicating astrocyte Ca^{2+} signals in the relaxation of vascular tone as a model for functional hyperemia was conducted by the Carmignoto laboratory. In a subset of their experiments, blood vessels were preconstricted by incubating the brain slices with L-NAME to block NOS and reduce the level of endogenous NO to ultimately enhance the vasodilations observed. This is interesting compared to the results obtained by the MacVicar laboratory, in which arterioles in untreated slices always displayed constriction when astrocytes were stimulated. Only when Mulligan and MacVicar incubated in L-NAME did they observe vasodilations in response to mGluR agonist t-ACPD.

On the basis of these dichotomous results, Metea and Newman speculated that the levels of NO dictated the type of vasomotor response. This was hypothesized to occur via regulation of the enzymatic conversion of AA to either EET or 20-HETE. Consistent with this idea, vessels that upon first test showed dilations were

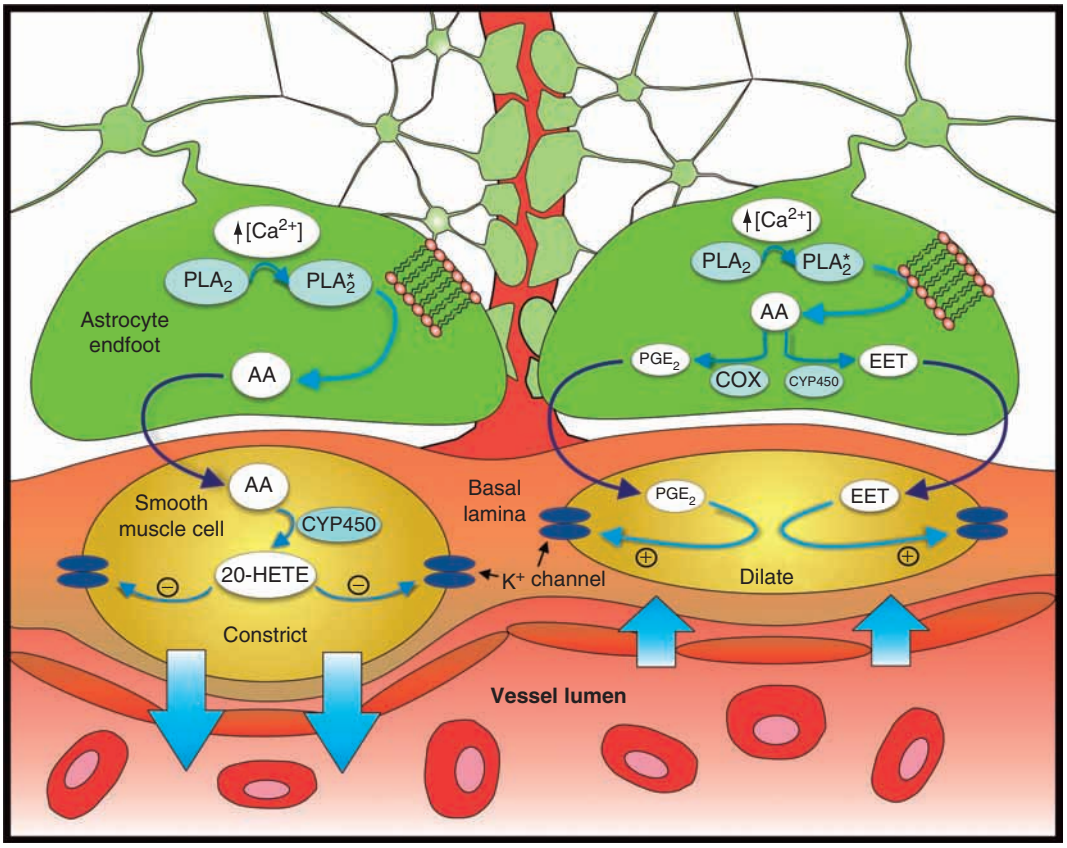

Figure 1 Astrocyte endfeet control of cerebrovasculature diameter by the generation of astrocyte AA via calcium-sensitive PLA_2. Two separate endfeet making contact with an arteriole show possible constriction (left endfoot) and dilation (right endfoot) mechanisms. Reproduced from Glia, Vol. 55, No. 12, 2007, pp. 1214–1221. Copyright 2008 Wiley-Liss, Inc. Reprinted with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

transformed into constrictions in the presence of the NO donor SNAP, an outcome thought to be due to the NO sensitivity of the EET producing enzyme CYP450. In the presence of the NO scavenger PTIO, the opposite was true: Vasoconstrictions were converted to vasodilations. In line with the latter experiment in which NO is limiting, blocking all forms of NOS with L-NAME produced only vasodilations when the preparation was stimulated, similar to the results of the Carmignoto and MacVicar laboratories.

An *in vivo* study conducted by the Nedergaard laboratory also examined a role for NO in astrocyte-mediated dilations by applying L-NAME to block NOS. This treatment failed to affect the degree of dilation induced when Ca^{2+} was uncaged in astrocyte endfeet. This negative result may be explained by a lower level of endogenous NO in the intact animal. However, NO donors were not tested, and it would be interesting to determine if by elevating NO levels *in vivo* the vasomotor response changes polarity from dilation to constriction when astrocytes are stimulated – as has been demonstrated in *in vitro* experiments.

The work outlined previously shows that astrocytes wield the necessary physiology to initiate both vasoconstriction and vasodilation mechanisms and that the level of NO or the degree of tonic constriction may dictate the polarity of the vasomotor response. However, there are many unknowns in the role of NO in mediating the previous processes. First, little is known about the importance of astrocyte-derived NO compared to that of neurons and endothelial cells in functional hyperemia. Second, studies testing the impact of NO on synaptic transmission have demonstrated that NO enhances the release of several neurotransmitters, including glutamate. This notion also corresponds with data collected *in vivo* showing a prominent role for enhanced NO release in function hyperemia. These demonstrations need to be reconciled both with the *in vitro* data showing that higher NO levels promote astrocyte-mediated vasoconstriction and with the *in vivo* data showing that astrocyte-mediated vasodilations do not rely on NO production.

K^+ and Vascular Control by Astrocytes

The first hypothesis to point to the glial cell as an essential controller of vascular tone was proposed by

Newman approximately 20 years ago. It was based on his observation that the endfeet of Muller cells express high levels of Kir channels, which could potentially contribute to vessel dilation by effluxing K^+ around the closely apposed vessel SMCs. However, the Newman laboratory disproved this hypothesis with two separate experiments. First, the direct depolarization of single glia that were in close proximity to a vessel did not elicit changes in vessel diameter. Second, the extent of K^+-induced vasodilations was seemingly unabated in a Kir knockout mouse compared to that of wild type.

An alternative astrocyte-mediated mechanism that takes advantage of the ubiquitous K^+ ion in the regulation of vascular tone has been proposed by the Nelson laboratory. Their work shows that a Ca^{2+} rise in astrocyte endfeet activates large-conductance Ca^{2+}-activated K^+ channels (BK channels), which upon opening allow the efflux K^+ into the extracellular space. They demonstrated that modest increases in external K^+ promote vasodilation by hyperpolarizing SMCs through an enhanced Kir conductance. This leads to decreased Ca^{2+} entry, SMC relaxation, and the consequent dilation (Figure 2). However, from the data presented, K^+ channels are not the only contributors to K^+- or activity-induced vasodilations because a fraction of the dilation response was blocked by inhibitors of cyclooxygenases, suggesting that vasoactive PGEs were also involved. Furthermore, it should be noted that the role on SMC Kir channels here likely does not conflict with Newman et al.'s data showing that there is no difference in K^+-induced dilations between wild-type and Kir knockout mice because of the difference in the Kir subtypes involved: Kir 2.0 and Kir 4.1 for SMCs and astrocyte endfeet, respectively.

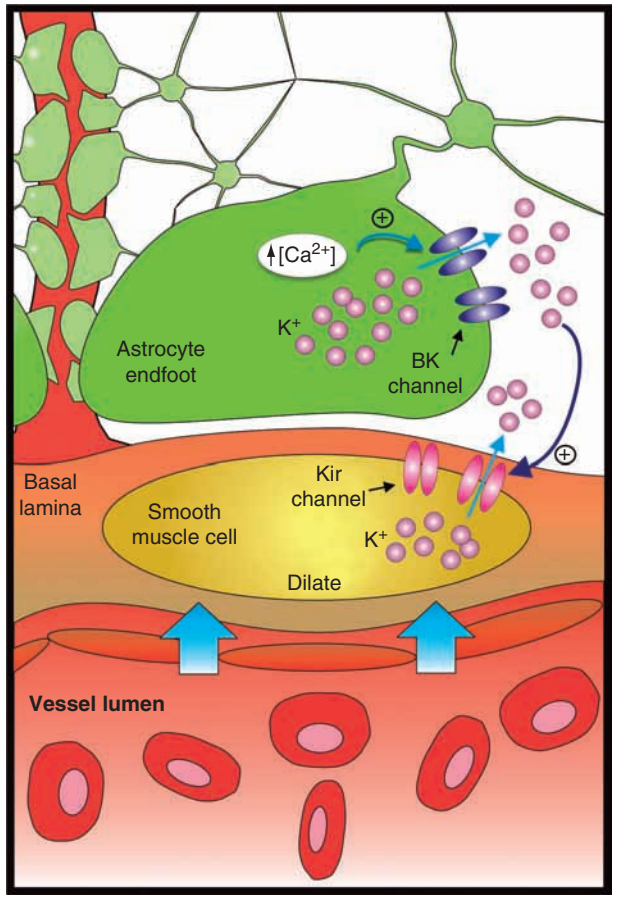

Figure 2 Astrocyte endfeet control of cerebrovasculature dilation by the efflux of potassium from BK channels. Elevated extracellular potassium activates Kir channels on SMCs, which causes dilation via hyperpolarization and subsequent SMC relaxation. Reproduced from Glia, Vol. 55, No. 12, 2007, pp. 1214–1221. Copyright 2008 Wiley-Liss, Inc. Reprinted with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

Norepinephrine and Astrocyte-Mediated Cerebrovascular Control

Vasomotor motor responses cannot be completely abolished by mGluR antagonists when synapses become activated, suggesting that transmitters other than glutamate can participate in neurovascular coupling. Along this line, an electron microscope examination of arterioles has revealed that the majority of noradrenergic terminals originate from the locus coeruleus that associate with vessels synapse on astrocyte endfeet rather than SMCs. Previous *in vivo* work demonstrated that activation by norepinephrine (NE) causes a decrease in CBF, an effect that may help maintain CBF at a constant rate at higher blood pressures. *In vitro* work has shown that NE triggers robust intracellular Ca^{2+} increases in astrocytes via activation of α_1 and β adrenergic receptors. Consistent with these results, experiments conducted by Mulligan and MacVicar showed that NE-mediated Ca^{2+} increases within astrocyte endfeet temporally preceded prominent vasoconstrictions. When astrocytes were loaded with BAPTA-AM to chelate rises in intracellular Ca^{2+}, vascular constrictions generated by NE were drastically reduced, suggesting that Ca^{2+} was critical for the astrocyte-mediated effect. These data add another level of complexity to neurovascular coupling by expanding the realm of vasoactive transmitter beyond that of glutamate.

Enigmatic Astrocyte Ca^{2+} Signals: Functional Significance?

From the studies described previously, it is clear that astrocytic Ca^{2+} transients, particularly those within the endfeet, have important albeit complex actions on vascular tone. Of paramount importance to all this work, however, is whether a rise in astrocyte

intracellular Ca^{2+} is indeed the signal responsible for relaying information on the metabolic state of neurons to the vasculature in order to affect vessel diameter and ultimately CBF. There are several observations concerning the very nature of astrocyte Ca^{2+} signaling that make it a worthwhile endeavor to entertain this query. Because CBF changes can occur in a graded manner to match the corresponding subtlety of ongoing synaptic communication and neuronal activation, it is expected that astrocyte Ca^{2+} will also mirror these changes if this signal is indeed the key factor in mediating functional hyperemia and other forms of neurovascular coupling. However, there are several lines of evidence that reveal a curious inconsistency with this idea. First, the Helmchen laboratory has described a novel two-photon scanning method that is capable of measuring Ca^{2+} signals at frequencies up to 10 Hz in hundreds of cells simultaneously. This technique has demonstrated that neuronal activity fails to consistently increase astrocyte Ca^{2+} with a temporal profile that follows the timing and patterning of neuronal activation. Second, Ca^{2+} increases often occur independent of neuronal activity. These spontaneous Ca^{2+} oscillations have been reported both *in vivo* and *in vitro*. Finally, the astrocyte Ca^{2+} waves thought to transduce important information toward vessels are not faithfully observed *in vivo*, suggesting that this signaling process may actually be attributable to aspects of the slice or culture condition, or that it manifests more easily in the pathophysiology of epilepsy rather than being the 'normal' method of signaling for neurovascular coupling. These discrepancies, however, may not justify the hypothetical scenario, we have just considered, that astrocyte Ca^{2+} signaling (endfeet excluded) does not constitute a mechanism utilized by the neurovascular unit. Instead, they may simply reflect our ignorance of why such timing differences arise and how the vastly different astrocyte signals preserve neural information, which we cannot currently decode. In either case, meaning versus no meaning, it is clear that more experimentation is needed to tease out the true nature of astrocyte Ca^{2+} signals in cerebrovascular control.

Because elevations in endfeet-free Ca^{2+} can trigger vasoconstriction or vasodilation, there must be different physiological conditions that permit the utilization of each astrocyte-mediated mechanism. Factors other than NO that dictate the polarity of the vasomotor response may include the vessel distance from the center of augmented neuronal activity. Work from the Dale laboratory has shown that cerebral vessels located within a brain volume undergoing an elevation in activity display, as expected, dilation, but this space is circumscribed by a surrounding brain volume in which the vessels contained therein display constriction. The result of this effect is a surround inhibition of vessel diameter, which is proposed to facilitate the enhanced CBF to the central active region that requires additional blood supply. This may constitute the physiological circumstance in which both astrocyte-mediated constrictions and dilations of cerebral vessels are utilized simultaneously directly adjacent to one another. How vasoconstriction is recruited over vasodilation as the distance from the center of the elevated activity increases is not known.

Conclusion

The work described here indicates that astrocytes are capable of eliciting changes in vessel diameter in both directions. Although there appears to be an important initiating role for glutamate through the activation of group I mGluRs that raises intraglial Ca^{2+}, other inputs and transmitters are likely involved in neurovascular coupling. An increase in intracellular free Ca^{2+} and the subsequent activation of Ca^{2+}-sensitive PLA_2 in astrocytes can trigger a surprisingly diverse array of vasoactive metabolites after the initial production of AA. Constriction occurs when AA is converted to 20-HETE, whereas dilation results from the conversion of AA to PGE_2 or EET. The enzymes governing the production of these vasoactive products are sensitive to NO, suggesting that NO levels may dictate the direction of the vessels' response. In addition, a role for Ca^{2+}-activated K^+ channels in astrocyte endfeet and the efflux of K^+ has also been suggested to relax vascular tone by hyperpolarizing SMCs via Kir channels. Astrocytes are proving to be important mediators of neurovascular coupling, but a comprehensive understanding of their role is far from complete. Novel experiments and techniques are unfurling much complexity in processes such as functional hyperemia, which historically was thought to occur by a simple correspondence to neuronal activation.

See also: Astrocyte: Calcium Signaling; Astrocyte: Response to Injury; Calcium Homeostasis in Glia; Cerebrovascular Disease; Gap Junctions and Hemichannels in Glia; Glial Growth Factors; Potassium Homeostasis in Glia.

Further Reading

Devor A, Ulbert I, Dunn AK, et al. (2005) Coupling of the cortical hemodynamic response to cortical and thalamic neuronal activity. *Proceedings of the National Academy of Sciences of the United States of America* 102(10): 3822–3827.

Filosa JA, Bonev AD, and Nelson MT (2004) Calcium dynamics in cortical astrocytes and arterioles during neurovascular coupling. *Circulation Research* 95(10): e73–e81.

Filosa JA, Bonev AD, Straub SV, et al. (2006) Local potassium signaling couples neuronal activity to vasodilation in the brain. *Nature Neuroscience* 9(11): 1397–1403.

Gobel W, Kampa BM, and Helmchen F (2007) Imaging cellular network dynamics in three dimensions using fast 3D laser scanning. *Nature Methods* 4(1): 73–79.

Haydon PG and Carmignoto G (2006) Astrocyte control of synaptic transmission and neurovascular coupling. *Physiological Review* 86(3): 1009–1031.

Koehler RC, Gebremedhin D, and Harder DR (2006) Role of astrocytes in cerebrovascular regulation. *Journal of Applied Physiology* 100(1): 307–317.

Metea MR, Kofuji P, and Newman EA (2007) Neurovascular coupling is not mediated by potassium siphoning from glial cells. *Journal of Neuroscience* 27(10): 2468–2471.

Metea MR and Newman EA (2006) Glial cells dilate and constrict blood vessels: A mechanism of neurovascular coupling. *Journal of Neuroscience* 26(11): 2862–2870.

Mulligan SJ and MacVicar BA (2004) Calcium transients in astrocyte endfeet cause cerebrovascular constrictions. *Nature* 431(7005): 195–199.

Roy CS and Sherrington CS (1890) On the regulation of the blood supply of the brain. *Journal of Physiology* 11: 85–108.

Simard M, Arcuino G, Takano T, Liu QS, and Nedergaard M (2003) Signaling at the gliovascular interface. *Journal of Neuroscience* 23 (27): 9254–9262.

Takano T, Tian G-F, Peng W, et al. (2006) Astrocyte-mediated control of cerebral blood flow. *Nature Neuroscience* 9(2): 260–267.

Zonta M, Angulo MC, Gobbo S, et al. (2003) Neuron-to-astrocyte signaling is central to the dynamic control of brain microcirculation. *Nature Neuroscience* 6(1): 43–50.

Glial Cells: Astrocytes and Oligodendrocytes during Normal Brain Aging

I Blasko and C Humpel, Innsbruck Medical University, Innsbruck, Austria
B Grubeck-Loebenstein, Institute for Biomedical Aging Research of the Austrian Academy of Sciences, Innsbruck, Austria

© 2009 Elsevier Ltd. All rights reserved.

Overview

Astrocytes are the most abundant cells in the brain, and together with oligodendrocytes, they are multifunctional housekeeping cells that allow neurons to become progressively specialized for the tasks of information processing. Astrocytes not only organize the structural architecture of the brain, but also its communication pathways and plasticity. Neurons co-cultured with astrocytes develop approximately sevenfold more synapses and display a sevenfold increase in synaptic efficacy, compared with neurons cultured in the absence of astrocytes. Astrocytes modulate synaptic strength and activity and regulate neurogenesis of the adult human brain. Oligodendrocytes are responsible for the production and maintenance of myelin, which is essential for normal brain function and signal transmission speed. A series of recent studies clearly demonstrate reciprocal paracrine interactions between astrocytes, endothelial cells, and ependymal cells and also between oligodendrocytes and neurons. In fact, astrocyte-derived signals induce neuronal differentiation of oligodendrocyte precursor cells. Coincident with these findings, it has become obvious that both cell types are substantially influenced during the aging process.

Role of Astrocytes in the Aging Process

Astrocytes in the Aging Brain

In the past decade considerable attention has been focused on understanding the role of astrocytes in aging and age-related neurodegenerative disorders. During normal aging, astrocytes display increased activation ('reactive gliosis'), with progressive increases of glial fibrillary acidic protein (GFAP) beginning before midlife (Figure 1(a)). Similar to the concept of age-related microglial dystrophy, analogous astrocytic dystrophy might be relevant in normal aging of astrocytes. Neurite outgrowth by embryonic neurons is markedly decreased when co-cultured with confluent astrocytes derived from old (24 months) versus young (3 months) animals. Overexpression of GFAP in young astrocytes models the effects of aging by reducing neurite outgrowth. These results implicate increased astrocytic GFAP expression that leads to functional defects, which may trigger neuronal dysfunction and support the development of neuronal atrophy during normal aging.

The ability of astrocytes to protect neurons or stimulate neurogenesis is compromised during aging. It is well known that activated astrocytes can produce and release cytokines, but they also produce neurotrophins such as nerve growth factor (NGF) and can enhance the survival of neurons in the brain (Figure 1(c)). There is evidence that multiple hippocampal stem cell/progenitor proliferation factors exhibit an early decline during aging, an event that seems to be linked to age-related alterations in hippocampal astrocytes. Neurogenesis in the dentate gyrus plays an important role for learning and memory and declines dramatically by middle age. Studies in rat astrocytes show that several growth factors – fibroblast growth factor-2 (FGF-2), insulin-like growth factor-1, and vascular endothelial growth factor – decline in concentration considerably by middle age but remain at steady levels between middle age and old age. Decreased concentrations of FGF-2 during aging are associated with decreased numbers of FGF-2-positive astrocytes. Quantification of GFAP and FGF-2-positive cells reveals that aging does not decrease the total number of astrocytes, but that some astrocytes express lower FGF-2 levels by middle age. Thus, dramatically decreased neurogenesis by middle age is likely linked to reduced concentrations of growth factors in the hippocampus.

Astrocytes regulate synthesis and extracellular levels of the excitatory amino acid glutamate via the glutamine cycle and strongly contribute to brain homeostasis and regeneration. Neurotrophic factors, transporter molecules, and enzymes involved in the metabolism of excitatory amino acids play a role in neuroprotection of neurons by regulating the production of potentially toxic substances (e.g., radicals). It has been found that mitochondrial membrane potentials in astrocytes isolated from old mice are more depolarized than are those from their younger counterparts and display a greater sensitivity to the oxidant hydrogen peroxide. ATP-induced Ca^{2+} responses in astrocytes isolated from old mice are consistently larger in amplitude and more frequently oscillatory than is seen in younger astrocytes, which may be attributable to lower mitochondrial Ca^{2+} sequestration. Finally, NGF-differentiated PC12

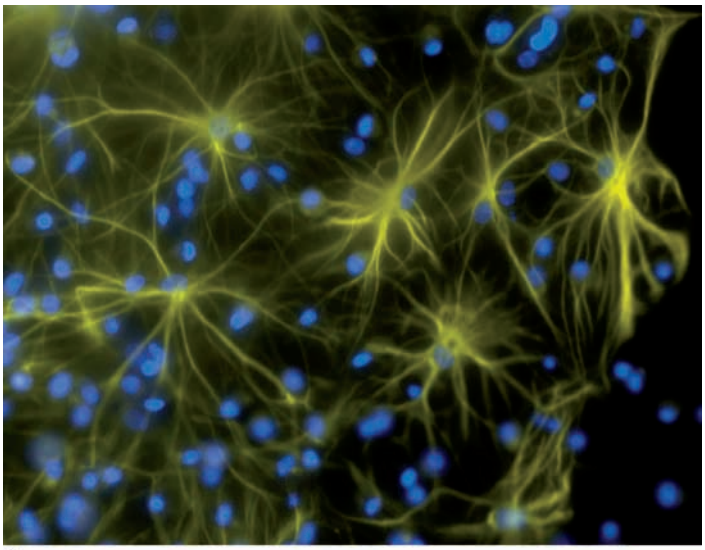

a

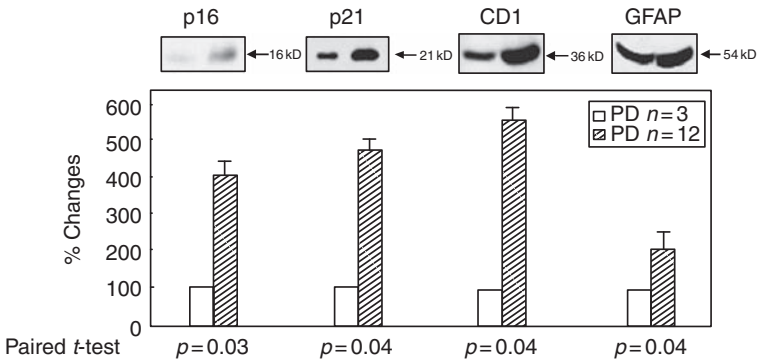

b

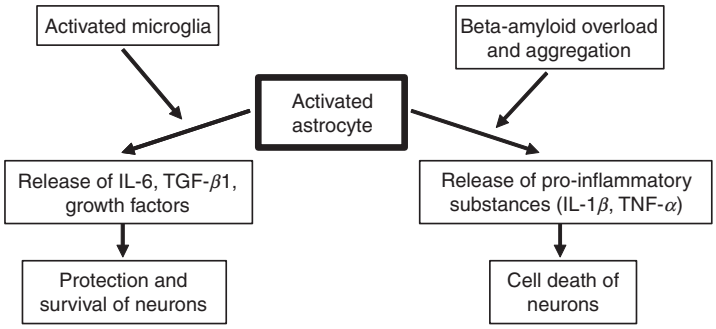

c

Figure 1 (a) Fluorescence immunohistochemistry of astrocytes in the brain expressing high levels of glial fibrillary acidic protein (GFAP). The cells were counterstained with the nuclear dye 4′,6-diamidino-2-phenylindole (blue). (b) Expression pattern of the senescence markers p16, p21, cyclin D1 (CD1), and the differentiation marker GFAP. Each blot shown in the upper part represents one characteristic Western blot experiment. The bars in the lower part represent mean ± SEM ($n = 3$ in each group). The cells from early passage (population doubling (PD) number 3) and late passage (PD 12) are compared. Marker expression levels in PD 3 were considered as 100%. PD 12 results are presented as percent change compared to early passage cells. The levels of significance between both groups are indicated (Student's *t*-test). (c) Schematic role of activated astrocytes during aging and neurodegeneration. Activated microglia stimulate astrocytes to produce growth factors and cytokines, which support survival of neurons (IL-6, interleukin-6; TGF-β1, transforming growth factor-β1; TNF-α, tumor necrosis factor-α). β-Amyloid overload and aggregation induce astrocytes to release proinflammatory cytokines, which induce neuronal cell death. (b) Reproduced from Blasko I, Stampfer-Kountchev M, Robatscher P, et al. (2004) How chronic inflammation can affect the brain and support the development of Alzheimer's disease in old age: The role of microglia and astrocytes. *Aging Cell* 3(4): 169–176, with permission from Blackwell Publishing.

cells co-cultured with old astrocytes are significantly more sensitive to hydrogen peroxide treatment than are co-cultures of NGF-differentiated PC12 cells with young astrocytes.

Senescence of Astrocytes

Cellular senescence is a stress response phenomenon resulting in a permanent withdrawal from the cell cycle and the appearance of distinct morphological and functional changes. As astrocytes can be easily passaged in cell culture, replicative senescence of primary human astrocytes has been well studied. In analogy to the established '*in vitro* senescence' models of fibroblasts and endothelial cells, astrocytes isolated from postmortem brains of Alzheimer' disease patients proliferate rigorously in culture and undergo six to eight population doublings (PDs) before astrocytes stop to divide. Growth arrest can be reached after 9–10 PDs, and astrocytes are therefore referred to as 'late-passage astrocytes' after 8 PDs. The quantification of cellular GFAP content reveals that late-passage astrocytes express more GFAP in comparison to early-passage astrocytes (Figure 1(b)). In agreement with this observation, a higher GFAP content is found *in vivo* in aged rats and mice in comparison to their young littermates. Additionally, late-passage astrocytes exhibit a significantly increased expression of the senescent markers p16, p21, and cyclin D1 (Figure 1(b)). The number of PDs of gray-matter astrocytes obtained from frontal cortex is higher, compared to the number from white-matter astrocytes. Whether this increased growth rate depends on *in vivo* degenerative or inflammatory processes is not yet precisely known. In contrast to fibroblasts, but in accordance with endothelial cells, astrocytes are prone to undergo apoptosis after having reached the end of their replicative life span. Interestingly, in late-passage astrocytes, tumor necrosis factor-α and interferon-γ stimulate the production of β-amyloid, demonstrating that stimulation of β-amyloid production by inflammatory products does not decrease during aging.

Astrocytes and Age-Related Neurodegenerative Diseases

Astrocytes greatly outnumber microglia in the brain and play a more critical role in the development of age-related diseases than previously thought. Many metabolic products, depending on their aggregation and deposition status, can activate astrocytes. It is well known that astrocytes are involved in the degradation of damaging products. A disturbed balance between the production and the degradation of toxic metabolites triggers chronic inflammatory processes during aging. Chronically activated microglia and astrocytes can damage neurons via release of highly toxic products such as reactive oxygen intermediates, nitric oxide, inflammatory cytokines, proteolytic enzymes, complement factors, or excitatory amino acids (Figure (1c)). In addition, astrocytes appear to be first-line phagocytic cells of degenerating myelin in cerebral cortex, and microglia cells are not activated until degeneration of entire nerve fibers proceeds (see later).

Inflammation in the aging brain is most likely the result of accumulation of toxic metabolites or their decreased degradation. Glial products such as interleukin-1β or tumor necrosis factor-α induce the synthesis of β-amyloid binding proteins in astrocytes, microglia, or neurons. The chronic release of proinflammatory cytokines in the aging brain is likely to maintain a chronic acute-phase protein secretion favoring the formation of β-amyloid fibrils. The astrocytic synthesis of amyloid precursor protein (APP) can be stimulated by these proinflammatory cytokines through regulation of gene transcription at the promoter level. In combination, cytokines such as interleukin-1β or tumor necrosis factor-α and interferon-γ affect the metabolism of APP and stimulate the production of β-amyloid (1–40) and β-amyloid (1–42) peptides. This molecular mechanism is complex, but it seems likely that the maturation (i.e., proper glycosylation) of APP is disturbed. The chemokine monocyte chemoattractant protein-1 (MCP-1) is believed to play an important role in astrocytosis, as MCP-1 levels are increased in aging brain. It is also of interest that adult astrocytes do not respond to stimulation with β-amyloid by increasing their release of MCP-1 in the same way as young control cells do.

Astrocytosis is a typical morphological feature of the Alzheimer's disease brain and represents either proliferation of astrocytes, in an effort to replace dying neurons, or a reaction to degrade the increasing amounts of toxic β-amyloid peptides. There are clear indications that age-related changes of astrocytes may influence the development of age-related neurodegenerative disorders such as Alzheimer's disease, Huntington's disease, or Parkinson's disease. On the other hand, age-related degenerative changes can modulate astrocytes and their supportive function for neurons. Little is known about the molecular mechanisms that underlie these age-related changes in astrocytes. Alzheimer's disease may coincide with the decreased capacity of aged glial cells to degrade and/or store toxic products. It is now well understood that an early and invariant step in the pathogenesis of Alzheimer's disease is the overproduction and/or decreased degradation of β-amyloid, which leads

to the oligomerization of β-amyloid, both of which are regarded as pathogenetic hallmarks of Alzheimer's disease. In contrast to neonatal astrocytes, adult astrocytes efficiently clear exogenously added and surface-bound β-amyloid and are capable of removing β-amyloid deposited in transgenic APP overexpressing mice. In addition, in young animals, astrocytes produce low amounts of the β-amyloid-degrading enzyme neprilysin, but older animals have increased levels of both β-amyloid peptides and neprilysin.

Role of Oligodendrocytes in the Aging Process

Oligodendrocytes are highly specialized cells that support between 1 and 40 separate myelin sheaths, depending on the size of the axons with which they associate. The production and maintenance of myelin are essential for normal brain function, because myelination results in more than a tenfold increase of signal transmission speed. Faster conduction velocity facilitates information flow by allowing for precise temporal coding of high-frequency bursts of neuronal activity. The human brain, which is approximately 2% of the body weight, contains approximately 25% of the body's membrane cholesterol. Therefore, the high lipid and cholesterol content of myelin sheaths is extremely important. The total length of myelinated axons is reduced by 27–45% in old age in humans, mostly because of loss of fibers with a small diameter, which myelinate later in development.

Oligodendrocytes are the most vulnerable cells in the brain. Myelin production continues with age and this is evident when the average thickness of myelin sheaths in monkey visual cortex is compared in young and old animals. The mean number of lamellae in the sheaths in young and in old monkeys is 5.6 and 7.0, respectively. An increase in the number of myelin sheaths requires an active incorporation of newly synthesized myelin components. Incorporation and disappearance of the components are compartmentalized into two pools, rapidly and slowly exchanging, and there is evidence that through aging the metabolic turnover in rapidly exchanging myelin is slightly accelerated. Remyelination in the brain is very low due to the expression of several axonal growth-inhibiting molecules. However, astrocytes and macrophages influence remyelination and create together an environment that allows regenerative processes, by producing multiple signaling molecules. Measurement of white-matter volume demonstrates that frontal lobe white-matter volume of healthy men continues to increase into the fifth decade of life, reaching a maximum volume at the age of about 45 years, and then declining. The question of whether there is a turnover of oligodendrocytes has not yet been fully resolved, but, in contrast to neurons, there is some evidence that the number of oligodendrocytes increases.

Oligodendrocytes are markedly affected in the aging cortex, displaying enhanced dense cellular inclusions, swellings of processes, an enhanced turnover of myelin, splitting of myelin sheaths, and enhanced lysosomes and vacuoles. Furthermore, the formation of 'balloons' is found; these are round or spherical cavities and they bulge out and expand the myelin, representing degenerative alterations. The enhanced formation of sheaths with redundant myelin points to uncontrolled production of myelin during the aging process. Impairment of learning and memory is in proportion to the severity of myelin sheath degeneration. Interestingly, in experimental models of glucose applications, oligodendrocytes undergo a senescent program, displaying elongation and thickening of processes and shortening of telomeres. Aging affects several key genes that play a role in the myelin formation, most importantly the genes encoding myelin basic protein (MBP) and proteolipid protein (PLP). Constitutive levels of both of these proteins are decreased during aging. In addition, Gtx, a homeodomain transcription factor that regulates MBP and PLP, is reduced during the aging process. In this context, progesterone and progesterone derivatives counteract the age-associated abnormalities of myelin production and enhance myelin formation by regulating MBP gene expression. In fact, promyelination is enhanced twofold after injury, and interaction of macrophages and astrocytes, mediators of successful myelination, loses its precision as the brain ages. Studies have indicated that steroid hormones and their derivatives have beneficial effects on the age-related decline in maintenance of myelin integrity, and partially reverse the decline in the efficiency of myelin repair.

Oligodendrocytes are particularly susceptible to damage by toxic stimuli, including proinflammatory cytokines, nitric oxide, enhanced calcium, glutamate excitotoxicity, oxidative stress, or complement from macrophages. These degenerative events lead to changes in conduction velocity, formation of redundant myelin, and enhanced thickness of myelin sheaths. In addition, oligodendrocytes contain the highest iron content of all brain cell types, and as much as 70% of brain iron is associated with myelin. Cholesterol- and lipid-synthesizing enzymes in oligodendrocytes require iron for proper function, and age-related increases in iron levels in brain are established characteristics of aging. Since the levels of soluble β-amyloid peptides increase with age, and metals such as iron, copper, and zinc can promote β-amyloid oligomerization and

toxicity, these parameters are likely of high importance in the development of Alzheimer's disease.

The susceptibility of myelination processes over a lifetime to heterogeneous risk factors such as brain trauma, vascular disease, hypertension, nutritional deficiencies, hypercholesterolemia, and reduced hormone levels could explain the pattern of myelin breakdown in aging and age-related diseases. Age-related myelin breakdown is regionally heterogeneous, and differences in myelin properties make later-myelinating regions more susceptible to normal aging processes. Little information is available on the role of oligodendrocytes as myelin-producing cells in neurodegenerative pathologies such as Parkinson's or Alzheimer's disease. The best studied disease in regard to oligodendrocytes is multiple sclerosis, due to the severe damage to oligodendrocytes, but the majority of studies during aging have focused on Alzheimer's disease. In Alzheimer's disease, myelin breakdown is globally exacerbated, consistent with an extracellular deleterious process such as β-amyloid peptide toxicity. In addition, myelin breakdown during aging is associated with apolipoprotein E, which points to an additional role in the initiation of Alzheimer's disease.

See also: Astrocyte: Response to Injury; D-Serine: From its Synthesis in Glial Cell to its Action on Synaptic Transmission and Plasticity; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Glia Control of Blood Flow; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: T Cell Interactions; Microglia identification Methods; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Oligodendrocyte Morphology; Radial Glial Cells: Brain Functions; Transporter Proteins in Neurons and Glia.

Further Reading

Ando S, Tanaka Y, Toyoda Y, et al. (2003) Turnover of myelin lipids in aging brain. *Neurochemical Research* 28: 5–13.

Bartzokis G (2004) Age-related myelin breakdown: A developmental model of cognitive decline and Alzheimer's disease. *Neurobiology of Aging* 25: 5–18.

Blasko I, Stampfer-Kountchev M, Robatscher P, et al. (2004) How chronic inflammation can affect the brain and support the development of Alzheimer's disease in old age: The role of microglia and astrocytes. *Aging Cell* 3(4): 169–176.

Franklin RJM, Zhao C, and Sim FJ (2002) Ageing and CNS remyelination. *NeuroReport* 13: 923–928.

Peters A (2002) The effects of normal aging on myelin and nerve fibers: A review. *Journal of Neurocytology* 31: 581–593.

Goldman S (2003) Glia as neural progenitor cells. *Trends in Neurosciences* 26: 590–596.

Pfrieger FW and Barres BA (1997) Synaptic efficacy enhanced by glial cells *in vitro*. *Science* 277: 1684–1687.

Rozovsky I, Wei M, Morgan TE, et al. (2005) Reversible age impairments in neurite outgrowth by manipulations of astrocytic GFAP. *Neurobiology of Aging* 26: 705–715.

Song H, Stevens CF, and Gage FH (2002) Astroglia induce neurogenesis from adult neural stem cells. *Nature* 417: 39–44.

Wyss-Coray T, Loike JD, Brionne TC, et al. (2003) Adult mouse astrocytes degrade amyloid-beta *in vitro* and *in situ*. *Nature Medicine* 9: 453–457.

Glial Cells: Invertebrate

J A Coles, Grenoble Institute of Neuroscience and Université Joseph Fourier, Grenoble, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Glia are cells that contact and invest neurons, and interact with them, but play at most an ancillary role in rapid electrical signaling. They generally share a common ectodermal origin with neurons. In general, glia surround neurons; sometimes the area of contact is increased by glial projections that fill neuronal invaginations. Present in annelids and insects, these processes are usually called trophospongia (Figure 1(b)) or, at the synapses of insect photoreceptors, capitate processes. Like vertebrate glia, invertebrate glia often contain glycogen granules and show electrical coupling. They also have a prominent cytoskeleton, usually of microtubules in arthropods and of filaments in mollusks and annelids. Large lysosomes called gliosomes (diameter, 1–4 μm) are frequent, and gliograna (diameter, 0.3 μm) are also found.

Research on glial cells has overwhelmingly concentrated on the rat (with some interest in the mouse), and the amount of research done on the several million other animal species, almost all of which are invertebrates, is tiny in comparison. However, in terms of what has been learned about major principles of glial function, studies on invertebrates have made great contributions. Invertebrate glia are most often mentioned in work on the early development of the fruit fly, *Drosophila*. This is because *Drosophila* lends itself to studies based on analysis of mutants and because precursor cells give rise to glial cells and neurons in similar numbers. The way in which genes are switched on and off to determine whether a precursor cell becomes a neuron or a glial cell and also the study of the gene products involved in neuron–glia interactions are very active fields of research. This inextricable interplay of neurons and glia in development is only briefly outlined here. Instead, the emphasis is on the physiology of neuron–glia interactions in adult invertebrates.

From the 1960s, until more powerful techniques became available for studying mammalian glia in the 1980s, a small number of invertebrate preparations with experimentally accessible glia were intensively investigated with the expectation that the results would reveal principles underlying the physiology of glia in general. The giant glial cell of the segmental ganglion of several species of leech (an Annelid), the Schwann cells of giant axons of squid and crayfish, and the 'outer pigment cell' of the retina of the honey bee drone have probably given rise to the majority of publications with this approach. Analysis of the signaling between neurons and glia necessary to the development of a nervous system is under way, particularly in insect sensory systems. However, altogether, the physiology of probably fewer than ten types of invertebrate glial cell has been studied.

Work on these model preparations has been at the origin of several general and important concepts that have been subsequently applied to mammals, and these concepts can usefully be approached through the original invertebrate work.

The Anatomical Distribution of Glial Cells in Invertebrates

Cnidaria: Nervous Systems without Glia

The phylum Cnidaria includes sea anemones, hydra, and jellyfish, all of which have nervous systems without glial cells. One or two ring nerves are found in cubozoans (box jellyfish) and hydrozoan jellyfish. In addition, the small (≈1 cm diameter) cubozoan jellyfish, notably *Tripedalia cystophora*, have elaborate visual systems with up to more than 24 eyes, each with advanced neural circuitry composed of many hundreds of neurons. Despite the absence of glia, these circuits allow rapid and complex hunting behavior.

Bilateria (Triploblastica)

Glial cells are reported in all Bilateria species that have been examined.

Nematoda The hermaphrodite form of the nematode *Caenorhabditis elegans* (Figure 2(a)) has 56 glial cells. Each of the 24 sensilla has a sheath glial cell and a socket glial cell (or two socket cells for the phasmids). In addition, six glial cells, labeled GLR cells, extend sheetlike projections that contact muscle arms in the head. These associate with 302 neurons. The glia create a large part of the architecture of each sensillum. In the case of the amphid sensillum, the socket glial cell has a long process, which follows the dendrites of 12 neurons (only one of which is shown in (Figure 1(b)). At the cuticle, it takes a doughnut shape to create a pocket open to the exterior. At the inner end of the pocket, the sheath cell takes the form of a sieve with eight channels through which pass the sensory cilia of eight of the neurons. Laser ablation of

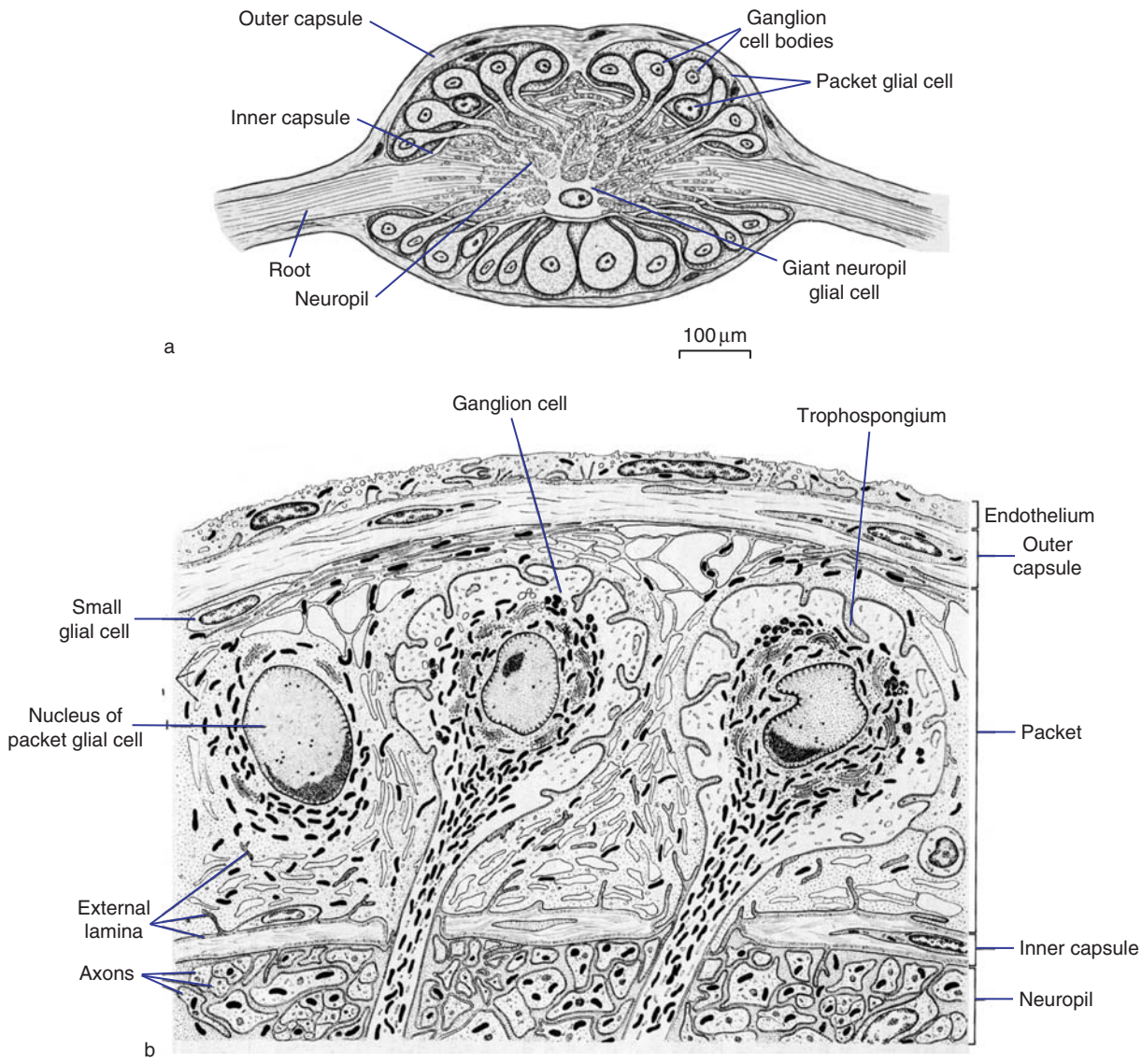

Figure 1 (a) Transverse section through a ganglion of the ventral nerve cord of the leech *Hirudo medicinalis*. Between the inner and outer capsules are six packet glial cells (five shown), each enclosing a packet of neuronal cell bodies. Two giant neuropil glial cells (one shown) lie on the axis of the cord and send processes among the neuropil within the inner capsule. (b) More detailed scheme showing the cell bodies of two packet neurons lying within a packet glial cell. From Coggeshall RE and Fawcett DW (1964) The fine structure of the central nervous system of the leech, *Hirudo medicinalis*. *Journal of Neurophysiology* 27: 229–289.

single cells is contributing to understanding how the cells act together to form the sensillum.

Annelida In leeches such as *Hirudo medicinalis*, the ventral nerve cord has 23 ganglia, each containing the cell bodies of monopolar neurons grouped in six packets. The neuronal somata and initial axon segments of each packet lie within a packet glial cell (Figure 1). Two further giant neuropil glial cells invest the neuropil in the center of the ganglion. The axons connecting neighboring ganglia are enclosed in a glial cell sheath. There is no myelin (here or in any invertebrate); the larger axons are separated one from another by glial processes, but smaller axons are in bundles with no intervening glia.

The neuropil glial cell has a cell body more than 100 μm across and extensive processes; it is probably the largest of all glial cells. For this reason, it was chosen in the 1960s by Stephen Kuffler and colleagues for the first systematic study of glial cell physiology. They found that the ionic basis of the membrane potential depended more strongly on potassium ions than is the case for neurons. They showed that glucose (and other compounds) can diffuse freely from

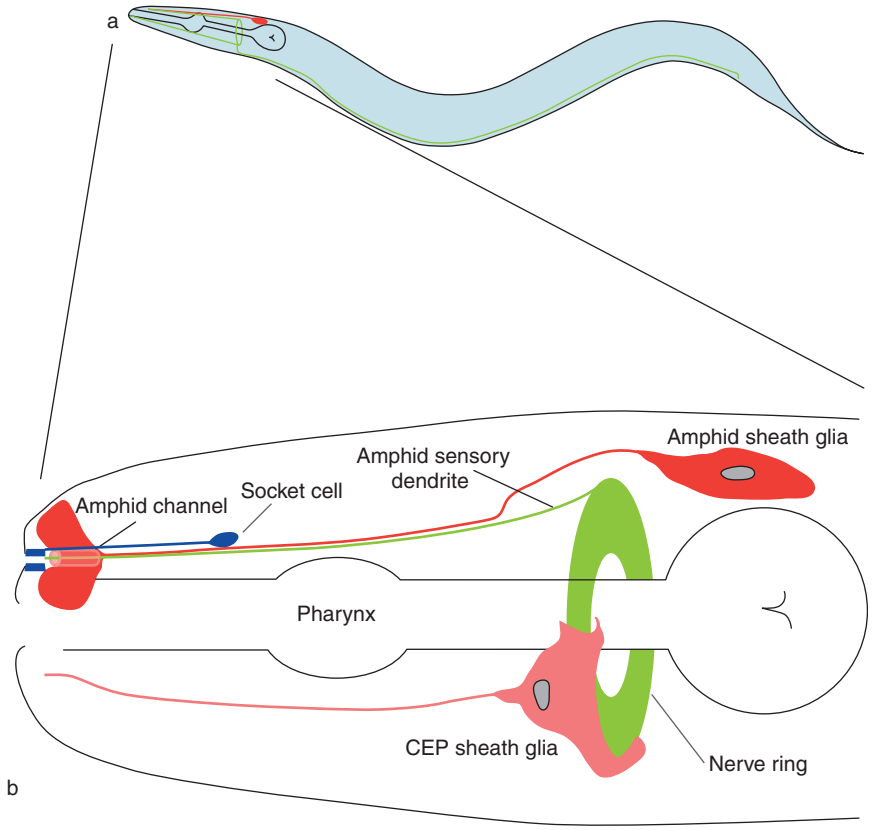

Figure 2 Glia in the nematode *Caenorhabditis elegans*. (a) Schematic of an adult hermaphrodite, approximately 1 mm long, showing the outline of the pharynx. Some neuronal tracts are shown in green, and one amphid sheath glial cell is shown in red. (b) Enlarged view of the anterior region. The nerve ring (green) is a neuropil where most synaptic interactions between neurons occur. Elements of one amphid sensillum are shown: one of the 12 neuronal dendrites (green), the sheath glial cell (red), and the socket glial cell (blue). Adapted from figures in www.wormatlas.org and from Shaham S (2005) Glia–neuron interactions in nervous system function and development. *Current Topics in Developmental Biology* 69: 39–66.

outside the ganglion to the surfaces of the neurons, and they concluded that these leech glial cells do not normally supply metabolic substrate to neurons. This is certainly not universally the case, and it may be exceptional. Over subsequent decades, it has been shown that these glial cells do, in fact, do a lot of interesting things (see Figure 3(a)).

Insecta In the central nervous system of insects (ventral nerve cord and brain) there are three main classes of glia: surface (or subperineurial) glia form a sheath around the entire central nervous system (Figure 4), cortex glia provide a matrix in which are embedded neuronal cell bodies, and neuropil-associated glia send processes into the neuropil. Unlike the neuropil glia of the leech (Figure 1), the cell bodies of insect neuropil-associated glia are numerous (25–30 in each neuromere segment of *Drosophila*) and are located outside the neuropil.

In addition, several types of glia in the periphery can be defined, such as the antennal nerve glial cells of *Manduca* and the outer pigment cells of *Apis* retina (Figure 5).

Urochordata

Ascidiacea (sessile tunicates) The nervous systems of the pelagic larval forms of sea squirts (*Urochordata* and *Ascidia*) are laid out much like those of vertebrates. Nonneural cells, usually called ependymal cells, outnumber neurons. When the larvae become sessile, the brain is remodeled, becoming an order of magnitude larger. In this adult form, glial cells may be absent: the axons are unsheathed, but a few nonneural cells are present within the brain.

Thaliacea (pelagic tunicates and salps) There are no glia in the peripheral nervous system. They are not in evidence in the brain: this is covered by a two-layered epithelium, but the cells of this differ markedly from most glia in that they conduct action potentials.

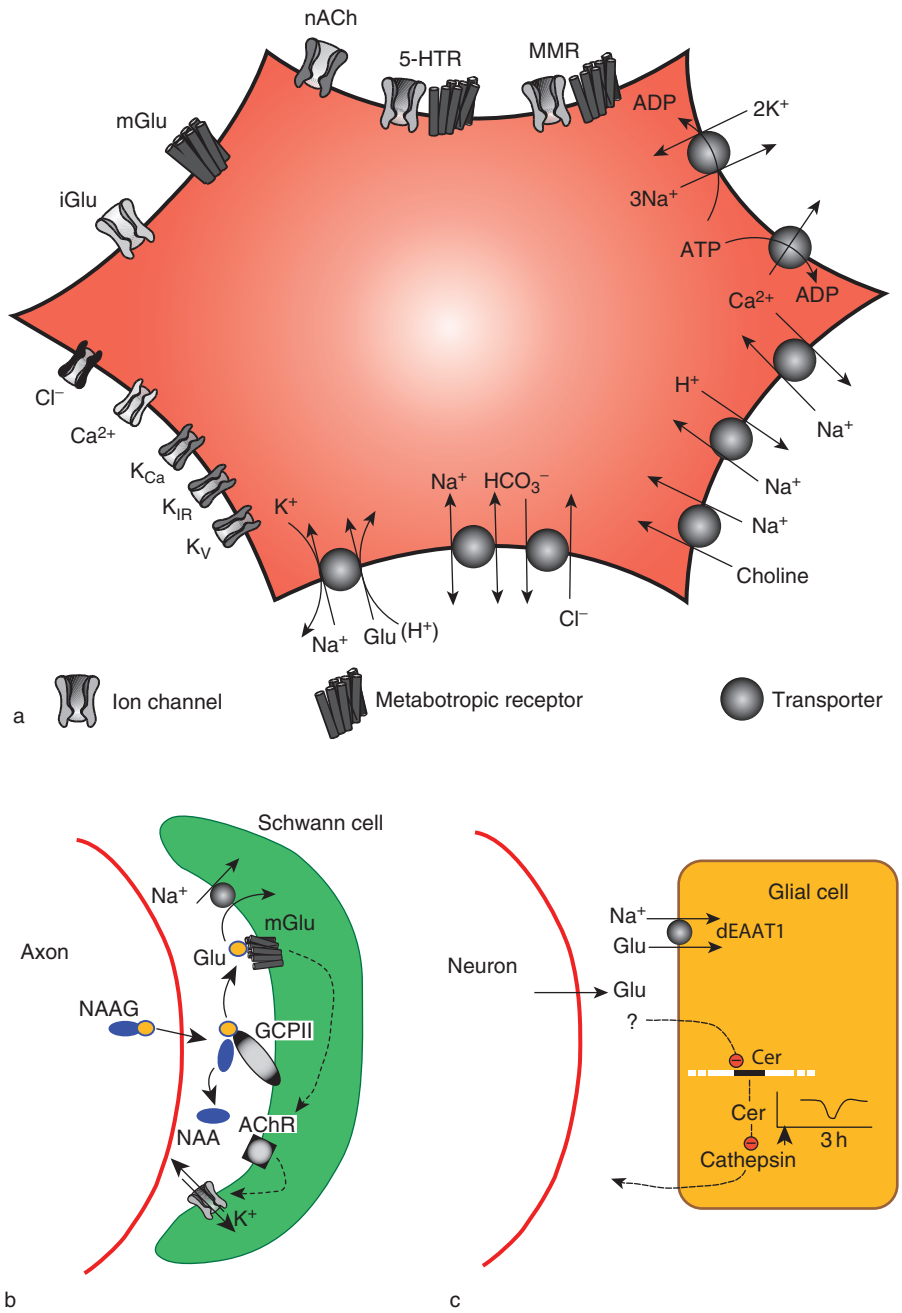

Figure 3 Invertebrate neuron–glia interactions are divers and usually have parallels in vertebrates. (a) Some of the receptors, channels, and transporters found on the giant neuropil glial cell of the leech. iGlu, ionotropic (AMPA/kainate) glutamate receptor; mGlu, metabotropic glutamate receptor; nACh, nicotinic acetylcholine receptor; 5-HTR, receptor for 5-hydroxytryptamine; MMR, receptor for myomodulin, a neurotransmitter in leech. (b) Signaling from giant axon to Schwann cell in the crayfish. Propagation of action potentials along the axon causes release of *N*-acetylaspartylglutamate (NAAG). NAAG is split into *N*-acetylaspartate (NAA) and glutamate (glu) by the extracellular enzyme glutamate carboxypeptidase II (GCPII). Glutamate acts on metabotropic glutamate receptors, which activate parallel signaling pathways, including signaling through extracellular acetylcholine. (c) On glia of *Drosophila* brain, glutamate transporters of the EAAT1 class reduce oxidative stress. In addition, the glia contain a cysteine protease, cathepsin, which is normally inhibited by Cer. Approximately 3 h after conditioned learning, the establishment of long-term memory involves a decrease in Cer concentration leading to increased cathepsin activity. (a) Modified from a scheme kindly supplied by JW Deitmer. (b) Based on results in Urazaev AK, Grossfeld RM, and Lieberman EM (2005) Regulation of glutamate carboxypeptidase II hydrolysis of N-acetylaspartylglutamate (NAAG) in crayfish nervous tissue is mediated by glial glutamate and acetylcholine receptors. *Journal of Neurochemistry* 93: 605–610 and papers cited therein. (c) Based on results in Rival T, Soustelle L, Strambi C, Besson MT, Iche M, and Birman S (2004) Decreasing glutamate buffering capacity triggers oxidative stress and neuropil degeneration in the *Drosophila* brain. *Current Biology* 14: 599–605, and Comas D, Petit F, and Preat T (2004) *Drosophila* long-term memory involves regulation of cathepsin activity. *Nature* 430: 460–463.

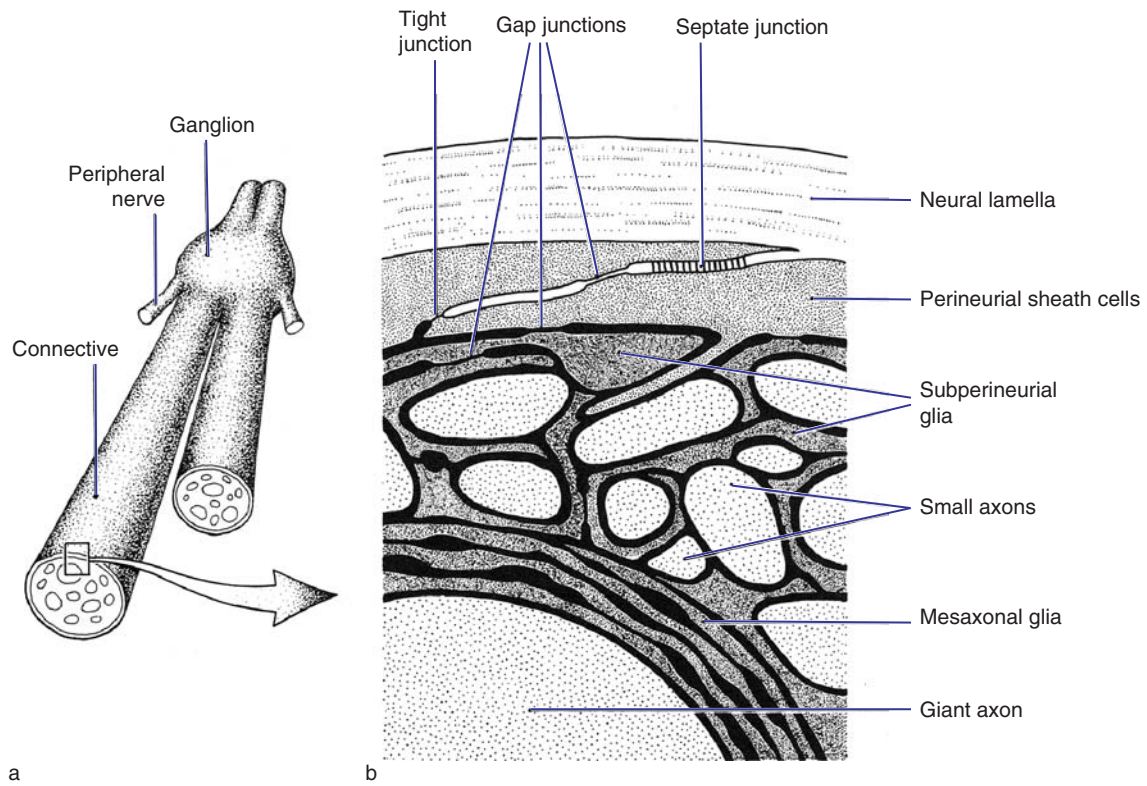

Figure 4 A glial blood–brain barrier. The ventral nerve chord (central nervous system) of the cockroach *Periplaneta americana*. (a) The connectives running between the ganglia are avascular and contain giant axons and small axons (all unmyelinated). (b) The neural lamella provides a connective tissue sheath. Some work suggests that the main barrier to extracellular diffusion of molecules to the axons is formed by tight junctions between the perineurial sheath cells, which are sometimes classed as glial cells. However, work on *Drosophila* suggests that these cells are of mesodermal origin and also that the barrier is constituted by the underlying glial cells. Adapted from Treherne JE and Scholfield PK (1981) Mechanisms of homeostasis in the central nervous system of an insect. *Journal of Experimental Biology* 95: 61–73.

Cephalochordata

The amphioxus or lancelet, *Branchiostoma lanceolatum*, has glial cells. Thin glial lamellae containing gliofilaments enclose the rhabdom of the Joseph photoreceptor cells.

Functions of Glia in Adult Invertebrates

Glia as a Blood–Brain Barrier

The nervous systems of invertebrates are not vascularized; thus, unlike vertebrates, if they have a blood–brain barrier, it is not between capillaries and the brain parenchyma but, rather, around the various elements of the more or less distributed nervous system. The structures usually responsible for occluding extracellular clefts are tight junctions. They are found between glia in the sheaths around the central nervous system of insects and at least some mollusks (e.g., cuttlefish). Insects have a remarkable blood–brain barrier that supports large differences in the concentrations of ions and molecules. In the ventral nerve cord of the cockroach, the barrier is ascribed to tight junctions between the outer perineural sheath cells (Figure 4(b)). Some authors call these cells glia, but examination of *Drosophila* mutants suggests that they are not of ectodermal origin. The retinas of insect compound eyes also have a glial blood retina barrier, and in *Apis*, at least, this barrier separates fluids with different concentrations of amino acids and carbohydrates (Figure 5(a)).

In contrast, in ganglia of the leech (Annelida) and the snail (Gastropoda), ions and molecules with a molecular weight of up to approximately 10 kDa can diffuse readily through the extracellular clefts of the glial sheath to reach the neurons.

Glia and Homeostasis of Extracellular Ion Concentrations

Although groups of short axons not separated by any glial sheath are found in animals from jellyfish to man, activity more subtle than the transmission of action potentials benefits from a regulated ionic environment. Potassium is the ion of most concern since

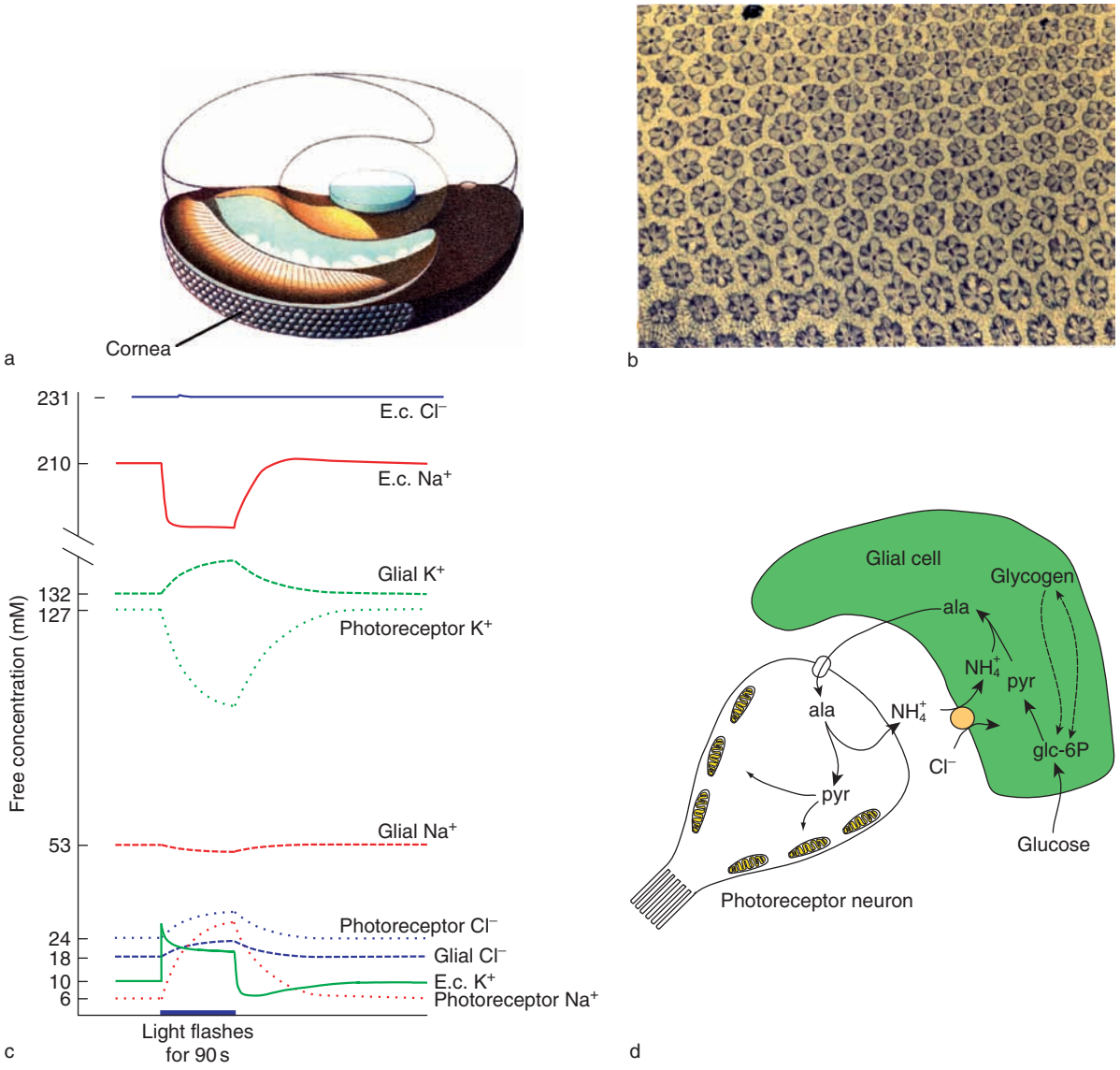

Figure 5 The retina of the drone (male) honey bee: a model for studying glia–neuron exchanges. (a) Drawing of a slice of the head, caudal upwards. Much of the chitin that normally covers the caudal surface of the head has been removed to expose the retina and brain. The facetted cornea covers much of the surface of the head (here, the facets are shown too large; there are approximately 10 000 in total). Behind each facet, the cell bodies of six large photoreceptor neurons extend for 400–500 μm to the basal membrane (brown line). (b) In section, the large photoreceptor cells appear clustered in sixes, like the petals of a rosette. The outlines the glial cells can be seen clearly in the lower left. The rosettes are approximately 25 μm in diameter. (c) Stimulation of the neurons with their physiological stimulus (light) causes large movements of ions. The initial and final values are the measured mean values; the time courses are schematic. Green, K^+; red, Na^+; blue, Cl^-. The volumes of the fluid in the three compartments have been estimated as follows: photoreceptors, 0.29 liters per liter of tissue; glial cells, 0.44 liters; and extracellular space, 0.05 liters. K^+ is removed from the extracellular space both by net uptake into the glia and by spatial buffering. (d) Metabolic exchanges. Glucose is taken up only by the glia. Pyruvate formed from glucose or glycogen is converted to alanine, which is transferred to the neurons. Ammonium, liberated when alanine is reconverted to pyruvate, returns to the glia being taken up on a specific Cl^-dependent cotransporter. (a, b) Adapted from Coles JA (1989) Functions of glial cells in the retina of the honeybee drone. *Glia* 2: 1–9. (c) Data from Coles JA, Orkand RK, Yamate CL, and Tsacopoulos M (1986) Free concentrations of Na, K, and Cl in the retina of the honeybee drone: Stimulus-induced redistribution and homeostasis. *Annals of the New York Academy of Sciences* 481: 303–317, and Coles JA, Orkand RK, and Yamate CL (1989) Chloride enters glial cells and photoreceptors in response to light stimulation in the retina of the honey bee drone. *Glia* 2: 287–297. (d) Based on Tsacopoulos M, Poitry-Yamate CL, and Poitry S (1997) Ammonium and glutamate released by neurons are signals regulating the nutritive function of a glial cell. *Journal of Neuroscience* 17: 2383–2390, and Marcaggi P, Thwaites DT, Deitmer JW, and Coles JA (1999) Chloride-dependent transport of $NH4^+$ into bee retinal glial cells. *European Journal of Neuroscience* 11: 167–177.

extracellular [K^+] affects the membrane potential of neurons, and active neurons release K^+, thus tending to increase extracellular [K^+]. Extracellular [K^+] is normally quite low, with the highest reported value being 10 mM (in bee retina), so addition of modest amounts of K^+ produces considerable fractional increases in [K^+]. Homeostasis of extracellular [K^+] has been studied particularly in the retina of the honey bee drone and in the giant nerve fiber of the squid. Regulation of extracellular [H^+] (pH) by transport of H^+ or HCO_3^- has been studied in ganglia of the leech, in which the presence of a glial Na^+-HCO_3^- cotransporter was discovered.

Ion homeostasis in the retina of the bee (*Apis*) The drone (male) honey bee has a highly performing eye, which it uses in pursuing queen bees. The functional units of the retina, the ommatidia, are arranged with an almost crystalline regularity imposed by the facets of the cornea (Figures 5(a) and 5(b)). The retina is composed essentially of two populations of cells – the photoreceptor neurons and the major glial cells, which were originally called 'outer pigment cells.' The drone retina is the only nervous tissue in which measurements have been made (using ion-selective microelectrodes) of the free concentrations of the major ions, K^+, Na^+, and Cl^-, in the three compartments – neurons, glial cells, and extracellular space. Stimulation of the photoreceptor neurons with light causes considerable shifts of ions between the neurons and glia, whereas the concentrations in the extracellular clefts show only modest changes (Figure 5(c)). It was found that the glial cells clear excess extracellular K^+ in two ways. The glial cells are connected to each other by gap junctions. When the activation of the neurons is nonuniform, extracellular [K^+] rises locally, and K^+ enters the glial cells and somewhat depolarizes the functional syncytium of glial cells. This depolarization drives K^+ out of the glial cells in places where extracellular [K^+] is only slightly raised. The process is called 'spatial buffering.' In addition, glial cells of drone retina also clear extracellular K^+ by taking it up together with Cl^- and water, which leads to swelling of the glial cell. This behavior is not shared by inactive neurons, which take up very little K^+. The situation appears to be similar in mammalian brain, in which raised extracellular [K^+] causes astrocytes to swell more than do neurons.

Possible role of glial cells in homeostasis of extracellular K^+ around the giant axon of the squid The giant axon of the squid (a cephalopod) is unmyelinated but achieves a high conduction velocity by having a large diameter, which can exceed 500 μm. It is surrounded by a thin continuous sheath composed of Schwann glial cells, each approximately 2 μm thick, 20 μm wide, and 100–600 μm long. This sheath is further surrounded by connective tissue. The ion channels of both the axon and the Schwann cells are well characterized.

Two Schwann cell conductances have been described – an L-type Ca^{2+} conductance and an outwardly rectifying K^+ conductance activated by entry of Ca^{2+}. These currents are functionally almost identical to conductances found on mammalian Schwann cells. Although propagation of action potentials along the axon causes release of K^+ through the axon membrane, little increase in [K^+] in the periaxonal space is observed in healthy squid nerves. The K^+ conductances of the Schwann cells are not suited to clearing K^+. Possibly, [K^+] does not rise excessively simply because the extracellular space is large.

Glia and the Removal of Neurotransmitter

Astrocytes in vertebrates take up neurotransmitter glutamate. The neuropil glial cells of the leech (Figures 1 and 3(a)), and glial cells of *Drosophila* (Figure 3(c)), are also equipped to do this. In the retina of the honey bee drone, the neurons release glutamate and the glial cells take it up in the absence of synapses – a situation with parallels to vertebrate nerve tracts, which also release glutamate. In *Drosophila*, the absence of the transporter dEEAT1 increases oxidative stress. Photoreceptors in insects and barnacle (*Balanus*, Crustacea) use histamine as their principal neurotransmitter. In barnacle, the glial cells make little or no contribution to recycling this histamine. In contrast, in *Drosophila*, glial cells are involved in at least two ways. Formation of neurotransmitter vesicles occurs on the neuronal membrane close to glial capitate processes that invaginate the neurons. Also, the enzymes coded by the genes *ebony* and *tan*, which are involved in histamine metabolism in the glial cells, are essential to normal neurotransmission.

Glia as a Site of Glycogen Metabolism

In many insects, reserves of glycogen and lipids (principally in the glia) are depleted during starvation, and the glial trophospongia which invaginate neurons are well placed to mediate transfer of substrate to the neurons. In isolated slices of bee head, not supplied with glucose or other exogenous metabolic substrate, glial glycogen is broken down to provide metabolic substrate (alanine) which is transferred to the photoreceptor neurons (Figure 5(d)). However, the metabolism of glycogen is active and regulated in complex ways, suggesting that it is not a passive

store but has physiological roles in normally fed animals. In isolated ganglia of the leech, electrical stimulation in the absence of exogenous glucose first causes a decrease, then an increase, and then a decrease in glycogen content. Light stimulation *in vivo* of *Apis* eye increases, rather than decreases, incorporation of glucose into glycogen in the retinal glia. In isolated leech ganglia, glycogen content is decreased by 5-hydroxytryptamine, dopamine, histamine, and octopamine and increased by GABA and glycine. In the retinas of bee drones, the quantity of glycogen in the glia is sufficient to support photoreceptor function for many hours, which seems a useless weight to be carried by an animal that does not leave its well-stocked hive for more than a few tens of minutes.

Glia as Suppliers of Fuel to Neurons

Do glial cells, in addition to providing fuel from glycogen during starvation, have a role in the normal supply of metabolic fuel to the neurons? A clear answer is available only for the retina of the drone honey bee. In slices of drone retina, glucose appears to be taken up exclusively by the glia. Pyruvate, formed by glycolysis, is converted to alanine, which leaves the glia and is taken up by the photoreceptors. In the photoreceptors, alanine is reconverted to pyruvate, liberating ammonium, which passes back to the glia, being taken up on a Cl^- cotransporter specific for NH_4^+ over K^+ (Figure 4(d)).

Glia and the Maintenance of Axon Function

Local synthesis of axon proteins It is quite simple to squeeze out most of the axoplasm from a squid giant axon, or to perfuse it, and the ability to do this has contributed greatly to understanding how proteins are maintained in axons. Proteins move down axons at a rate of 1–4 mm per day and most have half-lives of less than 1 or 2 weeks. Axons can be many meters long (as in giant squid); thus, at least in axons more than a few centimeters long, proteins have to be produced locally. It is well established that when squid axons, severed from their cell bodies but still ensheathed by glial cells, are bathed in radioactive amino acids, labeled proteins appear in the axoplasm. Furthermore, when the axon is bathed in a labeled RNA precursor ([^{3}H]uridine), labeled RNA appears in the cytoplasm of the axon. Transcription machinery appears to be present within the squid axon. The only known available DNA templates are in the glial cells; thus, presumably RNA is transferred from the glial cells to the axon. The mode of this transfer is unknown and the idea is controversial. Similar conclusions have been reached for the terminals of squid photoreceptors and for the giant Mauthner neuron of goldfish.

Rapid Signaling from Neurons to Glia

Neurotransmitter-mediated effects of neuronal activity on glia have been studied particularly in the neuropil glia of the leech (Figures 1 and 3(a)), in the glia surrounding the giant axons of squid and crayfish (Figure 3(b)), and in the glia ensheathing neurons of the snail. In all these cases, bursts of neuronal activity activate a Ca^{2+}-dependent K^+ conductance, usually leading to a hyperpolarization. This may improve the clearance of K^+ from the extracellular space and the efficacy of Na^+-dependent membrane transport, such as Na^+–HCO_3^- cotransport, and glutamate uptake.

A striking feature of these glial cells is the presence of receptors for a wide variety of transmitters. Leech neuropil cells, for example, respond to glutamate, 5-hydroxytryptamine, acetylcholine, adenosine phosphates, and myomodulin (a transmitter best known in gastropods).

The glial response to an agonist can be extremely complex. In the ventral nerve cord of the crayfish, the passage of a train of impulses along the giant axons causes the axons to release *N*-acetylaspartylglutamate (NAAG), which leads, by a cascade of events, to hyperpolarization of the periaxonal glia. The main pathway requires the presence of the extracellular enzyme glutamate carboxypeptidase, which splits off glutamate from NAAG, and metabotropic glutamate receptors. Acetylcholine plays a role in the cascade: it is released by the glial cells and acts back on glial acetylcholine receptors leading to activation of K^+ channels (Figure 3(b)). All the details of the overlapping pathways have not been worked out, but the analysis has been carried further than in vertebrates, in which elements of this signaling system have been found at the rat neuromuscular junction. At least some elements of this signaling are also found between the giant axon and its Schwann cells in the squid.

Signaling between glia and neurons in *Drosophila* In the mushroom bodies of *Drosophila* brain, *cer* encodes an inhibitor of a subfamily of cysteine proteinases named cathepsins. Three hours after conditioned learning, expression of *cer* in glial cells is transiently reduced (Figure 3(c)). Interference with this pattern of expression impairs long-term memory. In mammals, prolonged activation of cysteine proteinases is associated with neuronal degeneration in

Alzheimer's disease. This is an example of how the genetic manipulations that can be done relatively quickly in *Drosophila* can give clues to the functioning of mammals.

No evidence of multiple neuron–glia signals in bee retina In marked contrast to leech giant glial cells, snail glial cells, and the periaxonal glia of squid and crayfish, no neurotransmitter signaling to the glial cells of drone retina has been reported.

Glia as Actors in the Development of the Nervous System

For analyzing the genetic and molecular mechanisms involved in the development of a nervous system, the nematode *Caenorhabditis elegans* and a few insects have proved useful models. The development is rapid, the number of cells involved is not enormous, and genetic tools are available. At various stages, glial cells can regulate the proliferation, migration, and apoptosis of neurons.

Early Development of the Nervous System

For both *Caenorhabditis* and *Drosophila*, neurons and glia are roughly equal players in early development and are derived from common precursors. A landmark discovery was the role in *Drosophila* of the gene *glial cells missing* (*gcm*), whose expression determines the glial fate of all the glia except the midline glia, a small class of neuropil-associated glia. The gene product, Gcm, is a transcription factor that directly or indirectly activates transcription of approximately 50 genes, most of which have mammalian orthologs. Some Gcm-activated genes (e.g., *repo*, *loco*, and *pnt*) determine that a precursor becomes a glial cell rather than a neuron, whereas *tramtrack* represses neuronal fate. Other Gcm-activated genes are expressed only during a certain time or only in a subset of glial cells. This means that cofactors are required that determine the maturation and diversity of the glia.

Development of Sensory Systems

Most insects have sensory systems that are highly perfected yet relatively small, and some have been selected as models for a detailed analysis of their development. This can be illustrated by work on the olfactory system of the moth *Manduca sexta*.

The olfactory system of the moth, *Manduca sexta* The main olfactory receptor neurons in insects are in the antennae, and their axons extend to the olfactory lobe of the brain. In *Manduca sexta*, the receptor neurons for different molecular attributes of odorants are spatially intermixed in the antennal epithelium, but their axons, on entering the olfactory lobe, are sorted into fascicles according to their odorant specificity and terminate in glomeruli whose number increases during maturation to reach 64 ± 1 in the adult. This system develops over approximately 18 days during the metamorphosis from larva to adult. The axons of the olfactory receptor neurons initially grow down the antenna without glia (Figure 6(a)). On arriving outside the antennal lobe, the axons encounter sorting zone glial cells and reciprocal neuron–glia interactions occur. The axons send signals to the glia, which then increase in number. Also, the glia contribute to the sorting of the axons: On contacting glia, the growth cones enlarge, explore, turn, and find like axons (Figure 6(c), stages 4–6). Glial cells of a second kind, the olfactory ensheathing cells, migrate along the axons from the receptors and, later, enwrap bundles of axons.

In the olfactory lobe, the neuronal cell bodies are clustered in its periphery, outside an initially spherical zone of neuropil (Figure 6(a)). The cell bodies of the neuropil glial cells surround this neuropil. The outer parts of the neuropil progressively separate into glomeruli, and the layer of neuropil glial cells is deformed to follow this change of shape (Figure 6(b)). The neuropil glia show reciprocal interactions with the receptor cell axons. The axons induce properties of the neuropil glia, including expression of Ca^{2+} channels. Also, the neuropil glia induce synapse formation. Some elements of the neuron–glia signaling are known. In the case of the neuropil glia, acetylcholine released by the ingrowing receptor axon terminals depolarizes the glia. This activates voltage-dependent Ca^{2+} channels, and the resulting increase in intracellular free $[Ca^{2+}]$ in some way initiates glial cell migration (Figure 6(d)).

A glial function, myelination, is responsible for major differences in the sizes and life styles of vertebrates and invertebrates Electrical signaling by changing the electrical potential across a neuronal plasma membrane consisting of a lipid bilayer is slow and energetically expensive. This is because the capacitance of the lipid bilayer is very high ($1\,\mu F\,cm^{-2}$), which means that many ions have to cross it to change the potential. In many vertebrate nerves, the effective membrane capacity of the axon membrane is reduced by a myelin sheath, produced by glial cells. This makes possible the high nerve impulse conduction velocities necessary for large animals capable of rapid and complex movements, such as dinosaurs and humans. The structure most resembling myelin in invertebrates may be the several loosely wrapped layers of

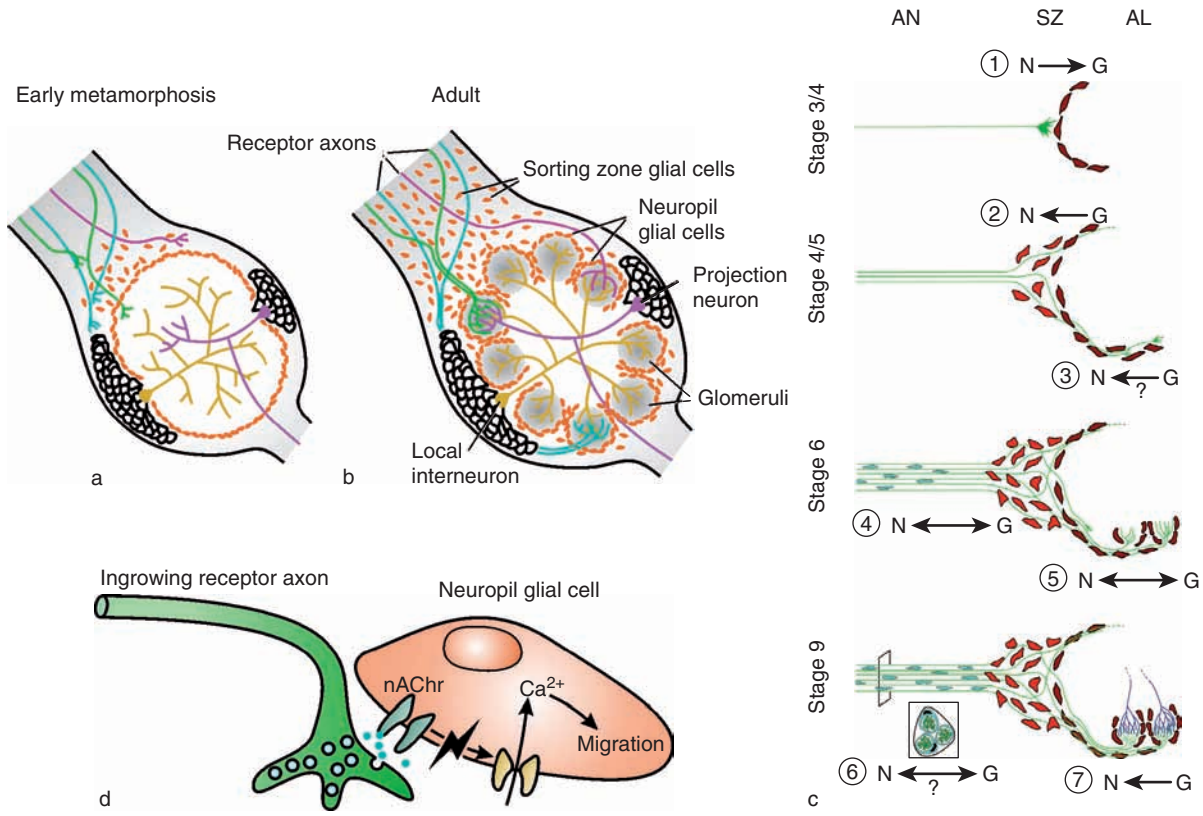

Figure 6 Glial cells contribute to the development of the olfactory system in the moth *Manduca sexta*. (a, b) The antennal lobe with the base of the antennal olfactory nerve (top left). During metamorphosis, axons from the olfactory epithelium (not shown) are sorted to appropriate odor-specific glomeruli. Probable neuron–glia interactions are numbered in (c). (1) The first axons to reach the lobe encounter central glial cells and induce them to proliferate. (2) Glia migrate into the nerve to become sorting zone glia. Axons encountering sorting zone glia form large growth cones. (3) *In vitro*, the growth cones of axons contacting neuropil glia cells become larger and more complex. (4) Glia arising in the olfactory receptor epithelium migrate along the olfactory nerve (not shown in b). (5) Receptor axons induce neuropil glial cells to extend stabilizing processes around protoglomeruli. (6) Antennal nerve glial cells enwrap bundles of receptor axons. (7) Neuropil glia define the borders of the glomeruli. (d) The signaling mechanism involved in step (5). Acetylcholine released by the axon terminal depolarizes the glial cell, and the depolarization opens voltage-gated Ca^{2+} channels, leading to an increase in intracellular $[Ca^{2+}]$. (a, b, d) Adapted from Lohr C and Deitmer JW (2006) Calcium signalling in invertebrate glial cells. *Glia* 54: 642–649. (c) Adapted from Tolbert LP, Oland LA, Tucker ES, et al. (2004) Bidirectional influences between neurons and glial cells in the developing olfactory system. *Progress in Neurobiology* 73: 73–105.

membrane that are sometimes found around axons of *Drosophila*, although *Drosophila* possesses homologs of only a few of the genes required for myelin. This wrapping is probably not useful for conducting active nerve impulses, but it would speed up passive, electrotonic conduction. The invertebrate solution to the problem of nerve conduction velocity is either to stay small (as are insects) or to make axons of very large diameter. These giant axons take up much space and their use is mainly limited to stereotyped escape reactions, as is the case in annelids and squid.

Lessons from Invertebrate Glial Cells

The conclusion from the small number of invertebrate glial cells whose functions have been studied is that in adult animals the proteins that serve as ion channels, transporters, receptors, and elements of intracellular signaling pathways have properties very similar to those of vertebrate glia. It is true that some invertebrate glia show behavior that has not been clearly described in vertebrates, such as a possible involvement of cathepsin in memory formation in *Drosophila*, or the choice of alanine, rather than lactate, as a substrate transferred from glia to neurons, as in bee retina. However, it would be premature to state that similar phenomena do not occur in vertebrates. The involvement of glial cells in the development of insect nervous systems also shows many close parallels with mammals, although many of the genes involved appear to have different functions in vertebrates. It should be borne in mind that

an invertebrate nervous system has evolved through many more life cycles than has that of *Homo sapiens*, even though the body shape of the animal may have changed little over several hundred million years.

See also: Glial Energy Metabolism: A NMR Spectroscopy Perspective; Learning and Memory in Invertebrates: *Drosophila*; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Olfaction in Invertebrates: *Drosophila*; Visual System Development: Invertebrates; Visual System: Invertebrates.

Further Reading

Carlson SD and SaintMarie RL (1990) Structure and function of insect glia. *Annual Review of Entomology* 35: 597–621.

Coggeshall RE and Fawcett DW (1964) The fine structure of the central nervous system of the leech, *Hirudo medicinalis*. *Journal of Neurophysiology* 27: 229–289.

Coles JA, Orkand RK, Yamate CL, and Tsacopoulos M (1986) Free concentrations of Na, K, and Cl in the retina of the honeybee drone: Stimulus-induced redistribution and homeostasis. *Annals of the New York Academy of Sciences* 481: 303–317.

Coles JA, Orkand RK, and Yamate CL (1989) Chloride enters glial cells and photoreceptors in response to light stimulation in the retina of the honey bee drone. *Glia* 2: 287–297.

Coles JA and Deitmer JW (2005) Extracellular potassium and pH: Homeostasis and signalling. In: Kettenmann H and Ransom BR (eds.) *Neuroglia,* 2nd edn., pp. 334–345. New York: Oxford University Press.

Coles JA (1989) Functions of glial cells in the retina of the honeybee drone. *Glia* 2: 1–9.

Comas D, Petit F, and Preat T (2004) *Drosophila* long-term memory involves regulation of cathepsin activity. *Nature* 430: 460–463.

Freeman MR, Delrow J, Kim J, Johnson E, and Doe CQ (2003) Unwrapping glial biology: Gcm target genes regulating glial development, diversification, and function. *Neuron* 38: 567–580.

Garm A, Ekström P, Boudes M, and Nilsson DE (2006) Rhopalia are integrated parts of the central nervous system in box jellyfish. *Cell and Tissue Research* 325: 333–343.

Gommerat I and Gola M (1996) Glial potassium channels activated by neuronal firing or intracellular cyclic AMP in Helix. *Journal of Physiology* 495: 649–664.

Inoue I, Tsutsui I, Abbott NJ, and Brown ER (2002) Ionic currents in isolated and *in situ* squid Schwann cells. *Journal of Physiology* 541: 769–778.

Ito K, Urban J, and Technau GM (1995) Distribution, classification, and development of *Drosophila* glial cells in the late embryonic and early larval ventral nerve cord. *Roux's Archives of Developmental Biology* 204: 284–307.

Lohr C and Deitmer JW (2006) Calcium signalling in invertebrate glial cells. *Glia* 54: 642–649.

Marcaggi P, Thwaites DT, Deitmer JW, and Coles JA (1999) Chloride-dependent transport of $NH4^+$ into bee retinal glial cells. *European Journal of Neuroscience* 11: 167–177.

Parker RJ and Auld VJ (2006) Roles of glia in the *Drosophila* nervous system. *Seminars in Cell & Developmental Biology* 17: 66–77.

Pentreath VW (1995) Metabolic interactions between neurons and glial cells in leech and snail ganglia. In: Vernadakis A and Roots BI (eds.) *Neuron–Glia Interactions during Phylogeny*, pp. 161–196. Totowa, NJ: Humana Press.

Pereneanu W, Shy D, and Hartenstein V (2005) Morphogenesis and proliferation of the larval brain glia in *Drosophila*. *Developmental Biology* 283: 191–203.

Rival T, Soustelle L, Strambi C, Besson MT, Iche M, and Birman S (2004) Decreasing glutamate buffering capacity triggers oxidative stress and neuropil degeneration in the *Drosophila* brain. *Current Biology* 14: 599–605.

Shaham S (2005) Glia–neuron interactions in nervous system function and development. *Current Topics in Developmental Biology* 69: 39–66.

Tolbert LP, Oland LA, Tucker ES, et al. (2004) Bidirectional influences between neurons and glial cells in the developing olfactory system. *Progress in Neurobiology* 73: 73–105.

Treherne JE and Scholfield PK (1981) Mechanisms of homeostasis in the central nervous system of an insect. *Journal of Experimental Biology* 95: 61–73.

Tsacopoulos M, Poitry-Yamate CL, and Poitry S (1997) Ammonium and glutamate released by neurons are signals regulating the nutritive function of a glial cell. *Journal of Neuroscience* 17: 2383–2390.

Urazaev AK, Grossfeld RM, and Lieberman EM (2005) Regulation of glutamate carboxypeptidase II hydrolysis of N-acetylaspartylglutamate (NAAG) in crayfish nervous tissue is mediated by glial glutamate and acetylcholine receptors. *Journal of Neurochemistry* 93: 605–610.

Relevant Websites

http://flybase.bio.indiana.edu – FlyBase.
http://www.sdbonline.org – Society for Developmental Biology.
http://www.wormatlas.org – Wormatlas. Altun ZF and Hall DH (eds.) 2002–2006.

Glial Cells: Microglia during Normal Brain Aging

M B Graeber

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

Background

Microglia were first described at the turn of the nineteenth century, independently by the neuropathologists Franz Nissl (in Munich) and William Ford Robertson (in Edinburgh). In Spain, in 1913, the Nobel prize-winning physician and histologist Ramón y Cajal described these cells as a *tercer elemento* ('third element') among central nervous system cells. However, it was Cajal's student, Pio del Rio-Hortega, who, around 1920, coined the term 'microglia' and conducted the first systematic studies on this cell type. Many of the observations of Rio-Hortega were of fundamental character and are still valid.

Morphology

After staining with Rio-Hortega's silver carbonate method, resting microglia show characteristic elongated, almost bipolar cell bodies with spinelike processes arising at a right angle and carrying short branches (Figure 1). The nucleus is large in comparison with the small perikaryon and contains heterochromatin apposed to the nuclear envelope. In electron micrographs, microglia show a dense cytoplasm, vimentin-positive intermediate filaments, a few microtubules, and prominent cisterns of granular endoplasmic reticulum, especially when activated, which run in parallel to the main axis of the cell (Figure 2). Lysosomes and Golgi complexes are also present.

Cellular Markers

After their discovery and description, there was decades-long controversy surrounding microglial identity and nature; the uncertainty about these cells, which lasted until the 1980s, can be explained by the remarkable phenotypic plasticity of microglia. Their appearance ranges from that of highly ramified cells in normal central nervous system (CNS) tissue (Figure 1) to rod cells in chronically inflamed cerebral cortex (Figure 3) to full-blown macrophages in tissue-destructive pathologies (Figure 4). The problem of defining the nature of microglia was solved when lectin and immunocytochemical markers became available, and when experimental models, including bone marrow chimeras, unequivocally demonstrated that microglia are closely related to cells of the peripheral monocyte–macrophage system. However, no marker possesses true, singular microglia specificity, and all known histological microglia markers are only microglia specific in the sense that they do not label other glia or neurons. Other macrophages that normally occur within the bony confinements of the CNS – for example, perivascular, plexus, and meningeal macrophages – typically cross-react with microglial detection methods, but many of the molecules expressed by the former are absent from microglia. Thus, the immunophenotype of microglial cells appears to be 'downregulated.' Before monoclonal antibodies became available, enzymes such as thiamine pyrophosphatase and nucleoside diphosphatase were used to identify microglial cells in tissue sections. Furthermore, the ectoenzyme 5′-nucleotidase is positive in the plasmalemma of microglia and strongly expressed in activated cells, but is not cell type specific. Nonspecific esterase, a monocytic marker, is negative.

Distribution

Resting microglia account for up to 20% of all glial cells in the CNS. Their almost uniform distribution varies rather little between brain regions, although there appear to be local differences in the expression of some molecules, such as major histocompatibility complex (MHC) class II antigens, which are present at higher levels in white matter and in the aged brain.

Origin

During vascularization of the CNS, cells related to the mononuclear phagocyte system invade the central nervous tissue and settle as resting microglia. They form the resident macrophage precursor cell pool of the brain, spinal cord, and retina. Renewal of juxtavascular (perivascular) and parenchymal microglia from the bone marrow occurs at an extremely low rate in the adult. Perivascular macrophages ('perivascular cells'), which are separated from the brain parenchyma by a basement membrane, are anatomically, immunophenotypically, and functionally distinct from microglia.

Function

The main function of microglia in the normal brain and spinal cord appears to lie in the formation of a network of highly responsive surveillance and defense cells. Accordingly, microglia serve as 'sensors of threats' (a term introduced by GW Kreutzberg) to the CNS. Microglia are potentially cytotoxic, but

healthy microglia do not attack their microenvironment except in the case of CNS infections. Thus, the metaphor of microglial 'guardians' that are armed to defend seems more appropriate than that of harmful 'killer cells.' In contrast to neuroectodermal glia, microglial cells are not electrotonically coupled via gap junctions. This may explain their very localized involvement in disease processes, which is of great diagnostic use. Antigen presentation is a likely function of microglia. Complement receptors are quickly upregulated in response to various stimuli. The mitotic activity of microglia is regulated by macrophage and granulocyte/macrophage colony-stimulating factors (M-CSFs and GM-CSFs) and the appropriate receptors. Other cytokines, such as interleukins (IL-1, IL-6, and IL-10) and transforming growth factor-β (TGF-β1), function as important signals in the control of microglial activation. Brain macrophages of microglial origin have a strong respiratory burst activity, meaning that they can produce oxygen radicals.

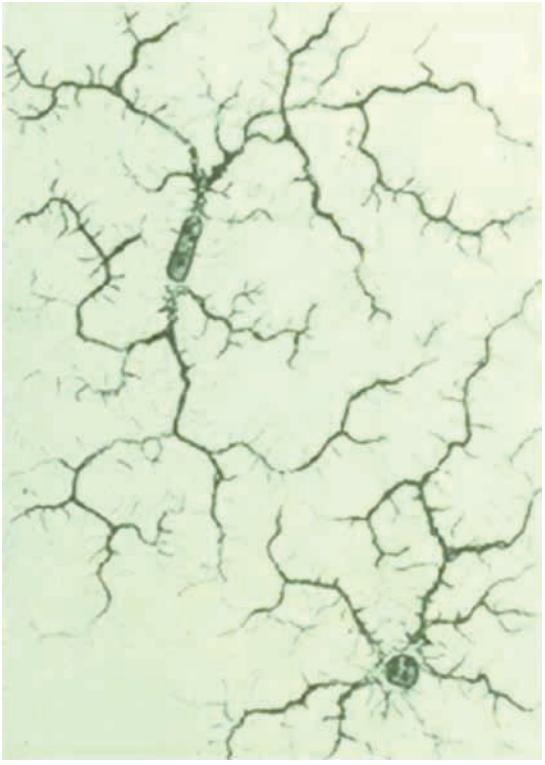

Figure 1 Microglia as illustrated by Pio del Rio-Hortega in 1920.

Pathology

More than a dozen names for microglia document their involvement in neuropathologies such as inflammation and vascular disease, and in removal of tissue debris after trauma, pointing to the phagocytic nature of the reactive microglial cells. Terms still in use are foam cells, gitter cells, rod cells, and fat granule cells. The term 'ameboid microglia' should be reserved to describe microglia as they appear in developing tissue. It is thought that microglial phagocytes become motile and migrate toward lesion sites, where they are loaded with tissue debris, which they then transport to the meninges or vascular system. Perivascular cuffs of fat-laden phagocytes can be seen many months and even years after acute tissue destruction. Axotomy of motor nerves leads not only to retrograde changes of the motor neurons but also to proliferation of perineuronal microglial cells. After a few days these microglia are found covering most of the surface of the regenerating motor neurons, thus displacing the synaptic terminals ('synaptic stripping'). In this situation no phagocytosis is observed. Based on observations under various pathological conditions, a concept of a graded response of microglial cells has been developed. The diagram in Figure 5 describes the stepwise activation of microglia from a resting to a phagocytic state.

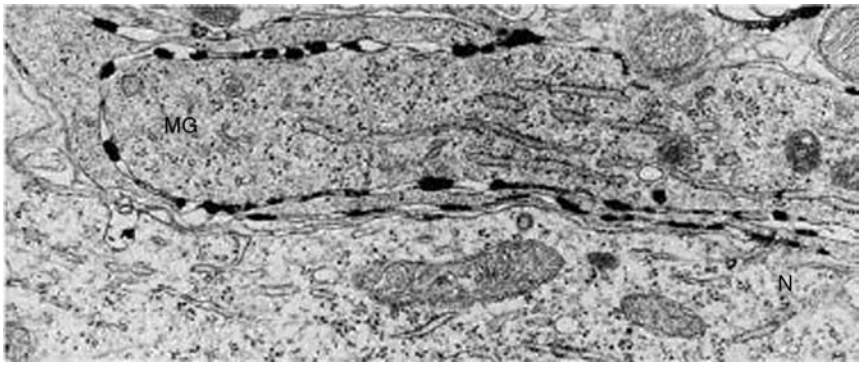

Figure 2 Activated perineuronal microglia (MG) cell covering the surface of a chromatolytic facial motor neuron (N) in the rat. The MG has a dense cytoplasm and an active rough endoplasmic reticulum arranged in parallel broad cisterns. Electron-dense reaction product indicates the presence of ecto-5′-nucleotidase in the plasmalemma of the MG. Electron micrograph 1:40 000. From Kreutzberg GW (1999) Microglia. In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience*, 2nd edn. Amsterdam: Elsevier.

Figure 3 Rod cells in the human cerebral cortex, stained for major histocompatibility complex class II molecules, in a case of subacute sclerosing panencephalitis, a condition in which rod cells are typically found.

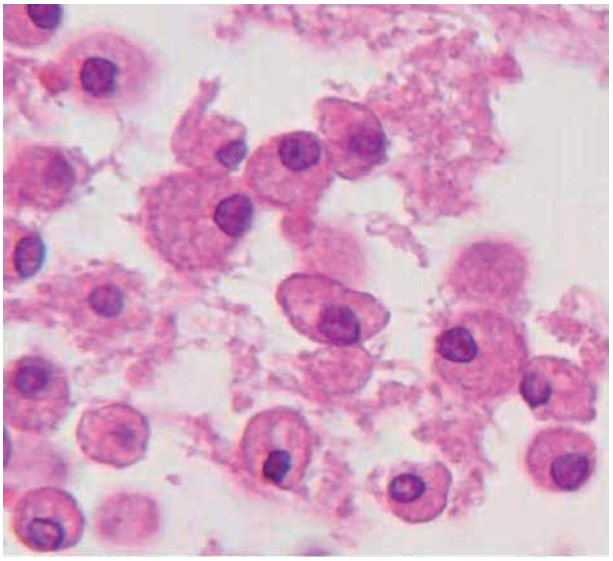

Figure 4 Brain macrophages in a cerebral infarct. Human brain; hematoxylin–eosin staining.

Figure 5 Microglial activation cascade. Activated microglia (middle) usually show a shortened, stout process but can also have the appearance of rod cells (**Figure 3**), depending on the anatomical region in which the cells reside. Activated microglia are not necessarily phagocytic (**Figure 2**).

See also: Bergmann Glial Cells; D-Serine: From its Synthesis in Glial Cell to its Action on Synaptic Transmission and Plasticity; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Gap Junctions and Hemichannels in Glia; Glia Control of Blood Flow; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Growth Factors; Glial Cells: T Cell Interactions; Macroglial Lineages; Microglia Properties; Microglia identification Methods; Microglial Response to Injury; Radial Glial Cells: Brain Functions; Transporter Proteins in Neurons and Glia.

Further Reading

Angelov DN, Walther M, Streppel M, et al. (1998) The cerebral perivascular cells. *Advances in Anatomy, Embryology & Cell Biology* 147: 1–87.

Flügel A, Bradl M, Kreutzberg GW, et al. (2001) Transformation of donor-derived bone marrow precursors into host microglia during autoimmune CNS inflammation and during the retrograde response to axotomy. *Journal of Neuroscience Research* 66: 74–82.

Graeber MB, Kreutzberg GW, and Streit WJ (eds.) (1993) *Microglia: Special issue* 7: 1–119.

Kreutzberg GW (1996) Microglia: A sensor for pathological events in the CNS. *Trends in Neurosciences* 19: 312–318.

Perry VH (1994) *Macrophages and the Nervous System*. Austin: R. G. Landes Co.

Rio-Hortega Pdel (1932) Microglia. In: Penfield W (ed.) *Cytology and Cellular Pathology of the Nervous System*, vol. 2, pp. 483–534. New York: Hafner [1965 facsimile of 1932 edition].

Glial Cells: T Cell Interactions

C Infante-Duarte and F Zipp, Charité – Universitätsmedizin Berlin, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

T cells are a major subset of lymphocytes that exert a central function during the specific immune response to pathogens. Interaction of T cells and cells of the central nervous system (CNS) is part of the normally occurring host defense against CNS pathogens and is crucial for combating infectious agents and preventing persistent infection inside the nervous system. However, an uncontrolled immune response inside the CNS may lead to serious pathologies, such as autoimmunity or neurodegeneration.

T Cell-Mediated Immunity

In order to successfully fight against pathogens, the mammalian immune system has developed into two arms – (1) innate and (2) adaptive or specific immunity. Innate immunity commences immediately upon pathogen entry and represents a very fast and generic response which does not lead to immunological memory. In contrast, the cells of adaptive immunity are able to recognize antigens specifically, to expand after exposure to the same antigen, and to develop into memory cells. T lymphocytes, together with the B lymphocytes, represent the cellular compounds of the adaptive immune system (Table 1). T cells originate in the bone marrow and mature in the thymus (T stands for thymus). T lymphocytes recognize specific antigens via the T cell receptor (TCR) and, in contrast to B cells that can recognize soluble antigens, respond only to antigens when they are presented by major histocompatibility complex (MHC) molecules on antigen-presenting cells (APCs).

There are two main subpopulations of T cells – cytotoxic T cells (CD8-positive) and helper T cells (CD4-positive). Cytotoxic T cells (CTLs) recognize the antigens in the context of MHC class I molecules, expressed on nearly all nucleated cells. Class I molecules present endogenous antigens, such as viral or tumor-derived peptides, which are indicative of intracellular alterations. CTLs are thereby able to identify and destroy infected or mutated cells. On the other hand, helper T cells (Th cells) (which can in fact also be cytotoxic) recognize antigens in the context of MHC class II molecules expressed on APCs. MHC class II molecules present exogenous antigens, such as microbial peptides, which result when APCs engulf and process pathogen-derived material. After interacting with the APC, Th cells become activated and are able in turn to trigger effector cells (e.g., B cells or CTLs), causing the destruction of the pathogen. Thus, T cells orchestrate the host immune system by secreting cytokines to regulate the immune response to infection. Cytokines are soluble proteinic factors which act on different cell types. They are redundant (most cytokine functions can also be carried out by other cytokines) and pleiotropic (a cytokine exerts many different functional effects), characteristics which make it difficult to understand their functional network. Cytokines function by binding to specific cell surface receptors on their target cells. Cytokine receptors are present not only on immune but also on various nonimmune cell types, such as astrocytes and oligodendrocytes.

Th cells can be subdivided into three distinct effector subpopulations according to their cytokine profile and effector functions. Th1 lymphocytes secrete interleukin (IL)-2, interferon gamma (IFN-γ), and tumor necrosis factor-β (TNF-β), and they induce cell-mediated immune responses; Th2 lymphocytes selectively produce IL-4, IL-5, IL-10, and IL-13 and promote humoral immune responses; whereas Th17 cells secrete IL-17, IL-6, and TNF-α and induce activation and recruitment of neutrophils. Regulatory T cells (T reg, Tr1, and Th3) are also CD4-positive T cells which, together with other immune cells, are responsible for controlling activation of the immune system once an antigen has been eliminated from the body, as well as for maintaining immune homeostasis (Figure 1). These cells also need to become activated via the TCR in order to exert their suppressor function. Failure of regulatory T cells to function properly may result in chronic inflammation or even in autoimmune disease.

Heterodimeric TCR is associated with proteins of the CD3 complex and with the CD4 or CD8 coreceptors on Th cells and CTLs, respectively. Interactions between the CD4 molecule and MHC class II or the CD8 molecule and class I MHC stabilize the binding between TCR and MHC molecules, thus accomplishing the process of antigen recognition. CD3 as well as CD4 and CD8 are therefore suitable markers for identifying and quantifying all T cells (CD3-positive) or their different subtypes. The ligation of TCR and its coreceptors (CD4 and CD8) with the MHC/antigen complex is necessary but not sufficient to induce a physiologic stimulation of T cells, a process which involves at least one additional activation signal. This second signal is dependent on either soluble factors such as APC-derived cytokines or the ligation of

Table 1 Natural and specific immunity

	Mechanisms involved	*General features*
Natural or innate immunity	Skin, mucous membranes Macrophages, neutrophils, NK cells Complement IFN-α/β, IFN-γ, TNF-α	Cannot distinguish between distinct antigens Are not able to expand after exposure to the same antigen Do not lead to immunological memory
Adaptive or specific immunity	T and B lymphocytes Antibodies Lymphocyte-derived cytokines	Specific for a given antigen Their production or expansion is stimulated specifically by an antigen Lead to immunological memory

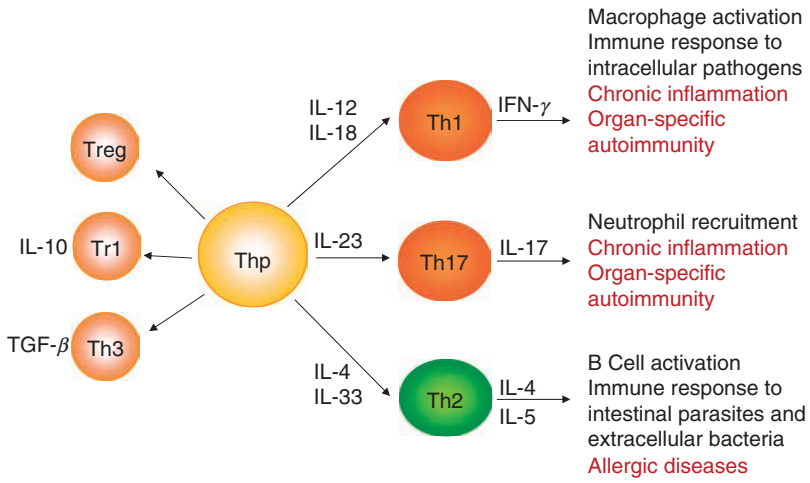

Figure 1 Phenotype development. Polarized Th cell subsets differentiated from a common Th precursor (Thp) in the periphery. Although differentiation is influenced by several factors, the most important of these are the cytokines present at the stage of TCR stimulation. Effector Th1, Th17, and Th2 cells protect against different microbial pathogens by producing cytokines able to mobilize different mechanisms of defense. On the other hand, an imbalanced Th effector response can promote different systemic and organ-specific disorders (text in red). In contrast to the effector Th cells, the factors influencing regulatory cell differentiation are not well defined.

cell surface proteins, the so-called costimulatory molecules, and it serves to initiate, maintain, and regulate the immune response. In line with this, T cells encountering antigens in the absence of adequate costimulation may remain unresponsive or anergic. Thus, costimulation has been postulated to play a critical role in determining T cell activation versus anergy. Throughout the years, several ligand–receptor pairs functioning as costimulators have been described; however, the best characterized costimulatory pathways remain the CD80/CD86 (B7)-CD28 and CD40-CD40L systems (Figure 2).

Immune Surveillance of the Central Nervous System

The CNS has often been considered as an 'immunologically privileged organ,' with limited prerequisites for mounting an immune response owing to, for example, the lack of a specialized lymphatic drainage, the absence of dendritic cells, the low expression levels of MHC class I and II, and the presence of the blood–brain barrier (BBB). The BBB is constituted by endothelial tight junctions and surrounding astrocytes which restrict the leukocyte traffic into the CNS. However, it has become evident that activated lymphocytes can cross the BBB, inspect the normal CNS in the absence of inflammation, and recognize target antigens within the CNS. This active immune surveillance of the CNS is essential to protect the brain from pathogen-induced damage. CNS infection and inflammatory disorders such as autoimmune reactivity promote the breakdown of the BBB, and under these conditions a much higher number of activated leukocytes penetrate the brain parenchyma to interact with neurons and glial cells.

Interaction of T Cells with Glial Cells

Glial cells represent the majority of cells in the nervous system (approximately 90% in some regions of

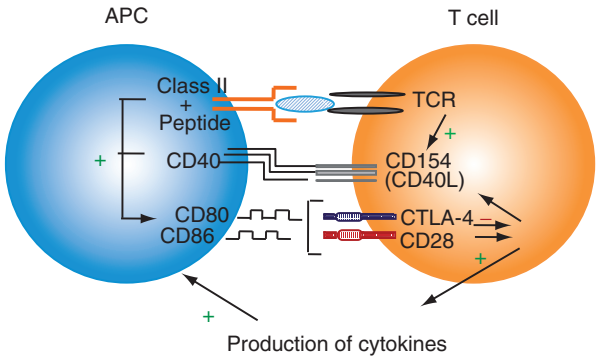

Figure 2 Scheme of the complex interactions between Th lymphocytes and APC. The presentation of antigen/MCH class II complexes to naive Th cells results in the upregulation of CD80/CD86 expression on APC and in the expression of CD40L on T cells. The ligation of this newly expressed CD40L with CD40 on APC also induces CD80 and CD86 on APC. At this point, the costimulatory signal provided by the APC is received by the T cells via the CD28/CTLA-4 receptors, which drive the T cells to enter into cell cycle, to complete T cell activation, and to produce cytokines. On the other hand, the signals originated through CD28 can also upregulate CD40L in the Th cells.

the human brain). They are nonneuronal cells that supply support and nutrition to neurons, maintain homeostasis, provide myelin sheaths, are integral parts of synapses, and participate in brain development. Here, we outline the interaction of T cells with the main glial populations in the CNS – that is, microglia, astrocytes, and oligodendrocytes.

Microglia

Microglia represent 10–20% of the glial cell population in the brain and are considered as phagocytic immune cells of the CNS. They share many features with macrophages and develop from a common myeloid precursor cell; therefore, they have been considered as potential resident APCs of the CNS. Microglia can be divided into two main populations: the parenchymal and the perivascular microglia. In contrast to the short-living macrophage-like perivascular microglia, parenchymal microglia can live for as long as the individual and express low levels of MHC class I or II in an inactive/resting state.

Microglia, and in particular perivascular microglia, may act as the professional APC population of the CNS, being involved in processing and presentation of target antigens to infiltrating T cells at the level of the BBB. Once an infection or inflammatory process is established inside the CNS, parenchymal microglia also become capable of presenting antigens. Parenchymal microglia become activated in response to many different inflammatory signals, such as IFN-γ, granulocyte–macrophage colony-stimulating factor (GMCSF), or TNF-α. Activated microglia express MHC class II and costimulatory molecules such as CD80, CD86, and CD40, and they are able of interacting with T cells in an antigen-dependent manner. This interaction directly affects the T cell effector function without inducing T cell proliferation, a mechanism which may support immune surveillance of the CNS, necessary for detecting and combating infections, and at the same time controls T cell expansion and subsequent inflammation, permitting a rapid T cell elimination.

It has been shown that activated microglia can regulate different effector Th responses, including Th1, Th2, and Th17, as well as regulatory Th responses. T cells activated by microglia are then able to produce effector cytokines, and perhaps also neurotrophic factors, which will support brain regeneration after an infection-promoted acute immune response. In response to pathogens or during inflammation (e.g., IFN-γ exposure), microglia secrete high amounts of IL-1, TNF-α, and IL-12, a cytokine that induces Th1 phenotype differentiation. This may help microglia modulating the development of a Th1 versus Th2 response necessary for the elimination of certain pathogens. At the same time, however, if the immune response is not properly controlled, this Th1-dominated response may support microbial-induced tissue damage or the development of autoimmune processes. In contrast, interactions with Th1 and Th2 cells, or interaction with T cells in the absence of CD40–CD40L ligation, induce the production of anti-inflammatory prostaglandin in microglia, a substance which may be neurotoxic but necessary for limiting T cell response. On the other hand, rat microglia activated *in vivo* by graft versus host disease do not express the costimulatory molecules CD80 and CD86, are unable to induce T cell expansion, and promote T cell apoptosis. Thus, the ability of microglia to influence T cell effector function is dependent both on their state of activation and on the state of activation and differentiation of the Th cells, indicating that T cell–microglia interaction consists of bidirectional communication in which both cell types regulate each other. As mentioned previously, the T cell-derived cytokine IFN-γ is a strong factor in promoting parenchymal microglia maturation into APC. Th1-derived IFN-γ and interaction of microglial CD40 with its ligand CD40L on T cells are known to induce IL-12 secretion by microglia even in the absence of antigenic presentation. In organotypic cultures, Th1 cells switch the expression of costimulatory molecules on microglia from CD86 to CD80, whereas Th2 cells decrease CD86 expression without affecting the expression of CD80. Furthermore, Th2 cells can modulate APC activity of microglia by

enhancing, for instance, their IL-10 production or by downregulating microglial expression of the intercellular adhesion molecule 1 (ICAM-1).

In summary, in the case of pathogenic insult, T cells patrolling the CNS can recognize antigen on antigen-presenting perivascular microglia. Inside the brain parenchyma, these reactivated antigen-specific T cells will, in turn, activate the parenchymal microglia and promote their maturation to APC, which is necessary for supporting a T cell response and eliminating pathogens. Under normal conditions, this could induce an acute but controlled inflammatory cascade, which might promote some self-resolving CNS injury. Thus, whether an infection will result in a chronic pathology or will remit after acute immune response depends on the balance of pro-versus anti-inflammatory and neurotoxic versus neurotrophic signals derived primarily from the interaction of T cells and microglia.

Astrocytes

Astrocytes represent the most abundant glial cell type. Astrocytes are responsible for maintaining homeostasis in the CNS, participate in the formation and maintenance of the BBB, function as detoxificating cells in the neuronal microenvironment, and produce a variety of neurotrophic and neuroprotective factors. Furthermore, they exert an ill-defined immune function inside the CNS. Inactive astrocytes do not express MHC class I or class II molecules. Similar to microglia, astrocytes are able to secrete IL-1, IL-6, and TNF-α in response to a diversity of stimuli and to present antigen to T cells after becoming properly activated by IFN-γ, for example. However, since reports on astrocytic immune functions are very controversial, it is under debate whether they can act as APCs at all. On the one hand, it has been shown that murine resting astrocytes constitutively express CD86 and are able to promote expansion of activated Th1 cells, whereas treatment with IFN-γ induces astrocytic expression of MHC class II and the costimulatory molecules CD80 and CD86. Thus, IFN-γ-treated astrocytes may be capable of antigen processing and presentation as well as of priming naive T cells. On the other hand, a large body of evidence indicates that astrocytes do not express costimulatory molecules and may deliver an inhibitory or tolerogenic signal to T cells, inducing anergy or apoptosis instead of proliferation. In addition, it has been postulated that activation of Th cells in the absence of costimulation could promote Th2 phenotype differentiation or activate exclusively Th2 cells, implicating astrocytes in the control/suppression of undesired inflammatory immune responses. In line with this, we have previously shown that adult astrocytes are capable of inducing T cell apoptosis via the CD95/CD95 ligand pathway. Since astrocytic endfeet contribute to the formation of the BBB, this depletion mechanism may play an important role as the brain's first line of defense against activated T cells.

Oligodendrocytes

Oligodendrocytes are glial cells specialized in the elaboration and maintenance of myelin in the CNS. Oligodendrocytes do not express MHC class I or MHC class II constitutively. Thus, under normal conditions these cells are unable to interact with T cells in an antigen-dependent manner. It has been shown that although few oligodendrocytes express MHC class I in the healthy human brain, this expression becomes upregulated during inflammation, such as in multiple sclerosis (MS). In contrast, oligodendrocytes seem not to express MHC class II molecules under any condition investigated so far, and consequently they cannot present antigens to Th lymphocytes. However, the fact that oligodendrocytes express Th2-like cytokine receptors, such as IL-4 receptor (R) and IL-10R, even under noninflammatory conditions, points to a possible communication path between both cell types. Additionally, during inflammation oligodendrocytes are able to express additional cytokine receptors, such as IL-12R or IFN-γR. This cytokine-mediated communication appears to be essential in the case of pathogenic insult affecting oligodendrocytes. In this context, it has been reported that a successful clearance of mouse hepatitis virus (JHMV strain) from rat oligodendrocytes requires expression of IFN-γR.

In addition, interactions between T cells and oligodendrocytes represent an important piece of the immune response in demyelinating pathologies such as MS. These aspects of T cell–glial cell interaction in inflammatory pathologies are described in more detail in the next section.

Interaction of T Cells with Glial Cells in Different Disorders

This section focuses on pathological T cell–glial cell interactions in acute and chronic inflammatory diseases, exemplified by human immunodeficiency virus type 1 (HIV-1)-induced encephalitis and MS, respectively, as well as in three examples of primarily noninflammatory neurodegenerative disorders: adrenoleukodystrophy (ALD), stroke, and Alzheimer's disease (AD).

Acute Inflammation

Encephalitis is an acute inflammation of the brain, commonly caused by a viral infection. In encephalitis

caused by HIV-1, microglia cells are the primary target of the infection and serve as a viral reservoir. Microglial infection leads to severe tissue damage in the CNS and ultimately to HIV-associated dementia, characterized by progressive motor disturbances, impaired memory and cognitive function, as well as behavioral abnormalities.

It is assumed that HIV enters the CNS very early after peripheral infection of circulating T cells and monocytes, probably as a passenger inside the infected cells. In the CNS, the most infected cells are the perivascular microglia, perhaps because they become replenished by infected circulating monocytes. The role of infected CD4-positive T cells inside the brain is unclear, and the question of whether parenchymal microglia become infected is still open. A hallmark of HIV encephalitis is the formation of multinucleated giant cells resulting from the fusion of infected and uninfected microglia/macrophages. The first stage of CNS infection with HIV appears to be characterized by a potent T cell response against the virus. The presence of large perivascular infiltration of CTLs, and the fact that during infection activated microglia produce lymphoattractant chemokines such as CCL3, indicates that at this stage a microglia–T cell interaction takes place. This strong immune response seems to be involved in HIV-induced neurodegeneration in a bystander way; that is, neurons are most likely damaged as a consequence of uncontrolled inflammatory processes. HIV-infected microglia express high amounts of chemokines and proinflammatory cytokines (mainly IL-1β and TNF-α), which may induce activation of other uninfected cells, alter the permeability of the BBB, and promote the migration of HIV-specific activated T cells into the brain. These cells may contribute to the amplification of the immune response and to HIV-induced neurotoxicity. The latter is mediated mainly by neurotoxic factors such as TNF-α, TNF-related apoptosis-inducing ligand (TRAIL), arachidonic acid, nitric oxide (NO), or quinolinic acid (probably generated by HIV-induced indoleamine 2,3-dioxygenase, which is produced by microglia and astrocytes). On the other hand, as demonstrated for other neurodegenerative diseases, it has been shown that the interaction between lymphocytes and microglia/macrophages is essential for promoting tissue repair after virus clearance. Even factors such as TNF-α, which are known to be potent neurotoxins, can also manifest neurotrophic effects under certain neuroinflammatory conditions. Thus, the degree of neurotoxicity or neuroprotection depends on the balanced interaction between cells of adaptive and innate immunity.

Chronic Inflammation

By definition, chronic inflammation is an inflammatory response of prolonged, even indefinite, duration which inevitably causes tissue damage and is provoked by persistence of the causative stimulus in the tissue. The classic example is MS, considered an autoimmune disease in which the immune response is presumably directed toward myelin-related autoantigen peptides. Myelin-specific Th cells are believed to orchestrate an immune reaction leading to damage to myelin and oligodendrocytes and thus severe disability in young adults. Studies of MS and its animal model, experimental autoimmune encephalomyelitis (EAE), have demonstrated that in addition to myelin and oligodendrocyte damage, axonal injury and loss of parental cell bodies are involved in disease severity. MS and EAE have been associated with an aberrant Th1- and also with Th17-dominated response to autoantigens. Conversely, several anti-inflammatory cytokines produced by Th2 or regulatory T lymphocytes, such as IL-4, IL-10, and IL-13, contribute to controlling pathogenic T cell response. As the professional APCs of the CNS, microglia are considered to be directly involved in antigen presentation to pathogenic T cells. Interestingly, in early MS lesions the expression of CD80 is increased, suggesting that CD80 upregulation is somehow involved in pathogenic T cell–microglia interaction. Myelin-specific Th1 cells induced the microglia expression of MHC class II, CD40, CD80, and CD86, as well as the expression of IL-12, a cytokine that increases IFN-γ secretion in Th cells. In inflammatory MS lesions, microglia express ICAM-1, CD40, CD80, and CD86, suggesting that these cells are matured to full APCs (perhaps stimulated by myelin-specific Th1 cells) and become able to reactivate newly infiltrated antigen-specific Th cells. This would represent an inflammatory loop, in which both cell types are likely involved in tissue damage during MS. However, microglia appear to be not only involved in tissue injury through the secretion of neurotoxic factors but also involved in modulating the pathogenic immune response, controlling CNS inflammation, and inducing neuroregeneration. The same phenomenon appears to occur in the contribution of T cells to immune pathogenesis. T cells are clearly involved in disease pathogenesis; however, it has been demonstrated that myelin-reactive T cells produce neurotrophic factors and support neuronal survival. This protective effect seems to rely in part on microglia. To what degree microglia–T cell interaction can be pathogenic or protective is a result of the cross-regulatory interactions between both cell types. This becomes

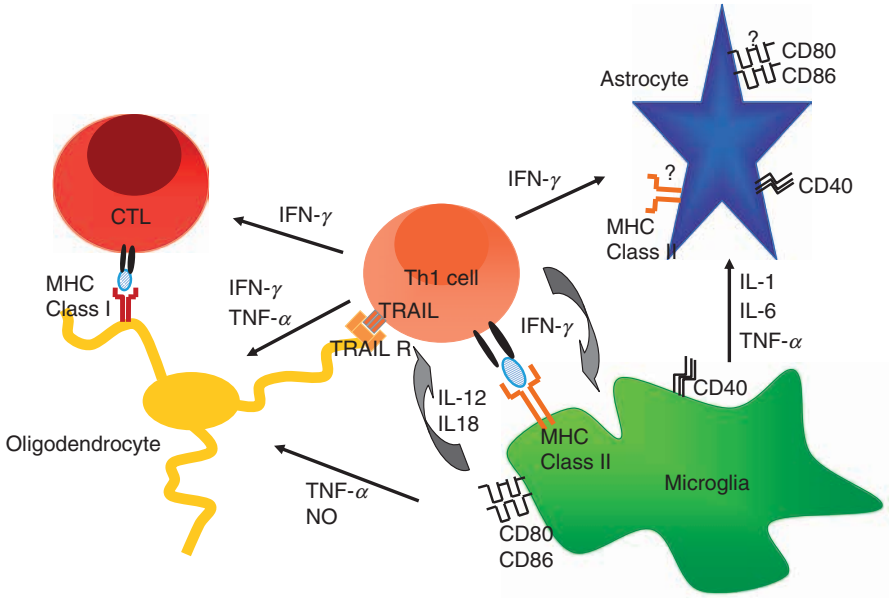

Figure 3 Possible interactions of T cells with glial cells in chronic CNS inflammation. Activated Th cells, such as myelin-specific Th1 cells, are reactivated inside the CNS by antigen-presenting perivascular microglia. Th1-derived IFN-γ activates parenchymal microglia expressing MHC class II, CD40, CD80, and CD86, as well as IL-12, a proinflammatory cytokine that further increases IFN-γ secretion by Th cells. Mature microglia are then able to reactivate newly infiltrated antigen-specific Th cells. Microglia-derived TNF-α and NO, and also Th-derived TNF-α or TRAIL can damage oligodendrocytes and interfere with myelin production. IFN-γ may also influence oligodendrocytes and astrocytes. Activated oligodendrocytes express MHC class I and can be attacked by antigen-specific CTL, which in turn is also stimulated by Th1-derived IFN-γ. Furthermore, the microglia-derived cytokines IL-1, IL-6, and TNF-α induce astrogliosis. In this inflammatory scenario, astrocytes may mature to APC and become capable of presenting antigen and of activating further myelin-specific T cells.

evident when considering that genetic susceptibility to MS has been associated with certain MHC class II molecules – that is, with particular interactions between APCs and T cells.

Not only microglia but also astrocytes have been considered a key cell type in initiating the autoimmune response. As previously mentioned, IFN-γ induces astrocytic expression of costimulatory and adhesion molecules. Some investigators also postulate that astrocytes of MS plaques express MHC class II molecules, although these data are conflicting and have not been confirmed by other authors. Furthermore, various studies failed to detect costimulatory molecules on astrocytes in MS plaques. Thus, the contribution of astrocytes to MS pathogenesis is still a subject of debate. Moreover, evidence indicates that astrocytes may instead be involved in neuroprotection and downregulation of the autoreactive T cell response. Astrocytic production of nerve growth factor (NGF) is increased during CNS inflammation, and Th cells significantly increase production of NGF in an antigen-dependent manner. Furthermore, astrocytes can suppress myelin-specific Th1 and Th2 cell activation, proliferation, and effector function, contributing perhaps to the immune privilege of the CNS. On the other hand, Th cells also modulate astrocytic function. Findings show that myelin-specific T cells inhibit glutamate uptake in astrocytes, which may contribute indirectly to glutamate-mediated axonal injury.

Oligodendrocytes are thought to represent the initial cellular target of the autoimmune attack during MS. Interactions of T cells with oligodendrocytes, via cell–cell contact and/or soluble factors, appear to be involved in oligodendrocyte injury but also in induction of remyelination during demyelinating pathology. The fact that oligodendrocytes upregulate MHC class I expression under inflammatory conditions supports the assumption that CTLs may be involved in myelin destruction during MS. Furthermore, in active MS lesions oligodendrocytes express not only Th2-like cytokine receptors but also receptors for IL-12, IL-18, and IFN-γ, indicating that they can respond to inflammatory cytokines produced by activated microglia as well as by Th1 lymphocytes (Figure 3).

Primarily Noninflammatory Disorders

Whereas neurodegenerative diseases are disorders of the CNS characterized by the gradual and progressive loss of neural tissue or nerve cells, stroke represents

an acute injury of the CNS. In the pathology of these diseases, the role of inflammation is only just beginning to be understood. Limited immune response in the CNS ('immune privilege') may permit the initiation of neurodegeneration, for example, upon aging. Moreover, breakdown of the immune privilege during disease course presumably contributes to severity. Whereas microglia seem to become activated in most cases of neurodegeneration, the contribution of T cell populations appears to vary between different disorders. In ALD, for instance, a T cell-mediated inflammatory process seems to determine clinical outcome. By contrast, in AD the role of T cells is apparently marginal, and inflammation may be caused mostly by cells of the innate immune system.

ALD is a rapidly progressive X-linked inherited recessive disorder characterized by the accumulation of saturated very long-chain fatty acids predominantly in the white matter of the brain and the adrenal glands. The disease leads to MS-like symptoms, but in a very rapid course, and is usually lethal in childhood. The cerebral lesions show massive breakdown of myelin and a perivascular lymphocyte infiltrate of mainly CD4-positive and CD8-positive T cells. Hematopoietic stem cell transplantation, whose effectiveness is most likely based on immunosuppression and then reconstitution of the immune system, is one of the current therapies in patients in the early stage of the cerebral inflammatory phenotype. The beneficial effect of this therapy indicates that immune cells, such as T cells present in the lesions, might be involved in disease pathogenesis. Although CD8-positive T cells are believed to be involved in oligodendrocyte cytolysis, it is still unclear what role is played by CD4-positive T cells. Neither the phenotype nor the specificity of these T cells has been defined, and it remains an open question whether these cells are able to interact with activated microglia in an antigen-dependent manner.

Stroke is one of the major causes of mortality worldwide, occurring when the blood flow to a part of the brain is suddenly blocked by occlusion and reduced delivery of oxygen causes the death of brain cells. Depending on the part of the brain that is affected, stroke can result in different clinical manifestations, such as weakness or paralysis of the arms, legs, and/or facial muscles and loss of vision or speech. The immune system apparently exacerbates the injury caused by the infarction since microglia become activated, express MHC molecules, and produce proinflammatory cytokines such as TNF-α and IL-1β. Activated microglia can recruit inflammatory leukocytes into the brain and might be able to present antigens and further activate CNS-specific inflammatory T cells, initializing an inflammatory cascade and leading to postischemic tissue destruction. This role of inflammation in stroke-related tissue destruction is supported by the fact that blockade of IL-1 or the death ligand TRAIL reduces brain damage in ischemia and by the beneficial effect of myelin-specific IL-10-producing Th cells in reducing stroke size.

AD is the most common neurodegenerative disorder, affecting approximately 13 million people worldwide and representing the major cause of dementia. Neurodegeneration in AD affects primarily hippocampal and neocortical brain regions, resulting in a progressive loss of cognitive and memory functions, finally leading to dementia. The neurodegenerative process is associated with deposits of β-amyloid (Aβ) in the brain as a consequence of a metabolic dysregulation of the Aβ precursor protein (βAPP). Cleavage of βAPP by the β and γ secretases releases both Aβ1–40 and Aβ1–42 peptide. The ratio between the two forms, as well as their capacity to precipitate, is determined by genetic factors such as gene mutations in the APP and presenilin genes or the presence of the ApoE4 allele. The fibrillar plaques formed by Aβ deposits are associated with dystrophic neurites as well as activated microglia (expressing high levels of MHC class II molecules) and astrocytes. Aβ-treated microglia express increased levels of proinflammatory cytokines and chemokines, including IL-1β, TNF-α, IL-12, IL-8, and CCL3; might be able to process and present peptides; and are capable of chemoattracting and activating Aβ-specific T cells. Nevertheless, apparently only few activated T cells are found in AD brains, and it is controversial whether and how these cells may contribute to disease pathogenesis. Although AD development seems to be independent of T cell-mediated inflammation, a large body of evidence indicates that disease development can be modulated by inducing an appropriated systemic T cell response to Aβ42 by vaccination. In an animal model, an Aβ42 vaccine prevented peptide deposits, induced clearance of the plaques, and improved cognitive deficits. This beneficial effect was triggered by antibody-mediated microglia activation. However, a placebo-controlled clinical trial designed to test the vaccine Aβ42 in patients had to be interrupted when aseptic meningoencephalitis was identified in 6% ($n = 18$) of the vaccinated patients. There is evidence that this severe adverse reaction was probably mediated by T cell-dominated immune reaction to Aβ and the consequent IFN-γ-induced activation of microglia. Despite the disappointing results of the clinical trial, this therapeutic strategy remains under intense investigation. Researchers are convinced

that the key question in AD vaccination is to find the mode of modulating the immune response to Aβ in a way that results in both suppression of Th1-dominated immunity and increment of Th2- and antibody-mediated response.

Perspectives

Although major advances in the understanding of immune cell interactions have occurred in the past few years, particularly regarding T cells and glial cells, we still have to contend with many conflicting and, in part, controversial data. Most of these data have resulted from the *ex vivo* evaluation of CNS tissue or have been obtained using *in vitro* models, and these methods may not reliably reflect the cell interactions taking place under *in vivo* conditions. This problem may be solved by molecular imaging, a newly emerging, methodology that combines tools from both molecular and cell biology for the purpose of imaging biological processes in living organisms. The field of molecular imaging encompasses sophisticated technologies, such as magnetic resonance imaging, positron emission tomography, single-photon emission computed tomography, two-photon microscopy, and bioluminescence or fluorescence. The further development and use of these novel methods will doubtless shed new light on molecular and cellular interactions in health and disease – knowledge that is indispensable for improved monitoring of disease, accurate diagnosis, and the development of novel therapeutic strategies.

See also: Alzheimer's Disease: Molecular Genetics; Astrocyte: Neurotransmitter and Hormone Receptors; Astrocyte: Identification Methods; Demyelinating Diseases; Encephalitis; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Glial Cells: Microglia During Normal Brain Aging; Inflammation in Neurodegenerative Disease and Injury; Microglia Properties; Microglia identification Methods; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neurotoxins and their Neurotoxicology; Oligodendrocyte Morphology.

Further Reading

Aloisi F (2001) Immune function of microglia. *Glia* 36: 165–179.

Becher B, Prat A, and Antel JP (2000) Brain–immune connection: Immuno-regulatory properties of CNS-resident cells. *Glia* 29: 293–304.

Blasko I, Stampfer-Kountchev M, Robatscher P, et al. (2004) How chronic inflammation can affect the brain and support the development of Alzheimer's disease in old age: The role of microglia and astrocytes. *Aging Cell* 3: 169–176.

Carson MJ (2002) Microglia as liaisons between the immune and central nervous systems: Functional implications for multiple sclerosis. *Glia* 40: 218–231.

Danton GH and Dietrich WD (2003) Inflammatory mechanisms after ischemia and stroke. *Journal of Neuropathology and Experimental Neurology* 62: 127–136.

De Keyser J, Zeinstra E, and Frohman E (2003) Are astrocytes central players in the pathophysiology of multiple sclerosis? *Archives of Neurology* 60: 132–136.

Dong Y and Benveniste EN (2001) Immune function of astrocytes. *Glia* 36: 180–190.

Frohman EM, Racke MK, and Raine CS (2006) Multiple sclerosis – the plaque and its pathogenesis. *New England Journal of Medicine* 354: 942–955.

Gonzalez-Scarano F and Martin-Garcia J (2005) The neuropathogenesis of AIDS. *Nature Reviews Immunology* 5: 69–81.

Jack C, Ruffini F, Bar-Or A, and Antel JP (2005) Microglia and multiple sclerosis. *Journal of Neuroscience Research* 81: 363–373.

Minagar A, Shapshak P, Fujimura R, et al. (2002) The role of macrophage/microglia and astrocytes in the pathogenesis of three neurologic disorders: HIV-associated dementia, Alzheimer disease, and multiple sclerosis. *Journal of Neurological Sciences* 202: 13–23.

Monsonego A and Weiner HL (2003) Immunotherapeutic approaches to Alzheimer's disease. *Science* 302: 834–838.

Moser HW, Raymond GV, and Dubey P (2005) Adrenoleukodystrophy: New approaches to a neurodegenerative disease. *JAMA* 294: 3131–3134.

Reinhardt RL, Kang SJ, Liang HE, and Locksley RM (2006) T helper cell effector fates – who, how and where? *Current Opinion in Immunology* 18: 271–277.

Zipp F and Aktas O (2006) The brain as a target of inflammatory processes: Common pathways link inflammatory and neurodegenerative diseases. *Trends in Neurosciences* 29: 518–527.

Glial Energy Metabolism: A NMR Spectroscopy Perspective

R Gruetter, Ecole Polytechnique Fédérale de Lausanne, University of Lausanne, Lausanne-Geneva, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During increased electrical brain activity, blood flow increases, tissue deoxyhemoglobin content decreases, glucose metabolism increases, and possibly oxygen metabolism increases. These changes in metabolism and the hemodynamic consequences are measured with contemporary brain imaging methods. However, the measurement of brain activity, be it through the hemodynamic correlates of increased electrical activity, increased oxygen consumption, or deoxyglucose uptake, is unable to depict the fundamentals of brain activity at the neuroglial level *in vivo*. The major mechanism of deactivation of the major excitatory neurotransmitter, glutamate, is glial uptake and conversion to electrophysiologically inactive glutamine, which is then transported back to the neuron to replenish the neurotransmitter pool of glutamate, completing the glutamate–glutamine cycle. It is the purpose of this article to summarize perspectives on *in vivo* metabolism in the glial compartment, relative to that in the neuronal compartment, that have been gained using an emerging method, namely nuclear magnetic resonance (NMR) spectroscopy. Most measurements are performed at metabolic steady state, defined as a condition of metabolic fluxes being constant.

NMR Spectroscopy of Brain Energy Metabolism: A Brief Primer

NMR spectroscopy of brain energy metabolism has focused chiefly on the signals from three NMR-active nuclei – the ubiquitously used ^{1}H, as well as ^{31}P and ^{13}C. Of the three, ^{1}H and ^{31}P are 100% naturally abundant, whereas ^{13}C is only 1.1% abundant in nature. Techniques to establish the quantitative measurement of energy metabolism depend critically on the specific signal measured from a given nucleus. In the following discussions, the current state-of-the-art measurements are given, without elaborating on methodological considerations in detail.

Proton NMR Spectroscopy

With suppression of the intense signal of water, which is used for magnetic resonance imaging (MRI), and elimination of the extracerebral fat signal by adequate localization, a multitude of resonances can be observed in the *in vivo* ^{1}H NMR spectrum; these are attributed to date to more than 20 compounds that can be quantitated with typical precision of ~10%, amounting to a neurochemical profile (Figure 1). Quantification is typically achieved by referencing the metabolite signal to an internal reference signal of known concentration, such as water. Metabolic rates are typically either inferred from measured metabolite concentrations or from their changes with time.

Phosphorus NMR Spectroscopy

The *in vivo* signal of the ^{31}P nucleus contains signals from the three phosphates of ATP – signals from inorganic phosphate and phosphocreatine, as well as signals attributed to phosphomonoester compounds, such as phosphoethanolamine and phosphocholine, and signals attributed to phosphodiesters, such as glycerophosphocholine and ethanolamine. The signals of ADP and UDP-glucose are in principle observable but are typically too low to be detected. The chemical shift of inorganic phosphate reveals intracellular pH, and by manipulating the magnetization of inorganic phosphate, phosphocreatine, or the γ-phosphate of ATP through magnetization transfer, the rates of the reactions of creatine kinase or ATPase can be assessed. Quantification is typically achieved by referencing relative to a brain ATP concentration of 3 mM.

Carbon-13 NMR Spectroscopy

The sensitivity of ^{13}C NMR spectroscopy can be improved by administering ^{13}C-labeled precursors, such as glucose, lactate, β-hydroxybutyrate, or acetate. Further sensitivity improvements are possible by indirectly detecting the ^{13}C label through vicinal ^{1}H nuclei. As in radiotracer methods, such as positron emission tomography (PET), the measurement of label incorporation into metabolic products allows the determination of metabolic rates using similar mathematical approaches. Energy metabolism has been assessed from the measurement of brain glucose, glycogen, lactate, glutamate, glutamine, and aspartate (Figure 2). Quantification is typically achieved by performing an external reference measurement.

Glycolysis

Glucose Transport

The brain relies on a continuous import of glucose from the blood across the blood–brain barrier. Glucose transport rates into the brain are thus indicative of the maximal sustainable rate of glucose consumption,

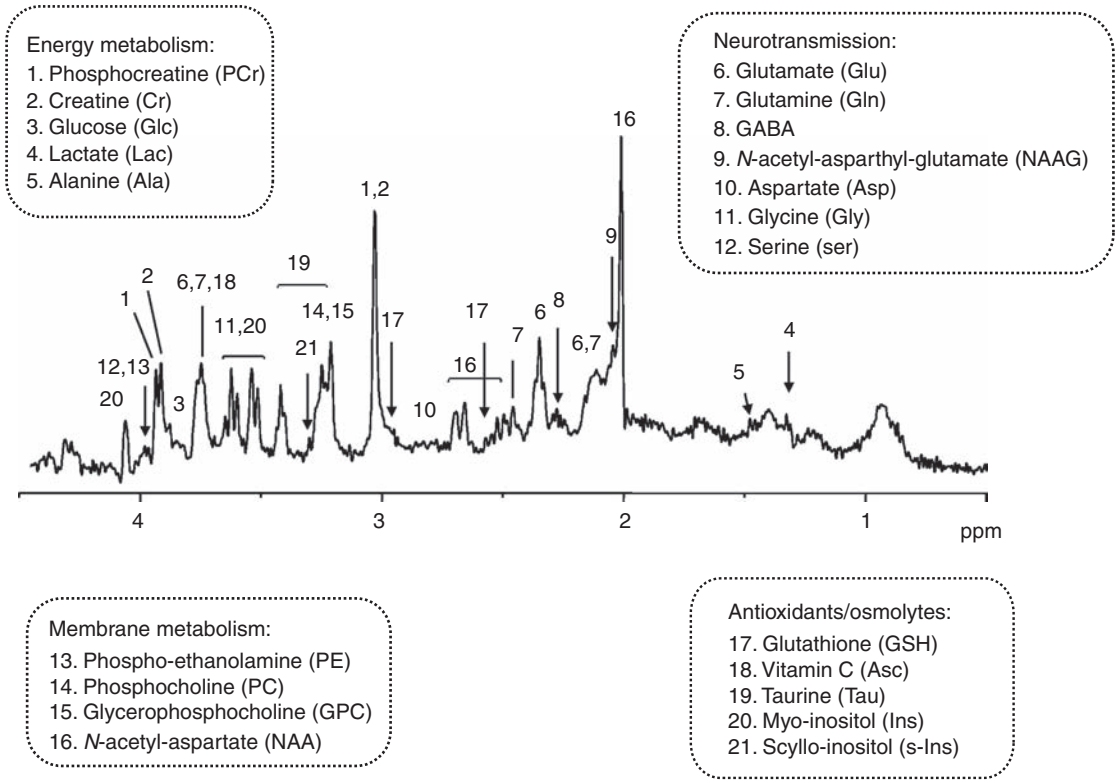

Figure 1 A neurochemical profile can be derived from the proton nuclear magnetic resonance spectrum of the brain (TE = 2.7 ms, 9.4 T). Shown are 21 compounds that can be measured *in vivo*, grouped according to their major significance. In addition, when present, acetate and β-hydroxybutyrate can also be measured.

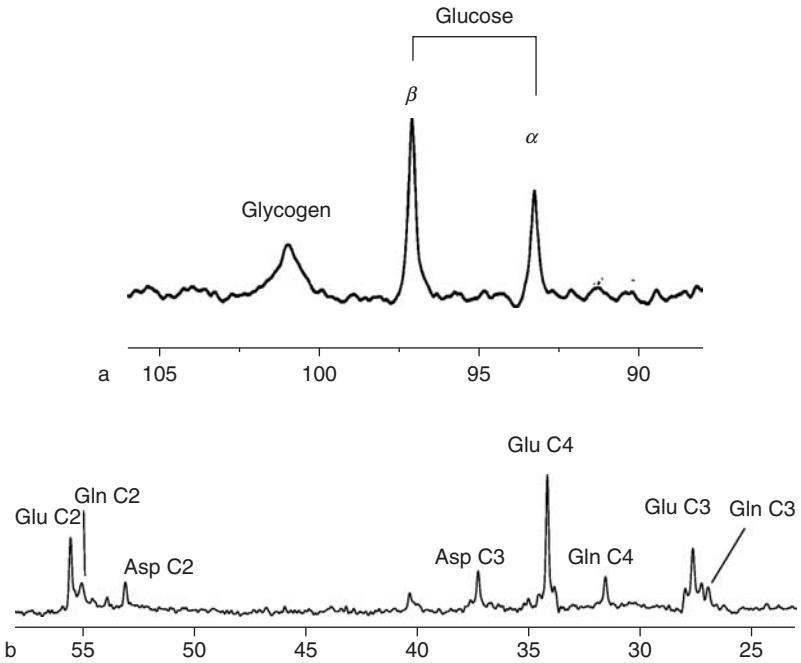

Figure 2 (a) Localized ^{13}C nuclear magnetic resonance spectroscopy of the C1 of glucose and glycogen in rat brain (9.4 T). (b) Localized ^{13}C nuclear magnetic resonance spectroscopy of the human brain (4 T). Spectra in both panels are shown after [1–^{13}C]glucose administration. Label incorporation is observed in the predominantly neuronal glutamate and aspartate and the predominantly glial glutamine, in multiple positions due to label scrambling in the tricarboxcylic acid cycle, and exchange with cytosolic amino acids due to, for example, the action of the malate–aspartate shuttle.

CMR_{glc}. Knowledge of the size of the brain glucose pool is important for the derivation of absolute metabolic fluxes and quantitative positron emission tomography studies. Steady-state glucose transport kinetics can be derived from the relationship between brain and plasma glucose. The former can be directly measured using NMR spectroscopy.

Brain glucose is distributed in the entire aqueous phase of the brain The volume in the brain into which glucose diffuses (i.e., the physical distribution volume, V_d) is critical in the derivation of glucose transport kinetics. The diffusion behavior of glucose, measured by diffusion-weighted NMR spectroscopy, reflects that ~80% of the glucose in brain tissue is intracellular, consistent with a high physical distribution volume for glucose, $V_d = 80\%$, which is approximately equal to the brain's water phase. Intracellular and extracellular glucose concentrations are approximately equal, suggesting that transport of glucose across the cell membranes is fast compared to the glucose consumption rate. Conversely, the limiting step for glucose entry into the brain is at the blood–brain barrier (BBB), comprising the endothelial cells.

Glucose content reflects the balance between transport and consumption The steady-state brain glucose concentration is a sensitive indicator of the ability of the BBB to transport glucose relative to the rate of glucose consumption, T_{max}/CMR_{glc}. The observed linear relationship between plasma and brain glucose is consistent with the need to take into account the reversibility of the transport process (Figure 3). The apparent K_m of glucose transport, K_t, is of the order of 5 mM. Decreased electrical activity and thus decreased energy metabolism result in increased brain glucose concentrations; yet a sizable concentration gradient is maintained at isoelectricity, implicating considerable net glucose uptake and thus substantial energy metabolism not related to signaling (Figure 4(a)). Increases in luminal glucose transporter (GLUT-1) result in an increase in the maximal transport capacity of the BBB, the consequence of which is an accordingly increased brain glucose content, suggesting that the luminal membrane is the major rate-determining step for glucose entry into the brain (Figure 4(b)).

Brain glucose becomes rate limiting for metabolism when its concentration approaches that of the K_m of the first step in its metabolism, which in the brain is

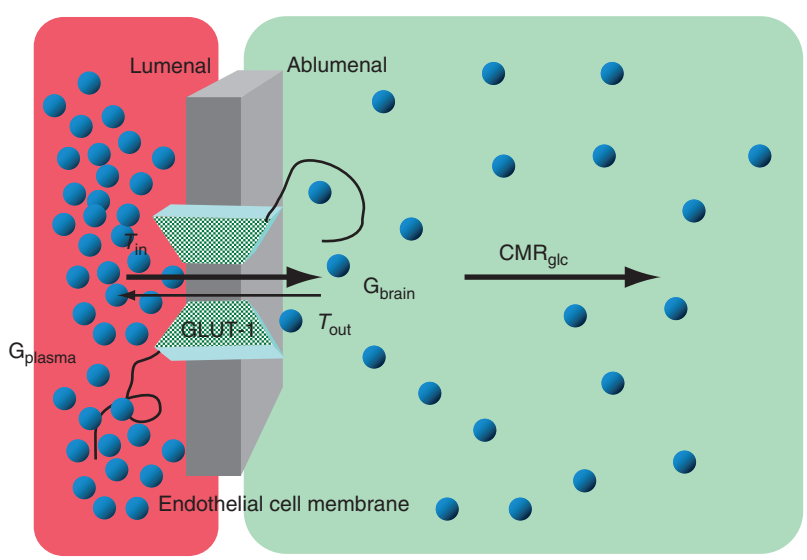

$$T_{in} = T_{out} + CMR_{glc}$$

$$T_{in} = T_{max} \frac{G_{plasma}}{(K_t + G_{brain}/V_d) + G_{plasma}} \qquad T_{out} = T_{max} \frac{G_{brain}}{(K_t + G_{plasma})V_d + G_{brain}}$$

Figure 3 Schematic of glucose (G) transport across the blood–brain barrier, depicting the steady-state situation and showing the reversible Michaelis–Menten kinetic expression for transport (CMR_{glc}, maximal sustainable rate of glucose consumption). Standard Michaelis–Menten kinetics neglect the effect of product concentration on reaction velocity (V), and expressions for the transport rates (T_{in} and T_{out}) are derived by assuming zero product concentration (i.e., by setting $G_{brain} = 0$ for T_{in} and $G_{plasma} = 0$ for T_{out}).

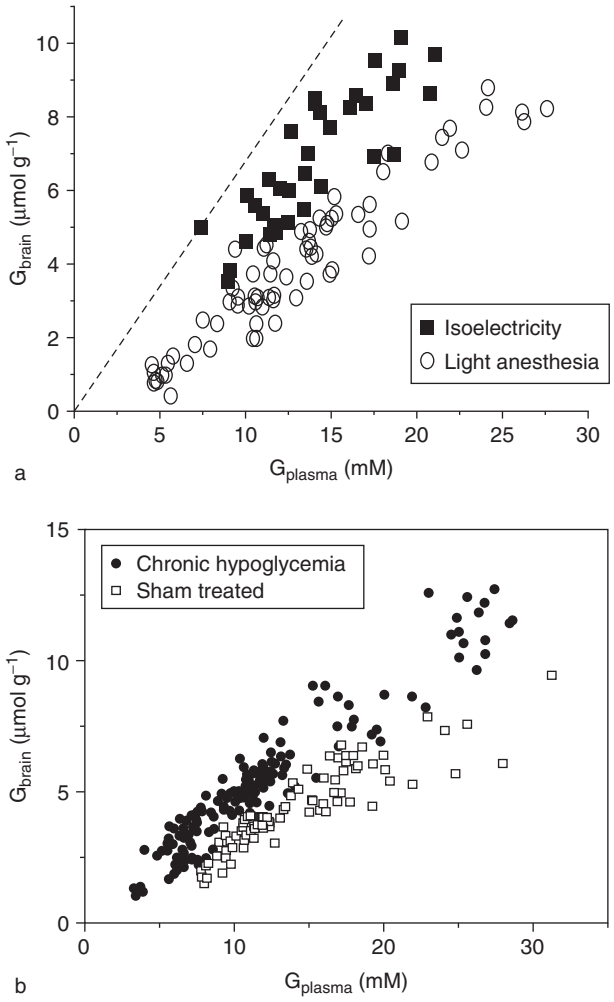

Figure 4 (a) Effect of reduced brain activity on brain glucose concentration in the rat brain. Reduction of electrical activity increases brain glucose content substantially, but a considerable gradient across the blood–brain barrier is maintained (the dashed line indicates the G_{brain}:G_{plasma} relationship in the absence of a gradient). (b) Effect of 50% increased blood–brain barrier glucose transporter GLUT-1 on brain glucose content. GLUT-1 was increased by inducing chronic hypoglycemia, which led to a substantial increase in brain glucose content at a constant brain glucose consumption rate.

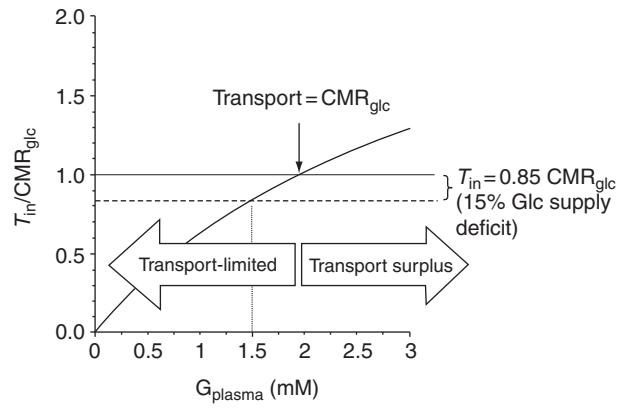

Figure 5 Glucose transport can become rate limiting for metabolism, but this is a gradually increasing deficit. The graph shows the unidirectional glucose influx rate, T_{in}, relative to the glucose consumption rate, CMR_{glc}. In this example glucose influx matches the consumption at $G_{plasma} \sim 2$ mM. As a result, even at substantial hypoglycemia (e.g., $G_{plasma} = 1.5$ mM), the glucose supply (i.e., T_{in}) accounts for ~85% of total glucose requirements.

mostly phosphorylation by hexokinase. Since the K_m of brain hexokinase is very low (~50 μM), brain glucose concentrations that are close to zero indicate metabolism that is limited by the glucose available to the brain cell (transport-limited metabolism). The substantial brain glucose content at euglycemia indicates that brain glucose transport is not rate limiting for metabolism under normal circumstances; the maximal sustainable rate of glucose consumption is approximately two to three times above the basal rate of glucose metabolism. Glucose consumption can, however, transiently exceed the maximum unidirectional transport into the brain, which may occur in conditions of extreme metabolic activation, such as during seizures. While plasma glucose drops, the brain traverses from excess glucose supply to a condition in which transport across the BBB supplies just enough glucose for metabolism, and eventually to a gradually increasing deficit in glucose supply (Figure 5).

During hypoglycemia, glucose metabolism tends to decrease and cerebral blood flow (CBF) increases when the brain glucose concentration approaches the K_m of hexokinase. The mechanism of such blood flow regulation is different from the neurovascular coupling in physiological activation, whereby glucose metabolism and CBF increase concomitantly.

Glial Glycogen – The Only Energy Reserve in Brain

Glycogen, ascribed to the glial compartment, is present in concentrations that exceed those of euglycemic brain glucose by severalfold. It thus comprises by far the most concentrated endogenous storage of glucose equivalents in the central nervous system (CNS), and the presence of glycogen stores is essential for brain function. Similar to glucose, brain glycogen is rapidly eliminated postmortem; therefore, its direct biochemical measurement is difficult. Localized ^{13}C NMR spectroscopy has the unique capability of following brain glycogen metabolism longitudinally, allowing measurement of the rate of change of metabolic rates and concentration changes of glycogen.

Brain glucose can regulate glycogen metabolism Glycogen metabolism in brain accounts for only a few percent of total glucose metabolism under resting conditions. In contrast to transport and metabolism of bulk brain glucose, glycogen metabolism is strongly insulin dependent and to a lesser degree its content

depends on the plasma glucose concentration. In deep anesthesia and hyperglycemia – conditions which are associated with increased brain glucose content – brain glycogen levels rise well above the normal concentration. Together with the apparent stability of glycogen in the nonstimulated brain at euglycemia or hyperglycemia, these data suggest that brain glucose plays an important regulatory role in cerebral glycogenolysis.

When the intracellular glucose concentration approaches the K_m of hexokinase, glycogen phosphorylase is activated and glycogen synthase is deactivated, resulting in net glycogenolysis to glucose-6-phosphate (and some free glucose, due to cleavage of the 1–6 bonds in glycogen by the debranching enzyme). Glucose-6-phosphate is eventually metabolized to pyruvate, which can either enter the tricarboxcylic acid (TCA) cycle or be converted by lactate dehydrogenase to lactate. Glial lactate can be exported via the action of monocarboxylate transporters at the glial and neuronal cell membranes.

Glycogen stores are neuroprotective By providing lactate to the extracellular fluid, astrocytes can sustain neuronal energy metabolism of neighboring neurons in conditions of fuel deprivation, such as hypoglycemia. During glycogenolysis, two ATPs per glucosyl unit are generated in glycolysis, which can provide energy to support uptake of synaptic glutamate and conversion into electrophysiologically inactive glutamine. Astrocytes thus can remove excitotoxic glutamate from the synaptic cleft in energy-deprived situations, which may be one mechanism by which glial glycogen exerts its neuroprotective effect (Figure 6).

Following glycogen depletion, be it induced by neurotransmitters, by hypoglycemia, by hypoxia, or by sleep deprivation, brain glycogen increases above the basal level and beyond in the presence of ample insulin and glucose, resulting in 'supercompensation' similar to that observed for glycogen in noncerebral tissues. Increased astrocytic glycogen levels reduce neuronal damage following subsequent fuel deprivation. Brain glycogen metabolism may be a factor involved in the mechanism of the hypoglycemia unawareness syndrome observed clinically in patients with insulin-treated diabetes, probably through the enhanced neuroprotective effect of increased brain glycogen.

Glycogen is an important energy reserve Glycogen serves as an energy reservoir of glucose equivalents.

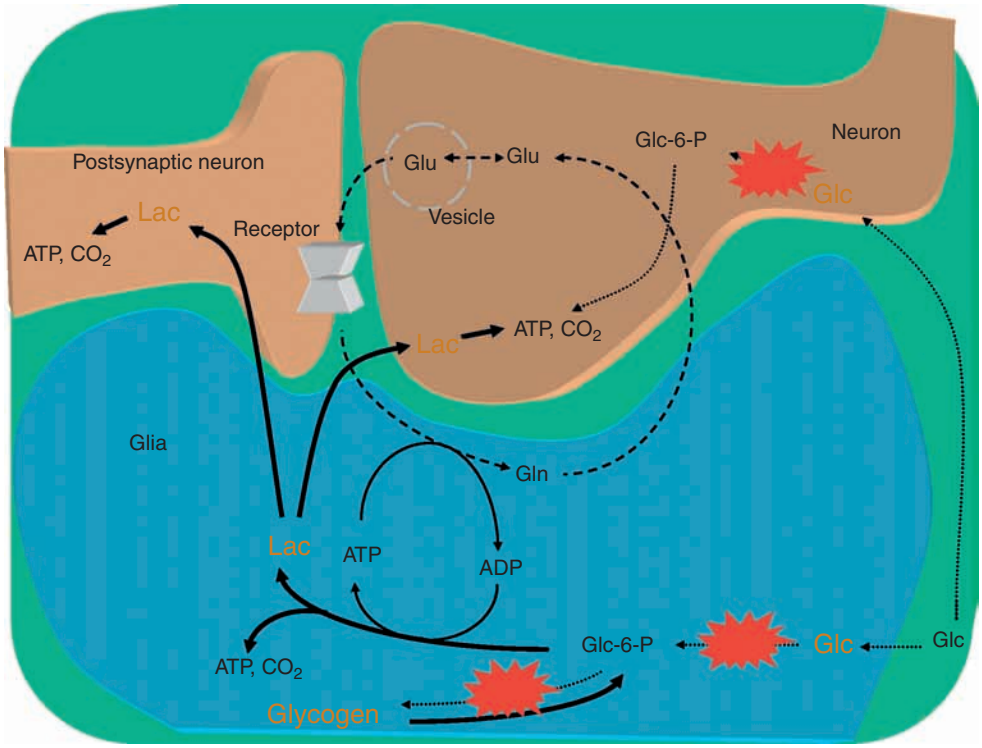

Figure 6 Schematic showing the neuroprotective role of glial glycogen and its ability to nourish neighboring neurons with substrates, such as lactate (Lac). In glucoprivation, such as hypoglycemia, glucose phosphorylation impaired, leading to insufficient glucose-6-phosphate (Glc-6-P) generation, deactivated glycogen synthase (red stars), and activated glycogen phosphorylase. The ensuing glycogen breakdown, through its metabolic pathways (solid thick lines), ultimately produces fuel for neighboring neurons. Glial glycogen is probably also neuroprotective through its ability to minimize excitotoxicity by providing energy for removal of glutamate from the synaptic cleft in fuel-deprived conditions.

For example, during aglycemia, when the supply of glucose to the brain cells is completely abolished, glycogen would sustain total glucose metabolism for about 20 min. However, aglycemia is difficult to achieve in the living brain, and the reduction of plasma glucose in reality is gradual in time, resulting in a gradually increasing glucose supply deficit (Figure 5). For example, in order to cover a partial supply deficit of the metabolic rate of glucose of 0.05 μmol g^{-1} min^{-1} in the awake human brain (corresponding to a deficit of 14% of a basal metabolic rate of 0.35 μmol glucose g^{-1} min^{-1}), the cerebral content of glycogen of 3.5 μmol glucosyl units g^{-1} can account for this fuel deficit for more than 1 h. The rate of brain glycogen degradation during hypoglycemia accounts for the majority of the glucose supply deficit during the hypoglycemic period, and following extended periods of moderate hypoglycemia, substantial glycogen reserves remain. The time estimated for complete glycogen depletion is 4 h, which corresponds to the time period that clinically results in residual neurological symptoms after an episode of hypoglycemia.

In chronic hypoglycemia, brain glucose content is elevated (Figure 4(b)), resulting in ambient intracellular glucose content that appears normal, despite the prevalent hypoglycemia. In this condition, consistent with the upregulation of brain glucose content, astrocytic 45 kDa GLUT-1 and neuronal GLUT-3 are unaltered, as are brain glycogen concentrations.

Glycogen is an important store of glucose equivalents in the brain, having a significant neuroprotective effect; glycogen metabolism is affected not only by tissue glucose levels and plasma insulin, but also by other hormones, neurotransmitters, and second messengers.

Lactate Metabolism – An Integral Part of Normal Brain Function

Brain lactate is present at near 1 μmol g^{-1} concentration under normoxic conditions *in vivo*, and is continuously synthesized and broken down. Glucose and glycogen are the major metabolic precursors for lactate, which can be exported and serve as fuel for other cells. A small lactate efflux from the brain is observed normally. Rapid but small increases in brain lactate are observed in focal activation, consistent with a transient uncoupling of oxygen and glucose metabolism. Similar to that for glucose, lactate diffusion behavior is restricted, indicating that most tissue lactate is intracellular. Lactate is almost evenly distributed in the brain's aqueous phase, owing to the large transport capacity of the abundant monocarboxylate transporters at the neuronal and glial cell membranes.

Tricarboxylic Acid Cycles

Glutamate Turnover – How Oxygen Metabolism Is Measured by NMR Spectroscopy

When using a substrate labeled at a suitable position, be it glucose or other suitable substrates, the measurement of oxygen metabolism by NMR spectroscopy can be accomplished by measuring the rate of labeling of glutamate. For simplicity we consider glucose labeled at the 1 and 6 positions, which labels the C3 of pyruvate and ultimately the C2 of acetyl-CoA, which condenses with oxaloacetate to form citrate. In the course of the metabolic reactions of the TCA cycle, which occur at a rate denominated as V_{TCA}, label arrives at the C4 of 2-oxoglutarate. Oxoglutarate is exchanged with glutamate through transamination and thus readily labels mitochondrial glutamate. However, in the brain *in vivo*, most of the tissue glutamate is of cytosolic origin, given the small volume fraction occupied by mitochondria. Therefore, when label incorporation into glutamate is observed, it reflects a signal ascribed mainly to the cytosol (Figure 7(a)). The label scrambling, however, occurs in the mitochondria. Carbon-13 label is transferred to the cytosolic glutamate by label exchange across the inner mitochondrial membrane. Processes such as the malate–aspartate shuttle (Figure 7(b)), which transfer reducing equivalents from the cytosol to the mitochondrion, play a key role in label exchange across the mitochondrial membrane. Therefore, the rate of labeling of glutamate is both a function of the TCA cycle rate (V_{TCA}) as well as of isotope exchange across the inner mitochondrial membrane (V_x).

When considering the further fate of label in the C4 of glutamate (and conversely the C4 of 2-oxoglutarate), the investigational power of NMR spectroscopy can be demonstrated. The label at C4 of 2-oxoglutarate proceeds along the TCA cycle to scramble equally between the C2 and C3 of succinate (a symmetric molecule), which ultimately leads to a symmetric labeling of the C2 and C3 of 2-oxoglutarate and thus by the same exchange mechanism (V_x in Figure 7(a)) the C2 and C3 of glutamate. Hence, as a result of metabolic reactions ascribed to the TCA cycle, label from a singly labeled acetyl-CoA molecule results in labeling of glutamate in three positions in two turns of the TCA cycle. Whereas the labeling of the C4 of glutamate takes up a simple analytical form of an exponential function, the rate of which is the geometric average of V_{TCA} and V_x divided by the glutamate pool size, the labeling of the C2 and C3 of glutamate is more complex. Briefly, at the C3 position, glutamate receives label not only from acetyl-CoA, but since some of the label escapes the oxoglutarate pool into glutamate (with a probability determined by the relative rate of V_{TCA} and V_x), its C4 position increasingly serves as

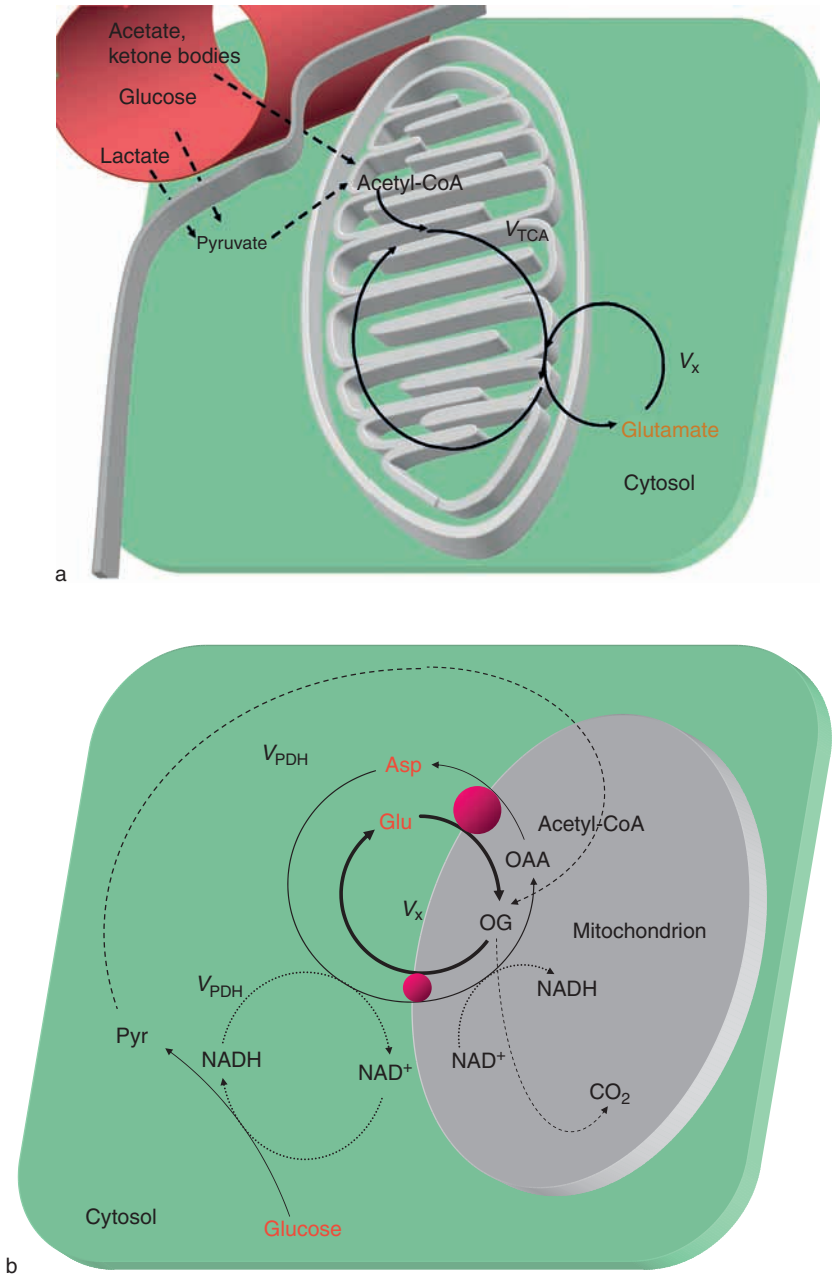

Figure 7 The flow of label from glucose to glutamate: (a) illustration of the flow of label from acetyl-CoA to glutamate; (b) the metabolic reactions involved in the malate–aspartate shuttle. The malate–aspartate (Asp) shuttle allows the cell to combust glucose oxidatively, by transferring the reducing equivalent generated per pyruvate (Pyr) molecule oxidized to the mitochondrion. This shuttle is critical in permitting respiring neural cells to metabolize glucose. When pyruvate is the sole fuel for the tricarboxcylic acid (TCA) cycle, the TCA cycle rate $V_{TCA} = V_{PDH}$ (PDH, pyruvate dehydrogenase; V_x, isotope exchange across the inner mitochondrial membrane; OG, oxoglutarate; OAA, oxaloacetic acid).

label substrate for the C3. Since the rate of labeling of Glu_4 is delayed compared to that of acetyl-CoA, due to the time required to completely label the glutamate pool, the labeling of Glu_3 is an increasingly concave function when increasing transmitochondrial exchange rate V_x, which thus can be assessed, allowing important insights into the regulation of oxidative metabolism and mitochondrial integrity *in vivo*. The labeling of glutamate by itself does not permit to study compartmentalization between astrocytes and neurons, however, as glutamate is ubiquitously present in many cells, including astrocytes.

The Glutamate–Glutamine Cycle

Glial-specific tracers Acetate and glucose do not share the same product-precursor relationship.

Today it is well established that acetate is selectively taken up by astrocytes, and hence labeled acetate can be used to 'trace' glial metabolism, whether it is by using radiotracers, as in ^{14}C autoradiography or ^{11}C (using PET), or by using [^{13}C]acetate in conjunction with NMR spectroscopy. The advantage of using acetate as a tracer lies in its exclusive uptake into the astrocytes, and thus from the turnover of glutamate, astrocytic TCA cycle flux can be assessed.

Labeled acetate will thus label glial glutamate much in the same manner as discussed previously; however, the action of glutamine synthetase and the diffusion of glutamine from the astrocyte into the presynaptic terminal result in label being scrambled from glutamine to neuronal glutamate, where it is being diluted with unlabeled carbon derived from the neuronal TCA cycle; this leads to a lower isotopic enrichment of tissue glutamate, compared to tissue glutamine (resulting in an anomalous product-precursor relationship). In general, glutamate and glutamine are not labeled to an equal extent, reflecting the substantial activity of the glial TCA cycle (see later).

When electrical activity ceases (isoelectricity) the glutamate–glutamine cycle activity is close to zero, but substantial energy is still expended, powering ion gradients, among others. This results in a substantial glucose concentration gradient at isoelectricity and active oxygen consumption (Figure 4(a)). With increasing electrical activity, the rate of glutamine synthesis increases but the glutamate–glutamine cycle does not operate stoichiometrically.

Glutamine Synthetase and Pyruvate Carboxylase

The synthesis of glutamine is catalyzed by glutamine synthetase (EC 6.3.1.2), which is attributed to the glial compartment (Figure 8). Hence the *in vivo* observation of glutamine labeling by ^{13}C NMR is a direct manifestation of glial metabolism. In principle, the simultaneous observation of labeling of glutamate, mainly ascribed to the neuronal compartment, and glutamine, a glial reaction, would thus follow the path of acetyl-CoA → glutamate → glutamine → glutamate, and thus would represent a direct manifestation of the glutamate–glutamine cycle, which is the dominant mechanism of glutamate neurotransmission. However, additional labeling of glutamine occurs via the glial TCA cycle, as astrocytes contain mitochondria and thus the necessary enzymatic predisposition for oxidative generation of energy (e.g., a high cytochrome oxidase content).

The presence of significant glial oxidative metabolism results in incorporation into glutamine label that does not match that of glutamate. Reactions associated with CO_2 fixation, such as the major anaplerotic enzyme, pyruvate carboxylase (EC 6.4.1.1), have been localized to astrocytes. *De novo* synthesis of

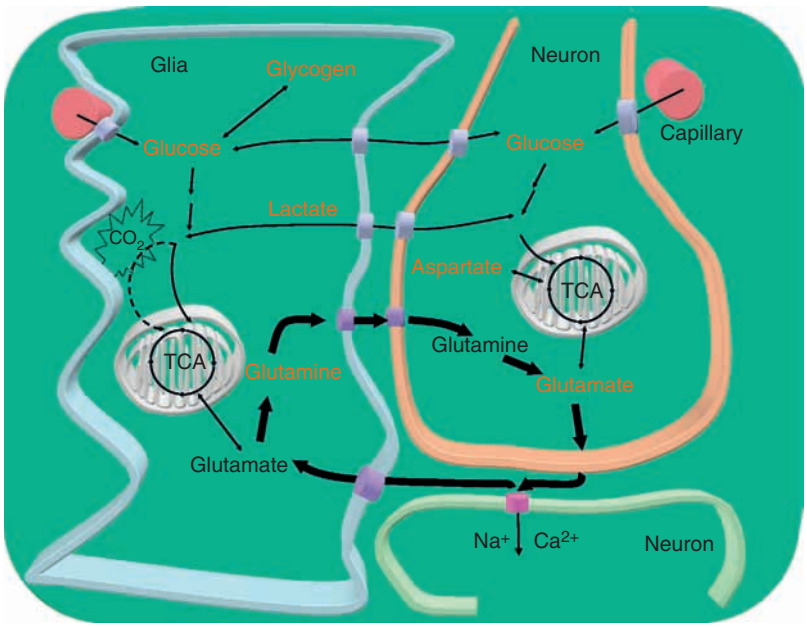

Figure 8 Schematic depiction of the metabolic relationship between neurons and glia. The compounds in red are those measured by nuclear magnetic resonance spectroscopy *in vivo*. Deactivation of neurotransmitter glutamate is primarily by uptake into the glia and conversion to glutamine, which at least in part diffuses back into the presynaptic terminal, to be converted to glutamate by phosphate-activated glutaminase, completing the glutamate–glutamine cycle (thick arrows; TCA, tricarboxcylic acid).

amino acids is thus a unique capability of the glial compartment, given that transamination which occurs in both astrocytes and neurons by itself cannot provide a net increase in tissue TCA cycle intermediate and hence amino acid concentration. In the reaction catalyzed by pyruvate carboxylase, pyruvate is combined with CO_2 to form oxaloacetate (OAA). As previously, we consider glucose labeled at the 1 and 6 positions as the precursor molecule, which results in label incorporation in the C3 of oxaloacetate from the C3 of pyruvate, which in turn labels the C2 of oxoglutarate, and thus the C2 of glutamate. As a result, the C2 of glutamate is labeled to a larger extent than is the C3. By comparing the labeling of glutamate in C2 and C3, the quantitative presence of anaplerotic reactions can be estimated *in vivo* by NMR, although some equilibrating flux back to fumarate may obscure the true extent of CO_2 fixation present. Anaplerotic reactions are a substantial metabolic activity in the brain, amounting to approximately one-third of glutamine synthesis, and are linked to brain activity. Anaplerotic reactions are decreased in metabolically depressed states, resulting in reduced glutamine synthesis and lowered brain glutamate and glutamine concentrations in hibernation, for example. Net synthesis of glutamine from glucose produces approximately ten ATPs per glutamine molecule, from the generation of several NADH molecules in the process, and thus is energetically favorable.

Summary

At rest, most of the energy requirements in the brain are matched by oxidative metabolism of glucose. Oxygen metabolism is uniquely compartmentalized between neurons and glia, the latter comprising about 25% of the energy metabolism. The consumption of acetate and fixation of CO_2 in glutamine are oxygen-dependent glial reactions. At least two-thirds of the ATP generated in astrocytes is by oxidative processes. Astrocytes thus can, in a unique way, buffer metabolic demand of the tripartite synapse. Glia exhibit a large capacity to phosphorylate glucose and can store it in the form of glycogen, the glucose equivalents of which can be exported in the form of lactate to neurons, providing neuroprotection. Astrocytes have a substantial anaplerotic activity, as *de novo* synthesis of glutamine by itself is an energetically favorable process. Glia therefore adapt the fuel demands to the specific requirements encountered in activation, hypoglycemia, and deactivation, nurturing surrounding neurons with substrates, while at the same time eliminating neurotransmitter glutamate from the synaptic cleft.

See also: Activity-Dependent Metabolism in Glia and Neurons; Glial Energy Metabolism: Overview; Glial Glycogen Metabolism; Glial Glutamate and GABA Metabolism; Glial Plasticity and Neuroendocrine Regulation; Glutamatergic and Gabaergic Systems; Magnetic Resonance Spectroscopy; Perfusion MRI.

Further Reading

Bouzier-Sore AK, Merle M, Magistretti PJ, et al. (2002) Feeding active neurons: (Re)emergence of a nursing role for astrocytes. *Journal of Physiology Paris* 96: 273–282.

Brown AM (2004) Brain glycogen re-awakened. *Journal of Neurochemistry* 89: 537–552.

Cerdan S, Rodrigues TB, Sierra A, et al. (2006) The redox switch/redox coupling hypothesis. *Neurochemistry International* 48: 523–530.

Gruetter R, Adriany G, Choi IY, et al. (2003) Localized *in vivo* ^{13}C NMR spectroscopy of the brain. *NMR in Biomedicine* 16: 313–338.

Gruetter R, Seaquist ER, and Ugurbil K (2001) A mathematical model of compartmentalized neurotransmitter metabolism in the human brain. *American Journal of Physiology (Endocrinology and Metabolism)* 281: E100–E112.

Hyder F, Patel AB, Gjedde A, et al. (2006) Neuronal–glial glucose oxidation and glutamatergic–GABAergic function. *Journal of Cerebral Blood Flow and Metabolism* 26: 865–877.

McKenna MC, Waagepetersen HS, Schousboe A, et al. (2006) Neuronal and astrocytic shuttle mechanisms for cytosolic-mitochondrial transfer of reducing equivalents: Current evidence and pharmacological tools. *Biochemical Pharmacology* 71: 399–407.

Morris P and Bachelard H (2003) Reflections on the application of ^{13}C-MRS to research on brain metabolism. *NMR in Biomedicine* 16: 303–312.

Oz G, Berkich DA, Henry PG, et al. (2004) Neuroglial metabolism in the awake rat brain: CO_2 fixation increases with brain activity. *Journal of Neuroscience* 24: 11273–11279.

Waagepetersen HS, Sonnewald U, and Schousboe A (2003) Compartmentation of glutamine, glutamate, and GABA metabolism in neurons and astrocytes: Functional implications. *Neuroscientist* 9: 398–403.

Zwingmann C and Leibfritz D (2003) Regulation of glial metabolism studied by 13C-NMR. *NMR in Biomedicine* 16: 370–399.

Glial Energy Metabolism: Overview

L Pellerin, Université de Lausanne, Lausanne, Switzerland
P J Magistretti, Ecole Polytechnique Fédérale de Lausanne (EPFL) and Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the late nineteenth century, neuroanatomists described the morphological features of glial cells and their cytoarchitectural relationships with other elements of the central nervous system. Among glial cells, astrocytes are the most numerous and exhibit a number of striking characteristics. First, they occupy a strategic position, being usually located between blood vessels and neurons. Second, they possess specialized processes called end-feet, and these processes come into close contact with blood vessels, covering a large part of their surface. Third, they also ensheathe numerous synapses. Altogether, these characteristics suggested to those early neuroanatomists, and among them Camillo Golgi, that astrocytes could play a critical role in the allocation of metabolic substrates to neurons. This view was nicely formulated by Andriezen in 1893:

> The development of a felted sheath of neuroglia fibers in the ground-substance immediately surrounding the blood vessels of the Brain seems therefore ... to allow the free passage of lymph and metabolic products which enter into the fluid and general metabolism of the nerve cells.

It is only recently, however, that such a theory experienced a revival through a series of experimental demonstrations that benefited from the advances made in the isolation of specific brain cell types and their use *in vitro* to study their metabolic characteristics. The current understanding of brain energy metabolism is still incomplete, but in recent years insights have been gained into a number of aspects related to glial energy metabolism.

Aerobic Glycolysis: A Versatile Metabolic Feature of Astrocytes to Cope with the Burden of Glutamate Recycling

Several studies have demonstrated the large glycolytic capacity of astrocytes that is accompanied by a sizable production and release of lactate despite the sufficient amount of oxygen present and active oxidative phosphorylation. Not only is their basal glycolytic rate elevated, but astrocytes also possess a large glycolytic reserve that can be mobilized under various conditions. This is the case under hypoxia or if oxidative phosphorylation is inhibited. But more important, this reserve can be used in specific physiological situations. Thus, it has been shown that enhanced sodium entry within an astrocyte, causing an elevation of intracellular sodium concentration, will lead to an increased glycolytic rate. Such a situation occurs notably when astrocytes are exposed to the excitatory neurotransmitter glutamate. Glutamate transport occurs in astrocytes via two glutamate transporters, glutamate aspartate transporter (GLAST) and glutamate transporter 1 (GLT1). Both depend on the Na^+ gradient and carry glutamate, together with three Na^+ ions, inside the cell. The increase in intracellular Na^+ concentration that accompanies glutamate transport will activate Na^+/K^+ adenosine triphosphatase (ATPase), and as a consequence of adenosine triphosphate (ATP) consumption, an activation of glucose utilization will take place, together with enhanced lactate production (Figure 1). Such an activation of aerobic glycolysis in astrocytes appears to be facilitated by a few molecular characteristics. First, it has been shown that astrocytes have a large part of their pyruvate dehydrogenase enzyme present in a phosphorylated and thus inactive form. This characteristic may limit the entry of pyruvate into the tricarboxylic acid (TCA) cycle (Figure 2). Second, astrocytes appear to have very reduced levels of an essential component of the malate–aspartate shuttle called ARALAR 1. As a consequence, the transfer of reducing equivalents from the cytosol to the inner mitochondria is reduced, and cytosolic nicotinamide adenine dinucleotide (NAD^+) regeneration is limited. To maintain a high glycolytic rate as observed in astrocytes, which requires a rapid regeneration of NAD^+ levels, the solution is to convert pyruvate to lactate via the lactate dehydrogenase enzyme. To facilitate this task, astrocytes prominently express the lactate dehydrogenase M isoform that is not inhibited by pyruvate and sustains high enzymatic rates. Moreover, astrocytes were found to express the monocarboxylate transporter (MCT) isoforms MCT1 and MCT4 that are particularly suited for lactate export from glycolytic tissues because of high K_m values.

An important question is whether such a metabolic response of astrocytes can be triggered by other neuroactive substances. The main inhibitory neurotransmitter, γ-aminobutyric acid (GABA), is an interesting case since it is also transported by a Na^+-dependent mechanism in astrocytes and it is the most important neurotransmitter after glutamate. However, it was

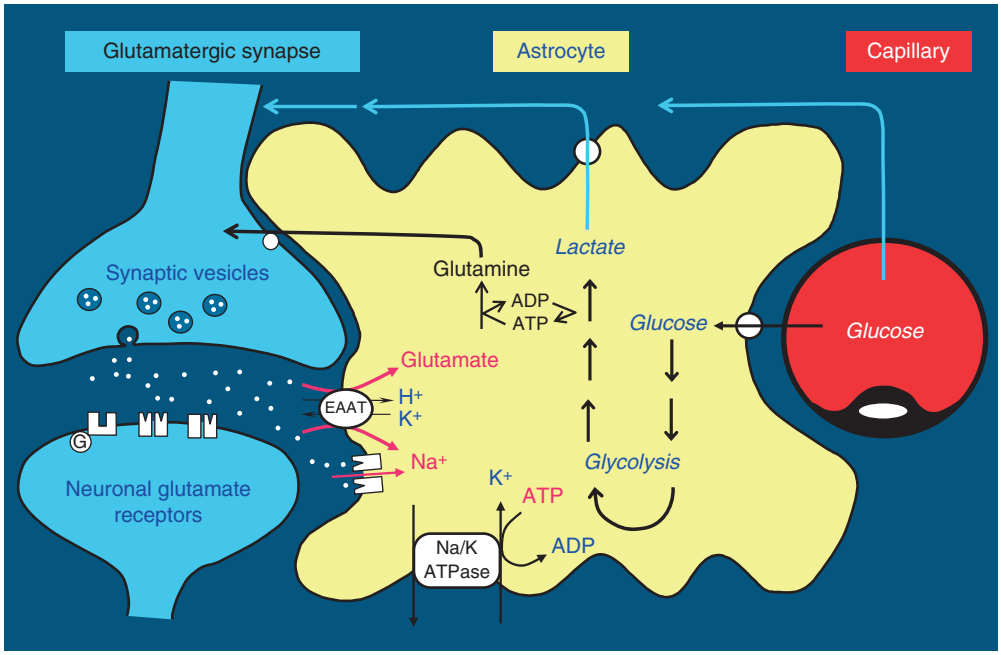

Figure 1 Schematic representation of the mechanism for glutamate-induced glycolysis in astrocytes during physiological activation. EAAT, excitatory amino acid transporter. Modified from Pellerin L and Magistretti PJ (1994) Glutamate uptake into astrocytes stimulates aerobic glycolysis: A mechanism coupling neuronal activity to glucose utilization. *Proceedings of the National Academy of Sciences of the United States of America* 91: 1065–10629.

shown that GABA does not cause a similar enhancement of aerobic glycolysis in astrocytes. The explanation resides in a rapid reduction in GABA transport with increasing intracellular Na^+ concentration before it reaches the threshold to activate glycolysis, a phenomenon not observed with glutamate transport. Several other neurotransmitters and neuromodulators have been tested, but all failed to activate glycolysis except noradrenaline. However, the mechanism by which noradrenaline stimulates glycolysis in astrocytes is different from that of glutamate but remains unclear. Another intriguing issue is whether the effect of glutamate is linked only to its uptake via glutamate transporters or whether activation of glutamate receptors (GluRs) can also participate in the metabolic effect. Indeed, astrocytes express both ionotropic and metabotropic glutamate receptors. Among ionotropic receptors, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)/kainate receptors are abundantly expressed, and all four subunit isoforms (GluR1–4) have been detected in astrocytes. Although activation of AMPA/kainate receptors with the specific agonist AMPA leads to a transient increase in intracellular Na^+ concentration, this is not sufficient to cause an activation of aerobic glycolysis. In this case, rapid desensitization of the receptors prevents the Na^+ influx from activating metabolism. Kainate causes a more important intracellular Na^+ change and a small but significant stimulation of aerobic glycolysis. This effect could be explained by the fact that kainate, in contrast to

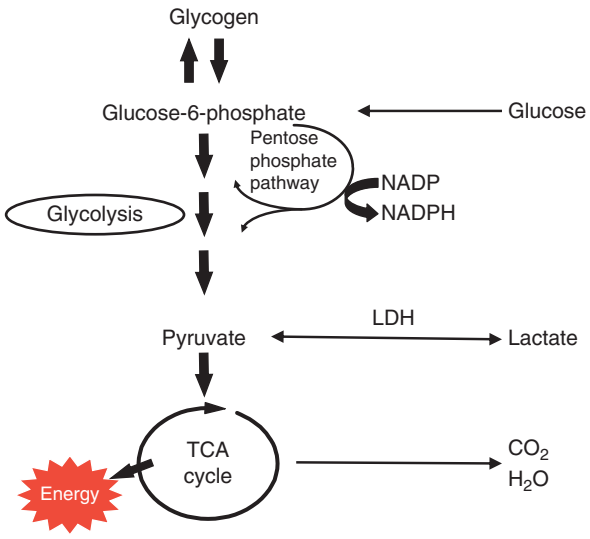

Figure 2 Metabolic pathways of glucose. LDH, lactate dehydrogenase; NADP, oxidized nicotinamide adenine dinucleotide phosphate; NADPH, reduced nicotinamide adenine dinucleotide phosphate; TCA, tricarboxylic acid cycle.

AMPA, leads to only a partial desensitization of AMPA/kainate receptors. One particularity of AMPA/kainate receptors is that they contain allosteric sites allowing certain agents to modulate their degree of desensitization/deactivation. In recent years, a series of compounds have been developed on the basis of their ability to prevent desensitization/deactivation of AMPA/kainate receptors. It is quite interesting that they were shown to exhibit neuroprotective and cognitive enhancement effects in various models showing impaired cognitive functions. Although the main explanation proposed for these effects is related to their action on neuronal AMPA/kainate receptors, it has also been suggested that a potentiation of glutamate-induced aerobic glycolysis in astrocytes occurs and participates in the beneficial effects of these drugs. Thus, enhancing the glycolytic response of astrocytes might represent a valuable therapeutic target for the future.

A new dimension of the role of enhanced aerobic glycolysis in astrocytes was revealed recently. It has been known for a while that astrocytes communicate with each other through Ca^{2+} waves that propagate in a regenerative manner along the astrocytic network. The specific role of these traveling Ca^{2+} waves is still not entirely clear. But in parallel with Ca^{2+}, Na^{+} waves have been described. These Na^{+} waves arise from glutamate released by one astrocyte on elevation of intracellular Ca^{2+} and taken up by its neighbors. As a consequence of this Na^{+} influx, a metabolic wave of enhanced aerobic glycolysis will ensue among the astrocytic network. It is purported that such a mechanism could be used to coordinate metabolism and provide adequate energy substrate supply to activated neuronal ensembles, irrespective of their neurotransmitter phenotypes. If confirmed *in vivo*, such a phenomenon would have important implications not only for an understanding of neuroenergetics but also for the analysis of images obtained with brain imaging techniques that rely on metabolic signals.

Glycogen: Helping to Face Increased Energy Demands

Glycogen has been viewed traditionally as an energy reserve maintained to face periods of low energy supply. This notion has been inherited from studies in peripheral organs such as the liver and muscle. However, the amount of glycogen contained in the brain is relatively small: approximately 10 times less than in muscle and 100 times less than in liver. Moreover, under physiological conditions, energy supply to the brain is maintained within narrow limits such that the idea of maintaining an emergency reserve for pathological situations appears unlikely. Rather, a series of observations argues in favor of a role for glycogen as an energy buffer rather than an energy reserve in the central nervous system.

Glycogen turnover was shown to be enhanced *in vivo* under sustained stimulation. In contrast, glycogen was found to accumulate in the brain under anesthesia. These observations suggest that glycogen utilization is tightly linked to neuronal activity. It is interesting that at the cellular level, glycogen is found almost exclusively in astrocytes. Since it is purported that the vast majority of energy expenditures are the consequence of neuronal activity, it may seem paradoxical that the sole energy reserve is located in a nonneuronal cell type. However, several lines of evidence indicate that astrocytes respond to neuronal signals and mobilize their glycogen stores to release a metabolite for use by neurons as an energy substrate. It has been shown that the brain, and in particular astrocytes, do not possess glucose-6-phosphatase and thus cannot release glucose as a product of glycogenolysis. Rather, it has been demonstrated that lactate constitutes the main substrate formed from glycosyl residues arising from glycogenolysis and released by astrocytes. Among the various neuroactive substances shown to induce glycogenolysis are the neurotransmitters noradrenaline and serotonin, the neuropeptide vasoactive intestinal peptide (VIP), and the neuromodulators adenosine and ATP. It has also been shown that elevated K^{+}occurring as a consequence of neuronal activity can efficiently cause glycogenolysis in astrocytes. In addition to glycogenolysis, several of these neuroactive substances promote a delayed but massive resynthesis of glycogen in astrocytes that takes place over several hours. Such resynthesis is critically dependent on the expression of a key protein called protein targeting to glycogen. This protein serves as a scaffold that binds not only glycogen but also all the enzymes involved in the regulation of glycogen synthesis and degradation. It is also massively induced in astrocytes following treatment with noradrenaline or VIP (Figure 3). Although its purpose is still a matter of debate, this mechanism might be essential to replenish glycogen stores and thus ensure adequate levels in order to face subsequent periods of high activity. In the optic nerve, glial glycogen has been shown to play a critical role in maintaining neuronal activity, in particular by providing lactate to meet the energetic demands of the axon.

The Pentose Phosphate Pathway: Energy Fluxes and Neuroprotection

ATP is not the only form of metabolic energy, as reducing power is needed in addition to ATP to energize

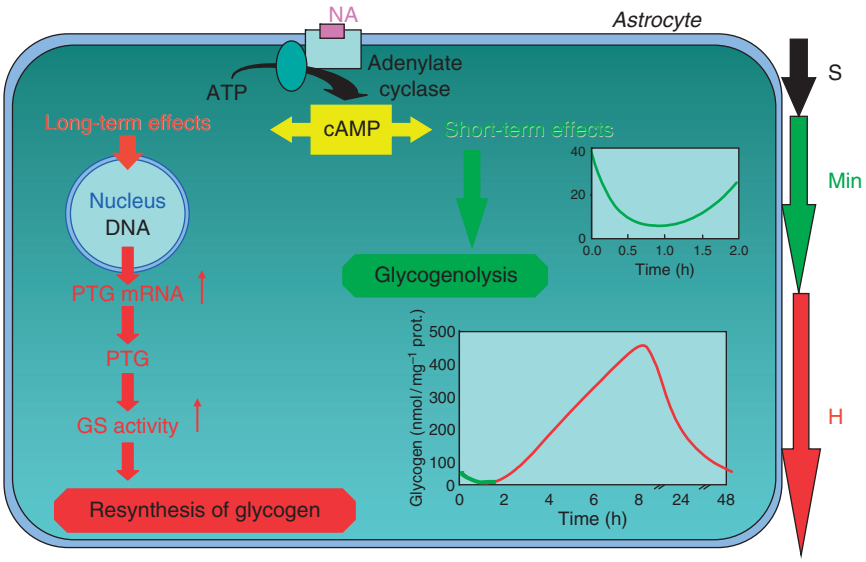

Figure 3 Regulation of glycogen metabolism by noradrenaline. Similar regulatory mechanisms are activated by vasoactive intestinal peptide and by adenosine. ATP, adenosine triphosphate; cAMP, cyclic adenosine monophosphate; GS, glutamine synthase; mRNA, messenger RNA; NA, nicotinamide adenine; Prot, protein; PTG, protein targeting to glycogen.

several important cellular processes. This reducing power is provided by the reduced form of nicotinamide adenine dinucleotide phosphate (NADPH). The processing of glucose through the pentose phosphate pathway produces NADPH (Figure 2).

A critical cellular process for which NADPH is needed is the scavenging of reactive oxygen species (ROS). The superoxide radical anion (O_2^-) hydrogen peroxide (H_2O_2), and the hydroxy radical (HO.) are three ROS generated by the transfer of single electrons to molecular oxygen as by-products of several physiological cellular processes. A considerable contribution to the generation of ROS is the oxidative metabolism of glucose taking place in the mitochondrial electron transfer chain associated with oxidative phosphorylation see below). Other ROS-generating reactions include the activities of monoamine oxidase, tyrosine hydroxylase, nitric oxide synthase, and the eicosanoid-forming enzymes lipoxygenases and cyclooxygenases. Reactive oxygen species are highly damaging to cells because they can cause DNA disruption and mutations, as well as activation of enzymatic cascades, including proteases and lipases, that can eventually lead to cell death. The coordinated activity of NADPH and glutathione is essential in protecting cells against ROS-mediated damage, or oxidative stress. Scavenging of ROS is ensured by the sequential action of superoxide dismutase (SOD) and glutathione peroxidase. Thus, two superoxide anions are converted by SOD into H_2O_2, still an ROS. Glutathione peroxidase converts H_2O_2 into H_2O and O_2 at the expense of reduced glutathione, which is regenerated by glutathione reductase in the presence of NADPH.

Scavenging of ROS provides another example of metabolic cooperation between neurons and astrocytes. Glutathione is a tripeptide (γ-L-glutamyl-L-cysteinylglycine (GSH)) synthesized through the concerted action of two enzymes, γ-GluCys synthase, which combines glutamate and cysteine to yield the dipeptide γ-Glu Cys, and glutathione synthase, which adds a glycine to the dipeptide to yield GSH.

The glutathione content and reducing potential are considerably higher in astrocytes than in neurons; this fact, combined with the much higher oxidative activity of neurons than of astrocytes, makes neurons more vulnerable to oxidative stress as well as highly dependent on astrocytes for protection. Indeed, a cooperativity between astrocytes and neurons appears to exist for glutathione metabolism; astrocytes release GSH, which is cleaved by the ectoenzyme γ-glutamyltransferase, releasing CysGly. The dipeptide is transported into neurons (note that neurons cannot take up GSH), providing two precursors for GSH synthesis. Glutamate, the third precursor of GSH, is also provided by astrocytes to neurons in the form of glutamine, from which glutamate is produced through the action of glutaminase.

This astrocyte–neuron metabolic cooperation aimed at controlling the damaging effect of ROS generated by cellular activities may be critical for neuroprotection. Indeed, neurons, being highly oxidative cells

dependent on astrocytes for scavenging ROS, are at high risk for neurodegeneration. Indeed, dysfunction of such neuroprotective mechanisms has been described in certain neurodegenerative diseases, such as a familial form of amyotrophic lateral sclerosis, due to an SOD mutation. Evidence for a decrease in GSH content in the substantia nigra has also been described in Parkinson's disease.

The TCA Cycle, Glutamate Metabolism, and Anaplerosis

The main metabolic pathway for ATP production is the TCA cycle. While this pathway is active in both neurons and astrocytes, neurons, being richer than astrocytes in mitochondria where the TCA cycle occurs, are overall more oxidative than astrocytes. In addition to its central role in ATP production through its coupling to oxidative phosphorylation, the TCA cycle contributes to another critical aspect of astrocyte–neuron metabolic cooperation, related to glutamate metabolism.

Synaptically released glutamate is removed rapidly from the extracellular space by a transporter-mediated reuptake system that is particularly efficient in astrocytes. This mechanism contributes in a crucial manner to the fidelity of glutamate-mediated neurotransmission. Indeed, glutamate levels in the extracellular space are low ($<3\ \mu mol^{-1}$), allowing for optimal glutamate-mediated signaling after depolarization while preventing overactivation of glutamate receptors, which could eventually result in excitotoxic neuronal damage.

What is the metabolic fate of glutamate taken up by astrocytes? Glutamate cannot be stored like carbohydrates or lipids. In astrocytes, one pathway involves the TCA cycle, namely, the transfer of the α-amino group of glutamate to oxaloacetate to yield α-ketoglutarate (α-KG) and aspartate in a reaction catalyzed by aspartate amino transferase (AAT). The α-KG generated is an intermediate of the TCA cycle and is therefore oxidized further. Another transamination reaction catalyzed by alanine amino transferase (ALAT) transfers the α-amino group of glutamate to pyruvate, resulting in the formation of alanine and α-KG.

Two other pathways exist in astrocytes to metabolize glutamate. First, glutamate can be converted directly to α-KG through an NAD-requiring oxidative deamination catalyzed by glutamate dehydrogenase (GDH). Glutamate, by entering the TCA cycle indirectly (through AAT or ALAT) or directly (through GDH), is an energy substrate for astrocytes. Second, the quantitatively predominant metabolic pathway of glutamate in astrocytes is its amidation to glutamine, an ATP-requiring reaction in which an ammonium ion is fixed on glutamate. This reaction is catalyzed by glutamine synthase, an enzyme almost exclusively localized in astrocytes, and provides an efficient means of disposing not only of glutamate but also of ammonium (Figure 1). Glutamine is released by astrocytes and is taken up by neurons, where it is hydrolyzed back to glutamate by the phosphate-dependent mitochondrial enzyme glutaminase. This metabolic pathway, often referred to as the glutamate–glutamine shuttle, is a clear example of cooperation between astrocytes and neurons. It allows the removal of potentially toxic excess glutamate from the extracellular space while returning to the neuron a synaptically inert (glutamine does not affect neurotransmission) precursor with which to regenerate the neuronal pool of glutamate.

However, not all glutamate is regenerated through the glutamate–glutamine shuttle because some of the glutamate released by neurons enters at the α-KG level, the TCA cycle in astrocytes; therefore, *de novo* synthesis is required to maintain the neuronal glutamate pool. Glutamate can be synthesized through NADPH-dependent reductive amination of α-KG catalyzed by GDH (note here another role of NADPH, in addition to ROS scavenging, in cooperation with glutathione). For this pathway leading to the synthesis of glutamate, glucose provides the carbon backbone as α-KG does through the TCA cycle, while plasma leucine, taken up by astrocytic end-feet, provides the nitrogen required for net glutamate synthesis from α-KG. Because this reaction, catalyzed by leucine transaminase, takes place in astrocytes to replenish the neuronal glutamate pool, the astrocytes export glutamate as glutamine.

As can be appreciated, because α-KG is used for glutamate synthesis, metabolic intermediates downstream of α-KG must be available to maintain a sustained flux through the TCA cycle in astrocytes. This need is met by the activity of the enzyme pyruvate carboxylase, which fixes CO_2 on pyruvate to generate oxaloacetate, which, by condensing with acetyl-CoA, maintains the flux through the TCA cycle. The carboxylation of pyruvate to oxaloacetate is referred to as an anaplerotic (Greek for 'fill up') reaction. It is interesting that like glutamine synthase, pyruvate carboxylase is selectively localized in astrocytes. The fact that these two enzymes are localized in astrocytes in conjunction with the existence of a glutamate–glutamine shuttle stresses that astrocytes are essential for maintaining the neuronal glutamate pool used for neurotransmission.

The Astrocyte–Neuron Metabolic Unit

In conclusion, the axon terminals at glutamatergic synapses, which represent more than 80% of all

synapses in the cerebral cortex, and the astrocytic processes that surround them should be viewed as a metabolic unit in which the neuron furnishes the activation signal (glutamate) to the astrocyte, and the astrocyte provides, not only the precursors needed to maintain the neurotransmitter pool (glutamine), but also the energy substrate (lactate) and a neuroprotective agent (glutathione) to scavenge ROS formed by the active oxidative phosphorylation of neurons. Thus the efficacy of the predominant excitatory synapse in the brain, the glutamatergic synapse, cannot be maintained without a close astrocyte–neuron interaction.

See also: Activity-Dependent Metabolism in Glia and Neurons; Activity-Dependent Regulation of Glucose Transporters; Glial Glutamate Transporters: Electrophysiology; Glial Glutamate Transporters; Glial Ion Homeostasis: A Fluorescence Microscopy Approach; Glial Glycogen Metabolism; Glial Glutamate and GABA Metabolism; Glial Energy Metabolism: A NMR Spectroscopy Perspective; Ionic Channels in Glia.

Further Reading

Beal FM (2005) Mitochondria take center stage in aging and neurodegeneration. *Annals of Neurology* 58: 495–505.

Bernardinelli Y, Magistretti PJ, and Chatton J-Y (2004) Astrocytes generate Na^+-mediated metabolic waves. *Proceedings of the National Academy of Sciences of the United States of America* 101: 14937–14942.

Brown AM (2004) Brain glycogen reawakened. *Journal of Neurochemistry* 89: 537–552.

Dringen R (2000) Metabolism and functions of glutathione in brain. *Progress in Neurobiology* 62: 649–671.

Erecinska M and Silver IA (1990) Metabolism and role of glutamate in mammalian brain. *Progress in Neurobiology* 35: 245–296.

Hyder F, Patel AB, Gjedde A, Rothman DL, Behar KL, and Shulman RG (2006) Neuronal-glial glucose oxidation and glutamatergic-GABAergic function. *Journal of Cerebral Blood Flow and Metabolism* 26: 865–877.

Magistretti PJ (2003) Brain energy metabolism. In: Squire L (ed.) *Fundamental Neuroscience*, 2nd edn., pp. 339–360. San Diego: Academic Press.

Magistretti PJ (2006) Neuron-glia metabolic coupling and plasticity. *Journal of Experimental Biology* 209: 2304–2311.

Magistretti PJ, Pellerin L, Rothman DL, and Shulman RG (1999) Energy on demand. *Science* 283: 496–497.

McEwen BS and Reagan LP (2004) Glucose transporter expression in the central nervous system: Relationship to synaptic function. *European Journal of Pharmacology* 490: 13–24.

Pellerin L and Magistretti PJ (1994) Glutamate uptake into astrocytes stimulates aerobic glycolysis: A mechanism coupling neuronal activity to glucose utilization. *Proceedings of the National Academy of Sciences of the United States of America* 91: 1065–10629.

Tsacopoulos M and Magistretti PJ (1996) Metabolic coupling between glia and neurons. *Journal of Neuroscience* 16: 877–885.

Glial Glutamate and GABA Metabolism

H S Waagepetersen and A Schousboe, University of Copenhagen, Copenhagen, Denmark

© 2009 Elsevier Ltd. All rights reserved.

Termination of Excitatory and Inhibitory Neurotransmission

The two amino acids glutamate and γ-aminobutyric acid (GABA), acting as an excitatory and inhibitory neurotransmitter, respectively, probably account for 80–90% of all neurotransmission signaling in the central nervous system (CNS). In the case of glutamate, termination of the transmission process is to a very large extent based on uptake of glutamate into astrocytes completely ensheathing the synapses. In GABAergic neurotransmission, on the other hand, astrocytic uptake of GABA seems less pronounced, and the majority of released GABA appears to be transported back into the presynaptic GABAergic nerve endings for recycling as a neurotransmitter. The fraction of released GABA which is taken up into surrounding astrocytes undergoes oxidative metabolism, as is delineated later. The balance between neuronal and glial GABA uptake is likely to play a fundamental role for the optimal function of inhibitory neurotransmission.

Glutamate Metabolism

Glutamate is taken up into astrocytes by the two transporters GLAST (EAAT1) and GLT-1 (EAAT2), which are abundantly expressed in astrocytic membranes, accounting for at least 10% of all membrane-associated proteins. This provides astrocytes with an enormous capacity for glutamate transport, a feature of great importance for maintaining a low extracellular glu tamate concentration, thereby preventing excitotoxic neuronal damage and cell death. As shown schematically in Figure 1, a complex set of enzymatically catalyzed reactions is responsible for glutamate metabolism in astrocytes. The major metabolic pathways are those catalyzed by the enzyme glutamate dehydrogenase (GDH) and aminotransferases such as aspartate aminotransferase (AAT), alanine aminotransferase (ALAT), and branched chain amino acid transaminase (BCAT), to mention the most important ones. Additionally, glutamate can be amidated in an ATP-requiring process catalyzed by glutamine synthetase (GS), an enzyme which is exclusively expressed in glial cells. This latter reaction produces glutamine which will be released from the astrocytes utilizing the glutamine transporters of system N and L. The former reactions lead to production of α-ketoglutarate which will enter the tricarboxylic acid (TCA) cycle for further metabolism, resulting in oxidation of the carbon skeleton, a process ultimately requiring pyruvate recycling to occur. Whether or not glutamate undergoes amidation (GS reaction) or oxidation (GDH/aminotransferases plus TCA cycle) is of fundamental importance for the maintenance of glutamatergic neurotransmission. Formation of glutamine and its subsequent release constitute the basis for the operation of the glutamate–glutamine cycle (Figure 2). If, however, the carbon skeleton of glutamate enters the TCA cycle, it is prevented from being part of the glutamate–glutamine cycle and glutamate is lost from the transmitter pool. Astrocytes are equipped with an alternative enzymatic machinery to compensate for this. Pyruvate formed from glucose in the glycolytic pathway can undergo carboxylation in an energy (ATP)-requiring process catalyzed by pyruvate carboxylase (PC), another astrocyte-specific enzyme. This may allow a net synthesis of either citrate or glutamine, of which particularly the latter can be used as precursor for glutamate in glutamatergic neurons. Citrate is also released from astrocytes, but the functional importance of this process is presently only poorly understood.

GABA Metabolism

The only GABA-metabolizing enzyme is GABA-aminotransferase (GABA-transaminase, GABA-T), which is expressed in all cells of the CNS as well as peripherally. It catalyzes the conversion of GABA to succinic semialdehyde (SSA) concomitantly with conversion of α-ketoglutarate to glutamate (Figures 3(a) and 3(b)). Subsequently, SSA is oxidized to succinate by the enzyme SSA-dehydrogenase (SSADH), which as a consequence prevents the otherwise freely reversible aminotransferase reaction from producing GABA due to a very low SSA concentration. Both GABA-T and SSADH are mitochondrial enzymes. The two metabolic schemes in Figures 3(a) and 3(b) illustrate that α-ketoglutarate, which is required as a co-substrate for GABA in the GABA-T reaction, may originate from the TCA cycle by two different series of reactions. It may be the result of conversion of succinate to oxaloacetate and further generation of α-ketoglutarate after condensation with acetyl CoA produced from glucose via pyruvate formed

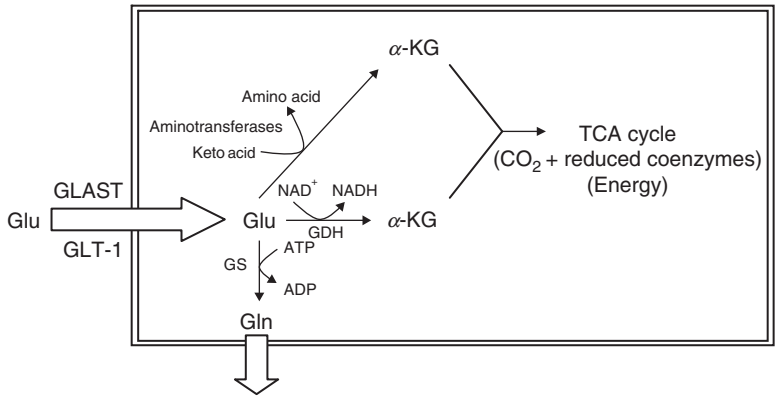

Figure 1 A schematic representation of glutamate (Glu) metabolism in astrocytes. Glu is taken up by high-affinity Glu carriers, that is, GLAST and GLT-1, and subsequently converted into either glutamine (Gln) by the operation of glutamine synthetase (GS) or α-ketoglutarate (α-KG) by glutamate dehydrogenase (GDH) or an aminotransferase. The latter reaction simultaneously converts a keto acid to the corresponding amino acid. α-KG may subsequently accelerate oxidation of acetyl CoA units in the tricarboxylic (TCA) cycle, producing energy. Gln is released from the astrocyte.

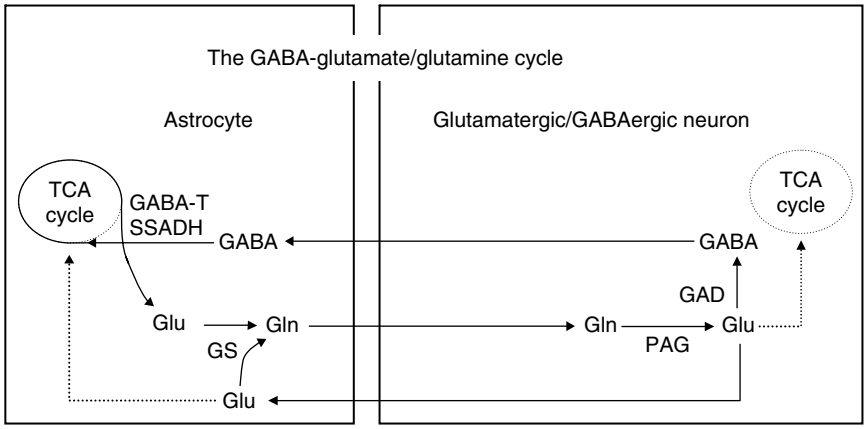

Figure 2 The GABA-glutamate (Glu)/glutamine (Gln) cycle depicts a metabolic interaction between astrocytes and glutamatergic or GABAergic neurons for maintenance of Glu and GABA homeostasis during neurotransmission. Glu and GABA are both synthesized from Gln by the concerted action of phosphate-activated glutaminase (PAG) and glutamate decarboxylase (GAD). Glu and GABA are released from the glutamatergic and GABAergic neurons, respectively, and subsequent to synaptic action, they are both partly taken up by surrounding astrocytes. GABA is degraded to succinate, a tricarboxylic (TCA) cycle intermediate, by the combined action of GABA-transaminase (GABA-T) and succinate semialdehyde dehydrogenase (SSADH). Succinate may after one turn in the TCA cycle serve as precursor for Glu and subsequently for Gln via glutamine synthetase (GS). Glu is either directly converted into Gln or metabolized in the TCA cycle. Gln is released from the astrocytes and taken up by the neuron completing the cycle.

by the glycolytic pathway (Figure 3(a)). This series of reactions allows glutamate produced from α-ketoglutarate in the GABA-T reaction to be amidated to glutamine by GS and thereby provides the means to generate glutamine from GABA. This allows release of glutamine for uptake into, for example, GABAergic neurons constituting the glu-GABA-glutamine cycle, which is an extension of the glutamate/glutamine cycle (Figure 2).

An alternative series of reactions is shown in Figure 3(b). Regeneration of α-ketoglutarate from glutamate in the GABA-T reaction takes place in an aminotransferase reaction catalyzed by alanine aminotransferase using pyruvate generated from glucose in the glycolytic pathway. This results in a net production of alanine. It should be noted that this alternative set of reactions does not lead to production of glutamine, excluding the operation of the GABA–glutamate–glutamine cycle. It should be noted that while the series of reactions shown in Figure 3(a) leads to production of glutamine without the need for anaplerosis involving pyruvate carboxylase, no glutamine production is associated with the reactions outlined in Figure 3(b). This means that in order to fulfill the requirement for glutamine production in and release from the astroglial compartment, anaplerosis must take place. These considerations are important in the light of the fact that glutamine is the

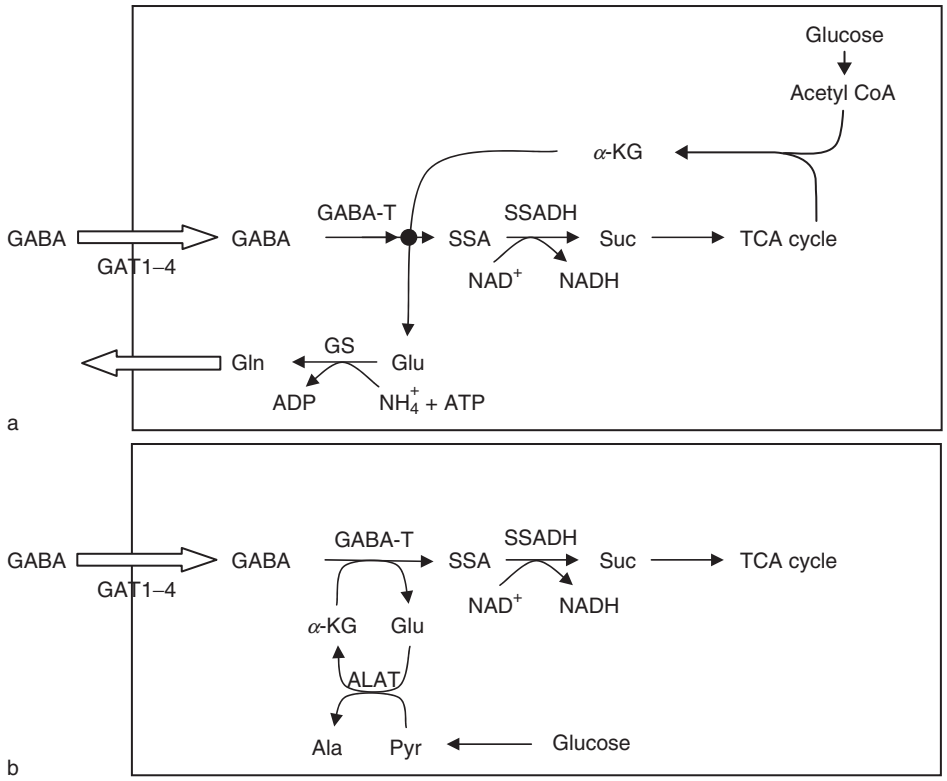

Figure 3 A schematic representation of GABA metabolism suggesting possible coupling to either glutamine (Gln) or alanine (Ala) synthesis in astrocytes. GABA is taken up by high-affinity GABA carriers, that is, GAT1–4, and subsequently degraded to succinate (Suc) via succinate semialdehyde (SSA) catalyzed by the combined operation of GABA-transaminase (GABA-T) and succinate semialdehyde dehydrogenase (SSADH). (a) The action of GABA-T simultaneously converts α-ketoglutarate (α-KG) to glutamate (Glu) being the precursor of releasable Gln via glutamine synthetase (GS). Suc may be precursor of this Gln, since α-KG can be formed from Suc via the tricarboxylic (TCA) cycle and condensation of one molecule of acetyl CoA possibly generated from glucose. (b) Glu formed in the GABA-T reaction may alternatively be reconverted into α-KG by the functioning of alanine aminotransferase (ALAT) concurrently forming Ala from pyruvate (Pyr), the latter being formed from glucose. This series of reactions thus leads to net synthesis of Suc, a TCA cycle constituent.

most important precursor for GABA synthesis in GABAergic neurons.

Pharmacology of GABA Metabolism

Contrary to glutamate metabolism, the GABA metabolic pathway constitutes an important pharmacological target. Inhibition of GABA degradation is theoretically an interesting way to elevate GABA levels. However, due to the fact that both the GABA-synthesizing enzyme glutamate decarboxylase (GAD) and the catabolic enzyme GABA-T utilize pyridoxal phosphate as a coenzyme, it is difficult to affect only one of the enzymes since carbonyl trapping agents will affect both enzymes. This problem was solved using γ-vinyl GABA, which is a catalytic substrate-directed suicide inhibitor specific for GABA-T. This compound was developed into a clinically active antiepileptic drug by the name vigabatrin. It was later discovered that it not only is specific for GABA-T but also affects particularly the neuronal (neurotransmitter) pool of GABA due to the presence of a vigabatrin transporter preferentially expressed in neurons.

See also: GABA Synthesis and Metabolism; Gamma-Aminobutyric Acid (GABA); Glial Glutamate Transporters: Electrophysiology; Glial Glutamate Transporters; Glutamate; Glutamatergic and Gabaergic Systems.

Further Reading

Bak LK, Schousboe A, and Waagepetersen HS (2006) The glutamate/GABA-glutamine cycle: Aspects of transport, neurotransmitter homeostasis and ammonia transfer. *Journal of Neurochemistry* 98: 641–653.

Danbolt NC (2001) Glutamate uptake. *Progress in Neurobiology* 65: 1–105.

Lippert B, Metcalf BW, Jung MJ, and Casara P (1977) 4-Aminohex-5-enoic acid, a selective catalytic inhibitor of 4-aminobutyric-acid aminotransferase in mammalian brain. *European Journal of Biochemistry* 74: 441–445.

McKenna MC, Gruetter R, Sonnewald U, Waagepetersen HS, and Schousboe A (2006) Energy metabolism of the brain. In: Siegel GJ, Agranoff BW, Albers RW, Fisher SK, and Uhler MD (eds.) *Basic*

Neurochemistry: Molecular, Cellular, and Medical Aspects, 7th edn., pp. 531–557. Philadelphia: Lippincott-Raven.

McKenna MC, Sonnewald U, Huang X, Stevenson J, and Zielke RH (1996) Exogenous glutamate concentration regulates the metabolic fate of glutamate in astrocytes. *Journal of Neurochemistry* 66: 386–393.

Norenberg MD and Martinez-Hernandez A (1979) Fine structural localization of glutamine synthetase in astrocytes of rat brain. *Brain Research* 161: 303–310.

Sarup A, Larsson OM, and Schousboe A (2003) GABA transporters and GABA-transaminase as drug targets. *Current Drug Targets CNS and Neurological Disorders* 2: 269–277.

Schousboe A, Larsson OM, and Seiler N (1986) Stereoselective uptake of the GABA-transaminase inhibitors gamma-vinyl GABA and gamma-acetylenic GABA into neurons and astrocytes. *Neurochemical Research* 11: 1497–1505.

Schousboe A, Sarup A, Bak LK, Waagepetersen HS, and Larsson OM (2004) Role of astrocytic transport processes in glutamatergic and GABAergic neurotransmission. *Neurochemistry International* 45: 512–527.

Schousboe A, Sarup A, Larsson OM, and White HS (2004) GABA transporters as drug targets for modulation of GABAergic activity. *Biochemical Pharmacology* 68: 1557–1563.

Schousboe A and Waagepetersen HS (2006) Glial modulation of GABAergic and glutamatergic neurotransmission. *Current Topics in Medicinal Chemistry* 6: 929–934.

Sonnewald U, Westergaard N, and Schousboe A (1997) Glutamate transport and metabolism in astrocytes. *Glia* 21: 99–105.

Tapia R (1975) Biochemical pharmacology of GABA in CNS. *Handbook of Psychopharmacology* 4: 1–58.

Van den Berg CJ and Garfinkel D (1971) A simulation study of brain compartments: Metabolism of glutamate and related substances in mouse brain. *Biochemical Journal* 23: 211–218.

Westergaard N, Banke T, Wahl P, Sonnewald U, and Schousboe A (1995) Citrate modulates the regulation by Zn^{++} of NMDA receptor-mediated channel current and neurotransmitter release. *Proceedings of the National Academy of Sciences of the United States of America* 92: 3367–3370.

Westergaard N, Drejer J, Schousboe A, and Sonnewald U (1996) Evaluation of the importance of transamination versus deamination in astrocytic metabolism of [U-^{13}C]glutamate. *Glia* 17: 160–168.

Yu ACH, Drejer J, Hertz L, and Schousboe A (1983) Pyruvate carboxylase activity in primary cultures of astrocytes and neurons. *Journal of Neurochemistry* 41: 1484–1487.

Glial Glutamate Transporters

A Nieoullon, CNRS – Université de la Méditerranée, Marseille, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

At glutamatergic synapses, excitatory amino acid (EAA) transporters are involved in the removal of the neurotransmitter glutamate (Glu) from the synaptic cleft to stop intercellular signaling following the release of the neurotransmitter and action on receptors (Figure 1). EAAs are major brain neurotransmitters. They are primarily involved in the so-called excitatory 'fast signaling,' which supports rapid transfer of excitatory information in neuronal networks. Glu is the main EAA. It is generally accepted to be present in approximately 30–40% of total brain synapses and probably much more in selected cortical and hippocampal areas involved in cognitive processes. One of the most interesting properties of the EAA synapse, however, is its contribution to neuronal plasticity through cellular mechanisms known as long-term potentiation and long-term depression processes, which are generally accepted to represent basic mechanisms for learning and memory. Indeed, numerous reports have shown specific impairments of learning and memory when synaptic transmission involving the NMDA receptor subtypes is altered *in vivo*.

In addition to the key contribution to the transfer of excitatory information in the brain, EAAs and especially Glu exhibit the unique property of representing a very potent endogenous neurotoxin. Pharmacological stimulation of EAA receptors using various Glu agonists has been shown both *in vitro* and *in vivo* to induce neuronal death, suggesting that an overconcentration of EAAs in the synaptic space is deleterious for the cells due to an excess of excitation. This process is called excitotoxicity. However, excitotoxicity is not the only mechanism that induces neuronal death; the depolarizing action of Glu is concomitant with a massive increase in Na^+ in the cell, which can lead to cell death through swelling of the cell in the case of overstimulation of the EAA receptors. Therefore, although such a consideration of neuropathology can be viewed as highly reductionistic, dysfunctions of brain neurotransmission involving EAAs have been reported in select brain diseases with neuronal death and their putative contribution to many others is often proposed. However, what is the basic nature of the mechanisms involved in the clearance of EAAs at the synapse? What is the role of the transporters? What could be their real contribution to neurodegenerative diseases?

The EAA Transporter Family as Key Membrane Proteins

Five members of the Glu transporter family have been characterized. The human subtypes are designed EAAT1–EAAT5. EAAT1, -2, and -3 are the human homologs, respectively, of GLAST, GLT-1, and EAAC1, the three main Glu transporters in the rat brain cloned in 1992. GLT-1 was initially purified as a 73 kDa glycoprotein and GLAST as a 66 kDa protein. When comparing the different Glu transporters, the identity of amino acid sequence is in the range of 44–55%. *Drosophila* EAAT homologs dEAAT1 and dEAAT2 have also been identified in the central nervous system. However, only dEAAT1 can transport Glu since dEAAT2 was characterized as an aspartate/taurine transporter.

EAAT1/GLAST and EAAT2/GLT-1 are primarily astrocytic, although some authors have suggested a limited neuronal expression of EAAT2/GLT-1. Conversely, EAAT3/EAAC1 and EAAT4 and -5 are considered to be neuronal transporters. However, EAAT3/EAAC1 expression was proposed to occur in some astrocytes and even oligodendrocytes, but such glial expression is considered as exceptional and limited.

Genetic approaches have been used following biochemical characterization of the transporters. Five different genes related to the solute carrier family 1 (SLC1) genes have been identified in humans. For example, the human EAAT2 gene is localized on chromosome 11p13–12 and consists of approximately 100 kb of DNA. The gene contains 11 exons. At least three different splice variants of the EAAT2/GLT-1 protein are issued from the reading of the sequence as the result of alternative splicing. In this respect, it has been shown that in an animal model of amyotrophic lateral sclerosis (ALS) exhibiting a mutation in the superoxide dismutase 1 (SOD1) gene associated with some forms of the human disease, the expression of five different splice variants of EAAT2/GLT-1 is affected.

Regarding the neuronal transporters, very little is known about EAAT4 and -5. The contribution of EAAT3/EAAC1 to glutamatergic transmission is still under investigation. Briefly, its postsynaptic localization and early expression in the fetus suggest a contribution to brain development. Because of its apparent capacity to also contribute to neuronal uptake of

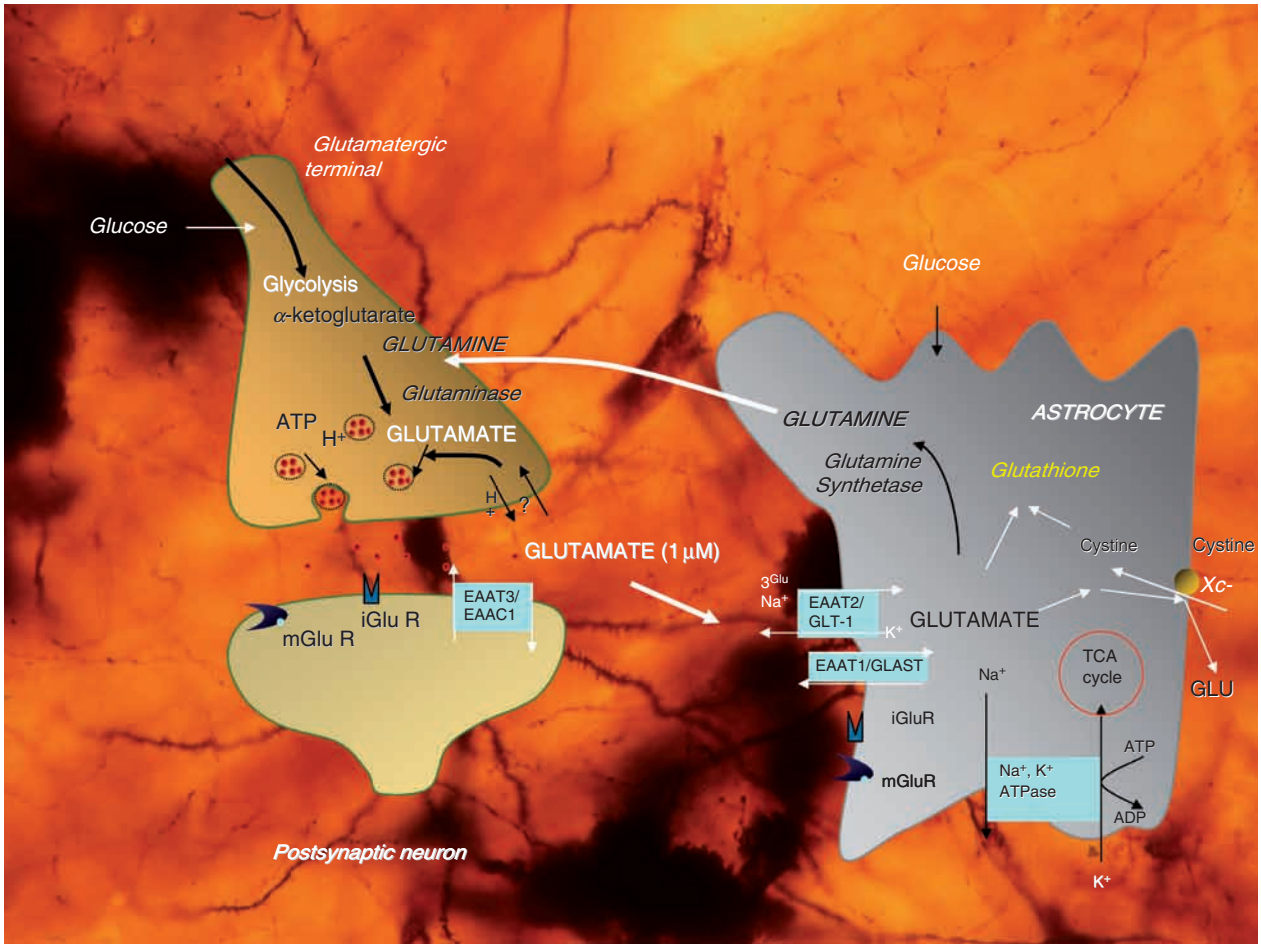

Figure 1 Main features of a brain synapse using excitatory amino acids as neurotransmitter. The neurotransmitter glutamate (Glu) is released from nerve terminals by an active process. In the synaptic space, Glu acts on receptors of both ionotropic (iGluR) and metabotropic (mGluR) subtypes. The termination of Glu action is assumed by quick removal from synaptic space either by the neuronal EAAT3/EAAC1 transporter, primarily located on the postsynaptic membrane, or by astrocytic EAAT1/GLAST and EAAT2/GLT-1 transporters. The question of the presence of a presynaptic Glu transporter is still a matter of debate. The diagram also shows the ion dependence of Glu transport into cells through transporter systems and the destination of Glu in the astrocytes. Glu is first subjected to transformation into glutamine, which acts as a Glu precursor when transported into the neuron (referred as the Glu–glutamine cycle). Glu is also transformed into glutathione, which constitutes one of the main antioxidative defenses of astrocytes. In fact, glutathione is even more dependent on Glu since Glu is also involved in the uptake of cystine through the transporter Xc-, which also acts as a precursor for glutathione. Newly synthesized Glu from glucose also contributes to neurotransmitter biosynthesis. The electrochemical gradient is maintained by Na^+/K^+ ATPase.

cysteine, which acts as a key substrate for the synthesis of glutathione, such a transporter may contribute to neuroprotective mechanisms against oxidative stress. Interestingly, numerous regulatory mechanisms involving both extracellular signals and intracellular signaling cascades have been evidenced, thus suggesting that the fine tuning of EAA neuronal uptake is a highly regulated process, which influences glutamatergic excitatory transmission.

EAA Transport Shows Selective Ion Dependence

One of the main characteristics of the transport involving members of the SLC1 gene family is the dependence of the transport of extracellular sodium concentration. The electrogenicity of the transport of Glu was first demonstrated approximately 30 years ago using brain synaptosomes, although in such a preparation the contribution of each selective ion was not easy to establish. Data from studies using cell expression of the cloned transporters and patch clamp analysis have identified the precise stoichiometry of the transport. For each Glu molecule transported into the cell, the transport involves two or three sodium ions (Na^+) and one hydrogen (H^+) ion entering the cell together with the molecule of Glu, in exchange for one potassium (K^+) ion. A sodium-activated chloride (Cl^-) conductance is also mediated through EAATs, but such a conductance is

not thermodynamically linked to Glu translocation into the cell. Site-directed mutagenesis has been used to identify amino acid residues related to ion transports associated with EAA translocation, verifying that Cl^- movement is not necessary for EAA transport.

Consequently, it appears that the transport of EAAs into the cells is driven by the electrochemical gradient existing through the membrane. Such a transport is highly efficient to concentrate the neurotransmitter into the cell. The transport is electrogenic, generating a net depolarization of the membrane, which can be measured by using electrophysiological recordings. Detailed analysis of currents produced by the transport has led to the proposal of hypothetic kinetic models which describe the different steps of Glu transport assuming that EAATs behave like sequential exchangers. Differences may be related to variations in the cycle and efficacy in the transport regarding the heterogeneity of the transporters. In this respect, although there are some differences, Glu transport appears to be rather slow since one complete transport cycle is believed to range from 60 to 80 ms.

Molecular Structure and Presumed Topology of EAATs

After purification of the first EAATs was achieved, sodium dodecyl sulfate–polyacrylamide gel electrophoresis demonstrated that these proteins have a tendency to form oligomers. In contrast to the pentameric structure proposed for EAAT3/EAAC1, the glial transporter contributing to the maturation of the pore was proposed to involve a trimeric structure. Such a proposal was verified by the crystallization of bacterial homologs of the SLC1 Glu transporters. The Glu transporters in eukaryotic cells are in fact members of a more general membrane transport protein family also involving prokaryotic transporters. Both types of proteins share a high degree of amino acid homology, similar substrate transport ionic specificity, and close transmembrane topology. The crystallized transporter from *Pyrococcus horikoshii* (Glt_{Ph}) shows a 37% sequence identity with the human EAAT2/GLT-1. In agreement with the previously proposed multimeric structure, which emphasized that trimers and dimers predominate, in the presence of Glu the Glt_{Ph} transporter appears to consist of three functional units, also in agreement with previous results obtained from the purification of Glu transporters from *Bacillus caldotenax* and *Bacillus stearothermophilus*.

The high-resolution topology of each subunit, however, remains to be clarified. Although two main models have been proposed based on studies using site-directed-mutagenesis, the controversy persists. The latest developments suggest that the protein, which exhibits intracellular localization of N- and C-terminals, consists of six transmembrane (TM) domains thought to be organized as α-helices, followed by two reentrant loops corresponding to hairpin loops in the proximity of the C-terminal part of the protein (Figure 2). Regarding GLT-1 topology, several amino acid residues have been shown to be critical for Glu transport, for example, at the levels of the two hairpin loops. The protein can be glycosylated in the large extracellular hydrophilic loops between TM3 and TM4 (two sites) and shows putative consensus phosphorylation sites in its intracellular part, suggesting the possibility of posttraductional regulation of activity. Furthermore, because of interactions between the transporters and proteins located in the immediate vicinity, there is a strong possibility that such partners contribute to intracellular trafficking to and from the plasma membrane. Regarding EAAT2/GLT-1, the transporter was shown, for example, to interact at its N-terminal ending with a protein of the LIM family Ajuba, which may be involved in regulation of intracellular signaling cascades or acting on the cytoskeleton. Similarly, the C-terminal part of the EAAT1/GLAST protein may be preferentially linked to the protein septine 2, a member of the GTPase family, and such an interaction may be involved in the intracellular trafficking of the transporter and, specifically, its presence at the membrane. Finally, because depletion of membrane cholesterol by methyl-β-cyclodextrin induced primarily marked reduction of glial Glu uptake in primary cortical cultures, it has been proposed that the Glu transporters, including EAAT3/EAAC1 but especially EAAT2/GLT-1, are associated with cholesterol-enriched lipid raft microdomains, and that such a localization is important for function.

Glial EAATs Are a Target for Drugs: Toward a New Molecular Pharmacology?

Because of high structural homology, it is difficult to selectively act on the different transporter subtypes. The initial generation of compounds that inhibit EAA transport of Glu focused on competitive substrates such as pyrrolidine decarboxylate (PDC) derivatives, which induce an efflux of endogenous Glu due to heteroexchange. Further development of nontransportable compounds, such as DL-TBOA or dihydrokainic acid (DHK), contributed to a more precise identification of transporter properties. DHK is one of the most interesting compounds because of its rather selective binding to EAAT2/GLT-1 but not to EAAT1/GLAST or EAAT3/EAAC1. However, it has the major drawback of also binding to the kainic acid EAA receptor subtype. Conversely, the serine derivative serine-O-sulfate was proposed to inhibit preferentially

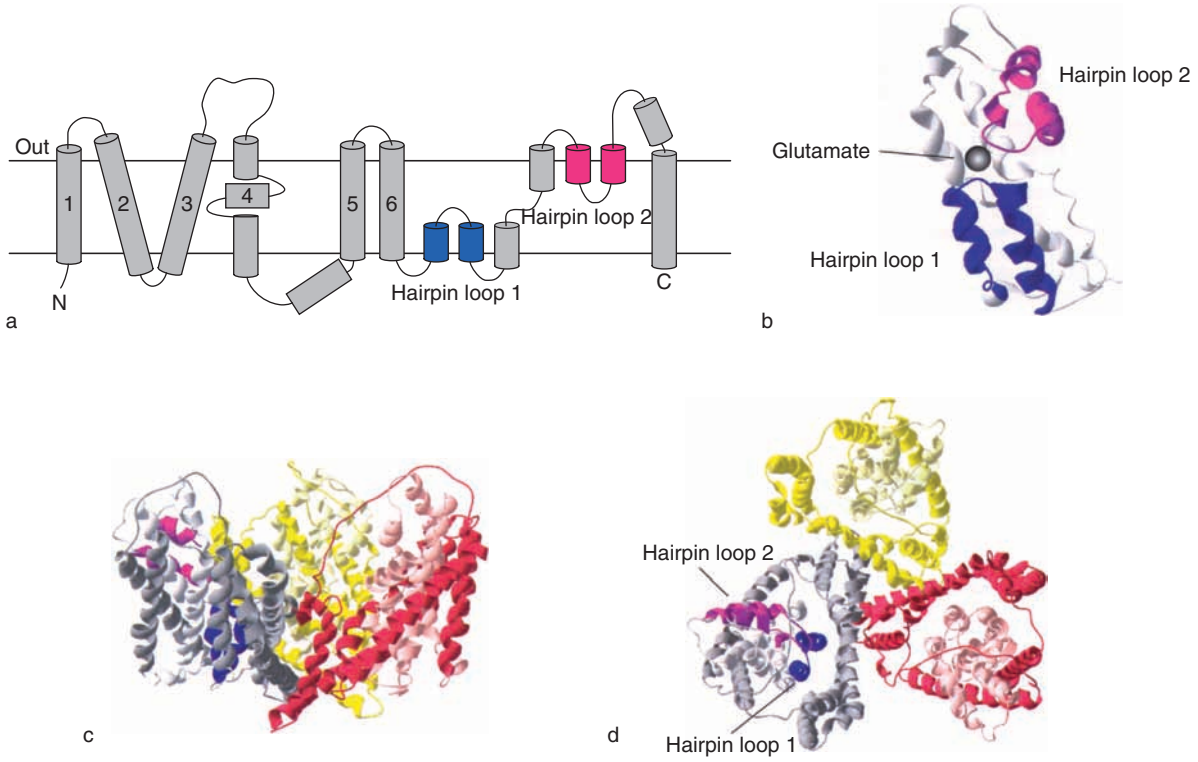

Figure 2 High-resolution structure of the Na^+-coupled glutamate transporter homolog (SLC1 family) from *Pyrococcu horikoshii* (Glt_{Ph}). (a) Schematic representation of membrane topology. 1–6, the transmembrane segments of the protein. (b) Predicted form of the glutamate binding site bordering the two hairpin loops. The hairpin loops could form the gate at both extracellular and intracellular levels to control access to the substrate binding site. (c) View parallel to the membrane. (d) View from the extracellular side. Reproduced from Gether U, Andersen PH, Larsson OM, and Schousboe A (2006) Neurotransmitter transporters: Molecular function of important drug targets. *Trends in Pharmacological Sciences* 27: 375–383, with permission from Elsevier.

EAAT1/ GLAST and EAAT3/EAAC1, and L-β-benzylaspartate was shown to act as a preferential inhibitor of EAAT3/EAAC1 compared to glial transporters.

The Brain Regional Distribution of EAATs Is Uneven and Their Expression Is Developmentally Regulated

In addition to brain expression, mRNA encoding EAAT1/GLAST and, more frequently, EAAT2/GLT-1, has been detected in the heart, kidney, pancreas, bone cells, and mammary glands. However, the functional role of such tissue expression is not completely understood. In the brain, since the demonstration that most Glu uptake activity is attributable to the glial transporters, the key role of astrocytes in EAA transport has been emphasized because EAAT1/GLAST and EAAT2/GLT-1 proteins are selectively expressed by these glial cell subtypes in the adult. Some ependymal cells and neurons in the retina, however, have also been shown to express these transporters. Thus, the debate is still open regarding especially EAAT2/GLT-1 since mRNA cellular localization has been described to occur not only in astrocytes but also in neurons from the adult brain in hippocampus and cortical areas. Moreover, transient EAAT2/GLT-1 expression has been reported in Purkinje cells or in neurons in the spinal cord during development. This is apparently not the case in rat cortical primary neuronal cultures: Astrocytes selectively express EAAT1/GLAST and EAAT2/GLT-1, whereas EAAT3/EAAC1 is expressed in the neurons (Figure 3). Finally, a truncated variant of the EAAT2/GLT-1 protein cloned from mouse liver (GLT-1v) is possibly expressed in astrocytes and certain neurons. However, the level of such neuronal expression of EAAT2/GLT-1 is low, and there is a consensus that EAAT2/GLT-1 expression is generally restricted to astrocytes in the adult brain.

The regional brain distribution of EAAT glial transporters is uneven, in contrast to that for EAAT3/EAAC1. In the rat brain, EAAT1/GLAST is primarily expressed in the cerebellum. The olfactory bulb, hippocampus, and striatal and cortical areas also express the transporter. Moreover, the distribution of the transporter is not restricted to brain areas; the EAAT1/GLAST is also present in the inner ear, for example. EAAT1/GLAST expression is restricted to

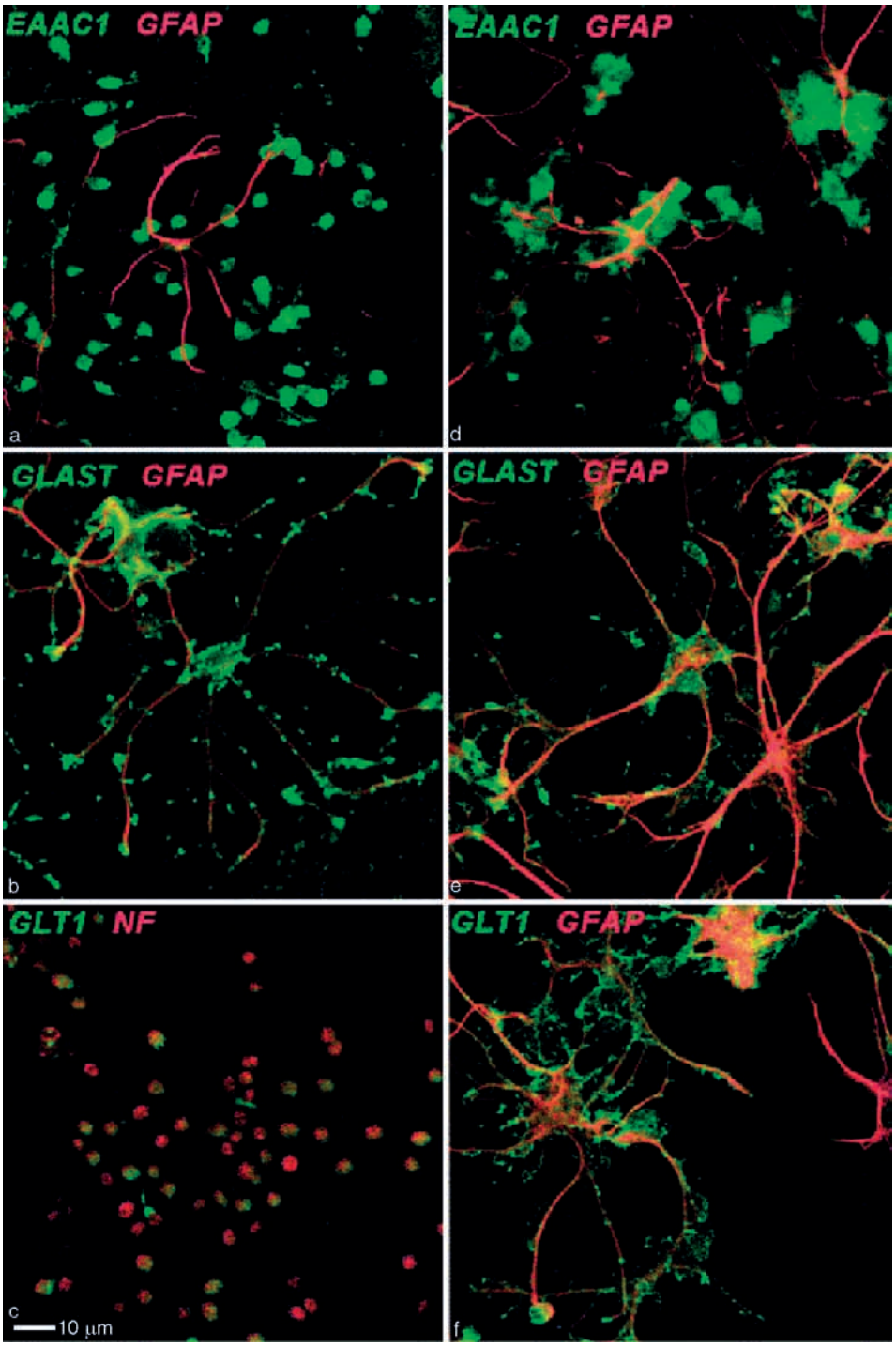

Figure 3 Cell selectivity and developmental profile of the expression of different EAATs in rat primary cortical cultures. Confocal double-labeling immunocytochemistry was used to study the localization of the three main EAA transporter proteins to neurons identified by neuronal neurofilament (NF) labeling or to astrocytes labeled by the glial fibrillary acidic protein (GFAP). (a–c) Data from cortical cultures on day *in vitro* (DIV) 7 treated for double labeling using GFAP or NF antibodies, which are combined on the same preparation with antibodies EAAT3/EAAC1 (a), EAAT1/GLAST (b), and EAAT2/GLT-1 (c), respectively. Data from similar experiments performed on DIV 14 (d–f) confirmed the selective neuronal expression of EAAT3/EAAC1 and astrocytic expression of EAAT1/GLAST and EAAT2/GLT-1. Interestingly, the expression of EAAT2/GLT-1 was only detected at DIV 14, in agreement with data from *in vivo* experiments showing delayed expression of the transporter. Reproduced from Guillet B, Lortet S, Masmejean F, Samuel D, Nieoullon A, and Pisano P (2002) Developmental expression and activity of high affinity glutamate transporters in rat cortical primary cultures. *Neurochemistry International* 40: 661–671, with permission from Elsevier.

astrocytes and Muller cells in the retina. Developmental analysis *in vivo* shows a progressive increase in EAAT1/GLAST expression from birth to adulthood. Conversely, in the newborn rodent EAAT2/GLT-1 brain expression is undetectable, whereas EAAT1/GLAST is present. Thus, the expression of both transporters increases dramatically with synaptogenesis during the first month after birth, which supports the proposal for the essential contribution of EAA to brain maturation. Interestingly, the expression of EAAT1/GLAST in cortical primary cultures occurs in the early stages of the cultures *in vitro* (day *in vitro* (DIV) 3), whereas EAAT2/GLT-1 expression in similar conditions occurs later (DIV 10). In such an experimental model, EAAT3/EAAC1 expression occurs as early as DIV 1. EAAT2/GLT-1 is actually the major EAAT expressed in the forebrain, especially in cortical areas, hippocampus, and striatal areas. However, the transporter is expressed at significant levels in the entire brain. EAAT2/GLT-1 is also present in the retina, where it is expressed in photoreceptor cells and in some bipolar ganglionic neurons. As mentioned previously, brain EAAT2/GLT-1 is expressed at very low levels at birth, and it progressively increases during the second week of life and reaches the adult level after approximately 1 month, in agreement with what is observed in primary cultures. The same astrocytes were shown to be able to simultaneously express EAAT1/GLAST and EAAT2/GLT-1.

Finally, regarding the subcellular localization of the glial transporters in astrocytes, basal expression is preferentially membranar (70%). The density of the proteins is also thought to be higher in the vicinity of glutamatergic synapses. However, this changes with stimulation or pharmacological manipulation, thus showing dynamic regulation of trafficking of the transporters.

Glial EAATs Are Essential Actors of the Excitatory Synapse: Evidence for Multiple Functional Implications

Importance of Glu Transport in Maintaining Synaptic Glu Concentration within Physiological Limits

As mentioned previously, glial EAATs are major contributors to the removal of EAAs and especially Glu from the synaptic space. The relative contribution of the different transporters to Glu uptake was examined *in vivo* using specific antisense oligonucleotides administered in the rat brain to transiently block the expression of the transporter subtypes. Data showed that the extracellular concentration of Glu is highly increased in selected brain structures (hippocampus and striatum). Comparative studies found that EAAT1/GLAST and especially EAAT2/GLT-1 are the major actors of Glu uptake, contributing approximately 80% of EAA uptake in the striatum and 60% in the hippocampus, although the resulting increase in extracellular Glu could be due to cytotoxic effects related to elevated Glu and subsequent excitotoxicity.

Glu transport is believed to be primarily involved in regulating excitatory transmission. Because of the desensitization properties of EAA receptors involved in fast signaling, the rapid removal of the neurotransmitter from the synapse is an essential step of neurotransmission to ensure the functionality of intercellular communication. Interestingly, Glu uptake is thought to also play a key role in limiting the spillover of the excitatory transmitter to EAA receptors located at a distance from the releasing site. Two hypotheses have been proposed. First, limiting the spillover of Glu could contribute to limit activation of receptors of the NMDA subtype present at neighboring synapses and thus to focus excitatory transmission, which may only involve active synapses and not neighboring, so-called 'silent synapses.' Second, because extrasynaptic NMDA activity would presumably contribute to cytotoxicity and neuronal death in contrast to synaptic NMDA receptors, limiting Glu spillover through primary EAA glial uptake could be neuroprotective. In this respect, *in vivo* experiments in the rat using antisense oligonucleotides administered in the brain with miniosmotic pumps showed that the loss of the glial transporters EAAT1/GLAST and EAAT2/GLT-1 produced neurodegeneration in the striatum or hippocampus similar to that observed in excitotoxicity. The progressive motor deficit observed in the treated animals was considered to be the consequence of neuronal cytotoxicity.

However, it has to be taken into account that the lack of glial Glu uptake could have differential effects in the brain regarding regional brain differences in neuronal–astroglial relationships. For example, in the cerebellar cortex, the excitatory synapses on Purkinje cells are closely surrounded by Bergmann glia. Thus, glial processes are in a position to efficiently limit Glu extracellular synaptic diffusion and receptor influence. In contrast, this is not the case at the cortical level and in the hippocampus since very few synapses are entirely surrounded by astrocytes. Consequently, limiting Glu uptake will influence EAA transmission in these brain structures less than in the cerebellar cortex, further suggesting that following synaptic release Glu can rapidly diffuse out of the synaptic cleft and act at a distance.

Glial EAATs Provide Neurons with Glu

Glu uptake into astrocytes provides a major contribution to the so-called 'Glu–glutamine cycle' (Figure 1). Numerous data indicate that Glu released in the synaptic space is taken up by surrounding astrocytes and subsequently transformed in glutamine by glutamine synthetase. The astrocytes release glutamine into the extracellular space, and glutamine is secondarily taken up by excitatory nerve terminals. In the nerve terminals, glutamine is transformed in Glu by the enzyme glutaminase phosphate, which contributes to increase the cytosolic and vesicular pool of the neurotransmitter. Approximately 30% of total Glu entering astrocytes is subjected to oxidative processes and 70% contributes to the Glu–glutamine cycle, which indicates that biosynthesis of Glu continuously occurs in the neurons, supplying nerve terminals with sufficient neurotransmitter. Interestingly, glutamine is not selectively taken up by excitatory terminals: Inhibitory GABA-containing neurons can also take up glutamine. In the inhibitory nerve terminals, glutamine is metabolized into Glu, which represents a metabolic precursor of GABA synthesized from Glu by glutamate decarboxylase. Thus, through the contribution of glial EAATs, the Glu–glutamine cycle contributes to biosynthesis of both the excitatory neurotransmitter Glu and the inhibitory transmitter GABA.

EAA Glial Transporters as Active Contributors to Energetic and Metabolic Processes in Astrocytes and Neurons

It is worth discussing another possible major destination of Glu taken up into astrocytes through glial EAATs. Glu transformed into α-cetoglutarate can be considered as an energetic substrate for the cell. Multiple transamination and oxidative deamination processes contribute to the production of this energetic substrate from Glu. Thus, when necessary, pyruvate produced in the astrocytes can be transferred as lactate to neurons through monocarboxylate transporters and used as pyruvate again in the tricarboxylic acid cycle as an energy source for the cell. The increase in glial Glu uptake occurs concurrently with glucose uptake and consumption and an increase in lactate release, which supports the concept of a net increase in Glu biosynthesis from glucose in astrocytes and increased metabolism when Glu uptake is activated. Lactate provides the energy to the cells. Glu uptake activates the Na^+/K^+-ATPase α_2 subunit, which is specifically expressed in the astrocytes. This mechanism has been suggested to contribute to coupling excitatory neuronal transmission to an increase in astrocytic glucose metabolism. Such a mechanism emphasizes the contribution of glial EAATs to results obtained with functional brain imaging methods (functional magnetic resonance imaging or positron emission tomography). The metabolic response to sensory stimuli is reduced in transgenic mice in which the expression of EAAT1/GLAST or EAAT2/GLT-1 is suppressed or following reduced expression of EAAT1/GLAST using antisense oligonucleotides.

The Glial EAATs as Key Actors for Antioxidative Defenses in Astrocytes

The glial EAATs may contribute to antioxidative defenses in astrocytes. Glu acts as a metabolic precursor of glutathione (GSH), the major neuroprotective agent against the deleterious action of reactive oxygen species (ROS). Glu is metabolized into γ-glutamylcysteine and then into GSH by selective enzymes. Interestingly, the contribution of Glu to GSH biosynthesis is also indirect since Glu is involved in the incorporation of cystine into the astrocytes from outside through the Xc- transport system. The incorporation of cystine provides astrocytes with cysteine used to synthesize GSH. In such a transport system, incorporation of cystine is associated with an efflux of Glu. Consequently, a decreased concentration of Glu in astrocytes may significantly effect the antioxidative defenses of the cell. This occurs when the transport of EAA is pharmacologically blocked: In cell cultures, we measured a progressive and massive death of astrocytes in the presence of PDC, which was correlated with increased production of ROS and further depletion in GSH. We verified that the death of astrocytes in the presence of PDC is neither due to excitotoxicity nor due to oxidative properties of Glu because increasing extracellular Glu concentrations, for example, is protective. Moreover, *N*-acetylcysteine was also shown to exert a protective action against PDC, which further supports the hypothesis that the key contribution of Glu transport into astrocytes is to antioxidative defenses through GSH production.

Further Evidence for an Essential Contribution of Glial EAATs to Brain Homeostasis

Evidences for a functional role of glial EAATs has been obtained from mutations inducing loss of function of the transporters. Mice lacking expression of the Glu transporter EAAT2/GLT-1 showed marked neurodegeneration in the hippocampus and spontaneous seizures. These mice generally died within 6 weeks after birth, further suggesting that such a glial transporter has a key role in controlling excitatory synaptic activity and EAA-induced cytotoxicity. The phenotype resulting from mutation of EAAT1/GLAST showed impaired

motor coordination and increased susceptibility to cerebellar injury, illustrating the importance of the transporter in cerebellar physiology, at least in mice. These experiments emphasize the major contribution to brain functioning of EAAT2/GLT-1 compared to EAAT1/GLAST or EAAT3/EAAC1. Furthermore, in the SOD1 mutant mouse developed as an animal model of some forms of ALS, in which different splice variants of EAAT2/GLT-1 are affected, histological analysis showed alterations in astrocytes concomitant with progressive hind limb paralysis and animal death.

Does Inversion of Glu Uptake Represent a Physiological Process Contributing to Regulation of Excitatory Transmission?

Since not only neurons but also astrocytes exhibit vesicles that contain neurotransmitter and can apparently contribute to exocytosis, the previous observations raise the question of the possible contribution to physiological processes of Glu released by the glial cells. The actual contribution of astrocytes to exocytosis is still under investigation. However, there is another way to obtain an increase in extracellular Glu from these cells. Depending on the electrochemical gradient across the membrane, the transport of Glu into astrocytes can be reversed. Although such a process was primarily assumed to be involved in pathological conditions, there is also evidence that this occurs in physiological conditions in intact cells. The transmembrane gradient of amino acids may be the driving force of the transport, and the reversion of the transport likely results from weakening of such a driving force. The reversed transport in this situation is independent of both ATP and Ca^{2+} ions, and it contributes to increasing extracellular EAA levels, further reinforcing the neurotransmitter action on synaptic receptors and excitatory synaptic signaling when necessary. However, the reversion of Glu transport may be more frequently relevant in pathological conditions.

EAA Transport in Glial Cells Is a Highly Regulated Cellular Process

Fine-Tuning of Glu Uptake Rate at the Excitatory EAA Synapse

In the early 1980s, our laboratory developed the concept of a possible short-term regulation of Glu uptake rate with regard to changes in neuronal activity. Such a concept of activity-dependent regulation of neurotransmitter uptake rate was further extended since modulations of Glu uptake have been shown in the rat, for example, in relation to stress or even estrous cycle. Numerous studies have supported the proposal for a fine-tuning of EAA transport, especially in astrocytes. Regarding EAAT1/GLAST and EAAT2/GLT-1, numerous data have been obtained suggesting regulatory processes which involve both transcriptional and posttraductional mechanisms. This latter hypothesis was also supported by experiments from our group showing that Glu uptake may be modulated in response to activation of neurotransmitter receptors. Furthermore, numerous trophic factors have been shown to increase Glu uptake. Arachidonic acid, NO (indirect action through cytokines production), Zn^{2+} ions, as well as pharmacological activation or inhibition of various protein kinases were conversely shown to decrease Glu uptake rate, supporting the concept of posttraductional regulation of the activity of transporter molecules. Such a concept was further supported by the demonstration that EAATs exhibit putative consensus phosphorylation sites in their molecular structure. Finally, it is worth mentioning the special role of oxidative conditions in altering Glu transport. Oxidative conditions severely decrease transporter activity, possibly through oxidization of the cysteine residues of EAAT1/GLAST and EAAT2/GLT-1. This occurs in the presence of ROS in the cells, further emphasizing the deleterious role of oxidative processes on the functionality of the transporter molecules in addition to the action of cell integrity.

Regulation of Cell Expression of Glial EAATs

In addition to evidence for the regulation of Glu transport rate, numerous experiments have also shown that EAA transport can be influenced at the transcriptional level. Astrocytes in culture provided a useful model to assess such regulatory influence on EAAT expression. Data have shown, for example, that EAAT2/GLT-1 expression is induced by various trophic factors, such as epidermal growth factor, transforming growth factor-α, fibroblast growth factor-2, platelet-derived growth factor, or even brain-derived neurotrophic factor, suggesting a contribution of tyrosine kinase pathways. Interestingly, insulin-like growth factor-2 was evidenced to maintain cellular EAAT1/GLAST and EAAT2/GLT-1 expression. Transcriptional effects on EAAT genes thus primarily involve protein kinase (PK) activation. For example, in the HEK293 cell line, increasing PKC activity induced a decreased expression of EAAT1/GLAST. Conversely, activation of PKC in HeLa cells induced an increase in EAAT2/GLT-1 expression; in this case, the authors originally demonstrated that the transcriptional effect was due to direct phosphorylation of the transporter. Co-culturing astrocytes together with neurons also increased the expression of the two glial transporters. Different mechanisms have been proposed to explain such

a neuronal influence on astrocytes, possibly involving cAMP, CREB transcription factor, MAP/Erk kinase signaling pathway, or pituitary adenylate cyclase-activating peptide.

Additional Regulatory Processes of Glial EAATs Acting on Intracellular Trafficking of the Transporter Protein and Membrane Stabilization

A recent view of the processes involved in the regulation of EAA transport is related to changes in the trafficking and stabilization of the transporter proteins at the membrane from intracellular stores. Such a mechanism is well documented for the neuronal transporter EAAT3/EAAC1, which is subjected to numerous influences. For example, PKC activation facilitates the association of the transporter with the αPKC subunit, which contributes to stabilization of the transporter at the membrane. Biotinylation and immunoblotting experiments show similar processes for the glial transporters. Interestingly, EAAT1/GLAST and EAAT2/GLT-1 are differentially regulated but possibly expressed in the same astrocyte. For example, inhibition of PKA activity using the compound H89 or stimulation of PKC using PMA both resulted in decreased EAAT1/GLAST at the astrocytic membrane, whereas wortmannin acting as a specific inhibitor of PI3 kinase produced the opposite response. Conversely, H89 induced an increase in EAAT2/GLT-1 at the membrane, and PMA and wortmannin produced the opposite effect. These results illustrate differential posttraductional regulation of the two transporters. Interestingly, assuming that increasing the presence of the transporter at the membrane correlates with increased EAA uptake activity, in our experimental model *in vitro* (primary cultures from rat cerebral cortex) PI3K inhibition using wortmannin is only active to reduce the uptake process at 14 DIV, when EAAT2/GLT-1 is fully active. At 7 DIV, wortmannin is inactive but EAAT2/GLT-1 is not yet expressed by the astrocytes.

The molecular mechanisms by which the different PKs contribute to transporter trafficking, membrane stabilization, and cytosolic sequestration remain obscure. As in the case of EAAT3/EAAC1, regulatory cytosolic proteins may be involved. Since the protein septin 2 acting as a GTPase has been shown to interact with the C-terminal part of the EAAT1/GLAST protein, septin 2 may represent one of the cytosolic regulatory factors for stabilization of the transporter at the membrane and activation of astroglial Glu uptake rate. Similarly, the protein LIM associated with the N-terminal part of EAAT2/GLT1 may contribute to changes in its intracellular trafficking.

In summary, the concept of fine tuning the uptake of EAAs to further regulate extracellular concentrations of the neurotransmitter with regard to nerve activity is well documented. Regulatory processes involve both transcriptional and posttraductional mechanisms and have been demonstrated to influence Glu uptake rate. Interestingly, the activity of the excitatory neurons may be a critical factor for regulatory processes. Such a proposal is further reinforced by the fact that treatment of cell cultures with high concentrations of NMDA to activate Glu transmission downregulates EAAT2/GLT-1 expression but upregulates EAAT1/GLAST expression, further illustrating differential regulatory mechanisms for both transporters. Conversely, low extracellular concentration of Glu increases EAAT2/GLT-1 expression in astrocytic cultures. Possible EAAT2/GLT-1 expression in neurons has also been reported, but because the uptake process is not sensitive to DHK, it is likely that the transporter is inactive or represents an inactive variant of EAAT2/GLT-1.

Glial EAATs Are Likely Involved in Neurodegenerative Diseases

Numerous observations primarily from animal models have implicated alterations of Glu transport in various brain diseases (Table 1). EAAs have been associated with acute brain lesions related to stroke (ischemic, anoxic, and traumatic events) and to the so-called functional diseases involving an excess of excitatory transmission, such as different states of epilepsy or, possibly, some forms of schizophrenia. EAAs have also been proposed to be involved in numerous neurodegenerative diseases, such as Huntington's disease, Parkinson's disease, or even Alzheimer's disease, although direct evidence for the contribution of EAAs is generally lacking, except for ALS. The involvement of EAAs in brain pathological events has also been proposed in HIV-associated dementia, malignant glioma, or hepatic encephalopathy due to an increased level of ammonia in the brain since decreased expression of EAAT2/GLT-1 has been reported.

Putative Contribution of EAAT2/GLT-1 to ALS

Approximately 10 years ago, a selective loss of EAAT2/GLT-1 isoforms in motor and sensory cortical areas and concomitant decreased Glu transport activity in patients suffering sporadic ALS was shown. Aberrant splicing of the EAAT2/GLT-1 transcript, and also vulnerability of the protein to oxidative stress, was suggested to be the cause of decreased EAAT2/GLT-1 expression/activity. Furthermore, a mutation which substitutes an asparagine for a serine at position 206 in the EAAT2/GLT-1 gene (mutation N206S) was reported in sporadic ALS. Because this

Table 1 Putative implication of excitatory amino acid transporters and brain pathologies highlighted from experimental approaches

Disease	*Experimental model*	*Transporter subtype*	*Resulting phenotype*
ALS	SOD1 mutants	EAAT2/GLT-1 reduced	Progressive hind limb paralysis and death
	EAAT2/GLT-1 antisense strategy in rat	EAAT2/GLT-1 reduced	Progressive motor neuron loss and hind limb paralysis
Alzheimer's disease	Mutant APP over expression in mice	EAAT1/GLAST and EAAT2/GLT-1 protein reduced (mRNA levels normals)	Amyloid plaque formation; behavioral changes
	Cultured rat astrocytes	Peptide β-amyloid reversal of Glu uptake	
Parkinson's disease	Nigro-striatal dopamine denervation in rat	L-DOPA induced increase in striatal EAAT2/GLT-1 protein	
	Striatal dopamine denervation	EAAT1/GLAST and EAAT2/GLT-1 striatal reduction	
Huntington's disease	Expression of mutant huntingtin	Reduction of EAAT2/GLT-1 mRNA and striatal protein	
Epilepsy	EAAT2/GLT-1 knockout mice	EAAT2/GLT-1 protein reduced	Seizures and death at 6 weeks
	EAAT1/GLAST knockout mice	EAAT1/GLAST protein reduced	Increased sensitivity to convulsive agents
	EAAT3/EAAC1 antisense strategy	Reduced EAAT3/EAAC1 and GABA	EEG episodes with seizures
Stroke	Seven-day-old-rat hypoxia–ischemia (cortex; hippocampus)	All the three main transporters decreased at 12 h	
	Hippocampal slices from EAAT2/GLT-1 knockout mice	Reduced Glu uptake but no reversal related to EAAT2/GLT-1	
	Rat MCA occlusion	Decreased EAAT2/GLT-1 and EAAT3/EEAC1	Contralateral hemiparesis Contralateral hemiparesis
	Rat MCA occlusion	EAAT1/GLAST increased	
	Rat MCA occlusion with antisense strategy to EAAT1/GLAST and EAAT2/GLT-1	The two transporter proteins are reduced	45% increase in stroke volume and increased mortality
	Hypoxic neonatal pig	EAAT2/GLT-1 and EAAT3/EAAC1 reduced at 24 h (striatum)	

Adapted from Maragakis NJ and Rothstein JD (2004) Glutamate transporters: Animal models to neurological diseases. *Neurobiology of Diseases* 15: 461–473.

mutation was demonstrated to alter Glu transport, it appeared that the excess of Glu in the extracellular space along with clearance deficiency of the neurotransmitter contributed to excitotoxicity and induced the death of motor neurons. The supranormal levels of Glu measured in the blood and possibly in cerebrospinal fluid from patients may be indicative of changes in EAA metabolism in ALS. The putative contribution of EAAs to neuronal death in ALS was further emphasized by the positive effect of the antiglutamatergic compound riluzole, which was shown to slow the rate of the disease.

The case of ALS pointed to the possible contribution of the EAAT system, and endogenous Glu, to the neurodegenerative processes. Because of the main glial expression of such transporters, results from ALS patients also emphasize the possible contribution of glial cells and especially astrocytes to motor neuron degeneration. Finally, the modest but real clinical effects of riluzole in ALS also point to the glutamatergic synapse as a possible target for development of further neuroprotective strategies. In the SOD1 mutant mice in which the function of EAAT2/GLT-1 is highly reduced, damaged astrocytes are detected in the spinal cord. Such astrocytic alteration is concomitant with motor neuron degeneration, which supports impaired locomotion. In this respect, as mentioned previously, increasing extracellular Glu levels using chronically implanted osmotic minipumps for local injections of the EAAT inhibitor PDC *in vivo* resulted in a loss of striatal neurons and astrocytes in the rat. The critical role of EAATs in neurodegeneration was also evidenced in the pioneer experiments of Rothstein and group, who elegantly showed that decreasing glial EAAT expression using antisense oligonucleotides in the rat resulted in elevated extracellular Glu levels, neurodegeneration, and progressive paralysis. As also mentioned previously, contribution of glial EAATs to

brain pathology was obtained from transgenic mice lacking expression of the transporters showing impaired motor coordination, cerebellar alterations, and spontaneous seizures depending on the transporter expressed. Interestingly, the decreased antioxidative defenses in astrocytes subjected to Glu depletion would explain the death of astrocytes. Therefore, such a hypothesis may be proposed for ALS in humans and for neurodegenerative processes in the SOD1 mutant mice. If such a hypothesis is correct, the death of motor neurons may therefore be secondary to lack of Glu transport in astrocytes and subsequent excitotoxicity.

Do the Glial EAATs Contribute to Alzheimer's Disease?

With the exception of ALS, the contribution of EAATs in other neurodegenerative diseases is not so evident. In Alzheimer's disease, it is generally accepted that the level of EAATs is normal, although specific reductions of EAAT2/GLT-1 have been reported in frontal cortex. However, abnormal neuronal expression of the two glial transporters has also been reported, notably in neurons concerned with tau protein accumulation. Interestingly, *in vitro* experiments have shown an inhibitory effect of the β-amyloid peptide on Glu transport, further indicating that in Alzheimer's disease the pathological process can contribute to neuronal death through impaired EAA transport. Consequently, the negative influence of the β-amyloid peptide on Glu transport may be viewed as initiating a 'cytotoxic cascade' which may be amplified by direct overstimulation of EAA receptors mediating neuronal death. The associated Glu astrocytic depletion, which may therefore decrease oxidative defenses of the cell, is then thought to induce astrocyte death and secondarily contribute to increase extracellular Glu concentrations since astrocytes cannot proceed to EAA uptake. However, such a hypothesis remains to be documented, and the contribution of EAAs to pathological processes in Alzheimer's disease has not been proved.

Brain Ischemia and Anoxia: Does Reversal of EAATs Represent One of the Key Steps of Degenerative Processes?

The involvement of EAATs in brain pathology may be related to a decreased uptake rate, decreased transporter tissue expression and/or altered trafficking of transporter proteins, and their putative implication in the transport of cytotoxic substances and also in the reversal of the transport direction, which likely contributes to increase extracellular Glu levels and decrease Glu astrocytic contents. EAA neuronal toxicity in severe ischemia is generally accepted to result from reversal of EAATs, especially the neuronal transporter EAAT3/EAAC1. However, the implication of the glial transporters is also evident since the driving force of transport is related primarily to the electrochemical gradient of ions associated with the uptake process in general. For example, the increase in extracellular Glu in experimental anoxic conditions is highly sensitive to DHK, suggesting a major contribution from EAAT2/GLT-1. Transient or more permanent disruptions of such a gradient in the case of deficient energy supply, as observed in ischemia or anoxia processes, will decrease EAA uptake and induce reversed Glu transport. Indeed, ATP depletion will result in rapid Na^+/K^+ ATPase loss of activity and reversal of the Na^+/K^+ gradient. The contribution of neurons to the increased extracellular Glu is related to their high content in the neurotransmitter compared to astrocytes, but depleting intracellular Glu stores in astrocytes is likely to induce a rapid decrease in GSH levels, thus reducing the antioxidative defenses of the cells. In this respect, in the ischemic area, cell death involves both neurons and glial cells as shown in many reports. Consequently, pharmacologically blocking the Glu transport to prevent reversal of the transport of Glu without altering uptake capacities may be of therapeutic interest to maintain low extracellular Glu in stroke.

Considerable delayed alterations of the EAATs have been observed in experimental models of ischemia following reestablishment of the blood supply. Although there is great variability due to major differences in animal models and experimental conditions *in vivo* and *in vitro*, the general feature is a decreased expression of most of the transporters, both glial and neuronal, until 72 h following the ischemic procedure. Interestingly, during the postischemic period abnormal neuronal expression of the glial transporters EAAT1/GLAST and EAAT2/GLT-1 as well as overexpression of EAAT1/GLAST have been reported in the region surrounding the ischemic areas, further supporting the idea that EAATs are involved either in long-lasting neurodegenerative processes following brain ischemia or in compensatory protective mechanisms that limit cellular damage.

Conclusion

Although direct evidence is lacking, numerous data from experimental models point to the contribution of EAAs to pathophysiological processes in several neurological and psychiatric disorders. However, the concept of the possible involvement of EAATs in such pathological processes is quite recent; for years, the involvement of EAAs was focused on excitotoxicity and the EAA receptors. Of course, excitotoxicity can

represent one of the final steps of neuronal death, but astrocytes are generally reported to be insensitive or less sensitive than neurons to excitotoxicity. Thus, the contribution to EAATs to degenerative processes in general is not restricted to simple impaired uptake into either neurons or glial cells. The dependence of astroglial cell survival on maintenance of sufficient intracellular Glu levels illustrates the key contribution of EAATs. Defective transport results in GSH depletion and increased susceptibility to oxidative stress, likely resulting in astrocytic death. Stimulating Glu uptake may therefore be a way to protect the neurons. Finally, because of the dependence of EAA transport on the electrochemical gradient, impaired energy results in reversal of the transport, which consequently can increase both excitotoxicity and astrocyte depletion in Glu. Therefore, the conservation of functional glial EAATs could contribute to neuroprotection. In this respect, developing new compounds aimed at keeping such EAATs active may be a useful strategy in the case of neurodegenerative diseases. Experiments in progress will contribute to developing this concept.

See also: Excitotoxicity in Neurodegenerative Disease; Glial Glutamate Transporters: Electrophysiology; Glial Ion Homeostasis: A Fluorescence Microscopy Approach; Glial Glutamate and GABA Metabolism; Glutamate; Ionic Channels in Glia.

Further Reading

Bergles DE, Diamond JS, and Jarh CE (1999) Clearance of glutamate inside the synapse and beyond. *Current Opinion in Neurobiology* 9: 293–298.

Bonvento G, Sibson N, and Pellerin L (2002) Does glutamate image your thoughts? *Trends in Neurosciences* 25: 359–364.

Bridges RJ and Esslinger CS (2005) The excitatory amino acid transporters: Pharmacological insights on substrate and inhibitor specificity of the EAAT subtypes. *Pharmacological Therapeutics* 107: 271–285.

Danbolt NC (2001) Glutamate uptake. *Progress in Neurobiology* 65: 1–105.

Dringen R, Gutterer JM, and Hirrlinger J (2000) Glutathione metabolism in brain metabolic interaction between astrocytes and neurons in the defense against reactive oxygen species. *European Journal of Biochemistry* 267: 4912–4916.

Gadea A and Lopez-Colome AM (2001) Glial transporters for glutamate, glycine and GABA: I. Glutamate transporters. *Journal of Neuroscience Research* 15: 453–460.

Gether U, Andersen PH, Larsson OM, and Schousboe A (2006) Neurotransmitter transporters: Molecular function of important drug targets. *Trends in Pharmacological Sciences* 27: 375–383.

Guillet B, Lortet S, Masmejean F, Samuel D, Nieoullon A, and Pisano P (2002) Developmental expression and activity of high affinity glutamate transporters in rat cortical primary cultures. *Neurochemistry International* 40: 661–671.

Hertz L and Zielke R (2004) Astrocytic control of glutamatergic activity: Astrocytes as stars of the show. *Trends in Neurosciences* 27: 735–743.

Hinoi E, Takarada T, Tsuchihashi Y, and Yoneda Y (2005) Glutamate transporters as drug targets. *Current Drug Targets – CNS Neurological Disorders* 4: 211–220.

Kanai Y and Hediger MA (2003) The glutamate and neutral amino acid transporter family: Physiological and pharmacological implications. *European Journal of Pharmacology* 479: 237–247.

Maragakis NJ and Rothstein JD (2004) Glutamate transporters: Animal models to neurological diseases. *Neurobiology of Diseases* 15: 461–473.

Nieoullon A, Canolle B, Masmejean F, Guillet B, Pisano P, and Lortet S (2006) The neuronal excitatory amino acid transporter EAAC1/EAAT3: Does it represent a major actor at the brain excitatory synapse? *Journal of Neurochemistry* 98: 1007–1018.

Pellerin L and Magistretti P (2004) Neuroenergetics: Calling upon astrocytes to satisfy hungry neurons. *Neuroscientist* 10: 53–62.

Ré DB, Boucraut J, Samuel D, Birman S, Kerkerian-LeGoff L, and Had-Aissouni L (2003) Glutamate transport alteration triggers differentiation-state selective oxidative death of cultured astrocytes: A mechanism different from excitotoxicity depending on intracellular GSH contents. *Journal of Neurochemistry* 85: 1159–1170.

Ré DB, Nafia I, Melon C, Shimamoto K, Kerkerian-LeGoff L, and Had-Aissouni L (2006) Glutamate leakage from a compartmentalized intracellular metabolic pool and activation of the lipoxygenase pathway mediate oxidative astrocyte death by reversed glutamate transport. *Glia* 54: 47–57.

Rothstein JD, Dykes-Hoberg M, Pardo CA, et al. (1996) Knockout of glutamate transporters reveals a major role for astroglial transport in excitotoxicity and clearance of glutamate. *Neuron* 16: 675–686.

Seifert G, Schilling K, and Steinhäuser C (2006) Astrocyte dysfunction in neurological disorders: A molecular perspective. *Nature Reviews Neuroscience* 7: 194–206.

Shigeri Y, Seal RP, and Shimamoto K (2004) Molecular pharmacology of glutamate transporters, EAATs and VGLUTs. *Brain Research Reviews* 45: 250–265.

Glial Glutamate Transporters: Electrophysiology

M P Kavanaugh, University of Montana, Missoula, MT, USA

© 2009 Published by Elsevier Ltd.

Molecular Diversity and Expression Patterns

The mammalian glutamate transporter gene family has five excitatory amino acid transporter (EAAT) members (*EAAT1–EAAT5*) that are expressed in the brain in highly specific regional and cell-type patterns (see Table 1). In some cases, transporter transcripts undergo alternative splicing, adding further complexity to these expression patterns. EAAT cDNAs have predicted amino acid sequences of approximately 500–600 residues, and subtypes share between 30% and 65% identity; typically greater than 80% identity is shared when comparing a particular subtype across species. Together with two structurally related neutral amino acid transporter genes (alanine/serine/cysteine transporter genes *ASCT1–ASCT2*), they make up a gene superfamily whose members are structurally distinct from other solute transporters, such as the vesicular glutamate transporters (GLTs) and the family of Na^+/Cl^--dependent neurotransmitter transporters responsible for uptake of GABA, glycine, serotonin, and catecholamines. Distinct EAAT subtypes are found on neurons and glia throughout the brain and spinal cord, and different functional roles are emerging for each. In the central nervous system (CNS), the vast majority of glutamate uptake activity is found on glial cell membranes (predominantly EAAT2/GLT-1 in forebrain and spinal cord, and EAAT1/GLAST in cerebellum). These glial transporters are densely packed in the membrane (up to $8500\,\mu m^{-2}$) and are often found in close apposition to neuronal synapses. Table 1 provides an overview of the nomenclature and expression patterns of the transporters.

Molecular Structure

Glt_{Ph} is an aspartate/glutamate transporter homolog from *Pyroccocus horikoshii* that exhibits approximately 35% identity to mammalian EAATs overall, with higher homology in domains that have been implicated in key functions such as substrate and ion binding. The three-dimensional structure of Glt_{Ph} has been determined by X-ray crystallography, revealing a rather complex trimeric bowl-shaped architecture. Each of the three subunits has eight transmembrane domains with two reentrant helical hairpin loops (HP1 and HP2) that dip into the membrane from opposite sides. Structural as well as functional evidence suggests that each subunit in the trimer binds substrate: in the crystal structure a molecule of bound aspartate is occluded between the tips of the two reentrant loops, with Na^+ ions above and below it. In an apo-state transporter crystallized without bound amino acid, the diffraction pattern reveals that the external reentrant loop (HP2) is swung outward approximately 10 Å, suggesting a possible docking trajectory and structural gating mechanism for the first hemicycle of an alternating access kinetic scheme (see later). In the occluded state, charge pairing between the bound aspartate and residues in the transporter stabilize the complex. A key determinant of substrate selectivity is a conserved arginine residue in transmembrane domain 8 (TMD8) that is coordinated with the distal carboxylate of the transported substrate (Figure 1). In the related ASC transporters, which selectively transport neutral rather than acidic amino acids, a cysteine or threonine residue is found in the analogous position in TMD8. Mutating the neutral residue to arginine in the ASC transporters or making the converse mutation in the EAATs switches the transporters' selectivity correspondingly.

All of the EAATs are stereoselective for L-glutamate, while both enantiomers of aspartate are transported. A number of glutamate and aspartate analogs, including some ring-constrained analogs such as pyrrolidine dicarboxylates (PDCs), are transported by the EAATs. More bulky aromatic analogs such as DL-*threo*-β-benzyloxyaspartate (TBOA) competitively inhibit glutamate transport but are not themselves transported. With the exception of kainate and dihydrokainate, which block EAAT2 with an affinity three to four orders of magnitude higher than the other transporter isoforms, there are no compounds available with which to distinguish the transporter subtypes with a high degree of selectivity *in vivo*.

Thermodynamics of Glutamate Transport

The influx of Na^+ and H^+ and the efflux of K^+ during glutamate uptake result in a net flow of positive charge into the cell that can be recorded and studied with a voltage-clamp circuit. A tight stoichiometric coupling is inferred from the effects of these cations' concentration gradients on the reversal potential of the pharmacologically isolated transport current. The transporter reversal potential (the equilibrium membrane potential at which there is no net transport) follows the predictions of the free energy equation for coupled transport of one glutamate molecule

Table 1 Glutamate transporter nomenclature and expression patterns

Human	*Other*	*IUBMB nomenclature*[a]	*Predominant expression pattern*
EAAT1	GLAST	SLC1A3	Astrocytic; cerebellum (Bergmann glia); retina (Müller cells); inner ear
EAAT2	GLT-1	SLC1A2	Astrocytic (most abundant transporter); brain and spinal cord, low levels in some neuron terminals
EAAT3	EAAC1	SLC1A1	Predominantly neuronal (widely expressed but levels much lower than EAAT1/2); predominant peripheral glutamate transporter
EAAT4	EAAT4	SLC1A4	Neuronal; cerebellum (Purkinje neurons)
EAAT5	EAAT5	SLC1A5	Retina: photoreceptors and bipolar cells, astrocytes (Müller cells)

[a]International Union of Biochemistry and Molecular Biology.

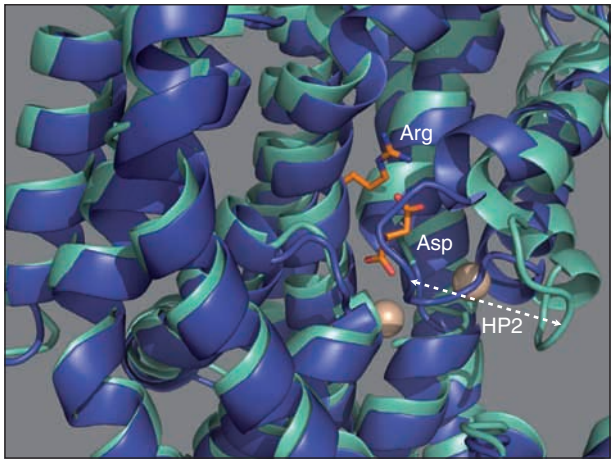

Figure 1 Superimposed structures of the *Pyroccocus horikoshii* aspartate/glutamate transporter homolog transport states (see text for discussion) T_o (green) and T_{occ} (blue), highlighting translation of hairpin 2 (HP2) loop going from empty T_o state to the occluded state. Bound substrate (aspartate) interaction with arginine residue in transmembrane domain 8 is shown, with two Na^+ ions (spheres) below. Python-enhanced molecular graphics tool representation of Protein Data Bank coordinates 2NWX and 2NWW.

with one proton and three Na^+ ions and countertransport of one K^+ ion during each uptake cycle. The equilibrium glutamate concentration gradient that can be maintained by stoichiometrically coupling movement of n Na^+, K^+, and H^+ ions is given by rearrangement of the zero-flux equation:

$$(\mathrm{Glu}_o/\mathrm{Glu}_i) = (\mathrm{Na}_i/\mathrm{Na}_o)^{nNa}(\mathrm{H}_i/\mathrm{H}_o)^{nH}(\mathrm{K}_o/\mathrm{K}_i)^{nK} \exp[(n_{\mathrm{Na}} + n_{\mathrm{H}} - n_{\mathrm{Glu}} - n_{\mathrm{K}})VF/RT] \quad [1]$$

where V= membrane potential, F = Faraday's constant, R = gas constant, and T = temperature. For astrocytes and neurons, this value exceeds 10^6; the free energy in the Na^+, K^+, and H^+ electrochemical gradients is sufficient to maintain nanomolar extracellular glutamate concentrations despite millimolar intracellular glutamate concentrations. Estimates of extracellular glutamate concentrations using dialysis probes in brain tissue suggest low micromolar values, but electrophysiological measurements in hippocampus are consistent with ambient glutamate levels below 100 nM. When ionic gradients are impaired, with elevated extracellular K^+ and intracellular Na^+, as can occur during severe ischemia, transporter reversal and efflux of intracellular glutamate may occur. Another distinct mode of transport, electroneutral glutamate exchange, can occur in the absence of K^+ when Na^+ and glutamate (or another substrate) are present on both sides of the membrane. This exchange can be monitored by measuring the unidirectional flux of radioactivity when radiolabeled substrate is added to one side of the membrane. The glutamate exchange mode can also be induced by mutations in the transporter near the substrate binding site that selectively interfere with K^+ binding (e.g., E404D; GLT-1 numbering).

Mechanism and Kinetics

Mechanistically, the observed transport modes are consistent with high-resolution glutamate transporter structures that appear to represent distinct kinetic states. Together, the structural and functional data suggest that the transporter cycles through three major states: one occluded state (T_{occ}), and two distinct states (T_o and T_i) that provide gated access of substrates to their binding sites from outside or inside the cell, respectively. The structures of the first two of these states have been solved (**Figure 1**) and the existence of the third state (T_i) is only inferred from kinetic data at this point. There is no evidence for a state with outer and inner gates simultaneously open (i.e., an open channel permeable to glutamate). Access for intra- and extracellular substrates appears to occur on a strictly alternating basis, providing a basis for tight thermodynamic coupling of glutamate and cation fluxes. A further tenet of the model supported by functional data is that Na^+ and K^+ binding are mutually exclusive. Na^+ and glutamate binding will occur in a coordinated fashion and their

occupancy precludes binding of a K^+ ion. Conversely, binding of a K^+ ion precludes Na^+ and glutamate binding. Kinetic evidence, primarily from the Na^+ and glutamate concentration-dependence of mammalian EAATs, suggests that binding of two Na^+ ions precedes glutamate binding, followed by the third Na^+ ion. The postulated sequence of events for one thermodynamic cycle (arbitrarily starting with the glutamate hemicycle) is as follows. Glutamate/Na^+/H^+ bind from outside (or inside) to a transporter in state T_o or T_i, respectively, thereby preventing binding of a competing K^+ ion. The two major states, T_o and T_i, are kinetically linked to each other through the occluded state. Unbinding of glutamate/Na^+/H^+ can occur from either state T_o or T_i after transition out of the occluded state; then a competition with K^+ for rebinding occurs. For a complete transport cycle to occur, the first hemicycle is completed when glutamate/Na^+/H^+ are translocated by unbinding from the side opposite to which they bound. A K^+ ion then binds and proceeds back through the same kinetic steps, completing the second hemicycle by its unbinding from the opposite side and restoring the initial state. The free energy change for the cycle is equal to the sum of the free energy changes for the individual ions. All of these kinetic transitions are reversible; the net flow through the states depends only on the transition rate constants, the concentration gradients for Na^+, glutamate, H^+, and K^+, and the membrane potential. The conformational transition of the unliganded transporter between the T_i and T_o states does not appear to proceed at a measurable rate; possibly occupancy by substrate is necessary to stabilize an unfavorable transition state. Overall, the mechanism bears notable similarity to the Na/K pump, with the primary difference being that the vectorial nature of transport in this case is a consequence of electrochemical gradients rather than ATP hydrolysis.

Studies of glutamate transport kinetics have utilized both steady-state and pre-steady-state techniques. The latter include analysis of transient currents following rapid concentration or voltage jumps. In transporters covalently labeled with fluorescent reporter probes, fluorescence changes can also be measured in response to perturbations of voltage or concentration. A prominent capacitive current (or fluorescence change) reflecting charge movements is present in the transporters that is analogous to gating charge movements seen in voltage-gated channels. This charge movement is Na^+ dependent, and it occurs in the absence of glutamate, providing evidence that Na^+ binding occurs prior to glutamate. The transient current or fluorescence change can be described by a Boltzmann function, similar to channel gating charge movements, and in the case of the transient current it can be used to estimate the number of transporters in a cell or patch. Using this approach, the turnover rate of the major glial transporter EAAT2 expressed in *Xenopus* oocytes is approximately $15\,s^{-1}$ at room temperature, or approximately $50\,s^{-1}$ at 37°, because of the transporter's temperature dependence. Because many rates in the transport cycle are voltage dependent, the turnover rate is also accelerated by hyperpolarizing membrane potentials (approximately e-fold per 75 mV in the case of EAAT2). Other techniques used to gain information about microscopic transport kinetics include rapid photo-uncaging or piezo-switched application of glutamate to cells or excised patches containing transporters. In the presence of extracellular Na^+ and intracellular K^+, a glutamate pulse produces a fast inward current that rapidly relaxes to a lower steady-state level (**Figure 2(a)**). This reflects fast glutamate binding ($\sim 10^7 M^{-1} s^{-1}$) and rapid charge movements across the electric field in the early steps of the transport cycle. Using paired concentration jumps applied to excised patches to measure the kinetics of recovery of the transient peak current, EAAT2 turnover rates exceeding $50\,s^{-1}$ have been estimated. This faster turnover (compared with steady-state oocyte measurements) likely reflects the absence of intracellular Na^+ and glutamate, which will slow the transport cycle in oocytes (as well as glia and neurons) because of the aforementioned microscopic kinetics. Even at 20 ms, the transport cycle time is roughly an order of magnitude slower than the estimated synaptic glutamate concentration decay. Nevertheless, the transporters near a release site may still theoretically influence the spatiotemporal profile of glutamate on a very fast timescale because the early steps in the transport cycle allow rapid glutamate capture and the transporters are present at high density

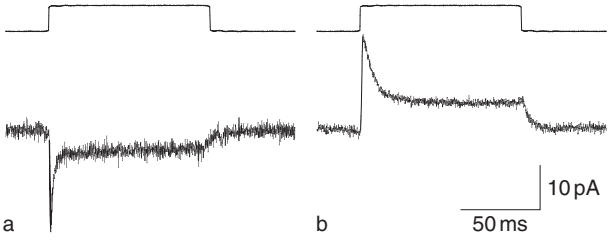

Figure 2 Outside-out patch recordings of glial transporter EAAT2, showing anion channel and glutamate transport kinetic responses to piezo-switched application of saturating (10 mM) glutamate. (a) Response predominantly reflects stoichiometrically coupled currents; (b) reflects anion currents carried by SCN^-. Pipette solution contained K-gluconate; bath contained NaCl (a) or NaSCN (b). Data are from the same patch held at −20 mV; trace above shows solution exchange time course measured following patch rupture.

(10^3–$10^4\,\mu m^{-2}$) on astrocyte membranes. Furthermore, more remote transporters surrounding synaptic release sites can play roles in restricting the heterosynaptic spillover of glutamate that occurs on a slower timescale (see later).

Chloride Channel Function

The aforementioned fluxes of stoichiometrically coupled ions through the transporter are not the only charge movements mediated by EAATs; chloride flux through the transporters also occurs. This was first suggested by the presence of a chloride current associated with activation of the native glutamate transporters in retinal neurons, and it has subsequently been demonstrated to occur with EAATs exogenously expressed in many different cell systems. Interestingly, this channel function can play roles in synaptic signaling that are entirely distinct from the effects of uptake on glutamate dynamics. Glutamate transporters (predominantly EAAT5) on the presynaptic terminals of rod bipolar cells mediate inhibitory signaling that serves to modulate the release of transmitter in a negative feedback loop. Several lines of evidence suggest that a chloride channel is intrinsic to all glutamate transporters. Mutation of residues clustered in TMD2 leads to discrete changes in anion selectivity. In addition, the net current activated by glutamate (reflecting both the stoichiometrically coupled and Cl^- currents) has a distinct reversal potential in each EAAT isoform, suggesting that each has a fixed and unique Cl^- current magnitude relative to the stoichiometrically coupled current. The relative magnitude of anion conductance to coupled current follows the sequence: EAAT4 ~ EAAT5 > EAAT1 > EAAT3 > EAAT2. In each transporter, the anion conductance displays a chaotropic selectivity sequence (e.g., SCN^- is approximately 70-fold more permeant than Cl^- is in EAAT1). The magnitude of the anion conductance through a particular EAAT is also substrate specific; for example, D-aspartate activates a relatively larger anion conductance through EAAT1 than does L-glutamate, resulting in different net current reversal potentials for these substrates. In the absence of glutamate, a small tonic anion leak is present that seems to require Na^+. The glutamate-dependent anion conductance is also affected by the identity of the alkali cation cotransported with glutamate; Li^+ can substitute for Na^+ in some isoforms to support glutamate transport but is much less efficacious at activating the anion conductance.

Importantly, in at least some of the EAAT isoforms, replacement of Cl^- with impermeant anions like gluconate does not affect glutamate flux. Cl^- flux thus is thermodynamically uncoupled from glutamate flux, and may involve a channel-like mechanism instead. Indeed, glutamate-dependent anion current fluctuations have been observed in whole-cell and excised patch recordings that are consistent with a stochastically gated channel mechanism related to but distinct from the alternating-access glutamate transport mechanism. Comparative analysis of the glutamate concentration dependence of both transport and chloride channel gating suggests that each subunit in the trimer harbors both a chloride channel and a glutamate transporter; furthermore, each subunit functions independent of the other two subunits. Activation of the anion current in response to rapid glutamate application reveals similar but somewhat slower dynamics than does the stoichiometrically coupled current (**Figure 2**). The structural determinants and kinetics of the anion channel and the transporter may indeed be intimately related despite the important differences between channel and carrier gating. Kinetic models may be able to reconcile these differences by representing a subset of the Markov states in the transport cycle as open-channel states. Covalent modification of several mutants with engineered cysteine residues (e.g., V417C; EAAC1 numbering) eliminates transport but does not prevent activation of the anion current by extracellular glutamate, suggesting that some or all of the open-channel states may be in the T_o or T_{occ} forms.

Synaptic Physiology

As previously discussed, the coupled fluxes of glutamate, Na^+, H^+, and K^+ would theoretically allow maintenance of ambient extracellular glutamate levels in the low nanomolar concentration range. However, estimates of actual glutamate levels in the brain tend to be significantly higher, and in some areas are above the threshold required to generate tonic activity of *N*-methyl-D-aspartate (NMDA) and/or metabotropic glutamate receptors. The tonic extracellular level reflects a nonequilibrium state determined by the rates of uptake and release. The latter includes both synaptic and nonsynaptic components. Factors influencing the extracellular glutamate concentration include the density of EAATs and the frequency of synaptic activity. Nonsynaptic routes of release, particularly from astrocytes, also contribute to the extracellular glutamate level. The nonsynaptic mechanisms are generally less well characterized, but appear to include swelling-activated anion channels, the system xc- glutamate–cystine exchanger, and possibly other routes.

In addition to playing a central role in determining ambient glutamate levels, EAAT activity can also influence the temporal concentration profile of synaptically

released glutamate. Following vesicle fusion, the glutamate concentration in the synapse briefly reaches millimolar values. The dynamics of the concentration decay are strongly influenced by the geometry of the synapse and surrounding barriers. Diffusion appears to account for the major component of clearance within the synapse, with a time constant of a few milliseconds. However, EAAT activity can also influence the kinetics of the postsynaptic response in a manner dependent not only on the synaptic and extrasynaptic geometry, but also on receptor affinity, receptor distance from release sites, and release frequency. Additionally, multivesicular release and/or heterosynaptic glutamate spillover between release sites can result in slower transmitter clearance and a larger relative contribution of EAATs to transmitter decay kinetics. There is a great deal of heterogeneity in the effect of transporter blockade on synaptic transmission at different synapses. In general, the influence of uptake on synaptic responses becomes more pronounced as postsynaptic receptor affinity for glutamate increases, receptor kinetics slow, and proximity to the release site decreases. Oftentimes there is a spectrum of effects on different components of the synaptic response according to these criteria. Thus, the response kinetics of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors within the synapse tend to be less influenced by glutamate transport than do the responses of extrasynaptic NMDA or metabotropic glutamate receptors. Synapse organization in the CNS exhibits a rich diversity, and the contributions of EAATs to synaptic signaling are correspondingly varied.

See also: Glial Glutamate Transporters; Glial Glutamate and GABA Metabolism; Glutamate; Glutamate Receptor Organization: Ultrastructural Insights; Ionic Channels in Glia.

Further Reading

Boudker O, Ryan RM, Yernool D, et al. (2007) Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter. *Nature* 445: 387–393.

Brasno G and Otis TS (2001) Neuronal glutamate transporters control activation of postsynaptic metabotropic glutamate receptors and influence cerebellar long-term depression. *Neuron* 31: 607–616.

Christie JM and Jahr CE (2006) Multivesicular release at Schaffer collateral–CA1 hippocampal synapses. *Journal of Neuroscience* 26: 210–216.

Danbolt N (2002) Glutamate uptake. *Progress in Neurobiology* 65: 1–105.

Grewer C and Rauen T (2005) Electrogenic glutamate transporters in the CNS: Molecular mechanism, pre-steady-state kinetics, and their impact on synaptic signaling. *Journal of Membrane Biology* 203: 1–20.

Huang YH and Bergles DE (2004) Glutamate transporters bring competition to the synapse. *Current Opinion in Neurobiology* 14: 346–352.

Zerangue N and Kavanaugh MP (1996) Flux coupling in a neuronal glutamate transporter. *Nature* 383: 634–637.

Relevant Websites

http://www.rcsb.org – Protein Data Bank (Research Collaboratory for Structural Bioinformatics).

http://sourceforge.net – Sourceforge.net. Python-enhanced molecular graphics tool.

Glial Glycogen Metabolism

I Allaman, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Glial Glycogen Metabolism: Introduction

The brain has a very high-energy metabolic rate and uses glucose as its main energy substrate. In most tissues glucose can be stored as glycogen, a highly branched polysaccharide. As a general and intuitive principle, glycogen is synthesized (glycogenesis) when glucose is in excess and is degraded (glycogenolysis) to meet increased energy needs. An interesting feature of glycogen is that it can be rapidly metabolized in a process that does not require ATP. Moreover, and in contrast to fatty acids, glycogen metabolization can yield ATP under anaerobic conditions. Thus, glycogen represents an advantageous means to store glucose. In the body, the major sites of glycogen deposition are the liver and the skeletal muscle. Glycogen reserves are also present in the brain but in a much lower concentration than in peripheral tissues. Indeed, normal glycogen concentrations in mouse liver, muscle, and brain are roughly in the proportion of 100:10:1. Despite its low level in the central nervous system (CNS), glycogen is the largest energy reserve of the brain. The precise functions of brain glycogen are still not fully characterized. Nevertheless, it has been shown that glycogen reserves are consumed during failure of energy supply and that its metabolism is modulated by neuronal activity in physiological conditions. The almost exclusive localization of glycogen in a particular type of glial cell, the astrocyte, in adult mammalian brain is further evidence for a significant metabolic interaction between neurons and glia.

Glycogen Metabolism

Most of our knowledge about the general pathways of glycogen metabolism originates from studies conducted in peripheral tissues, but these pathways are common to all tissues. Uridine diphosphate (UDP)-glucose is the source of all glucosyl residues added to glycogen (Figure 1). UDP-glucose is formed from the conversion of glucose-6-phosphate to glucose-1-phosphate, by phosphoglucomutase, which is then combined to UTP through the action of UDP-glucose pyrophosphorylase. This process is energy consuming, since regeneration of UTP from UDP requires formation of a high-energy phosphate bond. The first step in the biogenesis of glycogen is the autocatalytic attachment of glucosyl units to a single tyrosine residue of the enzyme glycogenin. Elongation of this 'primed' glycogenin is performed by glycogen synthase, which catalyzes the formation of an α-1,4-glucosidic linkage using UDP-glucose as a donor. When the elongating glycogen chain consists of at least 11 residues, a branching enzyme transfers the terminal oligoglycan (around six glucose units) onto a neighboring chain by an α-1,6-glucosidic bond. Thus, glycogen synthase elongates the glycogen chain, whereas the branching enzyme produces new branches, creating a mature glycogen molecule presenting so-called 'bush-like' structure. Branching is important because it increases the solubility of glycogen. Glycogen particles (β-particles) have a spherical shape, about 10–30 nm in diameter, and can contain up to 60 000 glucose units. In the liver, 20–40 β-particles are associated into larger complexes known as α-rosettes.

Glycogenolysis is a process that does not require energy; is not a reverse reaction of glycogen synthesis, but implies the concerted action of glycogen phosphorylase and the bifunctional debranching enzyme. In the presence of inorganic phosphate, glycogen phosphorylase cleaves the terminal glucose residue of an external chain and releases it as glucose-1-phosphate. Glycogen phosphorylase proceeds until the external chains have been shortened to four glucose units before a branch point. At this point, the transferase activity of debranching enzyme removes the outer three glucosyl units and attaches them through an α-1,4-glucosidic bond to the nonreducing end of the molecule, which then becomes available for action by glycogen phosphorylase. The single remaining α-1,6-linked glucose unit is then removed and liberated as free glucose through the amylo-1,6 glucosidase activity of debranching enzyme. Finally, glucose-1-phosphate is converted by phosphoglucomutase to glucose-6-phosphate, which is (with glucose) the final product of glycogen degradation. Glucose-6-phosphate can then reenter the glycolysis pathway or the pentose phosphate pathway, or can be released as glucose in the presence of glucose-6-phosphatase activity (Figure 1).

Regulation of Glycogen Metabolism

Glycogen is a highly dynamic pool of glucose, the metabolism of which is under complex regulation involving various allosteric factors as well as covalent modifications. In a general way, activation of glycogen synthesis results in the inhibition of the degradative pathway, and vice versa. Nevertheless, there is ample documentation demonstrating that

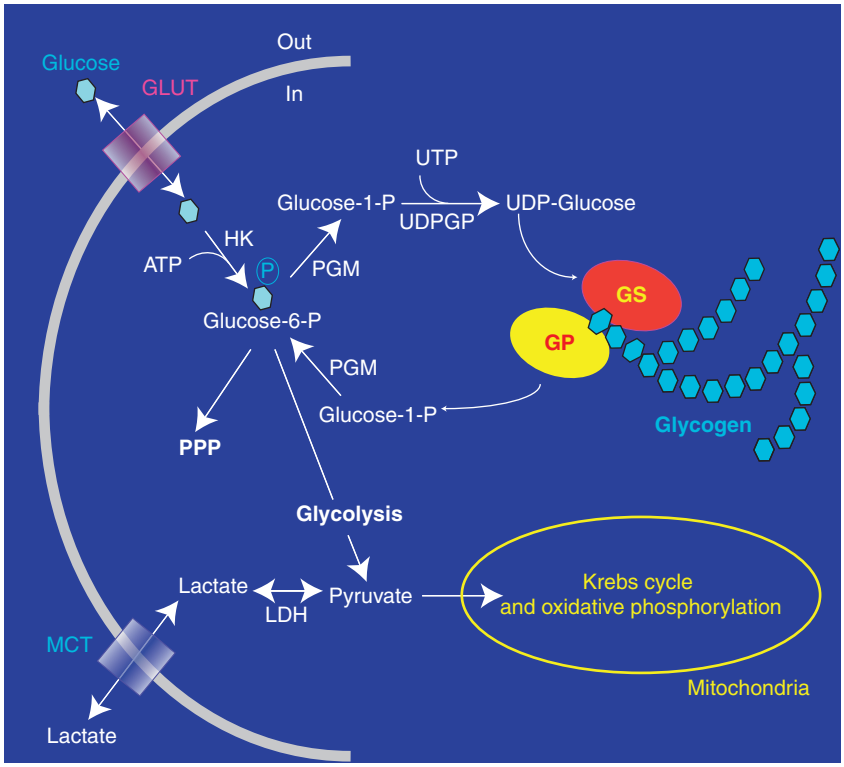

Figure 1 Schematic representation of glucose metabolism. Glucose enters cells through glucose transporters (GLUT) and is phosphorylated by hexokinase (HK) to produce glucose-6-phosphate (glucose-6-P). Glucose-6-P can be processed into different metabolic pathways: it can be metabolized through the pentose phosphate pathway (PPP) or through glycolysis, giving rise to pyruvate production. Pyruvate can enter mitochondria, where it is metabolized through the Krebs cycle and oxidative phosphorylation. Alternatively, pyruvate can be reduced to lactate by lactate dehydrogenase (LDH); lactate is released into the extracellular space through monocarboxylate transporters (MCT). Glucose-6-P can also be used to store glucosyl units as glycogen (see text for details). As a first step, glucose-6-P is converted to uridine diphosphate (UDP)-glucose through the action of phosphoglucomutase (PGM) and UDP glucose pyrophosphorylase (UDPGP). UDP-glucose is then incorporated into glycogen by glycogen synthase (GS). To meet energy needs, glycogen can be broken down by glycogen phosphorylase (GP) to produce glucose-1-P, which is converted back to glucose-6-P through the action of PGM.

simultaneous synthesis and degradation can take place at the tissue level. Such regulation is believed to involve distinct glycogen pools in individual or different cells, synthesis of one pool being in concurrence with degradation in another. It is generally accepted that glycogen synthase and phosphorylase represent two central elements of this regulation. Glycogen synthase is tightly controlled by the reversible phosphorylation of multiple serine residues. Generally, phosphorylation is associated with an inactivation of the enzyme (conversion of the active a-form to the b-form). *In vitro* studies have shown that glycogen synthase can be phosphorylated by various protein kinases, including protein kinase A (PKA), protein kinase C (PKC), phosphorylase kinase, casein kinase 1 (CK1), casein kinase 2 (CK2), calcium and calmodulin-dependent protein kinase II (CaMKII), glycogen synthase kinase-3 (GSK-3), and adenosine monophosphate (AMP)-activated protein kinase (AMPK). Moreover, glycogen synthase b is strongly activated by glucose-6-phosphate (high cellular energy status). Like glycogen synthase, glycogen phosphorylase is regulated by both ligand binding and phosphorylation. Glycogen phosphorylase is converted from the inactive b-form to the active a-form through phosphorylation of a single serine residue by phosphorylase kinase. Activation of phosphorylase kinase is mediated by increased cellular Ca^{2+} levels or phosphorylation by PKA. Glycogen phosphorylase is also allosterically activated by AMP (low cellular energy status) and inhibited by glucose-6-phosphate and ATP (high cellular energy status), which enables this enzyme to sense and respond to an intracellular need for energy. As described, kinase-catalyzed phosphorylation of both enzymes leads to glycogenolysis (by activation of glycogen phosphorylase and inactivation of glycogen synthase), whereas dephosphorylation favors glycogenesis.

Protein phosphatase 1 (PP1), upon its association with regulatory glycogen targeting subunits, appears to be responsible for dephosphorylation of glycogen synthase, glycogen phosphorylase, and phosphorylase kinase. Thus, both glycogen synthase and glycogen phosphorylase are under the control of extracellular

signaling (which controls phosphorylation state) as well as intracellular (allosteric) regulation (which drives synthesis and degradation of glycogen, depending on energy requirements). It is of importance to note that glycogenolysis is a very fast process (activated on a timescale of milliseconds) that can be used to match moment-to-moment cellular energy needs without requiring energy, in contrast to stimulation of glycolysis from glucose, which requires the investment of two molecules of ATP. Different isoforms of glycogen-related enzymes with different regulation properties are expressed in a tissue-specific manner to meet particular requirements. For instance, in liver, glycogenolysis leads to formation of glucose (due to the presence of glucose-6-phosphatase), which is released into the bloodstream when blood glucose levels fall; in skeletal muscle, glycogen is used as a local energy store during muscle contraction. Even though characterized in less detail than their peripheral counterparts, major components of glycogen metabolism have been described in mammalian brain; examples include glycogen phosphorylase and synthase, phosphorylase kinase, glycogenin, branching enzymes, debranching enzymes, and regulatory glycogen-targeting subunits of PP1. Moreover, our knowledge of isoform expression patterns and biochemical properties of enzymes involved in glycogen metabolism (in particular, glycogen synthase and phosphorylase) in the brain, as well as data on regulation of cerebral glycogen (see later) and the virtual absence of glucose-6-phosphatase activity, indicate that brain glycogen regulation closely resembles that in muscle cells.

Localization of Glycogen in the CNS

Glycogen molecules are electron-dense particles that are easily recognizable by electron microscopy. With this technique it has been possible to demonstrate that in the adult brain glycogen is almost exclusively localized in astrocytes, although its presence has also occasionally been shown in certain large neurons in the brain stem, as well as in choroid plexus and ependymal epithelia. Glycogen particles can be observed diffusely scattered through the cytoplasm of astrocytes and their processes (Figure 2). Developmental studies have indicated that in the immature CNS, glycogen particles are also present in relatively large amounts in neurons, but neuronal glycogen decreases with age and is practically absent in mature CNS neurons. The cellular expression patterns of glycogen phosphorylase and phosphorylase kinase show a very strong correlation with glycogen deposits. In contrast, the cellular distribution of glycogen synthase appears less selective, since, while present in astrocytes, it is also expressed in neurons.

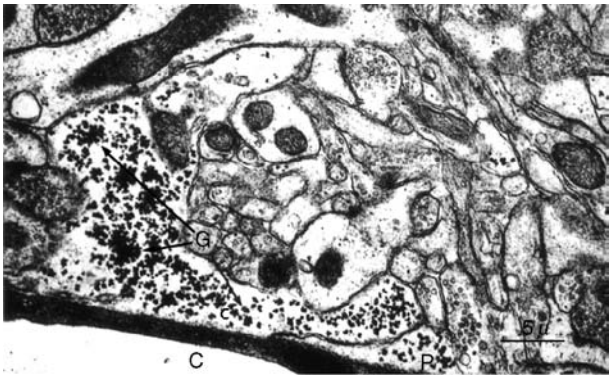

Figure 2 Visualization of brain glycogen by electron microscopy. Pericapillary astrocytic processes (P) laden with β-glycogen particles (G) bordering a capillary (C) in the hippocampus from an anesthetized mouse. Reproduced from Phelps CH (1972) Barbiturate-induced glycogen accumulation in brain: An electron microscopic study. *Brain Research* 39: 225–234, with permission from Elsevier.

Because ischemia is a strong activator of glycogenolysis and low levels of glycogen in the brain, precise postmortem measurements of glycogen levels are problematic. The lability of glycogen has been seen as a limitation in the determination of the regional localization and the role of this energy store. Technological breakthroughs, optimization of extraction procedures, and use of a high-energy focused microwave irradiation method to sacrifice animals (fast inactivation of glycogen metabolism) have partially overcome this problem. More recently, nuclear magnetic resonance (NMR) spectroscopy has emerged as a new powerful and noninvasive technique to follow glycogen content *in situ*. Using these new approaches, it has been shown that basal concentration of glycogen in rodent brain is in the range of 2–5 μmol g^{-1} of wet tissue, although higher values of up to 8–12 μmol g^{-1} of wet tissue have been reported. In rodent brain, glycogen presents regional variability, with higher levels in pons–medulla, cerebellum, and hippocampus, followed by hypothalamus, thalamus–midbrain, and cerebral cortex, which display similar concentrations, the lower concentration being found in striatum. Several studies tend to show that gray-matter structures possess higher levels of glycogen than does white matter, although a recent study indicates opposite results. In addition, variations of glycogen levels in particular structures have also been observed. For example, superficial layers of the mouse cerebral cortex possess higher glycogen content compared to the deeper layers.

Functions of Glycogen in the CNS

What are the functions of glycogen in the CNS? Is it used only as a local emergency reserve when there is a

decrease in brain energy supply, or is it mobilized to sustain normal brain functions, and if so, for what purposes? Despite numerous studies, these intriguing questions are still not fully answered. Concerning its role as an energy reserve, there is now no doubt that glycogen can be used to preserve cell integrity in pathological conditions such as hypoglycemia or ischemia. Nevertheless, several lines of evidence, obtained both *in vitro* and *in vivo*, also point to a physiological role of brain glycogen. First, glycogen content is under dynamic control by neurotransmitters, neurohumoral agents, and local energy status. Second, glial glycogen metabolism appears to be tightly coupled to neuronal activity.

Glial Glycogen Metabolism Regulation *In Vitro*

Numerous *in vitro* studies, conducted either in astrocyte cultures or in brain slice preparations, have shown that glycogen is under the dynamic control of neuroactive substances. Neurotransmitters such as noradrenaline (NA), serotonin, histamine, vasoactive intestinal peptide (VIP), and secretin have been shown to promote glycogenolysis through activation of the cyclic AMP (cAMP) signaling pathway (with PKA) or through elevation of intracellular Ca^{2+} levels. In addition to neurotransmitters, adenosine, ATP, and K^{+} also induce glycogenolysis. NA, VIP, and adenosine are of particular interest since they produce a biphasic effect in astrocyte cultures. Indeed, the rapid glycogenolysis induced by these neurotransmitters is transient and is followed by a delayed massive glycogen resynthesis (taking place several hours after stimulation) (Figure 3). Accordingly, this long-term effect is accompanied by dephosphorylation and activation of glycogen synthase. This effect also depends upon the activation of the cAMP signaling cascade and requires synthesis of new proteins. Glycogen resynthesis involves stimulation of expression of a regulatory glycogen-targeting subunit of PP1, named PTG (for protein targeting to glycogen). Thus, NA, VIP, and adenosine can cause glycogen mobilization and delayed replenishment of this energy store. Other compounds that favor the synthetic pathway include insulin, insulin-like growth factor 1 (IGF-1), and glutamate. The stress hormone glucocorticoids have been shown to exert long-term control on glycogen metabolism in astrocytes cultures by decreasing glycogen levels and by partial inhibition of glycogen synthesis induced by NA. Moreover, glycogen levels are very sensitive to energy supply or ATP production. Indeed, hypoglycemia, ischemia, and hypoxia are potent inducers of glycogen breakdown, whereas glycogen synthesis is promoted with increasing concentration of glucose. Interestingly, glucose-deprived astrocytes rapidly deplete their glycogen content, with the concomitant release of lactate but not glucose (Figure 1). This result not only argues for a glial glycogen metabolism comparable to its muscle counterpart as a local energy

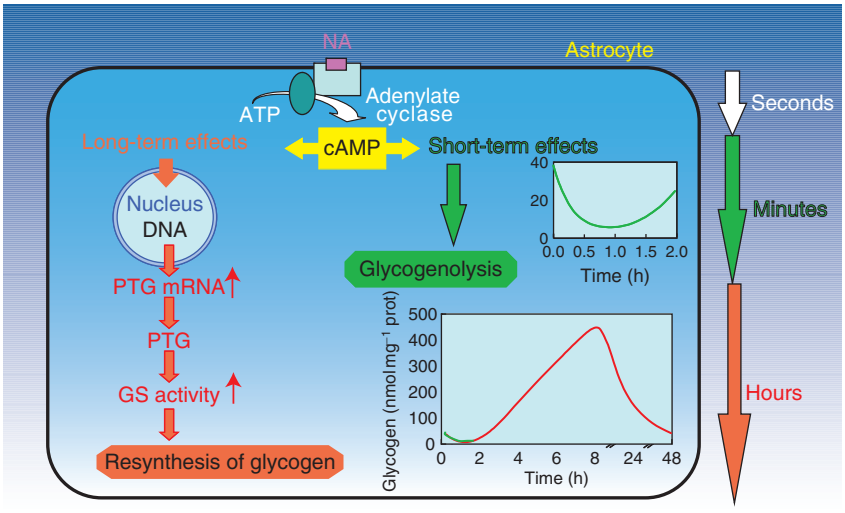

Figure 3 Differential modulation of glycogen metabolism in cultured astrocytes. Activation of cyclic adenosine monophosphate (cAMP)-dependent intracellular signaling by noradrenaline (NA), as well as by vasoactive intestinal peptide and adenosine, results in a short-term (seconds to minutes, in green) glycogenolysis and in a delayed (hours, in red) glycogen resynthesis. The long-term response requires induction of gene expression and is accompanied by stimulation of mRNA expression of protein targeting to glycogen (PTG), a member of the glycogen-targeting subunits of protein phosphatase 1, and by the activation of glycogen synthase (GS). Adapted from Magistretti PJ (2006) Neuron–glia metabolic coupling and plasticity. *Journal of Experimental Biology* 209: 2304–2311, with permission. Copyright (2006) by The Company of Biologists Ltd.

store, but also shows that glycogen mobilization can give rise to lactate release, which could be used by neighboring cells as an energy substrate.

A first line of evidence demonstrating the importance of glial glycogen to preserve neuronal integrity comes from studies on murine cortical cell cultures, containing both glia and neurons, in which depletion of astrocyte glycogen content reduces neuronal survival following glucose deprivation. This experiment also implies a metabolic coupling mechanism between both cell types. Interaction between neurons and astrocytes in terms of glycogen metabolism has also been demonstrated in honeybee retinal preparations, in which repetitive light stimulation induces glycogenolysis in Müller cells, retinal counterparts of astrocytes. Taken together, these observations are in agreement with the hypothesis that glycogen can play an active role *in vivo*, either during failure of energy substrate delivery or during physiological conditions. This conclusion is further supported by studies performed in rodent optic nerve preparations, representative of central white-matter tracts, showing that glial glycogen can sustain axonal function in either pathological or physiological conditions. Electrophysiological recordings demonstrate that functional activity of the optic nerve, as measured by the area under the supramaximal compound action potential (CAP) during aglycemic conditions, can be maintained for over 20 s, after which it rapidly fails, with a concomitant decrease in glycogen content. Manipulation of glycogen levels before induction of aglycemia demonstrates that there is a linear correlation between glycogen content prior to aglycemia and latency to functional CAP failure during aglycemia. Moreover, under normoglycemic conditions, glycogen content decreases during high-frequency axon discharge, while CAP function is fully maintained. During both aglycemia and high-frequency axon discharge, it has been shown that CAP failure is accelerated by pharmacological inhibition of glycogen phosphorylase. Moreover, in both of these paradigms, CAP failure is also accelerated when lactate transport is blocked, suggesting a net transfer of lactate as an energy substrate from astrocyte glycogen stores to the neuronal compartment. Overall, these studies demonstrate that astrocytic glycogen is not only available for use as an emergency reserve to preserve neuronal integrity and functions when energy supply is decreased, but that it can also be mobilized to support neuronal energy requirements during neuronal activity.

Pathological Conditions

In vivo, variation of glycogen levels has been studied in different pathological paradigms or pathologies. Even if these studies do not necessarily define a physiological role of glycogen, they have been important to highlight the fact that brain glycogen metabolism is sensitive to changes in the nature of the microenvironment in terms of energy supply, presence of neurohumoral agents, and in case of cell trauma or death. Thus, glycogen accumulation in astrocytes is observed in traumatic injuries at the periphery of tumors after X-ray irradiation, or in neurodegenerative disorders such as Creutzfeldt–Jakob disease and Alzheimer's disease. Moreover, pathologies such as glycogen storage diseases, with the exception of type 0, also lead to brain glycogen accumulation. *In vivo* experiments also support a role of glycogen to sustain cerebral integrity during periods of decreased energy availability.

Different studies demonstrate that insulin-induced systemic hypoglycemia reduces cerebral glycogen reserves. Using NMR spectroscopy, it has been possible to show in brain of anesthetized rats that depletion of glycogen induced by insulin occurs only when brain glucose levels decrease to very critical levels, approaching zero (with plasma glucose concentration below 2 mM). In this condition, a glycogen signal could still be observed even after 2 h of severe hypoglycemia, demonstrating that brain glycogen can survive extended periods of hypoglycemia. Thus, glycogen can serve as a significant energy reservoir during less severe hypoglycemia and can account for a partial deficit in glucose supply. Interestingly, following a single episode of hypoglycemia, hypoxia, or ischemia, brain glycogen concentration, measured several hours after the insult, exceeds preinsult levels, indicating a 'supercompensation' of glucose storage. The molecular mechanisms involved in this effect and its role are not well characterized. Nevertheless, this increase in glycogen deposition might correspond to a form of preconditioning. For instance, hypoxic preconditioning has been show to delay energy depletion from hypoxia–ischemia in the immature rat brain.

Physiological Conditions

Under physiological conditions, a close relationship between glycogen mobilization and neuronal activity has been demonstrated *in vivo*. Evidence linking brain glycogen metabolism and neuronal activity comes from the observation that glycogen levels increase under conditions of tonically decreased neuronal activity, such as anesthesia, hibernation, or sleep. It is well established that administration of general anesthetics (e.g., pentobarbital) to animals leads to great diminution of neuronal activity and increases glial glycogen levels. Interestingly, following pentobarbital anesthesia, changes in glycogen levels appear to be greatest in areas of high synaptic

density. Further evidence of a direct link between effect of anesthetics on neuronal activity and modulation of glycogen levels comes from the observation that anesthetics increase glycogen only when applied *in vivo*, and not in astrocyte cultures. Like anesthesia, hibernation is a state accompanied by deeply reduced neuronal activity and an increase in brain glycogen levels (e.g., in ground squirrels and hedgehogs). In addition, modulation of total brain glycogen levels has been described during the sleep–wake cycle, with maximum glycogen levels observed at the beginning of the sleep period. Since sleep is classically considered as an energy-saving state, it has been hypothesized that sleep may serve to restore brain glycogen levels that have decreased during wakefulness.

Direct evidence for an association between neuronal activation and glial glycogen metabolism has been obtained in awake rats following sensory stimulation or following generalized stimulation of sensory and motor pathways. Thus, tactile stimulation of the rat face and vibrissae accelerates local glycogen utilization in the somatosensory cortex as well as in the relevant thalamic nuclei (trigeminal, sensory, and motor nuclei) (**Figure 4**). These observations indicate that the physiological activation of a neuronal circuit results in the mobilization of glial glycogen stores. They also suggest that rapid increase in local energy demand during neuronal activation may be fueled by stimulation of glycogenolysis. A recent study has demonstrated that pharmacological inhibition of glycogenolysis in astrocytes interrupts memory consolidation in young chickens in a bead discrimination learning task. Whether similar phenomena occur in mammalian brain remains to be determined, but this presents an interesting challenge for the determination of brain glycogen functions. Such observations also lead to the question of why glycogen is mobilized in physiological conditions during neuronal activation. Indeed, in response to focal brain activation, local cerebral blood flow increases, adjusting glucose delivery to energy demands. This response is fast but not instantaneous. Thus, glycogen breakdown – which occurs in milliseconds and, unlike glucose utilization through glycolysis, does not require energy – may provide a readily available energy substrate to buffer any brief mismatch between metabolic demand and energy supply, suggesting a role of glycogen as a 'metabolic buffer.'

Another issue is the nature of the signal(s) triggering glycogen mobilization in relation to increased neuronal activity. *In vitro* studies (see earlier) demonstrate that glycogenolysis could be initiated by several neurotransmitters, such as VIP and NA. In view of the particular morphology of the neurons that contain VIP or NA, a function of these neuronal systems to locally regulate the availability of energy reserves has

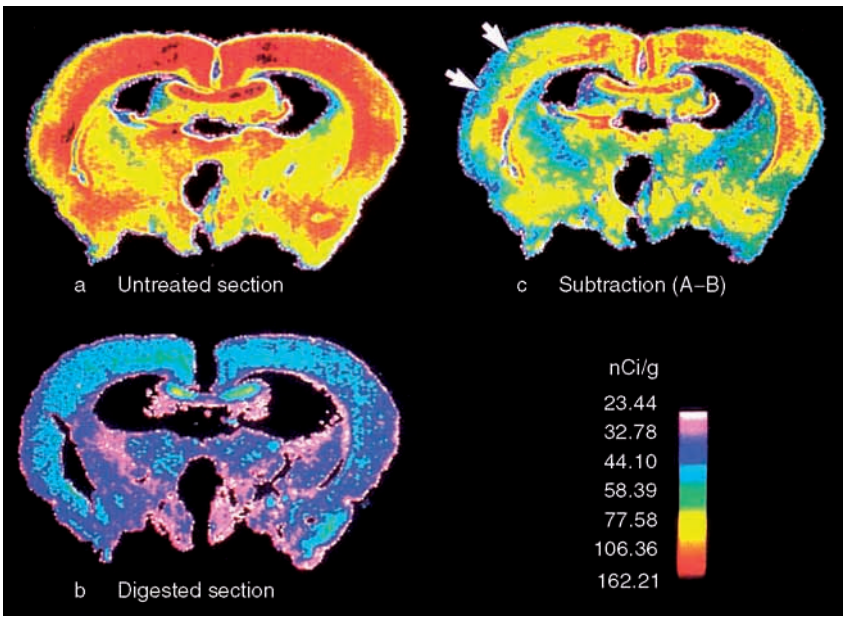

Figure 4 Glycogen utilization in the rat somatosensory cortex during tactile stimulation. [^{14}C]Glycogen labeling was obtained by intravenous injection of [^{14}C]glucose, and its utilization was determined by autoradiography. Autoradiographs were prepared before (a) and after (b) removal of glycogen from the section by enzymatic digestion. (c) Digital subtraction of autoradiograph (b) from autoradiograph (a) shows the distribution of [^{14}C]glycogen in the section. As shown in autoradiograph (c), unilateral tactile stimulation of face and vibrissae causes a loss of [^{14}C]glycogen in the somatosensory cortex (arrowheads) contralateral to stimulation. Reproduced from Swanson RA, Morton MM, Sagar SM, et al. (1992) Sensory stimulation induces local cerebral glycogenolysis: Demonstration by autoradiography. *Neuroscience* 51: 451–461, with permission from Elsevier.

been proposed. Indeed, VIP neurons are bipolar intracortical cells with an arborization pattern that defines cortical columns; these neurons receive direct thalamocortical and corticocortical inputs, implying that they are activated by modality-specific afferents to a given cortical area. The noradrenergic system originates in the locus coeruleus in the brain stem, the site from which fibers diffuse virtually to all areas of the brain. In the cortex, noradrenergic axons enter the frontal pole and diffusely project across a vast rostrocaudal expanse of the cortex, suggesting that a given set of cortical noradrenergic fibers can affect vast areas of the cortex simultaneously (Figure 5). Given these morphological characteristics, it has been proposed that VIP neurons, by promoting glycogenolysis, can control energy availability locally, within cortical columns, while noradrenergic fibers can metabolically 'prime' the neocortex.

Fate of Energy Stored in Glycogen

There is compelling evidence that astrocytes and neurons function as a cooperative metabolic unit. For example, astrocytes are responsible for the reuptake of neurotransmitters such as glutamate, as well as for the clearance of K^+ released in the extracellular space

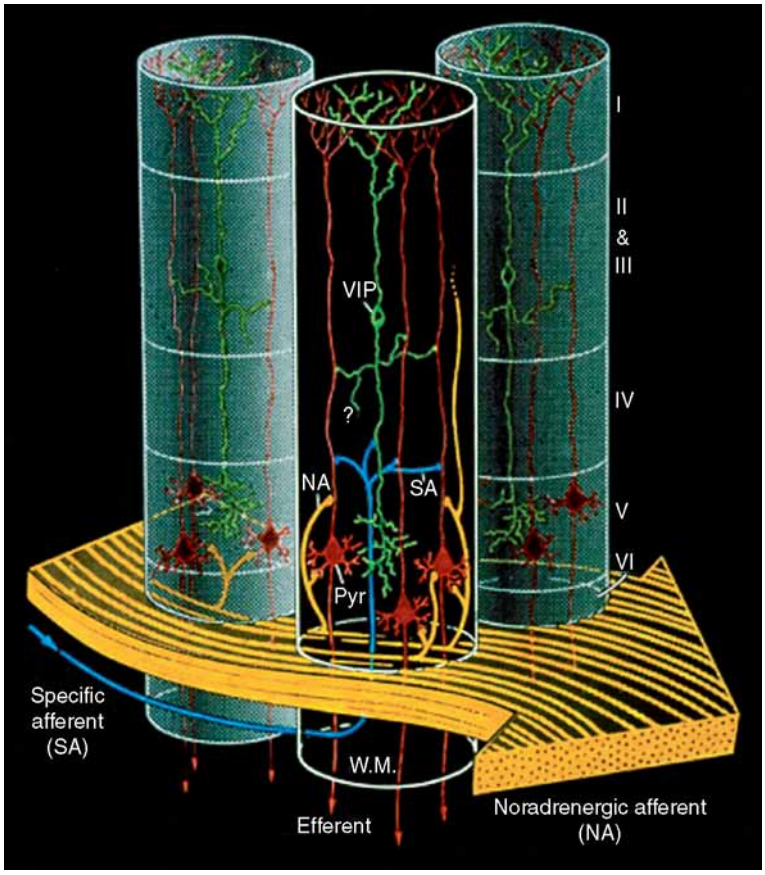

Figure 5 Diagrammatic representation of the anatomic substrate for the synergistic effects of noradrenaline (NA) and vasoactive intestinal peptide (VIP) following the concomitant activation of noradrenergic cortical afferents and a group of VIP-containing neurons. Such arrangement would allow for the establishment of temporary columnar 'hot spots.' Cortical layers are denoted by roman numerals. Note tangential orientation of NA fibers, and radially restricted domain of the VIP-containing neuron. The concomitant activation of noradrenergic fibers (large, horizontal yellow arrow at the base of the cylinders) by unexpected sensory stimuli, and of a group of VIP-containing intracortical neurons by specific thalamic inputs (small blue arrow, bottom left) would lead to a drastic increase of cyclic adenosine monophosphate levels within a discrete volume of somatosensory cortex, delineated here by the cylinder with the dark background. Adjoining gray-background cylinders represent nonactivated cortical volumes. For graphic clarity, only the VIP-containing (green) cells and the pyramidal cells (red) have been represented here. In particular, the principal target cells of the thalamocortical afferents (i.e., the small stellate cells in layer IV) have been omitted. However, any cell with the capacity for dendritic reception in layer IV may receive thalamic inputs. Furthermore, also for graphic clarity, only one VIP neuron per cylinder has been drawn, in representation of a group of VIP neurons. VIP, VIP-containing bipolar cell; NA, noradrenergic afferent; PYR, pyramidal cell furnishing major efferent projections; SA, specific afferent (i.e., from thalamus or from other cortical areas); W.M., subcortical white matter. Reproduced from Magistretti PJ and Morrison JH (1988) Noradrenaline- and vasoactive intestinal peptide-containing neuronal systems in neocortex: Functional convergence with contrasting morphology. *Neuroscience* 24: 367–378, with permission from Elsevier.

during neuronal activation. In parallel, astrocytes also release into the extracellular space energy substrates, such as lactate, which can be used by neurons to meet their own energy requirements. Considering that neurons are the prevalent site of energy consumption in the mature brain, this raises the question of the fate of energy stored in glial glycogen. Is it a 'neuronal' energy reserve or is it used by astrocytes for their own purposes? *In vitro* studies clearly show that glycogenolysis occurs in astrocytes, during glucose deprivation or following high neuronal stimulation, sustaining neuronal integrity and functions. Glycogen breakdown leads to the release of energy substrates (e.g., lactate) in the extracellular space, and these substrates are taken up and used by neurons to meet their energy needs.

Concluding Remarks

Glycogen represents the major energy store of the brain. During decades of research, glycogen has been perceived as an emergency reserve used in case of energy supply deficiency, but more recent data, obtained with the development of new techniques, now points to a physiological role for glycogen in relation to neuronal activation. Nevertheless, some questions as to glycogen physiological processes still need further exploration. In this context, the advent of noninvasive NMR spectroscopic methods may allow the development of very useful experimental approaches to investigate glycogen function in humans. The quite exclusive localization of glycogen in glia, and its mobilization upon neuronal activation, underline the tight metabolic cooperation occurring between neurons and astrocytes.

See also: Glial Energy Metabolism: Overview; Glial Glutamate and GABA Metabolism; Glial Energy Metabolism: A NMR Spectroscopy Perspective; Glycogen Metabolism in CNS White Matter.

Further Reading

Brown AM (2004) Brain glycogen re-awakened. *Journal of Neurochemistry* 89: 537–552.

Brown AM, Sickmann HM, Fosgerau K, et al. (2005) Astrocyte glycogen metabolism is required for neuronal activity during aglycemia or intense stimulation in mouse white matter. *Journal of Neuroscience Research* 79: 74–80.

Gruetter R (2003) Glycogen: The forgotten cerebral energy store. *Journal of Neuroscience Research* 74: 179–183.

Magistretti PJ (2003) Brain energy metabolism. In: Squire LR, Bloom FE, McConnell SK, Roberts JL, Spitzer NC, and Zigmond MJ (eds.) *Fundamental Neuroscience,* 2nd edn., pp. 339–360. San Diego: Academic Press.

Magistretti PJ and Morrison JH (1988) Noradrenaline- and vasoactive intestinal peptide-containing neuronal systems in neocortex: Functional convergence with contrasting morphology. *Neuroscience* 24: 367–378.

Magistretti PJ, Sorg O, and Martin J-L (1993) Regulation of glycogen metabolism in astrocytes: Physiological, pharmacological, and pathological aspects. In: Murphy S (ed.) *Astrocytes: Pharmacology and Function*, pp. 243–265. San Diego: Academic Press.

Magistretti PJ (2006) Neuron–glia metabolic coupling and plasticity. *Journal of Experimental Biology* 209: 2304–2311.

Phelps CH (1972) Barbiturate-induced glycogen accumulation in brain: An electron microscopic study. *Brain Research* 39: 225–234.

Raichle ME and Mintun MA (2006) Brain work and brain imaging. *Annual Review of Neuroscience* 29: 449–476.

Roach PJ (2002) Glycogen and its metabolism. *Current Molecular Medicine* 2: 101–120.

Sagar SM, Sharp FR, and Swanson RA (1987) The regional distribution of glycogen in rat brain fixed by microwave irradiation. *Brain Research* 417: 172–174.

Swanson RA (1992) Physiologic coupling of glial glycogen metabolism to neuronal activity in brain. *Canadian Journal of Physiology and Pharmacology* 70: S138–S144.

Swanson RA, Morton MM, Sagar SM, et al. (1992) Sensory stimulation induces local cerebral glycogenolysis: Demonstration by autoradiography. *Neuroscience* 51: 451–461.

Tekkök SB, Brown AM, Westenbroek R, et al. (2005) Transfer of glycogen-derived lactate from astrocytes to axons via specific monocarboxylate transporters supports mouse optic nerve activity. *Journal of Neuroscience Research* 81: 644–652.

Glial Growth Factors

B M Reuss, University of Göttingen, Göttingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Glial cells are a major source of growth factors in the healthy and injured central and peripheral nervous systems. Glia-derived growth factors regulate differentiation and function of glial cells in an autocrine fashion, and they modulate neuronal process formation, migration, and survival during development. An important function of glial growth factors is the regulation of cellular repair processes after injury, degeneration, and demyelination in the peripheral and central nervous systems. Changes in growth factor synthesis, release, and action are also associated with the formation of glia-derived brain tumors.

Expression of Growth Factors by Different Glial Subpopulations

Satellite Cells

Satellite cells are specialized glial cells in the sensory ganglia ensheathing cell bodies of pseudounipolar sensory ganglion cells. During development, satellite cell-derived nerve growth factor (NGF) is an important determinant for neurite outgrowth from sensory neurons since, in the rat, neurite outgrowth from neonatal sensory neurons after removal of satellite cells is induced by NGF.

After peripheral nerve injury in the rat, NGF and neurotrophin NT-3 induce sprouting of noradrenergic nerve terminals in the dorsal root ganglia. Satellite cell-derived NGF also affects expression of the nicotinic acetylcholine receptor in rat sensory neurons. NGF also seems to affect satellite cells since NGF depletion reduces reactivity of rat trigeminal satellite cells after inferior alveolar nerve injury.

Other growth factors released from satellite cells are vascular endothelial growth factor (VEGF), the transforming growth factor (TGF)-βs, and TGF-α. Thus, VEGF plays an important role in the establishment and maintenance of blood vessels in the dorsal root ganglia during development, whereas TGF-β signaling is altered after peripheral nerve transection as indicated by changes in the distribution of TGF-β and the TGF-β type I and type II receptors in peripheral nerves and mechanoreceptors after traumatic injury. TGF-α and its receptor reveal strong upregulation in satellite cells in response to peripheral nerve lesioning.

Schwann Cells

Schwann cells are the myelinating cells of the peripheral nervous system and thus ensheathe with their processes axons in peripheral nerves. Perturbed Schwann cell functioning is a key feature of several demyelinating disorders, the most prominent of which is multiple sclerosis (MS). Schwann cells synthesize and release a wide variety of growth factors, such as neurotrophins, neuregulatory cytokines, TGF-βs, glial cell line-derived neurotrophic factor (GDNF), epidermal growth factors (EGFs), and platelet-derived growth factor (PDGF).

High levels of NGF are found in Schwann cells *in vivo* and in primary cell cultures *in vitro*. NT-3 mRNA is expressed at low concentrations in cultured Schwann cells but is upregulated upon immortalization. In response to nerve injury, neurotrophins are differentially up- or downregulated in Schwann cells. Thus, low levels of NGF in normal Schwann cells are greatly increased after injury. Also, transection of the sciatic nerve leads to an increase in the mRNAs for brain-derived neurotrophic factor (BDNF) and NT-4 but to a decrease of that for NT-3.

Schwann cells of the sciatic nerve also express ciliary neurotrophic factor (CNTF) beginning on postnatal day 8 and reaching adult levels on day 21. CNTF expression in Schwann cells is regulated by axonal factors since after axotomy, CNTF expression recovers only after regeneration of the axonal processes. Whereas in cultured Schwann cells mRNAs for CNTF and leukemia inhibiting factor (LIF), another neuregulatory cytokine, are high, myelinating Schwann cells *in vivo* show only CNTF expression and levels of LIF mRNA are low. However, retrograde transport of LIF and CNTF is increased after nerve lesioning.

In Schwann cells and their precursors *in vivo*, both TGF-β2 and -β3 are expressed, whereas TGF-β1 is upregulated only after injury. Thus, in the transected sciatic nerve, TGF-β1 mRNA is induced in the distal nerve stump, whereas that for TGF-β3 is repressed. Also during axon regeneration, TGF-β1 mRNA is transiently increased. In contrast to the *in vivo* situation, in cultured Schwann cells expression of TGF-β1 mRNA is high but decreases when axonal contact is mimicked by the application of forskolin. A striking feature of the effects of TGF-β on Schwann cells is that application of TGF-β together with tumor necrosis factor-α (TNF-α) can induce Schwann cell death, whereas either cytokine alone is ineffective.

Whereas in normal circumstances GDNF expression in Schwann cells and satellite cells is low, it shows transient upregulation after injury. Thus, GDNF is

upregulated in the distal nerve stump of the transected sciatic nerve, as well as in the satellite cells of the trigeminal ganglion.

Schwann cells are also known to express TGF-α and other neuregulins, as well as the neuregulin receptor ErbB1, on their surface. In accordance with this, neuregulins act as autocrine stimulators of Schwann cell proliferation during development. Similar observations have been made for PDGF, the expression of which in rat Schwann cells is high at birth, followed by a continuous postnatal decrease in nonmyelinating but persisting high levels in myelinating Schwann cells.

Oligodendrocytes

Oligodendrocytes are the myelinating cells of the central nervous system (CNS) and thus, like Schwann cells in the peripheral nervous system (PNS), ensheathe the axonal processes by their processes. Oligodendrocytes and their precursor cells also provide an important source of growth factors such as the neurotrophins and several TGF-β superfamily members. Thus, they express NGF, NT-3, and BDNF, probably mediating the autocrine control of proliferation and survival of immature oligodendrocytes and of O-2A precursor cells. In O-2A precursor cells, different TGF-β isoforms, such as TGF-β1 and -β3, are expressed, whereas mature oligodendrocytes express TGF-β2 and-β3. In cultured oligodendrocytes, all three TGF-β isoforms are present.

Astrocytes

Astrocytes in the CNS mediate the transport of nutrients and waste products between blood vessels and neurons, and they also modulate neuronal activity and synaptic transmission. Due to their important role during gliosis and neuronal repair, growth factor synthesis by astrocytes has been intensely studied. Thus, astrocytes in the healthy CNS release NGF, NT-3, and NT-4, whereas reactive astrocytes after gliosis start to express BDNF together with NGF, a process which at least *in vitro* seems to depend on interleukin (IL)-1β and interferon-γ (IFN-γ).

Astrocytes also release high amounts of fibroblast growth factor-2 (FGF-2), mRNA of which can be found in cortical, hippocampal, and spinal cord astrocytes with a strong postnatal increase. FGF-2 seems to act in an autocrine manner on astrocyte differentiation since in FGF-2$^{-/-}$ mice astroglial gap junction coupling and neurotransmitter sensitivity are changed. Astroglial FGF-2 is also responsible for the induction of endothelial tight junctions and thus for the formation of the blood–brain barrier in the CNS.

Astrocytes express neuregulatory cytokines such as CNTF, which can be found in type 1 astrocytes of the optic nerve and the olfactory bulb. Also, primary cultures of astrocytes contain CNTF mRNA and protein. Astroglial CNTF expression is upregulated during brain lesions such as entorhinal cortex lesions, hippocampal deafferentiation, optic nerve transection, and ischemia. Also, LIF is expressed in cultured astrocytes of variable origins and is upregulated after trauma and during inflammation.

However, astrocytes do not express EGF itself but synthesize other members of this growth factor family, such as TGF-α and the neuregulins. TGF-α has been found in astrocytes of the corpus callosum, striatum, and globus pallidum. Astrocytes also express ErbB1, the receptor for TGF-α, and therefore it is thought to act in an autocrine manner. In reactive astrocytes, TGF-α and its receptor are transiently reduced. Other neuregulins which can be found in cultured astrocytes are highly upregulated during reactive gliosis. Thus, GGF-2 and NDF have been detected in human white matter astrocytes of the spinal cord and in the cortex.

Of the insulin-like growth factors (IGFs), only IGF-1 is expressed in astrocytes, which *in vitro* has been shown to promote astroglial proliferation in an autocrine manner. During remyelination, IGF-1 is upregulated in astrocytes, together with its receptor, and is also increased after ischemia, probably acting in a protective manner.

With regard to the TGF-β superfamily members, in astrocytes a highly heterogeneous expression pattern can be found. Thus, astrocytes in the uninjured brain express only TGF-β2 and -β3, whereas after lesioning TGF-β1 is also upregulated. In contrast, GDNF is not expressed in astrocytes of the adult brain but can be found during early postnatal development. However, GDNF is released from cultured fetal human astrocytes and in astrocytes *in vivo* after brain damage. GDNF is also highly upregulated in gliomas *in situ* and in glioma cell lines.

Ependymoglia

Ependymoglia line the ventricular wall and the choroid plexus and, thus, are important for substance transport and for establishing the blood–liquor barrier. FGF-2 is expressed in the choroid plexus epithelia of both lateral and third ventricles, in ependymal cells of the third ventricle, and along the lateral sides of the lateral ventricles. All positive cells reveal a similar distribution with apical labeling, as well as some cytoplasmic staining. This suggests a transepithelial transport of growth factors such as FGF-2 via ependymal cells into the underlying brain parenchyma. However, FGF-2 also seems to play a role in the autocrine regulation of ependymal cells since both EGF and FGF-2 have been shown to cause proliferation of

ependymal precursor cells in the adult rat spinal cord. Another growth factor which is expressed in ependymal cells, at least during development, is GDNF, which has been demonstrated in human prenatal cortical plate ependyme from the age of 10 weeks onward.

Müller and Bergmann Glial Cells

Müller glial cells are persisting radial glial cells in the retina and express a large variety of cytokines and their receptors. Thus, expression of NGF, BDNF, and NT-3, along with neurotrophin receptors trkB and trkC, but not of trkA, has been reported. Also, glial maturation factor-β, another cytokine, is synthesized in these cells. Human Müller cells *in vivo* express high amounts of FGF-2. In addition, FGF-2 has been shown to downregulate IGF-1 in Müller cells. *In vitro*, Müller cells release TNF-α and nitric oxide upon stimulation with lipopolysaccharide and IFN-γ. TGF-β2 and -β3 are also expressed in Müller cells. Glucose and pH influence expression of VEGF in Müller cells. VEGF183 has been reported as a Müller cell-specific splice variant of this growth factor. Müller cells also express CNTF and FGF-2.

Bergmann glia, the persisting radial glia in the cerebellar cortex, express some growth factors and growth factor-related proteins. Thus, IGFBP2 is expressed in these cells during postnatal cerebellar maturation in parallel to the expression of IGF-1 in cerebellar Purkinje neurons. Bergmann glia cells also express IL-6.

Role of Glial Growth Factors in Glial Cell Differentiation and Function and the Formation of Glial Brain Tumors

An important function of glial growth factors is para- and autocrine regulation of the differentiation and maturation of glial cells, a feature which is of utmost importance for pathological changes of the nervous system (Figure 1).

Autocrine Regulation of Astrocyte Differentiation

With regard to autocrine regulation of astrocyte differentiation, studies from the 1980s reported that astrocyte-conditioned medium is able to inhibit astrocyte proliferation compared to fresh medium – a feature that has been attributed to the secretion of growth inhibitory factors by these cells.

The complex network of para- and autocrine effects of glial cells during astrocyte development is highlighted by the fact that differentiation of O-2A progenitor cells into either oligodendrocytes or type 2 astrocytes seems to depend on growth factors released from type 1 astrocytes. Type 1 astrocytes have been shown to secrete PDGF, thus stimulating O-2A progenitor cell proliferation and oligodendrocyte differentiation. Later, type 1 astrocytes secrete CNTF, thus initiating differentiation of type 2 astrocytes.

The LIF-receptor and its ligands play an important role in astrocyte differentiation. Thus, neural precursor cells isolated from LIFR$^{-/-}$mice fail to generate glial fibrillary acidic protein (GFAP)-positive astrocytes, whereas precursors from heterozygous mice differentiate normally. Also, *in vivo* LIFR$^{-/-}$ mice show low levels of GFAP. In contrast, animals deficient in LIF show reduced GFAP levels only in some brain regions, suggesting that LIF is not the predominant endogenous ligand for the LIF receptor.

Region-specific effects of glia-derived factors on astrocyte differentiation can also be observed in the cerebellar cortex, in which granule cell precursors are differentiated into astroglial cells by sonic hedgehog and bone morphogenic peptides. Growth factor-stimulated cells initially express both GFAP and neuronal markers and later switch to S100-beta, a marker of differentiated astrocytes.

Despite their important role in neuronal stem cell proliferation and differentiation, with regard to astrocytes, FGF-1 and FGF-2 are primarily involved in the regulation of astrocyte proliferation. Nevertheless, GFAP immunoreactivity is reduced in mice with a genetic defect in FGF-2. Also, Fgf-8b, a splice variant of this growth factor, has been shown to promote astroglial differentiation of a subpopulation of E15 cortical precursor cells in culture.

Regulation of Oligodendrocyte and Schwann Cell Differentiation by Glial Growth Factors: Importance for Etiology and Treatment of MS

A central role in paracrine and autocrine regulation of oligodendrocyte differentiation is played by neuregulins, which in the normal human CNS are produced by astrocytes and neurons. Thus, NRG-1α and -1β and their receptors can be found in cultured oligodendrocytes from neonatal rat pups. Under these conditions, less differentiated oligodendrocytes contain both NRG isoforms in the cell bodies but not in the processes, whereas only NRG-1β was found in the nucleus. In contrast, differentiated oligodendrocytes contained both isoforms only in cytoplasm and cell processes. Similar effects could be observed for embryonic striatal oligodendrocyte precursor cells, which express Nrg-1 as well as its specific receptors, ErbB2 and ErbB4, but not ErbB3. Likewise, inhibition of Nrg-1 activity by the addition of soluble ErbB3 decreases the mitotic activity of Nrg-1 on oligodendrocyte precursor cells.

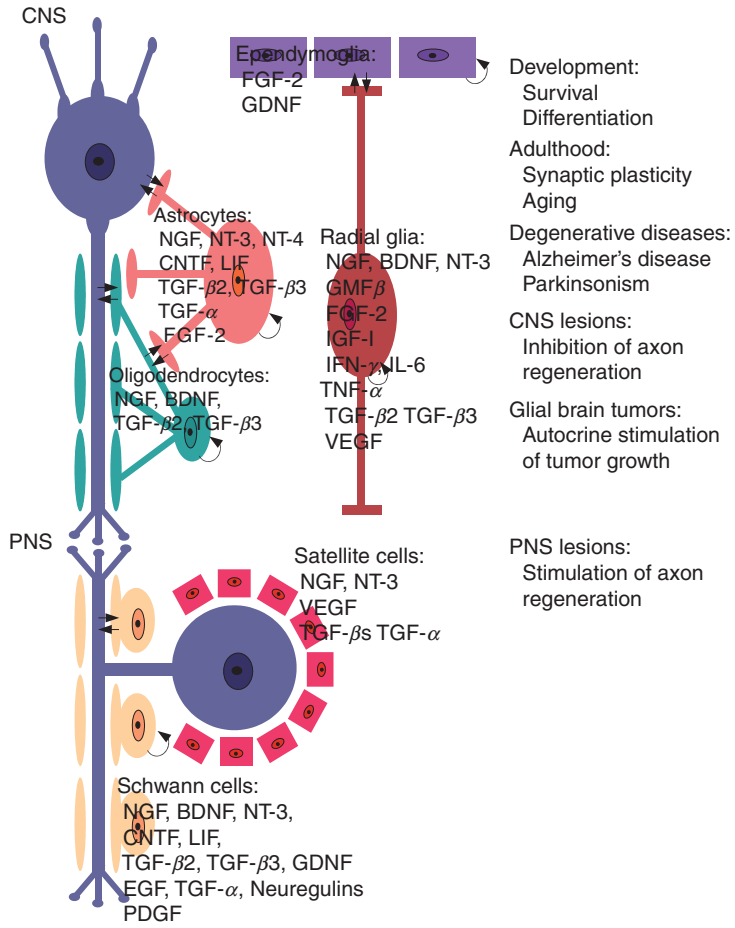

Figure 1 Graphical summary of expression and effects of glial growth factors in the central and peripheral nervous systems. A sophisticated network of glia–neuronal, interglial, and autocrine actions of glial growth factors regulate normal development and function of the nervous system, degeneration and regeneration after nerve injury, as well as brain tumor growth.

Despite these *in vitro* findings, neuregulins seem to also be important for myelination *in vivo* since mice with a null mutation in the nrg-1 gene show defects in oligodendrocyte differentiation. This also seems to be important for the pathology of MS since in active and chronically active MS lesions, expression of astrocyte-derived neuregulin is reduced and thus may contribute to impaired remyelination in MS patients.

Other glial growth factors influencing oligodendrocyte differentiation and myelination are PDGF and IGF-1. Thus, in spinal cord explants, NRG-1 and PDGF, but not LIF, enhance myelination. PDGF has also been shown to be a potent regulator of oligodendrocyte progenitor migration and proliferation in oligodendrocyte cultures. IGF-1 acts on myelin-forming cells to promote normal myelination and remyelination after injury. Thus, in experimental MS models, the neuregulin isoform GGF-2, IGF-1, and several neurotrophins promote remyelination following inflammatory demyelination.

Differentiation-stimulating effects of IGF-1 on oligodendrocyte precursors have also been detected *in vitro*, in which the addition of IGF-1 induced a high proportion of precursor cells to differentiate into galactocerebroside-positive oligodendrocytes, whereas the proportion of type 2 astrocytes was unaffected. In addition, IGF-1 promotes proliferation of O-2A precursor cells.

Also in myelinating Schwann cells, growth factors released from glial cells are involved in remyelination after damage. Thus, in neuron–Schwann cell cocultures, GGF-1, an NRG-1 isoform, inhibits myelination by preventing axonal segregation and ensheathment. In addition, treatment of established myelinated cultures with GGF-1 results in demyelination that frequently begins at the paranodes and progresses to the internodes. In contrast, FGF-2 and TGF-β1 inhibited myelination but did not cause demyelination, suggesting that this effect is specific to the NRGs. The NRG receptor proteins erbB2 and erbB3 are expressed on ensheathing and myelinating Schwann cells and rapidly phosphorylate upon GGF-1 treatment.

In Schwann cells, both NRG-1β and -1α are colocalized in cytoplasm and its processes. The Schwann

cell nucleus has weak immunoreactivity for both NRG-1 isoforms, although NRG-1β is predominant. ErbB2 and ErbB3 receptors, transducing the NRG-1 signal in Schwann cells, are found throughout cytoplasm and processes and are also localized in the nucleus. Stimulation of Schwann cells with mitotic agents induces nuclear translocation of NRG-1β.

Regulation of Radial Glia Differentiation

Little information is available on the role of glial growth factors in the differentiation of radial glia. Thus, in the mammalian retina proliferation of Muller glia is stimulated by EGF in a dose-dependent manner, whereas astrocyte proliferation is stimulated by FGF-1 and FGF-2. In contrast, proliferation of glial precursor cells is stimulated by FGF-1, FGF-2, and PDGF, but not by EGF. The inhibitory role of FGF signaling for Muller cell differentiation is further supported by the fact that inhibition of the endogenous FGF receptor by cotransfecting a dominant-negative form results in an increased number of Muller cells, suggesting a balance between FGF signaling and other signaling cascades to modulate retinal precursor cell fate.

Growth Factors in Glial Brain Tumors

Genetic changes in glial cells and their precursors lead to some of the most devastating human brain tumors, such as glioma, glioblastoma, and astrocytoma. During the process of tumor transformation, para- and autocrine stimulation of tumor cell proliferation is a key mechanism for tumor growth providing at the same time a starting point for pharmacological intervention of this process.

Glioma cells are among the oldest known sources for a number of different growth factors, such as FGF-2 and GDNF, which have also been shown to affect glial cell proliferation in an autocrine manner. Thus, glioma cells express high levels of FGF-2 protein and high-affinity FGF receptors and therefore are mitogenically responsive to FGF-2. Likewise, a knockdown of GDNF or its receptor, GFR-α1, by antisense RNA results in reduced proliferation of rat C6 glioma cells, suggesting a role in autocrine growth stimulation.

A growth factor that has been implicated in the regulation of glioblastoma growth is PDGF, which is coexpressed with its receptor in this tumor type and has been shown to regulate tumor cell proliferation. Because it is a secreted factor, PDGF not only has autocrine effects on producing cells but also paracrine actions on other tumor cells and on the tumor microenvironment. Thus, PDGF is involved in the regulation of tumor cell migration and tumor angiogenesis. In addition, human glioblastoma cells show a general increase in the expression of autocrine growth regulators such as TGF-α, TGF-β, and FGF-2 compared to normal human brain tissue.

Also in astrocytomas, activity and regulation of a number of mitogenic signaling pathways is aberrant. Thus, upregulation of growth factor receptors such as EGFR, PDGFR, and c-Met, as well as signaling intermediates such as Ras, protein kinase C, and others which have been shown to positively regulate tumor proliferation and cell cycle progression, has been demonstrated.

Effects of Glial Growth Factors on Differentiation and Function of Neurons in Both Healthy and Injured Nervous Systems

Glial growth factors are important regulators of normal neuronal differentiation and of neuron survival in the developing and adult brain. Disturbances of glial growth factor release are therefore supposed to play a role during CNS and PNS lesioning and in the etiology of various neurological disorders (Figure 1).

Regulation of Neuronal Differentiation during Development

Glial growth factors are important for the migration of neuronal progenitors during development. Thus, intraventricular injection of NT-4 or overexpression of BDNF lead to defects in cortical layering. Cortical layering defects can also be observed in FGF-2 knockout mice. GGF-1 promotes the migration of neuronal progenitors along radial glial fibers.

Glial growth factors can also influence axonal pathfinding, as shown for netrin-1 in the internal capsule. Netrin-1 is released from oligodendrocytes but not astrocytes. Together with netrin-1, engrailed-1 is directing axons of association neurons, projecting ipsilaterally to motor neurons in the spinal cord. Consistent with this, axonal pathways are altered in netrin-1-deficient mice. A number of glia-derived growth factors, such as TGF-α, have been shown to induce neuritogenesis.

An important role for glial growth factors is the regulation of neuron survival during periods of ontogenetic neuron death. Thus, FGF-2 can prevent neuron loss in the chick ciliary ganglion. Ciliary neurotrophic factor is also able to rescue motor neurons in the embryonic chick lumbar spinal cord during ontogenetic cell death. Moreover, TGF-βs enhance survival of chick ciliary ganglionic neurons *in vitro* synergistically with different neurotrophins and CNTF. Surprisingly, immunoneutralization of TGF-βs during chick development *in vivo* enhances neuronal survival.

Glia-Derived Growth Factors during Puberty and Aging

A late developmental event is puberty, during which gonadal steroids induce plastic changes in certain brain regions. In hypothalamic astrocytes, this is accompanied by the production of TGF-α and neuregulins, which elicit astroglial secretion of prostaglandin E2, stimulating the release of neuronal luteinizing hormone-releasing hormone. Hence, overexpression of TGF-α in the hypothalamus accelerates puberty, whereas blockade of TGF-α or neuregulin delays this process.

Growth factor synthesis by glial cells is also changed during aging, in which, together with astrogliosis, a continuous upregulation of TGF-β1 can be observed. Also, a decrease in NGF has been observed in a senescence accelerated mouse strain. Likewise, in aging mice an increase in IL-6 and a decrease in IL-10 release from glial cells can be observed.

Glia-Derived Growth Factors during Neurodegenerative Diseases

Neuronal degeneration during Parkinson's disease (PD) and Alzheimer's disease (AD) is accompanied by astrogliosis and thus upregulation of several glia-derived growth factors. On the other hand, some glial growth factors are downregulated in these diseases, such as FGF-2 expression, which is reduced in the substantia nigra of PD patients, and NGF expression, which is reduced in the hippocampus of both PD and AD patients. *In vitro* studies on growth factors such as FGF-2, GDNF, CNTF, and BDNF suggest that they are important for survival of dopaminergic midbrain neurons.

According to this, altered levels of glial growth factors are common in postmortem samples of the brains of AD patients. Thus, β-amyloid of senile plaques induces IL-1β and FGF-2, probably stabilizing the disease conditions. Whereas FGF-2 is expressed in astrocytes in the center of senile plaques, FGF-1 can be found in astrocytes surrounding the plaque. Also, other growth factors are upregulated in astrocytes of the brains of AD patients, such as endothelin-1, TGF-β2, IGF-1, and hepatocyte growth factor. In contrast, expression of BDNF is reduced in astrocytes surrounding senile plaques.

Role of Glia-Derived Growth Factors in Neural Regeneration in the Central Nervous System

It has long been known that neuron regeneration after lesioning is enhanced by soluble proteins released from glial cells (Table 1). Thus, NGF prevents death and promotes fiber growth of transected cholinergic neurons. Similarly, the release of IGF-1 from astrocytes is involved in neuron regeneration after cuprizone-elicited demyelination of the CNS. Notably, IGFs are also potent regulators of remyelination following CNS damage.

Glia-derived factors such as Nogo are also involved in the inhibition of axonal regeneration after spinal cord lesioning. Inhibition of the Nogo receptor could provide a means to overcome this obstacle to neuronal regeneration after spinal cord transection. However, Nogo is not the only axon outgrowth inhibitor

Table 1 Growth factor expression in different glial subpopulations

	SC		*Oligodendrocytes*		*Astrocytes*		*MG*	*BG*
	Normal	*Lesioned*	*Normal*	*Lesioned*	*Normal*	*Reactive*		
NGF	++	++++	+++	nd	+++	++++	++	nd
BDNF	+	+++	+++	nd	–	+++	++	nd
NT-3	++	–	nd	nd	++	nd	++	nd
NT-4	+	++	nd	nd	++	nd	nd	nd
CNTF	+++	+	–	–	+	+++	++	nd
LIF	+	++	nd	nd	+	+++	nd	nd
TGF-β1	–	+++	+	+++	–	+++	–	nd
TGF-β2	++	+	nd	nd	++	nd	++	nd
TGF-β3	+++	+	nd	nd	++	nd	++	nd
GDNF	+	+++	nd	nd	–	+++	nd	nd
TGF-α	++	++++	nd	nd	+	nd	nd	nd
Neuregulins	++	nd	nd	nd	++	nd	nd	nd
PDGF	++	nd	nd	nd	nd	nd	nd	nd
FGF-2	–	nd	nd	nd	+	+++	++	nd
IGF-1	nd	nd	nd	nd	nd	nd	++	nd

BDNF, brain-derived neurotrophic factor; BG, Bergann glia; CNTF, ciliary neurotrophic factor; FGF-2, fibroblast growth factor-2; GDNF, glial cell line-derived neurotrophic factor; IGF-1, insulin like growth factor 1; LIF, leukemia inhibitory factor; MG, Müller glia; nd, not determined; NGF, nerve growth factor; NT, neurotrophin; PDGF, platelet-derived growth factor; SC, Schwann cells; TGF, transforming growth factor.

since other astroglial molecules, such as collagen IV, are similarly active.

Also in brain regions in which axonal regeneration occurs throughout life, astrocyte-derived growth factors such as NGF, BDNF, GDNF, and neurturin seem to be involved since transplantation of astrocytes from the olfactory bulb to the spinal cord promotes axon regrowth after transection.

PDGF upregulation in astrocytes of the facial nucleus after transection of the facial nerve is important for neuronal regeneration. Similarly, astroglial CNTF expression is increased in fields of axonal sprouting in the deafferented hippocampus after entorhinal cortex lesioning. Likewise, BDNF, NT-3, and GDNF are induced in astrocytes of the spinal cord adjacent to a lesion.

Role of Glia-Derived Growth Factors in Regeneration of Peripheral Nerves

Transection of a peripheral nerve leads to apoptotic cell death of dorsal root ganglion neurons. This is probably due to a reduction in the expression of neurotrophins and neurotrophin receptors, and as a consequence, exogenously applied growth factors are able to counteract these effects. An important endogenous source for growth factor release in the distal nerve stump after peripheral nerve transection is Schwann cells, which release NGF, BDNF, NT-4/5, and p75. In addition to these factors, CNTF and LIF are needed for axonal regeneration, which act synergistically on nerve regeneration. Also, GDNF is upregulated in Schwann cells of the distal nerve stump after nerve injury and remains so for more than 5 months.

See also: Bergmann Glial Cells; Gap Junctions and Hemichannels in Glia; Glia Control of Blood Flow; Macroglial Lineages; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Schwann Cells and Axon Relationship; Schwann Cell Development.

Further Reading

Aloisi F (2003) Growth factors. *Neurological Science* 24 (Supplement 5): S291–S294.

Böttner M, Krieglstein K, and Unsicker K (2000) The transforming growth factor-betas: Structure, signaling, and roles in nervous system development and functions. *Journal of Neurochemistry* 75: 2227–2240.

Campbell K (2003) Signaling to and from radial glia. *Glia* 43: 44–46.

Cui Q (2006) Actions of neurotrophic factors and their signaling pathways in neuronal survival and axonal regeneration. *Molecular Neurobiology* 33: 155–179.

Decker L, Lachapelle F, Magy L, et al. (2005) Fibroblast growth factors in oligodendrocyte physiology and myelin repair. *Ernst Schering Research Foundation Workshop* 53: 39–59.

English AW (2003) Cytokines, growth factors and sprouting at the neuromuscular junction. *Journal of Neurocytology* 32: 943–960.

Miller RH (2002) Regulation of oligodendrocyte development in the vertebrate CNS. *Progress in Neurobiology* 67: 451–467.

Panickar KS and Norenberg MD (2005) Astrocytes in cerebral ischemic injury: Morphological and general considerations. *Glia* 50: 287–298.

Reuss B and von Bohlen und Halbach O (2003) Fibroblast growth factors and their receptors in the central nervous system. *Cell and Tissue Research* 313: 139–157.

Schwab ME (2002) Repairing the injured spinal cord. *Science* 295: 1029–1031.

Strelau J and Unsicker K (2003) Neuroregeneration. *Advances in Neurology* 91: 95–100.

Teismann P, Tieu K, Cohen O, et al. (2003) Pathogenic role of glial cells in Parkinson's disease. *Movement Disorders* 18: 121–129.

Glial Influence on Synaptic Transmission

C G Schipke and O Peters, Charité University Medicine, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Glia cells in the central nervous system are composed of three major cell types – microglia, oligodendrocytes, and astrocytes. Microglia are the innate immune cells of the brain, and oligodendrocytes are the myelin-forming cells; thus, these two cell types, due to their morphology and specialized role in the mammalian brain, do not influence synaptic transmission directly at the synapse. Nevertheless, microglia release cytokines and chemokines and thus influence synaptic transmission in large areas or the whole brain by altering the levels of these neuromodulatory substances. Oligodendrocytes have not been described to be directly involved in the influence on synaptic transmission, although they do play an important role in potassium buffering (along fiber tracts) and therefore contribute to ion homeostasis, another important factor for synaptic transmission.

Direct influence on synaptic transmission has only been described for astrocytes that have the appropriate morphological properties and thus comply with the requirements to be directly involved in the functional control of single synapses. During the mid-1990s, the first indications that astrocytes might be involved in the modulation of synaptic transmission came from a number of independent studies of cell cultures, in which neurons and astrocytes were co-cultured in a single petri dish. These studies described the general ability of astrocytes to release glutamate in a Ca^{2+}-dependent manner and to consecutively influence neuronal calcium activity. However, a substantial limitation of studies in co-cultures is the artificial environment and the altered morphology of cells, especially astrocytes in culture. For this reason, this article focuses on the mechanisms of glial influence on synaptic transmission that have been described from experiments using acute tissue preparations of the mammalian central nervous system, with the exception of the section of this article on the glia-derived acetylcholine-binding protein. Apart from the brain, a well-studied model system for neuron–glia and glia–neuron interactions is the mammalian retina with its layered structure. The mammalian retina comprises different types of glial cells and the glial influence on synaptic transmission in this preparation has been shown in some pioneering work for both excitatory and inhibitory neuronal synaptic transmission. This work and the emerging concepts are discussed later.

As previously mentioned, the astrocyte morphology in the intact brain tissue preparation is an important prerequisite for the ability of astrocytes to directly influence neuronal transmission at the synapse: with their fine appendages, often measuring only a few hundred nanometers in diameter, the astrocytes protrude into the extracellular space surrounding neuronal structures. These fine astrocytic structures seal the synapses from the extracellular space, but they contribute much more than just morphological support. The tight astrocytic coverage of synaptic structures limits the room for diffusion for secreted neurotransmitters, thereby influencing the active concentration of transmitters at the synapse. For the major excitatory neurotransmitter glutamate, astrocytes possess high-affinity transporters, thereby also contributing largely to the termination of the synaptic signaling.

Based on the first studies of astrocyte–neuron co-cultures, a completely new concept of synaptic function and plasticity has evolved and is currently well accepted – the concept of the 'tripartite synapse' (Figure 1). The traditional view of the functional synapse consisting of the presynaptic terminal releasing the neurotransmitter into the extracellular space and binding to postsynaptic neurotransmitter receptors of the adjacent neuron has been modified by adding a third, functionally important element – the astrocyte. It is now clear that astrocytes can release a number of neurotransmitters and neuromodulatory substances, especially glutamate, ATP, and D-serine, thereby directly influencing neuronal signal transmission at single synapses. Most often, the release of neurotransmitters from astrocytes is accompanied by changes of their intracellular Ca^{2+} concentration in response to neuronal activity. Thus, the activation of astrocytes by neuronal activity and the regulation and function of calcium excitability in astrocytes are extremely important for any kind of astrocyte–neuron communication. This article focuses on the mechanisms that influence synaptic transmission following initial activation of astrocytes, measured as changes in the intra-astrocytic calcium level.

It is important to take into consideration that a single hippocampal astrocyte in the CA1 area of the hippocampus reportedly occupies a neuropilar volume

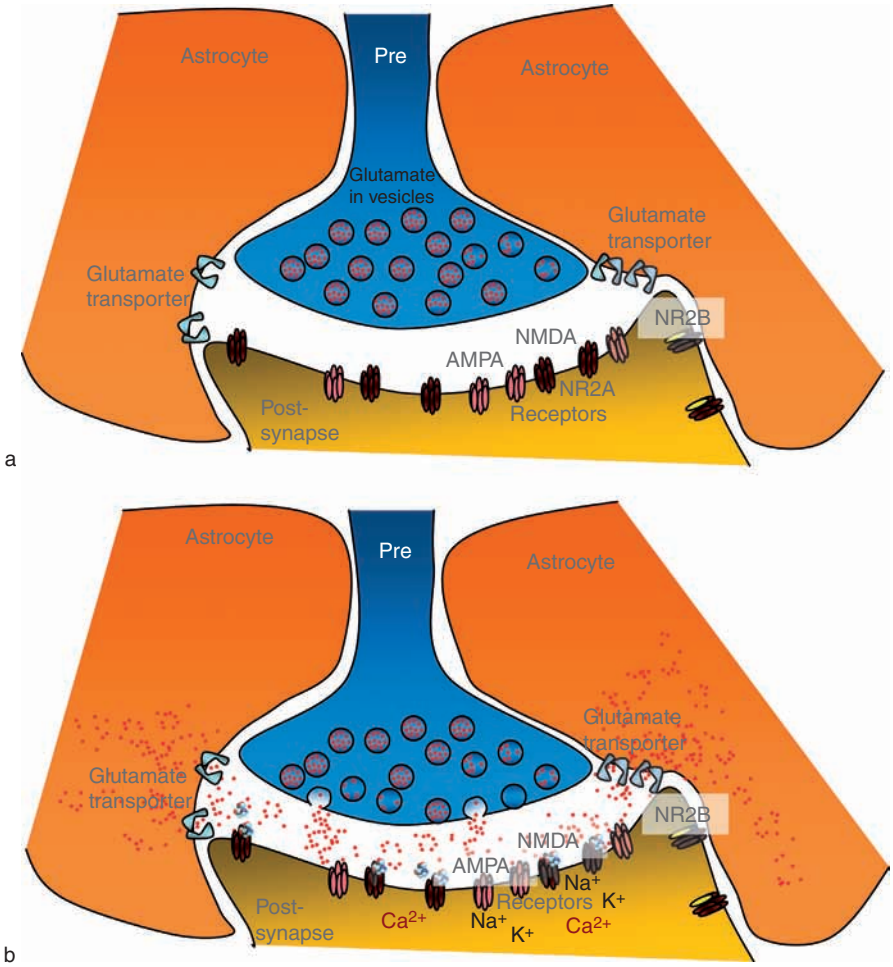

Figure 1 A glutamatergic synapse in the central nervous system. (a) The functional tripartite synapse consists of the presynapse (blue), the postsynapse (yellow), and the astrocytic coverage (orange). The presynapse is loaded with synaptic vesicles containing glutamate. On the postsynapse, different subtypes of glutamate receptors, AMPA- and NMDA-type receptors, are expressed. Also, extrasynaptic glutamate receptors are found on the postsynaptic element. The astrocyte is equipped with highly efficient glutamate transporters. (b) Upon the arrival of an action potential, glutamate is released into the synaptic cleft, binds to postsynaptic glutamate receptors, and excites the postsynapse. The influx of Na^+, K^+, and Ca^{2+} ions through glutamate receptors depolarizes the postsynapse. This depolarization is then propagated within the neuron. Glutamate is transported away from the synaptic cleft by the activity of astrocytic glutamate transporters. This transport determines the time course of the neuronal excitation because excitation stops when all glutamate is removed from the synaptic cleft. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; NMDA, *N*-methyl-D-aspartate.

of approximately 65 000 μm^3, thereby ensheathing thousands of synapses. Values for other brain regions have not been determined, but there is no indication that these values differ substantially. These thousands of astrocyte-ensheathed synapses can be of different kinds in terms of transmitters secreted. Thus, a single astrocyte is in position to locally react to neuronal activity but also to serve as an integrator of synaptic activation. In addition, astrocytes communicate over long distances via the intra-astrocytic propagation of calcium waves.

Taken together, astrocytes are in the position to react accordingly and are equipped with the instrumentation to modify synaptic transmission at synapses remote from those where the activity was sensed originally.

Influence of Glia-Derived Cytokines on Synaptic Transmission

It is unknown whether glia are involved in the rapid continual maintenance of synaptic strength. Synaptic strength is determined not only by the amount of transmitter released into the synaptic cleft or by the balance of excitation and inhibition but also by the availability of functional neurotransmitter receptors at the postsynapse. An influence on synaptic transmission via controlling the turnover of receptors cannot be exerted directly at the synapse but, rather, via influencing the receptor turnover by interfering with the neuronal function in general. In hippocampal slices, a cytokine produced by glia cells, tumor

necrosis factor-α (TNF-α), has been shown to be critically involved in the maintenance of synaptic strength by controlling the surface expression of AMPA-type glutamate receptors. Thus, only the continual presence of glia-derived TNF-α secures glutamatergic synaptic transmission at a constant level, thereby also introducing a possible way for glia modulation of synaptic transmission. A different form of influence on synaptic transmission is involved in the modulation by glia-derived interferon-γ (IFN-γ). The long-term presence of this proinflammatory cytokine influences the receptor clustering of AMPA-type glutamate receptors at the postsynapse, which modulates synaptic efficacy. Thus, by controlling the extracellular level of certain cytokines, glia have a role in the short-term (via TNF-α) and long-term (IFN-γ) modulation of synaptic transmission.

Astrocytic Modulation of Synaptic Transmission in the Retina

Since the beginning of research on the functional interactions at the tripartite synapse, the retina has been a very valuable preparation to study and describe the different mechanisms in glia–neuronal cross-talk. Numerous mechanisms that were first described in the retina were later demonstrated to also be functional in the brain. Astrocytes first activated by neuronal activity have been shown to be able to signal back and to modulate neuronal activity in this preparation. Three different, independent mechanisms of glial influence on synaptic transmission in this preparation have been demonstrated to be functional: inhibitory modulation of neuronal spiking, excitatory modulation of neuronal spiking, and hyperpolarization of neurons and thus inhibition of spontaneous spiking activity of neurons. The last mechanism has been shown to be dependent on ATP release from glia cells. The first two mechanisms need to be elucidated in detail, but it is clear that glial interaction with neuronal excitatory, as well as inhibitory signal transmission is involved.

Morphological Plasticity of Astrocytes as a Modulator of Synaptic Transmission

Although the morphology of astrocytes with regard to coverage at single synapses suggests a role for morphological plasticity in modulating synaptic transmission, there are few experimental paradigms to study this assumption. This is mainly because the size of a single synaptic structure is much smaller than the resolution of a conventional microscope. However, within the female rodent brain a remarkable form of glial morphological plasticity can be observed within a well-defined brain region, the supraoptic nucleus (SON), a structure within the hypothalamus. The morphological plasticity in this particular brain region is induced by strong secretion of neurohypophysial hormones in lactating animals and includes rapid and reversible changes in the astrocytic coverage of the neurons and their synapses. Consequently, the SON serves as a structure that allows the comparison of synaptic transmission in a specified subgroup of neurons under different conditions of astrocytic neuronal wrapping. As previously mentioned, an important functional role of astrocytes covering synaptic structures is to remove the neurotransmitter glutamate from the synaptic cleft, as well as limit the diffusion of neurotransmitters into the extracellular space. Thereby astrocytes influence glutamatergic neurotransmission. Astrocytes control the intersynaptic cross-talk with their fine processes. Furthermore, they are morphologically positioned to sense communication mediated by extrasynaptic transmission. A number of electrophysiological and histological studies have found that, indeed, astrocytes contribute to synaptic transmission in the SON by controlling the clearance and diffusion of glutamate in the extracellular space. Glutamate which is synaptically released in the SON not only binds to the postsynaptic ionotropic glutamate receptors but also binds to presynaptic modulatory metabotropic glutamate receptors (mGluRs). Activation of mGluRs leads to a reduction of glutamate release from glutamatergic synapse by a negative feedback loop. The activation of presynaptic mGluRs is a tonic action, thereby rendering this mechanism extremely sensitive to changing levels of glutamate. When neuronal signal transmission in virgin and lactating animals is compared, experiments clearly show that in lactating animals more glutamate is available in the extracellular space when astrocytic processes are farther away from the active glutamatergic synapses and thus are less effective in removing glutamate. This phenomenon can be measured as a strong reduction of efficacy at excitatory synapses accompanied by higher glutamate levels, leading to a stronger activation of presynaptic mGluRs, which in turn leads to a reduction of glutamate release. However, it appears that this is not the only influence of glia on synaptic transmission in the SON. In lactating animals, under conditions of reduced glial coverage of neurons and synapses, the intersynaptic cross-talk between glutamatergic and GABAergic synapses is also modulated. Glutamate which is released into the synaptic cleft also binds to presynaptic mGluRs at neighboring GABAergic terminals, leading to a potentiation of transmitter release and to a phenomenon called heterosynaptic depression. Heterosynaptic depression is much more pronounced in lactating animals, in which glutamate can easily diffuse to neighboring synapses and is cleared less

efficiently, than in virgin animals. Whether this effect is a consequence of less effective glutamate uptake or of facilitated diffusion of glutamate in the extracellular space is not clear. Taken together, these results from SON clearly demonstrate that alterations in the morphological relationship between astrocytes and neurons strongly affect synaptic function.

Release of Neurotransmitters and Neuromodulators from Astrocytes to Modulate Synaptic Efficacy

The discovery that astrocytes are capable of releasing neurotransmitters and that this process is a regulated action led to the assumption and, later, experiments clearly demonstrating that astrocytes are capable of modulating synaptic transmission and neuronal excitability via the release of gliotransmitters into active neuronal synapses. These gliotransmitters are not different from neurotransmitters; they are almost identical molecules. The term gliotransmitter is therefore mainly used to emphasize that a given molecule is released from glia cells. Generally, modulation of synaptic transmission via the release of neuroactive substances can be achieved in two ways:

1. Directly via the release of neurotransmitters, thereby either enhancing the stimulus or stimulating modulatory receptors at the active synapse.
2. Via the release of neuromodulatory substances, thereby modulating the function and efficacy of neurotransmitter and neurotransmitter receptors.

As mentioned previously, not all gliotransmitters are particular to astrocytes. Astrocytes use two of the most widely distributed transmitters – glutamate and ATP – to exert their influence on synaptic transmission. A third important gliotransmitter is D-serine. This potent modulator of synaptic transmission is exclusively synthesized by and released from astrocytes. It is therefore a special gliotransmitter in that neurons rely on astrocytes for the synthesis and release of this neuromodulatory molecule. D-Serine is a coagonist and potent modulator of *N*-methyl D-aspartate (NMDA) receptors. NMDA receptors are key players in glutamatergic excitatory synaptic transmission and have been implicated in many physiological processes, including learning and memory, but also in pathophysiological processes. Thus, influencing NMDA receptor function is a very potent way of modulating synaptic transmission with regard to the physiological process implicated in the basis of memory formation and learning. This might explain why a number of studies have demonstrated a direct action of astrocytes on neuronal NMDA receptors, either by activating or by modulating their function.

Release of Glutamate from Astrocytes to Influence Synaptic Transmission

One of the first molecules proposed to be released from astrocytes as a gliotransmitter was glutamate. The regulation and mechanism of its secretion from astrocytes have been well studied. Glutamate is released from vesicles in the same manner as it is released from presynaptic terminals. However, astrocytic release sites usually contain less vesicles and do not have a classical postsynapse as the target counterpart. Nevertheless, glutamate released from astrocytes exerts a number of important functions influencing and synchronizing neuronal activity. After the first results were obtained from *in vitro* co-culture (astrocyte plus neurons) models, the concept that astrocyte-derived glutamate can modulate synaptic transmission was first substantiated *in vivo* in the hippocampus. In these experiments, it was demonstrated that GABAergic activation of astrocytes, leading to an increase in the intracellular calcium concentration, can cause an NMDA receptor-dependent increase in miniature inhibitory postsynaptic current frequency detected in pyramidal neurons and a strengthening of certain inhibitory synapses. A number of experiments have shown that glia-derived glutamate acts on extrasynaptic ionotropic glutamate receptors or on presynaptic mGluRs to modulate neuronal transmitter release. A direct activation of neuronal NMDA receptors following glutamate release from astrocytes was first shown in the thalamus, in which spontaneously occurring Ca^{2+} elevations in astrocytes led to the induction of NMDA receptors-dependent currents in neighboring neurons. This again proves the Ca^{2+}-dependent glutamate release from astrocytes. Later, this finding was substantiated with comparable findings in the hippocampus, in which the astrocyte-mediated influence on synaptic transmission was analyzed in more detail. In these studies, a variety of stimuli, each of which led to astrocytic calcium oscillations, all caused NMDA receptor-mediated slow inward currents (SICs) in area CA1 pyramidal neurons. Detailed analysis of the pharmacology and kinetics of these currents and immunoelectron microscopy clearly demonstrates that the glutamate released from astrocytes selectively acts on extrasynaptic NMDA receptors that contain the NR2B subunit. The currents induced in neurons following astrocytic glutamate release are large and can lead to the modulation of neuronal function by influencing intraneuronal signaling pathways, such as modulating the pathways controlled by the cAMP response element binding protein. However, glial glutamate not only may be important for the function of single neurons but also may serve to synchronize neuronal activity as demonstrated in the hippocampus. A single astrocyte

can evoke simultaneous glutamate-mediated SICs in neighboring neurons. The functional consequences of this concerted excitation of neurons by astrocytes have yet to be determined, but it again proves the integrative role of astrocytes not only at single synapses but also across the multicellular network of the functional brain. Thus, glutamate released from astrocytes onto a number of different neuronal glutamate receptors is a potent way of influencing synaptic transmission either short term via modulating synaptic efficacy, synchronizing neuronal activity, or long term via modulating intraneuronal signal transduction pathways (Figure 2).

Release of D-Serine from Astrocytes to Influence Synaptic Transmission

D-Serine is emerging as one of the most important gliotransmitters because it strongly influences glutamatergic synaptic transmission and is exclusively released from astrocytes (Figure 3). The discovery and mode of action of D-serine are described elsewhere in this encyclopedia; however, the action of astrocyte-derived D-serine on neuronal NMDA receptors is briefly described here. The NMDA-type glutamate receptors require the binding of cofactors to allow the opening of the ion pore. Glycine is such a molecule, binding to the well-defined glycine binding site of the NMDA receptor as a required cofactor. In the past 15 years, it has become clear that D-serine is also an endogenous ligand at the glycine binding site of the NMDA receptor. D-Serine, the only D-amino acid to exert an important function in the human body, is solely produced in astrocytes. Astrocytic serine racemase converts the proteinogenic L-serine into D-serine. Evidence suggests that in some brain regions D-serine, and not glycine, is the only endogenous cofactor at the glycine binding site of NMDA receptors. Thus, secreted D-serine enables astrocytes to influence synaptic transmission by controlling the availability of D-serine and consequently the opening of NMDA receptors. Indeed, in hippocampal cultures devoid of astrocytes, the addition of D-serine is necessary to enable the induction of NMDA receptor-mediated synaptic plasticity, and in classical long-term potentiation (LTP) experiments the addition of D-serine can be a way to overcome the impairment of LTP induction in older animals. Interestingly, a role for astrocyte-derived D-serine in controlling metaplasticity has been demonstrated in the SON. In this brain structure, depending on the amount of glial coverage of synapses, the amount of D-serine released onto glutamatergic synapses is controlled. Because D-serine is the only cofactor for NMDA receptors in this brain region, astrocytes are able to determine whether a synapse exhibits LTP or long-term depression (LTD). When the astrocytic processes retract and the level of synaptic D-serine is reduced, LTD is induced, whereas in virgin rats that have a high degree of synaptic coverage, the same stimulus induces LTP. Thus, astrocytes also control forms of metaplasticity, influencing the overall signal integration in whole brain areas.

Release of Purines from Astrocytes to Influence Synaptic Transmission

In addition to glutamate, astrocytes also release the neurotransmitter ATP. There are much fewer studies

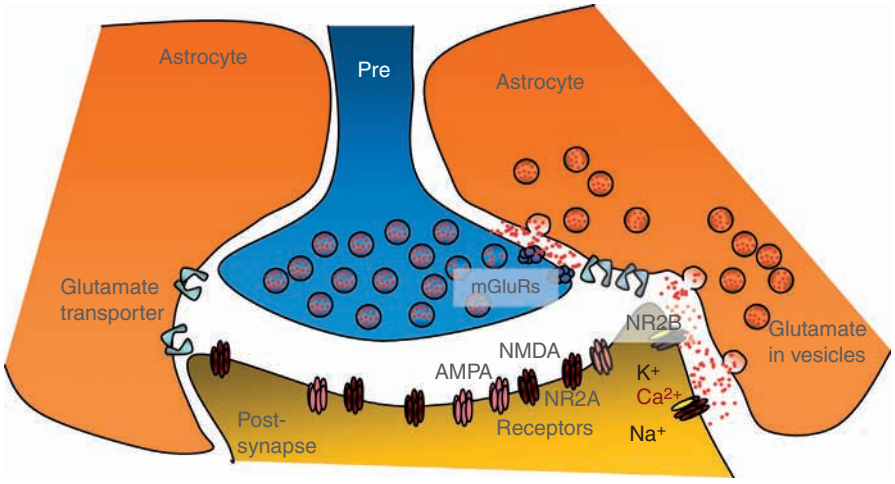

Figure 2 Glutamate released from astrocytes influences synaptic transmission. Astrocytes can release glutamate from vesicles in a Ca^{2+}-dependent manner. Following an increase in the cytosolic calcium concentration, glutamate is released and binds to presynaptic metabotropic glutamate receptors (mGluRs) and neuronal extrasynaptic NMDA receptors of the NR2B type. The activation of these receptors leads to modulation of synaptic transmission. Activation of mGluRs modulates the amount of transmitter released, and activation of NR2B receptors leads to activation of second messenger pathways that have long-term consequences in the signal integration in the neuron.

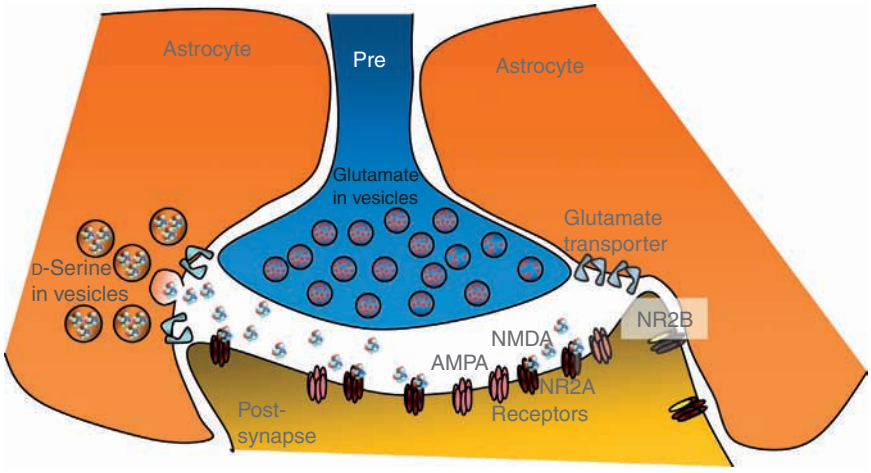

Figure 3 Astrocytes release D-serine. NMDA-type glutamate receptors require the binding of a cofactor to the so-called glycine site. This cofactor allows the receptor to open when glutamate binds. D-Serine is such a cofactor. D-Serine is exclusively synthesized in astrocytes and released into the synaptic cleft in a calcium-dependent manner. In certain brain regions, such as the supraoptic nucleus of the hypothalamus, D-serine is the only cofactor for the glycine site of NMDA receptors. Thus, the release of D-serine from astrocytes determines whether NMDA receptors on the postsynapse can open and transmit a signal when glutamate is released from the postsynapse.

on the release mode, as well as the mode of action of glia-derived ATP than studies on glutamate. Nevertheless, it has become clear, from studies of the retina and also from studies of the hippocampus, that glia-derived ATP very strongly modulates neuronal excitability. Purines in general, released from astrocytes, are emerging as the type of gliotransmitter that exerts its action even on remote synapses, at a distance far from synaptic activity that originally induced astrocytic activity. Purinergic modulation of synaptic transmission is mainly brought about by adenosine, although astrocytes originally release ATP. ATP is then rapidly degraded by different types of ectonucleotidases in the extracellular space to ADP, AMP, and, finally, adenosine. All these substances potentially activate different subtypes of purinergic receptors. ATP might be released in concentrations that are insufficient to activate neuronal P2 receptors but high enough to allow an accumulation of adenosine that will activate neuronal A1 receptors. Direct influence on synaptic transmission has been demonstrated for glia-derived ATP, degraded to adenosine in the hippocampus. It has been known for some time that presynaptically located adenosine A1 receptors cause presynaptic inhibition of transmitter release, a signaling pathway that is implicated in the physiological basis of heterosynaptic depression: active synapses that are potentiated as a consequence of their activation lead to the depression of neighboring synapses, thereby enhancing the contrast between different potentiated and unpotentiated pathways. It is clear that astrocyte-derived ATP, degraded to adenosine in the extracellular space, mediates this form of neuronal plasticity. A dramatically reduced level of heterosynaptic depression has been measured in the hippocampus of transgenic mice in which astrocytes are impaired in their ability to fuse vesicles at the membrane. Since the addition of ATP or adenosine rescues the effect, it is evident that vesicularly released ATP from astrocytes is the substance allowing the establishment of heterosynaptic depression. Astrocytes are activated by GABA following neuronal activation. GABA acts on astrocytic $GABA_B$ receptors and induces a calcium signal. The calcium signal is communicated in between astrocytes via gap junctions. Astrocytes remote from the synapses originally activated then release ATP, which, degraded to adenosine, leads to the presynaptic inhibition of transmitter release. This multicellular network, required for the induction of heterosynaptic depression, includes astrocytes as an important signaling partner; thus, the basis for functional modulation of synapses is not one astrocyte or two neurons but, rather, a network of astrocytes integrated into the neuronal signaling pathways (Figure 4).

Modulation of Synaptic Transmission Mediated by Astrocyte-Derived Acetylcholine-Binding Protein

A different way to influence synaptic transmission is to control the availability of active transmitter in the synaptic cleft by the presence of transmitter binding and thus neutralizing molecules. Such a molecule is known for acetylcholine. It is a protein that is secreted and binds acetylcholine. The acetylcholine-binding

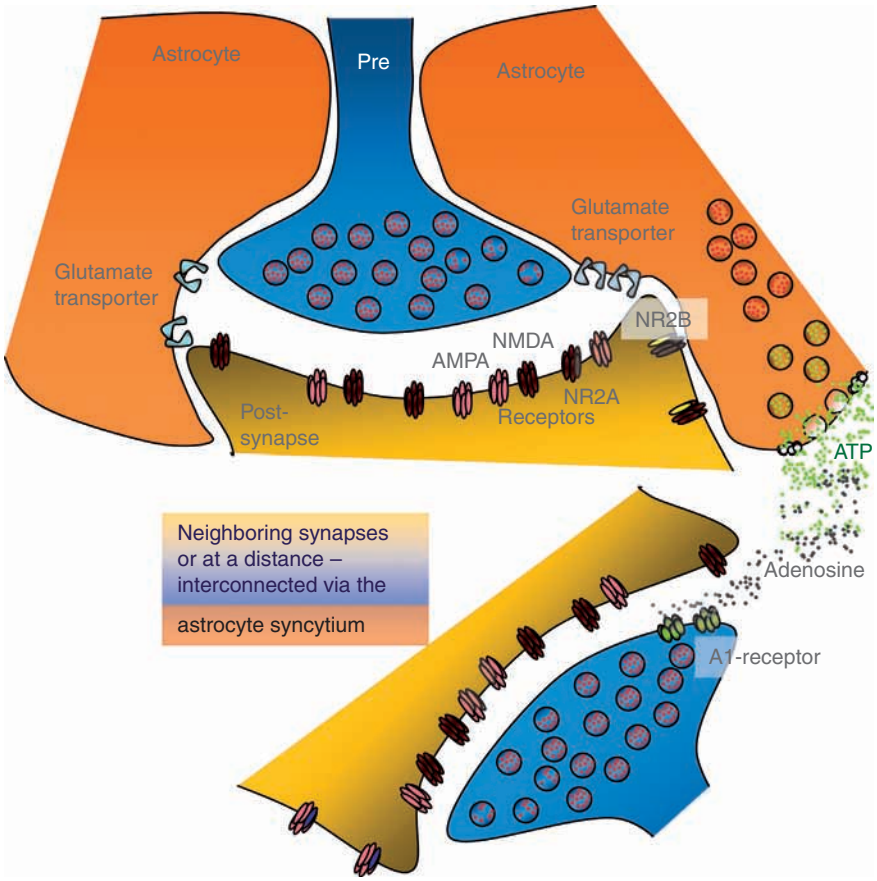

Figure 4 Astrocytes release ATP and influence synaptic transmission at neighboring synapses. Astrocytes release ATP from vesicles and possibly also via other mechanisms. In the extracellular space, ATP is rapidly degraded by ectonucleotidases to ADP, AMP, and, finally, adenosine. Adenosine activates presynaptically located A1 receptors. The activation of these receptors leads to an inhibition of neurotransmitter release from the presynapse. In the hippocampus, this mechanism is the physiological basis of heterosynaptic depression, a phenomenon involved in synaptic plasticity. Following communication through the astrocytic syncytium, synapses distant from the original point of activity can also be modulated.

protein (AChBP) is secreted from glia cells in the central nervous system of the freshwater snail, *Lymnaea stagnalis*, in which it modulates synaptic transmission. Like the nicotinic acetylcholine receptors, this astrocyte-derived molecule assembles into a heptamer with ligand-binding characteristics typical of a nicotinic receptor. Presynaptic release of acetylcholine induces the secretion of this AChBP from astrocytes, and once in the synaptic cleft, it acts as a molecular trap, binding the transmitter and reducing its availability at the synapse. Evidence for the functionality of this influence on synaptic transmission at mammalian synapses is lacking. Nevertheless, an imbalance of cholinergic signal transmission in a number of diseases, such as Alzheimer's disease and schizophrenia, is well described. The functionality of the AChBP-mediated pathway of glial influence on synaptic transmission is proposed as an explanation for the unresolved issue of the profound imbalance of the cholinergic system in these diseases.

See also: Adenosine Triphosphate (ATP); Adenosine Triphosphate (ATP) as a Neurotransmitter; Calcium Homeostasis in Glia; Calcium Waves in Glia; D-Serine: From its Synthesis in Glial Cell to its Action on Synaptic Transmission and Plasticity; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Glutamate Transporters: Electrophysiology; Glutamate; Microglia identification Methods; Neurotransmitter Release from Astrocytes; Oligodendrocyte and Schwann Cell Identification Methods; Potassium Homeostasis in Glia.

Further Reading

Auld D and Robitaille R (2003) Glial cells and neurotransmission: An inclusive view of synaptic function. *Neuron* 40: 389–400.

Chadwick DJ and Goode J (eds.) (2006) *Purinergic Signalling in Neuron–Glia Interactions*, Novartis Foundation Symposia. Chichester, UK: Wiley.

Hatton GI and Parpura V (eds.) (2004) *Glial–Neuronal Signaling*. Amsterdam: Kluwer.

Haydon PG and Carmignoto G (2006) Astrocyte control of synaptic transmission and neurovascular coupling. *Physiological Reviews* 86: 1009–1031.

Kang J, Jiang L, Goldman SA, et al. (1998) Astrocyte-mediated potentiation of inhibitory synaptic transmission. *Nature Neuroscience* 1: 683–692.

Kettenmann H and Ransom B (eds.) (2004) *Neuroglia*, 2nd edn. Oxford: Oxford University Press.

Newman EA and Zahs KR (1998) Modulation of neuronal activity by glial cells in the retina. *Journal of Neuroscience* 18: 4022–4028.

Panatier A, Theodosis DT, Mothet JP, et al. (2006) Glia-derived D-serine controls NMDA receptor activity and synaptic memory. *Cell* 125: 757–784.

Pascual O, Casper KB, Kubera C, et al. (2005) Astrocytic purinergic signaling coordinates synaptic networks. *Science* 310: 313–316.

Volterra A, Magistretti PJ, and Haydon PG (eds.) (2002) *The Tripartite Synapse*. Oxford: Oxford University Press.

Glial Ion Homeostasis: A Fluorescence Microscopy Approach

J-Y Chatton, University of Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Glial cells and, in particular, astrocytes have been assigned the main role of supporting neurons for their function in the brain, including buffering the extracellular potassium and providing energetic support. Advances in glial research have demonstrated a more active behavior of astrocytes, which are able to perform functions traditionally held by neurons, such as intercellular communication or neurotransmitter release, some of them directly or indirectly modulating neuronal network activity. To a large extent, the development of methodological approaches centered around fluorescence imaging has critically contributed to the new understanding of glial functions. This article discusses progress in imaging technology and then reviews some aspects of the current understanding of ion homeostasis in glial cells and their relationship with integrated cellular and network functions.

Technological Approaches

Fluorescence Imaging and Electrophysiology

Because glial cells do not generate action potentials and have a membrane potential that is generally maintained stable by an abundance of K^+ channels at their membrane, they have not attracted the attention of electrophysiologists as much as neurons have. Moreover, when studied in primary culture, astrocytes generally adopt an extremely flat shape with a thickness of a few micrometers, rendering them difficult to impale with microelectrodes. On the other hand, in the past decade a number of innovative imaging approaches of both high sensitivity and resolution – both temporal and spatial – that perfectly complement electrophysiological signal measurements have been developed.

The temporal resolution of electrophysiological techniques along with their exquisite sensitivity, which makes them useful for monitoring single channel kinetics, are currently not matched by those of imaging techniques, although the gap has become increasingly smaller. However, imaging approaches have several decisive advantages over electrophysiological techniques, the most obvious being the ability to monitor parameters with high spatial resolution. It is thus possible to simultaneously monitor responses from many cells or many sites within one cell. This approach has enabled several discoveries that encompass the long-range intercellular mode of communication called Ca^{2+} waves in glial cells.

The resolution of the optical microscope (~0.3 μm laterally and 0.5–1.0 μm axially) enables it to resolve subcellular structures such as organelles or large secretary vesicles. It can also detect, without spatially resolving them, a signal originating from smaller structures or even single molecules. Combining these two features has enabled the measurement of ion changes in microdomains barely accessible to microelectrodes, such as the mitochondrial matrix. Another advantage is that more than one fluorescent probe can be placed in a cell, which allows the monitoring of more than one parameter simultaneously, such as intracellular Ca^{2+} and Na^+ concentrations, or Ca^{2+} concentration changes in the cytosol and in organelles such as mitochondria. This feature provides a decisive advantage for a more complete and integrated understanding of cellular functions.

Organic Fluorescent Indicators

Since their first appearance in the late 1980s, a wide selection of ion-sensitive fluorescence molecules have been synthesized and made commercially available. A major advancement in the development of these probes was the idea of producing membrane-permeant analogs that render them capable of accumulating in cells. Cation-sensitive dyes are usually weak acids containing carboxyl groups and are normally negatively charged at physiological pH. Esterification of these molecules (using ethyl esters or acetoxymethyl ester) allows them to undergo passive diffusion across the cell membrane, and once inside the cell, esterases with broad selectivity hydrolyze the esters, releasing the free cation-sensitive and electrically negative molecule that remains trapped inside the cell. The following are major advantages of this loading method: (1) it is strictly noninvasive; (2) a large number of cells can be loaded simultaneously; (3) because the ester analog is continuously hydrolyzed inside the cell, there is an actual accumulation of the dye well above the concentration of the compound in the bath. Acetoxymethyl ester probes preferentially load into astrocytes in brain slices rather than neurons, which is a significant advantage for glial research.

Most of the probes discussed here can be purchased as membrane-permeant analogs from several sources. However, for fluorescent probes that do not exist as membrane-permeant precursors, other more invasive approaches have to be used, such as microinjection,

incorporation of the dye through the pipette in whole cell patch clamp studies, or pinocytosis.

pH Intracellular pH (i.e., H^+ concentration) was one of the first parameters successfully monitored by fluorescence imaging since the fluorescein molecule and several of its derivatives display intrinsic changes in their fluorescence characteristics (intensity and spectral shifts) depending on ambient pH. This opened the possibility of studying important membrane transporter activities involved in intracellular pH homeostasis. The most widely used pH dye is 2′,7′-bis-(2-carboxyethyl)-5- (and -6)-carboxyfluorescein, which has a p*K*a of ~7.00. This is close to the cellular pH and enables the monitoring of pH in an optimal dynamic range. This dye has an excitation maximum at a wavelength of ~490 nm, which shows pH sensitivity. When excited at ~440 nm, the fluorescence is no longer influenced by pH. This particular point in the spectrum is called the isosbestic point and is used as an internal reference point allowing correction for dye concentration changes or path length/thickness differences. This dye is called an excitation ratio probe. Other pH dyes, such as SNARF-1, display a similar pH dependency within their emission spectra and can be similarly used in the so-called emission ratio mode, which is particularly useful with confocal microscopy.

Ca^{2+} Intracellular Ca^{2+} has been recognized as a second messenger of prime importance critically influencing a number of normal cellular functions and playing a central role in pathological conditions, such as ischemia. The resting cytosolic Ca^{2+} concentration in glia is in the range of 100–200 nM and can reach low micromolar levels during transients. Its detection *in situ* by fluorescent probes is probably one of the elements responsible for the tremendous success of fluorescence imaging techniques. Ca^{2+} has thus become the most studied cation in cell biology, which is reflected by the wide array of probes made commercially available.

Ca^{2+} probes can be classified into low and high Ca^{2+} affinity probes, and they can also be classified according to their spectral properties (excitation in the ultraviolet (UV) or visible spectrum). The most widely used Ca^{2+} probes are Fura-2 (high affinity, UV) and Fluo-3 or Fluo-4 (high affinity, visible). Low affinity probes have a lower tendency to buffer intracellular Ca^{2+} changes and are more suited to resolve large amplitude Ca^{2+} transients. Examples of low-affinity probes are Fluo-5F (low affinity, visible) and Mag-Fura-2 (low affinity, UV). As a nonexcitable cell, the astrocyte does not usually generate large Ca^{2+} spikes, and high-affinity dyes are more often used.

In addition, changes in Ca^{2+} concentration in the microdomains facing the cell membrane (i.e., close to the channel mouth) are of different amplitudes and kinetics compared to bulk cytosolic Ca^{2+} changes. Lipophilic derivatives of Ca^{2+} probes (e.g., fura-C18 and FFP-18) that bind to the membrane surface can be used to selectively respond to Ca^{2+} changes only in the immediate vicinity of membrane channels and transporters. Finally, a class of Ca^{2+} probes can report changes in mitochondrial Ca^{2+}. These molecules (Rhod-2 AM and derivatives) harbor an intrinsic positive charge and preferentially localize in mitochondria of living cells, including glia and neurons.

Mg^{2+} Intracellular Mg^{2+} plays an important role as a mediator of enzymatic reactions, DNA synthesis, hormonal secretion, and muscular contraction. Cytosolic Mg^{2+} concentrations found in cells typically range from approximately 0.1 to 2 mM. Several probes exist with relatively good Mg^{2+} selectivity. As for Ca^{2+} probes, Mg^{2+} fluorescent indicators are available for visible or UV excitation. The most used Mg^{2+} probes are Mag-Fura-2 (UV) and Magnesium Green (visible). However, most Mg^{2+} probes exhibit substantial Ca^{2+} sensitivity, which makes them good low-affinity Ca^{2+} dyes but limits their application as Mg^{2+} probes in situations in which important intracellular Ca^{2+} changes are expected.

Monitoring of intracellular free Mg^{2+} has been used as an indirect but real-time *in situ* measurement of ATP hydrolysis. At the origin of this strategy is the fact that ATP displays a ~10-fold higher affinity for Mg^{2+} than ADP and binds a large proportion of cellular Mg^{2+}. As a consequence, upon hydrolysis of ATP into ADP, one Mg^{2+} ion is released, leading to a detectable increase in the free Mg^{2+} concentration.

Na^+ The development of fluorescent probes for measuring Na^+ has been more challenging than that of Ca^{2+} probes. The most widely used probe, sodium-binding benzofuran isophthalate (SBFI), is a UV excitable dye with similar spectral characteristics as the Ca^{2+} probe Fura-2. Its dissociation constant for Na^+ in the presence of physiological concentrations of K^+ is ~11 mM, whereas its selectivity for Na^+ is approximately 18-fold higher than that for K^+. Na^+ probes with visible excitation currently available (i.e., Sodium Green and CoroNa Green) have provided mixed results depending on the preparation studied. A red fluorescence emitting Na^+-sensitive dye, CoroNa Red, has been shown to localize in mitochondria and therefore may be used to monitor Na^+ concentration changes occurring in the mitochondrial matrix.

Cl^- Chloride has traditionally been difficult to monitor by fluorescence microscopy. Available dyes include MEQ, SPQ, and MQAE, which have in common the fact that their fluorescence decreases by collisional quenching when the intracellular Cl^- concentration increases. In contrast to cation-sensitive dyes, Cl^- dyes are not available in a membrane-permeant form (i.e., acetoxymethyl esters). However, MEQ can be loaded in cell populations in a noninvasive manner when administered in its chemically reduced form dihydroMEQ, which readily crosses cell membranes. Once inside the cell, it undergoes oxidation into a polar form that remains trapped inside the cell. However, because the reduced form is extremely labile, MEQ is not commercially available in this membrane-permeant form so that it is necessary to perform the chemical reduction reaction immediately before loading.

Genetically Engineered Sensors

One of the elements that considerably extended the range of applications of live-cell imaging was the development of the green fluorescent protein (GFP) of the jellyfish *Aequora Victoria*. This fluorescent label can be incorporated into proteins by genetic fusion. Since the original studies describing the use of GFP, several improved variants have been produced that have higher fluorescence quantum yield or improved stability. A powerful feature of these reporter proteins is the ability to target their expression to given cellular compartments such as mitochondria or endoplasmic reticulum, as well as to given cell types in living organisms. In particular, constructs bearing the glial fibrillary acidic protein promoter have been produced to specifically direct the expression of the fluorescent protein constructs to astrocytes in the brain and, as an extension, transgenic animals with this type of mutation have been produced. In parallel with the development of these reporters used as morphological markers, fluorescent protein constructs have also been modified to reflect the chemical environment. For instance, GFP mutants have been produced to report pH changes within acidic compartments such as secretory vesicles.

Another strategy has been employed to build biosensors, namely to rely on sensing elements (e.g., protein sequences) that are the result of the evolution of living organisms which use them to selectively detect changes in the environment. These sensing elements can be coupled to fluorescent proteins by genetic engineering to produce an optical biosensor. The optical detection relies on the fact that two fluorescent molecules with overlapping fluorescence spectra may transfer the absorbed energy from one to the other in a radiationless process called fluorescence resonance energy transfer (FRET). This occurs when the two molecules are situated within a distance of 1–10 nm. Based on this general scheme, an increasing number of optical sensors have been developed to detect ions (Ca^{2+}, H^+, and Cl^-), membrane potential, as well as compounds such as inositol-1,4,5-trisphosphate, cyclic nucleotides, or sugars. For example, the first generation of genetically encoded Ca^{2+} indicators made use of conformational changes in calmodulin induced by Ca^{2+} binding in the presence of a tethered calmodulin-binding peptide M13. The read-out was achieved by detecting changes in the FRET efficiency between two spectral variants of fluorescent proteins, namely cyan and yellow fluorescent proteins, linked to the calmodulin and the M13 part of the construct, respectively.

This approach for detecting ion changes has many advantages, but it suffers from several drawbacks in comparison with classical organic probes. For instance, the dynamic range and amplitude of the response is generally much smaller in FRET Ca^{2+} sensors than with dyes, and transfecting cells can be more invasive than dye loading using membrane-permeant indicators. In addition, protein sensors obviously have a much larger mass than conventional nonprotein indicators, and steric interactions can potentially perturb the normal physiology of the living cell. Nevertheless, the prospect of having transgenic animals expressing such biosensors in given cell types of the brain will push the development of these approaches.

Fluorescence Imaging and Detection Methods

Fluorescence microscopy approaches are generally classified as either wide-field (or conventional) imaging or confocal imaging. In both cases, however, the basic design is based on the epifluorescence principle, in which the strong excitation light is separated from the much weaker emitted fluorescence light by means of a dichroic mirror placed at 45° into the optical path of the microscope. Current research microscopes make extensive use of photonics technologies and routinely use lasers, acousto-optical devices for wavelength selection, and are equipped with high-performance interference filters.

Wide-field imaging In wide-field epifluorescence, the entire field of view is illuminated by the excitation light and fluorescence is visually observed through the oculars or detected using a camera. The main advantages of this approach are that it allows fast detection and requires lower levels of light, which are both crucial parameters for fluorescence imaging of living specimens. Until recently, low-light imaging

devices used for fluorescence microscopy belonged to the class of extremely sensitive and fast intensified cameras, with limited resolution, or to the digital charged-coupled device (CCD) cameras, with excellent resolution but lower sensitivity and reduced speed of acquisition at low light levels. Although digital CCD cameras have recently dominated the field of fluorescence imaging, the so-called electron multiplying CCD, the sensitivity and speed of which are similar to those of the intensified cameras, is becoming the camera of choice.

Confocal imaging In confocal microscopy, the excitation light originating from a laser beam is focused by the objective lens onto a diffraction-limited spot in the specimen plane, with the consequence that fluorescence is excited only in this specific point of the specimen (Figure 1). To obtain an image of the specimen, the whole field has to be exposed to the excitation light generally using a scanning device. Several variations of this basic principle are currently implemented in confocal microscopes, using either a single point scanner or multiple point parallel illumination (e.g., using Nipkow disks, line and slit scanners, array scanners). The second main characteristic of confocal imaging is that, to be detected, the fluorescent light has to pass through a so-called pinhole (i.e., a small aperture diaphragm) placed in the conjugate plane before the detector. Because light emanating from above or below the focal plane in the specimen will be rejected by the pinhole, this arrangement enables one to obtain sharp optical sections of the specimen that otherwise appear blurred in wide-field fluorescence imaging. In thick specimens, such as brain slice preparations, confocal microscopes enable one to obtain sharp images of fluorescent cells or cells loaded with indicators relatively deep into the tissue. This major advantage of confocal microscopy has to be balanced with the disadvantage of being slower and requiring much more intense illumination than its wide-field counterpart. It is beyond the scope of this article to describe the wide array of technological approaches used by confocal microscopes available today.

Advanced imaging and optical approaches In conjunction with or as extensions of wide-field or confocal imaging, several specialized optical approaches have become invaluable tools for neuroscience research.

Two-photon microscopy is a modification of confocal microscopy that considerably improves imaging in thick living tissue, such as brain slices or *in vivo* brains. The basic principle is that instead of exciting fluorophores at the maximum of absorption (e.g., 480 nm for GFP), a wavelength approximately twice this wavelength is used (e.g., ~930 nm for GFP), which corresponds to infrared light. Photons of this wavelength having approximately half the required energy would not be able to excite the fluorophore under normal conditions. However, when two such photons are delivered simultaneously (i.e., within 10^{-15} s) on the fluorophore, they can interact with it in a nonlinear manner to excite electrons of the molecule to an excited state molecular orbital. Relaxation from the excited state to the ground state occurs with the emission of a photon, as for conventional fluorescence. An extremely high density of photons needs to be produced by a pulsed infrared light source, and when focused by the microscope objective lens, multiphoton absorption can occur only at the focal point where light is both spatially and temporally compressed to have two infrared photons reaching the same molecule simultaneously. There are two major advantages to this method: Fluorescence will be emitted only from the focal point, eliminating the problem of out-of-focus background signal (Figure 2), and infrared light allows better penetration compared to

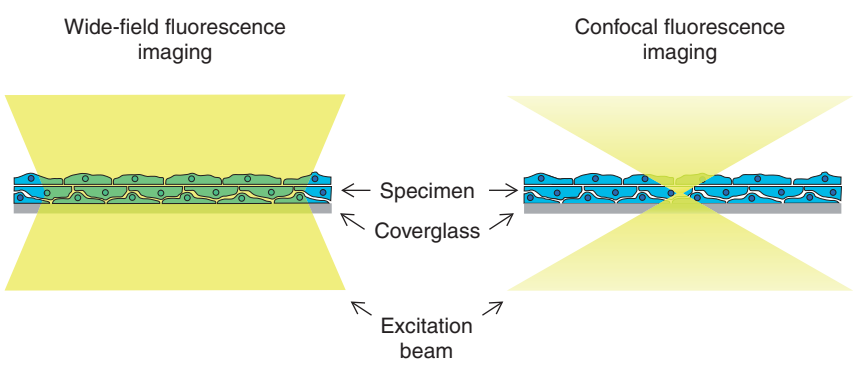

Figure 1 Comparison of fluorescence illumination in wide-field versus confocal microscopy. Whereas in wide-field fluorescence microscopes, the whole field of view is homogenously illuminated, point scanning confocal microscopes illuminate only a diffraction-limited spot focused by the objective lens. In this configuration, the intensity of excitation is maximal at the focal plane.

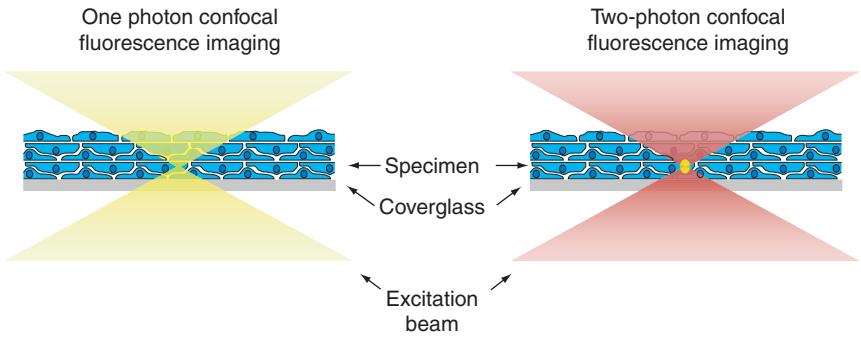

Figure 2 Comparison of fluorescence illumination in one-photon versus two-photon confocal microscopy. In two-photon microscopy, fluorescence excitation only occurs in a small focal volume where the light density is sufficient for simultaneous absorption of two infrared photons. No fluorescence is excited above or below this focal point, considerably reducing the background signal and allowing direct optical sectioning. Two-photon imaging is particularly useful for imaging in scattering media and thick specimens, such as a brain slice preparation or the whole living brain.

visible or ultraviolet light. Two-photon microscopy therefore enables imaging fluorescent reporter molecules or ion-sensitive probes deeper into the tissue. Although this technology has recently become somewhat easier to implement, the high cost of the laser source (generally a femtosecond Ti:Sapphire laser) remain a significant hurdle for its widespread use. Moreover, several fluorophores suffer from poor two-photon absorption capability.

Flash photolysis allows the fast and spatially defined application of bioactive molecules from photoactivatable inert precursors and is therefore a powerful tool for the dynamic study of the molecular mechanisms underlying physiological processes at the cellular level. The photoactivatable molecules, called caged compounds, are composed of active molecules that are bound to a photoabsorbing group resulting in a photolabile, biologically inert molecule. Upon brief and intense ultraviolet illumination, the photolabile caged compound releases the free, biologically active molecule along with the free caging group (Figure 3). It can be combined with electrophysiological techniques (e.g., patch clamp) or cellular imaging techniques (e.g., fluorescence microscopy) to monitor cellular responses to photoactivation of the caged molecules. Several classes of compounds, such as neurotransmitters, nucleotides, Ca^{2+} and Ca^{2+} chelators, fluorescent dyes, or second messengers, are commercially available as caged compounds and can be used for a wide variety of biological applications in neuroscience research.

Total internal reflection microscopy uses the property that a light beam reaching the interface between two media with different refractive indexes with an incident angle exceeding a critical value is totally reflected. Only a small portion of the radiation, called the evanescent wave, propagates into the distal medium. The evanescent wave will continue to travel into the medium of higher refractive index, but its strength will decay exponentially. Light escaping in this restricted region of the specimen immediately adjacent to the glass–water interface has a thickness usually less than 200 nm and can serve to selectively illuminate and excite fluorophores. This technique is particularly useful with cultured cells to visualize processes such as vesicle exocytosis and trafficking occurring in the close vicinity of the cell membrane facing the coverslip. For instance, it has been used to demonstrate that astrocytes possess a vesicular pool involved in glutamate release.

Glial Transporters and Channels

Glial cells express a large array of transporters and channels at their membrane that allow them to play a central role in regulating the extracellular milieu in the brain. The main primary element energizing these transmembrane ion fluxes comes from the Na^+/K^+-ATPase, which is abundantly expressed by glial cells. This pump utilizes the energy liberated by the hydrolysis of ATP into ADP to export three Na^+ ions and takes up two K^+ ions per cycle, thereby enabling the buildup of Na^+ and K^+ gradients in opposite directions. Because glial cells are not excitable, the Na^+ gradient has the main purpose of serving as the driving force of secondary active transport systems (i.e., cotransporters and antiporters). It is estimated that Na^+/K^+-ATPase accounts for approximately half of the total cellular ATP consumption.

Transporters

Glial cells, and astrocytes in particular, express several Na^+-coupled transport systems, such as those involved in pH regulation (Na^+/H^+ exchanger and

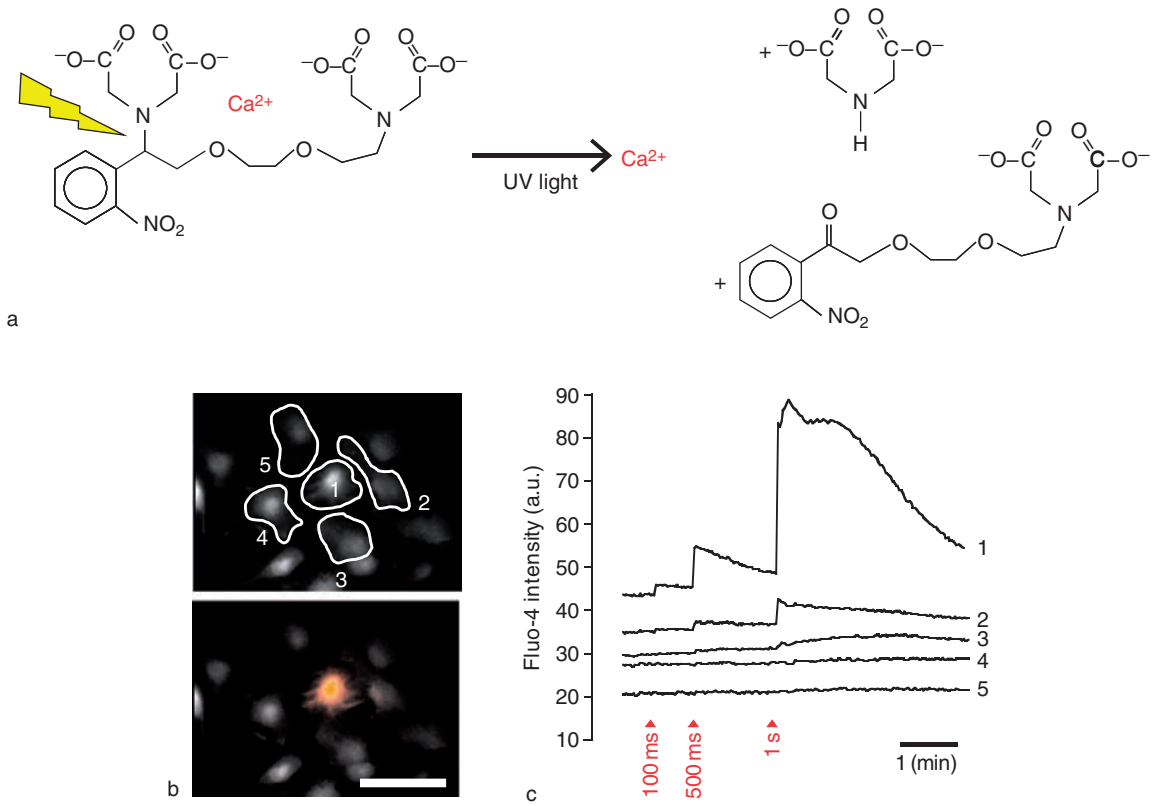

Figure 3 Intracellular Ca^{2+} photo release by ultraviolet flash photolysis. (a) Upon intense illumination by ultraviolet light, the caged Ca^{2+} molecule, nitrophenyl EGTA, with a bound Ca^{2+} ion undergoes a photochemical reaction leading to the breakdown of the molecule into two parts, liberating the free Ca^{2+} ion. (b) Ca^{2+} uncaging performed in astrocytes loaded with the caged Ca^{2+} compound NP-EGTA AM and the Ca^{2+} probe Fluo-4 AM (top). The UV beam, produced by a powerful light-emitting diode, was aimed at cell 1 in the center. The false color overlay in the bottom image shows that only the targeted cell 1 exhibited a significant elevation of cytosolic Ca^{2+} shown as Fluo-4 fluorescence intensity. (c) Fluo-4 fluorescence intensities plotted against time, showing that repetitive Ca_i^{2+} responses can be triggered by NP-EGTA uncaging in the targeted cell. Scale bar = 100 μm. (b, c) Adapted from Bernardinelli Y, Haeberli C, and Chatton JY (2005) Flash photolysis using a light emitting diode: An efficient, compact, and affordable solution. *Cell Calcium* 37: 565–572, with permission from Elsevier.

Na^+/HCO_3^- cotransporter), Ca^{2+} homeostasis (Na^+/Ca^{2+} exchanger), neurotransmitter uptake (glutamate, GABA, serotonin, and glycine), and amino acid uptake, as well as the $Na^+/K^+/2Cl^-$ cotransporter. They also express anion exchangers (Cl^-/HCO_3^- and Cl^-/OH^-) with a role in pH regulation.

Ion channels

Several ion conductances have been described in glial cells. The most studied are K^+ channels, which represent the most abundant conductance detected in glia and dominate all other ion conductances. They have been assigned the main role of buffering increases in interstitial K^+ concentration resulting from neuronal excitation. In addition, voltage-gated Ca^{2+} channels or voltage-gated Na^+ channels (TTX sensitive) have been reported in several studies, although it is unclear what role this latter channel could play in nonexcitable glial cells. Finally, anion channels are also expressed in glial cells and probably involved in volume regulation.

Ionotropic Receptors

Initially thought to be a genuinely neuronal attribute, ionotropic neurotransmitter receptors have been found to be functionally expressed in glia, such as non-NMDA glutamate receptors (AMPA and kainate receptors) and acetylcholine nicotinic receptors. NMDA receptors appear to be primarily associated with oligodendrocytes rather than astrocytes. No clear consensus exists on the purpose of their presence in glial cells. $GABA_A$ receptors as well as nicotinic acetylcholine receptor currents have also been observed.

Gap Junctions and Connexins

Connexins are proteins that form hexameric structures in the plasma membrane, called hemichannels, which harbor a central pore. When hemichannels of two adjacent cells dock to each other, they form a gap junction channel that enables the intercellular

diffusion of ions and molecules of up to 1 kDa. It is generally assumed that gap junctions contribute to long-range intercellular signaling as well as tissue homeostasis, including the transport of nutrients across the glial network and the spatial K^+ buffering. The gap junction channels can be gated by several stimuli, such as low pH and various pharmacological agents, and hemichannels not apposed to another cell are normally closed. However, under certain conditions, such as low extracellular Ca^{2+} levels, they can open and mediate the release of molecules such as ATP or glutamate from the cell as well as equilibrate ion gradients. Whether such hemichannel openings have significance in nonpathological conditions is currently debated.

Investigating Ion Homeostasis in Relation to Cellular Functions

The availability of fluorescent indicators coupled to sensitive detection devices has allowed the exploration of detailed mechanisms underlying the regulation of H^+, Na^+, Ca^{2+}, and, to a lesser extent, Mg^{2+}. In several cases, a link could be drawn to their role in certain functions at the level of the cell or the tissue. A few examples are discussed next.

Ca^{2+} Homeostasis and Signaling

Researchers began to recognize the decisive advantage of fluorescence imaging technology with the development of the ability to monitor intracellular Ca^{2+} with spatial resolution using indicators that reliably report changes in free Ca^{2+} concentration in the nanomolar to submillimolar range. Imaging techniques have demonstrated that glial cells, although not electrically excitable, show a complex repertoire of Ca^{2+} responses to various electrical, mechanical, and chemical stimuli. For example, astrocytes are able to generate Ca^{2+} signals by Ca^{2+} entry via voltage-gated Ca^{2+} channels, via AMPA receptor activation, and by IP_3-induced Ca^{2+} release following metabotropic receptor stimulation. Ca^{2+} transients in astrocytes have been shown to trigger the release of various transmitters such as glutamate or ATP, which in turn can drive neuronal excitation. In addition, astrocytes are also involved in long-range signaling because they can initiate and propagate the intercellular spreading of Ca^{2+} signals, known as Ca^{2+} waves (Figure 4). Until recently, this type of communication was thought to be an exclusive property of neurons. This form of excitability, also called calcium excitability, required optical approaches to be uncovered.

pH Regulation

Glial cells play a prominent role in overall H^+ homeostasis in the nervous system. Fluorescent indicators are a considerable improvement over ion-sensitive microelectrodes since they are not as invasive as cell impalements, and they provide spatial information on intracellular pH distribution. Studies using fluorescent indicators have provided crucial details on cytosolic H^+ buffering, transmembrane movements of acid–base equivalents, metabolic production of H^+, or sequestration of H^+ into intracellular organelles. Intracellular pH is regulated passively by the buffering capacity of the cell and, most important, actively by mechanisms that glial cells share with other cells types, such as exchangers (Na^+/H^+ or Cl^-/HCO_3^-) or cotransporters (Na^+/HCO_3^-). Changes in glial pH following various physiological stimuli can last from seconds to minutes or longer. They may also occur in pathological situations such as ischemia or epilepsy.

Activation of neurotransmitter receptors and transporters expressed by glial cells can induce pH transients. For instance, stimulation of $GABA_A$ receptor-coupled chloride channels, which present significant permeability to bicarbonate and are expressed in astrocytes, leads to cellular alkalinization because the HCO_3^- gradient is outwardly directed. Conversely, glutamate causes intracellular pH to drop by activating AMPA and kainate receptors, as well as by activation of Na^+–glutamate transporters which cotransport with one proton (or exchange a hydroxyl ion). pH changes in glial cells are thought to be important regulatory signals for function and metabolism, such as the glutamate–glutamine conversion.

Na^+ Homeostasis and Energy Metabolism

Studying Na^+ homeostasis with high spatial and temporal resolution, which cannot be achieved using radioactively labeled tracers or ion-sensitive microelectrodes, can be performed using fluorescent indicators. The transmembrane Na^+ gradient generated by Na^+/K^+-ATPase plays a central role in providing the driving force for secondary transport systems, such as the Na^+/H^+ or Na^+/Ca^{2+} exchangers, which are important regulators of intracellular pH and Ca^{2+}, respectively. One of the most specific functions of glial cells is the recapture of extracellular glutamate by Na^+–glutamate cotransporters, thereby preventing its buildup to excitotoxic levels during neuronal activity. Because of the prominent activity and density of Na^+-dependent glutamate transporters in glia, glutamate transport is accompanied by a substantial Na^+ influx that causes a three- or fourfold increase in the intracellular Na^+ concentration

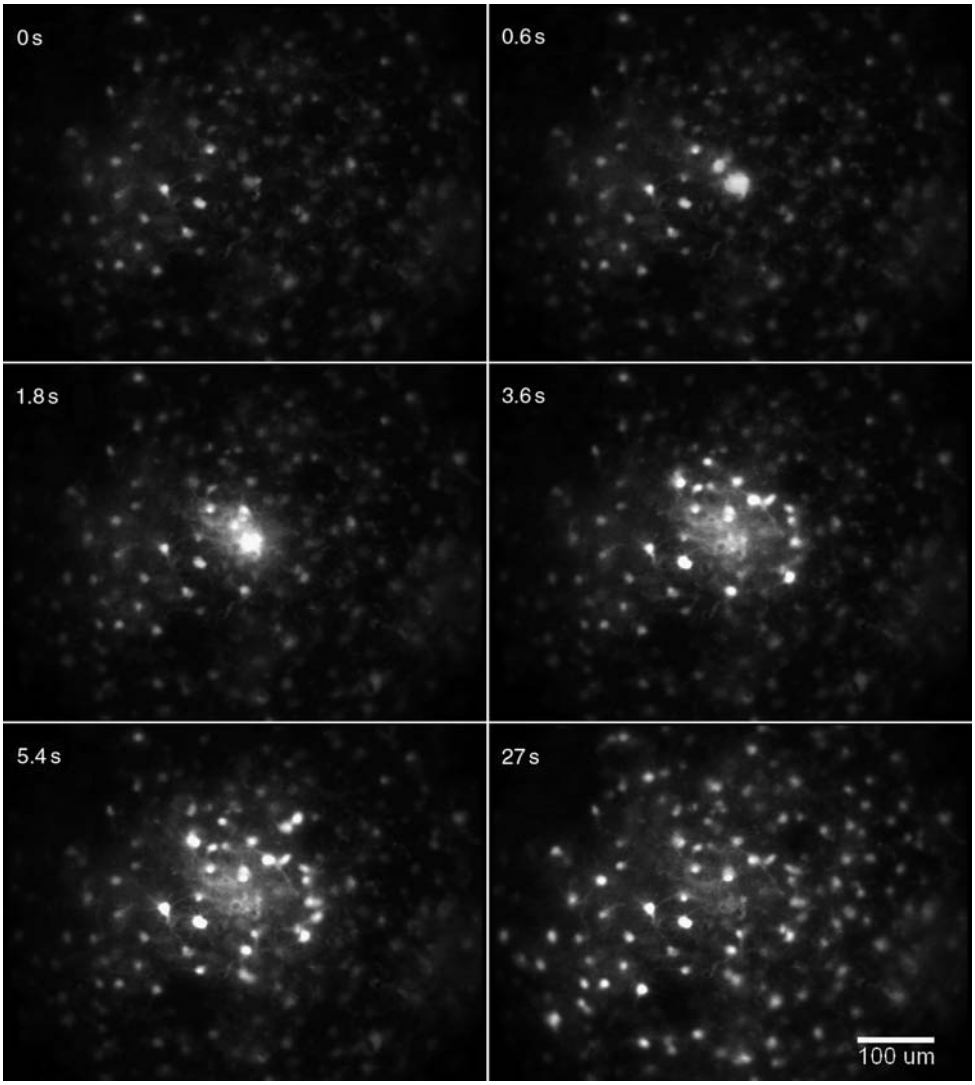

Figure 4 Intercellular Ca^{2+} wave in cultured astrocytes. Sequence of images of mouse astrocytes in primary culture loaded with the Ca^{2+} probe Fluo-4 AM. A Ca^{2+} response was triggered using a stimulation microelectrode placed 20 μm above the surface of one cell in the center of the field. The time course of the Ca^{2+} wave spreading from the center can be seen.

(Na_i^+), which is detectable with good accuracy using dyes, particularly SBFI. Such measurements are therefore an indirect means of detecting the spatial distribution of glutamate transport activity in real time. Studies have shown that Na_i^+ is the result of the balance between the rate of Na^+ influx and extrusion by Na^+/K^+-ATPase. Fluorescence imaging studies have demonstrated that glutamate transport activity causes a substantial energy burden to the cell by increasing the overall activity of Na^+/K^+-ATPase by a factor of two or three, increasing in turn its associated ATP consumption by a corresponding factor.

See also: Astrocyte: Identification Methods; Calcium Homeostasis in Glia; Calcium Waves in Glia; Fluorescence Microscopy in the Neurosciences; Gap Junction Communication; Ionic Channels in Glia; Neuron-Glia pH Regulation; Potassium Homeostasis in Glia; Two-Photon Imaging.

Further Reading

Bennett MV, Contreras JE, Bukauskas FF, and Saez JC (2003) New roles for astrocytes: Gap junction hemichannels have something to communicate. *Trends in Neurosciences* 26: 610–617.

Bernardinelli Y, Haeberli C, and Chatton JY (2005) Flash photolysis using a light emitting diode: An efficient, compact, and affordable solution. *Cell Calcium* 37: 565–572.

Brookes N (1997) Intracellular pH as a regulatory signal in astrocyte metabolism. *Glia* 21: 64–73.

Griesbeck O (2004) Fluorescent proteins as sensors for cellular functions. *Current Opinion in Neurobiology* 14: 636–641.

Inoué S and Spring KR (1997) *Video Microscopy: The Fundamentals*. New York: Plenum.

Knopfel T, Diez-Garcia J, and Akemann W (2006) Optical probing of neuronal circuit dynamics: Genetically encoded versus classical fluorescent sensors. *Trends in Neurosciences* 29: 160–166.

Pawley JB (ed.) (2006) *Handbook of Biological Confocal Microscopy.* New York: Plenum.

Verkhratsky A, Orkand RK, and Kettenmann H (1998) Glial calcium: Homeostasis and signaling function. *Physiological Reviews* 78: 99–141.

Verkhratsky A and Steinhauser C (2000) Ion channels in glial cells. *Brain Research Reviews* 32: 380–412.

Volterra A, Magistretti PJ, and Haydon PG (eds.) (2002) *The Tripartite Synapse*. Oxford: Oxford University Press.

Yuste R, Lanni F, and Konnerth A (eds.) (2000) *Imaging Neurons: A Laboratory Manual*, Cold Spring Harbor, NY: Cold Spring Harbor Press.

Glial Plasticity and Neuroendocrine Regulation

D T Theodosis, Inserm U862, University Victor Segalen, Bordeaux, France

© 2009 Elsevier Ltd. All rights reserved.

Astrocytic Processes Ensheathe Hypothalamic Neurons and Synapses

As in all brain areas, astrocytes are the most numerous glial cells in the hypothalamus. They have a variable, complex morphology, consisting of relatively small somata, several thick processes extending into the neuronal network (Figure 1), and numerous thin processes (Figure 2). At times, the latter consist merely of two membranes enclosing little cytoplasm and few organelles (Figures 3 and 4). These lamellipodia-like processes ensheathe neuronal somata, dendrites, and synapses, to a varying degree. Bulges at their tips can surround several synapses, thereby creating glial microdomains which permit neuron–glia interactions in a restricted microenvironment. In addition to supportive and nutritive functions, astrocytic processes, by their mere presence, act as a physical barrier to restrict spillover and diffusion of locally released neuroactive substances in the extracellular space (Figure 5).

These include ions (e.g., K^+), neurotransmitters (e.g., glutamate and γ-aminobutyric acid (GABA)), steroids, various neuropeptides (e.g., oxytocin and vasopressin), and amino acids (e.g., taurine and D-serine). On the other hand, since astrocytic processes carry many neurotransmitter receptors, transporters, and ion channels on their surfaces, they can intervene actively and rapidly in neuronal communication. For example, the numerous transporters on their surface (Figure 2) assure quick and efficient removal of glutamate released from synapses, thereby contributing to the regulation of presynaptic glutamate receptors on these and neighboring GABA terminals (Figure 6). This and similar observations have led to the concept of the synapse as a tripartite structure, of which astrocytes play a role as necessary as that of their pre- and postsynaptic neuronal components.

Astrocytes respond rapidly to neuronal activity by increases in free intracellular Ca^{2+}; Ca^{2+} increases, in turn, set off different intracellular signaling mechanisms and gene activation. Such activity-dependent Ca^{2+} signaling ultimately regulates glial structure and function. Ca^{2+} signaling can be bidirectional, since increased astrocytic Ca^{2+} in turn can enhance cytoplasmic levels of calcium in neurons. Astrocytes can thus potentially modulate neuronal activity. Moreover, they can release a variety of neuroactive molecules, termed gliotransmitters (Figure 7), including glutamate, D-serine, and adenosine triphosphate (ATP), that in turn affect synaptic function and neuronal excitability.

Astrocytic Ensheathement of Neurons Varies with Neuronal Activity

Observations from hypothalamic centers such as the supraoptic nucleus (SON), paraventricular nucleus (PVN), and arcuate and preoptic nuclei have revealed that astrocytes are not static cells, since they can undergo significant morphological transformations under different physiological conditions. This structural glial plasticity occurs together with remodeling of adjacent neurons and their afferent inputs and highlights the adult brain's remarkable capacity for restructuring to meet particular functional requirements.

The cytoskeletal protein, glial fibrillary acidic protein (GFAP), accumulates in astrocytic somata and thick processes, and its modified expression may indicate gross restructuring of the cells. Such a phenomenon appears to occur in the SON, where large glial processes, visualized by their immunoreactivity for GFAP (Figure 1), appear to change from a vertical to horizontal orientation during stress conditions (e.g., dehydration), which strongly activate SON neurons. In the arcuate nucleus, the neurons of which participate in the regulation of reproduction, there are significant changes in the expression of GFAP in relation to fluctuating systemic levels of the sex hormone, estrogen. Probably the most striking example of modified GFAP expression in the adult hypothalamus is offered by the suprachiasmatic nucleus, which regulates the diurnal clock and in which GFAP levels and distribution change dramatically in relation to circadian rhythmicity. Nevertheless, it must be kept in mind that since fine astrocytic processes are devoid of GFAP, changes in GFAP expression offer no clue as to their motility.

It is the rapid motility of the lamella-like distal processes of astrocytes that appears initiated by neuronal activity and which is the basis of glia–neuron crosstalk. Convincing evidence for this has been offered by morphological and electrophysiological observations of the SON and of the magnocellular portions of the PVN. These hypothalamic areas are composed of magnocellular neurons that secrete the neurohormones oxytocin and vasopressin and that show enhanced activity during conditions such as parturition, lactation, and osmotic regulation. In the SON,

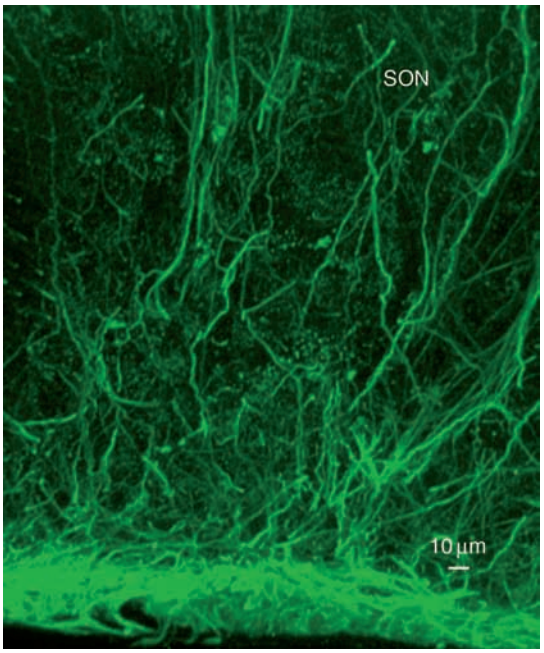

Figure 1 The majority of astrocytes in the rat supraoptic nucleus (SON) of the hypothalamus have a radial morphology, with somata aligned along the base of the brain, giving rise to long, thick processes traversing the nucleus. Confocal microscopy of a coronal section after immunofluorescence labeling for the cytoskeletal protein, glial fibrillary acidic protein. Adapted from Bonfanti L, Poulain DA, and Theodosis DT (1993) Radial glia-like cells in the supraoptic nucleus of the adult rat. *Journal of Neuroendocrinology* 5: 1–5, with permission.

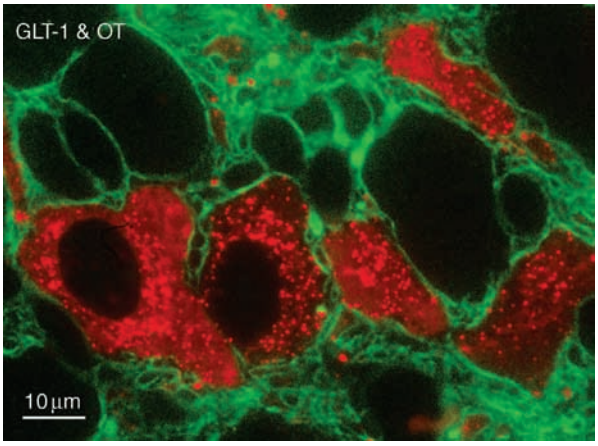

Figure 2 Fine astrocytic processes (green), rich in glutamate transporters (GLT-1), surround oxytocinergic (red) and putative vasopressinergic (black) neuronal somata and dendrites in the supraoptic nucleus. Confocal microscopy after double immunofluorescence labeling for GLT-1 and oxytocin (OT). Adapted from Oliet SHR, Piet R, Poulain DA, et al. (2004) Glial modulation of synaptic transmission: Insights from the hypothalamic supraoptic nucleus. *Glia* 47: 258–267, with permission.

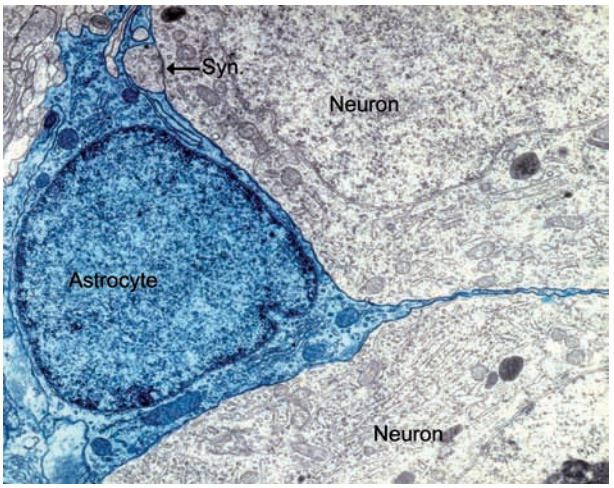

Figure 3 Electron micrograph showing an astrocyte and two magnocellular neurons in the supraoptic nucleus. Note that the astrocyte gives rise to a fine, lamellopodial-like process that separates the neuronal somata. Fine astrocytic processes also ensheathe a synapse (syn) contacting (arrow) one of the neurons. Adapted from Theodosis DT and Poulain DA (1993) Activity-dependent neuronal–glial and synaptic plasticity in the adult mammalian hypothalamus. *Neuroscience* 57: 501–535, with permission from Elsevier.

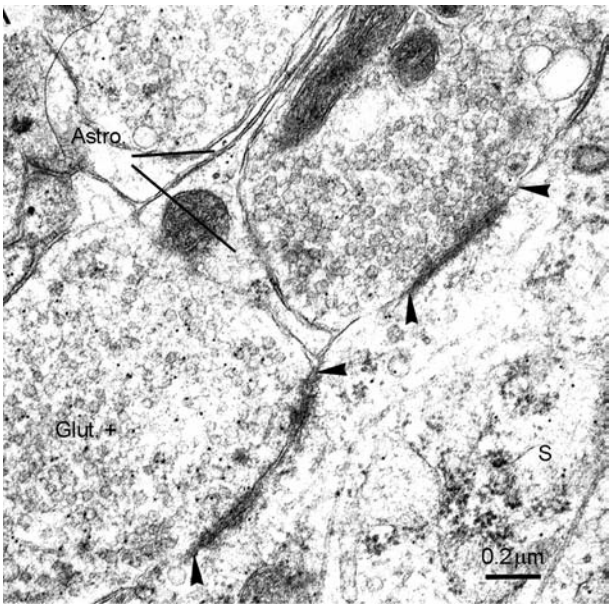

Figure 4 Electron micrograph of glutamatergic (glut+) terminals in the supraoptic nucleus after ultrastructural immunogold labeling for glutamate. The terminals make synaptic contact (at arrowheads) on a neuronal soma (S). They are ensheathed by fine astrocytic (astro) processes. Adapted from Oliet SHR, Piet R, Poulain DA, et al. (2004) Glial modulation of synaptic transmission: Insights from the hypothalamic supraoptic nucleus. *Glia* 47: 258–267, with permission.

glial remodeling is detectable within a few hours of the onset of parturition or within 1 h of exposure to oxytocin, and results in a significant reduction of astrocytic coverage of oxytocin neurons and their synapses. Thus, under normal conditions, about 90% of any SON neuronal soma is surrounded by astrocytic processes, but at lactation, this is reduced to 70% for oxytocin somata, the surfaces of which become directly

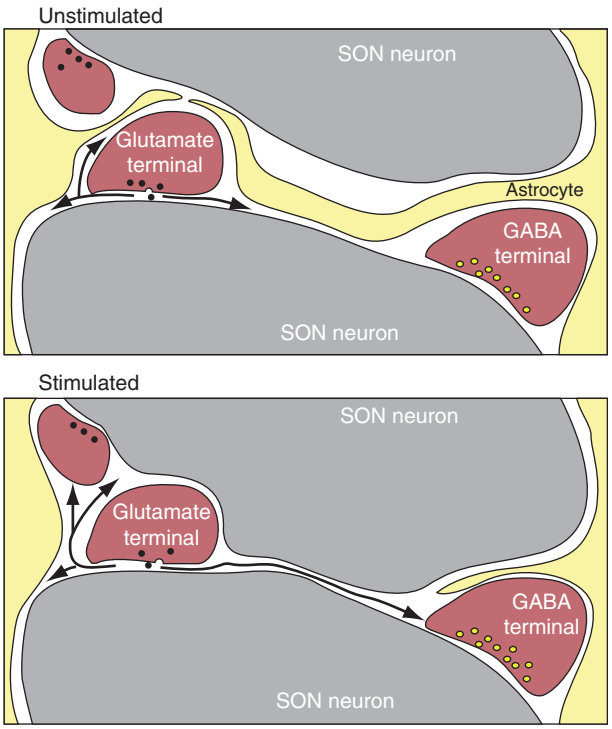

Figure 5 Schema depicting the presence of fine astrocytic processes between supraoptic nucleus (SON) neurons and around axonal terminals in unstimulated (e.g., virgin and postweaned rats) and stimulated (e.g., lactating and dehydrated rats) conditions of neurosecretion. Under stimulation, retraction of astrocytic processes permits diffusion (black arrows) of synaptically released glutamate into the extracellular space; glutamate can then act on the same and/or adjacent synapses (homo- and heterosynaptic interactions; GABA, γ-aminobutyric acid). Adapted from Theodosis DT, Piet R, Poulain DA, and Oliet SH (2004) Neuronal, glial and synaptic remodeling in the adult hypothalamus: Functional consequences and role of cell adhesion and extracellular matrix adhesion molecules. *Neurochemistry International* 45: 491–501, with permission.

juxtaposed (Figure 5). Concomitantly, glial coverage of their dendrites is diminished, which gives rise to the appearance of numerous bundles of directly juxtaposed dendrites. There is also an important reduction in astrocytic ensheathement of synapses impinging on their somata and dendrites. The axons of these neurons project to the neurohypophysis, where significant glial rearrangements take place concurrently. During lactation, for example, there is a reduced ensheathement of neurosecretory axons, retraction of glial processes from the basal lamina around fenestrated capillaries, multiplication of neurosecretory terminals, and, ultimately, an increased neurovascular contact zone. A similar remodeling occurs in the hypothalamic magnocellular nuclei and neurohypophysis in response to persistent osmotic stimulation and, interestingly, this affects essentially the oxytocin system. These morphological changes represent physiological plasticity, since glial coverage is reestablished rapidly upon arrest of activation, as after weaning the young, rehydration, or reduction of ambient oxytocin levels.

Rapid, activity-dependent structural plasticity occurs in the basal hypothalamus as well. In the rodent arcuate nucleus, which, as noted earlier, is composed of neurons that participate in reproductive and metabolic functions, the proportion of neuronal somatic surface covered by astrocytic processes fluctuates with changing sexual steroid levels, being high at estrus when plasma estrogen is elevated, and low 24 h later at metestrus when estrogen levels have diminished. Administration of estrogen to ovariectomized animals mimics the effects of the estrus cycle, and it is possible to detect a modified glial coverage of neurons within 2 h of steroid injection. The infundibulum and preoptic

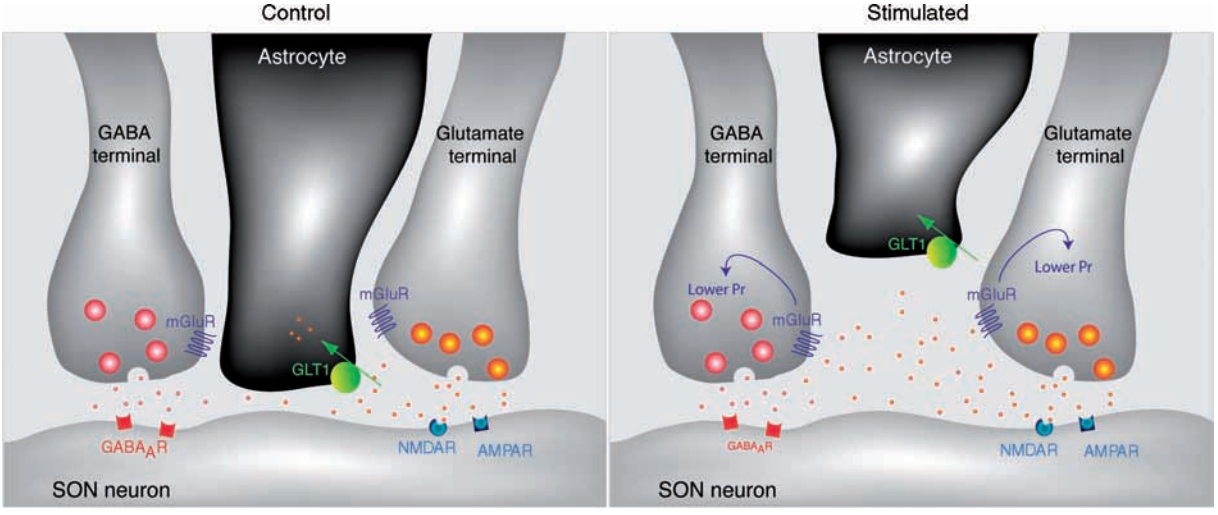

Figure 6 Schematic illustration of intersynaptic cross-talk in the supraoptic nucleus (SON). Under unstimulated (control) conditions of neurosecretion, glutamate transporters (GLT-1) on astrocytic processes ensure clearance of synaptically released glutamate. Under stimulated conditions (lactation, chronic dehydration), astrocytic withdrawal results in accumulation of glutamate in the extracellular space. The amino acid then reaches and activates presynaptic metabotropic receptors (mGluR) on glutamate terminals, thereby diminishing release probability (Pr) of glutamate. In addition, lack of astrocytic processes promotes spillover of glutamate, which activates mGluR on adjacent γ-aminobutyric acid (GABA)ergic terminals; NMDAR, *N*-methyl-D-aspartate receptor; AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor. Diagram courtesy of SHR Oliet.

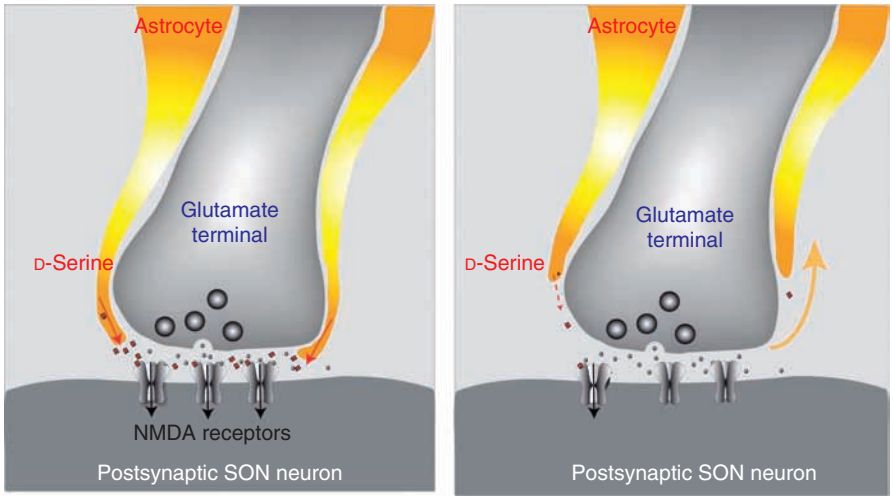

Figure 7 Schema illustrating gliotransmission in the supraoptic nucleus (SON). D-Serine (red squares) is released from astrocytes and, together with synaptically released glutamate (gray dots), activates *N*-methyl-D-aspartate (NMDA) receptors on (SON) neurons. Note that after retraction of astrocytic processes, there is less D-serine for NMDA receptor activation. Diagram courtesy of A Panatier and SHR Oliet.

areas in primates are analogous to the rodent arcuate and in these structures, glial ensheathement of gonadotropin-releasing hormone (GnRH) neurons controlling gonadotropin secretion in the pituitary varies significantly in relation to fluctuating levels of circulating sexual steroids. The axons of basal forebrain neurons project to the external layer of the median eminence where changing sex steroid levels give rise to dynamic axon–glial changes similar to those characterizing the neurohypophysis during different conditions of neurosecretion. Thus, in the median eminence, when plasma GnRH levels are elevated, there is retraction of glial end-feet from the perivascular zone, and this exposes GnRH terminals to fenestrated portal capillaries. Both in the neurohypophysis and median eminence, therefore, glial rearrangements ultimately permit direct access of neurosecretory terminals to perivascular areas, an anatomical configuration that will favor release of neurosecretory products into the circulation.

Consequences of Glial Plasticity

On Extracellular Homeostasis

Astrocytes remove ions from the extracellular milieu, thereby contributing to fluid homeostasis. A lack of intervening astrocytic processes during enhanced neuronal activity can result, therefore, in the accumulation of ions such as K^+, which can facilitate further neuronal excitability. Such increases have been recorded in the SON and neurohypophysis coincident with activation of oxytocin neuron firing during lactation.

On the other hand, the extracellular space is a diffusion channel for molecules implicated in extrasynaptic or volume transmission. If we consider astrocytic processes as a physical barrier in this channel, then, in their absence, diffusion of neurotransmitters (e.g., glutamate, GABA, and catecholamines) will be greatly facilitated (Figures 5 and 6). This then promotes actions distant from their release sites and allows cross-talk between neighboring synapses and neurons (Figure 6). Likewise, locally released neuropeptides (e.g., oxytocin and vasopressin) may diffuse more freely in the absence of intervening glial processes, and this will potentiate their autocrine and paracrine actions.

The extracellular matrix is composed of glycoproteins – cell adhesion molecules, proteoglycans, and tenascins – that are secreted by astrocytes. Some of these molecules are considered permissive factors for glial and neuronal remodeling. However, they can intervene in neuron–glia interactions more directly as well. They are large, complex molecules, potentially hindering access to receptors and transporters. In addition, they can interact heterophilically with other extracellular matrix components to affect neuronal activity directly. The expression of these matrix molecules is often reduced or absent when there are no intervening astrocytic processes between neuronal elements, and this will affect molecular interactions in the external milieu and, therefore, intercellular communication.

On Glutamate and GABA Neurotransmission

Astrocytic processes in the vicinity of glutamatergic synapses limit glutamate spillover and are therefore critical for ensuring point-to-point synaptic transmission. Their absence will enhance levels of ambient glutamate, since the amino acid escapes

reuptake (Figure 6). Glutamate then reaches and stimulates inhibitory metabotropic receptors at presynaptic sites of glutamate and neighboring GABA synapses, leading to homo- and heterosynaptic depression of neurotransmitter release, respectively. The resulting reduction in synaptic efficacy will also diminish background synaptic noise, which renders the neuron more electrically compact, increasing input resistance and the gain of its input–output response. Such changes can be critical for tuning neurons to respond to pertinent synaptic information. Overall, these effects will reduce tonic excitatory input to postsynaptic neurons. On the other hand, the attenuation of inhibitory GABA neurotransmission will ultimately increase their excitability.

On Gliotransmission

The amino acids glutamate and D-serine are produced by astrocytes and have been identified as gliotransmitters mediating astrocyte–neuron cross-talk. Since they are co-agonists in the activation of *N*-methyl-D-aspartate (NMDA) receptors (Figures 6 and 7), the degree of astrocytic coverage of neurons will govern the availability of these receptors for activation. In addition to affecting rapid transmission at glutamatergic synapses, there will be effects on long-term synaptic changes, the direction – long-term potentiation (LTP) or long-term depression (LTD) – and magnitude of which depend on the number of activated NMDA receptors. These phenomena are clearly illustrated by the glial changes in the SON. Since SON astrocytes release D-serine, the levels of the amino acid and its occupancy of NMDA receptors are closely related to the degree of astrocytic coverage. In this nucleus, then, NMDA-dependent phenomena such as LTP and LTD are modified by glial remodeling. By simply getting close to or away from the synaptic cleft, astrocytic processes change the direction of these long-term functional changes. Thus, at the time of glutamate activation, their proximity will make D-serine available, thereby favoring LTP, while in their absence, D-serine levels are reduced, which will favor LTD.

Another identified gliotransmitter is the purine ATP. In the PVN, noradrenalin increases the amplitude of postsynaptic excitatory currents via increased insertion of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors due to a concomitant release of ATP from adjacent astrocytes. In the absence of astrocytic processes, there will be reduction in ATP and, in consequence, reduction of postsynaptic efficacy at these glutamatergic synapses.

On Structural Synaptic Plasticity

Because of their contiguous positioning with respect to neuronal elements and because of their capacity for rapid morphological changes, astrocytes can intervene in synaptic remodeling. Observations from several neuroendocrine systems in which circuits remain plastic throughout life support such a contention. Here also, the oxytocin (OT) system is exemplary, since its astrocytic transformations are invariably accompanied by changes in afferent inputs controlling its activity. Coincident with reduced astrocytic coverage of its somata and dendrites, an increased density of inhibitory (GABA) and excitatory (glutamate, noradrenalin) synapses is always found; synaptic numbers return to unstimulated levels on reestablishment of glial coverage. Both synapse formation and elimination occur rapidly, in a matter of hours.

A similar rapid synaptic remodeling, closely correlated to astrocytic ensheathement of postsynaptic elements, occurs in the rodent arcuate nucleus, where the number of inhibitory axosomatic synapses diminishes at proestrus and is reestablished at estrus, effects that can be mimicked by gonadectomy or estrogen replacement. Likewise, the number of inhibitory synapses on GnRH neurons is modified in relation to puberty in primates and seasonal breeding in ewes. It is noteworthy that synaptic rewiring has been described in the basal hypothalamus of rodents under varying conditions of energy expenditure and food intake, and it is not unlikely that it is accompanied by astrocytic transformations like those in the arcuate nucleus and GnRH system during different reproductive states.

How astrocytic processes actually participate in structural synaptic plasticity remains unknown. Nevertheless, the mere interposition of even the thinnest astrocytic lamella between pre- and postsynaptic partners can prevent synapse formation; their retraction is necessary to free neuronal surfaces on which new synapses can be formed. On the other hand, the reinsertion of astrocytic processes at the end of a period of neuronal activation can facilitate detachment of synaptic terminals from postsynaptic sites (Figure 8), and this will contribute to reestablish synaptic numbers to prestimulation levels.

Glial Participation in Neuroendocrine Regulations

Parturition and Lactation

These key events in mammalian reproduction are under the control of oxytocin acting on the uterus and mammary gland, respectively. The pulsatile release of the neurohormone, before the birth of each offspring and before each reflex milk ejection during lactation, is due to an intermittent, high-frequency firing of oxytocin neurons, synchronized throughout the population in the hypothalamus. At first sight, it

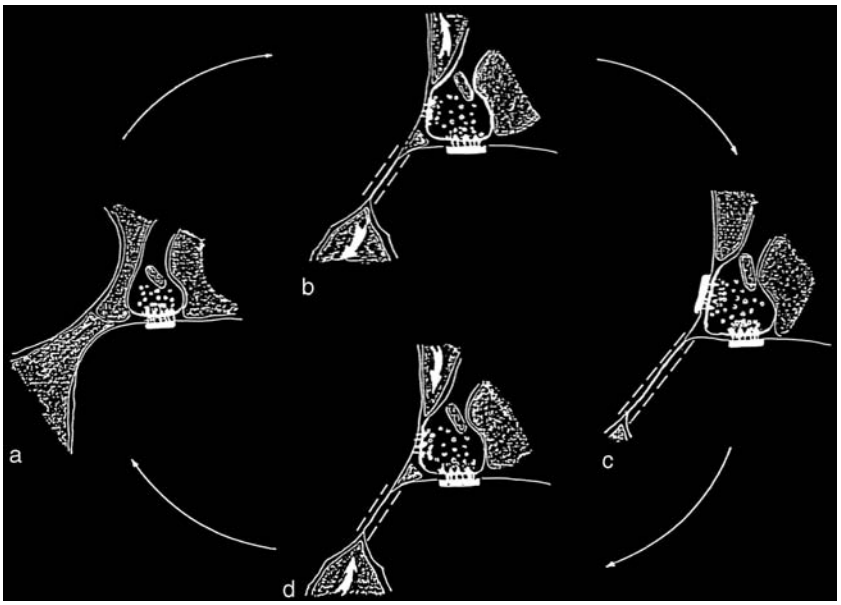

Figure 8 Schema showing possible cell mechanisms that may underlie astrocytic participation in synaptic remodeling in the adult supraoptic nucleus. (a) Synapses may result from hyperactivity in certain terminals that then induce postsynaptic density formation in adjacent neuronal elements. (b) Withdrawal of fine astrocytic processes (arrow) will facilitate this process by freeing neuronal surfaces for new synapse formation. (c, d) Upon diminution of synaptic activity, astrocytic processes reinsert themselves and can thus participate in the undoing of synaptic contacts. Adapted from Theodosis DT and Poulain DA (1993) Activity-dependent neuronal–glial and synaptic plasticity in the adult mammalian hypothalamus. *Neuroscience* 57: 501–535, with permission from Elsevier.

may appear paradoxical that, concurrent to this neuronal excitation, there is a reduction in astrocytic ensheathement of glutamatergic synapses which diminishes the efficacy of excitatory neurotransmission by activating their presynaptic metabotropic receptors (Figure 6). Nevertheless, the action of glutamate on its own release may depend on the level of presynaptic activity, so that during high-frequency firing, an increase in transmitter release probability due to calcium accumulation in the terminals overcomes metabotropic receptor-mediated inhibition. Thus, information transmitted via high-frequency firing would be less affected by inhibitory feedback in contrast to information from inputs displaying low-frequency activity. Such a high-pass filter would serve to increase the signal-to-noise ratio for information transmitted by inputs displaying high-frequency activity, as is the case of glutamatergic inputs driving oxytocin neurons when they are activated. To these effects of glutamate must be added the consequences of intersynaptic cross-talk, which is also facilitated by a reduced astrocytic coverage. As described previously (Figure 6), during intense glutamatergic activity, the amino acid activates metabotropic receptors on neighboring GABA terminals, and this heterosynaptic inhibition of GABA release will disinhibit oxytocin neurons and further facilitate their bursting firing. Changes in gliotransmission will be of consequence to this activity as well. In the absence of astrocytic processes, there is less D-serine in the vicinity of glutamatergic synapses, and this will diminish NMDA receptor activation (Figure 7). Since activation of these receptors is essential for long-term synaptic changes, reduction in glial coverage shifts their activity dependence toward higher activity levels, thereby favoring potentiation of glutamatergic synapses displaying high activity.

Removal of astrocytic processes from between oxytocin neurons will promote the diffusion of other molecules, such as oxytocin produced by oxytocin neurons, which facilitates their bursting and synchronized activity. It is obvious that in the absence of intervening astrocytic processes, the range of action of the peptide will be enhanced and this will result in the recruitment and synchronization of additional bursting neurons.

Finally, there will be facilitation of oxytocin neuron firing because of the increased number of synapses associated with reduced glial coverage. It is obvious that an increase in glutamatergic inputs will potentiate such firing, to which can be added that provided by additional excitatory noradrenergic synapses. As for inhibitory GABAergic inputs, at first glance it might appear incongruous that an increase in their numbers could facilitate enhanced neuronal firing. Nevertheless, the increase in inhibitory input may serve to prevent oxytocin neurons from being activated by factors other than those relevant to parturition and lactation. This kind of 'filtering' action may

contribute to maintain the neurons at a level of depolarization that would facilitate their subsequent synchronous activation just before milk ejection.

To these phenomena in the hypothalamus must be added the facilitatory action on oxytocin secretion provided by glial rearrangements in the neurohypophysis; these result in a significant reduction in glial processes in the vicinity of fenestrated capillaries through which the hormone diffuses. It is obvious that the mere absence of glial processes or end-feet from the perivascular space will promote diffusion of secreted hormones into the circulation, as well as any paracrine actions they may produce within the gland. In addition, the enlarged neurovascular contact area represents both an increased neuronal surface for exocytotic release of the neurohormones and a readier exposure of neurosecretory terminals to capillaries and therefore to the general circulation.

Osmotic Regulation

Oxytocin neurons do not display intermittent high-frequency discharges in hyperosmotic conditions, but they do develop enhanced continuous firing during persistent dehydration. Astrocytic coverage of oxytocin neurons is reduced during chronic dehydration to the same extent as that during lactation, and it is not unlikely that many of aforementioned mechanisms serve to facilitate this kind of activated firing as well.

In addition, there will be effects from other neuroactive molecules known to intervene in osmoregulation, like the amino acid taurine. Taurine is released by astrocytes of the magnocellular nuclei in an osmodependent fashion. It acts on neuronal glycine receptors and, at high concentrations, GABA-A receptors, contributing to mechanisms that inhibit neuronal activity. This kind of regulation may be affected significantly by a reduced glial ensheathement, since the distance between the release site of taurine and its target receptors will be enhanced, thereby reducing the level of glycine and GABA-A receptor activation. Such a reduction may then account, even partly, for the reduced osmosensitivity of oxytocin neurons noted in lactating animals.

The Ovarian Cycle and Reproduction

The neurohormone GnRH, synthesized in neuronal somata scattered in the basal forebrain, is secreted from neurosecretory terminals into the portal circulation of the median eminence to reach the pituitary, where it modulates the secretion of gonadotropins essential to gonadal function and reproduction. As described earlier, the astrocytic ensheathement/synapse relationship of GnRH neurons in various species varies in relation to fluctuating levels of gonadal steroids in the circulation. Even though we lack direct evidence, it is highly likely that astrocytes contribute to estrogen-dependent increases in neuronal firing that are correlated in time with GnRH release. Enhanced levels of the steroid augment astrocytic coverage of GnRH neurons, and this could contribute to an elimination of inhibitory GABAergic inputs that takes place concurrently. A greatly reduced inhibitory input will then certainly increase neuronal excitability and ultimately, secretion of GnRH. As in the hypothalamic–neurohypophysial system, concurrent morphological changes in the median eminence will further contribute to enhance such secretion, since retraction of glial end-feet from the perivascular area exposes GnRH terminals directly to capillaries, thereby facilitating passage of GnRH into the circulation.

See also: Astrocyte: Response to Injury; Astrocyte: Neurotransmitter and Hormone Receptors; Astrocyte: Identification Methods; Calcium Homeostasis in Glia; Cell–Cell Communication Through the Extracellular Space; D-Serine: From its Synthesis in Glial Cell to its Action on Synaptic Transmission and Plasticity; Glia and Synapse Formation: An Overview; Potassium Homeostasis in Glia; Transporter Proteins in Neurons and Glia.

Further Reading

Bonfanti L, Poulain DA, and Theodosis DT (1993) Radial glia-like cells in the supraoptic nucleus of the adult rat. *Journal of Neuroendocrinology* 5: 1–5.

Dityatev A and Schachner M (2003) Extracellular matrix molecules and synaptic plasticity. *Nature Reviews Neuroscience* 4: 456–468.

Garcia-Segura LM, Chowen JA, and Naftolin F (1996) Endocrine glia: Roles of glial cells in the brain actions of steroids and thyroid hormones and in the regulation of hormone secretion. *Frontiers in Neuroendocrinology* 17: 180–211.

Grosche J, Matyash V, Möller T, et al. (1999) Microdomains for neuron–glia interaction: Parallel fiber signaling to Bergmann glial cells. *Nature Neuroscience* 2: 139–143.

Haydon PG (2001) Glia: Listening and talking to the synapse. *Nature Reviews Neuroscience* 2: 185–193.

Hussy N, Deleuze C, Desarménien MG, et al. (2000) Osmotic regulation of neuronal activity: A new role for taurine and glial cells in a hypothalamic neuroendocrine structure. *Progress in Neurobiology* 62: 113–134.

Jourdain P, Israel J-M, Dupouy B, et al. (1998) Evidence for a hypothalamic oxytocin-sensitive pattern-generating network governing oxytocin neurons *in vitro*. *Journal of Neuroscience* 18: 6641–6649.

Laming PR, Kimelberg H, Robinson S, et al. (2000) Neuronal–glial interactions and behaviour. *Neuroscience and Biobehavioral Reviews* 24: 295–340.

Ludwig M and Leng G (2006) Dendritic peptide release and peptide-dependent behaviours. *Nature Reviews Neuroscience* 7: 126–136.

Oliet SHR, Piet R, and Poulain DA (2001) Control of glutamate clearance and synaptic efficacy by glial coverage of neurons. *Science* 292: 923–926.

Oliet SHR, Piet R, Poulain DA, et al. (2004) Glial modulation of synaptic transmission: Insights from the hypothalamic supraoptic nucleus. *Glia* 47: 258–267.

Panatier A, Theodosis DT, Mothet J-P, et al. (2006) Glial-derived D-serine controls NMDA receptor activity and synaptic memory. *Cell* 125: 775–784.

Sykova E (2004) Extrasynaptic volume transmission and diffusion parameters of the extracellular space. *Neuroscience* 129: 861–876.

Theodosis DT (2002) Oxytocin-secreting neurons: A physiological model of morphological neuronal and glial plasticity in the adult hypothalamus. *Frontiers in Neuroendocrinology* 23: 101–135.

Theodosis DT and MacVicar BA (1996) Neuron–glia interactions in the pituitary and hypothalamus. *Trends in Neurosciences* 19: 363–367.

Theodosis DT and Poulain DA (1993) Activity-dependent neuronal–glial and synaptic plasticity in the adult mammalian hypothalamus. *Neuroscience* 57: 501–535.

Volterra A and Meldolesi J (2005) Astrocytes, from brain glue to communication elements: The revolution continues. *Nature Reviews Neuroscience* 6: 626–640.

Glial Responses to Injury

P G Popovich, L B Jakeman, and D M McTigue, The Ohio State University College of Medicine, Columbus, OH, USA

© 2009 Published by Elsevier Ltd.

Introduction

Glia of the central nervous system (CNS) have largely been considered as support cells that, if injured, become activated but do not necessarily affect neurological function directly. Our current understanding of astrocytes, microglia, and oligodendrocytes reveals that this is a dramatic oversimplification. Collectively, glia outnumber neurons and are tightly integrated with all aspects of neurological function. Moreover, injury-induced activation of glia can directly and indirectly cause both neuronal death and repair. Here, we present the principal glial components of the CNS, their roles in homeostasis, and their responses to injury.

Astrocytes in Normal CNS

Astrocytes were named in the late nineteenth century for their characteristic starlike morphology revealed by silver staining of the intermediate filament cytoskeleton. Dye-filling techniques have revealed a more complete picture of astrocyte morphology. Specifically, astrocytes have a bushy appearance and exhibit larger and more elaborate processes than revealed by silver stains. Astrocytes are the most abundant cells in the CNS and are homogeneously distributed throughout the brain, spinal cord, and retina.

During development, astrocytes arise from progenitor cells found within the ventricular zone and subventricular zone (SVZ). Throughout life, the SVZ is maintained as a thin layer of proliferating cells that can give rise to neurons and glia. Astrocytes and their precursors help sustain neuronal growth and migration by providing essential guidance cues and structural and trophic support.

Astrocytes also help regulate neural and vascular homeostasis. Astrocytic processes closely appose synapses, blood vessels, and the pial surface. Consequently, astrocytes are ideally situated to respond to perturbations. For example, as neurons fire, extracellular K^+ rises, causing astrocytes to depolarize and increase intracellular glycogen reserves. Subsequently, astrocytes release glucose, which supports neuronal energetics. Astrocytes also modulate excitatory neurotransmission by quickly removing glutamate from synapses and converting it to glutamine. Astrocytes interact with endothelia to form and maintain the blood–brain barrier and water balance. The latter is mediated by aquaporin 4 channels on astrocytic end feet found along microvessels. Because mature astrocytes are coupled to one another through gap junctions, they are highly sensitive to physiological and pathological alterations in the CNS. In response to injury, astrocytes represent a syncitium of multifunctional cells that coordinate cellular responses at each interface of the CNS.

Astrocytes in the Injured CNS

CNS injury kills neurons and glia at the site of trauma due to shearing, loss of membrane integrity, and failure to restore ionic homeostasis. However, many astrocytes, microglia, and oligodendrocytes surrounding the site of injury survive. It is the collective interactions of these surviving glia that influence repair and glial-mediated injury processes in the CNS.

The signals that activate astrocytes are multifaceted. Tissue culture experiments show that mechanical stretching or cytokines can elicit a coordinated wave of calcium influx in adjacent astrocytes. Once thought to occur exclusively through gap junctions, this form of intercellular communication requires astrocyte-derived ATP acting at P2X7 receptors. Astrocyte ATP also activates adjacent microglia. As neurons and glia die, astrocytes sequester extracellular K^+ and glutamate through specific transporters. Water uptake is also enhanced through upregulation of aquaporin channels associated with blood vessels. Subsequent failure in Na^+/K^+ exchange causes intracellular solute concentrations to increase and astrocytes to swell. In cases of mild injury, astrocytic swelling and enhanced glutamate transport promote amino acid metabolism, glycogen synthesis, and glutathione release. These compensatory changes evoked by reactive astrocytes constitute a beneficial response to injury. However, more severe CNS damage can overwhelm the buffering capacity of astrocytes, causing them to lose their ability to remove extracellular glutamate. In fact, astrocytes can kill neurons and glia by releasing glutamate and oxidative molecules including nitric oxide. Astrocyte swelling and subsequent loss of contact with endothelia may also exacerbate vasogenic edema, whereas release of K^+ and nitric oxide contributes to vasodilation at the injury site. Activated astrocytes also regulate microglial activation and subsequent neuroinflammation through release of ATP, cytokines, and chemokines. In time, the inflammatory

environment – rich with microglia, macrophages, and the molecules they release – causes astrocytes to transition from an activated to a reactive state.

Reactive Astrocytes and the Glial Scar

Reactive astrocytes are most often identified at the injury site by increased synthesis of glial fibrillary acidic protein (GFAP) and changes in expression of other intermediate filaments, including vimentin and nestin. Astrocyte hypertrophy, visible using GFAP immunohistochemistry, occurs in gray and white matter within a few days after injury and can persist for weeks or months (Figures 1(a) and 1(b)). Astrocytes migrate toward the lesion, where they extend their processes toward and along the borders of the

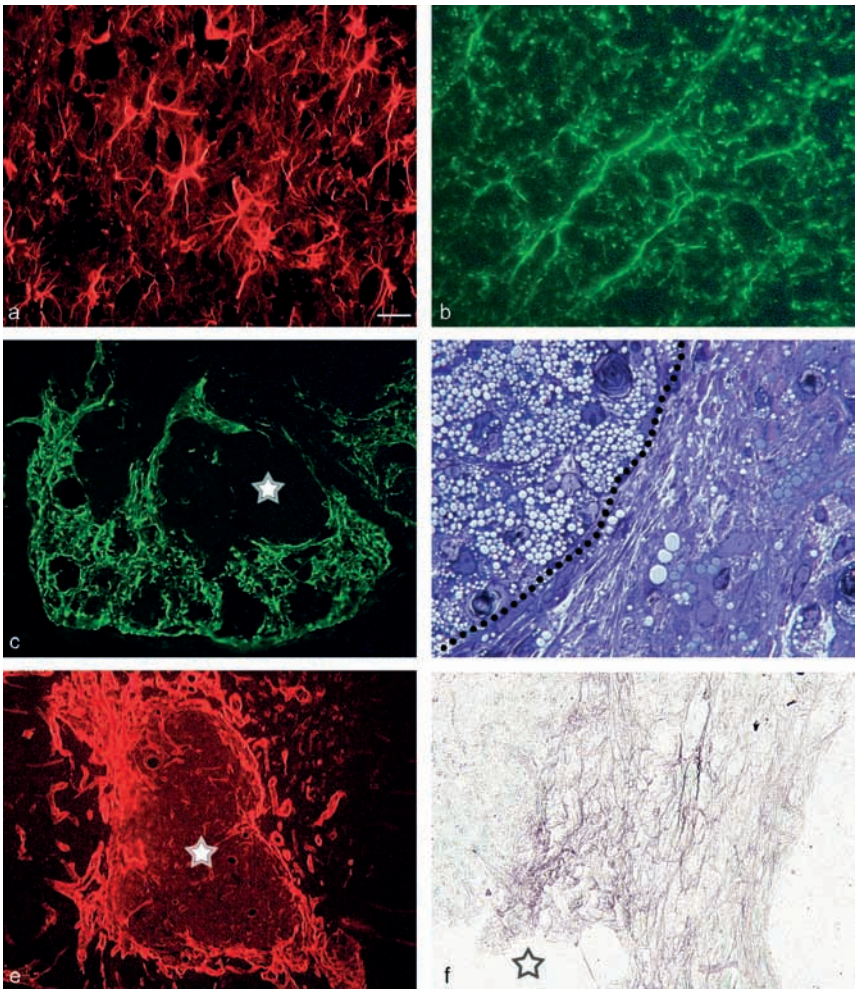

Figure 1 Astrocytes. (a) Reactive astrocytes in mouse spinal cord gray matter ~7 mm rostral to the site of contusion injury (2 weeks postinjury). Stained with rabbit anti-GFAP and Texas Red donkey anti-goat antibodies. (b) Reactive astrocytes in transverse section of mouse spinal cord white matter ~2 mm rostral to the site of a contusion injury (6 weeks postinjury). Stained with rabbit anti-GFAP and FITC goat anti-rabbit antibodies. (c) Transverse section through spinal cord ventromedial white matter at the site of a moderate contusion injury (6 weeks postinjury). Reactive astrocytes circumscribe the lesion in spared tissue but do not enter into the lesion center (star, unstained regions). Stained with rabbit anti-GFAP and FITC goat anti-rabbit antibodies. (d) Toluidine blue-stained 1 μm-thick section from the glial border of a moderate contusion injury in C57Bl/6J mouse (6 weeks postinjury). Macrophage clusters occupy the lesion site (left of dotted line) and contain degenerating myelin. Closely aligned and interwoven astrocyte processes are evident along the borders of the macrophage-dense region (right of dotted line). (e) Longitudinal section through the site of a spinal contusion in mouse (2 weeks postinjury). A dense border of laminin reactivity is seen along the edges of the injured spinal cord separating the central core of injury (star) from spared or residual tissue. This illustrates the formation of the glial limitans at the lesion edge. Stained with rabbit anti-laminin and Texas Red goat anti-rabbit antibodies. (f) Immunostaining for CSPG glycosaminoglycans (GAGs) on longitudinal sections from an injured mouse spinal cord (1 week postinjury). Note the increased CSPG GAG staining in the glial scar immediately surrounding the lesion (star), with little staining in the lesion center. Positive staining is associated with cellular processes resembling astrocytes and with the surrounding extracellular matrix. Stained with CS56 and goat anti-mouse antibodies. Scale bar = 20 μm (a); scale bar from (a) = 10 μm (b, d), 125 μm (c), 250 μm (e), 100 μm (f).

injury (Figure 1(c)). As these processes encounter activated inflammatory cells and invading mesenchymal cells, they circumscribe the zones of inflammation but do not advance beyond the lesion edge (Figure 1(d)). There, they secrete extracellular matrix (ECM) molecules required for the assembly of a basal lamina or a reconstituted glial limitans (Figure 1(e)). Reactive astrocytes are also seen in distal regions of axonal degeneration, where cell–cell communication with axons, neurons, or microglia may be the triggering factor. In cases of severe injury, astrocyte proliferation can occur, although most proliferation is associated with other cell types. Together with microglia and $NG2^+$ cells, the reactive astrocyte syncitium forms the glial scar.

Increased GFAP expression after injury is a stereotypical response associated with numerous signaling cascades, which are triggered by cytokines, growth factors, hormones, and extravasated blood components such as thrombin. This cardinal feature of reactive astrocytes has been exploited experimentally. By placing transgenes under the control of the GFAP promoter, various adhesion molecules, growth factors, cytokines, and transcription factors have been deleted or overproduced, thereby revealing their biological potential at sites of CNS injury.

Consequences of the Glial Scar

The glial scar serves a dual role in CNS repair. Historically, reactive astrocytes were thought to form a mechanical barrier to axonal growth. Indeed, axons are often found growing along, but not through, an intact or reconstructed glia limitans. However, because reactive astrocytes and other glia in the region produce ECM molecules that cause axon repulsion or die-back, the glial scar is now considered to be primarily a molecular barrier to axonal growth.

Chondroitin sulfate proteoglycans (CSPGs) are a family of ECM molecules upregulated by reactive astrocytes surrounding an injury site (Figure 1(f)). During CNS development, CSPGs are essential because they restrict diffusion of growth factors and limit sprouting and cell migration into inappropriate regions. These same properties can impede regeneration after CNS injury. Indeed, numerous studies have shown that CSPGs contribute to abortive sprouting and failed CNS axon regeneration.

Although the glial scar serves as a barrier to regeneration, experimental data also show that astrogliosis and glial scarring are essential for restoring tissue homeostasis and limiting secondary neurodegeneration after CNS injury. For example, if glial scarring is inhibited by killing reactive astrocytes or by preventing their migration to the injury site, traumatic lesions become larger, neuronal death is greater, neuroinflammation is enhanced, blood–brain barrier function is impaired, and functional recovery can be reduced. Thus, although activated astrocytes are an essential component of the CNS injury response, their growth-inhibitory effect must be titrated or fundamentally changed in order to promote axonal growth.

Microglia

In the early 1900s, del Rio-Hortega recognized microglia as a distinct glial subpopulation. Using a stab wound model in brain, he concluded that microglia had properties similar to macrophages. Nearly 75 years later, microglia are known to be resident CNS macrophages and are often referred to as the brain's immune system. Once activated, microglia rapidly change shape, newly express surface antigens that facilitate cell–cell interactions, and release molecules with neurotrophic and neurotoxic effects. This latter feature of activation has stoked a controversy regarding the biological effect of activated microglia in the injured CNS. Indeed, countless *in vitro* studies have shown that microglia can kill neurons and glia; however, few *in vivo* studies have proven such a cause–effect relationship. This is because it is difficult to determine which process occurs first – neuronal injury or microglial activation. Newer *in vivo* imaging techniques and conditional knockout or transgenic mice will undoubtedly help answer these questions. Despite these ambiguities, our understanding of microglia at rest and in response to injury has changed dramatically during the past 20 years. The following section provides a summary of the various concepts and primary research data that have shaped how we think about microglia after CNS injury.

Microglia in the Normal CNS

Microglia comprise up to 20% of all glia and are homogeneously distributed throughout the CNS (Figure 2(a)). Historically, microglia were believed to have little or no physiological role in the normal CNS. However, using multiphoton imaging to visualize 'green' microglia in the uninjured brain of genetically modified mice, we now know that microglia constantly extend and retract filopodia in a random pattern around their cell bodies. This dynamic surveillance mechanism provides a complete scan of the extracellular space every few hours and is dependent on ATP derived from astrocytes. Clearly, complex interactions between astrocytes and microglia help to maintain CNS homeostasis.

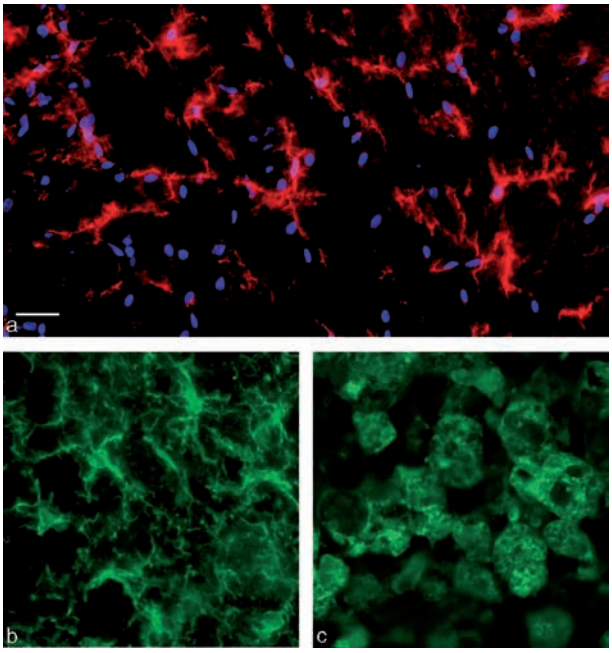

Figure 2 Microglia. (a) Microglia in intact rat spinal cord white matter. OX42 labeling with DAPI counterstain. (b) Activated microglia in lumbar spinal cord of C57BL/6 mouse with an active CNS inflammatory disease (EAE). Microglia labeled with Iba-1 antibody. (c) Phagocytic microglia/macrophages at the center of a mouse spinal contusion lesion (7 days postinjury). Macrophages labeled with FA/11 antibody specific for CD68. Scale bar = 20 μm (a–c).

Microglia in the Injured CNS

Almost immediately after CNS injury, microglia extend filopodia toward the site of damage and undergo a dramatic morphologic and phenotypic transformation. The stellate cell that predominates prior to injury becomes hypertrophied and starts to proliferate (Figure 2(b)). At sites of blood–brain barrier injury, microglia interact with newly recruited phagocytes (Figure 2(c)). In fact, using conventional histology or immunohistochemistry, it is not possible to distinguish between phagocytic microglia and macrophages derived from blood monocytes. As microglia change shape and proliferate, they begin to newly express adhesion molecules, major histocompatability complex antigens, and cytokine/growth factor receptors. These surface proteins facilitate intercellular communication and are closely coupled to the transformation of microglia into immune effector cells capable of secreting an array of bioactive molecules with neurotrophic and neurotoxic effects.

Cytokines, Chemokines, and Microglial-Mediated Injury and Repair

Given their rapidity of activation, their homogeneous distribution in the CNS, and their ability to effectively engage infiltrating leukocytes by the release of cytokines and chemokines, microglia are pivotal in coordinating pathological and repair sequelae after CNS injury. Of the cytokines that are released by microglia after injury, interleukin-1β (IL-1β) and tumor necrosis factor-α (TNF-α) have received the most experimental scrutiny. Still, our understanding of these molecules is incomplete.

Logically, it is difficult to envision that microglia are programmed to release cytokines with the intent to kill. It makes little intuitive sense that the CNS would develop with a well-distributed mechanism for self-destruction. Rather, the magnitude and timing of cytokine release and the microenvironment in which target cells (e.g., neurons and oligodendrocytes) encounter cytokines will determine death or survival. For example, before injury, physiological levels of TNF-α help control synaptic activity by modulating AMPA receptors. In the initial phases of injury, low but rising levels of TNF-α can protect neurons from oxidative and excitotoxic injury; however, as TNF-α levels surpass an undefined threshold, cell death occurs. The differential effects of TNF-α may also be dictated by whether target cells preferentially express TNFR1 (p55) or TNFR2 (p75) receptors. TNF-α binding to TNFR2 initiates signaling pathways that promote cell survival. Activation of TNFR2 also promotes proliferation of oligodendrocyte progenitor cells (OPCs) and subsequent remyelination. Although IL-1β can exert similar reparative effects on neurons and oligodendrocytes, the predominant view is that IL-1β potentiates or causes neurotoxicity. For this reason, analogs of the naturally occurring IL-1 receptor antagonist (IL-1ra) are being tested in clinical trials to treat stroke and are under consideration for use in other CNS disorders.

Data indicate that fractalkine (CX3CL1), a chemokine released exclusively by neurons, influences and regulates microglial activation at rest. As neurons die, their ability to control microglia is impaired. This specificity of control is due to the fact that microglia are the predominant cells expressing fractalkine receptor (CX3CR1) in the CNS. Using CX3CR1-deficient mice, investigators revealed the importance of this ligand–receptor interaction in controlling microglia-mediated neurotoxicity. In response to an inflammatory challenge, neuronal degeneration was enhanced when microglia lacked CX3CR1. Surprisingly, the magnitude of microglial activation and neuronal degeneration was not changed in models of mechanical or traumatic nerve injury (e.g., facial nerve axotomy). These disparate results illustrate the complexity of microglial activation and subsequent cytokine synthesis and emphasize our incomplete understanding of the role of microglia after CNS injury.

Microglial-Derived Growth and Trophic Factors

In addition to their role in amplifying and coordinating inflammation, microglia also release classical neurotrophins, including brain-derived neurotrophic factor (BDNF), nerve growth factor (NGF), and NT-3. Consequently, activated microglia have been associated with improved neuronal survival and axonal elongation in the injured CNS. Despite the inherent advantages of having a resident source of neurotrophins, the uncoordinated release of these molecules in an injured microenvironment can also evoke maladaptive responses. For example, focal activation of microglia with subsequent increases in NGF or BDNF can elicit aberrant sprouting of sensory afferents or alter neuronal firing in dorsal horn neurons, causing neuropathic pain.

Transforming growth factor-β1 (TGF-β1) is another microglial-derived cytokine with pleiotropic effects on glia, leukocytes, endothelia, and neurons. The biological importance of this cytokine in peripheral wound healing and scarring is well established. Although data indicate that microglial TGF-β1 may support neurogenesis, this cytokine/growth factor is consistently implicated in promoting glial scarring and the deposition of an extracellular matrix (ECM) that impairs axon regeneration.

Microglia and Extracellular Matrix Molecules

After CNS injury, there is a dramatic change in ECM composition within and around the site of injury. Although astrocytes are pivotal in forming this matrix, components of the ECM activate microglia. Microglia in turn regulate production and degradation of select ECM proteins via the release of cytokines (e.g., TGF-β1) and metalloproteinases. After injury, OPCs and a subpopulation of microglia also upregulate expression of NG2. Like other CSPGs, NG2 inhibits axon growth *in vitro* and perhaps *in vivo*.

Oligodendrocytes and OPCs in the Normal CNS

Oligodendrocytes were first described by Robertson in 1899 as small cells with few processes. In 1921, Rio Hortega officially named these cells oligodendrocytes, which means 'few tree cells.' Like other glia, oligodendrocytes are found throughout the CNS gray and white matter, where individual oligodendrocytes can myelinate as many as 40–50 axons. This important function enables rapid saltatory conduction throughout the CNS. Although myelination is the typical, and sometimes only, function attributed to oligodendrocytes, these cells do more than just ensheath axons. Oligodendrocytes can influence neuron survival, axon transport, axon caliber, and ion channel clustering along axons. Many of these effects are independent of myelination. For instance, simple wrapping of axons by oligodendrocyte processes in the absence of compacted myelin triggers local neurofilament accumulation and axon caliber expansion. In addition, studies of transgenic mice in which oligodendrocytes lacked proteolipid protein or cyclic nucleotide phosphodiesterase revealed marked axon degeneration in the absence of demyelination.

Because oligodendrocytes are postmitotic, trauma or disease of the CNS can cause demyelination with subsequent loss of neurological function. However, in some circumstances, oligodendrocytes may be replaced by $NG2^+$ OPCs, which exist throughout the adult CNS. Currently, the complete functional repertoire of $NG2^+$ cells is unknown; however, it is clear that at least a subset function as OPCs and contribute to oligodendrocyte replacement in the injured or diseased CNS.

Oligodendrocytes and OPCs in the Injured CNS

Oligodendrocytes and OPCs are particularly vulnerable to CNS injury and subsequent secondary injury cascades. For instance, the release of proinflammatory cytokines such as TNF-α and interferon-γ (IFN-γ) from microglia and leukocytes can induce oligodendrocyte apoptosis. OPCs and oligodendrocytes are also extremely sensitivity to oxidative stress, due to low glutathione levels and high intracellular iron content.

In injury or disease states, activated macrophages, dying neurons, and dysfunctional or overloaded astrocytes contribute to the rise in extracellular glutamate. Because oligodendrocytes express AMPA, kainite, and NMDA glutamate receptors, they can succumb to glutamate excitotoxicity. Excess glutamate also decreases OPC proliferation and ultimately causes OPC death. Since these progenitors are needed to produce new oligodendrocytes, the health and survival of OPCs is paramount if endogenous remyelination is to occur.

As oligodendrocytes die, their myelin sheaths unravel and myelin breakdown products accumulate, which can lead to demyelination and functional loss. In addition, degenerating myelin can impair axonal regeneration. Several studies have shown that components of CNS myelin (e.g., myelin-associated glycoprotein (MAG) and Nogo) inhibit growing axons. As described previously, axonal growth inhibition is also mediated in part by reactive astrocytes and the glial scar. Ironically, as the glial scar inhibits growth, it may also protect oligodendrocytes and OPCs by

sequestering these cells from the toxic lesion environment. Moreover, through direct contact and by secretion of growth factors (e.g., ciliary neurotrophic factor, basic fibroblast growth factor (bFGF), and platelet-derived growth factor (PDGF)), astrocytes can promote oligodendrocyte health and replacement.

OPC Responses

OPCs are activated rapidly by CNS injury. Within 24 h, OPCs in and around areas of CNS injury hypertrophy and increase expression of NG2 (Figures 3(a) and 3(b)). OPCs also proliferate and migrate into and around lesioned CNS tissue. On a molecular level, OPCs increase expression of developmentally important transcription factors, such as Olig2 and Nkx2.2. Although the mechanisms responsible for OPC activation are not well defined, *in vitro* data have shown that OPCs are responsive to growth factors and inflammatory by-products produced within the lesion environment. For instance, proliferation is promoted by leukemia inhibitory factor and the chemokine CXCL1, which acts through the CXCR2 receptor on OPCs to stimulate mitosis and prevent migration. OPC survival and differentiation can be enhanced by IL-4. Classical proinflammatory cytokines such as TNF-α and IFN-γ are deleterious to OPC survival, an effect which can be potentiated by the high intracellular iron levels within OPCs.

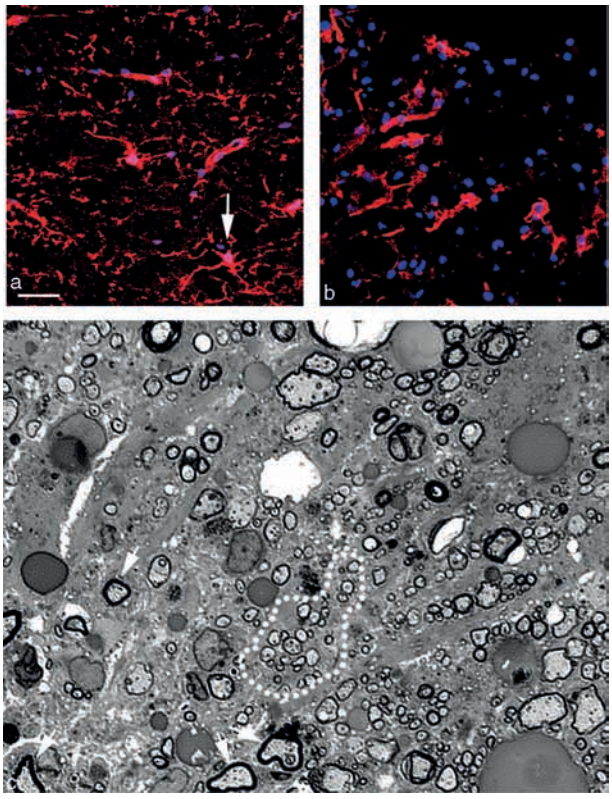

Figure 3 Oligodendrocytes and NG2+ cells. (a) In normal CNS tissue, NG2 cells display small cell bodies with elongated processes. Anti-NG2, red; DAPI, blue nuclei. (b) NG2 cells within a spinal cord lesion site are hypertrophied with fewer processes. Anti-NG2, red; DAPI, blue nuclei. (c) Oligodendrocyte remyelination after spinal cord injury (70 days postinjury). Note that the thickness of the myelin sheath in the cluster of axons remyelinated by oligodendrocytes (dotted line) is consistently thinner than that of normal intact oligodendrocyte myelin (arrows). Scale bar = 20 μm (a, b).

For effective remyelination to occur, proliferating OPCs must undergo a differentiation process. As OPCs differentiate, they upregulate expression of myelin proteins (e.g., PLP) while downregulating NG2 and PDGFαR. Functional maturation is considered complete when the cell processes of an oligodendrocyte engage bare axons and form compacted myelin. Although this can occur in the adult CNS, the thickness of oligodendrocyte remyelination does not recapitulate the preinjury state (Figure 3(c)). Still, relative to a demyelinated axon, thin oligodendrocyte myelin sheaths support efficient conduction of action potentials. Thus, manipulating OPCs is a theoretically feasible mechanism for promoting functionally significant CNS repair. However, in practice this may be difficult. Indeed, the repair potential of OPCs has been largely defined and examined using chemically induced demyelination in rodent CNS where remyelination is typically complete. The situation is less optimistic in human CNS. Studies examining human multiple sclerosis (MS) plaques have shown that OPCs often infiltrate demyelinated plaques but they do not always differentiate into remyelinating oligodendrocytes. This may be due to a lack of appropriate growth factors in MS tissue or to active inhibition of this process by the bare axons, nearby astrocytes, or infiltrating inflammatory cells.

OPC responsiveness to changes in the environment requires appropriately timed signals from the surrounding tissue. Although many of these signals have been identified in the developing CNS, those required for adult remyelination are still under investigation. Likely candidates include PDGF and bFGF. Inflammation and activated microglia are also likely contributors to remyelination. Specifically, there is a surprising impairment of remyelination in the pathological CNS of animals deficient in TNF-α or MCH class II. Depletion of peripheral macrophages also reduces remyelination after toxin-induced demyelination. The situation is not completely straightforward, however, because the same depletion of macrophages promotes sparing of myelinated axons and neurons in models of CNS autoimmune demyelination, stroke, and trauma.

Potential Roles for OPCs Other Than Oligodendrocyte Replacement

Because NG2 cell processes contact nodes of Ranvier, blood vessels, and synapses, it has been proposed that OPCs serve functions other than predecessors of remyelinating oligodendrocytes. In the hippocampus, NG2 cells have been shown to express glutamate receptors and 'synapse' with glutamatergic neuron terminals. After electrical stimulation of neurons, ion fluxes have been described within NG2 cells. Thus, NG2 cells, along with astrocytes, may help maintain synaptic function.

NG2 cells also may represent an endogenous CNS stem cell. *In vitro* studies have shown that under certain culture conditions, NG2 cells can be induced to differentiate into astrocytes, oligodendrocytes, or neurons. Currently, there is no independent confirmation that astrocyte or neuron formation from $NG2^+$ cells occurs *in vivo*. However, based exclusively on the *in vitro* data, there is hope that boosting the spontaneous gliogenic or neurogenic potential of OPCs may be a viable approach for cell replacement in the injured CNS.

Conclusion

CNS injury triggers a stereotypic cascade of glial activation. The consequences of this response are complex and involve the coordinated interaction of astrocytes, microglia, and oligodendrocytes. Apparent in every aspect of gliosis is the potential for repairing the injured CNS; however, this potential appears to be inefficient and constantly at odds with the ability of all glia to either cause damage (e.g., astrocytes and microglia) or inhibit endogenous repair (e.g., glial scar and myelin inhibitors).

See also: Astrocyte: Response to Injury; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Plasticity and Neuroendocrine Regulation; Microglia Properties; Microglial Response to Injury.

Further Reading

Arnett HA, Mason J, Marino M, Suzuki K, Matsushima GK, and Ting JP (2001) TNF alpha promotes proliferation of oligodendrocyte progenitors and remyelination. *Nature Neuroscience* 4: 1116–1122.

Bergles DE, Roberts JD, Somogyi P, and Jahr CE (2000) Glutamatergic synapses on oligodendrocyte precursor cells in the hippocampus. *Nature* 405: 187–191.

Davalos D, Grutzendler J, Yang G, et al. (2005) ATP mediates rapid microglial response to local brain injury *in vivo*. *Nature Neuroscience* 8(6): 752–758.

Kettenmann H and Ransom BR (2005) *Neuroglia*, 2nd edn. New York: Oxford University Press.

Microglia (special issue) (1993) *Glia* 7(1): 1–118.

Microglia (special issue) (2002) *Glia* 40(2): 131–269.

Nimmerjahn A, Kirchhoff F, and Helmchen F (2005) Resting microglial cells are highly dynamic surveillants of brain parenchyma *in vivo*. *Science* 308(5726): 1314–1318.

Pekny M and Nilsson M (2005) Astrocyte activation and reactive gliosis. *Glia* 50: 427–434.

Silver J and Miller JH (2004) Regeneration beyond the glial scar. *Nature Reviews Neuroscience* 5: 146–156.

Sofroniew MV (2005) Reactive astrocytes in neural repair and protection. *The Neuroscientist* 11: 400–407.

Relevant Website

http://www.microglia.net – The Microglia Home Page.

Glial Responses to Virus Infection

J Martín-García, Drexel University College of Medicine, Philadelphia, PA, USA
F González-Scarano, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The nonneuronal cells of the central nervous system (CNS) that provide support to and promote normal neuronal function are termed glia or glial cells. As described elsewhere in this encyclopedia, astrocytes and oligodendrocytes arise from ectodermal tissue just like neurons. Microglia, the immune cells of the CNS, are derived from the bone marrow and migrate into the nervous system at various points during embryonic development and beyond. Many viruses target cells of the central and/or peripheral nervous systems, and glial cells, infected or not, respond with a battery of mechanisms that to a large extent determine, together with the specific characteristics of the pathogen, the eventual outcome of the infection. The final result depends on both the cell type and the virus causing the infection; different cell types have a variable response to different pathogens, and even to the same pathogen, and different pathogens induce different responses in various cell types. We provide an overview of these responses together with a more detailed description of the glial response to human immunodeficiency virus infection of the CNS.

Glial Cells

Glial cells, neuroglia, or simply glia, the most abundant cells in the nervous system, have been described as nonneuronal cells that provide support and nutrition, regulate the internal environment of the brain (especially the fluid surrounding neurons and their synapses), form myelin, and participate in signal transmission in both the peripheral nervous system and the CNS. They also have important developmental roles and are active participants in synaptic transmission (e.g., regulating clearance of neurotransmitters from the synaptic space). Information regarding the biology and functions of astrocytes or astroglia, the most abundant type of glial cell, and oligodendrocytes is provided in other articles in this encyclopedia. Microglia are the resident brain macrophages and, together with the perivascular macrophages that are found around blood vessels in the brain, they perform immune surveillance in the CNS and are capable of phagocytosis and initiating the inflammatory and immune response to most insults. Despite their mesodermal origin, they are commonly categorized as glia because of their supportive role for neurons. Their roles are discussed in more detail later. Perivascular macrophages and microglia arise from bone marrow elements that settle in the CNS at various times during embryonic development and throughout adult life. Perivascular macrophages and perhaps macrophages associated with other CNS regions such as the choroids plexus and the meninges have a relatively short life span within the CNS and, therefore, a considerable turnover in adult life. The parenchymal microglial cells are long-lived and estimated to have an extremely low turnover on the basis of murine and human studies. This has important implications considering that pathogens that establish persistent infections in cells such as microglia ensure their replication for long periods of time after the initial infection.

Glial Responses

Microglial cells are quiescent in the normal CNS and are generally considered to be unable to perform effector functions such as antigen presentation to other immune cells until they are activated by injury or infection. At that time, microglia become the first line of defense. *In vitro* studies have shown that microglia can present antigens and perform other effector functions when activated by proinflammatory cytokines. In addition, infection of mice with Theiler's murine encephalomyelitis virus (TMEV; a natural mouse pathogen that causes an autoimmune demyelinating disease associated with a persistent infection of the CNS and that is used as a model for multiple sclerosis) has been shown to be persistent in murine microglial cells and perivascular macrophages. In this model, chronic demyelinating disease is dependent on $CD4^+$ T cells responding to myelin epitopes presented by microglial cells, suggesting that microglia most likely play an effector role and are critical in establishing an inflammatory response in autoimmune diseases and during viral infection.

The innate immune response is the initial antigen nonspecific response to insults such as traumatic injury and viral or microbial infection. It leads to the rapid production of chemokines and proinflammatory cytokines and subsequent expansion of the intra-CNS immune response. Cells of the innate immune system use the toll-like receptors (TLRs) to recognize 'danger' signals, microbial structures, and pathogen-associated molecular patterns (PAMPs), which are molecules or molecular structures expressed by various classes of pathogenic organisms that are not normally expressed

by eukaryotes. TLRs are members of an ancient superfamily of proteins, also present in invertebrates and plants, that not only activate the innate immune response but also orchestrate the development of adaptive immune responses. By discriminating between diverse stress or danger molecules, microbial structures, and PAMPs, TLRs are able to mount tailored or signal-specific cellular immune responses.

At least 11 such TLRs have been identified in humans and mice, and some of their known ligands are listed in Table 1. TLR2 recognizes peptidoglycan (PG) and associates with TLR1 to recognize bacterial triacyl lipopeptides and with TLR6 to recognize diacyl lipopeptides. TLR3 recognizes double-stranded RNA (dsRNA) which is produced during the replication of many RNA viruses, whereas TLR4 binds to lipopolysaccharide (LPS) from gram-negative bacteria in association with the LPS-binding protein, CD14. TLR9 recognizes unmethylated CpG DNA primarily found in bacteria. Upon engagement of the TLRs by their ligands, TLR-induced signaling leads to the induction of NF-κB, ultimately resulting in upregulation of transcription of chemokines, proinflammatory cytokines, and cell surface molecules involved in the initiation of adaptive immune responses to pathogens. TLRs are present on a wide variety of cells, but specific cell types show preferential expression of certain TLR molecules, probably related to the specific functions that the particular cell types will execute. Monocytes/macrophages have been shown to express most TLRs (with the likely exception of TLR3) during different stages of maturation, whereas dendritic cell expression of TLRs seems to be dependent on both the cell subset and the maturation state: immature dendritic cells express TLR1, -2, -4, and -5 (responding to LPS and PG with upregulation of expression of proinflammatory cytokines such as interleukin (IL)-12 and costimulatory molecules B7-1 and B7-2) and their expression decreases upon maturation, whereas TLR3 expression is increased. Finally, in addition to the recognition of dsRNA by TLR3, other TLRs have been identified as being capable of sensing and mediating responses to viral infections. Thus, TLR2 binds to a cytomegalovirus protein and to the hemagglutinin protein of wild-type, but not vaccine strain, measles virus; TLR4 interacts with the fusion protein of respiratory syncytial virus and with the envelope protein of mouse mammary tumor virus; TLR9 binds to herpesviral genomic dsDNA in mouse cells; and TLR7 in murine dendritic cells and TLR8 in human dendritic cells and macrophages bind to viral ssRNA.

Most TLR molecules signal through a common adaptor protein, MyD88, which binds to the conserved

Table 1 Toll-like receptors in the central nervous system

TLR	*Ligands*	*CNS cell types*		
		Microglia	*Astrocytes*	*Oligodendrocytes*
TLR1	Triacyl lipopeptides (associated with TLR2)	+++	++	
TLR2	MALP-2 (mycoplasma)	+++	++	+
	LAM (mycobacteria)			
	Lipoproteins (gram-positive bacteria)			
	Zymozan (yeast)			
	Hemagglutinin protein of wild-type measles virus[a]			
	Virion protein of cytomegalovirus[a]			
TLR3	Viral dsRNA[a]	++	+++	+
TLR4	LPS (gram-negative bacteria)	+++	++	
	Hsp60, fibronectin, hyaluronan (host-derived factors from stressed or damaged cells)			
	Protein F of respiratory syncytial virus[a]			
	Envelope protein of mouse mammary tumor virus[a]			
TLR5	Bacterial flagellin	++	++	
TLR6	Diacyl lipopeptides (associated with TLR2)	++	+	
TLR7	Small antiviral compounds (e.g., Imidazoquinolines)			
	ssRNA (in murine dendritic cells)[a]	+++	+	
TLR8	GU-rich ssRNA (in human dendritic cells and macrophages)[a]	++	−	
TLR9 (intracellular)	Bacterial unmethylated CpG DNA	+	++	
	Herpesviral genomic dsDNA (in mice)[a]			
TLR10	Unknown	+/−	+	
TLR11	Unknown	?	?	

[a]Viruses or viral molecules that bind to these receptors.

cytoplasmic domain on the TLRs. Following stimulation, MyD88 activates the IL-1R-associated kinase, which associates with tumor necrosis factor receptor (TNFR)-associated factor 6, leading to activation of the NF-κB family of transcription factors. MyD88-independent pathways unique to TLR3 and TLR4 have also been identified, including the activation of interferon regulatory factor-3 and subsequent upregulation of expression of several interferon-inducible genes. Thus, engagement of the TLRs leads to transcriptional activation of cytokines, chemokines, effector molecules, costimulatory molecules, and major histocompatibility complex (MHC) class I and II molecules.

Research has begun to elucidate the expression of TLRs in microglia, as well as their activation by various stimuli. Studies conducted directly on human tissue showed that the brain expressed all TLR molecules except for TLR6 and TLR8, although at levels lower than those observed in a peripheral organ such as the spleen. Slightly different results have been obtained with isolated adult human microglia, which expressed all TLR molecules except for TLR9. Human brain tissue specimens also showed that astrocytes and oligodendrocytes only expressed TLR2 and TLR3. Similarly, it has been shown that cultured neonatal murine microglia expressed all the TLR molecules, although expression in *in vitro* culture and in the CNS environment may vary, thus explaining the slight differences. *In vitro* stimulation of quiescent murine microglia with various TLR agonists, including LPS (TLR4), PG (TLR2), polyinosinic–polycytidylic acid (poly(I:C)) (TLR3), bacterial CpG DNA (TLR9), as well as infection with TMEV, induced the cells to upregulate innate and effector immune cytokines and chemokines (e.g., CCL3/MIP-1α, CCL2/MCP-1, and CCL5/RANTES) that are required for attraction of peripheral macrophages and T cells into the CNS. In addition, TLR stimulation upregulated MHC class II and some costimulatory molecules (B7-1, B7-2, CD40, and ICAM-1), enabling the microglia to effectively present myelin peptide to $CD4^+$ Th1 cells.

Interestingly, microglial cells apparently develop a stronger response to infection with viable TMEV than when stimulated directly with TLR ligands, including poly(I:C), which signals through TLR3. This suggests that the virus may activate innate immunity via multiple TLRs or through other mechanisms that trigger innate immune recognition, such as through dsRNA-activated protein kinase, which binds intracellular viral dsRNA and induces expression of type I interferons. TLR molecules other than TLR3 have been shown to recognize viruses, viral proteins, or viral nucleic acids. Examples are the TLR4-dependent inflammatory immune response to respiratory syncytial virus protein F and to mouse mammary tumor virus envelope protein, or the TLR2/CD14-dependent inflammatory cytokine production in response to human cytomegalovirus infection. Therefore, it is apparent that a more robust innate immune response can be achieved by stimulation/recognition through several TLR molecules (although a combination of LPS and poly(I:C) does not seem to result in a stronger response than either ligand alone in microglial cells). Genetic susceptibility to herpes simplex virus encephalitis has been mapped to UNC93B, a protein involved in the mediation of the response to TLR3, TLR7, and TLR8, and it has been suggested that a deficient interferon response is specifically responsible for this phenotype. The available data therefore suggest that microglia play a key role in both innate and adaptive immune responses, providing the CNS with the means to rapidly and efficiently respond to a wide range of pathogens.

Astrocytes and oligodendrocytes may also aid in the innate and adaptive immune responses, although microglial cells appear to be the most potent CNS resident antigen presenting cells. Studies on oligodendrocytes are more limited, but TLR expression seems to be low and limited to TLR2 and TLR3. In comparison with microglia, astrocytes also express a somewhat limited set of TLR molecules, with predominant expression of TLR3; adult human astrocytes in culture also express TLR4 to high levels.

The astrocytic response to TLR ligands is similar to that of microglia, with some notable differences. For example, TLR-mediated activation of microglial cells results in positive feedback for TLR2 and TLR3 expression but in downregulation of TLR4 expression. However, astrocytic TLR3 ligation and subsequent activation of astrocytes has been shown to result in upregulation of expression of all TLR2, TLR3, and TLR4. In addition, TLR3 can be found in astrocytes both within the cytoplasm and at the cell surface, whereas it is localized almost exclusively in the former in microglial cells. Another striking difference is that although other TLR molecules may be expressed on astrocytes, cellular activation of *in vitro* systems is limited to TLR3 ligands. The underlying reason for the absence of an astrocytic response to other TLR ligands is unknown, but it may be related to expression levels. Finally, activation of adult human astrocytes through TLR3, but not through TLR4, has resulted in suppression of astrocyte growth, promotion of endothelial cell growth, and enhancement of neuronal survival, providing a comprehensive neuroprotective response rather than a polarized proinflammatory reaction. Some differences in TLR expression and responsiveness have been observed in studies on murine and human astrocytes, although data on murine and human microglia appear to be much more consistent.

Thus, the evidence suggests that more cell type-specific, and possibly species-specific, differences in the responses mediated by TLRs are likely to be found as knowledge accumulates. In addition, in a disease context, multiple TLRs may be engaged by multiple ligands in multiple cell types and the ultimate functional outcome will be dependent on the timing of stimulations and cross-regulation of TLR-mediated signaling.

Viral Infections of the CNS

Viruses from many diverse RNA and DNA families and genera can infect the CNS. Some are only neurotropic – they can infect the nervous system – but are not neuropathogenic in normal circumstances, for example, in individuals who have intact immune systems. Others target the CNS principally, and their importance as pathogens is limited to their neurological disease. Some of these are listed in Table 2. In many ways, viral infection of the CNS is more complex than that of peripheral tissues. First, the presence of the blood–brain and blood–cerebrospinal fluid barriers limits viral trafficking into and out of the CNS. Primarily for this reason it was thought for many years that the neural route was the major route of viral entry into the nervous system, and experimental systems documented the retrograde axonal transport along the peripheral nerves of rabies virus, herpes simplex virus, and some nervous system-adapted strains of polioviruses. The olfactory mucosa is a unique anatomical site where distal processes of neural cells are directly exposed to the environment while the processes at the opposite site of the same neural cells establish connections within the CNS. This extremely specialized mucosal site is exploited by some aerosol-borne and respiratory infections to spread to the olfactory bulbs directly from the olfactory mucosa. Importantly, in most natural and experimental infections, viruses invade the brain from the bloodstream (in the so-called hematogenous route). When present at sufficiently high titers in the blood, viruses may infect endothelial cells and spread from them into the nervous tissue, be transported across the endothelial cells within vesicles, or be carried inside the brain within infected cells such as leukocytes migrating into the brain across the blood–brain barrier (BBB).

Once in the CNS, some viral infections in the brain, such as rabies and polioviruses, involve only neurons, and spread within the CNS most likely occurs through dendrites and other neuronal processes. The type and extent of neuropathology also relate to the target cell population and type of infection. Thus, polioviruses infect motor neurons and cause flaccid paralysis, whereas rabies virus infects neurons of the limbic system and cortical neurons leading to behavioral abnormalities in addition to eventual paralysis. Others, such as herpes simplex virus, infect both neurons and glial cells and cause a more focal encephalitis, preferring the temporal lobes. In contrast, some arboviruses (arthropod-borne viruses) cause generalized encephalitis by widespread infection of neurons and other cells. Murine and human cytomegaloviruses preferentially cause lytic infection in immature glial cells during brain development but mainly establish persistent infection in neuronal cells. Finally, the polyomavirus JC virus is an example of selective infection of oligodendrocytes that can lead to a focal or multifocal demyelinating disease, as seen in progressive multifocal leukoencephalopathy.

Cell damage due to viral infection in the nervous system may occur through necrosis induced directly by the virus, as a by-product of the immune response to the infection, or through apoptosis – a mechanism of programmed cell death that is critical in the normal development of the nervous and lymphoid systems, two organs in which excess cells are generated. At the same time, postmitotic cells such as many neuronal populations are not replenished easily and have mechanisms that are designed to protect them from

Table 2 Representative neurotropic viruses

DNA viruses
α, β, and γ herpesviruses
Herpes simplex virus
Cytomegalovirus
Epstein–Barr
Human herpesvirus 6
JC virus (polyomavirus)
RNA viruses
Retroviruses
HIV-1
HIV-2
Human T lymphotropic virus type 1
Coronaviruses
Enteroviruses (polioviruses, coxsackieviruses, cardioviruses)
Arboviruses
Alphaviruses
Western/Eastern equine encephalitis
Flaviviruses
West Nile virus, St. Louis encephalitis virus
Bunyaviruses
La Crosse virus, Rift Valley fever virus
Arenaviruses
Lymphocytic choriomeningitis virus, Lassa fever virus
Paramyxoviruses
Mumps virus (rubulavirus)
Measles virus (morbillivirus)
Henipaviruses (Nipah virus, Hendra virus)
Rabies viruses (Lyssavirus)
Rubella virus (Rubivirus)

premature apoptosis, such as expression of antiapoptotic genes. Moreover, they do not express histocompatibility antigens, which exempts them from most T cell-mediated cytotoxicity. In many instances, the balance between these two competing effects results in moderate or no cytopathology, leading to mild or slowly progressing disorders instead of a rapidly fatal encephalitis. In contrast to neurons, virus-infected glial cells can be destroyed not only by the virus but also by immune-mediated mechanisms since they have the capacity to express viral antigens on the surface and are subject to both humoral and cell-mediated immune responses. Neurons have been shown to initiate an innate immune response through poly(I:C) and viral dsRNA interaction with neuronal TLR3 with subsequent induction of β-interferon.

Finally, the association of HIV-1 infection with neurological diseases has raised the significance of viral infections of the nervous system to a new level. The extent of the HIV-1 epidemic and the fact that the virus is present in the CNS of the majority of infected patients have made HIV-1 the most prevalent neuropathogenic viral infection ever recognized. HIV/AIDS-related neurological complications are among the most frequent neurological diseases worldwide. In the developed world, highly active antiretroviral therapy has decreased the incidence of, but not eradicated, the most severe forms (e.g., HIV-1-associated dementia and vacuolar myelopathy), whereas the prevalence of milder cognitive disorders remains high and may even be increasing due to the longer survival and more advanced age of HIV-infected individuals. Because of the intensity and breadth of the many studies that have analyzed HIV infection of the CNS, this lentiviral complication now serves as a paradigm for the relationship between a virus and brain cells, specifically glia, since HIV does not infect neurons in significant levels.

Neurotropism of HIV-1

The CNS is susceptible to infection by retroviruses of various species and in particular by members of the lentivirus family, such as HIV-1. The specific requirements for entry into the brain and the variety of cell types within the CNS increase the complexity of virus–cell interactions in this organ, but the BBB is not an absolute barrier for many viral pathogens, such as HIV-1, since it is selectively permeable rather than impermeable. Cells of the immune system can cross the BBB and enter the brain within the context of immune surveillance in a process that is carefully regulated. Using the lentivirus visna virus of sheep as a model, Haase first proposed that HIV-1 and other lentiviruses enter the CNS as passengers in cells trafficking into the brain as part of normal immune surveillance mechanisms (the 'Trojan horse' hypothesis). For HIV-1, which infects $CD4^+$ T cells and monocytes in the circulation, this model is the most intuitively appealing and, unlike alternative mechanisms, has the most compelling evidence.

Figure 1 schematically shows the potential mechanisms of entry of HIV-1 into the brain. *In situ* hybridization and immunohistochemical analyses have demonstrated virus accumulation in perivascular regions, principally in $CD14^+$ cells. CD14 is an LPS receptor present as a membrane-bound form at the surface of myeloid cells, including monocytes, macrophages, and granulocytes. Therefore, the bulk of HIV-1 entry into the brain takes place within infected monocytes crossing the BBB and migrating into the brain to replenish the population of perivascular macrophages (Figure 1, A). In addition, phenotypic analysis of HIV-1-infected monocytes demonstrates an activation pattern that promotes migration into tissues such as the brain. The infected monocytes differentiate into macrophages and release new virions into the brain's extracellular space, spreading infection to additional perivascular macrophages and microglial cells present within the brain parenchyma (Figure 1, B). Microglial cell activation may also play a role in neuropathogenesis.

T cells also travel into the brain as part of the immune surveillance process, and HIV-1-infected T cells could therefore bring virus into the brain (Figure 1, C). However, the contribution of virions produced by infected $CD4^+$ T cells migrating into the brain to the pool of replicating virus present in the CNS may not be significant since genotypic and phenotypic analyses show that HIV strains from brain are closer to those from monocytes/macrophages than to those from T cells.

There is very limited pathological evidence of HIV-1 infection of endothelial cells, although BBB abnormalities due to HIV infection have been demonstrated (Figure 1, D). Other mechanisms, such as macropinocytosis, transcytosis, or localized loss of BBB functionality, might constitute additional pathways for viral penetration of the BBB, but they are unlikely to account for the bulk of the virus entering the brain.

Several mechanisms for HIV-1 infection of astrocytes have been described (Figure 1, E); however, it is accepted that infection of these cells is mostly nonproductive and that they do not contribute significantly to viral replication in the brain since few astrocytes express viral antigens or RNA. The mechanism of viral entry into astrocytes is also unclear since these cells are CD4 negative and do not express large amounts of the major HIV coreceptors (although detection of corresponding mRNAs has been reported). However,

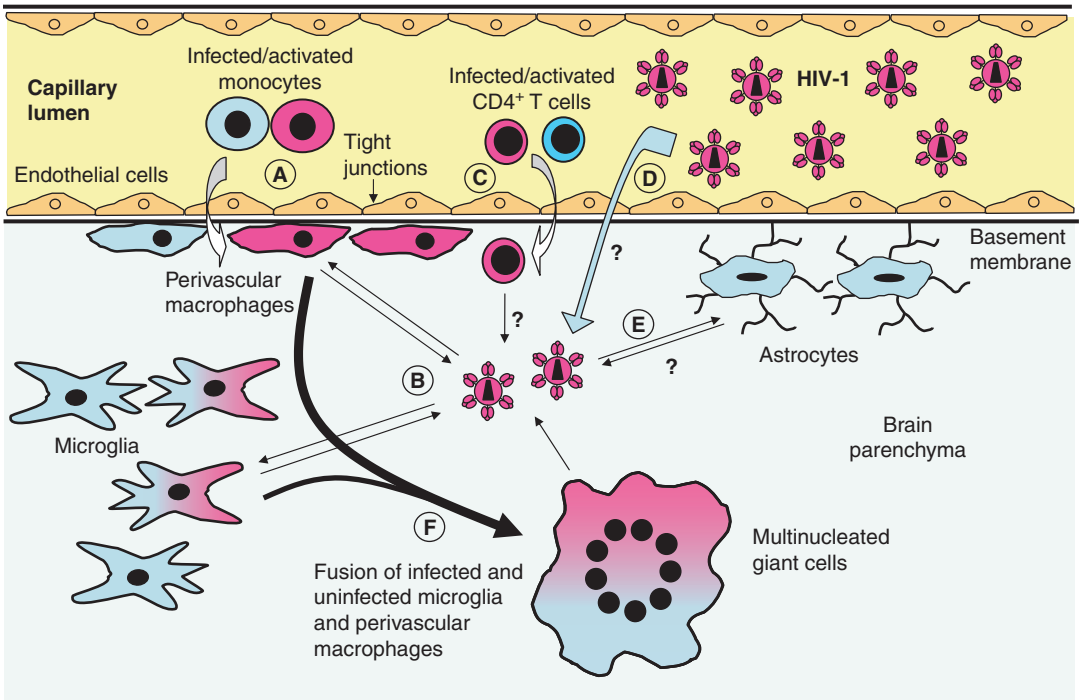

Figure 1 HIV-1 entry into the CNS, spread, and multinucleated giant cell formation.

it seems likely that the very limited HIV-1 gene expression in astrocytes, the astrocytosis induced in HIV-associated dementia, and the deregulation of cytokine and chemokine signals due to infection of macrophages and microglia all play a role in the alteration or loss of astrocytic functions and brain homeostasis leading to neuropathogenesis.

The HIV-1 envelope glycoproteins expressed on the surface of infected cells mediate cell-to-cell fusion with receptor- and coreceptor-positive cells; in the brain, cell-to-cell fusion involves macrophages and microglia and results in the formation of large multinucleated giant cells or syncytia that also produce virus before they eventually die (Figure 1, F). Their life span after undergoing fusion is unknown, but it is estimated to be days to weeks.

There is no *in vivo* evidence of HIV infection of oligodendrocytes, although *in vitro* infection has been reported. However, binding of gp120 to galactosylceramide or other proteoglycans present on the membranes of oligodendrocytes may reduce myelin synthesis and increase intracellular Ca^{2+} levels, which could lead to apoptosis.

Finally, neurons, as major effectors of cognitive and motor function, must be involved in HIV neuropathogenesis, and indeed there is significant accumulated neuronal cell death in HIV-infected brains. However, neuronal infection has been reported only by some investigators using very sensitive techniques, in agreement with the fact that neurons lack expression of the main HIV-1 receptor CD4. However, different neuronal subsets express some of the chemokine receptors that serve as coreceptors for HIV-1 infection, where they may be involved in cell migration.

The Chemokines and Chemokine Receptors Network

In the CNS, chemokines and their receptors are involved in cell migration, differentiation, activation, and proliferation of glial and neuronal cells. They are essential components of glial and neuronal physiology and play a crucial role in the balance between neuroprotection and neurodegeneration by counteracting the apoptotic signals that may be produced in response to injury and infection. CXCR4 and its only ligand, CXCL12/SDF-1α, are both widely expressed in the normal CNS and may be implicated in neuropathogenesis since they appear to be elevated in the cerebrospinal fluid and brain of HIV-1-infected individuals. Members of the CCR subfamily of receptors have also been detected in physiological conditions and upregulated during inflammatory and neurodegenerative diseases (including in the CNS of simian immunodeficiency virus (SIV)-infected macaques and HIV-1-infected patients), mostly in perivascular macrophages and microglia. Although β chemokines are only weakly or not expressed at all in the normal human brain, altered expression is observed in SIV and HIV encephalitis and could contribute to neuronal

degeneration and dysfunction. However, no neurotoxic effects of β chemokines have been demonstrated. In contrast, several studies have shown that CCL4/MIP-1β and CCL5/RANTES can protect cultured neurons against apoptosis induced by gp120. Therefore, reported alterations in expression of chemokines and their receptors in the lentivirus-infected CNS may have either protective or deleterious effects.

The primary protective role of β chemokines and CCR5 in viral infections of the CNS seems to be better established in the case of West Nile virus (WNV) encephalitis. CCR5 presumably functions normally in antimicrobial host defense because it mediates leukocyte chemotactic responses; however, evidence of antimicrobial functions for CCR5 in humans has been elusive. In a mouse model of WNV infection, CNS expression of CCR5 and its ligand CCL5/RANTES were prominently upregulated by the virus, and this was associated with CNS infiltration of $CD4^+$ and $CD8^+$ T cells, natural killer cells, and macrophages expressing CCR5. Infection of mice engineered to knock out CCR5 expression resulted in increased viral burden and was rapidly and uniformly fatal. This indicated that CCR5 is a critical antiviral and survival determinant in WNV infection of mice that acts by regulating trafficking of leukocytes to the infected brain. In addition, an increased frequency of homozygosity for CCR5Δ32, a defective CCR5 allele found predominantly in Caucasians, has been described in two independent cohorts of patients with laboratory-confirmed, symptomatic WNV infection (4% and 8% vs. 1% in a healthy control population), suggesting that CCR5 also mediates resistance to symptomatic WNV infection in humans. Thus, the same chemokine receptor can protect from fatal encephalitis or mediate infection, depending on the virus. Following WNV infection of the CNS, inflammatory chemokines are expressed by trafficking leukocytes and resident cells (microglia and astrocytes), and chemokine-dependent T cell recruitment to infected CNS tissues is likely to modulate viral pathogenesis. WNV infection also seems to induce neuronal production of CXCL10, which will recruit effector $CD8^+$ T cells through its cognate receptor, CXCR3. Also, a genetic deficiency in CXCL10 has been shown to correlate with reduced T cell trafficking to the CNS, greater viral loads, and enhanced mortality. Although CCR5 is a logical target for drug development in HIV/AIDS, blocking CCR5 could result in an increased risk of WNV disease for these patients.

HIV-1 Neuropathogenesis

Despite highly active antiretroviral therapy, HIV-associated dementia and especially milder psychomotor and cognitive disorders remain an important complication of HIV-1 infection in the developed world and even more so in the developing world, where access to therapy is limited. One potential explanation for the increased prevalence of mild to moderate HIV-associated neurological disorders in developed countries is that low-level viral replication, consistent with all but the most successful antiretroviral regimens, together with the longer life span of treated patients and the insufficient penetration of at least some of the currently used antiretroviral drugs into the brain, leads to slow, progressive neurodegeneration. Absent direct infection of neurons, microglia/brain macrophages and astrocytes will be the key cell types involved in HIV-1-associated neurodegeneration. Two non-mutually exclusive, probably coexisting mechanisms have been proposed (Figure 2):

1. The direct injury mechanism suggests that the production of viral proteins such as gp120, Tat, or Vpr from infected microglia/perivascular macrophages and perhaps astrocytes can cause disruption of normal neurological functions.
2. The indirect or bystander mechanism proposes that neurons are damaged mainly as a consequence of the inflammatory processes triggered in the brain by infected and activated cells, and it has been suggested that once the inflammatory process has taken hold, it may be a self-sustaining chain reaction leading to neurodegeneration.

Direct Injury Mechanism

In vitro experiments have shown that the HIV-1 surface glycoprotein gp120 seems to interact with chemokine receptors in neurons leading to neuronal injury. However, this interaction likely requires previous CD4 engagement, at least for infection, and CD4 is absent in neurons, putting into question its *in vivo* relevance. Alternative scenarios have been proposed since gp120 may induce neuronal cell death through direct interaction with the *N*-methyl-D-aspartate receptor or by further activation/stimulation of macrophages/microglia leading to increased TNF-α production and triggering of caspase-mediated cell death in neurons.

Two other viral proteins, Tat and Vpr, have also been suggested as potential effectors in HIV neuropathogenesis. Tat, a transcriptional transactivator, if secreted from infected cells (as shown *in vitro*) could alter tight junction protein expression and BBB function, upregulate expression of inflammatory mediators such as CCL2/MCP-1 promoting monocyte infiltration, and result in neurotoxicity by multiple intracellular signaling pathways. Viral protein Vpr is involved in cell cycle arrest, transcription, and

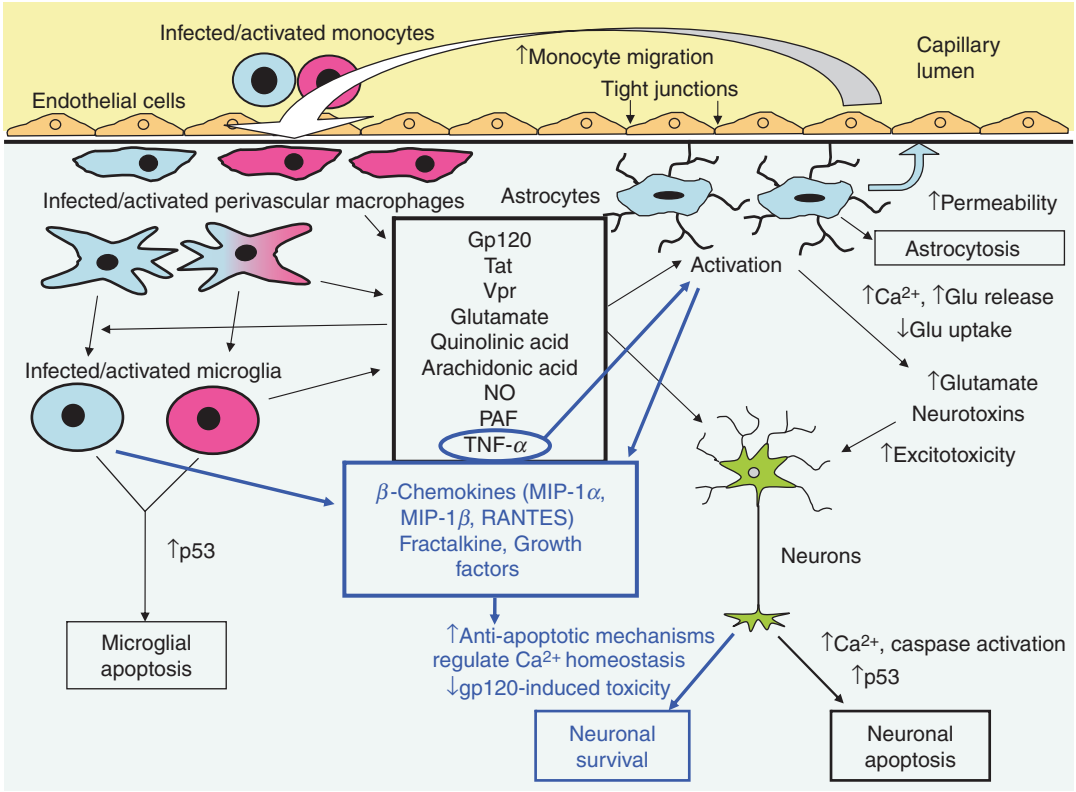

Figure 2 Potential mechanisms for HIV-1 neuropathogenesis and neuroprotection.

integration, and it has been shown to activate replication in latently infected cells and to induce caspase-dependent apoptosis of bystander cells such as neurons.

Bystander Effects: Toxicity of Glial Products

HIV encephalitis is associated with immune activation that is seemingly out of proportion to the amount of virus present, due to the alteration of the secretory functions of microglia and brain macrophages. Glutamate, cytokines such as TNF-α and IL-1β, chemokines and other soluble products, macrophage activation markers, quinolinic and arachidonic acids and related metabolites, as well as free radicals, have all been related with astrocytic proliferation and apoptosis and with the activation of uninfected cells, probably contributing to the amplification of HIV-induced neurotoxicity. In addition, some of these macrophage products modify the permeability of the BBB and promote further monocyte migration into the brain.

Glutamate, the major excitatory neurotransmitter, and other excitatory molecules are thought to be involved in HIV-induced neurotoxicity. Increases in extracellular glutamate concentration mediate uncontrolled Ca^{2+} influx and loss of cellular homeostasis, leading to neuronal death through excitotoxicity in a mechanism common to many neurodegenerative disorders. The clearance of extracellular glutamate involves high-affinity uptake by several excitatory amino acid transporters (EAATs), which are mainly expressed in astrocytes which therefore play a key role in neuroprotection against glutamate excitotoxicity. However, in the course of HIV infection, expression of certain astrocytic EAATs is reduced, resulting in a greatly decreased glutamate uptake by astrocytes. On the other hand, expression of certain EAATs has been described in human macrophages and macrophages and microglia in rat models and in simian and human AIDS, suggesting that they may also exhibit neuroprotective properties. This expression seems to be modulated by macrophage/microglia activation and infection, but it suggests that macrophages/microglia can have a protective role against glutamate excitotoxicity during HIV infection, contributing to the maintenance of homeostasis and the prevention of neuronal degeneration.

TNF-α seems to be clearly involved in neuropathogenesis, primarily by promoting additional production of inflammatory cytokines by macrophages and astrocytes, which results in amplification of glutamate release and decrease of glutamate uptake mainly by astrocytes. However, *in vitro* studies have shown that TNF-α can be neurotrophic as well. It has been reported that TNF-α has a neuroprotective role by preventing or attenuating Ca^{2+} accumulation in the

cytosol of neurons as well as through the activation of the antiapoptotic kinase Akt (a major component of pro-survival signaling pathways) and NF-κB signaling (which promotes expression of genes vital for the response to insults).

Finally, β chemokines and CX_3CL1/Fractalkine have also been shown to protect neurons from gp120 toxicity upon receptor interaction, again through Akt and NF-κB, suggesting that neuronal chemokine receptors mediate the neurotrophic effects of β chemokines and Fractalkine. Since TNF-α induces production of β chemokines by activated glial cells and of Fractalkine by astrocytes, it is likely that these cytokines and chemokines are involved in an extremely complex network of both paracrine and autocrine interactions between neurons and glial cells, with outcomes and functional consequences that are difficult to predict.

See also: Astrocyte: Identification Methods; Axon Guidance by Glia; Glia and Stroke; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Cells: Microglia During Normal Brain Aging; Glial Responses to Injury; Inflammation in Neurodegenerative Disease and Injury; Microglia identification Methods; Microglia Properties; Microglial Response to Injury; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines.

Glial Steroid Metabolism

R C Melcangi, University of Milan, Milan, Italy
L M Garcia-Segura, Instituto Cajal, CSIC, Madrid, Spain

© 2009 Elsevier Ltd. All rights reserved.

Concepts of Steroid Hormones, Neurosteroids, and Neuroactive Steroids

The finding that the nervous system is a target for steroid hormones is a well-ascertained concept. Indeed, several observations have shown that steroid hormones may regulate the development and function of the nervous system and affect mood, behavior, and cognition. It has been construed for a long time that steroids exert their actions in the nervous system exclusively by a classical endocrine mechanism involving steroid production by endocrine glands; steroids secreted by the adrenals and gonads, for example, enter the blood stream, cross the blood–brain barrier, and then regulate the central nervous system (CNS) and peripheral nervous system (PNS) in various ways. However, further findings have suggested that the effects of steroids in the nervous system are also due to paracrine and/or autocrine mechanisms. This concept is based on observations obtained by Baulieu and co-workers in 1981 indicating that the synthesis of steroids is not only a peculiarity of peripheral glands but may also occur in the CNS (i.e., the formation of the so-called neurosteroids). More recent observations by the same group indicate that also PNS synthesizes neurosteroids. On the other hand, because in many circumstances it is difficult to discriminate whether the steroid effect is due to *in situ* synthesis (i.e., neurosteroids), to the peripheral hormones (i.e., steroid hormones), or to an enzymatic conversion of steroids in metabolites which are more active and in some cases utilize a different mechanism of action, some investigators in this field use now the term 'neuroactive' steroids.

Steroid Synthesis and Metabolism by Glial Cells

Glial cells play a central role in the synthesis, metabolism, and action of steroids in the nervous system. It has been demonstrated that in the rodent and human CNS, astrocytes and oligodendrocytes express several enzymes involved in steroidogenesis (Figure 1), such as, cytochrome P450 side-chain cleavage (P450scc), 17α-hydroxylase/C17,20-lyase (P450c17), 3β-hydroxysteroid dehydrogenase (3β-HSD), and 17β-hydroxylase/C17,20-lyase (17β-HSD), and consequently these types of glial cells are able to produce different kinds of steroids. In particular, astrocytes appear to be the most active steroidogenic cells in the brain, since cultures of these cells produce pregnenolone, progesterone, dehydroepiandrosterone, androstenedione, testosterone, and estradiol. On the contrary, oligodendrocytes seem to be able to form only pregnenolone and androstenedione. The steroidogenic activity of microglia has not been assessed. However, systemic macrophages that share many properties and a common origin with microglia express aromatase and produce androgenic and estrogenic derivatives.

The capability to synthesize steroids seems not to be a peculiarity of the CNS, since the PNS also expresses cytochrome P450scc (i.e., the enzyme converting cholesterol to pregnenolone), and 3β-HSD, which converts pregnenolone into progesterone. Interestingly, both P450scc and 3β-HSD are also present in dorsal root ganglia (DRG) neurons. Indeed, P450scc-immunoreactive cells are localized in DRG neurons, and incubation of DRG tissue homogenates with labeled cholesterol has yielded data indicating formation of different steroids, including pregnenolone. Moreover, *in situ* hybridization of small and large sensory neurons from thoracic and lumbar DRG of adult male rats shows expression of 3β-HSD. Furthermore, in co-culture conditions, sensory neurons and Schwann cells are able to stimulate each other via diffusible molecules, representing the expression and activity of enzyme 3β-HSD. The capability to produce steroids is also present in human PNS. Indeed, very high levels of pregnenolone have been measured in human sciatic nerves.

Steroidogenic Acute Regulatory Protein, Peripheral Benzodiazepine Receptor, and Mitochondrial Transformation of Cholesterol in Pregnenolone

The synthesis of pregnenolone from cholesterol by P450scc is the first enzymatic step of steroidogenesis. However, there is a previous step that is rate limiting and hormonally regulated: the transfer of cholesterol from the outer to the inner mitochondrial membrane, where the P450scc enzyme is located. In this mechanism, two proteins located in the mitochondrial membranes, such as the peripheral benzodiazepine receptor (PBR) and the steroidogenic acute regulatory protein (StAR), take an active role (Figure 2). PBR was initially described on peripheral tissues as a second binding site for diazepam, which binds with higher affinity to γ-aminobutyric acid ($GABA_A$)

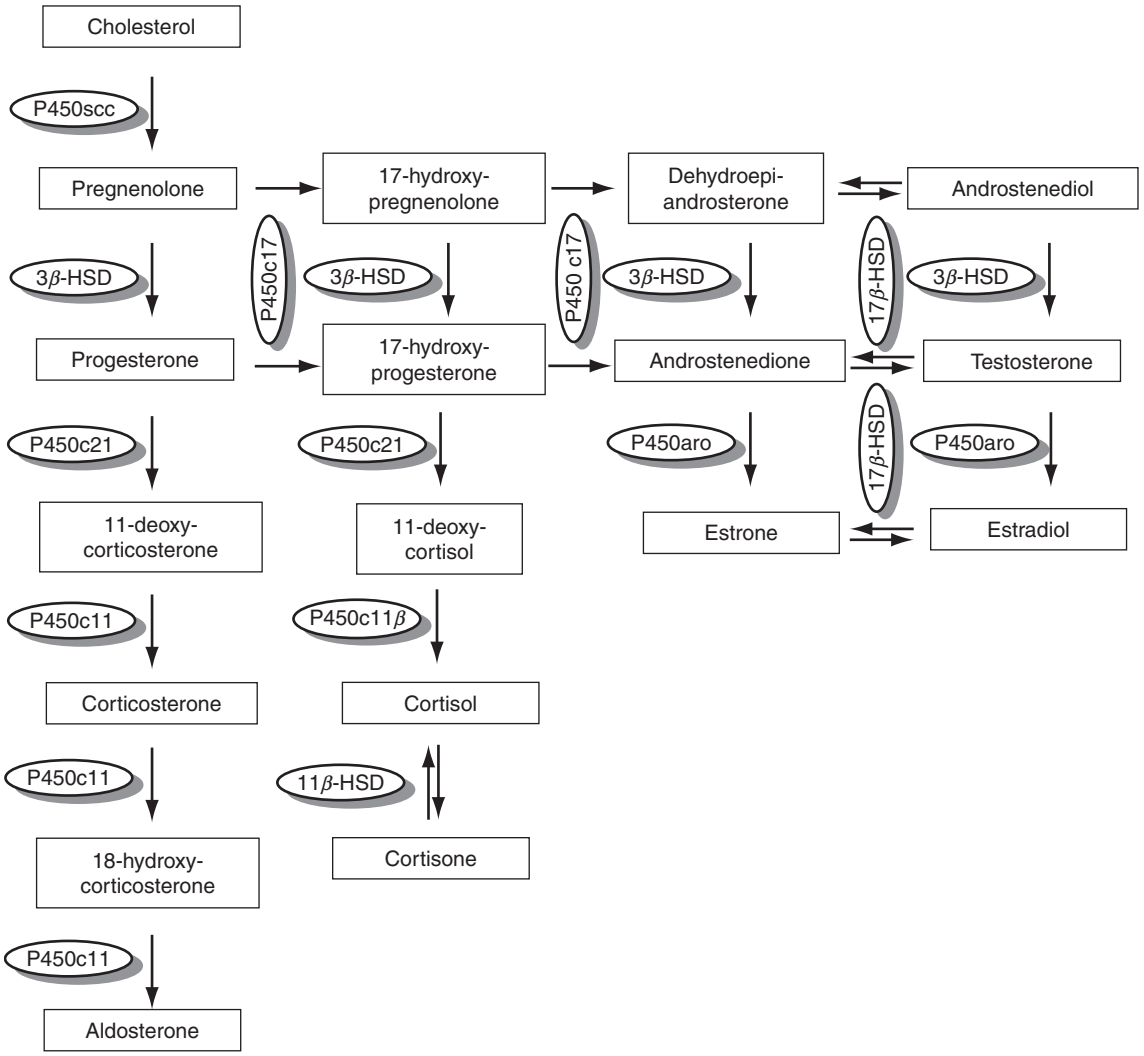

Figure 1 Biosynthesis of steroids in glial cells. Cholesterol is converted to pregnenolone, catalyzed by cytochrome P450 side-chain cleavage (P450scc), and pregnenolone is converted to progesterone, catalyzed by 3β-hydroxysteroid dehydrogenase (3β-HSD). Pregnenolone and progesterone can be further converted, via 17-hydroxypregnenolone and 17-hydroxyprogesterone, respectively, to dehydroepiandrosterone and androstenedione, catalyzed by 17α-hydroxylase/C17,20-lyase (P450c17), and from there to androstenediol and testosterone, respectively, catalyzed by 17β-hydroxylase/C17,20-lyase (17β-HSD). Androstenedione and testosterone can be converted to estrone and estradiol, respectively, catalyzed by aromatase (P450aro). In addition to these pathways, which produce androgens and estrogens, progesterone and its derivative 17-hydroxyprogesterone can be converted to the glucocorticoids corticosterone and cortisol, respectively. Cortisol can be further converted to the glucocorticoid cortisone, catalyzed by 11β-hydroxysteroid dehydrogenase (11β-HSD), and corticosterone, via 18-hydroxycorticosterone, can be converted to the mineralocorticoid aldosterone.

receptors on the nervous system. Further studies have then demonstrated that the PBR is pharmacologically and structurally distinct from the central benzodiazepine/$GABA_A$ receptors. The PBR is an 18 kDa peptide located predominantly in the mitochondrial membranes and represents a critical component of the permeability transition pore, a multiprotein complex implicated in the regulation of apoptosis. In addition, PBR has been related with the regulation of several physiological events, including the control of steroidogenesis. PBR is expressed in the nervous system and predominantly in glial cells of the CNS and PNS. Neural injury and different neuropathological conditions result in the induction of the expression of this molecule. Indeed, PBR basal expression is upregulated in gliomas, in neurodegenerative disorders (e.g., Alzheimer's disease), as well as in various forms of brain injury and inflammation induced by neurotoxins. The induction of PBR expression after injury in the CNS is mainly restricted to microglia and astrocytes. A very similar effect occurs in the PNS. Indeed, the expression of the PBR in Schwann cells increases after nerve lesion and returns to normal levels when regeneration is completed.

StAR is formed as a 37 kDa protein; it is rapidly transported into mitochondria, where it is cleaved, generating a mature 30 kDa intramitochondrial StAR protein that is inactive. This molecule appears to be widely distributed throughout the brain, although different levels of expression have been detected between different brain areas. Moreover, StAR expression seems to be restricted to very specific neuronal and astroglial populations in each brain area, although it is predominantly expressed by neurons. In the PNS, StAR is expressed by Schwann cells. Therefore, both neurons and glial cells may participate in the conversion of cholesterol into pregnenolone.

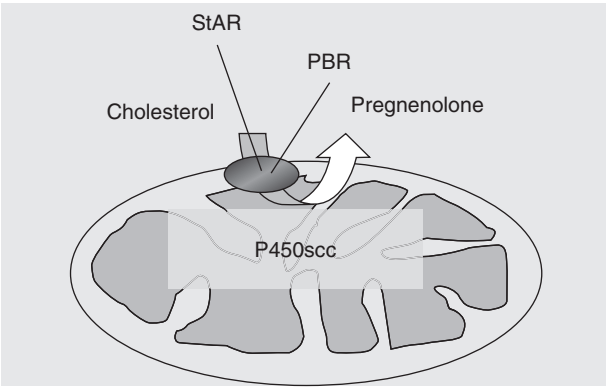

Figure 2 The synthesis of pregnenolone from cholesterol by the cholesterol side-chain cleavage enzyme (P450scc) is the first enzymatic step of steroidogenesis. However, there is a previous step that is rate limiting and hormonally regulated: the transfer of cholesterol from the outer to the inner mitochondrial membrane, where the P450scc enzyme is located. Proteins located in the mitochondrial membranes, such as the peripheral benzodiazepine receptor (PBR) and the steroidogenic acute regulatory protein (StAR), allow cholesterol to cross the hydrophilic intermembrane space. Both PBR and StAR expression have been detected in glial cells.

Synthesis and Metabolism of Progesterone

As mentioned previously, glial cells of the CNS and PNS may convert pregnenolone in progesterone. Moreover, these cells also possess several enzymes able to transform native steroids into neuroactive steroids. In particular, glial cells are able to metabolize native steroid hormones into their 5α- and 3α-hydroxy-5α-reduced derivatives via the enzymatic complex formed by the 5α-reductase (5α-R) and the 3α-hydroxysteroid dehydrogenase (3α-HSD) (Figure 3). This enzymatic complex is very versatile, since every steroid possessing the Δ^4-3-keto configuration may be first 5α-reduced and subsequently 3α-hydroxylated. Progesterone is a substrate for this enzymatic complex. Namely, in the CNS, astrocytes and oligodendrocytes convert this steroid, by the action of 5α-R, to dihydroprogesterone (DHP) and subsequently, via the action of 3α-HSD, to tetrahydroprogesterone (THP), also known as allopregnanolone. However, among glial cells the conversion of DHP to THP seems to be different. Indeed, observations obtained in culture have indicated that this enzymatic conversion is very high in type 1 astrocytes (which probably correspond *in vivo* to protoplasmatic astrocytes) in comparison to what observed in type 2 astrocytes (which probably correspond *in vivo* to fibrous astrocytes) and in oligodendrocytes. However, the conversion of DHP to THP in oligodendrocytes is significantly higher than that in type 2 astrocytes.

These enzymatic conversions may have an important impact on the mechanism of action of progesterone. In fact, DHP, like progesterone, acts on a classical steroid receptor, such as the progesterone receptor, whereas THP acts on a nonclassical steroid receptor, like the $GABA_A$ receptor. Progesterone

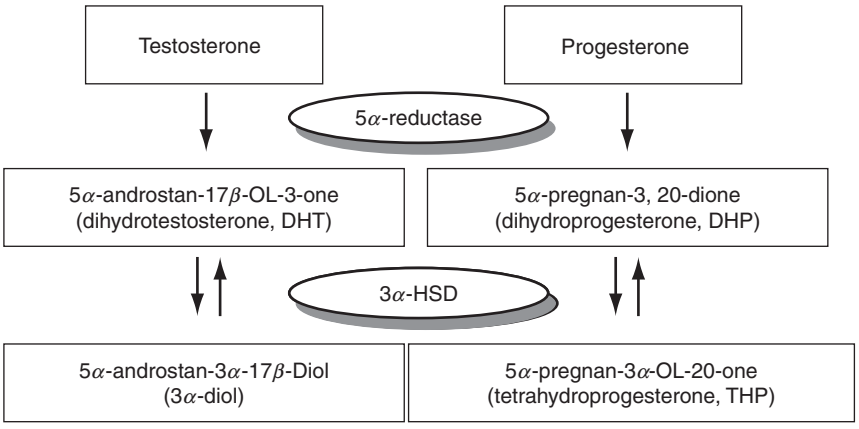

Figure 3 Metabolism of testosterone and progesterone by the action of the enzymes 5α-reductase and 3α-hydroxysteroid dehydrogenase (3α-HSD).

metabolism by glial cells is developmentally regulated. Namely, the formation of DHP in fully differentiated oligodendrocytes is fivefold higher than in oligodendrocyte pre-progenitors and in oligodendrocyte progenitors. On the contrary, the formation of THP is higher in oligodendrocyte pre-progenitors and decreases with oligodendrocyte differentiation. These findings emphasize that not only differentiated CNS cells possess the 5α-R/3α-HSD system, since considerable enzymatic activities converting progesterone are also present in undifferentiated cells, as shown in studies performed on stem cells originating from the mouse striatum.

The ability to convert steroid hormones into 5α- and 3α-hydroxy-5α-reduced derivatives is also present in myelin-forming cells of PNS (i.e., Schwann cells). However, the extent of progesterone metabolism seems to be different with respect to what is observed in the cellular counterpart in the CNS. Therefore, the conversion of progesterone to DHP as occurs in Schwann cells is at least 4 times higher than that in oligodendrocytes, whereas conversion of DHP to THP is lower than that in oligodendrocytes.

Metabolism of Androgens

Testosterone is a substrate for the action of 5α-R and 3α-HSD enzymes. Similar to the aforementioned case for progesterone, testosterone can be converted to dihydrotestosterone (DHT) by the enzyme 5α-R and then to 5α-androstane-3α,17β-diol (3α-diol; also known as androstanediol) by the enzyme 3α-HSD (Figure 3). Both enzymatic steps occur in glial cells. However, again, similar to what was reported for progesterone, these enzymatic steps are different, depending on the types of glial cells considered. For instance, the conversion of testosterone to DHT in type 2 astrocytes and in oligodendrocytes occurs at statistically higher levels than occurs in type 1 astrocytes. Moreover, the conversion of DHT into 3α-diol appears to be mainly localized in type 1 astrocytes. Also, in this case, enzymatic conversions of DHT to its reduced metabolites may have an important impact on the mechanism of action of the steroid. While DHT, like its precursor testosterone, is able to bind the classical steroid receptor for androgens, 3α-diol is able to bind to a nonclassical steroid receptor (e.g., the $GABA_A$ receptor).

Testosterone metabolism by glial cells is also affected by cellular interactions, since the presence of neurons stimulates the activity of 5α-R and 3α-HSD enzymes in astrocytes. The nature of these neuronal signals has been investigated. In mixed glial cultures the formation of DHT is increased by exposure to cAMP analogs (e.g., 8-Br-cAMP). In agreement with this finding, in cultures of glioma cells the stimulation of the β-adrenergic receptor induces activation of the cAMP/protein kinase A-mediated signaling pathway and a rapid elevation of 5α-R mRNA levels. The other alternative metabolic pathway of testosterone is its conversion to estradiol, as discussed in the following section.

Synthesis of Estradiol

The enzyme aromatase (P450aro), which is able to convert testosterone into estradiol, is not expressed by glial cells of the CNS of rodents under normal circumstances; only neurons typically possess such activity. However, rodent astrocytic cultures from the cerebral cortex of neonatal rats are able to produce estradiol and estrone. It is possible that specific culture conditions may induce aromatase expression in astrocytes. Indeed, stressful conditions, such as serum deprivation, induce aromatase expression in cultured rat astrocytes. Furthermore, aromatase is expressed by astrocytes in the brain of birds and mammals after brain injury, indicating that the enzyme may be induced *de novo* in these cells under specific circumstances. In contrast to rodents, aromatase is expressed in radial glia in the brain of teleost fish under normal conditions and also in a subpopulation of astrocytes in the human brain. Aromatase expression in glia seems to be regulated by steroid levels. A first demonstration in testosterone- and estradiol-treated females of a species of vocal (generation of acoustic signals) fish, the plainfin midshipman, shows in the preoptic area and dimorphic sonic/vocal motor nucleus an increase of aromatase mRNA levels.

Functional Implications of Glial Steroidogenesis

Glial cells are not only able to synthesize neurosteroids and to metabolize these or steroid hormones into neuroactive steroids, but these cells are also targets for their own actions. This is not surprising, because it has been clearly demonstrated that central (oligodendrocytes and astrocytes) and peripheral (Schwann cells) glia possess classical intracellular receptors for many families of hormonal steroids (e.g., receptors for glucocorticoids, mineralocorticoids, androgens, estrogens, and progesterone). Moreover, central and peripheral glial cells also express nonclassical steroid receptors – for instance, γ-aminobutyric acid type A and type B receptors ($GABA_A$ receptor, $GABA_B$ receptor), *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), and an atypical intracellular receptor like the sigma-1

protein, which have also been demonstrated to be a target for neuroactive steroid actions.

Schwann cells, oligodendrocytes, astroglia, and microglia are all affected by neuroactive steroids. Namely, these molecules affect proliferation, gene expression, differentiation, metabolism, and morphology of glial cells. For instance, estradiol and progesterone enhance the proliferation of Schwann cells and of oligodendrocyte precursors, but decrease in females the astrocyte proliferation that occurs following brain injury in the cerebral cortex and the hippocampus. Estradiol also decreases proliferation and increases cell death in primary cortical astrocytic cultures, while glucocorticoids decrease oligodendrocyte precursor proliferation.

Examples of effects on gene expression are those exerted by progesterone and derivatives (DHP and THP) or testosterone and derivatives (DHT and 3α-diol) on the expression of myelin proteins in Schwann cells and oligodendrocytes. Moreover, these neuroactive steroids also modulate the expression of transcription factors regulating expression of myelin proteins in Schwann cells.

Steroid effects on glial gene expression and morphology have also been extensively studied in astroglia. Corticosterone and dexamethasone inhibit glial fibrillary acidic protein (GFAP) expression in neonatal and adult rat brain. Pregnenolone, dehydroepiandrosterone, and sex steroids affect astroglia cell shape and regulate GFAP in different brain areas. However, it is important to highlight that the effects of steroids may be different, depending on whether they are acting under physiological or pathological conditions. For instance, estradiol increases GFAP expression and promotes extension of GFAP-immunoreactive processes under physiological conditions, while the same steroid decreases GFAP and vimentin expression in gliotic injured tissue.

Examples of other molecules affected by steroids in astrocytes and microglia are apolipoprotein E, heat shock proteins, metalloproteinase-9, aquaporin-4, glutamate transporters GLUT-1 and GLAST, and a variety of growth factors and cytokines, such as transforming growth factor-α, transforming growth factor-β1, transforming growth factor-β2, basic fibroblast growth factor, insulin-like growth factor 1, tumor necrosis factor-α, and interleukin-12. Furthermore, steroids may also affect neuronal function by the modification of glial metabolism. For instance, glucocorticoids are able to influence the metabolism of glutamate in astrocytes and Schwann cells, increasing expression of both glutamine synthetase and glutamate dehydrogenase in astrocytes and expression of glutamine synthetase and cytosolic aspartate aminotransferase in Schwann cells.

Altogether, these observations indicate that glial steroidogenesis has important functional implications, including regulation of synaptic function and connectivity and neuroendocrine regulation. It has been demonstrated that sex steroids affect synthesis, accumulation, and release by hypothalamic astrocytes and tanycytes of different growth factors that regulate neurons synthesizing and secreting the luteinizing hormone-releasing hormone (i.e., the hormone deputed to the neuroendocrine control of reproductive functions). Moreover, another important functional implication is based on the effects exerted by neuroactive steroids on myelin formation and neuronal survival. These promising effects have recently raised the hypothesis that neuroactive steroids may be considered not only as key regulators of nervous function, but also as protective agents. Examples supporting this hypothesis have been obtained in an experimental model of Niemann–Pick type C disease (a fatal, autosomal recessive childhood neurogenerative disease), Alzheimer's disease, Parkinson's disease, ischemic or spinal cord injury, and acquired (after trauma, during aging, etc.) or inherited peripheral neuropathy, such as Charcot–Marie–Tooth disease.

A possible therapeutic strategy for neurodegenerative events might be treatment with neuroactive steroids, or with synthetic ligands for their classical or nonclassical steroid receptors. However, taking into consideration the *in situ* formation and conversion of steroids at the levels of the glial compartments might be extremely important. In fact, as demonstrated in several experimental models, a local increase in the levels of steroids, or activation of steroidogenic enzymes, may be observed during neurodegenerative events. These findings have suggested that an increase in steroidogenesis is part of an overall mechanism used by the nervous tissue to cope with neurodegenerative conditions. Consequently, modulation of steroidogenesis or of enzymes able to increase neuroactive derivatives (e.g., 5α-R, 3α-HSD, aromatase) at the level of glial compartments might be also considered as an interesting therapeutic strategy. Recent observations have indicated that ligands of PBR – which, as already mentioned, is involved in the first step of steroidogenesis – exert protective effects both in the CNS and PNS, and that *in situ* inhibition of aromatase activity increases neurodegeneration.

Summary

The observations here reported indicate that glial compartments (i.e., astrocytes, oligodendrocytes, Schwann cells, and microglia) play a fundamental role in the production of steroids in the CNS and PNS. Steroids so formed, or activated in glial

compartments, may then act by autocrine pathway on the same cells, or by paracrine pathway on different glial types or on neuronal compartments. Indeed, classical and nonclassical steroid receptors have been reported in both neuronal and glial cells. Paracrine and autocrine effects are then integrated with classical endocrine mechanisms exerted by steroid hormones. Thus, the nervous system is a target for two different pools of steroids, one coming from the peripheral glands and the second one originating directly in the nervous system. The finding that neuroactive steroids may be considered as signaling molecules in glia–neuron and glia–glia cross-talk, together with the effects exerted by these molecules in physiological and pathological conditions, strengthens the concept that they are important key regulators of nervous function.

See also: Astrocyte: Neurotransmitter and Hormone Receptors; Estrus and Menstrual Cycles: Neuroendocrine Control; Gene Expression Regulation: Steroid Hormone Effects; Glial Cells: Microglia During Normal Brain Aging; Gonadotropin, Neural and Hormonal Control; Gonadotropin-Releasing Hormone: GnRH-1 System; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Neurosteroids; Radial Glial Cells: Brain Functions.

Further Reading

Belelli D and Lambert JJ (2005) Neurosteroids: Endogenous regulators of the GABA(A) receptor. *Nature Reviews Neuroscience* 6: 565–575.

Garcia-Ovejero D, Azcoitia I, DonCarlos LL, et al. (2005) Glia–neuron crosstalk in the neuroprotective mechanisms of sex steroid hormones. *Brain Research Reviews* 48: 273–286.

Garcia-Segura LM and McCarthy MM (2004) Minireview: Role of glia in neuroendocrine function. *Endocrinology* 145: 1082–1086.

Ibanez C, Shields SA, El-Etr M, et al. (2003) Steroids and the reversal of age-associated changes in myelination and remyelination. *Progress in Neurobiology* 71: 49–56.

Jordan CL (1999) Glia as mediators of steroid hormone action on the nervous system: An overview. *Journal of Neurobiology* 40: 434–445.

Melcangi RC and Panzica GC (2001) Steroids in the nervous system: A Pandora box? *Trends in Neurosciences* 24: 311–312.

Melcangi RC and Panzica GC (2006) Neuroactive steroids: Old players in a new game. *Neuroscience* 138: 733–739.

Melcangi RC, Martini L, and Galbiati M (2002) Growth factors and steroid hormones: A complex interplay in the hypothalamic control of reproductive functions. *Progress in Neurobiology* 67: 421–449.

Melcangi RC, Cavarretta IT, Ballabio M, et al. (2005) Peripheral nerves: A target for the action of neuroactive steroids. *Brain Research Reviews* 48: 328–338.

Mellon SH, Griffin LD, and Compagnone NA (2001) Biosynthesis and action of neurosteroids. *Brain Research Reviews* 37: 3–12.

Ojeda SR, Lomniczi A, Mastronardi C, et al. (2006) Minireview: The neuroendocrine regulation of puberty: Is the time ripe for a systems biology approach? *Endocrinology* 147: 1166–1174.

Papadopoulos V, Lecanu L, Brown RC, et al. (2006) Peripheral-type benzodiazepine receptor in neurosteroid biosynthesis, neuropathology and neurological disorders. *Neuroscience* 138: 749–756.

Rupprecht R (2003) Neuroactive steroids: Mechanisms of action and neuropsychopharmacological properties. *Psychoneuroendocrinology* 28: 139–168.

Schumacher M, Guennoun R, Mercier G, et al. (2001) Progesterone synthesis and myelin formation in peripheral nerves. *Brain Research Reviews* 37: 343–359.

Vallee M, Mayo W, and LeMoal M (2001) Role of pregnenolone, dehydroepiandrosterone and their sulfate esters on learning and memory in cognitive aging. *Brain Research Reviews* 37: 301–312.

Glioma

N J Ernest and H Sontheimer, University of Alabama at Birmingham, Birmingham, AL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Primary brain tumors are abnormal tissue expansions that result from the uncontrolled proliferation of brain cells. According to the Central Brain Tumor Registry of the United States, approximately 14 per 100 000 people in the United States (i.e., 40 000 patients) will be diagnosed with a primary brain tumor each year. On average, half of these tumors are benign and respond well to treatment. The other half, however, are malignant brain tumors with a poorer prognosis. The vast majority of these (78%) are called gliomas because they appear to be derived from cells of glial origin. These include primarily oligodendrocytes and astrocytes but also the ependymal cells lining the ventricles. Gliomas account for less than 2% of adult cancers. However, despite their rarity, gliomas are the leading cause of cancer-related deaths in men 20–39 years of age and the second most common cause of cancer-related deaths in children. Furthermore, they represent the most common solid tumor in children. The incidence rate for gliomas is highest in patients older than 55 years, for whom the prognosis is also poorest. Patients diagnosed with gliomas typically present with signs of increased intracranial pressure (e.g., headache, vomiting, and papilloedema) and local or generalized brain dysfunction (e.g., seizures and alterations in consciousness). Although a few patients may develop a tumor as a result of an uncommon genetic disorder such as neurofibromatosis types I and II, Li–Fraumeni syndrome, or tuberous sclerosis, or as the result of exposure to ionizing radiation at a young age, for most patients a risk factor for developing a glioma cannot be identified.

Classification

The term 'glioma' refers to a collection of tumors that can be stratified into four well-defined tumor grades based on histopathological criteria. These have been developed by the World Health Organization (WHO) and are now commonly adopted by scientists and physicians. According to this classification system, gliomas are graded on a scale of I to IV, with higher numbers representing more malignant tumors. The grade is based on the presence or absence of the following features: atypical morphology of cells, nuclear atypia, mitotic activity, the presence of necrotic tissue, and the presence of newly formed blood vessels (Table 1). At least three of these features characterize the highest grade IV gliomas, typically necrosis, neovascularization, and enhanced mitosis. The two features that characterize grade III gliomas are most typically mitosis and nuclear atypia. Low-grade, I and II gliomas do not show mitosis but present primarily with cellular and nuclear atypia. The former may be the only feature present in grade I gliomas. Histopathological tumor grading requires a tissue biopsy, ideally taken from the most atypical region of a patient's tumor. As the disease progresses, it is not uncommon for tumor cells to transform into more malignant cells over time. Indeed, many grade II gliomas progress toward a grade IV cancer over a period of years. Tumor grading is not merely academic but instructive with regard to treatment strategies. Low-grade gliomas (I and II), which are less common and more typically diagnosed at a younger age, carry a more favorable prognosis than high-grade tumors. Indeed, patients diagnosed with a grade I astrocytoma have a 5-year survival rate of greater than 90%, and many may not need to receive treatment at all. By contrast, patients diagnosed with a grade IV astrocytoma have a survival rate of less than 4% despite aggressive chemo and radiation therapy. Although lower grade tumors typically have a lower mitotic activity than high-grade tumors, both low- and high-grade gliomas can be highly infiltrative and therefore are rarely surgically curable.

Ependymomas

Tumors deriving from the linings of the ependymal cells lining the ventricle are called ependymomas and also classified as gliomas. Although these tumors are the most common gliomas encountered in the spinal cord and filum terminale, they represent only 5 or 6% of all gliomas and occur predominantly in children and adolescents. Histopathologically, ependymomas often contain tapered, elongated cells (sometimes referred to as 'carrot' or 'ice cream cone' cells) that are arranged around miniature ventricles (forming 'rosettes') or around blood vessels (forming 'pseudorosettes') (Figure 1). Ependymomas tend to carry a more favorable prognosis, with 10-year survival rates greater than 60%.

Oligodendrogliomas

Oligodendrogliomas account for 9 or 10% of all gliomas. They are characterized by an absence of markers associated with astrocyte differentiation and a uniform population of cells that contain round nuclei and very little cytoplasm (the classic 'fried egg' or

Table 1 Characteristics of WHO grade I–IV gliomes

WHO grade	*Characteristics*	*Examples*
I	Low mitotic activity; little nuclear atypia	Pilocytic astrocytoma; demoplastic infantile astrocytoma
II	Low mitotic activity; nuclear atypia; may be infiltrative; may recur	Diffuse astrocytoma (fibrillary, gemistocytic, or protoplasmic); xanthoastrocytoma
III	High mitotic activity; infiltrative; cellular and nuclear atypia; often recur	Anaplastic astrocytoma; anaplastic oligodendroglioma
IV	High mitotic activity; high degree of infiltration; cellular and nuclear atypia; necrosis; neovascularization	Glioblastoma multiforme

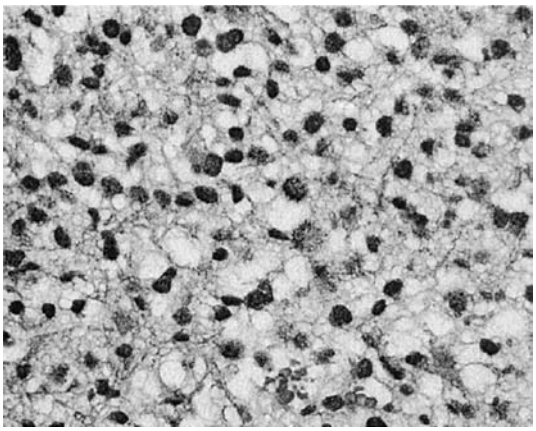

Figure 2 Cells of an anaplastic oligodendroglioma demonstrate the classic 'fried egg' appearance: round nuclei with clear, perinuclear halos. Reproduced from Engelhard H, Stelea A, and Cochran E (2002) Oligodendroglioma: Pathology and molecular biology. *Surgical Neurology* 58: 111–117, with permission.

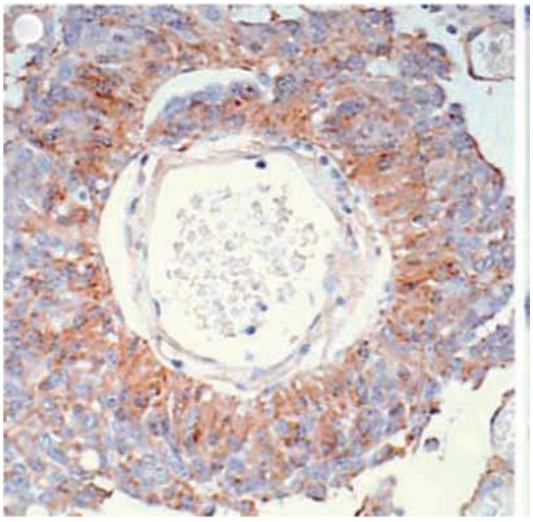

Figure 1 Ependymoma cells forming a perivascular pseudorosette demonstrate GFAP immunoreactivity. Reproduced from Lehman N, Jordan M, Huhn S, et al. (2002) Cortical ependymoma: A case report and review. *Pediatric Neurosurgery* 39: 50–54, with permission.

'fish eye' appearance) (Figure 2). Occasionally, these tumors may contain branching vessels that appear to divide the tumor into pseudolobules. Areas of hemorrhage and calcification are more commonly observed in oligodendrogliomas compared to astrocytomas, which may help in making a diagnosis. It is important to distinguish oligodendrogliomas from other gliomas because they are more likely to respond to chemotherapy than other gliomas of the same grade and are less likely to progress to a higher grade tumor. One indication that an oligodendroglioma will respond to chemotherapy is the loss of the 'short arm' of chromosome 1 (1p). This genetic feature along with loss of the 'long arm' of chromosome 19 (19q) are considered to be early events in the oncogenesis of oligodendrogliomas. These deletions are found in 88–100% of tumors of oligodendroglial origin, and they are less frequently found in mixed oligoastrocytomas (tumors that demonstrate histopathological features of both oligodendrogliomas and astrocytomas). Oligodendrogliomas occur more frequently in middle-aged adults. When pediatric tumors do occur, loss of 1p and 19q is less frequently observed, suggesting that these tumors may have a different genetic origin. Oligodendrogliomas most commonly form in the cerebral hemispheres, often expanding along gyri, obscuring the white–gray matter junction.

Astrocytomas

In contrast to oligodendrogliomas, low-grade astrocytomas (WHO grades I and II) are typically found within the margins of white matter. Tumors of this type more frequently show evidence of mutations in p53, a protein responsible for inducing apoptosis and limiting progression through the cell cycle when a cell is damaged. Higher grade astrocytomas may also exhibit amplification of cyclin-dependent kinases and MDM2 (a protein responsible for inhibition of p53) as well as high expression of platelet-derived growth factor ligands and receptors. The end effect of these genetic alterations is an upregulation of survival pathways and a decrease in pro-apoptotic pathways. Low-grade astrocytomas include pilocytic or diffuse astrocytomas, pleomorphic xanthoastrocytomas, and subependymal glial cell astrocytomas. Pilocytic astrocytomas are grade I, noninfiltrating tumors that account for approximately 8% of all gliomas and

are the most common gliomas found in children. These tumors may be fully enclosed by cysts and are sometimes fully resectable. Hence, the prognosis for these tumors is very good because they are often cured with surgery alone. Tumors that contain cells resembling both oligodendrocytes and astrocytes are referred to as mixed oligoastrocytomas. These WHO grade II gliomas carry a better prognosis and longer median survival than pure astrocytomas of the same grade but a less optimistic prognosis than pure oligodendrogliomas. Pure grade II astrocytomas with minimal cellular and nuclear atypia are classified as diffuse astrocytomas. The peak incidence of these tumors occurs in people 25–50 years of age. There are several variants of diffuse astrocytomas, including fibrillary, protoplasmic, and gemistocytic (Figure 3). The cells of fibrillary astrocytomas look similar to astrocytes, with many multipolar processes that look like many small fibers under a microscope. The nuclei of these cells may be cigar shaped and are surrounded by limited cytoplasm. Fibrillary tumors are often immunoreactive for GFAP, vimentin, and S-100 (markers associated with astrocytes in various stages of differentiation). The cells of protoplasmic astrocytomas are more stellar, have smaller processes, and are less likely to be GFAP reactive. Their nuclei often appear vesicular with a prominent nucleolus. Gemistocytic astrocytomas progress more rapidly to grade III and IV astrocytomas. The cells of these tumors contain extensive eosinophilic cytoplasm and odd nuclei. Grade III astrocytomas, anaplastic astrocytomas, account for 8% of all gliomas. Compared to lower grade astrocytomas, these tumors have increased cellularity, obvious nuclear atypia, and enhanced mitotic activity. However, these tumors often do not show evidence of necrosis or microvascular proliferation. The prognosis for this grade of astrocytoma is not as optimistic as that for lower grade astrocytomas because these tumors tend to recur, often as a grade IV glioma.

Glioblastomas

Grade IV gliomas are more commonly referred to as glioblastoma or glioblastoma multiforme (GBM). These most malignant tumors account for 23% of all primary brain tumors and 50% of all gliomas. Although GBMs are more common in older adults, they do account for 8% of all childhood brain tumors. Histopathologically, these tumors are characterized by a lack of uniformity and boundaries. They may contain cysts, calcium deposits, blood vessels, and a mixture of cell types, including giant cells. Typically, these tumors contain focal areas of necrosis with a border of pseudopalisading cells (Figure 4). GBM cells are highly infiltrative and often migrate far from the main tumor site. GBMs may develop from previously diagnosed lower grade tumors. When this occurs, the tumor (sometimes referred to as a 'secondary

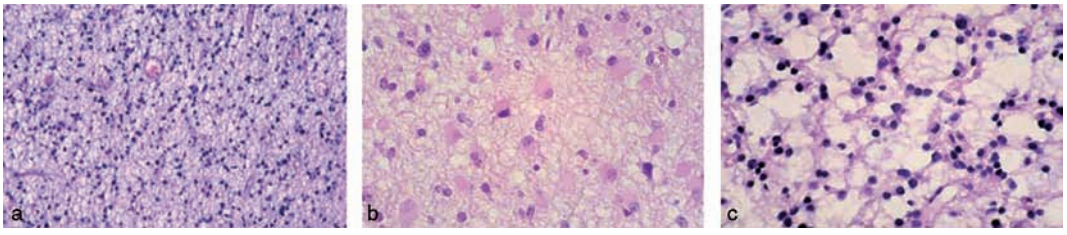

Figure 3 Typical examples of fibrillary (a), gemitocytic (b), and protoplasmic (c) astrocytomas. Reproduced from Reifenberfer G and Collins V (2004) Pathology and molecular genetics of astrocytic gliomas. *Journal of Molecular Medicine* 82: 656–670, with permission.

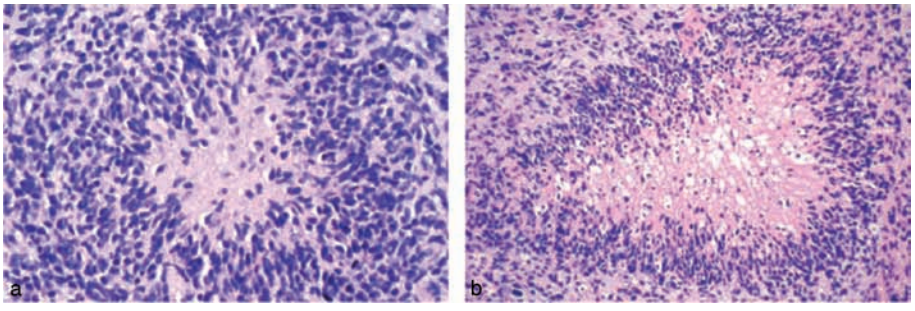

Figure 4 Examples of small (a, <100 μm) and medium (b, 200–400 μm) pseudopalisades in glioblastoma. Note the central necrosis in the larger pseudopalisade (b). Reproduced from Brat D, Castellano-Sanchez A, Hunter S, et al. (2004) Pseudopalisades in glioblastoma are hypoxic, express extracellular matrix proteases, and are formed by an actively migrating cell population. *Cancer Research* 64: 920–927, with permission.

GBM') is more likely to have evidence of mutations in p53. In addition, secondary GBMs are also frequently observed to have mutations in Rb1 (a protein important in cell cycle progression), in the PI3K/Akt pathway (a pathway that promotes survival), and in growth factor receptors. In fact, amplification of the epidermal growth factor receptor occurs in up to 35% of GBMs, although it is less common in patients younger than 35 years of age. Some mutations occur more frequently in GBMs that appear to arise *de novo* (also known as 'primary GBMs'). One example is a mutation in PTEN, a protein involved in the PI3K/Akt pathway, which appears to occur almost exclusively in primary GBMs.

The prognosis for patients diagnosed with this most common glioma is very poor. The average survival for newly diagnosed individuals is 9–12 months, with a 5-year survival rate of less than 4%.

Other Gliomas

Less common tumors account for the remaining 15% of gliomas. Examples of some less common gliomas include vestibular schwannomas and ganglioneuromas. Vestibular schwannomas are most commonly benign tumors surrounding the nerves supplying the inner ear. Bilateral vestibular schwannomas may occur in patients with neurofibromatosis type II. One of the rarest gliomas observed is a ganglioneuroma. This slow-growing tumor is actually composed of cells with both glial and neuronal characteristics.

Biology

Proliferation

Cancers only develop from cells that are intrinsically able to divide. Consequently, in the brain, neurons which are postmitotic are excluded from this process. However, nonmalignant glial cells appear to retain their ability to proliferate such that they can respond to injury by forming a glial scar that seals off the wound. Furthermore, the brain contains a population of progenitor cells that are also able to proliferate. Although these processes appear well controlled, tumor cell proliferation is uncontrolled due to the underlying genetic mutations that regulate the cell cycle. In fact, several of the genetic mutations found in human gliomas, including mutations in p53, Rb1, and the PI3K/Akt pathways, primarily affect cell cycle progression or the 'brakes' that would normally halt cycle progression (often leading to apoptosis) in mutated cells. The increased proliferative rate of glioma cells allows these cancers to rapidly expand. This causes unique problems as physical tumor expansion in the central nervous system is limited by the confines of the bony skull and spine. Although gliomas can, and do, expand into the ventricles, there are few other spaces into which the tumor mass can grow. One of the means by which gliomas overcome this problem is by destroying surrounding nervous tissue. Peritumoral neuronal necrosis is indeed a characteristic feature of more mitotic gliomas and may be the result of excitotoxicity via an increase in extracellular glutamate. Unlike nonmalignant glial cells which take up glutamate, glioma cells actually release glutamate through the system X_c exchanger, and this glutamate may trigger peritumoral seizures and ultimately the demise of neurons in the vicinity (Figure 5).

Angiogenesis

Expansion of tumor cells can also be limited by the availability of oxygen and nutrients from nearby blood vessels. Glioma cells located ten cell layers or more away from a vessel become hypoxic, which may lead to cell death and would limit the growth of the tumor. To compensate for this problem, more mitotic and atypical gliomas (grade III and IV) form new blood vessels through a process known as angiogenesis. In fact, angiogenesis is one of the criteria used by pathologists to classify grade III and IV tumors. Angiogenesis, or the formation of new blood vessels from preceding vessels, is a well-characterized process that involves increased vascular permeability, extravasation of plasma proteins, breakdown of the existing basement membrane, deposition of proangiogenic matrix molecules, and the subsequent proliferation and migration of endothelial cells. This cascade of events can be initiated by the signaling of vascular endothelial growth factors (VEGFs) through

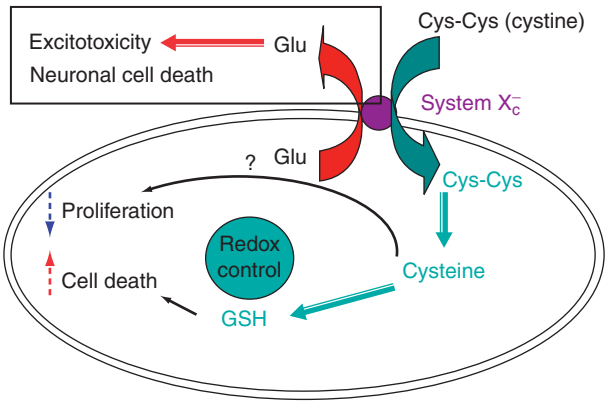

Figure 5 Role of cystine–glutamate exchange (system X_c) in glioma cell biology. Cystine uptake through system X_c serves as precursor for the synthesis of the antioxidant glutathione. Hence, tumor growth is dependent on the continued supply of cystine via this pathway. Glutamate release is an obligatory by-product of the cystine import and may cause excitotoxic neuronal cell death in peritumoral regions.

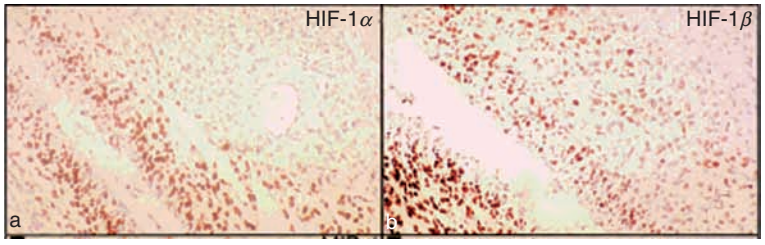

Figure 6 HIF-1α (a) and -β (b) are expressed in pseudopalisading cells surrounding central areas of necrosis. Reproduced from Zagzag D, Zhong H, Scalzitti J, Laughner E, Simons J, and Semenza G (2000) Expression of hypoxia-inducible factor 1α in brain tumors. *Cancer* 88: 2606–2618, with permission.

their receptors (VEGFRs). In glioma cells, the vascular endothelial growth factors VEGF-A and PDGF (platelet-derived growth factor) are upregulated. These ligands bind to and activate the VEGF receptors VEGFR-1 and VEGFR-2, which are located in endothelial cells and also appear to be upregulated in the endothelial cells associated with gliomas. One mechanism for the upregulation of VEGFs and their receptors (especially VEGFR-1) is through the transcriptional regulator hypoxia inducible factor-1 (HIF-1). Although HIF-1 can be degraded under normoxic conditions, the hypoxia that occurs as a tumor mass outgrows the normal blood supply can stabilize HIF-1, allowing it to bind to the hypoxia responsive elements located on several genes, resulting in enhanced transcription of these genes. This transcription factor has been detected in the pseudopalisading cells that form the leading edge of GBMs (Figure 6).

Angiopoietins have also been implicated in glioma angiogenesis. These proteins act by signaling through the receptor tyrosine kinase, Tie-2, on endothelial cells. Ang-1 signaling through this receptor is necessary for the stabilization and maturation of blood vessels. By contrast, Ang-2 signaling occurs during early angiogenesis and may block the stabilizing effect of Ang-1, allowing for the disruption of the contacts between endothelial cells and pericytes, thus permitting the sprouting of new vessels. Ang-2 and Tie-2 expression is absent in normal endothelium but upregulated, along with Ang-1, in many gliomas.

Migration

Unlike tumors of the body, primary brain tumors rarely metastasize by distribution through the blood or lymph vessels. Instead, glioma cells infiltrate the normal brain tissue, often traveling great distances along myelinated axons, the basement membranes of blood vessels, or tracking along the ependyma of ventricles. This aggressive cell migration is highlighted by case studies in which patients with high-grade gliomas underwent hemispherectomies and were still found to harbor tumor cells that had extended beyond the limits of the surgical resection. Migration, however, is not a property that arises *de novo* in tumor cells but characterizes many nonmalignant cells during normal brain development or their response to injury. Migration in any context requires a complex interaction with extracellular matrix (ECM) molecules, reorganization of the cytoskeleton, and the creation of an intercellular space through which a cell invades.

In all cell types, adhesion to the ECM occurs through interactions between ECM proteins and cell surface adhesion molecules. A number of ECM molecules have been demonstrated to be particularly efficient at promoting the migration of glioma cells. These include the proadhesive collagen, fibronectin, and laminin as well as the antiadhesive SPARC, tenascin-C, and thrombospondin. These proteins are often located in the perivascular space, which may be why glioma cells prefer to migrate along blood vessels. In addition to utilizing the native ECM molecules, glioma cells may synthesize and deposit matrix proteins such as tenascin-C, which has been shown to be overexpressed in these tumors. The cell surface adhesion receptors that interact with the ECM include lectins, cadherins, and integrins. Binding to these receptors promotes the rearrangement of the cytoskeleton and upregulation of prosurvival (antiapoptotic) pathways. Several antibodies and competitive substrates for cell surface adhesion molecules have been used to limit the migration of glioma cells *in vitro*.

In addition to being able to form attachments to the ECM, glioma cells must also be able to extend and retract processes in order to move in a particular direction. This involves the polarization of the actin cytoskeleton so that actin is concentrated at the leading edge of the cell. From this region, filopodia, invadopodia, lamellopodia, and pseudopodia extend and form new adhesion complexes. The microtubule cytoskeleton is also important in retracting the rear end of a cell.

The extent of glioma cell migration is surprising in light of the physical constraints imposed by the surrounding brain tissue and skull/spine. Indeed, the extracellular space of the brain has been estimated to be approximately 20 nm in width, which is only a fraction of the cell's diameter. The invasion of glioma

cells, therefore, must also involve mechanisms of creating intercellular spaces or shrinking to fit within already existing narrow spaces. Degradation of the extracellular space is one mechanism by which a glioma cell may create a space to invade. This can be accomplished through the expression and secretion of enzymes such as matrix metalloproteinases and cathepsin-B. Indeed, expression of these molecules appears to be upregulated in glioma cell lines and patient biopsies. Another option employed by glioma cells to invade through narrow spaces is for the cells to become smaller through shrinking and elongation. Electron micrograph images of glioma cells confirm that these cells do assume an elongated morphology as they invade through brain tissue (Figure 7). The process of cell shrinkage, in any cell type, requires the movement of osmolytes (most commonly Cl^- and K^+) out of a cell, followed by the concomitant movement of water. Glioma cells have been shown to express voltage-gated chloride channels (*in vitro* and *in vivo*) and GABA-gated chloride channels (*in vivo*). In addition, chloride channels have been shown to be clustered on the protruding edges of the glioma cell membrane, and movement of chloride through channels has been demonstrated to be necessary for glioma migration in transwell migration assays. However, it is not clear if the volume changes that would accompany chloride movement out of the cell are global or restricted to the protruding process of the cell. Interestingly, chlorotoxin, a drug known to block Cl^- channels, is currently in clinical trials to treat malignant gliomas.

Although gliomas contain the machinery permissive for migration, one must ask the question, what instructs a glioma cell to migrate? The cues that signal a glioma cell to migrate are potentially complex and are not completely understood. One of these signals may involve glutamate-induced Ca^{2+} oscillations, similar to those observed in the well-characterized migration of cerebellar granular cells. Unlike neurons, however, glioma cells do not express NMDA receptors. Instead, glioma cells have been shown to express a Ca^{2+}-permeable AMPA receptor. As mentioned previously, glioma cells efflux glutamate, thereby increasing its concentration in the extracellular space. Increased glutamate has been demonstrated to cause Ca^{2+} changes in glioma cells. Both the glutamate-induced Ca^{2+} changes and the ability of the glioma cells to grow can be inhibited by rendering the AMPA receptor Ca^{2+} impermeable through the adenoviral expression of GLUR2, suggesting that these Ca^{2+} alterations are indeed necessary for glioma cell migration. Although glutamate may serve as an agonist for glioma migration, other signals are also likely to be involved.

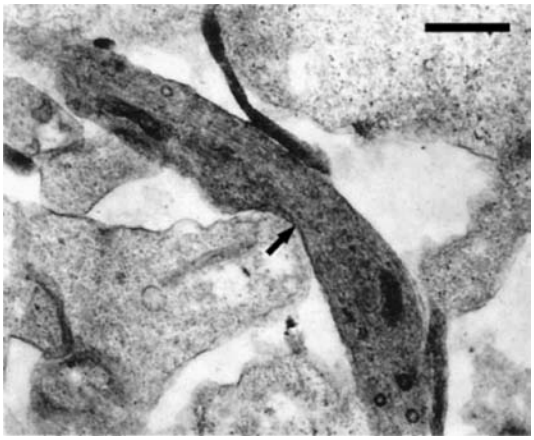

Figure 7 An electron micrograph shows an elongated glioma cell advancing through two normal rat brain cells. Scale bar = 1μm. Reproduced from Soroceanu L, Manning T, and Sontheimer H (1999) Modulation of glioma cell migration and invasion using Cl (−) and K(+) ion channel blockers. *Journal of Neuroscience* 19: 5942–5954. Copyright 1999 by the Society for Neuroscience.

Cell Death

Tissue necrosis is a characteristic feature of GBMs and a defining feature in diagnosing the highest grade tumors. The underlying cellular death occurs by both programmed (mainly apoptosis) and nonprogrammed (mainly necrosis) cell death mechanisms. Necrosis describes the cell death that occurs when the nutritional demands of a growing tumor exceed the nutrient supply from the vasculature. An imbalance in the nutritional supply and demand can occur when the tumor outgrows the existing blood supply or when the supply is reduced as might occur, for example, when a vessel collapses or a clot is formed. This imbalance results in alterations in energy metabolism (specifically reduced oxidative phosphorylation), cell swelling, loss of membrane integrity, and the activation of proteases and other destructive enzymes. In GBMs, necrosis does not occur in single cells but, rather, in a focal population of cells. When imaged by magnetic resonance imaging, intravenous administration of a contrast agent will enhance these foci (Figure 8).

By contrast, apoptosis occurs in single cells within a tumor. Apoptosis is more common in higher grade, more proliferative gliomas but accounts for only a very small fraction of the dead cells in a tumor. However, understanding the mechanisms of apoptosis is important because suppression of apoptosis may contribute to tumorigenesis. Furthermore, because apoptosis does not trigger an inflammatory response, it is the ultimate goal of many anticancer therapies. Apoptosis describes the type of cell death that occurs when specific pathways are activated by either

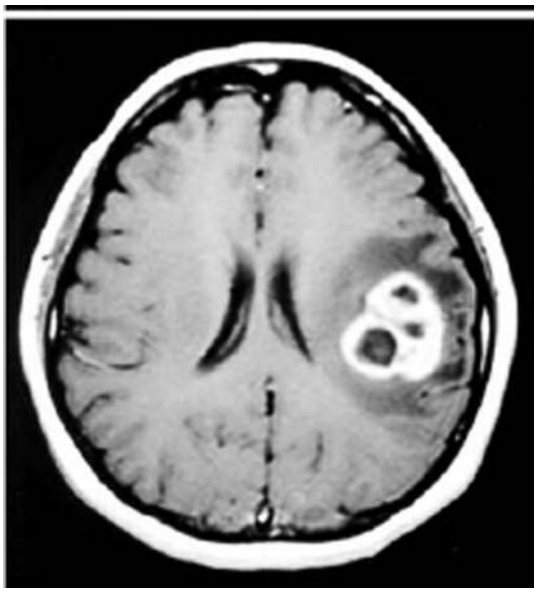

Figure 8 Post-contrast T1-weighted MRI image of a glioblastoma. Reproduced from Okamato K, Ito J, Takahashi N, et al. (2002) MRI of high-grade astrocytic tumors: Early appearance and evolution. *Neuroradiology* 44: 395–402, with permission.

'extrinsic' (e.g., extracellular ligands binding to receptors) or 'intrinsic' (e.g., DNA damage) signals. This type of death is characterized by cell condensation, chromatin fragmentation, and the formation of membrane-bound apoptotic bodies that are cleared by nearby cells through phagocytosis. In the development of gliomas, genetic mutations that enhance survival (e.g., mutations in the PI3K/Akt pathway) or disrupt the intrinsic signaling pathways (e.g., p53 mutations) may disrupt the balance between proliferation and cell death, favoring cell growth and mass formation. However, although apoptosis is not favored in glioma cells, it is able to occur in specific environments. For example, the pseudopalisading cells that surround the necrotic foci of GBMs show an increased percentage of apoptotic cells than the rest of the tumor. This phenomenon demonstrates that the apoptotic signaling pathways of glioma cells are still intact and can be manipulated. This is the objective of many anticancer treatments. Most of these therapies act via the intrinsic apoptosis pathway by producing DNA damage, inhibiting the synthesis of new proteins, or a number of other mechanisms. Other targets for the induction of apoptosis include signaling through extrinsic pathways such as CD95 and DR4 (death receptor 4) and DR5.

The Origin of Glioma Cells

No discussion of gliomas is complete without a brief mention of the two most popular hypotheses concerning the origin of these tumors. Before it was discovered that populations of neural stem cells exist in the adult human brain, it was believed that all gliomas arose from fully differentiated glial cells that had undergone neoplastic transformation. This hypothesis is supported by the fact that the overwhelming majority of primary brain tumors appear to contain cells of glial, rather than neuronal, lineage. Because glia cells are the only differentiated cells of the brain capable of dividing, they are most susceptible to transformation. However, tumors of mixed origin (e.g., oligodendrogliomas) are less likely to develop unless the transformed cell was multipotent, having the ability to differentiate into one of several cell types. Although mature glial cells are not known to be multipotent, investigators have been able to transform them into less differentiated cells using different growth factor treatments and expression systems. Therefore, it is conceivable that a mature, fully differentiated glial cell could be de-differentiated into a less mature, multipotent cell capable of forming a tumor.

Since the discovery of populations of neural stem cells and glial progenitor cells in the adult brain, however, other investigators have hypothesized that it is more conceivable that these cells are responsible for the origin of gliomas. Glioma cells do share characteristics with stem cells, including their affinity for blood vessels and white matter tracts, ability to proliferate, and migratory capacity. In addition, stem cells and progenitor cells are able to differentiate into a limited variety of cell types, accounting for the existence of 'mixed' gliomas. In further support of this hypothesis, many gliomas are located near the ventricles, an area of the brain that is enriched with progenitor cells.

In conclusion, experimental evidence exists to support both of these hypotheses, and the cellular origin of human gliomas will continue to be debated.

See also: Apoptosis in Nervous System Injury; Neural Stem Cells and CNS Diseases; Opsoclonus/Myoclonus and Neuroblastoma; Programmed Cell Death; Stem Cells and CNS Repair; Tumors of the Brain and the Spinal Cord.

Further Reading

Bögler O and Weller M (2002) Apoptosis in gliomas and its role in their current and future treatment. *Frontiers in Bioscience* 7: e339–e353.

Brat D, Castellano-Sanchez A, Hunter S, et al. (2004) Pseudopalisades in glioblastoma are hypoxic, express extracellular matrix proteases, and are formed by an actively migrating cell population. *Cancer Research* 64: 920–927.

Engelhard H, Stelea A, and Cochran E (2002) Oligodendroglioma: Pathology and molecular biology. *Surgical Neurology* 58: 111–117.
Fischer I, Gagner JP, Law M, Newcomb EW, and Zagzag D (2005) Angiogenesis in gliomas: Biology and molecular pathophysiology. *Brain Pathology* 15: 297–310.
Hulleman E and Helin K (2005) Molecular mechanisms in gliomagenesis. *Advances in Cancer Research* 94: 1–27.
Ichimura K, Ohgaki H, Kleihues P, and Collins VP (2004) Molecular pathogenesis of astrocytic tumors. *Journal of Neuro-Oncology* 70: 137–160.
Jeuken JWM, von Deimling A, and Wesseling P (2004) Molecular pathogenesis of oligodendroglial tumors. *Journal of Neuro-Oncology* 70: 161–181.
Kleihues P and Cavenee WK (eds.) (2000) *World Health Organization Classification of Tumours – Pathology and Genetics: Tumours of the Nervous System*. Lyon, France: IARC Press.
Kleihues P, Louis DN, Scheithauer BW, et al. (2002) The WHO classification of tumors of the nervous system. *Journal of Neuropathology and Experimental Neurology* 61: 215–225.
Lehman N, Jordan M, Huhn S, et al. (2002) Cortical ependymoma: A case report and review. *Pediatric Neurosurgery* 39: 50–54.
Markert JM, Devita VT, Hellman S, and Rosenberg SA (eds.) (2005) *Glioblastoma Multiforme*. Boston: Jones & Bartlett.
Ohgaki H (2005) Genetic pathways to glioblastomas. *Neuropathology* 25: 1–7.
Okamato K, Ito J, Takahashi N, et al. (2002) MRI of high-grade astrocytic tumors: Early appearance and evolution. *Neuroradiology* 44: 395–402.
Reifenberfer G and Collins V (2004) Pathology and molecular genetics of astrocytic gliomas. *Journal of Molecular Medicine* 82: 656–670.
Sanai N, Alvarez-Buylla A, and Berger M (2005) Neural stem cells and the origin of gliomas. *New England Journal of Medicine* 353: 811–822.
Sontheimer H (2003) Malignant gliomas: Perverting glutamate and ion homeostasis for selective advantage. *Trends in Neuroscience* 26: 543–549.
Soroceanu L, Manning T, and Sontheimer H (1999) Modulation of glioma cell migration and invasion using Cl(−) and K(+) ion channel blockers. *Journal of Neuroscience* 19: 5942–5954.
Zagzag D, Zhong H, Scalzitti J, Laughner E, Simons J, and Semenza G (2000) Expression of hypoxia-inducible factor 1α in brain tumors. *Cancer* 88: 2606–2618.

Relevant Website

http://www.cbtrus.org – Central Brain Tumor Registry of the United States.

Glutamate

S P H Alexander, University of Nottingham Medical School, Nottingham, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Glutamate is easily the most abundant transmitter in the nervous system and is reported to be the transmitter at 40% of all synapses in the brain. Indeed, it has been said that if a neuron fails to respond to glutamate, it is either not a neuron or it is dead. In itself, therefore, this abundance can lead to some issues. For example, a wide distribution of multiple pharmacological targets suggests that agents targeting the glutamate system will either lack target selectivity or impact on multiple bodily functions. However, it has become clear in recent years that the numerous receptors for glutamate and their localized expression have allowed the development of several potentially exciting therapeutic agents, although relatively few are sufficiently advanced as to reach a clinical setting.

In common with many other neurotransmitters (γ-aminobutyric acid (GABA), acetylcholine, 5-hydroxytryptamine, and adenosine 5′-triphosphate), the action of glutamate is mediated via both rapidly acting transmitter-gated channels (ionotropic receptors, Table 1) and more slowly acting seven-transmembrane (7TM), G-protein-coupled (metabotropic, Table 2) receptors.

Metabolism, Uptake, and Release

Since glutamate is a component of every cell, there is no clear synthetic route that sets apart neurotransmitter glutamate. One intriguing aspect of glutamate metabolism, however, is glutamic acid decarboxylase activity. This enzyme performs an elegant switch in emphasis, by catalyzing conversion of the major excitatory neurotransmitter (glutamate) to the major inhibitory neurotransmitter (GABA).

Glutamate Transport: Plasma Membrane and Vesicular Mechanisms

Due to the major roles of glutamate as an extracellular signaling molecule and as an intracellular building block for proteins combined with the toxicity associated with elevated extracellular levels, effective mechanisms exist for removal of glutamate from the extracellular fluid and also for concentrating the amino acid in synaptic vesicles. Thus, at least five cell-surface glutamate (and aspartate) transporters have been defined as members of the solute carrier family 1 (SLC1), while three vesicular transporters (of the SLC17 family) have also been identified. The cell-surface transporters are multisubunit, homo- or heteropentameric structures in which each subunit appears to have 8TM spanning domains and at least one re-entrant loop that 'dips' into the plane of the membrane. One of the physiological roles for the avid glial glutamate transport system is to allow glial conversion of glutamate to glutamine, which is then exported via the extracellular space into neurons as a means of regenerating neurotransmitter glutamate.

Determination of glutamate uptake in tissue slices or cultured cells is most often accomplished through the use of radiolabeled glutamate, as metabolism to other entities may confound direct assessment of glutamate accumulation.

Detection of Glutamate Release

In order to define a role for glutamate as a mediator of a physiological or pathophysiological mechanism, it is fundamental to be able to measure changes in extracellular glutamate in nervous tissues. *In vitro*, this is commonly accomplished through the use of [^{3}H]-D-aspartate as a surrogate marker, since the D-isomer is less likely to be metabolized. However, there are doubts as to whether [^{3}H]-D-aspartate accumulates in synaptic vesicles and thus the assay may more accurately reflect cytoplasmic glutamate release. It is, therefore, preferable to assess glutamate release directly. One means is to label tissue with [^{3}H]-L-glutamine, with a view to labeling preferentially the neuronal glutamate precursor pool. However, this approach has the drawback that a chromatographic separation of glutamate from glutamine and GABA is required, although it may be useful in some cases to identify simultaneously levels of both excitatory glutamate, as well as inhibitory GABA. A means of directly measuring glutamate release *in vitro* is through the use of high-performance liquid chromatography (HPLC) separation of extracellular fluid following amino acid derivatization with o-phthaldehyde. This has the potential advantage of allowing levels of glutamate and GABA (as well as other bioactive released amino acids such as glycine, aspartate, and taurine) to be assessed concurrently, although it has the disadvantage that a certain period of accumulation (usually 15–20 min) is required to obtain sufficient material to detect. Two 'real-time' *in vitro* assays for glutamate have been described: an enzyme-coupled spectrophotometric version making use of exogenous glutamate dehydrogenase activity and a glutamate-selective electrode.

Table 1 Ionotropic glutamate receptors: composition and pharmacology

	AMPA	*Kainate*	*NMDA*	
Other names	Quisqualate			
Subtypes	GLU_{A1}, GLU_{A2}, GLU_{A3}, GLU_{A4}	GLU_{K1}, GLU_{K2}, GLU_{K5}, GLU_{K6}, GLU_{K7}	GLU_{N1}, GLU_{N2A}, GLU_{N2B}, GLU_{N2C}, GLU_{N2D}, GLU_{N3A}, GLU_{N3B}	
Composition	Homo- or hetero-tetramers	Homo- or hetero-tetramers	Heterotetramers	
Ions gated	Na^+, K^+, (Ca^{2+})	Na^+, K^+	Na^+, K^+, Ca^{2+}	
Channel blockers	Polyamines, argiotoxin, Joro toxin	Polyamines	Dizolcipine, ketamine, phencyclidine, memantine, amantidine	
			Glutamate site	*Glycine site*
Endogenous ligand(s)	L-Glutamate	L-Glutamate	L-Glutamate, L-Aspartate	Glycine, D-Serine
Selective agonists	AMPA, (S)-fluorowillardiine	ATPA, (S)-5-iodowillardiine, (2S,4R)-4-methylglutamate, LY339434, domoic acid	DL-(tetrazol-5-yl) glycine, homoquinolinic acid	(+)HA966
Selective antagonist	NBQX, ATPO, LY293558, GYKI53655	UBP302	D-AP5, CGS19755	5,7-DCKA, L689560, L701324

In vivo detection of extracellular glutamate has been conducted through the use of microdialy coupled to HPLC separation and o-phthaldehyde detection or alternatively through the use of a glutamate-sensitive electrode.

Receptors and Signaling

Twenty-four separate gene products which make up the cell-surface receptors for glutamate have been identified and are divided into two major classes both structurally and functionally; rapidly responding 'ionotropic' transmitter-gated channels and slower responding 'metabotropic' 7TM receptors (Figure 1). The ionotropic receptors consist of three groups of receptors (yellow highlight in Figure 1), identified by the selective synthetic agonists α-amino-3-hydroxy-5-methyl-4-isoxazole proprionic acid (AMPA), kainate, and *N*-methyl-D-aspartate (NMDA). Metabotropic glutamate (mGlu) receptors also consist of three groups as identified by similarities in primary sequence (blue highlight in Figure 1), known simply as group I, group II, and group III.

Ionotropic Glutamate Receptors: AMPA, Kainate, and NMDA Glutamate Receptors

Rapid responses (<1 s) to glutamate in neurons are mediated through cell-surface transmitter-gated channels. These are tetrameric proteins arranged in a radially symmetrical fashion around a central pore, which is permeable to cations (Table 1). Each subunit comprises a large extracellular N-terminus (S1), which interacts with an extracellular loop (S2) to form the agonist binding site. There are three transmembrane segments (TM1, TM3, and TM4) with a single channel lining re-entrant segment (termed MD2 or p-loop) that inserts into the plane of the membrane from the cytoplasmic face (Figure 2) between TM1 and TM3. The cytoplasmic C-terminus is the site for interaction of a number of protein partners (*vide infra*).

Nomenclature and composition As ionotropic glutamate receptors are multi-subunit proteins with few selective pharmacological ligands, nomenclature of the individual subunits is based on physical (i.e., primary sequence) distinctions rather than pharmacological means. Of the eight subunits identified as NMDA receptor components, a minimal configuration in heterologous expression of the NMDA receptor appears to be two pairs of GLU_{N1} and GLU_{N2} subunits, although it appears more likely that, physiologically, NMDA receptors are mixtures of all three subunit groups (GLU_{N1}, GLU_{N2}, and GLU_{N3}, encoded by the genes *GRIN1*, *GRIN2A/2B/2C/2D*, and *GRIN3A/3B*). The GLU_{N3} subunits appear to have a generally inhibitory influence on NMDA receptor function. Many of the genes encoding NMDA receptor subunits undergo alternative splicing, which can generate multiple isoforms with differing pharmacological profiles.

AMPA receptors appear to be made up of any permutation of GLU_{A1}, GLU_{A2}, GLU_{A3}, and GLU_{A4} subunits (derived from the genes *GRIA1–4*, respectively); although GLU_{A2} subunits appear to require co-expression with other AMPA receptor subunits for cell-surface localization. A feature of the AMPA receptor subunit GLU_{A2} is the ability to function as two versions inter-converted following RNA editing (resulting in a switch from a glutamine to arginine residue; a Q/R switch) from a channel highly permeable

Table 2 mGlu receptors: biochemistry, localization and pharmacology

	Group I		*Group II*		*Group III*			
Signalling	$G_{q/11}$, PLC-β, Ca^{2+}		$G_{i/o}$, ↓AC, ↓cAMP, ↑K^+		$G_{i/o}$, ↓AC, ↓cAMP, ↑K^+			
Group selective ligands								
Endogenous	L-Glutamate, L-Aspartate		L-Glutamate, L-Aspartate, NAAG		L-Glutamate, L-Aspartate, L-Serine-*O*-phosphate			
Synthetic agonists	Quisqualate, 3.5-DHPG		DCG-IV, 2R,4R-APDC, LY354740		L-AP4, (RS)PPG			
Antagonists	LY393675		EGLU, LY307452		MAP4, MSOPPE			
Subtypes	*$mGlu_1$*	*$mGlu_5$*	*$mGlu_2$*	*$mGlu_3$*	*$mGlu_4$*	*$mGlu_6$*	*$mGlu_7$*	*$mGlu_8$*
Protein partners	Homer, TRPC1, PIKE	Homer, FMR1, PIKE						
Selective agonists		CHPG		NAAG	(−)−PHCCC	Homo-AMPA		
Selective antagonists	3-MATIDA	ACDPP	PCCG-4			THPG		MPPG
Allosteric enhancers	Ro67-4853	DFB	LY487379				AMN082	
Allosteric inhibitors	CPCCOEt	MPEP, MTEP						
Impact of gene disruption	Altered synaptic transmission in the hippocampus and cerebellum; symptoms of cerebellar ataxia; increased pain tolerance	Disruption of hippocampal synaptic transmission & LTP; weakened spatial learning and memory acquisition; reduced cocaine self-administration; a deficit in prepulse inhibition	Reduced hippocampal LTD associated with an absence of agonist-induced inhibition of glutamate release		Disruption of spatial learning and memory acquisition; loss of efficacy of the anti-epileptic agent baclofen	Normal retinal structure but marked impairments in visual receptiveness	A shortened lifespan associated with epileptic seizures	Heightened anxiety in behavioural tests

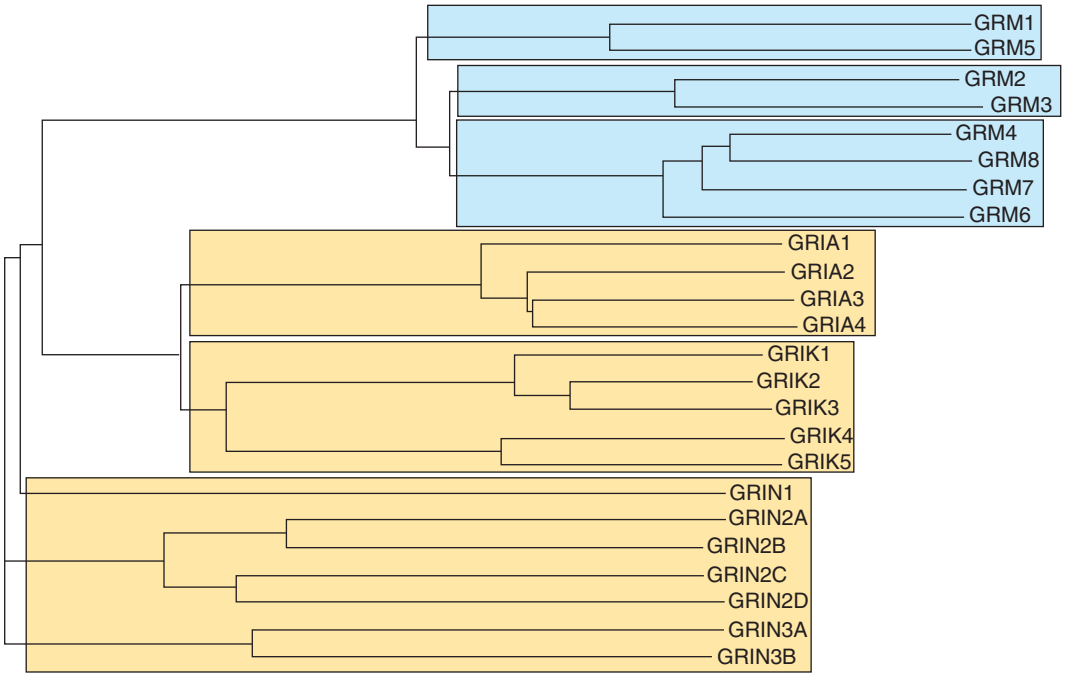

Figure 1 Phylogram of the human glutamate receptor subunit peptide sequences. Picked out in yellow highlight are the three families of ionotropic receptor (AMPA, kainate, and NMDA receptors). Blue highlighting identifies the three groups of mGlu receptor.

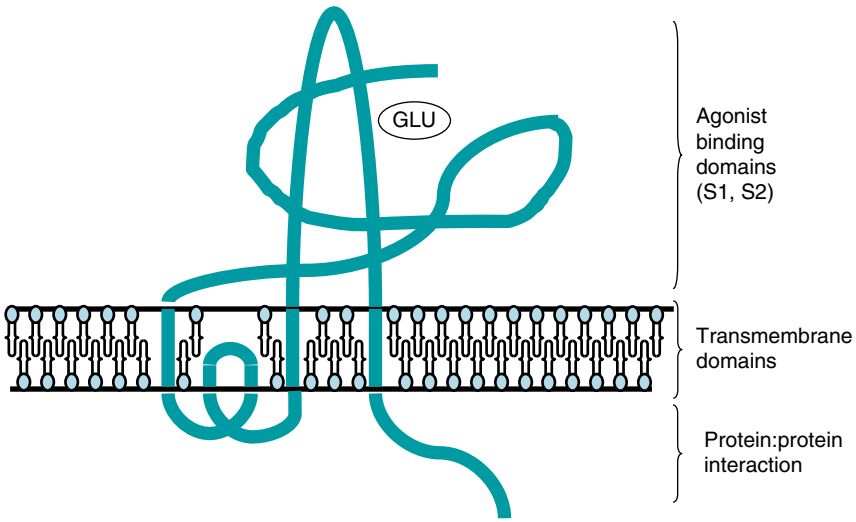

Figure 2 An ionotropic glutamate receptor subunit. Indicated is an extracellular N-terminus agonist binding site (S1, S2), the transmembrane domains (TM1, TM3, TM4, and re-entrant loop) and the intracellular C-terminus.

to calcium ions, to one much less permeable. All AMPA receptor subunits are subject to alternative splicing, which results in two variants termed 'flip' and 'flop' with distinct desensitization characteristics.

Of the kainate receptors, GLU_{K1} and GLU_{K2} subunits (from the genes *GRIK1* and *GRIK2*, respectively) are a subfamily which exhibits high-affinity agonist binding without function and which appear to require co-expression of members of the other kainate receptor subfamily, GLU_{K5}, GLU_{K6}, or GLU_{K7} subunits (from *GRIK3*, *GRIK4*, and *GRIK5* genes, respectively) to show functional activity. In contrast, the latter subfamily of subunits appears capable of generating functional kainate receptors as homotetramers.

Pharmacology and signaling The NMDA receptor is primarily located postsynaptically and is only activated in the presence of both glutamate and glycine, this dual requirement being unique amongst transmitter-gated channels. Glutamate acts on GLU_{N2}

subunits, while glycine binds to both GLU_{N1} and GLU_{N2} subunits, with receptor activation requiring two molecules of each ligand. Under resting conditions, the NMDA receptor ion channel is blocked by Mg^{2+} ions, which dissociate when the cell is depolarized. The open channel exhibits a relatively high permeability to Ca^{2+}, along with the monovalent cations Na^{+} and K^{+}. Similar to the $GABA_A$ receptor, the NMDA receptor exhibits multiple allosteric modulatory sites, including inhibitory responses to Zn^{2+} and protons, as well as subunit-selective potentiatory responses to neurosteroids and polyamines.

Selective agonists exist for both glutamate- and glycine-binding sites of the NMDA receptor (Table 1 and Figure 3), as well as selective antagonists. However, the most widely employed antagonists function as channel blockers and include dizocilpine (previously MK-801), ketamine (Figure 4), and phencyclidine (Table 1). It is thought that the abuse of the latter two agents (known as 'special K' and 'angel dust' or PCP, respectively), originally produced as dissociative anesthetics, is mediated at the molecular level through blockade of the NMDA receptor. Antagonists at the strychnine-insensitive glycine co-agonist binding site of the NMDA receptor, such as 5,7-dichlorokynurenate (Figure 4), are effective *in vitro* and *in vivo* inhibitors of NMDA responses.

AMPA receptors can be activated by high concentrations of kainate (>30 μM), while activation of kainate receptors by AMPA is dependent on the constituent subunits, with the GLU_{K2} subunit appearing to confer AMPA sensitivity (Figure 3). The quinoxalinedione antagonists, such as NBQX (Figure 4) and CNQX, show some limited selectivity in distinguishing AMPA from kainate receptors (Table 1). Synthetic allosteric inhibitors of AMPA and kainate receptor function (GYKI5246 and NS3763, respectively) have also been described. Intriguingly, a number of natural toxins (e.g., from Joro spiders, Figure 4) have been described which function as channel blockers for AMPA (and, to a lesser extent, kainate) receptors. Presumably, these have evolved as a means of predators disabling insect prey, since glutamate is the neurotransmitter at the invertebrate neuromuscular junction, acting on postsynaptic AMPA-like receptors.

In addition, aniracetam, cyclothiazide, and diazoxide are widely used to interfere with the channel desensitization characteristics of AMPA (and, to a lesser extent, kainate) receptors through an undefined allosteric interaction.

Protein partners A key area of regulation of (particularly) AMPA and kainate receptors is desensitization and intracellular trafficking. An intracellular protein

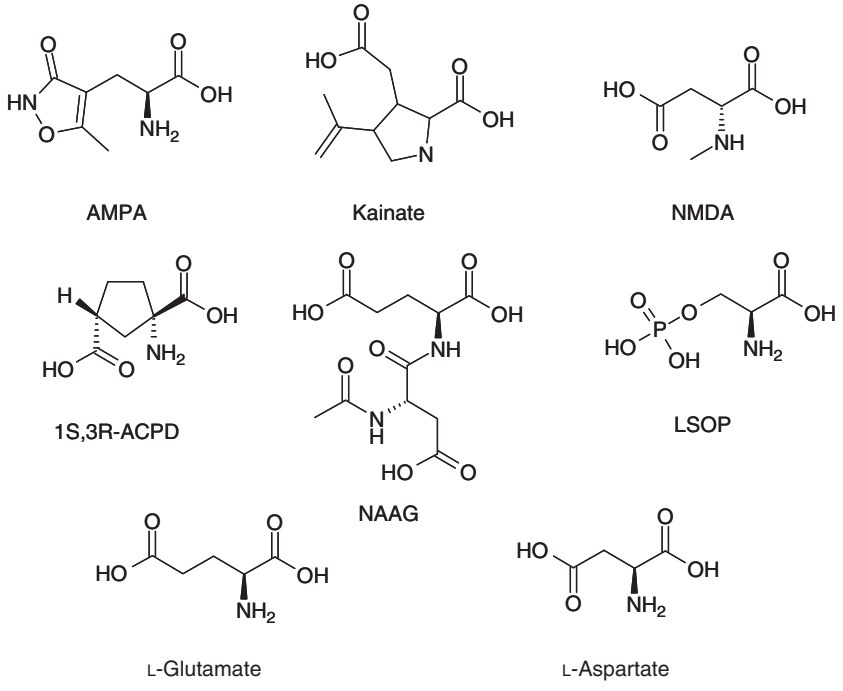

Figure 3 Glutamate receptor agonists. The top row indicates the eponymous agonists at AMPA, kainate, and NMDA ionotropic glutamate receptors. The middle row indicates mGlu receptor agonists: the nonselective agonist 1S,3R-ACPD, the mGlu3-selective NAAG, and the mGlu4/7/8-selective L-serine-*O*-phosphate. The bottom row shows the endogenous excitatory amino acids, L-glutamate and L-aspartate.

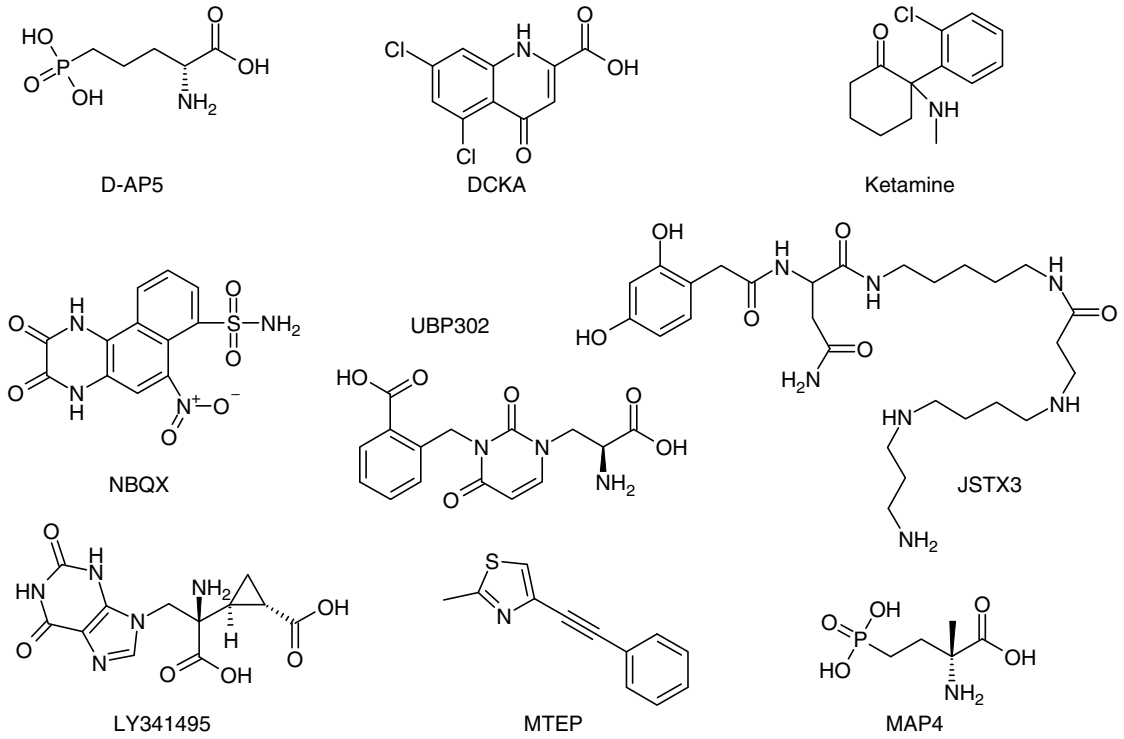

Figure 4 Glutamate receptor antagonists. The top row illustrates the three classes of NMDA receptor antagonist: antagonists competing for glutamate (D-AP5) and glycine (DCKA) binding sites, as well as an uncompetitive (open channel blocker, ketamine) antagonist. The middle row presents competitive AMPA (NBQX) and kainate (UBP302) receptor antagonists and a spider venom-derived AMPA channel blocker (JSTX3). The bottom line illustrates a nonselective mGlu receptor competitive antagonist (LY341495) and noncompetitive mGlu5 (MTEP) and competitive mGlu4/6 (MAP4) receptor antagonists.

called glutamate receptor-interacting protein (GRIP) appears to allow interaction between GLU_{A2} subunits and other proteins, including protein interacting with C kinase (PICK, which recruits protein kinase C) and kinesin (an intracellular molecular motor associated with the cytoskeleton). AMPA receptors also interact with Lyn, a protein kinase, which in turn signals via the extracellular signal-regulated kinase signaling pathway, leading to an increased expression of brain-derived neurotrophic factor. This growth factor is associated with neurogenesis in the adult brain and its expression is reduced following exposure to stress.

mGlu Receptors

mGlu receptors are members of the family C of 7TM receptors, which also includes $GABA_B$ and calcium-sensing receptors. This family is made up of long polypeptides (~1000 aa) and is characterized by the presence of a large extracellular N-terminus (~600 aa, Figure 5). This portion of the receptor contains the agonist binding site and a cysteine-rich domain, which appears to stabilize the three-dimensional structure of the receptor. The 7TM domains bear little sequence homology with the classical biogenic amine receptors (members of the rhodopsin class, family A), which make up the ligand-binding site in the plane of the plasma membrane. The N-terminus binding site is often described as a 'Venus fly-trap' arrangement in which the ligands bind within the hinge region (Figure 5). Conceptually, there is thus an obstacle to overcome as to how ligand binding at some distant extracellular point is able to influence G-protein activity across the hydrophobic plasma membrane. The means by which this appears to be mediated is through dimerization of the receptor, thereby allowing conformational changes in the extracellular region to be transduced through the membrane. A further consequence of the extended structure of the C family of 7TM receptors is the regulation of receptor activity through ligands binding at sites other than the cognate ligand-binding site. Thus, for many of the mGlu receptor subtypes, selective positive and/or negative allosteric modulators have been described which extends the potential for pharmacological/therapeutic exploitation considerably (*vide infra*). The cytoplasmic C-terminus is also enlarged in comparison with most members of the rhodopsin group of 7TM receptors and this is the site for interaction with a number of intracellular protein partners (*vide infra*).

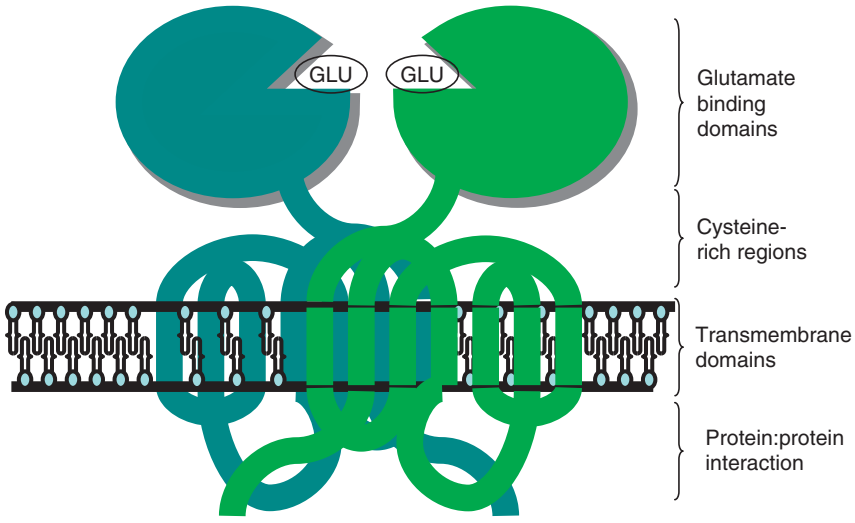

Figure 5 A dimeric mGlu receptor. Indicated are the extracellular N-terminus glutamate binding site, cysteine-rich conformation stabilizing regions, the transmembrane domains (TM1–7) and the intracellular C-terminus site for protein:protein interactions.

Eight subtypes of mGlu receptor have been identified. They are usually classified into three groups on the basis of similarities in primary sequence (Figure 1), signaling characteristics and pharmacology. Thus, group I mGlu receptors ($mGlu_1$ and $mGlu_5$) couple primarily to elevation of calcium, while group II ($mGlu_2$ and $mGlu_3$) and group III ($mGlu_4$, $mGlu_6$, $mGlu_7$, and $mGlu_8$) receptors evoke an inhibition of cAMP.

Localization mGlu receptors are located throughout the brain and spinal cord, and are also found in the retina ($mGlu_6$), pancreas ($mGlu_1$, where they appear to subserve a role in the monitoring of nutritional status), and in taste buds ($mGlu_4$, where they mediate sensation to the 'umami' taste). In neural tissues, there is evidence for differential neuronal and astrocytic locations of mGlu receptors. $mGlu_1$ receptors appear to be exclusively expressed in neurons, while $mGlu_5$ receptors are found on both astrocytes and neurons. Similarly, $mGlu_2$ receptors are selectively neuronally expressed, while $mGlu_3$ receptors have been identified on both neurons and astrocytes. Of the group III receptors, $mGlu_6$ receptors appear to be exclusively retinal, while the remainder are primarily neuronally located.

Pharmacology and Signaling

Although glutamate is an endogenous ligand at all of the mGlu receptors, the potency of glutamate varies widely. $mGlu_1$, $mGlu_3$, and $mGlu_5$ receptors respond to submicromolar glutamate, while $mGlu_2$, $mGlu_4$, $mGlu_6$, and $mGlu_8$ receptors respond to glutamate at concentrations up to $\sim 10\,\mu M$. In contrast, $mGlu_7$ receptors only respond to concentrations of glutamate approaching the millimolar range, leading to the suggestion either that glutamate is not the principal endogenous ligand or that the $mGlu_7$ receptor acts presynaptically as a glutamate autoreceptor, sensing synaptic release of glutamate and switching off further release as 'toxic' levels are reached.

Metabotropic receptors may be activated selectively compared to the ionotropic glutamate receptors by 1S,3R-ACPD (Figure 3), while LY341495 is a useful antagonist with a similar selectivity (Figure 4). Selective agonists and antagonists exist for the three groups of, if not the individual, mGlu receptors (Table 2). Thus, quisqualate is a high potency group I agonist, although it also activates AMPA receptors potently, while LY393675 acts as a group I-selective antagonist. A useful group II agonist is LY354740, while EGLU antagonizes group II receptors, albeit with low affinity. L-AP4 is a nonselective group III agonist, while MAP4 blocks most of these receptors (Figure 4).

Some members of the mGlu receptor family also respond to other endogenous ligands, raising the possibility of native activation in the absence of glutamate. Thus, *N*-acetylaspartylglutamate (also known as spaglumic acid) is a selective $mGlu_3$ receptor agonist, while L-serine-*O*-phosphate activates group III mGlu receptors (Table 2 and Figure 3).

A relatively unusual feature of the family C of 7TM receptors is the regulation of receptor activity through allosteric sites. Thus, both positive and negative allosteric modulators have been described for $mGlu_1$, $mGlu_2$, $mGlu_4$, $mGlu_5$, and $mGlu_7$ receptors. This means (for example) that $mGlu_5$ receptors can be activated by 2-chloro-5-hydroxyphenylglycine, a response which can be enhanced by 3,3′-difluorobenzaldazine,

which alone is ineffective. Activation of the $mGlu_5$ receptor can be inhibited either competitively (3-hydroxy-6-methyl-*N*-(6-methylpyridin-2-yl)pyridine-2-carboxamide) or noncompetitively (3-[(2-methyl-1,3-thiazol-4-yl)ethynyl]pyridine, Table 2).

The group I mGlu receptors, $mGlu_1$ and $mGlu_5$, couple primarily via activation of members of the $G_{q/11}$ family of G-proteins to phospholipase C-β and, hence, to elevation of intracellular calcium ion levels. The primary mode of signaling of the remaining mGlu receptors is via pertussis toxin-sensitive $G_{i/o}$ family members. This results in inhibition of adenylyl cyclase activity and the subsequent reduction in cyclic AMP levels. Arguably of more importance in neuronal cells is the activation of potassium channels and the subsequent hyperpolarization, resulting (typically) in a reduction in transmitter release. The $mGlu_6$ receptor, expressed in ON retinal bipolar cells, couples to activation of PDE6, a cyclic GMP-selective phosphodiesterase, thereby leading to the closure of cGMP-activated cation channels.

Protein partners It is apparent that the expanded C-terminus of the mGlu receptors has a role in the interaction with intracellular proteins, beyond the archetypal G-proteins. In particular, the group I mGlu receptors are associated with an immediate early gene product, Homer, which is enriched at the postsynaptic density. Homer appears to play an anchoring role, bringing together elements of the signaling cascade, including $mGlu_1$ and $mGlu_5$ receptors, phospholipase C, cell-surface transient receptor potential canonical (TRPC) calcium channels, intracellular inositol 1,4,5-trisphosphate receptors, and the phosphatidylinositol 3-kinase enhancer protein, PIKE. It has been suggested that the major function of Homer is to act as a 'brake' on group I mGlu receptor activity through the calcium mobilization pathway and to enhance coupling via PI 3-kinase.

These protein partners may also be involved in bringing together mGlu and NMDA receptors, leading to a functional interaction, such that $mGlu_5$ receptor activation enhances NMDA receptor function, while NMDA receptor activation reverses $mGlu_5$ receptor desensitization.

Physiological Functions

The primary physiological action of glutamate in the central nervous system (CNS) is as an excitatory transmitter, allowing neuronal pathways to function throughout the brain. At the same time, glutamate is thought to be the major transmitter underlying the acquisition of memory. The molecular correlates of memory acquisition appear to differ according to the synapse and (possibly) the maturity of the animal. The phenomenon of long-term potentiation (LTP), which has been suggested to occur at every excitatory synapse in the CNS, is demonstrated as an increased efficacy of synaptic transmission following a period of high frequency stimulation. It is well described at synapses of the Schaffer collateral–commisural pathway in the hippocampus, in which high-frequency stimulation of the perforant path input from the entorhinal cortex causes a long-lasting enhancement of low-frequency activation of the CA1 pyramidal cell output pathway. Cerebellar long-term depression (LTD) is a model of synaptic memory in which the efficiency of synaptic transmission (activation of Purkinje neurons) is reduced following high-frequency stimulation (coincident activation of parallel and climbing fiber presynaptic pathways).

One explanation of these changes in efficacy of synaptic transmission, which does not hold true at all synapses, is that the underlying mechanisms involve trafficking of AMPA receptors, such that LTP may be associated with an increase in postsynaptic AMPA receptor expression (possibly resulting from recruitment of cryptic receptors), while LTD may result from phosphorylation and internalization of AMPA receptors, reducing postsynaptic receptor numbers. This plasticity, therefore, is consistent with an increased and decreased efficiency of synaptic conductance, respectively.

One of the most insightful means of determining physiological roles of particular proteins is though the use of gene disruption models, although a caveat with embryonic knockout models is that the possibility exists for compensatory mechanisms to lessen or alter the true role of the gene and protein in question. Table 2 identifies some of the findings from such gene disruption studies focusing on mGlu receptors. A common theme is an alteration in synaptic plasticity, observed with disruption of $mGlu_1$, $mGlu_2$, $mGlu_4$, and $mGlu_5$ receptors, further underlining the importance of neurotransmitter glutamate in memory acquisition.

Glutamate as a Peripheral Neurotransmitter

There is good evidence that glutamate subserves an excitatory transmitter function in the gastrointestinal tract, with indications of glutamate release, ionotropic and mGlu receptors, and glutamate transporters in enteric ganglia. In addition, application of glutamate and its analogs has been shown to regulate gastrointestinal motility and solute secretion. Synapses in the enteric nervous system undergo

LTP-type plasticity, and it is thought that group I mGlu receptors mediate this phenomenon, at least in part.

Pathology

Stroke, Ischemia, and Neurodegenerative Diseases

An overabundance of extracellular glutamate is associated with excitotoxicity of neurons, which appears to involve activation of NMDA receptors. The mechanism of excitotoxicity is not absolutely defined, but there is a crucial role of a prolonged elevation of intracellular calcium ions, leading to mitochondrial damage and apoptosis. The NMDA receptor antagonist dizocilpine reached an advanced preclinical level as a potential antistroke therapy, failing on the basis of damage to unstressed brain areas due to a narrow therapeutic window. This drug may have been a victim of its own success, in that the high affinity and low reversibility displayed as an open channel blocker allowed little differentiation between positive and harmful effects. The generation of a lower-affinity agent, memantine, which exhibits greater reversibility and an increased tolerability has allowed approval for its use in man. The current indication for memantine is for treatment of Alzheimer's disease. The rationale for its use is that the prolonged presence of glutamate (albeit at lower concentrations) leads to protracted NMDA receptor activation and excitotoxicity of neurons and a cognitive deficit. By reducing NMDA receptor activation, it is hoped to reduce the excitotoxicity and reduce the cognitive deficit.

The accumulation of plaques positive for β-amyloid in Alzheimer's disease is associated with neurodegeneration and, since processing and secretion of γ-secretase (an enzyme which cleaves the amyloid precursor protein) is enhanced by activation of group I mGlu receptors, these receptors have also become a focus for research in this area.

Another neurodegenerative disorder in which glutamate overactivity has been implicated is Parkinson's disease. The glutamatergic pathways involved may be either thalamostriatal or corticostriatial, and so it has been suggested that group II agonists or $mGlu_5$ antagonists might be therapeutically useful by inhibiting glutamate release or action, respectively, at these synapses. Indeed, LY354740 has been observed to display efficacy in animal models of Parkinson's, although the $mGlu_5$ antagonist 2-methyl-6-(phenylthynyl) pyridine was ineffective.

A further potential exploitation of glutamate signaling in combating the effects of neurodegenerative diseases, such as Alzheimer's and Parkinson's, is the use of 'AMPAkines,' agents which enhance AMPA receptor function. As described above, a number of agents interfere with AMPA receptor function as allosteric modulators or by impeding desensitization. Aniracetam is an example of such an agent, which has been exploited as a nootropic substance, also known as 'smart drugs.' These agents have a positive profile in animal models and those in clinical and 'street' use are reported to enhance cognitive powers (concentration, memory, etc.); their use has yet to be approved in the treatment of Alzheimer's, although piracetam is used to treat the involuntary twitches suffered by Parkinson's patients.

Epilepsy

In vitro and *in vivo* focal application of glutamate results in epileptiform activity. This action appears to be mediated by the ionotropic receptors, and postmortem examination of the brains of epileptics suggests a neurodegenerative profile consistent with overactivity of these receptors. One potential therapeutic avenue is the exploitation of $mGlu_4$ and $mGlu_7$ receptors, which have been suggested to be glutamate autoreceptors, acting presynaptically to reduce glutamate release at the synapse. As noted in Table 2, disruption of either of these receptors leads to changes in seizure susceptibility further highlighting the potential for exploitation of glutamate signaling in treating this disorder.

Anxiety and Depression

In preclinical assessments for leads as anti-anxiety/anti-depressant action, there are promising results for both $mGlu_2$-selective agonists and $mGlu_5$-selective antagonists. It is hypothesized that the $mGlu_2$ agonists function by reducing frontal cortex transmitter release presynaptically, while the $mGlu_5$ antagonists regulate postsynaptic glutamate excitability.

Pain

Several observations suggest that mGlu receptors represent an excellent target for the treatment of pain. Members of all three groups of mGlu receptors are present at the primary afferent-spinal cord synapse or the subsequent ascending pathway. There is particular interest in group I mGlu receptors, partly because of evidence for diminished pain sensitivity in transgenic mice (Table 2), but also because of the efficacy of antagonists in multiple pain models.

Schizophrenia

Schizophrenia is suggested to be the consequence of an imbalance of glutamatergic and dopaminergic signaling systems. Phencyclidine, which blocks

NMDA receptors (as well as dopamine uptake), induces schizophrenia-like symptoms in man. In animal models, the group II agonist LY354740 was observed to reduce some of the behavioral effects of PCP, with a distinct pattern to traditional neuroleptics. In addition, the observation that $mGlu_5$-knockout mice have an altered behavioral trait observed in schizophrenics (Table 2) has led to heightened interest in the therapeutic exploitation of these receptors for this condition.

Addiction

There is extensive interest in the use of group I antagonists and group II agonists as potential therapeutic agents in the treatment of drug addiction. In animal models, these agents reduce the symptoms of withdrawal from nicotine and morphine. In addition, there is evidence for sensitization of group II mGlu receptors in particular brain areas associated with drug dependence, following chronic exposure to morphine.

Cerebellar Ataxia

Animal models of disrupted $mGlu_1$ receptors suggest many similarities with cerebellar ataxia, but of arguable greater significance is the observation of auto-antibodies against $mGlu_1$ receptors in patients suffering from paraneoplastic cerebellar degeneration. These antibodies were observed to inhibit cerebellar function *in vitro* and *in vivo* arguing for a pivotal role for $mGlu_1$ receptors in this disorder.

Concluding Remarks

Although there has been an abundance of research investigating the physiological and pathophysiological roles of glutamate, as well as extensive medicinal chemistry programs to develop selective ligands, there is still an under-representation of therapeutic agents focusing on this pivotal transmitter.

See also: AMPA Receptors: Molecular Biology and Pharmacology; Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin; Glutamate Receptor Organization: Ultrastructural Insights; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Metabotropic Glutamate Receptors (mGluRs): Functions; Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology.

Further Reading

Alexander SPH, Mathie A, and Peters JA (2006) Guide to receptors and channels, 2nd edn. *British Journal of Pharmacology* 147 (supplement 3): S1–S183.

Chen PE and Wyllie DJ (2006) Pharmacological insights obtained from structure–function studies of ionotropic glutamate receptors. *British Journal of Pharmacology* 147: 839–853.

Dingledine R, Borges K, Bowie D, and Traynelis SF (1999) The glutamate receptor ion channels. *Pharmacological Reviews* 51: 7–61.

Hinoi E, Takarada T, Tsuchihashi Y, and Yoneda Y (2005) Glutamate transporters as drug targets. *Current Drug Targets CNS and Neurological Disorders* 4: 211–220.

Kew JNC and Kemp JA (2005) Ionotropic and metabotropic glutamate receptor structure and pharmacology. *Psychopharmacology* 179: 4–29.

Lujan R, Shigemoto R, and Lopez-Bendito G (2005) Glutamate and GABA receptor signalling in the developing brain. *Neuroscience* 130: 567–580.

Madden DR (2002) The structure and function of glutamate receptor ion channels. *Nature Reviews Neuroscience* 3: 91–101.

Matute C, Domercq M, and Sanchez-Gomez MV (2006) Glutamate-mediated glial injury: Mechanisms and clinical importance. *Glia* 53: 212–224.

Niswender CM, Jones CK, and Conn PJ (2005) New therapeutic frontiers for metabotropic glutamate receptors. *Current Topics in Medicinal Chemistry* 5: 847–857.

Robbins TW and Murphy ER (2006) Behavioural pharmacology: 40+ years of progress, with a focus on glutamate receptors and cognition. *Trends in Pharmacological Sciences* 27: 141–148.

Swanson CJ, Bures M, Johnson MP, et al. (2005) Metabotropic glutamate receptors as novel targets for anxiety and stress disorders. *Nature Reviews Drug Discovery* 4: 131–144.

Watkins JC and Jane DE (2006) The glutamate story. *British Journal of Pharmacology* 147(supplement 1): S100–S108.

Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin

M C Chang, R W Cho, and P F Worley,
Johns Hopkins University School of Medicine,
Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Glutamate Receptor Clusters

Clustering of receptors is important for synapse formation and plasticity. In order to achieve efficient synaptic transmission, neurons must aggregate neurotransmitter receptors at the active zone of the synapse, and, in turn, stabilize and limit the lateral mobility of the clustered receptors. The most well-studied example of this is found in the peripheral nervous system where the postsynaptic membrane of the neuromuscular junction (NMJ) is enriched with a density of acetylcholine receptors (AChRs) $>10\,000\,\mu m^2$, 1000-fold higher than the density of extrasynaptic AChRs. Agrin, a heparan sulfate proteoglycan, mediates postsynaptic differentiation by clustering and enriching the postsynaptic membrane with AChRs. Agrin has been shown to initiate postsynaptic signaling through a postsynaptically localized muscle-specific receptor tyrosine kinase (MuSK) that leads to the clustering of AChRs. Mice deficient in agrin exhibit low levels of AChRs as well as an altered postsynaptic distribution within the myotubes. Some synapses were found to lack postsynaptic specializations entirely within these mutant mice. Thus, agrin clustering of AChRs contributes to the organization and resulting function of the postsynaptic membrane to respond rapidly and reliably to neurotransmitter released from the overlying nerve terminal.

Analogous extracellular and transmembrane proteins in the central nervous system (CNS) that contribute transynaptically to receptor clustering and synapse formation have been sought and a few candidate receptor clustering molecules within the CNS have been identified. EphB2 receptor, stargazin, and Narp are three well-studied examples of proposed receptor clustering molecules that function in the CNS (Figure 1).

EphB2/Ephrin

The first Eph receptor discovered, EphA1, was identified from a screen of a human cDNA library using the viral oncogene v-*fps* as a probe for low-stringency cross-hybridization to other tyrosine kinases. Since then, at least 14 Eph receptors have been identified in vertebrates, making them the largest subfamily of receptor tyrosine kinases. Eph receptor ligands are ephrins, which are tethered to the cell membrane of opposing Eph receptor-expressing cells by a transmembrane domain or a glycosylphosphatidylinositol anchor.

The protein structure of EphB2 is similar to that of the canonical Eph receptors. EphB2 is a type I transmembrane protein. The N-terminal extracellular domain contains motifs typically found in adhesion molecules: a putative immunoglobulin (Ig) domain, a cysteine-rich region, and two fibronectin type III repeats. The cysteine-rich region contains two C*n*C*x*C*n*C repeats ($n = 12$–15 acids and x is any amino acid) that are frequently found in epidermal growth factor (EGF)-like repeats of extracellular matrix proteins. The EphB2 cytoplasmic domain contains a tyrosine kinase domain that is conserved in all Eph receptors. EphB2 contains a potential postsynaptic density/disk/zonula occludens-1 (PDZ) binding domain at the C-terminus, possibly linking EphB2 to cytoskeletal and/or signaling proteins. A yeast two-hybrid screen using this domain has identified two synaptic PDZ proteins implicated in α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type glutamate receptor trafficking, PICK1 and GRIP, as potential binding partners.

Since all Eph receptors are expressed in the developing nervous system, they were originally hypothesized to play a role in development. Indeed, Eph/Ephrin signaling complexes have been implicated as regulators of contact-mediated axon growth cone guidance and neuronal cell migration. Recent evidence, however, suggests that Eph receptor, specifically EphB2, and their cognate ligands, the ephrin-B subgroup of ephrin ligands, play an additional role in synaptic development and plasticity. In the hippocampus, EphB2 is localized postsynaptically at excitatory neurons of the CA1 region of the hippocampus. Ephrin-B is expressed presynaptically and binding to EphB2 receptors induces clustering and activation of the receptors.

The formation of EphB2 clusters is important for signal initiation that is activated by ephrin binding. Unlike most other ligands for receptor tyrosine kinases, ephrins do not act as soluble mediators of receptor activation. Instead, ephrins are membrane bound and require membrane linkage that is necessary for cluster formation. Interaction of membrane ephrin-B and EphB2 at sites of cell–cell contact initiates cluster formation of both molecules in their respective plasma membranes. Ephrin-B binding to EphB2 results in the assembly of higher-order tetramers composed of ephrin-B/EphB2 dimers resulting from high-affinity protein–protein interactions between specific domains. Other mechanisms that contribute to cluster formation include enrichment of ephrin-B and EphB2 in lipid rafts as well as EphB2 interaction with PDZ

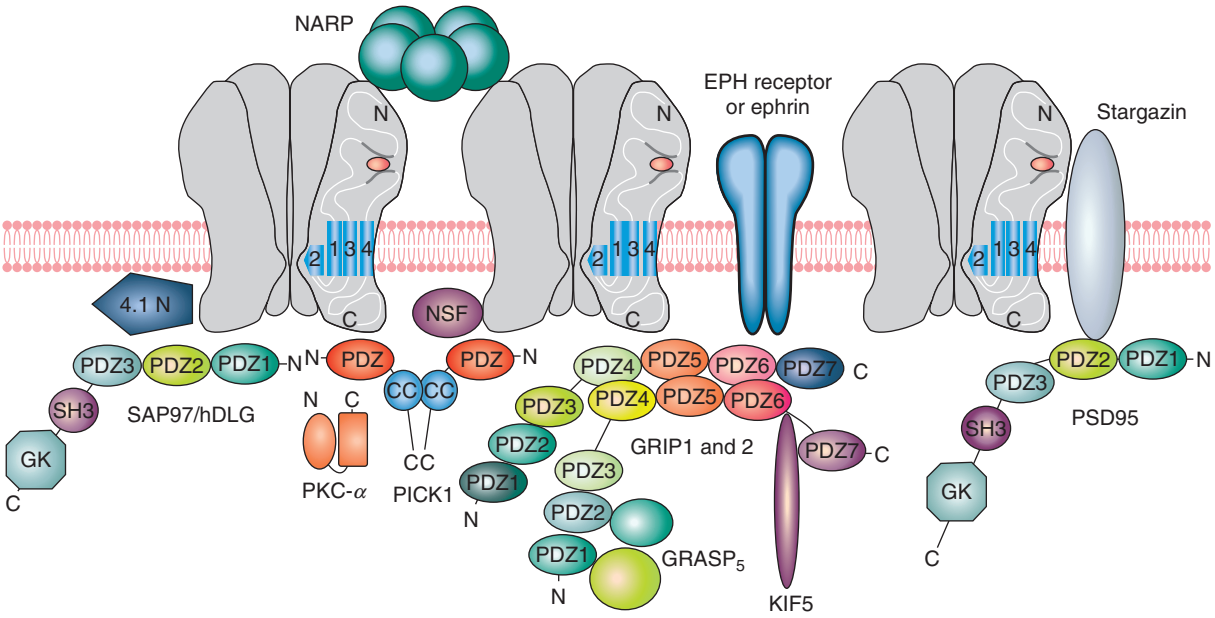

Figure 1 Receptor clustering. See text for discussion.

proteins. The ephrin-B/EphB2 clustered complex results in the initiation of intracellular signal transduction by EphB2 receptor tyrosine kinase domain. Interestingly, cluster formation also initiates 'reverse signaling,' in that cluster formation also results in signal transduction activation in ephrin-expressing cells.

A possible role for EphB2/ephrin-B complex in synapse function and formation is suggested by the interaction between EphB2 and *N*-methyl-D-aspartate (NMDA)-type glutamate receptors (NMDARs). EphB2 interacts directly with the NMDAR subunit NR1 via extracellular domain interaction. Cultured neurons treated with preclustered recombinant ephrin-B ligands to mimic *in vivo* activation of EphB2 receptors result in an increase in the number of endogenous EphB2 clusters. Increased clustering of EphB2 increases the number of co-clusters with NR1 and the number of synapses. Structure–function analyses in combination with experiments that block EphB2 signaling using dominant negative constructs demonstrate distinct function of the EphB2 extracellular domain and the intracellular kinase domain. That is, NMDAR cluster formation by EphB2 requires only the extracellular domain interaction, but subsequent stages of synapse assembly require the kinase domain. Although the kinase domain is not required for NR1 clustering, it has been shown to be required for NMDAR function, synaptic plasticity paradigms such as long-term potentiation (LTP) and long-term depression (LTD) at the mossy fiber/CA1 synapse of the hippocampus, and spine morphogenesis.

The role of EphB2 in clustering of NR1 and synapse formation and plasticity has been confirmed in mice lacking EphB2. Although these mice exhibit deficits in axon guidance and boundary formation in some regions of the brain, the fundamental circuitry and structure of the hippocampus are unaffected. Mice that lack EphB2, however, exhibit a 40% reduction in the number of NR1 receptors at synapses. Electrophysiological recordings performed at the Schaffer collateral/CA1 synapse in the hippocampus of these mice show that they have normal basal transmission. Studies examining synaptic plasticity within these mice show that although the early phase of LTP is also normal, LTP is not maintained over 100 min. Similarly, LTD is attenuated in EphB2 null mice. Surprisingly, mice expressing EphB2 lacking the intracellular catalytic domain exhibit normal LTP and LTD at this synapse, suggesting that extracellular EphB2 interaction with ephrin-B and NMDAR is sufficient to maintain synaptic plasticity.

EphB2 null mice exhibit deficits in spatial memory as measured by the Morris water maze, suggesting a defect in hippocampal memory function. However, it is difficult to interpret these results because the EphB2−/− mice exhibit a significant reluctance to participate in the task and a reduced swim speed.

Stargazin

Interest in stargazin began when a mutant mouse spontaneously arose within the A/J inbred strain at the Jackson Laboratory. These 'stargazer' mice exhibit mild gait ataxia beginning early in development, and later demonstrate a periodic tossing of their head backwards, so as to stare up and gaze at the stars.

Electrocorticographic recordings of these mice revealed prolonged, frequent spike-wave seizures with behavioral arrest, characteristic of absence epilepsy. Mapping of the stargazer gene (*stg*) led to the discovery of an early transposon insertion into the second intron at the voltage-dependent calcium channel $\gamma 2$ subunit locus (*Cacng2*) on chromosome 15, which results in decreased Cacng2/stg/$\gamma 2$ mRNA expression. Further characterization of this gene revealed a 36-kDa integral membrane protein, named stargazin, which contains four membrane-spanning domains, with both N- and C-termini present on the cytoplasmic side.

Weak protein sequence identity (25%) with the γ-1 calcium channel subunit of muscle suggested a possible role as a neuronal voltage-dependent calcium channel (VDCC) regulator. However, there is only modest evidence supporting the role of stargazin in regulating VDCCs, and the magnitude of the contribution is still in question. Alternatively, stargazer mice display an impairment of AMPA-type glutamate receptor (AMPAR) function, suggesting a role for stargazin in AMPAR regulation. Consistent with this notion, the cerebellum of *stg* mutant mice exhibits a lack of functional AMPARs within granule cells of the cerebellum. Functional studies of these *stg* mutant mice show an impairment in acquisition of eye-blink conditioning. Interestingly, although stargazin is expressed in other various regions throughout the brain, including the olfactory bulb, cerebral cortex, thalamus, and hippocampus, *stg* mice possess functional AMPARs within these areas.

Stargazin is a member of the family of transmembrane AMPAR regulatory proteins (TARPs), which includes $\gamma 3$, $\gamma 4$, and $\gamma 8$. All members of this family are expressed in brain but exhibit differential distribution patterns and are capable of direct interaction with the AMPAR subunits – GluR1, GluR2, or GluR4. Chimeric analysis between the TARPs and an inactive TARP homolog, $\gamma 5$, which cannot associate with AMPARs, revealed that the interaction between AMPARs and TARPs requires both the first extracellular loop as well as a 50-amino-acid region within the C-terminal intracellular domain of stargazin. Disruption of either of these domains, either through deletion, insertion, or $\gamma 5$ domain-swapping, leads to abolishment of AMPAR binding.

Stargazin facilitates the trafficking and clustering of AMPARs (but not kainite- or NMDA-type glutamate receptors) at excitatory synapses in two distinct ways. First, stargazin mediates delivery of functional AMPAR from the Golgi complex to the plasma membrane. Co-microinjection of stargazin and GluR1 into *Xenopus laevis* oocytes results in a greatly increased surface expression of AMPARs, when compared to oocytes injected with the GluR1 subunit alone. This function is dependent upon binding of a 50-amino-acid region within the C-terminal intracellular domain of stargazin to the AMPAR, as disruption or abolishment of this domain results in the impaired insertion of AMPAR into the membrane. In addition, binding of glutamate to AMPAR results in disruption of the stargazin/AMPAR interaction, perhaps via a conformational change in AMPAR. This is consistent with the notion that stargazin plays a role in activity-dependent cycling of AMPARs at the synapse. Surprisingly, although the first extracellular domain of stargazin has been shown to be necessary for AMPAR interaction, this domain is not necessary for the trafficking of AMPAR to the surface.

Second, the final four amino acids of stargazin (TTPV) constitute a type I PDZ-binding domain, and thus stargazin is capable of binding PDZ domain-containing proteins. The proper targeting and clustering of AMPAR to synapses is dependent upon binding of stargazin with the PDZ domain-containing synaptic scaffolding protein PSD-95. Deletion of the last four amino acids in the C-termini results in loss of AMPAR at the synapse. Despite the loss of synaptic AMPAR, however, trafficking of AMPAR to the plasma membrane is unaltered. This PSD-95/stargazin interaction is regulated by protein kinase A (PKA). Phosphorylation of the PDZ-binding domain by PKA results in the inability of stargazin to bind PSD-95. This in turn results in the loss of AMPAR at the synapse, and consequent attenuation of synaptic AMPAR currents.

In addition to trafficking AMPARs, stargazin has recently been shown to modulate AMPA receptor gating. Whereas AMPARs normally deactivate and desensitize quickly after activation, co-transfection of stargazin with AMPAR in *Xenopus* oocytes results in a slower deactivation and desensitization, an increased amplitude of steady-state currents in response to both glutamate and especially kainate, and an increased open channel probability. This effect is independent of the increased trafficking effect of stargazin on AMPAR. Subsequent chimeric analysis between stargazin and the $\gamma 5$ subunit (an inactive TARP homolog) found that the first extracellular domain of stargazin is necessary for this AMPA channel regulation.

Narp

Neuronal activity-regulated pentraxin (Narp; also termed NP2) binds and clusters AMPARs. Narp is a member of the neuronal pentraxin family of secreted, calcium-dependent lectins, originally named for their ability to form pentamers of five noncovalently bound subunits. Members within this family include

neuronal pentraxin 1 (NP1) and neuronal pentraxin receptor (NPR). Other nonneuronal pentraxins include apexin, serum amyloid P component (SAP), and C-reactive protein (CRP), which are involved in the innate immune response. Narp is a 432-amino-acid protein with an N-terminal signal sequence, followed by two coiled-coil domains and a 'pentraxin signature' sequence of H*x*C*x*S/TW*x*S in the C-terminus. Narp was identified using a differential and subtractive cloning strategy to isolate genes that are rapidly upregulated in response to neuronal activity, and that contribute to protein synthesis-dependent synaptic plasticity. Narp mRNA expression increases following maximal electroconvulsive seizure (MECS) with a time course that peaks at 2 h and returns to pre-MECS levels by 24 h. In addition to induction via MECS, Narp is upregulated by experimental forms of synaptic plasticity, including long-term potentiation. Moreover, high-level expression in rodent cortex is maintained in a steady-state manner by natural synaptic activity, since blockade of input from the eye via monocular injection of tetrodotoxin (TTX) results in a rapid reduction of Narp mRNA within the contralateral visual cortex. Narp protein is highly expressed in both brain and testes. In brain, Narp is enriched within layers 2/3 and 5/6 of the cortex as well as in the CA3 and dentate gyrus of the hippocampus. Narp accumulates at dendritic excitatory synapses on both excitatory and inhibitory neurons and can be secreted from both the pre- and postsynaptic membranes.

The role of Narp in the formation and remodeling of synapses was first proposed based on the observation that Narp co-clusters with AMPARs in cultured hippocampal neurons. Follow-up co-immunoprecipitation experiments demonstrated that Narp interacts directly with all AMPAR subunits, presumably via one of the extracellular AMPAR regions. Narp cluster formation is correlated with its ability to act as a synaptogenic agent. Overexpression of Narp increases the number of excitatory synapses in cultured spinal neurons. Conversely, dominant-negative Narp mutants that cause the rapid degradation of native Narp reduce the number of AMPAR clusters, suggesting a role for Narp in synapse formation.

The molecular basis of Narp's ability to form clusters has been examined. Narp assembles into higher-order quaternary structures of hexameric units, involving three position-specific intermolecular disulfide bonds as well as intermolecular protein–protein interactions. Narp contains three cysteine residues located within or near the N-terminal coiled-coil domains (residues 14, 26, and 79). Cysteine-14 and cysteine-26 mediate formation of hexameric units of Narp assembly; cysteine-79 mediates linkages between Narp hexamers. These specific disulfide linkages as well as protein–protein interactions by the coiled-coil domains mediate Narp cluster formation. The other neuronal pentraxins, NP1 and NPR, form assemblies via the same mechanism. Moreover, the neuronal pentraxins can form disulfide-linked heteromultimers that function cooperatively at excitatory synapses.

Conclusion

The molecular mechanisms of neurotransmitter receptor clustering are fundamental for the modification of synaptic strength and maturation of synapses. Although many of the molecular mechanisms of synaptic maturation and plasticity might be shared, they are distinct processes occurring at different time periods. Here, we have discussed the known roles that EphB2 receptor, Narp, and stargazin play in receptor clustering. The future challenge will be to more clearly define the individual and concerted roles of these proteins, their family members, and other yet-to-be identified proteins in receptor clustering during synaptic maturation and synaptic plasticity.

See also: AMPA Receptors: Molecular Biology and Pharmacology; Eph Receptor Signaling and Spine Morphology; Glutamate; Glutamate Receptor Organization: Ultrastructural Insights; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Chen L, Chetkovich DM, Petralia RS, et al. (2000) Stargazin regulates synaptic targeting of AMPA receptors by two distinct mechanisms. *Nature* 408: 936–943.

Contractor A, Rogers C, Maron C, et al. (2002) Trans-synaptic Eph receptor-ephrin signaling in hippocampal mossy fiber LTP. *Science* 296: 1864–1869.

Dalva MB, Takasu MA, Lin MZ, et al. (2000) EphB receptors interact with NMDA receptors and regulate excitatory synapse formation. *Cell* 103: 945–956.

Grunwald IC, Korte M, Adelmann G, et al. (2004) Hippocampal plasticity requires postsynaptic ephrinBs. *Nature Neuroscience* 7: 33–40.

Grunwald IC, Korte M, Wolfer D, et al. (2001) Kinase-independent requirement of EphB2 receptors in hippocampal synaptic plasticity. *Neuron* 32: 1027–1040.

Henderson JT, Georgiou J, Jia Z, et al. (2001) The receptor tyrosine kinase EphB2 regulates NMDA-dependent synaptic function. *Neuron* 32: 1041–1056.

Letts VA (2005) Stargazer – a mouse to seize! *Epilepsy Currents* 5: 161–165.

Murai KK and Pasquale EB (2004) Eph receptors, ephrins, and synaptic function. *Neuroscientist* 10: 304–314.

O'Brien R, Xu D, Mi R, et al. (2002) Synaptically targeted narp plays an essential role in the aggregation of AMPA receptors

at excitatory synapses in cultured spinal neurons. *Journal of Neuroscience* 22: 4487–4498.

O'Brien RJ, Xu D, Petralia RS, et al. (1999) Synaptic clustering of AMPA receptors by the extracellular immediate-early gene product Narp. *Neuron* 23: 309–323.

Tomita S, Adesnik H, Sekiguchi M, et al. (2005) Stargazin modulates AMPA receptor gating and trafficking by distinct domains. *Nature* 435: 1052–1058.

Tomita S, Fukata M, Nicoll RA, et al. (2004) Dynamic interaction of stargazin-like TARPs with cycling AMPA receptors at synapses. *Science* 303: 1508–1511.

Torres R, Firestein BL, Dong H, et al. (1998) PDZ proteins bind, cluster, and synaptically colocalize with Eph receptors and their ephrin ligands. *Neuron* 21: 1453–1463.

Tsui CC, Copeland NG, Gilbert DJ, et al. (1996) Narp, a novel member of the pentraxin family, promotes neurite outgrowth and is dynamically regulated by neuronal activity. *Journal of Neuroscience* 16: 2463–2478.

Xu D, Hopf C, Reddy R, et al. (2003) Narp and NP1 form heterocomplexes that function in developmental and activity-dependent synaptic plasticity. *Neuron* 39: 513–528.

Glutamate Receptor Organization: Ultrastructural Insights

Y Fukazawa and R Shigemoto, National Institute for Physiological Sciences, Okazaki, Japan

© 2009 Elsevier Ltd. All rights reserved.

L-Glutamic acid, glutamate, is a main excitatory neurotransmitter in the vertebrate central nervous system. Various subtypes of ionotropic and metabotropic glutamate receptors are expressed depending on the types of synaptic connections and regions of the brain. The precise and quantitative localization of glutamate receptor subtypes is thus indispensable for understanding the mechanisms of excitatory synaptic transmission, information processing, memory formation, and neuronal degeneration ascribed to excessive glutamate receptor activation. However, the small size of synapses (~1 μm wide, with a surface area of ~0.5 μm^2) hampers analysis of their precise localizations by light microscopy, which has limited spatial resolution (~500 nm). Electron microscopy with immunogold labeling has revealed highly differentiated distributions of glutamate receptors in the neuronal plasma membrane with a resolution up to 20 nm. Three immunogold labeling techniques – preembedding, postembedding, and freeze-fracture replica labeling – have different advantages and technical limitations and are used in combination to elucidate distinct distribution patterns and quantitative localizations of the receptors.

Localizations of Glutamate Receptors in Distinct Membrane Compartments of Postsynaptic Neurons

Twenty-two subtypes of glutamate receptors (GluRs) have been identified in the mammalian genome, and these are categorized into two major classes – ionotropic glutamate receptors (iGluRs) and metabotropic glutamate receptors (mGluRs). The former class has been further categorized into four types based on their pharmacological properties: the *N*-methyl-D-aspartate receptors (NMDARs) type, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors (AMPARs) type, kainic acid receptors (KARs) type, and orphan glutamate receptors (δ_1 and δ_2 glutamate receptors). The latter class has also been categorized into three groups, groups I–III, based on their pharmacological properties and second messenger signaling pathways. Ultrastructural localizations of these receptors have been conducted extensively in the hippocampus and cerebellum. NMDARs, AMPARs, δ_2 GluR, and group I mGluRs ($mGlu_1$ and $mGlu_5$) have been localized in the excitatory postsynaptic plasma membrane at the ultrastructural level, whereas group II and III mGluRs have been localized preferentially in the presynaptic neuronal plasma membrane. The ultrastructural localizations of KARs and δ_1 GluR have not been well demonstrated.

The postsynaptic dendritic plasma membrane can be subdivided into three subregions: postsynaptic membrane specialization and perisynaptic and extrasynaptic regions. The excitatory postsynaptic membrane specialization has postsynaptic density and directly faces the presynaptic active zone, where the release of neurotransmitter occurs. At the electron microscopic level, the postsynaptic membrane specialization can be identified by the asymmetrical appearance of electron densities between pre- and postsynaptic membranes (type I synapses). The perisynaptic region surrounds the postsynaptic membrane specialization and is 60–200 nm wide, with no clear boundary with the extrasynaptic region. NMDARs, AMPARs, and δ_2 GluR have been shown to accumulate over the postsynaptic membrane specialization, whereas group I mGluRs are mainly enriched in the perisynaptic plasma membrane (Figure 1). Expression of NMDARs/AMPARs and group I mGluRs has also been found in the extrasynaptic plasma membrane with lower density than in synaptic or perisynaptic regions, respectively.

Intrasynaptic Distribution of iGluRs

Tangential distribution of NMDARs and AMPARs in the excitatory synapse has been investigated in several brain regions. NMDARs have been shown to be relatively abundant in the center of the postsynaptic membrane specialization of pyramidal cells in the hippocampus and neocortex. On the other hand, AMPARs have been found to be localized in the synaptic specialization in various patterns depending on the types of synapses. AMPARs in the mossy fiber–pyramidal cell synapses and Shaffer collateral/commissure fiber–CA1 pyramidal cell (SC–Py) synapses in the hippocampus have been shown to be localized evenly over the synaptic specialization, whereas those in the synapses of entorhinal and neocortical cells are reported to be prevalent in the center and periphery of the specialization, respectively. On the other hand, microcluster-like distribution of AMPAR labeling within synapses has also been reported in the hippocampal synapses by postembedding immunogold labeling. Freeze-fracture replica labeling, which visualizes two-dimensional localization of membrane

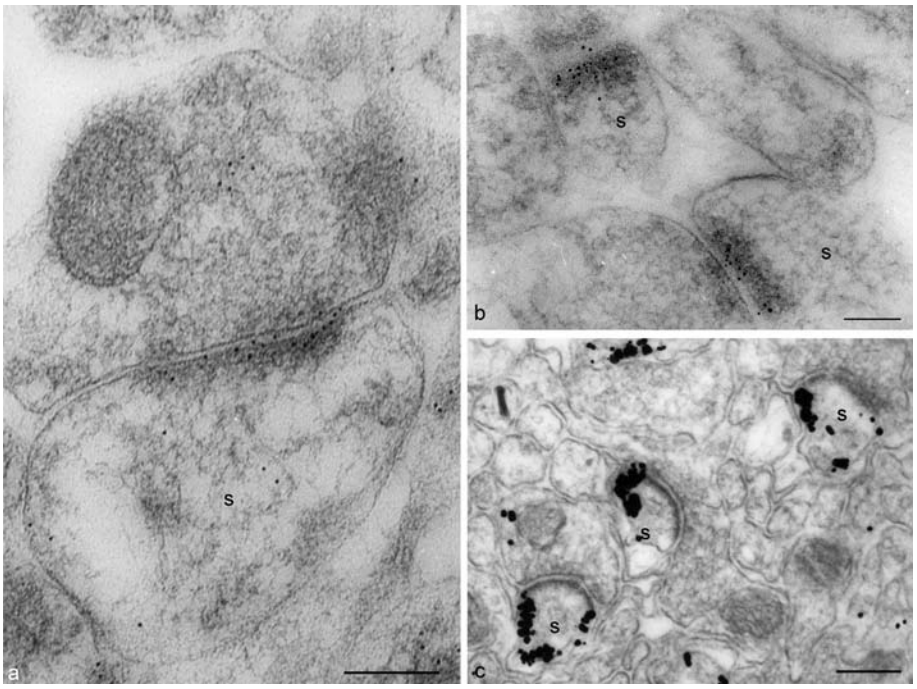

Figure 1 Differential distribution of three types of glutamate receptors at synapses on dendritic spines (s) in the cerebellar cortex (a and c) and in stratum radiatum of the CA1 hippocampal area (b). The AMPA-type glutamate receptor is enriched in the postsynaptic membrane specialization (a; GluR 2/3 subunits, postembedding immunogold reaction, 5 nm particles). The NMDA-type glutamate receptor is also enriched in the postsynaptic membrane specialization (b; NR1 subunit, postembedding, 5 nm immunogold reaction). The mGluR1α-type glutamate receptor is enriched in the perisynaptic region of parallel fiber–Purkinje cell synapses (c; preembedding silver intensified immunogold labeling). Scale bar = 0.1 μm in (a, b), 0.2 μm (c).

molecules with high resolution, has also revealed AMPAR microclusters within the postsynaptic specialization of parallel fiber–Purkinje cell (PF-PC) synapses in the cerebellum (Figure 2). These clusters have various subsynaptic localizations depending on individual synapses. Such microclusters have not been detected in the climbing fiber–Purkinje cell synapses, which have denser and more homogeneous localization of AMPARs (Figure 2). δ_2 GluRs have been shown to localize evenly over the synaptic specialization of the PF–PC synapses.

Taken together, it seems likely that NMDARs are preferentially distributed in the center of the postsynaptic membrane specialization in a wide variety of glutamatergic synapses, whereas intrasynaptic distribution of the AMPARs is regulated in highly differentiated forms in individual synapses. It is not clear how these highly differentiated distributions of iGluRs develop and take part in shaping characteristic features of synaptic transmission at each type of synaptic connection. It seems likely that synaptically localizing scaffold proteins associating with specific types of glutamate receptors are involved in the regulation of postsynaptic localization of glutamate receptors.

Glutamate Receptors and Their Associating Proteins

Biochemical and molecular biological studies have revealed large numbers (~500) of cytoplasmic and membrane-spanning proteins in the postsynaptic density fraction containing the postsynaptic membrane specialization. These proteins are thought to form a large protein complex as visualized as postsynaptic density in electron microscopic images. Some of these cytoplasmic protein members act as scaffold proteins for glutamate receptors at the postsynapse, taking part in controlling receptor localization, function, and signaling. MAGUK proteins such as postsynaptic density-95 (PSD-95; also known as SAP90) and Shank proteins (also known as ProSAP, SSTRIP, CortBP, Synamon, and Spank) are thought to be key proteins for regulating glutamate receptor localizations in the synapses. They have multiple protein–protein interaction domains and can act as a hub of protein networks in the postsynaptic density. MAGUK proteins directly interact with NMDARs and also interact with AMPARs through transmembrane AMPA receptor regulatory proteins (TARPs) assembled in the AMPA

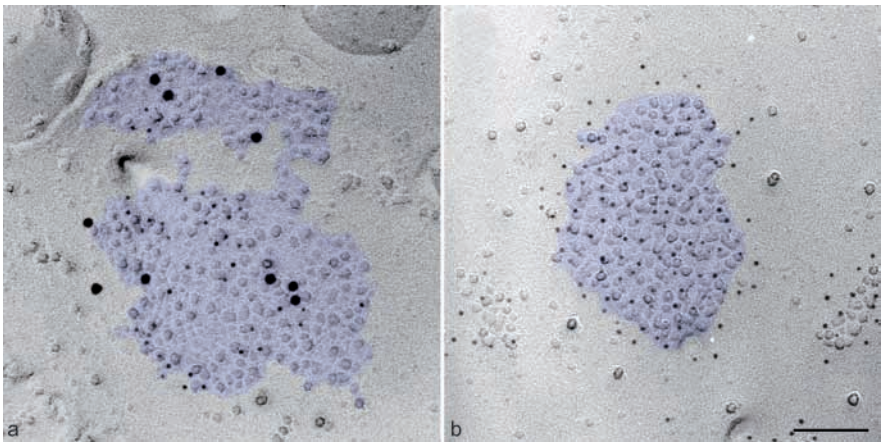

Figure 2 Two distinct patterns of intrasynaptic distribution of AMPA-type glutamate receptors. Distribution of the AMPA-type glutamate receptor in two types of synapses in the cerebellum was visualized by SDS-digested freeze-fracture replica labeling. (a) Microcluster-like aggregations of immunoparticles (5 nm) for AMPA-type receptors are found within the postsynaptic membrane specialization (blue) of the parallel fiber–Purkinje cell synapse identified by immunoparticles (15 nm) for the δ_2 glutamate receptor. (b) AMPA receptor immunoparticles (5 nm) in the climbing fiber–Purkinje cell synapse were distributed evenly over the postsynaptic specialization. Scale bar = 100 nm.

receptor complex. Shank proteins interact with group I mGluRs through Homer proteins (also known as Vesl, PSD-ZIP45, and Cupidin) and also interact with MAGUK proteins through GKAP. As a result, these key proteins assemble a large protein complex beneath the postsynaptic membrane specialization and provide scaffolds for GluRs regulating their intrasynaptic localization. Localization of these key proteins beneath the postsynaptic membrane specialization in the hippocampal synapses has been shown with immunoelectron microscopy. However, tangential localization of these proteins accountable for the distinct localization of NMDARs, AMPARs, and group I mGluRs has not been well documented.

The iGluRs also interact with various cytosolic proteins through their cytoplasmic domains. These interactions regulate expression levels of the receptors in the postsynaptic plasma membrane during development and synaptic plasticity via trafficking mechanisms of receptor-containing vesicles. Therefore, interaction of GluRs with these cytosolic proteins in the postsynaptic membrane specialization can account for differential localization of iGluRs. For example, adaptor protein complex-2 (AP-2) is known to couple NMDARs to the clathrin-mediated endocytic machinery, and these AP-2–NMDAR complexes are retrieved from the plasma membrane. It has been reported that the introduction of a point mutation in the NR2B subunit of NMDAR at a tyrosine phosphorylation site, which is necessary for interaction with α-actinin2, results in displacement of NMDARs toward the peripheral region of the postsynaptic membrane specialization. Similarly, a protein interacting with C-kinase 1 (PICK1) is known to interact with the GluR2 subunit of AMPAR and trigger internalization of the receptor complexes via endocytotic mechanisms. Knockin mice expressing a mutated GluR2 subunit lacking PICK1 interaction show altered distribution of the AMPARs in the postsynaptic plasma membrane of PF–PC synapses.

These observations clearly indicate that the postsynaptic distribution of GluRs is dynamically regulated through interactions with their associate proteins.

Expression Levels of GluRs Are Governed by Both Pre- and Postsynaptic Factors

Glutamatergic transmissions show large variability in their functional characteristics within and across populations of synapses, and changes in their efficacy are thought to be regulated by the different numbers of receptors in individual synapses.

Moreover, glutamatergic synaptic specializations in certain brain regions also show large variability in size. The quantity of receptors expressed within and across populations of synapses is related to anatomical parameters such as synapse sizes. Ultrastructural investigation has demonstrated not only the distinct spatial distribution patterns of receptors but also the number and density of receptors, which are differentially regulated at individual synapses.

NMDARs are expressed at all SC–Py synapses in the CA1 region of the hippocampus, and the expression level has no correlation with the size of the postsynaptic membrane specialization. Moreover, the density of synaptic NMDARs has shown a negative correlation with the size of synapse. On the other hand, SC–interneuron synapses in the same region express NMDARs in proportion to the synapse size,

maintaining constant densities at individual synapses. These observations indicate that the levels and the mode of the NMDAR expression are different among postsynaptic cell types.

Expression of AMPARs shows very different patterns from that of NMDARs. Among the SC–Py synapses, those smaller than 180 nm in diameter have been shown to be AMPAR immunonegative (AMPAR-lacking synapses comprise 50, 25, or 15% of all SC–Py synapses, depending on individual reports). In larger synapses, the level of AMPAR expression has been shown to be positively correlated with the synapse size. Thus, the ratio of AMPARs to NMDARs seems to be positively correlated with the synapse size. Similarly, the synapses between CA3 pyramidal cells (association/commissure (A/C) fiber–Py synapses) in the hippocampus have been shown to be AMPAR immnonegative in 15% of the population. Also, similar to the CA1 synapses, AMPAR-immunopositive synapses have shown a strong positive correlation between the expression level of AMPARs and the synapse size. Synapses made on the same type of cells but with input from the dentate gyrus granule cells (mossy-fiber synapses) have been found to be AMPAR immunopositive in all synapse populations and also show a strong positive correlation between the expression level of AMPARs and the size of synapse. These facts clearly indicate that the level of AMPAR expression depends on the identity of the presynaptic input. Moreover, comparison of AMPAR expression in A/C fiber synapses between pyramidal cells and GABAergic interneurons in the CA3 region revealed that all of the latter synapses express AMPARs and the expression levels, on average, were four times higher than those in the former synapses. Thus, the AMPAR expression level is governed by both pre- and postsynaptic factors.

Although perisynaptic $mGlu_5$ expression has been detected in 50% of SC–Py synapses, quantitative localization of group I mGluRs in respect to anatomical parameters has not been well elucidated. The level of mGluR expression at individual synapses also seems to be regulated by unknown mechanisms.

Conclusions

Ultrastructural investigation of the glutamate receptor organization in the postsynaptic plasma membrane has revealed both a unified view of highly compartmentalized and subtype-dependent localization and a highly variable view of their quantities within individual synaptic populations. The unique and highly variable distribution of each glutamate receptor subtype may reflect prerequisites for their proper functioning, which is crucial for controlling excitability of neurons in given neuronal networks and for processing and storage of information in the brain.

See also: AMPA Receptors: Molecular Biology and Pharmacology; AMPA Receptor Cell Biology/Trafficking; AMPA Receptors: Disease; Dendritic Spine History; Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin; Glutamate; Glutamatergic and Gabaergic Systems; Kainate Receptors: Molecular and Cell Biology; Kainate Receptor Functions; Metabotropic Glutamate Receptors (mGluRs): Functions; Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology; NMDA Receptors and Development; NMDA Receptors and Disease + C464; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation; Postsynaptic Specialization Assembly.

Further Reading

Baude A, Nusser Z, Molnar E, McIlhinney RAJ, and Somogyi P (1995) High-resolution immnogold localization of AMPA type glutamate receptor subunits at synaptic and non-synaptic sites in rat hippocampus. *Neuroscience* 69: 1031–1055.

Kharazia VN and Weinberg RJ (1997) Tangential synaptic distribution of NMDA and AMPA receptors in rat neocortex. *Neuroscience Letters* 28: 41–44.

Lujan R, Nusser Z, Roberts JDB, Shigemoto R, and Somogyi P (1996) Perisynaptic location of metabotropic glutamate receptors mGluR1 and mGluR5 on dendrites and dendritic spines in the rat hippocampus. *European Journal of Neuroscience* 8: 1488–1500.

Masugi-Tokita M, Tarusawa E, Watanabe M, Molnar E, Fujimoto K, and Shigemoto R (2007) Number and density of AMPA receptors in individual synapses in the rat cerebellum as revealed by SDS-digested freeze-fracture replica labeling. *Journal of Neuroscience* 27: 2135–2144.

Nusser Z (2000) AMPA and NMDA receptors: Similarities and differences in their synaptic distribution. *Current Opinion in Neurobiology* 10: 337–341.

Nusser Z, Lujan R, Laube G, Roberts JDB, Molnar E, and Somogyi P (1998) Cell type and pathway dependence of synaptic AMPA receptor number and variability in the hippocampus. *Neuron* 21: 545–559.

Okabe S (2007) Molecular anatomy of the postsynaptic density. *Molecular and Cellular Neuroscience* 34: 503–518.

Racca C, Stephenson FA, Streit P, Roberts JDB, and Somogyi P (2000) NMDA receptor content of synapses in stratum radiatum of the hippocampal CA1 area. *Journal of Neuroscience* 20: 2512–2522.

Steinberg JP, Takamiya K, Shen Y, et al. (2006) Targeted *in vivo* mutations of the AMPA receptor subunit GluR2 and its interacting protein PICK1 eliminate cerebellar long-term depression. *Neuron* 49: 845–860.

Takumi Y, Ramírez-León V, Laake P, Rinvik E, and Ottersen OP (1999) Different modes of expression of AMPA and NMDA receptors in hippocampal synapses. *Nature Neuroscience* 2: 618–624.

Glutamate Regulation of Dendritic Spine Form and Function

R A McKinney, McGill University, Montreal, QC, Canada
S M Thompson, University of Maryland School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During synaptogenesis, pyramidal cell dendrites in the hippocampus extend numerous filopodia. These filopodia are highly motile and can extend up to 10 μm; in comparison, spine extension is generally <2 μm. It has been suggested that a filopodium initiates contact with a growing axon nearby. This contact results in the accumulation of active zone exocytotic proteins in the axon at the site of contact, followed by the accumulation of postsynaptic density (PSD) proteins in the filopodium or at the shaft to which the filopodium retracts. It is therefore thought that the contact between a filopodium and an axon is the initial step in the formation of a mature axospinous synapse. The filopodium then evolves directly into a spine or facilitates the formation of a shaft synapse from which a spine arises at later stages of synaptogenesis. Dendritic spines are small protrusions of the dendritic membrane; they represent the postsynaptic component of the majority of the excitatory synapses in the central nervous system (Figure 1). More than 90% of the excitatory presynaptic terminals impinge on postsynaptic parts that reside on spines. A human brain thus contains over 10^{13} spines. Dendritic spines are found on excitatory and inhibitory neurons, including glutamatergic pyramidal neurons of the neocortex and hippocampus as well as γ-aminobutyric acid (GABA)ergic cerebellar Purkinje neurons. They were first described over a century ago by Santiago Ramón y Cajal. However, their ultrastructural and molecular organization has only recently become accessible, due to advances in microscopy and molecular biology.

Spine Shape and Size Are Correlated with Synaptic Strength

There is tremendous heterogeneity in the shapes and sizes of dendritic spine heads and in the lengths and thicknesses of the spine neck (Figure 2). Considerable progress has been made in correlating dendritic spine morphology with the strength of the synapses formed with the spines. Differences in the size of the spine head are correlated with differences in the size of the postsynaptic density. Because of the molecular links between the proteins of the postsynaptic density and glutamate receptors, it is postulated that the size of the spine head is correlated with the number of glutamate receptors expressed.

The relationship between spine size and sensitivity to glutamate has been addressed recently using state-of-the-art photorelease of chemically caged glutamate. This technique permits the release of glutamate in a small volume, comparable to the size of a single spine head. It is thus possible to stimulate α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid-preferring glutamate receptors (AMPARs) on single dendritic spines and to measure the resulting postsynaptic currents with whole-cell voltage-clamp recording. A strong correlation between spine head volume and the amplitude of the excitatory postsynaptic current has been observed, suggesting that larger spine heads do contain more AMPARs.

Electrophysiological and immunocytochemical studies have demonstrated that *N*-methyl-D-aspartate (NMDA) and non-NMDA receptors are not colocalized at all synapses. Rather, it appears that a significant proportion of synapses express only functional NMDA, and not AMPA, receptors. Because the NMDA receptors do not produce substantial current flow at the resting membrane potential, these synapses can appear 'silent.' The proportion of such silent synapses is high in early postnatal life and declines gradually during development. Synaptic maturation and induction of long-term potentiation (LTP) are accompanied by the appearance of functional AMPA receptors at previously silent synapses. The morphological correlates of silent synapses are not known. One hypothesis is that morphological maturation of spines, from long, thin filopodia with small heads to adultlike spines, occurs in parallel to insertion of AMPA receptors at silent synapses.

The diameter of the spine neck has also been predicted to affect synaptic efficacy. Long, narrow necks should result in a region of high electrical resistance between the synapse in the spine head and the parent dendrite. Shorter, fatter necks should have a lower electrical resistance, and the lower resistance might facilitate the transfer of dendritic depolarization for a given size of synaptic current. Other modeling studies have suggested that the neck resistance should have only an insignificant effect on synaptic transfer, however. Use of glutamate photostimulation techniques has allowed this question to be addressed directly. In cortical pyramidal cells, the length of the spine neck is inversely correlated with the amplitude of the electrical response to photostimulation. Remarkably, stimulation of those spines with the longest, thinnest

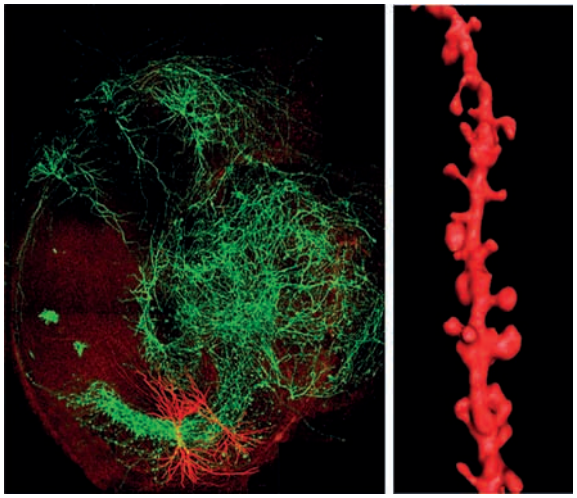

Figure 1 Morphology of dendritic spines. (Left) Confocal micrograph of hippocampal slice cultures prepared using a variegated transgenic mouse line labeled with a membrane-bound marker with enhanced green fluorescent protein) under the Thy1 promoter, so that only a subset of cells is labeled (green); a fluorescent dye was used to label CA3 pyramidal cells (red). (Right) High-magnification image of a dendritic segment of the pyramidal neuron shown in the left-hand panel, showing distinct spines of the various shapes. Micrographs courtesy of Sylvain Hugel and Anne McKinney.

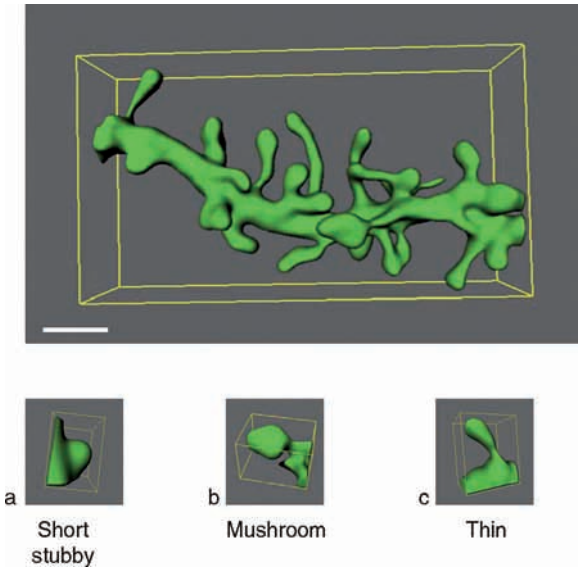

Figure 2 Diversity of spine shape. Dendritic spines are classified into three main types: (a) short, stubby spines (<0.5 μm in length), (b) mushroom-type spines, consisting of a short neck and mushroom-shaped head, or (c) thin, long spines with an elongated neck and small head. (Top) Three-dimensional volume-rendered portion of the dendritic tree of a CA1 pyramidal cell from a green fluorescent protein-expressing mouse brain.

necks (>2 μm in length) fails to induce a detectable depolarization at the cell soma, even though positive evidence of responses at the spine has been obtained. The spine neck thus appears to possess a significant electrical resistance that shapes the transfer of synaptic depolarization generated in the spine head to the parent dendrite and cell body.

The realization that voltage-dependent channels are also located in spines raised the possibility that spine neck resistance can influence the probability that a synaptic current would depolarize the spine sufficiently to trigger the activation of these channels. Lower spine neck resistance (i.e., shorter, fatter spine necks) should act to minimize spine depolarization and hence decrease synaptic activation of voltage-dependent conductances. In addition, the placement of synapses on spines may serve to keep the resistance of the dendritic shaft relatively constant and high so that it serves as an effective 'cable' from the synapses to the soma. If synapses were formed with the dendritic shaft directly, activation of synaptic conductances would decrease dendritic resistance, thereby diminishing its electrotonic properties.

Spine Shape and Intracellular Ca^{2+}

The influx of Ca^{2+} via NMDAR channels is the key triggering event in the induction of LTP and long-term depression (LTD). The positioning of synapses on spine heads may reflect the function of spines to act as 'compartments' that isolate and confine synaptically generated signals, such as second messengers, activated enzymes, and NMDAR-mediated Ca^{2+} influx, to the synapse at which they have been generated. Changes in the size of spine necks and heads might thus affect the induction of conventional NMDA receptor-dependent synaptic plasticity. In particular, because short spine necks should allow synaptically generated intracellular Ca^{2+} transients to readily diffuse out of spines and into dendrites, it has been theorized that LTP induction should be hampered at synapses formed with such spines. It has been determined that the narrow channel of cytoplasm within the spine neck slows diffusion of cytosolic molecules by about 100-fold, compared to free diffusion. This diffusional barrier allows ions in the spine head cytosol to be at a significantly different concentration than they are in the dendritic shaft, for >100 ms. The mechanisms governing the recovery of basal Ca^{2+} concentration in the spine after synaptic NMDAR activation remain controversial. There is evidence favoring both diffusion through the neck into the dendritic shaft and active extrusion.

It is also interesting to note that the spine neck also acts as a barrier to the rapid diffusion of plasma membrane proteins. Studies of the movement of fluorescently labeled plasma membrane molecules indicate that lateral diffusion is tenfold slower from the dendritic shaft to the spine head than it is for

diffusible cytosolic proteins. For integral membrane proteins, such as receptors, rates of diffusion have been measured at 0.01–0.4 $\mu m^2 s^{-1}$, and for proteins anchored at the intracellular surface of the membrane rates vary between 0.03 and 0.5 $\mu m^2 s^{-1}$. These values are in good agreement with measurements of the diffusion of labeled synaptic glutamate receptors (0.06 μm^2 s^{-1}), which is fivefold less than in the extrasynaptic regions. In this case it is believed that anchoring proteins link receptors to the cytoskeleton and reduce their mobility.

Maintenance of Spine Shape and Number

As dendritic spines are thought to be the sites of memory storage, any loss of a dendritic spine would result in the loss of a memory. It is important to determine how dendritic spines can be maintained. The influence of synaptically released glutamate on postsynaptic structure has been investigated by comparing the effects of deafferentation, receptor antagonists, and blockers of glutamate release in hippocampal slice cultures. CA1 pyramidal cell spine density and length decreased after lesioning of Schaffer collaterals and after application of AMPA receptor antagonists or blockade of spontaneous glutamate release by the addition of botulinum toxin to unlesioned cultures. These results suggest that the lack of AMPA receptor activation following degeneration of presynaptic afferents initiates the process of spine retraction in postsynaptic CA1 pyramidal cells. Loss of spines and synapses induced by lesion or by botulinum toxin could be prevented by simultaneous AMPA application. Synaptically released glutamate thus exerts a trophic effect on spines by acting at AMPA receptors, sufficient to maintain dendritic spines. Interestingly, the larger spine numbers do not change after chronic AMPA receptors blockade, and there are more asymmetric shaft synapses. It still remains to be determined if the thin spines retract with their presynaptic terminal attached, to form an asymmetric shaft synapse. Nonetheless, despite this loss of synapses and spines, the remaining synapses have no significant change in frequency of miniature excitatory postsynaptic currents, indicating a compensatory rearrangement to preserve synaptic function.

Myristoylated alanine-rich C kinase (MARCKS) is a major target of activity-dependent phosphorylation of protein kinase C, and this function is thought to be important in synaptic plasticity and in memory. Surprisingly, ablating MARCKS expression by small interfering RNA (siRNA) leads to destabilization of spine morphology, as exhibited by a decrease in spine number and length. There is no effect on dendritic arborization, suggesting that MARCKS has a specific role in dendritic spine stability. Interestingly, overexpressing wild-type MARCKS also reduces spine number and length, indicating that the level of MARCKS is critical for spine maintenance.

In adult visual cortex (V1), brain-derived neurotrophic factor (BDNF) has been shown to play a fundamental role in dendritic spine maintenance through activation of its receptor, TrkB. Staufen 2, a member of a double-stranded RNA-binding protein family, has also been shown to play an important role in spine maintenance. Interfering with Staufen 2 expression in neurons results in neurons with reduced numbers of dendritic spines. More recently, the Ca^{2+}/calmodulin-dependent protein kinase (CaMK) has been shown to act as a structural protein that can bind to F-actin through a very specific and stoichiometric interaction associated with the B subunit, in addition to its well-known role as a signaling protein critically involved in synaptic plasticity in the brain. This binding is necessary for the maintenance of dendritic spines.

The Rho family of GTPases, a subgroup of the Ras superfamily, are key regulators of the actin cytoskeleton. The Rho GTPases RhoA, Rac, and Cdc42 all have been investigated with regard to their role in spine formation and shape determination. Rac and Rho have opposite effects on spine number. Constitutively active RhoA decreases spine length and number, whereas constitutive expression of an active Rac1 leads to the formation of many small dendritic spines. The roles of some of the guanine nucleotide exchange factors (GEFs) or GTPases in activating proteins that regulate Rho family GTPases have recently been elucidated. One such example, the GEF kalirin-7, activates Rac1, as well as its downstream effector p21-activated kinase (PAK), a protein kinase. PAK phosphorylates and inactivates ADF/cofilin, which alters spine actin turnover and thereby alters spine morphology.

Morphological Changes Accompanying Synaptic Plasticity and Learning: LTP

It has long been speculated that lasting memories must be encoded as a quasi-permanent morphological change in the activated synapses. There has been an extensive search for correlations between morphology and changes in synaptic function during learning, with the hope that some spines would be shaped differently because they took part in the storage of some piece of critical information. An increase in the strength of the synaptic connection might be encoded morphologically as (1) an increase in the number of contacts between two cells, (2) an increase in the size of the spine head to make room for more receptors, or (3) a dilation of the spine neck, so that currents generated in the spine head more readily depolarize the

dendrite. There are at least some data indicating that all three changes accompany induction of LTP in brain slices or learning in whole animals. Most persuasively, recent advances in fluorescent microscopy, utilizing two-photon and laser-scanning instruments, for example, have made it possible to obtain high-resolution images of fluorescently labeled dendritic spines in living *ex vivo* brain slices at the same time as the induction and expression of LTP.

An increase in the number of synapses should result in an increase in spine density, and several studies have found strong evidence of an increase in the number of dendritic protrusions following LTP induction protocols in the CA1 pyramidal cells. In some studies these processes appear to be mature spines from the start. In others, they first resemble filopodia, only later maturing into spines, suggesting a recapitulation of the normal developmental sequence of events (**Figure 3(a)**). The induction of new dendritic processes shares some of the same requirements as does the induction of LTP and learning: induction is largely NMDAR dependent and requires activation of CaMK. These data are consistent with *in vivo* spine density changes.

Hebb's postulate states that learning a particular bit of information depends on the correlated action potential discharge of two cells. In order for the new spines to strengthen existing connections, they must be formed with presynaptic partners whose activity was correlated with the postsynaptic cell. Some data suggest that the new spines are formed by the splitting of existing pre- and postsynaptic partners, whereas other data suggest that new spines grow out and preferentially contact presynaptic terminals whose activity induced the new spine to form. Indeed, there is strong evidence that glutamate may act as a chemoattractant for growing dendritic processes, which then grow preferentially toward the active presynaptic terminal releasing glutamate.

There is equally compelling evidence indicating that LTP induction is accompanied by changes in the size of dendritic spines. Induction of LTP using photostimulation or tetanic synaptic stimulation, while visualizing spines loaded with fluorescent dyes, is accompanied by a 50% increase in spine head volume within 1–5 min after stimulation. The degree of spine head enlargement depends strongly on the initial, naive spine size (**Figure 3(b)**). Although almost all spines show a transient enlargement, only small spines become persistently increased in volume after stimulation with caged glutamate. Like LTP, spine head enlargement requires activation of NMDARs and CaMK. These data suggest that the spine head must become bigger in order to accommodate newly inserted AMPARs. Spine head growth may not be an obligatory concomitant of LTP, however, since it is not seen in all studies, despite equivalent LTP induction (**Figure 3(c)**).

Morphological Changes Accompanying Synaptic Plasticity and Learning: LTD

The strength of naive and previously potentiated synapses can be reduced in the process known as long-term depression, usually induced with prolonged periods of stimulation at moderate frequencies (1–10 Hz). The induction of LTD has recently been shown to be accompanied by morphological changes that are opposite to those accompanying LTP induction, including decreases in spine number and spine head size.

Loss of preexisting dendritic spines in response to tetanic stimulation reportedly occurs within 3 h of the stimulation and results in the loss of two to three spines per 100 μm of dendrite, with any one spine retracting over the course of 10 min to 1 h. Like LTD, this stimulation-induced spine retraction is prevented by blocking NMDARs. Interestingly, the loss of spines associated with LTD becomes progressively less pronounced with increasing age. The most spine loss is triggered in tissue slices from animals that are 12–20 days old. This finding supports the concept of a developmentally defined critical period in which morphological plasticity can most easily be induced, and that the loss of this plasticity contributes to brain maturation and stabilization.

Induction of LTD with low-frequency stimulation also triggers a decrease in spine head diameter of about 70%, which persists for >30 min. Induction of LTD thus triggers morphological changes that oppose those observed to accompany LTP. Indeed, low-frequency stimulation can reverse the increase in spine head volume triggered with high-frequency stimulation, and high-frequency stimulation can enlarge spine heads previously reduced by low-frequency stimulation.

Spine Shape and Cytoskeletal Rearrangement

The cytoskeleton of the dendritic spine is characterized by a loose network of actin filaments. Interestingly, dendritic spines lack microtubules, unlike the cytoskeleton in the dendritic shafts, except for the occasional microtubule in highly complex spines. It has been shown that the spine cytoplasm is composed of actin and actin-regulating proteins which are longitudinally situated in the spine neck and organized into a dense lattice surrounding the smooth

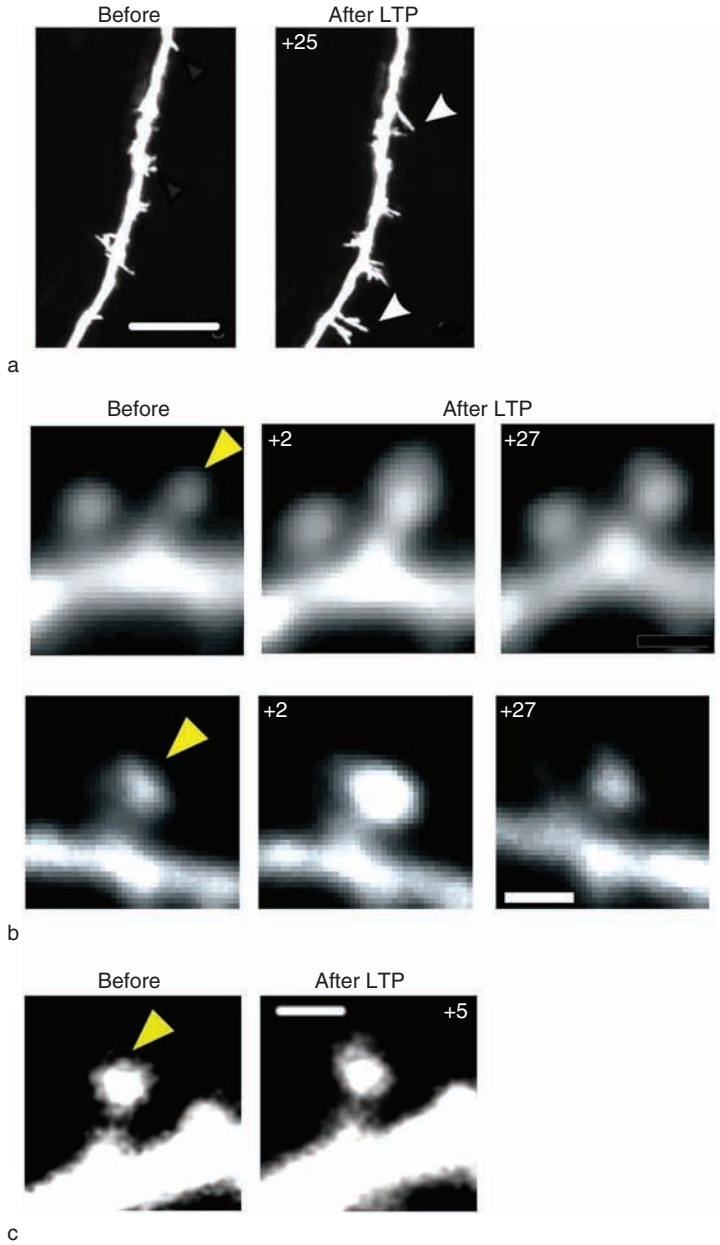

Figure 3 Spine changes in long-term potentiation (LTP). Images of fluorescent dye-filled dendrites of CA1 pyramidal cells before and after induction of LTP with synaptic stimulation (a) or photolysis of chemically caged glutamate (b, c). Time after induction, in minutes, is indicated in images. (a) LTP induction leads to outgrowth of dendritic filopodia, which are later replaced by spines. (b, c) Photostimulation was used to induce LTP at a single dendritic spine, indicated by a yellow arrowhead. (b) A rapid, transient enlargement of the spine head was seen in both examples, but persisted only in the small spine (upper row), not in the larger spine (lower row). (c) No enlargement of the spine head was seen, despite a similar level of potentiation. Scale bars = 10 μm (a), 1 μm (b, c). (a) Reprinted from Maletic-Savatic M, Malinow R, and Svoboda K (1999) Rapid dendritic morphogenesis in CA1 hippocampal dendrites induced by synaptic activity. *Science* 283: 1923–1927. (b) Reprinted from Matsuzaki M, Honkura N, Ellis-Davies GC, et al. (2004) Structural basis of long-term potentiation in single dendritic spines. *Nature* 429: 761–766. (c) Reprinted from Bagal AA, Kao JPY, Tang CM, et al. (2005) Long-term potentiation of exogenous glutamate responses at a single dendritic spine. *Proceedings of the National Academy of Sciences of the United States of America* 102: 14434–14439.

endoplasmic reticulum (SER) or spine apparatus in the head. This organization of the actin filaments suggests that they provide the scaffolding for the basic spine structure, as well as mediating changes in spine shape. Many actin-binding proteins, such as Arp2/3, cortactin, ADF/cofilin, drebrin, profillin, gelsolin, spectrin, and neurabin, have been identified within dendritic spines.

Increasing evidence suggests that inactivity extending for days induces changes in the structure and

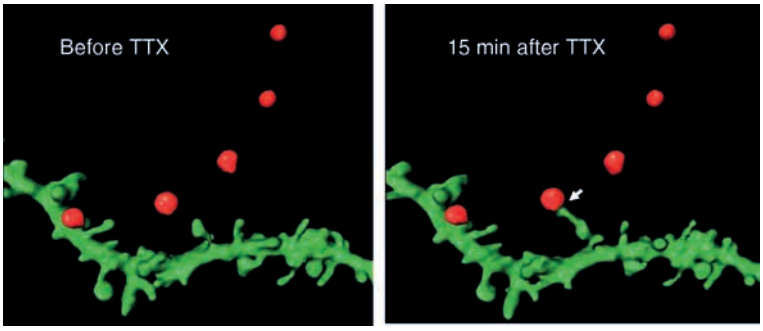

Figure 4 The formation of spine head protrusions. Tetrodotoxin (TTX; 1 μM) induces the formation of spine head protrusions. Example of a portion of dendrite double-labeled with membrane-bound green fluorescent protein (green) and FM 4–64 (red) at the beginning (left panel) and after a 15-min treatment with 1 μM TTX (right panel). Arrows indicate the location of a spine that converts into a spine head protrusion. Reprinted from Richards DA, Mateos JM, Hugel S, et al. (2005) Glutamate induces the rapid formation of spine head protrusions in hippocampal slice cultures. *Proceedings of the National Academy of Sciences of the United States of America* 102: 6166–6177.

function of hippocampal synapses, although the characteristics of these reported changes differ. For instance, long-term exposure of neurons to tetrodotoxin (TTX) leads to increases in spine density in the lateral geniculate nucleus, and in the number of dendritic filopodia on CA1 pyramidal cells. Under conditions of reduced synaptic input, spines establish new contacts with nearby boutons in a glutamate-receptor-dependent fashion, suggesting that this may occur in response to glutamate spillover (Figure 4). The formation of spine head protrusions is blocked by the AMPA kainate receptor antagonist 2,3-dihydroxy-6-nitro-7-sulfamoyl-benzo[*f*]quinoxaline-2,3-dione (NBQX), which prevents spines from sensing glutamate via AMPA receptors, and also by AMPA, which stabilizes spines due to the depolymerization of actin. These spine head protrusions 'reach out' to contact nearby boutons and appear to form new synapses. This is further strengthened by the observation that exogenously applied glutamate triggers the formation of spine head protrusions. Further investigation will determine whether these branched spines are forerunners to boutons splitting and new spines emerging. These findings suggest that if a spine has not been recently activated by glutamate released by its associated bouton, it can respond to glutamate released by a neighboring synapse. Furthermore, if there is too much activity or release of glutamate, the dendritic spines can retract and lose their contacts. Spines can thus 'compare' their individual recent history with the level of activity of neighboring synapses and modify hippocampal microcircuitry accordingly.

Conclusion

A variety of psychiatric and neurological diseases are associated with abnormalities in dendritic spine morphology or deficient spine density (notably, in virtually every disease in which cognitive performance is impaired). This spectrum includes epilepsy, stroke, trauma, schizophrenia, dementia, major depression, and chronic substance abuse, as well as normal aging. For the majority of these conditions it is not known whether spine abnormalities are a direct cause of synaptic dysfunction, resulting in cognitive deficits, or are secondary to some other event. Although still far from complete, the molecular understanding of the pathogenetic process seems to be most advanced in some hereditary forms of mental retardation, such as fragile X syndrome, in which proteins that regulate spine development and maintenance have been found to be affected.

A better understanding of the role dendritic spines play in determining synaptic efficacy and the expression of synaptic plasticity will provide a better insight into the relationship between cognitive deficits and changes in spine morphology, shape, and number in these various diseases.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axonal and Dendritic Identity and Structure: Control of; Dendritic Spine History; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; NMDA Receptors and Development; Postsynaptic Density/Architecture at Excitatory Synapses; Spine Plasticity; Ultrastructural Analysis of Spine Plasticity.

Further Reading

Araya R, Jiang J, Eisenthal KB, et al. (2006) The spine neck filters membrane potentials. *Proceedings of the National Academy of Sciences of the United States of America* 103: 17961–17966.

Bagel AA, Kao JPY, Tang CM, et al. (2005) Long-term potentiation of exogenous glutamate responses at a single dendritic spine.

Proceedings of the National Academy of Sciences of the United States of America 102: 14434–14439.

Bloodgood BL and Sabitini BL (2007) Ca(2+) signaling in dendritic spines. *Current Opinion in Neurobiology* 17: 345–351.

Bourne JN, Sorra KE, Hulbert J, et al. (2007) Polyribosomes are increased in spines of CA1 dendrites 2h after the induction of LTP in mature rat hippocampal slices. *Hippocampus* 17: 1–4.

Calabrese B, Wilson MS, and Halpain S (2006) Development and regulation of dendritic spine synapses. *Physiology (Bethesda)* 21: 38–47.

Holcman D, Korkotian E, and Segal M (2005) Calcium dynamics in dendritic spines, modeling and experiments. *Cell Calcium* 37: 467–475.

Holtmaat A, Wilbrecht L, Knott GW, et al. (2006) Experience-dependent and cell-type-specific growth in the neocortex. *Nature* 441: 979–983.

Matsuzaki M, Honkura N, Ellis-Davies GC, et al. (2004) Structural basis of long-term potentiation in single dendritic spines. *Nature* 429: 761–766.

Matus A (2005) Growth of dendritic spines: A continuing story. *Current Opinion in Neurobiology* 15: 67–72.

McKinney RA, Capogna M, Durr R, et al. (1999) Miniature synaptic events maintain dendritic spines via AMPA receptor activation. *Nature Neuroscience* 2: 44–49.

Park M, Salgado JM, Ostroff L, et al. (2006) Plasticity-induced growth of dendritic spines by exocytic trafficking from recycling endosomes. *Neuron* 52: 817–830.

Richards DA, Mateos JM, Hugel S, et al. (2005) Glutamate induces the rapid formation of spine head protrusions in hippocampal slice cultures. *Proceedings of the National Academy of Sciences of the United States of America* 102: 6166–6177.

Tada T and Sheng M (2006) Molecular mechanisms of dendritic spine morphogenesis. *Current Opinion in Neurobiology* 16: 95–101.

Zuo Y, Lin A, Chang P, et al. (2005) Development of long-term dendritic spine stability in diverse brain regions of cerebral cortex. *Neuron* 46: 181–189.

Glutamatergic and Gabaergic Systems

J-M Billard and B Potier, INSERM, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurotrophins (NTs), including brain-derived neurotrophic factor (BDNF), nerve growth factor (NGF), and neurotrophin-3 and neurotrophin-4 (NT-3, NT-4), were originally viewed as endogenously secreted glycoproteins that functioned in facilitating long-term neuronal survival and in stabilizing the phenotype of neurons. In the mammalian central nervous system (CNS), NTs have emerged in recent years as being more particularly involved in the modulation of activity-dependent processes, such as synaptic development and plasticity. The close apposition of NT signaling, which involves activation of the tropomyosin family of receptor tyrosine kinases (Trk) and the p75 NT receptors, with glutamate and γ-aminobutyric acid (GABA) synapses, as well as short- and long-term changes in excitatory and inhibitory synwaptic transmission induced by NTs, have identified this family of proteins as potent modulators of synaptic strength in the developing and mature brain. Given that the inability of neuronal networks to adapt is likely to impair cerebral function, it is not surprising that abnormal NT signaling has been linked to a wide range of neurological and psychiatric disorders. Herein, we review recent advances relative to the role of NTs in the developmental control and function of glutamatergic and GABAergic systems and discuss the possible mechanisms involving impaired trophic support in normal aging and pathological states, focusing on epilepsy and schizophrenia.

Developmental Regulation of Glutamatergic and GABAergic Systems by NTs

Proper function of the CNS results from the establishment of precise synaptic connections within neuronal networks during development. This occurs through a complex, multistep process that involves neurogenesis, migration, differentiation (dendritic and axonal growth), and synaptogenesis. Although NTs and their receptors are highly expressed in the developing brain during times of active neuronal growth and differentiation, their direct involvement in the maturation of excitatory synapse has remained elusive for a long time. BDNF facilitates the formation of excitatory synapses *in vitro* and promotes new synapses *in vivo*, but reduced levels of cerebral BDNF in mutant mice do not significantly affect synapse density, spine morphology, or synaptic vesicle distribution in cortical circuitry. NTs were initially proposed to indirectly act on excitatory connections through homeostatic mechanisms resulting from an alteration in inhibitory synapse formation, because GABAergic interneurons are established well before pyramidal neurons are, and GABA receptors and synapses are operative before glutamatergic synapses become operative during development in all brain structures and species studied to date. However, evidence for a direct effect of NTs on excitatory synapses has recently emerged using genetic manipulations or time-lapse imaging in dissociated cultures of cortical neurons. Knocking out the receptor TrkB specific for BDNF decreases the number of glutamatergic synapses, while conditional deletion of this receptor specifically at presynaptic or postsynaptic site clearly assigns a cell-autonomous role of TrkB signaling for normal development of both components of the hippocampal excitatory synapses. Visualization of the subcellular localization of TrkB receptors during synaptogenesis also provides convincing clues for a significant role of NTs as molecules dynamically involved in the formation of excitatory synapses between cortical neurons. During synapse maturation, TrkB receptors, which are highly mobile, concentrate in axonal growth cones and dendritic filopodia, thus actively participating in setting excitatory connections. At the postsynaptic level, BDNF, NT-3, and NT-4 control the length and complexity of dendrites of pyramidal neurons in slices of neonatal neocortex, and removal of TrkB receptors induces extensive dendritic retraction. Basal, but not apical, dendrites of neurons respond more strongly to a single NT, and increase or decrease in dendritic elongation and branching depends on cell layer. The growth and differentiation of dendritic processes induced by BDNF may be altered by NT-3 and reciprocally. This spectrum of neurotrophic actions and the laminar specificity of these effects in mammalian neocortex implicate endogenous NTs as regulatory signals in the development of specific patterns of dendritic trees of glutamatergic neurons.

Compared to glutamatergic systems, much more evidence indicates a link between NTs and many aspects of GABAergic interneuron development in the rodent CNS. Different proliferative regions of the subpallium telencephalon give rise to interneurons populating the neocortex and the hippocampus. At the earliest developmental stages, the medial ganglionic eminence (MGE) and the lateral ganglionic eminence (LGE) have been identified to be the sources of cortical and striatal GABAergic interneurons. The MGE is also the source of hippocampal interneurons,

but only those that will migrate to hippocampal layers of the cornus ammoni. Interneurons of the dentate gyrus (but also GABAergic cells of the cortex, striatum, amygdala, and nucleus accumbens) originate from the caudal ganglionic eminence (CGE). BDNF and NT-4 promote the migration of GABAergic precursors into the cortex and hippocampus. NT-4 is expressed along the migratory path of MGE-derived cells and increases the motility of stationary and migrating interneurons. BDNF and NT-4 stimulate tangential migration, whereas pharmacological inhibition or genetic deletion of TrkB receptors significantly decreases the number of cells migrating tangentially in the developing cortex. The effects of BDNF and NT-4 in MGE cells are mediated through the activation of the phosphatidylinositol 3-kinase (PI3K), inhibition of which also alters interneuron migration.

Once interneurons have migrated accurately to their final destination, they extend processes that form elaborate dendritic trees. As already mentioned for glutamatergic neurons, NTs promote dendritic elongation and branching in the developing interneurons. NTs not only induce the morphological but also the neurochemical maturation of hippocampal or neocortical interneurons. In cultured hippocampal neurons, exogenous BDNF application increases the level of glutamic acid decarboxylase 65 (GAD65), one of the two isoforms of GABA synthesis enzymes; GAD65 resides mainly in GABAergic terminals. BDNF also exhibits important effects on the neuropeptide expression that co-localizes with GABA within inhibitory interneurons. Application of NT-3 increases neuropeptide levels in neuropeptide Y (NPY)-containing and somatostatin-containing hippocampal interneurons, whereas NGF has no effect. BDNF and NT-4 enhance expression of NPY but exert no conspicuous effect on somatostatin content. Furthermore, levels of NPY are reduced in interneurons of BDNF-deficient mice.

A final crucial step in neuronal development is synaptogenesis. BDNF and NT-4 promote the formation of functional inhibitory synapses in cell cultures. *In vivo*, BDNF overexpression accelerates the maturation of GABAergic synaptogenesis in the visual cortex. Neuronal activity also enhances synaptogenesis of inhibitory connections, and antibodies against BDNF or NT-4 applied in cerebellar or hippocampal slice cultures prevent or reduce this activity-dependent maturation of GABAergic synapses. Finally, NTs optimize synaptic function of newly established synapses during development. At presynaptic level, NTs modulate the release of GABA. BDNF increases the vesicle release probability and vesicle replenishment with no effect on quantal size, the number of release sites, and the readily releasable pool. In addition, levels of several presynaptic proteins implicated in the mechanisms of release, such as synaptobrevin, synatophysin, and syntaxin, are upregulated. NTs also regulate the number of postsynaptic GABA receptors, and BDNF has been shown to increase the formation of $GABA_A$ receptor clusters at synapses.

Appropriate development of the GABAergic system is highly important for neonatal network activity, as GABA is the principal neurotransmitter that induces depolarization of the postsynaptic neurons up to postnatal day 10 in the rodent brain. The conversion of GABAergic transmission from excitatory to inhibitory is dependent on the progressive expression of the chloride extruder, the K^+/Cl^- cotransporter 2 (KCC2), occurring from embryonic stages to postnatal day 15 (P15). This operation of GABA as excitatory before and around birth may represent the first network activity necessary to drive the activity-dependent release of NTs, and therefore to play a role in the maturation of glutamatergic transmission.

Regulation of Glutamatergic and GABAergic Transmission and Plasticity by NTs

Brain expression of NTs and of their receptors is not limited to development; neurotrophic factors are still present and Trk receptors largely expressed in almost all neurons and inhibitory interneurons in adult CNS. These proteins therefore have functions clearly extending the maturation of synapses. An extensive body of literature indicates that NTs can modulate several distinct aspects of synaptic transmission, including glutamatergic and GABAergic transmission. NTs, both *in vivo* and *in vitro*, induce changes in glutamatergic synaptic transmission in the adult brain through a positive feedback modulation, which is mediated by the activation of Trk, but not p75, receptors. Brief application of NTs enhances the efficacy of synaptic transmission in a tetanized pathway without altering synaptic strength of an untetanized set of synapses monitored simultaneously, indicating that NTs act preferentially on active synapses. Reciprocally, the expression and release of NTs are tightly regulated by electrical activity. This close relationship between the level of neuronal activity and NT effects points out for a role of neurotrophic factors in long-lasting activity-dependent synaptic changes, such as long-term potentiation (LTP). LTP is an experience-dependent increase in synaptic strength involved in memory formation, which is experimentally induced by a high-frequency stimulation (HFS) of excitatory afferents, and it can last for weeks and months *in vivo*. Glutamatergic *N*-methyl-D-aspartate receptors (NMDARs) play a pivotal role in an early phase of LTP lasting for a couple

of hours, by rapidly elevating intracellular Ca^{2+} in postsynaptic dendritic spines, although other effectors such as voltage-gated Ca^{2+} channels or metabotropic glutamate receptors may also be involved. Early LTP depends on the phosphorylation of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionate subtype of glutamate receptors (AMPARs), already present at neuronal membranes and trafficking of new receptors at synapses, thus increasing the strength of synaptic transmission (Figure 1). Stronger induced HFS promotes longer-lasting LTP, referred to as late LTP or synaptic consolidation. Late LTP requires the activation of cyclic adenosine monophosphate (cAMP)-dependent protein kinase A (PKA) and extracellular signal-regulated kinase (ERK), acting through phosphorylation of the transcription factor cAMP/calcium response element-binding protein (CREB), leading to gene transcription and protein synthesis. Results from genetic deletion and pharmacological studies indicate multiple contributions of NT signaling to early and late LTP. First, NTs facilitate the induction of LTP through a permissive effect, which is characterized by the property of BDNF to prevent the decrease in synaptic transmission normally occurring during repetitive stimulations, namely synaptic fatigue. Following successive trains of stimulation, BDNF synthesis is stimulated in dendritic processes and is rapidly released from secretogranin II-positive secretory granules. This release evokes a TrkB receptor-dependent Ca^{2+} entry into presynaptic glutamatergic terminals through activation of voltage-gated Ca^{2+} channels, or through intracellular Ca^{2+} release from inositol 1,4,5-trisphosphate (IP_3)-gated Ca^{2+} stores. Phosphorylation of synapsin I, the membrane-associated protein of transmitter vesicles, by mitogen-activated protein kinases (MAPKs) is then promoted, which releases synaptic vesicles from the cytoskeleton and facilitates their entry into the exocytosis cycle. Additionally, NTs provide relieve of synaptic fatigue by promoting the docking of synaptic vesicles to presynaptic membranes.

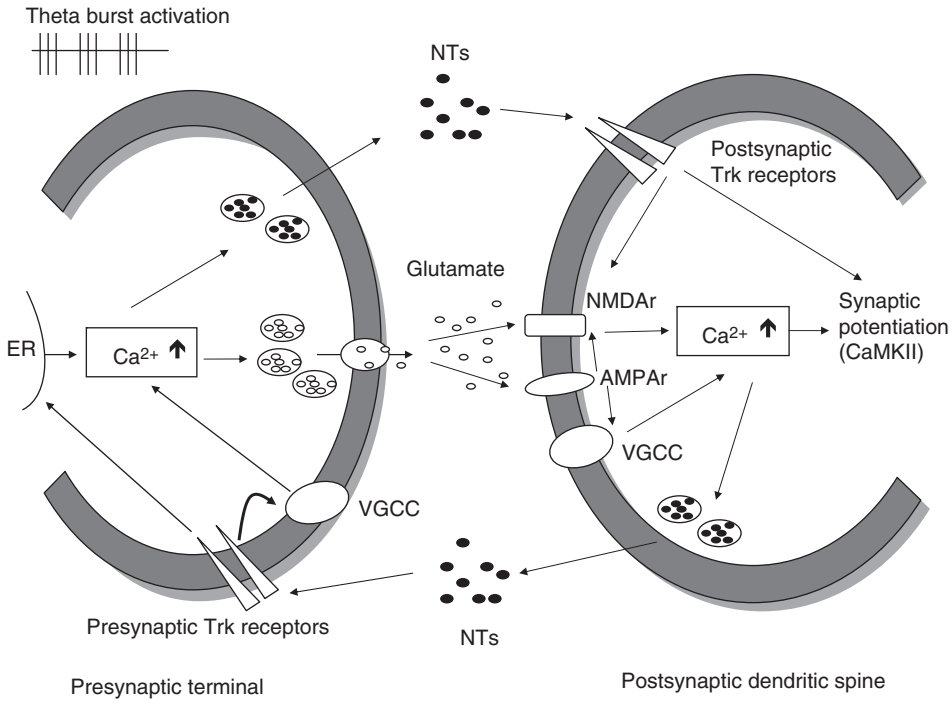

Figure 1 Neurotrophin (NT) signaling and synaptic potentiation. In response to repetitive presynaptic activation of afferent fibers (theta burst activation), glutamate is released and binds mainly on postsynaptic α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs). Membrane depolarization is generated, increasing intracellular Ca^{2+} concentration by Ca^{2+} entry through *N*-methyl-D-aspartate receptors (NMDARs) and voltage-gated calcium channels (VGCCs). The increase in intracellular Ca^{2+} activates intracellular pathways, such as the Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) pathway, inducing synaptic potentiation. In addition, it promotes the postsynaptic release of NTs into the synaptic cleft, where NTs bind to tyrosine Kinase (Trk) receptors located on glutamate fiber terminals. Activation of Trk receptors results in presynaptic increase in intracellular Ca^{2+} through the recruitment of VGCCs or Ca^{2+} release from the endoplasmic reticulum (ER), which facilitates subsequent release of glutamate from the vesicle pool. This mechanism allows NTs to relieve the synaptic fatigue that normally occurs during repetitive activation. On the other hand, increase in presynaptic intracellular Ca^{2+} also promotes release of NTs from presynaptic terminals. Activation of postsynaptic Trk receptors then enhances NMDAR activity, which reinforces the increase in postsynaptic intracellular Ca^{2+} and, consecutively, synaptic potentiation. It also directly modulates synaptic potentiation by locally facilitating the synthesis of proteins such as CaMKII.

Independently of preventing synaptic fatigue, NT signaling is also causally involved in the induction of LTP through an acute instructive effect. Brief application of BDNF induces a calcium rise in dendritic spines and a burst of action potentials in the cell body; these events trigger LTP when pairing with a weak HFS. However, these effects are prevented by NMDAR antagonists, suggesting that NTs are able to directly affect functional properties of this glutamatergic receptor subtype necessary for synaptic plasticity. Accordingly, BDNF, but not NT-3 or NGF, reversibly enhances the magnitude of the NMDAR-mediated synaptic currents, whereas AMPAR-mediated responses are not affected. This facilitation of NMDAR currents specifically involves the phosphorylation of the NR2B subunit through activation of the tyrosine kinase Fyn. In addition to a direct modulation of functional properties of NMDAR, activation of TrkB-mediated signaling also enhances the number of NMDAR clusters at postsynaptic membranes as well as the proportion of clusters apposed to presynaptic terminals.

Besides its modulation of early NMDAR-mediated LTP, NTs also interact with protein synthesis-dependent late LTP through a delayed instructive effect. Pharmacological and genetic manipulations of NT signaling alter synaptic consolidation. TrkB antibodies prevent development of late LTP but only when they are applied during a specific time delay window after LTP induction (between 30 and 60 min, and not later), indicating that TrkB signaling is critically involved at a specific step in the expression of synaptic consolidation. In the absence of any HFS, a persistent potentiation of synaptic transmission is induced by a long-lasting application of BDNF or NT-3 (but not NGF) in the hippocampus. This pharmacological form of LTP is occluded during late LTP, revealing that both NT-dependent LTP and synaptic consolidation share common mechanisms. Accordingly, BDNF-mediated LTP induced at medial perforant path–granule cell synapses *in vivo* also requires ERK activation and *de novo* gene expression. How NTs contribute to mechanisms of synaptic consolidation remains to be definitively determined. One possibility should be a role in synaptic tagging. HFS sets a synaptic tag that allows the capture of proteins required for late LTP. A recent postulate considers a critical role of TrkB receptor activation, promoting translation of newly synthesized immediate-early genes or regionally stored mRNAs, such as those for Ca^{2+}/calmodulin kinase II, and tagging dendritic spines that could locally enhance protein synthesis and generate a pool of proteins required for synaptic consolidation.

Modulation of synaptic strength by NTs in neuronal networks also involves changes in GABAergic transmission. The effects of NTs on inhibitory synaptic currents are age dependent, as they enhance GABAergic transmission in immature neurons but reduce inhibitory currents in adult CNS. BDNF and NT-3 acutely suppress GABAergic transmission in mature synapses, by decreasing the frequency of spontaneous GABAergic activity, indicating a presynaptic locus of action. BDNF inhibits GABA release via activation of TrkB receptors, but independently of Ca^{2+} entry in presynaptic GABA terminals through voltage-gated Ca^{2+} channels, and rather by altering the activity of GABA transporters.

In addition to presynaptic effects on GABA release, NTs also act via postsynaptic mechanisms. TrkB signaling modulates the clustering of postsynaptic $GABA_A$ receptors that controls a subsequent upregulation of NMDAR clusters. These NT-mediated effects on postsynaptic clustering participate in the balance of excitatory and inhibitory synaptic transmissions in mature neuronal circuitry. In adult isolated hippocampal neurons, application of GABA evokes a fast inward current that is inhibited by BDNF. Combined electrophysiology and cell surface staining indicate that this inhibitory effect depends on a weaker phosphorylation and a rapid internalization of $GABA_A$ receptors. In addition, BDNF downregulates Cl^- transport, and both BDNF and NT-4 alter the neuronal capacity of Cl^- extrusion by reducing the expression of KCC2, the K^+/Cl^- cotransporter that maintains an effective Cl^- gradient. In nociceptive dorsal root ganglion neurons in culture, NGF treatment facilitates desensitization and deactivation of $GABA_A$ receptor-mediated currents that underlie the algogenic action of the NT.

Results from long-term treatments with BDNF suggest that the modulation effects of NTs on $GABA_A$ receptor-mediated responses may be a complex mechanism. Indeed, chronic application of BDNF on cortical and hippocampal adult neurons enhances rather than decreases GABA release probability. In these conditions, AMPAR-mediated currents, but not NMDAR-triggered responses, are increased in neocortical GABAergic neurons in culture. In parallel, the amplitudes, but not frequencies, of spontaneous miniature excitatory postsynaptic currents are elevated in these interneurons. These findings indicate that chronic treatment with NTs may also enhance the input sensitivity of GABAergic neurons, facilitating their inhibitory function in the neocortex.

Contribution of NTs in CNS Pathology

In addition to trophic effects on target neurons, NTs appear to be part of a general mechanism for activity-dependent modification of synapses in the developing and adult nervous system. Thus, diseases of abnormal

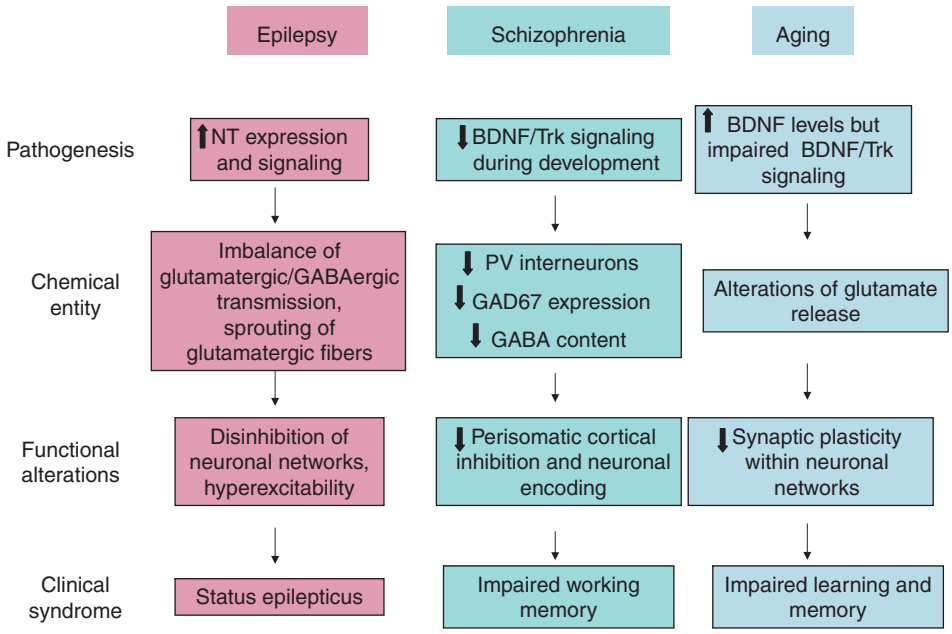

Figure 2 Schematic model outlining the possible contribution of abnormal neurotrophin (NT) signaling in epilepsy, schizophrenia, and aging. Increased NT signaling in response to various stimuli (e.g., ischemia, traumatic brain injury) may impair the glutamate/γ-aminobutyric acid (GABA) balance, leading to hyperexcitability of neuronal networks and ultimately to seizures. Decreased NT (BDNF, brain-derived neurotrophic factor; Trk, receptor tyrosine kinase) signaling in the prefrontal cortex during development may alter parvalbumin (PV)-containing GABAergic interneurons (and expression of glutamic acid decarboxylase isoform 67; GAD67), inducing an inhibition of pyramidal neurons and an impaired working memory, contributing to schizophrenia.

excitability (e.g., epilepsy and central pain sensitization), which are characterized by a disruption of the balance between excitatory and inhibitory synaptic transmissions, may be related to impaired NT signaling (Figure 2). On the other hand, synaptic plasticity, such as LTP, is known to underlie the acquisition and/or consolidation of certain forms of cognitive processes (e.g., learning and memory). Considering that synaptic plasticity is tightly regulated by NT signaling through multiple mechanisms, the question arises as to whether abnormal trophic support could be involved in cognitive deficits associated with normal aging and with brain pathologies such as schizophrenia.

A Role in Epilepsy?

Epilepsy is a cerebral disorder that is characterized by the recurrent appearance of spontaneous seizures due to neuronal hyperactivity. Since the discovery that limbic seizures increase NGF or BDNF expression, much has been learned about the contribution of abnormal NT signaling to the lasting structural and functional changes underlying epileptogenesis. The 'kindling' model of epilepsy – generated *in vivo* by repeated induction of pathological activity in the form of focal seizures – has identified a causal role for Trk receptors in epileptogenesis. Seizures are absent after hippocampal conditional deletion of TrkB receptors and a BDNF scavenger (TrkB–imunoglobulin G, which inhibits TrkB activation) suppresses kindling development. The exact nature of the endogenous ligand mediating Trk receptor activation during limbic epileptiform activities still remains to be determined. BDNF mRNA and protein are upregulated in areas implicated in epileptogenesis, while kindling development is reduced in mice in which one BDNF allele has been inactivated by gene targeting. In addition, epileptic seizures are enhanced in mice overexpressing BDNF. However, the results of these genetic manipulations must be regarded with caution, given potential developmental effects of altered BDNF levels. Indeed, only a modest effect on seizures is found in the hippocampal conditional BDNF knockout mice, despite a marked reduction of protein expression. Also, NT-4 null mutant mice are not affected with respect to any facet of the development or persistence of kindling. Endogenous NT-3 levels can influence the rate of epileptogenesis but NT-3 mRNA levels are significantly decreased in kindling. Evidence has been provided for a role of TrkA receptors in the development of epileptiform activities, since specific inhibition of TrkA–NGF binding and intraventricular infusion of anti-NGF antisera delay perforant path kindling progression.

The permanent hyperexcitability of the hippocampal network that generates status epilepticus occurs and is maintained through an imbalance between

glutamatergic-dependent excitation and GABAergic-mediated inhibition. Part of this imbalance may be dependent on altered NT signaling, since seizures have been shown to stimulate the expression of mRNA and protein for NTs and Trk receptors. Decrease in inhibition in hippocampal granule cells is reported in patients with medial temporal lobe epilepsy. A reduction of inhibition can trigger epileptiform activities or can sensitize the tissue to seizure activity, and may therefore represent one of the critical mechanisms allowing NTs to contribute to epileptic activity. It has been recently reported that seizure activity results in expression of synaptic $GABA_A$ receptors with an increased pool of α4 subunits; these are normally expressed only in extrasynaptic receptors with different functional properties. Early growth response factor 3 (Egr3) is the plasticity trigger for aberrant α4 subunit gene expression after status epilepticus. Interestingly, BDNF is the endogenous signal that induces Egr3 expression via a protein kinase C/MAPK-dependent pathway. This finding provides additional support for a role of NTs in epileptogenesis through disinhibition of neuronal networks.

NTs not only contribute to the maintenance of epileptic activity through changes in neuronal excitability but they also promote dendritic and axonal sprouting, morphological alterations, and synapse formation on a slower timescale. This abnormal remodeling, which reinforces the hyperexcitability of neuronal networks, is illustrated by the clear-cut sprouting of hippocampal mossy fibers in the epileptic brain. Mossy fiber sprouting is inhibited by either selective TrkA–NGF or p75–NGF antagonists, and it is also dependent on NT-3 signaling. On the contrary, BDNF administration has no effect on remodeling of the mossy fiber system in naive or kindled brain, suggesting that this NT acts on epileptogenesis by its effects on cell excitability.

A Role in Aging?

Aging is associated with defects in learning and memory, one of the most frequently noted features in the elderly. In early studies, cognitive impairments were linked to an age-dependent decline in forebrain cholinergic function, which was intensively investigated as being related to NT dysregulation. This hypothesis was eventually weakened, based on findings that age did not change the prevalence or regional distribution pattern of BDNF or TrkB mRNA throughout the life span. Learning and memory deficits in aged individuals are now known to be closely related to impaired induction of NMDAR-mediated LTP as well as of PKA-dependent synaptic consolidation. BDNF-induced LTP and the associated signaling are significantly impaired in the dentate gyrus of aged rats. In addition, recent evidence indicates that an increase in endogenous BDNF following treatment with a positive AMPAR modulator, AMPAkine, can reverse LTP deficit in aged rodents. Also, the stable LTP in AMPAkine-treated aged animals can be eliminated by infusion of the extracellular BDNF scavenger TrkB–Fc. These results indicate that managing BDNF signaling during aging can offset deficits of synaptic plasticity and strongly suggest that NTs may significantly contribute to age-related memory impairments. Accordingly, frontal cortex NGF and BDNF levels are positively, and hippocampal NT-3 levels are negatively, correlated with number of errors made as working memory load increases in aged animals, thus confirming that cerebral NTs may relate to working memory proficiency in aging.

A Role in Schizophrenia?

Schizophrenia is a devastating illness that affects approximately 1% of the world's population. Impairments of specific cognitive functions, such as prefrontal cortex-mediated working memory, are core features of schizophrenia. One of the most favored hypotheses for the etiology of schizophrenia is the neurodevelopmental hypothesis. Preclinical and clinical data indicate that dysfunction of NGF, BDNF, and NT-3 signaling may contribute to impaired brain development, neuroplasticity, and synaptic dysfunction, leading to the schizophrenic syndrome. Recent findings indicate that impairment in GABA-mediated inhibition in the prefrontal cortex may provide a critical mechanism for the disturbances in working memory. Postmortem tissue concentrations of the mRNA that encodes the 67 kDa isoform of glutamic acid decarboxylase (GAD67) are reduced in prefrontal cortex of individuals with schizophrenia, whereas mRNA and protein levels of GAD65 remain unchanged. This specific alteration may be critical, considering that genetic deletion of the GAD65 gene in mice does not alter cortical levels of GABA, whereas loss of the GAD67 gene is associated with a marked decrease in cortical GAD activity and GABA content. Because signaling through TrkB receptors stimulates the expression of GAD67, the reduced mRNA and protein levels for both BDNF and TrkB receptors that occurs in the prefrontal cortex of individuals with schizophrenia may thus contribute to the development of the illness through impaired GABAergic activity. Not only the mRNA and protein levels of GAD67 are altered in schizophrenia, but also the expression of parvalbumin (PV), a calcium-binding protein specifically localized in certain types of GABAergic cells. This effect is selective for PV-containing interneurons, as TrkB receptors are predominantly expressed by GABA neurons that contain PV, but not by calretinin-containing GABAergic interneurons. Since

stimulation of TrkB signaling increases the expression of PV, it has been suggested that deficiency in signaling through TrkB receptors possibly alters GABA synthesis in the PV-containing inhibitory interneurons in the prefrontal cortex of patients, thus disrupting perisomatic inhibition of pyramidal cells. This could induce a diminished capacity of the neuronal encoding that is required for working memory, thus contributing to the disease. In contrast to BDNF and TrkB signaling, the mRNA expression level for TrkC, the receptor tyrosine kinase for NT-3, is unchanged in schizophrenia. Therefore, alterations in the BDNF/TrkB system expression may specifically underlie the reduction of GABA-related genes (GAD67; GAT-1, the GABA transporter; and PV) observed in the prefrontal cortex in schizophrenia.

Conclusions

Many roles of NTs in glutamatergic and GABAergic systems have been elucidated. These roles include development, neuronal survival, maintenance of differentiated neuronal phenotype, regulation of synaptic connections, activity-dependent synaptic transmission, and plasticity in the adult brain. Evidence has accumulated indicating that alteration of one or several of these roles significantly contributes to the impaired brain functions that occur during normal aging, leading to cognitive defects, as well as to the impaired brain functions in pathological states such as schizophrenia and epilepsy. Interestingly, these roles have been complicated by the recent discovery that unprocessed forms of NTs (the pro-NTs) often induce effects that oppose those of mature NTs. Interaction of mature NTs with Trk receptors leads to cell survival and synaptic potentiation, whereas binding of pro-NTs to p75 NT receptors leads to apoptosis and synaptic depression. To what extent the proteolytic cleavage of pro-NTs determines the direction of NT regulation therefore represents a new challenge for scientists in the search of pharmacological treatments reducing NT-dependent brain pathologies.

See also: Autonomic Neuroeffector Junction; BDNF in Synaptic Plasticity and Memory; Calcium-Calmodulin Kinase II (CaMKII) in Learning and Memory; Cotransmission; Gamma-Aminobutyric Acid (GABA); Glutamate; Neurotransmitters and Growth Factors: Overview; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Neurotrophins: Physiology and Pharmacology; Synaptic Depression.

Further Reading

Bimonte HA, Nelson ME, and Granholm AE (2003) Age-related deficits as working memory load increases: Relationships with growth factors. *Neurobiology of Aging* 24: 37–48.

Binder DK, Croll SD, Gall CM, et al. (2001) BDNF and epilepsy: Too much of a good thing? *Trends in Neuroscience* 24: 47–53.

Bramham CR and Messaoudi E (2005) BDNF function in adult synaptic plasticity: The synaptic consolidation hypothesis. *Progress in Neurobiology* 76: 99–125.

Chao MV (2003) Neurotrophins and their receptors: A convergence point for many signalling pathways. *Nature Reviews Neuroscience* 4: 299–309.

Elmariah SB, Crumling MA, Parsons TD, et al. (2004) Postsynaptic TrkB-mediated signaling modulates excitatory and inhibitory neurotransmitter receptor clustering at hippocampal synapses. *Journal of Neuroscience* 24: 2380–2393.

Gomes RA, Hampton C, EL-Sabeawy F, et al. (2006) The dynamic distribution of TrkB receptors before, during, and after synapse formation between cortical neurons. *Journal of Neuroscience* 26: 11487–11500.

Gooney M, Messaoudi E, Maher FO, et al. (2004) BDNF-induced LTP in dentate gyrus is impaired with age: Analysis of changes in cell signaling events. *Neurobiology of Aging* 25: 1323–1331.

Gorski JA, Zeiler SR, Tamowski S, et al. (2003) Brain-derived neurotrophic factor is required for the maintenance of cortical dendrites. *Journal of Neuroscience* 23: 6856–6865.

Hashimoto T, Bergen SE, Nguyen QL, et al. (2005) Relationship of brain derived neurotrophic factor and its receptor TrkB to altered inhibitory prefrontal circuitry in schizophrenia. *Journal of Neuroscience* 25: 372–383.

Huang EJ and Reichardt LF (2001) Neurotrophins: Role in neuronal development and function. *Annual Review of Neuroscience* 24: 677–736.

Lessmann V, Gottmann K, and Malcangio M (2003) Neurotrophin secretion: Current facts and future prospects. *Progress in Neurobiology* 69: 341–374.

Lu B (2003) BDNF and activity-dependent synaptic modulation. *Learning and Memory* 10: 86–98.

Lu B, Pang PT, and Woo NH (2005) The yin and yang of neurotrophin action. *Nature reviews Neuroscience* 6: 603–614.

Luikart BW, Nef S, Virmani T, et al. (2005) TrkB has a cell-autonomous role in the establishment of hippocampal Schaffer collateral synapses. *Journal of Neuroscience* 25: 3774–3786.

McAllister AK, Katz LC, and Lo DC (1997) Opposing roles for endogenous BDNF and NT-3 in regulating cortical dendritic growth. *Neuron* 18: 767–778.

Rex CS, Lauterborn JC, Lin CY, et al. (2006) Restoration of long-term potentiation in middle-aged hippocampus after induction of brain-derived neurotrophic factor. *Journal of Neurophysiology* 96: 677–685.

Singh B, Henneberger C, Betances D, et al. (2006) Altered balance of glutamatergic/GABAergic synaptic input and associated changes in dendrite morphology after BDNF expression in BDNF-deficient hippocampal neurons. *Journal of Neuroscience* 26: 7189–7200.

Wirth MJ, Brun A, Grabert J, et al. (2003) Accelerated dendritic development of rat cortical pyramidal cells and interneurons after biolistic transfection with BDNF and NT4/5. *Development* 130: 5827–5838.

Woo NH and Lu B (2006) Regulation of cortical interneurons by neurotrophins: From development to cognitive disorders. *The Neuroscientist* 12: 43–56.

Xu B, Zang K, Ruff NL, et al. (2000) Cortical degeneration in the absence of neurotrophin signaling: Dendritic retraction and neuronal loss after removal of the receptor TrkB. *Neuron* 26: 233–245.

Zakharenko SS, Patterson SL, Dragatsis I, et al. (2003) Presynaptic BDNF required for a presynaptic but not postsynaptic component of LTP at hippocampal CA1–CA3 synapses. *Neuron* 39: 975–990.

Glycine Receptors: Molecular and Cell Biology

C Vannier and A Triller, INSERM U789, Ecole Normale Supérieure, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The amino acid neurotransmitters glycine and γ-aminobutyric acid (GABA) mediate inhibition in the vertebrate mature central nervous system by activating chloride conductance. Their ionotropic receptors belong to the canonical nicotinic acetylcholine receptor superfamily mediating fast synaptic transmission. Two features of this family, also known as the Cys-loop receptors, are its members' heteropentameric structure delineating the channel pore, and their assembly from distinct gene products or subunit splice variants that all possess a strictly conserved disulfide loop in their extracytoplasmic domain. Upon transient neurotransmitter binding, a pore opens in the pentamer that is responsible for the selective movement of ions through the plasma membrane according to their electrochemical gradient. In fully mature neurons, glycine binding to its receptor mediates inhibition by eliciting an influx of chloride ions. Whereas GABAergic transmission is widespread in the nervous system, glycine-mediated inhibition and the glycine receptor (GlyR) are predominantly encountered in the brain stem and spinal cord.

In the mid-1980s, GlyR was the first CNS receptor shown, by Triller and co-workers, to accumulate at postsynaptic differentiations (PSDs), in front of presynaptic release sites. Later on, variable subcellular organizations were demonstrated for many other receptors, including $GABA_A$ receptors (GABARs), which were analyzed by electron microscopy. Inhibitory synapses are intriguing, as both glycine and GABA open chloride channels and have a similar effect on the postsynaptic membrane. In the spinal cord, as well as in the brain stem and cerebellum, these synapses are either GABAergic, glycinergic, or mixed (i.e., respond to both neurotransmitters). The respective receptors accumulate in microdomains enriched with GABAR, GlyR, or a mixture of both, and it is the identity of the presynaptic element which imposes the composition of the postsynaptic membrane.

Another protein, named gephyrin, has been copurified with GlyR on an affinity column. Antisense experiments and work with knockout mice demonstrate that gephyrin is required for the synaptic localization of both GlyR and GABAR. The GABAR γ2 subunit is influential in synaptic insertion, since its absence in mouse results in a reduced number of GABA receptor clusters at synapses. The γ2 subunit role is mediated through gephyrin, but no direct link between GABARs and gephyrin has been demonstrated.

Molecular Biology and Topology of the Glycine Receptor

GlyR was the first neurotransmitter receptor to be isolated from the mammalian brain more than two decades ago. The purification of the native molecule was achieved in H Betz's laboratory, using affinity chromatography and taking advantage of the high-affinity interaction ($K_D \approx 1$–10 nM) of the receptor with its main antagonist, the well-known convulsant plant alkaloid, strychnine. The receptor appeared to be a complex of three proteins of 48 (α), 58 (β), and 93 kDa, in various vertebrate species.

Several biochemical studies of purified (or the membrane-anchored) form of GlyR have established that the integral membrane glycoproteins α and β represent the constitutive subunits of the receptor. Cross-linking techniques showed that α and β assemble to build up the channel-containing transmembrane core of GlyR. The size of the complex (250 kDa) logically suggested a pentameric assembly of the subunits, a quaternary structure well established for other Cys-loop receptors (Figure 1).

Unlike the α and β subunits, the 93 kDa co-purifying protein, gephyrin, is a nonglycosylated polypeptide able to interact reversibly with the α/β pentamer. Gephyrin is a cytoplasmic extrinsic membrane protein which binds tubulin dimers *in vitro* in a cooperative interaction of high affinity in the nanomolar range. It can also link brain microtubules to GlyR in a copolymerization assay. This finding underpinned the important notion that gephyrin could be a physical link between microtubules and GlyR in the postsynaptic membrane, as discussed in the following sections.

The Glycine Receptor as a Member of the Ligand-Gated Ionotropic Receptor Family

Analysis of the primary structures of the α and β proteins revealed the high homology of GlyR subunits, with amino acid sequence identity close to 50%. The GlyR subunit domain organization is comparable to that of other subunits of the Cys-loop receptor family. Hydropathy analysis of the mature polypeptide predicted a conserved region of four hydrophobic segments (M1–M4) able to cross the membrane bilayer at positions identical to those of transmembrane segments of acetylcholine and GABAR subunits (Figure 1(a)). GlyR subunits are polytopic type I membrane proteins. In contrast to M4, the

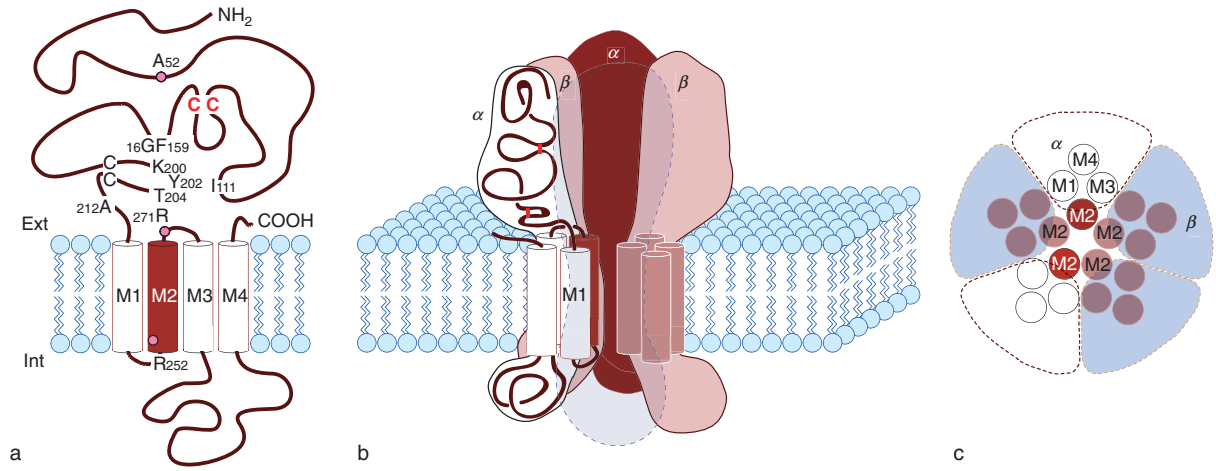

Figure 1 Structure of the glycine receptor. (a) Membrane topology of the α subunit, indicating the position of disulfide bridges, including the common Cys-loop delineated by Cys138 and Cys152 (red). Amino acid residues functionally important for agonist binding or modulation of the channel activity are shown. (b) Arrangement of α and β subunits in the mature $\alpha_2\beta_3$ receptor. One α subunit is depicted in cross-section, and one β subunit is rendered transparent to show the four packed M1–M4 transmembrane regions (cylinders). The pore-forming domain M2 of the α subunit is dark brown. (c) The α/β complex viewed from the extracellular space along the fivefold symmetry axis. Shown are the N-terminal regions of the five protomers (dashed lines) the counterclockwise organization of the four transmembrane domains.

transmembrane domains M1–M3 are highly conserved among the various subunits of GlyR. All subunits possess an extended N-terminal domain, bearing potential glycosylation sites, which were detected early by immunocytochemistry in the synaptic cleft. Finally, a large region of significantly lower homology is represented by a hydrophilic loop between the M3 and M4 segments; this loop protrudes into the cytoplasm (M3–M4 loop). Thus GlyR exhibits the molecular design and transmembrane topology of the acetylcholine receptor superfamily, based on four membrane-spanning domains able to form α helices (M2) and/or β strands (M1, M3, M4), and is assumed to function as an allosteric protein.

Our knowledge of this structure arose from extensive biochemical and electron microscopy data on the nicotinic acetylcholine receptor; these findings were recently refined by electron diffraction studies. The tertiary and quaternary structure of the soluble pentameric acetylcholine-binding protein (AChBP), which shares 20–24% and 17% sequence homology with nAChR and GlyR, respectively, has been useful in confirming or determining structure–function parameters for receptors of the family. Aside from an understanding of glycine and strychnine binding (see later), AChBP crystal structure provides a powerful model of the N-terminal domain of the Cys-loop receptor family. It shows that it is mainly a sandwich of antiparallel β sheets positioning conserved residues in order to stabilize the protomers, whereas variable residues are at their interface.

The topology of the transmembrane domains (Figure 1) delineates the ion permeation pathway away from the hydrophobic core of the phospholipid bilayer. For all Cys-loop receptors, the subunit's α-helical M2 domain lines the central water-filled pore, while M1, M3, and M4 form the interface with the lipids and isolate M2 from a hydrophobic environment. Recent studies on the GlyR α1 subunit have challenged the four-helix model and provide evidence that whereas M2 and M4 are entirely helical, M1 and M3 also contain β strands.

Assembly of GlyR

Initial studies of purified and membrane-anchored GlyR had suggested that α1 and β subunits assemble as a pentameric complex of α3β2 stoichiometry (Figures 1(b) and 1(c)). Short amino acid sequences (named assembly boxes), all located in the N-terminal extracytoplasmic domain of the β subunit and corresponding to three diverging motifs in α and β subunits, have been identified as having a role in precisely governing receptor assembly. Replacement of these motifs in the β subunit by the corresponding α1 motifs results in the loss of the subunit ratio in α/β oligomers, suggesting that different amino acid positions are determinants in the early step of subunit–subunit interaction. On the other hand, α1 and α2 co-assemble at variable subunit ratios. Importantly, these residues impose a mutually exclusive mode of assembly, either in complexes of invariant α/β stoichiometry or in homo-oligomers. A different stoichiometry (α2β3) has recently been established using affinity purification of expressed engineered tandem subunits. It confers on the β subunit a dominant role in the agonist binding properties of hetero-oligomeric α1/β GlyR, since ionic

interactions at the α/β interface are required to stabilize glycine. Such a structure for GlyR is consistent with the concept that neurotransmitter binding in the Cys-loop family requires extracellular segments from two adjacent subunits.

Molecular Diversity and Expression of Subunit Isoforms

The initial discovery of glycine receptor subtypes resulted from biochemical and immunochemical studies during rat spinal cord development. A comparison of GlyR immunoreactivities in neonatal and adult membrane extracts revealed a distinct form of GlyR predominating around birth (neonatal receptor, GlyRN) that displays a relatively low strychnine-binding affinity and only one major polypeptide band of 49 kDa. The discovery of functional $\alpha2$ pentamers predominantly found during the fetal and neonatal stages supported the idea of a developmental switch from GlyRN to the adult form, GlyRA, corresponding to the formation of $\alpha1/\beta$ heteromers. This switch, taking place within 3 weeks after birth, is not as complete in other regions of the nervous system where mixtures of $\alpha1$, $\alpha2$, and β subunits remain throughout adulthood. One current view is that postsynaptic GlyR corresponds to a mixture of $\alpha2$ pentamers and $\alpha2/\beta$ heteromers in immature neurons, while $\alpha1/\beta$ heteromers predominate in mature synapses.

A greater diversity of GlyR subunits than anticipated was established and includes not only the original $\alpha1$ and $\alpha2$ but also the $\alpha3$ and $\alpha2^*$ subunits. A murine $\alpha4$ subunit has also been identified which is absent in rat and human. These isoforms are highly homologous, with amino acid sequence identities on the order of 80%, and display a very low degree of interspecies variation (identity close to 99%). These isoforms can form glycine-gated chloride channels of comparable strychnine sensitivity, except $\alpha2^*$, which is more than 99% identical to human $\alpha2$ but exhibits a $\approx$500-fold lower strychnine sensitivity. Further generation of diversity resides in alternative exon usage. For the $\alpha1$ subunit, a variant mRNA encodes the form denoted '$\alpha1$ins,' a subunit containing eight additional amino acids. This sequence, inserted in the M3–M4 loop, bears a serine residue, representing a potential phosphorylation site and allowing a functional modulation of the $\alpha1$ subunit. Alternative splicing was also found for the $\alpha2$ and $\alpha3$ subunits.

The distribution of individual GlyR subunits was determined using *in situ* hybridization. However, transcription is not necessarily correlated with surface expression of functional receptors and the identification of the actual composition of GlyR in particular areas is not yet firmly established. The gene encoding the $\alpha1$ subunit and that encoding the $\alpha3$ subunit, though to a lesser extent, are mainly transcribed in spinal cord and brain stem at later postnatal stages. The $\alpha3$ subunit mRNA is present in the infralimbic system, the hippocampal complex, and the cerebellar granular layer. Expression levels of the $\alpha2$ subunit, by contrast, are high in the embryonic and perinatal stages but barely detectable in the adult brain, with some expression in higher cortical regions. The $\alpha4$ subunit, which is expressed at low levels in the adult but forms functional GlyRs, is restricted to the spinal cord and the sympathetic nervous system. The β subunit mRNA is more widely expressed throughout the embryonic and adult nervous system as compared to α subunit transcripts, and is found in brain loci devoid of strychnine-binding sites or GlyR immunoreactivity. The physiological significance of this widespread expression of the β subunit transcript is not understood, as so far there is no evidence that it assembles with α subunits of other receptors.

Gephyrin and Gephyrin-Binding Proteins

The Gephyrin Polypeptide

Gephyrin is the best characterized GlyR-interacting protein. Biochemical characterization and heterologous expression have shown that gephyrin binds cooperatively to tubulin dimers with a high affinity, and to GlyR via an 18-amino-acid sequence in the β subunit M3–M4 cytoplasmic loop. It was therefore postulated that gephyrin is a linker of the receptor to microtubules, adapted to anchor GlyR at synapses via the cytoskeleton. Gephyrin exists as several splice variants, as indicated by the characterization of the mouse gephyrin single gene and of its exonic structure, and the cloning of several different full-length coding sequences. Gephyrin variants, differing by the presence of distinct nucleotide sequences termed cassettes, are differentially expressed in numerous tissues, including brain, heart, skeletal muscle, and kidney, liver, lung, spleen, or testis, illustrating the ubiquitous nature of gephyrin.

Tertiary and Quaternary Structures of Gephyrin

The gephyrin polypeptide originated during evolution from the fusion of two genes of bacterial origin, respectively encoding the MogA and MoeA proteins involved in the biosynthesis of the molybdenum cofactor in *Escherichia coli*. These proteins are homologous to the gephyrin N- and C-terminal domains (G- and E-domains), respectively, flanking a 170-residue unique linker region. Intriguingly, in gephyrin these domains are enzymatically functional, and

disruption of the gephyrin gene leads to molybdoenzyme activity deficiency. These results indicate that gephyrin maintains MogA and MoeA activities in various tissues. Gephyrin tertiary and quaternary structures were determined following the crystallization of MogA and MoeA, and of gephyrin G- and E-domains. The G- and E-domains are compact structures forming stable trimers and dimers, respectively, which can be detected in crystals. If conserved in the gephyrin molecule, these oligomerization properties allow the construction of a two-dimensional hexagonal gephyrin lattice. This oligomerization pattern confers on gephyrin its function as a scaffold core protein and its ability to recruit GlyR at postsynaptic localizations. The GlyR-binding activity of gephyrin has been assigned in the E-domain to a pocket adjacent to the dimer interface.

Gephyrin-Interacting Proteins

Gephyrin is the core of the inhibitory postsynaptic scaffold (Figure 2). Binding partners for gephyrin other than GlyR have recently been identified, but their exact functions in receptor postsynaptic clustering have not yet been determined. Gephyrin may also be implicated in the modulation of signaling pathways. Gephyrin could thus participate in translational controls by binding to the RAFT1 kinase (a rapamycin and FKB12 target protein). The dynein light chains 1 and 2 (Dlc1/2) were recently identified as gephyrin-binding proteins. Further, gephyrin interaction with collybistin, a GDP/GTP exchange factor for GTPases of the Rho/Rac family, or with proteins such as profilin and Mena/VASP, could be a determinant for microfilament organization. This hypothesis

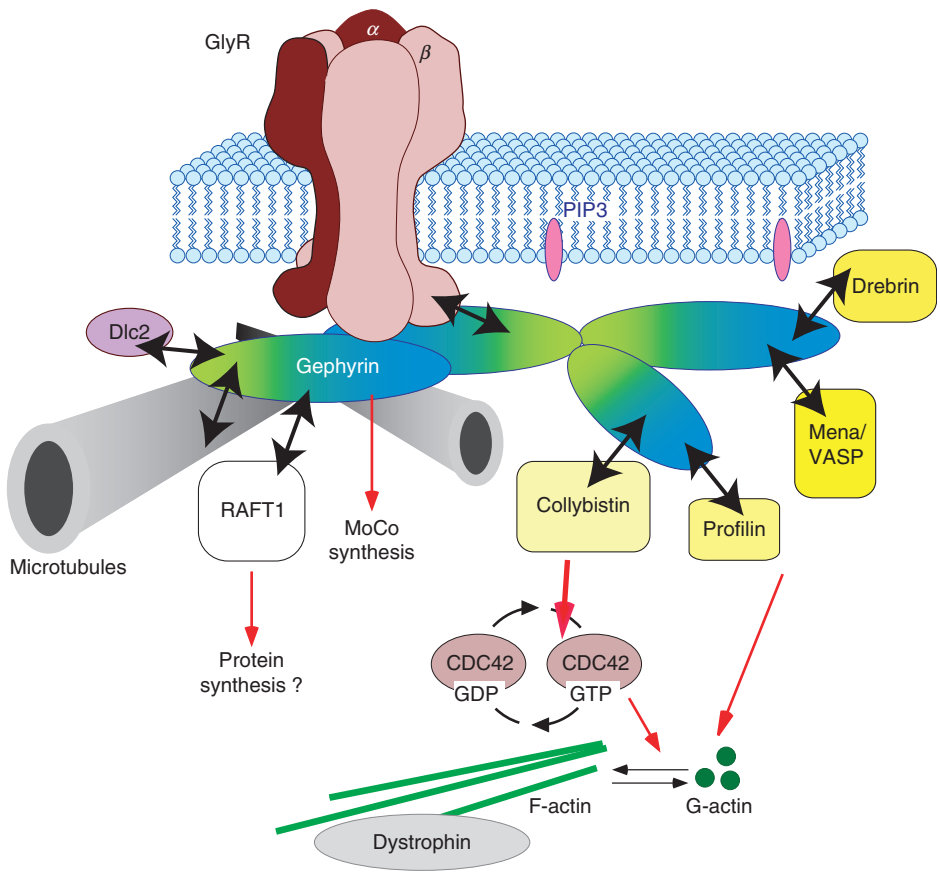

Figure 2 Molecular partners of the glycine receptor (GlyR)/gephyrin complex. Gephyrin is the only glycine receptor-interacting protein known so far. At synapses, it could form intermolecular trimers and dimers via its G- and E-domains, respectively. The localization of the receptor at synaptic loci relies on the ability of the former to interact directly with microtubules and indirectly with other proteins (collybistin and actin-binding proteins; yellow) able to modulate the dynamics of microfilaments. Until now the existence at synapses of a complex built up via simultaneous interactions involving all of these proteins has not been demonstrated. Double arrows indicate interactions disclosed by biochemical or two-hybrid strategies. A rapamycin and FKB12 target protein (RAFT1) is a translational activator which could mediate the synthesis of proteins required at synapses (Dlc2, dynein light chain 2; MoCo, molybdenum cofactor). Membrane binding sites for some partner proteins such as collybistin and profilin can be provided by phosphatidylinositol 3,4,5-trisphosphate (PIP3). Functions fulfilled by components of the proposed interaction network are shown by red arrows.

is consistent with results supporting the role of microfilaments in the control of postsynaptic gephyrin–GlyR co-cluster density. A feature of gephyrin is the presence of sequence stretches predicted to be WW (class IV) domain interaction motifs (ELM database search). A WW domain which is encountered in dystrophin/utrophin polypeptides might mediate phosphorylation-dependent binding to gephyrin. Interestingly, dystrophin, which is co-localized with dystroglycans, is present in a subset of inhibitory GABAergic synapses in the hippocampus, where it could be involved in a link between the extracellular matrix and the cytoskeleton.

Glycine Receptor Dynamics

Postsynaptic membrane domains and associated PSDs have long been viewed as stable multimolecular assemblies ensuring that a specific set of functions is performed locally. However, in recent years the notion has emerged that the postsynaptic domain is a highly dynamic protein complex adapted to the mechanisms underlying the plasticity of synaptic contacts. On the one hand, as for cellular multimolecular assemblies, postsynaptic complexes that contain GlyR correspond to the equilibrium between the exo- and endocytic delivery routes for membrane proteins. On the other hand, interactions involving cytoplasmic scaffolding proteins are pivotal in this steady-state balance that can be modified by cell differentiation or receptor activity.

Activity and GlyR Stability at Synapses

Receptor stabilization and aggregation, mediated by interactions with scaffold elements, are coupled at synaptic loci. The formation of GlyR clusters retained in the postsynaptic membrane in relation with receptor activation has been studied using an activity blockade by strychnine of cultured spinal cord neurons. This blockade leads to a reversible intracellular redistribution of GlyR which disappears from synaptic loci without alteration of gephyrin. Therefore, aggregation of gephyrin and that of GlyR are governed by distinct mechanisms. GlyR redistribution results from a defect in biosynthetic routing. At the steady state, the spinal cord neurons incorporate only a minor fraction of newly synthesized GlyR in synapses. It turns out that the stability of glycinergic synapses is due to a tight control of the relative rates of receptor synthesis, synaptic anchoring, and degradation. In order to explain selective positioning of active GlyR in front of glycinergic terminals, it has been proposed that glycine, which depolarizes immature neurons, triggers a calcium influx promoting the assembly of the postsynaptic scaffold.

Ubiquitination has recently been proposed to account for the regulation of cell surface GlyR $\alpha 1$ subunit endocytosis and degradation, thus determining postsynaptic receptor density. This hypothesis is based on the demonstration that ubiquitination of homopentameric GlyR occurs almost exclusively at the plasma membrane rather than during intracellular transport of the receptor.

Routing of Newly Synthesized GlyR to Synapses

The mechanisms of selective accumulation of receptors at synaptic sites are not fully understood. However, newly synthesized homomeric receptors spontaneously form microaggregates devoid of gephyrin in the neuronal plasma membrane. Although they can also in part remain in the extrasynaptic space, they associate with endogenous synaptic gephyrin according to a mechanism depending on the formation of the inhibitory innnervation. Interestingly, newly synthesized GlyR molecules are inserted in the plasma membrane as stable clusters which initially appear in the extrasynaptic space of the soma and dendritic proximal segments. Subsequently, they are redistributed over the dendritic tree. This occurs by diffusion of cell surface molecules, but not via exocytosis of vesicles routed to dendrites. Therefore, GlyR postsynaptic accumulation results from a diffusion/retention mechanism. This posits the crucial role of the retention signal, and in turn the anchoring role of gephyrin emerges. The complexity of the mechanisms of postsynaptic anchoring process, which relies on the exocytosis site, diffusion velocities, and on the concentration of synaptic binding sites, is illustrated by glutamate receptors of different compositions which accumulate at excitatory synapses with different apparent rates.

Gephyrin associates with vesicular compartments of GlyR intracellular trafficking and is a mediator of the insertion of newly synthesized receptor molecules in the plasma membrane, a process dependent on microtubule integrity. It has been further proposed that a nonsynaptic gephyrin/GlyR complex can be recruited to microtubules by dynein for retrograde transport along neurites. The existence of intracellular gephyrin/GlyR complexes indicates a dual function in trafficking and in assembly of postsynaptic scaffolds, as for proteins of excitatory synapses.

Local dendritic insertion of newly synthesized receptors is also a means of changing receptor numbers at synapses via mRNA localization. The discovery of the dendritic localization of GlyR α subunit mRNAs supported the view that local synthesis plays a role in synaptic plasticity. Remarkably, a microsecretory apparatus able to participate in the translation of membrane proteins is present in the subsynaptic cytoplasm.

Diffusion Properties of Cell Surface GlyR

In addition to membrane traffic, lateral diffusion within the plasma membrane has been demonstrated for GlyR. This concept has been extended to almost all excitatory and inhibitory receptors so far studied. The single-particle tracking (SPT) approach showed that receptor movements display interspersed periods of high and low diffusion rates in the extrasynaptic and synaptic membrane, respectively. Whereas periods of fast diffusion correspond to Brownian movement, those of slow mobility mainly result from transient association of GlyR with clusters of gephyrin. This association, which reduces diffusion coefficients from up to 0.5 $\mu m^2 s^{-1}$ to less than $10^{-2} \mu m^2 s^{-1}$, imposes a restricted area of exploration (confinement) similar to that generated for other proteins by insertion into lipid rafts or by contact with protein fences or obstacles. The fact that the receptors are not irreversibly retained in gephyrin domains, but can also exit from them, is a key finding which sheds light on how receptor clusters are formed or modified during plastic processes. It favors the new notion that any postsynaptic cluster of gephyrin can behave as donor or/and acceptor of GlyR. Therefore, the dynamic properties of GlyR–gephyrin interactions are also a means whereby the number of synaptic receptors can be regulated. It remains to be seen how the stabilizing scaffold molecules regulate these transitions, both in synaptogenesis and synaptic plasticity.

Because of its interaction with scaffold proteins, the cytoskeleton is a component of the postsynaptic differentiation, at least during differentiation or plastic changes. Because gephyrin is thought to link GlyR to microtubules, depolymerization of the latter would disrupt GlyR clusters. Interestingly, microtubule depolymerization alters the number and density of synaptic GlyR/gephyrin clusters, but the same treatment has no effect on GABA receptor/gephyrin clusters. This discrepancy may reflect the structural heterogeneity of inhibitory synapses that likely relies on receptor-specific assemblies of cytoplasmic proteins.

Pathology – Molecular Biology of Spastic Syndromes

Defects in glycinergic transmission are the cause of several neurological diseases. Some hereditary motor disorders in humans, mice, horses, and cattle result from the expression of mutated glycine receptor genes. The hallmark of these disorders is an exaggerated startle reflex in response to sudden stimuli, giving rise to rapid generalized responses in the form of hypertonia. Startle syndromes related to distinct genetic defects in glycine-mediated neurotransmission correspond either to impairment of GlyR agonist binding function or to reduced expression of functional channels. Both mechanisms lead to inefficient glycinergic inhibition and increased muscle tone.

Human Hyperexplexia

Hyperexplexia is a disorder in which unexpected sensory stimuli provoke exaggerated startle reflexes. Hereditary hyperexplexia, or Kok's disease or familial startle hyperexplexia (STHE), is a rare autosomal dominant neurologic disorder characterized by marked and continuous, sometimes fatal, muscle rigidity in infancy which progressively evolves into transient massive muscle contractions, a reaction that persists throughout adulthood.

Several missense mutations identified in the α1 subunit of GlyR can be responsible for higher glycine EC50 and/or lower channel conductance. The defective phenotype often results from mutations within, or in the vicinity of, the M2 transmembrane domain.

Startle Syndromes in the Mouse

Studies of mutants in the mouse reveal that defects in two genes can cause recessive disorders with clinical and pharmacological features reminiscent of STHE. The inherited spasmodic and spastic phenotypes are autosomal diseases characterized by muscle tremor, hypertonia, and pronounced startle reactions. They are rarely lethal, with the exception of the oscillator phenotype, and during the first 2 weeks after birth homozygous mutant mice have a normal phenotype.

The phenotypic similarity of spasmodic (spd) phenotype and STHE prompted the identification of the GlyR α subunit as the defective protein. A missense mutation, A52S, close to the protein N-terminus, decreases glycine sensitivity without changing the affinity for strychnine, in agreement with the normal binding of the antagonist in the *spd* spinal cord. The mechanism of action of this mutation is unknown, since the region of the mutation is not involved in ligand binding, receptor assembly, or gating.

The phenotype of the mutant mouse 'oscillator' (spdot) is characterized by a fine motor tremor and muscle spasms. Oscillator is allelic with spasmodic but corresponds to a microdeletion coincident with the cytoplasmic extremity of the M3 transmembrane segment. A translational frameshift generates a truncated form of the α1 subunit lacking the M3–M4 loop and the M4 domain.

The highly penetrant recessive 'spastic' mutation is associated with a dramatic reduction in the number of expressed GlyRs in the adult spinal cord and brain stem, but residual receptors exhibit normal subunit

structure and function. The delayed onset of the spasticity coincides with the developmental switch of GlyR, suggesting a selective lack of the adult form, in the absence of transcriptional alteration of the $\alpha 1$ subunit. A mutation affects a β subunit gene allele, which results from the intronic insertion (intron 5) of a LINE-1 transposable element. As a consequence, full-length β mRNAs are barely detectable, and truncated β subunits are generated by creation of premature stop codons.

The general loss of surface receptors in 'spastic' animals is consistent with the role of the β subunit, and results from the lack of a gephyrin-binding site in (hetero-)oligomers. However, it should be noted that a reduction in transport to, and insertion in, specific membrane areas may alternatively arise from an altered quaternary structure that does not allow exit from the endoplasmic reticulum and leads to degradation of the misassembled receptor.

Concluding Remarks

GlyR, which plays essential roles in inhibitory neurotransmission via a heterogeneous composition, has been the center of numerous studies during recent years. Better understanding has been gained at various levels, from structure–function relationships to the cell biology of synapses, including the associated postsynaptic density built up on gephyrin. Remarkably, pioneering work exploiting GlyR as a model molecule has helped develop new concepts concerning not only the biology of neurotransmitter receptors, but also the characterization of the dynamic equilibrium of GlyR-enriched postsynaptic membrane domains; such equilibrium accounts for the strength of synapses via control of their receptor number. GlyR and gephyrin are central to the assembly and activity of a nanomachine, which exploits the diffusion properties of the receptors in order to participate in the excitation–inhibition balance that is modified during plastic events within neuronal networks.

See also: GABA Synthesis and Metabolism; Gamma-Aminobutyric Acid (GABA); Startle Response.

Further Reading

Aguayo LG, van Zundert B, Tapia JC, et al. (2004) Changes on the properties of glycine receptors during neuronal development. *Brain Research: Brain Research Reviews* 47: 33–45.

Cascio M (2004) Structure and function of the glycine receptor and related nicotinicoid receptors. *Journal of Biological Chemistry* 279: 19383–19386.

Choquet D and Triller A (2003) The role of receptor diffusion in the organization of the postsynaptic membrane. *Nature Reviews Neuroscience* 4: 251–265.

Colquhoun D and Sivilotti LG (2004) Function and structure in glycine receptors and some of their relatives. *Trends in Neuroscience* 27: 337–344.

Dahan M, Levi S, Luccardini C, et al. (2003) Diffusion dynamics of glycine receptors revealed by single-quantum dot tracking. *Science* 302: 442–445.

Hanus C, Vannier C, and Triller A (2004) Intracellular association of glycine receptor with gephyrin increases its plasma membrane accumulation rate. *Journal of Neuroscience* 24: 1119–1128.

Kim EY, Schrader N, Smolinsky B, et al. (2006) Deciphering the structural framework of glycine receptor anchoring by gephyrin. *The EMBO Journal* 25: 1385–1395.

Kneussel M and Betz H (2000) Clustering of inhibitory neurotransmitter receptors at developing postsynaptic sites: The membrane activation model. *Trends in Neuroscience* 23: 429–435.

Laube B, Maksay G, Schemm R, et al. (2002) Modulation of glycine receptor function: A novel approach for therapeutic intervention at inhibitory synapses? *Trends in Pharmacological Sciences* 23: 519–527.

Legendre P (2001) The glycinergic inhibitory synapse. *Cellular and Molecular Life Sciences* 58: 760–793.

Lynch JW (2004) Molecular structure and function of the glycine receptor chloride channel. *Physiological Reviews* 84: 1051–1095.

Moss SJ and Smart TG (2001) Constructing inhibitory synapses. *Nature Reviews Neuroscience* 2: 240–250.

Rosenberg M, Meier J, Triller A, et al. (2001) Dynamics of glycine receptor insertion in the neuronal plasma membrane. *Journal of Neuroscience* 21: 5036–5044.

Triller A and Choquet D (2005) Surface trafficking of receptors between synaptic and extrasynaptic membranes: And yet they do move! *Trends in Neuroscience* 28: 133–139.

Vannier C and Triller A (1997) Biology of the postsynaptic glycine receptor. *International Review of Cytology* 176: 201–244.

Glycogen Metabolism in CNS White Matter

A Brown, University of Washington, Seattle, WA, USA; and University of Nottingham, Nottingham, UK
S Baltan and B R Ransom, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Glycogen is the readily mobilized storage form of glucose in most mammalian tissues, including the brain. It is a branched chain polysaccharide molecule that can contain thousands of glucose residues. Glucose can be stored in the form of glycogen during periods of adequate exogenous glucose availability, to be liberated for use during periods of glucose need. The presence of glycogen in the brain has been known for many years, where it is estimated to contribute about 0.1% of the total brain weight (0.5–1.5 g). This is a very low concentration compared to skeletal muscle and liver, calling into question whether brain glycogen is functionally important. Glycogen is unlikely to act as an energy reserve in the conventional sense, since calculations have shown that there is only enough glycogen present to support brain activity for a few minutes in the total absence of glucose (see the section titled 'Roles of brain glycogen in neural function'). A more likely role is to provide energy substrate during short periods of increased demand when the supply of glucose does not match the increased demand by the brain.

Another reason that brain glycogen has been neglected is that it is contained exclusively in astrocytes. While glycogen 'turns over' quite actively in astrocytes, intact glucose residues derived from glycogen do not appear outside of astrocytes. Thus, it was concluded that brain glycogen would be of no benefit for neurons, which are known to be glucose dependent. A number of developments, including the discovery that neurons can generate ATP when provided substrates other than glucose, led to a resurgence of interest in brain glycogen. This molecule is highly regulated, and its concentration and availability is under physiological control. This article summarizes the biochemistry of brain glycogen and offers a current view of its roles in brain function, topics once again dynamic, after decades of dormancy.

Brain Glycogen: Macroscopic Issues

Regional Variations in Glycogen Content

Glycogen content displays regional variability, with gray matter structures having higher levels than white matter structures. Caudal, primitive, gray matter structures have the highest glycogen content (e.g., pons/medulla, cerebellum), more than twice as much as rostral areas (e.g., cerebral cortex and striatum). Even at the level of individual brain structures there is variation in content, which may be related to the density of glycogen-containing cells within the structure. For example, the dentate gyrus possesses twice the glycogen content of the rest of the hippocampus; in mouse, cerebral cortex glycogen content is higher in the superficial layers than in the deeper layers; and in four areas of medulla oblongata glycogen levels vary significantly.

Cellular Localization of Brain Glycogen

Glycogen is located predominantly in the cytoplasm of astrocytes. During development, it appears in both neurons and glial cells, but thereafter only a few neurons in the brain stem express glycogen. Some glycogen is also found in ependymal cells, including the specialized ependyma that help form the choroid plexus; ependymal cells are considered to be a specialized type of astrocyte. The expression of the enzymes that control glycogen synthesis and breakdown correlates with the astrocytic localization of glycogen. The brain isozyme of glycogen phosphorylase, the enzyme involved in glycogen breakdown, is located in astrocytes. Glycogen synthase, the enzyme that controls glycogen synthesis, is also present in astrocytes (Figure 1). This enzyme is also found in neurons but its function is not well understood and appears not to involve glycogen metabolism. Since astrocytes are found in all brain regions, this cellular localization does not dictate regional distribution; rather, it may explain the variation in glycogen content within distinct brain areas. The highest concentration of glycogen occurs in areas of high synaptic density, which immediately suggests a role in neuronal (synaptic) activity.

Measurement of Brain Glycogen

As mentioned, central nervous system (CNS) glycogen is present in much lower concentrations than liver or skeletal muscle. The concentration of glycogen in the rodent brain is 2–12 μmol g^{-1} of wet tissue. The presence of glycogen in tissue can be ascertained using techniques that visually identify the presence of glycogen, and quantitative techniques that estimate or measure absolute glycogen content. The most common staining method relies on the interaction of glycogen and osmium tetroxide, resulting in glycogen appearing as electron dense particles, 10–40 nm in diameter. This technique is used to confirm the

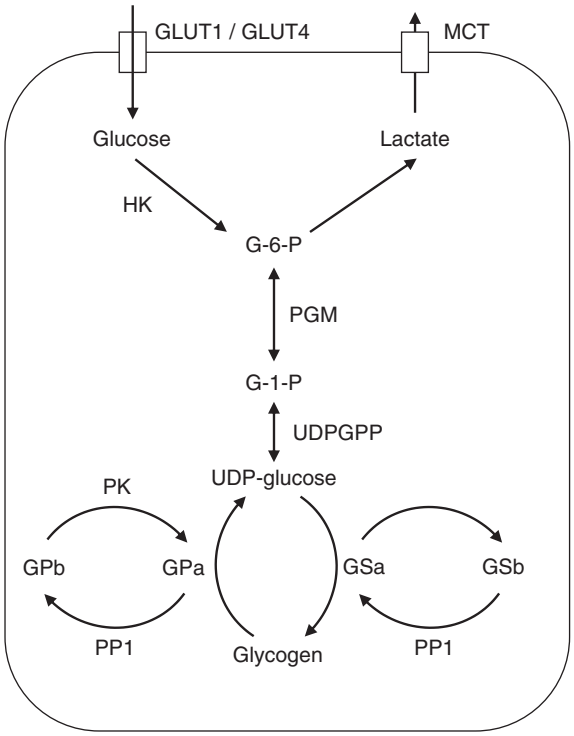

Figure 1 Glucose entry and glycogen synthesis in the brain. Glucose is transported via the insulin-insensitive GLUT1 transporter (GLUT1) and converted to glucose-6-phosphate (G-6-P) by hexokinase (HK), then to glucose-1-phosphate (G-1-P) phosphoglucomutase (PGM). G-1-P is converted to uridine diphosphate (UDP)-glucose by UDP glucose pyrophosphate (UDPGPP). UDP-glucose is then converted by glycogen synthase (GS) to glycogen. Protein phosphatase 1 (PP1) converts the inactive GSb to the active GSa, promoting glycogen accumulation. Glycogen is broken down by glycogen phosphorylase (GP), which exists in the active phosphorylated form (GPa) or the inactive dephosphorylated form (GPb). Phosphorylase kinase (PK) dephosphorylates GPb to the active form.

presence of glycogen and cannot yield quantitative information.

Quantitative measurements of glycogen can be performed in live or postmortem tissue. The majority of reported measurements of glycogen were obtained using biochemical assays on postmortem tissue. Sacrifice using very high-intensity microwave radiation denatures enzymes responsible for glycogen metabolism, thus preserving glycogen content. Alternatively, tissue can be rapidly frozen to minimize glycogen breakdown. The most accurate and common assay technique for measuring glycogen in postmortem tissue is to use enzymes that metabolize glycogen to glucose, which is subsequently metabolized via the use of enzymes whose ultimate reaction yields the reduced form of nicotinamide adenine dinucleotide (NADH), which can be measured fluorometrically. Comparison with glucose standards is used to yield estimates of glycogen content. Glycogen is usually expressed as glucosyl units, as the debranching enzymes yield one glucose molecule per glucosyl unit of the glycogen molecule.

Nuclear magnetic resonance (NMR) spectroscopic detection of labeled glucose (^{13}C) incorporation into the glycogen can be used to estimate glycogen content *in vivo* in both laboratory animals and humans. ^{13}C-labeled glucose is systemically injected, resulting in slow incorporation of the labeled glucose into the brain glycogen pool, which produces a detectable glycogen signal that can be continuously monitored. Although used for investigational purposes at present, this technology could be used clinically to measure brain glycogen levels under both pathological and physiological conditions.

Roles of Brain Glycogen in Neural Function

Glycogen Function in Liver and Skeletal Muscle

The two main stores of glycogen in the body are contained in the liver (~100 g of glycogen in the well-fed liver) and skeletal muscle (400 g in resting skeletal muscle), and it is from these two tissues that the majority of our knowledge about glycogen derives. Excess systemic glucose is stored in the liver to act as a buffer against systemic hypoglycemia. During periods of hypoglycemia, liver glycogen is metabolized to glucose, which is released into the blood to maintain euglycemia. The liver glycogen can support euglycemia for about a day during starvation, after which fat stores are used. The liver glycogen can be viewed as an egalitarian molecule ensuring that euglycemia is maintained for the benefit of the entire body. In reality, liver glycogen is used to support brain function, which is far more sensitive to hypoglycemia than other organs; that is, brain function deteriorates during hypoglycemia, whereas organs such as the heart are unaffected. Skeletal muscle also contains significant reserves of glycogen, although at a concentration less than liver (30–100 μmol g^{-1} vs. 100–500 μmol g^{-1}; glycogen accounts for 1–2% of skeletal muscle weight and 6–8% of liver weight). However, given the much greater total body weight of skeletal muscle compared to liver, there is more glycogen contained in skeletal muscle than liver. Skeletal muscle glycogen is used to fuel muscle cells during exercise. The glycogen is metabolized to lactate, which is disposed of by the systemic circulation. This seems rather wasteful, as the relative energy yield of glycolysis to oxidative phosphorylation is low; however, oxygen uptake cannot keep up with the rate of lactate production. Thus, the dogmatic view is that skeletal muscle glycogen is used solely to fuel skeletal muscle cells.

Recent data, however, have indicated that lactate produced in skeletal muscle can by extracted from the systemic circulation and metabolized by the brain.

Review of Glycogen Dynamics

Fate of lactate, a glycogen breakdown product Because glycogen is located predominantly in astrocytes and is too big to be transferred across cell boundaries, a glycogen-derived intermediary must be employed as an energy conduit. It is widely agreed that the ultimate breakdown product of astrocytic glycogen is lactate, since astrocytes do not express glucose-6-phosphatase and are therefore incapable of releasing glucose. When astrocytes are deprived of glucose, glycogen rapidly disappears and nearly all of the glycogen-derived carbon fragments found in extracellular fluid are in the form of lactate.

It was realized over 100 years ago that the anatomical arrangement of astrocytes, capillaries, and neurons suggested that astrocytes mediated the transfer of metabolic substrate from capillary to neuron. The endothelial cells and astrocytic endfeet express the GLUT1 glucose transporter facilitating glucose uptake from capillaries into astrocytes. Astrocytes possess the isoform of the lactate dehydrogenase enzyme (LDH5) that preferentially converts pyruvate to lactate. Neurons stain exclusively for the LDH1 enzyme that preferentially oxidizes lactate to pyruvate. Consequently, astrocytes readily convert pyruvate to lactate, and neurons oxidize lactate to pyruvate. Monocarboxylate transporters (MCTs) transport lactate across cell membranes using proton symport. There are several isoforms of MCT, which vary in their lactate binding affinities. The form expressed by astrocytes, MCT1, excels at lactate export. The MCT isoform expressed in neurons, MCT2, is a preferential lactate importer. Studies on the rodent optic nerve provided proof of principle that a glycogen-derived monocarboxylate, most probably lactate (although pyruvate is also remotely possible) is transported from astrocytes to axons during both aglycemia and increased tissue energy demand.

Turnover Glycogen turnover in the brain can be activated selectively. For example, glycogen breakdown is increased in 'barrel fields' in rat somatosensory cortex when the facial whiskers that project to this area are manually stimulated. Glycogen turnover is reported as rapid or relatively slow, depending on the methods employed. Irrespective of turnover rates, during quiescence it is obvious that glycogen metabolism can be rapidly activated. Once glycogen has been depleted it can be replenished, depending upon the rates of synthesis and degradation. The resynthesis of glycogen depends on protein synthesis, and it appears that neurotransmitters modify this interaction. *In vivo* rat studies demonstrated decreases in brain glycogen content by insulin-induced hypoglycemia followed by increased glycogen content above baseline levels on restoration of normoglycemia. This increase in glycogen content above baseline has been termed supercompensation and appears to be a standard compensatory mechanism, possibly to prepare the tissue for anticipated future metabolic challenges.

Physiological conditions that modify glycogen content
Sleep The effects of sleep, an index of brain activity, may provide clues to glycogen's role during physiological activity. Early studies demonstrated that brain glycogen accumulates during sleep, is depleted by sleep deprivation, and is mobilized on waking. In rodents, glycogen increased by 70% above waking levels within a few minutes of induction of slow-wave sleep. There is a correlation between glycogen phosphorylase activity and levels of noradrenaline and serotonin (5-HT), which stimulate glycogenolysis. Thus, it is hypothesized that sleep is a state in which activity is decreased and glycogen is replenished, and that glycogen depletion slowly occurs during the increased activity associated with waking. The expression of glycogen synthase and glycogen phosphorylase varies with sleep, with maximal activity occurring in the middle of the light period and minimal activity occurring in the middle of the dark period.

General anesthesia General anesthesia resulted in elevations of glycogen dependent on the depth and duration of anesthesia. Electron microscopic studies demonstrated increases in glycogen levels in the hippocampus, cortex, and molecular layer of the cerebellum, areas of high synaptic density. This immediately suggests a correlation between the ability of anesthetics to reduce neuronal activity and to elevate glycogen content. The effect of anesthetics on brain glycogen is related to neuronal activity since anesthetics added to astrocytes in culture did not affect glycogen content. This ability of anesthetics to elevate *in vivo* glycogen content indicates that in the fully conscious state neurons utilize glycogen, but during quiescence the neurons reduce their utilization, allowing glycogen accumulation.

Can Glycogen Serve as an Energy Substrate for Adjacent Neural Elements?

Hypoglycemia The introduction of insulin-induced hypoglycemia to treat schizophrenia prompted some of the earliest studies on brain glycogen, which attempted to study the effect of limiting glucose delivery on brain biochemistry. Measurements of glycogen

in the adult dog subjected to varying degrees of hypoglycemia indicated that glycogen content fell first in areas of the brain with the highest metabolic rates, demonstrating a connection between glycogen and brain function. During insulin-induced hypoglycemia, when the brain's demand for energy exceeds blood-borne glucose supply, glycogen is degraded to support brain function. However, once glycogen content is depleted during constant hypoglycemia brain function fails. More recent data using NMR spectroscopy to measure ^{13}C glucose incorporation into brain glycogen have demonstrated similar results; insulin-induced hypoglycemia depleted brain glycogen levels in anaesthetized rats, but only after the brain glucose concentration had fallen to zero. Thus, for the first time a link was established between glycogen metabolism and brain function, albeit under pathological conditions.

A more direct clue emerged from tissue culture studies on hypothalamic neurons, which were cultured in the absence or presence of glial cells. Neurons grown with glia survived for longer periods than neurons grown in the absence of glial cells, implying that neuronal survival in glia-confluent conditions was due to a metabolic coupling of unknown mechanism between the cell types. A similar study showed that it was the glycogen content of astrocytes that was the determining factor in neuronal survival in mixed glial–neuronal cultures. The benefit of glycogen remained unclear. Was it promoting neuronal survival by allowing more robust astrocyte metabolism, perhaps assisting glutamate uptake, or was this energy supply somehow 'shared' with neighboring neurons?

Studies using intact preparations yielded more evidence of a link between glycogen content with brain function. Stimulus-evoked excitatory postsynaptic potentials (EPSPs) in the CA1 region of the hippocampus were sustained in artificial cerebrospinal fluid (aCSF) containing 10 mM glucose. Removal of glucose from the aCSF resulted in a delayed decrease in the EPSP slope. This limited duration of function in aglycemia may be due to metabolism of the glycogen stores to support function. Supporting this theory, experimental maneuvers that deplete glycogen result in accelerated EPSP decrease during aglycemia.

Studies on the rodent optic nerve established a relationship between glycogen content and axon function (Figure 2). In the presence of astrocyte glycogen, axon function loss was delayed during aglycemia. Axon function remained stable for 20–30 min. Increasing tissue metabolic demand resulted in accelerated axon function loss, verifying that the metabolic rate determines the rate of glycogen breakdown. Upregulation or downregulation of glycogen content modified latency to functional loss accordingly. Furthermore,

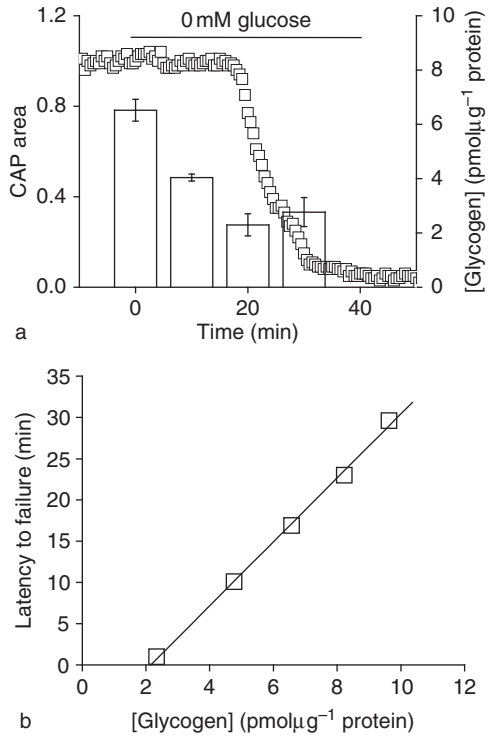

Figure 2 Glycogen content determines latency to axon function failure during aglycemia in mouse optic nerve (MON). (a) In MON, aglycemia results in delayed axon function loss. Axon function is quantitatively monitored as the area under the stimulus-evoked compound action potentials (CAPs). On exposure to aglycemia, glycogen content falls until it reaches its lowest point in 20 min, which coincides with CAP failure. The left axis indicates normalized CAP area (open squares) and the right axis indicates the glycogen content of the MONs (open columns ± S.E.M.) (b) The straight line is a best fit of data points and indicates the relationship between glycogen content and latency to CAP failure after aglycemia onset. (a, b) From Brown AM (2004) Brain glycogen re-awakened. *Journal of Neurochemistry* 89(3): 537–552, with permission from Blackwell Publishing.

blocking lactate uptake into axons using D-lactate, a competitive blocker of L-lactate uptake, accelerated axon function loss during aglycemia. Taken together, these data indicate that it is the product of glycogen metabolism, lactate, and its subsequent transfer to axons that sustain function. Therefore, it is the glycogen content of the tissue at the onset of aglycemia that determines the latency to functional loss. These results may have profound implications in understanding brain energy metabolism in patients suffering from type 1 diabetes mellitus.

Intense neural activity During high-frequency stimulation of a CNS axon pathway, glycogen content falls but axonal discharge is well maintained. When lactate transport is blocked by D-lactate during this high-frequency discharge, the axon function fails, supporting the idea that lactate is being transported

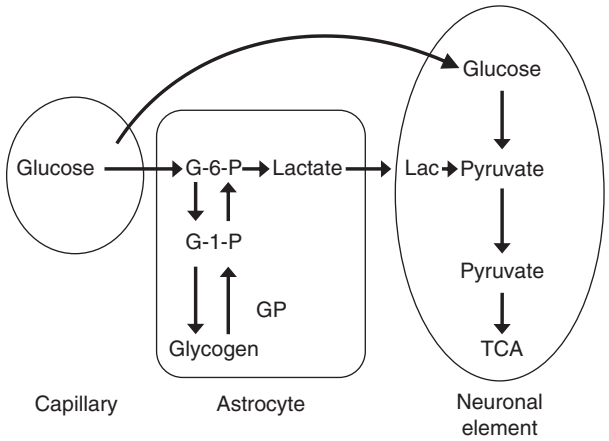

Figure 3 The role of brain glycogen. Glucose is transported into astrocytes then is phosphorylated to glucose-6-phosphate (G-6-P), which is the substrate for both glycogenesis and glycolysis. Glycogen can be rapidly degraded to G-1-P by glycogen phosphorylase (GP) in response to increased demand. G-6-P is then glycolytically metabolized ultimately to lactate, which is transported to neuronal element where it is oxidatively metabolized. TCA, tricarboxylic acid.

from astrocytes to axons under these conditions to assist axon ATP production (recall that the two molecules of lactate generated from one glucose molecule yield more than 90% of the ATP that would be generated if glucose itself were oxidized; Figure 3). Additionally, blocking glycogen phosphorylase during increased tissue energy demand causes axon function failure. Thus, glycogen appears to play a role in both pathological and physiological conditions, and astrocytes can share fuel with their neural neighbors. A possible advantage of this arrangement is that lactate taken up by neurons is metabolized without the initial expense of the ATP necessary to phosphorylate glucose. Since nearly every neuron/axon has astrocyte neighbors, the delivery of lactate may be more rapid, efficient, and reliable than diffusion of glucose from more distant capillaries.

A Physiological Role for Brain Glycogen

In summary, a role for brain glycogen is suggested based on the following: (1) glycogen can act as an energy substrate to support neural function during aglycemia; (2) glycogen content increases in the brain during quiescence are consistent with the conscious, awake brain utilizing glycogen; (3) glycogen metabolism is controlled by neurotransmitters and hormones, whose levels fluctuate, as well as local energy status; and (4) excess glucose uptake into the brain is disposed of by the venous circulation, implying that brain glycogen does serve a specific function. Stimulation of the rat face resulted in focal decreases in glycogen, implying local glycogenolysis, linking physiological neuronal activity with glycogen utilization. Handling of rats, which equates to elevated sensory activation, reduced brain glycogen compared to control unhandled animals. These studies demonstrate that sudden increases in local energy demand may be met by increased glycogenolysis in the brain.

In addition, studies on CNS white matter establish that astrocyte glycogen can support high-frequency axonal discharge by transferring glycogen-derived lactate to the intensely discharging axons. This 'physiological' role of glycogen was supported by the following observations: (1) inhibition of lactate transport out of astrocytes or into axons reduced function; (2) downregulation of glycogen content impaired high-frequency firing during normoglycemia; and (3) inhibition of glycogen phosphorylase activity by isofagomine reduced function. Thus, a role for glycogen as an energy buffer, rather than an energy reserve, has been proposed. In this role, brain glycogen provides energy substrate in the form of glycogen-derived lactate even in the presence of normoglycemic concentrations of glucose to support function when ambient glucose alone is insufficient to meet immediate tissue energy demands. Providing tissue with unlimited energy substrate allowed indefinite maintenance of function during high-intensity stimulus, indicating that the ability to support function during increased tissue energy demand is dependent on the availability of suitable energy substrate.

See also: Activity-Dependent Metabolism in Glia and Neurons; Glial Energy Metabolism: Overview; Glial Glycogen Metabolism; Glial Glutamate and GABA Metabolism; Glial Energy Metabolism: A NMR Spectroscopy Perspective.

Further Reading

Bittar PG, Charnay Y, Pellerin L, Bouras C, and Magistretti PJ (1996) Selective distribution of lactate dehydrogenase isoenzymes in neurons and astrocytes of human brain. *Journal of Cerebral Blood Flow and Metabolism* 16: 1079–1089.

Bouzier-Sore AK, Merle M, Magistretti PJ, and Pellerin L (2002) Feeding active neurons: (Re)emergence of a nursing role for astrocytes. *Journal of Physiology (Paris)* 96: 273–282.

Broer S, Rahman B, Pellegri G, et al. (1997) Comparison of lactate transport in astroglial cells and monocarboxylate transporter 1 (MCT 1) expressing Xenopus laevis oocytes. Expression of two different monocarboxylate transporters in astroglial cells and neurons. *Journal of Biological Chemistry* 272: 30096–30102.

Brown AM, Tekkok SB, and Ransom BR (2003) Glycogen regulation and functional role in mouse white matter. *Journal of Physiology* 549: 501–512.

Brown AM, Westenbroek RE, Baltan Tekkok S, and Ransom BR (2002) Functional insulin receptors are selectively expressed on CNS astrocytes. *Society of Neuroscience Abstracts* #581.512.

Chih CP, Lipton P, and Roberts EL Jr. (2001) Do active cerebral neurons really use lactate rather than glucose? *Trends in Neuroscience* 24: 573–578.

Chih CP and Roberts EL Jr. (2003) Energy substrates for neurons during neural activity: A critical review of the astrocyte–neuron lactate shuttle hypothesis. *Journal of Cerebral Blood Flow and Metabolism* 23: 1263–1281.

Dringen R, Gebhardt R, and Hamprecht B (1993) Glycogen in astrocytes: Possible function as lactate supply for neighboring cells. *Brain Research* 623: 208–214.

Forsyth R, Fray A, Boutelle M, Fillenz M, Middleditch C, and Burchell A (1996) A role for astrocytes in glucose delivery to neurons? *Developmental Neuroscience* 18: 360–370.

Koehler-Stec EM, Simpson IA, Vannucci SJ, Landschulz KT, and Landschulz WH (1998) Monocarboxylate transporter expression in mouse brain. *American Journal of Physiology* 275: E516–E524.

Pellerin L and Magistretti PJ (2003) Food for thought: Challenging the dogmas. *Journal of Cerebral Blood Flow and Metabolism* 23: 1282–1286.

Schurr A, West CA, and Rigor BM (1988) Lactate-supported synaptic function in the rat hippocampal slice preparation. *Science* 240: 1326–1328.

Swanson RA (1992) Physiologic coupling of glial glycogen metabolism to neuronal activity in brain. *Canadian Journal of Physiology and Pharmacology* 70: S138–S144.

Tekkok SB, Brown A, Westenbroek R, Pellerin L, and Ransom BR (2005) Transfer of glycogen-derived lactate from astrocytes to axons via specific monocarboxylate transporters supports mouse optic nerve activity. *Journal of Neuroscience Research* 81: 644–652.

Wender R, Brown AM, Fern R, Swanson RA, Farrell K, and Ransom BR (2000) Astrocytic glycogen influences axon function and survival during glucose deprivation in central white matter. *Journal of Neuroscience* 20: 6804–6810.

Glycosylation: General Overview

J Jaeken and H Carchon, University Hospital Gasthuisberg, Leuven, Belgium

© 2009 Elsevier Ltd. All rights reserved.

Protein Glycosylation

Congenital disorders of glycosylation (CDG) are a rapidly growing family of genetic diseases due to defects in the synthesis of the glycan moiety of glycoconjugates. This article is limited to genetic defects of protein glycosylation.

There are two main types of protein glycosylation: N-glycosylation and O-glycosylation. N-glycosylation (N-glycans attached to an aminogroup of asparagine of proteins) comprises an assembly part and a processing part, and extends over three cellular compartments: the cytosol, the endoplasmic reticulum (ER), and the Golgi. The assembly part of the N-glycosylation starts on the cytosolic side of the ER with the transfer of *N*-acetylglucosamine (GlcNAc) phosphate from Uridine diphosphate (UDP)–GlcNAc to membrane-bound dolichyl monophosphate (Dol-P), forming GlcNAc-pyrophosphatedolichol (GlcNAc-PP-Dol). One GlcNAc and five mannose (Man) residues are subsequently attached to this lipid-linked monosaccharide in a stepwise manner. The donor of these mannoses is a nucleotide-activated sugar, guanosine diphosphate (GDP)-Man, which is synthesized from fructose 6-phosphate, an intermediate of the glycolytic pathway. The lipid-linked heptasaccharide $Man_5GlcNAc_2$ is translocated by a flippase across the ER membrane and is elongated at the lumenal side by the attachment of four further mannose residues and, subsequently, three glucose residues. The four mannosyltransferases and three glucosyltransferases involved require dolichyl phosphate-bound monosaccharides (Dol-P-Man and Dol-P-Glc). The completed $Glc_3Man_9GlcNAc_2$ oligosaccharide is then transferred to selected asparagine residues of the nascent proteins by the oligosaccharyltransferase complex.

The processing part of the N-glycosylation starts in the ER by trimming the glucoses (catalyzed by glucosidases I and II) and one mannose (catalyzed by α-mannosidase I). The residual glycoprotein intermediate is directed to the *cis*-Golgi where the processing pathway branches. A minor branch targets glycoproteins to the lysosomes (after the action of a GlcNAc-phosphotransferase and removal of the GlcNAc residues, leaving high-mannose glycoproteins capped with Man 6-P). The main branch leads to further trimming of mannoses (leaving a trimannosyl core) and addition of, respectively, GlcNAc, galactose, and eventually, sialic acid in the medial- and *trans*-Golgi. Another modification of many N-glycoproteins in the Golgi is the attachment of fucose to the GlcNAc residue that is linked to asparagine. O-glycosylation (O-glycans attached to the hydroxylgroup of threonine or serine of proteins) has no processing part and thus only consists of assembly. Contrary to N-glycosylation, this assembly mainly occurs in the Golgi (Figure 1). O-glycan structures show a greater diversity than N-glycans. Examples of important O-glycans are *O*-*N*-acetylgalactosaminylglycans (mucin-type glycans), *O*-xylosylglycans (glycosaminylglycans), *O*-mannosylglycans, and *O*-fucosylglycans.

Genetic Diseases of Protein N-Glycosylation

Eighteen diseases are known in protein N-glycosylation: 12 assembly defects (CDG-I group) designated CDG-Ia to CDG-Il, and six processing defects (CDG-II group) designated CDG-IIa to CDG-IIf (Table 1).

Phosphomannomutase2 Deficiency (CDG-Ia)

Phosphomannomutase (PMM)2 deficiency is a (cytosolic) defect in the second step of the mannose pathway (transforming mannose-6-phosphate into mannose-1-phosphate), which normally leads to the synthesis of GDP-Man. This nucleotide sugar is the donor of the mannose used in the ER to assemble the dolichol-pyrophosphate oligosaccharide precursor. Deficiency of GDP-Man causes hypoglycosylation of numerous glycoproteins, including serum proteins, lysosomal enzymes, and membranous glycoproteins.

The diagnosis of CDG-Ia (and of congenital disorders of N-glycosylation in general) is usually made by isoelectrofocusing (IEF) and immunofixation of serum transferrin (Figure 2). Normal serum transferrin is mainly composed of tetrasialotransferrin and small amounts of mono-, di-, tri-, penta-, and hexasialotransferrins. The partial deficiency of sialic acid (a negatively charged and end-standing sugar) in CDG causes a cathodal shift. Two main types of cathodal shift can be recognized: type 1 is characterized by an increase of both disialo- and asialotransferrin and a decrease of tetrasialotransferrin; in type 2 there is also an increase of the tri- and/or monosialotransferrin bands. In PMM2 deficiency, a type 1 pattern is found. Capillary zone electrophoresis of total serum was introduced in the mid-2000s for the diagnosis of CDG (Figure 2). In addition to the above-mentioned serum glycoprotein abnormalities, laboratory findings

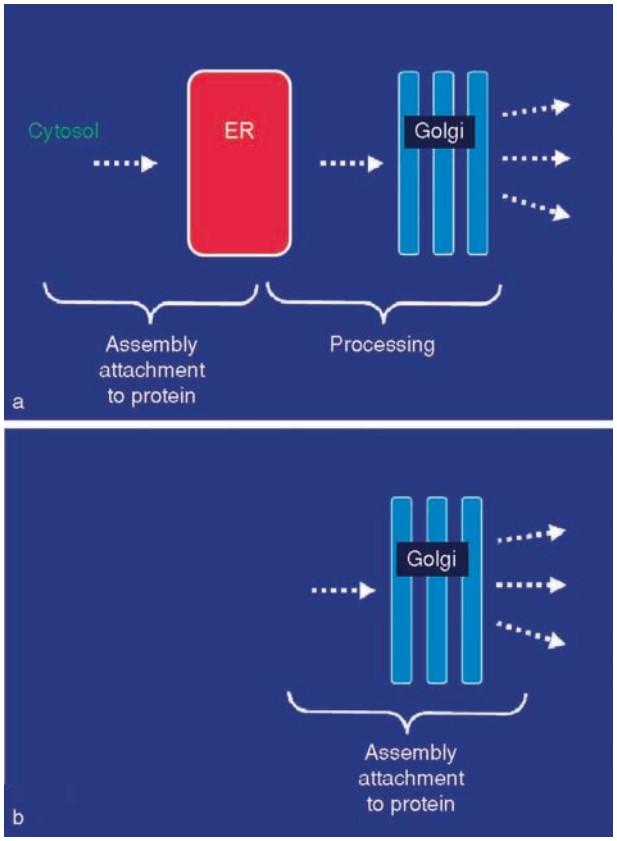

Figure 1 General scheme of the cellular localization of (a) N-glycosylation and (b) O-glycosylation.

Table 1 Human hereditary N-glycosylation disorders

Name	*Defective gene*	*Defective protein*
CDG-I		
CDG-Ia	PMM2	Phosphomannomutase 2
CDG-Ib	*MPI*	Phosphomannose isomerase
CDG-Ic	*hALG6*	Glucosyltransferase I
CDG-Id	*hALG3*	Mannosyltransferase VI
CDG-Ie	*DPMI*	Dolichol-P-Man synthase I
CDG-If	*Lec35*	Lec35
CDG-Ig	*hALG12*	Mannosyltransferase VIII
CDG-Ih	*hALG8*	Glucosyltransferase II
CDG-Ii	*hALG2*	Mannosyltransferase II
CDG-Ij	*DPAGT1*	UDP-GlcNAc: dolichol phosphate *N*-acetylglucosamine 1-phosphate transferase
CDG-Ik	*hALG1*	Mannosyltransferase I
CDG-Il	*hALG9*	Mannosyltransferase VII/IX
CDG-II		
CDG-IIa	*MGAT2*	*N*-Acetylglucosaminyltransferase II
CDG-IIb	*GLS1*	Glucosidase I
CDG-IIc	*FUCT1*	GDP-fucose transporter I
CDG-IId	*B4GALT1*	β-1,4-Galactosyltransferase 1
CDG-IIe[a]	*COG7*	Conserved oligomeric Golgi complex, subunit 7
CDG-IIf	*SLC35A1*	CMP-sialic acid transporter

[a]Preliminary assignment, combined N- and O-glycosylation disorder.

include elevation of serum transaminase levels, hypoalbuminemia, hypocholesterolemia, and tubular proteinuria. To confirm the diagnosis, the activity of PMM2 should be measured in leukocytes or fibroblasts. Prenatal diagnosis is possible by enzymatic analysis of amniocytes and chorionic villus cells; this should be combined with mutation analysis of the PMM2 gene.

CDG-Ia is by far the most frequent protein N-glycosylation disorder (some 600 patients known). The clinical spectrum is very broad: the nervous system is affected in all patients, and most other organs are involved in a variable way. The neurologic picture comprises alternating internal strabism and other abnormal eye movements, axial hypotonia, psychomotor retardation, ataxia, and hyporeflexia. After infancy, symptoms include retinitis pigmentosa, often stroke-like episodes, and sometimes epilepsy. During the first year(s) of life, there are variable feeding problems (anorexia, vomiting, diarrhea) that can result in severe failure to thrive. Other features are a variable dysmorphy (large, hypoplastic/dysplastic ears), abnormal subcutaneous adipose-tissue distribution (fat pads, inverted nipples), mild to moderate hepatomegaly, H skeletal abnormalities (including atlantoaxial subluxation), and hypogonadism. Some infants develop pericardial effusion and/or cardiomyopathy. At the other end of the clinical spectrum are patients with a very mild phenotype (no dysmorphy, slight psychomotor retardation). Patients often have an extroverted and happy appearance. There is a substantially increased mortality in the first years of life due to vital organ involvement or severe infection.

CDG-Ia is an autosomal-recessive disease due to mutations of *PMM2* on chromosome 16p13. At least 60 mutations have been identified (mainly missense mutations), the most frequent being the R141H mutation.

Phosphomannose-Isomerase Deficiency (CDG-Ib)

The defect in phosphomannose-isomerase (PMI) deficiency is in the first step in the biosynthesis of the nucleotide sugar GDP-Man. The substrate of the enzyme, fructose 6-phosphate, does not accumulate since it is an intermediate of the glycolytic pathway. The blood biochemical abnormalities are indistinguishable from those found in CDG-Ia. The diagnosis is confirmed by finding a decreased activity of phosphomannose isomerase in leukocytes or fibroblasts and/or mutation(s) in *PMI*.

Some 20 patients have been reported with this mainly hepatic–intestinal disease. Together with CDG-Ih, it is the only known N-linked CDG without or with only minor neurological involvement. Symptoms started at ages between 1 month and 11 months and consisted of various combinations of

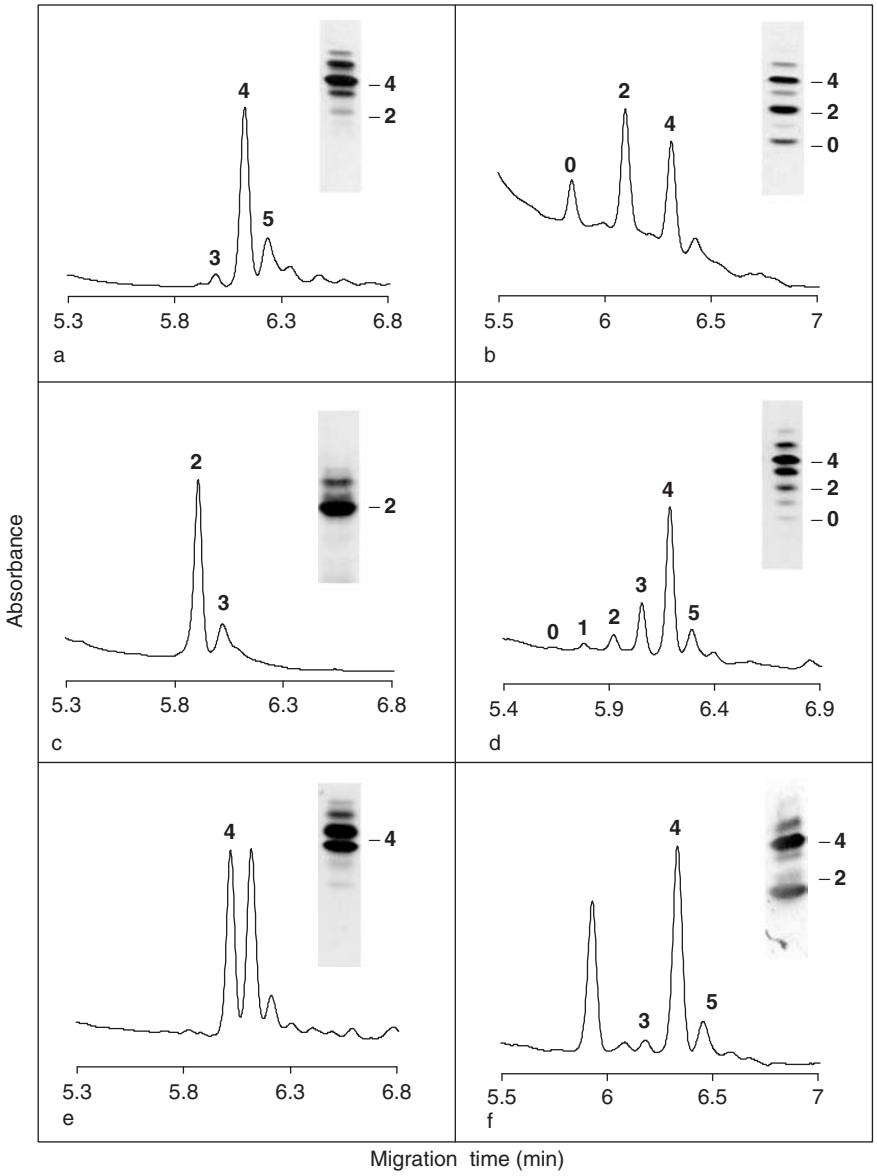

Figure 2 CE and IEF analysis of different types of serum transferrin profiles. (a) demonstrates a normal Tf profile, whereas (b–d) show the Tf profile corresponding to a CDG-Ia, CDG-IIa, and CDG-IIx, respectively. (e, f) both represent Tf protein variants, as the same profile was observed in one of the healthy parents and two major protein peaks remained after neuraminidase action.

recurrent vomiting, abdominal pain, protein-losing enteropathy, recurrent thromboses, gastrointestinal bleeding, liver disease, and symptoms of hypoglycemia. Several patients died.

This is the only known CDG that is efficiently treatable. The treatment is simple and consists of oral mannose (1 g kg^{-1} body weight per day, divided in 4–6 doses). The rationale for this treatment is that hexokinases phosphorylate mannose to mannose 6-phosphate, thus bypassing the defect.

Inheritance of CDG-Ib is autosomal recessive. The gene (*PMI*) has been localized to chromosome 15q22 and several (missense) mutations have been identified.

Glucosyltransferase I Deficiency (CDG-Ic)

Glucosyltransferase I deficiency is a defect in the attachment of the first glucose (of three) to the dolichol-linked $mannose_9$-*N*-$acetylglucosamine_2$ ER intermediate. It causes hypoglycosylation of serum glycoproteins, because nonglucosylated oligosaccharides are a suboptimal substrate for oligosaccharyltransferase.

This is the second most common protein N-glycosylation disease (some 30 patients identified). As in CDG-Ia, patients show hypotonia, strabismus, and seizures but psychomotor development is less retarded, there is less dysmorphy, and usually no retinitis pigmentosa or cerebellar hypoplasia. In one

patient, idiopathic intracranial hypertension and optic atrophy have been reported.

For unknown reasons, some of the glycoproteins have unusually low blood levels (particularly factor XI, and coagulation inhibitors such as antithrombin and protein C). The reason why the clinical picture in these patients is much milder than that of PMM-deficient patients may be because a deficiency in glucosylation of the dolichol-linked oligosaccharides does not affect the biosynthesis of GDP-Man and, hence, does not affect the biosynthesis of GDP-fucose or the biosynthesis of glycosylphosphatidylinositol-anchored glycoproteins. Long-term outcome is unknown since all reported patients have been children.

This disease illustrates that even in cases of mild psychomotor retardation without any specific dysmorphy, IEF of serum sialotransferrins has to be performed. When a type 1 pattern is found, PMM2 and PMI deficiency have to be considered first. If these enzymes show normal activities, the next step is analysis of the dolichol-linked oligosaccharides in fibroblasts. In CDG-I the major fraction of these oligosaccharides consists of nine mannose and two *N*-acetylglucosamine residues without the normally present three glucose residues. Glucosyltransferase I activity has to be measured in fibroblasts and/or mutation analysis performed.

Mannosyltransferase VI Deficiency (CDG-Id)

Three patients and a fetus have been reported affected with this deficiency. Also of the subsequently mentioned N-glycosylation disorders, only one or a few patients are known. Two patients showed a severe neurological syndrome with nearly absent psychomotor development, hypsarrhytmia, postnatal microcephaly, optic atrophy, iris coloboma, and brain and corpus callosum atrophy. In the third patient, there was hyperinsulinemic hypoglycemia with islet cell hyperplasia (also reported in CDG-Ia and CDG-Ib). IEF of serum transferrin showed a variant type 1 pattern without increase of asialotransferrin. The hallmark biochemical feature of this CDG (and of CDG-Ie) is an accumulation of dolichyl pyrophosphate-$Man_2GlcNAc_2$ in fibroblasts.

Dolichyl Phosphate-Mannose Synthase 1 Deficiency (CDG-Ie)

Five children have been reported with this defect in the synthesis of the mannose donor Dol-P-Man. These patients showed a severe neurological involvement as in CDG-Id, and a similar serum sialotransferrin pattern and accumulation of dolichyl pyrophosphate-Man_2 $GlcNAc_2$ in fibroblasts.

Lec35 deficiency (CDG-If) The deficient protein, Lec35, seems to be a chaperone involved in the utilization of both Dol-P-Man and Dol-P-Glc. The four reported patients showed severe encephalopathy and three of them had a scaly, erythematous skin disorder. The oldest patient (actually 20 years) had dwarfism with transient growth-hormone deficiency. Fibroblasts accumulated dolichyl pyrophosphate-$Man_5GlcNAc_2$ as well as dolichyl pyrophosphate-$Man_9GlcNAc_2$.

Mannosyltransferase VIII Deficiency (CDG-Ig)

In the five reported patients the clinical picture comprised facial dysmorphy, psychomotor retardation, and hypotonia. The three boys showed hypogenitalism and in two patients there were inverted nipples and subcutaneous fat pads. In the three patients tested, serum IgG was found to be decreased. The typical biochemical feature of this disorder is an accumulation of dolichyl pyrophosphate-$Man_5GlcNAc_2$ in fibroblasts. Eight different mutations were found in this small patient group.

Glucosyltransferase II Deficiency (CDG-Ih)

Five patients have been reported from four families. In four, there was no neurological involvement. Three showed a severe disease with dysmorphy and multiorgan failure resulting in death between three days and three months. The fourth patient had a milder presentation with hepatomegaly, protein-losing enteropathy, severe hypoalbuminemia, edema, and ascites. The fifth patient showed prominent central nervous system involvement (severe microcephaly, epilepsy, leukoencephalopathy) in addition to protein-losing enteropathy, and hepatic and renal disease.

Dolichyl pyrophosphate-$Glc_1Man_9GlcNAc_2$ was the accumulating lipid-linked oligosaccharide in fibroblasts.

Mannosyltransferase II Deficiency (CDG-Ii)

The clinical picture in the one reported patient was characterized by severe psychomotor retardation, infantile spasms, bilateral iris coloboma, and unilateral cataract. Brain magnetic resonance imaging showed severe dysmyelination. This is the first reported CDG in one of the early steps of dolichol-linked oligosaccharide biosynthesis at the cytosolic side of the endoplasmic reticulum. Accumulation of $Man_1GlcNAc_2$-PP-dolichol and $Man_2GlcNAc_2$-PP-dolichol, and severely reduced activity of the mannosyltransferase elongating $Man_1GlcNAc_2$-PP-dolichol were found in fibroblasts.

Deficiency of UDP–GlcNAc:Dolichol Phosphate *N*-Acetylglucosamine 1-Phosphate Transferase (CDG-Ij)

This is a defect in the very first step of dolichol-linked oligosaccharide biosynthesis. It was identified in a girl with intractable infantile spasms, severe psychomotor retardation, hypotonia, and microcephaly. Due to the nature of the defect there is no accumulation of lipid-linked oligosaccharides in fibroblasts.

Mannosyltransferase I Deficiency (CDG-Ik)

The four reported patients were boys showing epilepsy, severe psychomotor retardation, and variable features such as dysmorphy, liver dysfunction, cardiomyopathy, nephrotic syndrome, hypogonadism, and depletion of β-cells. Lipid-linked oligosaccharide analysis showed an accumulation of $GlcNAc_2$-PP-dolichol in the patients' fibroblasts.

Mannosyltransferase VII/IX deficiency (CDG-Il)

This is a defect in the only mannosyltransferase with a dual function. The reported patient presented with psychomotor retardation, microcephaly, seizures, axial hypotonia, and hepatomegaly. The lipid-linked oligosaccharide profile revealed an accumulation of the precursors $Man_6GlcNAc_2$-PP-dolichol and Man_8-$GlcNAc_2$-PP-dolichol.

N-Acetylglucosaminyltransferase II Deficiency (CDG-IIa)

Four patients have been reported with this first identified N-glycan processing defect. Besides neurological involvement (psychomotor retardation, epilepsy, behavioral disturbances) they presented with craniofacial dysmorphy, skeletal abnormalities, gastrointestinal disturbances, and growth retardation. Laboratory investigation showed increased glutamic oxaloacetic transaminase and normal glutamic pyruvic transaminase, a feature which is also seen in other CDG-II group patients. Isoelectrofocusing of serum transferrin showed a type 2 pattern (with nearly absent tetrasialotransferrin) and structural analysis of the transferrin oligosaccharides indicated a monoantennary *N*-acetyllactosamine type glycan. Thus, contrary to the structurally normal sialotransferrins in the CDG-I group or assembly disorders, the sialotransferrins of this and some other CDG-II diseases have abnormal glycan structures.

Glucosidase I Deficiency (CDG-IIb)

This CDG was discovered in an infant with dysmorphy, hypotonia, and epilepsy. The urine showed an abnormal oligosaccharide identified as the tetrasaccharide [Glc(α1-2) Glc(α1-3) Glc(α 1-3) Man]. This suggested a defect in glucosidase I, the first step in the N-glycan processing. Remarkably, IEF of serum transferrin was normal, probably due to a proven upregulation of ER mannosidase.

Golgi GDP-Fucose Transporter Deficiency (CDG-IIc)

This is the first reported CDG due to a nucleotide sugar transporter defect, specifically in a Golgi-localized, GDP-fucose transporter. The few reported patients suffered from leukocyte adhesion deficiency type II syndrome (LAD II) comprising craniofacial dysmorphy, severe psychomotor, and growth retardation as well as recurrent bacterial infections with persistent highly elevated leukocytosis. Due to hypofucosylation, neutrophils of these patients lack sialyl-Lewis X, a fucose-containing carbohydrate ligand of the selectin family of cell adhesion molecules that is required for the recruitment of neutrophils to infection sites. This hypofucosylation only affects N-glycans because GDP-fucose transport and O-fucosylation also occur in the ER. In one of the patients, fucose administration corrected the hypofucosylation and resulted in clinical improvement while other patients did not respond. Three kinds of molecular defects have been reported: mutations that affect the region for Golgi localization, mutations that affect the region necessary for the enzymatic function, and a truncation defect involving both regions.

β-1,4-Galactosyltransferase 1 Deficiency (CDG-IId)

Deficiency of this β-1,4-galactosyltransferase isoenzyme has been reported in a child with psychomotor retardation, Dandy–Walker malformation, and myopathy.

Golgi CMP–Sialic Acid Transporter Deficiency (CDG-IIf)

The one reported patient presented a very restricted and unusual phenotype with massive bleeding of the posterior chamber of the right eye and cutaneous bleeding as well as macrothrombocytopenia, neutropenia, and complete lack of the sialyl-Le antigen on polymorphonuclear cells. He also suffered from episodes of respiratory distress and opportunistic infections. Complications after bone marrow transplantation led to death at 37 months. Remarkably, plasma transferrin and other glycoproteins showed a normal glycosylation pattern possibly due to different degrees of altered splicing of the Golgi CMP–sialic acid transporter in different cells, but isoelectrofocusing of serum apolipoprotein CIII (containing only O-glycans) was abnormal.

Genetic Diseases of Protein O-Glycosylation

Defects have been identified in *O*-*N*-acetylgalactosaminylglycan synthesis (familial hyperphosphatemic tumoral calcinosis), in *O*-xylosylglycan synthesis (progeroid variant of Ehlers–Danlos syndrome and hereditary multiple exostoses), in *O*-mannosylglycan synthesis (Walker–Warburg syndrome and muscle–eye–brain disease), and in *O*-fucosylglycan synthesis (spondylocostal dysostosis type 3). They will be shortly summarized further.

Familial Hyperphosphatemic Tumoral Calcinosis

Patients suffer from recurrent, painful calcified subcutaneous masses that can be as large as 1 kg. This can be complicated by secondary infections and incapacitating mutilations. The hyperphosphatemia is due to increased renal phosphate retention. The physiopathology is not clear but is probably secondary to hypoglycosylation of a phosphatonin. The basic defect is in isoform 3 of the *N*-acetylgalactosaminyltransferase family comprising more than twenty members.

Progeroid Variant of Ehlers–Danlos Syndrome

This defect in β-1,4-galactosyltransferase 7 has been discovered in three patients from two families with a premature aging phenotype and hyperelastic skin in addition to microcephaly and joint hyperlaxity. The defect disrupts the trisaccharide linker region of glycosaminoglycans (O-linked xylose–galactose–galactose), specifically in the attachment of the first galactose to xylose.

Hereditary Multiple Exostoses

This is a long known disease and the only known CDG with autosomal dominant inheritance. It is characterized by osteochondromas of the ends of long bones. These tumors are often present at birth, their growth slows during adolescence, and stops in adulthood. Only a small percentage of these lesions show malignant degeneration. Compression of peripheral nerves and blood vessels may cause complications. The basic defect is in the Golgi-localized EXT1/EXT2 complex, which has both glucuronyltransferase and *N*-acetyl-D-hexosaminyltransferase activities involved in the polymerization of heparan sulphate.

Walker–Warburg Syndrome

This is a rare neuronal migration disorder characterized by brain and eye involvement associated with congenital muscular dystrophy. Male patients often have testicular defects. It is a very severe disease that usually runs a fatal course before the age of one year. Psychomotor development is absent. The brain lesions consist of cobblestone lissencephaly, agenesis of the corpus callosum, cerebellar hypoplasia, hydrocephaly, and sometimes an encephalocoele. In this disorder there is an aberrant glycosylation of α-dystroglycan, an external membrane protein expressed in muscle, brain, and other tissues. Most glycans of this heavily glycosylated protein seem to be O-linked via mannose and they control the interaction with extracellular matrix proteins. Disrupted glycosylation of α-dystroglycan (and probably other glycoproteins) results in loss of this interaction and hence in progressive muscle degeneration and abnormal neuronal migration (overmigration) in the brain. A minority of patients with this syndrome show mutations in the *POMT1* gene coding for protein *O*-mannosyltransferase 1. This enzyme catalyzes the first step in the synthesis of the *O*-mannose-linked core Galβ1-4GlcNAcβ1-2Man-*O*-Ser/Thr.

In some patients with this disorder, mutations have been found in the gene for protein *O*-mannosyltransferase 2, and in others mutations in the fukutin and the fukutin-related protein genes. Together these four defects explain only about 30% of patients with Walker–Warburg syndrome. On the other hand it should be noted that *POMT1* mutations not only cause Walker–Warburg syndrome but also limb girdle muscular dystrophy with mild mental retardation.

Muscle–Eye–Brain Disease

This is a neuronal migration/congenital muscular dystrophy syndrome similar to Walker–Warburg syndrome but less severe. As in Walker–Warburg syndrome, skeletal muscle shows almost no α-dystroglycan by western blotting and immunohistochemical analysis with an antibody that recognizes a glycosylated epitope. These patients show mutations in the gene encoding protein *O*-mannosyl-β1,2-*N*-acetylglucosaminyltransferase I, catalyzing the second step in the synthesis of the *O*-mannosylglycan core.

Spondylocostal Dysostosis Type 3

This type of spondylocostal dysostosis has very recently been identified as a notch pathway defect more specifically in Lunatic Fringe, an *O*-fucose-specific β-1-3-N-acetylglucosaminyltransferase. Patients show a severe vertebral phenotype with malsegmentation due to disruption of somitogenesis. Somites are precursors of the axial skeleton and associated musculature. Spondylocostal dystosis type 1 and 2 are due to mutations in other members of the notch pathway (*DLL3* and *MESP2*) indicating that this pathway is absolutely required for correct patterning of the axial skeleton.

Table 2 Human hereditary O-glycosylation disorders

Name	*Defective gene*	*Defective protein*
Defect in O-N-acetylgalactosaminylglycan synthesis	*GALNT3*	ppGaNTase-T3[a]
Familial tumoral calcinosis		
Defects in O-xylosylglycan synthesis	*B4GALT7*	β-1,4-Galactosyltransferase 7
Progeroid variant of Ehlers–Danlos syndrome	*EXT1/EXT2*	Glucuronyltransferase/*N*-acetyl-D-hexosaminyltransferase
Hereditary multiple exostoses		
Defects in O-mannosylglycan synthesis	*POMT1*	*O*-Mannosyltransferase 1
Walker–Warburg syndrome	*POMGnT1*	*O*-Mannosyl-β-1,2-*N*-acetylglucosaminyltransferase 1
Muscle–eye–brain disease		
Defects in O-fucosylglycan synthesis	*LFNG*	*O*-Fucose-specific β-1,3-*N*-acetylglucosaminyltransferase
Spondylocostal dysostosis type 3		

[a]UDP-*N*-acetyl-α-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferase 3.

Genetic Diseases of Protein N- and O-Glycosylation

Conserved Oligomeric Golgi Complex Subunit 7 Deficiency (CDG-IIe)

COG7 deficiency was identified in two siblings with dysmorphy (particularly of the face), encephalopathy, and cholestatic liver disease. They died at 5 and 10 weeks, respectively. Serum transferrin isoelectrofocusing showed a type 2 pattern.

The discovery of this disease opens a new avenue in the CDG field because it affects a protein complex (COG: conserved oligomeric Golgi complex, containing 8 subunits) involved in trafficking and function of the glycosylation machinery. Studies of fibroblast glycoproteins showed a partial N- and O-glycosylation defect caused by a decreased transport of CMP–sialic acid and UDP–galactose into the Golgi as well as a reduced activity of two glycosyltransferases involved in the galactosylation and sialylation of O-glycans. Localization of the COG subunits was investigated by indirect immunofluorescence and an abnormal distribution of COG5, COG6, and COG8 (diffuse cytoplasmic staining instead of Golgi localization) was found. The basic defect was eventually localized to the COG7 subunit (Figure 2 and Table 2).

Conserved Oligomeric Golgi Complex Subunit 1 Deficiency

The one reported patient, a girl, had a mild phenotype compared to the severe neurological and hepatic involvement in the COG7-deficient siblings (mild psychomotor retardation, slight brain hypotrophy, and minor dysmorphy). As the latter, this patient showed hyposialylation as well as hypogalactosylation.

See also: Congenital Muscular Dystrophy; Peroxisomal Disorders and Neurological Disease.

Further Reading

Foulquier F, Vasile E, Schollen E, et al. (2006) Conserved oligomeric golgi complex subunit 1 deficiency reveals a previously uncharacterized congenital disorder of glycosylation type II. *Proceedings of the National Academy of Sciences of the United States of America* 103: 3764–3769.

Freeze HH (2002) Human disorders in N-glycosylation and animal models. *Biochimica et Biophysica Acta* 1573: 388–393.

Jaeken J, Vanderschueren-Lodeweyckx M, Casaer P, et al. (1980) Familial psychomotor retardation with markedly fluctuating serum proteins, FSH and GH levels, partial TBG-deficiency, increased serum arylsulphatase A and increased CSF protein: A new syndrome? *Pediatric Research* 14: 179.

Jaeken J, Van Eijk HG, van der Heul C, et al. (1984) Sialic acid-deficient serum and cerebrospinal fluid transferrin in a newly recognized genetic syndrome. *Clinica Chimica Acta* 144: 245–247.

Jaeken J, Schachter H, Carchon H, et al. (1994) Carbohydrate-deficient glycoprotein syndrome type II: A deficiency in golgi localised *N*-acetyl-glucosaminyltransferase II. *Archives of Disease in Childhood* 71: 123–127.

Jaeken J, Matthijs G, Saudubray JM, et al. (1998) Phosphomannose isomerase deficiency: A carbohydrate-deficient glycoprotein syndrome with hepatic-intestinal presentation. *American Journal of Human Genetics* 62: 1535–1539.

Jaeken J, Matthijs G, Carchon H, and Van Schaftingen E (2001) Defects of N-glycan synthesis. In: Scriver CR, Beaudet WS, Sly WS, and Valle D (eds.) Childs B, Kinzler KW, and Vogelstein, B (assoc. eds.) *The Metabolic and Molecular Bases of Inherited Disease,* 8th edn., pp. 1601–1621. New York: McGraw-Hill.

Jaeken J (2003) Congenital disorders of glycosylation (CDG): It's all in it! *Journal of Inherited Metabolic Disease* 26: 99–118.

Jaeken J and Carchon H (2004) Congenital disorders of glycosylation: A booming chapter of pediatrics. *Current Opinion in Pediatrics* 16: 434–439.

Lübke T, Marquardt T, Etzioni A, Hartmann E, von Figura K, and Körner C (2001) Complementation cloning identifies CDG-IIc, a new type of congenital disorders of glycosylation, as a GDP-fucose transporter deficiency. *Nature Genetics* 28: 73–76.

Martinez-Duncker I, Dupré T, Piller V, et al. (2005) Genetic complementation reveals a novel human congenital disorder of glycosylation of type II, due to inactivation of the golgi CMP-sialic acid transporter. *Blood* 105: 2671–2676.

Marquardt T and Denecke J (2003) Congenital disorders of glycosylation: Review of their molecular bases, clinical presentations and specific therapies. *European Journal of Pediatrics* 162: 359–379.

Matthijs G, Schollen E, Pardon E, et al. (1997) Mutations in PMM2, a phosphomannomutase gene on chromosome 16p13, in carbohydrate-deficient glycoprotein type I syndrome (Jaeken syndrome). *Nature Genetics* 16: 88–92.

Matthijs G, Schollen E, and Van Schaftingen E (2004) The prenatal diagnosis of congenital disorders of glycosylation (CDG). *Prenatal Diagnosis* 24: 114–116.

Niehues R, Hasilik M, Alton G, et al. (1998) Carbohydrate-deficient glycoprotein syndrome type Ib: Phosphomannose isomerase deficiency and mannose therapy. *Journal of Clinical Investigation* 101: 1414–1420.

Panneerselvam K and Freeze HH (1996) Mannose corrects altered N-glycosylation in carbohydrate-deficient glycoprotein syndrome fibroblasts. *Journal of Clinical Investigation* 97: 1478–1487.

Schenk B, Imbach T, Frank CG, et al. (2001) *MPDU1* mutations underlie a novel human congenital disorder of glycosylation, designated type If. *Journal of Clinical Investigation* 108: 1687–1695.

Sparrow B, Chapman G, Wouters MA, et al. (2006) Mutation of the *Lunatic Fringe* gene in humans causes spondylocostal dysostosis with a severe vertebral phenotype. *American Journal of Human Genetics* 78: 28–37.

Topaz O, Shurman DL, Bergman R, et al. (2004) Mutations in GALNT3, encoding a protein involved in O-linked glycosylation, cause familial tumoral calcinosis. *Nature Genetics* 36: 579–581.

Wu X, Steet RA, Bohorov O, et al. (2004) Mutation of the COG complex subunit gene COG7 causes a lethal congenital disorder. *Nature Medicine* 10: 518–523.

Relevant Website

http://www.kuleuven.be – Katholieke Universiteit Leuven.

Goal-Directed Behavior Theories

S B Ostlund and B W Balleine, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The capacity for goal-directed action allows us and other animals to develop new behavioral strategies to control events in our environment so as to gain access to valued goals and avoid those we no longer value or that we find aversive. The adaptive significance of this capacity becomes most clear when these processes break down, as they do in disorders of 'voluntary' movement (e.g., Parkinson's disease) and in disorders involving intrusive involuntary movement (e.g., Tourette's syndrome, obsessive compulsive disorder, Huntington's disease, and in the behavioral stereotypes associated with clinical syndromes such as schizophrenia). In spite of the obvious importance of this capacity, the determinants of goal-directed action in humans are not fully understood. Many contemporary theorists have adopted a cognitive approach and developed accounts in terms of agents planning various behavioral strategies on the basis of their knowledge of action–outcome relationships, and selecting courses of action by some rational criterion after an evaluation of various alternatives. In fact, there is very little objective evidence that these kinds of factors play a direct role in human action. The main source of data regarding the fundamental determinants of goal-directed action is studies of animal models, perhaps the best known of which is the instrumental conditioning paradigm.

Goal-Directed, Instrumental Actions

The Contingency Criterion

In instrumental conditioning, animals are taught to perform arbitrary actions in order to gain access to some or other valued goal. In the paradigm case, hungry rats are taught to press a freely available lever to gain access to a food reward. But what is learned in instrumental conditioning? Is lever pressing by rats truly determined by the animals' knowledge of the relationship between the action and its consequences? Is it truly goal directed or is it merely a stimulus-elicited reflex? What tests should we apply to lever pressing – indeed, to any putative goal-directed action – in order to answer this question?

Such questions have been addressed in recent years using a variety of experimental approaches. For example, if the performance of an action is determined by the belief that a specific action is required in order to gain access to some goal or other, then the action should never be acquired or should be completely withheld in a situation in which performing the action actually precludes access to the goal; that is, performance should adjust appropriately to the imposition of an omission schedule. Similarly, to be regarded as goal directed, actions should surely respect the causal relation between the action and its consequences: any action acquired and/or maintained in a situation in which the occurrence of its particular outcome is equally likely whether the action is performed or not should not be considered goal directed.

There is, in fact, considerable evidence that free operant lever pressing in rats is sensitive to the causal relation between the action and its consequences. Not only will rats stop responding if lever pressing no longer delivers a valued food, they will stop responding even faster if their performance cancels otherwise freely available food (i.e., leads to the omission of the outcome). Assessing the ability of rats to encode the causal relationship between an action and its consequences has been difficult to achieve because research on causal judgment has been developed in humans often using verbal measures of causality rating. Nevertheless, causal relationships have been argued to approximate the contingency between a cause and an effect derived from the interaction of two probabilities: the probability of the effect given that the cause has been presented, and the probability of the effect given the absence of the cause. In instrumental conditioning these probabilities become (1) the probability of an outcome (O) given that an action (A) is performed, or $p(\mathrm{O/A})$, and (2) the probability of an outcome given that an action has not been performed, or $p(\mathrm{O/noA})$ (see Figure 1). In this situation, the net positive contingency between an action and an outcome ($V_{\mathrm{A-O}}$) is given by

$$V_{\mathrm{A-O}} = p(\mathrm{O/A}) - p(\mathrm{O/noA})$$

Using this approach, numerous experiments have demonstrated both that lever press performance is increased as $p(\mathrm{O/A})$ is increased and, conversely, that lever pressing is reduced as $p(\mathrm{O/noA})$ is increased. Furthermore, a number of studies have confirmed that rats are sensitive to a selective change in one action–outcome contingency in a situation in which another contingency is maintained intact. For example, in one experiment, hungry rats were trained to perform two actions, lever pressing and chain pulling, with one action earning food pellets and the other earning a starch solution, with a fixed probability in

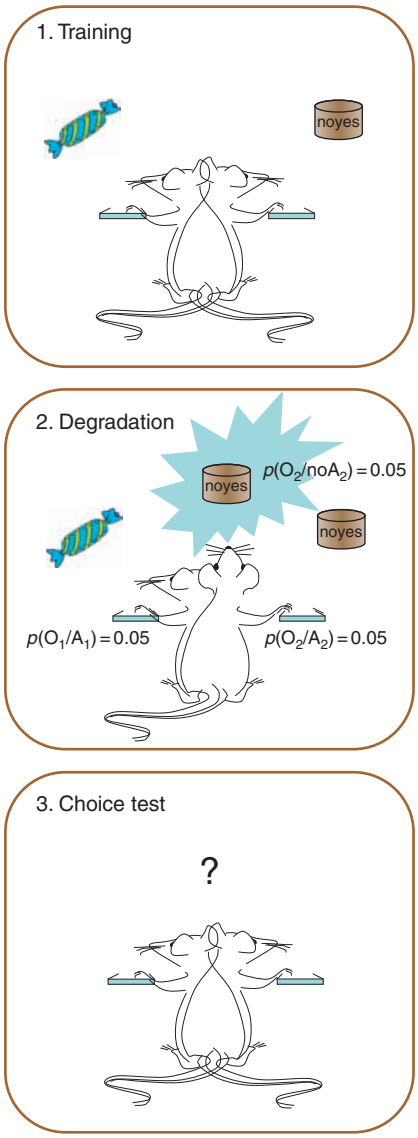

Figure 1 Contingency degradation. Studies in animals examining the sensitivity of an action to changes in the causal relation to its consequences typically modify the contingency between an action (A) and its outcome (O). Although not strictly identical, contingency degradation approximates changes in the causal relationship. Rats are typically trained (panel 1) on two actions (e.g., two levers), each earning a different outcome (grain pellets and sugar), with each action delivering its appropriate outcome at a fixed probability, $p(O_1/A_1)$. The contingency between one action and its outcome is then degraded (panel 2) by delivering that outcome at a fixed probability when the action is not performed, $p(O_1/noA_1)$. Finally, the effect of the degradation is assessed in a choice test (panel 3) in which neither the earned nor the free outcome is delivered.

each second in which a response was performed. After this training, in addition to being earned by performing its associated action, one of the two outcomes was also delivered in the absence of its associated action and at the same probability. As a consequence, the delivery of that outcome was equally probable in each second of the session, regardless of whether its associated action had been performed. If, in this situation, lever pressing and chain pulling were determined by the causal relationship between the actions and their consequences, then degrading one of the action–outcome contingencies in this way should result in a reduction in the performance of that specific action relative to performance of the other action. In fact, this result has been reported on many occasions.

Reward Value

The Goal Criterion

The foregoing discussion suggests that, in instrumental conditioning, animals encode the relationship between specific actions and their outcomes and remain sensitive to changes in this causal relationship. It has long been recognized, however, that encoding an action–outcome relationship is not sufficient to determine the performance of an action. Any learning that takes the form 'action A leads to outcome O' can be used both to perform A and to avoid performing A. It is now well established that the rats' evaluation of the reward value of the outcome – that is, its affective and motivationally relevant properties – can influence the performance of actions instrumental to outcome delivery. This has been referred to as the 'goal criterion of goal-directed action'; surely, if an action were to be maintained despite a significant reduction in the value of the goal, then we should doubt that the action was goal directed.

There is considerable evidence from studies of instrumental conditioning that actions such as lever pressing satisfy the goal criterion. This evidence comes largely from studies assessing posttraining devaluation of the instrumental outcome. For example, in a number of studies of this kind, hungry rats have been trained to press a lever and pull a chain, with one response earning, say, food pellets and the other, say, a starch solution. After this training, the reward value of one of the two outcomes is then reduced – typically by specific satiety, in which the rats are allowed to eat a large quantity of one or other outcome, but also by taste aversion learning, in which consumption of one or other outcome is followed by the induction of illness using an injection of lithium chloride. Both of these treatments are well known to reduce the consumption of a specific food relative to other, non-pre-fed, or unpoisoned, foods. After the devaluation treatment, the rats are then given a choice test on the levers and chains conducted in extinction (i.e., in the absence of the delivery of either food

outcome; see Figure 2). If performance of the actions during training is maintained by the motivational value of their respective outcomes, then devaluation of one outcome should reduce subsequent performance of its associated action in the extinction test relative to the other action. This is exactly the result that is commonly reported; devaluing one of the food rewards by specific satiety or taste aversion learning selectively and profoundly reduces the performance of the action that, in training, delivered that outcome, compared to the action trained with the nondevalued outcome.

Evidence for this claim has also come from studies assessing the effect of shifts in primary motivation on instrumental performance. Posttraining shifts, such as from hunger to satiety or from hunger to thirst, often

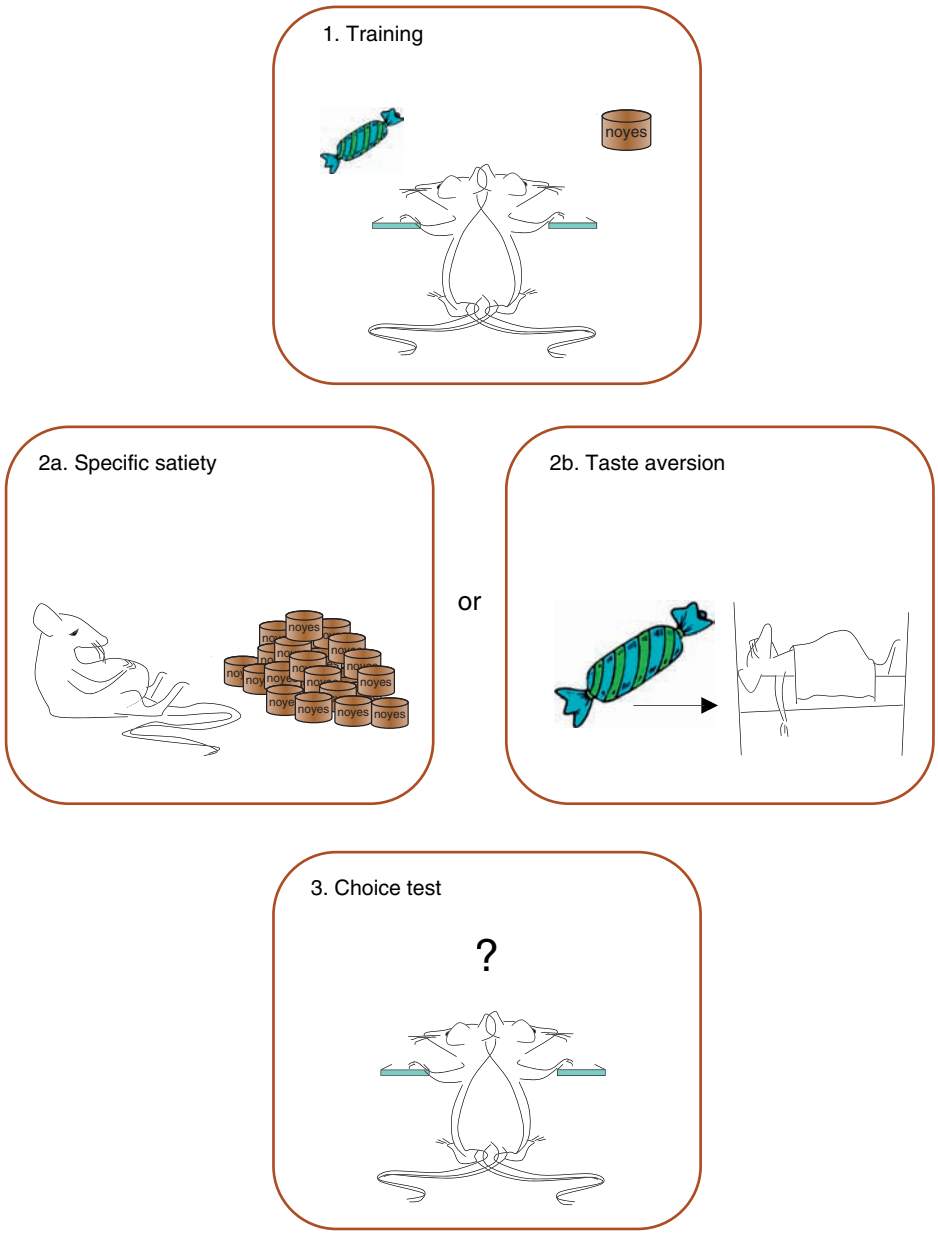

Figure 2 Outcome devaluation. Studies assessing effects of changes in the value of the goal on the performance of goal-directed actions typically use some form of outcome devaluation procedure. Rats are trained on two actions, each earning a different outcome (panel 1). After this phase, one or other outcome is devalued either by specific satiety (panel 2a) or taste aversion learning (panel 2b). When the outcome is no longer consumed, the tendency of the rats to perform the two actions is tested back in the training situation, except that now responding on the lever does not deliver the food. Hence, changes in the tendency to perform one or other action must reflect the integration of the training and devaluation phases.

have very little direct impact on the performance of goal-directed actions until the effect of the motivational shift on the reward value of goal has been made explicit through direct consummatory contact in the new motivational state (i.e., by allowing animals to engage in what is referred to as incentive learning). Satiety, for example, appears to reduce the performance of goal-directed actions that gain access to specific food rewards, not because it reduces general drive or arousal, but rather because it makes that food less palatable when it is contacted in the sated state (i.e., it reduces the reward value of that food). Likewise, a dry food reward appears to become aversive when it is contacted by animals that are thirsty, whereas a nutritive fluid reward, such as a starch solution or a sugar solution, becomes more highly valued, and the performance of actions trained with either the dry or liquid nutrients has been found to reflect this experienced value, even when assessed in a choice extinction test.

The general nature of the role of incentive learning in instrumental performance has been confirmed for a number of different motivational systems and in a number of devaluation paradigms. For example, in addition to the effects of satiety on the value of nutritive outcomes, incentive learning has been found to play a role in changes in the value of specific outcomes produced by (1) conditioned taste aversion, (2) posttraining shifts in food deprivation, (3) posttraining shifts in water deprivation, (4) changes in outcome value mediated by drug states, and (5) changes in the value of thermoregulatory and sexual rewards. In all of these cases, evidence suggests that, after a shift in primary motivation, learning about the effect of that shift on the incentive value of an instrumental outcome through consummatory contact increases the influence of this shift on subsequent performance. With respect to the processes controlling performance, this analysis suggests that, in instrumental conditioning, rats encode the relationship between actions and goals, as well as the current incentive value of the goal, and, more importantly, that they are able to integrate these sources of information to select a course of action. Indeed, it is in the evidence for the control over performance exerted by the integration of these two sources of information that the fundamentally goal-directed quality of instrumental conditioning in rats is most forcefully revealed.

Goal-Directed Learning: Outcome Encoding and Action Selection

The contingency and goal criteria that, when met, justify the description of an action as goal directed, claim that such actions respect both their causal relation to the goal and the reward value of the goal. Hence they tend to imply that the critical learning process that mediates the acquisition of goal-directed actions involves the association of an action with the outcome. Beyond that, however, they do not constrain theories of this learning process, and, unsurprisingly, a number of different theories have been offered over the years to describe the goal-directed nature of instrumental performance. These theories can be divided into two broad categories based on the their assumptions about how the outcome is encoded.

The first category, derived from traditional stimulus–response theory, focuses on the outcome as a stimulus and assumes that stimuli signaling the training outcome can acquire the capacity to elicit an expectation composed solely of that outcome's sensory features (e.g., texture, shape, flavor, and so on). When reinforced by the delivery of food, or some other positive event, the instrumental response becomes associated with this outcome expectation and, as a consequence, anything that alters outcome expectancy should also produce a concomitant change in the performance of the action associated with this expectancy. A variant on this suggestion argues that, rather than an expectancy of some subsequent outcome, actions become associated with the sensory features of any outcome earned prior to their performance, something that allows that outcome itself, or any stimulus that activates the representation of that outcome, to exert control over the performance of the action. According to this account, therefore, it is the antecedent relationship between a previously earned outcome and the response performed next that is encoded during instrumental training (i.e., an outcome–response, or O–R, association; see **Figure 3**). Note, however, that, since this account assumes that the outcome is encoded purely as a sensory event, it also tends to predict that, unless it is agreed that a change in the incentive value of the outcome modifies the ability of an animal to process those sensory features, outcome devaluation will have no effect on the outcome's ability to influence subsequent response selection.

Alternatively, a distinct category of theory holds that it is the consequent relationship, between a response and the outcome that it earns, that is learned during instrumental conditioning, and that mediates the acquisition of goal-directed action (i.e., a response–outcome, or R–O, association; see **Figure 3**). In general, R–O theories emphasize the status of the outcome as a goal and so often make the explicit assumption that the influence of the outcome on response selection depends on its current incentive value. Although it is possible to imagine a version of the O–R theory that can account for outcome

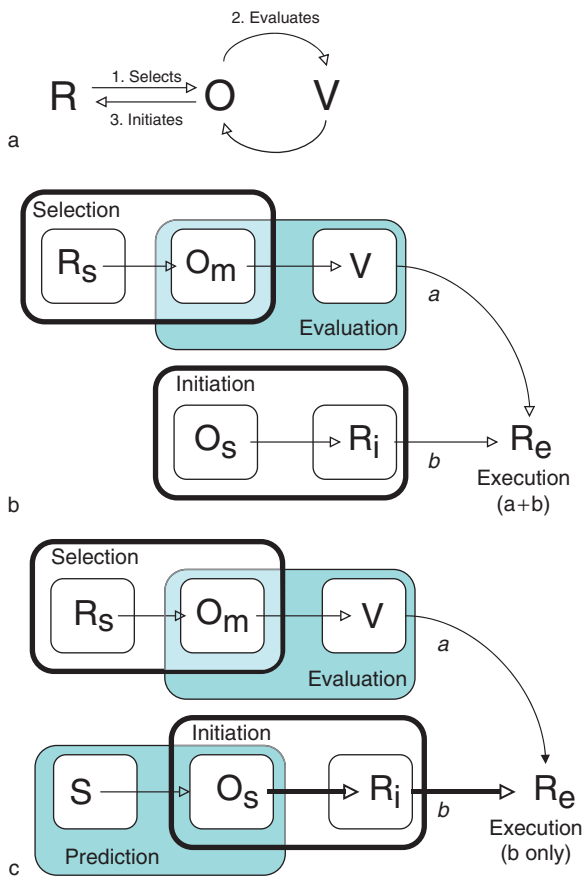

Figure 3 The associative structure of goal-directed action. Associations that current evidence suggests are formed between environmental stimuli (S), actions (R), and outcomes or goals (O) and goal value (V) during the course of acquisition or performance of goal-directed action. (a) Current research provides evidence for both R–O and O–R associations in the control of performance. Whereas the O–R association appears not to engage or be influenced by changes in outcome value, the R–O association strongly engages these evaluative processes. (b) The joint reliance on R–O (pathway *a*) and O–R (pathway *b*) processes suggests that these processes interact to control response execution (R_e). (c) The impact of pathway *a* can be overcome by increasing the contribution of pathway *b* to performance either by presenting a stimulus separately trained with the outcome (i.e., S–O, as in Pavlovian-instrumental transfer experiments) or by strengthening the association itself by selective reinforcement through overtraining. R, response; R_i, initiated response; R_s, selected response; R_e, executed response; O, outcome; O_s, sensory properties of the outcome; O_m, motivational properties of the outcome; V, value.

devaluation, the well-established sensitivity of instrumental performance to outcome devaluation is clearly consistent with the R–O account. Indeed, experiments contrasting the relative contribution of O–R and R–O relationships to outcome devaluation have generally confirmed that this effect is dominated by the R–O component.

The outcome devaluation effect is not, however, the only means of evaluating the nature of outcome encoding. Other findings relevant to an assessment of theories of goal-directed learning have come from studies investigating the facilitation of instrumental performance resulting from the noncontingent presentation of either the training outcome itself (i.e., the reinstatement effect) or a stimulus that predicts that outcome (i.e., the Pavlovian-instrumental transfer effect). Generally, if it is assumed that the connections between action and response are bidirectional, both the O–R and the R–O account of outcome encoding predict that the noncontingent presentation of the training outcome, or a stimulus that predicts that outcome, will trigger the selection of its associated action.

More recently, attempts to differentiate the two accounts have been developed by explicitly training hungry rats to use stimuli (S1 or S2) paired with specific outcomes (O1 or O2) as discriminanda signaling which of two actions (R1 or R2) will be rewarded either with the same outcome anticipated by the stimulus (the congruent case) or with a different stimulus (the incongruent case):

$$\text{Congruent: S1} - \text{O1}, \text{S2} - \text{O2}$$

$$\text{S1: R1} - \text{O1}, \text{R2}-; \text{S2: R1}-, \text{R2} - \text{O2}$$

$$\text{Incongruent: S1} - \text{O1}, \text{S2} - \text{O2}$$

$$\text{S1: R1} - \text{O2}, \text{R2}-; \text{S2: R1}-, \text{R2} - \text{O1}$$

In line with predictions from O–R theory, discriminative control has been found to be greater in the former case – that is, when the outcome anticipated by the cue is also used to reward the performance of the correct action. Nevertheless, devaluation of one or other outcome (O1 or O2) has been found to affect the response with which it is associated as an outcome – that is, if, say, O1 is devalued, R1 is reduced relative to R2 in the congruent case but R2 is reduced relative to R1 in the incongruent case.

A similar approach using the outcome itself as the discriminative stimulus has been applied to this problem:

$$\frac{\text{Congruent: O1: R1} - \text{O1}; \text{O2: R2} - \text{O2}}{\text{Incongruent: O2: R1} - \text{O1}; \text{O1: R2} - \text{O2}}$$

Thus, during training, each action is either signaled and followed by the same outcome (congruent training) or signaled and followed by different outcomes (incongruent training). As in the previous study, the performance of the incongruent group can be used to contrast the relative contribution of O–R and R–O relationships to the control of instrumental performance. It has been shown that, after such training, the devaluation of one or other outcome (e.g., O1)

tends to reduce the performance of the action that had been followed by that outcome (i.e., R1) relative to the action that had been signaled by that outcome (R2). This approach has also been used to assess the associative structure underlying the outcome's ability to reinstate extinguished instrumental performance. For example, it has been shown that after incongruent training and subsequent extinction on both actions, the noncontingent delivery of one or other outcome (O1) preferentially reinstates the performance of the action that had been signaled by that outcome (R2) relative to the outcome that had been followed by that outcome (R1). Again, these studies provide further evidence for both R–O and O–R processes.

Further evidence of variation in outcome encoding is provided by the Pavlovian-instrumental transfer effect. A number of studies have shown that a stimulus paired with a particular outcome can exert a selective excitatory effect over the performance of an action that has also been trained with that outcome, relative both to baseline performance and to performance of an action trained with a different outcome. Although transfer effects are compatible with O–R theory, they can also be explained in terms of action–outcome associations again, if it is assumed that stimuli associated with the outcome retrieve the action through a bidirectional action–outcome association. However, a growing body of evidence suggests that transfer effects are mediated by an associative structure different from that mediating the influence of the reward value of the goal on instrumental performance. For instance, a number of experiments have now been conducted assessing the effects of outcome devaluation on Pavlovian-instrumental transfer and have found that, although outcome devaluation reliably suppresses the performance of actions that deliver the now devalued outcome, it has no detectable effect on the facilitation of performance produced by a stimulus that predicts that outcome (i.e., the transfer effect). Using a similar approach, we have recently observed that the outcome selectivity of instrumental reinstatement is similarly insensitive to outcome devaluation.

This pattern of results has largely resisted interpretation in terms of a single associative process. Rather, the kind of model suggested reflects a combination of R–O and O–R associative processes, the R–O process mediating the acquisition and selection of actions based on the value of their consequences, and a separate O–R process mediating the initiation and execution of the action. This view of the relationship between the action selection, outcome evaluation, and action initiation processes is illustrated in **Figure 3**. **Figure 3**(a) illustrates the general principle captured by this combined model; the forward R–O association engages an evaluative process supporting the subsequent initiation and execution of the response. In **Figures 3**(b) and **3**(c) the relationship between these processes is more clearly explicated. In this view, actions can only be executed when motor excitation increases above a threshold. In the usual case, this requires both the selection of a valued action (i.e., one that delivers a rewarding outcome) and the initiation of that action through the O–R process. These two processes summate to drive activity in the motor system above threshold. The situation is slightly different in cases in which a stimulus independently associated with the outcome is presented. In this case, and consistent with the observed failure of outcome devaluation to affect transfer effects, the stimulus can drive the activation of the motor system by the O–R association sufficiently strongly to execute the action without the need of the R–O association. In fact, under certain conditions, the O–R association can grow sufficiently strong in and of itself to produce action execution in the absence of the R–O association. These cases are described briefly in the next section.

Actions and Habits

As mentioned earlier, the outcome devaluation and instrumental contingency degradation procedures provide useful diagnostic tests for whether or not a particular learned response should be considered a goal-directed action. Generally, the very great majority of the evidence suggests that free operant performance in rats meets these two criteria. Reports of negative results have not been uncommon, however. Nevertheless, these negative results have proved useful in establishing the boundary conditions determining when an action is goal directed and when it is not (i.e., when it is automatic, inflexible, stimulus bound, and habitual). For example, prior to the early 1980s, mostly failures to find an effect of outcome devaluation had been reported in the animal literature. One reason for these early failures appears to have been the use of interval schedules of reinforcement during training; early successful attempts to demonstrate the devaluation effect used either continuous or ratio schedules of reinforcement. And, indeed, in experiments comparing the impact of ratio and interval schedules, reliable devaluation emerged only when rats were trained on ratio schedules. Even with the use of a ratio schedule, however, instrumental performance may be rendered insensitive to devaluation and contingency degradation through repeated, unchanging training at asymptote – referred to as overtraining.

Although instrumental conditioning very often possesses the characteristics of a goal-directed action,

certain training conditions can facilitate the development of those responses into habits, the performance of which is insensitive to the consequences. One interpretation of this complex pattern of data is that, although instrumental performance is initially dominated by action–outcome learning, extensive training or training on interval schedules of reinforcement, by constraining experienced changes in the relationship between response rate and reinforcement rate, supports the acquisition of stimulus–response (S–R) associations and, hence, weakening in the relative control of performance by the R–O association. According to the S–R view, the outcome is thought to serve merely as a catalyst for strengthening the association between situational cues and the instrumental response and, when dominated by S–R associations of this kind, the outcome is prevented from playing a role in response selection.

Multiple Learning Systems?

Inevitably, the evidence referred to in the preceding discussion leads to the general conclusion that goal-directed action is multiply determined; whereas goal-directed learning involves encoding the relation between an action and its consequences, the overall rate of performance can also reflect input from S–R or (more specific) O–R processes. These two influences on performance may be thought to reflect control by two distinct kinds of representational structure: one truly goal directed, based on the encoding of the relationship between an action and it consequences (often described as a declarative representation), and a second sensorimotor or S–R association (described as a procedural representation because it directly reflects the use to which the representation is put in controlling behavior). The interaction of these two representational systems in the control of performance is, as described in Figure 3, likely to reflect a complex interaction between distinct learning processes at a number of levels. One of these has already been described in the way that response execution can be seen as reflecting summation of both R–O and O–R processes. Yet another is the interaction of reward and reinforcement processes. The outcome of a goal-directed action provides both an associate of the action in the formation of the R–O process and also a reinforcement signal that strengthens S–R associations. On the one hand, the outcome is represented as a valuable goal in terms of its specific motivational and emotional properties that allow it to serve an incentive. On the other hand, the outcome provides a general positive (or negative) signal that strengthens (or weakens) sensorimotor connections monotonically whenever it occurs, making a specific motor act more (or less) probable in the presence of that sensory event. Generally, therefore, it may be argued that these distinct learning systems differ not only in terms of the representational content of the learning they support, but also in terms of the distinct learning rules they obey: action–outcome causality (or contingency) in the case of goal-directed actions and response–reinforcer contiguity in the case of habits.

See also: Basal Ganglia: Habit; Conditioned Reflex; Conditioning: Theories; Neuropsychology of Primate Reward Processes; Operant Conditioning of Reflexes; Procedural Learning: Classical Conditioning; Psychopharmacology of Reward and Appetite in Rats; Reward and Learning; Reward Decision-Making; Social Interaction Effects on Reward and Cognitive Abilities in Monkeys.

Further Reading

Balleine BW (2001) Incentive processes in instrumental conditioning. In: Mowrer R and Klein S (eds.) *Handbook of Contemporary Learning Theories*, pp. 307–366. Hillsdale, NJ: Lawrence Erlbaum Associates.

Balleine BW (2004) Incentive behavior. In: Whishaw IQ and Kolb B (eds.) *The Behavior of the Laboratory Rat: A Handbook With Tests*, chap. 41, pp. 436–446. Oxford: Oxford University Press.

Balleine BW (2005) Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology & Behavior* 86: 717–730.

Balleine BW and Dickinson A (1998) Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37: 407–419.

Colwill RC and Rescorla RA (1986) Associative structures in instrumental learning. In: Bower GH (ed.) *The Psychology of Learning and Motivation,* vol. 20, pp. 55–104. New York: Academic Press.

Corbit LH and Balleine BW (2003) Pavlovian and instrumental incentive processes have dissociable effects on components of a heterogeneous instrumental chain. *Journal of Experimental Psychology: Animal Behavior Processes* 29: 99–106.

Dickinson A (1989) Expectancy theory in animal conditioning. In: Klein SB and Mowrer RR (eds.) *Contemporary Learning Theories*, pp. 279–308. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dickinson A (1994) Instrumental conditioning. In: Mackintosh NJ (ed.) *Animal Cognition and Learning*, pp. 45–79. London: Academic Press.

Dickinson A and Balleine BW (1994) Motivational control of goal-directed action. *Animal Learning & Behavior* 22: 1–18.

Dickinson A and Balleine BW (2002) The role of learning in motivation. In: Gallistel CR (ed.) *Learning, Motivation & Emotion, Volume 3 of Steven's Handbook of Experimental Psychology,* 3rd edn., pp. 497–533. New York: Wiley.

Holland PC (2004) Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavioral Processes* 30: 104–117.

Rescorla RA (1992) Response-outcome versus outcome-response associations in instrumental learning. *Animal Learning & Behavior* 20: 223–232.

Rescorla RA (1994) Transfer of instrumental control mediated by a devalued outcome. *Animal Learning & Behavior* 22: 27–33.

Gonadal Steroid Actions on Brain

S A Tobet and R J Handa, Colorado State University, Fort Collins, CO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Androgens and estrogens affect many structures in the brain and other areas of the nervous system. These gonadal steroids also undergo metabolism in several brain sites, and the interactions of the resulting array of steroids influence brain development, behavior, and neuroendocrine function. Since it consists of varied regions and cell types, which exhibit a spectrum of metabolic activities, the brain has been viewed as a second endocrine system in which steroid conversions and cellular actions occur. Many expressed proteins and molecular interactions in the nervous system resemble or are the same as those of classically defined endocrine tissues. These include enzymes that metabolize steroids, intracellular receptor proteins that are selective for different steroids, and regulatory proteins that help mediate steroid action.

Intracellular steroid receptors in neural tissues exhibit properties and characteristics similar to those in other tissues. In the absence of ligand, these proteins are soluble and are detectable in the cytosolic fractions of tissue homogenates following differential centrifugation. After binding to steroids, these receptors are less soluble and are more likely to be detectable in the pellets of sedimented tissue homogenates. Biochemical evidence indicates that both androgen and estrogen receptors have significant affinities for DNA, whether or not steroid is bound. However, when ligand bound, the equilibrium for the receptor to bind sites in cell nuclei is shifted toward higher affinity. Live cell imaging of chimeric receptor proteins that are tagged with fluorescent molecules shows that this change targets the receptors to specific sites in cell nuclei, correlating with selective gene expression. In general, many of these gonadal steroid receptors reside in cell nuclei even in the unoccupied state. However, steroid receptors are more and more frequently being found in additional cellular locations, particularly plasma membranes.

Experiments have revealed, in addition to actions in the nucleus, rapid electrical responses by neurons in the brain to specific steroids. There are also instances in which signal transduction cascades that begin at the cell surface are activated, leading to the phosphorylation of key intracellular proteins that regulate functions in cytoplasmic and nuclear compartments. Increased interest in finding receptors for steroids in neural membranes has led to a growing number of examples. In parallel with studies of these putative receptor sites is evidence for interactions of gonadal steroids with ion channel molecules. A class of steroid-related compounds called neurosteroids is produced in the brain, and these compounds interact with channel proteins, other neurotransmitter receptors, and possibly other classes of receptors. This continues to be an active and expanding area of investigation.

Investigators now know many details of the development, localization, and dynamics of interactions of the classical intracellular steroid receptors with physiologically circulating steroids. The proteins are products of a superfamily of nuclear receptors and related molecules. More than 300 sequences have now been attributed to this family of proteins, and a unifying system of nomenclature has been proposed. Public databases, the Nuclear Receptor Database and the Nuclear Receptor Signaling Atlas, have been created for cataloging and disseminating the rapidly expanding knowledge in this field. There are similarities in the gene sequences and protein structures for parts of these receptor molecules, and it is believed that androgen and estrogen receptors have evolved along a common phylogenetic path. Recent studies, however, indicate that they were first utilized as transcription factors and later evolved ligand dependence. Since estrogens are metabolized from androgens, it is reasonable to hypothesize that the hormonal system of agonist and receptor for one served as forerunner for the other, resulting in a more complex, interrelated system for regulation by androgens and estrogens. A major recent realization, and an expanding area of investigation, is the many forms of estrogen receptors. The original estrogen receptor is now termed alpha, as a beta form was discovered in 1996. Alternate forms of both of these have been described based on alternative splicing. Additional novel forms of estrogen receptors have been suggested, and estrogen binding by proteins not considered 'receptors' has also been reported.

In the brain, as elsewhere in the body, two primary synthetic routes metabolize androgens (**Figure 1**). These convert testosterone (T) to alternate, active androgen metabolites, including estrogens. The enzyme 5α-reductase catalyzes addition of two hydrogens across the delta^4 double bond in the A-ring of T, forming the major reduced form of androgen, 5α-dihydrotestosterone (DHT). Like T, DHT can bind to androgen receptors (ARs). The enzyme aromatase

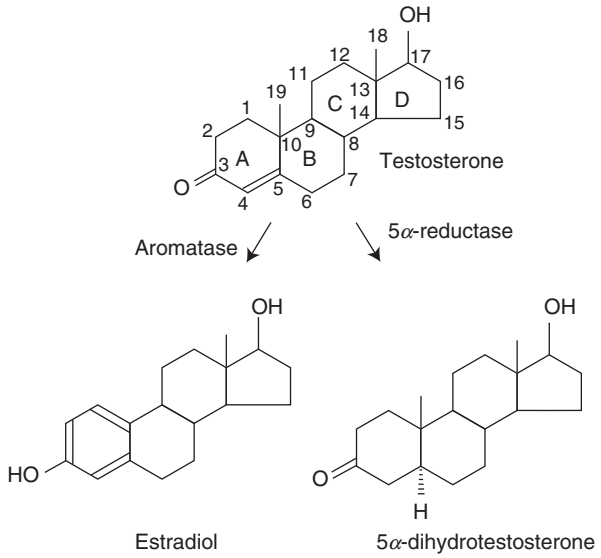

Figure 1 Testosterone metabolism to estradiol (aromatization) and dihydrotestosterone (5α reduction). Letters A–D denote the rings comprising the steroid nucleus, while the numbers denote identified carbon atoms among the rings of testorterone.

converts the A-ring of T to a phenolic one – an aromatic ring with a hydroxyl forming the major aromatized product, estradiol (E). In mammalian brains, E can accumulate from the circulation or can be formed *in situ* from circulating androgen. Aromatase is a member of the cytochrome P-450 family of enzymes that share a highly conserved heme-binding region and require nicotinamide adenine dinucleotide phosphate as a cofactor. The location of aromatase in brain has been examined by biochemical, molecular, and anatomical methods that indicate its selective expression in particular brain regions associated with gonadal steroid actions. The location of 5α-reductase in brain has received less attention, but biochemical methods have suggested a significantly more widespread distribution throughout the brain relative to aromatase.

Popularly, androgens and estrogens have been referred to as male and female hormones, respectively, since they were found originally to be the primary gonadal steroid hormones of the male (testis) and female (ovary) gonads. The experiments elucidating the interrelationships of androgens and estrogens in brain inform us that both of these hormones are functionally important sex steroids in both sexes. Accordingly, it can be misleading to refer to them as the respective male or female gonadal steroids, since they are critical to normal physiology in both sexes. These classifications are likely to change further as we learn more about the functions of sex steroids and their roles in different species.

Levels of androgen metabolism and quantities of intracellular receptor proteins to which they bind vary among species and change with age within species. Steroids and their receptors are found widely throughout the phylogenetic spectrum. Though utilized differently in the highly variable reproductive schemes among species, these hormones function intracellularly by similar routes to partially determine the development of reproductive tissues and the neural components regulating behavioral and reproductive physiology. Androgens and estrogens and their receptors are important in reproductive physiology during development and adulthood.

Steroids have been shown to act epigenetically at critical times in development to cause stable or imprinted differences in adult neural functions. These include effects on gene expression, behaviors, neuroendocrine regulation, and formations of several morphological traits that typically differ between males and females. Usually these developmental effects persist for the life of the animal, without requiring continuous exposure to hormone, and are called organizational effects. Commonly, these are believed to be permanent changes, though in some cases they can be altered later in life. Other responses occur repeatedly in direct response to steroid hormone activation whenever in the life span the appropriate signal or signals are given, and these are called activational effects. An example of this is the neuroendocrine signals causing release of some pituitary hormones. They also include acute effects on gene expression in sets of neurons. For some responses, both organizational and activational actions of steroids are required. Among species, the details of reproductive anatomy and physiology vary to a high degree. Presumably, the strong selective pressure on reproduction during evolution results in the high level of heterogeneity of this vital system among species or organisms, although the same steroidal hormones continue to be used as signals.

Estrogen Action

Organizational Effects

In most male mammals, during perinatal development, estrogens help determine differentiation of certain brain nuclei at the same time that they affect the differentiation of male-typical functional characteristics. It is presumed that this structural differentiation leads to mature males' potential to display masculine neuroendocrine and behavioral regulation. Given artificially to female fetuses, estrogen induces many of the same structural and functional changes, supporting the presumption of cause and effect between affected brain structures and elicited behaviors. If agents are used to block the action of aromatase, most of the actions of androgen in developing male

brain are prevented, supporting the role of estrogen as an essential determinant. The optimal ages for these effects to occur under the influence of physiological hormone levels in normal males, or in response to exogenous hormones in females, are generally restricted to certain perinatal ages, termed the 'critical period' for sexual differentiation of brain.

The timing of the onset of these estrogenic effects correlates with the final development of the gonads and the appearance of neural receptors for estradiol (ERs) in subcortical brain regions, including preoptic area, hypothalamus, and amygdala, and in certain cortical regions. The nervous system itself is the primary source of estrogen that interacts with ERs in developing male brains. For each vertebrate species studied, the specific activity of neural aromatization of androgen to estrogen (**Figure 1**) is several fold higher during perinatal life than in adulthood, with the activity of aromatase in preoptic area and amygdala being higher in males than in females at certain times during perinatal development. Concentrations of circulating T, the major substrate for brain aromatase, also are higher in male than in female vertebrates during critical periods of brain sexual differentiation. These result in increased biosynthesis of estrogen at specific sites and times in the developing male brain. The availability of androgen influences estrogen biosynthesis in two ways: (1) as the substrate for the enzyme and (2) via its action during adulthood to increase levels of aromatase.

During sexual differentiation, male mammals normally acquire the capacity for displaying masculine sexual behavior (the process of coital masculinization) and, depending on the species, lose the capacity to display feminine patterns of sexual behavior (the process of coital defeminization). In nonprimate species perinatal exposure of the male to T defeminizes both the neural mechanism controlling feminine sexual behavior and the brain mechanism controlling the cyclic release of luteinizing hormone-releasing hormone (also called gonadotropin-releasing hormone (GnRH)) from the median eminence into the hypophyseal portal vessels, thereby controlling the release of luteinizing hormone (LH) from the pituitary. In rats, both of these defeminizing actions of T in the male are caused by estrogen, which is formed in the developing brain from circulating androgen of testicular origin. Likewise, in species such as rat and hamster, the process of coital masculinization also depends to a high degree on the action of estrogen formed in neural tissues directly from circulating T, although there is evidence for ARs playing a role in this process. Some evidence has demonstrated that estrogens can exert developmental effects in the brains of higher mammals, including carnivores and primates. Experimentation supports the notion that both estrogen and androgens are required for masculinization, perhaps in a sequential manner, with estrogen acting before androgen. Nonetheless, significant data from human studies, including those with gene disruptions in either the estrogen receptor (alpha) or aromatase, indicate that investigators are a long way from transferring their understanding of the contributions of estrogen in rodents to these processes in humans.

Neural aromatization of androgen to estrogen also contributes to the formation of typically masculine patterns of neuroanatomical organization in both rats and ferrets. In rats, the medial nucleus of the preoptic area/anterior hypothalamus (POA/AH) is several times greater in volume in males than in females, and prenatal administration of either T or E to females augments the size of this nucleus toward that of males. By contrast, ER antagonists block normal differentiation of this nucleus in male rats. Similarly, in male ferrets a discrete nucleus is present in the dorsal portion of the POA/AH and is not apparent in females. Prenatal administration of a drug that inhibits brain aromatase activity prevents the development of this dorsal nucleus in male ferrets, whereas prenatal treatment with an androgen receptor antagonist does not.

Estrogen formed directly in neurons via aromatization of circulating androgen is a crucial source of estrogen to developing brains in male mice and rats. In rats and mice, but not in all species, the circulating liver-secreted protein α-fetoprotein has the property of binding estrogen with moderate affinity and high capacity. In these two rodents, α-fetoprotein concentrations are high enough to reduce the availability of estrogen from the general circulation. The recent creation of α-fetoprotein knockout mice provided for the demonstration that one role of α-fetoprotein is to prevent masculinization and defeminization in normal females. Whether α-fetoprotein also has other functions during brain development in these and other animals continues to be investigated. In other mammalian species, α-fetoprotein does not bind estrogen, so levels of circulating estrogen must be low enough in fetal female brains of those species at ages of high sensitivity so that they do not cause defeminization.

As in their male counterparts, the developing brains of female mammals produce considerable estrogen locally. Although estrogen receptors are prevalent in numerous subcortical and cortical brain regions during female development, lower numbers of estrogen receptors appear to be occupied by estrogen in developing females compared with males. Some investigators have hypothesized that some of this estrogen is available for potential interaction with neural ERs

and direct organization of developing female brains. This is in contrast with the prevailing dogma that all morphological and functional aspects of female brain development occur normally in the absence of any active steroidal signal. In rats, for example, perinatal exposure to very low levels of estrogenic stimulation can augment the ability of females to display sexually receptive behavior in adulthood; similar findings have been obtained in female ferrets. However, in the latter instance, severe perinatal deprivation of estrogenic stimulation to the developing female central nervous system (CNS) did not eliminate the ability of females to copulate after adult exposure to estrogen. In female rats the volume of the medial nucleus of the POA/AH, which normally is smaller in females than in males, is reduced by perinatal treatment with an estrogen receptor antagonist. Thus, differences between adult males and females in neural and functional characteristics may depend, in part, on the relative amount of estrogenic stimulation occurring in particular neural sites during development.

Activational Effects

Acting via intraneuronal ERs, estrogen also affects the function of the adult male and female CNS. In a wide variety of mammalian species, rising levels of estrogen, secreted by granulosa cells of developing ovarian follicles, promote the cyclic secretion of GnRH from the median eminence of the hypothalamus into the hypophyseal portal vessels, which conduct it to gonadotrophs in the anterior pituitary, which in turn release LH in quantities sufficient to cause ovulation. At other times during the female reproductive cycle, circulating estrogen acts to restrain the pulsatile secretion of GnRH from nerve terminals abutting portal capillaries of the median eminence. Other effects of estrogen in adult female rodents involve the priming of hypothalamic mechanisms to later respond to progesterone (P) stimulation, which in turn activates feminine sexual behaviors, including proceptive (solicitation of male partners) and receptive (lordosis) responses. Each of these effects of estrogen depends on the stimulation of protein synthesis at the neural sites being affected by the steroid.

In male mammals, estrogens also affect neuroendocrine function. For example, the expression of masculine sexual behavior in male rats, hamsters, and ferrets depends in part on the neural aromatization of circulating T. Likewise, estrogens formed locally in the hypothalamus and preoptic area of adult male mammals contribute to the inhibitory regulation of GnRH secretion, which in turn leads to the regulation of LH. Finally, exposure of either the perinatal or adult brain to supraphysiological concentrations of estrogen, either in the form of the natural estrogens such as E or in the form of synthetic estrogens such as diethylstilbestrol, can lead to the death of neurons in the arcuate region of the hypothalamus. Because estrogen has potent effects on males and females during development and in adulthood, there is considerable and growing concern for exposure to environmental estrogens. Substances with estrogenic actions are plentiful in many plants, and consequently in food sources, and others including insecticides and herbicides are released into the environment for commercial reasons. As risk assessment for these effects becomes more sensitive and selective, investigators are finding examples of neurobiological effects of many environmental estrogens. Some of these effects may be subtle, affecting normal events such as masculinization and defeminization, and others may be profoundly destructive to particular brain structures.

Parallel to their effects on gonadotropin secretion, estrogens and androgens modulate the brain areas controlling secretion of other anterior pituitary hormones such as growth hormone (GH), thyroid stimulating hormone, and adrenocorticotropic hormones. But unlike effects on gonadotrophins, some of these may not necessarily require sexual differentiation of the brain during development. For example, male rats display a low-frequency, high-amplitude GH secretory pattern with bursts of GH occurring every 3–4 h and separated by prolonged GH trough periods during which GH levels are very low or undetectable. Females show a high-frequency, low-amplitude pattern of GH secretion, with pulses of GH occurring every 1–2 h. Individual peak amplitudes are lower, and GH trough values higher and of shorter duration than those of male rats. Depending on the endocrine state of the animal, these parameters can be altered and sex-reversed in adults by administration of estrogens and androgens.

Androgen Action

Genetics

For the brain, less is known about effects of androgens that are directly mediated by ARs than about effects involving androgen conversion to estrogen and subsequent mediation by ERs. However, since androgen actions are required for normal male development and reproduction, but not for individual viability, animals with occurrences of genetic defects in androgen action survive and can be studied. Naturally occurring mutations can affect either the synthesis of androgens or the receptor proteins that mediate their actions. It had been assumed that the paucity of

similar estrogen receptor mutations was attributable to the likelihood of their lethality to the fetus. Recently, however, examples have been detected in humans for receptors or for aromatase. Also, through molecular genetic manipulations to produce 'knock-out' animals lacking receptors, viable mice have been obtained without the normal complement of ER alpha and/or beta. Since these receptor proteins have many functions, the genes are pleiotropic, and compensatory changes may occur during development. Therefore, while they are very suggestive of testable mechanisms, it is still challenging to discern causal relationships between these mutations and altered functions.

AR mutations affect all tissues in the body, including neural, and cause the syndrome of testicular feminization (TFM). In humans, rats, and mice, TFM is X-linked. Although X/Y animals with TFM have testes, which produce androgens, these animals do not respond normally to androgen and therefore develop with deficiencies of masculinization. Highly affected animals, including many humans with TFM, are externally female in appearance; other particular androgen-resistant animals, including humans with incomplete TFM, are virilized to variable extents. These mutants aid in determining the roles ARs play during development and behavior. Male sexual behavior can be elicited with estrogen or androgens in adult TFM rats, which are partially responsive to androgens, as it can be in normal males. Presumably the normal ERs in these animals and the partial complement of ARs are sufficient for masculinization during differentiation and activation of behavior in adults. By contrast, the TFM^H allele in mice results in profound insensitivity to those actions of androgen that require ARs rather than conversion to estrogen and binding to ERs. When given estrogen or androgen, a few of these adults showed partial male copulatory behavior, but most did not respond. These results are consistent with the concept that ARs, along with ERs, are involved both in sexual differentiation of mouse brain and in eliciting behavior.

Mutations in the gene for the enzyme 5α-reductase (Figure 1) affect androgen action. Humans with deficits in this enzyme are externally undervirilized to varying extents, depending on the specific mutation and protein defect. Certain of these patients appear externally as females at birth, and though raised as girls, they undergo virtually complete virilization over a period of years at pubertal ages. These individuals can then assume typical male gender roles, including marriage. It has been assumed that this transformation occurs because high levels of T at puberty are able to mimic the normal action of dihydrotestosterone at earlier ages. The question that is raised regarding behavior is: Does androgen act prenatally to masculinize the brains of human males as it does for rodents? Or is sexual behavior in humans less dependent on hormonal than on experiential and other influences? The sequence in the patients mentioned earlier is consistent with either hypothesis. On one hand, T is active in the development of brains of other animals without its reduction to DHT, so it may function normally in the brains of these individuals. When their genitalia become virilized, they would already have masculine brains. On the other hand, gender identity in humans is dependent on the perception of gender, so a person eventually appearing male might acquire the potential to assume male gender behaviors without specific hormonal effects on the brain. This dichotomy illustrates how mutations can result in ambiguous hypotheses about mechanisms of development and behavior.

Additional steps in metabolism of sex steroids in the brain might be important and are being investigated. One of these provides a third hypothesis regarding sexual differentiation of the human brain: that it is capable of occurring years after birth under the influence of androgen. The enzyme 5α-reductase catalyzes the reduction of several $delta^4$ steroids, including P and T. Unlike most other steroid receptors that are highly selective for a particular ligand, ARs bind both T and DHT, but with higher affinity for DHT, and they have moderately high affinities for E and P. It has been suggested that during gestation, P concentrations are high enough that in some tissues they might potentially inhibit androgen binding to ARs. However, the 5α-reduced product, 5α-pregnane-3,20-dione (5α-dihydroprogesterone, DHP), is less inhibitory than P. Accordingly, 5α-reductase could normally reduce T and P to DHT and DHP, respectively, thus allowing greater binding of androgen to AR. In some patients with 5α-reductase deficiency, P would achieve its potential for blocking AR, thus inhibiting its normal action. The virilization that occurs in these 5α-reductase-deficient patients could occur when androgen levels rise during puberty but without high levels of P that occur during gestation. The brain may be one organ in which this role of 5α-reductase is important. This explanation is plausible if the human brain is susceptible to change by hormones at near-adult ages.

Analogs

Analogs of androgens and estrogens which are agonists or antagonists for specific receptors or specific steps in metabolism help evaluate the involvement

and mechanisms of action of these hormones in the brain. If an effect of T involves aromatization to E and binding to ERs, then inhibitors of aromatase activity can prevent the response, whereas synthetic estrogens may mimic T. DHT is unlikely to be active for responses where strong activation of ER-alpha is needed, since reduced androgens are not aromatized to estrogens. Increasing understanding of ER-beta (and possibly other ERs) will likely engender an increasing interest in how DHT metabolites such as androstenediols may have estrogenic actions. If the effect of T is mediated by ARs, then DHT may be as agonistic as T. Nonetheless, situations occur in which T is active, aromatization and ERs are not indicated as involved, and yet DHT is not effective. This can occur if administered DHT is very quickly metabolized to inactive products before reaching target cells. To avoid this complication, synthetic androgens like methylenetrienolone (R1881) can be used. In adult castrated male rats and hamsters, administration of this synthetic steroid duplicates the rather minimal stimulatory action on mating that occurs in response to a naturally occurring androgen, DHT. When given concurrently with estrogen, either R1881 or DHT causes essentially a complete restoration of sexual activity in castrated males, equivalent qualitatively and quantitatively to the behavior seen after T treatment. This suggests that some actions of T involve both ERs and ARs. It has also been hypothesized that some responses to androgen occur efficiently only if T, but not DHT, binds to ARs.

A growing number of substances are being characterized as endocrine disruptors on the basis of their influences on estrogenic or androgenic targets. Some of these compounds also might have protective actions against some hormone-dependent cancers, increase emotional well-being, and inhibit other pathologies. Interest in the use of some of these compounds, known as phytoestrogens, as dietary supplements is increasing, and the basis for this is their estrogen-like action. There are three main classifications of phytoestrogens: (1) isoflavones, derived principally from soybeans; (2) lignans, found in flaxseed; and (3) coumestans, derived from alfalfa. Of these, consumption of isoflavones has the largest impact on human health because of its availability in a variety of food products containing soy. Of the isoflavones, genistein and daidzein are considered the most potent estrogens and thus have historically attracted the most attention. However, phytoestrogens (e.g., equol, a metabolite of the isoflavone daidzein) can also have other mechanisms of action, including antiandrogenic effects.

Other steps in sex steroid metabolism that are possibly involved in brain function have been investigated. Much interest has been generated in a possible chemical relationship between steroids and neurotransmitters. E can be hydroxylated to catechol estradiol (2-hydroxyestradiol). While considerable data demonstrate that certain brain tissues can catalyze the formation of this and related products, and that they can be potent when administered, many investigators now are uncertain that these have a significant physiological role. Another product, 5β-dihydrotestosterone (the 5β-reduced product of T reduction, rather than the more common 5α isomer), has been observed in avian species. Whether it is part of an alternative degradative, and perhaps regulatory, catabolic step or is another active steroid, as has been suggested for steps in blood production, is unknown. As methods for steroid and receptor biochemistry become increasingly refined and are applied toward the level of single cells, further steroid products may be discovered that play specific roles in brains of different species. Much has been learned in recent decades about the dynamics of androgens, estrogens, and their receptors in the brain and the effects these have on development and mature CNS. Current and future studies are revealing the changes in cellular and molecular biology that underlie these influences.

See also: Adrenal Steroids: Biphasic Effects on Neurons; Estrus and Menstrual Cycles: Neuroendocrine Control; Gene Expression Regulation: Steroid Hormone Effects; Gonadotropin-Releasing Hormone: GnRH-1 System; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Hormones and Behavior; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Neurosteroids.

Further Reading

Bakker J, DeMees C, Douhard Q, et al. (2006) Alpha-fetoprotein protects the developing female mouse brain from masculinization and defeminization by estrogens. *Nature Neuroscience* 9: 220–226.

Balthazart J and Ball GF (2006) Is brain estradiol a hormone or a neurotransmitter? *Trends in Neurosciences* 29: 241–249.

Carani C, Qin K, Simoni M, et al. (1997) Effect of testosterone and estradiol in a man with aromatase deficiency. *New England Journal of Medicine* 337: 91–95.

Couse JF and Korach KS (1999) Estrogen receptor null mice: What have we learned and where will they lead us? *Endocrine Reviews* 20: 358–417.

Gerall AA, Moltz H, and Ward IL (eds.) (1992) *Handbook of Behavioral Neurobiology, Vol. 11: Sexual Differentiation.* New York: Plenum Press.

Gottlieb B, Pinsky L, Beitel LK, and Trifiro M (1999) Androgen insensitivity. *American Journal of Medical Genetics* 89: 210–217.

Keay J, Bridgham JT, and Thornton JW (2006) The Octopus vulgaris estrogen receptor is a constitutive transcriptional activator: Evolutionary and functional implications. *Endocrinology* 147: 3861–3869.

Lund TD, Munson DJ, Haldy ME, Setchell KD, Lephart ED, and Handa RJ (2004) Equol is a novel anti-androgen that inhibits prostate growth and hormone feedback. *Biology of Reproduction* 70: 1188–1195.

McLachlan JA (2001) Environmental signaling: What embryos and evolution teach us about endocrine disrupting chemicals. *Endocrine Reviews* 22: 319–341.

Nuclear Receptors Nomenclature Committee (1999) A unified nomenclature system for the nuclear receptor superfamily. *Cell* 97: 161–163.

Qiu J, Bosch MA, Tobias SC, et al. (2006) A G-protein-coupled estrogen receptor is involved in hypothalamic control of energy homeostasis. *Journal of Neuroscience* 26: 5649–5655.

Ronnekleiv OK and Kelly MJ (2005) Diversity of ovarian steroid signaling in the hypothalamus. *Frontiers in Neuroendocrinology* 26: 65–84.

Schlinger BA and London SE (2006) Neurosteroids and the songbird model system. *Journal of Experimental Zoology. Part A, Comparative Experimental Biology* 305: 743–748.

Smith EP, Boyd J, Frank GR, et al. (1994) Estrogen resistance caused by a mutation in the estrogen-receptor gene in a man. *New England Journal of Medicine* 331: 1056–1061.

Tobet SA and Hanna IK (1997) Ontogeny of sex differences in the mammalian hypothalamus and preoptic area. *Cellular and Molecular Neurobiology* 17: 565–601.

Toran-Allerand CD (2005) Estrogen and the brain: beyond ER-alpha, ER-beta, and 17beta-estradiol. *Annals of the New York Academy of Sciences* 1052: 136–144.

Usui T (2006) Pharmaceutical prospects of phytoestrogens. *Endocrine Journal* 53: 7–20.

Wilson JD (1999) The role of androgens in male gender role behavior. *Endocrine Reviews* 20: 726–737.

Relevant Websites

http://www.endo-society.org – Endocrine Society.

http://www.nap.edu – National Academies Press site to buy *Exploring the Biological Contributions to Human Health: Does Sex Matter?*

http://www.ens-lyon.fr – Nuclear Receptor Database (NUREBASE) at the Ecole Normale Superieure de Lyon.

http://nrr.georgetown.edu – Nuclear Receptor Resource.

http://www.nursa.org – Nuclear Receptor Signaling Atlas (NURSA).

http://www.sbne.org – Society for Behavioral Endocrinology.

http://www.womenshealthresearch.org – Society for Women's Health Research.

http://www.epa.gov – US Environmental Protection Agency, Endocrine Disruptor Research Initiative.

Gonadotropin, Neural and Hormonal Control

I J Clarke and A J Tilbrook, Monash University, Clayton, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Reproduction is vital to the survival of species, and highly tuned systems have evolved to ensure success. The process is totally dependent on the function of gonadotropin-releasing hormone (GnRH) neurons, which provide the necessary drive to the gonadotropes of the pituitary gland, through neurosecretion. In turn, the secretion of gonadotropins maintains the function of the gonads. Although various factors may influence the function of the gonadotropes and the gonads, normal function is dictated by the function of GnRH neurons. Thus, pulsatile secretion of GnRH into the hypophysial portal system maintains synthesis and secretion of gonadotropins. Without GnRH function, the gonadotropes cease to function, and reproduction is not possible.

The phasic pattern of GnRH secretion is a fundamental property that allows high-fidelity control of downstream processes. Whereas the notion of a 'pulse generator' in the brain to control the phasic pattern of secretion has been proposed, such a center has not been identified, nor has any type of cell in the brain been ascribed such a function. On the other hand, it has been proposed that GnRH neurons possess an inherent ability to operate in a phasic manner; this allows modulation of secretion by various factors, including feedback of gonadal hormones. If one allows that GnRH neurons have inherent oscillatory function, then one can invoke control mechanisms that involve changes in phase (pulse frequency) and strength (pulse amplitude). There is a direct inverse relationship between the frequency and amplitude of GnRH and luteinizing hormone (LH) pulses such that a slow rate of pulses tends to increase the amplitude of the pulses and vice versa. Whereas this has been modeled for the release of LH, showing that the relationship reflects a releasable pool of hormone in the gonadotropes, there is no good understanding of whether such a relationship prevails in the neurosecretory terminals of GnRH neurons. Such changes are seen in the patterns of secretion that prevail across the female estrous/menstrual cycle and with changing function of reproduction in males, especially in seasonal breeders.

The GnRH neurons integrate information from other neuronal systems, which define the pattern of GnRH secretion from the brain. This hierarchical system allows various points of control. Figure 1 emphasizes the point that the function of GnRH neurons is affected by gonadal hormone feedback, stress, nutritional status, and mood in humans, as well as season in subhuman species. GnRH neurons receive information from a range of neuronal systems. These arise from the vegetative regions, such as the lateroventral medulla, the serotoninergic system, and various regions of the limbic system. These brain regions probably provide 'drive' to the GnRH neurons and process information from environmental cues (season and stressors) as well as intrinsic feedback. In particular, the GnRH system is regulated by the feedback effects of sex steroids from the gonads. The final common output of the brain that controls the reproductive system is GnRH secretion, which represents the integration of all this information from various brain regions.

GnRH Control of the Gonadotropes via the Hypophysial Portal

The neuroendocrine, hypothalamopituitary axis has special anatomical and functional features that optimize function and fine control. GnRH neuronal cell bodies are found in the preoptic area of most species but are predominantly located in the basal hypothalamus of primates. The cells are not as closely grouped as other neuroendocrine systems, such as the corticotrophin-releasing hormone neurons of the paraventricular nucleus. Nevertheless, the GnRH cells operate in a coordinated manner, possibly due to cell–cell communication. This synchronized pattern of activity of GnRH neurons may be due to coordinate function of afferents that arise from elsewhere in the brain, providing 'drive' to the system.

The GnRH cells project to the median eminence to provide a dense plexus of neurosecretory terminals in the external zone of this structure. Secretion into the hypophysial portal system reflects the phasic activity of the neurons and represents conversion from neural transmission to an endocrine signal. Since the hypophysial portal system is a relatively short one (<1 cm) and relatively low in volume, the secretion of femtomolar quantities of releasing factors (such as GnRH) from neuronal terminals can present physiologically relevant concentrations (picograms per milliliter) of the hormone to the relevant cells (gonadotropes) of the pituitary gland. The portal provides optimal economy of delivery to the pituitary gland since different releasing factors are carried in the same milieu, but selectivity of function is achieved because each relevant cell type of the pituitary gland possesses receptors specific

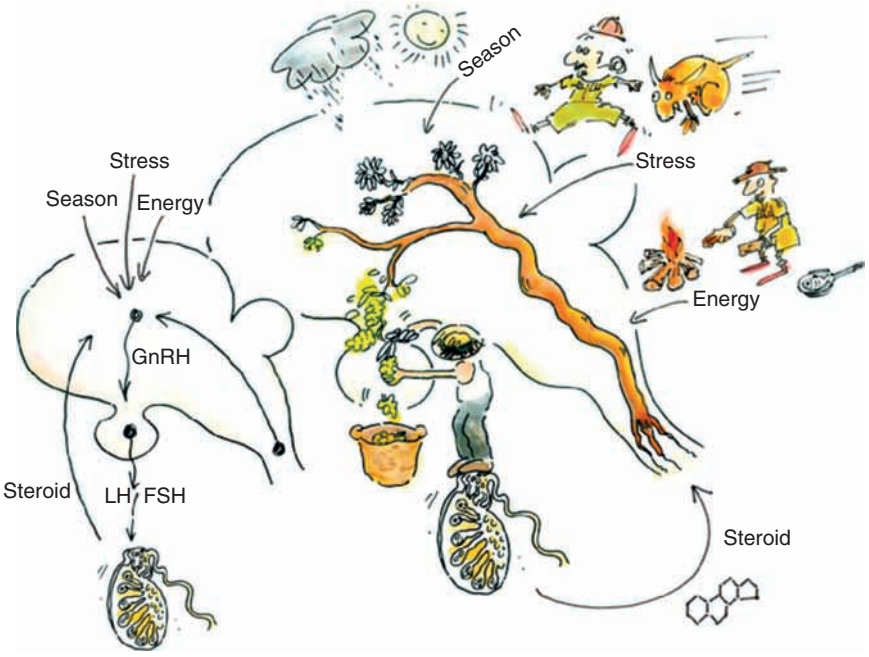

Figure 1 Schematic representation of the integration of information from various neural systems to regulate the function of gonadotropin-releasing hormone (GnRH) neurons. The GnRH cells are influenced by neurons that receive steroidal feedback signals from the gonads, centers that process information on energy balance, season, and stress. Thus, the secretion of GnRH and the gonadotropins provides drive to the reproductive system, in relation to environmental and internal homeostatic systems. FSH, follicle-stimulating hormone; LH, luteinizing hormone.

for a particular function. Thus, gonadotropes possess GnRH receptors, but corticotropes, lactotropes, and so forth, do not. Another special feature of this neuroendocrine architecture is that responses of the relevant pituitary cells may amplify the neuroendocrine signal manifold. Thus, the GnRH signal, which is at a concentration of picograms per milliliter in a few milliliters of hypophysial blood, acts on gonadotropes, which amplify the signal and secrete sufficient gonadotropin to achieve peripheral circulating concentrations of the order of nanograms per milliliter.

Relationship between GnRH, LH, and FSH Secretion

Of all the neuroendocrine systems studied, the relationship between GnRH and LH provides the best example of a one-to-one relationship. It was originally proposed by GW Harris that endocrine function of the anterior pituitary cells was a direct reflection of neurosecretion of relevant factors from the hypothalamus, and this hypothesis has been tested for all systems (reproductive axis, stress axis, thyroid axis, etc.). In particular, the pulsatile secretion of LH is a direct function of GnRH secretion, although some modulation of the output (LH) may occur by modulation of gonadotrope function. On the other hand, the secretion of follicle-stimulating hormone (FSH) from the same gonadotropes is not such a simple reflection of GnRH secretion. This is because the secretion of LH in a pulsatile manner is entirely dependent on GnRH action, whereas FSH secretion can occur without the secretory stimulus of GnRH. The difference in the means by which the secretion of the two gonadotropins is supported may be explained by the fact that secretion *per se* of LH requires a stimulus, whereas secretion of FSH may be passive. Regarding the latter, GnRH supports FSH synthesis, and by virtue of this, secretion occurs; secretion of this gonadotropin does not have the absolute requirement for acute stimulus. In this context, it may be considered that the pulsatile secretion of LH is the primary driver of endocrine function of the gonads, which is reflected in pulsatile secretion of estrogen from the ovary and testosterone from the testis. The function of FSH primarily supports gamete development, although this gonadotropin also stimulates inhibin production.

Feedback Control of GnRH Secretion

The arrangement of the GnRH system provides various control points, as outlined in Figure 2 for the feedback effects of estrogen. Similar 'wiring diagrams' may be constructed for other feedback regulators, and for purposes of simplification, the figure does not integrate all neuronal systems involved in the feedback regulation of estrogen.

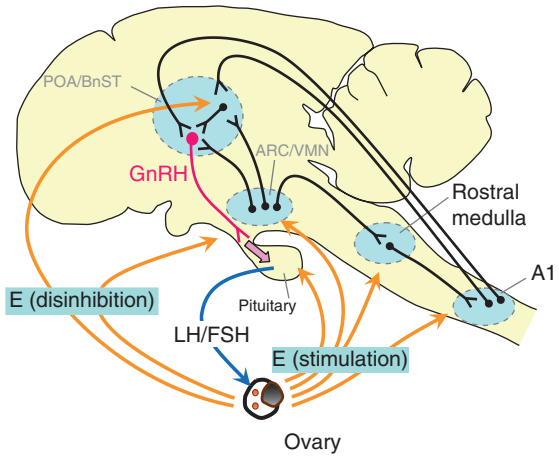

Figure 2 Estrogen (E) is able to negatively and positively regulate the secretion of gonadotropin-releasing hormone (GnRH) and the gonadotropins. These feedback effects are mediated via neurons in the brain that possess estrogen receptor-α, which may provide either direct input to GnRH neurons or interneuronal input via relay neurons. Positive feedback may also be assisted by disinhibition of neurons that are inhibitory to GnRH function. Various centers of the brain contain concentrations of neurons that are estrogen-responsive and have been shown to project to the preoptic area (POA) or the GnRH neurons. Main subgroups that provide input to the POA and GnRH neurons are (1) brain stem (especially the A1 cell group of the ventrolateral medulla; (2) the arcuate nucleus (ARC) and the ventromedial nucleus (VMN); and (3) the bed nucleus of the stria terminalis (BnST) and subgroups of the POA, including an area (the anteroventral periventricular nucleus) that displays sexual dimorphism and is known to be important in the generation of the positive feedback effect of estrogen that causes the preovulatory surge in rodent species. Stimulation and/or disinhibition may also regulate GnRH secretion by action at the level of the GnRH terminals in the neurosecretory zone of the median eminence. Estrogen may also act on hypothalamic cells that project to the brain stem, thus modulating the function of noradrenergic neurons that then project to the POA. Estrogen also acts on the pituitary gonadotropes so that there is coordinated action. For example, to generate the preovulatory surge in GnRH/luteinizing hormone (LH), estrogen acts to increase the secretion as well as the action of GnRH in a coordinated manner. This positive feedback action may involve similar time-delayed signaling processes in neurons and in gonadotropes. FSH, follicle-stimulating hormone.

These points of control are at various levels of the brain and allow the integration of information from environmental factors as well as intrinsic systems. There is significant experimental evidence to support the notion of a serial and converging 'wiring pattern' to GnRH neurons (Figure 3). Synaptic input to GnRH neurons has been described for a wide range of systems, but the extent to which each of these may be important and how they interact require further study. Whereas such synaptic input of a given element to GnRH cells may be demonstrable, the magnitude of influence of any one system is more difficult to ascertain. Nevertheless, evidence is emerging for a highly organized network of neuronal control that may involve direct input from 'major' control systems and 'modulatory' input from other systems. One potential model is an 'on–off' system, which is outlined in Figure 4; this is a model of multiple integration at a point upstream of the GnRH neuron (an on–off switch) and presumes that glutamate is the main stimulatory neurotransmitter and γ-aminobutyric acid is the main inhibitory neurotransmitter in the neuroendocrine system. Certainly this notion of primary stimulator/inhibitor has been proposed by various investigators, and GnRH cells receive inputs from both glutamate and γ-aminobutyric acid systems, but there is also strong evidence for other systems that exert major control. For example, the noradrenergic system of the brain stem provides input to GnRH cells, and this may provide a mechanism of primary drive. The proposed model is attractive because it allows the serial and converging input to GnRH neurons to be integrated into simple stimulatory or inhibitory signals, prior to final 'instruction' to the GnRH neurons. The integrated information from all afferent systems is reflected in output (GnRH secretion).

Neuronal afferents may regulate the function of the GnRH system by pre- or postsynaptic communication of afferent input, and there is also significant support for the notion that GnRH may be influenced by action of neurotransmitters at the level of the cell bodies or at the level of the terminals in the median eminence (Figure 5).

Steroid Feedback to the GnRH Cells

The function of GnRH cells is strongly influenced by the feedback of gonadal steroids on the brain. This dictates the cyclic pattern of reproductive function in the female and also maintains a closed-loop system in the male. Such a system allows for constant maintenance of finely tuned feed-forward (GnRH secretion) by feedback regulation. Although the gonads also secrete protein hormones, such as inhibin, the feedback action of these factors appears to be at the level of the pituitary gonadotropes, influencing GnRH action rather than secretion. Although powerfully influenced by the feedback effects of gonadal steroids, the GnRH cells do not express the relevant steroid receptors at a level that is of major functional relevance. Thus, GnRH cells do not express estrogen receptor-α (ER-α), androgen (AR), or progesterone receptors, but do express estrogen receptor-β (ER-β). Since the major feedback effects of estrogen are transmitted via ER-α, the expression of ER-β in GnRH cells is probably of minor consequence. Various cell types express the relevant steroid receptors and transmit information to GnRH cells; a good example of such a system is the recently discovered kisspeptin cells of the hypothalamus and preoptic area, which express ER-α and powerfully regulate

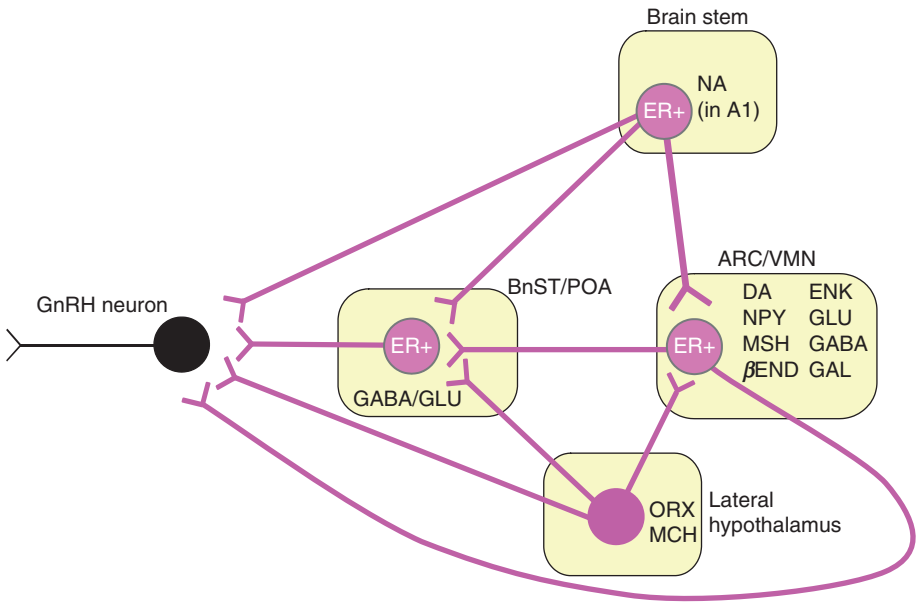

Figure 3 The gonadotropin-releasing hormone (GnRH) neuron is influenced by serial and converging series of inputs from various brain regions that support and modulate function. The figure indicates major centers in the brain from which cells project to the GnRH cell bodies. These may be involved in steroid feedback, and the figure indicates centers that contain estrogen receptor-α (ER+) and respond to estrogen, as well as centers where afferents do not express ER. The latter probably convey information other than steroid feedback to GnRH cells. For example the orexin (ORX) and melanin concentrating hormone (MCH) cells of the lateral hypothalamus may transmit information on metabolic state to the GnRH cells. These centers of the brain may also intercommunicate, as indicated. The 'wiring' may be directly to the GnRH neurons or may involve interneurons in the bed nucleus of the stria terminalis (BnST) or the preoptic area (POA). Serial input via chains of ER+ neurons may allow amplification of estrogen signaling. NA, noradrenaline; A1, noradrenergic cells of the caudal ventrolateral medulla; ARC, arcuate nucleus of the hypothalamus; VMN, ventromedial nucleus; DA, dopamine; ENK, enkephalin; NPY, neuropeptide Y; GLU, glutamate; MSH, melanocyte stimulating hormone; GABA, γ-aminobutyric acid; βEND, β-endorphin; GAL, galanin. For simplification, this figure represents a few pathways in order to indicate a general organization, but other pathways have also been described.

GnRH cells. Nevertheless, electrophysiological, single-cell studies have shown some direct effects of steroids in GnRH cells, so it remains possible that some level of direct control occurs. In spite of this, overwhelming evidence shows that the feedback effects of gonadal steroids are transmitted by neurons that express the relevant receptors at high levels and transmit information to GnRH cells. This pattern can be demonstrated by implantation of steroids in discrete regions of the brain and subsequent measurement of GnRH or gonadotropin secretion. In this way, various brain regions have been shown to participate in the steroid feedback regulation of GnRH cells.

Steroidal feedback regulation of GnRH cells may be negative or positive, although the latter is only seen in the female. This is due to sexually dimorphic patterns of brain development. The sexually dimorphic anteroventral periventricular nucleus of the female rodent brain is the center which appears responsible for generation of positive feedback signals of estrogen that produce the preovulatory surge in GnRH and LH. In other species, the relevant 'surge center' is found elsewhere; in sheep it resides within the mediobasal hypothalamus.

The means by which steroids such as estrogen may exert both negative and positive feedback actions on GnRH secretion is not well deciphered. Further work is required to examine the subcellular functions of estrogen within particular cell types. Another feature of the feedback systems is that different temporal patterns of response (in terms of GnRH secretion) may be observed. Thus, in a gonadectomized animal that is deprived of steroids, challenge by estrogen first causes a short-term negative feedback response, followed by a positive feedback response, followed by a prolonged negative feedback response. Active research is in progress to understand this time-related response.

Feedback Control of the Female Reproductive Cycle

Females undergo cyclic patterns of ovarian function that reflect cyclic patterns of GnRH and gonadotropin secretion. Ovulation is caused by a positive feedback action of estrogen, which causes a surge in secretion of GnRH. This involves the time-delayed response of neural systems mentioned above. Once ovulation has occurred, a corpus luteum forms on the

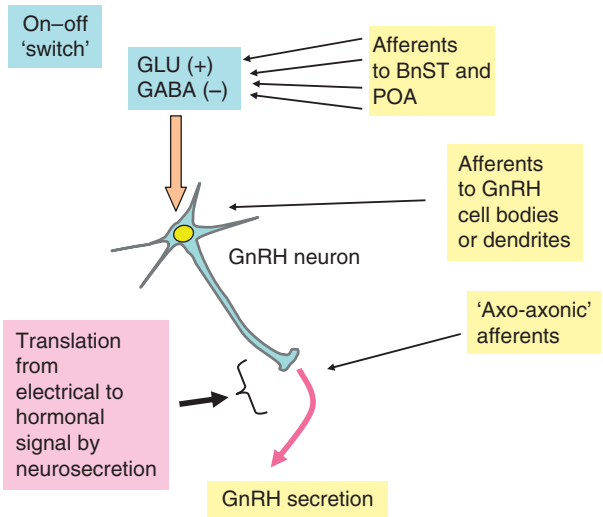

Figure 4 A proposed model for control of gonadotropin-releasing hormone (GnRH) neurons that involves simple on–off regulation via the glutamate (GLU) and γ-aminobutyric acid (GABA) systems, respectively, of the bed nucleus of the stria terminalis (BnST) or the preoptic area (POA). This model also incorporates input to the GnRH terminals that may modulate GnRH secretion. The evidence for direct 'axo-axonic' input to GnRH terminals is not strong, but various systems provide input to the neurosecretory zone of the median eminence and modulate secretion by 'volume control,' which does not require synaptic contact. The simplified 'on–off' model also incorporates direct input to GnRH cell bodies and dendrites, which is well described.

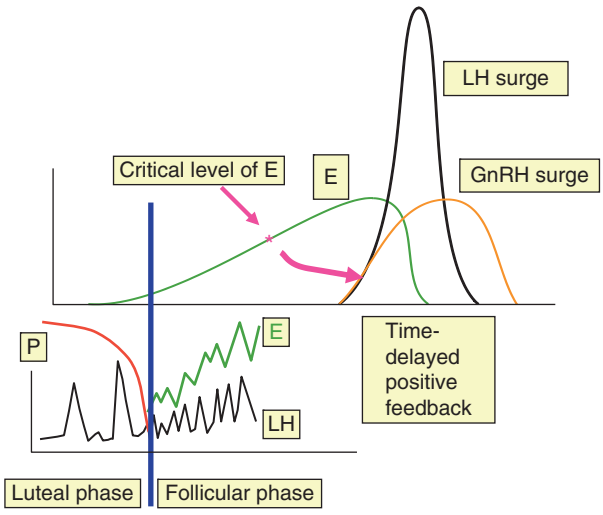

Figure 5 The female estrous/menstrual cycle is characterized by changing patterns of gonadotropin-releasing hormone (GnRH) and luteinizing hormone (LH) pulse frequency and amplitude, which reflect the feedback actions of progesterone (P) and estrogen (E). During the luteal phase, combined action of both steroids causes low-frequency, high-amplitude GnRH and LH pulses. With the decline in P at the end of the luteal phase, the frequency of GnRH and LH pulses increases, and the response of the ovarian follicles in terms of E secretion increases. At a critical level of E, a positive feedback effect on GnRH and LH secretion is initiated, culminating in the preovulatory surge. Note that the surge in GnRH secretion is more prolonged than the surge in LH secretion.

ovary, and plasma levels of progesterone rise. Progesterone acts on the brain to slow the rate of pulsatile release of GnRH, with an associated increase in pulse amplitude. Since estrogen (and androgens) is released by the ovary during the luteal phase of the cycle, the negative feedback control of GnRH neurons probably represents a combination of signaling from cells in the brain that are receptive to both steroids. In fact, work in various species has shown that all cells that express progesterone receptors also express estrogen receptors, so combined action of both steroids may be transmitted by the same cells. As the cycle progresses and if pregnancy does not ensue, the corpus luteum regresses, and progesterone levels fall. Associated with the demise of the corpus luteum, progesterone feedback is lost, and the restraint on GnRH cells is lifted. This leads to increased frequency of pulsatile secretion of GnRH and LH, with associated reduction in amplitude. The frequency of GnRH pulses increases throughout the follicular phase of the cycle, and a GnRH surge ensues that causes another reproductive cycle.

At some point in the follicular phase of the female cycle, the positive feedback mechanism is activated, probably in response to a critical circulating level of estrogen. Recent studies indicate that this phenomenon requires a minimum duration of estrogen action, but the cellular mechanisms involved are not defined. This is not an easy task, since it appears that the positive feedback stimulus occurs at a time when some degree of negative feedback prevails, that is, during the midfollicular phase of the ovarian cycle. It is possible that the effect of estrogen in this particular phase of the cycle is to set in train a sequence of events that is manifest as negative response followed by positive response, involving the same subcellular mechanisms. It is important to note that the negative effect of estrogen in the follicular phase of the cycle involves reduction in pulse amplitude but not frequency. Indeed, there is evidence that GnRH/LH pulse frequency increases during the follicular phase of the cycle, and this may be the primary determinant of rising estrogen levels.

Feedback Control of the Male Reproductive Process

Neural control of male reproductive function is subject to the feedback actions of testicular steroids. As in the female, the gonadal protein hormone inhibin acts on pituitary gonadotropes and not the brain. In particular, testosterone acts to negatively regulate the secretion of GnRH. The sites of negative feedback action of testicular steroids have been most extensively studied in

sheep and primates, indicating that the major effects are at the level of the brain, with minimal effects on pituitary gonadotropes. Control of GnRH function in males differs from control in females in that it is a closed system, with no cyclic changes and no positive feedback system. The mechanisms of negative feedback actions of sex steroids have not been as well researched in males, but as in females, the GnRH neurons do not possess steroid receptors.

In most species the predominant steroid produced by the testes is testosterone. In many tissues, including the brain, testosterone can be converted via 5α-reduction to the more potent androgen dihydrotestosterone, or via aromatization to estrogen (Figure 6). Thus, the negative-feedback actions of testosterone to regulate GnRH cells could be due to direct actions or via its primary metabolites. Both AR and ER exist in the male brain regions that control GnRH neurons. Progesterone receptors exist in the same regions, but the significance of this steroid in the control of GnRH neurons in males is not clear. The relative roles of testosterone and its metabolites in the regulation of GnRH neurons are not well understood, but it seems most likely that conversion of testosterone to estradiol accounts for at least some of the negative feedback effects on the secretion of GnRH. In male sheep, a key region of action of estradiol is the arcuate nucleus–ventromedial region. In male rats, androgens stimulate aromatase activity in the preoptic area and hypothalamus, which would stimulate conversion of testosterone to estradiol-17β. The importance of conversion of testosterone to dihydrotestosterone with respect to negative feedback regulation of GnRH is essentially unknown. If this conversion is important, the action would be mediated by AR. The locations of AR-containing cells involved in this action have not been identified, but evidence suggests that they are not in the arcuate nucleus–ventromedial region.

The neuronal systems relevant to the negative feedback actions of testicular steroids are largely unidentified. In male sheep, GnRH neurons receive input from various parts of the hypothalamus, with the major contribution from the medial or periventricular region, which is a region containing AR and ERα.

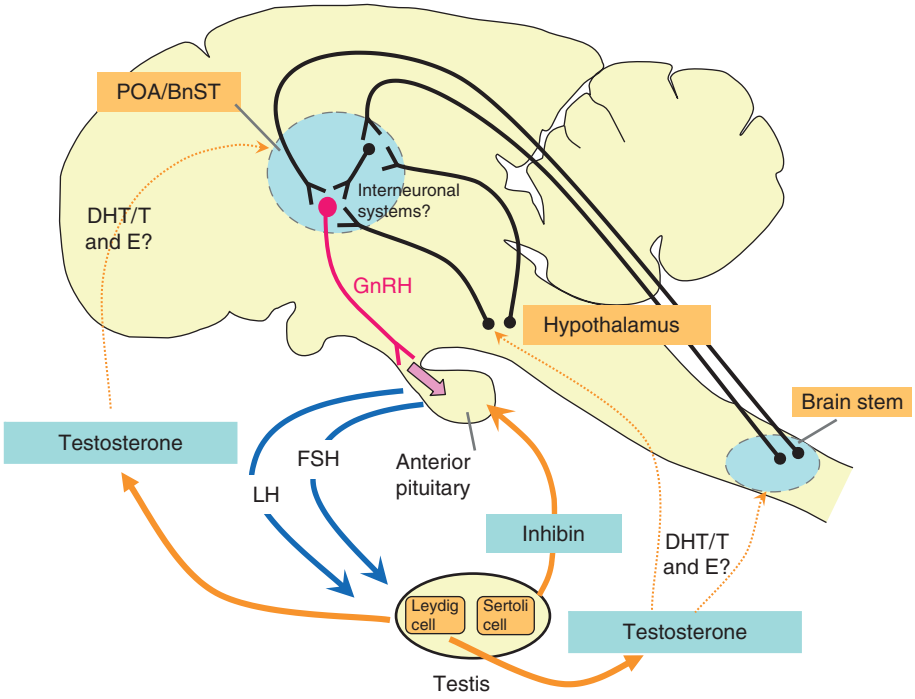

Figure 6 Schematic representation of the hypothalamopituitary-testicular axis. Luteinizing hormone (LH) acts on Leydig cells in the testes to stimulate the synthesis of testosterone (T), and follicle-stimulating hormone (FSH) acts on Sertoli cells to stimulate the secretion of inhibin. Testosterone has negative feedback actions in the central nervous system to suppress gonadotropin-releasing hormone (GnRH) secretion, but the specific sites of action have not been determined. Direct pituitary actions of testosterone are negligible. Testosterone may exert negative-feedback actions on GnRH neurons after conversion to either or both of its primary metabolites, dihydrotestosterone (DHT) or estradiol-17β (E). The neural pathways that mediate the negative-feedback effects of steroids in males have not been determined with certainty but may include γ-aminobutyric acid, opioids, and cathecholamines. Inhibin has negative feedback actions directly at the anterior pituitary (not the brain) to regulate FSH secretion. BnST, bed nucleus of the stria terminalis; POA, preoptic area.

Various hypothalamic nuclei also appear to transmit the feedback effects of steroids to the GnRH neurons. Nevertheless, the identity of most of these cells and the extent to which particular inputs relay the negative feedback effects of steroids are unknown. The brain stem also provides input to GnRH neurons in the male brain, mostly arising from noradrenergic fields. In essence, the effects of sex steroids on the male brain are not well understood.

Summary

GnRH neurons control reproduction, but synthesis of the peptide and secretion is regulated by feedback effects of gonadal steroids acting in specific regions of the brain. Since GnRH cells do not possess the relevant receptors, the feedback signals are relayed to GnRH cells via cells that are steroid-responsive. These steroid-receptive cells provide serial and converging pathways to GnRH neurons. In females, both negative and positive feedback systems operate, but the male system is controlled only by negative feedback. Reproduction is also modulated by stress, energy balance, and season, via neuronal systems in the brain that transmit information to the GnRH neurons. Thus, GnRH secretion represents the output of neurons that integrate information from a variety of control systems.

See also: Adrenal Steroids: Biphasic Effects on Neurons; Estrus and Menstrual Cycles: Neuroendocrine Control; Gonadal Steroid Actions on Brain; Gonadotropin-Releasing Hormone: GnRH-1 System; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Sexual Behavior: Neuroendocrine Control; Stress Response: Sex Differences.

Further Reading

Clarke IJ (1996) Effector mechanisms of the hypothalamus that regulate the anterior pituitary gland. In: Unsicker K (ed.) *The Autonomic Nervous System*, pp. 45–88. Harwood: London.

Clarke IJ (1996) The hypothalamo-pituitary axis. In: Hillier SG, Kitchener HC, and Neilsen JP (eds.) *Scientific Essentials of Reproductive Medicine*, pp. 120–132. London: Saunders.

Clarke IJ and Pompolo S (2005) GnRH synthesis and secretion of GnRH. *Animal Reproduction Science* 88: 29–55.

Gore AC (2002) *GnRH: The Master Molecule of Reproduction*. Boston: Kluwer.

Herbison AE (1998) Multimodal influence of estrogen upon gonadotropin-releasing hormone neurons. *Endocrine Reviews* 19: 302–330.

Smith JT and Clarke IJ (2007) Kisspeptin expression in the brain: Catalyst for the initiation of puberty. *Reviews in Endocrine and Metabolic Disorders* 8(1): 1–9.

Tilbrook AJ and Clarke IJ (2001) Negative feedback regulation of the secretion and actions of GnRH in males. *Biology of Reproduction* 64: 735.

Tilbrook AJ and Clarke IJ (2006) Neuroendocrine mechanisms of innate states of attenuated responsiveness of the hypothalamo-pituitary adrenal axis to stress. *Frontiers in Neuroendocrinology* 27: 285–307.

Tilbrook AJ, Turner AI, and Clarke IJ (2002) Stress and reproduction: Central mechanisms and sex differences in non-rodent species. *Stress* 5: 83–100.

Gonadotropin-Releasing Hormone: GnRH-1 System

S Wray, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

The neuroendocrine gonadotropin-releasing hormone (GnRH) system (also known as the luteinizing hormone-releasing system) is essential for vertebrate reproductive function. In mammals, cells within this system usually express the 'mammalian form' of the GnRH transcript. These cells undergo an unusual developmental journey, migrating from the nasal placode into the central nervous system (CNS), where they then become integral members of the hypothalamic–pituitary–gonadal axis (Figure 1). Throughout this article, these cells will be referred to as GnRH-1 cells. There are between 800 and 2000 neuroendocrine GnRH-1 cells in mammals. Within the rodent forebrain, the neuroendocrine GnRH-1 neurons are distributed in a bilateral continuum extending from the olfactory bulbs to the hypothalamus, with a large number of cells located in the diagonal band of Broca/preoptic area (Figure 1). Rostrally, the GnRH-1 cells are located on either side of the midline, in the dorsal half of the forebrain and close to the midline pial surface. Caudally, the GnRH-1 cells are found in the ventral half of the forebrain in more-lateral areas (approaching the hypothalamic sulcus). Independent of location, the majority of GnRH-1 cells send their axons to the median eminence, where they access (via fenestrated capillaries) the pituitary portal capillary system (Figure 1). Release of GnRH-1 at this site is pulsatile and activates the GnRH receptor (GnRHR) that is expressed on the surface of pituitary gonadotrope cells and subsequently affects the synthesis and secretion of gonadotropins (luteinizing hormone (LH) and follicle-stimulating hormone (FSH)) and consequently gonadal function. Genetic evidence from the hypogonadal mouse, which is sterile, established that the GnRH-1 system is essential for reproduction. The hypogonadal mouse has a mutation in only the mammalian GnRH-1 gene, and reproduction can be rescued by GnRH-1 gene therapy.

Although located within the CNS postnatally, neuroendocrine GnRH-1 cells in mammals are first detected in the nasal placode (Figure 2). To date it is still unclear whether GnRH-1 cells differentiate from precursors in the nasal placode or are derived from a stage in development prior to placode formation. It should also be noted that in mammals, GnRH-1 expression appears in the embryo preimplantation, and immunoreactive GnRH-1 is present in the morula and blastocyst stages. In chicken, GnRH-1 expression was found at gastrulation, when expression was detected in cells along the primitive streak. The role of GnRH-1 in early development is not completely understood and, in some cases, may represent transitory expression or a role in nonneuronal cell types.

The migration of neuroendocrine GnRH-1 into the forebrain during prenatal development is universal (Figure 2). GnRH-1 cells migrate on a subset of olfactory/vomeronasal axons, that is, they are neurophilic (also termed axonophilic) in nature and not gliophilic (Figure 3). The developmental relationship of the GnRH-1 system to the olfactory system is illustrated by Kallmann syndrome. Humans with the X-linked form of this disease have a monogenetic mutation that results in anosmia (lack of smell) and hypogonadism (failure to undergo pubertal development). Examination of Kallmann fetal material revealed lack of olfactory axon ingrowth into the developing forebrain and GnRH-1 cells 'stuck' on the CNS–side of the cribriform plate in a tangle of olfactory axons in one case and GnRH-1 cells in only nasal regions in a second case. Although the migration of GnRH-1 cells from nose to brain constitutes the best-characterized example of neurophilic migration in the forebrain, many questions still remain concerning the molecules and cues directing GnRH-1 cells to their final CNS location. But it is clear that disruption of either the development or the regulation of the GnRH-1 system results in reproductive dysfunction.

GnRH-Expressing Cell Groups within the CNS

A total of 24 distinct 'GnRH' forms have been described. Discussion of GnRH cell groups has been confounded by the classification of GnRH genes because of variants in specific species, developmental origin, cell location, and function of the gene products. Certainly, as complete sequence information becomes available for more species, classification may be revised, but several generalizations can be made. In vertebrates, a maximum of three GnRH forms have been detected in any given species, each encoded by a separate GnRH gene: GnRH-1 (human chromosome 8), GnRH-2 (human chromosome 20), and GnRH-3 (primarily in teleosts). Duplication of the genome has been proposed to account for the emergence of multiple GnRH genes. All the genes possess four exons and three introns, and the products

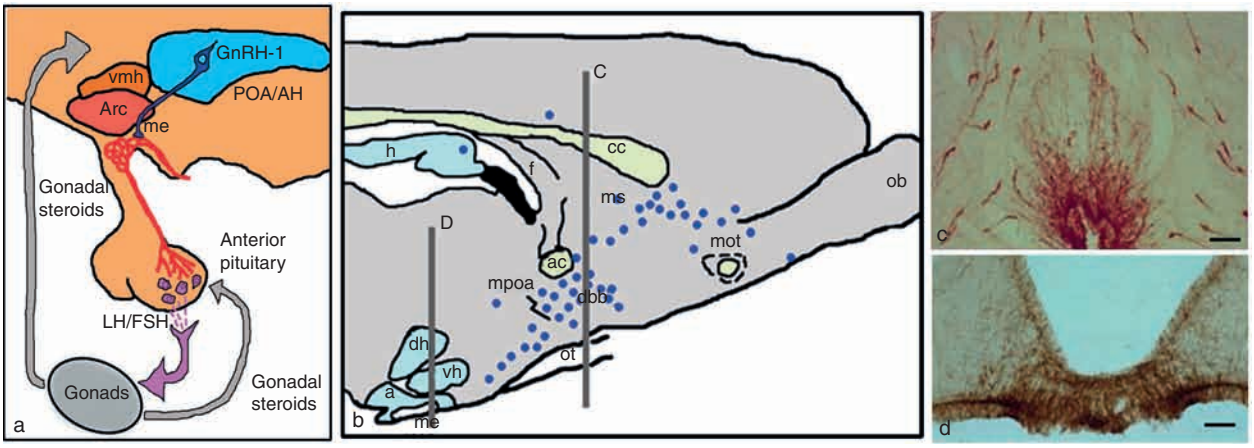

Figure 1 Overview of gonadotropin-releasing hormone-1 (GnRH-1) neuroendocrine system. (a) Hypothalamic–pituitary–gonadal axis. The majority of GnRH-1 cells send axons to the median eminence (me), where GnRH-1 is secreted into the portal capillary system. It is via this system that GnRH-1 cells affect gonadotropes of the anterior pituitary and subsequently gonadal function. POA/AH, preoptic area/anterior hypothalamus; vmh, ventromedial hypothalamus; Arc, arcuate nucleus; LH, luteinizing hormone; FSH, follicle-stimulating hormone. (b) Schematic of a parasagittal section of a rodent brain indicating the location of GnRH-1 cells (blue dots). GnRH-1 cells are distributed bilaterally in a continuum from the olfactory bulbs to the caudal hypothalamus; h, hippocampus; f, fornix; ob, olfactory bulb; mot, medial olfactory tract; ms, medial septum; dbb, diagonal band of Broca; mpoa, medial preoptic area; dh, dorsomedial hypothalamus; ot, optic tract; vh, ventromedial hypothalamus; a, arcuate nucleus; me, median eminence. (c, d) Photomicrographs of sections ((c) = DBB; (d) = ME); immunocytochemically stained for GnRH-1 (at levels indicated in (b)). Scale bar = 50 μm (c) and 330 μm (d).

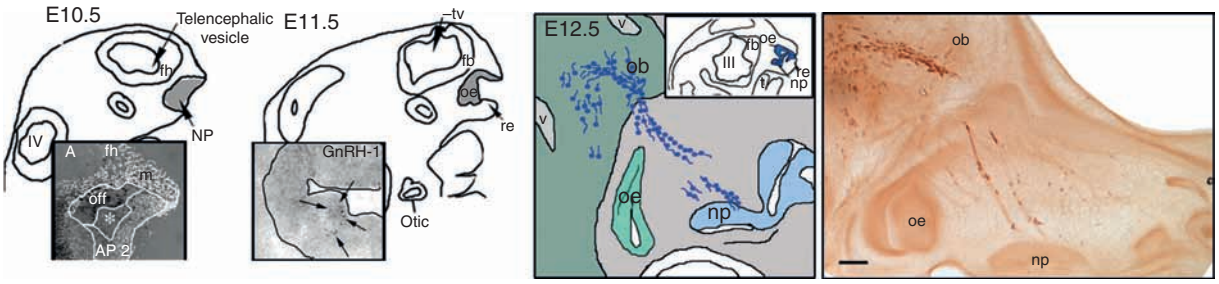

Figure 2 Development of the nasal placode (NP) and migration of gonadotropin-releasing hormone-1 (GnRH-1) neurons. Camera lucida drawing from embryonic day (E) 10.5–E12.5 mouse embryos showing invagination of the nasal placode and the development of the respiratory epithelium (re), olfactory epithelium (oe), and -nasal pit (np). tv, telencephalon; III, third ventricle; IV, fourth ventricle; t, tongue; fb, forebrain; ob, olfactory bulb. Parasagittal section immunocytochemically stained for GnRH-1 shows cells migrating from the nasal pit/vomeronasal organ, apposed to axons, through the nasal forebrain junction into the forebrain. Scale bar = 100 μm.

from these genes show a similar tripartite structure. Exon 1 encodes the 5′UTR, and exon 2 encodes a signal sequence, the GnRH decapeptide and the N-terminal region of gonadotropin-associated peptide (GAP, usually 50–60 amino acids in length). Exon 3 encodes the majority of GAP residues, and exon 4 encodes the C-terminal GAP residues and the entire 3′UTR. The function of GAP is unknown in any species. All known GnRH forms are ten amino acids in length, and the most common variation between forms occurs in amino acid positions 5–8.

GnRH-1 is in most vertebrates and clearly is the form in mammals that controls reproductive function. GnRH-1 is present in axons in the median eminence, where GnRH-1 is secreted into the portal capillary blood and thus affects pituitary function (Figure 1). Although the location of the GnRH-1 cells postnatally is species-dependent, they are often referred to as preoptic area/hypothalamic GnRH cells. The peptides encoded by the GnRH-1 gene can vary by as much as 50% among species.

GnRH-2 is highly conserved (this decapeptide is the most ancient and conserved form of GnRH) and, when present, is consistently found in cells in the midbrain. The function of GnRH-2 is unclear, but roles for GnRH-2 in regulation of puberty, reproductive behavior, feeding, and energy balance have been proposed in mammals. In some species, immunocytochemical and biochemical data suggested widespread expression of GnRH-2, with GnRH-2-expressing cell bodies within

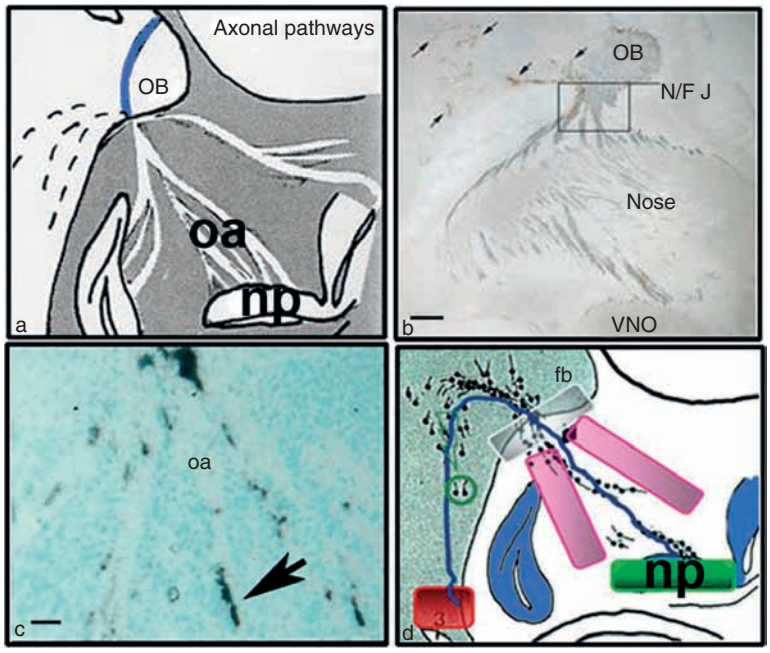

Figure 3 Migration of gonadotropin-releasing hormone-1 (GnRH-1) cells is neurophilic. GnRH-1 cells migrate along olfactory axons (a–c) and are exposed to multiple regulatory domains along the migratory route (d). (a) Schematic of the nasal area with olfactory axons (oa) crosssing to the developing forebrain. (b) Mouse section stained for GnRH-1 (brown) and peripherin (blue, a marker for sensory axons). (c) Cells expressing GnRH-1 transcript (black grains, arrow) apposed to cell poor areas (olfactory axons, oa). (d) Schematic highlighting domains which may produce factors that regulate GnRH-1 neuronal movement (light green; nasal pit, blue; olfactory epitelium, pink; nasal mesenchyme, white; nasal forebrain junction, blue line; transient pathway, red; basal forebrain, dark green; GnRH-1 cells themselves or other migrating cells along pathway). fb, forebrain; OB, olfactory bulb; N/FJ, nasal forebrain junction; VNO, vorneronasal organ; 3, third ventricle.

the hippocampus, thalamus, and hypothalamus. However, recent data indicate that the preproGnRH-2 gene has, during evolution, been deleted from the genome of rats and mice and inactivated in chimpanzee, sheep, and bovine. In humans, GnRH-2 messenger RNA (mRNA) may have stability and/or translation changes due to modifications. Thus, a physiological role for GnRH-2 remains controversial, especially as more information on the receptor becomes available.

GnRH-3, prevalent in fish, exhibits functional, anatomical, and developmental similarities to GnRH-1, depending on the species studied, and in some fish, it has the strongest projection to the pituitary, suggesting a primary function in reproduction. Many fish express all three forms of GnRH (GnRH-1, -2, and -3). In teleosts, when two forms of GnRH are expressed, one is GnRH-2 and the other is (1) GnRH-1 in ancient teleosts but (2) GnRH-3 in modern teleosts (zebra fish), with these fish appearing to have lost the mammalian GnRH-1 ortholog. Usually GnRH-3 is found in the terminal nerve/ganglia and rostral telencephalon. If GnRH-1 is absent, then GnRH-3 can also be found in areas normally populated with GnRH-1 cells.

Multiple GnRHRs have also been identified, with mammals exhibiting primarily two GnRHRs (GnRHR, human chromosome 4, and GnRHR2, human chromosome 1). Both genes encode a seven-transmembrane G-protein-coupled receptor. GnRHR is found on cells in the pituitary and is 'tailless' compared with GnRHR2. In many species that exhibit multiple forms of GnRH and GnRHRs, specificity of receptor to a specific GnRH ligand is often not observed. In general, though, GnRHR has a higher affinity for GnRH-1, and GnRHR2 has a higher affinity for GnRH-2. Evidence that GnRHR2 has been inactivated or deleted from the genome during evolution exists for rats, mice, cow, sheep, chimpanzee, and humans. Thus, even if GnRH-2 is present in mammals, functional analysis of GnRHR2 needs to be evaluated at the species level. The mammalian GnRHR appears to be unique to mammals and cannot be found in amphibians and fish.

Although in the past, GnRH-1 cells and GnRH-2 cells in mammals were easily discernible based on anatomical location and developmental origin, the scenario has become more complex as (1) populations of forebrain cells that transiently express GnRH-1 transcript have been identified, (2) cells expressing GnRH-1 transcript but truncated peptide (amino acids 1–5) have been found in extrahypothalamic areas, (3) GnRH-1-expressing cells have been characterized in the developing tooth 'placode,' (4) an early group of GnRH-1 cells was identified that populated forebrain areas but also appeared to arise from the nasal placode area (reminiscent of GnRH-1 and

GnRH-3 in some fish), and (5) GnRH-2-expressing cells have expanded to nonmidbrain regions. One can initially separate and identify GnRH-1 cells by referring to the neuroendocrine GnRH-1 cells. These cells project to the median eminence, alter pituitary function, and are essential for reproduction. Thus, there are three forebrain populations of 'nonneuroendocrine' GnRH-1-expressing neurons that originate in the brain and express GnRH-1 transiently during development and are not 'related' to neuroendocrine GnRH-1 cells *per se*, although they clearly express the correct combination of transcription factors for GnRH-1 production. The function, if any, of GnRH-1 in these nonneuroendocrine brain cells is unknown. Nonneuroendocrine GnRH-1-expressing cells in rodent teeth may provide developmental information relevant to the neuroendocrine GnRH-1 population since development of the tooth is a dynamic process that involves the interaction of ectoderm-derived dental epithelial cells and underlying neural crest-derived mesenchymal cells – an event that has similarities to invagination of the olfactory and respiratory epithelia in the nasal placode. In the tooth, the GnRH-1 cells are in epithelial-derived structures and not in neural crest-derived structures (see the section titled 'Development of GnRH-1, GnRH-2, and GnRH-3 cell groups'). Also noteworthy are several independent reports documenting dental agenesis in patients with Kallmann syndrome. Whether these occurrences are due to overlaps in craniofacial factors and/or GnRH-1 factors is not known. The occurrence and function of GnRH-2 cells outside mesencephalic areas in mammals awaits further research. Although certainly more complicated, the GnRH-1 and GnRH-2 systems are clearly discernible on the basis of sequence differences, postnatal anatomical distribution, different receptor specificity, and developmental history.

Development of GnRH-1, GnRH-2, and GnRH-3 Cell Groups

This section discusses only cells expressing both transcript and peptide for GnRH-1, -2, or -3, pre- and postnatally.

In 1989, two groups independently showed that neuroendocrine GnRH-1 neurons in mice arise from an extra-CNS origin and suggested the olfactory placode (now referred to as nasal placode) as the source since the GnRH-1 cells were first detected in the evaginating vomeronasal organ/nasal pit (Figure 2). Classically, it was thought that the olfactory placode gives rise to nonsensory respiratory and sensory olfactory epithelium (OE; Figure 2). The OE produces the main olfactory (chemosensory) epithelia, as well as the vomeronasal organ epithelia (pheromone reception). The location of the GnRH-1 cells led to the hypothesis that GnRH-1 cells originate in an area of the olfactory placode associated with OE-derived structures. However, ablation experiments in the chick, subdividing the presumptive olfactory and respiratory areas prior to placode formation, suggested that the GnRH-1 progenitors were more closely related to respiratory areas. Removal of the posterior area eliminated olfactory epithelial structures but spared the GnRH-1 nasal population. Conversely, removal of the anterior area eliminated GnRH-1 cells and respiratory epithelial cells but spared olfactory epithelial structures. Subsequent studies in chicken and mouse argued that OE progenitors were distinct from GnRH-1 progenitors. These studies included (1) screening of early-expressing mouse GnRH-1 cells for the olfactory epithelial lineage markers Mash-1, Math4A, Math4C/neurogenin1, and NeuroD (the markers were not detected in GnRH-1 cells) and (2) identification of 'ectopic' GnRH-1 cells emerging from the respiratory epithelium in a mutant mouse and in the respiratory epithelium of normal chick. These studies argue that GnRH-1 progenitors are closely associated with anterior or respiratory regions rather than with olfactory regions and that GnRH-1 neurons diverge from olfactory sensory cells during early stages of nasal placode differentiation. Therefore the more generic name, nasal placode, was used in mouse to delineate the structure from which neuroendocrine GnRH-1 cells arose. In addition, in mouse nasal placode (Figure 2), an intermediate area was identified between respiratory cells and sensory cells, and it has been suggested that this intermediate region normally gives rise to GnRH-1 cells and hence is the location of GnRH-1 progenitor cells. However, the lineage and location of the precursors of neuroendocrine GnRH-1 cells still remain to be determined, but experiments in most species document expression of neuroendocrine GnRH-1 cells first in the nasal placode area.

The embryonic origin of forebrain GnRH-1 neurons is perhaps more 'controversial' in nonmammalian taxa, with most work being performed in fish. Many different species have been examined, and different sites of origin have been reported for GnRH-1 and-3. By the use of molecular techniques in fish with one rostral forebrain population (either GnRH-1 or GnRH-3), the cells are first detected in the nasal epithelium, similar to mammals. If the fish has two rostral GnRH forebrain populations (both 1 and 3), then the neurons can share a nasal placode developmental origin or exhibit different developmental origins, nasal placode and telencephalon. These studies are based on the expression of appropriate GnRH mRNAs and/or green fluorescent proteins driven by specific GnRH promoters. As in mammals, it is still unclear whether

the precursors are located within the area of first detection or end up in this region as the placode fields form.

In zebra fish, after exhaustive searches, only GnRH-2 and GnRH-3 have been molecularly cloned, and it has been shown that all the rostral forebrain cells in zebra fish express GnRH-3 transcript. However, studies using lineage tracer dyes and immunocytochemistry or gene knockdown fish and immunocytochemistry have proposed that in zebra fish, GnRH-1 and GnRH-3 cell populations arise from two separate regions of the neural plate: the adenohypophysis (anterior pituitary) region of the anterior neural plate and cranial neural crest, respectively, and that GnRH-2 cells of the midbrain share a common neural crest origin with GnRH-3 cells (terminal nerve). Here, the terms GnRH-1 and GnRH-3 refer to the location of the cell population, with GnRH-1 cells eventually residing within the hypothalamus and GnRH-3 cells being associated with the terminal nerve and rostral telencephalon. Although these results have raised new questions about the origin of GnRH cell groups, the results are hard to reconcile with molecular work showing only GnRH-2 and one forebrain GnRH gene in zebra fish (GnRH-3, but not GnRH-1), and GnRH-3 cells' (both those of the terminal nerve and hypothalamus) sharing a common origin: nasal epithelium. The differences observed in fish, and often the same species of fish, remain to be resolved.

Thus, while the exact precursors of GnRH-1, -2, and -3 are still elusive, the consensus is that GnRH-1 and GnRH-3 cells are associated with the nasal placode as they turn on GnRH expression. Both GnRH-1- and GnRH-3-expressing cells thereafter migrate, to various extents, within the olfactory system and into rostral forebrain areas. Again, in most mammals, GnRH-3 has been lost, and the function of these cells may have been incorporated into the GnRH-1 group. The cells discussed, for the most part, in the remainder of this article express the GnRH-1 transcript, originate in placodal regions, migrate into the CNS, synthesize and process the decapeptide GnRH, and are essential for reproduction.

Migration of GnRH-1 Cells into the CNS

Nasal Regions

It is clear from studies examining GnRH-1 cell movement in nasal regions that the development of the olfactory system and the development of the GnRH-1 neuroendocrine system are intimately entwined (Figure 3). In all species examined to date, GnRH-1 cells migrate across the nasal region on axons originating from cells in the olfactory pit/vomeronasal organ (Figure 3). Several of the other phenotypes localized to the nasal placode (γ-aminobutyricacidergic (GABAergic), somatostatin-, neuropeptide Y-, tyrosine hydroxylase-, and calbindin D-28k-expressing cells) have also been shown to migrate along these same axonal pathways in nasal regions. A variety of molecules expressed on olfactory/vomeronasal axons mark these pathways, but to date none has exclusively been limited to the route on which GnRH-1 cells move in nasal regions. However, what is clear is that when the course of the olfactory/vomeronasal axons is altered, so is the migration of the GnRH-1 cells (or GnRH-3 in the case of zebra fish).

Prior to GnRH-1 migration, a group of 'pioneer' olfactory axons crosses the nasal region and targets the base of the developing telencephalon. The cues that direct this group of pioneer axons may be critical for the guidance and targeting of all future olfactory/vomeronasal axons and subsequently the appropriate migration of GnRH-1 cells to the developing forebrain. In addition, migratory neuronal progenitor cells and neuronal precursors have been identified leaving the nasal placode prior to pioneer axonal outgrowth in nasal regions. Thus, exactly how guidance cues are established in the developing olfactory system is still unclear. Once guidance cues are in place, however, it is apparent that the GnRH-1 cells use a subset of olfactory/vomeronasal axons as their pathway across the nasal region. At the light microscopic level, GnRH-1 cells have been shown to be associated with axons expressing the intermediate filament peripherin, which is a marker for all olfactory/vomeronasal axons. When examined at the electron microscopic level, GnRH-1 cells in the submucosa of the nasal septum were always found within axonal fascicles and never appeared in direct contact with the nasal mesenchymal cells. The association of GnRH-1 cells with peripherin-positive nasal axons, and not mesenchymal cells, led to the hypothesis that a cell-adhesion molecule or molecules may be involved in the movement of GnRH-1 cells in this region.

Kallmann Syndrome

Although a variety of molecules are expressed on olfactory/vomeronasal axons, none has been limited exclusively to the GnRH-1 cell route, and thus the molecular guides that lead to this pathway choice are unknown. Recently, genetic work (by several groups) on Kallmann (anosmic) and normosmic hypogonadotropic hypogonadism patients has yielded molecules that influence GnRH-1 development, including anosmin-1, FGFR1, GPR54, NELF, and prokineticin. All but GPR54 appear to act prenatally and indirectly via olfactory axon outgrowth. The *Kal* gene has been lost from the mouse/rat genome. In medaka, a fish with both GnRH-1 and GnRH-3, knockdown

of the Kal ortholog (Kal 1.1) disrupted migration of both GnRH cell populations, while in zebra fish, with only GnRH-3, knockdown of this same gene (*Kal1.1*) disrupted cell migration into the hypothalamus but did not affect those remaining in the terminal nerve. The *C. elegans* Kal-1 has also been cloned and shown to interact with multiple heparan sulfate proteoglycans to promote neuroblast migration. Where examined, Anosmin-1, FGFR1, NELF, and Prokinectin are not limited to nasal regions during development, often being found in areas actively showing cell/neuronal migration. In patients with Kallmann syndrome, other deficits occur, both within the CNS and in other organs. Whether deficits occur in other regions in vertebrates in general, after knockdown on these genes, remains to be determined. As genetic details emerge from work in human patients who come to the clinic because of pubertal delay, one can expect other genes involved in the development or migration of GnRH-1 neurons to be discovered.

Nasal–Forebrain Junction

The crossover from nasal region to CNS is a dramatic change in environment. At this region, GnRH-1 cells appear to pause, and this pause appears to be regulated, in part, by GABA via $GABA_A$ receptors. The reason for this pause is unclear, but it may ensure (1) maturation of GnRH-1 neurons, (2) establishment and/or targeting of the migratory pathway to appropriate brain regions, and/or (3) changes in the extracellular milieu composition. Chemokines, growth factors, and neurotransmitters localized to, or forming gradients at, the nasal forebrain junction have been shown to alter movement of GnRH-1 cells into the forebrain or out of nasal regions.

Pathway in CNS

It appears that a single pathway is used by GnRH-1 cells through nasal regions into the developing hypothalamus in mammals. The majority of peripherin-positive olfactory axons project to the main and accessory olfactory bulbs. However, in the embryonic mouse, peripherin has been shown to highlight a pathway that enters the CNS and turns caudally toward the developing hypothalamus. GnRH-1 cells have been shown to be associated with this caudally projecting peripherin-positive axonal pathway. This pathway (often termed the nervus terminalis or the caudal branch of the vomeronasal nerve) has also been identified by the markers TAG-1 (an axonal surface glycoprotein that is transiently expressed) and deleted in colorectal cancer (DCC; a receptor for the guidance molecule netrin-1). These two markers are expressed on olfactory/vomeronasal axons in the nasal region but are low, or not expressed, on the branches of these nerves that enter the forebrain and proceed rostrally toward the olfactory bulb/accessory regions. Certainly, guidance of this caudal axonal pathway is an important prerequisite for the establishment of the adultlike GnRH-1 cell distribution. In DCC knockout mice, the trajectory of the caudal branch of the vomeronasal nerve is abnormal, turning toward the cerebral cortex instead of the hypothalamus. GnRH-1 cells follow this pathway and therefore do not reach the hypothalamus but migrate into the cortex. GnRH-1 cells must differentiate between axons ending within olfactory regions and those reaching the hypothalamus. Some molecules appear to be expressed on the rostral olfactory/vomeronasal axons and to be absent from the caudal hypothalamic projection. In addition, some molecules are downregulated on GnRH-1 neurons as they enter the forebrain. This raises the possibility that a combination of molecules, some turning on while others turn off, play an important role in establishing the appropriate GnRH-1 adultlike distribution by ensuring that most GnRH-1 cells do not end up in forebrain olfactory regions. Clearly multiple regulatory domains exist as GnRH-1 (and GnRH-3) neurons migrate from their origin to their final destination within the CNS (**Figure 3**).

See also: Adrenal Steroids: Biphasic Effects on Neurons; Estrus and Menstrual Cycles: Neuroendocrine Control; Gonadal Steroid Actions on Brain; Gonadotropin, Neural and Hormonal Control; Olfactory Neuron Patterning and Specification; Sexual Behavior: Neuroendocrine Control; Stress Response: Sex Differences.

Further Reading

Cadman SM, Kim SH, Hu Y, Gonzalez-Martinez D, and Bouloux PM (2006) Molecular pathogenesis of Kallmann's syndrome. *Hormone Research* 67(5): 231–242.

Fernald RD and White RB (1999) Gonadotropin-releasing hormone genes: Phylogeny, structure and functions. *Frontiers in Neuroendocrinology* 20: 224–240.

Herbison AE (2006) Physiology of the GnRH neuronal network. In: Neill JD (ed.) *Physiology of Reproduction*, pp. 1415–1482. San Diego, CA: Academic Press.

Morgan K and Millar RP (2004) Evolution of GnRH ligand precursors and GnRH receptors in protochordate and vertebrate species. *General and Comparative Endocrinology* 139: 191–197.

Pitteloud N, Meysing A, Quinton R, et al. (2006) Mutations in fibroblast growth factor receptor 1 cause Kallmann syndrome with a wide spectrum of reproductive phenotypes. *Molecular and Cellular Endocrinology* 254–255: 60–69.

Skynner MJ, Slater R, Sim JA, Allen ND, and Herbison AE (1999) Promoter transgenics reveal multiple gonadotropin-releasing hormone-I-expressing cell populations of different embryological origin in mouse brain. *Journal of Neuroscience* 19: 5955–5966.

Tobet SA and Schwarting GA (2006) Minireview: Recent progress in gonado-tropin-releasing hormone neuronal migration. *Endocrinology* 147: 1159–1165.

Whitlock KE, Illing N, Brideau NJ, Smith KM, and Twomey S (2006) Development of GnRH cells: Setting the stage for puberty. *Molecular and Cellular Endocrinology* 254–255: 39–50.

Wierman ME, Pawloski JE, Allen MP, Xu M, Linseman DA, and Nielsen-Preiss S (2004) Molecular mechanisms of gonadotropin-releasing hormone neuronal migration. *Trends in Endocrinology and Metabolism* 15: 96–102.

Wray S (2002) Development of gonadotropin-releasing hormone-1 neurons. *Frontiers in Neuroendocrinology* 23: 292–316.

Gravitational Effects on Brain and Behavior

L R Young, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Balance on Earth

Our sense of the direction of the vertical is vital to avoid falling. On earth, the vertical is sensed by a combination of many different sensory organs. Visual, vestibular, haptic, proprioceptive, and auditory cues are combined with knowledge of voluntary movement and the sense of body axis to produce a single, usually consistent, perception of body orientation and motion. Similar processes are involved in the stabilization of the head and the eye, in order to avoid slip of the retinal image of the seen world with every step. Consider the simple act of tilting one's head to the shoulder. The voluntary movement is confirmed by the muscle and joint receptors at the neck, by the tilted visual scene imaged on the retina, by the increased pressure on one foot, and by two sets of signals from the balance organ in the inner ear. The semicircular canals sense the angular motion. The otolith organs, which are sensitive to gravity and linear acceleration, sense that the head is tilted relative to the vertical.

Orientation in Weightlessness

In the weightless free-fall condition of orbital space flight, the synergy among the sensors is drastically altered. The otolith organs no longer indicate anything meaningful concerning the static orientation of the head, since the dense otoconia on the otolithic membrane no longer pull the membrane toward the vertical, thus bending the hair cells of the maculae when the head is tilted. In weightlessness, the otolith organs act like any other accelerometer, or a pendulum for that matter, and only respond to linear acceleration as the astronaut moves about. Once deprived of the usual static orientation information from the otolith organs, the brain must turn to other sources to erect a reference frame with respect to which an astronaut can judge his or her orientation. According to the otolith–tilt-translation–reinterpretation (OTTR) theory, the brain learns to reinterpret novel otolith quasi-static signals to represent linear acceleration in space, rather than the usual interpretation of tilt relative to the earth vertical. As a consequence astronauts may feel themselves moving linearly to the side when tilting their head shortly after return to earth, and they may also have difficulty walking around corners.

Weightlessness and Space Motion Sickness

Motion sickness, whether it is sea sickness, car sickness, or even 'space sickness,' is generally associated with a sensorimotor conflict. When the nervous system is presented with orientation signals that are inconsistent with one another or with the motor commands associated with motion, the conflict stimulates the autonomic nervous system to produce a pattern of responses, from nausea to vomiting.

In space, upon initial head movements in weightlessness, the conflict between the unexpected otolith signals and the remaining senses at first can produce such a malady, known as 'space motion sickness,' which occurs in about two-thirds of all space travelers and can last from hours to a few days. It is not evident how to predict who will be sick, although repeat travelers are less likely to have difficulty. In space, the presence of unusual visual orientations, such as seeing a fellow crew member upside down (relative to the viewer and the spacecraft) or seeing the earth at the top of a window, can suddenly bring on vomiting. Space motion sickness is often treated by an intramuscular injection of promethazine, which seems to produce fewer side effects of drowsiness in space than on the ground.

Posture in Space

When a new astronaut is first exposed to space, the postural reactions are inappropriate. Legs keep getting in the way and there is a tendency to stand against the floor, which, in turn, propels the astronaut headward toward the ceiling. After a few hours, however, space locomotion is learned, and after several days most astronauts comfortably and efficiently move about, using hand-holds or precise forces against the walls. Many enjoy working 'inverted,' although their preflight training may be of limited use to them.

Visual Reorientation Illusions

Early in a space mission most astronauts tend to substitute visual orientation cues for the misleading otolith signals, and become more sensitive to self-motion created by moving visual fields. As described by Charles Oman (director of the Man/Vehicle Laboratory at the Massachusetts Institute of Technology), visual reorientation illusions are commonly reported, in which the spacecraft surface closest to being beneath the astronaut's feet suddenly becomes the

apparent 'floor,' even if it is a ceiling or wall. The subjective frame may suddenly shift, analogous to a figure reversal illusion such as the Necker cube, without any sense of being rotated. Entering a part of the spacecraft in an unexpected orientation, or seeing a crewmate 'upside down,' may trigger a sudden visual reorientation illusion. Localized tactile cues, such as foot pressure applied by standing against stretched elastic cords, may accelerate the switch of reference frame and then stabilize the sense of orientation, and may enhance the illusion that the surface beneath the floor feet is 'down.' Outside the spacecraft, during extravehicular activity, the visual reorientation illusion is even more striking (see Figure 1). When working close to some surface, such as the outside of a space station or the open cargo bay of the Space Shuttle, astronauts will normally adopt the local surface beneath their feet as the 'down' reference. When looking off into space with the earth in view, however, our home planet becomes the vertical reference, and astronauts frequently report height vertigo and a fear of falling, causing them to clutch their hand-hold strenuously.

Another common phenomenon, distinct from the visual reorientation, is the 'inversion illusion,' typically occurring shortly after entry into weightlessness and persisting even with eyes closed. The usual description of the inversion illusion includes a momentary sensation of tumbling immediately upon entering weightlessness, followed by a sensation of being upside down, which generally lasts from seconds to minutes but which can, in some cases, remain for days. It is not influenced by visual inputs. After 4–7 days, however, some astronauts seem to adopt an internal reference system, relying upon their own body axis rather than any external cues, and their orientation illusions are either eliminated or cease to be disturbing. Many others adopt a local reference frame whereby the floor of the module in which they are working represents a universal 'down' direction – at least until they move into another module.

Figure 1 NASA photo of an astronaut performing extravehicular activity, hanging from the International Space Station (ISS), with earth below.

Orientation to Gravity and to Free Fall

The role of gravity in the interpretation of a visual upright is well known and appears in the Aubert illusion, whereby a sole vertical line appears tilted when the head is tilted sideways. For most stationary subjects, on Earth, the subjective visual vertical lies between the direction of the true vertical, as measured by the graviceptors, and that defined by familiar objects in the visual scene. The tendency toward visual dominance is often measured by the rod-and-frame test, in which the surrounding lighted frame influences the perceived visual vertical. In weightlessness, absent any orientation information from the graviceptors, the subjective visual vertical normally aligns with either the visual vertical or the body axis.

Vection

When a large visual scene is moved in front of a stationary subject on earth, it generally causes a sensation of self-motion, or 'vection,' after several seconds. In space, absent the expected motion cues from the semicircular canals or otolith organs, which might otherwise inhibit or eliminate such illusions, the vection is strengthened and occurs more rapidly (see Figure 2). This can, however, be reduced by the provision of static tactile cues to the feet, when the individual stands upright against the force of stretched elastic cords.

Proprioception

Proprioception is altered in space because of the absence of static muscle tension required to support the weight of a limb, but accurate pointing is quickly learned. The proprioceptive feedback signaling ankle, knee, and hip flexion is disturbed during extended weightlessness and is likely a significant source of the inability of some astronauts to walk and run normally following return to earth. The ability to

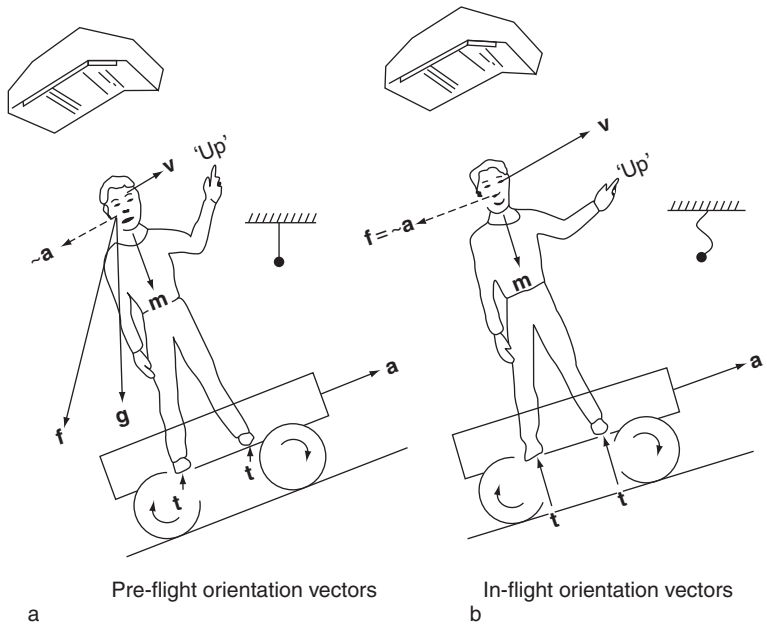

Figure 2 (a) On earth, the typical individual points upward, based largely on otolith cues measuring the direction of **f**, the specific force vector **g–a**. The direction is influenced also by the visual orientation **v** of apparently vertical surfaces, such as the panels of a work station, by the tactile cues **t,** coming from forces on the feet, and by an internal ideotropic vector **m,** aligned with the body axis. A ball held by the individual would hang straight down. (b) In weightlessness, however, the otolith measure of **f** indicates only acceleration and is largely ignored, in favor of increased weighting of the visual, tactile, and body vector signals. The indicated direction of 'up' now aligns more nearly with the direction of the visual ceiling.

estimate the mass of objects, in the absence of their weight, is also reduced. The otolith–spinal reflex, which normally prepares one for a landing from a fall, is also reduced in space, as well as is the Hoffman reflex measurement of spinal cord excitability during transient acceleration.

Vestibuloocular Reflex

Ocular stabilization during head movements, the vestibuloocular reflex, is disturbed in some individuals during space flight. Vertical and horizontal vestibuloocular reflex gains to head turns vary with duration in weightlessness. The time constant characterizing the decay of postrotatory nystagmus (oscillatory eye movements following prolonged turning) is thought to consist of a shorter time, about 5 s, characterizing the mechanical return of the displaced cupula, and a longer period, sometimes referred to as 'velocity storage,' which involves the vestibulocerebellum. Early on during space flight the time constant is shortened, representing a loss of the higher-order velocity storage mechanism. During prolonged flight, however, the time constants return to close to their preflight levels. An interesting gravity-related phenomenon is the sudden decrease in postrotatory nystagmus, known as 'nystagmus dumping,' normally produced by pitching the head forward. This dumping is absent in weightlessness and therefore is attributed to the effect of gravity on the otolith organs. The 'caloric response,' a nystagmus induced by applying hot or cold fluid in one ear, is a common clinical procedure for diagnosing peripheral vestibular loss. It is still present, to a lesser extent, in space, indicating that it is partially caused by some direct thermal effect rather than entirely by gravity-driven convection. Ocular counter-rolling of the eyes, in a direction toward stabilizing the eye relative to the vertical, is a sensitive measure of the response of the utricular otoliths to gravitational stimulation. Most measurements of this response postflight show a decrease in gain for astronauts, and frequently the appearance of some directional asymmetry not seen preflight. This may be interpreted as the emergence of an underlying uncompensated reflex. (No such preflight/postflight difference was found for the four astronauts tested on the 16-day National Aeronautics and Space Administration Neurolab mission, however, nor did these astronauts show any inflight changes in ocular counter-rolling, as tested on a short-radius centrifuge.)

Reentry and Earth Sickness

Although the adaptive behavior of the neurovestibular system is appropriate to living in weightlessness, it introduces problems during reentry and when back

on the ground. Head movements during entry can cause the visual scene to swim around (oscillopsia) and may produce significant spatial disorientation. Such head movements are therefore avoided by reentering pilots. Once back on the ground, many astronauts experience a period of 'earth sickness,' especially when making head movements in a limited visual field. Unlike space motion sickness, however, earth sickness generally disappears in less than a day. Locomotion takes a longer time to readapt. Returning space travelers will usually exhibit a wide stance and unsteady gait after landing. Their ability to maintain balance is impaired, especially when visual and ankle proprioceptive cues are eliminated, forcing them to rely on their vestibular system to remain steady. The ability to stand quietly for 10 min is severely challenged, presumably because of orthostatic hypotension associated with cardiovascular regulation, loss of venous tone, and fluid loss, but also possibly due to altered vestibular inputs to the sympathetic nervous system. (Reduced blood supply to the brain, associated with cardiovascular deconditioning, may contribute to the unsteadiness.) Return to normal spatial orientation, eye movements, and posture may take several days or even weeks following long space flights.

End Organ, Peripheral, and Cerebellar Changes

The end organs may show some dramatic changes in weightlessness. Although the otoconia appear unchanged, the peripheral synaptic connections apparently are altered significantly in space, at least in rats. Monkey eye movements show normal responses in space despite some apparent alterations in the gain or time constants of the basic semicircular afferent responses to rotation. Fish, when deprived of normal gravity cues, will perform looping swimming motions which may serve a postural reaction to remain level. Fish eye movements and cerebellar activity also support a shift toward visual dominance over vestibular in space.

The Neurolab space mission included a number of experiments aimed at uncovering the underlying neurophysiological sources of neurovestibular adaptation to weightlessness. Purkinje cells in the nodulus regions of the cerebellar cortex, receiving input from the otolith organs, showed significant changes in space, suggesting their role in the short-term plasticity seen in response to weightlessness. In an attempt to identify the areas of the brain involved in the plasticity of balance, the rat brain was studied for changes in Fos and fos-related antigens (FRAs) indicating change in immediate-early genes. Such gene expression was found in the vestibular nuclei, the nucleus of the tractus solitarius, the area postrema, and the central nucleus of the amygdala, as well as in certain cortical and subcortical areas and regions of the hypothalamus and brain stem. The peripheral nerve connections to the hair cells which detect movements of the otolithic membrane are subject to extensive changes during exposure to weightlessness. In rats examined early during space flight and again after 2 weeks, the number of cross-links between adjacent hair cells was extensively increased, relative to ground controls. This suggests that at least some of the sensory adaptation to weightlessness occurs at the periphery, changing the sensitivity and even the function of the otolith organs, rather than occurring entirely centrally.

Orientation

Three-dimensional navigation in weightlessness is challenging for humans, and may be related to our inability to maintain a vertical reference frame. Astronauts frequently experience disorientation when moving from one module of a space station to another one with a different orientation of floor and ceiling. Limited investigation of rat place cells during space flight, and head direction cells during short-duration parabolic flights, support the conclusion that two-dimensional navigation, as carried out on earth, requires the existence of an assumed vertical reference. In its absence, the well-known directional tuning could either disappear or show evidence of a shifted reference frame.

Development

The role of gravity in the normal development of the vestibular system and in the development of balance and locomotion remains a subject of importance, still lacking in significant experimental data in higher animals. Chicks developed in space show head stability problems on return to earth, and the postural reactions of rats after 2 weeks in space show marked deficit, from which they generally recover. However, the question remains open as to the existence of a critical period for exposure to gravity in the development of postural stability and spatial orientation.

See also: Balance and Posture Control; Canal–Otolith Interactions; Eye and Head Movements; Motion Sickness; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vestibular System; Vestibulo-Autonomic Responses.

Further Reading

Berthoz A (ed.) (1998) *Special Issue: Gravitational Environments? Brain Research Reviews* 28(1–2): 1–490.

Moore S, Cohen B, Clément G, et al. (2003) Ocular counter-rolling during centrifugation and static stilt. In: Buckey J and Homick J (eds.) *The Neurolab Spacelab Mission: Neuroscience Research in Space*, pp. 11–17. [Results from the STS-90, Neurolab Spacelab Mission. NASA SP-2003–535]. Washington, DC: US Government Printing Office.

Oman C (2007) Spatial orientation and navigation in microgravity. In: Mast FW and Jäncke L (eds.) *Spatial Processing in Navigation, Imagery and Perception*, pp. 209–248. New York: Springer.

Reschke MF, Bloomberg JJ, Harm DL, et al. (1994) Space flight and neurovestibular adaptation. *Journal of Clinical Pharmacology* 34: 609–617.

Spacelab-1 Mission (1986) Special issue. *Experimental Brain Research* 64(2): 237–392.

Spacelab Life Sciences Mission SLS-1 (1996) Special issue. *Journal of Applied Physiology* 81: 1–207.

Young LR, Oman CM, Watt DG, et al. (1984) Spatial orientation in weightlessness and readaptation to Earth's gravity. *Science* 225: 205–208.

Growth Cones

J A Raper, University of Pennsylvania, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons communicate with one another through synaptic connections on their dendritic and axonal processes. These processes can be highly elaborate, and axons in particular can grow long distances along highly precise routes to acquire their appropriate targets. At the tip of each growing process is a specialized motile structure called the growth cone. The growth cone contains the cellular machinery that drives process extension. This extension is not random but is guided by signals in the surrounding cellular and extracellular environment. These signals are called guidance cues. Guidance cues are recognized by receptors whose activation controls the direction of growth cone advance through a complex network of signaling pathways that is still being elucidated.

Growth Cone Structure

A growth cone's shape varies greatly depending on its rate of advance, the physical properties of the substratum on which it is growing, and its responses to guidance cues in the outside environment (Figure 1). A growth cone extending on a flat, featureless, and adhesive surface *in vitro* has a characteristic shape (Figure 2). At the tip of a growing axon is a fan-shaped structure which can be divided into three regions: the central, transitional, and peripheral domains. The central domain is closest to the axon and set back from the leading edge. It is rich in microtubules that splay out from the axon as they enter the growth cone. The peripheral domain is at the leading edge of the growth cone. It is composed of thin lamellar sheets and long thin protrusions called filopodia. The predominant cytoskeletal component of lamellae and filopodia is fibrillar actin. Actin is tightly bundled into a rodlike arrangement that fills each filopodium. Actin is organized into a highly cross-linked and branched network with the growing ends pointed predominantly toward the growth cone's leading edge in lamellipodia. The transition zone, as its name implies, is the region between the microtubule-rich central domain and the actin-rich peripheral domain.

Growth Cone Motility

Axonal extension is driven by the motile force generated by the cytoskeletal machinery within the growth cone. Growth cone motility is thought to be the result of force that is generated by a cycle of actin polymerization at its leading edge, myosin-driven retrograde flow of actin polymers through the peripheral domain, depolymerization of actin as it approaches the transition zone, diffusion of the released actin monomers back to the leading edge, and their repolymerization there (Figure 3). The resulting actin cycle generates a tank-tread-like flow of polymerized actin from the front to the back of the growth cone. When the actin-based tank tread gets a grip on the substratum, it can advance the leading edge and draw the growth cone forward. Coupling of the cytoskeleton to the substratum is accomplished by specialized receptors on the surface of the growth cone that recognize and bind to specific molecules in the outside environment. These receptors are indirectly linked to the actin cytoskeleton within the growth cone by cross-linking proteins. Receptors that provide this motility-enhancing function include integrins that recognize extracellular matrix components like laminin, and cadherins or immunoglobulin (Ig) superfamily members that recognize homophilic or heterophilic cadherins or Ig superfamily members expressed on adjacent cell surfaces.

Growth Cone Steering

Time-lapse observations of growth cones *in vitro* and *in vivo* demonstrate that both filopodia and lamellipodia are highly motile structures. Individual lamellae or filopodia appear to advance stochastically, at times and directions that are difficult to predict. Those that do not establish a secure contact with the substratum are soon withdrawn and resorbed, whereas those that establish a secure contact are more likely to become stabilized and persist. Growth cones facing a sharp boundary between a more permissive and a less permissive substratum tend to consolidate advances on the more permissive area while withdrawing more often from the nonpermissive area. In this way, differential adhesion experienced by spatially distinct lamellar and filopodial processes can steer growth cones onto a more permissive area of the substratum.

The growth cone is a highly polarized structure and advances only at its leading edge. Its ability to make sharp turns is limited by this extreme polarization, and it is relatively difficult to induce a turn, by any means, that is greater than about 90°. Growth cones are sensitive to the topology of the surface on which they extend. They tend to orient in the same direction as a grooved or filamentous substratum on which they are growing. A growth cone generates tension on the axon as it advances, and this tension reorients and stimulates additional growth in the

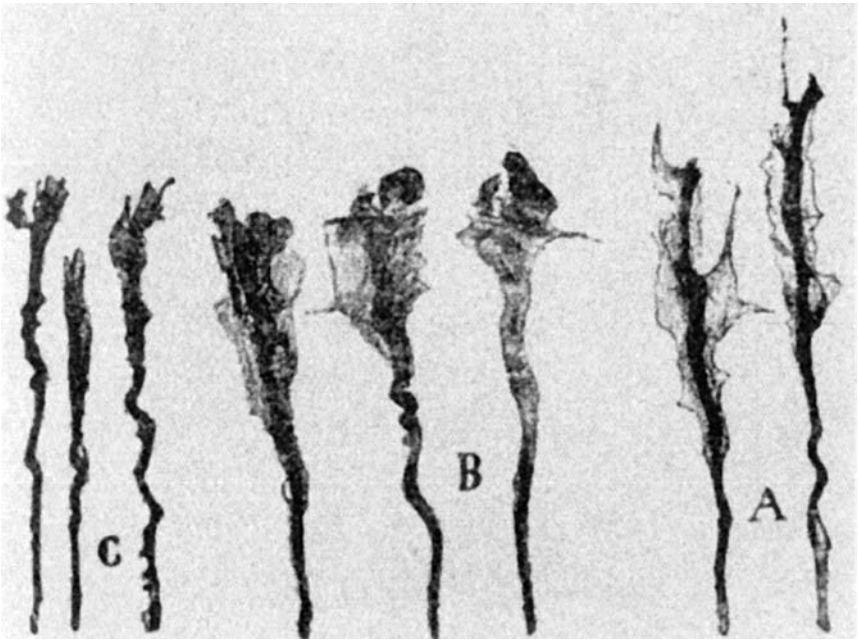

Figure 1 Growth cones were originally described by the famous neuroanatomist Ramón Y Cajal more than a century ago. This is a drawing he made to show the many different shapes they assume (grouped as A, B, and C by Ramón Y Cajal). Some of the variability comes from the stochastic manner in which individual filopodial and lamellar processes advance and withdraw, some is the result of growing in very different environments, and some is due to inherent differences in the overall motile state of the growth cone. From Ramón Y Cajal S (1890) Sur l'origine et les ramifications des fibres nerveuses de la moelle embryonaire. *Anatomischer Anzeiger* 5: 609–613. In Ramón Y Cajal S (ed.) (1995) *Histology of the Nervous System* (Swanson N and Swanson LW, trans.). New York: Oxford University Press. Originally published 1909.

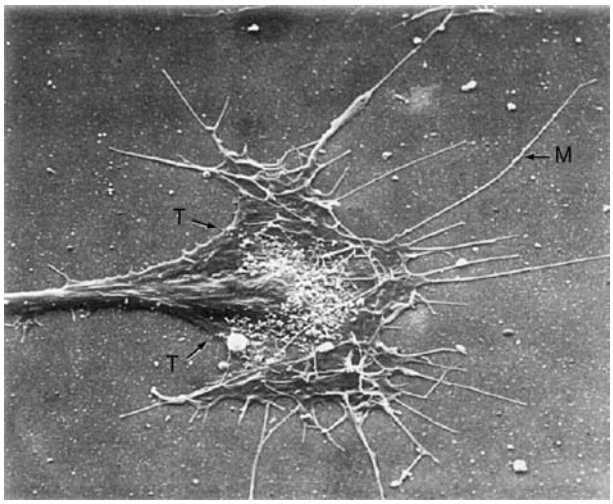

Figure 2 A scanning electron micrograph of a cultured growth cone. Note the thin lamellae and long, thin filopodia (M) projecting radially away from the leading edge and sides of the growth cone. In contrast, few filopodia originate from the traiting edges (T) of the growth cone. This growth cone is approximately 20 μm across. From Wessells NK and Nuttall RP (1978) Normal branching, induced branching, and steering of cultured parasympathetic motor neurons. *Experimental Cell Research* 115(1): 111–122.

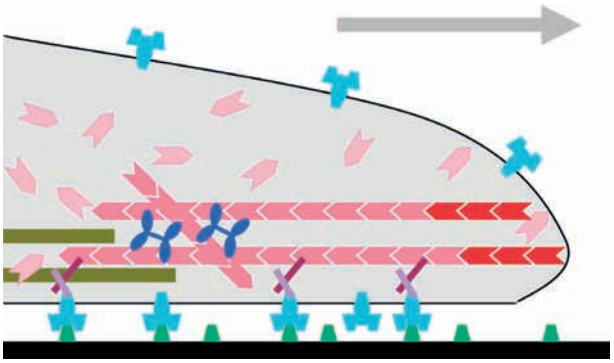

Figure 3 A schematic of the motile machinery at the leading edge of a growth cone that is advancing toward the right. Microtubles (olive) are confined to the proximal portion of the growth cone while fibrilar actin (rose) reaches into the distal most leading edge where it extends by polymerization (red). The resulting actin fibrils are contracted and moved backward by myosins (blue). A continuous cycle of actin polymerization at the leading edge, retrograde flow, and depolymerization at the rear is established. This cycle drives the growth cone forward over a substratum (black) decorated with specific permissive molecules (green) that are recognized by complementary cell surface receptors (light blue) which are linked to fibrilar actin (purples).

same orientation. This is likely to represent a reinforcing mechanism by which growth cones are steered toward more-permissive substrata as processes better stabilized against the substratum are likely to generate more tensile force than poorly stabilized processes. This differential tensile force would be expected to stimulate further additional advance toward a point of contact.

The consolidation of adhesive gains or responses to tensile force can be related to changes in the

organization of the growth cone cytoskeleton. Microtubules and fibrillar actin are in a kind of dynamic equilibrium that maintains their relatively high respective concentrations in the central and peripheral domains of the growth cone. The retrograde flow of actin in the peripheral domain tends to carry or push microtubules away from the leading edge. Some individual microtubules polymerize more rapidly than the retrograde flow of actin for brief periods and advance into the peripheral domain, but they cannot sustain high rates of polymerization long enough to remain there, and they get pushed or carried back to the central domain. In this way, a small subset of microtubules constantly probe the more peripheral regions of the growth cone. When filopodia or lamellae make contact with a highly permissive substratum, for example a preferred cell surface, the retrograde flow of actin is slowed by strong coupling to the substratum. Microtubules are then able to advance toward the site of contact (Figure 4). This invasion of microtubules stabilizes the advanced process and appears to bias further extension in the same direction. Treatments that prevent microtubules from polymerizing quickly make growth cones largely insensitive to external guidance cues. This suggests that stabilization of advancing microtubules as they probe the peripheral domain may help lock in actin-based advances at the leading edge.

A second method by which growth cones can be steered is by modulating the propensity of the leading edge to advance. Any localized change that promotes actin polymerization at the leading edge should tend to advance that region of a growth cone, while conversely, localized signals that inhibit actin polymerization or accelerate retrograde transport should induce retreat. Repellent guidance cues like semaphorin 3A can shut down actin polymerization at the leading edge, halting forward advance. When applied locally, semaphorin 3A discourages the advance of nearby lamellae more effectively than those at a distance. Lamellae that are less subject to the repellent signal advance more successfully than those that are close to the repellent stimulus. The growth cone consequently turns away from the localized source of repellent. Many attractants can be expected to work in a converse manner, locally promoting actin polymerization and the local advance of the leading edge of the growth cone.

Guidance Cues and Their Receptors

Growth cone advance in the developing nervous system is not random but highly stereotyped and organized. Axons generally follow reproducible routes to their appropriate targets, indicating that they are guided there by information in their surrounding environment. To be useful, this information must be distributed in a reproducible and spatially heterogeneous pattern. In consideration of what is known about growth cone motility and steering, this information is likely to fall into at least three general categories. The first is physical features that make advance easier in one

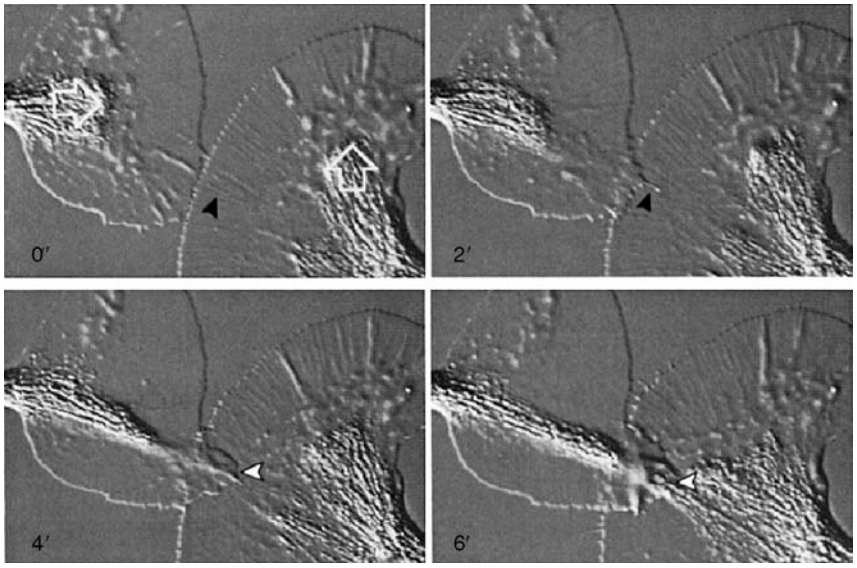

Figure 4 Microtubules within a growth cone are reoriented by contact with another cell surface. A sequence of differential interference contrast photographs of cultured snail neurons begins at 0′ with the initial contact between two growth cones. At 2′, the actin-rich leading edge has advanced markedly at the point of contact. Microtubules in the central domain already appear to be turning toward the point of contact. By 4′, microtubules in the left-most growth cone have advanced dramatically toward the portion of the leading edge that is growing rapidly onto the other cell surface. Black arrowheads mark a filopodium from the left-most growth cone extending on the second growth cone, white arrowheads mark a growing lamellum extending from the left-most growth cone onto the other growth cone. From Lin CH and Forscher P (1993) Cytoskeletal remodeling during growth cone-target interactions. *Journal of Cell Biology* 121(6): 1369–1383.

direction than another. For example, axons are sometimes canalized into a specific route by dense tissues on either side of their path. The second is permissive molecules whose distribution can guide axons by making some pathways better for outgrowth than others. This would in principle be a powerful method of determining where nerve tracts form; however, permissive molecules have rarely been found to be expressed in a prepattern that correlates well with axonal outgrowth. Permissive molecules are often more broadly distributed than the trajectories of forming axonal tracts. Finally, signaling molecules can steer growth cones by locally altering cytoskeletal dynamics. Thus far, it is these signaling cues that seem to play the most important role in guiding axons *in vivo*.

A wide range of disparate signaling molecules have been shown to help guide axons *in vivo*. Some of the first to be characterized include members of the semaphorin, netrin, ephrin, and slit families. A large proportion of these family members can act as guidance molecules, although it would be wrong to think of them as being dedicated to this function alone since many contribute to other developmental and signaling events. In addition to these well-recognized guidance families, signaling molecules that were originally identified as morphogens have been strongly implicated as guidance cues, including members of the Wnt, BMP, and hedgehog families. Additional guidance cues have been identified in the cadherin and Ig superfamilies, among others.

For the most part, each guidance cue has a well-defined receptor or receptor complex which must be expressed by a growth cone for it to be sensitive to that particular cue. For example, each semaphorin is recognized by specific members of the plexin family, ephrins by the members of the eph family of transmembrane tyrosine kinases, and slits and netrins by select members of the Ig superfamily. Most often a traditional ligand–receptor pairing triggers a signaling event within the growth cone that alters its direction, but in some cases, homophilic interactions underlie important guidance events, for example between Down syndrome cell adhesion molecule Ig superfamily members in invertebrates. It is interesting that many of the ligands that act as guidance molecules have some of the same general characteristics as receptors, and in some instances it can be shown that 'reverse signaling' occurs where the presumptive ligand acts as a receptor and the presumptive receptor acts as its ligand. Since a growth cone's response to a particular cue in its environment depends on its ability to recognize the cue, some guidance decisions depend critically on the expression of accessory proteins that control receptor expression. For example, whether an axon chooses to cross the midline of the fly central nervous system can be determined by the expression of a protein called commissureless which in turn controls the axonal expression of a receptor for midline repellents.

Signaling molecules that guide growth cones can be further subdivided into those that serve as attractants or repellents. Attractive interactions can be detected in several different ways. *In vitro*, attractants can sometimes promote axonal outgrowth. A surer test for attractant activity is to demonstrate that a localized source of signal turns or attracts axons (Figure 5). Similarly, *in vivo*, ectopic expression of an attractant will orient sensitive axons toward higher levels of expression, while knockdowns of an attractant are expected to reduce axonal trajectories toward the normal source of signal. Repellents are easily detected *in vitro* by their ability to quickly paralyze growth cone motility. A localized source of soluble or surface-bound repellent can redirect growth cones away from a localized source of signal. *In vivo*, misexpression of a repellent will exclude axons from an area they would normally enter or when knocked down and will encourage sensitive axons to invade an area they would normally avoid. It is interesting that the same guidance cue can act as either an attractant or a repellent. This is observed in

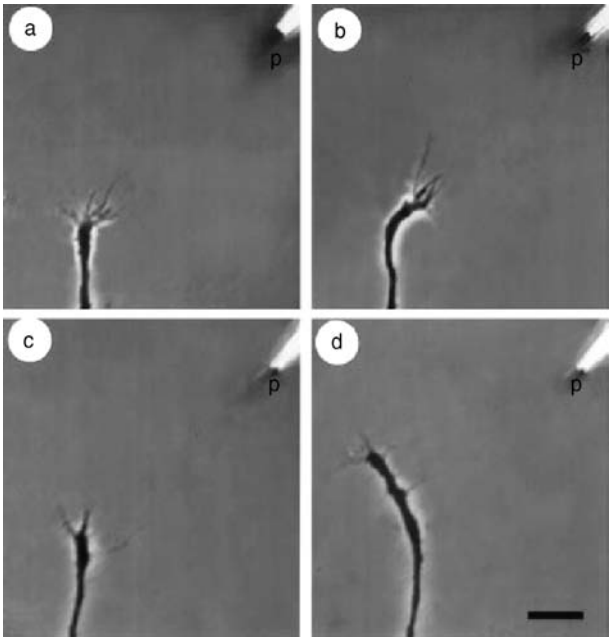

Figure 5 The conversion of an attractant into a repellent induced by a cyclic adenosine monophosphate (cAMP) antagonist (a, b). A cultured frog growth cone is induced to turn toward a source of brain-derived neurotrophic factor (BDNF) ejected from a pipette (p). (c, d) After treatment with the cAMP antagonist Rp-cAMPs, growth cones turn away from a source of BDNF. Scale bar = 20 μm (d). From Song HJ, Ming GL, and Poo MM (1997) cAMP-induced switching in turning direction of nerve growth cones. *Nature* 388(6639): 275–279.

the case of netrin-1. Attraction to netrin-1 is mediated by the expression of the DCC receptor, and repulsion is mediated by expression of the unc-5 receptor.

Guidance cues may act either on contact or at a modest distance. Many guidance cues are transmembrane proteins or are coupled directly to lipids on the cell surface. Unless cleaved by a protease, these cues cannot diffuse away from the cell that produces them and are recognized only by growth cones that touch them. Other guidance cues are secreted and can act at a distance. Thus far it appears that this distance is relatively small, perhaps only a few hundred micrometers or so. This relatively limited range may be partially determined by the required steepness of the concentration gradient required to orient growth cones. An additional limiting factor is that some secreted guidance cues are charged and diffuse poorly through the extracellular matrix. Thus, most growth cone guidance decisions are largely local in nature and are determined by the expression of guidance cues on immediately adjacent surfaces or nearby tissues.

Receptor Signaling

Since many of the receptors for guidance cues discovered over the past decade are novel, the signaling pathways through which they control growth cone motility and orientation are only now being determined. Presumably, their signaling must converge on common pathways that control actin polymerization, retrograde flow, and microtubule dynamics. Among the key downstream signaling intermediaries are the Rho family of small guanosine triphosphatases, RhoA, Rac, and CDC42, which are well known to affect actin polymerization and organization in non-neuronal cells. Activation of many of the known guidance receptors stimulates or inhibits the activity of one or more Rho family members. These in turn modulate the activity of a complex circuit of downstream effector proteins which reorganize the cytoskeleton. A second signaling intermediary of some importance is the divalent cation calcium. Its overall intracellular concentration has been shown to affect growth cone motility, while localized intracellular perturbations in calcium levels can induce localized changes in the cytoskeleton and steering of the growth cone. The activation of some but not all guidance receptors induces significant changes in intracellular calcium that are important in the growth cone's response. A third key signaling component appears to be cyclic nucleotides such as cyclic adenosine monophosphate (cAMP). Altering their internal concentrations can alter how growth cones respond to a given guidance cue. In the most dramatic examples, altering cAMP levels can convert an attractant into a repellent or a repellent into an attractant (Figure 5). From these observations it is evident that the internal state of the growth cone affects how it responds to outside guidance cues.

The growth cone has been a source of fascination ever since it was first described by the famous neuroanatomist Ramón Y Cajal more than a century ago. During the intervening years, great progress has been made in elucidating the basic mechanism through which growth cones advance and identifying many of the guidance cues that steer growth cones. Even so, much remains to be discovered. The signaling pathways through which guidance receptors control the cytoskeleton are only partly known, and a fuller understanding of this signaling circuitry will be essential in understanding how growth cones integrate and respond to multiple competing guidance cues. From a developmental perspective, much of the information about where and when guidance cues affect axon trajectories is fragmentary and incomplete. A more systematic, systems-based examination of all the guidance cues involved in determining any given axonal trajectory will be required to fully explain how growth cones are guided to their correct targets.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Pathfinding: Netrins; Axonal Pathfinding: Guidance Activities of Sonic Hedgehog (Shh); Axonal Pathfinding: Extracellular Matrix Role.

Further Reading

Charron F and Tessier-Lavigne M (2005) Novel brain wiring functions for classical morphogens: A role as graded positional cues in axon guidance. *Development* 132(10): 2251–1262.

Dickson BJ (2002) Molecular mechanisms of axon guidance. *Science* 298(5600): 1959–1964.

Huber AB, Kolodkin AL, Ginty DD, and Cloutier JF (2003) Signaling at the growth cone: Ligand-receptor complexes and the control of axon growth and guidance. *Annual Review of Neuroscience* 26: 509–563.

Lin CH and Forscher P (1993) Cytoskeletal remodeling during growth cone-target interactions. *Journal of Cell Biology* 121(6): 1369–1383.

Luo L (2002) Actin cytoskeleton regulation in neuronal morphogenesis and structural plasticity. *Annual Review of Cell and Developmental Biology* 18: 601–635.

Song HJ, Ming GL, and Poo MM (1997) cAMP-induced switching in turning direction of nerve growth cones. *Nature* 388(6639): 275–279.

Suter DM and Forscher P (2000) Substrate-cytoskeletal coupling as a mechanism for the regulation of growth cone motility and guidance. *Journal of Neurobiology* 44(2): 97–113.

Wessells NK and Nuttall RP (1978) Normal branching, induced branching, and steering of cultured parasympathetic motor neurons. *Experimental Cell Research* 115(1): 111–122.

Growth Factors: Neuronal Atrophy

M von dem Bussche and M H Tuszynski, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Growth factors are proteins that bind to specific receptors on the cell surface, with the primary effect during nervous system development of stimulating cell proliferation, neuronal differentiation, or neurite outgrowth. Expression of growth factors and their receptors continues in the nervous system throughout adulthood, where the growth factors act to modulate the functional state of neurons, influence synaptic efficacy, and potentially modify neuronal structure. Growth factors also exert potent effects on cell survival and axonal outgrowth in animal models of nervous system disease or age-related atrophy. This article focuses on the roles of growth factors in affecting neurons in a state of atrophy arising from either aging or disease.

Classification

More than 50 proteins with growth factor-like properties have been identified in the nervous system. Many of these are grouped into related families of molecules, including the neurotrophins, the glial cell line-derived neurotrophic factor (GDNF) family of ligands (GFLs), the cytokine growth factors, the insulin-like growth factors (IGFs), the fibroblast growth factors, and the hepatocyte growth factor family. These are summarized in Table 1. Effects of individual growth factors, with specific reference to models of aging and neuronal injury, are described in greater detail in the following sections.

The Neurotrophin Family

Nerve growth factor (NGF) The first growth factor to be identified was nerve growth factor (NGF), originally described by Rita Levi-Montalcini and Viktor Hamburger more than 50 years ago. NGF supports the survival of sensory and sympathetic neurons of the peripheral nervous system and establishes the patterning and timing of cholinergic innervation of the hippocampus during development. Of particular interest in the context of neuronal atrophy and aging, NGF promotes the survival and function of cholinergic neurons of the central nervous system (CNS).

Degeneration of basal forebrain cholinergic neurons (BFCNs) occurs in the aged brain and is an early and prominent feature of several age-related disorders of the nervous system, including Alzheimer's disease (AD) and such 'system degenerations' as progressive supranuclear palsy. Considerable evidence from animal studies suggests that NGF ameliorates basal forebrain cholinergic atrophy in models of both aging (Figure 1) and disease, with further evidence that NGF attenuates cognitive deficits. This evidence has emerged from both rodent and primate models. Further, elimination of NGF-producing cells in the hippocampus of intact adult rats by excitotoxic lesions induces atrophy of BFCNs, indicating an ongoing requirement for NGF to support cholinergic neuronal function in the intact brain.

Aging and NGF. With normal aging, cholinergic neurons of the rat and primate brain become atrophic, exhibiting reductions in neuronal size and function. However, there is little if any detectable death of cholinergic neurons as a function of aging, based on 'stereological' methods of quantification that correct for variations in cell size, brain volume, and sampling error when calculating total neuronal number. Yet aged cholinergic neurons are clearly hypofunctional, reflected by shrinkage in cell size and reduction in a variety of biochemical measures, including but not limited to choline acetyltransferase, high-affinity choline uptake, acetylcholine release, and expression of neurotrophin receptors. There is also evidence that age-related alterations in the electrophysiological properties of cholinergic systems occur, including perturbations in the ability of cholinergic systems to synchronize theta activity with activity generated in hippocampal targets. Collectively, there is evidence suggesting that hypofunctionality of the cholinergic system contributes to the extent and severity of cognitive decline with aging.

One of the earliest-described and most striking effects of growth factors in the adult nervous system came with the report of Fischer et al. in 1987 that NGF administration to aged rats reversed cognitive deficits and restored morphological features of cholinergic neurons. NGF was infused intracerebroventricularly into aged rats, and within a 2-week time period the animals exhibited significant improvement in spatial memory, functioning at the level of young rats in the Morris water maze (Figure 2). Anatomical analysis demonstrated a reversal of cholinergic neuronal atrophy and restoration of labeling for choline acetyltransferase to levels equaling those in young animals. This, together with earlier evidence that NGF prevents cholinergic neuronal death in a fimbria-fornix lesion model in young animals, ushered in an era of intense interest in and research on the effects of

Table 1 A partial list of identified growth factors

Family	*Growth factor*	*Preferred receptor(s)*
Classic neurotrophins	Nerve growth factor (NGF)	TrkA, $p75^{NTR}$
	Brain-derived neurotrophic factor (BDNF)	TrkB, $p75^{NTR}$
	Neurotrophin-3 (NT-3)	TrkC, $p75^{NTR}$
	Neurotrophin-4 (NT-4)	TrkB, $p75^{NTR}$
Glial cell-derived	GDNF	Ret, GFRα-1
neurotrophic factor (GDNF)	Neurturin	Ret, GFRα-2
family of ligands (GFLs)	Artemin	Ret, GFRα-3
	Persephin	Ret, GFRα-4
Cytokine family	Ciliary neurotrophic factor (CNTF)	gp130, LIFRβ, CNTFRα
	Leukemia inhibitory factor (LIF)	gp130, LIFRβ
	Cardiotrophin-1 (CT-1)	gp130, LIFRβ, CT1R
	Oncostatin-M (OSM)	gp130, OM1Rα, β
	Interleukin (IL)-6	gp130, IL-6Rα
	Interleukin (IL)-2–7,-9,-11,-12,-15	gp130, ILRs
	Granulocyte-colony stimulating factor (G-CSF)	G-CSFR
	Granulocyte-macrophage colony-stimulating factor (GM-CSF)	GM-CSFRα, β_C
	Leptin	Leptin receptor
Insulin-like growth factor	IGF-1	IGFR-1
(IGF) family	IGF-2	IGFR-1,2
Fibroblast growth factor	FGF 1 (aFGF)	FGFR1–4
(FGF) family	FGF 2 (bFGF)	FGFR1–3
Hepatocye growth factor	HGF	Met
(HGF) family	Macrophage-stimulating protein (MSP)	Ron
Epidermal growth factors (EGFs)	Neuregulin (ARIA)	NEU/ERBB2,3 RTK
	Neuregulin 3	ERBB4 RTK
	Netrin 1,2	Unc/DCC families
	TGF-α	EGFR
	Amphiregulin	EGFR
Others	Bone morphogenic proteins (BMPs)	BMPRs
	Midkine (NEGF2)	Syndecan-1
	Stem cell factor	c-kit
	Platelet-derived growth factor (PDGF)	PDGFR α, β
	Saposin C	?
	Semaphorins	Neuropilin, plexins

growth factors in the adult nervous system. The initial report of Fischer and colleagues was confirmed by several subsequent reports.

Following these findings in aged rodents, Smith et al. in 1999 and Conner et al. in 2001 reported that aged rhesus monkeys also exhibit atrophy in morphological features of cholinergic neurons and that NGF administration restores these markers to levels of young primates. Stereological quantification indicated that, as a function of normal aging, rhesus monkeys exhibit a 40% reduction in the number of cells expressing cholinergic markers and a 10% shrinkage in the remaining cholinergic neurons. Quantification of Nissl-stained sections in the same subjects indicated that cholinergic neurons were atrophic but not dead, paralleling observations of age-related atrophy rather than death of cholinergic neurons in rodent studies. NGF gene delivery to the brains of aged monkeys restored the morphological parameters of cholinergic neurons to levels of young monkeys, including the number of neurons labeled for choline acetyltransferase and the low-affinity neurotrophin receptor p75, and the size of neurons. Conner and colleagues subsequently reported that aged rhesus monkeys also exhibit a 25% reduction in cholinergic terminals innervating cortical target regions, including frontal, temporal, and parietal cortices. NGF gene delivery to the cholinergic somata restored the density of cortical cholinergic axons in the cortex to levels of young monkeys.

Thus, a variety of age-related neurodegenerative effects are sensitive to growth factors in rodent and primate models. However, age-related atrophy of these systems does not appear to be caused by deficient growth factor production, because NGF protein and mRNA levels do not decline in aging.

Other studies provide further support for the trophic effects of NGF on cholinergic systems of the basal forebrain. Hefti, Kromer, Williams, and their colleagues in the mid-1980s separately reported that NGF administered to the brains of adult rats prevented the death of basal forebrain cholinergic neurons after

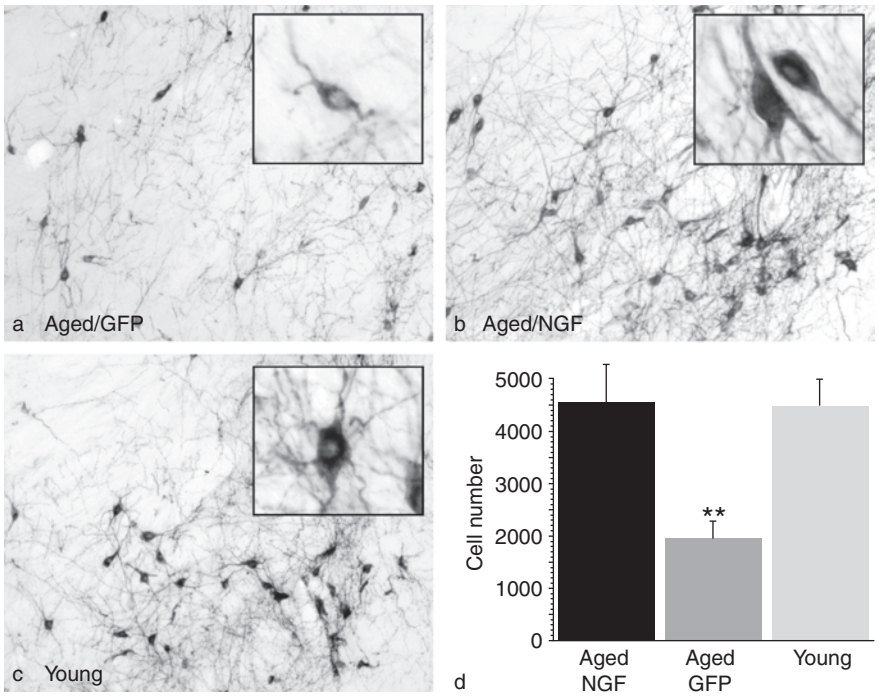

Figure 1 NGF ameliorates age related neuronal atrophy. (a–c) Immunolabling for p75, a marker of cholinergic neurons, in aged-control, aged-NGF-treated, and young rats, respectively. Aged rats show a reduction in functional cholinergic markers, and reversal of age-related atrophy with lentiviral NGF gene transfer. Insets, higher magnification. (d) Quantification of labeled neuron number using stereological methods. Scale bar = 35 μm. Reproduced by permission of A Blesch, Department of Neurosciences, UCSD.

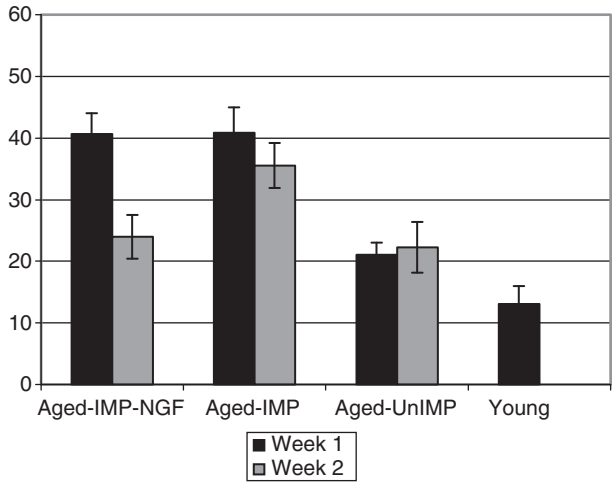

Figure 2 Behavioral performance on the Morris water maze of young, aged-impaired, aged non-impaired, and NGF-treated aged impaired rats (mean +/− s.e.m.). Performance is measured as swim distance to reach the hidden platform. The NGF-treated group alone improved significantly in week 2 compared to week 1 (cross). For week 2, the NGF-treated aged impaired group differed significantly from the control aged impaired group (asterisk). Adapted by permission from Macmillan Publishers Ltd: *Nature* (Fischer W, Wictorin K, Bjorklund A, Williams LR, Varon S, and Gage FH (1987) Amelioration of cholinergic neuron atrophy and spatial memory impairment in aged rats by nerve growth factor. *Nature* 329: 65), copyright (1987).

fimbria-fornix lesions. Whereas this lesion model ordinarily results in the death of 75–90% of cholinergic neurons over a 1-month period, NGF infusions rescue 90–100% of the cells. Tuszynski et al., Koliatosis et al., and Kordower et al., in separate studies, confirmed the neuroprotective effects of NGF after fornix lesions in the brains of adult primates. Excitotoxic mechanisms of injury to cholinergic neurons of the nucleus basalis of Meynert in rats were also prevented by NGF administration. Mobley and colleagues further reported that trisomy 16 mice, a model of Down syndrome in which an extra copy of the amyloid precursor protein (APP) gene is expressed, exhibit spontaneous atrophy of cholinergic neurons. NGF administration reversed cholinergic atrophy in these mice and improved learning and memory. Potential NGF effects have also been explored in transgenic mice either overexpressing APP or expressing mutant forms of APP; however, stereological quantification to date has disclosed no loss of cholinergic neurons in the brains of APP transgenic mice (reflecting the imperfect nature of APP overexpression as a complete model of AD).

Thus, NGF exerts potent effects on the survival and functional state of BFCNs. In AD, the most common human neurodegenerative disorder, basal forebrain cholinergic neuronal degeneration is an early and

prominent feature, likely contributing to cognitive impairment. Although NGF levels do not decline in either the aged or the AD brain, NGF transport from the cortex to the cholinergic basal forebrain is impaired. These findings collectively constitute a rationale for testing the hypothesis that NGF delivery in AD might improve cognition or cognitive decline. This hypothesis has undergone preliminary testing in human clinical trials.

Brain-derived neurotrophic factor (BDNF) and neurotrophin-4 (NT-4) Brain-derived neurotrophic factor (BDNF) and neurotrophin-4 (NT-4) are related growth factors sharing more than 50% amino acid homology that exhibit preferential binding to the tyrosine kinase receptor B (trkB) neurotrophin receptor. During nervous system development, BDNF and NT-4 influence the survival and patterning of several neuronal projections, and homozygous BDNF gene deletion in mice leads to early postnatal death. Notably, BDNF, NT-4, and trkB are extensively expressed in the adult and aged brain, suggesting ongoing roles throughout life in modulating neuronal function. Indeed, the existing evidence supports a direct and important role for these growth factors in modulating neuronal plasticity and function throughout life.

Several studies have examined the role of BDNF in the adult brain. BDNF is produced in most cortical regions and the hippocampus at all stages of life, and is actively trafficked along neuronal circuits in both anterograde and retrograde directions. Thus, BDNF may act as a molecular signal to support neuronal activity. Further, mRNA for BDNF is trafficked to dendrites, suggesting that its short-term availability in regions of synaptic activity may influence brain function. Notably, BDNF application to hippocampal slices can induce long-term potentiation and facilitate neurotransmitter release, confirming a role in modulating synaptic activity. Indeed, recent studies have indicated that BDNF availability is essential in at least two brain circuits to establish learning or experience: the amygdala and dopaminergic projections to cortex that influence addiction.

There are contradictory reports regarding the fate of BDNF in the aged brain. Some studies indicate that BDNF mRNA and protein levels are stable with aging, arguing against a direct role of this growth factor in events leading to age-related neuronal decline. Other studies suggest modest reductions of BDNF levels in the hippocampus or temporal cortex with aging. The question awaits resolution. It is clear, however, that BDNF levels decline in the temporal cortex in AD, potentially contributing to neuronal degeneration. Experiments in progress are exploring the potential protective effects of BDNF in animal models related to AD.

A number of experimental models have also reported neuroprotective effects of BDNF on cell degeneration in the striatum, substantia nigra, and motor neurons. BDNF has also been reported to induce neurogenesis in the adult brain. The possibility exists, therefore, that BDNF augmentation could be of therapeutic benefit in human disease.

Glial Cell-Derived Neurotrophic Factor Family

GDNF and neurturin A second major family of growth factors in the adult CNS is the GDNF family of ligands, or GFLs. This family consists at present of GDNF, neurturin (NTN), artemin, and persephin. GDNF and NTN have been the most extensively studied. Among the populations of neurons influenced by these growth factors are nigral and motor neurons and neurons comprising the parasympathetic nervous system.

The effects of GDNF and NTN on nigral neurons have generated much interest in their potential use as therapeutic agents in the treatment of Parkinson's disease. Following 6-OH-dopamine lesions in rats and 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) lesions in primates, both GDNF and NTN potently prevent nigral neuronal degeneration and promote striatal reinnervation, leading to significant improvement on behavioral assays that assess motor function.

GDNF and NTN also reverse age-related morphological and functional decline of nigrostriatal function. Aged rhesus monkeys exhibit a reduction in cellular and biochemical indices of striatal function, including reductions in levels of dopaminergic transmitters and their metabolites. These age-related atrophic events are paralleled by losses of striatal fluoro-dopa uptake on positron emission tomography (PET) scans. There are varying reports in the literature regarding the functional consequences of age-related nigrostriatal degeneration in primates, with some claims of functional integrity in motor tasks reflecting nigrostriatal functions in aged primates, and other reports of age-related decrements in motor function. Notably, provision of GDNF or NTN to the nigrostriatal system in aged primates enhances biochemical, anatomical, and PET-related measures. These effects are observed whether growth factors are administered by direct protein delivery or via gene therapy methods. Some studies also report benefits of growth factors

on motor performance in aged monkeys. Thus, paralleling observations of NGF on cholinergic systems, GDNF and NTN exhibit a capacity to directly amelioriate atrophic changes in nigrostriatal systems.

Clinical trials of GDNF, NTN, and other growth factors have been performed in Parkinson's disease. Early trials likely failed to deliver growth factors to the nigrostriatal system in sufficient doses to allow detection of potential beneficial effects. More recent trials, currently in progress, will provide the first true test of the hypothesis that growth factor delivery to the striatum will influence the clinical course of Parkinson's disease.

Insulin-Like Growth Factor-1

In the developing nervous system, IGF-1 promotes neurite outgrowth and can influence the survival of several neuronal populations, including primary motor neurons. The receptor for IGF-1 continues to be broadly expressed in the adult nervous system, and neuroprotective effects of IGF-1 have been reported in a variety of neuronal systems in the brain, including the cortex and hippocampus.

There has been a particular focus, however, on the possible use of IGF-1 as a treatment for motor neuronal degeneration that occurs in amyotrophic lateral sclerosis (ALS). In several models of motor neuron atrophy and degeneration, IGF-1 has been reported to protect motor neurons. These models include *in vitro* motor neuronal survival assays, motor nerve transection lesions that induce neuronal atrophy, and motor root avulsion lesions that lead to neuronal death. In all of these models, IGF-1 protects motor neurons. More recently, IGF-1 was reported to promote the survival of spinal cord motor neurons in transgenic mice expressing forms of superoxide dismutase mutation that occur in familial ALS.

A clinical program tested the ability of peripherally administered IGF-1 to alter the course of ALS. This program was in fact carried through two large phase three clinical trials: one trial showed efficacy, but the other did not, and the growth factor was not approved for human use. However, another National Institutes of Health (NIH)-sponsored multicenter controlled trial of IGF-1 in ALS is currently under way to establish definitively whether in fact IGF-1 will benefit the sporadic form of ALS. It was recently reported that gene delivery of IGF-1 to motor neurons utilizing retrograde transport of viral vectors may be a superior means of delivering the growth factor to necessary targets in the spinal cord. Additional preclinical studies are testing this possibility.

Summary

Neuronal atrophy is a feature of aging and a component of progressive neurodegenerative disorders and nervous system trauma. Many growth factors have been demonstrated to ameliorate neuronal atrophy, prevent cell death, and enhance the functional state of the neuron in models of disease. Indeed, neutralization of growth factor production or signaling in the normal brain appears to cause neuronal atrophy. Success in capturing the biological properties of the growth factors to treat human disease has been elusive, however, due to challenges in delivering growth factors into the CNS in a sustainable and safe manner. Nonetheless, such efforts are continuing. For the present, the study of growth factors remains an informative route to further our understanding of the signaling, structure, and function of the nervous system.

See also: Enteric Nervous System: Neurotrophic Factors; Glial Growth Factors; Glutamatergic and Gabaergic Systems; Nerve Growth Factor; Neurotransmitters and Growth Factors: Overview; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Neurotrophins: Physiology and Pharmacology.

Further Reading

Barde YA, Edgar D, and Thoenen H (1982) Purification of a new neurotrophic factor from mammalian brain. *European Molecular Biology Organization* 1: 549–553.

Conner JM, Darracq MA, Roberts J, and Tuszynski MH (2001) Non-tropic actions of neurotrophins: Subcortical NGF gene delivery reverses age-related degeneration of primate cortical cholinergic innervation. *Proceedings of the National Academy of Sciences of the United States of America* 98: 1941–1946.

Conner JM, Lauterborn JC, Yan Q, Gall CM, and Varon S (1997) Distribution of brain-derived neurotrophic factor (BDNF) protein and mRNA in the normal adult rat CNS: Evidence for anterograde axonal transport. *Journal of Neuroscience* 17(7): 2295–2313.

Fischer W, Wictorin K, Bjorklund A, Williams LR, Varon S, and Gage FH (1987) Amelioration of cholinergic neuron atrophy and spatial memory impairment in aged rats by nerve growth factor. *Nature* 329: 65–68.

Hefti F (1986) Nerve growth factor (NGF) promotes survival of septal cholinergic neurons after fimbrial transection. *Journal of Neuroscience* 6: 2155–2162.

Kaspar BK, Llado J, Sherkat N, Rothstein JD, and Gage FH (2003) Retrograde viral delivery of IGF-1 prolongs survival in a mouse ALS model. *Science* 301(5634): 839–842.

Kordower JH, Emborg ME, Bloch J, et al. (2000) Neurodegeneration prevented by lentiviral vector delivery of GDNF in primate models of Parkinson's disease. *Science* 290(5492): 767–773.

Lai EC, Felice KJ, Festoff BW, et al. (1997) Effect of recombinant human insulin-like growth factor-I on progression of ALS. A placebo-controlled study. The North America ALS/IGF-I Study Group. *Neurology* 49(6): 1621–1630.

Oppenheim RW, Houenou LJ, Johnson JE, et al. (1995) Developing motor neurons rescued from programmed and axotomy-induced cell death by GDNF. *Nature* 373: 344–346.

Sofroniew MV, Galletly NP, Isacson O, and Svendsen CN (1990) Survival of adult basal forebrain cholinergic neurons after loss of target neurons. *Science* 247: 338–342.

Tuszynski MH, Thal L, Pay M, et al. (2005) A phase 1 clinical trial of nerve growth factor gene therapy for Alzheimer disease. *Nature Medicine* 11: 551–555.

Yan Q, Matheson Q, and Lopez OT (1995) In vivo neurotrophic effects of GDNF on neonatal and adult facial motor neurons. *Nature* 373: 341–344.

Growth Hormone

L A Frohman, University of Illinois at Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C Carter-Su and K E Mayo © 2004. Elsevier B.V.

Introduction

Growth hormone (GH) is a single-chain protein of 191 amino acids with a molecular size of approximately 22 kDa that is synthesized by and secreted from the somatotroph cells of the anterior pituitary gland. GH exerts a broad spectrum of effects, leading to the promotion of growth and the regulation of carbohydrate, lipid, protein, and mineral metabolism. GH synthesis and secretion are regulated in a complex manner, primarily through the actions of hypothalamic neuropeptides that integrate hormonal, metabolic, and neurogenic signals. Alterations in GH secretion lead to disorders of growth and metabolism. This article addresses the identification and structure of this critical hormone, its expression and regulation in the pituitary gland, its mechanisms of action, both peripherally and within the central nervous system, diseases of GH secretion, and the therapeutic use of GH.

Identification and Structure of Growth Hormone

The first suggestion that the pituitary was a site of synthesis of a growth-promoting substance resulted from the observations of Marie in 1886 of pituitary gland tumors associated with acromegaly. Following this lead, Evans and Long directly demonstrated the ability of pituitary extracts to augment growth in the dog and rat, leading to the development of a rat tibial bioassay used to purify GH from the bovine pituitary. Determination of the primary structure of human GH revealed a protein that is similar in sequence to human chorionic somatomammotropin, also known as placental lactogen (Figure 1).

The primary structure of GH from numerous species was rapidly deduced through complementary deoxyribonucleic acid (cDNA) cloning techniques, beginning with the characterization of rat pituitary GH cDNA. The subsequent sequencing of human GH led to the production of recombinant hormone used for treating GH deficiency, and the sequencing of bovine GH has led to its use for enhancing milk production in cattle. Molecular cloning also led to the identification of the structure of the GH gene, which contains five exons separated by four introns. The human GH gene transcript undergoes alternative ribonucleic acid (RNA) processing to encode either the 22 kDa polypeptide or a shorter 20 kDa species, referred to as variant GH, which constitutes up to 20% of pituitary GH and may have altered biologic properties. The human GH gene, designated GH1 and which encodes growth hormone normal (GH-N), is part of a cluster of five related genes found in a 50 kb locus on chromosome 17. The locus also includes the genes for the chorionic somatomammotropins (CSH1, CSH2, and CSHP1) and for growth hormone variant (GH-V) (the GH2 gene), a form of GH expressed only in the placenta that becomes the predominant GH form during the latter half of pregnancy. GH-V differs from GH-N at 13 amino acid positions and exhibits most or all the biologic effects of GH-N, though its significance in fetal development is not well understood.

The crystallographic structure of GH was determined by radiographic diffraction techniques using genetically engineered porcine GH. The molecule consists of four antiparallel alpha helices with a unique connectivity, arranged in a left-twisted bundle. The other GH-related proteins are believed to share this same basic structure.

Expression and Regulation of GH

The GH gene is expressed in the pituitary gland during fetal development. Although GH gene expression in the rat is not detected until day 18 of gestation, in the human, serum GH is measurable by the end of the first trimester. GH is expressed exclusively in somatotrophs and, together with prolactin, in mammosomatotrophs. These cell populations vary depending on the species and the physiologic state and have been proposed to interconvert in the adult pituitary by transdifferentiation. The specific expression of GH in somatotrophs or mammosomatotrophs derives, in part, from the selective expression of the transcription factor Pit-1, a POU–domain protein found in somatotrophs, lactotrophs, and thyrotrophs. Pit-1 is required for transcription of the GH gene, where it appears to act together with a zinc finger protein, Zn-15. Mutations that disrupt the Pit-1 gene in the Snell or Jackson dwarf strains of mice and in some human patients with combined pituitary hormone deficiency result not only in a loss of GH expression but also in a failure of pituitary cells expressing Pit-1 to differentiate appropriately, resulting in pituitary hypoplasia. Another nuclear transcription factor, Prop1, is required to activate Pit-1 expression. Whereas Pit-1 mutations are rare, Prop1 mutations are the most common cause of multiple pituitary hormone deficiency in humans. The phenotype shows considerable variability but usually includes severe hypopituitarism owing to the lack of cells synthesizing

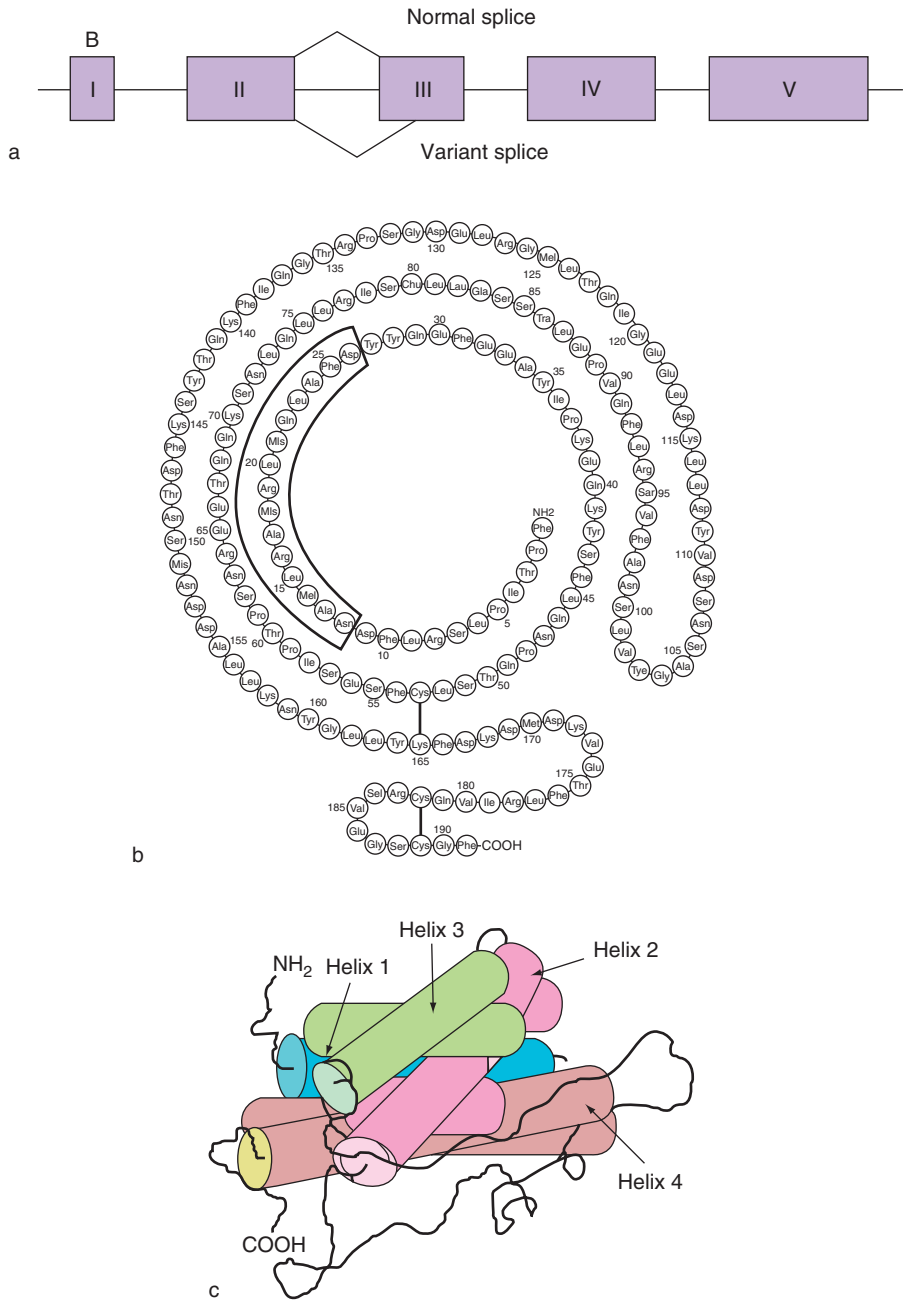

Figure 1 Structure of human growth hormone (GH) and its gene. Panel (a) shows the structure of the human GH gene, which includes five exons (I–V) and four introns. Alternative RNA processing generates the messenger RNA encoding the variant 20 kDa GH, as indicated. Panel (b) shows the primary sequence of human GH. The boxed residues are deleted from the 20 kDa variant GH. Panel (c) is a schematic representation of the crystallographic structure of porcine GH, a four-helix-bundle protein. Adapted from Abdel-Meguid SS, Shieh HS, Smith WW, et al. (1987) Three-dimensional structure of a genetically engineered variant of porcine growth hormone. *Proteomics* 84: 6434–6437.

GH, prolactin, and thyroid-stimulating hormone. Patients with Prop1 mutations may exhibit, in addition, gonadotropin deficiency. An animal model of Prop1 deficiency, the Ames dwarf mouse, has also been described. Other transcription factor mutations also cause GH and/or combined pituitary hormone deficiencies as a result of impaired pituitary gland development during fetal development.

Regulation of GH synthesis and secretion occurs through the integrated action of numerous peptide and nonpeptide hormones. The neuroendocrine control of GH is mediated predominantly by the hypophysiotropic hormones: GH-releasing hormone (GHRH) and somatostatin (Figure 2). These neuropeptides interact with specific G-protein-coupled receptors expressed by the somatotrophs, leading to activation

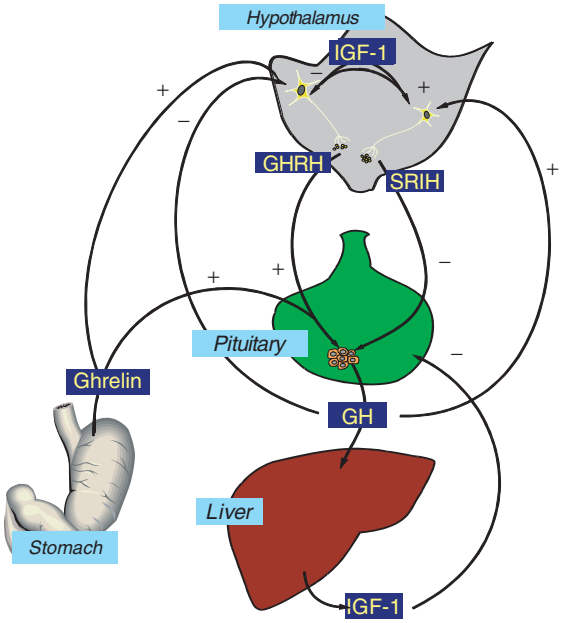

Figure 2 Neuroendocrine regulation of growth hormone (GH) secretion. GH-releasing hormone (GHRH) and somatostatin effects on GH synthesis and secretion are indicated. In addition to modulating proliferation, differentiation, and metabolic events in target cells, GH induces synthesis of insulin-like growth factor 1 (IGF-1), important for mediating many effects of GH on body growth and for negative feedback regulation, indicated by the arrows. Ghrelin also stimulates GH secretion at both pituitary and hypothalamic levels, though its role in physiologic regulation of the hormone is unclear.

(GHRH) or suppression (somatostatin) of adenylate cyclase and a corresponding change in intracellular cyclic adenosine monophosphate (cAMP) levels and an alteration in intracellular Ca^{2+}. GHRH stimulates GH synthesis and GH secretion, and chronic GHRH exposure results in somatotroph proliferation. Somatostatin exerts its effects predominantly at the level of GH secretion but also attenuates GHRH-induced GH gene expression and somatotroph proliferation. Many additional peptide hormones are reported to have modulatory effects on GH synthesis or secretion, including galanin, neuropeptide Y, and thyrotropin-releasing hormone, though their physiologic relevance is quite limited.

Potent synthetic GH secretagogues (GHSs) have been designed, initially based on a synthetic enkephalin-related hexapeptide (GHRP) and more recently on nonpeptide mimetics of GHRP. The GHRPs bind to the GHS receptor and appear to act at both the hypothalamic and the pituitary level. The endogenous ligand of the GHS receptor, ghrelin, has recently been identified and, like the GHRPs, potently stimulates acute GH release through both hypothalamic and pituitary actions. The stomach is the major site of ghrelin production, and its secretion is regulated by food intake. It may have a role in mediating the nutrient regulation of GH secretion although it does not appear to exert a role in other neuroendocrine control mechanisms for GH.

Nonpeptide hormones that modulate GH synthesis include glucocorticoids, which increase both GH gene transcription and messenger RNA (mRNA) stability, and thyroid hormones, which exert principally stimulatory actions on GH expression, although suppressive effects of thyroid hormones have been observed in experiments with the human GH gene.

GH secretion from the pituitary is pulsatile, with 8–12 peaks during a 24 h period, and the pulsatility is believed to be important for the physiologic actions of the hormone. GH secretion is augmented during sleep, by exercise, and during puberty, whereas it progressively decreases with aging. Negative feedback effects constitute a major component of GH secretory regulation. GH, GHRH, and somatostatin are all subject to feedback regulation, as is insulin-like growth factor 1 (IGF-1), the synthesis of which is induced in numerous target tissues in response to GH. IGF-1 exerts feedback effects at both the pituitary and the hypothalamic levels by suppressing GH and GHRH secretion and stimulating somatostatin secretion while GH feedback effects occur in the hypothalamus by stimulating somatostatin and inhibiting GHRH secretion.

Mechanism of Action of GH

GH exhibits a broad range of biologic actions which promote cellular differentiation and proliferation leading to increased growth and also exert non-growth-related metabolic effects. Many of the rapid actions of GH are believed to be direct, but a significant portion of its effects on body growth are mediated by IGF-1. GH induces IGF-1 generation in response to GH in numerous tissues including the liver, which is a major target of GH action. Circulating IGF-1 is derived primarily from hepatic secretion and, together with locally produced IGF-1 acting in a paracrine or autocrine manner, exerts a role in body growth.

Most GH in circulation is bound to a GH-binding protein (GH-BP), the N-terminus of which is identical to the cell surface GH receptor (GH-R), through which GH exerts its actions. Cloning of the common cDNA for GH-BP and GH-R revealed the receptor to be a protein of 638 amino acids that spans the plasma membrane, whereas the binding protein represents the extracellular domain of this receptor. The soluble binding protein is released from the mature receptor by proteolysis in some species (rabbit and human) and is encoded by a distinct mRNA

generated through alternative RNA splicing of the GH-R transcript in other species (rat and mouse). The radiographic crystallographic structure of the extracellular domain of the GH-R bound to GH consists of one molecule of GH interacting with two receptor molecules. Distinct domains within the GH molecule interact with unique sites in the two receptor molecules to form this complex.

The GH-R is a member of the cytokine receptor superfamily. The extracellular domains of these receptors include two pairs of conserved cysteine residues. The cytoplasmic domains are quite divergent but do share a proline-rich motif near the transmembrane domain. The initial step in GH signal transduction involves the binding of GH to its receptor followed by receptor dimerization. Janus kinase 2 (JAK2), a cellular protein tyrosine kinase that binds to the dimerized GH-R via the receptor's proline-rich motif, is activated in response to GH treatment. The activated JAK2 undergoes autophosphorylation and also phosphorylates the associated GH receptor on tyrosines which, in turn, recruit a variety of signaling proteins, many of which are then phosphorylated by JAK2. These signaling proteins include a group of cytoplasmic proteins – signal transducers and activators of transcription (STATs) – which translocate to the nucleus and regulate transcription, thereby mediating the actions of GH on gene expression. Among these, STAT5-b appears to exert a critical mediating role. The mitogen-activated protein kinases are also activated in response to GH, and roles for phospholipase C, protein kinase C, insulin receptor substrate/phosphatidyl inositol 3′ kinase/AKT pathways in GH signaling have been suggested (Figure 3). Additional adapter proteins (e.g., grb10, SH2-B), proteins that regulate the actin cytoskeleton and cell motility, and proteins in the JNK pathway have also been implicated in GH signaling.

Mutations of the GH receptor gene (Laron syndrome) or in genes encoding downstream signal transduction proteins (e.g., STAT5-b) lead to a GH insensitivity syndrome that consists of elevated GH levels, reduced IGF-1 concentrations, and impaired linear growth.

In addition to its growth-promoting effects, GH exhibits many effects on metabolic processes. Effects

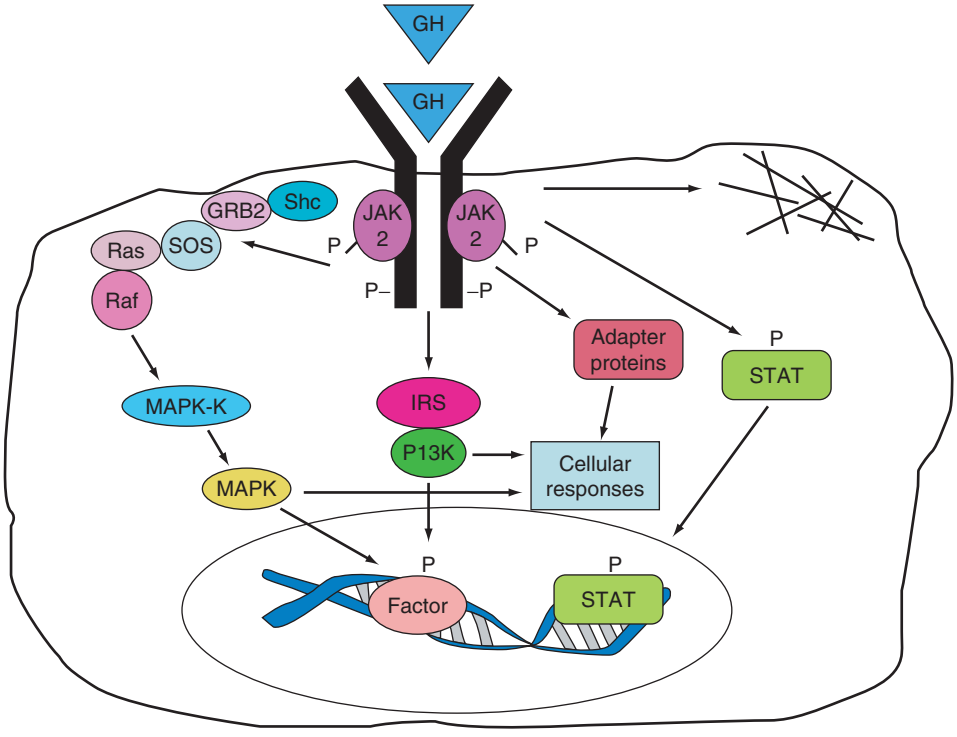

Figure 3 Signal transduction by growth hormone (GH) and its receptor. A single GH molecule binds through distinct sites on the hormone to two receptor molecules. The Janus kinase family member Janus kinase 2 (JAK2) that associates with the dimerized GH receptor is activated by GH binding, resulting in tyrosine phosphorylation of both JAK2 and the GH receptor. Activated JAK2 phosphorylates members of the signal transducers and activators of transcription (STAT) family of proteins, which then move to the nucleus and regulate the transcription of target genes. GH activates the mitogen-activated protein kinase (MAPK) pathway, perhaps through interaction of the phosphorylated receptorJAK2 complex with Shc. Insulin receptor substrates (IRS) and downstream signaling proteins including phosphatidyl inositol 3′kinase (PI3K) are also recruited into GH signaling pathways, as are a variety of other proteins, including adapter proteins. These pathways mediate GH regulation of gene transcription, the cytoskeleton, and other cellular responses.

on carbohydrate metabolism include both an enhancement of insulin release and an insulin-antagonistic effect, mediated in part by the hormone's marked lipolytic activity. The insulin resistance is responsible for the diabetogenic effects of the hormone. GH enhances amino acid uptake and protein synthesis in GH-responsive tissues such as muscle, liver, and kidney and is important in tissue hypertrophy. GH effects on mineral metabolism include enhancement of both bone turnover and bone mineral deposition and are important in bone growth and maintenance of normal bone integrity.

GH Effects in the Central Nervous System

Small amounts of GH are produced in several different brain regions, and there appear to be mechanisms for GH transport across the capillary bed-blood–brain barrier. GH exerts trophic effects within the brain in both embryonic and postnatal development and may also have some neuroprotective effects. However, many, if not all, of the effects of GH may be mediated by IGF-1 that is produced locally.

Disorders of GH Secretion and Therapeutic Uses of GH

Short stature resulting from GH deficiency occurs with a frequency ranging from 1 in 4000 to 1 in 10 000 births. Some cases are associated with midline developmental defects, and up to 30% may have a genetic basis, though the cause remains unknown in the majority of cases. The familial isolated GH deficiencies (IGHDs) include at least four distinct disorders. IGHD-1A is the most severe and is associated with deletions or inactivating mutations of the GH gene. In IGHD-IB, GH levels are low but detectable. Some cases appear to involve more-subtle mutations at the GH1 locus, but many cases are likely to involve mutations in other genes, such as the GHRH receptor gene. Both IGHD-IA and -IB exhibit autosomal recessive inheritance. IGHD-II patients have clinical characteristics similar to those with IGHD-IB but exhibit an autosomal dominant mode of inheritance. IGHD-II is most commonly caused by mutations in the third intron of the GH gene, leading to a splicing defect and loss of exon 3, though other missense mutations have also been observed. IGHD-III is an X-linked recessive disease that is clinically diverse and may represent mutations at multiple loci. Animal models of GH deficiency include the Snell and Jackson dwarf mice, which result from mutations in the transcription factor Pit-1 and represent a loss of the thyrotroph, somatotroph, and lactotroph lineages; the phenotypically similar Ames dwarf mouse, which is caused by a mutation in the Prop1 transcription factor gene; the Little mouse, which is a mutation in the pituitary receptor for GHRH and a model for IGHD-IB; the spontaneous dwarf rat, which results from abnormal splicing of the GH mRNA; and the Lewis dwarf rat, which most likely represents a defect in the GHRH signal transduction pathway.

Hypersecretion of GH results in gigantism when it occurs prior to closure of the long bone epiphysial growth plates, or acromegaly when it occurs afterward, and is most commonly associated with a GH-secreting pituitary tumor. Strong evidence exists for clonality of these somatotroph tumors. As many as 40% of such tumors are associated with activating mutations in the stimulatory G-protein alpha-subunit in somatotroph cells that result in the constitutive presence of a cAMP signal normally modulated by the actions of GHRH and somatostatin. A small minority of cases of acromegaly result from the hypersecretion of GHRH by a variety of endocrine tumor types and, rarely, by ectopic GH secretion. Animal models for GH hypersecretion include transgenic mice that grow to be twice normal size, resulting from transgenic overexpression of the GH. Transgenic mice that overexpress GHRH serve as a model for tumors with ectopic GHRH secretion.

GH therapy has been extensively used in the treatment of short stature associated with GH deficiency, particularly after the availability of recombinant human GH in 1986. Treatment with recombinant human GH produces few side effects. However, when administered in excess, GH causes fluid retention, swelling of hands and feet, arthralgias, and neural entrapment (carpal tunnel syndrome) similar to the early signs and symptoms of acromegaly. With the increased availability of GH, its use as a therapeutic agent has increased dramatically in the past decade. Recombinant human GH has proven an effective means of stimulating growth in children with primary GH deficiency or growth failure associated with Turner's syndrome, chronic renal failure, Prader-Willi syndrome, or intrauterine growth retardation. More recently, it has been used to stimulate growth in children with idiopathic short stature, unassociated with GH deficiency, who are presumed to have varying degrees of GH insensitivity.

In addition to its uses in the clinical management of growth deficiencies in children, GH is used for treatment of GH-deficient adults and has been evaluated in postsurgical and burn patients and in patients with AIDS, among others. The use of GH therapy in GH-deficient adults is controversial. In most studies, treatment has been reported to result in increases in lean body mass and bone density, decreases in body fat and serum cholesterol, and in some, an increased

quality of life. However, the increases in lean body mass do not appear to lead to increased muscle function. Some of these effects of GH therapy have also been reported in normal aging adults, where a progressive reduction in endogenous GH secretion occurs. However, a relatively high incidence of adverse treatment effects has limited overall enthusiasm in this population, and at present, the use of GH in aging must still be considered as investigational. Evidence that GH therapy restores the reduced life expectancy of hypopituitary patients to normal is lacking, and observations of mice with genetic models of GH deficiency (described earlier) as well as humans with genetic disorders of GH deficiency have found no shortening of life expectancy or premature atherosclerosis.

See also: Growth Factors: Neuronal Atrophy; Hypothalamo–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH; Insulin-Like Growth Factor Signaling and Actions in Brain; Neuroendocrine Aging: Pituitary Metabolism; Neuroendocrinology; Neurotransmitters and Growth Factors: Overview; Pituitary Gland (Cell Types, Mediators, Development).

Further Reading

Baumann G (1991) Growth hormone heterogeneity: Genes, isohormones, variants, and binding proteins. *Endocrine Reviews* 12: 424–449.

Cooke NE and Liebhaber SA (1995) Molecular biology of the growth hormone-prolactin gene system. *Vitamins and Hormones* 50: 385–459.

Frohman LA, Downs TR, and Chomczynski P (1992) Regulation of growth hormone secretion. *Frontiers in Neuroendocrinology* 13: 344–405.

Mayo KE, Miller T, DeAlmeida V, et al. (2000) Regulation of the pituitary somatotroph cell by GHRH and its receptor. *Recent Progress in Hormone Research* 55: 237–266.

Phillips JA III and Cogan JD (1994) Genetic basis of endocrine disease. 6. Molecular basis of familial human growth hormone deficiency. *Journal of Clinical Endocrinology and Metabolism* 78: 11–16.

Piwien-Pilipuk G, Huo JS, and Schwartz J (2002) Growth hormone signal transduction. *Journal of Pediatric Endocrinology and Metabolism* 15(6): 771–786.

Smith RG (2005) Development of growth hormone secretagogues. *Endocrine Reviews* 26: 346–360.

Veldhuis JD and Bowers CY (2003) Human GH pulsatility: An ensemble property regulated by age and gender. *Journal of Endocrinological Investigation* 26: 799–813.

Hair Cell Differentiation

C Puligilla and M W Kelley, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

The primary transducers for the sensations of hearing, balance, and acceleration are mechanosensory hair cells located within the auditory and vestibular regions of the inner ear. The auditory region comprises the spiral-shaped cochlea, which contains the organ of Corti, an elongated sensory epithelium with approximately 20 000 mechanosensory hair cells arranged into four or five distinct rows (Figure 1(a)). The vestibular system comprises three semicircular canals, each with associated ampullae and two saclike structures, the utricle and the saccule. The ampullae of each semicircular canal contain a sensory crista that detects angular acceleration, and the utricle and saccule each contain a macula for the detection of linear acceleration (Figure 1(b)). Regardless of location within the ear, all sensory epithelia are composed of two basic cell types, mechanosensory hair cells and nonsensory supporting cells, arranged in a pseudostratified configuration with hair cells located adjacent to the lumenal surface and single or multiple layers of supporting cells located between the hair cell layer and the basement membrane (Figure 1(c)). Because the epithelium is pseudostratified, all supporting cells extend processes that terminate at the lumenal surface. Mechanosensory hair cells derive their name from a group of modified group of microvilli, referred to as stereocilia, that extend from the lumenal surface of each cell. The group of stereocilia appears similar to a tuft of hair. Deflection of the stereociliary bundle in response to fluid movement in the lumenal space results in a depolarization of the resting potential and changes in the rate of neurotransmitter release.

Prosensory Specification

All hair cells are derived from one or two specialized regions, referred to as prosensory patches, which arise in the ventromedial wall of the bilateral otocysts, placodally derived structures that originate adjacent to the hindbrain. With continued development, the otocysts undergo a number of morphogenetic changes to give rise to all of the auditory and vestibular structures of the inner ear. At the same time, cells from these prosensory patches become distributed throughout the ear to form the various sensory epithelia just described. The factors that specify the prosensory patches have not been conclusively identified, but several genes that are known to regulate cellular identity in other systems are expressed at early times in prosensory patch development. These include *Jagged1, Sox2, Lunatic fringe (Lfng), Pax2, Bmp4,* and *FGF10* (Figure 2). Functional data concerning the roles of some of these factors are discussed next.

Role of Notch Signaling in Prosensory Specification

The Notch signaling pathway has been highly conserved throughout evolution and has been shown to regulate multiple events in embryonic development. In mammals, four Notch genes code for membrane-bound receptors that are activated on binding with any one of a family of membrane ligands that includes Jagged1, Jagged2, Delta1, and Delta3. Ligand binding results in the cleavage of the Notch protein to yield an intracellular component (Notch-ICD) that is then translocated to the nucleus. Notch1 is expressed throughout the developing inner ear, beginning at the early otic placode stage and persisting throughout development. Jagged1 expression begins about the time of placodal invagination and, although initially diffuse in expression, rapidly resolves to the developing prosensory patches. Historically, the primary role of Notch signaling has been considered to be to mediate lateral inhibitory interactions between cells. However, more recent data suggest that Notch signaling can have a positive, inductive effect on cell fate. The early expression of Notch1 and one of its ligands within developing prosensory patches is consistent with a role in the specification of these patches. And, as would be predicted, the deletion of *Jagged1* within the otocyst leads to a nearly complete loss of hair cells. However, an equivalent deletion of *Notch1* does not result in a similar phenotype. In fact, the loss of *Notch1* results, instead, in an overproduction of hair cells. The bases for these differences are only partially understood, but they strongly suggest that Jagged1 acts through the activation of another Notch gene. The potential role of Notch signaling in the specification of prosensory patches was recently supported by the demonstration that the transient overexpression of a constitutively active Notch1-ICD construct in chick otocysts results in the induction of ectopic sensory patches. Furthermore, the expression of serrate (the chick Jagged1 homolog) was upregulated within ectopic sensory patches, suggesting the existence of a positive feedback loop mediated through Jagged1/Notch activity leading to a lateral induction process.

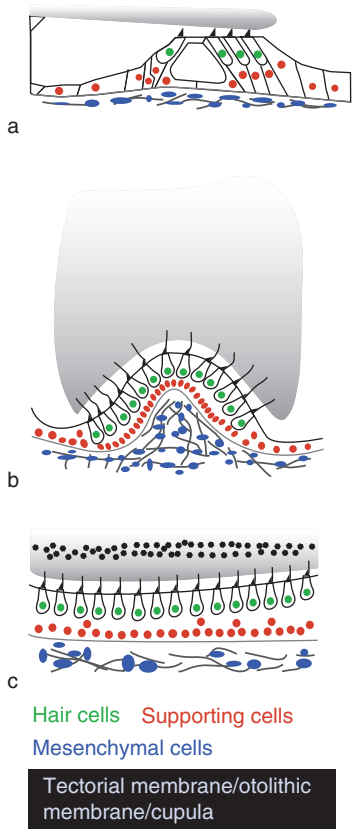

Figure 1 Diagrammatic cross-sections through the sensory organs of the mammalian inner ear: (a) the organ of Corti, the sensory epithelium of the mammalian cochlea; (b) a crista ampullaris, the sensory epithelium within a semicircular canal; (c) a vestibular sensory epithelium such as the sacculus or utriculus. In (a), the sensory epithelium of the organ of Corti is composed of two basic cell types, hair cells (green nuclei) and supporting cells (red nuclei) that lie on a basement membrane referred to as the basal lamina. Beneath the basal lamina is a layer of mesenchymal cells (blue nuclei). Hair cells have a specialized stereociliary bundle (black) on their apical surface that comes in contact with an overlying extracellular matrix called the tectorial membrane (gray). The displacement of this extracellular matrix relative to the hair cells provides the primary stimulus for mechanotransduction through deflection of the stereociliary bundle. In (b), as in the cochlea, the crista ampullaris is composed of hair cells and supporting cells. The hair cell stereocilia come in contact with an overlying extracellular matrix that is similar in composition to the tectorial membrane; it is referred to as a cupula. In (c), hair cells, supporting cells, and mesenchymal cells are present in a pattern similar to that in other hair cell sensory epithelia. Stereocilia project into an overlying extracellular matrix, the otoconial membrane, which also contains individual calcium carbonate-based otoconia (black hexagons).

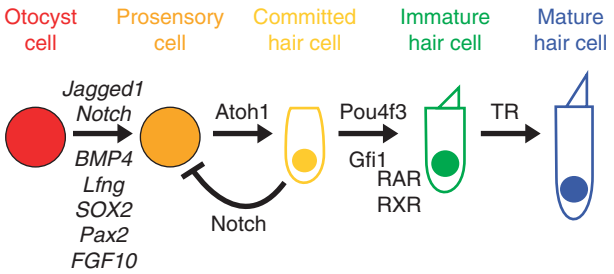

Figure 2 Summary of signaling molecules that are known to mediate hair cell commitment and differentiation. In the first step in hair cell formation, cells within the otocyst become specified to develop as prosensory cells. Although the factors that mediate this transition are not fully understood, a number of genes, including *Jagged1, Notch, Bmp4, Lnfg, Sox2, Pax2,* and *Fgf10,* are expressed in patterns that are consistent with a role in prosensory specification. Next, prosensory cells become committed to develop as hair cells through the expression of *Atoh1*. However, subsequent cell–cell interactions divert some prosensory cells from the hair cell fate through Notch-mediated lateral inhibitory interactions. During the first step of hair cell differentiation, committed hair cells begin to express Pou4f3 and Gfi1. In addition, early hair cell differentiation requires the activation of retinoic acid receptors (RAR and RXR). Finally, later differentiation requires signaling through thyroid hormone receptors (TR).

SOX2

The SRY-related transcription factor SOX2 is also expressed from the early otocyst stage, and its pattern of expression overlaps with Jagged1. However, expression of Sox2 is downregulated in *Jagged1* null mutants, suggesting that it is downstream of Notch signaling. Deletion of *Sox2* specifically within the developing inner ear results in a complete loss of hair cells and hair cell precursors, demonstrating the importance of this gene in hair cell development. A hypomorphic *Sox2* mutant with limited expression within the inner ear shows a partial loss of the prosensory patches, indicating a gene dosage effect.

BMP4

The secreted morphogen BMP4 is considered to be the earliest and most definitive marker for prosensory patches, even though it is expressed in all developing prosensory patches in birds but excluded from the saccular, utricular, and cochlear sensory patches in mammals. Consistent with a role in specification of the prosensory patch, treatment with BMP4 protein *in vitro* results in increased hair cell formation in both chicken otocyst cultures and mammalian cochlear cultures, whereas inhibition of BMP4 signaling using Noggin results in a decrease in the number of hair cells. However, it is important to consider that an additional study actually obtained contradictory results for the role of BMP4 and Noggin in chick otocysts, suggesting that the role of BMP4 may not be straightforward.

Pax2, *Gata-3*, and *Lfng*

As noted, several other molecules with known developmental roles, such as *Pax2*, *Gata3*, and *Lfng* are also expressed in patterns that are consistent with a role in prosensory specification. However, prosensory

patches and hair cells still form in mice with specific deletions in each of these genes, strongly suggesting either that these factors are dispensable for hair cell formation or that compensatory mechanisms exist either normally or in response to loss of gene function.

Exit from the Cell Cycle

Following specification of prosensory patches, individual progenitor cells must become specified to develop as either hair cells or supporting cells. The results of a number of studies have demonstrated a strong link between cell cycle exit and the onset of cellular commitment and differentiation. Within the inner ears of mammals, cell cycle exit occurs in precise patterns that are specific for each sensory epithelium. In the mouse cochlea, cell cycle exit occurs in a gradient that begins in the apex around E12.5 and proceeds toward the base over the next 72 h. In the vestibular epithelia, cell cycle exit typically occurs in a central to peripheral gradient in which cells in the center of the patch are the first to become postmitotic. Cell cycle exit in the inner ear is controlled through the cyclin-dependent kinase (Cdk) inhibitors $p27^{kip1}$ and $p19^{ink4dA}$ and the retinoblastoma tumor suppressor gene pRb, all of which act as negative regulators of proliferation. pRb appears to act as a primary regulator of cell cycle based on a pattern of expression that accurately precedes the pattern of terminal mitosis (except in the cochlea) and the fact that deletion throughout the otocyst results in ongoing proliferation of sensory progenitor cells and an overproduction of hair cells and supporting cells. In the cochlea, the pattern of expression of pRb does not match either the spatial or temporal gradient of terminal mitoses. Instead, cell cycle exit is preceded by expression of $p27^{kip1}$. Moreover, in *p27kip1* null mice, cell cycle exit is delayed, allowing one or two additional rounds of mitosis, and the pattern of terminal mitosis is altered to follow the pattern of expression of pRb. Finally, deletion of *p19Ink4d*, which is also expressed in hair cells, results in postnatal hair cells reentering the cell cycle, suggesting that hair cells rely on $p19^{Ink4d}$ to maintain a postmitotic state.

Commitment of Hair Cells

As discussed, cellular commitment typically follows closely behind terminal mitosis. Therefore, once cells within the prosensory patches exit the cell cycle, individual progenitor cells must be specified to develop as either hair cells or supporting cells. Based on morphological characteristics and expression of cell-specific markers, hair cells become specified first. The earliest indication of hair cell commitment is the expression of the basic helix–loop–helix transcription factor Atoh1 (formerly, Math1) (Figure 3). Atoh1 is initially expressed in a group of cells that appears to include progenitors that ultimately develop as both hair cells and supporting cells. However, as development continues, the expression of Atoh1 becomes restricted to cells that will develop as hair cells. The role of Atoh1 in hair cell development has been examined in a number of different studies. First, targeted deletion of *Atoh1* in a mouse model leads to a complete lack of hair cells in the both the auditory and vestibular sensory epithelia. In addition, the forced expression of Atoh1 in the developing cochlear sensory epithelium is sufficient to induce hair cell formation. And, surprisingly, forced expression of Atoh1 outside of the prosensory patches also leads to hair cell formation. The ability of Atoh1 to induce hair cell formation appears to persist even into the adult because the introduction of an *Atoh1*-expressing adenovirus into the ears of adult guinea pigs results in the formation of ectopic hair cells.

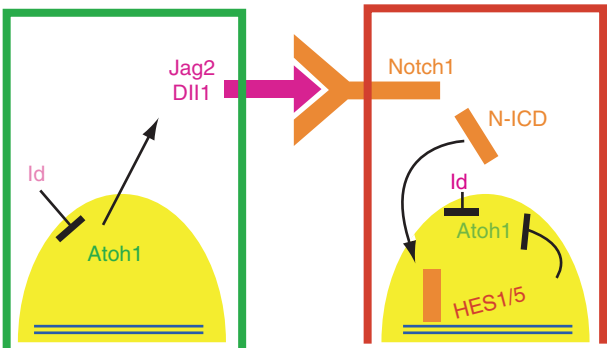

Figure 3 Regulation of hair cell fate through the Notch pathway. Prior to hair cell commitment, cells express Atoh1 along with members of the Id family. In response to downregulation of Ids, some cells are able to increase the expression of Atoh1, leading to the activation of Notch ligands, Jag2 and Dll1. Activated Jag2 and Dll1 bind to Notch1 receptors, leading to the generation of an intracellular fragment of Notch1, referred to as Notch-ICD (N-ICD). N-ICD is translocated to the nucleus, where it activates the expression of the inhibitory bHLH genes, *HES1* and *HES5*. HES1 and HES5 combine with Ids to downregulate the expression of Atoh1, leading to the inhibition of the hair cell fate.

As discussed, Atoh1 is apparently initially expressed in a larger number of cells than ultimately develop as hair cells. This observation suggests that some of the cells that initially become committed to a hair cell fate are subsequently diverted from that fate through cell–cell signaling. In all inner ear sensory epithelia, hair cells and supporting cells are arranged in a mosaic such that hair cells are completely surrounded by supporting cells. However, laser ablation of developing hair cells in either the embryonic mammalian cochlea or in mature amphibian lateral-line

neuromasts results in a change in the fate of neighboring cells, leading to a regenerative response. Based on the mosaic pattern and the observed effects of the removal of existing hair cells, a lateral-inhibitory mechanism was initially proposed to regulate the generation of the cellular mosaic. As previously discussed, the Notch signaling pathway was characterized based on its ability to mediate lateral inhibitory interactions. Therefore, a second role for Notch signaling during inner ear development seemed likely. Consistent with this hypothesis, developing hair cells in different species express one or more Notch ligands, including *Jagged2* and *Delta1*, whereas surrounding supporting cells express *Notch1*. In addition, supporting cells also express the Notch target genes *HES1* and *HES5* Therefore, the patterns of expression for all these genes are consistent with an inhibitory interaction in which developing hair cells use Jagged2 and Delta1 to activate Notch1, and subsequently HES1 and/or HES5, in neighboring cells with the effect of that activation being to inhibit those cells from developing as hair cells. As would be expected based on this hypothesis, the deletion of any of the genes in this pathway leads to an overproduction of hair cells. Similarly, the disruption of the Notch signaling pathway in zebra fish as a result of mutations in the *deltaa* gene or in *mindbomb*, a gene that encodes an E3 ubiquitin ligase that is responsible for targeting Notch for degradation, also leads to an increase in the number of hair cells with a concomitant decrease in the number of supporting cells. In addition, recent work has demonstrated that Atoh1 is a target of Notch1 in the ear and that overexpression of HES1 is sufficient to inhibit the ability of Atoh1 to induce hair cell formation. These results strongly suggest that developing hair cells directly inhibit their neighbors from developing as hair cells through a Notch-mediated inhibition of Atoh1.

One intriguing aspect of the role of Notch signaling in hair cell development is the observation that Jagged2 and Delta1 are not expressed in all Atoh1-positive cells. In many systems in which Notch signaling is used to sort cells into multiple fates, Notch and one or more Notch ligands are initially expressed throughout the population of progenitor cells with subsequent cell–cell interactions leading to some cells increasing the expression of Notch while other cells increase the expression of the Notch ligand. However, in the ear, only developing hair cells appear to express Notch ligands, suggesting that these cells may be biased toward the hair cell fate. Recently, a group of molecules referred to as inhibitors of differentiation (Ids) has been shown to play a role in this bias. Ids are helix–loop–helix molecules that are similar to bHLH molecules, except that they lack the basic domain that is required for DNA binding. As a result, Ids act as negative regulators of bHLH molecules, such as *Atoh1*, through competition for a common dimerization partner. The Id genes, *Id1*, *Id2*, and *Id3*, are expressed in the developing cochlea, but become specifically downregulated in developing hair cells. Moreover, prolonged expression of Id3 in individual cells within the ear leads to the inhibition of hair cell formation. Based on these results, it seems likely that downregulation of Id expression plays key step in the determination of which, and how many, cells within the ear develop as hair cells.

It is also important to consider that, with the exclusion of mammals, all vertebrates classes are able to generate new hair cells throughout the life of the organism, either as a result of normal turnover or as part of a regenerative response. In either case, the generation of new cells requires the specification of progenitors as hair cells. In most cases, the specification of new hair cells in adult animals involves existing cells within the epithelium, presumably supporting cells, reentering the cell cycle to divide and give rise to progenitor cells that then become specified as hair cells or supporting cells through a recapitulation of the developmental processes already described. However, there is also evidence indicating that under some circumstances supporting cells can develop as hair cells through a direct transformation referred to as transdifferentiation. During transdifferentation, supporting cells are presumed to downregulate supporting cell-specific genes and then to upregulate hair cell-specific factors, starting with cAtoh1 (the chick Atoh1 homolog). The subsequent events that lead to the development of these cells as hair cells have not been determined. Moreover, it is not clear how much of the subsequent signaling events described are recapitulated in transdifferentiating cells. For instance, would transdifferentiating hair cells need to activate the Notch signaling pathway to inhibit neighboring cells from developing as hair cells because these cells are already supporting cells? Clearly the answers to these questions, as well as, a more complete understanding of the differences between hair cell development and hair cell regeneration should have significant implications for the development of regenerative strategies.

Differentiation and Survival of Hair Cells

Following the commitment to a hair cell fate, the earliest molecule that is known to be expressed exclusively in differentiating hair cells is the unconventional myosin, myosin-VI. In the developing mammalian cochlea, myosin-VI can be detected in developing hair cells at almost the same time as

Atoh1. However, despite its early onset of expression, myosin-VI does not appear to play a role in hair cell differentiation because mutations in myosin-VI do not result in defects in hair cell development. The same is true for another unconventional myosin, myosin-VIIa, which is expressed soon after myosin-VI but is also not required for hair cell formation. At this point, the early functions, if any, of myosin-VI and myosin-VIIa remain unclear. Both molecules are required for the normal development and maintenance of the stereociliary bundle, but because the bundle does not develop until several days after initial commitment and differentiation, it is assumed that both myosins must play additional roles in early hair cell formation.

POU4f3, Gfi1, and Barhl1

Soon after the onset of expression of myosin-VI and myosin-VIIa, developing hair cells begin to express a group of transcription factors that are required for hair cell survival and possibly differentiation. The first of these factors is the class IV Pou-domain transcription factor, Pou4f3 (also known as Brn-3c or Brn-3.1). Expression of Pou4f3 is first detected approximately 24–48 h after the onset of expression of Atoh1. In cochleae from *Pou4f3*$^{-/-}$ animals, the initial development of hair cells appears normal, as determined by expression of myosin-VI and myosin-VIIa; however, there is a rapid degeneration leading to hair cell death and auditory and vestibular deficits, indicating that Pou4f3 is required for hair cell maintenance. Moreover, late-onset progressive hearing loss occurs in an Israeli Jewish family with an 8 bp deletion in the homeodomain of *POU4F3,* further demonstrating the importance of this gene in hair cell maintenance.

Following the onset of expression of Pou4f3, developing hair cells upregulate Gfi1, a zinc finger transcription factor that has been implicated in development of the hematopoietic system. As is the case for *Pou4f3*, the deletion of *Gfi1* leads to early degeneration of hair cells; however, the expression of Atoh1, myosin-VI, myosin-VIIa, and Pou4f3 is not disrupted, indicating that Gfi1 acts downstream of Pou4f3.

Beginning about the same time as Pou4f3, hair cells also express Barhl1, a member of the Bar family of homeobox genes. In Drosophila, barh1 and barh2 are required for the development of sensory organs. However, the development of hair cells in mice with a targeted deletion of *Barhl1* appears normal, with a full complement of hair cells formed at postnatal day (P)0. Although hair cells develop normally, signs of hair cell degeneration are present beginning at P6 and all hair cells ultimately die, but not until P300. Based on these results, it seems likely that *Barhl1* acts as a hair cell survival gene, whereas *Pou4f3* and *Gfi1* may act as both survival and differentiation genes. Because the hair cells die during the differentiation process in both *Pou4f3* and *Gfi1* mutants, it is not possible to determine whether survival, differentiation, or both are disrupted.

Role of Retinoic Acid Signaling in Hair Cell Differentiation

Another group of transcription factors that are expressed in developing hair cells soon after the onset of cellular differentiation is the nuclear receptors for retinoic acid (RA) and thyroid hormone (TH). Receptors for RA and TH belong to the steroid/thyroid receptor superfamily containing a zinc finger DNA binding domain and a ligand dimerization domain. RA signaling is mediated by two different forms of receptors, RARs (α, β, and γ) and RXRs (α, β, and γ), both of which are expressed in inner ear-sensory epithelia. The treatment of cochlear explant cultures with RA results in supernumerary hair cells and supporting cells, and it similarly induces hair cell differentiation in chicken otocyst cultures. The inhibition of RA synthesis or antagonism of RARα results in a decrease in the number of cells that develop as hair cells. However, Brn3.1 and myosin-VI are still expressed in hair cells, even in the presence of RA inhibitors, suggesting that initial differentiation of hair cells does not require RA. Based on these results, it was suggested that RA signaling through RARα plays a direct role in hair cell differentiation during early stages in development. However, a small number of hair cells are present in RARα/RARγ double-mutant mice (RARα mutant mice are phenotypically normal), suggesting that hair cell differentiation is not completely dependent on RA. However, considering that there are three *RAR* genes and three *RXR* genes, and that functional redundancy of RA receptors has been demonstrated in other systems, it would be informative to inactivate all three RAR genes or all six RA receptor genes within the inner ear.

Role of TH Signaling in Hair Cell Differentiation

Hypothyroidism or iodine deficiencies in pregnant women are known to cause hearing deficits in children, and induced hypothyroidism in animal models results in severe morphological abnormalities in the organ of Corti that appear consistent with a delay in cellular differentiation. Thyroid hormone signaling is mediated through three functional receptors, Trα1, Trβ1, and Trβ2, and a nonligand binding, Trα2.

All of these receptors are expressed within the developing inner ear; however, the expression of *Trβ1* and *Trβ2* is confined to the developing cochlear epithelium. Elimination of TRβ1 and TRβ2 through deletion of the *Trβ* gene results in animals that are deaf as a result of a defect in the maturation of a hair cell ionic conductance that is required for cellular function. However, the overall structure of the organ of Corti is not comparable with hypothryroid animals and, in fact, appears normal, suggesting a possible functional redundancy. Consistent with this hypothesis, the deletion of both *Trα* and *Trβ* results in a more profound hearing loss than in animals lacking just *Trβ*, and the morphology of the organ of Corti in these mice is consistent with the hypothyroid phenotype.

In both hypothyroid animals and *Trα/Trβ* double mutants, the phenotype of the organ of Corti is consistent with a delay or arrest in cellular differentiation. To determine whether this is in fact the case, the expression of a molecule that is turned on fairly late in the development of a subset of hair cells, Prestin, was examined in hypothyroid animals. As expected, Prestin expression is delayed in hypothyroid hair cells. Moreover, an analysis of the promoter region for the *Prestin* gene identified two specific TR binding elements, strongly suggesting that ongoing differentiation of hair cells is dependent on thyroid hormone signaling.

Summary

The specification of hair cells within the inner ear is dependent on the activation of a number of molecular signaling pathways that progressively direct the development of prosensory patches and subsequently direct individual cells within those patches to develop as hair cells. Although the factors that direct the specification of prosensory patches are not fully understood, the Notch signaling pathway and the transcription factor Sox2 clearly play important roles. Within each prosensory patch, the bHLH transcription factor Atoh1 is necessary and sufficient to induce progenitors to develop as hair cells, whereas subsequent Notch-mediated inhibitory interactions actually determine how many cells will develop as mature hair cells. Once committed, developing hair cells activate a number of genes, including *Pou4f3* and *Gfi1,* that are required for their survival and, probably, for differentiation. In addition, developing hair cells become dependent on external signaling factors, such as RA and TH, to stimulate further steps in their differentiation toward functioning hair cells.

See also: Cochlear Mechanics; Cochlear Development; Hair Cell Regeneration; Hair Cells: Sensory Transduction; Helix–Loop–Helix (bHLH) Proteins: Proneural; Notch Signal Transduction: Molecular and Cellular Mechanisms; Notch Pathway: Lateral Inhibition; Sensory Aging: Hearing; Vestibular System.

Further Reading

Barald KF and Kelley MW (2004) From placode to polarization: New tunes in inner ear development. *Development* 131: 4119–4130.

Bermingham NA, Hassan BA, Price SD, et al. (1999) Math1: An essential gene for the generation of inner ear hair cells. *Science* 284: 1837–1841.

Bryant J, Goodyear RJ, and Richardson GP (2002) Sensory organ development in the inner ear: Molecular and cellular mechanisms. *British Medical Bulletin* 63: 39–57.

Eatock RA, Fay RR, and Popper AN (2006) *Vertebrate Hair Cells*. New York: Springer.

Fekete DM and Wu DK (2002) Revisiting cell fate specification in the inner ear. *Current Opinions in Neurobiology* 12: 35–42.

Forrest D, Reh TA, and Rusch A (2002) Neurodevelopmental control by thyroid hormone receptors. *Current Opinions in Neurobiology* 12: 49–56.

Fritzsch B, Pauley S, and Beisel KW (2006) Cells, molecules and morphogenesis: The making of the vertebrate ear. *Brain Research* 1091: 151–171.

Kelley MW (2006) Hair cell development: Commitment through differentiation. *Brain Research* 1091: 172–185.

Kelley MW (2006) Regulation of cell fate in the sensory epithelia of the inner ear. *Nature Reviews Neuroscience* 7: 837–849.

Kelley MW, Wu DK, Popper AN, and Fay RR (2005) *Development of the Inner Ear*. New York: Springer.

Kiernan AE, Pelling AL, Leung KK, et al. (2005) Sox2 is required for sensory organ development in the mammalian inner ear. *Nature* 434: 1031–1035.

Lanford PJ, Lan Y, Jiang R, et al. (1999) Notch signalling pathway mediates hair cell development in mammalian cochlea. *Nature Genetics* 21: 289–292.

Moroy T (2005) The zinc finger transcription factor Growth factor independence 1 (Gfi1). *International Journal of Biochemistry and Cell Biology* 37: 541–546.

Rubel EW, Popper AN, and Fay RR (1998) *Development of the Auditory System*. New York: Springer.

Rusch A, Ng L, Goodyear R, et al. (2001) Retardation of cochlear maturation and impaired hair cell function caused by deletion of all known thyroid hormone receptors. *Journal of Neuroscience* 21: 9792–9800.

Hair Cell Regeneration

J R Meyers, University of Virginia, Charlottesville, VA, USA; and Colgate University, Hamilton, NY, USA
Z Hu, Z Lu, and J T Corwin, University of Virginia School of Medicine, Charlottesville, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Millions of lives are affected by hearing and balance deficits that arise when sensory hair cells in the inner ear are killed by loud sounds, head trauma, infections, autoimmune disorders, and certain drugs. Those deficits permanently affect humans and other mammals, but nonmammalian vertebrates can regenerate lost hair cells and recover their hearing and balance within a matter of weeks. Here we review the biology of hair cell regeneration and the prospects for stimulating the replacement of lost hair cells in humans and other mammals.

Hair Cell Structures and Functions

Sensory hair cells are named for a microscopic, directionally polarized hair bundle composed of 30–300 actin-filled modified microvilli, called stereocilia, and a single eccentrically positioned true cilium, the kinocilium. The hair bundle extends from the cell's apical surface into the fluid in the inner ear or the lumen of one of the lateral line organs on the head or body of a fish or aquatic amphibian (Figure 1). Hair cells detect sounds, accelerations of the head, and water motion around the body when small, stimulus-driven forces deflect their hair bundles. Deflections open and close mechanosensitive transduction channels, which leads to changes in transmembrane potential, Ca^{2+} entry, and the release of neurotransmitter.

Humans normally begin life with each ear's snail-shaped cochlea containing ~16 000 sound-sensitive hair cells precisely arranged in four rows in the organ of Corti sensory epithelium, which runs ~30 mm along the basilar membrane (Figure 1(b)). Those hair cells detect sound waves transmitted to the inner ear by the middle ear bones and provide the accurate frequency discrimination we rely on for communication by speech. Each ear also contains ~19 000 and ~33 000 hair cells in the gravity-sensitive saccule and utricle, respectively. Angular accelerations of the head are detected by the three semicircular canals, which contain ~7600 hair cells apiece (Figure 1(b)). The saccule, utricle, and semicircular canals constitute the vestibular division of the inner ear. The occurrence of age-related hearing impairments and the deterioration of visual stability reflexes and balance correlate with progressive age-related declines in the number of hair cells present in the human cochlea and vestibular epithelia.

The Developmental Production of Hair Cells

Mammals produce nearly all of their hair cells during embryonic development, but hair cell production can continue well beyond birth in other vertebrate classes. Like mammalian ears, the ears and lateral line organs of fish, amphibians, and reptiles develop reasonably large complements of hair cells during embryogenesis. However, the sensory epithelia in those animals continue to give rise to substantial numbers of new hair cells throughout life. For example, counts from scanning electron microscopy have shown that sharks add hundreds of thousands of hair cells in the inner ear as juveniles grow to adulthood. Sensory cell production continues even in the oldest animals within natural populations of fish and amphibians.

In the postembryonic sensory epithelia of those animals, proliferation and interstitial addition of hair cells occur at a low frequency throughout their maculae, but proliferating cells and newly produced hair cells are more frequently found near the peripheral edges of the sensory maculae. The occurrence of growth zones at the outer margins of these diverse types of hair cell epithelia suggests the possible existence of latent or slowly replicating progenitor cells that may be maintained near the progressively moving zone of transition between the expanding sensory epithelium and surrounding nonsensory epithelium. If such stem cells exist in growing hair cell epithelia, they are as yet unidentified, but there is the possibility that combinations of signals and cellular environment may support a stem cell niche at the edges of these sensory epithelia, similar to the reserve stem cell pools that persist at the outer edges of the retina in fish, amphibians, and birds.

Vestibular organs in mammals and both auditory and vestibular epithelia in chickens develop via a pattern of radial or appositional growth, similar to postembryonic growth in fish and amphibians. The first cells to withdraw from cycling are located near the center of each sensory epithelium in a zone of nonproliferation that expands radially outward until all the cells in the epithelium have been produced. Late in embryogenesis or soon after birth, cell production is believed to effectively cease in the

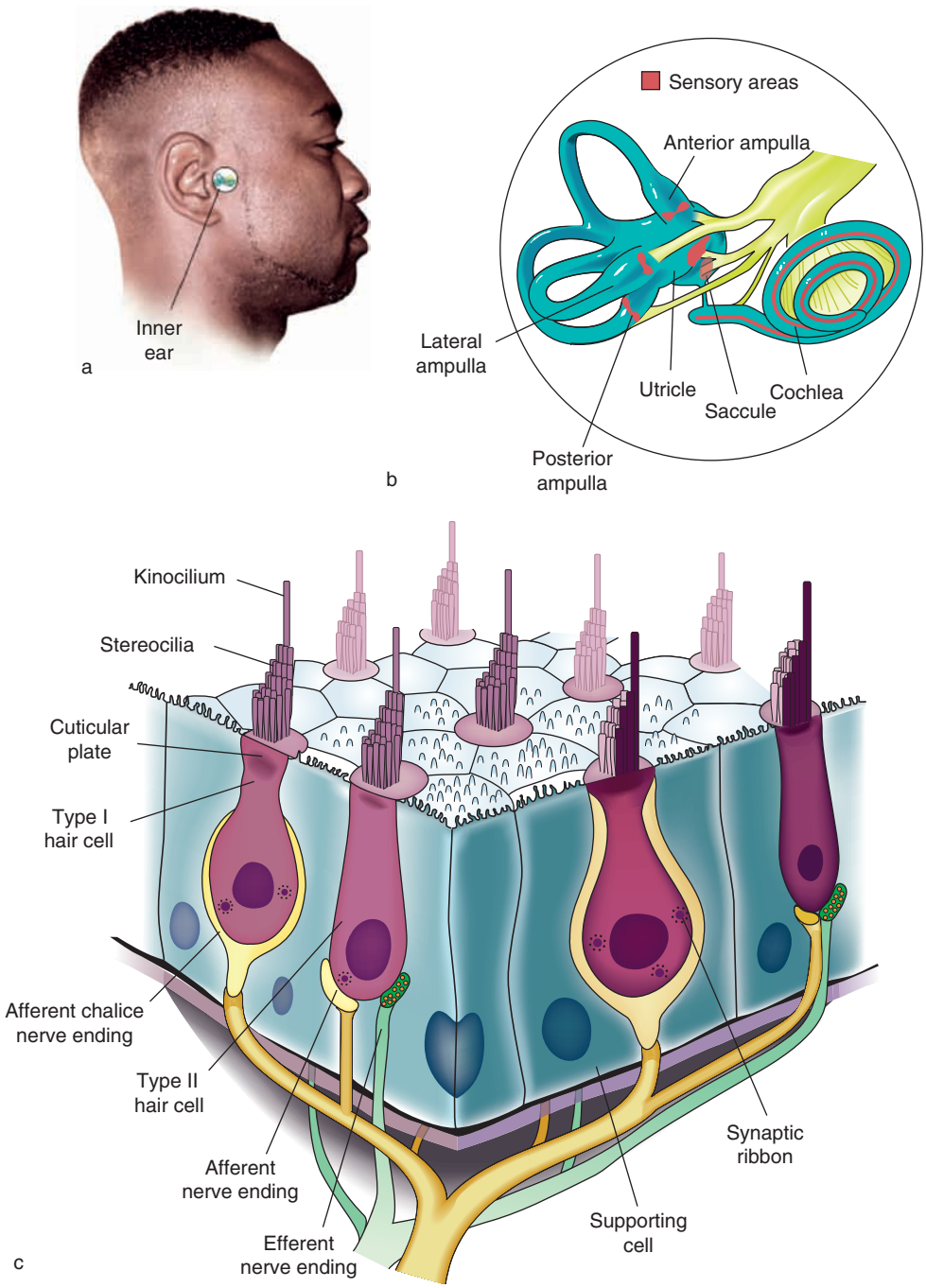

Figure 1 Schematic diagrams of the human inner ear: (a) approximate size and location of the sensory organs of the inner ear; (b) drawing of the auditory and vestibular labyrinth (the sensory patches within each organ are indicated by the dark coloration); (c) representation of a section of the sensory epithelium from a vestibular organ to demonstrate the general structure and organization of the sensory epithelium.

mammalian vestibular epithelia and the avian cochlea. During embryogenesis of those organs, however, mechanisms operate to sustain a yet-to-be-identified class of proliferative cells at the outer margins of growing hair cell epithelia, which may be widely conserved across all vertebrates and may persist throughout life in fish and amphibians.

The Regenerative Replacement of Hair Cells

Identification of the cells that give rise to replacement hair cells is critical to an accurate understanding of the events that lead to regenerative recovery of hearing, balance, and lateral line functions. Sensory

epithelia in the inner ear contain at least two distinct resident cell types – sensory hair cells and supporting cells (Figure 1(c)) – and lateral line epithelia contain hair cells and both internal supporting cells and mantle-type supporting cells.

Supporting cells occupy the spaces between neighboring hair cells, extending from the apical to the basal surfaces of the epithelium and ensheathing the sides and bases of the hair cells and nerve terminals. When hair cells are lost, surviving supporting cells divide and give rise to progeny that can continue to divide or can differentiate as replacement hair cells and supporting cells (Figure 2). Regeneration also can occur after the loss of entire lateral line neuromasts as well as in cases in which only hair cells have been lost from the neuromast.

In fish and aquatic-stage amphibians, entire groups of lateral line neuromasts regenerate after tail amputation. All the cells that make up those neuromasts arise from a cell lineage produced when mantle-type supporting cells located near the outer edge of the last remaining neuromast on the stump of the tail begin to divide. The epithelial progenitor cells produced from those cell divisions migrate out of the neuromast, divide further, and give rise to migratory cells that form a regenerative placode that extends posteriorly beneath the epidermis of the regenerating tail. As tail regeneration progresses, small groups of cells break away from the trailing anterior edge of the posteriorly moving regenerative placode. Those loose collections of progenitor cells progress through aggregation, self-organization into rosettes that will become replacement neuromasts, and the subsequent production of the first replacement hair cells, which arise in the very center of each radially symmetrical rosette.

In birds, the sensory epithelium of the cochlea forms in the embryo, and tracer experiments have shown that, under normal conditions, there is no replication of nuclear DNA in hair cells or supporting cells after hatching. When the sensory epithelium has been injured, however, many of the previously quiescent supporting cells that are in or near the lesion reenter the cell cycle, divide, and generate new cells that can quickly differentiate into replacement hair cells and supporting cells. Six days after the end of a 48 h exposure to a loud tone that kills hair cells, small hair bundles become visible within the confines of the former lesion. Those bundles grow rapidly and

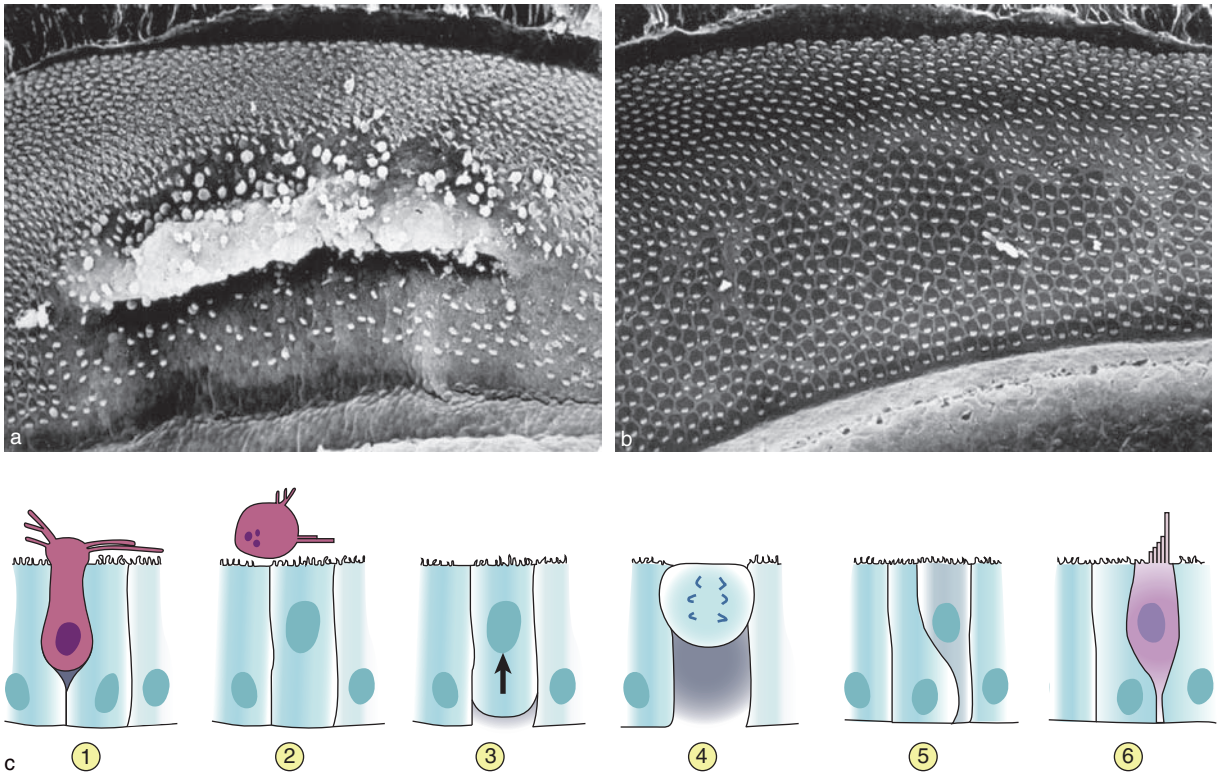

Figure 2 Hair cell regeneration in the chicken basilar papilla: (a) Following 48 h of acoustic overstimulation at 1.5 kHz, hair cells within the region of the cochlea tuned to 1.5 kHz are lost and can be seen extruding from the epithelium; (b) after being allowed 10 days to recover following acoustic overstimulation, the epithelium has been repaired and hair cells can be seen throughout the formerly damaged regions; (c) diagram of regenerative proliferation by supporting cells. In (c), in response to the loss of a hair cell (1), the supporting cell nucleus migrates toward the epithelial surface (2–3), where it undergoes mitosis (4). The progeny of that division then go on to form replacement hair cells or supporting cells (5–6). (a, b) Corwin and Cotanche (1998).

appear nearly mature by 10 days after the end of the sound exposure (**Figures 2(a) and 2(b)**).

Bundle form differs progressively along a gradient extending between the ends of the avian cochlea. The hair bundles at the high-frequency-sensitive proximal end of the chicken's cochlea contain over 200 short stereocilia, whereas the bundles at the low-frequency-sensitive distal end have less than 50 long stereocilia. Remarkably, the replacement hair cells develop hair bundles appropriate for the position the regenerated cells occupy, indicating that positional identity cues are maintained. How this occurs remains to be determined.

New cells that arise from the divisions of supporting cells appear to differentiate as hair cells by default when they develop in a region that does not already contain hair cells. Lateral inhibition by hair cells suppresses hair cell fate in immediate neighbors, setting up a mosaic pattern in which each hair cell is surrounded by supporting cells. Notch signaling may underlie the fate choices, as in the *mindbomb* mutant zebra fish line with disrupted Notch signaling, all of the progeny differentiate into hair cells and no supporting cells are produced.

The Innervation of Regenerated Hair Cells

In the weeks after regeneration of hair cells in birds, the replacement cells become innervated as synapses progress through stages of maturation similar to those that occur in embryos. In the cochlea, afferent terminals synapse on all the hair cells early in the regeneration process, but afferents subsequently withdraw from the short hair cells. After vestibular hair cell regeneration in birds, bouton, dimorphic, and calyceal afferent terminals grow slowly and progressively for up to 1 year, by which point the regeneration of those afferents has largely recapitulated the developmental stages that normally lead to their topographic organization.

Functional Recovery Follows Hair Cell Regeneration

After antibiotic-induced hair cell loss, thresholds for auditory-evoked potentials in the cochlear nuclei of chickens shift by 40–50 dB. By 16–20 weeks, they recover significantly in all but the high-frequency regions. Behavioral sound discrimination and precise vocal production in birds also recover after hair cell regeneration.

After streptomycin treatment, pigeons exhibit severe head and postural instability, with tremors, head shakes, staggering, and circling. These vestibular dysfunctions last for weeks, but gradual recovery that corresponds with the time for hair cell regeneration restores the capacity for spatial orientation and path navigation.

The Potential for Regeneration of Mammalian Hair Cells

There is little if any credible evidence for innate regenerative capacity in the mature mammalian organ of Corti. It appears likely that limits to regeneration relate to developmental controls inherent in the formation of the highly ordered arrangement of inner and outer hair cells and the specialization of the distinct types of supporting cells in the mammalian organ of Corti (Deiter's cells, Hensen's cells, Pillar cells, and phalangeal cells), which are unmatched in the ears of nonmammals. There are, however, limited regenerative responses in mammalian vestibular epithelia, which share structural similarities to their counterparts in the ears of fish, amphibians, and birds (**Figure 3**). When juvenile guinea pigs have been treated with aminoglycoside antibiotics, hair bundles are lost from certain regions of the utricle, but when other guinea pigs have been allowed to recover for 4 weeks after such treatments, their utricles contain many small hair bundles, reminiscent of bundles on regenerating hair cells in the avian ear. Small hair bundles are also present in very small but consistent numbers throughout the utricles of adult guinea pigs. Aminoglycoside antibiotics have been used in culture to kill the majority of the hair cells in utricles from adult guinea pigs and adult humans. The loss is followed by the proliferation of up to 100 or more supporting cells. However, the modest proliferative responses of mature mammalian vestibular epithelia are considerably below those observed in nonmammals.

Proliferation Controls in Hair Cell Epithelia

Treatments with transforming growth factor (TGF-)α, epidermal growth factor (EGF), insulin-like growth factor (IGF-1), insulin, glial growth factor (GGF2), and heregulin can increase mammalian supporting cell proliferation to varying degrees. The cells from the ears of young mammals are most responsive, and there is a significant decline in the number of supporting cells that respond to mitogenic growth factors as neonatal rodents mature during the first weeks after birth. Supporting cells continue to express growth factor receptors and small numbers of those cells can be induced to divide, but more research is needed to discover what limits the production of cells in mature mammalian hair cell epithelia.

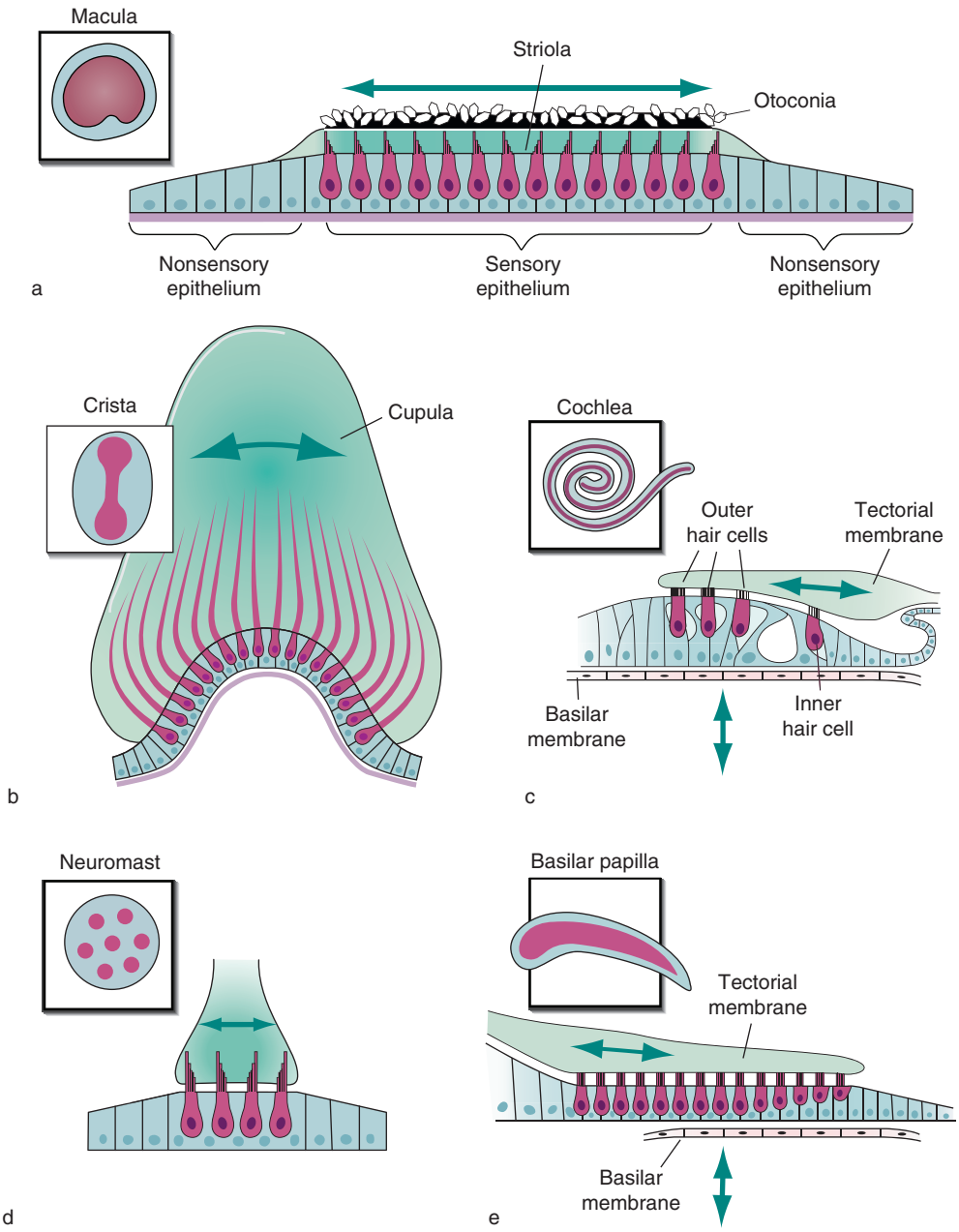

Figure 3 Diagrams showing the general structural organization of sensory epithelia: (a) organization of otolith macular organs in all vertebrates; (b) organization of the semicircular canal cristae; (c) organization of the mammalian cochlea; (d) the lateral line neuromasts found along the body of fish and amphibians; (e) the basilar papilla of birds. In (a), the sensory patch, indicated by the dark color in the inset, comprises hair cells and supporting cells in a mosaic pattern. The hair cells insert their hair bundles into a gelatinous matrix covered by otoconial crystals that provide inertial weight to acceleration of the head (arrow). The sensory epithelium is surrounded by a nonsensory epithelium, made up of supporting-like cells. In (b), the barbell-shaped sensory patch is organized very similarly to that of the otolith macular organs. The hair bundles, however, are inserted into a large cupula, which transduces rotational movement of fluid through the canal to the hair bundles (arrow). In (c), in contrast to the generally uniform population of supporting cells in vestibular organs, the cochlear supporting cells show discrete morphological specializations to form the unique structure of the organ of Corti. In (d), the lateral line neuromasts in fish and amphibians are organized in a way similar to vestibular organs, although there are many fewer hair cells per organ. In (e), the basilar papilla of birds shows some similarity to the vertebrate cochlea, although the supporting cells do not form morphologically discrete subtypes or show high levels of specialization.

The contributions of mitogen-activated protein kinase (MAPK), phosphatidyl inositol 3′ kinase (PI-3K), target of rapamycin (TOR), and protein kinase C (PKC) to the proliferative response appear similar in young birds and neonatal rodents, but they may differ as mammals mature. Tumor suppressors have important roles in the embryonic development of the ear, but it is unclear whether they are the limiting factors that restrict the regenerative replacement of hair cells in postnatal mammals. The cyclin-dependent kinase inhibitors (CDIs), such as $p27^{Kip1}$ and $p19^{Ink4}$, both appear to inhibit cell cycle progression in developing the mammalian ear. $p27^{Kip1}$ is selectively expressed in supporting cells, and the disruption of the gene allows cell proliferation to continue in postnatal mouse organ of Corti beyond the time when mitosis normally has ceased. It has not been determined whether differences in CDI regulation or expression exist between mammalian and nonmammalian ears. Additional research is needed to identify the upstream controls that restrict proliferation and to discover treatments that may overcome that restriction and promote effective regeneration in mammals.

Hair Cell Death and Regeneration

Hair cell death can occur through necrosis or apoptosis, but there are reasons to believe that clinically relevant forms of hair cell loss occur primarily through apoptosis. Unlike the populations of hair cells in the mammalian ear and those in the avian cochlea, the hair cell populations in the vestibular organs of birds turn over throughout life, with the hair cells typically living for only a few months. Experimental decreases in apoptotic hair cell death have been achieved using caspase inhibitors, and that results in reduced rates of ongoing proliferation. Thus, the rate of hair cell death regulates the rate of proliferation, but the specific factors that limit the life span of avian vestibular hair cells remain to be discovered. Identification of those factors could contribute to understanding the causes of age-related hair cell loss and other conditions that reduce the life span of human hair cells, whose longevity is believed to extend to a century or longer under normal conditions.

Potential Triggers for Regeneration

Although the loss of hair cells and the initiation of proliferation appear closely tied, the specific signals that trigger the regeneration responses remain to be identified. Loss of contacts between hair cells and supporting cells may result in signals transmitted from cell to cell through the subapical epithelial junctions or via gap junctions or Notch–Delta signaling. Laser microbeam ablations of hair cells in the avian cochlea have shown that the initial proliferative responses occur in supporting cells that lie up to ~200 μm beyond those cells that had direct contact with the lost hair cells. This suggests that the signals that trigger regenerative proliferation go beyond simple loss of cell–cell contact but are unlikely to be widely diffusing signals.

The onset of proliferation in the mantle-type supporting cells that contribute to the regeneration of hair cells in neuromasts in salamanders correlates with the recruitment of macrophages to the vicinity of the wound after tail amputation. TGFα and tumor necrosis factor (TNF-)α are cytokines secreted by macrophages and can promote the proliferation of avian and mammalian supporting cells. The inhibition of cytokine production reduces the proliferation of avian supporting cells *in vitro*, but the specific role that the immune response has in triggering hair cell regeneration *in vivo* remains to be determined.

Reserve Stem Cells versus Dedifferentiation

Persistent but unresolved questions revolve around the potential existence of a discrete subpopulation of reserve stem cells that might retain the capacity to produce new hair cells. The alternative question is whether all or nearly all supporting cells may be capable of dedifferentiating and contributing to regeneration. Currently, both possibilities derive support from indirect evidence. Neither has been ruled out, but the case for the importance of reserve stem cells is not as strong as that for dedifferentiation of supporting cells. Progenitor cells that have replicated DNA in preparation for cell division have thus far remained morphologically indistinguishable from other supporting cells in the ears of fish, amphibians, and birds. When treatment with a mitotoxin was used to eliminate initially dividing cells in experimentally damaged ears of fish, this did not prevent regeneration subsequent to a second round of damage. In chicken ears, the majority of supporting cells in regions of damage downregulate expression of G_0 markers, begin expressing markers for passage to the late G_1 phase, and begin to replicate their DNA. These data and the absence of immunocytochemical evidence for distinct supporting-cell subtypes are consistent with the possibility that the majority of supporting cells have the capacity to reenter the cell cycle and divide. Other evidence indicates that hair cell production in lateral line organs occurs via symmetrical division of a single progenitor to produce two hair cells during development and regeneration,

but still other evidence indicates that asymmetric divisions of supporting cells can produce both replacement hair cells and supporting cells.

Other experiments have suggested that a small number of cells from the adult mouse utricle may act as colony-forming stem cells whose progeny can differentiate into hair cells, neurons, muscle cells, kidney cells, liver cells, and skin cells. The suggestion that a handful of stem cells may exist within mature mammalian vestibular epithelia is enticing for its apparent consistency with observations that have shown only a limited number of dividing cells in cultures of mammalian vestibular epithelia, but such attractive interpretations should be taken cautiously until replicated by others.

Although most (if not all) supporting cells in nonmammalian vertebrates appear to be capable of returning to the cell cycle and dividing, it remains to be determined exactly what controls and limits the proliferation of supporting cells in the ears of mature mammals. As already noted, *in vitro* investigations have shown that treatments with appropriate growth factors and pharmacological agents can stimulate proliferation in the majority of supporting cells in epithelia cultured from the utricles of neonatal rodents. Thus, the observed proliferative capacity of supporting cells in the perinatal utricular epithelium appears to be more consistent with the hypothesis that nearly all supporting cells may be capable of proliferating as opposed to the hypothesis that proliferation depends on relatively rare reserve stem cells. Experiments that found age-related decreases in the response of mammalian supporting cells to growth factor stimulation suggest that maturing mammalian supporting cells may change in ways that restrict their capacity to act as progenitors. Whether the proposed stem cells in the mature mammalian inner ear are simply retained as less-differentiated supporting cells, whether there may be multiple subclasses of supporting cells that have different capacities for proliferation, or whether all mammalian supporting cells can be induced to return to the cell cycle remains to be determined.

Although evidence indicates that supporting cells can act as multipotential progenitors to produce new hair cells and new supporting cells, the supporting cells in many types of otic epithelia spend the majority of their life as mitotically quiescent, differentiated cells. For differentiated cells to reenter the cell cycle, they must first dedifferentiate. Some of the best-studied cases of dedifferentiation occur as urodele amphibians regenerate lost limbs or tails, in which cells from the neural tube, muscle, connective tissue, and Schwann cells have all been observed to dedifferentiate and contribute to the mass of proliferating multipotent cells that form the regeneration blastema. As regeneration progresses, the progeny from those cells differentiate into each of the cell types needed to replace all of the structures in the missing parts of the limb or tail. In a situation that appears somewhat similar to that in the inner ear, after damage amphibian myotubes readily dedifferentiate and contribute to the regeneration of lost structures, but mammalian myotubes remain differentiated and mitotically quiescent. Identification of the factors that restrict or control differentiation in the ear may provide fruitful targets for manipulation to promote the regeneration of hair cells.

Nonproliferative Recovery after Hair Cell Damage

Although proliferation of supporting cells and differentiation of the progeny into replacement hair cells and supporting cells seem to be the primary mechanisms for regeneration following injury, there is evidence that nonproliferative mechanisms may also contribute to recovery. In the mature mammalian vestibular system, there is evidence for significant recovery of new hair bundles but of only very limited proliferation of supporting cells. Two hypotheses have been put forth to account for this quantitative mismatch: repair of damaged hair cells and direct phenotypic conversion of supporting cells into hair cells without intervening mitosis (**Figure 4**).

In response to sublethal doses of aminoglycoside antibiotics, hair cells are able to survive the insult, with some losing their hair bundle and descending below the epithelial surface. Loss of a hair bundle does not always lead to hair cell death, which calls into question the conclusions of many studies that have relied on counts of hair bundles as the sole metric for hair cell loss and replacement.

Evidence for the nonproliferative conversion of supporting cells into hair cells is often less direct. During murine cochlear organogenesis, newly produced supporting cells convert directly into hair cells if the preexisting hair cells are lesioned through laser microbeam ablations, but this appears to occur only during the first 48 h after the production of the original hair cells. It has been proposed that supporting cells in other vertebrate sensory epithelia may be capable of phenotypic change, sometimes inappropriately termed transdifferentiation, long after organogenesis. When tracers for newly replicated DNA are administered after aminoglycoside-induced injury in birds and amphibians, the first hair cells that appear in the regions of damage have not incorporated the mitotic tracers, although in birds the later-produced hair cells do incorporate the mitotic tracers. Also, cells with some morphological characteristics intermediate between hair cells and supporting cells have been found in damaged sensory

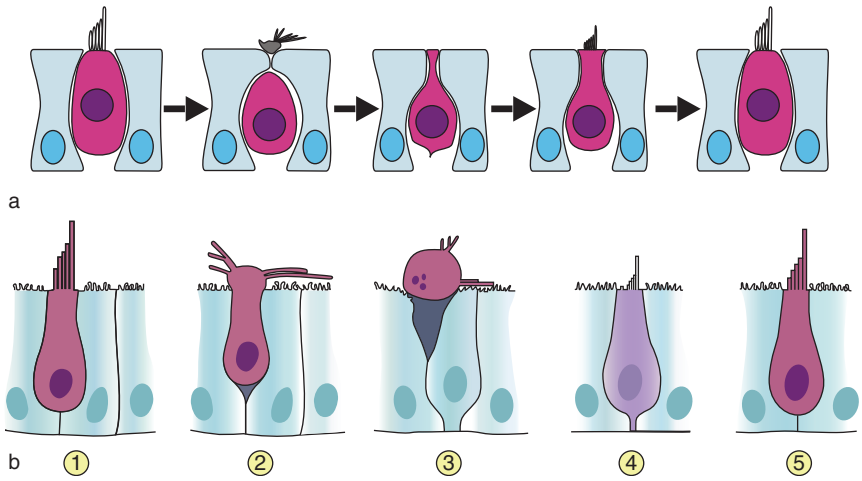

Figure 4 Diagrams of two proposed mechanisms of nonproliferative regeneration: (a) repair of damaged hair cells; (b) direct conversion of a supporting cell into a hair cell without an intervening mitosis. As shown in (a), in response to damage, the hair cell loses its hair bundle and apical cytoplasm, but over time is able to regenerate a new bundle that matures into a normal hair bundle. As shown in (b), when a hair cell is lost (1–2), one of the neighboring hair cells, no longer suppressed by an adjacent hair cell (3), is able to redifferentiate into a hair cell and develop a morphologically mature hair bundle (4–5).

epithelia. Finally, there can be recovery of hair bundles in damaged organs cultured in a mitotic inhibitor that prevent proliferation. However, it can be difficult to determine what contribution to recovery of hair bundles is due to intracellular repair and what is attributable to phenotypic conversion.

Significantly, the clinical correspondence between hair cell loss and the onset of permanent auditory and vestibular deficits in mammals suggests that any such nonproliferative mechanisms do not lead to the same level of functional recovery observed in nonmammalian vertebrates, in which robust regenerative proliferation occurs. The restriction in supporting cell proliferation, therefore, appears to be the primary factor that leaves mammals uniquely vulnerable to permanent hearing and vestibular deficits due to loss of hair cells.

Gene Therapy to Stimulate Conversion of Supporting Cells into Hair Cells

The hypothesis that supporting cells may have an endogenous capacity for direct phenotypic conversion into hair cells has received support from *in vitro* experiments that have established that supporting cells and other cells that reside just outside the sensory epithelia can be converted into hair cells through transfection and overexpression of the basic helix–loop–helix (HLH) transcription factor *Atoh1* (the mammalian homolog of *atonal*). Those cells take on the morphological characteristics of hair cells, express hair-cell-specific proteins, and develop hair bundles, suggesting that forced conversion could be used to promote hair cell recovery *in vivo*. In fact, it has been reported that, after aminoglycoside poisoning, viral delivery of the *Atoh1* gene *in vivo* leads to remarkably complete structural recovery of the four rows of hair cells in the rodent cochlea and recovery of hearing sensitivity, but replication and further investigation of these intriguing possibilities are necessary to understand the exact contribution that genetically converted hair cells make to the reported recovery.

Conclusion

Hair cell losses can produce severe hearing and balance deficits in mammals and nonmammals alike, but nonmammals recover within weeks because epithelial supporting cells divide and give rise to replacement hair cells. In mammalian ears, the capacity for cell production decreases soon after birth, but processes that do not depend on cell proliferation continue and may contribute to structural recovery after damage. However, such nonproliferative forms of repair may not normally bring about complete restoration of function, because hair cell losses in humans and other mammals typically result in permanent sensory deficits.

Balance epithelia cultured from the ears of rodents and adult humans respond to hair cell loss with low levels of supporting cell proliferation, and small hair bundles have been found throughout the vestibular epithelia of adult rodents, suggesting that the cellular machinery necessary for regeneration is present but held to a low level in the ears of mature mammals. Investigations are continuing to explore the mechanisms that underlie the robust hair cell regeneration that occurs in nonmammalian vertebrates, the mechanisms

that limit proliferation and regeneration in the mammalian ear, and experimental strategies that eventually may have the potential to stimulate effective regenerative responses and functional recovery in mammals.

See also: Cochlear Development; Deafness; Hair Cells: Sensory Transduction; Hair Cell Differentiation; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Vestibular System.

Further Reading

Bermingham-McDonogh O and Rubel EW (2003) Hair cell regeneration: Winging our way towards a sound future. *Current Opinion in Neurobiology* 13: 119–126.

Fritzsch B, Beisel KW, and Hansen LA (2006) The molecular basis of neurosensory cell formation in ear development: A blueprint for hair cell and sensory neuron regeneration? *Bioessays* 28: 1181–1193.

Gale JE, Meyers JR, Periasamy A, and Corwin JT (2002) Survival of bundle-less hair cells and subsequent bundle replacement in the bullfrog saccule. *Journal of Neurobiology* 50: 81–92.

Grant KA, Raible DW, and Piotrowski T (2005) Regulation of latent sensory hair cell precursors by glia in the zebrafish lateral line. *Neuron* 45: 69–80.

Gu R, Montcouquiol M, Marchionni M, and Corwin JT (2007) Proliferative responses to growth factors decline rapidly during postnatal maturation of mammalian hair cell epithelia. *European Journal of Neuroscience* 25: 1363–1372.

Hawkins RD, Bashiardes S, Powder KE, et al. (2007) Large scale gene expression profiles of regenerating inner ear sensory epithelia. *PLoS ONE* 2(6): e525[online].

Holley MC (2002) Application of new biological approaches to stimulate sensory repair and protection. *British Medical Bulletin* 63: 157–169.

Hu Z and Corwin JT (2007) Inner ear hair cells produced *in vitro* by a mesenchymal-to-epithelial transition. *Proceedings of the National Academy of Sciences of the United States of America* 104: 16675–16680.

Kelley MW (2006) Regulation of cell fate in the sensory epithelia of the inner ear. *Nature Reviews Neuroscience* 7: 837–849.

Laine H, Doetzlhofer A, Mantela J, et al. (2007) p19(Ink4d) and p21(Cip1) collaborate to maintain the postmitotic state of auditory hair cells, their codeletion leading to DNA damage and p53-mediated apoptosis. *Journal of Neuroscience* 27(6): 1434–1444.

Lee Y-S, Liu F, and Segil N (2006) Transcriptional regulation of p27Kip1 directs a morphogenetic wave of cell cycle exit in the developing organ of Corti. *Development* 133: 2817–2826.

Li H, Corrales CE, Edge A, and Heller S (2004) Stem cells as therapy for hearing loss. *Trends in Molecular Medicine* 10: 309–315.

Ma EY, Rubel EW, and Raible DW (2008) Notch signaling regulates the extent of hair cell regeneration in the zebrafish lateral line. *Journal of Neuroscience* 28: 2261–2273.

Matsui JI, Haque A, Huss D, et al. (2003) Caspase inhibitors promote hair cell survival and vestibular function following aminoglycoside treatment *in vivo*. *Journal of Neuroscience* 23: 6111–6122.

Meyers JR and Corwin JT (2007) Shape change controls supporting cell proliferation in lesioned mammalian balance epithelium. *Journal of Neuroscience* 27: 4313–4325.

Morest DK and Cotanche DA (2004) Regeneration of the inner ear as a model of neural plasticity. *Journal of Neuroscience Research* 78: 455–460.

Shou J, Zheng JL, and Gao W-Q (2003) Robust generation of new hair cells in the mature mammalian inner ear by adenoviral expression of *Hath1*. *Molecular and Cellular Neuroscience* 23: 169–179.

White PM, Doetzlhofer A, Lee YS, Groves AK, and Segil N (2006) Mammalian cochlear supporting cells can divide and trans-differentiate into hair cells. *Nature* 441: 984–987.

Zakir M and Dickman JD (2006) Regeneration of vestibular otolith afferents after ototoxic damage. *Journal of Neuroscience* 26: 2881–2893.

Hair Cells: Sensory Transduction

G S G Géléoc and J R Holt, University of Virginia, Charlottesville, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Structure of Sensory Epithelia

In each of the organs of the vertebrate inner ear, hair cells occur in specialized patches or strips of sensory epithelium, which generally contain just hair cells and supporting cells. Because the organs are formed during development by the infolding of an otic cyst, derived from surface ectoderm, the apical surfaces of hair cells face a closed compartment. Specialized epithelial cells supply this chamber with a unique fluid, termed 'endolymph,' which has a high potassium concentration and low sodium and calcium concentrations. Tight junctions ringing the apical surfaces of hair cells and supporting cells separate the endolymph from perilymph, a fluid similar to normal saline that bathes the basolateral surfaces of hair cells. In the mammalian cochlea, proton pumps in epithelial cells of the stria vascularis raise the potential of endolymph to about + 80 mV; elsewhere, however, the endolymph potential is within a few millivolts of zero.

Structure of the Hair Bundle

The mechanosensitive hair bundle comprises a few tens to a few hundreds of modified microvilli, termed 'stereocilia,' arranged in multiple rows of increasing height (Figures 1(a) and 1(b)). Adjacent stereocilia adhere to each other by means of interciliary links (Figures 1(c)–1(e)), so that when the bundles are deflected the stereocilia do not separate, but instead slide past one another: ankle links connect the stereocilia bases and lateral links couple the shafts of the stereocilia (Figure 1(c)), and tip links extend from the apical tip of a stereocilium to the side of its tallest neighbor (Figures 1(d) and 1(e)). The stereocilia have an internal structure similar to microvilli: they are bundles of several hundred actin filaments, tightly cross-linked by actin-bundling proteins espin and fimbrin. While ranging in length from about 1 μm (in mammalian cochlea) to over 60 μm (in semicircular canals), stereocilia have such stiff actin cores that they do not normally bend. The cores taper at their bases to a few actin filaments, which anchor the stereocilia into a cuticular plate just beneath the apical surface (Figure 1(f)). Since stereocilia have rigid cores and taper at their bases, they do not bend but instead pivot around their insertion point. Many hair bundles also contain a single, eccentrically placed true cilium, the kinocilium, which, in mature cells, has the 9 + 2 arrangement of microtubules characteristic of motile cilia (Figure 1(g)). The kinocilium may serve to organize the bundle during development (it is lost in mature hair bundles of the mammalian cochlea), and it couples the stereocilia to accessory structures in most other organs.

Axis of Polarity

The staircaselike variation in heights of stereocilia (Figures 1(a) and 1(b)) and the eccentric placement of the kinocilium at one edge of the bundle (Figure 1(g)) confer a morphological axis of polarity to the bundle. The positive direction is defined as toward the kinocilium or the tallest stereocilia (Figure 2(a)); negative is toward the shortest stereocilia. This also describes the physiological axis of the bundle: a positive deflection of the stereocilia evokes an excitatory response. It increases the membrane conductance by triggering the opening of nonselective cation channels (transduction channels), allowing positively charged ions to flow into the cell and thereby depolarize the cell. At the bundle's resting position, ~10% of the transduction conductance is active (Figure 2(b)). Thus, a negative deflection decreases the membrane conductance (closes channels open at rest) and hyperpolarizes the cell. This feature allows hair bundles to signal stimuli of either polarity.

The sensitivity of the hair bundle to mechanical stimuli is remarkable: a deflection of a few tenths of a micrometer in the positive direction is sufficient to activate the conductance completely, and deflections at the auditory threshold are estimated to be a few tenths of a nanometer. Side-to-side deflections of the bundle, even several micrometers, cause no conductance change, so that the cell has a vector sensitivity and only senses the component of the stimulus along the bundle's axis of polarity.

The Mechanism of Transduction

The past quarter century has provided a good understanding of the mechanoelectrical transduction process, whereby deflections of the stereocilia cause the opening of transduction channels. The process is extremely fast: at mammalian temperatures, channels begin to open within about 10 μs after deflection of the bundle. This can explain a sensitivity for frequencies as high as 100 kHz in some animals, but the speed also rules out transduction mechanisms involving an

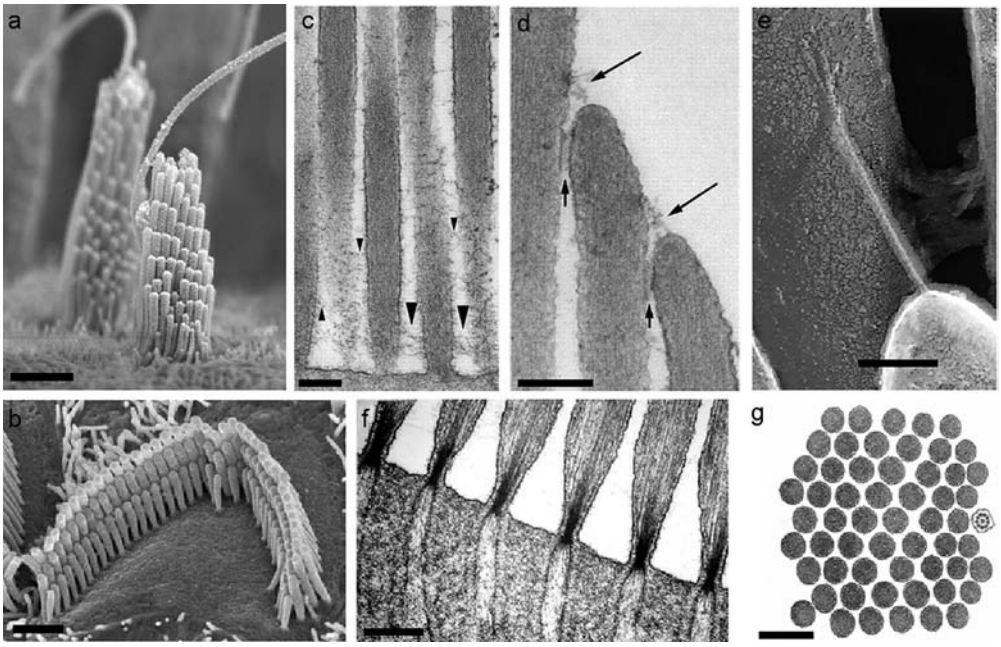

Figure 1 The structure of the vertebrate hair bundle. (a) Scanning electron micrograph of a hair bundle protruding from the apical surface of a mouse utricle hair cell. Utricular hair bundles are composed of ~65 stereocilia of graded heights and a single taller kinocilium. (b) Scanning electron micrograph of a guinea pig cochlea hair bundle. While a kinocilium is present in immature cochlear hair cells, it is lost after ~postnatal day 3. (c) Transmission electron micrograph of a chick utricle hair bundle. Just above the cuticular plate, the stereocilia are interconnected by ankle links (large arrowheads). Shaft connectors (small arrowheads) link the lateral aspects of the stereocilia. (d) At the tip of each stereocilium are tip links (long arrows) and horizontal top connectors (short arrows). (e) High-resolution freeze-etch image of a tip link from a guinea pig cochlea hair bundle. (f) Transmission electron micrograph of a longitudinal section of the apex of a hair cell from the saccule of a goldfish. The stereocilia taper at their bases and contain a large number of tightly packed actin filaments that converge into a dense bundle where they insert into the cuticular plate. (g) Thin cross-section of a hair bundle from the bullfrog saccule, which reveals the hexagonal arrangement of stereocilia with cores of actin and the single, microtubule-based kinocilium at the right. Scale bar = 5 μm (a), 1 μm (b), 200 nm (c, d, f), 100 nm (e), 500 nm (g). (a) Reproduced from Holt JR, Gillespie SKH, Provance DW, et al. (2002) A chemical strategy implicates myosin-k in adaptation by hair cells. *Cell* 108: 371–381, with permission of Elsevier. (b) Reproduced from Furness DN and Hackney CM (2006) *Springer Handbook of Auditory Research: Vertebrate Hair Cells*. Heidelberg: Springer, with permission of Springer-Verlag, Heidelberg. (c, d) Reproduced from Bashtanov ME, Goodyear RJ, Richardson GP, et al. (2004) The mechanical properties of chick (*Gallus domesticus*) sensory hair bundles: Relative contributions of structures to calcium chelation and subtilisn treatment. *Journal of Physiology* 559(1): 287–299, with permission of Blackwell Publishing. (e) Adapted from Kachar B, Parakkal M, Kurc M, Zhao Y, and Gillespie PG (2000) High-resolution structure of hair-cell tip links. *Proceedings of the National Academy of Sciences of the United States of America* 97: 13336–13341, with permission from PNAS. Copyright (2000) National Academy of Science, USA. (f) Reproduced from Fawcett DW (1981) *The Cell*, Philadelphia: WB Saunders, with permission of Elsevier. (g) Reproduced from Hudspeth AJ (1983) The hair cells of the inner ear. *Scientific American* 248(January 1983): 54–64, with permission of Dr. AJ Hudspeth and *Scientific American*.

enzyme cascade or second messenger molecules. It was proposed, instead, that deflections exert a force directly on transduction channels to open them, much as transmembrane voltage exerts a force directly on charged domains of voltage-gated channels. Several experiments have succeeded in measuring small movements of hair bundles that occur with channel opening. These experiments provide strong support for the idea of directly gated channels, as well as estimates of some of the mechanical properties of the transduction apparatus. The mechanical force required to open a single transduction channel is estimated to be about 2 pN and upon opening, the channel protein undergoes a conformational change of ~2 nm. Mechanosensitive transduction channels represent a third class of ion channel, in addition to those gated by voltage and ligand binding. A variety of mechanically gated channels have been identified in other systems. Properties of the hair cell transduction channel resemble those of ion channels of the transient receptor potential (TRP) family, inparticular TRPN1 (also known as NOMPC, absent from the mammalian genome) present in the *Drosophila* bristle. Recently, TRPA1 was presented as a candidate for the mechanosensitive channel of vertebrate hair cells, but lack of an auditory or vestibular deficit in TRPA1 knockout mice has called that suggestion into question. Thus, conclusive evidence of the identity of the hair cell transduction channel awaits.

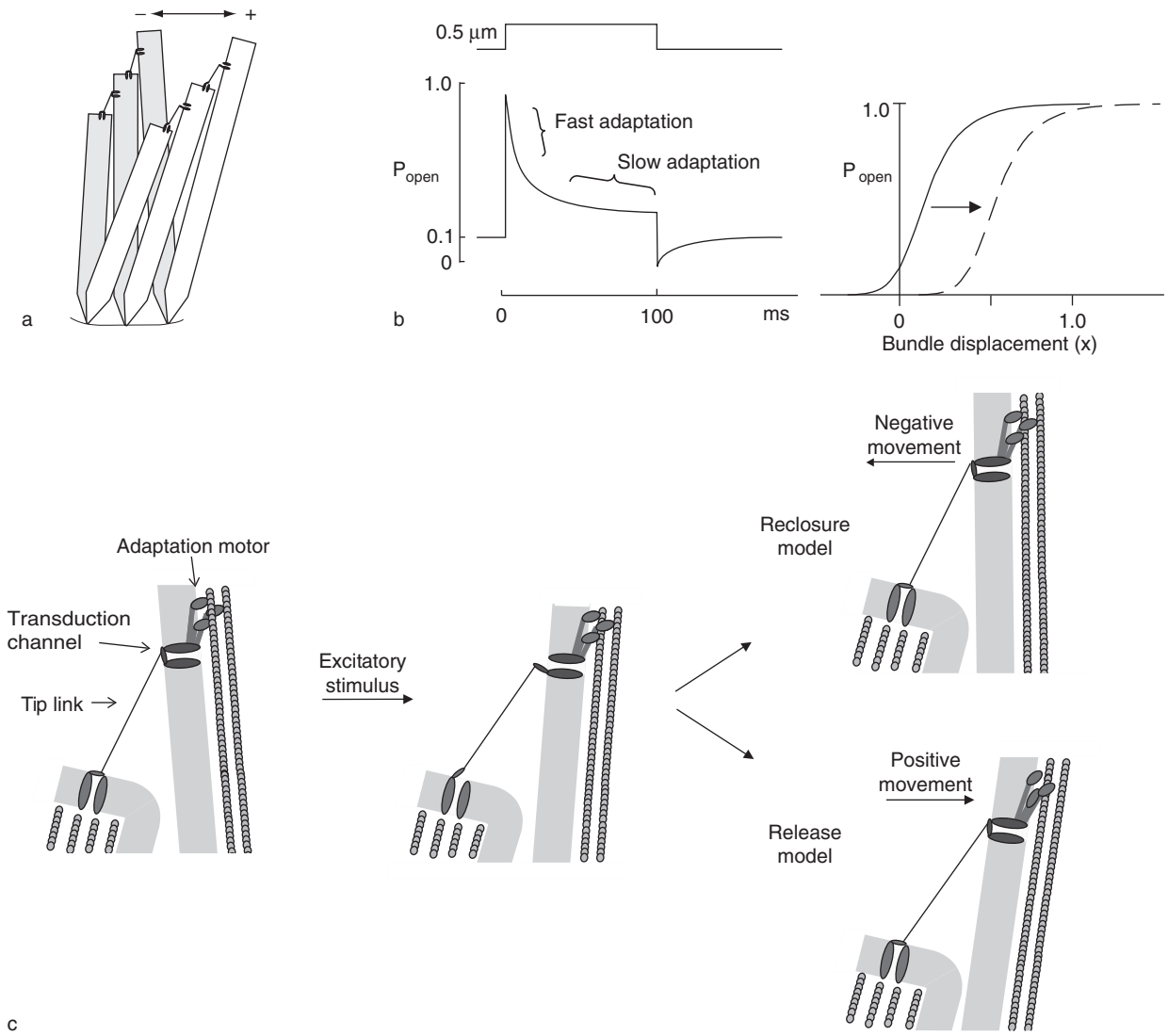

Figure 2 Models for hair cell transduction and adaptation. (a) Deflection of the hair bundle toward the tallest stereocilia opens mechanosensitive transduction channels. (b) A 0.5 μm step deflection rapidly opens channels. Over the subsequent 100 ms the open probability declines or adapts to a new steady-state level. The stimulus response relationship (right) shifts in the direction of the applied stimulus. In the example of a positive stimulus the curve shifts to the right (dashed line). The relationship is best described by a second-order Boltzmann function. (c) The prevailing model for adaptation posits that following an excitatory stimulus calcium enters open transduction channels and promotes a decline in open probability on two timescales. Fast adaptation occurs within the first few milliseconds and may result in change in the relationship between tension and open probability such that the channels are forced closed (reclosure model). This force is predicted to generate a negative movement in bundle position. Alternatively, fast adaptation may result from a release of tension within the transduction complex, which allows channels to close and the bundle to move farther in the positive direction (release model). Slow adaptation (not shown) follows fast adaptation and results from slipping of the adaptation motor down the side of the taller stereocilium, which also decreases tension and allows the channels to close. (b) Adapted from Pickles JO and Corey DP (1992) Mechanoelectrical transduction by hair cells. *Trends in Neuroscience* 15: 254–259.

Hair cell transduction channels are not particularly selective, passing all of the alkali cations, many divalent cations (including calcium), and small organic cations up to about 0.7 nm in diameter. Thus, the reversal potential is close to 0 mV. Taken together with a single-channel conductance of 90–100 pS and estimates of up to 200 channels per cell (1 or 5 per stereocilium), deflections of the hair bundle can evoke transduction currents of up to ~1 nA. Several independent experiments, including measurement of extracellular current and focal application of channel blockers, have localized transduction channels to the tips of stereocilia. This was confirmed by imaging calcium that flows into the cell through the transduction channels. Intracellular calcium was seen to rise first at the tips of the stereocilia and then more slowly

along their lengths. Furthermore, these experiments concluded that the channels could be found at either end and probably at both ends of tip links.

Hair cell transduction channels are opened by force applied to elastic elements, termed 'gating-springs,' that are stretched by positive deflections (toward the tallest stereocilia) of the hair bundles. Tip links seem ideally positioned to be gating-springs because they run parallel to the staircase arrangement of the hair bundle and join the top of one stereocilium to the side of its adjacent taller neighbor along the axis of polarity. The discovery that breaking tip links, by treatment with low concentrations of extracellular calcium, abolishes transduction provides compelling evidence that tip links are an essential element. However, the concept that tip links act as the gating-springs has recently been challenged. High-resolution images have revealed that tip links are rigid extracellular filaments, ~10 nm in diameter and 150–200 nm long (Figure 1(e)). They appear to consist of two to three coiled fibers with a helical period of ~25 nm and multiple globular domains ~4 nm in diameter. These measurements are consistent with the molecular dimensions of cadherin 23 (CDH23), which has 27 cadherin domains. CDH23 has been immunolocalized to the hair bundle and mutations in the gene that encodes CDH23 cause deafness and balance disorders in mice and humans. However, the notion that CDH23 is a component of the tip link remains controversial because hair bundle immunolocalization is lost in adult mice despite the persistence of tip links, which suggests that CDH23 may play a role in hair bundle development instead. Nonetheless, the rigid structure of the tip link suggests that the search for the molecular identity of biophysically defined gating-spring should be shifted to other elements, mechanically in series with the transduction channel.

The Process of Adaptation

Steady deflection of the bundle in the excitatory direction causes an initial activation of transduction channels, followed by a decline in the response toward the resting level as transduction channels close (Figure 2(b)). The instantaneous activation curve measured after 100 ms or so is shifted to the right, relative to its position at rest (Figure 2(b)). Tension on transduction channels relaxes during the deflection so that additional deflection can reactivate the channels. Similarly, a negative deflection that closes channels is followed by adaptation that increases the tension and shifts the activation curve to the left. This process, commonly referred to as adaptation, allows hair cells to retain high sensitivity over a broad operating range and filters out slower stimuli in favor of rapid stimuli. The current decline to constant stimuli typically exhibits a bimodal decay, which reveals two rates of adaptation: fast adaptation occurs within a few milliseconds or less while slow adaptation is about 10 times slower. The two forms of adaptation, although both dependent on calcium, reflect distinct mechanisms.

Circumstantial evidence has suggested that slow adaptation requires myosin molecules, perhaps in a cluster of a few dozen molecules. Members of the myosin family are the only molecules known to move on actin, and actin in the cores of stereocilia is polarized so that myosin would climb up to reset tension. The rate at which the climbing phase of adaptation (for negative stimuli) occurs and the rate at which muscle myosin-coated beads climb up the actin core are 1–2 $\mu m\, s^{-1}$, similar to the rate of myosin in other systems. Compounds that block myosin–actin interaction abolish adaptation.

Like myosin–actin interactions in other systems, the hair cell adaptation mechanism is regulated by calcium and the calcium-binding protein, calmodulin. The increase in intracellular calcium concentration that results from open transduction channels increases the rate of adaptation. When calcium levels rise, the myosin motors dissociate from actin and slip down the side of the stereocilium, which in turn relieves tip-link tension and allows the channels to close.

Out of ~40 members of the myosin superfamily, immunolocalization data focused the search for the adaptation motor on myosin 1c (formerly known as myosin 1-beta). A role for myosin 1c in adaptation was confirmed physiologically, using a mouse model in which myosin 1c was mutated, sensitizing it to a chemical inhibitor. When the inhibitor was added, both slow and fast adaptations were abolished in mutant but not wild-type vestibular hair cells.

Several other members of the myosin superfamily, including myosin 7a and myosin 15, appear to be critical for normal hair bundle development and function. Naturally occurring mutations in myosin 7a cause altered hair bundle morphology and, as a result, auditory and vestibular deficits in mice and humans. Taken together with localization data, this suggests that myosin 7a may be involved in holding the stereocilia together and maintaining the hair bundle within the properly erect mechanosensitive range.

It should be noted that slow (>10 ms) myosin-based adaptation has been most fully characterized in vestibular hair cells of bullfrogs and mice. Fast adaptation has been well characterized in hair cells of the turtle auditory organ and recently in the rat cochlea. In these cells, fast adaptation is also calcium dependent and causes a similar shift of the activation curve. It has been suggested that calcium entering through an open transduction channel can bind to

and force reclosure of the channel (reclosure model: Figure 2(c)). A second model for fast adaptation has been proposed that suggests that entering calcium triggers the release of a mechanical element in series with the transduction channel (release model: Figure 2(c)). The release reduces tension to the gating-spring and allows the transduction channels to close. Although both models explain the rapid decay in transduction current, they make opposite predictions on the movement of the hair bundle that occurs during fast adaptation. The reclosure model predicts that channel reclosure will increase tension in the gating-spring, which will pull the bundle in the negative direction (to the left in Figure 2(c)). The release model predicts that there will be a reduction in gating-spring tension which will allow the bundle to move further in response to a constant force (to the right in Figure 2(c)). Recent data from mice and rats are consistent with the release model, whereas data from turtle and some from frog hair cells support the reclosure model. The models are not mutually exclusive and both mechanisms may coexist in the same cell or to varying degrees in hair cells of different organs and species.

Cochlear Amplification

Mechanical amplification of sound stimuli in mammals requires the activity of outer hair cells which reside in three rows along the entire length of the mammalian cochlea. The process is known as cochlear amplification and functions to enhance sensitivity to faint sounds, and to tune or sharpen fine pitch perception. Two mechanisms have been proposed to underlie cochlear amplification: somatic motility, driven by the hair cell receptor potential, and active movements of the hair bundle, associated with fast adaptation.

Somatic motility requires the voltage-sensitive membrane protein, prestin, which is abundant in the lateral membranes of mammalian outer hair cells but absent in hair cells of lower vertebrates. The somatic motility theory for cochlear amplification posits that sound stimuli generate receptor potentials that evoke voltage-dependent conformational changes in prestin such that the protein contracts with depolarization and expands with hyperpolarization. Because the collective motion of the densely packed molecules occurs parallel to the lateral membranes, outer hair cells shorten and elongate in time with changes in membrane potential. The force produced by somatic motility in isolated outer hair cells is about 100 pN mV^{-1}. Forces of this amplitude are predicted to generate synchronous motion of outer hair cells; this, in turn, amplifies the overall motion of the basilar membrane in phase with the sound stimulus. However, because somatic motility depends on the hair cell receptor potential, the membrane time constant will likely limit force production at the high stimulus frequencies to which auditory organs are known to be sensitive. In addition, because cochlear amplification also occurs in lower vertebrates that lack outer hair cells and prestin, a second mechanism has been invoked.

The other model suggests that the forces generated by the hair bundle during adaptation are sufficient to drive amplification of sound stimuli in the auditory organs of many different species. Although the mechanism of fast adaptation is controversial, there is agreement that fast adaptation can drive active hair bundle movements. The active bundle movements are on the order of tens of nanometers. If the hair bundle movements are in phase with the sound stimulus, their collective motion is also predicted to drive amplification of basilar membrane oscillations. To tease apart the relative contributions of somatic motility and hair bundle motility will require their selective inactivation in live animals.

The Receptor Potential

Mature hair cells lack voltage-gated sodium channels, therefore they do not generate action potentials. Rather, changes in transduction channel activity evoke graded receptor potentials of up to 40 mV. The receptor potential is modified by the activity of voltage- and ion-sensitive conductances in the basolateral membrane. In some hair cells depolarizing receptor potentials elicit oscillations of the membrane potential that may serve to tune the cell to a particular stimulus frequency. These oscillations result from an interplay between activation of α-1D calcium channels and BK calcium-dependent potassium channels. Variation in the kinetics of BK channels results from both alternative splicing and co-assembly with β subunits and gives rise to membrane potential oscillations that vary in frequency among hair cells of the same organ.

The ability of hair cells to signal stimuli of either polarity is preserved at the afferent synapse. At the cell's resting potential (−50 to −70 mV), voltage-sensitive calcium channels are tonically active. This results in a steady release of neurotransmitter, which in turn results in a steady background firing in the fibers of the eighth nerve. Thus, receptor potentials of either polarity modulate calcium channel activity and hence transmitter release. A presynaptic electron-dense body (~0.5 μm in diameter) at each release site, around which vesicles cluster, is reminiscent of the synaptic ribbon in photoreceptors and may serve a

similar role in mediating continuous release. Neurotransmitter release is modulated with submillisecond precision to encode the frequency and intensity of sound and head movements. The excitatory amino acid glutamate has been shown to be released by hair cells, but it is not clear whether it is the primary transmitter or whether other transmitters are co-released.

The efferent synapse, carrying feedback control from the central nervous system (CNS) via the olivocochlear fibers, uses acetylcholine as the transmitter. The synapse may be excitatory in some hair cells, but in most organs efferent stimulation causes a slow inhibitory hyperpolarization. The inhibition is caused by activation of calcium-dependent potassium channels (SK). SK channels are voltage insensitive and can be activated by micromolar concentrations of calcium. The $\alpha 9$ and $\alpha 10$ nicotinic cholinergic subunits have been shown to form the receptor that mediates synaptic transmission between the efferent fiber and hair cells. The high degree of calcium permeability of this receptor may permit sufficient calcium entry to activate SK calcium-dependent potassium channels, and this, in turn, inhibits the cell.

See also: Cochlear Mechanics; Cochlear Development; Deafness; Hair Cell Regeneration; Hair Cell Differentiation; Sensory Aging: Hearing.

Further Reading

Bashtanov ME, Goodyear RJ, Richardson GP, et al. (2004) The mechanical properties of chick (*Gallus domesticus*) sensory hair bundles: Relative contributions of structures to calcium chelation and subtilisn treatment. *Journal of Physiology* 559(1): 287–299.

Eatock RA (2000) Adaptation in hair cells. *Annual Review in Neuroscience* 23: 285–314.

Eatock RA, Corey DP, and Hudspeth AJ (1987) Adaptation of mechanoelectrical transduction in hair cells of the bullfrog's sacculus. *Journal of Neuroscience* 9: 2821–2836.

Eatock RA, Fay RR, and Popper AN (eds.) (2005) *Springer Handbook of Auditory Research, Volume 27: Vertebrate Hair Cells.* New York: Springer.

Fawcett DW (1981) *The Cell.* Philadelphia, PA: WB Saunders.

Fekete DM (1999) Development of the vertebrate ear: Insights from knockouts and mutants. *Trends in Neurosciences* 6: 263–269.

Fettiplace R and Fuchs PA (1999) Mechanisms of hair cell tuning. *Annual Review of Physiology* 61: 809–834.

Fettiplace R and Hackney CM (2006) The sensory and motor roles of auditory hair cells. *Nature Reviews Neuroscience* 7: 19–29.

Fettiplace R, Ricci AJ, and Hackney CM (2001) Clues to the cochlear amplifier from the turtle ear. *Trends in Neurosciences* 24: 169–175.

Furness DN and Hackney CM (2006) *Springer Handbook of Auditory Research: Vertebrate Hair Cells.* Heidelberg: Springer.

Géléoc GS and Holt JR (2003) Auditory amplification: Outer hair cells press the issue. *Trends in Neuroscience* 26: 115–117.

Holt JR and Corey DP (2000) Two mechanisms for transducer adaptation in vertebrate hair cells. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11730–11735.

Holt JR, Gillespie SKH, Provance DW, et al. (2002) A chemical strategy implicates myosin-K in adaptation by hair cells. *Cell* 108: 371–381.

Hudspeth AJ (1983) The hair cells of the inner ear. *Scientific American* 248(January 1983): 54–64.

Hudspeth AJ (2005) How the ear's works work: Mechanoelectrical transduction and amplification by hair cells. *Comptes Rendus Biologies* 328: 155–162.

Hudspeth AJ, Choe Y, Mehta AD, et al. (2000) Putting ion channels to work: Mechanoelectrical transduction, adaptation, and amplification by hair cells. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11765–11772.

Kachar B, Parakkal M, Kurc M, Zhao Y, and Gillespie PG (2000) High-resolution structure of hair-cell tip links. *Proceedings of the National Academy of Sciences of the United States of America* 97: 13336–13341.

LeMasurier M and Gillespie PG (2005) Hair-cell mechanotransduction and cochlear amplification. *Neuron* 48: 403–415.

Parsons TD and Sterling P (2003) Synaptic ribbon. Conveyor belt or safety belt? *Neuron* 37: 379–382.

Pickles JO and Corey DP (1992) Mechanoelectrical transduction by hair cells. *Trends in Neuroscience* 15: 254–259.

Strassmaier M and Gillespie PG (2002) The hair cell's transduction channel. *Current Opinion in Neurobiology* 12: 380–386.

Wu YC, Ricci AJ, and Fettiplace R (1999) Two components of transducer adaptation in auditory hair cells. *Journal of Neurophysiology* 82: 2171–2181.

Headache

D G Jamieson, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Headaches are a universal human experience; almost everyone has had a headache at some time. Chronic headaches are one of the most common patient complaints in a general medical, and certainly in a neurological, practice. Although most headaches cause only self-limited head, neck, or face pain, some headaches may be a symptom of a life-threatening vascular, anatomic, or infectious disorder. In 2004 the International Headache Society (IHS) published definitions of headache types in an effort to standardize headache diagnosis and to improve recognition and treatment. The second edition of the International Classification of Headache Disorders (ICHD-II) divided headaches into primary and secondary types, with multiple subcategories within each type.

Primary Headaches

Primary headaches, the most common headache type, are not caused by anatomic, inflammatory, infectious, or physiological abnormalities. Approximately 98% of patients who present for medical evaluation will have a type of primary headache. The three major primary headache categories are migraine headache (with or without aura), tension-type headache (TTH), and cluster headache or other trigeminal autonomic cephalalgias, as described in Table 1. Other more unusual types of primary headaches occur less frequently, but are especially common in individuals who are prone to migraine headaches. The recognition of the primary headache type is useful for identifying the cause of the headache and directing appropriate treatment. However, the headache-prone patient may suffer from multiple primary headache types, which often respond to the same acute and preventative medications. Also, as useful as this categorization is in diagnosing headaches, patients' descriptions of their pain and associated symptoms may overlap the different categories.

Migraine

Most patients who seek medical attention for chronic headaches have migraine headaches. Almost 28 million Americans (21 million women and 7 million men) have frequent or severe migraines. Half of these individuals have functional disability as a result of these headaches. The American Migraine Prevalence and Prevention (AMPP) study estimated a 1-year period prevalence of 11.7% (5.6% for men and 17.1% for women) for migraine according to the strict migraine diagnostic criteria of ICHD-II. Headaches span the human life cycle, with young children and centenarians suffering from migraine headaches; however, the peak age for migraines is 25–55 years. According to the AMPP study, approximately a quarter (24.4%) of all women aged 30–39 years suffer from migraine headaches. The study found that migraine was more prevalent in women, in Caucasians more often than in African-Americans, and in those of lower, as compared to higher, incomes.

Migraine is a disabling, neurobiological disorder caused by complex and incompletely understood pathophysiological mechanisms. The majority of patients with migraine have a family history of headaches, inherited in an autosomal dominant pattern with unclear genetic linkage. A rare subtype of migraine with aura, familial hemiplegic migraine, has been mapped to specific genes, including a mutation of a P/Q voltage-gated Ca^{2+} channel gene (*CACNA1A*) on chromosome 19p13. A channel disturbance may be linked to the migraine aura, which is triggered by a reduced threshold for cortical spreading depression, the neuronal depolarization that may explain the evolution of the aura over time and space. The cranial vessels and their innervation through the trigeminal vascular system mediate migraine pain, which is characteristically throbbing and often unilateral. Stimulation of the dorsal horn neurons of the upper cervical ganglia, as well as of the trigeminal nucleus caudalis, explains the posterior head and neck pain frequently noted by migraine sufferers. The dura is innervated by the nerves of the ophthalmic (V-1) division of the trigeminal nerve, with trigeminal ganglion stimulation resulting in release of inflammatory neuropeptides and dilation of meningeal vessels. Interest in blocking the action of these neuropeptides has been disappointing for substance P antagonists, but calcitonin gene-related peptide antagonists show promising initial results in treating acute migraine. During a migraine attack, patients have a lower threshold for cortical activation (explaining migraine aura) and sensory stimuli (explaining sensitivity to light, sound, and smell). Patients also report that ordinarily innocuous stimuli can be perceived as painful, the definition of allodynia. Activation of first-order trigeminal sensory neurons (peripheral sensitization of dural nociceptors) causes the throbbing pain and exacerbation of head pain with movement. Cutaneous

Table 1 ICHD-II criteria for primary headaches

Migraine without aura
A. At least five attacks fulfilling B–D
B. Attacks lasting 4–72 h
C. At least two of the following four characteristics:
  1. Unilateral location
  2. Pulsating quality
  3. Moderate or severe intensity
  4. Aggravation by physical activity
D. Accompanied by at least one of the following signs:
  1. Nausea or vomiting, or both
  2. Photophobia and phonophobia

Episodic tension-type headache
A. At least ten episodes occurring on <1 day per month on average (<12 days per year) and fulfilling criteria B–D
B. Headache lasting from 30 min to 7 days
C. Headache has at least two of the following characteristics:
  1. Bilateral location
  2. Pressing/tightening (nonpulsating) quality
  3. Mild or moderate intensity
  4. Not aggravated by routine physical activity such as walking or climbing stairs
D. Both of the following:
  1. No nausea or vomiting
  2. No more than one of photophobia or phonophobia
E. Not attributed to another disorder

Cluster headache
A. At least five attacks fulfilling criteria B–D
B. Severe or very severe unilateral orbital, supraorbital, and/or temporal pain lasting 15–180 min if untreated
C. Headache is accompanied by at least one of the following:
  1. Ipsilateral conjunctival injection and/or lacrimation
  2. Ipsilateral nasal congestion and/or rhinorrhoea
  3. Ipsilateral eyelid edema
  4. Ipsilateral forehead and facial sweating
  5. Ipsilateral forehead miosis and/or ptosis
  6. A sense of restlessness or agitation
D. Attacks have a frequency from one every other day to eight per day
E. Not attributed to another disorder

From Headache Classification Committee of the International Headache Society (IHS) (2004) International Classification of Headache Disorders II. *Cephalalgia* 24(supplement 1): 9–160.

allodynia, present to some degree in up to 80% of patients with migraine, involves central sensitization of second-order neurons of the trigeminal nucleus caudalis and potentially the third-order neurons in the thalamus. Central modulation of migraine pain has been shown in positron emission tomography studies to occur in the dorsal midbrain, including the periaqueductal gray, and in the dorsal pons near the locus coeruleus. Stimulation of these and other pontine and midbrain regions may mediate cerebral blood flow, facilitating trigeminovascular pain processing. Pain relay networks connect the brain stem nuclei and the upper cervical ganglia to thalamic nuclei, particularly in the ventroposteromedial thalamus.

Migraine headaches may be triggered in susceptible individuals by environmental stimuli that are innocuous to nonmigraine sufferers. The most common migraine triggers include change in sleep habits (too much or too little sleep), stress or relief from stress, alcohol, change in weather, and hunger and/or dehydration. Some foods may trigger some headaches in some people at some time, but in general strict food avoidance has little effect on migraine attacks. The most common food triggers include processed or cured meats, moldy cheese, artificial sweeteners, and monosodium glutamate. In many women, their menstrual period is a very strong trigger for their most severe migraine headaches.

Women are particularly prone to migraine headaches, with different susceptibilities to headaches throughout their life influenced by hormonal (estrogen) changes. Migraines may start during puberty, at onset of menses. About 70% of women with migraine will have an exacerbation of their headaches associated with their menstrual cycle, with most of these menstrual migraines occurring in the 2–3 days of decreasing estrogen levels prior to menstrual flow. Migraines may occur in the first or second trimesters of pregnancy, but they improve as pregnancy progresses parallel with increasing estrogen levels. The general improvement in migraines in pregnancy, especially in the later months, is fortuitous, as absolutely risk-free options for treatment with acute and preventative medications during pregnancy are limited. About 40% of women will report their first migraine during pregnancy or shortly after delivery. Migraines may worsen after delivery as estrogen levels drop along with sleep deprivation and stress. Many women report an increase in frequency and severity of their headaches at the onset of perimenopause, but headaches generally lessen significantly after menopause and migraines are rare in elderly women and men.

The linkage between migraine and vascular disease is complex, as about one in three adults has some form of cardiovascular or cerebrovascular disease, and migraine afflicts a large number of adults who may also be at vascular risk. The relationship is more than a coincidental overlap, as migraine represents a vascular risk for both ischemic stroke and myocardial infarction. The Women's Health Study found that women who had migraine with aura had an increased risk of vascular events, including ischemic stroke, myocardial infarction, angina, and vascular death. This increased risk was not found in women who had migraine without aura. The highest association between migraine with aura and ischemic stroke was in younger women. This study confirmed that women with migraine with aura have a greater risk of cardiovascular, as well as cerebrovascular, disease; however the precise mechanisms to explain this association are

currently unknown. Since migraine without aura is far more common than migraine with aura, a majority of migraine patients are not at increased vascular risk. This mildly increased vascular risk should be considered when counseling and treating patients with migraine with aura.

Patients with chronic headaches often describe that they have sinus headaches; however, chronic headaches due to recurrent cranial sinus infections are rare, and nasal polyps and deviated septa do not cause chronic headaches. Headaches due to acute bacterial infections of the ethmoid or maxillary sinuses cause perinasal, periocular, or frontal facial pain with purulent nasal discharge. Sinus headaches occur with fever and evidence of significant sinus disease on imaging of the cranial sinuses. The great majority of patients with recurrent, intractable 'sinus headaches' have migraine or other primary headaches. Nasal congestion, rhinorrhea, and facial pain are common associated symptoms of migraine headaches. Change in weather and environmental allergies, often misperceived to be associated with chronic sinus infections, are common migraine triggers.

Migraine with aura Most migraines are not accompanied by any focal neurological symptoms; however, about 30% of migraine sufferers have some headaches that are presaged by a neurological aura which is most commonly visual. Focal neurological symptoms (visual disturbances, numbness, unilateral weakness, speech difficulty) occur prior to or during the head pain of migraine with aura. The visual aura generally lasts less than 60 min, with positive (photopsias, wavy, or zigzag lines) or negative (scotomata, tunnel vision) visual phenomena often described as lasting 15–30 min. The most commonly described visual aura is binocular alteration in vision with a scintillating scotoma in the peripheral vision that increases in size over time, and then disappears. There may be a symptom-free period of up to 60 min before the contralateral headache pain begins.

Probable migraine Most migraine suffers have some migraine-like headaches that do not conform to the strict IHS criteria for migraine, with or without aura. These probable migraine headaches respond to migraine treatment, both acute and preventative. The pain of a migraine headache may be sharp or pressure-like, rather than throbbing, or it may be bilateral, rather than unilateral. Nausea and vomiting, as well as sensitivity to light, sound, and movement, may not be present with every migraine headache in every migraine sufferer. The AMPP study found that probable migraine, defined as migraine fulfilling all but one criterion for migraine with or without aura, was a frequent, undertreated, and sometimes disabling condition, found more often in older men and women, ages 30–59 years. In contrast to migraine, probable migraine was more prevalent in African-Americans than in Caucasians in the AMPP study.

Cluster Headache and Other Trigeminal Autonomic Cephalalgias

Cluster headaches occur much more frequently in middle-aged men than in women, and may be associated with chronic smoking and excess alcohol use. Cluster headaches characteristically occur in the early morning hours, awakening the man from sleep with excruciating pain and autonomic symptoms. As compared to migraine suffers, who seek quiet and isolation during a migraine attack, men with cluster tend to be agitated or even violent. The attacks tend to occur daily, once to multiple times a day, for a period of weeks to months, and then they may remit for months to years. The attacks of cluster pain are relatively short in duration, but the pain is very intense, and the attacks are best treated with preventative medication, such as verapamil, steroids, or topiramate. Acute attacks may be treated with nasal oxygen, subcutaneous sumatriptan, and nasal triptans.

Other headache types present as attacks with pain and associated symptoms similar to those of cluster headaches, but are shorter lasting and more frequent. They occur more commonly in females than in males, and may respond to indomethacin. These headaches include episodic and chronic benign paroxysmal hemicrania. The multiple daily attacks of paroxysmal hemicrania are severe, unilateral, orbital, supraorbital, or temporal pain lasting 2–30 min. They are generally accompanied by conjunctival injection, nasal congestion, lid swelling or ptosis, or facial swelling. Paroxysmal hemicrania characteristically responds to daily indomethacin prophylaxis.

A syndrome with short-lasting attacks of unilateral pain lasting for 5–240 s, occurring from 3 to 200 times per day, is called SUNCT (short-lasting unilateral neuralgiform headache attacks with conjunctival injection and tearing). The pain is stabbing or pulsating in an orbital, supraorbital, or temporal location and may be refractory to acute and preventative treatment.

Tension-Type Headache

The TTH is the most common type of primary headache, with the lifetime prevalence in the general population of 30–78%. Most patients treat their TTH with over-the-counter medications and do not seek medical attention, but the disability from these headaches can be severe. There may be a neurobiological basis

for the more severe subtypes of TTH, but the exact pathophysiological mechanism of these headaches is not known. TTHs may be episodic or chronic. Diagnoses of chronic TTH and episodic TTH are based on the same criteria, except that the chronic TTH occurs ≥15 days per month on average for >3 months (≥180 days per year). Chronic tension-type headaches may evolve over time from episodic TTHs, especially with the overuse of acute pain medications.

Chronic Daily Headaches

Chronic daily headaches diminish the quality of life for patients and their families and pose a therapeutic challenge to physicians. A chronic daily headache is a headache of any type, either primary or secondary, that occurs for at least 15 days per month for longer than a month. Primary chronic daily headaches, which afflict up to about 5% of the population, are divided into short-duration (<4 h) and long-duration types. The most common long-duration types are chronic TTH, chronic migraine, and transformed migraine. The IHS classification emphasizes the increasing frequency of these headaches, which segue from episodic to chronic. Headache attributed to medication overuse, an important cause of primary chronic daily headache, can occur with excess and prolonged use of acute pain medication. The ICHD-II definition of this chronic headache type, which may show dramatic improvement with discontinuation of acute pain medications, is shown in Table 2. The use of a daily preventative medication, with judicious use of acute pain medication, can decrease primary chronic daily headaches, which may then evolve back to episodic headaches. The treatment of secondary chronic daily headaches depends on the underlying cause of the headaches.

Table 2 Headache attributed to medication overuse

A. Headache present on >15 days, fulfilling criteria C and D
- Characteristics of headache depend on drug

B. Regular overuse of a medication for >3 months
- Amount of overuse depends on drug
- Ergotamine, triptans, opioids, and combination analgesics, ≥10 days per month
- Over-the-counter analgesics, >15 days per month

C. Headache has developed or markedly worsened during medication overuse

D. Headache resolves or reverts to its previous pattern within 2 months after discontinuation of overused medication

E. Over the counter

From Headache Classification Committee of the International Headache Society (IHS) (2004) International Classification of Headache Disorders II. *Cephalalgia* 24(supplement 1): 9–160.

Other Primary Headache Types

Other primary headache types include heterogeneous, poorly understood, headaches. When the pain is identified as due to one of these relatively rare primary headaches, the sufferer can be reassured and appropriately evaluated and treated. Primary stabbing headaches consist of multiple episodes of transient stabbing pain, lasting seconds, in the distribution of the ophthalmic division (V–1) of the trigeminal nerve. Short-lasting stabbing pains are commonly described by patients with migraine. Exertion, such as exercise, cough, Valsalva, or orgasm, can precipitate a headache in patients who may or may not have a prior headache history. These headaches may be responsive to indomethacin or they remit on their own. Hemicrania continua, a unilateral continuous headache, with ipsilateral conjunctival injection, nasal congestion, or ptosis, is characteristically responsive to indomethacin. A new daily-persistent headache produces bilateral, mild/moderate pain which is daily and unremitting for greater than 3 months after its onset.

Secondary Headaches

Most secondary headaches are due to distinct and known anatomic, physiological, or infectious causes. While they are much less common than primary headaches, secondary headaches should be considered in patients who present for evaluation of either acute or chronic headaches. Patients who have a primary headache disorder may also develop unrelated secondary headaches, so primary headache patients should be questioned about a marked change in headache character or frequency that could indicate a new superimposed headache type. There are many types of secondary headaches, only some of which are mentioned here. Some of the major secondary headaches types according to the ICHD-II criteria are listed in Table 3.

Cerebrovascular disease is a major cause of secondary headaches, as outlined in Table 4. In a patient with ischemic stroke the complaint of headache is usually less prominent than the focal neurological symptoms. However, in a patient with a cerebellar infarct, a headache with nausea and vomiting may be a prominent symptom at presentation or it may be noted with the development of cerebellar edema and hydrocephalus. Although the headache due to a subarachnoid hemorrhage is most often described as being excruciatingly severe, a headache due to aneurysmal rupture is most remarkable for being maximally severe at onset, as opposed to the more gradual onset of pain with most other headache types.

Table 3 Secondary headaches

- Head and neck trauma (posttraumatic, intracranial hematoma, postcraniotomy)
- Vascular disorder (ischemic stroke, intracranial hemorrhage, vascular malformation, dissection, arteritis, venous thrombosis)
- Nonvascular intracranial disorders (increased or decreased cerebrospinal fluid pressure, inflammatory disease, neoplasm)
- Substance use or withdrawal (medications, illicit drugs, alcohol, toxins)
- Infection (intracranial infection, systemic infection, HIV/AIDS)
- Disorders of homeostasis (hypoxia, hypercarbia, dialysis, hypertension, hypothyroidism)
- Headaches and facial pain due to facial structures (cranial bones, neck, eyes, ears, temporomandibular joint, sinuses)
- Psychiatric disease (somatization disorder, psychotic disorder)
- Cranial neuralgias (trigeminal and other neuralgias, cranial nerve compression, central causes of facial pain)

From Headache Classification Committee of the International Headache Society (IHS) (2004) International Classification of Headache Disorders II. *Cephalalgia* 24(supplement 1): 9–160.

Table 4 Secondary headaches due to cerebrovascular disease

Subarachnoid hemorrhage (SAH)
Risk: women > men; cigarette smoking, OCP, head trauma, anticoagulation
Pathophysiology: ruptured cerebral aneurysm, trauma, venous hemorrhage
Symptoms: sudden onset of moderate/severe headache; maximal pain at onset, syncope, seizures, coma
Diagnosis: CT scan (<10% negative for blood), LP, MRI/MRA, CTA, angiogram
Treatment (very high morbidity/mortality): nimodipine, clip/coil aneurysm, ↑ BP, ↑ fluids
Warning: have a low threshold to investigate a sudden, severe headache to rule out SAH

Cerebral venous thrombosis (CVT) (sagittal sinus thrombosis)
Risk: women > men, head trauma, dehydration, infection, thrombophilias, pregnancy, OCP, HIV
Pathophysiology: clotting of cerebral venous sinuses (sagittal sinus)
Symptoms: moderate/severe headaches, sudden-onset headaches/seizures, coma
Diagnosis: enhanced CT scan (empty delta sign); MRI/MRV, cerebral angiogram
Treatment: heparin, ↑ fluids, antibiotics, intrasinus thrombolysis
Warning: intracerebral hemorrhage is not a contraindication to anticoagulation with CVT

Intracerebral hemorrhage (ICH)
Risk: HTN, head trauma, anticoagulation, vascular malformations (AVM, cavernous malformation), amyloid angiopathy, ↑ bleeding due to medications/conditions
Pathophysiology: rupture of normal or abnormal blood vessels with hemorrhage in characteristic locations for HTN (basal ganglia, thalamus, cerebellum, pons) and cerebral amyloid angiopathy (lobar)
Symptoms: sudden-onset moderate/severe headache; neurological deficits, seizures, coma
Diagnosis: hyperdensity on unenhanced CT scan; MRI/MRA, cerebral angiogram
Treatment: supportive care, BP stabilization, surgical evacuation
Warning: presentation of an ICH can be identical to that of an ischemic stroke

Subdural hematoma (SDH)
Risk: head trauma, dural metastases, medications/conditions that ↑ bleeding, intracranial hypotension
Pathophysiology: rupture of dural bridging veins with hemorrhage into dural space
Symptoms: new-onset mild/moderate/severe headache; neurological deficits, seizures, coma
Diagnosis: concave hyperdensity on unenhanced CT scan; MRI/MRA
Treatment: surgical evacuation, supportive care (small)
Warning: bilateral isodense SDHs with symmetric mass effect can be occult on CT scan

Epidural hematoma (EDH)
Risk: head trauma, skull fracture
Pathophysiology: rupture of menigeal arteries with rapid hemorrhage above the dura space
Symptoms: new-onset moderate/severe headache; syncope, neurological deficits, seizures, coma
Diagnosis: convex (lens) hyperdensity on unenhanced CT scan; MRI/MRA
Treatment: immediate surgical evacuation
Warning: transient period of consciousness followed by loss of consciousness

AVM, arteriovenous malformation; BP, blood pressure; CTA, computed tomography angiogram; HTN, hypertension; LP, lumbar puncture; MRI, magnetic resonance imaging; MRA, magnetic resonance angiography; MRV, magnetic resonance venography; OCP, oral contraceptive pill.

The most important cause of a 'thunderclap headache' is aneurysmal subarachnoid hemorrhage, since timely aneurysmal obliteration can decrease the significant mortality due to aneurysmal rebleeding. Other causes of the sudden onset of severe headache without gradual buildup of pain include cerebral venous thrombosis, pituitary apoplexy, and intracerebral hemorrhage. Despite these dire causes of

Table 5 Headaches due to disorders of cerebrospinal fluid dynamics

Increased cerebrospinal fluid pressure headache (pseudotumor cerebri; idiopathic intracranial hypertension)
Risk: obesity (in women), steroids, cerebral venous thrombosis (CVT)
Symptoms: holocephalgic pain, pulsatile tinnitus, VI palsy, transient visual obscurations
Diagnosis: ↑ opening pressure on LP; papilledema; normal or 'slit-like ventricles' on CT
Treatment: acetazolamide, topiramate, ↓ weight; optic nerve decompression, serial LPs
Warning: untreated, can result in visual loss from optic nerve compression
Decreased cerebrospinal fluid pressure headache (spontaneous intracranial hypotension)
Risk: lumbar puncture, trauma, spontaneous
Pathophysiology: leakage of cerebrospinal fluid through dural tear
Symptoms: holocephalgia pain worse upright, improved recumbent
Diagnosis: enhancement of leptomeninges on magnetic resonance imaging, bilateral subdurals
Treatment: lumbar blood patch, time, hydration, caffeine
Warning: may take weeks to months to resolve; may be associated SDHs

LP, lumber puncture; SDH, subdural hematoma.

thunderclap headache, crash migraine is the most common cause of the sudden onset of a severe headache. Migraine variants such as primary cough headache, exertion-related headache, and sex-related headache can also present with the sudden onset of severe head pain.

The etiology of a secondary headache may be deduced based on the patient's history and physical examination. Body habitus may point toward a headache cause; idiopathic intracranial hypertension and obstructive sleep apnea may result in headaches associated with obesity. Disorders of cerebral spinal fluid (CSF) dynamics may cause characteristic headaches (Table 5). The distinctive feature of a headache due to intracranial hypotension is the prominent increase of pain while upright, with marked improvement in pain while recumbent. This headache, due to dural leakage of CSF, can occur spontaneously or after a lumbar puncture and may be associated with subdural hematomas. Brain tumors are generally heralded by focal neurological signs and symptoms. Central nervous system infections, including meningitis, encephalitis, and brain abscess, may present with isolated headache, but are most often accompanied by fever, an alteration in consciousness, and focal neurological signs and symptoms. When there is a suspicion of a secondary headache, neuroimaging with a magnetic resonance imaging (MRI) scan (or computed tomography scan if MRI is contraindicated) may indicate the diagnosis.

Facial Pain Syndromes

Some primary headaches present with pain confined to the face and periorbital area. Other types of facial pain are considered secondary headaches. While most facial pain is not caused by an underlying anatomic lesion, new-onset facial pain that does not fit into an idiopathic pain syndrome should be investigated for a benign or metastatic neoplasm, especially if there is a focal neurological deficit.

Trigeminal neuralgia (TN) causes abrupt, paroxysmal pain that occurs in the unilateral distribution of one of the three branches of the trigeminal nerve. Patients characteristically discover trigger areas on the face or activities such as eating or brushing teeth that precipitate flashes of sharp stabbing electrical like pain. The lacinating pain of TN occurs frequently for days to months and then it may remit spontaneously, only to return years later. Medication and rarely surgery can offer significant relief, although some patients may be particularly difficult to treat. Postherpetic neuralgia is another cause of facial pain in older individuals. The pain after a herpes zoster infection is generally constant, not episodic like TN, and occurs at the location of the preceding cutaneous lesion.

Temporomandibular disorders cause facial pain and may trigger or perpetuate headaches. Inflammation of the temporomandibular joint (TMJ), as well as displacement of the articular surface of the joint, may lead to local or referred facial pain. Indications of TMJ disease include facial pain precipitated by joint movement, TMJ pain on joint palpation, joint noise, and limited range of motion. Treatment may include use of a splint therapy, anti-inflammatory medications, relaxation therapy, and rarely surgery. The contribution of temporomandibular and dental disorders to chromic primary headaches is incompletely understood.

Headaches at Different Ages

Children

Primary headaches are generally an adult affliction, but children also suffer from chronic headaches. Children may have migraines, tension-type headaches, idiopathic stabbing headaches, or non-classifiable headaches. Boys are more likely than girls to have preschool onset of primary headaches including migraine, but they have a higher remission rate as they get older. Children with a preschool age onset of headache generally have a relatively benign prognosis for remission or improvement, with the type of primary headache predicting outcome. A child with tension-type or idiopathic stabbing headaches is more

likely to have remission with age than is a child with migraine. Children with migraine without vomiting are more likely to have fewer headaches as they grow up than will children who have vomiting with migraines. Family history of headache is common in children, with early-onset headaches in a child correlating with a history of an early onset of headaches in the child's parents.

Between 4% and 11% of children ages 5–15 years have migraines. From 1% to 3% of even younger children, ages 3–7 years, have migraines, with a slight male predominance. The distinctive migraine symptoms and syndromes in children do not fit easily into the criteria for diagnosis of migraine in adults. Younger children may have difficulty describing the quality and localization of the pain. The accompanying nonpain migraine symptoms, such as photophobia and phonophobia, may have to be inferred from the child's behavior. Preschool children have shorter duration migraine and the child's accompanying gastrointestinal symptoms, including nausea, intractable vomiting, and abdominal pain, may be more prominent than is the head pain. Head pain, generally unilateral in adults, is commonly bilateral in young children. Localization of the migraine head pain, which is relatively variable in adults, is usually frontotemporal in children. Pediatric migrainous disorders include cyclic vomiting syndrome, abdominal migraine, and benign paroxysmal vertigo, which may overlap with migraine headaches.

The impact of chronic headaches, including migraine, in children is significant, with missed school and social activities compromising quality of life. Pulling children out of school because of frequent absences due to headaches may have irreparable social and academic consequences. The accurate diagnosis of primary headache disorders in children, including preschoolers, can lead to improved treatment and decreased disability for these youngest of headache sufferers.

Older Adults

Although primary headaches tend to remit with age, certain headache types are characteristically seen in individuals in their 50s and older. Hypnic headache, an idiopathic headache disorder with a characteristic nocturnal pattern, is seen in older men and women. The headache occurs multiple times a month, awakening the patient from sleep with moderate to severe dull or throbbing pain. The pain is generally diffuse or frontotemporal and while it may be accompanied by nausea, it usually lacks the other associated features of migraine headaches or trigeminal-autonomic cephalalgias. Because the patient experiences the headache at the same time each night, the hypnic headache may be a rapid eye movement (REM) sleep disorder related to melatonin deficiency. The hypnic headache may be difficult to treat, although some patients experience relief with lithium, which may indirectly increase melatonin levels. Direct treatment with melatonin is of variable benefit.

Giant cell arteritis (GCA) is a cause of headache of variable description in patients older than age 50 years. The headache may be associated with systemic symptoms, including polymyalgia rheumatica, malaise, jaw claudication, and weight loss. Untreated GCA may progress to monocular or binocular visual loss due to involvement of posterior ciliary and central retinal arteries. Ischemic strokes occur rarely, generally in the posterior circulation. The erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) levels are generally elevated, but the diagnosis is confirmed by the granulomatous changes with multinucleated giant cells and lymphocytes seen pathologically on temporal artery biopsy. Prednisone to decrease the ESR, CRP, and headache and systemic symptoms may be needed for months to years, as relapses may occur with steroid tapering.

Headache Evaluation

The evaluation of a patient with headache depends on obtaining a detailed history of the headaches and other medical complaints. Table 6 outlines some of the points that may lead to the appropriate evaluation, diagnosis, and treatment of a patient who presents with headache. While most patients will have a benign cause of headaches, the clinician evaluating the patient should be aware of symptoms or signs that are suspicious for headaches that deserve a more detailed investigation and specific treatment. In general a focused neurological examination, concentrating on the mental status (ascertained during the history taking), cranial nerve evaluation including visualization of the optic nerve, strength and coordination testing, and muscle stretch reflexes will suffice to rule out an abnormal neurological examination due to a headache-related lesion. The majority of the time evaluating the patient should be spent obtaining a history, realizing that the neurologist will have more time to talk with the patient than will the primary care practitioner or emergency medicine physician.

Most patients who present with chronic primary headaches do not need to have a neuroimaging procedure. If the history and physical examination are consistent with migraine or tension-type headaches, without concerning aspects to the evaluation, then brain imaging is not needed. Table 7 lists some aspects of the evaluation that may indicate a secondary cause of headaches for which brain imaging should be

Table 6 Headache history

History	*Details/comments*
Age at onset	Childhood and teens (TTH, migraines); young adulthood (TTH, migraines); elderly (GCA, hypnic headache)
Duration of headache history	The longer the duration the less the likelihood of secondary cause
Family history	Up to 80% of migraine patients have some family history of headaches
Migraine history accompaniments	Motion/car sickness; brain freeze; episodic vertigo/dizziness
Medical history	Depression/anxiety; irritable bowel; insomnia
Onset of pain	Sudden onset with SAH; gradual onset with migraine or tension-type headaches
Location of pain	Unilateral with some migraines; bitemporal with TTH (not specific); focal pain in specific neuralgias
Quality of pain	Throbbing with migraine; pressure with TTH (not specific)
Severity of pain	Milder with tension type, more severe with migraine (use 1–10/10 pain scale)
Pain-associated symptoms	Nausea; vomiting; sensitivity to light/sound/smell; migraine aura
Duration of pain	Minutes with some primary headaches, hours to days with migraine, hours to years with tension-type headache
Relief of pain (meds)	Over-the-counter medications for TTH; triptans for migraine
Relief of pain (non-meds)	Rest/quiet/dark/temporal pressure/ice/sleep with migraines
Frequency of medication use	Pills per headache, pills per week/month (get very specific answers)
Frequency of pain	Headache days per week/month
Disability of pain	Missed work/social activities; job or family concerns; anticipation of disability

TTH, tension-type headache; GCA, giant cell arteritis; SAH, subarachnoid hemorrhage.

Table 7 'Red flags' in the headache history

- Unusually severe headache of sudden, rather than gradual, onset
- Unexplained marked worsening or change in character of previously existing headaches
- Headache awakening the patient at night (as opposed to waking up with a migraine headache)
- Headache that worsens when recumbent, coughing, sneezing (may also be noted in migraine)
- New-onset headaches at an older age (>50 years) without prior or family history of headaches
- Headaches in patients with cancer known to metastasize to brain or skull
- Headaches in patients with HIV or other systemic infections
- Headache associated with unexplained systemic illness (e.g., fever, chills, rash, stiff neck)
- Abnormal neurological examination in a headache patient, including presence of papilledema
- Focal neurological signs that persist at the conclusion of headache pain
- Chronic headaches, transient visual abnormalities, pulsatile tinnitus in an obese young woman

arranged. If the patient does not respond as expected to the appropriate migraine or TTH headache treatment, then imaging, generally an MRI scan of the brain without contrast, may be obtained. The MRI of the brain in a patient with migraine may show changes that are remarkable but not clinically significant, leading to unnecessary anxiety and physician consultation. Findings noted on the brain MRIs of patients with migraines include atrophy greater than expected for age and white-matter lesions. Deep white-matter lesions have been shown to be increased in women with migraines, with or without aura, and are associated with increased migraine frequency. Cerebellar lesions consistent with infarcts have been noted associated with migraine with aura occurring more than once a month. The correlation of these MRI lesions with pathological damage is not clear, but their presence may indicate those patients who should be treated to decrease the frequency of their migraines with aura.

Headache Treatment

Acute Treatment of Primary Headaches

General principles of primary headache treatment include establishing an accurate diagnosis and educating patients about their condition, including its natural history and specific treatment. Realistic expectations about treatment of a chronic condition should be encouraged. Patients must be enlisted to participate in the management of their headaches, complying with the evaluation and the lifestyle and medication recommendations. The patient and physician should discuss the patient's treatment and medication preferences in order to encourage cooperation with the recommendations. A headache diary to monitor the frequency and severity of headaches is encouraged to document the disability of the headaches and the patient's response to treatment.

Patients often need multiple nonpharmacological and pharmacological therapies. The first step in headache prevention is regulation of lifestyle to avoid triggers. Unfortunately, many major migraine triggers

are difficult or impossible to avoid (e.g., stress and change in weather) and medication may be indicated to decrease the disability associated with frequent episodic or chronic headaches. Nonpharmacological treatment of primary headaches can include stress reduction, regular pleasurable exercise, biofeedback, relaxation techniques, and counseling. Multiple medication types – including acute treatment to treat the pain when it occurs, preventative medication to decrease the frequency and severity of the headaches, and rescue medication to treat acute intractable pain-are used to decrease headache disability. Patients may need multiple different strategies in order to choose the best treatment for each particular headache. Treatment often combines medication and withdrawal from stimulation (e.g., sleep or rest in a quiet, dark room for migraine sufferers), with patients discovering over time what works best for different headache types.

Patients with acute headaches generally take over-the-counter (OTC) pain relief medications, including aspirin, acetaminophen, ibuprofen, and naproxen, even before seeking medical consultation. Acute therapy should be appropriate for the degree of pain, with OTC therapy for mild or TTH headaches, and migraine-specific medication (generally triptans) for more severe migraine headaches. The specific 5-hydroxytryptamine (5-$HT_{1B/1D}$; serotonin) receptor agonists (triptans) and the nonspecific 5-HT agonists (ergotamine and dihydroergotamine) are specific for acute migraine pain relief. The multiple triptan medications that are available in the United States to treat acute migraine headaches are listed in Table 8. Barbiturates, opioids, and isometheptene may be combined with anti-inflammatory medications for treatment of pain, although their use should be monitored because of the risk of headache associated with medication overuse (Table 2). When acute medications (OTCs and triptans) are ineffective, rescue with barbiturates and opioids used sparingly may be necessary. A prolonged migraine that is intractable to oral medications can be treated with parenteral drugs, including intravenous valproate or dihydroergotamine, often in combination with intravenous hydration and antiemetic medication.

Table 8 Triptans for treatment of acute migraine[a]

Sumatriptan (Imitrex): oral, 25, 50, 100 mg; nasal, 5, 20 mg; autoinjector, 6 mg, 4 mg
Zolmitriptan (Zomig): oral, 2.5, 5 mg; ODT,[b] 2.5, 5 mg; nasal, 5 mg
Naratritpan (Amerge): oral, 1, 2.5 mg
Rizatriptan (Maxalt): oral, 5, 10 mg; ODT, 5, 10 mg
Almotriptan (Axert): oral, 6.25, 12.5 mg
Frovatriptan (Frova): oral, 2.5 mg
Eletriptan (Relpax): oral, 40 mg

[a]Triptans are serotonin (5-hydroxytryptamine; 5-$HT_{1B/1D}$) receptor agonists.
[b]ODT, orally disintegrating tablet.

Preventative Treatment of Primary Headaches

Migraine sufferers who have frequent disabling headaches may need a daily medication to decrease the frequency and severity of primary headaches and to decrease the use of acute medications. The decision to start a preventative medication is individualized, after discussion between the patient and physician, but there are guidelines to indicate the appropriate treatment of patients with chronic primary headaches. The AMPP study defined three groups of patients, depending on whether prevention should be offered, considered, or not considered. Prevention should be offered to migraine sufferers reporting six or more headache days per month, four or more days per month with at least some impairment, or three or more headache days per month with severe impairment. Prevention should be considered for migraine sufferers reporting either four to five headache days per month with normal functioning, three headache days per month with at least some impairment, or two headache days per month with severe impairment. Patients with less frequent or severe headaches would generally not be appropriate for preventative therapy, realizing that some patients' jobs or lifestyles may dictate prevention for even rare headaches. With most migraineurs having one to four headache attacks per month with severe impairment, the AMPP found that approximately a third of the patients who had never used a preventative medication met the criteria for offering or considering its use. Physicians should try to obtain a history indicating the headache frequency, severity, and level of disability in order to evaluate the appropriateness of preventative treatment.

Table 9 lists some types of medications that have been used to decrease the frequency and severity of headaches. The choice of preventative medication

Table 9 Preventative medications for treatment of chronic headaches

- Beta blockers (propranolol, atenolol, timolol, metoprolol)
- Calcium channel blockers (verapamil)
- Angiotensin receptor blockers/angiotensin-converting enzyme inhibitors (candesartan, olmesartan, lisinopril)
- Tricyclic antidepressants (nortriptyline, amitriptyline)
- Anticonvulsant medications (gabapentin, valproate, topiramate)
- Selective serotonin reuptake inhibitors
- Nonsteroidal antiinflammatory drugs
- Other therapies (riboflavin, Mg, feverfew, vitamin B_2, butterbur, botulinum toxin)

is often dictated by the migraine patient's other medical or psychiatric conditions. Common conditions found in migraine patients that may also respond to treatment with a headache preventative medication include depression, anxiety, obesity, sleep disorders, hypertension, and wrinkles. A tricyclic antidepressant dosed at bedtime may be particularly helpful in a migraine patient with insomnia. A selective serotonin reuptake inhibitor can be used when there is a significant component of depression or anxiety triggering headaches. A medication that also lowers blood pressure is appropriate for the hypertensive migraineur. The potential for weight loss with topiramate is attractive to many migraine patients, as many other migraine prevention medications may cause weight gain. Patients often report improved response to triptan medication when they are also taking a preventative medication. Injection of botulinum toxin into muscles of the face, head, and neck may have a role in headache prevention, with wrinkle prevention benefits as well.

A headache is an annoyance at best; rarely is it an indication of a life-threatening condition. The suffering associated with severe or frequent primary headaches can be extreme. Patients who have lived with headache pain for years often think that having headaches is a 'normal' part of life for which there is no treatment. Historical details, the neurological examination, and, when indicated, brain imaging provide guidance in diagnosing these headache patients. Recognition of the disability of headaches along with a compassionate, consistent approach to the care of headache sufferers can markedly improve the quality of life for headache sufferers, many of whom are undiagnosed and undertreated.

See also: Cerebrovascular Disease; Hemiplegic Migraine and Channel Disorders; Neuropathic Pain; Pain and Genes; Pain Management; Pain: Neuroimaging; Sensation from the Face.

Further Reading

Battistella PA, Fiumana E, Binelli M, et al. (2005) Primary headaches in preschool age children: Clinical study and follow-up in 163 patients. *Cephalalgia* 26: 162–171.

Buchanan TM and Ramadan NM (2006) Prophylactic pharmacotherapy for migraine headaches. *Seminars in Neurology* 26: 188–198.

Cutrer FM and Black DF (2006) Imaging findings in migraine. *Headache* 46: 1095–1107.

Evers S and Goadsby PJ (2003) Hypnic headache: Clinical features, pathophysiology, and treatment. *Neurology* 60: 905–909.

Goadsby PJ, Silberstein SD, and Dodick DW (eds.) (2005) *Chronic Daily Headache for Clinicians*. BC Hamilton, London: Decker Inc.

Goadsby PJ (2007) Recent advances in understanding migraine mechanisms, molecules and therapeutics. *Trends in Molecular Medicine* 13: 39–44.

Goadsby PJ, Cohen AS, and Matharu MS (2007) Trigeminal autonomic cephalalgias: Diagnosis and treatment. *Current Neurology and Neuroscience Reports* 7: 117–125.

Headache Classification Committee of the International Headache, Society (IHS) International Classification of Headache Disorders II. *Cephalalgia* 24(supplement 1): 9–160.

Kurth T, Gaziano JM, Cook NR, et al. (2006) Migraine and risk of cardiovascular disease in women. *JAMA: Journal of the American Medical Association* 296: 283–291.

Kruit MC, van Buchem MA, Hofman PA, et al. (2004) Migraine as a risk factor for subclincal brain lesions. *JAMA: Journal of the American Medical Association* 291: 427–434.

Lipton RB and Bigal ME (2006) Migraine and cardiovascular disease. *JAMA: Journal of the American Medical Association* 296: 332–333.

Lipton RB, Bigal ME, Diamond M, et al. (2007) Migraine prevalence, disease burden, and the need for preventative therapy. *Neurology* 68: 343–349.

Olesen J, Goadsby PJ, Ramadan NM, et al. (eds.) (2006) *The Headaches,* 3rd edn. Philadelphia: Lippincott Williams & Wilkins.

Rains JC and Poceta JS (2006) Headache and sleep disorders: Review and clinical implications for headache management. *Headache* 46: 1344–1363.

Schievink WI (2006) Spontaneous spinal cerebrospinal fluid leaks and intracranial hypotension. *JAMA: Journal of the American Medical Association* 295: 2286–2296.

Silberstein S, Loder E, Diamond S, et al. (2007) Probable migraine in the United States: Results of the American Migraine Prevalence and Prevention (AMPP) study. *Cephalalgia* 27: 220–229.

VanGijin J, Kerr RS, and Rinkel GJ (2007) Subarachnoid hemorrhage. *Lancet* 369: 306–318.

Ward TN and Levin M (2005) Headache in giant cell arteritis and other arteritides. *Neurological Sciences* 26: S134–137.

Weiller C, May A, Limmroth V, et al. (1995) Brain stem activation in spontaneous human migraine attacks. *Nature Medicine* 1: 658–660.

Weinberger J (2007) Stroke and migraine. *Current Cardiology Reports* 9: 13–19.

Welch KMA (2005) Brian hyperexcitability: The basis for antiepileptic drugs in migraine prevention. *Headache* 45: S25–S32.

Relevant Websites

http://www.americanheadachesociety.org – American Headache Society (AHS).

http://www.achenet.org – American Headache Society Committee on Headache Education.

http://www.i-h-s.org – International Headache Society (IHS).

Hearing and Echolocation in Dolphins

S H Ridgway, Space and Naval Warfare Systems Center, San Diego, CA, USA
W W L Au, University of Hawaii, Kailua, Hawaii

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Humans surpass most mammals in such important qualities of hearing as intensity and frequency discrimination and sound localization. During the past 40 years, research has shown that dolphins can localize sound and discriminate frequency and intensity as well as humans; moreover, it is now known that the range of dolphin hearing is about seven or eight times as broad as the range of human hearing which extends only to about 20 kHz. Audiograms have been done on several species of smaller odontocetes by both behavioral and electrophysiological techniques. *Tursiops* (Figure 1) responds to frequencies at least as high as 150 kHz with great sensitivity between about 10 and 120 kHz (Figure 2). Sensitive ears coupled with a massive auditory central nervous system are fundamental to the dolphin's extensive use of sound for communication and echolocation. As dolphins age, hearing sensitivity may degrade somewhat, especially in older males, where loss above 50 or 60 kHz appears to be common. Also, there is some evidence that long-term treatment with aminoglycoside antibiotics may degrade high-frequency hearing.

Open-Water Echolocation Studies with Trained Dolphins

Many echolocation studies have been done in the open waters of Kaneohe Bay, where a calm surface and muddy bottom help provide good acoustic conditions for the tests. In all dolphin echolocation experiments, here and elsewhere, the dolphins have emitted short (40–100 μs) pulses to search for underwater targets. The intense pulses (Figure 3), up to 230 dB re 1 μPa in peak-to-peak amplitude and with peak frequency often greater than 100 kHz, are directed forward in a narrow beam. With peak frequencies well outside the human hearing range, it is difficult for us to appreciate the intensity of dolphin echolocation pulses. An analogous airborne sound might be a pistol shot. At 1 m under water, dolphin pulses can have peak intensities as great or greater than the sounds of shots experienced by humans firing low-caliber rifles.

We still do not understand exactly how a dolphin can produce such loud pulses. The keen ears are located in the dolphin's head only a few centimeters away from pulse production mechanism. Anatomical structure, including sinuses containing highly reflective air, and various physiological mechanisms presumably help the animal avoid damaging its own ears.

Using these intense pulses, their sensitive hearing, and a nervous system specialized for rapid acoustic processing, dolphins can detect echoes from small objects at long ranges. For example, in behavioral studies, dolphins detected a 2.54 cm diameter steel ball 1.2 m under the surface to a range of 73 m and they detected a 7.62 cm water-filled sphere 113 m away. At 16 m range, dolphins distinguished objects of several different metals of identical size and shape. At the same range, dolphins detected a 0.23 mm difference in wall thickness of otherwise identical aluminum cylinders. Our studies have shown that over the range of 1–113 m, echolocating dolphins are superior to any human-made sonar for object detection and discrimination.

Studies of Dolphins in the Wild

It has only in recent times that detailed studies of dolphin echolocation in the wild have been undertaken using a multiple hydrophone array so that the echolocating animal could be localized. One remarkable finding is that the echolocation signals measured with five different species of odontocetes in widely separate locations suggest that these animals use a form of time-varying gain or automatic gain control. Time-varying gain is a technique used in technological sonar by which the gain of the receiver increases with time as a function of 40 log R, where R is the range in meters, to compensate for two-way spherical spreading losses that the signal and echo experience while propagating to the target and back. Dolphins instead vary the peak-to-peak source level (sound pressure level reference to 1 m in front of any projector) so that the further away a target is, the greater the output level of the echolocation signal, or, conversely, the closer the target the lower the output level. Variations in the pulse output level of five species of free swimming dolphins are shown in Figure 4. The solid line is the curve $K + 20 \log R$, that best fit the data with the r^2 value shown for each species. Not only are the data shown in Figure 4 from five different species of odontocetes, but the measurements were made at widely separated locations. The data for *Stenella frontalis* were obtained in the waters of the Bahama Islands, *Lagenorhynchus albirostris* in the waters of Iceland, *Orcinus orca* in the waters of

Figure 1 The bottlenose dolphin *Tursiops truncatus*. Note the absence of any signs of an external ear or pinna.

Vancouver Island, British Columbia, *Lagenorhynchus obscurus* in the waters of Kailoura, New Zealand, and *Neophocaena phocaenoides* in the Yangtze River of China.

The dynamic control of the echolocation source level is probably not the result of a cognitive process, at least not on a pulse-by-pulse basis, but rather a natural consequence of how echolocation clicks are produced. Dolphins typically emit echolocation clicks at a rate that allows the echoes to return to the animal before the next click is emitted. Consequently, the repetition rate increases as an animal closes on a target. The clicks are produced within the nasal system. Dolphins manipulate the air flow through the phonic lips within the nasal or blowhole complex. Dolphins initially pressurize their nasal system and then emit a click train with the individual clicks occurring at relatively low repetition rates. The animal continually adjusts the rate as targets are located. If the dolphin chooses to keep the amount of acoustic energy emitted relatively constant or within certain limits for each pressurization cycle, then the amplitude of the signal can be high when the repetition rate is low but the amplitude must continually decreases as the repetition rate increases. The data are consistent with the notion that there is a dynamic coupling between reception of echoes, repetition rate, and source level of transmitted signals.

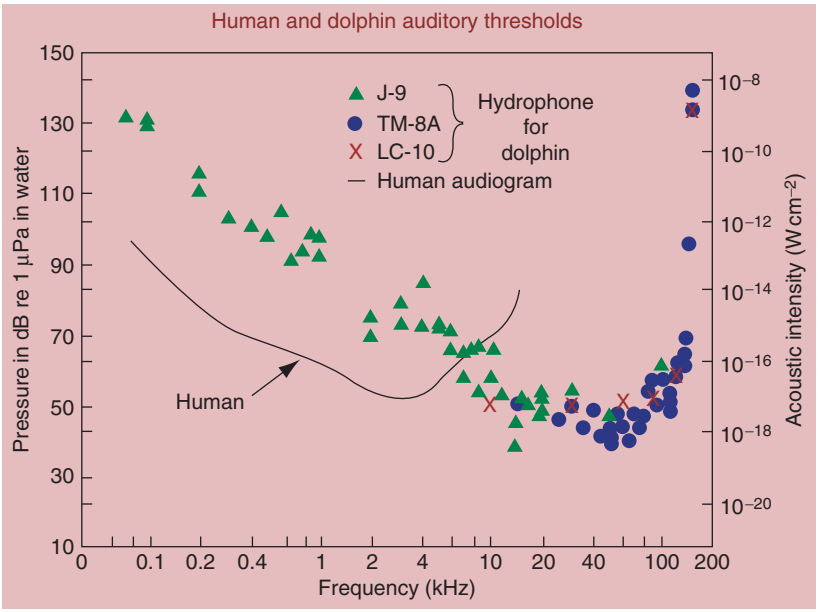

Figure 2 Auditory sensitivity of a bottlenose dolphin and human. The left ordinate indicates sound pressure level in dB re 1 μPa. The right ordinate indicates intensity in W cm^{-2} so that the audiogram in air (human) and water (dolphin) can be compared. From Johnson CS (1966) Auditory thresholds of the bottlenosed dolphin (*Tursiops truncatus* Montagu). US Naval Ordinance Test Station NOTS TP 4178, pp. 25. Johnson CS (1968) Relationship between absolute threshold and duration of tone pulse in the bottlenose dolphin. *Journal of Acoustical Society of America* 43: 757–763.

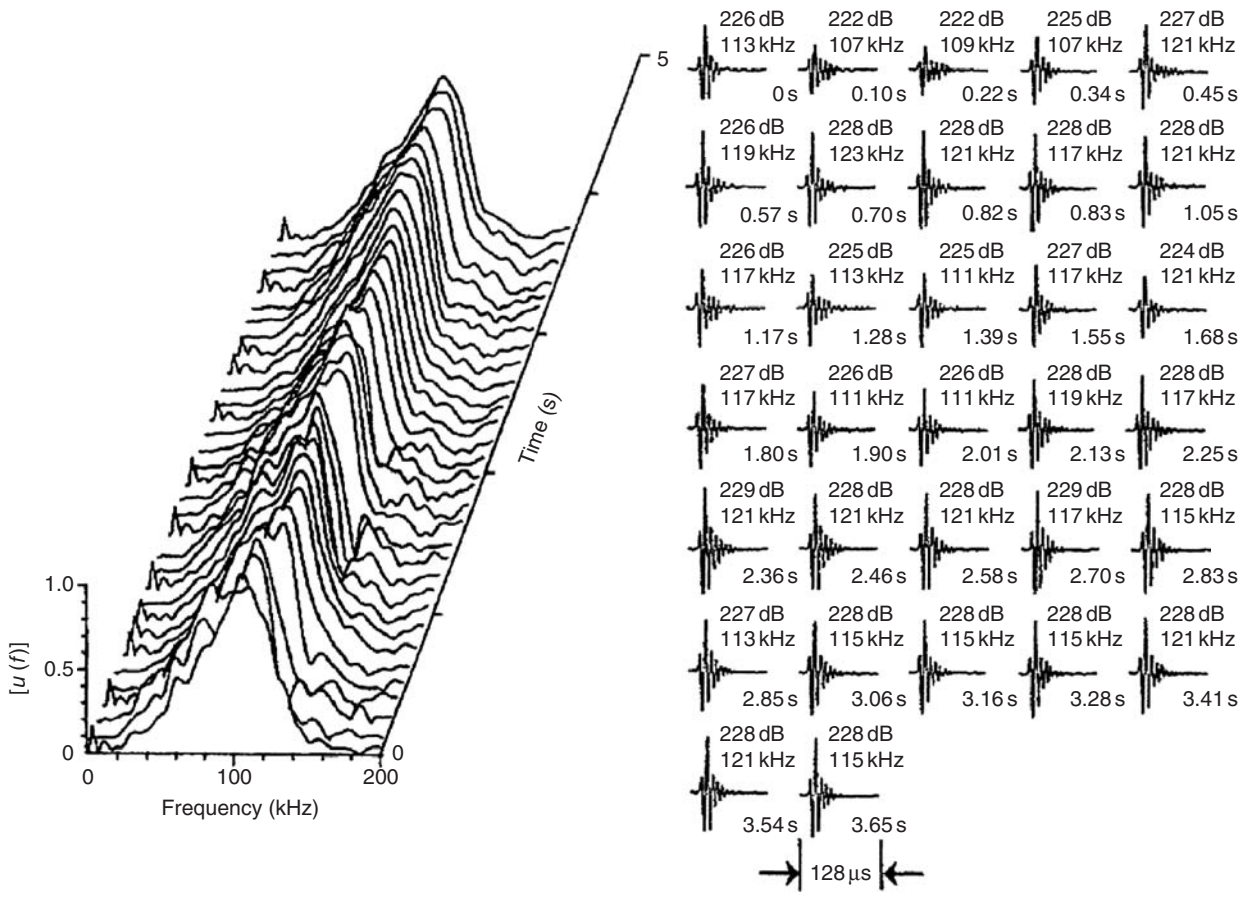

Figure 3 An example of a dolphin pulse train with the frequency vs. time plotted on the left and individual pulses presented with their peak frequencies and levels (dB re 1 μPa peak to peak) on the right. From a dolphin in an echolocation task in Kaneohe Bay, Hawaii.

High-Frequency, Narrow-Band Echolocators

Numerous species of bats have developed their own specialized echolocation signals to cope with various different ecological niches. A few small whale species such as porpoises of the family Phocoenidae, the genus *Kogia*, pygmy sperm whales, and dolphins of the genus *Cephalorhynchus* specialize in brief narrow-band, very high frequency, lower-intensity echolocation pulses with energy centered around 130 kHz. We suggest that as in the case of bat echolocators, ecological demands such as feeding advantage or predator avoidance have driven the evolution of specialized echolocation pulses in these small whale species. For example, specialized high-frequency, narrow-band echolocation pulses might not be heard by the killer whale, a potential cetacean predator that does not have significant hearing sensitivity at 130 kHz.

Communication Sounds

Although echolocation pulses may also be used in communication, the lower-frequency sounds such as whistles and bursts of pulses are often considered in the context of emotional or social communication. Whistles are tonal, frequency-modulated signals lasting roughly 0.5–5 s with fundamentals that are rarely outside a frequency range of about 3–30 kHz. Pulsed sounds described by terms such as 'buzzes,' 'squeaks,' 'yelps,' and 'squawks' are rapid trains of broadband pulses, usually 100–2000 s^{-1}. Such pulsed sounds usually contain energy from 1 kHz or less to over 150 kHz. On occasion, a dolphin can modulate the repetition rate of such pulses so that they sound humanoid. Unlike echolocation pulses mentioned above, these bursts of pulses often contain peak intensities below 20 kHz.

The Peripheral Auditory System

In keeping with a swimming-and-diving, entirely aquatic lifestyle, the pinna has disappeared (Figure 1), remaining only as a subcutaneous cartilagenous vestige within the dolphin's streamlined head. From a minute depression a few centimeters posterior to the eye, the vestigial external auditory meatus (EAM)

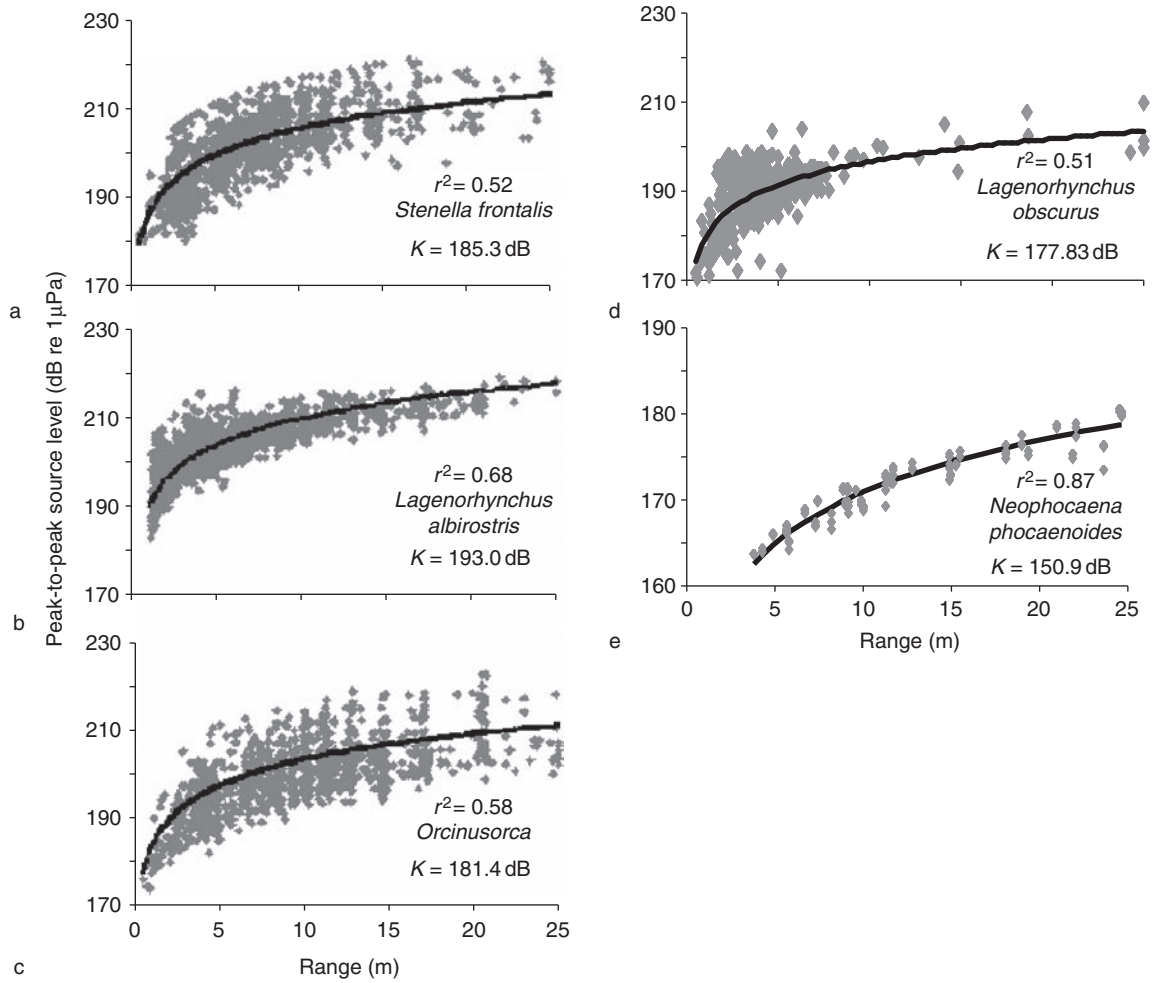

Figure 4 Source level of echolocation signals from five species of odontocetes as a function of range. (a–c) Data from Au WWL and Benoit-Bird KJ (2003) Automatic gain control in the echolocation system of the dolphin. *Nature* 423: 861–863. (d) Data from Au WWL and Wursig B (2004) Echolocation signals of dusky dolphins (*Lagenorhynchus obscurus*) in Kaikoura, New Zealand. *Journal of Acoustical Society of America* 115: 2307–2313. (e) Data from Li S, Wang D, Wang K, and Akamatsu T (2006) Sonar gain control in echolocating finless porpoises (*Neophocaena phocaenoides*) in open waters. *Journal of Acoustical Society of America* 120:1803–1806.

leads inward along a sinuous path to the tympanic bone, which lies deep in the lateral tissues of the head and is partly ringed by extensions of the exoccipital and basioccipital bones of the skull. Over much of this path the EAM lumen is tiny or even absent. Near the tympanic bone, the tiny meatus expands in a membranous funnel shape – the tympanic conus which attaches to the margins of a kidney-shaped opening in the dorsal surface of the tympanic bone. The same margins support the ligamentous tympanic membrane.

From the internal surface of the tympanic membrane, and from one edge of the bony margins, a thick tympanic ligament arises and runs anteroventrally to insert on the malleus. There appears to be no direct contact of the tympanic membrane with the malleus.

The ossicles are large and heavy. The malleus is rigidly attached to the bulla and to the incus; however, the joint between incus and stapes is movable. Like the ossicles, the temporal bones are unusually dense. Unlike their human homologs, dolphin ear bones do not touch the skull, being suspended by ligaments and tough fibrous tissue. The Eustachian tube feeds from the nasal cavity into the middle ear and auditory sinuses around the ear, creating an air cavity, especially on the medial and posterior sides, around much of the cochlea-containing periotic bone. Recently, computed tomography (CT) scans of living dolphins have been used to show the extent of these air cavities which no doubt play an important role in hearing.

The histological structure of the cochlea's organ of Corti is sturdy with large closely packed supporting cells giving it an appearance of great strength and rigidity. It contains the usual elements such as arch of Corti, Hansen's cells, Deiters' cells, and hair cells,

all overlain by the tectorial membrane, but the usual lymph spaces are much reduced. The basilar membrane of the cochlea shows a systematic variation in width and in the manner and rigidity of suspension. In the basal region, the basilar membrane is suspended by bony laminae on both the outer and inner edges. This suspension becomes progressively less rigid toward the apical end where the membrane is held by connective tissue. The basilar membrane varies in width by 14 times – ranging from 25 μm at the basal end to 350 μm in the upper apical. The human basilar membrane, by comparison, varies in width only 6.25 times and is not nearly so narrow at the basal end (Figure 5). EG Wever and colleagues compared the numbers of hair cells and ganglion cells and their distribution along the cochlea to values from humans. The human had about 3500 inner and 11 500 outer hair cells, whereas the dolphin had roughly 3500 inner and 14 000 outer hair cells. The ratio of ganglion cells to hair cells in the human cochlea is about 2:1, whereas the ratio is more than 5:1 in the dolphin.

Sound Conduction in the Dolphin Ear

McCormick et al. surgically placed an electrode on the round window of the cochlea in anesthetized dolphins and recorded the electrical potential of the cochlea in response to various sound stimuli; then, they observed changes produced in the cochlear potential due to manipulations of auditory structures including damping, immobilization, application of tension, or removal. Because such actions on parts of the outer ear produced no cochlear potential changes, the researchers concluded that the outer ear – EAM, tympanic conus, and tympanic membrane – plays no significant part in sound conduction to the dolphin cochlea in the frequency range of 20 kHz and above where the dolphin is most sensitive. The ossicles are involved, since 30 g of tension on the ossicular chain produced an 18 dB drop in the cochlear potential which was restored by relief of the tension. Severing the malleus produced only a 4 dB drop; therefore, a complete conductive chain is apparently not essential. These results were therefore interpreted to indicate that acoustic vibrations passing through the tissues produce a relative motion between the cochlear capsule and the ossicles – the inertia and stiffness of the ossicular chain result in relative movement by the stapes attached to the oval window – producing motion in the cochlear fluids which freely move because of air over the round window.

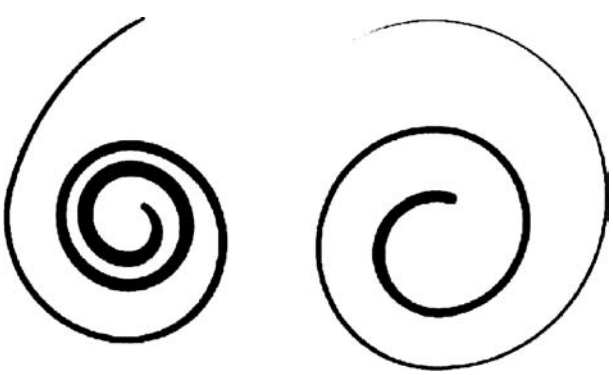

Figure 5 A tracing comparing length and width of dolphin (right) and human (left) basilar membrane. From Wever EG, McCormick JG, Palin J, and Ridgway SH (1971) Cochlea of the dolphin, *Tursiops truncatus*: The basilar membrane. *Proceedings of the National Academy of Sciences of the United States of America* 68(11): 2708–2711.

Air is some 700 times less dense than water. The acoustic impedance of the two media differs by a factor of about 3660. Sound is readily transmitted through the dolphin, because body tissues have a similar acoustic impedance to water; however, air adjacent to the ear bones is important for at least two reasons. First, compressible air makes possible movement of the cochlear fluid in response to relative forces between the stapes and cochlear capsule – the round window membrane is free to move in and out. The dolphin round window membrane is multilayered, interlaced with tough fibrous strands and many times thicker than the corresponding human structure. Second, the air–fluid interface is an excellent reflector of sound; therefore, the air around the ear bones may serve to protect the ears from the dolphin's intense pulses and to isolate one ear from the other. This idea is supported by body-mapping studies of sound-sensitive areas which showed a 12 dB difference in contralateral stimulation required to produce an equivalent cochlear potential in the opposite ear. This isolation of the ears aids in localization by directing the pattern of received sound.

Air compression during diving may be of some consequence, since dolphins are capable of diving to more than 300 m. The thorax and lungs are structured for collapse. The nasal system which pressurizes by muscle action during sound production connects directly to the ear cavities by the eustachian tube. At 300 m depth, a bottlenose dolphin's air supply is compressed to about 250 ml in the air cavities of the head. This small volume of air is apparently sufficient for sound production and hearing at depth. Even a 1 mm thickness of air serves as a very efficient reflector of sound and air adjacent to the round window membrane is sufficient for its elasticity. Hearing tests conducted with another member of the superfamily Delphinoidea, the white whale (*Delphinapterus leucas*), showed that hearing was just as acute at a depth of 300 m as near the surface.

The Path of Sound to the Ear

Human divers can hear ultrasonics underwater, but they cannot perceive pitch in the ultrasonic region above ~20 kHz. Even though humans hear ultrasound underwater, sensitivity is much reduced compared with hearing in air; moreover, human ability to localize sound is degraded when we go underwater. The evolution of the dolphin ear has apparently been a process of enhancing those characteristics that allowed for greater high-frequency hearing sensitivity and complex auditory processing in the underwater environment where sound travels about 5 times faster than in air. Early in the Eocene epoch, when the distant land-mammal ancestors of the cetaceans with their aerial ear entered the water, it is almost certain that they could immediately hear ultrasonics conducted through their tissues which are fairly well matched to the water for conducting sound.

Despite the acoustic conductance match between tissue and water, there has been a good deal of controversy about how sound reaches the dolphin ear. Many of the experiments on sound transmission have been done with dead specimens. Since acoustic impedance increases after death, these findings must be interpreted with caution. KS Norris, intrigued by the form and structure of the dolphin's lower jaw and its relation to the ear bones and 'acoustic fat,' has postulated that sounds enter through the thin flared posterior end of the mandible into the mandibular fat body and are transmitted to the middle ear. Experimental results seem to support Norris' theory on the role of the lower jaw in sound reception. Electrophysiological measurements of evoked potentials in the inferior colliculus and cochlear potentials indicated that maximum responses were obtained when a sound source was placed in the vicinity of the lower jaw. With the exception of air cavities, the impedance of dolphin head tissue is similar to that of water; thus, Johnson suggested that the directionality in dolphin sound reception may be caused not so much by discrete directional paths through the jaw as suggested by Norris but more by directional characteristics of the ear and its adjacent air cavities, and, possibly, by processing in the central nervous system.

Beam Patterns and Localization

Whatever the mechanism by which sound reaches the ear, we know that sounds approaching from certain directions are efficiently coupled to the ears, whereas sounds from other directions may be poorly coupled. Au and Moore measured the receiving beam pattern of a bottlenose dolphin in both the vertical and horizontal planes by determining the animal's masked threshold as the position of either the noise or signal sources varied in their angular position about the animal's head (Figure 6). The receiving beams are frequency dependent, narrowing beams in both planes at higher frequencies and broadening at lower frequencies. The horizontal beam points forward and is nearly symmetrical about the longitudinal axis of the dolphin. The vertical beam points upwards by 5–10° with respect to the posterior–anterior axis of the tooth row and is asymmetrical,

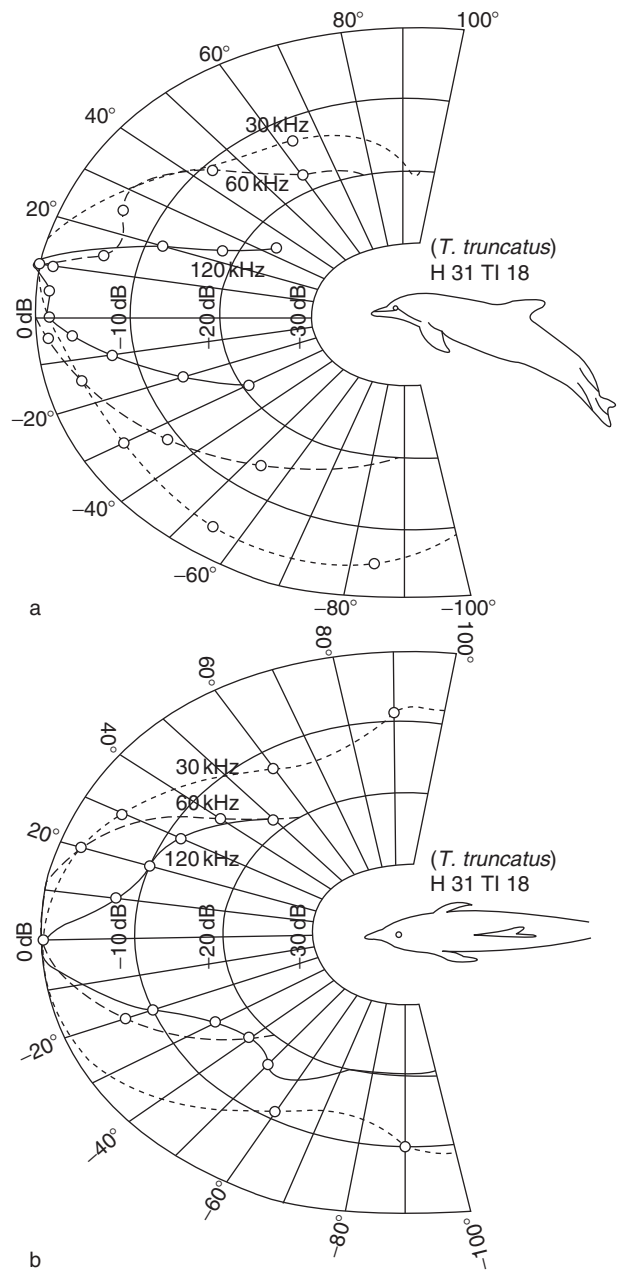

Figure 6 Sound reception characteristics of bottlenose dolphins. Receiving (a) vertical beam patterns and (b) horizontal beam patterns for frequencies of 30, 60, and 120 kHz.

decreasing faster as a function of angle above the animal's head than below. Most likely, as the angle of a sound source increases above a dolphin's head, the air spaces of the head shadow the ear and cause the dolphin's hearing sensitivity to decrease rapidly. The animals generally swim in a slightly beak-down attitude; thus the vertical beam points straight ahead of the swimming animal.

The energy from the outgoing echolocation clicks is also directed forward in a narrow pattern. Since dolphins, often constantly calling and clicking, frequently travel in large herds and feed in groups, the directionality of both transmission and reception systems seems to be critical. This directionality determines the amount of ambient and other noises that actually enter the animals' auditory system to mask echoes from its own clicks and other sounds of interest. The receiving beam can also affect a dolphin's ability to localize sound.

The high sound speed in water should make localization difficult because of the much reduced time-of-arrival difference between the two ears. Anatomical structures such as special 'acoustic lipids' that may help to focus sound and air sinuses that help to isolate the ears in some planes and to shadow the ears in other planes no doubt are critical in sound localization. Localization of sounds in the vertical plane is likely dependent on the width and shape of the vertical beam. In the horizontal plane, localization should depend on either time or intensity difference of a sound arriving at the two ears of the dolphin. However, the narrower the beam, the better the signal-to-noise ratio will be, which should lead to keener localization ability. A narrow beam will also reduce any front–back ambiguity in localizing sounds in the horizontal plane. Renaud and Popper measured the minimum-audible angles (MAAs) of a bottlenose dolphin in the horizontal and vertical planes. They measured MAAs between 1° and 3° in both planes for pure-tone signals with frequencies between 20 and 90 kHz. At 6 and 100 kHz, the MAA increased to approximately 4°. Using a simulated echolocation click with a peak frequency of 64 kHz, MAAs of 0.7–0.8° were measured. The underwater pure-tone localization capability of dolphins is about as good as the in-air capabilities of several mammals, including humans. Dolphins localize broadband clicks considerably better than other mammals.

Auditory Performance Characteristics

The dolphin and human audiograms are similar in shape, with the dolphin's shifted by a factor of about 8 to higher frequencies (Figure 2). The absolute sensitivity to sounds is comparable between the dolphin and human. Johnson also determined the hearing threshold of a bottlenose dolphin as the duration of the pure-tone stimulus was varied. The dolphin's threshold decreased as the tone duration increased to approximately 0.3 s, after which the threshold leveled off.

For the brief echolocation pulses, a concept of critical interval or integration time has been developed. Au et al. measured the integration time of the dolphin auditory system when receiving echolocation echoes to be approximately 264 μs. When delays between pulse pairs are reduced below this critical interval, the dolphin no longer perceives them as separate pulses. Comparable values for humans would be 1–5 ms. A delay of 264 μs is equivalent in echo delay from an object only about 20 cm from the pulse source within the dolphin's nasal system. Even though individual echolocation pulses may be much shorter, the integration time for pulses might be necessary to mitigate interference from echoes within the dolphin's own head.

For longer tonal signals, Johnson measured the integration time to be on the order of 100–200 ms. Bottlenose dolphins are sensitive in detecting frequency changes. The dolphin's best frequency discrimination occurred between 2 and 55 kHz, with relative difference limens (DLs) ranging from 0.2% to 0.4%. Relative DLs remained within 1% for all frequencies between 1 and 140 kHz (Figure 7). The frequency discrimination capabilities of dolphins compare well with the in-air capabilities of humans, but shifted toward higher frequencies. Intensity discrimination ability is also very good – on the order of 1 dB (apparently better than that of any other vertebrate studied except humans).

The Auditory Nervous System

In total diameter and in numbers of fibers, the eighth nerve, the largest of the dolphin's sensory cranial nerves, is much larger and contains several times as many fibers as the human auditory nerve. The larger fiber diameter has physiological significance since a doubling in fiber diameter of myelinated axons in laboratory mammals approximately doubles conduction speed. On a comparative basis, central auditory structures are even more enlarged. In the common dolphin, a small animal with a brain just over half as large as a bottlenose dolphin (800 vs. 1500 g), the cochlear nucleus and inferior colliculus are about 12 times as large as those of the human, the trapezoid is twice as large, the medial geniculate about 7 times as large, the superior olive is 150 times as large, and the nucleus of the lateral lemniscus is over 250 times as large as the equivalent structure in humans.

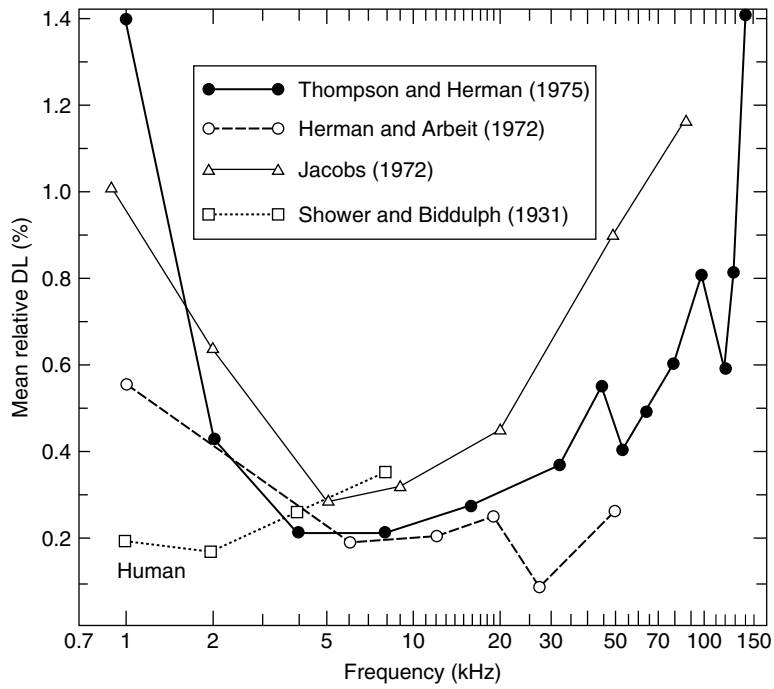

Figure 7 Frequency discrimination capability of bottlenose dolphin and human plotted in terms of the relative difference limen in percent. Thompson RKR and Herman LM (1975) Underwater frequency discrimination in the bottlenosed dolphin (1–140 kHz) and human (1–8 kHz). *Journal of Acoustical Society of America* 57: 943–948.

Physiological study of dolphin hearing and auditory processing has shown that temporal resolution of successive sounds is extremely rapid; the typical midbrain auditory evoked potential (AEP) was specialized for ultrasonic, ultrabrief, fast-rising, closely spaced sounds like the echolocation clicks. Physiological studies have also shown that the dolphin brain might possess two separate auditory systems, one specialized for ultrabrief, fast-rising sounds like echolocation clicks while the other appears to be specialized for longer, slower-rising sounds like dolphin whistles. Studies of the averaged brain stem response (ABR) have shown that clicks evoke a series of seven far-field waves cresting in the first 10 ms; these appear remarkably similar in several species. Dolphin ABRs are larger than those of other mammals previously investigated – about 10 times larger than similar waves recorded from the human scalp. In general, the ABR waves decrease in amplitude and increase in latency as click stimuli are attenuated. Brain stem transmission time (BTT) is considerably faster in the bottlenose dolphin, a species with mean brain weight around 1500 g, than in humans or domestic cats and is similar to that of the rat. Despite a much longer nerve pathway, BTT in the dolphin is equal to or faster than that in much smaller-brained species such as rats and cats.

Another characteristic of the dolphin ABR that differs from that in other mammals studied is the influence of stimulus repetition rate on the response. In adult humans, for example, latency increases by about 0.5 ms for each rate increase of $10\,s^{-1}$. The dolphin ABR latencies increase only slightly as stimulus rate is increased from 0.5 to $200\,s^{-1}$. In the cat, clicks at $80\,s^{-1}$ reduce all components to about 50% of recovered amplitudes, while dolphin ABR wave amplitudes show only little change at rates of $160\,s^{-1}$.

It is apparent that the dolphin brain is specialized for rapid processing of auditory stimuli. If given enough time, the human auditory system seems to perform as well on some echolocation tasks. When pulses similar to dolphin echolocation pulses were projected at targets at short range by instrumented divers and the received echoes stretched 128 times (tantamount to a slowed-down tape recording and therefore reduced equivalently in frequency), human divers performed with as few errors as *Tursiops* in distinguishing between paired metal targets of copper, brass, or aluminum and geometrical aluminum shapes covered with neoprene rubber. Studies with static metal targets may give only limited detail of the dolphin's echolocation processing ability. In addition, the static metal target discrimination may not be a very challenging echolocation task for a dolphin. However, if the findings are a fair comparison of human and dolphin echolocation ability, we must conclude that the major advantage in the sonar

processing component of the dolphin auditory system, in comparison with the human auditory system, is the ability to process high frequencies and to process sound rapidly. The large diameter of auditory nerve fibers, the short latency, and the relatively flat latency–intensity function of the dolphin ABR waves, and the rapid temporal resolution of successive sounds, are ample indication of an auditory system wired for rapid sound analysis. In fact, much of the great hypertrophy of the dolphin auditory system – and perhaps the entire cerebrum – may result from the animal's need for great precision and speed in processing sound.

See also: Auditory System: Central Pathways; Cetacean Brains; Cochlear Mechanics; Echolocation I: Behavior; Echolocation II: Neurophysiology; Language: Auditory Processes; Social Communication in Whales and Dolphins.

Further Reading

Au WWL (1993) *Sonar of Dolphins*. New York: Springer.

Au WWL (1997) Echolocation in dolphins with a dolphin–bat comparison. *Bioacoustics* 8: 137–162.

Au WWL and Benoit-Bird KJ (2003) Automatic gain control in the echolocation system of the dolphin. *Nature* 423: 861–863.

Au WWL and Moore PWB (1984) Receiving beam patterns and directivity indices of the Atlantic bottlenose dolphin (*Tursiops truncatus*). *Journal of Acoustical Society of America* 75: 255–262.

Au WWL and Wursig B (2004) Echolocation signals of dusky dolphins (*Lagenorhynchus obscurus*) in Kaikoura, New Zealand. *Journal of Acoustical Society of America* 115: 2307–2313.

Bullock TH and Ridgway SH (1972) Evoked potentials in the central auditory system of alert porpoises to their own and artificial sounds. *Journal of Neurobiology* 3: 79–99.

Dolphin WF (2000) Electrophysiological measures of auditory processing in *odontocetes*. In: Au WWL, Popper AN, and Fay RR (eds.) *Hearing by Whales and Dolphins*, pp. 294–329. New York: Springer.

Finneran JJ, Carder DA, Dear R, et al. (2005) Pure tone audiograms and possible aminoglycoside-induced hearing loss in the beluga (*Delphinapterus leucas*). *Journal of Acoustical Society of America* 117: 3936–3943.

Houser DS, Finneran J, Carder D, et al. (2004) Structural and functional imaging of bottlenose dolphin (*Tursiops truncatus*) cranial anatomy. *Journal of Experimental Biology* 207: 3657–3665.

Ivanov MP (2004) Dolphin's echolocation signals in a complicated acoustic environment. *Acoustical Physics* 50: 469–479.

Johnson CS (1966) Auditory thresholds of the bottlenosed dolphin (*Tursiops truncatus* Montagu). US Naval Ordinance Test Station NOTS TP 4178, pp. 25.

Johnson CS (1968) Relationship between absolute threshold and duration of tone pulse in the bottlenose dolphin. *Journal of Acoustical Society of America* 43: 757–763.

Johnson CS (1986) Dolphin audition and echolocation capacities. In: Schusterman R, Thomas J, and Wood F (eds.) *Dolphin Cognition and Behavior: A Comparative Approach*, pp. 115–136. New York: Lawrence Erlbaum Associates.

Li S, Wang D, Wang K, and Akamatsu T (2006) Sonar gain control in echolocating finless porpoises (*Neophocaena phocaenoides*) in open waters. *Journal of Acoustical Society of America* 120:1803–1806.

Madsen PT, Carder DA, Bedholm K, and Ridgway SH (2005) Porpoise clicks from a sperm whale nose – convergent evolution of 130 kHz pulses in toothed whale sonars? *Bioacoustics* 15: 195–206.

McCormick JG, Wever EG, Palin J, and Ridgway SH (1970) Sound conduction in the dolphin ear. *Journal of Acoustical Society of America* 48: 1418–1428.

Popper AN (1980) Sound emission and detection by delphinids. In: Herman LM (ed.) *Cetacean Behavior*, pp. 1–52. New York: Wiley.

Ridgway SH (2000) The auditory central nervous system of dolphins. In: Au WWL, Popper AN, and Fay RR (eds.) *Hearing by Whales and Dolphins*, pp. 273–293. New York: Springer.

Ridgway SH, Carder DA, Kamolnick T, Smith RR, Schlundt CE, and Elsberry W (2001) Hearing and whistling in the deep sea: Depth influences whistle spectra but does not attenuate hearing by white whales (*Delphinapterus leucas*, Odontoceti, Cetacea). *Journal of Experimental Biology* 204: 3829–3841.

Supin A Ya, Popov VV, and Mass AM (2000) *The Sensory Physiology of Aquatic Mammals*. Boston: Kluwer Academic Press.

Thompson RKR and Herman LM (1975) Underwater frequency discrimination in the bottlenosed dolphin (1–140 kHz) and human (1–8 kHz). *Journal of Acoustical Society of America* 57: 943–948.

Wever EG, McCormick JG, Palin J, and Ridgway SH (1971) The cochlea of the dolphin. *Tursiops truncatus*: Hair cells and ganglion cells. *Proceedings of the National Academy of Sciences of the United States of America* 68: 2908–2912.

Heart Rate Variability: A Neurovisceral Integration Model

J F Thayer, The Ohio State University, Columbus, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

> When the heart is affected it reacts on the brain; and the state of the brain again reacts through the pneumo-gastric (vagus) nerve on the heart; so that under any excitement there will be much mutual action and reaction between these, the two most important organs of the body. (Darwin, 1998, pp. 71–72; originally published in 1872)

Like many organs in the body, the heart is dually innervated. Although a wide range of physiologic factors determine cardiac functions such as heart rate (HR), the autonomic nervous system (ANS) is the most prominent. Importantly, when both cardiac vagal (the primary parasympathetic nerve) and sympathetic inputs are blocked pharmacologically (e.g., with atropine plus propranolol, the so-called double blockade), intrinsic HR is higher than the normal resting HR. This fact supports the idea that the heart is under tonic inhibitory control by parasympathetic influences. Thus, resting cardiac autonomic balance favors energy conservation by way of parasympathetic dominance over sympathetic influences. In addition, the HR time series is characterized by beat-to-beat variability over a wide range, which also implicates vagal dominance as the sympathetic influence on the heart is too slow to produce beat-to-beat changes. There is an increasing interest in the study of HR variability among researchers from diverse fields. Low heart rate variability (HRV) is associated with increased risk of all-cause mortality, and low HRV has been proposed as a marker for disease. The following sections discuss the assessment of HRV, the neural control of HRV, and the relationship of HRV to important aspects of self-regulation, including physiological, cognitive, and emotional regulation.

Heart Rate Variability

The basic data for the calculation of all measures of HRV is the sequence of time intervals between heartbeats (Figure 1, top trace). This interbeat interval time series is used to calculate the variability in the timing of the heartbeat. As mentioned previously, the heart is dually innervated by the autonomic nervous system such that relative increases in sympathetic activity are associated with HR increases and relative increases in parasympathetic activity are associated with HR decreases. Thus, relative sympathetic increases cause the time between heartbeats (the interbeat interval) to become shorter and relative parasympathetic increases cause the interbeat interval to become longer. The parasympathetic influences are pervasive over the frequency range of the HR power spectrum, whereas the sympathetic influences 'roll off' at approximately 0.15 Hz (Figure 1, bottom). Therefore, high-frequency HRV represents primarily parasympathetic influences, with lower frequencies (below approximately 0.15 Hz) having a mixture of sympathetic and parasympathetic autonomic influences. The differential effects of the ANS on the sinoatrial node, and thus the timing of the heartbeats, are due to the differential effects of the neurotransmitters for the sympathetic (noradrenaline) and parasympathetic (acetylcholine) nervous systems. The sympathetic effects are slow, on the timescale of seconds, whereas the parasympathetic effects are fast, on the timescale of milliseconds. Therefore, the parasympathetic influences are the only ones capable of producing rapid changes in the beat-to-beat timing of the heart (Figure 1, middle).

Measures of Heart Rate Variability

A variety of measures have been used to operationalize HRV. Long-term measures such as the standard deviation of all interbeat intervals in 24 h, short-term measures such as the standard deviation of 5 min intervals, and beat-to-beat measures such as the root mean square of successive RR differences (RMSSD) have all been used. Respiratory sinus arrhythmia (RSA) is another measure and is defined as the change in heart period corresponding with the inspiratory and expiratory phases of the respiratory cycle. In addition, power spectral analysis of interbeat interval time series is frequently used to quantify HRV. The power spectrum of short-term time series contains two major components – a high-frequency (0.15–0.40 Hz) and a low-frequency (0.01–0.15 Hz) component reflecting cardiac vagal tone and a mixture of vagal and sympathetic influences, respectively. RSA, RMSSD, and the high-frequency component of the power spectrum are closely related, and all reflect vagal cardiac influence. Measures derived from nonlinear dynamics have been used to describe aspects of HRV. One such measure is approximate entropy (ApEn). It quantifies the complexity or irregularity of time series data. This and other 'chaos'-derived parameters have proved valuable in the area of cardiovascular pathology, for example, in discriminating cardiac patient groups with respect to the dynamics of their cardiovascular system. The following sections review some of the many measures available.

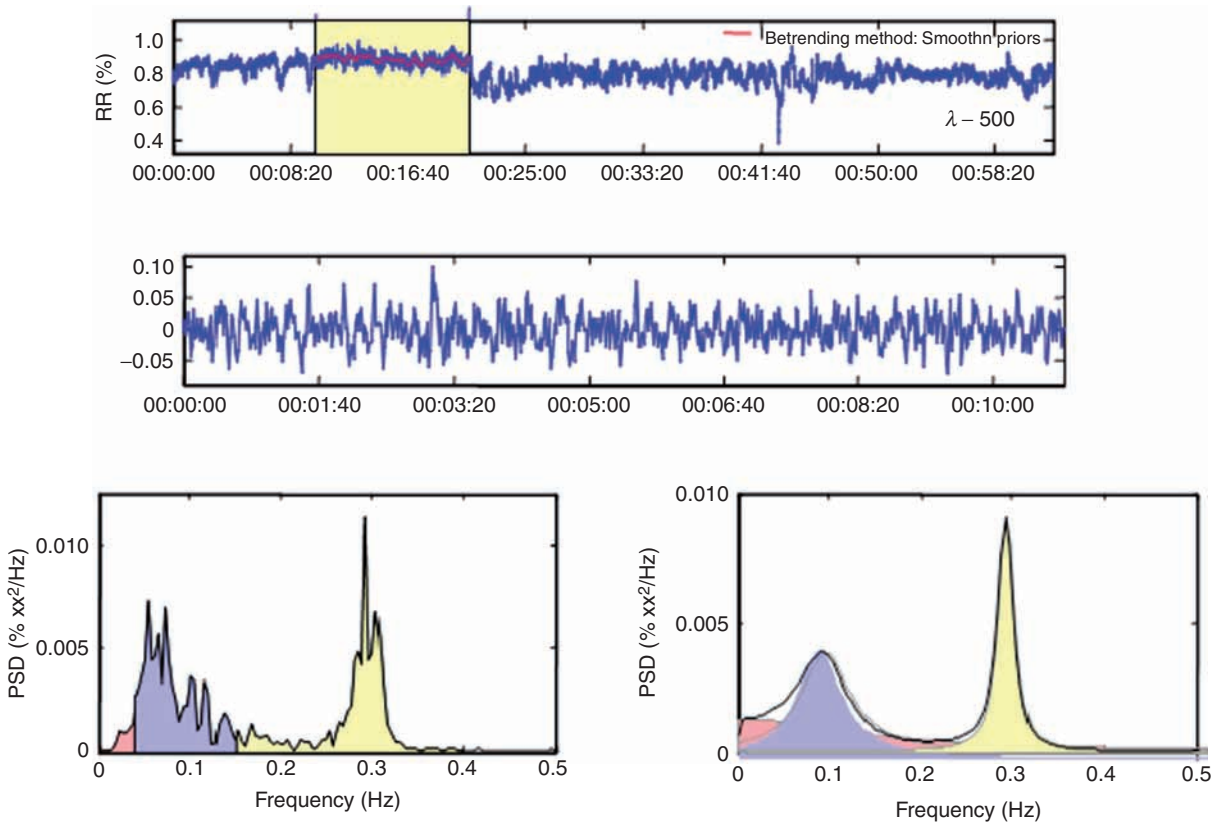

Frequency band	*Peak (Hz)*	*Power (ms²)*	*Power (%)*	*Power (n.u)*
VLF (0–0.04 Hz)	0.0391	23	3.5	
LF (0.04–0.15 Hz)	0.0547	321	49.6	51.4
HF (0.15–0.4 Hz)	0.2930	304	46.9	48.6
Total		648		
LF/HF		1.058		

Frequency band	*Peak (Hz)*	*Power (ms²)*	*Power (%)*	*Power (n.u)*
VLF (0–0.04 Hz)	0.0000	318	33.4	
LF (0.04–0.15 Hz)	0.0977	354	37.2	55.8
HF (0.15–0.4 Hz)	0.2930	281	29.4	44.2
Total		953		
LF/HF		1.263		

Figure 1 The top trace shows an IBI time series. The highlighted portion is shown in more detail in the second trace. The figure and values on the left are from an FFT analysis and those on the right are from an autoregressive analysis of the highlighted data. The tables show the frequency band, peak frequency, power, percentage power, and power in normalized units. Although the values from the two methods are similar, they are not identical. Adapted from the output of free software provided by the Department of Applied Physics, University of Kuopio, Finland. (See Niskanen JP, Tarvainen MP, Ranta-aho PO, and Karjalainen PA (2004) Software for advanced HRV analysis. *Computer Methods and Programs in Biomedicine* 76: 73–81).

Time Domain Indices of HRV

Time domain indices may be based on the interbeat intervals directly or on differences between successive interbeat intervals. In addition, there are both short-term and long-term indices. Short-term indices are based on recordings with a duration on the order of minutes. Long-term indices are usually based on 24-h recordings. Measures that are based on the interbeat intervals directly do not distinguish the autonomic sources of the variability. These indices include the standard deviation of all N–N (normal to normal) intervals (SDNN) and the standard deviation of the average of N–N intervals for each 5 min period over 24 h.

The second type of indices are those based on the differences between successive N–N intervals. These indices are based on the comparison of the lengths of adjacent cardiac cycles and include the percentage of adjacent cycles that are >50 ms apart (pNN50) and the root mean square of successive differences in milliseconds (RMSSD). RMSSD is the most commonly used measure derived from interval differences. This index uses what is called first-differencing in the econometrics literature and acts like a high-pass filter, thus removing long-term trends and slower frequency variability from the signal. Due to the frequency characteristics of the autonomic influences on the heart such that vagal influences cover the full

frequency range and sympathetic influences are primarily restricted to the lower frequencies, RMSSD reflects primarily vagal influences.

Frequency Domain Indices of Heart Rate Variability

Frequency domain analysis gives information about the amount of variance or power in the heart rate or heart period time series explained by periodic oscillations at various frequencies. Power spectral analysis of the time series provides basic information on the amount of variance or power as a function of frequency. Typically, four frequency bands can be distinguished from 24 h recordings and two frequency bands can be reliably discerned in short-term recordings. These are high-frequency (HF; 0.15–0.4 Hz), low-frequency (LF; 0.04–0.15 Hz), very low-frequency (VLF; 0.003–0.04 Hz), and ultra-low-frequency (ULF; $\leq$0.003 Hz). The HF and LF bands are prominent in short-term recordings, with a VLF or direct current (DC) component being associated with aperiodic and artifactual variance (Figure 1). Thus, the VLF and ULF bands are not usually interpretable in short-term recordings. It is important to note that the total spectral power is exactly equal to the time domain variance of the HR time series. This useful relationship can be helpful in troubleshooting spectral analysis algorithms. Moreover, this makes clear that spectral analysis is similar to analysis of variance in that both techniques are used to partition variance and attribute it to various sources.

HF power is primarily parasympathetically mediated. The HF band primarily reflects the respiration-mediated HRV at 0.15–0.4 Hz. The defined frequency band for this parameter usually encompasses the frequency range corresponding to the frequency of normal respiration. Thus, the HF power predominantly measures the variability in HR induced by respiration. This index of vagally mediated cardiac control correlates highly with the time-domain-based measure of RMSSD. A detailed accounting of the other frequency bands is given later.

The measurements of frequency power components described previously are usually expressed as absolute values of power (milliseconds squared). The absolute power measurements represent the variance of the measured signal about its mean value. The so-called normalized scores represent the relative value of each power component in proportion to the total power. In addition, the VLF or DC component is often subtracted from the total power in calculating the normalized values. The DC component is defined as the spectral components with a frequency $<$0.03 Hz. The LF-to-HF ratio has been proposed to reflect the sympathovagal balance.

There are numerous algorithms for doing spectral analysis. Historically, the fast Fourier transform (FFT) was used. However, there are several disadvantages to this approach, including the necessity of specifying the boundaries of the frequency bands in advance. This is called 'windowing' and can lead to serious misinterpretations of the data if they fail to capture the frequencies in which one is interested. For example, in doing spectral analysis on infants or other species such as rodents, the ranges of the frequency bands we have described will not apply. The frequency of respiration varies greatly from infancy to adulthood and across species, so to capture respiratory modulated HRV one has to adjust the 'windows' in an FFT. Whereas bioengineers have called the FFT the 'classical' approach to spectral analysis, more modern approaches based on linear prediction analysis have been developed. The major algorithm in this approach is the autoregressive algorithm, which overcomes the windowing problem by using a true components analysis in which the central frequency and width of the bands are determined empirically and on an individualized basis.

Still more recent developments include the use of wavelets and algorithms for time–frequency analysis. These approaches go even further and do not make assumptions concerning stationarity and thus allow for relatively continuous estimates of spectral power. In addition, numerous indices based on nonlinear dynamical systems or chaos theory have proved to be useful in the analysis of HRV. However, detailed discussion of these techniques is beyond the scope of this article. A particularly important resource for the analysis of complex signals including HRV, called PhysioNet, has been funded by the National Institutes of Health and provides numerous software solutions for the interested researcher.

Whereas the best signal source for the assessment of HRV is the ECG signal, sometimes a researcher will have an HR or interbreath interval (IBI) time series derived from a photoplethysmograph or from a pulse wave such as might be derived from a blood pressure recording. Research has shown that these devices can give adequate time series for the calculation of HRV under certain conditions.

One controversial issue in measuring HRV is the need for respiratory control. That is, given the close links between the cardiovascular and respiratory systems, some researchers suggest that measures of HRV need to take account of respiration. However, not all measures of HRV are correlated with respiration. It was shown in a large sample that ApEn was uncorrelated with respiration frequency. Moreover, this need for respiratory control is predicated on the idea that changes in respiration cause changes in HRV and thus might distort their interpretation. However, for

more than 100 years it has been known that HR might influence the initiation of inspiration. Thus, controls for respiration might throw out the baby with the bathwater, so to speak. Nonetheless, it is useful to examine respiration when interpreting the results of HRV analyses. If one does not have an independent measure of respiration, an index of breathing frequency can be derived from the HR or IBI time series. One way is via spectral analysis, where the central frequency of the high-frequency component of the HRV spectrum is highly correlated with strain gauge measures of respiratory frequency.

A brief overview of the various frequency bands associated with HRV is given in the following sections.

Ultra-Low-Frequency Band: ≤0.003 Hz

The ULF band has shown strong relationships with mortality and therefore an understanding of this band is important. Whereas, as noted previously, all frequencies of HRV reflect autonomic, particularly parasympathetic, influences to some degree, the utility of the ULF band to predict morbidity and mortality appears to lie in its relation with the functional capacity of the patient. Roach and colleagues have shown that SDNN and ULF power in 24 h recordings is related to the range of physical activity that patients engage in and not to differences in autonomic control of the heart *per se*. Given that approximately 1 h is needed to observe approximately ten cycles at this frequency, it is clear that this band cannot be resolved in short-term recordings and is thus not a meaningful index in such circumstances.

Very Low-Frequency Band: 0.003–0.04 Hz

The physiological origins of the VLF band have remained elusive. This band has been variously linked to fluctuations in the renin–angiotensin system and to thermoregulation. We undertook a study to investigate the relation between ambient temperature and HRV with an emphasis on the VLF band. In that study, with humidity controlled, we exposed participants in a within-subject design to both cold (55 °F) and hot (95 °F) conditions as well as to a control condition of 72 °F for 30 min each. Compared to the hot and control conditions, the cold condition was associated with significantly greater VLF power. This increase in VLF power was accompanied by an increase in diastolic blood pressure (DBP). Moreover, this pattern of an increase in VLF power and DBP was unique to the cold condition. These results suggested that the VLF band responded, at least in part, to variations in ambient temperature and thus was associated with thermoregulation. Again, because of the low frequency of these oscillations (approximately two per minute), this band is not well resolved in short-term recordings.

Low-Frequency Band: 0.04–0.15 Hz

There has been a great deal of controversy surrounding the meaning of variations in power in the LF band. Oscillations in this frequency band have been identified for more than 100 years and are associated with baroreflex-mediated blood pressure variations. In this context, it is important to appreciate that blood pressure is under complex control by the ANS. Mean arterial pressure is a function of cardiac output (CO) and total peripheral resistance (TPR). Furthermore, CO is a function of HR and stroke volume (SV). The major determinant of long-term BP regulation is fluid volume. However, the major determinant of short-term BP regulation is the ANS via the baroreflex. The most rapidly modulated component of BP is HR due to its primarily vagal control. As such, the most efficient way to modulate short-term BP is via changes in HR and therefore via changes in CO. SV is primarily regulated by changes in β-adrenergic control and in healthy individuals is relatively constant over a wide range of conditions. This leaves HR as the major factor in changes in CO. TPR is largely associated with α-adrenergic activity and thus is a relatively inefficient way to modulate short-term BP. Taken together, these relations suggest that variations in short-term BP that are thought to be associated with LF power will represent a complex mix of ANS influences, and thus earlier ideas that raw LF power represents activity of the sympathetic nervous system were clearly incorrect. However, this analysis does indicate that LF power should largely reflect parasympathetic influences. The empirical data support this analysis because LF power has been shown to correlate very highly with HF power (0.70 or higher) in numerous studies. Therefore, raw LF power most likely reflects significant parasympathetic nervous system activity. Given that approximately 2 min of data are necessary to provide adequate spectral resolution of this frequency (approximately ten oscillations), LF power can be reliably assessed in short-term recordings.

High-Frequency Band: 0.15–0.4 Hz

There is little controversy about the fact that power in the HF band represents respiratory-modulated HR variations. Moreover, given the transfer function characteristics of the ANS influences on the heart, there is little controversy that the HF band largely reflects parasympathetic influences.

Neural Control of Heart Rate Variability

The neurovisceral integration model has identified a flexible neural network associated with self-regulation and adaptability that might provide a unifying framework within which to view the diversity of observed responses across domains. In this model, a set of neural structures that regulates physiological, behavioral, emotional, and cognitive responses can be indexed via peripheral indices such as HRV. This model emphasizes the role of HRV as an index of self-regulation and its ability to reflect neural feedback mechanisms between the central nervous system (CNS) and the ANS.

At the core of this model is the central autonomic network (CAN). Functionally, CAN can be regarded as an integrated component of an internal regulation system, in which the brain controls visceromotor, neuroendocrine, and behavioral responses that are critical for goal-directed behavior, adaptability, and health. Structurally, CAN includes a number of components, such as the anterior cingulate, insular, orbitofrontal, and ventromedial prefrontal cortices; the central nucleus of the amygdala; the paraventricular and related nuclei of the hypothalamus; the periaqueductal gray matter; the parabrachial nucleus; the nucleus of the solitary tract (NTS); the nucleus ambiguus; the ventrolateral medulla; the ventromedial medulla; and the medullary tegmental field. These components are reciprocally interconnected such that information flows both from the bottom up and from the top down. Importantly for the current discussion, the primary output of CAN is mediated through the preganglionic sympathetic and parasympathetic neurons that innervate the heart via the stellate ganglia and the vagus nerve, respectively. The interplay of these inputs to the sinoatrial node of the heart generates the complex variability that characterizes the healthy HR time series. Furthermore, sensory information from the heart and other peripheral systems is fed back to CAN. Thus, HRV can be used to index a set of neural structures associated with self-regulation and health.

CAN is under tonic inhibitory control via the main inhibitory neurotransmitter within the CNS, γ-aminobutyric acid, in interneurons within the NTS. Disruption of this inhibitory pathway can cause, for example, hypertension and sinus tachycardia, which represents a disinhibition of sympathoexcitatory circuits within CAN.

Both animal and human data support this model. Retrograde viral staining studies in both primates and rodents have identified the central sources for the neural control of the heart. For example, Ter Horst and Postema identified frontal, insula, prelimbic, and anterior cortices as the forebrain concomitants of the parasympathetic influences on the heart in rodents. Although the viral labeling was bilateral, it was predominantly right sided. Similarly, Chuang et al. found predominantly right-sided viral labeling of parasympathetic neurons in the nucleus ambiguous in primates. In humans, using pharmacological blockade, it was found that deactivation of the anterior portions of the brain by sodium amobarbital produced HR increases and vagally mediated HRV decreases that were larger on the right side. In addition, numerous human neuroimaging studies have identified several prefrontal regions associated with HRV, including the medial prefrontal and anterior cingulate cortices.

Thayer and Lane suggested that a common reciprocal inhibitory cortico-subcortical neural circuit serves as the structural link between psychological processes, such as emotion and cognition, and health-related physiological processes, and that this circuit can be indexed with HRV. Thus, because of these reciprocally interconnected neural structures that allow prefrontal cortex to exert an inhibitory influence on subcortical structures, the organism is able to respond to demands from the environment and organize their behavior.

Physiological Regulation

The regulation of physiological systems that are important for health and disease has been linked to HRV. Thayer and Sternberg summarized data linking HRV to glucose regulation, hypothalamic–pituitary–adrenal axis function, and inflammation. In addition, Thayer and Lane reviewed the literature on the relationship between vagal function and the risk for cardiovascular disease (CVD) and stroke. The National Heart, Lung, and Blood Institute of the National Institutes of Health lists eight risk factors for heart disease and stroke (http://www.nhlbi.nih.gov). Six of these factors are considered to be modifiable. Three of these modifiable risk factors are associated with what could be called biological factors: high blood pressure (hypertension), diabetes, and abnormal cholesterol. Three others listed as modifiable could be considered lifestyle factors: tobacco use (smoking), physical inactivity (exercise), and overweight (obesity). Two factors are considered as nonmodifiable: age and family history of early heart disease or stroke. It is interesting to note that there are at least some data that suggest that each of these risk factors is associated with decreased vagal function as indexed by HRV. Furthermore, emerging risk factors for CVD and mortality, such as inflammation and psychosocial factors, are also associated with decreased HRV. In addition, several large epidemiological studies have shown that reduced HRV is a risk factor for all-cause mortality and morbidity. Thus, HRV, in part due

to its relation with physical health and physiological regulatory systems, may be of special relevance to researchers in medicine and biology. Another area of importance for neuroscience researchers is emotional regulation.

Emotional Regulation

We and others have reviewed the literature on the relation between HRV and emotional regulation. Emotional regulation is a valuable skill that has clear implications for health. Emotions represent a distillation of an individual's perception of personally relevant environmental interactions, including not only challenges and threats but also the ability to respond to them. Viewed as such, emotions reflect the integrity of one's ongoing adjustment to constantly changing environmental demands. When the affective system works properly, it promotes flexible adaptation to shifting environmental demands. In another sense, an adequate emotional response represents a selection of an optimal response and the inhibition of less functional ones from a broad behavioral repertoire, in such a way that energy use is matched to fit situational requirements. Resting HRV represents a type of resource that can be brought to bear in situations in which emotional regulation is called for. We have shown that individuals with higher levels of resting HRV, compared to those with lower resting levels, produce context-appropriate emotional responses as indexed by emotion-modulated startle responses and phasic heart rate responses in addition to behavioral and self-reported emotional responses. It has also been shown that phasic increases in HRV in response to situations that require emotional regulation facilitate effective emotional regulation. Therefore, the relation between HRV and emotional regulation will have important implications for those who study the link between emotional states and dispositions such as depression, anxiety, anger and hostility, alexithymia, and physical health. Yet another area of interest for neuroscientists involves research on cognitive functioning and HRV.

Cognitive Regulation

We have reviewed our research on the relationship between HRV and cognitive regulation. Attentional regulation and the ability to inhibit prepotent but inappropriate responses are important for health in a complex environment. Many tasks important for survival in today's world involve cognitive functions such as working memory, sustained attention, behavioral inhibition, and general mental flexibility. These tasks are all associated with prefrontal cortical activity. Deficits in these cognitive functions tend to accompany aging and are also present in negative affective states and dispositions such as depression and anxiety. In a series of studies, we showed that compared to individuals with low levels of resting HRV, those with higher resting levels performed better on tasks that required executive function, such as the Stroop task and the n-back task. Furthermore, we showed that when resting HRV levels were decreased by aerobic detraining, performance on executive function tasks suffered. Stress can also impair cognitive function and may contribute to the cognitive deficits observed in various mental disorders. It is also possible that autonomic dysregulation contributes to decline in attention and cognitive performance with aging.

Conclusions

The neurovisceral integration model proposes that the relation between HRV and the important physiological, cognitive, and emotional regulation functions is due to the ability of HRV to index activity in a flexible network of neural structures that is dynamically organized in response to environmental challenges. The neuroanatomical, pharmacological blockade, and neuroimaging studies provide support for the forebrain influences on HRV. It is hoped that this review will help neuroscientists better understand and contribute to this growing literature.

See also: Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Neuroanatomy; Autonomic Nervous System; Parasympathetic Nervous System; Sympathetic Nervous System.

Further Reading

Ahern GL, Sollers JJ, Lane RD, et al. (2001) Heart rate and heart rate variability changes in the intracarotid sodium amobarbital (ISA) test. *Epilepsia* 42: 912–921.

Benarroch EE (1997) The central autonomic network. In: Low PA (ed.) *Clinical Autonomic Disorders,* 2nd edn., pp. 17–23. Philadelphia: Lippincott–Raven.

Jose AD and Collison D (1970) The normal range and determinants of the intrinsic heart rate in man. *Cardiovascular Research* 4: 160–167.

Lane RD, Reiman EM, Ahern GL, and Thayer JF (2001) Activity in medial prefrontal cortex correlates with vagal component of heart rate variability during emotion. *Brain and Cognition* 47: 97–100.

Levy MN (1997) Neural control of cardiac function. *Bailliere's Clinical Neurology* 6: 227–244.

Pagani M, Lombardi F, Guzzetti S, et al. (1986) Power spectral analysis of heart rate and arterial pressure variabilities as a marker of sympathovagal interaction in man and conscious dog. *Circulation Research* 59: 178–193.

Pincus SM (2001) Assessing serial irregularity and its implications for health. *Annals of the New York Academy of Science* 954: 245–267.

Roach D, Wilson W, Ricthie D, and Sheldon R (2004) Dissection of long-range heart rate variability. *Journal of the American College of Cardiology* 43: 2271–2277.

Saul JP (1990) Beat-to-beat variations of heart rate reflect modulation of cardiac autonomic outflow. *News in Physiological Science* 5: 32–37.

Stein PK and Kleiger RE (1999) Insights from the study of heart rate variability. *Annual Review of Medicine* 50: 249–261.

Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology (1996) Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation* 93: 1043–1065.

Thayer JF and Brosschot JF (2005) Psychosomatics and psychopathology: Looking up and down from the brain. *Psychoneuroendocrinology* 30: 1050–1058.

Thayer JF, Hansen AL, Sollers JJ 3rd, and Johnsen BH (2005) Heart rate variability as an index of prefrontal neural function in military settings. *Proceedings of SPIE* 5797: 71–77.

Thayer JF and Lane RD (2000) A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders* 61: 201–216.

Thayer JF and Lane RD (2007) The role of vagal function in the risk for cardiovascular disease and mortality. *Biological Psychology* 74: 224–242.

Relevant Website

http://www.nhlbi.nih.gov – National Heart, Lung, and Blood Institute.

Hebbian Plasticity

T H Brown, Y Zhao, and V Leung, Yale University, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Hebbian synaptic plasticity is hypothesized to play an important role in storing information in the nervous system and in using pre- and postnatal experience to guide normal development. Named after the Canadian psychologist Donald Hebb, Hebbian plasticity has been securely demonstrated to exist in the hippocampus, neocortex, and amygdala as well as certain invertebrate circuits. Here we review the evolution of the current understanding of Hebbian plasticity – including the fundamental concepts, the neurophysiological evidence and mechanisms, and the resulting neuronal computations.

We start with antecedents to Hebb's conceptualization, showing that the concept of the Hebb synapse has a long evolutionary history. As we shall see, some neuroscientists, such as James Anderson, credit William James with proposing the essential ideas, 60 years before Donald Hebb's monograph. Others, such as Thomas Brown, also emphasize the obvious fact that contemporary quantifications of these earlier ideas are based on neurophysiological and theoretical results that were unavailable to Hebb. We discuss the history of neurophysiological experimentation on Hebbian plasticity within the broader framework of two synaptic phenomena that have been analyzed *in vitro* and *in vivo*: long-term potentiation (LTP), an activity-dependent, persistent, increase in synaptic strength, and long-term depression (LTD), a corresponding decrease in synaptic strength. We note that there are different varieties of LTP and LTD, only some of which are Hebbian.

Historical Antecedents

Contiguity and Associations

The concept of Hebbian plasticity reflects a natural evolution of ideas that have a long history in philosophy and psychology. More than 2000 years ago, Aristotle (*c.* 350 BC) proposed a theory of memory that held spatiotemporal 'contiguity' to be a fundamental principle governing the formation of associations. According to this principle, the closer ideas occur in space and time, the larger the probability of making an association between them. Aristotle's contiguity principle was later revived and extended by British associationists. In the late nineteenth century, contiguity was given a neurophysiological interpretation by the American philosopher and psychologist William James. According to James' 'law of neural habit,' which specifies the neurophysiological conditions that lead to a modification, "When two elementary brain processes have been active together or in immediate succession, one of them on recurring tends to propagate its excitement to the other" (William James, 1890).

Three years later, in 1893, the Italian neuroanatomist Eugenio Tanzi proposed that the functional junction between neurons is the locus of the physical modification responsible for learning. The hypothetical junction was named 4 years later when Charles Sherrington proposed the term 'synapse' (originally 'synapsis'). By the beginning of the twentieth century, the synaptic hypothesis had been embraced by the famous Spanish neuroanatomist Ramón y Cajal and was being presented in introductory psychology texts. At about the same time, Ivan Pavlov was describing the acquisition of conditional responses in theoretical terms that were similar to those of William James. The thoughts of Tanzi, James, Sherrington, Cajal, and Pavlov were combined and underwent further evolution with the publication of *The Organization of Behavior* by DO Hebb in 1949. In this influential monograph, Hebb proposed that the synapse is the locus of the physiological change that represents stored information and that the condition for causing the synaptic modification is the co-occurrence of activity between synaptically connected neurons.

Nothing in Hebb's monograph suggests an influence from what we regard today as his most obvious intellectual predecessors, including William James. At the network level James was thinking of associative learning as being driven by "a conjunctive learning rule that looks very much like the architectures for modern models of cognition," according to James Anderson, whose own work on associative memory structures has helped us better understand "what Hebbian synapses build." Donald Hebb evidently regarded his own ideas about synaptic plasticity to be among his less original contributions. According to Hebb's now-famous postulate:

> When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased. (Hebb, *The Organization of Behavior*, 1949, p. 62)

Although Hebb's postulate was not exceptionally novel, it did furnish a convenient qualitative starting point for others (such as Gunther Stent, John Hopfield, and James Anderson) who were interested in neurobiological or quantitative formulations of this idea. Hebb was necessarily imprecise about what was meant by 'firing' the cell, because his monograph predated any meaningful understanding of cellular neurophysiology. He knew nothing about the statistical and biophysical factors that contribute to the timing and modulation of the initiation, spread, or suppression of Na^+- and Ca^{2+}-dependent spikes along the various parts of a neuron.

Adaptive Computations

Hebb also did not attempt to quantify his postulate into a formal 'learning rule,' or algorithm. That task was left to the next group of theoreticians (such as David Willshaw, Christoph von der Malsburg, Terrence Sejnowski, Teuvo Kohonen, James Anderson, William Levy, and Erkki Oja), who began to appreciate and explore the computational implications of Hebbian plasticity for associative learning and self-organization. The earliest useful thoughts about how single neurons or groups of neurons might compute were advanced by Warren McCulloch and Walter Pitts in 1943 in their influential paper, "A logical calculus of the ideas imminent in nervous activity." A decade later, psychologist Frank Rosenblatt furnished an early demonstration of how circuits composed of adaptive, 'neuronlike' processing elements can exhibit an interesting form of learning.

Much of the subsequent research on adaptive neural computing can be seen as variations or improvements on Rosenblatt's 'perceptron learning.' In 1960, Bernie Widrow and Marcian Hoff developed a learning algorithm and associated mathematical theory that elucidates how activity input vectors can be adaptively mapped onto desired output vectors using the so-called delta rule for changing connection strengths. The delta rule changes connections as a function of an error term that is computed from the difference between the current output and a 'desired output.' This is called 'supervised learning,' because the desired output is furnished by an external 'teacher signal.' The versatile and popular backward-error-propagation algorithm is a mathematical extension of the delta rule.

Psychologists Robert Rescorla and Allan Wagner, in 1972, distilled into a simple 'learning equation' for Pavlovian conditioning a computationally simple version of the delta rule. The Rescorla–Wagner (R–W) learning rule furnished a valuable connection between error-driven learning algorithms and the variables and facts associated with Pavlovian conditioning. The R–W learning rule accounted for a surprisingly large body of experimental research and made novel and testable predictions. In the R–W rule, the unconditional stimulus (US) 'supervises' learning in that it furnishes a global error signal to populations of neurons. A most significant feature of the R–W equation is that this interpretation of the error signal later facilitated understanding how Pavlovian conditioning phenomena can be computed using Hebbian plasticity and conventional neurophysiological principles.

Modern Hebbian learning rules include bidirectional changes in synaptic strength, for theoretical and empirical reasons explained by Thomas Brown and co-workers in 1990. We often think of associative LTP working together with a form of LTD to store or represent some measure of the contiguity, covariation, or contingency between pre- and postsynaptic activity levels. In a sense, Hebb only addressed half of what we usually recognize as a useful learning rule – namely, the conditions that cause an increase in synaptic strength. Hebb's postulate is loosely captured by the mnemonic, 'cells that fire together, wire together.' Starting in the 1970s, theoreticians, such as Terrence Sejnowski, began to address the other half of the learning rule. Rules for decreasing synaptic efficacy are captured by the mnemonic, 'firing out of synch, weakens the synaptic link.'

Contemporary Understanding of Hebbian Plasticity

Neurophysiological studies and formal theoretical analysis combined to bring about further evolution of the idea of Hebbian plasticity. Today, the essence of a Hebbian computation is one that strengthens synaptic connections as a function of the conjunction or positive contingency or covariation between pre- and postsynaptic activities. As already noted, modern algorithms allow for bidirectional control of synaptic strength. The contemporary concept of Hebbian plasticity recognizes three defining features (interactive mechanism, time-dependent mechanism, and local mechanism) summarized in the following sections.

Interactive Mechanism

A Hebbian synaptic modification depends on the interaction between presynaptic and postsynaptic activity. An interactive mechanism means that whether a modification occurs depends on activity on both sides of the synaptic cleft. Changes that depend on either presynaptic or postsynaptic activity alone, or on the linear summation of changes attributable to activity in the presynaptic and postsynaptic sides of the synapses, would be termed 'noninteractive' and are non-Hebbian.

Time-Dependent Mechanism

Hebbian synaptic modifications are driven by a time-dependent mechanism, meaning that the change is produced by the co-occurrence of pre- and postsynaptic activity. Synaptic enhancement depends on the exact time interval between presynaptic and postsynaptic activities. This contrasts with learning algorithms in which modifications are only a function of the output error – not the timing of activity. In theoretical studies, the synaptic learning rule or algorithm is sometimes based on the covariance between presynaptic and postsynaptic activities. The critical 'activities' may be chemical or electrical. The timing of pre- and postsynaptic activity is what drives synaptic modification. The modification, in turn, can be regarded as representing the contiguity, covariation, or contingency – assessed over some time period – between pre- and postsynaptic activity.

Local Mechanism

Hebbian modifications are produced by a local mechanism in the sense that all of the information that is necessary to drive the synaptic change is available to the synapse as a consequence of ordinary synaptic transmission. This is in contrast to strictly supervised learning mechanisms, such as the 'back propagation' algorithm, which use a systems-level, external 'teacher' signal to instruct change on a synapse-by-synapse basis. Such learning schemes are not found in neurobiology. They would incur severe neuroanatomical overhead because they require independent presynaptic teaching signals to target every modifiable synapse. Furthermore, the environment usually does not afford an organism the information required to compute the error terms. On the other hand, there are neurobiological examples of 'diffuse' or 'global' reinforcement signals that control the induction or consolidation of plasticity in selected populations of synapses.

Hebbian Plasticity and Associative LTP

Technology for Finding Hebbian Plasticity

The experimental discovery of Hebbian plasticity was partly a by-product of research on LTP, evidence of which was first reported by Tim Bliss and Terje Lomo in 1973. They observed that repetitive stimulation of an afferent input to the dentate gyrus could cause a persistent increase in the size of the extracellular field potentials recorded in the dentate gyrus and evoked by the same afferent input. One limitation of this approach was that *in vitro* recordings of field potentials usually lack the spatiotemporal resolution necessary to make secure inferences about synaptic modifications.

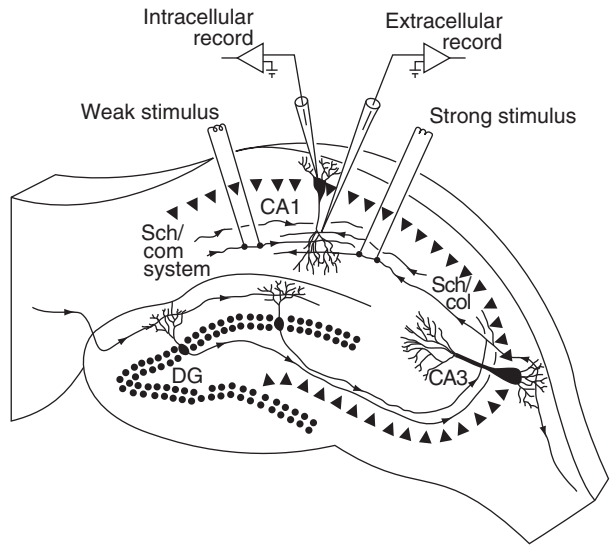

Figure 1 Associative long-term potentiation in the hippocampal brain slice preparation. Interactions between two excitatory afferent inputs to a CA1 pyramidal neuron were examined. By adjusting the strengths of the stimuli delivered to the two input pathways, one was made to generate a weak synaptic response and the other a strong synaptic response. Simultaneous tetanic stimulation of both input pathways resulted in subsequent long-term enhanced synaptic efficacy in the weak input under conditions in which the same tetanic stimulation of either input alone failed to have this effect. Sch/com, Schaffer collateral/commissural; DG, dentate gyrus. Adapted from Barrionuevo G and Brown TH (1983) Associative long-term potentiation in hippocampal slices. *Proceedings of the National Academy of Sciences of the United States of America* 80: 7347–7351, with permission. Copyright 1983 by National Academy of Sciences, USA.

Numerous *in vitro* studies conducted by Thomas Brown, Daniel Johnston, and co-workers, done in the 1980s, were designed to solve this problem by recording from individual neurons bathed in defined media.

The hippocampal brain slice proved to be particularly useful because much of the intrinsic local circuitry remains intact in a transverse slice. Figure 1 illustrates the classical 'trisynaptic circuit,' which emphasizes three projections: from entorhinal cortex to the dentate gyrus (DG), from DG to hippocampal region CA3, and from region CA3 to region CA1. Within the hippocampus proper, the most commonly studied synapse is the Schaffer collateral/commissural (Sch/com) input to pyramidal neurons of the CA1 region. These afferents arise from ipsilateral CA3 cells and from the ipsilateral and contralateral hippocampus. As discussed next, the Sch/com synapses were the first to reveal an underlying Hebbian mechanism.

Associative Nature of Hebbian LTP

LTP has been studied most extensively in the Sch/com synapses of the hippocampal formation (Figure 1). These synapses exhibit three classical properties (associativity, cooperativity, and input specificity) that

are signatures of Hebbian modifications. Associativity refers to a synaptic modification in one afferent input to a neuron that is conditional upon coactivity in a separate input to the same neuron. Cooperativity refers to the fact that the probability of inducing LTP increases with the number of stimulated fibers. The idea was that multiple axons 'cooperate' in triggering LTP. Input specificity refers to the fact that LTP induction is restricted to just those synapses that receive the necessary or critical activation at the right time. All three of these classical properties are signatures of a Hebbian modification.

The associative nature of LTP was first suggested in studies that William Levy and Oswald Steward performed *in vivo* on the perforant-path-to-granule-cell synapse (see Figure 1). Stimulation electrodes placed bilaterally in the entorhinal cortices elicited extracellular population field potentials in the granule cell layer of the dentate gyrus. Due to differences in the number of ipsilateral and contralateral axonal projections, stimulation of the entorhinal cortex ipsilateral to the recording site produced a large field potential response (strong input), whereas stimulation of the entorhinal cortex contralateral to the recording site produced a small field potential response (weak input). Stimulation of the (weak) contralateral input alone failed to induce LTP in that pathway. However, LTP could be induced in the contralateral input when it was stimulated concurrently with the converging (strong) ipsilateral input. Importantly, activity in the ipsilateral input, by itself, did not induce LTP in the contralateral input (a phenomenon called heterosynaptic LTP; see later, section titled 'Heterosynaptic LTP').

Associative LTP and Hebbian Plasticity

Using the hippocampal brain slice preparation, in 1983, German Barrionuevo and Thomas Brown demonstrated 'associative LTP' intracellularly in the Sch/com synaptic inputs to hippocampal CA1 neurons (see Figure 1). The experiment examined interactions between weak (W; small number of stimulated afferents) and strong (S; large number of stimulated afferents) inputs in a hippocampal brain slice. Tetanic (high-frequency) stimulation of the W input by itself failed to produce LTP in that pathway unless this stimulation was temporally paired with tetanic stimulation of the S input. LTP was induced in the W input only when its activity was associated with activity in the S input. S stimulation never induced LTP in the W input, meaning that heterosynaptic LTP (see later) did not occur in these synapses. The use of brain slices enabled pharmacological manipulations that ruled out the possibility that the overall increase in synaptic efficacy resulted from LTD in inhibitory interneurons.

In 1986, Stephen Kelso and Thomas Brown then examined the spatiotemporal specificity of associative LTP in these synapses, calling attention to the relationships between W and S inputs and Pavlovian conditioned and unconditioned stimuli (CSs and USs), respectively. W–S pairings were ineffective in inducing associative LTP when the periods of W and S stimulations did not overlap temporally, but were separated by a small fraction of a second. By using two W inputs (W1 and W2), these experiments simultaneously demonstrated the spatial specificity of the synaptic changes. Modifications only occurred in the synapses associated with the particular W input (W1 or W2) with an activity that overlapped temporally with that of the S input. More recently, Nelson Spruston and co-workers further refined the precise timing requirements for the induction of associative LTP.

The demonstration of associative LTP did not prove that a Hebbian mechanism was involved, because there was no direct demonstration that postsynaptic activity or depolarization *per se* was critical to the mechanism for LTP induction. A popular alternative possibility was that activity in the S input released a critical amount of an 'LTP factor' that enabled LTP in the coactive W input. Determining whether associative LTP in the Sch/com synapses resulted from a Hebbian mechanism required an experiment designed to force or prevent simultaneous presynaptic and postsynaptic activity. If a critical amount of postsynaptic depolarization or activity could be shown to be a necessary condition for inducing LTP in active synapses, this would be persuasive evidence in favor of a Hebbian mechanism.

There had been two previous failures to discover Hebbian plasticity. In 1982, a year before the experiments of German Barrionuevo and Thomas Brown, Holger Wigstrom and co-workers incorrectly concluded that the postsynaptic membrane potential had no effect on LTP induction. The error resulted from not appreciating the large difference that can exist between the voltage recorded by an intrasomatic microelectrode and the actual voltage across the dendritic membrane that is immediately postsynaptic to the active synapses. Two years later, Thomas Carew and co-workers repeated this same error in studies of synaptic plasticity in a ganglion of *Aplysia*.

In 1986, Thomas Brown and co-workers corrected the error in hippocampus. They furnished definitive experimental evidence for a Hebbian conjunctive mechanism in the Sch/com synapses of hippocampal region CA1 (Figure 1). The rationale for this study was simple. If the synapses were Hebbian, then it should be possible to induce LTP under experimental conditions, in which one substitutes for the usual S input direct depolarization of the postsynaptic

neuron via the recording microelectrode. Results are shown in Figure 2, which plots the mean amplitude of the excitatory postsynaptic potentials (EPSPs) evoked in the CA1 pyramidal neurons as a function of time and the various experimental manipulations.

The EPSPs were tested every 12 s by delivering a single extracellular stimulation to the Sch/com afferent inputs. Once a stable EPSP baseline was established, the postsynaptic neuron was depolarized alone (Figure 2, Depol. Alone) to demonstrate that this manipulation was without effect. Next, a tetanic stimulation (100 Hz) was given to the W input while the postsynaptic soma was maintained at −80 mV under voltage clamp conditions (Figure 2, 100 Hz + voltage clamp). Then the same presynaptic tetanic stimulation was given while the postsynaptic cell was depolarized under current clamp conditions (Figure 2, 100 Hz + depol.).

These studies demonstrated, unequivocally, that Hebbian plasticity does exist in the hippocampus and that this Hebbian mechanism could account for the spatiotemporal specificity of the activity–modification relationship. Using two separate weak inputs (W1 and W2) to the same cell, it was demonstrated that only the W input that was activated during postsynaptic depolarization exhibited LTP. Temporal specificity was evidenced by the fact that, to be effective, postsynaptic depolarization and presynaptic stimulation (at 100 Hz) had to occur nearly simultaneously.

In agreement with these findings, Holger Wigstrom and co-workers also published a brief note in 1986 suggesting that their previous conclusion had been erroneous. Starting in the 1990s, work by David Glanzman and co-workers corrected the earlier error in *Aplysia*. Glanzman's recent work suggests that Pavlovian conditioning in *Aplysia* results from a mechanism that requires postsynaptic Ca^{2+} influx and *N*-methyl-D-aspartate receptors (NMDARs). The mechanism in *Aplysia* appears to be remarkably similar to that responsible for NMDAR-dependent associative LTP in the Sch/com synapses.

Heterosynaptic LTP

Heterosynaptic LTP, in contrast to associative LTP, is not input specific. Heterosynaptic LTP implies that stimulation of one pathway causes synaptic changes in another, unstimulated pathway. For a variety of reasons, heterosynaptic LTP, which has been studied most extensively *in vitro* in *Aplysia*, is not considered to be a viable candidate synaptic substrate for associative learning and memory in mammals. However, David Glanzman and co-workers raise the interesting possibility that heterosynaptic plasticity in *Aplysia* could serve to modulate the induction and/or persistence of Hebbian synaptic modifications.

Anti-Hebbian Plasticity

As summarized, Hebbian plasticity is one that uses an interactive, time-dependent, and highly local mechanism to increase synaptic efficacy as a function of the contiguity, contingency, or association between pre- and postsynaptic activity. Synapses that decrease

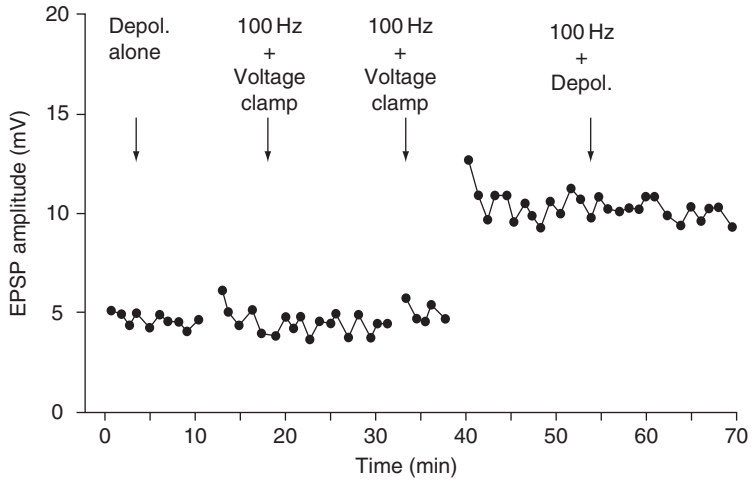

Figure 2 Demonstration of Hebbian plasticity in the hippocampus. Illustrated are excitatory postsynaptic potential (EPSP) amplitudes as a function of the time of occurrence (arrows) of three manipulations: the postsynaptic cell is first given an outward current step, resulting in postsynpatic spiking but no potentiation of the EPSP amplitude (depol. alone). Next, the postsynaptic soma is voltage clamped at −80 mV while delivering presynaptic stimulation trains of 100 Hz, resulting in no EPSP potentiation (100 Hz + voltage clamp), or an outward current step is delivered, causing the cell to fire, at the same time that the presynaptic input is stimulated at 100 Hz, resulting in potentiation of the EPSP amplitude (100 Hz + depol.). Each point is the average of five consecutive EPSP amplitudes. Adapted from Kelso SR, Ganong AH, and Brown TH (1986) Hebbian synapses in hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* 83: 5326–5330, with permission. Copyright 1986 by National Academy of Sciences, USA.

efficacy under the same conditions are sometimes termed 'anti-Hebbian.' Anti-Hebbian plasticity has been most commonly studied in certain neurons of the cerebellum that are important in adaptive motor changes. In these neurons, conjunctive presynaptic and postsynaptic activity results in long-term synaptic depression (LTD). The best understood example of associative LTD occurs at the parallel fiber-to-Purkinje cell synapse, where the mechanism is thought to involve an increase in the Ca^{2+} concentration ($[Ca^{2+}]_i$) in parts of the postsynaptic neuron. Anti-Hebbian synapses were anticipated by the theoretical studies of James Albus in 1971. Because the output from the Purkinje cells is inhibitory, associative LTD in the cerebellum results in disinhibition, or a net increase in excitatory drive on the ultimate target cells.

Induction of Associative LTP

Much evidence has accumulated suggesting a role for NMDARs in the induction of associative LTP. Two properties of the NMDAR are pertinent to a role in the induction of associative LTP. First, NMDARs are permeable to Ca^{2+} (in addition to Na^+ and K^+). This is significant because the intracellular Ca^{2+} concentration near the postsynaptic membrane is thought to play a critical role in inducing associative LTP. Second, the channel permeability is a function of both presynaptic and postsynaptic events, thus serving as a 'coincidence detector.'

The NMDAR conductance depends on coactivation of both pre- and postsynaptic neurons. Presynaptic activity is required for release of glutamate, the endogenous neurotransmitter that binds to NMDARs on the postsynaptic membrane. Postsynaptic activity is required to relieve a Mg^{2+} block of the NMDAR. The NMDAR-mediated conductance is therefore both ligand gated and voltage gated, allowing Ca^{2+} entry only when presynaptic release co-occurs with sufficient depolarization of the postsynaptic membrane to relieve the Mg^{2+} block. In some neurons, the elicitation of back-propagating action potentials (BAPs; see later) may also be necessary to achieve sufficient depolarization of the postsynaptic dendritic membrane.

We can now envision how these biophysical events could account for the properties of associativity, cooperativity, and input specificity. Active synapses release glutamate, which can bind to postsynaptic NMDARs, causing Ca^{2+} influx into dendritic spines on the postsynaptic neuron. Ca^{2+} influx into the spine acts on local enzymes that control LTP induction, resulting in input specificity. The Ca^{2+} influx into the spine occurs only when the synaptic input is strong enough to depolarize the postsynaptic membrane sufficiently to relieve the Mg^{2+} block, giving rise to cooperativity. Activity in a W input is not sufficient to depolarize the postsynaptic membrane sufficiently to remove the Mg^{2+} block, unless this activity is properly timed in relationship to activity in an S input to the same cell. If the S input elicits BAPs, this causes additional depolarization that can increase Ca^{2+} influx. Associativity results from the fact that the depolarization of the postsynaptic membrane resulting from combined activity in the W and S inputs is necessary to induce LTP in the W input.

Biophysical Models of Hebbian Plasticity

The laboratories of Thomas Brown and William Levy independently created biophysical models that elucidate the role of NMDARs in Hebbian plasticity. In both computational models, nearly concurrent presynaptic activity (producing glutamate release) and postsynaptic activity (relieving the Mg^{2+} block of the channel) allowed Ca^{2+} influx into the dendritic spine of the postsynaptic neuron. The increased $[Ca^{2+}]_i$ in a critical region of the dendritic spine, close to the NMDAR, activated Ca^{2+}-dependent enzymes, such as Ca^{2+}/calmodulin-dependent kinase II (CaMKII), that are thought to play a key role in LTP induction.

Both models wrestled with the fact that the voltage dependence of the NMDAR was insufficient to account for what is known about associative LTP. Both laboratories attempted to solve this problem by taking into account the additional nonlinearity in the relationship between postsynaptic depolarization and Ca^{2+}-dependent enzyme activation. Brown and co-workers also examined an additional nonlinearity – the effects of BAPs. The main finding was that back-propagating Na^+ spikes were only effective in promoting LTP induction if they in turn elicited dendritic Ca^{2+} spikes.

Dendritic Ca^{2+} spikes were included in the model for several reasons. First, Robert Wong and co-workers, in 1979, had shown that Ca^{2+} spikes can be elicited in the dendrites of hippocampal pyramidal neurons. Second, Stephen Kelso and co-workers, in 1986, noted that, whereas Hebbian LTP can be readily induced in hippocampal neurons under conditions in which Na^+ spiking was blocked, Ca^{2+} spikes continued to be elicited when the cell was sufficiently depolarized to enable LTP induction. Finally, in 1987, Thomas Richardson and associates provided evidence that BAPs can be readily elicited in hippocampal neurons by synaptic inputs. Their inference was based on extracellular recordings combined with current–source–density (CSD) analysis. The CSD data strongly

suggested that synaptic inputs can elicit spikes in the soma that then propagate into the dendrites. The role of BAPs in promoting or enabling Hebbian plasticity was only an interesting theoretical possibility in 1990.

However, further work by Johnston and co-workers provided compelling evidence in support of this possibility. Using patch clamp recordings from the soma and dendrites, combined with direct Ca^{2+} imaging, Johnston and colleagues furnished a detailed understanding of the elicitation and back-propagation of Na^{+} and Ca^{2+} spikes as well as the manner in which dendritic spikes are controlled by K^{+} currents. Even more important, in the present context, is the finding that the elicitation of BAPs appears to be essential for inducing LTP in some synapses. As noted previously, BAPs furnish an additional nonlinearity that can be computationally desirable. BAPs also introduce useful dynamics into synaptic activity–modification relationships in terms of the critical temporal relationship among W activity, S activity, and dendritic spiking.

Distribution and Varieties of LTP

Not all forms of LTP are Hebbian, nor do they all involve the NMDAR. In the peripheral nervous system, the rat superior cervical sympathetic ganglion exhibits a non-Hebbian form of LTP that relies on the nicotinic acetylcholine receptor. Robust and long-lasting LTP is also reported in the crayfish neuromuscular junction, where glutamate is the neurotransmitter. It is uncertain whether postsynaptic depolarization plays a role in this modification. As we have already noted, an NMDAR-dependent form of Hebbian LTP has also been described in *Aplysia* ganglia and fruit flies. Interestingly, the mossy fiber synapses of the hippocampus exhibit a Hebbian form of LTP that does not rely upon NMDARs.

Recent interests in NMDAR-independent forms of Hebbian LTP have focused on the role of Ca^{2+} influx triggered by BAPs. Timothy Teyler and co-workers have raised the interesting possibility that NMDAR-independent LTP is preferentially involved in longer-term information storage in cortex. Joseph LeDoux has made the same suggestion for information storage in the amygdala. Much more research is clearly needed on the critical question of how the various types of LTP are (or are not) related to specific aspects or forms of learning and memory.

Summary

In reviewing the evolution of the current understanding of Hebbian plasticity, including supporting neurophysiological evidence and biophysical mechanisms, it is seen that the suggestion by James and Hebb – that learning and memory emerge from a contiguity-based neuronal modification – dovetails nicely with current research on associative LTP and LTD. Today we are experiencing a grand confluence of ideas from philosophy, psychology, neurophysiology, and computation. Theoretical studies leave little doubt about the impressive learning and self-organizing capabilities of circuits containing Hebbian synapses. Experimental studies have shown that Hebbian plasticity is widespread throughout the nervous system and across phylogeny. A reasonable hypothesis is that Hebbian plasticity plays a fundamental role in some of the adaptive functions of the nervous system. The most pressing need is to elucidate precisely the relationships between various types of LTP (and LTD) and specific aspects or forms of learning and memory – what Daniel Johnston and co-workers have termed 'the missing link.'

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Depression: Cerebellum; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); NMDA Receptors and Development; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Mechanisms of Learning.

Further Reading

Anderson JA (1985) What Hebb synapses build. In: Levy WB, Anderson JA, and Lehmkuhle S (eds.) *Synaptic Modification, Neuron Selectivity, and Nervous System Organization*, pp. 153–173. Hillsdale, NJ: Erlbaum.

Barrionuevo G and Brown TH (1983) Associative long-term potentiation in hippocampal slices. *Proceedings of the National Academy of Sciences of the United States of America* 80: 7347–7351.

Bliss TVP and Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* 232: 331–356.

Brown TH, Byrne JH, LaBar K, et al. (2004) Learning and memory: Basic mechanisms. In: Byrne JH and Roberts JL (eds.) *From Molecules to Networks: An Introduction to Cellular and Molecular Neuroscience*. San Diego, CA: Academic Press.

Brown TH, Kairiss EW, and Keenan CL (1990) Hebbian synapses: Biophysical mechanisms and algorithms. *Annual Review of Neuroscience* 13: 475–511.

Golding NL, Staff NP, and Spruston N (2002) Dendritic spiking as a mechanism for cooperative long-term potentiation. *Nature* 418: 326–331.

Hebb DO (1949) *The Organization of Behavior*. New York: Wiley (Interscience).

James W (1890; 1950) *The Principles of Psychology*. New York: Dover Publications.

Kelso SR and Brown TH (1986) Differential conditioning of associative synaptic enhancement in hippocampal brain slices. *Science* 232: 8587.

Kelso SR, Ganong AH, and Brown TH (1986) Hebbian synapses in hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* 83: 5326–5330.

Levy WB (1985) Associative changes at the synapse: LTP in the hippocampus. In: Levy WB, Anderson JA, and Lehmkuhle S (eds.) *Synaptic Modification, Neuron Selectivity, and Nervous System Organization*, pp. 5–33. Hillsdale, NJ: Erlbaum.

Magee JC and Johnston D (2005) Plasticity of dendritic function. *Current Opinion in Neurobiology* 15: 334–342.

Murphy GG and Glanzman DL (1997) Mediation of classical conditioning in *Aplysia californica* by long-term potentiation of sensorimotor synapses. *Science* 278: 467–471.

Roberts AC and Glanzman DL (2003) Learning in *Aplysia*: Looking at synaptic plasticity from both sides. *Trends in Neurosciences* 26: 662–670.

Zador A, Koch C, and Brown TH (1990) Biophysical model of a hebbian synapse. *Proceedings of the National Academy of Sciences of the United States of America of the United States of America* 87: 6718–6722.

Helix–Loop–Helix (bHLH) Proteins: Hes Family

R Kageyama, T Ohtsuka, R Ohsawa, and J Hatakeyama, Kyoto University, Kyoto, Japan

© 2009 Elsevier Ltd. All rights reserved.

Structure and Transcriptional Activity

Hes genes are mammalian homologs of *Drosophila hairy* and *Enhancer of split* [*E*(*spl*)], which are known to inhibit neurogenesis. There are seven members in the Hes family, and each member has a conserved basic helix–loop–helix (bHLH) domain (Figures 1(a) and 1(b)), which is required for dimer formation and DNA binding. Hes1 forms homodimers and binds with high affinity to the DNA motif, the N-box (CACNAG) or the class C site (CACGCG), and with low affinity to the E-box (CANNTG). In contrast, many other bHLH factors, such as the proneural bHLH factors Mash1 and E47, form heterodimers and bind only to the E-box with a higher affinity. In the basic region, all Hes factors have a proline residue (Figures 1(a) and 1(b), asterisk), unlike other bHLH factors, suggesting that this proline residue might be responsible for the Hes-specific DNA binding activity. At the C-terminus, Hes factors have Trp-Arg-Pro-Trp sequence (the WRPW domain), which functions as a repression domain by recruiting the co-repressors TLE/Grg factors, homologs of *Drosophila* Groucho (Figures 1(a) and 2(a)). Thus, Hes factors act as transcriptional repressors, while proneural bHLH factors function as transcriptional activators. One of the targets of Hes1 is the proneural bHLH gene *Mash1*. There are class C sites in *Mash1* promoter, and Hes1 represses *Mash1* expression by directly binding to these class C sites. Hes1 also forms heterodimers with bHLH activators such as Mash1 and E47, but these heterodimers do not bind to the DNA, indicating that Hes1 inhibits the activities of proneural bHLH activators by forming non-DNA-binding heterodimers (Figure 2(b)). Thus, Hes factors repress gene expression by two different mechanisms: by recruiting co-repressors (active repression) and by forming non-DNA-binding heterodimers (passive repression) (Figure 2).

There are several Hes-related factors such as Hey/Hesr/Herp. They have a Hes-related bHLH domain and WRPW-related sequence and form heterodimers with Hes. Interestingly, Hes/Hey heterodimers bind to the class C site with a higher affinity and repress transcription more efficiently than Hes and Hey homodimers.

Between the bHLH and WRPW domains, Hes factors have the Orange domain (Figure 1(a)), which consists of two amphipathic helices. Some bHLH factors preferentially form heterodimers with each other while others do not, and this specificity for bHLH factor interaction is regulated by the Orange domain. This domain is also involved in transcriptional repression, although a co-repressor mediating this repression is not known.

Hes3 has two forms, Hes3a and Hes3b, which are generated by alternative promoters and first exons. Hes3a lacks the amino-terminal half of the basic region and thus cannot bind to the DNA, although it forms a non-DNA-binding heterodimer complex with other bHLH factors. Hes3a is specifically expressed by Purkinje neurons in the cerebellum, but the function of Hes3a in Purkinje neurons remains to be determined. In contrast, Hes3b has a complete basic region and represses transcription by binding to the N-box. Hes3b is expressed by neural progenitors (see following sections), unlike Hes3a. In the rest of this article, Hes3b is designated as Hes3.

Unlike other Hes factors, Hes6 does not bind to the N-box or the E-box, although it has a complete bHLH domain. Furthermore, Hes6 has an opposite activity to Hes1: Hes6 antagonizes Hes1 by physical interaction and supports Mash1 activity, thereby promoting neuronal differentiation. The loop region of Hes6 is shorter than that of the other Hes factors (Figure 1(b)), and it is likely that this short loop region is responsible for such unique activities of Hes6.

Regulation of *Hes* Gene Expression

Hes1 and *Hes5* expression is regulated by Notch signaling in the nervous system (Figure 3). The transmembrane protein Notch is activated by its ligands such as Delta. Upon activation, the intracellular domain of Notch (ICD) is cleaved from the transmembrane region and translocated into the nucleus. In the nucleus, the ICD forms a complex with the DNA-binding protein RBP-J, which binds to the *Hes1* and *Hes5* promoters. Before Notch activation, RBP-J interacts with a histone deacetylase co-repressor complex and functions as a transcriptional repressor, but Notch activation disrupts the repressor complex and leads to the formation of the ICD–RBP-J complex, which functions as a transcriptional activator. Thus, Notch activation induces expression of *Hes1* and *Hes5* genes. However, while *Hes5* expression is lost in the absence of Notch signaling, *Hes1* expression is not in some regions, suggesting that *Hes1* expression is additionally controlled

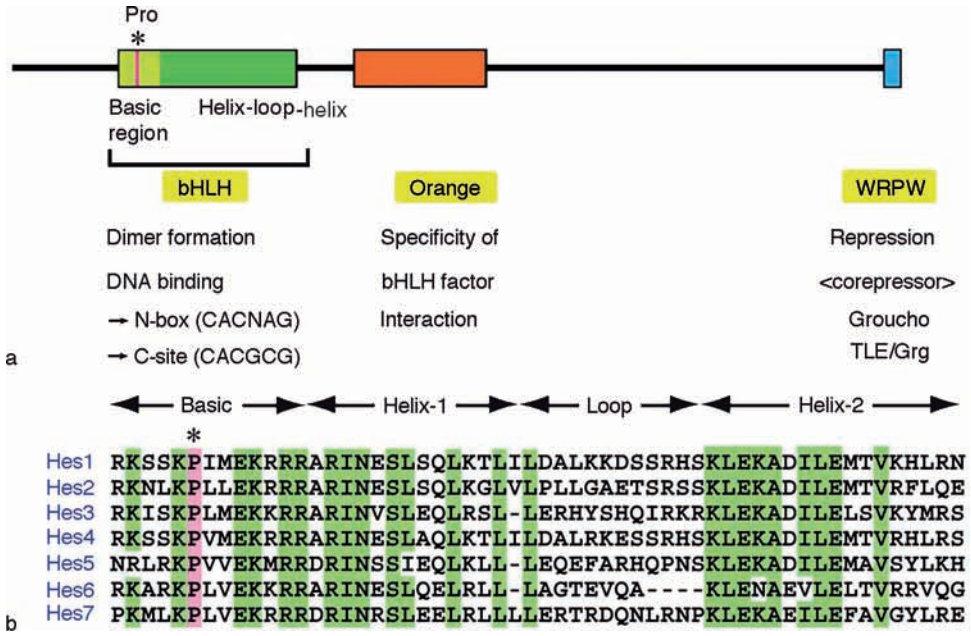

Figure 1 Features of Hes bHLH factors. (a) The bHLH, Orange, and WRPW domains and their functions. (b) Sequence alignment of the bHLH domain of Hes factors. Proline is conserved in the middle of the basic region (asterisk). The loop region of Hes6 is shorter than that of other Hes factors.

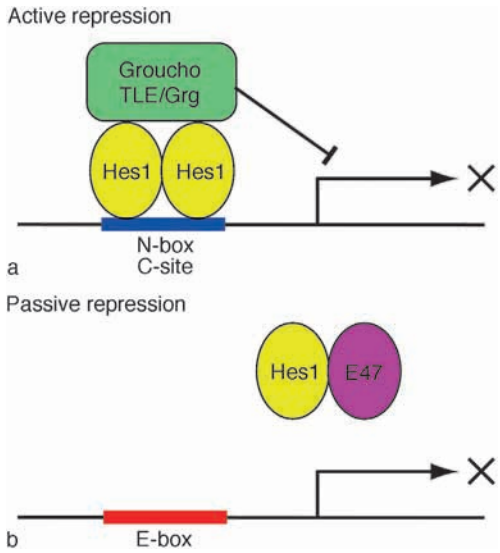

Figure 2 Two different mechanisms of Hes1-mediated transcriptional repression. (a) Active repression. Hes1 homodimer binds to the N-box and the class C site, and the co-repressors Groucho/TLE/Grg interact with the WRPW domain. This complex actively represses transcription by recruiting the histone deacetylase. (b) Passive repression. Hes1 forms a non-DNA-binding heterodimer with bHLH activators such as E47 and thereby represses transcription.

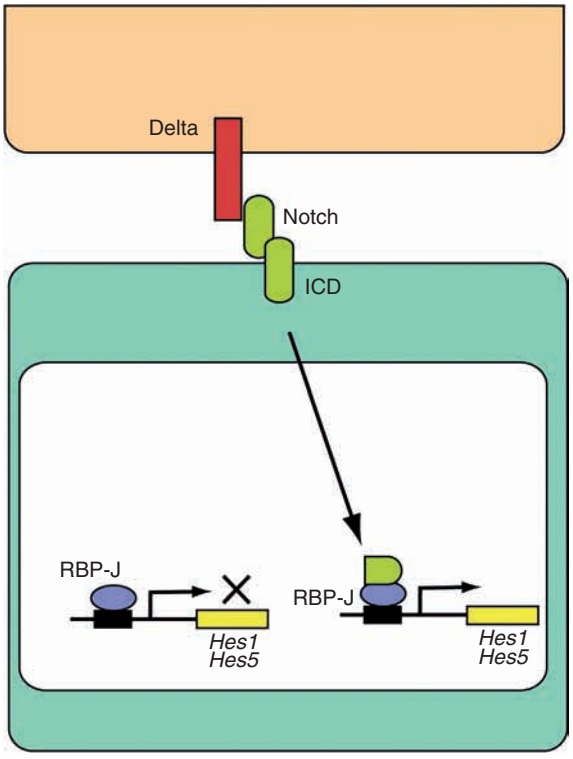

Figure 3 Notch signaling. The transmembrane protein Notch is activated by its ligands such as Delta. Upon activation, the intracellular domain (ICD) of Notch is cleaved and translocated into the nucleus. In the nucleus, the ICD forms a complex with the DNA-binding protein RBP-J, which binds to the *Hes1* and *Hes5* promoters. RBP-J alone is a transcriptional repressor, but the ICD–RBP-J complex is a transcriptional activator. Thus, Notch activation induces *Hes1* and *Hes5* expression.

by other factors. In agreement with this notion, Wnt, Shh, and BMP have been reported to upregulate *Hes1* expression. *Hey* gene expression is also induced by Notch signaling, indicating that Hes and Hey function together in Notch signaling.

Another important feature regulating *Hes* expression is that there are multiple N-box sequences in the *Hes1* promoter where Hes1 can bind and autorepress its own expression. Strikingly, when the promoter is activated, *Hes1* autonomously starts oscillatory expression with the periodicity of about 2 h (Figure 4). This oscillation depends on the negative feedback, as follows. *Hes1* promoter activation leads to accumulation of Hes1 protein, which represses its own transcription. This repression immediately leads to the disappearance of Hes1 mRNA and protein because they have extremely short half-lives (about 20 min). Disappearance of Hes1 protein then relieves the negative feedback, allowing the next round of transcription. In this way, *Hes1* displays oscillatory expression when the promoter is activated. Similarly, Hes7 expression also oscillates with the periodicity of about 2 h in the presomitic mesoderm. This oscillation is thought to be an integral component of the segmental clock that cycles every 2 h in the mouse, thus allowing another bilateral pair of somites to bud off at the anterior end of the presomitic mesoderm. Indeed, in the absence of Hes7, cycling of the clock is disrupted and somites are severely fused. It is likely that *Hes1* also acts as a 2-h-cycle biological clock in many cell types including neural progenitors. Supporting this idea, immunohistochemical analysis shows that Hes1 protein levels are variable in neural progenitors (see the following section), although further studies are required to determine whether or not *Hes1* expression oscillates in these cells *in vivo*.

Hes Expression in the Developing Nervous System

Differentiated neural cell types are generated from the neuroepithelial cells that line the ventricles of the developing brain and spinal cord. When they first appear in development, neuroepithelial cells initially undergo self-renewal by symmetric cell division (Figure 5). As development proceeds in the forebrain and midbrain, the neuroepithelial cells become radial glial cells, with cell bodies located to the ventricular zone and with long processes (radial fibers) reaching the pial surface (Figure 5). Radial glia have long been thought of as specialized glia (this is why they are named as such), but it has now been revealed that they are the second form of embryonic neural progenitors. Radial glial cells undergo asymmetric cell division, by which each radial glial cell produces one radial glial cell and one neuron (or neuronal precursor). Some radial glial cells undergo symmetric cell division, giving rise to two radial glial cells or two neurons. Neurons migrate along the radial fibers of radial glial cells toward the outer layers. Radial glial cells also give rise to ependymal cells, the internal

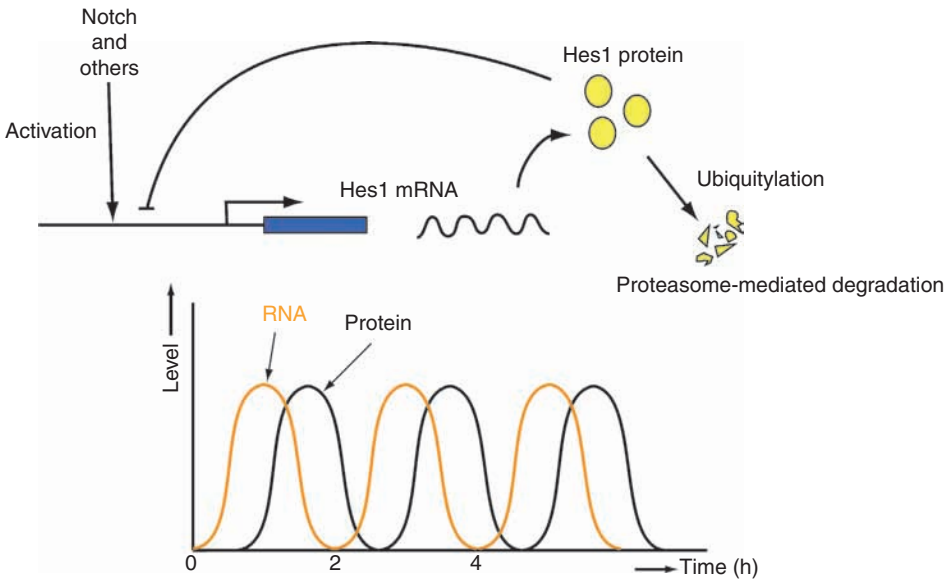

Figure 4 Oscillatory expression of Hes1. *Hes1* promoter activation leads to accumulation of Hes1 protein, which represses its own transcription. This repression immediately leads to disappearance of Hes1 mRNA and protein because they have extremely short half lives (about 20 min). Disappearance of Hes1 protein then relieves the negative feedback, allowing the next round of transcription. In this way, *Hes1* displays oscillatory expression when the promoter is activated.

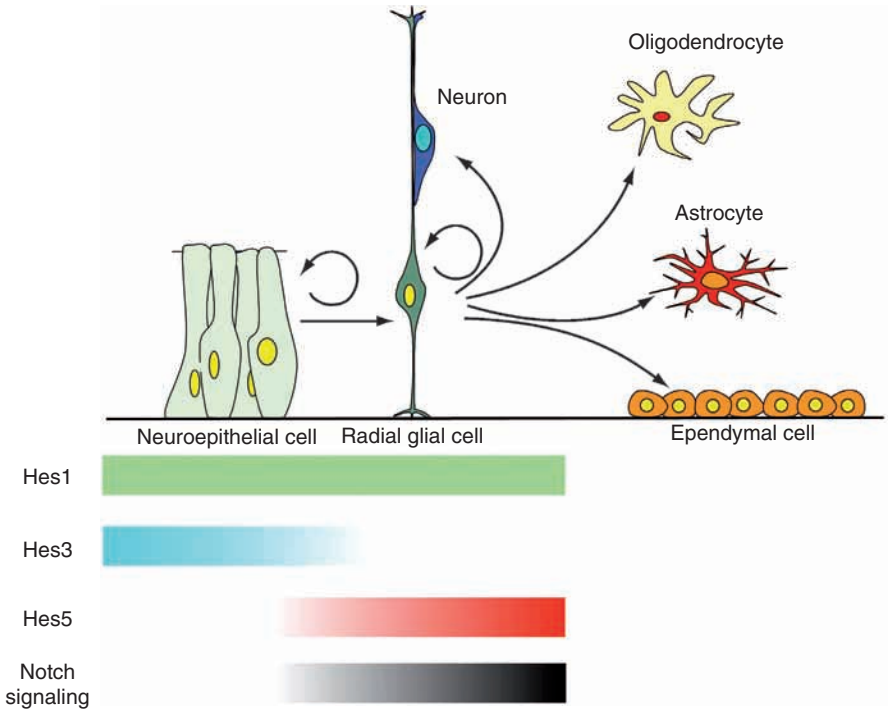

Figure 5 Expression patterns of *Hes* genes. In the developing nervous system, neuroepithelial cells are initially formed. Later, they become radial glial cells, which give rise to neurons, ependymal cells, oligodendrocytes, and astrocytes. *Hes1* and *Hes3* are expressed by neuroepithelial cells. *Hes1* is also expressed by radial glial cells while there is a switch from *Hes3* expression to *Hes5* expression during transition from neuroepithelial cells to radial glial cells. Notch signaling is also activated during this transition.

lining of ventricles (Figure 5). After production of neurons, radial glial cells are differentiated into glial cells (oligodendrocytes and astrocytes) (Figure 5). After birth, radial glial cells mostly disappear, but neural progenitors remain in a few neurogenic regions including the subventricular zone of the forebrain and the subgranular zone of the hippocampus. Thus, neural progenitors change their morphology and competence over time.

Initially, *Hes1* and *Hes3* are widely expressed by neuroepithelial cells (Figure 5). This expression occurs before *Notch1* and its ligands *Deltalike* (*Dll*) *1* and *Dll3* are expressed, indicating that *Hes1* and *Hes3* expression is independent of Notch signaling. During the transition from neuroepithelial cells to radial glial cells, *Hes3* expression is gradually downregulated and finally lost while *Hes5* expression starts together with *Notch1* and *Dll1* expression (Figure 5), suggesting that *Hes5* expression is controlled by Notch signaling. In contrast, Hes1 expression is maintained by radial glial cells (Figure 5) but becomes mostly complementary to *Hes5* expression (Figure 6). *Hes1* expression gradually overlaps with *Notch1* and *Dll1* expression, suggesting that *Hes1* expression is also controlled by Notch signaling at later stages. Interestingly, in the absence of *Hes1*,

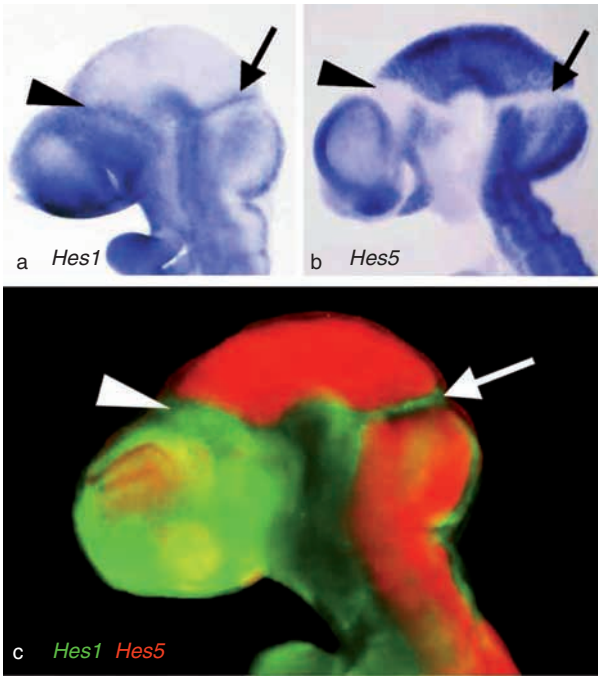

Figure 6 *Hes1* and *Hes5* expression in the developing brain. *Hes1* and *Hes5* expression are mostly complementary to each other at E9.5 in mouse embryos. *Hes1* is highly expressed in the zona limitans intrathalamica (arrowhead) and the isthmus (arrow) while *Hes5* is not.

Hes5 is expressed in the regions that normally express *Hes1* only, and vice versa, suggesting that *Hes1* and *Hes5* compensate each other. As development proceeds, *Hes1* and *Hes5* expression is downregulated but is retained in both the slowly dividing, niche localized neural stem cells, and the rapidly amplifying progenitors in the subventricular zone of the lateral ventricles, and the subgranular zone of the dentate gyrus. Thus, *Hes* genes are continuously expressed by neural progenitors. *Hes* genes are also expressed by subsets of glial cells such as Müller glia in the retina.

Hes Genes Regulate Maintenance of Neural Stem Cells

Many aspects of Hes function in the developing nervous system are explained by their ability to inhibit neurogenesis, thereby preventing neuronal differentiation of progenitors at earlier stages and promoting gliogenesis at later stages. Interestingly, in the absence of proneural bHLH genes such as *Mash1*, *Neurogenin* (*Ngn*), and *Math3*, progenitor cells that should normally become neurons are inhibited from neuronal differentiation and remain undifferentiated for a while but then adopt the glial fate prematurely. Thus, it is likely that proneural bHLH genes not only promote neuronal fate determination but also inhibit premature gliogenesis. In agreement with this notion, the proneural bHLH factor Ngn1 activates neuronal-specific gene expression and inhibits glial-specific gene expression by sequestering a p300/CBP–Smad complex. The phenotypes of proneural bHLH gene inactivation, inhibition of neurogenesis, and promotion of gliogenesis, are very similar to those of misexpression of *Hes* genes, suggesting that *Hes* genes exert their phenotypes by repressing proneural bHLH genes. Proneural bHLH genes are also required for *Hes* expression and maintenance of neural progenitors. These genes induce Delta expression, thereby activating Notch signaling in neighboring cells. In the absence of proneural bHLH genes, Notch is not activated, and *Hes* expression is downregulated, resulting in loss of neural progenitors. Thus, *Hes* and proneural bHLH genes form a network to regulate maintenance of neural progenitors and the formation of neurons and glial cells.

Hes1, Hes3, and Hes5 likely act to inhibit neurogenesis by repressing proneural gene expression directly. For example, the proneural bHLH gene *Mash1* contains class C binding sites in its promoter, and Hes1 represses *Mash1* expression by directly binding to these sites. In agreement with this result, *Mash1* expression is upregulated in the absence of *Hes1* and is even more severely upregulated in the absence of *Hes1*, *Hes3*, and *Hes5*. Furthermore, other proneural genes such as *Ngn* and *Math1* are also highly upregulated in the absence of these *Hes* genes, suggesting that Hes1, Hes3, and Hes5 negatively regulate proneural bHLH gene expression. Hes proteins might also inhibit neurogenesis through other targets, including the cyclin-dependent kinase (CDK) inhibitors p21 and p27, both of which inhibit cell cycle progression by retarding G1 phase. It is likely that Hes1, Hes3, and Hes5 promote cell cycle progression, and thus the maintenance of neural progenitors, by repressing p21 and p27. However, paradoxically, it has been also shown that persistent and high levels of Hes1 expression inhibit cell proliferation (see following section). Because p21 and p27 are also required for assembly of cyclin D1–CDK4 complex and nuclear import of cyclin D, complete repression of these CDK inhibitors by persistent and high levels of Hes1 may be inhibitory on cell cycle progression.

Based on their expression patterns and mutant phenotypes, the Hes1, Hes3, and Hes5 proteins subserve different although overlapping functions within the developing neuroepithelium. In the absence of *Hes1* and *Hes5*, radial glial cells prematurely differentiate into neurons and are depleted without generating later-born cell types. Premature depletion of radial glial cells leads to intermingling of neurons from the left and right walls of the neural tube and disruption of the neural tube structures (**Figures 7(f)–7(h)**, compare with **Figures 7(c)–(e)**), because radial glial cells have the junctional complex and the basal lamina at the apical and basal side, respectively, both of which serve as barriers to prevent neurons from spilling over. Conversely, transient misexpression of *Hes1* or *Hes5* in mouse brain at embryonic day 13.5 inhibits neuronal differentiation and maintains radial glial cells, leading to the increase of later-born neurons and glial cells. In contrast to the peripheral nervous system, where transient Notch activation instructs a glial fate (Schwann cells), transient Notch activation or transient misexpression of *Hes1* or *Hes5* in the central nervous system (CNS) does not seem to instruct gliogenesis. Instead, Hes1 and Hes5 seem to be required to maintain undifferentiated neural progenitors until later stages, when neurogenesis ceases and gliogenesis begins. Thus, *Hes1* and *Hes5* maintain neural progenitors without affecting their temporally changing competency in the CNS. *Hey1* and *Hey2* also have a similar activity: transient misexpression of *Hey1* or *Hey2* does not instruct gliogenesis but promotes maintenance of neural progenitors.

In the absence of *Hes1*, *Hes3*, and *Hes5*, not only radial glia but also neuroepithelial cells are not properly maintained and prematurely differentiated into neurons. However, even in *Hes1*:*Hes3*:*Hes5*

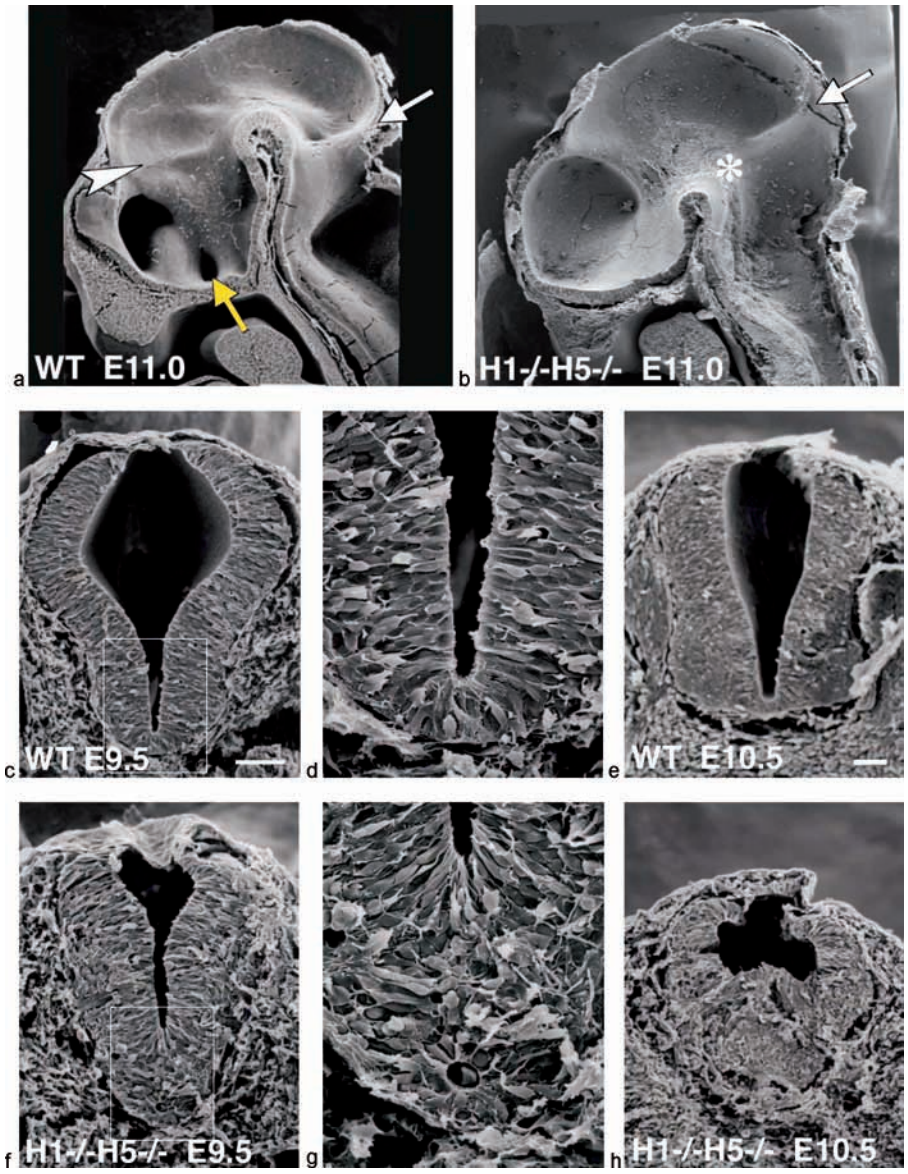

Figure 7 Structural defects of *Hes1*;*Hes5*-double-null mice. Mouse embryos were analyzed by scanning electron microscopy. (a) The optic vesicle (yellow arrow), the zona limitans intrathalamica (zli) (arrowhead), and the isthmus (white arrow) are formed in the wild type. (b) In the absence of *Hes1* and *Hes5*, the optic vesicle and the zli are not formed. The isthmus is only partially formed (asterisk and arrow). (c–e) Cells are aligned radially in the walls of the neural tube in the wild type. (f–h) Cells from the left and right walls are intermingled with each other at E9.5 in the ventral part of *Hes1*;*Hes5*-double-null embryo (f, g). This defect becomes severer at E10.5 (h). Boxed regions in (c) and (f) are enlarged in (d) and (g), respectively. Scale bar = 100 μm.

triple knockout mice, neuroepithelial cells are initially formed, indicating that *Hes* genes are required for maintenance of, but not formation of, neuroepithelial cells. It remains to be determined which factors are responsible for initial formation of neuroepithelial cells. Based on these phenotypes of *Hes*-mutant mice and expression patterns of *Notch* and *Hes* genes, the neuroepithelium develops stepwise in the following way: (1) an initial *Hes*- and *Notch*-independent phase, (2) a subsequent *Hes*-dependent but *Notch*-independent phase, and (3) a *Hes*- and *Notch*-dependent radial glial phase (Figure 8). During the second phase, the neuroepithelial cells go through a transitory stage, in which they usually do not give rise to any neurons, but without *Hes* genes, prematurely differentiate into neurons.

Based on the expression patterns, it is likely that *Hes1* and *Hes3* are mainly responsible for transitory neuroepithelial cells, while *Hes1* and *Hes5* are required for radial glial cells. *Hes1* and *Hes5* expression in radial glial cells may be controlled by Notch

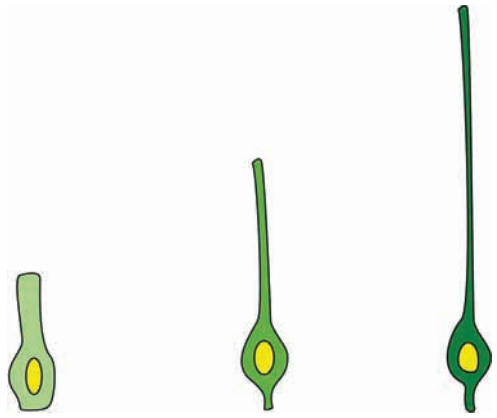

	Neuroepithelial cell	Transitory neuroepithelial cell	Radial glial cell
Hes dependency	–	+	+
Delta/*Notch*	–	–	+
Neurogenesis	–	–	+

Figure 8 Temporal changes of neural progenitor characteristics. Embryonic neural progenitors change their characteristics as follows: (a) the initial *Hes*- and *Notch*-independent neuroepithelial cells, (b) the subsequent *Hes*-dependent but *Notch*-independent transitory neuroepithelial cells, and (c) *Hes*- and *Notch*-dependent radial glial cells. *Hes*-dependent but *Notch*-independent transitory neuroepithelial cells prematurely differentiate into neurons in the absence of *Hes* genes.

signaling (Figure 5), but it remains to be determined which factors regulate *Hes1* and *Hes3* expression in transitory neuroepithelial cells. While transitory neuroepithelial cells usually undergo symmetric cell division and do not produce any neurons, radial glial cells give rise to neurons by asymmetric cell division. Interestingly, during the switch from transitory neuroepithelial cells to radial glial cells, expression of both *Notch* and its ligands begins together, suggesting that neural stem cells undergoing asymmetric cell division express Notch signaling molecules whereas those undergoing symmetric cell division do not. Thus, it is likely that expression of Notch signaling molecules is a key feature of the switch from symmetric to asymmetric cell division.

Eye formation also depends on *Hes1* and *Hes5*. In the absence of *Hes1*, neuronal differentiation is accelerated and retinal neural stem cells are prematurely depleted, resulting in disruption of the neural retina structures. Strikingly, in the absence of *Hes1* and *Hes5*, the optic vesicles are not formed (Figure 7(b), compare with Figure 7(a), yellow arrow), although the homeodomain gene *Rax*, which is essential for eye formation, is expressed normally in the prospective eye regions. Normally, no neurogenesis occurs during the optic vesicle formation, but without *Hes1* and *Hes5*, neurogenesis is severely accelerated in the prospective eye regions. Thus, *Hes1* and *Hes5* are required for maintenance of retinal progenitors from the earliest stage of eye development and seem to function together with *Rax*.

Hes Genes Regulate Boundary Formation

The CNS is partitioned into many compartments by boundaries (Figure 9), which are formed by specialized neuroepithelial cells or radial glial cells. Boundaries act as organizing centers which function by producing morphogens, thereby patterning nearby compartments and specifying regional aspects of neural differentiation. For example, the isthmus, a midbrain–hindbrain boundary, expresses Fgf8 and regulates neuronal specification in the midbrain and the hindbrain while the zona limitans intrathalamica (zli), a boundary between the dorsal and ventral thalamus, expresses Shh and regulates specification of dorsal and ventral thalamic neurons. Boundary cells restrict cell migration and do not proliferate as efficiently as neuroepithelial and radial glial cells of compartments. In addition, neurogenesis usually does not occur in boundaries. *Hes1* is expressed at high levels in most boundaries, while *Hes3* is expressed in the isthmus. *Hes5* is not expressed in boundaries, but it is ectopically expressed in some boundaries in the absence of *Hes1*.

Hes genes are not only required in neuropithelial cells within the compartments but also for the formation of neuroepithelial cells that comprise the boundaries. In the absence of *Hes1* and *Hes3*, the isthmus is not maintained and differentiated into neurons. As a result, Fgf8 expression is prematurely lost, resulting in defects of midbrain- and hindbrain-specific neuronal differentiation. Similar defects occur in the

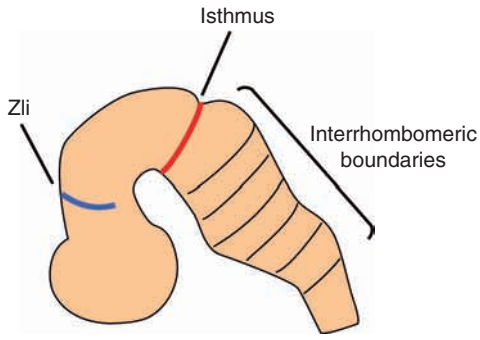

Figure 9 Boundaries in the developing CNS. The developing CNS is partitioned into compartments by boundaries, which function as organizing centers. Hes1 is highly expressed in boundaries and, without *Hes* genes, boundaries are prematurely lost or not formed. zli, zona limitans intrathalamica.

isthmus of *Hes1*;*Hes5*-double-null embryos (Figure 7(b), arrow and asterisk, compare with Figure 7(a), white arrow). In addition, in the absence of *Hes1* and *Hes5*, the zli is not formed (Figure 7(b), compare with Figure 7(a), arrowhead) and Shh expression does not occur. In the absence of *Hes1*, *Hes3*, and *Hes5*, ectopic neurogenesis and loss of the organizer activities occur in all boundaries, resulting in defects of regional-specific neuronal differentiation. *Hes1* is persistently expressed at high levels by many boundary cells (Figures 6(a) and 6(c), arrow and arrowhead). In contrast, *Hes1* expression is variable in compartments: some cells express Hes1 protein at high levels while others express at lower levels or do not express Hes1 protein. When Hes1 expression is reduced, proneural bHLH factor expression is upregulated, displaying an inverse correlation between Hes1 and proneural bHLH factor expression (Figure 10(a), upper panel). As stated above, Hes1 expression could oscillate in compartmental progenitors (Figure 10(a), lower panel). In boundaries, it is likely that persistent and high levels of Hes1 expression constitutively repress proneural bHLH gene expression and thereby block neurogenesis (Figure 10(b)). In agreement with this notion, inactivation of *Hes1*, *Hes3*, and *Hes5* leads to ectopic expression of proneural bHLH genes and premature loss of organizing activities in boundaries. Furthermore, when *Hes1* is misexpressed at persistent and high levels in compartmental neural progenitors, these cells encounter blockade of neurogenesis and reduction of proliferation rates, two important features of boundary cells. Thus, expression mode of Hes1 is different between compartments (variable) and boundaries (persistent) (Figure 10), and it is likely that this different expression mode regulates compartment versus boundary cell characteristics. Interestingly, both Hes1 mRNA and protein are highly expressed in boundaries, suggesting that the negative feedback does not work in

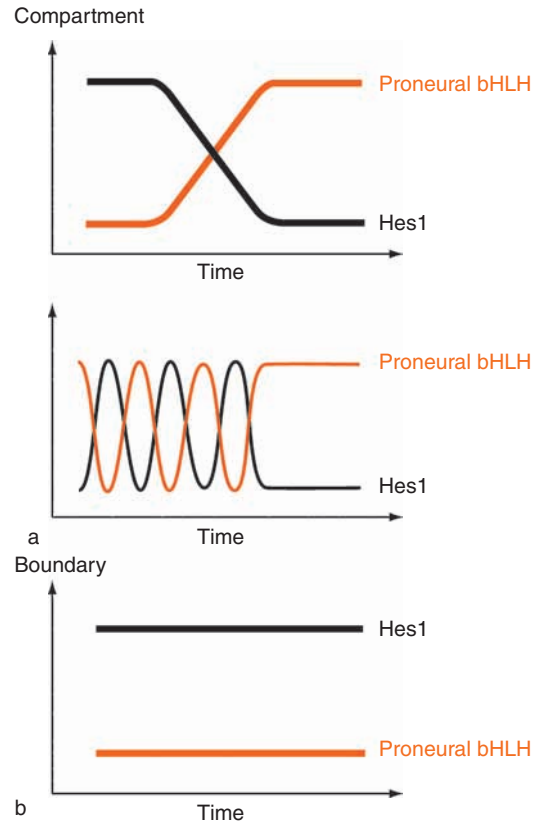

Figure 10 Different Hes1 expression modes in compartments and boundaries. (a) In compartments, Hes1 expression is variable: either gradual downregulation or oscillatory expression. Downregulation of Hes1 expression leads to upregulation of proneural bHLH gene expression. (b) In boundaries, persistent and high levels of Hes1 constitutively repress proneural bHLH gene expression, thereby maintaining neuron-free zones and organizing centers. It is likely that dynamic versus persistent expression of Hes1 regulates compartment versus boundary formation.

boundaries. Further studies are required to demonstrate how this unique expression mode is regulated.

Summary

In sum, *Hes* genes act to inhibit neurogenesis by repressing proneural bHLH genes but do so in different ways along the neuraxes. In the neurogenic compartments, *Hes* gene expression tends to be dynamic, and a loss of *Hes* gene function results in accelerated neuronal differentiation and a premature loss of neural progenitors at the expense of later-born cells. Thus, *Hes* genes are required in neurogenic regions to ensure the generation of differentiated cells in both the correct numbers and full diversity. In regions of the neuroepithelium that form boundaries, *Hes* gene expression persists at high levels and thereby constitutively represses proneural bHLH gene expression. In the absence of *Hes* genes, ectopic neurogenesis occurs in boundaries and their organizing activities are

lost. These different aspects of *Hes* gene function are likely to depend on the different modes by which their expression is regulated (i.e., dynamic versus persistent).

See also: Helix–Loop–Helix (bHLH) Proteins: Proneural; Notch Signal Transduction: Molecular and Cellular Mechanisms; Notch Pathway: Lateral Inhibition; Radial Glial Cells: Brain Functions; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Baek JH, Hatakeyama J, Sakamoto S, Ohtsuka T, and Kageyama R (2006) Persistent and high levels of Hes1 expression regulate boundary formation in the developing central nervous system. *Development* 133: 2467–2476.

Bertrand N, Castro DS, and Guillemot F (2002) Proneural genes and the specification of neural cell types. *Nature Reviews Neuroscience* 3: 517–530.

Davis RL and Turner DL (2001) Vertebrate hairy and enhancer of split related proteins: Transcriptional repressors regulating cellular differentiation and embryonic patterning. *Oncogene* 20: 8342–8357.

Fisher A and Caudy M (1998) The function of hairy-related bHLH repressor proteins in cell fate decision. *Bioessays* 20: 298–306.

Hatakeyama J, Bessho Y, Katoh K, et al. (2004) *Hes* genes regulate size, shape and histogenesis of the nervous system by control of the timing of neural stem cell differentiation. *Development* 131: 5539–5550.

Hirata H, Tomita K, Bessho Y, and Kageyama R (2001) *Hes1* and *Hes3* regulate maintenance of the isthmic organizer and development of the mid/hindbrain. *EMBO Journal* 20: 4454–4466.

Hirata H, Yoshiura S, Ohtsuka T, et al. (2002) Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science* 298: 840–843.

Ishibashi M, Ang S-L, Shiota K, Nakanishi S, Kageyama R, and Guillemot F (1995) Targeted disruption of mammalian *hairy* and *Enhancer of split* homolog-1 (*HES-1*) leads to up-regulation of neural helix-loop-helix factors, premature neurogenesis, and severe neural tube defects. *Genes and Development* 9: 3136–3148.

Iso T, Kedes L, and Hamamori Y (2003) HES and HERP families: Multiple effectors of the notch signaling pathway. *Journal of Cell Physiology* 194: 237–255.

Jarriault S, Brou C, Logeat F, Schroeter EH, Kopan R, and Israël A (1995) Signalling downstream of activated mammalian notch. *Nature* 377: 355–358.

Kageyama R, Hatakeyama J, and Ohtsuka T (2006) Roles of Hes bHLH factors in neural development. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 3–22. Weinheim: Willey-VCH.

Kageyama R, Ohtsuka T, Hatakeyama J, and Ohsawa R (2005) Roles of bHLH genes in neural stem cell differentiation. *Experimental Cell Research* 306: 343–348.

Ohtsuka T, Ishibashi M, Gradwohl G, Nakanishi S, Guillemot F, and Kageyama R (1999) *Hes1* and *Hes5* as notch effectors in mammalian neuronal differentiation. *EMBO Journal* 18: 2196–2207.

Ross SE, Greenberg ME, and Stiles CD (2003) Basic helix-loop-helix factors in cortical development. *Neuron* 39: 13–25.

Sasai Y, Kageyama R, Tagawa Y, Shigemoto R, and Nakanishi S (1992) Two mammalian helix-loop-helix factors structurally related to Drosophila *hairy* and *Enhancer of split*. *Genes and Development* 6: 2620–2634.

Helix–Loop–Helix (bHLH) Proteins: Proneural

J E Johnson and S M Glasgow, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

bHLH Transcription Factor Family

The basic helix–loop–helix (bHLH) family of proteins comprises transcriptional regulators of a variety of developmental processes including cellular differentiation and lineage commitment, not only in the nervous system but also in the pancreas, muscle, and heart. Data collected from several model organisms have implicated bHLH transcription factors in the determination of neuronal lineages and the specification of distinct cell fates. *Drosophila* studies have underscored the importance of bHLH factors in the acquisition of neuronal identity in the ectoderm. In this context, bHLH proteins are both necessary and sufficient to confer a neuronal fate to ectodermal cells; consequently, several bHLH factors, including achaete–scute and atonal, have been dubbed 'proneural genes,' while inhibitory bHLH factors such as hairy and enhancer-of-split suppress neurogenesis. Vertebrate homologs of the *Drosophila* proneural genes have been isolated and neural expression and function appears to be conserved.

The bHLH proteins mediate their effects by protein dimerization and DNA recognition. The binding to a core hexa-nucleotide e-box, CANNTG, is mediated by the ~60-amino-acid bHLH motif. The crystal structures of the bHLH proteins E47 (Tcfe2a, MGI) and Myod1 bound to DNA demonstrated that the bHLH region is composed of two α-helices separated by a loop. The basic region contacts DNA in the major groove, while the HLH domain is involved in protein dimerization. The bHLH proteins are classified by three criteria: (1) dimerization capabilities, (2) DNA binding specificities, and (3) tissue distribution. Two broadly defined groups are class A (class I) and class B (classes II–VII). Class A proteins are broadly expressed and are often called E-proteins. These E-proteins, including HEB (Tcf12, MGI), E2-2 (Tcf4, MGI), and the two splice variants of E2a (Tcfe2a, MGI), E12 and E47, in vertebrates, and Daughterless in *Drosophila*, form homodimers, as well as heterodimers with class B proteins. Class B bHLH proteins are further categorized into six subclasses termed classes II–VII based upon protein motifs and activity.

Different members of these subclasses of bHLH factors act to either inhibit or promote transcription of downstream targets. Particularly important for neural differentiation and cell-type specification is the neural expressed subset of class II bHLH factors that include Mash1 (Ascl1, MGI), Math1 (Atoh1, MGI), Math5 (Atoh7, MGI), Ngn1,2,3 (Neurog1,2,3, MGI), and Neurod1. These factors form heterodimers with E-proteins, bind to DNA at e-box elements, and act as positive regulators of neurogenesis and neuronal specification. Table 1 lists some class II bHLH factors chosen for their expression in progenitor zones in the neural tube. In each case, loss-of-function studies have determined their importance in the formation of discrete neural cell types. In addition, class VI factors that include Hes proteins (homologs to *Drosophila* hairy and enhancer-of-split) also play essential functions in these processes but they act to suppress neurogenesis. The class II and class VI bHLH factors are in a regulatory loop that involves cell–cell signaling via the Notch–Delta signaling pathway and serves to limit the number of cells differentiating from the progenitor domain (Figure 1). In one cell, the Class II proneural bHLH increases levels of Notch ligand Delta. Notch signaling is activated in the neighboring cell resulting in an increase in Hes levels. Hes proteins suppress levels of the proneural bHLH factor. The cell with high levels of the proneural bHLH will undergo neuronal differentiation, whereas the cell receiving Notch signaling is more likely to remain as a progenitor, or to later become an astrocyte. This pathway was worked out originally using *Drosophila* genetics, but many components and mechanisms appear to be conserved in vertebrates as well.

Neuronal Differentiation

Several lines of evidence have determined roles for bHLH factors in neural development, particularly those of the achaete–scute and atonal classes. Members of these subclasses of bHLH factors are transiently expressed in restricted progenitor domains throughout the developing nervous system. Loss-of-function studies have demonstrated their absolute requirement for the formation of specific subsets of neurons, while overexpression assays have been particularly important in revealing their function in neural differentiation. In *Drosophila*, the achaete–scute complex genes (AS-C) are required for external sensory organs, and atonal is required for chordotonal organs and photoreceptor cells. In the absence of these bHLH factors, neuroblasts fail to form. Conversely, overexpression of achaete, scute, and atonal led to neuronal hyperplasia.

Table 1 Vertebrate neural bHLH factors and neural tissues where they function

bHLH	*Population/region*	*References*
Math1 (Atoh1)	Cerebellar granule cells	Ben-Arie (1997) *Nature* 390: 169.
	Gut epithelium	Yang (2001) *Science* 294: 2155.
	Hair cells of the inner ear	Bermingham (1999) *Science* 284: 1837.
	Merkel cells of the skin	Ben-Arie (2000) *Development* 127: 1039.
	Pontine nuclei	Ben-Arie (2000) *Development* 127: 1039.
	Spinal cord dI1 interneurons	Gowan (2001) *Neuron* 31: 219.
Math3	Retinal horizontal cells	Akagi (2004) *Journal of Biological Chemistry* 279: 28492.
(Neurod4)	Retinal amacrine cells	Inoue (2002) *Development* 129: 831.
	Retinal bipolar cells	Hatakeyama (2001) *Development* 128: 1313.
Math5 (Atoh7)	Retinal ganglion cells	Wang (2001) *Genes Development* 15: 24.
		Kanekar (1997) *Neuron* 19: 981.
Mash1 (Ascl1)	Adrenal medulla-chromaffin cells	Huber (2002) *Development* 129: 4729.
	Gabaergic neurons cortex	Casarosa (1999) *Development* 126: 525.
		Fode (2000) *Genes Development* 14: 67.
	Enteric serotonergic neurons	Blaugrund (1996) *Development* 122: 309.
	Hindbrain serotonergic neurons	Pattyn (2004) *Nature Neuroscience* 7: 589.
	Noradrenergic neurons of brainstem	Hirsch (1998) *Development* 125: 599.
		Lo (1998) *Development* 125: 609.
	Noradrenergic neurons of PNS	Guillemot (1993) *Cell* 75: 463.
	Paracardiac ganglia	Guillemot (1993) *Cell* 75: 463.
	Primary olfactory neurons	Guillemot (1993) *Cell* 75: 463.
	Retinal neurons	Tomita (1996) *Genes to Cell* 1: 765.
	Spinal cord dI3 and dI5 interneurons	Helms (2005) *Development* 132: 2709.
	Spinal cord V2 interneurons	Li (2005) *Proceedings of the National Academy of Sciences of the United States of America* 102: 10688.
Ngn1/2	Distal cranial sensory ganglia (Ngn2)	Fode, *Neuron* 20: 483
(Neurog1/2)	Proximal cranial sensory ganglia (Ngn1) (1998).	Ma (1998) *Neuron* 20: 469.
	Glutamatergic neurons of cortex	Fode (2000) *Genes Development* 14: 67.
	Dorsal root sensory ganglia	Ma (1999) *Genes Development* 13: 1717.
	Motor neurons (with Olig2)	Mizuguchi (2001) *Neuron* 31: 757.
		Scardigli (2001) *Neuron* 31: 203.
	Spinal cord dI2 interneurons	Gowan (2001) *Neuron* 31: 219.
	Spinal cord ventral interneurons (V1–3)	Scardigli (2001) *Neuron* 31: 203.
	Ventral midbrain dopaminergic neurons	Kele (2006) *Development* 133: 495.
Ngn3 (Neurog3)	Enteroendocrine cells of stomach	Lee (2002) *Genes Development* 16: 1488.
	Pancreatic Islet of Langerhans	Gradwohl (2000) *Proceedings of the National Academy of Sciences of the United States of America* 97: 1607.
	Ventral spinal cord glial cell differentiation	Lee (2003) *Developmental Biology* 253: 84.
Olig1/2	Motor neurons	Novitch (2001) *Neuron* 31: 773.
		Mizuguchi (2001) *Neuron* 31: 757.
	Oligodendrocytes of spinal cord and brain	Zhou (2001) *Neuron* 31: 791.
		Zhou (2002) *Cell* 109: 61.
Olig3	Spinal cord dI1–3 interneurons	Müller (2005) *Genes Development* 19: 733.
Ptf1a	GABAergic neurons of the cerebellum	Hoshino (2005) *Neuron* 47: 201.
	Spinal cord GABAergic dI4 and dIL^A interneurons	Glasgow (2005) *Development* 132: 5361.
	Pancreatic exocrine cells	Krapp (1998) *Genes Development* 12: 3752.

A role for vertebrate neural bHLH factors in inducing neuronal differentiation has been repeatedly demonstrated in multiple paradigms across a variety of species. Overexpression of Mash1, Ngn1, Neurod1, Neurod2, and Math1 conferred neuronal properties to the mouse embryonal carcinoma cell line, P19, by inducing cell cycle exit and expression of neuronal specific genes. In *Xenopus*, injection of mRNA for XNgnr1, Xneurod, Xath1, and Xash1/3 increased primary neurogenesis to varying extents at the expense of ectodermal cells. Forced expression of Ngn1 in primary rat cortical cultures induced neurogenesis while suppressing gliogenesis. Additionally, forced expression *in ovo* of Mash1, Math1, Ngn1, and Ngn2 in chick neural tube has clearly demonstrated the activity of these factors in driving neuronal differentiation. In these experiments, the proliferating progenitors in the ventricular zone were forced out of the cell cycle, moved laterally out of the ventricular zone into the marginal zone, and began expressing neuronal specific genes. In some cases, the interpretation of loss-of-function mutations

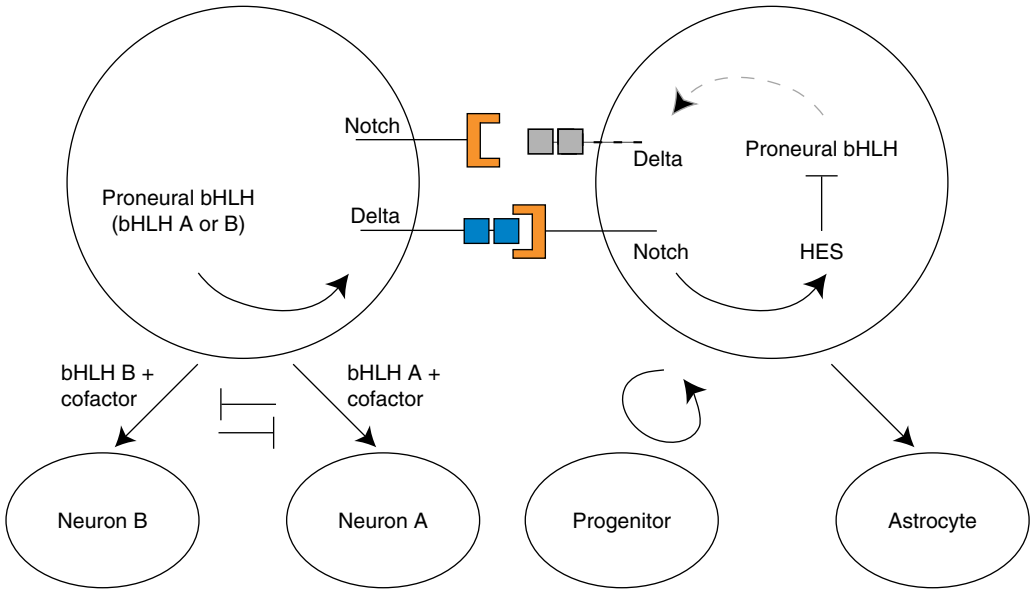

Figure 1 Proneural bHLH transcription factors and the Notch–Delta signaling pathway interact to control neuronal differentiation. In a cell with high levels of a proneural bHLH protein, expression of Delta ligand is enhanced. This in turn signals the neighboring cell, through the Notch receptor, to turn on the Hes bHLH transcription factors that repress the proneural bHLH proteins in that cell. The cell with high levels of Notch signaling and low levels of the proneural bHLH will remain as a progenitor cell or ultimately differentiate to an astrocyte. The cell with low levels of Notch signaling and high levels of the proneural bHLH will undergo neuronal differentiation. The type of neuron that is generated depends on the combination of the proneural subtype plus additional transcription factors.

in mouse are more difficult to interpret in regards to neuronal differentiation, since in the absence of one neural bHLH factor, neighboring or co-expressed bHLH factors can compensate. While compensation is a major issue, there are many examples where loss of one bHLH factor has dramatic consequences for neurons in a specific region (see Table 1).

Neurons versus Glia

The bHLH factors are not only involved in the differentiation of various neuronal lineages but they are also intrinsic mediators of the decision between neuronal and glial cell fates. As mentioned above, in cortical cultures, Ngn1 induces neurogenesis while suppressing astrocyte development. *In vivo*, neuronal and glial cell generation are temporally separated, where neurons are generated first followed by overlapping waves of oligodendrocytes and astrocytes. The loss of neural bHLH factors such as Mash1, Math3 (Neurod4, MGI), and Ngn1/2 results in an increase in astrocytes. This is seen in the cortex in the absence of Mash1 and Ngn2; likewise, loss of Mash1 and Math3 in the midbrain and hindbrain results in an increase in astrocytes. Single mutants do not clearly show the increase in astrocytes, probably due to redundant function of these bHLH factors in neurogenesis. Indeed, in mouse models where Ngn2 replaces Mash1 or vice versa, to a first approximation their neurogenesis activity is interchangeable. Studies in retina have also contributed to establishing a role for several bHLH factors in promoting neuronal fate and inhibiting glial fate. Overexpression of Math3 or Mash1 in retinal explant cultures increased neuron production while inhibiting the generation of Müller glial cells. Consistently, the absence of Mash1 results in an increase in the production of Müller glial cells. Together, these results demonstrate the requirement for class II neural bHLH factors to produce neurons; in their absence, glial cells form. As opposed to the function of the class II, class VI bHLH factors, particularly Hes1 and Hes5, suppress neuronal differentiation. In mutants for this class of bHLH proteins, premature neuronal differentiation is detected.

Oligodendrocytes are another major class of glia cells, and they also require bHLH proteins. Olig1 and Olig2 are required for oligodendrogenesis and maturation in brain and spinal cord. In the developing ventral spinal cord, Olig2 and Ngn2 are involved in the specification of oligodendrocytes and motor neurons. A precursor cell expressing Ngn2 and Olig2 will generate motor neurons. However, when Ngn2 levels are reduced, the Olig2 cells generate oligodendrocytes. Motor neurons and oligodendrocytes are not generated from a common Olig1/2-expressing progenitor; instead, these divergent cell types appear to be generated sequentially from neuroepithelial stem cells.

As stated previously, multiple lines of evidence place Mash1 as a neuronal differentiation factor.

However, there are increasing reports that Mash1 functions in oligodendrocyte formation as well. Mash1 is present in progenitors to neurons early and then oligodendrocytes later in the spinal cord. Mash1 is also present in the subventricular zone (SVZ) and rostral migratory stream (RMS) in the telencephalon in progenitors to neurons and oligodendrocytes. When cells from this region were cultured from Mash1 mutants, there was a loss of neurons and oligodendrocytes but not astrocytes relative to that seen with wild-type cultures. Together, these studies demonstrate the importance of some bHLH factors for generation of neurons, such as Ngn1 and Ngn2, and also reveal the importance of bHLH factors in oligodendrocyte development, such as Olig1, Olig2, and Mash1. In the absence of these bHLH factors, excess astrocytes are formed. This balance between neurons and oligodendrocytes on the one hand, and astrocytes on the other, involves the activation of Notch signaling (described above), a pathway known to be activated by the proneural bHLH factors (Figure 1). Through Notch signaling, the Hes bHLH factors are induced, which in turn suppress the proneural bHLH factors, and, thus, suppress neurogenesis.

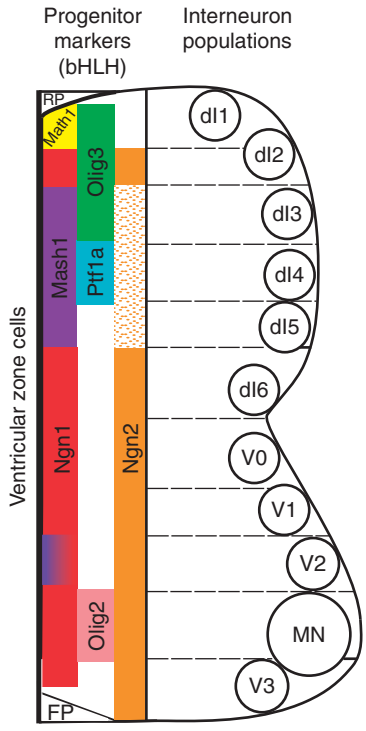

Figure 2 Class II bHLH factors are present in distinct populations of neuronal progenitor cells that are required for neuronal differentiation and cell-type specification. The pattern of expression of a subset of bHLH factors is shown in the ventricular zone of the spinal neural tube. Some progenitor cells express multiple bHLH factors. The dorsal interneurons (dI1–dI6), ventral interneurons (V0–V3), and the motor neurons (MN) require a distinct combination of bHLH factors to develop. RP, roof plate; FP, floor plate.

Neuronal Subtype Specification

Once a cell has committed to the neuronal lineage, it is specified to a certain neuronal cell type. Neural bHLH proteins in some regions of the nervous system have been shown to play roles in this aspect of nervous system development as well (see Table 1). In many regions of the developing embryo, the expression domains of the bHLH factors and the neuronal populations they specify are tightly linked. Expression of Mash1, Math1, and Ngn1 throughout the neural tube is nonoverlapping. Loss-of-function mutations in each of these bHLH factors have established their requirement for formation of specific neuronal populations. For example, in the dorsal spinal cord where interneuron populations are defined by the combination of homeodomain transcription factors they express, Math1 is required for dI1 neurons, Ngn1 is required for dI2 neurons, Mash1 is required for dI3 and dI5 neurons, and Ptf1a is required for dI4 neurons (Figure 2). Furthermore, ectopic expression of Math1, Ngn1, and Mash1 in ventricular zone cells in the chick neural tube results in an increase of a specific neuronal subtype at the expense of the others. This neuronal specification function of the bHLH factors is thought to work in combination with other factors. For example, when Math1 is ectopically expressed in the dorsal neural tube, it appears to convert dI2 and dI3 neurons to dI1, but when it is expressed ventrally it does not induce dI1 neurons, suggesting that a factor present dorsally but not ventrally is required for dI1 cell specification in combination with Math1. In addition, ectopic expression of Mash1 with another bHLH Olig3 increases dI3 neurons while Mash1 plus the homeodomain factor Lbx1 increases dI5. Thus, the generation of each interneuron population in the dorsal neural tube requires a distinct neural bHLH factor, but to specify the specific neuronal cell type, additional factors are required.

Neurotransmitter phenotype is a major facet of neuronal subtype that has also been shown to be specified, at least in part, by bHLH factors. GABAergic neurons in the dorsal horn of the spinal cord and in the cerebellum have an absolute requirement for the bHLH factor Ptf1a. However, the relationship between the requirement for a specific neural bHLH factor and a specific neurotransmitter phenotype is usually not this clear. For example, Mash1 is required for adrenergic neurons in the sympathetic nervous system and in the hindbrain, serotonergic neurons in the enteric nervous system and hindbrain, and GABAergic neurons in the cortex. Similarly, Ngn1

and Ngn2 are required for glutamatergic neurons in the cortex, sensory ganglia, and cholinergic neurons in the ventral spinal cord. Although the bHLH subtype does not appear linked to the neurotransmitter subtype when assessed in this global manner, regionally, neurons of specific neurotransmitter subtype require distinct bHLH factors. In the cortex, the GABAergic neurons migrate from ventral regions and are Mash1 derived, while glutamatergic neurons are generated dorsally and are Ngn1 and Ngn2 derived.

The requirement for factors in addition to the bHLH proteins for neuronal subtype specification, particularly the homeodomain class of transcription factors, is best illustrated from studies of retina development. In the retina, as in the brain and spinal cord, both gain- and loss-of-function studies have demonstrated the importance of the bHLH factors in both neuronal differentiation and in subtype specification. And although specific retinal cell types are lost or reduced when neural bHLH factors are absent, ectopic expression of these factors is not sufficient to specify the correct cell type. In addition, even with loss of function, often more than one bHLH must be mutated to lose a cell type. Retinal ganglia cells require Math5, amacrine cells require both NeuroD and Math3, and bipolar cells require Mash1 and Math3. The requirement for the concerted activity of bHLH and homeodomain factors for specification of these cell types was shown in overexpression assays. For example, to obtain bipolar cells in retinal explants requires the activity of the homeodomain containing factor Chx10 together with Math3 or Mash1. A similar situation arises in the specification of amacrine cells where the homeodomain factors Pax6 or Six3 are required with Math3 or Neurod1 to get specification to this cell type. Thus, although bHLH factors are important components of the code that determine the neuronal subtype, they appear to work in combination with homeodomain factors. The mechanism by which these two classes of transcription factors interact is largely unknown. Synergistic interaction between LIM homeodomain proteins and two bHLH factors, Ngn2 or NeuroM, that activate transcription of the Hb9 enhancer may play a role in motor neuron specification.

Control of bHLH Protein Expression

Controlling expression of the bHLH factors is critical for normal nervous system formation. The different neural bHLH factors can be classified into those expressed in mitotically active cells of the ventricular zone, to those expressed in newly postmitotic cells, to those that are maintained in mature neural cells in the adult. This article highlights the factors expressed and functioning at the early stages of neural differentiation (Table 1). In some cases, the later-expressed bHLH factors are downstream targets of the early factors, as is seen in Ngn1 and Ngn2 regulation of the postmitotic bHLH factor NeuroD. There are many additional mechanisms controlling bHLH expression including autoregulation, repression by Notch signaling (Figure 1), response to signaling pathways, and cross-repression between different bHLH factors. Autoregulation includes positive autoregulation seen with Math1 and Math5, and negative autoregulation seen with Mash1. The latter likely involves a regulatory loop with Notch/Delta signaling where Mash1 induces Delta, which in turn activates Notch signaling in the neighboring cell, resulting in Hes repression of Mash1 (see Figure 1). As discussed earlier, this may control the timing of expression of a bHLH factor that drives neuronal differentiation. Patterning signals influence the spatial expression pattern of a bHLH factor. For example, bone morphogenetic protein (BMP) signaling is required for expression of Math1 and Ngn1 in the dorsal neural tube, and Math1 in the cerebellum. And finally, cross-repression of expression was demonstrated in the dorsal neural tube where Math1 and Ngn1 suppressed expression of neighboring bHLH factors. This mechanism ensures strict boundaries of expression that may be important in generating the correct composition of neurons.

Concluding Remarks

Overwhelming evidence implicates bHLH transcription factors in the neuronal differentiation and cell-type specification in multiple regions of the developing nervous system. The bHLH family of transcription factors includes factors that are positive regulators of neurogenesis, negative regulators of neurogenesis, and positive regulators of oligodendrogenesis. Even with these different subclasses of bHLH factors, there are complexities that arise from formation of homodimers and heterodimers across these distinct subclasses of bHLH proteins. These complexes may bind DNA and activate or repress transcription, they may form inactive complexes that bind DNA but have no transcription activity, or they may not bind DNA at all. Other intricacies that modulate the function of the neural class II factors include the formation of higher-order complexes. For instance, the bHLH factor Ptf1a that is required for GABAergic neurons in the dorsal horn and cerebellum, is known to form a trimer complex in pancreas with an

E-protein (HEB, E12, or E2-2) and Rbpsuh, an effector of Notch signaling. Another example is the putative complex formed between bHLH heterodimers, and tetrameric or hexameric complexes containing Lim homeodomain factors. Interactions between these classes of transcription factors are suggested by their synergistic activity in the ventral spinal cord in generating motor neurons and ventral interneurons. Together, there is extensive evidence supporting essential roles for neural bHLH transcription factors in the timing and process of neuronal differentiation, and in the specification of neuronal subtype. However, the mechanistic complexities involved in these functions remain a fertile area for future studies.

See also: Drosophila Apterous Neurons: from Stem Cell to Unique Neuron; Helix–Loop–Helix (bHLH) Proteins: Hes Family; Notch Signal Transduction: Molecular and Cellular Mechanisms; Notch Pathway: Lateral Inhibition; Oligodendrocyte Specification.

Further Reading

Bertrand N, Castro DS, and Guillemot F (2002) Proneural genes and the specification of neural cell types. *Nature Reviews Neuroscience* 3: 517–530.

Caspary T and Anderson KV (2003) Patterning cell types in the dorsal spinal cord: What the mouse mutants say. *Nature Reviews Neuroscience* 4: 289–297.

Castro B, Barolo S, Bailey AM, and Posakony JW (2005) Lateral inhibition in proneural clusters: *Cis*-regulatory logic and default repression by suppressor of hairless. *Development* 132: 3333–3344.

Fode C, Ma Q, Casarosa S, Ang SL, Anderson DJ, and Guillemot F (2000) A role for neural determination genes in specifying the dorsoventral identity of telencephalic neurons. *Genes Development* 14: 67–80.

Gowan K, Helms AW, Hunsaker TL, et al. (2001) Crossinhibitory activities of Ngn1 and Math1 allow specification of distinct dorsal interneurons. *Neuron* 31: 219–232.

Hatakeyama J and Kageyama R (2004) Retinal cell fate determination and bHLH factors. *Seminars in Cell and Developmental Biology* 15: 83–89.

Jan YN and Jan LY (1994) Neuronal cell fate specification in *Drosophila*. *Current Opinion in Neurobiology* 4: 8–13.

Lai EC (2004) Notch signaling: Control of cell communication and cell fate. *Development* 131: 965–973.

Lee S-K and Pfaff SL (2003) Synchronization of neurogenesis and motor neuron specification by direct coupling of bHLH and homeodomain transcription factors. *Neuron* 38: 731–745.

Massari ME and Murre C (2000) Helix-loop-helix proteins: Regulators of transcription in eucaryotic organisms. *Molecular and Cellular Biology* 20: 429–440.

Nieto M, Schuurmans C, Britz O, and Guillemot F (2001) Neural bHLH genes control the neuronal versus glial fate decision in cortical progenitors. *Neuron* 29: 401–413.

Novitch BG, Chen AI, and Jessell TM (2001) Coordinate regulation of motor neuron subtype identity and pan-neuronal properties by the bHLH repressor Olig2. *Neuron* 31: 773–789.

Parras CM, Schuurmans C, Scardigli R, Kim J, Anderson DJ, and Guillemot F (2002) Divergent functions of the proneural genes *Mash1* and *Ngn2* in the specification of neuronal subtype identity. *Genes and Development* 16: 324–338.

Tomita K, Moriyoshi K, Nakanishi S, Guillemot F, and Kageyama R (2000) Mammalian achaete–scute and atonal homologs regulate neuronal versus glial fate determination in the central nervous system. *EMBO Journal* 19: 5460–5472.

Zhou Q and Anderson DJ (2002) The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification. *Cell* 109: 61–73.

Hemiplegic Migraine and Channel Disorders

P J Goadsby, Institute of Neurology, London, UK; and University of California at San Francisco, San Francisco, CA, USA
A M J M van den Maagdenberg and M D Ferrari, Leiden University Medical Center, Leiden, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Migraine: General Principles

Migraine is a common primary headache disorder, typically characterized by disabling attacks of severe throbbing unilateral headache, accompanied by nausea and sensitivity to sound and light and head movement, lasting approximately 1 day. In one-third of patients there is a preceding aura that typically lasts 20–60 min. It usually consists of homonymous visual symptoms, such as flashing zigzag lights and visual loss, which begin paracentrally and slowly expand over minutes as a hemifield defect. Migraine aura may also include other transient focal neurological symptoms. Table 1 summarizes the diagnostic criteria for migraine according to the International Headache Society. The disease is broadly classified into (1) migraine with aura, in which at least some of the attacks are temporally associated with distinct transient focal neurological aura symptoms; and (2) migraine without aura, in which there are no associated neurological symptoms of a focal nature. Many patients believe that their attacks are precipitated by specific trigger factors, as listed in Table 1. A small group of patients have limb weakness with their attacks, which is termed hemiplegic aura.

Hemiplegic Migraine

Patients may present isolated (sporadic) or with a family history of similar attacks (familial) of an aura that involves weakness in the limbs, termed hemiplegic migraine (Table 2). Although most patients have weakness (hemiparesis) as opposed to paralysis (hemiplegic), the term hemiplegic is used to indicate either clinical manifestation. These patients have substantially similar headache phases, the pathophysiology of which is discussed elsewhere in this encyclopedia. Here, we discuss

- clinical features of familial hemiplegic migraine (FHM),
- the genetics of hemiplegic migraine,
- functional consequences of FHM mutations,
- the pathophysiology of aura – animal and human studies, and
- clinical features of FHM.

FHM is a rare, severe, monogenic subtype of migraine with aura characterized by at least some degree of hemiparesis during the aura. The hemiparesis may last from minutes to several hours or even days. Patients are frequently initially misdiagnosed with epilepsy. Apart from the hemiparesis, the other headache and aura features of the FHM attack are identical to those of attacks of the common types of migraine. Perhaps a major difference between the aura symptoms and the attack proper is that the attack itself is effectively treated with relatively specific antimigraine treatments, the serotonin 5-$HT_{1B/1D}$ receptor agonists (the triptans), whereas there is no established treatment that can abort the aura phase. Moreover, since hemiplegic migraine was systematically excluded from clinical trial programs with triptans, they are explicitly contraindicated in hemiplegic migraine. The most rational reason to continue this prohibition derives from a placebo-controlled trial that demonstrated that acute triptan administration during aura was ineffective for both aura and consequent headache. It is widely considered that the prohibition should be based on safety, but since triptans do not measurably alter brain blood flow, this is a theoretical issue. In addition to attacks with hemiparesis, the majority of FHM patients also experience attacks of 'normal' migraine with or without aura. As in the common forms of migraine, attacks of FHM may be triggered by mild head trauma. Thus, from a clinical standpoint, FHM seems a valid model for the common forms of migraine. Major clinical differences, apart from the hemiparesis, include that FHM in 20% of cases may also be associated with cerebellar ataxia and other neurological symptoms, such as epilepsy, mental retardation, brain edema, and (fatal) coma. To date, three genes for FHM have been published, but based on unpublished linkage results in several families, there are more.

Genetics of FHM

The FHM1 *CACNA1A* Gene

The first gene identified for FHM was the *CACNA1A* gene on chromosome 19p13. It is responsible for approximately 50% of all known families with FHM (Figure 1). The FHM1 gene encodes the ion-conducting, pore-forming α_{1A} subunit of $Ca_v2.1$ (P/Q-type), voltage-gated, neuronal calcium channels. The main function of neuronal P/Q-type calcium channels is to modulate release of neurotransmitters, both at peripheral neuromuscular junctions and at central synapses, mainly within the cerebellum, brain stem, and cerebral cortex. More than 50

Table 1 International Headache Society features of migraine and triggers believed to precipitate migraine attacks

Features of migraine	
Repeated episodic headache (4–72 h) with the following features:	
Any two of	Any one of
Unilateral	Nausea/vomiting
Throbbing	Photophobia and phonophobia
Worsened by movement	
Moderate or severe	
Triggers believed to precipitate migraine attacks	
Altered sleep patterns: becoming tired or oversleep	
Skipping meals	
Overexertion	
Weather change	
Stress or relaxation from stress	
Hormonal change, such as menstrual periods	
Excess afferent stimulation: bright lights, strong smells	
Chemicals: alcohol or nitrates	

Table 2 Familial hemiplegic migraine and sporadic hemiplegic migraine

Familial hemiplegic migraine

Description: Migraine with aura including motor weakness and where at least one first- or second-degree relative has migraine aura including motor weakness.

Diagnostic criteria

A. At least two attacks fulfilling B–E.
B. Fully reversible motor weakness and at least one of the following other aura symptoms: visual, sensory, or speech disturbance.
C. At least two of the following:
   1. At least one aura symptom develops gradually over $\geq$5 min and/or different symptoms occur in succession.
   2. Each aura symptom lasts less than 24 h.
   3. Headache that meets criteria B–D for migraine without aura (1.1) begins during the aura or follows aura within 60 min.
D. At least one first- or second-degree relative has migraine attacks with aura including motor weakness.
E. Not attributed to another disorder.

Sporadic hemiplegic migraine

Description: Migraine with aura including motor weakness but no first- or second-degree relative has aura including motor weakness.

Diagnostic criteria

A. At least two attacks fulfilling B–D.
B. Fully reversible motor weakness and at least one of the following other aura symptoms: visual, sensory, or speech disturbance.
C. At least two of the following:
   1. At least one aura symptom develops gradually over $\geq$5 min and/or different symptoms occur in succession.
   2. Each aura symptom lasts less than 24 h.
   3. Headache that meets criteria B–D for migraine without aura (1.1) begins during the aura or follows aura within 60 min.
D. No first- or second-degree relative has migraine attacks with aura including motor weakness.
E. Not attributed to another disorder.

CACNA1A mutations have been associated with a wide range of clinical phenotypes (Figure 1). These include pure forms of FHM, combinations of FHM with various degrees of cerebellar ataxia or fatal coma due to excessive cerebral edema, and disorders not associated with FHM such as episodic ataxia type 2, progressive ataxia, spinocerebellar ataxia type 6, and absence and generalized epilepsy. Interestingly, in several FHM families, FHM1 *CACNA1A* mutations were also found in family members who only had 'normal' nonparetic migraine aura. This suggests that gene mutations for FHM may also be responsible for the common forms of migraine, probably due to different genetic and nongenetic modulating factors.

The FHM2 *ATP1A2* Gene

The *ATP1A2* FHM2 gene on chromosome 1q23 encodes the α_2 subunit of a Na^+, K^+ pump ATPase. This catalytic subunit binds Na^+, K^+, and ATP, and it utilizes ATP hydrolysis to exchange Na^+ ions out of the cell for K^+ ions into the cell. Na^+ pumping provides the steep Na^+ gradient essential for the transport of glutamate and Ca^{2+}. The gene is predominantly expressed in neurons in neonates and in glial cells in adults. In adults, an important function of this specific ATPase is to modulate the reuptake of potassium and glutamate from the synaptic cleft into the glial cell. Mutations in the *ATP1A2* gene are responsible for at least 20% of FHM cases (Figure 2) and have been associated with pure FHM and FHM in combinations with cerebellar ataxia, alternating hemiplegia of childhood, benign focal infantile convulsions, and other forms of epilepsy. In an Italian family, a variant in the *ATP1A2* gene segregated with basilar-type migraine, a subtype of migraine with aura characterized by aura symptoms attributable to the brain stem and both occipital lobes. In two non-FHM migraine families, *ATP1A2* variants were identified, suggesting that this gene may be involved in the susceptibility to common forms of migraine.

The FHM3 *SCNA1* Gene

The *SCNA1* gene on chromosome 2q24 encodes the α subunit of a neuronal voltage-gated sodium ($Na_v1.1$) channel. The $Na_v1.1$ channel is mainly responsible for the generation and propagation of neuronal action potentials. Different mutations in this gene are known to be associated with epilepsy and febrile seizures. A novel Q1489K mutation has been identified in three German FHM families of common ancestry. Another new *SCNA1* mutation was found in a North American FHM family, confirming the relationship between *SCNA1* and FHM3 (Figure 3).

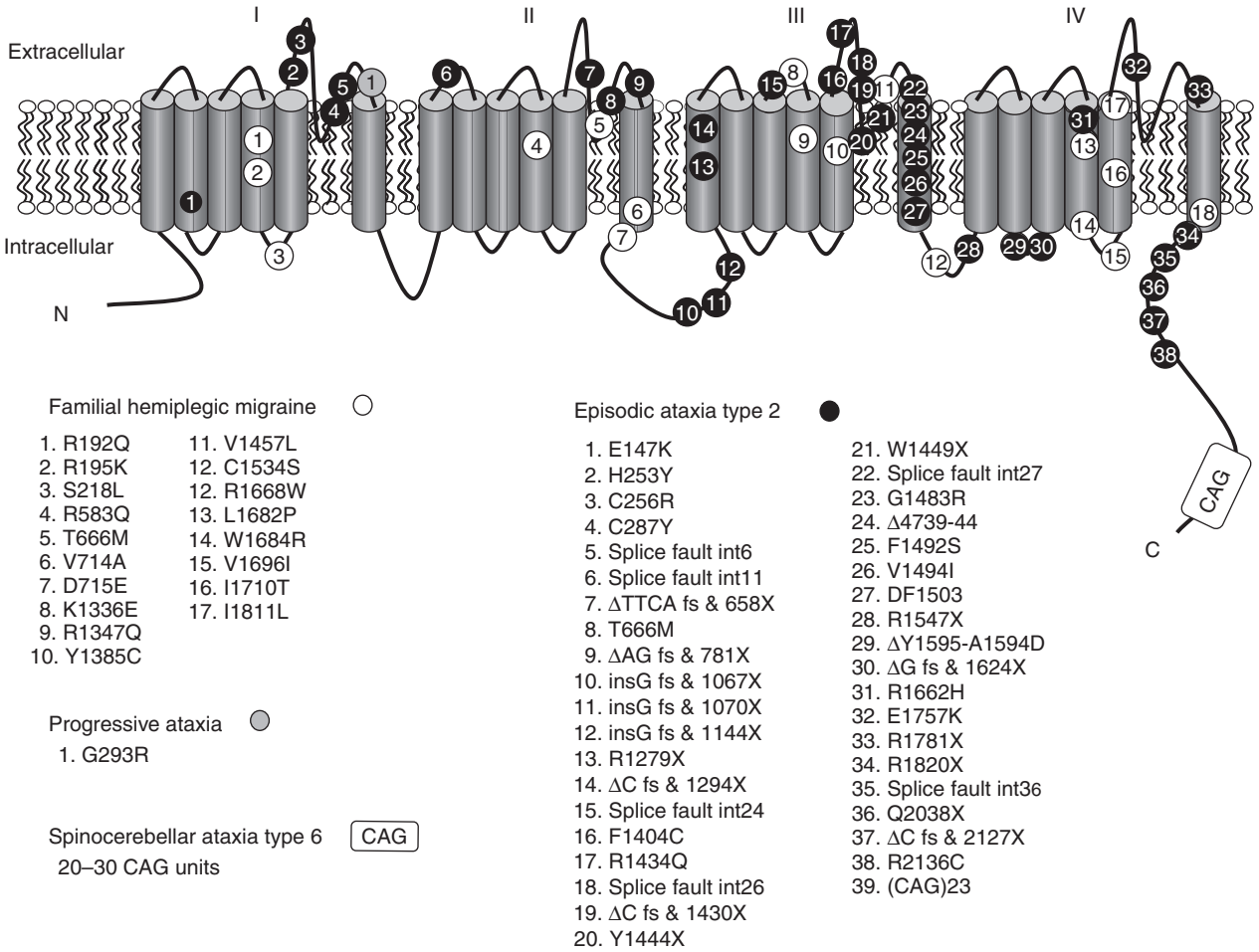

Figure 1 The *CACNA1A* gene with mutations. The $Ca_v2.1$ pore-forming subunit of P/Q-type voltage-gated calcium channels is located in the neuron membrane and contains four repeated domains each encompassing six transmembrane segments. Positions of mutations and associated clinical phenotypes are depicted in the schematic representation of the protein. *CACNA1A* ref. seq.: GenBank Ac. nr. X99897.

Sporadic Hemiplegic Migraine

Hemiplegic migraine patients are not always clustered in families. Sporadic patients, without affected family members, are often seen and may sometimes represent the first 'FHM patient' (*de novo* mutation) in a family. Apart from sharing the clinical phenotype with the common forms of migraine, similar to FHM, sporadic hemiplegic migraine (SHM) and normal migraine also show a remarkable genetic epidemiological relationship. SHM patients have an highly increased risk of also suffering from typical migraine with aura, and their first-degree relatives have a highly increased risk of both migraine with aura and migraine without aura. Although in an initial study *CACNA1A* mutations were found in only 2 of 27 SHM patients, a much larger study did find mutations in the FHM genes in a higher proportion of SHM patients, confirming a genetic relationship between FHM and SHM.

Functional Consequences of FHM Gene Mutations

Understanding the functional consequences of gene mutations is crucial to the understanding of the disease pathways.

Functional Studies in Cellular Models for FHM1 *CACNA1A* Mutations

Several FHM and episodic ataxia type 2 mutations have been analyzed with electrophysiological techniques in neuronal and nonneuronal cell models. Whereas episodic ataxia type 2 *CACNA1A* mutations all show a dramatic decrease or even complete loss of current density, FHM1 mutations cause different effects on channel conductance, kinetics, and/or expression in transfected cells. The most consistent change found with FHM1 mutations, when tested in a single channel configuration, was a hyperpolarizing

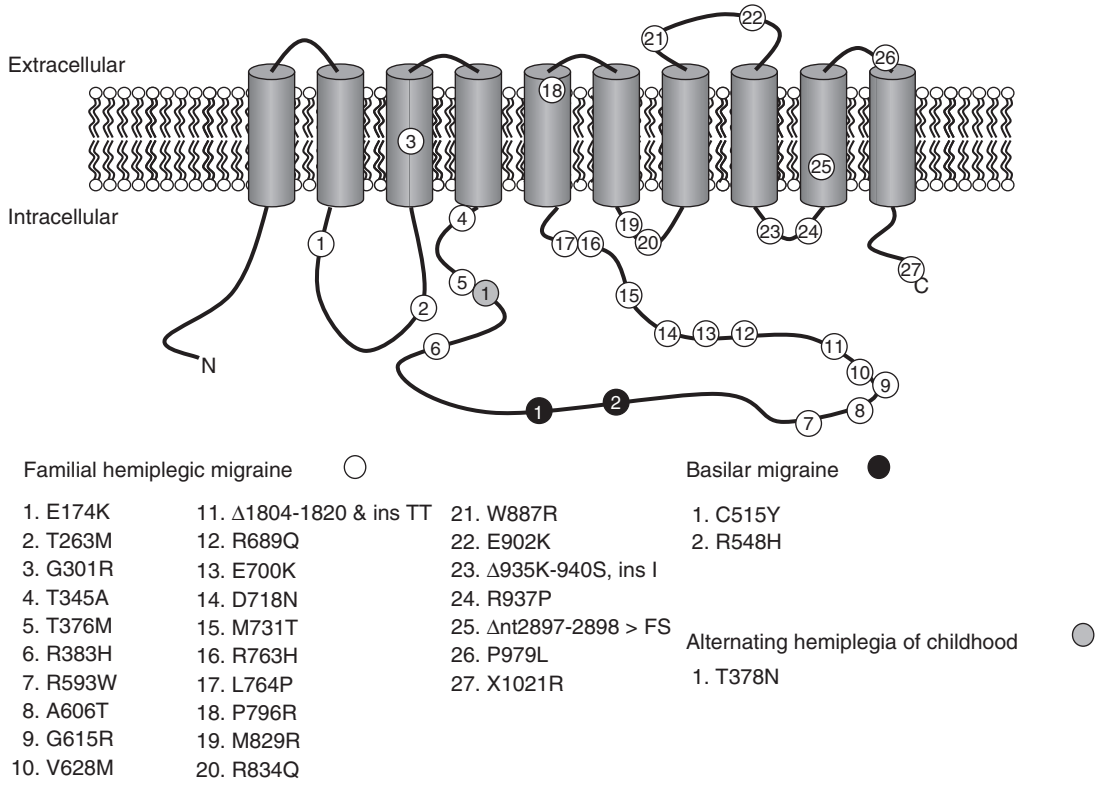

Figure 2 The *ATP1A2* gene with mutations. The α_2 subunit of sodium potassium pumps is located in the plasma membrane and contains 10 transmembrane segments. Positions of mutations and associated clinical phenotypes are indicated in the schematic drawing of the protein. *ATP1A2* ref. seq.: GenBank Ac.nr. NM_000702.

shift of approximately 10 mV of the activation voltage. The change in calcium influx could, however, alter during high neuronal activity. Mutant T666M and V714A channels have a low conductance mode that may sometimes switch to the wild-type state. For other FHM mutations, such as R583Q and D715E, accumulation of inactivated channels was observed during repetitive stimulation. Such phenomena could contribute to the paroxysmal presentation of symptoms.

The gain-of-function effects found in single channel test models in theory will lead to an easier opening of channels in neurons. The overall change in calcium influx, however, is difficult to predict and seems to depend, at least in part, on the model that is being used. For instance, evidence for reduced calcium influx at the whole cell level in transfections of cultured mouse hippocampal neurons has been reported. What will happen in the mutant brain will be determined by the delicate interplay between the functional effects of a particular mutation, the different channel properties and density, the different channel subunits, and the direct and indirect cellular environment. The observation that different auxiliary β subunits of calcium channels can modulate the consequence of FHM mutations adds to the complexity of predicting calcium channel functioning. Therefore, it seems more appropriate to study the functional consequences of gene mutations in knockin mouse models carrying human pathogenic mutations.

Functional Studies in Naturally Occurring *Cacna1a* Mouse Mutants

Naturally occurring mouse mutants with *Cacna1a* missense mutations (Tottering, Rocker, and Rolling Nagoya) or *Cacna1a* truncation mutations (Leaner) display different combinations and severities of various types of epilepsy and ataxia. A reduction in calcium current density appears to be the main effect of these mutated P/Q-type channels, with a change in channel kinetics for the Leaner and Rolling Nagoya mutants. Two *Cacna1a*-null (knockout) mouse models were generated showing ataxia, dystonia, and lethality at a young age. The total Ca^{2+} influx in cerebellar cells and neurotransmission at the neuromuscular junction was reduced in these mice. Loss of P/Q-type channels could be (partly) compensated for by N-, R-, and-L-type channels. Moreover, Leaner mice showed an increased threshold for cortical spreading depression and a reduced release of cortical glutamate.

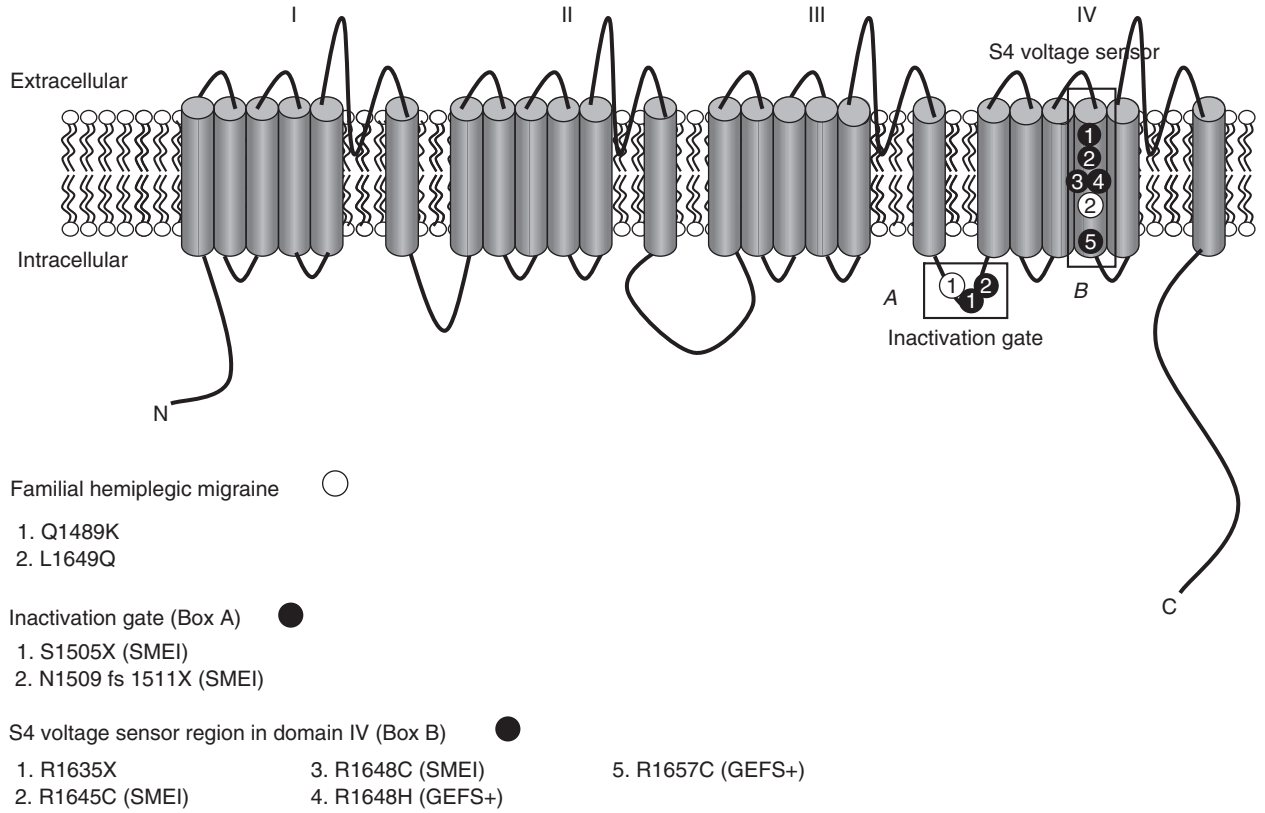

Figure 3 The *SCN1A* gene with mutations. The pore-forming subunit of $Na_v1.1$ sodium channels is located in the neuron membrane and contains four repeated domains each encompassing six transmembrane segments. More than 150 mutations have been identified for severe myoclonic epilepsy in infancy (SMEI) or generalized epilepsy with febrile seizures plus (GEFS+). For clarity, only SMEI and GEFS+ mutations are shown that reside in two functional domains (e.g., inactivation gate and D4/S4 voltage sensor) where familial hemiplegic migraine mutations were identified. Positions of mutations and associated clinical phenotypes are depicted in the schematic representation of the protein. *SCN1A* ref. seq.: GenBank Ac.nr. NM_006920.

Functional Studies in FHM1 *Cacna1a* Knockin Mouse Mutants

A knockin mouse model was generated carrying the human FHM1 R192Q mutation. Unlike the natural *Cacna1a* mutant mouse models, transgenic R192Q mice exhibit no overt clinical phenotype or structural abnormalities. This is very similar to the human situation: the R192Q mutation causes only mild FHM attacks, very similar to the common forms of migraine, albeit with hemiparesis, without other neurological symptoms. Extensive functional analysis revealed multiple gain-of-function effects. These include increased Ca^{2+} influx in cerebellar neurons; increased release of neurotransmitters at the neuromuscular junction, both spontaneous and upon stimulation at low Ca^{2+}; and, in the intact animal, a reduced trigger threshold for cortical spreading depression that propagates with increased velocity. It seems that whole-animal studies may be better suited to dissect the effects of mutations and to understand the integrated physiology of the disease. Other studies aiming at functional changes within the brain stem are under way and will shed more light on the important question of whether migraine-related mechanisms within the trigeminocervical complex are also affected.

Functional Studies in Cellular Models for FHM2 *ATP1A2* Mutations

Functional analysis of mutated proteins revealed inhibition of pump activity, decreased affinity for K^+, and decreased catalytic turnover resulting in reduced reuptake of K^+ and glutamate into glial cells.

Functional Studies in *Atp1a2* Knockout Mouse Models

Two groups have generated α_2 subunit-deficient (*Atp1a2*-null) mice. These mice died immediately after birth because of severe motor deficits and absent respiration. *Atp1a2*-null fetuses of 18.5 days revealed selective neuronal apoptosis in the amygdala and piriform cortex in response to neural hyperactivity. *Atp1a2*-null mice on 129sv genetic background

displayed frequent generalized seizures and died within 24 h after birth. Epilepsy is also a feature of the clinical phenotype in humans with *ATP1A2* mutations. Heterozygous *Atp1a2*+/_ mice are viable. Their heart shows a hypercontractile state with positive inotropic response and resembles what is typically seen after the administration of cardiac glycosides. In addition, they revealed enhanced fear and anxiety behaviors after conditioned fear stimuli; this is probably due to neuronal hyperactivity in the amygdala and piriform cortex. No *Atp1a2* knockin models are available.

Functional Studies in Cellular Models for the FHM3 Mutation

Functional consequences of the Q1489K *SCN1A* mutation have been analyzed in the highly homologous SCN5A protein. This missense mutation is located within a domain critical for fast inactivation of the sodium channel. It causes a two- to fourfold increased acceleration of recovery from fast inactivation. This would predict enhanced neuronal excitation and release of neurotransmitter. No knockout or knockin mouse mutants are available for this gene.

Pathophysiology of Aura

Cortical Spreading Depression in Experimental Animals

It is now well accepted that the migraine aura is not due to reactive vasoconstriction, as was previously believed for several decades, but rather is neurally driven and most likely caused by the human equivalent of the cortical spreading depression (CSD) of Leao. In experimental animals, CSD is a short-lasting (<1 min) intense and steady depolarization of neuronal and glial cell membranes that spreads into contiguous areas of brain cortex at a rate of 2–5 mm min^{-1}, regardless of functional cortical divisions or arterial territories. It is accompanied by a transient total loss of neuroglial membrane integrity, a massive influx of Ca^{2+} and Na^{+}, and a massive efflux of K^{+} causing highly elevated extracellular K^{+} levels. This results in a spreading wave of brief excitation followed by a longer lasting inhibition of spontaneous and evoked neuronal activity that traverses the cortex at a rate of approximately 3–5 mm min^{-1}. In experimental animals, the electrophysiological changes are associated with characteristic triphasic cerebral blood flow (CBF) changes. Initially, there is a small and very brief reduction in CBF. This is followed by a profound increase in CBF for several minutes. The third phase consists of a reduction of CBF that may last for up to 1 h and is accompanied by a loss of the cerebrovascular response to hypercapnia. CSD appears to be a self-defense mechanism of the brain to strong stimuli and can be triggered by electrical stimulation of brain tissue, cortical trauma, cerebral ischemia, or cortical application of high concentrations of K^{+} or neuroexcitatory amino acids such as glutamate.

CSD and the Migraine Aura

There is a considerable body of clinical evidence that CSD is the likely basis of migraine aura. Visual aura symptoms typically spread or march from the center of the visual field to the periphery at a speed of approximately 3 mm min^{-1} when translated to the visual cortex. This is very similar to the propagation rate of CSD in experimental animals. The positive (e.g., scintillations and paraesthesias) and negative (e.g., scotomata and paresis) phenomena of the migraine aura could be well explained by the initial transient hyperexcitation front of CSD followed by neuronal depression. Most important, however, functional neuroimaging studies in humans have convincingly demonstrated that the CBF changes that occur during migraine aura are very similar to those observed in experimental animals during CSD. Using functional MRI, a focal increase in BOLD signal was identified spreading into the occipital cortex at a rate of 3.5 mm min^{-1}. The cortical direction and speed of the spread were congruent with the visual experiences of the patient. The increased BOLD signal was followed by a decrease. This pattern suggests an initial brief rise of CBF followed by a longer-lasting oligemia as is seen in experimental CSD.

Although the evidence that CSD causes the migraine aura is mounting, there is much debate as to whether CSD may trigger the rest of the migraine attack as well through activation of the trigeminovascular system.

Can CSD Trigger the Mechanisms for the Headache Phase?

Whether CSD can initiate the migraine headache cascade is controversial. KCl-induced CSD in the rat parietal cortex activates ipsilateral trigeminal nucleus caudalis neurons and CSD causes a long-lasting blood flow increase in the rat middle meningeal artery and a dural plasma protein leakage that can be inhibited by ipsilateral trigeminal nerve section. Chronic daily, but not acute, administration of migraine prophylactic drugs in rats dose- and duration-dependently suppressed KCl-induced CSD frequency by 40–80% and increased the triggering threshold for inducing CSD, although one study suggested topiramate can acutely inhibit CSD.

However, reviewing a number of mainly clinical arguments casts doubt on whether aura can trigger headache. The fact that migraine aura only occurs in up to one-fourth to one-third of migraine patients makes it difficult to explain how CSD might be involved in the majority of migraineurs in whom attacks are not associated with aura. The only explanation to rescue the 'CSD headache trigger' hypothesis would be that CSD might occur in clinically silent subcortical areas of the brain without propagating through all six cortical laminae within the neocortex. Direct human evidence for such events is difficult to obtain and, apart from possibly one case, is lacking. Other clinical observations that argue against the role of CSD in triggering the headache are that the headache may occur on the same side of the aura instead of the expected opposite side if CSD would have triggered the trigeminovascular system, and that aura may sometimes occur after the headache has started. Furthermore, aura appears to not be confined to migraine; it is reported with attacks of cluster headache, paroxysmal hemicrania, and hemicrania continua. Finally, intranasal ketamine could abort migraine aura without affecting the ensuing headache. From this discussion, it is clear that the exact role of CSD in triggering migraine headache mechanisms in humans remains unclear and needs further investigation.

A Mechanism for Triggering FHM Attacks

Three different genes encoding three very different proteins with apparent different mechanisms have been associated with FHM. How can we fit these apparently very diverting mechanisms into one (final) common pathway? Mutations in the FHM1 calcium gene cause increased neuronal release of neurotransmitters and, in the cortex, specifically the neuroexcitatory amino acid glutamate that can induce, maintain, and propagate CSD. Mutations in the FHM2 sodium/potassium pump gene cause reduced reuptake of K^+ and glutamate from the synaptic cleft into the glia cell. Mutations in the FHM3 sodium channel gene result in hyperexcitability and most likely increased release of neurotransmitters in the synaptic cleft. The overall result is increased levels of glutamate and K^+ in the synaptic cleft, resulting in an increased propensity for CSD. This would easily explain the aura of FHM attacks.

A most important question that remains to be answered is whether the same mechanisms are also involved in the common forms of migraine with and without aura. Interestingly, although drugs with migraine prophylactic activity belong to a wide range of pharmacological classes, such as antiepiletics and blockers of calcium channels and serotinergic, β-adrenergic, and histamine receptors, they all seem to share anti-CSD activity. CSD inhibition may thus be a promising model system to contribute to the development of preventive medicines. An interesting compound in this respect is Tonabersat, which has entered clinical trials in migraine. Tonabersat (SB-220453) inhibits CSD, CSD-induced nitric oxide (NO) release, and cerebral vasodilation. It does not constrict isolated human blood vessels, but it does inhibit trigeminally induced craniovascular effects. Tonabersat is inactive in the human NO model of migraine, as is propranolol, although valproate showed some activity in that model. If proven effective, it would be the first migraine prophylactic agent developed on the basis of the CSD hypothesis.

See also: Calcium Channels; Epilepsy: Channelopathies; Headache; Paroxysmal Paralysis; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Voltage-Gated Calcium Channels.

Further Reading

De Fusco M, Marconi R, Silvestri L, et al. (2003) Haploinsufficiency of ATP1A2 encoding the Na^+/K^+ pump α_2 subunit associated with familial hemiplegic migraine type 2. *Nature Genetics* 33: 192–196.

Dichgans M, Freilinger T, Eckstein G, et al. (2005) Mutation in the neuronal voltage-gated sodium channel *SCN1A* causes familial hemiplegic migraine. *Lancet* 366: 371–377.

Dove LS, Abbott LC, and Griffit WH (1998) Whole-cell and single-channel analysis of P-type calcium currents in cerebellar Purkinje cells of leaner mutant mice. *Journal of Neuroscience* 18: 7687–7699.

Goadsby PJ (2005) Can we develop neurally-acting drugs for the treatment of migraine? *Nature Reviews Drug Discovery* 4: 741–750.

Goadsby PJ (2005) Migraine pathophysiology. *Headache* 45: S14–S24.

Goadsby PJ, Lipton RB, and Ferrari MD (2002) Migraine – current understanding and treatment. *New England Journal of Medicine* 346: 257–270.

Giffin NJ, Ruggiero L, Lipton RB, et al. (2003) Premonitory symptoms in migraine: An electronic diary study. *Neurology* 60: 935–940.

Haan J, Kors EE, Vanmolkot KR, van den Maagdenberg AM, Frants RR, and Ferrari MD (2005) Migraine genetics: An update. *Current Pain and Headache Reports* 9: 213–220.

Hadjikhani N, Sanchez del Rio M, Wu O, et al. (2001) Mechanisms of migraine aura revealed by functional MRI in human visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 98: 4687–4692.

Headache classification committee of the international headache society (2004) The international classification of headache disorders (2nd edn.). *Cephalalgia* 24: 1–160.

Lance JW and Goadsby PJ (2005) *Mechanism and Management of Headache*. New York: Elsevier.

Lorenzon NM, Lutz CM, Frankel WN, and Beam KG (1998) Altered calcium channel currents in Purkinje cells of neurological mutant mouse leaner. *Journal of Neuroscience* 18: 4482–4489.

Mori Y, Wakamori M, Oda S, et al. (2000) Reduced voltage sensitivity of activation of P/Q-type Ca^{2+} channels is associated with the ataxic mouse mutation rolling Nagoya (tg(rol)). *Journal of Neuroscience* 20: 5654–5662.

Moskowitz MA, Bolay H, and Dalkara T (2004) Deciphering migraine mechanisms: Clues from familial hemiplegic migraine genotypes. *Annals of Neurology* 55: 276–280.

Olesen J, Tfelt-Hansen P, Ramadan N, Goadsby PJ, and Welch KMA (2005) *The Headaches*. Philadelphia: Lippincott Williams & Wilkins.

Ophoff RA, Terwindt GM, Vergouwe MN, et al. (1996) Familial hemiplegic migraine and episodic ataxia type-2 are caused by mutations in the Ca^{2+} channel gene CACNL1A4. *Cell* 87: 543–552.

Silberstein SD (2004) Migraine. *Lancet* 363: 381–391.

Silberstein SD, Lipton RB, and Goadsby PJ (2002) *Headache in Clinical Practice*. London: Martin Dunitz.

Wakamori M, Yamazaki K, Matsunodaira H, et al. (1998) Single tottering mutations responsible for the neuropathic phenotype of the P-type calcium channel. *Journal of Biological Chemistry* 273: 34857–34867.

Hemispheric Specialization and Cognition

M T Banich, University of Colorado at Boulder, Boulder, CO, USA

© 2009 Published by Elsevier Ltd.

Hemispheric Differences in Cognition

Discovery of Hemispheric Differences in Cognition

Paul Broca, a French neurologist and anthropologist, is generally considered the first person to have clearly illustrated hemispheric differences in cognitive function. As such, he is often said to have discovered the phenomenon. His critical insight derived from his visit with a patient who demonstrated an interesting dissociation in language abilities. Although able to understand what was said to him, the man had an inability to produce speech, being able only to utter the syllable 'tan.' Broca's postmortem analysis of his brain showed that damage was localized to the third convolution of the inferior frontal gyrus in the left hemisphere. Broca then went on to examine other patients who exhibited a similar cognitive profile – an inability to produce speech in the face of a retained ability to comprehend speech. In all cases, the damage was localized to the same region but, most important, always in the left hemisphere – a region now known as Broca's area. As a result of these findings, Broca proposed in 1863 that the left hemisphere is specialized, or dominant, for speech output. Thus, Broca's paper was the first systematic and compelling demonstration of hemispheric specialization of function. In fact, unlike most other aspects of hemispheric specialization, this one is absolute: the right hemisphere has no ability to control speech output in practically all right-handed individuals.

Evidence from Patients with Unilateral Brain Damage

As a result of Broca's discovery, the idea of cerebral dominance was overgeneralized. Probably because language was considered synonymous with thought, his work was interpreted to mean that the left hemisphere was dominant for all aspects of cognitive function. This idea only began to erode gradually over the next century as studies of patients with unilateral brain damage demonstrated different consequences depending on which hemisphere was damaged. As apparent to any neurologist or clinical neuropsychologist, left hemisphere damage usually results in deficits in the domains of verbal, sequential, and analytic processing. For example, aphasia is a common consequence of left hemisphere damage. In contrast, right hemisphere damage typically yields deficits in nonverbal, holistic, and Gestalt processing. For example, deficits in visuospatial processing are more often observed after right hemisphere damage.

Evidence from Split-Brain Patients

In the 1960s, research by Nobel laureate Roger Sperry and colleagues with split-brain patients dramatically demonstrated the relative specialization of the cerebral hemispheres. In these split-brain patients, the main nerve fiber tract connecting the cerebral hemispheres, the corpus callosum, is severed for the treatment of intractable epilepsy. As a result, higher order information, such as that about an item's identity (e.g., a car, the letter 'A,' and the face of Bill Clinton), cannot be transferred from one hemisphere to the other. Thus, information directed to a single hemisphere is functionally isolated to that hemisphere. This situation provides a unique opportunity to examine the relative specialization of the cerebral hemispheres because each hemisphere's capabilities can be examined in isolation from those of its partner. As a result, research with split-brain patients has yielded much important information about hemispheric specialization. Absolute differences have been demonstrated only for a couple of functions, namely speech output and phonological processing, which are under sole control of the left hemisphere. Both hemispheres can perform all other tasks, albeit with differing levels of ability and in different manners. Whereas the left hemisphere has a rich ability to perform most all language tasks, the vocabulary of the right hemisphere is much more limited, as is its ability to process complicated grammatical functions. On the other hand, the right hemisphere is superior at processing most types of spatial relationships, especially those involving three-dimensional relations or complicated geometries.

Perceptual Asymmetries in Neurologically Intact Individuals

The relative specializations of the cerebral hemispheres can also be demonstrated in neurologically intact individuals through the use of methods that essentially pit the hemispheres against one another. These methods, which include tachitoscopic presentation, dichotic listening, and dichaptic presentation, all take advantage of the neuroanatomical wiring of the human brain that transfers information from sensory receptors to the contralateral sensory cortex. In these methods, information is presented laterally so it is received either solely or predominantly by one hemisphere. Then behavioral performance, with

regard to either reaction time or accuracy, can be examined depending on which hemisphere initially received the sensory information. These behavioral measures are often referred to as perceptual asymmetries because they reflect the asymmetry in the perception of information depending on the hemisphere to which information was initially directed. Even though the corpus callosum in neurologically intact individuals contains more than 250 million fibers by which the hemispheres can communicate, differences in performance are nonetheless observed. Typically, these effects are in the range of a 10% difference in accuracy in performance or a 20–50 ms difference in reaction time. These studies provide converging evidence with data obtained from patients with unilateral brain damage and from split-brain patients. Myriads of studies have confirmed a left hemisphere superiority for processing verbal information and a right hemisphere superiority for processing nonverbal material, regardless of sensory modality – visual, auditory, or tactile.

Neuroimaging Studies of Hemispheric Specialization

The surge in neuroimaging during the past decade has also served to emphasize that the specializations of the hemispheres are more relative than absolute. These studies have shown that for most all tasks, activation is bilateral, although not necessarily of equal extent nor intensity. Even classic language tasks, such as verbal word reading, activate both hemispheres, although the activation is more left-sided than right-sided. In addition, the complimentarity of the hemispheres is revealed by these studies. For example, single-word processing leads to greater left than right hemisphere activation, but understanding the nonliteral meaning of language, such as analogy or the moral of a story, leads to greater right than left hemisphere activation.

Asymmetries Related to Emotion and Emotional Processing

Hemispheric specialization is not limited to cognitive function; it is found for emotional processes as well. The majority of evidence suggests that the right hemisphere is specialized for the interpretation of emotional information, including information contained in tone of voice and facial expression. Moreover, it is also specialized for the production of emotional cues that serve a communicative function (e.g., a smile that sends a communicative signal to someone else letting him or her know you are happy or pleased).

In contrast, lateralization of mood – that is, the subjective experience of one's internal emotional state – appears to rely on a pattern of brain activation across prefrontal and parietal regions. A large body of research has indicated that asymmetry of activation of frontal regions of the brain is linked to mood states. Greater activation of left than right frontal regions is associated with positive mood and approach behaviors. In contrast, greater activation of the right than left frontal regions is associated with negative mood and avoidance behavior. Moreover, individual differences in these asymmetries have been linked to differences in temperament in infants and to susceptibility to depression in later life. Overlaid on these effects of valence (positive and negative) are differences in activation of right parietal regions, which are linked to arousal. Depressed mood is associated with decreased activity of right parietal regions, whereas the panic associated with heightened anxiety is associated with increased activity of right parietal regions.

Models of Hemispheric Specialization

The broad body of work – from patients with unilateral brain damage, split-brain patients, and studies of perceptual asymmetries with neurologically intact individuals – was originally framed with regard to differences in the type of material that each hemisphere is specialized to process. Initially, the left hemisphere was considered specialized for processing verbal materials, whereas the right hemisphere was considered specialized for processing nonverbal materials. However, in a series of studies in the late 1960s and early 1970s, studies with split-brain patients clearly demonstrated that in many cases both hemispheres were capable of processing a given type of material. What differed, however, was the manner in which they processed that material. This led to a new rubric for understanding hemispheric specialization. The left hemisphere was conceptualized to analyze information in an analytic, piecemeal, and local manner, whereas the right hemisphere was conceptualized to analyze information in a holistic and Gestalt manner. For example, although the right hemisphere of a split-brain patient is superior to the left at identifying a previously viewed face, both can do so. The right hemisphere appears to analyze the overall configuration of the face, such as whether the face is long and narrow or wide and round, and whether the eyes are wide set compared to the width of the face. In contrast, the left hemisphere processes the features or local details, such as the shape of the chin or the eyes. This shift in conceptualizing hemispheric specialization was important because it explained a potential advantage of a specialized brain, namely the ability of simultaneous dual processing. Practically all information can be processed

independently and in a distinct manner by each hemisphere at the same time, providing two distinct ways of simultaneously understanding and interpreting the world.

With this conceptual shift, researchers began to explore hemispheric differences from a computational perspective, with the goal of determining the fundamental differences in computations performed by each hemisphere. One early theory suggested that the hemispheres differed in their ability to process low-level sensory information. In particular, the right hemisphere was conceived as being specialized for processing information of low visual spatial frequency – that is, information that does not shift from dark to light within a small degree of visual space. This information generally provides the general contours and outline of visual forms but not the details. In contrast, the left hemisphere was conceived as being specialized for processing information of high spatial frequency – that is, information that shifts quickly, within a given amount of visual space, from dark to light. Such information provides the detail in visual forms. Moreover, asymmetries in higher order cognitive function were posited to emerge from these sensory asymmetries. For example, the overall configuration of a face would be provided by visual information of low spatial frequency, whereas the detailed information would be provided by visual information of high spatial frequency.

This theory was highly influential and has been subsequently modified. Further research demonstrated that the hemispheres appear to be specialized not for the absolute frequency of sensory information but, rather, for the relative frequency, an effect that holds across different sensory modalities. For example, when individuals are presented with information of high auditory frequency, the left hemisphere exhibits a performance advantage for processing the higher half of those high auditory frequencies and the right hemisphere for processing the lower half of those high auditory frequencies. Findings such as these have been expanded to be accounted for by the double-filtering hypothesis, which argues that hemispheric differences arise from an attentional bias in the information that each processes. It is argued that the right hemisphere employs a low pass filter on incoming information, whereas the left hemisphere employs a high pass filter.

Other researchers have focused on why there might be a need to have distinct ways of processing information in each of the cerebral hemispheres. These theories have focused on how the hemispheres might insulate conflicting or independent processes from one another. For example, in the spatial domain it has been argued that the right hemisphere is specialized to coordinate spatial relationships, considered those that provide information about the distance between objects. In contrast, the left hemisphere is specialized for categorical spatial relations that describe the relationship between items (e.g., above, below, to the right of, and to the left of). These two ways of describing information are considered orthogonal to each other because knowing the coordinate information, such as that one item is 3 ft from another, provides no information about categorical information, such as if one item is behind the other. Computational models suggest that these two types of spatial processing are best supported by independent insulated processing systems because the representations required are mutually incompatible and/or create interference. Performance on spatial tasks is superior in a split rather than unitary computational model, suggesting that the hemispheres are specialized to allow for noninterference of processing.

Similar arguments have been made with regard to language processing. Even though the left hemisphere alone has control over speech output and phonological processing, its specialization for language is also relative. Whereas the left hemisphere has been found to be superior to the right hemisphere in aspects of grammar and syntax, the right hemisphere is superior at processing the nonliteral aspects of language, such as discourse, metaphor, and analogy. These differences may arise from incompatible means of semantic access and/or organization. The left hemisphere appears to access words in a very specific, precise, and local manner, whereas the right hemisphere accesses words in a more diffuse manner that allows activation for more far-flung associates. For example, if a paragraph was about gardening, the left hemisphere would quickly hone in on the particular meaning of 'bug' related to insects, whereas the right hemisphere would hold onto a more diffuse set of meanings, including not only the meaning related to insects but also that related to spying devices. As such, the right hemisphere would be better equipped to make connections across sentences and phrases that would allow for discourse and for nonliteral aspects of language comprehension.

Origins of Hemispheric Asymmetry

Although there was once speculation that cerebral asymmetry is a unique feature of the human brain, hemispheric asymmetry is observed in other species, including not only mammals but also fishes, reptiles, and amphibians. The asymmetries observed in behaviors range from those involved in courtship and copulation to escape behavior and limb use (i.e., 'handedness'). These asymmetries are found not only

for an individual organism (i.e., a right-sided limb preference in a given animal) but also at the population level. For example, apes tend to use the left hand for stabilizing objects and the right hand for fine motor manipulation. Likewise, 59% of certain species of toad prefer to use their right paw to remove an object affixed to their head.

Because hemispheric asymmetries are observed in many other species, much theorizing has focused on how language processing becomes lateralized. Some theories posit that the association of fine motor coordination of the right hand associated with tool use served as the platform for the fine motor control that is associated with the vocalizations that underlie human language. Other theories posit that a lateralized gestural system, when linked to vocalizations, led to language lateralization. Cross-species support for such an idea is provided by research showing that cells in area F5 of the macaque monkey brain, a region homologous to Broca's area in humans, fire when a monkey sees someone else perform a grasping action similar to one which it has just performed. It has been argued that these 'mirror' neurons underlie the ability to form, through gesture, a common communicative system between individuals. Such a common communicative system would then have evolved to include vocalization. Supporting this idea, gestures linked to speech are produced with much more frequency by the right hand in humans and chimpanzees, whereas gestures that do not have a communicative value (e.g., straightening one's clothes) are not produced asymmetrically.

Developmental Issues and Hemispheric Specialization

Given the evidence for an evolutionary history of lateralization of functioning, it is not surprising that hemispheric asymmetries exist at birth. That is not to say, however, that this pattern cannot be modified by environmental factors. Evidence for an inborn pattern of asymmetry comes from numerous sources. First, gyral and sulcul patterns, which differ between right- and left-handers (who, as discussed later, differ in behavioral asymmetry) are present before birth and not modified thereafter. Second, the effects of hemispherectomy at birth differ depending on whether the left or right hemisphere is removed. Although individuals with only one hemisphere acquire both verbal and nonverbal skills, the degree to which these skills are acquired varies by the hemisphere removed. Individuals with only a right hemisphere perform, on average, better on spatial tasks than those with only a left hemisphere, whereas individuals with only a left hemisphere perform better on verbal tasks than those with only a right hemisphere. Third, asymmetries can be observed in the newborn. These include motoric asymmetries, behavioral asymmetries, and asymmetries in brain responses. For example, a larger evoked response is recorded over the left hemisphere to verbal materials and over the right hemisphere to nonverbal materials.

Thus, it appears that the basic blueprint of hemispheric specialization exists at birth. However, the nature of that blueprint can be altered by environmental factors. For example, it has been well documented that after damage to the left hemisphere during approximately the first 2 years of life, the right hemisphere can acquire the ability to control speech. Somewhat more ambiguous are findings that experience can influence the degree of perceptual asymmetry that is observed. For example, a greater left visual field advantage is observed for holistic processing of faces from a racial group with which one is familiar than for faces of a group with which one is not familiar. Whether this result indicates that the nature of hemispheric specialization is changed by experience, or whether more experience tends to engage more specialized processors in the brain, remains unclear.

The Effect of Handedness

One individual difference, that of handedness, has been clearly linked to patterns of hemispheric specialization for cognitive and emotional function. The relative specializations of the cerebral hemispheres described previously appear to hold only for individuals who are right-handed. In contrast, left-handed individuals, who comprise approximately 10% of individuals worldwide, can have a diverse pattern of lateral organization. This difference is well-known by neurologists and neuropsychologists who have long observed that handedness is an important factor in predicting the types of deficits and the amount of recovery that is likely to be observed after unilateral brain damage. However, there is much variability among left-handers regarding the type of lateralized brain organization they display. In some cases, it is similar to that of right-handers, with the left hemisphere specialized for verbal functions and the right for nonverbal functions. In other cases, it is the opposite, with the right hemisphere specialized for verbal function and the left for nonverbal function. It is estimated that such a brain organization is found in only 1% or 2% of right-handers. Finally, other left-handers exhibit a pattern in which both hemispheres appear to be able to process both verbal and nonverbal information, including speech output. Although much research has attempted to isolate a

factor that can predict the type of brain organization a given left-hander will exhibit, for the most part, these efforts have failed. Currently, the leading model of the genetic basis of handedness and its relationship to lateralized brain organization suggests that handedness (and hence brain organization) is randomly distributed, unless one inherits a right-shift allele, which shifts handedness to the right hand and language to the left hemisphere. Such a model assumes that it was evolutionarily advantageous to have motor control of the right hand and control of language co-lateralized to the same hemisphere.

Interhemispheric Integration

After the explosion of research examining hemispheric specialization in the 1970s and 1980s, recent work has examined how the processing of the hemispheres is coordinated and the effects of integration of information between the cerebral hemispheres.

Interhemispheric integration occurs mainly via the massive bundle of nerve fibers connecting the cerebral hemispheres, the corpus callosum. Because information from the peripheral nerves is directed contralaterally, at least for vision, motor, and somatosensory information, integration of information across the hemispheres allows information from each sensory half-field to be bound together. For example, in the visual domain, information is represented contralaterally in V1, V2, and V3a/VP. In fact, there are few, if any callosal connections in BA 17. Representation of information from the ipsilateral visual field first appears at the level of V3a and V4v, which are brain regions whose cells have receptive fields that span the midline. This critical role of the callosum in fusing the sensory worlds is made apparent by split-brain patients who cannot bind information from different sensory fields. For example, if fixated on a central point and shown two items, one on each side visual midline so that they are projected to opposite hemispheres, a split-brain patient cannot determine if they are the same or different.

Although the callosum acts as a mechanism to transfer information between the hemispheres, this transfer involves time costs as well as degradation of information. Transfer of information via the corpus callosum takes approximately 5 ms via the large myelinated fibers that connect sensory and motor areas and 20–50 ms for small unmylineated fibers. The fidelity costs are apparent in old/new memory paradigms. If an item is presented in one visual field, correctly recognizing of that item as previously viewed is poorer if presented to the hemisphere that did not initially view the item as compared to the hemisphere. Such findings are consistent with theorizing that the representations supported by each of the hemispheres are different and are not completely interchangeable.

Integration of information across the cerebral hemispheres plays an additional role above and beyond that of binding together the visual world – that of attentional control. Evidence suggests that the cerebral hemispheres dynamically couple and decouple to meet task demands. When tasks involve relatively low attentional demands, performance is better when all the critical information needed for a decision (e.g., are these two items physically identical?) is directed to a single hemisphere compared to divided between them. It has been proposed that under such conditions, the processing capacity of a single hemisphere is adequate to meet the demands and the cost of interhemispheric communication takes a toll on performance. In contrast, when attentional demands are high, dividing processing between the hemispheres enables more resources to be brought to bear and the advantage provided by these additional resources more than outweighs the cost of callosal transfer. Such an outcome can occur because, as described previously, the specialization of the hemispheres is not absolute but relative, allowing both hemispheres to process almost all types of information. Thus, the additional resources provided by the partner hemisphere, even if not specialized for the task, can nonetheless have a major impact on performance. These findings are consonant with data from individuals in whom the corpus callosum is severed or damaged, such as occurs in multiple sclerosis. A common consequence of callosal damage or insufficiency is problems in attentional control. Although the neural mechanisms that allow the callosum to play such a role remain somewhat obscure, connectionist modeling suggests that the advantage afforded by dividing processing between the hemispheres is an emergent phenomenon of having two somewhat insulated and distinct processors.

Summary

Each of the human cerebral hemispheres has a distinct manner of processing information, with the left hemisphere more adept at attending to and processing more fine-grained information and the right hemisphere more adept at attending to and processing more coarse-grained information. This distinction holds for all types of information, whether verbal or spatial. The complimentarity of these modes of processing means that the human brain contains two processors, each of which provides a unique manner of understanding and interpreting the world.

Moreover, this dichotomy and isolation of processing also provides for a system that can dynamically reconfigure to act in isolation or in tandem depending on attentional demands.

See also: Brain Asymmetry: Evolution; Cognition: An Overview of Neuroimaging Techniques; Dichotic Listening Studies of Brain Asymmetry; Emotional Hormones and Memory Modulation; Memory: Genetic Approaches; Multisensory Convergence and Integration; Split-Brain Patients.

Further Reading

Annett M (2002) *Handedness and Brain Asymmetry: The Right Shift Theory.* New York: Psychology Press.

Banich MT (1998) The missing link: The role of interhemispheric interaction in attentional processing. *Brain and Cognition* 36: 128–157.

Beeman MJ and Chiarello C (1998) Complementary right- and left-hemisphere language comprehension. *Current Directions in Psychological Science* 7: 2–8.

Corballis MC (2003) From mouth to hand: Gesture, speech, and the evolution of right-handedness. *Behavioral and Brain Sciences* 26: 199–260.

Davidson RJ (1992) Anterior cerebral asymmetry and the nature of emotion. *Brain & Cognition* 20: 125–151.

Hellige JB (2001) *Hemispheric Asymmetry: What's Right and What's Left.* Cambridge, MA: Harvard University Press.

Hugdahl K and Davidson RJ (eds.) (2003) *The Asymmetrical Brain.* Cambridge: MIT Press.

Ivry RB and Robertson LC (1998) *The Two Sides of Perception.* Cambridge: MIT Press.

Keller J, Nitschke JB, Bhargava T, et al. (2000) Neuropsychological differentiation of depression and anxiety. *Journal of Abnormal Psychology* 109: 3–10.

Kosslyn SM (1987) Seeing and imagining in the cerebral hemispheres: A computational approach. *Psychological Review* 94: 148–175.

Levy J, Trevarthen C, and Sperry RW (1972) Perception of bilateral chimeric figures following hemispheric deconnexion. *Brain* 95: 61–78.

Monaghan P and Pollmann S (2003) Division of labor between the hemispheres for complex but not simple tasks: An implemented connectionist model. *Journal of Experimental Psychology: General* 132: 379–399.

Sergent J (1985) Influence of task and input factors on hemispheric involvement in face processing. *Journal of Experimental Psychology: Human Perception and Performance* 11: 846–861.

Trevarthen C (1996) Lateral asymmetries in infancy: Implications for the development of the hemispheres. *Neuroscience & Biobehavioral Reviews* 20: 571–586.

Zaidel E and Iacoboni M (eds.) (2003) *The Parallel Brain: The Cognitive Neuroscience of the Corpus Callosum.* Cambridge: MIT Press.

Hepatic Encephalopathy

A S Basile, DOV Pharmaceutical, Inc., Somerset, NJ, USA
K D Mullen, Case Western Reserve University, Cleveland, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Description

Hepatic encephalopathy (HE) is a complex neuropsychiatric syndrome characterized by the global depression of central nervous system (CNS) activity, with a gradual deterioration of higher mental and motor functions (Table 1). The syndrome is associated either with acute liver failure, commonly precipitated by drug overdose or viral hepatitis, or with chronic liver failure, frequently caused by alcoholic cirrhosis. Liver disease is the fourth largest cause of death in Americans between 45 and 54 years of age, and HE (including minimal HE) is associated with 50–70% of all patients with cirrhosis. Moreover, HE is more prevalent in nations with higher rates of liver disease than occur in the United States.

Although HE has been recognized since the beginning of recorded medical history, the precise pathogenic mechanisms responsible for the syndrome remain complex and poorly defined. Thus, the modalities for treating HE are still evolving. A profusion of metabolic abnormalities and gut-derived toxins result from hepatic failure, resulting in significant alterations in neurotransmission and CNS energy metabolism, superimposed upon a characteristic astrocyte pathology (Alzheimer type II degeneration). Thus, it is difficult to ascribe the neurological and psychiatric manifestations of HE to the presence of a single agent or to changes in a particular metabolic pathway. However, since HE is a metabolic encephalopathy, with few neuroanatomical abnormalities observed in patients who die of either acute or chronic liver failure, it has the potential to be completely reversible.

The neurological manifestations of HE (Table 1) are determined by two principal components of the underlying liver disease: hepatocellular failure and systemic shunting of portal venous contents (Figure 1). Normally, potentially neurotoxic substances are absorbed from the gut and detoxified by the liver. However, in hepatocellular failure these agents may avoid catabolism by passing through the diseased liver unmetabolized and into the systemic circulation (portal–systemic shunting). Alternatively, these toxins may bypass the liver altogether and enter the systemic circulation through portal-venous shunts that are either surgically constructed, or formed spontaneously as a result of portal hypertension associated with cirrhosis. Once these gut-derived substances reach the brain, they modify CNS function either by altering metabolic processes such as oxidative metabolism or neurotransmitter synthesis, or by interacting directly with neurotransmitter receptors.

Hyperammonemia and HE

Ammonia is produced systemically by deamination of amino acids, in the gut wall by the action of glutaminase, and by degradation of amines, amino acids, and urea by gut bacteria. Ammonia diffuses into the portal circulation, where it would normally be converted to glutamine and urea via the Krebs–Henseleit cycle. However, this process is impaired in liver failure, resulting in elevated concentrations of ammonia in the blood. Ammonia in the blood readily diffuses across the blood–brain barrier in its uncharged form. Levels of ammonia in the cerebrospinal fluid (CSF) of patients and animals with clinically significant manifestations of encephalopathy secondary to liver failure range from 70 to 850 μM, compared to concentrations <50 μM in normal humans. Brain ammonia concentrations may reach as high as 1–5 mM in animal models of acute liver failure. While total ammonia levels in the blood are not well correlated with the severity of HE, brain ammonia uptake, as indicated by glutamine levels in CSF and by ^{1}H magnetic resonance spectroscopy, is well correlated with the severity of HE associated with cirrhosis.

Acute exposure to elevated levels of ammonia causes a number of biochemical and electrophysiological derangements in the CNS, including suppression of oxidative metabolism, inhibition of Na^{+}/K^{+}-ATPase activity, inactivation of Cl^{-} extrusion pumps, and decreases in the equilibrium potential of the neuronal inhibitory postsynaptic potential (IPSP). More recent investigations indicate a significant impact of ammonia on astrocytic function. In the presence of ammonia, astrocytes swell (Figure 2), contributing to cerebral edema in acute hepatic failure, while persistent elevations of brain ammonia levels, as observed in chronic hepatic failure, are associated with the development of Alzheimer type II astrocytosis. In the absence of a functional hepatic mechanism for converting ammonia to urea, astrocytes remove ammonia by converting it to glutamine. Exposing astrocytes *in vitro* to elevated concentrations of either ammonia or glutamine results in increased free radical production, possibly by activation of the mitochondrial permeability transition.

Table 1 Clinical stages of HE

Stage	*Mental state*	*Neuromuscular state*
I	Mild confusion, euphoria, depression, decreased attention, slowed analytical ability, irritability, sleep inversion	Mild incoordination, impaired handwriting
II	Drowsiness, lethargy, gross deficits in analytical ability, obvious personality changes, inappropriate behavior, intermittent disorientation; electroencephalogram abnormalities (high-amplitude, low-frequency waves with nonfocal changes)	Asterixis, ataxia, dysarthria, paratonia, apraxia
III	Somnolent but rousable, unable to perform analytical tasks, disorientation with respect to time and/or place, amnesia, rage, slurred speech	Hyperreflexia, muscle spasticity, Babinski's sign present, seizures (rare)
IV_A	Coma	Thalamic–subcortical spasticity of entire body; no response to painful stimuli
IV_B	Deep coma	Death imminent

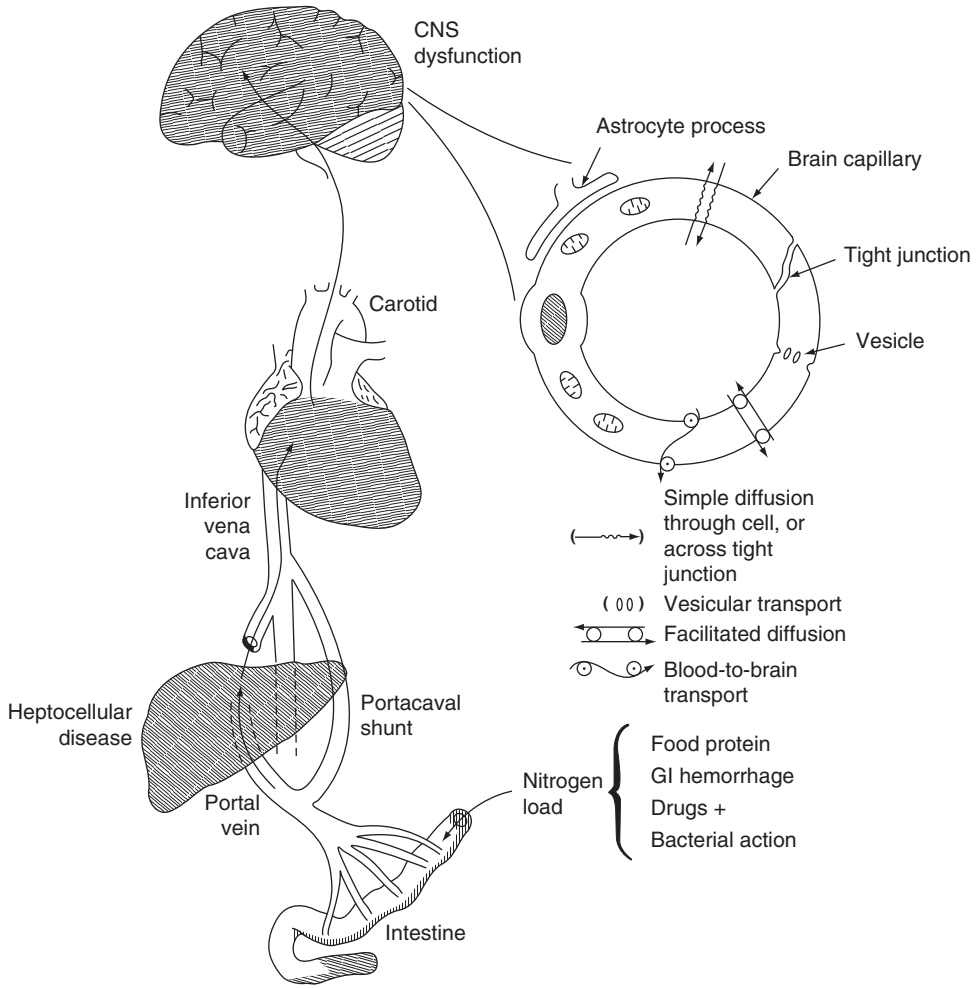

Figure 1 The concept of portal–systemic shunting as it contributes to the development of HE. Nitrogenous substances arising from the gastrointestinal (GI) system normally are absorbed into the portal venous system and delivered to the liver, where they are extracted and metabolized. However, in liver disease these substances may pass directly into the systemic circulation or bypass the liver completely via portal–venous shunts. Once these compounds enter the systemic circulation, they may gain entry to the central nervous system (CNS) by a variety of mechanisms, including high-capacity active transport or facilitated diffusion, resulting in abnormal brain function. Reprinted from Basile AS, Jones EA, Skolnick P, et al. (1991) The pathogenesis and treatment of hepatic encephalopathy: Evidence for the involvement of benzodiazepine receptor ligands. *Pharmacological Reviews* 43: 27–71, with permission of the American Society for Pharmacology and Experimental Therapeutics.

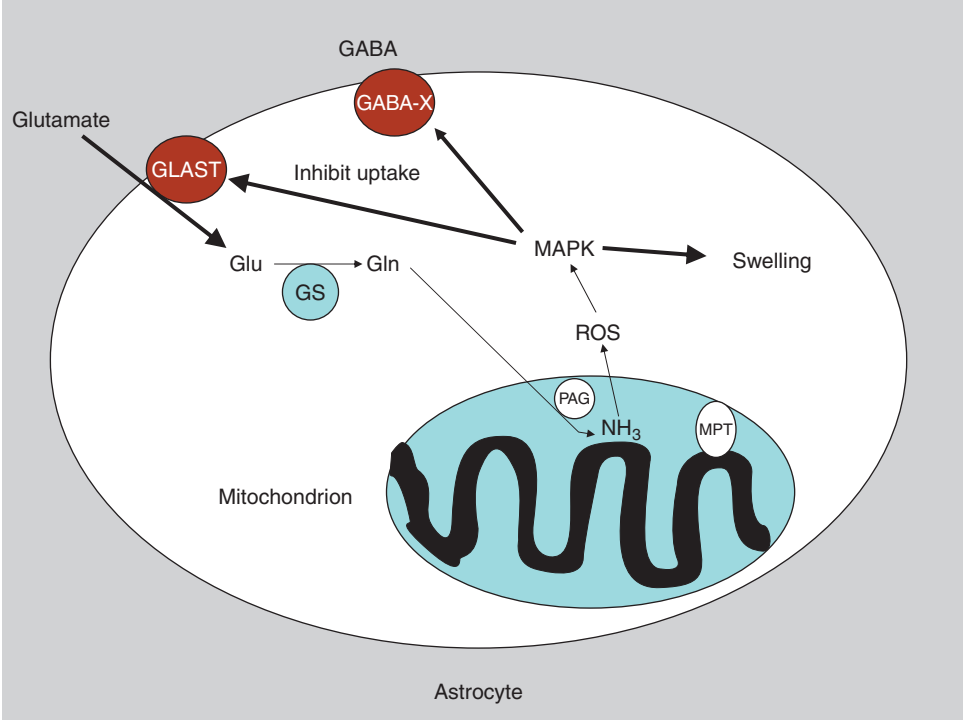

Figure 2 Alterations in astrocytic function induced by hyperammonemia. Alzheimer type II astrocytosis and cerebral edema are often observed in acute liver failure and contribute to the end stages of hepatic encephalopthy. Exposing astrocytes in culture to elevated concentrations (5 mM) of ammonia (NH_3) for 1–3 days causes them to swell. Astrocytes will try to reduce ammonia levels by fixing ammonia to glutamate (Glu) using glutamine synthase (GS), yielding glutamine (Gln). However, ammonia may be released from glutamine by mitochondrial phosphate-activated glutaminase (PAG), causing the production of reactive oxygen species (ROS) by the mitochondria and induction of the mitochondrial permeability transition (MPT). Elevated ROS levels increase the phosphorylation of a number of mitogen-activated protein kinases (MAPK), such as ERK-1/2, $p39^{MAPK}$, and JNK-1/2/3. Treating astrocytes with appropriate MAPK inhibitors or antioxidants will prevent the astrocytic swelling, glutamate uptake inhibition, and decline in glutamate–aspartate transporter (GLAST) mRNA and protein expression. Suppression of GABA transport (GABA-X), which is reported in experimental hyperammonemic conditions, may also be influenced by these processes and contribute to some of the manifestations of hepatic encephalopathy.

The resulting increase in oxidative stress causes phosphorylation of extracellular signal-regulated kinase-1/2 (ERK-1/2), p38 mitogen-activated protein kinase ($p38^{MAPK}$), and c-Jun N-terminal kinase-1/2/3 (JNK-1/2/3). Ammonia-induced astrocytic swelling, which can be blocked *in vitro* by antioxidants and MAPK inhibitors, may have a number of secondary effects on neuronal function and neurotransmitter levels, as detailed in following sections.

Other Neurotoxins

Hyperammonemia plays a significant role in the pathogenesis of HE, but other conditions/agents are present in hepatic failure/portal–systemic shunting that can alter CNS function. Early work attempting to identify and characterize the potential neurotoxins involved in HE focused on the following classes of agents: mercaptans, fatty acids, imbalances in the ratio of branched-chain to aromatic amino acids, and catecholamines/false neurotransmitters. The relevance of these agents to the pathogenesis of HE is now mostly historical. Mercaptans, phenol, and medium-chain-length fatty acids are present in elevated concentrations in patients with liver disease. However, the neurotoxic concentrations of these substances are 8–15 times higher than the levels commonly observed in patients with HE. While it has been proposed that the toxicity of these agents is enhanced in the presence of ammonia, the concept of 'synergistic neurotoxicity' clearly applies to the effects of all of the metabolic abnormalities and toxic compounds that arise in liver failure, and is not unique to these agents.

CNS levels of aromatic amino acids (e.g., tryptophan, phenylalanine, and tyrosine) are increased between 70% and 300% in patients with cirrhosis. These amino acids are precursors to several monoaminergic neurotransmitters, including serotonin (Table 2). The rate of serotonin turnover in discrete brain regions

Table 2 CNS levels of selected neurotransmitters in HE[a]

Neurotransmitter	*Control*	*Liver failure*	*Liver failure with HE*
Dopamine[b]	2698 ng g^{-1} caudate	2852	4774
Noradrenaline[b]	92 ng g^{-1} caudate	94	154
Serotonin[c]	0.281 ng mg^{-1} protein		1.397
Glutamate[d]	9.54 μmol g^{-1} brain		7.59**
γ-Aminobutyric acid	1.46 μmol g^{-1} brain[d]		1.21
	0.38 μmol l^{-1} plasma[e]		3.37**
Benzodiazepine receptor ligands[f]	78 ng DZ equiv g^{-1} brain		405*

[a]Values represent mean concentrations per unit of brain obtained during autopsy, or plasma from patients with acute or chronic liver failure, with or without HE, or from patients with diseases other than liver failure (controls). DZ equiv., diazepam equivalents. *, **, $P < 0.05$, 0.01 vs. control, respectively.
[b]Levels measured in samples of frontal cortex (Cuilleret et al., 1980).
[c]Levels measured in samples of caudate nucleus (Bergeron et al., 1989).
[d]Levels measured in samples of prefrontal cortex (Lavoie et al., 1987).
[e]Median values of GABA levels measured in plasma samples from controls and patients with acute liver failure, stage IV HE (Levy and Losowsky, 1989).
[f]Levels measured in samples of frontal cortex from patients with acute liver failure (Basile et al., 1991).

increases in liver failure, possibly as a result of increased concentrations of tryptophan associated with liver failure. However, serotonin may play only a minor role in the development of HE (possibly in the development of sleep inversion), since there is no correlation between the severity of HE and the concentration of serotonin or its metabolites. Alternatively, it has been proposed that the elevations in CNS tyrosine and phenylalanine concentrations could enhance the synthesis of 'false' neurotransmitters, such as octopamine or tyramine, which may serve to deplete catecholamine neurotransmitters. However, the levels of the false neurotransmitter octopamine decrease in the brains of patients with HE, while noradrenaline and dopamine levels are generally unchanged or increased (Table 2).

Magnetic resonance imaging (MRI) of some cirrhotic patients reveals hyperintense T_1 signals in the globus pallidus, suggestive of manganese accumulation. Subsequent analysis of necropsy samples from cirrhotics indicated a 128% increase in manganese levels. Manganese is neurotoxic and causes Alzheimer type II astrocytic changes. However, the degree of pallidal hyperintensity resulting from liver failure does not, in itself, appear to be associated with the severity of HE.

Enhancement of GABAergic Neurotransmission and HE

γ-Aminobutyric acid (GABA) is the primary inhibitory neurotransmitter in the CNS, and overactivation of this pathway results in behavioral changes consistent with the primary manifestations of HE, specifically motor dysfunction, cognitive deficits, and altered consciousness. No consistent changes in the $GABA_A$ receptor complex are observed in animal models of HE or in humans with liver failure. While plasma GABA levels are reportedly increased in liver failure (Table 2), they may not impact basal levels of free GABA in the CNS. Nonetheless, there is evidence that synaptic concentrations of GABA may be increased, possibly as a result of ammonia-induced decreases in glial reuptake (Figures 2 and 3) or alterations in release parameters.

GABAergic neurotransmission may also be enhanced by the presence of endogenous, positive modulators of the $GABA_A$ receptor, including agonists of the benzodiazepine-binding site on the $GABA_A$ receptor complex and neurosteroids. Benzodiazepine binding site agonists are increased in several animal models of the syndrome and in humans suffering from HE due to either acute or chronic liver failure (Table 2, Figure 3). Tissues and body fluids from humans and animals with HE, who were otherwise never exposed to 1,4-substituted benzodiazepines, contain numerous benzodiazepine binding site ligands, as determined by radiometric techniques. Only two of these were chemically and immunochemically identified as 1,4-benzodiazepines. However, these compounds comprised less than 20% of the total concentration of benzodiazepine binding site ligands present in HE in both animal models and humans (Table 2, Figure 3). The identification and source of the remaining benzodiazepine binding-site ligands is unclear, but metabolites of hemoglobin, including hemin and protoporphyrin IX, are possible candidates.

Alterations in astrocytic function induced by ammonia may increase the production of neurosteroids such as allopregnanolone, a potent inhibitor of neuronal activity acting through the $GABA_A$ receptor.

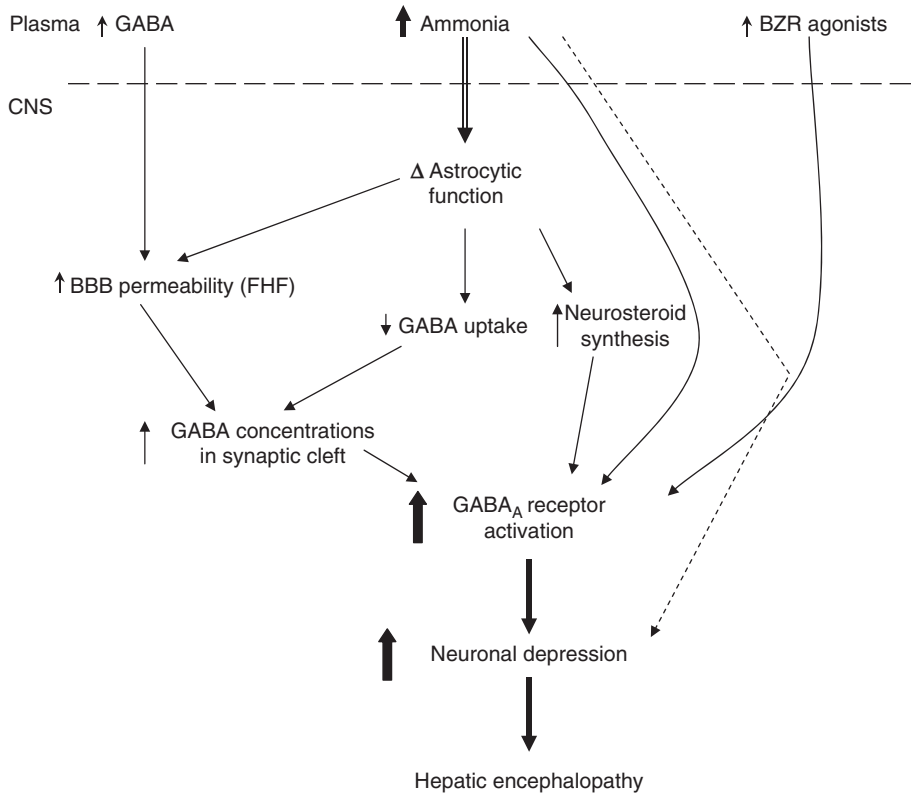

Figure 3 GABAergic neurotransmission is enhanced in hepatic failure through a variety of mechanisms. γ-Aminobutyric acid (GABA), benzodiazepine receptor (BZR) agonists, and particularly ammonia, originating from outside of the brain (top tier), make their way into the central nervous system (CNS) either through diffusion or via a permeabilized blood–brain barrier (BBB). Elevations in CNS ammonia levels alter astrocytic function, which may lead to swelling and permeabilization of the BBB and fulminant hepatic failure (FHF), often observed in acute liver failure. Astrocytes in a hyperammonemic state also show a reduction in GABA uptake. The increased access of plasma GABA to the CNS, combined with decreased GABA uptake by astrocytes and evidence of enhanced GABA release, increases the availability of GABA at the synapse. In addition, the presence of positive modulators of the GABA$_A$ receptor, such as neurosteroid and benzodiazepine receptor agonists, as well as ammonia itself, yields a net increase in GABAergic neurotransmission. These changes, in combination with the suppression of oxidative metabolism of neurons by elevated ammonia levels, cause neuronal depression, which is expressed as the core manifestation of HE (see **Table 1**).

The peripheral-type benzodiazepine receptor (PBR), located on glial mitochondria, plays a role in neurosteroid synthesis. The density of PBRs is increased in animal models of HE, and in postmortem samples from humans with hepatic failure, but interestingly is not observed using positron emission tomography (PET) imaging in patients with HE. Nonetheless, allopregnanolone has been measured in postmortem samples of brain from patients with HE in concentrations sufficient to enhance GABA-gated chloride currents, which may serve to depress neuronal activity.

Evidence also indicates that ammonia can enhance GABAergic neurotransmission through direct actions at the GABA$_A$ receptor. An electrophysiological study using isolated cortical neurons indicated that the maximal level of GABA-activated current is enhanced by pathophysiologically relevant concentrations of ammonia, as is the potency of GABA. Subsequent investigations of the effects of ammonia on radioligand binding to sites on the GABA$_A$ receptor confirmed and extended these observations. Pathophysiologically relevant concentrations of ammonia selectively increased the density of agonist (flunitrazepam) ligand binding to the GABA-binding site of the GABA$_A$ receptor complex, with no effect on antagonist (flumazenil) ligand binding. Further, ammonia increased the affinity of agonist binding to the benzodiazepine-binding site on the GABA$_A$ receptor complex, without altering antagonist binding. The latter effect was further enhanced (threefold) in the presence of GABA. It appears that the ability of ammonia to modulate ligand binding to the GABA$_A$ receptors (a heterooligomeric complex composed of members from five families of subunits) depends on the presence of the β subunit, as ammonia can enhance [^{3}H]flunitrazepam binding to recombinant GABA$_A$ receptors composed of $\alpha_1\beta_2\gamma_2$ subunits (the pharmacology of which most resembles that of the native GABA$_A$ receptor) but has no effect on binding

to $\alpha_1\gamma_2$ constructs. Taken together, the numerous neurochemical changes present in hepatic failure may enhance GABAergic neurotransmission, leading to the motoric and behavioral changes characteristic of HE.

Suppression of Glutamatergic Neurotransmission and HE

Contrasting with GABA, glutamate is the primary excitatory neurotransmitter. Accumulating evidence suggests that there are alterations in glutamatergic neurotransmission in acute liver failure. Total brain glutamate levels are decreased under hyperammonemic conditions, including liver failure, consistent with the conversion of glutamate to glutamine by astrocytes in order to reduce ammonia levels. Basal extracellular glutamate levels increase in models of acute hepatic failure, but not in the portal–systemic shunted rat model of chronic hepatic failure. In the latter model, stimulated release of glutamate is variably increased. Glial glutamate transporter (GLT-1 and/or GLAST) expression is transiently decreased in portal–systemic shunted rats and is more consistently decreased in rats with acute liver failure. Glutamate uptake by astrocytes *in vitro* is also reduced in the presence of ammonia. Thus, there is evidence for increased extracellular glutamate levels in acute liver failure, resulting from either increased release or decreased reuptake. There is no consistent evidence of a change in *N*-methyl-D-aspartate (NMDA) receptor density in animal models of acute liver failure.

The specific contribution(s) of the alterations in glutamatergic neurotransmission to the manifestations of HE is unclear. It has been proposed that increased glutamate levels may contribute to glial swelling and cerebral edema, but there is no correlation between the onset of glutamate elevations and increased intracranial pressure. Similarly, changes in glutamate levels may contribute to the sleep-inversion, extrapyramidal disorders and to the cognitive deficits observed in HE, particularly when associated with chronic liver failure. However, the evidence for alterations in extracellular glutamate levels in animal models of chronic liver failure is variable, and the role of glutamatergic neurotransmission in the aforementioned behavioral changes is weak. Nonetheless, the significance of glutamatergic neurotransmission to CNS function is so great as to warrant continued investigation of its role in HE.

Therapy

Therapeutic strategies for allaying the manifestations of HE may focus on the target organ, the CNS, and the gut. Current, standard therapeutic regimens are directed at reducing ammonia production and absorption from the gut. Restricting oral protein consumption in order to reduce a source of ammonia was a primary treatment for decades, but is now discouraged because it aggravates lean body wasting in cirrhosis. Another approach to reducing gut ammonia production from dietary sources is to prescribe less encephalopathogenic, vegetable-based protein diets. Nonabsorbable antibiotics and cathartics (lactulose and lactitol) have also been used to reduce ammonia production and absorption. In addition to its antibiotic actions, neomycin may reduce ammonia production by inhibiting intestinal glutaminase. Bacterial overgrowth in the small intestine and other disturbances of the microecology of indigenous gut flora are common in cirrhosis, and may impact toxin/ammonia production. The ammonia-lowering actions of lactulose/lactitol may be due in part to their effects on gut flora. However, despite their widespread use, controlled trials have indicated that lactulose/lactitol and neomycin therapies are not superior to placebo in the treatment of HE. The utility of these agents in the treatment of HE is currently being reassessed using randomized, placebo-controlled trials. Similarly, the efficacy of modulating intestinal flora, such as through the administration of *Enterobacter faecium SF 68*, in treating HE is receiving renewed interest.

Another mode of therapy is to enhance the excretion of waste nitrogen from the body in patients with liver failure. Sodium benzoate, phenylacetate, and L-ornithine-L-aspartate (LOLA) promote this process through a variety of mechanisms. LOLA may improve HE by lowering ammonia levels and suppressing ammonia-induced neurotoxicity by stimulating urea cycle function. Moreover, LOLA is more effective than placebo in clinical trials.

Other therapies for HE shown to be effective in placebo-controlled trials include acarbose and the centrally acting benzodiazepine receptor antagonist, flumazenil. There are a number of anecdotal reports of the efficacy of flumazenil in treating patients suffering from HE. It has been claimed that the reversal of coma by flumazenil in these reports reflects the abuse of benzodiazepines by patients with liver disease. However, these claims do not take into account that endogenous benzodiazepine receptor ligands have been found in multiple animal models of HE. Moreover, a number of placebo-controlled trials of the efficacy of flumazenil in controlled patient populations have been conducted. Typically, low doses (1–5 mg) of flumazenil reverse HE in 30% of patients within 0.5–5 min of administration. In a meta-analysis of 12 randomized trials, eight with cross-over design, flumazenil was associated with a significant improvement of HE (30%) compared to placebo (7%), particularly in a subpopulation of patients

with low-grade encephalopathy. While no follow-up therapy was employed in these studies, long-term suppression of HE has been reported with chronic oral administration of flumazenil (25 mg). Development of a long-acting, orally available form of a benzodiazepine antagonist should enhance the use of this therapy for HE.

Finally, can the evidence implicating ammonia in the enhancement of GABAergic neurotransmission be used to develop therapeutic modalities for treating HE? An agent that can reverse ammonia's enhancement of GABA and benzodiazepine receptor function with minimal convulsant activity would be desirable. Such an agent may be an antagonist of the neurosteroid-binding site on the $GABA_A$ receptor complex, such as pregnenolone. As previously noted, ammonia enhances the binding of [^{3}H]flunitrazepam. Performing these assays in the presence of pregnenolone sulfate completely abolished the ability of ammonia to enhance the binding of [^{3}H]flunitrazepam. Similar effects were seen on ligand binding to the GABA site on the $GABA_A$ receptor complex. These preliminary observations suggest a potential use for neurosteroid antagonists in reversing some of the psychomotor symptoms of HE. Interestingly, present therapeutic protocols aimed at reducing ammonia levels by restricting dietary intake and modifying colonies of enteric bacteria may also reduce the synthesis or uptake of benzodiazepine receptor ligands or their precursors.

Conclusion

In summary, multiple factors contribute to the pathogenesis of HE, including agents that impair oxidative metabolism or neurotransmitter synthesis in the CNS, or which interact directly with neurotransmitter receptors. The principal effect of this is to suppress neuronal activity, in part by enhancing GABAergic neurotransmission. The renewed emphasis on the role of abnormal CNS function in the manifestations of HE suggests that future therapeutic modalities for this disorder will focus upon the brain as well as the gut.

See also: GABA-A Receptors: Developmental Roles; Gamma-Aminobutyric Acid (GABA); Glutamatergic and Gabaergic Systems; Neural Stem Cells and CNS Diseases; Neurotoxins; Neurotoxins and their Neurotoxicology.

Further Reading

Als-Nielsen B, Kjaergard LL, and Gluud C (2004) Benzodiazepine receptor antagonists for acute and chronic hepatic encephalopathy. *Cochrane Database of Systematic Reviews* 2CD002798.

Bai G, Rama Rao KV, Murthy CRK, et al. (2001) Ammonia induces the mitochondrial permeability transition in primary cultures of rat astrocytes. *Journal of Neuroscience Research* 66: 981–991.

Barbaro G, DiLorenzo G, Soldini M, et al. (1998) Flumazenil for hepatic encephalopathy grade III and IVa in patients with cirrhosis: An Italian multicenter double-blind, placebo-controlled, cross-over study. *Hepatology* 28: 374–378.

Basile AS (2002) Direct and indirect enhancement of GABAergic neurotransmission by ammonia: Implications for the pathogenesis of hyperammonemic syndromes. *Neurochemistry International* 41: 115–122.

Basile AS, Hughes R, Harrison P, et al. (1991) Brain concentrations of 1,4 benzodiazepines are elevated in patients with fulminant hepatic failure. *New England Journal of Medicine* 325: 473–478.

Basile AS, Jones EA, and Skolnick P (1991) The pathogenesis and treatment of hepatic encephalopathy: Evidence for the involvement of benzodiazepine receptor ligands. *Pharmacological Reviews* 43: 27–71.

Bergeron M, Reader TA, Layrargues GP, et al. (1989) Monoamines and metabolites in autopsied brain tissue from cirrhotic patients with hepatic encephalopathy. *Neurochemical Research* 14: 853–859.

Conn HO and Burcher J (1989) *Hepatri Encephalopathy: Management with Lactulose and Related carbohydrates*. East lansing, MI: MediEd Press.

Cooper AJ and Plum F (1987) Biochemistry and physiology of brain ammonia. *Physiological Review* 67: 440–519

Cuilleret G, PomierLayrargues G, Pons F, et al. (1980) Changes in brain catecholamine levels in human cirrhotic hepatic encephalopathy. *Gut* 21: 565–569.

Jayakumar AR, Panickar KS, Murthy CRK, et al. (2006) Oxidative stress and mitogen-activated protein kinase phosphorylation mediate ammonia-induced cell swelling and glutamate uptake inhibition in cultured astrocytes. *Journal of Neuroscience* 26: 4774–4784.

Klos KJ, Ahlskog JE, Josephs KA, et al. (2005) Neurologic spectrum of chronic liver failure and basal ganglia T1 hyperintensity on magnetic resonance imaging: Probable manganese neurotoxicity. *Archives of Neurology* 62: 1385–1390.

Lavoie J, Giguere JF, Layrargues GP, et al. (1987) Amino-acid changes in autopsied brain tissue from cirrhotic patients with hepatic encephalopathy. *Journal of Neurochemistry* 49: 692–697.

Levy LJ and Losowsky MS (1989) Plasma gamma aminobutyric acid concentrations provide evidence of different mechanisms in the pathogenesis of hepatic encephalopathy in acute and chronic liver disease. *Hepatogastroenterology* 36: 494–498.

Norenberg MD (2001) Astrocytes and ammonia in hepatic encephalopathy. In: de Vellis J (ed.) *Neuroglia in the Aging Brain*, pp. 477–496. Totowa, NJ: Humana Press.

Olasmaa M, Rothstein JD, Guidotti A, et al. (1990) Endogenous benzodiazepine receptor ligands in human and animal hepatic encephalopathy. *Journal of Neurochemistry* 55: 2015–2023.

Rose C, Butterworth RF, Zayed J, et al. (1999) Manganese deposition in basal ganglia structures results from both portal-systemic shunting and liver dysfunction. *Gastroenterology* 117: 640–644.

Ruscito BJ and Harrison NL (2003) Hemoglobin metabolites mimic benzodiazepines and are possible mediators of hepatic encephalopathy. *Blood* 102: 1525–1528.

Scollo-Lavizzari G and Steinmann E (1985) Reversal of hepatic coma by benzodiazepine antagonist (Ro 15-1788). *Lancet* 1: 1324.

Herbal Products and GABA Receptors

G A R Johnston, M Chebib, R K Duke, S P Fernandez, J R Hanrahan, T Hinton, and K N Mewett, University of Sydney, Sydney, NSW, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

This article profiles ten examples of chemicals derived from herbal preparations that influence ionotropic receptors for the brain's major inhibitory neurotransmitter γ-aminobutyric acid (GABA). Many herbal medicines are used to influence brain function in order to treat anxiety, cognitive disorders, and insomnia. Such disorders are considered likely to involve GABA receptors, particularly ionotropic receptors that are responsible for fast synaptic transmission via ion channels. The evidence of efficacy, or otherwise, in the use of herbal medicines to treat central nervous system (CNS) disorders remains to be established unequivocally in many cases. There is no doubt, however, that these products contain chemicals that in their pure form can be shown to influence brain function in specific ways. The structural diversity of these chemicals provides valuable leads for the development of new therapeutic agents and tools with which to investigate nervous system function.

Herbal medicines are usually complex mixtures of chemicals. The combined actions of these chemicals may contribute to the overall effects of the medicines. Thus, it is important to study interactions among the individual chemicals. With respect to ionotropic GABA receptors, there are examples of synergistic interactions between two chemicals and examples of what has been termed second-order modulation, in which a second modulator enhances the action of a first-order modulator. This may represent a new form of drug action and lead to the identification of new drug targets. Investigating the complex ways in which these chemicals interact could lead to a better understanding of the function of ionotropic GABA receptors.

GABA itself is an important plant constituent, and thus it should not be surprising that plants contain a range of other chemicals that can influence GABA function. In addition, there are examples of plant-derived GABA mediating some form of communication among animals, bacteria, fungi, and plants. Thus, GABA may represent a conserved and ubiquitous biological signaling molecule.

Quality control of herbal medicines is a significant problem. Identification of key active ingredients in herbal medicines and the interactions among them may provide the means to better quality control using functional assays in addition to the currently used chemical assays. Many of the advances in studies of herbal products and GABA receptors have come from ligand binding studies, for example, from the binding of benzodiazepines to brain membranes and, more recently, from functional studies in which GABA receptors of known subunit composition are expressed in cells that do not usually express such receptors (e.g., frog oocytes).

Ionotropic GABA Receptors

The discovery that the plant alkaloid bicuculline could antagonize certain inhibitory actions of GABA in the CNS provided a vital pharmacological agent with which to probe GABA-mediated inhibition. This observation was a key component of the pharmacological basis for the original classification of GABA receptors into $GABA_A$ and $GABA_B$, with $GABA_A$ receptors being sensitive to bicuculline antagonism and insensitive to the $GABA_B$ agonist baclofen and $GABA_B$ receptors being insensitive to bicuculline and sensitive to baclofen. Subsequently, a third class of GABA receptors was described, $GABA_C$ receptors, that were insensitive to both bicuculline and baclofen. $GABA_A$ and $GABA_C$ receptors are ionotropic receptors, being ligand-gated ion channels, whereas $GABA_B$ receptors are metabotropic G-protein-coupled receptors.

Ionotropic receptors for the inhibitory neurotransmitter GABA are found on most, if not all, neurons in the CNS. They mediate fast neurotransmission via a central ion pore constituted by five surrounding protein subunits that, on activation by GABA, are permeant to chloride ions. They belong to the nicotinicoid superfamily of ligand-gated ion channels (also called cys-loop receptors) that includes nicotinic acetylcholine, strychnine-sensitive glycine, and serotonin 5-HT_3 receptors. The family of ionotropic GABA receptors may be divided into two subfamilies, $GABA_A$ and $GABA_C$ receptors, on the basis of their ability to form endogenous functional heteromeric and homomeric receptors, respectively, and of significant differences in their physiological and pharmacological properties.

The heteromeric $GABA_A$ receptors are made up of different protein subunits (e.g., a common makeup involves two $\alpha1$, two $\beta2$, and one $\gamma2$ subunits). There are 16 different subunits making up the $GABA_A$ receptor family in human brain: $\alpha1$–6, $\beta1$–3, $\gamma1$–3, δ, ε, π, and θ. In addition, there are splice variants of many of these subunits. Although the potential combinatorial diversity of $GABA_A$ receptors is huge,

studies of native $GABA_A$ receptors suggest that there may be fewer than 20 widely occurring $GABA_A$ receptor subtype combinations. There is less diversity in the homomeric $GABA_C$ receptors in that they are made up exclusively of either $\rho1$, $\rho2$, or $\rho3$ subunits, although pseudoheteromeric $GABA_C$ receptors have been described *in vitro*.

Enhancing the action of GABA on $GABA_A$ receptors is a key property of several classes of important therapeutic agents, including the benzodiazepines, barbiturates, and many general anesthetics. With advances in our understanding of the molecular diversity of $GABA_A$ receptors, there is an urgent need for the development of agents acting selectively on subtypes of these receptors.

Agents that enhance the action of GABA on $GABA_A$ receptors are known as positive modulators and are generally considered to involve action at allosteric sites on $GABA_A$ receptors remote from the GABA recognition sites (orthosteric sites). Such allosteric sites are highly valued as targets for the development of subtype-specific drugs because there is generally greater diversity between receptor subtypes in amino acid sequence at allosteric sites than at orthosteric sites. Agents that reduce the action of GABA on $GABA_A$ receptors are known as negative allosteric modulators (once known as inverse agonists because they have the opposite actions to those of the classical benzodiazepines). Agents that block the actions of both positive and negative modulators are known as neutralizing allosteric modulators (e.g., the classical benzodiazepine antagonist flumazenil).

$GABA_C$ receptors are much less widely distributed than $GABA_A$ receptors. There is evidence for functional $GABA_C$ receptors in the retina, spinal cord, superior colliculus, pituitary, and gastrointestinal tract. Given the lower abundance and less widespread distribution of $GABA_C$ receptors compared to $GABA_A$ receptors, $GABA_C$ receptors may be a more selective drug target than $GABA_A$ receptors. The major indications for drugs acting on $GABA_C$ receptors are in the treatment of visual, sleep, and cognitive disorders. Agents acting on $GABA_C$ receptors may be useful for the treatment of myopia. A recent study has linked $GABA_C$ receptors to the possible treatment of Alzheimer's disease by providing evidence that the stimulation of $GABA_C$ receptors has a neuroprotective action against β-amyloid protein.

GABA Receptor Antagonists/Negative Modulators

Many of the chemicals currently used to study ionotropic GABA receptors, including the antagonists bicuculline and picrotoxin, are of plant origin. Bicuculline is used extensively to characterize $GABA_A$ receptor function, but has no known therapeutic use. Picrotoxin was used as an analeptic to treat barbiturate overdose and may have therapeutic use in the treatment of vertigo. Other plant-derived antagonists include bilobalide and the ginkgolides from *Ginkgo biloba*, extracts of which are used to treat mild cognitive impairment, and α-thujone, one of the active constituents of the psychotropic beverage absinthe prepared from *Artemisia absinthium*.

Bicuculline

Bicuculline (**Figure 1(a)**) is a phthalide isoquinoline alkaloid first isolated from the plant *Dicentra cucullaria* and subsequently from a variety of *Corydalis*, *Dicentra*, and *Adlumia* species. The discovery of the GABA antagonist action of bicuculline came from a systematic study of convulsant alkaloids following the discovery of the glycine antagonist action of strychnine. Extensive studies of convulsant alkaloids showed that, whereas many isoquinoline alkaloids are convulsants, most are glycine antagonists, with GABA antagonism being restricted to the phthalide isoquinoline alkaloids that have the 1S,9R configuration (i.e., bicuculline, corlumine and (+)-hydrastine).

Bicuculline is a competitive antagonist of $GABA_A$ receptor activation. There is substantial evidence that bicuculline interacts with the same binding sites as agonists. Point mutations of $\alpha1$ subunits of $GABA_A$ receptors alter both agonist and competitive antagonist properties, suggesting a close structural association of $\alpha1$ Phe64 with agonist/antagonist binding sites.

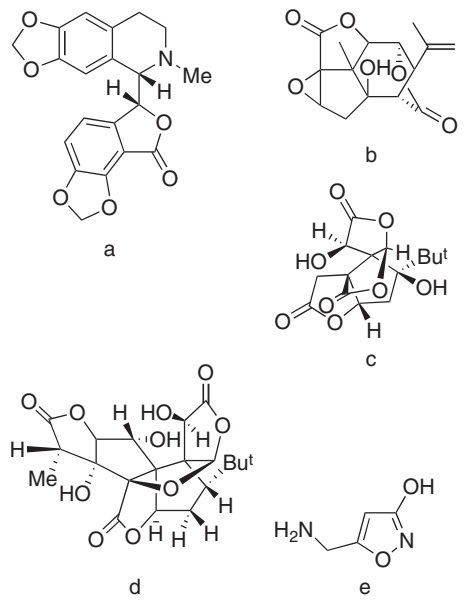

Figure 1 Plant-derived agonists and antagonists of ionotropic GABA receptors: (a) bicuculline; (b) picrotoxinin; (c) bilobalide; (d) ginkgolide B; (e) muscimol.

Substitution of this Phe by Leu results in a large decrease in the apparent affinity for both GABA and bicuculline.

Quaternary salts of bicuculline, such as bicuculline methochloride and bicuculline methiodide, are easier to use than the ternary salts such as the hydrochloride in that they are more water soluble and more stable; however, the lactone ring can still open in these derivatives, rendering them inactive as GABA antagonists. Unlike bicuculline, the quaternary salts do not pass the blood–brain barrier on systemic administration. The quaternary salts are more potent inhibitors of acetylcholinesterase than the hydrochlorides. Other actions of bicuculline and its derivatives include effects on 5-HT_3, nicotinic, and *N*-methyl-D-aspartate (NMDA) receptors, as well as voltage-activated potassium channels.

Picrotoxin

Picrotoxin is an equimolar mixture of picrotoxinin and picrotin isolated from *Anamirta cocculus* and related poisonous plants of the moonseed family. Picrotoxinin (**Figure 1(b)**) is a relatively potent convulsant and a noncompetitive $GABA_A$ receptor antagonist, whereas picrotin is some 50 times less active than picrotoxinin. Picrotoxinin is one of a number of structurally related convulsants of plant origin, including coriamyrtin and tutin, that act as $GABA_A$ receptor antagonists. Most of the development of picrotoxinin-related compounds has been directed toward the discovery of new insecticides, exploiting differences between insect and mammalian ionotropic GABA receptors.

Picrotoxinin is relatively nonspecific in that it is a potent antagonist at $GABA_A$ and $GABA_C$, moderate at glycine, and weak at $5HT_3$ receptors. Interestingly, picrotin is equipotent with picrotoxinin as an antagonist of glycine receptors, indicating substantial differences in the picrotoxinin binding sites of ionotropic GABA and glycine receptors.

Unlike bicuculline, picrotoxinin can act as a GABA antagonist when administered intracellularly, probably by entering open chloride channels. It is clear that bicuculline and picrotoxinin act at different sites to antagonize GABA. The actual mechanisms by which picrotoxinin blocks ionotropic GABA receptors are complex. Site directed mutagenesis studies show that picrotoxinin acts at sites within the chloride channels of $GABA_A$, $GABA_C$, and glycine receptors.

Because the antagonist action of picrotoxinin is directed toward the GABA-activated chloride channels rather than the GABA recognition sites on ionotropic GABA receptors, picrotoxinin should perhaps be classified more correctly as a negative allosteric modulator, even though it is traditionally described as an antagonist.

Bilobalide and Ginkgolides

Bilobalide (**Figure 1(c)**), a sesquiterpenoid lactone from *Ginkgo biloba* that bears some structural similarities to picrotoxinin, including a lipophilic side chain and a hydrophilic cage, acts as an antagonist at $GABA_A$ and $GABA_C$ receptors. Like picrotoxinin, bilobalide appears to act at sites in the chloride channels of ionotropic GABA receptors and is thus more appropriately considered as a negative allosteric modulator.

There is, however, a major difference between the effects of bilobalide and picrotoxinin *in vivo*. Whereas picrotoxinin is a convulsant, bilobalide is an anticonvulsant. The lack of convulsant action in an agent that reduces GABA action may be important for the enhancement of cognition. The lack of convulsant action of bilobalide may result from inhibition of glutamate release, thus reducing synaptic excitation and outweighing reductions in synaptic inhibition. Bilobalide has a neuroprotective action in a variety of animal models.

The structurally related ginkgolides, especially ginkgolide B (**Figure 1(d)**), also act as negative modulators at $GABA_A$ receptors. They also inhibit strychnine-sensitive glycine receptors and platelet-activating factor. Bilobalide and the ginkgolides reduce barbiturate-induced sleeping time in mice, an effect perhaps relevant to the clinically observed vigilance-enhancing and antidepressant-like actions of *Ginkgo* extracts.

There are extensive, well-controlled clinical trials on *Ginkgo* extracts, particularly the standardized leaf extract code named EGb761, for the treatment of cognitive and vascular disorders. This extract contains 24% flavonoids and 6% terpenoids (including bilobalide and ginkgolides). It exerts multiple effects including reducing apoptosis, lipid peroxidation, inflammation, and β-amyloid aggregation, in addition to effects on GABA receptors and glutamate release. The cognition-enhancing effects of *Ginkgo* extracts may be partly mediated by bilobalide acting to enhance hippocampal pyridamidal neuronal excitability.

GABA Receptor Agonists

Two well-known agonists at ionotropic GABA receptors are found in plants and fungi: GABA itself, which occurs in all plants, and muscimol, a psychoactive agent from the mushroom *Amanita muscaria*.

GABA

As an important plant constituent, GABA is part of our regular diet. In recent years, chocolate, tea, wine, and yoghurt rich in GABA have been developed and promoted as being beneficial to health. Such products are reported to reduce blood pressure and increase mental alertness. Pure GABA is promoted for body building, acting via human growth hormone to increase muscle and reduce fat.

Because GABA is not considered to pass the blood–brain barrier on systemic administration, dietary GABA may be acting on GABA systems in the periphery. $GABA_A$ and $GABA_B$ receptors have been described on peripheral smooth muscle, cerebral blood vessels, and the sinus node of the heart. Alternate transcripts of the ε subunit of $GABA_A$ receptors have been found in human and rat heart; the effects of GABA and related compounds on blood pressure may be mediated in part through $GABA_A$ receptors in the heart containing ε subunits that have not been studied extensively to date. Bicuculline methiodide, a quaternary salt of bicuculline that does not cross the blood–brain barrier, increases mean arterial pressure and heart rate in rats on intravenous administration, thus providing functional evidence for bicuculline-sensitive GABA receptors outside the CNS that influence blood pressure.

Plant-derived GABA is a confounding factor in the bioactivity-guided fractionation of extracts of herbal preparations using ligand binding or functional assays of GABA receptors.

Muscimol

Muscimol (**Figure 1(e)**) is one of the most widely used agonists in the investigation of ionotropic GABA receptors. It is a more potent agonist at $GABA_C$ receptors than at $GABA_A$ receptors. The agonist action of muscimol at $GABA_C$ receptors is not blocked by bicuculline but is sensitive to picrotoxinin.

Muscimol is a conformationally restricted analog of GABA in which a hydroxyisoxazole moiety replaces the carboxyl group of GABA. The 3-hydroxyisoxazole is recognized as a carboxyl group equivalent by $GABA_A$ and $GABA_C$ receptors but not by $GABA_B$ receptors. The neuronal GABA uptake system recognizes the 3-hydroxyisoxazole moiety, in that muscimol is a weak inhibitor of GABA uptake but is neither an inhibitor of nor a substrate for GABA aminotransferase, indicating that this enzyme does not interact with the 3-hydroxyisoxazole moiety.

Muscimol became a prototype substance for the design and development of a range of isoxazoles with varying activities on GABA systems. THIP (Gaboxadol) is a conformationally restricted analog of muscimol with analgesic and sleep-promoting properties. It was investigated for the treatment of insomnia, but was recently withdrawn from phase III clinical trials due to efficacy and side-effect problems. THIP shows selectivity for $GABA_A$ receptors that contain a δ subunit, at which it is a more efficacious agonist than GABA. At similar concentrations, THIP is a $GABA_C$ receptor antagonist.

GABA Receptor Modulators

Flavonoids are responsible for many of the brilliant colors of fruits and vegetables and are important constituents of red wine, green tea, and many herbal preparations. More than 5000 different flavonoids have been identified. It has been estimated that the average daily intake of flavonoids is 1–2 g. Their antioxidant properties are regarded as important in our diet. They have a wide variety of biological activities and are being studied intensively as anticancer agents and for their effects on the vascular system.

Flavonoids have a range of activities on $GABA_A$ receptors and have been described as a new family of benzodiazepine receptor ligands. They were first linked to $GABA_A$ receptors when three isoflavans isolated from bovine urine were shown to inhibit diazepam binding to brain membranes. These isoflavans were most probably derived from plant sources in the bovine diet. Subsequently, many flavonoids directly isolated from plants were shown to influence benzodiazepine binding. A low-affinity benzodiazepine site has emerged as a possible target for the modulatory action of some flavonoids. This site is insensitive to the classical benzodiazepine antagonist flumazenil and has been described on a wide range of $GABA_A$ receptors, including those made up of only $\alpha1\beta2$ subunits. Flumazenil-insensitive effects of flavonoids on $GABA_A$ receptors have been extensively described.

Terpenoids are widespread in plants, especially in what are known as essential oils that can be extracted from plants. They have a wide range of uses from perfume constituents to paint thinners. Terpenoids are oxygenated products formally derived from C5 units and are classified by the number of C5 units in their structure. The sesquiterpenoid lactone picrotoxinin is the most widely used terpenoid in studies on ionotropic GABA receptors. However, a number of other terpenoids are of interest for their actions as positive modulators of $GABA_A$ receptors.

Another important class of herbal products that act as positive modulators of $GABA_A$ receptor function are α-pyrones, especially the kava lactones found in kava-kava (*Piper methysticum*). There are many random controlled trials of extracts of kava-kava that show a reduction in anxiety relative to placebo.

Apigenin

Apigenin (Figure 2(a)) is a common flavonoid found in a range of plants, including chamomile (*Matricaria recutita*). The traditional use of chamomile tea as a treatment for insomnia and anxiety led to investigations of its active constituents, including apigenin. Apigenin was found to have anxiolytic properties, and it competitively inhibited the binding of flunitrazepam to brain membranes without influencing the binding of muscimol to $GABA_A$ receptors.

Recent studies on human recombinant receptors in oocytes have shown that apigenin inhibited the activation of $GABA_A$ receptors in a flumazenil-insensitive manner and had a similar effect on $GABA_C$ receptors. Other studies on recombinant $GABA_A$ receptors also found an inhibitory effect of apigenin on GABA responses and, in addition, an enhancement of the diazepam-induced positive allosteric modulation of GABA responses by lower concentrations of apigenin (1 μM), described as a second-order modulation by apigenin of the first-order modulation by diazepam.

The flumazenil-sensitive anxiolytic effects of apigenin may be the result of apigenin enhancing a subthreshold effect of an endogenous benzodiazepine system. Evidence for physiologically relevant endozepines has come from the discovery of a mutant $GABA_A$ receptor in childhood absence epilepsy and febrile seizures that has diminished sensitivity to benzodiazepines with no other apparent alteration in function.

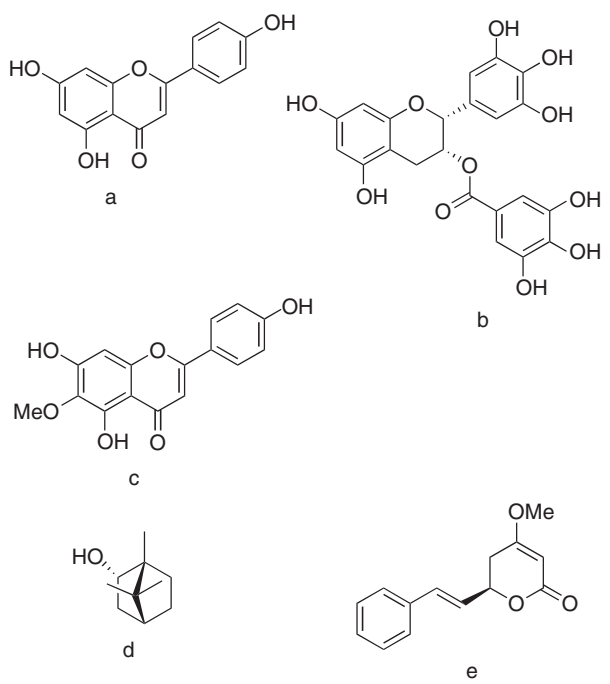

Figure 2 Plant-derived positive modulators of the activation of $GABA_A$ receptors by GABA: (a) apigenin; (b) epigallocatechin gallate; (c) hispidulin; (d) (+)-borneol; (e) kawain.

Overall, it seems that the effects of apigenin on $GABA_A$ receptors are complex and involve both flumazenil-sensitive and flumazenil-insensitive components, and that other receptors could be involved in the behavioral effects of apigenin. Like most flavonoids, apigenin is known to have a wide variety of biological actions, including effects on adenosine receptors and progestational activity. Of particular interest are the findings that apigenin, at concentrations that inhibit $GABA_A$ and $GABA_C$ receptors, also inhibits NMDA receptors. Such an action could contribute to the flumazenil-insensitive sedative actions of apigenin.

Epigallocatechin Gallate

Green tea polyphenols are being considered as therapeutic agents in well-controlled epidemiological studies aimed to alter brain-aging processes and to serve as possible neuroprotective agents in progressive neurodegenerative disorders such as Parkinson's and Alzheimer's diseases.

Epigallocatechin gallate (Figure 2(b)) (EGCG) is the major polyphenol in green tea (*Camellia sinensis*). Studies on $\alpha1\beta2\gamma2L$ $GABA_A$ receptors showed that EGCG at low concentrations (0.1 μM) has a potent second-order modulatory action on the first-order modulation by diazepam but inhibits the action of GABA at higher concentrations (>1 μM). EGCG was an order of magnitude more potent than apigenin in acting as a second-order modulator. In addition, it has been found that EGCG, at concentrations that have no influence on the activation of $GABA_A$ receptors by GABA, was able to reverse the negative modulation of such receptors by methyl β-carboline. This indicates that EGCG may act as a second-order modulator with respect to the first-order modulation by both positive and negative modulators that act on benzodiazepines sites on $GABA_A$ receptors.

The novel second-order modulation by EGCG and apigenin of the maximum first-order modulatory action of diazepam of the activation by GABA of $GABA_A$ receptors observed in these studies may result from altering the coupling of the benzodiazepine allosteric sites with the orthosteric GABA sites on $GABA_A$ receptors. There is evidence from binding studies that the nexus between the benzodiazepine and GABA sites on $GABA_A$ receptors is complex and involves other factors such as phospholipids. EGCG and apigenin may serve as lead compounds for the development of more selective agents for the second-order modulation of benzodiazepine enhancement of the action of GABA on $GABA_A$ receptors.

There is much interest in the anticancer and antitumor properties of EGCG associated with the consumption of green tea. EGCG has anticancer effects on ovarian carcinoma cell lines, is a selective inhibitor of COX-2 expression, and induces apoptosis in monocytes. Little is known about the CNS actions of this flavan. However, it is found in the brain after gastric administration to mice and is neuroprotective in rats on intraperitoneal (IP) injection after focal ischemia and in a mouse model of Parkinson's disease. This neuroprotective action may be associated with its antioxidant properties, but enhancement of $GABA_A$-mediated synaptic inhibition could also contribute. In addition, EGCG is known to reduce glutamate-induced cytotoxicity via intracellular Ca^{2+} ion modulation, suggesting that other neurotransmitter systems may be involved.

Hispidulin and Related Flavonoids

Hispidulin (Figure 2(c)) (the 6-methoxy derivative of apigenin) was isolated together with apigenin from *Salvia officinalis* (sage) recently using a benzodiazepine binding assay-guided fractionation. Hispidulin was some 30 times more potent than apigenin in displacing flumazenil binding. Preparations of sage have been used in herbal medicine to assist memory, and an extract of *Salvia lavandulaefolia* (Spanish sage) has been shown to enhance memory in healthy young volunteers. Unlike apigenin, hispidulin has been shown to act as a positive allosteric modulator of $GABA_A$ receptors. The positive modulatory action of 10 μM hispidulin at $\alpha1\beta2\gamma2S$ receptors was reduced from 47% to 17% by flumazenil, indicating that sites other than classical flumazenil-sensitive benzodiazepine sites were involved in the action of hispidulin. Because hispidulin did not influence the action of GABA on $\alpha1\beta2$ $GABA_A$ receptors, this indicates that it did not interact with low-affinity flumazenil-insensitive benzodiazepine sites. Hispidulin was shown to have an anticonvulsant action in seizure-prone Mongolian gerbils and to pass the blood–brain barrier.

Flavonoids structurally related to hispidulin that influence benzodiazepine binding have been isolated from *Scutellaria baicalensis*, an important herb in traditional Chinese medicine. Oroxylin A (5,7-dihydroxy-6-methoxyflavone, i.e., hispidulin lacking the 4′-hydroxy group) inhibits flunitrazepam binding at 1 μM, and on oral administration acts as a neutralizing modulator blocking the anxiolytic, myorelaxant, and motor incoordination effects but not the sedative and anticonvulsant effects elicited by diazepam. 6-Methylapigenin isolated from *Valeriana wallichii,* a known sedative herb, influences benzodiazepine binding at 0.5 μM in manner suggesting it may be a positive modulator of $GABA_A$ receptors. 6-Methylapigenin has anxiolytic properties, and it can act synergistically to potentiate the sleep-enhancing properties of hespiridin, a flavanone glycoside from *Valeriana*.

Thus, flavones substituted in the 6 position with a methoxy or methyl substituent have interesting effects on $GABA_A$ receptor function and may contribute to the properties of some herbal preparations. Several flavonoid glycosides, including goodyerin, linarin, and hesperidin, are also being studied as sedative and anticonvulsant agents likely to interact with $GABA_A$ receptors.

(+)-Borneol

(+)-Borneol (Figure 2(d)) is a bicyclic monoterpenoid found in high concentrations in extracts of *Valeriana officinalis* that are widely used to reduce the latency of sleep onset and to increase the depth of sleep and the perception of well-being. Extracts of *Valeriana* are known to contain a large number of constituents, including flavonoids and terpenoids, many of which are considered to be active at $GABA_A$ receptors.

(+)-Borneol was shown to be a flumazenil-insensitive positive allosteric modulator of human recombinant $\alpha1\beta2\gamma2L$ $GABA_A$ receptors of low potency (median effective concentration (EC_{50}) 250 μM) but very high efficacy, producing a ten-fold enhancement of the action of 10 μM GABA at a concentration of 450 μM. (−)-Borneol showed similar positive modulatory properties to (+)-borneol, whereas isoborneol, (−)-bornyl acetate, and camphor (a known convulsant) were much less active. The relatively rigid cage structure of these bicyclic monoterpenes and their high efficacy may aid in a greater understanding of the molecular aspects of positive modulation.

Kava Lactones

Kava lactones (α-pyrones) have been recognized as the active constituents responsible for the reported anxiolytic and sedative effects of kava-kava. The six major kava lactones are kawain (Figure 2(e)), dihydrokawain, methysticin, dihydromethysticin, yangonin, and demethoxyyangonin, and they occur in varying proportions in different cultivars.

There is evidence to indicate that kava lactones may mediate their anxiolytic and sedative effects via positive modulation of $GABA_A$ receptor function acting at sites different from the classical benzodiazepine binding sites. Individual kava lactones have been shown to enhance the binding of bicuculline methochloride and muscimol to brain membranes without

influencing the binding of diazepam or flunitrazepam, indicating that they interact with the agonist/competitive antagonist binding sites on $GABA_A$ receptors. The enhancement of muscimol binding showed regional selectivity to be highest in the hippocampus, amygdala, and medulla, with minimal effect on the cerebellum and frontal cortex, suggesting that the kava lactones may show subtype specificity for particular $GABA_A$ receptors. Functional studies of recombinant $\alpha1\beta2\gamma2L$ $GABA_A$ receptors have shown that kava lactones do act as positive modulators, but such studies need to be extended to investigate different subunit combinations.

See also: Anxiety Disorders; GABA Synthesis and Metabolism; GABA-A Receptors: Developmental Roles; GABA-A Receptor Synaptic Functions; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-A Receptors and Disease; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism.

Further Reading

Beaubrun G and Gray GE (2000) A review of herbal medicines for psychiatric disorders. *Psychiatric Services* 51: 1130–1134.

Campbell EL, Chebib M, and Johnston GAR (2004) The dietary flavonoids apigenin and (−)-epigallocatechin gallate enhance the positive modulation by diazepam of the activation by GABA of recombinant $GABA_A$ receptors. *Biochemical Pharmacology* 68: 1631–1638.

Chebib M, Hanrahan JR, Mewett KN, Duke RK, and Johnston GAR (2004) Ionotropic GABA receptors as therapeutic targets for memory and sleep disorders. *Annual Reports in Medicinal Chemistry* 39: 13–23.

Ernst E (2006) Herbal remedies for anxiety – a systematic review of controlled clinical trials. *Phytomedicine* 13: 205–208.

Fernandez SP, Wasowski C, Loscalzo LM, et al. (2006) Central nervous system depressant action of flavonoid glycosides. *European Journal of Pharmacology* 539: 168–176.

Granger RE, Campbell EL, and Johnston GAR (2005) (+) and (−)-Borneol: Efficacious positive modulators of GABA action at human recombinant $\alpha1\beta2\gamma2L$ $GABA_A$ receptors. *Biochemical Pharmacology* 69: 1101–1111.

Huang SH, Duke RK, Chebib M, et al. (2004) Ginkgolides, diterpene trilactones of *Ginkgo biloba*, as antagonists at recombinant $\alpha1\beta2\gamma2L$ $GABA_A$ receptors. *European Journal of Pharmacology* 494: 131–138.

Johnston GAR (2005) $GABA_A$ receptor channel pharmacology. *Current Pharmaceutical Design* 11: 159–164.

Johnston GAR, Hanrahan JR, Chebib M, Duke RK, and Mewett KN (2006) Modulation of ionotropic GABA receptors by natural products of plant origin. *Advances in Pharmacology* 54: 285–316.

Kavvadias D, Sand P, Youdim KA, et al. (2004) The flavone hispidulin, a benzodiazepine receptor ligand with positive allosteric properties, traverses the blood–brain barrier and exhibits anticonvulsive effects. *British Journal of Pharmacology* 142: 811–820.

Leung JWC and Xue H (2003) GABAergic functions and depression: From classical therapies to herbal medicine. *Current Drug Targets – CNS & Neurological Disorders* 2: 363–373.

Marder M and Paladini AC (2002) $GABA_A$-receptor ligands of flavonoid structure. *Current Topics in Medicinal Chemistry* 2: 853–867.

Möhler H (2006) $GABA_A$ receptors in central nervous system disease: Anxiety, epilepsy, and insomnia. *Journal of Receptors and Signal Transduction* 26: 731–740.

Shelp BJ, Bown AW, and Faure F (2006) Extracellular γ-aminobutyrate mediates communication between plants and other organisms. *Plant Physiology* 142: 1350–1352.

Hereditary Spastic Paraplegia

J K Fink, University of Otago, Christchurch, New Zealand

© 2009 Published by Elsevier Ltd.

Following normal prenatal and early childhood development, subjects with uncomplicated hereditary spastic paraglegia (HSP) develop leg stiffness and gait disturbance (stumbling and tripping) due to difficulty dorsiflexing the foot and weakness of hip flexion. Although the majority of patients experience symptom onset in the second through fourth decades, symptom onset from infancy through age 85 has been observed. Gait disturbance progresses insidiously without exacerbations, remissions, or saltatory worsening. Subjects may experience paresthesiae below the kness. Urinary urgency progressing to urinary incontinence is a frequent, although variable, late manifestation. It should be noted that some autosomal dominant, uncomplicated HSP kindreds exhibit onset of progressive spastic paraplegia in childhood (before age 6 years) and relatively little progression of symptoms beyond adolescence. These patients often do not experience urinary bladder disturbance and generally remain ambulatory with assistance. Age of symptom onset, rate of progression, and extent of disability may be variable both within and between families.

Neurologic abnormalities in uncomplicated HSP are limited to the lower extremities which demonstrate increased muscle tone and weakness that is particularly noted at iliopsoas, tibialis anterior, and, to a lesser extent, hamstring muscles. Distal muscle wasting has been reported in uncomplicated HSP although is usually mild and limited to atrophy of the shins in wheelchair-dependent elderly patients. Mildly diminished vibratory sense in distal lower extremities is often noted. Although the combination of mildly impaired vibratory sense and lower extremity spasticity is helpful in recognizing HSP, vibration perception may be normal in mildly affected subjects. Slight terminal dysmetria is observed occasionally on finger-to-nose testing in older affected subjects. Deep tendon reflexes are pathologically increased in the lower extremities. Crossed adductor reflexes, ankle clonus, and extensor plantar responses are present. Gait is abnormal with circumduction owing to hip flexor and ankle dorsiflexion weakness. Pes cavus is generally present in older affected subjects.

Currently, there is no specific treatment to prevent, retard, or reverse HSP's progressive disability. Symptomatic reduction in spasticity often occurs with oral and intrathecal lioresal and oral dantrolene. Bladder spasticity has been improved with oxybutynin.

Nerve conduction studies in uncomplicated HSP are generally normal although subclinical sensory impairment has been reported. Lower extremity somatosensory evoked potentials show conduction delay in dorsal column fibers. Cortical evoked potentials, used to measure neurotransmission in corticospinal tracts, show reduced corticospinal tract conduction velocity and amplitude of evoked potential in muscles innervated by lumbar spinal segments but normal or only mildly reduced conduction velocity in muscles innervated by cervical segments. These findings indicate that there are fewer corticospinal tract axons reaching the lumbar spinal cord and that conduction velocity is reduced in those that remain.

Neuropathology

Autosomal dominant, uncomplicated HSP is characterized by axonal degeneration that is maximal in the terminal portions of the longest descending and ascending tracts (crossed and uncrossed corticospinal tracts to the legs, fasciculus gracilis fibers, and to a lesser extent, spinocerebellar fibers). Neuronal cell bodies of degenerating fibers are preserved. Some cases of HSP exhibit mild loss of anterior horn cells. Dorsal root ganglia, posterior roots, and peripheral nerves are normal. There is no evidence of primary demyelination.

Differential Diagnosis

Diagnosing HSP is straightforward when the family history indicates autosomal dominant inheritance of progressive spasticity as an isolated symptom. Documenting associated deficits (such as ataxia, retinitis pigmentosa and periphereal neuropathy and extrapyramidal signs) is important to classify HSP patients and to distinguish them from other spinocerebellar syndromes including Machado–Joseph disease. Other diagnostic considerations are most relevant when examining patients with apparently sporadic disorders. The primary differential diagnosis of insidiously progressive, spastic para- or tetraparesis in children and young adults includes structural spinal cord disturbance from tethered cord syndrome, Arnold-Chiari malformations, spondylosis, neoplasm, multiple sclerosis, and arteriovenous malformations of the spinal cord. Myelopathies associated with vitamin B12 deficiency, vitamin E deficiency, and abetalipoproteinemia (Bassen–Kornzweing disease) are typically accompanied by peripheral neuropathy. It is important to consider progressive dystonia with diurnal variation in all individuals with progressive

gait disturbance, as this disorder is completely treatable with low-dose levodopa-carbidopa. Although corticospinal tract signs occasionally precede weakness and wasting in amyotrophic lateral sclerosis (ALS), electromyography readily distinguishes ALS from HSP. Rare, adult variants of adrenomyeloneuropathy (ALD/AMN), globoid cell (Krabbe) leukodystrophy, and metachromatic leukodystrophy (MLD) can be evaluated with serum long-chain fatty acid, leukocyte galactocerebrosidase, and arylsulfatase A assays (respectively). Tropical spastic paraparesis due to HTLV I infection and acquired immunodeficiency syndrome (AIDS) related myelopathy may be evaluated by appropriate antibody titres, although each typically has a subacute course. Similarly, myelopathy as a paraneoplastic syndrome usually has a subacute presentations in contrast to HSP's insidious progression.

Genetic Analysis

Uncomplicated autosomal dominant, autosomal recessive, and X-linked HSP are genetically heterogeneous: different genes cause clinically indistinguishable disorders. A locus for autosomal recessive HSP is on chromosome 8q. Loci for autosomal dominant HSP have been identified on chromosomes 2p, 14q, and 15q. Currently, linkage to chromosome 2p is the most frequent among autosomal dominant HSP families linked to any chromosomal locus. For many HSP families, however, the disorder is excluded from known HSP loci. This indicates that the disease is linked to additional, as yet unidentified loci in approximately 45% of autosomal dominant HSP kindreds.

X-linked uncomplicated HSP is rare and mapped to a locus on Xq22 in two families. A missense mutation in the proteolipoprotein (PLP) gene coding sequence was found in affected males of one of these families. This indicates that *uncomplicated* X-linked HSP in this family is allelic to a form of *complicated* X-linked HSP and Pelizaeus–Merzbacher disease, as they all result from PLP gene mutations.

Genetic Counseling

With few exceptions, uncomplicated HSP exhibits nearly complete, age-dependent genetic penetrance. This is an important consideration when counseling subjects from autosomal dominant HSP kindreds, since first-degree relatives of affected subjects will have approximately 50% lifetime risk of developing the disorder. Genetic counseling must consider that although HSP generally exhibits complete genetic penetrance, age of symptom onset, rate of symptom progression, extent of urinary bladder involvement, and functional disability may vary significantly both within and between families. One cannot predict with certainty that a wheel chair will be required or that urinary incontinence will inevitably occur. A cautious wait and see approach is advised. Genetic counselors and physicians must note that the disorder in some HSP kindreds begins in childhood and is relatively nonprogressive after approximately age 10 years. Patients in these kindreds often remain ambulatory.

While the availability of genetic markers tightly linked with HSP loci greatly improves genetic counseling, their use is limited to counseling informative subjects from kindreds linked to these loci who are at risk of transmitting or inheriting this disorder. Although limited to informative members of kindreds linked to known HSP loci, tightly linked markers can be used prenatally to assess the chance that the fetus has inherited HSP. Inheritance of tightly linked markers alone should not be used for early diagnosis of HSP, however. It is important that each patient suspected of having HSP undergo thorough neurologic examination and often neuroimaging to exclude other neurologic disorders. Extreme caution must be used in counseling unaffected parents of a child with suspected HSP about the risk of HSP in other progeny. The disorder in this case could be sporadic (low recurrence risk), due to nonpaternity (uncertain recurrence risk), or autosomal recessive (25% recurrence risk). Furthermore, given the variation in age of symptom onset, observations of genetic anticipation in autosomal dominant HSP, and rare occurrences of incomplete penetrance, it is possible that the disorder is autosomal dominant, inherited from a presymptomatic parent, and that the risk of recurrence in other progeny approaches 50%.

Conclusion

Very strong clinical and pathologic similarity between HSP kindreds linked to different genetic loci suggests that the abnormal gene products from these HSP loci interact in a common biochemical cascade. Disturbance of this pathway results in a degeneration of the distal portions of the longest CNS axons. Identifying the precise gene defects responsible for HSP will advance our knowledge of HSP's pathophysiology and will facilitate rational design of treatments to prevent, retard, or reverse this disorder.

See also: Spasticity; Statistical Tests and Inferences.

Further reading

Behan W and Maia M (1974) Strumpell's familial spastic paraplegia: Genetics and neuropathology. *Journal of Neurology, Neurosurgery and Psychiatry* 37: 8–20.

Fink JK, Sharp G, and Lange B (1995) Autosomal dominant hereditary spastic paraparesis, type I: Clinical and genetic analysis of a large North American family. *Neurology* 45: 325–331.

Fink JK, Wu CB, and Jones SM (1995) Familial spastic paraplegia: Tight linkage to chromosome 15q. *American Journal of Human Genetics* 56: 188–192.

Fink JK and Heiman-Patterson T (1996) Hereditary spastic paraplegia: Advances in genetic research. *Neurology* 46: 1507–1514.

Harding AE (1993) Hereditary spastic paraplegias. *Seminars in Neurology* 13: 333–336.

Hazan J, Lamy C, Melki J, Munnich A, de Recondo J, and Weissenbach J (1993) Autosomal dominant familial spastic paraplegia is genetically heterogeneous and one locus maps to chromosome 14q. *Nature Genetics* 5: 163–167.

Hentati A, Pericack-Vance MA, and Hung W-Y (1993) Linkage of the pure recessive familial spastic paraplegia to chromosome 8 markers and evidence of genetic locus heterogeneity. *Human Genetics* 53: 1013.

Hentati A, Pericak-Vance MA, and Lennon F (1994) Linkage of the late onset automal dominant familial spastic paraplegia to chromosome 2p markers. *Human Molecular Genetics* 3: 1867–1871.

McKusick VA (1988) Spastic paraplegia. *Mendelian Inheritance in Man.* Baltimore, MD: John Hopkins University Press.

Polo JM, Calleja J, Combarris O, and Berciano J (1993) Hereditary pure spatic paraplegia: Study of nine families. *Journal of Neurology, Neurosurgery, and Psychiatry* 56: 175–181.

Heritable Microgyrias

B S Chang and C A Walsh, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Among the syndromes of human brain malformation that result from disruptions in cerebral cortical development, those conditions involving microgyria, or polymicrogyria (PMG), are among the most clinically common. PMG, generally defined as the presence of an excessive number ('poly') of small ('micro') gyri, and usually associated with an abnormal pattern of cortical lamination, can be focal or generalized, unilateral or bilateral. It can be present in isolated form or can be associated with other brain malformations. Its neurological manifestations can include developmental delay in both the cognitive and motor spheres, focal deficits such as weakness on one side of the body, and epilepsy of varying types and severities. PMG can be associated with known environmental insults during embryonic and fetal development, such as intrauterine cytomegalovirus (CMV) infection or loss of oxygenated blood flow to the developing brain, or it can be part of an inherited syndrome with a germline genetic basis. In this article, we focus on the major heritable microgyria syndromes and review their clinical characteristics, radiological appearance, and genetic basis, when known.

Bilateral Frontoparietal Polymicrogyria

Bilateral frontoparietal polymicrogyria (BFPP) is the only known isolated PMG syndrome whose causative gene has been identified. The study of BFPP serves as an instructive example of the region-specific nature of many inherited PMG syndromes, the importance of brain development genes in not only the patterning of the cortex but also the proper development of subcortical and hindbrain structures, and the utility of detailed genotype-phenotype analyses to the practicing neurological clinician.

Clinical Characteristics

BFPP is usually characterized clinically by moderate to severe delay in both the cognitive and motor spheres, resulting in mental retardation and the spastic form of cerebral palsy. Refractory epilepsy, with focal and generalized seizures, is usually present in these children. Most children die before adulthood either from infections or from complications of seizures, although some patients with BFPP have survived into their 20s. Two distinctive clinical features help to differentiate this condition from similar bilateral PMG syndromes. First, a majority of patients have dysconjugate gaze (usually esotropia, or crossed eyes); many have had corrective surgery for strabismus in early childhood. Second, uncoordinated limb movements (cerebellar ataxia) are often present, although this finding may be hard to detect in the presence of severe spasticity.

Radiological Appearance

The 'classic' BFPP syndrome has a very stereotypical radiological appearance (Figure 1). PMG is present in a widespread distribution on both sides of the brain, but with a prominent anterior–posterior gradient such that the frontal and parietal lobes are most severely affected, while the occipital lobes are least affected or can on occasion even appear completely spared. The PMG is typically diagnosed based on an apparent increase in the thickness of the cerebral cortex (representing 'heaped-up' microgyri superimposed on one another without distinguishable intervening sulci), and an irregular, scalloped junction between gray matter and white matter. In addition to the frontoparietal PMG, however, the classic BFPP syndrome is also characterized by patchy areas of signal change in the white matter on T2-weighted magnetic resonance imaging (MRI), suggestive of abnormal myelination or maldevelopment of the white matter. There is also usually underdevelopment of the brain stem (resulting in a flattened-appearing pons) and cerebellum (affecting both the vermis and hemispheres).

Genetic Etiology

BFPP, originally described in multiple consanguineous pedigrees of Middle Eastern and South Asian origin, is inherited in an autosomal recessive manner. It is now known to be associated with mutations in *GPR56*, a gene that encodes an orphan G-protein-coupled receptor whose exact role in brain development is still unknown. The GPR56 protein is expressed in the ventricular zones of the developing brain, suggesting that the molecular mechanisms underlying PMG may be ones affecting cell fate at an early stage of cortical development, rather than merely late stages of post-migrational cortical organization.

Genotype–Phenotype Analysis and Related Syndromes

The description and genetic characterization of *GPR56*-associated BFPP led to the identification of

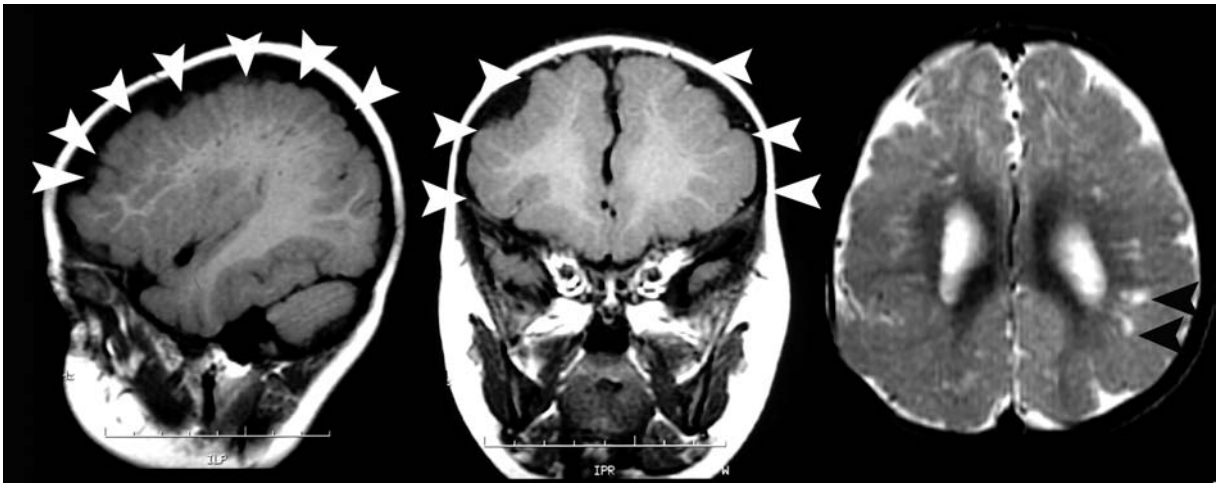

Figure 1 BFPP. T1-weighted sagittal (left), T1-weighted coronal (middle), and T2-weighted axial (right) brain MRI of a 1-year-old child with *GPR56*-associated BFPP. White arrowheads indicate the anatomical distribution of the PMG, which is widespread but has a decreasing anterior–posterior gradient of severity and relatively spares the occipital lobes. Black arrowheads indicate patches of T2-weighted signal change within the white matter, a characteristic radiological sign of this disorder.

multiple families with a very similar clinical and radiological condition now referred to as BFPP2, or BFPP-like syndrome. In this disorder, which also appears to be inherited in an autosomal recessive manner, the appearance of the cerebral cortex is essentially identical to that in *GPR56*-associated BFPP, but there are typically no underlying white matter signal changes and no abnormalities of the brain stem or cerebellum, unlike in BFPP. None of the individuals affected by BFPP2 have been found to harbor *GPR56* mutations to date, strongly suggesting that the role of *GPR56* in human brain development may extend beyond cortical patterning to also involve white matter and hindbrain development.

Bilateral Perisylvian Polymicrogyria

Bilateral perisylvian polymicrogyria (BPP) is the most commonly recognized inherited form of PMG. Historically, disorders with overlapping clinical features have been variously described as 'congenital bilateral perisylvian syndrome' and 'Worster-Drought syndrome,' but these and other terms are likely to be references to the same pathologic entity, since many cases have been characterized histologically by PMG lining the Sylvian fissures of the brain bilaterally. BPP is one of the rare cortical malformations for which the clinical features are so distinctive that experienced neurologists may suspect this specific malformation on clinical grounds alone.

Clinical Characteristics

The most distinctive clinical feature of BPP is a difficulty with coordinated oral movements that is commonly referred to as oromotor dyspraxia, or pseudobulbar paresis. This can range in severity from a mild speech impediment that is only evident in later childhood to a severe inability to feed, suck, or swallow in infancy, requiring tube feeding in some cases. In even mild cases, the tongue may have a noticeable limitation of coordinated movement, and children or adults who can speak often do so with significant slurring, or dysarthria. Cognitive disabilities and cerebral palsy are common as well, although the rate of frank mental retardation appears to be relatively low. A seizure disorder is present in most patients but can vary significantly in severity, ranging from the occasional partial seizure to medically refractory epilepsy.

Radiological Appearance

One of the more obvious radiological signs of BPP is an elongation and gross malorientation of the Sylvian fissure (Figure 2). On sagittal MRI images, the fissure often appears abnormally 'open,' without proper apposition of the two sides. The fissure may extend posteriorly and superiorly in an elongated fashion, reaching the vertex, rather than ending below the inferior parietal lobule as usual. On close inspection, the gyri in the perisylvian region, both superiorly and inferiorly, demonstrate PMG with increased apparent cortical thickness and an irregular gray–white junction. In some instances, the PMG can be more widespread, extending into dorsolateral parts of the frontal lobes or posterior parietal regions, but the predominance of the PMG in the perisylvian area, the contiguity of the malformed regions of cortex, and the abnormally oriented Sylvian fissure, if present, usually help to identify these cases as variants of the fundamental BPP syndrome.

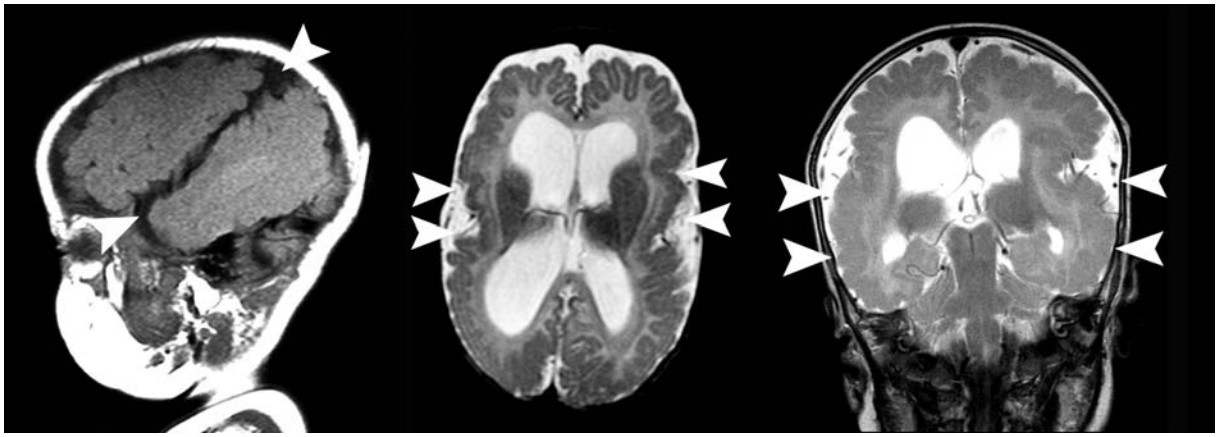

Figure 2 BPP. T1-weighted sagittal (left), T2-weighted axial (middle), and T2-weighted coronal (right) brain MRI of a 3-month-old child with BPP. White arrowheads demarcate the two ends of the abnormally open and maloriented Sylvian fissure (left), as well as the presence of PMG along the Sylvian fissures bilaterally, particularly inferiorly (middle and right).

Genetic Etiology

BPP appears to be genetically heterogeneous. One study mapped several families with an inherited form of BPP to a locus at chromosome Xq28, but no causative gene has yet been identified. Other reports describe BPP families whose pedigrees appear consistent with autosomal recessive or dominant inheritance.

Susceptibility of Perisylvian Region

Large series of PMG cases demonstrate that the perisylvian region is the area of cortex most commonly affected by this malformation, either alone or in combination with other areas. Why the perisylvian cortex should be more susceptible to PMG is unclear, but potential factors include region-specific differences in gene expression, differential timing of gyral development, and the particularities of its arterial supply. The abnormally open and maloriented appearance of the Sylvian fissure in many cases of BPP does suggest that the malformation has its basis in an earlier stage of cortical development than simply postmigrational organization.

Bilateral Generalized Polymicrogyria

Bilateral generalized polymicrogyria (BGP) is a unique syndrome of widespread PMG that appears to be inherited in an autosomal recessive manner.

Clinical Characteristics

BGP patients have cognitive delay of variable severity, with about half exhibiting severe delay. Problems with motor development are also nearly universal, and were reported to be severe in just fewer than half of the originally described cases. Seizures are present in the majority of cases, and are occasionally refractory. No specific oral motor dysfunction is seen, unlike in BPP, and no specific eye movement abnormalities or cerebellar dysfunction are seen, unlike in BFPP. Macrocephaly (large head size), rarely seen in other PMG syndromes, was reported in a quarter of the originally described cases.

Radiological Findings

In BGP, the PMG is seen in a generalized distribution in all cases, but about two-thirds of the time there is a perisylvian predominance (Figure 3). Some investigators classify cases such as these as a subtype of BPP that includes extension of the PMG all the way to the frontal and occipital poles, but a distinct BGP syndrome that includes enlarged ventricles, reduced white matter volume, and macrocephaly in some cases, appears to exist. In addition, the Sylvian fissure does not usually extend abnormally to the vertex in BGP, as it does in most cases of BPP. Some BGP patients do have signal change in the white matter on T2-weighted MRI images.

Genetic Etiology

The genetic etiology of BGP is not known. Pedigrees of the original seven families suggest that the disorder is inherited in an autosomal recessive manner. Analyzed pedigrees have not shown evidence of linkage to the BFPP locus on chromosome 16q, despite the widespread nature of the PMG in both syndromes, and *GPR56* mutations have not been found in patients with BGP.

Bilateral Frontal Polymicrogyria

Bilateral frontal polymicrogyria (BFP) was first described as a discrete brain malformation in 2000. Its

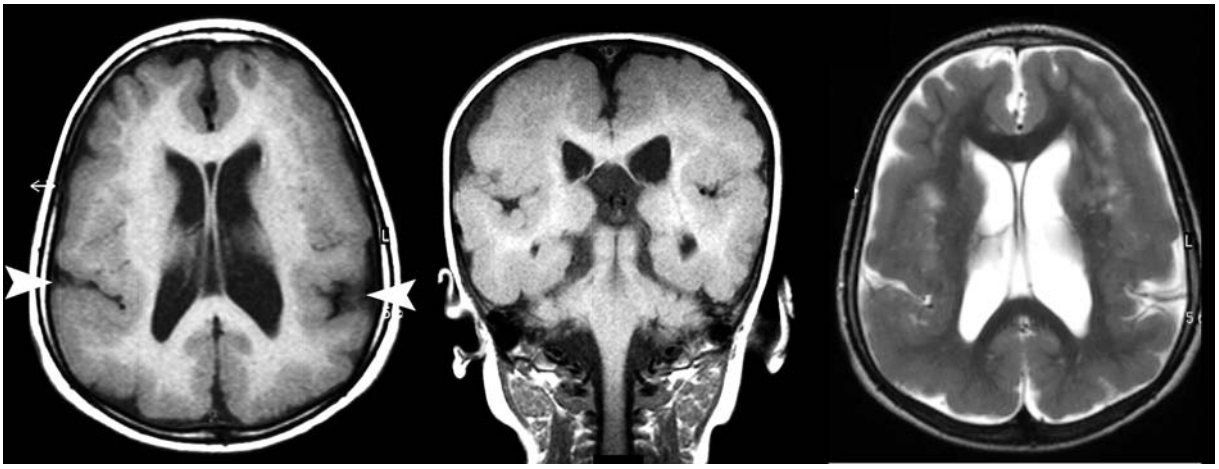

Figure 3 BGP. T1-weighted axial (left), T1-weighted coronal (middle), and T2-weighted axial (right) brain MRI of a 7-year-old child with BGP. In all three images, the PMG is widely distributed throughout the cortex, without any particular anterior–posterior gradient, but the perisylvian regions are prominently affected (white arrowheads). Ventriculomegaly, as seen in this patient, is commonly observed in BGP.

restriction to the frontal cortex distinguishes it from the other heritable bilateral PMG syndromes discussed here so far.

Clinical Characteristics

All of the originally described patients with isolated BFP had clinical signs of developmental delay and mild spastic weakness of all extremities. Language impairment was nearly universal, although the extent varied, and mental retardation of varying severity was similarly present in most patients. A minority of the patients had seizures, with partial, generalized tonic-clonic, atypical absence, and clonic seizure types all represented. Oral motor problems are not typically seen in BFP, nor have eye movement abnormalities or cerebellar signs been generally reported.

Radiological Findings

BFP is radiologically characterized by irregular cortical infoldings with an increase in apparent cortical thickness, diagnostic of PMG, in the frontal lobes bilaterally (Figure 4). Most commonly, the malformation in BFP extends from the frontal poles posteriorly to the precentral gyrus, and inferiorly to involve the frontal operculum. The infrasylvian cortex is not affected, unlike in BPP. A few of the originally described BFP patients had small foci of frontal white matter signal change and about half had mild enlargement of the ventricles. No brain stem or cerebellar abnormalities were seen.

Genetic Etiology and Related Syndromes

No known causative gene has been identified for BFP. The condition is presumed to be of autosomal recessive inheritance, given the absence of affected parents and the presence of consanguineity in two of the originally described families. BFP, in association with a thin corpus callosum, is also seen as a component of Warburg Micro syndrome, an autosomal recessive disorder associated with mutations in the *RAB3GAP* gene that is characterized by congenital eye anomalies, underdevelopment of the genitals, and microcephaly (small head size).

Familial Unilateral Right Polymicrogyria

Although unilateral PMG is commonly assumed by clinicians to arise from environmental causes, some cases of unilateral PMG have clear genetic etiologies (e.g., those associated with chromosome 22q11 deletions), and more recently, a familial syndrome of unilateral right-sided PMG has been described in four pedigrees, each of which includes two affected members, suggesting a germline genetic etiology for this condition in at least some cases.

Clinical Characteristics

Patients with familial unilateral right-sided PMG generally present with the clinical triad of cognitive difficulties, left-sided weakness, and partial epilepsy with a seizure focus in the right side of the brain. Most patients have low normal intelligence or mild mental retardation, and the left-sided weakness is generally mild to moderate in severity. The seizure disorder is refractory in nature about half of the time.

Radiological Findings

All four families with this syndrome have a very stereotyped appearance to their unilateral right

PMG, which affects the frontal and perisylvian cortex (Figure 5). In all cases, the Sylvian fissure on the right is deeply invaginated at its posterior extent, and PMG extends from the frontal pole posteriorly to the posterior end of the Sylvian fissure. In addition, the right cerebral hemisphere is smaller than the left hemisphere in all cases, although this asymmetry varies in severity among the families.

Genetic Etiology

No genetic etiology has yet been identified for familial unilateral right-sided PMG. The disorder may in fact be genetically heterogeneous, as three of the four reported families include two affected siblings born to healthy unrelated parents (suggestive of autosomal recessive inheritance) while the fourth family includes an affected mother and son (compatible with autosomal dominant or X-linked inheritance). Unilateral right PMG has been seen in the setting of chromosome 22q11 deletions, usually in the context of associated facial and cardiac anomalies, although sometimes in isolation. It has also been observed in association with an unbalanced translocation resulting in partial 1q monosomy and 12p trisomy,

Figure 4 BFP. T1-weighted axial (left), T1-weighted sagittal (middle), and T2-weighted coronal (right) brain MRI of an 18-year-old woman with BFP. White arrowheads demarcate the extent of the PMG, which involves the frontal lobes bilaterally but does not extend posteriorly to the parietal lobes or inferiorly to the temporal lobes. The coronal image through the frontal lobes (right) demonstrates that the PMG affects most of the frontal cortex equally but relatively spares the orbitofrontal regions.

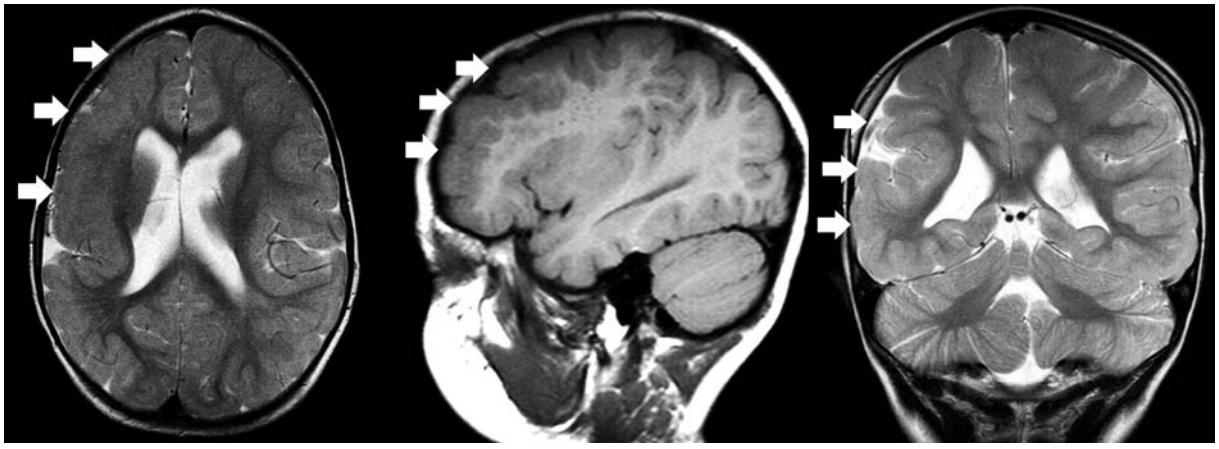

Figure 5 Familial unilateral right-sided PMG. T2-weighted axial (left), T1-weighted sagittal (middle), and T2-weighted coronal (right) brain MRI of a 4-year-old child with familial unilateral right-sided PMG. White arrows indicate the anatomical extent of the malformation, which involves the frontal and perisylvian cortex, extending from the frontal pole to a deep posterior infolding of the Sylvian fissure. The affected right hemisphere is smaller than the left hemisphere in this syndrome, as demonstrated in these images.

although those cases were also characterized by additional congenital abnormalities outside the central nervous system.

Lessons from Heritable Microgyria Syndromes

Among the human brain malformations, PMG has played an important role in our exploration of the genetic and molecular mechanisms of cerebral cortical development. In particular, the recognition that there are multiple region-specific bilateral and unilateral PMG syndromes offers the opportunity to understand better the unique roles of individual genes involved in cortical patterning and development. In addition, the presence of associated subcortical and hindbrain abnormalities keyed to a particular genotype raises questions about the mechanistic relationship, if any, between PMG and other malformations (such as cobblestone lissencephaly syndromes) that also feature abnormalities in these brain structures.

With many PMG genes yet to be identified, multiple potential genotype–phenotype correlations remain possible in this field – a one-to-one correspondence between discrete PMG syndromes and unique causative genes; genetic heterogeneity, with multiple genes linked to a clinically and radiologically uniform syndrome; or phenotypic variability, with one causative gene for multiple distinct but related PMG syndromes.

Besides the instructive nature of its study from a neurobiological viewpoint, PMG also plays a large role in clinical practice, being one of the more common cortical malformations and potentially presenting with a variety of different clinical manifestations depending on its anatomical location and extent. Thus, a detailed understanding of both the genotype–phenotype relationships and the clinico-anatomical correlates underlying various PMG syndromes will be of use to practicing neurologists seeking to diagnose, treat, and counsel patients with these conditions.

Acknowledgments

B S Chang and C A Walsh were supported by the National Institute of Neurological Disorders and Stroke (K23 NS049159 and R37 NS35129, respectively). C A Walsh is an Investigator of the Howard Hughes Medical Institute.

See also: Cerebral Cortex; Epilepsy; Memory Consolidation: Cerebral Cortex; Orofacial Motor Control.

Further Reading

Andermann F (2000) Cortical dysplasias and epilepsy: A review of the architectonic, clinical, and seizure patterns. *Advances in Neurology* 84: 479–496.

Barkovich AJ (2005) Congenital malformations of the brain and skull. In: *Pediatric Neuroimaging,* 4th edn. Philadelphia: Lippincott Williams and Wilkins.

Barkovich AJ, Hevner R, and Guerrini R (1999) Syndromes of bilateral symmetrical polymicrogyria. *American Journal of Neuroradiology* 20: 1814–1821.

Chang BS, Piao X, Bodell A, et al. (2003) Bilateral frontoparietal polymicrogyria: Clinical and radiological features in 10 families with linkage to chromosome 16. *Annals of Neurology* 53: 596–606.

Gaitanis JN and Walsh CA (2004) Genetics of disorders of cortical development. *Neuroimaging Clinics of North America* 14: 219–229.

Guerreiro MM, Andermann E, Guerrini R, et al. (2000) Familial perisylvian polymicrogyria: A new familial syndrome of cortical maldevelopment. *Annals of Neurology* 48: 39–48.

Guerrini R and Filippi T (2005) Neuronal migration disorders, genetics, and epileptogenesis. *Journal of Child Neurology* 20: 287–299.

Harding B and Copp AJ (1997) Malformations. In: Graham DI and Lantos PL (eds.) *Greenfield's Neuropathology,* 6th edn. London: Arnold.

Jansen A and Andermann E (2005) Genetics of the polymicrogyria syndromes. *Journal of Medical Genetics* 42: 369–378.

Piao X, Hill RS, Bodell A, et al. (2004) G protein-coupled receptor-dependent development of human frontal cortex. *Science* 303: 2033–2036.

Villard L, Nguyen K, Cardoso C, et al. (2002) A locus for bilateral perisylvian polymicrogyria maps to Xq28. *American Journal of Human Genetics* 70: 1003–1008.

Hibernation

T S Kilduff, SRI International, Menlo Park, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, © 2004, Elsevier B.V.

Hibernation in mammals is an adaptation for energy conservation during periods of cold and limited food availability that involves seasonal modulation of behavioral state and physiological control systems. Contrary to popular belief, animals do not remain continuously torpid throughout the hibernation season; rather, they typically cycle between euthermia and deep hibernation attaining body temperatures (T_b) close to freezing many times during the winter. Individual bouts of torpor can last from several hours to several weeks in length and are punctuated by periods of euthermia (the interbout intervals). These recurring patterns of increased and dramatically decreased T_b and accompanying changes in behavior and physiology are referred to as the hibernation cycle. It is thought that these dramatic changes in arousal state evolved from normal euthermic arousal state control mechanisms. Therefore, the molecular, neurobiological, and endocrine mechanisms regulating the events of the hibernation cycle are likely to be of considerable significance in understanding mammalian arousal state transitions in general.

Four major phases of the hibernation cycle can be identified (Figure 1). The 'euthermic phase' during the interbout intervals is characterized by a high proportion of nonrapid eye movement (NREM) sleep and body temperature regulation between 36 and 38 °C. During 'entrance' to hibernation, an active resetting of the hypothalamic 'thermostat' which regulates T_b occurs, resulting in a decrease in metabolism that approaches 1/30 to 1/100 of the euthermic value at the same ambient temperature. This dramatic decline in T_b and metabolism usually occurs over the course of 12–24 h and is characterized by a predominance of NREM sleep, with a suppression of both wakefulness and REM sleep. The 'deep hibernation phase' is characterized by a low T_b that can either be in equilibrium with environmental temperature or be regulated at a level between environmental and euthermic body temperature. T_b can be as low as 1.2 °C in the ground squirrel *Citellus lateralis*, and there is evidence for subfreezing T_b values in Arctic species. Thus, hibernation is not simply passive hypothermia; rather, T_b is determined by an active thermoregulatory system. Dramatic changes occur in other regulated physiological control systems as well: as T_b declines from euthermic values of 37 to 5 °C in ground squirrels, heart rate decreases from 200–300 beats min^{-1} to 7–10 beats min^{-1}, and respiratory rate drops from 100–150 breaths min^{-1} to 1–2 breaths min^{-1} with long periods of apnea. During the 'arousal' from hibernation, animals return to euthermic T_b by a combination of shivering and nonshivering thermogenesis; the mobilization of energy stores in the brown fat is an important contributor to heat production during the arousal. The specific factors which terminate the deep hibernation phase and provoke the arousal to euthermia are unknown, but recent studies summarized below suggest that sleep homeostasis may be important. The neurobiological mechanisms underlying the dynamic phases of the hibernation cycle, the entrance and arousal, are of interest in understanding transitions between arousal states in mammals.

There are considerable changes in arousal state distribution across the hibernation cycle. Hibernation is entered primarily through NREM sleep and is a regulated lowering of T_b. A downward resetting of the hypothalamic T_b regulator is a general feature of NREM sleep in all mammals studied. These observations have led to the suggestion that NREM sleep and hibernation in mammals are homologous adaptations for energy conservation. Thus, mammalian hibernation may have evolved in species which faced limited food availability as selection favored the energy conservation resulting from the downregulation of T_b that accompanies NREM sleep. Decreasing T_b during NREM sleep and prolonging sleep could have enabled certain species to exploit highly seasonal environments.

Behavioral and electrophysiological studies show that hibernation is not a homogeneous behavioral state. Strumwasser noted that, during deep hibernation ($T_b = 6$ °C), *Citellus beecheyi* made two to three postural adjustments per hour. These adjustments consisted of (1) raising the head and forepart of the body, (2) burrowing deeper into the nesting material, and (3) shifting the position of the body in the nest. In addition, *C. beecheyi* was capable of vocalization and orientation to sounds during deep hibernation. Marmots are also capable of complex behaviors during the entrance into hibernation. For example, they perform stereotyped nest-building behaviors, move in circles in the nest, and curl the body and move the head during the entrance into hibernation. Such complex behaviors generally occur in awake euthermic mammals. The presence of complex behaviors during entrance into, as well as deep hibernation, suggests that their occurrence may be coincident with epochs of wakefulness interspersed throughout a period in which a NREM-like state predominates.

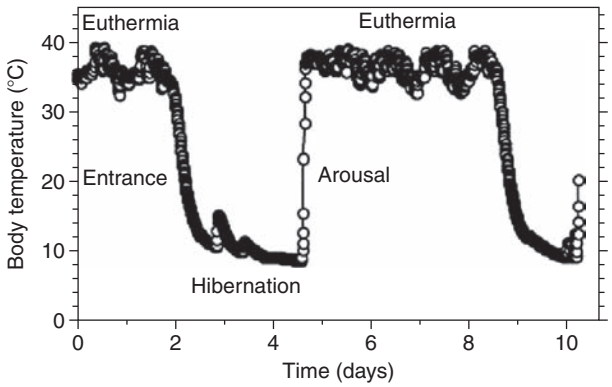

Figure 1 Body temperature recorded every 10 min from a ground squirrel during two consecutive hibernation bouts during the hibernation season. The four major phases of the hibernation cycle are labeled.

Electrophysiological observations support the contention that hibernation is not a homogeneous behavioral state. Above a $T_b = 25\,°C$, hibernating mammals cycle through the range of electrographically defined behavioral states exhibited by euthermic mammals, although both total sleep and NREM sleep time increase while REM sleep time greatly diminishes. During the latter stages of entrance and deep hibernation, cortical activity declines so greatly that standard EEG criteria for state scoring cannot be used below a T_b of about 25 °C. Therefore, arousal state has been assessed in deep hibernation by a combination of thalamic single unit activity and alterations in electromyographic (EMG) activity. At T_b values above 25 °C, arousal states can be clearly distinguished by the activity of single units and the EMG. The same patterns of unit activity and EMG can be used to make arousal state distinctions at lower T_b values. Although the overall activity of thalamic neurons progressively declines as animals enter deep hibernation, spontaneous alterations in the activity of thalamic units and the EMG still occur. These observations have led to the conclusion that animals in deep hibernation (below $T_b = 25\,°C$) continue to cycle through electrophysiologically defined states that are believed to be comparable to euthermic NREM sleep and wakefulness, with NREM predominating. Although REM sleep appears to be completely absent during hibernation, REM sleep occurrence has not been directly addressed by unit studies of subcortical regions associated with REM sleep during hibernation. Thus, both behavioral and cellular neurophysiological studies suggest that a bout of hibernation is mostly, but not exclusively, homologous with NREM sleep.

Although the experiments described previously suggest that hibernation and NREM sleep may be controlled by similar neuronal mechanisms, recent experiments have led to the startling conclusion that animals are actually sleep-deprived during hibernation. The delta range of the EEG (approximate frequencies of 0.5–4 Hz) contains the slow wave activity and can be quantified using spectral analysis of the EEG. The spectral power in the delta range (EEG 'delta power') increases as a function of prior wake duration in many species, including humans, and declines monotonically during the sleep period. These observations have led to the idea that EEG delta power reflects the restoration from preceding waking activities which occurs during sleep. Based on the homology between NREM sleep and hibernation, EEG delta power would be expected to (1) be low right after arousal from hibernation and (2) increase during euthermia because of intervening wakefulness. Furthermore, it might be expected that the longer a hibernation bout, the less the need for subsequent recovery sleep.

The results obtained by two laboratories working independently on different hibernating species were contrary to expectations. Not only did NREM sleep predominate during the interbout periods of euthermia, but EEG delta power was maximal after arousal from hibernation. Delta power declined monotonically until the initiation of the next entrance. Moreover, the longer the hibernation bout, the greater the EEG delta power following arousal. Both the time course of EEG delta power and the high amount of NREM sleep during the interbout euthermic period suggest that animals are sleep-deprived during hibernation and that sleep restoration occurs during the interbout euthermic periods. Although NREM sleep and hibernation appear to be homologous in terms of neuronal control, one possible interpretation of these results is that the low temperatures of torpor might inhibit the underlying cellular and biochemical mechanisms of recovery sleep. Thus, one function of the periodic bouts of euthermia during the hibernation season might be to allow sleep restoration at normothermic brain temperatures. Since the neural bases of the sleep recovery mechanism(s) remain to be elucidated, hibernation should be a valuable model system in which to examine the biochemical and molecular bases of these processes.

Given the complex changes in behavior and physiology described previously, it is evident that mammalian hibernation must involve widespread changes in the activity of the nervous system. The neural events underlying the dramatic arousal state changes and physiological processes occurring during various phases of the hibernation cycle are of direct relevance to understanding arousal state control mechanisms. Depth EEG studies have revealed that brain regions become electrically quiescent in a replicable sequence during entrance into hibernation: the cortex becomes electrically quiescent first at a T_b around 20 °C, followed by

the thalamus and then limbic structures such as the hippocampus. Electrical activity is maintained in the hypothalamus throughout hibernation, whereas intermittent activity occurs in the amygdala. Lesion studies have indicated that the hypothalamus is an important brain region in the regulation of hibernation, as well as in euthermic T_b regulation. Thermoregulatory studies have demonstrated that the regulator of T_b located in the preoptic anterior hypothalamic (POAH) area is continuously reset during entrance to hibernation and remains active during deep hibernation. Pharmacological studies have also suggested that the hypothalamus plays an important role in regulating the entrance to, and arousal from, hibernation.

In order to obtain a more comprehensive view of neural activity during various phases of hibernation than is possible using conventional electrophysiological methods, a series of autoradiographic studies of cerebral metabolic activity was conducted using 2-[^{14}C]deoxyglucose (2-DG). In these studies, two brain nuclei underwent particularly remarkable changes in their metabolic activity relative to other brain structures. The paratrigeminal nucleus (PTN), a little-studied medullary region of unknown function, was the only brain region to significantly increase its relative 2-DG uptake in hibernation when compared to euthermia. A more extensive study across the entire hibernation cycle found that the PTN underwent a state-dependent increase in its relative 2-DG uptake as T_b declined during entrance to hibernation. Consequently, the PTN went from the least metabolically active nucleus during euthermia of the 96 brain regions examined to become the most metabolically active region in hibernation. Although the function of the PTN is unclear at this time, it has been proposed to have thermoafferent and/or chemosensory functions.

The suprachiasmatic nucleus (SCN) of the hypothalamus was the other brain region that underwent a remarkable change in its activity across the hibernation cycle. Like the PTN, the SCN also underwent a state-dependent increase in its relative 2-DG uptake and rank among metabolically active brain regions as T_b declined. The SCN are well known to regulate circadian rhythmicity in mammals and their activity during hibernation may represent maintenance of circadian function during this state. Indeed, circadian rhythms of T_b have recently been shown to persist in the ground squirrel and to be responsible for the timing of arousals, even in the apparent absence of action potentials within the SCN.

To obtain a global view of brain metabolic activity across the hibernation cycle, relative 2-DG uptake in 96 brain regions was measured across seven phases of the hibernation cycle. Statistical analyses revealed that the entrance to hibernation is mediated by two processes: an inhibition of activity in cortical regions and a relative activation of hypothalamic structures. These results are consistent with previous descriptions of a progressive decline in electrical activity in the cortex as entrance to hibernation proceeds and sustained electrical activity in the hypothalamus. In contrast, analysis of the 2-DG results revealed that the arousal from hibernation was primarily mediated by hypothalamic nuclei. Curiously, the four hypothalamic nuclei implicated in the arousal from hibernation are the SCN and three nuclei which receive efferent projections from the SCN. In summary, the entrance to hibernation involves both an inhibition of cortical activity and activation of hypothalamic regions whereas the arousal from hibernation is primarily a hypothalamic function. These results are consistent with those from lesion, physiological, and microinjection studies which have implicated the POAH region as being important in regulating hibernation.

In addition to the central nervous system (CNS), the parasympathetic nervous system appears to play an important role in the entrance to hibernation. Asystoles are commonly observed in the electrocardiogram during entrance to hibernation and can be blocked by intravenous administration of atropine. Whereas high doses of atropine can prevent hibernation from occurring, lower doses administered during hibernation produce a higher resting heart rate than in control animals, further suggesting an increased parasympathetic drive during entrance to hibernation. As the deep hibernation phase is reached, asystoles occur even in the atropinized animal, indicating a decrease in sympathetic drive.

Neurochemical studies of catecholamine levels are consistent with the notion of decreased sympathetic drive during hibernation. In five species studied to date, either no change or decreased levels of norepinephrine (NE) and epinephrine have been found in various brain regions during hibernation, although there is one report of increased levels of NE in the pons/medulla of hibernating *C. lateralis*. Plasma levels of catecholamines also decrease in hibernating woodchucks. Turnover studies indicate either a greatly decreased or complete cessation of catecholamine turnover in the hibernating brain. In contrast, catecholamines appear to play an important role in the arousal process. Both whole brain and hypothalamic NE levels increase during this phase of the hibernation cycle. Arterial infusion of L-DOPA results in arousal in the Turkish hamster; the rate of arousal in golden hamsters is retarded by injection of a tyrosine hydroxylase inhibitor. The crucial brain region for the involvement of NE in the arousal process may be the POAH since microinjections of NE into this area

during deep hibernation evoke a full arousal, whereas control injections into the midbrain reticular formation have no effect. Repeated injections of NE into the lateral ventricles facilitate the rate of arousal.

Changes in brain serotonin (5-HT) metabolism occur prior to and during entrance to hibernation. In the golden hamster, the 5-HT turnover rate increases 18 times in winter compared to summertime levels and increases even further in hibernation. Serotonin levels in several brain areas have been reported to increase greatly during entrance to hibernation. On the other hand, 5-HT levels during deep hibernation have variously been reported to increase, decrease, or not change, but the concentration of the 5-HT metabolite, 5-hydroxyindole acetic acid, has uniformly been found to increase, consistent with the notion of increased 5-HT turnover in this state. Depletion of endogenous 5-HT by injections of parachlorophenylalanine temporarily abolishes the ability to hibernate in the ground squirrel and destruction of at least 30% of the median raphe nuclei completely eliminates hibernation. Hamsters fed a tryptophan-rich diet to increase endogenous 5-HT levels spend a greater amount of time in hibernation. In the squirrel, these changes in serotonin metabolism may act directly upon the thermoregulatory system to decrease T_b, since intracerebroventricular (i.c.v.) 5-HT injections during the hibernation season activate heat loss and inhibit heat production mechanisms. Furthermore, repeated i.c.v. injections of 5-HT during arousal from hibernation slow the rate of arousal by increasing heat loss. Thus, 5-HT appears to promote entry to hibernation and interfere with arousal from this state by affecting thermoregulatory responses.

Acetylcholine (ACh) may be involved in the arousal process as microinjections into either the POAH or midbrain reticular formation can result in a full arousal. The hibernating animal shows a 'progressive irritability' to such ACh injections as a hibernation bout proceeds. Early in a bout, a given dose may have no effect, while the same dose administered progressively later in a bout may result in a thermogenic response or a full arousal to euthermia.

The hippocampus of the hibernator is exquisitely sensitive to injections of thyrotropin-releasing hormone (TRH). Doses as low as 0.1 ng result in arousal, whereas doses in the microgram range have no effect at control sites. Since TRH has been shown to inhibit neuronal activity in several systems, it is thought that TRH blocks the inhibition in the hippocampus normally characteristic of hibernation. Interestingly, microinjections of TRH into the hippocampus during the nonhibernating state have opposite effects on T_b and arousal state, depending on whether injections occur during wakefulness or slow wave sleep.

Endogenous opioids have also been implicated in hibernation. Intraperitoneal injections of naloxone and naltrexone shorten the hibernation bout. I.c.v administration of naloxone decreases bout length in a dose-dependent manner. Increased brain concentrations of both Met- and Leu-enkephalin occur in hibernating squirrels. Lastly, animals administered morphine during hibernation fail to develop physical dependence on morphine, suggesting that a higher proportion of opiate receptors are occupied by endogenous opiates during hibernation.

The pineal gland and its hormone melatonin appear to have different roles in permissive and obligate hibernators. Pinealectomy reduces the incidence of torpor in a photoperiodic species, *Peromyscus leucopus*, whereas melatonin administration increases torpor incidence. It has been suggested that these effects may be secondary to gonadal effects, since testosterone implants block hibernation in two other permissive hibernators, the Turkish and golden hamsters. In obligate hibernators such as the Richardson's and 13-lined ground squirrels, neither melatonin nor pinealectomy influence the incidence or duration of hibernation. However, the second hibernation season after pinealectomy is shortened in the pinealectomized golden-mantled ground squirrel; these animals also do not 'progress' into deeper bouts during the hibernation season. Hence, the pineal may be playing a role of phasing circannual rhythms in obligate hibernators. Interestingly, in the three hibernating species examined to date, the circadian rhythm of pineal or plasma-borne melatonin which peaks at night in the euthermic state is completely abolished during hibernation; melatonin levels remain low throughout a hibernation bout. This may be related to the fact that 2-DG uptake is high in the SCN during hibernation, since there is generally an inverse relationship between the rhythm of 2-DG uptake in the SCN and the rhythm of pineal melatonin.

The biochemical and molecular changes in organ systems which enable hibernation are likely to occur with two distinct time courses. Longer-term biochemical and endocrine changes that are preparatory for hibernation occur on a seasonal basis and have been documented in several organ systems; the numerous examples of involution of endocrine organs prior to the hibernation season are undoubtedly accompanied by major changes in gene expression. Once these seasonal changes in organismal physiology have occurred and the animal is fully prepared for hibernation, the occurrence of a bout of hibernation on any given day is likely to be accompanied by shorter-term changes in gene expression. It likely that shorter-term changes in gene transcription in the CNS are of significance in regulating the various phases of the

hibernation cycle. Because of their central role in regulating numerous downstream 'target' genes, initial studies of gene expression in the CNS across the hibernation cycle have focused on expression of immediate early genes (IEGs). The results published to date indicate a phase-dependent change in IEG mRNA expression in the hypothalamus across the hibernation cycle, with the greatest increases observed during the arousal phase. Furthermore, the anatomical resolution provided by *in situ* hybridization studies indicate that, within the hypothalamus, one locus of the increased c-*fos* expression is the SCN, the same region implicated by the 2-DG studies as undergoing a significant change in its metabolism during arousal. Molecular neurobiological studies of hibernation are likely to be fruitful avenues for future research.

See also: Autonomic Nervous System: Central Thermoregulatory Control; Dopamine Control of Arousal; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Pineal Gland and Melatonin; Sleep Deprivation: Neurobehavioral Changes; Sleep–Wake State Regulation by Acetylcholine; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Beckman AL and Stanton TL (1982) Properties of the CNS during the state of hibernation. In: Beckman AL (ed.) *The Neural Basis of Behavior.* New York: Spectrum.

Carey C, Florant GL, Wunder BA, and Horwitz B (1993) *Life in the Cold Ecological: Physiological and Molecular Mechanisms*. Boulder: Westview Press.

Kilduff TS, Dement WC, and Heller HC (1990) Functional mapping of brain activity during arousal states: Mammalian hibernation as a model system. In: Inoue S and Krueger JM (eds.) *Endogenous Sleep Factors*, pp. 229–239. The Hague: SPB Academic Publishing BV.

Lyman CP (1984) Pharmacological aspects of mammalian hibernation. *Pharmacological Therapy* 25: 371–393.

Lyman CP, Willis JS, Malan A, and Wang LCH (1982) *Hibernation and Torpor in Mammals and Birds*. New York: Academic Press.

Malan A and Canguilhem B (1989) *Living in the Cold II*. Montrouge (France): Colloque INSERM/John Libbey Eurotext Ltd.

Wang LCH (1985) Mammalian hibernation. In: Morris GJ and Grout B (eds.) *The Effects of Low Temperature on Biological Systems*. London: Edward Arnold.

Hippocampus

R L Isaacson, Binghamton University, Binghamton, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

The word 'hippocampus' is used in several different ways. One way is to describe portions of the nervous system of vertebrates that meet certain anatomic requirements, including the presence of 'double' pyramidal cells and interconnections with the septal area and the hypothalamus. In mammals, the area of the hippocampus is an anatomically 'primitive' cortical structure lying largely within the lateral ventricles of the brain. This intraventricular anatomic area is called the 'hippocampus proper' or 'Ammon's horn,' a term coming from the appearance of the dissected hippocampal formation in the human in which two large and distinctive areas spread out into the temporal lobes after meeting near the middle of the brain. This structure vaguely resembles the horns of a ram. Probably at some time in the Middle Ages or earlier, the name of the Egyptian god with a ram's head, Ammon, was given to what is now called the hippocampus.

Another way that the word hippocampus is used is as a loose description of a set of anatomic regions closely connected to the hippocampus proper and whose members have functional interactions with each other and with other brain regions. The best descriptive term for this collection of interconnected brain regions is really the 'hippocampal formation.' When the term hippocampus is used in this article, it refers to the hippocampus proper 'and' the dentate gyrus. Combining these closely related regions seems the simplest approach, even though some authors tend to consider the dentate gyrus as separate from the hippocampus.

In the mammalian brain, the hippocampus is prominent as a ridge along the internal edges of the lateral ventricles. These ridges occupy substantial portions of the lateral ventricles. In the human, the hippocampus is prominent from the anterior termination of the lateral ventricles near the most forward part of the temporal lobes to the merger of the lateral ventricles at the intraventricular foramen. Because of its large size and dorsal extension toward the center of the brain, the hippocampus should not be considered to be only a temporal lobe structure. This is significant because even the most extensive temporal lobe resections in the human that are undertaken for the relief of intractable epilepsy involve only the most anterior portions of the hippocampus, although they involve substantial portions of the hippocampal formation. A photograph of the human hippocampus as seen in a section through the temporal lobe is shown in Figure 1. Note the presence of substantial amounts of choroid plexus that exist in the lateral ventricles.

The hippocampus is a layered structure. Usually, it is considered to have between three and five layers of cells and fibers, depending on the criteria used for distinguishing layers. Myelinated axons cover its ventricular surface. As the hippocampus is exposed through the lateral ventricle during surgery, its surface appears shiny white because of the myelin coating of the axons on its outer surface (the alveus). Most of these axons are efferent from the hippocampal formation and combine to form the fimbria-fornix as they leave the surface of the hippocampus. The more compact tract of the fornix arises from the fimbria as the hippocampus diminishes in size near its dorsomedial termination above the thalamus. In this area, neuronal fibers from each side cross to the other in the hippocampal commissures.

In addition to the major portions of the hippocampus found adjacent in the lateral ventricle, in most mammalian brains, there are a few cells with 'hippocampal characteristics' found in the cortex below the genu of the corpus callosum (the anterior hippocampal rudiment) and a few such cells (the indusium griseum) scattered over the dorsal aspects of the corpus callosum. These areas are connected with the main parts of the hippocampal formation by a small group of fibers that are in, or stretch over the top of, the corpus callosum (the dorsal fornix and the stria of Lancisi).

Internal Structure

The most obvious internal features of the hippocampus are rows of densely packed neurons that are readily apparent in histological sections stained with common Nissl dyes or with noble metals. The cells in these rows have a characteristic appearance in many species. In mammals, the compact rows of cells that define the hippocampus proper begin suddenly at the boundary with the allocortex that forms the transition to the neocortex. Indeed, the areas that contain these distinctive rows of densely packed cells define the hippocampus proper. Other well-defined lines of cells are apparent in the granule cell layers of the dentate gyrus.

The prominent bands of closely packed cells that define the hippocampus proper and the dentate gyrus exist in rows or bands and range from three or four to more than ten cells deep. The width of the bands is

Figure 1 A photograph of a section through the side of the human brain at a level that cuts through the midsection of the temporal lobe. The hippocampus as it exists at this level is shown in the lateral ventricle and is marked by the black arrows. Just above the hippocampus a portion of the choroid plexus of the lateral ventricle can be seen. Adapted from an original photograph by Donald Dewsbury, University of Florida.

Figure 2 A drawing of the bushy dendrites processes of a 'double pyramidal' cell found in prominent rows through the CA1, CA2, and CA3 areas of the hippocampus. The cell body lies in the narrow middle of the dendritic arborizations that extend toward the outer surface and toward deeper layers of the hippocampus. Adapted from Freund TF and Buzsaki GY (1996) Interneurons of the hippocampus. *Hippocampus* 6: 347–470.

related both to the location within the hippocampus and to the species being examined. There is some evidence that there are 'gap junctions' (physical channels) connecting the internal contents of at least some adjacent cells. In the hippocampus, the large, principal cells found in distinctive bands are called pyramidal cells. Many of them differ from pyramidal cells found in the neocortex or other areas in that many of them have bushy dendritic arborizations extending from both the top and the bottom of the cell. These dendritic processes extend toward the surface and depths of the structure at right angles to the direction of the cell layer; they are called 'double pyramidal cells' (Figure 2).

The so-called granule cells of the dentate gyrus, the principal cells of that area, do not resemble the single or double pyramidal cells of the hippocampus. They are smaller, tightly packed cells arranged in a layered formation. These cells have dendritic 'bushes' that extend in only one direction: toward the outer, molecular layer of the gyrus. An illustration of a dentate granule cell is shown in Figure 3. Recently, multipotential cells, possibly stem cells, have been reported localized beneath the granule cells of the hippocampus. There is strong evidence that some of these cells have the ability to divide and differentiate in new neurons throughout life. They appear to differ from granule cells formed at the normal developmental period after birth in that have a rather limited life span. Their functional significance can not be determined at this time.

In mammals, the granule cells of the dentate appear as a U formation at the end of the band of hippocampal pyramidal cells called 'the hilar region.' The hippocampal pyramidal cells and interneurons become diffusely distributed in the middle part of the dentate gyrus.

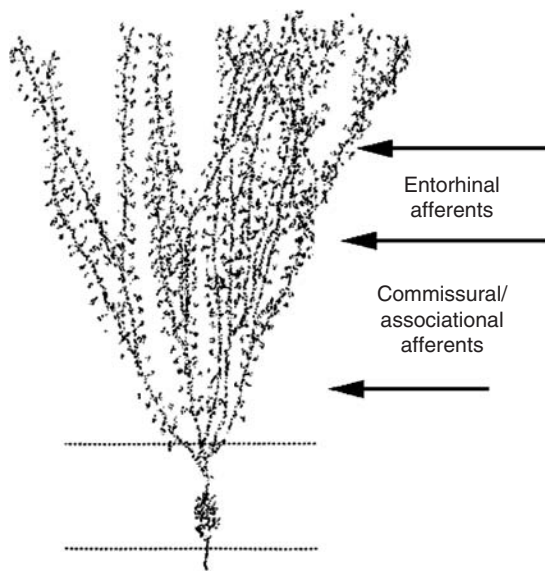

Figure 3 A schematic drawing of a granule cell of the dentate gyrus drawn to emphasize the large number of synaptic contacts made on the dendritic arborizations that extends toward the molecular layer. As can be seen, the dendritic tree extends only in one direction. The areas reached by afferents from associational areas, including from related areas on the opposite side of the brain, and cells from the entorhinal indicate areas are indicated by arrows on the right side of the figure. The two broken horizontal lines near the bottom of the figure represent the approximate area of the granule cell layer. Adapted from Freund TF and Buzsaki GY (1996) Interneurons of the hippocampus. *Hippocampus* 6: 347–470.

The Labeling of Cytoarchitectural Areas in the Hippocampus

There have been many attempts to provide a useful nomenclature for the areas of the hippocampus and the dentate gyrus. Perhaps the simplest subdivision of the hippocampus is a division into a superior and an inferior region. The superior region begins with a compact band of cells that originates in the boundary area between the subiculum and the hippocampus. Progressing away from the transitional cortex, the band of pyramidal cells becomes somewhat broader and the cells a bit larger and more scattered. This marks the beginning of the inferior region, which continues until the band of cells breaks up into a diffuse group of cells in the middle of the dentate gyrus.

The anatomist Lorente de Nó divided the hippocampal pyramidal cell systems into four subfields. He gave each the prefix CA to indicate they were part of Ammon's horn (*cornu ammonis* in Latin): CA1, CA2, CA3, and CA4. In his system, the superior region is designated CA1, the inferior region CA3. Between CA1 and CA3 is a small region designated CA2. Region CA4 is the area of more diffused patterned large cells found in the open U of the dentate gyrus after the strict laminar pattern of pyramidal cells in CA3 has disappeared. This area is also known as the hilus (or hilar) region of the dentate gyrus.

There has been some question as to whether a CA2 region should be considered as distinct from the adjacent CA1 or CA3 areas. In stained sections of the adult rodent brain, it is difficult to detect a CA2 segment. However, ontogenetically, the CA2 region develops 2 to 3 days earlier than CA1 and CA3. It also shows a higher resistance to destruction by the administration of certain toxins than the adjacent CA areas. This implies that the CA2 area has some special qualitative or quantitative neurochemical differences from its neighbors.

Alternative descriptions of hippocampal areas have been offered from time to time, some of which are still used. One of these is that of the anatomist, M Rose, who in the late 1920s described the hippocampus in terms of 'h' areas. In his terminology, h1 is equivalent to CA1 and to a small portion of the transitional cortex leading into CA1, h2 is equivalent to CA2 and some adjacent regions of CA3, h3 and h4 to remaining regions of CA3, and hs to CA4. The h system is sometimes used by neuropathologists who use it to define the regions affected by damage due to epilepsy or ischemia. This usually involves the CA1 area and some nearby transitional cortex. This damage-susceptible region is sometimes called Sommer's sector. This term refers to the h1 area that includes CA1 and, in addition, some parts of the presubiculum and subicular regions. Some of these designations are illustrated in Figure 4.

The two limbs of the U of the dentate gyrus have been given separate names. Usually, that part of the U closest to the hippocampal fissure is called the 'suprapyramidal limb,' and the other limb, the 'infrapyramidal.' The hippocampal formation folds back on itself so that the outer (molecular) layer of the structure is continuous with the outer, molecular layer of the neocortex. This can be seen by following the outer, molecular layer (layer 1) in Figure 4 into the hippocampus. Note that at the hippocampal fissure, it merges with the molecular layer (stratum moleculare lacunosum) of the hippocampus. This is really the 'outer layer' of the hippocampus. Correspondingly, the stratum oriens is the deepest layer.

Regardless of the designations used, physiological differences are also found in regions of the hippocampus. Cells in the several CA areas and in the the two blades of dentate gyrus differ in their sensitivity to ischemia and certain neurotoxins, nature of membrane ion channels, responsivity to steroid hormones, and content of several types of neuron-related growth factors. Biochemical differences between areas also exist. As one example, immunocytochemical and *in situ* hybridization techniques indicate protein kinase Cα is more heavily concentrated in the pyramidal neurons of CA2 and CA3 than in CA1 pyramids, whereas the opposite profile exists for protein kinase Cβ. Many other genomic and biochemical differences have been noted between various hippocampal areas.

Distinctions within the Transitional Cortical Areas

The subiculum and the presubiculum are adjacent to the CA1 field of the hippocampus. These areas have a simple anatomic structure that becomes increasingly more complex at greater distances from CA1. Although some anatomists have made further subdivisions among subicular regions, the significance of these differences is doubtful. Farther away from CA1 and the subiculum, the cortex becomes increasingly laminated. This marks the beginning of the presubiculum. The presubiculum merges with the retrosplenial cortex (behind the caudal termination of the corpus callosum, the 'splenium') that is found between the presubiculum and the parasubiculum. The latter merges with the entorhinal area. As this transition occurs, the layers of the neocortical tissue become much more distinct. This region also contains a very dense fiber plexus in layers II and III.

Species Variations in Hippocampal Components

The cells in the various subfields of Ammon's horn and transitional cortical areas vary in amount and

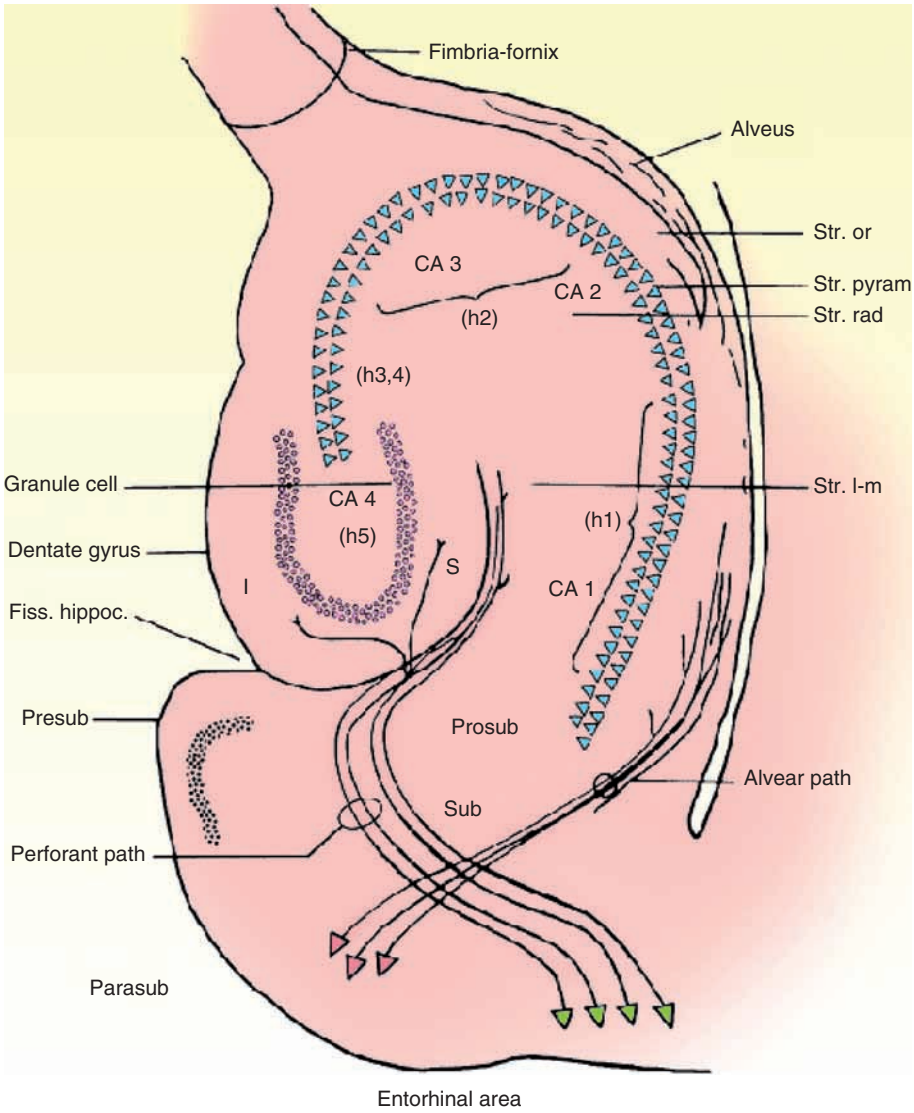

Figure 4 A schematic diagram of a horizontal cross-section of a rat brain showing details of the hippocampal formation. Areas CA1, CA2, CA3, and CA4 refer to the designations of Lorente de Nó. The numbers in parenthesis with a preceding 'h' refer to subareas of the hippocampus used by M Rose. The 'S' and 'I' in the dentate gyrus refer to the superior and inferior blades of the granule cell rows. Str. l-m, stratum lacunosum-molecular; Str. rad, stratum radiatum; Str. or, stratum oriens; Str. pyram, pyramidal cell layer in the hippocampus; Parasub, parasubiculum; Presub, presubiculum; Prosub, prosubiculum; Sub, subiculum; Fiss. hippoc, hippocampal fissure. Alvear and perforant pathways are also shown. Adapted from figure 10-4 in Brodal A (1969) *Neurological Anatomy,* 2nd edn. Oxford: Oxford University Press.

shape at different places along the extent of the hippocampus. These changes in character and distinctness are altered over the course of development. There is also considerable variation in appearance and size of the cells among animal species. For example, the CA1 region in the human makes up about 50% of the volume of the hippocampus, whereas it makes up only about 25% of hippocampal volume in the rat. Because the number of synapses made in a cortical area is closely tied to the volume of that area, this suggests that the CA1 may be of greater functional significance in the human than in the rat. In primates, the main source of neuronal input to CA1 arises from neocortex, especially from areas closely related to the processing of sensory information. Despite quantitative differences, the primate CA1 retains essentially the same types of afferents from subcortical regions as are found in the rat. Cortical input to the rat CA1 is restricted to afferents from entorhinal and perirhinal transitional cortical areas. The suggestion is that input to the rodent CA1 is weighted toward information from subcortical systems. The relative expansion of the primate CA1 may reflect in increase in the amount of information coming from the 'higher centers' of the brain, that is, the neocortex.

The projections from the CA1 area are quite diverse. They include axons that reach the subiculum, the entorhinal area, the lateral septal area, portions of nucleus accumbens, the olfactory bulb, and several limbic areas of the forebrain.

Interneurons

Neurons that cannot be classified as pyramidal cells in the hippocampus or as granule cells in the dentate gyrus generally are grouped together as interneuron's. Almost all of these cells synthesize γ-aminobutyric acid (GABA) and use it as their primary transmitter substance. Their terminals generally end on GABA receptors that have inhibitory actions on the recipient cells. This has led the to the use of an alternate descriptive term for interneurons: 'GABAergic nonprincipal cells.' Within the past few years, it has become apparent that these cells play crucially important roles in all hippocampal activities and in activities in many other brain areas as well. The wide rage of interneuron cell types and their many extensive and diverse patterns of dendritic and axonal projections require a re-evaluation of hippocampal functions. Examples of the diversity of such projections are shown in Figure 5.

In the rat, and likely in other species as well, about 10% of the neurons in the hippocampus are interneurons. These cells exert inhibitory influences on the pyramidal cells in the structure of both nearby cells and othrers at a considerable distance. It seems likely that a significant number of axons from the interneurons extend beyond the hippocampus proper into areas of the hippocampal formation and beyond. Interneurons can be found throughout the hippocampus. Although many are diffusely scattered throughout the hippocampus, they also can be found mixed in with principal cells in all the CA and dentate granule cell regions. Although interneurons make up only 10% of the neurons in the hippocampus, the physical extent of the distributions of their neural processes and the number of synaptic contacts made by each cells makes their influence of staggering importance. For example, some axons of interneurons of the CA3 area have axons that extend to more than 50% of the hippocampus and make contact with between 20 000 and 60 000 other cells.

Some of the interneuron axons make contact with the initial segments of axons arising from principal cells in the hippocampus. These are called 'chandelier cells.' Their axons form synaptic contacts with the axons of principle cells after running parallel to these axons for some distance. These cells help regulate the axons of the principal cells and by so doing influence their output. Another major group of interneurons are basket cells. The processes of these cells surround the soma and basal dendrites of other cells, both principal cells and other interneurons. Axons of types of interneurons also access the cell bodies and portions of dendrites of other interneurons.

Interneurons are remarkable in their biologic makeup as well. Although virtually all synthesize and utilize

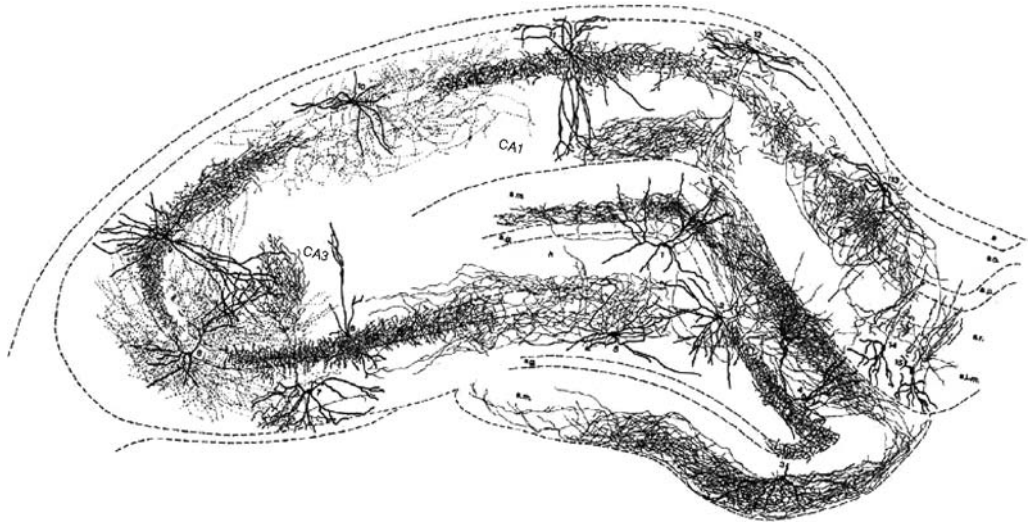

Figure 5 A drawing of the hippocampus and 15 interneurons of different types. It shows the extensive distributions of axonal and dendritic arborizations of these cells and their extensiveness through many different areas of the structure. The lines representing dendrites are thicker than lines representing axons. a, alveus; s.o., s.p., s.r., s.l-m, strata oriens, pyramidale, radiatum, and lacunosum-moleculare of the hippocampus; s.g, s.m, h, strata granulosum, moleculare, and hilus of dentate gyrus. The CA subfields are as described in the caption for Figure 4. From Freund TF and Buzsaki GY (1996) Interneurons of the hippocampus. *Hippocampus* 6: 347–470. figure 18, p. 382.

GABA for transmission purposes, many other proteins and peptides are colocalized with them. These include several types of calcium binding proteins, neurotrophic factors, neuronal nitric oxide synthase, and many of the peptides known to have importance synergistic activity with neuronal transmitters throughout the brain (e.g., cholecystokinin, substance P, and the enkephalins, somatotropin, vasoinhibitory peptide, and neuropeptide Y). Interneurons have receptors for many different transmitters, including muscarinic and nicotinic receptors for acetylcholine.

One mystery that remains unresolved is how can small groups of neurons in the brain have such profound and wide effects throughout the structure. For example, the cells of origin of all of the biogenic amines are relatively few and are generally localized to relatively small areas. Despite this, they influence activities throughout the brain. It is possible that they do so because their primary targets in the brain are interneurons, including those of the hippocampus, whose neural processes extend so widely. Although the importance of inhibitory neurons in the brain has been discussed for many years, it is only recently that the actual anatomy and physiology of the systems has been elaborated.

A systems Analysis

The Temporoammonic System

The entorhinal cortex is an important source of neocortical input to the hippocampus and provides most of the neocortical input to the hippocampus in rodents. Neuronal pathways originate from cells in all major sensory regions of the neocortex and converge in the entorhinal region, albeit by indirect routes. The entorhinal cortex may be thought of as modulating hippocampal activity on the basis of neocortical activities. Efferents from the lateral entorhinal areas reach the outer portions of the dendrites in the molecular layer of the dentate gyrus. Efferents from the medial portions of the entorhinal cortex reach the middle zones of the dentate. These entorhinal-dentate gyrus connections are called the 'perforant paths.' The axons in the perforant paths are generally glutaminergic and excitatory in function. Axons from the granule cells, in turn, give rise to the 'mossy fiber' system, which is made up of thin axons arising from the granule cells of the dentate gyrus that terminate on the proximal dendrites of neurons in the CA3 region. Axons from cells in the CA3 area leave the hippocampus via the fimbria and fornix, largely terminating in the lateral septal area. However, branches of these efferent axons form the Schaffer collateral system that project to dendrites of the CA1 neurons. Neural outputs from the CA1 cells then exit into the alveus and join the fornix, projecting on the one hand to the septal area and, on the other, to the subicular region. The major output of the hippocampal formation over the fornix system to the forebrain areas arises from subicular areas. Efferent fibers in the fornix distribute to many locations including the nucleus accumbens, the anterior and lateral nuclei of the thalamus, the lateral and preoptic hypothalamic areas, the mammillary bodies, the arcuate and ventromedial hypothalamic nuclei, and the bed nucleus of stria terminalis.

This description of the passage of activity from the entorhinal area through to the subiculum is much oversimplified. For example, some axon collaterals from the granule cells of the dentate and from principal cells in CA3 and CA4 actually turn back to reach dendrites in the outer portions of the molecular layer of the dentate gyrus. There is much yet to be learned about the finer connections among the principal cells.

The principal cells in the CA4 region of the hippocampus, the hilar area, are scattered throughout the area within the U formed by the blades of the dentate gyrus. Many of the larger cells play an important role in modulating activity in the dentate gyrus and in the CA3 region. They are thought to be 'misplaced' pyramidal cells. They are glutaminergic and produce excitatory effects on neurons with which they are connected.

Activity arising from neurons in the entorhinal cortex may or may not progress in a linear fashion through these hippocampal circuits. Whether information reaching the dentate gyrus gets forwarded to the next stations, such as CA3 and CA1, depends on the activity occurring in the hippocampus when the afferent signals arrive. Transmission through this multisynaptic circuit seems least impeded during slow-wave sleep but is restricted at particular points along the pathway during waking states and rapid eye movement sleep. Transmission through the hippocampal formation depends in great part on how each 'station' along the way is being modulated by the diverse influences acting on them.

The Hippocampus as a Chemical Sensor

It is likely that the hippocampus evolved from the same primary sensory epithelium as did the olfactory system, although in living mammals, it is physically remote from the primary odor recptors. Some investigators have suggested that the hippocampus may have turned from the detection of air-borne chemicals to those carried by the certain fluids of the body, including the blood and the cerebrospinal fluid. It is a structure that is well suited for the detection of

changes in the makeup of cerebrospinal fluid because it lies within the lateral ventricle and close to the choroid plexus. It is possible that the hippocampus can detect such chemical changes and can initiate events that are likely to restore an optimal composition by influencing biologic and behavioral actions. Supporting this hypothesis is the fact that receptors for more than 60 hormones, peptides, metabolites, and other chemical ligands of importance have been found in the hippocampus. It has more of these sorts of receptors than any other brain area.

Modulation of Hippocampal Activities

Modulation of activities in the hippocampus and in the 'neural chain' that begins at the entorhinal cortex and runs through the CA areas to the subiculum may occur through influences mediated by the diffusely projecting biogenic amine systems, afferent cholinergic influences, and both localized and systemic neuropeptide and hormonal systems. There are vast numbers of potential modifying influences. Norepinephrine, serotonin, histamine, VIP, cholecystokinin, substance P, somatostatin, the enkephalins, and the hypocretins influence neuronal activity in the hippocampus. In addition, a significant dopaminergic input arising primarily from the ventral tegmental area reaches the hippocampus and the entorhinal cortex.

The substantial norepinephrine projections received by the hippocampus come from cells in the anterior part of the locus coeruleus. These cells are the sole source of the noradrenergic input to the hippocampus, but collaterals of these locus coeruleus cells reach other areas, including the neocortex, thalamus, and spinal cord. Some cells in this nucleus project bilaterally to the hippocampus. The serotonergic fibers that reach the hippocampus arise in large part from the median raphe cell group of the brain stem. Some fibers also reach the hippocampus from the dorsal raphe. These fibers terminate in the dentate gyrus and reach the molecular layers of the CA1 and CA3 to a greater extent than do noradrenergic fibers.

Septohippocampal Interactions

The hippocampal formation and the septal area are reciprocally connected. Principal neurons of the medial septal area and the diagonal band of Broca project to the hippocampus, and neurons of the hippocampus project to the lateral septal area. The projections from the septal region and the diagonal band to the hippocampus are predominately cholinergic and GABAergic. Both are frequently co-localized with various peptides and calcium binding proteins. Based on choline acetyltransferase histochemistry, the cells containing acetylcholine that project to the hippocampus are located in the medial septal area and both the vertical and the horizontal limbs of the diagonal band of Broca. There is a topographic pattern to this projection such that axons originating from the medial septal nucleus and the ventromedial diagonal band enter the fornix and the fimbria, whereas axons originating in the dorsolateral diagonal band reach the hippocampus via the supracallosal stria. The former axons terminate in the anterior three quarters of the hippocampus, and the latter axons terminate in the temporal regions of the hippocampus. Additionally, it is likely that some axons from both pathways terminate in the subiculum and the entorhinal cortex. Although differences in projection patterns exist between individuals and among species, the cholinergic cells of the basal forebrain should probably be regarded as part of a single system, the medial septal area and the diagonal band being continuous with the cholinergic cells of the ventral pallidum and the lateral hypothalamus. To a limited extent, these regions also contribute to the cholinergic innervation of the hippocampus. It is generally agreed no neurons in the hippocampus synthesize acetylcholine. Both the cholinergic and the GABAergic projections to the hippocampus significantly influence the electrical rhythms generated in the hippocampal formation and the overall excitability of the neurons in the structure.

One of the most intensively studied biologic characteristics of the hippocampus, and one that likely exerts a strong modulatory influence on its functions, is the binding of corticosteroids to membrane, cytoplasmic, and nuclear sites of its cells. Both mineralocorticoid and glucocorticoid receptors are found in large numbers in cells in the hippocampus. The net effect of the adrenal steroids on hippocampal function likely depends on the relative occupancy rates of the two receptor types. Adrenalectomy results in impaired performance in behavioral tests, some of which are similar to impairments found after hippocampal damage. However, adrenalectomy produces a highly unusual condition in which both glucocorticoids and mineralocorticoids are largely or entirely absent. Behavioral and functional anomalies are found with both abnormally high and abnormally low adrenal corticoid hormone levels.

Electrical Rhythms of the Hippocampus

In rodents and certain other species, a number of different types of electrical activities can be recorded

from the hippocampus. Probably the most intensively studied hippocampal electrical rhythms are those of steady oscillations that are very regular and of high amplitude. Rhythmic slow activity (RSA) in the rat that varies from 5 to 10 Hz is known as the theta rhythm. This frequency RSA is not in the frequency range of the theta rhythm described in the human. In people, theta rhythms are considered to be oscillations in the 4–7 Hz range. In the rat, theta rhythms, as well as the faster beta (15–50 Hz) and gamma rhythms (40–100 Hz), are observed during directed movements and during rapid eye movement sleep. During periods of immobility, consummatory behaviors and slow-wave sleep, bursts of very rapid neuronal discharges occur that can be called 'sharp-waves' and abbreviated as 'SPW.' Recently, ultrafast rhythms (>60 Hz) in dense dentritic areas of the hippocampus have been subjects of investigation. The oscillations of electrical rhythms in the hippocampal neurons are controlled by the activity of networks of inhibitory interneurons. In addition, ultra-slow oscillations in the excitability of hippocampal neurons have been reported in the rat, and similar changes have been noted in some human patients with epilepsy. Peak-to-peak periods of 40–50 s have been observed. Evidence from animal studies indicates these slow changes originate in neocortical areas and are projected to the hippocampus by relays in the entorhinal cortex.

A significant pacemaker for RSAs of the theta frequency in rats comes from the medial-basal forebrain cholinergic systems. However, although medial septal lesions abolish RSAs, not all RSAs in the freely moving animal depend on cholinergic activities. Atropine, for example, abolishes RSAs only under certain circumstances. A separate, noncholinergic source of RSAs probably comes from the entorhinal cortex. It is possible that some RSAs are produced by cholinergic fibers that are directly exciting inhibitory interneurons, whereas entorhinal input directly excites interneurons of the hippocampus to produce RSAs.

At one time, it was thought that certain frequencies of rhythmic activities were closely tied to certain behaviors being exhibited by the animals. However, it now seems that the rhythmic activity of different forms can occur during various behaviors and probably are a method of coordinating activity among the principal neurons of the hippocampus and among various neuronal nets of the interneurons. However, any universal association between electrical activities and behavior is questionable because many species differences exist in the occurrence of RSAs. They are seldom observed in humans or nonhuman primates but are quite variable among members of some species.

Functional Considerations

Based on observations of patients with temporal lobe epilepsy and especially since the late 1950s, it has been popular to associate damage to the hippocampus with some forms of memory disturbances. This is due in large part to the remarkable recent memory deficiencies found in a patient of Dr. William Scoville, case H.M. While in his 20s, this patient underwent bilateral temporal lobe extirpations for the relief of intractable temporal lobe epilepsy. Ever since, he has demonstrated an extreme form of recent memory loss, a loss far more extreme than that found in any other patient who has undergone temporal lobe surgery for the relief of this disorder. Recent memory loss and some forms of amnesia after temporal lobe damage is most likely to be the consequence of damage to temporal lobe tissue medial and ventral to the hippocampus. Damage to the hippocampus, itself, may not be the critical insult. The actual lesion to the hippocampus in case H.M. was far less than originally reported, and equivalent damage in at least one other patient did not produce a similar memory impairment. Studies with nonhuman primates also indicate that destruction of the hippocampus and/or the amygdala also fails to produce memory anomalies, whereas damage to parahippocampal regions does.

Despite current beliefs associating the hippocampus with memory, either in information storage or in retrieval, a necessary association between the hippocampus and memory is not likely in people or in experimental animals. In rats, the idea that the integrity of the hippocampus is essential for the effective use of spatial information in certain kinds of test situations has been popular for many years. Nevertheless, after complete destruction of the hippocampus in rats, the animals are able to use many kinds of spatial and contextual information for the attainment of desired goals. Still, it is certain that there will be a continuing discussion of the role of the hippocampus in the use of acquired information for many years to come. What is certain is that no single theory has been found to adequately explain all of the available experimental and clinical results.

At the present time, the functional contributions of the hippocampus to mental and behavioral phenomena remain obscure. As we learn more about the functional contributions of the neocortex, basal ganglia, and other neural structures, the contributions of the hippocampus may become clearer.

See also: Entorhinal Cortex; Gamma-Aminobutyric Acid (GABA); Glutamate Receptor Organization: Ultrastructural Insights; Hippocampus and Neural Representations; Hippocampus: Computational Models; Hippocampus: Molecular Anatomy; Intermediate Filaments.

Further Reading

Aggleton JP and Brown MW (1999) Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral Brain Science* 22: 425–489.

Alyan SH, Jander R, and Best PJ (2000) Hippocampectomized rats can use a constellation of landmarks to recognize a place. *Brain Research* 876: 225–237.

Anderson MF, Aberg MA, Nilsson M, and Eriksson PS (2002) Insulin-like growth factor-I and neurogenesis in the adult mammalian brain. *Brain Research Developmental Brain Research* 134: 115–122.

Brodal A (1969) *Neurological Anatomy,* 2nd edn. Oxford: Oxford University Press.

Corkin S, Amaral DG, Gonzales RG, et al. (1997) H. M.'s medial temporal lobe lesion: Findings from magnetic resonance imaging. *Journal of Neuroscience* 17: 3964–3979.

Freund TF and Buzsaki GY (1996) Interneurons of the hippocampus. *Hippocampus* 6: 347–470.

Isaacson RL (2002) Unsolved mysteries: the hippocampus. *Behavioral and Cognitive Neuroscience Reviews* 1: 87–107.

Lamas M, Gomez-Liva G, and Gutierrez R (2001) Vascular GABA transporter mRNA expression in the dentate gyrus and in mossy fiber synaptosomes. *Molecular Brain Research* 93: 209–214.

Murray EA and Mishkin M (1998) Object recognition and location memory in monkeys with excitotoxic lesions of the amygdala and the hippocampus. *Journal of Neuroscience* 18: 6568–6582.

Hippocampus and Neural Representations

E I Moser and M-B Moser, Norwegian University of Science and Technology, Trondheim, Norway

© 2009 Elsevier Ltd. All rights reserved.

Discovery of Place Cells

Hippocampal representations have been associated with bodily spatial location almost since the first single cells were recorded from this brain area. In 1971, O'Keefe and Dostrovsky reported that a major proportion of complex-spiking neurons in the rat hippocampus had spatial receptive fields. These 'place cells' fire when the rat is in a particular place in its environment (the 'place field' of the cell), but are virtually silent when the rat is in other positions (Figure 1). Firing rates in the place field are often 10–100 times higher than the background rate. Place fields reflect the animal's location as such and not the behaviors emitted at these locations. Different cells have place fields at different places and collectively they cover the entire test environment. Unlike neurons in the sensory cortices, place cells in the hippocampus have nontopographic firing correlates: place fields of neighboring cells are apparently not closer than those of distant cells, and all places seem to be represented within local place cell ensembles in any part of the hippocampus. The size of the place fields appears to increase, however, as recording positions shift from the septal to the temporal pole of the structure.

The discovery of place cells led O'Keefe and Nadel to suggest that place cells form a distributed maplike representation of the spatial environment and that the animal can use this 'map' for navigation. Spatial firing in the hippocampus was thought to reflect the integration of inputs from a general metric navigational system, possibly outside the hippocampus, with stimuli that were specific to the environment. The metric system might maintain a dynamic representation of the animal's position by constantly integrating the distance and direction of its own movement, a process now referred to as path integration. A number of studies from McNaughton and others have supported a role for path integration in the control of firing location in hippocampal place cells. They showed that place fields were, under some conditions, determined primarily by the animal's movement. When a rat was released from a movable start box at an unpredictable location on a linear rail with a fixed goal at the end, place cells fired in relation to the distance from the start box for the first few seconds, before the external visually dominated reference frame took over. These studies imply that output from a path-integration mechanism is sufficient to drive place-selective firing in the hippocampus, although the place-field map is normally calibrated against the external landmarks of the environment. The locations and mechanisms of these proposed computations were not known at the time when O'Keefe and Nadel proposed their 'cognitive map' theory for the hippocampus. An important step toward the identification of the underlying computations was taken in 1984, when Jim Ranck reported the discovery of another component of the brain's navigational system, the head-direction cell, in the dorsal presubiculum. Its extrahippocampal location suggested that spatial representation and navigation might depend on a widespread brain circuit.

Place Cells – More Than Place

During the decades that passed after the discovery of place cells, it became apparent that hippocampal neurons respond to a number of influences, some of which are not associated with particular locations in the environment. Eichenbaum and colleagues showed that hippocampal neurons may fire in response to specific odors, regardless of the location where the odors were encountered, and Deadwyler and colleagues found hippocampal neurons that fired only during specific phases of an operant delayed-nonmatching task, even when the rat was more or less in a constant place. Many of these neurons may have exhibited place-specific firing in addition, but the studies clearly suggested that nonspatial information was represented as well. These findings implied that the hippocampus must represent experience more broadly, with spatial location being only one dimension of the memory stored in the hippocampal network.

Place cells do not merely mirror the spatial and nonspatial sensory input that the animal receives. At least three types of observations point to a role for intrinsic dynamics in the control of hippocampal firing patterns. First, place fields develop slowly. When a rat enters a new environment, place fields stabilize only after some minutes, at a much slower rate than the rate at which sensory information about the specific features of the environment is passed to the hippocampus. Second, changes in place fields cannot be predicted from the amount of change in the physical properties of the environment. Place fields are often maintained after significant landmarks are removed

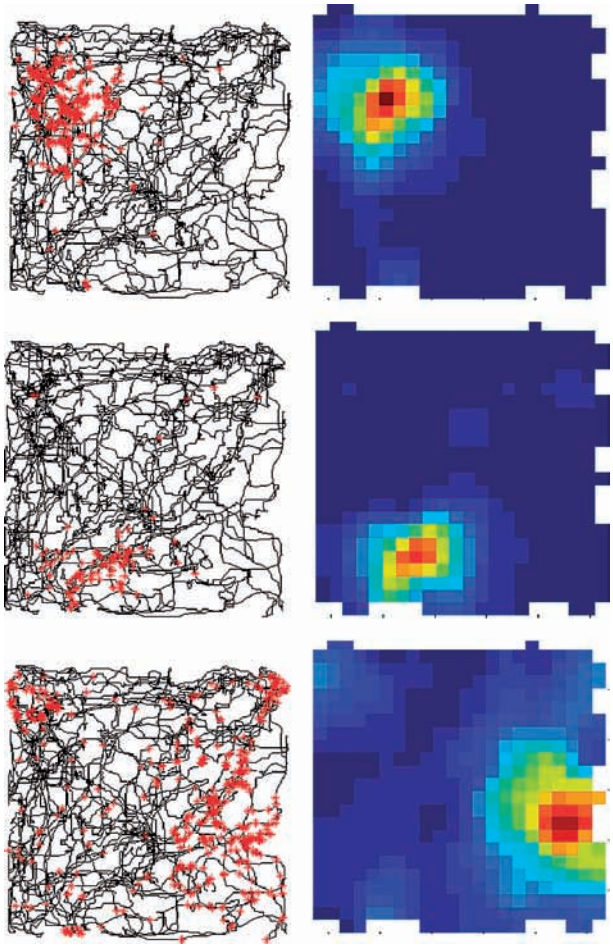

Figure 1 Place cells in the hippocampus. The firing fields of three place cells are shown, either as maps of the recording enclosure with spike locations (red) superimposed on the animal's trajectory (black) (left), or as color-coded rate maps in the same enclosure (right). The color scale is from blue (silent) to red (peak rate). Peak rates are 10–20 Hz.

from the environment, and if an animal is first trained in two different environments, and one of them is then smoothly transformed into the other, animals quite often keep the representation from the first environment throughout the test sequence, suggesting that firing is determined more by the history of testing than by the actual sensory stimuli impinging on the animal. Third, different locations in the environment are represented by different numbers of active cells; some places are associated with more neuronal firing than are others. In enclosed chambers, place fields appear to be somewhat more common near edges and walls than in the center, and in environments with distinct reward locations, a larger number of cells may have activity fields at the goal location than at other locations. Taken together, these observations suggest that hippocampal neurons do not only map the physical structure of the environment, but also express information about stimuli and events that the animal experiences or has experienced there.

Recent studies have raised the possibility that hippocampal cell assemblies, especially in the CA3 subfield, have different coding schemes for locations and stimuli or events associated with the locations. Global remapping, expressed as a complete redistribution of firing locations and firing rates among simultaneously recorded place cells, is nearly always induced when an animal is moved between two distinct locations. Under some circumstances, global remapping may be induced also at a single location, presumably when the defining features of the local environment change rather extensively. Rate remapping, in contrast, is expressed as a redistribution of the firing rates without an accompanying change in the preferred firing locations of the hippocampal neurons. This form of remapping is often induced after replacement of a salient cue in an otherwise constant environment. The presence of such apparently independent coding schemes suggests that spatial and nonspatial information may be expressed by the same neuronal populations. Changes in spatial location are represented as changes in firing fields, whereas changes in cues within an environment are represented by changes in firing rate. The combination and integration of spatial and nonspatial information in the output from the hippocampus may form the neural basis for the role of the this brain region in episodic memory.

Attractor Dynamics in the Hippocampus

One of the greatest challenges of the brain's episodic memory system is to recognize and encode a novel input without causing interference with other items that are already stored in memory. Each day we meet friends and colleagues in a variety of contexts, but we generally recognize them as the same individuals even when we do not see details because of poor illumination or when their identity is masked by sunglasses, a hat, or new clothes (Figure 2). With such an extraordinary ability to generalize, how can the brain at the same time identify an unfamiliar person as distinct from the wide variation around templates already stored in memory? Theoretical models have suggested that the right balance between fault tolerance and prevention of interference can be reached in a dynamic attractor-based type of neural network.

An attractor network has one or several preferred positions or volumes in state space, such that when the system is started from any outside location, it will evolve until it reaches one of the attractor basins. It will then stay there until the system receives new input. In a neural network with such properties, stored memories (the preferred positions of the

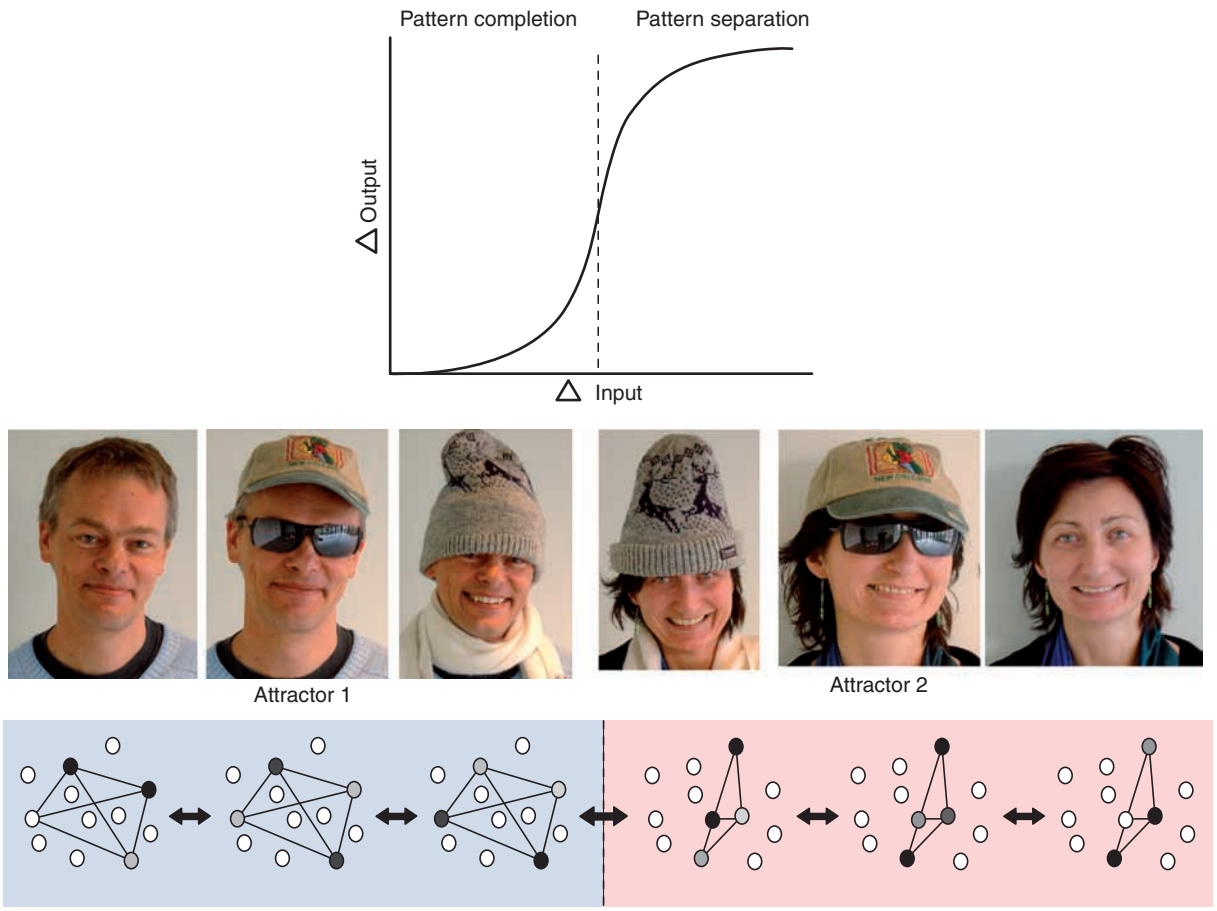

Figure 2 Schematic illustration of attractor dynamics. During retrieval, inputs to the hippocampus converge to one of several preestablished patterns, yielding a similar output for a range of input values (pattern completion). At the same time, they diverge away from other interfering patterns (pattern separation). The nonlinear nature of this process can be illustrated by the fact that multiple images of the same face are recognized as variants of the same person (one of the authors) but as different from images of another person (the other author), even when a number of cues are shared. The nonlinearity of the process (the absence of an intermediate representation) is consistent with the idea that the two faces are represented by discrete attractor states. A similar nonlinearity is thought to be reflected in the firing of hippocampal place cells, as illustrated by the sharp transition between ensemble representations under some test conditions. (Top and bottom rows) Reproduced from Leutgeb JK, Leutgeb S, Treves A, et al. (2005) Progressive tranformation of hippocampal neuronal representations in 'morphed' environments. *Neuron* 48: 345–358, with permission from Elsevier.

system) may be recalled from degraded versions of the original input. This is referred to as pattern completion. At the same time, the retrieved pattern is not mixed up with other memories (other preferred positions) in the same network. This reflects a process referred to as pattern separation. There is considerable evidence for both pattern completion and pattern separation in the hippocampus. Pattern completion is expressed in place cells as persistent firing in the cell's place field after removal of a significant subset of the landmarks that originally defined the environment. Taking away a cue card or turning off the room lights, for example, rarely causes much change in the firing locations of hippocampal place cells. Pattern separation is expressed in place cells as the tendency to undergo 'remapping' after only minor changes in the sensory input, such as a change in the color or the shape of the recording enclosure. Remapping is a property of the cell population; when remapping occurs, it is generally expressed in all place cells.

In agreement with attractor network models of the hippocampus, place cells show abrupt shifts between representations under conditions where inputs change to the point where a different representation provides a better match to the inputs experienced by the animal. O'Keefe and colleagues trained rats in a square and a circular version of a morph box with flexible walls, until place cell representations in the two environments were very different. The rats were then exposed to several intermediate shapes. A sharp transition from squarelike representations to circlelike representations was observed near the middle between the familiar shapes, as predicted if the place cell network had discrete attractor states corresponding to these

shapes. Discrete attractors may not be ideal for storing sequential information, however. If hippocampal networks consisted exclusively of discrete attractors, the ensembles would probably fail to encode the continuous stream of information that is characteristic of experience; the network might not be able to maintain a unified representation of the flow of experience if it fluctuated between preformed states whenever the input was intermediate to any set of previously stored conditions. Recent observations suggest that under conditions where the spatial reference frame is constant, place cells assimilate gradual or moderate deviations in the input into preexisting representations. Stable states can be attained along the entire continuum between two preestablished representations in a single context. This ability to represent continua may allow hippocampal networks to encode and retrieve consecutive inputs as uninterrupted, distinguishable episodes.

Anatomical Location of Attractor Networks

While hippocampal cell ensembles have been known for a while to exhibit some of the key features of an attractor network, the anatomical location of the attractors is just beginning to be unfolded. Attractor dynamics may originate in one of the densely interconnected networks of the hippocampal formation, or these properties may be imposed on hippocampal neurons by other brain areas. Somewhat paradoxically, until a few years ago, nearly all knowledge about hippocampal ensemble dynamics came from population recordings in the CA1 subfield, probably because of the relative ease with which this area could be accessed. Theoretical models of hippocampal function, however, were generally based on the neural architectures of the dentate gyrus and CA3, their dense and modifiable recurrent circuitry, their specific interconnections, and the sparse firing of principal neurons in these networks. Recordings in CA1 might pick up the results of neural computations in these subfields, but because place-selective firing in CA1 is also controlled by direct input from the entorhinal cortex, the use of CA1 as a window to the attractor dynamics of the hippocampus is somewhat questionable.

In the hippocampal formation, the most extensively interconnected associative network is found in CA3. Widespread collaterals interconnect pyramidal cells along nearly the entire length of the subfield. On average, each pyramidal neuron makes synapses with as many as 4% of the pyramidal cells in the ipsilateral CA3, and more than three quarters of the excitatory synapses on a CA3 pyramidal cell are from other CA3 pyramidal neurons. These recurrent synapses show long-term plasticity, allowing the formation of an extensive number of possible cell assemblies in the network. The possibility that attractor dynamics originates in CA3 is supported by a growing body of experimental evidence. The CA3 matrix performs both pattern completion and pattern separation. Pattern completion is apparent in CA3 pyramidal neurons as a remarkably persistent and coherent firing map after removal of significant parts of the original sensory input. *N*-Methyl-D-aspartate (NMDA)-receptor-dependent plasticity in CA3 is necessary for the ability to retrieve spatial memories from small subsets of the cues that originally defined a target location. Pattern separation is apparent in the same network as a nearly complete replacement of the active cell population when animals move between enclosures with common features. Representations in CA3 may undergo complete orthogonalization under conditions where place cells in CA1 retain considerable similarity.

Although the CA3 is by far the most densely interconnected structure of the hippocampal formation, recurrent connections exist also in other subfields. Granule cells in the dentate gyrus, for example, form synapses on the proximal dendrites of the hilar mossy cells, which in turn have excitatory synapses on the dendrites of the granule cells in the molecular layer. Pyramidal cells in the subiculum have extensive axon collaterals that terminate on other pyramidal cells across a substantial part of the subiculum. The degree of interconnectivity in these subfields, as estimated by the number of locally distributed axon collaterals and the density of varicosities on them, is lower than in CA3, however, and it is conceivable that CA3 is responsible for much of the attractor dynamics thought to underlie hippocampal memory.

The Origin of Pattern Separation

A critical step in the encoding of a new episodic memory is the minimization of similarities between the new representation and representations that already exist in the network. When this decorrelation, or pattern-separation, process fails, and the same retrieval cues get associated with multiple episodic memories, retrieval breaks down. Rats with lesions in the hippocampal formation specifically fail to discriminate between contexts or locations with different reinforcement histories when the test stimuli are highly similar. This suggests that a neuronal pattern-separation mechanism may be located in this brain region, but the origin within the hippocampal circuitry has not been determined.

The strong orthogonalization of place-field representations in area CA3 suggests that a major function

of the early stages of the hippocampal formation may be to enhance small differences in information coming into the hippocampus, such that memories can subsequently be stored with minimal overlap. However, although pattern separation is expressed in CA3, theoretical models have for a long time pointed to the dentate gyrus, one synapse upstream, as a more likely origin of pattern separation in the hippocampal formation. These models were, to a large extent, based on analogies between the granule cell layers of the hippocampal formation and the cerebellum. The circuit architecture of the cerebellum provided the first clues about how the brain might decorrelate or orthogonalize inputs. As proposed by David Marr almost 40 years ago, in the cerebellum, incoming sensory information is projected onto an expanded intermediate layer of granule cells before the information reaches the associative synapses of the Purkinje cells. The number of cerebellar granule neurons by far exceeds the number of the input neurons in the brain stem. This allows different incoming signals to be dispersed onto largely nonoverlapping cell populations, such that each granule cell carries only a small and distinct fraction of the total information passed into the network over a period of time. Combined with the sparse firing of the granule cells, the divergence is thought to enable efficient decorrelation of overlapping input patterns. A similar recoding mechanism was later proposed to exist for the hippocampus. Like in the cerebellum, input is spread out on an expanded intermediate layer of granule cells. The number of granule cells of the dentate gyrus is about 5 times larger than the number of afferent projection neurons in layer II of the entorhinal cortex (approximately 1000×10^3 vs. 200×10^3, respectively, in the rat). In addition, the firing of the granule cells is sparse and each granule cell makes synapses with only a limited number of CA3 pyramidal cells, suggesting that a combination of mechanisms might potentially contribute to decorrelation of cortical information in the dentate gyrus. Whether these hypothetical mechanisms represent the origin of pattern separation in the hippocampus remains to be determined.

Off-Line Activity in Place Cells

After the initial encoding of new information, memories are thought to undergo further consolidation off-line when the subject is engaged in other activities. Although much of this activity appears to take place during slow-wave sleep, the specific cascade of events that leads to consolidation of hippocampal memory is not well understood; little is known about whether consolidation has distinct stages, and, if is has, which brain areas and which neuronal computations are involved in each stage. Ensemble recordings from the hippocampus in sleeping rats provide some clues. Cells that are coactivated in the hippocampus during awake behavior exhibit correlated firing during sleep episodes subsequent to the behavioral session. The order of firing is similar but the sequence is replayed at a faster rate. This form of reactivation is temporally specific. It is strongest shortly after behavioral experience and decays during subsequent hours. Reactivation is associated with sharp waves, which are bursts of synchronous activity in hippocampal pyramidal cells during slow-wave sleep and awake rest. These bursts have the capacity to induce plasticity in downstream areas and may therefore potentially be involved in the consolidation of long-term memory within the hippocampus or in the neocortex.

Although reactivation during sleep may provide a mechanism for consolidation of recent memories, the delay between behavior and sleep is often too long to account for the fact that memories are maintained for hours or more independently of whether subjects get an opportunity to take a nap. Recent studies have indicated that replay may also take place during 'interleaved' sharp waves in the waking states, when the animal is resting, eating, drinking, or just stopping briefly as it is exploring the environment. During such 'interleaved' sharp waves, spikes are replayed in reverse order. When a rat is running on a linear track, for example, place cells fire in a certain order as the animal runs across the track. As it reaches the end of the track, where food is available and the rat stops to eat, sharp waves appear and the cells that were active on the track begin firing. The order of firing is backward with respect to the order of firing on the track. Reverse replay may facilitate the storage of behavioral sequences that lead to rewards, and this may take place only seconds after the experience is first imprinted as a pattern of activity in the hippocampal network.

Sequence Coding and Temporal Organization in Hippocampal Representations

The Hebbian postulate for how associations occur in the cortex, through correlated presynaptic and postsynaptic activity, confers a temporal constraint on how neuronal activities can mediate memory encoding. If Hebb's theory is right, areas involved in these processes must be in temporal synchrony during encoding. Hippocampal networks display characteristic oscillatory activity during spatial learning, but the exact significance of temporal firing and coherence patterns is not yet understood. One of the most intriguing observations is the compression of spike sequences into individual theta cycles during phase

precession in hippocampal place cells. Phase precession refers to the fact that as an animal moves through the place field of a cell, the cell fires at a progressively earlier phase of the theta cycle. Because of this precession, the sequence of firing in a set of place cells is reflected at a compressed timescale within each individual theta cycle. This mechanism orders the spiking of sequentially active neurons within the time range of spike-dependent plasticity, and so may result in associative storage of sequences that are longer than a single cycle of the theta oscillation.

Place Cells – Where Do They All Come From?

Place cells exist in all subfields of the hippocampus and in the dentate gyrus, the subiculum, and the entorhinal cortex. The most specific and confined firing fields are observed in the hippocampus proper. The first single-unit recording studies in the subiculum and the entorhinal cortex in the early 1990s reported quite dispersed firing fields, although only limited parts of these regions were sampled. Based on the striking contrast between the weak spatial modulation of entorhinal input neurons to the hippocampal formation and the high spatial information content of principal neurons in the hippocampus, investigators thought that place-specific firing originated from computations within the hippocampal network, presumably in the dentate gyrus and the CA3. This view was challenged, by the observation that CA1 neurons exhibit place-specific firing even after relatively selective lesions of the dentate gyrus or the CA3. These lesions interrupted the trisynaptic circuit, such that all remaining excitatory input to CA1 must have been mediated by the direct perforant-path input. Place fields in CA1 of animals with CA3 lesions were slightly broader than in untreated animals, but the firing was still confined to quite small regions of the area that the animal explored, suggesting that the spatial information might not originate within the hippocampus itself. These studies were followed by work showing that neurons have strong spatial correlates also upstream of the hippocampus, in the medial entorhinal cortex, the major interface between the hippocampus and the neocortex. Many principal neurons in this brain area were shown to have multiple sharply tuned firing fields that collectively signaled the rat's constantly changing position with the same precision as place cells in the hippocampus. The sharpest firing fields were observed in the dorsocaudal part of the entorhinal cortex, which contains the bulk of the entorhinal neurons projecting to the well-characterized place cells of the dorsal hippocampus. As the electrodes were moved toward more ventromedial parts of the entorhinal cortex, the firing fields became wider and more dispersed; in the intermediate-to-ventromedial bands, where most of the earlier recordings were made, neurons fired almost in the entire recording box, with no detectable spatial correlates at the most ventral locations. Spatially modulated neurons were only observed in the medial subdivision of the entorhinal cortex. The localized firing of the entorhinal cells was maintained without an intact hippocampus, suggesting that the map of the rat's current position is established before any signals are passed through this brain region. These data point to a new function for the entorhinal cortex (spatial computation) and suggest that metric spatial signals in the hippocampus are derived mostly from afferent areas rather than being computed in the hippocampus itself. This discovery 'releases' the hippocampus from an exclusive spatial domain and suggests that its primary role is the encoding and storage of associative memory, including both spatial and nonspatial elements.

The principal element of the entorhinal neural map is the 'grid cell' (Figure 3). The multiple firing locations of this cell type, which may comprise the

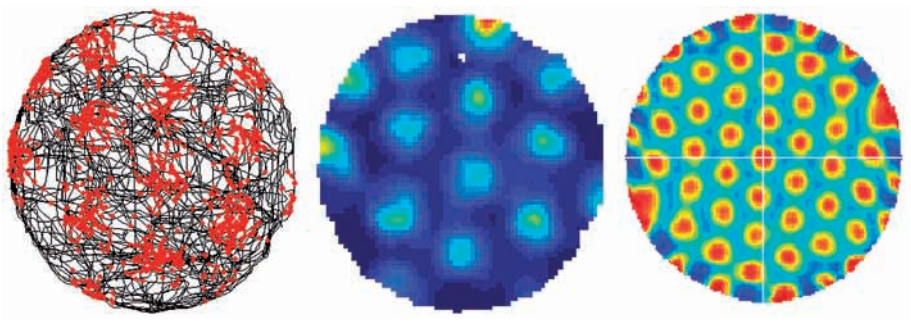

Figure 3 Periodic firing fields of a grid cell in the medial entorhinal cortex. The firing fields form a periodic triangular matrix tiling the entire environment covered by the animal. Left: Trajectory of the rat, with spike locations superimposed. Middle: Color-coded rate map, showing distribution of firing rate in the same cell (blue, silent; red, peak rate). Right: Autocorrelation matrix for the rate distribution in the two other diagrams (blue, −1; green, 0; red, +1). Place cells in the hippocampus are thought to reflect the integration of output from multiple grid cells with different spacing. Adapted from Hafting T, Fyhn M, Molden S, et al. (2005) Microstructure of a spatial map in the entorhinal cortex. *Nature* 436: 801–806, with permission from Nature.

majority of the neurons in layer II of the medial entorhinal cortex, form a remarkably regular and repetitive pattern. For each cell, the set of active locations defines a regular array that tiles the entire environment of the rat, like the cross-points of graphic paper rolled out over the test arena. However, unlike the rectangular lattices emerging from a Cartesian coordinate system, the entorhinal grid is triangular, with a pair of oppositely oriented equilateral triangles as the repeating unit. Grids of different cells are offset relative to each other such that each position in the environment can be identified, within a certain radius, from the activity of a fairly limited number of adjacent cells. These observations point to ensembles of entorhinal grid cells as elements of a possible neuronal coordinate system used by mammals for metric spatial navigation.

Several observations suggest that the entorhinal spatial map is the path-integration-based metric navigation system proposed as one of the key inputs to the hippocampal cognitive map of O'Keefe and Nadel. The entorhinal map is universal, operating at all times in all kinds of environments. Unlike place cells in the hippocampus, ensembles of grid cells are activated in a stereotypic manner, regardless of the unique landmarks of the environment. Information about position is apparently computed from the animal's own movement, using distance and direction of movement as the primary inputs. Because of its grid-like nature, the map can potentially represent places not visited just as well as places that have been visited. The grid map may thus be infinite, allowing the inclusion of all places and potential places. The existence of a single map that can be applied anywhere is economical and avoids the enormous capacity problems that might arise if the brain were to store rules about spatial interrelations for every single spatial context that the subject experienced.

Grids of neighboring cells have a common spacing and a common orientation. The spacing of the grid increases progressively as one gets deeper into the brain. By integrating inputs from grids at different scales, neurons in the hippocampus or other target areas may provide an unambiguous high-resolution representation of the spatial environment. If all grid cells had a single, common scale, the population code would repeat itself at intervals corresponding to a single period of the grid of individual cells. If activity were integrated across grids with different spacing and orientation, however, the cycle for repetition would get very large, enabling each position within this radius to be expressed by a unique pattern of collective activity in the grid network. The gradual increase in the scale of the firing fields and the distributed nature of the phase of the grid (the apparently random displacement of grid vertices among neighboring cells) are both mirrored downstream in the hippocampus.

Conclusion

Neural representations in the hippocampus carry strong spatial signals. These are expressed reliably throughout the hippocampal neuronal population, and they are likely to originate by integration of outputs from path-integration-based dynamic representations of the animal's location in the medial entorhinal cortex. On top of the spatial signal, hippocampal neurons carry information about other features of the environment, allowing neuronal ensembles in the hippocampus to store a number of discrete or continuous representations for each place or context experienced by the animal. With new powerful techniques that allow neuronal computation to be studied at the ensemble level, the study of place cells has become one of our most important tools to for understanding of how memory and cognition are implemented in the brain.

See also: Hippocampus; Neural Coding of Spatial Representations; Shape Representation in Inferotemporal Cortex; Spatial Memory: Assessment in Animals; Spatial Cognition and Executive Function; Synaptic Plasticity and Place Cell Formation.

Further Reading

Bostock E, Muller RU, and Kubie JL (1991) Experience-dependent modifications of hippocampal place cell firing. *Hippocampus* 1: 193–205.

Brun VH, Otnass MK, Molden S, et al. (2002) Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. *Science* 296: 2243–2246.

Foster DJ and Wilson MA (2006) Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* 440: 680–683.

Fyhn M, Hafting T, Treves A, et al. (2007) Hippocampal remapping and grid realignment in entorhinal cortex. *Nature* 446: 190–194.

Fyhn M, Molden S, Witter MP, et al. (2004) Spatial representation in the entorhinal cortex. *Science* 305: 1258–1264.

Hafting T, Fyhn M, Molden S, et al. (2005) Microstructure of a spatial map in the entorhinal cortex. *Nature* 436: 801–806.

Lentgeb JK, Leutgeb S, Moser MB, and Moser EI (2007) Pattern seperation in the dentate gyrus and CA3 of the hippocampus. *Science* 315: 961–966.

Leutgeb JK, Leutgeb S, Treves A, et al. (2005) Progressive transformation of hippocampal neuronal representations in 'morphed' environments. *Neuron* 48: 345–358.

Leutgeb JK, Moser MB, and Moser EI (2005) Place cells, spatial maps and the population code for memory. *Current Opinion in Neurobiology* 15: 738–746.

Leutgeb S, Leutgeb JK, Barnes CA, et al. (2005) Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science* 309: 619–623.

Marr D (1971) Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 262: 23–81.

McClelland JL and Goddard NH (1996) Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus* 6: 654–665.

McNaughton BL and Nadel L (1989) Hebb–Marr networks and the neurobiological representation of action in space. In: Gluck MA and Rumelhart DE (eds.) *Neuroscience and Connectionist Theory*, pp. 1–63. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nakazawa K, Quirk MC, Chitwood RA, et al. (2002) Requirement for hippocampal CA3 NMDA receptors in associative memory recall. *Science* 297: 211–218.

O'Keefe J and Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research* 34: 171–175.

O'Keefe J and Nadel L (1978) *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

O'Keefe J and Recce ML (1993) Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* 3: 317–330.

Samsonovich A and McNaughton BL (1997) Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience* 17: 272–275.

Wills TJ, Lever C, Cacucci F, et al. (2005) Attractor dynamics in the hippocampal representation of the local environment. *Science* 308: 873–876.

Wilson MA and McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* 265: 676–679.

Wood ER, Dudchenko PA, and Eichenbaum H (1999) The global record of memory in hippocampal neuronal activity. *Nature* 397: 613–616.

Hippocampus: Computational Models

R A Koene and M E Hasselmo, Boston University, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The hippocampus appears to be involved crucially in the acquisition and intermediate term maintenance of novel episodic event memories. For example, in the famous case of patient H.M. the anterior hippocampus and adjacent structures were bilaterally removed surgically. H.M. was consequently unable to form new event memories, but event memories prior to the surgery were preserved. H.M. also retained the ability to converse and perform computations when uninterrupted. In order to understand the functions of the hippocampus and the specific properties of memory acquired and maintained there, it is necessary to understand intrinsic properties of neurons and the complex interactions in networks of neurons. The search for such understanding is greatly aided by computational modeling, due to the ability to make explicit and precise observations in each component of a simulation and the ability to exert similarly precise and explicit control over model structure and parameters.

The following article reviews neural models of hippocampal memory function. (1) We present an anatomical summary of the neural circuitry that is involved. (2) We compare attractor network models of memory retrieval with computational models that achieve memory retrieval in terms of sequences of activity patterns. (3) We discuss recent models of spatial navigation. (4) We highlight the significance of the 'theta' brain rhythm in recent models of hippocampal function, especially during the sequential readout of memory. This includes a recent model mechanism that achieves rapid switching between memory encoding and memory retrieval, thereby enabling specific hippocampus-dependent memory tasks. (5) This leads to the discussion of a model that proposes a mechanism for temporal context-dependent retrieval of episodic memory. (6) We present advances in computational models of proposed short-term buffering in parahippocampal regions, which provide the buffer requirements for associative synaptic storage in the hippocampus. (7) We explicitly state the importance of context for recognition versus recall in memory tasks. (8) We conclude by reviewing models that address the transfer and consolidation of memories stored in the hippocampus into long-term memory and semantic memory.

The Standard Model Based on Anatomical Structure

Recent models have directly addressed data from human cognitive memory tasks using representations based on extensive physiological and anatomical data concerning the hippocampal formation. These hippocampal models utilize many features discussed in priormodels of the hippocampus. Individual stored items such as words are represented as patterns of active and inactive neurons in the entorhinal cortex, which provides input to the hippocampus and which in turn receives converging input from a broad range of neocortical structures. Activity spreads from this input layer into subregions of the hippocampus, including a structure with extensive excitatory recurrent connections – region CA3. Strengthening of the recurrent synapses connecting active neurons within region CA3 provides a mechanism for associating different components of each stored pattern. The experimental phenomenon of long-term potentiation provides support for this mechanism of synaptic modification. Associative storage in region CA3 has been utilized in most models of episodic memory function, which differ primarily in details of the model dynamics. These models allow behavioral features of memory to be related to specific neural substrates.

The dentate gyrus (DG) receives input from entorhinal cortex (EC) and sends output to CA3. Connections diverge from a small number of EC neurons to a large number of DG neurons, resulting in a sparser pattern of activity in DG and CA3 which allows more effective associative storage. In addition, mossy cell and granule cell layers of the dentate gyrus together form a second recurrent network in the hippocampus, which may allow associative encoding in DG. Encoding of novel information may be facilitated by extensive neurogenesis observed in the human adult DG.

Region CA1 of the hippocampus receives input from layer III of entorhinal cortex (ECIII) at distal dendrites via the perforant path synapses and receives input from region CA3 at proximal dendrites through the Schaffer collaterals. Input from ECIII cannot elicit activity in CA1 independently, depending on depolarization by input from CA3. ECIII provides an active context to pyramidal neurons in CA1. By contrast, input via the Schaffer collaterals from CA3 can in some cases cause spiking in CA3 without excitatory input from ECIII. Activity in CA1 forms the output of hippocampal processing, projecting out of the hippocampus via the subiculum. In spatial tasks involving rodents, CA1 pyramidal neuron spiking has been

found to be place specific (place cells), and subiculum activity is direction specific. During theta rhythm in linear track experiments, place cell firing exhibits phase precession.

Models Based on Attractor Network Dynamics

Many early models of the hippocampus represent memories with a single spatial pattern distributed across a population of neurons. During retrieval, activity induced by a memory cue causes activity to evolve toward one of these single spatial patterns – an 'attractor' state. This final attractor state persists indefinitely. Because of this single stable final state, these networks are referred to as fixed-point attractor systems.

Fixed-point attractors can be obtained in networks with extensive excitatory feedback connections and could therefore occur in region CA3 of the hippocampus, or in neocortical structures. One danger of such strong excitatory feedback connections is the possibility that activity can increase exponentially within the network. In early models, units were prevented from firing at rates higher than a particular maximum value, and memory states commonly involved firing at these high rates. More realistic networks obtained attractor dynamics with lower rates of firing by balancing feedback excitation with different types of inhibition, including shunting inhibition, subtractive inhibition, or normalization of total activity. Though fixed-point attractors in models can persist indefinitely, it is likely that neural circuits come under the influence of individual attractors for only brief periods. In very detailed biological models with spiking neurons, attractors require larger numbers of units to be stable, but can be obtained by using very specific point-to-point inhibitory connectivity or saturating synapses.

Sequence Models

Memories can also be represented as sequences of activity patterns within a network. In this framework, each pattern of activity on a population of neurons such as region CA3 is associated with a different subsequent pattern during encoding. During retrieval, presentation of an early pattern then elicits a chain of different subsequent patterns in the network which could potentially be repeated in a limit cycle. Sequences provide a simple means of representing interitem associations in memory tasks, as well as pathways through the environment. Recently, models of sequence storage in region CA3 of the hippocampus have been used to address behavioral tasks, including the transitivity task studied by Bunsey and Eichenbaum and spatial navigation tasks.

An important focus of recent models concerns the phenomenon of theta-phase precession. This experimental evidence suggests sequence storage within the hippocampus. As a rat runs along a continuous track, individual neurons (place cells) in its hippocampus will fire as the rat traverses a location specific to that cell (the place field). The firing of these cells has been measured relating to the phase of a high-amplitude oscillation in the hippocampal electroencephalogram (EEG) called the theta rhythm. As the rat enters the place field associated with a particular place cell, the firing of that cell will occur late in the theta cycle. As the rat crosses and leaves the place field, the firing of the place cell will occur earlier and earlier, suggesting that the cell was initially the end of a sequence being read out in the hippocampus, and as the rat crosses the field the cell becomes an earlier component of the sequence.

Several models of theta phase precession have been published. These models all involve a readout of sequences across time, but two of them involve slow readout of sequences across the full cycle of the theta rhythm (which has a period of about 200 ms). In a model by Tsodyks, Skaggs, McNaughton, and Sejnowksi, this slow readout is obtained with very weak excitatory connections. In a model by Jensen and Lisman, this slow readout is obtained with the slow dynamics of the *N*-methyl-D-aspartate (NMDA) receptor. Recently, Lisman modified this approach by stipulating sequence readout in which the retrieval of successive items is delayed by the time taken to correct an item representation through auto-associative pattern completion. Retrieval is further delayed by feeding back the resulting activity as a cue for the subsequent readout of the next item, using the reciprocal fiber loop that is formed by the pathway from hippocampal region CA3, via the mossy cells to the dentate gyrus, and the mossy fiber pathway from dentate gyrus to hippocampal region CA3. This change in the model was motivated by data showing that NMDA antagonists did not eliminate the phase precession phenomenon.

In contrast, more rapid readout with α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor kinetics is used in a separate model. In this model by Hasselmo, the theta phase precession is obtained by readout of sequences to different lengths during different phases of the theta cycle, due to phasic changes in depolarization or presynaptic regulation of synaptic strength. This phenomenon could allow selective context-dependent retrieval of sequences. Finally,

still other models have proposed that the theta phase precession does not result from sequence readout, but from a precession due to theta oscillations running at different frequencies in the soma versus the dendrites of pyramidal cells. Several of these hypotheses can be tested with pharmacological investigation of the phenomenon of theta phase precession.

Models of Spatial Navigation

The hypothesis that pathways through the environment are stored as sequences of place cell activity gives rise to another prediction, that as a pathway becomes familiar, place fields should expand and move backward along the path. This prediction has recently been confirmed in experimental measurements. Recent models by Burgess, Sharp, Hasselmo, Koene, and Gorchctchnikov have proposed that representations of neural space involve learning of multiple different pathways within that space, which can then be effectively integrated in a flexible relational structure. Many models start with an array of simulated place cells which encode the environment, and then modify the connections between these cells to learn gradients toward specific goals. This can be viewed as instantiating the assumption that space is the dominant parameter for neuronal response, although it is possible that place cells arise from a generic sparse conjunctive coding scheme. One recent model by Samsonovich and McNaughton assumes that generic two-dimensional maps of space are precoded in region CA3, and learning of a new environment involves modification of excitatory input to individual maps, rather than modification of recurrent connections within a map. In contrast, models which start with individual sequences can draw on a range of features in each sequence, building a response to task elements beyond just the spatial layout. Most models use a particular goal to influence the activity of other neurons in order to direct network activity toward a particular location. This remains an important issue, as experiments have not demonstrated 'goal cells' in any particular structure. Functional models of navigation provide an important means of interpreting available physiological evidence from this important experimental paradigm.

The Significance of Hippocampal Theta Rhythm

Recent computational models address the role of the hippocampal theta rhythm. During the peaks and troughs of the theta rhythm, differential modulation of membrane potential and synaptic transmission results in phasic changes in the magnitude of synaptic currents induced in region CA1 by input from EC and CA3 during different phases of theta rhythm. Empirical data show that the phase of theta rhythm in hippocampal regions also affects the strength of synaptic modification by long-term plasticity (LTP). The strongest LTP is elicited by stimulation at the peak of theta rhythm, as recorded in the hippocampal fissure. Combined, these modulatory effects led to the presupposition that opposing phases of theta rhythm may impose rapidly alternating modes that differentially favor encoding (strong input from EC, strong LTP) and retrieval (strong internal spread arising from CA3, weak LTP) in subregions of the hippocampus.

Computational models that propose a mechanism of ongoing alternating encoding and retrieval during task performance allow encoding of new information, interleaved with retrieval of prior knowledge to perform the task. Recent simulations of spatial navigation behavior used a mechanism of alternating encoding and retrieval within each theta cycle. These simulations were based on the hypothesis that information needed to perform the task is retrieved during each theta cycle in hippocampal region CA1. It was proposed that retrieval activity in CA1 is elicited by a convergence of excitatory input from pyramidal neurons in entorhinal cortex layer III through the perforant path and of excitatory input from pyramidal neurons in hippocampal region CA3 via Schaffer collaterals. The models suggested three important uses of theta rhythm. (1) Theta modulation enables rapidly alternating encoding and retrieval modes. (2) Offset phases of theta modulation in the connected networks of the entorhinal cortex and hippocampus enable synchronous operations. For example, the readout of buffered activity in the entorhinal cortex can occur reliably for presentation as input during the encoding phase of activity at the recurrent fiber synapses of hippocampal region CA3. (3) The modulation of membrane potential at theta rhythm in layer II of the entorhinal cortex may be an essential component of a nonsynaptic buffer that can maintain short sequences of spiking activity.

Recent computational modeling proposes explicit mechanisms for the retrieval of episodic memory in a specific temporal context. This effectively models data from Wood and Eichenbaum showing that units fire in a selective context-dependent manner in region CA1 of the hippocampus. Inputs representing forward retrieval could be gated by inputs representing temporal context via convergence in region CA1. Simulations demonstrated performance in a delayed spatial alternation task in which temporal context-specific retrieval is necessary for successful performance of the task.

Short-Term Buffers

Experimental observations and computer modeling by Alonso and Fransen demonstrated that certain pyramidal neurons in the entorhinal cortex exhibit afterdepolarization and the ability to maintain a stable pattern of intrinsic spiking even after input has terminated. In layer II of the entorhinal cortex (ECII), the intrinsic spiking frequency is similar to the frequency of the theta brain rhythm. The persistent firing behavior of those neurons led to proposals for short-term buffers of spike sequences, which can allow the acquisition and encoding of novel input in the parahippocampal and hippocampal regions. The buffer of spike sequences could allow (1) time compression between patterns of spikes to enable spike timing-dependent potentiation, (2) multiple cycles of spike pattern reactivation to repeatedly strengthen active synapses, and (3) maintenance of sequence order for the encoding of episodic memory. Another important element of these models is network-wide competitive inhibition, which serves to suppress activity following a pattern of spikes and thereby guarantee the separation of successive buffered spike patterns. This competitive inhibition and the detection of spiking at a phase of the theta rhythm that indicates a full short-term buffer are the proposed components of a mechanism that enables first-in–first-out replacement of buffered spike patterns with novel spiking input patterns. In addition to its role during sequence encoding, a short-term buffer for short sequences of representative spike patterns may gather and time compress a short sequence of spikes elicited by input that provides a temporal context cue for episodic retrieval in the hippocampus.

Recognition versus Recall

Most humans are familiar with the difference in effort required for recognizing a name versus recalling a name. Recall is more sensitive than recognition to injections of the acetylcholine receptor blocker scopolamine before encoding, and the cellular basis of this sensitivity has been analyzed in a model. In this model, neurons in entorhinal cortex represent the subject's memory for experimental context – that is, the testing room and apparatus – while separate neuronal populations in this area represent individual items – that is, words on a list. During encoding, activity spreads through the dentate gyrus into region CA3, where connections are strengthened among the neurons representing context, among the neurons representing individual words, and between the context and word neurons. The experimentally demonstrated effects of scopolamine were represented by reducing the rate of synaptic modification, reducing the depolarization of neurons, and increasing feedback excitation.

Simulation of scopolamine effects impaired subsequent recall but not recognition. During free recall, a subject is asked, "What words were on the list?" In the model, this is simulated with activation of entorhinal context units, which activate the context representation in region CA3. This context representation then sequentially activates the representations of individual words via strengthened connections (competition between words prevents simultaneous recall). During recognition, subjects are presented with individual words. In the model, neurons representing individual words are activated. If spread of activity across strengthened connections is sufficient to evoke activity in the context neurons, then the item is counted as correctly recognized. Because the context is present more frequently than the words during encoding, it has stronger excitatory feedback and is easier to activate, allowing recognition to persist even when slower synaptic modification during encoding prevents effective recall. This model suggests specific parameters of memory function which should be affected by specific drugs. Drug effects on conditioning phenomena in rats have also been modeled, but space does not allow review of neural models of conditioning.

The model of recognition described previously used one representation of recognition processes – but humans may use two different processes in recognition: (1) a process of rich recollection, in which the details of the specific item are recalled in a specific context and (2) a general sense of familiarity based on component features of an item. The rich recollection process may depend on the hippocampus, whereas the familiarity sense may depend on other cortical structures, including parahippocampal structures. Experimental data on these differences have been addressed in a recent paper which models the dynamics of recollection in a simulation of the hippocampal formation. The activation of representations in region CA3 of this model depends upon a conjunction of cues in the entorhinal cortex. This requirement of a specific conjunction of features prevents recollection from being induced by items containing only a portion of the features of a particular memory. For example, the model was trained on patterns representing word pairs such as 'window-reason' and 'car-oyster' and then tested on repaired lures such as 'window-oyster.' The model was able to recollect studied word pairs and reject many re-paired lures by retrieving the correct pair. The simulation also demonstrated the experimentally observed property that recollection-based recognition decreases with number of words on a list, whereas there is an increase in false alarms

due to the more vague familiarity-based recognition. This familiarity-based recognition may take place outside the hippocampus and contribute to many aspects of memory function. Lesions of the hippocampus alone cause significant impairments in episodic memory tasks including recognition memory, but research indicates that lesions including structures adjacent to the hippocampus cause stronger overall memory impairments.

Consolidation

Lesions of the hippocampal formation do not appear to impair preexisting semantic memory, but there is some loss of episodic memories from the time before the lesion, and more recently stored information appears to be affected more strongly – a phenomenon termed temporally graded retrograde amnesia. This suggests that the hippocampus mediates the gradual formation of neocortical memory representations. Models of the formation of semantic memory demonstrate that gradual, interleaved learning of new episodic information with existing semantic representations is essential to prevent distortion of previously stored semantic representations. Thus, the hippocampus may provide a temporary store for associations which then gradually modify neocortical representations.

The potential effect of the loss of hippocampal training on semantic memory has been investigated in studies of children with perinatal damage to the hippocampus. These subjects show a profound impairment of episodic memory, and as might be expected from the model, their development of semantic memory requires extensive training over a longer period – the external world must take the place of an internal mechanism for interleaved learning. The phenomenon of temporally graded retrograde amnesia does not appear consistently in all behavioral tasks, but in the case of human subjects, temporally graded retrograde amnesia appears to be particularly true for patients with damage selective to the hippocampus proper.

Potential physiological mechanisms for a two-stage model of memory formation have been proposed. The initial encoding in the hippocampus has been proposed to occur during theta rhythm dynamics, and the subsequent transfer to the neocortex to occur during sharp waves in quiet waking and slow-wave sleep. The dramatic decrease in acetylcholine levels during slow-wave sleep could contribute to these different dynamical states, as it will greatly enhance the strength of excitatory feedback in the hippocampus. This two-stage system allows rapid encoding of episodic memories, with slower consolidation of more complex semantic representations.

See also: Animal Models of Amnesia; Declarative Memory System: Anatomy; Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Episodic Memory: Assessment in Animals; Episodic Memory; Executive Function and Higher-Order Cognition: Computational Models; Hippocampus and Neural Representations; Hippocampus; Memory: Computational Models; Spatial Memory: Assessment in Animals.

Further Reading

Burgess N, Recce M, and O'Keefe J (1994) A model of hippocampal function. *Neural Networks* 7(67): 1065–1081.

Eichenbaum H and Cohen N (2004) *From Conditioning to Conscious Recollections Memory Systems of the Brain*. Oxford: Oxford University Press.

Gluck M and Myers C (2000) *Gateway to Memory: An Introduction to Neural Network Modeling of the Hippocampus and Learning*. Boston: MIT Press.

Hasselmo M, Bodelon C, and Wyble B (2002) A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation* 14(4): 793–817.

Hasselmo M and Eichenbaum H (2005) Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural Networks* 18(9): 1172–1190.

Hasselmo M and Wyble B (1997) Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research* 89: 1–34.

Jensen O and Lisman J (1996) Hippocampal CA3 region predicts memory sequences: Accounting for the phase precession of place cells. *Learning & Memory* 3: 279–287.

Jensen O and Lisman J (2005) Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends in Neurosciences* 28(2): 67–72.

Koene R, Gorchetchnikov A, Cannon R, and Hasselmo M (2003) Modeling goal-directed spatial navigation in the rat based on physiological data from the hippocampal formation. *Neural Networks* 16(5–6): 577–584.

Koene R and Hasselmo M (2006) First-in–first-out item replacement in a model of short-term memory based on persistent spiking. *Cerebral Cortex* 17: 1766–1781.

Levy W (1996) A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus* 6(6): 579–590.

Lisman J (1999) Relating hippocampal circuitry to function: Recall of memory sequences by reciprocal dentate–CA3 interactions. *Neuron* 22: 233–242.

Lisman J (2005) Hippocampus, ii: Memory connections. *American Journal of Psychiatry* 162(2): 239.

Lisman J and Idiart M (1995) Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science* 267: 1512–1515.

McClelland J, McNaughton B, and O'Reilly R (1995) Why there are complementary learning systems in the hippocampus

and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102: 419–457.

Norman K and O'Reilly R (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychology Review* 110 (4): 611–646.

Reed J and Squire L (1997) Impaired recognition memory in patients with lesions limited to the hippocampal formation. *Behavioral Neuroscience* 111: 667–675.

Rolls E and Treves A (1998) *Neural Networks and Brain Function.* New York: Oxford University Press.

Sharp P (1991) Computer simulation of hippocampal place cells. *Psychobiology* 19: 103–115.

Hippocampus: Molecular Anatomy

C L Thompson and E S Lein, Allen Institute for Brain Science, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

The hippocampus has long been a major focus of neuroscience research, in large part due to its highly distinctive cytoarchitecture and its widely accepted functional role in learning and memory. Analysis of gene expression in the hippocampus has been a popular means to understand the cellular diversity of the hippocampus and genetic underpinnings of phenotypic properties of these cells. These approaches range from global genomic analyses of hippocampus in its entirety to cellular resolution immunocytochemical analysis.

Genomic Analysis of Total Hippocampal Gene Expression

The availability of methodologies for large-scale transcriptional analysis has allowed genome-scale profiling in the complete hippocampus or in subdivisions of the hippocampus, most notably using serial analysis of gene expression (SAGE) and DNA microarrays. These approaches have yielded insight into the types and relative abundances of genes expressed in the hippocampus and also the genetic relationship between the hippocampus and other brain regions. Based on these techniques, at least half of known genes are expressed at detectable levels in the hippocampus. Less than 0.5% of these genes, largely composed of mitochondrial genes and genes involved in energy metabolism, make up >20% of the total mRNA transcripts in the hippocampus. A wide variety of neurotransmitter receptors, G-protein-coupled receptors, ion channels, and synaptic components representing neuronal gene expression are expressed at lower levels. Gross analysis of the entire hippocampus also reflects nonneuronal populations such as oligodendrocytes and astrocytes since these cell types are intermingled with or adjacent to the major neuronal populations.

Global patterns of gene expression in adult brain regions bear the imprint of their developmental origins. Hippocampal gene expression is most similar to that of other forebrain regions such as the neocortex and amygdala, less similar to the evolutionarily older brain stem regions, and even more dissimilar to regions such as the retina and pituitary gland. Interestingly, global gene expression patterns in different subfields of the hippocampus are more similar to one another than to other brain regions, despite the major differences in predominant neuronal types between hippocampal subfields (pyramidal vs. granule) and apparent morphological similarities of cell types in different brain regions (e.g., pyramidal cells in the hippocampus and neocortex).

Histochemical analysis of gene localization in the hippocampus has produced a wealth of cellular resolution gene expression data in the hippocampus. Centralized databases have become available that provide detailed gene expression data on a genomewide scale as well, allowing a global approach to correlating underlying gene expression to the unique phenotypic properties of particular cell types in the brain.

Subregional Gene Expression in Excitatory Hippocampal Neurons

The contemporary delineation of the hippocampus includes the dentate gyrus and Ammon's horn (cornu ammonis), which is divided into four subfields – CA1–CA4 – based on the nomenclature originally proposed by Lorente de Nó. CA4, now generally referred to as the hilus or polymorphic layer. The dentate gyrus is composed largely of small, densely packed granule neurons, whereas the CA fields are dominated by a dense band of pyramidal neurons that vary in their morphological and physiological characteristics from field to field. There is some debate regarding the distinct nature of the CA2 subfield in the rodent because it is not possible to delineate boundaries on the basis of cytoarchitecture alone, whereas single neurons in CA2 bear a strong resemblance to neurons in either CA3 or CA1.

Many studies have demonstrated striking regional gene expression in the different subfields of the hippocampus, discovered either on a gene-by-gene basis or through targeted profiling using DNA microarrays on microdissected hippocampal subregions. Pyramidal cells in CA1–CA3, hilar neurons, and dentate gyrus granule cells are each uniquely delineated by expression of individual genes (Figure 1). A fairly large number of genes expressed selectively in CA2 (or specifically excluded from CA2) have bolstered the argument for its identity as a discrete hippocampal subfield, although no distinct function has been ascribed to it. Visualized with specific markers, CA2 forms a thin, uniform band between, and nonoverlapping with, both CA1 and CA3 throughout the entire septotemporal extent of the hippocampus (Figure 2). Interestingly, co-expression of genes in CA2 and CA3 occurs more frequently than expression in either region separately, indicating that pyramidal cells in these regions are functionally very similar to one another.

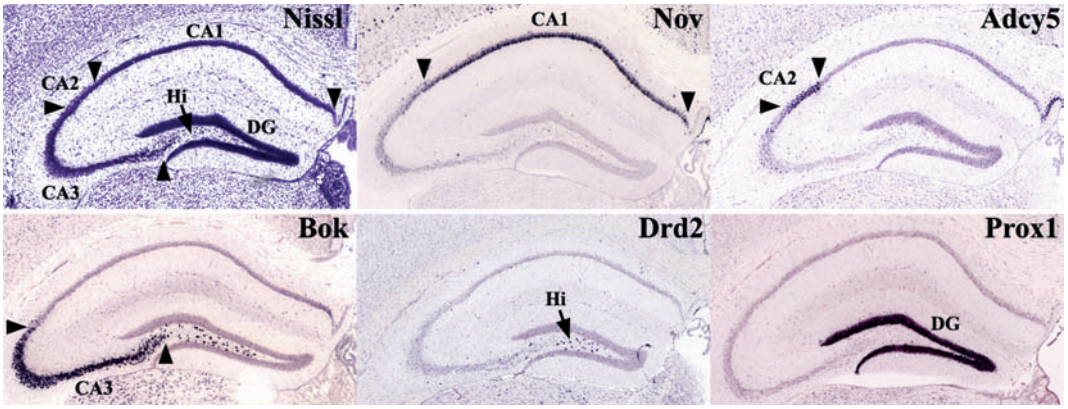

Figure 1 Highly specific gene expression in hippocampal subfields. Excitatory neurons in each of the major hippocampal subfields are delineated by specific patterns of gene expression. Adcy5, adenylate cyclase 5; Bok, Bcl-2-related ovarian killer protein; CA1–CA3, pyramidal cell layers of Ammon's horn; DG, dentate gyrus; Drd2, dopamine receptor 2; Hi, hilus of the dentate gyrus; Nov, Nephroblastoma overexpressed protein; Prox1, Prospero-related homeobox 1. Data from the Allen Brain Atlas project at http://www.brain-map.org.

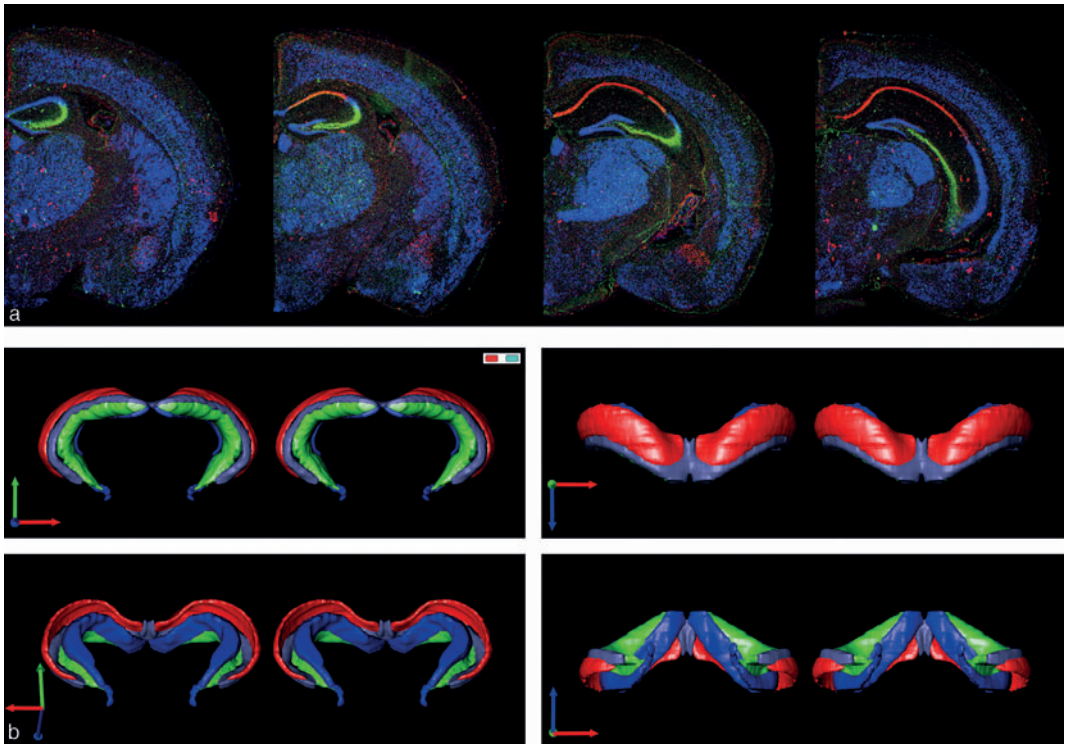

Figure 2 Nonoverlapping ion defines hippocampal subfields CA1–CA3 and the dentate gyrus. (a) Pseudo-colored overlays of *in situ* hybridization data for mannosidase 1, alpha (red; CA1), Bcl-2-related ovarian killer protein (green; CA3), and Purkinje cell protein 4 (blue; dentate gyrus + CA2) on coronal sections of mouse brain at different rostral–caudal locations. (b) Stereo pairs of three-dimensional reconstruction of serial sections for these same genes demonstrating the full extent of each subfield throughout the entire hippocampus. CA2 is colored purple to distinguish it from the dentate gyrus. Three-dimensional orientation bars in b: dorsal (green), rostral (blue), medial/lateral (red). Reprinted from Lein ES, Callaway EM, Albright TD, and Gage FH (2005) Redefining the boundaries of the hippocampal CA2 subfield in the mouse using gene expression and 3-dimensional reconstruction. *Journal of Comparative Neurology* 485(1): 1–10, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

Anatomical and functional heterogeneity within hippocampal subregions has been described at a number of levels that are likely to be reflected at the level of gene expression. First, efferent, afferent, and intrahippocampal connections are topographically organized along both the septotemporal and the proximal–distal planes. Several members of the Ephrin family, demonstrated to mediate the formation of

topographically specific connections in other structures, are expressed in gradients along the septotemporal axis of the hippocampus. There is also a functional differentiation along the septotemporal axis of the hippocampus, with the dorsal portion being more tightly linked to spatial learning and the ventral portion more closely linked to anxiety-related behaviors. Dorsal–ventral (septotemporal) variation in CA1 gene expression has been described that may correlate with these functional differences.

Genes that display regional specificity within the hippocampus are enriched for certain types of molecular functions. Signal transduction, transcriptional regulation, ion channels or transporters, and cytoskeletal or structural molecules are highly represented, as are growth factors and calcium-binding proteins. These classes of genes are likely to give rise to the different morphological and dynamic signaling properties that differentiate excitatory hippocampal neurons. Subfield-specific splice variance of ion channel subunits has also been described, adding another level of complexity to the relationship of gene expression to the electrophysiological properties of specific hippocampal neurons.

Gene expression at the regional and cellular level is highly complex and combinatorial. At the gene level, few genes are expressed in a single subfield. The majority of genes differentially regulated in the hippocampus are expressed in more than one subfield, and all possible combinations of subfields are seen for particular genes. At the cellular level, neurons in each subfield express unique combinations of genes that overlap strongly but differentially with neurons in the other subfields. Ultimately, it will be necessary to understand larger genetic interactions between cohorts of co-expressed genes at the cellular level to predict the full spectrum of phenotypic differences between different cell types in the hippocampus.

Hippocampal Interneurons

Inhibitory interneurons that use γ-aminobutyric acid (GABA) as their neurotransmitter form a diverse set of hippocampal cell types. At least 16 subtypes of interneurons exist in CA1 alone, with a remarkable diversity of morphological and physiological properties as well as distinct patterns of connectivity. For example, axo-axonal cells innervate the axon initial segment of large numbers of pyramidal neurons, putting them in position to strongly modulate action potential generation in those neurons. Several other cell types selectively innervate other interneurons, presumably evoking disinhibition in their target neurons.

Different interneuron subtypes are also distinguished by differential gene expression. Calcium-binding proteins, neuropeptides, and certain neurotransmitter receptors in particular are robustly and differentially expressed in different interneuron populations. CA1 expression patterns for most of the defined interneuron markers are shown in Figure 3, illustrating specificity of somal locations in the different sublaminae of CA1 but with highly overlapping patterns. Few, if any, interneuron classes are uniquely identified with current genetic markers, however, and are typically distinguished by a combination of characteristics including soma location, axonal and dendritic patterns of connectivity, and physiological response properties in addition to marker expression. These findings indicate, as for excitatory neuronal populations, that cellular phenotype is the product of the expression of complex combinations of gene products. Although it is possible that there exist single genetic markers that specifically delineate different interneuron populations, it seems highly likely that a set of genes may be necessary to identify a particular neuronal class.

Nonneuronal Cell Populations

In addition to the neuronal populations of the hippocampus, several prominent nonneuronal populations exist in the hippocampus (Figure 4). Oligodendrocytes and astrocytes comprise the major glial populations. Oligodendrocytes are found mainly in the fimbria and alveus, with smaller numbers in stratum lacunosum-moleculare and the dentate gyrus (the mossy fiber pathway from the dentate gyrus to CA3 is unmyelinated). Oligodendrocytes express very high levels of myelin-associated genes, including myelin basic protein and proteolipid protein, which are among the highest expressing hippocampal genes identified by SAGE analysis of total hippocampus. Astrocytes are more uniformly distributed throughout the hippocampus, with some enrichment in stratum lacunosum-moleculare. There appears to be some diversity among astrocyte populations defined by molecular markers, as demonstrated by the different expression patterns of glial fibrillary astrocytic protein and the gap junctional protein connexin-43, two well-characterized markers for astrocytes (Figure 4).

The hippocampus also contains dense vasculature. Vascular endothelial cells making up the vascular walls express a distinct set of genes, such as vascular endothelial protein tyrosine phosphatase, whereas hemoglobin expression in red blood cells is prominent in nonperfused tissue preparations.

Finally, the subgranular zone of the dentate gyrus is a region of intense study as one of two regions in which neurogenesis persists in the adult brain. This zone contains a precursor population that continues

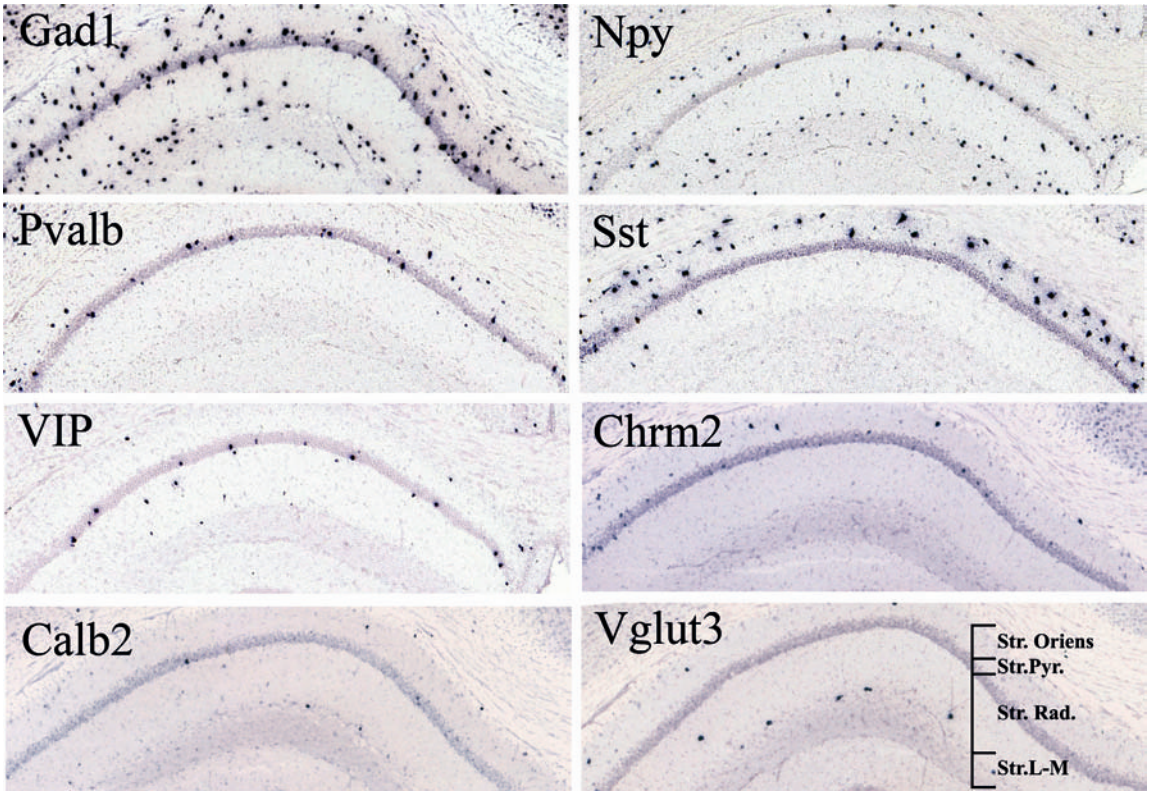

Figure 3 Interneuron markers in CA1. Cellular localization patterns for a pan-GABAergic interneuron marker (Gad1) and a series of markers commonly used to define subtypes of hippocampal interneurons. Calb2, calretinin; Chrm2, cholinergic receptor, muscarinic 2, cardiac; Gad1, glutamic acid decarboxylase 1; NPY, neuropeptide Y; Pvalb, parvalbumin; Sst, somatostatin; VIP, vasoactive intestinal peptide; Vlut3, vesicular glutamate transporter 3. Sublaminae of CA1: stratum oriens, stratum pyramidale, stratum radiatum, and stratum lacunosum-moleculare. *In situ* hybridization data on adult mouse brain from the Allen Brain Atlas project at http://www.brain-map.org.

to divide and generate new granule neurons that migrate into the overlying granule cell layer of the dentate gyrus. Cell cycle-associated genes such as mKi67 are specifically expressed in these dividing cells, and they are detectable as sparse cells in the subgranular zone as well as in dividing cells in the walls of the lateral ventricles.

Developmental Hippocampal Gene Expression

Gene expression profiling has also been applied to hippocampal development, identifying stage-specific genetic programs associated with hippocampal neurogenesis, differentiation, and maturation. For example, at prenatal ages, genes involved in cell cycle, histone regulation, and DNA replication are expressed at high levels, correlating with the period of neurogenesis. Later, genes associated with neuronal maturation are expressed at high levels, including actin and tubulin isoforms. Finally, in later development genes associated with synaptogenesis are upregulated. Subregion-specific gene expression provides useful markers to track the specification of hippocampal cell types. For example, the subfields of the hippocampus can be differentiated by region-specific gene expression as early as embryonic day 15.5, prior to the arrival of major hippocampal afferents.

Summary

Despite its apparent cytoarchitectural simplicity, the hippocampus contains a wide variety of neuronal and nonneuronal cell populations. This cellular diversity is reflected at the level of gene expression, which provides the genetic substrate for phenotypic differences between these cell classes. The availability of techniques for global gene expression analysis holds great promise for understanding structure–function relationships within the hippocampus, particularly as they begin to be applied to specific cellular populations. Genome-scale cellular resolution gene expression data are now available as well, holding the promise of defining comprehensive molecular taxonomies at the cellular level.

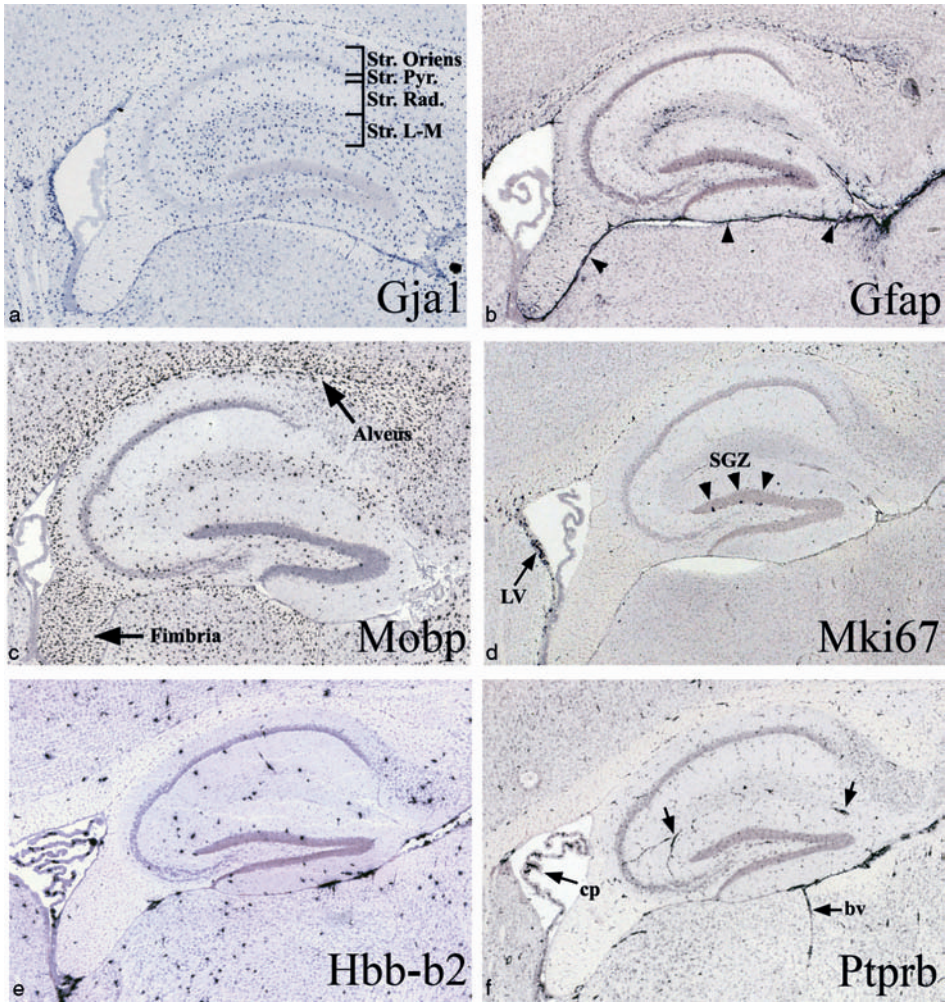

Figure 4 Nonneuronal cell types in the hippocampus. *In situ* hybridization for genes expressed in nonneuronal cell types. Astrocytes: gap junction membrane channel protein 1 (Gja1) and glial fibrillary acidic protein (Gfap); oligodendrocytes: myelin-associated oligodendrocytic basic protein (Mobp); dividing cells: antigen identified by monoclonal antibody Ki 67 (Mki67); vascular endothelium: protein tyrosine phosphatase receptor type B (Ptprb); red blood cells: hemoglobin beta adult minor chain (Hbb-b2). Sublaminae of CA1: stratum oriens, stratum pyramidale, stratum radiatum, and stratum lacunosum-moleculare. bv, blood vessels; cp, choroids plexus; LV, lateral ventricle; SGZ, subgranular zone of the dentate gyrus. Arrowheads in b delimit endothelial lining between hippocampus and thalamus. Data from the Allen Brain Atlas project at http://www.brain-map.org.

See also: Corticothalamic Connections: Ultrastructure; Corticothalamic Connections: Structure and Function; Gene Expression in Normal Aging Brain; Gene Expression Regulation: Activity-Dependent; Gene Expression in the Evolution of the Human Brain; Hippocampus.

Further Reading

Coultrap SJ, Nixon KM, Alvestad RM, Valenzuela CF, and Browning MD (2005) Differential expression of NMDA receptor subunits and splice variants among the CA1, CA3 and dentate gyrus of the adult rat. *Brain Research: Molecular Brain Research* 135: 104–111.

Freund TF and Buzsaki G (1996) Interneurons of the hippocampus. *Hippocampus* 6: 347–470.

Grove EA and Tole S (1999) Patterning events and specification signals in the developing hippocampus. *Cerebral Cortex* 9: 551–561.

Lein ES, Zhao X, and Gage FH (2004) Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput *in situ* hybridization. *Journal of Neuroscience* 24: 3879–3889.

Mody M, Cao Y, Cui Z, et al. (2001) Genome-wide gene expression profiles of the developing mouse hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* 98: 8862–8867.

Somogyi P and Klausberger T (2005) Defined types of cortical interneurone structure space and spike timing in the hippocampus. *Journal of Physiology* 562: 9–26.

Sugino K, Hempel CM, Miller MN, et al. (2006) Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature Neuroscience* 9: 99–107.

Sunkin SM (2006) Towards the integration of spatially and temporally resolved murine gene expression databases. *Trends in Genetics* 22: 211–217.

Zapala MA, Hovatta I, Ellison JA, et al. (2005) Adult mouse brain gene expression patterns bear an embryologic imprint. *Proceedings of the National Academy of Sciences of the United States of America* 102: 10357–10362.

Relevant Website

http://www.brainatlas.org – Allen Brain Atlas.

Histamine Receptors and their Ligands: Mechanisms and Applications

H Timmerman and H van der Goot, Amsterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Histamine was known as a chemical before its presence in biological systems was confirmed. In 1907 Windaus and Vogt described the synthesis of 2-(imidazole-4-yl) ethylamine; much later this compound received the name histamine (Greek *histos*, 'tissue'), because of the relatively high concentrations found in tissues. In 1910 Dale and Laidlaw isolated the compound from animal tissues, but originally it was thought to be there just as a metabolite of histidine, having no physiological significance of its own. During the 1920s it became clear that histamine evokes several pharmacological activities, resembling very much the actions of the (at that time) not well-defined 'H substance,' a more or less hypothetical compound thought to be released from injured tissue. Eventually it was found that H substance was a complex mixture of compounds, of which histamine was only one, playing a major role in anaphylaxis. Because of the role of histamine in anaphylaxis an intensive search for histamine blocking agents was started; the first so-called antihistamines were found in the late 1930s, and in the 1940s the antihistamines were introduced as antiasthmatics and as antiallergics in general.

Very many antihistamines, of which diphenhydramine may be considered as being the archetype, have been developed since. They are all of limited use in asthma, but are quite effective in treating other allergic responses, such as hay fever and skin rashes. For many years it remained unclear why certain effects of histamine (e.g., stimulation of gastric acid secretion or increased heart rate) were not blocked by highly active antihistamines. In the 1960s Ash and Shild proposed that a second type of histamine receptor might exist in heart and stomach tissues. Subsequently, in 1974, Black et al. reported for the first time on selective blockers of the actions of histamine in both the heart and the stomach. The new type of receptor was named H_2; the receptor discovered earlier was designated H_1 from then on. As with the antihistamines – now renamed H_1 blockers – several selective H_2 blockers followed the first H_2 antagonist burimamide; cimetidine and ranitidine, antigastric ulcer agents, have become blockbuster medicines (Tagamet and Zantac, respectively).

In the 1980s it became clear that histamine not only has the characteristics of a local hormone, but is a neurotransmitter as well. Shortly thereafter, Schwartz and group in Paris identified a third class of histamine receptors. These H_3 receptors occur mainly as presynaptic receptors, functioning both as auto- and heteroreceptors by regulating the release of histamine (autoreceptor) and other neurotransmitters (heteroreceptor). Pretty soon thereafter a large number of selective H_3 agonists and antagonists were found; so far none of these compounds has been developed for clinical application; several antagonists, however, are undergoing clinical testing for different central nervous system (CNS)-related diseases.

During a search for new G-protein-coupled receptor (GPCR) genes, using the sequence of the H_3 receptor gene as template, a fourth type of histamine receptor was found in the early 2000s. This H_4 receptor is present especially in peripheral organs, on immunoactive cells, where it likely mediates the proinflammatory properties of histamine. Histamine-induced inflammation is not blocked by H_1, H_2, or H_3 antagonists. There is a low level of homology between the H_1, H_2, and H_3/H_4 receptor proteins. The latter pair has a relatively high degree of homology, explaining why imidazole-derived ligands of the H_3 receptor typically have a considerable affinity for the H_4 receptor as well; the reverse is also true. Selective nonimidazole H_3 and H_4 ligands are available, though.

In the following sections, the several ligands of the four histamine receptors are discussed, regarding both their medicinal (or potential medicinal) application and their use as tools in pharmacological research. Although it is beyond the scope of this article to discuss the many pharmacological technologies applied to characterize the different ligands, the most frequently used tests are summarized in the final section.

The Histamine H_1 System

The Receptor, Mechanisms, and Role in Physiology and Pathology; H_1 Blockers as Medicines

Originally the H_1 receptor was known as 'the' histamine receptor, and its ligands were viewed as (anti) histamines. Like all identified histamine receptors, the H_1 receptor belongs to the large family of G-protein-coupled receptors. The receptor is via a G_q-protein coupled to phospholipase C; receptor stimulation leads to the hydrolysis of phosphatidylinositol 4,5-bisphosphate, yielding increased levels of inositol 1,4,5-trisphosphate (IP_3) and 1,2-diacylglycerol and eventually of increased Ca^{2+} levels and activation of protein kinase C.

The receptor is present in many tissues, both in peripheral organs and in the CNS. The first

described function of the H_1 receptor was its role in allergy-related conditions – contraction of soft tissue (in airways) and increased permeability of the capillaries of the cardiovascular system. It was known that histamine release from mast cells occurred during allergen–antibody reactions; thus the idea emerged that H_1 antagonists would be effective antiallergics. These expectations were not completely fulfilled, especially because histamine is only one of the factors released in allergic responses and also because side effects (sleepiness) occur at H_1 blocker doses lower than needed for an antiallergic effect. The several H_1 blockers that reached the market are, however, of use in conditions such as hay fever, but are only weakly effective in treating asthma.

Although sleep-inducing properties of H_1 blockers were seen for virtually all of the many compounds that were marketed during the 1950s and 1960s, it had not been proved that these effects were related to the H_1 blockade. There were intensive efforts to develop so-called non-sleep-inducing H_1 blockers, without any true success, though. The application of so-called *ex vivo* experiments using radiolabeled ligands showed a very good correlation between occupation of H_1 receptors in the brain and induction of sleep in animals, suggesting that the induction of sleep by H_1 blockers is caused by H_1 blockade. When, moreover, later on it was found by molecular biology approaches that H_1 receptors in peripheral tissues are identical with the receptors in the CNS, it became obvious that nonsedating H_1 blockers could only be based on the absence of penetration of a given compound into the CNS.

Early attempts to reduce brain penetration by making less lipophilic derivatives failed. In the 1980s, rather by chance, a number of less CNS-penetrating H_1 blockers were found (terfenadine, cetirizine, and loratadine are examples); these compounds led to a new class of a second-generation non-sleep-inducing (or less sleep-inducing) antiallergic medicines that were quite active, especially in hay fever patients. The warning 'no driving when using this medicine,' applicable to the early antihistamines, could be rescinded for the new medicines.

During the 1990s a cardiotoxity, a prolongation of the QT interval, was observed for some members of the new class of medicines. This serious side effect, caused by a blockade of K^+ channels of the human *eag*-related gene (HERG) type on cardiac tissue, led to the removal from the market of a few medicines (e.g., the first successful second-generation blocker, terfenadine), but the cardiotoxicity was found not to be a class effect. From animal experiments it became clear that some of the early H_1 blockers – and even those available as over-the-counter medicines – showed signs of the same toxicity, but such compounds remained freely available. In addition, there is ample evidence that histamine plays a role via H_1 receptors in seizures and epilepsy, and Japanese investigators published data on the proconvulsant activity of H_1 blockers. These observations have not, however, led to serious concern about this potential side effect of the H_1 blocking agents.

As is the case for many receptor systems, the H_1 receptor shows constitutive activity; this means that under certain conditions the receptor is in an active conformation, thus producing the second messenger intracellularly, in the absolute absence of an agonist. Equally important is that histamine blocking agents should be considered as being either true antagonists (binding only to the receptor; no intrinsic activity) or as inverse agonists (reducing the constitutive activity upon binding to the receptor). All clinically applied H_1 blockers tested for these properties have been identified as inverse agonists.

Compounds; Structure–Activity Relationships

H_1 agonists Histamine has a very simple chemical structure (Figure 1). The compound has a low affinity for the H_1 receptor (K_D around 4.5 nM), and as the ED_{50} is around 6.5 nM, there is a high spare receptor capacity; only about 1% occupation of the receptor leads to the maximal level of receptor activation.

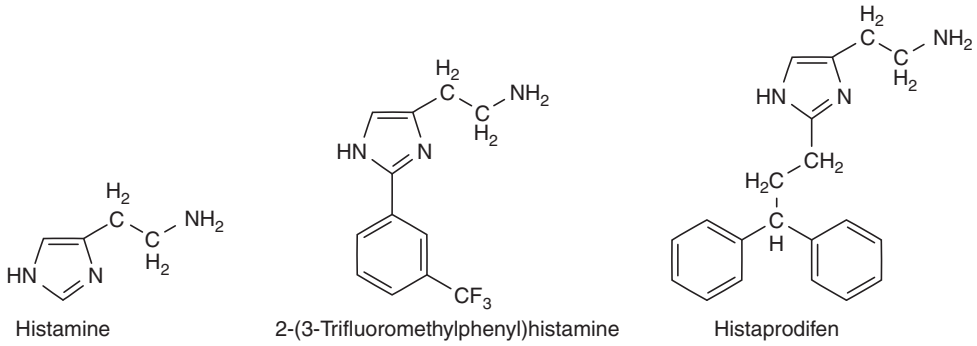

Figure 1 Histamine and some representative selective H_1 agonists.

The agonistic activity is very much dependent on the structure. Introduction of even small substituents, either in the side chain or at the imidazole nucleus, causes a strong reduction in activity, except at position 2 of the imidazole ('between the two nitrogens') group. At this position certain substitutions favor the activity. Especially aromatic groups and ω-benzhydryl alkylene groups at this C2 position may lead to compounds with a much higher (and, moreover, more selective) H_1 activity (see Figure 1). As a rule of thumb it may be said that H_1 agonism may be expected from heteroaromatic compounds in which an ethylamine moiety is placed on the aromatic nucleus at a position adjacent to a nitrogen atom. Recently, our group identified the ergot derivative lisuride, a compound with a potent H_1 agonistic activity and an unexpected structure.

H_1 blockers During the late 1930s the first experimental H_1 blockers were found during random screening procedures. The intravenous LD_{50} of histamine was determined in untreated mice; when mice were pretreated with a test compound, the LD_{50} was found to be higher, thus a histamine blockade was established and a blocker was identified. Though not the first H_1 blocker to be used therapeutically, diphenhydramine became the archetype of the very many derivatives that appeared in the scientific literature, several of which were marketed. Figure 2 depicts the structures of a number of H_1 blockers.

The H_1 blockers compounds reaching the marketplace since the late 1940s and the 1950s have been much less successful than anticipated. Besides their lack of activity, the high incidence of CNS-related side effects (sedation to a level which made them attractive as sleeping pills, thus ruling out their use when driving) limited their usefulness. Many researchers have tried to design so-called nonsedating H_1 blockers. Originally it was thought that the sedation induction had nothing to do with H_1 blockade, and nonsedating compounds were searched for more or less at random among series of H_1 blockers. When it was eventually found that the sedation is a consequence of H_1 receptor blockade in the CNS and that these CNS receptors are not different from the ones found in peripheral systems, the focus became the identification of compounds that do not penetrate into the brain. Attempts to reach this objective by selecting hydrophilic blockers were unsuccessful. The goal was finally reached by chance, when the diphenhydramine derivative terfenadine was found to be an active H_1 blocker, but did not penetrate into the CNS. This first example of this new class of second-generation (nonsedating) H_1 blockers was followed by a number of other effective derivatives, including loratadine and cetirizine (Figure 2). Only later was it found that the beneficial property (no CNS penetrance) is a result of the property of these compounds to be substrates of the permeability glycoprotein (PgP) transporter system. A number of these second-generation compounds have achieved major market status as antiallergic medicines.

H_1 Ligands as Research Tools

For any molecular pharmacology study, the availability of selective ligands as tools is essential. In this section and in subsequent sections on H_2, H_3, and H_4 receptors, ligands that afford such utility are presented.

H_1 agonists Until rather recently very few selective H_1 agonists were available, and those that were had an activity level much lower than that of the nonselective histamine, making them of limited use only. The most widely used agonists have been the 2-thiazolyl and the 2-pyridyl analogs (Figure 3). The histamine derivatives shown in Figure 1, when carrying on position C2 an aromatic group with an alkylene linker (Figure 3), have greatly improved the matter; compounds such as histaprodifen and superhistaprodifen have become important tools for studying the role of H_1 receptors in physiology and pathology.

H_1 antagonists; chiral ligands Although many selective H_1 antagonists have been described, only a few have been selected as tools for experimental studies; these compounds are shown in Figure 4. The most frequently used compound is likely mepyramine, but triprolidine and chlorpheniramine offer certain advantages. The structure of triprolidine is interesting in that it forms two geometric isomers. The E-isomer (triprolidine) is about a 1000 times more active (as measured in guinea pig ileum) than the Z-isomer, which makes triprolidine very useful for studying the binding mode of the compound to the receptor and the reason why it is frequently used in molecular modeling studies.

Chlorpheniramine, having a chiral structure, is also very suitable for many pharmacological studies. One of its two isomers is a much more potent H_1 blocker than the other one; thus the two isomers can be used to determine whether an effect of an H_1 agonist is caused by an interaction with the H_1 receptor. A nice example of such an application has been given by Leurs and colleagues, in studies proving that a binding site for mepyramine on liver cells was not an H_1 receptor but rather was a P450 enzyme.

Radiolabeled ligands Studying the affinity of ligands for any binding system can best be done by experiments

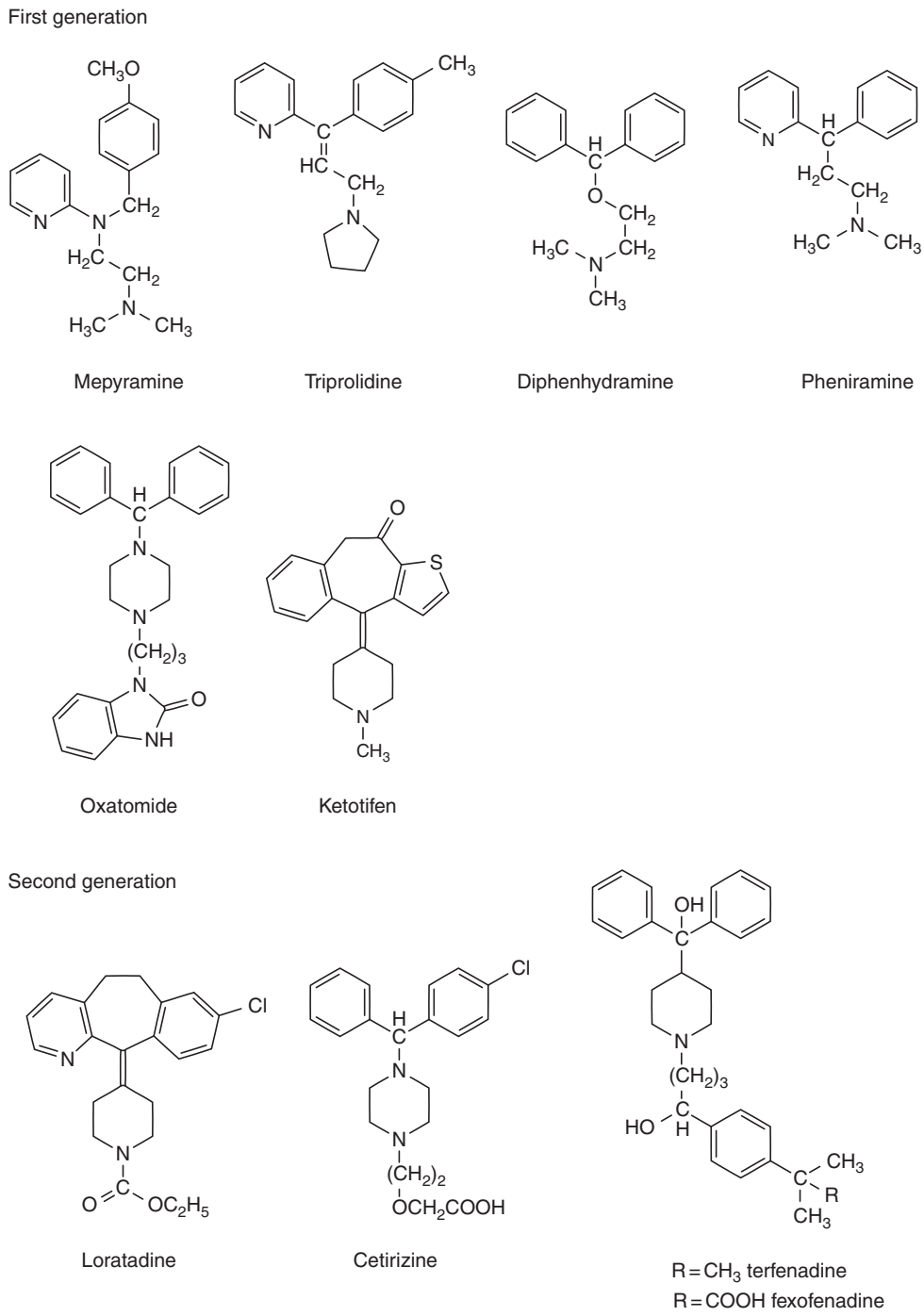

Figure 2 Representative blockers of first-generation (sedating) and second-generation (nonsedating) antihistamines.

in which the replacement of a selective, radiolabeled ligand, bound to that binding system, is measured. These radiolabeled ligands are also very useful for studying the distribution and localization of H_1 receptors throughout biological systems. Several H_1 blockers labeled by a radioactive isotope have been developed and are readily available. [^{3}H]Mepyramine is widely used, and [^{3}H]doxepin is used to a lesser extent. More useful, however, because of a higher specific radioactivity, are ^{125}I-labeled ligands; compounds of this type include [^{125}I]iodobolpyramine and [^{125}I] iodoazidophenpyramine (Figure 4).

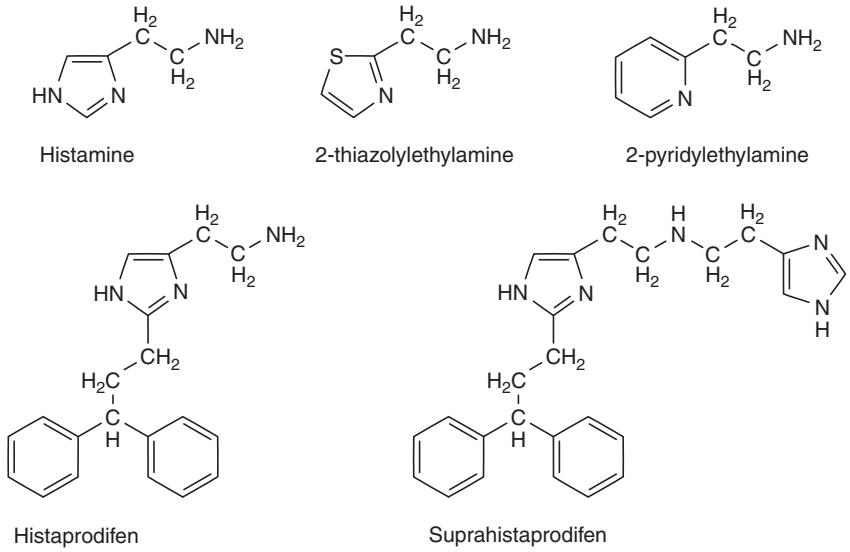

Figure 3 H_1 agonists used as pharmacological research tools.

Ligands for positron emission tomography studies The introduction of positron emission tomography (PET) made it possible to study the interaction of ligands with their binding sites in human experiments. For histamine H_1 blockers this technology has been used to prove that the sedating effect of H_1 blockers is directly related to their binding to receptors. For this purpose the physicochemical properties of the ligands need to fulfill precise requirements: good penetration into the particular biosphere is crucial. A number of derivatives have been developed: [^{11}C]doxepin and [^{11}C]mepyramine have been successfully used to study the role of histamine receptors in brain functions.

The Histamine H_2 System

The Receptor, Mechanisms, and Role in Physiology and Pathology; H_2 Ligands as Medicines

The H_2 receptor is a G_s-protein-coupled receptor; upon receptor stimulation, through activation of an adenylyl cyclase, the intracellular concentration of c-AMP is increased. For about two decades after the introduction of antihistamines (i.e., H_1 blockers) for therapeutic purposes, it remained unexplained why active blockers antagonized the effects of histamine in the soft tissues (e.g., the airway system and blood vessels), but not the effects of histamine in the heart (increase in rate) and stomach (gastric acid production). After the success story of the distinction of α- and β-adrenergic receptors by Ash and Schild, and following the introduction of selective ligands for these receptors, Ash and Schild proposed the theory of the existence of two classes of histamine receptors. Subsequently, Black (who also worked on selective ligands for the β-receptors) proved that this concept was right, by identifying selective ligands for the two types of receptors; from then on these receptors were called histamine H_1 and H_2 receptors. The classical antihistamines were found to be selective blockers of the H_1 receptors, which were identified in the airway system and on small blood vessels; via these receptors histamine causes bronchoconstriction and an increased permeability of capillaries of the vascular system. Black's concept became generally accepted when his group synthesized selective H_2 blockers; later on, most of these compounds were identified as inverse agonists.

Histamine H_2 agonists have some medicinal application for inducing gastric acid production in patients with an achlorhydria; the potential use of H_2 agonistic impromidine analogs as cardiac stimulants (see later), as an alternative for β-agonists, has not moved beyond experimental status. The major obstacle is the intrinsic side effect of an increase in gastric acid secretion; some selectivity (heart vs. stomach) has proved possible, though. Another problem is caused by a species difference in receptor affinities of the analogs, making the selection of the best candidate for clinical evaluation difficult.

Histamine H_2 blockers are very effective in reducing gastric acid secretion. As antiulcer medicines (cimetidine, ranitidine, famotidine) they achieved major market blockbuster status in the 1980s, having the highest sales ever for a single type of drug during that time. In the 1990s this status was taken over by the

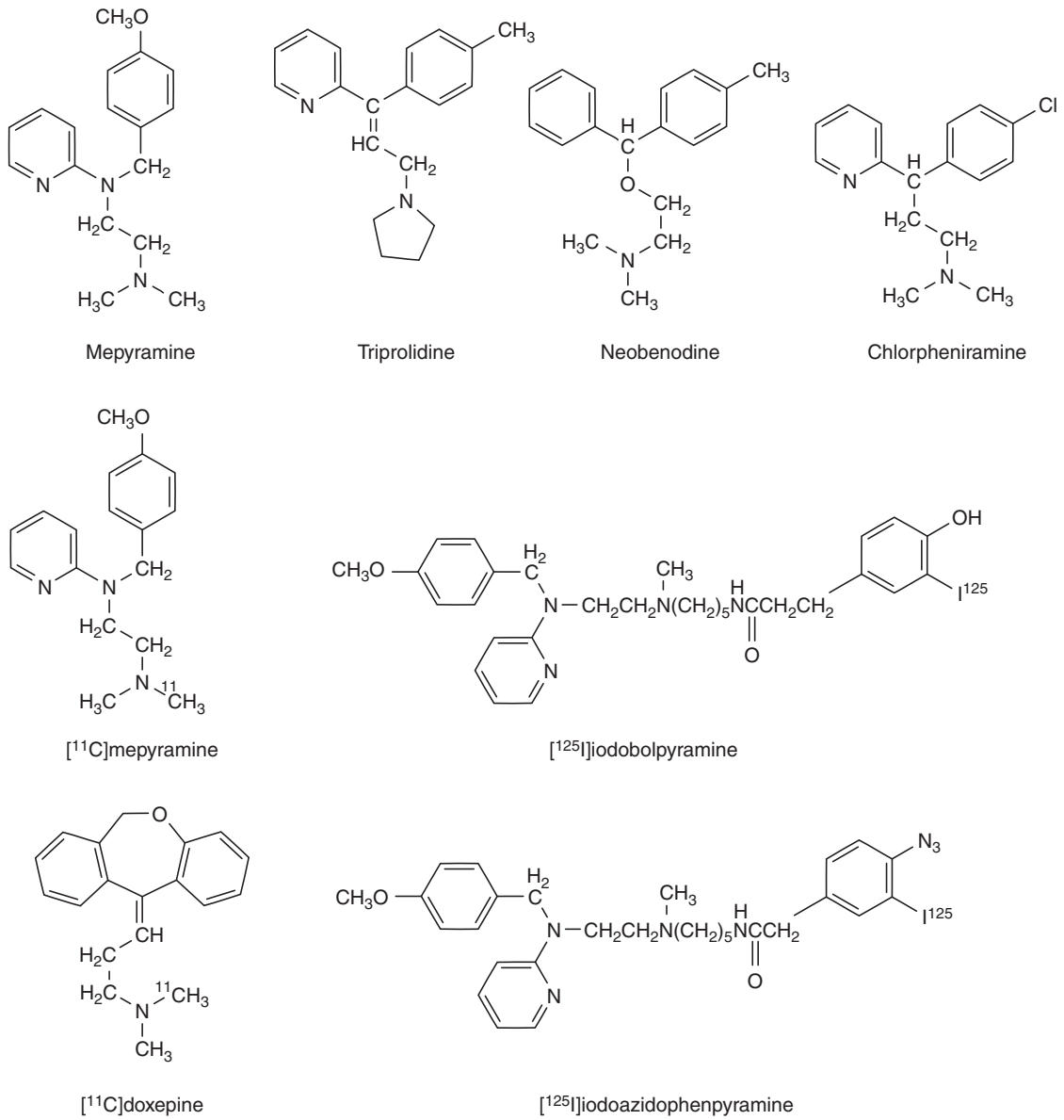

Figure 4 H_1 blockers used as pharmacological research tools.

proton pump inhibitors and anti-*Helicobacter* agents. Collectively, the highly effective antiulcer agents have more or less definitely solved the problem of gastric ulcers, and surgery is no longer necessary.

Histamine H_2 blockers do not influence heart rate; histamine does not play an important physiological role in this respect. Although H_2 receptors are present in several CNS areas and whereas brain-penetrating H_2 blockers have been found, no useful property for medicinal application in brain diseases has been identified for H_2 blockers.

The H_2 receptors show constitutive activity and it has been found that most H_2 blockers are inverse agonists, with burimamide, the first H_2 blocker reported, being an exception; this derivative is a true, neutral antagonist, antagonizing both H_2 agonists and H_2 inverse agonists with the same level of activity.

Compounds; Structure–Activity Relationships

Agonists Histamine has a low affinity for the H_2 receptor (K_D around 4.5 nM); as has been seen for H_1, activity is observed at much lower concentrations, with a pD_2 of about 6.2 at heart H_2 receptors, causing a high incidence of spare receptors at H_2 (the pD_2 value is the negative log of the molar EC_{50} value of an agonist). Again, as established for H_1 agonism, the structure of the physiological stimulant of the

H_2 receptor is very sensitive to changes. A certain selectivity over H_1 and H_3 (but not H_4; see later) with a minimal loss of agonistic efficacy has been reached by introducing a methyl group on position 5 of the imidazole nucleus (Figure 5). Empirically it was established that the tautomeric equilibrium of the imidazole nucleus is favorable for H_2 stimulatory activity (Figure 6). Derivatives with a substituent on the imidazole shifting the tautomeric equilibrium to one side have much reduced H_2 agonistic properties; 5-NO_2-imidazole is an example. It was proposed that upon binding of the agonist to the receptor, this equilibrium induces a proton shift in the receptor protein subsequently causing a change in receptor conformation and finally activation of the receptor.

By empirical approaches several H_2 agonists were established and have become important research compounds. The first two were dimaprit and impromidine (Figure 7). Both compounds fulfill what was considered an important criterion – that a tautomeric equilibrium would be needed for H_2 agonism. When the isothiourea group of dimaprit was 'closed,' thiazole derivatives were obtained, which on proper substitution, yielded a new class of selective H_2 agonists, of which amthamine became an important research tool in pharmacology (Figure 7). The agonistic activity of compounds like amthamine was reason to abandon the 'proton shift' theory. In contrast with histamine, no spare receptor capacity has been found for amthamine; this finding shows nicely that

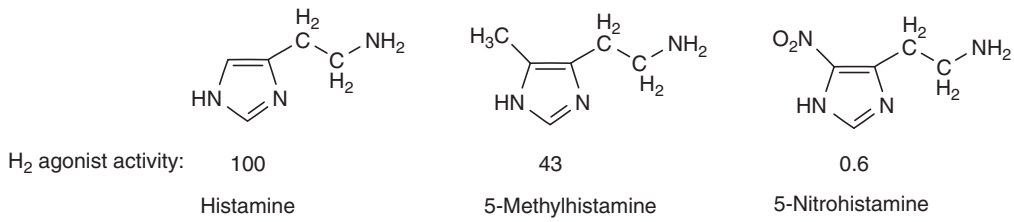

Figure 5 The influence of substituents (histamine derivatives) on H_2 agonistic activity.

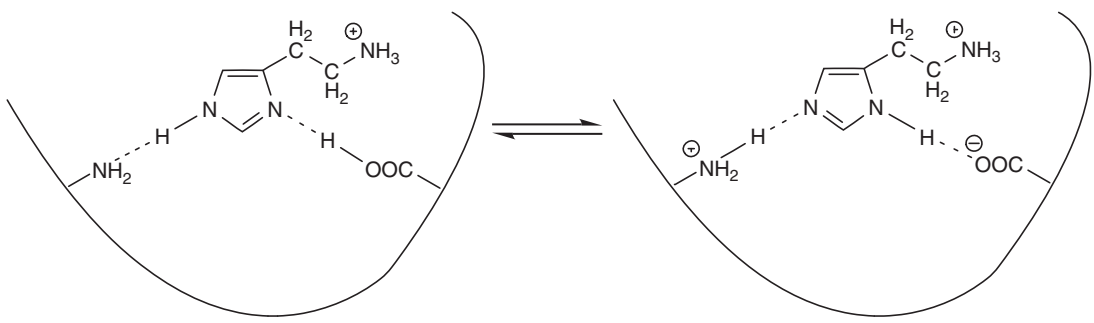

Figure 6 The tautomerism of the imidazole nucleus of histamine for activating the H_2 receptor.

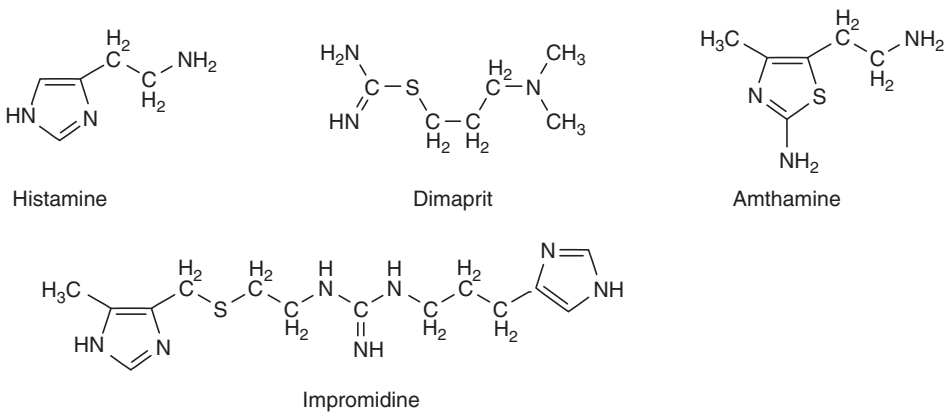

Figure 7 Histamine and selective ($H_2 > H_1$) H_2 agonists.

a spare receptor capacity is a property of a compound and not of a receptor system.

The molecular structure of impromidine is of particular interest; at regions I, II, and III (Figure 8)

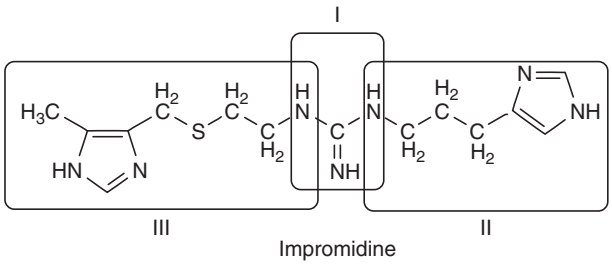

Figure 8 The different regions of the impromidine molecule and their roles in activating the H_2 receptor: I is for binding to a negative locus at the receptor, II is for activating, comparable to the imidazole nucleus of imidazole, and III contributes extra affinity for binding.

tautomeric equilibriums are possible; by structural modifications it has been established that region II is essential for H_2 agonism. This observation has led several investigators to modify region III in such a way that compounds with dual activity are obtained, yielding H_2 agonists with an extra (e.g., an H_1 blocking) property (Figure 9).

Histamine H_2 blockers Whereas H_1 blockers were initially found by applying random screening procedures, H_2 blockers were first identified by a more rational approach. Black et al. had found β-adrenergic antagonists by masking the hydroxyl groups in adrenaline. A comparable attempt using histamine led to inactive derivatives; when Black and group then changed the side chain of histamine, they eventually found a class of H_2 blockers, of which

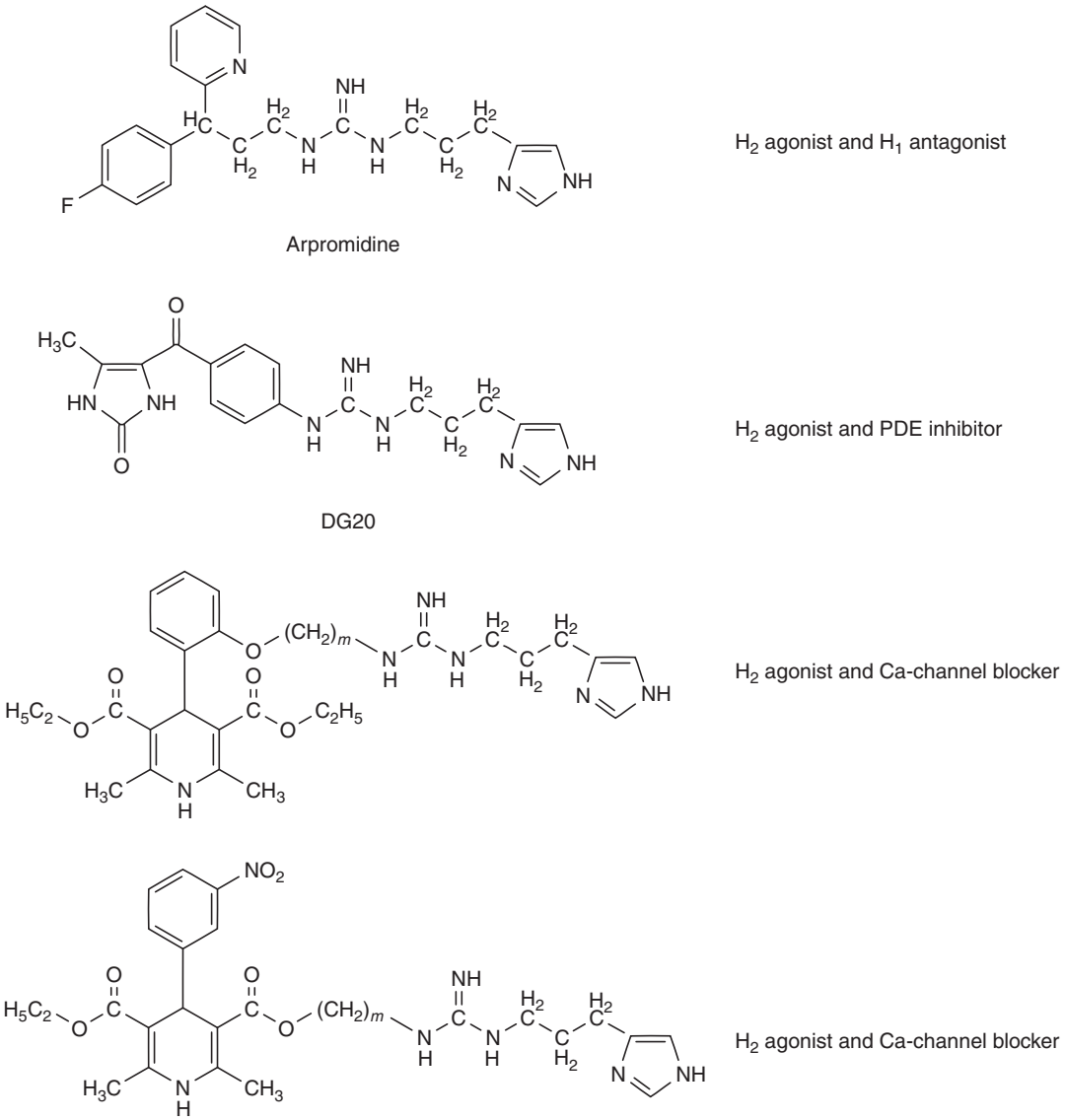

Figure 9 Impromidine derivatives with additional pharmacological properties (PDE, phosphodiesterase).

cimetidine (Figure 10) was the first to reach the market as an antiulcer agent. The development of cimetidine was based on the assumption that the imidazole part of the molecule was crucial to the blocking properties; this assumption was not true, as became clear when it was found that nonimidazole ligands such as ranitidine and famotidine (Figure 10) had higher H_2 blocking activities, and this became important as an active principle regarding medicines effective against stomach ulcers. Because the imidazole group – as in cimetidine – causes an interaction with P450 enzymes, the favored market position of cimetidine, which had become rapidly a blockbuster medicine, was gradually taken over by the nonimidazole agents.

H_2 Ligands as Research Tools

H_2 agonists When compounds such as dimaprit, impromidine, and amthamine became available, the classic histamine derivatives (e.g., 5-methylhistamine) became obsolete as H_2 agonists. In addition to an H_2 agonistic effect, both dimaprit and impromidine also have an affinity for the H_3 and the H_4 receptors, and the selective amthamine offers advantages for pharmacological studies. Amthamine has no or very weak effects on other receptor systems; for example, the compound shows no adrenergic effect *in vitro* or *in vivo*.

H_2 blockers Compounds such as cimetidine, ranitidine, and famotidine are attractive therapeutically and are very valuable research tools. The compounds are rather selective, having no significant affinity for the other three histamine receptors. There is no example available of an active H_2 blocker carrying a chiral center.

Radiolabeled ligands When the H_2 system had been identified, there was an immediate need for a radiolabeled blocker for pharmacological studies. Cimetidine was labeled with tritium, but this ligand was unsuitable as a research tool due to extreme levels of aspecific binding, not only to proteins but also to glass surfaces. Tritium-labeled tiotidine (Figure 11) was somewhat better, but far from ideal. A very useful derivative came from a new class of blockers synthesized by Schunack and group; iodoaminopotentidine, labeled with ^{125}I, has a very high H_2 affinity (pK_D around 9.5) and a low level of specific binding. Interestingly, the iodoazido analog is an irreversible H_2 blocker and is also useful in receptor studies.

Ligands for PET studies As no use has been found for H_2 ligands in the CNS, no H_2 PET ligand has been developed.

The Histamine H_3 System

Receptors, Mechanisms, and Role in Physiology and Pathology; Potential Therapeutic Applications of H_3 Ligands

The histamine H_3 receptor was first identified during the last years (the early 1980s) of the pregenomic era, by classical pharmacological investigations, as a presynaptic autoreceptor, influencing both histamine synthesis in and release from histaminergic CNS neurons. It was subsequently found that the receptor occurs also as a heteroreceptor on other types of neurons (e.g., those releasing dopamine, serotonin, acetylcholine) in the CNS. The CNS H_3 receptors reach in high densities in distinct areas (basal ganglia, hippocampus, and cortex) but occur in certain peripheral systems as well (gastrointestinal tract, blood vessels, and airways).

After the receptor was described, it took relatively long before the H_3 gene was cloned. The gene was

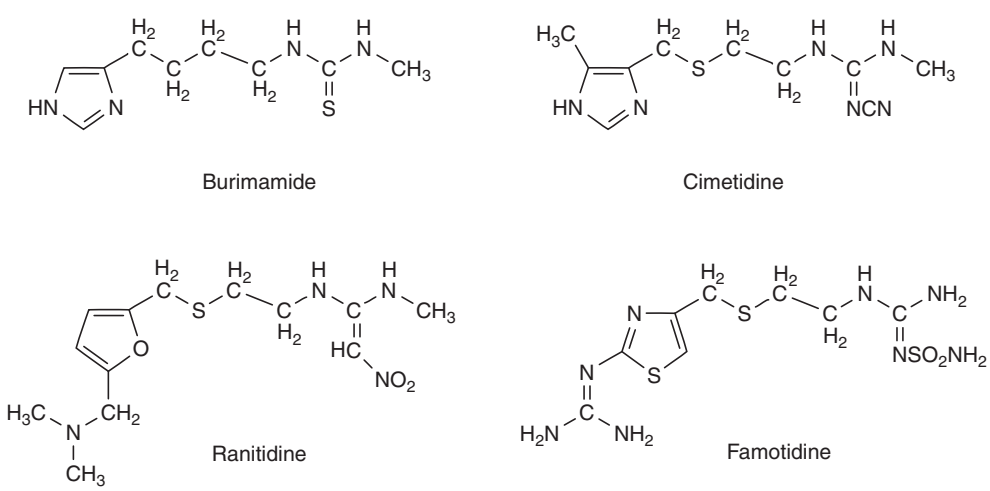

Figure 10 Selective H_2 blockers.

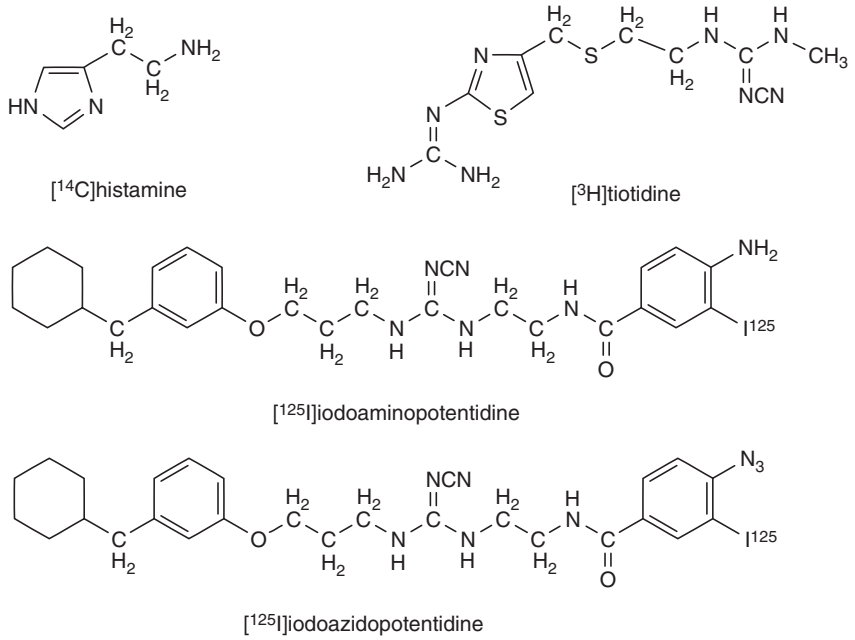

Figure 11 Radiolabeled ligands of the H_2 receptor.

identified by researchers at Johnson & Johnson using an expressed sequence tag (EST) database. The receptor belongs to the GPCR family, being linked to $G_{i/o}$, leading to a lowering of the intracellular cAMP levels. Relatively high degrees of species differences have been established, making predictive pharmacology complicated. This became very apparent when it was found that the H_3 receptor shows constitutive activity. Indeed, agonists, inverse agonists, and (true) antagonists are available. Certain ligands are classified as so-called protean agonists, meaning that in some test systems they are agonists and in others they are inverse agonists. An example is proxyfan (Figure 12). Given all of these different aspects, the prediction of clinical applications for both agonists and antagonists is extremely difficult; this matter has been nicely reviewed by Letavic and colleagues. Interestingly, in this case inverse agonism has been shown to also occur *in vivo*. In animal experiments, both thioperamide and clobenpropit act as inverse agonists. Various isoforms of the H_3 receptor have been identified in guinea pigs, rats, and humans. These isoforms, resulting from alternative splicing, show distinct expression in several brain nuclei and signaling routes.

An extremely wide variation of biological effects mediated via H_3 receptors emerged when selective ligands for the new receptor became available. Based on the pharmacological effects, many potential therapeutic applications have been proposed, but despite intensive research programs by a large number of pharmaceutical companies, no H_3 ligand has yet reached the market as a medicine. Agonists have been

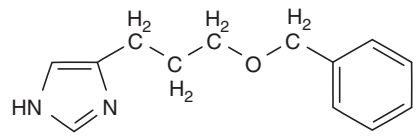

Figure 12 Proxyfan, a protean H_3 agonist.

suggested as having potential benefits in treating asthma, pain, and insomnia; antagonists are thought to be of value in treating cognitive problems, obesity, neuropathic pain, allergic rhinitis, narcolepsy, and schizophrenia. Maikel Wijtmans and colleagues have recently published a comprehensive survey on therapeutic roles that have been proposed for H_3 ligands.

Ligands of the H_3 receptor carrying an imidazole moiety likely also have considerable affinity for H_4 receptors. Although this should be seriously considered in experimental studies, the implications for potential therapeutic use seem to be limited, as H_3 ligands are targeted mainly to the CNS, and so far no role for H_4 in the CNS has been reported. However, the search for nonimidazole ligands has been very intensive and successful; because such compounds are less likely to interact with P450 enzymes, there is major interest in continuing research on their therapeutic application.

Compounds; Structure–Activity Relationships

H_3 agonists Comparing the activity of histamine on H_3 receptors with its activities on H_1 and H_2 receptors indicates that H_3 activity occurs at relatively

low concentrations of histamine; the observed pD_2 is about 7.5, being a log unit higher than the comparable parameters for H_1 and H_2 (the pD_2 values were measured as a reduction of the outflow of histamine upon electrical stimulation from isolated neuronal tissue). Even more striking is that substitutions in the side chain of histamine with small alkyl groups, which reduce both H_1 and H_2 agonistic properties, strongly increase H_3 agonism. Surprisingly, a basic moiety is not required for H_3 agonism: several 4-alkyl-substituted imidazoles have a moderate H_3 receptor-activating property. From the data of Table 1 it is apparent that H_3 agonists have no or low spare receptor capacities, a feature of many H_1 and H_2 agonists. The H_3 agonist α-methylhistamine has a pD_2 value of about 8.5. An additional feature of the latter compound is the presence of a chiral center, making this derivative very useful as a pharmacological tool (see later).

Table 1 H_3 agonism of histamine and some 4-substituted imidazoles, including compounds without a basic side chain

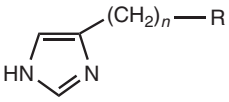

R	*n*	$pK_i \pm SEM$[a]	$pEC_{50} \pm SEM$[b]
NH_2 (histamine)	2	8.0 ± 0.1	8.0 ± 0.0
NH_2	3	7.3 ± 0.0	7.4 ± 0.0
OH	3	6.5 ± 0.1	6.1 ± 0.1
CH_3	3	7.6 ± 0.0	7.5 ± 0.3
CH_3	4	7.6 ± 0.1	7.3 ± 0.3
CH_3	5	7.5 ± 0.1	7.6 ± 0.4

[a]Displacement of [^{3}H]N^{α}-methylhistamine from human H_3 receptors expressed on SK-N-MC cells.
[b]Inhibition of cAMP production in β-galactosidase assay (SK-N-MC cells expressing human H_3 receptor).

Following identification of α-methylhistamine as a research tool (the compound has, however, also been investigated in several disease models), many other agonists selective over H_1 and H_2 were identified, several with a high level of activity; the most interesting ones are summarized in Table 1 and Figure 13. Because of its selectivity (H_3 over H_4), immethridine is of particularly great interest, as is methimepip. All of the listed compounds are substituted imidazoles and thus are not attractive as medicinal agents, because of poor brain penetration capacities and, even more importantly, because imidazole moieties are likely to interact with P450 enzymes, with consequences for drug–drug interactions. The poor brain penetration has been overcome by making a prodrug of α-methylhistamine (Figure 14); designed by Walter Schunack and group (Berlin), BP 2-94, upon oral administration in animals, reaches active levels of the parent α-methylhistamine in the CNS. In animals, the compound also has interesting anti-inflammatory and antinociceptive effects (no clinical data are available).

H_3 antagonists Shortly after the identification of the first H_3 agonists, reports of a number of active antagonists were published. The archetype H_3 antagonist is thioperamide (Figure 15), introduced by Jean-Charles Schwartz and co-workers in 1987.

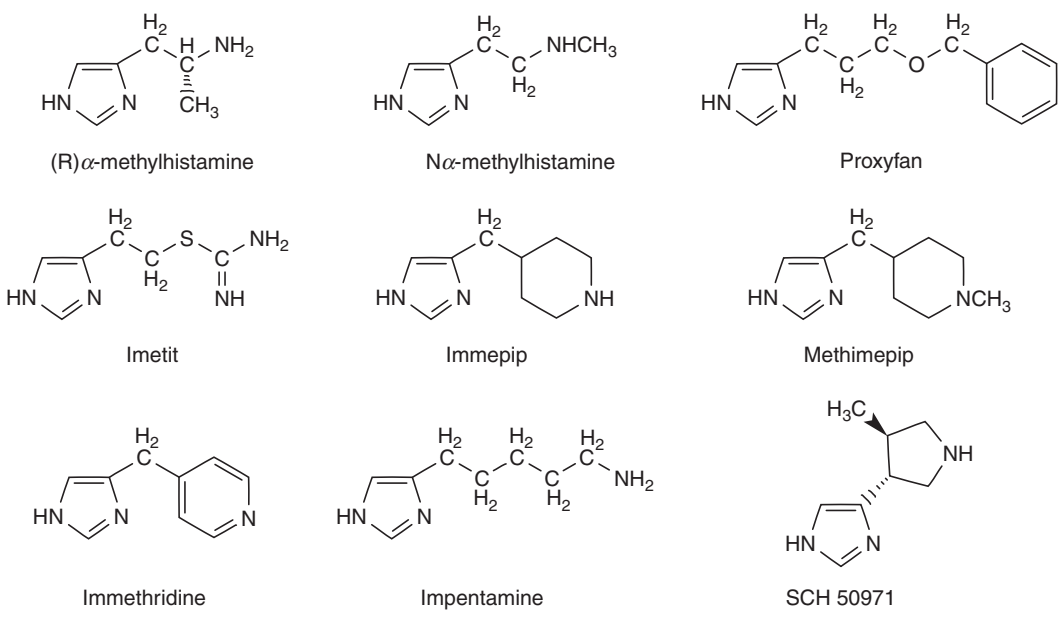

Figure 13 Examples of selective H_3 agonists.

Soon thereafter many other H_3 blockers appeared on the scene. A recent review has been published by Wijtmans et al. All early compounds were imidazole derivatives, causing the same limitations for practical use as mentioned previously for the agonists. Highly potent blockers eventually became available, with K_D values in the low nanomolar range. A selection of such ligands is shown in Figure 15.

The H_2 agonists dimaprit and impromidine (Figure 7) both show appreciable H_3 blocking effects. The route of synthesis of the impromidine structure is complicated, whereas dimaprit analogs are easy to prepare; consequently, large numbers of the latter type of compound have been investigated and many highly active derivatives have been identified (Figure 15). The active compounds share a high level of lipophilicity; this feature, together with the presence of the imidazole nucleus and an isothiourea moiety (known to be toxic), make this type of compound unattractive for therapeutic application. These compounds have proved to be highly useful research tools, though. It is again striking that no basic group is required in the side chain. An H_3 blocker with a somewhat unusual structure is the compound known as GT-2331 (Figure 16). This derivative resulted from an optimization program screening potential lead compounds from natural sources; it was tried as a blocker in CNS indications, but failed. The cause of the failure is likely due to species differences; the compound was identified in rats as a blocker, but in humans it functions as an H_3 agonist.

For obvious reasons many nonimidazole H_3 blockers are under development and are being intensively investigated for several (mainly brain) disorders; comprehensive reviews by Wulff and Wijtmans are available. Most interesting are the nonimidazole derivatives; a selection of representative examples is shown in Figure 17. There is also a certain interest in ligands which have both H_1 and H_3 antagonistic properties; such compounds are thought – on the basis of their pharmacological profile – to be of use in certain allergic conditions. An example is the compound shown in Figure 18, which has nanomolar affinities for both H_1 and H_3 receptors.

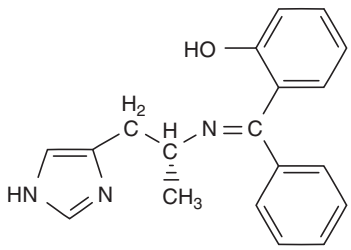

BP 2–94

Figure 14 The potent H_3 agonist (*R*)α-methylhistamine prodrug, BP 2-94.

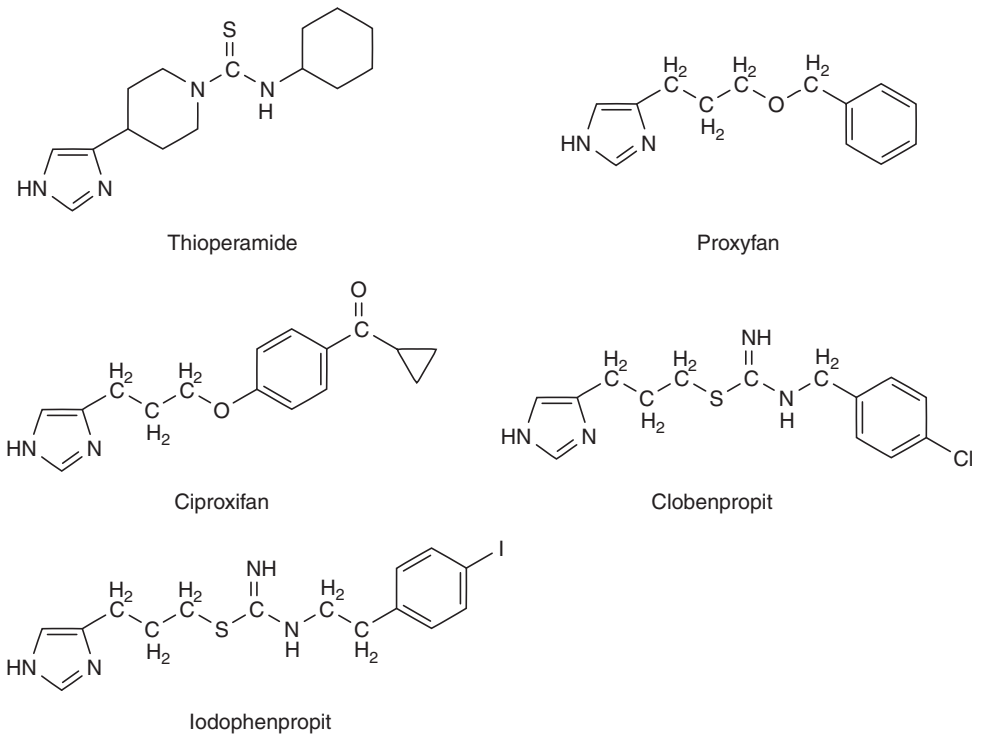

Figure 15 H_3 antagonists of the imidazole type.

A very complicating feature of several ligands is the extreme level of species differences; some compounds (impentamine, for example, or the already mentioned GT-2331) identified in rat studies as antagonists appear to be agonists on the (cloned) human receptor; currently the human receptor is used as the standard model for establishing an H_3 profile.

H_3 Ligands as Research Tools

H_3 agonists As already discussed, a large number of active and selective (H_3 over H_1 and H_2) agonists is available. A striking feature of the structure–activity relationship of the agonists is the role of chirality. The availability of pairs of active and almost inactive ligands such as (*R*)α-methylhistamine and (*S*)α-methylhistamine has been very instrumental in unraveling the several roles of H_3 receptors. Other widely used H_3 agonists in pharmacological studies are immepip and imetit.

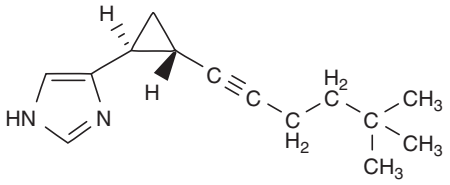

GT-2331

Figure 16 GT-2331, an H_3 antagonist with an unusual structure.

H_3 antagonists Blockers of the H_3 receptors are also readily available. Compounds such as thioperamide, clobenpropit, and ciproxifan (Figures 15 and 16) are frequently used as highly potent H_3 blocking agents. Pairs of active and inactive enantiomers of H_3 blockers are not available, but this has not hampered pharmacological investigations of the H_3 system.

Radiolabeled H_3 ligands Soon after the identification of α-methylhistamine as an H_3 agonist, a tritium analog, labeled at the methyl group, was introduced as an attractive tool for ligand displacement studies. The N^{α}-methyl derivative with a tritium label in the methyl group also became available, but [^{3}H] (*R*)α-methylhistamine became the standard ligand

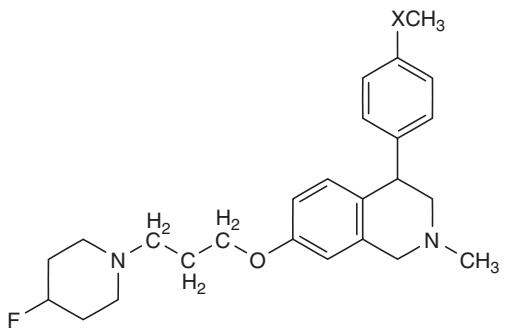

X = O, S

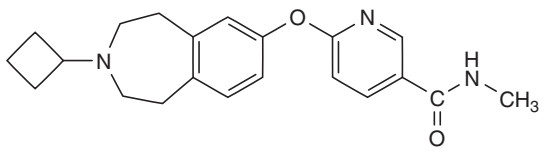

GSK 189254

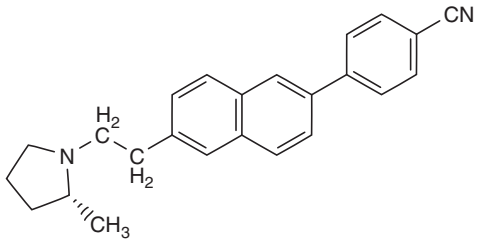

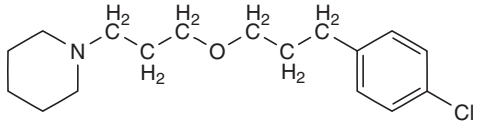

BF 2.649

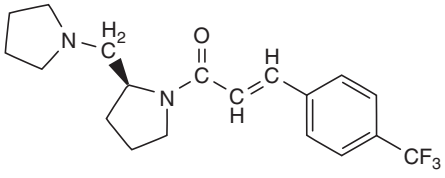

NNC 38–1202

Figure 17 Nonimidazole H_3 antagonists.

(Figure 19). The use of an agonist as radiolabeled ligand in ligand displacement studies has an important disadvantage that the agonistic activities may cause induction of different activation states (high and low affinity) of the receptor. Several research groups have therefore tried to identify, for the purpose of radiolabeling, suitable H_3 blockers. Successful compounds include the ^{125}I analogs of iodophenpropit and iodoproxyfan (Figure 19) (complicating this effort, iodoproxyfan has been found to be an agonist for human H_3 receptors). Besides having the advantage of being antagonists, these compounds are of great use because of the high specific radioactivity of the iodo label; a disadvantage of these compounds is their high level of lipophilicity, causing nonspecific binding to biological preparations. Both ^{125}I derivatives have also been applied to visualize H_3 receptors in the CNS. Nonimidazole inverse agonists have also been radiolabeled (tritium) for use in receptor characterization studies.

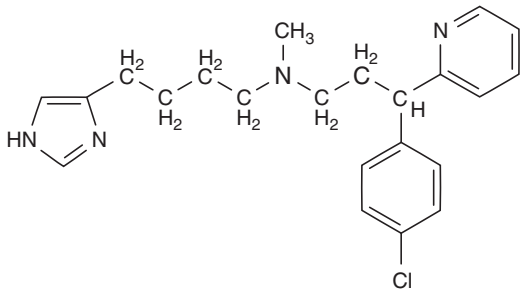

Figure 18 An H_3 blocker with additional H_1 blocking activity.

H_3 PET ligands There was an early demand for suitable PET ligands to study the pharmacology of H_3 ligands and to identify their potential medical applications. It proved difficult to find ligands that fulfill the criteria of high affinity, good CNS penetration, fast on/off kinetics at the receptor, and a 'doable' route of synthesis; there were several failures in pursuit of these characteristics. Recently, however, an interesting ligand was reported by research teams at Johnson & Johnson and the Vrije Universiteit Amsterdam. This derivative (Figure 20) could become an attractive tool to further unravel the role of H_3 receptors in healthy brains and in pathological conditions, as it fulfills all of the aforementioned criteria.

The Histamine H_4 System

The H_4 Receptor, Mechanisms, and Role in Physiology and Pathology; Potential Use of Ligands as Medicines

The newest histamine receptor to date, H_4, was not first identified by pharmacological means, as the other three receptors were. The gene for the H_4 receptor was found in 2000 by Tamaki Oda and colleagues by searching a gene bank using the sequence of the H_3

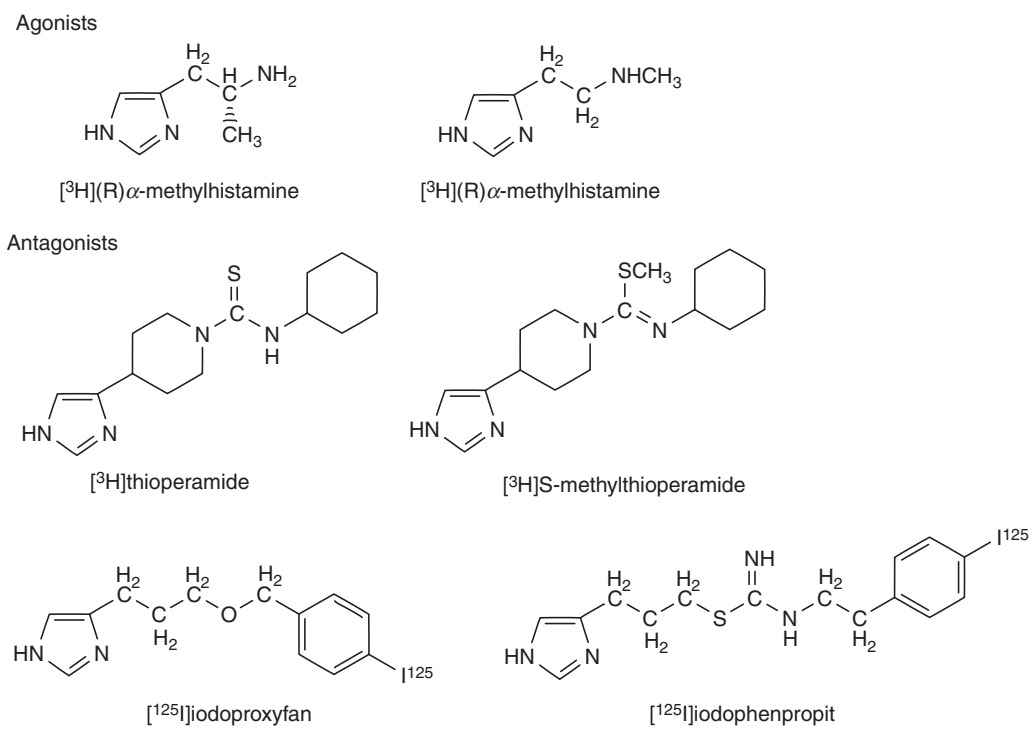

Figure 19 Radiolabeled ligands of the H_3 receptor.

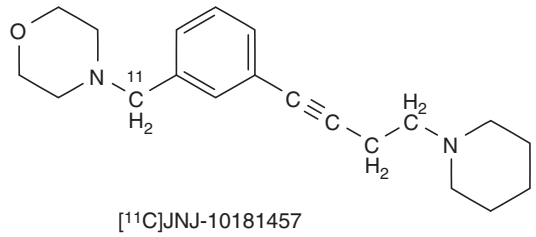

Figure 20 A positron emission tomography (PET) ligand of H_3 receptors.

gene; they subsequently identified H_4 as a new type of histamine receptor. The receptor is linked to a $G_{\alpha i/o}$ signaling system, inhibiting intracellular cAMP synthesis. In eosinophils and mast cells H_4 activation leads to increased intracellular Ca^{2+} levels through a phospholipase-mediated process. As for the other three histamine receptors, constitutive activity is observed for H_4.

The H_4 receptor (protein) shows a relatively high level of homology with the H_3 receptor. Comparable to what has been observed for H_3, spliced isoforms occur for H_4, and as is the case for H_3, important species variation characterizes the H_4 system; consequently, ligand activities vary from species to species. As an example, histamine is about 25 times more active on the human receptor than on the rat H_4 receptor. As a consequence of the H_3/H_4 homology, several H_3 ligands show effects on H_4 as well (this is especially true for imidazole derivatives); the opposite also applies.

The H_4 receptor is especially expressed in bone marrow and peripheral leukocytes; the mRNA is found in eosinophils, mast cells, spleen, and, at a lower level, in the small intestine. Perhaps because methods of analysis are not absolutely reliable, conflicting results have been published on the distribution of the mRNA. It has long been recognized that histamine has proinflammatory properties; some, but not all, H_3 blockers antagonize to a certain extent the histamine-induced inflammation, but no effect is observed from H_1 or H_2 blocking agents. From these observations and based on the H_4 receptor distribution, it is attractive to consider this receptor as being the pathway by which histamine obtains its proinflammatory properties. So far, pharmacological data are available only for H_4 ligands.

There is ample evidence from *in vitro* experiments that H_4 receptors mediate processes playing a role in inflammation. It is quite remarkable that the CC chemokine ligand CCL16 also interacts directly with H_4 receptors. The pathways involved seem to be very complex and are not yet completely understood. There are also indications from pharmacological experiments that H_4 receptors may be viewed as targets for medicines to be used in inflammatory bowel disease, rheumatic arthritis, and asthma. Whereas H_1 blockers are intensively applied for treating itching, such preparations are effective only in allergy-induced itching, and not when other mechanisms are involved. In a mouse model in which itching is induced by histamine or an H_4 agonist, H_4 blockers show efficacy.

There is increasing evidence that H_4 receptors mediate chemotaxis for cells that play important roles in allergy and inflammation. H_4 receptor-mediated chemotaxis has been described for mast cells and eosinophils.

A role for histamine in malignant cell growth has long been proposed. The presence of H_1, H_2, and H_4 receptors has been confirmed on cancer cells from the colon. For example, in Caco-2 cells, histamine induces an upregulation of cyclooxygenase (Cox-2) expression which can be antagonized by H_2 and H_4 blockers; based on a prostaglandin E2 assay (PGE2), it is speculated that production of vascular endothelial growth factor and tumor growth may be influenced by a mechanism regulated by H_2 and H_4 receptors.

Compounds; Structure–Activity Relationships

H_4 agonists The reported structural resemblance of the H_3 and H_4 proteins explains a remarkable overlap in affinities of ligands for the two receptors. Most, if not all, H_3 ligands of the imidazole-type derivatives also have affinities for the H_4 receptor. As an example, the H_3 blocker thioperamide also blocks the H_4 receptor, whereas the H_3 blocker clobenpropit behaves as an agonist at H_4 receptors.

A very remarkable, nonselective H_4 agonist is clozapine, a compound that behaves as an antagonist at a variety of GPCRs. Clozapine also blocks the H_1 receptor; the clozapine analog VUF 6884 (Figure 21) is also an H_4 agonist, having a tenfold higher affinity for H_4 compared to clozapine. 5-Methylhistamine is a selective H_4 agonist; this compound was previously described as a moderately active but selective H_2 agonist. The selectivity for H_4 over H_2 is, however, at least 100-fold, with a reported K_i of 50 nM.

Compounds with a dual H_4 and H_3 agonistic activity have been found in a series of dimaprit analogs, originally developed as H_2 agonists. The guanidine analog of dimaprit, VUF 8430 (Figure 21), has comparable affinities for H_3 and H_4, having only very weak affinities for H_1 and H_2, and differs therefore in this sense from 5-methylhistamine; VUF 8430 may serve as an interesting tool in pharmacological research.

H_4 antagonists Soon after the identification of the H_4 receptor it became clear that the imidazole type

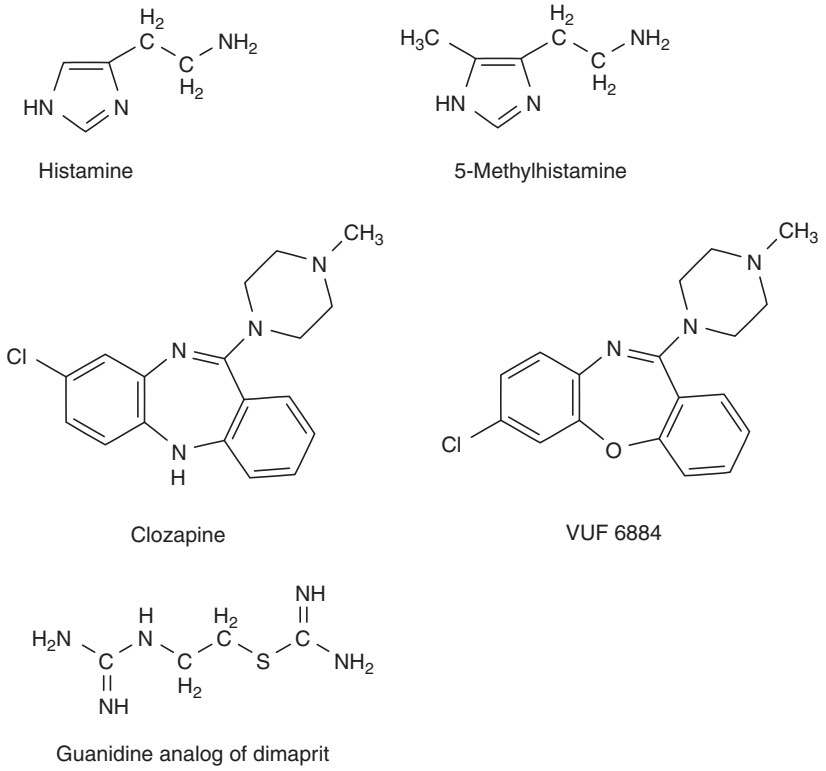

Figure 21 Selected H_4 agonists.

of H_3 ligand binds to the H_4 receptor as well. From the beginning, the search for selective H_4 blockers has therefore been focused on nonimidazole derivatives – in this way avoiding some other disadvantages of imidazoles. Highly active compounds have been found among a series of indoles and benzimidazoles, but other heterocyclic compounds (quinoxalinones, aminopyrimidines) have also been reported; some selected compounds are shown in Figure 22. Most of the published data are available so far in patent literature only, and it is therefore as yet difficult to decide on the practical value of these compounds.

H_4 ligands as tools The pharmacological approaches in H_4 research are rather special, as no simple and easy functional model is available and selective ligands are scarce as well. Cell lines with the expressed human H_4 receptor are the most frequently used model to establish either agonistic (with a reporter gene system) or blocking properties. For confirming agonistic effects, a blockade with an antagonist as JNJ 7777120 (Figure 22) may be applied.

To establish affinity, a ligand displacement technique is used, with tritium-labeled histamine as the hot ligand. This aselective ligand is applied to a population of cells expressing H_4 receptors, but lacking other histamine receptors. In general, an antagonist is preferred as the hot ligand; a study using an H_4 antagonist with a radiolabel has been published, but there is only limited information on this ligand available so far.

Pharmacological Methods to Characterize Selective H_1, H_2, H_3, and H_4 Ligands

The way in which ligands for the four histamine receptors have been identified illustrates in a very nice way the evolution that experimental pharmacology has undergone during the twentieth century – namely, from whole animal pharmacology, through molecular pharmacology with isolated organs, through radioligand displacement technology, to application of cells with expressed (human) receptors, including reporter gene technologies for measuring functional activity. In the following sections the most frequently used methods are briefly presented.

H_1

When the role of histamine in allergy became accepted in the 1930s, and when the idea emerged that histamine blocking agents might be beneficial for treating allergic patients, an intensive search for so-called antihistamines was initiated. The intravenous LD_{50} of histamine was used as a study parameter: any test compound that increased the LD_{50} was considered

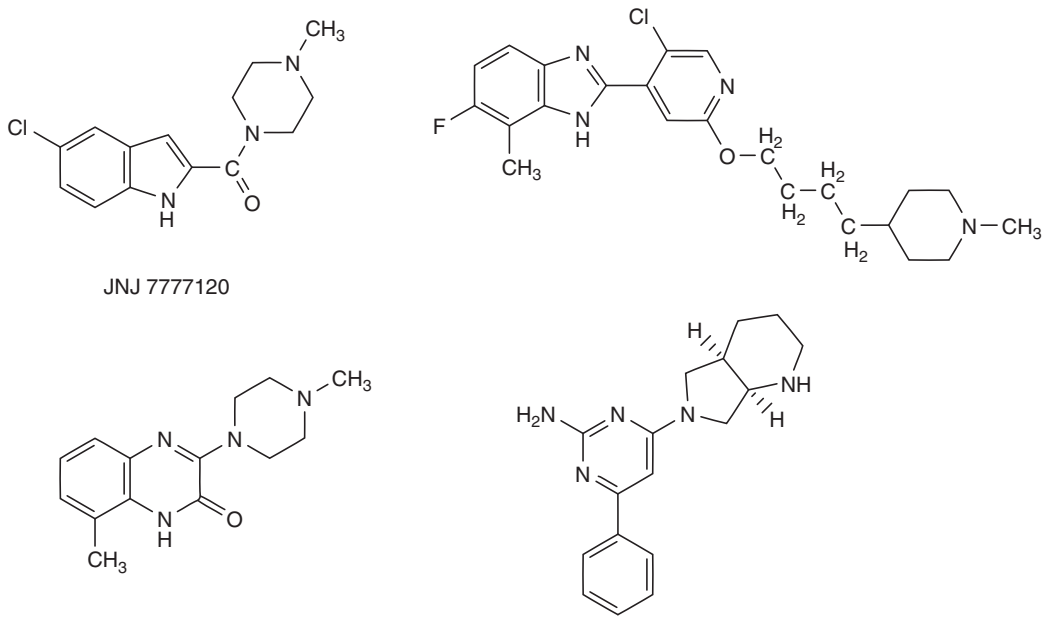

Figure 22 Selected H_4 antagonists.

to be an antihistamine. Because acute death caused by histamine is mediated via H_1 receptors, only H_1 blocking agents were identified; these compounds are currently known as H_1 blockers.

When pharmacology went molecular (1950s and 1960s), the isolated gut of guinea pigs, contracted by histamine, became the standard for pD_2 values, and pD_2 (together with 'intrinsic activity') and pA_2 values became the typical parameters for agonists and antagonists. Somewhat later, isolated cells, first with endogenous receptors and later (1990s) with expressed receptors, became the standard for screening for affinity; functional activity remained to be done with the isolated organ technique. Because the radioligand mepyramine became the standard, the measured affinities, however, were independent of the applied radioligand. Currently, the functional activity of H_1 receptor ligands can be measured on isolated cells by measuring the production of the second messenger IP_3 or Ca^{2+}; a reporter gene (nuclear factor-κB) approach is also available.

H_2

The theory of the existence of more than one class of histamine receptors was originally tested on organs on which histamine was known to exert influences (the heart, increased rate; the stomach, enhanced acid secretion) but which were unaffected by the available antihistamines. Isolated organ preparations of heart and stomach, especially from the rat, have thus been used for a relatively long time and prior to availability of radioligands for reliable ligand displacement studies. The problem with the organ studies was that most H_2 blockers proved to have a high level of aspecific binding, even to the glass of the test tubes used experimentally. The problem was solved with the introduction of radiolabeled aminopotentidine (Figure 11). Another method for functional activity determinations on isolated cells involves cAMP measurement; here too a reporter gene assay (CRE) has been applied.

H_3

H_3 agonism by histamine was first identified by measuring the reduction of the outflow of (radiolabeled) histamine from neuronal tissue, preloaded with radiolabeled L-histidine. Upon stimulation by K^+, antagonism was established by measuring a reduction of the influence of the agonist. As this setup is somewhat tedious, it was replaced by a method in which the relaxation of a piece of isolated gut (guinea pig) contracted by electrical stimulation, as induced by an agonist, is established. For *in vivo* experiments, the reduced outflow of histamine, following application of an agonist in a microdialysis setup, proved to be very useful. The same is true for a method in which the reduced outflow of histamine, achieved by applying an agonist, is followed by measuring the concentration of the level of the major metabolite of histamine, N^{τ}-methylhistamine.

For ligand displacement studies the tritium-labeled agonist α-methylhistamine is widely used; the radiolabeled antagonists (^{125}I label) iodophenpropit and iodoproxyfan are also often applied. Due to the large species differences seen for this receptor, the cells with expressed human H_3 receptors have a strong preference. Currently, determination of functional activity is

often accomplished using an assay in which the binding of GTPγS is measured.

H_4

The specific distribution of the H_4 receptor constitutes a specific problem, especially for screening procedures. In such screening procedures cells with expressed H_4 receptors are widely used; in ligand displacement studies, labeled histamine is applied as hot ligand. For establishing agonistic activity, a reporter gene assay is available; in this approach an increased level of β-galactosidase is used as the parameter. Other pharmacological effects are of a rather specific nature and not very suited for screening systems. A reporter gene assay using a G_i-coupled CRE gene, with increased production of cAMP being the parameter, is applied as well. Functional activity can be established by GTPγS binding.

See also: Peptidergic Receptors; Sleep and Sleep States: Histamine Role.

Further Reading

Arrang J-M, Garbarg M, and Schwartz J-C (1983) Autoinhibition of brain histamine release mediated by a novel class (H_3) of histamine receptor. *Nature* 302: 832–837.

Back JW, Duncan WAM, Durant GJ, et al. (1972) Definition and antagonism of histamine H2-receptor. *Nature* 236: 385–390.

Cvetovic M, Leake B, and From MF (1999) OATP and p-glycoprotein transporters mediate the cellular uptake and excretion of fexofenadine. *Drug Metabolism and Disposition* 27: 866–871.

Hill JS, Ganellin CR, Timmerman H, et al. (1997) International Union of Pharmacology XIII. Classification of histamine receptors. *Pharmacological Reviews* 49: 253–278.

Letavic MA, Barber AJ, Dvorak CA, et al. (2006) Recent medicinal chemistry of the histamine H_3 receptor. *Progress in Medicinal Chemistry* 44: 181–206.

Leurs R and Timmerman H (eds.) (1998) *The Histamine H_3 Receptor; a Target for New Drugs*. Leiden-Amsterdam: Elsevier.

Lim HD, Smits RA, Leurs R, et al. (2006) The emerging role of the histamine H_4 receptor in anti-inflammatory therapy. *Current Topics in Medicinal Chemistry* 6: 1365–1373.

Lovenberg TH, Roland BL, Wilson SJ, et al. (1999) Cloning and functional expression of the human histamine H_3 receptor. *Molecular Pharmacology* 55: 1101–1107.

Oda T, Morikawa N, Saito Y, et al. (2000) Molecular cloning and characterization of a novel type of histamine receptor preferentially expressed in leukocytes. *Journal of Biological Chemistry* 275: 36781–36786.

Smit MJ, Leurs R, Alewijnse AE, et al. (1997) Inverse agonism of histamine H2 antagonists leads to antagonist-induced upregulation of the histamine H2 receptor. *Proceedings of the National Academy of Sciences of the United States of America* 93: 6802–6807.

Tagliatela M, Timmerman H, and Annunziata L (2002) Cardiotoxic potential and CNS effects of first generation antihistamines. *Trends in Pharmacological Science* 21: 52–56.

Timmerman H (2006) Histamine H1 blockers: From relative failures to blockbusters within series of analogues. In: Fischer J, and Ganellin CR (eds.) *Analogue-based Drug Discovery*, pp. 401–416. Weinheim: Wiley-VCH.

Van der Goot H and Timmerman H (2000) Selective ligands as tools to study histamine receptors. *European Journal of Medicinal Chemistry* 35: 5–20.

Wijtmans M, Leurs R, and de Esch I (2007) Histamine H_3 receptor ligands break ground in a remarkable plethora of therapeutic areas. *Expert Opinion on Investigational Drugs* 16: 967–985.

History of Neuroscience: Early Neuroscience

C G Gross, Princeton University, Princeton, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Ancient Egypt

The first written description of the cerebral cortex and the first indications that the site of brain injury can determine the nature of neurological symptoms are found in the Edwin Smith Surgical Papyrus. Written in about 1700 BCE, it is a copy and gloss of a much older treatise dating back to about 3000 BCE. Legend ascribes the original treatise to Imhotep, grand vizier of the 3rd dynasty pharaoh, Zoser, and later deified as an Egyptian god of medicine.

In describing a case of skull fracture, the author noted the pulsations of the exposed brain and compared the surface of the brain to the rippling surface of copper slag (which indeed has a gyral–sulcal pattern). The laterality of injury was related to the laterality of symptom, and both aphasia ("he speaks not to thee") and seizures ("he shudders exceedingly") after head injury were described.

Overall, the treatise is a coolly empirical and practical handbook, perhaps for a battlefield surgeon. It contrasts with most subsequent ancient Egyptian medical writings, which are amalgams of mysticism, superstition, and elaborate speculation. In them, the heart, not the brain, is the most important organ, the seat of the mind and the center of intellectual activity. This hegemony of heart over brain was the prevailing view in the Ancient Near and Far East, and advocates of this opinion can be found in Europe into the seventeenth century.

Classical Greece and Rome

Formal, self-conscious science began in the sixth century BCE. with the naturalistic Ionian philosophers (such as Thales and Anaximenes) and the more mystical Pythagoreans. The Ionians, particularly, were responsible for the idea that the universe could be understood by reason: that it consisted of a set of mechanisms that worked in a consistent fashion according to fixed rules.

Several of these pre-Socratic philosophers, such as Empedocles and Democritus, were particularly interested in the sense organs and perception. However, the pre-Socratic most important for the history of neuroscience was Alcmaeon of Croton. He was the first writer to champion the brain as the seat of the senses and the central organ of intellect. According to ancient authorities, he believed.

> the seat of sensations is in the brain. This contains the governing faculty. All the senses are connected in some way with the brain; consequently they are incapable of action if the brain is disturbed . . . the power of the brain to synthesize sensations makes it also the seat of thought: the storing up of perceptions gives memory and belief and when these are stabilized you get knowledge.

Alcmaeon dissected the eye, described the optic nerves and Eustachian tubes, and recognized arteries and veins as blood vessels. More generally, he stressed that health was the state of harmony and disease that of discord of the elements composing the body.

Perhaps the most famous treatise of the Hippocratic corpus and certainly the most relevant one for neuroscience is *On the Sacred Disease* written in the fifth century BCE. The so-called sacred disease is epilepsy, but the author is clear about its natural origins:

> I do not believe that the sacred disease is any more divine or sacred than any other disease, but, on the contrary, has specific characteristics and a definite course . . . It is my opinion that those who first called it sacred were the sort of people we call witch-doctors, faith-healers, quacks and charlatans . . . By invoking a divine element they . . . conceal their ignorance of its nature.

The author is equally clear about the importance of the brain:

> It ought to be generally known that the source of our pleasure, merriment, laughter, and amusement, as of our grief, pain, anxiety, and tears, is none other than the brain. It is specially the organ which enables us to think, see, and hear, and to distinguish the ugly and the beautiful, the bad and the good, pleasant and unpleasant . . . It is the brain too which is the seat of madness and delirium, of the fears and frights which assail us . . . it is there where lies the cause of insomnia and sleep-walking, of thoughts that will not come, forgotten duties, and eccentricities.

The Hippocratic doctors did not practice dissection, and their knowledge of anatomy was slight. Although they made many perceptive observations and even devised some reasonable treatments, their conception of mechanisms was largely a mixture of false analogy and speculation. Epilepsy, for example, was believed to be caused by phlegm (one of the four humors) descending from the brain and preventing pneuma (roughly, 'vital air') from entering the blood vessels.

The influence of both Plato and Aristotle on the development of neuroscience was essentially negative, but in different ways.

Plato (427–247 BCE) argued against every kind of science, preferring pure reason over observation and experimentation, seeking divine principle rather than

natural law, and advocating the study of ideas or ideal forms rather than actual objects. His cosmological schema of the universe and the body, particularly as set forth in the *Timaios*, had a particularly strong and nefarious influence. In it the soul is divided into three parts. The immortal soul, responsible for reason, resides in the head. The superior part of the mortal soul is in the heart, which receives input from sense organs and is the executor of reason. The inferior part, controlling animal desires and emotions, is placed in the liver. This concept of localization of mental functions continued to reverberate for about 2000 years.

Aristotle (384–322 BCE) was the greatest biologist of antiquity, the founder of comparative anatomy, the first embryologist, the first evolutionist, and the first systematic student of animal behavior. Yet he dismissed the brain as wet and cold and devoid of sensation. Rather, the heart was the center of sensation, intellect, and movement. Why did Aristotle reject the views on the hegemony of the brain held by his predecessors such as Alcmaeon, his contemporaries such as Hippocrates, and even his collaborator and successor as head of the Lyceum, Theophrastus (371–287)? Aristotle produced physiological, comparative, embryological, and introspective arguments for his view of brain function. But the crucial approach in his time was the study of brain-injured humans. Both Alcmaeon and the Hippocratic writers were practicing physicians and their evidence on brain function was strictly clinical. Perhaps because Aristotle's father was a physician and he had been slated to become one himself, medicine was one of the few intellectual endeavors in which Aristotle never showed any interest, and clinical medicine was the key to brain function in fourth-century Greece.

The great Museum at Alexandria has been compared to the National Institutes of Health. The Museum was a vast state-supported institute for research including an astronomical observatory, a zoo, botanical gardens, dissecting and operating rooms, more than 100 professors and that wonder of the ancient world, the Library. In two ways it was a continuation and expansion of Aristotle's Lyceum. First, the founder and chief patron of the Museum was the first Ptolemy, who had been, together with his boyhood friend Alexander, a young pupil of Aristotle. Second, its main scientific founders were Demetrios and Strato, who had been students of Theophrastus at the Lyceum. Strato himself was interested in the brain and taught that it and not the sense organs (let alone the heart) was the seat of sensation.

The Greek reverance for the body had made human dissection impossible, or at least illegal, in Athens, but in Alexandria, where dissection of the dead for embalming had been practiced for centuries, this restriction was absent and the systematic study of the human body flourished.

The two great neuroanatomists at the Museum were Herophilos and Erasistratos (both second century BCE). Herophilos distinguished the cerebrum and the cerebellum, described the meninges and sinuses, and provided the first clear description of the ventricles. He believed all nerves originated in the brain and traced their courses to sense organs and muscles. Herophilos distinguished sensory and motor nerves by noting that damage to the former produced loss of sensation and damage to the latter loss of movement.

Erasistratos continued Herophilos's anatomical work and was particularly interested in the application of contemporary physical ideas to the study of neural function. He attributed human superior intelligence to the greater number of convolutions in brain and seems to have carried out experiments on the living brain, perhaps including the human brain. He believed the psychic pneuma entered the brain through hollow sensory nerves providing sensation and that movement was caused by the expansion of muscles due to pneuma carried to them from the brain through the motor nerves. Both views were dominant well into the eighteenth century.

Galen (129–199) was the most important figure in Classical medical science. He represents its culmination, and his views on the body and brain dominated Western thought for more than 1500 years. Galen's work on the nervous system included detailed dissections of the brains of a variety of animals, extensive clinical investigations, and a series of systematic studies of the effects of experimental lesions of various parts of the nervous system.

Galen regarded the brain as the site of sensation and thought and the controller of movement. The sensory nerves were 'soft' and came from the anterior portions of the brain, whereas the motor nerves were 'hard' and came from the posterior portions, particularly the cerebellum. Thus, sensation was a 'central process,' and the characteristics of nerves derived from their central connections.

Galen's studies of the cranial nerves and spinal cord were outstanding. He described seven of the cranial nerves and experimentally determined the functions of several of them. He studied the effects of transections of the spinal cord at various levels and concluded that the spinal cord was an extension of the brain and the conduit of sensory signals from and motor commands to the body below the head. He noted that specific spinal nerves controlled specific

muscles, and had the idea of the reciprocal action of muscles. For the next advance in understanding spinal function we must await Bell and Magendie in the nineteenth century.

Galen's writings on mental and nervous diseases, although often bizarre from our point of view, were no less influential. Mental diseases were attributed to the obstruction of the passage of pneuma in the brain by one of the four humors or sometimes to an excess of a particular humor. For example, melancholia was due to an excess of black bile and epilepsy to an excess of phlegm. He attempted to attribute particular nervous diseases to dysfunctions of specific brain regions. Epilepsy was supposed to involve the posterior ventricles and apoplexy the posterior cerebral matter itself. He ridiculed Erasistratus's correlation of intelligence with the cerebral convolutions, and in general, minimized the importance of the cerebral cortex, a view that remained prevalent until Gall. Although Galen believed the ventricles were important for the passage of pneuma, he vigorously denied that the soul was situated in them, because animals and humans could survive damage to the ventricles.

After his death, Galen's ideas became frozen dogma. Rather than extend his discoveries and experimental innovations, his successors in Europe accepted as undisputed and indisputable his views in every branch of medicine.

Medieval Europe

The advances in understanding the brain in medieval Europe are easy to summarize: there were none. The central feature of the medieval view of the brain was the localization of the mental faculties in the ventricles of the brain. The church fathers were very much concerned with the nonmaterial nature of the soul. Thus, rather than localize the soul, they ascribed loci for Aristotle's classifications of its functions, namely, the functions of the mind such as sensation, reasoning, and memory. Furthermore, they felt that brain tissue was too mundance to act as an intermediary between the earthly body and the heavenly soul, so they placed the mental faculties in the empty spaces in the brain – the ventricles. Sensation was assigned to the anterior ventricle because according to Galen, the anterior cerebrum was soft and impressionable. Cognition and reason were in the middle ventricle, an appropriate central place for the 'digestion' of the sensations. Memory was in the posterior ventricle, as that region of the brain was hard and thus a good place for storage.

During the European Middle Ages, however, much of Classical medicine was kept alive and further developed in Islamic civilization. Thus, the writings of Hippocrates, Galen, and other Greek medical figures usually came to Europe through Syrian, Persian, Hebrew, or Arabic translations and then into Latin.

Vesalius and the Rebirth of Neuroscience

The first person to break the stronghold of Galen and rekindle neuroscience was Andreas Vesalius (1514–64) – the founder of modern anatomy and, with Nicholas Copernicus (1473–1543), the initiator of the scientific revolution. By the beginning of the sixteenth century, the foundations for Vesalius's achievement were well under way. Galen's original works had become available, and the invention of movable type helped disseminate them, as well as more recent anatomical studies. Naturalism in Renaissance art had generated interest in at least superficial human anatomy. Public dissections of the human body in medical schools, originally for forensic purposes, were common and drew large audiences.

Those dissections, however, were largely demonstrations, if not ceremonial celebrations of Galenic anatomy, not investigations of the human body. Indeed, only gradually did even Vesalius realize, and then only partially, that Galen's descriptions tended to be of the brain and body of the ox or barbary ape, rather than of humans. It is only with Vesalius's great work, *On the Fabric of the Human Body* (1543), that the structure of the human body began again to be studied systematically and directly.

Vesalius ridiculed the crude medieval drawings of ventricular localization of mental function, saying, "Such are the inventions of those who never look into our Maker's ingenuity in the buildings of the human body." Although largely accepting Galen's pneumatic physiology, he was modest about the potential of anatomy for understanding brain function: "How the brain performs its functions in imagination, in reasoning, in thinking and in memory ... I can form no opinion whatsoever. Nor do I think that anything more will be found out by anatomy."

The Reflex: From Descartes to Marshall Hall

René Descartes (1596–1650) combined Galenic physiology with a conception of the body as a machine to provide the first idea of reflex action. Within a nerve, Descartes thought, there are thin threads attached at one end to the sense organs. External stimuli pull on the threads to open little gates to the ventricles allowing pneuma to flow back out of the ventricles (reflected) through the same hollow nerves, causing movement by inflating the muscles. The flow of the pneuma is

directed by the pineal gland, extending from the midline into the ventricles. In animals this is a strictly mechanical process. However, in humans, who unlike animals have a soul, the soul interacts with the body at the pineal gland and thus can influence the flow of pneuma to the muscles.

Thomas Willis (1621–75), anatomist, physician, and Oxford professor, took this idea further and related it to actual brain structures. His *Cerebri Anatome* (1664), illustrated by Christopher Wren (1632–1723), was the most complete description of the brain to that date. Sense impressions, Willis speculated, were carried by pneuma within the nerves to the *sensus communis* in the corpus striatum and then on to the corpus callosum and the cerebral cortex, where they were perceived and remembered. However, some were 'reflected' back to the muscles by way of the cerebellum. Thus, voluntary movement was controlled by the cerebrum and involuntary, or 'reflex,' movement by the cerebellum.

The first actual experiments on neural mechanisms of reflexes were carried out by Robert Whytt (1714–66) of Edinburgh. Using frogs, he showed that the spinal cord, indeed only a segment of the cord, was necessary and sufficient for reflex responses to stimulation of the skin. He also demonstrated that the pupillary reflex was dependent on the midbrain. He stressed, more than Descartes had, the protective function of reflexes. Whereas movement was strictly mechanical for Descartes, Whytt believed it was dependent on 'sentient principle,' even when involuntary or reflex, an idea that persisted into the nineteenth century.

The modern concept of a reflex largely began with the English physiologist and physician, Marshall Hall (1790–1857). He used Charles Bell (1774–1842) and Francois Magendie's (1783–1855) distinction of sensory and motor roots to develop the idea of the reflex arc. Reflexes were now, by definition, dependent on the spinal cord, independent of the brain, and strictly unconscious and involuntary. Hall also described the excitation or inhibition of reflex movements by various drugs. Finally, he was the first to use reflexes in medical diagnosis and treatment. The next significant development was CS Sherrington's (1857–1952).

Bioelectricity: From Frog's Legs to Action Potential

By the middle of the eighteenth century the stage was set for Galvani and the beginning of electrophysiology. The Leyden jar as a source of electricity and the electroscope to measure electricity had been invented, the electrical nature of lightning was known, and ideas about animal electricity were widespread among the intelligentsia. Using his frog nerve–muscle preparation, Luigi Galvani (1737–98) produced muscle contraction by stimulation with an electrostatic machine, by atmospheric electricity, by hanging the frog's legs by brass hooks on an iron railing, and by placing the cut end of a nerve on the muscle. He interpreted these observations as reflecting intrinsic animal electricity. Alessandro Volta (1745–1827), on the other hand, in a long and acrimonious controversy with Galvani, explained these and similar results as due to electricity produced by currents generated by two dissimilar metals and quite independent of any animal electricity.

Frederick von Humboldt (1769–1859) finally sorted out the controversy and demonstrated that Galvani was dealing with two independent phenomena, intrinsic animal electricity and bimetallic electricity. Carlo Matteucci (1811–68) went a step further to detect current flow in muscle both with an electroscope and another nerve muscle preparation.

Soon after, Emil du Bois-Reymond (1818–96) differentiated nerve from muscle current. Du Bois-Reymond's theory of electromotive particles lined up along the surface of nerve and muscle was further developed by his student Julius Bernstein (1839–1917) into the beginning of the modern membrane theory of the action potential. Further development awaited JZ Young's (1907–97) discovery of the squid giant axon and AL Hodgkin (1919–98) and AF Huxley's (1917–) brilliant use of it to elucidate the biophysical basis of the action potential.

From Globules to Neurons

Cells were first described (in plants) by Robert Hooke (1635–1703) and in what he called nerve fibers by Anton van Leeuwenhoek (1632–1723), but it was not until the development of achromatic lenses and methods of embedding and staining in the middle of the nineteenth century that the study of the fine structure of the nervous system could begin. Until then, and even somewhat after, nerves were considered to be hollow as Galen had suggested, gray matter was made up of 'globules,' and the relationship between the fibers of white matter and globules was unclear.

Although MJ Schleiden in 1838 pointed out that cells were the basic unit of plant life and T Schwann in 1839 extended this idea to animals, the nervous system resisted an interpretation in terms of their cell theory for at least another 60–70 years. Even after RA Von Kolliker (1817–1905) demonstrated that nerve fibers came from nerve cells and OFK Deiters (1834–63) distinguished axons from dendrites, the nervous system was thought to consist of an

interconnected network. Nerve cells were thought to be nodes of this reticular structure and the fibers originating from them to anastomose completely. J. Gerlach (1820–96) is usually considered the founder of this reticular doctrine, and Camillo Golgi (1844–1920) its most famous advocate. Yet, by using Golgi's silver stain the great Spanish anatomist Salvadore Ramón y Cajal (1852–1936) was able to demonstrate that each nerve cell with its dendrites and axon is an independent unit. This extension of cell theory to the nervous system is known as the 'neuron doctrine'. Cajal further demonstrated that neurons come in a great variety of specific shapes that are characteristic of their location and often constant from animal to animal and species to species. When they shared the Nobel Prize in 1906, Golgi and Cajal's addresses were violent attacks on one another, and the reticular concept did not entirely disappear for a few more decades. Indeed, the definitive disproof of the reticular doctrine and confirmation of the neuron doctrine had to await the electron microscope.

See also: Cajal's Place in the History of Neuroscience; Cognitive Neuroscience: An Overview; Neuron Doctrine: Historical Background; Neurophysiology: Past and Present; Phrenology.

Further Reading

Fearing F (1970) *Reflex Action: A Study in the History of Physiological Psychology*. Cambridge: MIT Press.

Finger S (1994) *Origins of Neuroscience: A History of Explorations into Brain Function*. Oxford: Oxford University Press. New York.

Gross CG (1998) Brain, Vision, Memory: Tales in the History of Neuroscience. Cambridge: MIT Press.

Liddell ET (1970) *The Discovery of Reflexes*. Oxford: Clarendon.

McHenry LC (1969) *Garrison's History of Neurology*. Thomas Springfield.

Singer C (1957) *A Short History of Anatomy from the Greeks to Harvey*. New York: Dover.

Hodgkin–Huxley Models

F Santamaria, The University of Texas at San Antonio, San Antonio, TX, USA
J M Bower, The University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

© 2009 Published by Elsevier Ltd.

> Fitting theoretical equations to biological processes is often not a very profitable business, but in this case Huxley and I had a particular reason for carrying out such an analysis. (AL Hodgkin, *The Sherrington Lectures VII*, 1964)

Introduction

In 1963, Alan Lloyd Hodgkin and Andrew Fielding Huxley shared the Nobel Prize with Sir John Carew Eccles for their efforts to elucidate the electrical process by which neurons process information and then send signals down their axons as action potentials. Hodgkin and Huxley's contribution specifically involved the first clear elucidation of the ionic mechanisms underlying the action potential (Figure 1). The methods they employed involved a combination of precise and carefully controlled biophysical experiments and a careful mathematical and statistical analysis of the data. The results of these biophysical experiments were then captured in a series of nonlinear differential equations, now commonly referred to as the Hodgkin–Huxley model for action potential generation. Although first developed in the context of the movement of sodium (Na) and potassium (K) ions through the cell membrane of the squid giant axon, the basic structure of these equations has been used to represent a broad range of ion membrane conductances found not only within cells of the nervous system but also in many other electrically active cell types (e.g., the heart and kidney). Although alternatives to the Hodgkin–Huxley model have been proposed for specialized modeling purposes, Hodgkin and Huxley's basic equations have provided the foundation for literally tens of thousands of computational studies, and their achievement is arguably the foundation for computational approaches to biological systems.

Historical Context

Hodgkin and Huxley's quantitative description of membrane currents and their application to conduction and excitation in nerves was published in the *Journal of Physiology* in 1952 as the fifth in a series of papers, the first four of which were experimental. The biophysical studies that formed the foundation for the theoretical model were conducted in the giant axon of the squid, whose unusually large size allowed electrodes to be placed inside its cell membrane. The potential value of the squid giant axon for these types of studies was first suggested by JZ Young in 1936, and Hodgkin and Huxley first demonstrated the ability to record electrical potentials from inside the axon in a short report in the journal *Nature* published in 1939. Interrupted by World War II, Hodgkin and Huxley returned to their biophysical studies at the Marine Biological Station in Plymouth, UK, following up their original description of the squid action potential with a longer paper in the *Journal of Physiology* in 1945. These studies and those of their American peers, Kenneth Cole and Howard Curtis, put to rest the view that the action potential was due to the simple breakdown of the cell membrane, and they laid the groundwork for a series of papers suggesting that the electrical properties of nerves were active and not passive. Hodgkin and Huxley's five papers in 1952 were the capstone to this effort.

The first of the five papers provided a description of the electrical behavior of the membrane of the squid giant axon under control conditions. The second paper examined the effect of changing the concentration of Na on the action potential, providing evidence that the ionic currents' association with the action potential could be resolved into those related to Na flowing into the cell and K flowing out of the cell. The third paper reported on voltage clamp techniques (first applied by Cole and Curtis in 1947) used to demonstrate the effects of rapid changes in membrane potential on these ionic conductances. The fourth paper described the conduction inactivation process that they had discovered applied to the Na conductance. The final paper reassembled all these pieces together into a set of mathematical equations whose principal parameters were fit with experimental data on the voltage- and time-dependent permeability of the membrane to Na and K.

Hodgkin and Huxley's efforts represent not only arguably the first but also the gold standard for computational studies in biology. They demonstrated that a series of equations tuned using data obtained by artificially isolating and controlling the electrical response of the squid axon could, without further tuning, reproduce the naturally occurring action potential. Thus, the action potential was an emergent property of equations tuned to replicate individual ionic conductance data. In their words, these equations predicted

> with fair accuracy many of the electrical properties of the squid giant axon: the form, duration, and amplitude of the spike, both 'membrane' and propagated; the

conduction velocity; the impedance changes during the spike; the refractory period; ionic exchanges, subthreshold responses; and oscillations.

The Hodgkin–Huxley Equations

This section describes the set of Hodgkin–Huxley equations in the context of what was known about the biology of the squid axon and neuronal processes at the time. Although the key step in parameterizing the equations involved curve fitting experimental data, the form of the equations was strongly influenced by the researchers own intuition and speculations regarding the underlying physical mechanism.

The Membrane

Biologists at the time knew that cell membranes had passive electrical properties that could be characterized using basic equations derived from the physics of electrodynamics and electrical circuits. Thus, in the Hodgkin–Huxley model, the membrane of the squid axon is represented by an electrically equivalent circuit with values for membrane resistance, R_m, and capacitance, C_m. In addition, they knew that cell membranes maintain a difference in potential between the inside and the outside of the axon. Accordingly, the electrical equivalent for the membrane includes a battery (Figure 2(a)). By their own convention, the resistance is in units of $[1/G_m] = [R_m] = \text{ohm} \times \text{cm}^2$, $[C_m] = \text{F cm}^{-2}$, $[V] = \text{mV}$, and $[I_m] = \text{A}$. It was also known, of course, that the membrane potential of the axon changed with time, and that this change in time involved both the movement (or conductance) of ions through the membrane and a capacitive component due to the membrane. Thus, the overall flow of current was represented by the following form:

$$I_m = G_m(V - V_m) + C_m(\mathrm{d}V/\mathrm{d}t) \quad [1]$$

Note that according to this equation, holding V constant (thus, $\mathrm{d}V/\mathrm{d}t = 0$) allows a direct measure of I_m, which was the critical insight for the experimental use of the voltage clamp technique.

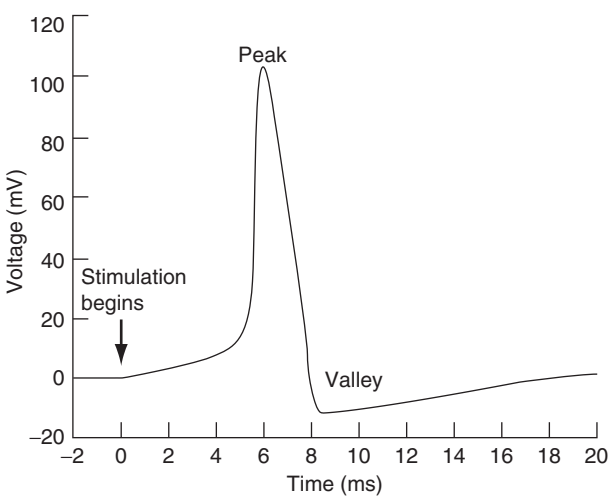

Figure 1 The action potential is the electrical discharge of a neuron. Action potentials are usually generated in the area connecting the soma and the axon (axon hillock). The plot shows the calculated voltage values of an implementation of the Hodgkin–Huxley model in which the axon was stimulated with a constant step current of 0.025 nA starting at $t = 0$.

Membrane Currents

In the Hodgkin–Huxley formulation, ionic currents are modeled as conductances (g_i) in parallel to a leak conductance through the membrane as shown in Figure 2(b) [eqn 2.1]:

$$I_m = g_m(V - V_m) + C_m \frac{\mathrm{d}V}{\mathrm{d}t} + \sum_i g_i(V - V_i) \quad [2.1]$$

$$g_i = \bar{g}_i f_i(V, t) \quad [2.2]$$

Equation [2.2] shows the net conductance given the constant $\bar{g}_i$, which represents the maximum conductance of the channels, and V_i, which is the

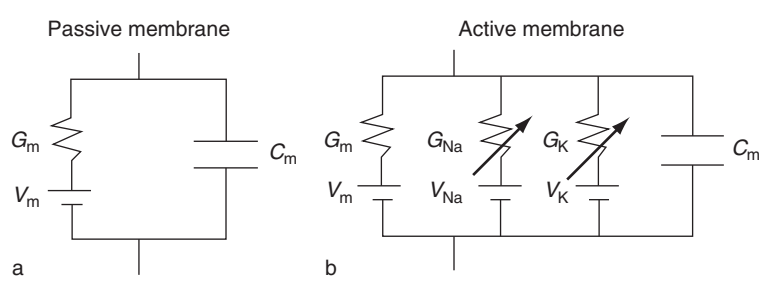

Figure 2 The membrane of a neuron has electrical properties. (a) The passive membrane properties include the constant negative voltage that the membrane keeps between the inside and the outside of the cell, modeled by a battery (V_m). The membrane also acts as a capacitor by separating different concentrations of charges on both sides (C_m), and the membrane allows the flow of current, characterized by a membrane resistance (R_m) or conductance ($G_m = 1/R_m$). (b) The active properties of a membrane are modeled similarly as the passive properties using conductances that depend on time and membrane voltage for each individual modeled channel. Note that each active current has a different reversal potential (V_{Na} and V_K).

reversal potential for the respective currents. Using the voltage clamp technique, Hodgkin and Huxley determined how g_i varied as a function of voltage and time, and fitted a biophysical model to $f_i(V, t)$. This function is also a function of temperature, but here we focus on its temporal and voltage dependence.

Adding Ion Currents

It was known at the time that the action potential was certainly associated with the flow of Na into the cell and K out of the cell. What was not well known was the specific biophysical mechanism of these conductances or their detailed relationship to time, voltage, or each other.

Potassium current Using voltage clamp techniques and manipulating ion concentrations, Hodgkin and Huxley showed that the K conductance generated during an action potential was turned on by depolarization and remained on as long as the stimulation remained constant. They modeled this K conductance as a first-order differential equation. The actual mathematical model they chose was based on kinetic reaction equations that had been intensely studied during the early nineteenth century by Leonor Michaelis, Maud Menten, and others. Using this form, if n is the kinetic variable, the reaction is described by eqn [3.1], with n having values between 0 and 1, and with α_n and β_n being rate functions dependent on voltage. To properly curve fit the experimental results, the activation term was raised to the fourth power, although Hodgkin and Huxley also provided a possible physical interpretation for this value (see below):

$$\frac{dn}{dt} = \alpha_n(1-n) - \beta_n n \qquad [3.1]$$

$$f_k(V) = n^4 \qquad [3.2]$$

Using the voltage clamp technique to fix the voltage of the axon at a specific value, two parameters were measured: the final value of the conductance and the time constant required to reach that value, with the rate functions then expressed as in eqns [4.1] and [4.4] (**Figures 3(a) and 3(b)**):

$$\text{Steady state: } 0 = \alpha_n(1-n_\infty) - \beta_n n_\infty \qquad [4.1]$$

$$n_\infty = \frac{\alpha_n}{\alpha_n + \beta_n} \qquad [4.2]$$

Temporal response, α_n and β_n constant

$$n_\tau = e^{-t(\alpha_n+\beta_n)} + \frac{\alpha_n}{\alpha_n + \beta_n} \qquad [4.3]$$

With time constant

$$\tau_n = \frac{1}{\alpha_n + \beta_n} \qquad [4.4]$$

The total current passing through the membrane associated with K was therefore described by

$$I_m = g_m(V - V_m) + c_m \frac{dV}{dt} + \bar{g}_k n^4 (V - V_k) \qquad [5]$$

Sodium current Hodgkin and Huxley discovered that the other conductance associated with the action potential – that due to Na – was somewhat more complex. Unlike the K conductance, the Na conductance does not maintain a continuous level with maintained depolarization but instead appears to decrease on its own after a period of time. Hodgkin and Huxley therefore postulated that in addition to a K-like activation, the Na conductance also had an intrinsic inactivation phase related to time. To describe the Na conductance, Hodgkin and Huxley therefore had to introduce two voltage-dependent variables, one for activation (m) and one for inactivation (h), as shown in eqn [6.1]:

$$f_i(V, t) = m^3 h \qquad [6.1]$$

$$\frac{dm}{dt} = \alpha_m(1-m) - \beta_m m \qquad [6.2]$$

$$\frac{dh}{dt} = \alpha_h(1-h) - \beta_h h \qquad [6.3]$$

As in the equations for K (eqns [4.2] and [4.4]), the kinetics of these variables were governed by the rate variables m_∞ and h_∞, with a reverse dependence with respect to depolarization voltage for activation (m) kinetics and inactivation (h) kinetics (**Figures 3(c)–3(f)**). Accordingly, the values of m_∞ increase as voltage rises, whereas the values of h_∞ decrease (**Figures 3(d) and 3(f)**). The multiplication of both variables ensures that the sodium current is only active in a small window of voltage and time after depolarization, and the inactivation term means that the conductance decreases even in the presence of a sustained depolarization (generated, for example, by the voltage clamp). Combining the Na and K conductance terms, the final form of the Hodgkin–Huxley model is

$$I_m = g_m(V - V_m) + C_m \frac{dV}{dt} + \bar{g}_k n^4 (V - V_k) + \bar{g}_{Na} m^3 h (V - V_{Na}) \qquad [7]$$

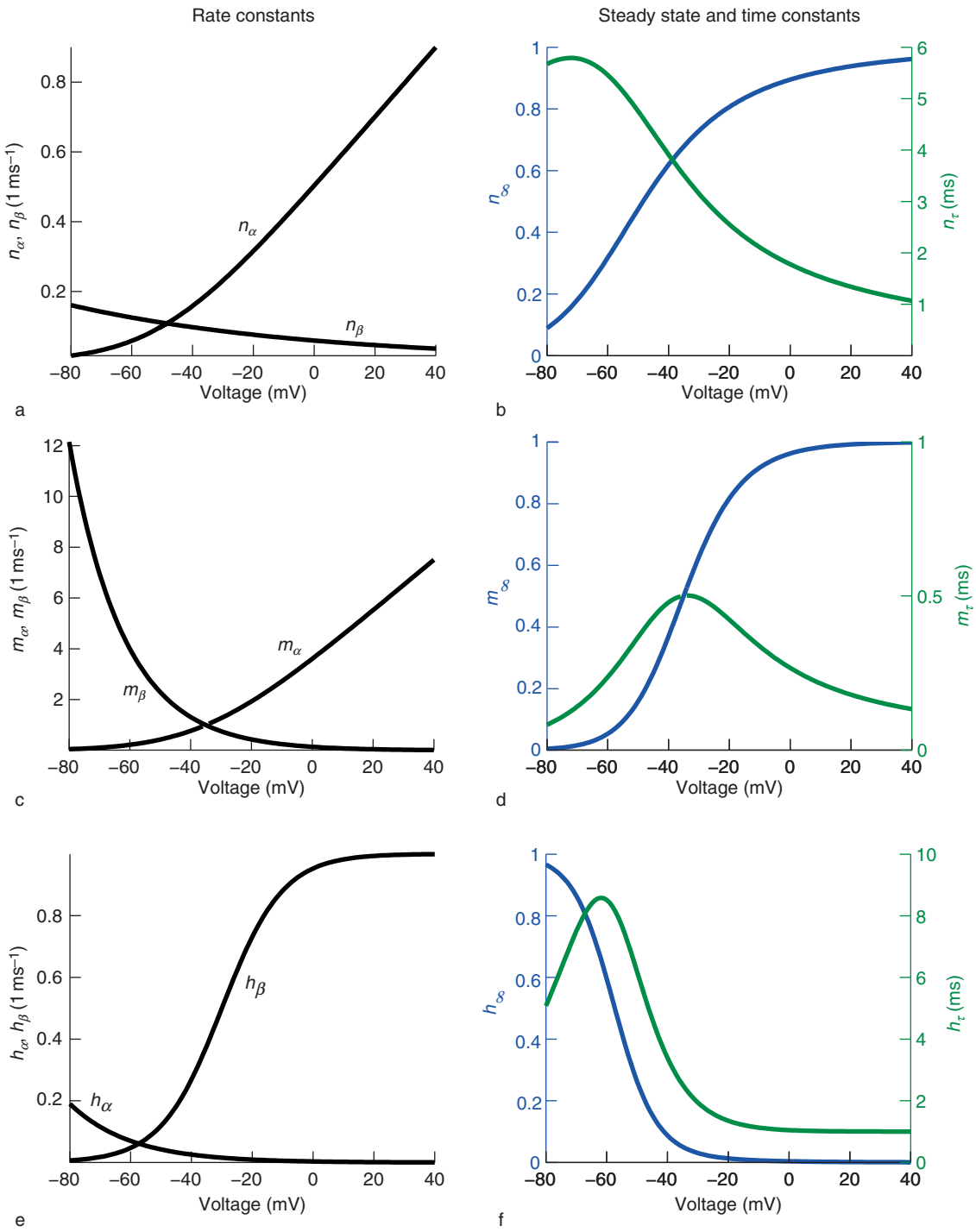

Figure 3 Rate functions used to reproduce the action potential in the Hodgkin–Huxley model. (a) Forward (α_n) and backward (β_n) rate constants used to model the potassium conductance: $dn/dt = \alpha_n(1-n) - \beta_n n$. (b) An alternative way to describe the kinetics of the model that allows measuring their values experimentally is to use the steady-state value of the rate variable, $n_\infty = \alpha_n/(\alpha_n + \beta_n)$, and time constant, $\tau_n = 1/(\alpha_n + \beta_n)$. (c–f) The sodium current requires two particles, one of which models the fast activation (m) and the second one models the inactivation (h). Note that the inactivation particle behaves inversely as a function of voltage compared to the activation particle.

Emergent Properties

Although the ability of this equation to replicate an action potential "agreed satisfactorily with the time course of the action potential recorded experimentally" was the evidence cited as support for the model in the original 1952 paper, another measure of the success of their efforts was in the other properties of axons and neurons which this equation also illuminated (and replicated). These are discussed next.

Firing threshold and firing frequencies One way to determine the minimum current required to generate an action potential in the model and the experiments is to inject a step current of different amplitudes. Figure 4 shows that for low levels of current amplitude, the model does not generate action potentials (black trace). When the amplitude of the step current is approximately 0.02 nA, the model produces only one action potential. A single action potential is evoked by stimulation between 0.02 and 0.045 nA. After 0.045 nA, the model generates an uninterrupted train of action potentials. After that, the increase in firing frequency with increased current injection follows a logarithmic function. Thus, overall, a constant current of sufficient amplitude, injected into a neuron and the model, results in repetitive firing after reaching a threshold. However, the Hodgkin–Huxley model also accounts for the fact that the onset of firing is dependent on the rate of depolarization of the membrane. As shown in Figure 4, two types of current stimulus induce different current threshold for firing. The red trace shows a similar simulation in which the step current was preceded by a 100 ms ramp. Note that not only has the threshold for generating a repetitive train of action potentials been shifted to higher currents (0.08 nA) but also there is no generation of a single spike for lower values of stimulation. The same is true for voltage. Figure 5 shows that an instantaneous depolarization of the membrane results in a voltage threshold between 6 and 7 mV. If, instead of a sudden instantaneous depolarization, the membrane voltage is clamped to a ramp, then the threshold to generate an action potential varies as a function of the duration of the ramp. If the duration of the ramp is longer than 10 ms, then no action potential is evoked. The model suggests that this is due to the increasing inactivation of the sodium conductance under this type of stimulus. Thus, an action potential is generated at a

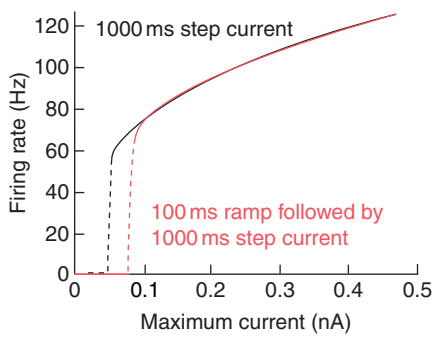

Figure 4 Action potential frequency and threshold as a function of injected current. The model was stimulated with different constant values of current starting at $t = 0$ for 1000 ms (black). We then recorded the number of action potentials generated during that period and calculated the firing frequency. The red curve shows the average firing rate calculated over the same window of time as the black curve but preceded by a ramp current over 100 ms.

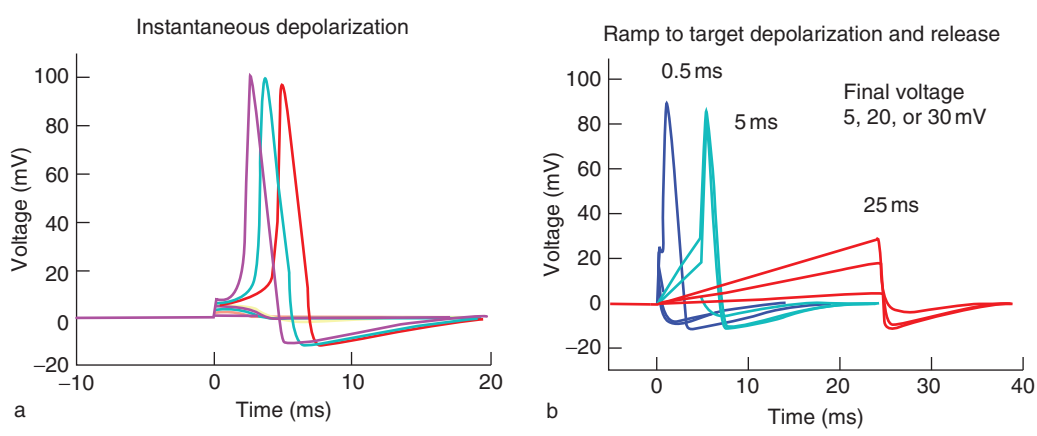

Figure 5 The Hodgkin–Huxley model does not have a fixed voltage threshold. (a) Response of the model to increasing instantaneous depolarizations. At $t = 0$, the voltage is fixed at different depolarizing values for a single time step (0–8 mV). The threshold is between 6 and 7 mV. (b) The voltage threshold depends on the stimulus. We simulated several voltage clamp experiments in which the axon was depolarized following a ramp function starting at $t = 0$ for three different windows of time. After the voltage clamp ramp reached the target final voltage, the simulation was allowed to continue freely. The final voltage was 5, 20, or 30 mV. Depending on the length of the ramp stimulation, the axon never generated an action potential.

different level of depolarization for a step voltage pulse than for a voltage current. In this sense, Hodgkin and Huxley demonstrated that neurons do not, in fact, have a fixed threshold for firing, but both the timing and the rate of firing depend on the type of depolarization.

Refractory period After an action potential is generated, there is a window of time in which an identical stimulus fails to generate a second action potential. The Hodgkin–Huxley model suggests that this refractory period is due to the combination of the inactivation of Na channels and the activation of K conductances (Figure 6).

Propagation velocity Hodgkin and Huxley related eqn [7] to the speed of action potential propagation by noting that the membrane current (I_m) is proportional to the speed of voltage change:

$$I_m = \frac{a}{2R_a v^2}\frac{\partial^2 V}{\partial t^2} \qquad [8]$$

where v is the propagation velocity, a the fiber diameter, and R_a is the cytoplasmic axial resistance. Hodgkin and Huxley found that only the correct value of v makes the shape of the action potential stable. Using an iterative method, they calculated the value of v that was within 10% of the experimentally found value.

Anode break A feature of many neurons is that an action potential is generated after being hyperpolarized. In the Hodgkin–Huxley model, this is due to the sudden change in voltage and the fast time constant of the sodium channels. Combined, they are capable of allowing a sudden influx of current that results in an action potential (Figure 7).

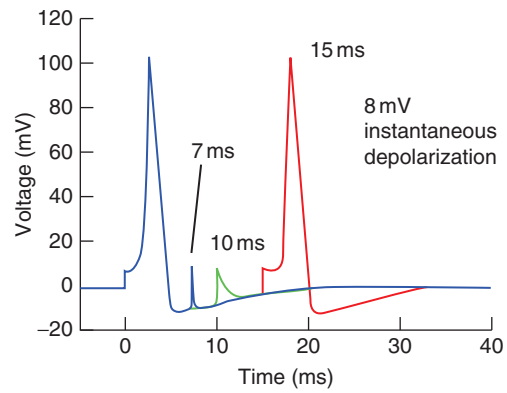

Figure 6 The action potential has a refractory period. The plot shows an identical simulation as shown in Figure 5(a), where the model was instantaneously stimulated at $t = 0$ to 8 mV, generating an action potential. At different times after the initial stimulus, a second identical stimulation was delivered ($t = 7$, 10, and 15 ms). The plots show that it is not until the second stimulus is 15 ms apart that a full-blown action potential is evoked.

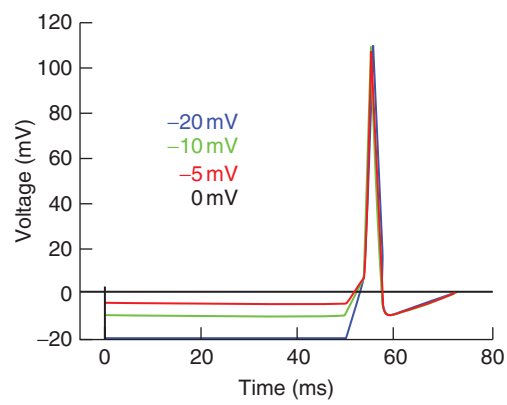

Figure 7 Anode break response or excitation after release from inhibition. The model generates an action potential after being voltage clamped at relative negative potentials (−20, −10, and −5 mV) for 50 ms and then allowed to continue freely without any further external stimulation.

Physical Interpretations of the Hodgkin–Huxley Equations

Rigorously, the Hodgkin–Huxley equations are not a model of action potential generation since their derivation does not start from an actual biophysical model of the components (now known to be actual channels in the membrane) that are responsible for current flow. This was recognized by the authors in their original paper when they noted that "an equally satisfactory description of the voltage clamp data could no doubt have been achieved with equations of a different form, which would probably have been equally successful in predicting the electrical behavior of the membrane."

There is no question, however, that the choice of the form of the equations Hodgkin and Huxley used reflected their own guess as to the mechanisms involved. In particular, Hodgkin and Huxley knew from their previous experimental work that the movement of ions through the membrane was an active process and not, as had previously been thought, simply a general breakdown of the membrane. Therefore, it was natural for them to suggest that some active agent, in this case transporter particles, might be responsible for actively facilitating the movement of ions through the membrane. Hodgkin and Huxley interpreted the exponents in the rate equations (n^3, m^4, and h) to reflect the number of particles necessary to move a single ion.

Today, we know that Na and K conductances (and almost all ion conductances) are dependent on actual

physical channels that can have multiple open and closed states. The actual flow through these channels is not facilitated; once the channel opens, the ions flow along their concentration and potential gradients. However, the fact that Hodgkin and Huxley constructed their equations based on first-order reaction kinetics allowed multiple mechanistic interpretations to be applied, including the probability of open and closing of a 'gate' in the channel. The fact that the form of the equations was so well matched to the form of the experimental data also meant that Hodgkin and Huxley captured the dynamics, independent of the specific mechanism.

Modern Interpretation of Hodgkin–Huxley Equations

It is another tribute to Hodgkin and Huxley's work, and the close interaction between their experimental and theoretical efforts, that the form of their equations can be reinterpreted in the context of modern channel data, even though the authors did not know about channels. Specifically, we now know that current flows along the electrochemical gradient of a pore created by a membrane channel. The opening and closing of a channel is a stochastic event that depends on membrane voltage and time. Each channel is composed of internal states. For example, the exponent of the activation particle of the K conductance (n^4) is now interpreted to mean that each ion channel has four independent internal states, and it is only when the four states are in the active position that ions actually flow through the membrane. Similarly, for the Na conductance, a single channel requires three internal states to be in the open position (m^3) and the inactivation variable to be closed (h) for Na to flow. The rate functions α_i and β_i determine the transition between open and closed states.

Limitations in the Hodgkin–Huxley Formulation

Although the Hodgkin–Huxley model has been tremendously successful, there are limitations. The fact that Hodgkin and Huxley found it expedient to work in the squid giant axon, for example, has had consequences for our understanding of the biophysics and evolution of ion channels. Specifically, the squid giant axon is a highly specialized biological device whose sole purpose is to rapidly and reliably activate the muscles of the squid's mantle to generate its jet propulsion mechanism. As a result, the ion conductances of the axon have evolved for speed and consequently are less complex than those found in many other neurons. We now know that most other neurons are much more complex, especially with regard to the number and diversity of potassium conductances. Nevertheless, these more complicated conductances are still generally represented computationally using a Hodgkin–Huxley formulation.

With an interest in reducing the computational complexity of their models, others have argued that the Hodgkin–Huxley formulation is more complex than necessary to replicate the basic phenomenon it describes. In their original 1952 paper, Hodgkin and Huxley made the point that there might be other equations which capture the dynamics of the action potential. In 1961 and 1962, Fitzhugh and Nagumo introduced one such formulation, which is now used extensively in simplified neural network-type models to represent the spiking behavior of neurons. This simplified model is computationally more compact than the Hodgkin–Huxley model, although it loses several of the subtleties of the Hodgkin–Huxley form, including the precise reproduction of the shape of the action potential; however, it preserves the timing of firing when stimulated by a constant current.

With respect to real biological mechanisms, there is good evidence that the Hodgkin–Huxley formulation may not be complex enough. It has been known for some time, for example, that the Hodgkin–Huxley equations cannot in their traditional form account for the variability of spike initiation or the speed of the rising phase of the action potential, especially in mammalian cortical neurons. One explanation for these two properties relies on reconsidering the fundamental assumption of the Hodgkin–Huxley equations that individual ionic conductances act independently of each other. There is now evidence, although controversial, for cooperativity among sodium channels which appears to account for the variability and speed of the action potential but requires a new form of equations.

Thus, since its original introduction, the Hodgkin–Huxley model has been changed by some to be less complex and by others to be more complex. Yet, for more than 50 years these equations have provided the basis for the majority of modeling efforts in neuroscience. The close and quantitative connection Hodgkin and Huxley made between careful experimental studies and intuitively driven mathematical structures continues to provide the foundation for computational neuroscience and our evolving understanding of the function of the nervous system.

See also: Action Potential Initiation and Conduction in Axons; Computational Methods; Ion Channel Localization

in Axons; Neural Oscillators and Dymanmical Systems Models; Spiking Neuron Models.

Further Reading

Cronin J (2003) *The Mathematical Aspects of Hodgkin and Huxley Neural Theory*. Cambridge, UK: Cambridge University Press.

De Schutter E (ed.) (2001) *Computational Neuroscience: Realistic Modeling for Experimentalist*. Boca Raton, FL: CRC Press.

Hille B (2001) *Ion Channels of Excitable Membranes*, 3rd edn. Sunderland, MA: Sinauer.

Hodgkin AL and Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* 117: 500–544.

Johnston D and Wu SM (1994) *Foundations of Cellular Neurophysiology*. Cambridge, MA: MIT Press.

Koch C (1999) *Biophysics of Computation: Information Processing in Single Neurons*. New York: Oxford University Press.

McCormick D, Yousheng S, and Yuguo Y (2007) Neurophysiology: Hodgkin and Huxley model – still standing? *Nature* 445(7123): E1–E3.

Naundorf B, Wolf F, and Volgushev M (2006) Unique features of action potential initiation in cortical neurons. *Nature* 440(7087): 1060–1063.

Shepherd GM (2003) *The Synaptic Organization of the Brain*, 5th edn. New York: Oxford University Press.

Stuart G, Spruston N, and Häusser M (eds.) (1999) *Dendrites*. New York: Oxford University Press.

Relevant Websites

http://www.genesis-sim.org – Genesis.

http://neuron.duke.edu – Neuron.

http://www.mcell.cnl.salk.edu – Salk Institute for Biological Studies, MCell: A Monte Carlo Simulator of Cellular Microphysiology.

Holoprosencephaly

J A Golden, Children's Hospital of Philadelphia and the University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Holoprosencephaly is a dramatic central nervous system (CNS) malformation highlighted by a partial or complete failure of the brain to separate into right and left hemispheres, forming a continuum of cortex and other structures across the midline. Patients with this anomaly frequently exhibit specific craniofacial anomalies, including midline facial clefts, cyclopia, and nasal anomalies. Although relatively rare, holoprosencephaly is the most common anomaly involving the brain and the face, with an incidence of approximately 0.48–0.88 per 10 000 live-born children with normal chromosomes and an estimated 40 per 10 000 cases among human abortuses, indicating a high rate of fetal loss. The identification of mutations in multiple genes coupled with advancements in understanding neural tube induction and patterning have provided a somewhat unique opportunity to begin understanding the neurodevelopmental basis of a human malformation. This article reviews the pathology, genetics, and current understanding of neural development as it relates to holoprosencephaly.

Classification and Associated Anomalies

The cardinal feature of holoprosencephaly is a single telencephalic ventricle with continuity of the cerebral hemisphere across the midline (Figure 1). Holoprosencephaly is conventionally divided into three subtypes, lobar, semilobar, and alobar, based on the extent of the malformation. In reality these three subtypes represent a continuous spectrum rather than clearly distinguishable forms of the same disorder. In addition to the CNS anomalies, specific craniofacial abnormalities are associated with holoprosencephaly and together constitute the holoprosencephaly sequence or syndrome. Similar to the brain anomalies, the craniofacial anomalies vary in extent, from the most severe, cyclopia with a single proboscis (nasal anlagen) located above the midline eye, to the mildest facial abnormality, consisting of a single midline maxillary incisor. While some correlation exists with the extent of the brain anomaly to the severity of the craniofacial anomalies, there is not a perfect association.

Environmental Causes of Holoprosencephaly

An outbreak of cyclopia in lambs in the early 1960s led to the identification of cyclopamine, a steroidal alkaloid, as the causative agent. Ewes grazing on *Veratrum californicum*, a plant species containing large amounts of cyclopamine, were found to give birth to malformed offspring. Further studies determined that cyclopamine and a related compound, jervine, can cause a similar spectrum of anomalies in chickens, rats, mice, hamsters, and rabbits. The pathogenetic mechanism through which these alkaloids exert their effect was linked to molecular signaling along the sonic hedgehog (SHH) pathway, a signaling pathway involved in many developmental processes, including neural induction and the pathogenesis of holoprosencephaly.

Genetics

Approximately 25% of patients with holoprosencephaly have a recognizable syndrome. As will be discussed, the Smith–Lemli–Opitz syndrome, which is due to a defect in cholesterol biosynthesis at the level of 3β-hydroxy-steroid-Δ^7-reductase, is of particular interest because it also has a link to the SHH signaling pathway. A number of nonrandom chromosomal abnormalities have also been found in patients with holoprosencephaly. Chromosomal anomalies have been reported in 24–45% of live-born infants, the most common of which is trisomy 13. Based on nonrandom cytogenetic rearrangements, at least 12 chromosomal regions on 11 chromosomes may contain genes that play a role in the pathogenesis of holoprosencephaly.

Positional cloning studies, cloning of translocation sites, and candidate gene screenings have resulted in the identification of mutations in six genes associated with holoprosencephaly. The first identified was *SHH*; subsequently, mutations in *SIX3*, *ZIC2*, *TGIF*, *PTCH1*, *GLI2*, and *TDGF1* have been identified (Table 1). All but one of these genes can be placed in a complex signaling pathway, suggesting a common pathogenesis for holoprosencephaly (Figure 2).

Confirmation that *SHH* plays a role in the pathogenesis of holoprosencephaly has come through the generation and analysis of mice that are homozygous mutant for *Shh*. These mice show cyclopia and CNS features that resemble holoprosencephaly. Why only homozygous null mutant mice have features of the holoprosencephaly sequence, whereas humans heterozygous

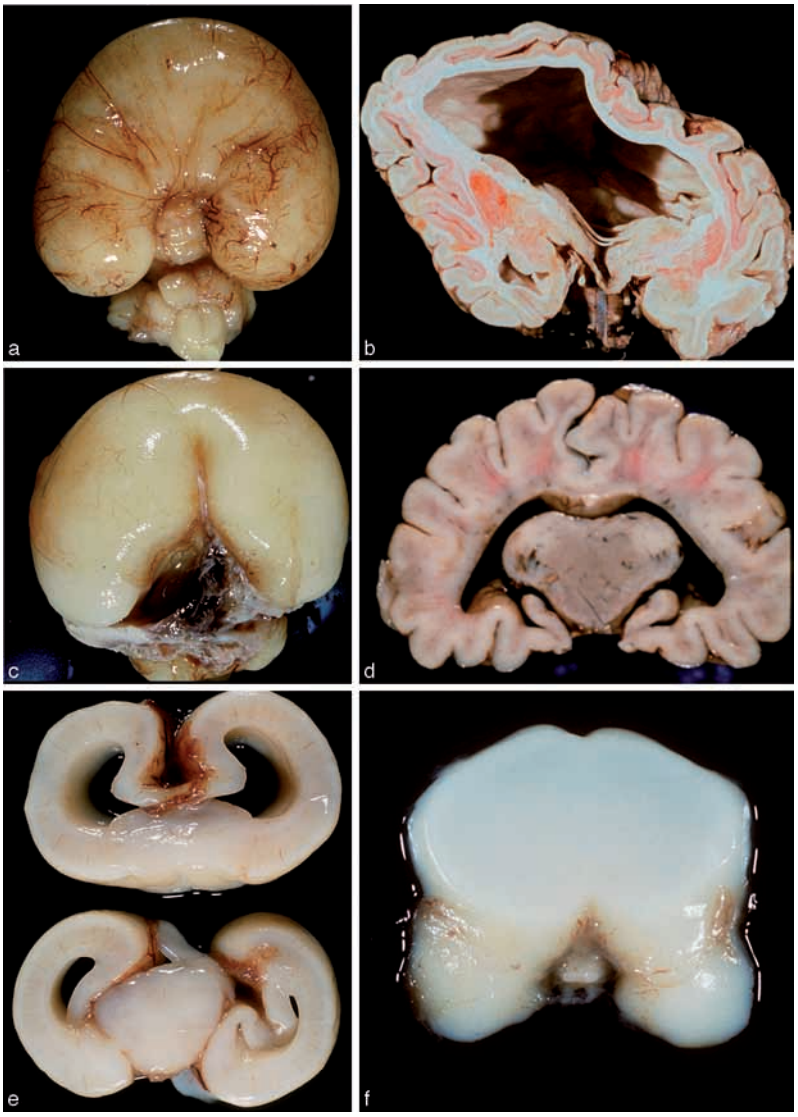

Figure 1 Alobar, semilobar, and lobar holoprosencephaly. (a) Frontal view of the brain from a 20-week fetus with alobar holoprosencephaly; note there is no fissure separating the two hemispheres. (b) Brain from an adult with alobar holoprosencephaly, showing continuity of the gray matter across the dorsal midline. This patient had minimal ventral abnormalities, which is atypical for most cases of alobar holoprosencephaly. (c) Semilobar holoprosencephaly with a small cleft partially separating the dorsal–caudal cerebral hemispheres (caudal view, 19-week fetus). (d) Brain from a full-term child with semilobar holoprosencephaly. The diencephalic structures show an abnormal structure spanning the midline and there appears to be a sagittal fissure; however, gray matter is seen crossing the midline and no corpus callosum is present. (e) Brain of a 20-week fetus with lobar holoprosencephaly. The cerebral hemispheres appear separated in both anterior (top) and posterior (bottom) sections, although the diencephalon shows a massively enlarged massa intermedia or an abnormal developed midline structure (bottom section). (f) Coronal section from the same brain as in (e), showing a continuum across the midline in the frontal cortex only.

Table 1 Holoprosencephaly genes

Gene	*Abbreviation*	*Chromosomal location*
Sonic hedgehog	*SHH*	7q36
Patched-1	*PTCH1*	9q22.3
Sine oculis homeobox homolog 3	*SIX3*	2p21
ZIC family member 2	*ZIC2*	13q32
5′-TG-3′-interacting factor	*TGIF*	18p11.3
GLI-Kruppel family member GLI2	*GLI2*	2q14
Teratocarcinoma-derived growth factor 1	*TDGF1*	3p21.31

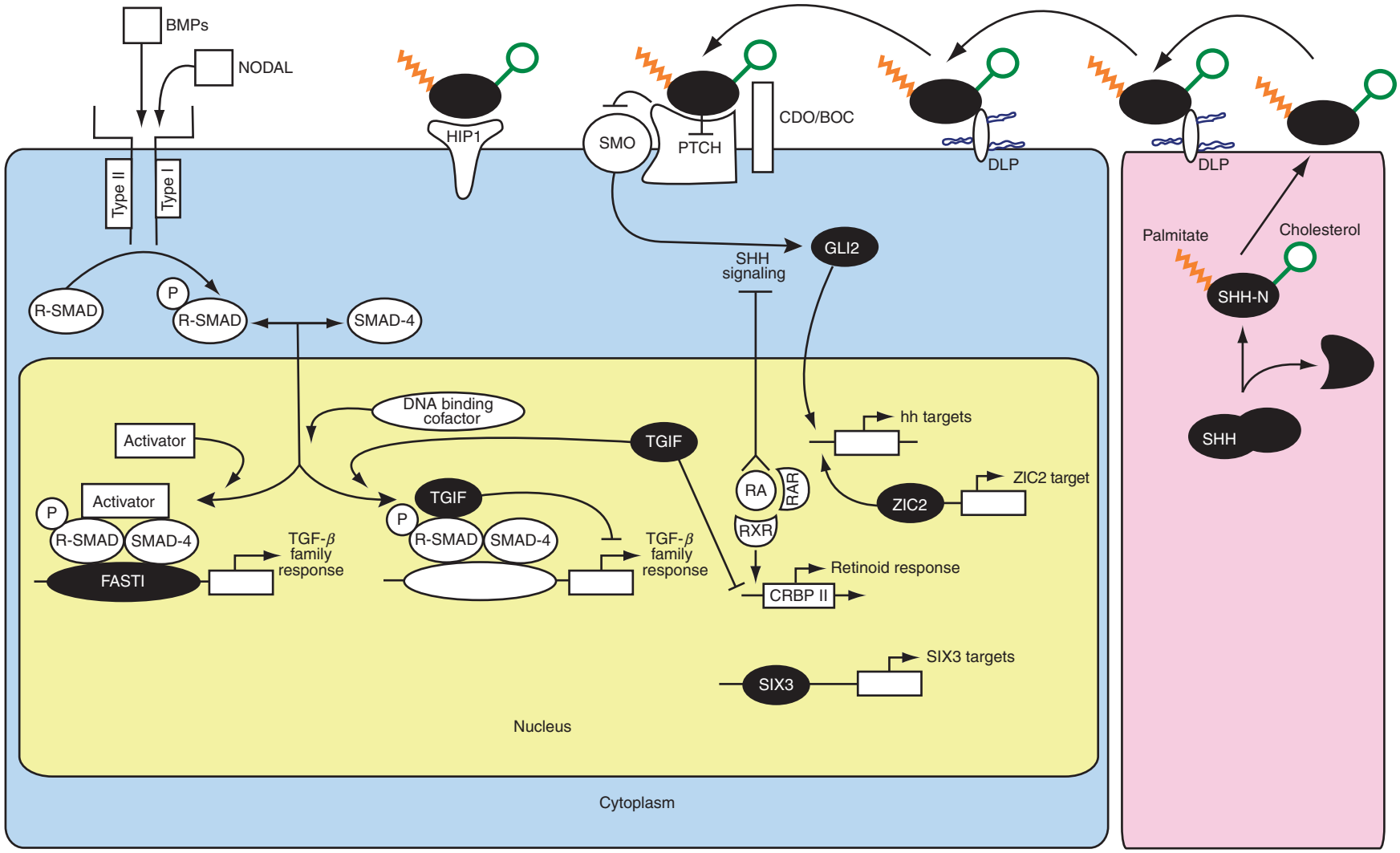

Figure 2 Cellular signaling pathways implicated in holoprosencephaly. Schematic representation of two cells (pink and blue), with the nucleus (yellow) highlighted in one cell. All factors shaded in black have known mutations correlating to holoprosencephaly (see text). SHH, sonic hedgehog; SHH-N, N-terminal of SHH; DLP, Dally-like protein; PTCH, Patched; SMO, Smoothened; CDO/BOC, cell adhesion molecule-related/downregulated by oncogenes/brother of CDO; GLI2, GLI-Kruppel family member 2; ZIC2, ZIC family member 2; RA, retinoic acid; RAR, retinoic acid receptor; RXR, retinoid X receptor; CRBPII, cellular retinol-binding protein 2; SIX3, sine oculis homeobox homolog 3; HIP1, hedgehog-interacting protein 1; BMP, bone morphogenetic protein; Type I, BMP receptor type I; Type II, BMP receptor type II; SMAD, mothers against decapentaplegia; R-SMAD, receptor SMAD; P, phosphorylation; FAST1, forkhead activin signal transducer 1 (forkhead box H1); TGIF, 5′-TG-3′-interacting factor.

for mutations in *SHH* show holoprosencephaly, remains unknown. Presumably, point mutations in the human *SHH* gene result in haploinsufficiency. Further studies will be required to sort out this discrepancy. The analysis of other mouse mutants also supports the role in holoprosencephaly, although this was not true for at least one mouse model (*Tgif*), confusing its role in the pathogenesis. Interestingly, extreme heterogeneity exists in the clinical expression of the holoprosencephaly sequence within families with known mutations. In some of these families, rare polymorphisms or mutations in a second holoprosencephaly gene have been identified, supporting a complex genetic origin in some cases.

Neuropathology

Hans Kundrat first designated this class of malformation as arrhinencephalies, considering the absence of the olfactory bulbs and tracts as the cardinal feature. Paul Yakovlev later pointed out that these malformations were not of the rhinencephalon (end of the forebrain) alone, but instead were a defect of the entire telencephalon. Recognizing the single telencephalic ventricle, which Yakovlev called a holosphere, he named this class of malformation 'holotelencephaly.' The subsequent recognition that the thalamus and other diencephalic-derived structures were also affected in these patients led William DeMeyer to coin the term 'holoprosencephaly' to account for a defect in the entire prosencephalon. Holoprosencephaly has since been the preferred terminology.

In all forms, the brain is small, averaging less than 100 g at full term for the alobar and semilobar forms of holoprosencephaly. The classification of alobar, semilobar, and lobar forms of holoprosencephaly is primarily based on the presence or absence of a midline fissure separating the two cerebral hemispheres and the extent of the fissure. In reality, these three forms are really a spectrum of the same entity, and this classification is useful for anatomic and pathologic purposes but does not appear to have clinical relevance.

In alobar holoprosencephaly, no midline fissure separates the two cerebral hemispheres that morphologically exhibit one holosphere (**Figure 1**). The Sylvian fissures are not formed and there is often no distinction between the lobes of the cerebral hemisphere. In infants and older individuals, the gyri are usually well developed; however, the course of the gyri does not resemble the normal surface anatomy. Instead, there are frequently sets of medial and paramedial gyri. The most medial gyri frequently contain the hippocampus and entorhinal cortex and course along the lateral aspect of a cystlike membrane located in the dorsal–posterior aspect of the brain. The posterior aspect of the cerebrum is shaped like a horseshoe with the posterior–dorsal rim is composed of the thin, cystlike membrane. This membrane is actually the posterior roof of the single telencephalic ventricle. The basal ganglia and thalamus are immediately under the cystlike membrane and vary from distinct left and right structures to disorganized masses that span the midline. The corpus callosum and anterior commissure are usually absent; however, occasionally rudimentary callosal plates may be identified. Of interest, Probst's bundles have not been identified in the brains of patients with holoprosencephaly, which may reflect the absence of normal cortical projections (see later). The brain stem and cerebellum are most often grossly normal, although the corticospinal tracts are usually hypoplastic or absent.

Semilobar holoprosencephaly represents an intermediate degree of the same CNS malformation. The distinction between semilobar and alobar holoprosencephaly is the presence of a partial interhemispheric fissure separating the posterior extent of the cerebral hemisphere (**Figure 1**). In semilobar holoprosencephaly, the midline fissure does not extend to the rostral end of the brain and on cross sections; at the level of the midline fissure, cerebral cortex crosses the midline in continuity. Again the gyral pattern is disorganized.

The lobar form of holoprosencephaly is generally considered the least severe in the spectrum and is characterized by an interhemispheric fissure that divides virtually the entire brain, with the exception of the most rostral and ventral regions of the frontal lobes. Despite the apparent interhemispheric fissure, cerebral cortex is still found crossing the midline along virtually the entire anterior–posterior axis. Occasionally, a small part of the most posterior region of the brain will have a complete separation of the cerebral hemispheres and may even show a well-developed corpus callosum in the region of the separated cerebral hemispheres. The lobes of the cerebral hemispheres may become recognizable in the mildest cases, along with the semblance of a normal gyral pattern. Cross sections of the brain continue to show the single telencephalic ventricle (holosphere) with disorganized basal ganglia, commonly restricted to the most rostral region, forming a single structure across the midline. The thalami may be less involved, although an enlarged massa intermedia is quite common.

Some controversy exists regarding the cytoarchitecture of the cerebral cortex. All cytoarchitectural areas of the cerebral cortex have been identified in some cases, whereas others have reported the

organization of the cerebral cortex to be markedly disrupted. Large periventricular and/or white matter glioneuronal heterotopia are encountered in occasional cases. The defects in cortical organization that are more frequently observed may represent secondary injury to the cerebral cortex or an abnormality in connections into and out of the cerebral cortex. The pia overlying the cerebral hemisphere often shows extensive glioneuronal rests that form a 'crust' over the surface of the brain. This is not ordinarily seen at midgestation, but is common in late gestation and postnatally, suggesting a secondary change or late-developing anomaly.

The hippocampus is virtually always present, although most often developmentally incomplete or abnormal, particularly in its medial structures. The histopathology of the basal ganglia and the thalamus varies, frequently in direct proportion to the severity of the cerebral hemispheric malformation. In the most severe cases (alobar holoprosencephaly), the basal ganglia may be unidentifiable or recognized only as a midline mass with disorganized cytoarchitecture that has only a superficial resemblance to the caudate or putamen. The globus pallidus is often not recognizable. In less severe forms of holoprosencephaly, all components of the basal ganglia are present, although the head of the caudate may be a single structure spanning the midline and the septal region is not identifiable as such. The cerebellum may show various degrees of cortical dysplasia and/or heterotopia. This is particularly true for those cases associated with trisomy 13 and other cytogenetic abnormalities. Abnormalities of the brain stem are frequently limited to hypoplastic or absent corticobulbar and corticospinal tracts.

The olfactory bulbs and tracts are absent in virtually all of the cases. The optic nerves are often hypoplastic, while the remaining cranial nerves are usually normally formed. The course of middle and anterior cerebral arteries is usually abnormal. Most often, a disorganized cluster of vessels (rete mirable) extends from the region of the internal carotid arteries anteriorly over the ventral and frontal regions of the brain. A single anterior cerebral artery may be found coursing along the midline over the rostral aspect of the cerebral hemispheres. A variety of other less commonly associated anomalies of the brain (e.g., encephaloceles, myelomeningoceles, arachnoid cysts, porencephaly) may be seen in occasional cases of holoprosencephaly.

Pathogenesis

When the arrhinencephalies were first defined as a distinct entity, the embryologic origin was postulated to be a failure of the most rostral forebrain and the olfactory bulbs and tracts to develop. This remained the prevailing theory until 1959, when Paul Yakovlev first postulated that these anomalies resulted from a failure of the simple rostral neural tube to cleave into two cerebral hemispheres. In humans this was believed to occur shortly after neural tube closure, approximately 38–45 days postfertilization. The failure-of-cleavage hypothesis was later amended to include the possibility that the two cerebral hemispheres either represented a failure of cleavage of the simple neural tube, or that the two cerebral hemispheres failed to grow out from the neural tube. This suggested a failure in regional cell proliferation or potentially excessive cell death. Given that brains showing holoprosencephaly are always very small, averaging less than 100 g at full term, a failure to grow, or increased cell death, appeared to be a reasonable hypothesis. Over the past decade, the focus of the anomaly has shifted from the failure of cleavage or outgrowth of the rostral neural tube to a defect of the prechordal mesendoderm and the mediobasal forebrain. These data have led to the refined hypothesis that holoprosencephaly results from a primary defect of induction and patterning of the rostral neural tube, occurring in the first 4 weeks of gestation.

The classic experiments of Spemann and Mangold demonstrated that the neural tube, and, in fact, the entire body axes, could be induced by a small region of embryonic tissue, which has become known as the Spemann organizer. At the site of induction, an organized process of cell movement known as gastrulation occurs, resulting in the formation of a primitive streak and the migration of a population of cells to form the notochord. The notochord induces the overlying ectoderm, fated to become the neuroectoderm, to become the ventral neural tube. At the level of the spinal cord, the induction is mediated through the secreted protein SHH. First expressed by the notochord, SHH induces the midline of the overlying neural plate to become a specialized region of the neural tube known as the floor plate. The floor plate, once induced, also expresses SHH, leading to the induction of several transcription factors in the adjacent ventral columns. In turn, these give rise to the various classes of ventral interneurons and motor neurons, with each class dependent on the expression of specific transcription factors.

Development of the dorsal neural tube appears to be regulated by molecules produced initially in the adjacent nonneural ectoderm and later by the dorsal midline (roof plate). The molecular regulators of this process include bone morphogenetic proteins (BMP) and WNT proteins. Multiple members of the *BMP* and *WNT* gene families are expressed in the dorsal midline of the CNS. Loss-of-function studies

for several members of the *WNT* family and BMP signaling through the BMP receptors support a role for these proteins in dorsal neural tube development.

In contrast to the pathways that are being elucidated for dorsal–ventral patterning in the more caudal regions of the neural tube, study of the molecular basis for dorsal–ventral patterning of the rostral neural tube are more complex and less complete. Similar to the spinal cord, SHH participates in ventral patterning of the midbrain and the forebrain. SHH is expressed in two domains of the prosencephalon; initially it is restricted to the tuberal and mammillary hypothalamic primordia, and later to a small region in the rostrobasal telencephalon. Ectopic expression studies in the forebrain indicate that SHH is sufficient to induce gene markers of the tuberal and mammillary hypothalamus and basal telencephalon. Mice homozygous mutant for *Shh* lack ventral forebrain development and have cyclopia and holoprosencephaly. *Shh* has also been shown to be necessary and sufficient for forebrain development in zebra fish, complementing the data from mammals.

The genetic pathway(s) involved in dorsal induction and patterning of the prosencephalon are beginning to be understood, although they are currently less well defined than those for ventral induction. Similar to more caudal regions of the neural tube, BMPs also appear to participate in dorsal forebrain patterning. *In vitro* studies with forebrain tissue explants indicate that BMPs can induce genes that are expressed in the roof plate, such as *Msx1*, and repress *BF1*, which is expressed in the dorsal lateral forebrain. Ectopic localization of recombinant BMP5 or BMP4 into the chick embryo prosencephalon also results in holoprosencephaly, cyclopia, and a proboscis. Evaluation of these embryos has shown that dorsally expressed genes are still expressed in the forebrain. However, genes normally expressed in the ventral (mediobasal) forebrain are not seen or their expression is markedly reduced. Further evaluation of these embryos reveals evidence of cell death in the ventral (mediobasal) forebrain alone. Development of the holoprosencephaly sequence in these chick embryos is of further interest because the recombinant BMPs were introduced after the normal timing of prechordal plate induction on the forebrain.

In summary, the data on dorsal and ventral patterning of the forebrain are beginning to shed light on the molecular basis of holoprosencephaly. Development of this malformation may result from an initial failure of the prechordal plate to induce the overlying neuroectoderm to become the ventral forebrain. Some have hypothesized that the underlying defect in this malformation is the prechordal plate. Support for this concept comes from the *Shh* homozygous mutant mice, in which signaling from the entire notochord and prechordal plate fails to induce ventral markers and cell types, although these mice still express dorsal markers. Studies of two zebra fish mutants, *Cyclops* and *one-eye pinhead*, also support a primary defect in signaling from the prechordal plate as the pathogenesis of this malformation. The *cyclops* mutant has recently been shown to represent a defect in either nodal-related protein Ndr1 or Ndr2. The Nodal family of proteins are members of the transforming growth factor-β (TGF-β) superfamily, similar to the BMPs. The zebra fish *cyclops* mutants have mutations in *Ndr1* or *Ndr2* and show abnormal maintenance of the prechordal plate cells. In another mutant, *one-eyed pinhead*, the notochord develops but not the prechordal plate. The *one-eyed pinhead* mutant fish have cyclopia and are missing the ventral forebrain. A novel epidermal growth factor (EGF)-like molecule was recently identified as the *one-eyed pinhead* gene product. These data all support a primary defect of ventral forebrain induction by the prechordal plate. Data from ectopic localization of BMPs in chick embryos suggest that the holoprosencephaly sequence can occur in the presence of normal prechordal plate signaling if the responding cells (ventral forebrain) are absent. Further work will be required to define the pathogenesis of this sequence and the many genes involved. These data are likely to come from reverse genetics, experimental embryology, and the identification and characterization of additional mutant from multiple species.

Recent work has demonstrated that the holoprosencephaly sequence can arise not only from abnormal dorsal–ventral patterning of the forebrain, but also from defects in cholesterol biosynthesis. These abnormalities in cholesterol biosynthesis may have effects on the SHH signaling pathway. SHH is synthesized as an approximately 45 kDa precursor that undergoes an autocatalytic processing event that produces an approximately 19 kDa N-terminal product (SHH-N) and an approximately 25 kDa C-terminal fragment (SHH-C). SHH-N is responsible for all signaling by binding to Patched (PTCH), the primary receptor for SHH signaling. Binding of SHH-N to PTCH results in the suppression of PTCH, which normally functions to suppress SHH pathway activity. SHH-C's function is to perform the autocatalytic cleavage of the precursor protein and mediate the covalent modification of the N-terminal fragment by cholesterol. A second lipid modification subsequently results from attachment of palmitate at the N-terminus of SHH-N. Through this pathway it can be seen how cholesterol is required for normal SHH signaling. A second role for cholesterol is in the translocation of Smoothened (SMO) upon derepression

from PTCH. How this occurs is less clear but may depend on specialized lipid rafts. Lipid rafts have also been postulated to regulate other membrane components of SHH signaling, such as hedgehog interacting factor (HIF) and Dispatched (DISP) (see Figure 1).

Given these apparent roles for cholesterol in SHH processing, it is reasonable to speculate on the observed experimental overlap between SHH signaling and perturbations in the cholesterol biosynthesis pathways. As already noted (see discussion of environmental factors), blockers of cholesterol biosynthesis by naturally occurring compounds or specific drugs give rise to experimental animals with cyclopia and, in some cases, holoprosencephaly. It has also been shown that disruptions in genes involved in cholesterol transport can also phenocopy mutations in SHH. Furthermore, holoprosencephaly may occasionally be seen in the Smith–Lemli–Opitz syndrome, which is caused by defective cholesterol biosynthesis and can result in severe hypocholesterolemia. While it is enticing to implicate the cholesterol modification of SHH as the mechanism causing the phenocopy of SHH signaling abnormalities and cholesterol biosynthesis defects, the pathogenesis remains unknown. In fact, experimental blockage of the cholesterol biosynthesis pathway in mice resulted in cyclopia but not in the sterol modification of SHH-N. In contrast, tissues treated with these same compounds were unresponsive to cholesterol. This has led to the hypothesis that the cholesterol biosynthesis is primarily important for the response of target tissues (i.e., effects on SMO).

Finally, it is worth at least speculating on the pathogenesis of the frequently associated craniofacial and CNS anomalies. As discussed previously, the prechordal plate plays essential roles in induction of the ventral forebrain. SHH, which is expressed in the prechordal plate, appears to mediate this induction. SHH also appears to directly participate in craniofacial development, and the initial signaling for this is also likely to come from the prechordal plate. Eye development is also under the control of SHH, and this signaling is absent in cyclopic zebra fish. Normal eye development requires repressing the midline of a single eye field, which results in two lateral eye fields and thus two lateral eyes. The midline repression appears to be under the control of the prechordal plate. In the absence of this repression, a single eye field remains and results in the development of a single midline eye. Given that this eye field is located ventral to the site of nasal placode induction, the location of a proboscis above a single midline eye seems intuitive. Whether this is the embryologic origin of the craniofacial features in some cases of holoprosencephaly remains speculative at this time.

See also: Floor Plate Patterning of Ventral Cell Types: Ventral Patterning; Forebrain Development: Holoprosencephaly (HPE); Forebrain: Early Development; Olfactory Neuron Patterning and Specification; Sonic Hedgehog and Neural Patterning.

Further Reading

Chiang C, Litingtung Y, Lee E, et al. (1996) Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. *Nature* 383: 407–413.

Golden JA (1998) Holoprosencephaly: A defect in brain patterning. *Journal of Neuropathology & Experimental Neurology* 57(11): 991–999.

Golden JA (2004) Midline patterning defects. In: Golden JA and Harding BN (eds.) *Developmental Neuropathology*, pp. 14–25. Basel: ISN Neuropath Press.

Hahn JS and Barkovich AJ (2006) Factor analysis of neuroanatomical and clinical characteristics of holoprosencephaly. *Brain & Development* 28: 413–419.

Hayhurst M and McConnell SK (2003) Mouse models of holoprosencephaly. *Current Opinion in Neurology* 16(2): 135–141.

Jellinger K, Gross H, Kaltenbäck E, et al. (1981) Holoprosencephaly and agenesis of the corpus callosum: Frequency of associated malformations. *Acta Neuropathologica (Berlin)* 55(1): 1–10.

Jessell T and Lumsden A (1997) Inductive signals and the assignment of cell fate in the spinal cord and hindbrain. An axial coordinate system for neuronal patterning. In: Cowan W, Jessell T, and Zipursky S (eds.) *Molecular and Cellular Approaches to Neural Development*, pp. 290–333. New York: Oxford University Press.

Li H, Tierney C, Wen L, et al. (1997) A single morphogenetic field gives rise to two retina primordia under the influence of the prechordal plate. *Development* 124(3): 603–615.

Mann RK and Beachy PA (2004) Novel lipid modifications of secreted protein signals. *Annual Review of Biochemistry* 73: 891–923.

Ming JE and Muenke M (2002) Multiple hits during early embryonic development: Digenic diseases and holoprosencephaly. *American Journal of Human Genetics* 71(5): 1017–1032.

Norman M, McGillivray B, Kalousek D, et al. (1995) *Congenital Malformations of the Brain: Pathological, Embryological, Clinical, Radiological and Genetic Aspects*. New York: Oxford University Press.

Roessler E and Muenke M (2003) How a Hedgehog might see holoprosencephaly. *Human Molecular Genetics* 12(Spec No 1): R15–R25.

Takahashi TS, Kinsman S, Makris N, et al. (2004) Holoprosencephaly–topologic variations in liveborn series: A general model based upon MRI analysis. *Journal of Neurocytology* 33: 23–35.

Wallis DE and Muenke M (1999) Molecular mechanisms of holoprosencephaly. *Molecular Genetics and Metabolism* 68(2): 126–138.

Yakovlev P (1959) Pathoarchitectonic studies of cerebral malformations: III. Arrhinenecephalies (holoprosencephaly). *Journal of Neuropathology & Experimental Neurology* 18: 22–55.

Homeostasis at Multiple Spatial and Temporal Scales

A R Houweling, Humboldt University, Berlin, Germany
A van Ooyen, Vrije Universiteit, Amsterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

For neural circuits to function normally, many properties of the nervous system require precise regulation. Temperature, pH, and ionic gradients, for instance, need to be maintained at relative constant levels, despite fluctuations in the cellular environment. Complex homeostatic mechanisms are at work to keep them at their desired values. In general, a property is said to be under homeostatic control if a desired output is restored after a perturbation. Typically, homeostatic control is characterized by a desired output level (the set point), sensors that report a mismatch between the actual output and the set point, and mechanisms that restore the output to the set point level.

Like temperature and pH, neural activity appears to be under homeostatic control to prevent activity in neural circuits from becoming too high or too low. Throughout life, neural circuits are constantly perturbed by changes in synaptic strength, connectivity, morphology, sensory input, and protein turnover. When left unchecked, these disturbances may alter neural activity levels beyond the range in which neural circuits can operate properly. For example, for controlled activity propagation in feed-forward networks as well as stable ongoing activity in recurrent networks, synaptic strengths should be just so that each action potential triggers on average one action potential in another neuron; otherwise, activity will explode or die out. Without normalizing mechanisms, changes in synaptic strengths – brought about by Hebbian synaptic plasticity, for example – will thus destabilize network activity.

Changes in Neural Circuits during Development and Adulthood

During development, the number and strength of synapses alter dramatically. Once initial synaptic connections are established during the early stages of development, refinement of connections takes place by elimination of axons that project to the wrong targets and further elaboration of axons that project to the correct targets. This process of fine-tuning connectivity thus involves both synapse elimination and continued synapse formation. In many cases, the net effect is that the total number of synapses onto a postsynaptic cell increases during the period in which elimination occurs. For example, during the elimination of polyneuronal innervation in the neuromuscular junction, the total synaptic area on the muscle endplate becomes larger. Similarly, during the refinement of connections onto the lateral geniculate nucleus and cortex that creates ocular dominance columns, the number of synapses onto postsynaptic cells increases considerably. Despite these changes in synapse numbers during development, neural circuits have to remain functional.

In adulthood, synaptic strengths are also not constant but increase or decrease as a result of activity-dependent synaptic plasticity. In neural network models, correlation-based synaptic plasticity rules have been shown to lead to a hyperexcitable network without a stabilizing mechanism such as, for example, conservation of the average level of neural activity.

Likewise, the number of neurons is not constant but changes as new cells are generated (neurogenesis) and existing cells die (apoptosis). Neurogenesis occurs not only during development but, in brain areas such as the dental gyrus of the hippocampus, also in adulthood. Homeostatic mechanisms will therefore be necessary to regulate the changes in neural activity resulting from the insertion or deletion of neurons.

During development, neurons can experience a great amount of structural growth, which will affect their electrophysiological properties. The neural circuits in which they operate nevertheless have to remain functional. This requires that activity-dependent regulatory processes, targeting, for example, synaptic or intrinsic properties, must tune circuits to maintain their appropriate patterns of network activity during periods of neuronal growth.

Other changes during development that can affect electrical activity include switches in the subunit composition of neurotransmitter receptors. For example, the γ-aminobutyric acid $(GABA)_A$ receptor, the main inhibitory receptor in the mammalian brain, shows a subunit switch in the neonatal visual cortex around the time of eye opening. As a result of this switch, the duration of the inhibitory postsynaptic current decreases, effectively reducing the level of inhibition.

Experimental Evidence for Homeostasis of Neural Activity

Several experimental observations support the idea of homeostatic regulation of neural activity. First, in neuronal cultures, firing rates are restored after manipulation of the global level of activity or the

activity level in single cells. In cultured cortical networks composed of excitatory and inhibitory neurons, application of bicuculline (which reduces inhibition) initially raises the activity. After 2 days in bicuculline, however, activity returns to control levels. Conversely, blockade for 2 days of spiking activity with tetrodotoxin generates a hyperactive network when tetrodotoxin is removed. In this network, the firing rates of both excitatory and inhibitory neurons are increased. Suppressing activity in a single hippocampal neuron by overexpressing an inward rectifier K^+ channel initially reduces firing rates. After a few days, however, the activity of the neuron is restored to control levels. Second, at the neuromuscular junction, changes in presynaptic efficacy induce compensatory changes in postsynaptic efficacy and vice-versa, effectively keeping neuromuscular transmission at a relatively constant level. For example, reducing the number of synapses causes a compensatory increase in the postsynaptic response to a single vesicle of neurotransmitter. Conversely, hyperpolarization of muscle fibers induces an increase in presynaptic release. Third, the removal of neuromodulatory synaptic inputs to stomatogastric ganglion (STG) neurons initially slows or abolishes the typical STG pyloric rhythm of alternating bursts of activity. However, after a period of days, normal rhythmic burst firing resumes. The presence of a wide variety of activity-dependent mechanisms that counteract changes in activity forms further evidence for the homeostatic control of neural activity.

The Set Point

In the context of homeostatic control of neural activity, the set point represents the desired level of neural activity to which neurons or neural networks strive to return after experiencing a perturbation. There are, however, many open questions regarding the nature of such a set point:

1. What is the spatial scale of the set point? Do neurons or perhaps subneuronal structures (such as neurites) have their own individual set points, or is the set point specified at the level of neuronal populations? Under some conditions, homeostasis of neural activity can be a cell-autonomous process. Hyperpolarizing single hippocampal pyramidal cells in neuronal cultures by overexpressing an inward rectifier K^+ channel initially decreases firing rates, but after a few days, firing rates are restored to normal levels. The scaling of excitatory synapses on inhibitory neurons is dependent on the activity-dependent release of brain-derived neurotrophic factor from nearby pyramidal neurons, which perhaps suggests that homeostasis of neural activity in interneurons is not cell-autonomous. An important question is whether cell-autonomous homeostatic processes are capable of achieving homeostasis of activity at the network level. Some studies suggest that this may indeed be the case, but there are also examples showing that synaptic and cellular homeostatic rules do not achieve homeostasis of neural activity at the network level, at least under some conditions (see the section titled 'Discussion').

2. Does the set point involve activity as it occurs over the entire day or only during certain periods of the day? The temporal range of the set point may span long periods of ongoing activity or perhaps only periods during which a neuron or a neural circuit actively processes particular synaptic inputs. Another possibility is that the set point involves only neural activity during sleep. In neuronal cultures, homeostasis appears to involve ongoing activity. At the neuromuscular junction, peak depolarization is precisely restored after postsynaptic activity manipulations, suggesting that the set point involves an adequate response to synaptic input.

3. What aspects of neural activity are under homeostatic control? At the level of individual neurons, the possibilities that have been suggested include a temporal average of firing rate, membrane potential, intracellular Ca^{2+} concentration, amount of excitatory synaptic receptor activation, or a particular pattern of spiking activity. At the level of neuronal populations, a spatial average of firing rate or a particular spatial pattern of activity could be under homeostatic control. In STG neurons the set point may involve a precise pattern of burst firing. In neuronal cultures the set point may involve the average firing rate of neurons. Furthermore, perhaps there is not a single set point but a range of activities that is allowed.

4. How is the set point encoded? The set point may be encoded as an explicit cellular property, or alternatively it could come about as an emergent property of a large number of interacting processes. Although it may not be explicitly encoded as such, the attractor state of a complex system may function as a set point. For example, spike timing-dependent plasticity (STDP) contributes to stabilizing neural activity levels, but it does not contain an explicit set point. In this form of long-term synaptic plasticity, presynaptic action potentials that precede postsynaptic spikes strengthen a synapse, whereas presynaptic action potentials that follow postsynaptic spikes weaken it. STDP has the effect of keeping the total synaptic input to the neuron roughly constant, independent of the presynaptic firing rates. In the Bienenstock, Cooper, and Munro model, the set point of postsynaptic activity is not explicitly specified but arises as a

result of the local synaptic learning rule. If the threshold value of postsynaptic activity above which long-term potentiation of synapses occurs increases faster than linearly with the average postsynaptic activity, the synaptic strengths will adjust to keep the postsynaptic activity near a particular value.

5. How are differences between the actual activity and the set-point activity level sensed? Many experimental studies indicate that intracellular Ca^{2+} concentration is a good indicator of neural activity, and cells appear to be capable of sensing the difference between actual and desired Ca^{2+} levels. The intracellular Ca^{2+} concentration increases in response to activity (e.g., through enhanced influx through *N*-methyl-D-aspartate channels and voltage-sensitive Ca^{2+} channels) and decreases during inactive periods. Intracellular Ca^{2+} is an important messenger signal for a wide variety of cellular processes and can modulate, for example, levels of gene expression. Different firing patterns, such as bursting or regular spiking, result in different temporal profiles of Ca^{2+}. These profiles can be captured by integrating Ca^{2+} on different timescales.

Mechanisms Contributing to the Homeostasis of Neural Activity

A variety of mechanisms operating at different spatial and temporal scales have been suggested to contribute to the homeostasis of neural activity.

Synaptic Scaling

In cortical cultures, changing the level of network activity causes dynamic adjustments in the relative strengths of excitatory and inhibitory connections. Raising activity results in reduced feedback excitation and increased feedback inhibition onto pyramidal neurons. Lowering activity produces the opposite set of changes. A few days of pharmacological activity blockade in cortical cell cultures increases the amplitude of miniature excitatory postsynaptic currents (mEPSCs) and evoked EPSCs in pyramidal cells. Conversely, prolonged enhanced activity reduces the amplitude of mEPSCs. Similar bidirectional activity-dependent changes in mEPSC size have been observed in spinal cell cultures. It is interesting that activity blockade scales down the amplitude of miniature inhibitory postsynaptic currents (mIPSCs), in the opposite direction of excitatory currents. The expression site of homeostatic excitatory synaptic plasticity in cortical cultures is a controversial issue. Synaptic scaling of mEPSCs occurs in part postsynaptically by changes in the number of glutamate receptors, but also presynaptically by changes in release probability. The pre- or postsynaptic expression site has recently been shown to depend on the time neurons have spent *in vitro*. Synaptic scaling of mIPSCs is accompanied by a change in the number of $GABA_A$ receptors clustered at synaptic sites. Chronic suppression of activity in a single hippocampal neuron increases excitatory synaptic strengths on the neuron, but not sufficiently to alter inhibitory synaptic strengths.

It remains to be explored whether homeostatic synaptic plasticity operates *in vivo* and in cell cultures similarly. Recent evidence suggests that excitatory synapses in intact rodent visual cortex are upregulated after sensory deprivation. Two days of monocular deprivation scales up mEPSC amplitudes by 15–30% in a layer- and age-dependent manner. However, these changes are smaller than those seen in culture. One possible explanation is that sensory deprivation only partially reduces activity compared with the complete activity block in cultures.

An important issue is the conditions under which homeostatic synaptic plasticity can cause the network to settle into a desired activity level. It seems rather obvious that an individual neuron in isolation may be able to keep its activity close to a set point value by increasing synaptic excitation and decreasing synaptic inhibition. Neurons in the brain, however, are not isolated but embedded in circuits with recurrent synaptic connectivity. Such neural networks are known to be very sensitive to the balance of synaptic excitation and inhibition. An altered balance of excitation and inhibition may affect synchrony in the network and in extreme conditions result in hyperactivity or other nonphysiological oscillatory states. Thus, scaling excitatory and inhibitory synaptic strengths in opposite directions may have undesirable effects for the stability of network activity. This is an important issue that can be explored in neural network models.

Intrinsic Excitability

Intrinsic excitability is also regulated by activity. Chronic activity blockade in cortical cell cultures enhances Na^+ currents and reduces K^+ currents in pyramidal cells, resulting in an enhanced responsiveness to current injections. Fast Na^+ conductances in pyramidal cells are upregulated by approximately 30% and K^+ conductances (of the delayed rectifier and a persistent outward current) are downregulated by about 30–60% after 2 days of activity blockade. Inhibitory neurons also display an increased excitability after chronic activity blockade, but through a different mechanism. In somatostatin-positive inhibitory neurons, the densities of a passive leak current and the hyperpolarization-activated (H-) current are downregulated, which increases excitability mostly

through a decrease in membrane input conductance. Intrinsic excitability may also be modified in an adaptive manner at the very short timescale of minutes. In hippocampal CA1 pyramidal neurons, an increase in H-current can be observed within tens of minutes of enhanced activity, which reduces firing rates in response to current injections. Reducing background activity results in an increased excitability within tens of minutes through a decrease in a sustained K^+ conductance.

In model neurons (and networks) with activity-dependent regulation of maximal conductances of ionic currents, conductances may adjust to restore a given pattern of electrical activity following a perturbation. For example, in a multicompartmental model neuron, regulation of some of the maximal conductances (of the fast Na^+ current, the delayed rectifier K^+ current, and the transient K^+ current) by the local intracellular Ca^{2+} concentration acts as a gain control mechanism to restore the neuron to its optimal firing range after potentiating a set of synapses. Another example concerns a model of a bursting STG neuron, in which the maximal conductances of seven voltage-dependent membrane currents are regulated by Ca^{2+} sensors acting on different timescales. When the electrical activity of this model neuron is perturbed, such as by changing the reversal potential for the K^+ currents or by applying hyperpolarizing current, the maximal conductances adjust to restore the original bursting pattern. Under some conditions, activity-dependent regulation of intrinsic excitability, operating at the level of individual neurons, may stabilize network activity in models of synaptically connected neurons. In a model of the neocortex, neurons fire initially asynchronously and at low firing rates. Removing the external synaptic inputs strongly reduces the firing rates of all neurons. Activity-dependent regulation of some of the maximal conductances in pyramidal cells (of the fast and persistent Na^+ currents and two types of K^+ current) restores an irregular state with low firing rates comparable to that in the intact model. Another example concerns a three-neuron network model of the pyloric circuit of the STG. The small network model recovers rhythmicity after blockade of neuromodulatory synaptic inputs using activity-dependent regulation of some of the maximal conductances (of a K^+ current and a Ca^{2+} current).

Structural Plasticity

Neurite elongation and branching are under the control of the growth cone (a specialized structure at the tip of a growing neurite) and are sensitive to the Ca^{2+} concentration inside the growth cone. Electrical activity in general increases the Ca^{2+} influx through voltage-gated Ca^{2+} channels. In neuronal cultures, raising the level of neural activity has been found to cause retraction of neurites, whereas lower levels allow further outgrowth. Model studies have shown that this activity-dependent neurite outgrowth will contribute to the homeostasis of neural activity, both in purely excitatory and in mixed networks of excitatory and inhibitory cells. In purely excitatory networks, for example, when the level of activity is too high, the neuron will withdraw neurites and, consequently, lose connections with other neurons, resulting in a decrease in neural activity. Conversely, when the activity is too low, neurites will grow out and new connections will be formed, resulting in an increase in neural activity. Experimental work on the Xenopus retinotectal system suggests that neuronal activity, via activation of Ca^{2+}/calmodulin-dependent protein kinase, slows the outgrowth of axonal and dendritic arbors and controls synapse formation. In the developing spinal cord, the intracellular Ca^{2+} in growth cones has been found to regulate axon elongation, with the rate of axon elongation being inversely proportional to the frequency of Ca^{2+} transients. Many experimental studies have demonstrated considerable structural plasticity in the adult brain, including synapse formation and elimination and remodeling of axons and dendrites, which might be influenced by activity. Large-scale changes in dendrites can occur in response to environmental enrichment and with pharmacological manipulations. In the superior cervical ganglion of the spinal cord, highly dynamic dendritic branches have been found. Recent long-term imaging studies in the *in vivo* adult neocortex have found high levels of structural plasiticity of axonal branches and boutons. In addition, *in vivo* and *in vitro* studies have revealed rich dynamics of postsynaptic dendritic spines, including spine formation, retraction, and shape changes. Changes in spine number are likely to indicate synapse formation and elimination.

If the size of the dendritic tree increases or decreases, then the activity of the neuron can change not only as a result of loss or formation of synaptic connections, but in principle also directly as a consequence of altered morphology. In many cell types – including pyramidal cells in visual cortex, somatosensory cortex, prefrontal cortex, and hippocampus – firing patterns (both spike frequency and spike pattern) are correlated with dendritic morphology. Results from modeling studies suggest a causal relationship between dendritic morphology and firing patterns. For example, models of neocortical neurons that have the same channel densities and kinetics but differ in their dendritic shape and size can generate firing patterns ranging from regular spiking to bursting when stimulated at the soma with a fixed current

injection. Both dendritic size and dendritic branching structure may influence firing patterns, especially the cell's propensity for generating bursts (the occurrence of clusters of spikes with short interspike intervals).

Number of Inhibitory and Excitatory Cells

Insertion of new excitatory neurons into adult neuronal networks, as takes place in the dentate gyrus of the hippocampus, requires the existence of homeostatic mechanisms to regulate the resulting changes in neural activity. One of the possibilities, for which there is some experimental evidence, is that not all the newly generated neurons become excitatory granule cells but that some take on features typical of inhibitory GABAergic basket cells. A related mechanism is the conversion of preexisting granule cells to the GABAergic phenotype. Into maturity, granule cells express low levels of the inhibitory transmitter GABA.

Discussion

Besides keeping neural activity in the range where neural circuits can operate properly, homeostasis of neural activity has been postulated to lead to some additional beneficial properties. For example, homeostatic synaptic plasticity may prevent the runaway of synaptic strengths during Hebbian modifications or facilitate the competition between synapses during development.

Modeling studies have shown that homeostatic activity-dependent neurite outgrowth leads to a number of interesting emergent phenomena, including a transient phase of high network connectivity during development, size differences between the spatial extent of neurites of excitatory and inhibitory neurons, and compensatory sprouting of remaining cells following cell death. Furthermore, activity-dependent neurite outgrowth has been found to construct networks with a critical connectivity level in which the frequency of spontaneous network bursts of activity shows a power law relationship with burst duration and size. This critical connectivity has been shown to optimize information transmission while preventing runaway network excitation.

It is generally believed that homeostatic synaptic plasticity acts to restore a target pattern of activity in networks after a perturbation. However, under some conditions, rather than restoring a desired activity state, homeostatic plasticity rules may instead lead to pathological forms of neural activity. In a computer model of the neocortex with a biologically based homeostatic plasticity rule that operates to maintain firing rates, pathological neural activity has been found to develop after complete cortical deafferentation (such as may occur after severe head trauma). Homeostatic synaptic plasticity increases the network excitability to a level at which only synchronized bursting activity becomes possible.

Under some conditions, model networks consisting of excitatory and inhibitory cells in which individual neurons adapt their intrinsic excitability may become unstable. The stability of the adaptation at the network level critically depends on the relation between the cellular adaptation rates of both neuron populations, combined with their respective gains. If input–output relations of both excitatory and inhibitory cells have equal gain, the inhibitory cells need to adapt more slowly than the excitatory cells for a network to remain stable and attain its target state. If this condition is not met, the activity level in the network keeps increasing until the network abruptly falls silent, after which activity gradually builds up again.

Models exploring the consequences of homeostatic, activity-dependent neurite outgrowth have shown that networks with both excitatory and inhibitory cells can under some conditions develop into a pathological network state. This state, in which the average activity is still at set point level, is characterized by a high level of network connectivity and fast, epileptiform oscillations in electrical activity. Blocking electrical activity in the network for a certain period of time during development can induce the network to evolve into this state. It is interesting and perhaps counterintuitive that the higher the level of inhibition (e.g., the number of inhibitory cells) during the early stages of development, the more likely the network is to end up in the pathological state. As a consequence of the interactions between excitation, inhibition, and outgrowth, too much inhibition prevents the normal pruning of exuberant connections, resulting in a network with epileptiform activity. This model result may shed light on experimental observations that in rat pups with hypoxic–ischemic encephalopathy, epileptiform activity later on in life appears to be due to the preferential survival, not loss, of inhibitory elements.

See also: Activity-Dependent Remodeling of Presynaptic Boutons; Calcium Homeostasis in Glia; Dendrite Development, Synapse Formation and Elimination; Neural Oscillators and Dymanmical Systems Models; Neurogenesis in the Intact Adult Brain; Neuronal Plasticity after Cortical Damage; Plasticity of Intrinsic Excitability; Potassium Homeostasis in Glia; Spike-Timing-Dependent Plasticity Models; Spiking Neuron Models; Synaptic Plasticity: Neuronal Sprouting.

Further Reading

Burrone J and Murthy VN (2003) Synaptic gain control and homeostasis. *Current Opinion in Neurobiology* 13: 560–567.

Corner MA, Van Pelt J, Wolters PS, Baker RE, and Nuytinck R (2002) Physiological effects of sustained blockade of excitatory synaptic transmission on spontaneously active developing neural networks. *Neuroscience and Biobehavioral Reviews* 26: 127–185.

Davis GW (2006) Homeostatic control of neural activity: From phenomenology to molecular design. *Annual Review of Neuroscience* 29: 307–323.

Goldberg DJ and Grabham PW (1999) Braking news: Calcium in the growth cone. *Neuron* 22: 423–425.

Houweling AR, Bazhenov M, Timofeev I, Steriade M, and Sejnowski TJ (2005) Homeostatic synaptic plasticity can explain post-traumatic epileptogenesis in chronically isolated neocortex. *Cerebral Cortex* 15: 834–845.

Marder E and Goaillard JM (2006) Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience* 7: 563–574.

Meltzer LA, Yabaluri R, and Deisseroth K (2005) A role for circuit homeostasis in adult neurogenesis. *Trends in Neurosciences* 28: 653–660.

Miller KD (1996) Synaptic economics: Competition and cooperation in synaptic plasticity. *Neuron* 17: 371–374.

Remme MWH (2006) *Scaling of Excitability in Neurons and Networks*. PhD Thesis, University of Amsterdam.

Turrigiano GG and Nelson SB (2004) Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience* 5: 97–107.

Van Ooyen A (1994) Activity-dependent neural network development. *Network* 5: 401–423.

Van Ooyen A (2001) Competition in the development of nerve connections: A review of models. *Network* 12: R1–R47.

Van Ooyen A (ed.) (2003) *Modeling Neural Development*. Cambridge, MA: MIT Press.

Zhang W and Linden DJ (2003) The other side of the engram: Experience-driven changes in neuronal intrinsic excitability. *Nature Reviews Neuroscience* 4: 885–900.

Homology and Homoplasy

A B Butler, George Mason University, Fairfax, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The word 'homology' literally translates as 'the study of sameness.' In biology, it refers to the relationship of a character in one species to what is believed to be the same character in another species. A character can be any definable attribute, such as an organ, part, feature, trait, structure, developmental process or pathway, or morphogenetic field. In 1843, the British comparative anatomist Richard Owen proposed the concept of homology and defined it as "the same organ in different animals under every variety of form and function." Thus, although the original definition recognized that the 'same' character could vary in both its morphology and function across species, it did not specify the criteria for recognizing sameness. Our concept of homology has therefore changed in step with advances in understanding the processes by which morphology is generated.

Phylogenetic Homology

Owen's definition of homology preceded the announcement of the Darwin–Wallace theory of evolution by 16 years. The concept of homology was subsequently linked to the Darwinian perspective, and those studying phylogeny, or systematics, redefined Owen's original notion. The criteria for recognizing homology became phylogenetic: homologous characters were those that had been inherited from a common ancestor. A commonly cited definition in the current literature is that of Edward Wiley:

> A character of two or more taxa is homologous if this character is found in the common ancestor of these taxa, or, two characters (or a linear sequence of characters) are homologous if one is directly (or sequentially) derived from the other(s).

Thus, the character must have been expressed in the phenotype of the common ancestor and been present in all subsequent members of each descendant lineage. This type of homology is now referred to as phylogenetic, or historical, homology. It is the opposite of homoplasy, which is the independent evolution of apparently similar characters and comprises three types: parallelism, reversal, and convergence.

Analysis of phylogenetic homology is particularly challenging for the study of the nervous system, since most neural structures do not fossilize. Only some surface features of the brain can be detected from fossilized endocast material. For inferring the phylogeny of internal brain structures, the most common method, called cladistic analysis, is to examine the distribution of a character across the members of a taxon for which the phylogenetic relationships have been established by nonneural characters. If the neural structure in question is present in all or most members of the taxon, then it is presumed to have been present in the common ancestor, according to the principle of parsimony. All living species are a mosaic of characters, some of which were present in a common ancestor and thus are plesiomorphic (primitive, or ancestral) characters and some of which are more recently evolved and thus are apomorphic (specialized, or derived) characters.

Shared apomorphic characters (synapomorphs) are used to recognize sets of species that are more closely related to each other than to other species that lack these characters. This method is called outgroup comparison and is used to establish phylogenies as well as to hypothesize homology. Neural characters can be evaluated across such phylogenies. For example, consider a character that occurs in the brains of reptiles and birds but is absent in the brains of mammals. Was this character present in the common ancestor of all amniotes and subsequently lost in the ancestral line to mammals, or was it gained only in the common ancestor of reptiles and birds after the mammalian line diverged? Presence of the character in the brains of amphibians, as the outgroup to amniotes, would indicate the former, while it absence in amphibians would indicate the latter.

For recognizing a character in the first place as being the 'same' character in different species, multiple criteria can and should be used. Similar features are most commonly used to identify potentially homologous characters. The probability that the character is the same across species increases with the number of points of similarity. Homologous genes and developmental pathways that produce the characters are of great current interest in this regard, but it is important to weigh other points of similarity as well. For neural structures, location, cellular architecture, histochemical traits, and afferent and efferent connections are all informative. Also, the basis of the homology needs to be specified unless the specification is implicit and unambiguous. When comparing one nucleus with another nucleus, for example, the specification is implicit: the two nuclei are being compared as nuclei. In many other cases, the specification needs to be stated: the wing of a bird and the wing of a bat are not homologous as wings; they are, however, homologous as forelimbs.

Homoplasy

Since similarity can arise from several different causes, one must distinguish between similarity due to inheritance from a common ancestor and similarity gained independently of such ancestry. The term 'homoplasy' was coined by Lankester in 1870 to refer to the latter circumstance – the independent gain of similar characters in different members of a taxon. It is the opposite of phylogenetic, or historical, homology. Homoplasy traditionally comprises three types: parallelism, reversal, and convergence. George Gaylord Simpson also distinguished analogy, mimicry, and chance similarity, although these are most often considered as different instances of convergence. Recent advances in our understanding of the relationship of embryological development to evolution have necessitated a reevaluation of the adjective 'independent,' particularly concerning parallelism and reversal, as further discussed in the following text.

Homoplasy presents a challenge to the task of reconstructing the evolutionary history of neural structures. Since, as already noted, the internal components of the nervous system do not fossilize, one cannot utilize the fossil record in order to find out what neural structures the ancestor of any extant lineage had or whether a particular structure has been phenotypically expressed with consistency along a lineage. The degree of similarity may help in some situations to differentiate convergence from phylogenetic homology, but, as will be further discussed, among closely related species, either parallelism or reversal could easily be mistaken for phylogenetic homology.

Convergence was defined by Edward Wiley as "the development of similar characters from different preexisting characters." It is usually assigned to cases of similarity in animals with only a remote common ancestry, and the similarity is usually superficial, such as with the resemblance of the wing of a bird to the wing of a bat. Both are aeronautically suited to flight, but the use of feathers versus membrane and different modifications of the bony elements of the hands belie the overall, superficial similarity. Comparison of the anatomical features makes it difficult to imagine a common ancestral condition that could have given rise over evolution to the two different types of wings, and of course the fossil record confirms that the common amniote ancestor had an upper limb that then independently evolved into the two types of wings in the two separate lineages. While the wings of bats and birds are phylogenetically homologous as upper limbs, they are convergently homoplaseous as wings. Likewise, the wings of flying fishes, derived from their forelimbs, are convergent, as are the wings of insects, which demonstrate extreme differences in their anatomical components. The study of convergence addresses how similar selective pressures can result in the evolution of similar solutions that are truly independently arrived at.

In contrast to characters evolved by convergence, characters that result from parallelism and reversal exhibit minute and extensive similarities. Parallelism has been defined by Edward Wiley as "the independent development of similar characters from the same plesiomorphic [i.e., ancestral] character." Likewise, David Wake has defined parallelism as "the production of apparently identical traits by the same generative system." Parallelism usually occurs in closely related animals, although in some instances extensive similarity occurs for characters scattered across wide phylogenetic gaps. Reversal is the reacquisition within a single lineage of a developmental or structural character of its ancestral condition, with the intermediate member(s) of the lineage having lacked the character or exhibited a different condition.

Homoplasy is generally considered to be the result of adaptive evolution, with different lineages responding in similar ways to similar adaptive pressures. However, the similarities seen in cases of convergence, parallelism, and other instances of homoplasy may also be the result of limits of design. For example, only a circumscribed set of options exists for creating a wing surface that will permit flight – the feathers of birds and the membranous surfaces of bats, flying fishes, pterosaurs (the extinct, flying reptiles), and insects. Extending the body surface by limb extension and/or the deployment of membranous structures is also used for parachuting or gliding by other animals, including flying squirrels and some lizards, but solutions other than membranes or feathers have not appeared.

The most frequently used tests to distinguish homoplasy from phylogenetic homology are evaluating the degree of similarity and determining whether congruence occurs, that is, whether the distribution of the character in question is congruent with a phylogeny established using other characters not related to the character in question. Cases of convergence fail both of these tests. In contrast, some cases of parallelism may exhibit extensive similarity but fail the test of congruence. Cases of parallelism in closely related lineages as well as cases of reversal often pass both the similarity and congruence tests and can thus be mistaken for phylogenetic homology if the fossil record cannot be used to illuminate the situation.

Evolutionary-Developmental Approaches to Homology

New information on the genetic bases for characters and the stunningly widespread distribution of many

of the patterning genes across almost all lineages of multicellular animals has challenged the dictum that homology must be phylogenetic and meet the requirements of Edward Wiley's definition. The concept of homology has been reexamined in light of the new synthesis of evolutionary and developmental biology. While embryological development does not recapitulate evolution *per se*, the sequence of development is from the general to the specific, as Karl Ernst von Baer recognized in the nineteenth century, so that the early steps in the developmental sequence are those that are shared across all groups within a large taxon, while the later steps produce the more specialized characters that uniquely characterize a particular genus or species.

Almost all characters that are phylogenetically homologous are produced in the phenotype as the result of the same genes and generative pathways that were inherited from a common ancestor. In some instances, a different generative event can be substituted for the inherited one in part of the pathway for production of the character. The evolution of generative pathways is of great interest in regard to how characters become stabilized and conserved within a taxon and how they vary so that selection can act on them. Günter Wagner's concept of biological homology for characters that share a set of developmental constraints (specified by the generative pathways) addresses these mechanisms. Biological homology seeks to understand the patterns of character diversification and thus focuses on the mechanisms of character evolution; in most cases, it overlaps with phylogenetic homology.

The issue of shared genes and generative pathways as a basis for comparison is a pivotal one, since it has long been recognized that instances of parallelism and reversals are both the result of the inheritance of a shared generative basis, inherited with continuity from the common ancestor. Various other terms for parallelism (which is more frequently encountered than reversals) thus have been offered to denote this situation. Gavin de Beer proposed the term 'latent homology' for characters that have a high degree of similarity but occur only within some members of a taxon. Latent homology implies a 'genetically based homology' that allows for the expression of the character in selected descendant species or groups of species, even though the character itself was not a 'visible phenotypic' character in the common ancestor or in many of the descendant taxa. Likewise, the terms 'homoiology,' or 'deep homology,' apply to similarity evolved from a homologous base through parallel evolution. For example, the wings of birds, bats, and flying fishes – even though often cited as examples of convergence – are homoiologous, since they all exhibit the 'tendency' of the vertebrate upper limb to form extended surfaces usable for flight, presumably using a shared genetic basis for at least part of this morphological evolution, in contrast to the much more phylogenetically distant wings of insects. The more information comes to light about the commonalities of patterning genes across all multicellular animals, the more cases of parallelism and reversal and even some apparent cases of convergence fail the criterion of being 'independent.' Even between taxa as diverse as insects and mammals, some of the same patterning genes specify the dorsoventral and rostrocaudal divisions of the brain, for example.

Thus, in contrast to cases of true convergence, it is now clearly acknowledged that most cases of phylogenetic homology as well as parallelism and reversal share the evolutionary commonality of inherited genetic basis. Due to this recent illumination of the processes for the distribution of various characters and the shared genetic basis in many cases of homoplasy, the previous lumping of the process of convergence with the processes of parallelism and reversal under this rubric has become conceptually untenable. Parallelism and reversal are distinguished from phylogenetic homology only by the distribution of the character across the taxon. A concept that unites these three ways of generating sameness is that of generative homology, or syngeny, which literally means 'same genes' or 'same genesis.' Syngeny comprises all instances of characters produced by the same genes, regardless of phylogenetic distribution. It applies to characters that are produced by generative bases inherited from a common ancestor, whether the character was expressed in the common ancestor (and with continuity along a lineage of descent) or not. In contrast, the complementary concept of allogeny (meaning 'different genes' or 'different genesis') is the basis for true cases of convergence.

Syngeny comprises almost all cases of phylogenetic homology, as well as situations where the phylogenetic expression of a character is sporadic, either across a taxon of extant species or along a lineage. For example, eyes occur in several very distantly related groups of animals – arthropods (including insects such as flies), cephalopods (octopus, squid, etc.), and vertebrates. The structure of these eyes varies considerably, but the same patterning gene, *Pax6*, is one of a few essential genes for initiating the generative cascade that produces the eyes in both fruit flies and mammals (and presumably in cephalopods as well). This gene is known to be expressed in all light-detecting, rostrally lying structures examined to date, from the eye spots of planarian worms and those along the mantles of clams to the compound eyes of flies and the retinal eyes of

mammals. While the large, whole eyes of flies and mammals cannot be considered homologous in the phylogenetic sense, they are, however, syngenous – homologous in the generative sense – as whole eyes initiated by *Pax6*. In contrast, the finer details of these eyes – ommatidia in flies versus the retinal eye of mammals – to the extent that these features are truly independently evolved, result from the actions of different patterning genes and are thus allogenous.

The concepts of biological and generative homology, as well as those that more specifically recognize a genetic basis for parallelism and reversals, are aligned with the new approaches to evolution that integrate developmental and evolutionary perspectives. As noted by Axel Meyer,

> The biological basis for these seemingly disparate kinds of 'sameness' in evolution [including all forms of homoplasy] may in some, or even most, instances not be all that different and may be based on the same principle – the long evolutionary retention of genes, gene interactions and developmental mechanisms. . . . Therefore, the biological basis of both [historically] homologous traits (those that are evolutionarily always expressed) and homoplasious traits (those that are not always 'on,' but are 're-awakened' during evolution) might not be so different, and the distinction between homology and some forms of homoplasy may be somewhat artificial.

Other Applications of the Homology Concept

Several other applications of the concept of homology – whether phylogenetic, biological, or generative – need to be mentioned here. The concept of field homology has enjoyed a renewed interest recently, particularly in regard to comparisons of some sets of neural populations across different vertebrate taxa. Hobart Smith defined field homology as the "derivation of structures, however similar or dissimilar, from a common anlage." Anlagen, or morphogenetic fields, are defined by Scott Gilbert as "discrete [modular] units of embryological development . . . produced by the interactions of genes and gene products within specific bounded domains . . .[and]. . . the morphogenetic field . . . is seen as a major unit of ontogeny whose changes bring about changes in evolution [and] homologous morphogenetic fields can exist . . . between different organisms." Some opponents of applying this concept in neurobiology argue that it compares adult characters in one taxon to embryonic ones in another. Those that find the concept useful consider it to compare a set of adult characters to another set of adult characters, each set derived from a specified homologous morphogenetic field. Specification of the field in question allows for this concept to apply broadly to a large set of structures, such as all parts of the forelimb as derivatives of the forelimb morphogenetic field, or to a much more restricted set, such as the digits as derivatives of the morphogenetic field for the hand.

Serial, or iterative, homology is also a generally recognized concept, albeit not universally accepted. Some take the position that iterative structures are not homologous but simply iterative. The concept applies to similar units, such as segments, that repeat along an axis of the body. Vertebrae are an excellent example of iteration or serial homology. Concerning development, Scott Gilbert has noted that "homologous morphogenetic fields can exist within the same organism (serial process homology)." It is important to note, however, that this phenomenon is not simply pleiotropy, the actions of a gene in multiple tissues. Many of the patterning genes that are shared across most multicellular animal taxa are involved in more than one role at different levels of the body. A single patterning gene may be involved in specifying clearly different structures, such as parts of the nervous system in the head and the heart in the thorax, but that does not confer any homologous or even iterative relationship for such disparate structures. Only structures with multiple points of similarity that repeat along an axis within an individual are candidates for serial homology.

The question of whether homology can be partial is a hotly debated one. David Wake convincingly defends the case for partial homology, citing the system of *Hox* patterning genes. He notes that "all homology among genealogical entities is partial at some level, except between the closest relatives." The concept of partial homology dovetails with the concepts of syngeny and allogeny, since, like eyes, virtually all structures across members of a taxon will have some genes and generative pathways in common and be syngenous, as specified, in that regard and at the same time have some genes and generative pathways that differ, causing the finer levels of variation among them, and be allogenous, as specified, in that regard.

Finally, the concept of homology has been applied to behaviors, with particular import to the field of neuroethology. Some would assert that only those behaviors that are generated by homologous structural characters could be themselves homologous, but others have argued that the inheritance and continuity of a behavior itself can be preserved in some cases even though it is generated by nonhomologous structures. This situation is similar to that of substitutions of part(s) of the generative pathways for the production of homologous structural characters in the phenotype.

Our understanding of homology – at all levels – is crucial to our appreciation of how the nervous system has evolved in its many different guises across the kingdom of multicellular animals. As David Wake has stated, homology is "the central concept for *all* of biology." The new insights into the interactions of embryological development and evolution continue to challenge and redefine our comprehension of sameness.

See also: Autonomic Nervous System: Neuroanatomy; Comparative Neurobiology: History; Computational Neuroethology; Neuroethological Perspective.

Further Reading

Bock GR and Cardew G (eds.) (1999) *Homology. Novartis Foundation Symposium 222*. Chichester: Wiley.

Campbell CBG and Hodos W (1970) The concept of homology and the evolution of the nervous system. *Brain Behavior and Evolution* 3: 353–357.

Gilbert SF, Opitz JM, and Raff RA (1996) Resynthesizing evolutionary and developmental biology. *Developmental Biology* 173: 357–372.

Hall BK (ed.) (1994) *Homology: The Hierarchical Basis of Comparative Biology*. San Diego, CA: Academic Press.

Meyer A (1999) Homology and homoplasy: The retention of genetic programmes. In: Bock GR and Cardew G (eds.) *Homology. Novartis Foundation Symposium 222*, pp. 141–153. Chichester: Wiley.

Puelles L and Medina L (2002) Field homology as a way to reconcile genetic and developmental variability with adult homology. *Brain Research Bulletin* 57: 243–255.

Roth VL (1984) On homology. *Biological Journal of the Linnean Society* 22: 13–29.

Simpson GG (1961) *Principles of Animal Taxonomy*. New York: Columbia University Press.

Smith H (1967) Biological similarities and homologies. *Systematic Biology* 16: 101–102.

Van Valen LM (1982) Homology and causes. *Journal of Morphology* 173: 305–312.

Wagner GP (1989) The biological homology concept. *Annual Review of Ecology and Systematics* 20: 51–69.

Wake DB (1994) Comparative terminology. *Science* 265: 268–269.

Wake DB (1996) Introduction. In: Sanderson MJ and Hufford L (eds.) *Homoplasy: The Recurrence of Similarity in Evolution*, pp. xvii–xxv. San Diego, CA: Academic Press.

Wake DB (1999) Homoplasy, homology and the problem of 'sameness' in biology. In: Bock GR and Cardew G (eds.) *Homology. Novartis Foundation Symposium 222*, pp. 24–33. Chichester: Wiley.

Wiley EO (1981) *Phylogenetics*. New York: Wiley-Liss.

Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

Energy homeostasis – the balance between energy intake and energy expenditure – is the key principle underlying weight fluctuation in humans. Until the mid-1990s, very little was understood about the regulation of energy homeostasis. However, since the discovery of leptin in 1994, a network of hormonal mechanisms has been unraveled that links the control of appetite and energy expenditure to feeding behavior and adiposity. A host of signals produced in the digestive system, adipose tissue, and brain interact to modify behavior and physiological responses. Malfunctions in this delicately balanced system can lead to metabolic disorders and obesity.

Major Hormones Involved in Appetite Regulation

Leptin

Leptin was the first hormone shown to have a direct role in the regulation of food intake. It was quickly recognized as a circulating index of peripheral energy stores that could directly signal hypothalamic neurons in a position to effect appropriate adjustments in energy intake and expenditure. Long before its discovery in 1994, researchers had been using a homozygous mutant strain of mice called *obese* (*ob*/*ob*), which was characterized by excessive eating (hyperphagia) and obesity. When the *ob* gene was cloned, it was found to encode a circulating hormone, which was named leptin, secreted by adipocytes. The absence of leptin in *ob* mice was proposed to cause their characteristic phenotype since injections of recombinant leptin restored them to normal. Furthermore, the injection of leptin into wild-type mice resulted in fasting (hypophagia) and weight loss, leading to the proposal that leptin signals to the brain when excess fat is stored and promotes a reduction in food intake. In support of this conclusion, plasma leptin levels are usually proportional to total fat content, and the amount of leptin entering the central nervous system (CNS) is proportional to circulating levels.

The leptin receptor gene encodes a protein with six known isoforms, but only one of these (LepRb) has been shown to have an active intracellular signaling domain. Leptin receptors are richly expressed in the hypothalamus, which is the primary brain area responsible for energy homeostasis. This discovery correlated well with the results of earlier surgical experiments showing that lesions and electrical stimulations in different parts of the hypothalamus had a strong impact on food intake. Lesions in the ventromedial hypothalamus (VMH) and arcuate nucleus, or electrical stimulation of the lateral hypothalamus (LH), produce animals that tend toward morbid obesity and hyperphagia, whereas LH lesions or stimulation of the VMH produces unusually lean animals that are reluctant to eat. Such experiments formed the basis of Stellar's dual-center model, in which the LH acts as the feeding center and the VMH as the satiety center. Leptin receptors are expressed in all these regions, as well as in the dorsomedial hypothalamus. There are extensive reciprocal connections among these brain regions and with other parts of the brain, including the nucleus tractus solitarius (NTS) and various higher brain centers.

Since the discovery of leptin and other hormones that regulate appetite and food intake, the dual-center model has been progressively refined to include multiple hypothalamic nuclei and hormonal signaling pathways, as well as multiple roles for many of the regulators. For example, more-recent work with leptin has shown that its role is more complex than initially understood since it is also an indicator of general nutritional health and has a specific role in sexual development and the onset of fertility.

Research on leptin has focused on the arcuate nucleus, a structure within the mediobasal hypothalamus, adjacent to the third ventricle and the median eminence. This placement is probably significant since it makes the arcuate nucleus especially sensitive to leptin and other energy homeostasis regulators and indicators. The arcuate nucleus has the highest density of leptin receptors in the brain, and injections of leptin directly into this structure result in profound hypophagia, an effect which can be completely ameliorated by surgical ablation. Two groups of neurons have attracted particular interest, those containing neuropeptide Y (NPY) and those containing proopiomelanocortin (POMC). Centrally projecting NPY neurons, found in the most ventromedial part of the nucleus, project to the LH and paraventricular nucleus (PVN). Activation of these neurons stimulates the appetite, and they are acutely sensitive to leptin and ghrelin (see the section titled 'Ghrelin'). Centrally

projecting POMC neurons have widespread projections to many brain areas, including all hypothalamic nuclei. Activation of these neurons suppresses the appetite, and they are also responsive to leptin and ghrelin. There are direct projections between the NPY and POMC neurons.

Insulin

Insulin is a pancreatic hormone that regulates blood glucose levels by stimulating the conversion of glucose to glycogen. In addition to this role in carbohydrate metabolism, insulin suppresses the appetite, as shown by the effect of direct insulin injections into the brain. Insulin receptors are distributed in the hypothalamus and NTS and are particularly concentrated on NPY and POMC neurons in the arcuate nucleus, where their distribution is similar to that of the leptin receptor. This additional role for insulin may be related to its function in lipid homeostasis since, as well as promoting the storage of carbohydrates, insulin promotes the storage of fats. Overall levels of insulin correlate with the degree of adiposity in a manner similar to the correlation shown by leptin.

Ghrelin

Ghrelin was first identified as a growth hormone-releasing peptide (hence its name) but was later found to also have a role in energy homeostasis. Ghrelin is a circulating hormone and acts on the arcuate nucleus and hypothalamus in a manner antagonistic to leptin. Consistent with this role, injection of ghrelin either directly into the brain or into the blood results in increased food intake, and co-injection of ghrelin and leptin attenuates the effects of leptin. Ghrelin is produced by the empty stomach, specifically by P/D1 cells, and stimulates food intake. Levels of ghrelin fall when the stomach is full after a meal. However, it appears that ghrelin release is inhibited, not by stretching of the stomach, but by the presence of energy-rich food. Strategies to combat obesity such as gastric bypass work not only by limiting the capacity of the stomach but also by reducing the amount of circulating ghrelin. A treatment for obesity recently developed by the Scripps Research Institute is based on the use of antibodies against ghrelin to prevent its uptake by the CNS.

The ghrelin receptor is expressed in the arcuate nucleus and is found on both NYP and POMC neurons. It is also expressed strongly in the VMH.

NPY

Injections of NPY into the brain result in ravenous feeding, weight gain, and obesity, indicating that NPY is a potent appetite stimulator. In addition, NPY reduces sympathetic impulses to brown adipose tissue, thus reducing the metabolic rate and increasing energy efficiency. NPY is produced by neurons in the arcuate nucleus, which project strongly to the LH and PVN. The synthesis and release of NPY by these neurons is strictly regulated according to the energy balance, and hypothalamic neurons have NPY receptors, indicating a complex feedback system is in place to control this pivotal component of energy homeostasis.

Surprisingly, knockout mice lacking the ability to produce NPY appear to feed normally both under typical conditions and after a period of fasting, and they respond normally to chemical treatments that induce obesity. In *ob/ob* mice, NPY deficiency results in significant weight loss (although not a complete restoration to wild-type feeding behavior), which indicates that NPY may have a role in balancing the activity of leptin and may contribute to the obesity phenotype when leptin is absent. This is supported by the exaggerated effects of leptin injections in *npy/npy* mice.

Five major subtypes of NPY receptors have been described. Studies with antagonists suggest that NPY receptors 1 and 5 are the main receptors involved in the stimulation of food intake, although the putative autoreceptors 2 and 4 are also highly expressed in the hypothalamus, including both arcuate NPY and POMC neurons. Paradoxically, mice lacking the NPY1, NPY2, or NPY5 receptors have mild obesity phenotypes and varying degrees of hyperphagia. The failure to see a lean phenotype after the removal of a potent stimulator of feeding such as NPY suggests that the pathways involved in the control of food intake, like other hypothalamic systems, are sufficiently redundant to compensate for the loss of a major signaling system.

Alpha-Melanocyte-Stimulating Hormone

The *POMC* gene encodes a precursor polypeptide that gives rise to multiple peptides in different cell types, depending on the way it is cleaved. In the arcuate nucleus, the main products of pro-opiomelanocortin are two melanocyte-stimulating hormones (αMSH and γMSH) and the opioid β-endorphin. Only αMSH has a major role in energy homeostasis, and it acts as a potent inhibitor of food intake when injected into the brain. Furthermore, the effects of αMSH on sympathetic outflow to brown adipose tissue are opposite those of NPY; that is, it increases energy expenditure and results in net weight loss. Mice expressing an antagonist of αMSH are morbidly obese, and several genetic forms of human obesity have been linked to mutations in the *POMC* gene.

There are five melanocortin receptors, and it appears that the effects of αMSH are mediated mainly through MC4R. This has been deduced primarily by studies of *mc4r* knockout mice, which are severely hyperphagic and obese, but it is also apparent that approximately 5% of cases of childhood-onset obesity are associated with mutations in the human *MC4R* gene. It is likely that αMSH may be partly responsible for the loss of appetite seen in chronic illnesses such as cancer, since *mc4r* knockout mice with cancer appear to feed normally, as do mice expressing an antagonist of αMSH.

The specific role of MC4R, as opposed to the other receptors, is shown in experiments in which *mc4r* knockout mice are treated with αMSH agonists, which would increase signaling through the other receptors but still fail to influence feeding behavior. A mild obesity phenotype is seen in *mc3r* knockouts, but this is accompanied by feeding aversion, suggesting the underlying effect concerns energy utilization rather than food intake.

Major Control Circuits Involved in Appetite Regulation

The NPY and POMC neurons in the arcuate nucleus represent the primary integration site for circulating hormones and central signals involved in energy homeostasis (Figure 1). They carry a similar repertoire of receptors for the regulatory molecules discussed above, and they also project into many of the same brain regions, including the LH, PVN, and NTS, but they have antagonistic effects through the activities of NPY and αMSH. In addition to their nominal roles, NPY neurons produce agouti-related peptide (AGRP), which is an endogenous antagonist of the melanocortin receptors, and the inhibitory neurotransmitter γ-aminobutyric acid. POMC neurons also produce cocaine- and amphetamine-regulated transcript (CART).

Although NPY and POMC neurons have receptors for the major regulatory hormones, the effect of hormone signaling is distinct in each case. Leptin acts to

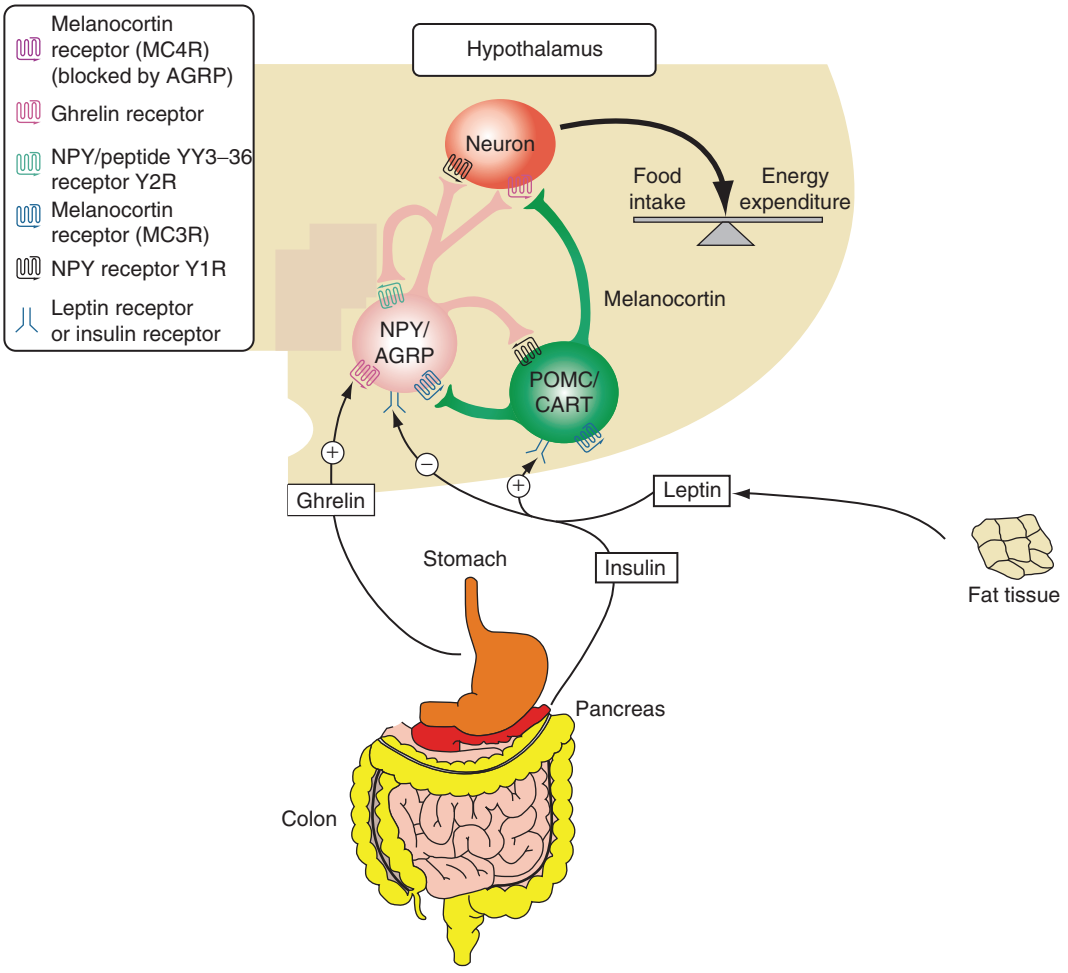

Figure 1 Simplified representation of the core regulatory circuit in energy homeostasis. AGRP, agouti-related peptide; CART, cocaine- and amphetamine-regulated transcript; NPY, neuropeptide Y; POMC, pro-opiomelanocortin.

reduce food intake by stimulating POMC neurons (thus causing the release of αMSH) while inhibiting NPY neurons (and limiting the release of NPY). Conversely, ghrelin treatment stimulates the production of NPY and AGRP but has no impact on *POMC* messenger RNA (mRNA) levels. This is thought to reflect different intracellular signaling responses in each cell type. In the case of leptin signaling, binding to the leptin receptor activates the immediate early gene c-*fos* in POMC neurons, but not in NPY neurons. The activation of c-*fos* generally correlates with neuronal excitation. The ability of leptin to increase the activity of POMC neurons and decrease the activity of NPY neurons has been confirmed directly through electrophysiological recordings.

There are also reciprocal connections between the NPY and POMC neurons, indicating they may directly modulate each other's ability to release peptides. However, there is a further control mechanism mediated through the release of AGRP by NPY neurons. Like NPY, AGRP stimulates food intake and is thought to do so by blocking the effects of melanocortins. Thus, the release of both NPY and AGRP by NPY neurons has two simultaneous effects: a direct effect on downstream targets through the activity of NPY and indirect effects through the inhibition of αMSH and any other melanocyte-stimulating hormones released by POMC neurons.

The PVN is a major candidate downstream target of the POMC and NPY neurons since both NPY and αMSH have been shown to influence food intake following direct injections into this structure, and thyrotropin-releasing hormone neurons located therein express the pivotal αMSH receptor, MC4R. It is possible that the inhibitory effects of αMSH on food intake are mediated by direct signaling to these neurons and the stimulatory effects of leptin reflect at least in part the blocking of melanocortin activity via AGRP. Regulation of thyrotropin-releasing hormone and other PVN neurons would regulate both the pituitary and the thyroid axes, thereby providing a mechanism by which leptin and the melanocortin system may modulate both food intake and energy homeostasis. Other potential downstream targets of the POMC and the NPY neurons are discussed in the next section.

Downstream Factors Involved in Appetite Regulation

Factors That Promote Food Intake

A number of additional hormones and regulators have been shown to promote feeding when administered to rodents, and research over the past few years has allowed many of these to be slotted into the primary control circuit involving leptin, ghrelin, NPY, and αMSH. Some of these factors are likely to be directly involved in energy homeostasis, while others have peripheral roles, such as general roles in arousal or in mediating the rewarding aspects of food intake.

One of the principal downstream targets of NPY is the release of melanin-concentrating hormone (MCH), a potent stimulator of feeding when injected directly into the brain. *MCH* mRNA levels increase during fasting and decrease after meals (and after the administration of leptin). The hormone is found exclusively in so-called MCH neurons in the LH, which project to other hypothalamic nuclei as well as to the NTS and cortex. These neurons are likely to be targets of the NPY and POMC neurons in the arcuate nucleus, as discussed earlier, and possibly act as a relay system in the NPY circuit. Tampering with gene expression levels artificially has predictable effects: transgenic mice overexpressing the *MCH* gene are hyperphagic and obese, while *mch*/*mch* knockouts are hypophagic and lean. This contrasts with the phenotype of *npy*/*npy* mice and suggests that the functional redundancy exhibited in the NPY system is not present in its downstream target, MCH.

Another possible target for NPY is the release of galanin, a 30-amino-acid neuropeptide, from discrete populations of neurons in the arcuate nucleus, dorsomedial hypothalamus, and PVN. Galanin neurons project to multiple areas of the hypothalamus and other brain regions involved in food intake, and they have synaptic links with both NPY and POMC neurons. Galanin levels increase during fasting, and the injection of the mature galanin peptide into the PVN, LH, VMH, and NTS stimulates food intake in rodents, although not to the levels seen with NPY injections. The precise role of galanin remains unclear since *gal*/*gal* knockout mice are neither obese nor unusually lean.

Finally, it is likely that some of the effects of leptin are mediated by endogenous cannabinoids, at least two of which, anandamide and 2-arachidonoylglycerol, are normally present in the hypothalamus. Supportive evidence includes the increased levels of endogenous cannabinoids in the hypothalami of *ob*/*ob* mice and the reduction in levels that follows administration of leptin. Also, injections of cannabinoid agonists stimulate feeding, whereas mice lacking the CB1 cannabinoid receptor are hypophagic after fasting (although without fasting they eat normally). Research has focused on the CB1 receptor because prolonged administration

of a CB1 antagonist also results in hypophagia and leanness. Rimonabant (Acomplia) is a cannabinoid receptor antagonist which has been prescribed as an appetite suppressant in the countries of the European Union since 2006.

Factors That Inhibit Food Intake

A number of additional factors have also been described which inhibit feeding and generate a negative energy balance, and again it is possible to slot many into the central regulatory circuit controlling energy homeostasis. CART is a neuropeptide co-expressed with αMSH in POMC neurons of the arcuate nucleus. Like αMSH, CART can inhibit food intake when injected directly into the brain. CART expression is also dependent on leptin, which marks it out as an important component of the energy homeostasis network. However, mice deficient in the expression of this peptide show no significant changes in feeding behavior or body weight, indicating that whatever role CART plays in the system, it can be replaced by a redundant component. CART is expressed in several additional sets of neurons that do not express POMC, so it is possible these neurons have secondary roles in appetite control.

While CART is expressed in the CNS, cholecystokinin (CCK) is a satiety factor expressed primarily in the gut. CCK is secreted by the gut after a meal, and its release stimulates CCK receptors located on the vagal nerve, which transmits the satiety signal through the NTS. There are two classes of CCK receptor, and studies using specific antagonists show that the satiety signal is transmitted primarily through the class A (CCKA) receptor. Such antagonists inhibit feeding after a period of fasting. Injections of CCK result in sudden but short-lived food aversion, which is enhanced in the presence of leptin. Additional studies have shown that CCK and its receptors are also produced in the brain, indicating a central role through direct action on neurons in the NTS and hypothalamus. The precise role of CCK is clouded by different results from studies in different animals. For example, CCKA knockout mice show no obvious feeding phenotype, whereas rats with the orthologous gene knocked out show mild obesity. Another gut peptide with activity similar to CCK's is bombesin, and mice lacking the receptor BBR3 are obese. Finally, enterostatin is a peptide, produced by the pancreas and gastrointestinal tract, which reduces insulin secretion, increases sympathetic outflow to brown adipose tissue, and stimulates adrenal corticosteroid secretion. The overall impact of enterostatin secretion is a sensation of fullness leading to a selective reduction in fat intake. It is not yet clear how this peptide interacts with the other components of the energy homeostasis system.

Several other peptides are known to inhibit food intake following direct injection and to be regulated by leptin but to have little effect on feeding in rodents with the corresponding genes knocked out. Corticotropin-releasing hormone (CRH), for example, is produced mainly by neurons in the PVN and has a number of diverse effects, including the release of adrenocorticotropic hormone (another product derived from POMC) from the pituitary gland. CRH levels are regulated by leptin, and injections of CRH into the brain reduce food intake, even after induction by NPY, and stimulate thermogenesis via sympathetic pathways, but mice lacking the corresponding receptors, CRHR1 and CRHR2, have normal weight and feeding behavior. Urocortin is a related peptide which has a more potent anorexigenic effect than CRH's when injected, but because it signals through the same receptors as CRH and there is no impact on feeding when these receptors are missing, its precise role is unclear. Another potent anorectic is glucagon-like peptide 1 (GLP-1), a gastrointestinal peptide that is released in response to food intake. GLP-1 plays an important role in glucose homeostasis by inhibiting the secretion of glucagon and augmenting glucose-induced insulin secretion. In the context of this article, GLP-1 is also proposed to act as a satiety factor. Consistent with this hypothesis, peripheral administration of GLP-1 inhibits food intake. In addition to being produced in the intestine, GLP-1 is expressed at hypothalamic sites, the caudal portion of the NTS, and the forebrain. Central administration of GLP-1 inhibits food intake through actions in the hypothalamus, including the PVN. Conversely, GLP-1 antagonists stimulate feeding in satiated rats. GLP-1 antagonists also attenuate the effects of both leptin and CCK to inhibit food intake. Leptin has been shown to increase hypothalamic GLP-1 levels and to activate GLP-1 neurons in the NTS, providing further evidence of a role for GLP-1 in food intake and energy homeostasis. However, mice lacking the GLP-1 receptor do not show any feeding abnormalities.

Summary

The regulation of appetite and food intake is a complex process. It involves a coordinated response to many orexigenic and anorexigenic factors in multiple brain regions. Significant progress has occurred during the past decade, including the discovery of the adipostatic hormone leptin and some of the pathways required for its actions. Leptin produces its effects in part by a coordinated action to simultaneously

activate POMC and inhibit NPY neurons in the arcuate nucleus to alter food intake and energy homeostasis. Downstream targets of the NPY and POMC neurons include specific neurons in the PVN, LH, NTS, and other brain regions. However, much is still unknown, and it is likely that some critical appetite factors remain to be discovered. Furthermore, as is the case with other functions regulated by the hypothalamus, there appears to be some redundancy between these appetite factors.

See also: Appetitive Systems: Amygdala and Striatum; Eating Disorders; Energy Homeostasis: Adiposity Signals; Energy Homeostasis: Hypothalamic Development; Energy Homeostasis: Paraventricular Nucleus (PVN) System; Energy Homeostasis: Thermoregulation; Food and Water Intake: Regulation; Gastrointestinal Signals: Stimulation; Gastrointestinal Signals: Satiety; Neuroendocrine Control of Energy Balance (Central Circuits/Mechanisms).

Further Reading

Balthasar N (2006) Genetic dissection of neuronal pathways controlling energy homeostasis. *Obesity* 14: 222S–227S.

Butler AA and Cone RD (2001) Knockout models resulting in the development of obesity. *Trends in Genetics* 17: S50–S54.

Dham S and Banerji MA (2006) The brain-gut axis in regulation of appetite and obesity. *Pediatric Endocrinology Reviews* 3: 544–554.

Friedman JM (2000) Obesity in the new millennium. *Nature* 404: 632–634.

Halford JC (2001) Pharmacology of appetite suppression: Implication for the treatment of obesity. *Current Drug Targets* 2: 353–370.

Inui A (2001) Ghrelin: An orexigenic and somatotrophic signal from the stomach. *Nature Reviews Neuroscience* 2: 551–560.

Klok MD, Jakobsdottir S, and Drent ML (2007) The role of leptin and ghrelin in the regulation of food intake and body weight in humans: A review. *Obesity Reviews* 8: 21–34.

Murphy KG and Bloom SR (2006) Gut hormones and the regulation of energy homeostasis. *Nature* 444: 854–859.

Spiegelman BM and Flier JS (2001) Obesity and the regulation of energy balance. *Cell* 104: 531–543.

Hormones and Behavior

N S R Sanderson and D Crews, University of Texas at Austin, Austin, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

There are some stimuli encountered by animals to which the most appropriate response is invariant. If a large predator appears, seeking refuge almost always benefits a foraging rodent. Then there are other stimuli and situations to which the most appropriate response is dependent on the internal state of the individual. A hungry animal, for example, may consume food that a well-fed animal would ignore. There are many such internal or physiological variables that influence animals' behavior, and some change rapidly, whereas others change more slowly. Using the example of feeding, in addition to the difference in behavior between an animal that has just fed and an animal that has not fed for several hours, two animals with equally full stomachs may behave differently if one is fat and the other is lean. Furthermore, some of these physiological factors change unpredictably, whereas others change cyclically across a range of timescales from minutes to years. Such cyclic variability is typical of the reproductive states of many animals – for example, the ovarian cycle of most vertebrates.

For many of these phenomena, changes in a particular physiological parameter are associated with changes in the levels of one or more signaling molecules produced by the body. For several kinds of behavior, described in this article, it has been possible to show that a particular signaling molecule or hormone is the critical agent inducing the behavior appropriate for the physiological state associated with that hormone. Hormones, in other words, can be thought of as messengers that inform the brain of the state of the rest of the organism and enable the appropriate behavior. This article presents an overview of current knowledge about the relationships between environment, hormones, and the brain. Then, for three particular kinds of behavior (i.e., copulatory behavior, affiliative behavior, and feeding behavior) it describes in more detail the endocrine and neural systems that are thought to underlie these relationships.

Research

To say that a hormone influences a behavior is to say that an organism in one hormonal state responds to a given stimulus with one response, whereas the same organism in a different hormonal state exhibits a different response to the same stimulus. The elucidation of this relationship involves the following questions:

- What are the hormonal states involved, in terms of chemical identity, concentrations, and timing of hormonal exposure?
- What is the stimulus?
- What are the alternative behavioral responses?
- Which cells in the nervous system are involved in conveying information about the stimulus to the effectors that mediate the responses?
- On which of these cells does the hormone(s) have its effect?
- What kinds of receptors in these cells bind the hormone?
- Following hormone receptor binding, what kinds of changes occur in each cell, in terms of second-messenger activation, gene expression, and so on?
- How do these changes in each cell produce changes in cell-to-cell neurotransmission so that the stimulus–response relationship is altered?

Hormones

Historically, for a substance to be considered a hormone, it had to satisfy three criteria: synthesis by a specified kind of cell or tissue, release into and transport by the blood, and action on a specified target cell or tissue type(s). With increasing appreciation of physiological signaling systems, the second of these criteria has been become more complicated, and a system of adjectives has been developed to describe signaling molecules that act at various distances. Molecules that act within the cell that produces them are called intracrine signals, those that are secreted but act on the secreting cell are called autocrine, and those that act on adjacent cells are called paracrine. This article deals principally with endocrine systems – that is, those involving signaling molecules secreted from the producing cell into the intracellular fluid and exerting a physiological effect on other cells. However, several kinds of signaling molecules satisfy this definition that are not thought of as hormones. In particular, signaling molecules such as dopamine that are released by neurons and influence large numbers of other neurons are known as neuromodulators and not as hormones. Furthermore, some substances such as the cytokines modulate interactions between various components of the immune system and satisfy all three criteria perfectly well but are not called hormones, perhaps because

they are the historical purview of immunologists rather than endocrinologists.

Most of the well-studied hormones belong to one of four chemical classes, and as a result of their biochemical properties, members of a particular class tend to serve similar functions and engage similar transduction machinery in their target cells.

Steroids

Steroids, such as the gonadal hormones estrogen, progesterone, and testosterone, and the adrenal steroids corticosterone and aldosterone are synthesized from cholesterol and have a four-ring structure similar to the estrogen molecule shown in Figure 1. Their hydrocarbon-dominated structure makes their solubility in water low and that in lipids high. Consequently, they pass easily though cell membranes, and a steroid secreted in one part of the body will gain access to all tissues. The concentration of hormone reaching target tissues, however, is limited by the low carrying capacity of the blood, although for many steroids, this is augmented by specialized carrying proteins. Many of the effects of steroid hormones on behavior are mediated by proteins belonging to the nuclear receptor family that dimerize on binding ligand and then associate with the DNA of regulatory regions of genes to influence gene transcription. There is also accumulating evidence that steroids sometimes interact with membrane-associated receptors, and such effects are generally more rapid than the hours or days during which the nuclear receptors manifest their influence.

Prostaglandins

Prostaglandins are signaling molecules derived from fatty acids and are common mediators of physiological processes. Evidence of direct involvement in vertebrate behavioral regulation is not abundant, but, for example, prostaglandin E has been implicated as a key component of the developmental process of masculinization of the brain and subsequent behavior.

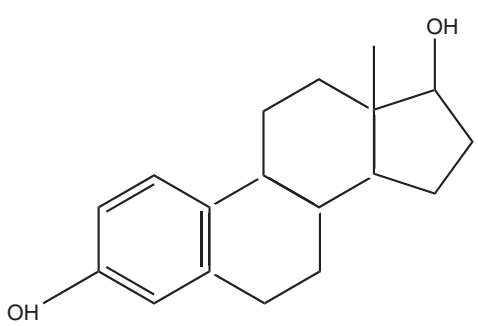

Figure 1 17β estradiol. Estradiol and other sex steroids share the basic structure of three six-carbon rings and one five-carbon ring, and they are interconverted by modifications of the substituent groups. Estradiol, for example, is formed from testosterone by the aromatization of the first six-carbon ring and the reduction of the ketone group on this ring to the hydroxyl shown here.

Monoamines

Monoamines, such as epinephrine, serotonin, and melatonin, are small molecules derived from single amino acids. They generally exert their effects through membrane-bound receptors with an extracellular ligand-binding site, often a G-protein-coupled receptor (GPCR), whose action on ligand binding is often to influence the production of a second messengers such as cyclic AMP.

Peptides

Peptide hormones are polymers of small numbers of amino acids (from fewer than ten to a few hundred); in other words, they are small proteins. Like monoamines, they generally utilize membrane-bound receptors, often GPCRs. The complexity of peptide hormones means that they often show variation in exact structure according to the organism examined. Steroids and monoamines are chemically identical from species to species (e.g., testosterone in a fish is identical in structure to testosterone in a human), although there are species differences in the nature and pattern of synthesis and release of the hormone (e.g., in many male fish 11-ketotestosterone is the dominant circulating androgen and is produced by steroidogenic enzymes not present in humans).

Behaviors Influenced by Hormones

Sexual Behaviors

There are three very important roles for hormones in modulating reproductive behavior. First, males and females have different roles in reproduction, and gonadal sex steroids tell the brain whether male-typical or female-typical behavior is appropriate. Second, animals do not reproduce throughout their lives but, rather, have a reproductively quiescent period before puberty, and, according to the species, sometimes breeding and nonbreeding seasons; hormones serve to encode reproductive status of the animal. Third, in females, sexual behavior is often restricted to narrow windows of the ovarian cycle, which are also associated with particular hormonal states. In general, testicular androgens enable the expression of male-typical behaviors such as competitive aggression, mounting, and intromission, whereas ovarian hormones, particularly estrogen, enable the display of female-typical sexual behaviors such as estrous and maternal behaviors. However, the hormones do not

act as simple on/off switches but, rather, are players in a complicated program of interaction between the developing animal and the environment, acting at multiple time points in both the brain and the periphery to coordinate the critical function of reproduction.

Development A central concept in the behavioral neuroendocrinology of reproduction is the organization–activation hypothesis. This idea was originally formulated to describe the development of the peripheral reproductive organs. Mammalian embryos are initially bipotential, possessing gonadal and genital primordia of both sexes. Working with rabbit embryos *in utero* in the 1940s and early 1950s, Alfred Jost recognized that early in development, castration of a male embryo would result in the development of the female-appropriate genital tract and the disintegration of the male genital tract, whereas a similar operation on an embryonic female did not affect the development of the female genital primordia or the disintegration of the male genital tract. This organization (i.e., the irreversible selection of which sexual structures are retained and which are lost) is irreversible and is restricted to a critical developmental time period. Later, once puberty occurs and levels of sex steroids increase again, the sexual structures develop further from their prepubescent state to their reproductively functional state. This activation of the previously organized structures is generally reversible, and in seasonally breeding animals it is commonly executed and reversed each breeding season.

Experiments conducted by William C Young's group in the 1950s suggested that the same kind of relationship between embryonic and adult effects of hormones might hold in the brain. Female guinea pigs that had been exposed to androgens *in utero* were much more likely to express male-like copulatory behavior when exposed to androgens as adults, and they were also less likely to express female-typical receptive behavior when treated with estrogen and progesterone. These two effects of prenatal androgenization, known as masculinization and defeminization, together constitute the process of behavioral organization and are experimentally separable by the timing of the critical developmental window and the nature of the hormones involved. During normal development, it appears that both components are mediated by circulating testosterone, although according to the species involved, the local conversion (i.e., at the site of action in the brain) of testosterone to estrogen, or aromatization, may be more or less critical. In rats, for example, administering an aromatase inhibitor prenatally can block masculinization. A schematic depiction of the development of sexual dimorphisms in anatomy, physiology, and behavior related to reproduction is presented in Figure 2. In addition to the behavioral motor patterns of copulation, such as female-typical receptivity and male mounting and intromission, other aspects of sexually dimorphic behavior can be subject to the organizational effects of embryonic hormone exposure. Examples include sex-typical play patterns, aggressiveness, and adult preference for sexual partners of the opposite sex.

Site of action It is presumably not a coincidence that the same hormones that control the development of reproductive organs also control the expression of behaviors utilizing those organs. In vertebrates of both sexes, the gonadal steroids that regulate the production of gametes also enable mating behavior, a kind of pleiotropy with obvious adaptive significance. More subtly, a given hormone will often have many sites of action within the brain for two reasons. First, expression of one kind of behavior may be associated with suppression of another kind of behavior. For example, the rise in progesterone associated with early pregnancy in female rabbits not only induces nest-building behavior but also suppresses mating behavior. Second, behavioral responses involve many different areas of the brain. A large number of experiments have identified specific nuclei of the limbic forebrain to be critically involved in the display of sexual behavior. Furthermore, various methods reveal these to be reciprocally interconnected, to contain sex steroid hormone receptors, and to be sexually dimorphic in their volume and synaptic organization as a consequence of the nature and frequency of sex steroid hormones secreted perinatally. Importantly, these properties appear to be evolutionarily conserved. Furthermore, these limbic nuclei form integrated circuits that are sensitive to both somatosensory and hormonal stimuli with overlapping functions that subserve all sex steroid hormone-modulated social and reproductive behaviors. In this manner, social and reproductive behaviors emerge from the activity of a unitary neuroanatomical framework rather than being the product of activity of a single brain area.

Interaction with the environment With increases in our understanding of the genetic determination of gonadal sex, it is becoming possible to map out, albeit with gaps, a pathway from the sex chromosomes an individual inherits, through the development of gonads, the secretion of sex-typical patterns of hormones, the sex-typical development of the brain, to the expression of sex-typical behavior. However, an individual vertebrate is not a hormone-driven automaton, blindly executing whatever behavioral program its gonads dictate, any more than the shape of its

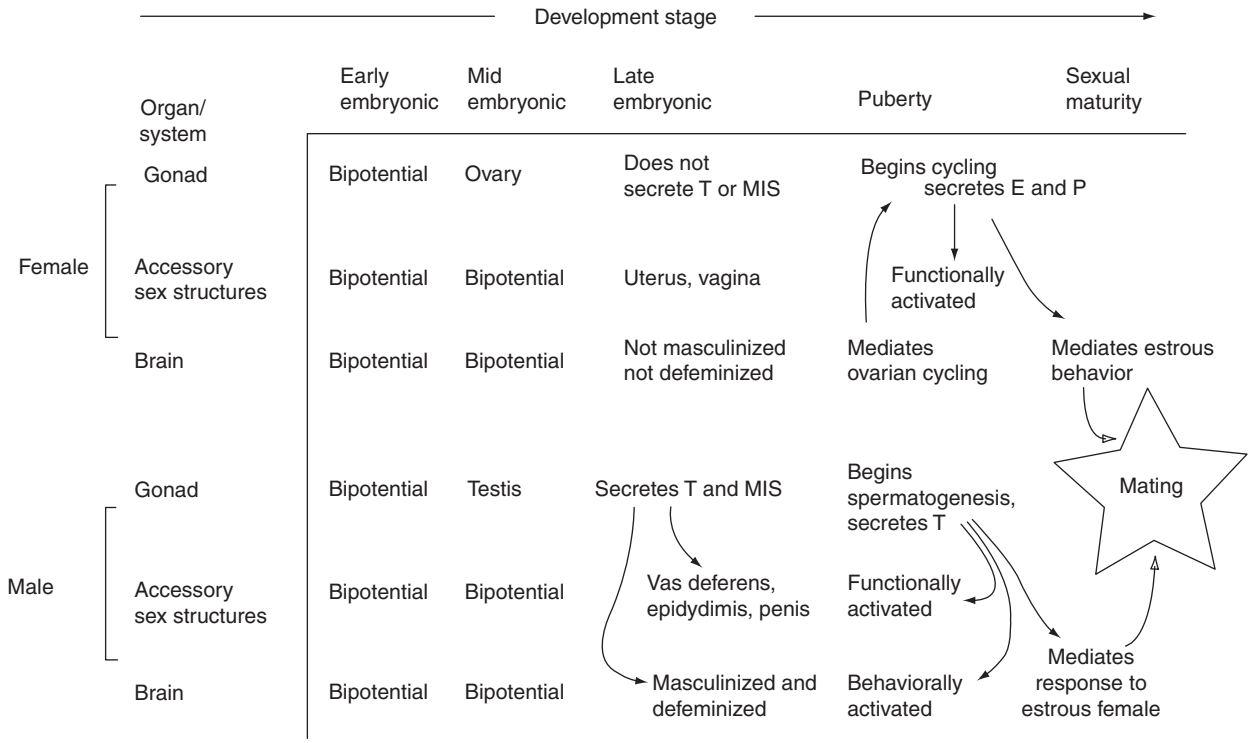

Figure 2 Sex steroids act in various ways throughout development to coordinate physiology and behavior. This figure is a schematic depiction of important events in the development of female and male mammals, illustrating the interaction between the three organ systems (gonad, peripheral sex organs, and brain) and ultimately between the two sexes. In mammals, the sex is fixed at conception, and a molecular cascade will determine the sexual fate of the gonad. Secretions of the fetal gonad are then principally responsible for determining the sexual phenotype of the other two systems. This includes the accessory sex structures where testosterone maintains the Wolffian duct and enables them to develop into the epididymis and vas deferens, whereas Müllerian inhibiting substance (MIS), a protein signaling molecule secreted by the testis, induces the disintegration of the Müllerian ducts, which otherwise (i.e., in the developing female) develop into the uterus, cervix, and parts of the vagina. Fetal testosterone is also responsible for masculinizing (i.e., enabling future masculine) and defeminizing (i.e., disabling future feminine) behavior by its actions on the developing brain. At puberty, the body and brain thus organized during fetal development are activated by increasing levels of sex hormones, and animals of either sex become physiologically capable of reproduction and behaviorally able to mate. Arrows signify the actions of secreted molecules. E, estrogen; P, progesterone; T, testosterone.

development is determined entirely by its genome. Rather, each individual develops and behaves in continual interaction with other individuals and with other aspects of its environment. Such environmental influences on the development of the hormone–behavior relationship include hormone exposure *in utero* from mother and siblings, postnatal interaction with mother and siblings, the type and availability of food, and sexual and aggressive interactions with other individuals in adulthood, all of which can act to increase, decrease, or qualitatively change the expression of sex-typical as well as sex-atypical behaviors.

Copulatory behavior Mammalian male sexual behavior is dependent on testicular androgens so that castration eliminates the behavior in the majority of individuals, and administration of exogenous testosterone will reinstate it. In naturalistic settings, the expression of male sexual behavior is generally determined by the sexual behavior of females. Female sexual behavior is largely under the control of ovarian steroids and is therefore tightly linked to the ovarian cycle, of which a schematic is presented in Figure 3. In the rat, the cycle is usually 4 days long, with estrogen levels rising continuously until ovulation on the afternoon of the fourth day and then falling back to baseline, whereas progesterone levels remain low except on the day of ovulation. The periovulatory increase in progesterone following a period of sustained estrogen exposure is the hormonal antecedent to behavioral estrus, including increased attractivity, receptivity, and proceptivity, and can be mimicked in ovariectomized rats by an injection of estrogen, followed 48 h later by an injection of progesterone. In the wild, this cycle is unlikely to persist for long since behavioral estrus will normally involve mating and pregnancy, and the sustained high levels of progesterone

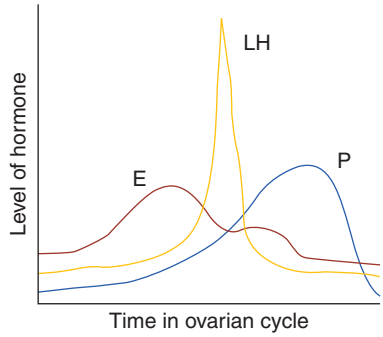

Figure 3 Hormonal changes across the ovarian cycle. Levels of three hormones are shown, scaled arbitrarily on the vertical axis and plotted against the time of the ovarian cycle. The horizontal axis would thus represent 4 days for a rat, or 28 days for a human. Luteinizing hormone (LH) is shown because it is a reliable marker of ovulation, rather than for its behavioral importance. In rodents, the rise in estrogen (E) levels before ovulation followed by a rise in progesterone (P) levels are critical determinants of estrous behavior.

originating from the corpora lutea suppress further ovarian cycles until gestation is complete. Other species have cycles of different lengths, but the general pattern of estrogen levels rising gradually before ovulation and then progesterone levels rising during and after ovulation is conserved. There are also some species such as cats, in which there is no regular ovarian cycle but, rather, ovulation is induced by copulation.

In 1976, Frank A Beach proposed that the sexual behavior of female mammals could generally be thought of as including three components, namely attractivity, proceptivity, and receptivity. The attractivity of a female (many aspects of which are obviously nonbehavioral) represents the degree to which a male will seek and elicit copulation with her. Proceptivity is the degree to which she will actively seek out a sexual partner and attempt to initiate copulation. Receptivity is the degree to which a female will allow the courtship and intromission of a male that attempts to copulate with her. These three components are still extremely useful heuristically in describing the copulatory sequence of mammals.

In rats, attractivity is assessed largely through olfactory and pheromonal cues (Figure 4(a)). Little is known about the neural mechanisms by which females assess a male's value as a mate, but the neuroanatomy underlying male processing of female pheromonal signals is known in some detail (Figure 5(a)). Volatile odorants are detected at the olfactory epithelium and larger, less volatile pheromones by the vomeronasal organ. These sensory structures project respectively to the main and accessory olfactory bulbs and then via the amygdala (particularly medial division) and the bed nucleus of the stria terminalis to the medial preoptic area (MPOA). The MPOA plays a

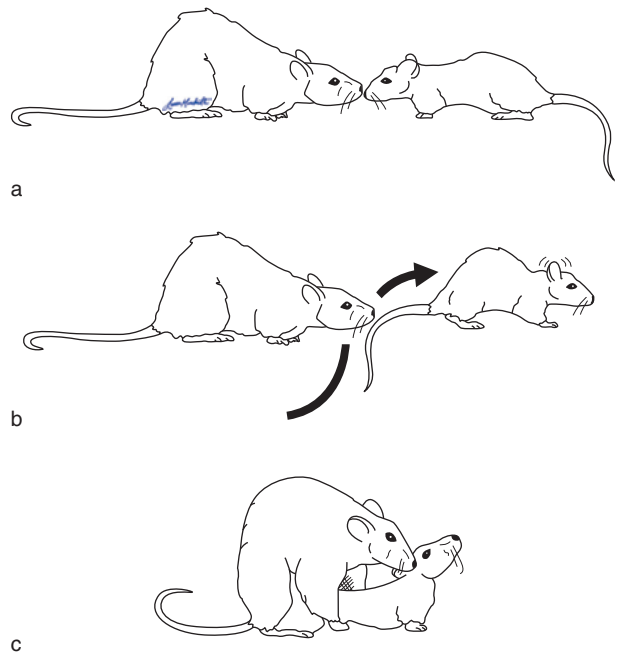

Figure 4 Characteristic behaviors in the rat mating sequence. Rat social interactions of all kinds, including sexual, often involve mutual olfactory examination, particularly of the facial area (a). When the interaction involves an estrous female and a suitable mate, she will express a suite of proceptive behaviors that signal her sexual status to the male, including rapid approach and withdrawal, hopping and darting, and rapidly moving her ears (b). A sexually active male will follow and mount her (c). If receptive, she will express the flattened posture of lordosis, with head and rump elevated and back depressed, enabling him to intromit, thrust, and ejaculate. Illustrations by Lauren Munchrath.

particularly important role in integrating male-typical copulatory behavior, especially the hormonal gating thereof. Although prolonged systemic exposure to androgens is necessary for the development of sexual structures such as the penis, it seems that expression of copulatory behavior as an adult is simply determined by exposure of a small number of brain areas to hormone, particularly the MPOA.

An estrous female will solicit the male's attention with proceptive behaviors such as approaching him and rapidly withdrawing with a characteristic dart-hopping movement, ear wiggling, and other behaviors (Figure 4(b)). Comparatively little effort has been made to elucidate the neuroanatomical substrates of this behavior as distinct from those subserving the receptive behavior that follows.

A sexually active male will follow a proceptive female and attempt to mount her. The stimulation from his snout and forepaws on her rump and flank induces the estrogen- and progesterone-dependent reflexive receptive posture known as lordosis shown in Figure 4(c), and whose neural substrates are schematized in Figure 5(b). The essential reflex arc flows

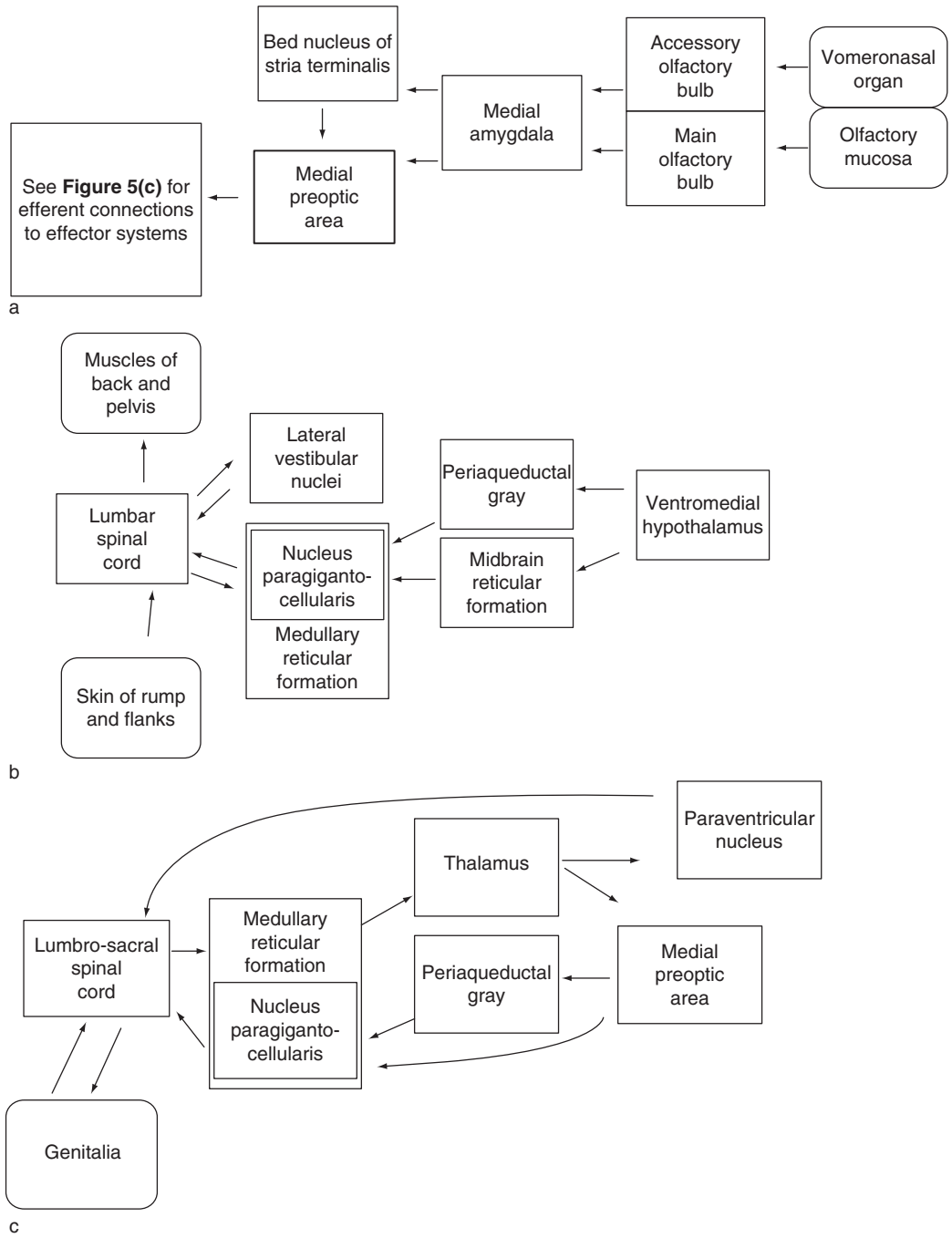

Figure 5 Neuroanatomical pathways mediating copulatory behavior. Brain nuclei known to be important in the expression of the various behaviors are depicted as square boxes, and the projections between them (of which some are excitatory and some inhibitory) are shown as arrows. Sensory and effector organs are shown as rounded boxes. Brain nuclei critical for hormonal control over behavior are shown as bold boxes. The neuroanatomy involved in processing sexually relevant chemosensory information has been well studied in the male rodent (a) and involves the pathway from the main and accessory olfactory bulbs to the amygdala, particularly the medial division, and thence via the bed nucleus of the stria terminalis and an alternative route via the ventral amygdalofugal pathway to the medial preoptic area. Exactly what happens to this information once it reaches the medial preoptic area is unknown, but it is presumed to result in the decision to attempt to mount. Once the mount is established, events in both the female and the male involve reflex arcs mediated by well-characterized neural circuits, shown by bold arrows. (b) Lordosis in the female involves the transfer of sensory information from the male's mounting and thrusting movements to the lumbar spinal cord, whence it ascends to the brain stem motor nuclei responsible for integrating the muscular motor pattern of lordosis. Descending control over this reflex arc is exerted by the ventromedial hypothalamus via the periaqueductal gray and the midbrain reticular formation. The medial preoptic area plays an inhibitory role. (c) Erection and ejaculation in the male are spinal reflexes, albeit subject to descending influences from higher in the brain, whereas other motor patterns of male-typical copulation are mediated by brain stem motor areas. The periaqueductal gray exerts an inhibitory influence on these lower copulatory reflexes, and a major role played by the medial preoptic area is probably disinhibition achieved by inhibiting the periaqueductal gray.

from the sensory receptors on the skin via the pudendal nerve to the lumbar spinal cord. Then it is relayed to the brain stem motor centers of the lateral vestibular nuclei and the nucleus paragigantocellularis of the medullary reticular formation. These motor centers project back to the lumbar spine and orchestrate the motor pattern of lordosis via their control over the musculature of back and pelvis. Hormonal control of this behavior is achieved by higher centers in the brain, particularly the ventromedial hypothalamus, which influences the nucleus paragigantocellularis via the periaqueductal gray. A similar system of nuclei mediates the motor and other physiological responses of the mounting male (**Figures 4(c) and 5(c)**). Sensory information from the penis is relayed via the pudendal nerve to the sacral and lumbar spinal cord. Some important components of the male sexual response, such as erection and ejaculation, are spinal reflexes and are observed even in rats spinally transected above the thoracic level. Sensory information, however, is conveyed to the medullary reticular formation, whose nucleus paragigantocellularis integrates reflexive motor patterns such as thrusting, and then via the thalamus to higher areas in the brain. Descending signals from the MPOA reach the nucleus paragigantocellularis directly or via the periaqueductal gray. Other aspects of female sexual behavior, such as mate choice, the rewarding properties of copulation in various different contexts, copulation for other reasons than reproduction, and an individual's preference for one particular mate over others, are also beginning to be studied.

Aggression The ecology of sexual aggression is complex, with a number of factors influencing which individuals are aggressive to which others under which circumstances in which species. This complexity in the nature of the behavior is reflected in the relationship between the behavior and its hormonal antecedents, and general principles are difficult to extract. However, with respect to the endocrinology of male reproductive aggression, observations from a number of species can be reconciled in a conceptual framework known as the 'challenge hypothesis,' originally elaborated by John Wingfield and colleagues to explain some of the available observations concerning testosterone and aggression in songbirds. This hypothesis posits that testosterone levels are low during nonreproductive periods, rise to intermediate levels during breeding to support reproductive physiology such as spermatogenesis and copulatory behavior, and increase to especially high levels during periods characterized by frequent agonistic interactions with other males, sometimes apparently as a result of such agonistic interactions.

Again, possibly because of the complexity of the phenomenon, the neuroanatomical substrates of aggression have been less thoroughly elucidated than those subserving copulatory behavior, but some kinds of aggression seem to involve the lateral septum and the nucleus accumbens, in addition to the amygdala and various subdivisions of the hypothalamus. These areas express androgen and/or estrogen receptors and also various types of serotonin receptor. Serotonin appears to occupy a central role in the control of aggressive behavior, with reduced levels of serotonin receptor activation being associated with increased aggressiveness in a range of contexts in a range of species. This is not a completely valid generalization, however, since maternal aggression is associated with increased serotonin levels.

Parental behavior Parental behavior obviously includes both maternal and paternal behaviors, but since paternal behavior is not common other than in birds and certain fish taxa, this discussion is largely limited to females. Once again, rats have been the main focus of research, and important kinds of maternal behavior in this species include nest building, licking and grooming, nursing, pup retrieval, and maternal aggression. These behaviors are exhibited at different points in the reproductive cycle (or at least exhibited at different frequencies), and their expression is partly dependent on the endocrine events associated with pregnancy, lactation, and birth. The end of pregnancy is characterized in rats by high estrogen levels, plunging progesterone levels, and a sharp spike in prolactin levels, whereas the event of parturition is associated with a peak of oxytocin secretion. All of these hormonal parameters appear to be involved in the onset of maternal behavior. Expression and maintenance of the behavior is not only partly controlled by hormones but also largely dependent on stimulation from the young so that nursing female rodents are not exceptionally aggressive except in the presence of their pups.

Pair bonding In some species, males and females form stable reproductive pairs in order to raise young. Oddly, the neuroendocrine mechanisms underlying this behavioral phenomenon have received little attention in birds, in which the behavioral pattern is common, and instead have been most extensively investigated in microtine rodents. These studies have implicated corticosteroids and the peptide hormones oxytocin and vasopressin in mediating pair-bond formation in these animals. In the prairie vole, socially naive animals will form monogamous pairs with the partners with whom they first mate. The genesis of this pair bond is

dependent on copulation-induced oxytocin release in females and on a combination of vasopressin and corticosterone in males.

Other Hormonally Influenced Behavior

Feeding is important, but there are times, such as when an animal has a full stomach, that it is inappropriate or, when the animal is fat, that it is less important than other activities such as reproduction. Accordingly, vertebrates have evolved a network of signaling systems that keeps the brain informed about the animal's short-term and long-term state of nutrition and enables appropriate feeding behavior. In addition to direct neural pathways to the brain, the gastrointestinal tract produces a range of peptide hormones that encode information about the quantity and quality of food being ingested and processed. The majority of these hormones, such as cholecystokinin and glucagon-like peptide 1, appear to be satiety inducing, rising after feeding and reducing food intake when administered exogenously. An interesting exception is ghrelin, which appears to peak before (i.e., in anticipation of) feeding and increases food intake. Just as other behaviorally important hormones often mediate nonbehavioral aspects of the processes with which they are associated, these gastrointestinal peptides are intricately involved in the regulation of gut function in addition to their role in regulating feeding behavior.

In the longer term, the brain receives information about nutritional status via adiposity signals such as leptin secreted by adipocytes and insulin from the pancreas. In the central nervous system, the arcuate nucleus appears to play a central role in regulating feeding via projections to the lateral, paraventricular, and ventromedial nuclei of the hypothalamus. The various peripherally generated nutrition-related signaling molecules reach the arcuate nucleus via the median eminence and are transduced into two kinds of signals – either suppressing food intake and involving the expression of pro-opiomelanocortin and cocaine- and amphetamine-related transcript or augmenting food intake and involving the expression of neuropeptide Y and agouti-related peptide.

It is important to recognize the degree to which separate hormone–behavior systems are interrelated. For example, in some species (e.g., elephant seals), males do not feed during the mating season, and in most species the onset of puberty is dependent on nutritional status.

Summary

Several kinds of behavior are dependent on or influenced by circulating hormones. Many, but not all, of these are related to reproduction and are accordingly influenced by the same suite of steroid and peptide hormones that serve to regulate the other aspects of reproductive physiology. Sex differences in vertebrate hormone metabolism occur throughout the development of the individual, from the differentiation of the fetal gonad until death. For much of this period, these hormonal differences have effects, both subtle and profound, on the brain and on behavior. In many cases, the systemic hormonal antecedents to particular behaviors in adults are well known, although in the brain, the localized release of various peptide hormones and the localized metabolism of steroids make it a much more difficult task to identify which hormone is contributing how much to which effect in which target cell. For some behaviors, particularly those such as rodent lordosis in which the stimulus–response relationship is a simple one, the anatomy of the neural circuits responsible for responding transducing hormone–behavior relationships is fairly thoroughly elucidated. In others, such as female proceptive behaviors, the neural substrates are more of a mystery. In almost all cases, our knowledge of neural structure stops at the level of anatomical connectivity. Despite the fact that the first sex difference in neural structure was reported at the ultrastructural level, we still know very little about how hormones alter the flow of information from one neuron to another, which is presumably a prerequisite for changing behavior. This last lacuna in our knowledge of hormone–behavior relationships is perhaps the most inevitable, given that the same problem exists throughout behavioral neuroscience, and for the same reason it is one of the most challenging.

See also: Aggression: Hormonal Basis; Development of Behavior; Emotional Hormones and Memory Modulation; Estrus and Menstrual Cycles: Neuroendocrine Control; Gene Expression Regulation: Steroid Hormone Effects; Gonadal Steroid Actions on Brain; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Hormones and Memory; Neuroendocrinology of Social/Affiliative Behavior; Sexual Behavior: Neuroendocrine Control.

Further Reading

Dominguez JM and Hull EM (2005) Dopamine, the medial preoptic area, and male sexual behavior. *Physiology and Behavior* 86: 356–368.

Nelson RJ (2005) *An Introduction to Behavioral Endocrinology*. Sunderland, MA: Sinauer.

Nelson RJ (ed.) (2005) *Biology of Aggression*. New York: Oxford University Press.

Paredes RG (2003) Medial preoptic area/anterior hypothalamus and sexual motivation. *Scandinavian Journal of Psychology* 44: 203–212.

Pfaff D, Arnold AP, Etgen AM, Fahrbach SE, and Rubin RT (eds.) (2002) *Hormones, Brain and Behavior*, 5 vols. New York: Academic Press.

Rosenblatt JS (1992) Hormone–behavior relations in the regulation of parental behavior. In: Becker JB, Breedlove SM, and Crews D (eds.) *Behavioral Endocrinology*, pp. 219–259. Cambridge, MA: MIT Press.

Strader AD and Woods SC (2005) Gastrointestinal hormones and food intake. *Gastroenterology* 128: 175–191.

Wingfield JC, Hegner RE, Dufty AM, and Ball GF (1990) The 'challenge hypothesis': Theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. *American Naturalist* 136: 829–846.

Wynne K, Stanley S, McGowan B, and Bloom S (2005) Appetite control. *Journal of Endocrinology* 184: 291–318.

Young LJ and Wang Z (2004) The neurobiology of pair bonding. *Nature Neuroscience* 7: 1048–1054.

Hormones and Memory

J S Janowsky and D R Roalf, Oregon Health and Science University, Portland, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

Background

Compelling neurobiological evidence from animal models suggests that estrogen and testosterone influence the structure and function of brain regions important for memory. These regions include the hippocampus, amygdala, basal ganglia, and prefrontal cortex, all of which have traditional androgen or estrogen receptors or have what are considered to be rapid 'membrane mechanisms' of hormone action. A variety of physiological changes occur in these regions when sex hormone levels are manipulated. These findings have shaped studies of sex hormone actions on memory in humans; however, findings from human studies are as yet ambiguous. The literature can best be characterized as the 'the good, the bad, and the ugly.' There are several situations through which hormone action in humans has been studied. These include effects on memory of hormone changes across the menstrual cycle, age and menopause-related hormone loss, hormone therapy in healthy men and women, hormone supplementation in neurodegenerative disease, and hormone loss for prostate cancer treatment in men. Recent studies of memory effects from these models will be briefly reviewed here with attention to the specific forms of memory affected. That is the good news. Most studies of animal models use 17-β-estradiol, testosterone, or a metabolite of these sex steroids. Human studies must utilize medications and application methods approved for human use. Thus, the studies reviewed examined effects from a variety of medications that modify the hormonal milieu of the brain including estradiol, conjugated equine estrogen, testosterone, gonadotropin-releasing hormone agonists, administered by pill, injection, patch, gel, or after surgical oopherectomy. Differences in memory effects among these different methods of hormone intervention are not well established in the literature. That is the bad news. Finally, data come from naturalistic, cross-sectional, and placebo-controlled trials. Thus, differences in effects or lack of effects among studies may well be due to the difference in medications, sample, and design. That is ugly, but part and parcel of the constraints of human studies.

Estrogen and Declarative Memory

Effects of Estrogen on Memory during Normal Menstrual Cycling

Definitive studies of memory effects during menstrual cycling and their relationship to sex hormone levels in young women are sparse. The phase of the menstrual cycle (e.g., high estradiol and progesterone in the luteal phase versus low sex hormones in the menstrual phase) does not modify memory in women with mild or severe premenstrual symptoms nor does it modify cortisol-induced memory impairments. Similarly, explicit verbal memory is not affected by menstrual phase in healthy young women despite effects in other cognitive domains including visual priming and mental rotation. Visual memory is poorer at the low estrogen menstrual phase as opposed to the luteal phase, but other measures of memory are unaffected. Factors that qualify studies in the literature are the lack of confirmation of menstrual phase with estradiol levels, very small sample sizes, and oral contraceptive use by women in some studies. The lack of definitive effects of estradiol modulation, particularly of verbal memory in younger women, is in contrast to effects in aging.

Effects of Estrogen in the Perimenopausal Period

During the transition to menopause, estrogen production gradually declines and women report memory difficulties. This transition causes physiological symptoms, such as hot flushes and sleep disturbances. Estradiol therapy relieves these physiological symptoms. Studies suggest that women on estrogen therapy to relieve symptoms of the menopause and particularly after surgical menopause, when hormone levels drop precipitously to negligible levels, have superior recall of verbal information. What is still unclear is whether this is a critical period for long-term estrogen effects in aging, or whether the critical factor is the severity or rate of hormone loss. A placebo-controlled trial showed that women (mean age 51.2 years) on conjugated equine estrogen for 21 days had numerically higher verbal memory performance than controls, but this did not reach significance when scores were corrected for general intelligence differences among treatment groups. Furthermore, five other randomized clinical trials show that women under the age of 65 receiving estrogen are more likely to show better verbal memory performance later in life than those not on estrogen. In addition, a novel intranasal estradiol

preparation administered for 2 years did improve women's perception of their memory in a dose-dependent manner. These studies are in contrast to no effects or poorer performance in studies where hormone therapy is initiated late in life, after the age of 65. Verbal memory does not differ in women at different stages of the menopausal transition, nor between current versus past hormone users in a population-based observational study. Similarly, a placebo-controlled study finds that the severity of menopausal symptoms, their relief, or estradiol treatment during the transitional phase of natural menopause affects memory. However, a significant association occurs such that those who began hormone use during the menopausal transition have better memory subsequently than those who begin hormone use only after the final menstrual period. Neither verbal nor nonverbal memory differed among women who were never, current, or past users of hormone therapy when the data were adjusted for age, education, and other cognitive performance differences among hormone-use groups in another cross-sectional study. Together these studies suggest that verbal memory is affected by estrogen when it is begun during but not after the menopausal transition and this may be a narrow window. Thus, there may be a critical period during which the influence of estrogen on memory, particularly verbal memory, is most significant. Similar temporally limited effects of estrogen replacement after gonadectomy are found in rodent models. This would suggest that memory areas change in their responsiveness to estradiol depending on age or the prior hormonal milieu of the brain.

Effects of Estrogen on Older (Postmenopausal) Women

Hormone therapy in older postmenopausal women may have a protective effect against age-related memory decline, and/or the promotion of memory performance over and above age-related changes in memory. For instance, an early randomized controlled study found that a year of estradiol therapy improved performance on the Wechsler memory scale as compared to a placebo control group or pretreatment performance in older women confined to a nursing home. In contrast, recent data from women in their 70s find that there is no effect of estrogen in combination with progesterone on verbal or nonverbal (figural) memory approximately 3 years after initiation of hormone therapy. However, estrogen and progesterone treatment are negatively associated with verbal memory and positively associated with nonverbal memory rates of change over the approximately 3-year follow-up interval. Estradiol alone for a shorter duration (20 weeks) in a similar sample of older women does not affect cognition. Blinded controlled data on the long-term effects of estradiol alone are as yet not available. The female advantage for verbal memory tasks remains when older men and women with similar estradiol levels are examined. This suggests that sex differences are not dependent on circulating estradiol levels in adulthood.

Case-control studies compare women who are users versus nonusers of hormone therapy and who are matched for a variety of other demographic variables. This is the only method that is available to understand whether estradiol has effects in older women when used over much of their postmenopausal life. Healthy, older postmenopausal women on estrogen therapy perform significantly better than nonusers of estrogen on word-list recall and on both immediate and delayed paragraph recall. Users of hormone therapy had better visual memory on the Benton visual retention task as compared to women not using hormone therapy. Visual memory, as measured by the visual reproduction test of the Weschler memory scale, is also significantly better in women on hormone therapy. However, not all studies find such effects. For instance, users of hormone therapy perform similarly to never users of hormone therapy on a test of story memory in a large sample of healthy older women. Hormone use is not related to rate of memory decline in older women over a 2-year study period, and in fact hormone users have a numerically faster rate of memory decline compared to never users.

Estradiol levels are positively related to verbal memory in elderly women (age 61–91 years) even if they are not on hormone therapy but are not associated or are negatively associated with visual memory. The negative association of estradiol levels and visual memory in hormone users as opposed to the positive effects observed in hormone users suggests that there may be a threshold over which estradiol has positive effects on nonverbal memory in elderly women.

Brain activity during memory tasks is also associated with hormone therapy in older women. For example, measurements of blood flow (using positron emission tomography (PET)) in brain regions that subserve verbal and nonverbal memory processing, including the parahippocampus and right frontal regions, are higher in women using hormone therapy as compared to women without hormone therapy. A 2-year follow-up PET study finds increases in regional cerebral blood flow (rCBF) in estrogen users as compared to nonusers in areas associated with a verbal recognition memory test, including the hippocampus, parahippocampus, and middle temporal lobe. This parallels findings of improved memory performance over an 18-month period in hormone users as compared to nonusers.

Estrogen and Working Memory

Estrogen does not affect working memory performance in middle-aged women (mean age 50.8 years) but does result in increased brain activity in the inferior parietal lobe during storage of verbal information and decreased brain activity in the same region during storage of nonverbal material. Another placebo-controlled study finds no influence of estrogen therapy on working memory in older women using the subject-ordered pointing test. Cross-sectional studies find middle-aged women on estrogen perform better than nonusers on working memory tasks comprised of both verbal and spatial components. Young women have better verbal working memory (span task) during the high estrogen phase as opposed to the low estrogen phase of the menstrual cycle as inferred by body temperature, while the performance of women on birth control, and estradiol levels, do not vary but remain constant.

Estrogen and Implicit Learning

Higher estradiol levels or performance at the high estrogen phase of the menstrual cycle in young, premenopausal women is related to faster motor dexterity on the grooved pegboard task, and motor sequence learning, but not simple reaction time using a task associated with the striatum. Visual priming may be lessened at the high estrogen phase, although this was confounded with the order of testing across menstrual phase.

Sex Hormone Effects in Men

Testosterone supplementation is most commonly used to treat hypogonadism in older men. In addition, testosterone can be converted to estrogen in the brain. Thus, studies using inhibitors to block the metabolism to estrogen, or studies of complete blockade of the androgen pathway, as is a common treatment to prevent the progression of prostate cancer, will result in very low estradiol in men. Very high-dose estradiol has also been studied as a novel treatment for prostate cancer. Studies using these models show that significant changes of testosterone or estradiol levels affect memory in older men. Testosterone deprivation results in poorer verbal memory in younger men, as well as older men on hormonal therapy for prostate cancer. Replacement with very high-dose estradiol, that also maintains an androgen blockade, results in improved verbal memory in men with prostate cancer. Similarly, testosterone supplementation in hypogonadal men results in improved verbal memory, while dihydrotestosterone, an androgen that does not metabolize to estradiol, results in improved spatial memory. Blocking the conversion of testosterone to estradiol blocks the enhancement of verbal memory in older men. Together these studies suggest that estradiol is important for verbal memory enhancement, at least in older men. Memory is not affected by testosterone supplementation in men with congenital disorders that cause hypogonadism. Very brief testosterone supplementation by a single injection or dose does not affect memory in younger men, although object location memory is improved in younger women.

Working memory as assessed with the subject-ordered pointing test is improved with testosterone supplementation in older men, and cross-sectional studies show that higher free (unbound) testosterone levels are related to lower rates of decline in verbal and nonverbal memory as well as a working memory measure, Trails B, performance as men age. The relationship between testosterone levels and memory performance gets stronger as men age. Like estrogen, cross-sectional studies suggest that low free testosterone is a premorbid risk factor for Alzheimer's disease in men and is associated with poorer performance on screening measures for dementia in women. Together these studies suggest that testosterone or its metabolites may play a role in maintaining brain health in aging.

Concluding Comments on Sex Hormones and Memory

Estradiol does not affect declarative memory in younger cycling women, and the effects on memory in older women are unclear. The timing, dose, and form of sex hormones influence the outcome of studies. Still unclear is whether these hormones activate memory performance or have a neuroprotective role when cognition declines. If it were the latter, then effects would only be seen with comparisons among people with age-related or disease-related cognitive decline. These lifetime factors complicate the ability to study effects in humans versus effects found in animal models. Regardless, the effects in humans are subtle as compared to the effects of gonadectomy and hormone replacement in animal models. This suggests that the day-to-day role of sex hormones in humans is to modulate memory formation, while severe and sudden hormone deprivation, particularly in critical periods for age-related decline, may amplify memory impairments and risk for neurodegenerative disease.

See also: Aging and Memory in Animals; Aging and Memory in Humans; Emotional Hormones and Memory

Modulation; Estrus and Menstrual Cycles: Neuroendocrine Control; Procedural Learning in Animals.

Further Reading

Cherrier MM, Matsumoto AM, Amory JK, et al. (2005) The role of aromatization in testosterone supplementation: Effects on cognition in older men. *Neurology* 64(2): 290–296.

Henderson VW, Guthrie JR, Dudley EC, Burger HG, and Dennerstein L (2003) Estrogen exposures and memory at midlife: A population-based study of women. *Neurology* 60(8): 1369–1371.

Janowsky JS (2006) Thinking with your gonads: Testosterone and cognition. *Trends in Cognitive Science* 10(2): 77–82.

Janowsky JS, Chavez B, and Orwoll E (2000) Sex steroids modify working memory. *Journal of Cognitive Neuroscience* 12(3): 407–414.

LeBlanc ES, Janowsky J, Chan BK, and Nelson HD (2001) Hormone replacement therapy and cognition: Systematic review and meta-analysis. *JAMA* 285(11): 1489–1499.

Maki PM (2006) Hormone therapy and cognitive function: Is there a critical period for benefit? *Neuroscience* 138(3): 1027–1030.

McEwen B (1999) The molecular and neuroanatomical basis for estrogen effects in the central nervous system. *Journal of Clinical Endocrinology and Metabolism* 84(6): 1790–1797.

Moffat SD, Zonderman AB, Metter EJ, Blackman MR, Harman SM, and Resnick SM (2002) Longitudinal assessment of serum free testosterone concentration predicts memory performance and cognitive status in elderly men. *Journal of Clinical Endocrinology and Metabolism* 87(11): 5001–5007.

Moffat SD, Zonderman AB, Metter EJ, et al. (2004) Free testosterone and risk for Alzheimer disease in older men. *Neurology* 62(2): 188–193.

Muller M, Aleman A, Grobbee DE, de Haan EH, and van der Schouw YT (2005) Endogenous sex hormone levels and cognitive function in aging men: Is there an optimal level? *Neurology* 64(5): 866–871.

Parducz A, Hajszan T, Maclusky NJ, et al. (2006) Synaptic remodeling induced by gonadal hormones: Neuronal plasticity as a mediator of neuroendocrine and behavioral responses to steroids. *Neuroscience* 138(3): 977–985.

Resnick SM, Maki PM, Rapp SR, et al. (2006) Effects of combination estrogen plus progestin hormone treatment on cognition and affect. *Journal of Clinical Endocrinology and Metabolism* 91: 1802–1810.

Sherwin BB (2003) Estrogen and cognitive functioning in women. *Endocrine Reviews* 24(2): 133–151.

Sherwin BB (2006) Estrogen and cognitive aging in women. *Neuroscience* 138: 1021–1026.

Yonker JE, Eriksson E, Nilsson LG, and Herlitz A (2003) Sex differences in episodic memory: Minimal influence of estradiol. *Brain Cognition* 52(2): 231–238.

Hox Genes Expression

M Parrish, C Nolte, and R Krumlauf, Stowers Institute for Medical Research, Kansas City, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

What Are *Hox* Genes?

Hox genes encode a family of homeodomain-containing transcription factors required for patterning regional properties along the rostral–caudal (R/C) axis during embryogenesis. The *Hox* gene family is highly conserved in organisms ranging from *Drosophila* to humans. In most vertebrates, except ray-finned fishes, there are 39 members segregated into four tightly clustered gene arrays (A–D) on four separate chromosomes (Figure 1). The *Hox* complexes are believed to have arisen by tandem duplication of a single gene to create a cluster, followed by duplication and divergence of the ancestral cluster associated with large-scale genomewide duplications in vertebrate evolution. This gene family is defined by sequence conservation, their genomic organization, and the presence of a homeobox or homeodomain motif within the protein. This domain is a helix–turn–helix protein motif responsible for sequence-specific DNA binding. These genes were initially identified in fly mutant screens as genes important in regulating segmental identity. One of the most famous examples is a gain-of-function mutation of the *Antennapedia* (*Antp*) gene, which transforms fly antenna into legs (Figure 2). Further investigation led to the discovery of 8 Hox proteins in *Drosophila*, each with roles in patterning and R/C-restricted patterns of expression.

Hox Gene Expression

Initial experiments demonstrating dramatic phenotypes associated with *Hox* gene deficiencies stimulated exploration of their expression patterns in *Drosophila* and other organisms. Surprisingly, developmentally regulated, nested and ordered patterns of expression were observed during embryogenesis. This further elevated interest in this family as developmental regulators because their patterns of expression along the embryonic R/C axis correlated with their gene order along the chromosome, suggesting they might provide a molecular means of defining head from tail or different structures from a common tissue. In the mouse, *Hox* gene expression begins at approximately embryonic day 6 (E6), one-third of the way through mouse development, after gastrulation has begun and the R/C axis has extended and become patterned. During the middle third of mouse development, *Hox* genes are highly expressed in the developing neural tube, often in dynamic patterns. The pattern of expression of individual *Hox* genes generally correlates both temporally and spatially with the location of the gene within its complex, a phenomenon called collinearity. Genes at the 3′ end of the complexes (paralog group 1) are expressed earliest with the most rostral boundaries of expression, and these two characteristics progressively change in successive genes in the complex, such that the most 5′ genes (paralog group 13) are only expressed caudally and have a delayed onset of expression compared to 3′ genes (Figures 1, 3 and 4). Although the spatial/temporal aspects of expression patterns progressively change between adjacent paralog groups, within paralog groups expression characteristics and even protein functions are often similar. Although the focus of this article is on the role of *Hox* genes in R/C patterning, they are expressed and have important functions in many other tissues. *Hox* genes have roles in hematopoiesis; lung, thyroid, parathyroid, kidney, breast, and muscle development; and in development of the structure of the head and neck and in the gastrointestinal and urogenital systems, where they have at least partially collinear patterns of expression. A common feature of *Hox* genes is that they provide a means of specifying different regional properties along the axis of a tissue. Within these systems they can do so by regulating a broad range of cellular activities, such as proliferation, differentiation, migration, patterning, adhesion, and cell death. In fact, part of the excitement about *Hox* genes is that they provide insight into the cellular and molecular mechanisms that use common building blocks to create different tissue structures (leg vs. antenna, hindbrain vs. spinal cord, and rib vs. tail bone).

Hox Gene Function

Since their initial discovery and analysis in *Drosophila*, Hox proteins have been demonstrated to be crucial molecules in regulating R/C patterning and defining segmental identity in vertebrates as well. Targeted mutagenesis of *Hox* genes in the mouse or manipulation of *Hox* gene expression in chickens via *in ovo* electroporation often alter segmental identity in the hindbrain and spinal cord as demonstrated by changes in motor neuron differentiation and identity. For instance, when either *Hoxa1* or *Hoxb1* is deleted by gene targeting, it perturbs the development of rhombomere-specific motor neuron populations as evidenced by altered motor neuron migration and/or projection patterns. When either of these is ectopically expressed in rhombomere 1 (normally a region devoid of *Hox*

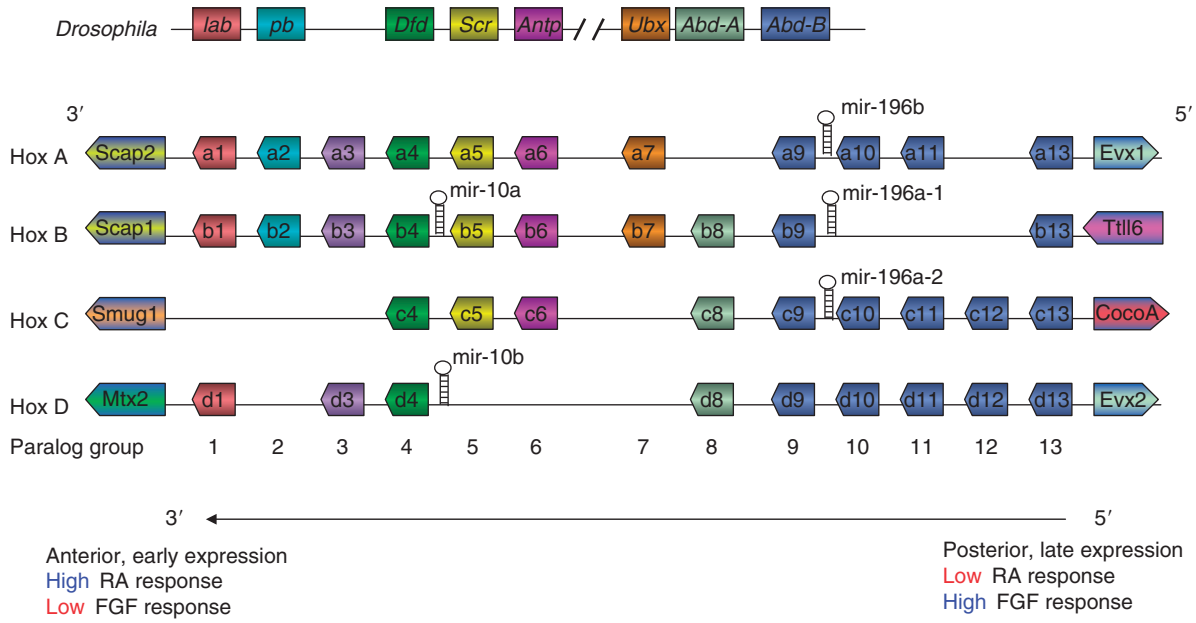

Figure 1 Schematic diagram of the split *Drosophila Hox* complex and its mammalian homologs. Homologs are indicated in the same color as their *Drosophila* homolog, listed in 3′ to 5′ orientation, and transcriptional direction is indicated by the pointed edge of each box. Paralog groups are indicated by numbers below each column. miRNAs embedded within clusters are indicated. Next adjacent genes 3′ and 5′ of each complex are also listed.

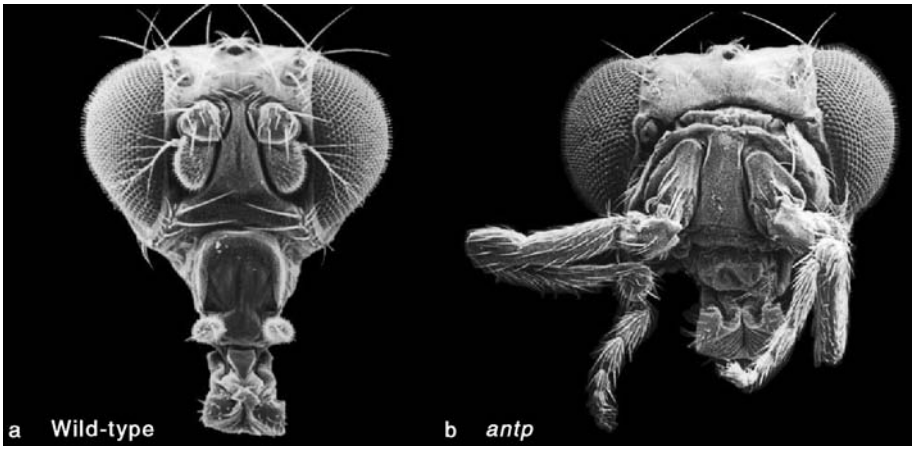

Figure 2 Scanning electron micrographs of wild-type (a) and *Antennapedia* (*antp*) gain-of-function mutant (b). Ectopic expression of *antp* in the antenna primordium transforms them into leglike structures. Courtesy of F Rudi Turner, Indiana University.

expression), it triggers ectopic development of motor neuron types typically found in more posterior segments. Similar experiments manipulating *Hox* gene expression in the developing chick spinal cord have shown that the products of *Hox* genes are the primary factors defining motor neuron pool subdivisions which innervate specific limb muscles. These studies emphasize that within the hindbrain and spinal cord, Hox proteins play a pivotal role in developmental patterning by defining segmental identity and controlling segment-specific structures and cell types. Pharmacological studies also support this role in that agents that perturb *Hox* regulatory pathways result in corresponding shifts in *Hox* expression profiles and regularly lead to alterations in segmental identity. Hence, detailed knowledge of the mechanisms of *Hox* gene regulation is fundamental to understanding how segmental identity is established and maintained. The following sections summarize the current understanding of basic mechanisms of *Hox* gene regulation within the hindbrain and spinal cord and provide specific examples to demonstrate key points.

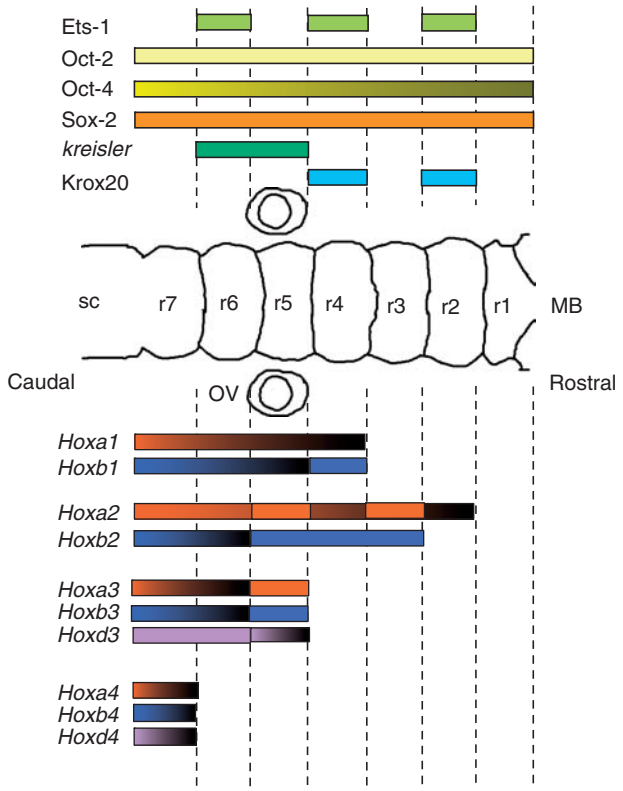

Figure 3 Expression patterns of *Hox* gene regulatory transcription factors and *Hox* genes are indicated by colored bars above and below the schematic of the developing mouse hindbrain at E9.5. Color gradients indicate relative expression levels or caudally regressing patterns of expression. r1–r7, rhombomeres 1–7; MB, midbrain; SC, spinal cord. Dashed lines indicate rhombomere boundaries.

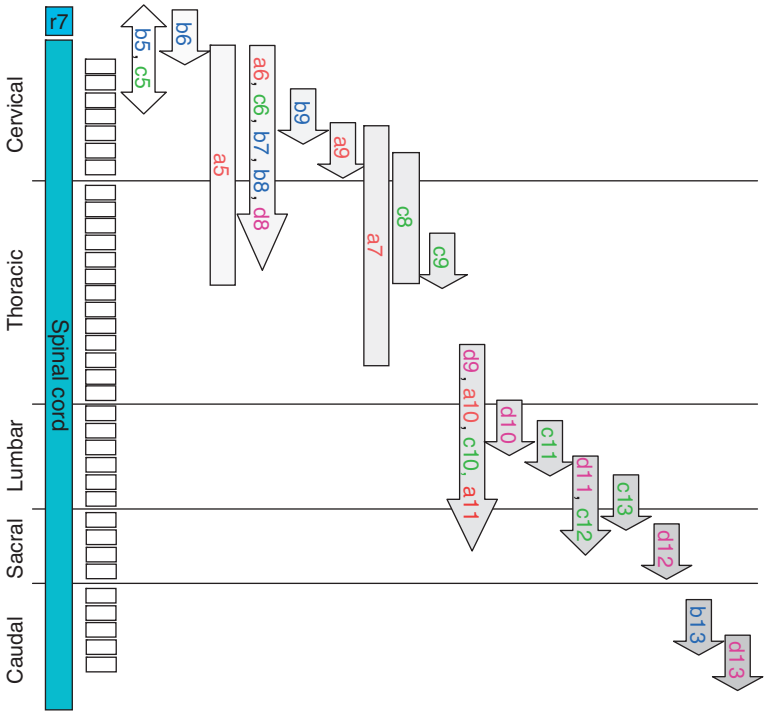

Figure 4 Expression patterns of caudally expressed *Hox* genes between E11.5 and E12.5. Spinal cord regions and associated segments are indicated on the left. The blunt ends of arrows indicate expression boundaries and the pointed ends signify that expression continues rostrally or caudally in the direction of the arrows. Expression patterns that have both anterior and posterior boundaries within the limits of the diagram are indicated as boxes.

Signaling Pathways Are Important Regulators of *Hox* Gene Expression

The retinoic acid (RA) and fibroblast grow factor (FGF) signaling pathways regulate significant aspects of *Hox* gene expression patterns. Manipulation of either of these pathways results in both alterations in *Hox* gene expression patterns and changes in R/C segmental identity. This has been demonstrated by adding exogenous RA or FGF, generating vitamin A-deficient (VAD) embryos, employing chemical agonists/antagonists, or expressing dominant negative FGF receptors. Additionally, mice engineered to be deficient in Cdx transcription factors, one of the integrators of FGF signaling, also display phenotypes which are strikingly similar to mice bearing homozygous *Hox* gene deficiencies. The teratogenic effects of RA exposure on human development and the similarity of the results to those of experiments done in mouse, chick, frog, and fish model systems corroborate the importance of *Hox* genes in segmental identity and the importance of RA in *Hox* gene regulation.

Regulation of *Hox* Genes by RA

RA, the active form of vitamin A, is produced by an enzymatic pathway that culminates in the conversion of retinaldehyde to RA by retinaldehyde dehydrogenase (RALDH). Lipophilic RA then diffuses or is transferred into surrounding tissues and cells so that it is able to bind to retinoic acid receptors (RARα, -β, and -γ and RXRα, -β, and -γ). These nuclear receptors exist as heterodimers that together bind retinoic acid response elements (RAREs) within the genome and modulate gene transcription. The binding of RA to these receptors converts these bound heterodimers to transcriptional activators via increasing their affinity for coactivators and decreasing their affinity for co-repressors. RA signaling relevant for R/C patterning of the hindbrain and spinal cord is initially regulated by controlling RA production via *Raldh2* expression in somitic mesoderm. After RA has been produced, local concentrations are modulated by cytoplasmic retinoic acid binding proteins (CRABP1 and -2), clearance by the blood, isomerization from 9-*cis* to the less potent all-*trans* RA, and, finally, degradation via the Cyp26 family of enzymes. This provides a complex metabolic mechanism to ensure the levels of this important signal are very carefully modulated in a tissue-specific and temporally specific manner. A transgenic mouse line bearing an RA responsive marker gene shows that these mechanisms synergize to produce substantial RA signaling in the eye, telencephalon, and throughout the spinal cord and adjacent mesoderm/somites from the level of the caudal hindbrain throughout the majority of the trunk, with the lowest levels at the node and tailbud at later stages.

RA Response Specificity and Diversity

Many *Hox* genes have been shown to be RA responsive, and several functional RAREs have been identified near *Hox* genes demonstrating that they are direct targets of the signaling pathway. Multiple DR5 (direct repeat + 5-base pair spacer; i.e., AGGTC AnnnnnAGGTCA) and DR2 RAREs have been identified within the *Hox* clusters and associated with either activation or repression of one or more neighboring genes. For example, transcription of both *Hoxb3* and *Hoxb4* is activated by a single DR5 RARE located in the intergenic region between these two genes. *Hoxb1*, on the other hand, is associated with three separate RAREs, all of which have distinct influences on its expression. A DR5 RARE 5′ of *Hoxb1* mediates repression of *Hoxb1* in rhombomeres 3 and 5, whereas a DR2 RARE and a DR5 RARE 3′ of *Hoxb1* both activate expression of *Hoxb1* but in different spatial patterns (**Figure 5**). Several other 3′ *Hox* genes not mentioned here also have RAREs with similar organization and function as those mentioned previously. Thus, RA affects a broad range of *Hox* expression patterns, and although these responses can be regulated both directly and indirectly, it is clear that *Hox* genes are often direct targets of RA activity. Additional modes of regulation continue to be investigated.

FGF Regulates *Hox* Expression

FGFs are also important regulators of *Hox* gene expression. FGFs are an expanding family of secreted signaling molecules that includes more than 20 members. After secretion, these soluble proteins bind to transmembrane tyrosine kinase receptors, FGFR1–4. These ligand-bound receptors dimerize, autophosphorylate themselves, and transduce intracellular signals via mitogen-activated protein kinase, phosphatidylinositol 3-kinase, or protein kinase C/protein lipase C (PLCγ), among others. Activation of these pathways can result in a broad spectrum of responses depending on the context; however, activation of *Cdx1–4* homeobox gene expression is one of the primary modes of influencing *Hox* gene expression. Cdx homeodomain transcription factors, homologs of the *Drosophila* caudal protein, are capable of binding *Hox* gene enhancers and directly regulating *Hox* transcription. Hence, FGFs activate *Hox* gene expression indirectly via the *Cdx* family of genes. FGF8 expressed in the isthmus (a thin region that separates the midbrain from the hindbrain) and the node (a signaling center and region of posterior

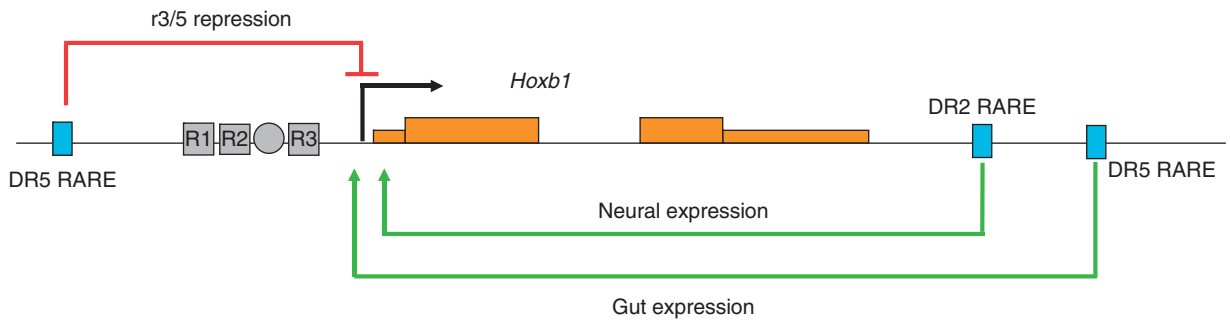

Figure 5 RA regulation of *Hoxb1*. *Hoxb1* has three retinoic acid response elements (RAREs) indicated by blue boxes. Two of these activate expression of *Hoxb1* (green arrows) and one mediates repression (red line). The *Hoxb1* autoregulatory *Hox* response elements are indicated as gray squares and circles. DR2 and -5, direct repeat 2- and 5-base pair spacers.

elongation) has significant patterning influences on the R/C axis. FGF8 from the node influences *Hox* gene expression throughout much of the spinal cord during elongation of the axis. Isthmic sources of FGF8 are critical for maintaining the identity of rhombomere 1, the first hindbrain segment, but its role in *Hox* gene repression in this uniquely Hox-free region is poorly understood.

FGF and RA Regulate Opposing Groups of *Hox* Genes

Although both RA and FGF exhibit dose-dependent activation of *Hox* gene expression, they display opposing preferences for which *Hox* genes are activated by each. RA preferentially activates *Hox* genes in the 3′ portion of the complex (*Hox1–5*) and has little influence on the expression of 5′ genes (*Hox6–13*). For instance, expression of a dominant negative RAR diminishes *Hoxb4* expression in the developing spinal cord but does not alter *Hoxb9*. Functional RAREs responsible for this preferential activation have been identified in the 3′ portions of the complexes. In contrast, FGF preferentially influences *Hox* genes in the 5′ portion of the complex (*Hox6–13*) and can expand their rostral boundaries of expression. FGF specificity appears to be determined, in part, by *Cdx* expression domains, which supports the hypothesis that Cdx proteins play an important role in integrating FGF signaling. FGFs are capable of activating expression of both 5′ and 3′ *Hox* genes and expanding their expression domains, but activation of 3′ *Hox* genes in the hindbrain is limited by the availability of Cdx. Whereas exogenous FGF can extend the expression of *Hoxb9* from its caudal boundary up to the level of the otic vesicle, *Hoxb4* expression that normally terminates just posterior to the otic vesicle cannot be expanded by additional FGF alone. Ectopic expression of a dominant active Cdx (XcadVP16) in the presence of exogenous FGF, however, can expand both *Hoxb4* and *Hoxb9* expression past the otic vesicle and into the hindbrain.

FGF and RA Generate Opposing Gradients

RA and FGF signaling oppose each other on multiple levels. They preferentially influence opposing groups of *Hox* genes and also positively and negatively cross-regulate the opposing pathway. For example, FGF can activate transcription of RARs, activate expression of *Cyp26* (the enzyme responsible for RA degradation), and block expression of *Raldh2* and thus attenuate RA production. Conversely, RA can activate *FGFR* expression, block expression of fgf8, block expression of *Cdx1*, and stimulate its own degradation by upregulation of *Cyp26*. These diverse mechanisms work together to produce opposing gradients of FGF and RA signaling activity, which allows a wide range of concentration-dependent, segmentally restricted responses to FGF and RA, including their regulation of *Hox* gene expression (Figure 6).

Additional Signaling Pathways

Although FGF and RA appear to be the most influential signaling pathways regulating *Hox* genes, other pathways do appear to impinge upon *Hox* gene expression to a lesser extent. Gdf11, a secreted transforming growth factor-β homolog, can influence *Hox* gene expression and sensitize neural tube progenitors to FGF8 emanating from the node. The Wnt signaling pathway is another critical signal that influences regional character. Wnts can modulate Cdx1 activity and have an indirect input on *Hox* expression via this route. This further illustrates the importance of understanding Cdx regulation because it appears to integrate RA, FGF, and Wnt signaling in the regulation of *Hox* gene expression in the developing central nervous system and axial skeleton.

Regulation of *Hox* Genes by Transcription Factors

Regulation by transcription factors serves to establish, stabilize, and define the boundaries of *Hox* gene

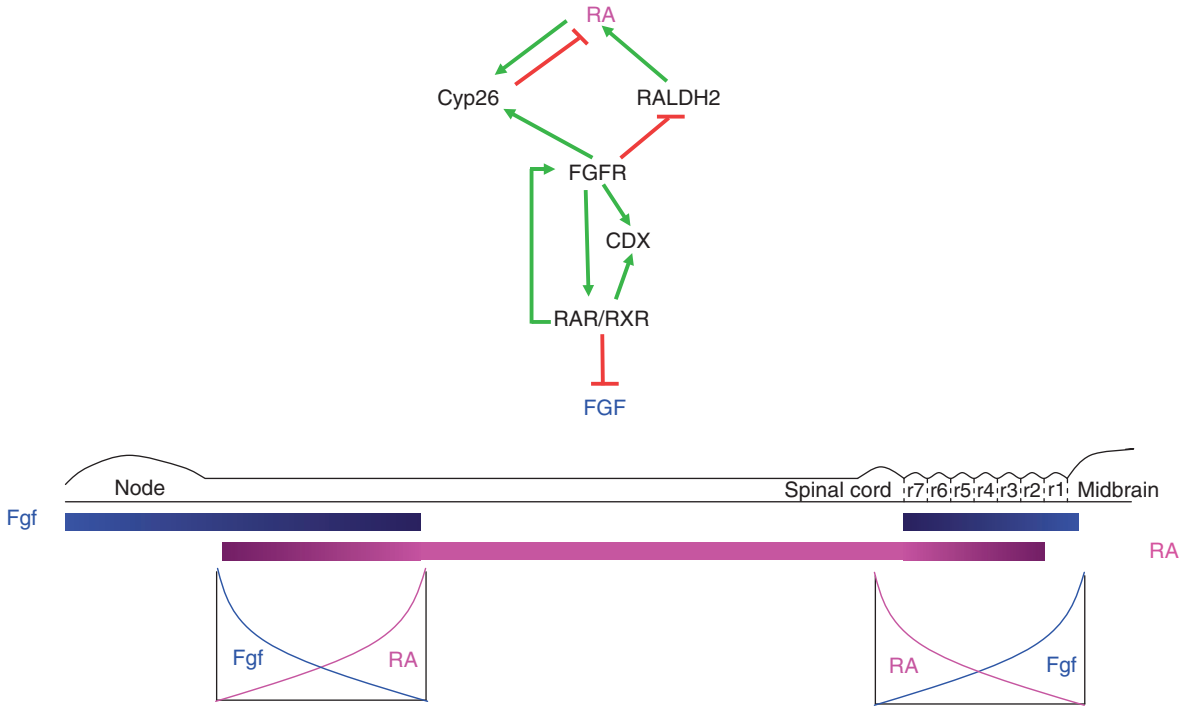

Figure 6 FGF and RA form opposing gradients through cross-regulation. (Top) Regulatory relationships between elements of the FGF and RA pathways. Activation is indicated by green arrows and repression by red lines. (Bottom) FGF (blue) and RA (purple) gradient formation along the rostral–caudal axis of the developing mouse embryo.

expression initiated by the previously mentioned signaling systems. Some of these upstream transcription factors activate *Hox* gene transcription in a defined way, limited by the boundaries of their own expression. Many of the transcription factors that regulate *Hox* gene expression, however, are ubiquitously expressed and presumably rely on association with additional proteins to achieve specificity. Exploring these cooperative mechanisms of specificity is an important area of current investigation.

Four spatially restricted and developmentally regulated transcription factors have been shown to be important regulators of *Hox* expression patterns: Krox 20, Kreisler (MafB), and Sox/Oct. *Krox 20* encodes a zinc finger transcription factor expressed in rhombomeres 3 and 5. Binding of Krox 20 to its target sequences upstream of *Hoxa2* and *Hoxb2* and between *Hoxb3* and *Hoxb4* activates expression of these genes in rhombomeres 3 and 5. *Kreisler,* encodes a bZIP leucine zipper containing transcription factor expressed in rhombomeres 5 and 6, is required to activate transcription of *Hoxa3* and *Hoxb3*. Two other transcription factors important in stem cell maintenance, Sox2 and Oct4, may form a complex and participate in regulating *Hoxb1* in rhombomere 4.

Polycomb group (PcG) proteins and trithorax group (TrxG) proteins are two classes of chromatin-modifying proteins that epigenetically regulate *Hox* gene expression. These groups form multiprotein chromatin remodeling complexes that either repress transcription (PcG) or activate transcription (TrxG). Although these complexes are ubiquitously expressed and regulate many genes and cellular behaviors, mutations in members of both of these groups can alter *Hox* gene expression patterns and compromise R/C patterning in ways similar to both *Hox* gene deficiencies and abnormal FGF or RA signaling. PcG regulation of *Hox* expression may be direct because the PcG protein Phc2 can bind directly to the *Hoxb8* promoter. Additional regulation of this system appears necessary, however, because Phc2 can be observed at the *Hoxb8* promoter in tissues with high *Hoxb8* expression and also in tissues with low *Hoxb8* expression levels. PcG mutations also alter *Cdx* expression, and thus may have additional indirect inputs into *Hox* regulation. Several other ubiquitous transcription factors have been shown to be capable of regulating *Hox* genes. Factors such as YY1, Sp1, USF, NFY, and AP2 can modulate *Hox* expression in some contexts. The mechanism by which these ubiquitous transcription factors regulate specific aspects of *Hox* gene expression remains unclear; however, it may be related to their association with other factors. Sp1 can interact with AP2 and indirectly cooperate with Hox proteins

themselves. USF and NFY can interact with each other, and YY1 and Plzf can both interact with PcG components, although the roles of these interactions *in vivo* remain poorly characterized.

Auto-/Cross-Regulation

In addition to regulation by numerous other transcription factors, *Hox* genes also exhibit extensive autoregulation. Hox proteins can bind to *Hox* responsive elements (HREs), which consist of a *Hox* target site and usually include an adjacent binding site for one or more cofactors, such as Pbx, Prep, or Meis. Among the numerous possibilities, HRE-mediated *Hox* regulation usually functions in three main ways to refine and maintain *Hox* expression: *Hox* genes may maintain their own expression (autoregulation), activate expression of other *Hox* genes expressed in the same segment (cross-activation), or have antagonistic interactions in which one gene represses the expression of another *Hox* gene (cross-repression). These are all important processes in constructing and maintaining the complement of *Hox* genes expressed at each axial level and in defining boundaries between adjacently expressed genes.

The most extensively studied HRE is that of *Hoxb1*, which contributes to its expression in rhombomere 4. This element contains multiple Hox/cofactor binding sites and exhibits both autoregulation by Hoxb1 and cross-activation by Hoxa1. RA signaling and RAREs are directly responsible for the initial activation of *Hoxb1* expression. However, at later stages of development the HRE assumes regulatory control of *Hoxb1* transcription and Hoxb1 maintains its own expression. Genetic studies in mouse reveal that both the *Hoxa1* and *Hoxb1* genes are required for normal expression of *Hoxb1* in rhombomere 4. HREs have been identified either functionally, by sequence comparison, or inferred by transcriptional responses to *Hox* overexpression and are associated with the regulation of at least one-fourth of all *Hox* genes. Auto- and cross-regulatory interactions between the *Hox* genes appear to be fundamental mechanisms for modulating Hox transcriptional activity. Although both autoregulatory and cross-activation functions of the *Hoxb1* HRE can be clearly demonstrated, these regulatory interactions become less distinguishable for *Hox* members more 5′ in each complex because of the increasing number of *Hox* genes expressed at more posterior levels and the potential for indirect relationships. For 5′ *Hox* genes, it is likely that HREs are utilized to produce defined boundaries between different *Hox* gene expression domains by cross-repression. Such relationships have been demonstrated for *Hoxc5–Hoxc8*, *Hoxc6–Hoxc9*, and *Hox4–Hoxa7* such that ectopic expression of one gene of the pair represses the expression of the second gene, producing precise boundaries between expression domains of the two genes. Many of these relationships have not been demonstrated to be direct; however, several of the *Hox* genes involved are associated with functional HREs (Figure 7).

Hox Cofactors

Although isolated Hox proteins show sequence-specific binding to a 4-base pair A/T-rich target site (ATTA or TAAT), their affinity for this site is usually low and the length of the site allows little specificity in target site selection. It is thought that associations with cofactors may confer specificity to the Hox proteins in target site selection. HREs that contain both a Hox binding site and cofactor target sites can dramatically increase affinity of Hox + cofactor complexes for the site, specificity of the response, and the ability of the complex to stimulate transcription. The primary Hox cofactors are of the Meinox/TALE (three amino acid loop extension) family of proteins which include Pbx1–3, related to Drosophila Extradenticle (Exd), and Prep1-4/Meis1-3, related to *Drosophila* Homothorax (Hth). Pbx proteins directly bind to Hox proteins via a hexapeptide motif, X (Y/F) (P/D) WM (K/R), N-terminal to the Hox homeodomain, whereas a C-terminal portion of the homeodomain participates in stabilizing this association. The interaction of Hox and Pbx expands the target site specificity to a consensus ATGATTNATNN, where the Hox protein binds to the TNATNN portion of the half site

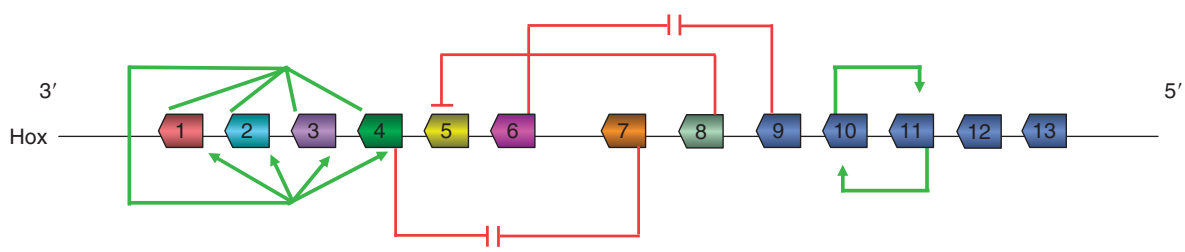

Figure 7 *Hox* gene auto-/cross-regulatory relationships. Regulatory relationships between paralog groups are indicated by green arrows, indicating positive loops, or red lines, indicating repressive relationships.

and Pbx binds to the ATGAT element. The binding selectivity of a Hox–Pbx target site is determined, in part, by site preferences conferred by the identity of the Hox protein present in the heterodimer. Prep/Meis cofactors initially bind to Pbx, which subsequently binds Hox proteins to form a ternary complex. These ternary complexes bind expanded target sites and activate transcription at a level higher than Hox/Pbx dimers and to a much greater extent than Hox proteins alone. Hox paralogs 1–10 can bind Pbx or Pbx–Prep/Meis heterodimers, paralogs 9 and 10 can bind Meis in the absence of Pbx, and groups 11–13 lack the Pbx interacting motif but can form complexes with Prep/Meis. Pbx/(Prep/Meis) heterodimers may also bind HREs in the absence of Hox and repress transcription. Both *in vitro* biochemical experiments and *in vivo* mouse deletions of these cofactors demonstrate their importance in *Hox* gene regulation and the function of HREs. Whereas the formation of Hox–Pbx heterodimeric complexes is DNA dependent, the formation of Pbx–Prep/Meis heterodimers is not. Pbx–Prep heterodimers can form in the cytosol and heterodimerization facilitates the nuclear localization of Pbx. Understanding these protein complexes and the events that modulate Hox/cofactor interactions is important in building a clear understanding of how Hox proteins find and act on their target sites *in vivo*.

Although the functional significance of Hox protein interactions is only beginning to be explored, several other developmentally important transcription factors have been shown to be capable of physically interacting with Hox proteins and these may also be relevant *in vivo* as cofactors or modulators. Modulation of Hox protein function may explain additional aspects of Hox protein target site specificity and switch Hox transcriptional activity between activation and repression. The transcription factors Sp1, Btg1, Smads, and members of the Maf family can interact with Hox proteins in some contexts. Although many of these interactions were investigated in artificial systems, these proteins were capable of interacting with Hox proteins and affecting their ability to transactivate reporter expression. Sp1 and Hoxa13, known regulators of BMP4, cooperate in activating transcription from the BMP4 promoter. Btg1 and-2 have been identified as Hoxb9 interacting proteins by the yeast two-hybrid assay and work together to activate transcription from a *Hoxb9* response element-driven reporter. Although Maf proteins have been shown to interact with several Hox proteins, these experiments did not reveal the function of these interactions. Kriesler, a member of this Maf family of proteins, has been shown to both regulate *Hox* gene expression and be involved in R/C patterning of the hindbrain; thus, it seems unlikely that these interactions would be without functional consequence during hindbrain development. These potential Hox cofactors participate in many different systems, but the importance of their interaction with Hox proteins remains unknown.

Initiation Versus Maintenance

Regulatory influences on *Hox* gene expression often result in expression patterns controlled in two separate phases, an early initiation phase and a later maintenance phase. Many of the 3′ *Hox* genes are regulated primarily by a combination of retinoic acid response elements and autoregulatory HREs. RA functions early in initiating expression in the absence of active *Hox* gene expression, and as Hox proteins accumulate within cells, the HREs may then assume regulatory influence over these genes. This mechanism is critical in maintaining *Hox* expression because expression patterns of *Cyp26* expand to encompass much of the hindbrain, reducing RA there and decreasing RARE-mediated *Hox* gene expression. Experiments in which the DR2 RARE 3′ of *Hoxb1* was deleted demonstrate this phenomenon very clearly. They show that the DR2 RARE 3′ of *Hoxb1* is required for correct temporal and spatial initiation of *Hoxb1* expression, and the remaining autoregulatory elements are capable of directing accurate *Hoxb1* spatial expression patterns at later time points.

Competition and Sharing

Unlike most other genes, *Hox* genes are organized into dense arrays/clusters. Nine to 11 genes are clustered within each ~130 kb genomic locus and are all transcribed off the same DNA strand. Mechanisms that actively isolate individual *Hox* genes from the influences of other nearby *Hox* gene enhancers have been found in the *Drosophila* HOM-C cluster; however, such elements do not appear to be present, as a general rule, in vertebrates. Additionally, in mouse, an entire *Hox* cluster is approximately the size of a single *Hox* gene, such as *Antp* in *Drosophila*. As a consequence of this close proximity, adjacent *Hox* gene promoters often appear to share or compete for enhancers. Transgenic mouse experiments exploring the regulatory landscape around *Hoxb4* and *Hoxb5* revealed that enhancer sharing, selectivity, and competition are all mechanisms regulating the expression of these two genes. Multiple elements between *Hoxb4* and *Hoxb5* participated in the regulation of both genes, which is largely responsible for overlapping *Hoxb4* and *Hoxb5* expression domains. One control element, equally capable of activating *Hoxb4* and *Hoxb5*, was preferentially recruited to

activate the *Hoxb4* promoter compared to that of *Hoxb5* when given a choice. Although few additional examples of promoter competition exist, either because of their rarity or because of their experimental in tractability, enhancer sharing appears to be common between neighboring *Hox* genes. In the case of *Hoxb3* and *Hoxb4*, at least two distinct elements in the intergenic region between these genes influence the expression of both genes. Similarly, shared elements have also been identified between *Hoxb5* and *Hoxb6*. The compactness of the *Hox* complexes and the absence of insulator elements blocking the influence of nearby regulatory elements suggest that enhancer sharing must be a common feature of *Hox* gene regulation.

Posttranscriptional Regulation of *Hox* Genes

Although mRNA levels and spatial/temporal localization are normally good indicators of protein distribution, it is becoming clear that they may be misleading in relation to Hox protein distribution. Because antibodies to Hox proteins have been particularly difficult to produce, expression analysis has often relied on *in situ* detection of mRNA. Several cases in which immunodetection of Hox protein has been achieved have demonstrated that protein expression domains often do not encompass the entire domain in which mRNA is expressed. Careful analysis of the regulation of the *Hoxb4* gene allowed the observation of a difference in reporter gene expression patterns dependent on whether the reporter utilized the endogenous *Hoxb4* 3′untranslated region (UTR) and polyadenylation signal or used an exogenous polyadenylation signal. Although the *Hoxb4* 3′ UTR was present in each of these reporter constructs, its inclusion into the mRNA transcript produced altered patterns of expression. This clearly demonstrated that *Hoxb4* was posttranscriptionally regulated and identified the domain responsible for this regulation. Similarly, mRNA expression and protein levels of some of the *Hoxc* genes in the spinal cord have been shown to be discordant, indicating that translational regulation may indeed be an additional mechanism modulating Hox expression.

Concurrent with these observations, microRNAs (miRNAs) (small regulatory RNAs that can mediate both pre- and posttranscriptional regulation through sequence-specific binding to target mRNAs) were being investigated as *Hox* regulatory molecules. Multiple miRNAs have been identified embedded within *Hox* complexes. mir-196a1/a2/b, which are located 5′ of the *Hox9* paralog in *Hox* B, C, and A clusters, respectively, are predicted to target *Hoxa7*, *Hoxb8*, *Hoxc8*, and *Hoxd8* transcripts for degradation. Verification of miRNA targets *in vivo* and assessment of their relevance has just begun; however, degradation products resulting from mir-196 targeting of *Hoxb8* have been isolated from mouse embryos, demonstrating that this process is functional *in vivo*. *Hox* genes have been identified as the targets of these and other miRNAs. Posttranscriptional regulation may be a consistent mechanism by which *Hox* genes are regulated; however, the importance of miRNAs in this process has not been demonstrated.

Global Regulatory Elements

By far the most elusive *Hox* gene regulatory process is the mechanism by which collinear expression of *Hox* genes is achieved. One hypothesis focuses on long-range elements outside of the complex (global regulatory elements) that mediate processive regulatory influences on *Hox* genes that travel along the DNA through the *Hox* complex. Experiments in which a *Hoxb1* reporter was inserted 5′ of *Hoxd13* showed that some aspects of the reporter expression were affected by the location of the reporter in the complex which were not seen when the reporter was inserted randomly, suggesting that location within the complex could influence expression pattern. Separate experiments showed both a sequential de-compaction of *Hox* complexes during development and partial collinearity of 'active' chromatin marks spreading through the *Hoxb* locus during differentiation, further supporting the global element plus spreading model. Although some elements capable of long-range influences on multiple *Hoxd* genes have been identified for the limb expression, global elements capable of long-distance, sequential regulation of entire complexes in most other sites of *Hox* expression have yet to be identified. In contrast, local elements capable of specific and ordered expression for many of the individual genes have been well characterized. Thus, it remains possible that many aspects of collinear expression could be achieved through the collective activity of locally positioned internal elements rather than through a single global modulator of each cluster. Clearly, if global control regions are a common feature regulating collinear *Hox* gene expression, they are not solely responsible for this phenomenon. Rather, it seems likely that diverse mechanisms are used to open the clusters to make them available for transcription that vary at different times and locations. A variety of locally recruited activators and repressors (including the Hox proteins) may work in concert with local and long-range chromatin remodeling complexes to impart the ordered domains of *Hox* expression.

Conclusions

Hox proteins play a fundamental role in the development of R/C identity in diverse tissues in addition to many less appreciated roles in organogenesis and cell differentiation. They participate in giving R/C segments of the hindbrain and spinal cord their unique identity through their nested expression patterns within these structures. The host of Hox proteins present at each axial level then conveys identity onto that segment observable by segment-specific gene expression patterns, neuronal phenotypes, and cell migratory behavior. Appreciation of Hox gene collinear expression patterns and their regulation are essential aspects of understanding mechanisms of R/C polarity within the hindbrain and spinal cord. *Hox* genes are regulated on multiple levels. Expression is initiated by transcription factors and secreted molecule signaling systems and is subsequently maintained via Hox-mediated auto- and cross-regulatory mechanisms. Experiments have begun to unveil an additional level of *Hox* gene regulation that functions posttranscriptionally; however, the full significance of this has yet to be determined.

A burgeoning area of exploration is the role of Hox cofactors and other interacting proteins. These experiments open the possibility of an additional level of modulation of Hox protein binding, activation/repression, and target site selection. Although much of these data have yet to be shown relevant *in vivo* during development, they could vastly expand our understanding of the role of Hox proteins and their mechanisms of action. More attention must be directed toward this avenue of investigation, but it promises to be an exciting future for Hox gene study.

Although it is clear that Hox proteins play a critical role in R/C patterning and identity, several questions remain about how this occurs. Altering *Hox* gene expression patterns within the hindbrain and spinal cord often results in dramatic alterations in cell identity or behavior; however, how Hox proteins regulate target genes to achieve these effects is largely unknown. A few downstream target genes activated or repressed by Hox proteins have been identified, but our understanding is too limited to create a picture of the pathways under Hox regulation that impose diverse regional characters. This is an exciting area for future study. Systems biology approaches and genomic analyses are all aimed at understanding how the alteration of one gene can impact on the outcome of cell behaviors to result in differences such as a leg versus an antenna. This is an elusive basis of morphogenesis, and understanding these mechanisms and their outcomes opens the possibility of a wealth of knowledge about the genetic programs that control normal development, how they are perturbed in human diseases, and how they might be involved in producing diversity. *Hox* genes have fundamental roles in these processes and provide an opportunity for exciting insight into the mechanisms responsible for building the basic body plan of animals.

See also: Anterior-Posterior Spinal Cord Patterning of the Motor Pool; Neural Patterning: Midbrain–Hindbrain Boundary; Retinoic Acid Signaling and Neural Patterning; Segmentation: Spinal Cord Segmentation and A–P Somite Patterning; Transcriptional Networks and the Spinal Cord; Wnt Pathway and Neural Patterning.

Further Reading

Akin ZN and Nazarali AJ (2005) Hox genes and their candidate downstream targets in the developing central nervous system. *Cellular and Molecular Neurobiology* 25: 697–741.

Bel-Vialar S, Itasaki N, and Krumlauf R (2002) Initiating Hox gene expression: In the early chick neural tube differential sensitivity to FGF and RA signaling subdivides the HoxB genes in two distinct groups. *Development* 129: 5103–5115.

Brend T, Gilthorpe J, Summerbell D, and Rigby P (2003) Multiple levels of transcriptional and post-transcriptional regulation are required to define the domain of Hoxb4 expression. *Development* 130: 2717–2728.

Chambeyron S and Bickmore WA (2004) Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & Development* 18: 1119–1130.

Chan SK, Ryoo HD, Gould A, Krumlauf R, and Mann RS (1997) Switching the *in vivo* specificity of a minimal Hox-responsive element. *Development* 124: 2007–2014.

Chang CP, Brocchieri L, Shen WF, Largman C, and Cleary ML (1996) Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. *Molecular and Cellular Biology* 16: 1734–1745.

Dasen JS, Tice BC, Brenner-Morton S, and Jessell TM (2005) A Hox regulatory network establishes motor neuron pool identity and target-muscle connectivity. *Cell* 123: 477–491.

Diez del Corral R and Storey KG (2004) Opposing FGF and retinoid pathways: A signalling switch that controls differentiation and patterning onset in the extending vertebrate body axis. *Bioessays* 26: 857–869.

Gavalas A and Krumlauf R (2000) Retinoid signalling and hindbrain patterning. *Current Opinions in Genetics and Development* 10: 380–386.

Hunt P, Gulisano M, Cook M, et al. (1991) A distinct Hox code for the branchial region of the vertebrate head. *Nature* 353: 861–864.

Lee T, Jenner R, Boyer L, et al. (2006) Control of developmental regulators by polycomb in human embryonic stem cells. *Cell* 125: 301–313.

Maconochie M, Nonchev S, Morrison A, and Krumlauf R (1996) Paralogous Hox genes: Function and regulation. *Annual Reviews in Genetics* 30: 529–556.

Mangelsdorf DJ and Evans RM (1995) The RXR heterodimers and orphan receptors. *Cell* 83: 841–850.

Mann RS and Chan SK (1996) Extra specificity from extradenticle: The partnership between HOX and PBX/EXD homeodomain proteins. *Trends in Genetics* 17: 258–262.

Moens CB and Selleri L (2006) Hox cofactors in vertebrate development. *Developmental Biology* 291: 193–206.

Niederreither K, Subbarayan V, Dolle P, and Chambon P (1999) Embryonic retinoic acid synthesis is essential for early mouse post-implantation development. *Nature Genetics* 21: 444–448.

Nolte C and Krumlauf R (2006) Expression of Hox genes in the nervous system of vertebrates. In: Papageorgiou S (ed.) *Hox Gene Expression*, pp. 14–41. Austin, TX: Landes Bioscience.

Pasquinelli AE, Hunter S, and Bracht J (2005) MicroRNAs: A developing story. *Current Opinions in Genetics and Development* 15: 200–205.

Pearson JC, Lemons D, and McGinnis W (2005) Modulating Hox gene functions during animal body patterning. *Nature Reviews Genetics* 6: 893–904.

Sharpe J, Nonchev S, Gould A, Whiting J, and Krumlauf R (1998) Selectivity, sharing and competitive interactions in the regulation of Hoxb genes. *European Molecular Biology Organization Journal* 17: 1788–1798.

Yekta S, Shin IH, and Bartel DP (2004) MicroRNA-directed cleavage of HOXB8 mRNA. *Science* 304: 594–596.

Relevant Websites

http://www.eurekah.com – Eureka Bioscience Database.

http://research.nhgri.nih.gov – National Human Genome Research Institute Homeodomain database.

http://www.iephb.nw.ru – The Hox Pro database managed by Alex Spirov.

Human Depth Electrodes

I V Viskontas, University of California at Los Angeles, Los Angeles, CA, USA
I Fried, University of California at Los Angeles, Los Angeles, CA, USA; and Tel-Aviv Medical Center and Tel-Aviv University, Tel-Aviv, Israel

© 2009 Elsevier Ltd. All rights reserved.

The language of the nervous system and its building blocks, the neurons, is that of changing electrical potentials. The opportunity to record this activity directly inside the human brain is, however, limited to a few clinical situations. The newer methods of neural imaging, such as functional magnetic resonance imaging (fMRI), though immensely popular, are indirect ways of imaging neuronal activity and are based on measurements of regional hemodynamic changes in the brain. The major limitations of the fMRI technique are its spatial and temporal resolutions: Although fMRI has provided a wealth of information about brain regions, it cannot demonstrate how individual neurons and small neuronal assemblies contribute to cognitive processes, and it lacks the millisecond resolution which is often important in probing the timing of neuronal activity. This article describes the contributions of a technique that allows investigators to observe the activity of individual neurons deep within the human brain. This relatively small area of scientific investigation is situated between animal extracellular neurophysiology, a technique that allows investigators to observe single neuron activity during behavior, and the human neuroimaging techniques that provide images of neuronal activity from whole brains. Extracellular neurophysiology in animals, however, cannot inform science about cognitive processes that are unique to humans and accessible via conscious report, whereas fMRI has limited resolution and is an indirect measure of neural activity. Depth electrode recordings in humans have two primary advantages: unlike fMRI, recordings have excellent spatial and temporal resolution and they provide a direct measure of neuronal activity; unlike animal extracellular neurophysiology, recordings are performed on conscious subjects who can describe their experiences.

Methodological Considerations

Invasive depth electrode recordings are only performed for clinical reasons, predominantly in patients with movement disorders (e.g., Parkinson's disease (PD)) or pharmacologically resistant epilepsy. In these contexts, depth recordings are used to diagnose disorder subtypes and define appropriate treatment in particularly difficult cases. Today's electrodes can measure both local field potentials, which provide information on the averaged dendrosomatic synaptic activity of large cell assemblies, and action potentials of individual neurons isolated with embedded microelectrodes. Both types of information are useful, and each is discussed in this article. Depth electrode recordings in humans, then, allow investigators to compare neural activity in human brains with the wealth of information garnered from recordings in other animals and to evaluate the validity of animal models with respect to human brain function.

Depth Electrodes in the Basal Ganglia and Related Structures

We begin with a discussion of the major findings stemming from direct recordings in patients with movement disorders. Today, the most common use of depth electrode recordings in the context of movement disorders is to localize the target sites for therapeutic lesions or, more commonly, for the placement of a deep-brain stimulating (DBS) electrode. In patients with PD, DBS electrodes are used to inhibit the overactive cells in the globus pallidus (GP) or subthalamic nucleus (STN) that are normally inhibited by dopaminergic cells in the substantia nigra. Since PD is caused mainly by the degeneration of the substantia nigra, DBS electrodes relieve symptoms by acting directly on the GP or STN. In patients with PD, surgeons may implant recording electrodes in the GP or STN to aid the implantation of DBS electrodes. While the surgeons are targeting these two regions, they have the capability of observing neural activity at the target sites. Patients remain fully conscious throughout the surgery, allowing surgeons to test their motor functions and to observe the effects of electrical stimulation in the targeted regions.

Local Field Potentials Recorded from the Basal Ganglia and Related Structures

In one study using depth electrode recordings, in patients who were not on medication and were in the akinetic 'off' state (i.e., they were in the process of showing symptoms of PD), neural activity coherence between the basal ganglia structures and other cortical regions was found to be dominated by tremor (3–8 Hz) and beta-band (15–30 Hz) frequencies. In previous studies using noninvasive recordings, investigators had failed to find the putative source of neural activity in the tremor frequencies, thought to underlie the resting tremor seen in PD patients.

Furthermore, stimulation of those sites where beta-band coherence was the highest yielded the greatest reduction in PD symptoms. Similar approaches have allowed investigators to evaluate the effects of certain drug treatments on these coherence peaks in the hope of finding new pharmacological treatments for PD and other movement disorders.

Insights from Single-Unit Recordings in Subcortical Structures

Recordings from single units in the STN in particular provide a unique complement to electrophysiology in animals. For example, single-unit depth electrode recordings in humans have delineated the somatotopic organization of the STN, previously only described in animals. In addition, the cellular correlates of resting tremor have also been described: cells that show synchronous rhythmical discharges associated with tremor have been found in the external pallidum, GP and STN, and motor thalamic nuclei. These findings support a model in which the tremors that characterize PD are driven by a central pacemaker which may be modulated by sensory inputs, thereby providing a model for how sensory stimuli may alleviate PD symptoms.

Single-cell recordings have also provided support for a 'dynamic' model of basal ganglia function. This model posits that neurons oscillate at tremor frequencies and in the beta band and that these firing patterns tend to be synchronized along the basal ganglia loop, leading to abnormally patterned firing in motor cortex. The model further suggests that the normal dopaminergic system supports a functional segregation of subcircuits in the basal ganglia. A breakdown of this independent processing leads to abnormal temporal coupling and the emergence of the symptoms of PD. The existence of such oscillatory neurons and the synchrony in their firing patterns has been confirmed using invasive depth electrode recordings.

Combining the findings from local field potentials (LFPs) and single units in the basal ganglia has resulted in a model in which the basal ganglia, thalamus, and cortex interact in different frequency bands, modulating basal ganglia functions in a task- and state-dependent way. For example, slow oscillations at tremor frequencies or in the beta band that result from dopamine depletion may disrupt normal motor function.

Future Directions

As our understanding of the neural basis of movement disorders such as PD becomes more complete, more effective treatments can be devised. Until then, however, depth electrode recordings in these regions have provided critical information concerning the functions of brain regions underlying movement. Future studies may include the application of DBS to other diseases, such as major depression, obsessive–compulsive disorder, and epilepsy, as invasive electrode recordings uncover the ways in which DBS can affect brain function. Moreover, modified DBS could become effective treatment for people afflicted with paralysis and motor coordination problems.

Depth Electrode Recordings in Patients with Pharmacologically Resistant Epilepsy

Patients with intractable epilepsy can benefit from surgical resection of the small area implicated in seizure onset. If scalp electroencephalogram (EEG) recordings fail to definitively localize the seizure focus, patients may be implanted with depth electrodes for further monitoring. Most of these patients have foci in the medial temporal lobe (MTL), which includes the hippocampus, amygdala, entorhinal and posterior parahippocampal cortices, as well as other extratemporal lobe sites. Electrodes are generally implanted bilaterally so that physicians can monitor the spread of seizures and thus confirm the putative focus. Once implanted, patients are monitored on the ward for several days and sometimes weeks. During this time, patients may consent to participate in behavioral experiments that provide invaluable information about single- and multiunit activity in many aspects of brain function. These neurons can also be monitored during different states of consciousness, such as sleep or during seizures. In general, several types of electrophysiological data can be obtained from depth electrodes. Most depth electrodes used in epilepsy have contacts for recording of intracranial EEG along the shaft of the electrodes at distances of 5–10 mm between adjacent contacts. If behavioral paradigms are used as well, event-related potentials (ERPs) can then be recorded with respect to timed stimuli or behavioral events. In some centers, depth electrodes are also fashioned with the capability of recording local field potentials and multi- and single-unit activity via microelectrodes, usually microwires, which are situated either at the tip of the electrode or along the shaft of the electrode (Figure 1).

In addition to chronic recordings with depth electrodes, EEG and even single-unit activity can be recorded acutely during epilepsy surgery via electrodes placed on the surface of the cortex or inserted into the cortex. Acute single- and multiunit recordings in conjunction with language and memory tasks have been performed during epilepsy surgery under local anesthesia, mainly by Ojemann and colleagues. This

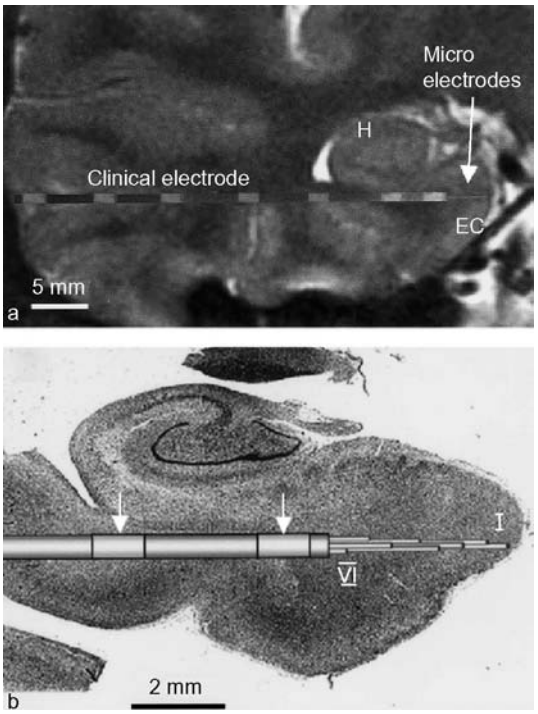

Figure 1 Placement of depth electrodes. Some electrodes are fitted with microwires that are capable of recording electrical activity at the level of single neurons. (a) MRI and CT scans may be fused in order to localize the anatomical placement of electrodes. (b) A Nissl-stained image of the MTL shows how protruding microwires may be arranged. Reproduced from Bragin A, Wilson CL, Staba RJ, Reddick M, Fried I, and Engel J Jr. (2002) Interictal high frequency oscillations (80–500 Hz) in the human epileptic brain: Entorhinal cortex. *Annals of Neurology* 52: 407–415, with permission.

article, however, focuses on data obtained from implantation of depth electrodes for chronic recordings over a period of several days while patients are monitored in a hospital ward environment for spontaneously occurring seizures and can perform behavioral tasks outside the acute environment of the operating room.

LFPs in Patients with Epilepsy

In addition to single- and multiunit activity, depth electrodes in the MTL can be used to record LFPs or, when observed during behavior, ERPs. These ERPs are commonly recorded from EEG contacts along the shaft of the electrode – the same contacts that are also used to monitor seizure onset and spread patterns. Historically, ERP studies have used noninvasive scalp recordings in healthy participants to correlate behavior with changes in neural activity. Many of these studies have investigated changes in scalp ERPs during memory processes. One reliable finding is the presence of the P3 component (the third positive peak after stimulus onset, also called P300 to denote that it occurs approximately 300 ms after stimulus onset) thought to represent the end of a salient event and the updating of information in working memory. The origin of this component, however, cannot be precisely defined using scalp EEG. Instead, depth electrode recordings have shown that P3 is generated in the hippocampus, superior temporal sulcus, and the amygdala, indicating that polymodal association areas interact with the limbic and memory systems in order to produce this marker of working memory function. A similar methodology has been used to investigate the origin of averaged ERPs as it reflects the activity of underlying neural assemblies. Depth electrode recordings have shown that two processes operate to generate late ERP components in cognitive tasks: event-related activation of neural assemblies and phase resetting of ongoing activity. These two mechanisms may serve cognitive processes in different ways, and they need to be further investigated as potentially independent signals.

LFP insights into cognitive processes Another body of work has shown that the formation of declarative memories, or consciously retrievable memories for facts and events, correlates with phase synchronization of rhinal–hippocampal EEG activity in the gamma range (~40 Hz) and an increased EEG coherence in the low-frequency range. Specifically, the process generated by the rhinal cortex, which occurs within 300 ms of stimulus onset, precedes the hippocampal process, occurring approximately 300 ms later. Furthermore, the rhinal process correlates with semantic preprocessing, whereas the hippocampal process correlates with a purely mnemonic operation. Current knowledge about phase synchrony in gamma activity suggests that synchronization facilitates communication between the rhinal cortex and the hippocampus, whereas decoupling terminates this communication. This phase synchrony may accomplish Hebbian synaptic plasticity, as neurons that fire in synchrony become associated, and establish the initial memory trace.

Another controversy in neuroscience has been the precise function of theta (3–8 Hz) oscillations in the MTLs. A hypothesis concerning the function of theta oscillations, or neural synchrony in general, is known as the sensorimotor integration hypothesis, which suggests that theta oscillations provide constantly updated feedback from sensory regions to voluntary motor regions to ensure correct motor output and performance of a task. Findings from depth electrodes have shown that theta oscillations in the neocortex increased when patients were actively engaged in a virtual reality game, supporting the sensorimotor integration hypothesis.

Findings from Single-Unit Recordings in the Human MTL and Other Regions

In a small number of hospitals, depth electrodes are fitted with microelectrode wires (40 μm) protruding from the tips, allowing investigators to record individual cells. Recordings from single units deep within various regions have provided insights into various aspects of brain function, including basic mechanisms of epilepsy, sleep, memory, perception, and other aspects of cognition and motor function. Because many of the patients undergoing implantation of depth electrodes are suspected of having temporal lobe epilepsy originating in the MTL, the MTL has been most accessible to investigations at the single-unit level.

Insights into basic mechanisms of epilepsy and sleep behavior From a clinical standpoint, recordings from isolated single neurons have provided important information regarding seizure physiology. For example, neurons in epileptogenic regions show higher baseline firing rates, more bursting and stronger synchrony during slow-wave sleep, yet show no differences from cells in nonepileptogenic regions during awake behaviors. It may be that during wakefulness, epileptogenic neurons show higher firing rates and more bursting when there is synchronous firing of neighboring cells, indicating that there may be a difference in synchrony during sleep and wakefulness that influences seizure activity.

In addition, rodent epilepsy models suggest a central role for fast ripple oscillations (200–500 Hz) in seizure activity. In humans, investigators have observed both fast ripple oscillations and ripples (80–200 Hz) chiefly during non-rapid-eye-movement sleep (NREM). Whereas ripples represent normal activity in the human hippocampus, fast ripple oscillations have been shown to be more prevalent in epileptogenic areas than in healthy tissue. Furthermore, the highest rates of fast ripple oscillations have been found during NREM sleep. These findings support the view that fast ripple oscillations are the product of pathological neuronal hypersynchronization associated with seizure-generating areas.

Insights into memory function Studies of patients with epilepsy have provided some of the most significant findings concerning the neural basis of memory. Beginning with the patient H.M., who had bilateral MTL resection to treat his seizures, numerous studies have identified the critical role that this brain region plays in consciously accessible memory for facts and events. Therefore, invasive depth electrode recordings in the MTL can provide information about the manner by which memories are represented at the level of individual neurons.

Early single-unit studies found that cells in the hippocampus responded with changes in firing patterns to specific items during word or picture recognition tasks. During repeated performance of the task, however, many of these cells changed their 'preferred' stimulus, suggesting that neurons are flexible in terms of their response patterns and may only temporarily represent an item during learning or encoding. These early studies also failed to find a correlation between individual firing patterns and performance.

As the technology used to record cells improved, studies have dramatically increased sample sizes, and a link between neural firing and memory performance has indeed been made. In the hippocampus, for instance, responses during encoding have been shown to predict whether patients subsequently remembered word pairs, whereas in the entorhinal cortex correct performance was predicted by activity during retrieval. Depth electrode recordings have also shown that neurons in the MTL are capable of detecting novelty or familiarity immediately on the second presentation of a stimulus. In addition, a substantial number of cells in the hippocampus and amygdala respond to conjunctions of features rather than individual stimuli. In order to interpret these results, a discussion of current models of MTL function in memory is warranted.

Memories are composed of a set of associations between different elements in the original sensory representation. The pattern of activation that underlies the set of associations is strengthened as the elements of which it is composed 'fire together' and create Hebbian-style synapses (named for the neuroscientist who first envisioned the process). During retrieval, when a subset of elements is reactivated, the entire pattern is retrieved, strengthening associations between elements. The number of connections that this model requires between elements quickly escalates to a biological implausibility. Furthermore, as new associations are created, they may disrupt existing ones or result in overlapping representations. If the number of overlapping representations becomes large, the reactivation of a particular pattern may no longer be unique. This situation is called catastrophic interference. Given the problem of catastrophic interference, the models propose that a separate processor, namely the hippocampus, could rapidly store the events and facilitate gradual transfer of these memories to the neocortex. To enable this rapid storage and avoid interference, it has been proposed that only a very small subset of cells encode any given element; this is called sparse coding. Strikingly, in one study, a few cells in both the hippocampus and the parahippocampal cortex responded to novel stimuli with an

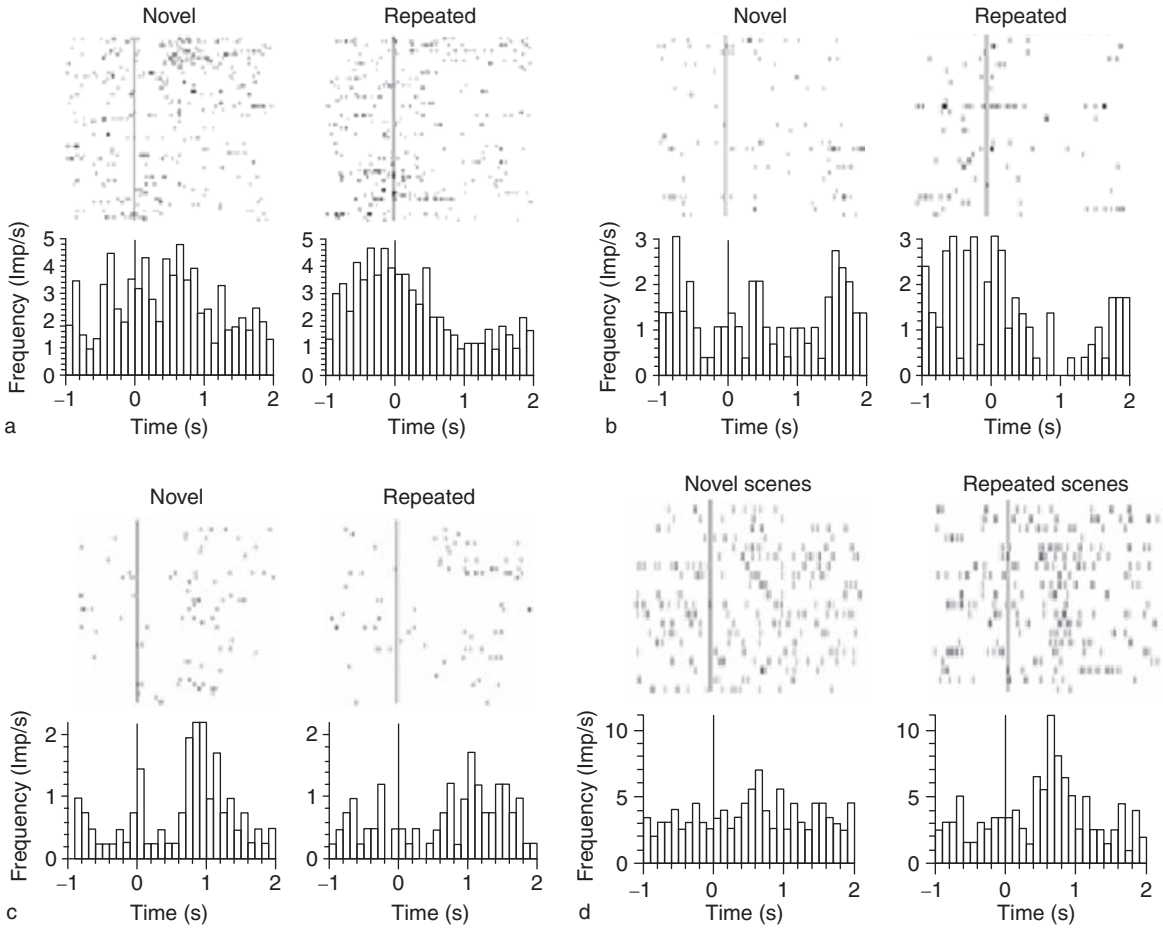

Figure 2 Cells coding mnemonic properties of stimuli. This figure shows raster plots and firing frequency histograms for four individual neurons sorted by the mnemonic property of the stimuli. (a and b) Responses of two hippocampal cells that decrease firing from baseline to repeated stimuli. (c) The response of a novelty preferring cell in the entorhinal cortex and (d) the cell in the entorhinal cortex that is sharply tuned to the second presentation of scenes. Reproduced from Viskontas IV, Knowlton BJ, Steinmetz PN, and Fried I (2006) Differences in mnemonic processing by neurons in the human hippocampus and parahippocampal region. *Journal of Cognitive Neuroscience* 18: 1654–1662, with permission.

increase in firing (excitation). However, on the second presentation of a stimulus, neurons in these regions already showed very different firing patterns; whereas in the parahippocampal region there was a dramatic decrease in the number of cells responding to the stimuli, in the hippocampus a large subset of neurons showed inhibitory responses (firing decreased from baseline) (Figure 2). Therefore, in line with this view, wide-scale inhibition observed by depth electrode recordings in the hippocampus may provide a mechanism by which cells in the human hippocampus increase the signal-to-noise ratio, allowing those few cells that represent a given item to be 'heard' by higher cognitive regions, thereby supporting sparse coding in memory processes.

Another role for the hippocampus: Human spatial navigation The discovery of 'place cells' in the rodent hippocampus led to the theory that this structure is involved in spatial navigation. Place cells respond with a change in their firing patterns when the animal is in a particular location in space. Single-unit depth electrode recordings in humans have confirmed the presence of place cells in the human hippocampus as well (Figure 3). In these studies, patients engaged in a virtual-reality computer game in which they could 'move' through a town. Investigators found that certain cells responded in highly reproducible ways when the patient was 'in' a specific location in the town. Importantly, place cells were more likely to be in the hippocampus than in the surrounding parahippocampal cortex, and a subset of place cells were modulated by the patient's goal in navigation. For example, some cells would only respond with excitation if the patient was in a particular place and had a specific goal in mind, such as locating a virtual person. By contrast, cells in the parahippocampal cortex were more likely to respond

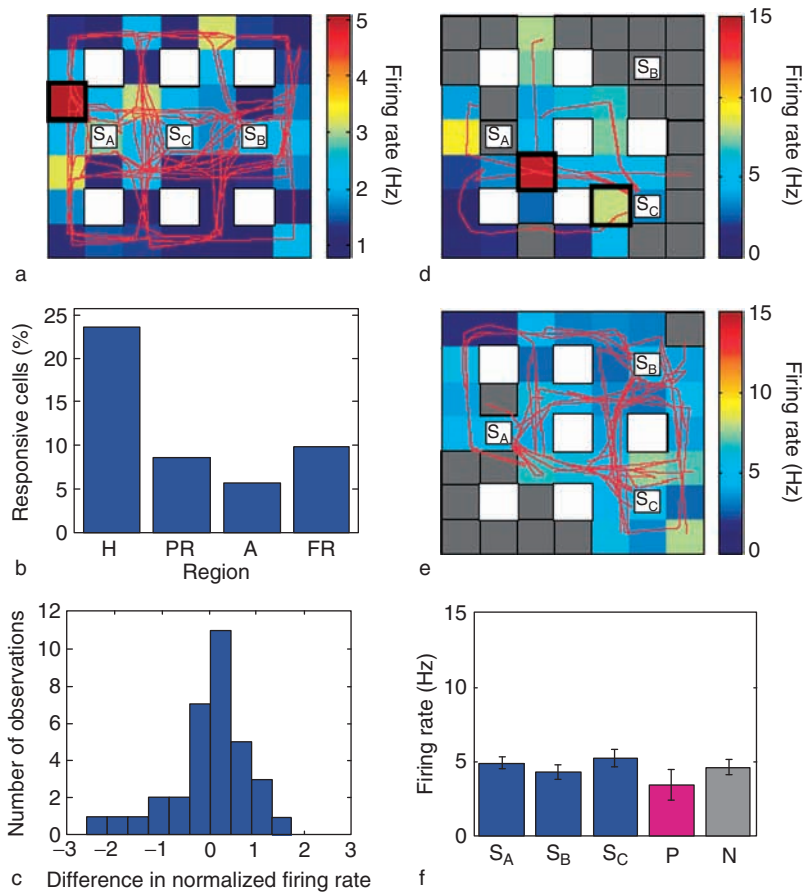

Figure 3 'Place' cells. (a) Firing rate map of a hippocampal cell showing significant place selectivity. Red lines indicate the subject's trajectory, and black-rimmed squares indicate regions of significantly high firing rate (all examples, $p = 0.01$). (b) Place-responsive cells were clustered in the hippocampus (H) compared with amygdala (A), parahippocampal region (PR), and frontal lobes (FRS). (c) Regions of high firing included high numbers of traversals in different directions. The distribution of firing rate differences across these traversals was centered on zero and normal. (d) Firing rate map of a right hippocampal cell showing significant place selectivity when searching for shop SC but no such specificity when searching for other goals. (e) Areas with two traversals were excluded. (f) This cell similarly showed no effect of viewing specific targets. P, viewing passengers; N, a control background view. Reproduced from Ekstrom AD, Kahana MJ, Caplan JB, et al. (2003) Cellular networks underlying human spatial navigation. *Nature* 425: 184–188, with permission.

predictably when the patient was viewing a specific landmark, even if that landmark was viewed independently of the person's 'location' in the game. Thus, these results confirm the presence of place cells in the hippocampus but also underscore another functional difference between the hippocampus and its surrounding cortices. Moreover, findings from this study suggest that although some aspects of MTL function in humans are highly similar to those in nonhuman animals, there are important differences which single-unit recordings in humans can help define.

Visual perception Another remarkable finding that has been reported from depth electrode recordings in human MTL is the existence of category-specific cells and cells that respond to a particular concept, such as a person or a landmark (e.g., Bill Clinton or the Sydney Opera House), even when the sensory stimuli that represent the concept are dissimilar (e.g., a pictorial and a semantic cue). In the case of a category such as specificity, neurons were found that responded exclusively to visual images of animals, human faces, outdoor scenes, or inanimate objects such as cars or foods. Even though the actual pictures that represented the category items were quite dissimilar, the neurons were capable of separating and classifying the percepts, developing firing patterns specific to a single category. Furthermore, the investigators found that category-specific cells would respond even when the patients were asked to 'imagine' a stimulus within the cell's preferred category. This finding, as well as findings from another study using flash suppression, suggest that MTL cells respond to a percept rather than to the image falling on the retina.

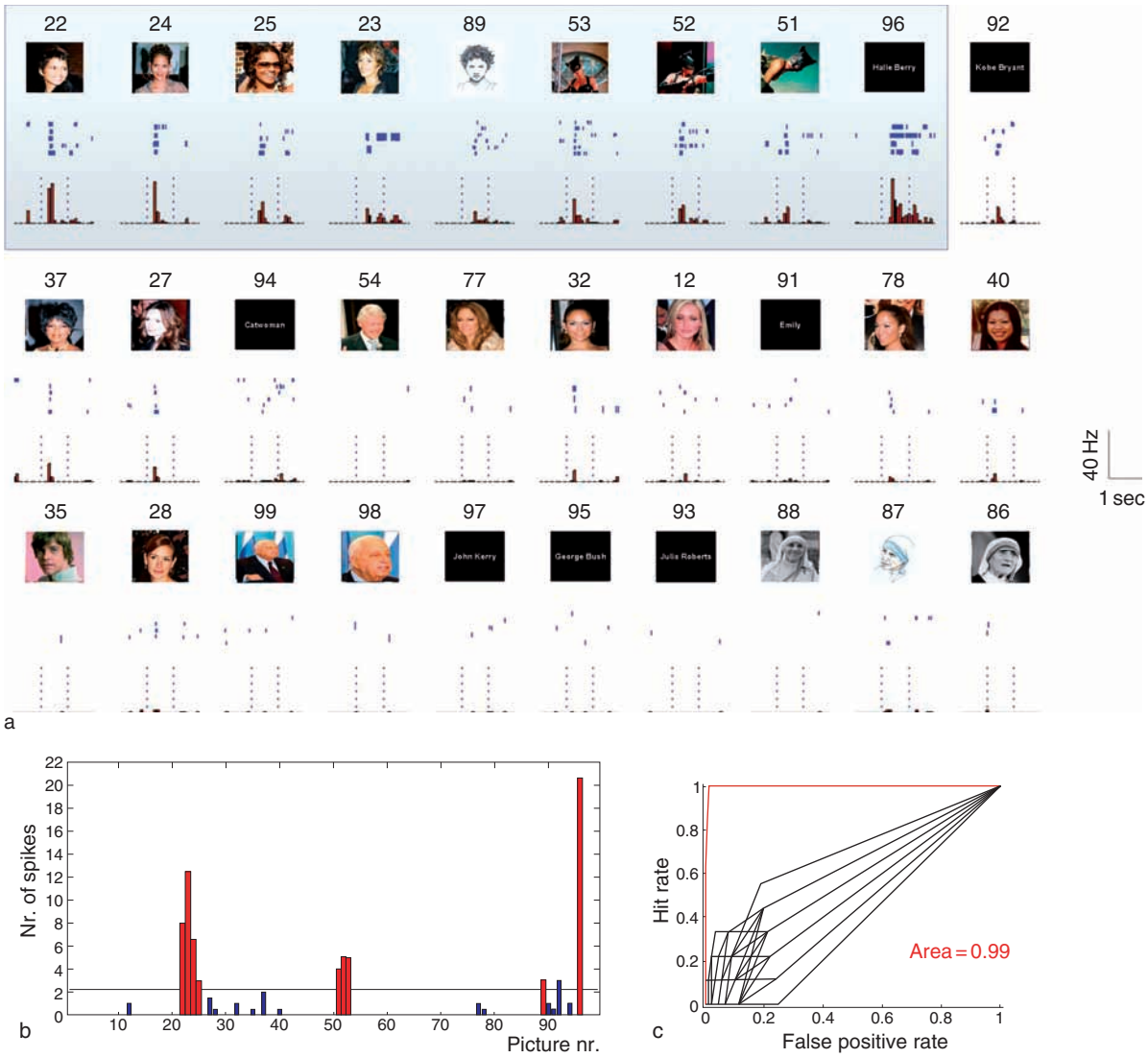

Figure 4 Cell responding to images and the name of Halle Berry. This cell showed an increase from baseline firing when the patient viewed images of Halle Berry but not other celebrities. Reproduced from Quiroga RQ, Reddy L, Kreiman G, Koch C, and Fried I (2005) Invariant visual representation by single neurons in the human brain. *Nature* 435: 1102–1107, with permission.

In addition to a remarkable specificity of responses, such as the ability to respond preferentially to a particular person but not to hundreds of other people, MTL cells also exhibit invariance in their responses; that is, they show the same response to strikingly different exemplars of the same preferred stimulus (Figure 4). This property suggests that single neurons code for an abstraction of the stimulus rather than its physical characteristics.

Perception of speech and music Intracranial depth electrode recordings have also been used in investigations of auditory perception. Neurons in the primary auditory cortices have been shown to respond to human speech, music, and other sounds. For example, auditory-evoked potentials in the auditory cortex (Heschl's gyrus) have been shown to display greater phase-locked synchronous activity during the perception of dissonant chords, such as minor and major seconds, than during the perception of consonant chords, such as octaves and major fifths, thus replicating findings in monkeys. Ojemann and colleagues have also used single- and multiunit recordings in the operating room to show specific neuronal responses to language, music, and visuospatial stimuli in temporal neocortex.

Findings related to emotional processes In addition to models of cognitive processes, depth electrode recordings have also been applied to the study of the neural

basis of emotional processes. Two independent groups have reported that stimulation of the pre-supplementary motor area via depth electrodes results in feelings of mirth and laughter. In another study of emotional processes, investigators found that individual neurons in the ventral and medial prefrontal cortex, in both hemispheres, responded selectively to aversive, pleasant, or neutral stimuli, yet most of the neurons that showed responses in these regions preferred aversive stimuli. At the local field potential level, investigators using depth electrode recordings have also found a preference for aversive stimuli in the amygdala – a finding that supports a large body of animal work implicating the amygdala in fear responses.

Motor planning and execution Occasionally, depth electrodes need to be placed in the medial frontal lobe. Placement in this region represents an opportunity to record neuronal activity related to motor planning. Recording from medial frontal cortex, investigators have found neurons that responded selectively to specific features of a motor plan, such as which hand performed the motor activity or the complexity of the sequence. These results support the notion that the medial frontal cortex is involved in the temporal organization and planning of movement in humans, and that individual neurons can distinguish different aspects of a particular actions.

Investigating the relationship between the blood oxygen level-dependent signal and neuronal activity Another important use for depth electrode recordings has been the investigation of the extent to which the blood oxygen level-dependent (BOLD) response used in fMRI correlates with actual neural activity. fMRI uses a measure of cerebral blood flow and the assumption that neural activity is accompanied by increased blood flow to the active area. The extent to which this assumption is valid in the human brain, however, remains under debate. Depth electrode recordings in the human auditory cortex have shown that the BOLD signal is in fact highly correlated with activity even at the level of single neurons. Data from this study showed that when patients viewed a 9 min movie clip several times, neurons in the auditory cortex (Heschl's gyrus) showed reproducible firing patterns in response to the auditory stimulation. Furthermore, the fMRI BOLD signal that was recorded in the auditory cortex of a different set of healthy participants who watched the same movie clip showed a high correlation with both activity recorded from signal neurons and high-frequency local field potentials in the same region. This study demonstrates the value of depth electrode recordings in bridging the gap between neuronal activity and the signals measured by fMRI. Additional studies using recordings from other brain regions should be performed to further investigate these issues, because the BOLD signal and the neural code underlying various cognitive processes may vary among different-regions. For instance, correlation between single-unit activity, LFPs, and the BOLD signal may be different for regions and cognitive processes dominated by sparse versus distributed coding.

Conclusions and Future Directions

Depth electrode recordings represent a unique approach to understanding how neural activity supports human brain function at multiple levels. It is important to keep in mind, however, that these studies are performed in patients with neurological disorders. Nevertheless, they represent the only opportunity in modern-day neuroscience to probe human cognitive function at the level of the single neuron. Therefore, it is important to use this rare opportunity in conjunction with paradigms which are based on well-established findings from the animal and nonhuman primate literature in particular. At the same time, the opportunity to record neural activity directly in humans challenges us to extend our investigations to include questions which cannot be answered using animal studies, especially those paradigms in which conscious behavioral reports are invaluable. Although experimental design should be comparable to work in animal neurophysiology, investigators should also consider findings from studies performed in humans using fMRI, scalp-recorded ERPs, and magnetoencephalography.

The opportunities to record intracranial signals at the single neuron and neuronal assembly level include mainly the surgical treatment epilepsy and movement disorders. The rapidly evolving field of clinical investigation of brain–machine interfaces may, however, provide more opportunities. Already, several studies point to the feasibility of decoding movement and percepts from neural activity at the single- and multi-unit level and at the level of LFPs. Closed-loop devices utilizing intracranial EEG to deliver targeted stimulation to stop epileptic seizures are also under development and in clinical trials. Therefore, the role of brain-derived signals from depth electrodes is likely to play an important role not only in understanding basic mechanisms of cognition but also in designing neural prostheses and other clinically relevant devices.

See also: Epilepsy; Event-Related Potentials (ERPs) and Cognitive Processing; fMRI: BOLD Contrast; Hippocampus; Magnetic Resonance Spectroscopy; Neuroimaging; Perfusion MRI; Positron Emission Tomography (PET).

Further Reading

Benabid AL, Koudsie A, Benazzouz A, et al. (2000) Subthalamic stimulation for Parkinson's disease. *Archives of Medical Research* 21: 282–289.

Bragin A, Wilson CL, Staba RJ, Reddick M, Fried I, and Engel J, Jr. (2002) Interictal high frequency oscillations (80–500 Hz) in the human epileptic brain: Entorhinal cortex. *Annals of Neurology* 52: 407–415.

Brugge JF, Volkov IO, Garell PC, Reale RA, and Howard MA (2003) Functional connections between auditory cortex on Heschl's gyrus and on the lateral superior temporal gyrus in humans. *Journal of Neurophysiology* 90: 3750–3763.

Crick F, Koch C, Kreiman G, and Fried I (2004) Consciousness and neurosurgery. *Neurosurgery* 55: 273–282.

Ekstrom AD, Kahana MJ, Caplan JB, et al. (2003) Cellular networks underlying human spatial navigation. *Nature* 425: 184–188.

Engel AK, Moll CKE, Fried I, and Ojemann GA (2005) Invasive recordings from the human brain: Clinical insights and beyond. *Nature Reviews Neuroscience* 6: 35–47.

Fell J, Klaver P, Elger CE, and Fernandez G (2002) The interaction of rhinal cortex and hippocampus in human declarative memory formation. *Reviews in the Neurosciences* 12: 299–312.

Fishman YI, Volkov IO, Noh MD, et al. (2001) Consonance and dissonance of musical chords: Neural correlates in auditory cortex of monkeys and humans. *Journal of Neurophysiology* 86: 2761–2788.

Fried I, MacDonald KA, and Wilson CL (1997) Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron* 18: 753–765.

Kreiman G, Koch C, and Fried I (2000) Imagery neurons in the human brain. *Nature* 16: 357–361.

McCormick DA and Contreras D (2001) On the cellular and network bases of epileptic seizures. *Annual Review of Physiology* 63: 815–846.

Mukamel R, Gelbard H, Arieli A, Hasson U, Fried I, and Malach R (2005) Coupling between neuronal firing, field potentials, and fMRI in human auditory cortex. *Science* 309: 951–954.

Ojemann GA, Ojemann SA, and Fried I (1998) Lessons from the human brain: Neuronal activity related to cognition. *Neuroscientist* 4: 285–300.

Quiroga RQ, Reddy L, Kreiman G, Koch C, and Fried I (2005) Invariant visual representation by single neurons in the human brain. *Nature* 435: 1102–1107.

Rosenow JM, Mogilnert AY, Ahmed A, and Rezai AR (2004) Deep brain stimulation for movement disorders. *Neurology Research* 26: 9–20.

Rutishauser U, Mamelak AN, and Schuman EM (2006) Single-trial learning of novel stimuli by individual neurons of the human hippocampus–amygdala complex. *Neuron* 49: 805–813.

Staba RJ, Wilson CL, Bragin A, Fried I, and Engel J, Jr. (2002) Sleep states differentiate single neuron activity recorded from human epileptic hippocampus, entorhinal cortex, and subiculum. *Journal of Neuroscience* 22: 5694–5704.

Viskontas IV, Knowlton BJ, Steinmetz PN, and Fried I (2006) Differences in mnemonic processing by neurons in the human hippocampus and parahippocampal region. *Journal of Cognitive Neuroscience* 18: 1654–1662.

Human Haptics

S J Lederman, Queen's University, Kingston, ON, Canada
R L Klatzky, Carnegie Mellon University, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Haptic Perception of Surface and Object Properties

People use their hands to obtain information about a variety of physical properties of surfaces and objects. They are particularly effective at haptically processing dimensions that pertain to the materials used in the construction of surfaces and objects. The most notable dimensions relate to texture, compliance, thermal qualities, and weight, the last being a hybrid of material and geometry. People are considerably less effective at haptically evaluating and discriminating an object's geometric features, such as orientation, shape, and size. In the following sections, we discuss research on the principal dimensions of haptic perception.

Roughness

To date, most research on haptic texture perception has focused on the perception of surface roughness. Provided there is relative motion between skin and surface and that both speed and force are equated, roughness perception is equivalent whether the hand is stationary and the surface moves or vice versa. Thus, cutaneous information is critical for perceiving roughness. Perceived roughness of very fine surfaces (i.e., those with spatial periods of less than ~200 μm in size) is likely based on intensive coding of vibration by Pacinian or fast adapting type II (FAII) peripheral tactile units. For spatial periods ranging from ~200 μm to 2 mm, perceived roughness is neurally represented by the stochastic variation in the spike rates of slow-adapting type I units (SAI) innervated by Merkel complexes. Beyond a spatial period of ~2 mm, perceived roughness is processed spatially by SAI tactile units.

Compliance

Considerably less is known about the perception of compliance or other related physical properties, such as elasticity and viscosity. Early research investigated how expert bakers perceived four primary dimensions of dough: stickiness, elasticity, firmness or toughness, and extensibility (i.e., the extent to which dough may be stretched before it breaks). The bakers were most sensitive to variation in stickiness, an attribute judged to be quite unpleasant; however, they were relatively insensitive to changes in either elasticity or viscosity.

More recent research has shown that cutaneous information alone is both necessary and sufficient to judge the relative softness of rubber specimens with deformable surfaces. The discharge rate of SAI units increases with increasing average pressure during indentation and, thus, could serve as a peripheral neural code for perceived softness independent of the velocity of indentation. Because single-fiber response depends on indentation velocity as well as on surface compliance, this code must be based on a population response, particularly in SAI units. To differentiate the stiffness of compliant objects with planar rigid surfaces, cutaneous cues on their own are insufficient and must be combined with kinesthetic cues. Figure 1 shows the two different kinds of compliant stimuli used in this work.

Thermal

The resting temperature of the skin (25–36 °C) is usually higher than the ambient temperature of materials in the environment. Thermal cues, which are sensed by peripheral populations of thermoreceptors (warm and cool), may potentially contribute to the haptic identification of different materials by producing different thermal transients as the skin interacts with the material during contact. People can identify materials such as copper, stainless steel, granite, and foam merely by contact, but not equally well. Perceptual confusions among materials are probably the result of similar thermal transients.

Weight

Weight is optimally perceived when the observer actively lifts the object away from a supporting surface, thereby providing information from both cutaneous and kinesthetic mechanoreceptors, as well as from the associated neural correlates of the descending motor commands. Humans use more than object mass to haptically assess weight. For example, a small object feels heavier than a large object of the same mass, a very powerful illusion known as the haptic size–weight illusion.

Opinions vary as to the nature of the processing involved in haptic weight perception. Some researchers have argued that haptic observers use cognitive expectations based on prior experience of the correlation between weight and size (or density) to judge object weight. Others suggest that observers incorporate sensory cues (cutaneous and/or kinesthetic) to object volume into their weight judgments. Still others argue that weight, both veridical and illusory,

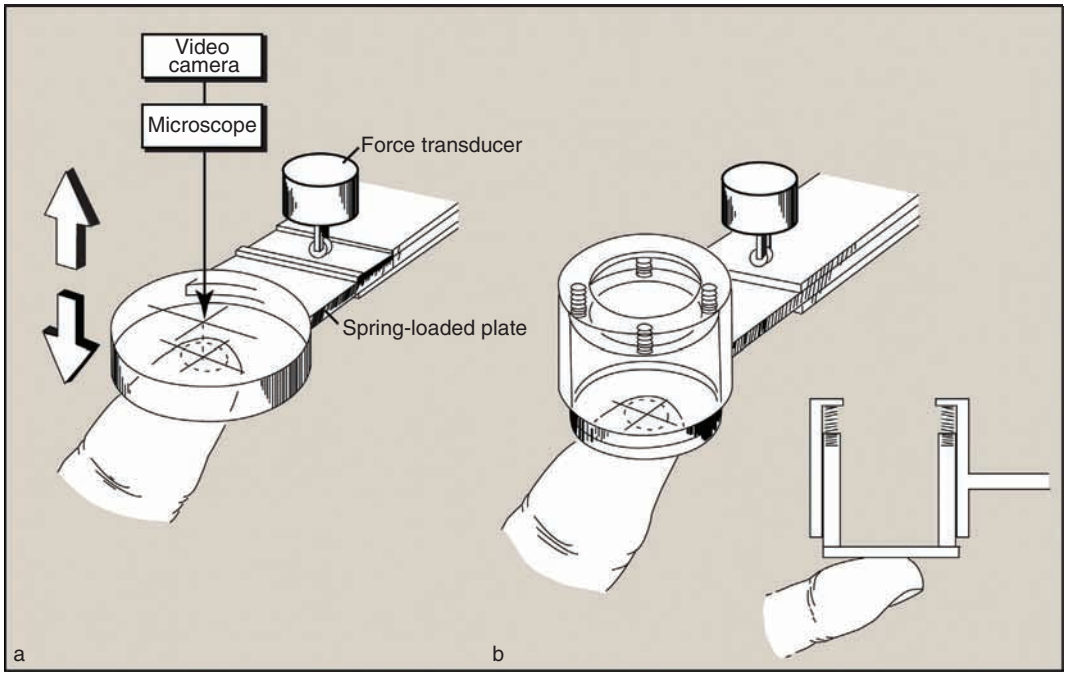

Figure 1 Schematics of apparatus for experiments: (a) with transparent rubber specimens of variable compliance; (b) on deformable objects with planar rigid surfaces. In (a), the specimen is mounted on the spring-loaded plate projecting out of a computer-controlled tactile stimulator. The plate is in contact with a force transducer that is able to measure forces of contact between the specimens and finger pads under active or passive touch conditions. The contact regions are videotaped using a dissection microscope fitted with a video camera. Shown in (b) is a spring cell mounted on the same spring-loaded plate and in longitudinal section. Each specimen consisted of two telescoping hollow cylinders such that the inner cylinder could move freely within the outer cylinder. Four springs glued to the bottom plate of the outer cylinder and connected to the inner cylinder determined the compliance of the specimen. Reproduced from LaMotte MA and Srinivasan RH (1995) Tactual discrimination of softness. *Journal of Neurophysiology* 73(1): 88–101, with permission from American Physiological Society.

is assessed more directly in terms of differences in the magnitude of the moments of inertia, that is, resistance to rotational forces produced as the observer wields an object. This last interpretation constitutes an example of dynamic touch, a haptic subsystem that uses deformations of muscles, tendons, and ligaments to perceptually judge the attributes of objects wielded in the hand (e.g., length, weight, width, shape of the object tip, and the orientation of the hand and object relative to one another).

Curvature, Angle, and Orientation (Tilt)

Humans can haptically detect curvature over a wide range of spatial scales and haptic sensory inputs, according to the size of the explored object. The relative contribution of the cutaneous and kinesthetic signals varies with the spatial scale of the curved surfaces.

Humans are capable of haptically differentiating 2-D angular differences of 5.2° between pairs of relatively large angles with an accuracy level of 75%, with performance depending on both cutaneous and kinesthetic (shoulder) inputs.

Finally, haptic perception of orientation has been measured by requiring that a test bar be positioned parallel to a reference bar presented together within a planar (horizontal) work space. The perception of orientation within this peripersonal space is systematically biased by several factors, particularly the location of the bars within the plane (e.g., distance from the midline of the body).

Shape of Two- and Three-Dimensional Forms Larger than a Fingertip

Spatial patterns that fit within the fingertip are sensed on the basis of cutaneous cues alone, and are thus not discussed further here. Humans haptically process relatively ineffectively 2-D and 3-D shapes of larger objects that are made of a single material. At this time, little is known about the underlying neural codes.

Purposive manual exploration critically influences how objects and their properties are haptically processed and cognitively represented. More specifically, the spatially restricted inputs derived from contacting objects with multiple fingers impose constraints on spatial integration. Furthermore, the sequential nature of manual exploration renders it slow to perform, creating notable demands on memory and

temporal integration. Humans initially process an object haptically more by its local features than by its global shape, particularly when exploration time is limited.

Size

To date, most research has focused on the haptic perception of object length or width, as opposed to its area or volume. Haptic perception of the length of rods statically held at their ends between the two index fingers varies in a 1:1 fashion with physical length. Unlike most other sensory judgments, then, there is excellent correspondence between the percept and the physical property. However, in some cases the haptic perception of size may be systematically distorted. For example, people are haptically susceptible to a variant of the horizontal–vertical illusion, in which a vertical line is perceived to be longer than the same line oriented horizontally, in both T- and L-shaped configurations of this illusion. The horizontal–vertical illusion is believed to consist of two separate components resulting from bisection and radial–tangential effects. The bisection effect consists of the bisecting line being overestimated relative to the bisected line. However, the radial–tangential effect is an anisotropy of haptic space in which radial movements toward and away from the observer's body are overestimated relative to those that are tangential to the body.

Haptic Space Perception by Sighted and Blind Observers

How spatial locations are coded haptically and which errors are induced by the coding system adopted are issues that have been extensively investigated. The focus here is on space within human reach, that is, manipulatory space. Do people represent the location of their hands in space by the kinesthetic posture needed to place them there or by some extrinsic spatial framework? A posture-matching strategy is adopted in tasks wherein the observers' hands are passively moved to a location to which they must reach again. If this strategy is prevented, for example, by having the observer reach the second time with a pointer, performance by posture matching becomes relatively error-prone and variable.

A frame of reference may also be defined extrinsically with respect to the body when space is haptically explored. For example, observers who attempt to match the tabletop position to which one hand is moved by placing the other hand in the corresponding location underneath the table (Figure 2) appear to make spatial errors consistent with using an extrinsic frame of reference in which the haptic egocenter (origin) is located at the shoulder of the arm used to mark the position on the tabletop. The position of this haptic egocenter is not fixed; for example, it reverses when the roles of the hands are switched.

Still other research has shown that spatial coding can also occur using an extrinsic spatial frame of reference outside the body. For example, consider when an observer's left or right finger is passively guided to a specific location. The observer must then actively reach to that location with the right hand. Systematic biases (constant errors) shift the endpoints outward laterally along the frontal plane and compress them sagittally and vertically. More significantly, variable errors are essentially independent of whether the observer uses the same hand or switches hands between initial exposure and test. The postures for the responding (right) arm depend little on the arm that originally locates the target. Therefore, target position after passive kinesthetic exposure can be represented in an amodal frame of reference that is not referred to reaching postures; rather, it is defined by extrinsic space and is shared by vision and kinesthesis. This representation is not always observed when relatively few target locations and responses are used.

It has been further proposed that multiple spatial representations are formed after haptic exposure in a temporal cascade. These begin with kinesthetic reference frames and proceed to extrinsic frames that might be egocentric or environmental.

Once spatial locations are encoded by means of reaching, they can be updated when people move in the absence of vision. Studies investigating nonvisual updating in manipulatory space have found that sighted and blind observers can update their representation of touched locations after moving their finger to a new reference location or after walking to a new location. However, updating is superior when observers can actually touch the target at the updated location rather than just point to it, suggesting that a specifically haptic code is updated more successfully during movement than is an abstract representation of location.

Haptic Recognition of Objects

Humans are remarkably effective at recognizing common objects using only their hands. When asked to identify them, participants attain close to 100% accuracy, most often within only 2–3 s.

Role of Hand Movements

Active manual exploration may play a critical role in the normal development of somatosensory perception in mammals. Nonetheless, as long as there is

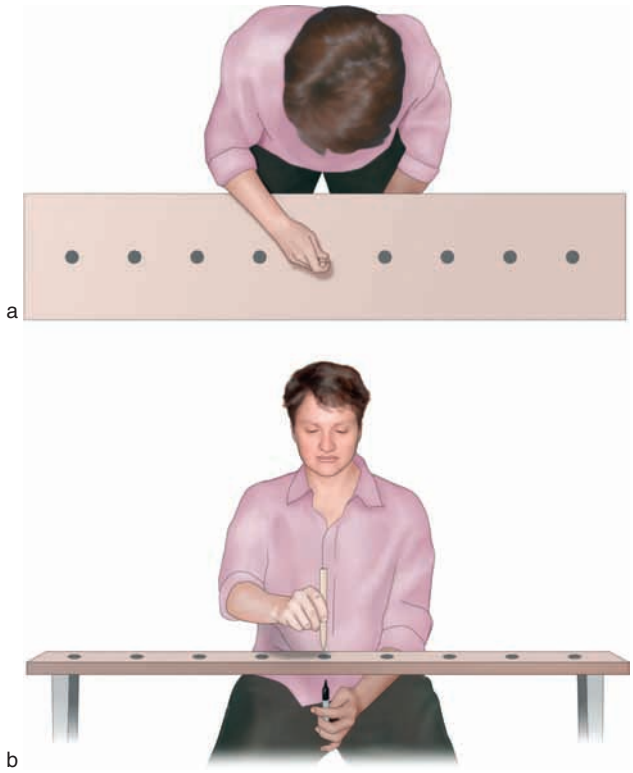

Figure 2 Locating the haptic egocenter: (a) top view; (b) front view. One hand places a stylus on the target on the upper table surface, and the other hand attempts to match up underneath the table in the corresponding location. Reproduced from Wolfe JW, Levi D, Kluender K, Bartoshuk L, Herz R, Klatzky RL, and Lederman S (2005) *Sensation and Perception*. Sunderland, MA: Sinauer, with permission from Sinauer Associates, Inc.

relativemotion between the surface and the skin, for fingertip-size stimuli such as Braille characters and textures, it is not critical whether the hand moves over a stationary surface or the surface moves across the stationary hand.

Manual exploration contains the key to understanding why humans are so skilled at recognizing common objects haptically. People manually explore such multi-attribute objects and surfaces in highly systematic ways by executing a number of stereotypical hand-movement patterns know as exploratory procedures. Each procedure is voluntarily selected in association with one or more object properties to which the person is selectively attending.

The exploratory procedures that have most frequently been the focus of scientific investigation are shown in Figure 3 in association with the property (ies) with which they have been most closely linked. Lateral motion involves a back-and-forth shearing motion across a surface and provides the most precise information about the textural properties of a surface. Pressure involves the application of force normal to a surface or a torque about some axis of the object and provides the most precise information about properties relating to compliance. Static contact, as its name implies, involves making simple static contact with a surface and is used to extract optimal information about thermal qualities. Unsupported holding involves lifting the object away from a supporting surface, often jiggling it, to learn best about its weight. Enclosure involves molding the fingers to the object's contours

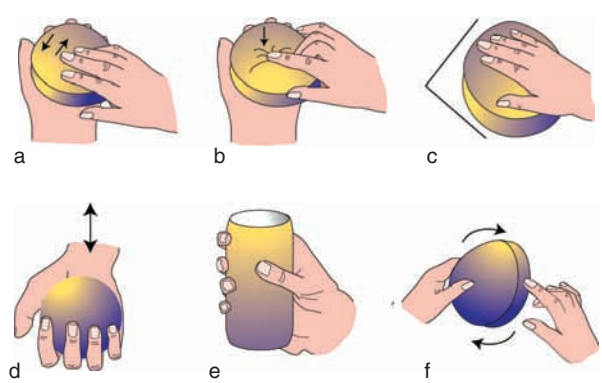

Figure 3 Six haptic exploratory procedures and the object property(ies) with which each is most closely associated: (a) lateral motion, texture; (b) pressure, hardness; (c) static contact, temperature; (d) unsupported holding, weight; (e) enclosure, global shape and volume; (f) contour following, global shape and exact shape. Reproduced from Lederman SJ and Klatzky RL (1987) Hand movements: A window into haptic object recognition. *Cognitive Psychology* 19(3): 342–368, with permission from Elsevier Ltd.

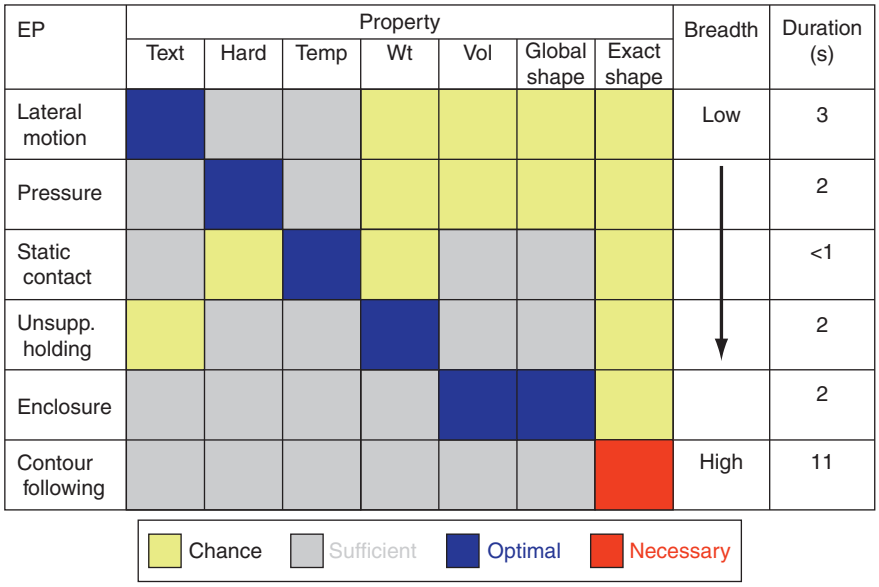

Figure 4 Costs and benefits as influenced by three relative performance characteristics of haptic exploratory procedures: precision of information (chance, sufficient, optimal, or necessary), breadth of sufficiency or generality, and average execution duration. EP, exploratory procedures; Temp, temperature; Unsupp., unsupported; Vol, volume; Wt, weight. Reproduced from Lederman SJ and Klatzky RL (1987) Hand movements: A window into haptic object recognition. *Cognitive Psychology* 19(3): 342–368, with permission from Elsevier Ltd.

and provides the most precise information about an object's size and general shape. Finally, contour following involves tracing around an object's edges or contoured surfaces and provides the most precise spatial details about an object.

These six exploratory procedures have been further compared in terms of four different performance characteristics: precision of information, breadth of object property information, average duration, and co-executability. A summary of the first three relative performance characteristics is presented in Figure 4.

Relative precision of information The empty cells indicate that the designated exploratory procedure (row) does not provide any information about the target property (column). Such is the case when unsupported holding is used to provide information about texture. The gray cells indicate that the designated exploratory procedure provides only coarse information about the target property, such as when pressure is used to extract texture information. The blue cells indicate that the designated exploratory procedure provides the most precise information about the target property, as is the case when lateral motion is used to extract texture. Finally, the red cell indicates that the designated exploratory procedure not only provides the most precise information about the target property (exact shape) but that this information is necessary for the haptic perception of fine shape details.

Breadth of property information The exploratory procedures also vary in terms of the number of properties about which information is made available, minimally at a coarse level. Thus, lateral motion and pressure, the least general hand-movement patterns, provide only three types of property information that are all related to an object's material. In contrast, enclosure and contour following each offers information about both material and geometric object properties.

Duration Most exploratory procedures are relatively quickly performed, as is the case for those that provide relatively precise information about material. In contrast, although relatively quickly executed, enclosure provides only coarse information about an object. Contour following takes notably longer to perform than any other exploratory procedure.

Co-executability A fourth relative performance characteristic relates to the extent to which exploratory procedures can be co-executed. To the extent that this is possible, the observer gains information that each exploratory procedure is not only optimal (or necessary), but also sufficient, at providing.

Implications for Haptic Object Classification

The relative performance characteristics of exploratory procedures just discussed have both costs and benefits for the haptic perception of the physical

properties of the concrete world and for object recognition and identification.

First, material properties of objects tend to be more salient (i.e., more active) in our conscious experience than others. Consider what happens when observers are biased by instructions to focus on either haptic or visual information when sorting multi-attribute objects into several piles. The most similar objects are to be placed within the same pile. When objects are constructed so that material (e.g., texture and compliance) and geometric (e.g., shape and size) dimensions vary in perceptually equivalent (i.e., discriminable) ways, the importance of material properties for people's judgments of object similarity increases when they are given haptic-bias instructions compared to when they are given vision-bias instructions (Figure 5). The reverse occurs for the importance of geometric properties when people are given vision-bias (as opposed to haptic-bias) instructions. This result may be explained by the relative efficacy with which people perform material, as opposed to geometric, exploratory procedures.

Second, when vision is available during object recognition, humans tend to use their sense of touch primarily when the most diagnostic object property relates to material, as opposed to geometric, properties (e.g., sandpaper vs. spoon, respectively). As shown in Figure 6, when people are instructed to make difficult discriminations between pairs of common objects whose most diagnostic property is based on material (e.g., texture and compliance) or geometry (e.g., shape complexity and size), they choose to also haptically explore the objects for which material differences are most diagnostic. This is much less the case for highly discriminable pairs, presumably because information is semantically available.

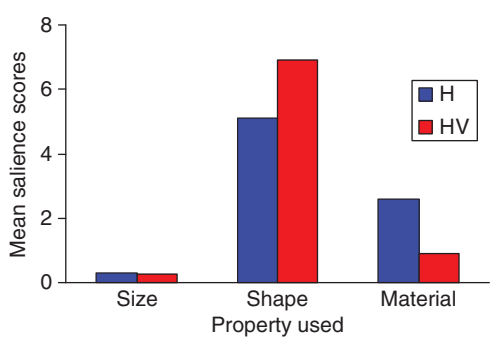

Figure 5 Relative cognitive salience of geometric (shape and size) and material object properties. Subjects were instructed to sort multi-attribute objects into piles based on cues available from haptic (H) or haptic + visual (HV) exploration. The salience score (out of 8) were calculated on the basis of subjects' sorting judgments Based on Lederman S, Summers C, and Klatzky RL (1996) Cognitive salience of haptic object properties: Role of modality-encoding bias. *Perception* 25(8): 983–998.

Third, when people haptically recognize an object, manual exploration proceeds in a stereotyped way that is determined by the relative breadth of property information afforded by exploratory procedures and by their associated durations. Thus, humans begin their manual exploration by grasping (enclosure) and lifting (unsupported holding) the object. Both movement patterns provide information about many object properties and are relatively quickly executed. This highly standard routine is followed, albeit less frequently, by executing the exploratory procedure that provides optimal information, that is, that offers the most precise information about the property that is highly diagnostic of the object.

Fourth, humans may capitalize on the fact that multiple properties often predict an object's identity. For example, a pencil is long and narrow; in addition, it frequently has a section at one end consisting of an interior metal band adjacent to a rubber end. With the haptic system, the perceptual gain resulting from property redundancy(ies) is potentially constrained by manual exploration. The addition of a second predictive property does not reduce the time needed to categorize objects if the two exploratory procedures that are optimal for those properties cannot be co-executed, either because of biomechanical constraints or because the second property is available at a region that cannot be explored simultaneously with the first.

Consequences of Constraining Manual Exploration in Space and/or Time

Restricting manual exploration helps to reveal how object properties are haptically encoded. More specifically, limiting spatial access to objects highlights the significance of both shape and material information

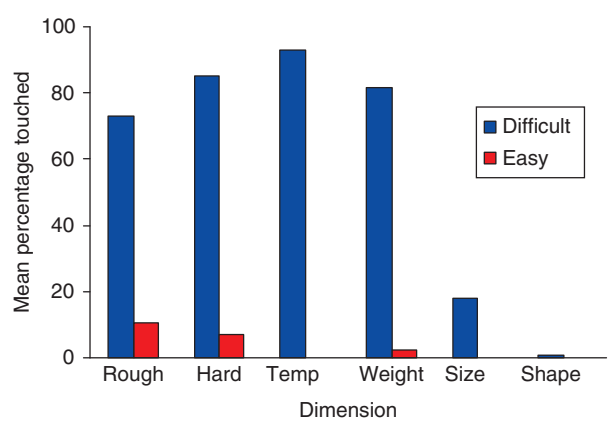

Figure 6 Frequency of touching objects during haptic discrimination of various object properties when vision is available. Temp, temperature. Reproduced from Klatzky RL, Lederman SJ, and Matula D (1993) Haptic exploration in the presence of vision. *Journal of Experimental Psychology: Human Perception and Performance* 19(4): 726–743, with permission from APA.

for object identification. Limiting temporal access speaks, rather, to the sequence in which object properties emerge.

Spatial constraints Small-scale, cutaneous spatial constraints may be imposed by covering the fingertip in a rigid sheath. Without the spatially distributed force array that is normally available, observers can access only the summed forces stimulating receptors in the skin, muscles, tendons, and joints. Under such circumstances, they retain their ability to sense vibration and remain sensitive to coarse pressure differences; however, they are unable to discriminate fine spatial patterns within the scale of the fingertip.

It is also possible to constrain larger-scale spatial variation in objects that are typically explored using contour following and/or enclosure exploratory procedures. Consider, for example, what happens when observers are required to identify a set of common objects selected because they are all rigid, fairly smooth, and fixed to a supporting surface. A thick glove is worn to reduce the cutaneous information from the spatially distributed deformation patterns and vibrations normally created during contact. Under such circumstances, kinesthetically derived shape information becomes critical for object identification. Performance may be compared across gloved haptic exploration conditions that vary in the extent to which the kinesthetic inputs from muscles, tendons, and joints are spatially constrained. For example, the kinesthetic cues may remain intact. Alternatively, kinesthetic information may be reduced by wearing finger splints to prevent any flexion at the joints or by restricting the number of fingers used during exploration. By comparing performance across conditions, we can assess the extent to which reduced kinesthesis impairs object recognition when cutaneous sensing is minimized. Regardless of the method used to reduce kinesthetic information, performance declines, and it worsens as the number of exploration constraints is increased. When a rigid interface (i.e., sheath) between hand and object or a rigid, pencil-shaped probe is used, performance declines still further.

Temporal constraints The costs and benefits of exploratory procedures imply that features become available at different points in time, depending on the duration of haptic exploration and the processes (e.g., intensive and spatial) used to encode the explored features. Haptic access to information pertaining to material, to the presence of edges, and to geometric features has been shown to differ in terms of the relative availability of information about these feature subclasses over time. More specifically,

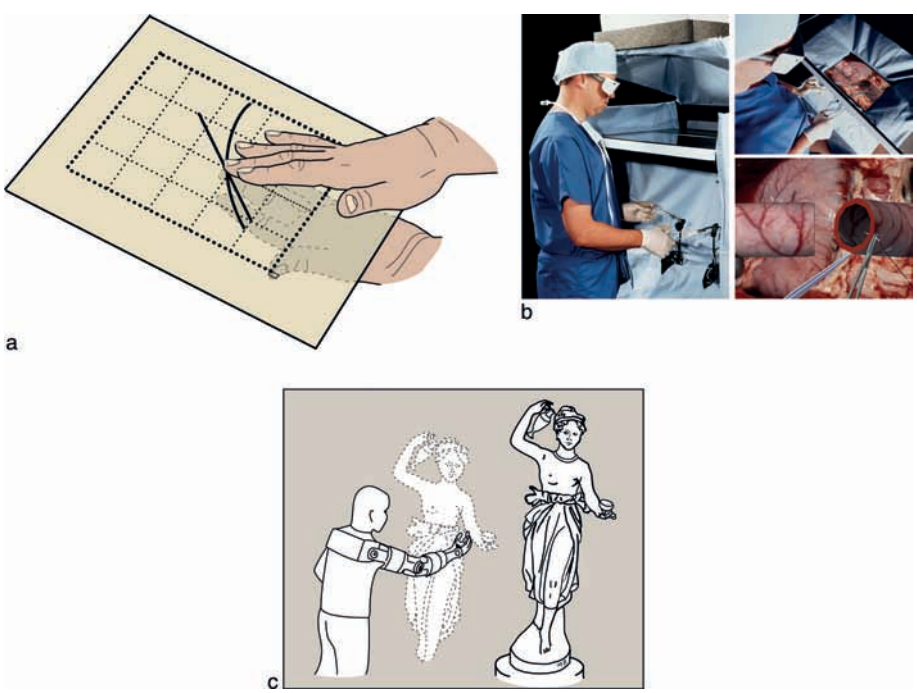

Figure 7 Examples of haptic application domains: (a) designing tangible graphics displays for the blind; (b) virtual surgical trainer; (c) virtual haptic sculptures. (a) Reproduced from Jones LA and Lederman SJ (2006) *Human Hand Functon*. New York: Oxford University Press, with permission of Oxford University Press. (b) Reproduced by permission of Boston Dynamics. (c) From Copyright 2003 from *Visual Impairment Research* by Jansson G, Bergamasco M, and Frisoli A. Reproduced by permission of Taylor & Francis Group, LLC., http://www.taylorandfrancis.com.

material attributes and edges, which are processed intensively, emerge earlier than do the spatial arrangement of an object's geometric features, even when those features lie within the span of a fingertip.

Applications

The results of studies based on the scientific investigation of human haptic processing of surfaces, objects, and their properties have been applied to a number of application domains. Examples include the design of tangible graphics displays (maps, drawings, graphs, etc.) for the blind that are apprehended by hand as opposed to eye, the design of haptic and multisensory interfaces for exploring and acting on remote real and virtual environments, and haptic art for the blind and sighted. Several examples are presented in Figure 7.

See also: Cross-Modal Interactions Between Vision and Touch; Machine Haptics; Mechanoreceptors; Perception of Surfaces and Forms; Proprioception; Reaching and Grasping; Sensorimotor Control of Manipulation; Shape Representation in Inferotemporal Cortex; Somatosensory Perception; Somatosensory Cortex: Functional Architecture; Somatosensory Pathways (Ascending): Functional Architecture; Spatial Transformations for Eye–Hand Coordination; Tactile Coding in Peripheral Neural Populations; Tactile Texture; Temperature Sensation.

Further Reading

Bensmaïa S, Hollins M, and Yau J (2005) Vibrotactile intensity and frequency information in the Pacinian system: A psychophysical model. *Perception & Psychophysics* 67(5): 828–841.

Burdea G and Coiffet P (2003) *Virtual Reality Technology*, 2nd edn. Hoboken, NJ: Wiley-Interscience.

Connolly KJ (ed.) (1998) *The Psychobiology of the Hand*. London: MacKeith Press.

Gibson JJ (1966) *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Hatwell Y, Streri A, and Gentaz E (2003) *Touching for Knowing: Cognitive Psychology of Haptic Manual Exploration*. Amsterdam: John Benjamins.

Johnson KO (2001) The roles and functions of cutaneous mechanoreceptors. *Current Opinions in Neurobiology* 11: 455–461.

Jones LA and Lederman SJ (2006) *Human Hand Function*. New York: Oxford University Press.

Katz D (1989) *The World of Touch*, Krueger L (trans.) Mahwah, NJ: Lawrence Erlbaum (Original work, *Der Aufbau der Tastwelt*, published 1925).

Klatzky RL and Lederman SJ (2003) Touch. In: Healy AF and Proctor RW (eds.) *Handbook of Psychology, Vol. 4: Experimental Psychology*, pp. 147–176. New York: Wiley.

Klatzky RL, Lederman SJ, and Matula D (1993) Haptic exploration in the presence of vision. *Journal of Experimental Psychology: Human Perception and Performance* 19(4): 726–743.

La Motte MA and Srinivasan RH (1995) Tactual discrimination of softness. *Journal of Neurophysiology* 73(1): 88–101, with permission from American Physiology Society.

Lederman SJ and Klatzky RL (1987) Hand movements: A window into haptic object recognition. *Cognitive Psychology* 19: 342–368.

Lederman S, Summers C, and Klatzky RL (1996) Cognitive salience of haptic object properties: Role of modality-encoding bias. *Perception* 25(8): 983–998.

Rieser D, Ashmead F, Ebner, and Corn A (eds.) (2007) *Blindness and Brain Plasticity* in Navigation and Object Perception. Mahwah, NJ: Lawrence Erlbaum.

Turvey MT (1996) Dynamic touch. *American Psychologist* 51: 1134–1152.

Wolfe JW, Levi D, Kluender K, Bartoshuk L, Herz R, Klatzky RL, and Lederman S (2005) *Sensation and Perception*. Sunderland, MA: Sinauer.

Human Methods: Psychophysics

B A Schneider, University of Toronto at Mississauga, Mississauga, ON, Canada
S Parker, The American University, Washington, DC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Psychophysical methods have been developed by researchers in an attempt to determine how the mind encodes and represents environmental events – that is, to determine the relationship between the mind ('psycho') and the world ('physics'). To investigate such relationships, methods were developed in which observers were required to make judgments concerning the presence or absence of stimuli, or the sensory relationship of one stimulus to another (e.g., whether stimulus one is brighter than stimulus two). These judgments were then used to determine how the stimulus is encoded and represented in the minds of the observers. Note that the nature of this representation is inferred from behavioral judgments. Hence psychophysical methodology is a way of indirectly investigating how information about the physical world is organized and represented in the nervous system.

Although these methods were initially used to investigate the relationship between a physical aspect of the stimulus and how it is experienced by the observer (e.g., how the loudness of a sound is related to its physical intensity), they are now used to investigate how information in general is encoded and represented. Hence, psychophysical methodology can be, and is, used not only to investigate how a psychological attribute varies as a function of some physical variable, but also to determine, among other things, the psychological utility of objects, the prestige of various occupations, or the discriminability between classes of words previously presented to observers and semantically related words that were not previously presented. Note that in these latter three examples, the psychological property under investigation is not directly related to any specific physical aspect of the stimulus.

In approaching psychophysical methodology we begin here by describing the theory upon which many of these methods are based. We then describe the range of psychophysical methods that have been developed to assess how information is encoded and represented, along with examples of their application.

Theoretical Concepts

Representation of Information

Psychophysical methodologies assume that the activity a stimulus evokes in the nervous system can be represented as a point in an n-dimensional space. Consider, for example, a pure tone, which has four physical dimensions, intensity, frequency, duration, and phase. If each of these dimensions had a psychological counterpart, we would need four dimensions to represent the manner in which an observer experiences that tone. However, because observers are not sensitive to monaural phase, it is likely that there are only three dimensions to the psychological experience: loudness, pitch, and perceived duration. In Figure 1, the loudness, pitch, and perceived duration evoked by single presentations of each of three tones are indicated as points in a three-dimensional (3-D) space. Because a stimulus will not give rise to exactly the same experience each time it is presented (because of the presence of external noise or intrinsic variability in the nervous system), repeated presentations of a stimulus are depicted as giving rise to a cloud of points clustered around a mean position in this 3-D space.

To better illustrate the degree of variability in the evoked experience, Figure 2 plots the probability density functions that result when three pure tones having different intensities and frequencies, but identical durations, are presented repeatedly to an observer. (Because the mean perceived duration is the same for all three tones, variation along the perceived duration dimension is not illustrated.) As in Figure 1, variations in loudness and pitch are assumed to be independently and normally distributed with standard deviations of 1.0. In this example, because the tones are closely spaced together, the distributions exhibit some degree of overlap.

It is important to note that, theoretically, we could assume any two-dimensional probability function for the stimuli in this space, and we need not assume that all probability density functions are equally variable. However, the default assumption is usually that the distributions are normal or Gaussian in shape, and that the variation of stimuli along any one dimension is constant. (In Figure 2 the variances are shown as equal across the two dimensions.) In addition, in the vast majority of cases in which psychophysical methods are employed, it is assumed or shown that the stimuli being judged can be represented along a single

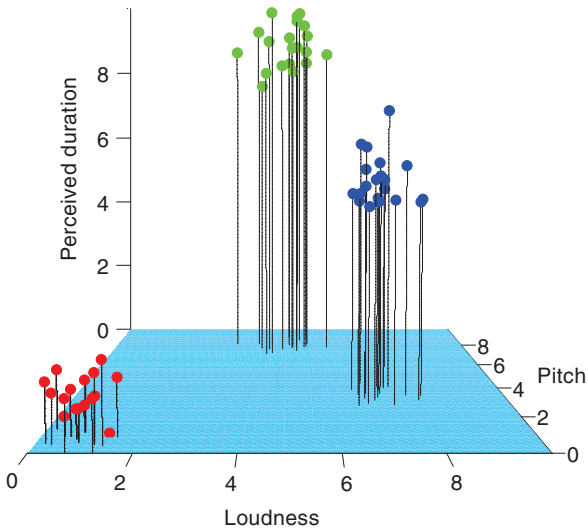

Figure 1 Each tone is presumed to evoke a sensation of loudness, pitch, and perceived duration. The subjective experience corresponding to a tone presentation is represented as a point in a 3-D perceptual space with axes of loudness, pitch, and perceived duration. Because of variability in the stimulus, and the presence of internal noise, the location of a point varies from trial to trial, even when the stimulus remains the same. In generating a cluster of points for a tone, it was assumed that the loudness, pitch, and perceived duration values were independently and normally distributed with a standard deviation equal to 1.0. Shown here are hypothetical clouds of points resulting from 20 repetitions of three tones. The mean locations of the three tones, for loudness (L), pitch (P), and perceived duration (PD), are depicted as follows: red ($L=1$, $P=1$, PD$=1$), green ($L=5$, $P=8$, PD$=8.5$), and purple ($L=7$, $P=4$, PD$=4.5$).

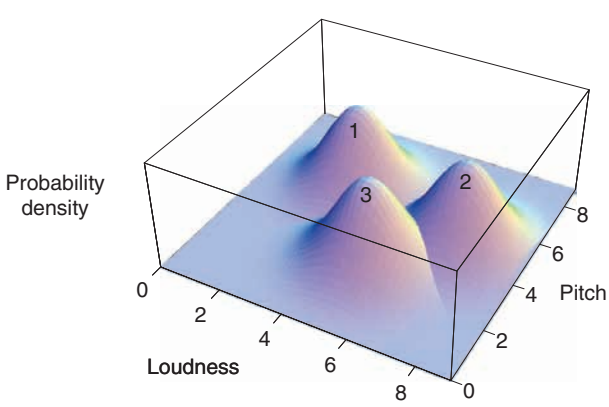

Figure 2 Probability density functions for three tones having identical perceived durations but different loudness and pitch values. Loudness (L) and pitch (P) values are as follows: tone 1 ($L=3$, $P=6$), tone 2 ($L=7$, $P=5$); and tone 3 ($L=5$, $P=3$). The loudness and pitch values for each of the tones are assumed to be independently and normally distributed with standard deviations equal to 1.0.

psychological dimension. For example, were we to restrict our set of pure tones to have the same mean pitch and perceived duration, so that they differed reliably only in loudness, we would choose to represent

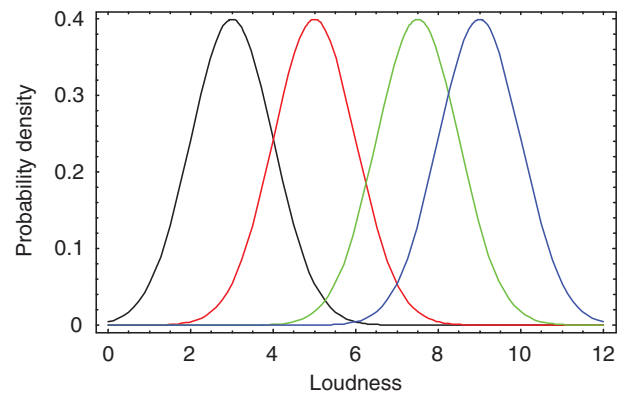

Figure 3 Hypothetical distributions of loudness values for four pure tones in the equal-variance Gaussian model.

their density functions along a single psychological dimension, such as the one shown in Figure 3. Here, the Gaussian probability density functions for all tones have the same variance, and differ only in their mean values. We refer to this situation as the unidimensional, normally distributed, equal-variance case. In describing the various psychometric methods, we restrict our discussion herein to this case.

Relationship to the Goal of Psychophysics

As mentioned in the Introduction, psychophysical methods were developed to explore how sensory information is structured in the mind, and to relate this structure to the physical world. For instance, if Figure 3 truly represented how the loudness of four different tones was experienced, then one could relate the mean loudness values of these stimuli to their physical intensities, thereby establishing a psychophysical relationship. Psychophysical methods, based on observers' judgments, were developed to establish how the physical properties of stimuli were experienced, so that psychophysical relationships, such as the one between loudness and sound intensity, could be established. However, these methods can be used to determine how other psychological phenomena are represented, such as the utilities of a collection of objects, even though it may be difficult to determine exactly how the utility of an object is related to its physical properties. In all cases, these psychophysical methods are based on judgments about relationships among stimuli. Here we consider different types of judgments and what can be learned from them.

Psychophysical Methodologies

Detection Experiments

In many instances, the experimenter wants to know the minimum value of a stimulus that can be detected. For example, a vision scientist might want to know

the smallest degree of contrast that can be detected in a sinusoidal grating (see Figure 4, top row), and the simplest way of determining this is to ask the observer to report the presence or absence of contrast on a trial. There are two classes of psychophysical methods that have been developed for the detection experiment, the Yes/No method and the *m*-interval forced-choice method.

The Yes/No method In the simplest version of this method there are *n* trials and a single stimulus. The occurrence of a trial is signaled in some fashion, and there are two types of trials: those in which a stimulus appears and those in which no stimulus is presented. For example, consider an experiment in which the stimulus is a low-contrast sinusoidal grating (Figure 4, top row), and the observer's task is to detect the presence of this stimulus when it is added to a field of Gaussian noise (Figure 4, middle row). The occurrence of a trial is signaled by the presentation of a fixation cross in the middle of the screen, and a trial of fixed duration begins either shortly after the fixation cross occurs, or shortly following a button press by the observer. The trial consists of a presentation of the noise field alone (N trial; Figure 4, middle row), or the sum of the sinusoidal grating (S) and the noise field (signal + noise, or SN, trial; Figure 4, bottom row). The probability of an SN trial, $p(\mathrm{SN})$, is fixed throughout the session, and the observer's task to indicate whether there is a signal present (a Yes response) or not (a No response). Correct detection of a signal is typically referred to as a Hit (H); responding Yes when there is no signal is referred to as a False Alarm (FA). If there are $n(\mathrm{N})$ noise trials and $n(\mathrm{SN})$ signal trials, the critical data in this experiment are the proportion of Hits, $n(\mathrm{H})/n(\mathrm{SN})$, versus the proportion of False Alarms, $n(\mathrm{FA})/n(\mathrm{N})$. These two proportions are used to estimate the detectability of the sinusoidal grating, assuming that the N and SN stimuli give rise to two different distributions of events along a unidimensional decision axis (e.g., perceived contrast). As mentioned previously, it is usually assumed that these two distributions are Gaussian and have equivalent standard deviations (see Figure 5(a)).

Interpretation of the results of this experiment, however, depends on whether or not one assumes that the threshold for detecting a grating is fixed or variable. Figure 5 illustrates what we would expect if the threshold was fixed. Here, the observer would report the presence of a grating whenever the occurrence of a trial produced an event along the decision

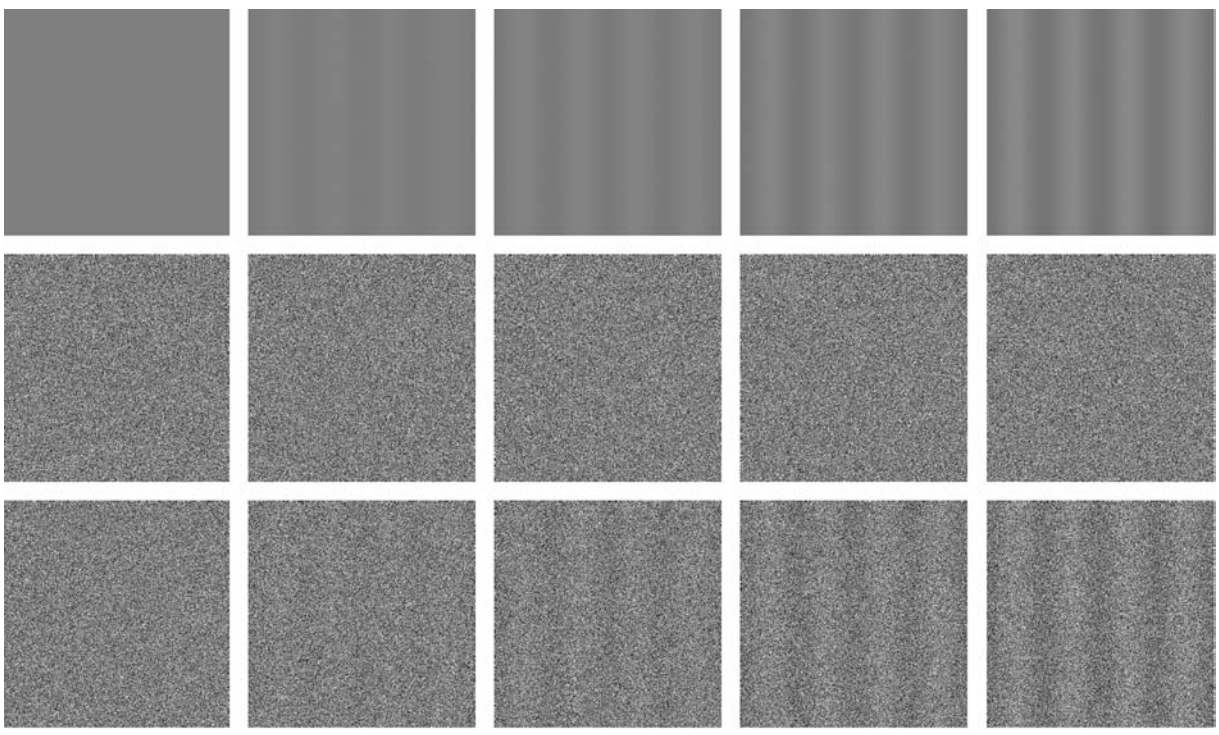

Figure 4 Images of signals and noises. (Top row): Signals; vertically oriented sine wave gratings (four cycles/image). The leftmost image has a contrast of 0 (no grating, i.e., uniform gray); the contrasts of the other images in the row increase from left to right. (Middle row): Noises; five independent samples of Gaussian noise. (Bottom row): Signal + noise; here, the gratings in the top row have been added to independent samples of noise. Note that there is no sinusoidal grating in the leftmost square of this row. Hence, this is technically a noise trial. The other four squares constitute signal + noise events.

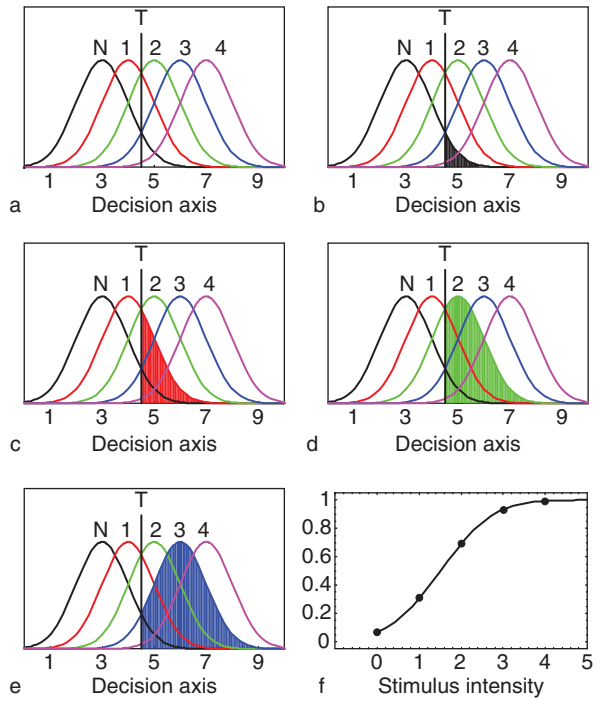

Figure 5 (a) Noise (N) and four signal + noise distributions (1, 2, 3, 4) along a unidimensional decision axis; T represents the point on the decision axis corresponding to a fixed threshold. (b) The portion of the N distribution that exceeds threshold. (c–e) The corresponding portions of the first three signal + noise distributions (1–3) that exceed threshold. (f) Probability of a Yes response as a function of stimulus intensity. This function assumes that the mean location of a stimulus event along the decision axis is linearly related to stimulus intensity. If the relation mapping stimulus intensity onto the decision axis is nonlinear, the shape of this function will change accordingly.

axis (the neural representation of contrast) to the right of the threshold, T. The area under an SN distribution to the right of the threshold would correspond to the probability of a Hit in an SN trial. Increasing the degree of contrast would shift the mean of the SN distribution to right, thereby increasing the proportion of the distribution to the right of T (compare areas to the right of T for Figures 5(b)–5(e)). Hence, by presenting different stimulus contrasts over successive experimental sessions one could map out a psychometric function relating the proportion of Yes responses to the degree of visual contrast as indicated in Figure 5(f). Note that as visual contrast approaches 0, p(H) approaches p(FA). The reason for this is that the N distribution extends to the right of T, so occasional events occur along the decision axis that exceed this fixed threshold even in the absence of a stimulus. Hence, p(FA) provides a floor for p(H), as stimulus contrast becomes smaller and smaller.

Note that if the threshold is fixed, the probability of a False Alarm will be independent of the stimulus contrast in the signal. Therefore, rather than varying stimulus contrast across sessions (where each session consists of N and SN trials), it would be more efficient to simply vary stimulus contrast within a session, with one of the several contrast values being zero (an N trial). This method is called the 'method of constant stimuli,' and the asymptotic value of the psychometric function obtained when stimulus magnitude approaches zero is sometimes referred to as the 'probability of guessing.' Under fixed-threshold theory, it would be interpreted as a measure of the location of the threshold relative to the null event (no stimulus presentation).

Assuming a fixed threshold and a unidimensional decision axis, one can measure the location of each of the stimuli along the decision axis with reference to the null event, measured in standard deviation units. Assume for the moment that we use the method of constant stimuli (in which one of the stimuli was the null stimulus) to map out a psychometric function relating p(Yes) to stimulus contrast. To do this we first rescale the x axis so that all distributions have a standard deviation of 1.0. Second, we determine the z-score corresponding to the location of the threshold T with respect to the mean, μ_N, of the N distribution ($z_{N,T} = \mu_N - T$). Third, we determine the z-score for T relative to the mean for an SN stimulus (e.g., $z_{SN1,T} = \mu_{SN1} - T$). Note that p(FA) (see Figure 5(b)) is the probability that an event elicited on an N trial, $z_{N,}$, exceeds $z_{N,T}$ – that is, $p(FA) = p(z_N > z_{N,T})$ – and that p(H) on an SN_1 trial (see Figure 5(c)) is the probability that an event elicited on an SN_1 trial, z_{SN1}, exceeds $z_{SN1,T}$ – that is, $p(H) = p(z_{SN1} > z_{SN1,T})$. Hence, p(FA) and p(H) can be used to find $z_{N,T}$ and $z_{SN1,T}$, respectively. Now if we subtract $z_{N,T}$ from $z_{SN1,T}$ we obtain $(\mu_{SN1} - T) - (\mu_N - T) = \mu_{SN1} - \mu_N$. Hence, the difference between the z-transformation of p(H) when SN_1 is presented, and the z-transformation of p(FA), is an estimate of the distance between the means of the two distributions when the abscissa has been rescaled so that all distributions have a standard deviation of 1.0. This difference between the two means is called d' (d prime). Therefore, we can determine the separation between the means of the signal distributions with respect to the mean of the N distribution in units of d'. Note that to do this, we need to assume that the decision axis is unidimensional and that the distributions corresponding to different stimuli are equal-variance normal distributions. Increasing the separation between the distributions in Figures 5(a)–5(e) will increase d' and the rapidity of the growth of the function in Figure 5(f). An individual who has a large value of d' has good resolving power – a good ability to discriminate between presentations of N and of SN.

Note that up to this point we have assumed a fixed threshold. A number of experiments have suggested that the location of the decision point along the decision axis is variable and can be set by the observer. For example, in a simple Yes/No experiment, we can vary the probability of an SN trial across sessions and/or we could create variable rewards for Hits and penalties for False Alarms. If the threshold is truly fixed, then p(FA) and p(H) should be independent of both p(SN) and the payoff structure of the experiment. However, in general they are not, and there is every indication that observers can adjust location of the decision point (the 'criterion') to improve outcomes in the experiment.

Consider the case in which the probability of an SN trial is 0.5, and the observer is penalized \$2.00 for every False Alarm, and rewarded with \$1.00 for every Hit. Assume that the N and SN distributions are separated by 1.5 standard deviation units ($d' = 1.5$). Figure 6 shows five different possible locations (panels a–e) of the criterion along the decision axis. Each of these criterion locations will give rise to different proportions of False Alarms and Hits in an experiment. Assume 1000 (500 N and 500 SN) trials. The amount of money that a person will earn is $\$1.00 \times n(\text{H}) - \$2.00 \times n(\text{FA})$, where n(H) and n(FA) are the numbers of Hits and False Alarms. The amount of money earned for each of the criterion locations is also shown in Figures 6(a)–6(e). Figure 6(f) shows how the expected payoff for this experiment varies as a function of criterion location. Observers in this experiment will maximize their payoffs if they locate their criterion approximately 5/6 of the way between the mean of the N distribution and the mean of the SN distributions. Clearly, if possible, observers should position their criterion to maximize payoff. A similar argument holds when p(SN) is varied.

By convention, the unbiased condition occurs when the observer locates the criterion midway between the two distributions. Under these conditions, p(FA) will equal $1 - p$(H). Changing the location of the criterion away from the unbiased location will change p(FA) and p(H) in a systematic way, but will not change d'. A plot of p(H) versus p(FA) for a constant d' is called a 'receiver operating characteristic' (ROC). Figure 7 shows ROCs for $d' = 0$ (the diagonal line), $d' = 0.5$, $d' = 1$, $d' = 1.5$, and $d' = 2.0$.

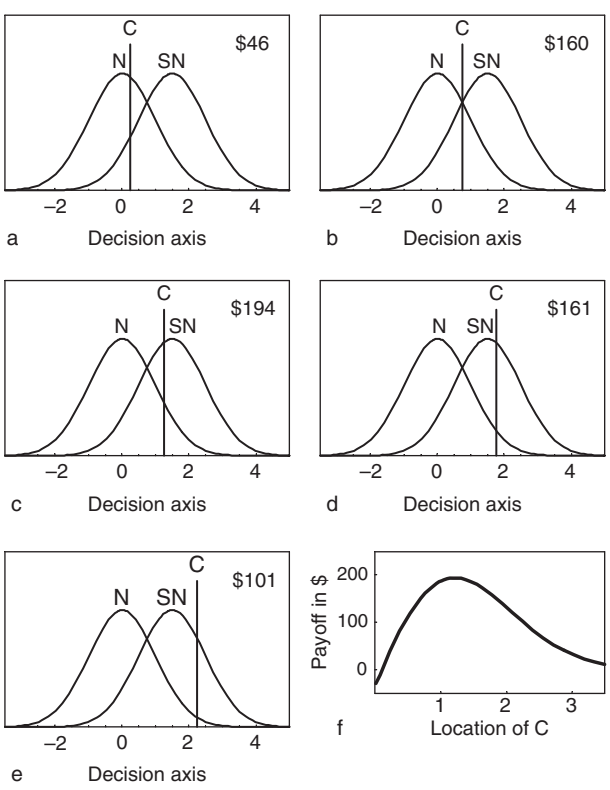

Figure 6 (a–e) Different criterion (C) locations for detecting a signal in noise, where d' between the noise (N) and signal + noise (SN) distributions is 1.5. The dollar amount is the expected payoff if the observer is given \$1 for each hit and penalized \$2 for each false alarm (number of trials = 1000). (f) The expected payoff for this experiment as a function of criterion location.

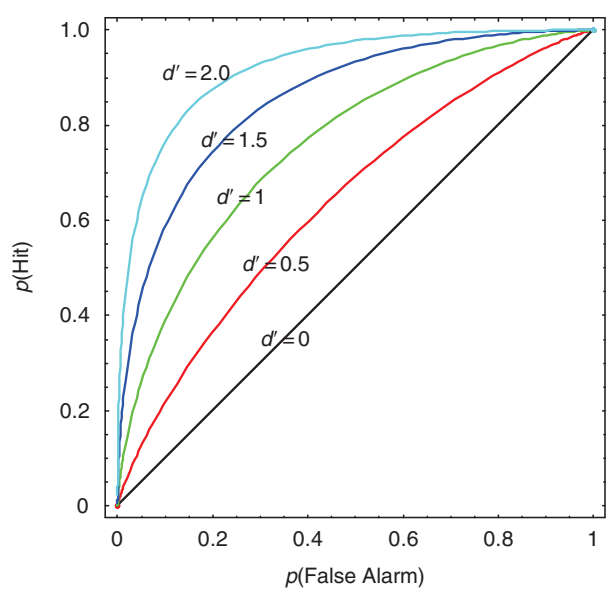

Figure 7 Receiver operating characteristics for a Yes/No paradigm, assuming that repeated presentations of noise give rise to a normal distribution of events along a decision axis, and that repeated presentations of signal + noise give rise to another normal distribution of events along the decision axis with the same variance as that of the noise distribution. The probability p of a Hit is plotted as a function of the probability of a False Alarm in this paradigm for four different values of d'. Each curve is generated by determining and plotting p(Hit) and p(False Alarm) as the location of the criterion is swept along the decision axis.

It is important to note that signal-analysis techniques based on the Yes/No paradigm can also be applied to neural data. For instance, suppose a researcher is measuring the response rate of a single unit in visual cortex that is sensitive to luminance contrast. Using the method of constant stimuli, the researcher could repeatedly present the each of the five gratings shown in the top row of Figure 4 an equal number of times, and record the response rate during that presentation period. Distributions of response rates could then be determined for each of the stimuli, and a threshold response could be set by the experimenter. This would be equivalent to setting the location of T in Figure 5. The experimenter could then determine the area under each distribution to the right of the threshold criterion and then plot this area (proportion of times threshold is exceeded) as a function of stimulus contrast, thereby mapping out a function like the one shown in Figure 5(f). In this way, the experimenter can determine the shape of the function relating d' to physical contrast. Note, however, that in this kind of application, it is the experimenter, rather than the subject of the experiment, who is setting the location of the criterion.

The advantage of a signal-detection analysis of a Yes/No experiment is that it allows the experimenter, in analyzing the data, to separate shifts of criterion location (also referred to as response biases) from sensitivity. Both changes in d' and shifts of criterion location will affect both $p(\mathrm{H})$ and $p(\mathrm{FA})$ – the observer's performance will be sensitive to both. However, to estimate both accurately requires the collection of enough data to plot an ROC curve. Often, experimenters are primarily interested in sensitivity measures (such as d') and not interested in determining response criteria. In such circumstances, an m-interval or m-alternative forced-choice paradigm is typically used.

***m*-Interval forced-choice method** In an m-interval forced-choice paradigm, a trial consists of the presentation of m intervals separated by a short but constant interinterval time period. Most experimenters prefer to use two intervals only, and two stimulus values (N and SN), so that this method is referred to as a two-interval, two-alternative, forced-choice (2I2AFC, or sometimes just 2IFC) procedure. In a 2IFC paradigm, SN is randomly assigned to one of the two intervals, with N occurring in the other. Hence, there are two types of trials, N_SN and SN_N, with the probability of each trial type typically being set to 0.5. The observer's task is to decide which type of trial it is. The ideal strategy is for the observer to note the value of the event that occurs along the decision axis in each of the intervals and subtract one from the other. Let the difference between the means of the N and SN distributions be d'. Suppose the observer subtracts the observation in interval one from the observation in interval two. On N_SN trials, this strategy will give rise to a normal distribution with a mean of d' and a standard deviation $= \sqrt{2}$. On SN_N trials, this strategy will give rise to a distribution with a mean $= -d'$ and a standard deviation of $\sqrt{2}$. If these two distributions are then normalized so that their standard deviations are 1.0, the separation between the means of these two distributions is $\sqrt{2}d'$. The distribution of this difference score is shown for N_SN and SN_N presentations in Figure 8, where the d' value separating the original N and SN distributions in the Yes/No task is 1.0. Note that if the observer chooses to respond that the signal is in interval 2 if the observation in interval 2 is greater than the observation in interval 1, this places the location of the criterion, C, for deciding which type of trial type occurred, midway between the distributions for N_SN and SN_N (i.e., the decision is unbiased). Provided that the two types of trials occur equally frequently and the payoff for correct identification of both kinds of trials is unbiased, there is every reason to believe that observers will set their criterion to zero, i.e., to the unbiased position. Hence, the 2IFC is said to minimize the possibility of the occurrence of response biases. Because the expectation in a 2IFC experiment is that the observer is unbiased, d' values

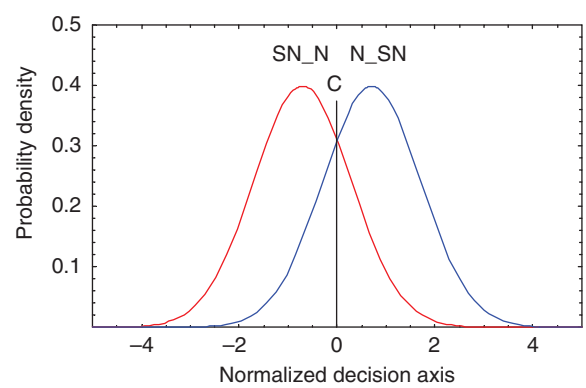

Figure 8 Shown here are hypothetical distributions of events along a normalized decision axis in a two-interval forced-choice experiment. For noise (N) and signal + noise (SN), both the SN_N distribution (signal in first interval, noise in the second interval) and the N_SN distribution (noise in the first interval, signal in the second) were generated by assuming that the observer subtracts an observation along the decision axis during the first interval from an observation along the decision axis during the second interval on both kinds of trials. In other words, the observer responds that the signal is in the second interval when this difference observation is >0. The two distributions of difference judgments corresponding to the two types of trials (SN_N and N_SN) have been normalized to have unit variances, with the separation between the means of the N and SN distributions along the original decision axis being equal to 1 ($d'_{\mathrm{Yes/No}} = 1$); C, criterion location.

are typically not computed, and experimenters usually plot percent correct identification as a function of stimulus intensity.

A very significant advantage of the 2IFC procedure is that it is amenable to adaptive techniques in which the intensity of the signal is changed from trial to trial in an attempt to rapidly converge to the stimulus intensity corresponding to a particular point on the psychometric function relating percent correct to stimulus intensity. Consider the following rule for adjusting the stimulus contrast on successive trials in a 2IFC task. If the observer is correct on three successive trials, lower the contrast. However, if the observer makes a mistake on a trial, increase the contrast. Typically, the amount by which the contrast is increased or decreased becomes smaller each time there is a reversal in direction. Figure 9 plots the performance of an observer in such a task when the initial stimulus contrast was set to 50%, and the first step size is a 10% change in contrast, with the step size decreased by one-half with each reversal of direction, until a minimum step size is reached. The session is usually terminated after a certain number of reversals have occurred, and the threshold is estimated by averaging the last few reversals in intensity. For an attentive observer this '3 down, 1 up' rule will converge on the 79% point of the psychometric function.

Many different algorithms have been developed, including some based on maximum-likelihood procedures, to achieve efficient and unbiased convergence on the targeted point on the psychometric function using the 2IFC procedure. In general, the algorithms that provide the most rapid convergence are sensitive

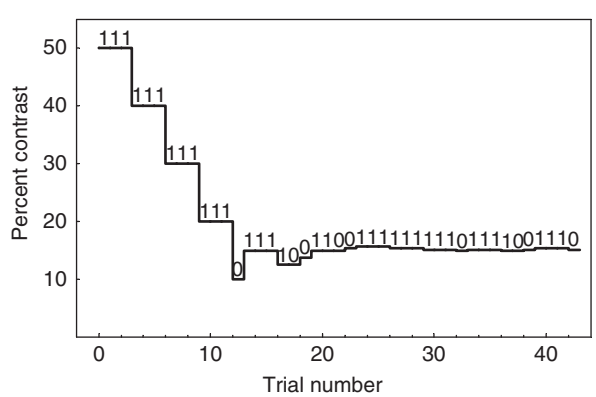

Figure 9 Performance of a hypothetical observer in a method of adjustment paradigm employing a '3 down, 1 up' procedure in a two-interval forced-choice experiment paradigm. In this procedure, three correct responses in a row (a correct response is denoted by the numeral 1) reduce the contrast by a fixed amount on the next trial. If an incorrect response (0) occurs, the contrast is incremented on the next trial. On each reversal of direction of change, the size of the increment is halved on the next trial until a minimum step size for a contrast change is reached. The session is terminated when performance has stabilized.

to lapses in attention. Therefore, the appropriate method to choose will depend on the motivation and attentiveness expected of the observers.

Vigilance paradigm In Yes/No and 2IFC methods, the trial periods are discrete and usually short. In the vigilance paradigm, there is usually one long trial period, during which stimuli are occasionally and infrequently presented. In order to perform well on this task, the observer has to remain focused on the task (i.e., remain vigilant). The dependent measure is the proportion of times the observer detects the presentation of the signal. The vigilance paradigm is typically employed in an attempt to simulate situations in which constant monitoring is required, and the occurrence of signals is unpredictable and infrequent (e.g., monitoring a radar screen to detect intruder aircraft).

Tracking techniques In the techniques just described, stimulus presentations are usually brief, intermittent, and under the control of the experimenter. One can determine the threshold value of a stimulus, however, by using a tracking technique in which stimulus intensity is controlled by the observer. For instance, observers can be told that as long as they press down on a button, the intensity of a tone will continue to decrease and will increase only when the button is released. They are then told to hold down the button when they can hear the tone and to release it when the tone becomes inaudible. In this way, the threshold can be tracked during the course of a session.

Discrimination Experiments

In addition to determining the detectability of a stimulus, one often wants to determine how discriminable pairs of the stimuli are. Here, rather than having a no-signal condition (with or without external noise) versus a signal condition (again, with or without noise), the two conditions become stimulus 1 versus stimulus 2 (S1 and S2). Because of inherent variability in the nervous system, repeated presentations of the two stimuli are assumed to give rise to dispersions of events along a decision axis (even in the absence of external noise), presumably with Gaussian distributions. Hence, by simply substituting S1 for N, and S2 for SN, stimulus discriminability can be investigated using either the Yes/No or the 2IFC method.

Scaling Experiments

In many cases the experimenter would like to know how a dimension of stimulus experience is represented in the nervous system. Two classes of procedures exist for determining this: direct scaling methods and indirect scaling methods. In direct scaling methods the

observer is asked to directly report, say, on the loudness of a sound by assigning a number to it, or by indicating where the loudness of that sound would lie along some scale. The most commonly used direct methods require observers to assign numbers to the stimuli that reflect the magnitude of a sensory property of a stimulus. For example, one could present observers with pure tones ranging from 40 to 100 dB, and ask them to assign numbers to these tones such that the numbers are directly proportional to the loudness they experience (magnitude estimation). Or, one could ask them to assign a number that represents the ratio of the loudness of one tone to another (ratio estimation), or the difference in loudness between two tones (interval estimation), or one could ask them to rate loudness on a 1 to 10 scale (category estimation). The assumption in such techniques is that observers can assign numbers in such a way that they are proportional to sensory experience – that is, $N_L = k \times L$, where N_L is the numerical estimate of loudness generated by the observer, and L is the magnitude of the psychological loudness experience evoked by the stimulus. These techniques are easy to employ, require relatively few trials, and typically generate reliable and replicable relationships between numerical estimates and stimulus intensity. Their weakness is that they assume that observers can accurately, and without bias, assign numbers to reflect stimulus qualities.

Indirect scaling methods derive sensory scales without assuming that the observer can directly assign numbers to sensory experience. For example, consider a set of 12 pure tones differing only in intensity. Suppose each tone gives rise to a value along a unidimensional loudness scale, these 12 tones form $12 \times 11/2 = 66$ distinct pairs. Suppose the observer is asked, when presented with two pairs of tones, to indicate whether the loudness difference between the tones in pair 1 was larger or smaller than the loudness difference between the tones in pair 2, and these judgments are based on a comparison of the psychological distance between the tones in pairs 1 and 2. It turns out that there is sufficient information in the comparisons of pairs to permit the experimenter to recover the location of the tones along this subjective dimension such that the recovered locations, P, are related to the loudnesses of the tones, L, by $P = k \times L + b$. In Figure 10, 12 points were randomly chosen along the line segment from 0 to 1. The 66 interpoint differences between the points were then computed and used to decide which of two intervals was larger. These binary comparisons of intervals were used to recover the location of the original points along a line segment from 0 to 1. Figure 10 plots the recovered locations of the points as a function of their original values, and shows that a comparison of intervals contains sufficient information to recover the original points, such that $y = k \times x + b$, where y is the recovered value of the point, and x is the original value of the point. When 12 stimuli are involved, and the observer can accurately judge which interval is larger, agreement between the recovered and original points is almost perfect.

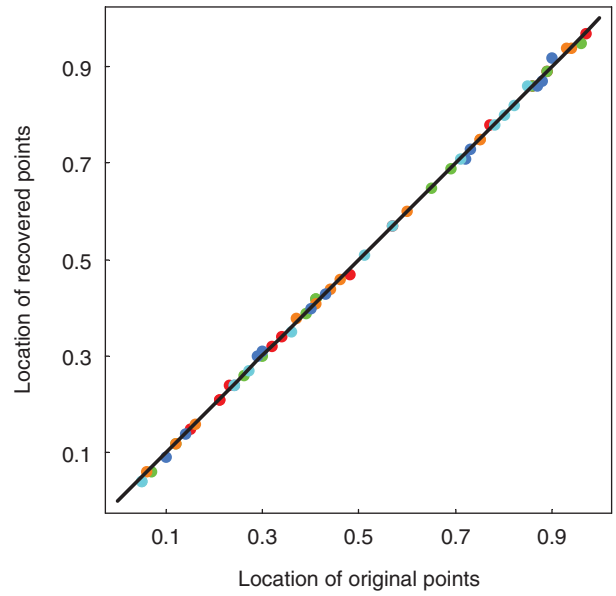

Figure 10 Five simulations of the application of a nonmetric scaling procedure to recover the location of a set of 12 points, randomly sampled from the line segment (0, 1), from comparisons of the intervals between pairs of points. Different colors are assigned to the five different sets of 12 points. In each simulation, all pairwise distances were determined among the set of 12 points. These distances were then used to determine whether the interval between a pair of points (a, b) was greater than, equal to, or less than the interval between another pair of points (c, d). These inequality judgments served as input to a nonmetric scaling program, which returned the best fitting locations of the points along the line segment (0, 1). These recovered points are plotted as a function of the original points. The squared correlation coefficients between the recovered and original points exceed 0.999 in all five cases.

See also: Decision-Making and Vision; Ideal Observer Theory; Information Coding; Statistical Tests and Inferences.

Further Reading

Bissett RJ and Schneider B (1991) Spatial and conjoint models based on pairwise comparisons of dissimilarities and combined effects: Complete and incomplete designs. *Psychometrika* 56: 685–698.

Dunn-Rankin P, Knezek GA, Wallace S, et al. (2004) *Scaling Methods,* 2nd edn. Mahwah NJ: Lawrence Erlbaum Associates.

Gescheider G (1997) *Psychophysics: The Fundamentals,* 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Goldstein EB (2005) Cross-talk between psychophysics and physiology in the study of perception. In: Goldstein EB (ed.) *Blackwell Handbook of Sensation & Perception*, pp. 1–23. Malden, MA: Blackwell.

Macmillan NA (2001) Psychometric functions and adaptive methods [Special issue]. *Perception & Psychophysics* 63(8).

Macmillan NA and Creelman CD (2005) *Detection Theory: A User's Guide*, 2nd edn., Mahwah NJM: Lawrence Erlbaum Associates.

Marks LE and Gescheider GA (2002) Psychophysical scaling. In: Pashler H and Wixted J (eds.) *Stevens' Handbook of Experimental Psychology: Volume 4. Methodology in Experimental Psychology,* 3rd edn., pp. 91–138. New York: Wiley.

Schneider B, Parker S, and Stein D (1974) The measurement of loudness using direct comparisons of sensory intervals. *Journal of Mathematical Psychology* 11: 259–273.

Taberner AM and Liberman MC (2005) Response properties of single auditory nerve fibers in the mouse. *Journal of Neurophysiology* 93: 557–569.

Wickens TD (2002) *Elementary Signal Detection Theory.* New York: Oxford University Press.

Humans

A P Shimamura, University of California, Berkeley, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

The frontal lobes comprise roughly a third of the human cerebral cortex. Through evolution this brain region has increased disproportionately, and as a result of this expansion it has been viewed as the basis for human intellect and thought. Studies of individuals with brain injury have suggested that frontal lobe damage affects many aspects of cognition, including attention, memory, language, problem solving, emotion, and decision making. Yet, its role as the seat of human reasoning and thought became suspect, as findings indicated that frontal lobe damage did not grossly affect measures of general intelligence. Now, there is general consensus that the frontal lobes – or more specifically the anterior portion of the frontal lobes identified as the prefrontal cortex (PFC) – contributes to the regulation or control of mental processing.

The PFC can be parceled into five general regions – the dorsolateral (dlPFC), ventrolateral (vlPFC), posterior (pPFC), anterior (aPFC), and ventromedial (vmPFC) prefrontal regions. These first four regions are spatially located as north, south, east, and west regions within the PFC when the left hemisphere is viewed from the side (i.e., laterally). The fifth region, vmPFC, refers to the lowest and medial (i.e., inner) portions of the PFC. These regions are intricately connected to regions outside the PFC (i.e., in the posterior cortex) and also have interconnections among themselves. These regions are distinct both in terms of anatomical and functional characteristics. In particular, functional neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), have been successful in demonstrating that these brain regions play different roles in the control of cognitive (and emotional) processes. With respect to memory, the dlPFC, vlPFC, and aPFC have been most associated with control processes associated with encoding, organizing, and retrieving memory.

Consider the following question: What did you eat for dinner 3 days ago? It is unlikely that you were able to answer this question immediately. However, with some time and searching through your memory, you may have been able to retrieve the information. Most of us would begin by trying to figure out the day of the week. "Let see, today is Wednesday, three days ago would be Tuesday, Monday, Sunday... yes Sunday, what was I doing on Sunday...." If you ultimately retrieved the information – or even if you tried but were unsuccessful – you likely engaged in a variety of mental processes, particularly those involved in selecting, maintaining, organizing, and ultimately retrieving the desired memory. These processes enabled you to guide and control your mental path toward successful retrieval. In psychological terms, these functions enabled executive control, a process that is critically dependent on the PFC. By analogy, the PFC is your brain's chief executive – it functions to select information, monitor its use, disregard irrelevant information, make decisions, and initiate actions. Without it, your memory – just like a poorly run business – is disorganized, inefficient, and more susceptible to disruptions.

Stimulus Overload and Working Memory

One reason for the necessity of executive control is that we cannot possibly hold in mind everything that we know or perceive. Indeed, it would be an inordinate load on our cognitive system if we even attempted to consider everything that we experienced at any given moment. Because of this stimulus overload, we need to select certain information to be analyzed. Other information is filtered or ignored. For instance, at this moment you are reading and presumably paying attention to this sentence. Yet, pause briefly and pay attention to other sensory information. Perhaps, you can now hear surrounding noises such as a fan or people talking. You can direct your attention to other things, such as the rhythmic changes of pressure on your chest as you breathe. Clearly, at any given moment we experience a multitude of sensations. Not only can we select aspects from the environment, we can select and consider a plethora of memories. It almost seems incredible that in the midst of this information overload, we are able to select, maintain, and control what to bring into our realm of consciousness. The term working memory is used to describe the information that we choose to keep in mind at any given moment. Working memory is transient, and executive control is the dynamic process by which information is selected, maintained, updated, and organized in working memory.

Patients with damage to the PFC (more specifically damage to dlPFC and vlPFC) exhibit a syndrome in which executive control is severely impaired. These patients have problems in paying attention, holding things in mind, organizing thoughts, and retrieving memories. They experience particular problems when

confronted with multiple sensations or choices and must decide which ones to select and which ones to disregard. One particular problem concerns their ability to hold information in mind. A simple task to test this ability is the digit span test, in which individuals are asked to report a short string of digits immediately after presentation. Most individuals can hold about six or seven digits in mind, just as we might when we have to hold a phone number in mind before placing a call. Patients with PFC damage have reduced digit span – they have difficulty keeping information in mind. This deficit pervades many domains, including the ability to maintain locations, colors, or sounds in mind.

Neuroimaging studies have confirmed the role of the PFC in working memory. When individuals must hold or maintain information in mind for a brief delay, prefrontal activity is increased. Interestingly, different prefrontal regions appear to be responsible for maintaining different kinds of information. For example, increased activity in left vlPFC is observed when individuals are asked to hold names of objects in mind, whereas increased activity in right vlPFC is observed when individuals are asked to keep spatial locations in mind. In many instances, these prefrontal activations are linked to activations in posterior regions of the brain, suggesting that the holding of information in working memory involves a brain circuit that includes the information to be activated (presumed to be stored in posterior cortex) and the executive control process that keeps the information active (presumed to be established by PFC).

In addition to its role in maintaining information in mind, the PFC is also involved in updating and manipulating information in working memory. In many everyday situations, it is necessary to rearrange or update information in mind. In laboratory studies, this aspect of working memory has been investigated by asking individuals to reorganize information, such as putting a random string of letters in alphabetical order. Another manipulation task is called the self-ordered pointing task in which individuals are presented with an array of stimuli (e.g., eight objects) and asked to point to one of the stimuli. On subsequent trials, the same objects are shown in different positions, and individuals are asked to point to a different object on each presentation. This task not only requires the holding of objects in mind but also remembering which objects have already been selected. Thus, after each presentation, the contents in working memory must be updated or reorganized to include another item. Patients with PFC damage have difficulty performing this task. Also, neuroimaging studies suggest that the dlPFC is particularly active during such manipulations of information in working memory compared to conditions in which only the maintenance of information is required. Taken together, findings suggest that the mere maintenance or holding of information involves the vlPFC, whereas the manipulation of that information additionally recruits the dlPFC.

Encoding Memories

Executive control is also critical for efficient learning – that is, the encoding of new memories. For example, how would you learn the following series of words: peach, bus, shirt, car, apple, socks, train, shoe, pear? Perhaps, you continually repeated the words in your mind. Alternatively, you may have noticed that the words could be grouped into three meaningful categories: vehicles, fruits, and clothing. In fact, when learning such word lists, many people organize words during encoding and then report the words by grouping all of the words in each category – such as reporting the vehicles first, then fruits and clothing. This kind of organizational strategy would involve maintaining, manipulating, and updating items while they are being presented. Thus, efficient encoding of memories would depend critically upon the kind of executive control processes involved in working memory tasks, such as digit span and the self-ordered pointing task.

Patients with frontal lobe lesions exhibit poor organizational strategies during learning. They fail to elaborate information, such as grouping information into meaningful categories. Moreover, long-lasting memory formation requires the integration of the new information with existing knowledge. This interchange between new and existing information is called elaborative encoding and depends upon the PFC for efficient analysis. Neuroimaging studies corroborate the role of the PFC in elaborative encoding. For example, the left vlPFC has been shown to be particularly active when individuals are asked to consider the meaning of items, such as determining whether a word is concrete (e.g., dollar) or abstract (e.g., freedom) compared to conditions in which they determine superficial features, such as the number of vowels in a word.

The extent of PFC activity during encoding can predict the success of later retrieval. In neuroimaging studies, activations during the initial encoding of items were analyzed separately on the basis of whether the items were later remembered or forgotten. Remembered items were associated with greater prefrontal activity (generally in left vlPFC) during encoding compared to forgotten items. That is, if you recruited PFC during the encoding of an item, you increased your chances of later remembering that item. These findings further support the notion that

efficient encoding depends upon prefrontal activation during learning. The method of backsorting brain activations during learning in terms of whether items are subsequently remembered or forgotten offers a useful means by which to relate efficient encoding strategies to long-lasting memory retrieval.

Another aspect of executive control is related to the suppression of distracting or irrelevant information. In such situations, multiple activations of information place heavy demands on updating and manipulating information in working memory. Proactive interference is the term used to describe instances when earlier information impedes or interferes with the encoding of new information. For example, in memory tests individuals are asked to learn word associates, such as thief–crime, which are then tested by presenting the first word in a pair and asking for the second word, such as theif– ? Proactive interference is assessed by presenting a second learning phase involving the same cues but different responses (e.g., thief–bandit). Patients with PFC damage exhibit particular impairment when they try to recall the second set of responses. That is, they exhibit heightened proactive interference due to the learning of the first set of associates. Evidence for interference is demonstrated by the fact that these patients make many intrusion errors – that is, they use words from the first set during testing of the second set. These problems can be explained by a lack of inhibiting or suppressing the activation of related but now irrelevant information. With respect to executive control, the mitigation of proactive interference requires selecting and updating relevant information. The importance of the PFC in inhibitory control of distracting information has also been demonstrated in fMRI analyses using similar learning paradigms. For example, when individuals are asked to learn word associates such as dog–boxer then later dog–Labrador, they exhibit increased left dlPFC activity when attempting to learn the second related word pair.

Selecting Semantic Memory

A hallmark feature of PFC damage is a problem in searching through memory and selecting specific information. This deficit can be demonstrated in the verbal fluency task, in which individuals are given a minute to retrieve words from a specific semantic category (e.g., 'animals') or retrieve words that begin with a specific letter (e.g., 'A'). This task requires a search through semantic memory – that is, our database of memory for facts and knowledge. Similar to tasks involving the manipulation of working memory or the encoding of new information, efficient retrieval involves the development of strategies to control what words are generated. For example, after trying to come up with just any animal, it would be useful to consider subcategories of animals, such as pets, farm animals, or reptiles. Such strategies are efficient because they facilitate the selection of different words and prevent the report of words already generated. Patients with PFC damage have difficulty controlling their memory searches on this task. They report only 10–12 animal names in a minute, whereas most individuals would be able to retrieve twice as many animal names. Moreover, patients with PFC damage tend to repeat words over and over, as if they have difficult in suppressing the activation of words that were already selected.

The verbal fluency task appears to be a rather easy one, though it is important to appreciate the role of executive control in selecting and updating such information. Not only is it necessary to generate and select information, it is also necessary to monitor responses so that one avoids repeating the same word. Thus after each response all prior responses must be kept in mind so that they will not be repeated. As such, this task has some similarities with the self-ordered pointing task and tasks associated with proactive interference, in which prior information interferes with current performance. In these tasks, it is necessary to monitor and control the activation of previously relevant but currently irrelevant information.

In neuroimaging studies, vlPFC is particularly active during the generation and selection of semantic knowledge. One often-used task to assess retrieval from semantic memory is the verb generation task, in which individuals are presented a noun cue (e.g., 'nail') and asked to generate an associated verb (e.g., 'pound'). As in the verbal fluency task, verb generation requires a search and selection process through one's semantic database. Also, it is necessary to control competing or interfering items during retrieval. In fact, vlPFC activity increases to the extent that there are many competing responses from which to choose. Such regulation of semantic retrieval extends to other linguistic tasks, such as making decisions about the conceptual relatedness between items or interpreting difficult or ambiguous sentences.

The retrieval of general factual knowledge can be assessed by tests of remote pubic events (e.g., "Who shot John Lennon?" [answer: Chapman]). In one study, patients with PFC lesions and control subjects were asked to recall public events or faces of famous people. These tests assessed recent to very remote semantic knowledge. Thus, exposure to these events and people occurred well before the onset of neurological damage. As such, it can be presumed that deficits in performance are due to problems in retrieval rather than encoding. Patients with PFC damage exhibited

recall impairment that was similar across all time periods tested. However, the patients were not impaired on tests of recognition memory, in which they selected the correct answers from several choices. The disproportionate impairment on tests of recall compared to recognition memory suggests that the PFC is involved in controlling the search through semantic memory rather than impairment in the knowledge itself. That is, selecting, updating, and manipulating information in working memory play a significant role in strategic retrieval of information, such as recalling a public event or a famous face. Such retrievals require searching through semantic knowledge, generating relevant information, and disregarding irrelevant information. Recognition memory can be based on a familiarity judgment. That is, on recognition tests it is simply necessary to determine which choice seems most closely related or familiar.

Another example of the distinction between impaired in control of semantic memory and damage to semantic memory itself can be demonstrated in other analyses of recognition versus recall. Consider again the verbal fluency task, in which patients are asked to generate animal names. This task requires recall of information, and patients with PFC damage are impaired. Yet, these patients exhibit intact semantic knowledge of animals when asked to make recognition judgments of animal characteristics. In one study, patients were shown three animal names, such as dog, cow, and pig, and asked to determine which two animals go together. They were told that there were no correct answers and to simply make decisions based on which two go together. The patients exhibited intact semantic memory ability in that they based their decisions on appropriate subcategories of animals, such as associating farm animals (e.g., indicating that cow and pig go together in the example above). In such studies, executive control is minimized by having individuals make simple choice decisions rather than searching, updating, and organizing self-generated retrievals.

Retrieving Autobiographical Memories

Another aspect of memory that requires extensive executive control is the retrieval of autobiographical memory – that is, recollecting an event or experience from the past. The opening exercise of remembering what you had for dinner 3 days ago is an example of retrieving autobiographical memory. Closely related to this form of memory is the notion of source recollection, in which individuals are asked to recollect specific time and place features of an event, such as when an event occurred or who was present during an event. Remembering such information requires executive control in that it is likely that you will have to make a guided analysis to reconstruct an autobiographical memory.

Neuropsychological studies have also demonstrated a prominent role of the PFC in mediating aspects of autobiographical retrieval. Although patients with PFC damage do not exhibit severe amnesia, they are impaired when asked to recollect past experiences. As in the retrieval of semantic knowledge, patients with PFC damage exhibit a particular impairment of recall compared to recognition memory. With respect to source recollection, patients with PFC damage exhibit particular impairment in remembering details of a learning event, such as where and when some information was presented. Such findings are consistent with disorders in memory for the temporal order of events, a common disorder in patients with PFC damage.

In neuroimaging studies, the dlPFC and aPFC are particularly involved in retrieval of autobiographical memory. Studies of autobiographical memory have assessed brain activity during the reminiscence of life experiences, reexperiencing spatial locations in virtual realities, and recognizing photographs taken by oneself or others. Such retrieval tasks have implicated both right and left PFC along with a host of other brain regions, including parietal cortex, medial temporal cortex, and retrospenial cortex. Similarly, studies of source recollection have shown similar brain activations. In such studies, individuals are asked to recollect features of a learning event, such as the color a stimulus, its spatial location, or its temporal order. In studies of source recollection, left PFC appears to be more active than right PFC.

To the extent that the PFC is involved in the control of autobiographical retrieval – rather than in the actual storage of such information – it is necessary to demonstrate that this kind of retrieval places particular demands on executive control. That is, it must be demonstrated that the selecting, maintaining, updating, and organizing of memory is particularly important for retrieval of details about one's past experiences. This issue was addressed with a task that facilitated executive control in patients with PFC damage. Individuals were presented objects and were asked to perform an action with each one (e.g., bounce the ball). This manipulation was presumed to increase sensorimotor activation during encoding and thereby increase the distinctiveness of stimuli. Memory for the temporal order of items was assessed by presenting pairs of objects and asking individuals to determine which of the two was presented more recently. Performance by patients with PFC damage did not differ from performance by control subjects on this test. However, the patients exhibited significant impairment on a control task in which they were

presented the objects but did not manipulate them. This benefit appeared to be rather selective to patients with PFC damage because performance by amnesic patients with temporal lobe lesions was not facilitated by the manipulation condition.

A Board of Executives

How can one interpret the role of the PFC in memory processes? As described here, the PFC monitors and controls information processing thus enabling executive or 'top-down' control. Top-down control refers to the manner in which thoughts and knowledge control lower levels of analysis. In one view of top-down processing, called metacognitive control, basic processes – such as object recognition, speech analysis, and semantic memory access – are defined as 'object-level' processes, which are controlled by 'meta-level' processes. Meta-level processes monitor object-level processes, and – as a result of this monitoring – initiate control (see Figure 1). The PFC fits well in this model as it can serve as the meta-level processor to basic processors. It is presumed that top-down control is implemented by reciprocal neural pathways between PFC and posterior regions. Control occurs as a result of selecting and maintaining task-relevant activity (activation) and by filtering irrelevant activity (inhibition).

As suggested by findings presented above, different PFC regions serve different control functions, such as selecting information from the environment, maintaining information in working memory, accessing semantic knowledge, and retrieval. Thus, rather than one executive controller, there appear to be numerous controllers that monitor and control different aspects of information processing. By analogy, one could view the brain as having a board of executives, with each one assigned to a specific function. Executives interact with each other, though each one is responsible for monitoring and controlling a particular part of a business. It is known that different PFC regions are connected to distinct regions in the posterior cortex. By way of these reciprocal interconnections, PFC has the capability of influencing many object-level processes in posterior cortex.

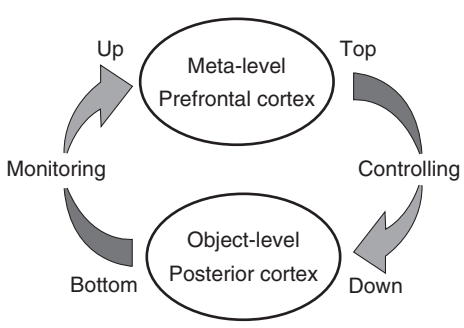

Figure 1 Metacognitive model of executive control. Monitoring of object-level processes is achieved by bottom-up pathways from posterior cortex to PFC. Meta-level processors in the PFC initiate control by top-down pathways to object-level processes. Control of memory is enabled in this model by the role of the PFC in selecting, maintaining, accessing, and retrieving object-level processes for efficient learning and memory.

Physiological evidence exists for the role of the PFC in metacognitive control of neural activity. Studies using electroencephalogram (EEG) scalp recordings assess neural activity in response to stimulus presentations. Patients with PFC lesions exhibit abnormal increases in neural response in posterior cortex. Thus, as a result of PFC damage, bottom-up neural signals are disinhibited due to a failure to modulate or control posterior cortical activity. In PET and fMRI studies of normal individuals, increased activity in the dlPFC is often correlated with decreased activity in posterior cortical regions, as if PFC can not only select posterior activity, but it can also suppress irrelevant activity.

In summary, the frontal lobes play a significant role in human memory. Rather than being the storehouse of semantic knowledge or autobiographical memories, the PFC monitors and controls processes associated with accessing, encoding, and retrieving memories. PFC initiates and channels information so that it can be held in working memory and efficiently manipulated or updated. PFC facilitates encoding and organization during learning, and access and recall during retrieval. Without it, thoughts are disorganized, and memories are poorly learned and retrieved. Executive control, top-down processing, and metacognitive control are terms used to describe how PFC monitors and controls information processing. In the service of memory processes, PFC is critical for efficient access to and retrieving from our storehouse of knowledge, as well as adding new knowledge to it. It is as critical for memory as an efficient librarian is for a library.

See also: Cognition: An Overview of Neuroimaging Techniques; Episodic Memory; Executive Function and Higher-Order Cognition: Neuroimaging; Frontal Lobe Syndrome; Memory Representation; Memory Consolidation: Systems; Memory Disorders; Prefrontal Cortex: Structure and Anatomy; Recognition Memory; Semantic Memory; Short Term and Working Memory.

Further Reading

Baddeley A (1986) *Working Memory.* Oxford: Oxford University Press.

D'Esposito M, Postle BR, and Rypma B (2000) Prefrontal cortical contributions to working memory: Evidence from event-related fMRI studies. *Experimental Brain Research* 133: 3–11.

Duncan J (2001) An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience* 2: 820–829.

Johnson MK, Hashtroudi S, and Lindsay DS (1993) Source monitoring. *Psychological Bulletin* 114: 3–28.

Luria AR (1973) *The Working Brain.* New York: Penguin Books.

Maguire EA (2001) Neuroimaging studies of autobiographical event memory. *Philosophical Transactions of the Royal Society of London B* 356: 1441–1451.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24: 167–202.

Paller KA and Wagner AD (2002) Observing the transformation of experience into memory. *Trends in Cognitive Science* 6: 93–102.

Roberts AC, Robbins TW, and Weiskrantz L (eds.) (1998) *The Prefrontal Cortex: Executive and Cognitive Function.* Oxford: Oxford University Press.

Rugg MD, Otten LJ, and Henson RNA (2002) The neural bases of episodic memory: Evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society B* 357: 1097–1110.

Shimamura AP (2000) The role of the prefrontal cortex in dynamic filtering. *Psychobiology* 28: 207–218.

Squire LR and Schacter DL (eds.) (2002) *Neuropsychology of Memory,* 3rd edn. New York: The Guilford Press.

Teuber H-L (1964) The riddle of frontal lobe function in man. In: Warren JM and Akert K (eds.) *The Frontal Granular Cortex and Behavior,* pp. 410–444. New York: McGraw-Hill.

Thompson-Schill SL, Bedny M, and Goldberg RF (2005) The frontal lobes and the regulation of mental activity. *Current Opinion in Neurobiology* 15: 219–224.

Wagner AD (2002) Cognitive control and episodic memory. In: Squire LR and Schacter DL (eds.) *Neuropsychology of Memory,* 3rd edn., pp. 174–192. New York: The Guilford Press.

Huntington's Disease

J F Gusella and M E MacDonald, Massachusetts General Hospital, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

In 1872, George Huntington, a Long Island physician, drew on his own experience and that of his father and grandfather to provide a graphic description of the neurodegenerative genetic disease that now bears his name. Originally known as 'Huntington's chorea,' this autosomal disorder is inherited in a completely dominant fashion and has a prevalence of about 1 in 10 000. Huntington's disease (HD) is characterized by progressive involuntary choreiform (dancelike) movements, psychologic disturbance, cognitive decline, and a distinctive neuropathology. The onset of symptoms can occur at any time in life but most commonly begins during the fourth to fifth decade (mean ~38 to 40 years). In a small percentage of cases, termed juvenile onset HD and usually inherited paternally, the first symptoms occur before age 15. In typical adult-onset HD, the initial signs of motor disturbance are slight, such as awkwardness of gait, clumsiness, facial twitching, or subtle involuntary movements of the fingers. Progression of the disorder makes the involuntary movements more frequent and exaggerated, impairing and eventually climinating normal day-to-day activity by interfering with the ability to walk, stand, write, speak, and swallow. In some cases, particularly those of juvenile onset, extreme rigidity is seen rather than chorea. Emotional and behavioral changes often accompany or precede the onset of movement disorder and similarly can have a profound impact on ability to function and on interaction with family members. Impulsive, erratic behavior, impaired memory, poor concentration, and moodiness can all be early signs of HD. In particular, chronic depression occasionally precedes motor symptoms by several years. The distinctive symptoms of HD are caused by the loss of neurons that is most prominent in the striatum but is seen also in the globus pallidus and more diffusely in the cortex. Medium spiny neurons appear to be the first target of the genetic defect, and their loss follows a characteristic gradient proceeding along posteroanterior, dorsoventral, and mediolateral axes of the striatum. There is no effective treatment for preventing the onset or delaying the progression of HD. After a course of 10 to 20 years that leads to severe emaciation and complete incapacitation, the afflicted individual usually succumbs to heart disease, choking, or pneumonia secondary to aspiration.

In 1983, the Huntington's disease gene (*HD*) was mapped to chromosome 4 by tracking the segregation of polymorphic DNA markers in families with the disorder. This landmark study was the first successful application of this nascent technology, and it set off a torrent of similar investigations that has led to the mapping of dozens of other human disease genes. It also acted as the foundation for a decade-long quest to isolate and characterize the genetic defect based on its chromosomal location. This location-cloning strategy passed through a number of stages as the technologies available for human genome analysis evolved rapidly. Somatic cell hybrid mapping panels that dissected 4p 16.3 supported the isolation of large numbers of restriction fragment length polymorphism (RFLP), variable number of tandem repeat (VNTR), and simple sequence repeat (SSR) markers. These were used to create detailed genetic linkage maps and to anchor the construction of physical maps and overlapping clone sets. A progressive refinement of the location for the HD defect was provided by recombination analysis in disease pedigrees and by linkage disequilibrium between the disorder and individual markers. Ultimately, through a novel strategy employing haplotypes of multiallele markers, a small region shared by one-third of disease chromosomes with a common ancestral mutation was identified. Detailed inspection of candidate genes merged with continued search for informative polymorphisms to reveal a polymorphic CAG trinucleotide repeat in the 5′ end of a long, novel gene (*ITI5*), which was given the locus name *HD*. The recognition that this CAG repeat is expanded and unstable on HD chromosomes signaled that the marathon search for the defect was over and a new quest for understanding of mode of action had begun.

On normal chromosomes, the *HD* gene's CAG repeat varies from 6 to 34 trinucleotide units and is inherited in a Mendelian fashion. On HD chromosomes, the repeat is expanded, with 35 to more than 100 units, and is unstable, changing in length in most meiotic transmissions. Usually, the repeat increases or decreases in length by only a few units, but occasionally the increase are larger, sometimes more than doubling the number of CAGs. These larger jumps occur preferentially in male gametogenesis. About 50% of the variability in disease severity is explained by the inherited CAG repeat, the length of which is inversely correlated with age at onset of motor and psychiatric symptoms and age at death. The remaining variability is not accounted for by somatic variation in CAG repeat length, which is extremely limited in magnitude and may be due to modifying genes,

environmental factors, or stochastic processes. The inverse correlation of repeat length and age at onset, combined with the occurrence of occasional large repeat length increases in male transmissions, explains the bias for paternal inheritance of juvenile onset disease.

All cases of HD share the same mechanism of disease, expansion of the CAG segment, although haplotype analysis indicates that there have been multiple independent origins of this HD defect. New mutations to HD have been identified as cases in which expansion of a CAG repeat occurs into the HD size range in a family with no history of the disorder. In such families, the allele subject to expansion typically possesses a repeat length intermediate between the normal and HD size ranges. In some cases, CAG alleles in the intermediate range of 34 to 39 units may be nonpenetrant, with the carrier not displaying HD symptoms in a normal lifespan, again presumably due to modifying genes that moderate the effects of the expanded CAG allele.

The genetic linkage study that mapped *HD* to chromosome 4 introduced the potential for presymptomatic diagnosis of 'at risk' individuals. The linkage test was necessarily cumbersome, expensive, and of limited applicability, because it required the participation of several affected and unaffected family members. However, its advent prompted a cautious, careful assessment of the difficulties inherent in delivering potentially disastrous genetic information many years before any sign of disease in an untreatable disorder with a strong psychiatric component. This experience led to formal guidelines for presymptomatic testing that should be applicable to other late-onset disorders and that are now being applied in HD with increased frequency, because the discovery of the genetic defect has produced a direct test to replace the linkage assay. Assay of CAG repeat length by polymerase chain reaction amplification represents a less costly, more accurate alternative that is applicable to any at risk individual. It can also be used to confirm HD in cases difficult to diagnose or in individuals with no family history.

The question in HD of how the expanded CAG repeat triggers the characteristic pattern of neuronal death remains open. The CAG repeat is located in the 5′ end of the coding sequence of a ~348 kDa protein of unknown function that has been named *huntingtin*. The HD mutation does not simply eliminate the normal function of huntingtin, because individuals with one copy of the gene inactivated by deletion or translocation do not display HD-like symptoms. Indeed, the effect of the expanded CAG stretch must be dominant to explain the similar clinical picture of HD heterozygotes and HD homozygotes. Huntingtin is expressed in the cytoplasm of a wide variety of cells in both the nervous system and peripheral tissues, in contrast to the specific pattern of cell death in HD. The expanded CAG stretch of the disease allele does not eliminate huntingtin expression but alters its structure by extending a stretch of polyglutamine near the N-terminus of the protein. The absence of any obvious effects on the pattern of expression of huntingtin early in the disease suggests that the defect probably acts via this polyglutamine segment as a gain-of-function mutation by conferring a new property on huntingtin that is peculiarly detrimental to striatal neurons. Several other neurodegenerative disorders with different patterns of neuronal loss, including spinobulbar muscular atrophy, spinocerebellar ataxias 1, 2, 3, 6, 7, and 17, and dentatorubropallidoluysian atrophy, have all been found to be caused by expanded, translated CAG trinucleotide repeats. Thus, all of these disorders and HD may share a common mechanism of neuronal toxicity, in which the susceptibility of particular neuronal populations is determined by the context in which the lengthened polyglutamine segment is presented.

Investigations of HD and the other polyglutamine disorders have led to a number of criteria for the mechanism that triggers the pathogenic pathway leading to disease onset: (1) a threshold polyglutamine length below which the disease does not occur in a normal lifespan, (2) progressive severity with increasing polyglutamine length, (3) dominance over the normal copy of the gene present in most heterozygous HD patients, (4) relatively greater sensitivity to increasing polyglutamine length than to the increased disease protein concentration produced from two disease alleles in homozygotes, and (5) specificity of neuronal loss that is influenced by the nature (structure, function, or localization) of the protein containing the polyglutamine tract. These criteria have provided a means for judging potential mechanisms and have led to the notion that the deleterious property of the polyglutamine tract may involve an altered conformation of huntingtin's N-terminal region. This alteration is demonstrated most easily by the capacity of an expanded polyglutamine tract, in the context of a small N-terminal huntingtin fragment, to promote aggregation and conversion to an amyloid structure. Expression of truncated mutant huntingtin also leads to aggregates *in vivo*. For example, the first HD mouse model involved the expression of such an N-terminal fragment of mutant huntingtin producing large intranuclear aggregates in neurons and other cells. Similar aggregated material has been detected in affected brain postmortem, leading to the suggestion that aggregate toxicity might be the

cause of pathogenesis. However, the overexpression of mutant N-terminal fragments of huntingtin are not universally toxic to cells. Moreover, expression of full-length mutant huntingtin does not lead directly to aggregate formation, which occurs only after activation of proteolytic enzymes such as caspases, calpain, and others leads to huntingtin cleavage.

While aggregate formation is a clear phenomenon late in disease pathogenesis, the question remains as to how the cascade of pathogenesis is first triggered by the expanded polyglutamine tract. Generation of mouse models that express mutant huntingtin at physiologic levels suggest that the deleterious conformational property has its first physiologic consequences due to the polyglutamine tract in huntingtin when it is still full length, many months before the appearance of truncated fragments or aggregates. The nature of the effects of mutant huntingtin early in pathogenesis remain the subject of intensive investigation, with early results suggesting that altered intracellular signaling and energy metabolism indicate subtle neuronal dysfunction that progresses only much later to neuronal loss. The full-length mutant huntingtin mouse models also suggest that the late onset of HD results from a pathogenic pathway that begins early in life and progresses in a time-dependent and polyglutamine length-dependent manner, rather than beginning in midlife shortly before the onset of symptoms.

The application of recombinant DNA techniques during the past 20 years has moved our understanding of HD from descriptive knowledge of the clinical and pathologic features of the disorder to identification of the molecular trigger for neuronal cell loss. It has also acted as a model for applying similar approaches in other disorders. Investigations of HD are now targeted at delineating the pathway that leads from the expression of the CAG repeat to the eventual death of specific neurons. As each step is identified, it may present an opportunity to develop a rational treatment that effectively derails initiation of the disorder or halts the currently inescapable progression of pathologic changes that bring about the patient's decline and demise. Consequently, altered properties of mutant huntingtin and the alterations that these cause in biochemical, cellular, and animal models are all being used as assays to screen chemical libraries for small compounds worthy of testing in clinical trials as potentially effective drugs to prevent the onset or delay the progression of HD.

See also: Axonal Transport and Huntington's Disease; Cell Replacement Therapy for Huntington's Disease; Huntington's Disease: Neurodegeneration; Transgenic Models of Neurodegenerative Disease; Triplicate Repeats: Huntington's disease.

Further Reading

Beal MF (1994) Neurochemistry and toxin models in Huntington's disease. *Current Opinion in Neurology* 7: 542–547.

Folstein SE (1989) *Huntington's Disease. A Disorder of Families.* Baltimore: Johns Hopkins University Press.

Gines S, Seong IS, Fossale E, et al. (2003) Specific progressive cAMP reduction implicates energy deficit in presymptomatic Huntington's disease knock-in mice. *Human Molecular Genetics* 12: 497–508.

Gusella JF and MacDonald ME (1993) Hunting for Huntington's disease. *Molecular Genetics Medicine* 3: 139–158.

Gusella JF and MacDonald ME (2000) Molecular genetics: Unmasking polyglutamine triggers in neurodegenerative disease. *Nature Reviews Neuroscience* 1: 109–115.

Gusella JF and MacDonald ME (2006) Huntington's disease: Seeing the pathogenic process through a genetic lens. *Trends in Biochemical Science* 31: 533–540.

Gusella JF, MacDonald ME, Ambrose CM, and Duyao MP (1993) Molecular genetics of Huntington's disease. *Archives of Neurology* 50: 1157–1163.

Gusella JF, Wexler NS, Conneally PM, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306: 234–238.

Hersch S, Jones R, Koroshetz W, and Quaid K (1994) The neurogenetics genie: Testing for the Huntington's disease mutation. *Neurology* 44: 1369–1373.

Huntington G (1872) On chorea. *Medical Surgery Reporter* 26: 317–321.

Huntington's Disease Collaborative Research Group(1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72: 971–983.

Li JL, Hayden MR, Warby SC, et al. (2006) Genome-aide significance for a modifier of age at neurological onset in Huntington's disease at 6q23–24: The HD MAPS study. *BMC Medical Genetics* 7: 71.

Persichetti F, Ambrose CM, Ge P, et al. (1995) Normal and expanded Huntington's disease gene alleles produce distinguishable proteins due to translation across the CAG repeat. *Molecular Medicine* 1: 374–383.

Seong IS, Ivanova E, Lee JM et al. (2005) HD CAG repeat implicates a dominant property of Huntingtin in mitschondrial energy metabolism. *Human Molecular Genetics* 14: 287–288.

Vonsattel JP, Myers RH, Stevens TJ, et al. (1985) Neuropathological classification of Huntington's disease. *Journal of Neuropathology and Experimental Neurology* 44: 559–577.

Wheeler VC, White JK, Gutekunst CA, et al. (2000) Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knock-in mice. *Human Molecular Genetics* 9: 503–513.

Huntington's Disease: Neurodegeneration

M F Chesselet, M A Hickey, and C Zhu, The David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
M S Levine, Semel Neuroscience Institute, The David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Huntington's Disease

Huntington's disease (HD) is characterized by progressive motor dysfunction, cognitive decline, and psychiatric disturbances. The mutation causing the disease is an unstable expansion in the number of CAG repeats in exon 1 of the gene that encodes huntingtin (Figure 1). Whereas the normal number of CAG repeats varies from 6 to 35, repeat lengths of 36–41 lead to an increased risk for the disease, and greater lengths cause HD with full penetrance. Although there is a direct correlation between repeat length and age of onset, with juvenile forms being usually caused by repeat lengths of over 60, there is considerable variation in age at onset for the same number of CAG repeats in the mid-range of the expansions.

Neurodegeneration in HD

Neuronal degeneration is the pathological signature of HD. Progressive striatal and cortical atrophy can be detected by brain imaging *in vivo* and is striking at autopsy (Figure 2). The most clearly affected neuronal population is the medium-sized spiny GABAergic efferent neurons of the striatum, which comprise 95% of neurons in this region. Neuronal loss in the striatum proceeds in a medial to lateral gradient that led to the classification of the disease progression into four stages or grades. In contrast to efferent neurons, striatal interneurons are relatively spared, although whether this extends to all classes of interneurons remains a matter of debate. Although pathology is mostly restricted to the striatum and cortex at early stages, other brain areas such as the globus pallidus, hippocampus, thalamus, and cerebellum may be affected by grades 3 and 4. The involvement of the hippocampus and the cerebellum is particularly noticeable in cases of juvenile HD. Thus, an important point about neurodegeneration in HD is that although striatal efferent neurons are clearly more vulnerable than other neuronal types, this 'selective vulnerability' is not absolute, and greater CAG repeat burden seems to lead to the involvement of a wider range of neurons.

Ultrastructural studies in postmortem human brains, although notoriously difficult, revealed a loss of dendrites, dendritic spines, and synaptic connections in surviving neurons, with indentations of the nuclear membrane, disorganization of the nucleolus, and depletion of the ribosomes of the rough endoplasmic reticulum. Increased spine density and curved, elongated dendrites have also been observed, suggesting proliferative changes in spared neurons. These observations indicate that even surviving neurons are affected by the disease process, and these morphological changes may be related to neuronal dysfunction which occurs in humans long before evidence for neuronal degeneration. Similarly, decreases in neurotransmitter peptides in the axons of striatal efferent neurons have been detected in the few cases of presymptomatic carriers examined at autopsy. Finally, recent imaging studies revealed that alterations in white matter tracts could precede neuronal loss, as assessed by brain atrophy. It is striking, however, that surviving neurons even in advanced grades have high levels of the corresponding mRNA and that degenerating neurons appear to be intermingled with spared ones.

Neuropathological studies in humans have not fully resolved the mechanisms of neuronal cell death occurring. It has been suggested that apoptosis is the primary mechanism of cell death in HD. Evidence, however, remains inconclusive. Indeed, although there is no doubt that DNA nick labeling techniques revealed DNA damage in HD tissue, examination of these DNA fragments by gel electrophoresis did not reveal the characteristic apoptosis profile. This is not incompatible with the histological data as TUNEL or other DNA damage stains are by no means specific for apoptotic cell death. Nevertheless, a number of factors may hinder the detection of apoptotic cell death in postmortem brain, such as postmortem delay, the transient nature of apoptosis, or a low quantity of DNA fragments being present in tissue. Morphological changes are the most reliable method of detecting apoptosis, but they are difficult to assess in postmortem human tissue. Interestingly, caspase and calpain family members are increased in HD tissue. However, the extensive and early astrocytosis observed in HD brains argues against an apoptotic mechanism.

In conclusion, neurodegeneration is a critical aspect of HD but its mechanism remains uncertain. It affects primarily but not exclusively the striatum, and proceeds according to a well-characterized gradient

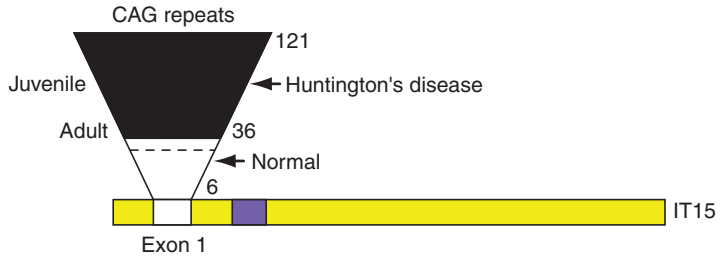

Figure 1 Cartoon of the gene encoding human huntingtin. In the normal population, the CAG repeats in exon 1 range from 6 to 35. CAG repeat lengths of 42 and above cause HD while repeat lengths between 36 and 41 increase the risk of the disease.

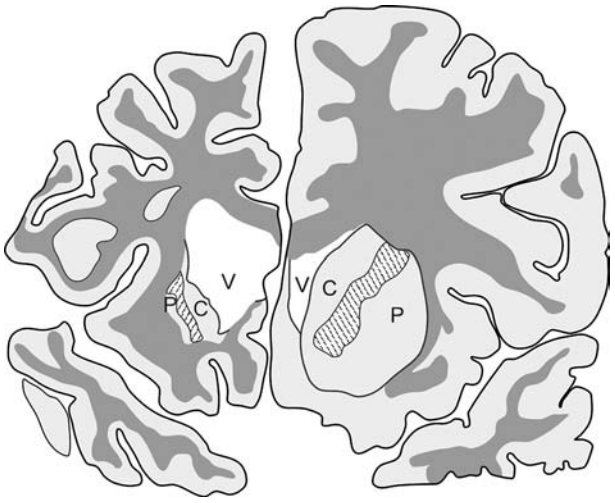

Figure 2 Representation of degeneration in a human HD brain (left) compared with a normal brain (right). Representations are of coronal sections at the level of frontal lobe. Note the profound atrophy of the caudate nucleus, putamen and the cortex, and the ventricular enlargement in the HD brain. C, caudate nucleus; P, putamen; V, lateral ventricle. Adapted from Hickey MA and Chesselet MF (2003) The use of transgenic and knock-in mice to study Huntington's disease. *Cytogenetics and Genome Research* 100(1–4): 276–286, with permission from S. Karger AG, Basel.

in this region, where it mostly affects efferent neurons. Neuronal loss can be followed during the patient's life by assessing regional atrophy with structural brain imaging techniques but a period of neuronal dysfunction is likely to precede cell death, which appears to be stochastic even in the most affected regions. These observations point to the importance of understanding the early mechanisms of neuronal dysfunction caused by the HD mutation, rather than cell death, in order to identify most effective therapeutic targets.

Neurodegeneration in *In Vitro* Models of HD

Since the identification of the HD-causing mutation in 1993, much effort has been devoted to reproducing HD in cellular and animal models. Expression of mutated huntingtin in yeast and mammalian cells leads to neurodegeneration under certain experimental conditions. In general, neurodegeneration is promoted by smaller fragments of mutated huntingtin compared to the full-length protein, suggesting that protein cleavage is an essential step in the pathophysiology of the disease. *In vitro* studies have revealed an association between cell death and the interaction of mutated huntingtin with transcriptional coactivators and transcriptional dysregulation. They also challenged the idea that neurodegeneration is directly linked to the formation of insoluble aggregates of mutated huntingtin. Indeed, aggregates can be neuroprotective despite frequent correlative evidence that neurodegeneration parallels the formation of aggregates. Several *in vitro* studies have suggested that mutated huntingtin can induce cell death by triggering apoptotic mechanisms, with evidence of DNA fragmentation, caspase activity, and cytochrome *c* release. However, the relevance of these mechanisms to the disease situation remains unclear, due to the usually high levels of mutant protein in cell lines.

Neurodegeneration in *In Vivo* Models of HD

The first example of neurodegeneration in an animal model of HD was obtained in *Drosophila*. Expression of mutated huntingtin (exons 1–3 and part of 4) under the eye-specific expression construct pGMR led to progressive degeneration of ommatidia with reduced age of onset of degeneration and increased nuclear accumulation with increased polyglutamine length. Similarly, neurodegeneration was observed in *Caenorhabditis elegans*. These models were instrumental in demonstrating the ability of the chaperones to prevent HD-induced neurodegeneration. They also played a key role in furthering our understanding of mutation-induced mechanisms of neurodegeneration (see below).

Paradoxically, neurodegeneration has proved to be challenging to demonstrate in genetic mammalian

models of HD. Prior to the identification of the HD mutation, neurodegeneration was created in the rodent or monkey striatum by local or peripheral injections of the neurotoxins quinolinic acid or 3-nitropropionic acid. The relevance of these models to HD was inferred from the similarity of their effects to the pattern of neuronal death observed in HD, with preferential loss of efferent neurons and relative sparing of interneurons. These models pointed to the potential role of excitotoxicity for inducing neurodegeneration in HD, an idea that remains timely (see below). Since the discovery of the HD-causing mutation, genetic mammalian models are a more direct route to understand HD pathophysiology, and many different mouse and even rat models have been created.

The first published transgenic mouse model of HD, the R6 line generated by the laboratory of Gillian Bates, remains the best characterized. The R6/2 line expresses exon 1 of the HD gene with approximately 145 CAG repeats. These mice have severe motor anomalies, cognitive decline, weight loss, and seizures; they die around 4 months of age. Levels of mRNA and protein for neuronal molecules present in striatal medium-sized spiny neurons are decreased in a 4-week-old mice, but evidence of neurodegeneration is only obtained late in disease development (Figure 3). This observation indicates that, similar to humans, neurological symptoms precede neurodegeneration in mice, indicating a protracted period of neuronal dysfunction before cell death occurs. Similar observations were made in a number of other mouse models of HD, including both transgenics, and mice with extended CAG repeats inserted in their own HD gene (knock-in mice). Most of these mice show brain atrophy at least at late stages of the disease, and some present evidence of 'dark cells' that are interpreted as degenerating neurons. Few models, however, show a decrease in the number of neurons in the striatum when examined with stereological methods.

Contrary to expectation, work in mice did not unequivocally clarify the mechanism of cell death induced by the HD mutation. In R6/2 mice, 'dark' neurons were found in the striatum and had dense nuclear condensation but no evidence of apoptotic bodies. The absence of DNA fragmentation with either TUNEL staining or *in situ* nick translation does not support an apoptotic mechanism, but some authors did report DNA fragmentation at the very late stages of disease. Also arguing against an apoptotic mechanism, the levels of proteins involved in apoptotic pathways were not changed in R6/2 mice, although increases in these proteins were reported in another R6 line, R6/1, which also showed 'dark' neurons. Increases in apoptotic proteins, TUNEL staining, and apoptotic profiles were reported in another full-length transgenic line, which is no longer in existence. It should be noted that these differences among transgenic models parallel the severity of the phenotype, with the mice that exhibit a more prolonged life and milder phenotype showing more evidence for apoptosis than mice with a more severe course of disease. A recent study proposed that mutated huntingtin causes neurodegeneration through a novel mechanism of cell death related to reduction in Pol II-dependent transcription. The cell death was prolonged, morphologically unlike apoptosis, necrosis or autophagy, and modulated by isoforms of yes-associated protein (YAP), and these isoforms reduced degeneration in an HD *Drosophila* model.

Although most knock-in and transgenic models show little cell loss, viral vectors encoding long

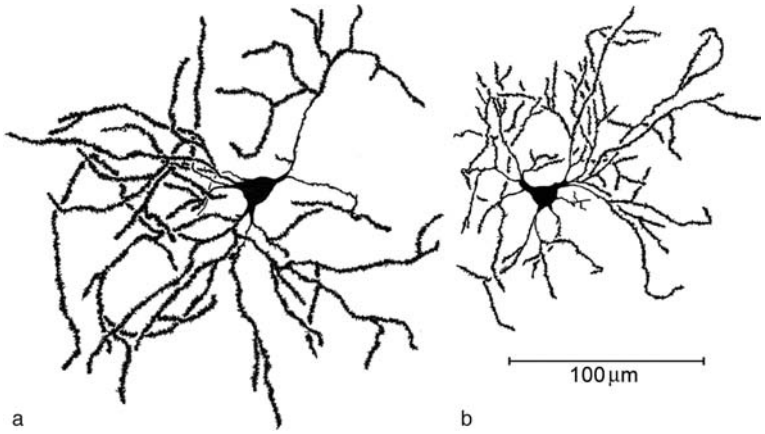

Figure 3 Camera lucida drawings of Golgi impregnated medium spiny neurons in wild-type (a) and R6/2 transgenic HD (b) mice at 12 weeks of age. Note the loss of spines and reduction in the diameter in the dendritic field in the R6/2 transgenic neuron. Adapted from Klapstein GJ, Fisher RS, Zanjani H, et al. (2001) Electrophysiological and morphological changes in striatal spiny neurons in R6/2 Huntington's disease transgenic mice. *Journal of Neurophysiology* 86(6): 2667–2677, with permission from the American Physiological Society.

CAG repeats induce apoptotic cell death *in vivo* when introduced into the striatum. This may be due to the very high expression levels of the abnormal polyglutamine tract in this model.

Pathophysiological Mechanisms of Neurodegeneration in HD

As indicated earlier, the precise mechanism of cell death in HD remains unclear. However, much information has accumulated in the last 13 years on the various ways mutated huntingin alters cellular function. Evidence that the lack of huntingtin is embryonic lethal and that mutated huntingtin can rescue this phenotype suggests that the mutation causes an abnormal gain of function. What is not excluded, however, is that the detrimental effects of this gain of function are compounded by a loss of function of the normal protein, in particular its role in stimulating the production of brain-derived neurotrophic factor (BDNF), an essential growth factor for striatal medium-sized spiny neurons.

A compelling abnormal property of mutated huntingtin is its ability to form insoluble aggregates both in the cytoplasm and in the nucleus. These aggregates are usually detected with antibodies against the N-terminal portion of the protein (which contains the polyglutamine expansion), but they can also contain full-length huntingtin. Aggregates, which range in size from microaggregates to large nuclear inclusions, were described in human brain and later found to contain huntingtin. After their detection in the R6/2 line, they were described in all animal models of HD (Figure 4), as well as in many cellular models. Normal huntingtin does not form aggregates, although it can be recruited into aggregates of mutated protein. Aggregates are present in all diseases of CAG repeat expansion and can be formed by long polyglutamine repeats, either isolated or inserted in a nonrelevant gene.

Despite their obvious association with pathology, aggregates may not be pathological themselves. *In vitro* studies have shown that they do not always cause cell death or they can even be neuroprotective. A lack of causal relationship between aggregates and neuronal dysfunction has been demonstrated in several mouse models where a careful analysis has shown that behavioral deficits precede detectable aggregates, and in a recent description, mice expressing exon 1 and 2 of mutant huntingtin show widespread aggregate development, but no atrophy or motor impairments. Even in postmortem human brains, aggregate location does not correlate with neurodegeneration. Finally, several compounds have beneficial effects in models of HD without causing a parallel decrease in aggregates.

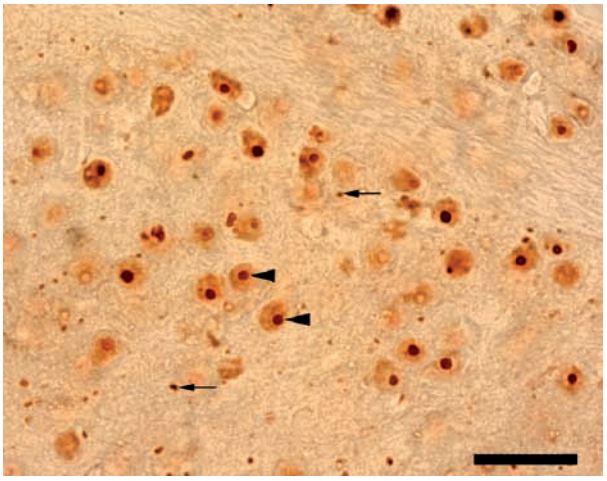

Figure 4 Photomicrograph of aggregated huntingtin in striatal neurons in knock-in (CAG 140) mouse brain at 13 months of age (Zhu and Chesselet, unpublished). Mutant huntingtin forms nuclear inclusions (arrowhead) and neuropil aggregates (arrows) in striatal neurons. Immunohistochemistry with polyclonal EM48 (Gift from Dr. X.-J. Li, Emory University). Scale bar = 20 μm.

Although huntingtin aggregates do not appear to directly cause neurodegeneration, they could contribute to pathology in different ways. Aggregated huntingtin may interfere with axonal transport, block proteasomal degradation of other proteins, and sequester proteins that are important for cell functioning. Indeed, transcriptional dysregulation appears to be a critical pathophysiological mechanism in HD and histone deacetylase inhibitors, which reverse this deficit, have beneficial effects in cellular and animal models. Interactions with transcription coactivators, however, may depend on huntingtin misfolding and/or truncation and do not always parallel aggregate formation.

The form of huntingtin that initiates the aggregation process (or is pathological in itself) is not fully identified. As indicated earlier, truncated huntingtin is more toxic than the full-length protein, suggesting that protein cleavage is important. Critical proteases that have come to light include several caspases and calpains. Indeed, recent data show that preventing cleavage of mutant huntingtin at the caspase 6 cleavage site (as opposed to the caspase 3 site) protects transgenic mice from neurodegeneration and motor impairments.

Whatever the pathophysiological mechanisms leading to neurodegeneration in HD, a main question that remains is the relative selective vulnerability of certain cell types despite the widespread expression of huntingtin. Recent studies in mice with restricted expression of the mutated protein suggest that the

process may be initiated not in the striatum but in cortical interneurons, which in turn affect corticostriatal projections and ultimately result in the death of striatal neurons. A role for glutamatergic cortical inputs in HD was suspected since the experiments with quinolinic acid mentioned earlier. For example, cellular pathology caused by mutant huntingtin in presynaptic corticostriatal terminals might result in an increased release of glutamate. Alternatively, impaired clearance of glutamate from the synaptic cleft might increase glutamatergic neurotransmission. In both cases, striatal excitotoxicity could occur. Intracerebral microdialysis has shown that depolarizing concentrations of potassium chloride increase the extracellular concentrations of glutamate substantially more in HD mutant than in control mice, and, in particular, the glial glutamate transporter is downregulated before any evidence of neurodegeneration. This indicates that there is an impairment in glutamate transport and glutamate–glutamine cycling, and suggests that a defect in astrocytic glutamate uptake could contribute to the phenotype.

One of the first indications of electrophysiological changes in the corticostriatal pathway was the observation that the stimulus intensity necessary to evoke an excitatory postsynaptic potential (EPSP) in medium-sized spiny neurons was significantly increased in symptomatic HD transgenic mice. Another observation was that synaptic responses in transgenic mice displayed slower rise and incomplete decay of the EPSP. This phenomenon was hypothesized to indicate a larger contribution of a slower kinetic current, such as the one mediated by activation of *N*-methyl-D-aspartate (NMDA) receptors. This idea is supported by the occurrence of enhanced NMDA receptor-mediated responses. In addition to alterations in evoked synaptic responses, both transient and progressive changes in spontaneous synaptic currents occur. Spontaneous excitatory postsynaptic currents show a progressive reduction in frequency that becomes more evident as the HD phenotype becomes more severe. This effect was interpreted as a progressive disconnection between the striatum and its cortical inputs. There is also a transient expression of complex, large amplitude synaptic currents that coincide temporally with the onset of the behavioral phenotype. These large currents probably reflect dysregulation of glutamate release and/or an increase in cortical synchronization.

Changes in the corticostriatal pathway and glutamate receptors alone are probably insufficient to produce neurodegeneration. However, numerous membrane properties of striatal medium-sized spiny neurons are also altered. One of the earliest and most consistent alterations in the basic membrane properties is an increase in input resistance. This increase probably reflects loss of conductive membrane channels due to morphological changes such as reduced membrane area, possibly as a consequence of the loss of spines. Consistent with this observation, cell capacitance is significantly reduced in symptomatic animals. The increase in membrane input resistance could also be due to alterations in the number and/or properties of K^+ channels. As a consequence, many medium-sized spiny neurons have a depolarized resting membrane potential and are less able to repolarize. These alterations are particularly relevant because membrane depolarization can remove the Mg^{2+} block of the NMDA receptor, causing neurons to become more depolarized when glutamatergic inputs are activated. These neurons will stay in a depolarized state for longer periods of time, thus predisposing them to excitotoxicity. Other voltage-gated conductances may also be affected as alterations in firing patterns occur. For example, there is a reduction in high voltage-activated Ca^{2+} conductances in symptomatic HD mice but an increase in these currents in presymptomatic mice. Taken together, the complex set of alterations in voltage-gated conductances of medium-sized spiny neurons will produce dysfunctional cells that have altered responses to inputs. Early increases in Ca^{2+} conductances will predispose cells to become more easily depolarized while subsequent decreases might be protective. Early increases in input resistance will decrease electrotonic decay of conductances allowing peripheral inputs to depolarize over longer distances while specific changes in K^+ channel function can also predispose neurons to remain depolarized and more excitable. In particular, decreased inward rectification could amplify excitatory inputs. These alterations emphasize that striatal medium-sized spiny neurons are dysfunctional and may be primed to be affected by abnormal cortical inputs. Thus, selective vulnerability in HD may result from a combination of changes in intrinsic properties of neurons and extrinsic inputs.

Conclusion

Thirteen years after the identification of the mutation that causes HD, much has been learned about the disease but the exact mechanisms leading to neurodegeneration remain unclear. Although cell loss remains a hallmark of advanced HD, extensive evidence that neuronal dysfunction precedes cell death has shifted emphasis from preventing neurodegeneration to maintaining neuronal health. Even apoptotic mechanisms may contribute to pathology by cleaving the mutant protein into more toxic forms rather than

leading directly to cell death. Current data point to a number of avenues for therapeutic development that are being actively pursued. Some successes have been obtained in animal models but the challenge remains to develop compounds that are safe enough for long-term use and target the most critical mechanisms rather than peripheral or secondary effects of the mutation.

See also: Animal Models of Huntington's Disease; Apoptosis in Neurodegenerative Disease; Axonal Transport and Huntington's Disease; Axonal Transport and Neurodegenerative Diseases; Cell Replacement Therapy for Huntington's Disease; Huntington's Disease; Oxidative Damage in Neurodegeneration and Injury; Transgenic Models of Neurodegenerative Disease; Triplicate Repeats: Huntington's disease.

Further Reading

Arrasate M, Mitra S, Schweitzer ES, Segal MR, and Finkbeiner S (2004) Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature* 431(7010): 805–810.

Brouillet E, Conde F, Beal MF, and Hantraye P (1999) Replicating Huntington's disease phenotype in experimental animals. *Progress in Neurobiology* 59(5): 427–468.

Cattaneo E, Zuccato C, and Tartari M (2005) Normal huntingtin function: An alternative approach to Huntington's disease. *Nature Reviews Neuroscience* 6(12): 919–930.

Cha JH (2000) Transcriptional dysregulation in Huntington's disease. *Trends in Neurosciences* 23(9): 387–392.

DiFiglia M (1990) Excitotoxic injury of the neostriatum: A model for Huntington's disease. *Trends in Neurosciences* 13(7): 286–289.

Harper PS (1996) *Huntington's Disease*. London: Saunders.

Hickey MA and Chesselet MF (2003) Apoptosis in Huntington's disease. *Progress in Neuropsychopharmacology and Biological Psychiatry* 27(2): 255–265.

Hickey MA and Chesselet MF (2003) The use of transgenic and knock-in mice to study Huntington's disease. *Cytogenetics and Genome Research* 100(1–4): 276–286.

Klapstein GJ, Fisher RS, Zanjani H, et al. (2001) Electrophysiological and morphological changes in striatal spiny neurons in R6/2 Huntington's disease transgenic mice. *Journal of Neurophysiology* 86(6): 2667–2677.

Levine MS, Cepeda C, Hickey MA, Fleming SM, and Chesselet MF (2004) Genetic mouse models of Huntington's and Parkinson's diseases: Illuminating but imperfect. *Trends in Neurosciences* 27(11): 691–697.

Mangiarini L, Sathasivam K, Seller M, et al. (1996) Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* 87 (3): 493–506.

Menalled LB (2005) Knock-in mouse models of Huntington's disease. *NeuroRx* 2(3): 465–470.

Menalled LB and Chesselet MF (2002) Mouse models of Huntington's disease. *Trends in Pharmacological Sciences* 23(1): 32–39.

Saudou F, Finkbeiner S, Devys D, and Greenberg ME (1998) Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. *Cell* 95(1): 55–66.

Vonsattel JP and DiFiglia M (1998) Huntington disease. *Journal of Neuropathology and Experimental Neurology* 57(5): 369–384.

Hyperacuity

G Westheimer, University of California, Berkeley, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Acuity

Acuity as a term in a sensory system is traditionally used to describe the limit of the capacity to resolve two stimuli as separate. In vision it has long been of interest in connection with distinguishing between single and binary stars. The notion has full underpinning from studies in physics, anatomy, and physiology.

Microscopy of the retina reveals that individual receptors are packed in the fovea in a tight hexagonal mosaic. Each acts as a discrete unit, indivisible, with a unique position label within its matrix, called local sign. Although they can be singled out nowadays in ophthalmoscopic view, under ordinary visual conditions receptors are not stimulated just one at a time because the eye's optics cannot generate an image sufficiently compact. Even an ideal optical system cannot recreate a point or infinitely narrow slit; these become diffraction patterns, whose diameter is at a minimum of the order of λ/a in angular measure (radians), where λ is the wavelength of light and a the aperture of the pupil, both in the same units. When inevitable aberrations are factored in, the narrowest possible light spread on the retina of a light point or star covers at least a dozen receptors (see Figure 1).

When two such objects are placed close together in object space (unless both arise from the same laser source), light in their images will intermingle indistinguishably. For them to be resolved, their joint light distribution needs to have a detectable dip that enables the segregation into two components (Figure 1). In practice, this is the case when the objects are separated by about half the width of the distribution. This distance is called the Rayleigh limit after the nineteenth-century British physicist who first used this approach to analyze the fine structure of lines in spectroscopes. Moreover, because of compartmentalization of light processing in receptors, the three salient features of the distribution (viz., the two peaks and the intervening trough) have to be captured each by a separate receptor. Finally, the light intensity incident on these receptors has to be such that their output to the next neural stage is distinguishably different. For large fields, signal summation yields signal-to-noise ratios good enough to enable intensity differences of 1% or less to be detected, but when it comes to adjoining individual receptors it requires a difference of approximately 20%, and this is just about the peak/trough ratio of an optical image at the Rayleigh limit.

In the human eye, there is a satisfactory match between this Rayleigh limit, the spacing of the retinal receptor elements, and the experimentally determined resolution limit of good human observers in sharp focus. This, then, is the widely accepted teaching of visual acuity: three separate lines of enquiry – optics, anatomy and psychophysics – converge in a single limit, which is about 1 arcmin of visual angle (2 mm at a viewing distance of 6 m or 20 ft). Optometric acuity charts using letters with a 2 mm stroke width to be seen at 20 ft embody this standard (20/20).

The anatomical projection from the retina preserves and conveys to the brain the responses from individual foveal cones separately, even in the presence of abundant neural interconnectivity in the retina and later stages of the visual stream. The large number of neurons in the primary visual cortex, in relation to the number of incoming optic nerve fibers, would seem to ensure that no information loss is incurred in this juncture, although difficulties of recording have so far prevented the demonstration of a cortical spatial grain matching that of the retina.

Better Than Acuity

The solidity of the teaching of visual acuity, which has served so well as a clinical tool and as a showcase example of how behavior ultimately must obey the limits set by physics and anatomy, deflected attention from findings, just as valid and venerable, that appear to be at variance on their face. Soon after Weber enunciated his rule about the constancy of the ratio between the detectable difference and the intensity of a sensory stimulus, $\Delta I/I = \text{constant}$, experiments were performed to document it for an observer's judgment of the distance between two points in visual space. In general, no disagreement was found, but a careful analysis of the situation for very short separations revealed that the detectable difference in separation has values lower than the resolution limit. This, together with the observation that a movement was detectable when an object was shifted through a distance smaller than the separation needed to tell the endpoints as separate when they were shown simultaneously, at first was regarded as throwing doubt on the idea of a discrete receptor mosaic in the retina. But by the end of the nineteenth century, retinal histology and careful measurements of the precision with which two abutting lines could be aligned had sharpened the issue. This vernier procedure, which

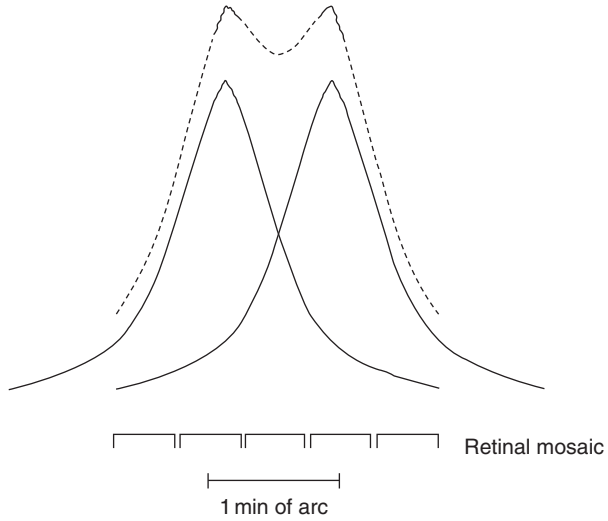

Figure 1 Schematic diagram illustrating the factors operating when two stars can just be detected as separate in the human fovea. Each of the two sources gives rise to a light distribution that spreads over several foveal cones (solid curves). Optical quality of the eye, grain of the retinal mosaic, and the sensitivity to light differences in the visual system all have converged to a limit at a critical separation, where the two peaks and the intervening trough in the summed distribution (dashed curve) each covers a single receptor and the peak/trough light ratio is just detectable. This is the basis of the eye's resolving power to which the term acuity is applied.

had had practical utility for centuries, was measured by Wülfing in 1892 to have a thresholds in the human fovea to be of the order of 10 arcsec, at least one-third of the diameter of a foveal cone and of the half width of the eye's diffraction image. This observation prompted Hering to formulate the concept that, whereas each cone had its own unique and indivisible local sign (as needed for visual acuity), a more precise spatial value could be arrived at by averaging the local signs of all the cones along the image of a line.

Hering's notion appeared to constitute a compelling bridge between two important research directions: on the one side, strong psychophysical findings based in the traditional views of retinal image width and in cone local sign and spacing, and on the other, averaging, an acceptable kind of neural processing. A challenge by Ludvigh in 1953, who claimed that alignment can be just as good when points were substituted for lines, was laid aside because his experiment was performed on only one observer with methods that were questioned at the time. Moreover, his 2-arcsec thresholds are almost impossibly low, even by present standards.

Because of the very small distances and laborious experimental procedures, new investigations had to await the advent of computer-driven visual displays and research apparatus. With these, it quickly became apparent not only that the earlier suggestions of visual spatial thresholds below the resolution limits were well founded but also that Hering's proposition had to be dropped. Alignment thresholds with point targets can be just as good and sometimes even better than with lines, and many other spatial visual thresholds have values smaller than the elements of the retinal mosaic and are of a kind for which averaging of signals along image contours cannot be invoked as a plausible explanation.

Clearly, room had to be made for a class of sensory discriminations in which the performance limit is not set by the element size of the receptor grain. The task requires more sophisticated neural processing than reliance on the detection of a differentiable signal in the middle one of three adjoining individual receptors. The word visual acuity had been preempted for over 100 years by its identification with a concept centered on the eye's resolving power, diffraction limit, dimensions of the elements of the receptor mosaic, and letter eye charts based on form discrimination and two-line resolution. Thus a new word, hyperacuity, was coined, implying that ordinary acuity is being transcended. Superresolution (in German, *Überauflösungsvermögen*) had already been used in radar in an entirely different connection, and, in any case, it still has the connotation of detecting separability.

The nomenclature helps to emphasize that two fundamentally different tasks are involved. In one task, ordinary acuity such as tested in the Snellen Letter Chart, one is dealing with the ability to detect feature components as separate, requiring differentiable retinal image components such as separable peaks and troughs. In the other, the task is to identify the location of a feature, the exact nature of whose retinal image is not at issue. Acuity and hyperacuity are now used as the terms applied to the limits, respectively, of resolution and localization (**Figure 2**). Although the resolution limit must perforce conform, at a minimum, with the diffraction equations, no laws of physics are disobeyed when the location of an object is specified to a much smaller value. The detection of a single photon from a point object will, to be sure, have location uncertainty given by the diffraction pattern (an expression of the uncertainty principle). But if stability is retained between source and receiver, the accumulation of enough photon events will enable the identification of the source location with arbitrary precision.

If the optical stimulation and the subsequent processing were entirely punctate (i.e., if all information were directed to, accepted by, and subsequently handled entirely in discrete units without overlap), the tiling of the visual space by these units would indeed constitute the inescapable lower bound of spatial differentiation. The information flow from each

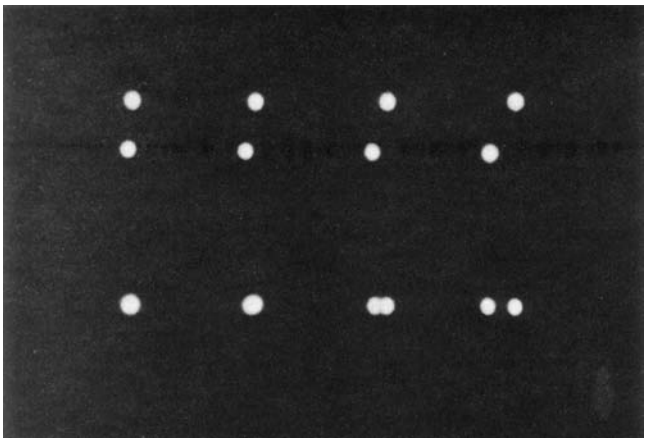

Figure 2 Acuity (bottom row) and hyperacuity (top rows). The relative location of two light sources (top rows) can be detected when the shift is much smaller than the one required to resolve them as separate (bottom row). Adapted from Whitaker D and Buckingham T (1987) Theory and evidence for a clinical hyperacuity test. *Ophthalmic and Physiological Optics* 7: 431–435.

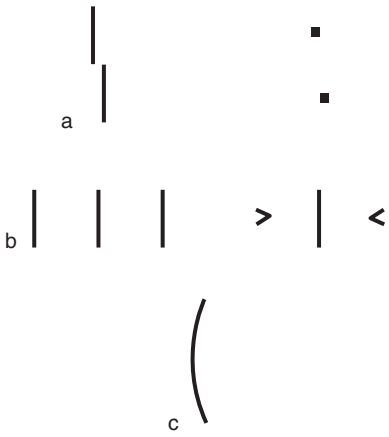

Figure 3 Discrimination thresholds in the hyperacuity range: (a) detection of the misalignment in the direction orthogonal to the line joining two features (typical example: vernier acuity); (b) discrimination of the separation of two or more features (typical example: bisection); (c) form discrimination (typical example: detection of curvature). Thresholds are generally robust to the nature of the features but critically depend on distance factors.

receptor would be independent of that from the others. But this is far from the case. As is apparent in Figure 1, the spread of light from a point source covers to varying degrees many receptors. There is considerable convergence and overlap of signals within the retina, so that the impulses carried by the optic nerve from individual ganglion cells, although their number in the fovea exceeds that of receptors, is a composite of the light from a wider spatial region and the output of many receptors. Subtle processing can then tease out location parameters enabling exceedingly fine discriminations (Figure 3).

The neural mechanism by which this is achieved still needs detailed elucidation. It has been convincingly demonstrated that responses of single neurons in the cat and primate retina can differentiate between target locations just a hyperacuity distance apart. This provides a necessary, but not sufficient, neural substrate because such response differences can also arise from light intensity differences. For single small sources, it seems clear that there is an apparatus which arrives at the position of the centroid of the retinal light distribution as it emerges after passage through the individual receptors and their retinal connections, probably by way of a population vector. The equivalent concept in statistics is the calculation of the mean of a histogram. Edge location, which is just as good as points', is thought to involve a mechanism of fixing the position of the inflexion points of the light distribution, which mathematically is defined as zero-crossing of their second derivative.

Even though all the physical preconditions have been met and there is the physiological apparatus for it, localizing a single target in isolation is difficult for an observer because eye position is neither maintained with the requisite stability nor can it be read out with sufficient precision. However, if the task is relative localization, then long-term stability is not required. This view of the situation is accentuated by the fact that for good visual alignment and separation acuity, exposure of the whole pattern needs to be simultaneous; onset asynchronies of as little as 20–50 ms of configuration components leads to a reduction of performance (Figure 4).

A critical experiment was the demonstration that the relative localization of two visual targets in the hyperacuity range is robust to movements imposed on them jointly, up to a velocity of several degrees per second. In practice, this means that the distance between two lines or edges, or the degree of their alignment, can be judged with a precision of a small fraction of the diameter of the elements of the retinal

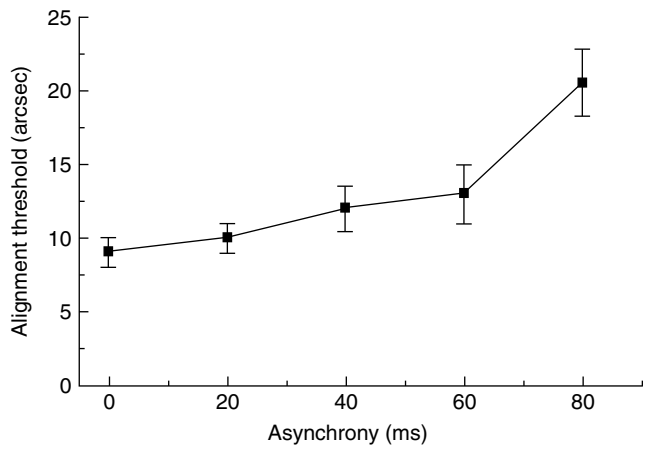

Figure 4 Performance is impaired when there is asynchrony of the presentation of the two components of a vernier alignment task, an indication of the processing mechanism for hyperacuity.

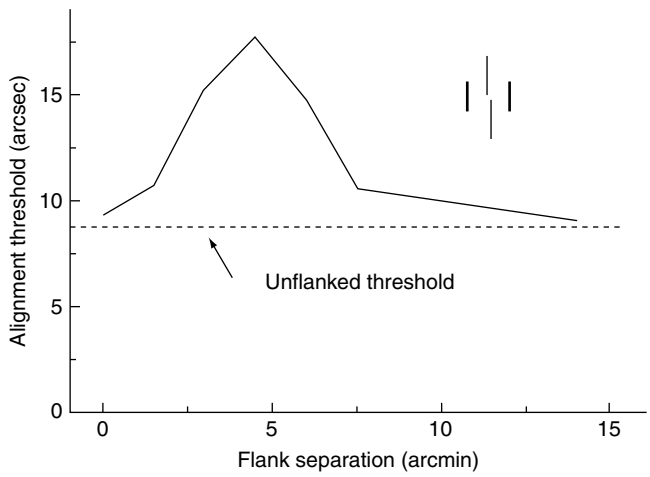

Figure 5 Laterally placed flanks interfere with the ability to detect the misalignment of a vernier configuration. The effect is maximal when the flanks are a few arcminutes distant and are clearly seen as separate, and it persists when the target and flanks are presented to different eyes (dichoptic presentation). This and other contextual effects implicate cortical processing.

mosaic when the whole configuration sweeps across the retina at the rate of 100 receptors in 200 ms. A sophisticated cortical apparatus is clearly at play.

Some information is available to narrow down the parameters of its operation. For one thing, it can operate on target components of a variety of shapes and in directions both along the line joining target components and orthogonal to it (Figure 3). Although it is better for the vertical and horizontal meridians, performance is still commendable in the obliques. This suggests that it is not simply a matter of using the known receptive field properties of neurons in the primary visual cortex.

As distinct from ordinary visual acuity, which needs exposures of more than 100 ms duration for optimal values, hyperacuity can be quite good at very short durations. Further information about the processing apparatus is provided by the fact that hyperacuity thresholds are very susceptible to interference by neighboring stimuli, a phenomenon called masking (Figure 5). It seems that, to arrive at local signs of feature elements, or the difference between their local signs, signals are used that cover a larger spatial territory than the area of the individual targets, something that might be expected if the very precise location value arises from processing of a population of neurons.

Again, as might be inferred from the more central processing that is involved, hyperacuity thresholds can be improved by practice, a phenomenon that goes under the rubric of perceptual learning. Two regimes can be distinguished here. Because the measurement requires some familiarity with the targets and the registration of responses, threshold values improve, often by a factor of two, in the first few sessions or the first few hundred responses in

psychophysical procedures. This fast perceptual learning should be distinguished from true perceptual learning, which takes place over several thousand responses extending over several days and represents long-lasting modification of circuitry in the cortex. This phenomenon can also serve to illuminate the processing apparatus by estimating the transfer of training to neighboring location and to related tasks.

Lest it be thought that even within the category of spatial visual abilities all hyperacuities are similar, it should be pointed out that stereoacuity, the facility to determine location differences in the third dimension, does not share some of its important characteristics with, say, vernier acuity, to which it is closely related. Its distance and temporal summation properties are different, although its amazingly low thresholds, depending on the detection of binocular disparity, are tiny compared with the size of retinal receptors even in the fovea.

General Considerations

Hyperacuity refers to sensory abilities in which the whole organism's performance transcends the grain imposed by the anatomical structure and physiological organization of the sensory apparatus. It has been most closely studied in spatial vision, in which the size of the retinal receptors and the indivisibility of their local sign had been used previously to lay the foundation for a highly successful interpretation of visual acuity, a major functional attribute of clinical and scientific import.

The understanding of neural processing derived from the distinction between visual acuity and hyperacuity are of more general value in the study of neural circuitry. The perception of the orientation of contours in visual space – for example, whether a picture frame is hung accurately – is usually thought to arise at the level of orientation-selective neurons in the primary visual cortex, associated with their discoverers Hubel and Wiesel, and concentrated in cortical columns of shared orientation. For a while, the question about the minimum orientation difference between adjoining cortical columns was asked, with the implication that there was need to seek a tiling of visual space into a set of discrete orientations. The analogy with visual acuity becomes helpful here. Most observer can quickly learn to distinguish between orientations 0.5° apart. Yet the orientation tuning of the cortical cells (i.e., the range of contour orientations to which they will respond) is many degrees, even many tens of degrees, wide. The domain of contour orientation is covered by neural elements with widely overlapping acceptance lobes, and it is patent that the sharply defined orientation of a seen contour must emerge from or, if this concept is preferred, be computed by, a circuit using the output of a range of such elements. Such a notion can then progress effortlessly to the study and understanding of masking phenomena, due to activation by adjoining contours of inhibitory connections, and of contour grouping, due to potentiating influences from collinear or even top-down connections.

For the chromatic rather than spatial attribute of visual perception, the physical basis is the wavelength acceptance of the three visual pigments which overlap widely. The highly refined discrimination of very small chromaticity steps is, in a way, a hyperacuity. We do not see just three basic hues, one for each of the pigment types, but many thousand of shades, responding in a nuanced manner to differential patterns of output from the three classes of cones.

Auditory, Tactile and Other Senses

These considerations and observations in the visual sense have found resonance in other sense modalities. In the tactile sense, Loomis has shown that cutaneous patterns can be distinguished in a hyperacuity range, as in reading Braille, and the underlying peripheral neural mechanism has been explored. Similarly, it has been demonstrated that many auditory responses, when analyzed psychophysically and in the firing pattern of cortical neurons, fit the classification of hyperacuity.

Finally, the concept of discrimination of only a fraction of the size of the elementary unit of a particular dimension has been applied also in the temporal domain, to the discharge pattern of neurons. Refractory periods set an upper limit of a few hundred hertz to the firing frequency of neurons, yet meaningful temporal information has been assigned to epochs of only a very few milliseconds and, in the case of echolocation of bats and jamming avoidance of electric fish, even much less.

See also: Neural Patterning: Eye Fields; Perception and Eye Movements; Stereoscopic Vision; Vision for Action and Perception.

Further Reading

Altes RA (1989) Ubiquity of hyperacuity. *Journal of the Acoustical Society of America* 85: 943–952.

Crick FHC, Marr DC, and Poggio T (1981) An information-theoretical approach to understanding the visual cortex. In: Schmitt FO, Worden FG, Adelman G, and Dennis SG (eds.) *The Organization of the Cerebral Cortex*, pp. 505–533. Cambridge, MA: MIT Press.

Crist RE, Kapadia MK, Westheimer G, and Gilbert CD (1997) Perceptual learning of spatial localization: Specificity for

orientation, position, and context. *Journal of Neurophysiology* 78: 2889–2894.

Heiligenberg W (1991) Sensory control of behavior in electric fish. *Current Opinions in Neurobiology* 1: 663–667.

Kashimori Y, Inoue S, and Kambara T (2001) A neural mechanism of hyeraccurate detection of phase advance and delay in the jamming avoidance response of weakly electric fish. *Biological Cybernetics* 85: 117–131.

Lee BB, Wehrhahn C, Westheimer G, and Kremers J (1993) Macaque ganglion cell responses to stimuli that elicit hyperacuity in man: Detection of small displacements. *Journal of Neuroscience* 13: 1001–1009.

Loomis JM (1979) An investigation of tactile hyperacuity. *Sensory Processes* 3: 289–302.

McAnally KI and Calford MB (1990) A psychophysical study of spectral hyperacuity. *Hearing Research* 44: 93–96.

Simmons JA, Ferragamo MJ, and Sanderson MI (2003) Echo delay versus spectral cues for temporal hyperacuity in the big brown bat *Eptisicus fuscus*. *Journal of Comparative Physiology A* 189: 693–702.

Shapley RM and Victor J (1986) Hyperacuity in cat retinal gnaglion cells. *Science* 231: 999–1002.

Westheimer G (1975) Editorial: Visual acuity and hyperacuity. *Investigative Ophthalmology* 14: 570–572.

Westheimer G (1981) Visual hyperacuity. *Progress in Sensory Physiology* 1: 1–30.

Westheimer G (1994) The Ferrier lecture, 1992: Seeing depth with two eyes: Stereopsis. *Proceedings of the Royal Society (London) B Biological Sciences* 257: 205–214.

Wheat HE, Goodwin AW, and Browning AS (1995) Tactile resolution: Peripheral neural mechanisms underlying the human capacity to determine positions of objects contacting the finger pads. *Journal of Neuroscience* 15: 5582–5595.

Whitaker D and Buckingham T (1987) Theory and evidence for a clinical hyperacuity test. *Ophthalmic and Physiological Optics* 7: 431–435.

Hypocretin/Orexin and MCH and Receptors

A Adamantidis and L de Lecea, Stanford University School of Medicine, Palo Alto, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Discovery

The hypocretins (Hcrts) were discovered as part of a subtractive hybridization study to identify messenger RNAs restricted to discrete nuclei within the rat hypothalamus. Among these was a species, named prepro-hypocretin, whose expression was restricted to a few thousand neurons that were bilaterally distributed within the dorsolateral hypothalamus. The sequence of preprohypocretin suggested that it encoded two putative neuropeptides, named hypocretin-1 (Hcrt1) and Hcrt2, with significant similarity to the incretin peptide family.

A large collaborative study to identify endogenous ligands for orphan G-protein-coupled receptors (GPCRs) discovered the peptides independently. This group referred to the peptides as orexins because they stimulated acute food intake when administered to rats during the daytime. In this article, we refer to the peptides by their first-used name, hypocretins (from the sequence similarities with various members of the incretin family and its hypothalamic localization), but both names have been used extensively in the large literature that has grown up around the peptides.

Melanin-concentrating hormone (MCH) was originally discovered in salmon pituitaries, in which it acts as a circulating hormone and lightens skin in response to the background color change or during stress reaction. MCH acts antagonistically to α- and β-melanocyte-stimulating hormone (MSH) and induces pigment aggregation of melanin granules in skin melanophores *in vitro* and *in vivo*. In fish, MCH is a cyclic neuropeptide of 17 amino acids which is generated along with other peptide (neuropeptide EI (NEI)) by cleavage of a preprohormone. Subsequent to its identification in fish, MCH immunoreactive-like labeling in rat brains led to the identification of this peptide, its gene, and its tissue localization in this species. In vertebrates, MCH-producing neurons are restricted to the lateral hypothalamus (LH) and zona incerta (ZI).

Structure

Both Hcrt and MCH peptides are derivatives of a longer precursor of 130 and 165 residues, respectively, with an apparent signal sequence.

The C-terminal 19 residues of these two putative peptides – Hcrt2 (28 residues; RPGPPGLQG RLQRL-LQANGNHAAGILTM-amide) and Hcrt1 (33 residues; EPLPDCCRQKTCSCRLYELLHGAGNHAAGILTL-amide) – share 13 amino acid identities, suggesting that the peptides have related structures and functions. This region of Hcrt2 contains a seven-amino-acid match with secretin. Hcrt1 contains two intrachain disulfide bonds. Human Hcrt1 is identical to the rodent peptide, whereas human Hcrt2 differs from rodent Hcrt2 at two residues. The nonamidated forms of the peptides are not electrophysiologically active.

The MCH intrachain disulfide bond leads to a cyclic peptide (19 residues; DFDMLRCMLGRVY RPCWQV). Strong sequence similarity has been found with the somatostatin peptide family.

Distribution

Hcrt and MCH neurons are morphologically very similar, multipolar or fusiform in shape, with two to five primary dendrites that are either smooth or sparsely invested with dendritic spines. In rats (3000 neurons) and humans (50 000–80 000 neurons), cells highly positive for Hcrt mRNA and immunoreactivity are located between the rat fornix and the mammillothalamic tracts (Figure 1). In addition to their close localization to the Hcrt neurons, MCH-producing neurons are also found in the ZI. Transcripts and immunostaining for MCH in a small set of neurons of the olfactory tubercules and pontine tegmentum have also been reported.

The projections of Hcrt and MCH cells are widely distributed within the central nervous system (CNS) in accordance with their respective receptors' transcripts distribution. Sparse anatomical and physiological evidence supports the existence of subpopulations of Hcrt or MCH neurons. Based on c-fos profile expression, Harris and colleagues proposed two subpopulations of Hcrt neurons related to arousal and consummatory behaviors. Subpopulations of MCH neurons have been identified with the presence of the cocaine- and amphetamine-regulated transcript (CART) peptide immunoreactivity and neuroanatomical projections.

There are several reports of Hcrt, MCH, and receptors expression in the periphery, including in the enteric nervous system, pancreas, kidney, stomach, and ileum. Hcrt expression has also been detected in the human retina, anterior pituitary, and adrenal gland.

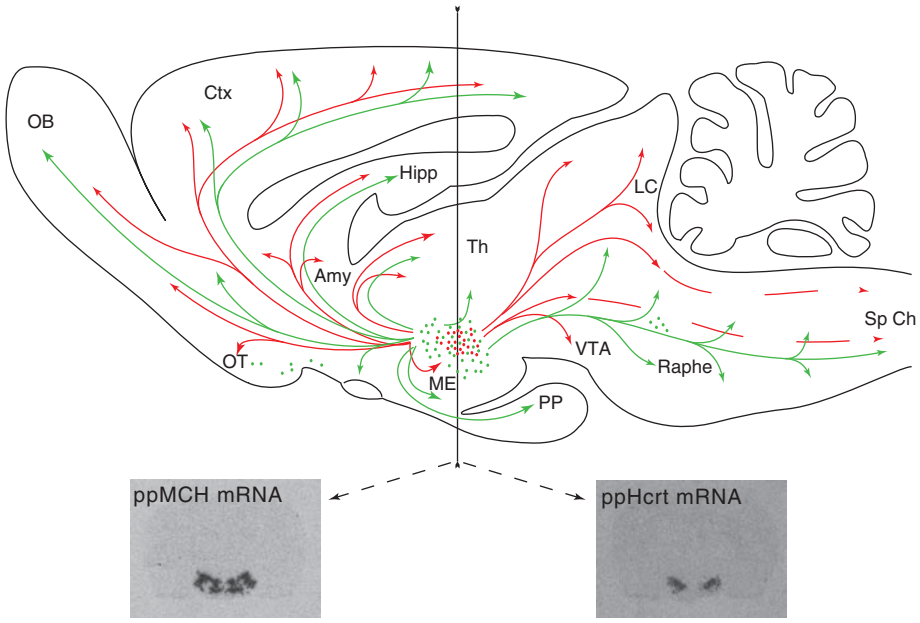

Figure 1 Neuroanatomy of the hypocretin (Hcrt) and melanin-concentrating hormone (MCH) systems. Schematic drawing of a saggital section through the rat brain showing the organization of the MCH and Hcrt systems. Dots indicate the relative location and abundance of MCH (green)- and Hcrt (red)-expressing cell bodies. Arrows point out some of the more prominent terminal fields. *In situ* hybridization of both precursor mRNA (ppMCH and ppHcrt) in mice is shown below in a coronal brain section made at the level of the vertical black line. Amy, amygdala; Ctx, cortes; Hipp, hippocampus; LC, locus coeruleus; ME, median eminence; OB, olfactory bulb; OT, olfactory tubercule; PP, posterior; Sp, Ch, Spinal chord; Th, thalamus; VTA, ventro-tegmental area.

Receptors and Signaling Cascades

The initial orphan GPCR, Hcrt receptor 1 (Hcrtr1) (also referred to as OX1R), binds Hcrt1 with high affinity, but binds Hcrt2 with 100- to 1000-fold lower affinity. A related GPCR, Hcrtr2 (OX2R), sharing 64% identity with Hcrtr1, was identified by searching database entries with the Hcrtr1 sequence and has a high affinity for both Hcrt2 and Hcrt1. These two receptors are highly conserved (95%) across species. Radioligand-binding studies and Ca^{2+} flux measurements have shown Hcrt1 to have an equal affinity for Hcrtr1 and Hcrtr2, whereas Hcrt2 has ~10-fold greater affinity for Hcrtr2 than Hcrtr1.

The mRNAs that encode the two hypocretin receptors and the receptor proteins themselves, detected by immunohistochemistry, are both enriched in the brain and moderately abundant in the hypothalamus. But they have different distributions within the brain.

The distribution of Hcrt receptors is largely consistent with Hcrt axon innervation patterns. The composite distribution of the two Hcrt receptors strongly resembles the distribution of MCH receptor 1 (MCHR1). In the locus coeruleus (LC), amygdala, and other brain stem noradrenergic groups, MCHR1 mRNA distribution is similar to that of the Hcrtr1. In regions such as the septum, hypothalamus, and much of the brain stem, the distribution of MCHR1 mRNA resembles that of the Hcrtr2.

Interestingly, *in silico* analysis of the human genome sequences revealed a second MCH receptor (MCHR2), which shares 38% identity with MCHR1. However, the MCHR2 is only functional or encoded in the dog, ferret, monkey, and human, whereas it is absent or encoding a nonfunctional truncated protein (pseudogene) in the mouse, hamster, rat, guinea pig, and rabbit. In addition to a different gene structure (five exons instead of the two in MCHR1) MCHR2 expression is restricted to the cortex, hippocampus, amygdala, and hypothalamus in the CNS. The Hcrt and MCH receptors are also widely expressed in the periphery, especially in endocrine tissues, including the pituitary, adrenal gland, testis, gastrointestinal tract, pancreas, and pineal gland.

Both Hcrt1 and Hcrt2 act through a family of guanosine triphosphate (GTP)-binding proteins (G_q) that activate protein kinase C (PKC) and the mobilization of intracellular Ca^{2+}. G_q-activated signaling cascades result in phosphorylation of Ca^{2+} channels, which can increase Ca^{2+} conductance and neuronal excitability.

Probably depending on the tissue considered, the MCHR1 GPCR could be coupled to $G_{i/o}$, G_α, G_s, or G_q via different intracellular cascade including mitogen-activated protein kinase (MAPK) and extracellular signal-regulated kinase ERK(1/2). In contrast, the MCHR2 receptor has been reported to act only through G_q.

Activation of specific intracellular pathways by the MCH may rely on the modification of MCHR1 (or MCHR2) binding properties through interactions with cytosolic or membrane protein. Recently, using a yeast two-hybrid approach, the group of Bachner identified several proteins (MCHR1-interacting zinc-finger protein (MIZIP), periplakin, and neurochondrin), which interact with the C-terminus of the MCHR1 receptor. They may be involved in the regulation of MCHR1 signaling and may play a role in modulating MCH-mediated physiological functions *in vivo*.

Input/Output

Hcrt and MCH neurons in the LH are highly interconnected with a network of glutamatergic, GABAergic, dopaminergic, and cholinergic neurons (Figure 2).

Glutamate, excitatory amino acid transporter (EAAT)3, and the vesicular glutamate transporters VGLUT1 and VGLUT2 are expressed by Hcrt neurons; thus, Hcrt neurons are likely to be glutamatergic. Other proteins detected in Hcrt neurons include dynorphin; $GABA_A$ receptor epsilon subunit; serotonin receptor (5-HT-1A); μ-opioid receptor; pancreatic polypeptide Y4 receptor; adenosine A1 receptor; leptin receptor; signal transducer and activator of transcription (STAT-)3; and the neuronal pentraxin Narp, implicated in clustering of ionotropic glutamate receptors.

In the rat, CART has been show to be coexpressed mainly in the MCH neuron of the rostro-median ZI. Moreover, part of these CART-expressing MCH neurons also express the glutamate decarboxylase enzyme, confirming that part of the MCH neurons are GABAergic. Recently, substance P has been co-localized with few MCH neurons.

Part of the MCH neuronal population expresses several membrane proteins, including neurokinin receptor (NK)3, the receptor to the chemokine stromal cell-derived factor-1α (CXCR4), leptin receptor (Ob-R, undefined subtype), glutamate and GABA receptors, adrenoreceptor (α2), muscarinic and serotoninergic receptors, and HcrtR1 and/or -2. Interestingly, neuropeptide Y (NPY) has been shown to inhibit MCH neurons by pre- and postsynaptic mechanisms, whereas the melanocortin agonist MTII and melanocortin antagonist SHU9119 were without effect.

Biological Actions within the Brain

Experimental administration of MCH and Hcrt peptides to animals stimulates food intake, affects motivation and rewarding behavior and modifies arousal. It is noteworthy that in mammals MCH and Hcrt follow a diurnal pattern of gene expression and probably peptide release. *In vivo* recordings of Hcrt neurons revealed that these cells do not fire tonically but, rather, show bursts of activity at the transitions between sleep and wakefulness and during exploratory or consummatory behavior.

Feeding and Metabolism

Sakurai and colleagues found that intracerebroventricular (ICV) administration of either Hcrt1 or Hcrt2 increased short-term food consumption in rats. Furthermore, rats that had been deprived of food for 48 h showed increased concentrations of Hcrt mRNA and peptides in the hypothalamus. Feeding responses can be elicited by the local administration of Hcrt1 to the paraventricular nucleus, the dorsomedial nucleus, the LH, or the perifornical area. ICV administration of Hcrt2 also increases food intake in sheep and goldfish. Many observations leave little doubt that the Hcrt system influences and is influenced by primary nutritional homeostasis circuits. For instance, Hcrt neurons are sensitive to glucose, leptin, and triglyceride concentrations. Hcrt neurons are contacted by NPY-containing neurons in the arcuate nucleus, and much of the food intake increase elicited by Hcrt1 seems to be mediated by NPY.

Other findings suggest that the Hcrts are not critical players in feeding activities but, rather, play roles in increasing arousal and motivation levels so that feeding can take place. Continuous administration of Hcrt1 for 7 days in rats does not significantly alter daily food intake, body weight, blood glucose, total cholesterol, or free fatty acid levels, suggesting that many of Hcrt's effects may be limited to the short-term immediate stimulation of feeding behavior consequent to the increased wakefulness. If Hcrt has a direct role on food intake, we expect a lean phenotype in Hcrt-deficient mice, as has been described for MCH. Hcrt knockout mice show very modest, if significant, differences in food intake. Hcrt ataxin-3 mice, which are genetically depleted of Hcrt neurons and would be expected to be lean, show obesity and hypolocomotion, and this effect appears to be dependent on diet and genetic background. During fasting, Hcrt1 accumulation in the cerebrospinal fluid (CSF) does not exceed concentrations normal for the waking period. Also, Hcrt ataxin-3 mice, unlike wild-type animals, do not show an increase in locomotor activity after fasting. All these data suggest that some of the food-uptake effect may result from arousal rather than direct feeding pressure.

Among the putative roles of MCH neurons, their involvement in feeding behavior, energy homeostasis, and body weight are by far the best documented. Chronic intracerebral injection increased food intake

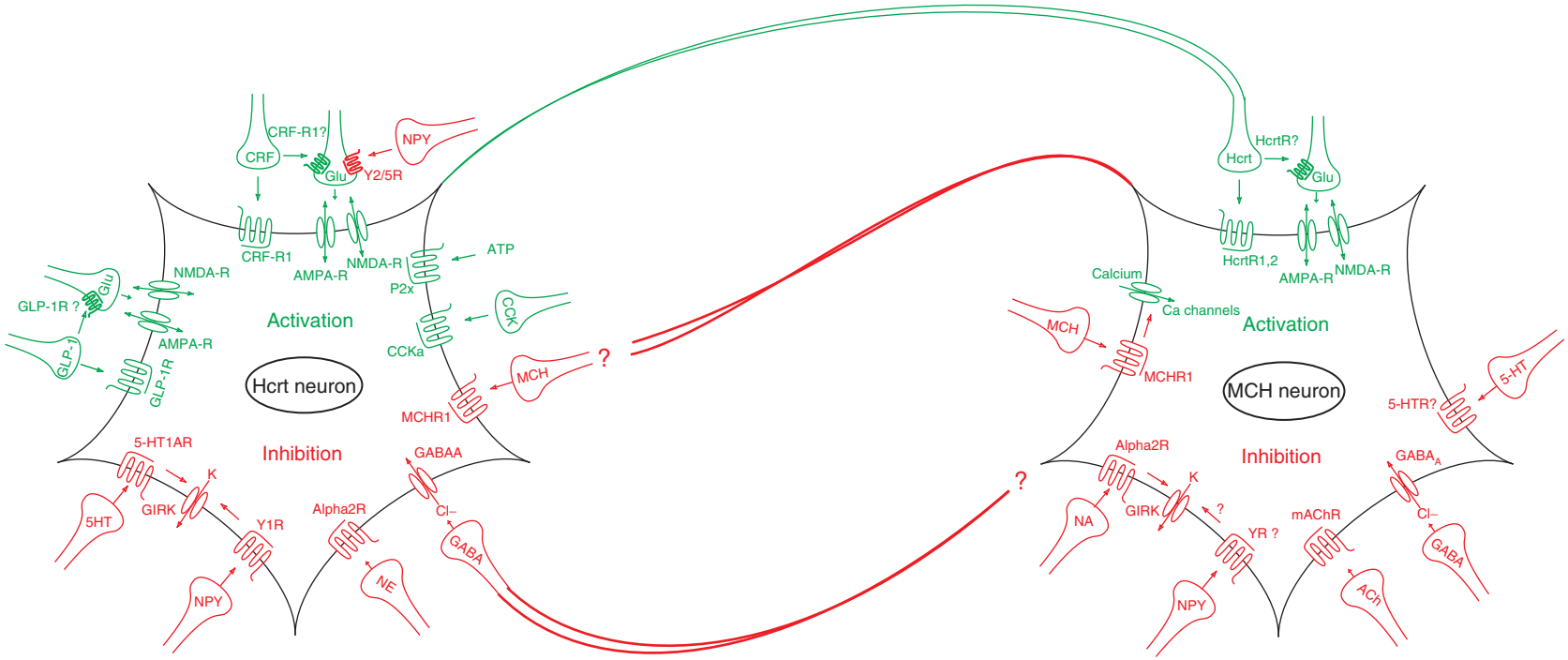

Figure 2 Inputs to hypocretin (Hcrt) and melanin-concentrating hormone (MCH) neurons. Schematic showing electrophysiologically demonstrated excitatory (green) and inhibitory (red) inputs on Hcrt and MCH neurons. Note that some transmitters have pre- and postsynaptic mechanisms on glutamatergic and GABAergic interneurons, and therefore they can be excitatory and inhibitory. Hcrt neurons are known to contact and excite MCH cells, and in the reciprocal connection it is likely that MCH project back onto Hcrt neurons. Hcrt1 excites Hcrt neurons via glutamatergic interneurons; direct changes in membrane potential have not been observed.

during the hours following injections without modifying the ingested food amount in 24 h or the body weight. Also, both ppMCH mRNA and MCH peptide increase after fasting, as does the MCHR1 mRNA. Mice overexpressing MCH develop mild obesity and hyperphagia and MCH knockout mice are lean but hyperphagic, probably because of their increased nocturnal locomotor activity and consequent increased metabolic rate. Accordingly, transgenic mice depleted of MCH neurons also develop late-onset leanness. Finally, leptin has been shown to depress MCH gene expression in a series of obese leptin-deficient animal models (ob/ob knockout mouse, fa/fa rat, and yellow agouti mice, in which animals develop overfeeding). At the cellular level, glucose dose-dependently enhances the electrical excitability of MCH neurons by inducing depolarization and increasing membrane resistance, but it has the opposite effect on the electrical activity of Hcrt cells. As a result of this participation in food intake behavior and energy homeostasis, a plethora of anorexigenic MCHR1 antagonists has been developed in recent years.

Motivation and Addiction

Intracranial self-stimulation (ICSS) is a well-characterized paradigm for quantifying brain reward. Infusion of Hcrt1 in the brain ventricles elevates ICSS thresholds, indicating that the peptide decreases brain reward function in a way that is similar to stress.

Hcrt neurons are highly responsive to morphine, and their activation is linked to preferences for cues associated with drug and food reward and also naltrexone-precipitated withdrawal. Moreover, Hcrt reinstates extinguished drug-seeking behavior, an effect blocked by an Hcrt1 antagonist. Hcrt1 infusion into the ventricle or the ventral tegmental area leads to a dose-dependent reinstatement of drug-seeking behavior, and it also increases ICSS threshold. Hcrt-induced reinstatement of cocaine seeking was prevented by the blockade of noradrenergic and corticotropin-releasing hormone systems. Accordingly, Hcrtr1 antagonist blocks footshock-induced reinstatement of previously extinguished cocaine-seeking behavior, leading to the conclusion that Hcrts are a major gate in driving stress-mediated drug-seeking behavior. Hcrt knockout mice exhibit dramatically attenuated morphine-withdrawal symptoms.

However, the response of these neurons is heterogeneous, suggesting that there might be different subpopulations of Hcrt cells. Expression of pepro-Hcrt increases after precipitated withdrawal. Hcrt neurons express the μ-opioid receptor; hence their response may be directly related to morphine and naltrexone. These observations might explain why animals self-administer heroin to the LH. Hcrt neurons have extensive projections to the mesolimbic dopamine and noradrenergic (LC) pathways, regions well studied for their roles in drug addiction. These neurons also project to and inhibit nucleus accumbens neurons, resulting in a negative regulation of the activity of brain reward circuitries.

In contrast, MCH neurons are not activated during acute and chronic morphine treatment or following naltrexone-induced withdrawal. Interestingly, MCH injection in the nucelus accumbens has been shown to induce food intake, whereas an MCHR1 antagonist has the opposite effect. Theses data support the idea that, like Hcrt peptide function, the feeding response relies at least partly on activation of brain reward mechanisms.

Furthermore, MCHR1 knockout mice fail to develop behavioral (locomotor) sensitization to cocaine and do not exhibit any sign of cocaine conditioning during a saline challenge. It is speculated that MCHR1 may contribute to the neurobiological mechanisms of conditioned cocaine-induced psychomotor effects, possibly to these underpinning sensitizations, and to a lesser extent to those subserving acute pharmacological cocaine action.

Arousal

Rodent models Continuous recording of the behavior of Hcrt gene-deleted mice (Hcrt knockout mice) revealed periods of ataxia, which were especially frequent during the dark period. Electroencephalograph (EEG) recordings showed that these episodes were not related to epilepsy, and that the mice suffered from cataplectic attacks and their EEGs showed episodes of direct transition from wakefulness to rapid-eye movement (REM) sleep, the hallmarks of narcolepsy. Similar observations were made in rats in which the Hcrt neurons of the LH were inactivated by saporin targeting, although in this model, cataplexy was not observed. Mice with an inactivated Hcrtr2 gene had a milder narcoleptic phenotype than the Hcrt knockouts; Hcrtr1 knockouts exhibited only a sleep-fragmentation phenotype, whereas double Hcrtr1/Hcrtr2 mutants recapitulated the full Hcrt knockout phenotype, suggesting that signaling through both receptors contributes to normal arousal, although the role of Hcrtr2 is greater than that of Hcrtr1.

Transgenic rats and mice depleted of Hcrt neurons have been generated by expressing a mutant form of ataxin 3 in Hcrt neurons. These animals show a narcolepsy-like phenotype as well as reduced locomotor activity on fasting, strongly suggesting a key role of Hcrt neurons in linking metabolic needs with arousal.

Human narcolepsy Nishino and colleagues studied Hcrt concentrations in the CSF of normal controls

and patients with narcolepsy by radioimmunoassay. In control CSF, Hcrt concentrations were highly clustered, suggesting that tight regulation of the substance is important. However, of nine patients with narcolepsy, only one had a Hcrt concentration within the normal range. One patient had a greatly elevated concentration, and seven patients had no detectable circulating Hcrt. In an expanded study, Hcrt was undetectable in 37 of 42 narcoleptics and in a few cases of Guillain–Barré syndrome. CSF Hcrt was in the normal range for most neurological diseases, but was low, although detectable, in some cases of CNS infections, brain trauma, and brain tumors.

Peyron, Thannikal, and their teams of collaborators found that, in the brains of narcolepsy patients, they could detect few or no Hcrt-producing neurons. Whether the Hcrt neurons are selectively depleted, as is most likely, or only no longer expressing Hcrt is not yet known, although one report showed some indications of gliosis. The codistributed MCH neurons were unaffected. Furthermore, a single patient with a non-human leukocyte antigen (HLA)-linked narcolepsy carries a mutation within the hypocretin gene itself. The mutation results in a dominant negative amino acid substitution in the secretion signal sequence that sequesters both the mutant and heterozygous wild-type Hcrt nonproductively to the smooth endoplasmic reticulum. These findings leave no doubt as to the central role of the Hcrt system in narcolepsy. Because most cases are sporadic, mutations in the Hcrt gene or those for its receptors can account for no more than a small subset of the human narcolepsies. The HLA association, loss of neurons with signs of gliosis, and age of disease onset are consistent with autoimmune destruction of the Hcrt neurons accounting for the majority of narcolepsy, although a non-immune-mediated degenerative process has not been ruled out.

As mentioned previously, the MCH cell population is still intact in narcoleptic patients, and no mutation in the ppMCH or MCHR1/2 genes has been found or linked to arousal pathology. MCHR1 antibody has been found in vitiligo patients, an autoimmune pathology causing skin depigmentation. On the other hand, it has been proposed that MCH neurons act as the counterpart of Hcrt cells. In this way, Verret and collaborators showed in rats that, during a REM sleep rebound induced by specific REM sleep deprivation, a large population of MCH neurons were immunoreactive for the c-FOS protein, a marker of neuronal activity. In addition, after ICV infusion of the peptide in rats, REM sleep and non-REM sleep amounts were increased, leading the authors to propose a hypnogenic role for MCH neurons.

Concluding Remarks

The anatomical and electrophysiological data suggest that Hcrt neurons integrate complex information about circadian, limbic, and metabolic variables. This information is transmitted into a coherent output that excites arousal centers depending on the physiological demand. *In vivo* recordings of Hcrt neurons have revealed that these cells are active in anticipation of wakefulness and during consummatory behaviors but are mostly silent. Together the data gathered thus far indicate that the Hcrts provide an alarm signal and stability to arousal centers. In contrast, MCH neurons, which strongly innervate Hcrt cells and regulate Hcrt activity, appear to be modulated mostly by energy demand. The cross-talk between these two peptidergic neurons in the LH provides a balance system to modulate complex behaviors dependent on metabolic needs.

See also: Addiction: Neurobiological Mechanism; Drug Addiction: Cellular Mechanisms; Energy Homeostasis: Hypothalamic Development; Hypothalamic Structure–Function Relationships; Narcolepsy; Neuropeptides: Discovery; Neuropeptides: Food Intake; Reticular Activating System; Sleep and Sleep States: Hypothalamic Regulation.

Further Reading

Bittencourt JC, Presse F, Arias C, et al. (1992) The melanin-concentrating hormone system of the rat brain: An immuno- and hybridization histochemical characterization. *Journal of Comparative Neurology* 319: 218–245.

de Lecea L, Kilduff TS, Peyron C, et al. (1998) The hypocretins: Hypothalamus-specific peptides with neuroexcitatory activity. *Proceedings of National Academy of Sciences of the United States of America* 95: 322–327.

de Lecea L and Sutcliffe JG (eds.) (2005) *Hypocretins: Integrators of Physiological Functions*. New York: Springer.

Sakurai T, Amemiya A, Ishii M, et al. (1998) Orexins and orexin receptors: A family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* 92: 573–585.

van den Pol AN, Acuna-Goycolea C, Clark KR, and Ghosh PK (2004) Physiological properties of hypothalamic MCH neurons identified with selective expression of reporter gene after recombinant virus infection. *Neuron* 42: 635–652.

Hypothalamic Structure–Function Relationships

J P Card, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The mammalian hypothalamus is integral to regulatory systems essential for survival of organisms and their species. The influences of this small subdivision of the diencephalon upon physiology and behavior are far-reaching and emerge directly from the unique functional properties of its constituent neurons, its extensive interconnections with other regions of the neuraxis, and structural specializations that permit it to exert regulatory control over peripheral organ systems through the pituitary gland. The goal of this article is to place these essential functions within a structural framework that illustrates how the hypothalamus accomplishes its many and varied responsibilities. Rather than being comprehensive, the discussion focuses upon three aspects of hypothalamic function that are representative of its diverse functional capacity. These include (1) the unique timekeeping capabilities that allow it to impose temporal organization upon behavior and hormone secretion, (2) the means through which it orchestrates reproduction of the species, and (3) the neural systems that are involved in the control of feeding. These discussions are introduced with a short consideration of fundamental organizational principles that enable hypothalamic functions.

Functional Organization of the Hypothalamus

The limits of the hypothalamus are clearly defined by distinctive landmarks on the medial and ventral surfaces of the brain (Figures 1 and 2). These include the optic chiasm, the tuberal eminence and its associated pituitary stalk, and the mammillary bodies. In classical literature these landmarks were used to divide the overlying hypothalamus into chiasmatic, tuberal, and mammillary subdivisions. Although this nomenclature is still commonly used as a reference frame to localize hypothalamic cell groups, contemporary scientific literature has advanced organizational principles that are functionally based. Particularly influential in this regard are studies that have defined hypothalamic phenotype, circuit organization, and patterns of neuronal activation that occur in response to homeostatic challenges. Collectively, this literature suggests that hypothalamic nuclei are organized into three columns that extend through the longitudinal axis of this subdivision of diencephalon (Figure 3). Important to this organizational framework is the recognition that major neural inputs to the hypothalamus project onto these columns. This arrangement is central to hypothalamic organization in that it allows cell groups mediating similar functions to share relevant sensory input. In this regard it is useful to note that the control of motivated behavior is common to a number of hypothalamic functions (e.g., foraging, defensive responses, reproduction) and that, while these functions are localized to different hypothalamic cell groups, they often draw upon the same sensory input. Thus, there is functional utility for the distribution of sensory information among the functionally distinct cell groups that form hypothalamic columns. This is a fundamental tenet upon which contemporary hypotheses of the functional organization of the hypothalamus are based.

The hypothalamus imposes profound influences upon physiology and behavior through classical synaptic mechanisms and through hormonal regulation. The synaptology of neural systems that mediate these effects is complex and often involves synaptic relays that have not been fully defined. For example, although much is known of the identity of hypothalamic cell groups that impose temporal organization upon behavior, the neural circuitry through which this is accomplished is only understood in general terms. Similarly, multiple cell groups participate in the control of the autonomic nervous system through descending projections that often involve brain stem relays that serve as integrative centers. The use of relays to impose regulatory influence upon other systems, which is a characteristic feature of hypothalamic projection systems, serves the important purpose of allowing the hypothalamus to orchestrate integrated control of homeostasis and behavior.

A much broader influence upon peripheral physiology derives from neuroendocrine regulation of the pituitary gland. The pituitary controls peripheral organ systems through the release of hormones into the vascular circulation, and this release is similarly regulated by hormone release from neurons that reside in the hypothalamus (Figure 4). This 'neuroendocrine regulation' is achieved through a dedicated portal vasculature that transports bioactive molecules released from the hypothalamus to the pituitary. The absence of a blood–brain barrier at the contact zone where neural processes of hypothalamic origin terminate upon the fenestrated vasculature is an essential component of this system. The access to the vasculature afforded by these 'circumventricular organs' also

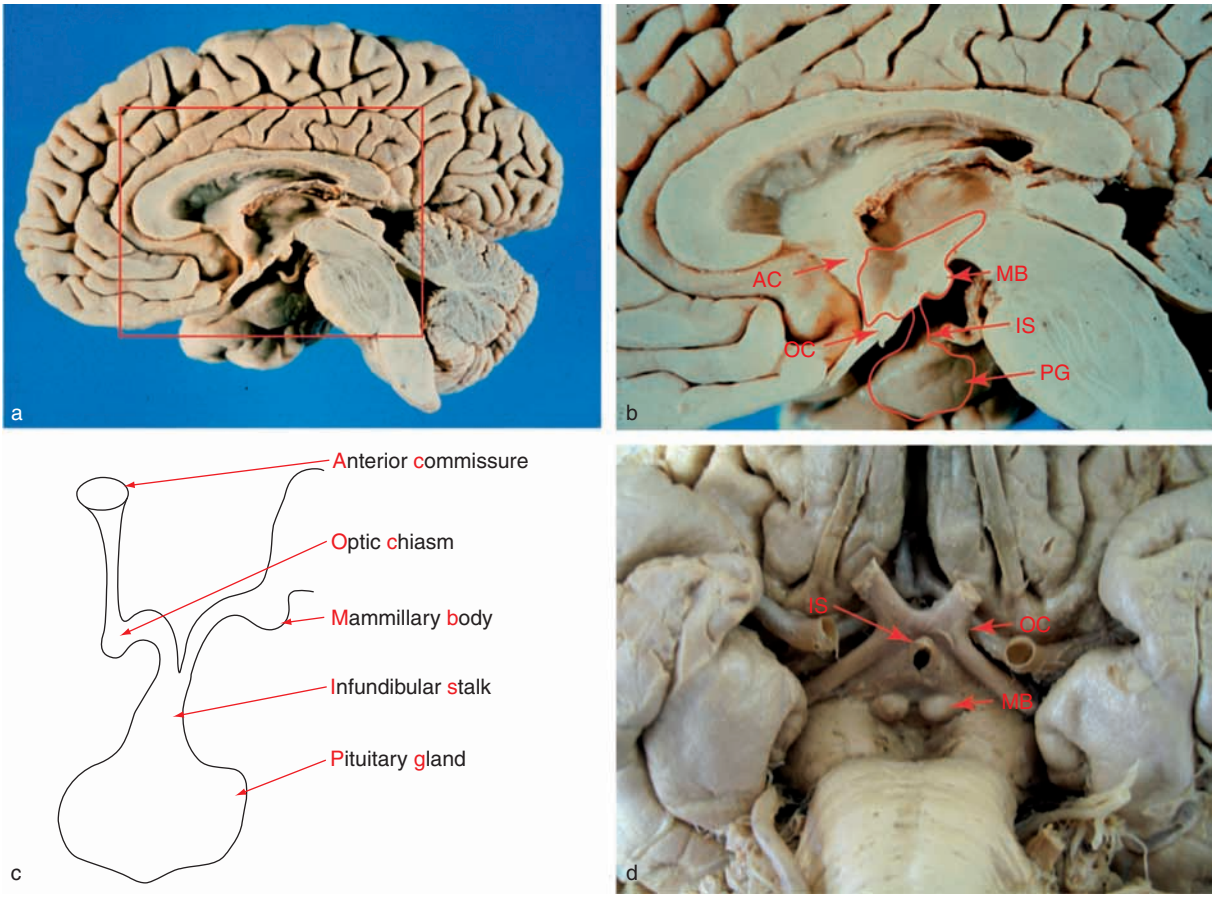

Figure 1 The position of the hypothalamus is illustrated in sagittal (a, b) and ventral (d) exposures of the human brain. The boxed area in panel (a) is shown at higher magnification in panel (b). (c) A schematic illustrating the relationship of the pituitary gland to ventral hypothalamus. This relationship is mapped onto the sagittal exposure of the hypothalamus in panel (b), which also defines the rostral, dorsal, and caudal boundaries of the hypothalamus.

allows circulating molecules to influence the activity of hypothalamic systems. This capacity for 'feedback regulation' is integral to the tightly regulated function of the hypothalamus. Thus, the hypothalamus exerts regulatory influence over disseminated systems, both central and peripheral, and is in turn regulated by the output of those systems.

Hypothalamic Influences upon the Temporal Organization of Behavior

The hypothalamus has long been implicated in the control of rhythmic aspects of behavior and hormone secretion. The early evidence in this regard derived from alterations in behavioral state documented in clinical and experimental studies that involved damage to the hypothalamus. Baron Constantin von Economo provided one of the more interesting and historically significant observations through his studies of the neural basis of encephalitis lethargica. The pandemic that gave rise to this disease occurred in the early 1900s and was characterized by high fever, double vision, headache, and lethargy. The most serious cases caused persistent coma, and postmortem analysis of the brains of these individuals by von Economo revealed lesions that involved the posterior hypothalamus. This observation took on added significance in the 1950s with the experimental demonstration that hypothalamic lesions produced either coma or insomnia, depending upon the location of the damage. These findings led Walle Nauta to postulate that the hypothalamus contains 'sleep centers' responsible for the control of the sleep–wake cycle. Contemporary studies have validated the role of the hypothalamus in behavioral state regulation and identified four functionally distinct cell groups that participate in this control.

The neurons of the hypothalamic suprachiasmatic nuclei (SCN) (Figure 3) provide one of the most prominent influences upon behavioral state. This small population of neurons overlying the optic chiasm plays a fundamental role in generating the daily oscillatory pattern of activity and rest. The SCN was initially

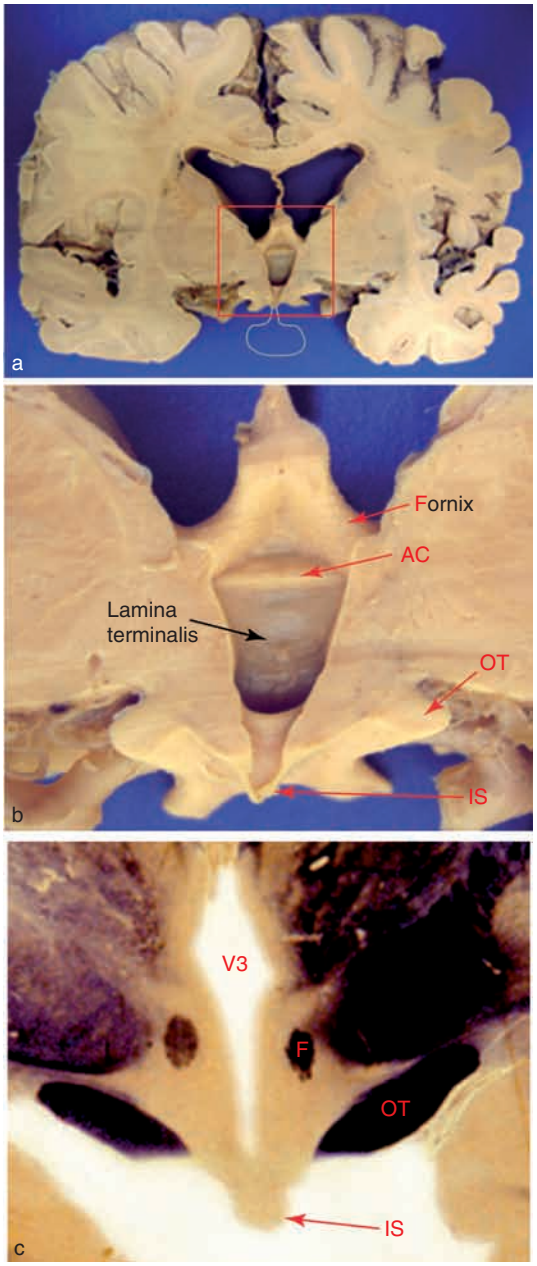

Figure 2 Landmarks defining the position of the human hypothalamus in coronal sections through the pituitary (infundibular) stalk are illustrated. The boxed area in panel (a) is illustrated at higher magnification in panels (b). Panel (c) illustrates a section through the same approximate area stained for myelin (blue-black). The rostral limit of the hypothalamus (lamina terminalis) is visible at the rostral extent of the third ventricle (V3). The columns of the fornix are also prominent at the dorsal extent of the diencephalon. This projection pathway is an important source of sensory information essential to the ability of the hypothalamus to orchestrate motivated behaviors. AC, anterior commissure; OT, optic tract; IS infundibular stalk; F, fornix.

implicated in behavioral state regulation on the basis of anatomical studies that demonstrated that it received a direct projection from the retina (subsequently coined the 'retinohypothalamic tract (RHT)') and on the basis of lesion experiments that eliminated the rhythmic aspects of the sleep–wake cycle and hormone secretion. Subsequently it was established that the activity of SCN neurons oscillates over a period approximating 24 h and that this circadian (*circa*, 'about'; *diem*, 'day') rhythm is genetically determined rather than being the emergent property of a network. That is, each SCN neuron is a circadian oscillator and the synchronous activity of the population is achieved through local circuit connections among the constituent neurons. Additionally, the rhythmic activity of the pacemaker is entrained to the environmental cycle of light and dark through the RHT, an observation with functional implications for the treatment of jet lag and some sleep disorders. Interestingly, the retinal ganglion cells that give rise to this projection are functionally distinct from those involved in visual perception and contain light-sensitive opsins. Thus, these retinal neurons represent a component of the visual system dedicated to the temporal organization of behavior rather than to visual perception.

Lesion studies in rodents have proven to be particularly illuminating in establishing the role of the SCN in the sleep–activity rhythm. Following destruction of the SCN, activity and sleep are distributed in small bouts across the 24 h period rather than being consolidated in periods of activity and rest. Importantly, the sum total of these bouts of activity and rest over a 24 h period in animals with complete SCN lesions approximates that observed in animals with an intact SCN. Thus, it is apparent that the SCN imposes temporal organization upon rest and activity rather than generating the behavior. Available evidence supports the conclusion that the temporal influences of the SCN are exerted through relays that wire the pacemaker into a larger polysynaptic circuit of neurons responsible for generating rest and activity. Support for this hypothesis can be found in studies demonstrating that caudal hypothalamic neurons that receive SCN input project to the locus coeruleus, a brain stem nucleus with activity that is correlated with arousal and vigilance. Functional studies indicate that the SCN functions through pathways such as this to promote arousal, which opposes the homeostatic drive for sleep.

The larger circuit of neurons devoted to the control of behavioral state includes other hypothalamic cell groups with activity that is associated with either arousal or somnolence. An interesting set of studies has demonstrated that neurons in the ventrolateral preoptic nuclei (VLPO) become active at the onset of sleep. In addition, it is known that these inhibitory neurons synapse upon caudal hypothalamic cell groups that are known to promote arousal. The latter include histaminergic neurons in the tuberomammillary

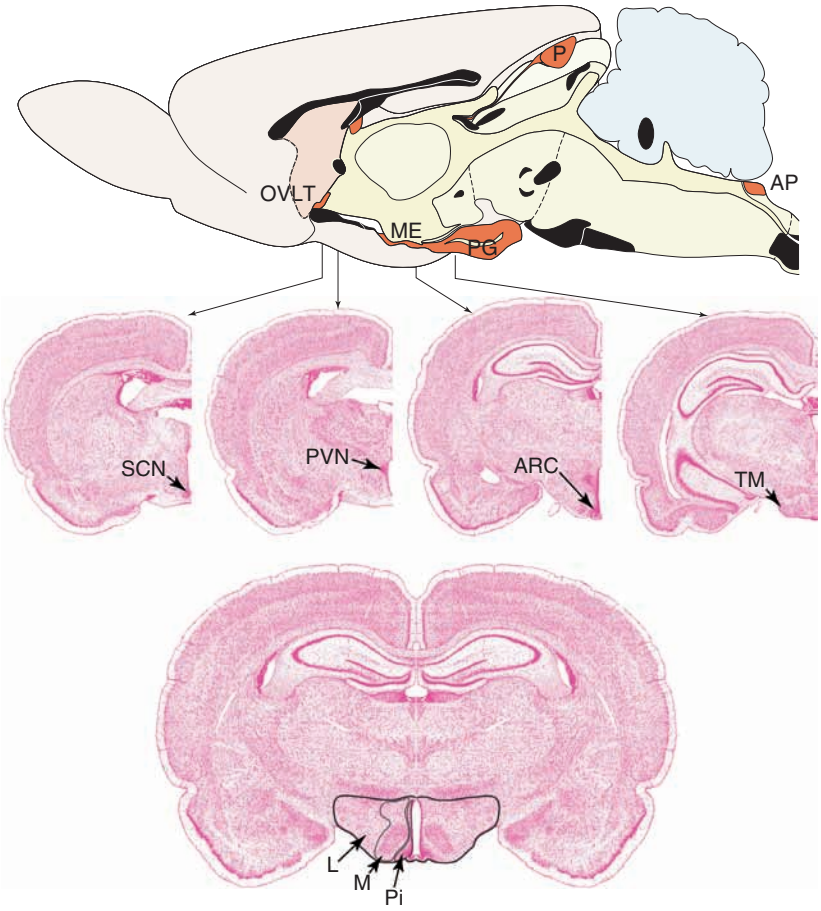

Figure 3 The locations of internal structures of the hypothalamus in the rat brain are shown in this figure. The locations of circumventricular organs lacking a blood–brain barrier are shown in red on the midsagittal schematic at the top of the figure. The cytoarchitecture of the hypothalamus and the positions of hypothalamic cell groups discussed in the text are illustrated in Nissl-stained sections in the middle of the figure. These sections illustrate half of the brain, presented in coronal section. The lower figure illustrates a coronal section through the portion of the hypothalamus that is connected to the pituitary gland (not shown); the heavy black outline defines the boundaries of the hypothalamus on either side of the third ventricle. The thin black outline on the left defines the boundaries of hypothalamic columns that are distinguished by cytoarchitecture, phenotype, and connections. These columns (Pi, periventricular; M, medial; L, lateral) extend through the rostrocaudal extent of the hypothalamus and are recipient of topographically organized sensory hippocampal input that is relayed through the septum. AP, area postrema; ARC, arcuate nucleus; ME, median eminence; OVLT, organum vasculosum of lamina terminalis; P, pineal gland; PVN, paraventricular nucleus; SCN, suprachiasmatic nucleus; TM, tubermammillary nucleus; PG, pituitary gland. Adapted from Swanson LW (1998) *Brain Maps: Structure of the Rate Brain*, 2nd edn. New York: Elsevier.

nuclei (Figure 3) and hypocretin-containing neurons in the lateral hypothalamic area that project diffusely to the cortical mantle and other subcortical cell groups known to promote arousal (e.g., locus coeruleus). These observations illuminate the neural circuitry underlying the 'sleep centers' advanced by Nauta in the 1950s, and indicate that the rostral and caudal hypothalamic cells groups function cooperatively to alter behavioral state. That is, the consequence of the increasing activity of VLPO neurons at the onset of sleep is to inhibit the hypocretin and histaminergic neurons that are known to produce arousal. While it is important to emphasize that these cell groups function within a larger circuit of neurons outside of the hypothalamus, this literature makes the important point that the hypothalamus is an essential component of this system.

The cooperative interaction of these functionally distinct hypothalamic cell groups, with one another and with the larger system of central nervous system (CNS) neurons that control behavioral state, is apparent in studies of the neural basis for narcolepsy. Narcolepsy is a neurological disorder characterized by excessive daytime sleepiness, abnormal rapid eye

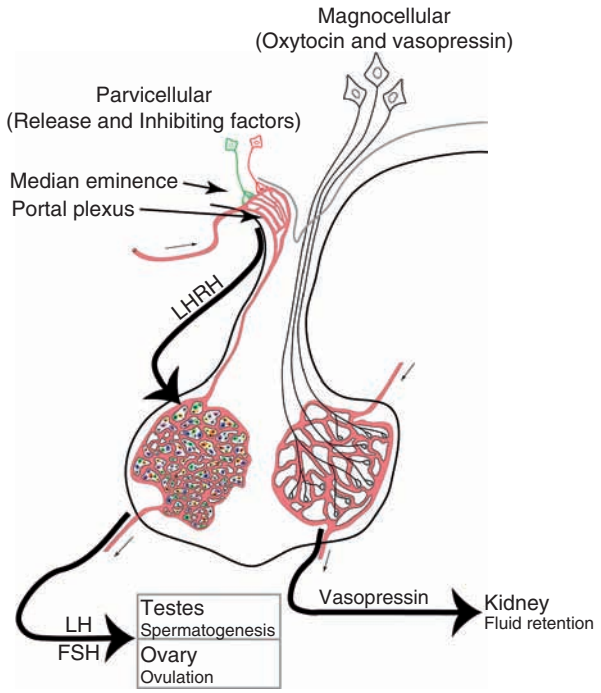

Figure 4 The means through which the hypothalamus imposes regulatory influences over the secretory activity of the pituitary gland is illustrated. The gland is connected to the hypothalamus via a stalk that serves both as a conduit for axons that terminate with a vascular plexus in the anterior caudal portion of the pituitary and as a conduit for long portal vessels that connect a fenestrated vascular plexus in the median eminence with vascular plexus in the anterior pituitary that surrounds hormone-producing cells. Magnocellular neurons provide the only direct neural connection between the hypothalamus and pituitary. The axons of these neurons release vasopressin or oxytocin into the vasculature, to influence the activity of peripheral targets. For example, vasopressin release (antidiuretic hormone) into the circulation acts on the kidneys to reduce fluid excretion. Control over the hormone secretion of cells in the anterior pituitary is achieved through the portal plexus connecting the median eminence to this portion of the gland. Bioactive molecules released into the portal vasculature either stimulate or inhibit the secretion of hormones from cells in the anterior lobe. The role of luteinizing hormone (LH) releasing hormone (LHRH) in the production of gametes provides a good example of this neuroendocrine regulation (FSH, follicle-stimulating hormone; see the text for a more detailed description).

movement (REM), or 'dream' sleep, and cataplexy. Narcoleptic patients typically exhibit multiple unregulated episodes of REM sleep during the waking hours but do not sleep more than normal individuals. In this respect their behavior is similar to experimental animals with complete SCN lesions. However, available evidence, derived from experimental studies of animal models and postmortem clinical analysis, indicates that the pathology underlying narcolepsy involves hypocretin neurons rather than neurons in the SCN. Cerebrospinal fluid levels of hypocretin are considerably reduced in patients with narcolepsy, and postmortem analysis reveals substantial reductions in the number of hypocretin cells as well as their terminal fields in areas such as the locus coeruleus. These observations are in accord with animal models in which alterations in behavioral state regulation similar to the symptoms of narcolepsy result from interference with hypocretin signaling. Considered collectively, these data demonstrate the importance of hypothalamic neurons to the CNS systems responsible for the control of behavioral state.

Hypothalamic Regulation of Reproductive Function

The hypothalamus contains the neural systems essential for reproduction. These systems exert regulatory control over the production of gametes (gameteogenesis; generation of ova and sperm) as well as the gender-specific motivated behaviors associated with successful mating. Separate populations of hypothalamic neurons, the activity of which must be coordinated for successful fertilization to occur, administer these interdependent components of reproduction. Thus, there is adaptive significance to these functionally distinct cell groups being located in close proximity to one another so that they can efficiently integrate sensory information relevant to reproductive function.

Hypothalamic cell groups responsible for the control of ovulation and spermatogenesis do so through control over the secretory activity of the anterior lobe of the pituitary gland (Figure 4). The pituitary exerts its control over peripheral targets such as the ovaries and testes through the regulated secretion of hormones, and this regulation is controlled by the hypothalamus through the release of bioactive molecules (peptides and neurotransmitters) into the fenestrated portal vasculature. These peptides and neurotransmitters stimulate or inhibit the release of hormones from the pituitary gland and are therefore more commonly known as release factors or inhibiting factors. In the case of gametogenesis, release of luteinizing hormone releasing hormone (LHRH) into the portal vasculature elicits the release of luteinizing hormone (LH) and follicle-stimulating hormone (FSH) from the pituitary (Figures 4 and 5). In males, these hormones act upon different cells in the testes in support of spermatogenesis. In addition, LH and FSH also stimulate the production of hormones (testosterone and inhibin, respectively) that provide negative feedback control over the release of LH and FSH from the pituitary. In females, FSH stimulates the growth of follicles in the early stages of ovulation and also is integral to the production of the hormone

Figure 5 Illustration of the organization of phenotypically defined rat median eminence neural systems that participate in the control of the pituitary gland. The top diagram is a sagittal schematic (OVLT, organum vasculosum of lamina terminalis; P, pineal; ME, median eminence; PG, pituitary gland; AP, area postrema). (a) Vasopressin-immunoreactive cells and processes in a sagittal section through the hypothalamus, corresponding to the boxed area in the sagittal schematic diagram. Vasopressin-containing magnocellular neurons in the upper left of the micrograph are the source of axons that pass through the intermediate layer of the median eminence into the pituitary stalk, to gain access to the posterior lobe of the pituitary. (b) The position of these axons in the intermediate zone of the median eminence is shown in the coronal plane. (c, d) The dense distribution of hypothalamic parvicellular axons in the external 'contact' zone of the median eminence is seen. These axons release their peptides into the fenestrated portal vessels for transport to the anterior pituitary. The distributions of somatostatin (growth hormone inhibiting factor)-containing axons (c) and of LHRH-containing axons (d) are illustrated. Note that these axons are differentially distributed within the external zone. Schematic diagram adapted from Swanson LW (1998) *Brain Maps: Structure of the Rat brain*, 2nd edn. New York: Elsevier.

estradiol. Levels of estradiol are further enhanced through LH-stimulated secretion of testosterone in the ovary, which is converted to estradiol. Elevated levels of estradiol in the circulation increase the sensitivity of the pituitary to LHRH (positive feedback), which ultimately produces a surge in LH release from the pituitary that results in ovulation. This simplified overview of neuroendocrine regulation of gameteogenesis illustrates the general mechanisms through which the hypothalamus exerts hormonal control over peripheral target organs, emphasizing the preeminent role of hypothalamus in the control of body homeostasis.

Successful reproduction depends upon coordinating reproductive physiology (gametogenesis) with the behaviors that give rise to copulation and fertilization. Reproductive behaviors are under the control of hypothalamic cell groups that exhibit sexual

dimorphism in their structure and connectivity. The experimental demonstration that the cytoarchitecture and synaptology of hypothalamic nuclei involved in reproductive behaviors can be modified through manipulation of sex steroid exposure during different developmental stages was very influential in establishing this dimorphism. Thus, a defining feature of these cell groups is the fact that they express receptors that render them responsive to circulating sex steroids. It is also important to note that brain circuitry involved in reproductive behavior includes cell groups outside of the hypothalamus that are steroid sensitive and project upon the hypothalamus. Chief among the hypothalamic nuclei implicated in the control of sexual behavior are the medial preoptic nuclei and the ventromedial hypothalamic nuclei. These cell groups reside within the same hypothalamic columns that control gameteogenesis, providing a neural substrate for sharing sensory information necessary for coordination of reproductive physiology and behavior. Among the neural pathways that exert a prominent influence upon the activity of these nuclei are relayed projections from the hippocampal formation conveying multimodal sensory information, components of the olfactory system carrying pheromonal information, and projections from the amygdala. While details concerning the organization and dimorphism of this circuitry continue to evolve, it is clear that the hypothalamus is an essential regulatory component of this system.

Hypothalamic Regulation of Feeding

The historical literature on the neural control of food intake is remarkably similar to that of the neural control of behavioral state, in that lesion studies were influential in identifying the hypothalamus as an important regulator of energy homeostasis. In this regard, investigations conducted in the 1940s documented substantial increases in food intake and obesity following medial hypothalamic lesions, and reductions in food intake and weight loss after lesions of the lateral hypothalamus. Contemporary literature has confirmed a preeminent role for the hypothalamus in the control of feeding and has identified a complex circuitry that is responsive to sensory measures of nutrient state. In fact, many of the recent advances in understanding of this system can be traced to the identification of circulating molecules linked to nutritional status that act upon the nervous system to produce compensatory alterations in food intake. This circuitry consists of neurons with opposing influences upon food intake and is a component of a larger regulatory network responsible for reinforcement of behavior. The system is also heavily dependent upon sensory feedback conveying environmental information that supports successful foraging. While the details of how the components of these systems function in a cooperative fashion continue to evolve, much has been learned regarding the identity of hypothalamic systems that participate in this regulation and how they sample the nutritional state of the parent organism.

Given that survival is dependent upon adequate ingestion of nutrients, it is not at all surprising that food and fluid intake is tightly regulated and involves redundant systems. The hypothalamus contains at least four phenotypically distinct populations of neurons that provide regulatory control over nutrient intake through dedicated systems that stimulate or inhibit ingestion of food and fluids. Sensory measures of gastrointestinal function, circulating nutrients, and nutrient stores determine the activation state of components of this system. These sensory measures of nutritional state reach the hypothalamus through neural circuits relayed from the viscera as well as through blood-borne signals. Thus, the access to the vasculature afforded by circumventricular organs in the hypothalamus and brain stem is an essential component of the modulation of hypothalamic feeding circuits.

The arcuate nucleus (**Figure 3**) contains two groups of neurons that are particularly influential in the control of feeding. This became apparent in early studies of rodents in which chemically induced lesions of the arcuate nucleus were shown to produce obesity. The basis for this lesion-induced obesity became apparent in the 1990s following identification of a blood-borne signal, the protein leptin, which acts upon the hypothalamus to modify feeding behavior. It is now firmly established that this protein is secreted into the circulation by white adipose tissue and that levels of leptin provide a relative measure of body fat stores. Circulating leptin acts upon arcuate neurons that express neuropeptide Y (NPY) and agouti-related peptide (AgRP), to stimulate feeding, and inhibits feeding through arcuate neurons that express proopiomelanocortin (POMC) and the cocaine and amphetamine-related transcript (CART). Both of these populations of arcuate neurons exert their influence through projection targets that have not been fully defined, but include neurons in the paraventricular hypothalamic nucleus and neurons in the lateral hypothalamic nucleus. The latter includes neurons that express hypocretin (also known as orexin) and melanin-concentrating hormone (MCH). Both of these cell groups give rise to widespread projection patterns and induce feeding. POMC/CART neurons are thought to exert their inhibitory influence through CNS neurons expressing melanocortin receptors. Investigations of the synaptology of these systems continue to reveal how they function to modify

appetite and feeding and have also raised the possibility that these circuits reorganize in response to changes in metabolic demand. Clearly, much has been learned about the identity of neural systems that control feeding, and this has made it abundantly clear that the hypothalamus is central to this regulation. Identification of these systems has also created many opportunities for exploring how they function in the control of energy balance.

As noted previously, blood-borne leptin plays an important role in modulating the activity of central circuitry controlling food intake. However, leptin is not the only circulating molecule that modulates this circuitry. The gastrointestinal tract is the source of other important signaling molecules that can modulate these circuits either directly, at the level of the hypothalamus, or indirectly, by acting upon brain stem neurons that project to the hypothalamus. All of these molecules exert their influence through circumventricular neurons that are lacking a blood–brain-barrier. Ghrelin, a hormone secreted by the stomach, is one such molecule that increases feeding. Increased circulating levels of ghrelin occur in response to fasting and NPY/AgRP arcuate neurons are among those in the CNS that express ghrelin receptors. Circulating molecules secreted by the viscera also act upon neurons outside of the blood–brain-barrier in the area postrema of the caudal brain stem. These include the hormones glucagon-like peptide-1 (GLP-1) and cholecystokinin (CCK). CCK and GLP-1 activate area postrema neurons, which, in turn, activate neurons in the adjacent nucleus of the solitary tract (NTS) that project to the hypothalamus. Interestingly, GLP-1 is also produced in a subset of the NTS neurons that project to the hypothalamus. The central effects of CCK and GLP-1 function to reduce food intake. Experimental evidence has also implicated GLP-1 signaling in conditioned taste avoidance as well as nausea. These observations are in accord with the important role of the NTS in taste.

The aforementioned hypothalamic cell groups interact with a larger constellation of neural circuits that contribute to the maintenance of adequate nutritional state. These include systems involved in control of behavioral state, feeding behavior (foraging), and the rewarding aspects of nutrient intake. Again, the way in which these systems interact is only understood in general terms, and defining these organizational principles remains an important goal.

Summary and Conclusions

There can be little debate over the assertion that the hypothalamus exerts essential regulatory control over the maintenance of homeostasis and behavioral state. This control is dependent upon structural specializations that enable circulating molecules to influence the activity of hypothalamic neurons. Structural specializations that permit the hypothalamus to exert neuroendocrine control over pituitary secretion are also fundamental to the regulatory capacity of this small but influential subdivision of the diencephalon. The use of relays to distribute information from hypothalamic cell groups is also a fundamental feature of the means through which hypothalamic neurons exert their control. This use of relays suggests that the output of functionally defined hypothalamic neurons is often integrated with that of other systems. This hypothesis is parsimonious, with the conclusion that physiology and behaviors orchestrated by the hypothalamus are relevant to a variety of systems and must be integrated with their outputs. Collectively, these observations reveal the hypothalamus to be a dynamic regulatory unit with integrated activity that is essential to an organism's abilities to function productively in their environment.

See also: Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Gastrointestinal Control; Circumventricular Organs; Energy Homeostasis: Hypothalamic Development; Narcolepsy; Sleep and Sleep States: Hypothalamic Regulation.

Further Reading

Hobson JA, Pace-Schott EF (2003) Sleep, dreaming and wakefulness. In: Squire L, Bloom FE, McConnell SK, Roberts JL, Spitzer NC, and Zigmond MJ (eds.) *Fundamental Neuroscience*, 2nd edn., ch. 41, pp. 1085–1108. New York: Academic press.

Horvath TL and Diano S (2004) The floating blueprint of hypothalamic feeding. *Nature Reviews Neuroscience* 5: 662–667.

Moore RY (2003) Circadian timing. In: Squire L, Bloom FE, McConnell SK, Roberts JL, Spitzer NC, and Zigmond MJ (eds.) *Fundamental Neuroscience*, 2nd edn., ch. 41, pp. 1067–1084. New York: Academic press.

Saper CB (2001) The central autonomic nervous system: Conscious visceral perception and autonomic pattern generation. *Annual Review of Neuroscience* 25: 433–469.

Saper CB, Chou TC, and Elmquist JK (2002) The need to feed: Homeostatic and hedonic control of eating. *Neuron* 36: 199–211.

Sawchenko PE (1998) Toward a new neurobiology of energy balance, appetite, and obesity: The anatomists weigh in. *Journal of Comparative Neurology* 402: 435–441.

Swanson LW (1998) *Brain Maps: Structure of the Rate Brain*, 2nd edn. New York: Elsevier.

Simerly RB (2002) Wired for reproduction: Organization and development of sexually dimorphic circuits in the mammalian forebrain. *Annual Review of Neuroscience* 25: 507–536.

Swanson LW (1998) *Brain Maps: Structure of the Rate Brain*, 2nd edn. New York: Elsevier.

Swanson LW (2000) Cerebral hemisphere regulation of motivated behavior. *Brain Research* 886: 113–164.

Hypothalamic–Pituitary–Adrenal (HPA) Axis

N Pecoraro and M F Dallman, University of California at San Francisco, San Francisco, CA, USA

© 2009 Published by Elsevier Ltd.

The Hypothalamic–Pituitary–Adrenal Axis

The hypothalamic–pituitary–adrenal (HPA) axis keys reactive and predictive physiological and mental adjustments to environmental demand. Understanding HPA axis complexity requires viewing it as the final common output pathway of diverse inputs – including circadian pattern generators; systems sensitive to metabolic, immune, and nociceptive stimuli; cognitive assessments; and intrahypothalamic networks coordinating neuroendocrine and other motor outflows – and recognizing HPA axis feedback and feed-forward effects on brain.

Motor neurons in the paraventricular nucleus (PVN) of the hypothalamus project to the median eminence and release corticotropin-releasing factor (CRF) into a portal vasculature perfusing anterior pituitary glandular corticotropes that release adrenocorticotropic hormone (ACTH) into the systemic circulation. Systemic ACTH stimulates adrenal cortex to release glucocorticoids (GCs) into the systemic circulation to act on peripheral tissues and organs, and brain regions involved in cognition, memory, motivation, and primary motor outflows, including HPA axis outflow (see Figure 1).

The Paraventricular Nucleus of the Hypothalamus

The PVN integrates diverse inputs and contains distinctive neuroendocrine, behavioral, and autonomic cell groups. Neurosecretory regions include parvicellular neurons containing CRF/arginine vasopressin (AVP) peptides that are co-packaged in vesicles, and co-secreted. Synthesis is drive dependent. CRF neurons arc through perifornical hypothalamus exhibiting boutons of passage, sweeping ventrally to the median eminence and anterior pituitary. Magnocellular cells containing oxytocin (OXY) or AVP reside laterally to CRF neurons and project to the neural lobe of the pituitary.

Nonneuroendocrine, descending PVN divisions control behavioral and autonomic outflows. Preautonomic cell groups include the 'dorsal cap,' and a group at the caudolateral tip of the PVN extending caudally into zona incerta. A related retrochiasmatic group participates in a putative pattern generation network (described later).

Pituitary Corticotropes

Proopiomelanocortin (POMC) molecule cleavage products (e.g., ACTH) are stored in intracellular corticotrope vesicles and secreted in response to CRF associating with CRF receptor (CRF_1R) via G-protein/cyclic adenosine monophosphate (c-AMP)-mediated mechanisms. AVP weakly stimulates ACTH secretion, via a receptor ($AVP_{1b}R$)-mediated mechanism, but synergizes with CRF, dramatically increasing ACTH secretion. Corticotropes contain glucocorticoid receptors (GRs) and corticosteroid-binding globulin (CBG, or transcortin). GR activation inhibits synthesis and secretion of ACTH through a negative glucocorticoid recognition element (GRE) on POMC DNA, but CBG binds GCs with high affinity. Tonic inhibitory action of GCs on corticotropes requires high concentrations, exceeding CBG capacity, making ACTH secretion vulnerable to high doses of exogenous GCs – for example, during treatment of autoimmune diseases, in patients with pituitary ACTH-secreting adenomas (Cushing's disease), or when warding off transplant rejection.

Adrenal Zona Fasciculata

Unlike pituitary corticotropes that store hormone, the adrenal cortex zones synthesize and secrete hormones on ACTH demand. ACTH acts via c-AMP at the rate-limiting step of cholesterol transfer from cytosolic esters to mitochondria, where steroid synthesis begins. ACTH increases other enzymes involved in GC synthesis. All three zones of cortex (zona glomerulosa, zona fasciculata, and, in primates, zona reticularis) respond to ACTH, acting through melanocortin receptors (MCR2). In primates, injection of ACTH stimulates secretion, within minutes, of aldosterone (glomerulosa), cortisol (fasciculata), and dehydroepiandrosterone sulfate (reticularis). The major steroids (cortisol in primates; corticosterone in most rodents) synthesized in the zona fasciculata are most important in HPA function. ACTH transiently stimulates zona glomerulosa cells, but the major role of ACTH is sustained stimulation of GCs.

CBG (Transcortin) Binding

CBG (transcortin), synthesized and secreted from liver, is a high-affinity binding protein for GCs, and, under basal conditions, when adrenal secretion is low, 95–99% of the secreted steroid is bound to CBG without diffusion to targets. CBG–GC is in reversible equilibrium with free GCs. CBG is a saturable reservoir and buffer for biologically active GCs, so that considerable free steroid is able to diffuse to target

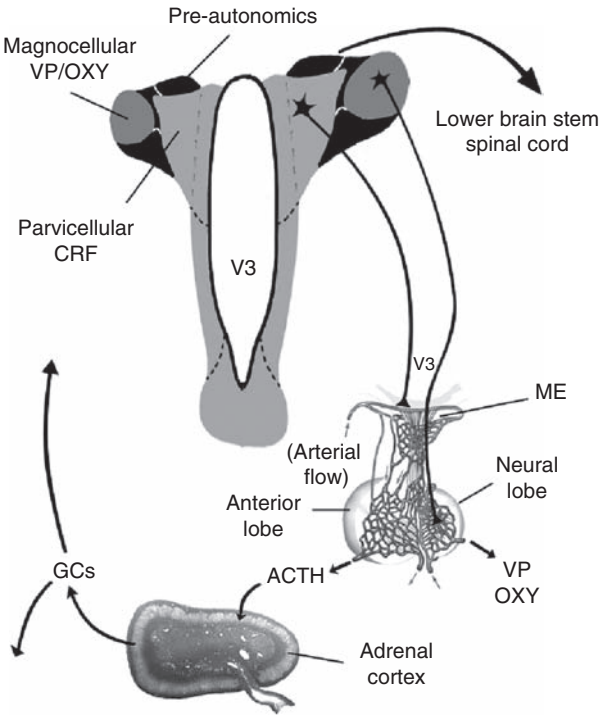

Figure 1 The hypothalamic–pituitary–adrenal axis. The paraventricular nucleus of the hypothalamus is depicted as a bilateral coronal section in an oblique view, with the third ventricle (V3) indicating the midline. The periventricular region (light gray) contains parvicellular neuroendocrine cells with corticotropin-releasing factor (CRF)/arginine vasopressin-containing neurons in the dorsolateral region; these project to the median eminence (ME) and portal vasculature to the anterior pituitary. The magnocellular neuronal division (medium gray) consists of cells containing oxytocin (OXY) and vasopressin (VP); these project directly to the posterior, neural lobe of the pituitary. The descending division of the paraventricular nucleus (black) contains preautonomic neurons projecting to lower brain stem areas and spinal cord, involved in both behavioral and autonomic regulation. CRF/arginine vasopressin release into the portal vasculature of the anterior pituitary stimulates glandular corticotropes to release adrenocorticotropin hormone (ACTH) into the systemic circulation. Plasma ACTH then stimulates adrenocortical cells to synthesize and release glucocorticoids (GC) into the systemic circulation to act on peripheral tissues and organs and the central nervous system.

cells during stress. Because GCs inhibit hepatic CBG synthesis, when tonic GC concentrations are low, circulating CBG is high, and vice versa. Because total GC concentrations in plasma (bound and free) are usually measured, it is worth knowing the history of GC secretion when collecting acute samples.

Inhibition of the HPA Axis by GC

GC elevations inhibit activity in PVN CRF cells and pituitary corticotropes, first by membrane GR-mediated effects, then by GC-associated GR protein–protein interactions altering gene expression. Feedback ensures controlled neuroendocrine responses. Longer term feedback effects exist in both brain and pituitary (e.g., through GC–GR complexes interacting with a negative GRE on POMC in corticotropes). Chronic GCs diminish HPA axis outflow in proportion to exposure, provided no concurrent stress exists, making the axis unresponsive to new stressors and thus explaining the requirement for extra GC coverage during major surgery in patients on chronic GC therapy.

There is rapid inhibitory GC feedback on both CRF and ACTH secretion. In hypothalamic slices *in vitro*, GC alters neuronal activity, including CRF-synthesizing cells, within 3 min. GC stimulates endocannabinoid secretion, presynaptically inhibiting glutamate-stimulated miniature excitatory postsynaptic currents (mEPSCs) via endocannabinoid EC-1 receptors. GCs infused directly into neurons does not have this effect, and rapid GCs actions on pituitary ACTH secretion *in vitro* are not inhibited by exposure to protein or by RNA synthesis inhibitors, suggesting membrane-mediated feedback.

Distinctive HPA Axis Drivers

Distinctive brain systems impinge on HPA axis outflow (Figure 2), including two circadian oscillators, brain stem, vasculature, and circumventricular organ systemic networks, cortico-striato-pallidal-brain stem cognitive networks, and an intrahypothalamic network that coordinates other drivers and motor outflows.

Periodic/Circadian Drivers of the HPA Axis

In constant darkness, HPA axis activation exhibits free-running circadian oscillations that entrain to light/dark cycles. Plasma GC concentrations increase at least tenfold around the time of waking. GCs decline throughout the activity phase, propped up by postprandial elevations, before they reach a nadir in early rest. When restricted feeding time is pitted against a light/dark cycle, a separate food-entrainable oscillator allows the HPA axis to separately track mealtimes.

The suprachiasmatic nuclei The suprachiasmatic nuclei (SCN) sit atop the optic chiasm in the basal hypothalamus, serving as the master, light-entrainable oscillator (LEO) in mammals. The SCN receive transduced light stimuli via the retinohypothalamic tract to entrain endogenous pacemaker activities of the SCN, driving the HPA axis and PVN preautonomic neurons. The SCN indirectly influence PVN CRF neurons via substantial projections to the subparaventricular zone, which targets the PVN with γ-aminobutyric acid (GABA)ergic inputs. The SCN project substantially to the dorsomedial hypothalamus (DMH), which

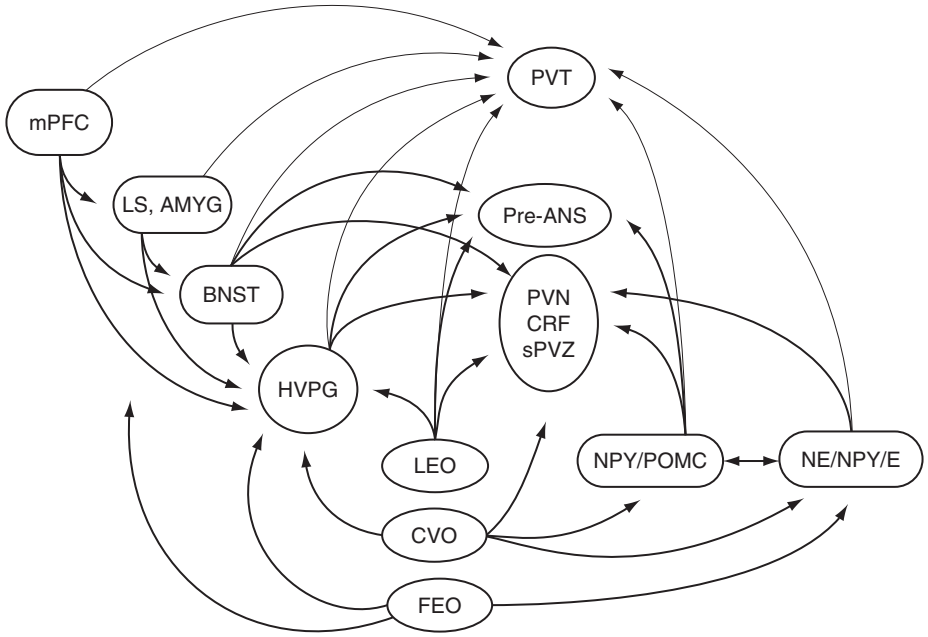

Figure 2 This highly simplified schematic depicts the major drivers of the hypothalamic–pituitary–adrenal (HPA) axis, with some indications of their coordinated drive on the autonomic outflow (pre-ANS; preautonomic nervous system) and on one another. For simplicity, paraventricular nucleus (PVN) corticotropin-releasing factor (CRF)-containing cells are grouped with the heavily interconnected subparaventricular zone (sPVZ). The suprachiasmatic nuclei, the master, light-entrainable oscillator (LEO), provide endogenous circadian drive to the PVN CRF-containing cells, the PVN preautonomics, and the hypothalamic visceromotor pattern generator (HVPG). During restricted, scheduled feedings a second food-entrainable oscillator swings into motion to track mealtimes and entrain food anticipatory rhythms. Systemic threats to homeostasis are signaled through circumventricular organs (CVO), the melanocortin system (NPY/POMC; neuropeptide Y/proopiomelanocortin) in the arcuate nucleus of the hypothalamus, and the ascending medullary catecholaminergic neurons (NPY and NE/E; norepinephrine, epinephrine). Psychogenic drive on the HPA axis is represented by multilayered cortico-striato-pallido-hypothalamic networks, in this case showing the visceromotor medial prefrontal cortex (mPFC) projecting to the lateral septum (LS) and amygdala (AMYG), bed nucleus of the stria terminalis (BNST), and HVPG. Virtually all HPA axis drivers also project to the 'neuroendocrine' paraventricular thalamus (PVT). Also indicated is the food-entrainable oscillator (FEO).

also projects to neuroendocrine and preautonomic divisions. AVP levels in and around the PVN vary on a circadian basis, being highest at rest, during basal HPA axis activity. AVP infusions into brain cause rapid decreases in GCs, and lesser decreases in ACTH. The DMH is the most effective infusion site, providing substantial GABAergic inhibitory fibers to CRF neurons. Disinhibition of HPA axis activity by AVP antagonists occurs only in SCN-intact animals, suggesting origination in the SCN. The SCN project to the medial preoptic (MPO) hypothalamus, an area involved in coordination of outflows. Outside hypothalamus, SCN project to the paraventricular thalamus (PVT), a PVN target involved in both acute and chronic stress networks. Pseudorabies virus injected into the adrenals is retrogradely propagated transsynaptically via the intermediolateral column to PVN preautonomics in the SCN. This pathway controls adrenocortical sensitivity to ACTH. In early darkness, bright light inhibits GCs, but not ACTH, secretion, only in SCN-intact rats. Thus, the SCN influence HPA axis output through CRF motor neurons and by neuronally controlling adrenal sensitivity to ACTH.

The food-entrainable oscillator A separate, food-entrainable circadian oscillator (FEO) prevails during restricted, scheduled feedings to entrain activity, feeding, drinking, body temperature, HPA axis activity, and peripheral metabolic rhythms. Entrainment is robust in SCN-lesioned animals that would otherwise be arrhythmic under a light/dark cycle when freely feeding. In intact rats, SCN-coupled rhythms free-run (e.g., 26 h period) in constant dark, but animals can independently and simultaneously anticipate scheduled meals at different periods (e.g., every 24 h) from the free-running rhythms. It takes 3–13 days for anticipatory activity to precede scheduled meals, but evidence suggests that single meals can reset food-entrainable rhythms. Food anticipatory activity continues for several days in the absence of the food zeitgeber prior to damping.

The anatomical locus of the FEO is unknown. Ascending medullary norepinephrine/neuropeptide Y (NE/NPY) systems entrained by scheduled meals drive prefeeding peaks in HPA axis output and feeding rhythms. Lesions of the ascending ventral noradrenergic bundle (VNAB) prevent food-anticipatory

elevations in plasma GCs. Striatal and pallidal inputs to the HPA axis – namely, the central nucleus of the amygdala (CeA) and bed nucleus of the stria terminalis (BNST) – exhibit phase shifts in clock gene expression during food entrainment. Like SCN, both the CeA and the BNST oval (BNST-ov) nucleus exhibit circadian rhythms in the clock gene period2 (per2) that peak at activity onset under *ad libitum* feeding. Restricted feeding forces per2 rhythms in CeA and BNST-ov out of phase relative to SCN. Constant light and free feeding abolish per2 rhythms in all three areas; these are reestablished by restricted feeding, such that peak per2 is in antiphase to feeding. Adrenalectomy (ADX) also abolishes per2 rhythms in limbic regions in a GC-reversible fashion. Lesions of the DMH (or insular cortex) prevent preprandial temperature elevations, and severely damp locomotor activity, pointing to integrated output from the DMH. Both systemic and limbic networks are candidate mediators of HPA axis entrainment by food.

Homeostatic Inflow: Systemically Responsive Networks Driving the HPA Axis

The HPA axis is highly sensitive to energy balance and other systemic threats, such as nociception and immune challenge, which have distinct and shared pathways to the HPA axis.

The arcuate NPY/melanocortin system The melanocortin system (Figure 2) includes cell groups in the arcuate nucleus (ARC) of the hypothalamus and the medullary nucleus of the tractus solitarius (NTS) sensitive to energy balance, and expressing POMC (usually co-expressing cocaine and amphetamine-related transcript (CART)) or neuropeptide Y (co-expressing agouti-related peptide (AgRP) in the ARC). POMC neurons mainly reside in the ARC, projecting primarily to stress-sensitive sites, including the PVN, perifornical lateral hypothalamus, NTS, reticular nuclei (RET), lateral parabrachial nucleus, nucleus accumbens (NACC), lateral septal nucleus, BNST, CeA, and medial amygdala (MeA). Tyrosine hydroxylase (TH)-negative POMC-expressing neurons also originate from the NTS. ARC NPY/AgRP neurons have similar projection patterns (e.g., PVN, DMH, PH, PVT, MPO, lateral septal nucleus). Energy balance signals may be transmitted neurally and through proximal circumventricular organs (CVOs).

ARC POMC. ARC POMC cleaves into α-melanocyte-stimulating hormone (α-MSH), a potent agonist at melanocortin-3 and -4 receptors (MCR3 and MCR4). Intracranial administration of leptin activates POMC neurons, suppresses feeding, and increases autonomic outflow. CRF mediates the inhibitory effects of α-MSH on feeding. Like α-MSH, central administration of CART also inhibits feeding, whereas fasting reduces CART mRNA expression. Three populations of CART neurons contribute innervation to the PVN, including medullary and arcuate POMC neurons and a third unknown source.

NPY neurons. NPY neurons directly innervate PVN CRF neurons. Increases in NPY elevate CRF mRNA expression, ACTH and GC secretion, and food intake, and suppress sympathetic outflow to brown fat. In contrast to α-MSH, AgRP is an antagonist at MCRs. Negative energy balance (e.g., low insulin and leptin) activates NPY neurons, including inhibitory GABAergic fibers to POMC neurons. POMC and NPY neurons are 'dueling rheostats' on energy balance.

GC effects on POMC/NPY. POMC/CART and NPY/AgRP neurons in the ARC express GRs. ADX or inhibition of GC synthesis reduces expression of CART mRNA. Though CART appears anorectic, CART-expressing medullary neurons co-express an epinephrine-synthesizing enzyme, phenylethanolamine-*N*-methyltransferase (PNMT), which is involved in feeding responses to glucoprivation. In addition, CART phosphorylation of cAMP response element-binding protein (CREB) in PVN CRF neurons should have additive effects with α-adrenoreceptor activation of CRF expression. Despite opposite signaling functions of CART and GCs, evidence suggests some positive feedback. GCs also stimulate NPY gene expression, peptide synthesis, receptor activity, and NPY-induced feeding. Here, positive feedback biases the system to correct further energetic disturbances. Clear evidence for functional interplay between GC and melanocortin systems exists, but the details require further elucidation.

Systemic threats: The medullary NE/E/NPY system Stimuli signaling actual threats to homeostasis are categorized as systemic stressors; these include including ether, hemorrhage, glucoprivation, and nociception, and are transduced by various lower brain stem NPY/catecholaminergic cell groups, including the dorsomedial medulla (NTS; A2–C2 cell groups) and ventrolateral medulla (A1–C1 cell groups). Ninety percent of all catecholaminergic innervation ascends to the PVN via the ventral noradrenergic bundle (VNAB). Immunostaining for dopamine β-hydroxylase, a marker for epinephrine (E) and norepinephrine (NE), shows extensive innervation of the PVN. NPY immunostaining shows similar, but more extensive, patterns in the PVN. NPY co-localizes in lower brain stem with PNMT in most adrenergic neurons (C1, C2, C3), but shows less co-localization in noradrenergic A1 neurons, about 60% rostrally and many fewer caudally. Some ascending VNAB fibers arrive at the median eminence rather than at the PVN,

possibly controlling exocytosis. Eergic, NEergic, and NPYergic stimulation of the PVN increases HPA axis outflow and feeding. Disrupting ascending inputs through chemical lesions, inhibiting transmitter synthesis in originating neurons, α-1 or β receptor blockade in PVN, or neutralization of CRF neurons, all disrupt HPA axis responses to systemic threats.

Glucoprivation. Acute glucoprivation by peripheral administration of insulin or 2-deoxyglucose (2DG) results in rapid counterregulatory HPA axis and feeding responses. Destruction of medullary Eergic and NEergic neurons via retrograde transport of selective neurotoxins delivered to PVN profoundly blunts counterregulatory responses, but has no effect on HPA responses to psychogenic stress or circadian variations in GCs, thus validating a systemic/psychogenic distinction.

Pain. Nociceptive stimuli (e.g., sharp pinch, thermal, or visceral pain) stimulate the HPA axis through systemic and psychogenic networks. Spinohypothalamic pathways transmit pain stimuli to the lateral hypothalamus (LH), which projects ipsilaterally to PVN CRF neurons. The spinothalamic pathway projects from spine to ventral posterolateral thalamus, then to sensory cortical areas. The spinoreticulothalamic pathway projects from spinal cord to reticular formation, then to midline and intralaminar thalamic nuclei, and on to limbic cortical areas (cingulate, piriform, and entorhinal cortices). Vagal and glossopharyngeal nerves convey visceral pain to the NTS, then to the parabrachial nucleus, thalamus, and visceromotor prefrontal cortex. A2 and C2 cell groups also project to A1 and C1 neurons in the ventrolateral medulla, locus coeruleus, amygdala, and PVN, which shows strong bilateral activation of CRF parvicellular neurons in response to many noxious stimuli. VNAB hemisections reduce PVN activation primarily on the hemisected side. Noncatecholaminergic neurons in the NTS, terminating in the PVN, also express fos after nociceptive stimuli. (Cognitive control of the HPA axis is discussed later.)

Proinflammatory stimuli. Systemic inflammatory responses to immune challenges are powerful stimuli to autonomic outflow and HPA axis activation, which inhibits proinflammatory processes. Proinflammatory molecules include tumor necrosis factor-α (TNF-α), interleukin-1β (IL-1β), and interleukin-6 (IL-6). Brain vasculature expresses receptors for TNF-α, IL-1β, and IL-6 under basal conditions; further induction occurs following immune challenge when plasma cytokine concentrations increase. Activation of cytokine receptors in the blood–brain barrier results in the transcription of various substances, including prostaglandins (PGs) following signaling cascades involving cyclooxygenase-2 (COX-2) and nuclear factor-kappa B (NF-κB). The brain vasculature and irrigation systems are thought to be the primary source of central PGs following systemic cytokine treatment, as COX-2 expression is highly co-expressed in areas staining for von Willebrand factor, a marker for cerebral endothelium. Activation is particularly striking in blood vessels, choroid plexus, and the CVOs.

Prostaglandin E_2 (PGE_2) is involved in neuronal responses to systemic immune challenge. Injection of IL-1 results in PGE2 expression in CVOs, preoptic nuclei, PVN, dorsal hippocampus, and lateral ventricles. Central administration of PGE_2 increases HPA axis and autonomic outflows. Febrile responses may originate in the MPO, whereas HPA axis activation may originate in the PVN, as prostaglandins (PGs) induce CRF mRNA and type 1 receptors in the PVN. Immune challenge activates specific receptors for PGE_2 (subtypes EP1–EP4) in brain regions controlling HPA axis and autonomic outflows. Neurons expressing EP1–EP4 particularly show c-fos activation after systemic IL-1β administration in CVOs and central stress networks, with highest expression in PVN CRF neurons. Central HPA axis and autonomic outflows following systemic immune challenge appear to be coordinated by classic central stress response networks, in part by activation of brain endothelial and leptomeningeal tissues, resulting in signal transduction to neuronal tissues.

Immune signaling may also originate in liver, as hepatic vagotomy eliminates febrile responses, and blunts brain stem c-fos responses, PVN noradrenergic levels, and ACTH responses to endotoxin. Endotoxin-clearing Kupffer cells also release PGE_2 in response to immune challenge stimulating excitatory vagal afferents, relaying signals to regions involved in febrile (MPO) and HPA axis (PVN) responses. Rapid febrile responses to immune challenges occurring prior to cytokine elevations utilize this pathway. Whether the argument holds for rapid HPA axis activation has not been thoroughly tested. Hepatic vagotomy has reduced HPA axis output by about half in response to endotoxin, but had a much smaller effect on responses to IL-6. Unfortunately, HPA axis responses were not measured until 2 h after injections, and not all studies agree on the effects of vagotomy. Inactivating the Kupffer cell phagocytotic activity using gadolinium chloride failed to reduce HPA axis responses to lipopolysaccharide (LPS), but again, ACTH and corticosterone were measured many hours later, long after cytokines were produced and acted in brain.

The Hypothalamic Visceromotor Pattern Generator

Recent anatomical evidence suggests the existence of a hypothalamic visceromotor pattern generator (HVPG)

located between hypothalamic neuroendocrine, preautonomic, and behavioral motorneuron pools, involving a DMH 'hub' and five preoptic nuclei. Figure 3 shows a schematic of connections between the HVPG and the PVN. Core nodes of the HPVG are massively interconnected, with each node idiosyncratically projecting to neuroendocrine and preautonomic cell groups. Notable among the few extrahypothalamic targets are the BNST, lateral septum (LS), periaqueductal gray (PAG), subfornical organ (SFO), and the PVT. The coordination of visceromotor outflow likely includes not only sympathetic, but also parasympathetic, outflow via strong connections between the DMH and parasubthalamic nucleus, which projects heavily to the dorsal vagal complex.

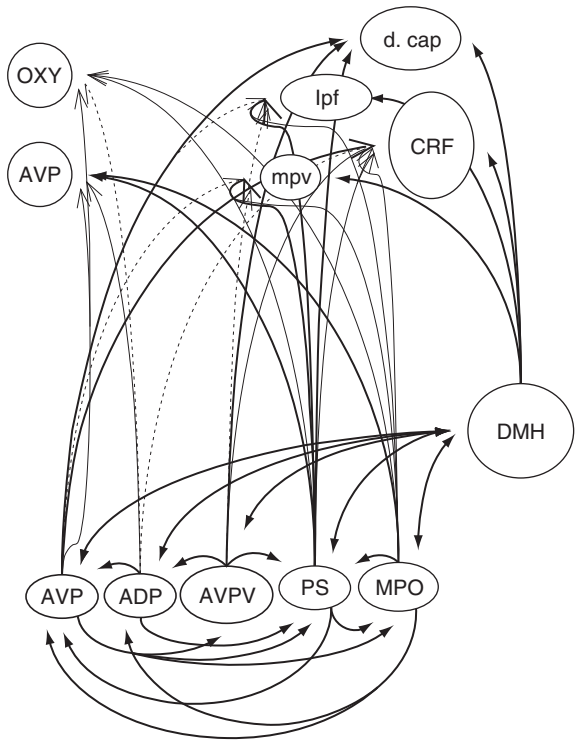

Figure 3 The hypothalamic visceromotor pattern generator. This schematic is based on Thompson's and Swanson's hodological analysis of relations among various hypothalamic nuclei with respect to neuroendocrine and autonomic divisions of the paraventricular hypothalamus. The hypothalamic visceromotor pattern generator consists of the dorsomedial hypothalamus (DMH), its major hub, and five preoptic nuclei, including the anteroventral preoptic (AVP), anterodorsal preoptic (ADP), anteroventral periventricular preoptic (AVPV), parastrial nucleus (PS), and medial preoptic nucleus (MPO). Magnocellular oxytocinergic (OXY) and vasopressinergic (AVP) neurons are shown in the upper left. Parvicellular corticotropin-releasing factor (CRF)-containing neurons are shown in the upper right. The descending division of the paraventricular hypothalamus includes preautonomic neurons in the dorsal cap (d. cap), the caudolateral perifornical (lpf) paraventricular hypothalamus, and the medial parvicellular (mpv) neurons residing between neuroendocrine cell groups. Heavier projections are indicated by heavier lines. Note the massive interconnectivity between the DMH and preoptic nuclei, and their more idiosyncratic relations to neuroendocrine and autonomic divisions of the paraventricular hypothalamus.

Inputs to the DMH include top-down inputs from cortico-striato-pallidal circuits, including visceromotor cortex (infralimbic), ventral subiculum, LS, MeA, as well as BNST; medial hypothalamic nuclei involved in basic behavioral functions such as reproduction, feeding, and defense; sensory nuclei, such as the parabrachial nucleus, relaying viscerosensory information, and the SFO, relaying humorosensory information; and substantial input from SCN and the subparaventricular zone. Hodological analysis strongly suggests that the HVPG integrates cognitive, systemic, and circadian inputs to coordinate neuroendocrine, autonomic, and some behavioral outflows.

The specific coordinating functions are uncharacterized, but good evidence indicates that the DMH drives the HPA axis and autonomic outflows. Disinhibiting the DMH using $GABA_A$ antagonists (or stimulating the DMH, using excitatory amino acids) increases sympathetic outflow and plasma catecholamines, heart rate, blood pressure, and blood flow to skeletal muscle while reducing blood flow to viscera. Blood pressure increases resulting from DMH stimulation can be blocked by inactivating the rostral ventrolateral medulla (RVLM), whereas tachycardia is likely relayed through descending projections to PAG or raphe pallidus. Lower doses of excitatory amino acids are required to obtain visceromotor outflow in the DMH compared to the PVN. Likewise, the blunting of autonomic and HPA axis outflows to psychogenic stress is greater after DMH than after PVN inactivation by muscimol. Neurons in the DMH expressing c-fos following footshock have been retrogradely labeled from the PVN and are glutamatergic. Muscimol microinfusion into the DMH reduces PVN activation to air-puff stress, but not to hemorrhage, although this should not be interpreted to mean that the DMH plays no role in systemic threats, as hemorrhage is a powerful stimulus.

Psychogenic Networks Driving the HPA Axis

Psychogenic stress is distinct from systemic stress; knife cuts of the ascending VNAB reduce PVN activity after nociception, but not after restraint stress. Top-down pathways to the PVN and surrounding networks consist of excitatory cortical, inhibitory striatal, and disinhibitory pallidal inputs operating both serially and in parallel. Medial prefrontal cortices (mPFCs) are the highest level of motor control over visceromotor outflow. Violations of cortical predictions or changes in the valence of experience alter prefrontal activity

and downstream HPA axis and autonomic outflows. Cognitive networks are sensitive to feedback from the visceromotor processes they control, and exhibit stress-dependent plasticity. Hypofunctional cortical states are associated with HPA axis dysregulation.

The mPFC shows substantial immediate early gene activation following stressors, high levels of GR staining, and ADX-reversible glutamate release following a stressor. Inputs to prefrontal cortex include ascending dopaminergic, noradrenergic, and serotonergic inputs, all of which are modulated by HPA axis function. Both dorsal and ventral regions of the mPFC may be involved in HPA axis outflow, but the ventral regions, including infralimbic (IL) cortex and ventral prelimbic cortex, contribute to HPA axis activation and project directly to brain stem preautonomic centers.

After receiving highly processed cortical input, the largely ipsilateral projections of the IL cortex to cerebral nuclei include the CeA, MeA, and LS in the striatum, BNST, and substantia innominata in the pallidum. In hypothalamic brain stem, the IL cortex projects to key nodes of the HVPG, including MPO, lateral preoptic (LPO), DMH, and perifornical LH, all areas projecting to the PVN, but the IL has negligible direct projections to the PVN. In lower brain stem there are moderate to light projections to autonomic areas, such as the parabrachial nucleus (PBN) and PAG, Barrington's nucleus, the DMV, NTS, nucleus ambiguous, and the RVLM. Like nearly every driver of the HPA axis, this also projects substantially to the PVT in the dorsal brain stem. Although some projections to brain stem visceromotor centers may be light, subcortical targets of the IL cortex also project to these centers. Thus, the IL cortex drives a cascade of direct excitatory projections, inhibitory striatal projections, and disinhibitory pallidal projections, layers of stimulation aimed at visceromotor processes.

Chemical, electrical, and lesioning manipulations of mPFC alter HPA axis and autonomic outflows. Corticosterone implants into dorsal mPFC inhibit HPA axis output. Infusions of GABA agonists and *N*-methyl-D-aspartate (NMDA) antagonists into mPFC exaggerate adrenocortical responses to cage-change stress. In contrast, electrical stimulation of IL cortex results in stimulation of the HPA axis, whereas lesions inhibit stress-induced HPA axis activation. It appears that dorsal mPFC tends to inhibit the HPA axis, whereas ventral activation stimulates. Issues of laterality exist, insofar as right hemispheric lesions blunt stress responses, whereas those in the left do not. Unfortunately, large lesions involving dorsal and ventral regions were used.

Subcortically, stimulation of LS or BNST, a pallidal gate of the CeA, also stimulates HPA axis outflow. The BNST also gates input from MeA, as MeA stimulation in anesthetized rats stimulates the HPA axis in intact animals, but not those bearing BNST lesions. The same lesions do not prevent ether-induced HPA responses. BNST and CeA CRF-containing neurons both converge on the PVT, which is involved in changed HPA axis output following chronic stress.

GCs in Central Motive States and Central Plasticity

HPA axis control cannot be viewed without regard for GC effects on brain outside of classical inhibitory feedback mechanisms. GCs also feed forward in brain, facilitating HPA axis outflow, and affect aversive and appetitive motive states, learning and memory, and gross neuronal remodeling, ultimately affecting HPA axis functioning.

Corticotropin-Releasing Factor

Central CRF systems are critical to aversive responding. Although ADX upregulates CRF in the PVN in a GC-reversible fashion, GC pellets implanted in dorsal CeA increase anxiety, stress-induced elevations in CRF mRNA, ACTH and GC secretion, colonic motility, and visceromotor responses to colorectal distension. Chronic stress or GC treatment also increases CRF mRNA in BNST. Chronic intracerebroventricular (icv) infusions of GCs enhance basal and stress-induced HPA outflow. Chronic elevations of GCs in brain upregulate central stress networks and increase HPA, autonomic, and behavioral outflows.

Depolarization of sheep PVN CRF neurons through stress, chemical, or electrical means increases venous cortisol and extracellular amygdalar CRF. Amygdalar CRF response to psychogenic stress exhibits a bimodal pattern: one immediate stressor-related peak, and a second peak resulting from GCs entering the brain. GCs infused intravenously causes one rise in amygdalar CRF following the venous cortisol signal. Repeated activation of amygdalar CRF by GCs causes potentiated acute CRF responses to novel stressors. Exposing sheep once daily to predator stress increases amygdalar CRF upon each exposure. When a novel forelimb shock is later applied, animals show facilitated amygdalar CRF responses, unless daily predator exposures were accompanied by cortisol synthesis inhibition, which prevents GC-mediated facilitation. GC exposure recruits both acute and chronic central stress networks.

Norepinephrine

The ascending dorsal NE system is sensitive to stress and central CRF. Cognitive stressors virtually always activate locus coeruleus (LC) NE neurons, in association

with increased NE release in LC targets, such as mPFC and BNST, arousal and anxiety, and HPA axis outflow. Depletion of LC NE blunts HPA axis outflow. CeA and BNST CRF neurons target LC dendrites, and CRF infusions into LC increase LC discharge rates, alter cortical electroencephalogram (EEG), and behavioral, autonomic, and HPA axis responses. CRF effects on LC are dose dependent and are blocked by CRF antagonists. Although tyrosine hydroxylase-immunopositive neurons in LC stain for GRs, the role of GCs in LC is unclear. ADX increases LC discharge rats in a GC-reversible fashion, but also causes metabolic disturbances that disinhibit brain stem TH networks. It is unclear if GC inhibition of LC is due to direct action or to peripheral metabolic corrections.

Dopamine

GCs and CRF target mesocorticolimbic DA systems are associated with multiple appetitive, aversive, and reinforcement mechanisms. GCs stimulate dopamine (DA) efflux in DAergic terminal fields, including nucleus accumbens, prefrontal cortex, BNST, and amygdala. Adrenalectomized rats given low levels of GCs show greatly reduced DA efflux compared to intact rats after tail pinch, cocaine, or morphine treatments. Rapid DA efflux occurs if sufficient GC concentrations are achieved. GR co-localize with D1 receptors in prefrontal pyramidal neurons, substantia nigra, and the ventral tegmental area (VTA). Stress or GCs enhances VTA discharge rates stimulated by excitatory ligands, a GR antagonist-reversible effect. GCs given alone have no effect, suggesting enhanced glutamatergic transmission is required. GC antagonists also block plastic processes in the VTA. Stress-related interactions between prefrontal glutamatergic neurons and VTA DAergic neurons play important roles in stress-related pathologies, including drug abuse. Central CRF systems also target the VTA, and CRF receptors co-localize with DAergic neurons. Like GCs, CRF potentiates excitatory signaling in VTA DAergic neurons, and contributes to neuronal sensitization. Footshock increases CRF and glutamate efflux in the VTA and DA efflux in terminal fields, and promotes reinstatement to drug seeking, a CRF antagonist-reversible effect.

Endogenous Opioid Peptides

Endogenous opioid peptides (EOPs) affect central stress networks state dependently. CRF and AVP both stimulate EOP secretion, and enkephalinergic neurons in the RVLM project to PVN and LC dendrites, overlapping with CRF-containing fibers. The PVN also receives enkephalinergic (ENK) innervation from the MPO and DMH. Stress increases ENK mRNA in the PVN, but downregulates it in the RVLM. Administration of opioid agonists and antagonists can cause both activation and inhibition of the HPA axis. In resting sheep, β-endorphin causes a naloxone-reversible increase in ACTH, whereas in chronically stressed animals it causes a naloxone-reversible inhibition of HPA output. Opioid-induced increases in ACTH are abolished by pretreatment with CRF antiserum. Thus, in a state of stress, opioids may act as a counterregulatory limb of the central stress response, whereas under basal conditions they are stimulatory, as if CRF and EOPs exist as opponent processes, a contention supported by the fact that opioids inhibit LC discharge rates. The effect is again state dependent, as opioid antagonists have no effect on basal LC discharge, but do increase LC discharge under stress, suggesting stress-induced recruitment of opioid signaling. Likewise, excess opioid signaling recruits CRF/NE systems, and dramatic increases in LC discharge rates occur when opioid-dependent animals are given opiate antagonists. Complex sequelae unfold in the HPA axis following intermittent opioid treatment and withdrawal, indicating initial recruitment of HPA axis facilitators leading to HPA axis hyperresponsiveness to stress, followed by recruitment of some mechanism that makes the HPA axis hyporesponsive to stress.

Memory

GCs influence memories of aversive and appetitive events, but the direction of the effect is a structure- and task-dependent event. In the hippocampus, correlates of learning exhibit an inverted-U-shaped dependency on GC concentrations, such that moderate levels promote correlates of learning, whereas absence of or excess GCs degrade them. On the other hand, ADX, GC synthesis inhibition, or selective blockade of GRs all impair memory for aversive tasks, such as inhibitory avoidance. ADX-induced deficits are reversible by GC administration. GCs are most constructive memorially when given immediately posttraining, whereas they impair retrieval when given prior to testing. One locus of positive GC effects is the basolateral amygdala (BLA). GC microinfusions into the BLA, but not its striatal target, the CeA, improve memory consolidation, whereas BLA lesions abolish facilitating effects of high levels of systemic steroid. Improved consolidation is partly mediated by NE, as blockade of β-adrenoreceptors within the BLA prevents the facilitating effects of high steroid levels. The source of NE may be medullary, as GR agonists-infused into NTS improve memory which is blocked by β-adrenoreceptor blockade in BLA. Posttraining CRF infused into the BLA also improves memory. As the BLA contains CRF receptors, but few CRF-immunopositive fibers, CRF probably diffuses into

the BLA from the CeA. Dopamine agonists given simultaneously into the BLA and the accumbens shell region also enhance memory for conditioned fear.

GC-mediated conditioned responses feed back on HPA axis and autonomic outflows. The CeA is essential for expressing behavioral and neuroendocrine measures of conditioned fear, whereas lesions of the LH prevent autonomic but not behavioral components, and lesions of the PAG interfere with behavioral but not autonomic responses. For the HPA axis, lesions of the CeA disrupt conditioned responses to both discrete (e.g., light, tone) and contextual (e.g., room cues) stimuli, whereas lesions of the BNST only disrupt responses to contextual, but not discrete, stimuli, suggesting alternative routes to the HPA axis output, depending on the modality of association.

Neuronal Remodeling

GCs remodel gross brain structure. Acute GC elevations rapidly induce spinogenesis in hippocampal slices in the absence of protein synthesis. Chronic stress, GCs, and ADX inhibit neurogenesis in the dentate gyrus, suggesting that moderate steroid level is permissive to normal structure. Chronic stress reduces dendritic arborization in CA3 fields and total hippocampal volume, effects reversed by stress-free periods. Similar events occur in the mPFC. In contrast, a regimen of repeated immobilization stress increases dendritic length and spinogenesis in the BLA. Acute stress only increases spinogenesis, which increases, rather than decreases, during stress-free periods. The BNST shows some similar effects. Structure-dependent effects of stress on neuronal remodeling are consistent with the direction of correlates of learning in these structures. GCs favor memorial and structural modifications related to emotional processing at the expense of cognitive executive processing.

Conclusion

Although the HPA axis is a relatively simple final common output pathway, its complexity increases when properly viewed in terms of its integration of diverse central drivers and its feedback and feedforward effects on the brain. Diverse central drivers include two distinct circadian oscillators controlling daily cycles of energy balance, bottom-up inputs controlling systemic homeostatic pressures, such as the melanocortin system and the ascending catecholaminergic system, and top-down cortico-striato-pallido-hypothalamic circuits controlling psychogenic responses, and a putative intrahypothalamic pattern generator that coordinates HPA axis outflow with behavioral and autonomic outflows. In addition, a system of circumventricular organs and perhaps signals from the hepatic vagus allow signaling between the immune system and the HPA axis. GCs enter the brain to have both immediate, membrane-mediated effects on various transmitter and neuromodulatory systems and more delayed, genomically mediated effects. The net effect of GC feedback in brain depends on the parameters of stimulation, such as whether a stressor is acute or chronic, with some effects being time or state dependent. GC actions in brain can lead to both inhibition and facilitation of further HPA axis outflow, affecting other motor outflows and influencing mood, incentive, memorial, and structural systems relevant to adaptation. The HPA axis defends the body against systemic derangements, while sculpting and mounting brain networks to anticipate and avoid further challenge.

See also: Apoptosis in Neurodegenerative Disease; Corticomotoneuronal System; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Energy Homeostasis: Hypothalamic Development; Hypothalamo–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH; Melanocortins: Brain Effects; Norepinephrine: CNS Pathways and Neurophysiology; Proopiomelanocortin; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress, the HPA Axis and Depressive Illness.

Further Reading

Antoni FA (1993) Vasopressinergic control of pituitary adrenocorticotropin comes of age. *Frontiers in Neuroendocrinology* 14: 76–122.

Buijs RM and Van Eden CG (2000) The integration of stress by the hypothalamus, amygdala and prefrontal cortex: Balance between the autonomic nervous system and the neuroendocrine system. *Progress in Brain Research* 126: 117–132.

Cone RD (2005) Anatomy and regulation of the central melanocortin system. *Nature Neuroscience* 8(5): 571–578.

Cook CJ (2002) Glucocorticoid feedback increases the sensitivity of the limbic system to stress. *Physiology & Behavior* 75: 455–464.

Dallman MF, Akana SF, Cascio CS, et al. (1987) Regulation of ACTH secretion: Variations on a theme of B. In: Clark JH (ed.) *Recent Progress in Hormone Research*, vol. 43, pp. 113–173. Orlando: Academic Press.

Dallman MF, Akana SF, Jacobson L, et al. (1987) Characterization of corticosterone feedback regulation of ACTH secretion. *Annals of the New York Academy of Sciences* 512: 402–414.

Dallman MF, Pecoraro N, Akana SF, et al. (2003) Chronic stress and obesity: A new view of 'comfort food.' *Proceedings of the National Academy of Sciences of the United States of America* 100(20): 11696–11701.

De Kloet ER, Vreugdenhil E, Oitzl MS, et al. (1998) Brain corticosteroid receptor balance in health and disease. *Endocrine Reviews* 19(3): 269–301.

Erickson K, Drevets WC, and Schulkin J (2003) Glucocorticoid regulation of diverse cognitive functions in normal and pathological emotional states. *Neuroscience & Biobehavioral Reviews* 27: 233–246.

Lamont EW, Robinson B, Stewart J, et al. (2005) The central and basolateral nuclei of the amygdala exhibit opposite diurnal rhythms of expression of the clock protein Period2. *Proceedings of the National Academy of Sciences of the United States of America* 102(11): 4180–4184.

LeDoux J (2003) The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology* 23(4–5): 727–738.

McEwen BS and Magarinos AM (2001) Stress and hippocampal plasticity: Implications for the pathophysiology of affective disorders. *Human Psychopharmacology* 16(S1): S7–S19.

Mistlberger R (1994) Circadian food-anticipatory activity: Formal models and physiological mechanisms. *Neuroscience & Biobehavioral Reviews* 18(2): 171–195.

Pacak K and Palkovits M (2001) Stressor specificity of central neuroendocrine responses: Implications for stress-related disorders. *Endocrine Reviews* 22(4): 502–548.

Perreau-Lenz S, Pevet P, Buijs RM, et al. (2004) The biological clock: The bodyguard of temporal homeostasis. *Chronobiology International* 21(1): 1–25.

Piazza PV and LeMoal M (1997) Glucocorticoids as a biological substrate of reward: Physiological and pathophysiological implications. *Brain Research Reviews* 25: 359–372.

Przekop F, Mateusiak K, Stupnicka E, et al. (1990) Suppressive effect of beta-endorphin and naloxone on the secretion of cortisol under stress conditions in sheep. *Experimental and Clinical Endocrinology* 95(2): 210–216.

Ritter S, Watts AG, Dinh TT, et al. (2003) Immunotoxin lesion of hypothalamically projecting norepinephrine and epinephrine neurons differentially affects circadian and stressor-stimulated corticosterone secretion. *Endocrinology* 144(4): 1357–1367.

Rivest S (2001) How circulating cytokines trigger the neural circuits that control the hypothalamic–pituitary–adrenal axis. *Psychoneuroendocrinology* 26(8): 761–788.

Roozendaal B (2003) Systems mediating acute glucocorticoid effects on memory consolidation and retrieval. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 27(8): 1213–1223.

Saal D, Dong Y, Bonci A, et al. (2003) Drugs of abuse and stress trigger a common synaptic adaptation in dopamine neurons. *Neuron* 37: 577–582.

Saper CB, Lu J, Chou TC, et al. (2005) The hypothalamic integrator for circadian rhythms. *Trends in Neuroscience* 28(3): 152–157.

Sullivan RM and Gratton A (2002) Prefrontal cortical regulation of hypothalamic–pituitary–adrenal function in the rat and implications for psychopathology: Side matters. *Psychoneuroendocrinology* 27: 99–114.

Swanson LW (2000) Cerebral hemisphere regulation of motivated behavior. *Brain Research* 886(1–2): 113–164.

Thompson RH and Swanson LW (2003) Structural characterization of a hypothalamic visceromotor pattern generator network. *Brain Research: Brain Research Reviews* 41(2–3): 153–202.

Valentino RJ and Van Bockstaele EJ (2001) Opposing regulation of the locus coeruleus by corticotropin-releasing factor and opioids. *Psychopharmacology* 158: 331–342.

Vyas A, Mitra R, Shankaranarayana Rao BS, et al. (2002) Chronic stress induces contrasting patterns of dendritic remodeling in hippocampal and amygdaloid neurons. *Journal of Neuroscience* 22(15): 6810–6818.

Yoshihara T, Honma S, and Honma K (1996) Prefeeding release of paraventricular neuropeptide Y is mediated by ascending noradrenergic neurons in rats. *American Journal of Physiology* 270 (4, part 1): E596–E600.

Hypothalamic–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH

R M Lechan, Tufts Medical Center, Boston, MA, USA
A Hollenberg, Beth Israel Deaconess Medical Center, Boston, MA, USA
C Fekete, Hungarian Academy of Sciences, Budapest, Hungary; and Tufts Medical Center, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The maintenance of thyroid function is dependent on a complex interplay between the hypothalamus, anterior pituitary, and thyroid gland, as well as other factors that influence the function of these organ systems. This function is critical to achieve a constant level of free thyroid hormone in the bloodstream to support peripheral tissues by affecting protein synthesis and/or altering the metabolic activity of cells, as well as for brain development during fetal growth and early infancy. The hypothalamic peptide primarily responsible for this action by stimulating the biosynthesis and release of thyrotropin (TSH) from the anterior pituitary gland is thyrotropin-releasing hormone (TRH). TSH then stimulates the release, from the thyroid gland into the circulation, of the thyroid hormones, thyroxine (T4) and tri-iodothyronine (T3), which feed back on the hypothalamic TRH neurons and anterior pituitary thyrotropes, completing what is recognized as a classic example of a negative feedback loop system (Figure 1).

Biosynthesis and Processing of TRH

TRH (pyroGlu–His–ProNH2, MW 362) arises by a messenger RNA (mRNA)-directed ribosomal mechanism and posttranslational processing of a larger precursor prohormone. The complete sequence of the TRH precursor, deduced from its complementary DNA, has been elucidated for frog, rat, mouse, and human and is illustrated schematically in Figure 2. Common to each prohormone are multiple copies of a progenitor sequence for TRH, Gln–His–Pro–Gly, flanked on either side by paired basic amino acids, Lys–Arg or Arg–Arg, that are processing signals for carboxypeptidase B-like enzymes. The C-terminal glycine functions as a substrate for α-amidating enzymes that convert TRH-Gly to TRH, and the N-terminal glutaminyl is modified by glutaminyl cyclase to result in the fully mature and biologically active TRH. Presumably, the repeating progenitor sequences of TRH in the prohormone are of importance to magnify the response to signals that trigger TRH secretion. Pre-proTRH$_{141–149}$ (referred to as human thyrotropin-releasing hormone-associated peptide 3 (hTAP-3)) is also present in human serum and corresponds in position to the intervening sequence preproTRH$_{160–169}$ in the rat prohormone.

In addition to the sequences that give rise to TRH, the TRH precursor contains several cryptic peptides between the TRH progenitor sequences, and N- and C-terminal flanking peptides that may be biologically active. At least one peptide, preproTRH$_{160–169}$, a decapeptide that follows the third TRH sequence in the rat prohormone, has little activity on anterior pituitary secretion by itself, but in the presence of TRH, it potentiates the secretion of TSH by increasing TSHβ mRNA in thyrotropes. As receptors for preproTRH$_{160–169}$ have been identified only in folliculostellate cells, specialized glial cells of anterior pituitary origin, the effect of preproTRH$_{160–169}$ to potentiate the action of TRH on TSH secretion is presumed to be indirect via paracrine interactions between the processes of folliculostellate cells and anterior pituitary thyrotropes.

PreproTRH$_{178–199}$ separates the fourth and fifth progenitor TRH sequences and is also released from the rat median eminence, but as yet, no receptors for this peptide have been identified. Claims for biologic activity of preproTRH$_{178–199}$ were based on its ability to act as a corticotropin-inhibiting factor, blocking both the release of adrenocorticotropic hormone and the transcription of POMC mRNA, but the data are controversial. PreproTRH$_{178–199}$ may also exert other pituitary actions, including inhibition of growth hormone and prolactin secretion.

In the hypothalamus, preproTRH is fully processed to yield all five copies of TRH. Two enzymes that are critical for the processing of proTRH and present in most TRH neurons in the hypothalamus are the proconvertase enzymes, PC1/3 and PC2. Processing begins in the trans-Golgi network and continues in secretory granules. Both PC1/3 and PC2 are inversely regulated by thyroid hormone in hypophysiotropic TRH neurons, thereby contributing to thyroid hormone action on the regulation of TRH.

TRH Gene and Promoter Regulation

The murine, rat, and human TRH genomic structures are identical, each containing three exons and two introns (Figure 3(a)); exon 1 encodes the 5′ untranslated region, whereas exons 2 and 3 contain the

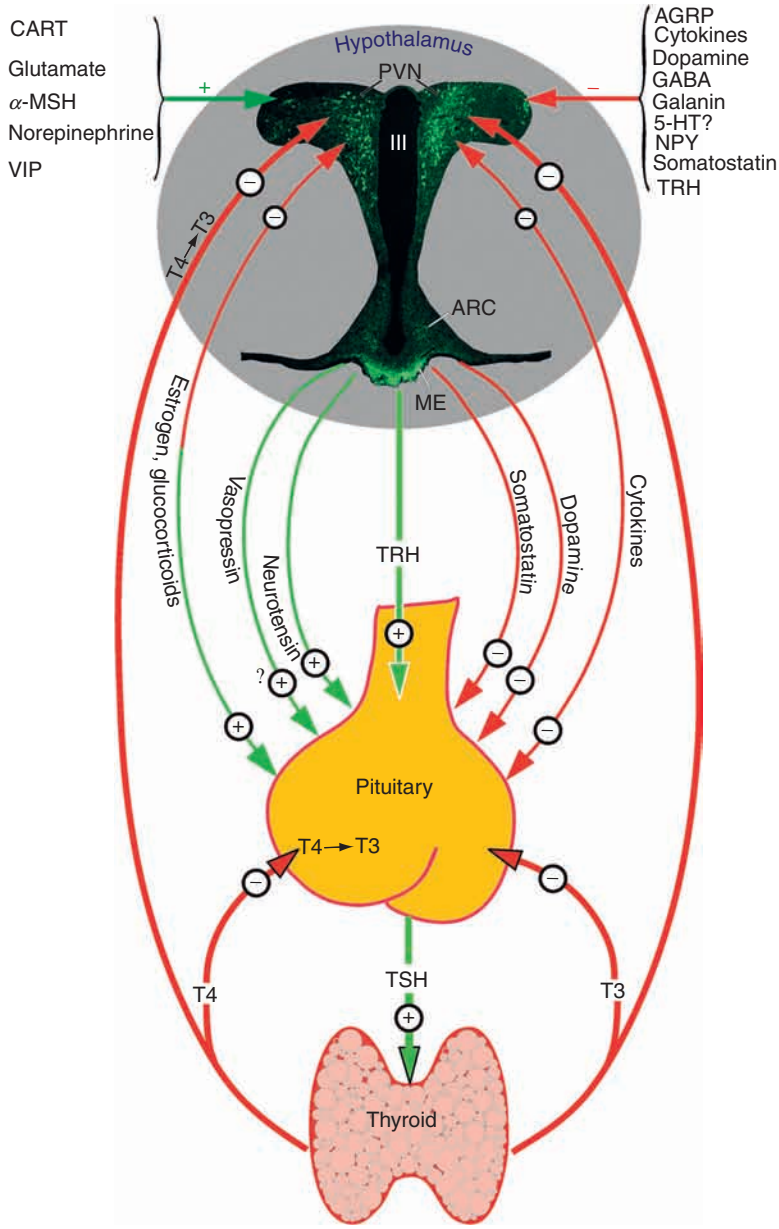

Figure 1 Neuroregulatory control systems involved in the secretion of thyroid hormone. Bold lines denote the negative feedback loop of thyroid hormone on thyrotropin-releasing hormone (TRH) secretion from the hypothalamus and thyroid-stimulating hormone (TSH) secretion from the anterior pituitary. Both the hypothalamic TRH neurons and anterior pituitary thyrotrops are impinged on by numerous other potential regulatory influences that are activated under specific physiological or pathological conditions. CART, cocaine- and amphetamine-regulated transcript; α-MSH, α-melanocortin-stimulating hormone; VIP, vasoactive intestinal polypeptide; AGRP, agouti-related peptide; GABA, γ-aminobutyric acid; 5-HT, 5-hydroxytryptamine; NPY, neuropeptide Y; PVN, paraventricular nucleus; ARC, arcuate nucleus; ME, median eminence; T4, thyroxine; T3, tri-iodothyronine.

sequence that encodes the preproTRH message. The promoter region of the TRH gene is located immediately 5′ to exon 1 and is presumably the major locus where regulation of the gene takes place. *In vivo* studies in transgenic mice using TRH promoter fragments linked to a luciferase reporter gene suggest that sequences within the proximal promoter, as well as the first exon, are critical for TRH expression in the hypothalamus. It remains unclear, however, which regions of the TRH gene allow for its expression in individual hypothalamic nuclei or other tissues, such as the pituitary, pancreas, or heart.

The TRH promoter region is heralded by the presence of a conserved TATAA box, allowing for transcriptional initiation by RNA polymerase II. Just proximal to the TATAA box is an important regulatory

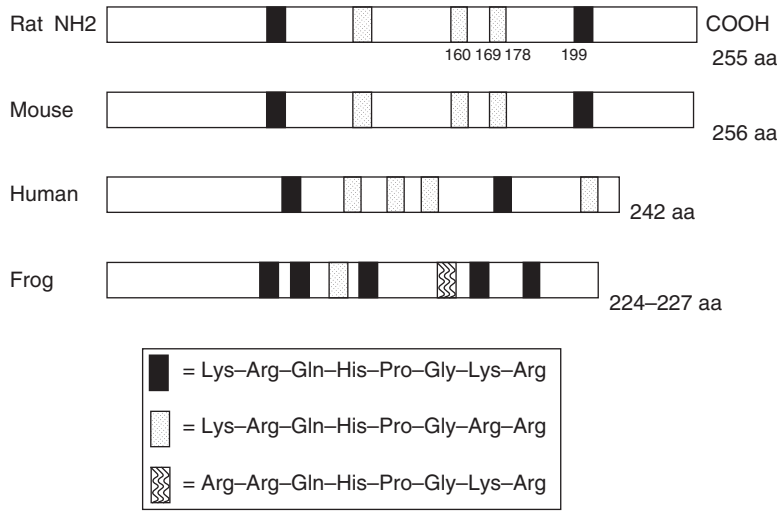

Figure 2 Schematic representation of rat, mouse, human, and frog prepro thyrotropin-releasing hormone (TRH). Each sequence contains multiple copies of a TRH progenitor sequence (Gln–His–Pro–Gly) flanked by paired basic residues. The cryptic peptides, preproTRH$_{160–169}$ and preproTRH$_{178–199}$, are located between the third and fourth progenitor TRH sequence, and the fourth and fifth progenitor TRH sequence in rat preproTRH, respectively. aa, amino acid.

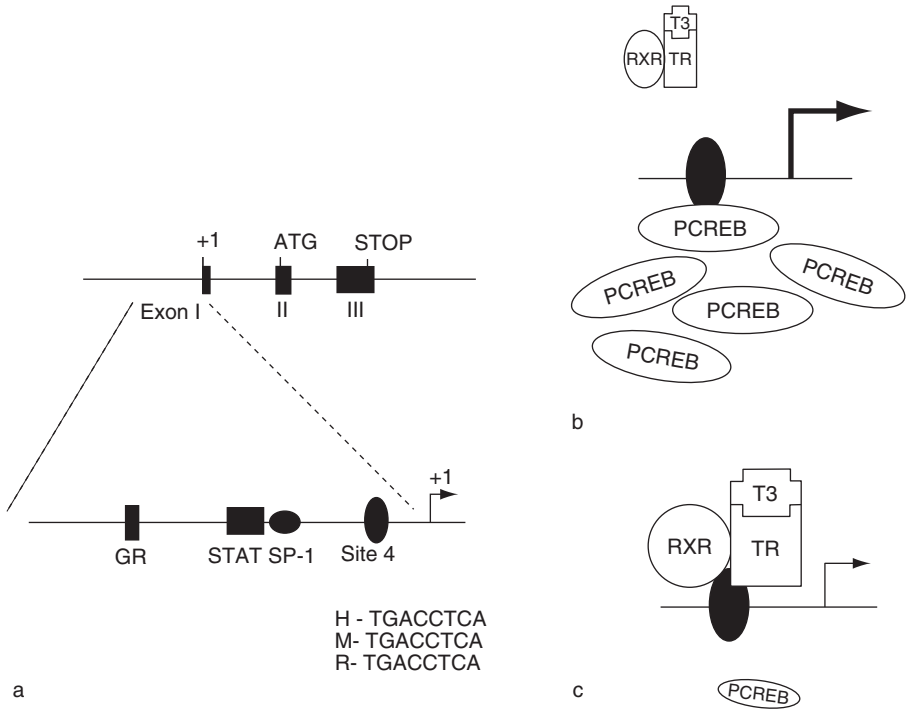

Figure 3 (a) Genomic and promoter structure of thyrotropin-releasing hormone (TRH). The murine, rat, and human TRH genes are composed of three exons and two introns. The coding sequence for the precursor protein is present on exons II and III. As depicted, the TRH promoter region precedes the transcription start site in exon I. The proximal 250-bp sequences of the human (H), mouse (M), and rat (R) promoters are similar and share the indicated transcription factor binding sites. The locations of the cyclic adenosine monophosphate-responsive element binding protein (CREB) binding site (Site 4) across animal species are indicated. The locations of the signal transducer of activated transcription (STAT) element and glucocorticoid receptor (GR) response element are also shown. (b, c) Hypothesized schematic representation of the interaction between phosphorylated CREB (PCREB) and the thyroid hormone receptor at Site 4. (b) In the presence of abundant PCREB, there may be less availability for binding of the thyroid hormone receptor/tri-iodothyronine (T3) complex; hence, an increase in TRH gene transcription. When PCREB concentrations fall as shown in (c), increased binding of the thyroid hormone receptor/T3 complex reduces TRH gene transcription as illustrated by the thicker arrow. RXR, retinoic acid X receptor; TR, thyroid hormone receptor. From Lechan RM and Fekete C (2006) The TRH neuron: A hypothalamic integrator of energy metabolism. *Progress in Brain Research* 153: 209–235.

element termed Site 4 (TGACCTCA) that is conserved in the mouse, rat, and human promoters (**Figure 3(a)**). This sequence is important for basal expression of the TRH promoter in many different mammalian cell lines when they are assessed by means of transfection studies. In addition, Site 4 serves as a multifunctional binding site for the transcription factor cyclic adenosine monophosphate (cAMP) regulatory response element binding protein (CREB) and for thyroid hormone receptors (TRs), including TR homodimers, as well as heterodimers of TR and the retinoic acid X receptor (RXR). It appears that Site 4 cooperates with a second, weak TR binding site, 11 bp 3′ to Site 4, that binds both TR homodimers and TR/RXR heterodimers. Altogether Site 4 appears to be a critical sequence within the TRH promoter, is preserved across species, and controls both basal and regulated expression of the TRH gene.

The importance of Site 4 in the regulation of TRH expression was first demonstrated in functional studies in mammalian cell lines. These studies showed that it was required for downregulation of TRH promoter constructs by T3 in the presence of the TR. T3 mediates its effects on gene expression via the three TR isoforms (TRα_1, TRβ_1, and TRβ_2). TRα_1 is expressed in many tissues but plays little role in negative regulation by T3. The TRβ isoforms are products of a single gene and differ only in their N-terminal regions. TRβ_1 mediates the effect of T3 in the inner ear, liver, and kidney, and recent genetic studies have confirmed that TRβ_2 is critical for negative regulation of the TRH and the TSHα and β subunit genes by T3.

The molecular mechanism governing negative regulation of the TRH by T3 has not been firmly established. Site 4 and its surrounding region can bind TRβ_2, which appears to be necessary for negative regulation. However, it is not clear which cofactors are involved. On genes positively regulated by T3, the TR isoforms are bound constitutively to thyroid hormone response elements. In the absence of T3, they recruit nuclear corepressors that repress transcription of target genes. The presence of T3 causes the release of corepressors and the recruitment of coactivators such as SRC-1, which lead to transcriptional activation. The role of corepressors and coactivators in the negative regulation of the TRH gene remains less clear, given that the recruitment of coactivators by TRβ_2 in the presence of T3 would be expected to activate rather than repress TRH expression. Surprisingly though, they also appear to be important as mutations in TRs in knock-in mice that prevent binding of coactivators without altering DNA binding are unable to mediate negative regulation *in vivo*. Furthermore, mice that lack SRC-1 are also unable to downregulate TSH subunit gene regulation. Thus, coactivators play a role in negative regulation *in vivo*, though the exact mechanism remains to be determined.

Site 4 can also be bound by CREB. CREB is bound constitutively to target promoters, usually as a homodimer or as a heterodimer with other members of the ATF-1 family. When phosphorylated by protein kinase A after activation of the cAMP signaling cascade by activators such as forskolin, PCREB recruits the coactivator CREB binding protein, thereby allowing for transcriptional activation of the TRH gene via modifications of the histone structure surrounding the gene. Mutation or deletion of Site 4 prevents activation of TRH promoter constructs by cAMP. The phosphorylation of CREB, therefore, is likely a key mechanism for positive regulation of TRH gene expression and, by competing with PCREB for Site 4, TRβ_2, could prevent activation of the TRH promoter (**Figures 3 and 3(c)**). Binding of PCREB to Site 4 also has an important role in mediating the effects of melanocortin signaling on the TRH gene (see the section titled 'Regulation of hypophysiotropic TRH by fasting and infection (nonthyroidal illness syndrome)').

Activation of the TRH gene also occurs via a canonical signal transducer of activated transcription (STAT) binding site in the murine and rat promoter 5′ to Site 4. A similar site in the human promoter located between −150 and −125 has also been identified, suggesting the importance of this sequence across animal species (**Figure 3(a)**). Tyrosine phosphorylation allows phospho-STAT to dimerize and enter the nucleus to bind to its target regulatory elements, such as the one found in the TRH promoter. STAT activates transcription by recruiting coactivators such as CREB binding protein. The leptin receptor activates transcription via the JAK-STAT signaling pathway and is able to significantly enhance TRH promoter activity in transfection experiments in mammalian cells. Furthermore, the TRH promoter is able to recruit STAT3 *in vivo* in chromatin immunoprecipitation experiments. The presence of such a STAT site also suggests that the TRH promoter is poised to respond to other cytokine signaling pathways that could engage TRH neurons in the PVH or elsewhere.

The TRH promoter also contains binding sites for the glucocorticoid receptor and SP-1. The glucocorticoid receptor response element is located just upstream of the STAT element and appears to be functional in the pituitary, where glucocorticoids enhance TRH gene expression (**Figure3(a)**). However, glucocorticoids appear to be inhibitory to TRH gene expression in the paraventricular nucleus (PVN). There are multiple potential SP-1 sites within the promoter, but one located adjacent to the STAT site is required for full induction by leptin.

Anatomical Organization and Central Mediators of the Hypothalamic TRH Tuberoinfundibular System

The origin of TRH-containing neurons that regulate anterior pituitary TSH secretion is the hypothalamic PVN, a triangular, midline nuclear group symmetrically located on either side of the dorsal portion of the third ventricle (Figures 4(a)–4(c)). The PVN is composed of two major parts: magnocellular neurons, located in more lateral portions of the nucleus, and parvocellular neurons, located in more medial portions of the nucleus. The magnocellular component gives rise to the hypothalamic neurohypophysial tract that carries vasopressin and oxytocin to the posterior pituitary. The parvocellular component has a number of subcompartments, including anterior, medial, periventricular, ventral, dorsal, and lateral parvocellular subdivisions. TRH-synthesizing neurons are found primarily in the anterior, medial, and periventricular parvocellular subdivisions and to a lesser extent in the dorsal and lateral subnuclei (Figures 4(d) and 4(e)). However, only TRH neurons in medial and periventricular parvocellular subdivisions project to the median eminence (and are often referred to as hypophysiotropic neurons as they are involved in regulation of the anterior pituitary), and thereby they are functionally distinct from the TRH neurons in the anterior parvocellular subdivision. Functional heterogeneity between the two major groups of TRH neurons in the PVN is further indicated by the co-expression of cocaine- and amphetamine-regulated transcript (CART) in medial/periventricular parvocellular TRH neurons but not in anterior parvocellular TRH neurons.

Axons from hypophysiotropic TRH neurons in the PVN project to the median eminence, one of several so-called 'circumventricular organs' in the mammalian brain that lie outside the blood–brain barrier. Here, they terminate in the external zone of the median eminence (Figure 4(f)), where they have access to the fenestrated capillaries of the portal system. Some fibers also descend in the pituitary stalk to terminate in the posterior pituitary and may provide an alternative pathway by which TRH can regulate anterior pituitary secretion through the vascular channels (short portal veins) that connect the posterior pituitary to the anterior pituitary.

By residing in the PVN, TRH-producing neurons are situated in a region highly enriched by afferent inputs from other regions in the brain. These inputs are important to establish the set point for feedback regulation of hypophysiotropic TRH by thyroid

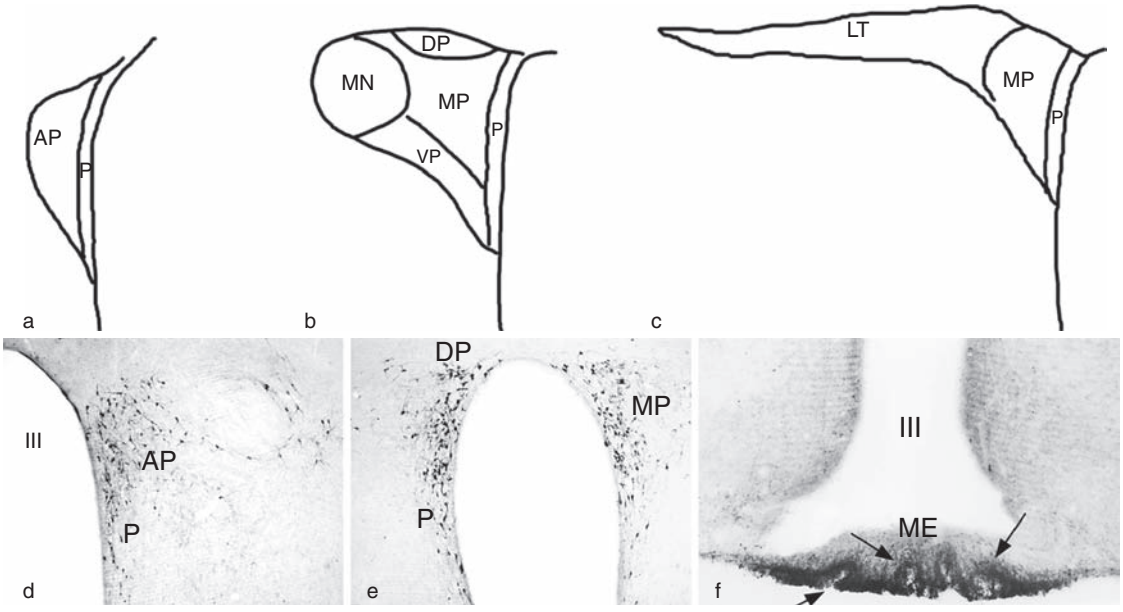

Figure 4 (a–c) Location of the major subdivisions in the rat paraventricular nucleus (PVN). AP, anterior parvocellular subdivision; DP, dorsal parvocellular subdivision; LT, lateral parvocellular subdivision; MN, magnocellular division; MP, medial parvocellular subdivision; P, periventricular parvocellular subdivision; VP, ventral parvocellular subdivision. Thyrotropin-releasing hormone (TRH) neurons are located in the anterior (d), medial, and periventricular (e) parvocellular subdivisions of the PVN, but only those in the medial and periventricular parvocellular subdivisions terminate in (f) the external zone of the median eminence (ME). Note close contact of TRH-containing axon terminals with capillaries of the portal vascular system (arrows). III, third ventricle. (a–c) Adapted from Sarkar S, Wittmann G, Fekete C, and Lechan RM (2004) Central administration of cocaine- and amphetamine-regulated transcript increases phosphorylation of cAMP response element binding protein in corticotropin-releasing hormone-producing neurons but not in prothyrotropin-releasing hormone-producing neurons in the hypothalamic paraventricular nucleus. *Brain Research* 999: 181–192.

hormone under basal conditions and during times when it is necessary to increase or decrease circulating levels of thyroid hormone (described in the section titled 'Feedback regulation of hypophysiotropic TRH by thyroid hormone'). Central modulation of hypophysiotropic TRH neurons through afferent projections may also occur in the external zone of the median eminence that could influence tuberoinfundibular TRH release by axo-axonal associations, but true axo-axonic synapses have only rarely been recognized in the mammalian median eminence.

While several regions of the brain project to the PVN, at least three regions provide the major input to hypophysiotropic TRH neurons, A1,2/C1–3 catecholamine neurons in the medulla of the brain stem, the hypothalamic arcuate nucleus, and the hypothalamic dorsomedial nucleus (Figure 5(a)). Axon terminals containing catecholamines give one of the most conspicuous innervations to TRH neurons in the PVN (Figure 5(c)) arising from catecholamine cell groups in the medulla. These axons establish mostly asymmetric synapses with both perikarya and dendrites of TRH neurons, suggesting an excitatory function. Noradrenergic axon terminals have also been identified in close apposition to TRH axons in the external layer of the rat median eminence, indicating that catecholamine inputs may influence TRH hypophysiotropic neurons, not only through direct effects on their perikarya and dendrites, but also on their axon terminals. The trajectory of catecholamine-containing axons from the brain stem is to course rostrally in the dorsal and ventral noradrenergic bundles, ultimately to enter the median forebrain bundle in the hypothalamus before terminating in the PVN (Figure 5(b)). Transection of the noradrenergic bundle at the level of the midbrain results in marked depletion of the catecholamine input to the PVN, in keeping with its origin from the brain stem. Approximately

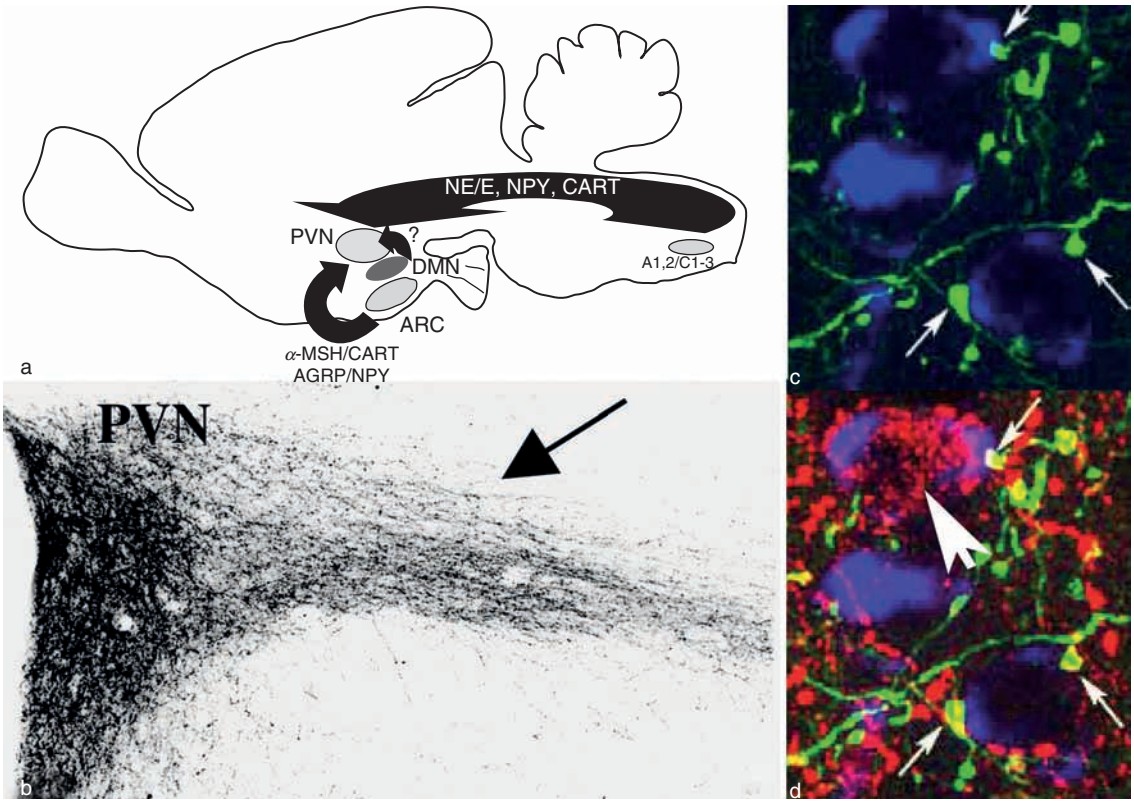

Figure 5 (a) Sagittal drawing of the rat brain showing three major sources of innervation of thyrotropin-releasing hormone (TRH) neurons in the hypothalamic paraventricular nucleus (PVN): the brain stem A1,2/C1–3 cell groups that carry the catecholamines NE and E, as well as neuropeptide Y (NPY) and cocaine- and amphetamine-regulated transcript (CART), to the PVN; the hypothalamic arcuate nucleus (ARC) that carries the neuropeptides NPY, agouti-related peptide (AGRP), α-melanocortin-stimulating hormone (α-MSH), and CART to the PVN; and the dorsomedial nucleus (DMN), for which the neurotransmitter mediators are unknown. (b) Neurons arising in each of the three regions innervate the PVN by traversing or joining fibers in the medial forebrain bundle (arrow). (c) Arrows denote catecholamine-containing fibers (green) innervating TRH neurons (blue) in the PVN. (d) CART-containing fibers (red) also innervate TRH neurons in the PVN. Some of these fibers co-contain catecholamines and CART and appear yellow due to color mixing. Note co-expression of CART immunoreactivity in the cytoplasm of one of the TRH neurons (arrowhead). Modified from Lechan RM and Fekete C (2006) The TRH neuron: A hypothalamic integrator of energy metabolism. *Progress in Brain Research* 153: 209–235.

20% of the neuropeptide Y (NPY) and more than 60% of the CART input to TRH neurons (Figure 5(d)) in the PVN also derive from catecholamine-producing brain stem neurons.

At least four different peptides of arcuate nucleus origin, α-melanocortin-stimulating hormone (α-MSH), CART, NPY, and agouti-related peptide (AGRP), are contained in axon terminals in synaptic contact with TRH hypophysiotropic neurons (Figure 5(a)). Neurons of the arcuato-PVN pathway project their axons primarily ipsilateral to the PVN and, like the catecholamine axon trajectory from the brain stem, course in the medial forebrain bundle prior to entering the PVN. In the rat, α-MSH and CART arise from the same neurons located in lateral portions of the arcuate nucleus, whereas NPY and AGRP are co-contained in a distinct population located in more medial portions of the nucleus, although both neuronal populations can have similar terminal fields on TRH neurons in the PVN. A similar organization is also present in the human brain, suggesting the evolutionary importance of this neuroregulatory control system. Indeed, in the rat and mouse, pharmacologic ablation of the arcuate nucleus not only results in profound reduction in the number of axons contacting TRH neurons in the PVN that contain these neuropeptides but also reduces circulating levels of TSH and thyroid hormone. Similar to the catecholamine innervation, asymmetric specializations between axons containing α-MSH/CART and TRH neurons in the PVN are observed, suggesting an excitatory function, while axon terminals containing NPY and AGRP establish primarily symmetric synaptic specializations, suggesting an inhibitory function. The latter associations are sometimes so dense as to completely envelop the perikarya and first-order dendrites of TRH neurons.

While the arcuate nucleus contributes 100% of the AGRP and α-MSH innervation to hypophysiotropic TRH neurons and approximately 80% of the NPY innervation, it contributes only less than 34% of the CART innervation. Brain stem catecholaminergic neurons provide the remaining 20% of the NPY innervation, particularly to the distal dendrites of TRH neurons, and approximately 66% of the remaining CART innervation originates from the brain stem, primarily adrenergic C1–3 neurons in the medulla.

The hypothalamic dorsomedial nucleus also gives rise to a major projection pathway to the PVN, coursing in the medial forebrain bundle before entering the caudal pole of the PVN. Practically all TRH neurons in the PVN are innervated by the dorsomedial nucleus arising from neurons in the dorsal and ventral subdivisions. While the chemical mediators remain unknown, evidence that the dorsomedial nucleus gives rise to a prominent galaninergic innervation of the PVN and that galanin-containing axon terminals densely innervate TRH neurons in the PVN provides circumstantial evidence for a role of this neuropeptide in mediating the actions of the dorsomedial nucleus on hypophysiotropic TRH. The dorsomedial nucleus may also contribute to the γ-aminobutyric acid and glutamate innervation of hypophysiotropic TRH neurons. As retrogradely transported marker substances injected into the PVN concentrate in neurons in the dorsomedial nucleus that are densely innervated by axon terminals containing α-MSH, it is conceivable that the arcuate nucleus regulates TRH neurons in the PVN by both direct and indirect pathways, using the dorsomedial nucleus as a relay center ultimately to affect hypophysiotropic TRH neurons in the PVN.

Feedback Regulation of Hypophysiotropic TRH by Thyroid Hormone

The anatomical specificity of the TRH neuronal population in the medial and periventricular parvocellular subdivisions of the PVN is associated with a functional specificity of these cells with respect to feedback regulation by thyroid hormone. When circulating levels of thyroid hormone fall below normal values (hypothyroidism), perikaryal size and the content of proTRH and proTRH mRNA increase in these neurons but remain unchanged in other TRH neurons in the forebrain. This increase is accompanied by a decline in the content of TRH in the median eminence due to an increase in the secretion of TRH into the portal blood for conveyance to the anterior pituitary. Conversely, increased circulating levels of T4 cause marked suppression of proTRH mRNA in the PVN and a reduction in the secretion of TRH into the portal plexus, establishing an inverse relationship between thyroid hormone and the biosynthesis and secretion of hypophysiotropic TRH.

The increase or decrease in the amount of hypophysiotropic TRH secreted into the portal system for conveyance to the anterior pituitary as dictated by variations in circulating levels of thyroid hormones, is important to establish the set point for feedback regulation by thyroid hormone on anterior pituitary TSH secretion. When portal blood TRH concentrations are low, TSH can be suppressed by less T4 circulating in the bloodstream (reduced set point), whereas high portal blood TRH concentrations raise the set point for feedback regulation by thyroid hormone. Thus, thyroidectomized animals with lesions of the hypothalamus that include the PVN can suppress TSH with smaller doses of T4 than can animals

with an intact hypothalamus, whereas chronic intrathecal infusion of TRH can elevate TSH and circulating free and total thyroid hormones.

The selectivity of TRH neurons in the medial and periventricular parvocellular subdivisions of the PVN to negative feedback regulation by thyroid hormone is an intrinsic property of these cells, mediated by the presence of thyroid hormone receptors in these neurons and the high vascular flow to the PVN. Stereotaxic implantation of microcrystals of T3 adjacent to one side of the PVN in hypothyroid animals, therefore, leads to marked inhibition of proTRH mRNA only on that side, but not on the opposite side.

The source of nuclear T3 responsible for feedback regulation of TRH neurons in the PVN differs from the source of T3 in other regions of the CNS, such as the cerebral cortex and anterior pituitary, where the majority of T3 arises from the intracellular monodeiodination of T4 to T3 by type 2 iodothyronine 5′-monodeiodinase (D2). This is because the PVN contains little, if any, D2 activity or D2 mRNA, indicating that cells in this region are not capable of intracellular conversion of T4 to biologically active T3. Thus, hypophysiotropic TRH neurons in the PVN must receive T3 from an exogenous source, either directly from the bloodstream, the cerebrospinal fluid (CSF), or adjacent glia or by transneuronal transport. The possibility that feedback regulation by thyroid hormone on TRH neurons in the PVN is mediated exclusively by circulating levels of T3 alone is unlikely since the systemic infusion of graded doses of T3 to hypothyroid animals that restore plasma levels of T3 to normal does not suppress proTRH mRNA levels in the PVN to euthyroid levels in the absence of T4. Only by increasing plasma T3 into the supranormal range (approximately 1.6 times euthyroid levels) is there an apparent reduction in hybridization signal for proTRH mRNA to euthyroid levels, suggesting that like other regions of the brain, feedback regulation of thyroid hormone on TRH neurons in the PVN is partially dependent on circulating T4.

One potential source of conversion of T4 to T3 is the base of the third ventricle, where among the ependymal cells are specialized glial cells, tanycytes, that contain one of the highest concentrations of D2 mRNA in any region of the brain (**Figure 6**). These cells also express the monocarboxylate transporter 8, recently demonstrated to be an important thyroid hormone transporter and necessary for normal feedback regulation of hypophysiotropic TRH. Tanycytes extend apical villus-like projections into the CSF and cytoplasmic extensions into the neuropil that envelop the portal capillaries in the median eminence and blood vessels in the hypothalamic arcuate nucleus. Thus, tanycytes may provide a cytoplasmic conduit between the bloodstream and the CSF, allowing bidirectional movement of substances between the vascular and CSF compartments. 'D2 tanycytes,' therefore, may have an important role in the conversion of T4 to T3, either by the uptake of T4 from the CSF or from vascular compartments in the median eminence/arcuate nucleus, and then in the delivery of T3 to the CSF. By volume transmission, T3 would have access to extracellular spaces in the brain, such as the hypothalamic PVN, where it could mediate actions directly on hypophysiotropic TRH neurons. Alternatively, tanycytes may release T3 into the median eminence, where it can be taken up by axon terminals of hypophysiotropic TRH neurons and transported retrogradely or released into the arcuate

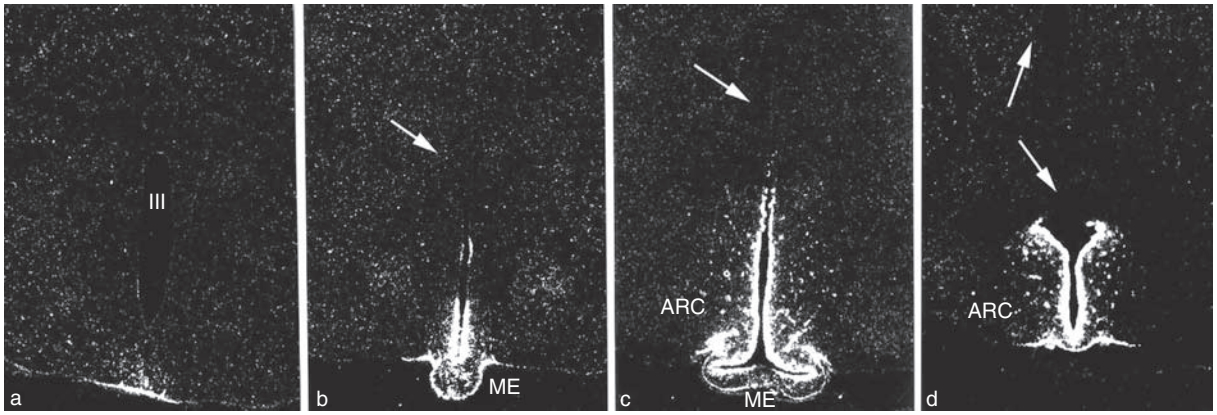

Figure 6 Rostral–caudal distribution of type 2 iodothyronine deiodinase (D2) messenger RNA in the mediobasal hypothalamus (darkfield illumination). In the most rostral section (a), hybridization is present in the external zone of the median eminence (ME) but is absent from walls of the third ventricle (III). In more caudal sections (b–d), intense hybridization is seen over cells lining the floor and infralateral walls of the third ventricle and in the arcuate nucleus (ARC) and ME. Hybridization is absent in the dorsal portions and roof of third ventricular wall in (b–d) (arrows). Adapted from Tu HM, Kim SW, Sulvatore D, et al. (1997) Regional distribution of type 2 thyroxine deiodinase messenger ribonucleic acid in rat hypothalamus and pituitary and its regulation by thyroid hormone. *Endocrinology* 138: 3359–3368.

nucleus through tanycyte–neuronal interactions, where it could influence the activity of arcuate nucleus neurons that have known projections to TRH neurons in the PVN.

Regulation of Hypophysiotropic TRH by Cold Exposure and Suckling

Thyroid hormone levels and TSH are acutely elevated by cold exposure in several animal species because of the increased secretion of TRH in the median eminence. This response is accompanied by a rapid increase in the transcription of TRH in the PVN, resulting in increased secretion of TRH into the portal capillaries from axon terminals in the median eminence, and it can be abolished with TRH antiserum or ablation of the PVN. Particularly intriguing is that proTRH mRNA is increased at a time when circulating levels of thyroid hormone are high. This is contrary to the mechanism of inverse feedback regulation by thyroid hormone, discussed in the section titled 'Feedback regulation of hypophysiotropic TRH by thyroid hormone,' in which proTRH mRNA would be expected to be inhibited. Thus, cold exposure may override the normal mechanism for feedback regulation by altering the set point for inhibition of TRH gene expression by T3.

There is strong evidence to suggest a role of catecholamines in altering the set point for feedback inhibition by thyroid hormone during cold exposure. Norepinephrine stimulates the release of TRH from hypothalamic preparations in culture, and depletion of catecholamines with inhibitors of catecholamine biosynthesis such as α-methyl-paratyrosine or selective catecholamine neurotoxins such as 6-hydroxydopamine abolishes the response of the hypothalamic–pituitary–thyroid (HPT) axis to cold exposure. In addition, the TSH response to acute cold exposure does not occur within the first 10 days after birth in the rat, when the hypothalamic norepinephrine innervation is still immature. Because CART is co-contained in greater than 50% of catecholamine axons in contact with TRH neurons in the PVN and increases TRH gene expression and monoamine accumulation in several regions of the brain, CART may also participate in activation of the HPT axis during cold exposure, either through direct effects on hypophysiotropic TRH neurons or by potentiating catecholamine secretion.

The molecular mechanisms that explain the ability of catecholamines to override the feedback effects of thyroid hormone on TRH gene expression may involve the increased availability of PCREB for binding to Site 4 in the TRH promoter as a result of the binding of catecholamines to G-protein-coupled α receptors on TRH neurons. Presumably, the inhibitory feedback of thyroid hormone bound to its receptor on the TRH promoter may be reduced under these conditions due to increased competition by elevated intranuclear concentrations of PCREB.

Suckling also increases proTRH mRNA in the PVN within 30 to 60 min of the response but for the primary purpose of increasing prolactin secretion from the anterior pituitary. TSH and thyroid hormone levels do not increase, perhaps due to the simultaneous release of CART co-expressed with TRH in hypophysiotropic neurons that can attenuate TRH-induced TSH secretion. While serotonergic afferents from the midbrain raphe may mediate this response, the mechanism for the suckling-induced increase in TRH gene expression remains elusive.

Regulation of Hypophysiotropic TRH by Fasting and Infection (Nonthyroidal Illness Syndrome)

Fasting and infection also alter the normal mechanisms of feedback regulation by circulating levels of thyroid hormone on hypophysiotropic TRH. Under these circumstances, there is a fall in thyroid hormone levels yet a reduction in TRH gene expression in the PVN rather than the anticipated increase that would occur in primary hypothyroidism. Presumably, these conditions override the normal feedback mechanism described earlier by reducing the set point for feedback inhibition by circulating thyroid hormone and thus inducing a state of central hypothyroidism, commonly referred to as 'nonthyroidal illness' or the 'sick euthyroid syndrome' in humans. By reducing thyroid thermogenesis and preserving nitrogen stores, this mechanism is an important adaptive response to reduce energy expenditure until the adverse stimulus has been removed.

The state of central hypothyroidism induced by fasting is orchestrated by a circulating cytokine of white adipose tissue origin, leptin, which declines with fasting and is restored to normal levels by refeeding. If leptin is administered systemically or intracerebroventricularly (icv) to fasting animals, the reduction in circulating levels of thyroid hormone, TSH, and hypophysiotropic proTRH mRNA in the PVN is prevented (Figure 7). Two actions of leptin on hypophysiotropic TRH have been proposed, a direct action as a result of leptin receptor-induced phosphorylation of STAT3 within the hypophysiotropic TRH neurons and an indirect action mediated by neurons in the hypothalamic arcuate nucleus. If the arcuate nucleus is destroyed, not only is the response of the thyroid axis to fasting abolished, but the response of the thyroid axis to the exogenous administration of leptin is lost as well. The arcuate

nucleus, therefore, would appear to serve as a critical locus to mediate the effects of leptin on the HPT axis and in so doing, presumably establishes the set point for feedback sensitivity of proTRH-producing neurons in the PVN to thyroid hormone, lowering the set point at which TRH levels are suppressed during fasting.

At least two anatomically distinct populations of neurons present in the arcuate nucleus with opposing functions, α-MSH-producing neurons that co-express CART and NPY-producing neurons that co-express AGRP, are responsible for the actions of leptin on hypophysiotropic TRH. Both α-MSH and CART have activating effects on hypophysiotropic TRH neurons, and when administered intracerebroventricularly, they have potent effects to restore the fasting-induced suppression of proTRH mRNA in hypophysiotropic neurons to levels in *ad lib* feeding animals (Figures 8(a) and 8(c)). The activating mechanism for α-MSH on the TRH promoter is likely mediated by the phosphorylation of CREB, and

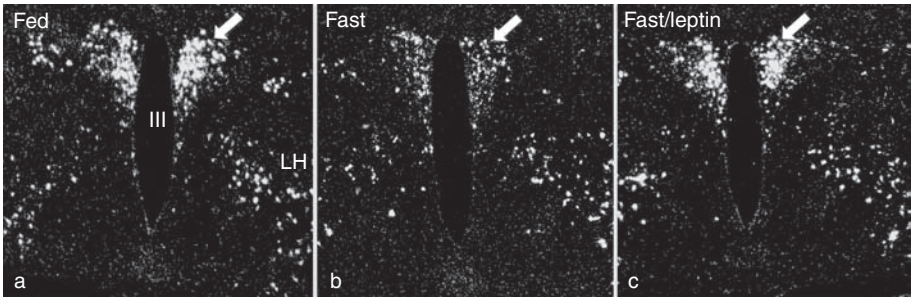

Figure 7 *In situ* hybridization autoradiographs of prothyrotropin-releasing hormone (proTRH) messenger RNA (mRNA) in the rat paraventricular nucleus (PVN; arrow) of (a) normal fed and (b) fasting animals. Note the marked reduction in hybridization signal for the fasting animal. (c) ProTRH mRNA levels are restored to normal in fasting animals administered leptin. From Legradi G, Emerson CH, Ahima RS, Flier JS, and Lechan RM (1997) Leptin prevents fasting-induced suppression of prothyrotropin-releasing hormone messenger ribonucleic acid in neurons of the hypothalamic paraventricular nucleus. *Endocrinology* 138: 2569–2576.

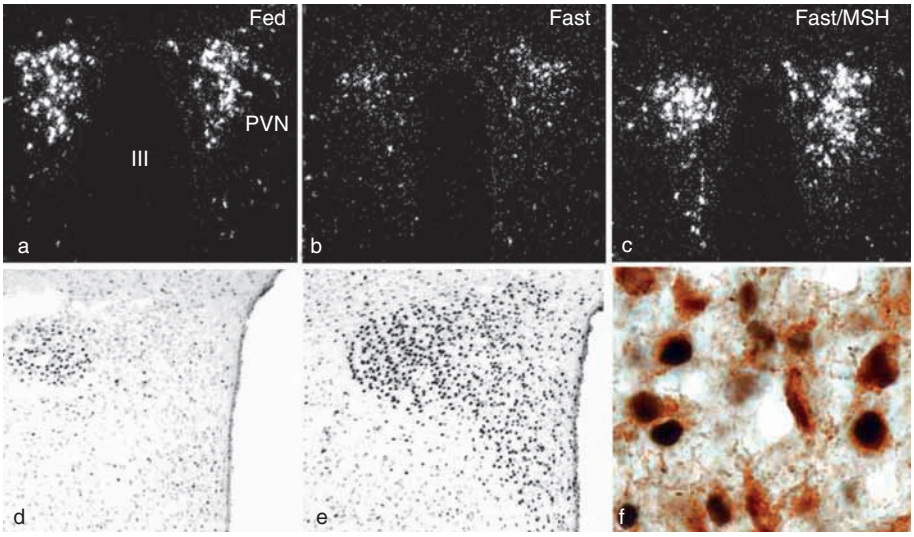

Figure 8 (a–c) *In situ* hybridization autoradiograms of prothyrotropin-releasing hormone (proTRH) messenger RNA (mRNA) in the medial and periventricular parvocellular subdivisions of the rat paraventricular nucleus (PVN) of (a) fed, (b) fasted, and (c) fasted animals receiving α-melanocortin-stimulating hormone (α-MSH) icv. Fasted animals receiving α-MSH show a marked increase in proTRH mRNA that is similar to that of the fed control. (d)–(f) Effect of icv α-MSH on cyclic adenosine monophosphate-responsive element binding protein (CREB) phosphorylation (PCREB) in the PVN of fasted animals. (d) Fasted animals show little PCREB in the medial and periventricular parvocellular subdivisions but (e) a marked increase following α-MSH administration. (f) Double-labeling immunocytochemistry shows that following α-MSH administration, the majority of TRH neurons (brown cytoplasm) contain PCREB (black nucleus). III, third ventricle. (a–c) Modified from Fekete C, Legradi G, Mihaly E, et al. (2000) Alpha-melanocyte-stimulating hormone is contained in nerve terminals innervating thyrotropin-releasing hormone-synthesizing neurons in the hypothalamic paraventricular nucleus and prevents fasting-induced suppression of prothyrotropin-releasing hormone gene expression. *Journal of Neuroscience* 20: 1550–1558. (d–f) From Sarkar S, Legradi G, and Lechan RM (2002) Intracerebroventricular administration of alpha-melanocyte stimulating hormone increases phosphorylation of CREB in TRH- and CRH-producing neurons of the hypothalamic paraventricular nucleus. *Brain Research* 945: 50–59.

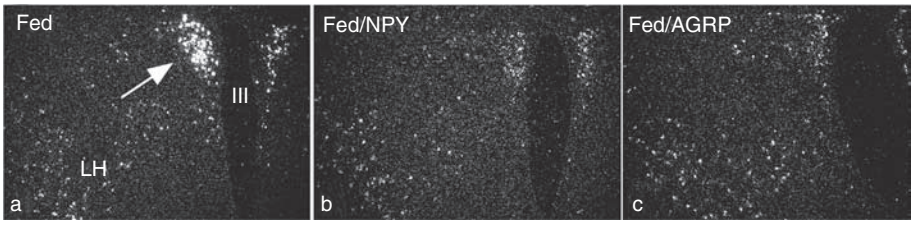

Figure 9 Darkfield illumination photomicrographs of prothyrotropin-releasing hormone (proTRH) messenger RNA (mRNA) in the medial parvocellular subdivision of the hypothalamic paraventricular nucleus (PVN; denoted by arrow) in (a) artificial CSF-treated control *ad lib*-fed animals, (b) *ad lib*-fed animals treated with neuropeptide Y (NPY) intracerebroventricularly (icv), and (c) *ad lib*-fed animals treated with agouti-related peptide (AGRP) icv. Note marked reduction in silver grains over neurons of the PVN in the AGRP- and NPY-treated groups that are selective for the PVN. No effects of NPY or AGRP are seen on proTRH mRNA expression in the lateral hypothalamus (LH). III, third ventricle. Modified from Fekete C, Sarkar S, Rand WM, et al. (2002) Agouti-related protein (AGRP) has a central inhibitory action on the hypothalamic–pituitary–thyroid (HPT) axis *Endocrinology* 143: 3846–3853.

following intracerebroventricular administration, α-MSH markedly increases the number of TRH neurons in the PVN that contain PCREB in their nucleus (Figures 8(d)–8(f)). In addition, mutation of Site 4 in the TRH promoter or introduction of a dominant negative inhibitor of CREB that cannot be phosphorylated in an *in vitro* expression system prevents activation of the TRH promoter by α-MSH. The mechanism by which CART activates TRH neurons in the PVN remains unknown.

Conversely, both NPY and AGRP, when administered icv to *ad lib* feeding animals (Figure 9), have profound inhibitory effects on TRH gene expression in hypophysiotropic neurons, creating a state of central hypothyroidism that closely compares with that observed in fasting animals. Presumably, AGRP produces its inhibitory effects on TRH gene expression by antagonizing the activating effects of α-MSH at the melanocortin 4 receptor and/or by suppressing constitutively active melanocortin receptors by functioning as an inverse agonist. The inhibitory effect of NPY is mediated through both Y1 and Y5 receptors and is likely due to inhibition of cAMP and hence a reduction in the accumulation of PCREB in the nucleus of TRH neurons.

Thus, during fasting when circulating leptin levels decline, there is simultaneous inhibition of α-MSH and CART production and increase in AGRP and NPY production in arcuate nucleus neurons. Under these circumstances, the end result for hypophysiotropic TRH neurons in the PVN would be a decline in CREB phosphorylation, reducing the availability of PCREB for binding to Site 4 in the TRH promoter. Therefore, the inhibitory feedback effects of thyroid hormone bound to its receptor on the TRH promoter may become enhanced due to reduced competition by lower intranuclear concentrations of PCREB, essentially lowering the set point for feedback inhibition of the TRH gene by thyroid hormone. Activation of D2 in tanycytes may also contribute to fasting-induced inhibition of hypophysiotropic TRH that may be dependent on leptin and glucocorticoid signaling.

Paradoxically, suppression of the HPT axis following endotoxin administration is associated with increased circulating leptin levels, increased melanocortin signaling, and increased CART gene expression in the arcuate nucleus that would predict activation of hypophysiotropic TRH gene expression. Nevertheless, endotoxin suppresses TRH gene expression in the PVN, resulting in central hypothyroidism similar to that observed with fasting. Presumably, therefore, a set of regulatory controls different from that observed during fasting must come into play under these conditions, superseding all other regulatory effects on hypophysiotropic TRH. The observation that endotoxin markedly increases D2 gene expression and enzymatic activity in tanycytes has raised the possibility that central hypothyroidism associated with infection may be secondary to an endotoxin-induced increase in local tissue levels of T3 in the mediobasal hypothalamus. This could have the result of suppressing hypophysiotropic TRH by inhibitory feedback effects on TRH gene expression, as schematically shown in Figure 10. The mechanism for endotoxin-induced increase of D2 activity in tanycytes is unknown but is unlikely related to the fall in circulating levels of thyroid hormone. Thyroid hormone replacement to thyroidectomized animals treated with endotoxin still shows an increase in tanycyte D2 activity, whereas the T4 clamp blocks the endotoxin-induced increase of D2 activity in the anterior pituitary and cerebral cortex. In addition, elevated levels of corticosterone probably have little role because corticosterone-clamped adrenalectomized animals still show a D2 response to lipopolysaccharide. Rather, the effects of endotoxin on tanycyte D2 activity may be mediated by endotoxin itself or by endotoxin-induced cytokines via the activation of nuclear factor-kappa B. This hypothesis is based on the observation that the

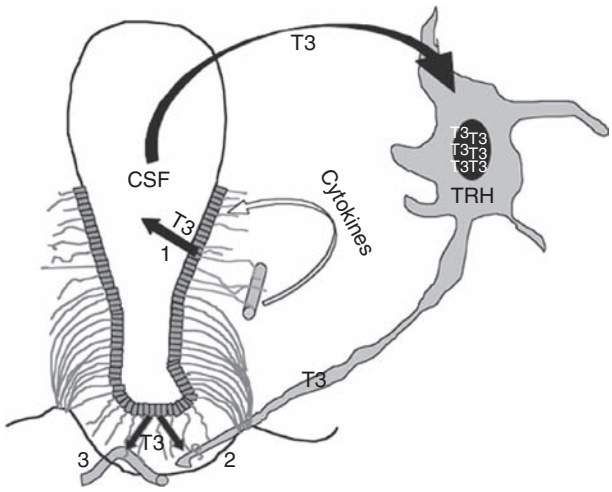

Figure 10 Proposed mechanism for type 2 iodothyronine deiodinase (D2) regulation of the hypothalamic–pituitary–thyroid axis following the administration of lipopolysaccharide (LPS). LPS increases D2 activity in tanycytes, resulting in increased thyroxine (T4) to tri-iodothyronine (T3) conversion. (1) T3 is released from tanycyte apical processes into the cerebrospinal fluid (CSF) for conveyance to the paraventricular nucleus (PVN) or (2) taken up from hypophysiotropic thyrotropin-releasing hormone (TRH) axonal processes in the median eminence and transported retrogradely to its cell body in the PVN. (3) T3 may also be released into the portal capillary system and directly inhibit the secretion of thyroid-stimulating hormone. Local tissue hyperthyroidism inhibits TRH in the PVN. From Lechan RM and Fekete C (2005) Role of thyroid hormone deiodination in the hypothalamus. *Thyroid* 15: 883–897.

D2 promoter contains nuclear factor-kappa B binding sites that can cause profound activation of the human D2 promoter in several cell lines. Inhibitory effects of endotoxin on the HPT axis may also be mediated directly on hypophysiotropic TRH neurons by cytokines such as interleukin-1, interleukin-6, tumor necrosis factor, and interferon, which rise dramatically following endotoxin administration and abolish both the normal circadian variation and pulsatile secretion of TSH.

See also: Hypothalamic–Pituitary–Adrenal (HPA) Axis; Neuroendocrine Aging: Pituitary Metabolism; Neuroendocrine Peptide Processing; Neuroendocrinology; Thyroid Hormone and Transcriptional Regulation in the CNS.

Further Reading

Abel ED, Ahima RS, Boers M-E, Elmquist JK, and Wondisford FE (2001) Critical role for thyroid hormone receptor $\beta 2$ in the regulation of paraventricular thyrotropin-releasing hormone neurons. *Journal of Clinical Investigation* 107: 1017–1023.

Coppola A, Liu ZW, Andrews AM, et al. (2007) A central thermogenic-like mechanism in feeding regulation: An interplay between arcuate nucleus T3 and UCP2. *Cell Metabolism* 5: 21–33.

Fekete C, Kelly J, Mihaly E, et al. (2001) Neuropeptide Y has a central inhibitory action on the hypothalamic–pituitary–thyroid axis. *Endocrinology* 142: 2606–2613.

Fekete C and Lechan RM (2006) Neuroendocrine implications for the association between CART and hypophysiotropic TRH. *Peptides* 27: 2012–2018.

Fekete C, Legradi G, Mihaly E, et al. (2000) Alpha-melanocyte-stimulating hormone is contained in nerve terminals innervating thyrotropin-releasing hormone-synthesizing neurons in the hypothalamic paraventricular nucleus and prevents fasting-induced suppression of prothyrotropin-releasing hormone gene expression. *Journal of Neuroscience* 20: 1550–1558.

Fekete C, Sarkar S, Rand WM, et al. (2002) Agouti-related protein (AGRP) has a central inhibitory action on the hypothalamic-pituitary-thyroid (HPT) axis *Endocrinology* 143: 3846–3853.

Fliers E, Alkemade A, Wiersinga WM, and Swaab DF (2006) Hypothalamic thyroid hormone feedback in health and disease. *Progress in Brain Research* 153: 189–207.

Guo F, Bakal K, Minokoshi Y, and Hollenberg AN (2004) Leptin signaling targets the thyrotropin-releasing hormone gene promoter *in vivo*. *Endocrinology* 145(5): 2221–2227.

Harris M, Aschkenasi C, Elias CF, et al. (2000) Transcriptional regulation of the thyrotropin-releasing hormone gene by leptin and melanocortin signaling. *Journal of Clinical Investigation* 107: 111–120.

Lechan RM and Fekete C (2005) Role of melanocortin signaling in the regulation of the hypothalamic–pituitary–thyroid (HPT) axis. *Peptides* 27: 310–325.

Lechan RM and Fekete C (2005) Role of thyroid hormone deiodination in the hypothalamus. *Thyroid* 15: 883–897.

Lechan RM and Fekete C (2006) The TRH neuron: A hypothalamic integrator of energy metabolism. *Progress in Brain Research* 153: 209–235.

Lechan RM and Fekete C (2007) Infundibular tanycytes as modulators of neuroendocrine function: Hypothetical role in the regulation of the thyroid and gonadal axis. *Acta Bio-Medica: Atenei Parmensis* 78(supplement 1): 84–98.

Lechan RM and Toni R (2004) Functional anatomy of the hypthalmus and pituitary. In: Grossman A (ed.) *Neuroendrocrinology, hypothalamus and Pituitary.* http://www.endotext.org/neuroendo/neuroendo3b/neuroendoframe3b.htm (accessed jul. 2008).

Legradi G, Emerson CH, Ahima RS, Flier JS, and Lechan RM (1997) Leptin prevents fasting-induced suppression of prothyrotropin-releasing hormone messenger ribonucleic acid in neurons of the hypothalamic paraventricular nucleus. *Endocrinology* 138: 2569–2576.

Nillni EA and Sevarino KA (1999) The biology of pro-thyrotropin-releasing hormone-derived peptides. *Endocrine Reviews* 20: 599–648.

Perello M and Nillni EA (2007) The biosynthesis and processing of neuropeptides: Lessons from prothyrotropin releasing hormone (ProTRH). *Frontiers in Bioscience* 12: 3554–3565.

Sanchez E, Uribe RM, Corkidi G, et al. (2001) Differential responses of thyrotropin-releasing hormone (TRH) neurons to cold exposure or suckling indicate functional heterogeneity of the TRH system in the paraventricular nucleus of the rat hypothalamus. *Neuroendocrinology* 74: 407–422.

Sarkar S, Lagradi G, and Lechan RM (2002) Intracerebroventricular administration of alpha-melanocyte stimulating hormone increases phosphorylation of CREB in TRH- and CRH-producing neurons of the hypothalamic paraventricular nucleus. *Brain Research* 945: 50–59.

Sarkar S, Wittmann G, Fekete C, and Lechan RM (2004) Central administration of cocaine- and amphetamine-regulated transcript increases phosphorylation of cAMP response element binding protein in corticotropin-releasing hormone-producing neurons but not in prothyrotropin-releasing hormone-producing neurons in the hypothalamic paraventricular nucleus. *Brain Research* 999: 181–192.

Singru P, Fekete C, and Lechan RM (2005) Neuroanatomical evidence for participation of the hypothalamic dorsomedial nucleus (DMN) in regulation of the hypothalamic paraventricular nucleus (PVN) by α-melanocyte stimulating hormone. *Brain Research* 1064: 42–51.

Trajkoviac M, Visser TJ, Mittag J, et al. (2007) Abnormal thyroid hormone metabolism in mice lacking the monocarboxylate transporter 8. *Journal of Clinical Investigation* 117: 627–635.

Tu HM, Kim SW, Sulvatore D, et al. (1997) Regional distribution of type 2 thyroxine deiodinase messenger ribonucleic acid in rat hypothalamus and pituitary and its regulation by thyroid hormone. *Endocrinology* 138: 3359–3368.

Relevant Website

http://endotext.org – Endotext.org: Your Endocrine Source.

Ideal Observer Theory

D Kersten, University of Minnesota, Minneapolis, MN, USA
P Mamassian, CNRS and Université Paris 5, Paris, France

© 2009 Published by Elsevier Ltd.

Introduction

An ideal observer is ideal in the sense that it achieves statistically optimal performance for a specified task. It makes the best decisions or estimates given uncertain sensory information and prior knowledge. What is best is specified in terms of costs and benefits. Ideal observer models can be used to study different types of real test observers, such as in the psychophysical analysis of the input/output behavior of humans and spike-train analysis of information carried by single neurons. If human perceptual performance approaches that of an ideal observer, this can, in principle, rule out a large class of candidate neural mechanisms that are suboptimal relative to the ideal. Although the ideal observer concept can be traced to earlier work, the modern notion of an ideal observer was developed in the 1950s by Peterson, Birdsall, Fox, Tanner, Swets, and others in the context of a general theory of signal detectability. The theory and its initial applications to human auditory and then visual sensitivity are described in a classic 1966 book by Green and Swets. In the late 1970s, Horace Barlow applied ideal observer analysis to higher-level perceptual tasks, such as the perception of symmetric patterns. By the 1990s, comparisons of human and ideal performance were extended to an increasingly wider set of problems, including object and motion perception, perceptual organization, reading, and motor control. Ideal observer analysis has shown, for example, that the way in which human observers integrate depth or motion information takes into account uncertainty in the input image cues, as we might expect from an ideal observer doing the same task.

Defining the Ideal Observer

The key ingredients in defining an ideal observer are (1) the generative model; (2) the task requirements, including a measure of performance; and (3) an observer model that specifies the optimal action rule for the given task requirement. The ideal observer can be defined in terms of four classes of random variables on a directed graph (i.e., one with arrows) that represents how the variables influence one another (see Figure 1). These variables represent states of the world s, observations (data) x that result, actions a (e.g., decision or estimate) on the data, and the losses (costs) $L(a, s)$ of action a, given the true state s. The spaces can be discrete or continuous. For example, a state of the world could mean one of two positions of a light switch or the distance of an object. An action may be only indirectly related to the state, for example, when a person throws a ball, the speed of release of the ball aimed at a distant target versus an estimate of the state itself (distance). We now go over the elements in greater detail.

The Generative Model

The generative model specifies how the states of the world s determine the observed data x, that is, sensory input (e.g., the pattern of image intensities). The state space can be discrete or continuous. In general, states and data are multidimensional, $s = (s_1, s_2, s_3, \ldots)$ and $x = (x_1, x_2, x_3, \ldots)$, and the relationship between states and data can have complex dependencies (a simplified example is shown in Figure 1(b)). The generative model consists of the prior probability of the state s, $p(S = s)$ and the probability of the observation x given s, $p(X = x|s)$. The prior probability can be simple, representing the probability of a hypothesis with just two values, such as a light switch being set to high $S = s_{\mathrm{H}}$ rather than low $S = s_{\mathrm{L}}$. But the prior probability on s can also be used to model complex pattern regularities that take into account covariation among elements of s, such as the probability of a surface tending to have smoothly varying, rather than rapidly changing, depths. In this case, validating the model for the prior from real-world data can be a significant problem by itself; however, in the lab, an experimenter can specify the prior. The probability of an observation x given s describes how the signal s is encrypted in the data and is the likelihood of the hypothesis (the state s) given the data. The description could be as simple as $x = s + n$, where x is light intensity, s is the average level of one of two fixed light intensities, and n is a sample of Gaussian noise. Then $p(X = x|s) = p(x - s)$. Or $p(X = x|s)$ could be derived from a more complicated function $x = \phi(s, n)$, where x is a vector representing an image, and ϕ is a function that describes how a shape s together with a confounding variable n, such as viewpoint (or slant angle in the next example), determines the image observation. In the context of a psychophysical experiment, the generative model can be thought of as a probabilistic description of how the stimuli are generated.

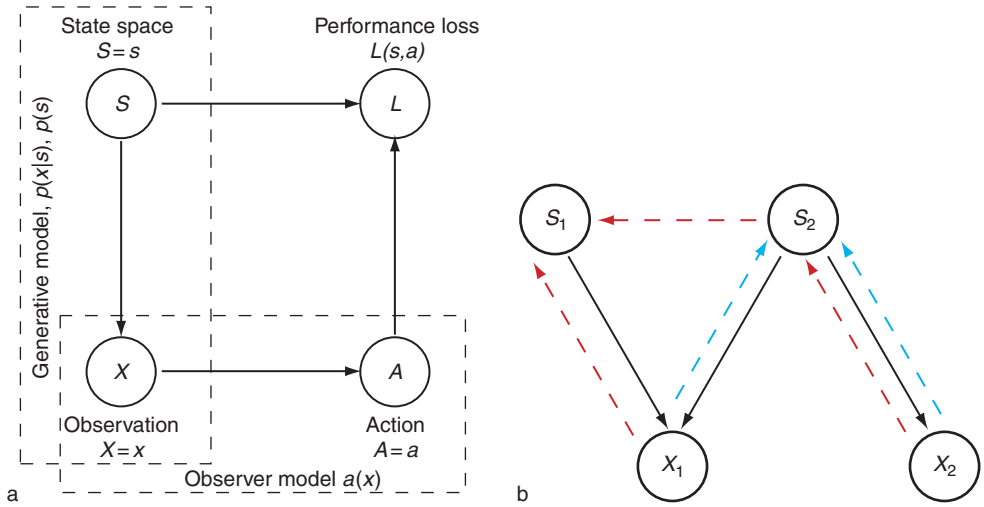

Figure 1 State and observation spaces in an ideal observer model: (a) components of the model; (b) an example in which state variables S_1 and S_2 both influence observation X_1 but only S_2 directly influences X_2. In (a), the nodes represent random variables, and the arrows indicate how they influence one another. The arrows in the graph (formally, a directed acyclic graph) can be interpreted as describing the dependencies among variables. The graph structure determines how to factor the joint distribution for the task, $p(s, x, l, a) = p(l|a, s)p(a|x)p(x|s)p(s)$. From this, we can formalize the problem of estimating components, such as the probability of a loss value $p(L = \mathrm{l})$ (i.e., probability of error $L = 1$) and other performance measures, such as the hit rate $p(L = 0|s_{\mathrm{H}})$, or false positive rate $p(L = 0|s_L)$ for signal state s_{H} and noise state s_{L}. As shown in (b), the state and observation spaces can have complex dependencies. Solid lines indicate the generative model, and the dashed lines the directions of (inverse) inference for two tasks. Suppose the goal is to estimate S_2; then both observations determine the optimal estimate (blue dashed lines). This is an example of cue integration, in which the two observations are conditionally independent of S_2. If the goal is to estimate S_1, both observations again determine the estimate, but X_2 indirectly affects beliefs about S_1 through S_2 (red dashed lines). This is an example of explaining away. The graph structure is equivalent to the factorization $p(s_1, s_2, x_1, x_2) = p(x_1|s_1, s_2)p(x_2|s_2)p(s_2)p(s_1)$.

The Task Requirements

The task requirements are twofold: (1) decide what is to be done with the data, for example, detect the presence or absence of the signal (e.g., high vs. low contrast? or animal present or not?), identify the signal (e.g., which letter, A, B, or C?), and estimate a continuous value (e.g., what is the depth of the object?); and (2) determine the value of achieving or not achieving the goal. Not all elements of the state variable, $s = (s_1, s_2, s_3, \ldots)$, are equally important and need to be determined accurately. In classic signal detection theory, state variables are divided into two types – the relevant variables to be inferred accurately (signals) and those to be discounted (noise) – with associated losses for right and wrong answers. Generally, the value of accuracy can be expressed in terms of a loss function (or, alternatively, a gain function, equal to one minus the loss function). For example, different degrees of relevance can be placed on the state variables by specifying a loss function $L(a_1, a_2; s_1, s_2)$, which puts a cost on choosing actions a_1, a_2 when the true states are s_1 and s_2. If the action is an estimate of s, $a = \hat{s}$, loss may be expressed in terms of differences, $L(\hat{s} - s)$.

Optimal Action or Decision Rule

Of the many possible observer models, the ideal observer is defined as one that uses an action rule $a(x)$ that minimizes the risk for the task. The expected risk is defined as the loss averaged over both state and observation variables: $R(a|x) = \sum_{s,x} L(s, a(x))p(s, x)$. For a given observation x, the decision rule is to choose action a so as to minimize $R(a|x) = \sum_{s,x} L(s, a(x))p(s|x)$. Often the action space has a simple relationship with the signal state space. For example, in a detection or discrimination task, the state space is discrete and binary, and the action is to decide whether the signal was sent or not (answer yes or no to questions such as 'Is Thomas there?' or 'Is Thomas bigger than Simon?'). For identification, the state space is discrete with multiple values ($S = \{Thomas, Simon, Hermann, Jacob\}$), and the actions are decisions based on x ('Is it Thomas? Simon? Hermann? Jacob?'). For estimation, the action is an estimate of the state variable, $a = \hat{s}$ from observation x (e.g., 'How tall is she?'). If the loss is uniformly high for all wrong decisions and low for the correct decision, then the optimal action is to pick the most probable state s given the data x. This is the maximum *a posteriori* (MAP)

observer: $a(x) = \text{argmax}_s p(s|x)$. This rule yields the smallest average error. If the loss is the squared error between the state and estimated state variables, then the optimal action is to choose the mean of the posterior. Next, we illustrate task requirements with examples for detection and estimation tasks.

Applications of Ideal Observer Theory

We can distinguish two types of applications of ideal observer theory. For the purposes of this article, we refer to the first as ideal observer analysis; in this application, the test observer (humans or neural system) competes against the ideal observer in the same well-defined laboratory task. Because the experimenter controls the conditions for the stimuli and task, no observer can do better (on average) for the specified task. In this case, the performance of an ideal observer can be thought of as a benchmark against which to measure human performance. The emphasis is on departures from ideal performance, which provides clues to the underlying, generally subideal mechanisms. Ideal observer analysis is a useful addition to the experimenter's toolbox alongside linear and nonlinear systems analysis. Ideal observer analysis has been used to show that humans perform some tasks with strikingly high efficiency relative to the ideal observer but perform others quite poorly. One of the historic success stories was to show that the human ability to discriminate light intensity under nighttime viewing conditions can be extremely high. Following on the work of Hecht, Schlaer, and Pirenne in the 1940s, Horace Barlow showed in the early 1960s that human quantum efficiencies were sufficiently high as to rule out any explanation that required the retinal transduction of more than a few photons.

The second application of ideal observer theory is as an approximate theory or model of performance. The primary distinction between the two applications is that the second typically makes simplifying assumptions about the generative model (see the section titled 'Estimation'), including the nature of the input representation (e.g., the ideal observer may have geometrical features as input, whereas the human observer receives image intensities as input, from which geometrical information is derived) or how the data are caused by the states of the world (e.g., Gaussian noise is added, whereas the actual noise may be non-Gaussian). In other words, the ideal observer is optimal with respect to a generative model that may differ in detail from the true generative model, either as defined in a particular laboratory study or in a real-world task. The primary reason for these simplifying assumptions is practical – the true ideal observer is too hard to calculate. Even though humans are in general suboptimal, an ideal observer model may nonetheless go a long way in explaining the observed behavior. In this second type of application, ideal observer theories are equivalent to Bayesian theories of perceptual performance. As a modeling tool, a general advantage of the Bayesian ideal observer approach is that it avoids a commitment to untestable mechanistic details, such as a particular neural architecture, while still providing quantitative predictions of behavior (Table 1).

Table 1 Ideal observer model equations

Terms	*Equations*
Conditioning	$p(x\|y) = p(x,y)/p(y)$
Marginalization	$p(x) = \int p(x,y)\mathrm{d}y$
Bayes's rule	$p(s\|x) = p(x\|s)p(s)/p(x)$
Signal-to-noise ratio for equal variance Gaussian case	$d'_{\text{snr}} = (\mu_{\text{s}} - \mu_{\text{n}})/\sigma$
Relation of z and P	$z(P) : p = (1/\sqrt{2\pi}) \int_x^\infty e^{-x^2/2}\mathrm{dx}$
Sensitivity from false alarm and hit rates	$d'_{\text{perf}} = z(P_{\text{FA}}) - z(P_{\text{H}})$
ROC in terms of z values for unequal variance case	$z(P_H) = (\sigma_n/\sigma_s)z(P_{FA}) - [(\mu_s - \mu_n)/\sigma_s]$
d' for 2AFC from proportion correct	$d'_{\text{2afc}} = -\sqrt{2}z(P_{\text{c}})$
Efficiency	$E = (d'_{\text{snr}_{\text{I}}}/d'_{\text{snr}_{\text{T}}})^2$

2AFC, two-alternative forced-choice task; ROC, receiver operating characteristic; snr, signal-to-notice ratio.

Examples of Ideal Observer Analysis

Signal Detection

Figure 2 illustrates a classic task in signal detection theory in which a switch is set to high or low and the observer has to guess the setting based on an observation, in this case a measurement of light intensity.

The generative model The switch setting is the state variable. If the two switch settings are equally likely, independent of the data, then the prior is constant and equal to one-half, $p(S = s) = 1/2$. The second part of the generative model, the likelihood, specifies how the observations depend on the state. Suppose that when the switch is set to high the average light intensity is higher than when the switch is set to low. However, because of trial-to-trial fluctuations (noise), the measured or observed light intensity varies. So sometimes the measured light intensity is higher for the low setting than for the high switch setting. Note that in ideal observer analysis, we may be able to test the validity of the generative model, apart from questions of optimal inference. Thus, for example, the

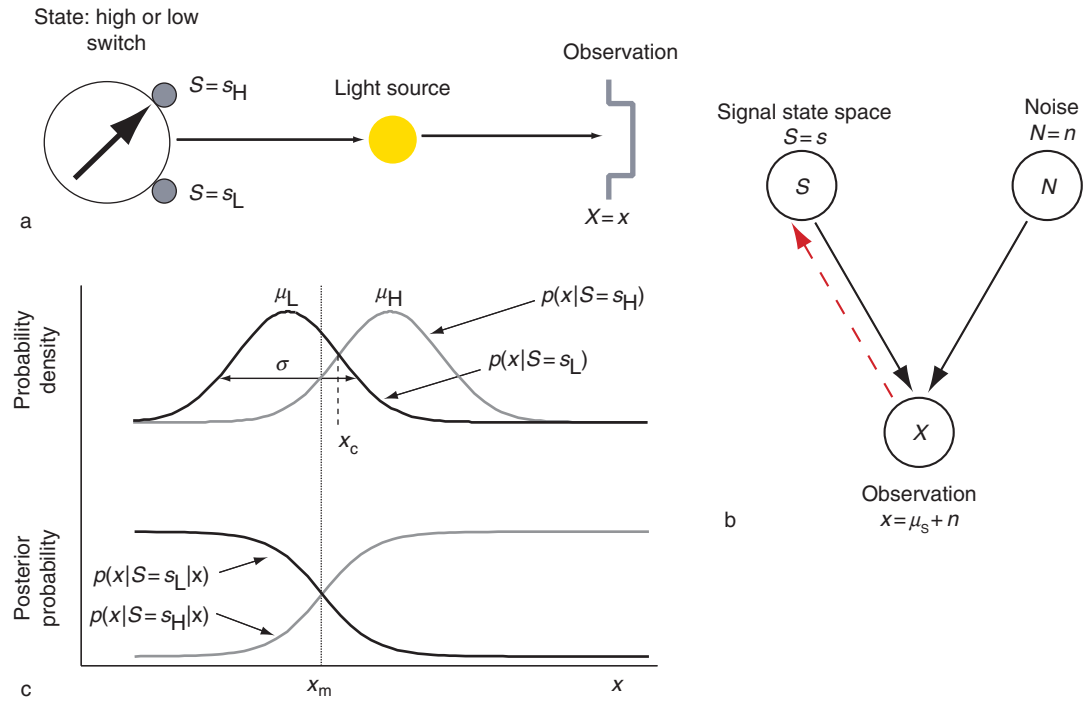

Figure 2 Example of a simple ideal observer for a binary task, signal detection: (a) schematic of the physical model in which a light source produces an observation x, of light intensity; (b) graph showing causal dependencies corresponding to factorization in terms of the components of the generative model: $p(x, s, n) = p(x|s, n)p(s)p(n)$ (the direction of the inference for the detection task is shown by the dashed arrow); (c) Gaussian probability densities of the observation under the two hypotheses (switch states), also known as the likelihood functions corresponding to different means, μ_L, μ_H (top) and graphs of the posterior probabilities (bottom). In (c, top), the standard deviations are the same, σ. In (c, bottom), when the switch probabilities are not equal, choosing the highest likelihood no longer produces the lower error rate. The ideal observer guesses the state with highest posterior probability.

light switch model assumes additive Gaussian noise (although at low intensities at which fluctuations in photon emission and absorption may dominate; a Poisson model is a better choice). Figure 2(c) (top) shows the Gaussian distributions of observations under the two possible state conditions. The noise is additive because the high switch setting produces the same distribution as that of the lower setting, except that the mean is shifted by an added amount. The standard deviation remains the same.

The task Suppose that given an observation the observer has to decide whether it was due to a high or a low switch setting. Further, suppose that the task requires that the probability of error $p(L=1)$ be as small as possible. Loss is represented by $1-\delta_{s,\hat{s}}$ (the Kronecker delta function, $\delta_{s,\hat{s}}=1$ if $s=\hat{s}$, and $\delta_{s,\hat{s}}=0$ otherwise). Another common task in psychophysics is the two-alternative, forced-choice task (2AFC). The test observer (e.g., a human subject) and ideal observer share the same state, observation, and action space, as well as loss function (scoring system). In general, they differ in their mappings of observations to actions, that is, their decision rules.

The optimal decision rule Intuitively, we guess that the best strategy might be to choose s_H if $p(x|s_H) > p(x|s_L)$, and, in fact, this is the optimal rule for the constant prior case if we want to have the smallest error rate. In our example, this rule is equivalent to choosing s_H if $x > x_c$, where x_c is the point where the two curves in Figure 2(c) (top) cross one another. However, the general rule for minimizing error is the MAP rule, which takes into account the prior probability and picks the state with the higher posterior probability; that is, it chooses s_H if $p(s_H|x) > p(s_L|x)$. The posterior $p(s|x)$ is determined by the prior and likelihood through Bayes's rule: $p(s|x) = p(x|s)p(s)/p(x)$. If, for example, the prior probability of the switch setting high is bigger than the probability of low, this is equivalent to moving the criterion to x_m (i.e., choosing s_H if $x > x_m$; compare the upper and lower panels of Figure 2(c)). But there are other rules depending on the loss function we choose. The loss function is a 2×2 payoff matrix with loss values for hit (true positive), false alarm (false positive), correct rejection (true negative), and miss (false negative). We can assign distinct losses to each of these outcomes, in which, for example, false

alarms are more or less costly. In general, this results in a different rule r based on minimizing the risk – choose s_H if $r(s_H|x) > r(s_L|x)$. Standard results in signal detection theory show that for this problem the optimal rule is equivalent to: choose s_H if $x > x_m$, where the value of x_m (the criterion) depends on the loss matrix.

How to compare test and ideal observers? The test observer can perform no better than the ideal and in general does worse because of suboptimal action rules, such as systematically failing to use all the relevant information in the input stimulus. In this specific example, an observer can be suboptimal by putting the criterion x_m at the wrong place – a response bias. As the criterion shifts, the values of the hit and false alarm rates also shift. A graph of hit versus false alarm rates is known as the 'receiver operating characteristic' (ROC). A key contribution of signal detection theory was to show that underlying human sensitivity to a sensory signal (d') could be teased apart from the value of the criterion.

In theory, the ideal and test observers can be compared in terms of their scores (average loss, i.e., proportion of errors) or d'_{perf}, given the same states and resulting observations. However, it is often not practical to directly compare error rates (when the ideal is making a modest number of mistakes, the test observer may be near chance and would thus require too many measurements to estimate d'_{perf_T}); alternatively, we can compare state parameters in the generative model that produce identical performances (e.g., sensitivities d'_{perf_T}, d'_{perf_I} for test and ideal in an equal-variance Gaussian noise-detection task are determined by the signal-to-noise ratios d'_{snr_T} and d'_{snr_I}). Suppose, for example, that a human and ideal observer both have the same proportion correct in a 2AFC task; then, the ideal observer can be used as a benchmark, where a standard measure of comparison is statistical efficiency $E =$ (Number of samples required by the ideal observer/Number of samples required by test observer). In the case of additive Gaussian noise, this is equivalent to $E = (d'_{\text{snr}_I}/d'_{\text{snr}_T})^2$. If the noise is the same, efficiency takes on the simple form of the ratio of the squared threshold values of the ideal observer to the human observer. In the case of detecting an image pattern in additive Gaussian noise, an observer can be suboptimal by failing to use all the image samples (pixels) or by performing as if there were additional noise. These two sources of suboptimality have been teased apart in psychophysical experiments, measured in terms of calculation (or central) efficiency or transduction efficiency. Efficiency provides a unit-free measure of performance that allows for the stimulus complexity and task constraints.

Estimation

Perception is critical to a wide range of tasks, of which signal detection is just one. Consider the problem of determining the dimensions of a three-dimensional object from a two-dimensional projected image. In this case, a major source of uncertainty results from the loss of information due to projection. Figure 3 illustrates a Bayesian ideal observer model for a simple task to estimate the height and slant of a class of elliptical disks. Given an elliptical shape on the retina, what object could have produced it? It could be an elliptical disk, but a flat circular disk could also produce an elliptical image.

The generative model For simplicity, assume that the world consists of elliptical disks of width equal to 1 but with various heights h. These disks can have various slants, α, with respect to a viewer. The states are continuous and can be represented by $s = (s_1, s_2) = (h, \alpha)$. We can hypothesize a prior, $p(s) = p(h, \alpha)$ on these values. Figure 3 shows the graph of a bivariate Gaussian prior that assumes that the average slant is $\pi/4$ and the average height is 1. If viewed from a far distance, the width of the projection of any disk from this state space is a constant value (say, 1); but the height of the projection depends on both the true height of the disk and the slant. An observation of the height x of the retinal image provides data that constrains but, without additional information, does not uniquely determine the possible states. In the signal-detection example, uncertainty resulted from added Gaussian noise; here, although we assume some additional measurement noise, the critical uncertainty results from the property that the surface could be slanted away from the viewer by various unknown degrees. For a distant view, the image observation x can be approximated as $x \sim h\cos(\alpha) + n$. Assuming Gaussian noise, this determines the likelihood function, $p(x - h\cos(\alpha))$, where $p(n) = e^{-n^2/2\sigma^2}/(\sigma\sqrt{2\pi})$ is the standard formula for a Gaussian distribution with 0 mean.

The task Given an observation, say $x = 1/2$, the ideal observer needs to determine the height and slant with the least risk. Figure 3 illustrates the steps. It first computes the likelihood of the observation for all possible values of the state variables. The likelihood has a high ridge of constant height, meaning that there are infinitely many pairs of height and slant values that are equally likely. Thus, for example, a height of 4 inclined at 83° is no less likely than a head-on slant of 0 together with a height of 1/2.

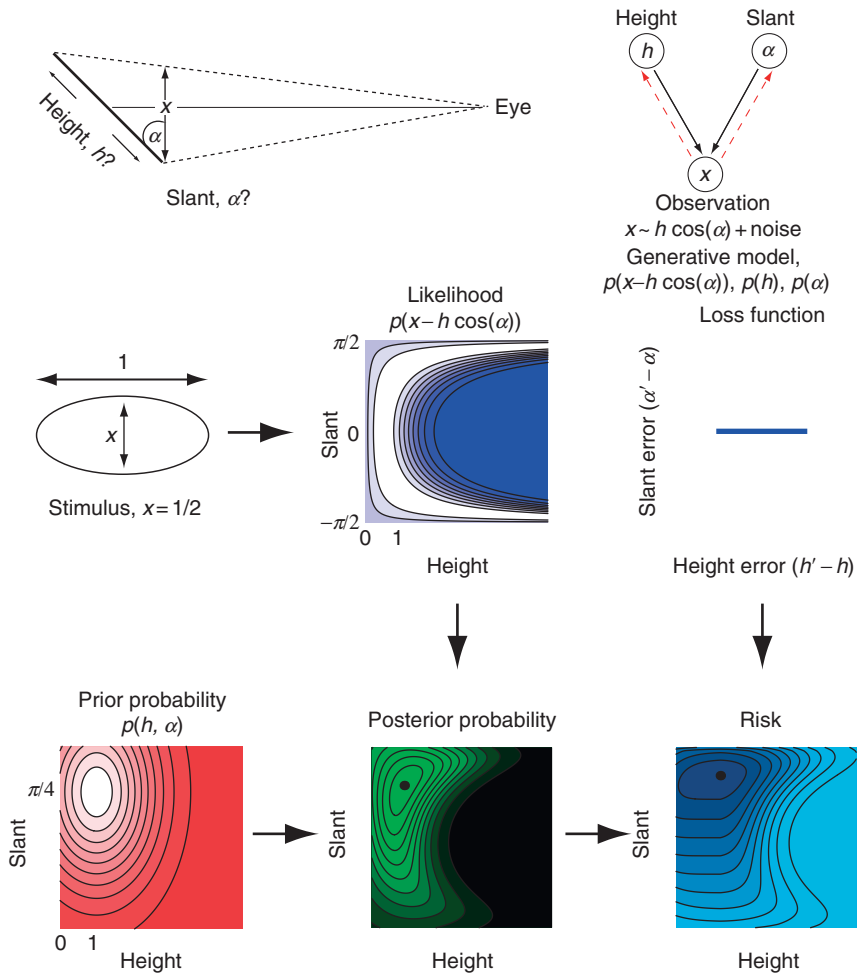

Figure 3 Example of ideal observer theory applied to object perception; a Bayesian ideal observer model for estimating the height and slant of a class of elliptical disks.

This ambiguity can be resolved by multiplying the likelihood by the prior distribution to obtain the posterior distribution, $p(h, \alpha|x)$. At this point, we could stop and pick off the peak values of the posterior as the most probable values of height and slant. This corresponds to the decision rule for the MAP observer. However, suppose the task is more complex, requiring greater precision, say, in estimating the slant than the height.

The optimal decision rule In this more general case, we use a rule that minimizes risk. A simple model of risk is to define loss in terms of the absolute values of the errors, that is, the difference between the true and estimated state variables. We then assign a uniformly low loss to errors of height that fall within a wide range, but penalize slant errors more narrowly, as shown by the loss function in Figure 3. The risk is then a convolution of the loss function with the posterior, $R(h, \alpha|x) = \int p(h', \alpha'|x) L(h - h', \alpha - \alpha') \mathrm{d}h' \mathrm{d}\alpha'$, and the decision rule is to pick the pair of state variables corresponding to the lowest values of R. It is straightforward to show that if the tolerance to errors in height and slant are infinitely wide and narrow, respectively, then height effectively drops out of the risk, and minimizing risk corresponds to the MAP estimation on the marginal posterior, $p(\alpha|x)$.

How to compare test and ideal observers? Even if human observer estimates differ in detail from an ideal observer model, we can still test the degree to which information is optimally combined. Consider the posterior term, which is proportional to the product of the likelihood and prior terms. The ideal observer combines the image cues and the prior information in such a way as to take into account the reliability of each of the sources. Thus, smaller variances on the prior bias the final estimate toward the prior means, and, conversely, highly reliable (e.g., noise-free) image data move the estimates of slant

and/or height toward values consistent with the data. Task differences also affect shifts in the optimal estimates of the state variables. Although this example illustrates the balance between the prior and likelihood, there is a similar trade-off in the case of cue integration. For example, if there is additional information for the slant of the surface (e.g., a texture gradient), then the ideal observer combines and weights the cues according to their reliabilities. For the case of conditional independence and Gaussian noise, the optimal cue weighting is given by $s_{\text{opt}} = [(s_1 r_1)/(r_1 + r_2)] + [(s_2 r_2)/(r_1 + r_2)]$, where s_i is the estimate of the state variable from cue i, and r_i is the reciprocal of the variance for that cue.

Related Areas

Ideal observer theory has close connections to several other areas of neuroscience and cognitive science. Optimal control theory has a similar structure to that illustrated in Figure 1(a). In applications to motor control, the state variables change continuously in time and correspond to internal physical parameters that influence a movement. The Bayes's risk function is replaced by a cost function (e.g., physical energy of a movement), and the action rule is replaced by a control law.

Ideal observer analysis can be naturally extended to optimal learning. Bayes's optimal learning corresponds to updating the parameters (e.g., mean and variance) of the posterior distribution as new data come in. Rather than using the data to infer the causes or states of the generative model, we use Bayes' rule to learn the values of the parameters that determine the posterior.

Computational techniques such as expectation maximization (EM) and Bayesian belief propagation, have provided the means to deal with more complicated and non-Gaussian generative models, with applications to inference and learning. Future work should produce a growing number of applications to the experimental study of cognitive and neural processes.

Comparisons of ideal and human perceptual behavior inevitably lead to the question of the nature of the underlying neural mechanisms that might support ideal-like computation. One of the central issues is how information about uncertainty may be represented and processed in the nervous system.

See also: Active Perception; Bayesian Cortical Models; Bayesian Models of Motor Control; Decision-Making and Vision; Human Methods: Psychophysics; Motor Psychophysics; Neural Integrator Models; Psychophysics of Attention; Statistical Analysis of Visual Perception.

Further Reading

Barlow HB (1962) A method of determining the overall quantum efficiency of visual discriminations. *Journal of Physiology* (*London*) 160: 155–168.

Doya K, Ishii S, Pouget A, and Rao RPN (eds.) (2007) *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT Press.

Duda RO, Hart PE, and Stork DG (2001) *Pattern Classification*, 2nd edn. New York: Wiley.

Eckstein MP (1998) The lower efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science* 9: 111–118.

Ernst MO and Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415(6870): 429–433.

Geisler WS (2003) Ideal observer analysis. In: Chalupa L and Werner J (eds.) *The Visual Neurosciences*, pp. 825–837. Boston: MIT Press.

Green DM and Swets JA (1966) *Signal Detection Theory and Psychophysics*. New York: Wiley.

Griffiths TL and Yuille AL (2006) Technical introduction: A primer on probabilistic inference. UCLA. Department of Statistics Papers no. 2006010103. UCLA, Los Angeles, CA. http://repositories.cdlib.org.

Kersten D, Mamassian P, and Yuille A (2004) Object perception as Bayesian inference. *Annual Review of Psychology* 55: 271–304.

Knill DC and Richards W (1996) *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press.

Körding KP and Wolpert DM (2006) Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* 10(7): 319–326.

Pelli DG, Burns CW, Farell B, and Moore-Page DC (2006) Feature detection and letter identification. *Vision Research* 46(28): 4646–4674.

Tanner WPJ, Birdsall TG, and Clarke FR (1960) The concept of the ideal observer in psychophysics. Technical Report no. 98. Communications and Signal Processing Laboratory (CSPL), University of Michigan, Ann Arbor. http://hdl.handle.net.

Watson AB, Barlow HB, and Robson JG (1983) What does the eye see best? *Nature* 31: 419–422.

Weiss Y, Simoncelli EP, and Adelson EH (2002) Motion illusions as optimal percepts. *Nature Neuroscience* 5(6): 598–604.

Wickens TD (2002) *Elementary Signal Detection Theory*. Oxford: Oxford University Press.

Imaging Studies Using Reporter-Gene Transgenic Rats

A Kimura and E Kobayashi, Jichi Medical University School of Medicine, Tochigi, Japan

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In recent efforts to create specifically designed models for imaging experiments, mouse and rat have been essential. Mouse is an excellent experimental model that has well-developed genetic tools, but rat models also show greater benefit in surgical interventions due to their larger body size. Indeed, surgical manipulation of rats has provided excellent experimental models of neurological disorders, including brain infarction, Parkinson's disease, and spinal cord injury. These rat models have been valuable not only in elucidating disease processes but also in evaluating experimental therapies. We have created genetically engineered rats and tested their application in imaging experiments. Here, we outline the use of reporter gene-transgenic rats in imaging studies for neurological research.

Reporter Genes

There are a number of different methods for labeling cells of interest so that they can be tracked *in vivo*. Labeling with fluorescent dyes is an easy method; however, there is a drawback in that the signal intensity decreases during cell division and/or by photobleaching. In this regard, reporter genes that can be stably integrated into the genome have a distinct advantage over conventional methods, especially in long-term studies.

LacZ (β-Galactosidase)

For over two decades, lacZ has been the reporter gene of choice for a variety of biological experiments. LacZ is a bacterial gene that encodes hydrolase enzyme β-galactosidase, which breaks down the colorless substrate X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) into galactose and a blue insoluble product. LacZ is useful for labeling distinct populations of cells in chimeras or for reporting the activity of specific promoters; however, this reporter has a major disadvantage in that it requires tissue fixation and chemical staining for visualization. Other histochemical marker genes, such as alkaline phosphatase (AP) and β-glucuronidase (GUS), have a similar pitfall.

Green Fluorescent Protein and Spectral Variants

Green fluorescent protein (GFP), a novel fluorescent genetic reporter derived from the bioluminescent jellyfish *Aequorea victoria*, currently represents a unique alternative to other conventional reporter genes. Because wild-type GFP has unsatisfactory properties with regard to fluorescent intensity, folding properties, and biphasic excitation spectrum, many efforts were undertaken to optimize the properties of this protein by introducing point mutations, which finally led to the generation of enhanced green fluorescent protein (EGFP), with brighter fluorescence and a single excitation peak at 488 nm. The direct visualization of EGFP under excitation light has great advantage over conventional markers such as β-galactosidase (LacZ) and alkaline phosphatase, which require tissue fixation and chemical staining for visualization.

The improved properties of EGFP have brought innovations in a variety of biological experiments over the past decade. In principle, application of EGFP can be broadly categorized into two approaches. First, it can be used as a tag in a fusion protein to visualize its localization in time-lapse imaging of living cells. EGFP is usually much less harmful when illuminated in living cells than most small fluorescent molecules such as fluorescein isothiocyanate (FITC), which induces strong phototoxicity caused by the generation of free radicals. Imaging studies using EGFP-tagged proteins have provided valuable information about complex cellular processes that had been analyzed with fixed specimens in the past. Second, it can be used as a genetically encoded reporter gene to identify cells of interest *in vivo*. This latter application has become increasingly popular in imaging studies with the emergence of transgenic animals, in which EGFP is genetically encoded under the control of various promoters. EGFP-transgenic animals can be a source of EGFP-labeled cells or organs that can be transplanted into experimental disease models and be directly visualized under fluorescent light.

Spectral variants of EGFP have been developed by genetic manipulation of the EGFP gene; these include blue fluorescent protein (BFP), cyan fluorescent protein (CFP), and the yellow fluorescent protein (YFP), emitting blue, cyan, and yellow fluorescence, respectively. Furthermore, a red fluorescent protein was obtained by cloning a coral fluorescent protein (*Discosoma* sp.), and this was named DsRed. These spectral variants allowed the tracking of two, three, or even more distinct targets simultaneously.

Luciferase

Luciferases comprise a group of enzymes that emit light in the presence of oxygen and a substrate (luciferin). Such a luciferin–luciferase system is found in

nature, for example, in bacteria (*Vibrio harveyi*), dinoflagellates (*Gonycaulax*), and the firefly (*Photinus pyralis*). These luciferases, in particular the eukaryotic firefly luciferase (Luc), have been commonly used as a light probe in a number of biological experiments, including promoter activity assays. Furthermore, thanks to recent advances in high-sensitivity photon-counting video cameras, a number of studies have employed Luc as a reporter gene in real-time noninvasive *in vivo* imaging. This *in vivo* imaging allows the visualization of the spatial and temporal behavior of Luc-expressing cells in living animals; for example, growth of tumor cells, trafficking of immune cells, and migration of transplanted cells. These imaging studies using Luc-labeled cells have offered new insights that could not have been gained using conventional histological approaches.

Transgenic Animals Carrying a Reporter Gene

Establishment of Transgenic Animals

Transgenic animals are generally established by traditional zygote-injection-based technology or embryonic stem (ES) cell-mediated technology. However, transgenic rat models are established only by the former technology because of a lack of rat ES cell lines. Interestingly, in both technologies, established transgenic animals show variable intensity of reporter expression, depending on the animal lines, organs, phenotypes of cells, and developmental stages. Because the reporter-encoding DNA is randomly integrated into the genomes, the expression of a reporter gene is largely influenced by the integration site. Therefore, careful selection of lines is required to obtain animals with sufficient and stable reporter expression. The variability in gene expression caused by the positional effect of integrated DNA is a weak point of conventional transgenic technology; however, once an animal line with preferential reporter expression is established, the phenotype is stably inherited by the progeny. In addition, the intensity and specificity of reporter expression depends largely on the type of promoter that drives a reporter gene. The promoters are broadly divided into two categories: (1) tissue-specific promoters and (2) ubiquitous promoters.

Tissue-Specific Promoters

In neurological disease research, specifically labeled neurons, glial cells, and immune cells are extremely valuable in elucidating disease processes and in evaluating experimental therapies. Transgenic animals with tissue-specific promoters enable the selective labeling of cells or tissues *in vivo*. In particular, fluorescent proteins driven by the Thy1 promoter specifically label neurons and even individual organelles in neurons; that is, the cytoplasm, membranes, mitochondria, and axon tracts can all be selectively labeled. Specifically labeled transgenic mouse models are further advanced than similar rat models. Thy1-driven fluorescent protein transgenic mice have been used to explore the formation of synapses and the regeneration processes of severed axons. Other cellular elements of the central nervous system, including astrocytes, oligodendrocytes, and microglia, can be selectively labeled with fluorescent proteins under the control of the specific promoters, such as glial fibrillary acidic protein (GFAP) promoter, proteolipid protein (PLP) promoter, and CX_3CR1-locus S100 promoter, respectively. These transgenic mice, combined with emerging optical imaging technologies including multiphoton microscopy, have the potential to elucidate the complicated biological processes occurring in neuroinflammatory lesions.

Ubiquitous Promoters

Housekeeping-type promoters, such as cytomegalovirus (CMV) enhancer/chikin β-actin (CAG), thymidine kinase (TK), Simian virus 40 (SV40), and ROSA26, drive general expression of reporter genes. There are a variety of transgenic lines that express a reporter gene extensively in their organs, driven by ubiquitous promoters. This type of transgenic animal model has been developed not only in mice but also in animals with a larger body size, such as rats, rabbits, and pigs. The mouse represents the premier mammalian model organism; however, its use for biomedical research involving surgical manipulation may be restricted due to limitations imposed by its body size. In contrast, rats are suitable for surgical manipulation because of their larger body size, which also makes it easy to produce disease models and to test various therapeutic interventions in clinical medicine. Therefore, we have focused on establishing transgenic rats with generalized reporter gene expression driven by a ubiquitous promoter. These rats were designed for the use of imaging experiments, especially to trace the fate of cells involved in transplantation therapy. Here, we describe some examples of the use of transgenic rat lines in investigating stem cell-based therapy.

Tracking of Green Fluorescent Protein-Labeled Neural Stem/Progenitor Cells

Transplantation of neural stem/progenitor cells (NSPCs) has emerged as a potential therapeutic strategy for various disorders of the central nervous

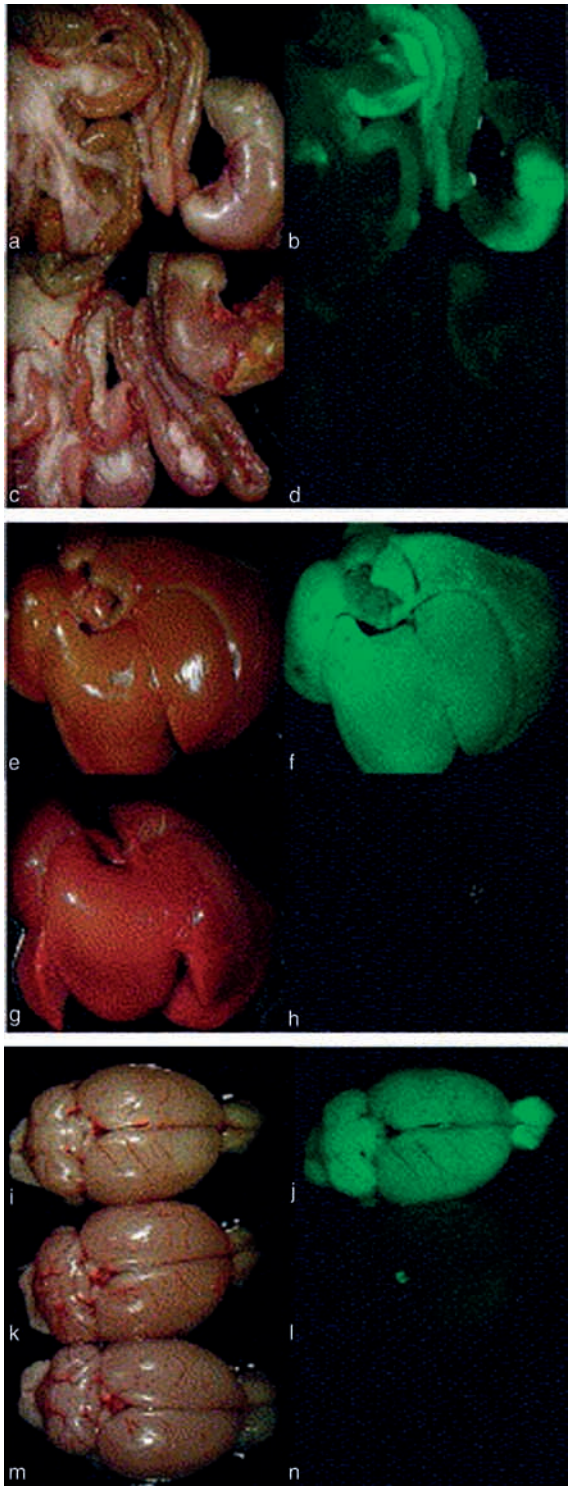

Figure 1 Differential GFP-expression pattern between CAG/GFP-Wistar, CAG/GFP-LEW Tg rats, and wild-type LEW rats: (a–d) small intestine; (e–h) liver; (i–n) brain. Representative organs were removed from CAG/GFP-LEW (panels (a), (b), (e), (f), (i), and (j)) and CAG/GFP-Wistar Tg rats (panels (c), (d), (g), (h), (k), and (l)) and examined under visible (panels (a), (c), (e), (g), (i), (k), and (m)) or 489 nm excitation light (panels (b), (d), (f), (h), (j), (l), and (n)). Panels (m) and (n) are from wild-type LEW rats (as a negative control). CAG, CMV enhancer/chikin β-actin; GFP, green fluorescent protein; LEW, Lewis; Tg, transgenic.

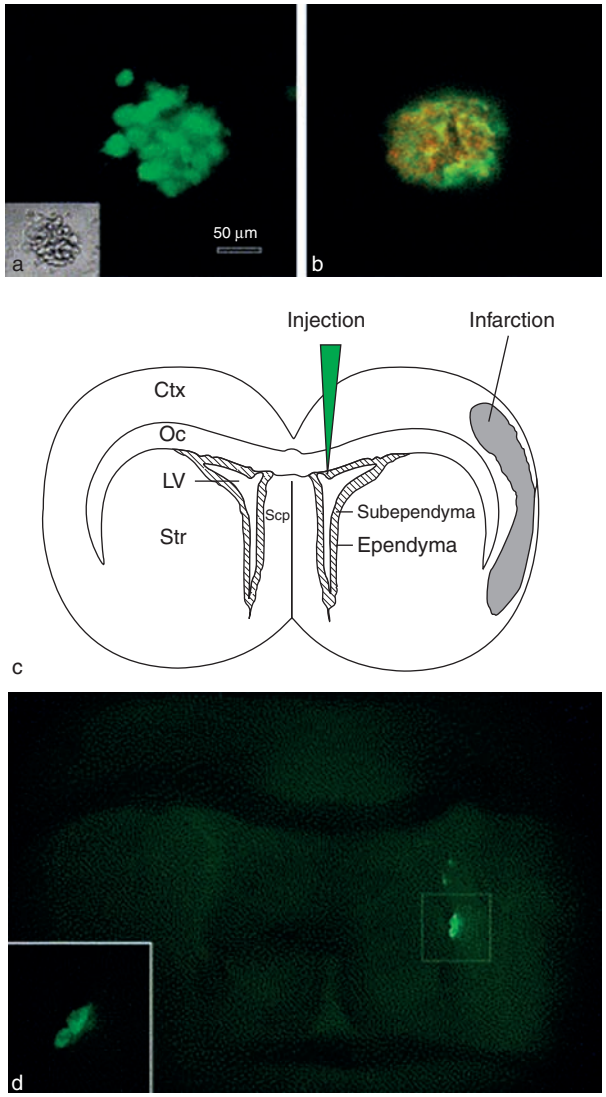

Figure 2 Migration of neural progenitor-derived cells to the cerebral infarction area: (a) neural progenitor cells from CAG/GFP-LEW Tg rats expressing substantial levels of GFP, as seen under 489 nm excitation light 40× magnification); (b) nestin (red color) expressed in neural progenitor cells from CAG/GFP-LEW Tg rats (40× magnification); (c) representative scheme of stereotactic microinjection into the cerebral ventricle; (d) accumulation of GFP^+ neural progenitor-derived cells in the area of the cerebral infarction. CAG, CMV enhancer/chikin β-actin; CC, corpus callosum; Ctx, cortex; GFP, green fluorescent protein; LEW, Lewis; LV, left ventricle; Scp, superior cerebellar peduncle; Str, striatum; Tg, trangenic.

system. Following the fate of transplanted cells (e.g., survival, proliferation, migration, and differentiation) with appropriate markers is indispensable for elucidating uncharacterized regeneration processes involved in stem cell-based therapies. Transgenic animals carrying a fluorescent protein can be an ideal source of genetically labeled NSPCs. Recently, our group established a new GFP-Lewis rat

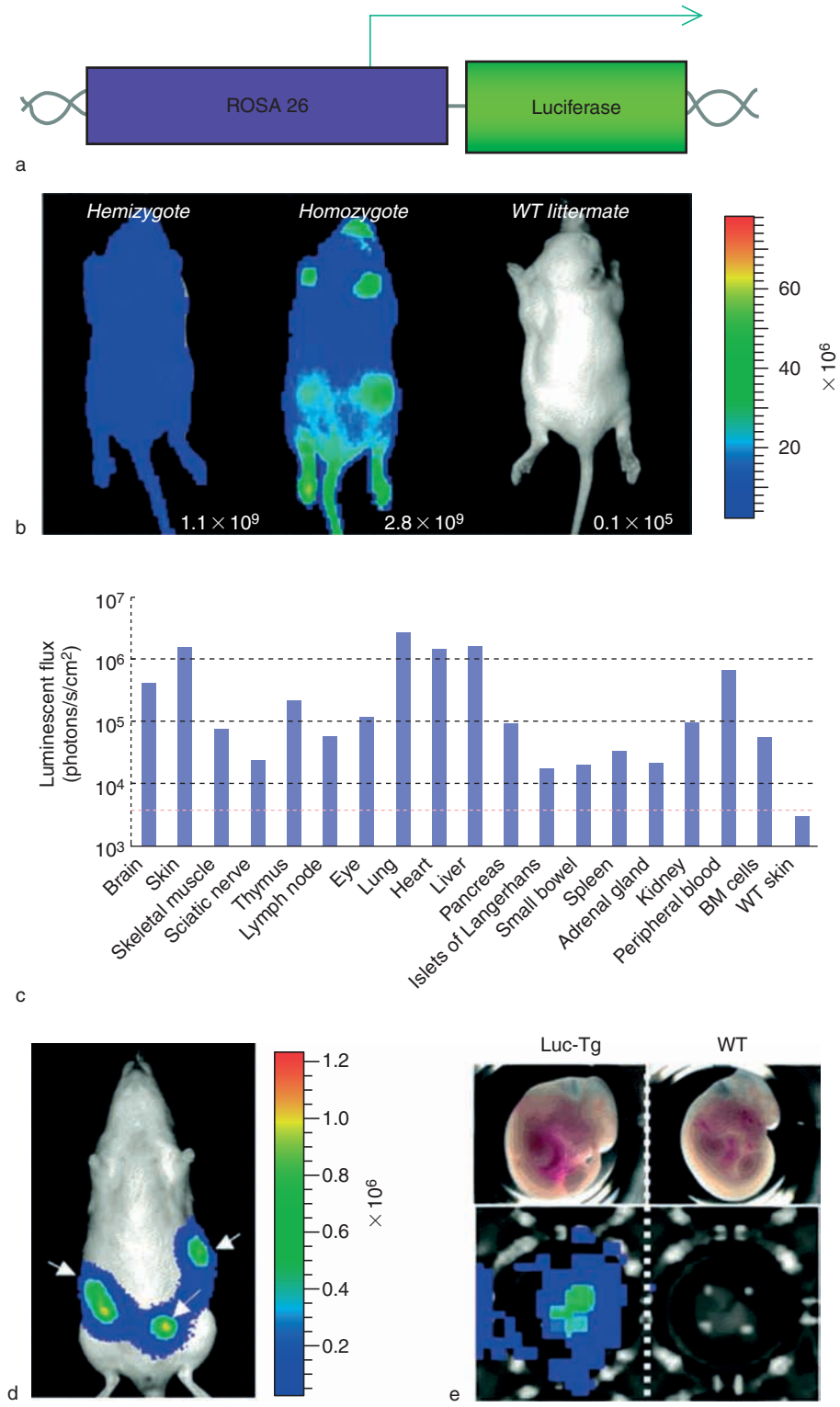

Figure 3 Establishment of Luc-Tg rats: (a) representative scheme of transgene composition; (b) representative luminescent images of Luc-Tg rats (hemizygote, homozygote, and wild-type (WT) littermate); (c) luciferase expression profile in various tissues from the Luc-Tg rats; (d) luciferase activity in Luc-Tg embryos in the pregnant rat; (e) isolated rat embryos emitting luminescent light in the presence of D-luciferin (upper photos: appearance of rat embryos; lower photos: luminescent imaging). As shown in (a), the transgene consists of the ROSA26 promoter and a firefly luciferase gene. In (b), luciferase activity in the Luc-Tg homozygote rat was almost twice that in the hemizygote rat. The values in the right lower corners represent total photons per animal (in photons per second). As shown in (c), luciferase expression is ubiquitous and is particularly strong in the liver, lung, brain, skin, thymus, and eye. The red dotted line represents the background level (WT rat skin). In (d), the WT LEW female rat was mated with the Luc-Tg male rat, and embryos obtained at 14.5 embryonic days were safely visualized using luminescent imaging. Arrows indicate the location of Luc-Tg embryos in a pregnant rat. BM, bone marrow; LEW, Lewis; Luc, luciferase; Tg, trangenic.

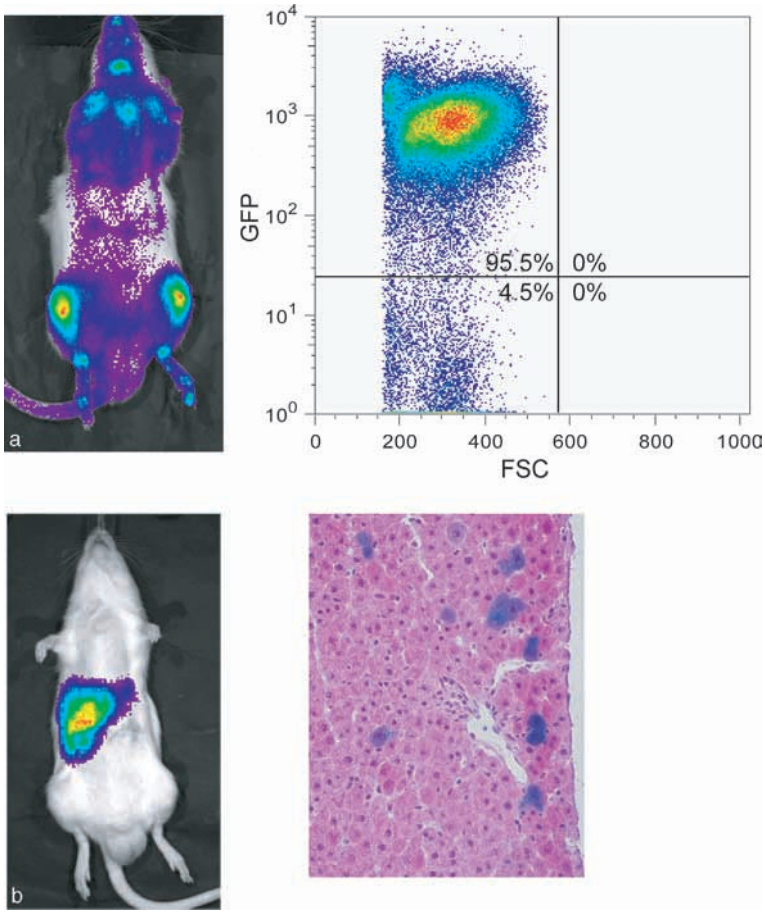

Figure 4 *In vivo* imaging using dual-reporter transgenic rats: (a) BMCs obtained from the (Luc × GFP) F1 rat transplanted into lethally irradiated wild-type Lewis rats; (b) hepatocytes obtained from the (Luc × LacZ) F1 rat transplanted into the portal vein of hepatectonized wild-type Lewis rats. Shown in (a) are a representative scheme of luminescent *in vivo* imaging (left) and flow cytometry analyses of GFP-positive cells in peripheral blood 30 days after transplantation (right). Shown in (b) are a representative scheme of luminescent *in vivo* imaging (left) and X-gal staining of the recipient liver (right). BMCs, bone marrow cells; GFP, green fluorescent protein.

(CAG/EGFP-Lewis rat) that expresses GFP strongly and ubiquitously in most organs, including the central nervous system, compared with the previously established GFP-Wistar rat (CAG/EGFP-Wistar rat) (Figure 1). NSPCs derived from fetal brains of GFP-Lewis rats also expressed high levels of GFP (Figure 2(a)). Most of the GFP^+ NSPCs were positive for the neural stem cell marker Nestin (Figure 2(b)), and they differentiated into neurons, astrocytes, and oligodendrocytes after the removal of growth factor and addition of serum. To investigate the potential of GFP^+ NSPCs to be tracked *in vivo* after cell transplantation, we employed a rat cerebral infarction model. Briefly, a rat with a cerebral infarction received a stereotactic injection of $EGFP^+$ NSPCs into the cerebral ventricle. Four weeks later, the accumulation of $EGFP^+$ NSPCs around the cerebral infarction area was examined (Figure 2(c)). The GFP fluorescence was sufficient to allow visualization of cells without the need for immunohistochemical amplification.

In Vivo Tracking of Luciferase-Labeled Cells

A growing number of studies have employed Luc as a biological light probe for real-time *in vivo* imaging. Luc-labeled cells can be visualized externally in living animals using a luminescent imaging device because luminescence from the Luc–luciferin reaction has significantly greater signal-to-noise ratios in mammalian tissues. *In vivo* imaging was applied to some experimental models of neurological diseases in relation to NSPCs. Previous studies have visualized the recruitment of Luc-labeled NSPCs toward sites of brain infarction and brain tumors in living animals. Accumulation of transplanted Luc-labeled NSPCs in the injured spinal cord was also visualized using time-lapse *in vivo* imaging studies, in which Luc was genetically encoded by a virus-mediated approach.

Recently, our group established a Luc-transgenic rat (ROSA/Luc-Lewis rat) that expresses Luc in

almost all organs, including the central nervous system (Figure 3). NSPCs were cultured from fetal brains of Luc-transgenic rats, and their expression of Luc was verified *in vitro* (data not shown). Therefore, this transgenic rat model, combined with *in vivo* imaging technologies, may be useful for research of neural stem cell-based therapies.

Imaging Studies Using Dual-Reporter Genes

Real-time *in vivo* imaging using the Luc–luciferin system is an attractive tool for a variety of biomedical research, but it has some pitfalls. First, the luminescence signal emitted from a small number of cells cannot be detected externally in living animals, even under highly sensitive imaging devices, whereas the minimum number of cells required for detection is influenced by the intensity of Luc expression. Second, Luc-labeled cells are portrayed as poorly marginated areas in *in vivo* imaging systems. Histological examinations after *in vivo* imaging experiments may partially resolve these problems; however, the enzymatic activity of Luc is lost in fixed specimens. Third, Luc-labeled cells are not appropriate for flow cytometry analysis, in contrast to GFP-labeled cells.

To address these problems, we developed dual-reporter gene transgenic rats. Our transgenic rats, which express single reporter genes such as Luc, LacZ, and GFP, are inbred Lewis rat lines. Thus, we can obtain dual-labeled rats by mating these lines with one another. When Luc-transgenic Lewis rats were mated with GFP-transgenic Lewis rats, the hybrid progeny ((Luc × GFP) F1 rat) represented a source of Luc-GFP-dual-labeled cells. These dual-labeled cells can be applied not only to *in vivo* imaging but also to flow cytometry analysis (Figure 4(a)). In the same way, (Luc × LacZ) hybrid rats provide Luc-LacZ-dual-labeled cells that can be visualized by X-gal staining after the period of *in vivo* imaging (Figure 4(b)).

See also: Alzheimer's Disease: Transgenic Mouse Models; BAC Transgenesis: Cell-Type Specific Expression in the Nervous System; Neural Stem Cells and CNS Diseases; Stem Cells and CNS Repair; Transgenic Models of Neurodegenerative Disease.

Further Reading

Bareyre FM, Kerschensteiner M, Misgeld T, and Sanes JR (2005) Transgenic labeling of the corticospinal tract for monitoring axonal responses to spinal cord injury. *Nature Medicine* 11: 1355–1360.

Francis JS, Olariu A, Kobayashi E, and Leone P (2007) GFP-transgenic Lewis rats as a cell source for oligodendrocyte replacement. *Experimental Neurology* 205(1): 177–189.

Greer LF III and Szalay AA (2002) Imaging of light emission from the expression of luciferases in living cells and organisms: A review. *Luminescence* 17: 43–74.

Hakamata Y, Murakami T, and Kobayashi E (2006) "Firefly rats" as an organ/cellular source for long-term *in vivo* bioluminescent imaging. *Transplantation* 81: 1179–1184.

Hakamata Y, Tahara K, Uchida H, et al. (2001) Green fluorescent protein-transgenic rat: A tool for organ transplantation research. *Biochemical and Biophysical Research Communications* 286: 779–785.

Inoue H, Ohsawa I, Murakami T, et al. (2005) Development of new inbred transgenic strains of rats with LacZ or GFP. *Biochemical and Biophysical Research Communications* 329: 288–295.

Kulbatski I, Mothe AJ, Keating A, Hakamata Y, Kobayashi E, and Tator CH (2007) Oligodendrocytes and radial glia derived from adult rat spinal cord progenitors: morphological and immunocytochemical characterization. *Journal of Histochemistry and Cytochemistry* 55: 209–222.

Misgeld T and Kerschensteiner M (2006) *In vivo* imaging of the diseased nervous system. *Nature Reviews Neuroscience* 7: 449–463.

Mothe AJ, Kulbatski I, van Bendegem RL, et al. (2005) Analysis of green fluorescent protein expression in transgenic rats for tracking transplanted neural stem/progenitor cells. *Journal of Histochemistry and Cytochemistry* 53: 1215–1226.

Murakami T and Kobayashi E (2005) Color-engineered rats and luminescent LacZ imaging: A new platform to visualize biological processes. *Journal of Biomedical Optics* 10: 41204.

Okada S, Ishii K, Yamane J, et al. (2005) *In vivo* imaging of engrafted neural stem cells: Its application in evaluating the optimal timing of transplantation for spinal cord injury. *FASEB Journal* 19: 1839–1841.

Immune Function During Sleep and Sleep Deprivation

J M Mullington, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Context and Questions

Physicians through the ages have recommended rest and sleep to assist in the healing and recovery process. Most of us have experienced a time when we were overcome by sleepiness due to an illness and were unable to function, leaving little alternative but sleep. Once infection is established, it seems that sleep aids in the healing process, but does it actually help to clear the body of pathogens and, if so, how? It is also common experience that when we do not get adequate sleep for a prolonged period, we become 'worn down' and feel susceptible to common colds and other ailments. It is clear that sleep responds to challenges to the host defense system (the body's immune protective and defensive systems), but does sleep actually serve a function in defending against microbial invasion? Are host defenses bolstered in some way by sleep? Do adequate quantity and quality of sleep assist in building defenses against future attacks? Does poor sleep, or going without sufficient sleep for a period, compromise the early host defense system or the development of immunological memory? Does sleep support good health in a more general way, such as supporting homeostatic endocrine and metabolic systems that facilitate a balanced and measured host defense? Host defense mechanisms are highly integrated in physiology, and so the answers to these questions will not be simple. Controlled studies designed to test these intuitive assumptions have been sparse. An overview of the major findings is provided in this article.

Sleep Responds to Experimental Immune Challenge

Animal Models

Studies carried out in many species have shown that animals need sleep in order to survive. If deprived of sleep, rats will die within approximately 3 weeks. In studies in which infectious agents were administered to rabbits, the probability of survival was increased in animals that slept more and whose sleep following challenge showed enhanced slow-wave activity. The animals that were able to increase their slow-wave sleep were not only more likely to survive but also had higher increases over baseline in their white blood cell count (immune participating) and a lower triglyceride response to challenge (a marker of the acute phase response or early immune response to infection or trauma) compared to the animals that died. Sleep, it seems, does help the recovery process.

There are specialized receptors for the outer membrane lipopolysaccharide (LPS) fragment of gram-negative bacteria, called CD14 receptors, that work through a JAK-STAT (Janus kinase and signal transducers and activators of transcription) pathway to activate the production of inflammatory protein mediators such as interleukin-1β (IL-1β), tumor necrosis factor-α (TNF-α), interleukin-6 (IL-6), and others within immune-competent cells such as the monocyte, macrophage, and endothelial cell in the periphery and glial cells in the brain. In animals, levels of IL-1β and TNF-α in the brain vary with the sleep–wake cycle and are increased following host challenge.

Several studies in animal models have shown that IL-1β and TNF-α are key stimulators of the central nervous system (CNS) production of sleep and enhanced slow-wave activity, and they are therefore thought to be important for physiological sleep regulation. If the production of these proinflammatory cytokines is inhibited, the sleep-inducing and -enhancing effects are also inhibited. In fact, blocking TNF-α and IL-1β inhibits normal increases in non-rapid eye movement (NREM) sleep that occur after sleep deprivation. Anti-inflammatory factors such as glucocorticoids or counterinflammatory cytokines such as IL-10 also reduce the normal NREM rebound following sleep deprivation.

The CNS response to immune challenge is not linear. Sleep may aid in recovery in animals experimentally challenged with pathogens, but there are clear dose-dependent effects in the response to challenge. Rabbits and rats show increased NREM sleep despite moderate fever. Beyond a certain level of inflammatory system activation and fever, sleep onset is delayed and/or fragmented. The NREM slow-wave enhancing properties of this CNS response to the acute phase reaction is not simply a by-product of fever because studies that have blocked the fever response to challenge have shown that sleep is nonetheless enhanced. Further evidence comes from species-specific differences – for example, the human sleep response to LPS.

Human Models

As a species, we are exquisitely sensitive to gram-negative LPS and achieve fevers after being challenged with very small amounts of endotoxin. The preparations used for research in humans and those used in animal research are very different. When prepared for human use, endotoxin is stripped

of protein so that real infection does not develop. It is the outer membrane component of the gram-negative bacteria, the LPS fragments, that cause the pyrogenic response. Approximately 1 h after injection, TNF-α is already elevated and temperature begins to increase. Unlike the typical NREM sleep response in rabbits and rats that occurs in the context of moderate fever, humans show increased NREM slow-wave sleep only when the dose of LPS is subpyrogenic and when it is too low to increase cortisol. This subpyrogenic dose of LPS is nonetheless able to cause manyfold increases in key cytokines such as TNF-α and its receptors, p55 and p75. IL-1β is generally not measured in peripherally sampled plasma due to insufficient assay sensitivity. Based on animal studies, TNF effects on sleep appear to be mediated through the TNF type 1 receptor (p55) and not its type 2 receptor (p75).

Effects of Sleep Deprivation on Host Defense

Animal Models

The role of sleep in host defense mechanisms is complex. Although there is ample evidence that sleep is necessary for survival, there is also apparently contrary evidence suggesting it is not so important for adequate host defense. For example, although the likelihood of survival is increased by slow-wave sleep following lethal challenge in rabbits, sleep deprivation in rabbits for 4 h by gentle handling, prior to an endotoxin challenge, leads to higher fevers than found in animals that were permitted to sleep prior to challenge. If animals were similarly deprived after challenge, there were no differences in fever response. In fact, except for the fever response, sleep deprivation did not alter the severity of clinical parameters in response to infection. Blood cells, triglycerides, and stress hormones (cortisol and corticosterone) were not negatively impacted by sleep deprivation. The deprivation used may not have been severe enough to cause the expected changes in host response, and the time window with which to see the changes may not have been ideally placed. However, another study using an influenza challenge also failed to demonstrate a sleep advantage compared with sleep deprivation that preceded and/or followed inoculation. Taking yet another approach, tumors implanted in rats that are subsequently sleep deprived shrink in size, counter to what would be expected if sleep loss was compromising host defense. Furthermore, one study in rats failed to find any positive effects of sleep versus 72 h of sleep deprivation on wound healing. Thus, it seems that sleep deprivation, for at least variations of relatively plausible duration, do little to compromise the host defense response to various forms of challenge. This does not mean that sleep is not involved in supporting host response systems, but it does indicate that the systems function very well and hold up for reasonably long deprivation periods. This is perhaps not surprising, given the central importance of the host defense system for survival and the evolutionary necessity to have the ability to maintain wakefulness for prolonged periods in order to flee danger.

As noted previously, sleep deprivation in animal models does lead to death, usually within 2 or 3 weeks of total sleep deprivation and up to 6 weeks when REM sleep is selectively deprived. Although the ability of animals to withstand the deprivation varies greatly, the pattern of physiological changes is quite consistent. Norepinephrine levels increase early in both total and REM sleep deprivation. Body temperature is increased during the first half of the survival period during total sleep deprivation, but the second half of REM and total sleep deprivation periods are marked by decreases in temperature. Animals undergoing sleep deprivation lose weight despite increased food consumption. This increased energy expenditure and thereby negative energy balance are accompanied by changes in metabolic markers, including decreases in total and free thyroxine and triiodothyronine and reductions in growth hormone, insulin-like growth factor 1, prolactin, and leptin. Fur becomes dull and unkempt in appearance, although grooming time is maintained; tail and paw lesions form.

Leukocytes and some mediators of the immune system (innate and cellular) have been reported to increase under conditions of total sleep deprivation in rats. In addition, consistent with a model of sepsis, bacterial pathogens have been found in organs and tissues of sleep-deprived animals, and these findings have been interpreted by some as evidence of failure of host defense systems. This interpretation is not without opposition, however, because antibiotic treatment is quite successful at controlling aerobic bacteria during deprivation but the animals proceed to accumulate a negative energy balance and die within the same time frame as untreated animals. However, human sepsis, once begun, is often treatment resistant in spite of antibacterial intervention.

In interpreting the findings, it is important to remember that host defense systems are complex and broad in terms of systems mobilized. Sleep supports very complex homeostatic systems, and therefore short-term deprivation and a long-term sleep deficit may have different effects. Very few data are available to enable comparison of the effects of experimental chronic partial sleep deprivation, in which

sleep is reduced to a percentage of normal sleep diet for a prolonged period, and total sleep deprivation, in which no sleep is permitted.

Normal Rhythms of Immune-Active Cells and Modulating Hormones in Humans

A diurnal rhythm in human circulating blood cells has long been established. In Figure 1, the diurnal pattern of white blood cells and cortisol is illustrated for 25 healthy male and female volunteers between the ages of 25 and 53 years. Participants slept between 11:00 p.m. and 7:00 a.m. before and during the blood sampling period. Blood was sampled every 4 h through an indwelling forearm catheter and results are shown for a 36 h period. The clearly established rhythm of cortisol can be seen. The time of day of lowest circulating levels is morning to early afternoon for lymphocytes, monocytes, and eosinophils, and through the sleep period for neutrophils. Several hormones that show diurnal variation have been reported to modulate immune activity, most notably cortisol, growth hormone, and prolactin. Studies have reported a diurnal rhythm in several circulating cytokines as well, with IL-6 consistently showing an early morning nadir and a nocturnal peak. TNF-α and its receptors have also been reported to have diurnal rhythms, but the peaks and troughs of these rhythms are less consistent across studies.

Cortisol is a clear modulator of immune response. It is used at pharmacological doses to suppress immune function, and at physiological levels it participates in the modulation of the early host response to endotoxin challenge in humans.

Sleep Deprivation in Humans

An interplay of balance between cellular and humoral immune system responding has been described in which type 1 responses are predominantly cellular and type 2 predominantly humoral. The hormonal and cytokine milieu associated with healthy sleep supports growth, probably including development of immunological memory because monocyte production of IL-12, a key cytokine involved in the development of cellular (type 1) memory, is increased during sleep in humans.

Total and partial sleep deprivation studies have been carried out investigating levels of circulating white blood cells and inflammatory mediators. In humans, total sleep deprivation studies have found monocytes and neutrophils to be increased and lymphocytes decreased. Studies of prolonged partial sleep deprivation have also found increased white blood cells due to deprivation. Both total and partial sleep deprivation studies have found that circulating levels of IL-6 and C-reactive protein (RCP) are increased in healthy volunteers. The soluble TNF receptor 1 has been shown to increase during acute total sleep deprivation. These changes may indeed be related to autonomic changes because some studies have shown norepinephrine to be increased in sleep deprivation,

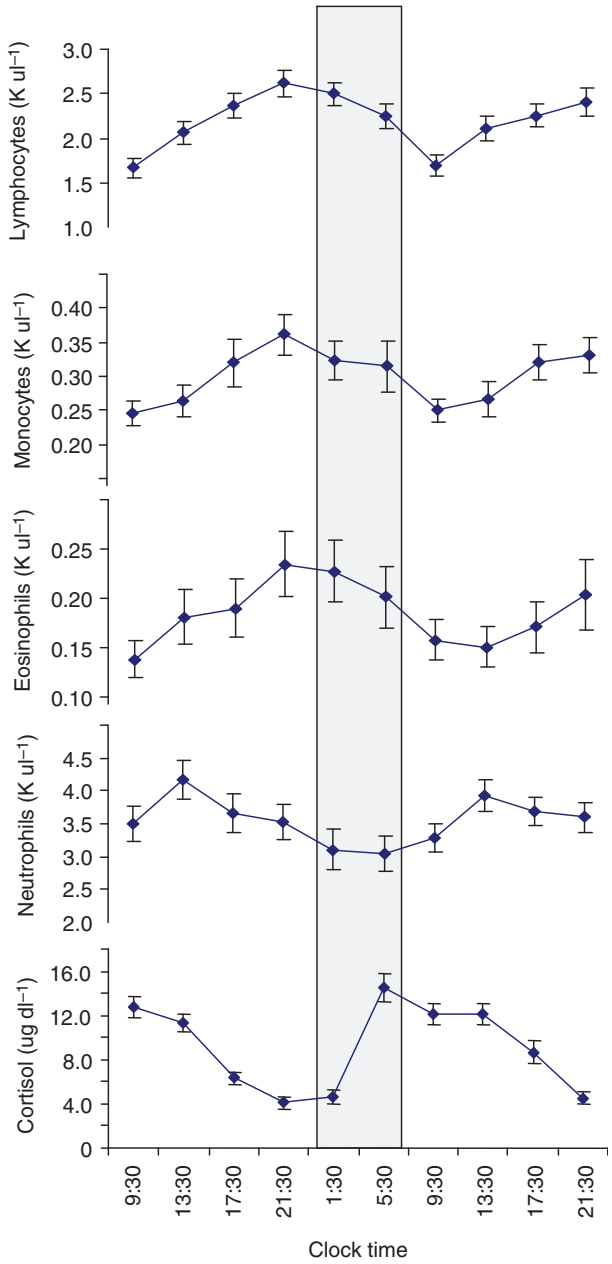

Figure 1 Diurnal rhythm in cortisol and white blood cells. Twenty-five healthy men and women slept between 11:00 p.m. and 7:00 a.m., and blood was drawn through intravenous forearm catheter at 4 h intervals for 36 h. The shaded bar illustrates the time of the sleep period, and the 4 h measurements are shown with standard error bars.

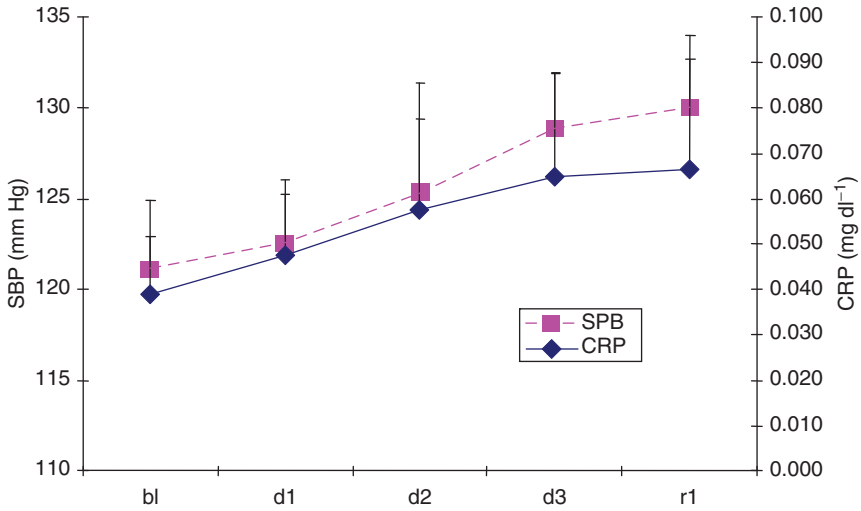

Figure 2 Effects of acute sleep deprivation on C-reactive protein (CRP) and systolic blood pressure (SBP). The average data are shown for the baseline (bl) day, days 1–3 of wakefulness (d1–d3; 88 h total), and the recovery day (r1). Data from Meier-Ewert HK, Ridker PM, Rifai N, et al. (2004) Effect of sleep loss on C-reactive protein, an inflammatory marker of cardiovascular risk. *Journal of the American College of Cardiology* 43: 678–683.

and several studies have shown blood pressure to be increased during sleep deprivation. As can be seen in Figure 2, systolic blood pressure has been found to increase with a time course similar to that of CRP. Although the IL-6 and CRP elevations are small, they may be very important in long-term health.

IL-6 and CRP are important acute phase markers of the innate immune system (the type of immunity that humans have without requiring exposure and development of immunological memory) known to be elevated in overweight individuals and which have been linked to development of future cardiovascular disease and diabetes, even independent of their relationship with adiposity. It is important to understand that even very small subclinical elevations in inflammatory mediators have been linked with future risk for the development of disease. It has only been possible to detect these small basal shifts with improvements in assay technology that were made about a decade ago. For this reason, epidemiological studies have only recently been uncovering important predictive relationships that not long ago were undetectable. What used to be thought of as subclinical from the perspective of the acute phase response is now understood to be meaningful from the perspective of understanding risk and disease vulnerability. Do these shifts in inflammatory mediators seen during sleep deprivation translate into increased risk for future disease development? It is too early to know. However, epidemiological studies show that there are risks associated with habitual short sleep duration.

Epidemiological studies show that individuals who sleep less than 6 h per night are at increased risk for developing future cardiovascular disease and diabetes, independent of other known risk factors such as body mass index. Individuals with sleep-disordered breathing have higher levels of inflammatory mediators, including IL-6 and CRP. Of particular importance in this regard, healthy asymptomatic individuals with very low-level basal increases in CRP have a higher likelihood of developing future cardiovascular disease. This association has also been described for other proinflammatory cytokines such as IL-6, but CRP is particularly useful to examine in population-based studies because it is quite stable and has a long half-life (between 15 and 19 h) and has been shown not to have a diurnal rhythm.

Individuals with disturbed sleep due to disorders such as sleep apnea have been reported to show elevated basal levels of proinflammatory cytokines. It is less clear whether treatment of these disorders resolves the elevated basal inflammatory mediators and to what extent these changes are related to metabolic changes. However, sleep deprivation alters basal metabolic function, which in itself is likely to contribute to changes in immunological homeostasis. Sleep plays a key role in maintaining homeostasis of physiological systems – most important, metabolic and endocrine systems. It is reassuring that host defense mechanisms are intact in the face of acute short-lasting deprivation, but in the long term, sleep is essential for life and both chronic partial sleep deprivation and severe acute sleep deprivation are associated with increased inflammation. Elevations in inflammatory mediators are associated with the development of many diseases, and further investigations are needed to determine the contribution of insufficient sleep to their etiology.

See also: Autonomic Dysregulation During REM Sleep; Endocrine Function During Sleep and Sleep Deprivation; Metabolic Syndrome and Sleep; Sleep Apnea; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Born J, Lange T, Hansen K, Mölle M, and Fehm LH (1997) Effects of sleep and circadian rhythm on human circulating immune cells. *Journal of Immunology* 158: 4454–4464.

Dinges DF, Douglas SD, Zuagg L, et al. (1994) Leukocytosis and natural killer cell function parallel neurobehavioral fatigue induced by 64 hours of sleep deprivation. *Journal of Clinical Investigation* 93: 1930–1939.

Everson CA (2005) Clinical assessment of blood leukocytes, serum cytokines, and serum immunoglobulins as responses to sleep deprivation in laboratory rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 289: R1054–R1063.

Everson CA and Toth LA (2000) Systemic bacterial invasion induced by sleep deprivation. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 278: R905–R916.

Kuhn E, Brodan V, Brodanova M, and Rysanek K (1969) Metabolic reflection of sleep deprivation. *Activitas Nervosa Superior* 11: 165–175.

Meier-Ewert HK, Ridker PM, Rifai N, et al. (2004) Effect of sleep loss on C-reactive protein, an inflammatory marker of cardiovascular risk. *Journal of the American College of Cardiology* 43: 678–683.

Obal FJ Jr. and Krueger JM (2003) Biochemical regulation of sleep. *Frontiers in Bioscience* 8: 520–550.

Pollmächer T, Mullington J, Korth C, et al. (1996) Diurnal variations in the human host response to endotoxin. *Journal of Infectious Diseases* 174: 1040–1045.

Rechtschaffen A and Bergmann BM (2002) Sleep deprivation in the rat: An update of the 1989 paper. *Sleep* 25: 18–24.

Rechtschaffen A, Bergmann BM, Everson CA, Kushida CA, and Gilliand MA (1989) Sleep deprivation in the rat: X. Integration and discussion of the findings. *Sleep* 12: 68–87.

Renegar KB, Floyd RA, and Krueger JM (1998) Effects of short-term sleep deprivation on murine immunity to influenza virus in young adult and senescent mice. *Sleep* 21: 241–248.

Shearer WT, Reuben JM, Mullington JM, et al. (2001) Soluble TNF-alpha receptor 1 and IL-6 plasma levels in humans subjected to the sleep deprivation model of spaceflight. *Journal of Allergy and Clinical Immunology* 107: 165–170.

Toth LA, Opp MR, and Mao L (1995) Somnogenic effects of sleep deprivation and *Escherichia coli* inoculation in rabbits. *Journal of Sleep Research* 4: 30–40.

Toth LA, Tolley EA, and Krueger JM (1993) Sleep as a prognostic indicator during infectious disease in rabbits. *Proceedings of the Society for Experimental Biology and Medicine* 203: 179–192.

Immune System–Neuroendocrine Interactions

S Rivest, Laval University, Québec City, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Inflammation is a general name for reactions occurring after most kinds of tissue injuries or infections or immunologic stimulations as a host defense against foreign or altered endogenous substances. The local inflammatory reaction is characterized by an initial increase in blood flow to the site of injury, enhanced vascular permeability, and selective accumulation of different effector cells from the peripheral blood to injured regions. These cells, mostly circulating neutrophils and monocytes and locally resident macrophages, together mount a rapid inflammatory response that is characterized by the secretion of cytokines, among other features (Figure 1). Their secretion into the bloodstream is a key step in triggering the neuronal activity and subsequent neurophysiological responses that take place during systemic and localized tissue insults. Cytokines influence many neuroendocrine systems, the most prominent of which is the hypothalamic–pituitary–adrenal (HPA) axis, activation of which results in the release of adrenocorticotrophic hormone (ACTH) and glucocorticoids. Once present in the bloodstream, these cytokines also have the ability to increase body temperature (fever) and cause symptoms of sickness. The proinflammatory cytokine interleukin-1 (IL-1), especially its β form (IL-1β), is probably the most important molecule capable of modulating cerebral functions during systemic and localized inflammation. Systemic IL-β injection activates the neurons involved in the control of autonomic functions, and neutralizing antibodies or IL-1 receptor antagonists are capable of preventing numerous responses to inflammatory stimuli. Studies in IL-1β-deficient mice have provided solid and compelling evidence supporting the critical role played by this cytokine in neural–immune interaction.

Other cytokines implicated in neuroendocrine and febrile responses include tumor necrosis factor-α (TNF-α) and IL-6. Like IL-1β, intravenous TNF-α injection has a profound influence on the brain, on fever, and on hormones of the HPA axis. In contrast, the role of IL-6 in neuroendocrine and febrile responses is quite controversial. From the observation that IL-6-deficient mice were unable to develop normal fever in response to lipopolysaccharide (LPS) and IL-1β, IL-6 has been defined as one of the principle endogenous pyrogens. It has also been demonstrated that prostaglandins (PGs) mediate IL-6-induced fever and activation of the HPA axis, but IL-6 is unable to stimulate PG formation in the brain. However, a specific antiserum to IL-6, IL-6AS, was able to prevent LPS-induced febrile response, and cyclooxygenase 2 (COX-2)-deficient mice do not develop fever in response to icv IL-6 injection. The mechanisms by which IL-1β and IL-6 activate the autonomic functions are likely to be quite different, and IL-1β is clearly more potent than IL-6 in triggering PGE_2 synthesis in the central nervous system (CNS).

Possible Pathways by Which Circulating IL-1β Signals the Brain

The transport of soluble substances across vascular compartments occurs via either paracellular or transcellular mechanisms. Within the cerebrovasculature, the paracellular route is particularly impermeable because of the blood–brain barrier (BBB). The BBB consists primarily of nonfenestrated endothelial cells that are connected by tight junctions. Furthermore, the large molecular sizes (13–15 kDa) and hydrophilic nature of cytokines preclude their transcellular movement by simple diffusion to any appreciable extent, and early studies concluded that the BBB was impermeable to IL-1β. It is nevertheless possible that endogenous cytokines diffuse across the BBB during extreme periods of fever or in presence of high levels in plasma during a long period of time.

Regions relatively devoid of BBB permit cytokine interaction with the neuronal elements. The circumventricular organs (CVOs) contain capillaries with rather greater permeability than the rest of the CNS, and the vascular density in these regions is extraordinarily high. Therefore, the CVOs have been proposed as potential sites of action, in particular the structure lining the anteroventral third ventricle region, namely the organum vasculosum of the lamina terminalis. The subfornical organ, median eminence, and area postrema are the other potential target CVOs. On the other hand, a medullary group of cells may play an important role in processing visceral sensory information carried out by the vagus and glossopharyngeal nerves. The vagus nerve has indeed been proposed to provide a rapid communication pathway for cytokine signaling between the periphery and brain, but this pathway depends on dosage and route of administration.

On the other hand, a strong body of evidence now supports the concept that cytokines, particularly IL-1, act on the microvasculature, with consequent release of local signaling molecules, namely PGs. Indeed, brain microvessels exhibit constitutive expression levels of IL-1 type 1 receptor transcript, and systemic

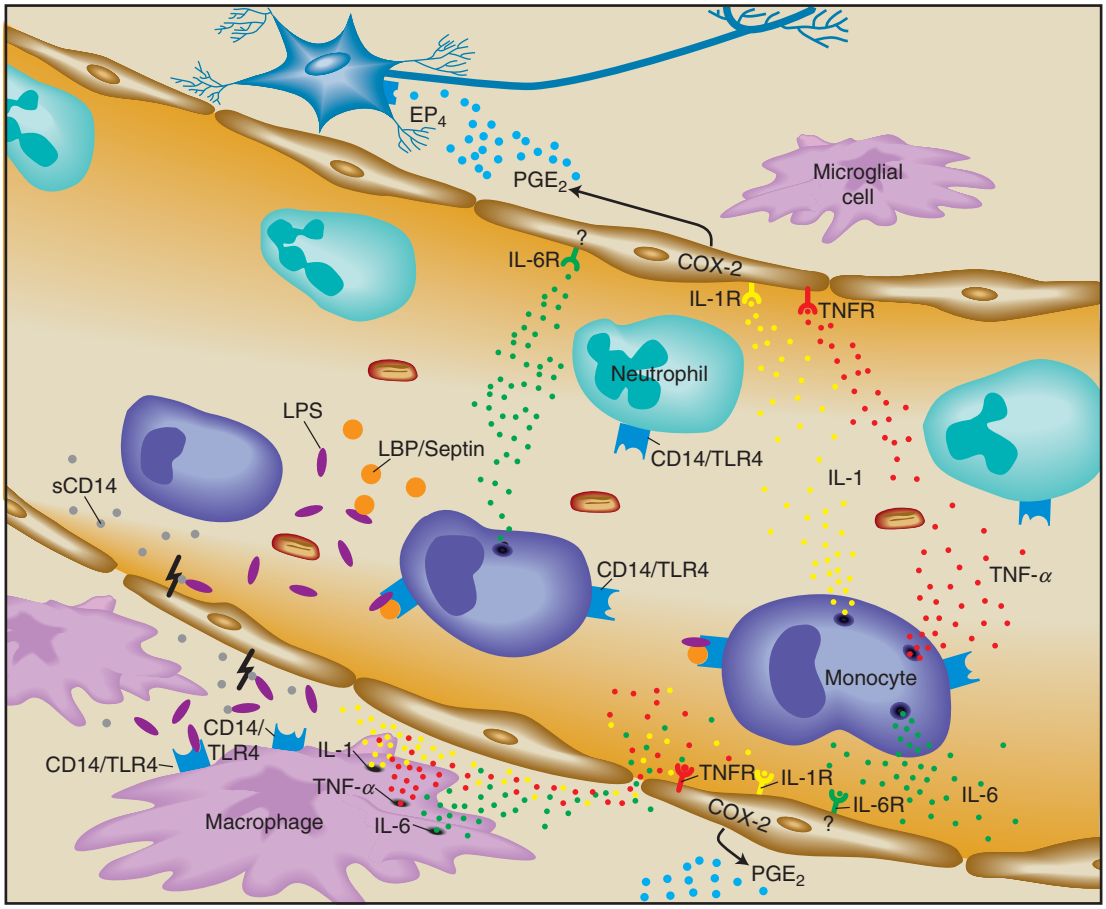

Figure 1 Innate immune response evoked by endotoxic lipopolysaccharide (LPS). LPS is a major component of the outer membranes of Gram-negative bacteria, which is the best-characterized example of innate recognition associated with a robust inflammatory response by phagocytic cells. Secretion of cytokines by circulating monocytes/neutrophils and tissue macrophages by LPS requires a series of mechanisms in cascade, the first step being the binding of the LPS with the serum proteins LPS-binding protein (LBP) or septins. The newly formed complex may then activate different populations of cells by binding to its CD14 receptor and toll-like receptor 4 (TLR4). The latter is the actual signaling receptor for LPS. Two forms of CD14 receptors can be found. The first is present on the surface of myeloid cells (mCD14) and acts as a glycosyl-phosphatidylinositol (GPI)-anchored membrane glycoprotein; the other form is soluble in the serum (sCD14) and lacks the GPI properties, though it can bind LPS to activate cells devoid of mCD14, such as endothelial cells. LBP is not essential for LPS signaling, but the LPS–LBP complex is particularly powerful in activating cells of myeloid origin, that is, neutrophils, monocytes, macrophages, and microglia. One of the well-known consequences of such activation is the production of proinflammatory cytokines, such as interleukin (IL)-1β (IL-1β), tumor necrosis factor alpha (TNF-α), and IL-6. These cytokines may, in turn, bind to their cognate receptors expressed on the surface of cells forming the blood–brain barrier, but they are not essential for mediating the effects of LPS in the central nervous system. LPS is an exogenous ligand for cerebral tissue that expresses both LPS receptors. Such an innate immune/inflammatory response also takes place during tissue injuries and in presence of more complex pathogens (e.g., bacteria and viruses). See text for details. COX-2, cyclooxygenase 2; EP_4, a prostaglandin E receptor; PGE2, prostaglandin E_2; R, receptor.

IL-1β injection causes a rapid and profound transcriptional activation of the gene encoding COX-2 in cells lining the CNS blood vessels. Circulating IL-1 also induces microsomal PGE synthase (mPGES-1), which co-localizes with COX-2 in the perinuclear region of the cerebral endothelium. Of great interest are the data that mPGES-1-deficient mice failed to develop fever in response to either systemic LPS and IL-1, but not to centrally administered PGE_2. These data clearly underline the critical role of the PGE_2 machinery in the endothelium of the brain's blood vessels to activate the autonomic functions during systemic inflammation.

Biosynthesis of PGs

Prostanoids are a group of 20 carbon-unsaturated fatty acid derivates that are produced via a complex enzymatic cascade. First, COX activity adds molecular oxygen to the unsaturated fatty acid arachidonic acid, generating PGG_2. PGG_2 is then converted to PGH_2 by the peroxidase activity of the enzyme.

Once generated, PGH_2 is rapidly converted to PGs (PGD_2, PGE_2, and $PGF_{2\alpha}$), prostacylin, and thromboxane A_2 by tissue-specific synthases. PGE synthase has recently been identified and designated as mPGES and cytosolic PGES (cPGES). These molecules or their derivatives interact with specific receptors to modulate cell function. The diversity of the tissue-specific synthases and receptors gives rise to a wide range of potential biologic functions for the prostanoids. PGG, PGH, prostacyclin, and thromboxane A are chemically unstable and are degraded into inactive products under physiological conditions, with a half-life of 30 s to a few minutes. Other PGs, although chemically stable, are metabolized quickly. It is therefore believed that prostanoids work locally, acting only in the vicinity of the site of production to serve as potent autocrine and paracrine mediators in a wide variety of physiological processes. Although PGE_2 rises in blood promptly after the entry of microorganisms or in response to the endotoxin LPS and cytokines, it is now generally accepted that the PGE_2 detected in the brain is not derived from the blood but rather is produced directly within the CNS.

It is therefore proposed that circulating LPS and cytokines bind to their cognate receptors onto endothelial and/or monocytic cells lining the BBB, which leads to proinflammatory signaling and transcription of the enzymes responsible of the PGE_2 formation in the cerebral tissue. It is interesting to note that systemic inflammatory insults induce COX-2 and mPGES in a rather unspecific manner across the cerebral blood vessels and small capillaries, whereas the neuronal activity is limited to selective nuclei, including the endocrine hypothalamus. It is thus possible that expression of specific PGE_2 receptors within parenchymal cells adjacent to the site of production determines the action of the PG in the brain.

PGE_2 Sites of Action

Classic prostanoid receptors comprise a family of eight encoding transmembrane G-protein-coupled receptors. These receptors are classified on the basis of selective affinities for naturally occurring prostanoids. There are distinct receptors for thromboxane A_2, prostacylin, $PGF_{2\alpha}$, and PGD_2 and four different receptors for PGE_2 (EP_{1-4}). Multiple alternatively spliced isoforms exist for the PGE_2 EP_3 receptor ($EP_{3\alpha}$, $_{3\beta}$, $_{3\gamma}$). They share common extracellular and membrane-spanning regions but differ in intracellular and C-terminal domains. Each receptor is associated with a unique G-protein and consequently a unique second messenger system, namely, elevation of intracellular Ca^{2+} (EP_1) and stimulation (EP_2 and EP_4) or inhibition of (EP_3) of adenylate cyclase. Despite the presence of some conserved sequences, overall homology among the prostanoid receptors is quite limited, ranging from 20–30%. On the other hand, the homology of a given type or subtype of receptor among various species is considerably higher. Each of the eight types and subtypes of receptors shows selective ligand binding specificity that distinguishes it from the others. In addition to transmembrane receptors, the peroxisome proliferator-activated receptor-γ is a member of the nuclear receptor family of transcription factors that can be activated by binding to PGD derivates, such as 15-deoxy-Δ12,14 PG J_2.

In 1988, the distribution of [^{3}H] PGE_2-binding sites, presumably PGE_2 receptors, was first demonstrated in the monkey diencephalon, and more-detailed analysis of [^{3}H] PGE_2-binding sites in rat brain ensued. PGE_2-binding sites were located in a number of discrete brain regions, including thalamic and hypothalamic nuclei, ventral hippocampus, central gray, superior colliculus, parabrachial nucleus, locus coeruleus, raphe nuclei, spinal trigeminal nuclei, and nucleus of the solitary tract (NTS). *In situ* hybridization was thereafter used to determine the exact distribution of each PGE_2 receptor subtype. The hypothalamic paraventricular (PVN) and supraoptic nuclei exhibited EP_1-expressing cells, but a wide distribution was found for the gene encoding the EP_3 subtype. Investigators have reported a very distinct pattern of EP_2- and EP_4-expressing neurons throughout rat brain; that is, EP_2 receptor messenger RNA (mRNA) was detected in the bed nucleus of the stria terminalis, lateral septum, subfornical organ, ventromedial hypothalamic nucleus, central nucleus of the amygdala, locus coeruleus, and area postrema, while EP_4 receptor transcript was located mainly in regions involved in the control of neuroendocrine and autonomic activities. Moderate doses of LPS or IL-1β activate EP_4 neurons, and this activation is prevented when animals are pretreated with COX inhibitors.

All four PGE_2 EP receptor mRNAs are expressed in the anteromedial preoptic region that plays a crucial role in the febrile response. Among these, only EP_4 receptor mRNA is strongly expressed throughout the PGE_2-sensitive regions, including the organum vasculosum of the lamina terminalis, ventromedial preoptic nucleus (VmPO), and median preoptic nucleus. These EP_4-expressing neurons are also activated by systemic inflammation, whereas EP_2- and EP_3-positive neurons do not respond. EP_1 receptor mRNA is present in PGE_2-sensitive regions, but its expression level is weak. Thus, EP_4 may be the key binding and functional receptor for PGE_2 in the brain to activate the circuits involved in autonomic control.

Nevertheless, pharmacological and genetic mutation experiments suggest otherwise. Drugs with

agonist and antagonist properties for each EP receptor have been used in rats. Injection icv of 17-phenyl-ω-trinor-PGE_2 (an EP_1 and $EP_{3\alpha}$ receptor agonist), but not butaprost, M&B28767, or 11-deoxy-PGE_1 (EP_2, $EP_{3\alpha}$, and EP_4 receptor agonists, respectively), induced fever, and SC19220 (an EP_1 receptor antagonist) prevented the febrile response to PGE_2. Another pharmacological study indicated that EP_2 or EP_3 receptor might be the receptors necessary to produce fever. In mice bearing genetic deletions of the EP_{1-4} receptors, only EP_3-knockout animals failed to show the early phase of fever (up to 1 h) after LPS iv or PGE_2 icv. It is important to note, however, that EP_4-deficient mice do not survive and have to be intercrossed in a different background. It is therefore quite difficult to compare them with the other gene-deficient mice, and inducible EP_4 knockout mice will be essential to clearly define the role of this receptor in the brain of mature animals. The relative lack of specificity of EP_4 antagonists and agonists in the species studied may also explain the pharmacological data. It is nevertheless possible that the anatomy and circuits unraveled by functional indices of neuronal activity have little to do with the physiological outcomes, namely, fever and increased activity of the HPA axis.

Functional Circuits by Intraparenchymal PGE_2

Without such anatomical work, however, it is not possible to clarify the exact pathways and groups of neurons involved in neural–immune interaction. Indeed, the final picture is far from complete. The following cascade of events is based on the data described above, but a few steps still remain to be fully validated and should therefore be considered working hypotheses:

1. Endotoxin LPS reaches the bloodstream and binds to the serum protein LPS-binding protein (LBP). The newly formed complex then binds to membrane CD14 and TLR4 receptors located on the surface of myeloid cells, triggering NF-κB signaling, nuclear translocation, and cytokine gene transcription (Figure 1). Pro-IL-1β and-TNF-α have to be cleaved by specific cytosolic enzymes before being transported into the extracellular environment. Recent data suggest a critical role of extracellular adenosine triphosphate and the P2X7 receptor in this final secretion process. Once in the bloodstream, both IL-1β and TNF-α can act on distant organs to trigger different physiological responses by binding to their respective type I receptors.

2. The endothelial and perivascular cells are the most obvious target cells in the CNS because they constitutively express numerous cytokine receptors and they are highly sensitive to low circulating levels of proinflammatory molecules. It is important to note that circulating LPS is also a direct ligand for these cells and that the release of circulating cytokines is not a prerequisite for LPS activation of the endothelial and perivascular cells lining the BBB.

3. Circulating LPS/LBP, IL-1, and TNF initiate COX-2 and mPGES transcriptions through a NF-κB and mitogen-activated protein kinase (MAPK) pathway, which is activated within seconds on ligand–receptor interaction. This leads to PGE_2 synthesis and release by the endothelium and pericytes of the brain capillaries. The PG may therefore diffuse across the brain parenchymal elements and bind to specific EP receptors at the surface of the neurons controlling the autonomic functions.

4. Whether one or more EP receptors participate in the neuronal response to PGE_2 is still a matter of great debate, but the neuroanatomy weighs in favor of the EP_4 receptor subtype. This receptor is found in key regions that control fever, HPA axis activity, and other autonomic functions. EP_4-expressing neurons are activated (as revealed by spontaneous induction of the immediate-early gene c-*fos*) in response to centrally injected PGE_2 and systemically administered IL-1 and LPS. Selective and nonselective COX inhibitors have the ability to prevent such functional activation of these neurons. EP_4-expressing neurons display an integrated circuit associated with the control of autonomic functions. The EP_4 receptor is associated with G_s-protein and stimulation of cyclic adenosine monophosphate (cAMP), which is the most potent transduction pathway involved in the release of corticotropin-releasing factor (CRF) from the PVN and increase in HPA axis activity. There is also evidence in support of cAMP signaling in thermosensitive neurons of the VmPO.

5. On the other hand, PGE_2 may inhibit major inhibitory neurons (e.g., γ-aminobutyric acid (GABA)) projecting to the nuclei that control the autonomic circuits. This would explain the data that only EP_3-deficient mice fail to exhibit a febrile response to inflammatory stimuli. This particular receptor is associated with a G_i-coupled protein inhibiting cAMP on its activation, which may take place in GABA neurons that innervate the VmPO and PVN. Obviously, these neurons would not express c-*fos* or other indices of neuronal activation in response to PGE_2. Functional anatomical data supporting this concept are nevertheless lacking, because there are as yet no reliable tools to assess inhibition rather than stimulation *in vivo*. If interaction of PGE_2 with its EP_3 receptor blocks the activity of GABA neurons innervating the VmPO and PVN, then the spontaneous induction of immediate-early genes in these regions may be indirectly attributable to GABA inhibition rather than to a direct interaction of PGE_2 with its EP_4 receptor. This is a chicken-and-egg problem that will have to be resolved with appropriate tools and animal models.

Alternative Pathways

PGE_2 production in the CNS may rely on two different pathways: (1) a fast transient production in endothelial cells and CVOs dependent on activation (possibly through MAPK p38 phosphorylation) of available constitutive PGE_2 synthesis enzymes ($cPLA_2$, COX-1, cPGES, and mPGES-2), independent of transcription and NF-κB activation (Figure 2), or (2) a slow, sustained PGE_2 synthesis in the endothelial cells and perivascular microglia (equipped to respond to cytokines and pathogens) involving the robust induction of the principle enzymes involved in PGE_2 synthesis (especially secretory phospholipase A_2 ($sPLA_2$-IIA), COX-2, and mPGES-1) that are NF-κB- and MAPK-mediated gene products. The slow transcription-mediated response would be massive but could offer a way through which glucocorticoids would then exert their negative feedback on NF-κB activity to shut down the CNS production of PGE_2. The fast response could be involved in priming the later delayed response and prevent drastic changes to homeostatic processes that the delayed response would have problems overcoming. A study looking at the effects of specific COX-1 and COX-2 inhibitors supports such a role for the COX-1-dependent early response since the ablation of COX-1 activity following LPS administration led to a fast hypothermic response, which could not be overcome by the later COX-2-dependent fever response. This study emphasizes the importance of the coordination between the fast and delayed responses.

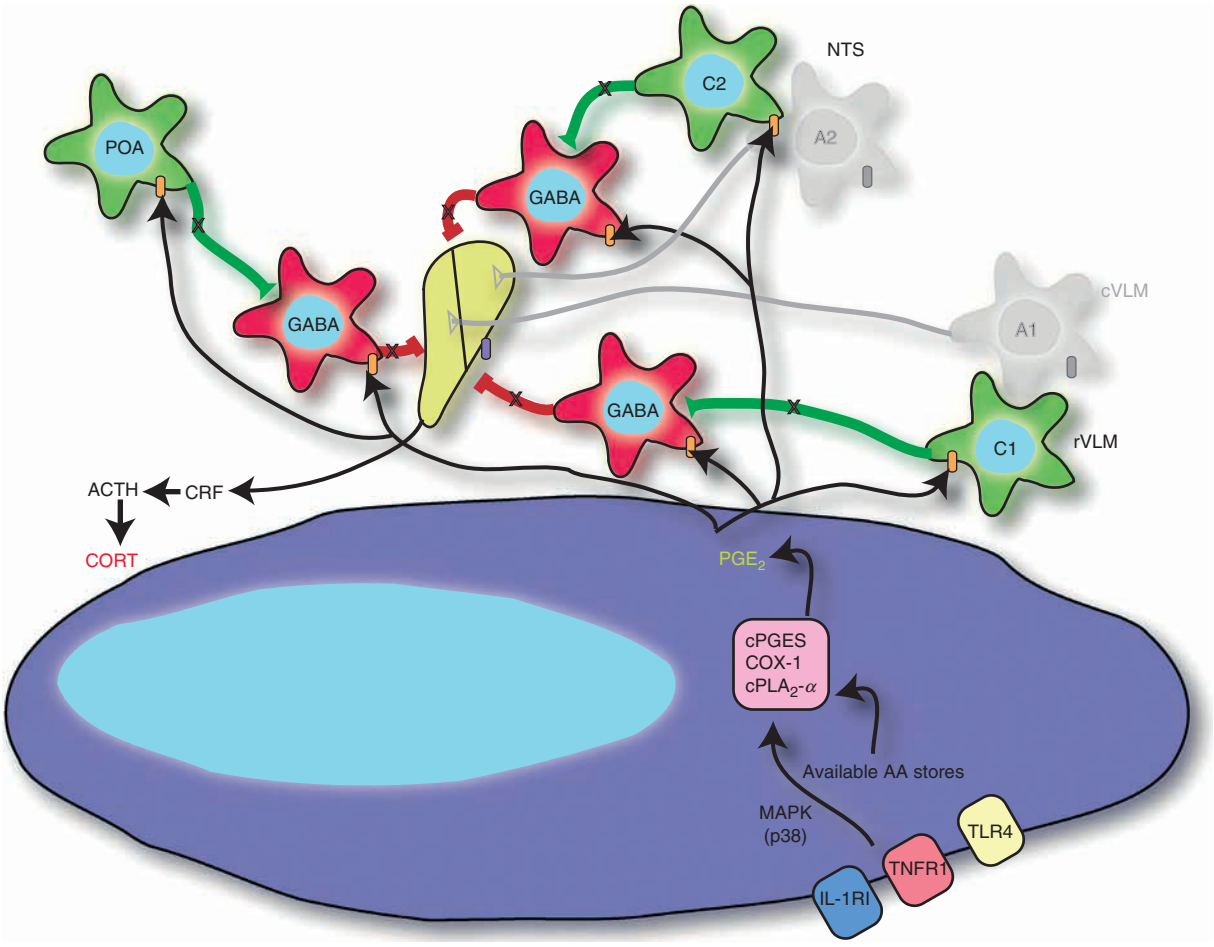

Figure 2 Hypothetical model for the fast prostaglandin E_3 receptor (EP_3)-dependent prostaglandin E_2 (PGE_2) response in the brain. Circulating inflammatory factors (interleukin (IL)-1, tumor necrosis factor (TNF), lipopolysaccharide (LPS)) bind to their receptors expressed on the surface of endothelial cells of the brain microvasculature. This leads to a rapid use of available constitutive enzymes (cyclooxygenase (COX)-1, cytosolic phospholipase A_2 ($cPLA_2$), cytosolic prostaglandin E synthase (cPGES)) and arachadonic acid (AA) stores to transiently produce and release PGE_2 (possibly through mitogen-activated protein kinase (MAPK) p38 mechanisms). This PGE_2 acts on its EP_3 receptor on γ-aminobutyricacidergic afferences to the hypothalamic paraventricular nucleus, which rapidly turns off the inhibition of the corticotropin releasing factor (CRF)-releasing neurons. A similar scenario may happen at the medial preoptic area (POA), C1 catecholaminergic neurons of the rostral ventrolateral medulla (VLM), and C2 catecholaminergic neurons of the solitary tracts nucleus, where the binding of PGE_2 to the EP_3 receptor inhibits excitation of these GABAergic projections of PVN neurons. ACTH, adrenocorticotropic hormone; CORT, corticosterone; GABA, γ-aminobutyric acid; R, receptor; TLR4, toll-like receptor 4.

The population of cells containing the enzymes responsible for PGE_2 synthesis during systemic inflammation is still debated. Numerous papers have provided clear evidence that both COX-2 and mPGES-1 transcripts and proteins are expressed essentially within the endothelium of the brain blood vessels. On the other hand, few studies have shown that perivascular myeloid cells are responsible to deliver PGE_2 in response to systemic inflammatory stimuli. The action of this PG may therefore be dependent on its cellular source, the endogenous pool of PGE_2 in endothelial cells, and the receptor subtypes expressed on their cellular targets. It has been shown that EP_4 is present on the postsynaptic sites of hypothalamic stimulatory neurons, whereas GABA neurons contain presynaptic EP_3 receptors in their terminals. Activation of both EP_3 and EP_4 receptors is actually involved in the excitatory regulation of supraoptic neurons by PGE_2. These data may suggest that both receptors contribute to the activation of PVN neurons during systemic inflammation, but at different times. Stimulation of the $cPLA_2$–COX-1–cPGES pathway would lead to rapid secretion of PGE_2 from endothelial cells and then cause an immediate activation of the CRF neurons through inhibition of GABA afferents that contain EP_3 receptors (Figure 2). The hypothalamic PVN receives robust innervation from GABA neurons that play a critical role in the control of the HPA axis during stress. Inhibition of GABAergic neurons by EP_3 may take

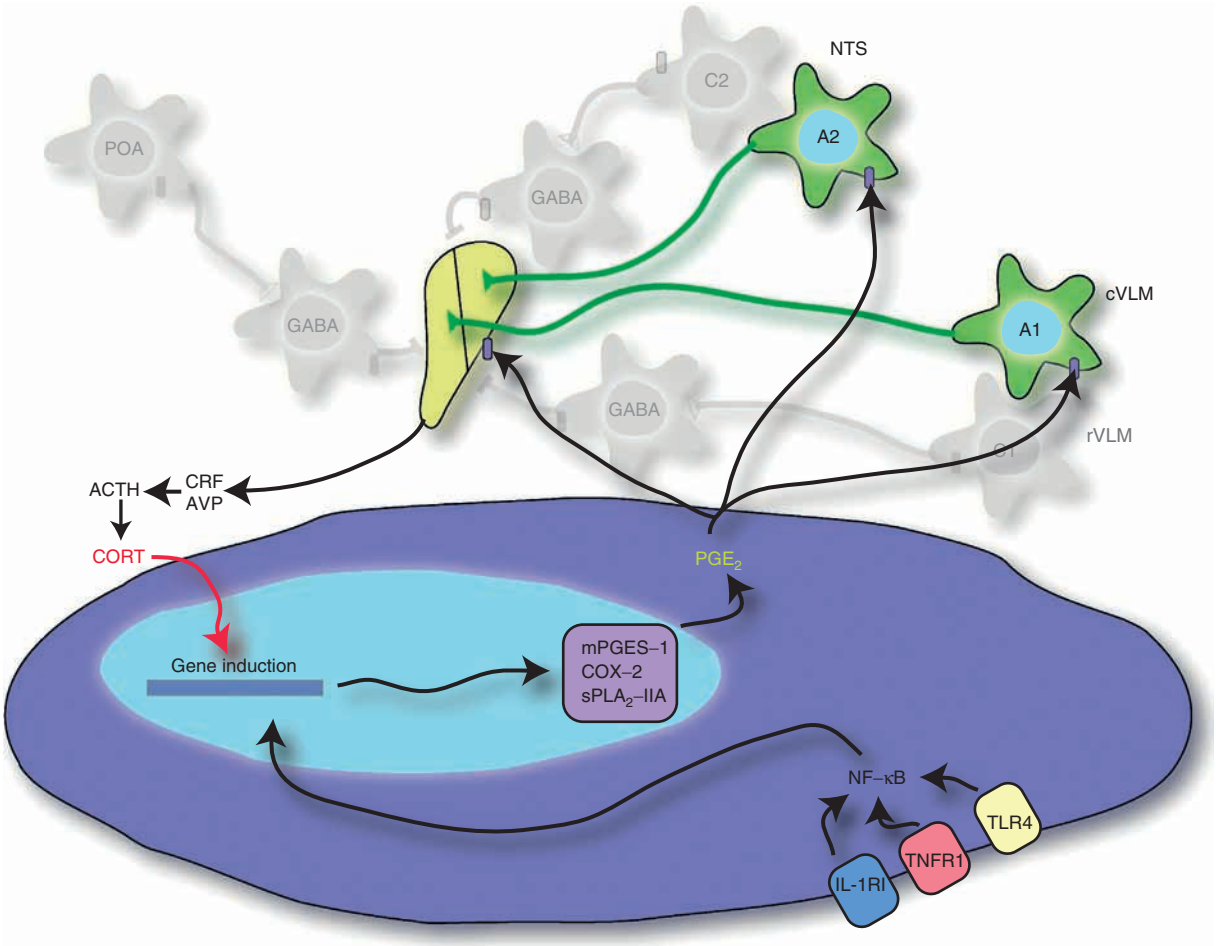

Figure 3 Delayed prostaglandin E_4 (EP_4)-dependent prostaglandin E_2 (PGE_2) response in the brain during systemic immune insults. The delayed PGE_2 response requires transcription of the inducible enzymes (secretory phospholipase A_2 ($sPLA_2$-IIA), cyclooxygenase (COX)-2, microsomal prostaglandin E synthase (mPGES)-1) through nuclear factor-kappa B (NF-κB) and area postrema-1-mediated transcriptional mechanisms. These enzymes permit the massive production of PGE_2 that acts on the EP_4 receptors in the paraventricular nucleus and at the adrenergic A1 region of the caudal ventrolateral medulla (cVLM) and A2 region of the nucleus of the solitary tract (NTS) to enhance corticotropin-releasing factor (CRF) and vasopressin (AVP) release into the infundibular system. This leads to the release of adrenocorticotrophic hormone (ACTH) and subsequently biosynthesis and secretion of corticosterone (CORT) from the adrenal medulla. The latter hormone acts as negative feedback on the transcription of the enzymes responsible for the production of PGE_2 and prevents overproduction of inflammatory molecules during innate immunity. IL-1, interleukin-1; R, receptor; TLR4, toll-like receptor 4; TNF, tumor necrosis factor.

place locally at the level of the nerve terminals within the PVN or through distal inhibition of GABA projections originating from different areas, such as the preoptic area, parabrachial nucleus, rVLM, and/or NTS.

PGE_2 synthesis under the control of the $sPLA_2$-IIA–COX-2–mPGES-1 pathway would take the relay and trigger different populations of neurons (Figure 3). Enzymes of this pathway are not constitutively present and have to be induced by the proinflammatory signal transduction pathways. It is also possible that both endothelial and microglial cells contribute to the PGE_2 production in the delayed phase. Once secreted, PGE_2 would maintain CRF activity via EP_4 stimulation either directly or through activation of the A1 and A2 catecholaminergic circuits from the caudal ventrolateral medulla and NTS, respectively. These mechanisms fit nicely with the hypothesis of a fast and delayed PGE_2 response, in which the fast phase could be associated with the disinhibition of CRF-releasing neurons of the PVN, without the need for time- and energy-hungry transcription. This response could also prime the PVN to become more receptive to the massive release of PGE_2 by endothelial and perivascular microglial cells with their newly transcribed PGE_2-synthesizing enzymes. In support of this hypothesis is the specific increase in EP_4 mRNA levels in CRF neurons and catecholaminergic afferents to PVN following a systemic immune challenge. The sustained glucocorticoid release provides direct inhibitory feedback on these events at the level of transcriptional machinery, which is the most powerful endogenous immunosuppressive mechanism. This innovative dual regulation of PGE_2 production and action in the brain could have important implications for an appropriate control of fever, HPA axis, and other autonomic functions.

Concluding Remarks

More work is clearly needed to unravel the fine-tuning of the circuits and signal transduction pathways that participate in the neural–immune interface. There is, however, no doubt that PGE_2 plays a determinant role in activating or inhibiting the key populations of neurons that are together involved in engaging the physiological responses necessary for the organism to restore health during illness. An appropriate febrile response and a time-dependent release of glucocorticoids are major phenomena in which the brain has a direct impact on the systemic innate immune system. Imbalances in these two regulatory systems are actually becoming hallmarks of autoimmune diseases and neurodegenerative disorders. Clarifying the exact mechanisms and circuits that allow such fine neurophysiological outcomes will obviously be essential for designing appropriate therapeutic strategies when dysfunction occurs during the acute-phase response. The best guess at this time points to the key kinases controlling the proinflammatory signal transduction pathways in cells of the BBB, PGE_2 biosynthetic machinery and release, and specific EP receptor subtypes expressed on the surface of the neuronal populations that control endocrine and autonomic functions.

See also: Circumventricular Organs in Neuroendocrine Control; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Inflammation in Neurodegenerative Disease and Injury; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neuroimmune System: Aging; Stress Response: Neural and Feedback Regulation of the HPA Axis.

Further Reading

Ek M, Engblom D, Saha S, Blomqvist A, Jakobsson PJ, and Ericsson-Dahlstrand A (2001) Inflammatory response: Pathway across the blood–brain barrier. *Nature* 410: 430–431.

Elmquist JK, Scammell TE, and Saper CB (1997) Mechanisms of CNS response to systemic immune challenge: The febrile response. *Trends in Neurosciences* 20: 565–570.

Engblom D, Saha S, Engström L, et al. (2003) Microsomal prostaglandin E synthase-1 is the central switch during immune-induced pyresis. *Nature Neuroscience* 6: 1137–1138.

Laflamme N, Lacroix S, and Rivest S (1999) An essential role of interleukin-1β in mediating NF-κB activity and COX-2 transcription in cells of the blood–brain barrier in response to systemic and localized inflammation, but not during endotoxemia. *Journal of Neuroscience* 19: 10923–10930.

Narumiya S, Sugimoto Y, and Ushikubi F (1999) Prostanoid receptors: Structures, properties, and functions. *Physiological Reviews* 79: 1193–1226.

Nguyen MD, Julien JP, and Rivest S (2002) Innate Immunity: The missing link in neuroprotection and neurodegeneration? *Nature Reviews Neuroscience* 3: 216–227.

Rivest S (1999) What is the cellular source of prostaglandins in the brain in response to systemic inflammation? Facts and controversies. *Molecular Psychiatry* 4: 501–507.

Schiltz JC and Sawchenko PE (2003) Signaling the brain in systemic inflammation: The role of perivascular cells. *Frontiers in Bioscience* 8: s1321–s1329.

Turrin NP and Rivest S (2004) Unraveling the molecular details involved in the intimate control of the immune and neuroendocrine systems: A revised and innovative hypothesis. *Experimental Biology and Medicine* 229: 996–1006.

Ushikubi F, Segi E, Sugimoto Y, et al. (1998) Impaired febrile response in mice lacking the prostaglandin E receptor subtype EP3. *Nature* 395: 281–284.

Yamagata K, Matsumura K, Inoue W, et al. (2001) Coexpression of microsomal-type prostaglandin E synthase with cyclooxygenase-2 in brain endothelial cells of rats during endotoxin-induced fever. *Journal of Neuroscience* 21: 2669–2677.

Zhang J and Rivest S (2003) Is survival possible without arachidonate metabolites in the brain during systemic infection? *News in Physiological Sciences* 18: 137–142.

Induced Seizures as Therapy in Man

M Fink, State University of New York, Stony Brook, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Of all the peculiar therapies suggested for the relief of madness, few match the bizarre nature of inducing seizures. For centuries seizures have been viewed as signs of individual wickedness and of God's displeasure, and efforts have been aimed at their extinction. Yet, ever since reports appeared showing that chemically induced seizures relieved catatonic schizophrenia, a small cadre of neurologists and psychiatrists have induced seizures in consenting patients to defeat the depressions of melancholia, the agitated and excited states of delirious mania, the mutism and negativism of catatonia, and the delusions of acute psychoses. Despite doubts of the professions and challenges from the public, despite terrifying images in films, despite legislative restrictions, and despite challenges from device manufacturers and their consultants who hawk nonseizure brain stimulation methods as replacements, seizures continue to be induced for the relief of madness. Indeed, interest in convulsive therapy is undergoing a renaissance in Europe and Asia, with the next generation of clinicians enthusiastically offering the treatment to an ever-growing public who is desperate for the relief of severe psychiatric disorders that are not alleviated by the plethora of medications, ever-more imaginative psychotherapies, and electronic devices.

Soon after Meduna's induction of seizures with analeptic agents (camphor, pentylenetetrazol), an electrical induction was developed and has been the mainstay of the treatment. Shock therapy, electroshock, electroconvulsive therapy (ECT), electroseizure therapy (EST), and convulsive therapy are terms for inductions in humans. Electroconvulsive shock (ECS) is the term applied to seizures induced in animals.

The repeated induction of bilateral grand mal seizures is the core process. Patterned electroencephalogram (EEG) changes and the release of brain hormones into the blood and cerebrospinal fluid (CSF) define effective inductions. EEG seizure durations longer than 20 s with characteristic buildup of amplitudes, spike and wave formations, runs of high-voltage slow waves, and a sharp end to the discharge define the seizure. Elevated serum levels of peptides derived from the hypothalamus and pituitary, such as prolactin and adrenocorticotropin (ACTH), reflect effective inductions.

A successful treatment course in patients with mood disorders requires from 8 to 12 treatments, spaced 48–72 h apart. For patients with more severe psychoses, 12–25 induced seizures may be necessary. In malignant catatonia (MC) and delirious mania, three to five daily treatments are often sufficient.

The principal benefits are the rapid relief of disordered mood, improvement in vegetative functions, and reduction in suicidal drive. Seizure inductions are remarkably safe and are well tolerated at all ages, in pregnancy, and in the systemically ill. The principal risks of the seizure are fractures, headache, and acute deliria. These risks are relieved by anesthesia and muscle relaxation. In long courses of treatment and in the elderly, cognitive impairment (memory loss) is the most widely cited complaint. For many patients, especially those whose illness is characterized by dementia, cognition improves after successful treatment.

History

In the 1920s, reports that epilepsy and schizophrenia were rarely found in the same patient suggested a biological antagonism between the two disorders. Glial concentrations in postmortem brain samples of schizophrenic subjects were lower than in normal subjects, and overabundant in the brains of epileptics. These observations led Ladislas Meduna, a Hungarian neuropsychiatrist, to develop convulsive therapy. The first patient was treated on 24 January 1934, using intramuscular camphor injections to induce seizures. In 1938, the Italian psychiatrists Ugo Cerletti and Lucio Bini demonstrated the merits of inducing the seizure through alternating electrical currents through bitemporal electrodes.

Convulsive therapy was introduced at the same time as insulin coma therapy (ICT) and leucotomy (lobotomy, psychosurgery). The efficacy of ICT was well established for the more severe forms of schizophrenia. Antipsychotic drugs replaced its use. A 10–20% of comas were accompanied by grand mal convulsions, and its mode of action was based on these events and not a direct effect of cerebral hypoglycemia. Leucotomy was a gross procedure whose efficacy was not demonstrated; it was also replaced by antipsychotic drugs.

In ECT, brief pulse stimuli replaced alternating electrical currents for lesser confusional effects. In the mid-1950s, seizure inductions using the inhalant flurothyl (Indoklon) were equally effective, but more cumbersome. In 1961, seizures induced through unilateral electrode placement over the nondominant hemisphere were associated with less immediate cognitive effects and seeming equivalent efficacy. The

placement was widely adopted. To achieve equivalent efficacy requires excessively high energies, at least 6–8 times the calibrated seizure threshold. At these high energy levels, the cognitive advantages of unilateral electrode placement are lost.

Indications

The American Psychiatric Association's classification of mental disorders (DSM-IV) is the standard nomenclature for psychiatric disorders. The disorders for which ECT is acknowledged to be effective (Table 1) and ineffective (Table 2) are listed.

Table 1 DSM-defined clinical diagnoses in which ECT is effective

Major depressive disorder	
Single episode	[296.2x]
Recurrent	
Bipolar disorder	[296.3x]
Mania	
Depressed	[296.0x]
Mixed type	[296.5x]
Not otherwise specified	[296.70]
Atypical psychosis	[298.90]
Schizophrenia	
Catatonic	[295.2x]
Schizophreniform	[295.40]
Schizoaffective	[295.70]
Catatonia	
Schizophrenia, catatonic type	
Catatonic disorder due to a medical condition	[293.89]
Malignant catatonia	[293.89]
Neuroleptic malignant syndrome	
Delirium	
Due to a general medical condition	[293.0]
Due to substance intoxication	[291.0]

Table 2 DSM diagnoses in which ECT is ineffective

Dementia and amnestic disorders	[293.0, 290.xx, 294.xx]
Substance-related disorders	[303.xx, 291,x, 304.x, 292.x]
Anxiety and somatiform disorders	[300.xx]
Factitious disorders	[300.xx]
Dissociative disorders	[300.1x, 300.6]
Sexual dysfunctions	[302.xx, 625.8, 608.89, 607.84, 608.89, 625.8]
Sleep disorders	[307.xx, 780.xx]
Impulse disorders	[312.3x]
Adjustment disorders	[309.xx]
Personality disorders	[301.xx]

Depressive Mood Disorders

Disorders dominated by sadness, hopelessness, fears, and thoughts that life is no longer worth living respond rapidly and well to ECT. At least four varieties are recognized: melancholia (dominated by motor and vegetative abnormalities), psychotic depression (psychosis), pseudodementia (cognitive deficit), and catatonia.

Melancholia A well-established syndrome of depressive mood dominated by motor signs (retardation or agitation) and vegetative symptoms of inability to sleep, feeding, and weight loss. Suicide thoughts are common. Hypercortisolemia, measured by the dexamethasone suppression test (DST), is characteristic as the patients exhibit elevated serum levels of cortisol, obtunded diurnal rhythmicity, and failure of serum levels to fall with an administered dose of dexamethasone. The DST normalizes with adequate treatment and becomes abnormal again with relapse.

In the 1990s, Parker and his associates focused attention on motor signs as melancholia's primary expression. They described a useful rating scale of motor behaviors to define the syndrome. After an extensive review of the literature, Taylor and Fink concluded that classifying mood disorder patients as either melancholic or nonmelancholic offered more homogeneous populations with better outcomes with tricyclic antidepressants (TCAs) and ECT than did the DSM classification of major depression and bipolar disorder. In their formulation, melancholia is a syndrome of depressive mood, with motor and vegetative abnormalities, and with evidence of cortisol abnormality.

Delusional (psychotic) depression Patients who are overwhelmed by feelings of helplessness, hopelessness, and worthlessness, believing others are watching or talking about them, report voices when no one is present, and imagine events depicted on a television or movie screen apply directly to them are suffering from psychotic depression. The syndrome was demarcated when depressed patients failed to respond to adequate dosing of antidepressants but did respond to ECT.

Many reviews verify that psychotic depression has a more severe pathophysiology and using the same treatments as for nonpsychotic depression, even at much higher doses, is not adequate. Yet, bitemporal ECT is remarkably effective. In a four-hospital ECT study, of 253 patients with unipolar major depression, 77 were psychotic depressed. Their remission rate was 95% compared to 83% for the nonpsychotic depressed, with the speed of response faster for the psychotic depressed patients.

Pseudodementia (reversible dementia) Because the depressed patient ignores daily events, little of what happens to him is registered and memory is compromised. The condition is hardly distinguishable from Alzheimer's dementia. It should be considered in any patient with the rapid onset of severe dementia, especially if the patient had prior depressive episodes. The dementia responds to effective antidepressant medications and to ECT.

Because the syndrome is not well known, patients are often sent to nursing home care. An example of a 58-year-old woman who developed a reversible dementia and was not adequately treated for 8 years is reported. Antidepressant treatment relieved the syndrome and returned the patient to a more normal family life.

Catatonia When the patient is mute, sitting rigidly in a chair or lying motionless on his bed, and unresponsive to questions and commands, appearing as if in a stupor, the state is catatonia or depressive stupor. Catatonia is seen among patients with many DSM diagnoses. It is discussed below.

Manic Mood Disorders

A mood disorder dominated by grandiosity, expansiveness, and increased energy and excitement may last for hours, days, or weeks. The disorder alternates or combines with depression. When switches between mania and depression occur within one or a few days, the experience is labeled rapid cycling, a malignant form of the illness. Bipolar disorder is the label applied to both mania and mixed forms of the illness.

Disturbances in eating and sleeping, thinking, memory, and movement are features of mania. Patients do not sleep, eat poorly, lose weight, and concentrate thoughts poorly. Memory is impaired, often severely, and patients may appear demented and delirious. Variations of melancholia, psychosis, pseudodementia, and catatonia occur.

Delirious mania (manic delirium) is a striking form. An otherwise normal person becomes excited, restless, sleeps poorly, and is fearful that he is being watched. He hides in the house, dresses inappropriately, and wanders about the streets. Confusion, mutism, posturing, and repetitive movements, as well as physical exhaustion to the point of death occur.

ECT was the first effective treatment. Of manic patients treated with ECT, 371/562 (66%) remitted or showed marked clinical improvement. The introduction of chlorpromazine replaced ECT for its ease of use. Effective doses were high, resulting in sudden death, neuroleptic malignant syndrome (NMS), tardive dyskinesia, and tardive dystonia. Lithium was next introduced followed by anticonvulsant drugs. Despite these agents, many patients fail medications. At such times, ECT is the effective treatment. For patients in delirious mania, ECT is a primary indication.

Catatonia

Catatonia is a motor syndrome associated with thought and mood disorders. Muscular rigidity, posturing, negativism, mutism, echolalia, echopraxia, and stereotyped mannerisms are characteristic signs. Catatonia is recognized in patients with depression and mania, in patients with systemic disorders, and in toxic brain states caused by hallucinogenic drugs.

For decades, the prevailing (and erroneous) belief was that each instance of catatonia represented schizophrenia. The DSM and International Classification of Disease (ICD) classification systems assigned patients with catatonia to the diagnosis of schizophrenia, catatonic type, with antipsychotic medications as the logical treatment. In the 1970s, various reports described catatonia in patients with manic-depressive illness and toxic states. The association with schizophrenia is unfortunate, since anticonvulsant sedatives (barbiturates, benzodiazepines) and ECT are more effective and safer but these are not usually considered for schizophrenia.

Catatonia is defined by the persistence of two or more characteristic motor signs for more than 24 h in a patient with a mental disorder. While posturing and staring can be observed, most signs are elicited in examinations described in catatonia rating scales. An intravenous challenge of lorazepam or amobarbital verifies the diagnosis in more than two-thirds of patients; a positive test response augurs well for high-dose benzodiazepine therapy. When this treatment fails, ECT is effective, although the treatment schedule may require daily treatments.

Catatonia may be transitory or may persist for months or years. It appears in many guises. Prominent examples are MC (pernicious) with a high risk of death and the NMS that follows on the administration of neuroleptic drugs.

Patients who develop an acute febrile delirium with excitement or stupor, catatonia, and vegetative dysregulation suffer from MC. ECT is the life-saving treatment. The NMS is a variant of MC. It is elicited by all neuroleptics, most commonly with the high-potency agents like haloperidol, fluphenazine, and thiothixene, but also with atypical neuroleptics. The toxic serotonin syndrome (TSS), an acute febrile and autonomic response with prominent catatonia features with prominent gastrointestinal symptoms, is another MC variant. The diagnosis and treatment follows the protocol for MC.

Psychosis

Delusions are a feature of many psychiatric conditions, notably psychotic depression, postpartum depression, toxic psychosis, and delirious mania. For delusions encapsulated within the DSM diagnostic system, the treatment is defined by the diagnosis, and the role of ECT may or may not be considered. Yet, the efficacy of ECT in relieving delusions (and hallucinations and other disorders of thought) compels its consideration as an effective antipsychotic treatment regardless of the associated mood, motor, or cognitive signs. Despite the efficacy of antipsychotic agents, including clozapine, a cadre of 'medication-resistant' psychotic patients has developed. For these patients, augmentation of the antipsychotic with ECT is often effective.

Delirium

Acutely ill psychotic patients exhibit disturbances in consciousness and appear confused. Delirium is common in toxic states, either drug induced (alcohol being the most common), or secondary to drug withdrawal, or associated with systemic illnesses. Delirium is a feature of acute manic states (e.g., delirious mania). With few resources to treat acute psychoses, ECT was applied with favorable results.

Mode of Action

The mode of action of induced seizures in relieving psychiatric disorders is an unanswered question. Much is known about the process, sufficient to define a strategy to understand the mechanism.

The Role of the Seizure

The induction of a grand mal seizure is essential to effective treatment. Two measures of this effect are known. Characteristic ictal EEG patterns are the best index that have a direct effect on the brain. Persistent interseizure EEG effects are indices of the long-term effects. The release of hypothalamic and pituitary peptides in the CSF and blood also are evidence of effective treatments.

Neither the motor elements of the convulsion nor the physiologic effects (heart rate, blood pressure) nor the psychological aspects (fear, cognition) of the intervention are essential to outcome. Although modern treatment depends on an anesthetic and muscle relaxant to minimize the risks, these are not in the therapeutic chain, as unmodified inductions are clinically effective. Although electricity is the most widely used induction agent, it offers little benefit over chemical inductions other than ease of use, ready availability, and least expense.

A single seizure is rarely successful. Repeated seizures from 7 to 25 inductions are effective in acute courses, varying with psychopathology. Once to twice weekly in the elderly depressed to daily for MC is the range of inductions. Many illnesses resolve quickly, indicating that the mechanism is of immediate moment. Relapse is frequent, indicating that the mechanism needs continuous reinforcement. Medication combinations are useful in continuation, but optimal results require continuation ECT. Durations vary from a few months to years.

While the focus is on the seizure, the therapeutic effect depends on the brain effects for which the seizure is the outward sign. Manipulations of the seizure characteristics by varying electrode location and stimulus strength affect treatment efficacy. Seizures induced through unilateral electrode placement and electrical doses just over threshold (to minimize effects on memory that reside in the dominant temporal lobes) are minimally clinically effective. They affect the EEG asymmetrically and elicit smaller brain hormone discharges. Once energies are elevated to levels sufficient to affect brain hormones, the advantages in cognitive effects are lost. The difference in efficacy between unilateral stimulation (RUL ECT) and bilateral stimulation (BT ECT) encourages the conclusion that the benefits of the seizure reside in the brain's centrencephalic regions.

Recent attempts to stimulate brain regions by repeated magnetic impulses (rTMS) or retrograde stimulation of the left vagus nerve (VNS) or stimulation through brain-implanted electrodes (DBS) have not approached the therapeutic effect of ECT. These experiences are based on the expectation that stimulation of brain loci or at levels not sufficient to elicit a bilateral grand mal seizure will be as effective as inducing a seizure. Similar earlier attempts have also failed to replicate the effects of convulsive therapy: subconvulsive electrical stimulation, nonseizure doses of analeptic agents, prolonged isoelectric brain stimulation under isoflurane anesthesia; nonconvulsive electrotherapy, electroanesthesia, and direct current cerebral stimulation.

What Does the Seizure Signify?

Seizures are a characteristic of all mammalian forms. Their biological significance is unknown. In convulsive therapy, the electrical patterns of a grand mal seizure are markers of brain changes that are essential to behavioral change. The motor and systemic aspects of the seizure do not have the same significance.

Neuroendocrine features are central The release of peptides from the hypothalamus and pituitary into the CSF and measurable in the blood are essential to

clinical outcome. The metabolism of peptides is measurable in hours and days. While epinephrine, serotonin, and dopamine brain concentrations change with seizures, their levels persist for very short periods and are localized to the brain regions of stimulation. At one time, acetylcholine was considered in the therapeutic chain in convulsive therapy. Changes in cholinesterases were associated with behavioral change and behaviors were modified by cholinergic and anticholinergic substances.

The blood–brain barrier The blood–brain barrier is compromised with seizures, suggesting that substances released into the brain and CSF more easily leak to other body tissues.

Continuation treatments are necessary The changes in behavior are measurable in hours after one or two seizures, in days with more seizures, and persistent after a successful series of induced seizures. But the measures of change (EEG and serum peptides) revert to baseline levels in days or weeks after a series of seizure inductions. Such reversion is associated with behavioral relapse, seen in more than half the patients when reexamined over 6 months. Seizure inductions must be continued for periods up to 6 months in most patients and lifelong in some to sustain a clinical benefit.

The diabetes analogy While seizures are induced under anesthesia, giving the appearance of a surgical procedure, ECT is not comparable to surgery in that a single treatment is not successful. A more comparable analogy is that of diabetes; the failure of the pancreas to deliver adequate amounts of insulin is a chronic condition that is best alleviated by repeated administrations of insulin or insulin-active substances. A converse of this analogy is that the disorders in mood, thought, and motor functions result from deficiencies in brain hormones, most likely of hypothalamic pituitary origin.

The Neuroendocrine Theory of the Mode of Action of ECT

The neuroendocrine explanation views the responsive behavior disorders as evidence of deficiencies in the release of hypothalamic–pituitary peptides. Inducing seizures stimulates central brain nuclei to release peptides in greater amounts. Falls in CSF and serum calcium are indirect evidence that hormone release is massive. With repeated stimulation, brain and tissue peptide levels normalize. If these become abnormal again, relapse occurs. The compromise of the blood–brain barrier allows greater penetration of the newly formed peptides. A useful index of abnormality in depressed patients is elevated serum cortisol, reflected in measurement of the DST. Acetylcholine release and the response in cholinesterases are another evidence of the brain abnormality and normalization.

Alternative Theories

Many hypotheses of the mode of action have been postulated, beginning with memory change and the suppression of unconscious conflicts. Of the alternative hypotheses, the interest in the stimulation of brain cell growth, as in increase in neurogenesis, is now the most active.

Procedures

Recommendation for Treatment

Referral of patients for ECT is generally a last resort, except for the most acutely ill hospitalized patients. Professional inexperience, lack of facilities, and public repugnance for electroshock make referral difficult. Its use is impeded by the demand that patients explicitly sign a consent document that defines the aims of the treatment, the reasons it is now recommended, who will administer the treatments, how many treatments are accepted, and the explicit statement that the patient may discontinue treatment at any time. For patients who do not comprehend the seriousness of their illness and cannot voluntarily consent, calls on the state courts for consent, a complex and expensive procedure, must be made. Such involuntary application was less than 5% of treatments in the US in the 1990s.

Procedure

An intravenous line is established; an anesthetic (methohexital, propofol, etomidate, ketamine), and the muscle relaxant succinylcholine are administered; an airway is established; and the patient is ventilated with 100% oxygen. Electrodes are applied either bitemporally or bifrontally. Current dosages are defined as about 50% above the minimum necessary to achieve a seizure.

Seizure duration is monitored by the EEG, the motor activity in a limb isolated from succinylcholine by a blood pressure cuff inflated above the systolic pressure, or by heart rate. Seizure durations less than 25 s are considered inadequate and are usually repeated. Durations over 180 s are considered prolonged and are terminated by intravenous diazepam.

Treatments are given either 3 or 2 times a week, with full courses ranging from 6 to 25 treatments, with an average of seven to nine for mood disorders and larger series for psychoses.

Devices used to stimulate a seizure are now limited to brief pulse, square wave currents, with parameters

of 0.5–1.0 ms pulse width, 30–70 Hz frequency, for stimulus durations of 0.5–8.0 s, delivering between 25 and 500 millicoulombs of energy. Alternating current devices are associated with greater memory impairment and are no longer endorsed.

Following a successful treatment course, continuation treatment for 3–6 months is recommended. It is usually accomplished by antidepressant drugs, lithium, or continuation ECT, with the latter increasingly favored.

Consequences

The principal behavioral changes are a heightening of mood and affect, relief of abnormal mood, and suicidal ideation. Delusional ideation gradually disappears. Motor excitement, withdrawal, and apathy are relieved, and vegetative functions improve.

Bone fracture, organic mental syndrome, and cardiovascular complications leading to death are the principal risks. Fractures are rare under present conditions of anesthesia; as is death, which is reported in less than 4 of 100 000 treatments. Cognitive impairment, seen as persistent complaints of memory loss, is reported in about 1% of patients. The incidence is greater in patients treated with sinusoidal current devices and bilateral electrode placements. Recall of recent events around the period of hospitalization and of the acute illness is most disrupted. Remote memory, learned skills, and memory of events after the treatment course are not affected.

Neuropsychological tests, including memory, motor functions, reaction time, and perception, are progressively impaired during the treatment course, but recover within 4–6 weeks after the last seizure. In patients who are relieved of their depressive illness, motor performance and cognitive tests improve over pretreatment values.

During each seizure, the EEG exhibits high-voltage, high-frequency spike activity with slow-wave bursts. In interseizure records, frequencies slow, amplitudes increase, and fast frequencies are reduced. Following treatment, slow-wave frequencies are reduced and synchronized alpha activity increases.

During each seizure, there is transient hypertension, tachycardia, and in many subjects, arrhythmias.

Transient increases in cerebral catecholamines, acetylcholine, and peptides, including prolactin, vasopressin, adenocorticotropic hormone, and thyrotropin releasing hormone (TRH) occur during each treatment. These values return to baseline levels within 2 h in the first treatments but persist for longer periods with increased number of inductions. Increases in cholinesterase and monoamine oxidase in the CSF and a reduction of calcium in plasma, CSF, and blood may persist for longer periods. Catecholamine turnover and receptor sensitivity increase, particularly dopamine synaptic activity. In patients with evidence of neuroendocrine dysregulation (as abnormal cortisol and failure to suppress with dexamethasone, abnormal growth hormone response to apomorphine, or thyrotropin-stimulating hormone response to TRH), these responses normalize with treatment.

Social Factors

ECT is remarkably effective and, with modern methods of anesthesia, remarkably safe. For many severe mental illnesses, it is more effective and safer than alternate therapies. While 8–10% of patients admitted to academic in-patient psychiatric services in the US receive ECT, the treatments are infrequently used in municipal, state, and federal facilities. This disparity in access reflects the treatment's 'bad press,' as it has been the target of the antipsychiatry movements throughout the world.

See also: Bipolar Disorder; Depression and the Brain; Electroencephalography (EEG); Epilepsy; Epilepsy: Channelopathies; Functional Amnesia; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome.

Further Reading

Abrams R (2001) *Electroconvulsive Therapy,* 4th edn. New York: Oxford University Press.

Abrams R and Swartz C (1990) *The Technique of ECT.* Educational videotape. Somatics Inc., Lake Bluff, IL.

American Psychiatric, Association (2002) *Electroconvulsive Therapy: Recommendations for Treatment, Training and Privileging.* Washington, DC: American Psychiatric Association.

Beyer JL, Weiner RD, and Glenn MD (1998) *Electroconvulsive Therapy. A Programmed Text,* 2nd edn. Washington, DC: American Psychiatric Association Press.

Dukakis K and Tye L (2006) *Shock: The Healing Power of Electroconvulsive Therapy.* New York: Penguim Group.

Fink M (1984) Meduna and the origins of convulsive therapy. *American Journal of Psychiatry* 141: 1034–1041.

Fink M (1986) *Informed ECT for Patients and Families.* Informational videotape. Somatics Inc., Lake Bluff, IL.

Fink M (1986) *Informed ECT for Health Professionals.* Informational videotape. Somatics Inc., Lake Bluff, IL.

Fink M (1991) Impact of the anti-psychiatry movement on the revival of ECT in the US. *Psychiatric Clinics of North America* 14: 793–801.

Fink M (1997) Prejudice against ECT: Competition with psychological philosophies as a contribution to its stigma. *Convulsive Therapy* 13: 253–265.

Fink M (1999) *Electroshock: Restoring the Mind.* New York: Oxford University Press.

Fink M and Taylor MA (2003) *Catatonia: A Clinician's Guide to Diagnosis and Treatment.* Cambridge: Cambridge University Press.

Kellner CH, Pritchett JT, Beale MD, and Coffey CE (1997) *Handbook of ECT.* Washington, DC: American Psychiatric Association Press.

Kneeland TW and Warren CAB (2002) *Pushbutton Psychiatry. A History of Electroshock in America.* Westport CT: Praeger.

MECTA Corporation (2003) Electroconvulsive Therapy. Educational videotape. Dartmouth Hitchcock Medical Center.

Ottosson J-O and Fink M (2004) *Ethics in Electroconvulsive Therapy.* New York: Brunner-Routledge.

Scott AIF (ed.) (2004) *The ECT Handbook,* 2nd edn. London: Royal College of Psychiatrists.

Shorter E and Healy D (2007) *Shock Therapy: The History of Electroconvulsive Treatment in Mental Illness*. New Brunswick, NJ: Rutgers University Press.

Taylor MA and Fink M (2006) *Melancholia: The Diagnosis, Pathophysiology and Treatment of Depressive Illness.* Cambridge, UK: Cambridge University Press.

Infectious Agents in Neurodegenerative Disease

M J Bellizzi and H A Gelbard, University of Rochester Medical Center, Rochester, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuronal apoptosis has been recognized for more than 10 years as one of the pathologic hallmarks of neurodegenerative disease associated with HIV-1 infection of the central nervous system (CNS), but mounting evidence suggests it has little to do with symptoms of neurologic disease associated with neurodegeneration. Studies of HIV-1-associated dementia (HAD) suggest that neurologic and cognitive function becomes impaired primarily as a result of injury to dendrites and synapses, often occurring long before cell death. HAD appears to develop not from loss of neurons but because of loss of communication between neurons and perhaps glia as well. This distinction has important implications for disease progression and treatment; dead neurons are not likely replaceable, except perhaps in areas of the CNS that may benefit from neurogenesis, but dendrites and synapses can recover from severe structural and functional injury as long as the neuronal cell body survives. With synaptic injury and clinical symptoms often preceding the extensive neuronal loss of late-stage disease by months or years, therapeutic strategies that provide synaptic protection have potential to improve neurologic function and perhaps reverse the course of the disease.

Dramatic recovery of neurologic function in patients with HAD beginning antiretroviral therapy demonstrates that a temporal window for reversibility of neurologic deficits does exist. HAD develops following HIV-1 infection of brain macrophages and microglia, and cognitive and motor symptoms include impaired concentration, apathy, memory loss, psychomotor slowing, weakness, tremor, and ataxia. These symptoms can be severely debilitating, but antiretroviral therapy that decreases HIV-1 burden in the brain and the periphery can ameliorate and even reverse many of these symptoms, at least temporarily. For example, an early report described a patient diagnosed with HIV-1 infection and HAD after presenting with 6 months of progressive memory and attentional deficits that had become so severe that she could no longer work, dress herself, or be left unsupervised. After 12 weeks of antiretroviral therapy that decreased viral burden and associated inflammatory markers in the plasma and cerebrospinal fluid, her performance on cognitive and motor testing had dramatically improved and most, although not all, her presenting deficits were reversed. Similar clinical recovery, accompanied by significant improvement of cerebral atrophy documented by neuroimaging, was reported after antiretroviral therapy in a HIV-1-infected child with language, attention, and motor deficits that had caused him to lose developmental milestones and prevented him from communicating with others at school and at home. Studies following cohorts of patients beginning highly active antiretroviral therapy (HAART) provide further evidence of a large reversible component in HAD.

HAART modifies the disease course but does not cure or prevent HAD. The prevalence of HAD and other neurologic disease in patients receiving HAART is the same as in the pre-HAART era. HAART has changed the phenotype of the disease, causing it to develop more slowly and shifting the pattern of clinical deficits, perhaps reflecting involvement of different brain areas: functional neuroimaging studies in the HAART era suggest substantial cortical involvement in a disease once considered to be primarily subcortical. The persistent 30–40% prevalence of HAD in HIV-positive cohorts despite HAART indicates that therapies targeting mechanisms that are independent of the viral life cycle with the goal of preventing synaptic injury are likely necessary for long-term treatment of HIV-1-associated neurologic disease.

How HIV-1 damages synaptic architecture and function is not known, and understanding mechanisms of synaptic injury in HIV-1 neurologic disease is necessary for identifying therapeutic approaches to protect synaptic function. Similarities in disease pathology and neurotoxic mechanisms suggest that this understanding might provide insight into treatment for other neurodegenerative disorders with an infectious etiology.

HIV-1-Associated Dementia and Prionoses Are Diseases of Synaptic Architecture

Evidence that neurologic impairment in HAD is caused by synaptic injury comes largely from correlations of neuropathologic studies with neurologic and neuropsychological testing. HIV-1 infection is associated with neuronal apoptosis, dramatic decreases in numbers of presynaptic terminals and postsynaptic dendritic spines, dendritic injury, and signs of HIV encephalitis, which include HIV-1-infected monocyte–macrophages, multinucleated giant cells, activated

microglia, astrocytosis, and myelin pallor. Only synaptic loss and microglial activation correlate with the presence of neurologic deficits. While the multinucleated giant cells and myelin pallor of HIV encephalitis were seen only in patients with dementia, 50% of patients developed HAD in the absence of either of these findings. Similarly, premortem neurocognitive function did not correlate with the presence of microglial nodules, astrocytosis, or HIV-infected mononuclear cells in autopsy tissue.

In contrast, loss of pre- and postsynaptic structures correlates well with neurologic impairment across a range of severity of HIV-1 neurologic disease. Studies of patients with HAD and milder forms of HIV-1-associated neurologic impairment, as well as unimpaired HIV-1-seropositive controls, found that global ratings of neuropsychologic function derived from an extensive battery of tests significantly correlated with synaptic density estimated by presynaptic synaptophysin immunoreactivity as well as with postsynaptic microtubule-associated protein 2-positive dendritic area. Loss of synapses was not simply a result of neuronal death; neuronal loss consistently failed to correlate with neurologic impairment in a number of studies. While neuronal death clearly has an impact on neurologic function, these studies suggested that neuronal loss occurs late in the disease and that synaptic injury prior to neuronal death is the major determinant of HIV-associated neurologic disease. This is supported by morphologic studies that show extensive injury to the dendritic arbor, including focal swelling, or beading; loss of dendritic branches; and 40–60% reductions in densities of dendritic spines along remaining branches, all occurring on surviving neurons in HAD. The additional correlation between neurologic impairment and microglial activation, a general marker of neuroinflammation, suggests that inflammatory processes might contribute to synaptic injury.

Similar evidence suggests that synaptic injury underlies neurologic deficits in other neurodegenerative diseases including the prionoses. Loss of synapses preceding neurodegeneration has been demonstrated in scrapie-infected murine hippocampus, with the major neuropathologic finding of abnormal pyramidal synapses in the scrapie-infected murine hippocampus. Loss of synapses and neurologic impairment correlates in the ME7 murine model of prion disease, and studies demonstrate similar synaptic pathology in patients with Creutzfeldt–Jakob disease (CJD) at both a light microscopic and ultrastructural level. Indeed, one of the most striking ultrastructural findings in a study of the CJD prionoses was dark profiles or 'tombstones' of synaptic terminals.

In addition to loss of synaptic structures, functional impairment of existing synapses might further disrupt neural processing in HAD. In addition, electrophysiologic studies in brain slices prepared from several different animal models of HAD have revealed deficits in synaptic plasticity, often occurring early in the disease and before overt neuropathological changes. It is interesting that in the ME7 model of prionoses, abnormalities in synaptic plasticity, while temporally contiguous with deposition of protease-resistant protein, scrapie-type (PrP^{Sc}), occur before overt changes in behavior or frank neuropathologic hallmarks of the prionoses such as vacuolation are evident.

While much attention has been focused on neuroimaging techniques, including diffusion weighted images and fluid-attenuated inversion recovery images for the diagnosis of the prionoses, no studies have characterized progression of neurologic deficits and related it to changes in functional neuroimaging parameters, probably because of the difficulty in diagnosing this class of diseases in the asymptomatic phase and the rapid, inexorable progression of the disease once neurologic symptoms are apparent. In contrast, functional studies in patients with HIV-1 are consistent with impairment of synaptic processing even when there is no overt sign of neuropsychologic or neurologic deficit. HIV-1-positive patients with varying degrees of neurologic impairment showed increased areas of brain activation on functional magnetic resonance imaging during attention and working memory tasks. During simpler tasks, brain activation increased but performance did not deteriorate, whereas during more-complex tasks, larger increases in brain activation correlated with decreased performance. This finding supports the hypothesis that patients with impaired processing recruited additional neural substrate, or cognitive reserve to complete the tasks. It is interesting that cognitively asymptomatic HIV-1-positive patients recruited larger brain areas, in the same regions as patients with HAD, compared to HIV-1-seronegative controls, suggesting that the former might already have had impairment of neural processing which was compensated by increased recruitment of cognitive reserve. Electroencephalographic studies during an auditory processing task support this conclusion. Patients with HIV-associated neurologic disease had longer latencies of event-related potentials that reflect impaired frontal cortical processing. Latencies increased with more-severe neurologic deficits and were also significantly longer in cognitively intact HIV-seropositive patients compared than in seronegative controls, again suggesting subclinical functional impairment during the asymptomatic phase of the disease.

Taken together, these diverse studies suggest that neural processing is impaired in preclinical stages of

HAD and worsens throughout disease progression in tandem with the development of progressively more severe synaptic pathology. Neuronal cell death likely becomes significant in late-stage disease, well after the onset of neuropsychologic impairment. These findings strongly support the hypothesis that synaptic injury is the primary determinant of neurologic impairment in HAD and very possibly the substrate for early neurologic disease in the prionoses. Thus, understanding mechanisms of synaptic injury and dysfunction in these diseases might be crucial to preserving neurologic function.

Inflammation and Excitotoxic Synaptic Injury

How synapses are injured in prionoses is not well understood. Evidence for excitotoxic damage to synapses and neuronal cell bodies in the prionoses is equivocal, arguably because of the difficulty in understanding how PrP^{Sc} interacts with different cellular moieties germane to glutamate-mediated neurotransmission, and the finding that mice with a targeted deletion of the normal cellular prion protein do not display abnormalities in the kinetics of γ-aminobutyric acid and glutamate receptor-mediated currents. It is interesting that plasminogen, a pro-protease that may participate in glutamate-mediated excitotoxicity, selectively binds PrP^{Sc}. However, neuropathologic studies of expression of excitatory amino acid transporter-1 (EAAT-1) in brain macrophages and microglia in patients with prion disease were unable to establish a clear relationship between EAAT-1's ability to clear extracellular glutamate and the extent of neuronal apoptosis or spongiform neuropathologic changes.

In contrast, *in vitro* studies have revealed many potential neurotoxic mechanisms for the pathogenesis of HAD, but investigation of how these contribute to synaptic dysfunction and destruction *in vivo* has been hampered by the slow progression of synaptic injury in these diseases. Nevertheless, multiple lines of evidence suggest that synaptic damage in HAD is likely due to chronic neuroinflammation associated with mitochondrial dysfunction, disrupted intracellular calcium homeostasis, and oxidative stress in vulnerable neurons that increase synaptic vulnerability to glutamate excitotoxicity.

HIV-1 rarely, if ever, infects postmitotic neurons. It exerts its neurotoxic effects largely by triggering brain inflammation. By infecting brain macrophages, microglia, and to a lesser extent astrocytes and secondarily activating uninfected microglia and astrocytes, HIV-1 causes release of cytokine and other inflammatory mediators, including tumor necrosis factor-α (TNF-α), interleukin-1β (IL-1β), interleukin-6, arachidonic acid metabolites, platelet-activating factor (PAF), and complement proteins, in addition to secreted HIV-1 proteins. This inflammatory milieu is toxic to neurons *in vitro*, as demonstrated by experiments with cerebrospinal fluid (CSF) from patients with HAD and conditioned media from HIV-infected macrophages. In patients with HIV, brain inflammation appears to relate closely to cognitive dysfunction: severity of dementia correlated well with numbers of activated microglia and macrophages and with TNF-α mRNA expression in the brain, but not with numbers of HIV-1-infected cells, and decreases in CSF TNF-α protein levels coincided with neurologic recovery in a patient beginning HAART.

Evidence from *in vitro* studies suggests that HAD inflammatory mediators damage neurons by making them vulnerable to Ca^{2+}-dependent glutamate excitotoxicity. Supernatants containing the soluble mediators released by HIV-1- or amyloid-β-activated macrophages and microglia can kill cultured neurons in a manner that depends on activation of *N*-methyl-D-aspartate (NMDA)-type glutamate receptors. Many factors may contribute: proinflammatory molecules PAF, TNF-α, and IL-1β, as well as amyloid-β and the secreted HIV proteins gp120 and Tat, each exert neurotoxic effects largely by promoting glutamate excitotoxicity. Several studies are consistent with a role for excitotoxicity *in vivo*: NMDA receptor-dependent neuronal loss and neurotransmitter depletion have been observed, respectively, in rats injected with amyloid-β and a retroviral (LP-BM5) murine model of HAD. It is intriguing that studies with the LP-BM5 murine leukemia retrovirus (MuLV) model demonstrate production, probably by molecular mimicry, of virally induced auto-antibodies to alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor subunits, which are responsible in part for the neurodegenerative disease in this model. Furthermore, NMDA receptor antagonists prevent decline in synaptic function in the severe combined immunodeficiency mouse models of HIV-1 encephalitis and HAD. Glutamate excitotoxicity is widely believed to depend on Ca^{2+} influx into neurons via ionotropic glutamate receptors and subsequent uptake by mitochondria, leading to mitochondrial depolarization and dysfunction, including free radical generation and dysregulation of intracellular Ca^{2+} homeostasis, which can culminate in neuronal injury or death. HIV mediators likely contribute to this process at multiple stages. HIV-1, gp120, TNF-α, and amyloid-β can all inhibit glutamate uptake by cultured astrocytes, leading to higher extracellular glutamate concentrations in the brain. TNF-α, which can have toxic or protective effects depending on which receptors it activates, likely augments AMPA receptor-mediated glutamate toxicity

in neurons by increasing the membrane expression of Ca^{2+}-permeable AMPA receptors. PAF increases neuronal glutamate release and promotes coupling of NMDA receptor activation with production of nitric oxide (NO), which links Ca^{2+} influx to mitochondrial dysfunction. Amyloid-β and HIV proteins can increase NO production and can lead to mitochondrial dysfunction and free radical generation in cultured neurons. Signs of mitochondrial dysfunction and oxidative stress have been found in patients with HAD compared with cognitively normal controls. In addition, treatment of cultured neurons with CSF from HIV-infected patients disrupts intracellular Ca^{2+} homeostasis, with higher Ca^{2+} concentrations at baseline and dramatically increased Ca^{2+} responses to glutamatergic stimulation.

Impaired uptake of extracellular glutamate and increased sensitivity of neurons to glutamate-triggered Ca^{2+} influx and its downstream sequelae provide a compelling basis for the proposal that glutamate excitotoxicity mediated by ionotropic glutamate receptors might be a final common pathway to neuronal injury in HAD. This has led to considerable interest in therapeutic interventions that target NMDA receptor activation. However, the sequence of events leading to excitotoxic injury in chronic neurodegenerative disease is far from clear and could have important implications for the ability of glutamate receptor antagonists to mitigate excitotoxic injury. Many investigators have speculated that NMDA receptors are tonically activated in these diseases by chronically elevated glutamate concentrations in the synaptic cleft due to impaired astrocytic uptake of glutamate and/or by partial depolarization of neuronal membranes, perhaps due to mitochondrial dysfunction and impaired energy metabolism, that eliminates the voltage-dependent Mg^{2+} blockade of NMDA receptors and leads to sustained Ca^{2+} influx. This scenario provides the rationale for the therapeutic use of memantine, an uncompetitive NMDA receptor antagonist that blocks tonic but not transient activation. Direct evidence for this scenario *in vivo* is scarce, but indirect support comes largely from efficacy of treatment with memantine in disease models and in patients. Memantine has been shown to improve synaptic function and learning, respectively, in mouse models of HAD, but its ability to ameliorate symptoms of HAD is unclear.

One way of explaining this surprising conundrum is that at early disease stages, neurons and astrocytes are not sufficiently dysfunctional to cause the tonic NMDA receptor activation that is most amenable to memantine blockade. Alternatively, effects of HAD inflammatory mediators might promote synaptic injury following more physiologic glutamate release. Recent work has demonstrated that high-frequency synaptic stimulation, sufficient to generate long-term potentiation (LTP) in hippocampal slices, can instead initiate calcium- and caspase-dependent dendritic injury, with failure of LTP in slices exposed to cPAF, a PAF analog resistant to catabolism by tissue acetylhydrolases. Patch clamp recordings from postsynaptic neurons show no change in duration or extent of glutamate receptor activation during stimulation in cPAF-treated or control slices, suggesting that similar patterns of glutamatergic activity might induce physiologic plasticity or synaptic injury depending on whether neuroinflammation is present. Because we found that dendritic injury could be prevented and LTP restored by preconditioning with a mitochondrial adenosine triphosphate-sensitive K^+ channel antagonist, and that blockade of Ca^{2+} influx or postsynaptic caspase activation prevented structural injury but failed to restore LTP to cPAF-treated slices, we have concluded that abnormal NMDA receptor activation might not be sufficient to protect against synaptic injury during neuroinflammation. Rather, we believe that chemical preconditioning might provide a promising strategy for preventing activity-dependent synaptic injury while preserving physiologic plasticity.

To date, most preclinical studies of neurotoxic mechanisms in HAD have used neuronal death as their endpoint. Thus, it would not be surprising if these studies offered more insight into end-stage disease, where cell death is far more prominent, than into earlier stages, where injury is restricted to synapses. Thus, development of model systems to investigate perturbation of synaptic structure and function in the absence of neuronal death might be important for identifying therapeutic strategies to prevent synaptic degeneration before it occurs irreversibly.

Synaptic Degeneration and Recovery

Neuropathologic findings make clear that synapses degenerate before neurons die, but this progression is not inevitable. Abundant experimental evidence demonstrates that neurotoxicity can originate in and can be confined to pre- and postsynaptic elements and that synaptic structure and function can largely recover after removal of the neurotoxic insult. Molecular events associated with cell death, including caspase activation, mitochondrial calcium uptake and depolarization, free radical generation, loss of membrane asymmetry, and membrane blebbing, can occur locally in synapses exposed to glutamate, amyloid-β, prion protein, staurosporine, or Fe^{2+}-induced oxidative stress. Similar events may occur in presynaptic terminals and axons following loss of postsynaptic targets.

Apoptotic responses can be generated in the synapses and dendrites, independent of the cell body,

as demonstrated by application of toxins to isolated synaptic compartments or restricted portions of neurites. Though experiments with cytosolic extracts from synaptosomes show that synapse-derived signals can trigger apoptotic changes in cell nuclei, studies in intact neurons suggest that they remain localized to sites of excitotoxic exposure. Synapses and dendrites appear to be especially vulnerable to injury as cellwide exposures to sublethal doses of glutamate, amyloid-β, and prion protein can cause dendritic beading, caspase activation, and loss of membrane asymmetry in the dendrites while leaving the neuronal cell bodies relatively unaffected.

Injured neurites have a remarkable capacity for recovery. In cultured neurons and hippocampal slices, sublethal doses of excitotoxins glutamate, kainate, or NMDA can dramatically disrupt dendritic architecture with large focal swellings and retraction of spines, accompanied by loss of synaptic transmission and changes in mitochondrial structure and function that can all recover following removal of the excitotoxin. Dendritic injury in these experiments reflects in part swelling due to ion influxes and osmotic changes, which can recover within minutes, and in part Ca^{2+}-mediated toxicity that can last from hours to days; nevertheless, dendrites can recover from both when the cell body is spared.

Synaptic repair correlating with neurologic recovery has been observed in chronic neuroinflammatory disease models. In mice with experimental autoimmune encephalitis, a model for multiple sclerosis, beading of motor neuron dendrites and loss of immunostaining for pre- and postsynaptic markers accompanied episodes of motor dysfunction, and all returned to normal with remission of neurologic deficits.

The remarkable regenerative potential of dendrites and synapses suggests a likely basis for the reversibility of HAD neuropsychologic symptoms in patients beginning HAART. Although such reversal has not been demonstrated in patients with other neurodegenerative diseases, or even in patients with HAD who have received long-term treatment with HAART, the potential for functional neurologic recovery in this patient group warrants further investigation with neuroprotective strategies that either reduce neuroinflammatory burden on synapses or precondition them to withstand this type of stress. This is even more compelling when one realizes that HAART fails to prevent HAD in the long run. Identifying potential targets for such therapies will require a better understanding, not of pathways to neuronal death, but of how HIV and inflammatory neurotoxins lead to synaptic injury and loss of function.

See also: Apoptosis in Neurodegenerative Disease; Excitotoxicity in Neurodegenerative Disease; Inflammation in Neurodegenerative Disease and Injury; Intracellular Calcium and Neuronal Death; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Prion Diseases.

Further Reading

Adle-Biassette H, Chretien F, Wingertsmann L, et al. (1999) Neuronal apoptosis does not correlate with dementia in HIV infection but is related to microglial activation and axonal damage. *Neuropathology and Applied Neurobiology* 25: 123–133.

Anderson ER, Boyle J, Zink WE, Persidsky Y, Gendelman HE, and Xiong H (2003) Hippocampal synaptic dysfunction in a murine model of human immunodeficiency virus type 1 encephalitis. *Neuroscience* 118: 359–369.

Bellichenko PV, Brown D, Jeffrey M, and Fraser JR (2000) Dendritic and synaptic alterations of hippocampal pyramidal neurons in scrapie-infected mice. *Neuropathology and Applied Neurobiology* 26: 143–149.

Bellizzi MJ, Lu SM, Masliah E, and Gelbard HA (2005) Synaptic activity becomes excitotoxic in neurons exposed to elevated levels of platelet-activating factor. *Journal of Clinical Investigation* 115: 3185–3192.

Chang L, Speck O, Miller EN, et al. (2001) Neural correlates of attention and working memory deficits in HIV patients. *Neurology* 57: 1001–1007.

Chiti Z, Knutsen OM, Betmouni S, and Greene JR (2006) An integrated, temporal study of the behavioural, electrophysiological and neuropathological consequences of murine prion disease. *Neurobiology of Disease* 22(2): 363–373.

Cunningham C, Deacon R, Wells H, et al. (2003) Synaptic changes characterize early behavioural signs in the ME7 model of murine prion disease. *European Journal of Neuroscience* 17: 2147–2155.

Ellis R, Langford D, and Masliah E (2007) HIV and antiretroviral therapy in the brain: Neuronal injury and repair. *Nature Reviews Neuroscience* 8: 33–44.

Ernst T, Chang L, Jovicich J, Ames N, and Arnold S (2002) Abnormal brain activation on functional MRI in cognitively asymptomatic HIV patients. *Neurology* 59: 1343–1349.

Everall IP, Heaton RK, Marcotte TD, et al. (1999) Cortical synaptic density is reduced in mild to moderate human immunodeficiency virus neurocognitive disorder. *Brain Pathology* 9: 209–217.

Gendelman HE, Zheng J, Coulter CL, et al. (1998) Suppression of inflammatory neurotoxins by highly active antiretroviral therapy in human immunodeficiency virus-associated dementia. *Journal of Infectious Diseases* 178: 1000–1007.

Hasbani MJ, Hyrc KL, Faddis BT, Romano C, and Goldberg MP (1998) Distinct roles for sodium, chloride, and calcium in excitotoxic dendritic injury and recovery. *Experimental Neurology* 154: 241–258.

Jeffrey M, Halliday WG, Bell J, et al. (2000) Synaptic loss associated with abnormal PrP precedes neuronal degeneration in the scrapie-infected murine hippocampus. *Neuropathology and Applied Neurobiology* 26: 41–54.

Masliah E, Heaton RK, Marcotte TD, et al. (1997) Dendritic injury is a pathological substrate for human immunodeficiency virus-related cognitive disorders. *Annals of Neurology* 42: 963–972.

Mattson MP, Keller JN, and Begley JG (1998) Evidence for synaptic apoptosis. *Experimental Neurology* 153: 35–48.

Inflammation in Neurodegenerative Disease and Injury

M G Tansey and S G Kernie, The University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

Features of Neuroinflammation

Inflammation can be defined in clinical, pathological, and molecular terms. The four cardinal signs of inflammation described by Celsius in AD 25, dolor, tumor, calor, and rubor (i.e., pain, swelling, heat, and redness), are not present in patients with neurodegenerative disease. Neuropathological and neuroradiological studies indicate that inflammation in neurodegenerative diseases is characterized by a glial response that in certain diseases may be present prior to significant loss of neurons. It was once believed that the blood–brain barrier (BBB) prevented access of immune cells to the brain and, as a result, the immune and central nervous system were relatively independent of each other. However, it has become clear that the permeability of the BBB can be regulated under normal conditions and may increase or become dysregulated in disease states. The inflammatory reaction in the brain associated with most acute or chronic neurodegenerative diseases is often termed 'neuroinflammation.' Neuroinflammation consists mainly of an innate immune response involving activation of glial cells and central production of cytokines, chemokines, prostaglandins, complement cascade proteins, and reactive oxygen species and reactive nitrogen species (ROS/RNS) in response to a central or peripheral immune challenge. The features of neuroinflammation include the presence of reactive astrocytes and activated microglial cells in the central nervous system (CNS). In certain neurodegenerative diseases, infiltration of lymphocytes (B and T cells) and polymorphonuclear cells occurs when the re is extensive breakdown of the BBB.

Glia: Microglia and astrocytes Microglia are the monocyte-derived resident macrophages of the brain responsible for immune surveillance. An initial physical or pathogenic event in the CNS is expected to elicit activation of microglia and secretion of factors that promote neuronal survival, with the purpose of limiting injury, protecting vulnerable neuronal populations, and aiding in repair processes. When the permeability of the BBB increases, recruitment of peripheral macrophages to the CNS results in expansion of the brain microglia population. Activated microglia produce ROS and RNS such as nitric oxide (NO) and secrete prostaglandins, chemokines, and cytokines, as well as neurotrophic factors (including glial-derived neurotrophic factor family ligands, or GFLs) that promote neuronal survival. If glial activation persists for extended periods, a cycle of chronic neuroinflammation may result and lead to elevation of local levels of cytokines, chemokines, prostaglandins, and other inflammatory mediators that can contribute to tissue damage. The mechanisms that regulate microglial activities in the brain are poorly understood. In recent years, a more thorough understanding of the functional link between microglial activation markers and specific cellular functions has evolved. New information about the mechanisms that regulate the balance of neuroprotective microglial activities (e.g., neurotrophic factor production, phagocytosis of debris, amyloid or senile plaque removal) and neurotoxic microglial activities (i.e., ROS/RNS, prostaglandin, cytokine, and chemokine overproduction) will be required to understand how dysregulation of this balance hastens the death of vulnerable neuronal populations.

A second type of glial cell, the astrocyte, resides in the CNS and is important for regulation of the permeability of the blood–brain barrier. The cellular processes of astrocytes make intimate contact with essentially all areas of the brain and are functionally coupled to neurons, oligodendrocytes, and other astrocytes. This tight coupling enables astrocytes to regulate the homeostatic environment (including regulation of the excitatory neurotransmitter glutamate) and promote proper functioning of the neuronal network. Inflammation in the CNS may disrupt this process either transiently or permanently, as cytokines have important roles in reprogramming gene expression in glia (both in microglia and in astrocytes) during injury and recovery. Upregulation of glial fibrillary acidic protein (GFAP) in astrocytes denotes reactive astrogliosis in the CNS and may be part of a response to acute injury or neurodegeneration that can lead to permanent scarring and tissue damage. It is becoming increasingly evident that astrocytes also have important roles in neurological disorders. Mouse models of disease have revealed astrocyte-specific pathologies that contribute to neurodegeneration. Development of animal models in which astrocyte-specific proteins and pathways have been manipulated has contributed to our understanding of the normal biology of astrocytes. It is expected that development of targeted therapies for pathways in which astrocytes participate to the detriment of neurons will be possible in humans and should ultimately influence the clinical treatment of neurodegenerative disorders.

Cytokines, chemokines, and lymphokines Cytokines, chemokines, and lymphokines are multifunctional immunoregulatory proteins secreted by cells of the immune system. Within the CNS, cytokines, most notably tumor necrosis factor-α (TNF-α), the interleukin-1 (IL-1) and interleukin-6 (IL-6) families, interferon-γ (IFN-γ), and transforming growth factor-β (TGF-β) act in context-dependent ways to modulate inflammatory processes that affect the permeability of the BBB. Cytokines can exert neuroprotective effects independent of their immunoregulating properties and some of them (e.g., TNF-α) have been shown to modulate neurotransmission. However, under certain conditions, cytokines and chemokines promote apoptosis of neurons, oligodendrocytes, and astrocytes as well as damage to myelinated axons. The outcome of cytokine action in the brain depends on the dynamics, cellular source, degree of compartmentalization of cytokine release, pathophysiological context, and presence of co-expressed factors. The levels of several cytokines (including TNF-α, IL-1, and IL-6) become elevated in a wide range of CNS disorders, including ischemia, trauma, multiple sclerosis, Alzheimer's disease (AD), and Parkinson's disease (PD). While there is no evidence to support a role for any of them in triggering any of these diseases, cytokine-driven neuroinflammation and neurotoxicity may modify disease progression in a number of neurodegenerative diseases (see later).

What Triggers Neuroinflammation?

Neuroinflammation may be triggered by protein aggregation and formation of inclusions arising from mutations (i.e., α-synuclein) or disruption of the ubiquitin–proteasome system (UPS), immunologic challenges (bacterial or viral infections), neuronal injury (brain trauma or stroke), or other epigenetic factors. Chronic inflammatory syndromes (rheumatoid arthritis, Crohn's disease, and multiple sclerosis) and environmental toxins (pesticides and particulate matter) are the most notable epigenetic factors likely to trigger neuroinflammation. These insults can increase the permeability of the BBB to allow infiltration of lymphocytes and macrophages into the brain parenchyma. The immune reactions initiated by viruses and bacteria may contribute to latent vulnerabilities which could become manifest with future immunologic challenges. Depending on context, duration, and type of inflammatory response, inflammation may be detrimental or beneficial to the individual. During trauma to the CNS, acute inflammatory responses are mounted to limit injury and to aid in neuronal repair. However, when inflammation becomes chronic, it may lead to tissue damage and is likely to be detrimental to neuronal survival. Acute injury in most tissues is followed by release of histamine, with resulting increases in vascular permeability (i.e., exudation of fluid into the injured tissue and neutrophil infiltration). In the brain, such a response could result in increased intracranial pressure and would be detrimental within the confines of the skull. Thus, it is possible that the downgraded response of resident microglia is an evolved response that permits the CNS to respond to injury in a way that will result in minimal brain damage.

How Does Neuroinflammation Affect Neurons?

At the molecular level, acute inflammation in the CNS is believed to involve the neuroprotective actions of microglia, and to be beneficial because it acts to limit tissue damage (as in ischemic stroke). In contrast, chronic neuroinflammation is more likely to involve microglial activities that promote ROS/RNS generation, enhance the unfolded protein response (UPR), and contribute to neurotoxicity. Inflammatory challenges may trigger clinical symptoms of neurological or psychiatric conditions. High levels of certain microglial-derived inflammatory mediators (e.g., TNF-α) often lead to enhanced oxidative stress in neurons because they contribute to glutathione depletion. In addition,, several of the downstream signaling pathways (including caspases) activated by inflammatory factors may converge to elicit mitochondrial dysfunction and death of vulnerable neuronal populations through apoptosis or other forms of programmed cell death.

Evidence for Inflammation in Neurodegenerative Diseases

Neurodegenerative diseases are characterized by the loss of specific neuronal populations and often by intraneuronal as well as extracellular accumulation of fibrillary materials. Formation of intracellular inclusion bodies may result from abnormal protein–protein interactions, aberrant protein folding, and/or dysregulation of the UPS. These conditions, often referred to as 'proteinopathies,' are thought to have a central role in the neuronal dysfunction that characterizes several neurodegenerative diseases. One common way in which a number of divergent molecular or cellular events (e.g., mutations, oxidation, protein misfolding, truncation, or aggregation) may all contribute to death of neurons is via activation of resident microglial populations in specific brain regions. If the host is unable to eliminate the initiating trigger (i.e., proteinopathy caused by a mutation), chronic activation of microglia and chronic neuroinflammation-derived

oxidative stress may contribute to progressive neuronal dysfunction. If neuroinflammation is in fact contributing to neurodegeneration of vulnerable populations, then timely use of certain anti-inflammatory drugs should offer neuroprotection and limit progression of disease. However, human clinical trials with anti-inflammatory drugs have had very limited success in the management of several neurodegenerative diseases to date. Possible reasons for the failure of these trials could be the timing of the treatment and the dosing regimens chosen for the trials; further investigation in this area is warranted before dismissing the possibility of potential long-term benefits of anti-inflammatory drugs.

Clinical and Epidemiological Evidence for Inflammation in PD

Over the past decade, a great wealth of new information has emerged to suggest that inflammation-derived oxidative stress and cytokine-dependent neurotoxicity are likely to contribute to nigrostriatal pathway degeneration and lead to idiopathic PD in humans. Postmortem analyses of brains from PD patients have confirmed the presence of inflammatory mediators in the area of substantia nigra (SN), where maximal destruction of vulnerable melanin-containing dopamine (DA)-producing neurons occurs in PD patients. Signs of inflammation included activated microglia and accumulation of cytokines (including TNF-α, IL-1β, IL-6, and IFN-γ). The advent of technologies (such as positron emission tomography brain scans) that can be conducted in living humans confirms that patients with idiopathic PD have markedly elevated microglia activation in the pons, basal ganglia, striatum, and frontal and temporal cortical regions, irrespective of the number of years with the disease, compared to age-matched healthy controls. As stated earlier, activated microglia overproduce cytokines (including TNF-α, IL-6, and IL-1β) which exert neurotoxic effects on DA neurons; and SN dopaminergic neurons may be uniquely vulnerable to neuroinflammatory insults that enhance oxidative stress. Elevated levels of cytokines may act in an autocrine manner to maintain activation of abundant numbers of microglia in the midbrain, potentiating inflammatory responses (autoamplification of ROS, NO, and superoxide radicals to form highly oxidizing peroxynitrite species) and creating an environment of oxidative stress which further hastens damage of DA neurons. Although the genes for various cytokines, chemokines, and acute-phase proteins have been surveyed and individual reports demonstrate that certain single nucleotide polymorphisms are overrepresented in certain cohorts of individuals affected with PD, most of these findings have not been replicated in independent studies. Meta-analyses of multiple such association studies are crucial to assess the overall genetic effect of cytokine gene polymorphisms on disease.

Evidence for Oxidative Stress and Inflammation in Experimental Models of PD

A better understanding of the extent to which environmental factors that elicit chronic inflammation modify the known genetic risks for developing PD is needed. The two neurotoxins most widely used to induce nigral DA neuron loss in rodents are 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) and 6-hydroxydopamine (6-OHDA). These neurotoxin models are accompanied by a robust glial reaction. Studies with nonhuman primates exposed to MPTP indicate that the cycle of neuroinflammation triggered by these neurotoxins persists long after the initial insult abates and may contribute to the progressive degeneration of the nigrostriatal pathway. Although each neurotoxin may trigger different initial cascades of events, they all consistently involve oxidative stress as the critical mechanism that elicits DA neuron death. In addition, there are endogenous toxins that can contribute to nigral DA neuron loss and to the risk of developing PD. These include tetrahydroisoquinolines (TIQs), which are produced via the reaction of DA with aldehydes and act to decrease dopaminergic neurotransmission, generate ROS, and trigger apoptosis. Nitric oxide (NO) also contributes to DA neuron death through mechanisms that, while not completely understood, are likely to involve mediation of excitotoxicity, activation of poly(ADP ribose) polymerase-1 (PARP-1), DNA damage, activation of caspase-dependent and-independent cell death, and/or nitrosylation of proteins, including α-synuclein and Parkin. Thus, general synthetic antioxidants as well as natural products with antioxidant properties (including green tea, omega oils, curcumin, and various varieties of berries) are being intensely investigated for their ability to provide DA neuroprotection in experimental models of PD and in the clinic.

Findings obtained from histopathologic, genetic, and pharmacologic studies in animals implicate a role for cytokine-driven inflammation and neurotoxicity in loss of nigral DA neurons induced by MPTP, 6-OHDA oxidative neurotoxin, and bacteriotoxin. Rapid increases in cytokine levels are detectable in rodent midbrain SN within hours of *in vivo* administration of 6-OHDA or MPTP. Pharmacologically, chronic infusion of various anti-inflammatory compounds, including cyclooxygenase-2 (COX-2)-selective nonsteroidal anti-inflammatory drugs (NSAIDs) or soluble TNF-selective inhibitors, rescues nigral DA

neurons from progressive degeneration and death. The higher sensitivity of nigral DA neurons to injury induced by neuroinflammatory mediators may be secondary to reduction of endogenous antioxidant capacity (i.e., glutathione depletion). In summary, an overwhelming number of studies now indicate that chronic elevation of certain cytokines in the CNS can contribute to progressive degeneration of nigral DA-producing neurons. These findings raise the interesting possibility that environmental triggers initiate cytokine-driven neuroinflammation and contribute to the development of PD in humans.

Anti-inflammatory Drugs and PD

Additional strong evidence that inflammatory processes modulate PD risk came from a prospective study of hospital workers; the study indicated that the incidence of sporadic PD among chronic users of over-the-counter NSAIDs was 46% lower than in age-matched non-NSAID users. NSAIDs scavenge free oxygen radicals and inhibit cyclooxygenase-2, which is known to participate in oxidation of DA. Although the specific NSAID target(s) that mediate(s) the effect associated with lower PD risk remain(s) unidentified, these findings suggest that neuroinflammatory processes contribute to DA neuron loss and development of PD in humans. In concert with findings derived from experimental models, these epidemiological findings raise the possibility that anti-inflammatory therapy could delay or prevent onset of PD.

Steroids have anti-inflammatory properties and have been shown to rescue DA neurons in rodent models of PD; however, because of their well-established side effects, they are not likely to be suitable for long-term use in humans. Clinical trials in PD patients using short-term administration of immunophilin ligands (which lack immunosuppressive properties of the parent compounds) have also had limited success. At the time of this writing, minocycline, the tetracycline derivative which readily crosses the BBB and is well-tolerated in humans, was under evaluation in phase III clinical trials to determine if it could alter long-term progression of PD and other neurodegenerative diseases. A reexamination of available classes of anti-inflammatory agents may be needed to identify other compounds that can cross the BBB. Alternatively, novel modes of therapeutic delivery for introduction of promising anti-inflammatory compounds that cannot readily cross the BBB will need to be developed. Time will tell if successful development and timely use of anti-inflammatory approaches can modify the progressive course of this disease and afford therapeutic benefit in patients with early signs of PD.

Inflammation in Acute CNS Injury: Traumatic Brain Injury

Inflammation is a hallmark of all forms of acquired brain injury (i.e., traumatic brain injury (TBI), cerebral ischemia or stroke, and spinal cord injury). Worldwide, millions of people are treated each year for significant head injury; a substantial proportion die, and many more are disabled. The progressive tissue damage that occurs following TBI has been attributed to cerebral inflammation and apoptotic cell death. Strategies to inhibit one or both of these pathways are being investigated as potential therapies for TBI patients (see later), though current therapy for TBI remains entirely supportive. The rationale behind mechanism-based drug design is to provide some degree of neuroprotection by preventing injured or vulnerable nerve cells from dying and/or to stimulate regenerative processes that could lead to the restoration or the formation of new connections lost on account of the injury. However, in the case of TBI, although it is clear that microglia activation occurs, the role of microglia in limiting versus promoting injury has not been clearly established. To add to an already complex scenario, significant activation of astroglia also occurs in TBI and recent evidence suggests that astrocytes may also participate in the neuroinflammatory process by regulating expression of heme oxygenase-1 (HO-1) in microglia and limiting ROS/RNS production.

Although there are no clinical treatments yet available to enhance repair of the damaged brain after trauma, a number of potential therapies have been investigated with varying degrees of success. For many years, corticosteroids had been used to treat head injuries in the emergency room. This popular practice was based on the rationale that if these drugs could reduce mortality by 1–2% they would have a significant impact given the large numbers of people that incur TBI each year. However, the largest head injury trial conducted to date has provided conclusive data to indicate that global immunosuppression of innate immune responses with corticosteroids actually negatively affects outcome after TBI. In 2001, after a successful pilot study, the Medical Research Council of Great Britain funded the CRASH (Corticosteroid Randomization after Significant Head Injury) trial to examine the effect of corticosteroids on death and disability after injury. Data obtained for nearly 10 000 patients at 6 months postintervention showed that the risk of death was higher in the corticosteroid group than in the placebo group, as was the risk of death or severe disability, with no evidence that the effect of corticosteroids differed by injury severity or time since injury. On the basis of this definitive study, it was concluded that corticosteroids

should not be used routinely in the treatment of head injury. Another large randomized, controlled trial using therapeutic hypothermia in the setting of moderate to severe TBI also failed to show an improvement. However, recent data suggest that one way that hypothermia may be helpful is when it is administered early enough to be an immune modulator. Based on this, there is currently another National Institutes of Health (NIH)-funded phase I clinical trial looking at the efficacy of hypothermia initiated early following TBI in children.

Given that TBI can induce progressive neurodegeneration in association with chronic inflammation and that chronic treatment with the nonsteroidal anti-inflammatory drug, ibuprofen, had been shown to improve functional and histopathologic outcome in mouse models of AD, ibuprofen was also investigated for its ability to improve long-term outcome following experimental TBI. Disappointingly, chronic ibuprofen administration was found to worsen cognitive outcome following TBI. These data may have important implications for TBI patients who are often prescribed NSAIDs for chronic pain.

Minocycline, a tetracycline derivative with anti-inflammatory properties, has been intensely studied and found to be therapeutically effective in various models of CNS injury and disease via mechanisms involving suppression of inflammation and apoptosis. Although individual studies have shown some positive effects with minocycline, further studies on minocycline in TBI are necessary in order to consider it as an effective therapy for brain-injured patients.

The neuroprotective properties of progesterone (PROG) and its enantiomers (ent-PROG) or its metabolite (allopregnanolone) have been investigated after TBI. Studies have shown that a single injection of progesterone can attenuate cerebral edema when administered during the first 24 h after TBI in rats, but this regimen did not always produce functional benefits. In contrast, steady-state progesterone treatment after TBI decreased edema and anxiety and increased activity, thus enhancing behavioral recovery. Therefore, chronic pharmacological administration of progesterone may prove to be more beneficial in translational and clinical testing compared to bolus injections over the same period of time. In addition, ent-PROG treatment has been shown to decrease cerebral edema, cell death mediators, inflammatory cytokines, and reactive gliosis, and to increase antioxidant activity in experimental models of TBI. Based on these compelling experimental data supporting the neuroprotective properties of progesterone, a phase II randomized, double-blind, placebo-controlled clinical trial was conducted at an urban level I trauma center to assess the safety and potential benefit of administering progesterone to patients with acute TBI. The study showed that patients with moderate TBI who survived their injury were more likely to have a moderate to good outcome if they received progesterone than were those who received the placebo.

The pro-apoptotic cytokine TNF as well as the impact of TNF neutralization has been extensively investigated following TBI in a number of different experimental models (e.g., percussion injury and controlled cortical injury) with a number of different outcomes, depending on the timing of the intervention. Overall, TNF was found to have dual effects: in certain instances it enhanced survival and in others it promoted apoptosis during secondary damage after TBI. These findings are not surprising given that TNF signaling activates numerous downstream pathways that have the capacity to differentially regulate activation of microglia – nuclear factor-kappa B (NF-κB)- and c-Jun N-terminal kinase (JNK)/AP-1-dependent gene transcription. Multiple studies have suggested that the protective effects of hypothermia following TBI may be mediated via TNF-R1 activation of the JNK and caspase-3 pathways.

Last, it is worth noting that repetitive TBI has been linked to the development of AD. Development of diffuse β-amyloid (Aβ) deposits and tau pathology approximately 1–3 weeks after head trauma has been reported clinically, and head trauma is the environmental factor most consistently associated with AD or PD risk. The mechanisms underlying this process are not clear but are likely to be influenced by apolipoprotein E (apoE) gene status. The ε4 allele of the *APOE* gene (*APOE* ε4), which confers higher AD risk to individuals, has been associated with both the vascular risk factors and the consequences (cognitive impairment, dementia) of white-matter lesions (WMLs), suggesting that *APOE* ε4 carriers may have diminished capacity for neuronal repair after TBI. Further studies are warranted to clarify this relationship.

Conclusions

Over the past decade, the concept of the brain as an immune-privileged organ has been rejected in light of unequivocal evidence that the permeability of the blood–brain barrier can be regulated under normal conditions and may become dysregulated in chronic and acute neurodegenerative conditions. Reactive astrocytes, activated microglial cells, and overproduction of inflammatory mediators orchestrate neuroinflammatory responses that either promote or compromise neuronal survival, depending on type, duration, and context. In general, acute inflammatory responses during trauma to the CNS involve the neuroprotective actions of microglia aimed at limiting injury and enhancing neuronal repair. When

inflammation becomes chronic, microglial activities create an environment that often enhances oxidative stress in neurons, contributes to neuronal dysfunction, and hastens neurodegeneration. While there is no evidence to support a role for any specific inflammatory mediator in the etiology of any neurodegenerative disease, cytokines in the brain and/or cerebrospinal fluid become elevated in a wide range of CNS disorders, including ischemia, trauma, multiple sclerosis, AD, and PD, raising the possibility that they modify (i.e., accelerate) disease progression. Collectively, whether cytokine action has beneficial or harmful outcomes in the brain depends on the dynamics, cellular source, degree of compartmentalization of cytokine release, pathophysiological context, and presence of co-expressed factors. Nigral DA neurons may be uniquely vulnerable to neuroinflammatory insults that enhance oxidative stress; and over the past decade, a great wealth of new information has emerged to suggest that inflammation-derived oxidative stress and cytokine-dependent neurotoxicity may contribute to nigrostriatal pathway degeneration and increase the risk for development of idiopathic PD in humans. Following acute TBI, cerebral inflammation and apoptotic cell death contribute to progressive tissue damage and may increase the risk for development of AD by promoting diffuse deposits of Aβ. No highly effective clinical treatments are available to enhance repair of the damaged brain, and use of high-dose steroids is now contraindicated in the management of TBI. Further basic science and more clinical research are needed to establish a causal link between neuroinflammation and neurodegeneration. If one is found, timely use of anti-inflammatory drugs may prove effective in slowing down the progressive neuronal death in chronic neurodegenerative disease and acute CNS injury.

See also: Animal Models of Parkinson's Disease; Astrocyte: Response to Injury; Brain Trauma; Cell Replacement Therapy: Parkinson's Disease; Glial Responses to Injury; Glial Glutamate Transporters; Microglial Response to Injury; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neural Stem Cells and CNS Diseases; Oxidative Damage in Neurodegeneration and Injury; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Arai H, Furuya T, Mizuno Y, et al. (2006) Inflammation and infection in Parkinson's disease. *Histology and Histopathology* 21: 673–678.

Block ML and Hong JS (2005) Microglia and inflammation-mediated neurodegeneration: Multiple triggers with a common mechanism. *Progress in Neurobiology* 76: 77–98.

Chen H, Zhang SM, Hernan MA, et al. (2003) Nonsteroidal anti-inflammatory drugs and the risk of Parkinson disease. *Archives of Neurology* 60: 1059–1064.

Chen H, Jacobs E, Schwarzschild MA, et al. (2005) Nonsteroidal antiinflammatory drug use and the risk of Parkinson's disease. *Annals of Neurology* 59: 988–989.

Edwards P, Arango M, Balica L, et al. (2005) Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury-outcomes at 6 months. *Lancet* 365: 1957–1959.

Hald A and Lotharius J (2005) Oxidative stress and inflammation in Parkinson's disease: Is there a causal link? *Experimental Neurology* 193: 279–290.

Hirsch EC, Hunot S, and Hartmann A (2005) Neuroinflammatory processes in Parkinson's disease. *Parkinsonism & Related Disorders* 11(supplement 1): S9–S15.

Jellinger KA (2004) Traumatic brain injury as a risk factor for Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry* 75: 511–512.

McGeer PL, Itagaki S, Boyes BE, et al. (1988) Reactive microglia are positive for HLA-DR in the substantia nigra of Parkinson's and Alzheimer's disease brains. *Neurology* 38: 1285–1291.

McGeer PL, Schwab C, Parent A, et al. (2003) Presence of reactive microglia in monkey substantia nigra years after 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine administration. *Annals of Neurology* 54: 599–604.

Minghetti L (2005) Role of inflammation in neurodegenerative diseases. *Current Opinion in Neurology* 18: 315–321.

Mrak RE and Griffin WS (2005) Glia and their cytokines in progression of neurodegeneration. *Neurobiology of Aging* 26: 349–354.

Sanchez-Pernaute R, Ferree A, Cooper O, et al. (2004) Selective COX-2 inhibition prevents progressive dopamine neuron degeneration in a rat model of Parkinson's disease. *Journal of Neuroinflammation* 1: 6.

Wersinger C and Sidhu A (2002) Inflammation and Parkinson's disease. *Current Drug Targets – Inflammation & Allergy* 1: 221–242.

Relevant Websites

http://www.apdaparkinson.org – American Parkinson Disease Association Inc.

http://www.pdgene.org – Database for Parkinson's disease genetic association studies, Massachusetts General Hospital/Harvard Medical School, The Michael J. Fox Foundation and Alzheimer Research Forum.

http://www.braininjury.com – Medical, legal, and informational resources for persons dealing with traumatic brain injury.

http://www.parkinson.org – National Parkinson Foundation.

http://www.pdf.org – Parkinson's Disease Foundation.

http://www.braintrauma.org – The Brain Trauma Foundation.

http://www.michaeljfox.org – The Michael J. Fox Foundation for Parkinson's Research.

http://www.worldpdcongress.org – World Parkinson Congress.

Information Coding

B J Richmond, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Neuronal Codes

Neural coding describes the study of information processing by neurons. Such studies seek to learn what information is used, and how information is transformed as it passes from one processing stage to another. The field of neural coding seeks to synthesize information arising from many levels of analysis and to explain how integrated behavior arises from the cooperative activity of the neurons in the brain.

Much is known about the biophysics of neuronal responses – that is, how a spike is generated, how the spike acts on the axonal terminal to cause transmitter release, and how transmitters act on the target neuronal receptors. However, the impact of series of spikes is not well known, and it is even less clear what information has been encoded and how that information will be utilized in subsequent processing stages.

It is natural to consider the brain as an information processing machine. It takes in information from physical changes in the environment, including light, touch, and sound, integrates this information with remembered or genetically coded information and then produces organized behavior. In any information processing system, issues to be considered are what is being encoded, what is the code used to transmit the information, how reliable (noisy or not) is the code, and how the information is utilized or decoded.

At the base of this cascade of information processing is the information transformed and transmitted by single neurons. The signal on dendrites and cell body is reflected in fluctuations in the potential difference across the membrane. The signal that is transmitted down the axon is quite different; it is the action potential. This action potential propagates rapidly down the axon. At the axon terminal the arrival of the action potential generally causes the release of a transmitter that affects the membrane of the target neuron. The action potentials mark times at which the cell membrane at the neuronal cell body reaches the firing threshold. Thus, there is a sequence of action potentials containing information about the membrane potential at the neuronal cell body, and it is the information carried by the train of action potentials that provides information at the projection targets for the neurons. This train of action potentials (the spike train) can be considered as elements of a neural code.

Average Rate Codes

The number of action potentials, or 'spikes,' is the critical feature of the neural code. For example, for primary visual cortex, the number of spikes provides information about the orientation of a bar, as shown in Figure 1, or, for primary motor cortex, the direction of reaching. For the two examples, the tuning curve can be parameterized by some simple function (e.g., as has been done for reaching direction in primary motor cortex with a cosine or some other similar smooth, unimodal function). This 'tuning curve' then provides a means to estimate the direction of the bar (or movement) from the neural response (Figure 2). In higher brain regions (e.g., inferior temporal cortex of the monkey), the neurons respond most strongly to a particular complex object, sometimes a single face or a hand, and a tuning curve for this 'spike count code' can be determined empirically. So far it is generally not possible to parameterize the tuning curve as simply as is done with the aforementioned directional tuning examples. However, for considering the principles underlying neural codes, parameterizing the tuning curve is not critical. Decoding can be studied in any brain region for which the equivalent of this 'tuning curve' is available. The results of experiments using measured tuning curves only become easier to interpret as the knowledge about the tuning function improves. Hence, over the history of single neuronal recording, there has been appropriate emphasis on measuring such 'tuning curves.'

If every presentation of a stimulus elicited the same response, decoding the response would be straightforward; it would only be necessary to measure the response and look up which condition or conditions elicit that response. However, as can be seen in Figures 1 and 2, the responses are variable. Many of the possible responses emitted by the primary visual cortical neuron could have been elicited by any of the orientations of the bar (here measured in spike count). In the face of such variability, it is not possible to know unequivocally which stimulus elicited the response. It is possible, however, to determine how likely or probable any particular stimulus might be. In the case shown in Figures 1 and 2, one can pick the stimulus that is most likely to have elicited the particular response, and it then becomes possible to estimate how many neurons would be needed to decode the orientation of the stimulus. Thus, even though variability limits the information that a single neuron can encode, a probabilistic estimate of the orientation can be developed. Then, based on these probabilities, some rule for choosing a stimulus can be identified.

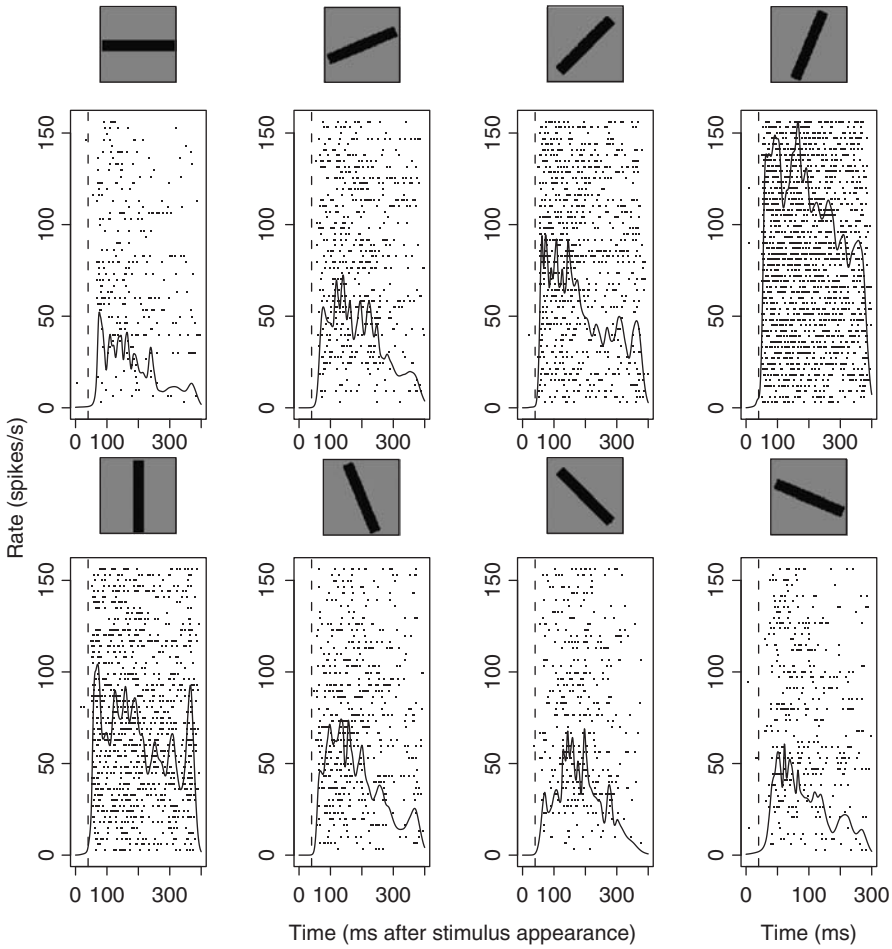

Figure 1 Example of orientation sensitivity of a neuron in primary visual cortex. The black bar at the top of each panel was flashed onto the receptive field (i.e., the position in the visual field that, for that neuron, is sensitive to visual stimulation) of the neuron for 600 ms. The dots in each horizontal row in the panels represent the times that the neuron emitted an action potential. Each row shows a different presentation of the stimulus. Time on the *x*-axis is relative to the stimulus appearance. The third black bar from the left (top and bottom) elicited the most intense response (the largest activity) from this neuron. The large variation in responsiveness from one stimulus presentation to the next can be seen in that successive rows of dots have different numbers of dots, and the dots are in different places for all of these black bars. The smooth black line in each panel is a smoothed average of the spikes across all of the stimulus presentations. This is taken as an estimate of the overall driving of this neuron by its inputs. Modified from Doya K, Ishii S, Pouget A, and Rao RPN (eds.) (2007) *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT Press.

The usual rule, and the easiest to implement, is to choose the stimulus that is most likely to have elicited the response; this is the maximum likelihood estimate.

It is because of the response variability that tools from signal processing, statistical pattern recognition, and information theory have become important for studying neural coding. Even though it is obvious that the pattern of spikes can carry additional information, it is easiest to start with the assumption that the spikes are distributed randomly through time, and that only the spike count changes. It then becomes straightforward to add information from the time sequence of the spikes after developing the principles by which to assess the code.

As a first approximation, assume that there is no information in the pattern of spikes. Then measure the information using the spike count alone as a neural code. The only process for which the spikes are equally likely at all times – that is, there is no correlation over time – and hence for which there can be no information in the timing is a Poisson process. A Poisson process is a special case of point process – that is, a process in which there are discrete events that can all be considered identical occurring over time. The only number needed to characterize a Poisson process is the average rate of the events. The variability is proportional to the mean firing rate (Figure 3), with the variability being higher when the firing rate is higher, although the mean is not equal to the variance, as would be the case for a pure Poisson process. The source of this variability is not really known, although if a

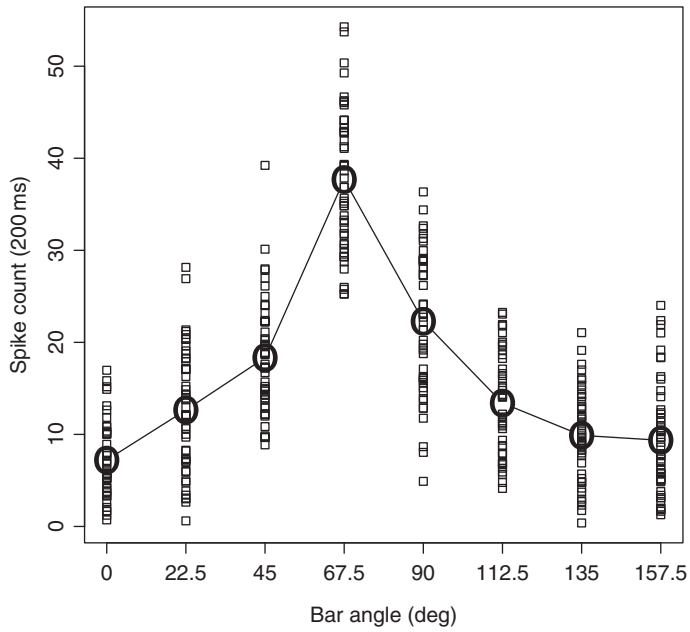

Figure 2 Orientation tuning for the neuron in **Figure 1**. The plot shows the bar angle on the *x*-axis and number of spikes on the *y*-axis. Each small square is the number of spikes emitted in response to one stimulus presentation. The large circles show the average, and the line connects the averages for each viewing. This illustrates the wide variation in the number of spikes emitted even for the same bar on different presentations (the squares in each column). Nonetheless, there is a clear and reliable tendency across trials. Trial-by-trial estimation of the bar orientation can only be done in a probabilistic way, making it necessary to consider how the responses of populations of neurons might be read and decoded. Modified from Doya K, Ishii S, Pouget A, and Rao RPN (eds.) (2007) *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT Press.

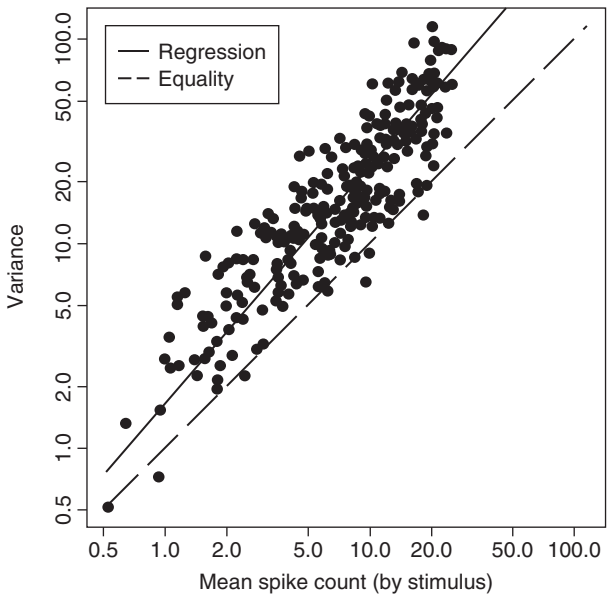

Figure 3 Scatter plot of means vs. variances. Each black dot represents the mean (*x*-axis) and variance (*y*-axis) elicited across many presentations of a single black and white stimulus on the receptive field of a V1 neuron. The solid line is the regression line, and the dashed line is the equality line where the points would lie if the responses arose from a Poisson process. The regression line is parallel to and above the line expected for a Poisson process. From this it can be concluded that the responses have some internal timing structure. From experience it is known that this type of overdispersion is related to point processes in which the points have a tendency to be in groups or 'bursts.'

visual stimulus changes very rapidly, responses can be considerably less variable. It is obvious that the pattern of the spikes over time might be carrying information, and simple considerations of the biophysics of information across synapses make it clear that this temporal 'coding' could easily influence the target neuron.

Two methods commonly used to quantify the effect of this variability are classical analysis of variance (ANOVA), with each stimulus being one level in the stimulus factor (or if a continuous parameter is used to characterize the signal, linear regression), and information theory, which estimates how much the uncertainty about the stimulus being shown is reduced by seeing the response. In ANOVA, in addition to the *p*-value (the probability that all of the levels are the same), the amount of variance explained by each factor (in our example, the orientation) is calculated, and the amount of variance that is left unexplained in this 'model' of the data is also given. Thus, ANOVA (or linear regression) allows the calculation of the signal-to-noise ratio (the explained variance divided by the residual variance). This is a powerful and straightforward approach to quantifying the relation between a set of experimental conditions and the discriminability among conditions that can be achieved using the neural responses. Thus, for the example in **Figures 1**

and 2, 70% of the variance can be accounted for, by orientation, and rest is unexplained (for, the purpose here, noise).

Temporal Codes

If it were certain that the responses could be characterized by the spike count or rate alone, tools from linear statistical modeling, such as regression or ANOVA (with possibly some generalization to other types of statistical linear modeling), would be adequate to quantify the relations between stimuli or experimental conditions and responses. However, the responses occur over time, and, as can be seen in Figure 4, the rate varies as a function of time. It is natural to ask how much better the stimulus can be determined by using the time course of response, as compared to using the response strength alone. In the example shown in Figure 4, the total number of spikes and the variances of these two sets of responses are the same when counted during the whole period of stimulus presentation. It is evident that the variability in both the number and timing of spikes will limit the information carried by the responses. However, it is also clear that the envelope of the firing rate (the low-pass filtered version of the spikes) is different across the two stimuli. Thus, even though these two sets of responses have the same number of spikes over the whole counting interval, the time course of when the spikes occur appears to be different.

There is an ongoing debate about how precisely the timing of spikes needs to be measured to capture all of the information in spike trains. The way for an analysis to become sensitive to the time course of the response is to break the response into epochs of different lengths, and ask whether the information increases as the epochs become smaller. This makes the measurement of the time course more and more precise – that is, as the measurement of the time course becomes more precise. As the epochs become shorter (and there are more of them), the variations in the response that is represented become finer: the response has more 'wiggles.' The difficulty with this approach is that as the epochs become shorter, there are more of them, and the estimate of the firing in each epoch becomes less certain.

One straightforward way to study the temporal pattern of the responses is to use the principal components of the set of responses. The principal components are the eigenvectors of the data domain covariance matrix, and form an orthogonal set of, in this case, temporal waveforms derived from the data. Each response can be represented as a weighted sum of principal components. Information theory is a valuable tool for assessing these types of multidimensional codes, or codes with more than one element. To use information theory, the basic definition of transmitted or mutual information is required. The mutual or transmitted information, $H(r, s)$, is the difference between the entropy of the spike train (or stimulus),

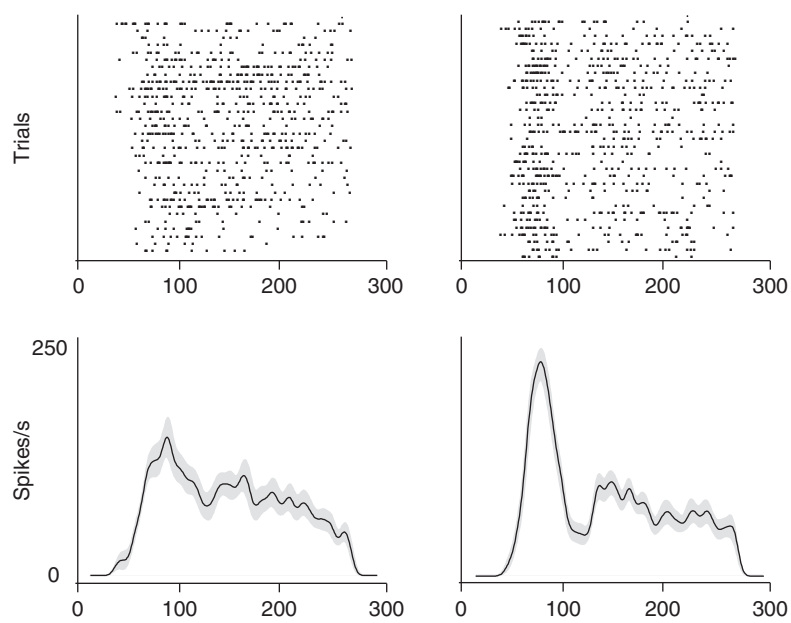

Figure 4 Raster dot diagrams and average spike densities for two stimuli eliciting responses with the same number of spikes. It is clear from this diagram that the pattern of spikes over time, or the pattern of the driving force over time, is different, and ought to be useful for using the spike trains to distinguish the stimuli.

$H(r)$, and the entropy of the response when a stimulus is presented, $H(r|s)$; $H(s, r) = H(s) - H(r|s)$. The entropy is

$$\sum_s p(s)\log(p(s))$$

and the entropy of the stimulus given the response is

$$\sum_s p(s,r)\log(p(s|r))$$

The r and s can be interchanged because mutual information is symmetrical with regard to the sender and receiver of a message, even if the physical channel might not actually be symmetrical (or bidirectional). Information has the strength that the representation of the stimulus and/or the response can be multivariate. With this approach it is possible to compare transmitted or mutual information from different codes. This, then, provides measurements to assess whether one assumed code carries more information than another assumed code.

The responses in Figure 4 are decomposed using their first few principal components; in Figure 5, it can be seen that the value of the first principal component is similar for the two responses, as the spike counts are. The coefficients of the second principal components, however, here have opposite signs. This is because the second principal component has a shape that readily represents the dip (Figure 4, left) or lack of it (Figure 4, right) seen in these two responses. The weights then become the code elements. The information from one, two, or more principal components can be used as candidate neural codes. Using more than the first two principal components adds little, if any, additional information in this example. For the cases where the principal

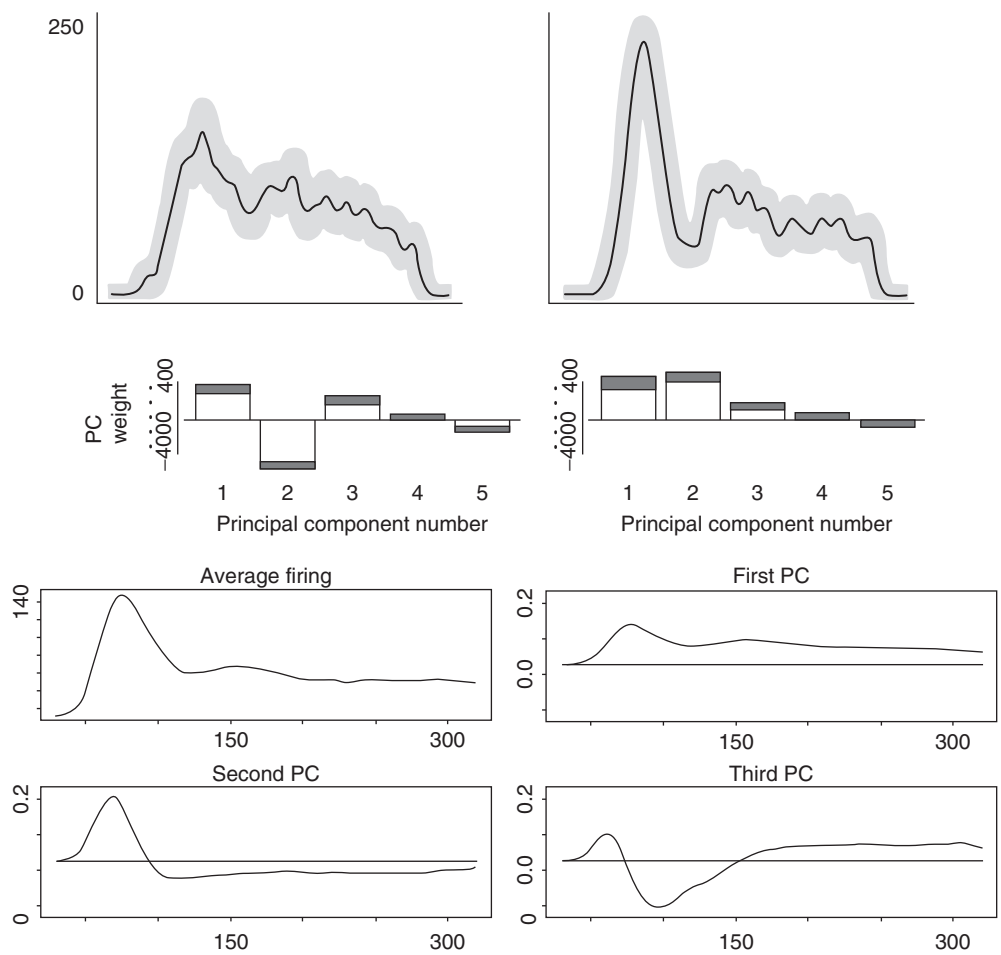

Figure 5 Principal decomposition of the two responses shown in Figure 3. The principal components were derived from the covariance matrix calculated from all of the responses to all of the patterns. The average spike densities are shown at the top. The middle panels show the weightings of the first five principal components (PC). The bottom panels show the shapes of the principal components (the eigenvectors of the covariance matrix), as modulations over time. The time-dependent mean (shown as the first waveform, labeled 'Average firing') was subtracted from each response before the covariance matrix was built. The second principal component (labeled 'Second PC') has a shape that, when added to the average, will accentuate the peak or blunt it (by subtracting or inverting this waveform).

components have been used to distinguish responses, the number of them that carry independent information over the preceding ones has been small, between 2 and 5.

In general, slow changes in the neuronal firing are represented by the lower numbered principal components, and as the principal component number goes up, the precision of the timing patterns increases. The need for this small number of principal components suggests that the spikes do not generally need to be represented very exactly in time – that is, blurring them out in time does not hurt the potential decodability of the responses. In the visual system, several experiments have shown that the precision of the spike timing is on the order of 10–20 ms, or longer in these experiments using stimuli that appear and remain on. In contrast to these measurements, if the stimulus is rapidly changing (i.e., if we repeatedly hit the system hard), the responses have spikes tightly time-locked to some temporal feature of the stimulus time series.

The need for more than one principal component suggests that more than one feature of the stimulus (or experimental condition) is encoded (i.e., information is multiplexed). One question is whether different parts of the neuronal responses can be identified with different stimulus characteristics. There are two types of response structures in primary visual cortex that have been given such identifications. These structures are single versus doublet spikes (i.e., whether spikes tend to be grouped into pairs or whether they occur singly). It has been shown that, for complex cells, the amount of clustering is related (tuned) for spatial frequency, and the single spikes are tuned for contrast. It has also been shown for complex cells that the response strength is closely tuned for orientation across a wide range of contrasts, whereas the latency changes substantially with contrast and response strength changes substantially less.

An alternative explanation for the variability is that it arises from encoding sources within the nervous system that do not covary with the stimulus (e.g., arousal or focal attention). These factors might then give rise to unaccounted for variability that is not true noise. At a minimum, the variability will have to be described and its influence on the organization of neuronal responses must be taken in account. The variability is seen both in the number and timing of spikes. Although it was for a long time not completely appreciated, it is now clear that the variability in the number of spikes influences our estimates of relating the number of spikes to their timing.

It is known that the variance becomes greater as the spike count increases. If we plot the coefficients of the variance of the second principal component against its mean for each stimulus, there is only a weak linear relation. The variance of the second principal component, however, is related to the spike count. This can be understood in stochastic terms. Consider stochastic sampling of a uniform point process. If the spikes are placed stochastically, the number of combinations of spikes that arise by chance is

$$\frac{n!}{k!(n-k)!}$$

or, n choose k, where n is the number of bins for the time period (usually taken to be 1 ms) and k is the number of spikes to be placed in the bins. If the number of spikes changes, the number of patterns that occur by chance necessarily also changes; n, the number of bins, is constant, and, k, the number of spikes, changes stochastically across stimulus presentations. For example, 200 choose 2 yields 19 900 combinations of spikes by chance, and 200 choose 3 yields 1 313 400 combinations. Since a small change in spike count can dramatically change the number of spike patterns expected by chance, any effort to describe what happens must first and carefully account for the variability in spike count. This then leads to a line of work to evaluate the impact of this necessary relation between spike count and numbers of spike patterns.

Finally, there is one more indication of this relation between the number of spikes and the number of spike patterns. If the coefficients of the first and second principal components from one neuron are placed on a scatter plot, they fill a cone, with the apex at the lowest spike count. This shows that the number of response types grows with the spike count (and would continue to grow until half the bins could be filled). Thus, when there are only a few spikes, there are only a relatively few spike patterns available for coding.

For a random process, even if not strictly Poisson, the null hypothesis that the spikes are randomly placed so that the overall rate is duplicated can be constructed using order statistics. Order statistics give the unconditional probability distributions for the kth event out of n events. For example, we can calculate the probability distribution for the third event or spike out 10. Order statistics explicitly needs the counts, the density of events (peristimulus time histogram, or PSTH), and the cumulative distribution, all of which we can estimate from data. However, we do not just want this unconditional distribution. Rather if the conditional distribution is available, a decoder or an accurate simulator can be constructed.

The basic equation of order statistics is as follows:

$$h(t) = \binom{n}{k} kF^{k-1}(t)f(t)[1 - F^{n-k}(t)]$$

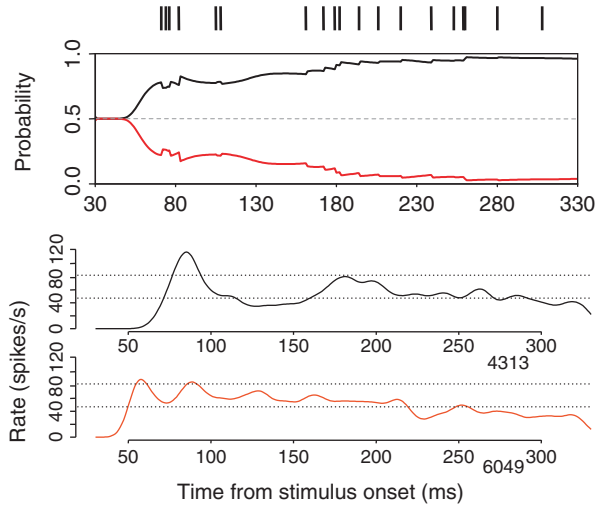

Figure 6 Decoding the spike train at the top for the two responses shown in Figure 4. The spike densities are shown again at the bottom in red and black. The *x*-axis on all three panels is time (in milliseconds) after stimulus appearance. The vertical axis on the top panel is the probability of guessing stimulus 1 (black) or stimulus 2 (red). For this spike train, shown by the black vertical tick marks above the top panel, the guessing probability rises quickly (and is correct). Note that the confidence of the guess can decrease at times. It turns out that the latency difference is a key feature for correct decoding in this example. Having a delay in the appearance of the first spike makes the likelihood of stimulus 1 very high.

where $h_1(t)$ is the probability of the spike k at time t; n choose k times k is a normalizing factor, and $F(t)$ is the cumulative distribution from $f(t)$, the rate function over time (estimated as the spike density or peristimulus time histogram). By working recursively, the conditional estimate, based on the time or spikes that have already gone by, can be constructed. This is simplified to be the probability of the first spike over time, $h_1(t)$.

With this recursive relation each successive spike can be considered the next first spike. Next, to construct the probability of the stimulus, regardless of the spike count, the order statistics are weighted by the probability of that spike count occurring, and these are summed. Then the probability of a stimulus s, at time t, through Bayes' theorem, can be estimated, and from this, the probability of the stimulus $p(s, t)$, at time t, can be calculated. Using these new probabilities as the new $p(s|t)$ for $p(s)$, decoding the spike train can be carried out spike by spike. Stepping through the spike train, $p(s)$ is updated. Using a maximum-likelihood rule, for example, a guess is made as to what stimulus is eliciting the spike train (Figure 6).

Population Coding

It is certain that activity must be coordinated among large ensembles of neurons, but it is not yet known how information is handled in ensembles. The simplest hypothesis is that the responses of neurons are summed or averaged, with the weight being the experimental condition that is weighted by the tuning in the tuning curve. This vector coding approach is the most influential model of population coding. It is relatively easy to compute. The experimenter can even measure the activity of neurons at different times. It is also straightforward to calculate how many neurons with independent noise are needed to decode from the responses in any condition. This approach is very important, also, for understanding the assumptions underlying mass population measurement approaches such as functional imaging (which actually measures blood flow changes in most cases) and field potentials. These are popular surrogates for having measurements of neural activity, spike trains, membrane potentials, and/or synaptic activity. These approaches assume that the average (presumably some sort of vector average, as in the vector coding approach) reveals functional properties.

The vector averaging code is powerful and is very convenient for computing the results from measuring the tuning curves of lots of neurons. However, it seems likely that the patterns of activity that occur simultaneously, or at least in a coordinated manner, might play an important role. The issue is whether spikes arriving from different inputs to a target neuron have different effects on the target, depending on their relative arrival times. The biophysics of neurons suggests that this should be the case. Before going forward, it is important to point out that the precision with which the coordinated activity must occur is not settled, so there are many investigators trying to make the needed measurements.

Another consideration is whether it matters which neuron carries the spikes. Can the spikes from different neurons be mixed without any loss of information? In the measurements made to this time, the information is greater when the spikes from two neurons are kept separate – that is, it helps to know which neuron carried the spike. This measurement has only been made a couple of times, and must be repeated because it suggests that population measures that only make use of the ensemble average activity will underestimate the information-carrying capacity of the tissue being considered.

A great deal of work has been carried out to look at coordination or synchronization of spikes. The question remains about how precisely the spikes must be represented to account for the information they carry. It is especially important but difficult to determine whether there is information that depends on codes across neuron (e.g., imagine a code based on the difference between the responses of two or more neurons). An intriguing hypothesis about cortical processing is

that there are patterns of spikes hidden within the apparently noisy responses of the aforementioned sort. This hypothesis, the synfire chain hypothesis, arises from strong theoretical considerations, in turn arising from straightforward, elegant considerations about cortical connectivity and neuronal biophysical properties. There is some evidence supporting this type of coding, but the number of these hidden patterns is small even in the best examples, so, although this is an elegant idea, the support is not yet very strong.

Another intriguing code that has been suggested is the order of neuronal recruitment. This code gets theoretical consideration, and there is a small amount of evidence that it could be used, at least in the peripheral somatosensory system. It is elegant and simple, but in the face of neurons with ongoing background activity it might be difficult to implement in practice, but it is mentioned here because it emphasizes that latency to onset of firing might (almost certainly does) play an important role in coordination of responses across neurons.

Finally, it seems plausible, and some would say even likely, that there might be dynamically changing neuronal ensembles, especially during learning, and under those circumstances there might be patterns of spikes within and across neurons that are strongly related to cognitive events. A lot of new work is being carried out to collect data from several neurons, even tens of neurons. The evidence about how signals are combined is not at all conclusive. With the advent of new tools from molecular genetics, it is becoming possible to visualize individual members of simple ensembles and make optical recordings of them. Now will come the exciting work of learning how the neural codes from different members of an ensemble are related.

See also: Decoding Neuron Transcriptome by SAGE; Natural Images: Coding Efficiency; Neural Synchrony and Feature Binding; Neural Coding of Spatial Representations; Neural Coding in Primary Motor Cortex; Neural Oscillators and Dymanmical Systems Models; Olfactory Coding; Perception of Surfaces and Forms; Perceptual Learning: Neural Mechanisms; Perceptual Learning and Sensory Plasticity; Prefrontal Contributions to Reward Encoding; Spiking Neuron Models; Vision: Mechanisms of Orientation, Direction and Depth; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging; Visual Attention; Visual Motion Models.

Further Reading

Abeles M (1991) *Corticonics*. Cambridge, UK: Cambridge University Press.

Optican LM and Richmond BJ (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *Journal of Neurophysiology* 57: 162–178.

Reich DS, Mechler F, and Victor JD (2001) Independent and redundant information in nearby cortical neurons. *Science* 294: 2566–2568.

Reich DS, Victor JD, Knight BW, et al. (1997) Response variability and timing precision of neuronal spike trains *in vivo*. *Journal of Neurophysiology* 77: 2836–2841.

Richmond BJ, Optican LM, Podell M, et al. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. I. Response characteristics. *Journal of Neurophysiology* 57: 132–146.

Rieke F, Warland D, de RuytervanSteveninck R, et al. (1996) *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Victor JD (2005) Spike train metrics. *Current Opinion in Neurobiology* 15: 585–592.

Wiener MC and Richmond BJ (2003) Decoding spike trains instant by instant using order statistics and the mixture-of-Poissons model. *Journal of Neuroscience* 23: 2394–2406.

Inherited Macular Degenerations: Animal Models

B J Song and S H Tsang, Columbia College of Physicians & Surgeons, New York, NY, USA
D B Farber, UCLA School of Medicine, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Macular degenerations constitute a heterogeneous group of diseases that involve mainly the macular region of the eye (Figure 1), ultimately causing a defect in central vision. In contrast to global retinal diseases and cone predominant degenerations, macular dystrophies tend to have normal full-field electroretinograms (ERGs; tests that measure the electrical responses generated by various cell types in the retina when it is stimulated by increasing amounts of light) and peripheral visual fields (the total area in which objects can be seen in the peripheral vision while the eye is focused on a central point). Pathologically, they are characterized by atrophy of the retinal pigment epithelium (RPE) and photoreceptor degeneration. As a whole, macular degenerations are the most common cause of irreversible blindness in the general population, particularly among the elderly in the Western world.

Recent technological advances have led to significant developments in genetic engineering and have allowed researchers to manipulate the expression of genes and generate accurate and reproducible disease models. This is critical for identifying the underlying mechanisms by which diseases occur. Here, we summarize the most significant macular degenerations that are currently known and describe the animal models that are available for their study.

Best Disease

Best disease, also known as vitelliform macular dystrophy, is an autosomal-dominant congenital disease that affects the central retina. Its name is derived from the appearance of the classic vitelliform (or egg yolk-like) lesion seen in the macula on clinical examination (Figures 2(a) and 2(b)), which results from excess lipofuscin (the name given to fine yellow-brown pigment granules composed of lipid-containing residues of digested cellular material) in the RPE cells across the fundus as well as the presence of degenerating RPE cells located between the foveal RPE and Bruch's membrane. In addition to the vitelliform lesion, there is a loss of foveal photoreceptors in this disorder.

The eye has a standing electrical potential across it, with the front of the globe positive and the back negative, that is generated largely by the transepithelial potential across the RPE. Retinal illumination causes an initial rapid fall in the standing potential (the fast oscillation), followed by a slow rise (the light response). The clinical electro-oculogram (EOG) measures the amplitude of the standing potential and light response. The ratio of the light peak to the dark trough is called the Arden ratio. The diagnosis of Best disease is based on a reduced Arden ratio on the EOG (Figure 2(c)).

It is now known that Best disease is caused by a mutation in the *VMD2* gene found on chromosome 11q13. *VMD2* encodes a protein called bestrophin, which acts as a chloride channel in the basolateral membrane of the RPE.

Model of Best Disease

Recently, an acute model of Best disease was generated in rats. Constructs containing mutant forms of bestrophin (W93C or R218C) in replication-defective adenoviral vectors were subretinally injected into rat eyes, resulting in transient overexpression of the bestrophin mutants in the RPE. The treated animals had all the electrophysiological characteristics of Best macular dystrophy – diminished light peak response in the absence of a-wave, b-wave, and c-wave abnormalities. Although *VMD2* has been identified as the major gene responsible for Best disease, a stable genetic model has yet to be reported.

Stargardt Disease

Unlike the other common macular dystrophies (most of which have a dominant inheritance pattern), Stargardt disease has an autosomal recessive mode of inheritance. It is characterized by an abnormally high accumulation of lipofuscin in the RPE that leads to atrophy and the appearance of a bull's eye-like lesion on the fundus (Figure 3). Yellowish flecks (usually called fish-tail flecks) are also seen at the level of the RPE – a phenomenon known as fundus flavimaculatus, another name given to Stargardt disease. *ABCA4*, also known as *ABCR*, is the gene responsible for Stargardt disease. The *ABCA4* gene encodes a protein called rim, which is a transmembrane protein involved in the transport of vitamin A intermediates, specifically N-retinylidine-phosphatidylethanol-amine (N-RPE), to the RPE. As its name suggests, the rim protein is expressed in the rims of photoreceptor disc membranes (Figure 4). An alteration in the normal function of *ABCA4* leads to the accumulation of N-RPE in the outer segments of the photoreceptor cells, which in turn causes the formation of the lipofuscin component, N-retinylidene-N-retinylethanolamine (A2E). As a result, high levels of lipofuscin accumulate in the RPE, causing the

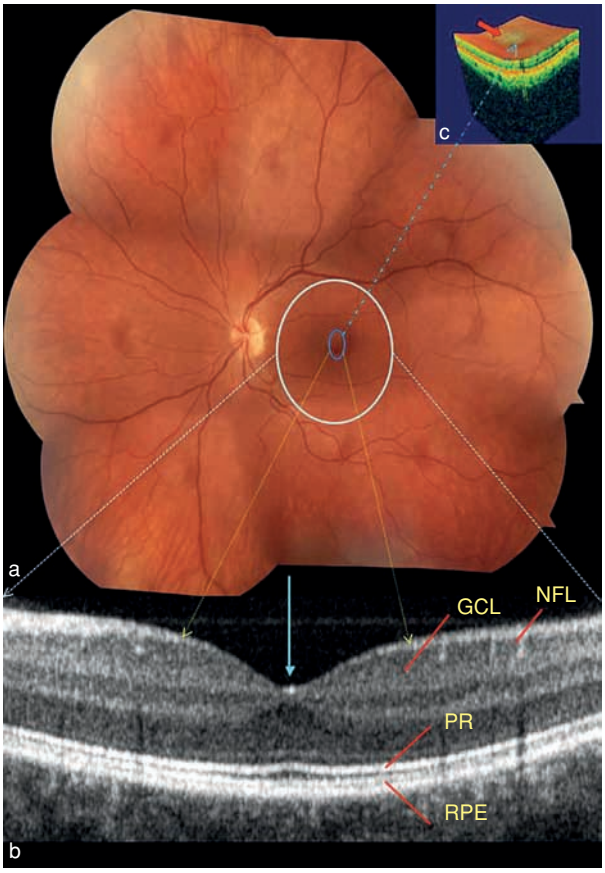

Figure 1 The target area of macular degenerations is the macula: (a) composite color fundus photograph of the eye; (b) retinal cross-sectional profile for one B-scan of the macular region, positioning the fovea and foveola (corresponds to the foveal minimum, light blue arrow); (c), spatially encoded frequency domain OCT image with 3-μm axial resolution (arrows point to the foveal depression). The macula is defined clinically as a central retinal area 5.5 mm in diameter within the boundaries of the superior and inferior temporal vascular arcades (white circle in (a)). The macula has the highest density of photoreceptor cells, both rods and cones. The fovea is a central depression within the macula approximately 1.5 mm in diameter, highly enriched in cones (blue smaller circle in (a)). The foveola is a rodless region within the fovea. Attached to the photoreceptors is a monolayer of pigmented epithelial cells, the retinal pigment epithelium (RPE), which is vital to the integrity and health of the visual cells. Bruch's membrane separates the RPE from the choroid (formed by a network of blood vessels, nerves, immune cells, and fibroblasts), which supplies the RPE and the photoreceptors with all their nutrients. Macular thickness can be determined using optical coherence tomography (OCT), a noninvasive technology that provides *in vivo* high-resolution measurements of the retina, nerve fiber layer, and optic disk. OCT is a clinical tool that is very useful in the diagnosis and monitoring of eye diseases such as macular degenerations. In OCT, a single point of light reflected off the retina forms an A-scan that contains information about the axial location of tissue interfaces within the eye. Data from these single points of light can be laterally aligned to form B-scan images (b). Currently, a spatially encoded frequency domain OCT device can capture 256 B-scans in rapid succession, resulting in a dense three-dimensional cube of OCT information (c). GCL, ganglion cell layer; NFL, nerve fiber layer; PR, photoreceptor inner and outer segments.

subsequent photoreceptor degeneration seen in Stargardt disease.

Model of Stargardt Disease

Based on the discovery of the role of *ABCA4* in Stargardt disease, an *abca4* knockout mouse was generated with the purpose of expanding our knowledge of this disease as well as identifying possible targets for therapeutic intervention. *Abca4* −/− mice lacking rim protein develop delayed dark adaptation and increased all-trans-retinaldehyde following light exposure. In addition, there is increased accumulation of both lipofuscin and A2E, the toxic lipofuscin fluorophore, within the RPE cells. Although in humans the compromise of the RPE due to A2E accumulation causes secondary degeneration of the photoreceptors, this has not been observed in the *abca4* −/− mouse. To date, this is the only genetic model of recessively inherited Stargardt disease.

Sorsby Fundus Dystrophy

Sorsby fundus dystrophy (SFD) causes visual loss in affected individuals beginning around the fourth decade of life. Transmission is autosomal dominant and has been linked to mutations in the tissue inhibitor of the metalloproteinases-3 (*TIMP3*) gene, which encodes a TIMP3 protein that inhibits the action of matrix metalloproteinases. Affected patients typically present with problems transitioning between light and dark, followed by central vision abnormalities and late loss of peripheral vision. Clinically, drusen (tiny yellow or white accumulations of extracellular material that build up in Bruch's membrane) are seen along with choroidal neovascularization (new blood vessels originating from the choroid) that grows into the subretinal space through a thickened Bruch's membrane (Figure 5). The presence of a few small (hard) drusen is normal with advancing age, and most people over 40 have some hard drusen. However, the presence of larger and more numerous drusen in the macula is a common early sign of age-related macular degeneration (AMD). In view of their location between the RPE and its vascular supply, the choroid, it is possible that drusen deprive the RPE and photoreceptors of oxygen and nutrients. As a result, SFD has garnered much interest because of its clinical similarities to the much more prevalent AMD.

Model of Sorsby Fundus Dystrophy

Currently, only one model of SFD has been reported, a knock-in mouse carrying the Ser156Cys mutation in the *TIMP3* gene, generated using site-directed mutagenesis and homologous recombination in embryonic stem cells. After 8 months of life, these mice showed abnormal morphology in both Bruch's membrane and

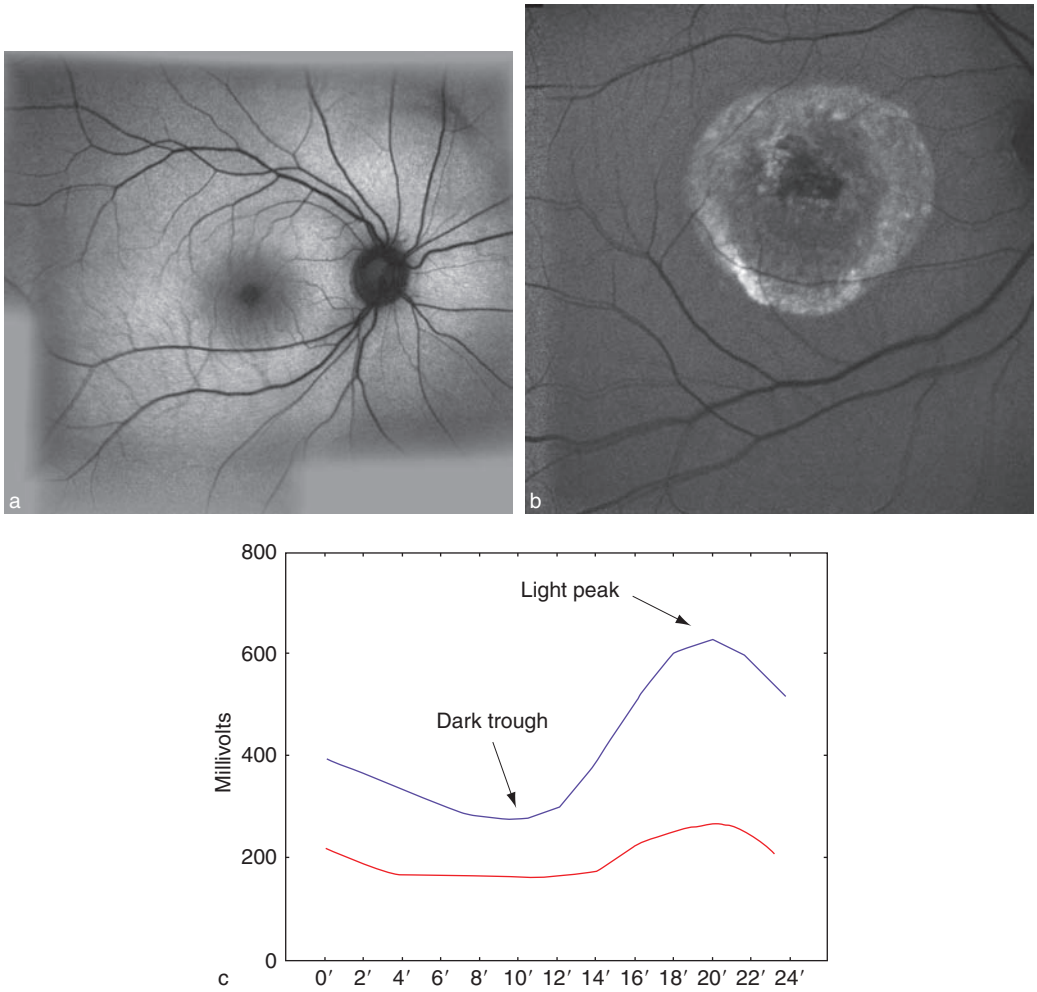

Figure 2 Best disease: (a) autofluorescence imaging of a normal right fundus; (b) autofluorescence imaging of the right fundus of a patient with Best disease with resolving egg-yolk lesion; (c) electro-oculogram (EOG) recordings from a normal individual (upper curve) and a patient affected with Best disease (lower curve). The control subject has an Arden ratio of 2.24, and the Best disease patient has an Arden ratio of 1.39.

the RPE, but no neovascularization. Scotopic and photopic ERGs were normal throughout the life span of affected *Timp3*$^{S156C/S156C}$ mice.

Stargardt-Like Macular Dystrophy

Stargardt-like macular degeneration is an autosomal-dominant variation of Stargardt disease. Like classic Stargardt disease, ERG findings can vary, but in general they are fairly normal. In contrast, the appearance of a dark choroid, commonly seen on fluorescein angiography in recessive Stargardt disease patients, is absent in Stargardt-like macular dystrophy. Stargardt-like macular degeneration is much less common than recessive Stargardt disease, and it was not until 2001 that *ELOVL4* was identified as the gene responsible for this disorder. *ELOVL4* encodes an enzyme that is highly expressed in photoreceptor cells and is involved in the elongation of very long-chain fatty acids, hence its name. In particular, it is hypothesized that the ELOVL4 protein is involved in photoreceptor function and membrane composition because of its role in the synthesis of the polyunsaturated fatty acids of the outer segment. Genetic analyses of various pedigrees of Stargardt-like macular dystrophy have demonstrated that a 5-bp deletion or two 1-bp deletions result in the formation of a similarly truncated ELOVL4 protein that causes macular disease.

Model of Stargardt-Like Macular Dystrophy

The identification of *ELOVL4* as the gene responsible for Stargardt-like macular dystrophy led to the development of a transgenic mouse that expresses the mutant form of human ELOVL4 resulting from a 5-bp deletion. Similar to the *abca4* knockout mouse,

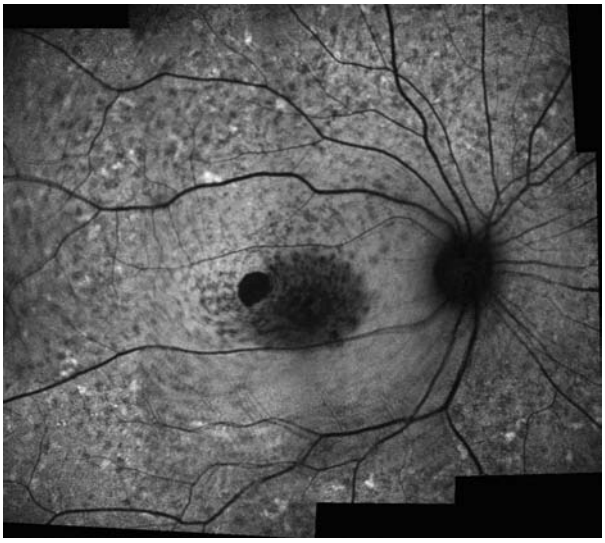

Figure 3 Stargardt disease. Autofluorescence imaging of right eye with hyperfluorescent flecks and hypofluorescent, diseased retinal pigment epithelium (RPE).

elovl4 transgenic mice show the accumulation of phagosomes (membrane vesicles found in RPE cells and formed by the inward folding of the cells' membranes to hold phagocytized photoreceptor outer-segment fragments) and lipofuscin in the RPE. However, photoreceptor degeneration and abnormal ERG changes are seen subsequently in these mice. As a result, the *elovl4* transgenic mouse model exhibits features of both Stargardt-like macular dystrophy and the atrophic form of AMD.

Age-Related Macular Degeneration

AMD is the leading cause of visual impairment in the elderly population of the United States and other developed countries. As its name suggests, this progressive disease is localized to the macular region of the retina. Its end result is the complete loss of central vision. AMD is broadly categorized into two types: atrophic (also known as dry or nonexudative) and exudative (also known as wet). The atrophic form is characterized by the deposition of drusen between the RPE and Bruch's membrane (Figure 6). The accumulation of A2E and its isomers and subsequent RPE death leads to secondary degeneration of the photoreceptors. Eventually, the retina thins and visual impairment occurs. Exudative or wet AMD (Figure 7) is characterized by choroidal neovascularization (CNV). Rapid visual impairment occurs when newly formed vessels leak into the subretinal space and damage the retina. The exudative form of AMD is responsible for most cases of severe visual loss, even though it is much less common than the dry form. Dry AMD does not usually progress to wet AMD. There is some debate as to whether both forms are related or represent separate disease processes; however, polymorphisms in complement factor H (CFH) have been linked to both types, suggesting an association between the two forms.

Although association studies have clearly implicated the role of genetics in the development of AMD, this disease is believed to have a multifactorial etiology with both genetic and environmental components (e.g., smoking and diet) contributing to it. One of the main hurdles in studying AMD has been the lack of precise animal models that would allow the elucidation of the genetic components involved.

Rodent Models of Age-Related Macular Degeneration

Rodents, specifically mice, are currently the primary species used to study AMD. Although rodents lack a macula, they offer several advantages for studying retinal disease. They are cheap and easy to handle, and, as a result, they can be used in large numbers for experiments. In addition, their relatively short life span allows for timely and efficient study of diseases of the elderly, such as AMD. A wide variety of hereditary retinal degenerations occur in mice, and many of the genes causing these diseases possess a corresponding human equivalent. Furthermore, genetic manipulation in mice is fairly simple, reproducible, and well established in the medical literature.

As already mentioned, the *abca4* $^{-/-}$ and the *elovl4* transgenic mice are useful models of Stargardt disease and Stargardt-like macular dystrophy, respectively. The pathologic features seen in these mice as well as studies showing the association of the human *ABCA4* and *ELOVL4* genes with AMD also make them useful for the investigation of AMD, particularly nonexudative AMD. In addition, partial loss of *abca4* in heterozygous +/− *abca4* mice constitutes another mouse model for AMD because these animals have a phenotype similar to those of Stargardt disease and AMD. For this reason, it has been suggested that the Stargardt disease carrier state may be a predisposing factor to the development of AMD. Furthermore, some researchers also consider the *Timp3* transgenic mouse generated to study SFD useful for AMD due to the overlap of clinical features in both diseases.

Hypercholesterolemia and diet have been implicated in the pathogenesis of AMD. As a result, it is not surprising that several models of AMD have been developed which involve impaired cholesterol and/or lipid metabolism. One model that used the concept of impaired lipoprotein metabolism is a knockout mouse of the very low-density lipoprotein

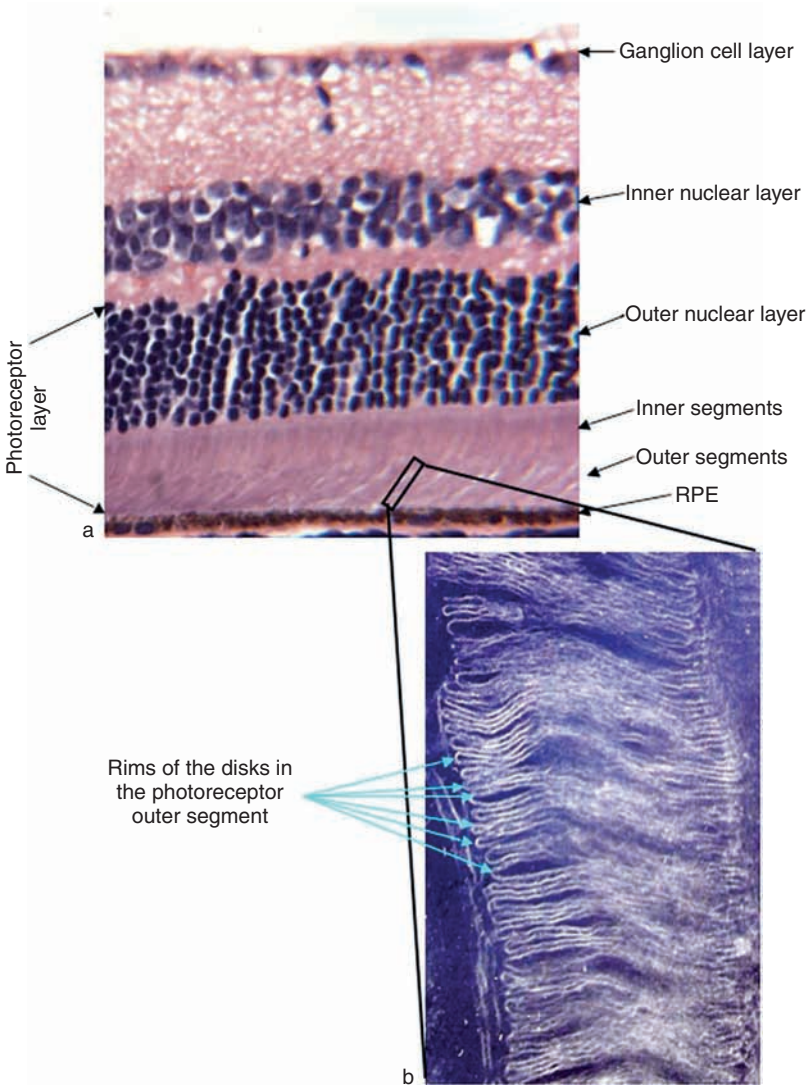

Figure 4 Micrographs of the retina and a photoreceptor cell: (a) light micrograph of a retina section, showing the retinal layers; (b) electronmicrograph of the outer segment of a photoreceptor cell, showing the rims of the membranous disks where the rim protein is localized. As shown in (a), the outer nuclear layer is formed by the nuclei of the photoreceptor cells, which also have inner and outer segments. RPE, retinal pigment epithelium.

(VLDL) receptor gene. A homozygous mutation in the VLDL receptor gene (*Vldlr*^{tm1Her}) causes retinal angiomatous proliferation, a form of occult CNV. In a similar model generated by deleting the low-density lipoprotein (LDL) receptor from the mouse genome, lipid particles accumulated in Bruch's membrane and increased with fat intake. In addition, vascular endothelial growth factor (VEGF) expression in the outer retinal layers of affected mice had a direct correlation with the amount of lipid particle deposition on Bruch's membrane. However, evidence of neovascularization has not been reported in this mouse model. Hyperlipidemia also developed in transgenic C57BL/6 mice that overexpressed the APO B100 gene, the major apolipoprotein associated with LDL cholesterol. When these mice were fed a high-fat diet and exposed to blue light, sub-RPE deposits developed. Deposits did not form in the absence of blue light exposure, and they were prevented by pretreatment with the antioxidant α-tocopherol, suggesting that oxidant injury is responsible for their presence. As confirmed by the examples described next, most experimental mouse models that use the impairment of cholesterol or lipid metabolism show the formation of basal laminar deposits or drusen.

The *ApoE* gene, which encodes apolipoprotein E, has been identified as a risk factor for AMD. As a result, transgenic mice carrying the human APO*E3-Leiden gene (which expresses a dysfunctional form of Apo-E3 discovered in a family from

Figure 5 Sorsby pseudoinflammatory dystrophy angiogram demonstrating classic choroidal neovacularization in the right eye.

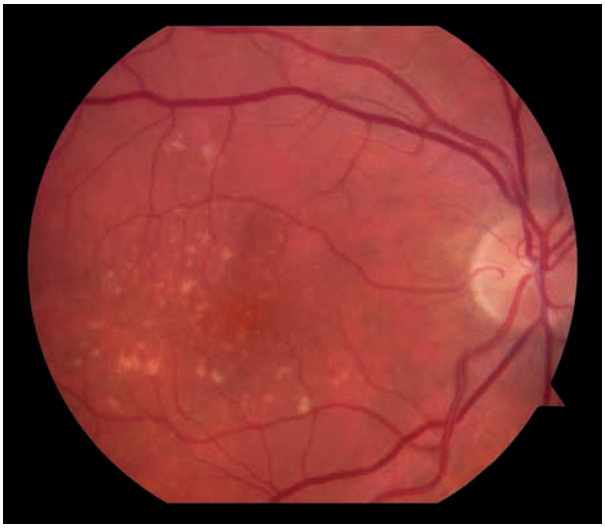

Figure 6 Nonexudative age-related macular degeneration (AMD). Color fundus photograph of the right eye showing drusen and nonexudative AMD.

Leiden) were developed and found to have basal laminar deposits that were more severe in mice consuming a high-fat and high-cholesterol diet. In another animal model for AMD, the *apoE* knockout mouse, electron-lucent deposits on Bruch's membrane observed at the ultrastructural level were similar to basal laminar deposits seen in humans with AMD. Very interesting results were obtained with *apoE* knock-in mice expressing *apoE2, apoE3,* or *apoE4*. These animals were used in a study that combined the presence of the *apoE* genotype with the known AMD risk factors of advanced age and a high-fat, cholesterol-rich diet. Mice deficient in *apoE4* were the most severely affected and developed many pathological changes characteristic of AMD. These included the thickening

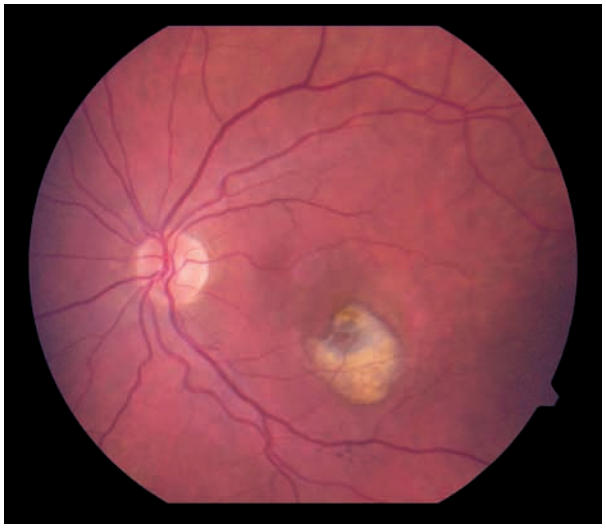

Figure 7 Exudative age-related macular degeneration (AMD). Color fundus photograph of a left eye with exudative AMD, showing a disciform scar and choroidal neovascularization.

of Bruch's membrane, atrophy and changes in the pigmentation of the RPE, drusenoid deposits, and, in some cases, CNV, a characteristic of wet AMD.

Based on the finding that iron levels are elevated in the RPE of AMD patients, a mouse deficient in both ceruloplasmin (Cp; a ferroxidase localized in mouse RPE that facilitates iron export from cells) and hephaestin (Heph; another ferroxidase of the RPE with 50% identity to Cp) was generated. RPE iron levels of *Cp*−/−*Heph-/Y* mice, as well as the levels of ferritin (an iron-storage protein) increased in an age-dependent manner, causing RPE hypertrophy and death, photoreceptor degeneration, and subretinal neovascularization. Thus, combined *Cp* and *Heph* deficiency leads to retinal degenerative changes similar to many of those that are features of AMD. However, the lack of clinically observable drusen is one shortcoming of this model.

Increased amyloid beta (Aβ) is typically associated with neurodegenerative diseases such as Alzheimer's. Because drusen from patients with AMD has been shown to contain Aβ, a neprilysin gene-disrupted model in mice was used to examine its role in the pathogenesis of AMD. Neprilysin is a zinc metallopeptidase that has a rate-limiting role in Aβ catabolism. Lack of this enzyme in senescent neprilysin −/− mice increased Aβ deposition, leading to RPE degeneration and basal laminar and linear deposits beneath the RPE. In addition, it decreased the expression of pigment epithelium-derived growth factor (PEDF) and CRALBP, a marker for the differentiation of RPE cells, but despite causing an upregulation of VEGF, there was no observable CNV. Thus, amyloid fibrils may be associated with pathological symptoms in

different age-related diseases, but their specific way of inducing these disorders may vary depending on the tissue. Nevertheless, some of the therapeutic approaches that are being used for Alzheimer's disease may be effective against the development of AMD.

Impairment of macrophage mobilization is another genetic modification that has been used to create models of AMD. Knockout mice deficient in monocyte chemoattractant protein-1 (Ccl2), as well as another line of mice deficient in C-C chemokine receptor-2 (Ccr2), showed phenotypic features similar to those in AMD, including thickening of Bruch's membrane, subretinal deposits, drusen, and lipofuscin accumulation. Immune complexes accumulated in the retinas of these mice, as in humans with AMD, and it is thought that abnormalities in macrophage trafficking contributed to the formation of drusen. In addition, *Ccl2*$^{-/-}$ and *Ccr2*$^{-/-}$ mice exhibited photoreceptor atrophy, outer nuclear layer cell loss, and increased amounts of A2E.

The RPE is rich in lysosomal enzymes that degrade the photoreceptor outer-segment fragments ingested daily as part of the renewal process of the visual cells. Cathepsin D (CatD) is one of these enzymes. Because the accumulation of debris resulting from the abnormal breakdown products of photoreceptor outer segments could produce changes in the RPE similar to those observed in AMD, a homozygous transgenic mouse (*mcd/mcd*) carrying a two-amino-acid deletion in CatD, was developed. Progeny of this mouse showed age-dependent RPE proliferation and atrophy, photoreceptor degeneration, shortened outer segments, appearance of lipofuscin and basal laminar and linear deposits, and a significant reduction in the ERG a- and b-wave amplitudes. Thus, this is another mouse model that presents several features of nonexudative AMD.

Choroidal neovascularization, one of the characteristics of wet AMD, has been associated with the expression of vascular endothelial growth factor (VEGF). VEGF is known to stimulate the proliferation of endothelial cells in blood vessels, and it is one of the major growth factors participating in angiogenesis (the physiological process involving the growth of new blood vessels from preexisting vessels). The choroidal vessels grow through breaks in Bruch's membrane and proliferate under the RPE and retina, leaking serum and blood. One of the first models of choroidal neovascularization was a transgenic mouse generated by introducing a murine *Vegf* cDNA coupled to the RPE-specific RPE65 promoter into the mouse genome. This mouse overexpressed Vegf in the RPE and choroid. However, although choroidal neovascularization and blood vessel leakage were observed, these vessels did not penetrate Bruch's membrane into the subretinal space.

Another mouse model of exudative AMD was developed using endocrine-gland-derived VEGF (EG-VEGF, also called prokineticin 1 (hPK1)) coupled to the rhodopsin promoter. This led to the expression of hPK1 in the retina of the resulting transgenic mouse. Because hPK1 is a mitogen of fenestrated endothelium, its expression caused the enlargement of the fenestrated choroidal vascular bed without affecting the nonfenestrated retinal vasculature. In addition, the hPK1 transgenic mice showed the accumulation of the major lipofuscin fluorophore, A2E, but not enough to cause visible RPE damage or the invasion of a CNV-like region into the subretinal space.

The most recent mouse model of AMD involves the impairment of superoxide radical dismutation in the retina, a tissue that is particularly susceptible to oxidative stress because of its high consumption of oxygen, high concentration of polyunsaturated fatty acids, and light exposure. Cu, Zn-superoxide dismutase 1 (SOD1) is one of the main antioxidant enzymes in the retina. Senescent SOD1 −/− mice developed thickening of Bruch's membrane, drusen, and choroidal neovascularization, all of which worsened over time. RPE cells showed signs of oxidative damage and degeneration on histology, and the localization of β-catenin and N-cadherin, adherence junction proteins that are involved in making up the barrier integrity of the RPE, was altered. Furthermore, photoreceptor cell loss was seen in a subset of affected mice, consistent with histopathological findings in human eyes with AMD. Overall, all the pathological features of the Sod −/− mouse model tend to be progressive with age, much like human AMD.

To date, only one rat model of long-lasting CNV has been developed by injecting an adeno-associated viral construct containing human VEGF into the subretinal space of rats. These rats developed extensive subretinal neovascularization, photoreceptor degeneration, and leaking blood vessels on fluorescein angiography, and they showed decreased ERG a- and b-wave amplitudes.

Nonrodent Models of Age-Related Macular Degeneration

One of the primary challenges in the study of macular diseases such as AMD, is the anatomic differences between human eyes and those of animals. In particular, the absence of a macula in nonprimate species makes them less than ideal for studying a geographical-specific disease such as AMD. However, the high costs of nonhuman primates and the difficulty in handling them compared to smaller species are some of the drawbacks in the use of primates as AMD models.

The only naturally occurring, nonrodent animals that were found to have abnormalities consistent with

human aging and AMD were reported to be in a colony of aged adult rhesus monkeys at the Caribbean Primate Research Center (CPRC). These monkeys had drusenlike deposits found in both the inner and outer collagenous zones of Bruch's membrane, and they have dense bodies found in both Bruch's membrane and the RPE cytoplasm. A follow-up study of the same monkey colony showed that none of the sampled animals showed signs of exudative AMD or disciform scarring. Thus, these animals are suitable only for the study of dry AMD. Because this is not a disease of simple Mendelian genetics, no gene was identified that could account for the monkeys' AMD phenotype. It should be noted that the relative long life span of monkeys requires that they be tracked over a lengthy period of time before they develop signs of macular disease. Although this monkey colony presents a unique opportunity to study AMD in a naturally occurring disease model, the drawbacks of using primates for research (already mentioned) limit the practicality of their use.

Conclusion

In conclusion, we can state that studies using animal models, in combination with progress in the field of molecular genetics, have improved our understanding of the pathways leading to macular degenerations. As future research uncovers more details about specific genetic defects and generates increasingly accurate animal models, we hope that novel therapeutic strategies will be developed to prevent, treat, and even cure these blinding and debilitating diseases.

See also: Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: Neurodegeneration; Alzheimer's Disease: An Overview; Retinal Development: An Overview; Sensory Aging: Vision; Vascular Issues in Neurodegeneration and Injury.

Further Reading

Allikmets R, Singh N, Sun H, et al. (1997) A photoreceptor cell-specific ATP-binding transporter gene (ABCR) is mutated in recessive Stargardt macular dystrophy. *Nature Genetics* 15: 236–246.

Bernstein PS, Tammur J, Singh N, et al. (2001) Diverse macular dystrophy phenotype caused by a novel complex mutation in the ELOVL4 gene. *Investigative Ophthalmology and Visual Science* 42: 3331–3336.

Chen LJ, Liu DT, Tam PO, et al. (2006) Association of complement factor H polymorphisms with exudative age-related macular degeneration. *Molecular Vision* 12: 1536–1542.

Hageman GS, Anderson DH, Johnson LV, et al. (2005) A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proceedings of the National Academy of Sciences of the United States of America* 102(20): 7227–7232.

Karan G, Lillo C, Yang Z, et al. (2005) Lipofuscin accumulation, abnormal electrophysiology, and photoreceptor degeneration in mutant ELOVL4 transgenic mice: A model for macular degeneration. *Proceedings of the National Academy of Sciences of the United States of America* 102: 4164–4169.

Malek G, Johnson LV, Mace BE, et al. (2005) Apolipoprotein E allele-dependent pathogenesis: A model for age-related retinal degeneration. *Proceedings of the National Academy of Sciences of the United States of America* 102: 11900–11905.

Marmorstein AD, Stanton JB, Yocom J, et al. (2004) A model of Best vitelliform macular dystrophy in rats. *Investigative Ophthalmology and Visual Science* 45: 3733–3739.

O'Gorman S, Flaherty WA, Fishman GA, and Berson EL (1988) Histopathologic findings in Best's vitelliform macular dystrophy. *Archives of Ophthalmology* 106: 1261–1268.

Petrukhin K, Koisti MJ, Bakall B, et al. (1998) Identification of the gene responsible for Best macular dystrophy. *Nature Genetics* 19: 241–247.

Polkinghorne PJ, Capon MR, Berninger T, et al. (1989) Sorsby's fundus dystrophy. A clinical study. *Ophthalmology* 96: 1763–1768.

Sorsby A, Mason MEJ, and Gardner N (1949) A fundus dystrophy with unusual features. *British Journal of Ophthalmology* 33: 67–97.

Sun H, Tsunenari T, Yau KW, and Nathans J (2002) The vitelliform macular dystrophy protein defines a new family of chloride channels. *Proceedings of the National Academy of Sciences of the United States of America* 99: 4008–4013.

Weber BH, Lin B, White K, et al. (2002) A mouse model for Sorsby fundus dystrophy. *Investigative Ophthalmology and Visual Science* 43: 2732–2740.

Weng J, Mata NL, Azarian SM, et al. (1999) Insights into the function of Rim protein in photoreceptors and etiology of Stargardt's disease from the phenotype in ABCR knockout mice. *Cell* 98: 13–23.

Zhang K, Kniazeva M, Han M, et al. (2001) A 5-bp deletion in ELOVL4 is associated with two related forms of autosomal dominant macular dystrophy. *Nature Genetics* 27: 89–93.

Inhibitory Control over Action and Memory

M C Anderson and C Weaver, University of Oregon, Eugene, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

Functions of Inhibitory Control

The ability to suppress strong habitual (prepotent) responses supports at least two broad computational functions necessary for the effective control of behavior: selection and stopping. Selection refers to the act of isolating one representation among competitors vying for the control of behavior or thought. Selection is necessary at all stages of cognitive activity. In memory, selection must occur to retrieve a target memory trace in the face of many competing memories activated by the cues guiding retrieval. In action, a stimulus may initiate multiple compatible physical actions, only one of which should guide behavior. Stopping refers to the need to either preempt or override a mental or physical activity that is inappropriate in a certain context or that, in the course of execution, has become undesirable. For instance, reflexively reaching to catch a falling object might be abruptly stopped when it is realized that the falling object is a hot pan or a potted cactus. The general models of both functions are depicted in Figures 1(a) and 1(b).

In both selection and stopping, the ability to limit the influence of unwanted representations is mediated by a top-down control process acting directly or indirectly on this entity, rendering it temporarily less accessible. In this article, we refer to the brain structure originating the top-down control signal as the source of the inhibitory control and the area representing the to-be-controlled representation or process as the site of inhibition. Broadly, research indicates regions in the lateral prefrontal cortex (either the ventrolateral prefrontal cortex (VLPFC), Brodmann's area 44/45, or the dorsolateral prefrontal cortex (DLPFC), Brodmann's area 9/46) as likely candidates for the source of control, whereas differing cortical and subcortical targets constitute control sites, depending on the content to be controlled.

Paradigms Used to Study Inhibitory Control

Behavioral paradigms used in research with human subjects have taken two approaches to measuring inhibitory control. Many studies measure subjects' ability to implement control by quantifying the accuracy or speed with which selection or stopping is achieved. This approach provides a straightforward measure of behavioral control. Other studies, however, focus on the aftereffects of selection or stopping on the stopped representations. If a response is suppressed, there should be measurable consequences on the ability to subsequently access the representation, reflecting persisting inhibition. For instance, if a motor action is inhibited, people should be slower to emit that same response later, or if a memory has been inhibited, it should become more difficult to retrieve. Methods that measure the inhibitory aftereffects of response override have the advantage of not presupposing inhibition as the mechanism. For example, although inhibition may be engaged to stop a motor action, stopping may instead be achieved by facilitating an antagonistic response that precludes the execution of the to-be-stopped action. Measures based on the first approach cannot conclusively confirm that inhibition is involved in stopping the rejected representation. Examples of both methods are discussed here.

Inhibitory Control of Action

Inhibitory control of actions has been studied in the contexts of selection and stopping, and it is typically measured with functionally oriented tasks. The use of these paradigms, accompanied by various imaging and electrophysiological techniques has established that inhibitory control engages the lateral prefrontal cortex (LPFC). Further, inhibitory control depends on the integrity of this region, as revealed by studies of patients with LPFC damage, studies using transcranial magnetic stimulation (TMS), and lesions in nonhuman primates.

Inhibitory Control in Selective Responding

A variety of paradigms have focused on the role of inhibitory control in selecting one response in the face of interference from prepotent competitors. One approach used to investigate response selection across a variety of stimulus modalities and response types is to manipulate the compatibility of perceived stimuli and their associated responses. Tasks measuring stimulus–response (S–R) compatibility effects have two conditions: (1) trials in which S–R mappings (i.e., associations) are compatible and exploit innate biases, and (2) trials characterized by incompatible S–R mappings which require that the tendency to make the natural response be overcome and the task-appropriate response selected. For example, subjects might see an arrow in the center of a screen pointing left or right and be required to indicate

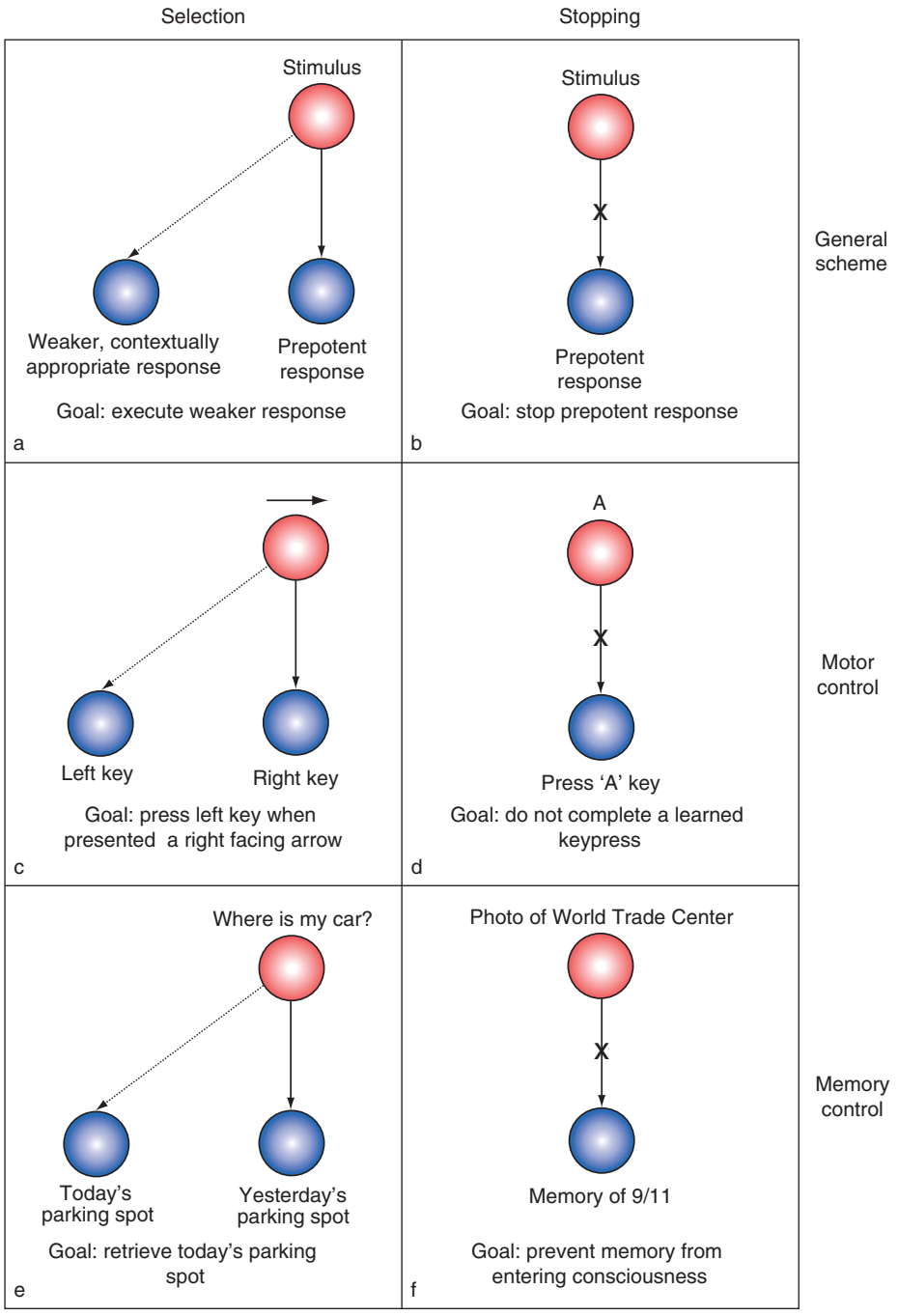

Figure 1 Two situations that require response override in human action and thought and commonly used paradigms: (a) selection; (b) stopping; (c) S–R compatibility paradigm; (d) stop-signal paradigm; (e) selection to inhibit the interference of a memory; (f) stopping the retrieval of a memory. The top row represents a schematization of these two situations. In each instance, a stimulus is associated with one or more responses, such that when the stimulus appears, the responses become active in proportion to their associative connection to the cue (represented by thickness of the line). In selection (a), the weaker response must be made, despite the existence of a strong competitor that becomes more active and threatens to capture control of behavior. In stopping (b), there is only one response, but it must be prevented. As shown in (c), sometimes the prepotent action is not the correct response, and a nondominant movement must instead be made (e.g., S–R compatibility and antisaccade tasks). As shown in (d), if no alternative response is warranted, the movement can simply be canceled (e.g., go/no-go, stop-signal, and countermanding saccade tasks). As shown in (e), sometimes inhibition must be initiated to selectively retrieve a memory with a weaker association to a cue that is shared by another trace (e.g., retrieving today's parking spot and suffering interference from the memory of yesterday's; RIF). In other circumstances (f), confronting a cue may activate an unwanted memory, leading the person to stop retrieval. For instance, when the sight of the World Trade Center initiates retrieval of the memory of the events of 9/11, retrieval might be stopped. This process can be assayed by TNT tasks. RIF, retrieval-induced forgetting; S–R, stimulus–response; TNT, think/no-think.

which direction the arrow is pointing by pressing one of two keys. If the instructions require subjects to press a key with their left hand in response to left-pointing arrows and a key with their right hand in response to right-pointing arrows, the S–R mappings are compatible. If, however, subjects must press a key with their right hand in response to a left-pointing arrow (doing the converse for the right-pointing arrow), the S–R mappings are incompatible. In a nonspatial variant, subjects are presented two consecutive color patches and asked to judge and report whether they are the same or different. In the incompatible mapping condition, subjects are instructed to say 'same' when the colors actually differ. The response selection requirement of such tasks is illustrated in Figure 1(c). The reduced accuracy and increased reaction time (RT) of stimulus-incompatible responses are the measurable costs of resolving the interference created by the prepotent compatible response.

Human imaging work suggests that the LPFC contributes to the resolution of this interference. Response selection appears to be subserved by both the DLPFC and VLPFC, together with a network of regions often observed in attentional control tasks, including the superior parietal cortex, anterior cingulate cortex (ACC), and lateral premotor cortex. A number of imaging studies have sought to differentiate the contributions of these areas, and converging evidence from S–R compatibility and other response selection designs is consistent with the view that the ACC contributes to the detection of response conflict, signaling the need for greater top-down control to the LPFC. In contrast, the LPFC – specifically, BA46 within the DLPFC – is thought to be instrumental in achieving response selection. Consistent with this, lesions and TMS of the human DLPFC compromise response selection. Intriguingly, cellular and neuroimaging studies in nonhuman primates emphasize the VLPFC over the DLPFC as the principal source of response selection. There is also conflicting evidence for the generality of response selection. Considerable behavioral evidence from dual-task interference indicates that response selection constitutes a central processing bottleneck spanning varied stimulus and response modalities. These findings are corroborated by neuroimaging findings that reveal overlapping neural activation across many stimulus and response demands. However, response selection appears in some situations to be lateralized according to stimulus characteristics, with spatial tasks associated with increased activation in the right hemisphere and nonspatial stimuli causing left lateralized activations. Thus, it is unclear whether response selection is a unitary process across all input and output modalities, although most research points to the importance of the LPC in achieving this function.

Inhibitory control over motor action has also been studied in the context of oculomotor response selection. In human subjects, saccade termination similarly appears dependent on the DLPFC. In oculomotor inhibition paradigms, subjects intentionally try to override a reflexive foveal orientation to a stimulus, and the number of failed suppressions is measured. For example, in antisaccade tasks, subjects are instructed to look away from visual targets when they abruptly appear instead of toward them. This requires the inhibition of the visually triggered saccade that ordinarily orients subjects' eyes to the onset stimulus. Although the DLPFC is not necessary to execute accurate saccades, numerous lesion studies show that damage to this area, and to BA46 in particular, markedly increases antisaccade errors, suggesting its importance in their cancellation. Similar patterns of errors have been reported in nonhuman primates following lesions of the DLPFC homolog (dorsal bank of the principal sulcus). Further, substantial neurophysiological work done in nonhuman primates has identified neurons within the DLPFC and frontal eye field (FEF) that show preferential activity during antisaccade trials. In human neuroimaging studies, increased activation for antisaccade trials has been found bilaterally in the DLPFC. Based on these findings, the DLPFC is believed to originate a top-down signal that acts on the FEF and possibly the superior colliculus (SC), which serve as the sites of the inhibition.

Taken together, research from these paradigms and others, including task-set switching, Stroop, and flanker tasks, consistently implicate the LPFC in implementing the top-down signal by which prepotent responses are suppressed to achieve selective motor response. Additional work is required to ascertain the relative involvement of the DLPFC and VLPFC, their respective roles, and the extent to which response selection is general or lateralized by content.

Inhibitory Control in Motor Stopping

Intended actions can be stopped, and some cognitive process must support this function, perhaps inhibition. To study stopping, many paradigms ask subjects to occasionally prevent a response they have been trained to complete, and then they measure the effectiveness or efficiency with which this is done. In the go/no-go (GNG) task, participants are presented with a stimulus to which they must make a motor response. On a minority of trials, a special no-go stimulus is presented, for which subjects must prevent their response. For example, people might be

presented with a series of letters on a screen. For every letter, they might be asked to press the same key as quickly as possible, unless they see the letter X, for which they must withhold their response. The percentage of no-go trials successfully withheld is a metric of inhibitory control ability. Another approach used to study motor stopping is the stop-signal paradigm developed by Gordon Logan, illustrated in Figure 1(d). In this paradigm, subjects typically perform a simple perceptual discrimination (e.g., deciding whether a letter is X or O), to which they must make a motor response (e.g., pressing the key corresponding to X or O) as quickly as possible. On a minority of the trials, a stop signal (typically a tone, but visual and tactile signals are sometimes used) occurs at some point after the stimulus, indicating that subjects should withhold their motor response (commonly a key press, although hand, arm, and eye movements; verbalization; and hand squeezes have been used). Over trials, the delay between the appearance of the stimulus and the stop signal is modulated by a dynamic tracking algorithm that adjusts the stop-signal delay (making it longer or shorter, depending on subjects' accuracy) until stopping performance is at chance. Using the average go trial RT and the average stop-signal delay latency required for an equal probability of successful and failed inhibition, the stop-signal RT (SSRT) can be computed. The SSRT, which is approximately 200 ms for manual responses, quantifies the speed of the stopping process itself and is commonly used as a measure of inhibitory control ability. Stop signal and GNG paradigms have been used to study inhibitory control across the life span, in neuropsychological patients, and as a way to study the neural basis of inhibitory control in nonhuman subjects such as monkeys and rats.

Considerable evidence from motor-stopping paradigms indicates that response inhibition is mediated by the VLPFC and DLPFC, with a right lateralization tendency, at least in humans. For example, work by Aron and colleagues showed that, in patients with damage to the right VLPFC or DLPFC (and not the left), SSRT is significantly slowed and the degree of slowing is correlated with the lesion volume. Work with a variety of other techniques yielded similar conclusions. For example, TMS over the right VLPFC disrupts motor stopping in human subjects, and lesions to a VLPFC homolog (the inferior frontal convexity) in nonhuman primates disrupts performance on GNG tasks. Electrophysiological studies in nonhuman primates have demonstrated a no-go potential originating in the dorsal bank of the principal sulcus (homologous to the DLPFC in humans) and in the rostroventral corner of the PFC (see Figure 2). The no-go potential is an electrical signature of the behavioral no-go response. Strikingly, when stimulation is applied during go trials in which the monkey should respond, the response is cancelled or delayed, concomitant with reduced electrical activity in the motor cortex. The strength of the suppressive effect depends on the stimulation timing, with the maximal suppressive effect observed when the no-go potential would have peaked (75–125 ms), and occurs exclusively in the regions of the PFC in which the no-go potential is generated. As might be predicted from the foregoing findings, response inhibition activates the DLPFC and VLPFC in neuroimaging studies, very often right laterally, although sometimes bilaterally, and frequently accompanied by a network of areas implicated in attentional control. Further, direct comparisons of the neural correlates of three inhibitory tasks (the GNG and two response selection tasks, a S-R compatibility task and a flanker task) within-subjects revealed the DLPFC to be a region activated by these distinct cognitive control tests. Taken as a whole, these findings indicate that both the DLPFC and VLPFC contribute to a top-down signal for inhibiting prepotent motor responses, although some investigators have emphasized the primacy of the right VLPFC.

Although the lateral PFC appears to be the source of a top-down signal that suppresses prepotent motor responses, the site of action, and the circuit through which that site is influenced depends on the nature of the response. In the case of manual motor responses, stopping activates the expected control areas of the VLPFC and dorsomedial PFC, and the regions on which that control is exerted within the motor system. Subcortically, the control of movement within the basal ganglia is conceptualized using the brake–accelerator model. The accelerator is the direct pathway in which the cortex acts on the striatum and internal pallidum, increasing the activation of the motor cortex by the thalamus. Braking is accomplished by the same structures with the critical addition of the subthalamic nucleus (STN), which indirectly inhibits the thalamus and hence the motor cortex. Recent work suggests that a faster inhibitory mechanism, the cortico-subthalamo-pallidal hyperdirect pathway, may be contributing to response stopping. In this conception, the VLPFC interprets the stop signal and (circumventing the striatum) increases the activation of the STN, and thus the internal pallidum, resulting in the inhibition of the thalamus and, consequently, the motor cortex. Consistent with this possibility, a recent stop-signal neuroimaging study conducted by Aron and Poldrack revealed that the amount of activation in the STN (and not the striatum) predicted SSRT. Thus, in the case of manual motor response inhibition, the VLPFC may control action by exciting structures that inhibit motor cortex.

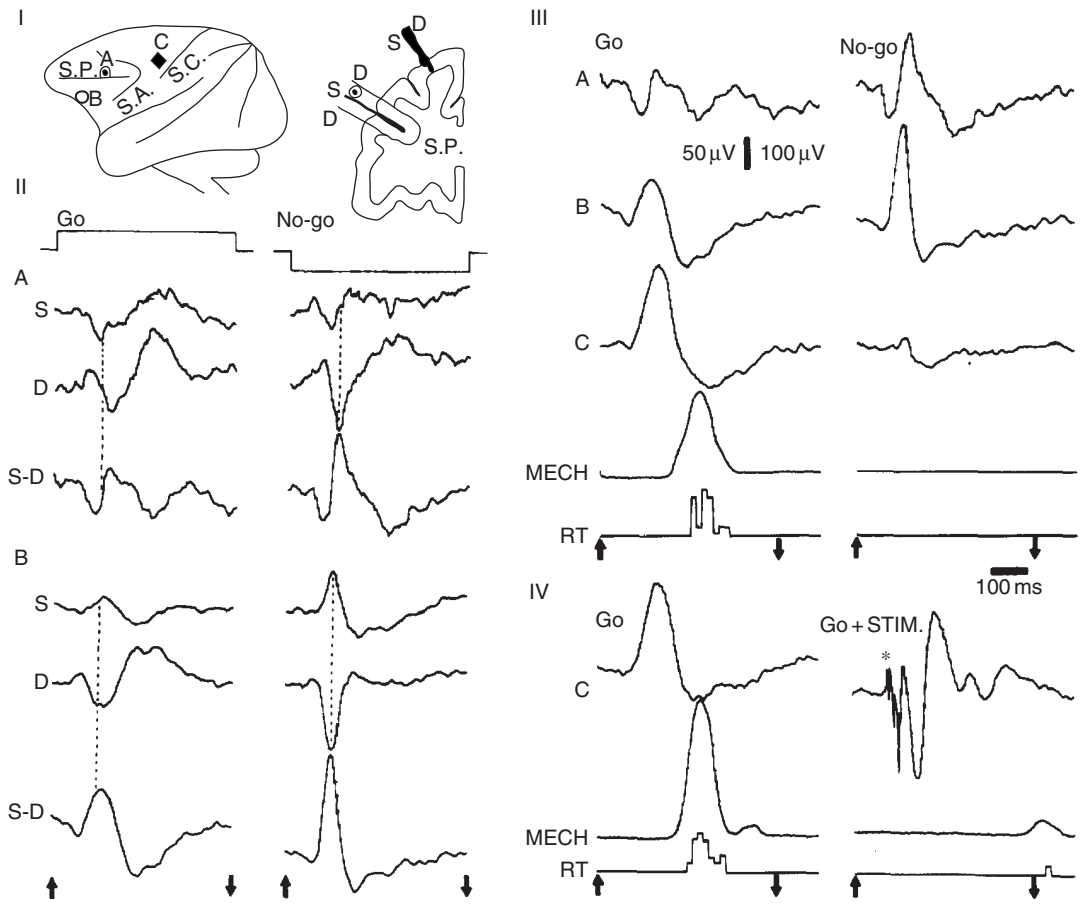

Figure 2 Stimulation of the DLPFC and VLPFC homologs suppresses motor responses in nonhuman primates. Monkeys were rewarded for completing manual responses to a light stimulus and learned to respond accurately. Responses to another light stimulus were unrewarded, and the monkeys learned to withhold their responses. No-go activity coincides with no-go behavior and is characterized by surface-negative, depth-positive field potentials generated 110–150 ms after presentation of the visual stimulus. Emergence of no-go potential is contingent on learned discrimination of light stimuli. Moreover, stimulation of specific lateral PFC loci induces no-go behavior. **(I)** Lateral (left hemisphere) and coronal (right hemisphere) views of the frontal lobe illustrate electrode localization. Surface (S) and depth (D) electrodes were placed in the monkey homologs of the DLPFC (dorsal bank of the principal sulcus; A, open circle with central dot), VLPFC (rostroventral corner of the prefrontal cortex; B, open circle), and forelimb area of the motor cortex contralateral to responding hand (C, filled diamond). (II) Field potentials consistent with no-go potential at S and D electrodes and S-D at prefrontal loci A and B during go (left column) and no-go (right column) trials for a single monkey. (III) S-D field potential data for each of the three recording regions across both conditions; A, B, and C in go (left column) and no-go trials (right column). Specificity of no-go potentials to regions and A and B, illustrated by lack of potential in area C. In addition to field potentials, movement (as measured by mechanogram (MECH)), and reaction times (RT) are also reported for each condition. (V) Presentation of go stimulus produces typical field potentials, movement, and reaction times, whereas the go stimulus coupled with bilateral stimulation of areas A and B prevented response, as is typical of no- to trials (compare to III). Arrows indicate onset and offset of visual stimulus. DLPFC, dorsolateral prefrontal cortex; PFC, prefrontal cortex; S.A., sulcus arcuatus; S.C., sulcus centralis; S.P., sulcus principalis; VLPFC, ventrolateral prefrontal cortex. Reproduced from Sasaki K, Gemba H, and Tsujimoto T (1989) Suppression of visually initiated hand movement by stimulation of the prefrontal cortex in the monkey. *Brain Research* 495(1): 100–107, with permission from Elsevier.

Suppressing unwanted eye movements, like suppressing manual responses, engages the LPFC, and in particular the DLPFC, as noted in the discussion of antisaccade research. The site acted on by LPFC, however, is different than in the case of manual responses. For instance, research on the cellular dynamics of motor inhibition with nonhuman primates points to the importance of the FEF. An adaptation of the stop-signal paradigm, called the countermanding saccade task, is often used. The monkeys learn a saccade RT task and are subsequently trained to stop their saccadic eye movements whenever an infrequent stop signal is administered. Thus, the performance of this task requires two processes: (1) initiation of the intended saccades and (2) revocation of the to-be-cancelled saccade. Both of these processes are believed to be supported by the FEF, although movement cancellation is also supported by the SC and the supplementary eye fields. In particular, within the FEF and SC, movement neurons drive go processes,

whereas fixation neurons are implicated in stop processes. Whether a saccade is executed or not depends on the relative activity of these types of neurons. Increasing activation and the rate of that increase in movement neurons are predictive of commitment to saccade initiation and the timing of the initiation, respectively. In contrast, fixation neurons demonstrate activity that is reciprocal to that of movement neurons; their activity is gradually reduced before saccade initiation and increases with saccade suppression. Thus, movement and fixation neurons implement saccade initiation and termination in the FEF. Together with the work on antisaccade performance in human subjects, this suggests that inhibitory control of eye movements is implemented by an interaction between the DLPFC and regions of the FEF involved in initiating and canceling eye movement responses.

Thus, whereas the source of inhibitory control in LPFC may well be shared across different motor-stopping tasks, the site of action depends on the effector system being controlled. Next we discuss evidence indicating that this principle is not limited to the control of physical actions but also applies to the control of internal cognitive acts, such as long-term memory retrieval.

Inhibitory Control in Long-Term Memory

Inhibitory control over memory retrieval provides a model system for the role of inhibitory control in cognition; it has also been studied in selection and stopping. Human behavioral studies have revealed that overcoming the influence of prepotent traces in long-term memory retrieval yields an inhibitory aftereffect such that suppressed traces are subsequently less accessible. Neurobiological research indicates that retrieval suppression, like motor stopping, engages the LPFC, which modulates the activation in the hippocampus and other medial temporal regions associated with memory and emotion. The similarities between the systems engaged by memory and action control suggest that the same or perhaps isomorphic computational solutions may exist for solving response-override problems for physical and cognitive actions.

Inhibitory Control in Selective Retrieval

One procedure that has been designed to study inhibitory control is known as the retrieval–practice paradigm, developed by Anderson and colleagues. In a typical version of this procedure, subjects encode six exemplars from each of eight taxonomic categories (e.g., fruits-orange, fruits-banana, and drinks-gin) in a study phase. Following the initial study, subjects perform directed retrieval practice on a subset of exemplars from some of the studied categories (e.g., three of the six fruits) via category-plus-stem cued recall tests (e.g., fruit or___). Subjects are then tested on all exemplars from all categories. On the final test, subjects can either be cued with the category name and asked to recall all studied exemplars or be given a recognition memory test on all items. Of interest are the effects of practicing some category exemplars on the ability to recall the remaining unpracticed exemplars of those categories. Repeatedly retrieving some exemplars of the category impairs the subjects' ability to recall the remaining unpracticed exemplars, relative to their performance on baseline categories that were also studied, but none of whose exemplars were retrieved in the interim. This effect, known as retrieval-induced forgetting (RIF) is thought to reflect the consequences of an inhibitory process that suppresses unpracticed competitors during retrieval practice, causing enduring impairment of those items on subsequent tests. RIF has been induced with verbal and visuospatial stimuli, and across levels of complexity ranging from simple word pairs to complex factual knowledge and even eyewitness events. RIF affects episodic, semantic, and implicit memory as measured by conceptual fluency and lexical decision. Further, RIF appears on tests of episodic recall and recognition memory. Thus, RIF may be an example of inhibitory control aiding in the function of selection during the retrieval process by suppressing interfering items during retrieval of a target. The selection requirements of RIF are illustrated in Figure 1(e).

RIF exhibits functional properties that firmly establish the involvement of an inhibitory control process that overcomes interference. First, it is retrieval-specific; repeated study of some associates of a cue does not impair the recall of related items, even when repeated retrieval of those same items does. Further, RIF is interference-dependent, with larger RIF effects observed for competitors that cause interference during retrieval, consistent with the view that inhibitory control is engaged to suppress interfering competitors. Finally, RIF is cue-independent; the impairment of a competing memory item generalizes to novel cues used to test the inhibited items, indicating a generalized reduction in the accessibility of the suppressed trace. Thus, in the domain of memory control, theoretically focused behavioral research based on inhibitory aftereffects has firmly established the involvement of inhibition.

A large body of work establishes the role of the VLPFC in selective retrieval of verbal semantic knowledge under conditions of interference. Moreover, patients with LPFC damage show heightened susceptibility to proactive interference in long-term episodic memory retrieval, as revealed in work by Shimamura

and colleagues. Nevertheless, only recently has work begun to tie the LPFC to inhibitory aftereffects measured in RIF. Consistent with executive control literature, the ACC appears central to the detection of conflict between competing memories associated with a cue, signaling the need for top-down control to resolve competition; in contrast, the VLPFC appears to initiate control processes that resolve that conflict through inhibition. These principles can be seen in a study by Kuhl and colleagues, who hypothesized that because RIF is a consequence of resolving interference from competing items during target retrieval, inhibited items should cause less interference on later retrievals of the target items. Indeed, more activation was observed in the bilateral VLPFC and right DLPFC during early than during late retrieval practice trials, and the more activation decreased over trials in the right VLPFC and ACC, the more memory inhibition was found. Moreover, subjects showing higher inhibition exhibited elevated activation in the ACC and hippocampus during early trials that gradually diminished with repeated retrievals; low-inhibition subjects did not show this pattern. Thus, subjects who experience the greatest initial competition between the medial-temporal representations of the target and competitors (as indexed by medial temporal lobe (MTL) and ACC co-activation) show the greatest engagement of control and, consequently, the greatest RIF.

Similar conclusions have been reached using other techniques, such as electrophysiology. For example, more positive going event-related potentials (ERPs) are observed over frontal electrode sites during selective retrieval of target items than in a passive practice condition with the same targets. The magnitude of this retrieval-specific ERP effect on the PFC predicts the amount of RIF observed on the final test. These findings indicate that some functional requirement uniquely associated with retrieval, such as the need to resolve competition, engages the PFC and induces RIF, as before. Mounting evidence points to the involvement of the LPFC in the induction of RIF and the importance of this process in achieving selective retrieval, echoing the role of inhibitory processes in selective motor responding.

Inhibitory Control in Retrieval Stopping

Another method used to study inhibitory control of memory is known as the think/no-think (TNT) procedure, developed by Anderson and colleagues. Modeled after the GNG task, TNT measures the inhibitory aftereffects of stopping memory retrieval. Subjects learn a collection of cue-target pairs, so that whenever they are presented with the cue, they can recollect its associated target item. Following encoding, subjects participate in the TNT phase, during which a single cue item from one of the earlier pairs is presented on each trial. For some of these cues, the subjects are directed to think of the associated target memory and keep it in mind for as long as the cue is on the screen. For other cues, subjects are directed to not think about the associated target memory, keeping it out of mind for as long as the cue is on the screen. The latter no-think task requires that the subject override the retrieval process and prevent the associated declarative memory from entering awareness, despite the tendency for the cue to elicit that memory (see **Figure 1(f)** for an example of memorial stopping). Subjects are then tested on their memory for all cue–target pairs. In the typical design, one-third of the original pairs participate in the think trials in the TNT phase, one-third in the no-think trials, and one-third are reserved as baseline items against which to measure the effects of retrieval and retrieval suppression on later memory performance.

Repeatedly presenting participants with reminders of an event and asking them to stop retrieval of that event has several notable effects. First, it eliminates the normally beneficial effects of reminders in enhancing a memory's later accessibility, so that reminders can be presented over a dozen times with little apparent benefit in accessibility of the associated traces. Second, stopping retrieval very often impairs the ability to recall the associated trace, relative to other baseline cue–reminder pairs that were trained at the same time but that did not have their reminders presented in the interim during no-think or think trials. Thus, repeated exposure to reminders coupled with an effort to stop declarative memory retrieval impairs access to the associated trace. Third, the impairment of the excluded trace occurs even when that trace is tested with a novel cue, indicating a generalized impairment of the trace, consistent with the idea that the trace has been inhibited. Most of these effects have been observed with both verbal cue–target pairs and visual pairs such as face–scene pairs, and the effects appear to generalize to target items with emotional content. Thus, stopping unwanted retrievals appears to be effected in part by suppressing the associated memory, consistent with inhibitory control. As such, the TNT paradigm provides an example of the role of inhibitory control that is analogous to the GNG and stop-signal paradigms.

Neuroimaging studies have established several key findings about the neural systems involved in stopping long-term memory retrieval. First, as can be seen in **Figure 3**, suppressing retrieval of an episodic memory in response to a reminder engages the LPFC, including the VLPFC and DLPFC, often bilaterally,

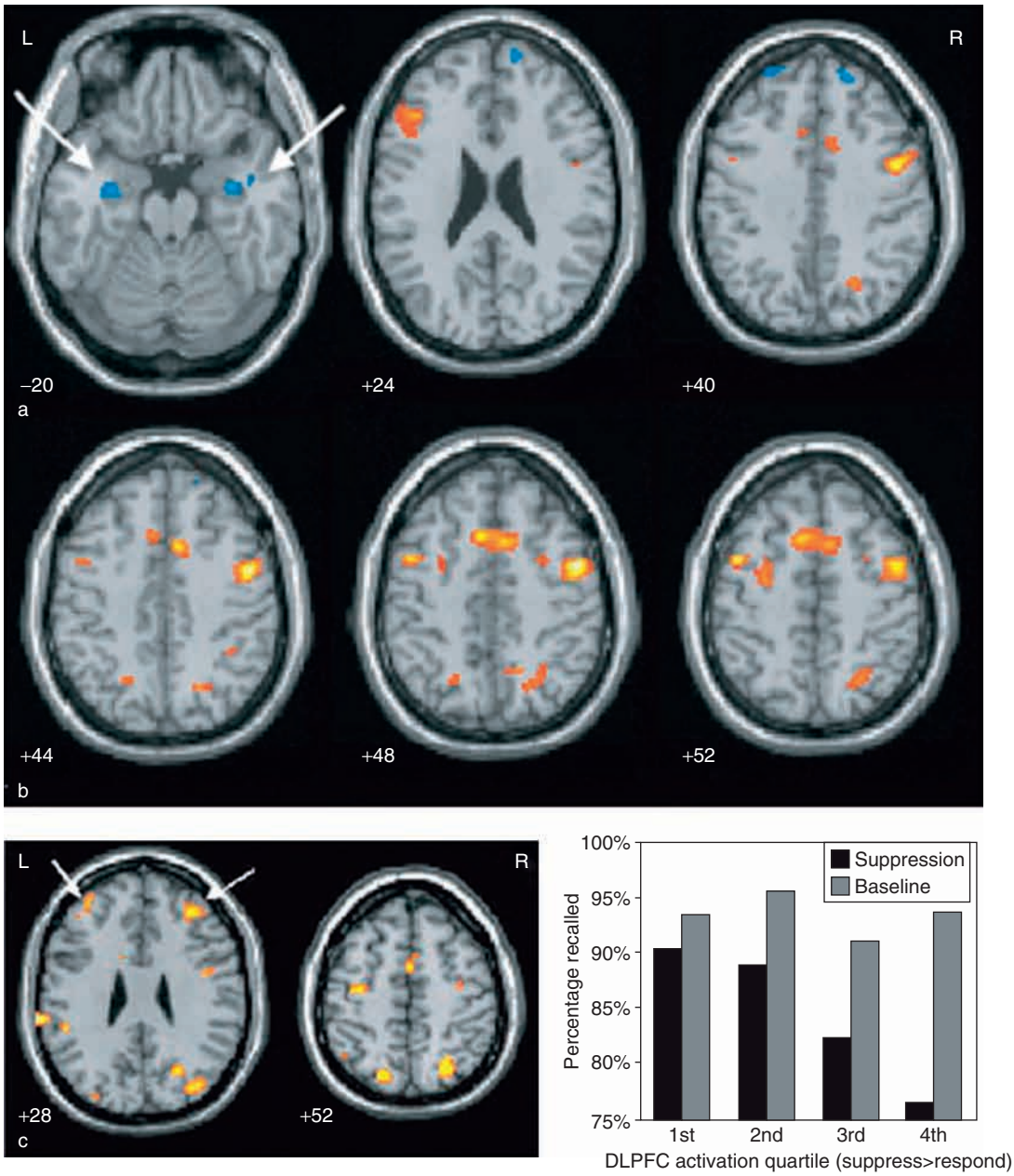

Figure 3 fMRI results during no-think trials and think trials: (a) brain regions that significantly differed in activation between the no-think trials and think trials during the TNT phase ($n = 24$); (b) successful recruitment of the DLPFC, which predicts individual differences in memory inhibition (the inhibitory aftereffect of stopping retrieval); (c) inhibitory aftereffects for four subject groups, differing in DLPFC activation. In (a), areas in yellow were more active during no-think trials than during think trials, whereas areas in blue were less active during no-think trials ($P < 0.001$). The white arrows highlight the reduced hippocampal activation in the no-think condition. In (b), plotted in warm colors are the regions that predict the magnitude of the inhibitory aftereffect observed on the final memory tests (the white arrows indicate the DLPFC; VLPFC was also observed in this study but not in a later study involving photographs). In (c), the greater the DLPFC activity shown by a subject group, the more impaired was the later recall of no-think items on the final test, compared to the recall of baseline items that were learned at the same time but that did not appear during the TNT phase. DLPFC, dorsolateral prefrontal cortex; fMRI, functional magnetic resonance imaging; TNT, think/no-think; VLPFC, ventrolateral prefrontal cortex. From Anderson MC, Ochsner KN, Kuhl B, et al. (2004) Neural systems underlying the suppression of unwanted memories. *Science* 303(5655): 232–235.

although sometimes with right lateralization. VLPFC engagement parallels the activations observed for RIF and for motor-response suppression, although the DLPFC appears to be more consistently engaged in retrieval stopping. Second, individual differences in the engagement of the DLPFC (and sometimes the VLPFC) during suppression predict memory inhibition, with inhibition increasing with increasing activation, as might be expected if the DLPFC originated a top-down signal responsible for memory

impairment. Third, suppressing retrieval reduces activation in the hippocampus, compared to engaging retrieval, suggesting that inhibitory control modulates functional activity in structures involved in declarative memory. Recent work indicates that activation is reduced below the levels observed during passive fixation, with the extent of this reduction predicting memory failure. Finally, retrieval suppression engages the full network of regions associated with endogenous attention, including the lateral premotor cortex, ACC, and intraparietal sulcus, indicating common brain systems may support controlled attention during perception and memory tasks. With emotional memories, suppression also significantly reduces amygdalar activation, probably reflecting control over emotional aspects of the negative memories. Taken together, these findings indicate that functional interactions between the LPFC and MTLs support inhibitory control over memory retrieval.

As already discussed, imaging studies indicate that suppressing retrieval of an episodic memory significantly reduces activation in the hippocampus, compared to retrieval and a passive fixation baseline. The hippocampus has widely acknowledged involvement in episodic memory, and many studies indicate that it is also involved in the recollection of recently encoded events. As such, reduced hippocampal activation may reflect the successful control of conscious recollection by inhibitory control processes. Electrophysiological studies using the TNT procedure also support the conclusion that conscious recollection (as measured by an ERP component, the parietal old/new effect) may not merely be reduced, but may be eliminated by retrieval suppression. The extent of the modulation of this component predicts subsequent forgetting of no-think items, consistent with the view that regulating recollection induces later forgetting. Thus, both hemodynamic and electrophysiological measures converge on the conclusion that inhibitory control can be engaged to modulate the neural machinery responsible for generating episodic recollection.

The foregoing research illustrates the strong functional and neurobiological similarities between the control of action and memory. Both engage a common frontal region – the LPFC – to override the influence of unwanted neural activity in support of selection and stopping. What differs between motor and memory control (and, indeed, between different forms of motor control) appears to be the sites of control and, perhaps, the means of achieving that control (see the section titled 'Neural architecture of inhibitory control'). The work on memory control adds the unique and theoretically significant contribution of showing that both selection and stopping induce inhibitory aftereffects on rejected representations, verifying and quantifying the involvement of inhibition in achieving control. To the extent that memory control may be used as a model system for cognitive control, these findings suggest the existence of a domain general system for controlling action and thought via inhibition. Consistent with this possibility, the LPFC is engaged in an extraordinarily diverse collection of cognitive tasks, and Duncan and colleagues have shown that it predicts individual differences in fluid intelligence (gF, or psychometric g). A common theme running throughout these tasks appears to be the management of interference, consistent with the putative role of this region in inhibitory control. Nevertheless, it remains possible that there is not one mechanism of control in the LPFC but many content-specific ones, localized similarly, the differences between which cannot yet be ascertained in the absence of direct comparisons between cognitive and motor control. Even, so, lateral prefrontal-posterior interaction remains a valid organizing principle that characterizes our understanding of the networks underlying inhibitory control.

Neural Architecture of Inhibitory Control

Many investigators agree that inhibitory control supports the regulation of behavior by limiting the influence of undesired representations on ongoing behavior by functionally altering the state of the undesired representation. Although the brain clearly solves the computational problems of stopping and response selection, there is not yet widespread agreement on how they are solved or on whether they are solved in a unitary manner. There are two general views on how inhibition may be achieved: direct inhibition and biased competition. Direct inhibition hypothesizes that an inhibitory process targets the unwanted representation or system (either directly or by exciting inhibitory interneurons), degrading its control over behavior. In contrast, biased competition hypothesizes that attentional control signals facilitate the representations of desired actions and that, as a result, competing representations are inhibited through local reciprocal inhibitory projections. The evidence base for biased competition is well developed, and Miller and Cohen have argued that this computational principle may be a primary means of implementing executive control. It is established, however, that both types of inhibitory control take place in the nervous system. For example, as reviewed by Sotres-Bayon and colleagues, excitatory projections from the medial frontal cortex terminate on inhibitory interneurons in the amygdala that modulate the functioning and output of that structure and, correspondingly, suppress conditioned emotional

responding. Similarly, the frontal cortex excites inhibitory neurons in the reticular nucleus of the thalamus, which inhibit the thalamus, gating sensory input to the cortex. The fronto-subthalamic circuitry underlying motor-response suppression (discussed earlier) may be another such example. Thus, inhibitory control may often be accomplished, not by facilitating a competing response but by inhibitory modulation of the brain system underlying the undesired process. The types of situations in which these mechanisms of control operate are an interesting area for future study.

Concluding Comments

The capacity to override unwanted actions and memories is a fundamental component of the effective control of behavior and thought. The pervasiveness of selection and stopping as computational problems has led to the proposition that individual variation in the efficiency with which people inhibit may underlie differences in intellectual functions of many sorts (e.g., attention, working memory, long-term memory, reading, reasoning, and behavioral control) and may even be an important contributor to some psychological disorders. When a population exhibits perseverative tendencies, difficulties managing interference, or distractibility, investigators often consider the potential contribution of frontal lobe dysfunction (and, consequently, inhibitory dysfunction). These proposals, often referred to as inhibitory deficit theories, have been advanced to explain a range of phenomena. For example, because of the late development and early decline of the PFC in the life span, alterations in inhibitory control efficiency have featured prominently in theories of cognitive development and cognitive aging. Clinically, it has been proposed that inhibitory deficits contribute to psychopathologies, including obsessive–compulsive disorder, depression, schizophrenia, attention deficit/hyperactivity disorder, addiction, and chronic posttraumatic stress disorder. It seems unlikely that these diverse conditions are produced by the same deficit and, indeed, that even the various deficits within each population are produced by a single underlying deficiency. Nevertheless, the overlap between systems engaged during motor and memory inhibition suggests that generalized inhibitory control deficits are possible. If so, elucidating the neural systems that subserve inhibitory control is not merely of theoretical interest but has the potential for broad impact on problems of clinical importance.

See also: Computational Approaches to Motor Control; Cortical Control of Eye Movements; Episodic Memory: Assessment in Animals; Episodic Memory; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Saccadic Eye Movements; Semantic Memory; Strategic Control of Memory.

Further Reading

Anderson MC and Green C (2001) Suppressing unwanted memories by executive control. *Nature* 410(6826): 366–369.

Anderson MC, Ochsner KN, Kuhl B, et al. (2004) Neural systems underlying the suppression of unwanted memories. *Science* 303 (5655): 232–235.

Aron AR (2007) The neural basis of inhibition in cognitive control. *Neuroscientist* 13(3): 214–228.

Aron AR, Robbins TW, and Poldrack RA (2004) Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences* 8(4): 170–177.

Bergstrom ZM, Velmans M, de Fockert J, and Richardson-Klavehn A (2007) ERP evidence for successful voluntary avoidance of conscious recollection. *Brain Research* 1151: 119–133.

Boucher L, Palmeri TJ, Logan GD, and Schall JD (2007) Inhibitory control in mind and brain: an interactive race model of countermanding saccades. *Psychological Review* 114(2): 376–397.

Bunge SA, Ochsner KN, Desmond JE, Glover GH, and Gabrieli JD (2001) Prefrontal regions involved in keeping information in and out of mind. *Brain* 124(part 10): 2074–2086.

Carlson SM and Moses LJ (2001) Individual differences in inhibitory control and children's theory of mind. *Child Development* 72(4): 1032–1053.

Dagenbach D and Carr TH (1994) Inhibitory Processes in Attention Memory and Language. San Diego, CA: Academic Press.

Depue BE, Curran T, and Banich MT (2007) Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *Science* 317: 215–219.

Diamond A (1990) Developmental time course in human infants and infant monkeys, and the neural bases of, inhibitory control in reaching. *Annals of the New York Academy of Sciences* 608: 637–669 [discussion 669–676].

Duncan J and Owen AM (2000) Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neuroscience* 23(10): 475–483.

Garavan H, Ross TJ, Murphy K, Roche RA, and Stein EA (2002) Dissociable executive functions in the dynamic control of behavior: inhibition, error detection, and correction. *Neuroimage* 17(4): 1820–1829.

Goldman-Rakic PS (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In: Mountcastle VB and Plum F (eds.) *Handbook of Physiology, Sec. 1: The Nervous System, Vol. 5: Higher Functions of the Brain, Pt. 1*, pp. 373–417. Baltimore, MD: American Physiological Society.

Gorfein DS and MacLeod CM (eds.) (2007) *Inhibition in Cognition.* Washington, DC: American Psychological Association.

Jiang Y and Kanwisher N (2003) Common neural substrates for response selection across modalities and mapping paradigms. *Journal of Cognitive Neuroscience* 15(8): 1080–1094.

Johansson M, Aslan A, Bauml KH, Gabel A, and Mecklinger A (2007) When remembering causes forgetting: electrophysiological correlates of retrieval-induced forgetting. *Cerebral Cortex* 17(6): 1335–1341.

Jonides J and Nee DE (2006) Brain mechanisms of proactive interference in working memory. *Neuroscience* 139(1): 181–193.

Knight RT, Staines WR, Swick D, and Chao LL (1999) Prefrontal cortex regulates inhibition and excitation in distributed

neural networks. *Acta Psychologica (Amsterdam)* 101(2–3): 159–178.

Levy BJ and Anderson MC (2002) Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences* 6(7): 299–305.

Logan GD (1994) On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In: Dagenbach D and Carr TH (eds.) *Inhibitory Processes in Attention, Memory and Language*, pp. 189–239. San Diego, CA: Academic Press.

Lustig C, Hasher L, and Toney ST (2001) Inhibitory control over the present and the past. *European Journal of Cognitive Psychology* 13(1–2): 107–122.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24: 167–202.

Munoz DP and Everling S (2004) Look away: The anti-saccade task and the voluntary control of eye movement. *Nature Reviews Neuroscience* 5(3): 218–228.

Proctor RW and Vu KL (2006) Stimulus-Response Compatibility Principles: Data, Theory, and Application. Boca Raton, FL: CRC Press.

Rowe JB, Toni I, Josephs O, Frackowiak RS, and Passingham RE (2000) The prefrontal cortex: response selection or maintenance within working memory? *Science* 288(5471): 1656–1660.

Sasaki K, Gemba H, and Tsujimoto T (1989) Suppression of visually initiated hand movement by stimulation of the prefrontal cortex in the monkey. *Brain Research* 495(1): 100–107.

Schumacher EH and Jiang Y (2003) Neural mechanisms for response selection: Representation specific or modality independent? *Journal of Cognitive Neuroscience* 15(8): 1077–1079.

Shimamura AP (2000) The role of the prefrontal cortex in dynamic filtering. *Psychobiology* 28(2): 207–218.

Sotres-Bayon F, Bush DE, and LeDoux JE (2004) Emotional perseveration: An update on prefrontal–amygdala interactions in fear extinction. *Learning & Memory* 11(5): 525–535.

Thompson-Schill SL, D'Esposito M, Aguirre GK, and Farah MJ (1997) Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences of the United States of America* 94(26): 14792–14797.

Insulin-Like Growth Factor Signaling and Actions in Brain

V C Russo and G A Werther, Royal Children's Hospital and University of Melbourne, Parkville, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

The Discovery of the Brain IGF System

In 1957, Salmon and Daughaday reported a serum factor (or factors) that mediated the cartilage sulfation and longitudinal bone growth activity of the somatotrophic hormone (growth hormone, GH). This factor was termed 'sulfation factor' and was produced by hepatic cells after exposure to GH (see Figure 1). These circulating factors, which also showed insulin-like activity not suppressible by anti-insulin antibodies, were later found to have a biochemical structure similar to that of the β chain of insulin. Their structural homology with proinsulin led to their current designation as insulin-like growth factors 1 and 2 (IGF-1 and IGF-2).

In the middle 1980s, a variant of IGF-1, then called des(1–3) or 'truncated' IGF-1, was identified in brain tissue. This truncated IGF-1 variant lacks the first three amino acids and is more potent than intact IGF-1 in various cell culture systems, probably due to its lower affinity for IGF-binding proteins (IGFBPs).

This finding suggested brain synthesis of IGF-1 or its truncated variant. Following the isolation of IGF-1 messenger RNA (mRNA) from postnatal rat brain, a number of *in situ* hybridization studies demonstrated that IGF-1 and IGF-1 receptor (IGF-1R) mRNA is synthesized in the rat brain in specific regions, namely, olfactory bulb (Figure 2), hippocampus, and cerebellum.

In addition, IGF-carrier proteins, later named as IGFBPs, were found to be expressed in similar brain regions. The finding of IGF-1 mRNA co-localization with IGF-1 receptors and the presence of IGFBPs suggested a paracrine or autocrine role for IGF-1, potentially modulated by IGFBP, in the developing brain (Figure 3).

IGF System Expression in the Nervous System

IGF Ligands

IGF-1 plays a key role in the development of the nervous system, with demonstrated effects in many stages of brain development, including cell proliferation, cell differentiation, and cell survival. Although recent reports have demonstrated that postnatal circulating IGF-1 might exert neurogenic and survival activity, systemic IGF-1 is not readily transported through the blood–brain barrier, and therefore local production of IGF-1 is considered the primary source of the ligand (autocrine and paracrine action) for brain cells.

During embryogenesis, IGF-1 mRNA expression is detectable in many brain regions, with its expression being particularly high in neuron-rich regions such as the spinal cord, midbrain, cerebral cortex, hippocampus, and olfactory bulb (Figure 1). IGF-2 mRNA is abundantly expressed in the embryonic rat central nervous system (CNS); however, data are conflicting on IGF-2 expression in cells of neuroepithelial origin. IGF-2 is the most abundantly expressed IGF in the adult CNS, with the highest level of expression found in myelin sheaths, but it is also expressed in leptomeninges, microvasculature, and the choroid plexus, all nonneuronal structures that enable diffusion of growth factors to their sites of activity.

IGFBPs

The mRNA expression profiles and location of the most abundant IGFBPs, 2, 4, and 5, in the normal developing and adult CNS are well defined, suggesting an important role for IGFBPs in the nervous system.

IGFBP-1 is not expressed in the CNS, but its expression is induced under certain experimental conditions. IGFBP-2 is expressed early in embryogenesis and by embryonic day 10 is highly expressed in discrete neuroectoderm structures. Later in development, IGFBP-2 mRNA is highly abundant throughout the brain, particularly in brain regions undergoing continuous remodeling, as is the olfactory bulb (Figure 2), the cerebellum, and the hippocampus. IGFBP-2 expression correlates with and complements that of IGF-2, and both IGFBP-2 and IGF-2 protein are highly abundant in the cerebrospinal fluid and choroid plexus.

IGFBP-2 associates to cell surface proteoglycans in rat brain tissue and neuronal cells, and IGF-1–IGFBP-2–proteoglycan complexes have been identified in rat brain tissue; their role is not yet well understood. Such findings suggest that IGFBP-2 may be involved in targeting IGFs to their cell surface receptors in the brain. The absence of a 'brain phenotype' in the IGFBP-2 –/– mouse suggests functional redundancy in the IGFBP family during development of the CNS.

IGFBP-3 is normally expressed at a low level in the CNS, mainly in nonneuronal structures, including epithelial cells. The effects of IGFBP-3 gene deletion on CNS are either unknown or have not been reported. IGFBP-3 is found to be upregulated in rat

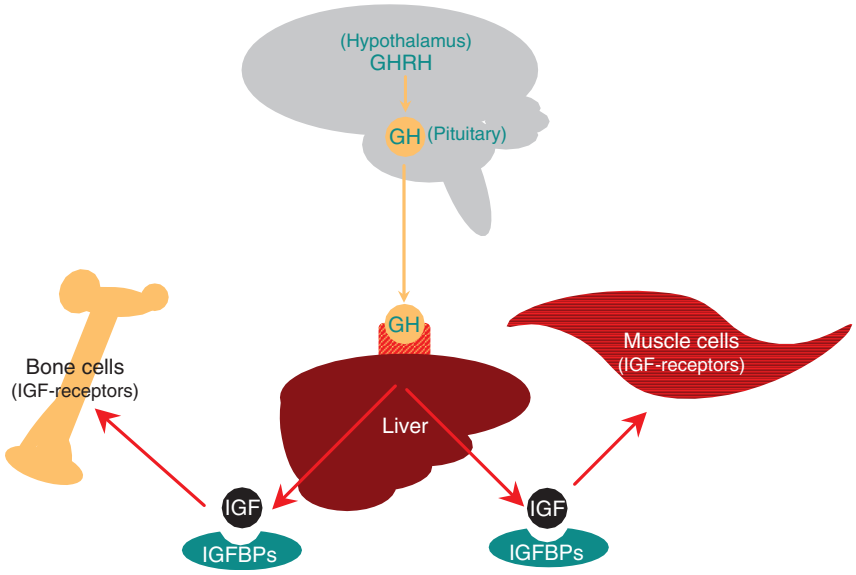

Figure 1 The somatomedin hypothesis. The circulating factor that was originally proposed to mediate the effects of growth hormone on longitudinal bone growth was subsequently found to be insulin-like growth factor 1 (IGF-1). However, the presence of IGF-1 and IGF-1 messenger RNAs in multiple tissues has necessitated the revision of the original 'somatomedin hypothesis' to include both autocrine and paracrine actions of IGF in addition to its classical endocrine aspects. The IGF-binding proteins (IGFBPs) present in serum and expressed in most tissues actively contribute to the endocrine, paracrine, and autocrine action of IGF-1. GH, growth hormone; GHRH, growth hormone-releasing hormone.

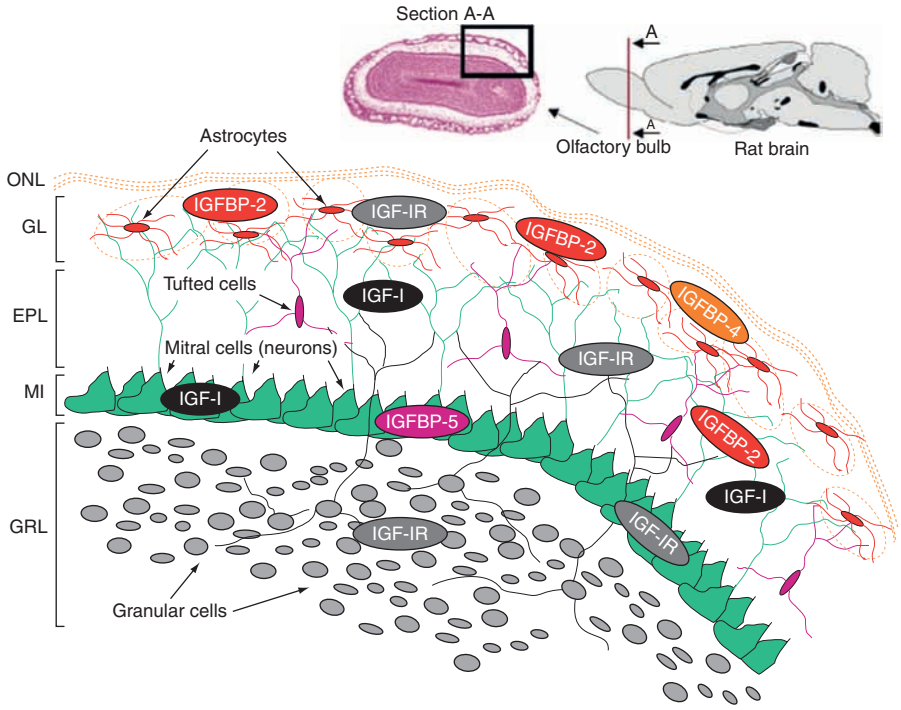

Figure 2 The insulin-like growth factor (IGF) system in the rat brain olfactory bulb. IGF-1 is expressed and synthesized in distinct areas of the rat brain, adjacent to regions rich in IGF-1 receptors (IGF-1Rs) and insulin-like growth factor binding proteins (IGFBPs; e.g., olfactory bulb, as shown). This provides strong evidence for an autocrine and paracrine action of the IGFs in the nervous system. IGF action is modulated by locally expressed IGFBPs. ONL, Olfactory nerve layer; GL, glomerular cell layer; EPL, external plexiform cell layer; MI, mitral cell layer; GRL, granular cell layer. Section A-A shows a hematoxylin–eosin staining of olfactory bulb. Area in the rectangle is enlarged and represented in the main drawing. Reproduced from Russo VC, Gluckman P, Feldman EL, and Werther GA (2005) The insulin-like growth factor system and its pleiotropic functions in brain. *Endocrine Reviews* 26(7): 916–943. Copyright 2005, The Endocrine Society.

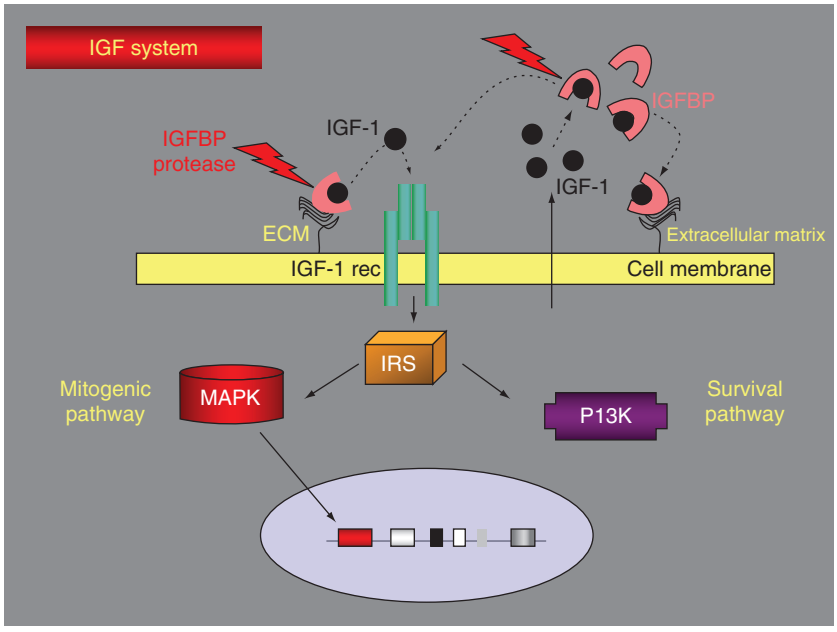

Figure 3 Circulating (endocrine) or locally expressed (i.e., in autocrine or paracrine system) insulin-like growth factor (IGF) complex. IGF-binding proteins (IGFBPs), in addition to regulating transport, diffusion, and half-life of IGFs, provide specific binding sites for the IGFs in the extracellular and pericellular space. This pericellular localization might contribute to the modulation of IGFs' interactions with their receptors and downstream cellular signaling. The effects of the IGFBPs are also regulated by the presence of specific IGFBP proteases, which cleave the binding proteins, generating fragments with reduced or minimal binding affinity for the IGFs. ECM, extracellular matrix; IRS, insulin receptor substrate; rec, receptor.

cerebral cortical cells following GH stimulation. GH promotes proliferation of neural precursors, neurogenesis, and gliogenesis; these responses appear to be mediated in part by locally produced IGF-1 and its modulator IGFBP-3. No change in brain growth or phenotype has been reported in IGFBP-3 transgenic mice.

IGFBP-4 is normally expressed at a very low level in the CNS, where its mRNA is found in a variety of brain cell types, including meningeal cells, astrocytes, and fetal neuronal cells. Whether ablation of the IGFBP-4 gene affects the CNS has not been reported.

IGFBP-5 gene expression is highly abundant during early brain development. In rodents, IGFBP-5 appears to be co-expressed with IGF-1 in principal neurons of sensory relay systems, cerebellar cortex, hippocampal formation, and many other neuron-rich regions, including the olfactory bulb (Figure 2). Furthermore, *in vivo*, IGFBP-5 expression is specifically regulated by IGF-1. In *Xenopus*, IGFBP-5, together with three IGFs expressed in the early embryo, promotes anterior development by increasing growth of the head region. Thus, both IGF signal activation and IGFBP-5 appear to be required for anterior neural induction in *Xenopus*. Whether IGFBP-5 has similar functions in early mammalian neural development is not known. The effects of IGFBP-5 gene ablation in the brain have not been reported.

IGFBP-6 is poorly expressed in the nervous system, and information regarding its mRNA expression distribution, in both the developing and adult nervous system, is limited. IGFBP-6's unique property of preferential binding to the IGF-2 ligand, coupled with the fact that this ligand is the most abundantly expressed IGF in the adult CNS, raises the possibility that the IGFBP-6–IGF-2 complex may have a unique role in modulating IGF-2 function in the adult brain. Preliminary analysis of the IGFBP-6 transgenic mouse shows reduced cerebellum size and weight, combined with altered differentiation of astrocytes. Abnormalities in the hypothalamus and pituitary were also reported. The effects on the brain of IGFBP-6 gene ablation remain unknown.

IGF Receptors

Type 1 IGF receptors are highly expressed in many specific brain regions and cell types, particularly in plastic regions such as the developing cerebellum, midbrain, and olfactory bulb (Figure 2) and in the ventral floor plate of the hindbrain. Given that IGF receptors are expressed from early embryogenesis and throughout life and that their ligands also show a similar 'temporal–spatial' pattern of expression, it is evident that local brain IGF circuits are crucial modulators of the processes activated during brain

development. The level of IGF-1R decreases to adult levels soon after birth but remains relatively high in the choroid plexus, meninges, and vascular sheaths. It is thus not surprising that knockout of the IGF-1R gene produced, in addition to *in utero* growth retardation, is a clear brain phenotype, namely, a small brain.

The type 2 IGF receptor, which has low affinity for IGF-1, but a high affinity for IGF-2, and also binds mannose-6-phosphate, is selectively expressed in all major brain regions. It is therefore possible that the IGF-2–mannose-6-phosphate receptor, in addition to its role in transporting lysosomal enzymes, might also participate in control of postnatal brain development and possibly repair.

Various atypical IGF-1R subtypes possessing distinct structures or binding properties have also been described (hybrid insulin–IGF-1Rs with the ability to bind insulin as well as IGFs). However, the physiological significance of hybrid and atypical IGF receptors is unclear.

Early studies demonstrated that some brain regions (i.e., olfactory bulb and choroid plexus) expressing both the IGF-1 and IGF-2 receptors are also rich in insulin receptors. Insulin receptor density is elevated in the choroid plexus, olfactory bulb, limbic system, and hypothalamus, all regions concerned with olfaction, appetite, and autonomic functions. More recently, ablation of the insulin receptor in nestin-positive neurons has suggested a role for insulin in control of appetite suppression and reproduction.

IGF Signaling and Action in Brain

Binding of IGF-1 (or IGF-2) to the α subunit of the type-IGF-1R (cysteine-rich domain) induces comformational changes of the intracellular β subunit with consequent induction of tyrosine kinase activity. Two major substrates, insulin receptor substrates 1 and 2 (IRS-1 and IRS-2), interact with the IGF-1R and become phosphorylated on multiple residues (Figure 4). Phosphorylated IRS-1 is then able to recognize and bind to the 'Src homology polypeptide domain' (SH domains) present in various signal transduction proteins (Figure 4). Molecules that bind IRS-1 include the growth factor receptor-bound (Grb-2), the adaptor oncogenic protein containing

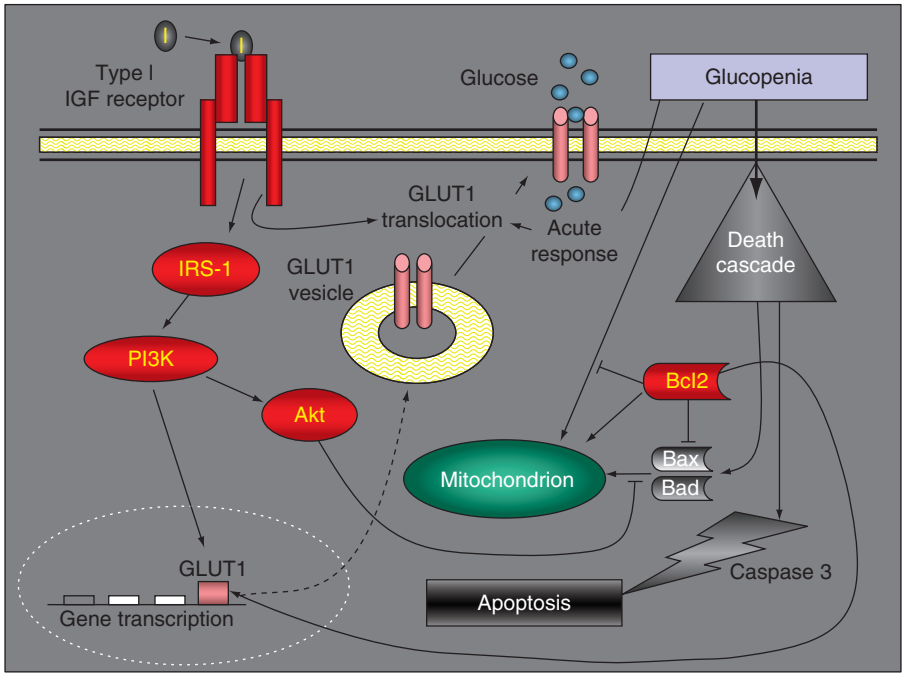

Figure 4 A model of key pathways and interactions involved between the insulin-like growth factor (IGF) and bcl-2 systems in glucose-deprived neuronal cells. Neuronal cell responses to glucose withdrawal (glucopenia) include activation of acute responses such as increased translocation of glucose transporter (GLUT) to cell surface and modulation of mitochondrial activities. These responses are followed by sustained activation of the extrinsic (death receptor cascade) and the intrinsic (mitochondrial cascade) proapototic pathways. These effects are potently regulated by IGF-1 via the phosphoinositide-3 kinase (PI3K) pathways, which lead to the upregulation of GLUT transcription and translocation, recovery of mitochondrial activities, and inhibition of intrinsic and extrinsic proapototic processes. A key element of IGF antiapoptotic action is Bcl-2, which promotes mitochondrial functions and potentiates glucose uptake. Akt, protein kinase B; IRS, insulin receptor substrate.

one SH2 and three SH3 domains, the protein tyrosine phosphatase Syp (SH-PTP2), and the p85 *β-subunit* domain of PI3K (see the section titled 'The mitogenic pathway').

The Mitogenic Pathway

The IGF-1R mitogenic pathway begins with Grb-2, which is a small cytoplasmatic protein and acts as an 'adapter molecule' that links the guanidine nucleotide exchange factor for $p21^{ras}$, termed mSOS (homologous to the *Drosophila* protein son of sevenless (SOS)), to IRS-1. The Grb-2–SOS complex can thus activate Rouss avian sarcoma (Ras), which recruits Raf (serin/threonin protein kinase), and they proceed to activate the mitogen-activated protein kinase (MAPK) pathway (MEK kinase, MEK, extracellular signal-regulated kinase (ERK) kinase, p90) and transcription factors and eventually induce activation of genes such as the early growth response genes *c-jun* and *c-fos*.

Therefore, whether the IGFs are produced locally or reach brain cells systemically, binding to their receptor activates intracellular signaling, promoting neurotrophic, neurogenic, and neuroprotective/antiapoptotic activities. A large number of *in vitro* studies have demonstrated that IGF-1 promotes mitogenesis and differentiation in glia, oligodendrocytes, neuronal cells, adult stem cells, and brain explants and regulates axon myelination via activation of MAPK, also known as ERK. There is also clear evidence that IGF-1 activation of the MAKP/ERK pathways enhances growth cone motility and promotes neurite outgrowth, suggesting that this pathway is likely to be involved in neurogenesis, migration, and repair.

Supportive evidence for the importance of IGF and IGF receptor signaling and activation of downstream effectors has come from a number of studies in IGF null (−/−) mice. These studies have clearly demonstrated that most of the IGF functions determined *in vitro* also apply to the *in vivo* situation, affecting a wide range of brain cells. Thus IGF-1 null mice have reduced brain size and altered brain structures, including alteration of myelination processes. Furthermore, ablation of the *Igf-1* gene has revealed deficits in the numbers of specific neurons and oligodendrocytes in the olfactory bulb, dentate gyrus, and striatum and in cochlear ganglion neurons. In addition to gross structural brain abnormalities, IGF null mice also show alteration of important brain metabolic functions such as reduced glucose uptake, the major source of energy for neuronal cells. These *in vivo* models also show that the IGF system affects several steps involved in development and organization of CNS architecture, including number of cells, connections, and regulation of the extracellular matrix.

It is important to emphasize that the IGFs do not act in isolation or only discretely activate specific pathways. It is well recognized that the actions and effects of IGF signaling are influenced by or affect a large number of other growth factor/cytokine sytems and that IGF-induced IGF-1R phosphorylation activates complex and integrated intracellular signaling.

The Survival and Metabolic Pathway

Binding of the IGFs to the IGF-1R may also activate signaling through the PI3K cascade (Figure 4). Protein kinase B (PKB), also known as serine/threonine kinase Akt, is one of the primary downstream targets of PI3K, and it is well characterized for its involvement in modulation of cell survival. Activated PKB/Akt in turn phosphorylates and inhibits several proapoptotic proteins, such as Bad, caspase-9, and the winged-helix family of transcription factors (FKHRL1, FKHR, and AFX), leading to cell survival in various systems, including hippocampal neurons.

Activation of PKB/Akt appears to be also involved in the modulation of brain metabolic events such as glucose transport and utilization in neuronal cells. As mentioned earlier, IGF-1 is a major regulator of glucose uptake and utilization in brain. In an *in vitro* model of glucose stress (glucopenia), mimicking *in vivo* hypoglycemia, activation of PKB/Akt by IGF-1 was involved in both survival (inhibition of proapoptotic genes) and upregulation of glucose transporters at transcriptional and posttranslational levels (**Figure 4**).

Another target of IGF-1, via the PKB/Akt pathway, is the glycogen synthase kinase 3β (GSK3β). IGF-1, similar to insulin, stimulates the phosphorylation of GSK3β, thus promoting glycogen and protein synthesis. Accumulation of glycogen, normally associated with astrocytes in mature brain, is abundant in IGF-1-positive projecting neurons and postnatally in neurons in IGF-1-rich areas, suggesting that local autocrine and paracrine IGF-1 signaling via the PI3K/Akt pathway is involved in glucose metabolism and storage.

The GSK3β substrate tau, involved in formation of the intracellular neurofibrillary tangles that contribute to the neuronal degeneration seen in Alzheimer's disease, is hyperphosphorylated in IGF-1-null mice. Furthermore, GSK3β promotes apoptosis in neurons via mechanisms involving phosphorylation of tau and β-catenin. Thus IGF-1-induced inhibition of GSK3β, via the PI3K/Akt pathway, appears to be a critical mechanism of IGF-1 action in the brain.

The IGFs Are Neurotrophic Factors

IGF-1 is involved in brain plasticity processes, and it specifically modulates synaptic efficacy by regulating synapse formation, neurotransmitter release, and neuronal excitability. IGF-1 provides constant trophic support to neuronal cells in the brain and in this way maintains appropriate neuronal function. Alteration of this trophic input may contribute to brain disease as seen in neurodegenerative disorders such as Alzheimer's disease, ataxia telangiectasia, Huntington's disease, and Parkinson's disease, all variably responsive to IGF-1 treatment. Recent studies have shown that neurogenesis declines in brains of aged mice but is efficiently restored following IGF-1 administration via intracerebroventricular infusion.

The IGF System and Neuroendocrine Signaling Cross Talk in Brain

In vivo, growth factors such as IGF-1 do not exist in isolation. Hence, the presence of other growth factors may further modulate IGF's biological activity and cellular responses. This modulation can be seen in IGF's neurogenic activity, which may be optimal only in the presence of such cytokines as fibroblast growth factor-2 (FGF-2) and a range of other local growth factors.

In primary cultures of hypothalamic cells, FGF-2 combined with IGF-1 significantly increased the number of neurite-bearing cells above that seen when either one of the two growth factors was added in isolation. Their effects appear to be synergistic since blocking IGF-1 from binding with an antibody abrogates FGF-2's effects. In contrast, a recent study in neuroblastoma cells demonstrated that FGF-2 overrides IGF-1 mitogenic and survival activity via induction of neuronal differentiation and blockade of IGF-1's antiapoptotic response. Interactions among local brain growth factors are thus complex and may vary under differing conditions.

Recent reports have highlighted the importance of a number of serum (in contrast to brain-produced) growth factors, which included IGF-1, FGF-2, and other blood-borne neurotrophic factors (neurotrophins), in neuroprotective surveillance. All these factors exerted a tonic trophic input on brain cells, providing defense mechanisms ranging from blockade of neuronal death to upregulation of neurogenesis.

Increased interest has thus recently been devoted to the interaction of IGF-1 with other 'classic' neurotrophic factors. Of a number of neuroprotective factors including IGF-1, glial cell-derived neurotrophic factor (GDNF), brain-derived neurotrophic factor (BDNF), neurotrophin-4 and -3 (NT-4/5, NT-3), and ciliary neurotrophic factor, only IGF-1, GDNF, and NT-4/5 were found to be potently neuroprotective of motor neurons. These findings point to the potential combined use of these neuroprotective factors in treatment of neurodegenerative diseases. Results from a number of phase III clinical trials with nerve growth factor, BDNF, and IGF-1 have indicated that these neurotrophic factors may find an application in degenerative disorders or injury of peripheral nerves and motor neurons.

IGF-1 and GH

Although GH gene expression also occurs in the central and peripheral nervous systems, with brain GH imunoreactivity not affected by hypophysectomy, the potential cross talk between the IGF system and GH in the nervous system has not been fully investigated. Several experimental models, *in vivo* and *in vitro*, support a key role for GH in brain development. The Snell dwarf mouse (Pit1) and the GH-releasing hormone receptor-deficient little mouse exhibit microcephalic cerebra with hypomyelination and retarded neuronal growth with poor synaptogenesis. Postnatal GH administration normalizes neuronal growth in these mice. Recent *in vivo* studies have demonstrated that GH is involved in neuroprotection during hypoxic or ischemic brain injury. In these studies, GH-like immunoreactivity was demonstrated on injured brain cells, and it was shown that GH administered intracerebroventricularly is capable of preventing brain cell loss. Whether the neuroprotective effects of GH involve IGF-1 induction is not clear. However, some recent *in vitro* studies have demonstrated that GH promotes proliferation and differentiation of fetal cerebral cortical cells in primary culture and that these effects are mediated by IGF-1.

There is some evidence that the GH-related hormone prolactin also possesses neurogenic properties. In female mice during pregnancy, prolactin stimulates production of neuronal progenitors in the forebrain subventricular zone. Whether prolactin's neurogenic properties are a direct effect of its action or are mediated by other factors, including IGF-1, remains to be determined.

IGF-1 and Erythropoietin

Erythropoietin (EPO) is traditionally known as a hematopoietic cytokine produced by the fetal liver and adult kidney in response to hypoxia. However, the expression of EPO and EPO receptors in the CNS and the upregulation of EPO by hypoxic–ischemic insult suggest that this cytokine is an important mediator of the brain's response to injury. In fact, *in vivo* EPO administration protects hippocampal CA1 neurons and retinal neurons from ischemic damage and prevents brain injury

following a number of 'insults' by mechanisms involving both inhibitions of apoptosis and neurotrophic actions. In the brain, EPO expression, regulated by hypoxia-inducible factor-1α (HIF-1α), is mainly found in astrocytes.

IGF-1 protection of primary neuronal cells exposed to low-oxygen concentration correlates with activation of HIF-1α expression. In the same studies using an *in vivo* model of hypoxic or ischemic brain injury, it was also shown that IGF-1 transcriptional activation correlates with that of HIF-1α, suggesting that HIF-1α might mediate some of the IGF-1 responses. Another study has shown that IGF-1 induces HIF-1α transcriptional activity in rat cerebral cortex and neuronal cells (PC12) and that this induction is abolished by a selective IGF-1R antagonist (JB-1). Furthermore, studies in hepatoma cells have shown not only that insulin, IGF-1, and IGF-2 induced HIF-1α expression but that HIF-1α is required for expression of IGF-2, IGFBP-2, and IGFBP-3.

These findings suggest the presence of a potential complex synergistic cross talk between the IGF and the EPO systems, involving both activation of common intracellular signaling pathways and regulation of gene expression. The presence of synergistic cross talk between the IGF and the EPO systems in neuronal cells has recently been demonstrated by showing that EPO can exert a more immediate neuroprotective action when administered in concert with IGF-1. The neuroprotective mechanism, following coadministration of EPO and IGF-1, involved synergistic activation of the PI3K/Akt pathway. These findings further support the concept that the coadministration of synergistic neuroprotective agents, rather than a single agent, might provide improved therapeutic outcome. Thus, treatment with appropriate combinations of EPO and IGF-1 could be a future therapeutic strategy for a variety of acute neurological events.

IGF-1 and Sex Steroids

Among the numerous endocrine systems regulating brain physiology, sex steroids play an important role. However, it appears that some of the effects of these hormones on the brain are mediated by trophic factors, including IGF-1. Estradiol and IGF-1 increase survival and differentiation of developing fetal rat hypothalamic neurons. In this case, estrogen-induced activation of the estrogen receptor requires the presence of IGF-1; both estradiol and IGF-1 use the estrogen receptor to mediate their trophic effects on hypothalamic cells. *In vivo* sex steroids affect IGF-1 levels in the endocrine hypothalamus, arcuate nucleus, and median eminence.

Furthermore, increased clinical and basic evidence suggests that gonadal steroids affect the onset and progression of several neurodegenerative diseases and schizophrenia, as well as recovery from traumatic neurological injury such as stroke. Similarly, in the brain, both estrogen synthesis and estrogen receptor expression are upregulated at sites of injury. The neuroprotective effects of estrogen may be exerted independent of the classical nuclear estrogen receptors, involving modulation of IGF-1 signaling. This is supported by the fact that estrogen receptors and IGF-1R interact in the activation of PI3K and MAPK-signaling cascades and possibly in the promotion of neuroprotection. It is therefore possible that the decrease in estrogen and IGF-1 levels with aging may thus result in an increased risk for neuronal pathological alterations. These findings thus suggest the presence of potential complex synergistic cross talk between the IGFs and sex steroids involving activation of common intracellular signaling pathways; however, the precise mechanism remains unclear.

Future Directions

Since 1957, a considerable body of investigation has been devoted to defining the role of the IGF system in many tissues, including the brain. There is no doubt that these pleiotropic factors, in concert with their receptors and binding proteins, are involved in controlling key processes in brain development and traumatic or degenerative disorders of the nervous system.

Although the development of *in vivo* models such as the transgenic and knockout mice or disease and injury models have contributed to providing unique information about the growth-promoting, neurogenic, and neuroprotective actions of the IGF system, many questions remain. In particular, analysis of the signaling pathways activated by IGFs and their integration and cross talk with other brain growth factor systems is critical. Such insights will potentially lead to the development of improved approaches to prevention, management, and effective treatment of destructive, neurodegenerative, and malignant process of the nervous system.

See also: Diabetic Neuropathy; Gonadal Steroid Actions on Brain; Growth Factors: Neuronal Atrophy; Growth Hormone; Immune System–Neuroendocrine Interactions; MAP Kinase Signaling in Learning and Memory; Nerve Growth Factor; Neuroendocrine Aging: Pituitary Metabolism.

Further Reading

Beilharz EJ, Russo VC, Butler G, et al. (1998) Co-ordinated and cellular specific induction of the components of the IGF/IGFBP axis in the rat brain following hypoxic–ischemic injury. *Brain Research* 59: 119–134.

Bondy CA and Cheng CM (2004) Signaling by insulin-like growth factor 1 in brain. *European Journal of Pharmacology* 490: 25–31.

Bridges RS and Grattan DR (2003) Prolactin-induced neurogenesis in the maternal brain. *Trends in Endocrinology and Metabolism* 14: 199–201.

Cardona-Gomez GP, Mendez P, DonCarlos LL, Azcoitia I, and Garcia-Segura LM (2001) Interactions of estrogens and insulin-like growth factor-I in the brain: Implications for neuroprotection. *Brain Research. Brain Research Reviews* 37: 320–334.

Carro E, Trejo JL, Nunez A, and Torres-Aleman I (2003) Brain repair and neuroprotection by serum insulin-like growth factor I. *Molecular Neurobiology* 27: 153–162.

Digicaylioglu M, Garden G, Timberlake S, Fletcher L, and Lipton SA (2004) Acute neuroprotective synergy of erythropoietin and insulin-like growth factor I. *Proceedings of the National Academy of Sciences of the United States of America* 101: 9855–9860.

Drago J, Murphy M, Carroll SM, Harvey RP, and Bartlett PF (1991) Fibroblast growth factor-mediated proliferation of central nervous system precursors depends on endogenous production of insulin-like growth factor I. *Proceedings of the National Academy of Sciences of the United States of America* 88: 2199–2203.

Edmondson SR, Werther GA, Russell A, Le Roith D, Roberts CT Jr., and Beck F (1995) Localization of growth hormone receptor/binding protein messenger ribonucleic acid (mRNA) during rat fetal development: Relationship to insulin-like growth factor-I mRNA. *Endocrinology* 136: 4602–4609.

Feldman EL, Sullivan KA, Kim B, and Russell JW (1997) Insulin-like growth factors regulate neuronal differentiation and survival. *Neurobiology of Disease* 4: 201–214.

Giacobini MM, Olson L, Hoffer BJ, and Sara VR (1990) Truncated IGF-1 exerts trophic effects on fetal brain tissue grafts. *Experimental Neurology* 108: 33–37.

Jones JI and Clemmons DR (1995) Insulin-like growth factors and their binding proteins: Biological actions. *Endocrine Reviews* 16: 3–34.

LeRoith D, Werner H, Faria TN, Kato H, Adamo M, and Roberts CT Jr. (1993) Insulin-like growth factor receptors: Implications for nervous system function. *Annals of the New York Academy of Sciences* 692: 22–32.

Russo VC, Gluckman P, Feldman EL, and Werther GA (2005) The insulin-like growth factor system and its pleiotropic functions in brain. *Endocrine Reviews* 26(7): 916–943.

Russo VC, Kobayashi K, Najdovska S, Baker NL, and Werther GA (2004) Neuronal protection from glucose deprivation via modulation of glucose transport and inhibition of apoptosis: A role for the insulin-like growth factor system. *Brain Research* 1009: 40–53.

Scheepens A, Williams CE, Breier BH, Guan J, and Gluckman PD (2000) A role for the somatotropic axis in neural development, injury and disease. *Journal of Pediatric Endocrinology & Metabolism* 13(supplement 6): 1483–1491.

Sepp-Lorenzino L (1998) Structure and function of the insulin-like growth factor I receptor. *Breast Cancer Research and Treatment* 47: 235–253.

Torres-Aleman I, Naftolin F, and Robbins RJ (1990) Trophic effects of basic fibroblast growth factor on fetal rat hypothalamic cells: Interactions with insulin-like growth factor I. *Brain Research. Developmental Brain Research* 52: 253–257.

Intermediate Filaments

D L Eng, Palo Alto, CA, USA
L F Eng, Stanford University School of Medicine, Palo Alto, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

In the nervous system, the shape of neurons and astrocytes is largely maintained by structures built with intermediate filaments (IFs). Intermediate filaments comprise a common gene family divided into five classes of structures which display such unusual degrees of cell specificity that they are often used as markers of cellular differentiation. All five gene classifications share homology in a core rod domain, which contains multiple α-helical domains that can form coiled coils. There are conserved regions with sequence homology of common domains that have such similarity that some antibodies can recognize all known intermediate filaments from mammals through a wide range of invertebrates. IFs are also similar in ultrastructural form between all five classes, and they form 8 to 10 nm ropelike filaments that may be several micrometers long. Exclusively expressed in neurons are class IV neurofilaments (NFs). The class III intermediate filament protein, glial fibrillary acidic protein (GFAP), is expressed in astrocytes and Schwann cells. Class III and IV intermediate filament proteins are the two main members of the cytoskeleton proteins that act together to form and maintain cell shape and facilitate the transport of particles and organelles within the cytoplasm. NFs are different from the other IF classes in that they have side arms that project from the surface. These side arms result in less dense packing and give spacing between NFs. In nonneuronal cells, IFs are seen as densely packed bundles. Vimentin is the most widely expressed type III IF and is found in a variety of cell types, such as fibroblasts, microglia, and smooth muscle cells, as well as in embryonic cell types, including neuronal and glial precursors. Vimentin is transiently expressed during oligodendrocyte cell development. Unlike NFs, the type III IF GFAP homopolymer is shorter and denser because it lacks side arms which project from the sides. Nestin is the most divergent of the IFs that form filaments and is sometimes considered to be a distinct IF type. Although nestin is specific to the nervous system, it is expressed in multipotent neuroectodermal precursors and suppressed during subsequent development. For this reason, it can be used as an early marker for differentiation of precursor cells. Keratins are type I and II intermediate filaments which are found in epithelial cells throughout the body in various stoichiometries characteristic to the individual tissue type. The type IV NFs are heteropolymers and share sequence homology in the rod domains and have common exon and intron structures different from the type I–III classes. Nuclear lamins are IFs, polypeptides which form the walls of the nuclear envelope structure, and are the type V class. Only cells with a nucleus (eukaryotes) express type V IFs. Bacteria (prokaryotes) do not have the class type V IFs. An entire phylum in the animal kingdom, arthropods, expresses only type V nuclear lamins, and arthropod cells have no IF cytoskeletal structures. Otherwise, nearly all somatic cells have IFs, although some are only expressed transiently during development.

Three other IFs are seen transiently in neuronal development: α-internexin, peripherin, and nestin. These three IFs are most prominently expressed during development and then downregulated in selected neurons and neuronal precursors. Based on sequence and gene structure, peripherin is a type III IF, whereas α-internexin is a type IV IF. Whereas the NF triplet proteins and α-internexin may be components of both central nervous system (CNS) and peripheral nervous system (PNS) neurons, peripherin appears to be expressed preferentially in PNS neurons. Both α-internexin and peripherin can coassemble with the NF triplet proteins *in vitro* and may do so *in vivo* but can also form homopolymeric filaments. Both are expressed at higher levels in a variety of developing neurons, and expression becomes more restricted as neurons mature. Many neurons cease to express α-internexin following maturation. However, a few neurons retain the expression of IFs containing α-internexin. Notably, IFs of parallel fibers in the cerebellar cortex contain only α-internexin subunits. α-Internexin continues to be expressed in some peripheral neurons into maturity. In contrast to NF subunits, α-internexin and peripherin self-assemble into homopolymers and coassemble with NF subunits.

Neuronal Intermediate Filaments

Many traditional histological techniques have visualized neurons using metal stains that have specific interactions with NFs. Using electron microscopy, invented in 1931 by Max Knoll and Ernst Ruska in Germany, IFs of all classes appear as 8–12 nm solid ropelike fibers and look densely packed in all cells except the neurons. NFs which are only found in neurons, as visualized by electron microscopy, appear as 10 nm fibers that can be hundreds of microns long. NFs have phosphorylated side arms that prevent dense packing, and these projections give axons

their cytoplasm volume. NFs constitute the most abundant structures in large myelinated neurons and can account for 12% or more of total protein in these neurons. NFs are formed from three subunit proteins known as the NF triplet: NF high-molecular-weight subunit (NFH; 180–200 kDa), NF middle-molecular-weight subunit (NFM; 130–170 kDa), and NF low-molecular-weight subunit (NFL; 60–70 kDa. Each of the NF subunit proteins is coded for by a separate gene. The NF triplet proteins have a characteristic domain structure. The N-terminal regions of all three subunits interact via α-helical coiled coils to form the core of the filament. NFM and NFH also have long C-terminal regions, which project from the core filaments as side arms. NFH and, to a lesser extent, NFM have a large number of these Lys-Ser-Pro (KSP) consensus phosphorylation sites for proline-directed kinases in this C-terminal extension (>50 on NFH and >10 on NFM in many species). Almost all (>99%) of the assembled NFs in myelinated internodal regions are known to be stoichiometrically phosphorylated in the KSP repeat domain; cell bodies have less phosphorylated NF. In large myelinated axons, most, if not all, of these sites are phosphorylated. This phosphorylation of NFH and NFM C-terminal NF side arms alters the charge density on the NF surface, repelling adjacent NFs with similar charge. Electrostatic repulsion of adjacent neurofilaments, the mutual repulsion by the side arms of NFs, is thought to be a major determinant of axonal caliber. Some neurons lack NFs. Finally, NFs have an unusual degree of metabolic stability which stabilizes and maintains neuronal morphology.

Neurofilament Response to Injury and Disease

The neuronal response to injury has been studied in some detail to gain a better understanding why mammalian CNS neurons regenerate poorly or not at all, whereas PNS neurons regenerate effectively. Although regeneration is often described as paralleling developmental growth of neurons, the distinctive pattern of changes in the neuronal cytoskeleton during regeneration differs in some important ways from that during development. First, there is a coordinate downregulation in NF subunit expression following injury. NFH, NFM, and NFL protein and mRNA levels in the perikarya decline rapidly after injury and do not recover until after reconnection with an appropriate target cell has occurred. Concurrent with the reduction in NF protein following injury, both tubulin and actin increase significantly. During the elongation phase of regeneration tubulin and actin protein expression and associated increased expression of microtubule-asssociated proteins (MAPs) and microfilaments are high, but the NF protein expressions are low.

Alterations in the cytoskeleton are frequent hallmarks in neuropathology. A definitive diagnosis for many neurological disorders depends on a histological examination. Abnormal accumulation of NF is detected in many human neurodegenerative disorders, including amyotrophic lateral sclerosis (ALS), Alzheimer's disease (AD), dementia with Lewy bodies, Parkinson's disease (PD), and diabetes. These disorders show accumulations of neurofilaments in neurons. The identifying characteristic for a number of neuropathological conditions is a disrupted or aberrant cytoskeleton. This is a feature of many toxic neuropathies and a number of neurodegenerative conditions, including motor neuron disease and AD. Although other neuronal functions may also be affected and the initial cause of the disease may not directly involve the cytoskeleton, the associated pathogenic disruption of cytoskeletal function may be a key element in the loss of neurons or neuronal function.

PD was first described by English physician James Parkinson in 1817. PD, the shaking palsy, is caused by degeneration of the pigmented neurons in the substantia nigra (SN) of the brain and results in decreased dopamine availability. Clinical symptoms include decreased spontaneous movement, gait difficulty, postural instability, rigidity, and tremor. The major neuropathological change in PD is an accumulation of somal protein in SN neurons called the Lewy body. The Lewy body is composed of several proteins, being the three NF subunits and α-synuclein protein. Abnormal phosphorylated NFs form a compacted skein in neuronal soma of affected SN neurons. Point mutations in the Parkin gene account for many familial PD cases. It seems that mutations in NF genes may account for some inherited forms of PD.

A key hallmark of ALS is intraneuronal aggregates of NFs in degenerating motor neurons. Several ALS-like diseases with different pathogenic mechanisms have been associated with disruptions of axonal NFs. NFs accumulate in the cell bodies of motor neurons in patients with ALS and related motor neuron diseases. The latter example is particularly interesting because the symptoms of motor neuron disease can be produced in animal models in several very different ways. Some of these animal models have little obvious relevance to the cytoskeleton, such as those resulting from point mutations in a superoxide dismutase gene. Others directly involve NFs, such as transgenic mice that express defective NF subunits or overexpress normal NF subunits. Regardless of how these animal models are produced, a characteristic feature of motor neuron pathology is the

accumulation of poorly phosphorylated NFs in cell bodies and initial segments of the axon. It has been proposed that alterations in NF homeostasis are directly relevant to the pathogenesis of ALS.

Other diseases with disruptions in NF organization include diabetic neuropathy and Charcot–Marie–Tooth disease. For these diseases, the disruption of NFs is a secondary effect. In most cases, neuronal degeneration is an eventual consequence, but neuronal function may be impaired prior to substantial loss of neurons. Generally, disruption of NFs has the most severe consequences in large motor neurons, which is consistent with the fact that the largest neurons have the highest NF expression.

Diabetes is a disease in which the body does not produce or properly use insulin. The cause of diabetes is unknown, although both genetics and environmental factors such as obesity and lack of exercise appear to play roles. Diabetes is associated with symmetrical distal axonal neuropathy. The neuropathy is predominately in sensory neurons or dorsal root ganglia. Drug-induced diabetes by streptozocin increases phosphorylation of NF specific to the lumbar region dorsal root ganglion neurons. These drug-induced changes are not seen in other types of neurons. Pathologically, neurites from dorsal root ganglia are swollen with disorganized aggregates of NF and other proteins. The neuropathy associated with diabetes includes well-documented impairment of axonal transport, a reduction in axon caliber, and a reduced capacity for nerve regeneration.

Phosphorylation of cytoskeletal proteins has also been linked to various neuropathological conditions. For example, the NF accumulations seen in many diseases have aberrant phosphorylation patterns. Another significant change may be the hyperphosphorylation of a MAP such as tau, which appears to be a key step in the formation of neurofibrillary tangles in AD. These tangles contain distinctive paired helical filaments that are extremely resistant to solubilization and analysis. They were initially thought to be aberrant NFs based on their dimensions, but a careful immunochemical and biochemical dissection showed that the primary polypeptide present in the tangles was tau, the axonally enriched MAP. The tau in neurofibrillary tangles is differentially phosphorylated, and this misphosphorylation is thought to play a role in the formation of tangles. Although the precise etiology of neurofibrillary tangles and their relationship to the deposition of amyloid are not certain, the appearance of these tangles is closely correlated with loss of neurons, and their number is a good indication of how far the disease has advanced. Clearly, disruption of the cytoskeleton is an important component of the AD neuropathology.

Glial Intermediate Filaments

The abundant and most extensively studied intermediate filament in astrocytes is GFAP. The term evolved from 'plaque' protein, glial fibrillary protein, to GFAP or GFA protein and was originally based on its high acidic amino acid content and its morphological fibrous appearance. Its rapid response to injury and insult, and its highly antigenic properties, has permitted extensive chemical, biological, and molecular biological investigations on CNS development, function, and disease.

Properties of GFAP

GFAP shows species-specific amino acid sequence heterogeneity, and its molecular weight ranges from 48 kDa for mouse to 49 kDa for human, 50 kDa for bovine, and 51 kDa for rat. The relatively slow metabolic turnover rate for GFAP is consistent with a structural role for glial filaments in the cytoplasm. GFAP is composed of three distinct regions: an N-terminal HEAD region, a central ROD region, and a C-terminal TAIL region. The ROD regions are highly conserved among IF proteins, whereas the size and amino acid sequences of the HEAD and TAIL regions vary. The head region of porcine GFAP is composed of 35-amino-acid residues and has β-turn structure. It contains eight arginine residues that characterize this highly basic region. The ROD region of GFAP is composed of 345 amino acid residues, containing long tandem repeats of seven amino acid sequences called 'heptad repeats.' This region is essential for the packing of the α-helical coil which is important for the formation of 10 nm filaments. The tail region of approximately 50-amino-acid residues has a globular structure and may be involved in the interactions between IF and other proteins.

Posttranslational modification to the N- and C-terminal domains of IFs are common, whereas those within the central α-helical domain are rare. The posttranslational modification of greatest functional and structural significance arises from the action of various kinases on serine and threonine embedded within specific recognition sequences. The phosphorylation events which have a regulatory role in the assembly and turnover of most IFs *in vivo* are confined to the N- and C-terminal domains of the chains. Phosphatases, which dephosphorylate the chains, have an equally important role in governing assembly and function. In cytokeratins, vimentin, and lamins, the effect of phosphorylation is to disassemble the IFs, whereas dephosphorylation allows spontaneous self-assembly to proceed. These changes typically occur during cell division. Phosphorylation and dephosphorylation of the IF HEAD regions have a critical

role for the maintenance and reconstruction of IFs. GFAP filaments disassemble into a soluble form when the amino acids in the HEAD are phosphorylated. The HEAD region of GFAP is positively charged due to the many arginine residues. The mechanism of GFAP assembly regulation may be related to regulation of the electric charge in the HEAD domain by phosphorylation. The sites of second messenger-dependent protein kinases have been identified in the HEAD region of GFAP. Four serines (Ser8, Ser13, Ser14, and Ser34) and one threonine (Thr7) are known to be phosphorylated by different protein kinases: cyclic-AMP-dependent protein kinase (A kinase), Ca^{2+} calmodulin-dependent kinase II (CaM kinase II), C kinase, and Rho kinase. A kinase phosphorylates Thr7, Ser8, Ser13, and Ser34 residues of the HEAD region. CaM kinase II phosphorylates Ser8, Ser17, and Ser34. cdc2 kinase phosphorylation of GFAP did not cause disassembly, whereas CaM kinase II phosphorylation did induce disassembly of the phosphorylation of vimentin, and Rho kinase inhibited GFAP filament formation *in vitro*.

Molecular Biology of GFAP

GFAP gene The molecular cloning of the mouse gene opened a new and productive area for GFAP studies. Genomic clones have been obtained from human, mouse, and rat GFAP genes. Each gene is composed of nine exons distributed over approximately 10 kb of DNA and yields a mature mRNA of approximately 3 kb. The coding sequences for the three genes are highly homologous. Strong homology also extends upstream of the RNA start site for approximately 200 bp, recurs between approximately-1300 and -1700 (RNA start point = +1), and is present in some intronic regions. The primary sites for the initiation of RNA and protein synthesis are essentially identical for the three genes, and each contains a TAT-like sequence (CATAAA or AATAA) in the expected 5′-flanking position. In addition to GFAP-α, two additional mRNAs that start at different sites have been identified, GFAP-β and GFAP-γ. The different tissue distributions of GFAP-α, -β, and -γ mRNAs suggest that the synthesis of each is subject to unique control. All transcriptional studies to date either have explicitly measured GFAP-α or have not distinguished among the possible mRNA isotypes. In the PNS, GFAP-β RNA is thought to be the predominant form of GFAP, whereas, GFAP-α is the predominant form in the CNS. The GFAP-β mRNA has been found in both CNS and non-CNS tissues, including mouse bone marrow and spleen. One report suggests the possible existence in rat brain of a novel GFAP mRNA isoform, GFAP-δ, that differs in the C-terminal tail domain. GFAP transgenes have been used extensively to study signaling pathways that operate during development, disease, and injury – all states that increase GFAP gene activity.

GFAP promoter A major contribution of GFAP transcription studies has been the identification of DNA sequences that can target transgene expression to mouse astrocytes. The entire GFAP genomic clone and a gfa2 promoter have been utilized. The use of the GFAP promoter has resulted in the initiation of a range of studies, including *in vivo* tests of various gene products on astrocyte function, use of astrocytes as factories for bioactive molecules, and creation of disease models.

GFAP and Vimentin Knockout Mice

GFAP knockout mice The technique of gene knockout has been used to examine intermediate filaments in mice and has provided the first evidence that intermediate filaments are directly involved in cell resilience and the maintenance of tissue integrity. To investigate the structural role of GFAP *in vivo*, mice carrying null mutations in GFAP have been created by disrupting the GFAP gene in embryonic stem cells. These mice were completely devoid of GFAP but exhibited normal development and showed no obvious anatomical abnormalities in the CNS. Studies examining the anatomy and synaptic function of the cerebellum of GFAP-deficient mice have shown that long-term depression (LTP) is deficient in the GFAP mutant cerebellum without any detectable anatomical abnormalities. Furthermore, GFAP mutant mice exhibited significant impairment of eyeblink conditioning without any detectible deficits in motor coordination tasks. Another study suggested that GFAP expression is essential for normal white matter architecture and blood–brain barrier integrity, and its absence leads to late-onset CNS dysmyelination. GFAP null mice also showed subtle changes in astrocyte morphology and enhanced LTP in hippocampal neurons. Mice lacking GFAP are hypersensitive to cervical spinal cord injury caused by sudden acceleration of the head.

Vimentin knockout mice Animals homozygous for the vimentin knockout mouse develop and reproduce without an obvious phenotype but show a cerebellar defect and impaired motor coordination. Fibroblast derived from these mice are also mechanically weak and severely disabled in their capacity to migrate and to contact a three-dimensional collagen network. Wounds in the vimentin null adult mice showed delayed migration of fibroblasts into the wound site. The GFAP network in the vimentin

null mice has disrupted GFAP and fails to assemble into a filamentous network in astrocytes that normally co-express GFAP and vimentin (i.e., corpus callosum and Bergmann glia).

GFAP and vimentin double knockout mice In studies on the reaction to injury in the CNS, glial scar formation appeared normal after spinal cord or brain lesions in GFAP null or vimentin null mice but was impaired in double null mice that developed less dense scars frequently accompanied by bleeding. These studies have suggested that GFAP and vimentin are required for proper glial scar formation in the injured CNS and that some degree of functional overlap exists between these IF proteins. Reactive astrocytes devoid of IFs exhibit clear morphological changes and profound defects in cell motility. Culture studies have shown that astrocytes in mice deficient for both GFAP and vimentin cannot form IFs even when nestin is expressed and are thus devoid of IFs in their reactive state. Additional studies with the double mutant have suggested a role for the cytoskeleton in astrocyte volume regulation and the involvement of intermediate filaments in the process. The rate of glucose uptake through facilitative hexose transporters was not affected by the depletion of GFAP or vimentin, nor was ascorbate uptake and efflux.

Alexander's Disease

Alexander's disease (AxD) is a rare disease of unknown origin which is often fatal to infants, juveniles, and adults. The pathological hallmark of AxD is the presence of Rosenthal fibers, cytoplasmic inclusions in astrocytes that contain increased amounts of GFAP in association with stress proteins. Six lines of transgenic mice have been generated that carry added copies of the human GFAP (hGFAP) gene in order to determine the properties of astrocytes containing increased GFAP without external stimulation and activation. The discovery that mice expressing an hGFAP transgene formed abundant Rosenthal fibers suggested that mutations in the GFAP gene were a cause of AxD. Sequencing results from a number of laboratories identified GFAP coding mutations in most cases of the disease, including the infantile and juvenile forms. These mutations were found in the 1A, 2A, and 2B segments of the conserved central ROD domain of GFAP and also in the variable TAIL region. All changes detected were heterozygous missense mutations, and none were found in any parent of a patient who had been tested. These data indicate that most cases of AxD arise through *de novo*, dominant, GFAP mutations. The authors suggest that the gain of function is due to a partial block of filament assembly that leads to accumulation of an IF that participates in a toxic interaction. AxD is the only known primary disease of astrocytes.

GFAP in Reactive Gliosis

Astroglia in the CNS and PNS reacts to injury by hypertrophy and, in some cases, proliferation. The functions of reactive astroglia are still not well understood. A common feature of these cells is expression of GFAP. In the CNS of higher vertebrates, following injury as a result of trauma, disease, genetic disorders, or chemical insult, astrocytes become reactive and respond in a typical manner, termed astrogliosis. Reactive astrogliosis is a prominent feature of astrocytes adjacent to injury. Astrogliosis is characterized by rapid synthesis of GFAP intermediate filaments. Numerous *in vitro* and *in vivo* studies on the molecular profiles of substances which are upregulated during astrocyte activation document the complex and varied responses of astrocytes to injury. Increased protein content or immunostaining of GFAP has been found in experimental models involving gliosis. These include the cryogenic lesion in the brain, stab wounds, experimental allergic encephalomyelitis (EAE), hyperthermia, electrically induced seizures, and toxic lesions. Increase of GFAP mRNA has been associated with scrapies, AD brains, and Creutzfeldt–Jacob disease and was observed in 6-hydroxydopamine lesion of the substantia nigra in rat, in a mechanical lesion to rat cerebral cortex, in entorhinal cortex lesions, in corticospinal axotomy, in EAE, in the lesioned dentate gyrus, and in cerebral freeze injury.

Three diseases which show intense gliosis are multiple sclerosis, an inflammatory demyelinating disease characterized by demyelinating plaques composed of bare axons and extensive reactive astrogliosis; adrenoleukodystrophy, a hereditary disorder of lipid metabolism; and AD, characterized by extensive gliosis and Rosenthal fiber deposits. Reactive gliosis occurs with any type of injury to the CNS. Neuropathological differences and severity of gliosis and time course of disease development are characteristic to the affliction and can be useful in differentiating between the following diseases: ALS, Gerstmann–Straussler syndrome, Huntington's disease, Wilson's disease, Pick's disease, PD, AD, and Down's syndrome. Astrocytic gliosis is a prominent neuropathologic change in AD, most frequently in association with neuritic plaques.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Amyotrophic Lateral Sclerosis (ALS); Astrocyte: Response to Injury; Cytoskeletal Interactions in the Neuron; Glial Responses to Injury; Neurofilaments: Organization and Function in Neurons; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Al-Chalabi A and Miller CJ (2003) Neurofilaments and neurological disease. *BioEssays* 25(4): 346–355.

Brady ST (1993) Motor neurons and neurofilaments in sickness and in health. *Cell* 73: 1–3.

Brady ST (1993) Axonal dynamics and regeneration. In: Gorio A (ed.) *Neuroregeneration*, pp. 7–36. New York: Raven Press.

Brenner M (1994) Structure and transcription of the GFAP gene. *Brain Pathology* 4: 245–257.

Brenner M and Messing A (1996) GFAP transgenic mice. *Methods: A Companion to Methods in Enzymology* 10: 351–364.

Eng LF, Ghirnikar RS, and Lee YL (2000) Glial fibrillary acidic protein: GFAP: Thirty-one years (1969–2000). *Neurochemical Research* 25: 1439–1451.

Eng LF and Lee YL (1995) Intermediate filaments in astrocytes. In: Ransom BR and Kettenmann H (eds.) *Neuroglial Cells*, pp. 650–667. New York: Oxford Press.

Fliegner KH and Liem RKH (1991) Cellular and molecular biology of neuronal intermediate filaments. *International Review of Cytology* 131: 109–167.

Fuchs E and Weber K (1994) Intermediate filaments: Structure, dynamics, function, and disease. *Annual Review of Biochemistry* 63: 345–382.

Lasek RJ (1988) Studying the intrinsic determinants of neuronal form and function. In: Lasek RJ and Black MM (eds.) *Intrinsic Determinants of Neuronal Form and Function*, pp. 1–60. New York: Liss.

Lee MK and Cleveland DW (1996) Neuronal intermediate filaments. *Annual Review of Neuroscience* 19: 187–217.

Lewis SA, Balcarek JM, Krek V, Shelanski M, and Cowan NJ (1984) Sequence of a cDNA clone encoding mouse glial fibrillary acidic protein: Structural conservation of intermediate filaments. *Proceedings of the National Academy of Sciences of the United States of America* 81: 2743–2746.

Li R, Messing A, Goldman JE, and Brenner M (2002) GFAP mutations in Alexander disease. *International Journal of Developmental Neuroscience* 20: 259–268.

Liu Q, Xie F, Siedlak SL, et al. (2004) Biomedicine and diseases: Review: Neurofilament proteins in neurodegenerative diseases. *Cellular and Molecular Life Science* 61: 3057–3075.

Manetto V, Sternberger NH, Perry G, Sternberger LA, and Gambetti P (1988) Phosphorylation of neurofilaments is altered in amyotrophic lateral sclerosis. *Journal of Neuropathology and Experimental Neurology* 47: 642–653.

Matsuzawa K, Kosako H, Azuma I, Inagaki N, and Inagaka M (1998) Possible regulation of intermediate proteins by Rho-binding kinases. *Subcellular Biochemistry* 31: 423–435.

McLean WG (1997) The role of axonal cytoskeleton in diabetic neuropathy. *Neurochemical Research* 22: 951–956.

Messing A, Head MW, Galles K, et al. (1998) Fatal encephalopathy with astrocyte inclusions in GFAP transgenic mice. *American Journal of Pathology* 152: 391–398.

Parry DAD and Steinert PM (1999) Intermediate filaments: Molecular architecture, assembly, dynamics, and polymorphism. *Quarterly Reviews of Biophysics* 32: 97–187.

Sternberger NH, Sternberger LA, and Ulrich J (1985) Aberrant neurofilament phosphorylation in Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America* 82: 4274–4276.

Interstitial Axon Branching/Collateral Elimination

K Kalil, University of Wisconsin–Madison, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During development, growing axons navigate over long distances to reach distant targets. They are guided by highly motile growth cones at axon tips that respond to attractive and inhibitory guidance cues in the environment. In this way, growth cones steer toward and recognize their correct targets. However, in many central nervous system (CNS) pathways, the growth cone of the primary axon does not extend directly into targets. Instead, after the primary growth cone has grown past the target, branches develop interstitially from the axon shaft at the correct locations to extend into their appropriate targets (Figure 1). Collateral branching, like growth cone turning, changes the direction of axon outgrowth and is thus an important form of axon guidance. Axon branching enables a single neuron to connect with multiple targets and is therefore essential for the assembly of complex neural circuits. Interstitial branching has been well documented in mammalian CNS pathways such as the corpus callosum and corticospinal tract that arise from the cerebral cortex and innervate the contralateral cortex and spinal cord, respectively.

Given the fundamental importance of axon branching for establishing specific synaptic connectivity, surprisingly little is known about how branching is regulated. However, recent evidence suggests that axon guidance at the growth cone and at branch points share common mechanisms. Guidance factors, acting through receptors in the cell membrane, are attractive or repellent to growth cones during pathfinding and also promote or inhibit axon branching. Directed growth from the growth cone and branching from the axon shaft are also mechanistically linked by similar changes in the dynamics of the actin and microtubule cytoskeleon and their associated intracellular signaling pathways.

Studies *in vivo* and *in vitro* have shown that axon branching often occurs after the primary axon has stopped elongating. Independent axon outgrowth and branching have been demonstrated in CNS pathways. For example, direct visualization of fluorescently labeled axons extending in living brain slices revealed that in the corpus callosum the primary axonal growth cone enlarges and pauses beneath the contralateral cortex while axon branches extend dorsally toward cortical targets. However, the mechanisms that slow or stop extension of the primary axon along its pathway while permitting growth of the axon's collaterals into targets are not well understood. Establishment of neural connectivity during development involves not only neurite outgrowth but pruning of excess or inappropriate axon branches by retraction or degeneration. Some axons in adult animals are also capable of robust sprouting after injury which may involve mechanisms similar to those that regulate CNS axon branching during development. However, the focus of this article is on development of axon collateral branches rather than on branch elimination or collateral sprouting after injury.

Effects of Growth and Guidance Factors on Axon Branching

Attractive and Inhibitory Guidance Molecules

Families of evolutionarily conserved guidance cues have been discovered that either attract or repel axonal growth cones by chemotropism, defined as guidance directed by a gradient of a diffusible factor. Repulsive guidance molecules cause growth cones to collapse, inhibiting their advance or causing growth cones to turn away from the source of the repellent cue. For example, Slit proteins repel axons at the midline in *Drosophila* and regulate midline crossing in regions of the mammalian CNS such as the optic chiasm and corpus callosum. Surprisingly, while Slit proteins collapse and repel growth cones, they can induce branching on dorsal root ganglion (DRG) sensory axons. Semaphorin 3A (Sema3A) has also been shown to promote branches from retinal axons after growth cone collapse. Both the Slit proteins and the semaphorins, particularly Sema3A, are expressed at high levels in the developing mammalian cortex and act as chemorepellants for cortical axons *in vitro*. Application of Sema3A to cultures of early postnatal cortical neurons not only collapsed growth cones but also reduced the development of axon branches by half. While Sema3A reduced branch length, it had no effect on the length of the primary axon. Thus, Sema3A specifically reduces cortical axon branching but has no effect on axon outgrowth *per se*. Interestingly, recent studies have shown that Sema3A and the Slit proteins repel cortical axons but respectively attract or promote the growth and branching of their dendrites. Thus, in many situations, the growth of

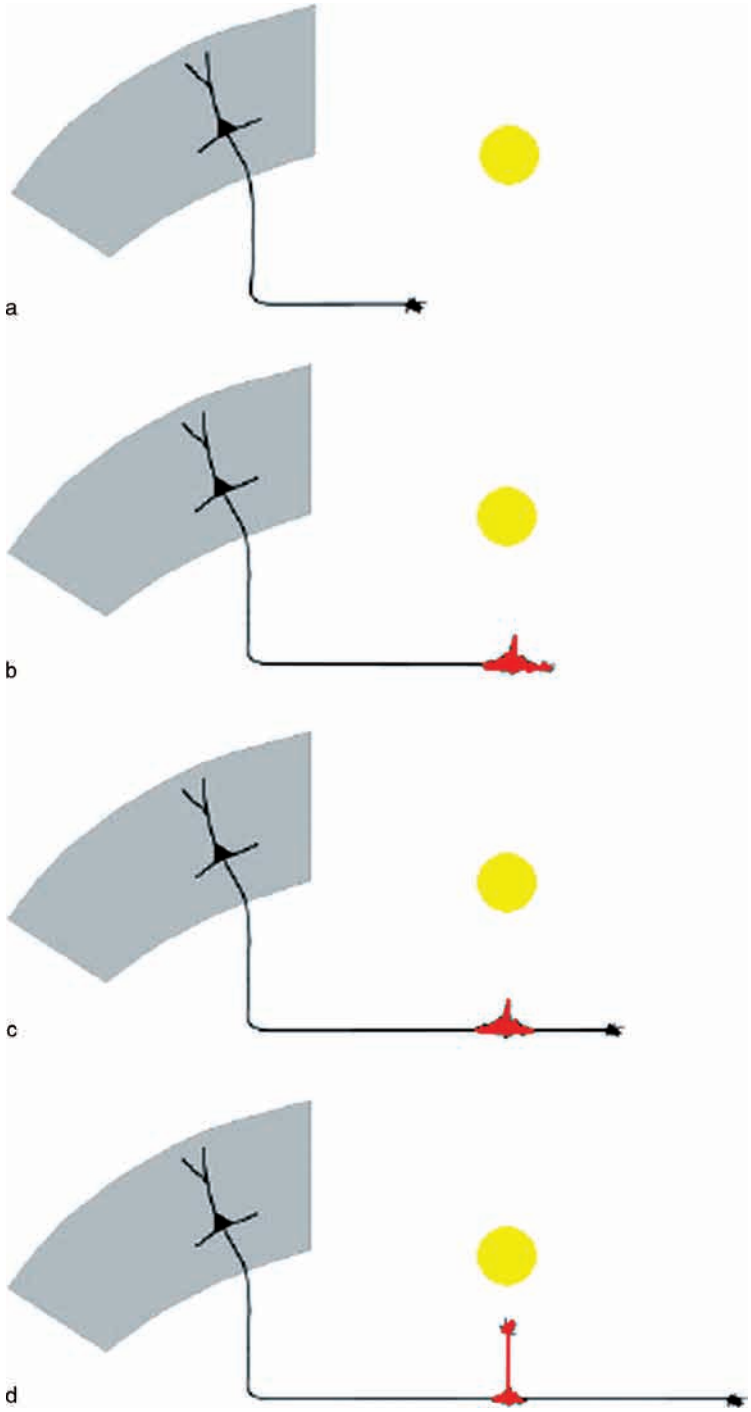

Figure 1 A model for interstitial axon branching. Schematics represent different stages in axon branching *in vivo*. (a) Growth cone of an axon is advancing along a pathway toward its target, indicated by the yellow circle. (b) In response to target-derived signals, the primary growth cone pauses in the vicinity of the target and enlarges. (c) After the primary growth cone has resumed forward advance, remnants of the reorganized growth cone appear as filopodial or lamellar activity along the axon shaft. (d) A branch from these active membrane regions extends toward the target after various time delays. Typically, branches extend interstitially some distance behind the primary growth cone.

dendritic versus axonal processes can be differentially regulated by a single guidance factor.

Other guidance molecules such as the ephrins can also induce growth cone repulsion and regulate axon branching. For example, members of the Eph family of receptor tyrosine kinases and their ephrin ligands play important roles in the development of the retinotectal projection. Retinal axons initially overshoot

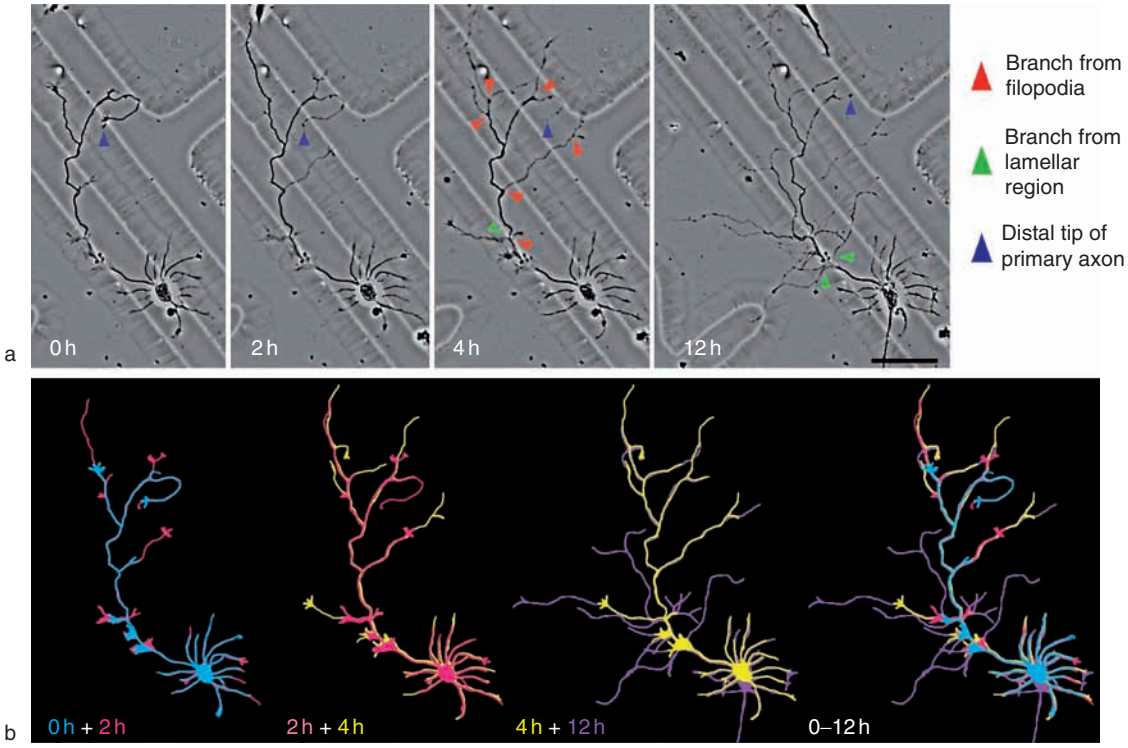

Figure 2 Netrin-1 induces axon branching from filopodial protrusions. (a) Phase-contrast image of a single postnatal day 1 cortical neuron observed over 12 h in time lapse. Branches emanating from filopodial protrusions are marked with an orange arrowhead; branches forming from spread lamellar regions of the axon shaft are marked with a green arrowhead. The distal tip of the primary axon is marked with a blue arrowhead in all frames. (b) Composite tracings of the neuron in (a) overlaid on the previous time point showing the progression of branch formation and extension. Note that the primary axon does not increase in length. The time points are color coded for clarity and correspond to the color of the tracing of the neuron. Scale = 50 μm.

their future targets in the optic tectum during development. Topographic specificity is achieved not by guidance of terminal growth cones to tectal target neurons but by interstitial axon branching of retinal axons at appropriate locations in the tectum. Gradients of ephrinAs regulate development of this topographic map, and map formation is thought to depend on a combination of branch-promoting and branch-suppressing activities. Studies using anatomical mapping of retinotectal connections in EphA7 mutant mice revealed that opposing gradients of ephrinAs and EphA7 are required for development of appropriate retinotectal topographic mapping. Formation of appropriate termination zones in the tectum depends upon suppression of retinal axon branching by EphA7 anterior to future termination zones.

The netrin family of guidance molecules can guide growth cones in many regions of the nervous system. Studies of cortical axon outgrowth in explant cultures, for example, have shown a potent effect of netrin-1 in attracting growing axons. Netrin-1 can also elicit a large and rapid increase in axon branching. In cultured cortical neurons, spontaneous development of axon branches typically requires several days and involves pausing and enlargement of the primary growth cone at future branch points. In contrast, netrin-1 applied to hippocampal or to cortical neurons promoted formation of filopodia directly from the axon shaft within minutes. Over several hours, many of these filopodia on cortical axons developed into branches tipped by growth cones (Figure 2).

Growth Factors

Growth factors such as nerve growth factor (NGF), fibroblast growth factor (FGF-2), brain-derived neurotrophic factor (BDNF), and neurotrophin-3 that promote axon outgrowth are also potent stimulators of axon branching. Injection of BDNF into the optic tectum of living tadpoles increased the branching and complexity of optic axon arbors, whereas neutralizing antibodies to BDNF reduced their arborization and complexity. These effects occurred rapidly and were specific to BDNF. Interestingly, BDNF had opposite effects on dendritic arbors of retinal ganglion cells. Injection of BDNF into the tadpole retina reduced dendritic arborization whereas function-blocking antibodies increased arbor complexity. Target-derived NGF also promotes axonal sprouting. Localized

sprouting of axon collateral branches, that began as filopodia, was induced on cultured DRG neurons when NGF covalently conjugated to polystyrene beads was applied directly to the axons. FGF-2 locally applied in the form of coated beads induced cortical axon branching in regions where the beads contacted the axon shaft. However, in contrast to the rapid effects of locally applied netrin-1, development of axon branches in close proximity to the FGF-2 coated beads required several days in culture.

While growth and guidance factors can evoke filopodial protrusions and branches directly from the axon shaft, observations in living cortical slices and in dissociated cell cultures have shown that pausing behaviors and enlargement of the primary growth cone can demarcate branch points along cortical axons from which branches later develop interstitially. Moreover, effects on branching by global application of FGF-2 and Sema3A to cortical neurons are mediated by effects on the primary growth cone. Thus, FGF-2 increased axon branching by inducing growth cones to pause and enlarge, whereas Sema3A collapsed cortical growth cones and the resulting absence of large pausing growth cones reduced branching by about 50%. Bath application of netrin-1 actually decreased the numbers of large paused growth cones, and many branches resulting from netrin-1 treatment extended directly from the axon shaft rather than from expanded branch points. The rapid growth of branches on axons of netrin-1-treated neurons, in contrast to the growth of branches over several days in control and FGF-2-treated cultures, suggests that different intracellular signaling mechanisms may be involved. Branching of cortical axons in response to cues from pontine targets can occur by extension of filopodial processes directly from the axon shaft. Recent evidence suggests that axon pruning also involves diverse axon remodeling processes including selective retraction and local degeneration of axon branches. It seems likely that in different contexts axon branches may also develop by different cellular mechanisms.

Many of the factors that influence axon branching have no effect on axon length. Thus, netrin-1 increased number of branches as well as branch length, whereas Sema3A decreased branch numbers and length, but neither treatment affected axon length. Although it is generally the case that cues attractive or repellent to cortical axons increase or decrease their branching, respectively, these results suggest that axon outgrowth and axon branching are independently regulated. This is consistent with the situation *in vivo* where axons can stop elongating while collaterals of the same axon are extending into their targets.

Cytoskeletal Reorganization in Growth Cones and Branches

Cytoskeletal Dynamics

As motile growth cones extend toward their targets they advance, retract, turn, and branch. Growth cone steering and extension of axon branches require reorganization of the microtubule and actin cytoskeleton. Microtubules are long hollow polymers of tubulin that form dense bundles in axons and dendrites. They not only lend structural support to neuronal processes but also serve as railways along which materials are transported from the cell body into neurites. The dynamic properties of microtubules, that is, their ability to grow and shrink in dynamic instability, enable them to explore growth cone lamellipodia and filopodia. In growth cones, the fast-growing (plus) ends of microtubules face outward toward the periphery. Actin filaments play an essential role in driving motility and guidance of the growth cone. In veil-like lamellipodia actin filaments form a meshwork, but within filopodia (the spiky finger-like protrusions from the growth cone periphery that are important for exploring the environment) actin filaments form straight bundles (Figure 3). Actin filament dynamics involve cycles of net assembly of actin filaments at the leading edge of the growth cone and disassembly of filaments proximally.

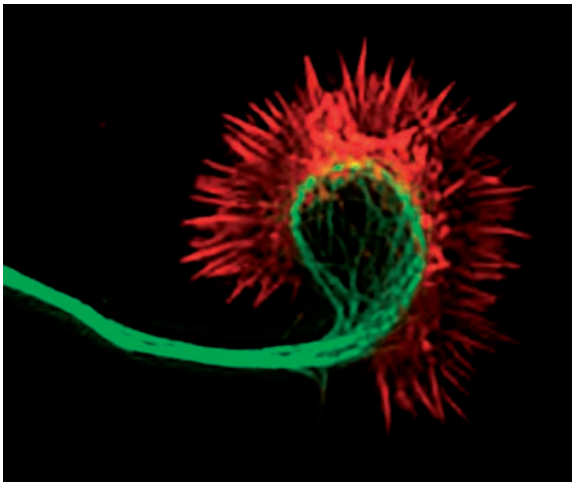

Figure 3 A large pausing cortical axonal growth cone *in vitro*. Microtubules, pseudocolored green, are stained with an antitubulin fluorescent antibody. Actin filaments, pseudocolored red, are stained with rhodamine phalloidin. Microtubules are bundled together in the axon, splay apart in the central region of the growth cone, and form a prominent loop surrounding the central region. Actin filaments extend into the lamellipodia and form straight bundles in spiky filopodia. In the transition region between the growth cone center and periphery, microtubules overlap (yellow) with actin filaments.

Visualizing fluorescently labeled microtubules and actin filaments in living neurons with high-resolution time lapse microscopy has revealed the dynamic cytoskeletal changes that underlie growth cone motility and guidance behaviors as well as axon branching. Microtubules are bundled together in the axon shaft and also extend into the central and peripheral regions of the growth cone. They are splayed apart in growth cones that are extending, but when growth cones stall microtubules form prominent loops surrounding the central region (Figure 3). Similarly, at axon branch points, microtubules undergo localized debundling. Microtubule reorganization at the growth cone and at axon branch points are strikingly similar. Along the axon shaft, microtubules splay apart and locally fragment prior to development of branches. During transitions of the growth cone from pausing to growth states, microtubules reorganize from looped to splayed configurations followed by fragmentation. Reorganization and fragmentation permit short microtubules to explore new directions of growth with forward and backward movements within the growth cone periphery and in branches extending from the axon shaft (Figure 4). Observations of long-term branching events showed that microtubules invaded and remained in branches favored for further growth but withdrew from branches that regressed, suggesting that microtubules can stabilize developing branches. Thus, cytoskeletal mechanisms underlying axon branching involve the reorganization of the microtubule array into a more labile form whereby microtubules can become debundled and fragmented. This plasticity would permit microtubules to explore the growth cone periphery and newly forming branches.

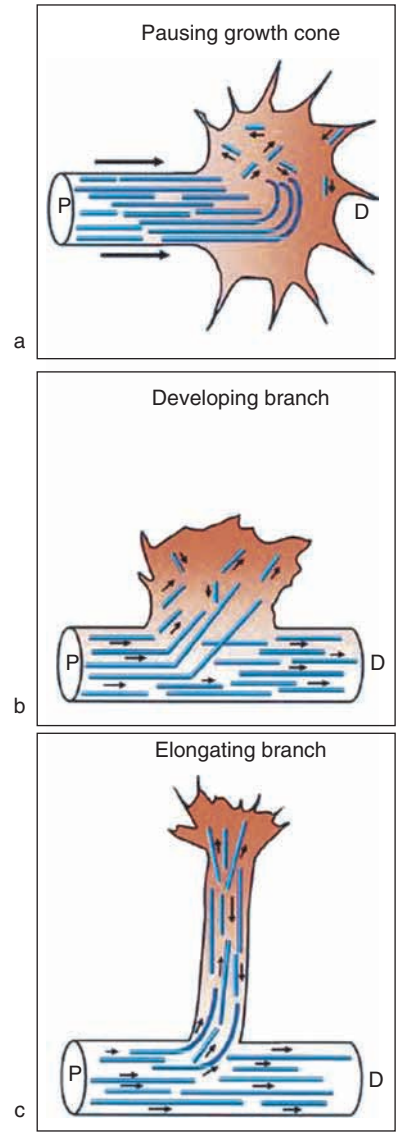

Figure 4 Schematic model for microtubule reorganization during axon outgrowth and branching. (a) Microtubules (blue) in the pausing growth cone form a loop in the central region. As microtubules splay apart in a region of the loop, short microtubules fragment from longer microtubules and explore the lamellipodium. (b) A branch is emerging from the axon shaft in a region where local splaying and fragmentation of microtubules occurs. Short microtubule fragments invade the developing branch. (c) The axon branch is extending as microtubules elongate within it. Arrows indicate directions of microtubule movement. P and D refer to proximal and distal segments of the axon.

Actin–Microtubule Interactions

Microtubules and actin filaments are highly dynamic, undergoing continual cycles of growth and shortening during polymerization and depolymerization. At the same time, actin filaments and microtubules can also move and undergo dramatic changes in their organization and location within the growth cone. In cortical neurons, branching from the growth cone and the axon shaft was preceded by splaying apart of looped or bundled microtubules accompanied by localized accumulation of F-actin. Dynamic microtubules co-localized with F-actin in transition regions of growth cones and axon branch points. Interactions between microtubules and actin filaments involved their coordinated polymerization and depolymerization. Drugs that attenuated either microtubule or actin dynamics abolished F-actin–microtubule interactions at the growth cone and at axon branch points, inhibiting axon branching without affecting axon length. Similar findings in the growth cones of invertebrates showed that actin–microtubule interactions are essential for initiating growth cone steering. These results imply that factors, such as extracellular cues that influence the dynamics of either F-actin or microtubules, may affect both cytoskeletal elements through bidirectional signaling pathways.

Effects of Growth and Guidance Molecules on Cytoskeletal Dynamics

Growth and guidance factors such as Sema3A, FGR-2, and netrin-1 regulate axon branching through changes in cytoskeletal dynamics. Sema3A applied to growth cones depolymerized bundled actin in filopodia and attenuated motility and protrusion of lamellipodia (Figure 5). Microtubules decreased their exploratory behaviors and collapsed centrally, suggesting that reduced actin dynamics attenuated microtubule dynamics. The failure of microtubules to invade the growth cone periphery and delineate new directions of growth decreased axon branching. FGF-2 which increased cortical axon branching and enlarged growth cones over time promoted actin polymerization and the formation of microtubule loops. Netrin-1 also increased actin polymerization causing microtubules to splay apart in the axon shaft and the growth cone, thereby permitting dynamic microtubules to interact with the newly polymerized actin bundles and initiate outgrowth in new directions. A recent study of KIF2A, a kinesin family member that interacts directly with microtubules by inhibiting their ability to elongate, showed the importance of cytoskeletal dynamics in regulating axon branching. When KIF2A was knocked out in mice, hippocampal axons had much longer collateral branches than in wild-type neurons. This suggests that KIF2A is required for suppressing supernumerary branches on hippocampal axons. In mice lacking KIF2A, microtubule depolymerizing activity was lower, suggesting that KIF2A normally depolymerizes microtubules and may suppress outgrowth of collaterals by regulating microtubule dynamics. Thus the absence of KIF2A would result in overextension of axon branches.

In neuronal growth cones and at axon branch points, the actin and microtubule cytoskeleton is the ultimate target of intracellular signaling pathways activated by extracellular guidance cues via cell surface receptors. Although these signaling pathways have been the subject of intense study, it is still not well understood how guidance cues modulate the dynamics of actin filaments and microtubules and regulate their interactions. However, it is known that, through the actions of cytoskeletal-associated regulatory proteins, reorganization of the cytoskeleton and changes in cytoskeletal dynamics underlie the ability of growth cones to turn toward gradients of attractive guidance cues (such as netrin-1) or turn away from repulsive cues (such as Sema3A). This model of growth cone guidance (Figure 6) could also apply to the promotion or inhibition of axon branching by growth and guidance cues.

Regulation of Axon Branching by Calcium Signaling

Axon Guidance and Branching

Growth and guidance cues are known to elicit calcium activity in neuronal growth cones. For example, in response to gradients of netrin-1, intracellular Ca^{2+} concentrations are elevated in the growth cone to regulate changes in guidance behaviors. Attractive turning is preceded by protrusion of lamellipodia and filopodia. These effects involve relative Ca^{2+} concentrations that form a gradient across the growth cone such that attractive turning may occur in the direction of higher Ca^{2+} concentrations (although growth cones can also be repelled toward higher Ca^{2+} levels). Rapid and extensive axon branching induced by netrin-1 was also found to occur through changes in Ca^{2+} activity in the form of repetitive transients. Netrin-1 evoked repetitive Ca^{2+} transients in restricted regions of cortical axons *in vitro*. Netrin-1 could also modulate endogenous Ca^{2+} activity by increasing the frequency of Ca^{2+} transients. These transients could be evoked in the entire axon or begin in localized axon regions and spread throughout the axon. Moreover, experimental reduction of Ca^{2+} levels in the culture medium revealed that Ca^{2+} signaling is essential for netrin-1-induced axon branching. Ca^{2+} transients often occurred in patterns, frequencies and amplitudes that were different in the axon and its branches, raising the possibility that Ca^{2+} transients of different frequencies and amplitudes could differentially regulate the growth of different branches on the same axon (Figure 7) or the growth of the primary axon and its branches. This is consistent with the finding in several studies that higher levels of electrical activity confer a competitive advantage on the growth of one axon branch over another. The role of netrin-1 in axon branching has not been demonstrated *in vivo*. However, it seems likely that similar calcium signaling mechanisms elicited by gradients of chemotropic target-derived cues such as netrin-1 may promote branching to topographically appropriate targets at specific locations along the axon.

Downstream Targets of Calcium Signaling during Axon Branching

Changes in the frequency of Ca^{2+} transients are known to activate different downstream intracellular targets. Different frequencies as well as patterns of Ca^{2+} transients may activate genes for different neurotransmitters and also target different kinases and phosphatases in the cytoplasm to elicit different

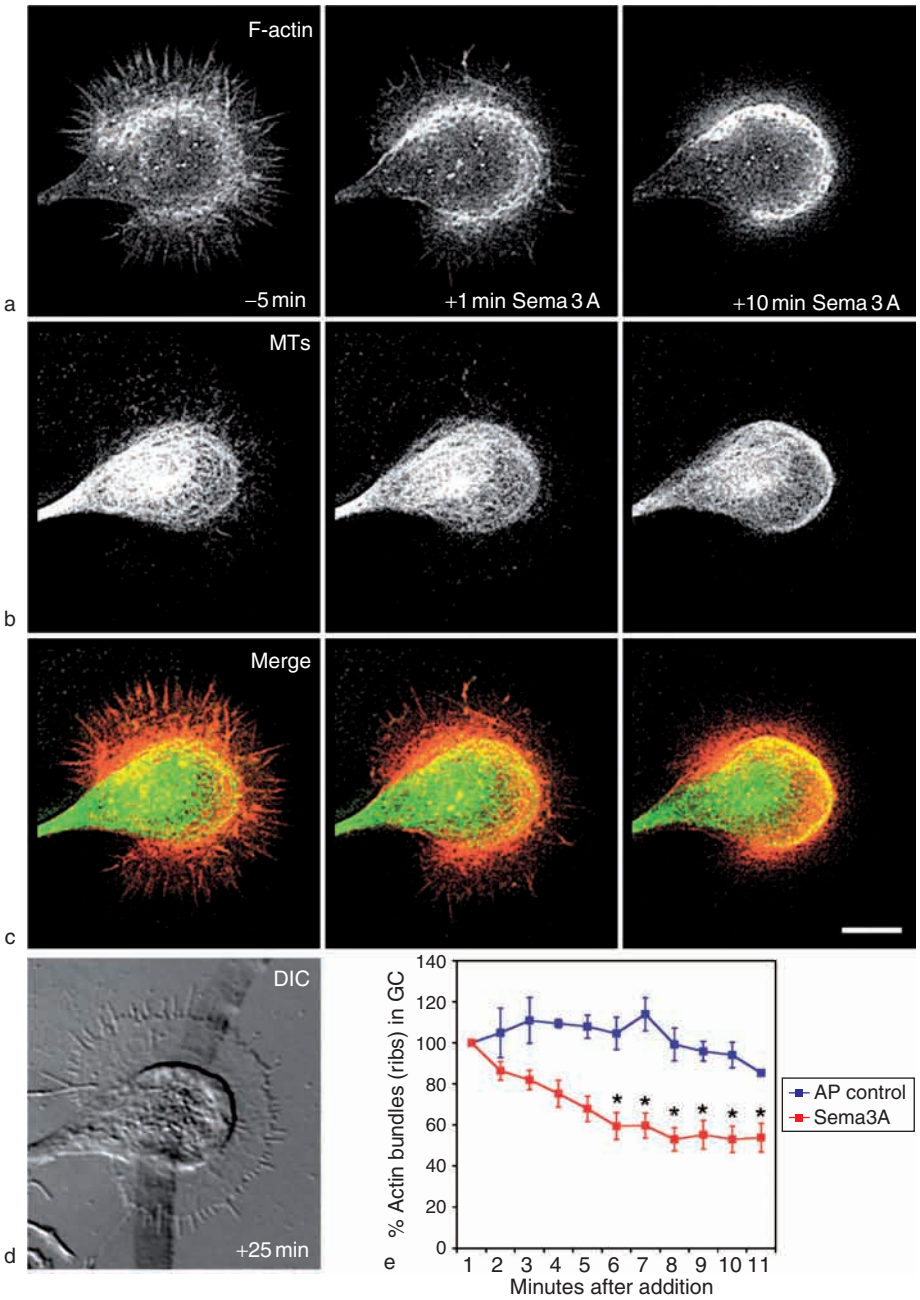

Figure 5 Sema3A rapidly decreases the number of F-actin bundles and dynamic microtubules in the peripheral region of the growth cone without collapse of the growth cone. (a, b) Images of a living large paused growth cone from a postnatal day 3 cortical neuron co-injected with Alexa488 phalloidin to label F-actin and with rhodamine tubulin to label microtubules (MTs). Within 1 min after application of Sema3A, many of the actin bundles in the growth cone disappear, concomitant with the loss of dynamic microtubules extending from the central microtubule loop. By 10 min after Sema3A application, the intensity of an F-actin ring surrounding the peripheral microtubule loop increases. (c) Merged images of the same growth cone in (a) and (b), where F-actin is pseudocolored red and microtubules are green. (d) A differential interference contrast (DIC) image of the same growth cone in (a)–(c) showing that the filopodia and broad lamellar region have not retracted. (e) Graph plotting the number of actin bundles over time. Semaphorin addition resulted in a significant decrease in actin bundles over time compared with supernatant collected from AP vector-infected cells (*$p < 0.05$; paired t test). All error bars are ±SEM. Scale = 10 μm.

cellular responses. Thus, specific cellular events may be regulated by calcium transients at optimal frequencies. For example, calcium/calmodulin-dependent protein kinase II (CaMKII) is known to function as a Ca^{2+} spike frequency detector. Mitogen-activated protein kinases (MAPKs) are also sensitive to intracellular Ca^{2+} changes. Ca^{2+} oscillations are self renewing, repetitive, and persistent. Moreover, the

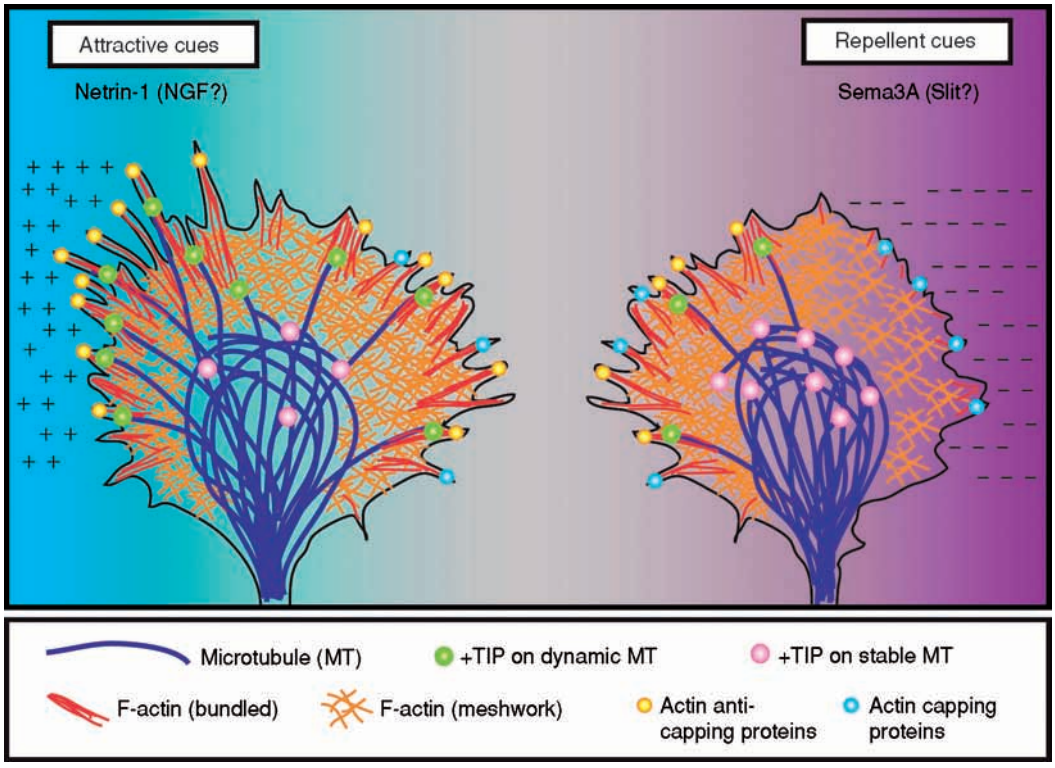

Figure 6 Model for cytoskeletal reorganization and dynamics in growth cones responding to attractive and repellent cues. The side of the growth cone facing away from guidance cues is the unstimulated side. The growth cone contains bundled stable microtubules in the center, a meshwork of actin filaments in the lamellipodium, actin filament bundles that penetrate filopodia, and dynamic microtubules that extend into the periphery to interact with actin filaments. The attractive cue netrin-1 increases the number of filopodia and promotes actin filament elongation by actions of anticapping proteins. (Capping proteins normally terminate actin filament elongation.) In the growth cone periphery, an increased number of actin filaments bundles promotes interaction with dynamic microtubules and their associated proteins (such as plus end tracking +TIP proteins that bind to and stabilize microtubules). These interactions could lead to attractive turning behavior. The repellent cue Sema3A causes a dissolution of actin filament bundles and the subsequent loss of dynamic microtubules, which leads to growth cone collapse and repulsive turning. Slit might induce similar microtubule looping leading to growth cone pausing and/or repulsion. This model is also applicable to the promotion or inhibition of axon branching by growth and guidance cues.

mechanism of frequency modulation to decode information from calcium signals has advantages over mechanisms based on changes in amplitude of Ca^{2+} concentration alone, because discrete Ca^{2+} oscillations can be more easily distinguished by calcium decoders such as CaMKII. CaMKII and MAPK, both of which are known to be influenced by frequencies of Ca^{2+} transients in functions related to memory and synaptic plasticity, were recently shown to be major components of the netrin-1 calcium signaling pathway essential for axon branching. Overexpression of CaMKII promoted cortical axon outgrowth as well as axon branching, whereas MAPK increased axon branching but not axon outgrowth. Activation of specific kinases could be a mechanism whereby netrin-1-induced Ca^{2+} signaling promotes axon branching independent of axon outgrowth. Accordingly, Ca^{2+} may activate different downstream targets to elicit opposing effects on the axon such as inhibiting axon outgrowth but promoting the development of new branches.

Effects of Calcium Signaling Pathways on the Cytoskeleton

Ultimately, signaling pathways activated by guidance cues must converge on the actin–microtubule cytoskeleton to induce changes in growth cone behaviors and axon branching. Axon branching requires dynamic actin–microtubule interactions and netrin-1 elicits filopodial protrusion through actin polymerization. Netrin-1 is known to activate Rho GTPases, which regulate actin polymerization, as well as activating MAP1B (microtubule-associated protein), which is thought to play a role in actin–microtubule interactions. Thus, netrin-1-induced calcium signaling could influence regulators of cytoskeletal dynamics required for axon branching. CaMKII is a major target of netrin-1-induced calcium signaling. Studies of regulation of dendritic morphology and plasticity suggest that CaMKII may be an important link to the cytoskeleton. In dendritic filopodia, resembling spines, αCaMKII was shown to interact with βCaMKII and

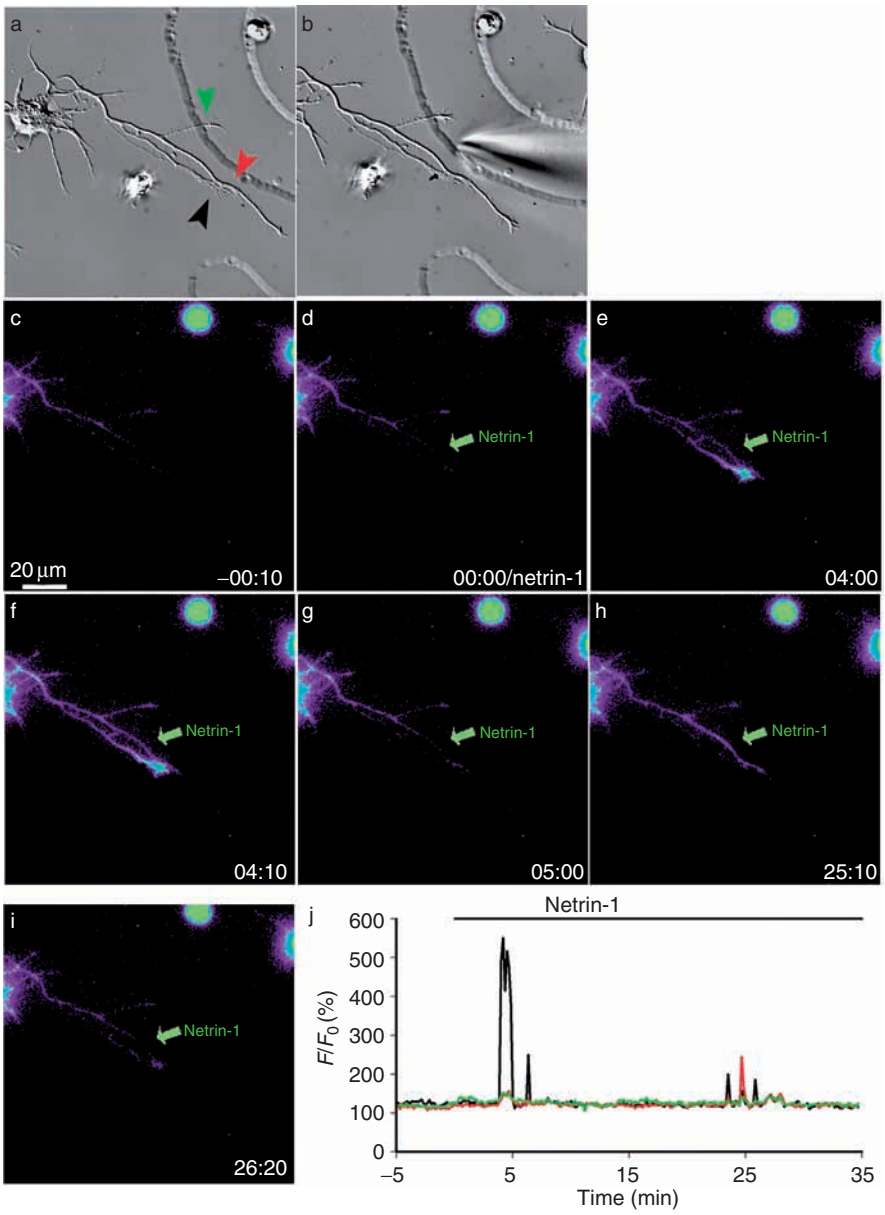

Figure 7 (a, b) DIC images and C-I time lapse images in pseudocolor showing changes in calcium activity in an axon (red arrowhead) and its two branches (green and black arrowheads) after local application of netrin-1 from the pipette shown in (b) and indicated by the green arrows in the fluorescence images. Localized calcium transients occur in one branch at 4 min (e), in another smaller branch at 4:10 (f), and in the primary axon at 25 min (h). (j) Measurements of calcium transients in the axon (red), a large branch (black), and a small branch (green) corresponding to the arrows in the DIC image in (a). Calcium transients induced by netrin-1 can occur at different frequencies and at different times in an axon and its branches.

translocate to dendritic spines, thereby targeting CaMKII to the actin cytoskeleton. This is consistent with the role of CaMKII in stabilizing dendritic arbors. In this way, repetitive calcium transients induced by netrin-1 could provide a mechanism for translocating CaMKII to actin filaments in nascent axonal filopodia. However, it is not completely understood how major targets of calcium signaling, such as CaMKII and MAPK, regulate cytoskeletal dynamics required for development of new branches.

Activity-Dependent Regulation of Axon Branching

The involvement of Ca^{2+} signaling, a measure of electrical activity, suggests a role for activity-dependent mechanisms in the regulation of axon branching. During development, electrical activity shapes the growth and branching of axon arbors and thus the formation of neural circuits. The importance of electrical activity in concert with molecular cues was

demonstrated in refining the retinoptopic map *in vivo*. In the absence of spontaneous retinal waves, the refinement of retinal ganglion axon arbors into a precise map in the superior colliculus was degraded. In newborn rat visual cortical slices, imaging of axons extending horizontally in layers 2/3 of the cortex combined with electrophysiological recording showed that the complexity of axon branching gradually increased over time with a concomitant dramatic increase in frequency of spontaneous firing. When spontaneous action potential firing was silenced with tetrodotoxin (TTX) or synaptic transmission was blocked with pharmacological manipulation of glutaminergic transmission, the numbers of axon branches fell to about half that of control slices, suggesting that neural activity plays a role in regulating the early development of branches on cortical neurons.

Electrical activity is known to shape axon arbors by regulating retraction or elimination of axon branches. Recently, direct observation has shown that activity also controls the extent and rate of branch formation. Imaging of retinal axon arbors *in vivo* in the developing optic tectum of zebra fish revealed how suppression of activity influences axon branching. Transfection of retinal ganglion cells individually with a specific type of potassium (K^+) channel made it possible to electrically silence the transfected cells by hyperpolarization. Overexpression of the inward rectifier K^+ channel was found to suppress calcium spiking associated with spontaneous electrical activity. In these neurons, arbor length and complexity were reduced in comparison with adjacent untransfected cells with normal electrical activity. These results show that competitive mechanisms based on relative activity influence the development of functional networks through growth of axon arbors. Importantly, the competition effect is expressed in the formation of new axon branches rather than on the stability of branches. Studies of axon branching *in vitro* also support a model of competitive growth of branches based on electrical activity. In cultures of sympathetic neurons in the presence of NGF, local depolarization of axon collaterals enhanced their growth whereas unstimulated inactive axon collaterals from the same or neighboring neurons were at a competitive disadvantage for growth. Depolarization by patterned electrical stimulation, resulting in pulses of increased intracellular calcium, mediated this effect through the activation of a CaMKII–MEK signaling pathway. These findings suggest a mechanism for selection of one axon branch over another through electrical activity that promotes the growth of one set of axon branches while inhibiting growth of unstimulated axon collaterals.

Conclusions

The growth of collateral axon branches into appropriate targets during development is essential for establishing CNS connectivity. Branches can arise along the length of the axon as interstitial collaterals as well as forming terminal arbors. Although pathfinding by axonal growth cones and branching from the axon shaft are independently regulated, both serve to steer axons in new directions toward their targets. Therefore it is not surprising that growth cone guidance and axon branching involve similar extracellular guidance cues, cytoskeletal dynamics, and intracellular calcium signaling pathways. The terminal growth cone and developing axon branches are often at long distances from each other and from the cell body. Many of the cytoskeletal and signaling events that cause the growth cone to turn or a branch to form are therefore highly localized. It is not well understood how gradients of target-derived cues elicit branching at correct locations or inhibit inappropriate branching along the axon to ensure appropriate target innvervation. The signaling events that transform the relatively static axon shaft to one of dynamic motile protrusions during axon branching await further discovery.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Regeneration: Role of Growth and Guidance Cues; Calcium and Signal Transduction; Growth Cones.

Further Reading

Bastmeyer M and O'Leary DD (1996) Dynamics of target recognition by interstitial axon branching along developing cortical axons. *Journal of Neuroscience* 16: 1450–1459.

Cohen-Cory S and Fraser SE (1995) Effects of brain-derived neurotrophic factor on optic axon branching and remodelling *in vivo*. *Nature* 378: 192–196.

Dent EW, Barnes AM, Tang F, and Kalil K (2004) Netrin-1 and semaphorin 3A promote or inhibit cortical axon branching, respectively, by reorganization of the cytoskeleton. *Journal of Neuroscience* 24: 3002–3012.

Dent EW, Callaway JL, Szebenyi G, Baas PW, and Kalil K (1999) Reorganization and movement of microtubules in axonal growth cones and developing interstitial branches. *Journal of Neuroscience* 19: 8894–8908.

Dent EW and Kalil K (2001) Axon branching requires interactions between dynamic microtubules and actin filaments. *Journal of Neuroscience* 21: 9757–9769.

Dickson BJ (2002) Molecular mechanisms of axon guidance. *Science* 298: 1959–1964.

Gomez TM and Zheng JQ (2006) The molecular basis for calcium-dependent axon pathfinding. *Nature Reviews Neuroscience* 7: 115–126.

Halloran MC and Kalil K (1994) Dynamic behaviors of growth cones extending in the corpus callosum of living cortical brain slices observed with video microscopy. *Journal of Neuroscience* 14: 2161–2177.

Homma N, Takei Y, Tanaka Y, et al. (2003) Kinesin superfamily protein 2A (KIF2A) functions in suppression of collateral branch extension. *Cell* 114: 229–239.

Hua JY, Smear MC, Baier H, and Smith SJ (2005) Regulation of axon growth *in vivo* by activity-based competition. *Nature* 434: 1022–1026.

Huber AB, Kolodkin AL, Ginty DD, and Cloutier JF (2003) Signaling at the growth cone: Ligand-receptor complexes and the control of axon growth and guidance. *Annual Review of Neuroscience* 26: 509–563.

Kalil K and Dent EW (2005) Touch and go: Guidance cues signal to the growth cone cytoskeleton. *Current Opinion in Neurobiology* 15: 521–526.

Katz LC and Shatz CJ (1996) Synaptic activity and the construction of cortical circuits. *Science* 274: 1133–1138.

O'Leary DD and Terashima T (1988) Cortical axons branch to multiple subcortical targets by interstitial axon budding: Implications for target recognition and 'waiting periods'. *Neuron* 1: 901–910.

Rodriguez OC, Schaefer AW, Mandato CA, Forscher P, Bement WM, and Waterman-Storer CM (2003) Conserved microtubule–actin interactions in cell movement and morphogenesis. *Nature Cell Biology* 5: 599–609.

Schaefer AW, Kabir N, and Forscher P (2002) Filopodia and actin arcs guide the assembly and transport of two populations of microtubules with unique dynamic parameters in neuronal growth cones. *Journal of Cell Biology* 158: 139–152.

Singh KK and Miller FD (2005) Activity regulates positive and negative neurotrophin-derived signals to determine axon competition. *Neuron* 45: 837–845.

Tang F and Kalil K (2005) Netrin-1 induces axon branching in developing cortical neurons by frequency-dependent calcium signaling pathways. *Journal of Neuroscience* 25: 6702–6715.

Uesaka N, Hirai S, Maruyama T, Ruthazer ES, and Yamamoto N (2005) Activity dependence of cortical axon branch formation: A morphological and electrophysiological study using organotypic slice cultures. *Journal of Neuroscience* 25: 1–9.

Yates PA, Roskies AL, McLaughlin T, and O'Leary DD (2001) Topographic-specific axon branching controlled by ephrin-As is the critical event in retinotectal map development. *Journal of Neuroscience* 21: 8548–8563.

Intracellular Calcium and Neuronal Death

M Gleichmann and M P Mattson, National Institute on Aging Intramural Research Program, Baltimore, MD, USA

Published by Elsevier Ltd.

Subcellular Electrochemical Ca^{2+} Gradients

The physiological extracellular Ca^{2+} concentration is 1.3 mM, whereas the resting cytosolic Ca^{2+} concentration in neurons ranges from 50 to 200 nM. In order to understand the importance of Ca^{2+} in neuronal death and survival, one has to know the molecular mechanisms that are involved in the regulation of cellular Ca^{2+} homeostasis, the most basic of which are briefly outlined in the following section.

Calcium ATPases and Sodium Calcium Exchangers

Plasma membrane Ca^{2+} ATPases (PMCA 1–4) pump Ca^{2+} out of the cytoplasm into the extracellular space at the expense of ATP. In addition, sodium calcium exchangers (NCX 1–3) use the Na^+ gradient to extrude Ca^{2+} from the cytosol to the extracellular space by in return allowing Na^+ into the cytosol (forward mode). Under conditions of high intracellular Ca^{2+}/low extracellular Na^+, they operate in reverse mode (i.e., pump out Na^+ for Ca^{2+}). As the extracellular/cytosolic Na^+ gradient is the main driving force for the NCX, its function is tightly coupled to other proteins that affect this gradient, like sodium–potassium ATPase and sodium-proton exchanger.

Voltage-Gated and Ligand-Gated Calcium Channels

Ca^{2+} can enter from the extracellular space into the cytoplasm via voltage-gated Ca^{2+} channels (L-, P-, Q-, R-, N-, and T-type Ca^{2+} channels) and ligand-gated cation channels such as glutamate receptors. Glutamate receptors can be separated (according to their selectivity for glutamate analogs) into *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), and kainate receptors. NMDA receptors are heterotetramers of an NR1-subtype dimer and another dimer composed of any combination of the other subtypes, NR2A-D and NR3A-B. They have a high permeability for Ca^{2+} and a low permeability for Na^+, and in neurons NMDA receptor activation will cause more Ca^{2+} entry than activation of voltage-gated Ca^{2+} channels. AMPA receptors are heterotetramers of subtypes GluR1–4 and generally have high conductance for Na^+ and low conductance for Ca^{2+} unless they do not contain subtype GluR2. GluR2 mRNA editing will result in an exchange of glutamine to arginine and this Q/R change will cause a block for Ca^{2+} ions. Kainate receptors are heterotetramers of subtypes GluR5–7 and/or KA 1–2 and, like AMPA receptors, have high permeability for Na^+ and low permeability for Ca^{2+}.

Endoplasmic Reticulum

The second main source for Ca^{2+} influx into the cytosol is the endoplasmic reticulum (ER), which constitutes a tubular organellar system throughout the neuron including dendrites, axons, and synapses. Ca^{2+} concentrations in the ER are variable (and may vary within subregions of the ER) but generally range from 0.5 to 1 mM. The ER membrane, in contrast to the plasma membrane, has considerable Ca^{2+} leakage, which is also a source of Ca^{2+} influx into the cytosol. Ca^{2+} is pumped into the ER by sarcoplasmic/endoplasmic reticulum ATPases (SERCA 1–3), which are activated by low Ca^{2+} concentrations in the ER. NCXs are not known to play a role in ER Ca^{2+} homeostasis. Ca^{2+} release occurs via ryanodine receptors (RRs) and inositol-3-phosphate receptors (IP_3Rs). Both RRs and IP_3Rs are regulated by Ca^{2+} itself in a way that they are activated by cytosolic Ca^{2+} increases (with the IP_3R1 isoform being inactivated at higher Ca^{2+} concentrations). This feedforward mechanism is known as calcium-induced calcium release and is believed to give the ER excitable and regenerative properties, similar to the plasma membrane, which may allow for Ca^{2+} release to spread throughout the ER within a cell. Notably, ER-associated Ca^{2+} oscillations can be found in neurons.

Mitochondria

Mitochondria are not believed to be a major Ca^{2+} source under physiological conditions. However, their proton gradient constitutes an electrochemical driving force for cations and, in situations of high cytosolic Ca^{2+} concentrations, mitochondria accumulate Ca^{2+}, thus acting as Ca^{2+} buffers. They are also closely connected to ER Ca^{2+} release sites and are believed to remove Ca^{2+} from those local sites.

Calcium as a Second Messenger in Physiological Conditions

The steep extracellular/cytosolic gradient makes Ca^{2+} a useful second messenger, especially in excitable cells. Depolarization-induced Ca^{2+} entry causes synaptic vesicle release at the synapse making Ca^{2+}

the integrating mechanism that controls neuronal signal propagation. In the cytosol, a variety of Ca^{2+}-binding proteins such as calbindin and calmodulin act as calcium sensors and via interaction with other proteins such as calcium/calmodulin kinases (CamK's) and/or the calcineurin phosphatase affect phosphorylation status of numerous other proteins (e.g., transcription factors, ion channels, other kinases). A basic feature of cytosolic Ca^{2+}-binding proteins is that they usually have a low K_D, thus also acting as Ca^{2+} buffers (calpains being an exception). Calcium-binding and buffering proteins also provide a rationale for the concept of different Ca^{2+} concentrations within the cytosol. In the vicinity of voltage-gated Ca^{2+} channels, the Ca^{2+} concentrations are believed to increase up to 100 μM under physiological conditions while calcium-binding proteins buffer Ca^{2+} at further distances.

CREB

One well-established Ca^{2+}-associated signaling cascade involves calmodulin kinase (CaMK) and cAMP response element binding protein (CREB). CREB is an important transcription factor for neuronal differentiation, survival, and long-term potentiation (LTP)-mediated synaptic plasticity. It is activated upon phosphorylation at Ser 133 by CaMKII and IV. The nuclear Ca^{2+} determines CREB phosphorylation status and transcriptional activity much more than cytosolic Ca^{2+} making somal Ca^{2+} rises more likely to affect CREB activity than Ca^{2+} in distant dendrites. The nucleus is surrounded by the nuclear envelope, from which the ER is believed to be produced. Therefore, Ca^{2+} at distant synapses could cause ER-mediated Ca^{2+} release, which in certain cases might spread from dendritic ER region all the way to the nuclear envelope, causing an increase in nuclear Ca^{2+} and thus nuclear CREB phosphorylation.

CREB activity is closely associated with neuronal activity and neuronal survival during development. During development neurons are generated in excess, and the early postnatal period of rodents is a phase where dispensable neurons are eliminated by apoptosis. During this developmental period, synaptic activity is the key determinant of a neuron's survival (as neurons without synaptic activity are unlikely to be important for CNS function, this survival criterion appears useful in the context of brain function). A very well established *in vitro* model of this paradigm is the potassium deprivation-induced cell death of cerebellar granule neurons, where granule neurons are isolated from the cerebella of 8-day-old rat pups and cultivated under moderately depolarizing conditions with elevated potassium. A switch to normal potassium concentrations will induce cell death in these granule neurons. In this model of developmental activity-dependent neuronal survival, it has been shown *in vitro* and *in vivo* by conditional knockout and use of dominant negative mutants that CaMKIV and (phospho)CREB activity are essential in conveying the survival message. However, if CREB is knocked out at later stages of life, rodents develop a neurodegenerative phenotype where neuronal loss in hippocampus and striatum happens at a slower pace than during CREB knockout in development. Interestingly, CREB has also been linked to rodent models of trinucleotide repeat diseases such as Huntington's disease in which a decreased expression of the transcriptional cofactor CREB binding protein (CBP) has been observed, which would result in decreased CREB transcriptional activity.

While CREB plays an important role in neuronal survival, Ca^{2+} and CaMK are not the only factors phosphorylating CREB. Activation of the Ras/extracellular receptor kinase (ERK) or the PI_3K/Akt cascades results in CREB phosphorylation as well. On the other hand, most signals that will activate ERK or Akt are closely linked to Ca^{2+}, although Ca^{2+} may not be required itself for activation. One example for this is NMDA receptor-mediated activation of ERK. The exact molecular mechanism of NMDA receptor-mediated ERK activation is not fully understood, but there is some evidence that it is not Ca^{2+} itself that is required, but rather molecular interactions and conformational changes mediated by other proteins that also bind to the scaffold proteins at the postsynaptic density, such as the postsynaptic density protein of 95 kDa (PSD95), via its so-called postsynaptic density and zonula occludens (PDZ) domains. This also offers an explanation for observed differences in synaptic versus nonsynaptic NMDA receptor-mediated signaling, with only the former activating ERK.

BDNF

Another activator of neuronal survival signaling, ERK and Akt phosphorylation, and Ca^{2+} is brain-derived neurotrophic factor (BDNF). BDNF is a neurotrophic factor expressed throughout the CNS and, just like CREB, its function is important for neuronal differentiation, survival, and synaptic plasticity. Moreover, CREB and BDNF are interconnected as BDNF signaling will result in CREB-mediated transcriptional activation and CREB activation will increase transcription from the BDNF gene. BDNF binding to its high-affinity receptor TrkB will cause receptor dimerization and autophosphorylation. The ensuing signaling cascade involves activation of Ras and ERK,

PI_3K and Akt, as well as phospholipase C (PLC). PLC activation will result in the generation of IP_3, which in turn will open IP_3Rs in the ER membrane causing Ca^{2+} release. Therefore, similar to NMDA receptors, BDNF/TrkB signaling is closely associated with Ca^{2+}, although not all downstream cascades strictly require Ca^{2+} for activation.

Neuronal Activity and Ca^{2+} Sustain Trophic Signaling Cascades

From the data gathered so far on Ca^{2+}, CREB, CaMK, and BDNF, a picture emerges with considerable feedforward potential. Thus, glutamate release will induce Ca^{2+} influx via NMDA receptors, leading to CaMKII+IV/CREB activation, ERK and Akt phosphorylation, increased production of BDNF, and via Ca^{2+}-mediated vesicle release BDNF secretion. BDNF in turn activates ERK + Akt and induces ER Ca^{2+} release, thus enhancing synaptic strength (and LTP) and further facilitating glutamate signaling. Glutamate will also activate the metabotropic glutamate receptors which results in PLC activation and ER Ca^{2+} release. In such a scenario, Ca^{2+} would sustain active synapses and overall para- and autocrine release of neurotrophic factors. It provides the rationale for a broader concept which sees neuronal activity as a key factor in neuronal survival and Ca^{2+} plays a central role in this concept. This concept is also reflected in less scientific slogans of brain function such as "brain jogging" or "use it or lose it". However, as appealing as these ideas may be, one has to keep in mind that apart from the developmental situation, such theories are not settled knowledge and more evidence is required to convincingly make the case for neuronal activity and Ca^{2+} as a neuroprotective strategy in aging and neurodegenerative disorders.

Ca^{2+} in Pathological Conditions

As with most aspects of biology, there is a flip side to the Ca^{2+} coin as well. Ca^{2+} can have detrimental effects on neurons under pathological conditions, and this is closely (though not exclusively) linked to the process of excitotoxicity. Excitotoxicity describes toxic effects of high intensity and/or sustained activation of excitatory neurotransmitter receptors on neurons.

Glutamate Excitotoxicity

In cell culture concentrations of 100 μM, glutamate will cause massive cell death. Because glutamate treatment only for a few minutes with subsequent glutamate washout (or glutamate receptor blockade) and switch to normal culture conditions will cause no prominent morphological changes in phase-contrast microscopy but later on will result in massive neuronal death, the term 'delayed glutamate toxicity' was coined, which may be somewhat misleading as glutamate excitotoxicity (both *in vitro* and *in vivo*) actually is a fast process.

NMDA receptors and Ca^{2+} are key mediators of glutamate toxicity The finding of glutamate excitotoxicity was followed by the discovery that in the absence of Ca^{2+} glutamate causes considerable less cell death, whereas in the absence of Na^+ it remains highly toxic, leading to the idea that high and sudden increases in cytosolic Ca^{2+} cause neuronal death, whereas a cytosolic Na^+ increase poses less of a problem. With the development of specific receptor antagonists it was shown that NMDA receptor blockade prevents glutamate excitotoxicity, whereas AMPA and kainate receptor blockade proves to be less effective. These findings still remain undisputed basic knowledge on glutamate excitotoxicity and make a lot of sense in the light of the ion-regulating physiology of neurons, since Ca^{2+} efflux has to proceed against a higher electrochemical gradient than Na^+ efflux and NMDA receptors are the main source for glutamate-induced Ca^{2+} entry. However, the downstream events in glutamate excitotoxicity are less clear.

Source specificity theory of glutamate toxicity Although the evidence for Ca^{2+} as a key mediator of glutamate excitotoxicity is compelling, a paradox remains. Cultured neurons can be depolarized by means of elevated extracellular potassium concentrations to the extent that cytosolic Ca^{2+} concentrations and total cellular Ca^{2+} uptake are equivalent to the amounts seen in glutamate toxicity, yet such elevated potassium concentrations do not cause cell death. This observation has led to the formulation of the source specificity theory of glutamate toxicity. It claims that not the total amount of Ca^{2+} entering the cell but rather the source of entry determines neuronal fate. However, elevated potassium causes Ca^{2+} influx only, whereas glutamate increases Na^+ and Ca^{2+}. This has not been fully appreciated in the source specificity theory.

Nitric oxide and reactive oxygen species Calcium influx through NMDA receptor channels is coupled to activation of neuronal nitric oxide (NO) synthase (nNOS), which is located close to synaptic NMDA receptors by binding to PSD95 via PDZ domains. Activation of nNOS might lead to excessive NO and reactive oxygen species (ROS) formation, thus

contributing to glutamate toxicity and potentially explaining the source specificity of glutamate excitotoxicity. Moreover, increased NO and ROS generation after glutamate treatment have been documented by many groups. However, although it is generally accepted that ROS formation contributes to glutamate toxicity, it is more controversial as to how much ROS are at the core of glutamate toxicity.

Mitochondrial Ca^{2+} overload Once glutamate receptors and voltage-gated calcium channels open and Ca^{2+} enters the cytosol in large amounts, the Ca^{2+} will enter the mitochondrial matrix due to the mitochondrial membrane potential which constitutes an electrochemical driving force for cations. In the mitochondrial matrix, Ca^{2+} will form a complex with phosphate ($CaPO_4$) so that the concentration of free Ca^{2+} remains constant and more Ca^{2+} can enter the mitochondria. In this way, mitochondria act as a sink for Ca^{2+} and their total Ca^{2+} uptake exceeds that of the free Ca^{2+} concentration both in cytosol and mitochondrial matrix by orders of magnitude. Mitochondrial Ca^{2+} uptake may continue to the point where $CaPO_4$ will precipitate from a gel-like to a crystalline form, causing the mitochondrial permeability transition pore to open. However, the latter mechanism does not explain the source specificity of excitotoxicity, and it is unclear to what extent such a mechanism applies to all types of neurons.

Calpain activation and Ca^{2+} efflux Another concept of glutamate toxicity focuses on calpain activation. Calpains are calcium-activated papain-like proteases. Unlike most other cytosolic calcium-binding proteins, calpains have K_D's in the micromolar and millimolar range (μ-calpain and m-calpain), concentrations which under physiological conditions are unlikely to occur anywhere other than in the close proximity to voltage-gated calcium channels or glutamate receptors, especially since endogenous inhibitors, calpastatins, prevent uncontrolled calpain activation. However, in cases of massive glutamate release, calpain activation increases far beyond levels observed under physiological conditions, where nontoxic NMDA receptor activation and processes such as neurite outgrowth are associated with limited calpain activation. Importantly, calpain inhibition decreases glutamate toxicity. Calpain activation leads to cleavage of the NCX3 isoform, thus impeding Ca^{2+} efflux and causing cell death. This concept puts Ca^{2+} efflux into the equation, an important aspect that so far has not received much attention.

The complexities of glutamate excitotoxicity The concepts outlined above represent only some of the glutamate excitotoxicity mechanisms, and so far no generally accepted chronological series of downstream events has been established in glutamate excitotoxicity. However, it is important to stress that the existing concepts are not mutually exclusive and it is quite possible that each concept covers part but not all of the story. For example, it is conceivable that calpain activation is specifically linked to glutamate receptor opening (if one assumes that the local Ca^{2+} concentrations in their vicinity are higher than at the calcium channels), thus linking the source specificity with calpain activation. Also, it is possible that death pathways may be multiple and redundant, such that some neurons die from mitochondrial Ca^{2+} overload while others die from calpain overactivation. In the latter scenario, inhibiting one pathway (e.g., calpain activation) might rescue only a small population of neurons, while other neurons undergo a different route toward cell death (e.g., mitochondrial Ca^{2+} overload). If this is the case, then downstream rescue strategies would require combination approaches, an area that so far has not been systematically tested. Interestingly, such a conversion from one cell death pathway to another exists between apoptosis and necrosis in glutamate excitotoxicity such that apoptosis inhibition will result in necrosis.

Ca^{2+}, Glutamate, and Stroke

Although glutamate excitotoxicity in its purest form is an *in vitro* model, it is believed to be of importance in many pathological *in vivo* situations, the most prominent of which is stroke. In stroke, neurons undergo energy depletion due to ischemia and this will result in collapse of the membrane potential and release of large amounts of glutamate, contributing to cell death in the so-called penumbra region where blood supply is not completely absent but critically reduced. Glutamate concentrations in stroke areas have been reported to reach levels of 100 μM, and NMDA receptor antagonists are protective in rodent stroke models. Also, many downstream events observed in glutamate excitotoxicity are found in rodent stroke models (and postmortem human brain tissue) including increased NO and ROS generation, calpain activation, and (so far only in rodent models) NCX3 cleavage. Similarly, the *in vitro* model of combined hypoxia and hypoglycemia shows a broad overlap with glutamate excitotoxicity when done in neurons. Unfortunately, NMDA receptor antagonists do not work in stroke therapy in humans, the problem being that they cause acute psychotic episodes, preventing the drug from being administered in the required dose.

So, while direct NMDA receptor antagonists are regarded as a dead end road in stroke treatment,

the jury is still out on treatment strategies that target more downstream events of glutamate excitotoxicity. However, targeting downstream events is no easy task because of the existence of multiple pathways that will likely require polypharmacy approaches.

Glutamate Excitotoxicity and Other Neurodegenerative Diseases

Although glutamate excitotoxicity is most closely linked to stroke, it is also believed to play a role in many other diseases. Importantly, certain features of glutamate excitotoxicity, including cellular calcium overload, calpain activation, and ROS generation, are observed in many different neurodegenerative disorders. However, it is not clear whether such changes reflect late events taking place after neurons have passed the point of no return, or actually are early pivotal events. While hippocampal damage in chronic epilepsy is a case where glutamate release (due to epileptic discharge) is located at the beginning of events (and NMDA blockade is helpful in rodent models), in other diseases like Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), HIV-associated dementia (HAD), or the trinucleotide repeat diseases the connection is not that obvious. One concept that involves glutamate excitotoxicity and Ca^{2+} deregulation and can be applied to a number of diseases centers around a vicious cycle of increased ROS generation, protein damage, mitochondrial dysfunction, reduced ATP generation, impaired Ca^{2+} homeostasis, and increased (synergistic) glutamate toxicity. In this scenario, it is conceivable that the disease process starts slow and progresses to a point where otherwise nontoxic glutamate levels cause cell death. More disease-specific mechanisms have been found that link glutamate excitotoxicity and neuronal death. For example, in the case of AD, a synergistic *in vitro* toxicity of amyloid beta and glutamate exists, as well as a specific impairment of the NCX (but not the sodium/potassium ATPase). Also, the NMDA receptor open channel blocker memantine is approved for AD therapy and may actually slow down the disease process. Memantine has also been approved for the treatment of HAD, and may also be a viable approach for the treatment of PD.

In ALS, glutamate excitotoxicity was at the center of research for a long time, although this changed to some degree with the discovery that mutations in the SOD1 gene cause some cases of inherited ALS. Nevertheless, glutamate levels in the cerebrospinal fluid (CSF) of ALS patients are increased and the only drug that has been approved so far in ALS, riluzol, has (among other properties) agonist properties for the excitatory amino acid transporter 2 (EAAT2) protein, which is expressed on glial cells and is supposed to scavenge excess glutamate from the extrasynaptic space. In addition, there appears to be a specific loss of expression of EAAT2 expression in ALS.

Calculating the Balance: Is Ca^{2+} Good or Bad?

Taken together, there exist two seemingly contradictory concepts, one of which stresses the role of Ca^{2+} and depolarization in sustaining neuronal function, the other pointing to Ca^{2+} as the key factor in excitotoxicity and associated neurodegenerative diseases. Both concepts have accumulated considerable data for their support, so one might ask if Ca^{2+} is preventing or causing neuronal death. To formulate the question in this way probably is not useful because one has to separate between early development and older age. Moreover, Ca^{2+} has multiple functions, the net effect of which likely depends on the molecular context of its signaling. For example, it seems that massive Ca^{2+} entry via NMDA receptors is more harmful than via voltage-gated channels, but the status of Ca^{2+} efflux capability and NCX function also have to be taken into account. On the other hand, NMDA receptors are more likely to cause ERK phosphorylation when located synaptically and nuclear Ca^{2+} appears to be particularly helpful in sustaining CREB transcriptional activity. If these concepts hold true, then simple approaches like activating or antagonizing NMDA receptors should not be the most effective approach. Instead, a combination of NMDA receptor antagonist plus a neurotrophic factor like BDNF would make more sense, as the neurotrophic factor might compensate for loss of signaling through ERK and Akt kinases. Similarly, one might consider combining partial NMDA or AMPA receptor agonists with antioxidants and think about ways to specifically target synaptical NMDA receptors. Future research will be required to establish whether such calcium-based approaches are effective.

See also: Calcium Channels; Excitotoxicity in Neurodegenerative Disease; Glutamate; Stroke; Stroke: Injury Mechanisms; Voltage-Gated Calcium Channels.

Further Reading

Aarts MM and Tymianski M (2004) Molecular mechanisms underlying specificity of excitotoxic signaling in neurons. *Current Molecular Medicine* 4: 137–147.

Arundine M and Tymianski M (2003) Molecular mechanisms of calcium-dependent neurodegeneration in excitotoxicity. *Cell Calcium* 34: 325–337.

Arundine M and Tymianski M (2004) Molecular mechanisms of glutamate-dependent neurodegeneration in ischemia and

traumatic brain injury. *Cellular and Molecular Life Sciences* 61: 657–668.

Choi DW (1992) Excitotoxic cell death. *Journal of Neurobiology* 23: 1261–1276.

Hardingham GE and Bading H (2003) The yin and yang of NMDA receptor signalling. *Trends in Neurosciences* 26: 81–89.

Mantamadiotis T, Lemberger T, Bleckmann, et al. (2002) Disruption of CREB function in brain leads to neurodegeneration. *Nature Genetics* 31: 47–54.

Mattson MP (2003) Excitotoxic and excitoprotective mechanisms: Abundant targets for the prevention and treatment of neurodegenerative disorders. *Neuromolecular Medicine* 3: 65–94.

Mattson MP and Chan SL (2003) Neuronal and glial calcium signaling in Alzheimer's disease. *Cell Calcium* 34: 385–397.

Nicholls DG (2005) Mitochondria and calcium signaling. *Cell Calcium* 38: 311–317.

Nicotera P and Melino G (2004) Regulation of the apoptosis-necrosis switch. *Oncogene* 23: 2757–2765.

Orrenius S, Zhivotovsky B, and Nicotera P (2003) Regulation of cell death: The calcium-apoptosis link. *Nature Reviews Molecular Cell Biology* 4: 552–565.

Vanhoutte P and Bading H (2003) Opposing roles of synaptic and extrasynaptic NMDA receptors in neuronal calcium signalling and BDNF gene regulation. *Current Opinion in Neurobiology* 13: 366–371.

Verkhratsky A (2005) Physiology and pathophysiology of the calcium store in the endoplasmic reticulum of neurons. *Physiological Reviews* 85: 201–279.

Yu SP, Canzoniero LM, and Choi DW (2001) Ion homeostasis and apoptosis. *Current Opinion in Cell Biology* 13: 405–411.

Zipfel GJ, Babcock DJ, Lee JM, and Choi DW (2000) Neuronal apoptosis after CNS injury: The roles of glutamate and calcium. *Journal of Neurotrauma* 17: 857–869.

Invertebrate Models to Study Learning and Memory: Lymnaea

P R Benjamin and G Kemenes, University of Sussex, Brighton, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Associative conditioning studies are predominant in the *Lymnaea* model system. Integration between behavioral, neural circuitry, and molecular biological approaches are key features of the *Lymnaea* work. Learning occurs within more complex circuits involving central pattern generators (CPGs) rather than the reflexive circuits often used in other mollusks. An attractive feature of the *Lymnaea* system is the ability to carry out one-trial reward conditioning. Mechanistic studies have focused on feeding and respiration building on the detailed information of the neural circuitries involved.

The Feeding Circuit

Feeding in *Lymnaea* consists of a sequence of repetitive movements called rasps (Figure 1(a)). During each rasp, the mouth opens and a toothed radula is scraped forward over the food substrate (protraction phase). Food is then lifted into the mouth (retraction phase), which closes while the food is being swallowed (swallow phase), and the sequence is repeated. Rhythmic movements of the feeding muscles are driven by a network of motor neurons (B1–B10) that, in turn, are driven by synaptic inputs from a feeding CPG network of interneurons. Each phase of the feeding rhythm is generated by one of three main types of CPG interneurons, N1 (protraction), N2 (retraction), N3 (swallow), providing sequences of excitatory and inhibitory synaptic inputs to motor neurons active in different phases of the feeding rhythm. CPG-driven rhythmic electrical activity can be recorded in the feeding network even in the absence of feeding muscles and this is called fictive feeding. Activity in the motor neurons and CPG neurons is modulated by identified higher-order interneurons, such as the cerebral giant cells (CGCs) and cerebral ventral 1 cells (CV1s). These higher-order neurons have been the focus of the learning and memory studies. The CGCs act as gating neurons in the feeding circuit. Increased CGC spiking activity during feeding facilitates feeding responses to food. The CV1 cells are members of a larger population of neurons called the cerebrobuccal interneurons (CBIs) that are command-like neurons involved in the activation of feeding.

Sucrose is an effective chemical stimulus for feeding and is therefore used as the unconditioned stimulus (US) in associative (Pavlovian) reward conditioning. At the cellular level, sucrose applied to the lips in semi-intact preparations induces fictive feeding in motor neurons and interneurons of the feeding network. As well as the CPG neurons, the modulatory CGC and CV1 cells are also activated in the sucrose-driven fictive feeding rhythm.

The conditioned stimulus (CS) for reward conditioning in *Lymnaea* is either a chemical (amyl acetate, pear drops) or a tactile stimulus (a gentle brush stroke applied to the lips). Recently, visual stimuli have also been used and the success of a black-and-white checkered pattern as a CS indicates that the eyes are involved. The chemical CS was previously thought to have no effect on feeding ('neutral stimulus'), either at the behavioral or electrophysiological levels, but has recently been shown to have stimulatory or inhibitory effects in naive animals depending on concentration. Touch to the lips, monitored either at the behavioral or the electrophysiological level, cannot initiate or maintain feeding but nevertheless produces a complex sequence of inhibitory and excitatory synaptic inputs on all neurons of the feeding network.

The Respiratory Network

The aquatic pulmonate snail *Lymnaea* can breathe either through the skin or through a simple lung (Figure 1(b)). As the name implies, *Lymnaea stagnalis* populations often live in stagnant water, and when the environment becomes hypoxic the snails float to the surface and perform rhythmic opening and closing movements of their pulmonary opening, the pneumostome. This aerial respiration behavior is controlled by a respiratory CPG, the three main components of which are the right pedal dorsal 1 (RPeD1), input 3 (IP3), and visceral dorsal 4 (VD4) interneurons. These provide synaptic inputs to identified motor neurons (I and J, opener; and K, closer) innervating pneumostome opener and closer muscles. The chemosensory stimulus (hypoxia) that triggers pneumostome opening first activates sensory cells in the pneumostome-osphradial area, which in turn provide excitatory afferent inputs to RPeD1. Through its synaptic connections with the other members of the CPG network, activation of RPeD1 initiates CPG activity which underlies the respiratory rhythm.

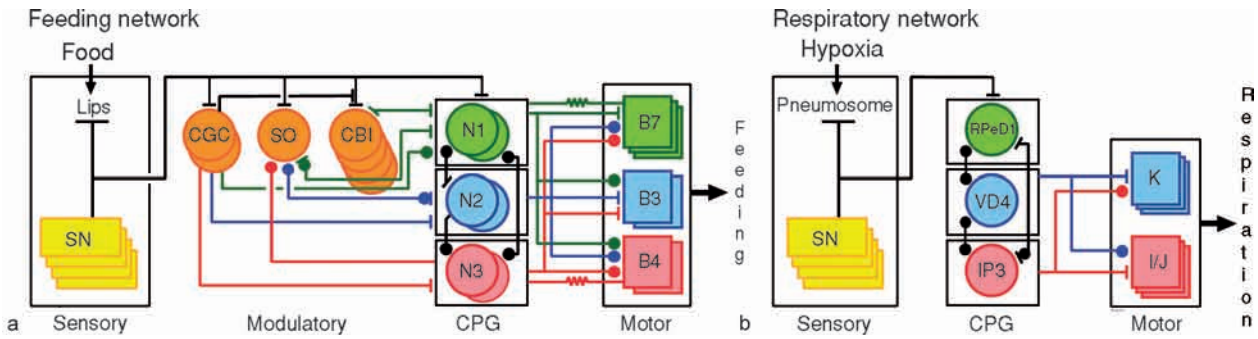

Figure 1 The neuronal networks for feeding and respiration in *Lymnaea*. (a) The neuronal network underlying rhythmic feeding behavior. Chemosensory neurons (SNs, yellow squares) located in the lip structures detect the presence of food or chemostimulants, such as sucrose. Excitatory inputs from the sensory pathways are distributed in parallel to modulatory neurons (orange circles), including the paired cerebral giant cells (CGCs), the single slow oscillator (SO) interneuron and the group of cerebral-buccal interneurons (CBIs), and also to N1 type CPG interneurons (green circles). The CGCs may presynaptically modulate certain types of chemosensory inputs to the CBIs (see Figure 4). The complex synaptic connectivity between the modulatory and N1 interneurons leads to activation of the whole CPG, with the protraction (N1), rasp (N2, blue circles), and swallow (N3, pink circles) phases of the feeding cycle following in sequence due to the synaptic connectivity within the CPG and their intrinsic properties. Several subtypes of each of the N cells have been characterized but because of the complexity of their synaptic connections with modulatory interneurons and motor neurons, only a generalized representative of each N-type is shown. The rhythmic pattern generated by the CPG drives protraction, rasp, and swallow phase motor neurons, such as B7 (green squares), B3 (blue squares), and B4 (pink squares), leading to sequences of muscular activity and feeding movements. (b) The neuronal network underlying respiratory behavior. Activation of sensory neurons (SNs, yellow squares) located in the pneumostome-osphradial area, leads to activation of the respiratory CPG, comprised of RPeD1 (green circle), VD4 (blue circle), and IP3 (pink circle). VD4 drives activity in K MNs (motor neurons, blue squares), responsible for pneumostome closing (inspiration) and IP3 drives activity in I/J MNs (pink squares), responsible for pneumostome opening (expiration). Filled circle, chemical inhibitory synapses; vertical bar, chemical excitatory synapses; horizontal bar, peripheral sensory endings; zig-zag lines, electronic connections.

Tactile stimulation of the pneumostome area evokes pneumostome closure and stops aerial respiratory behavior. This response, which is mediated by an identified whole-body withdrawal neuron (RPeD11), is an important component of the withdrawal response in *Lymnaea*, and therefore tactile stimulation of the pneumostome is used as an aversive US in the operant conditioning experiments.

Reward Classical Conditioning of Feeding to a Tactile CS

Lymnaea can be classically conditioned by repeatedly pairing touch to the lip with food (5–15 trials) (Figure 2(a)). This type of learning shares important characteristics with associative conditioning in vertebrates, such as stimulus generalization and discriminative learning, classical–operant interactions and strong dependence on both external and internal background variables, such as novelty versus familiarity of the environment or hunger versus satiety.

Two approaches have been used to investigate the neural basis of tactile reward classical conditioning in *Lymnaea*. One approach was based on the development of an *in vitro* preparation where electrophysiological manipulation of neuronal pathways aims to mimic the behavioral conditioning paradigm. In this study, a lip touch stimulus was paired with intracellular activation of the modulatory SO neuron, which can drive fictive feeding. After 6–10 pairings, presentation of the touch stimulus could activate a robust fictive feeding rhythm in feeding motor neurons.

A second approach used behavioral conditioning followed by electrophysiological analysis to record changes in electrical activity that follow long-term memory (LTM) formation. In the semi-intact preparation, electrophysiologically recorded fictive feeding patterns were used as a systems-level correlate of the repetitive feeding behavior seen in the intact snail (Figure 2(b1)). A feeding motor neuron, such as the B3 (Figure 1(a)), was routinely used as a monitor of fictive feeding in the whole network (Figure 2(b2)). This cell is easy to identify, plays no role in pattern generation, but receives synaptic inputs from all three types of CPG interneurons and so can be used as an indirect monitor of the conditioned activation of the CPG. It also receives an easy-to-identify early compound excitatory postsynaptic potential (EPSP) input that results from stimulation of the CS pathway. Other motor neurons such as the B1 salivary gland motor neuron were also recorded (Figures 2(b2) and 2(b3)) to check that all the feeding motor neurons are activated during conditioned fictive feeding. Most recently, recordings were made from modulatory interneurons such as the CV1 cells (Figures 2(b2) and 2(b3)) to see if they also undergo changes following conditioning.

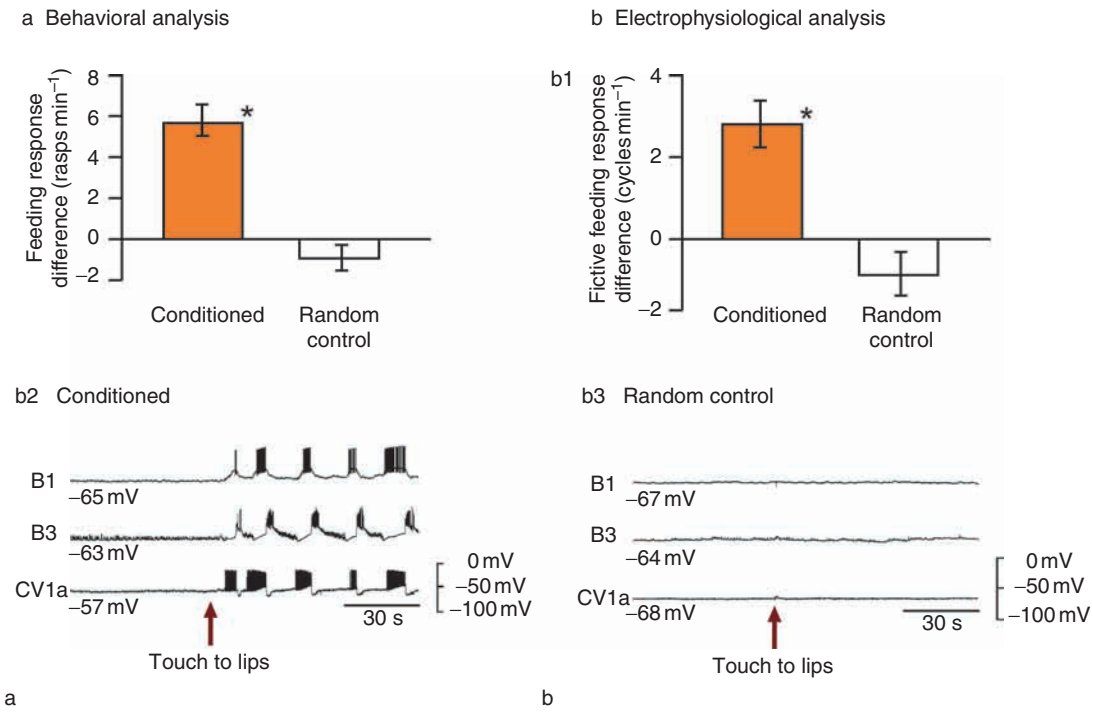

Figure 2 Electrophysiological correlates of the LTM trace after classical conditioning in *Lymnaea*. (a) The behavioral response to the touch CS, which was calculated as the postminus prestimulus difference in feeding activity, is significantly (*) stronger in the conditioned versus the control group. (b) Behavioral classical conditioning produces a CS-induced increase in the systems-level electrical activity (fictive feeding response). (b1) The electrophysiological response to the touch CS, which was calculated as the postminus prestimulus difference in fictive feeding activity, is significantly (*) stronger in the conditioned versus the control group. (b2) The lip touch CS evokes rhythmic fictive feeding activity in a CBI, CV1 (CV1a subtype) and motor neurons B1 and B3 in a preparation from a conditioned animal. (b3) The lip touch CS evokes no rhythmic fictive feeding activity in a preparation from a control animal. Note that CV1a is more depolarized in the preparation from the conditioned animal than in the preparation from the control animal, a consistent observation in this type of experiment.

Using the lip touch behavioral training protocol snails were subjected to 15 training trials over 3 days and then these and control animals were dissected for electrophysiological analysis, starting on the day after the last training trial. Touching the lips of the intact snails from the experimental group after training induced a pattern of feeding movements significantly greater than controls (Figure 2(a)). Similar significant differences (Figure 2(b1)) were seen between experimental (Figure 2(b2)) and control animals (Figure 2(b3)) at the level of the electrophysiologically recorded fictive feeding pattern in motor neurons, recorded from the same snails.

The CPG-driven activity in the motor neurons depends ultimately on activity of neurons at all levels of the feeding network, so the conditioned fictive feeding recorded in the motor neurons is a systems-level 'readout' of the memory trace in the whole feeding system. However, more detailed changes can also be recorded in different parts of the network. One of these is the early EPSP that occurs in the B3 motor neuron before the onset of the fictive feeding pattern. The amplitude, but not the latency and duration of the EPSP, was significantly enhanced after conditioning. In sated snails, the conditioned fictive feeding response to touch was lost but the increase in the EPSP amplitude persisted. This suggests that there is unlikely to be a causal link between increases in amplitude in B3 and generation of the fictive feeding pattern.

Electrical correlates of tactile conditioning were recorded also at other levels within the feeding circuit and these could all be potential sites of plasticity. That sites, quite early in the CS pathway, could be involved in conditioning was revealed by extracellularly recording mechanosensory fibers located in the connective between the cerebral and buccal ganglia. Tactile responses could be recorded in these fibers and, following conditioning, the number of spikes occurring early in this response increased, compared with controls.

Interestingly, a correlate of tactile conditioning could also be recorded in the CPG network. A long-lasting sequence of inhibitory synaptic inputs that occurs in the N1 CPG interneurons in response to lip touch in naive animals (Figure 3(b)) changes to a strong sustained depolarizing synaptic input after conditioning (Figure 3(a2)), and this drives a sustained

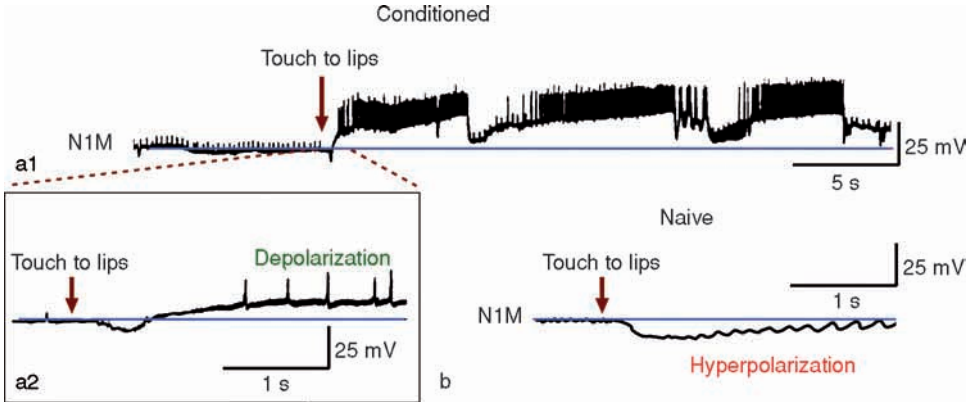

Figure 3 Learning-induced modification of mechanosensory inputs to a CPG interneuron of the *Lymnaea* feeding system. (a1) After behavioral classical conditioning with a tactile CS (touch to the lips) and food US (sucrose), the CS induces rhythmic activity in the CPG interneuron N1M, which is known to lead to fictive feeding activity in the entire feeding network. (a2) The CS-evoked inputs on an expanded time base. A brief hyperpolarization is followed by a strong maintained depolarization. (b) Synaptic inputs to the N1M in a preparation from a naive animal, evoked by the same touch stimulus that was used as the CS in the experiment shown in (a2). The tactile inputs have a strong maintained hyperpolarizing effect on N1M, a consistent observation in naive preparations.

plateauing pattern in the N1 cell (Figure 3(a1)). This is an example of synaptic plasticity affecting an important interneuronal component of the feeding network.

One candidate for initiating CPG activity following conditioning is the CBI cell type known as CV1. This neuron is capable of driving a fictive feeding pattern via its connections with the N1M cells of the CPG network, and activity in these cells normally accompanies unconditioned feeding patterns stimulated by sucrose. After conditioning, the CV1 cells are significantly more active following touch in conditioned snails (Figure 2(b2)) compared with controls (Figure 2(b3)), and they show the typical patterned activity seen with sucrose (US) application. More detailed experiments on the role of CV1 cells in tactile conditioning have revealed that nonsynaptic electrical changes play a role in memory. A long-lasting membrane depolarization of 11 mV on average was recorded in CV1s from conditioned compared with control snails that persisted for as long as the electrophysiological and behavioral memory trace. The depolarization makes the cells more responsive to the CS and can account for the activation of the feeding response after conditioning via the CV1 cell's strong excitatory synaptic connection with the CPG. The importance of this result is emphasized by experiments in which the membrane potential of the CV1 cells is manipulated to either reverse the effect of behavioral conditioning or mimic the effects of conditioning in naive snails. These experiments showed that the persistent depolarization of the CV1 cells was both sufficient and necessary for the conditioned response in the feeding network.

Reward Classical Conditioning of Feeding to a Chemical CS

In the original successful formulation of chemical reward conditioning in *Lymnaea*, snails were subjected to a chemical conditioning protocol using amyl acetate as the CS and sucrose as the US. Following training, the explicitly paired (CS–US) experimental group showed significantly greater feeding responses to amyl acetate over their own naive responses and all the standard control groups (random, explicitly unpaired, CS alone, US alone). As might be predicted for appetitive conditioning, both age and motivational state (hunger vs satiety) influenced learning. Both hungry and sated young snails could acquire the conditioned response, but in the latter group its expression was only apparent when the animals were starved before testing. On the other hand, old snails could only acquire the conditioned response if they were maintained in a hungry state during training. The significance of motivational state became even more apparent when it was realized that if snails were starved long enough (for 5 days) before and throughout the experiment, even a single pairing of amyl acetate and sucrose resulted in LTM which lasted for at least 19 days. This is a remarkable example of single-trial learning that is used now for analyses of the time course of the mechanisms underlying LTM formation.

An electrophysiological correlate of the conditioned response to amyl acetate was recorded as changes in the fictive feeding responses in motor neurons, as in tactile conditioning, but electrical activity following conditioning has also been recorded in other parts of

the feeding system in attempts to localize sites of plasticity. The cell bodies of chemosensory neurons are located in lip epithelial tissue and project to the cerebral ganglia via the lip nerves where they synapse with cerebral ganglion neurons like the CBIs. Extracellularly recorded spike responses to both the CS and US can be recorded in the lip nerves from naive animals and these responses do not change after conditioning. In contrast, neuronal output from the cerebral ganglia is significantly enhanced in response to the CS after conditioning. This indicates that chemical conditioning affects central but not peripheral processing of chemosensory information, with the cerebral ganglia being an important site of plasticity. The fibers that were recorded extracellularly to indicate cerebral plasticity were originating from the CBI interneurons, so their activation is particularly significant. Confirmation that the CBIs do increase their activity after conditioning was obtained by showing increases in feeding patterns to the CS in CV1 neurons, a specific CBI cell type. From this work, it is suggested that the synapses between the primary chemosensory neurons and the CBIs are a major site of plasticity. Two other CS pathways are present in naive animals but these are not affected by conditioning.

The current network model for chemical conditioning (Figure 4) also includes nonsynaptic plasticity. The CGCs are persistently depolarized by about 10 mV after behavioral conditioning. This indirectly increases the strength of postsynaptic responses to CGC stimulation by a process that involves an increase in intracellular calcium concentration. The local target for CGC depolarization are the CBI cells, and artificial depolarization of the CGCs in naive snails increases the response of the CBI cells to the CS, mimicking the effects of behavioral conditioning. It appears that the CGCs are increasing the strength of the CS to CBI synapse by presynaptic facilitation, as shown in Figure 4. The onset of the CGC depolarization is between 16–24 h after training and it persists for at least 14 days, as long as the behavioral memory trace was present. There is an early behavioral memory trace from 2 h after conditioning, so the CGCs cannot be involved in memory expression immediately after training. It is more likely that the CGCs are involved in the maintenance of the LTM after the trace has already been consolidated and encode information that is important for memory recall. Interestingly, the CV1 cells that show a persistent change in membrane potential after tactile conditioning show no change after chemical conditioning.

At the molecular level, we know that the protein synthesis-blocker anisomycin (ANI) injected 10 min (but not at 1, 2, 3, 4, 5, and 6 h) after conditioning blocks the 24 h memory trace and an earlier 5 h memory trace is blocked by both protein synthesis and RNA synthesis blockers (Actinomycin D, Act-D), confirming that LTM is present as early as 5 h after conditioning. Regulation of gene expression during memory consolidation is known to involve transcription factors like the cyclic AMP-responsive element-binding protein (CREB) and the highly conserved CREB gene and CREB-like proteins have been identified recently in *Lymnaea*. Consistent with a role for CREB in *Lymnaea* LTM is the observation that levels of phosphorylated CREB are increased in the CGCs following reward conditioning.

Consolidation of the 24 h memory trace is dependent on the nitric oxide (NO)-GMP signaling pathway. There is a critical period of sensitivity up to 5 h after conditioning when blocking this pathway by drug injection prevents behavioral LTM formation. Further evidence for a role of NO in LTM came from experiments on the CGCs. These neurons express mRNA transcripts from two related nNOS genes (*Lym-nNOS1* and *Lym-nNOS2*). Six hours after one-trial conditioning, *Lym-nNOS1* is upregulated

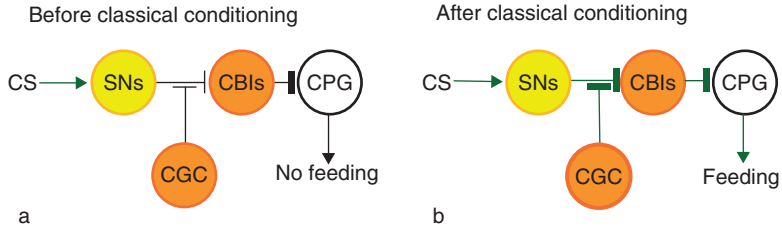

Figure 4 A cellular model of LTM based on the results from chemical classical conditioning in *Lymnaea*. (a) In naive animals, the excitatory connections between the sensory neurons (SNs) and the command-like CBIs are weak and the chemical CS (amyl acetate) cannot activate feeding. The CGCs are at normal membrane potential (~65 mV), and the presynaptic modulatory input from the CGCs to the SNs is inactive or weak. (b) In conditioned animals, the CGC soma is depolarized by ~10 mV compared to naive animals, and this leads to an enhancement of the CGC presynaptic modulatory inputs to SNs and a strengthening of the SN to CBI excitatory inputs, which enables the feeding response to the CS. Black lines, inactive connections; green lines, active connections. Thin bars, weak excitatory synapses; thicker bars, stronger/enhanced excitatory synapses. The thicker outline of the CGC in (b) indicates learning-induced nonsynaptic plasticity (membrane depolarization).

compared with controls. This upregulation of the NOS coding transcript may be due to an earlier downregulation of the *Lymnaea anti-NOS* transcript at 4 h that is known to be inhibitory on NOS transcript production.

Other highly conserved molecular pathways that have been implicated in LTM in *Lymnaea* are the protein kinase A (PKA) and mitogen-activated protein kinase (MAPK) signaling pathways. Inhibition of PKA catalytic subunit activity or MAPK phosphorylation without blocking sensory or motor pathways blocks LTM and 30 min after conditioning levels of both PKA activity and MAPK phosphorylation were increased, with PKA activation first detected in an even earlier time window (5 min). Whether there is any linkage between the NO, PKA, and MAPK signaling pathways is yet to be examined.

An interesting new observation is that memory reconsolidation after retrieval at 6 h posttraining is both PKA and protein synthesis dependent, whereas reconsolidation after retrieval at 24 h depends on protein synthesis but not on PKA activity. This finding indicates that depending on the 'age' of the consolidated memory, different molecular pathways are activated by memory retrieval and contribute differentially to memory reconsolidation.

Reward Classical Conditioning of Feeding to a Visual CS

Feeding also can be conditioned by pairing visual stimuli (CS) with sucrose (US). *Lymnaea* has a lens that is capable of forming an image on the retina under water and visual approach and reaching behavior can be evoked by a black-and-white check pattern. One-trial conditioning is possible when snails approaching a black panel along a gutter are stimulated with sucrose. After conditioning, approaching the black pattern elicited more rasping movements compared with un-reinforced trials. Successful reward conditioning was also obtained with a checkered pattern that was discriminated from a gray pattern of equal luminance, but in this experiment multiple trials were necessary (four trials per day for 4 days). A memory trace for this visual stimulus was recorded 4 days after conditioning.

Aversive Classical Conditioning of Feeding to a Chemical CS

Aversive classical conditioning in *Lymnaea* (conditioned taste aversion or CTA) is based on pairing sucrose with an aversive stimulus such as KCl, which inhibits feeding and evokes a withdrawal response. After eight trials, trained animals show a significantly weaker feeding response to sucrose than did controls and this memory lasted for over 1 month.

A neural analysis of CTA was carried out on isolated brains dissected from conditioned and control animals. In particular, the synaptic connection between the modulatory CGCs and the CPG interneuron, N1, was examined. In conditioned animals compared to controls, a significant increase in the size of IPSPs recorded in the N1 was observed following an artificial depolarization of the CGCs. Since the CGCs are known to play a critical gating role in feeding behavior and the N1 is a pivotal member of the feeding CPG, this enhanced IPSP may be an important cellular correlate of the CTA learning. Like reward chemical conditioning, aversive chemical conditioning appears to involve PKA and the CREB transcriptional regulatory system. The injection of cAMP or PKA into the soma of the CGCs led to a long-term enhancement of the synapse between the CGC and a follower motor neuron, B1. The injection of a CRE oligonucleotide into the CGCs 1 h prior to the injection of cAMP blocked the expression of long-lasting synaptic plasticity. CTA also has a requirement for the transcription factor, CCAAT/enhancer binding protein (C/EBP) and this was shown in the B2 gut motor neuron. Levels of phosphorylated LymC/EBP increased in B2 after CTA consolidation although paradoxically transcriptional levels were reduced in the same neuron.

Aversive Classical Conditioning of Whole-Body Withdrawal Responses to a Visual CS

Lymnaea do not normally respond to a light flash (CS) but after pairing with body rotation (US) the CS evokes a whole-body withdrawal response. This type of learning requires an interaction between two sensory organs, the eyes and the vestibular organs (the statocysts), and so this type of learning is called 'visuo-vestibular conditioning' or VVC. Spaced training, at least 10 trials over 3 days, is required, and it results in an LTM lasting for at least 1 week. The latency for a withdrawal response following the CS is measured and the reciprocal of the latency is used as the behavioral index for VVC and compared with response in controls where either random application of the CS and US are applied or naive animals receive no stimulus. Neural analysis has so far concentrated on analysis of sensory processing of the CS and US in the eye and statocyst, respectively, and their interactions in naive animals and this work forms the background for future mechanistic studies on LTM formation.

Operant Conditioning of the Respiratory Network to a Tactile CS

The aerial respiratory behavior of *Lymnaea stagnalis* has been used in a series of experiments investigating the behavioral and neuronal mechanisms of operant conditioning. Hypoxia triggers pneumostome opening and this was used as the operantly conditioned behavior. Tactile stimulation of the pneumostome area evokes pneumostome closure and stops aerial respiratory behavior. Animals were tested and trained in an artificially created hypoxic N_2-rich environment to increase the level of respiratory behavior. In the operantly trained group, a tactile stimulus was applied to the pneumostome area each time aerial respiration was attempted by the animal. Suitable yoked and hypoxic control groups were also used. The number of openings, latency to first opening, and total breath durations were recorded in pre- and posttraining periods. Only the operantly conditioned group showed significant changes between the pre- and posttraining behaviors, with significant reductions in openings and total breathing time and significant increases in the latency to first breath. It has since been demonstrated that a memory for this conditioned response could persist for at least 4 weeks when a spaced training procedure was used. Both intermediate memory (ITM) and LTM have been described based partly on the length of time the memory persists and also on the sensitivity to protein and mRNA synthesis blockers. ANI prevents the formation of both ITM and LTM, whereas Act-D only prevents LTM. Both reconsolidation and extinction have been studied following operant conditioning and both have been shown to be dependent on new RNA and protein synthesis. Extinction is viewed as a new type of associative memory that 'covers up' but does not replace the original memory. Thus, following extinction trials, the loss of memory at 2 h is followed by full 'spontaneous recovery' at 24 h.

Neural changes associated with this learned behavior have been identified in the isolated CNS derived from operantly conditioned animals. Specifically, spontaneous patterned activity in the IP3 interneuron, which is involved in pneumostome opening (see **Figure 1(b)**), showed a significant reduction compared to activity in the IP3 neurons of brains derived from yoked controls. Furthermore, a higher percentage of RPeD1 CPG interneurons, which are important in the onset of the respiratory cycle, were silent in conditioned versus control preparations. A reduction in the ability of the RPeD1 cells to induce IP3 activity was also observed. More direct evidence for the role of RPeD1 in memory formation came from somal ablation experiments. Removal of the soma 2 h prior to conditioning prevented LTM but had no effect on ITM, without affecting the ability of the snail to carry out respiratory behavior; this suggested that the RPeD1 soma was necessary for LTM formation. Removal of the RPeD1 soma 1 h after conditioning had no effect on LTM, indicating that the effects of soma ablation were not related to memory access or retrieval. Interestingly, both extinction and reconsolidation also require the presence of the soma of RPeD1, indicating that RPeD1 is involved in the formation of more than one type of memory trace.

Conclusions

Lymnaea can be trained to respond to a wide variety of sensory cues: visual, touch, and chemical. This pond snail is a generalist feeder, meaning that it is not restricted to any particular food and it would be an advantage for it to learn about potentially useful foods using several different modalities. Visual stimuli are provided by floating plants that the animal feeds on by locomoting to the water surface. Tactile cues potentially provide information about physical properties such as texture. Touch is known to provide a normal component of the food stimulus to the lips although touch alone is ineffective in initiating feeding in naive animals. Chemical cues provided by food are also useful indicators of nutritionally valuable material objects and presumably this is why they are such an effective CS. Both visual and chemical cues can be used in single-trial reward conditioning providing a useful model for studying the neuronal and molecular temporal changes occurring during and in the immediate period after training.

All the types of associative conditioning described here are capable of inducing LTM. ITM has been described for operant conditioning but has not yet been examined in other paradigms. Reconsolidation and extinction in operant conditioning have been shown to have similar general molecular requirements to LTM formation. More detailed molecular studies of LTM in chemical conditioning have emphasized that regulation of gene expression by transcription factors such as CREB and C/EBP and the PKA and MAPK signaling pathways. The importance of these molecular pathways in *Lymnaea* provides further evidence for the generality of these highly conserved mechanisms in molluskan learning. NO has been shown to be important in memory consolidation in chemical reward conditioning and this appears to involve regulation of nNOS gene expression in a specific cell type, the CGCs.

Neuronal analysis has emphasized that changes following learning can be recorded at multiple sites in the feeding and respiratory circuits. How

molecular and electrical changes in specific cells like RPeD1 in the respiratory circuit and the CGCs in the feeding circuit can translate to circuit level and behavioral changes is still a major question. This can be partly understood in the feeding circuit by recent results showing that persistent changes in membrane potential in the CGCs induced by conditioning can facilitate CS sensory synaptic pathways to command neurons for feeding. However, this is unlikely to be the only mechanism because other pathways originating in the buccal ganglia CPG are also likely to be important. The involvement of both synaptic and nonsynaptic electrical changes in reward conditioning of feeding is proving to be a major feature of LTM formation in other invertebrate and vertebrate models of learning and memory.

See also: Conditioning: Theories; Learning and Memory in Invertebrates: *Aplysia*; Learning and Memory in Invertebrates: *C. Elegans*; Learning and Memory in Invertebrates: *Drosophila*; Learning and Memory in Invertebrates: *Hermissenda*; Learning and Memory in Invertebrates: Honey Bee; Learning and Memory in Invertebrates: *Limax*; Learning and Memory in Invertebrate Models: Tritonia; Procedural Learning: Classical Conditioning.

Further Reading

Andrew RJ and Savage H (2000) Appetitive learning using visual conditioned stimuli in the pond snail, *Lymnaea*. *Neurobiology of Learning and Memory* 73: 258–273.

Benjamin PR, Staras K, and Kemenes G (2000) A systems approach to the cellular analysis of associative learning in the pond snail *Lymnaea*. *Learning and Memory* 7: 124–131.

Fulton D, Kemenes I, Andrew RJ, and Benjamin PR (2005) A single time window for protein synthesis-dependent long-term memory formation after one-trial appetitive conditioning. *European Journal of Neuroscience* 21: 1347–1358.

Hatakeyama D, Sadamoto H, Watanabe A, et al. (2006) Requirement of new protein synthesis of a transcription factor for memory consolidation; paradoxical changes in mRNA and protein levels of C/EBP. *Journal of Molecular Biology* 356: 569–577.

Ito E, Koboyashi S, Kojima S, Sadamoto H, and Hatakeyama D (1999) Associative learning in the pond snail, *Lymnaea stagnalis*. *Zoological Science* 16: 711–723.

Jones NG, Kemenes I, Kemenes G, and Benjamin PR (2003) A persistent cellular change in a single modulatory neuron contributes to associative long-term memory. *Current Biology* 13: 1064–1069.

Kemenes G, Kemenes I, Michel M, Papp A, and Müller U (2006) Phase-dependent molecular requirements for memory reconsolidation: Differential roles for protein synthesis and PKA activity. *Journal of Neuroscience* 26: 6298–6302.

Kemenes I, Kemenes G, Andrew RJ, Benjamin PR, and O'Shea M (2002) Critical time-window for NO-cGMP-dependent long-term memory formation after one-trial appetitive conditioning. *Journal of Neuroscience* 22: 1414–1425.

Kemenes I, Straub V, Nikitin ES, et al. (2006) Role of delayed nonsynaptic neuronal plasticity in long-term associative memory. *Current Biology* 16: 1269–1279.

Kojima S, Nanakamura H, Fujito Y, and Ito E (1997) Enhancement of an inhibitory input to the feeding central pattern generator in *Lymnaea stagnalis* during conditioned taste aversion. *Neuroscience Letters* 230: 179–182.

Korneev SA, Straub V, Kemenes I, et al. (2005) Timed and targeted differential regulation of nitric oxide synthase (NOS) and anti-NOS genes by reward conditioning leading to long-term memory formation. *Journal of Neuroscience* 25: 1188–1192.

Lukowiak K, Sangha S, Scheibenstock A, et al. (2003) A molluscan model system in the search for the engram. *Journal of Physiology (Paris)* 97: 69–76.

Ribeiro MJ, Schofield MG, Kemenes I, O'Shea M, Kemenes G, and Benjamin PR (2005) Activation of MAPK is necessary for long-term memory consolidation following food-reward conditioning. *Learning and Memory* 12: 538–545.

Ribeiro MJ, Serfozo Z, Papp A, et al. (2003) Cyclic AMP response element-binding (CREB)-like proteins in a molluscan brain: Cellular localization and learning-induced phosphorylation. *European Journal of Neuroscience* 18: 1223–1234.

Sakakibara M, Kawai R, Koboyashi S, and Horikoshi T (1998) Associative learning of visual and vestibular stimuli in *Lymnaea*. *Neurobiology of Learning and Memory* 69: 1–12.

Sangha S, Scheibenstock A, and Lukowiak K (2003) Reconsolidation of a long-term memory in *Lymnaea* requires new protein and RNA synthesis and the soma of right pedal dorsal 1. *Journal of Neuroscience* 23: 8034–8040.

Sangha S, Scheibenstock A, Morrow R, and Lukowiak K (2003) Extinction requires new RNA and protein synthesis and the soma of the cell right pedal dorsal 1 in *Lymnaea stagnalis*. *Journal of Neuroscience* 23: 9842–9851.

Straub VA, Kemenes I, O'Shea M, and Benjamin PR (2006) Associative memory stored by functional novel pathways rather than modifications of pre-existing neuronal pathways. *Journal of Neuroscience* 26: 4139–4146.

Straub VA, Styles BJ, Ireland JS, O'Shea M, and Benjamin PR (2004) Central location of plasticity involved in appetitive conditioning in *Lymnaea*. *Learning and Memory* 11: 787–793.

Invertebrate Neurohormone GPCRs

C J P Grimmelikhuijzen, G Cazzamali, M Williamson, M Schneider, and F Hauser, University of Copenhagen, Copenhagen, Denmark

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The invertebrate nervous systems use, in principle, the same classes of compounds for chemical neurotransmission as the vertebrate nervous systems do: amino acids, acetylcholine, adenosine triphosphate (ATP), monoamines, neuropeptides, and protein hormones. Whereas the first three compounds or compound classes are mainly used for fast neurotransmission, the last three are used for slow transmission, neuromodulation, and paracrine and endocrine signaling. This slow transmission is made possible by the relatively high metabolic stability of monoamines, neuropeptides, and protein hormones. Furthermore, it is intuitively clear that lower invertebrates (Figure 1(a)) are not fast movers. Many cnidarians, for example, obtained their names in the eighteenth century from flowers (e.g., sea anemone, sea pansy, and Anthozoa) because they were colorful and apparently immobile. Thus, it comes as no surprise that lower invertebrates mainly use biogenic amines, neuropeptides, and protein hormones as their neurotransmitters or neurohormones.

In this article, we confine ourselves to the receptors for biogenic amines, neuropeptides, and protein hormones. We call these three compound classes neurohormones because, in addition to their role as neurotransmitters in the central and peripheral nervous systems, these molecules mainly function as neurohormones, being released into the circulation and mediating the control of the nervous system on overall physiological processes such as development, reproduction, feeding, and homeostasis. The receptors for the three classes of neurohormones are nearly exclusively G-protein-coupled receptors (GPCRs), which are transmembrane proteins, consisting of seven transmembrane α-helices, an extracellular N-terminus, and an intracellular C-terminus (Figure 2). Ligand binding occurs at the extracellular parts of the receptor, whereas the intracellular parts transduce the receptor activation to the intracellular G-proteins, leading to second-messenger cascades. There exist several classes or families of GPCRs: family-A (also called rhodopsin family, or family-1), family-B (family-2), and family-C (family-3). Most neurohormone GPCRs belong to family-A or -B. The GPCR families differ by certain amino acid consensus sequences; for example, the family-A GPCRs have a characteristic DRY motif (or similar sequences such as the DRF or ERY sequences) in their second intracellular loop just after the third transmembrane α-helix (Figure 2).

Most animals belong to one of two evolutionary lineages: Protostomia and Deuterostomia (Figure 1(a)). Mammals, vertebrates, and some invertebrate groups (e.g., sea urchins) belong to the deuterostomian lineage, whereas most invertebrates (among them insects) belong to the protostomian lineage. Some very basal invertebrate groups such as cnidarians (e.g., *Hydra*, sea anemones, and corals) evolved before the split of Protostomia and Deuterostomia (Figure 1(a)).

If we consider the number of animal species living on Earth, it is evident that most (90%) of all animals are invertebrates (Figure 1(b)). Only 1–2% of all animal species are mammals, although it is in these species that most neuroscience research is being carried out. Therefore, addressing our attention to the invertebrates is clearly justified.

Seventy-five percent of all animals are insects. Thus, we mainly focus in this article on insect GPCRs and then relate these data to findings from some other invertebrate groups, such as mollusks, nematodes, and cnidarians (Figure 1).

Insect GPCRs

Insects are ecologically and economically extremely important because most flowering plants (and also food plants) depend on insects for their pollination. But insects can also be agricultural pests and vectors (intermediate pathogen carriers) for major diseases, such as malaria, elephantiasis, sleeping disease, Chagas disease, dengue fever, and yellow fever. During the last 8 years, the genomes from 24 insect species have been sequenced or are in the process of being sequenced (Figure 3). Some of these species are medically important, such as the malaria mosquito *Anopheles gambiae*, the yellow fever mosquito *Aedes aegypti*, and the blood-sucking bug *Rhodnius prolixus* (which transmits Chagas disease). Other insects are agriculturally important, and one insect, *Drosophila melanogaster*, is a well-known model organism.

The presence of these sequenced insect genomes has greatly facilitated the identification of insect neurohormone GPCRs. This is because gene-finding computer programs can detect most genes in a sequenced genome, and the characteristic GPCR

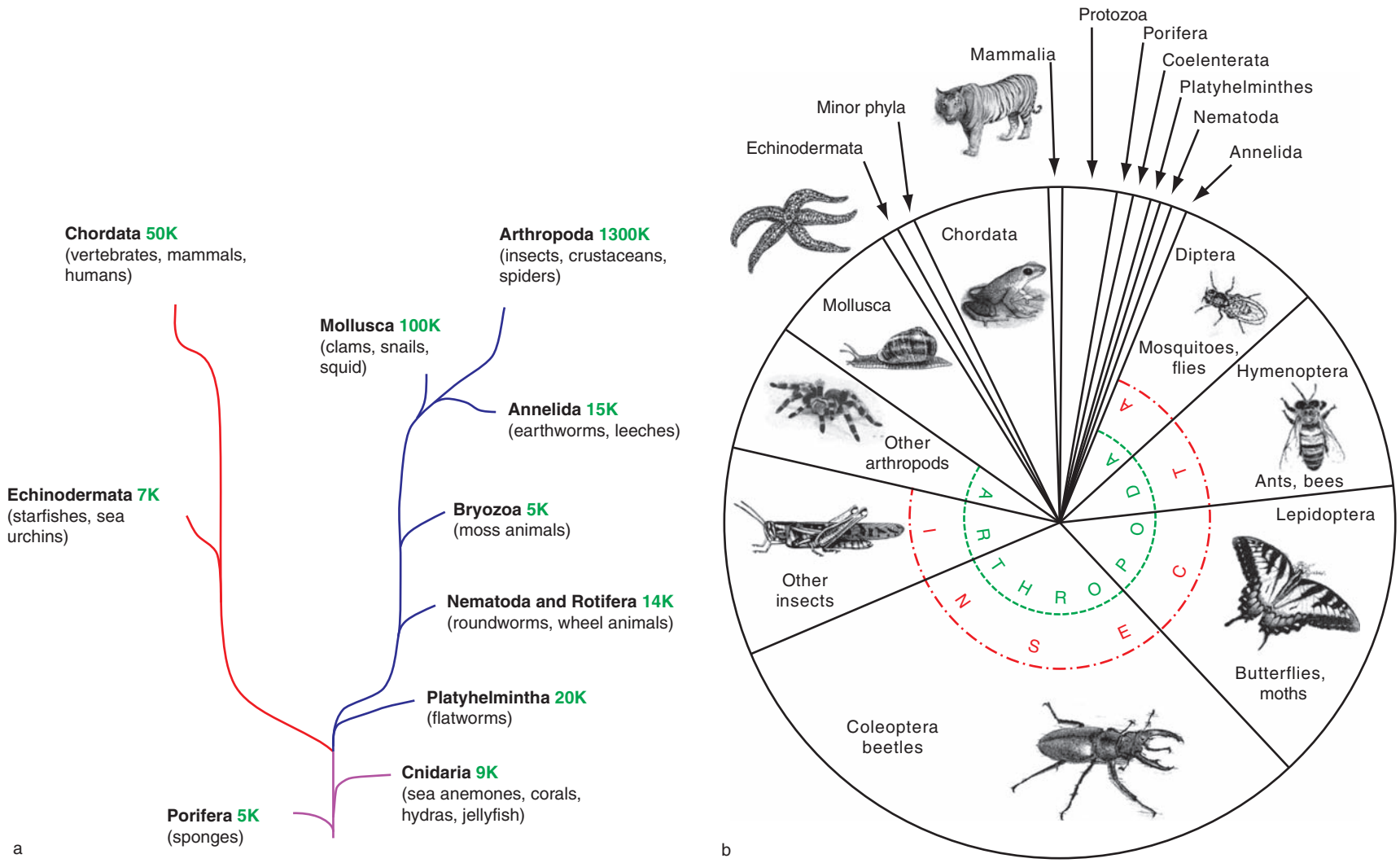

Figure 1 Animal evolution and number of animal species. (a) Phylogenetic tree, showing the two evolutionary lineages of multicellular animals that split shortly after the Cnidaria: Protostomia (blue) and Deuterostomia (red). The major phyla are highlighted in bold and their approximate species numbers are given in thousands of species (K). The majority of invertebrates belong to the Protostomia branch. Cnidarians are the first phylum, where a nervous system evolved; there are no nervous or endocrine cells in Porifera. (b) Schematic representation of the number of each species belonging to each animal phylum or order. Insects (class insecta or Hexapoda) account for 75% of all animal species living on Earth. Invertebrates account for approximately 90%, and mammals for only 1–2%.

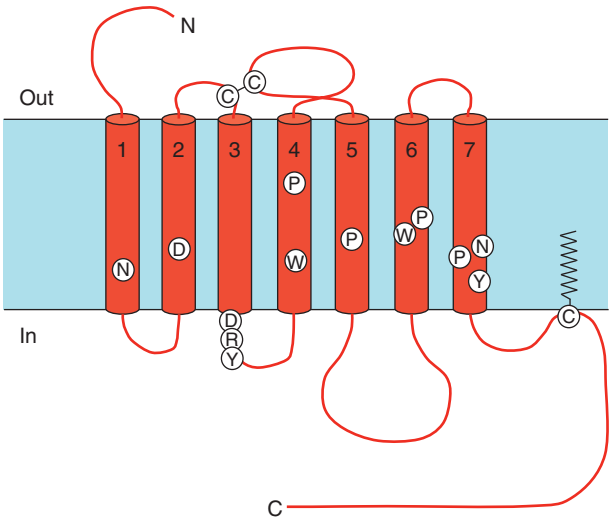

Figure 2 Topology of a neurohormone G-protein-coupled receptor (GPCR), belonging to family-A. This membrane protein has seven transmembrane α-helices, an extracellular N-terminus and an intracellular C-terminus. The extracellular loops 1 and 2 are connected by a cystine bridge. The DRY sequence shortly after transmembrane α-helix 3 is a hallmark for family-A GPCRs, as are other residues within several of the seven transmembrane α-helices. The C-terminus is often connected via a palmitoylated cysteine residue to the cell membrane. G-proteins interact with the three intracellular loops and the proximal region of the receptor C-terminus to initiate a second-messenger response after receptor activation.

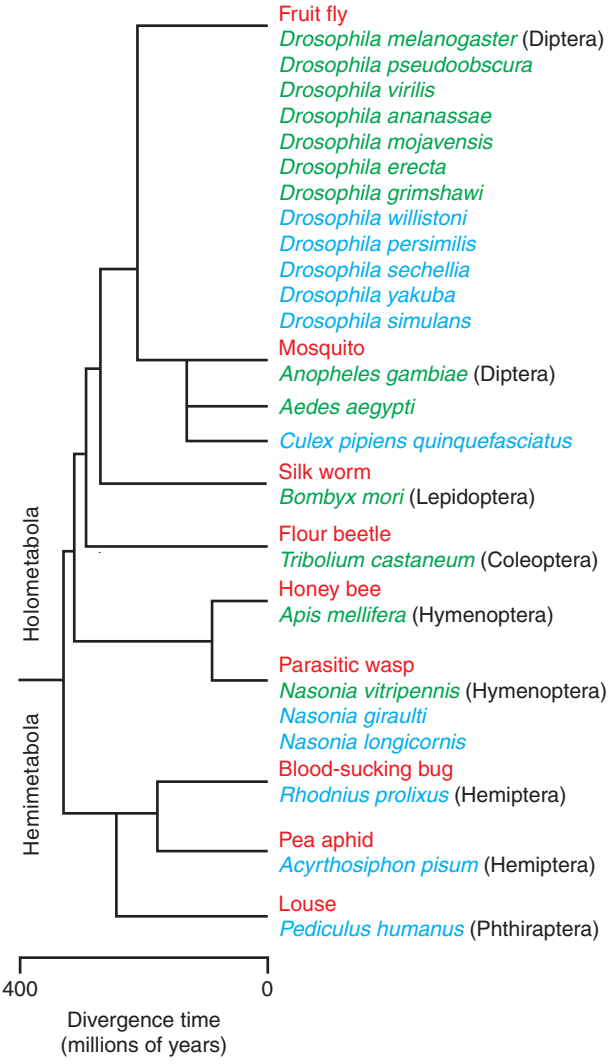

Figure 3 The evolutionary relationships of insects with a fully sequenced genome and those for which a whole genome-sequencing project is in progress. Insects for which the genome sequencing has been completed (at least 8–9 times coverage) are highlighted in green. Insects for which the genome sequencing is in progress are highlighted in blue. Adapted from Grimmelikhuijzen

topology (Figure 2) makes it possible to identify some of these genes as coding for neurohormone GPCRs. Of course, after the initial predictions (annotations), the identity of the predicted GPCR has to be experimentally confirmed (by cloning) and the cloned GCPR has to be deorphanized (its ligand to be identified). Ligand identification occurs after heterologous expression of the GPCR cDNA in cells in cell culture or in frog oocytes and the measurement of second messengers or second-messenger-mediated responses after addition of potential ligands.

Because *D. melanogaster* was the first insect whose genome was completely sequenced, most neurohormone GPCRs were first discovered and deorphanized in this fruitfly. As can be seen in Figure 4, *Drosophila* has 21 biogenic amine receptors (the *Drosophila* receptor genes are indicated by CG numbers; the Am numbers refer to the honey bee *Apis mellifera*). From these 21 receptors, 15 (71%) have been deorphanized (the identified ligand appears after the CG numbers in the figure). All the *Drosophila* biogenic amine receptors (Figure 4) are family-A GPCRs. The neuropeptide receptors given in Figure 5 are also family-A GPCRs; those in Figure 6 are family-B GPCRs. Together they show that *Drosophila* has 48 neuropeptide and protein hormone receptors. From these 48 GPCRs, 31 (65%) have been deorphanized. These deorphanized GPCRs are invaluable for our understanding of *Drosophila*, but they can also be used to annotate similar (ortholog) receptors in other insects with a sequenced genome.

Most insects belong to one of two evolutionary lineages: Holometabola (insects with a complete metamorphosis during development: larvae, pupae, adults) and Hemimetabola (insects with an incomplete metamorphosis: the early developmental stages, also called nymphae, resemble the adults). The holo- and hemimetabolous insects have a common evolutionary origin, which lies about 340 million years back in

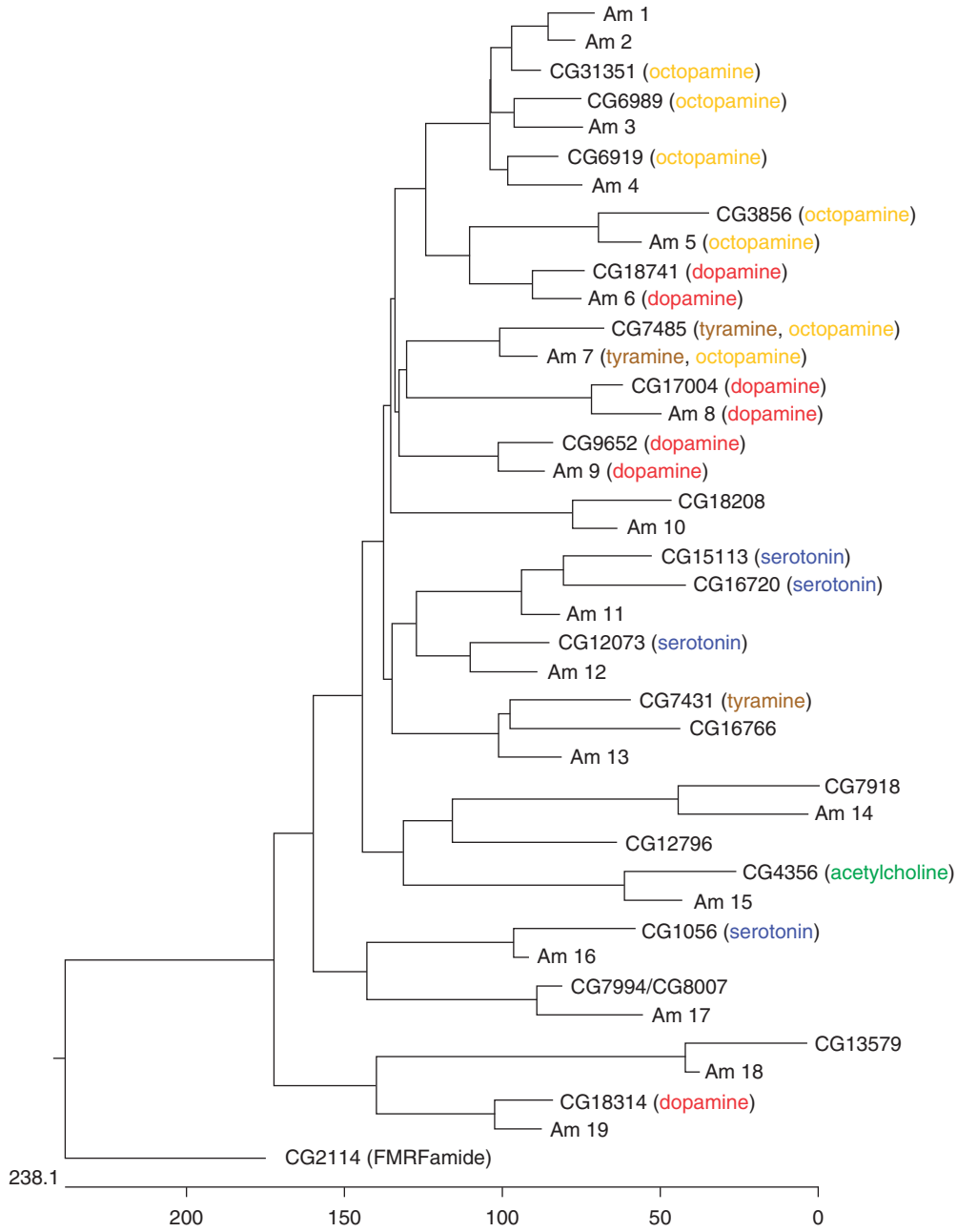

Figure 4 Phylogenetic tree analysis of the *Drosophila* and honey bee biogenic amine GPCRs. The *Drosophila* genes are indicated by their official CG identification numbers, whereas the honey bee genes are numbered Am 1–19. The tree is rooted by the *Drosophila* FMRFamide receptor gene, CG2114. The length of each branch represents the distance between each receptor and the common ancestor of that receptor and its neighbor. The units at the bottom represent the number of amino acid residue substitutions corresponding to this length. GPCR, G-protein-coupled receptor. Reproduced from Hauser F, Cazzamali G, Williamson M, Blenau W, and Grimmelikhuijzen CJP (2006) A review of neurohormone GPCRs present in the fruitfly *Drosophila melanogaster* and the honey bee *Apis mellifera*. *Progress in Neurobiology* 80: 1–19.

time (Figure 3). The holometabolous insects consist of four major orders: Diptera (flies), Lepidoptera (moths and butterflies), Coleoptera (beetles), and Hymenoptera (bees, wasps, and ants) (Figure 3). Figure 1(b) shows that more than half of all animals (more than 80% of all insects) are holometabolous insects. Recent genome comparisons show that Diptera are the most advanced and Hymenoptera the most basic holometabolous insects (Figure 3).

Because most available information on insect GPCRs is from *Drosophila*, we have used the *Drosophila* (Diptera) neurohormone GPCRs to

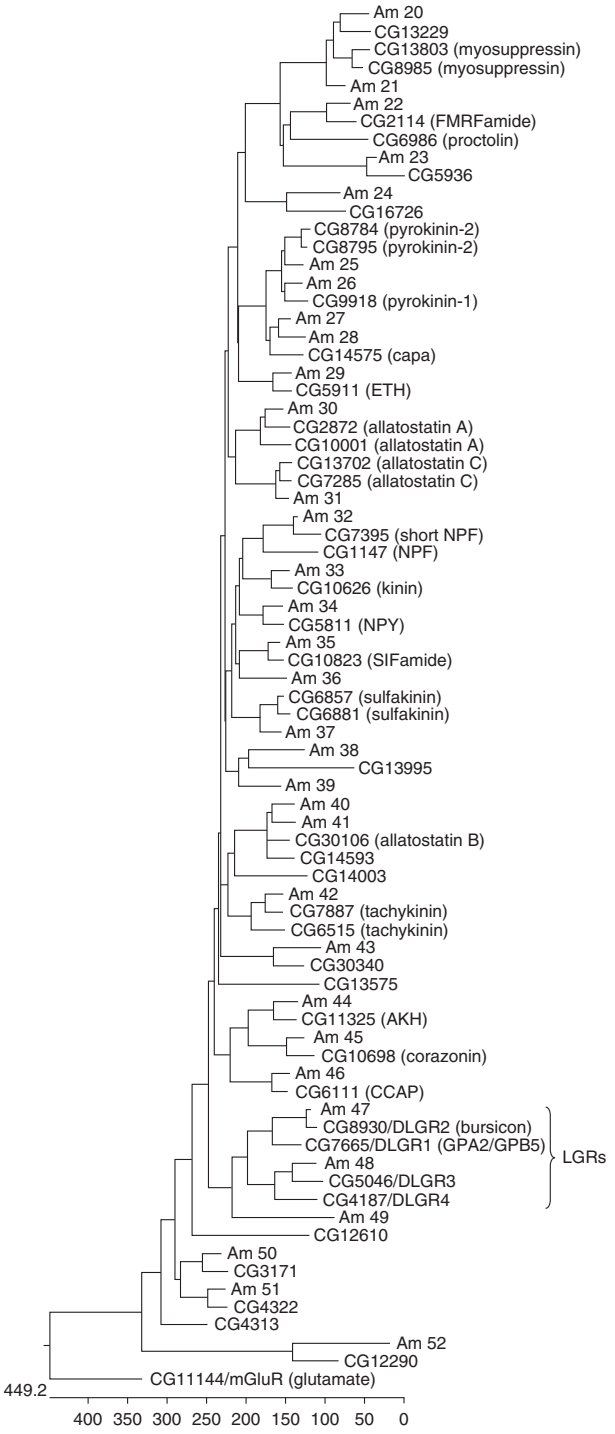

Figure 5 Phylogenetic tree analysis of the *Drosophila* and honey bee neuropeptide and protein hormone GPCRs belonging to family-A. The honey bee receptors are numbered Am 20–52. The tree is rooted by the *Drosophila* metabotropic glutamate receptor, CG11144. GPCR, G-protein-coupled receptor; LGR, leucine-rich repeats-containing GPCR. Reproduced from Hauser F, Cazzamali G, Williamson M, Blenau W, and Grimmelikhuijzen CJP (2006) A review of neurohormone GPCRs present in the fruitfly *Drosophila melanogaster* and the honey bee *Apis mellifera*. *Progress in Neurobiology* 80: 1–19.

annotate neurohormone GPCRs in the genome from the honey bee *A. mellifera* (Hymenoptera). The results of these comparisons are shown in Figures 4–6. Whereas *Drosophila* has 21 biogenic amine receptors, the honey bee has 19 (Figure 4). The 15 deorphanized *Drosophila* biogenic amine GPCRs all have a honey bee ortholog, so we know by this annotation effort what the likely ligands for these 15 honey bee GPCRs are.

A similar comparison for the peptide and protein hormone GPCRs shows that, whereas *Drosophila* has 48 GPCRs of this kind, the honey bee has 37 (Figures 5 and 6). The major reason for this larger number in *Drosophila* is gene duplication, which occurred during late fly evolution (e.g., note the two *Drosophila* allatostatin-A and two allatostatin-C receptors, which both have only one honey bee ortholog, Am 30 and Am 31, in Figure 5).

Evolution of Insect Neurohormone GPCRs and the Possibility of Co-Evolution with Their Ligands

From Figures 5 and 6, it is clear that the neuropeptide and protein hormone GPCRs occur in clusters with common origins. This suggests that during evolution new receptors originated by gene duplication. These duplicated receptors could use the same ligands. For example, the *Drosophila* GPCRs encoded by CG13803 and CG8985 (Figure 5, top) both use myosuppressin as their ligand. We could imagine that, in some cases, it would be impractical for the two new receptors to use the same ligand, because the two systems would interfere. In this case, a new ligand could be generated by gene duplication, followed by mutations (amino acid substitutions). Examples of this are the pyrokinin receptors (Figure 5, top). Pyrokinins are insect neuropeptides with the C-terminal consensus sequence **FXPRLamide**. *Drosophila* has two pyrokinins with slightly different C-terminal structures, pyrokinin-1 (TGPSASSGLW**FGPRLamide**) and pyrokinin-2 (SVP**FKPRLamid**e), and three different pyrokinin receptors, CG9918, CG8784, and CG8795. Pyrokinin-1 strongly activates its receptor CG9918 (Figure 5) but, although structurally related, it does not activate CG8784 and only very weakly CG8795. Similarly, pyrokinin-2 strongly activates its receptors CG8784 and CG8795 (Figure 5), but only weakly CG9918. Thus, for neuropeptide (and protein hormone) GPCRs, it appears to be possible to create new receptor–ligand couples by gene duplication of receptor and ligand, followed by mutations and subsequent amino acid substitutions.

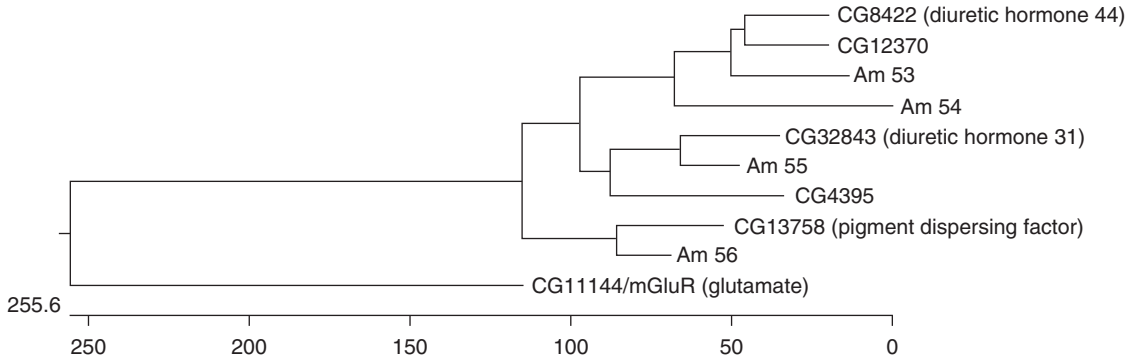

Figure 6 Phylogenetic tree analysis of the *Drosophila* and honey bee neuropeptide and protein hormone GPCRs belonging to family-B. The honey bee receptors are numbered Am 53–56. The tree is rooted by the *Drosophila* metabotropic glutamate receptor, CG11144. GPCR, G-protein-coupled receptor. Reproduced from Hauser F, Cazzamali G, Williamson M, Blenau W, and Grimmelikhuijzen CJP (2006) A review of neurohormone GPCRs present in the fruitfly *Drosophila melanogaster* and the honey bee *Apis mellifera*. *Progress in Neurobiology* 80: 1–19.

This type of co-evolution between receptors and ligands apparently does not occur during biogenic amine GPCR evolution (Figure 4). The reason for this is probably that the biogenic amine molecule is relatively small and that it is not as easy to modify this ligand as it is for peptides. For neuropeptides, it only takes one nucleotide mutation in its gene to change an amino acid residue, whereas for biogenic amines a new (decarboxylating, hydroxylating, etc.) enzyme system has to become active. Furthermore, the ligand binding pocket in the biogenic amine GPCR might be small and tolerate only little structural variation of the ligand. In short, what we see in Figure 4 is that gene duplications also occur during the evolution of biogenic amine GPCRs. Sometimes the same ligands are used by the duplicated (paralog) receptors; for example, CG15113 and CG16720 (Figure 4, middle) both use serotonin. However, the use of serotonin as a ligand is not confined to evolutionary-related GPCRs because serotonin is also used by a distantly related receptor CG1056 (Figure 4, bottom). Another example is dopamine, which is used by CG18314 (Figure 4, bottom) and two other groups of distantly related receptors: CG9652 + CG17004, and CG18741 (Figure 4, middle and top). Thus, during biogenic amine GPCR evolution, new GPCRs evolved that needed new ligands. Because of structural or biosynthetic constraints, the only way to obtain such new ligands was by borrowing from other systems. This caused a ligand hops during biogenic amine GPCR evolution.

When we look at honey bee biogenic amine GPCRs, we find that nearly every *Drosophila* receptor has a honey bee ortholog that probably uses the same ligand. For some instances, this has been confirmed by deorphanizing the honey bee receptor (e.g., Am 6, Am 8, and Am 9 use dopamine as a ligand; Figure 4). Thus, the ligand hops have occurred before the split of the evolutionary lineages leading to present-day flies and honey bees, which is about 340 million years ago (Figure 3).

The Bursicon Story

Only very few protein hormone GPCRs have been deorphanized in *Drosophila*, one of them being the receptor for bursicon. Bursicon was described more than 40 years ago as a hormonal activity from the insect brain that causes tanning and hardening of the soft exoskeleton (cuticle) in a newly hatched insect after adult ecdysis (molt). Later studies showed that bursicon also steers a motor program needed for the final phase of metamorphosis, including wing expansion, and that it causes apoptosis of the wing epithelial cells after wing expansion is completed. The neurohormone bursicon has recently been identified as being a heterodimer, consisting of two different but related cystine-knot proteins. Also, the *Drosophila* receptor for bursicon has recently been identified, CG8930 (Figure 5, bottom). This receptor belongs to the family of leucine-rich repeats-containing GPCRs (LGRs), of which we found four members in *Drosophila* (DLGR1–4). LGRs have a large leucine-rich repeats-containing, horseshoelike structure in the extracellular N-terminus, which represents the ligand-binding part of the receptor. The *Drosophila* LGRs have orthologs in mammals, where they are the receptors for follicle-stimulating hormone, luteinizing hormone/choriogonadotropin, thyroid stimulating hormone, and relaxin, which are all hormones involved in reproduction or development.

The honey bee has a close ortholog (Am 47) to the *Drosophila* bursicon receptor, which suggests that Am 47 (Figure 5, bottom) is the receptor for a honey bee bursicon. This is corroborated by the presence of a bursicon gene in the honey bee genome.

GPCRs in Other Invertebrates

Because LGRs are present in both insects and mammals, it can be concluded that they must have evolved before the split of Protostomia and Deuterostomia, in other words, very early in animal evolution (Figure 1(a)). This is exactly what we find because LGRs have also been cloned from cnidarians – in this case, from the sea anemone *Anthopleura elegantissima*. The ligand of the sea anemone LGR has not been identified yet, so the receptor is still an orphan. Other orphan LGRs have been reported in nematodes (the roundworm *Caenorhabditis elegans*) and mollusks (the snail *Lymnaea stagnalis*), so this family of protein hormone receptors appears to be ubiquitous in the Animal Kingdom.

Information on neurohormone GPCRs in invertebrates other than insects is rather scarce and fragmented, and this is probably due to the lack of completed genome projects. Although no molluskan genome project has been completed, approximately eight neuropeptide GPCRs have already been identified in this phylum, mainly in the pond snail *L. stagnalis*. Among them is the receptor for an FMRF-like peptide (LyCEP); this is not, however, the receptor for the authentic molluskan FMRFamide neuropeptide, which remains elusive. Both in *L. stagnalis* and in *Octopus vulgaris*, molluskan variants of oxytocin (OT) and vasopressin (VP) occur, and their corresponding GPCRs have been identified. Similar OT- and VP-like peptides and their GPCRs have been found in annelids. OT and VP also clearly occur in deuterostomians, which suggests that the OT-VP system has ancient evolutionary roots (before the split of Deuterostomia and Protostomia; Figure 1(a)).

The genome of a lower invertebrate and model organism, *C. elegans*, has been fully sequenced (together with three other *Caenorhabditis* species). Bioinformatic screening of the *C. elegans* genome yielded approximately 18 biogenic amine and 50 neuropeptide GPCRs. These are similar to the numbers found in insects. Still, no global annotation has been performed of neurohormone GPCRs comparable to the ones carried out in *D. melanogaster* and *A. mellifera* (Figures 4–6). Twenty-nine genes (*flp*-1 to *flp*-29) in *C. elegans* encode preprohormones for FMRFamide-like peptides (flp), yielding altogether 68 distinct flp variants. It comes as no surprise, therefore, that all four neuropeptide GPCRs cloned and deorphanized in this nematode are flp receptors. The gene of one of these flp receptors (NPR-1) has two allelic variants, of which one is responsible for social and the other one for solitary feeding.

In addition to the four flp receptors, nine biogenic amine GPCRs have been deorphanized in *C. elegans*: three for serotonin, three for acetylcholine, two for octopamine/tyramine, and one for dopamine. These biogenic amine receptors and their ligands control various aspects of behavior. The serotonin receptor SER-1, for example, controls egg laying and mating behavior. The information obtained from studies in *C. elegans* so far suggests that the set of neurohormone GPCRs in lower invertebrates is not much different from that in higher invertebrates (insects) (Figure 1(a)). The numbers of biogenic amine and neuropeptide GPCRs is about the same, and similar biogenic amines occur in insects as in nematodes (acetylcholine, serotonin, dopamine, tyramine, and octopamine). The only difference between lower and higher invertebrates may be that flp neuropeptides and their receptors are biased in nematodes. This is understandable, however, because this class of neuropeptides is abundant and widespread in basal invertebrates, such as cnidarians, suggesting that it is evolutionarily very old.

See also: Astrocyte: Neurotransmitter and Hormone Receptors; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Neuropeptide Signaling in Invertebrates; Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells; Octopamine and Other Monoamines in Invertebrates.

Further Reading

Cazzamali G and Grimmelikhuijzen CJP (2002) Molecular cloning and functional expression of the first insect FMRFamide receptor. *Proceedings of the National Academy of Sciences of the United States of America* 99: 12073–12078.

Cazzamali G, Torp M, Hauser F, Williamson M, and Grimmelikhuijzen CJP (2005) The *Drosophila* gene CG9918 codes for a pyrokinin-1 receptor. *Biochemical and Biophysical Research Communications* 335: 14–19.

Civelli O (2005) GPCR deorphanizations: The novel, the known and the unexpected transmitters. *Trends in Pharmacological Science* 26: 15–19.

Civelli O, Nothacker H-P, Saito Y, Wang Z, Lin SH, and Reinscheid RK (2001) Novel neurotransmitters as natural ligands of orphan G-protein-coupled receptors. *Trends in Neuroscience* 24: 230–237.

Egerod K, Reynisson E, Hauser F, Cazzamali G, Williamson M, and Grimmelikhuijzen CJP (2003) Molecular cloning and functional expression of the first two specific insect myosuppressin receptors. *Proceedings of the National Academy of Sciences of the United States of America* 100: 9808–9813.

Gether U (2000) Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocrine Reviews* 21: 90–113.

Grimmelikhuijzen CJP, Cazzamali G, Williamson M, and Hauser F (2007) The promise of insect genomics. *Pest Management Science* 63(5): 413-416.

Grimmelikhuijzen CJP, Williamson M, and Hansen GN (2002) Neuropeptides in cnidarians. *Canadian Journal of Zoology* 80: 1690–1702.

Hauser F, Cazzamali G, Williamson M, Blenau W, and Grimmelikhuijzen CJP (2006) A review of neurohormone GPCRs present in the fruitfly *Drosophila melanogaster* and the honey bee *Apis mellifera*. *Progress in Neurobiology* 80: 1–19.

Kanda A, Satake H, Kawada T, and Minakata H (2005) Novel evolutionary lineages of the invertebrate oxytocin/vasopressin superfamily peptides and their receptors in the common octopus (*Octopus vulgaris*). *Biochemical Journal* 387: 85–91.

Luo CW, Dewey EM, Sudo S, et al. (2005) Bursicon, the insect cuticle-hardening hormone, is a heterodimeric cystine knot protein that activates G protein-coupled receptor LGR2. *Proceedings of the National Academy of Sciences of the United States of America* 102: 2820–2825.

McVeigh P, Geary TG, Marks NJ, and Maule AG (2006) The FLP-side of nematodes. *Trends in Parasitology* 22: 385–396.

Mendive FM, Van Loy T, Claeysen S, et al. (2005) *Drosophila* molting neurohormone bursicon is a heterodimer and the natural agonist of the orphan receptor DLGR2. *FEBS Letters* 579: 2171–2176.

Mertens I, Meeusen T, Janssen T, Nachman R, and Schoofs L (2005) Molecular characterization of two G protein-coupled receptor splice variants as FLP2 receptors in *Caenorhabditis elegans*. *Biochemical and Biophysical Research Communications* 330: 967–974.

Nässel DR (2002) Neuropeptides in the nervous system of *Drosophila* and other insects: Multiple roles as neuromodulators and neurohormones. *Progress in Neurobiology* 68: 1–84.

Park Y, Kim YJ, and Adams ME (2002) Identification of G protein-coupled receptors for *Drosophila* PRXamide peptides, CCAP, corazonin, and AKH supports a theory of ligand-receptor coevolution. *Proceedings of the National Academy of Sciences of the United States of America* 99: 11423–11428.

Rogers C, Reale V, Kim K, et al. (2003) Inhibition of *Caenorhabditis elegans* social feeding by FMRFamide-related peptide activation of NPR-1. *Nature Neuroscience* 6: 1178–1185.

Savard J, Tautz D, Richards S, et al. (2006) Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Research* 16: 1334–1338.

Stables J, Green A, Marshall F, et al. (1997) A bioluminescent assay for agonist activity at potentially any G-protein-coupled receptor. *Analytical Biochemistry* 252: 115–126.

Staubli F, Jørgensen TJD, Cazzamali G, et al. (2002) Molecular identification of the insect adipokinetic hormone receptors. *Proceedings of the National Academy of Sciences of the United States of America* 99: 3446–3451.

Vanden Broeck J (2001) Neuropeptides and their precursors in the fruitfly. *Drosophila melanogaster. Peptides* 22: 241–254.

Inwardly Rectifying Potassium Channels

H T Kurata and C G Nichols, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

'Inwardly rectifying' K^+ (Kir) channels are expressed in a variety of cell types and are involved in regulating numerous physiological processes. Examples range from roles in regulating insulin release from pancreatic β-cells to regulating electrolyte balance in the kidney to maintaining a stable resting membrane potential in many tissues. In addition, spontaneous or inherited mutations in genes encoding Kir channels have been shown to underlie a number of genetic disorders, emphasizing the breadth and importance of the role of Kir channels in physiology and normal cellular function.

As a functional descriptor, the term inwardly rectifying broadly refers to the preferential conduction of currents in the inward direction. Most K^+ channels that exhibit inward rectification are members of the Kir channel family, can be categorized as Kir channels, which are distinguished from other classes of K^+ channels based on important features of their primary sequences and overall structures. Functionally, inward rectification of Kir channels arises from voltage-dependent channel blockade by intracellular cations such as Mg^{2+} or polyamines. Structurally, each subunit of a Kir channel comprises two transmembrane helices and a large cytoplasmic domain formed by the N- and C-termini. The absence of a voltage-sensing domain is an important structural distinction from the six transmembrane segments present in the voltage-gated class of K^+ channels (Kv channels). Although the gating behavior of Kir channels is not intrinsically voltage-dependent, this class of channels exhibits a rich tapestry of regulation by a number of physiological ligands/factors, including nucleotides, auxiliary subunits, protons/pH, phospholipids, and $G_{\beta\gamma}$ subunits. In this article, we review current understanding of the structural features of Kir channels and some of the mechanisms that underlie their regulation.

Molecular Architecture of Kir Channels and Channel Complexes

Like all members of the cation channel superfamily, the membrane-spanning region of Kir channels is a tetrameric arrangement of a core pore-forming unit. Kv channels contain additional transmembrane helices with charged residues that form a voltage-sensing domain and underlie the voltage-sensitivity of these channels. Two-pore K^+ (K2P) channels lack a voltage-sensing domain structure and comprise two pore-forming units arranged in tandem. Kir channel subunits, in contrast, contain only the core pore-forming domain in a single copy (Figure 1(a)). Crystal structures have been determined for a full-length bacterial inward rectifier (KirBac1.1) and for fragments of two eukaryotic Kir channels (Kir2.1 and Kir3.1), advancing our understanding of Kir structure to atomic resolution (Figures 1(b) and 1(c)). The two transmembrane helices of the core pore-forming domain are named M1 and M2 and correspond to the pore-forming S5 and S6 helices in Kv channels (Figure 1(b)). In some Kir channels, the M2 helix contains an acidic residue that has been identified as a critical determinant of polyamine sensitivity and is frequently referred to as the rectification controller (see the section titled 'Inward rectification'). Kir channels also exhibit the characteristic reentrant pore loop structure, containing a K^+-channel signature sequence that forms the selectivity filter of the channel. Within the selectivity filter, K^+ ions are dehydrated and stabilized by backbone carbonyl oxygen atoms contributed by each subunit of the channel. The selectivity filter underlies the ability of the K^+ channels to selectively permeate K^+ in the presence of other physiological cations such as Na^+. The rotationally symmetric arrangement of these pore-forming domains forms a single K^+-selective pore that spans the plasma membrane (Figure 1(c)).

Beyond the central transmembrane pore domain are cytoplasmic N- and C-termini. These two ends of the channel assemble into a large cytoplasmic domain consisting mostly of β-sheet structures, with the N-terminus of one subunit interacting primarily with the C-terminus of an adjacent subunit. This feature is illustrated in the cutaway image of a channel in Figure 1(b), where the N-terminus of the red subunit interacts with the large C-terminus of the blue subunit. The cytoplasmic domain extends the transmembrane pore-forming domain well beyond the width of the plasma membrane, such that the obligatory K^+ ion permeation pathway is approximately 70 Å in length. This extended pore region contains multiple pore-facing acidic residues (as many as four per subunit in Kir2.1), which play a role in determining both unitary conductance and sensitivity to various blocking cations. In addition, the Kir N- and C-termini together form binding sites for multiple physiologically

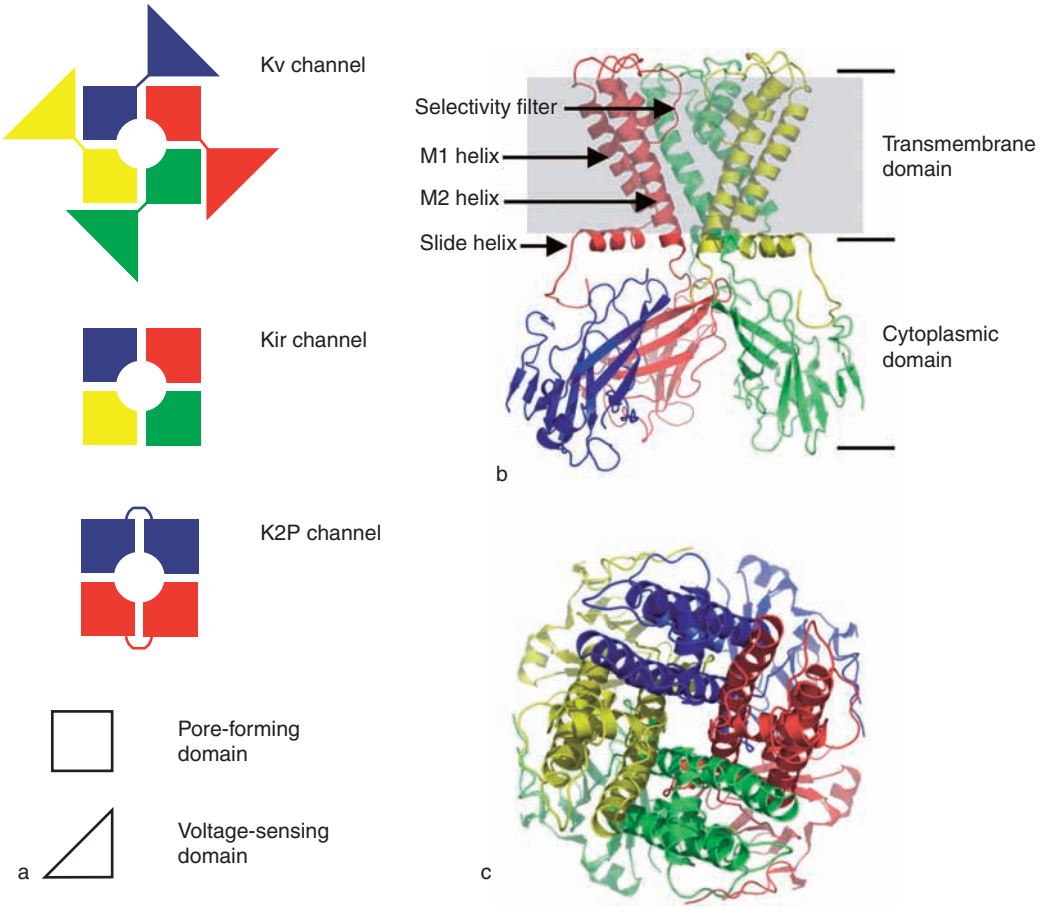

Figure 1 Kir channels: (a) molecular architecture; (b) side-view atomic resolution structure of a bacterial channel (KirBac1.1); (c) top-view atomic resolution structure of the bacterial channel. In (a), Kir channels are shown as tetrameric arrangements of a core pore-forming domain. Squares represent the canonical pore-forming K^+ channel domain, with each color representing a distinct subunit. K^+ channels exhibit rotational symmetry around a central K^+-selective pore. In Kv channels, an additional modular voltage-sensing domain is attached to the core pore-forming domain, whereas in two-pore (K2P) channels, two core domains are arranged in tandem. In (b), a cutaway side view of the channel is shown (the cytoplasmic domain of the yellow subunit and the transmembrane domain of the blue subunit have been removed for clarity). Together the primarily α-helical transmembrane domain and the β-sheet-rich cytoplasmic domain form a pore over 70 Å in length. The shaded box represents the approximate dimensions of the plasma membrane, with the slide helix located near the interface of the membrane and cytoplasm.

important channel ligands, including $G_{\beta\gamma}$ subunits in Kir3 channels and adenosine triphosphate (ATP) in Kir6 channels.

A final interesting structural feature of Kir channels is the slide helix, an α-helical segment that is apparent in the full-length KirBac 1.1 crystal structure. This segment precedes the M1 helix and runs parallel to the plane of the plasma membrane. Structures analogous to the slide helix were absent in the earliest K^+-channel crystal structures determined for KcsA and MthK, and at present a definite functional role for the slide helix has not been established. This structure could be considered equivalent to the S4–S5 linker in Kv channels, which has been proposed to play an important role in coupling of the pore-forming domain to conformational changes in the voltage-sensing domain. Along these lines, one possibility is that the slide helix plays a critical role in coupling ligand binding in the cytoplasmic domain to gating events in the transmembrane domain of Kir channels. A second possibility is that the slide helix underlies some channel interactions with membrane phospholipids.

Adding complexity to the molecular arrangement of the pore-forming Kir channel subunits is their assembly with various auxiliary proteins. Among the best-described examples of Kir auxiliary subunits is the obligate coassembly of Kir6 channel subunits with the sulfonylurea receptor (SUR). This coassembly is required for adequate trafficking of either binding partner to the plasma membrane, and it is thought to arise from mutual shielding of polybasic endoplasmic reticulum (ER) retention signals (RXR motifs) present in both Kir6 and SUR subunits. Apart from a trafficking

role, coassembly with SUR subunits plays a critical role in determining the response of the channel complex to various nucleotides. Other examples of channel assembly with associated proteins include the association of Kir2 channels with SAP97 and the association of Kir4.1 with the dystrophin–glycoprotein complex, although these binding partners are generally not considered to be channel subunits and their impact on channel gating has not been determined. Finally, recent work has suggested that the G-protein-regulated Kir3 subfamily channels (GIRKs) may remain intimately associated with heterotrimeric G-proteins before and after the activation of G-protein-coupled receptors (GPCRs), with this preassociation of signal complexes playing an important role in the generation of specificity of channel responses to extracellular signals.

Inward Rectification

Historically, inward rectification – the preferential conduction of currents in the inward direction – was the identifying functional property of Kir channels. It is now understood that inward rectification is not an intrinsic property of Kir channels but, rather, is a consequence of a voltage-dependent block by intracellular Mg^{2+} or polyamines (e.g., spermine, spermidine, cadaverine, and putrescine). As a consequence, removal of channels from the cellular environment (e.g., in an excised patch-clamp experiment) results in the loss of apparent rectification properties, and these can be restored by the application of exogenous polyamines or Mg^{2+} (Figure 2(a)). Significantly, the various cloned Kir channels exhibit a spectrum of the degree or strength of rectification. These differences in rectification arise due to different sensitivities to intracellular blocking cations and are reflected in the different functional roles of each channel type. In strongly-rectifying channel types, rectification is so dramatic that very little current flows through them at potentials more positive than the reversal potential of K^+ (Figure 2(a)). This classical, strong inward rectification has been most widely studied in channels from the Kir2 subfamily, although Kir4 channels and heteromultimers of Kir4 and Kir5.1 channels also exhibit a steeply voltage-dependent and highly potent block by spermine. The physiological relevance of strong inwardly rectifying K^+ channels is especially apparent in tissues characterized by rapid or repetitive excitation, such as skeletal myocytes and neurons in the central nervous system, and in tissues that exhibit prolonged action potentials, such

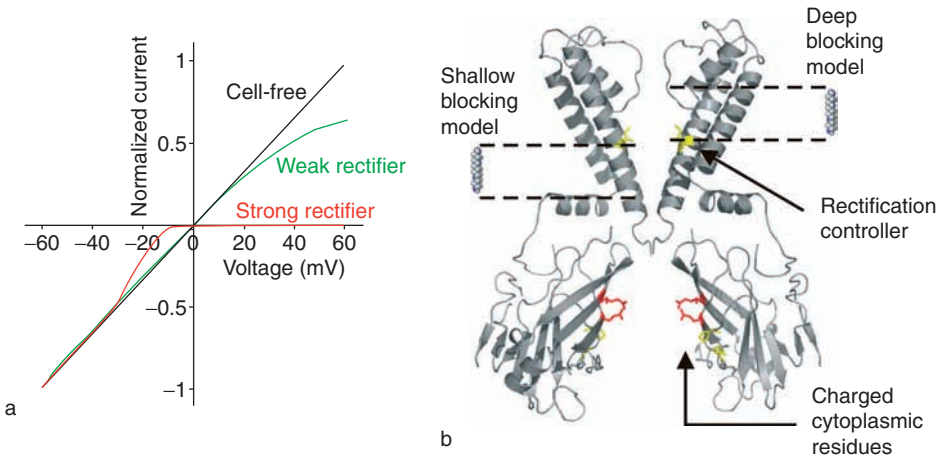

Figure 2 Mechanism of inward rectification: (a) normalized current–voltage relationships; (b) pore-lining acidic residues that interact with polyamines and potential spermine binding sites. In (a), the normalized current–voltage relationships are recorded from inside-out patches in symmetrical 150 mM K^+ solutions for a strongly rectifying Kir channel (Kir2.1, red trace) and a weakly rectifying Kir channel (Kir6.2, green trace) in the presence of 100 μM spermine applied to the cytoplasmic side of the membrane patch. In the absence of spermine, Kir channels behave essentially as ohmic resistors and exhibit little or no inward rectification (cell-free, black trace). In (b), strongly rectifying Kir channels possess a pore-lining acidic residue in the transmembrane domain, referred to as the rectification controller residue. Kir channels also possess multiple acidic residues in the cytoplasmic domain (charged cytoplasmic residues) that influence the kinetic and equilibrium features of polyamine block. Acidic residues that are present in wild-type KirBac1.1 channels are colored red, whereas acidic residues that are absent in KirBac1.1 but present in the strong rectifier Kir2.1 are colored yellow at the homologous position. Polyamines block Kir channels by entering and occluding the central permeation pathway, and several models have been suggested to describe the stable polyamine binding site underlying strong rectification. The shallow blocking model suggests that polyamines bind at a shallow site between the cytoplasmic charged residues and the rectification controller. Recent evidence favors the deep blocking model, which suggests that polyamines bind very deep in the pore, between the rectification controller and the selectivity filter.

as cardiac myocytes. In these cell types, channels exhibiting strong inward rectification underlie the maintenance of a stable resting potential, but the significant and rapid reduction of K^+ conductance on depolarization beyond the K^+ reversal potential allows action potentials to take place. Other subfamilies, such as Kir1, Kir6, and Kir7 channels, exhibit much weaker rectification and conduct substantial currents in the outward direction, even in the presence of high concentrations of spermine or other polyamines (Figure 2(a)).

Molecular Determinants of Polyamine Block

The early mutagenesis studies of inward rectification provided important clues to the molecular determinants of polyamine block, but the recent determination of Kir structures at atomic resolution has revitalized studies aimed at elucidating the detailed structural basis of inward rectification. Structure–function studies demonstrated that the negatively charged rectification controller residue, present in the pore-lining M2 helix, ensures steeply voltage-dependent and highly potent rectification in the presence of intracellular spermine (Figure 2(a)). Sequence homology with crystallized K^+ channels such as KcsA or KirBac suggests that this residue lines the pore and interacts electrostatically with positively charged amines in spermine or other blocking cations. In strongly rectifying channels, such as Kir2.1, neutralization of this residue (D172 in Kir2.1) invariably weakens rectification, but does not abolish the block by intracellular polyamines or Mg^{2+}. However, in weakly rectifying channels, the introduction of a negative charge at the equivalent position is generally sufficient to reconstitute the most important features of strong inward rectification (i.e., high voltage dependence and high spermine potency). This effect is apparent in the N171D mutation in Kir1.1 and the N160D or N160E mutations in Kir6.2. Furthermore, replacement of this residue with a positively charged lysine causes a loss of both polyamine- and Mg^{2+}-dependent inward rectification. Significantly, although the rectification controller residue is conserved among strongly rectifying channels, there does not appear to be a strict requirement for the location of the negative charge to confer strong rectification. In Kir6.2 channels, for example, the introduction of negatively charged glutamate residues at any pore-lining position in the M2 helix is sufficient to confer high polyamine sensitivity and steeply voltage-dependent polyamine block.

In addition to the rectification controller, mutagenesis studies and the recent structural determination of the cytoplasmic domain of Kir2.1 have identified negatively charged residues beyond M2, within the cytoplasmic domain of the channel, that affect sensitivity to blockade by polyamines, the kinetics of polyamine block, or both. In Kir2.1, at least four acidic pore-lining residues have been identified in the cytoplasmic domain (E224, D255, D259, and E299) that have a significant effect on the properties of spermine block (Figure 2(a)). In the tetrameric arrangement of an intact channel, each of these residues forms a ring of negative charge that lines the pore. Neutralizing or charge-reversing mutations at any of these sites can alter the kinetic and steady-state properties of the Mg^{2+} and polyamine block. In Kir6.2 channels, a nearby protonatable site (residue H216) also regulates the properties of the polyamine block, resulting in a weakened block at acidic pH. Significantly, the number of pore-lining negative charges in the cytoplasmic domain varies among different channel types, although all inward rectifiers (weak or strong) have at least one pore-lining acidic residue in the cytoplasmic domain. Therefore, the presence of negatively charged residues in the cytoplasmic domain is not sufficient to confer properties of strong rectification.

Characterization of the kinetic and steady-state features of spermine block of Kir channels demonstrated that there are at least two sequential steps involved. An initial, weakly voltage-dependent step leads to spermine binding at a shallow location in the pore, near the intracellular entrance to the channel, probably within the cytoplasmic domain. Kinetic modeling of experimental data indicates that the affinity of this shallow binding site is dramatically reduced by the neutralization of the acidic residues in this region. The shallow binding step is followed by a much more strongly voltage-dependent step, involving the entry of spermine into a stable binding site deep in the pore. It is this second deep-binding step that is responsible for most of the overall voltage dependence of spermine block. Significantly, although spermine is a polyvalent cation, the effective valence of the block by spermine and other polyamines in Kir2.1 exceeds the charge carried by the blocker. This observation suggests that much of the voltage dependence of spermine block does not arise from the blocker traveling across the membrane electric field but, rather, from coupled movement of occupant K^+ ions ahead of the blocker. The movement of these permeant ions through the selectivity filter accounts for a significant fraction of the voltage dependence of polyamine blockade.

With the identification of critical residues involved in the stabilization of polyamines in the Kir channel pore, many studies have attempted to determine the precise physical location of spermine binding,

particularly in its deepest, most stable binding site. Two models of spermine block have received the most attention. One model of shallow spermine binding suggests that the stable spermine binding site necessary for steep voltage-dependent rectification is formed by negative charges in both the inner cavity and cytoplasmic domain. In this model, spermine lies between the rectification controller (in the inner cavity) and the pore-lining acidic residues in the cytoplasmic domain (see Figure 2(b)). An alternative model of deep spermine block suggests that steep rectification involves spermine entry to a very deep site on the pore, near or within the selectivity filter (see Figure 2(b)). Although the location of stable spermine binding in the Kir pore is not yet established by a crystal structure, recent studies demonstrated that stably bound spermine (in Kir2.1 or Kir6.2 N160D channels) significantly interferes with methanethiosulfonate ethylammonium (MTSEA) modification of cysteine residues substituted between the rectification controller and the selectivity filter but not at sites closer to the cytoplasmic entrance of the channel. These data strongly support the second model of deep spermine binding, in which the leading amine of spermine could be stabilized within the selectivity filter, with the trailing amine(s) interacting with the rectification controller in the inner cavity.

Ligand Gating of Kir Channels

Apart from intracellular blockade by Mg^{2+} and polyamines, very critical physiological roles of Kir channels relate to their regulation by a variety of ligands. Many Kir channels are strongly inhibited at acidic pH, a feature most widely studied in the ROMK1 (Kir1.1) channel. Kir3 family members (or GIRK channels) are activated by interactions with $G_{\beta\gamma}$ subunits on activation of the GPCRs, whereas Kir6 family members are inhibited by intracellular nucleotides, particularly ATP. In general, the cytoplasmic domain provides a critical interface for the interactions of Kir channels with their ligands, although specific details vary between different channel subfamilies.

In Kir3 channels, much effort has been devoted to determining the structural basis of channel interactions with G-proteins. Segments in both the N- and C-termini have been shown to interact directly with $G_{\beta\gamma}$ subunits. In addition, binding of $G_{\beta\gamma}$ appears to be enhanced when both the N- and C-terminal fragments of the channel are present together. This finding probably reflects the close physical proximity of these channel segments in crystal structures (Figure 1(b)) and suggests that $G_{\beta\gamma}$ interactions involve multiple contacts with the cytoplasmic domains of Kir3 channels. Also consistent with this hypothesis is the identification of multiple segments of $G_{\beta\gamma}$ subunits required for functional interactions with channels. Among the most pressing questions regarding physiological regulation of GIRK channels is the determination of mechanisms underlying receptor specificity. It is now well known that Kir3 family channels can be activated by the cytoplasmic application of recombinant $G_{\beta\gamma}$ with little specificity. However, the activation of GIRK channels in native tissues by the stimulation of cell-surface receptors can only occur if the receptors couple to pertussis toxin-sensitive G-proteins ($G_{i/o}$). One proposed means by which this specificity is achieved is that there are direct interactions between the channel and G_α subunits. This proposal diverges from the now widely accepted view of GIRK channel activation, in which the activation of a GPCR results in the dissociation of G_α and $G_{\beta\gamma}$ subunits, allowing $G_{\beta\gamma}$ to diffuse and bind/activate GIRK channels.

Binding of ATP to Kir6 subunits similarly involves residues in both the N- and C-termini. Several residues contributing directly to an ATP-binding pocket have been identified by mutagenesis studies, including N-terminal residues R50 and Q52 and C-terminal residues G334, R201, K185, E179, and R301. These amino acids map to nearby locations in a molecular model of the cytoplasmic domain of Kir6.2, generating four ATP-binding sites in an intact channel. Each ATP-binding site comprises N-terminal residues of one subunit and C-terminal residues of an adjacent subunit. Binding of ATP is coupled to the closure of a ligand-operated gate that occludes K^+ flux through the channel.

Significantly, K_{ATP} channels exhibit additional complex regulation by other nucleotides, adenosine diphosphate (ADP) in particular, which governs the activity of these channels in the face of various cellular metabolic conditions. ADP (in the presence of Mg^{2+}) actually counters the effects of ATP and exerts a stimulatory action on K_{ATP} channels. This requires interaction with the SUR auxiliary subunit of Kir6 channels rather than the direct interaction of a nucleotide with the channel. K_{ATP} channels are present in all muscle cell types and in the brain, although the physiological role of the dueling influence of ATP and ADP is probably best understood in β-cells of the pancreas. In these cells, K_{ATP} channels are tonically active under resting conditions and contribute significantly to the maintenance of a stable negative resting-membrane potential. However, when blood glucose is elevated and intracellular ATP levels are increased, the ATP/ADP ratio is increased, leading to the inhibition of K_{ATP} channels, depolarization of the membrane, opening of Ca^{2+} channels, and release of insulin.

Overall, there is a general theme of ligand gating in which ligands bind to the cytoplasmic domain of Kir channels and influence the conformational state of a gate located along the permeation pathway. Depending on the channel and the ligand, interactions may cause the gate to open (as with GIRK channels and $G_{\beta\gamma}$ subunits) or to close (as with K_{ATP} channels and ATP). The conformational changes involved in ligand-dependent gating have been studied in multiple channel types. Although considerable insights have been gained, there remain uncertainties regarding the nature of the gate itself and the molecular details of opening and closing. It has been suggested that several structural motifs play a role in channel gating, including the selectivity filter and the crossing point of the M2 helices (the bundle crossing; Figure 3(a)). At present, there is a general consensus that rapid conformational changes around the selectivity filter underlie the brief flickering openings and closings that characterize the intra-burst behavior apparent in single-channel recordings of Kir channels. In contrast to intra-burst flickering, the burst duration depends very strongly on ligand interactions and probably depends on the conformation of the bundle crossing. In the crystal structure of KirBac1.1, the M2 helices form a constriction in the bundle-crossing region, with the narrowest point formed by the bulky hydrophobic side chains of residue F146. Hinged movements of the M2 helix are thought to allow the widening of this aperture and the opening of the permeation pathway (Figure 3(b)). By analogy to Kv channels, in which a constriction near the base of S6 (equivalent to M2 in Kir channels) constitutes a steric gate to ion entry into the channel pore, the constriction of M2 at the bundle crossing probably acts as the ligand-operated gate in Kir channels (Figure 3(b)). Considerable evidence supports this hypothesis because the accessibility of cysteine residues substituted in the inner cavity of Kir6.2 and Kir2.1 can be impeded by ligand-mediated channel closure.

Most recently, multiple structures of Kir channels or isolated domains also revealed considerable structural variability of the G-loop region in the cytoplasmic domain (Figure 3(a)). It has been speculated that the G-loop may constitute an additional gating structure in Kir channels because, in certain crystal structures, the G-loop forms a narrow constriction near the top of the cytoplasmic domain. Mutations in the G-loop of Kir2.1 have been shown to disrupt channel function, although at present there is little experimental evidence demonstrating that conformational changes in the G-loop can regulate K^+ permeation in an intact channel. The G-loop does appear to exhibit considerable structural flexibility, and future work will probably determine whether this structure behaves as a gate or, rather, is involved in the coupling

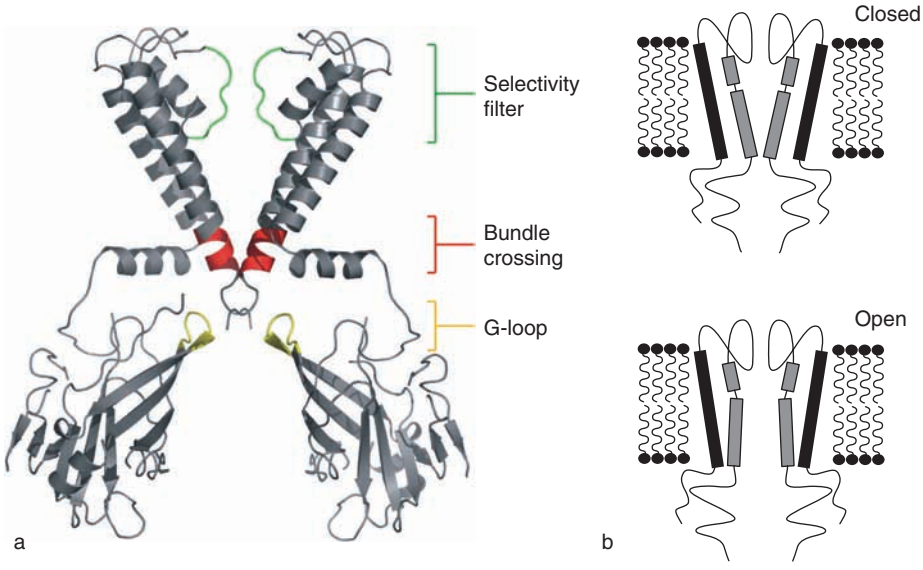

Figure 3 Ligand-dependent gating mechanisms of Kir channels: (a) structural motifs proposed to be involved in ligand-dependent gating of Kir channels; (b) sketches of structural changes involved in Kir channel opening and closing. In (a), the structural motifs have been highlighted using the KirBac1.1 crystal structure as a template. It is likely that conformational changes in the selectivity filter (green) or at the crossing point of the M2 helices (bundle crossing, red) can result in channel closure, with the bundle crossing being the primary mediator of ligand-dependent channel gating and the selectivity filter being involved in stochastic ligand-independent opening and closing. Recent work suggested that the G-loop structure (yellow) in the cytoplasmic domain is also involved in ligand-dependent gating. As shown in (b), similar to Kv channels, channel opening is thought to involve hinged motions of the M2 helix (gray) that widen the pore in the bundle-crossing region.

of ligand binding to conformational changes elsewhere in the channel.

Kir Channel Regulation by Membrane Phospholipids

A final critically important regulator of Kir channel function is the lipid composition of the plasma membrane. Although the various Kir channel subfamilies each exhibit unique ligand-dependent gating properties, a universal feature of eukaryotic Kir channels is activation by anionic membrane phospholipids such as phosphatidyl-inositide-4,5-bisphosphate (PIP2). In most cases, PIP2 or another anionic phospholipids appear to be important cofactors for channel function because experimental depletion by various measures can significantly run down channel activity; the application of exogenous PIP2 can rescue activity. Several studies have identified residues in both the N- and C-termini of Kir channels that are critical for interactions with PIP2 and other anionic lipids. Prominent are a number of basic amino acids that are critical for channel stimulation by PIP2, probably through electrostatic interaction between positively charged side chains and the negatively charged lipid head group.

Physiologically, PIP2 interactions with Kir channels modulate the apparent efficacy or potency of other ligands. In Kir6.2 channels, for example, the addition of PIP2 to the channel membrane can dramatically decrease the apparent ATP sensitivity. In this particular example, the molecular basis for this effect of PIP2 appears to be due to partial spatial overlap between the residues involved in PIP2 interactions and the ATP-binding residues. Similarly, PIP2 application or depletion can alter the Kir3 channel response to $G_{\beta\gamma}$ subunits. In Kir1.1, the introduction of positive charges at specific locations in the N-terminus can enhance PIP2 interactions and allow PIP2 application to modulate the channel response to changes in pH. Although experiments to date have demonstrated that the manipulation of PIP2 levels can dramatically alter channel function and channel interactions with other ligands, the impact of physiological variation of lipid content on Kir-channel activity in living organisms has been incompletely explored, and may provide a rich topic for future study.

See also: Potassium Channel Regulation; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Bichet D, Haass FA, and Jan LY (2003) Merging functional studies with structures of inward-rectifier K(+) channels. *Nature Reviews Neuroscience* 4: 957–967.

Domene C, Doyle DA, and Venien-Bryan C (2005) Modeling of an ion channel in its open conformation. *Biophysical Journal* 89: L01–L03.

Doyle DA (2004) Structural changes during ion channel gating. *Trends in Neuroscience* 27: 298–302.

Enkvetchakul D and Nichols CG (2003) Gating mechanism of KATP channels: Function fits form. *Journal of General Physiology* 122: 471–480.

Kuo A, Domene C, Johnson LN, Doyle DA, and Venien-Bryan C (2005) Two different conformational states of the KirBac3.1 potassium channel revealed by electron crystallography. *Structure* 13: 1463–1472.

Kuo A, Gulbis JM, Antcliff JF, et al. (2003) Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science* 300: 1922–1926.

Kurata HT, Marton LJ, and Nichols CG (2006) The polyamine binding site in inward rectifier K^+ channels. *Journal of General Physiology* 127: 467–480.

Lopatin AN, Makhina EN, and Nichols CG (1994) Potassium channel block by cytoplasmic polyamines as the mechanism of intrinsic rectification. *Nature* 372: 366–369.

Lu Z (2004) Mechanism of rectification in inward-rectifier K^+ channels. *Annual Review of Physiology* 66: 103–129.

Nichols CG (2006) KATP channels as molecular sensors of cellular metabolism. *Nature* 440: 470–476.

Nichols CG and Lopatin AN (1997) Inward rectifier potassium channels. *Annual Review of Physiology* 59: 171–191.

Nishida M and MacKinnon R (2002) Structural basis of inward rectification: Cytoplasmic pore of the G protein-gated inward rectifier GIRK1 at 1.8 Å resolution. *Cell* 111: 957–965.

Pegan S, Arrabit C, Zhou W, et al. (2005) Cytoplasmic domain structures of Kir2.1 and Kir3.1 show sites for modulating gating and rectification. *Nature Neuroscience* 8: 279–287.

Sadja R, Alagem N, and Reuveny E (2003) Gating of GIRK channels: Details of an intricate, membrane-delimited signaling complex. *Neuron* 39: 9–12.

Ion Channel Localization in Cell Bodies and Dendrites

R Westenbroek, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons are highly polarized cells made up of a soma, a large dendritic tree, and an axonal region. They have the ability to receive, process, and transmit information due to the structural and molecular uniqueness of each of these compartments. The unique complement of ion channels in the somatodendritic and axonal region is one of the important factors determining the functioning of an individual neuron. Most neurons possess dendrites that are the recipients of thousands of synaptic inputs, and the integration of these numerous inputs is required for the generation of an action potential threshold in a neuron. With the hippocampal pyramidal neuron as a model (Figure 1), this article discusses the localization of voltage-gated sodium (Na_V), calcium (Ca_V), and potassium (K_V) channels, the most clearly understood members of the voltage-gated ion channel family, and describes their role in the integration of signals in the somatodendritic region.

Subcellular Localization of Na_v Channels; Na_V1 Family

Na_V channels are specialized for electrical signaling and are responsible for the initiation and propagation of action potentials in nerve, muscle, and endocrine cells due to the rapid and transient influx of sodium ions through their pore region. Depending on their dendritic localization (Figure 1(a)), Na_V channels may be key components of excitatory postsynaptic potential amplification, dendritic spike initiation, or action potential propagation.

Within the central nervous system (CNS), investigators using immunocytochemical techniques have identified the localization of the pore-forming α subunit of $Na_V1.1$, $Na_V1.2$, $Na_V1.3$, $Na_V1.5$, and $Na_V1.6$ channel protein within distinct subcellular domains of adult rat and human neurons. $Na_V1.4$, $Na_V1.7$, $Na_V1.8$, $Na_V1.9$, and Na_X do not appear to be expressed in the CNS. $Na_V1.1$ channels are localized mainly in the cell bodies and dendrites of neurons in both rats and humans. $Na_V1.1$ channel staining is dense in the soma and proximal dendrites of pyramidal neurons in the hippocampus. Recent reports suggest $Na_V1.1$ may also be localized in axons. In contrast, $Na_V1.2$ channels are localized mainly in axons. Dense staining of axon-rich regions or layers has been observed in the hippocampus and cerebellum. $Na_V1.2$ densely stains the stratum radiatum, where the axons and terminals of Schaffer collaterals reside in CA1 and in the mossy fibers in the CA3 region. There is also dense staining of the stratum oriens in the CA1 region, where axons of the commissural/association pathway reside. These anatomical findings suggest that $Na_V1.1$ channels participate in propagating signals from the dendrites to the cell body and are presumably involved in the integration of electrical signals over the soma surface and that $Na_V1.2$ channels are important in action potential conductance along the axon. $Na_V1.3$ has a somatodendritic localization in the adult human brain. In rats a somatic, sometimes dendritic staining pattern in limited regions of the cerebral cortex, hippocampal formation, and cerebellar granule cells is observed, along with axonal staining in fiber tracts like the corpus callosum and anterior commissure. The expression of $Na_V1.3$ channels is high in embryonic brain and low in adult animals.

$Na_V1.6$ is localized predominantly in the soma and proximal dendrites of output neurons, such as the hippocampal pyramidal and granular neurons, the Purkinje cells of the cerebellum, and the pyramidal neurons of the dorsal cortex, where they are involved in the integration of synaptic input. In addition, $Na_V1.6$ channels are localized to the axon initial segment and are very dense at both central and peripheral nodes of Ranvier in adult rodents, sites known to be important for the initiation and propagation of action potentials. Human studies identified a predominantly somatodendritic distribution of $Na_V1.6$ in the projection neurons of the hippocampus, cerebral cortex, and cerebellum.

$Na_V1.5$, $Na_V1.8$, and $Na_V1.9$, the tetrodotoxin-insensitive members of Na_V channels, have not been detected in the brain by electrophysiological techniques. $Na_V1.5$ is mainly expressed in cardiac myocytes, with limited expression in the brain. In contrast, $Na_V1.7$, $Na_V1.8$, and $Na_V1.9$ are mainly expressed in dorsal root ganglion neurons located in the peripheral nervous system.

Subcellular Localization of Ca_V Channels

Ca_V channels play a pivotal role in electrical signaling. They act as signal transducers by converting depolarization of the cell membrane into an influx of calcium ions through their pore. These calcium ions in turn initiate neurotransmission, contraction, secretion, and other intracellular regulatory events. Ten human genes encode the pore-forming Ca_V α1

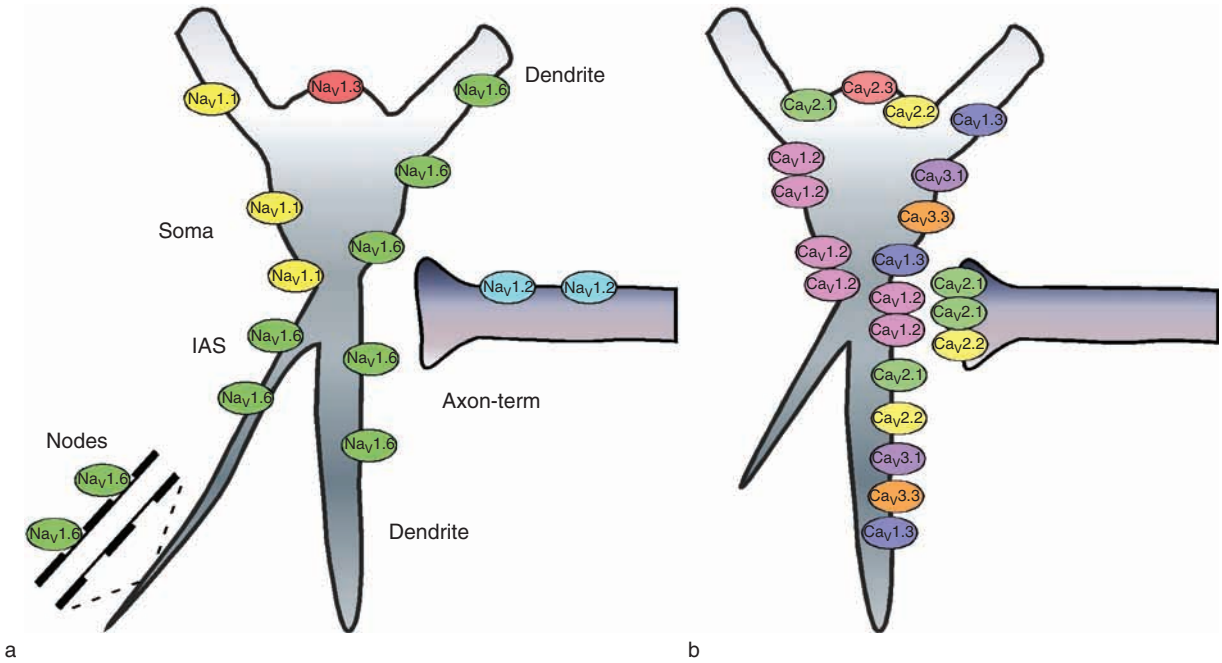

Figure 1 Schematic diagram illustrating the expression of sodium channels and calcium channels on the soma, dendrites, initial axon segment (IAS), nodes of Ranvier (Nodes) and/or local axon/terminal (axon-term) of a representative hippocampal pyramidal neuron. (a) Distribution of Na_V channels and (b) distribution of Ca_V channels.

subunits. These genes fall into three subfamilies with relatively distinct patterns of localization and function (Figure 1(b)).

Physiological studies suggest that members of the Ca_V1 and Ca_V2 subfamilies make up the high-voltage-activated (HVA) calcium channels observed in electrophysiological studies. These same electrophysiological studies provide evidence for HVA channels in dendrites of various neurons with roles in amplification of synaptic responses, generation of action potential bursts, dendritic spikes, or calcium-dependent dendritic transmitter release. In addition, entry of Ca^{2+} through dendritic Ca_V channels can serve as a second messenger in a variety of intracellular signaling pathways that may ultimately affect synaptic plasticity. Hence, HVA dendritic Ca_V channels can play important roles in synaptic integration and synaptic plasticity.

Ca_V1 Subfamily

The Ca_V1 subfamily has four members, with only $Ca_V1.2$ and $Ca_V1.3$ present in brain neurons. $Ca_V1.2$ staining is punctate while $Ca_V1.3$ channels have a smooth pattern of localization on the soma of neurons located in the hippocampus, cortex, and cerebellum. $Ca_V1.2$ channels are concentrated on the proximal dendrites while $Ca_V1.3$ is also present in more-distal dendrites in clusters at or near axosomatic and axodendritic synapses, suggesting different functional roles for each of these channels in both soma and dendrites of most neurons.

Ca_V2 Subfamily

Localization studies of the three neuronal members of the Ca_V2 subfamily of channels have revealed both synaptic and somatodendritic patterns of subcellular localization. In rat brain, $Ca_V2.1$ channel staining is prominent in regions of synaptic terminals, such as in the molecular layer of the cerebellum and in the mossy fiber region of the hippocampus, and distinct molecular isoforms are differentially localized on dendrites. Since $Ca_V2.1$ channels are known to regulate transmitter release, the prominent staining of terminals by both isoforms is most consistent with known channel function. Limited studies in human brain tissue reveal $Ca_V2.1$ protein is present in the synapse-rich molecular layer of the cerebellum and on the soma and dendrites of hippocampal pyramidal neurons.

$Ca_V2.2$ α subunits are most prevalent in synaptic structures in rat brain, such as the mossy fiber terminals and along dendrites of cortical neurons, which correlates with their physiological role in neurotransmitter release in central neurons. There is a less intense but smooth pattern of staining along the dendrites of cortical neurons in layers II/III and V and on the dendrites of pyramidal neurons in the CA1–CA3 regions of the hippocampus. From alternative splicing

at the C-terminus of $Ca_V2.2$ messenger RNA (mRNA), a long and a short isoform have been identified. Immunocytochemical studies in hippocampal cultures indicate that the short isoform is targeted to the somal and dendritic regions and the long isoform is present in the presynaptic terminals of axons. Low levels of $Ca_V2.2$ staining are present on the cell bodies of cortical pyramidal neurons, hippocampal pyramidal neurons, and Purkinje neurons. In limited human studies, $Ca_V2.2$ protein was less intense at the mossy fiber terminals, expressed at low levels in the cell bodies of the hippocampus, and stronger in synapse-rich layers (stratum lucidum and stratum radiatum) of the CA regions.

$Ca_V2.3$ channels are present throughout the rat brain and are localized mainly to the cell soma, with clear dendritic staining in some populations of cells. Staining of $Ca_V2.3$ on the cell soma has been observed in projection neurons of the hippocampus and the neocortex, and in both the cell bodies and dendrites of Purkinje neurons. The most prominent cell body and dendritic staining was in the midbrain and hindbrain structures. In human tissue, $Ca_V2.3$ is strong in the cell body and dendrites of Purkinje neurons but weaker in the cell bodies of principal neurons in the hippocampus, and relatively strong staining has been observed in the region of mossy fiber terminals. These channels have been associated with multiple physiological functions, including neurotransmitter release and repetitive firing.

Ca_V3 Subfamily

Physiological studies classify this family of Ca_V channels as the low-voltage-activated calcium channels (T-type). These studies also suggest that low voltage-activated calcium channels are expressed in the dendrites of hippocampal pyramidal neurons, thalamocortical relay neurons, and Purkinje cells, where they play important roles in burst firing and/or amplification of synaptic responses.

The subcellular distribution of $Ca_V3.2$ is still unknown; however, $Ca_V3.1$ protein is widely distributed throughout the brain, and this protein is localized in both the soma and the dendrites of large neurons in rats and mice, with some staining of the neuropil. There is a low to moderate level of $Ca_V3.1$ protein in the soma and dendrites of pyramidal neurons of the CA1–CA3 regions of the hippocampus, low levels in the Purkinje cells, and moderate levels of staining in many cortical cell types, with the strongest staining typically in the cortical pyramidal neurons. These channels are important in thalamic oscillations.

$Ca_V3.3$ immunoreactivity is localized to both the soma of neurons and their processes. In the cortex, $Ca_V3.3$ immunoreactivity has been found in all regions and in all layers. In the hippocampus, scattered interneurons have been labeled in the stratum lucidum of the dentate gyrus and in the CA3 and CA2 fields, with the highest being in the CA2 region. $Ca_V3.3$ staining of terminals is supported by dense staining of the inner and outer thirds of the molecular layer of the dentate gyrus. Dense staining is also present in the cerebellar molecular layer, where parallel fibers and their terminals reside. Similar to $Ca_V3.1$ channels, these channels have important physiological roles in thalamic oscillations.

Subcellular Localization of K_V Channels

The largest ion channel family of the VGL-chanome are the K_V channels, which are products of 40 genes in 12 subfamilies. The K_V1–9 subfamilies are the most closely related and are responsible mainly for setting the resting membrane potential, repolarization of cells during the falling phase of action potentials, termination of trains of action potentials, and hyperpolarization of cells between bursts of action potentials. Besides the K_V family, potassium channels also include the Ca^{2+}-activated (K_{Ca}), the inwardly rectifying (K_{ir}), and two-pore (K_{2P}) families (Figure 2).

In contrast to dendritic Na_V and Ca_V channels, whose presence can amplify synaptic signals, dendritic K_V channels in general tend to decrease the influence of synaptic inputs and decrease dendritic excitability. Since there are numerous subtypes of K_V channels that can occur in dendrites, these channels have been grouped electrophysiologically as either transient or sustained K^+ currents, both of which can be influenced by multiple second messengers that ultimately alter the influence of dendritic K_V channels on dendritic excitability.

K_V1 Subfamily

$K_V1.1$, $K_V1.2$, and $K_V1.4$ are the three most highly expressed subunits of the K_V1 ($K_V1.1$–1.8) family in rat brain (Figure 2(a)). $K_V1.1$ and $K_V1.2$ maintain membrane potential and modulate electrical excitability, and $K_V1.4$ participates in neuronal afterhyperpolarization. Expression of $K_V1.1$, $K_V1.2$, or $K_V1.4$ protein is mainly in axons and at or near the axon terminals. $K_V1.1$, $K_V1.2$, and $K_V1.4$ have been shown to associate in different combinations across the brain. Only $K_V1.1$ and $K_V1.4$ are present in the mossy fiber axons. There is a high density of $K_V1.1$ and $K_V1.2$ and a much lower density of $K_V1.4$ and $K_V1.$ 6 in the cerebellum. There is extremely dense labeling of

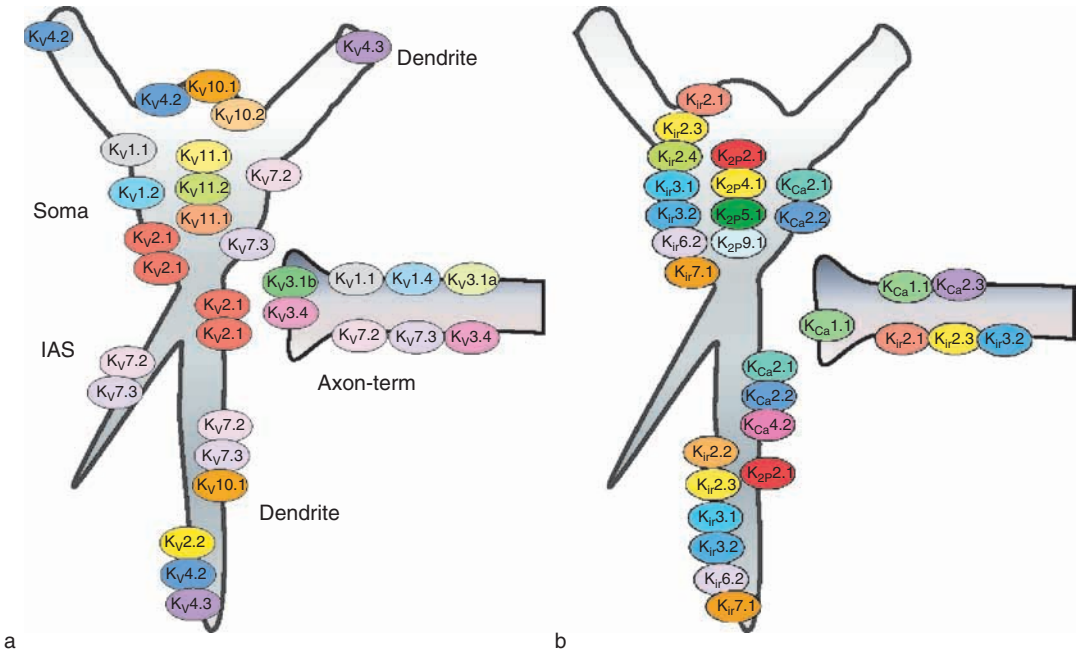

Figure 2 Schematic diagram demonstrating the localization of various potassium-selective channels on the soma, dendrites, and/or local axon/terminal (axon-term) of a representative hippocampal pyramidal neuron. (a) K_V channels and (b) K_{Ca}, K_{ir}, and K_{2P} channels: their known sites of expression.

$K_V1.1$ and $K_V1.2$ in the axon terminal plexuses of cerebellar granule cells that surround the initial segment of Purkinje cell axons. $K_V1.1$ is densely expressed in the molecular layer of the cerebellum, where afferent inputs to the cerebellar cortex are situated. $K_V1.4$ and $K_V1.6$ antibodies also weakly stain the molecular layer. There is low expression of $K_V1.1$ on the cell bodies and proximal dendrites of pyramidal neurons in the cortex, the pontine reticular nucleus, and the principal sensory trigeminal nucleus and low levels of protein staining for $K_V1.1$ and $K_V1.2$ in hippocampal neurons. One region where there is a high density of staining on the cell soma and proximal dendrites is on the mitral cells in the olfactory bulb with anti-$K_V1.2$ antibodies. $K_V1.1$ is expressed in the cell bodies of neurons in some discrete regions of the brain, including the neocortex, the pontine reticular nucleus, and the principal sensory trigeminal nucleus. However, the main staining for the K_V1 family is in axons. The majority of $K_V1.3$ staining in the brain is localized to the parallel fibers of the granule cells in the cerebellum, suggesting that $K_V1.1$ and $K_V1.3$ can form heteromeric channels in these axons.

The staining of $K_V1.6$, a regulator of membrane potential in neurons in rodent brain, is mainly in interneurons, with some staining of dendrites of principal neurons. Many of the medium to large interneurons in the CA1–CA3 fields are immunoreactive for $K_V1.6$. $K_V1.5$ has low expression in the brain and may be mainly located in glial cells; however, there are high levels of expression of $K_V1.5$ on the soma and dendrites of Purkinje neurons. $K_V1.7$ is not present in the brain but is expressed in skeletal muscle, heart, and pancreatic islets, and $K_V1.8$ is expressed in the brain; detailed subcellular localization is still lacking.

K_V2 Subfamily

In contrast to the K_V1 subfamily of potassium channels that are localized predominantly to axons, the K_V2 family is expressed in the somatodendritic compartments of neurons in the brain. $K_V2.1$ channels are localized to the majority of neurons in the brain. At the subcellular level, they are present only on the soma and proximal dendritic membranes. $K_V2.1$ channels form large clusters in these regions. Neurons with especially high levels of $K_V2.1$ expression include cortical pyramidal neurons in layers II/III and V and interneurons and neurons in the hippocampus in general. The highest expression in the hippocampus is found in the dentate granule cells and in the CA1 pyramidal neurons. $K_V2.2$ channels are localized in a smooth pattern along the length of the dendrite and are not found in clustered pattern in the proximal dendritic membranes, as is observed for $K_V2.1$ channels. $K_V2.2$ channels are found in many of the same neurons and interneurons that

contain $K_V2.1$ channels; however, the highest levels of $K_V2.2$ expression are observed in olfactory bulb neurons and in cortical pyramidal neurons.

K_V3 Subfamily

K_V3 mRNA has extensive alternative splicing that results in numerous splice variants that differ at their C-terminus. $K_V3.1a$ proteins are expressed mainly in the axons of parvalbumin-positive interneurons in the neocortex, the hippocampus, and the caudate-putamen and in neurons of the cochlear nucleus. In contrast, $K_V3.1b$ is localized to the soma and proximal dendrites of some of the same neurons as $K_V3.1a$. $K_V3.1b$ is expressed in many parvalbumin-containing basket cells but is not present in hippocampal or neocortical pyramidal neurons. Instead, these pyramidal neurons are surrounded by $K_V3.1b$-labeled presynaptic terminals from interneuron axons.

$K_V3.2$ channels are in a different population of interneurons that are characterized as nonfast-spiking, somatostatin-containing, and calbindin-containing interneurons. $K_V3.2$ proteins are prominently expressed in patches in somatic and proximal dendritic membrane, as well as in axons and presynaptic terminals of these γ-aminobutyricacidergic interneurons. To date, the precise localization of the individual splice variants of $K_V3.2$ ($K_V3.2a$, $K_V3.2b$, and $K_V3.2c$) has not been determined.

Limited immunofluorescence studies have revealed staining in the cell soma and dendrites of Purkinje neurons with $K_V3.3$ channel antibodies. $K_V3.4$ is localized to the axons and terminals of principal cells located in the hippocampus and neocortex. For example, $K_V3.4$ appears to be present in the mossy fiber axons and/or terminals, in terminals of the perforant path, and in cerebellar basket cell terminals.

K_V4 Subfamily

$K_V4.1$ is expressed minimally in the brain while the levels of mRNA expression for $K_V4.2$ and $K_V4.3$ is widespread throughout the brain. $K_V4.2$ channels are mainly in dendrites of principal cells, and $K_V4.3$ mainly in interneurons and a smaller population of principal neurons. $K_V4.2$ and $K_V4.3$ staining is present in the distal dendrites of hippocampal granule cells and the distal portion of both the basal and apical dendrites of CA3 pyramidal neurons, while only $K_V4.2$ is expressed on the soma of CA3 pyramidal neurons. $K_V4.3$ is localized to the majority of interneurons in the hippocampus. There is a distinct separation of $K_V4.2$ and $K_V4.3$ staining in the dorsal cortex, with $K_V4.2$ staining predominating in the pyramidal neurons in layer V and $K_V4.3$ mainly in the pyramidal neurons of layer II. A population of interneurons in layers II–VI of the cortex stains for $K_V4.3$.

K_V5, 6, 8, and 9 Subfamilies

The K_V5, 6, 8, and 9 subfamilies of the K_V family do not produce functional channels on their own. Instead, these subunits form heterotetramers with K_V2 family subunits. As a result, they function as modifiers and/or silencers of the K_V2 family channel and increase functional diversity within the K_V2 family.

K_V7 Subfamily

There are five known members of this subfamily, collectively known as the delayed rectifier potassium channels, or voltage-dependent potassium channels that activate slowly but show no inactivation. The $K_V7.1$ channels are not found on neurons in the CNS, whereas $K_V7.2$–$K_V7.5$ exhibit a variable distribution in the brain and are known also as the M channels that exert crucial influence over neuronal excitability. Studies of protein expression show overlapping $K_V7.2$ and $K_V7.3$ expression on both the somata and the dendrites of pyramidal neurons in the hippocampus and cerebral cortex. This staining in the somata and dendrites is low compared with axonal staining in these regions. $K_V7.2$ staining is also present in the initial segment of neurons in the hippocampus, neocortex, brain stem, and striatum and is often co-localized with $K_V7.3$ staining. There is dense expression for $K_V7.2$ on the mossy fibers in the hippocampus. $K_V7.4$ channels have been localized to brain stem auditory nuclei, outer hair cells of the cochlea, and type 1 hair cells of vestibular organs and are known to mediate efflux of potassium from outer hair cells. $K_V7.5$ channel mRNA is broadly expressed in the brain and functions to determine subthreshold excitability of neurons.

K_V10 Subfamily

In rat both $K_V10.1$ and $K_V10.2$ staining is localized to the cell somas of pyramidal neurons in the hippocampus and granule cells in the dentate gyrus. $K_V10.1$ was also detected in the surrounding neuropil, suggesting its presence along the dendrites of these neurons.

K_V11 Subfamily

By means of antibodies, high levels of protein expression for $K_V11.1$ and $K_V11.3$ were observed in the somas of hippocampal pyramidal neurons in CA1–CA3. In contrast, $K_V11.2$ localization is absent from the pyramidal neurons of the CA2 region but present in the CA1 and CA3 regions. In the cortex, $K_V11.1$ and $K_V11.3$ stain the cell bodies of neurons throughout layers II–VI, with $K_V11.3$ also staining dendrites of pyramidal neurons. $K_V11.2$ displays weak somal reactivity in the cortex except for layer V neurons. There is intense staining of the cell bodies of Purkinje

neurons with $K_V11.1$, $K_V11.2$, and $K_V11.3$ antibodies. $K_V11.2$ antibodies also stain the proximal dendrites of the Purkinje neurons.

K_V12 Subfamily

The mRNA for $K_V12.1$, $K_V12.2$, and $K_V12.3$ is abundant in many regions of the brain. Patterns of subcellular localization and physiological functions are unknown.

Subcellular Localization of K_{Ca} Channels

K_{Ca} channels are important in many cellular processes involving calcium-dependent signaling in both electrically excitable and nonexcitable cells. In addition to voltage-dependent gating, K_{Ca} channels sense other intracellular signals (i.e., Ca^{2+}, Na^+, and H^+) that mediate their function.

$K_{Ca}1$ Subfamily

In rats, high levels of $K_{Ca}1.1$ protein are targeted to regions rich in terminals and axons, such as the mossy fibers and perforant pathway of the hippocampal formation, and to terminal areas of fiber tracts in the substantia nigra pars reticulata, globus pallidum, and olfactory system (Figure 2(b)). Unlike most other brain regions, $K_{Ca}1.1$ expression is found on the cell body and proximal dendrites of Purkinje neurons. In mouse brain, $K_{Ca}1.1$ channels were found on cell bodies, dendrites, axons, and synaptic terminals. Their physiological functions in the brain include presynaptic regulation of neurotransmitter release and, when selectively coupled to with $Ca_V2.2$ channels, mediation of fast afterhyperpolarization in neurons.

$K_{Ca}2$ Subfamily

This subfamily of ion channels is made up of three members, $K_{Ca}2.1$, $K_{Ca}2.2$, and $K_{Ca}2.3$, whose mRNA patterns of distribution are both overlapping and complementary in the rat brain. The reported pattern of protein expression is somewhat variable depending on the species. In mouse, $K_{Ca}2.1$ and $K_{Ca}2.2$ staining was mainly localized to somatic and dendritic structures while $K_{Ca}2.3$ was mainly associated with varicose fibers. All three channels are present in the soma of Purkinje neurons, while only $K_{Ca}2.2$ is observed in hippocampal pyramidal neurons and layer V pyramidal neurons of the cerebral cortex. In rat, $K_{Ca}2.1$ and $K_{Ca}2.2$ staining in cell bodies was quite variable while dendritic staining was similar to that observed in the mouse. As in the mouse, staining for $K_{Ca}2.3$ was observed mainly in varicose fibers and axonal-rich regions such as the mossy fibers. Collectively, this group of channels is involved in the afterhyperpolarization in vertebrate neurons.

$K_{CA}3$ and 5 Subfamilies

These two subfamilies have one member ($K_{Ca}3.1$ and $K_{Ca}5.1$) each, and they have not been detected in the brain.

$K_{Ca}4$ Subfamily

The physiological function of this subfamily has not been identified yet. $K_{Ca}4.1$ protein staining was found mainly in the brain stem, with limited expression in the cortex and hippocampus. In the hippocampus, $K_{Ca}4.1$ immunocytochemical staining was limited to the soma and axonal projections of interneurons located in the CA3 region. $K_{Ca}4.1$ was also limited to the frontal cortex and was associated with the cell somas and axons. In the brain stem, $K_{Ca}4.1$ was localized to the cell bodies and axons. Like $K_{Ca}4.1$, $K_{Ca}4.2$ is widely distributed in the rat brain, with strong hybridization signals and immunoreactivity in the brain stem. However, protein and mRNA for $K_{Ca}4.2$ were observed in regions that did not contain $K_{Ca}4.1$ expression. These include the CA1–CA3 regions of the hippocampus, the dentate gyrus, and cortical layers II, III, and V, which suggest $K_{Ca}4.2$ functioning independent of $K_{Ca}4.1$ in some of these regions. Unlike $K_{Ca}4.1$, $K_{Ca}4.2$ is observed mainly in somatosensory cortex and visual cortex, with staining mainly on cell bodies and in fiber tracts within layer VI. Within the hippocampus, $K_{Ca}4.2$ was localized to the dendrites, with little staining of the soma. Very low levels of $K_{Ca}4.2$ staining have been associated with Purkinje neurons. In the brain stem, $K_{Ca}4.2$ staining was present mainly on cell somas or axons.

Subcellular Localization of K_{ir} Channels

K_{ir} channels have a variety of tasks, including control of the resting membrane potential, regulation of cardiac and neuronal electrical activity, and control of electrolyte balance by conduction of large inward currents at membrane voltages negative to the K^+ equilibrium potential. The mRNA for $K_{ir}1.1$, $K_{ir}3.3$, $K_{ir}4.1$, $K_{ir}4.2$, and $K_{ir}5.1$ has been detected in the brain, but studies of protein localization are still lacking.

$K_{ir}2$ Subfamily

In general, $K_{ir}2$ channels are localized in the soma and dendrites of many neurons in the CNS and are involved in maintenance of a resting membrane potential and modulation of cell excitability (Figure 2(b)). In the hippocampus there is variable expression of $K_{ir}2$ channels. $K_{ir}2.1$ protein expression is high in the granule cells, weak in the CA3 region except for the mossy fiber tract, and moderate in the pyramidal cells of the CA1–CA2 regions. $K_{ir}2.2$ is expressed mainly in

the dendritic fields of granule cells and the pyramidal neurons of CA1–CA3. The $K_{ir}2.3$ channels have moderate staining in the CA1–CA2 regions, weak staining in CA3, and weak staining of the mossy fibers. In contrast, the $K_{ir}2.4$ channels exhibit moderate immunoreactivity in the cell bodies of the CA1–CA3 pyramidal neurons and in dentate granule cells. In the cortex, $K_{ir}2.1$–2.4 channels are strongly localized to the cell soma of neurons in all cortical layers, with weaker staining in the dendritic processes. In the cerebellum, $K_{ir}2.1$ channels are weakly expressed in the soma of neurons in the molecular layer, $K_{ir}2.2$ channels have a strong pattern of staining in the soma of granule cells, and $K_{ir}2.3$ protein is intensely expressed in soma and dendrites of Purkinje neurons, but $K_{ir}2.4$ expression is extremely low, if present. In general, $K_{ir}2.1$ is localized in particular subsets of neurons throughout the whole brain, with highest density of staining in the olfactory bulb and superior colliculus. $K_{ir}2.2$ staining is highest in the cerebellum, hypothalamus, and spinal cord while $K_{ir}2.3$ is predominately expressed in the cortex, Purkinje cells, basal ganglia, and olfactory bulb. $K_{ir}2.4$ is expressed at low levels in most neurons.

$K_{ir}3$ Subfamily

This subfamily is distinquished by being G-protein-coupled members of the K_{ir} group of channels. $K_{ir}3.1$ immunoreactivity is located in selected regions of the rat brain, including cerebral cortex, cerebellum, hippocampus, thalamus, and the olfactory system. There is staining of $K_{ir}3.1$ in the soma and dendrites of granule cells in the cerebellum, and it is absent from Purkinje neurons. In the cerebral cortex, $K_{ir}3.1$ staining is complex, with labeling of the neuropil in layer IV, labeling in the soma and apical dendrites of layer V pyramidal neurons, and labeling of dendritic terminals in layer I which originate from the pyramidal cells of layers II, III, and V. In the hippocampus, $K_{ir}3.1$ staining is present in the soma and apical dendrites of pyramidal neurons and the granule cells of the dentate gyrus. These channels are responsible for receptor-dependent hyperpolarization of membrane potential. $K_{ir}3.2$ immunoreactivity is observed in many regions of the brain and is mainly in the somatodendritic subcellular compartment and sometimes in axonlike fibers and they participate in the formation of the slow inhibitory postsynaptic potential and probably in presynaptic inhibition. Highest density of staining was observed in the hippocampus, septum, granule cell layer of the cerebellum, amygdala, and substantia nigra pars compacta. In the cortex there is weak staining of $K_{ir}3.2$ neuronal cell bodies, with stronger staining in the large cell bodies and apical dendrites of layer V pyramidal neurons within the sensorimotor cortex and in layer V cell bodies adjacent to the white matter. Intense staining of the neuropil was evident in the hippocampal formation, along with labeling of the cell bodies and proximal dendrites of the CA1–CA3 pyramidal neurons and dentate granule cells. In the cerebellum there are low levels of staining of the granule cells and Purkinje cells and strong staining of the cerebellar glomeruli.

With $K_{ir}3.4$ antibodies, there is mainly staining of cell bodies and dendrites. In the cortex, $K_{ir}3.4$ was observed only in cells in the deepest portion of layer VI, in cells located in the medial habenula, posterior hypothalamus, superior colliculus, and restricted nuclei of the forebrain. The cell bodies and proximal dendrites of Purkinje neurons also stain positive for $K_{ir}3.4$. mRNA studies suggest the presence of $K_{ir}3.4$ in rat and human hippocampal neurons while protein studies have found a lack of expression in the hippocampus.

$K_{ir}6$ Subfamily

$K_{ir}6.1$ channels are present in astrocytes in most regions of the brain and are expressed in the cell soma of only a few distinct nuclei or neuronal populations at significant levels in cholinergic interneurons in the striatum, the dorsal medial and paraventricular hypothalamic nuclei, the zona incerta, and the substantia nigra pars compacta. In the hippocampus, cortex, and cerebellum, $K_{ir}6.1$ staining is weak to negligible in the soma of neurons. $K_{ir}6.2$ is broadly distributed in the brain and is mainly observed in the soma of neurons and their main dendrites. These channels are thought to act as an oxygen and glucose sensor in the brain and are involved in cytoprotection during cardiac and brain ischemia. $K_{ir}6.2$ is present in layer V pyramidal cells and their dendrites, the cell soma and dendrites of dentate granule cells, CA1–CA3 pyramidal cells of the hippocampus, and Purkinje neurons of the cerebellum. Some regions of the brain also show strong staining of the neuropil, suggestive of axonal staining.

$K_{ir}7$ Subfamily

There is only one known member of this family, and limited studies of the protein distribution of this channel have determined it to be localized to the soma and dendrites of cerebellar Purkinje neurons and pyramidal cells of the hippocampus, where it contributes to resting membrane potential.

Subcellular Localization of K_{2P} Channels

In general, mammalian K_{2P} channels are thought to be responsible for the background or leak potassium currents in neurons and function in a highly regulated process that controls cellular excitability by stabilizing membrane potential below firing threshold and

expediting repolarization. The physiological function of most of these channels is still poorly understood. *In situ* studies have demonstrated the presence of mRNA in the CNS for $K_{2P}1.1$, K_{2P} 6.1, $K_{2P}7.1$, $K_{2P}12.1$, $K_{2P}13.1$, and $K_{2P}15.1$, but their precise subcellular distribution is still unknown. $K_{2P}10.1$, $K_{2P}16.1$, and $K_{2P}17.1$ channels are not present in the CNS.

$K_{2P}2.1$ Channels

Immunocytochemical studies suggest $K_{2P}2.1$ protein is localized to the soma and dendrites of both projection neurons and interneurons that are mainly γ-aminobutyric acid positive. Immunoreactivity is widespread in the CNS, with staining observed in such regions as cerebral cortex (concentrated in layers III and VI), basal ganglia, hippocampus (CA2–CA3 and granule cells in dentate gyrus), and cerebellum (both Purkinje cells and Golgi cells in the granular layer).

$K_{2P}3.1$ Channels

Immunocytochemical studies in rat suggest that $K_{2P}3.1$ staining is mainly in the somas of cerebellar granule cells and the Purkinje neurons. Neuropil staining was also present in the cerebral cortex, basal ganglion, thalamic nuclei, and amygdala, suggesting expression in axons or dendritic processes.

$K_{2P}4.1$ Channels

Immunocytochemical studies using anti-$K_{2P}4.1$ antibodies in mouse brain revealed a broad pattern of staining throughout the brain, with highest levels of staining in the cells located in the cerebellum, neocortex, hippocampus, dentate gyrus, thalamus, hypothalamus, caudate, putamen, and olfactory bulb. The $K_{2P}4.1$ staining was observed mainly in the soma of neurons and to a lesser degree in the axons and/or dendrites of most neurons.

$K_{2P}5.1$ Channels

In studies using anti-$K_{2P}5.1$ antibodies, staining was mostly localized to cell bodies and was most prominent in the pyramidal cells of the hippocampus, granule cells of the dentate gyrus, most neurons in the cortex, and the soma and proximal dendrites of Purkinje neurons.

$K_{2P}9.1$ Channels

In human tissue, immunocytochemical studies suggest strong cell body staining of Purkinje neurons and less-intense staining in the granule cells of the cerebellum. In rats, antibody staining has revealed $K_{2P}9.1$ expression in the cell bodies of neurons located in the hypoglossal motor nucleus, locus coeruleus, anterodorsal thalamic nucleus, dorsal raphe nucleus, and striatum of the rat. Others report expression of $K_{2P}9.1$ protein in the soma of neurons located in the rat hypothalamus, hippocampus, and a subset of cortical neurons.

See also: Calcium Channels; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Dendritic Signal Integration; Inwardly Rectifying Potassium Channels; Ion Channel Localization in Axons; Ionic Channels in Glia; Large Conductance Calcium-Activated Potassium Channels; Neuromodulation of Sodium Channels; Neuromodulation of Calcium Channels; Potassium Channel Regulation; Sodium Channels; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Catterall WA, Goldin AL, and Waxman SG (2005) International union of pharmacolgy. XLVII. Nomenclature and structure–function relationships of voltage-gated sodium channels. *Pharmacological Reviews* 57: 397–409.

Catterall WA, Perez-Reyes E, Snutch TP, and Striessnig J (2005) International union on pharmacology. XLVIII. Nomenclature and structure–function relationships of voltage-gated calcium channels. *Pharmacological Reviews* 57: 411–425.

Goldstein SAN, Bayliss DA, Kim D, Lesage F, Plant LD, and Rajan S (2005) International Union of Pharmacology. LV. Nomenclature and molecular relationships of two-P Potassium channels. *Pharmacological Reviews* 57: 527–540.

Gutman GA, Chandy KG, Grissmer S, et al. (2005) International union of pharmacology. LIII. Nomenclature and molecular relationships of voltage-gated potassium channels. *Pharmacological Reviews* 57: 473–508.

Kubo Y, Adelman JP, Clapham DE, et al. (2005) International union of pharmacology. LIV. Nomenclature and molecular relationships of inward rectifying potassium channels. *Pharmacological Reviews* 57: 509–526.

Trimmer JS and Rhodes KJ (2004) Localization of voltage-gated ion channels in mammalian brain. *Annual Review of Physiology* 66: 477–519.

Wei AD, Gutman GA, Aldrich R, Chandy KG, Grissmer S, and Wulff H (2005) International union of pharmacology. LII. Nomenclature and molecular relationships of calcium-activated potassium channels. *Pharmacological Review* 57: 463–472.

Yu FH and Catterall WA (2004) The VGL-Chanome: A protein superfamily specialized for electrical signaling and ionic homeostasis. *Science STKE* 253: re15.

Yu FH, Yarov-Yarovoy V, Gutman GA, and Catterall WA (2005) Overview of molecular relationships in the voltage-gated ion channel superfamily. *Pharmacological Reviews* 57: 387–395.

Relevant Website

http://www.brain-map.org – Allen Institute for Brain Science.

Ion Channel Localization in Axons

M N Rasband, Baylor College of Medicine, Houston, TX, USA
J S Trimmer, University of California at Davis, Davis, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Types of Axonal Ion Channels

Ion Channel Overview

Voltage-gated ion channels are multisubunit protein complexes that respond to changes in membrane potential with conformational changes that lead to gating, or opening and closing, of an ion-selective transmembrane pore. Voltage-gated ion channels selective for sodium (Na_v channels) and potassium (K_v channels) ions are the major determinants of axonal electrical transmission, although other ion channels play additional roles in axonal physiology. The past 20 years has seen major advances in our understanding of the structure and function of the Na_v and K_v channels that are important to the function of mammalian axons. Also, regulatory motifs on Na_v and K_v channel proteins involved in mediating interactions with a diverse set of axonal proteins have been identified. These axonal proteins in turn participate in macromolecular complexes that in certain cases may also mediate intercellular interactions crucial to the generation and maintenance of cell–cell interactions.

Axonal Na_v Channels

Axonal Na_v channels provide all of the inward depolarizing current that initiates and propagates the orthodromic action potential. As reviewed elsewhere in this encyclopedia, Na_v channels comprise principal α subunits and auxiliary β subunits. There exist in the mammalian genome genes encoding nine different principal α subunits, $Na_v1.1$–$Na_v1.9$. Of these, $Na_v1.4$ and $Na_v1.5$ are predominantly expressed in skeletal muscle and heart, respectively. All of the remaining Na_v1 α subunits have been described on axons. $Na_v1.1$, $Na_v1.2$, $Na_v1.3$, and $Na_v1.6$ are the predominant Na_v channel α subunits expressed in central neurons, with $Na_v1.3$ being found at higher levels in neonates than in adult. $Na_v1.7$, $Na_v1.8$, and $Na_v1.9$ are the predominant Na_v channel α subunits expressed in specific classes of sensory peripheral neurons. However, $Na_v1.6$ also makes a significant contribution in adult peripheral neurons, and $Na_v1.2$ and $Na_v1.3$ make a significant contribution in younger animals.

The subcellular localization of Na_v channel α subunits differs between neuronal types and in different compartments of the same neurons. $Na_v1.2$ is expressed at high levels in unmyelinated axons in many brain regions, such as mossy fiber axons arising from hippocampal dentate granule cells and the parallel fibers of cerebellar granule cells. $Na_v1.6$ is found in myelinated fibers, being the predominant Na_v channel α subunit present at nodes of Ranvier of adult myelinated axons (Figure 1, green). $Na_v1.6$ is also present at terminal heminodes (the last segment of the axon prior to the nerve terminal proper) and at the axon initial segment (AIS; Figure 2, green). As such, $Na_v1.6$ is crucial to the first, last, and intermediate steps of axonal action potential conduction. In certain central neurons, there exists a subcompartmentalization of the AIS, with $Na_v1.1$ present in a small region more proximal to the soma, and $Na_v1.6$ more distal and over the bulk of the initial segment. $Na_v1.1$ is also expressed in somatodendritic domains, although most dendritic staining is associated with proximal and not distal dendrites. A small percentage of nodes of Ranvier in adult myelinated axons contain $Na_v1.2$. The situation is reversed in early development and/or in the absence of intact myelin, when $Na_v1.2$ is present at the majority of nodes of Ranvier. $Na_v1.2$ is also found in high levels at the AIS of developing neurons, but unlike the situation at the node of Ranvier, it is developmentally replaced by $Na_v1.6$ independent of myelination.

Axonal K_v Channels

K_v channels are critical to regulating the excitability of the axon, by setting the resting membrane potential, and in action potential repolarization. K_v channels are also important in suppressing rebound back-propagating or antidromic action potentials by their presence at initial and terminal segments of axons. The mammalian genome contains genes encoding 32 different K_v channel α subunits that have been grouped into 12 different subfamilies (K_v1–K_v12). K_v channel complexes generally contain homotetrameric or heterotetrameric assemblies of α subunits from within the same K_v channel subfamily. The exception may be so-called 'electrically silent' α subunits from the K_v5, K_v6, K_v8, and K_v9 subfamilies, which in heterologous cells do not form functional channels in themselves but assemble promiscuously with functional α subunits from other K_v families to impact the function of the resultant complexes. K_v channel complexes can also contain cytoplasmic or transmembrane auxiliary subunits. The precise

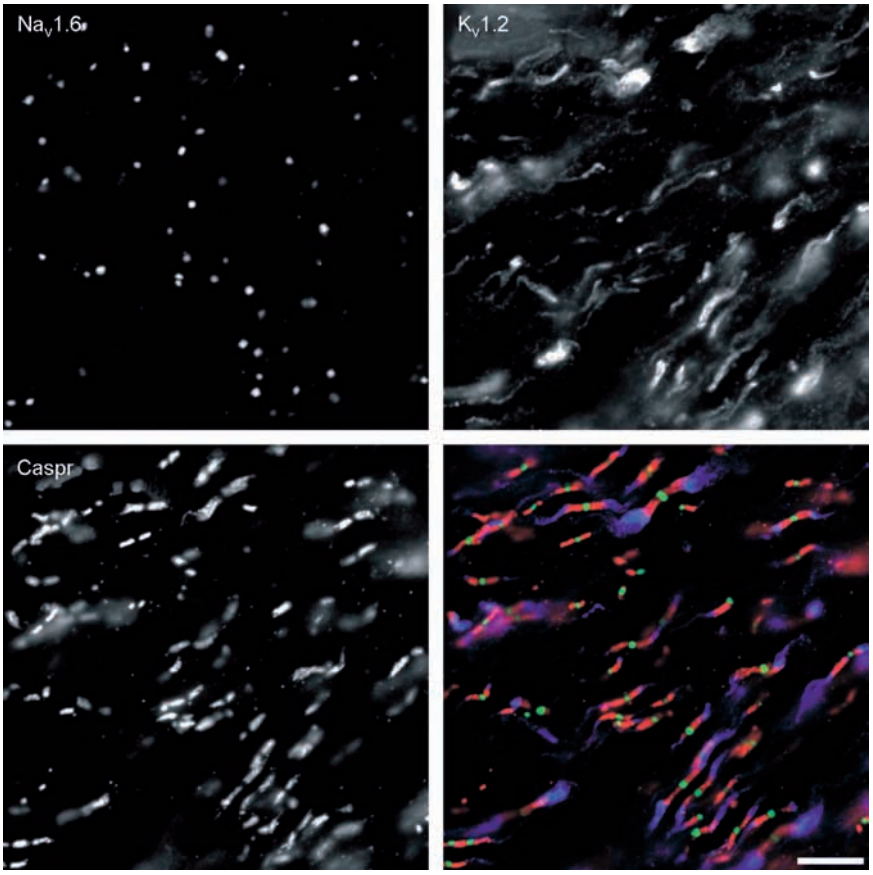

Figure 1 Rat optic nerve triple immunofluorescence stained for $Na_v1.6$ (green), $K_v1.2$ (blue), and contactin-associated protein (Caspr) (red). Scale bar = 20 μM.

stoichiometry of K_v channel principal and auxiliary subunits determines the intracellular trafficking and localization, biophysical properties, and pharmacology of the resultant channel complexes.

Of these 12 subfamilies of K_v channels, it is clear that K_v1, K_v3, and K_v7 family members are expressed on axons. The seven members of the K_v1 channel α subunit subfamily, of which $K_v1.1$–$K_v1.6$ are expressed in neurons, form low-voltage activated channels that are present in virtually all mammalian axons. K_v1 channels are predominantly found as heterotetramers assembled from more than one K_v1 α subunit (e.g., $K_v1.1 + K_v1.4$). K_v1 channel complexes commonly contain cytoplasmic $K_v\beta$ subunits. $K_v1.1$, $K_v1.2$, and $K_v\beta2$ are a prominent combination and found at juxtaparanodes of myelinated axons (Figure 1, blue), along unmyelinated axons, in nerve terminals or preterminal axon segments, and at AISs. Based on this subunit composition, these K_v1 α and $K_v\beta$ subunits would be expected to form low-voltage-activated, non- or slowly inactivating channels sensitive to the mamba snake neurotoxin α-dendrotoxin. Complexes of $K_v1.1$, $K_v1.4$, $K_v\beta1$, and $K_v\beta2$ predominate in unmyelinated axons, nerve terminals, and preterminal axon segments in many brain areas. These subunits would be expected to form low-voltage-activated, somewhat rapidly inactivating channels also sensitive to α-dendrotoxin. These channels would be expected to exhibit use-dependent suppression due to the rapid onset of and slow recovery from inactivation; this process may be important in the frequency-dependent facilitation of neuronal firing. More complex mixtures of subunits (e.g., $K_v1.1$, $K_v1.2$, $K_v1.4$, and $K_v\beta2$) are also present on certain axons. All of these axonal K_v1 channels are targets of neuromodulation by a diverse array of neurotransmitters.

The four members of the K_v3 channel α subunit subfamily ($K_v3.1$–$K_v3.4$) are all expressed in mammalian brain, with $K_v3.1$–$K_v3.3$ most prominent. Transcripts from each of the K_v3 channel α subunit genes are subjected to alternative splicing to generate K_v3 channel α subunits with alternative C-termini. Such splicing can impact subcellular localization and sensitivity to neuromodulation. K_v3 channel α subunits form homomeric and heteromeric high-voltage activated channels that, given their high threshold for voltage-dependent activation, are probably only activated in response to action potentials. Somatodendritic K_v3 channels are critical to setting neuronal

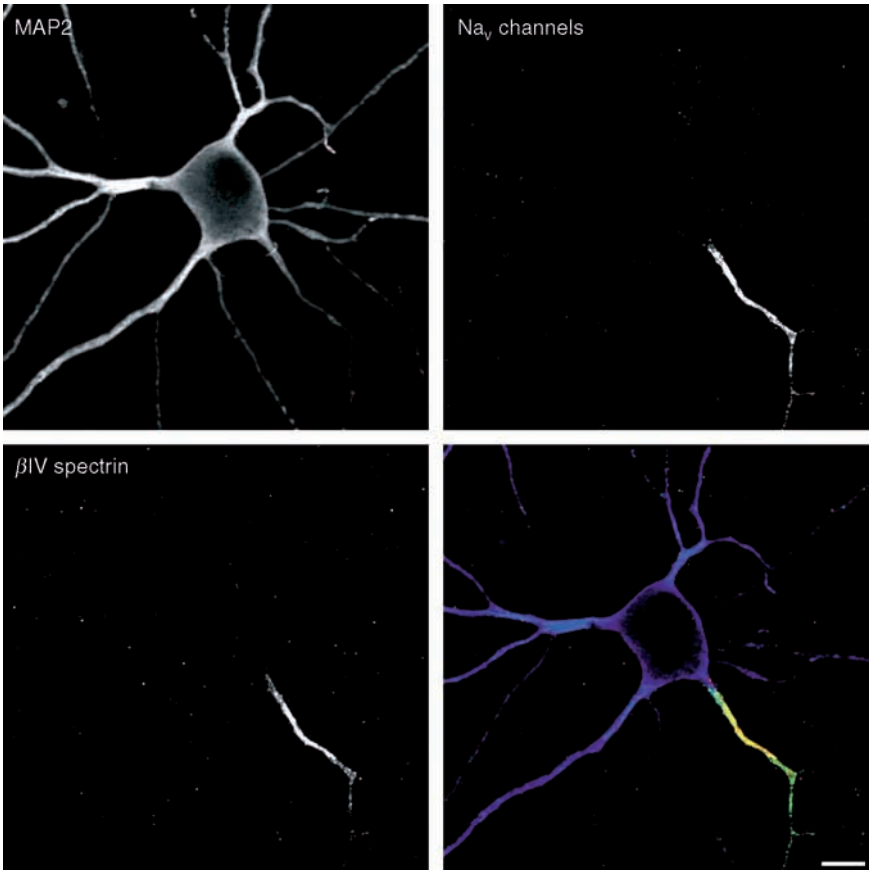

Figure 2 Cultured hippocampal neurons triple immunofluorescence stained for MAP2 (blue), Na_v channels (green), and βIV spectrin (red). Scale bar = 10 mM.

firing rates; in fact, virtually all fast-spiking cells in brain express somatodendritic K_v3 channels. The $K_v3.1b$ isoform is found predominantly on axons and in terminal fields. $K_v3.2$ and $K_v3.4$ are also present in certain axons and terminal fields, most prominently in basket cell terminal pinceau. Unlike K_v1 channels, which are generally found at the base of presynaptic terminals and not in the terminals, K_v3 channels are found in the nonrelease face of the nerve terminal. $K_v3.1b$ is also found at a subset of central nervous system nodes of Ranvier. Interestingly, $K_v3.1b$ can also be found at peripheral nervous system (PNS) nodes in an animal model of Charcot–Marie–Tooth disease type IA, a demyelinating neuropathy, suggesting that myelination may regulate the localization and/or expression of these channels.

There are five members of the K_v7 subfamily of K_v α subunits, of which four ($K_v7.2$, $K_v7.3$, $K_v7.4$, and $K_v7.5$) are expressed in neurons (these are referred to as KCNQ2, KCNQ3, KCNQ4, and KCNQ5, respectively, in many reports). K_v7 α subunits are expressed in both the somatodendritic and the axonal compartments, with especially prominent expression of $K_v7.2$ and $K_v7.3$ observed in axons. Within the axon, K_v7 α subunits exhibit a restricted distribution, with robust expression of $K_v7.2$ and $K_v7.3$ at AISs and at nodes of Ranvier, coinciding with that for Na_v channels. K_v7 channels are especially sensitive to neuromodulation, which is achieved through a wide variety of mechanisms.

Other Axonal Ion Channels

Large conductance calcium and voltage-activated potassium, BK, or $K_{Ca}1.1$ channels are formed as tetramers of principal α subunits encoded by a single mammalian gene, Slo1. However, Slo1 transcripts are subjected to extensive alternative splicing that generates distinct C-termini that influence diverse aspects of $K_{Ca}1.1$ channels, including subcellular localization. $K_{Ca}1.1$ channels also contain transmembrane β subunits that modulate expression, localization, and function. $K_{Ca}1.1$ α subunits are found robustly expressed on axons and within nerve terminals, where they co-localize with voltage-gated calcium or Ca_v channels in the active zone. There is no evidence

to suggest that $K_{Ca}1.1$ channels are enriched at other axonal sites (e.g., initial segment and nodes of Ranvier).

Small conductance calcium and voltage-activated potassium, SK, or $K_{Ca}2$ channels are formed as tetramers of three different principal α subunits – $K_{Ca}2.1$ (SK1), $K_{Ca}2.2$ (SK2), and $K_{Ca}2.3$ (SK3) – each encoded by different genes. Robust $K_{Ca}2.3$ expression is present in mossy fiber and recurrent hilar axons and/or presynaptic terminals of hippocampal dentate granule cells. No other SK3 protein is detected in hippocampus. $K_{Ca}2.1$ is expressed in mossy fiber axons/terminals but is also present at high levels on the dendrites of many mammalian brain neurons. $K_{Ca}2.2$ is predominantly somatodendritic.

Slack, Slo2.2, or $K_{Ca}4.1$ α subunits form tetrameric pH-activated potassium channels and are strongly localized to axons and nerve terminals of expressing central neurons. $K_{Ca}4.1$ is expressed at high levels in brain stem neurons, such as in trigeminal and auditory nuclei, where it is expressed in axons and in large Calyx of Held terminals of the ventral cochlear nucleus globular bushy cells. Slick, Slo2.1, or $K_{Ca}4.2$ α subunits form tetrameric sodium-activated potassium channels that are found in a variety of neuronal axons, although robust staining is also observed in somatodendritic compartments.

Inwardly rectifying potassium or K_{ir} channels have not been described at the immunohistochemical level as being prominent on axons, nor have cyclic nucleotide gated or CNG channels. Some transient receptor potential (Trp) channels have been localized to axons. For example, TrpV1 is highly enriched in nociceptive C-fibers, and TrpC5 is involved in axon guidance through effects on calcium influx into growth cones.

There are four distinct mammalian hyperpolarization activated cation or HCN channels, HCN1–HCN4. These form homotetrameric channels, and they may also form heterotetramers. Of the HCN channels, only HCN1 has been reported in axons and terminals. The best example is in cerebellar basket cell terminals that contain high levels of HCN1, although there also exists prominent somatodendritic staining in certain neurons. Other HCN channels exhibit prominent somatodendritic staining.

Voltage-gated calcium or Ca_v channels also exhibit pronounced targeting to nerve terminals where they are integral components of release machinery and mediate calcium influx triggering vesicle fusion and transmitter release. The role of Ca_v channels in the axon proper is less clear.

Figure 3 provides a summary of the various axonal ion channels and their localization as described previously.

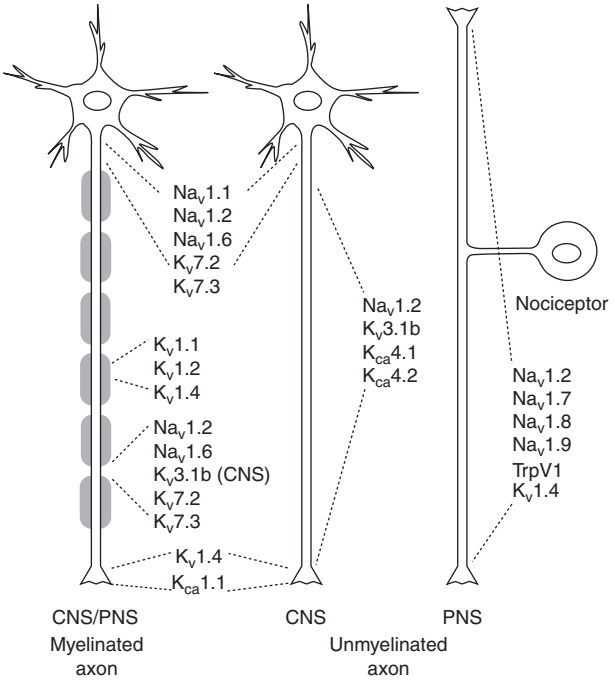

Figure 3 Schematic summary of the types and localization of axonal ion channels. CNS, central nervous system; PNS, peripheral nervous system.

Determinants of Ion Channel Localization

Intrinsic Determinants for Polarized Targeting of Ion Channels to Axons

A number of potential mechanisms exist to achieve localization of ion channels in axons, including direct targeting of channel-containing transport vesicles exclusively to the axon, general transport with selective insertion at axonal sites, and general transport and insertion with selective retention on axons and/or selective retrieval from dendrites. Which of these mechanisms is used for determining the axonal localization of the diverse ion channels cataloged previously is not clear. Work has begun to provide insights into molecular mechanisms leading to trafficking of ion channels to axons. Studies of Na_v channels by Dargent and colleagues have identified a motif that is in some contexts sufficient for ion channel targeting to axons. These workers made chimeric type I membrane proteins by appending different cytoplasmic fragments of axonal $Na_v1.2$ to the extracellular and transmembrane domains of CD4. They also made similar constructs using the analogous regions of a type II membrane protein, the transferrin receptor (TfR). Wild-type CD4 and TfR have nonpolarized distributions on both axons and dendrites. Cytoplasmic fragments of $Na_v1.2$, derived from the linkers between internally repeated domains I–II and

II–III, did not support the axonal localization of the resultant chimeras. However, inclusion of the cytoplasmic C-terminus of $Na_v1.2$ led to axonal localization of chimeras built on both CD4 and TfR backbones, due to the presence of a nine-amino-acid motif (CLDILFAFT) that contained a dileucine-like (IL) motif. Interestingly, even though this motif is 100% conserved in all neuronal Na_v channels (including $Na_v1.1$ and $Na_v1.6$), appending the C-terminus of $Na_v1.1$ and $Na_v1.6$ to CD4 did not yield a polarized expression on the axon. This suggests that other motifs found on the C-terminus of $Na_v1.2$, but not $Na_v1.1$ or $Na_v1.6$, are necessary for axonal localization of the reporter chimeras (although native $Na_v1.1$ and $Na_v1.6$ are both localized on axons). The role of the dileucine-like motif in axonal localization is in directing the selective endocytosis of the chimeras from the somatodendritic but not axonal plasma membrane. A second motif, in the cytoplasmic linker between internally repeated pseudosubunit II–III, also appears to mediate preferential somatodendritic endocytosis.

Two different C-terminal motifs can lead to axonal expression of CD4–K_v7 chimeras. As such, they are somewhat analogous in function to the C-terminal domains that direct axonal localization of Na_v channels. Interestingly, the K_v7 C-terminal domains have previously been shown to be required for subunit assembly, a requisite step for intracellular trafficking of multisubunit proteins. The mechanism (directed targeting or selective axonal retention/dendritic retrieval) has not been defined.

Axonal targeting of K_v1 channels has also been studies using a chimeric approach. In this case, the N-terminal T1 domain, which is highly conserved among K_v1 α subunits and directs subfamily-specific assembly, is also sufficient to direct axonal localization of CD4 and TfR chimeras. Unlike the case for Na_v and K_v7 motifs, the structure of the T1 domain has been solved by X-ray crystallography. Mutation of certain residues along a face of the domain that mediate the interaction with cytoplasmic auxiliary $K_v\beta$ subunits leads to loss of axonal localization of CD4 chimeras, consistent with a previous report that showed that $K_v\beta$ subunit coexpression was required for efficient axonal localization of exogenous K_v1 α subunits. The mechanism for T1-mediated localization of the chimeras to axons appears to at least in part be based on selective trafficking to axons because transport vesicles containing TfR are exclusively targeted to dendrites, whereas those containing the TfR–T1 chimera are found in both axons and dendrites. It is intriguing that the domains on K_v1 and K_v7 channels that can confer axonal localization to CD4 and TfR are unrelated at the level of primary structure but are both involved in assembly of these tetrameric K_v channel complexes.

Intrinsic Determinants for Clustering of Ion Channels at Precise Sites

Axonal ion channels can exhibit highly specific localizations within the axon (**Figures 1** and **2**). This suggests that subsequent to the mechanisms that lead to a biosynthetic enrichment of newly synthesized ion channels in the axonal plasma membrane, mechanisms must exist to selectively insert channel-containing vesicles at specific axonal sites and/or to capture plasma membrane ion channels at preferred axonal sites or selectively retrieve them from nonpreferred axonal sites. One of the best studied examples of the motifs that direct ion channels to specific axonal membrane domains is for the AIS because unlike other domains that require specific and complex cell–cell interactions (e.g., nodes of Ranvier), this structure forms spontaneously in pure cultures of neurons. Using the chimera approach described previously for defining axonal localization signals, a second motif on Na_v α subunits, located in the interdomain II–III linker, was identified that directed the enhanced targeting of $Na_v1.2$-containing chimeras to the AIS and its interaction with the scaffolding protein ankyrin-G (ankG). This motif (VPIALGESD) is highly conserved in all Na_v channels, and although not all axonal Na_v channels are restricted to sites enriched in ankG, they all bind to ankG *in vitro*. As such, we are left with a situation similar to the C-terminal dileucine-like motif; it is sufficient to direct a specific localization in certain contexts but it must be acting in conjunction with other unidentified determinants to yield the patterns of localization observed *in vivo*. Subsequently, an ankyrin-binding motif similar to that described in Na_v channels was identified on K_v7 channels, where it also directs localization to the initial segment. The importance of ankG for Nav and K_v7 channel clustering is further highlighted by the observation that Purkinje neurons lacking ankG fail to cluster these ion channels at the AIS.

Intrinsic determinants for localization of ion channels at specific sites outside of the initial segment are less well defined. K_v1 α subunits have at their C-termini active binding motifs (S/TXV) for class I PDZ domains, such as are present on membrane-associated guanylate kinases (MAGUKs) including PSD-95 and SAP97. PSD95 co-localizes with clustered $K_v1.1$, $K_v1.2$, and $K_v\beta$ subunits at juxtaparanodes of nodes of Ranvier of myelinated axons, as well as in cerebellar basket cell terminals. Thus, it seems likely that PSD95 plays some role in localization of K_v1 channels at these sites. However, genetic elimination of detectable PSD95 expression in transgenic mice does not affect K_v1

channel localization at these sites. Although there is no obvious upregulation of other MAGUKs at juxtaparanodes, it may be that other PDZ-containing proteins, or other less obvious compensatory mechanisms, are brought into play in these animals.

Extrinsic Determinants for Clustering of Ion Channels at Precise Sites

Rapid action potential propagation in mammals is a consequence of myelin. The traditional view has been that myelin passively alters the electrical properties of the axon by increasing the transverse membrane resistance and decreasing the membrane capacitance. Together, these properties limit the current necessary to depolarize the axonal membrane. However, it is now clear that myelin also actively regulates and promotes action potential propagation by directing a complex set of interactions (neuron–glia interactions) that ultimately result in the clustering of ion channels at and near nodes of Ranvier. Thus, in addition to the various intrinsic mechanisms described previously, the node of Ranvier is an outstanding example illustrating how complex intercellular interactions can modulate or even direct the localization of axonal ion channels. Most important, the active regulation of ion channel localization has profound effects on the biophysical and electrical properties of the axon.

As at the AIS, nodes of Ranvier have high densities of Na_v (primarily $Na_v1.6$) and K_v7 channels. Furthermore, both the AIS and the nodes contain a set of cell adhesion molecules (CAMs), including a neuronal splice variant of neurofascin (NF) called NF-186, and a set of cytoskeletal and scaffolding proteins, including ankG and βIV spectrin. Thus, nodes of Ranvier and initial segments share a common molecular organization. Despite these common constituents, one major difference between the nodes and the AIS is that the clustering of Na_v and K_v7 channels at the AIS only requires interaction with ankG, but clustering of these same proteins at nodes of Ranvier depends on the extrinsic influence of myelinating Schwann cells or oligodendrocytes in the peripheral and central nervous systems, respectively. How are ion channels clustered at nodes of Ranvier, and what is the role of myelinating glia in the formation of nodes?

Studies have shown that in the PNS, node formation is initiated by the recruitment and clustering of axonal NF-186 by a glial CAM called gliomedin. Axonal NF-186 binds to gliomedin, which is located at the tips of myelinating Schwann cells on small finger-like processes called microvilli. Binding to gliomedin causes NF-186 to accumulate in the axonal membrane beneath the microvilli. NF-186 then functions as a membrane attachment site for the recruitment of ankG; ankG binds to a C-terminal FIGQY motif conserved in all L1 CAMs. After ankG is correctly positioned along the axon, the process of ion channel clustering is thought to proceed in a manner similar to that for AISs: ankG functions as a scaffold, interacting with and anchoring Na_v and K_v7 channels to nascent nodal membrane domains, as well as the cytoskeletal protein βIV spectrin. βIV spectrin links the entire ion channel complex to the actin-based cytoskeleton. Ablation of either gliomedin or NF-186 results in failure to cluster ankG, βIV spectrin, and Na_v channels, emphasizing the crucial role that neuron–glia interactions play in defining sites of ion channel clustering in axons.

Nodes of Ranvier are flanked by the paranodal or axoglial junctions and are the locations where the myelin sheath attaches to the axon (**Figure 1**, red). These domains are composed of a tripartite CAM complex comprising axonal contactin and contactin associated protein (Caspr), as well as a glial splice variant of NF called NF-155. Together, this protein complex is thought to function as an adhesive attachment site for the myelin membrane. The paranodal junctions also contribute to ion channel function and localization by acting as both a membrane protein diffusion barrier and a paracellular ion diffusion barrier. Disruption of paranodal junctions results in mislocalization of nodal and juxtaparanodal ion channels.

Immediately adjacent to the paranodal junctions and beneath the myelin sheath is a domain called the juxtaparanode (**Figure 1**, blue). This region of the axon extends for 10–20 μm toward the internode and is characterized by a high density of K_v1 channels, including $K_v1.1$, $K_v1.2$ and, rarely, $K_v1.4$. The functions of these channels are not entirely clear, although data suggest they may play important roles during development in limiting repetitive firing of nodes due to re-entrant excitation of the nodal membrane. During disease, however, it is clear that these channels are functional since uncovering or redistribution of these channels from beneath the myelin sheath dramatically limits axonal excitability. Indeed, K_v1 channel blockers are currently in use to restore conduction in multiple sclerosis and spinal cord injury. Regarding the Na_v and K_v7 channels, juxtaparanodal K_v1 channel localization depends on neuron–glia interactions mediated by CAMs. Furthermore, the formation of this domain depends on a tripartite CAM complex similar to that found at the paranodal junction. In the axon, a heterodimer consisting of Caspr2 (a homolog of Caspr) and TAG-1 (a homologue of contactin) binds in trans with glial TAG-1. Caspr2 interacts indirectly with K_v1 channels, most likely through an unidentified PDZ domain-containing protein since both K_v1 channels and Caspr2 contain consensus PDZ binding motifs, and mutation of these

PDZ binding motifs blocks their association. Mice that lack either TAG-1 or Caspr2 have normal levels of K_v1 channels in the axon, but K_v1 channels fail to accumulate at the juxtaparanodes. Finally, Caspr2 interacts with the cytoskeletal linker protein 4.1B. This protein links the Caspr2/TAG-1/K_v1 channel protein complex to αII and βII spectrin and the underlying axonal cytoskeleton.

Together, these examples demonstrate that common molecular mechanisms are utilized to organize ion channels into discrete domains in myelinated axons. In general, ion channel clustering depends on interactions between glial and axonal CAMs which function as attachment points for scaffolding proteins to cluster and anchor ion channels in the membrane. Finally, these ion channel/scaffold/CAM protein complexes are stabilized through interactions with the axonal cytoskeleton.

See also: Action Potential Initiation and Conduction in Axons; Calcium Channels; Epilepsy: Channelopathies; Ion Channel Localization in Cell Bodies and Dendrites; Neuromodulation of Sodium Channels; Neuromodulation of Calcium Channels; Sodium Channels; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Catterall WA, Goldin AL, and Waxman SG (2003) International Union of Pharmacology: XXXIX. Compendium of voltage-gated ion channels: Sodium channels. *Pharmacological Reviews* 5: 575–578.

Delmas P and Brown DA (2005) Pathways modulating neural KCNQ/M (K_v7) potassium channels. *Nature Reviews Neuroscience* 6: 850–862.

Dodson PD and Forsythe ID (2004) Presynaptic K^+ channels: Electrifying regulators of synaptic terminal excitability. *Trends in Neuroscience* 27: 210–217.

Garrido JJ, Giraud P, Carlier E, et al. (2003) A targeting motif involved in sodium channel clustering at the axonal initial segment. *Science* 300: 2091–2094.

Gu C, Jan YN, and Jan LY (2003) A conserved domain in axonal targeting of Kv1 (Shaker) voltage-gated potassium channels. *Science* 301: 646–649.

Hedstrom KL and Rasband MN (2006) Intrinsic and extrinsic determinants of ion channel localization in neurons. *Journal of Neurochemistry* 98: 1345–1352.

Lai HC and Jan LY (2006) The distribution and targeting of neuronal voltage-gated ion channels. *Nature Reviews Neuroscience* 7: 548–562.

Pan Z, Kao T, Horvath Z, et al. (2006) A common ankyrin-G-based mechanism retains KCNQ and NaV channels at electrically active domains of the axon. *Journal of Neuroscience* 26: 2599–2613.

Rasband MN, Park EW, Zhen D, et al. (2002) Clustering of neuronal potassium channels is independent of their interaction with PSD-95. *Journal of Cell Biology* 159: 663–672.

Rudy B and McBain CJ (2001) K_v3 channels: Voltage-gated K^+ channels designed for high-frequency repetitive firing. *Trends in Neuroscience* 24: 517–526.

Schafer DP and Rasband MN (2006) Glial regulation of the axonal membrane at nodes of Ranvier. *Current Opinion in Neurobiology* 16: 508–514.

Trimmer JS and Rhodes KJ (2004) Localization of voltage-gated ion channels in mammalian brain. *Annual Review of Physiology* 66: 477–519.

Zhou D, Lambert S, Malen PL, Carpenter S, Boland LM, and Bennett V (1998) AnkyrinG is required for clustering of voltage-gated Na channels at axon initial segments and for normal action potential firing. *Journal of Cell Biology* 143: 1295–1304.

Ionic Channels in Glia

M L Olsen and H Sontheimer, University of Alabama at Birmingham, Birmingham, AL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Kuffler and colleagues performed the first series of physiological studies investigating the biophysical properties of glial cells in the early 1960s. Using sharp microelectrodes, they demonstrated that glial cells from the mud puppy, leech, and salamander shared several biophysical properties. One common feature of glia was a hyperpolarized resting membrane potential (RMP), more negative than that of a typical neuron. A second common feature was an absence of electrical excitability. Injection of a depolarizing current (up to 200 mV) caused a linear change in the RMP, and current did not pass from the glia to a neuron or vice versa. However, glial cells did depolarize following neuronal stimulation, and it was discovered that this depolarization was due to an accumulation of external K^+ ions released from neurons after an action potential. Indeed, Kuffler and colleagues showed that glial membranes are selectively permeable for K^+ ions and suggested that glia may serve to regulate the extracellular K^+ concentration.

With the introduction of the patch–clamp technique in the early 1980s, it became possible to examine glial membrane properties more rigorously, and numerous studies have investigated glial cells from different species and preparations. These studies show quite universally that although passive, glial cells express many of the same ion channels and respond to most of the same ligands and neurotransmitters as their excitable neighbor, the neuron. However, because glial cells do not experience large voltage fluctuations necessary for many of these channels to activate, nor do they have the ability to fire action potentials, it has been difficult to assign putative functions for many of these channels. Indeed the ubiquitous expression of ion channels in all cell types (excitable or not) alludes to crucial functions in cellular biology distinct from conferring excitability. This article highlights the major findings of the past 25 years regarding ion channel expression and function in macroglial cells.

Potassium Channels

Inwardly Rectifying Potassium Channels

Inward rectifier potassium channels (Kir) are unique among K^+ channels expressed in glial cells in that they show their highest open probability close to the normal RMP values for most differentiated glial cells (~ -80 mV). Unlike the 'classic' gating of voltage-activated ion channels, the pore of the Kir channel is blocked from the intracellular side by polyamines and Mg^{2+}. This block is released at negative potentials, allowing K^+ influx. Because more channels are open at rest, they are capable of conferring the high potassium permeability typical of astrocytes and are hence ideally suited to establish and maintain the hyperpolarized glial RMP. Initial whole-cell patch–clamp experiments in glial cells demonstrated currents mediated by Kir channels in cultured rat astrocytes. To date, as many as eight of the 16 cloned Kir channels have been identified at the protein level in glial cells, while gene transcripts for all Kir genes have been identified (Table 1). It is believed that Kir4.1 is a particularly important member of the Kir channel family in glial cells.

The functions of Kir channels were directly inferred from the first biophysical studies. As mentioned above, most differentiated glia maintain a hyperpolarized RMP, close to the potassium equilibrium potential (E_K). This potential is required for the normal function of many of the transporters found in astrocytes. Opening of Kir channels will serve to pull the membrane potential toward E_K. Furthermore, Kir channel conductance increases with increasing concentrations of external potassium ions (Figure 1(a)). Therefore, when extracellular potassium rises, as it does following neuronal activity, channel conductance increases accordingly. Hence this channel is ideally suited to take up K^+ in response to neural activity by a process called potassium buffering. The polarized distribution of Kir channels (Figure 1(b)), which tend to localize preferentially to processes in the vicinity of neurons on the one hand and blood vessels on the other, allows astrocytes to remove potassium ions at their site of release during neuronal activity and subsequently redistribute these ions to blood vessels or areas where the extracellular potassium concentration is lower. Such a model of K^+ redistribution is supported by data from Müller cells, astrocyte-like cells in the retina which demonstrate co-localization of a specific Kir channel subunit, Kir4.1, at the subvitreal and perivascular spaces with the water channel aquaporin 4. Together these two channels allow for the uptake of K^+ and water in areas where concentrations are high and redistribute them to areas where concentrations are lower.

It is interesting that Kir channel expression appears to be developmentally regulated, with immature proliferating cells expressing fewer, if any, Kir channels

Table 1 Molecularly identified ion channels in glial cells

Channel and Subunit	*Preparation*	*Detection method*	*Putative function*
K^+-channels			
$K_{ir}4.1$	Müller cells	Immunogold-EM	Ion homeostasis
$K_{ir}4.1$	Adult rat CNS gray matter astrocytes	Immunohistochemistry	Potassium buffering
$K_{ir}4.1$	White matter astrocytes	Immunohistochemistry	Maintenance RMP, potassium buffering
Kir4.1	Cultured and slice spinal cord astrocytes	Western blot/ immunocytochemistry Knockout	Maintenance RMP, potassium buffering
Kir4.1 and 2.1	Mouse retina	Immunohistochemistry	Spatial buffering
Kir2.1, 2.2, 2.4, 3.1, 3.2, 4.1 and 6.1	Retina and primary cultured Müller cells	RT-PCR	Possibly spatial buffering
Kir 6.1	Hippocampal, cortical and cerebellar astrocytes, tanycytes, Bergmann glial cells	Immunohistochemistry	Function during metabolic changes
Kv1.4, 1.5 and 1.6	Cultured oligodendrocyte progenitor cells	RT-PCR/ immunocytochemistry	Membrane conductance and proliferation rates
Kv1.2, 1.4, 1.5, 1.6	Cultured oligodendrocytes and oligodendrocyte progenitors	RT-PCR/Western blot	Proliferation/differentiation
Kv1.5	Cultured spinal cord astrocytes	Western blot	Channel activity correlates with proliferation
Kv1.6	Cultured mouse cortical astrocytes	RT-PCR/ immunocytochemistry	Regulating extracellular ion homeostasis
Kv1.2, 1.4, 1.5, 1.6	Cultured oligodendrocytes and oligodendrocyte progenitors	RT-PCR/Western blot	
SK	Astrocytes of rat supraoptic nucleous	Immunohistochemistry	Aid in buffering potassium when $[Ca^{2+}]_i$ increased
BK	Rat astrocytic perivascular endfeet	Immunohistochemistry	Volume changes and potassium homeostasis
BK	Myelinating Schwann cells in sciatic nerve	Immunocytochemistry	Potassium homeostasis
Chloride channels			
ClC2 and ClC3	Cultured cortical astrocytes	PCR	Regulatory volume decrease
ClC2	Astrocytes	Immunohistochemistry	Deliver Cl^- to GABAergic terminals
ClC2, 3, 4, 5, 7 VDAC/CFTR	Cultured cortical astrocytes	PCR/immunocytochemistry	Volume regulation
ClC2, 3, 5	Cultured cortical astrocytes	Western blot	Volume regulation
ClC2	Hippocampul astrocytes (endfeet)	Immunohistochemistry	Facilitating GABAergic transmission
ClC2	Mouse astrocytes acute slice	Electrophysiological study in control and ClC2 KO	Volume regulation and K^+ buffering
Na^+-channel			
Nav1.2–1.3, Na-G	Cultured rat spinal cord astrocytes	RNA blot hybridization	Not discussed
$Na_v1.1$–1.3	Rat optic nerve, cultured spinal cord astrocytes	PCR	Not discussed
Nav1.2, 1.3	Cultured Schwann cells	PCR	Not discussed
$Na_v1.2$	Adult rat spinal cord	Immunocytochemistry	Not discussed
$Na_v1.6$	Astrocytes and Schwann cell	RT-PCR, *in situ* hybridization	Not discussed
Na-G	Astrocytes, glial sodium channel subunit	Cloning, northern blot, RNase protection, *in situ* hybridization	Detected in brain, and outside the CNS, may be TTX-insensitive Na_v channel
$Na_v1.6$	Cultured spinal cord astrocytes	Immunocytochemistry	Contribute to electrical properties
$Na_v1.6$	Bergman glia	*In situ* hybridization Immunohistochemistry	Not discussed
Ca^{2+}-channels			
α2-L, α1C-L α1D-L, α1A-P/Q	Reactive astrocytes following injury	Immunohistochemistry	Ionic homeostasis after injury
α1D, α2 and β3	Human Müller cells	RT-PCR	

Continued

Table 1 Continued

Channel and Subunit	*Preparation*	*Detection method*	*Putative function*
α1 L-type	Cultured human astrocytes	Western blot	Regulation of nitric oxide synthase
α1A-P/Q, α1B-N, α1C-L, α1E-R	Cultured rat cortical astrocytes	RT-PCR	Regulation of calcium influx
α1B-N, α1C-L, α1D-L, α1E-R, α1G-T	Cultured rat cortical astrocytes	RT-PCR/Western blot Immunocytochemistry	Calcium homeostasis
CL-type	Reactive astrocytes after ischemic injury	Immunohistochemistry	Ischemia or glutamate toxicity
L-type, N-type	Astrocytes and oligodendrocytes in spinal cord white matter	Confocal immunohistochemistry	Postinjury axonal conduction failure
α1E-R type	Oligodendrocytes in rat brain	Immunohistochemistry	Myelinogenesis
TRPC channels			
TRPC1	Cultured mouse astrocytes	Immunocytochemistry Antisense knockdown	Capacitative calcium entry
TRP1,2,3,4, and 6	Cultured rat cortical astrocytes	RT-PCR	Capacitative calcium entry
TRPC3	Cultured rat cortical astrocytes	RT-PCR/antisense KO	OAG- induced Ca^{2+} oscillations
TRPC4	Cultured human fetal astrocytes	RT-PCR, WB/immucyto (confocal and EM)	Regulation of calcium ion homeostasis
TRPC3	Rat brain oligodendrocytes	Immunohistochemistry/EM	Calcium homeostasis important for myelination

than do differentiated astrocytes. As expected, the higher density of Kir channels correlates with more-negative RMPs. In the context of disease, a lower Kir channel density is also observed in areas of gliosis, in brain regions associated with epileptic seizures, and in highly proliferative astrocytic tumors. Here the homeostatic roles of astrocytes which require a hyperpolarized RMP, as well as the inability to clear excess potassium, may no longer function, and hence this glial dysfunction may exacerbate neuronal hyperexcitability and cell death.

Voltage-Activated K^+ Ion Channels

Voltage-dependent potassium channels were the first voltage-activated channels to be described in glial cells. These channels form the largest and most diverse family of ion channels. Collectively they efflux potassium ions in response to membrane depolarization. Classically they are divided into three families: the fast-inactivating potassium channels (Ka), delayed-rectifying potassium channels (Kd), and calcium-activated potassium channels (KCa). The expression of each channel subtype in glia and possible functions are discussed below.

Ka and Kd Ka and Kd channels comprise the voltage-activated potassium channel subfamily (Kv), which, to date, comprises 40 identified members (Kv1.x–Kv12.x). In excitable cells, each of these channels has a specific role in shaping the action potential or its propagation. Currents mediated by classic Ka and Kd channels are easily distinguished from one another on the basis of their activation properties. Ka channel activation occurs rapidly, and currents typically inactivate quickly (Figure 2) whereas currents mediated by Kd channels activate with a delay and rarely inactivate, even with sustained depolarizations. These two channel types can be readily distinguished pharmacologically. Ka channels are sensitive to low millimole concentrations of 4-aminopyridine while Kd channels are inhibited by low millimole concentrations of tetraethylammonium. It is interesting that neither Ka nor Kd channels are open at typical RMPs of glial cells but instead begin to activate at approximately −40 mV.

Biophysical studies have demonstrated some Kd channel activity in nearly all glial cells, although higher channel activity is associated with developing glia. When induced to proliferate following brain injury or in the context of disease processes such as epilepsy, channel activity is increased relative to normal surrounding tissue. The RMP of proliferating or immature glia is depolarized relative to differentiated astrocytes, increasing the likelihood of Kd channel. It has been demonstrated that Kd channel inhibition or protein knockdown is sufficient to decrease cell proliferation rates suggesting a role for this channel in cell division. However, the underlying mechanism linking Kd channel activity and cell proliferation is unknown.

It is interesting that in the 25 years since these two channel types were first described in astrocytes, very

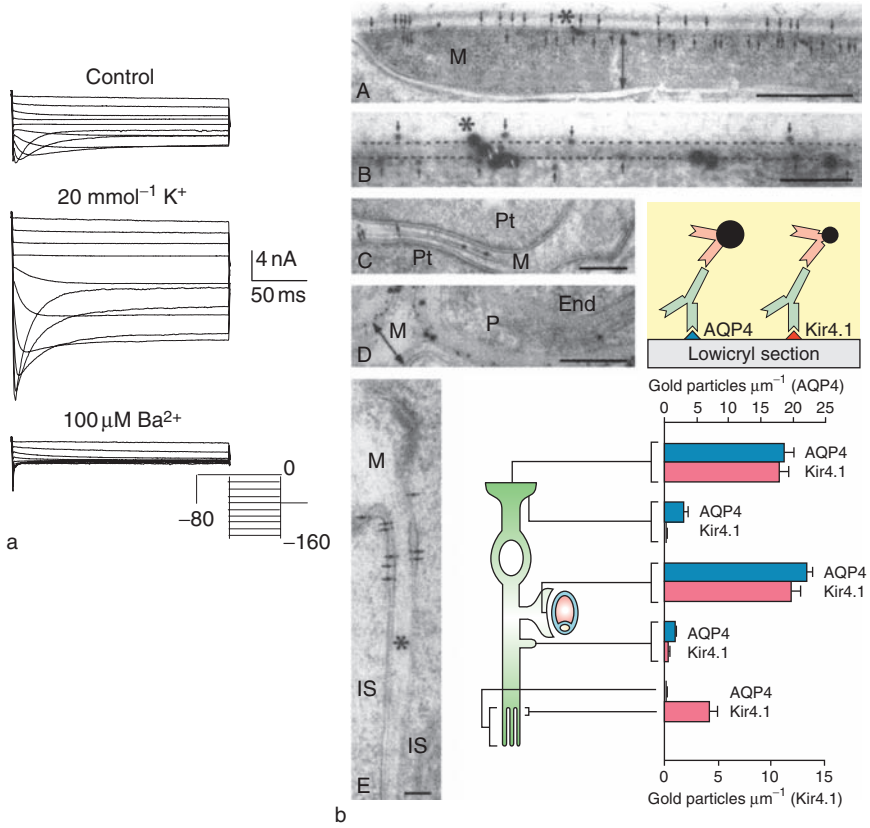

Figure 1 Inwardly rectifying potassium currents are robustly expressed in spinal cord astrocytes. (a) Typical voltage step response (protocol depicted in inset) in 5 mmol l^{-1} K^{+o} (top), 20 mmol l^{-1} K^{+o} (middle), and 5 mmol l^{-1} K^{+o} + 100 mmol l^{-1} Ba^{2+} (a nonspecific inward rectifier potassium (Kir) channel inhbitor; bottom). (b) Double immunogold labeling of Kir4.1 (10 nm particles) and aquaporin 4 (AQP4; 20 nm particles) in vitreal (A, B), perisynaptic (C), and perivascular (D) Müller cell (M) membranes. C and D are from the outer plexiform layer. (A, B) Note, in A, that the M end foot (M) is selectively labeled at its vitreal aspect (both aspects indicated by double arrow). The vitreal plasma membrane (between dashed lines in B) is obliquely cut, allowing labeling at the two sides of the section to be distinguished. Small arrows indicate Kir4.1 labeling. Asterisks indicate corresponding points at the vitreal surface (part of A is enlarged in B). C: Weak Kir4.1 (arrows) and AQP4 labeling is found in the thin M processes that surround photoreceptor terminals (Pt). D: Perivascular endfeet (M) show a polarized distribution of gold particles (double arrow; compare with A). End, endothelial cell; P, pericyte. E: Kir4.1 immunolabeling of an M microvillus (asterisk). The gold particles (arrows) are restricted to its basal part, which is devoid of large particles signaling AQP4. IS, inner segment of photoreceptors. Error bar = SEM (standard error of the mean). Graph shows linear densities of gold particles in different membrane domains of M. Note different scales for AQP4 (top) and Kir4.1 (bottom). Vitreal and perivascular endfeet membranes are coenriched in AQP4 and Kir4.1. Only in the microvilli is there a segregation between immunoreactivities for the two membrane molecules. Scale bars = 0.5 μm (a), 0.1 μm (b, e), 0.25 μm (c, d). (a) Reprinted from Olsen ML, Higashimori H, Campbell SL, Hablitz JJ, and Sontheimer H (2005) Functional expression of K_{ir} 4.1 channels in spinal cord astrocytes. *Glia* 53(5): 516–528, with permission of Willey–Liss, Inc., a subsidiary of John Wiley & Sons, Inc. (b) Reproduced from Nagelhus EA, Mathiisen TM, and Ottersen OP (2004) Aquaporin-4 in the central nervous system: Cellular and subcellular distribution and coexpression with KIR4.1. *Neuroscience* 129(4): 905–913, with permission from Elsevier.

little information has become available regarding possible functions. Conditions which depolarize the membrane sufficiently for channel activation are poorly understood. Although there is significant biophysical evidence to suggest that these channel types contribute to membrane conductance, possibly in the context of efflux during potassium buffering and/or maintenance of a hyperpolarized RMP, more-definitive functional studies are lacking. However, ubiquitous expression in nearly every glial preparation studied alludes to the importance of both Ka and Kd channels in glial cell biology.

Ca^{2+}-activated potassium channels Distantly related to the voltage-activated potassium channels are the calcium-activated potassium (KCa) channels. KCa channels can be grouped into three distinct subfamilies, large-conductance (BK), intermediate conductance (IK), and small conductance (SK) channels. As implied by their name, these channels differ in their unitary conductance. Unlike Kv channels, which activate solely in response to membrane depolarizations, all KCa channels require increases in intracellular calcium for their activation. BK channels also demonstrate voltage dependence not seen in

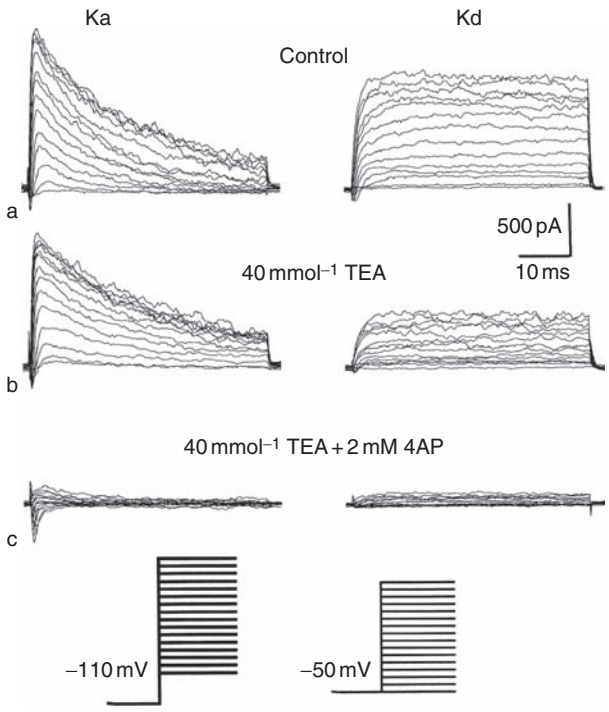

Figure 2 Pharmacological properties of K^+ currents observed in $BrdU^+$ spinal cord astrocytes. The isolated transient and sustained outwardly rectifying potassium currents from a typical scar-associated cell (6 h postinjury) recorded before and after the application of potassium channel blockers, as indicated. (a) In normal recording solution, large transient (left) and sustained currents (right) are recorded. (b) Application of 40 mmol l^{-1} TEA inhibited the sustained K^+ current by ~50% but had no effect on the transient current. (c) The addition of 4-aminopyridine (4AP; 2 mmol l^{-1}) completely inhibited the transient potassium currents and also inhibited the residual sustained K^+ current that was not inhibited by TEA. Ka, fast-inactivating potassium channels; Kd, delayed-rectifying potassium channels. Reproduced from MacFarlane SN and Sontheimer H (1997) Electrophysiological changes that accompany reactive gliosis *in vitro*. *Journal of Neuroscience* 17(19): 7316–7329, with permission from Elsevier.

either IK or SK channels and hence can be activated by $[Ca^{2+}]_i$, voltage, or both. Specific pharmacological channel inhibitors for each channel type are available, allowing for identification of currents mediated by each channel. BK and SK channels have been described biophysically in astrocytes, oligodendrocyte precursor cells, and Müller cells. Only a few studies have identified BK and SK channels *in situ* (Table 1).

Although glial cells rarely see depolarizations large enough to activate voltage-gated channels, there are numerous circumstances which raise $[Ca^{2+}]_i$ sufficiently to activate KCa channels. These include, but are not limited to, activation of purinergic receptors, G-protein-coupled receptors, transient receptor potential (TRP) channels, and mitogen signaling. It has been suggested that KCa channels may facilitate potassium efflux from glial cells. This interpretation would explain the localization of channels on glial endfeet that contact blood vessels. Furthermore, KCa channels may contribute to astrocyte–astrocyte signaling in response to $[Ca^{2+}]_i$ increases in the form of glial calcium waves. Depolarizations of the membrane activate calcium channels, and the combination of the calcium signal and membrane depolarization may activate either BK or SK channels. The resulting increase in extracellular potassium would serve to depolarize neighboring cells, thus further propagating the signal.

In some glial cells, BK channels display developmental regulation, with higher channel activity in proliferative cells. In fact, glioblastoma cells, or tumor cells thought to derive from glia, demonstrate robust BK channel activity. BK channels expressed in these cells demonstrate classic BK channel properties, including sensitivity to the specific BK channel inhibitor (Figure 3(a)) as well as a dependence on intracellular Ca^{2+} (Figure 3(b)). In general, proliferative cells maintain a more depolarized RMP than that of quiescent cells. The opening of any potassium channel will pull the RMP toward the potassium equilibrium potential or hyperpolarize the cell, suggesting that rather then remaining open, these channels open briefly in response to increases in intracellular calcium. This suggests that other channels mediate the resting conductance in proliferative cells. However, higher potassium channel activity in proliferative cells is commonly observed. Indeed, potassium channel activation has long been linked to cell cycle progression for nearly every cell type studied. However, whether glial potassium channels simply provide a permissive ionic environment for cell division or whether regulation of the RMP has a more instructive role remains to be seen.

Voltage-Activated Chloride Channels

Chloride channels are ubiquitously expressed in all cells and are typically involved in vital cellular functions that include cell volume regulation, cell proliferation, and salt secretion. As the chief physiological anion, chloride ions provide the counter ion for many cellular transport systems. Astrocytes and oligodendrocytes differ markedly in their chloride permeability, resting conductance, and resting chloride concentrations. Astrocytes have relatively low chloride permeability but actively accumulate chloride as their intracellular Cl^- concentration is much higher than predicted by the RMP. Consequently, the opening of astrocytic chloride channels results in chloride efflux. Oligodendrocytes demonstrate a higher permeability to Cl^- and have intracellular chloride concentrations favoring passive distribution.

Biophysical studies have described Cl^- channels in many glial cell preparations, yet the majority of these

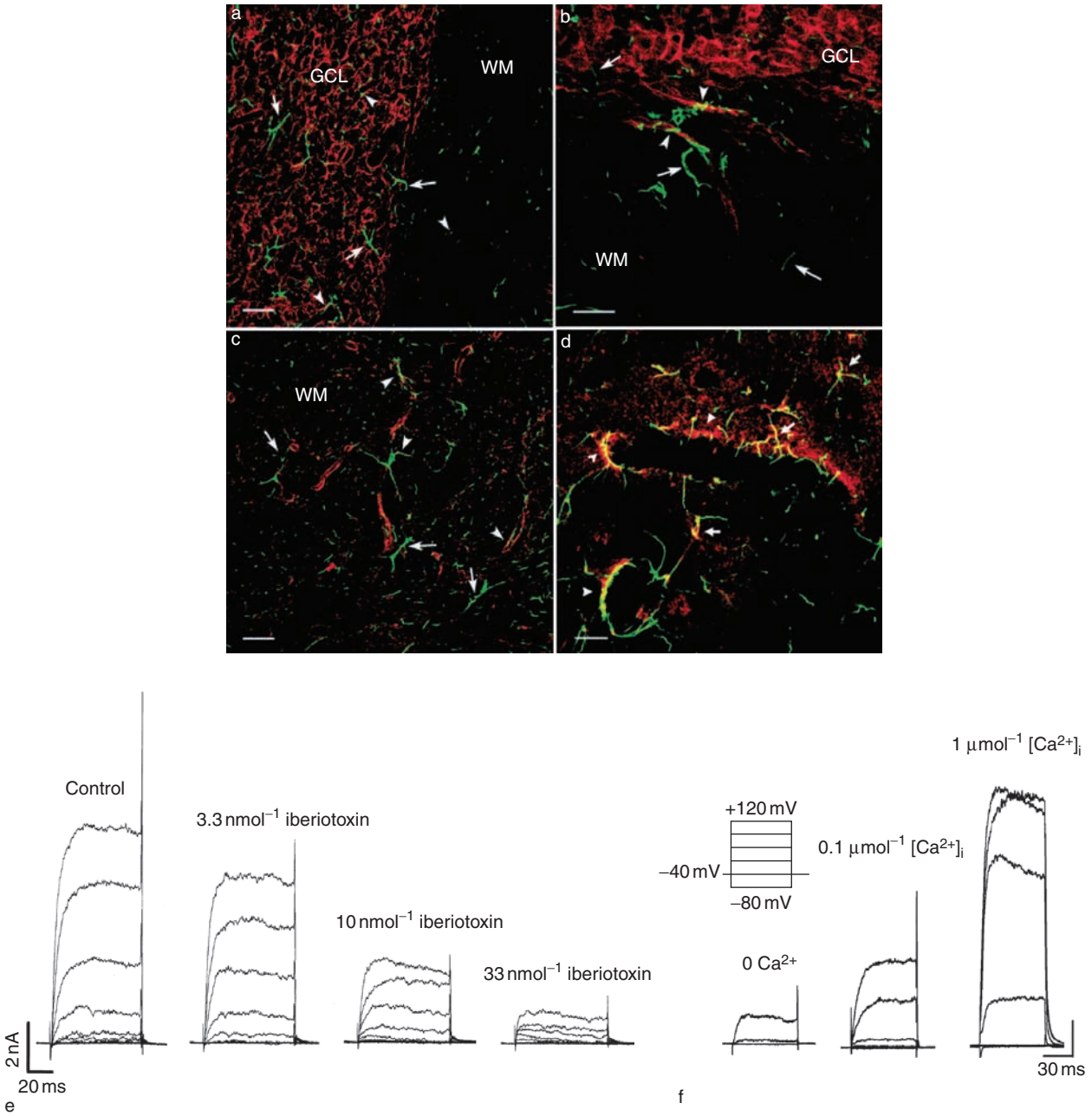

Figure 3 Ca-activated potassium channels are expressed in glial cells. (a) Co-localization of rat large-conductance (BK) channels and glial fibrillary acidic protein (GFAP) in the rat cerebellum and hippocampus. Confocal images of BK (in red) and GFAP (in green) immunolabeling of the cerebellar granule cell layer (GCL), white matter (WM) ((a) and (b)), and the deep cerebellar white matter (c). Note that most cerebellar astrocytic GFAP labeling has little BK labeling associated with it (arrows); however, GFAP and BK co-localization, when it occurs, is frequently found surrounding blood vessels (broad arrowheads). (d) Confocal image of BK (red) and GFAP (green) immunolabeling within hippocampal CA1 region. Hippocampal GFAP labeling overlaps with BK along astrocytic processes (arrows), with high densities of overlap occurring at perivascular end-foot regions (arrowheads). (a), (c), and (d) are single optical sections, and (b) is a maximum projection reconstructed from a z-series ($n = 16$). Scale bars = 20 μm (a–c), 10 μm (d). (e) Whole cell voltage-dependent currents in human glioma cells were inhibited by the indicated concentration of scorpion-venom peptide iberiotoxin. (f) BK channels in glioma cells are particularly calcium sensitive to elevated intracellular calcium. (a–d) Reproduced from Price DL, Ludwig JW, Mi H, Schwarz TL, and Ellisman MH (2002) Distribution of rSlo Ca^{2+}-activated K^+ channels in rat astrocyte perivascular endfeet. *Brain Research* 956(2): 183–193, with permission from Elsevier. (e,f) Reproduced from Ransom CB and Sontheimer H (2001) BK channels in human glioma cells. *Journal of Neurophysiology* 85(2): 790–803, used with permission from The American Physiological Society.

studies have been carried out on cultured astrocytes. Collectively, they demonstrate that chloride channel activation requires an osmotic challenge or a change in cell morphology. In contrast, gliotic astrocytes or astrocytes which have undergone neoplastic transformation have an elevated resting chloride conductance, and in these cells chloride currents can be recorded in the absence of an osmotic challenge.

Several biophysical studies have identified an inwardly rectifying chloride current (mediating an efflux of chloride ions) which has the electrophysiological signature of ClC-2. ClC-2 belongs to the family of voltage-gated chloride channels (ClC1–ClC-9). Consistent with this finding, a comparison of astrocytes in ClC-2 knockout mice and wild-type littermates suggests that astrocytes normally express functional ClC-2 channels. ClC-2 immunoreactivity has been identified in astrocytic endfeet or processes which localize with γ-aminobutyric acid (GABA)ergic axon terminals. Here, ClC-2 may facilitate GABAergic transmission by presenting chloride ions to axons involved in GABAergic transmission. It is interesting that mutations in the ClC-2 gene have been linked to several of the most common types of idiopathic generalized epilepsy. It is thought that two such mutations would lower the transmembrane chloride gradient essential for GABAergic inhibition.

The majority of studies examining chloride channel function have done so in the context of cell volume regulation (Figure 4). Astrocytic volume changes have been demonstrated to occur in response to physiological events, such as potassium buffering and in pathophysiological conditions such as brain edema. In the former, the movement of K^+ and Cl^- ions enters the cells along with water in regions of elevated potassium. There does not appear to be any long-term net gain in cell volume; rather K^+ and Cl^- and water would also be released at a distant site. In edematous brain, astrocytic swelling dramatically decreases the size of the extracellular space and in turn increases the concentration of ions and molecules in this space. In culture, swelling of astrocytes by a hypotonic challenge results in activation of chloride channels that subsequently mediate regulatory volume decrease. Application of chloride channel inhibitors prevents the swelling-induced regulatory volume decrease. Unfortunately, the absence of specific chloride channel inhibitors makes the identification of the underlying channels difficult, but many possible gene candidates have been identified (Table 1). There remains a controversy as to whether the volume-activated Cl^- channel, often called ICl_{swell}, belongs to the ClC family or has yet to be molecularly identified.

Voltage-Gated Calcium Channels

Intracellular calcium concentrations are maintained at least four orders of magnitude lower than that of extracellular calcium. This gradient makes calcium an exceptionally effective intracellular signaling molecule. Plasma membrane voltage-gated calcium channels constitute a significant mechanism for calcium signaling. Five types of voltage-gated calcium channels (L, T, P, N, and Q) can be distinguished on the basis of biophysical and pharmacological properties. Molecularly, the alpha subunits that comprise the main channel structure are grouped into three families. These are Ca_v1, which include L-type channels; Ca_v2, composed of L, T, R, N, and P/Q channel subunits; and Ca_v3, or T-type channels. These channels can be distinguished pharmacologically (Figure 5). Auxiliary subunits often associate with the alpha subunit, altering its properties. Regardless of the channel subtype, Ca_v channels always require a depolarization for their activation. Physiological activation of

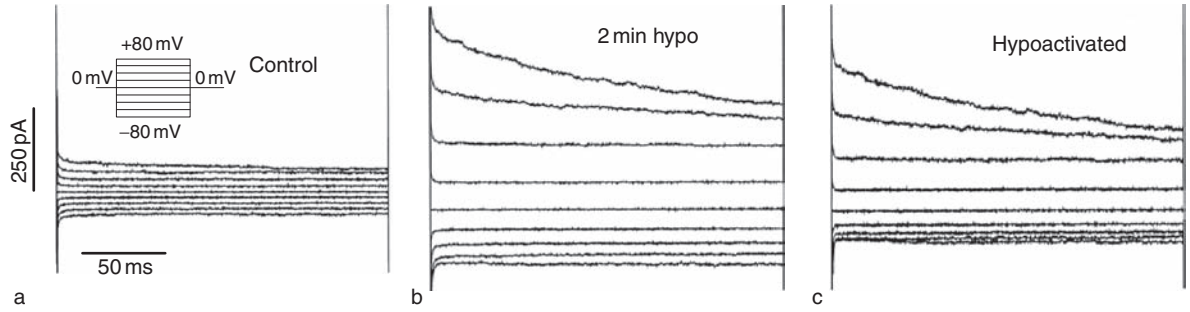

Figure 4 Hypotonic challenge activates whole-cell chloride currents in cultured cortical astrocytes. (a) Representative whole-cell recording of voltage steps from −80 mV to +80 mV for 200 ms from a 0-mV holding potential (inset) in choline chloride (ChCl) bath and pipette solutions before the osmotic challenge. (b) Voltage steps after 2 min in hypotonic ChCl bath solution shows increased current magnitudes. (c) Subtraction of (b) from (a) demonstrates hypotonically activated currents. Reprinted from Parkerson KA and Sontheimer H (2004) Biophysical and pharmacological characterization of hypotonically activated chloride currents in cortical astrocytes. *Glia* 46(4): 419–436, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley and Sons, Inc.

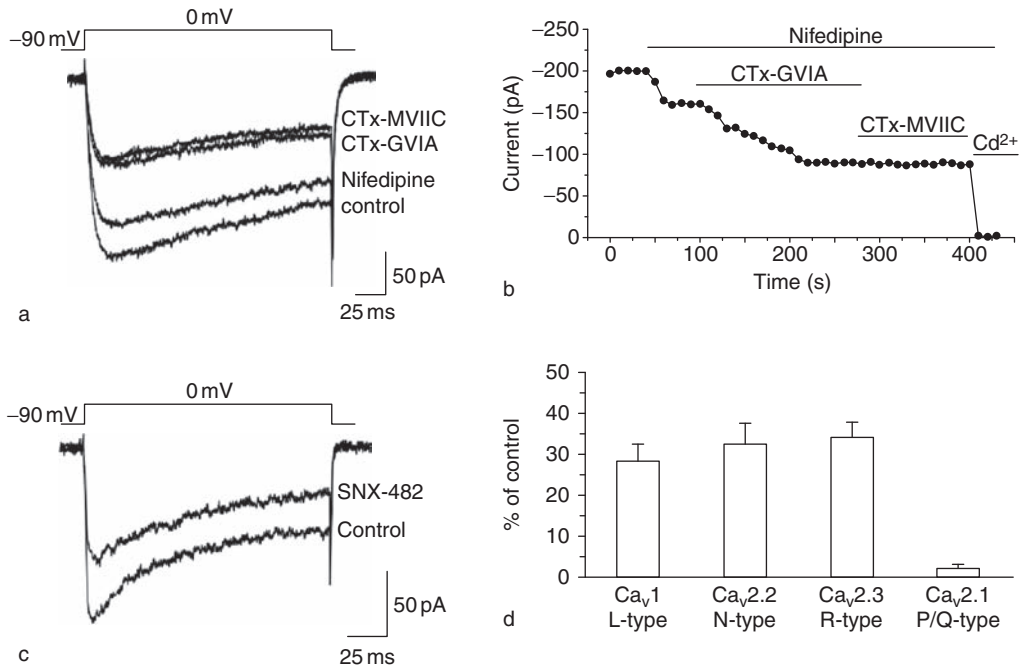

Figure 5 Voltage-gated calcium currents are expressed in rat cortical astrocytes. (a) High-voltage-activated (HVA) currents recorded before (control) and after application of 5 mol l^{-1} nifedipine (which blocks L-type channels), 3 mol l^{-1} CTx-GVIA (blockade of N-type channels), and 3 mol l^{-1} CTx-MVIIC (P/Q channel blockade). (b) Time plot of data from the experiment shown in (c) and blockade of residual current (i.e., resistant to L-, N-, and P/Q-channel blockade) by 200 mol l^{-1} Cd^{2+}. (c) Reduction of HVA current by the selective blocker of R-type channels SNX-482 (100 nmol l^{-1}). (d) Percentage contributions of the different types of voltage-gated calcium channels to the total HVA current recorded in cortical astrocytes. Reprinted from D'Ascenzo M, Vairano M, Andreassi C, Navarra P, Azzena GB, and Grassi C (2004) Electrophysiological and molecular evidence of L-(Cav1), N-(Cav2.2), and R-(Cav2.3) type Ca2+ channels in rat cortical astrocytes. *Glia* 45(4): 354–363, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley and Sons, Inc.

these channels in glial cells may be achieved by increases in extracellular potassium or by activation of other ligand-gated channels, which allows nonspecific cation influx to depolarize the membrane. In addition, glial cells respond to neurotransmitters such as glutamate, which results in calcium influx and membrane depolarization.

Voltage-activated calcium channels were first demonstrated in cultured astrocytes and oligodendrocytes. These initial investigations demonstrated that either coculture with neurons or exposure to agents that increase intracellular cyclic adenosine monophosphate is required to detect channel activity in cultured cells. Subsequently, several studies examined the expression of these channels in slices or acutely dissociated cells using calcium imaging techniques as well as whole-cell voltage clamp experiments. These investigations revealed that intracellular calcium increase caused by K^+-induced depolarizations was dependent on influx of calcium (independent of release from intracellular stores) and was inhibited with different Ca_v channel inhibitors. Although channel subtypes seem to vary in different preparations, the prevailing consensus is that glial cells express these channels *in situ*. Indeed, several Ca_v channel subunits have been molecularly identified in astrocytes, oligodendrocytes, and Müller cells (Table 1).

Under pathological conditions, plasma membrane depolarizations would likely be significant because of increases in extracellular K^+ and influx of calcium through other mechanisms. Indeed, many of the studies examining the functional role of Ca_v channels are in this context. Biophysical and molecular investigations have revealed that calcium influx through voltage-gated channels following injury or ischemia was detrimental and potentially contributed to glial injury. Furthermore, Ca_v channel immunoreactivity increased following injury. It is interesting that the regulatory volume decrease is severely impaired when astrocytic voltage-gated calcium channels are inhibited. Yet another curiosity pertains to Ca_v expression in the context of seizures. It has been shown that $[Ca^{2+}]_o$ (where 'o' denotes extracellular) levels are reduced below normal levels at seizure focal sites, and indeed, depletion of $[Ca^{2+}]_o$ is sufficient to produce seizurelike activity experimentally. This activity can be blocked with inhibitors of voltage-gated Ca^{2+} channels, suggesting that calcium channels present on neurons as well as glia may be involved in calcium uptake from the extracellular space. If true, glial

voltage-gated calcium channels may have a role in extracellular calcium homeostasis.

Voltage-Gated Sodium Channels

Voltage-gated sodium channels mediate the upstroke of the action potential in excitable cells. They are defined by an alpha subunit derived from one of nine genes ($Na_v1.1$–1.9). Auxiliary beta subunits typically associate with the alpha subunit and modify channel properties. The initiation and propagation of the action potential appears to be the only capacity in which these channels function. Yet glial cells do not fire action potentials. Not surprisingly, the first description of Na^+ channels in glia was greeted with suspicion. To date, glial cells from culture, acutely dissociated and in the acute slice, have been shown unequivocally to express voltage-gated sodium channels, albeit often at low density, insufficient to generate an action potential. Indeed, sodium currents reminiscent of those seen in neurons, which activate at approximately 60 mV, inactivate quickly, and are sensitive to low (10 nmol l^{-1}) tetrodotoxin, have been described in astrocytes (Figure 6(a)). A second type of current unique to glial cells has been described that demonstrates tetrodotoxin insensitivity (1000 nmol l^{-1}; Figure 6(b)). Sodium currents which display low tetrodotoxin sensitivity have been described in protoplasmic astrocytes as well as injured astrocytes *in vitro* (Figure 6). Studies from cultured spinal cord

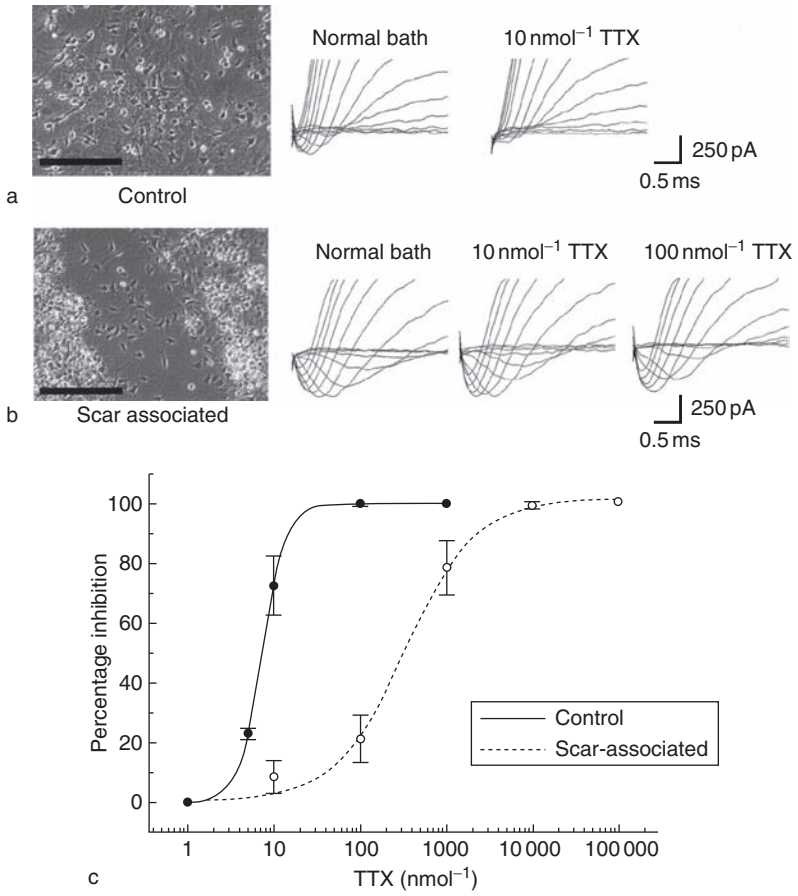

Figure 6 Spinal cord astrocytes display altered tetrodotoxin (TTX) sensitivity on injury *in vitro*. (a) Photomicrograph of control culture and representative Na^+ current traces for control cells. In the presence of normal bath solution, Na^+ currents are evoked in response to a voltage protocol that hyperpolarized the cell to 110 mV for 200 ms and then depolarized the cell for 8 ms to potentials ranging from 70 to 80 mV. Sodium currents in control cells are inhibited largely by 10 nmol l^{-1} TTX. (b) Photomicrograph of scar-associated cells and their representative current recordings. In response to the same voltage protocol (above), scar-associated cells exhibit Na^+ currents with twofold greater conductance than those of control cells. These currents are largely resistant to 10 and 100 nmol l^{-1} TTX. (c) Dose-dependent current inhibition for control vs. Scar-associated Na^+ currents. Apparent IC50 for scar-associated cells was 314 nmol l^{-1}, which is ~40-fold less sensitive than control cells (IC50 = 8 nmol l^{-1}). Scale bar = 100 μm. Error bars = SEM (standard error of the mean). Reproduced from Macfarlane SN and Sontheimer H (1998) Spinal cord astrocytes display a switch from TTX-sensitive to TTX-resistant sodium currents after injury-induced gliosis in vitro. *Journal of Neurophysiology* 79: 2222–2226, used with permission from The American Physiological Society.

astrocytes, A2B5$^+$ progenitor cells, and astrocytes from acute hippocampal slices have demonstrated that channel activity is dynamic and heterogeneous. These data are supported by immunocytochemical studies which demonstrate regional variations and suggest channel expression may be developmentally regulated. Additional data from reactive glia, which express ionic currents reminiscent of immature glia, in tissue associated with epilepsy and low-grade astrocytic tumors demonstrate these cells display higher-than-normal sodium channel activity.

As stated above, the identification of Na_v channels in glial cells prompted many investigators to search for some type of excitability in these cells. And in some cell types under specific conditions, large or suprathreshold stimulations resulted in an action potential-like response. However, even in cultured spinal cord astrocytes, which display a relatively high sodium channel density, several factors, including a high potassium resting conductance and low input resistance, preclude them from becoming excitable. Putative functions for sodium channels in glia are scarce. However, blockade of sodium channels in astrocytes reduced Na^+K^+ adenosine triphosphatase (Na^+K^+ATPase) activity, depolarized the RMP, and reduced intracellular sodium concentrations, suggesting glial sodium channels may provide a pathway for sodium influx necessary to maintain Na^+K^+ATPase activity. Conversely, inhibition of sodium channel inactivation in astrocytes from hippocampal slice led to a marked increase in intracellular sodium, suggesting these channels function in some biological capacity.

TRP Channels

TRP channels are a large family of sensory transduction ion channels that are activated in response to a number of stimuli, including mechanical stimuli, chemical messengers, heat, osmotic stress, and odorants, among others. By and large, these channels are voltage-independent, nonselective cation channels, and their activation causes membrane depolarization due to movement of positively charged ions into the cell. However, because of the disproportionate driving force for calcium ions to enter the cell, this particular ion garners much of the attention. TRPC (C = canonical) channels comprise a subfamily of this large group of proteins; they influx Ca^{2+} in response to activation of membrane receptors linked to phospholipase C. Of the near 30 distinct genes encoding mammalian TRP channels, there are seven TRPCs, and all but one (TRPC7) have been molecularly identified in astrocytes (Table 1). Although controversial, several TRPC channels have been suggested to be the store-operated channels involved in capacitative calcium entry (CCE) whereby depletion of endoplasmic reticulum calcium stores activates influx from the plasma membrane. CCE in turn is important in intracellular calcium regulation. It is in this capacity that TRPC channels have been most studied in astrocytes. Supporting the idea that TRPC channels promote CCE, several TRPC channels are found on the plasma membrane of cultured astrocytes, and knockdown of TRPC1 in mouse cortical astrocytes decreases extracellular calcium influx in response to depletion of intracellular stores. In contrast, it has been shown that two TRPC-permeable ions, Ba^{2+} and Sr^{2+}, do not enter astrocytes during CCE, arguing that TRPC channels are not involved. A second line of investigation has suggested that TRPC channels may be involved in astrocytic calcium oscillations. Although they are likely involved in intracellular calcium regulation, further study is needed to elucidate the role of TRPC channels in astrocytes and their possible role in CCE.

ATP-Gated Channels

ATP is a primary signaling molecule in astrocytes under physiological as well as pathological conditions. Ligand-gated channels that bind ATP and its metabolites are termed purinergic receptors. These can be further classified into purinergic ion channel receptors (P2X) and G-protein-coupled receptors. This article focuses on P2X receptors, or receptors that allow influx of sodium and calcium and efflux of potassium in response to extracellular binding of ATP. Activation of P2X receptors results in large increases in intracellular calcium because of the huge driving force to move calcium ions into the cell. These changes in intracellular calcium culminate in a primary form of signaling between adjacent and neighboring astrocytes as well as astrocytes and neurons. Furthermore, the increases in intracellular calcium on astrocytic P2X receptor activation have been associated with increased cell proliferation, ion channel activation, and countless other downstream signaling events.

Purinergic receptors were first described in astrocytes more than 10 years ago, when perforated patch–clamp experiments demonstrated that application of 100 μmol l^{-1} ATP induced an inward current that desensitized rapidly. This current was dependent on external sodium ions, and measurements with potassium-sensitive microelectrodes demonstrated a decrease in intracellular K^+. Furthermore, calcium imaging demonstrated a transient increase in intracellular calcium that appeared dependent on both internal and external calcium sources. Since these

initial studies, all seven P2X receptors (1–7) have been molecularly identified in astrocytes in many preparations (Table 1).

Data from multiple laboratories suggest that astrocytes from injured or ischemic tissue respond differently to ATP. In addition, it has been suggested that increased ATP concentrations following brain injury may stimulate gliosis via stimulation of purinergic receptors. After injury, changes in P2X receptor subunit expression were demonstrated in reactive astrocytes; however, particular attention has focused on the P2X7 receptor, which shows postischemic upregulation both *in vivo* and *in vitro*. This subunit has the unique ability to form a pore permeable to relatively large molecules on prolonged exposure to agonist and in the presence of low divalent cations, conditions experienced during ischemic insults and seizures. Typically, activation of the P2X7 receptor subunit requires high concentrations of agonist seen after brain injury. It is interesting that P2X7 pore formation has been coupled to nonvesicular, calcium-independent glutamate release both *in vitro* and *in vivo*. Although not selectively permeable to glutamate, in astrocytes this anion maintains the largest concentration gradient. Because glutamate can induce ATP release, glutamate efflux may act to amplify purinergic receptor signaling, possibly leading to glutamate-induced cell death. This unique pore may also represent a unique pathway for astrocytic ATP release following brain injury.

Concluding Remarks

Biophysical studies of glial cells have challenged the notion that voltage-gated ion channels function exclusively or even primarily to generate an electrical signal. Instead, roles for ion channels in glia must be considered that pertain to ion homeostasis and general cellular biology. In many instances, the true importance of these channels remains elusive. Recent molecular genetic techniques are now making a more rigorous assessment of ion channel function a reality.

See also: Calcium Channels; Calcium Homeostasis in Glia; Glial Ion Homeostasis: A Fluorescence Microscopy Approach; Large Conductance Calcium-Activated Potassium Channels; Neuromodulation of Sodium Channels; Neuromodulation of Calcium Channels; Potassium Channel Regulation; Potassium Homeostasis in Glia; Sodium Channels; Transient Receptor Potential (TRP) Channels; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

D'Ascenzo M, Vairano M, Andreassi C, Navarra P, Azzena GB, and Grassi C (2004) Electrophysiological and molecular evidence of L-(Cav1), N-(Cav2.2), and R-(Cav2.3) type Ca2+ channels in rat cortical astrocytes. *Glia* 45(4): 354–363.

MacFarlane SN and Sontheimer H (1997) Electrophysiological changes that accompany reactive gliosis *in vitro*. *Journal of Neuroscience* 17(19): 7316–7329.

MacFarlane SN and Sontheimer H (1998) Spinal cord astrocytes display a switch from TTX-sensitive to TTX-resistant sodium currents after injury-induced gliosis in vitro. *Journal of Neurophysiology* 79: 2222–2226.

Nagelhus EA, Mathiisen TM, and Ottersen OP (2004) Aquaporin-4 in the central nervous system: Cellular and subcellular distribution and coexpression with KIR4.1. *Neuroscience* 129(4): 905–913.

Olsen ML and Sontheimer H (2004) Voltage-activated ion channels in glial cells. In: Ransom BR and Kettenmann H (eds.) *Neuroglia,* 2nd edn., pp. 112–130. New York: Oxford University Press.

Parkerson KA and Sontheimer H (2004) Biophysical and pharmacological characterization of hypotonically activated chloride currents in cortical astrocytes. *Glia* 46(4): 419–436.

Price DL, Ludwig JW, Mi H, Schwarz TL, and Ellisman MH (2002) Distribution of rSlo Ca^{2+}-activated K^{+} channels in rat astrocyte perivascular endfeet. *Brain Research* 956(2): 183–193.

Ransom CB and Sontheimer H (2001) BK channels in human glioma cells. *Journal of Neurophysiology* 85(2): 790–803.

Sontheimer H, Black JA, and Waxman SG (1996) Voltage-gated sodium channels in glia: Properties and possible functions. *Trends in Neuroscience* 19: 325–331.

Steinhauser C and Seifert G (2002) Glial membrane channels and receptors in epilepsy: Impact for generation and seizure activity. *European Journal of Pharmacology* 447: 227–237.

Verkhratsky A, Orkand RK, and Kettenmann H (1998) Glial calcium: Homeostasis and signaling function. *Physiology Reviews* 78(1): 99–141.

Verkhratsky A and Steinhauser C (2000) Ion channels in glial cells. *Brain Research Reviews* 32: 380–412.

Wolfgang W (2002) Chloride ion channels in glial cell membranes. *Glia* 40: 1–10.

Wonderlin WF and Strobl JS (1996) Potassium channels, proliferation and the G1 progression. *Journal of Membrane Biology* 154: 91–107.

Joubert Syndrome

R J Ferland, Rensselaer Polytechnic Institute, Troy, NY, USA
C A Walsh, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Mutations in basic developmental genes are frequently observed in a variety of human neurodevelopmental syndromes. Disorders affecting the cerebellum (and brain stem) comprise some of these developmental disorders, which often result in clumsiness (ataxia), developmental delays, learning disabilities, mental retardation, and autistic-like behaviors. Thus, identification and characterization of the genes responsible for these neurodevelopmental abnormalities will be critical in our understanding of the basic mechanisms of cerebellum and brain stem development in addition to understanding these neurological disorders. Joubert syndrome is one such neurodevelopmental disorder in which the development of the cerebellum and brain stem is greatly affected in addition to the presence of abnormalities in the crossing of axons in the posterior midline of the brain.

Clinical Presentation

Joubert syndrome was originally described by Dr. Marie Joubert in 1968 in a family presenting with hyperventilation, abnormal eye movements, ataxia, and mental retardation with agenesis of the cerebellar vermis. Joubert syndrome can be divided into various subclassification groups based on the clinical presentation and the neuropathological profile. The originally described syndrome is commonly referred to as classic Joubert syndrome and primarily consists exclusively of central nervous system (CNS) involvement. Patients with classic Joubert syndrome first present with weakness (hypotonia) and abnormal, irregular breathing (episodic hyperpnea and/or apnea) during the neonatal period. Later, atypical eye movements (oculomotor apraxia, nystagmus, and strabismus), ataxia, and mirror movements (behavior whereby the movements on one side of the body are 'mirrored' by the other) may become apparent. Patients with Joubert syndrome can have developmental delay and a spectrum of cognitive abilities ranging from normal intelligence to severe mental retardation. Another common feature of Joubert syndrome is the presence of autistic-like behaviors.

Non-CNS abnormalities also can occur in Joubert syndrome, including cystic kidney disease (renal disease whereby fluid-filled sacs develop in the kidney forming cysts), nephronophthisis (a progressive renal disease affecting the kidneys symmetrically, particularly the tubules and glomeruli), hepatic fibrosis (disease affecting the liver where fiber-like connective tissue develops throughout the liver), soft-tissue tongue tumors, and polydactyly (too many fingers or toes). These non-CNS features, when they occur with the more standard classical Joubert syndrome brain phenotype, are evidence for the clinical heterogeneity that is apparent in Joubert syndrome.

Neuroanatomical Abnormalities

Although Joubert syndrome was first identified in 1968, most of our understanding of the neuropathology of Joubert syndrome comes from magnetic resonance imaging (MRI) studies since there are few neuropathological studies of individuals with Joubert syndrome. The available neuropathological and neuroradiological data suggest that individuals with Joubert syndrome have two major brain abnormalities: (1) malformations of the cerebellum, and (2) an interesting, but complex, neuroanatomical abnormality called the molar tooth sign.

The cerebellum is divided into four major lobes, the flocculonodular lobe, the two cerebellar hemispheres, and the vermis, which lies between the two cerebellar hemispheres. Individuals with Joubert syndrome specifically have abnormalities of the cerebellar vermis including absence (aplasia), reduced size (hypoplasia), or a disorganized structure (dysplasia) (compare Figure 1(b) to Figures 1(d), 1(f), and 1(h)). These abnormalities can be observed *in utero*, in the developing fetus with Joubert syndrome. With aplasia of the vermis, the two cerebellar hemispheres appose one another, but do not appear fused. On MRI, the cerebellar hemispheres and flocculonodular lobe seem relatively unaffected in Joubert syndrome; however, there have been neurohistological reports of Purkinje cell loss and proliferation of Bergmann glia in the cerebellum. Overall, these neuropathological findings suggest that individuals with Joubert syndrome have abnormalities in the patterning of the cerebellar vermis, the major midline structure in the cerebellum.

There are four deep nuclei in the cerebellum, the dentate, emboliform, globose, and fastigial nuclei. The most severely affected of the four cerebellar nuclei, in individuals with Joubert syndrome, are the dentate nuclei which give rise to the superior cerebellar peduncles. In Joubert syndrome, the dentate nuclei are poorly defined and are separated into small

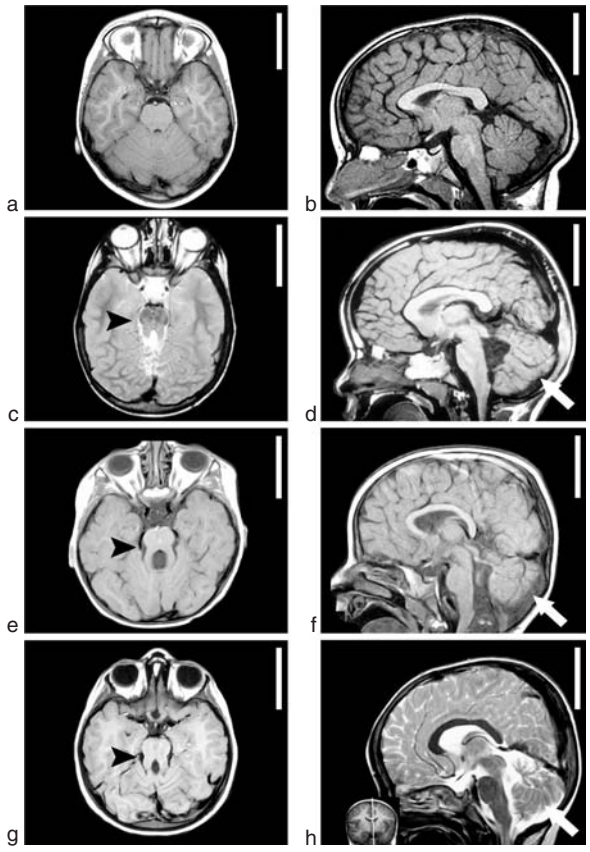

Figure 1 MRI of a normal brain (a, b) and the brains of individuals with *AHI1* mutations (c–h). MRIs are presented in an axial plane (a, c, e, g showing the cerebellum and upper pons) and in a sagittal plane (b, d, f, h showing a mid-sagittal section through the brain displaying the cerebral cortex, cerebellum, midbrain, and pons). Axial T1-weighted (a) and mid-sagittal T1-weighted (b) images of an unaffected child with a normal brain stem (a) and a normal cerebellar vermis (b). Axial T2-weighted (c) and mid-sagittal T1-weighted (d) MRI images of a male child at 6 years of age with *AHI1* mutations demonstrating the molar tooth sign ((c) black arrowhead) and the absence of the cerebellar vermis ((d) white arrow). Axial T1-weighted (e) and mid-sagittal T2-weighted (f) images of a male child with *AHI1* mutations at 4 years of age showing the molar tooth sign ((e) black arrowhead) and the absence of the cerebellar vermis ((f) white arrow). Axial T1-weighted (g) and mid-sagittal T1-weighted (h) images of a 2-year-old male child with *AHI1* mutations also noting the presence of the molar tooth sign ((g) black arrowhead) and a lack of a cerebellar vermis ((h) white arrow). In all of the axial images from affected individuals (c, e, g), the failure of the superior cerebellar peduncles to decussate can be appreciated by the presence of the peduncles traveling horizontal to the brain stem and not crossing the midline. Note that the brain structure that is present in the mid-sagittal sections (d, f, h) at the level of the cerebellum in individuals with Joubert syndrome is not the vermis but in fact is the right lobe of the cerebellum that is now in the plane of the mid-sagittal section because of the absence of the vermis. Scale bar = 5 cm. Adapted from Ferland RJ, Eyaid W, Collura RV, et al. (2004) Abnormal cerebellar development and axonal decussation due to mutations in *AHI1* in Joubert syndrome. *Nature Genetics* 36: 1008–1013, with permission.

islands; however, the dentate nuclei do appear to send out their axons forming the superior cerebellar peduncle.

The superior cerebellar peduncles are the major output (efferent) pathway of the cerebellum projecting to various thalamic, midbrain, and medullary structures. In normal brains, the superior cerebellar peduncle projections will cross the midline (or decussate) and project to the contralateral thalamus and midbrain. However, in Joubert syndrome, the superior cerebellar peduncle projections do not decussate in the dorsal midbrain, but remain ipsilateral to the dentate nuclei of origin. Moreover, the superior cerebellar peduncles are much more thickened, and run horizontally between the cerebellum and midbrain as compared to normal (compare **Figure 1(a)** to **Figures 1(c), 1(e), and 1(g)**). In addition to the absence of a decussation of the superior cerebellar peduncles, individuals with Joubert syndrome also lack the decussation of the central pontine tracts and the corticospinal tracts. These findings suggest that, in Joubert syndrome, there are defects in either axonal guidance or axonal crossing in the posterior midline of the brain.

The abnormally projecting superior cerebellar peduncles also are part of a complex brain stem malformation referred to as the molar tooth sign. The molar tooth sign, visible on radiographic imaging, consists of dysgenesis of the isthmic portion of the brain stem at the mesencephalon–metencephalon (pontomesencephalic) junction, enlargement of the 4th ventricle, vermic clefting, and the abnormally thickened and reoriented superior cerebellar peduncles (**Figures 1(c), 1(e), and 1(g)**). The molar tooth sign is one of the hallmark radiological features for the diagnosis of Joubert syndrome.

Other major neuroanatomical abnormalities that have been observed in Joubert syndrome are dysplasia/hypoplasia of the inferior olives, accessory olives, the solitary nuclei and tracts, and the nucleus and spinal tracts of cranial nerve V. In addition, elongation of the locus coeruleus has been consistently observed. Anomalies have been reported in the dorsal column nuclei, the descending trigeminal tract, the basis pontis, and the reticular formation in individuals with Joubert syndrome. Marked dysplasia of the caudal medulla has been noted specifically with neuronal swelling and aberrant axonal spheres (immunoreactive for neurofilaments) in the malformed nuclei gracilis and cuneatus. The posterior median sulcus also has been reported to be missing in individuals with Joubert syndrome.

Other less frequent neuroanatomical abnormalities in Joubert syndrome include retinal dystrophy, Leber

congenital amaurosis (retinal degenerative disorder), and ocular colobomas (defects in the lens, iris, retina, or optic disc). There are reported cases of individuals with Joubert syndrome that have focal regions of cortical dysplasia or polymicrogyria (a neuronal migration disorder characterized by many small folds on the surface of the cerebral cortex), but the incidences of these cortical findings in Joubert syndrome are currently unknown. However, these findings are mostly based on MRI, and subtle alterations in the cerebral cortex might be missed and need to be examined at the microscopic level. Moreover, it is important to note that many of the major abnormalities of the cerebellum, superior cerebellar peduncles, and corticospinal tract, discussed above, are primarily based on MRI imaging; however, there have been a few cases of Joubert syndrome that were analyzed histologically which verified these MRI findings. However, more neuropathological studies are needed in order to determine whether there are more subtle abnormalities in Joubert syndrome in other regions of the brain beyond the detection/resolution limits of MRI. As a whole, these neuroanatomical findings suggest that Joubert syndrome is primarily a disorder of the posterior midline of the brain involving defective patterning and axonal guidance in these regions.

Clinicopathologic Correlations in Joubert Syndrome

Given the knowledge of some of the brain anomalies and behavioral deficits in Joubert syndrome, we can begin to explain the neuroanatomic underpinnings of some of the abnormal behaviors associated with Joubert syndrome. The most obvious behavioral correlate for the cerebellar malformations in Joubert syndrome would be clumsiness (ataxia), since the cerebellum is critical for mediating the ability to maintain posture and movement. However, hypotonia is also observed in individuals with Joubert syndrome and the cerebellar malformations in Joubert syndrome could explain the hypotonia; however, it also might be explained by cerebral cortical malformations. The cognitive difficulty, ranging from mild to severe mental retardation, that is present in Joubert syndrome, could be explained by the cerebellar malformations given the role of the cerebellum in motor learning and the more newly evolving research demonstrating the importance of the cerebellum in cognition. Alternatively, the cognitive difficulty also might be explained by the presence of cerebral cortical abnormalities. The abnormal eye movements indicative of Joubert syndrome could result from the absence of the vermis since the anterior vermis is one of the brain structures important in eye movements; however, the extensive brain stem malformations in patients are an alternative explanation for the aberrant eye movements. That is because the tectoparapontine reticular formation pathway and the basis pontis nuclei, which are disrupted in Joubert syndrome, are important for the proper movement of the eyes. The nuclei gracilis and cuneatus in the brain stem are important structures in mediating normal breathing patterns. The irregular breathing (hyperpnea) that is found, especially in young children with Joubert syndrome, may be the result of malformations in these nuclei. Lastly, one other major behavioral manifestation that occurs in individuals with Joubert syndrome is mirror movements. The corticospinal fibers that normally decussate in the medulla are one of the main neuroanatomical substrates responsible for producing alternating movement of the arms and legs. Since the corticospinal tract fails to properly decussate in Joubert syndrome, this lack of midline crossing may be responsible for the mirror movements. However, higher cortical functioning also is important in mediating alternating motor behavior, so cerebral cortical involvement also has to be considered. Overall, the neuroanatomical abnormalities associated with Joubert syndrome can explain the behavioral manifestations present in this disorder.

Joubert Syndrome Is a Clinically Heterogeneous Syndrome

It is now becoming increasingly clear that certain clinical and neuropathological characteristics of Joubert syndrome can be observed in subclasses of Joubert syndrome in addition to other disorders that also involve the cerebellum and brain stem. As mentioned earlier, classic Joubert syndrome is characterized by mainly the molar tooth sign and varying degrees of cerebellar vermis hypoplasia/aplasia. In addition, there are at least three distinct subclasses of classic Joubert syndrome that include the hindbrain malformations and either (1) retinal dystrophy, (2) polymicrogyria, or (3) renal dysfunction (all discussed in the next section). These multiple subclasses of Joubert syndrome demonstrate that this syndrome is clinically heterogeneous.

However, the molar tooth sign is not only found in Joubert syndrome. It can also be observed in patients with COACH syndrome (Cerebellar vermis hypoplasia/aplasia, Oligophrenia (type of congenital mental retardation), Ataxia, ocular Coloboma, and Hepatic fibrosis); Dekaban–Arima syndrome (agenesis of the cerebellar vermis, ocular abnormalities (Leber

congenital amaurosis and colobomas), renal cysts, and liver disease); Malta syndrome (cerebellar vermis abnormalities, hydrocephalus, encephalocele (protusion of the brain through a fissure in the skull), Leber congenital amaurosis, and renal cysts); Senior–Loken syndrome (cerebellar vermis hypoplasia/aplasia, Leber congenital amaurosis, and nephronophthisis); and Varadi–Papp syndrome (or orofaciodigital syndrome VI: disorder comprised of cerebellar vermis abnormalities, psychomotor retardation, polydactyly, and cleft lip/palate). Whether these syndromes should be considered as additional subclasses of Joubert syndrome remains to be determined. However, as researchers continue to search for genes that can cause these disorders, defining the clinical and neuropathological sequelae in individuals with these various syndromes will be critically important so as to be confident in the pooling of various individuals for genetic mapping studies.

The Genetics of Joubert Syndrome: A Genetically Heterogeneous Syndrome

Genetic research has demonstrated that Joubert syndrome is genetically heterogeneous. The early studies of Joubert syndrome defined the mode of inheritance for this disorder as autosomal recessive. The incidence of Joubert syndrome is not definitively known, but best estimates range from 1:50 000 to 100 000 births. Most of the genetic research that aimed at identifying causative genes for Joubert syndrome utilized linkage analysis and positional cloning. These studies relied mostly on homozygosity mapping in consanguineous families (where first and second cousins produce offspring). The use of consanguineous pedigrees greatly increases the power of homozygosity mapping. Homozygosity mapping is based on the fact that recessive diseases, such as Joubert syndrome, require two mutant copies of a gene (allele) for the disease to become manifest. Because of linkage disequilibrium, genetic regions surrounding a disease locus will be homozygous in individuals with the disease. By identifying genomic regions that are homozygous in individuals with disease, one can begin to locate regions of the genome that may contain a mutated, causative gene for the disease.

The JBTS1 Locus

The first genetic loci that was linked to Joubert syndrome was identified on chromosome 9q34.3 in 1999 and is now referred to as JBTS1. The JBTS1 locus appears to correspond to classic Joubert syndrome involving the pure CNS phenotype of cerebellum and brain stem malformations. To date, no causative gene has been identified for this locus.

The JBTS2 Locus

The second genetic locus for Joubert syndrome (JBTS2) was found in 2003 on chromosome 11p12-q13.3. Individuals with Joubert syndrome that are linked to the JBTS2 locus have the characteristic brain features of classic Joubert syndrome, but also have renal, liver, and retinal involvement. No gene has yet been found for the JBTS2 locus.

The JBTS3 Locus

In 2004, the third locus for Joubert syndrome was identified on chromosome 6q23 and was called JBTS3. JBTS3 appears to represent classic Joubert syndrome with cerebellum and brain stem abnormalities is sometimes associated with cerebral cortical malformations, most commonly polymicrogyria. Reports of kidney disease (nephronophthisis) and retinal dystrophy have also been noted. In the JBTS3 loci, mutations were identified in the *Abelson helper integration site 1* gene (*AHI1*) in individuals with Joubert syndrome. Based on sequence analysis, AHI1 appears to be a cytoplasmic phosphoprotein with multiple protein-binding domains including an N-terminal coiled-coil domain, WD40 repeats, and a C-terminal SH3 domain. To date, mutations have been identified throughout the *AHI1* gene including the WD40 repeats and the SH3 domain; however, no mutations have been identified in the coiled-coil domain. Remarkably, even though these mutations occur in very different regions of *AHI1*, there do not appear to be any obvious genotype–phenotype relationships.

Human *AHI1* mRNA is highly expressed in the developing fetal brain and kidney. In adult human brain, *AHI1* mRNA is highly expressed in cerebellum and cerebral cortex, with weaker expression in the putamen. Murine *Ahi1* mRNA is found in the developing embryo as early as embryonic day 10.5 (E10.5) and continuing into adulthood; however, the expression level decreases significantly following the first week after birth. Mouse *Ahi1* mRNA expression is most pronounced in the developing and mature hindbrain and spinal cord (**Figures 2(a)–2(f)**). *Ahi1* expression is found in the cerebellar primordium at E12.5 to E16.5, peaking around E14.5 (**Figures 2(a) and 2(a′)**). The deep cerebellar nuclei, especially the dentate nuclei, have particularly high levels of *Ahi1* expression with the cerebellar hemispheres having proportionately weaker expression (**Figures 2(f) and 2(f′)**). The *Ahi1* expression patterns in the cerebellum remarkably parallel the cerebellar malformations seen in individuals with Joubert syndrome, specifically the severe abnormalities of the midline structures (cerebellar vermis and deep cerebellar nuclei). The

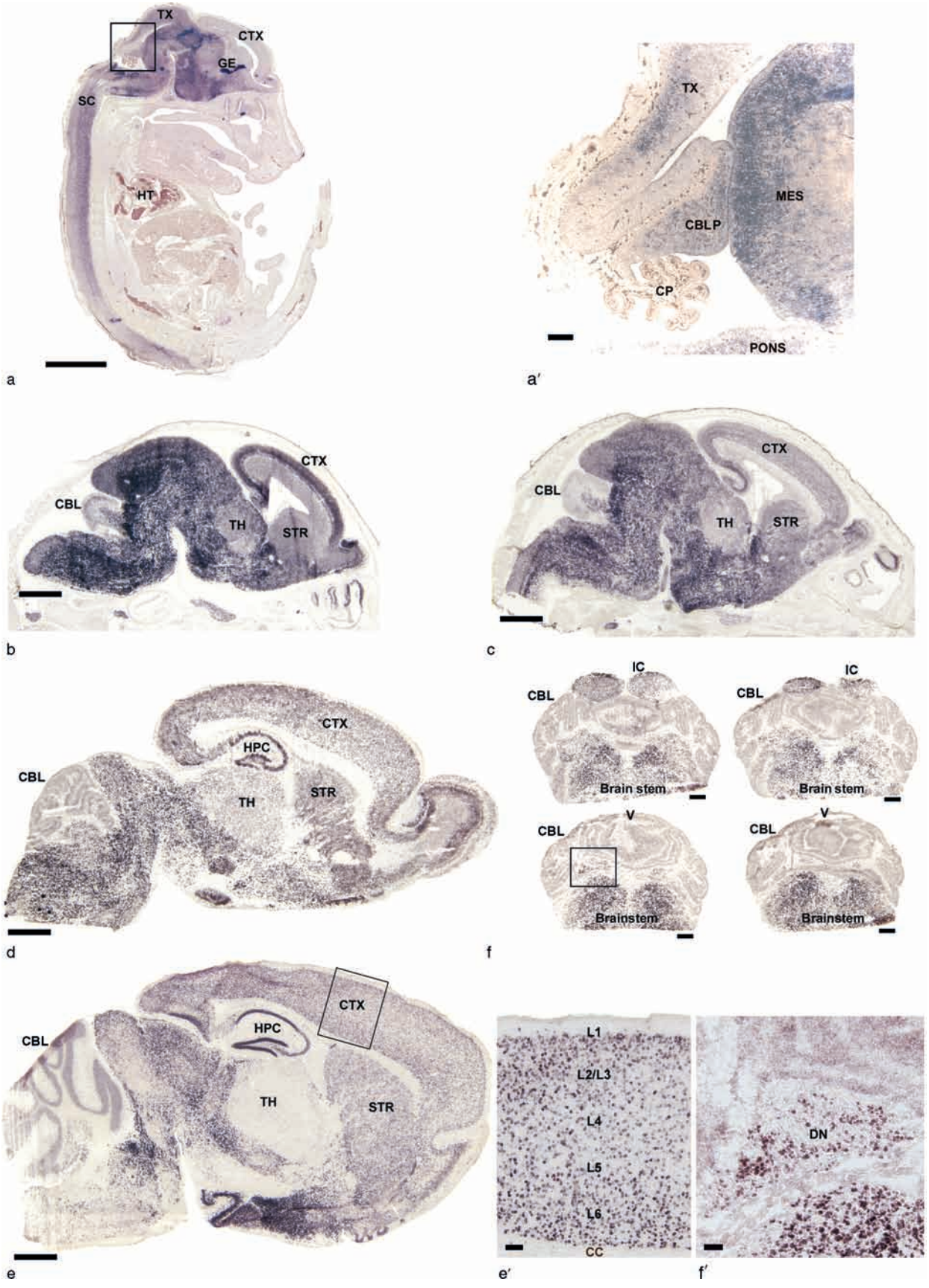

Figure 2 Continued.

low-level expression of *Ahi1* in the cerebellar hemispheres also parallels the Joubert syndrome pathology since the cerebellar hemispheres are largely spared. These developmental expression patterns suggest that *Ahi1* may be an important mediator in the patterning of the cerebellum, particularly the cerebellar vermis.

The cell bodies giving rise to the axonal tracts that fail to decussate at the midline in Joubert syndrome express *Ahi1* mRNA. Specifically, *Ahi1* mRNA expression is found in the layer V cell bodies of the cerebral cortex that give rise to the corticospinal tract (**Figures 2(e) and 2(e′)**). In Joubert syndrome, it is the corticospinal tract that fails to decussate and cross the midline in the medullary pyramids. In addition, *Ahi1* expression is also found in the cell bodies of the dentate nucleus of the cerebellum (**Figures 2(f) and 2(f′)**). These are the cells that give rise to the superior cerebellar peduncles whose axons fail to cross the midline in Joubert syndrome. Overall, these results suggest that *Ahi1* may be a critical cell autonomous modulator of axonal decussation, given its expression in the neurons which give rise to midline crossing, and since human *AHI1* mutations result in a lack of decussation in Joubert syndrome.

Expression of *Ahi1* mRNA also is observed in many other brain structures not always associated with Joubert syndrome, such as the hippocampus, basal ganglia, hypothalamus, and thalamus (**Figures 2(a)–2(f)**). Assuming that the human expression pattern parallels the mouse descriptions of *Ahi1* expression, it is then plausible that some of the behavioral features of Joubert syndrome might be mediated by structures that are not usually malformed in patients.

Another potential glimpse into the role of *AHI1* in brain development comes from work suggesting that *AHI1* may have been a target of evolutionary selection. Differences in the structure of the nervous system rely heavily on genes that control its development. Evolution allows for changes in such genes that specify the nervous system resulting in mutations occurring in genes that can alter the structure of the nervous system in either an advantageous or disadvantageous manner. Advantageous mutations presumably would result in a better functioning nervous system whereas disadvantageous mutations might result in poorer functioning nervous system. It is through the accumulation of these advantageous mutations in developmental control genes that our nervous systems have attained their current level of higher-order functioning. The *AHI1* gene is one such developmental control gene that has changed through evolution and may be involved in the evolution of human higher-order functioning, although there are many other genes that are also involved.

Comparisons of primate *AHI1* and rodent *Ahi1* resulted in the observation that whereas the primate protein contained a 140-amino acid N-terminal coiled-coil domain, the mouse protein was lacking this functional domain. Moreover, there does not appear to be a homologous protein similar to AHI1 in lower species such as *Caenorhabditis elegans* or *Drosophila*. As a whole, this analysis suggested that the N-terminal coiled-coil domain might be particularly dynamic evolutionarily.

Further analyses in which *AHI1* was sequenced and compared in various nonhuman primates and humans suggested that the full-length *AHI1* gene was under positive evolutionary selection in humans, with the 5′-half of the *AHI1* gene, which includes the coiled-coil domain, giving an evolutionary signature that was particularly suggestive of positive evolutionary selection along the human lineage. Positive evolutionary selection is when natural selection favors a single mutation and therefore continuously propagates this mutation in a species, presumably because it is adaptive. Overall, these studies suggest that given the aberrant axonal patterns and associated motor abnormalities in individuals with Joubert syndrome, combined with the molecular evolutionary pattern of selection, *AHI1* may be involved in some aspect of the evolution of the distinctive motor programs that characterize humans.

Figure 2 Spatial expression of murine *Ahi1* during brain development and postnatal development. *In situ* hybridizations for *Ahi1* mRNA in sagittal sections of mouse brain at E14.5 (a), E16.5 (b), P0 (c), P6 (d), and P21 (e). *Ahi1* mRNA is widely expressed throughout the entire brain with the greatest level of expression occurring in the hindbrain. Weaker expression also occurs in the developing cerebellar primordium (E12.5–E16.5) with the highest expression occurring at E14.5. *Ahi1* expression is also present in the olfactory bulb, hippocampus (dentate gyrus, CA1–3), neocortex, basal ganglia, and spinal cord. (f) *Ahi1* mRNA expression in sequential sections through the cerebellum of a P6 mouse brain demonstrating the relative weak expression of this gene in the cerebellar hemispheres following embryogenesis. However, *Ahi1* expression is high in the cerebellar dentate nuclei, the output nucleus of the cerebellum that forms the superior cerebellar peduncle (f, f′), and the deep layer cortical neurons (layer 5) that form the corticospinal tract (b–e, e′). Scale bar 6 = 1 mm (a–e), 500 μm (f), and 100 μm (a′, e′, and f′). CBL, cerebellum; CBLP, cerebellar primordium; CC, corpus callosum; CP, choroid plexus; CTX, cerebral cortex; DN, dentate nucleus of the cerebellum; GE, ganglionic eminences; HPC, hippocampus; HT, heart; IC, inferior colliculus; L, layers of the cerebral cortex; MES, mesencephalon; SC, spinal cord; STR, striatum; TH, thalamus; TX, tectum; V, cerebellar vermis. Adapted from Ferland RJ, Eyaid W, Collura RV, et al. (2004) Abnormal cerebellar development and axonal decussation due to mutations in *AHI1* in Joubert syndrome. *Nature Genetics* 36: 1008–1013, with permission.

The JBTS4 Locus

Since individuals with Joubert syndrome can also have nephronophthisis, mutational analysis was performed on one of the major nephronophthisis genes, *nephrocystin 1* (*NPHP1*), on chromosome 2q13. Homozygous deletions of *NPHP1* were identified in individuals with a mild form of Joubert syndrome with nephronophthisis. However, this deletion mutation in *NPHP1* is considered a relatively rare occurrence, but demonstrates the potential for the co-expression of both Joubert syndrome and nephronophthisis having a singular genetic cause. NPHP1 has been localized close to cell–cell junctions and can form complexes with proteins involved in focal adhesion formation and adherens junctions suggesting possible roles in cell proliferation and cell signaling. It will be important to establish the role of *NPHP1* in brain development, particularly in hindbrain development, and to determine whether there is an interaction between *AHI1* and *NPHP1*.

The JBTS5/6 Locus

Most recently, a new genetic locus for Joubert syndrome was identified on chromosome 12q21.32–12q21.33 called JBTS5/6. Individuals with Joubert syndrome that demonstrate linkage to JBTS5/6 have the brain phenotype consistent with classic Joubert syndrome, but also have severe retinal and renal abnormalities. Mutational analysis revealed multiple mutations in the *CEP290* gene encoding the centrosomal protein of 290kD. The *CEP290* gene is also called *NPHP6*. The CEP290 protein is composed of 13 putative coiled-coil domains, a region of homology to SMC (structural maintenance of chromosome) chromosome segregation ATPases, a bipartite nuclear localization signal, six RepA/Rep+ protein KID motifs (involved in chromosome segregation and cell cycle regulation), three tropomyosin homology domains (possible interactions with actin filaments), and an ATP/GTP binding site motif A with a P-loop (possibly involved in active transport).

Murine *Cep290* mRNA is expressed as early as E7 and persists into adulthood. Expression of *Cep290* mRNA appears to be widely expressed in the body, but is most prominent in the kidney and brain, particularly in the retina and hindbrain. *Cep290* mRNA *in situ* hybridization studies on the developing cerebellum demonstrates expression in both the external and internal granule cells. Since the external granule cells are one of the main proliferative cell types in the cerebellum, this suggests a role for *Cep290* in cerebellar neurogenesis. In further support of this role, *Cep290* mRNA is also present in the dividing cells of the telencephalon. Since the internal granule cell layer is one of the final resting places for cells that have migrated from the external granule cell layer, expression of *Cep290* mRNA in the internal granule cell layer might suggest a role of *Cep290* in granule cell migration.

As the name of the *CEP290* gene implies, it encodes a centrosomal protein. Experimental evidence suggests that *CEP290* is localized to the area surrounding the centrioles and colocalized with the centrosomal protein, γ-tubulin, all consistent with its presumptive role in centrosome function. Moreover, CEP290 localization to centrosomes is not only dynamic during the cell cycle, given that it redistributes to the cytosol in prometaphase, but its localization also appears to occur independently of microtubules and dyneins. These results suggest that CEP290 is a component of the centrosome, and therefore has at present an undetermined role in the microtubule-organizing center of the cell and in the progression of the cell cycle.

Phylogenetic analysis of *CEP290* determined that this gene was present in both zebra fish (*Danio rerio*) and sea squirts (*Ciona intestinalis*). Knockdown of *cep290* expression in zebra fish, by injection of *cep290* antisense morphilinos into the developing zebra fish embryos, resulted in defects in retinal, cerebellar, optic cavity, and renal (cyst formation) development. These abnormalities were noted to be reminiscent of the malformations occurring in humans with *CEP290* mutations.

Joubert Syndrome: A Disease of Primary Cilia (Ciliopathy)?

Given the clinical and anatomical overlap between Joubert syndrome and other related diseases, and the identification of common genes that can cause these disorders, it is becoming increasingly clear that cilia may be important in brain and renal development. The role of cilia in the development and functioning of the kidney has been appreciated for some time given the involvement of multiple genes, which localize to the cilia, in nephronophthisis and polycystic kidney disease. However, only until recently have studies begun to address the role of cilia in the brain.

Cilia are organelles that are extensions of the plasma membrane of cells. They are approximately 5–10 μm in length. The microtubule-organizing center, called the basal body in cilia, is located at the base of the cilia. These basal bodies are composed of centrioles. Some cilia constantly beat (motile cilia) and others that do not beat are thought to act as sensors (nonmotile cilia). Nonmotile cilia are the predominant form in the CNS. Cilia have an outer membrane surrounding a core called the axoneme. The axoneme contains nine pairs of microtubule doublets. Motile cilia have this

configuration of the axoneme, but also contain two additional microtubule singlets attached to the outer microtubule doublets (referred to as the 9+2 structure). Nonmotile cells, such as neural cells, just have the nine pairs of microtubule doublets with no singlets (referred to as the 9+0 structure). This 9+0 cilia structure is found in the outer segment of the rod photoreceptor in the eye, in the terminal fibers containing the odorant receptors in olfactory neurons, and in the Purkinje cells and granule cells of the cerebellum. Motile cilia have been observed in the brain, particularly along the ependymal surface of the lateral ventricles where these beating cilia are required for normal cerebrospinal fluid flow through the ventricles, for controlling gradients of guidance molecules in the cerebrospinal fluid, and for directional migration of neural cells. Based on this information, it is apparent that cilia have important roles in the development and functioning of the brain.

One of the proteins known to cause both nephronophthisis and Joubert syndrome, NPHP1, has been shown to localize to primary cilia, particularly in the basal bodies, in renal epithelial cells. Likewise, CEP290 is also localized to the base of the cilia. However, morpholino antisense knockdown of *cep290* expression in zebra fish did not produce any observable changes in either cilia length or cilia motility in the pronephroi of zebra fish. This suggests that *cep290* may not have a critical structural or mechanical role in cilia function, but may have an important role in the signaling of the cilia. Whether AHI1 is localized to the basal bodies of the cilia remains to be determined. In fact, all of the Joubert genes identified to date, *AHI1*, *NPHP1*, and *CEP290*, appear to be cytoplasmic proteins and not membrane-spanning proteins, further suggesting that these genes may be involved in the signal transduction events that occur following membrane receptor activation. In fact, the sonic hedgehog and Wnt signaling pathways have been implicated in the formation and functioning of cilia leading to critical roles in the patterning of the body plan in vertebrates. Future studies are needed which specifically examine whether any of the Joubert syndrome genes have important roles in the development of the CNS and kidney through a role in primary cilia function.

See also: Autism; Brainstem and Cranial Nerves; Brainstem Control of Eye Movements; Cerebellum and Oculomotor Control; Cerebellum: Clinical Pathology; Corticospinal Development; Fragile X Associated Tremor Ataxia Syndrome; Friedreich Ataxia.

Further Reading

Dixon-Salazar T, Silhavy JL, Marsh SE, et al. (2004) Mutations in the *AHI1* gene, encoding jouberin, cause Joubert syndrome with cortical polymicrogyria. *American Journal of Human Genetics* 75: 979–987.

Ferland RJ, Eyaid W, Collura RV, et al. (2004) Abnormal cerebellar development and axonal decussation due to mutations in *AHI1* in Joubert syndrome. *Nature Genetics* 36: 1008–1013.

Jiang X, Hanna Z, Kaouass M, et al. (2002) *Ahi-1*, a novel gene encoding a modular protein with WD40-repeat and SH3 domains, is targeted by the Ahi-1 and Mis-2 provirus integrations. *Journal of Virology* 76: 9046–9059.

Joubert M, Eisenring JJ, and Andermann F (1968) Familial dysgenesis of the vermis: A syndrome of hyperventilation, abnormal eye movements and retardation. *Neurology* 18: 302–303.

Keeler LC, Marsh SE, Leeflang EP, et al. (2003) Linkage analysis in families with Joubert syndrome plus oculo-renal involvement identifies the CORS2 locus on chromosome 11p12-q13.3. *American Journal of Human Genetics* 73: 656–662.

Lagier-Tourenne C, Boltshauser E, Brievik N, et al. (2004) Homozygosity mapping of a third Joubert syndrome locus to 6q23. *Journal of Medical Genetics* 41: 273–277.

Parisi MA, Bennett CL, Eckert ML, et al. (2004) The *NPHP1* gene deletion associated with juvenile nephronophthisis is present in a subset of individuals with Joubert syndrome. *American Journal of Human Genetics* 75: 82–91.

Parisi MA, Dohery D, Eckert ML, et al. (2006) *AHI1* mutations cause both retinal dystrophy and renal cystic disease in Joubert syndrome. *Journal of Medical Genetics* 43: 334–339.

Saar K, Al-Gazali L, Sztriha L, et al. (1999) Homozygosity mapping in families with Joubert syndrome identifies a locus on chromosome 9q34.3 and evidence for genetic heterogeneity. *American Journal of Human Genetics* 65: 1666–1671.

Sayer JA, Otto EA, O'Toole JF, et al. (2006) The centrosomal protein nephrocystin-6 is mutated in Joubert syndrome and activates transcription factor ATF4. *Nature Genetics* 38: 674–681.

Utsch B, Sayer JA, Attanasio M, et al. (2006) Identification of the first *AHI1* gene mutations in nephronophthisis-associated Joubert syndrome. *Pediatric Nephrology* 21: 32–35.

Valente EM, Marsh SE, Castori M, et al. (2005) Distinguishing the four genetic causes of Joubert syndrome-related disorders. *Annals of Neurology* 57: 513–519.

Valente EM, Brancati F, Silhavy JL, et al. (2006) *AHI1* gene mutations cause specific forms of Joubert syndrome-related disorders. *Annals of Neurology* 59: 527–534.

Valente EM, Silhavy JL, Brancati F, et al. (2006) Mutations in *CEP290*, which encodes a centrosomal protein, cause pleiotropic forms of Joubert syndrome. *Nature Genetics* 38: 623–625.

Yachnis AT and Rorke LB (1999) Neuropathology of Joubert syndrome. *Journal of Child Neurology* 14: 655–659.

Relevant Websites

http://www.joubertsyndrome.org – Joubert Syndrome Foundation and Related Cerebellar Disorders.

http://www.ninds.nih.gov – NINDS Joubert Syndrome Information Page, National Institute of Neurological Disorders and Stroke, National Institutes of Health.

Kainate Receptor Functions

J Lerma, Instituto de Neurociencias de Alicante Consejo Superior de Investigaciones Científicas-Universidad Miguel Hernandez, San Juan de Alicante, Spain

© 2009 Elsevier Ltd. All rights reserved.

After finding specific kainate receptors (KARs) in central neurons, physiologists suffered many problems trying to elucidate the role of these receptors in the nervous system, principally due to the lack of pharmacological tools to activate or antagonize KARs. Indeed, many of the pharmacological agonists and antagonists that activate KARs also interact with α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs). This lack of pharmacological specificity has hindered our understanding of KAR function for several years. Thus, the discovery that 2,3-benzodiazepines, and particularly GYKI 53655 (LY300168 or the active isomer LY303070), antagonize AMPARs but not KARs represented a significant breakthrough in the field. More recently, other compounds that specifically act on KAR subunits have been developed and mice deficient for KAR subunits have been generated, paving the way for the study of the synaptic physiology of KARs. As a result, KARs have been seen to play a role in synaptic transmission, influencing both neuronal excitability and information transfer in the brain. An interesting characteristic of KARs that has emerged is that they use two forms of signaling: a canonical pathway that involves ion flow and another, more unexpected noncanonical signaling pathway that links KAR activation to G-proteins and second-messenger cascades.

Elementary Properties of Kainate Receptor Channels

KARs are nonselective cation channels. Therefore, the single-channel conductance of some KAR assemblies has been determined. According to stationary noise analysis, the single-channel conductance of KARs lies in the picosiemen range with values of 2.9 and 5.4 for homomeric glutamate receptor (GluR)5(Q) or GluR6(Q), respectively. A more detailed study of channel openings revealed three subconductance levels that are approximately two or three times the size of the smallest current. Like AMPARs, the replacement of Q by R in the channel domain as the result of RNA editing reduces the single-channel conductance of KARs by more than one order of magnitude. Furthermore, the conductance is also altered by heteromerization. For instance, the elementary conductance of GluR5(R) channels is less than 200 fS, but it increases to 950 fS on co-assembly with KA2. Similarly, GluR6(R) homomers have a single-channel conductance between 230 and 260 fS, whereas the value for GluR6(R)/KA2 heteromers lies between 570 and 700 fS. The effect of KA2 on the conductance of unedited homomeric receptors is more subtle, with increases to 4.5 and 7.1 pS, for GluR5 and GluR6, respectively. Interestingly, native receptors display similar values, a conductance of 2–4 pS having been measured in dorsal root ganglion (DRG) cells and three subconductance levels, similar to those found for GluR5(Q) or GluR5(Q)/KA2, have been identified. These findings are in accordance with the abundant expression of GluR5 in DRG neurons. In addition, the conductance of KARs expressed in immature proliferating cerebellar granule neurons has been determined to be ~1 and ~4 pS, the channels spending most of their open time in the 4 pS state.

Desensitization of Kainate Receptor Currents

Rapid desensitization is one of the major characteristic features of KARs. The time course of current decay in the continued presence of an agonist follows a single or double exponential decay. The speed of desensitization is dependent on the receptor subtype and on the cell type being analyzed. GluR6 homomers and native hippocampal KARs have desensitization time constants of ~11–13 ms, a value similar to that found for recombinant channels in excised patches or lifted cells. However, KARs recover slowly from the desensitized state, and this recovery is dependent on the nature of the agonist. Indeed, whereas the recovery from desensitization by a pulse of kainate takes over 1 min, the receptor requires only 15 s when glutamate is the agonist. The subunit composition also affects recovery from desensitization; thus, whereas GluR5 homomeric receptors recover in about 1 min, the time constant for the recovery of GluR5/KA2 heteromers is just 12 s. These dramatic differences in the time scale of desensitization and recovery imply that the equilibrium between the two states is strongly displaced toward the inactive state. In other words, the receptors spend most of their time desensitized. This contradicts the fact that low concentrations of kainate and glutamate have a striking effect on native KARs, as seen in neurons in brain slices. There is still no clear explanation for this behavior, which is not predicted from rapid

activation and inactivation kinetics. However, it is possible that unknown proteins interact with native KARs, altering their gating properties, as seen for other receptors and channels.

Pharmacological Aspects

As already mentioned, a major difficulty in understanding KAR function has been the lack of a specific pharmacology for these receptors. For instance, kainate, albeit showing a clear preference for KARs, has a significant action on AMPARs at relatively low doses. Indeed, there is only a 5–30 times difference between the apparent affinity of these two receptors for kainate. There has also been some progress concerning selective KARs agonists. Some derivatives of willardiine have been developed as selective KAR agonists (Table 1). Among them, (*S*)-5-iodowillardiine shows a ~130-fold selectivity for KARs (median effective concentration (EC_{50}) = 140 nM). SYM 2081, the (2S-4R) diastereomer of 4-methylglutamate, displays a potency three orders of magnitude higher for KARs than for AMPARs, both in binding and in functional assays, but the selectivity of this molecule for kainate over *N*-methyl-D-aspartate receptors (NMDARs) is significantly lower. Although its pharmacological profile is incomplete, SYM 2081 does not seem to show the subunit specificity observed for ATPA and for (*S*)-5-iodowillardiine; it elicits rapidly desensitizing currents on GluR5 and GluR6 homomeric channels. Due to this property, this compound has also been used as a functional antagonist of the KARs because it efficiently desensitizes KARs at low concentrations (Table 2).

Similarly, the prototypic non-NMDAR antagonists, 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX), 6,7-dinitroquinoxaline-2,3-dione (DNQX) and 6-nitro-7-sulphamoylbenzo[f]quinoxaline-2,3-dione (NBQX), just show poor selectivity between AMPARs and KARs, although it has been shown that low concentrations of NBQX do not have any action on KARs. However, the separation of KAR-mediated responses from the AMPAR-mediated currents is possible by using 2,3-benzodiazepine GYKI 53655 (Lilly's code LY300168), which is selective for antagonizing AMPARs. Therefore, in the presence of GYKI53655, or its active isomer LY303070, CNQX could be used as a KAR antagonist (Table 2). Several other very useful compounds have been developed lately. In particular, LY382884 ((3S, 4aR, 6S, 8aR)-6-((4-carboxyphenyl) methyl-1,2,3,4,4a,5,6,7,8,8a-decahydro isoquinoline-3-carboxylic acid), seems to antagonize KARs at concentrations that do not affect AMPARs or NMDARs. Actually, LY382884 is a selective antagonist at neuronal KARs containing the GluR5 subunit (Table 2).

Table 1 Kainate receptor agonists

Compound	*Selectivity*
ATPA	GluR5-contaning; GluR6/KA2
S-5-iodowillardine	GluR5-containing; GluR6/KA2; GluR7/KA2
Kainate	All assemblies
Domoate	GluR5- and GluR6-containing.
AMPA	GluR5-containing; GluR6/KA2; GluR7/KA1
Me-Glutamate (SYM 2081)	GluR5- and GluR6-containing

AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; GluR, glutamate receptor.

Table 2 Kainate receptor antagonists

Compound	*Selectivity*	*Mechanism*
CNQX	All assemblies; more potent at AMPAR	Competitive
LY382884	GluR5-containing	Competitive
UBP296	GluR5-containing	Competitive
NS3763	Homomeric GluR5	Noncompetitive
LY377770	GluR5-containing	Competitive
Kynurenate	GluR5-containing and homomeric GluR6	Competitive
SYM2081	GluR5- and GluR6-containing	Desensitization

AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; CNQX, 6-cyano-7-nitroquinoxaline-2,3-dione; GluR, glutamate receptor.

Kainate Receptors Have a Fundamental Role in Controlling Synaptic Neurotransmitter Release

The discovery that GYKI 53655 (or LY303070, the active isomer) antagonizes AMPARs but not KARs paved the way for the study of the participation of KARs in synaptic responses. Nowadays, there is considerable evidence suggesting that KARs may be situated on both sides of the synapse. A considerable amount of effort has been devoted to determining the role of presynaptic KARs in controlling transmitter release, particularly in the hippocampus. In this structure, KARs are widely distributed in presynaptic buttons, and they can bidirectionally regulate the release of glutamate at the mossy fibers to CA3 synapses. The mild activation of KARs enhances the release of glutamate, whereas when they are more intensely activated glutamate release is inhibited. As such, a proportion of the high-frequency-dependent facilitation, a hallmark of the mossy fiber synapses, can be attributed to the activation of presynaptic KARs by synaptically released glutamate, as if low concentrations of exogenous kainate had been applied (Figure 1(b)). Therefore, the characteristic short-term plasticity of mossy

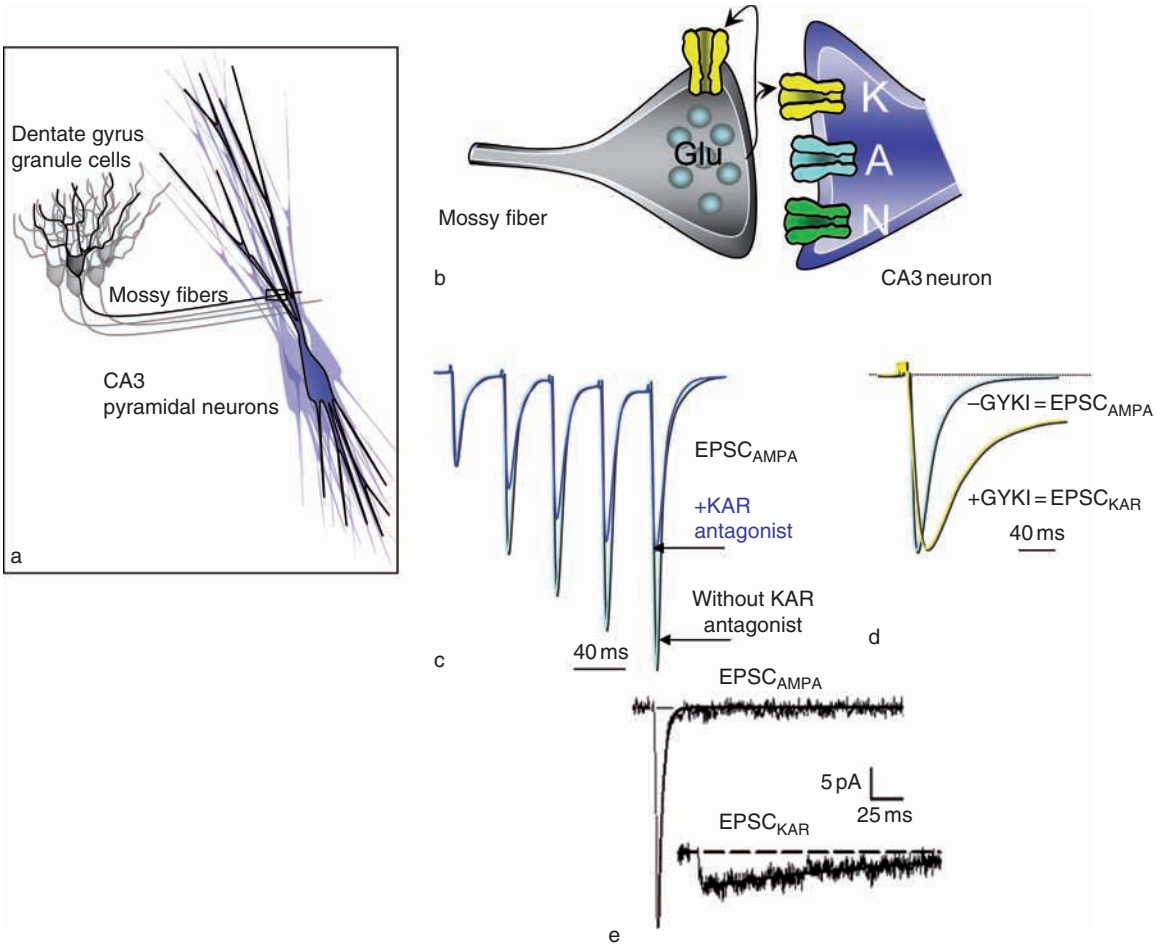

Figure 1 Proposed roles of kainate receptors (KARs) in synaptic transmission: (a) mossy fiber-CA3 synapses; (b) synaptic response of the mossy fibers to CA3 pyramidal cells; (c) presynaptic role; (d) postsynaptic current; (e) postsynaptic responses at the thalamo-cortical synapse. These can be isolated in the presence of APV (to block NMDA receptors) and GYKI53655 (to block AMPA receptors). In all cases, the KAR-mediated synaptic responses are much slower than those observed for AMPA receptors and usually present one-tenth their amplitude (in (d), both responses have been normalized). In (c), presynaptic KARs can enhance or inhibit glutamate release, depending of the degree of activation. The best-established example is the synapse made by mossy fibers on to CA3 pyramidal neurons, where KARs seem to be responsible of a part of the high-frequency facilitation, a characteristic of these synapses. KARs also mediate synaptic responses of Schaffer collateral to interneurons (not shown). In (e), EPSCs mediated exclusively by AMPA or KARs have been isolated by minimal stimulation. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; EPSC, excitatory postsynaptic current; Glu, glutamate; NMDA, *N*-methyl-D-aspartate. (d) Adapted from Lerma J (2003) Roles and rules of kainate receptors in synaptic transmission. *Nature Reviews Neuroscience* 4: 481–495. (e) From Kidd FL and Isaac JT (1999) Developmental and activity-dependent regulation of kainate receptors at thalamocortical synapses. *Nature* 400: 569–573.

fiber–CA3 synapses is partly mediated by the long-lasting activation of a kainate autoreceptor. This mechanism has been shown to impose associative properties on mossy fiber long-term potentiation (LTP) because the activity in neighboring mossy fiber synapses, or even associational/commissural synapses, influences the threshold for inducing LTP. However, higher concentrations of kainate, instead of facilitating mossy fiber synaptic transmission, depress it. Interestingly, this phenomenon is reproduced by synaptic activity, such that a brief conditioning of tetanus on associational-commissural fibers enhances the mossy fiber responses, whereas prolonged tetanus depresses it. The exact mechanism involved in the presynaptic facilitation of glutamate release is unclear. One explanation for the facilitation of glutamate release is that presynaptic KARs depolarize synaptic terminals to a level at which further release is enhanced. Also, it has been claimed that a Ca^{2+}-induced Ca^{2+} release from the intracellular stores is required for this phenomenon to occur. Therefore, such a mechanism remains unclear. How the larger activation of KARs might lead to the inhibition of glutamate release in mossy fiber buttons also remains a matter of debate. It could be the case that the large depolarization of synaptic terminals inactivates

the generation of action potentials, leading to a failure in release. However, recent data show that inhibition (but not facilitation) of release by presynaptic KARs depends on a G-protein-dependent mechanism. This indicates that KARs act independent of their ion channel activity. Therefore, it is possible that the threshold for activating one or other KAR signaling pathway could determine the physiological response.

It is now widely accepted that presynaptic KARs also control the release of the inhibitory neurotransmitter γ-aminobutyric acid (GABA). Although KAR stimulation may enhance the release of GABA at connections between the hippocampal interneurons, stimulation of these receptors inhibits GABA release at interneuron–pyramidal cell synapses. Thus, data indicate that the spillover of glutamate from adjacent terminals could activate KARs at presynaptic GABA terminals (Figure 2(b)). The specific role that the receptors controlling GABA release play in hippocampal synaptic networks remains unclear. However, the infusion of kainate *in vivo* reduces the efficiency of recurrent inhibition, producing a state of hyperexcitability that leads to the generation of recurrent epileptic spikes. Accordingly, KAR antagonists should be able to prevent and/or abolish epileptic

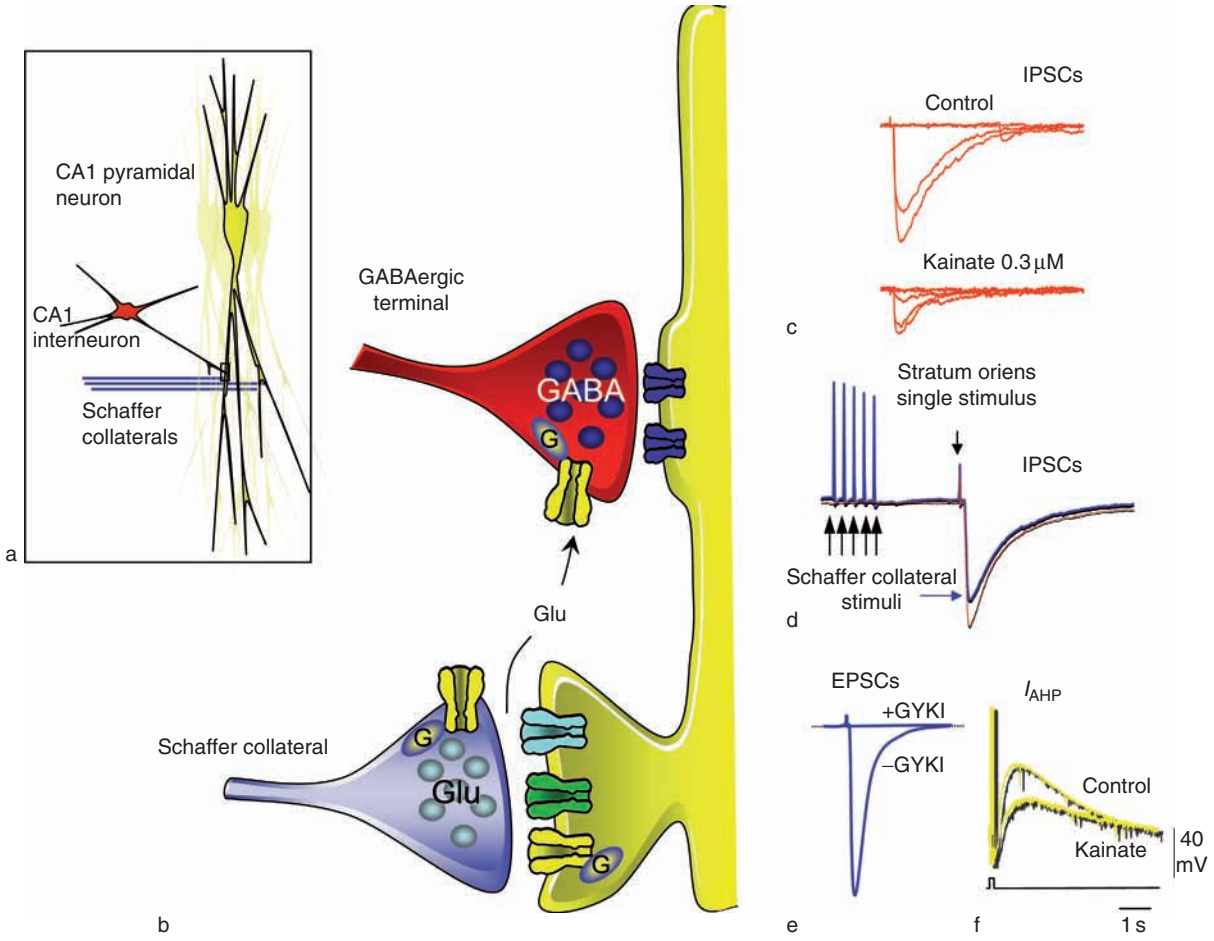

Figure 2 Role of kainate receptors (KARs) in neurotransmitter synaptic release: (a) CA1 pyramidal neurons receive excitatory inputs from the Schaffer collateral and inhibitory actions from interneurons; (b) illustration of an excitatory and an inhibitory synapse on a CA1 pyramidal cell dendrite; (c) a low concentration of kainate selectively reduces the GABA release, as evidenced by a reduction in the evoked IPSC, under pharmacological isolation of $GABA_A$-mediated responses; (d) glutamate could spill over to activate presynaptic KARs present in GABAergic terminals such that the IPSCs (induced by a single shocks to the stratum oriens) are reduced after a train of stimuli is applied on the Shaffer collateral pathway; (e–f) synaptically released glutamate could activate postsynaptic KARs. Synaptically released glutamate could also activate presynaptic (not shown) KARs. Although the activation of postsynaptic KARs does not induce any synaptic current ((e), +GYKI), the afterhyperpolarization current ((f), I_{AHP}) is depressed. KARs have been shown to exert all these actions by activating G-proteins. EPSC, excitatory postsynaptic current; GABA, γ-aminobutyric acid; Glu, glutamate; IPSC, inhibitory postsynaptic current. (d) Adapted from Min MY, Melyan Z, and Kullmann DM (1999) Synaptically released glutamate reduces gamma-aminobutyric acid (GABA)ergic inhibition in the hippocampus via kainate receptors, *Proceedings of the National Academy of Sciences of the United States of America* 96: 9932–9937. (f) From Lerma J (2006) Kainate receptor physiology. *Current Opinion in Pharmacology* 6: 89–97.

activity. This has indeed been demonstrated when epilepsy is induced in rats by a pilocarpine injection. In this model, KAR antagonists (e.g., LY377770, Table 2) not only abolish the established seizure activity but also prevent the development of seizures when they are applied beforehand. Therefore, these and other data indicate that the excitability of the network may be tightly controlled by the extracellular concentrations of glutamate through the transient and/or tonic activation of KARs.

One important issue that has remained a matter of intense debate over the last few years relates to the mechanism by which KARs affect release at these inhibitory synapses. There is evidence indicating that GABA release is inhibited by a G-protein- and protein kinase C (PKC)-dependent mechanism (Figure 2(b)). It is possible that the large barrage of inhibitory postsynaptic current (IPSC) received by pyramidal cells on kainate perfusion may cause some short circuiting, but it is now generally accepted that this is not sufficient to account for such inhibition. Indeed, it should be taken into account that KARs may regulate CA1 GABAergic circuits through two distinct and opposing mechanisms. Whereas the activation of somatodendritic KARs may increase the activity of GABA interneurons, presynaptic KARs may diminish inhibition. These two effects are mediated by receptor populations that are functionally, physically, and pharmacologically distinct, one of which is located in the somatodenritic and/or axonal compartments and the other at the presynaptic terminals directly inhibiting GABA release. Likewise, there is considerable data that point to a prominent role of KARs in the regulation of neuronal excitability through the modulation of the GABAergic system in a variety of brain nuclei, such as the amygdala, the neocortex, the striatum, and the cerebellum.

Kainate Receptors Mediate Part of the Postsynaptic Response at Certain Synapses

Synaptic release of glutamate can also activate postsynaptic KARs, as seen primarily in synapses established by mossy fibers on CA3 neurons. Subsequently, KAR-mediated synaptic responses were demonstrated in Schaffer collaterals–hippocampal CA1 interneurons, at the parallel fiber–Golgi cell synapses, in the basolateral amygdala, in Purkinje cells of the cerebellum, at thalamo-cortical connections on cortical interneurons (Figure 1(e)), in neocortical pyramidal neurons, and in the synapse established by cones on bipolar cells in the retina. Similarly, a synaptic component mediated by KARs has been found in the synapse between afferent sensory fibers and dorsal horn neurons in the spinal cord. With the exception of excitatory postsynaptic currents (EPSCs) in the retina, the KAR-mediated EPSCs all displayed common characteristics such as a low amplitude (approximately an order of magnitude lower than the component mediated by AMPA receptors) and a longer-lasting time course than the EPSC mediated by the AMPA receptor (Figure 1(d)).

What could be the functional significance of a small and long-lasting EPSC for neurons? Although it is not definitively determined, a slow small EPSC, such as that mediated by KARs, offers the possibility of integrating excitatory inputs over a wide time window. Thus, the relative contribution of KARs to synaptic drive could be significant. It has been calculated that KARs elicit sufficient charge transfer so as to exert a substantial impact on the activity of interneurons. Indeed, the kinetics of KAR-mediated EPSPs are sufficiently slow to allow substantial tonic depolarization even during modest presynaptic activity. Such tonic depolarization could account for the observation that single afferent inputs are very effective in driving interneuron spiking. Accordingly, KARs contribute to the ongoing ionotropic glutamatergic transmission in the hippocampus. Therefore, neurons may display unique integration properties in areas where KARs are expressed.

Why native and recombinant KARs present different kinetic properties has not been convincingly explained. However, KARs take direct part of the glutamatergic synapse because these receptors are activated by quantal release of glutamate at both mossy fibers–CA3 synapses and excitatory synapses on CA1 interneurons. In addition, there is evidence that AMPARs and KARs can be segregated at different synapses. This seems to be true at least in CA1 interneurons and at thalamo-cortical synapses, although not in individual mossy fiber synapses. These studies allowed the kinetics of elementary KAR-mediated EPSC in CA1 interneurons to be determined, revealing that they are still slower than the pure AMPAR-mediated EPSC. Therefore, there should be factors such as the dissimilar composition of the synaptic receptors or accessory proteins that may explain the difference between the response kinetics of the native and recombinant receptors studied so far.

The retina constitutes one of the systems in which KARs have been shown to display similar properties to those found in recombinant receptors, specifically at the synapse from cones to the off bipolar cells. In this system, the kinetics of the synaptic responses fits well with the activation and desensitization properties of KARs determined in recombinant systems. In these synapses, the slow kinetics of recovery from desensitization of KARs is exploited in processing the afferent signals rather than suffering the constrictions

of the faster time course of AMPARs. In the dark, a cone releases sufficient neurotransmitter to desensitize most postsynaptic KARs, leaving just a small persistent postsynaptic current. Because the postsynaptic KARs desensitize rapidly and recover from desensitization slowly, little recovery can occur during the brief hyperpolarizations that are produced in the cones by repetitive light stimuli. Therefore, the large increase in conductance that would saturate the membrane voltage if the rate of transmitter release by the cone was high is prevented. Actually, signaling between the cones and the three morphological types of off bipolar cells occurs through distinct postsynaptic receptors. Whereas two cell types use KARs with distinct properties, the other type uses AMPARs. Each receptor recovers from desensitization with a particular time course, and this property contributes to determining the temporal characteristics of synaptic transmission in the retina (Figure 3).

Recent studies have demonstrated that KARs are also involved in long-term synaptic plasticity (LTP). Indeed, mossy fiber LTP may be prevented by the GluR5 antagonist LY382884. This antagonist, however, is ineffective for preventing LTP induced by tetanic stimulation of associational/commissural fibers in the CA3 region or at CA1 synapses. Therefore, a role for KARs in the induction of LTP seems to be specific for the mossy fiber pathway, although pharmacological data are at odds with results obtained from knockout mice. The use of KAR subunit-deficient mice confirmed that KARs play a critical role in mossy fiber LTP. However, mossy fiber LTP was reduced, but not abolished, in GluR6-knockout mice, whereas GluR5-deficient mice exhibited normal LTP at these synapses. Clearly, more investigation is needed to understand the molecular mechanisms by which KARs may play a permissive role at mossy fibers to induce a permanent change in release properties. Also interesting is the fact that at thalamo-cortical synapses in the neonatal barrel cortex, LTP expression is associated with a reduction in the KAR-mediated component of the synaptic response. Thus, as recently confirmed, KAR-mediated synaptic transmission can undergo a form of long-term depression (LTD) that is mechanistically distinct from the LTD seen for AMPAR-mediated transmission.

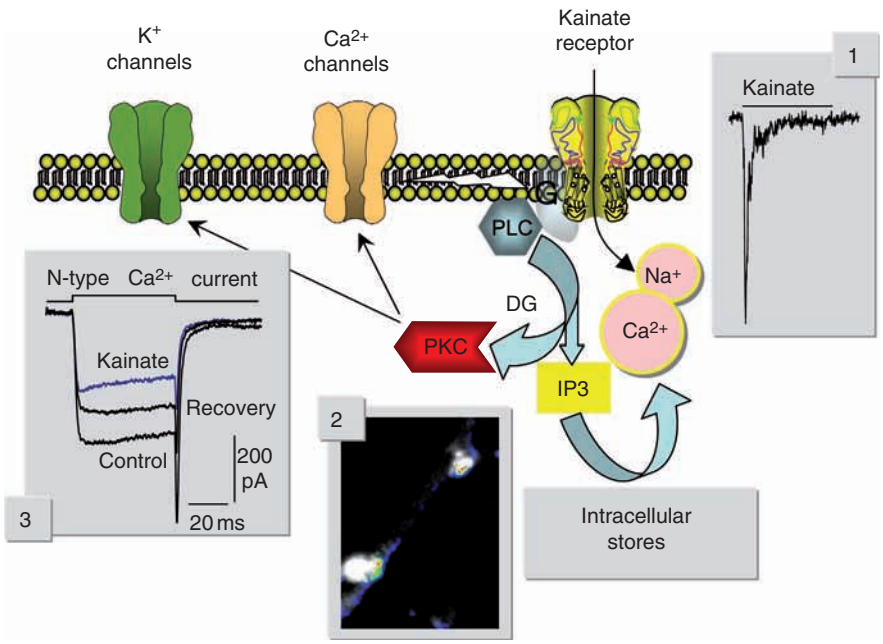

Figure 3 Canonical and noncanonical signaling of kainate receptors (KARs). As ligand-gated ion channels, the activation of KARs evokes an inward current, through which they depolarize the membrane (1). However, they can also set in motion a noncanonical signaling pathway triggered by the activation of a pertussis toxin-sensitive G-protein (G). The signaling pathway involves phospholipase C (PLC), leading to the synthesis of diacylglycerol (DG) and inositol 1,4,5, triphosphate (IP3). Although IP3 could release Ca^{2+} from intracellular stores, as revealed by confocal Ca^{2+} imaging (2), DG activates protein kinase C (PKC), which in turn inhibits Ca^{2+} channels, as illustrated by the direct recording of Ca^{2+} currents (3). This pathway is also responsible for the inhibition of the Ca^{2+}-dependent K^+ current, accounting for the hyperpolarization of the membrane after repetitive firing (afterhyperpolarization). These effects of kainate do not seem to require ion channel receptor activity, providing evidence for the existence of a metabotropic KAR that inhibits Ca^{2+} and K^+ channels. Adapted from Rozas JL, Paternain AV, and Lerma J (2003) Noncanonical signaling by ionotropic kainate receptors. *Neuron* 39: 543–553.

Dual Signaling of Kainate Receptors

It was recently demonstrated that KARs and other ionotropic glutamate receptors may transmit signals both through ion flux and by receptor coupling to G-proteins, which in turn may regulate voltage-dependent Ca^{2+} channels (Figure 3). This was initially described in the hippocampus, and it was later confirmed in other tissues. Indeed, activation of KARs in cultured DRG cells, which almost exclusively express GluR5 and KA2 KAR subunits, induces the release of Ca^{2+} from intracellular stores in a G-protein-dependent manner (Figure 3). Interestingly, the concomitant activation of PKC inhibited voltage-dependent N-type calcium channels. This noncanonical signaling is independent of ion channel activity and provides insights into the dual signaling of KARs because both pathways depend on a common ionotropic subunit, GluR5. But the way in which an ion channel becomes coupled to a G-protein is still unclear, although it seems likely that intermediate or linker proteins may assist in the process. Unfortunately, none of the proteins that interacts with KARs identified to date seem to carry out this role, and thus, these intermediate elements remain to be identified. In addition, this property does not seem to be restricted to a given KAR subunit. Postsynaptic KARs can also inhibit hyperpolarizing currents through similar signaling mechanisms (Figure 2(f)), and this seems to involve KA2 subunits rather than GluR6 or GluR5 subunits. There is compelling evidence indicating that synaptically released glutamate could activate postosynaptic KARs in CA1 and CA3 pyramidal neurons that do not involve the generation of inward currents. Rather, these receptors reduce the typical afterhyperpolarization current seen after cell firing (I_{AHP}). Inhibition of after hyperpolarization has a great impact on the neuron excitability, such that the response to a given input is drastically altered. Afterhyperpolarization is induced by the activation of Ca^{2+}-dependent K^+ channels, curtailing repetitive firing. The reduction of such a current causes the neuron to generate large burst of spikes, largely increasing the neuronal output. Together with the inhibition of GABA release, this represents an example in which KARs may exert a striking effect on neuronal excitability by a noncanonical signaling. Therefore, temporal integration of neuronal inputs may be modulated by KARs not acting as ion channels, just by attenuating hyperpolarizing conductances such as I_{AHP} and GABA receptor-mediated current. Although exogenous kainate could facilitate or inhibit transmitter release, depending on the concentration and type of synapse, cumulative data indicate that the rule is that the inhibitory activity of KARs is linked to this noncanonical signaling. However, the variety of signals activated by these metabotropic KARs indicates a certain lack of predictability in the coupling of KARs to G-proteins. Hence, much work remains to be done to clarify when KARs may work through classical or noncanonical signaling pathways. Although not without their problems, the availability of specific drugs as well as of knockout mice for each of the KAR subunits allows us to clarify which subunit participates in specific KAR-mediated responses.

Involvement of Kainate Receptors in Neuropathology

The implications of KARs for nervous system pathology is being underscored but not without uncertainties. For instance, a contribution of KARs to pain is supported by pharmacological data because the GluR5-selective antagonist LY293558 has demonstrated preclinical and clinical efficacy in models of pain. Indeed, GluR5 subunits are essential for KAR-mediated responses in DRG neurons and for presynaptic regulation of transmitter release in the spinal dorsal horn. However, nociceptive thresholds are largely unaffected in GluR5- and GluR6-deficient mice. Only one report found decreased response to capsaicin and formalin in GluR5-deficient mice. Such a profile raises the possibility that KAR antagonists could be effective for the treatment of certain forms of pain and acute migraine. Similarly, excitotoxicity is triggered by the enhanced activation of GluRs. Although kainate was demonstrated to be a neurotoxin long ago, there is no compelling evidence for a direct role of KARs in this process. The most sensitive area in the brain to kainate-induced excitotoxicity is the hippocampal CA3, a region rich in high-affinity KARs. However, in GluR6-knockout mice, kainate-induced CA3 toxicity is attenuated but not eliminated. Recently, KARs activation has been linked to the activation of c-jun N-terminal kinase 3 (JNK) and mixed lineage kinase 3 (MLK3) that seem to link GluR6-containing KARs to the cell death pathway activated after ischemic insults.

The best-established brain disorder in which KARs are implicated is the excitatory imbalance linked to epilepsy. Certainly, kainate injections have provided an animal model for studying human temporal lobe epilepsy. The inhibition of GABA release leading to recurrent epileptic activity may account for the acute epileptogenic effect of kainate. Recently, it has been observed that the synaptic response mediated by KARs provides a substantial component of the excitatory transmission at functional aberrant synapses formed by sprouting of glutamatergic fibers, a hallmark of human temporal lobe epilepsy. Thus, these

results raise the possibility of designing antiepileptic therapies based on KAR antagonists.

A number of studies have examined the link between the expression of KARs genes and/or single-nucleotide polymorphisms (SNPs) to disorders with a genetic background with disparate results. No strong associations have been found between the expression of genes coding for KAR subunits (GluR6 and GluR7 (*GRIK2* and *GRIK3* in human nomenclature)) and obsessive–compulsive disorder patients. Only *GRIK2* SNP I867 was transmitted less readily than expected. Interestingly, this variant has also been associated with autism. An allele of the GluR6 gene is associated with the early age of onset of Huntington's disease; GluR6-mediated excitotoxicity has been implicated in the pathogenesis of Huntington's disease. Allelic variants of *GRIK1* (GluR5), but not *GRIK2* (GluR6), appear to contribute a major genetic determinant to the pathogenesis of juvenile absence epilepsy and related phenotypes. Interestingly, the GluR5 gene is located at chromosame 21q22.1, and physical mapping situates it near the familial amyotrophic lateral sclerosis APP and SOD1 regions, making it a possible candidate for familial amyotrophic lateral sclerosis and other diseases. Whether GluR5 influences the phenotypes associated with partial trisomy or monosomy of chromosome 21 remains to be determined. Other recent results have indicated that there is a disequilibrium of *GRIK2* (GluR6) transmission in autism and schizophrenia. However, no linkage has been established with GluR5 (*GRIK1*) SNPs and their haplotypes in schizophrenia. Whether the described SNPs may impact the function and/or trafficking of KARs is yet to be determined. Therefore, a definitive role of KARs in these disorders is possible but not demonstrated.

See also: AMPA Receptors: Molecular Biology and Pharmacology; AMPA Receptor Cell Biology/Trafficking; Glutamate Receptor Organization: Ultrastructural Insights; Kainate Receptors: Molecular and Cell Biology; NMDA Receptors and Development; NMDA Receptors, Cell Biology and Trafficking.

Further Reading

Castillo PE, Malenka RC, and Nicoll RA (1997) Kainate receptors mediate a slow postsynaptic current in hippocampal CA3 neurons. *Nature* 388: 182–186.

Christensen JK, Paternain AV, Selak S, Ahring PK, and Lerma J (2004) A mosaic of functional kainate receptors in hippocampal interneurons. *Journal of Neuroscience* 24: 8986–8993.

Contractor A, Swanson G, and Heinemann SF (2001) Kainate receptors are involved in short- and long-term plasticity at mossy fiber synapses in the hippocampus. *Neuron* 29: 209–216.

Cossart R, Epsztein J, Tyzio R, et al. (2002) Quantal release of glutamate generates pure kainate and mixed AMPA/kainate EPSCs in hippocampal neurons. *Neuron* 35: 147–159.

Coussen F, Perrais D, Jaskolski F, et al. (2005) Coassembly of two GluR6 kainate receptor splice variants within a functional protein complex. *Neuron* 47: 555–566.

Epztein J, Represa A, Jorquera I, Ben-Ari Y, and Crepel V (2005) Recurrent mossy fibers establish aberrant kainate receptor-operated synapses on granule cells from epileptic rats. *Journal of Neuroscience* 25: 8229–8239.

Frerking M, Malenka RC, and Nicoll RA (1998) Synaptic activation of kainate receptors on hippocampal interneurons. *Nature Neuroscience* 1: 479–486.

Frerking M and Ohliger-Frerking P (2002) AMPA receptors and kainate receptors encode different features of afferent activity. *Journal of Neuroscience* 22: 7434–7443.

Kidd FL and Isaac JT (1999) Developmental and activity-dependent regulation of kainate receptors at thalamocortical synapses. *Nature* 400: 569–573.

Lerma J (2003) Roles and rules of kainate receptors in synaptic transmission. *Nature Reviews Neuroscience* 4: 481–495.

Lerma J (2006) Kainate receptor physiology. *Current Opinion in Pharmacology* 6: 89–97.

Lerma J, Paternain AV, Rodriguez-Moreno A, and Lopez-Garcia JC (2001) Molecular physiology of kainate receptors. *Physiological Reviews* 81: 971–998.

Melyan Z, Lancaster B, and Wheal HV (2004) Metabotropic regulation of intrinsic excitability by synaptic activation of kainate receptors. *Journal of Neuroscience* 24: 4530–4534.

Min MY, Melyan Z, and Kullmann DM (1999) Synaptically released glutamate reduces gamma-aminobutyric acid (GABA) ergic inhibition in the hippocampus via kainate receptors. *Proceedings of the National Academy of Sciences of the United States of America* 96: 9932–9937.

Mulle C, Sailer A, Perez-Otano I, et al. (1998) Altered synaptic physiology and reduced susceptibility to kainate-induced seizures in GluR6-deficient mice. *Nature* 392: 601–605.

Paternain AV, Morales M, and Lerma J (1995) Selective antagonism of AMPA receptors unmasks kainate receptor-mediated responses in hippocampal neurons. *Neuron* 14: 185–189.

Rodriguez-Moreno A, Herreras O, and Lerma J (1997) Kainate receptors presynaptically downregulate GABAergic inhibition in the rat hippocampus. *Neuron* 19: 893–901.

Rodriguez-Moreno A and Lerma J (1998) Kainate receptor modulation of GABA release involves a metabotropic function. *Neuron* 20: 1211–1218.

Rozas JL, Paternain AV, and Lerma J (2003) Noncanonical signaling by ionotropic kainate receptors. *Neuron* 39: 543–553.

Rubinsztein DC, Leggo J, Chiano M, et al. (1997) Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. *Proceedings of the National Academy of Sciences of the United States of America* 94: 3872–3876.

Schmitz D, Mellor J, Breustedt J, and Nicoll RA (2003) Presynaptic kainate receptors impart an associative property to hippocampal mossy fiber long-term potentiation. *Nature Neuroscience* 6: 1058–1063.

Smolders I, Bortolotto ZA, Clarke VR, et al. (2002) Antagonists of GLU(K5)-containing kainate receptors prevent pilocarpine-induced limbic seizures. *Nature Neuroscience* 5: 796–804.

Relevant Websites

http://www.ebi.ac.uk – European Bioinformatics Institute.

Kainate Receptors: Molecular and Cell Biology

J Lerma, Instituto de Neurociencias de Alicante Consejo Superior de Investigaciones Científicas – Universidad Miguel Hernandez, San Juan de Alicante, Spain

© 2009 Elsevier Ltd. All rights reserved.

It is now well established that there are three types of ionotropic glutamate receptors (GluRs), whose nomenclature reflects their preferred ligands: AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptors, NMDA (*N*-methyl-D-aspartate) receptors, and kainate receptors (KARs). The cloning of many GluRs at the beginning of the 1990s and the discovery of their structural relationships confirmed the legitimacy of this pharmacological subdivision. Likewise, their cloning has enabled us to advance our understanding of the biophysical properties and the physiological role of each receptor subtype in the mammalian brain. Indeed, analyzing just a few of these GluR subunits emphasized that true KARs are ion channels with a stronger preference for this agonist rather than for AMPA and that they display rapidly desensitizing responses. Thus, defining the molecular biology of KAR subunits represents a real breakthrough in the study of these receptors, establishing the foundations for us to better understand their physiology.

The Diversity of Kainate Receptors

It is now accepted that all ionotropic GluRs share a conserved transmembrane topology and stoichiometry. Like the AMPA receptors (AMPARs) and NMDA receptors (NMDARs), KARs are tetrameric combinations of the GluR5, GluR6, GluR7, KA1, and KA2 subunits. Of these, GluR5/6/7 can form functional homomeric or heteromeric receptors, whereas KA1 and KA2 participate in functional receptors only when accompanied by GluR5/6/7 subunits. In the ionotropic GluRs, each monomer carries its own ligand binding site and it contributes with a specific amino acid stretch to the channel lumen. In addition to this hydrophobic sequence, each subunit has three transmembrane segments (M1, M3, and M4) arranged in such a manner that the N-terminal domain of each protein lies extracellularly and the C-terminal region lies within the cell. As in the case of AMPARs, the glutamate binding site of KARs is formed by residues that are distributed throughout the distal N-terminal domain (called S1) and the loop between M3 and M4 (called S2). Indeed, this latter segment contains specific residues that confer the receptor with high or low sensitivity to various agonists (AMPA and kainate), as well as to ions that participate in the channel gating.

According to the rules governing subunit assembly and the different arrangements that give rise to functional receptors, a wide diversity of KARs may be expressed in the brain. Of the five KAR subunits that have been cloned so far, some combinations do not render functional channels and others are not specific to kainate. For example, the GluR5 subunit forms receptors that can be activated not only by glutamate but also by kainate, domoate, and AMPA. When activated by kainate these receptor channels desensitize relatively slowly, whereas when they bind glutamate their desensitization is faster and almost complete. In contrast, GluR6 subunits form homomeric channels with fast-desensitizing kinetics when activated by either agonist. Unlike the GluR5 receptors, GluR6 homomeric channels are not sensitive to AMPA or to its derivative ((RS)-2-amino-3(3-hydroxy-5-*tert*-butylisoxazol-4-yl) propanoic acid (ATPA) (**Figure 1(b)**). Functional homomeric receptors can also be formed by GluR7 subunits, these displaying a very low affinity for glutamate and complete insensitivity to AMPA and domoate. Each of these three subunits also form heteromeric assemblies with KA1 or KA2, and the resulting heteromeric receptors present slightly different biophysical and pharmacological channel properties. Moreover, GluR5, GluR6, and GluR7 may also combine with one another to form functional receptors. The sequence alignment reveals a homology of 75–80% between GluR5, GluR6, and GluR7, whereas KA1 and KA2 present 68% of similarity. Homology, however, drops to 45% when these two subfamilies of KARs are considered. The homology of KAR subunits with AMPARs is also approximately 40%, although they are not able to form heteromeric receptors.

Kainate Receptor Subunits Are Subject to Alternative Splicing

In addition to the inherent variation in KARs, alternative splicing generates a number of different isoforms of KAR subunits (**Figure 1(a)**). Indeed, two different GluR5 variants have been found, one that contains an extra segment of 15 amino acids in the N-terminal region (GluR5-1) and another that lacks this insert (GluR5-2). Furthermore, the GluR5-2 subunit presents four different C-terminal domains. The C-terminus of the 2a variant is 49 amino acids shorter than the variant originally described (which was then renamed 2b) due to the introduction of a premature stop codon. Variant 2c contains an extra in-frame exon

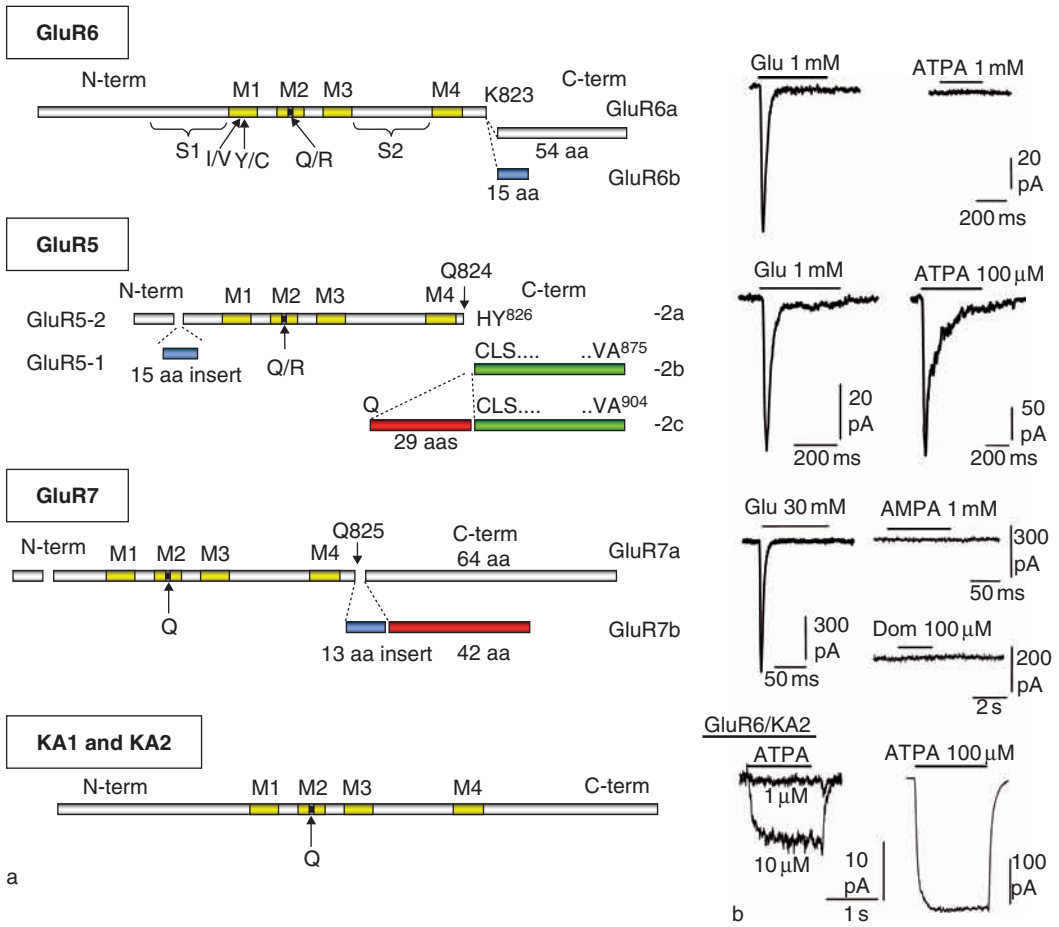

Figure 1 Kainate receptor (KAR) subunits: (a) structure of subunits; (b) electrophysiological recordings of responses to different agonists of the corresponding homomeric receptors recombinantly expressed in HEK293 cells. In (a), the KAR subunits are all made of a chain of approximately 900 amino acids (~100 kDa) that spans the membrane three times (hydrophobic segments M1, M3, and M4). Thus, the N-terminal domain of the protein is extracellular and the C-terminal domain lies intracellularly. The M2 hydrophobic sequence does not cross the membrane, but it is thought to dip into the membrane forming a loop and forming the pore of the channel. Two isoforms of GluR6 subunits exist, -a and -b, differing in their C-terminal domain. The GluR5 can be found as two different variants: GluR5-1, which contains 15 extra amino acids in the N-terminal region, and GluR5-2, which lacks this insert. GluR5-2 presents three additional possibilities generated by a splice of the C-terminal domain. GluR5-2a is 49 amino acids shorter than the GluR5-2b because of the introduction of a premature stop codon. GluR5-2c contains an extra in-frame exon that makes it 29 amino acids longer than the 2b isoform. GluR7 can be present as two isoforms with different C-terminals due to a 13-amino-acid insertion (blue segment), which produces a new open reading frame and a completely different C-terminus (red segment). In contrast, no splice variants have been found for the KA1 and KA2 subunits. GluR5 and GluR6 subunits undergo mRNA editing at the M2 (Q/R site), generating a mixture of edited and nonedited subunits. GluR6 presents two additional editing sites in M1. In every case, a single receptor channel is believed to be composed of four subunits. In all KARs, the binding domain is formed by two peptide segments, S1 and S2, which are situated before the M1 (S1) and linking M3 and M4 (S2). In (b), responses of GluR6 are also shown when this subunit heteromerizes with KA2 subunits (bottom). GluR, glutamate receptor. ATPA, (RS)-2-amino-3(3-hydroxy-5-*tert*-butylisoxazol-4-yl) propanoic acid; Dom, domoic acid.

that makes it 29 amino acids longer than the 2b isoform. Finally, the last 15 amino acids of variant 2d (not shown), from Q824 onward, are completely unrelated to any other variant, and significantly this last isoform was isolated from human tissue. Two splice variants of the murine GluR6 have also been found and, again, these differ in their C-terminal domains. Furthermore, in the rats and humans two splice variants of the GluR7subunit have been discovered. In this case, the insertion of 40 nucleotides out-of-frame in the C-terminal sequence of GluR7a leads to the production of a shorter protein, GluR7b, which bears no significant homology in this region to any of the known ionotropic GluRs. In contrast, the KA1 and KA2 subunits do not seem to undergo alternative splicing. Despite the existence of this repertoire of isoforms, the role of the different KAR splice variants (e.g., in receptor trafficking) is just starting to be determined.

Some Kainate Receptor Subunits Undergo mRNA Editing

The structural repertoire of KAR subtypes is even larger because the pre-mRNA of the KAR subunits GluR5 and GluR6 can undergo editing at the so-called Q/R site of the second membrane domain (M2). As in AMPARs, the Q to R substitution in homomeric KARs assembled either from GluR5 or GluR6 decreases the permeability of the receptor to calcium. Furthermore, this substitution transforms the rectification properties of these receptors from inwardly rectifying to linear or slightly outwardly rectifying, and it reduces the unitary conductance of the channels. As in AMPARs, inward rectification arises from the blockade of the ion channel by internal polyamines, such that the presence of an arginine (R) residue not only drastically reduces calcium permeability but also abolishes ion channel block by polyamines. In addition, GluR6 alone presents two further sites in the first transmembrane domain (M1) that are susceptible to editing, where I567 and Y571 could be changed to V and C, respectively. Although the role of these M1 editing sites remains unclear, editing within the M2 domain of GluR6 KARs has recently been shown to exert a significant effect on synaptic physiology.

Kainate Receptors Are Widely Distributed throughout the Brain

Most of the studies defining the distribution of KARs in the brain come from *in situ* hybridization analysis of mRNA. KAR subunits are widely expressed throughout the nervous system, although there are a few known cell types that exclusively express a given subunit. For instance, GluR5 and KA2 are the only subunits that have been detected in the dorsal root ganglion (DRG) neurons. GluR5 transcripts are also abundant in the Purkinje cells of the cerebellum as well as in the subiculum, the piriform and cingulate cortices, and the hippocampal and cortical interneurons. On the other hand, GluR6 is the most abundant subunit in cerebellar granule cells, the striatum, the dentate gyrus, and the CA1 and CA3 regions of the hippocampus. The mRNA encoding GluR7 is expressed at low levels throughout the brain, but it is particularly prominent in the deep layers of the cerebral cortex, in the striatum, and in the inhibitory neurons of the molecular layer of the cerebellum. The KA1 subunit is present at low levels in the dentate gyrus, in the amygdala, and in the entorhinal cortex, being more abundant in the CA3 region. In contrast, KA2 mRNA can be found in virtually every part of the nervous system. Although the different KAR subunits are already present in the embryo, most of these expression patterns emerge during the early postnatal period. Thus, the amount of GluR5 mRNA peaks between P0 and P5, and it then begins to decline toward the levels found in adults by P12. Similarly, the patterns of GluR6 and KA1 expression observed within the hippocampal formation begins to change at around P0 (GluR6) and P12 (KA1).

We are still awaiting the development of specific antibodies against various KAR subunits. Although *in situ* hybridization is informative, it is not capable of revealing the subcellular distribution of a given subunit. Relatively specific antisera against the KAR subunits GluR6/7 and KA2 have been used to detect KARs, and to date, the most convincing studies have been carried out in the rat retina. In this tissue, the localization of these subunits at the synapse has been investigated by electron microscopy and the selective synaptic distribution of KARs in both plexiform layers of the retina has been observed. Whereas GluR6/7 could be found in horizontal cell processes postsynaptic to both rod spherules and cone pedicles, the KA2 subunit was found only in postsynaptic densities of the cone pedicles in the dendrites of off-bipolar cells.

Anti-GluR6/7 and less specific anti-GluR5/6/7 antibodies have also been used to study other neurons in the brain. Although the limitation of these antibodies, particularly the latter, must be taken into account when interpreting the data obtained, they appeared to map KARs to fibers and synaptic terminals, as well as to dendrites and postsynaptic membranes. Hence, it seems that KARs may be found in both pre- and postsynaptic locations. In particular, GluR5/6 and KA2 have been found on parallel fibers and GluR5/6 immunoreactivity has been identified in a large population of terminals that form axospinal and axodendritic asymmetric synapses in the monkey striatum.

Regulation of the Surface Expression of Kainate Receptors

Although more recent experiments have been directed toward understanding how KARs are targeted to specific membrane domains, the mechanisms controlling surface expression remain unclear. Retention in the endoplasmic reticulum (ER) plays a rate-limiting role in KAR trafficking. As in most GluRs, the residues important for ER retention or exit are proximal to Psd-95, Dlg, and ZO1 (PDZ) binding motifs present in the C-terminal domain. A number of proteins containing PDZ motifs have been shown to interact with KAR subunits, although most of these are promiscuous, also binding to AMPARs and, possibly,

NMDARs. These include postsynaptic density protein (PSD-)95/ synapse-associated protein (SAP-)90, CASK, glutamate receptor interacting protein (GRIP), protein interacting with C kinase 1 (PICK1), and syntenin. However, PDZ proteins such as PICK1, GRIP, and PSD-95 do not appear to play a significant role in the exit of GluR5 or GluR6 from the ER but, rather, they have been implicated in the insertion and/or maintenance of AMPARs and KARs at the synapse. Indeed, it was recently shown that the number of KARs in the membrane surface can rapidly be altered through protein–protein interactions.

These PDZ proteins appear to regulate kainate and AMPARs in a distinct manner. Disrupting the interaction with GRIP decreases synaptic responses through KARs while concomitantly increasing the AMPAR-mediated component. Similarly, interfering with the interaction of PICK1 with KARs and AMPARs depresses the KAR-mediated excitatory postsynaptic current (EPSC), leaving the current mediated by AMPARs unaltered. It is likely that the synaptic expression of KARs is tightly regulated through interactions with still unknown scaffolding proteins, as well as through proteins that influence the trafficking and targeting of KARs. For instance, the rules governing the polarized targeting of KARs to axons and dendrites remain unknown. The study of mice that fail to express different subunits has not demonstrated a link between subunit composition and the compartment to which KARs are targeted. Nevertheless, a variety of trafficking determinants in kainate receptors may promote either membrane expression or intracellular sequestration. The identification of the proteins involved (chaperones) should shed light on the mechanisms that regulate KAR signaling and function. Indeed, a recent study has identified the co-atomer protein complex I (COPI) vesicle coat as a critical mechanism for the retention of KA2 subunits in the ER, such that assembly of KA2 subunits with other KAR subunits to form heteromeric receptors reduces their association with COPI.

The Atomic Structure of Kainate Receptor Binding Domain Has Been Resolved

One of the most exciting advances in the structure–function relationships of KARs has been the achievement of high-resolution crystal structures for the ligand-binding core of some subunits. This has provided a detailed understanding of agonist recognition not only for KARs but also for other ligand-gated ion channels. Indeed, the crystal structure of the GluR5 and GluR6 KAR ligand binding cores complexed with glutamate and other agonists, such as kainate and

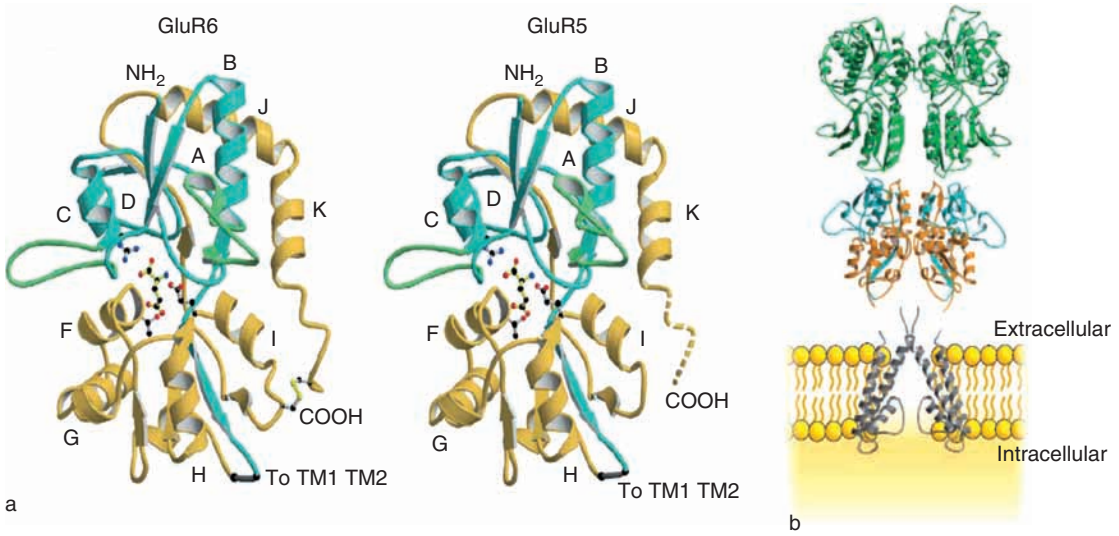

Figure 2 Atomic structure of glutamate-bound GluR6 and GluR5 binding domains: (a) ribbon representation of GluR6 and GluR5 crystal structures at resolution 1.65 and 2.1 Å, respectively; (b) suggested molecular architecture of glutamate ionotropic receptors. In (a), S1, S2, and loops 1 and 2 are colored blue, gold, and green, respectively. The linker shown in gray replaces the TM segments 1 and 2. The C-terminus of GluR5 appears partially disordered (discontinued line) in the crystal. The bound glutamate molecule is illustrated with ball and stick representations. The three ligand binding pocket side chains shown are conserved in all kainate receptors. Letters A through K name the different α-helical structures loops in the molecule. In (b), the architecture is composed of the crystal structures of the mGluR1 ligand-binding domain dimer (green), the GluR5 ligand-binding core dimer (S1, cyan; S2, orange), and two subunits from the KcsA potassium channel tetramer (gray). An intact glutamate receptor is believed to assemble as a dimer of dimers. GluR, glutamate receptor; mGluR, metabolic glutamate receptor. (a) Adapted from Mayer ML (2005) Crystal structures of the GluR5 and GluR6 ligand binding cores: Molecular mechanisms underlying kainate receptor selectivity. *Neuron* 45: 539–552. (b) Adapted from Mayer ML (2006) Glutamate receptors at atomic resolution. *Nature* 440: 456–462.

quisqualate, have been shown at a resolution of 1.65 (GluR6) and 2.1 Å (GluR5) (Figure 2(a)). As in the AMPARs and NMDARs, the structure of the KAR binding domain revealed a two-domain closed shell motif linked by β strands. Glutamate binds in a cavity that is isolated from the external solution. For example, agonist binding induces shell closure; this evokes a conformational rearrangement, leading to the opening of the channel. Interestingly, the cavity is smaller in GluR6 than in GluR5 but larger than that of the AMPAR subunit GluR2, due to the presence of unique side chains in KAR subunits. Nevertheless, the mechanism by which glutamate binds to GluR5, GluR6, and GluR2 is nearly identical. Like GluR2, the GluR6 domain associates as a dimer with many conserved interdimer contacts. Thus, an examination of the atomic structure explains why some residues are critical for GluR5-selective ligands (e.g., ATPA). Binding of these agonists to GluR6 is prevented by steric occlusion because the cavity in GluR6 is too small to accommodate the agonists that, like ATPA, bind to GluR5. In Figure 2(b), a recently suggested molecular structure for ionotropic glutamate receptors, including KARs, is presented. It seems that ionotropic GluRs are formed by the fusion of discrete segments, whose ancestors could be found as bacterial proteins. The structure has been built by taking the solved crystal structures of both the N-terminal domain of metabotropic GluRs and the agonist binding domain of KARs. So far, the structure of the GluR ion channel is unknown at the atomic resolution. However, based on the fact that the ion channel of ionotropic GluRs presents remarkable sequence homology with bacterial potassium channels, the crystal structures of a bacterial potassium channel, the KcsA potassium channel, has been incorporated as the channel domain. Although it is likely that a KAR presents a similar arrangement, much work remains to finally solve the structure of the GluR channel. It is expected that this rapidly developing area will soon provide the real structure at the atomic resolution.

See also: AMPA Receptors: Molecular Biology and Pharmacology; AMPA Receptor Cell Biology/Trafficking; Glutamate Receptor Organization: Ultrastructural Insights; Kainate Receptor Functions; NMDA Receptors and Development; NMDA Receptors, Cell Biology and Trafficking.

Further Reading

Bettler B, Boulter J, Hermans-Borgmeyer I, et al. (1990) Cloning of a novel glutamate receptor subunit, GluR5: Expression in the nervous system during development. *Neuron* 5: 583–595.

Jaskolski F, Coussen F, and Mulle C (2005) Subcellular localization and trafficking of kainate receptors. *Trends in Pharmacological Science* 26: 20–26.

Kohler M, Burnashev N, Sakmann B, and Seeburg PH (1993) Determinants of Ca^{2+} permeability in both TM1 and TM2 of high affinity kainate receptor channels: Diversity by RNA editing. *Neuron* 10: 491–500.

Lerma J, Paternain AV, Naranjo JR, and Mellström B (1993) Functional kainate selective glutamate receptors in cultured hippocampal neurons. *Proceedings of the National Academy of Sciences of the United States of America* 90: 11688–11692.

Lerma J, Paternain AV, Rodriguez-Moreno A, and Lopez-Garcia JC (2001) Molecular physiology of kainate receptors. *Physiological Reviews* 81: 971–998.

Madden DR (2002) The structure and function of glutamate receptor ion channels. *Nature Reviews Neuroscience* 3: 91–101.

Mayer ML (2005) Crystal structures of the GluR5 and GluR6 ligand binding cores: Molecular mechanisms underlying kainate receptor selectivity. *Neuron* 45: 539–552.

Mayer ML (2006) Glutamate receptors at atomic resolution. *Nature* 440: 456–462.

Paternain AV, Herrera MT, Nieto MA, and Lerma J (2000) GluR5 and GluR6 kainate receptor subunits coexist in hippocampal neurons and coassemble to form functional receptors. *Journal of Neuroscience* 20: 196–205.

Paternain AV, Morales M, and Lerma J (1995) Selective antagonism of AMPA receptors unmasks kainate receptor-mediated responses in hippocampal neurons. *Neuron* 14: 185–189.

Petralia RS, Wang YX, and Wenthold RJ (1994) Histological and ultrastructural localization of the kainate receptor subunits, KA2 and GluR6/7, in the rat nervous system using selective antipeptide antibodies. *Journal of Comparative Neurology* 349: 85–110.

Schiffer HH, Swanson GT, and Heinemann SF (1997) Rat GluR7 and a carboxy terminal splice variant, GluR7b, are functional kainate receptor subunits with a low sensitivity to glutamate. *Neuron* 19: 1141–1146.

Seeburg PH (1996) The role of RNA editing in controlling glutamate receptor channel properties. *Journal of Neurochemistry* 66: 1–5.

Swanson GT, Feldmeyer D, Kaneda M, and Cull-Candy SG (1996) Effect of RNA editing and subunit co-assembly single-channel properties of recombinant kainate receptors. *Journal of Physiology* 492: 129–142.

Wisden W and Seeburg PH (1993) A complex mosaic of high-affinity kainate receptors in rat brain. *Journal of Neuroscience* 13: 3582–3598.

Relevant Website

http://www.ebi.ac.uk – European Bioinformatics Institute.

Kinematics and Dynamics

A d'Avella, Santa Lucia Foundation, Rome, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

To behave adaptively under novel and changing circumstances, many animals extend and adapt their movement repertoire. Humans and other primates in particular have an extraordinary capacity for learning skilled arm and hand movements. Reaching and grasping are crucial components of the movement repertoire of all primates and the foundations of many sophisticated human skills. While these basic skills are learned early in life, changes in the motor apparatus due to growth or injury require continuously monitoring and recalibrating the neuronal control system for maintaining adequate performance. Moreover, using tools and performing sports often requires learning new reaching and grasping skills. Consider, for example, reaching with a cursor on a computer screen using a 'mouse,' grasping a small object with a forceps under a microscope, or hitting a ball with a racket. In this article, recent advances in the study of the characteristics and the neural basis of motor learning are reviewed. The focus is on learning and adaptation of reaching skills as a specific but very significant case of motor learning.

To control reaching, as for many other goal-directed movements, the central nervous system (CNS) has to map sensory input into motor output. Vision and audition provide information about the movement goal, that is, the spatial location to be reached. Vision and proprioception are used to determine the location of the hand and the arm joints configuration. The activations of shoulder and arm muscles constitute the motor output. This sensorimotor mapping involves both kinematic and dynamic transformations. Kinematics is the description of the motion of the limb joints or the motion of the limb endpoint, while dynamics is concerned with the forces acting on the limb segments and the segment masses that produce or affect those motions. The overall mapping can be decomposed into different transformations. First, proprioceptive information has to be aligned with visual information in a common reference frame. This alignment process is likely accomplished through transformation of proprioceptive information about the limb configuration into a vision-based coordinate frame and requires a direct kinematic transformation: the mapping of joint angles and muscle lengths into end-effector spatial location. Second, a motor error, formulated as a difference vector between the location of the target and the location of the hand, must be transformed into a motor plan, possibly expressed as the joint displacements necessary to zero the error and bring the hand to the target. This is an inverse kinematics transformation: from end-effector to joints. Finally, the motor plan must be mapped into muscle activations, an inverse dynamic transformation.

When learning a new reaching skill, such as moving a cursor on a computer screen by moving an object (a 'mouse') with the hand or reaching with a stick, a new kinematic mapping between the end-effector motor error and the motor plan must be acquired. For example, the motor error may be on a vertical plane (the computer screen) while the hand moves on a horizontal plane (the desk surface). In situations involving perturbations of the alignment of vision and proprioception, as when vision is distorted by prism glasses (see the section titled 'Adapting to prism glasses'), the existing kinematic mappings must be adjusted. Finally, when learning to reach while holding an object with unusual dynamic properties, a new dynamic mapping between motor plan and motor output must be acquired, or when adapting to novel dynamic behavior of the arm, the existing mapping must be modified.

Learning Kinematics

Adapting to Prism Glasses

The adaptation processes induced by a perturbation of the visual input has been investigated experimentally using prism glasses. When people wear glasses with wedge prisms in place of regular lenses, the path of light is deviated by a certain angle, and the visual field is rotated accordingly. Initially, participants are unable to reach a nearby target in front of them with their hand or to hit a more distant target by throwing a clay ball (Figure 1). For example, participants wearing prisms that bend the light path to their right, so that objects appear shifted to the left of their actual location, initially reach to the left of the target. With practice, participants progressively change the movement direction until they reach accurately in the direction of the target. When the prisms are removed, participants miss again the target by reaching to its right, a phenomenon called negative aftereffect. Within a few trials, this aftereffect disappears.

The existence of aftereffects indicates that some changes have occurred in the kinematic transformations used for controlling reaching movements in

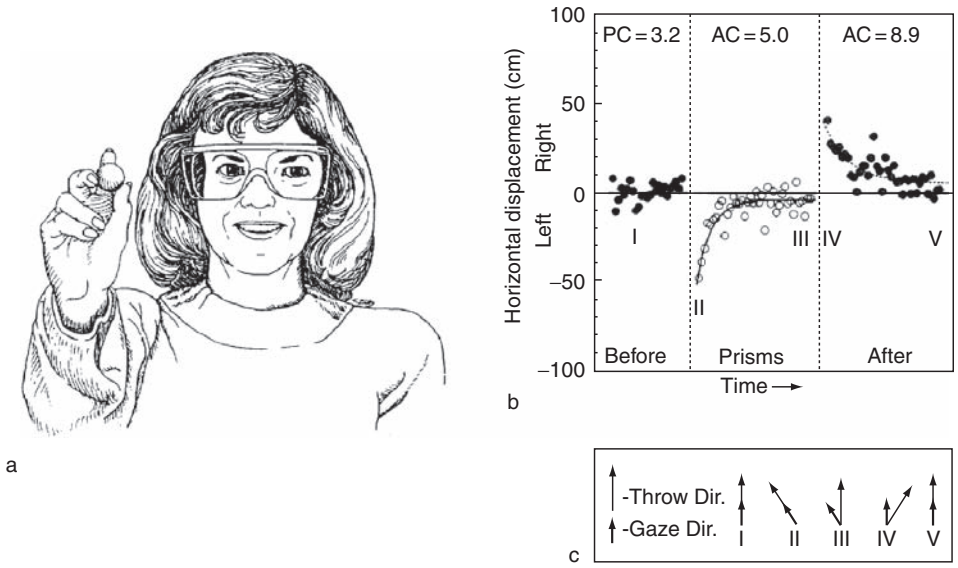

Figure 1 Adaptation to prism glasses. In the adaptation study illustrated here, a participant wearing wedge prisms bending the light path to the left (a) initially throws a ball to the left of the target ((b), II, where the horizontal displacement of the impact points from the target are shown as a function of the trial number). With practice, the participant progressively adapts to the visual perturbation and hits the target with the same accuracy ((b), III) than before wearing the prisms ((b), I). In the first trial after the removal of the prisms, the participant misses the target by throwing to its right ((b), IV), that is, making an error in the opposite direction to the error observed in the first trial with the prisms (aftereffect). Within a few trials without the glasses, the participant hits the target accurately again ((b), V). (c) Gaze and throw directions in the different phases of the experiment schematized with arrows. Adapted from Martin TA, Keating JG, Goodkin HP, Bastian AJ, and Thach WT (1996) Throwing while looking through prisms. II. Specificity and storage of multiple gaze-throw calibrations. *Brain* 119: 1199–1211, with permission from Oxford University Press.

normal conditions. The adaptation is specific to the trained arm: participants who have adapted and reach correctly with one hand while wearing the prisms do not show aftereffects when reaching with the other hand after removing the prisms. Thus, the adaptation cannot be accounted for by a shift of the mapping of the visual input to the reference frame where a motor error is encoded. Instead, the specificity of adaptation is compatible with a shift in the kinematic transformations that map proprioceptive input into a visual reference frame. This proprioceptive shift can be interpreted as a shift in the 'felt' hand location in the visual field. However, while this proprioceptive shift allows the CNS to compute the correct motor error even when the hand is not in sight, accurate reaching with the prisms requires that the kinematic mapping between motor error and motor plan be adjusted as well. In fact, the original mapping of motor errors in visual coordinates into joint displacements would generate a correct motor plan only if the physical world were shifted along the visual field. Thus, the kinematic mapping between motor error and motor plan is adjusted during adaptation to compensate for the shift of the visual field.

Learning Visuomotor Transformations

In the experiments with prism glasses, the CNS interprets a mismatch between vision and proprioception and between motor error and motor plan as an error in the implementation of the normal sensorimotor transformations used for reaching and recalibrates those transformations. However, in other conditions the CNS may interpret a visuomotor mismatch as due to the kinematic characteristic of an object or tool being manipulated. In this case, a new visuomotor transformation is learned instead of adapting the existing transformation used for normal reaching. For example, learning to reach with a computer mouse does not affect normal reaching movements, and it is possible to switch from one transformation to the other instantaneously. Visuomotor recalibration, as in the case of prism adaptation, and visuomotor skill learning, as in the case of tool-specific skills, generalize differently to untrained conditions. Generalization has been studied in experiments using novel visuomotor transformation. In these experiments, visual feedback provided by either a computer screen or a virtual reality display allowed simulating either a novel tool whose end-effector movements did not coincide with the hand movements, as in the case of a computer mouse, or a precisely controlled prismlike perceptual distortion of the movement of a person's own hand. In both cases, by training study participants in a single reaching condition, such as reaching in one direction, and testing them on multiple conditions, it was possible

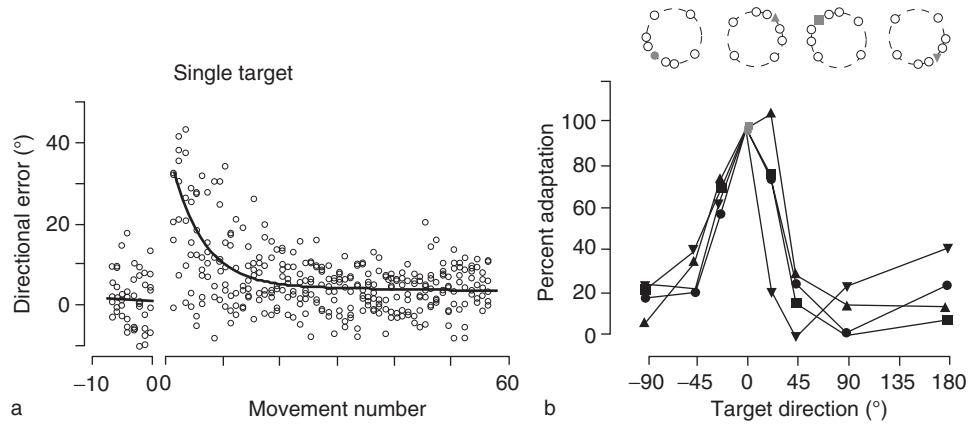

Figure 2 Visuomotor rotation learning. In this experiment, participants sat facing a computer monitor at eye level and controlled a screen cursor by moving a hand-held indicator in a horizontal digitizing tablet. Participants were required to move the cursor from one central location to different peripheral targets. After practicing without any visual perturbation, participants were trained to perform the task while the direction of cursor movement relative to the hand movement on the tablet was rotated counterclockwise by 30°. Training with a single target (a), the directional error was quickly reduced with practice. When generalization to other targets was tested, the amount of adaptation resulted very small except for directions very close to the trained direction ((b), where the curves representing the percent adaptation as a function of the angular difference between trained and tested targets, bottom, are shown with different symbols for different trained targets, top). Adapted from Krakauer JW, Pine ZM, Ghilardi MF, and Ghez C (2000) Learning of visuomotor transformations for vectorial planning of reaching trajectories. *Journal of Neuroscience* 20(23): 8916–8924, Copyright (2000) by the Society for Neuroscience.

to study how the learned visuomotor transformation generalizes to untrained conditions.

When the transformation between the hand movement on a horizontal plane, recorded by a mouse or a pen on a digitizing tablet, and the movement of a cursor on a vertically mounted computer display is altered by a rotation or a shift, with practice people learn the novel visuomotor transformation (Figure 2(a)). Learning this skill does not affect the accuracy of normal reaching movements, and if only one movement condition is trained, the generalization to other untrained conditions is modest (Figure 2(b)). In contrast, when a novel visuomotor transformation is introduced in a virtual reality display, where participants initially see a virtual representation of the position in space of their fingertip matching its actual location (Figure 3(a)), learning to reach in one movement condition generalizes to other conditions (Figures 3(b) and 3(c)). Moreover, immediately after the removal of the visual perturbation, normal reaching is affected and aftereffects are observed. Thus, the perception of the coincidence of the source of visual feedback and the end effector may be a crucial element in determining whether motor learning occurs through recalibration of an existing visuomotor mapping or by acquisition of a new mapping.

Learning Dynamics

When adapting to altered visual feedback or learning new visuomotor skills, the CNS needs to deal only with kinematic transformations. However, in many circumstances the forces necessary to generate a desired movement may change, and thus the CNS must recalibrate or acquire a new dynamic transformation that maps motor plans to muscle activations.

Adapting to Novel Inertial Forces

The dynamics of the arm change during trunk rotations. Moving the arm with respect to a rotating reference frame generates velocity-dependent inertial forces (Coriolis forces) that alter the hand trajectory generated by a given muscle activation pattern. This effect is fully compensated for when normal reaching movements are combined with active trunk rotations so that it is possible to reach accurately during whole-body turning. However, exposure to artificially generated Coriolis forces, during passive trunk rotations, causes reaching errors. The adaptation process required to compensate for these unusual inertial forces has been investigated in experiments in which participants performed reaching movements in a rotating room. If the heads of the participants are close to the rotation axis and the angular velocity is constant, participants do not perceive the effect rotation if they do not move. When a participant moves, Coriolis forces act on the whole arm without any external mechanical contact (Figure 4(a)). Initially, reaching toward a remembered location, without vision of the arm, is inaccurate (Figure 4(b)). The trajectory of the hand deviates from the straight line connecting start to target which would normally

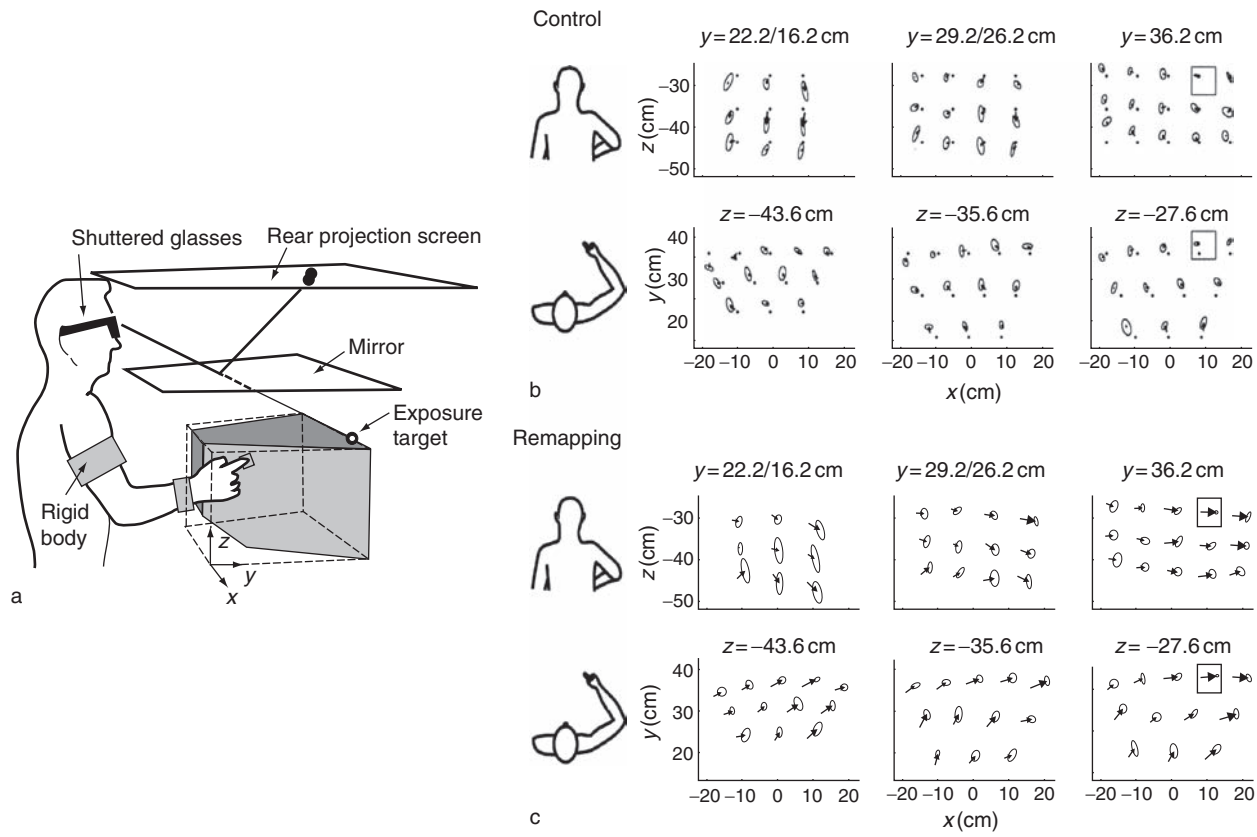

Figure 3 Learning a local visuomotor transformation in a virtual reality environment. A three-dimensional virtual visual feedback setup ((a), constituted by a mirror, a rear projection screen, and shuttered glasses synchronized with the display for stereoscopic perception) was used in this experiment to overlay images onto the arm workspace. Participants could not see their arm but were shown a target location and their finger location as a 1 cm-edge cube. A computer-controlled discrepancy between the finger and the cube position could be introduced. After a familiarization phase, during which participants pointed to 36 different targets with continuous veridical feedback, pointing errors without visual feedback of the finger location were assessed in a preexposure phase. During the following exposure phase, participants pointed to one target only while the finger position was displaced, gradually across trials, up to 6 cm to the right of the average preexposure position, and the visual feedback of finger position was displayed only within 3 cm from the target. As a control condition, visual feedback was altered so that participants had to point to their average preexposure position to see their finger on the target. Changes in pointing due to the exposure of the visuomotor remapping were assessed in a postexposure phase by testing all targets without visual feedback. In the control condition (b), changes in pointing due to the exposure to the remapping when pointing to the target indicated by a square were not significant. In contrast, remapping of the visual feedback at one target ((c), exposure target indicated by a square) induced significant changes over the whole workspace. Adapted from Vetter P, Goodbody SJ, and Wolpert DM (1999) Evidence for an eye-centered spherical representation of the visuomotor map. *Journal of Neurophysiology* 81(2): 935–939, used with permission from the American Physiological Society.

be followed in reaching movements, and the hand does not reach the planned location. Within a few trials, accuracy increases, and the trajectories in the rotating room resemble those in a stationary environment (Figures 4(b) and 4(c)). However, the first reaching movement performed after the room rotation is terminated is inaccurate. The hand trajectory is deviated in a direction opposite to the direction of the deviation induced by Coriolis forces. Such aftereffects disappear after a few trials.

As for the experiments with prism glasses, the existence of aftereffects indicates that the dynamic transformation used for reaching movements in normal conditions is adapted during the exposure to novel velocity-dependent noncontacting forces. Since the adaptation occurs without visual or tactile feedback, proprioceptive input is sufficient to adjust the dynamic transformation. A mismatch between expected and actual signals from the muscle spindles conveys the information about an incorrect execution of the motor plan. The resulting changes in the generation of the muscle activations allow for implementing the normal motor plan, a straight hand trajectory from start to target, in the dynamically altered conditions. This dynamic mapping is specific to the arm exposed to the dynamic perturbation, as shown by

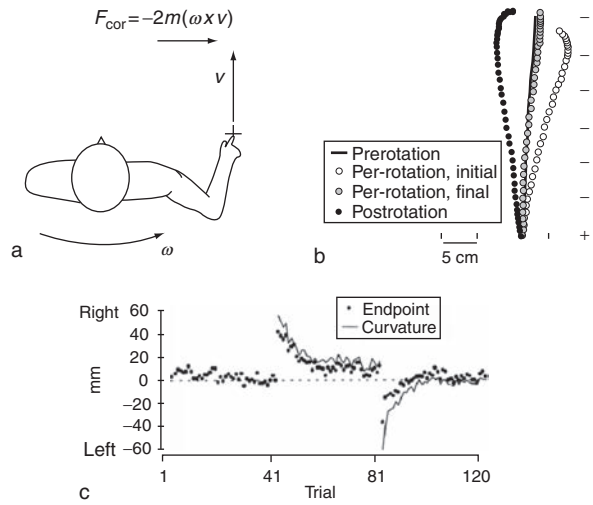

Figure 4 Adapting to novel inertial forces. (a) In this experiment, conducted in a dark room rotating counterclockwise (ω: angular velocity), Coriolis forces (F_{cor}) generated on the reaching arm (v: velocity relative to the torso; m: arm mass) are directed to the right. (b) The trajectory of the first movement to reach a visual target (extinguished at movement initiation) during rotation (10 rpm) was deviated in the direction of the Coriolis force (per-rotation, initial, average over 11 participants) and did not reach the target. However, at the 40th movement (per-rotation, final), the trajectory was not different from the trajectory in the stationary room (prerotation). The first movement after rotation (postrotation) showed mirror symmetric endpoint error and curvature to the first movement during rotation (aftereffect). (c) Summary of endpoint error and curvature before (trials 1–40), during (41–80), and after (81–120) rotation. Adapted from Dizio P and Lackner JR (1995) Motor adaptation to Coriolis force perturbations of reaching movements: Endpoint but not trajectory adaptation transfers to the nonexposed arm. *Journal of Neurophysiology* 74(4): 1787–1792, used with permission from the American Physiological Society.

the fact that the trajectory adaptation does not transfer to the other, unexposed arm.

Learning Novel Viscous Force Fields

The dynamics of the reaching arm is also altered when one is grasping an object or manipulating a tool. For objects and tools whose dynamic characteristics are well known, trajectory and endpoint accuracy of reaching while holding the object does not change with respect to reaching with the empty hand, indicating that the CNS is automatically selecting the dynamic mapping appropriate for each condition. The process of learning such tool-specific mappings has been studied experimentally by using a two-link planar robot manipulator (Figure 5(a)) that imposes novel velocity-dependent (viscous) forces on the arm (Figures 5(b) and 5(c)). In a velocity-dependent force field, as in the case of Coriolis forces, no perturbation is applied at the beginning and at the end of the movement, when the arm is stationary. In a typical experiment, participants make reaching movements while holding the end-effector of the manipulator. As for some of the visuomotor learning experiments described above, participants are first trained to move a cursor on a computer screen by moving the manipulator handle. Trajectories are approximately straight in this baseline condition. Next, a perturbing force is applied to the hand by the manipulator. This force has a magnitude proportional to the hand tangential velocity but a direction which is not always opposite to the movement direction, as is usual with viscous forces. The direction may be always perpendicular to the hand velocity (curl force field) or change as a function of the movement direction. When first exposed to the force field, participants make large errors. Without visual feedback, the trajectory is skewed and deviated from a straight line (Figure 5(d)). With practice, however, the trajectories become straight again, indicating that a dynamic mapping correctly implementing the motor plan has been acquired (Figure 5(e)). Finally, when the perturbing forces are removed, participants initially produce trajectories which are mirror images of the initial trajectories after force field exposure (Figure 5(f)). With practice these aftereffects also disappear.

Comparing the adaptation to viscous force fields delivered by a robot manipulator with the adaptation to the Coriolis forces, there are a number of differences. Full adaptation to the perturbation is accomplished after a few tens of repetitions with noncontacting Coriolis forces, but it requires hundreds of movements with the robot manipulator. Moreover, after the adaptation, normal reaching in free space is impaired in the case of Coriolis forces but not with the robot. Thus, the adaptation to noncontacting Coriolis forces and the adaptation to forces applied through mechanical contact appear as different processes. The former is more directly related to an internal calibration process of the dynamic transformation used for reaching with the free arm, whereas the latter is related to the acquisition of a new mapping useful to skillfully control a novel tool.

Neural Basis of Motor Learning

Neural plasticity is the likely physiological mechanism underlying the recalibration and acquisition of the kinematic and dynamic transformations involved in the control of reaching. Plasticity may be caused by unmasking of existing connections, synaptic changes, or growth of new connections. The areas of the CNS involved in motor learning are beginning to be identified through lesion and imaging studies. These studies

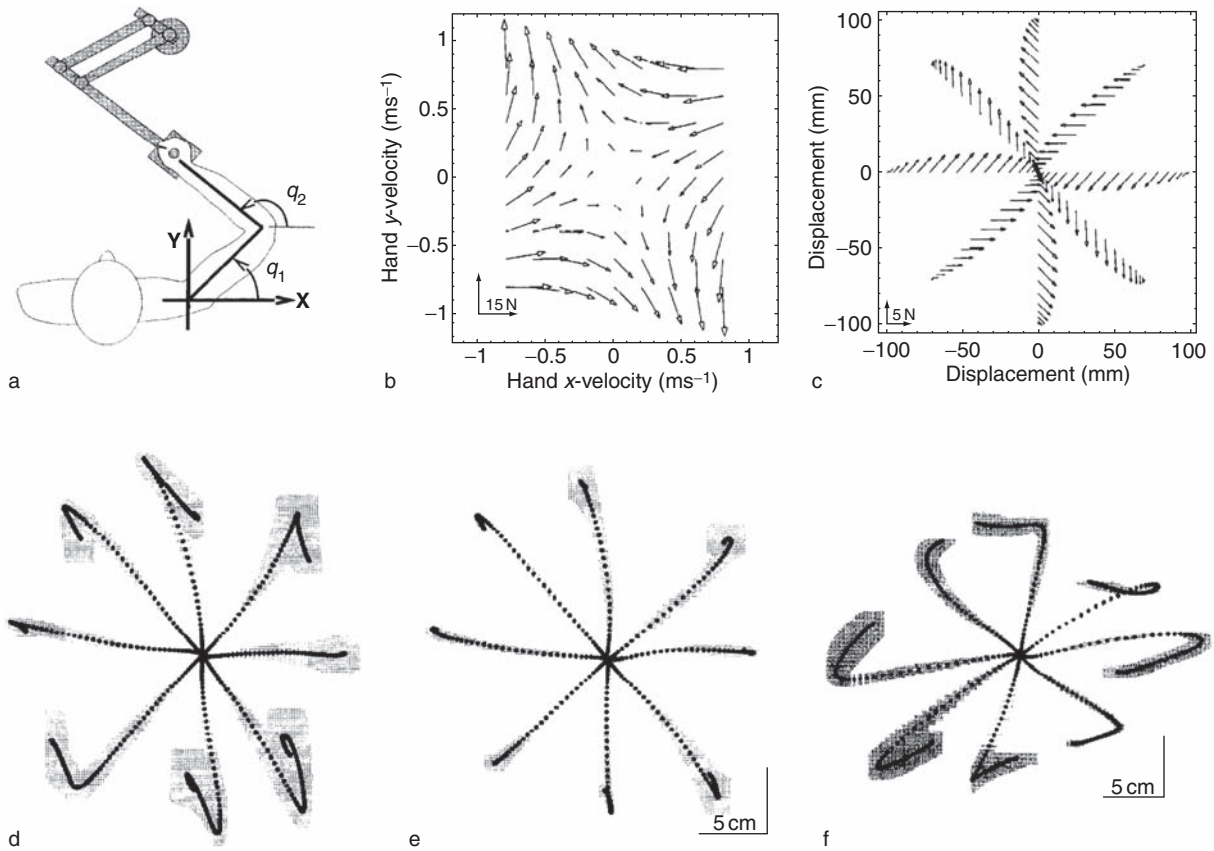

Figure 5 Learning novel velocity-dependent forces. Participants made reaching movements while grasping the handle of a two-link planar robot manipulator with their arm on a horizontal plane (a). The location of the handle on the plane was displayed on a monitor placed in front of the participants. The torque motors of the manipulator could be programmed to apply forces to the handle that depended on the handle velocity ((b), handle velocity and force in *x–y* Cartesian coordinates). Different from normal viscous forces, the applied force was not aligned with the movement direction but, instead, depended on it. This fact is illustrated by the forces acting on the hand during simulated center-out reaching movements from one central location to eight peripheral targets (c). Note that forces at the beginning and at the end of the movement vanish because the velocity is null. After practicing in the null field (torque motors off), participant performed reaching movements in the force field with randomly interspersed trials in the null field (catch trials). In all conditions, a fraction of the movements were performed without visual feedback. At the beginning of the exposure to the force field ((d), average and standard deviation of hand path without visual feedback recoded during the first block of 250 trials in the force field indicated by black dots and gray areas, respectively), the hand path deviated substantially from the null field straight-line trajectory, and there were corrections at the end of the movements. With practice, paths become progressively straighter ((e), average and standard deviation of hand path without visual feedback recoded during the fourth and last block of force field trials). However, during catch trials ((f), average and standard deviation of hand path without visual feedback in the null field recoded during the last block of trials), the hand paths showed aftereffects with deviations opposite to those observed during the initial force field exposure ((d)). Adapted from Shadmehr R and Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience* 14(5 Pt 2): 3208–3224, used with permission from Society for Neuroscience.

suggest that motor learning depends on multiple, interconnected cortical and subcortical systems, each contributing to specific aspects and different phases of the process. Moreover, the changes in the neural representation of movement parameters in some of the areas involved in motor learning have been investigated with single cell recordings in monkeys.

CNS Structures Involved in Motor Learning

Studies of patients with lesions of the CNS and imagining studies of normal participants have identified some of the cortical and subcortical areas involved in learning the kinematic and dynamic transformations involved in goal-directed arm movements. Cerebellar dysfunctions impair adaptation of both kinematic (prism) and dynamic (force field) transformations. Positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) studies have also found an increase in cerebellar activation during early phases of visuomotor rotation and force-field learning. The cerebellar involvement during early phases might be related to its role in processing the large execution errors present before adaptation

is complete. Activation of the posterior parietal cortex has been observed with PET during adaptation to prism glasses and during learning visuomotor hand–cursor rotations. A parietal activation has also been observed with fMRI during early stages of visuomotor learning. This parietal activation might be related to the remapping of the proprioceptive representation of the hand position in a visual reference frame containing the representation of hand and target position. During the learning of a force field perturbation, significant activations have been observed with PET in the dorsolateral prefrontal cortex during early learning phases and in the posterior parietal cortex, premotor cortex, and cerebellum during recall of the learned dynamic transformation.

Neural Correlates of Kinematic and Dynamic Learning in the Motor Cortex

Recordings of the activity of single cells in the premotor and primary motor cortex of macaque monkeys during reaching have shown that these cortical areas encode both kinematic and dynamic parameters. The firing rate of most cells in these areas is modulated by the movement direction, with a directional dependence well captured by a cosine function. Thus, the firing rate of each cell is maximal for some direction of movement, the preferred direction. The population vector, a simple algorithm for decoding the direction of movement from the activity of a population of neurons on the basis of the linear combination of the preferred direction of individual cells, indicates that kinematic parameters are accurately encoded by populations of cortical cells. However, the activity of many cells is also modulated by dynamic parameters, such as the force applied by the hand against a load and the orientation of the arm during reaching. It is likely that different populations of cortical neurons, in parietal and frontal cortical areas, constitute a multilayer recurrent network capable of transforming a population representation of the target and the current limb position, derived from visual and proprioceptive sensory input, into a population representation of the motor output, the appropriate muscle patterns for reaching the target.

When learning a new visuomotor mapping, such as the one required to reach with prism glasses or after a rotation of the joystick–cursor transformation, the neural representation of the mapping must also

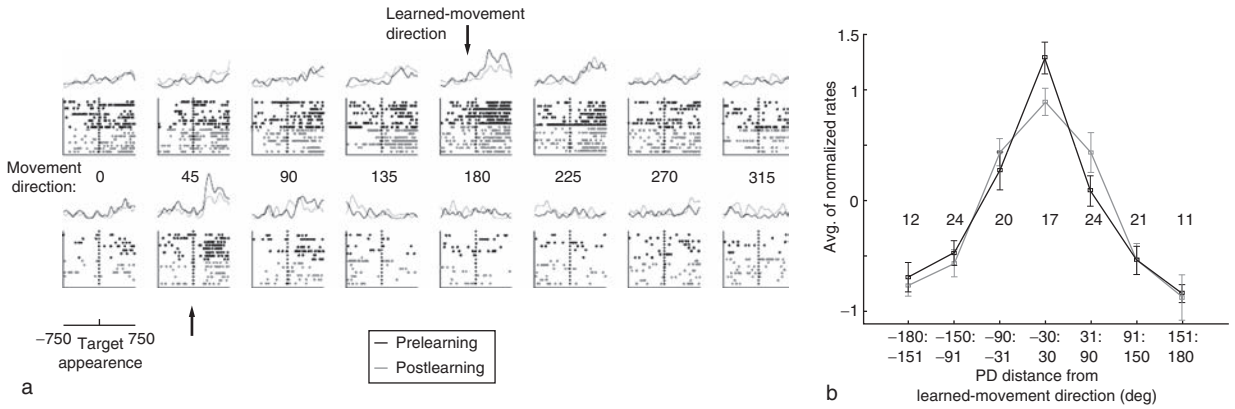

Figure 6 Amplitude modulation of the firing rate on neurons in the motor cortex during visuomotor learning. Macaque monkeys were trained to move a manipulandum on a horizontal plane to move a cursor on a vertical monitor from one central location to eight peripheral targets. In the default visuomotor mapping, a rightward (0°) hand movement (on the horizontal plane) corresponded to a rightward cursor movement (on the vertical plane) and a forward hand movement (90°) to an upward cursor movement. Monkeys were tested with this mapping before and after learning a novel visuomotor transformation. In the learning epoch, monkeys had to reach only the 90° target, but the direction of hand movement required to reach the target was changed on each experimental session (randomly chosen between 0°, 45°, 135°, and 180°, that is, requiring a rotation of −90°,−45°, 45°, or 90° between hand and cursor movement direction with respect to the default mapping). Neural activity in the default eight-target task was compared before and after learning. (a) Raster plots and peristimulus histograms of the activity of two cells for the eight directions of movements before learning (gray ticks in the raster plots, where each row represents a trial and each tick a spike; gray lines in the peristimulus histogram, showing the sum of the spikes across trials in each time bin after alignment with the onset of the target stimulus) and after learning (black ticks and lines). In the first cell (top row of raster plots and histograms), the learned movement direction necessary to reach the 90° target was 180° (indicated by an arrow) whereas in the second cell (bottom row), it was 45°. The firing rate during the preparatory period (after stimulus onset but before movement onset) is enhanced in the learned-movement direction. (b) Changes in the population averages of normalized firing rates during the preparatory activity before (gray) and after (black) learning. The abscissa shows the angular distance between the cells' preferred direction and the learned-movement direction; the number of cells in each bin is indicated. Only cells with a preferred direction close to the learned-movement direction showed a significant change in firing rate due to learning. Reprinted by permission from Macmillan Publishers Ltd. (Paz R, Boraud T, Natan C, Bergman H, and Vaadia E (2003) Preparatory activity in motor cortex reflects learning of local visuomotor skills. *Nature Neuroscience* 6(8): 882–890), copyright (2003).

change. Single unit recordings in the motor cortex during visuomotor learning have revealed a complex pattern of changes in the neuronal activity. A large fraction of the cells show a change in the magnitude of the activity modulation during adaptation, some increasing and others decreasing. When the visuomotor rotation to be learned is local, that is, in only one direction of movement, changes in the activity modulation are observed only in a subpopulation of cells with a preferred direction near the direction of the learned movement (Figure 6). Thus, it appears that the neural representation of the visuomotor transformation is accomplished by local elements, that is, subpopulations with a specific directional tuning, which can be adjusted independently to adapt to novel mappings. Moreover, the changes in neuronal activity in motor cortex during visuomotor learning occur during movement preparation and not during movement execution. This observation suggests that these changes involve the representation

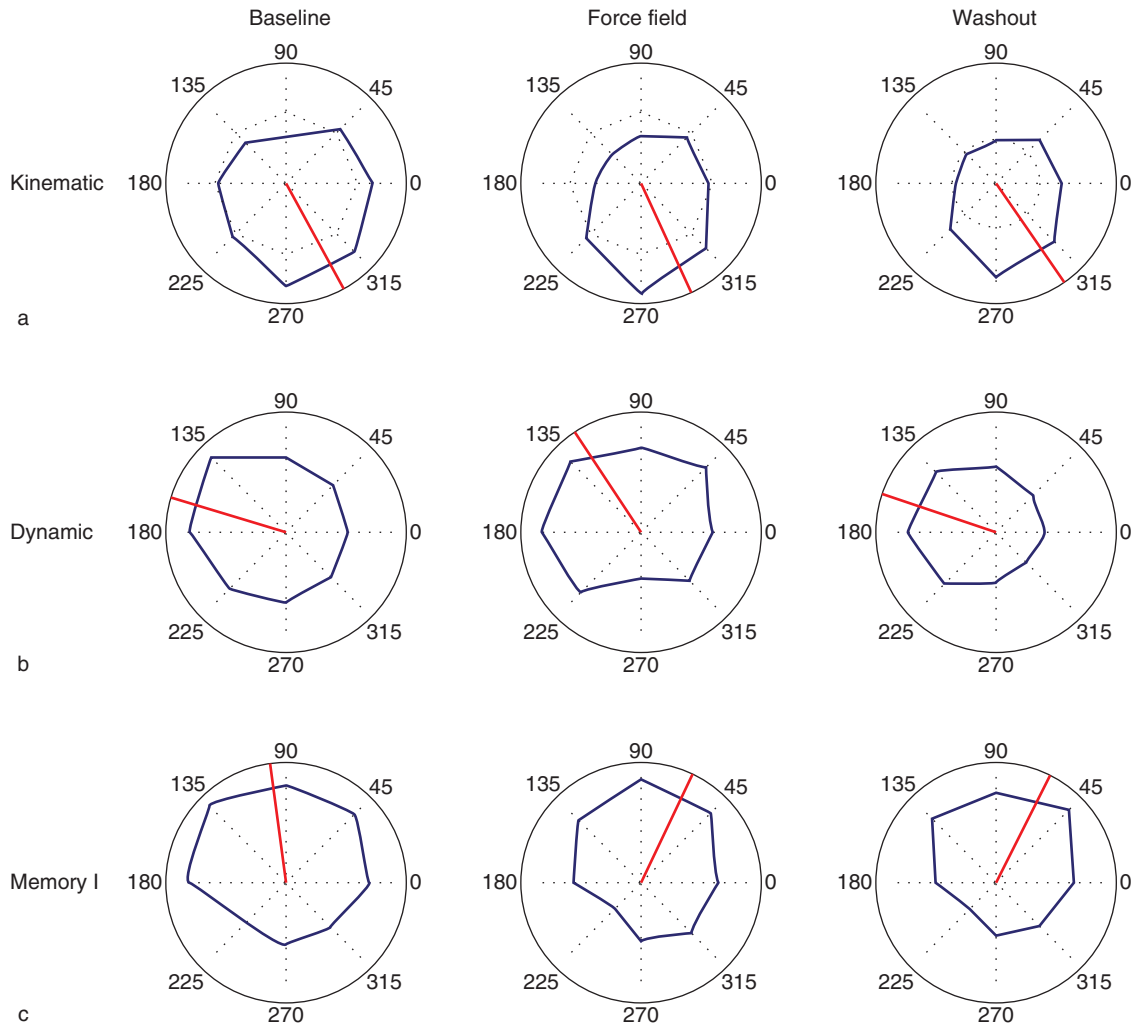

Figure 7 Rotation of the preferred direction for different types of cells in the motor cortex during force field learning. Macaque monkeys were trained to move a manipulandum on a horizontal plane to move a cursor on a vertical monitor from one central location to eight peripheral targets. As in the experiment described in Figure 6, a rightward hand movement (0°) was mapped to a rightward cursor movement and a forward (90°) hand movement to an upward cursor movement. However, in this experiment two torque motors acting on the manipulandum, during the 'force field' epoch, generated a clockwise velocity-dependent curl field, imposing a force on the hand perpendicular to the movement direction. (a) An example of the tuning curves of the movement-related activity of a 'kinematic' cell before, during, and after adaptation to the force field. The preferred direction (indicated by a red segment in the polar plot) of this cell does not change across epochs. (b) An example of a 'dynamic' cell: Its preferred direction rotates in the direction of the force perturbation during the force field epoch but returns to the initial direction in the washout epoch. (c) An example of 'memory' cell: The preferred direction rotates during the force field exposure and does not change back after the force field has been removed. Adapted from Li CS, Padoa-Schioppa C, and Bizzi E (2001) Neuronal correlates of motor performance and motor learning in the primary motor cortex of monkeys adapting to an external force field. *Neuron* 30(2): 593–607, with permission from Elsevier.

of the kinematic transformation of motor error into movement plan rather than the dynamic transformation of movement plan into muscle activations.

During learning of novel forces applied by a robot manipulator, the neuronal activity of single cells in the motor cortex during movement execution also shows a complex pattern of changes. Macaque monkeys, like humans, learn with practice to reach accurately under a viscous force field, and they show aftereffects when the perturbation is removed. Some cells show no change in magnitude of activity modulation and its directional tuning after learning the force field (**Figure 7(a)**). These cells might contribute to a kinematic representation of the motor plan. Some cells change their modulation magnitude and directional tuning only during the presence of the force field (**Figure 7(b)**). This group of cells appears to be involved in the representation of the motor output. In fact, the preferred direction of these cells rotates, similar to the muscles, in the direction of the force perturbation. Finally, other cells change their tuning during the force field but maintain this change after the perturbation is removed (**Figure 7(c)**). This group of cells might be encoding the learned novel dynamic mapping required for reaching in the force field. At the population level, this persistent change in the directional tuning after the perturbation is removed is balanced by a change in the tuning, in the opposite direction, of a fourth group of cells.

See also: Finger Movements: Control; Motor Skill Learning; Motor Sequences; Neural Coding in Primary Motor Cortex; Premotor Cortex in Primates: Dorsal and Ventral; Premotor Areas: Medial; Reaching and Grasping; Synaptic Mechanisms of Learning.

Further Reading

Atkeson CG (1989) Learning arm kinematics and dynamics. *Annual Review of Neuroscience* 12: 157–183.

Clower DM and Boussaoud D (2000) Selective use of perceptual recalibration versus visuomotor skill acquisition. *Journal of Neurophysiology* 84(5): 2703–2708.

Flanagan JR, Nakano E, Imamizu H, et al. (1999) Composition and decomposition of internal models in motor learning under altered kinematic and dynamic environments. *Journal of Neuroscience* 19(20): RC34.

Krakauer JW, Ghilardi MF, and Ghez C (1999) Independent learning of internal models for kinematic and dynamic control of reaching. *Nature Neuroscience* 2(11): 1026–1031.

Lackner JR and DiZio P (2005) Motor control and learning in altered dynamic environments. *Current Opinion in Neurobiology* 15(6): 653–659.

Martin TA, Keating JG, Goodkin HP, Bastian AJ, and Thach WT (1996) Throwing while looking through prisms. II. Specificity and storage of multiple gaze-throw calibrations. *Brain* 119: 1199–1211.

Shadmehr R and Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience* 14(5 Pt 2): 3208–3224.

Shadmehr R and Wise SP (2005) *The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge, MA: MIT Press.

Kinship Signals in Animals

J M Mateo, University of Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Biologists have long been interested in the evolution of sociality and the factors promoting cooperation and competition among individuals. Social relationships mediate speciation, population dynamics, mating systems, and reproductive success, and theoretical work has focused on the advantages of helping kin (nepotism), avoiding inbreeding, and engaging in affiliative and agonistic relationships with conspecifics as a function of relatedness. WD Hamilton's seminal work on inclusive-fitness theory, arguably the most noteworthy advance in evolutionary biology since Darwin's theory of natural selection, posited that individuals could increase their genetic fitness directly by producing offspring, as well as indirectly by helping collateral kin produce their offspring. Thus behaviors which are costly to an individual can evolve if they help kin to survive and to pass on genes shared in common. Since Hamilton's work, significant advances have been made in understanding the functions of kin recognition in facilitating nepotism (a specific form of altruism in which costly behaviors directly benefit kin) and mate choice (minimizing costs of inbreeding and outbreeding). Hamilton also noted that selection would favor the evolution of mechanisms that permit accurate discrimination among kin. The discussion here will address kin recognition from both of Tinbergen's levels of analysis, including proximate explanations (the developmental, physiological, and cognitive aspects of recognition occurring within an individual's lifetime) and ultimate explanations (the phylogeny and adaptive significance of kin recognition across generations).

Kin recognition is an internal process of assessing genetic relatedness that can be inferred through kin discrimination – the observable, differential treatment of conspecifics based on cues that correlate with relatedness. The recognition process typically involves three components (Table 1): the expression of unique phenotypic cues, or labels (such as odors, plumage patterns, or vocalizations; referred to as the production component), the perception of these labels and their degree of correspondence with a recognition template (the perception component), and the action taken by the animal as a function of the perceived similarity between its template and an encountered phenotype (the action component). The particular action taken, if any, depends on the degree of relatedness between individuals and the costs or benefits of the action. Production and perception components make up the mechanism of recognition, both of which are necessary for discrimination of conspecifics. That is, the action component requires this mechanism before any differential treatment of conspecifics can occur (e.g., nepotism, inbreeding avoidance, aggression, avoidance, cooperation).

Thus recognition labels and perceptual-matching abilities provide proximate or mechanistic explanations for kin recognition. The action component provides a functional or adaptive explanation for recognition because it involves behaviors with fitness costs or benefits (such as agonism or avoidance of inbreeding). As a consequence, recognition may not lead to discrimination; likewise, a lack of discrimination does not necessarily indicate a lack of recognition ability. It is important to note that each component of recognition is subjected independently to natural selection (and perhaps sexual selection), such that if kin labels and perceptual abilities have not both evolved, differential treatment of individuals cannot be selected, regardless of how favorable it might be. For example, in species in which there is little philopatry or overlap of generations, selection may not favor the evolution of a production or perception component to identify kin, because there is little risk of mating with kin. Similarly, nepotistic behaviors might be evident in one species but not in a closely related species if those animals do not live near kin or if the costs of such favoritism are not outweighed by the inclusive-fitness benefits.

The Production Component

Any phenotypic trait that correlates reliably with relatedness can serve as a kin label. In many taxa, recognition of conspecifics is mediated through chemical cues, including insects, amphibians, mammals, fish, and, perhaps, birds. Depending on the species, recognition odors can include urine and feces, cuticular hydrocarbons, glandular secretions, or secretions from unspecialized skin cells. Odors, or other kin labels, can have multiple functions, particularly in mammals, when odors are complex mixtures of several to hundreds of compounds. Not only can odors indicate an animal's individual identity, sex, or relatedness, they can also reveal its genetic quality, its reproductive status, its age, or even its location.

Although it is widely recognized that odors can mediate kin recognition, it is still unclear how odors

Table 1 The components of kin recognition

Component	*Description*
Production component	
Phenotypic trait which reliably correlates with relatedness ('kin label')	Chemical, visual, acoustic, tactile, behavioral, spatial, or mating cues
Perception component	
Prior association	Familiar individuals recognized as kin
Phenotype matching	Cues of unfamiliar individuals are matched to a recognition template and recognized as kin if match is close
Context based	Spatial location, mating exclusivity, or cohort sharing used as contexts for recognizing individuals as kin
Recognition alleles	Alleles control production of kin label, perception of label, and resulting cooperation
Action component	
Nepotism	Costly behaviors which benefit kin
Mate choice	Avoidance of extreme inbreeding and outbreeding

convey information about genetic relatedness. Assessment of relatedness via odors may be facilitated by the major histocompatibility complex (MHC), a large group of genes involved in vertebrate immune function, which also influences odor production. These highly variable genes, some having over 100 alleles per locus, regulate recognition of self versus non-self tissues and pathogens. Because of the large number of alleles involved, the likelihood of two random conspecifics having the same MHC haplotypes is very low, but since relatives, by definition, share many genes in common, the MHC can serve as an accurate proxy of relatedness. Further, the extreme variability of the MHC likely facilitates an individual's ability to resist a broad range of rapidly evolving parasites and pathogens, and therefore could be used in mate-choice contexts, not only to reduce inbreeding but also to promote heterozygosity in offspring if adults prefer MHC-dissimilar mates. In addition to coding for glycoproteins, the MHC also influences odor production, probably through an interaction between MHC by-products and bacteria on gland surfaces (leading to unique secretions) or in gastrointestinal tracts (creating unique urine and fecal odors). Thus families share both similar MHCs and similar odors; MHC-derived odors could serve as kin labels that are learned from family members during early development (see later, discussion of perception component).

Laboratory rodents (rats and mice) can discriminate between the odors produced by highly inbred animals that are genetically identical except for a single MHC allele. They also often prefer to mate with MHC-dissimilar individuals, although the strength of this preference varies with the sex of the choosing individual and the experimental methods. Little is known about whether MHC-derived odors mediate mate choice and nepotism in free-living outbred animals, although mate preferences are influenced by the MHC in some salmonids, lizards, jumping rats, pheasants, and, perhaps, humans. Further, because diet and health status can influence MHC-based odors of inbred laboratory rodents, it is important to determine if foraging habits and parasite loads affect perception of genetic differences among odors in free-living animals as well.

Much of the research on the production component in rodent kin recognition has emphasized MHC-based odors, but recent work has identified another compound that may be involved in social recognition. In house mice, major urinary proteins (MUPs) are large, nonvolatile compounds that bind and carry the more volatile odor molecules that are traditionally thought to mediate recognition. Each male excretes urine with 10–12 different genetically encoded MUPs, and the resulting distinct odors are sufficient for individual recognition.

Production Cues from Other Modalities

Because odors can vary with genetic relatedness and provide evolutionarily reliable cues to assess kinship, most studies of kin recognition have focused on this modality. However, relatedness could potentially be assessed through other modalities, such as vocalizations or visual markers. If the production of such cues is stable across time, then young animals could become familiar with them and use them for discriminating among kin classes. Nonolfactory kin labels might be important if chemical differences were inaccurate indicators of relatedness, such as when nests of social insects contain multiple matrilines but all workers acquire similar odors from the nest material upon eclosion. Yet if cues are to accurately mediate kin recognition of previously unfamiliar kin (via phenotype matching; see later), then differences among cues must be heritable. That is, behavioral cues such as vocalizations or visual displays may differ among families, but if these behaviors are plastic (whether

during early development or throughout the life span), then they are subject to drift and may not reliably co-vary with relatedness. For example, nestling long-tailed tits learn the contact calls of their adult caretakers, many of whom are kin. These calls can serve as kin labels for recognition of familiar family members, but since calls are learned, they unlikely to mediate recognition of unfamiliar kin, such as cousins or half-siblings.

Mechanisms of Kin Recognition: Perception Component

As noted previously, the evolution of kin-biased behaviors requires the evolution of mechanisms that permit recognition of kin. Several mechanisms for kin recognition have been described, and the use of particular mechanisms varies across and within species as a function of ecological factors, demographic patterns, and life-history variables. These mechanisms are not mutually exclusive or necessarily exhaustive, and individuals might use several mechanisms, depending on their age or the recognition context. Two of the mechanisms, prior association and phenotype matching, have been the focus of most empirical research and are described in detail in the following sections.

Prior-Association Mechanism

Recognition can be based on familiarity via prior association, when animals learn the phenotypes of individuals (often during early development) and later discriminate these familiar relatives from unfamiliar animals. For instance, animals might learn the traits of their parent(s) and same-aged siblings, as well as their older or younger siblings if dispersal is delayed, and recognize these individuals later, regardless of when or where they are encountered. Learning can occur through habituation, imprinting, or associative learning, although the actual type of learning involved is usually unknown. Prior association can also mediate recognition through with a third individual, such as when siblings born in different years both interact with their parent or when a parent's behavior toward the younger offspring promotes similar behavior by the older offspring (e.g., helpers at the nest). Note that with prior association, animals can recognize familiar individuals, but unfamiliar kin such as paternal half-siblings or cousins would be identified as non-kin.

As a proxy for relatedness, prior association or familiarization can be adequate for accurate recognition of kin when relatives reliably interact in the absence of non-kin during early learning, such as at nest sites or in exclusive home ranges. If young animals are precocial and mix with non-kin shortly after hatching or birth, learning should occur rapidly, to restrict establishment of familiarity with kin only.

Parent–offspring recognition and sibling–sibling recognition are often achieved through prior association when litter, brood, or clutch size is >1 or when young remain with their natal group. This mechanism also operates when there is little selection pressure to favor discrimination among equally familiar individuals (e.g., full maternal and half-maternal siblings) or equally unfamiliar individuals (kin and non-kin), such as when animals are short-lived with little overlap of generations or when there are few opportunities for social interactions with other kin. However, prior association will lead to recognition mistakes if non-kin are encountered during the learning phase (e.g., communal nesters; parasitized nests with eggs of other females), or if close kin are not encountered until some time later (e.g., young born in different years).

Phenotype-Matching Mechanism

Through a generalization from prior association, animals can learn their own phenotypes or those of their familiar kin, store these memories in a recognition template, and later compare or match the phenotypes of unknown animals to this template. Phenotype matching requires a correlation between phenotypic and genotypic similarity so that individuals with traits that most closely match an animal's template are its closest kin; that is, kin labels must be heritable. Unlike prior association, phenotype matching permits discrimination among unfamiliar kin and non-kin. That is, unfamiliar relatives are 'recognized' as belonging to a particular kin class. This recognition occurs through generalization from recognition templates, with the degree of match between an encountered phenotype and an individual's template indicating the degree of relatedness between the two animals.

Phenotype matching is favored when animals typically encounter unfamiliar kin, because it allows discrimination among kin and non-kin without prior direct contact with them. It is expected when males mate multiply (for recognition of paternal half-siblings), when females nest communally (so females can discriminate against familiar young that are not their own), when individuals disperse as juveniles or after successful breeding (allowing them to recognize their dispersed siblings or parents that they had not previously encountered), or when there is overlap of generations, particularly in long-lived species (allowing recognition of aunts, cousins, etc.). Self-referential phenotype matching, using one's own cues as a referent for recognizing kin, should be favored in species with multiple paternity or maternity to discriminate

among full siblings and half-siblings or when individuals commonly encounter older (or younger) siblings after dispersal. It would also be favored in brood parasitism or communal nesting, when individuals encountered early in life might not be kin and thus would be poor referents. It would be especially advantageous when highly accurate estimates of relatedness are required, because an individual's own cues most reliably reflect its genotype.

Phenotype-matching abilities can be demonstrated empirically by cross-fostering young at birth or hatching and later testing their ability to recognize unfamiliar siblings (this assumes that kin cues are not learned prenatally), or by testing the ability of normally reared individuals to recognize unfamiliar kin, such as older full siblings or paternal half-siblings. A more robust demonstration of phenotype matching would involve manipulation of recognition templates with subsequent changes in patterns of social interactions.

Context-Based Recognition

Some types of discrimination are based on contextual cues rather than on cues borne by individuals. First, animals may be recognized indirectly based on spatial cues, with individuals encountered in a particular area treated as kin regardless of true relatedness (e.g., parent–offspring recognition in nests). Spatial cues may correlate reliably with relatedness, such as when females lay their eggs or give birth in a nest to which other individuals do not have access, and when communal nesting and brood parasitism do not occur. A second contextual cue that can reliably correlate with kinship is mating effort (assuming internal fertilization). Here, recognition of offspring is based on the amount of time a male had exclusive mating access to a female, as a rule-of-thumb for estimating paternal certainty, rather than based on traits of offspring. This mating-effort mechanism of discrimination will be favored if it reliably leads to increased investment in own kin (albeit not exclusive), or in avoidance of or aggression toward non-kin. A third contextual cue, referred to as 'cohort sharing,' can allow young to recognize their paternal half-siblings (and perhaps other classes of unfamiliar kin), and can allow for adults to identify likely paternal half-siblings and perhaps even to assess paternity. If breeding in a social group is seasonal and is dominated by one or a few males, then most young born into a cohort are likely to be offspring of that one male or those few males. However, if cohort sharing is to be a reliable mechanism for recognition, the timing of births must be episodic enough for distinctions between separate cohorts to be clear to all in the group.

Recognition-Allele Mechanism

Finally, recognition could be mediated by so-called recognition alleles that cause expression of a phenotypic cue, recognition of that cue in others, and preferential treatment of individuals bearing the cue. This recognition mechanism has not been the focus of much research because, conceptually, it would lead to cooperation with non-kin who also express that cue, thereby selecting against such a single-allele cue-based cooperation system (Dawkins' green-beard effect). For example, planktonic larval ascidians (*Botryllus schlosseri*) settle near and eventually fuse with kin, more than with non-kin, and these kin-biased aggregations are beneficial because colonies composed of kin grow and reproduce faster than do aggregations of distant or non-kin. Larval settlement patterns are mediated by cellular recognition, as larvae fuse with other individuals if they share a histocompatibility allele but reject those with a different allele, even if those larvae are kin. Thus in this system the expression, perception, and action components all appear to be influenced by histocompatibility alleles.

As previously discussed, the vertebrate MHC, a large and highly polymorphic set of genes involved in immune functioning, influences the production of kinship-correlated odors. In addition, olfactory-receptor genes are found in the MHC (although it is currently unknown whether these genes are functional), which suggests that the MHC could serve as recognition alleles. To date there is no evidence that the MHC also causes preferential treatment of individuals with similar MHCs, so, although the MHC influences the production of genetically distinct odors, these odors were likely co-opted for kin-recognition purposes independently.

Action Component

Once an animal assesses its degree of genetic relatedness to another individual (kin recognition), it can then interact with that individual accordingly (kin discrimination); however, as noted earlier, recognition does not always lead to differential treatment. The action component generally takes two forms: nepotism (kin favoritism) and mate choice. Nepotistic behaviors include coalitions with kin, parental care of offspring, warning kin of predators, reducing or sacrificing reproduction in favor of kin, and selective withholding of aggression against kin. Nepotism is typically seen in group-living animals, particularly among the philopatric sex, such that animals live near and frequently encounter kin.

According to kin-selection theory, nepotism results in increased reproductive success of recipients; that is,

there are fitness costs to the nepotist and fitness benefits to the receiver. Specifically, nepotism is favored when Hamilton's rule, $rb - c > 0$, is satisfied, where r is the coefficient of relationship between two individuals, b is the fitness benefit to the recipient, and c is the fitness cost to the actor. This may result in a net inclusive-fitness benefit to the nepotist, if the action helps the receiver to reproduce and pass on genes the two individuals share in common. Because the benefits to the recipient are devalued by r, nepotism is typically directed toward close kin only.

Accurate kin recognition is favored in mate-choice contexts because inbreeding depression can be a potent selective force, according to theory and the available empirical data from both captive and free-living populations. The deleterious effects of inbreeding can be minimized through a variety of processes, such as multiple mating or extra-pair matings, postcopulatory mate choice, sex-biased breeding or natal dispersal, and, importantly, kin recognition with active avoidance of kin. Kin recognition might also facilitate mating preferences to optimize gene combinations. Although there is ample evidence that relatedness influences mate choice and reproductive success in captive animals, less is known about choice in free-living animals that must make trade-offs between kinship, mate quality, and search time. It has been argued that the kin-recognition mechanisms for nepotism and mate choice are organized similarly, but for most animals it is unknown whether there are indeed commonalities in the mechanisms used.

Figure 1 Adult female Belding's ground squirrel with two of her juvenile offspring. The left juvenile is engaged in a nasal investigation of its mother, smelling her oral glands, which provide information about individual and kin identity. Photo by JM Mateo.

Issues in Kin Recognition

Evolution of Recognition Components

Kin labels can evolve specifically for recognition purposes, or they can be artifacts of some other unrelated mechanism(s). For example, sexually selected traits in male birds, such as song repertoires, which are learned from male parents, might also serve as patrilineal markers. Oral- and dorsal-gland odors of Belding's ground squirrels (Figure 1) are both individually distinct and kin distinct, suggesting a generalized mechanism by which secretions reflect genetic variation among individuals as well as among kin classes. (The production of multiple, unique odors, or other cues, can facilitate accurate discrimination of conspecifics.) Although songs and odor cues probably did not evolve for recognition purposes, use of them to discriminate individuals and kin classes would provide a parsimonious mechanism for social recognition.

Kin-selection theory predicts that recognition components (production, perception, and action) will evolve with sociality, depending in part on species' social systems, the nature of interactions with kin, and costs and benefits of nepotism. For example, despite the wide range of social systems among voles in the genus *Microtus*, including group size, patterns of natal dispersal, and frequency of multiple mating, social interactions and reproductive behaviors of most species examined to date appear to be influenced by prior association rather than by relatedness (although there have been few empirical studies of their recognition abilities). Similarly, depending on the social system, kin-recognition abilities may not lead to differential treatment. Like Belding's ground squirrels, golden-mantled ground squirrels also produce kin-distinct oral odors and can discriminate among kin classes, even though this species shows no evidence of nepotistic behaviors except between mothers and their dependent young. However, these recognition abilities might be used in mate-choice contexts, to avoid mating with close kin.

Given the predominance of sex-biased dispersal among most vertebrates, some have claimed that there is no need for other inbreeding-avoidance mechanisms, such as kin recognition, yet few data are available to support this prediction. Dispersal does not

eliminate the risk of mating with half-siblings or cousins, In addition, adults may remain where they mate year after year and thus risk mating with their reproductively mature offspring which have not dispersed. Finally, if kin recognition can be used to avoid mating with, for example, cousins or half-siblings, it is parsimonious to expect that a similar mechanism is used to avoid very close inbreeding.

Model System for the Study of Recognition Components

Belding's ground squirrels are one of a few species for which all three recognition components have been investigated. In this species females are philopatric and can survive up to 12 years; the resulting kin clusters have favored the evolution of nepotism. Females with close kin (e.g., mothers, daughters, sisters) are more likely to give potentially fatal alarm calls than are females without close kin. During gestation and lactation, when females defend natal burrows against potential infanticide, they are less agonistic toward and more cooperative with close kin than are non-kin or distant kin (e.g., aunt, cousin; Figure 2). Maternal half-sisters are more agonistic than are full sisters, suggesting discrimination of equally familiar littermates that vary in relatedness. Males, which disperse as juveniles and rarely live near kin, are not nepotistic nor do they behave nepotistically. Mothers use spatial cues to discriminate among young, caring for any pup in their underground nests, until just before young leave the nest (Figure 3). After young emerge from the nest at about a month of age, spatial cues no longer reliably correlate with relatedness, because young can mix with litters in other territories. However, at this age offspring are beginning to produce unique recognition odors, and this corresponds with when mothers are able to use a prior-association mechanism for discriminating among own and unfamiliar young. Odors from both oral and dorsal glands are kin distinct, being more similar among close kin than among distant kin, which in turn are more similar than are those of non-kin (Figure 4(a)). Juveniles

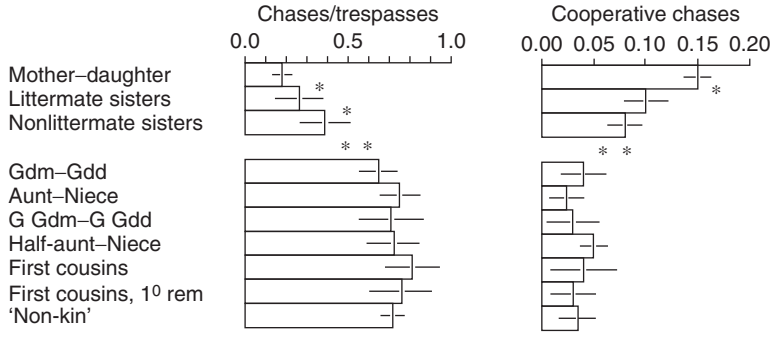

Figure 2 Nepotism among adult female Belding's ground squirrels. Proportion (mean and standard deviation) of times females chased (left) or assisted in chases with (right) various kin and non-kin classes from their territory during gestation and lactation. Asterisks indicate that close kin were tolerated and cooperated with more often than were distant or non-kin. Adapted from Sherman PW (1981) Kinship, demography, and Belding's ground squirrel nepotism. *Behavioral Ecology and Sociobiology* 8: 251–259, with kind permission of Springer Science and Business Media.

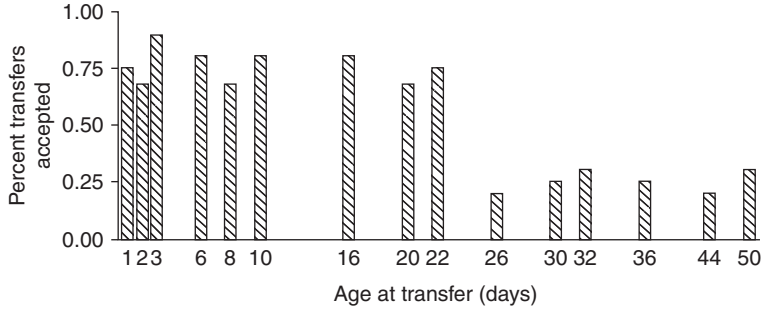

Figure 3 Proportion of young *Spermophilus beldingi* accepted by unrelated adult females in a field cross-fostering experiment. Free-living *S. beldingi* mothers readily retrieved into their natal burrow unrelated young placed at their burrow entrance, but only until their own young reached about 25 days of age, which coincides with when mothers' own young emerge aboveground from their natal burrow. Reproduced from Holmes WG and Sherman PW (1982) The ontogeny of kin recognition in two species of ground squirrels. *American Zoologist* 22: 491–517, copyright by American Society of Zoologists, 1982, with permission.

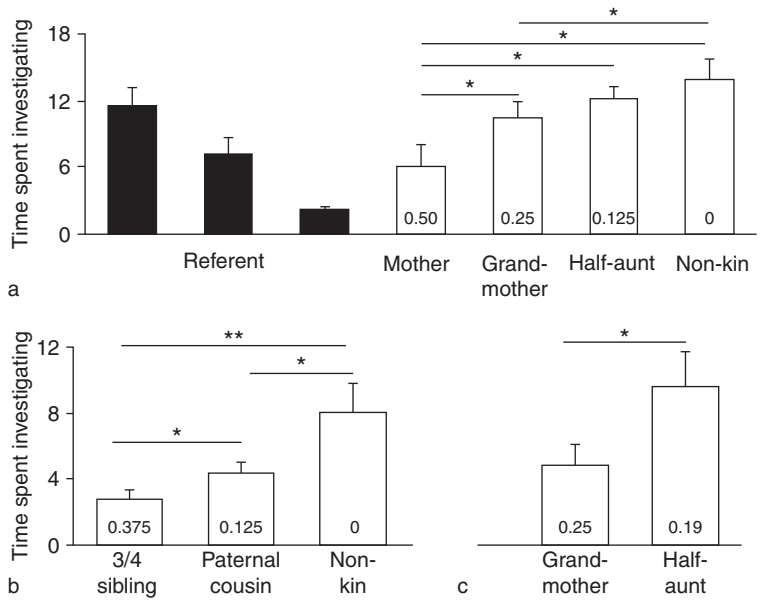

Figure 4 Tests of the production and perception components of kin recognition in Belding's ground squirrels. (a) Duration of investigation (mean second + SE) of dorsal-gland odors by 13 individuals in habituation–discrimination tasks designed to test whether odors vary with genetic relatedness. Solid bars represent investigation of the referent odor (three habituation trials) and open bars represent investigation of test odors (discrimination phase), collected from a referent's mother, maternal grandmother, half-aunt (mother's half-sister), and non-kin. Investigation of test odors increased as relatedness to the referent decreased, demonstrating that odors vary linearly with relatedness. (b, c) Duration of investigation (mean second + SE) of odors collected from unfamiliar relatives of seven individuals during preference tasks designed to determine whether the animals can spontaneously discriminate among odors of their unfamiliar kin. (b) Investigation of oral-gland odors of unfamiliar kin, collected from the individual's three-quarter sibling (offspring of two sisters mated with the same male), cousin (offspring of the individual's mother's brother), and non-kin. (c) Investigation of oral odors collected from individual's grandmother and aunt. An increase in investigation as relatedness of the odor donor to the individual decreases indicates that odors of distant kin are perceived as less familiar to individuals than are odors of close kin. Significant differences in investigation of odors from various kin classes indicate discrimination of those odors, and thus that the individual can recognize the difference between, for example, its distant kin and non-kin. Numbers inside bars are estimated coefficients of relationship between individuals and odor donors. Horizontal bars and asterisks represent significant differences in investigation of odors (*, $p < 0.05$; **, $p < 0.01$) based on repeated measures (ANOVAs or paired *t*-tests). Adapted from Mateo JM (2002) Kin-recognition abilities and nepotism as a function of sociality. *Proceedings of the Royal Society of London. Series B* 269: 721–727.

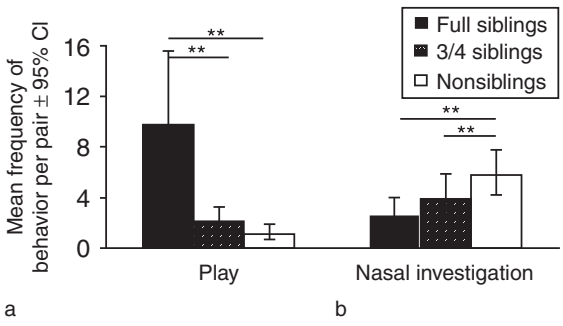

Figure 5 Mean frequencies (+ 95% CI) of (a) play and (b) nasal investigations between pairs of juveniles as a function of relatedness (three-quarter siblings are offspring of two sisters who mated with the same male). Horizontal bars and asterisks represent differences in behavior frequencies (*, $p < 0.05$; **, $p < 0.01$) based on Poisson regression analyses. Mean frequencies are adjusted statistically by the regression model for effects of sex of each juvenile in a pair and the weight difference between juveniles in each pair. Reproduced from Mateo JM (2003) Kin recognition in ground squirrels and other rodents. *Journal of Mammalogy* 84: 1163–1181, by permission of Alliance Communications Group, a division of Allen Press, Inc.; copyright by American Society of Mammalogists, 2003.

and adults can also use phenotype matching to discriminate among unfamiliar individuals based on genetic relatedness (Figures 4(b) and 4(c)). This recognition ability influences social behaviors, including nasal investigations (allowing perception of kin-distinct oral-gland odors) and play-partner preferences among juveniles (thought to influence the development of adult kin preferences; Figure 5). Discrimination among unfamiliar kin and non-kin, as evidenced by differential nasal investigation, therefore leads to preferential social interactions with kin. Both sexes are able to recognize their distant kin, indicating that the limited action component (nepotism among close female kin only) cannot be explained proximately by deficits in recognition ability. The importance of these kin recognition abilities in mate choice is currently under study.

Recognition Errors

Evolutionary theory predicts that animals will discriminate on the basis of relatedness if the benefits

outweigh the costs, such as preferential investment in own offspring. Yet in some instances adults appear to treat young indiscriminately, caring for their own young as well as so-called alien young. These parent–offspring recognition 'failures' should be explained at the ultimate level of analysis (e.g., evolutionary costs and benefits of recognition) as well as at the proximate level. That is, what recognition cues are produced and what perceptual mechanisms might be employed? There has been much theoretical and empirical interest in why cuckolded male birds invest in unrelated young, yet few studies investigate whether young birds even produce kin labels or whether males have the ability to discriminate among these labels. This recognition mechanism must evolve before preferential investment in one's own young (the action component) can be expressed. Further, if young animals benefit from kin-biased behaviors at an early age, then it follows that kin labels must develop early. For example, prairie vole mothers interact differentially with weanlings, but not with newborns, as a function of relatedness, and this result has been interpreted as a function of costs of investment in young across development. Yet a proximate explanation should be considered as well, such as whether newborns produce recognition cues.

The Meaning of 'Template'

The word 'template' is used to refer to memories of familiar kin, which are probably learned during early development. Here, a template is not thought of as a neural structure, such as face-recognition cells or song-nuclei circuits. Instead, it is typically used as a heuristic for conceptualizing recognition mechanisms, although it is still unclear how kin labels are remembered. That is, do templates consist of exemplars (specific memories of individuals' phenotypes) or prototypes (a single amalgam or gestalt of several individuals' phenotypes)? Many species can discriminate among specific kin, such as sibling A and sibling B, suggesting that templates contain some specific representation of these individuals. When an unfamiliar relative is encountered, such as a half-sibling, animals could refer to these separate representations for recognition (necessitating multiple comparisons between each representation and the stranger's phenotype), or they could use a gestalt template composed of those representations (requiring just one global comparison). Functionally, the outcome would be the same whether exemplars or prototypes are used, but these possibilities pose interesting neurobiological questions. Templates may be both exemplar based and prototype based, since some animals can both recognize individual kin (via prior association) and 'recognize' unfamiliar kin (via phenotype matching).

Neural Correlates of Perception

In most taxonomic groups, unfamiliar stimuli receive more attention than do familiar stimuli. This includes a higher frequency or duration of tactile inspection, odor investigation, or visual attention toward unfamiliar individuals or most distantly related kin, compared with familiar individuals or close kin, respectively (e.g., Figures 4(b) and 4(c)). These investigatory behaviors suggest that it takes longer to identify or recognize unfamiliar individuals, compared to familiar ones, as if additional time is needed for unknown cues to be compared or matched to templates. However, recent perceptual studies with laboratory rodents suggest otherwise, because trained animals take the same amount of time to discriminate simple odors, regardless of the similarity of the odor stimuli and thus the difficulty of the task. If these findings apply to other species, then why do animals attend to unfamiliar cues longer? It is possible that an unknown phenotype is perceived and characterized quickly (e.g., in the olfactory bulb or antennal lobe) by an animal, but it needs additional time to match it to the recognition template and accurately determine its degree of relatedness to the new individual. Familiar individuals would match the template, producing a threshold response (match or not match), whereas unfamiliar individuals would have to be evaluated against the template, with some decision algorithm used to produce a graded response according to relatedness.

Future Directions

In the four decades since Hamilton's seminal work on kin selection, most of the research on kin recognition has focused on the perception component, determining how animals use familiarity and relatedness to discriminate among conspecifics. More work is needed on the production component, determining the extent to which phenotypic traits vary with kinship and thus can potentially be used as kin labels. As labels are identified, our understanding of how they evolved would be enhanced by a quantitative assessment of the correlation between genetic similarity and cue similarity. For example, gas chromatography and mass spectrometry have revealed that beavers' (*Castor canadensis*) anal-gland secretions are more similar among kin than among non-kin, and can be used for phenotype matching; their castoreum is largely diet derived and does not co-vary with relatedness, and thus is not a reliable kin label. Finally, for a complete understanding of kin recognition, we need to determine whether discrimination and resulting kin-biased behaviors are indeed adaptive. That is, do individuals in fact increase their fitness by performing nepotistic

behaviors or by rejecting kin as mates? Evidence of a kin bias alone is insufficient to posit a function for kin-recognition abilities.

See also: Chemical Senses: Overview; Game Theory and the Economics of Animal Communication; Multimodal Signaling in Animals; Olfactory Cortex: Comparative Anatomy; Olfactory Neuron Patterning and Specification; Olfactory Coding; Olfactory Receptors; Pheromones and other Chemical Communication in Animals; Sexual Selection and the Evolution of Animal Signals.

Further Reading

Beecher MD (1982) Signature systems and kin recognition. *American Zoologist* 22: 477–490.

Fletcher DJC and Michener CD (1987) *Kin Recognition in Animals*. New York: Wiley.

Gamboa GJ, Reeve HK, and Holmes WG (1991) Conceptual issues and methodology in kin recognition research: A critical discussion. *Ethology* 88: 109–127.

Hamilton WD (1964) The genetical evolution of social behaviour, I. II. *Journal of Theoretical Biology* 7: 1–52.

Hepper PG (1991) *Kin Recognition*. Cambridge, UK: Cambridge University Press.

Holmes WG and Sherman PW (1982) The ontogeny of kin recognition in two species of ground squirrels. *American Zoologist* 22: 491–517.

Holmes WG (2001) The development and function of nepotism. In: Blass EM (ed.) *Handbook of Behavioral Neurobiology: Developmental Psychobiology*, vol. 13, pp. 281–316. New York: Kluwer Academic/Plenum Publishers.

Holmes WG and Sherman PW (1982) The ontogeny of kin recognition in two species of ground squirrels. *American Zoologist* 22: 491–517.

Mateo JM (2002) Kin-recognition abilities and nepotism as a function of sociality. *Proceedings of the Royal Society of London. Series B* 269: 721–727.

Mateo JM (2003) Kin recognition in ground squirrels and other rodents. *Journal of Mammalogy* 84: 1163–1181.

Mateo JM (2004) Recognition systems and biological organization: The perception component of recognition. *Annales Zoologici Fennici* 41: 729–745.

Mateo JM and Holmes WG (2004) Cross-fostering as a means to study kin recognition. *Animal Behavior* 68: 1451–1459.

Penn D (2002) The scent of genetic compatibility: Sexual selection and the major histocompatibility complex. *Ethology* 108: 1–21.

Pusey A and Wolf M (1996) Inbreeding avoidance in animals. *Trends in Ecology & Evolution* 11: 201–206.

Sherman PW (1981) Kinship, demography, and Belding's ground squirrel nepotism. *Behavioral Ecology and Sociobiology* 8: 251–259.

Sherman PW (1991) Multiple mating and kin recognition by self-inspection. *Ethology and Sociobiology* 12: 377–386.

Zelano B and Edwards VE (2002) An MHC component to kin recognition and mate choice in birds: Predictions, progress, and prospects. *The American Naturalist* 160: S225–S238.

Kisspeptins and their Receptors

M Tena-Sempere, University of Córdoba, Cordoba, Spain

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Reproductive function is sustained by a complex neurohormonal system, which conforms the gonadotropic or hypothalamic–pituitary–gonadal (HPG) axis. The main hierarchical element of this axis is the hypothalamic decapeptide, gonadotropin-releasing hormone (GnRH), which is secreted by a discrete neuronal population of the forebrain (the GnRH neurons) and operates as driving signal for the pulsatile release of the pituitary gonadotropins, luteinizing hormone (LH) and follicle-stimulating hormone (FSH). These are, in turn, responsible for the development and trophic maintenance of different gonadal functions, such as spermatogenesis, folliculogenesis, and ovulation. In addition, secretion of gonadal steroids and peptides is elicited by gonadotropins, thereby contributing to the dynamic regulation of this neuroendocrine axis via negative and positive feedback loops. Moreover, multiple central and peripheral regulators participate in the tuning of reproductive function, mostly acting primarily on hypothalamic GnRH neurons, which operate as integrators and a common output pathway for a wide diversity of modulatory factors. However, the mechanisms and neuronal networks projecting on to GnRH neurons and mediating such regulatory actions at different developmental and physiological conditions are far from being completely deciphered.

The KiSS-1/GPR54 System: Identification, Structural Features, and Initial Functions

The KiSS-1/GPR54 is a ligand–receptor system whose elements were sequentially identified between 1996 and 2001. First, by the use of subtractive hybridization in melanoma cell lines with different metastatic capacity, KiSS-1 was identified and originally cataloged as metastasis-suppressor gene, given that its expression was significantly attenuated in tumor cells with high invasiveness. Independently, in 1999, GPR54 was cloned in the rat as an orphan receptor with a partial sequence similarity (>40%) with the transmembrane regions of galanin receptors; the human ortholog of GPR54 was subsequently cloned and termed AXOR12 or hOT7T175. Finally, in 2001, the structural characterization of the peptide products of the KiSS-1 gene was completed. These turned out to be a family of structurally related peptides, globally termed kisspeptins, which result from the differential proteolytic processing of a common precursor. All kisspeptins were found capable of binding and activating GPR54, which was then deorphanized and cataloged as the cognate receptor of KiSS-1 peptides.

In the human, the KiSS-1 precursor is a 145-amino-acid protein, which contains a putative 19-amino-acid signal sequence, two potential dibasic cleavage sites (at amino acids 57 and 67), and one putative site for terminal cleavage and amidation (at amino acids 121–124). Its major proteolytic product appears to be a 54-amino-acid peptide, largely secreted by the placenta, termed metastin or kisspeptin-54. In addition, other peptide fragments of the KiSS-1 precursor have been identified, such as kisspeptin-14, kisspeptin-13, and kisspeptin-10. All kisspeptins share the C-terminal region of the metastin molecule, where they present an Arg-Phe-NH_2 motif distinctive of the RF-amide peptide family. The shorter 10-amino-acid fragment at the C-terminus (kisspeptin-10) retains maximal activity at the receptor level.

The biological functions initially assigned to the KiSS-1 system were restricted to its role in the control of tumor progression. Thus, the initial observations that loss of KiSS-1 expression is linked to increased invasiveness of melanoma cells were followed by a number of studies assessing the expression of KiSS-1 gene and/or the biological actions of kisspeptins in different tumors cell lines. In brief, those studies demonstrated that metastin and other kisspeptins are provided with antimetastasis activity in several tumors, and the loss of KiSS-1 gene expression was suggested as potential index for tumor progression and metastasis in different types of cancer. In any event, the actual relevance of this system in the control of cancer progression remains to be fully clarified.

In parallel to the characterization of their potential role in tumor progression, expression analyses of KiSS-1 and GPR54 genes in normal tissues from humans and rodents strongly suggested that the KiSS-1/GPR54 system is probably involved in additional biological functions. Thus, KiSS-1 gene expression was detected in human placenta, testis, small intestine, and brain, where KiSS-1 mRNA showed scattered distribution throughout the central nervous system. Similarly, GPR54 gene expression was demonstrated in a number of central and peripheral tissues, with maximal mRNA levels in the placenta, pancreas, pituitary, spinal cord, and various brain areas. Indeed, based on their patterns of distribution

in the central nervous system, it was suggested that kisspeptins and their receptor play a role in nociception, visceral regulation, and/or neuroendocrine control. Yet the functional evidence to support these assumptions was very limited, and in the case of neuroendocrine control, it was initially restricted to the demonstration of the ability of kisspeptin-10 to stimulate oxytocin release. At peripheral tissues, the observation that KiSS-1 gene expression is maximal in human placenta, together with the dramatic increase in serum metastin levels observed during pregnancy, led to the hypothesis that kisspeptins may participate in the control of key aspects of placental physiology. In this sense, kisspeptin-10 has been demonstrated to inhibit trophoblast invasion *in vitro*, although the physiological relevance of this function is still a matter of debate. Other potential peripheral functions of the KiSS-1/GPR54 system, in tissues such as the pancreas and gonads where expression of this system has been reported, remain mostly unexplored.

Distribution of KiSS-1 mRNA/Protein in the Brain

In this article, we focus on the analysis of the role of kisspeptins in the neuroendocrine control of the reproductive axis. As an introduction, it is considered relevant to provide a map of the known patterns of distribution of KiSS-1 expression (at the mRNA and peptide levels) in the central nervous system, with special attention to the hypothalamus. However, results from different studies do not totally overlap, and the mRNA and peptide data are not completely analogous.

Concerning KiSS-1 gene expression, initial analyses were conducted by means of reverse transcription polymerase chain reaction (RT-PCR) of isolated brain regions in humans. Those studies demonstrated a scattered pattern of expression, with moderate levels of KiSS-1 mRNA being detected throughout the central nervous system, including the basal ganglia and the hypothalamus. Likewise, KiSS-1 gene expression was also demonstrated in the rat brain, specifically in the hypothalamic area. These expression data have been recently refined by *in situ* hybridization analyses that demonstrated the presence of KiSS-1-expressing neurons at different sites along the rostral–caudal extent of the hypothalamus in the mouse, including the arcuate nucleus (Arc) and the anteroventral region of the periventricular nucleus (AVPV). In addition, KiSS-1 mRNA-expressing cells have been also detected in the anterodorsal preoptic nucleus, medial amygdala, and bed nucleus of the stria terminalis. In primates, expression of KiSS-1 mRNA has been also observed in the Arc.

Regarding distribution of KiSS-1 peptides, an exhaustive immunohistochemical study on the rat brain was initially reported, showing that metastin-like immunoreactivity (irMT) is widely detected throughout the central nervous system. Thus, irMT was observed in the dorsomedial hypothalamic nucleus, ventromedial nucleus, and Arc within the hypothalamus; nucleus of the solitary tract; and caudal ventrolateral medulla. In addition, fibers with different levels of intensity of irMT were detected in different areas of the telencephalon (amygdala, bed nucleus of stria terminalis, septal nuclei, nucleus accumbens, and basal ganglia), diencephalon (several hypothalamic regions, thalamus, and zona incerta), mesencephalon (periaqueductal gray and raphe nuclei), metencephalon, and myelencephalon (lateral parabrachial nuclei, locus coeruleus, spinal trigeminal tract, and medullary reticular nucleus). However, these immunohistochemical analyses were conducted using a polyclonal antiserum against human kisspeptin-10, which might cross-react with other peptides showing the terminal RF-amide motif. More recently, an additional study, using a different antibody raised against mouse kisspeptin-10, identified three major populations of hypothalamic neurons expressing kisspeptins located at the anteroventral and preoptic periventricular regions (AVPV/PeN), dorsomedial hypothalamus, and Arc. In addition, the latter study provided the first morphological demonstration of the close apposition between kisspeptin fibers and GnRH neuron cell bodies within the hypothalamus.

Finally, our knowledge of the pattern of expression of GPR54 in the central nervous system remains somewhat limited. Thus, although RT-PCR assays initially demonstrated that GPR54 mRNA is present in the spinal cord and various brain areas, including the hypothalamus, basal ganglia, amygdala, substantia nigra, and hippocampus, no immunohistochemical analysis of the distribution of GPR54 protein within the brain has so far been reported. Likewise, precise location analyses of GPR54 mRNA across the central nervous system are still pending. Yet the expression of GPR54 mRNA has been demonstrated in rat GnRH neurons by means of *in situ* hybridization.

Disclosure of the Reproductive Facet of KiSS-1/GPR54

Identification of the connection between the KiSS-1/GPR54 system and the neuroendocrine control of reproductive function did not take place until late

2003, when two research groups independently reported the presence of deletions and inactivating mutations of GPR54 gene in patients suffering familial or sporadic forms of idiopathic hypogonadotropic hypogonadotropism (HH). Since then, a limited number of additional loss-of-function mutations of the GPR54 gene have been described in humans affected by hypogonadism of central origin. Moreover, simultaneous with the description of the first cases of hypogonadotropism in humans bearing mutations in GPR54, mice engineered to lack functional GPR54 were reported to be a complete phenocopy of affected patients. Altogether, these data evidenced an indispensable, highly conserved role of GPR54 and its ligands in the control of reproduction in mammals.

Although from an epidemiological perspective the number of cases of HH caused by GPR54 mutations appears to be low, disclosure of the essential role of the KiSS-1/GPR54 system in the control of reproduction drew considerable attention from a physiological standpoint. Indeed, initial hormonal studies conducted in humans and mice with genetic inactivation of GPR54 strongly suggested that the KiSS-1 system is fundamental in the control of GnRH secretion at the hypothalamus. This assumption was based on the observation that GPR54 mutation did not disrupt the capacity of the pituitary to respond to GnRH. Similarly, lack of GPR54 signaling did not prevent normal migration of GnRH neurons from the olfactory placode at early stages of development, nor did it impair GnRH synthesis, as indirectly suggested by the demonstration of preserved GnRH content at the hypothalamus of GPR54 knockout mice. Obviously, the hypothesis that the KiSS-1 system operates as major regulator of GnRH secretion needed to be experimentally validated. Compelling evidence from molecular, physiological, and pharmacological studies, gathered during the last 3 years, has now substantiated that kisspeptins and their receptor are probably (one of) the most important gatekeepers of the function of GnRH neurons and, hence, of the gonadotropic axis. The most salient roles of KiSS-1 and GPR54 in the control of reproduction are reviewed in the following sections.

KiSS-1 System and the Control of Gonadotropin Secretion: Negative Feedback

In the context of reproduction, one of the facets of the KiSS-1 system that initially attracted more attention was its potential role in gonadotropin secretion. Thus, between 2004 and early 2005, pharmacological studies conducted in different species, such as the rat, mouse, sheep, and macaque, demonstrated that kisspeptins are very potent stimulators of LH and FSH release. More recently, the ability of metastin to stimulate LH secretion has been also shown in humans. Interestingly, some differences have been noticed in the patterns of LH and FSH responses to kisspeptins, such as their relative magnitude (higher for LH) and time course (faster for LH), as well as in the threshold and mean effective doses (two orders of magnitude lower for LH). In any event, it is becoming evident that kisspeptins are probably the most potent elicitors of the GnRH–gonadotropin axis known so far (see Figure 1). A particular feature of the gonadotropin-releasing effect of kisspeptins is that it is detectable even after the systemic administration of low doses of metastin or kisspeptin-10. This is in contrast with most of the neuropeptides involved in the central control of gonadotropin secretion, which are ineffective when administered systemically.

The mechanism whereby the KiSS-1 system activates the reproductive axis and stimulates gonadotropin secretion involves a primary action at the hypothalamus, where kisspeptins seem to operate as major synaptic excitatory signals for GnRH neurons. This contention was initially suggested (indirectly) by the demonstration that blockade of GnRH actions (by means of a specific antagonist) completely blunted the ability of kisspeptins to stimulate gonadotropin secretion *in vivo*. Thereafter, double-labeled *in situ* hybridization studies in the rat showed that >75% of GnRH neurons co-express GPR54 mRNA and that kisspeptin-52 (equivalent to metastin in the rat) is able to induce c-fos expression in >85% of GnRH neurons. Moreover, hormonal analyses

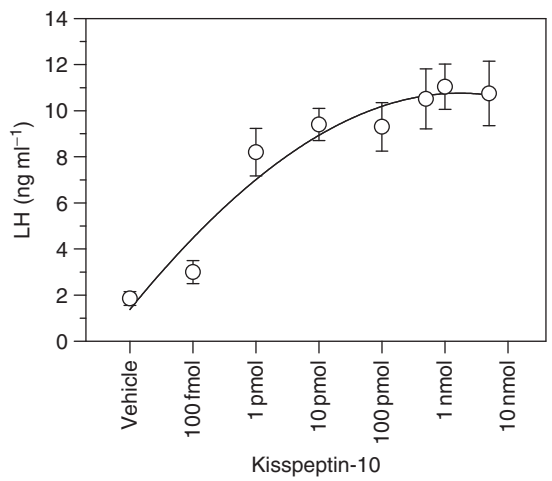

Figure 1 Potent LH releasing activity of kisspeptin after intracerebral administration. A range of doses of kisspeptin-10 (5 nmol, 1 nmol, 500 pmol, 100 pmol, 10 pmol, 1 pmol, and 100 fmol) were tested. Nonlinear regression allowed to estimate an ED_{50} of ~2 pmol for LH responses to central injection of kisspeptin. ED_{50}, mean effective dose; LH, luteinizing hormone.

demonstrated the ability of kisspeptins to stimulate GnRH secretion by rat hypothalamic explants *ex vivo*, as well as to induce the release of GnRH into the cerebrospinal fluid in the sheep after *in vivo* administration. In sum, these data are evidence that kisspeptins act primarily on hypothalamic GnRH neurons to elicit GnRH secretion, which in turn stimulates LH and FSH release from the pituitary.

Further evidence for the relevance of the KiSS-1 system in the physiological control of gonadotropin secretion came from expression analyses in the hypothalamus in models of gonadectomy in the rat. Thus, bilateral orchidectomy and ovariectomy induced a significant increase in KiSS-1 (and, to a lower extent, GPR54) mRNA expression in whole hypothalamic fragments. Such changes in mRNA levels closely paralleled the expected rise in circulating LH concentrations following gonadectomy. Moreover, sex steroid replacement fully prevented both hormonal and gene expression responses. These observations strongly suggested that the hypothalamic KiSS-1 system plays an essential role in relaying the negative feedback input of sex steroids to GnRH neurons. Indeed, location analyses by means of *in situ* hybridization fully confirmed these original observations and demonstrated that the subset of KiSS-1 neurons in the Arc, where the expression of KiSS-1 mRNA is enhanced after gonadectomy and decreased following androgen or estrogen supplementation, is responsible for conveying the negative feedback actions of sex steroids on gonadotropin secretion.

KiSS-1 System and the Female Gonadotropic Axis: Positive Feedback

In situ hybridization analyses of KiSS-1 mRNA expression at the hypothalamus of ovariectomized rats also provided experimental evidence for a higher degree of complexity of the KiSS-1 system and its regulation by sex steroids. Whereas KiSS-1 mRNA levels in the Arc appeared to be repressed by estrogen, KiSS-1 gene expression in the AVPV was suppressed by ovariectomy and enhanced by estrogen replacement. Considering that the AVPV had been previously identified as an hypothalamic area involved in the positive feedback control of gonadotropins, those observations raised the intriguing possibility that the population of KiSS-1 neurons at this hypothalamic site might be involved in the positive feedback effect of estradiol on gonadotropin secretion during the preovulatory phase. This hypothesis has been recently supported by the demonstration that KiSS-1 mRNA levels in the AVPV, but not in the Arc, significantly increase in the rat during the evening of proestrus (i.e., coinciding with the preovulatory surge of gonadotropins). Altogether, these data evidence that KiSS-1 neurons in the AVPV participate in the network responsible for the generation of the positive feedback effect of estradiol on gonadotropin secretion in the female, thus playing a pivotal function in the generation of the hormonal signal triggering ovulation. This is in good agreement with previous evidence showing that immunoneutralization of central metastin prevents the preovulatory surge of gonadotropins in the cyclic rat. Note that, although the existence of the positive feedback effects of estradiol on gonadotropin secretion had been known for decades, the actual neuronal circuitry responsible for such a phenomenon remained largely unknown. The contribution of KiSS-1 neurons in the AVPV and Arc in mediating the positive and negative feedback effects of estradiol is tentatively depicted in Figure 2.

KiSS-1 System and Onset of Puberty

In keeping with the ability of kisspeptins to stimulate gonadotropin secretion, the hypothalamic KiSS-1 system is also involved in the physiological mechanisms of the onset of puberty, as indirectly suggested initially by the demonstration of a state of impuberism in humans and mice with inactivating mutations of GPR54. Thus, expression analyses in hypothalamic

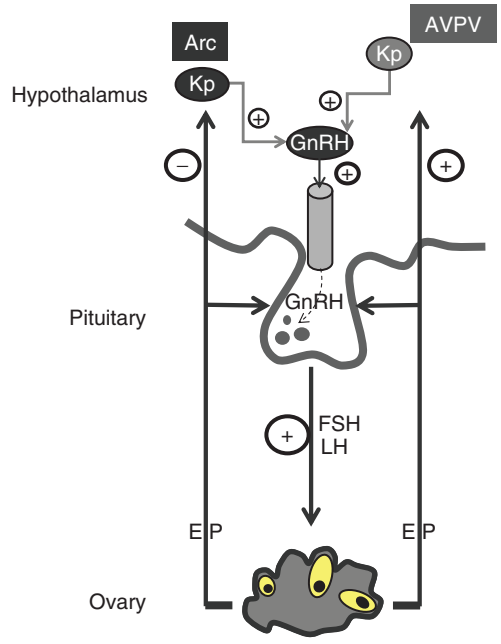

Figure 2 Tentative model for the differential role of KiSS-1 neurons at the arcuate nucleus (Arc) and the anteroventral periventricular nucleus (AVPV) of the hypothalamus in relaying the negative and positive feedback effects of sex steroids (estradiol) on GnRH neurons. +, positive feedback; −, negative feedback; EP, estradiol; FSH, follicle-stimulating hormone; GnRH, gonadotropin-releasing hormone; Kp, kisspeptin; LH, luteinizing hormone.

samples from male and female rats at different stages of postnatal development demonstrated a significant increase in KiSS-1 and GPR54 mRNA levels at the time of puberty onset, which was preceded by low expression levels during the prepubertal period. Likewise, hypothalamic levels of KiSS-1 and GPR54 mRNAs significantly increased during the transition from the juvenile to the mid-pubertal stage in monkeys. Moreover, repeated injections of kisspeptin-10 in immature female rats induced the precocious activation of the gonadotropic axis (with earlier vaginal opening, increased LH and estradiol levels, and elevated uterus weights). Similarly, chronic administration of kisspeptin-10 at the end of the juvenile phase evoked a sustained train of GnRH discharges, similar to that found at puberty, in primates.

More recently, functional and expression analyses have suggested that activation of the hypothalamic KiSS-1 system is not merely derived from the enhancement of KiSS-1 gene expression at the time of puberty. Thus, *in vivo* and *in vitro* studies, testing the effects of different doses of kisspeptin on LH secretion and its ability to activate GnRH neurons, have demonstrated that a significant increase in the sensitivity of GnRH system to kisspeptin activation takes place in rodents during puberty; a phenomenon which might involve the enhanced expression of GPR54 and increased efficiency in its coupling to intracellular effector systems. In addition, the maturational activation of the KiSS-1 system at puberty probably involves plastic changes in fiber projections and synaptic density (especially for KiSS-1 neurons at the AVPV) because a significant increase in the number of appositions between kisspeptin fibers and GnRH cell bodies has been demonstrated in the mouse along pubertal transition. Altogether, these observations demonstrate that the activation of the KiSS-1 system at the time of puberty is a complex phenomenon that involves changes in KiSS-1 and GPR54 gene/peptide expression, receptor signaling efficiency, and physical interactions between KiSS-1 and GnRH neurons. Such a sophisticated, multistep regulatory process clearly illustrates the fundamental role of the KiSS-1/GPR54 system in the activation of the reproductive axis at puberty.

KiSS-1 System and the Metabolic Control of Reproduction: Modulation by Leptin

In addition to its role in puberty onset and feedback regulation of gonadotropin secretion, the hypothalamic KiSS-1 system also has been involved in the neuroendocrine circuitries responsible for the metabolic control of reproductive function. This contention was originally proposed based on the observed changes in KiSS-1 gene expression in models of energy insufficiency linked to hypogonadotropic hypogonadism. Thus, short-term fasting, which induced the expected reduction in basal gonadotropin levels, also decreased KiSS-1 mRNA expression in the hypothalamus of male and female rats, 60 h after food deprivation. Similarly, in another model of metabolic alteration associated to hypogonadotropism, such as streptozotocin (STZ)-induced diabetes, decreased expression of KiSS-1 gene was observed in male and female rats, both in intact condition and after gonadectomy. Such changes in hypothalamic KiSS-1 expression closely paralleled the reduction of serum LH levels detected in STZ-injected animals. Altogether, these data suggest that the suppression of hypothalamic KiSS-1 system in states of energy deficit and metabolic disturbance is mechanistically relevant for the observed inhibition of the gonadotropic axis in those conditions.

Further evidence for the relevance of the KiSS-1 system in the metabolic regulation of reproduction came from functional studies analyzing the effects of administration of kisspeptins on different parameters of the gonadotropic axis in conditions of energy insufficiency. Thus, repeated administration of kisspeptin-10 to immature female rats subjected to a regimen of undernutrition (30% caloric restriction), which was able to prevent puberty onset, was sufficient to restore vaginal opening (as an external index of puberty) in a significant number of animals (60%) and to induce robust gonadotropin and estrogen responses. Likewise, despite the prevailing reduction in basal gonadotropin levels, kisspeptin administration in diabetic rats was capable of acutely eliciting maximal LH responses and to ameliorating several indices of reproductive dysfunction (such as reduced prostate and testis weight, as well as decreased LH and testosterone levels) after repeated administration. In sum, results from expression analyses and functional tests strongly suggest that the body energy status has an influence on the hypothalamic KiSS-1 system and that replacement with kisspeptin can rescue (at least partially) the hypogonadotropic state observed in conditions of undernutrition and metabolic alteration. In this context, the hypothalamic KiSS-1 neurons might be considered sensors for the energy status of the organism, thus contributing to the coupling of the state of energy reserves and the function of the reproductive axis in different physiologic and pathologic conditions.

The nature of the signal(s) responsible for conveying metabolic information to KiSS-1 neurons is being actively investigated. In this context, the adipocyte-derived hormone, leptin, has recently emerged as a major regulator of hypothalamic KiSS-1 expression.

Of note, leptin had been long known to operate as a key neuroendocrine integrator signaling the magnitude of body energy (fat) stores to multiple hormonal axes, including the reproductive system, on which leptin has been reported to play mostly a permissive/stimulatory role. Supporting the possibility of the direct modulatory effects of leptin, a substantial proportion of KiSS-1 neurons in the Arc have been recently shown to express leptin receptors in the mouse. In addition, leptin-deficient *ob/ob* mice were reported to have a lower expression of KiSS-1 gene in the hypothalamus, a phenomenon which could be rescued by exogenous leptin administration. Moreover, chronic intracerebral infusion of leptin was sufficient to normalize hypothalamic KiSS-1 mRNA levels, as well as LH and testosterone concentrations, in hypogonadotropic diabetic male rats, which are also severely hypoleptinemic. Altogether, these data suggest that circulating leptin modulates the functioning of the reproductive axis (and, specifically, of GnRH neurons), at least partially, via regulation of the hypothalamic KiSS-1 system. This indirect mode of action on GnRH neurons helps to explain the paradoxical observation that, although it is sensitive to the actions of leptin, this neuronal population does not apparently express leptin receptors.

Proposed Model: The KiSS-1 Neuron as Novel Gatekeeper of the GnRH System

The identification of the prominent reproductive roles of kisspeptins and their receptor has been defined as the most important finding in the field of neuroendocrinology of reproduction since the isolation of GnRH in early 1970s. These fundamental functions of kisspeptins involve key roles in the full activation of the gonadotropic axis at puberty, the regulation of gonadotropin secretion by sex steroids, the generation of the preovulatory surge of LH (as hormonal trigger for ovulation), and the metabolic control of reproductive maturation and fertility. In addition, environmental cues with prominent

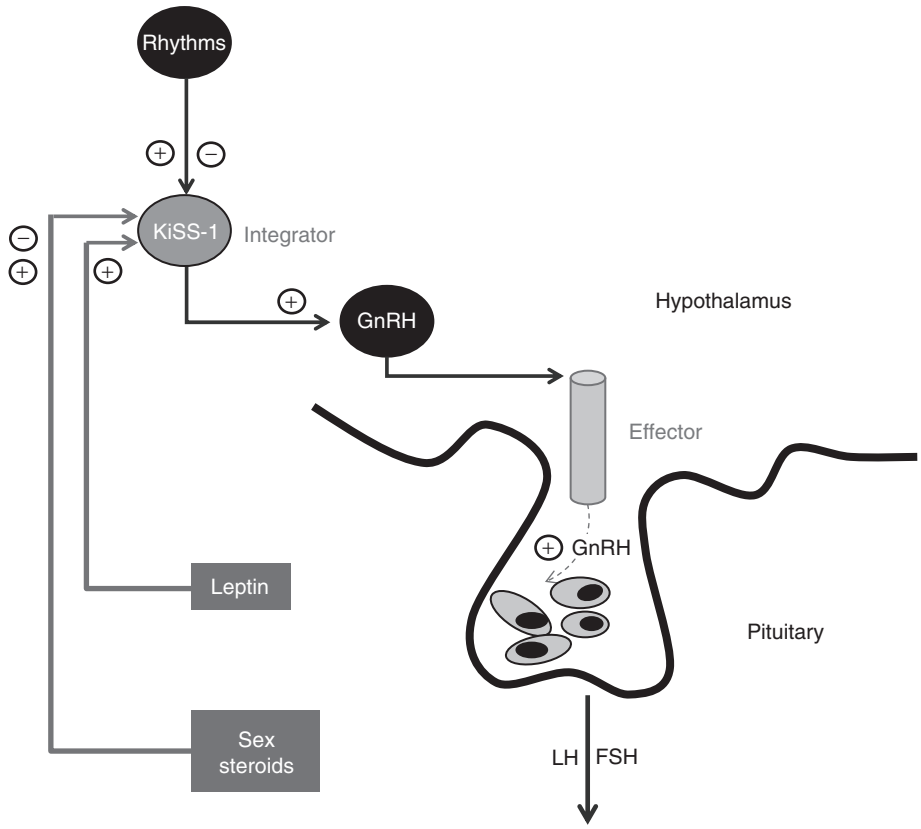

Figure 3 Tentative model for the central role of hypothalamic KiSS-1 neurons in the neuroendocrine control of reproduction. According to this model, KiSS-1 neurons are the target for the modulatory actions (stimulatory, +; inhibitory, −) of different regulators of the gonadotropic axis, such as sex steroids, leptin, and environmental signals, on which this information converges and integrates. In turn, KiSS-1 neurons operate as major regulators of GnRH neurons, which respond to changes in kisspeptin input with activation or, eventually, inactivation (if kisspeptin tone decreases); GnRH is the ultimate effector for kisspeptin actions on the gonadotropic axis. FSH, follicle-stimulating hormone; GnRH, gonadotropin-releasing hormone; LH, luteinizing hormone.

modulatory actions on the HPG axis, such as dark–light rhythms, have been recently suggested to operate via the hypothalamic KiSS-1 system. Considering the short time that has elapsed since the first discovery of the reproductive dimension of kisspeptins, it is understandable that several aspects of these functions remain to be fully characterized. In any event, the experimental evidence gathered in the last 2 years allows us to propose a tentative model for the integration of KiSS-1 neurons in the central networks governing reproductive function, as depicted in Figure 3. Thus, whereas it was conventionally assumed that GnRH neurons operate both as integrators and final output pathway for the array of regulatory signals of the gonadotropic axis, in this hypothetical model the KiSS-1 neuron functions as the center for the dynamic integration of the information conveyed by different central and peripheral modulators of the HPG system. In contrast, the GnRH neuron might play a more passive role, acting as the final transducer of the information relayed by the KiSS-1 system. This hypothetical model does not exclude the potential integration of additional excitatory and inhibitory inputs at the level of GnRH neurons. Notwithstanding, the KiSS-1 system can be now defined as a novel, essential gatekeeper of the reproductive axis. Further progress in our knowledge of the physiological roles of this system in the control of reproduction will require the complete characterization of the afferents to and projections of KiSS-1 neurons in the hypothalamus, as well as the identification of the major central and peripheral regulators of hypothalamic KiSS-1 expression and function, in what promises to be a very active area of research in neuroscience in the near future.

See also: Gonadotropin, Neural and Hormonal Control; Gonadotropin-Releasing Hormone: GnRH-1 System; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Neuroendocrine Control of Energy Balance (Central Circuits/Mechanisms); Neuroendocrinology of Puberty.

Further Reading

Castellano JM, Navarro VM, Fernandez-Fernandez R, et al. (2005) Changes in hypothalamic KiSS-1 system and restoration of pubertal activation of the reproductive axis by kisspeptin in undernutrition. *Endocrinology* 146: 3917–3925.

de Roux N, Genin E, Carel JC, et al. (2003) Hypogonadotropic hypogonadism due to loss of function of the KiSS1-derived peptide receptor GPR54. *Proceedings of the National Academy of Sciences of the United States of America* 100: 10972–10976.

Dungan HM, Clifton DK, and Steiner RA (2006) Kisspeptin neurons as central processors in the regulation of GnRH secretion. *Endocrinology* 147: 1166–1174.

Iovane A, Aumas C, and de Roux N (2004) New insights in the genetics of isolated hypogonadotropic hypogonadism. *European Journal of Endocrinology* 151: U83–U88.

Kotani M, Detheux M, Vandenbogaerde A, et al. (2001) The metastasis suppressor gene KiSS-1 encodes kisspeptins, the natural ligands of the orphan G protein-coupled receptor GPR54. *Journal of Biological Chemistry* 276: 34631–34636.

Messager S, Chatzidaki EE, Ma D, et al. (2005) Kisspeptin directly stimulates gonadotropin-releasing hormone secretion via G protein-coupled receptor 54. *Proceedings of the National Academy of Sciences of the United States of America* 102: 1761–1766.

Navarro VM, Castellano JM, Fernandez-Fernandez R, et al. (2004) Developmental and hormonally regulated messenger ribonucleic acid expression of KiSS-1 and its putative receptor, GPR54, in rat hypothalamus and potent luteinizing hormone-releasing activity of KiSS-1 peptide. *Endocrinology* 145: 4565–4574.

Navarro VM, Castellano JM, Fernandez-Fernandez R, et al. (2005) Characterization of the potent LH releasing activity of KiSS-1 peptide, the natural ligand of GPR54. *Endocrinology* 146: 156–163.

Ohtaki T, Shintani Y, Honda S, et al. (2001) Metastasis suppressor gene KiSS-1 encodes peptide ligand of a G-protein-coupled receptor. *Nature* 411: 613–617.

Seminara SB, Messager S, Chatzidaki EE, et al. (2003) The GPR54 gene as a regulator of puberty. *New England Journal of Medicine* 349: 1614–1627.

Seminara SB (2005) Metastin and its G protein-coupled receptor, GPR54: Critical pathway modulating GnRH secretion. *Frontiers in Neuroendocrinology* 26: 131–138.

Shahab M, Mastronardi C, Seminara SB, et al. (2005) Increased hypothalamic GPR54 signaling: A potential mechanism for initiation of puberty in primates. *Proceedings of the National Academy of Sciences of the United States of America* 102: 2129–2134.

Smith JT, Cunningham MJ, Rissman E. F, Clifton DK, and Steiner RA (2005) Regulation of Kiss1 gene expression in the brain of the female mouse. *Endocrinology* 146: 3686–3692.

Smith JT, Popa SM, Clifton DK, Hoffman GE, and Steiner RA (2006) Kiss1 neurons in the forebrain as central processors for generating the preovulatory luteinizing hormone surge. *Journal of Neuroscience* 26: 6687–6694.

Tena-Sempere M (2006) GPR54 and kisspeptin in reproduction. *Human Reproduction Update* 12: 631–639.

Relevant Websites

http://www.mgh.harvard.edu – Harvard University (puberty gene news release).

http://www.paradigm-therapeutics.co.uk – Paradign Therapeutics (news release).

http://www.sciencemag.org – *Science* Magazine.

Lamina-Specific Neuronal Connections

M Yamagata, Harvard University, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons are connected with a remarkable degree of synaptic specificity in neuronal circuits. One prominent feature of neuronal wiring is regional specificity, such as topographic maps, in which the spatial order of neurons maps systematically onto their axonal target. Similarly striking is laminar specificity of neuronal wiring in the orthogonal orientation. Many areas of the nervous system in both vertebrates and invertebrates consist of multiple neuronal cell layers, each forming lamina-specific connections with entering afferents and intrinsic interneurons. Thus, lamina-specific synaptic connectivity appears to be a major organizational principle of the central nervous system.

In vertebrates, clearly stratified parts include the cerebral cortex, olfactory bulb, hippocampus, lateral geniculate nucleus, superior colliculus (optic tectum), cerebellum, retina, and spinal cord. In these areas, each lamina bears a distinct series of neuronal subtypes, distinguishable by morphology, neurochemistry, and synaptic projections. The dendrites of such neurons span multiple laminae, but specific populations of input axons make arbors and synapses confined to particular laminae. Here, axons synapse onto neurons that are themselves restricted to particular laminae (**Figure 1(a)** also see later, **Figure 3(b)**). This specificity is seen in the cerebellum where climbing and pontine afferents synapse on Purkinje and granule neurons, respectively, which populate discrete layers. In other cases, inputs synapse on somata or lamina-confined segments of dendrites that span several laminae (**Figure 1(b);** also see later, **Figure 3(g)**). This class of subcellular domain specificity is spectacularly observed in the hippocampus, where different types of input fibers lay an array of synapses on distinct subdivisions of pyramidal and granule cell dendrites in a lamina-specific fashion. The concept of cellular and subcellular specificities can be also applied for lamina-specific connections between neurons wiring intrinsically in laminated structures. In many systems, however, lamina-specific synaptic connections involve a combination of both cellular and subcellular specificities.

During development, incoming axons form specific synaptic connections in the correct target laminae. Several mechanisms can be proposed that might explain how lamina-specific targeting, branching, and synapses are established (**Figure 2**). One possible scenario could be that the time of arrival of the distinct class of axons into developing laminae influences their lamina-specific pattern of innervation. However, recent studies have demonstrated a plethora of molecular and cellular mechanisms. These include (1) the diffusible and immobilized lamina-specific cues that attract or repel axons to target laminae (**Figure 2(a)**), (2) cell surface determinants displayed on lamina-specific postsynaptic neurons or their subcellular domain (**Figure 2(b)**), (3) nonneuronal cues associated with glial cells and/or extracellular matrix (**Figure 2(c)**), (4) guidepost cells that help navigate connections without directly participating in functional wiring (**Figure 2(d)**), and (5) activity-dependent mechanisms that sculpt initially diffuse projections (**Figure 2(e)**). Although accumulating evidence suggests that lamina-specific connections develop in different manners in different areas, it is likely that multiple mechanisms cooperate to generate lamina-specific synaptic connections in each system. In this article, these basic mechanisms are discussed in detail by citing some examples.

Lamina-Specific Signals

Growing axons arrive at the appropriate region after their journey. The growth cones still need to encounter a series of cellular and molecular cues until they are in the correct lamina at the apposition of synaptic targets: development of lamina-specific connections relies on a precisely organized spatiotemporal network of a panoply of lamina-restricted cues expressed along the axonal pathway (**Figure 2(a)**). These cues could be signals permitting or inducing axonal extension, nonpermissive or repulsive signals from inappropriate laminae, or branching or synaptogenic signals from target neurons to enable arborization or presynaptic differentiation in the target lamina. In recent years, much progress has been made in identifying molecules that play crucial roles in axon guidance. These molecules include the attractants and repellants that act in a long-range or in a contact-dependent manner. The similar molecular mechanisms can be envisioned to regulate lamina-specific responses of axons while passing through nontarget laminae, as well as after contacting the target lamina.

Several cues can be observed using organotypic cultures *in vitro*. In the chick retinotectal system, retinal ganglion cell axons enter the optic tectum through a superficial lamina, then turn into discrete deeper target laminae (**Figure 3(a)**). Using an

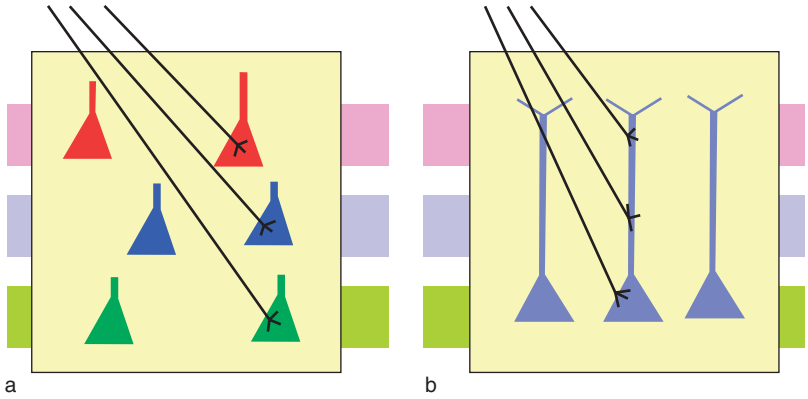

Figure 1 Lamina-specific neuronal connections. Entering inputs establish synaptic connections with specific types of neurons populating distinct laminae (a). In other cases, inputs terminate on discrete subdivisions of the target neuron's surface (b).

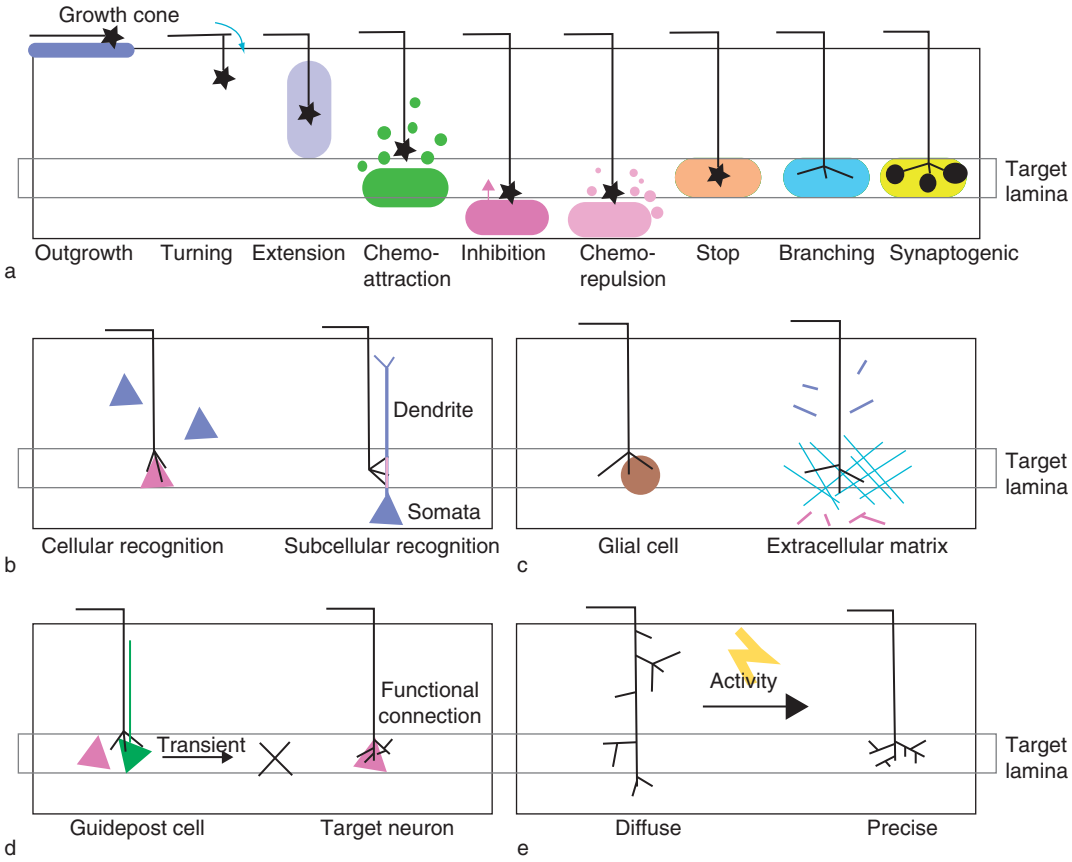

Figure 2 Developmental principles of lamina-specific neuronal connections. (a) Lamina-specific signals. Incoming axons outgrow, turn into the laminated structure, then extend orthogonally along several laminae until they contact the target lamina. The cues support a series of axonal behaviors, including axon outgrowth, turning into deeper laminae, intralaminar extension by contact guidance, chemoattraction mediated by diffusible factors, inhibition or collapse of outgrowth on inappropriate laminae, and chemorepulsion by nontarget laminae. Target-lamina-derived signals stop axonal growth, cause branching axon terminals, and induce synaptogenesis. (b) Target recognition. Cellular recognition involves direct choice of a correct target neuron populating the target lamina, whereas subcellular recognition involves selection of the particular subcellular domain of a target neuron (pink). (c) Nonneuronal cues. Glial cells and extracellular matrix can also regulate a mode of axon behavior, such as repulsion, attraction, and synapse formation. (d) Guidepost cells (green) are specific cellular cues used by axonal growth cones in axon pathfinding and targeting decisions. During development, incoming axons transiently contact guidepost cells, which help deliver the axons to final synaptic partners (pink). (e) Activity-dependent refinement. An initially diffuse pattern of connectivity is elaborated to establish a precise projection pattern by neural activity.

organotypic culture system, outgrowth, branching, and arborization patterns of retinal axons were recapitulated in the lamina-selective manner. The similar behavior of axons was essentially observed on chemically fixed tectal slices. Thus, each tectal lamina bears distinct immobilized cues that change lamina-selective behavior of retinal axons. In the cerebellum, cultured granule neurons arrested the outgrowth of their pontine mossy fiber afferents *in vitro*, suggesting that the cerebellar granule cell layer could halt growth of their afferents as a stop signal (Figure 3(b)). Although the molecular nature of these cues has not been fully characterized, these studies demonstrate that a panel of unique cues arranged in a lamina-restricted manner modulate a series of behavior of incoming axons.

In other systems, some of the lamina-specific cues have been molecularly identified. For instance, a role of lamina-specific expression of the cell surface ephrin/Eph family has been fairly well studied during early postnatal development of the mammalian cerebral cortex. In the six-layered cortex, thalamic axons make lamina-specific synaptic connections in the layer 4 (Figure 3(c)). Ephrin-A5 is strongly expressed in layers 4 and 6 when the thalamic afferents form synapses. *In vitro*, ephrin-A5 acts as a branch-promoting factor for thalamic axons: thalamic axons encountered on membrane substrates of layer 4 or ephrin-A5 develop more branches than when on layer 5 or on control substrate. Furthermore, *in vivo*, thalamic afferents from ephrin-A5-deleted mice stop normally in layer 4, but make less complex arbors. Thus, it appears that ephrin-A5 controls the arborization of thalamic axons in the target lamina. Neurotrophin-3 together with ephrin-A5 promote axonal branching of layer 6 neuronal axons, which target layer 4 expressing ephrin-A5. Importantly, ephrin-A5 inhibits branching of layers 2 and 3 neuronal axons, which normally avoid layer 4. Many other members of the Eph/ephrin family show a lamina-specific expression in other laminated structures as well as in the cortex, suggesting that the Eph/ephrin system plays a universal role in formation of lamina-specific connections.

Semaphorins comprise a large family of membrane-associated and secreted proteins that signal axon repulsion and/or attraction. Sema3A, one of the best characterized members in this family, is a diffusible molecule that induces growth cone collapse and axon repulsion of a variety of neuronal populations, through interaction with a receptor complex containing neuropilin-1/plexin-A1. One of its functions during development is thought to be the establishment of an appropriate pattern of lamina-specific axonal innervation within the dorsal horn of the spinal cord (Figure 3(d)). Sema3A functions to selectively repel axons from a subset of dorsal root ganglion neurons that are of small size, are nerve growth factor sensitive, and are responsible for nociception and thermoreception. Lamina-specific expression of Sema3A in the ventral horn during development is believed to restrict the growth of these axons to laminae I and II of the dorsal horn.

Target Recognition Molecules

The *sine qua non* of lamina-specific connections is to establish functional wiring by apposing two membranes of synaptic partners (Figure 2(b)). To give an explanation for development of specific neuronal wiring, in 1963, Roger Sperry formulated the 'chemoaffinity theory' in which complementary tags or adhesion molecules determine specificity of the 'lock' and 'key' kind. Consistent with this classic hypothesis, several cell surface molecules have been identified as potential recognition determinants in vertebrates and invertebrates.

Sidekicks (SDKs) have been considered as essential for the formation of lamina-specific synaptic connections in the retina. The vertebrate SDK molecules were identified in the course of a screen for genes expressed by subsets of chick retinal ganglion cells. Vertebrates have two SDKs, whereas insects and nematodes have only one SDK (rig-4 in *Caenorhabditis elegans*). From N- to C-terminus, each SDK protein contains a signal peptide sequence, six immunoglobulin C2 motifs, 13 fibronectin type III motifs, a single type II transmembrane domain, and an approximately 200-amino-acid cytoplasmic domain. In the chick retina, *SDK1* and *SDK2* genes are expressed by nonoverlapping subsets of the retinal ganglion cells, with each expressed in nearly a quarter of the cells there. The inner plexiform layer (IPL) of the retina is made from layers of axons and dendrites from the retinal ganglion cells, amacrine cells, and bipolar cells (Figure 3(e)). Distinct subsets of retinal ganglion cells confine their dendritic arbors to certain layers of the IPL, displaying distinct laminar specificity. In the IPL, a complementary localization of SDKs is also found in the synaptic regions. Dividing the IPL into five sublaminae (S1–S5), each occupying 20% of the IPL, showed that SDK1 is concentrated in S4 with lower levels in S2, whereas the converse is the case for SDK2. Surprisingly, ectopic expression of SDKs by in ovo electroporation of early chick embryos diverted projections of neurons from their normal position in SDK-negative sublaminae to SDK-positive sublaminae. Together with the homophilic adhesive interactions between each SDK, the gain-of-function experiment suggests that SDKs could function to target specific

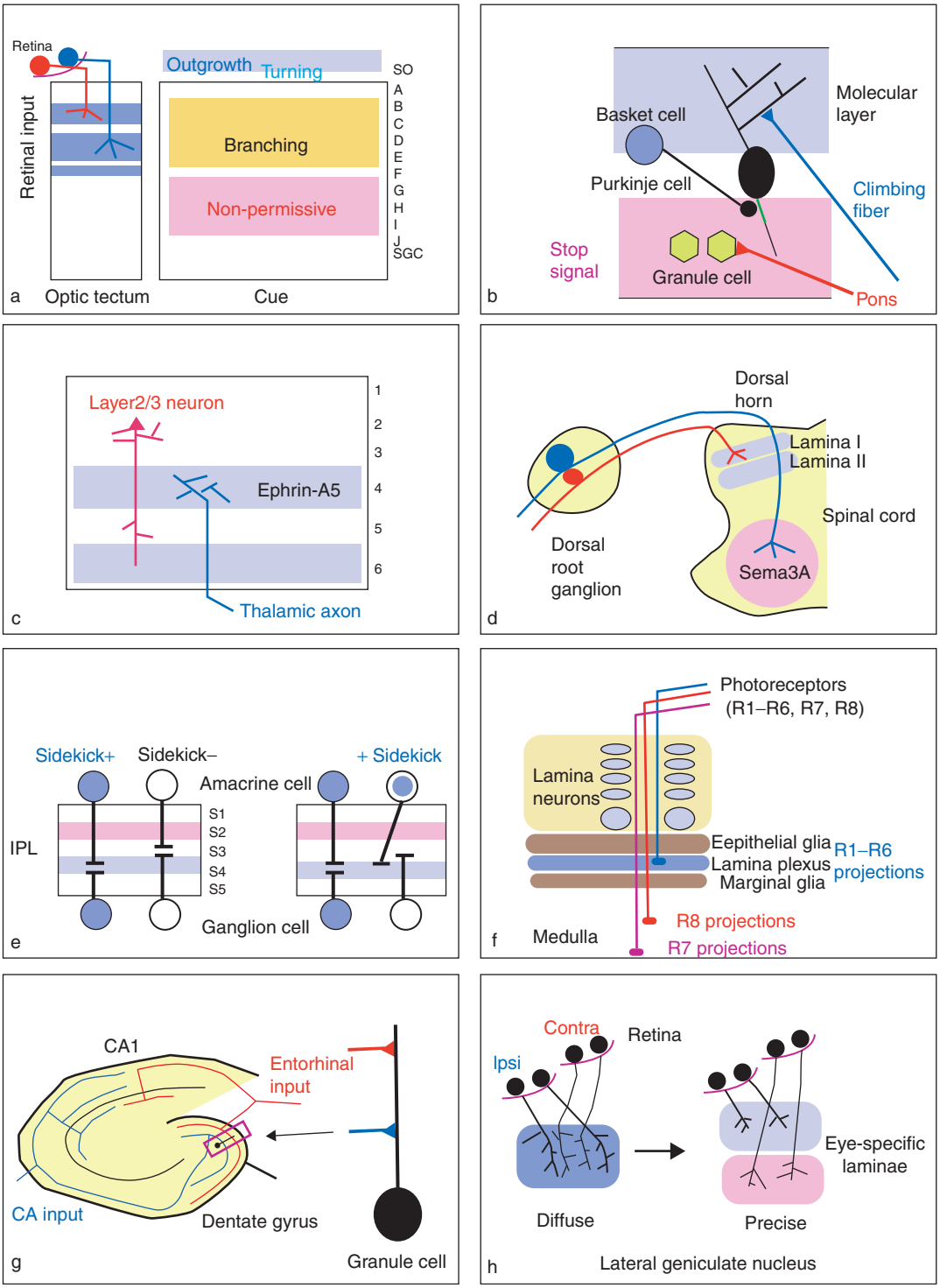

Figure 3 Examples of lamina-specific neuronal connections. (a) Retinotectal system: retinal ganglion cell axons project to only four retinorecipient laminae of the optic tectum, which contains >15 laminae. Experiments using an organotypic culture identified a series of lamina-selective cues which can control lines of axonal behavior. Distinct subsets of ganglion cells (red, blue) terminate in different laminae, although molecular and cellular mechanisms that underlie this fine laminar selection have not been elucidated. (b) Cerebellum: mossy fibers from the pons innervate granule cells in the granule cell layer, whereas climbing fibers from the inferior olive innervate onto dendrites of Purkinje cells in the molecular layer. This wiring pattern is thought to depend on the stop signal on granule cells. Basket interneurons make GABAergic ponceau synapses exclusively on the axonal initial segment (green) of Purkinje cells. (c) Cortex: thalamic inputs terminate in layer 4 of the six-layered cortex. This lamina-specific connection appears to be regulated by ephrin-A5 expressed in layers 4 and 6. By contrast, neurons in layers 2 and 3 develop their arbors on the outside of the ephrin-A5-expressing layers. (d) Spinal cord: distinct subpopulations of sensory neurons in the dorsal root ganglion project to different laminae of the spinal cord. Inputs from

neurites to certain sublaminae. SDKs are localized at the synapses and contain a postsynaptic density/disc/zonula occludens-1 (PDZ) domain binding motif in their cytoplasmic domains. It is likely that these molecules not only mediate the recognition events in synaptic specificity but also send a signal for inducing synapses, including presynaptic and/or postsynaptic differentiation. In addition to SDKs, a panel of synaptic adhesion molecules has been recently described in synapses. However, the combinatory roles of these molecules in synaptic specificity and formation remain still elusive.

The color-sensitive photoreceptors of the *Drosophila* visual system provide a genetically amenable, stereotyped, simple system in which to study layer-specific axon targeting (Figure 3(f)). The adult *Drosophila* visual system develops at the end of larval life. Each class of photoreceptor neuron (R1–R6, R7, and R8 cells) projects to a different synaptic layer in the brain. Several genes are known to contribute to the establishment of the layer-specific projection pattern of photoreceptors. Endogenous expression of the leucine-rich repeat cell adhesion molecule Capricious in R8 or its ectopic expression in R7 directs these photoreceptors to innervate the Capricious-positive layer. The interaction mediated by Capricious is homophilic, whereby the proteins interacting on the two opposing membranes are also homophilic, being reminiscent of SDKs in the vertebrate retina. In addition, the seven-pass transmembrane cadherin Flamingo is required for R8 targeting, whereas loss of either *Drosophila* neuronal (DN)-cadherin or the immunoglobulin superfamily (IgSF) receptor-type protein tyrosine phosphatases alters R7 to mistarget to the R8 layer. Thus, an array of these cell surface molecules and their ligands are candidates that directly drive the layer-specific assortment of distinct photoreceptor projections.

In addition to choice between correct and incorrect target neurons, lamina-specific connectivity involves subcellular specificity. One intriguing observation in the cerebellum clearly demonstrates a molecular mechanism that underlies this specificity. Basket neurons establish a distinct type of γ-aminobutyric acid (GABA)ergic synapse only on the axon initial segment of Purkinje neurons (Figure 3(b)). This subcellular specificity of synapse formation is mediated by a subcellular gradient of the IgSF molecule neurofascin 186 along the proximal–distal axis of Purkinje cells. This gradient is established by the cytoskelton protein ankyrin G, which binds neurofascin 186. Thus, the cytoskeleton-based recruitment of crucial adhesion molecules determines the subcellular or lamina-specific location of synapses.

Nonneuronal Cues

In addition to neurons, the nervous system contains numerous glial cells. Moreover, an extracellular matrix (ECM) produced by glial cells and/or neurons occupies the space between cells. Thus, the lamina-specific signals that affect axonal behavior can be associated with glial cells and the ECM (Figure 2(c)).

The importance of glial cells in lamina-specific innervations can be demonstrated by ablating glial cells. In the *Drosophila* visual system, when R cell axons enter the 'lamina' (note that 'lamina' is a nomenclature for the particular layer in this system), they encounter both glial cells and neuronal plexus (Figure 3(f)). R1–R6 axons stop in the flat lamina layer structure,' while R7 and R8 axons pass through this lamina layer and terminate in the medulla. The cellular requirement for R1–R6 axonal projection was determined using loss-of-function mutations affecting different cell types in the lamina layer. Deletion of glial cells but not neurons disrupts R1–R6 targeting. Thus, glial cells in this system provide the initial stop signal, promoting growth cone termination for R1–R6 axons in the lamina layer.

In the mammalian hippocampus, the commissural/associational (CA) fibers, which originate from the mossy cells, innervate the inner molecular layer. By contrast, entorhinal afferents, which originate from the entorhinal cortex, innervate the outer molecular

small-size sensory neurons specifically terminate in lamina I and lamina II, presumably by the repellant action of Sema3A. (e) Retina: the inner plexiform layer (IPL) of retina composed of an array of synapses formed between neurites of retinal ganglion cells and amacrine/bipolar cells. Homophilic sidekick adhesion molecules are expressed by specific subsets of retinal neurons which are making synaptic contacts in a subset of sublaminae. Overexpression of this adhesion molecule in a cell which does not normally express sidekicks diverts its neurite to sidekick-positive sublaminae. (f) Fly visual system: distinct types of photoreceptors project to distinct layers in either the lamina layer or the medulla layer. R1–R6 axons project to the lamina layer whereas R7 and R8 axons pass through the this layer and project to the medulla layer. R7 and R8 innervate the different layers of the medulla. (g) Hippocampus: entorhinal (red) and commissural/associational (CA) inputs (blue) terminate in the different laminae. In the hippocampus, the pattern of lamina-specific innervations is divergent, depending on the area. In the dentate gyrus, granule cells receive inputs on their different subdivision of dendrites, providing a good means to study subcellular specificity in lamina-specific neuronal connections. (h) Retinogeniculate system: axons from retinal ganglion cells in mammals terminate in different laminae of the lateral geniculate nucleus. Two inputs of the ipsilateral and contralateral eyes project to discrete laminae in adults as shown in this schematic. This eye-specific, lamina-specific connection is initially diffuse during development, and later is refined through a neural activity-dependent mechanism.

layer (Figure 3(g)). These lamina-specific axonal projections show a subcellular specificity of inputs: CA fibers synapse on proximal dendritic segments, whereas entorhinal fibers innervate only the distal dendritic segments of the same target neurons. In an adhesion assay using live hippocampal slices, dissociated entorhinal neurons selectively adhere to the outer molecular layer, raising the possibility that the laminar adhesive cue controls the laminar specificity. This lamina-specific adhesion was abolished by treating the slices with hyaluronidase, which degrades glycosaminoglycans. Furthermore, laminar specificity of entorhinal fiber ingrowth to the right target lamina was disturbed by treatment with *Streptomyces* hyaluronidase, which specifically degrades hyaluronan (hyaluronic acid). Interestingly, hyaluronidase treatment did not affect the laminar specificity of other afferent projections, such as the CA projection to the dentate inner molecular layer or the mossy fiber projection. Taken together, these data suggest that laminar specificity of entorhinal fibers that project to the correct target lamina requires hyaluronan or hyaluronan-associated ECM molecules. Major hyaluronan-associated molecules in the ECM are a lectican family of chondroitin sulfate proteoglycans such as aggrecan, neurocan, brevican, and versican. In this system, however, it is not known what types of neurons and glial cells express these proteoglycans. In the chick optic tectum, versican is expressed by lamina-specific neurons and deposited in a lamina-specific manner. In this system, versican appears to regulate maturation of synapses in the versican-positive laminae.

The other large ECM molecule, reelin, is expressed and secreted by Cajal–Retzius cells in the marginal zone of the hippocampus. Besides its role in neuronal migration, reelin has been shown to influence the branching pattern of entorhinal fibers projecting to the outer molecular layer. In the reeler mouse, entorhinal afferents are less branched than they are in wild-type mice, which suggests that reelin acts on entorhinal axons and influences their branching pattern. However, reelin is not required for the laminar specificity of entorhinal fibers *per se*, because in reeler mutants entorhinal fibers still innervate the correct target lamina. Instead, both ephrin/Eph and semaphorin/plexin systems play an important role in selection of this lamina-specific termination. Thus, this is an example of cooperation of multiple cues.

Guidepost Cells

Guidepost cells are specific cellular cues used by axonal growth cones in pathfinding and targeting choices (Figure 2(d)). For example, it has been experimentally demonstrated in insects that pioneer axons follow pathways marked by specific nonadjacent guidepost cells. In the developing mammalian cortex, pioneer neurons provide a transient scaffold for ingrowing axonal projections. A similar mechanism is also suggested for a synaptic target specificity in *C. elegans* which requires a heterophilic binding between two IgSF adhesion molecules, Syg-1 and Syg-2. Thus, it can be surmised that 'guidepost cells' are involved in formation of lamina-specific connections.

One example is seen in the hippocampus: transient Cajal–Retzius cells in the marginal zone were suggested to serve as intermediate targets for incoming entorhinal fibers. Transient synapses between entorhinal afferents and Cajal–Retzius cells were detected before the establishment of synaptic contacts between entorhinal fibers and their final targets, the granule cell dendrites. Furthermore, Cajal–Retzius cells were shown to make an early axonal projection to the entorhinal cortex, suggesting a role for these axons as a scaffold for the first incoming entorhinal axons that are seeking their correct target lamina. In fact, after ablation of Cajal–Retzius cells in organotypic slice cultures, entorhinal fibers no longer terminated in the target lamina, indicating that Cajal–Retzius cells play an indispensable role in development of this lamina-specific axonal connection.

Activity-Dependent Refinement

The last mechanism is based on a gamut of observations in which some precise lamina-specific connections arise from an inaccurate projection (Figure 2(e)). The most examined preparation is the retinogenicular projection. Retinal ganglion cells, particularly in higher mammals, send their axons to target neurons in the lateral geniculate nucleus in such a way that axons originating from the two eyes terminate in adjacent but nonoverlapping eye-specific laminae (Figure 3(h)). During development, initial projections from the ipsilateral and contralateral eyes are intermixed, and the adult precise pattern appears gradually, as axons from the two eyes mutually sort out into different laminae, making a lamina-specific termination. Several lines of evidence indicate that the segregation process, even though it occurs before vision develops in higher mammals, requires firing of ganglion cells, thereby neural activity plays a pivotal role in the lamina-specific segregation. The first compelling observation is that blocking ganglion cell action potentials with a voltage-dependent sodium-channel blocker inhibits the segregation of laminae. These action potentials are generated by the ganglion cells, which fire spontaneously and synchronously with each other, exhibiting waves that spread across the retina. A calcium imaging

study has shown that amacrine cells also participate in the correlated activity patterns, suggesting that a synaptic circuit within the developing retina is responsible for producing the retinal waves. Thus, it appears that, before the photoreceptors are functional, the spontaneous and synchronous activity produced in the retina is transmitted across central target synapses. Since spontaneously generated activity is present not only in the developing visual system but also widely in the developing nervous system, this process of activity-dependent neural circuit formation could be used to help elaborate other sets of lamina-specific neural connections.

Finally, even for activity-dependent refinement in the retinogeniculate system, recognition-molecule-mediated interaction also appears necessary. Class I major histocompatibility complex (MHC) molecules, known to be essential for immune recognition of antigen, are also expressed by neurons in an activity-dependent manner. In mice lacking class I MHC, lamina-specific connections in the retinogeniculate system are not perfect during development. These results demonstrate a potential role for class I MHC in the activity-dependent tuning of neuronal wiring.

See also: Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Pathfinding: Extracellular Matrix Role; Cell Adhesion Molecules at Synapses; Chemoaffinity Hypothesis: Development of Topographic Axonal Projections; Dendrite Development, Synapse Formation and Elimination; Spontaneous Patterned Activity in Developing Neural Circuits; Synapse Formation: Competition and the Role of Activity.

Further Reading

Benson DL, Colman DR, and Huntley GW (2001) Molecules, maps and synapse specificity. *Nature Reviews Neuroscience* 2: 899–909.

Bolz J, Uziel D, Mühlfriedel S, et al. (2004) Multiple roles of ephrins during the formation of thalamocortical projections: Maps and more. *Journal of Neurobiology* 59: 82–94.

Boulanger LM and Shatz CJ (2004) Immune signalling in neural development, synaptic plasticity and disease. *Nature Reviews Neuroscience* 5: 521–531.

Clandinin TR and Zipursky SL (2002) Making connections in the fly visual system. *Neuron* 35: 827–841.

Förster E, Zhao S, and Frotscher M (2006) Laminating the hippocampus. *Nature Reviews Neuroscience* 7: 259–267.

Huang ZJ (2006) Subcellular organization of GABAergic synapses: Role of ankyrins and L1 cell adhesion molecules. *Nature Neuroscience* 9: 163–166.

Katz LC and Shatz CJ (1996) Synaptic activity and the construction of cortical circuits. *Science* 274: 1133–1138.

Manzini MC, Ward MS, Zhang Q, et al. (2006) The stop signal revised: Immature cerebellar granule neurons in the external germinal layer arrest pontine mossy fiber growth. *Journal of Neuroscience* 26: 6040–6051.

Martínez A and Soriano E (2005) Functions of ephrin/Eph interactions in the development of the nervous system: Emphasis on the hippocampal system. *Brain Research Reviews* 49: 211–226.

Mumm JS, Godinho L, Morgan JL, et al. (2005) Laminar circuit formation in the vertebrate retina. *Progress in Brain Research* 147: 155–169.

Pielage J and Klämbt C (2001) Glial cells aid axonal target selection. *Trends in Neuroscience* 24: 432–433.

Raper JA (2000) Semaphorins and their receptors in vertebrates and invertebrates. *Current Opinion in Neurobiology* 10: 88–94.

Sanes JR and Yamagata M (1999) Formation of lamina-specific synaptic connections. *Current Opinion in Neurobiology* 9: 79–87.

Scheiffele P (2003) Cell-cell signaling during synapse formation in the CNS. *Annual Review of Neuroscience* 26: 485–508.

Shen K (2004) Molecular mechanisms of target specificity during synapse formation. *Current Opinion in Neurobiology* 14: 83–88.

Shinza-Kameda M, Takasu E, Sakurai K, et al. (2006) Regulation of layer-specific targeting by reciprocal expression of a cell adhesion molecule, capricious. *Neuron* 49: 205–213.

Skutella T and Nitsch R (2001) New molecules for hippocampal development. *Trends in Neuroscience* 24: 107–113.

Suto F, Tsuboi M, Kamiya H, et al. (2007) Interactions between plexin-A2, plexin-A4, and semaphorin 6A control lamina-restricted projection of hippocampal mossy fibers. *Neuron* 53: 535–547.

Tayler TD and Garrity PA (2003) Axon targeting in the *Drosophila* visual system. *Current Opinion in Neurobiology* 13: 90–95.

Uziel D, Garcez P, Lent R, et al. (2006) Connecting thalamus and cortex: the role of ephrins. *Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology* 288: 135–142.

Yamagata M and Sanes JR (1995) Lamina-specific cues guide outgrowth and arborization of retinal axons in the optic tectum. *Development* 121: 189–200.

Yamagata M, Weiner JA, and Sanes JR (2002) Sidekicks: Synaptic adhesion molecules that promote lamina-specific connectivity in the retina. *Cell* 110: 649–660.

Yamagata M, Sanes JR, and Weiner JA (2003) Synaptic adhesion molecules. *Current Opinion in Cell Biology* 15: 621–632.

Yamagata M, Weiner JA, Dulac C, et al. (2006) Labeled lines in the retinotectal system: Markers for retinorecipient sublaminae and the retinal ganglion cell subsets that innervate them. *Molecular Cell Neuroscience* 33: 296–310.

Yamamoto N (2002) Cellular and molecular basis for the formation of lamina-specific thalamocortical projections. *Neuroscience Research* 42: 167–173.

Landau Kleffner Syndrome

M A Mikati, R M Kurdi, and A N Shamseddine, American University of Beirut, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Definition and Epidemiology

Landau–Kleffner Syndrome (LKS), which was first described in 1957, is a childhood disorder manifesting two main symptoms: an acquired aphasia and a paroxysmal electroencephalogram (EEG) with spikes and spike slow waves that are often multifocal, most commonly seen in the temporal or temporo-parieto-occipital regions, and that are enhanced during sleep. This syndrome is also often associated with two other symptoms: behavioral problems and epileptic seizures. The International League Against Epilepsy (ILAE) defines this syndrome as "a childhood disorder in which an acquired aphasia, multifocal spikes, and spike and wave discharges are associated." Onset of LKS ranges from 3 to 8 years following a normal cognitive and language development. To date, the prevalence of LKS has not been determined. Eighty-one cases were reported between 1957 and 1980, and 117 cases between 1981 and 1991. It is worth noting that there are definitely many more cases than reported. During the past two decades, there has been increasing interest in LKS course, its pathogenesis, and treatment. Patry et al. in 1971 were the first to coin the term 'subclinical status epilepticus induced by sleep,' which was later referred to as 'electrical status epilepticus during slow-wave sleep' (ESES). Now it is known as epilepsy with 'continuous spikes and waves during slow-wave sleep' (CSWS). CSWS often occurs in LKS. The sex (male to female) ratio in LKS is 2:1. Family and medical history are usually, otherwise, negative, and there are usually no additional coexisting neurological symptoms other than those mentioned above (see Table 1).

Clinical Manifestations

Patients present with loss of expressive speech and auditory verbal agnosia. Behavioral and psychiatric disturbances are common. These include motor hyperactivity, impulsivity, and aggressive behavior. Anxiety can occur as a reaction to the sudden loss of understanding of spoken language that these patients experience. Language problems typically start as word deafness or auditory agnosia. Children do not respond to their parents' commands even with loud voices. The auditory agnosia may exacerbate to a stage that those children are unable to recognize even familiar sounds like bells or ringing phones. These patients are usually suspected to have hearing impairment. No abnormalities are found in audiograms or brain stem auditory evoked responses (BAERs). There are delays in long-latency cortical evoked responses which implicate posterior temporal regions' involvement. Permanent extinction of one ear contralateral to the involved temporal cortex is shown with dichotic listening tasks. Word deafness can extend to complete lack of response. Problems in expression such as frequent or even continuous misarticulations, use of telegraphic speech and flowing jargon, or even complete mutism often occur. The language problems of LKS resemble in many aspects those of the autism spectrum disorder. However, there are clear and well-defined differences. Autism patients suffer from problems in the development of spoken language and in the capability of starting a conversation. The language that the autistic child uses is stereotyped, repetitive, and characteristic. Confusion sometimes also arises from the fact that autistic patients can have seizures with frequent EEG discharges. In addition, at minimum, one-third of the autistic cases suffer from neurodevelopmental deterioration involving language, sociability, playing, and thinking skills. However, the language regression in autism occurs before 3 years of age, whereas the age of onset of LKS is typically 5–7 years. Autistic children suffer from early regression that involves losing single words, whereas children with LKS, who are older, suffer from more dramatic changes as they have acquired more vocabulary and language. Problems in reciprocal social relatedness and limited stereotypical forms of interests and behaviors that are associated with autism are not manifested in LKS patients. Patients who fulfill the criteria of LKS, but who also have the behavioral problems of autism, are known as patients with the Landau–Kleffner variant.

Epileptic Manifestations

Seventy percent of patients with LKS have seizures. One-third of them have one seizure or a single status epilepticus event, usually at the onset of the syndrome. The other patients usually have occasional seizures between 5 and 10 years of age. Sporadic seizures persist in one-fifth of the patients after the age of 10 years. Seizures rarely occur by the time patients reach the age of 15 years. Seizures in LKS are mostly nocturnal simple partial motor seizures. Complex-partial seizures occur infrequently. Atonic seizures are common. The occurrence of generalized

Table 1 Characteristic features of Landau–Kleffner syndrome, epilepsy with CSWS, and CSWS as an EEG finding

Characteristics	*LKS*	*ESES or epilepsy with CSWS*	*CSWS as an EEG finding*[a]
Age of onset	3–8 years	The first seizure occurs between the ages of 1 and 10 years, with a peak at 4–5 years	Usually in children
Symptoms	Auditory verbal agnosia and loss of expressive speech with behavioral psychiatric disturbances	Atypical absence seizures, seizures with falls, and sometimes absence status. Other seizure types commonly occur. Neuropsychological dysfunction as a rule is present	Depending on the epilepsy syndrome: LKS (in 80% of cases), ESES (in all cases), or BECTS (in occasional cases)
EEG features	Focal or multifocal epileptic discharges that occur most commonly in the temporal areas and that are consistently enhanced during sleep. EEG may also show CSWS (i.e., >85% of sleep dominated by continuous electrographic seizure activity) in 80% of cases.	In the waking state, focal and/or multifocal, and/or generalized diffuse spike wave activity. During sleep, the EEG shows continuous bilateral and diffuse slow spike wave activity persisting through all or most (>85%) the slow sleep stages (i.e., CSWS)	Spike wave activity constitutes more than 85% of slow-wave sleep time
Etiology	It is unknown in most cases (cryptogenic). Multiple etiologies: encephalitis, autoimmune mechanisms, and others	Posthemorhagic, postinfectious, hydrocephalus, focal cortical dysplasia, periventricular leukomalacia, perinatal occlusion of middle cerebral artery, arachoid cyst, congenital biopercular syndrome, and other etiologies	Depending on the epilepsy syndrome
Therapy	Anticonvulsants, corticosteroids, adrenocorticotropic hormone (ACTH), intravenous immunoglobulin, and surgical	Anticonvulsants, corticosteroids	Depending on the epilepsy syndrome
Prognosis	50% of the patients recover partially or suffer from permanent aphasia	In adulthood, 50% of patients suffer from speech abnormalities and behavioral problems	Depending on the epilepsy syndrome

[a]Can be seen in multiple epilepsy syndromes including BECTS, epilepsy with CSWS and LKS.

tonic-clonic seizures, atypical absences, and myoclonic-astatic seizures is noted less frequently. The prognosis is not dependent on the frequency and type of seizures. Seizures, but not aphasia, are usually controlled by the use of anticonvulsants.

EEG Findings

Focal or multifocal epileptic discharges commonly occur in the temporal areas. Paroxysmal activity is rarely precipitated by hyperventilation or photic stimulation, but is consistently enhanced during sleep, often leading to CSWS. Though common, the presence of CSWS is not essential to diagnose LKS as studies have shown that CSWS occurs in 80%, but not in all patients fulfilling the criteria of LKS. CSWS as an EEG finding is determined to be present when the spike wave activity constitutes more than 85% of slow-wave sleep time. The ILAE defines the epilepsy syndrome of "epilepsy with CSWS" as a condition that "results from the association of various seizure types, partial or generalized, occurring during sleep, and atypical absences when awake" and "despite the usually benign evolutions of seizures, prognosis is guarded because of the appearance of neuropsychologic disorders." The spike slow-wave activity is characterized by being bilateral and mainly generalized. In that respect, one should distinguish between epilepsy with CSWS as an epilepsy syndrome (as defined above) and CSWS as an EEG finding (also as defined above). The epilepsy syndrome of CSWS resembles LKS clinically, but the difference between them arises from the EEG (which may not show CSWS in LKS), from the general mental deficiency in epilepsy with CSWS, and the serious comprehension problems (auditory agnosia) in LKS. Also, the two syndromes can be differentiated by their clinical profile when patients reach adulthood, whereby patients with epilepsy with CSWS suffer from global cognitive abnormalities, and those with LKS have these deficits restricted to speech abnormalities and behavioral problems.

Etiology and Pathogenesis

It is probable that LKS is a syndrome of multiple etiologies which are yet to be fully elucidated. This

syndrome was initially assumed to be secondary to encephalitis because an initial report indicated the presence of encephalitic changes in a biopsy taken from the cortex of a case of LKS. Subsequent studies could not confirm the presence of encephalitic changes in pathological specimens taken from other patients. Autoimmune mechanisms have been reported to be potentially important, at least in some patients. Autoantibodies against each of the following – brain-derived neurotrophic factor, neuronal antigens, myelin, and brain endothelial cells – were found in the sera of some patients with LKS. Other possible etiologies have been implicated in different case reports. These include genetic predisposition, cerebral arteritis, toxoplasmosis, neurocysticercosis, temporal astrocytoma, temporal ganglioglioma, hemophilus influenzae meningitis, subacute sclerosing panencephalitis, inflammatory demyelinating disease, and abnormal zinc metabolism. The neurologic deficit in LKS is thought to be secondary to the frequent and at times continuous electroencephalographic discharges that inhibit the function of the speech centers. The pathophysiology of this electroencephalographic seizure activity is believed to be that thalamic-related neural generators of synchronous sleep-related EEG oscillations combine to promote electrographic seizure propagation during nonrapid eye movement sleep and drowsiness.

Laboratory and Radiological Studies

Cerebrospinal fluid (CSF), computed tomography (CT), and magnetic resonance imaging (MRI) usually show no abnormalities. In rare cases, there may be an elevation in CSF protein, changes in white matter, or structural lesions on CT or MRI scans. Occasionally, some enlargement or asymmetry of temporal horns may be present, and this is suspected to be due to the effects of long-term epileptic activity. Various studies using single photon emission computed tomography (SPECT) and positron emission tomography (PET) on LKS patients have demonstrated the presence of abnormalities in the temporal lobe in brain perfusion and in glucose metabolism. These include hypometabolism in middle temporal gyri on fluoro-deoxy glucose (FDG)-PET. Also, decreased temporoparietal perfusion was detected by SPECT. The exact relationship between LKS aphasia and hypometabolism or hypoperfusion of the temporal lobe is not known because these features are also seen in epileptic patients who do not manifest aphasia.

Diagnosis and Differential Diagnosis

Essentials for the diagnosis of LKS include (1) auditory agnosia with language regression and (2) epileptiform EEG abnormalities that exacerbate during sleep. The workup to define the underlying etiology, if one can be found, usually includes an MRI scan. On rare occasions, functional imaging, CSF studies, or arteriograms may be needed. These tests are, however, not essential for making the diagnosis of this syndrome. It is crucial to differentiate LKS and Landau–Kleffner variant from CSWS (see the earlier section 'EEG findings'), autism (see the earlier section 'Clinical manifestations'), and from benign rolandic epilepsy with centrotemporal spikes (BECTS) with exacerbation in sleep. In rare cases of BECTS, the strong activation during sleep of the centrotemporal EEG spike and wave activity results in cognitive decline and attention disorders. In the usual BECTS cases, the EEG shows normal background rhythms; spike foci can be unilateral or bilateral. It also shows typical morphology of high amplitude and diphasic waveforms. BECTS seizures are typically simple partial seizures that include speech arrest, excessive drooling, twitching of the face, stiffening of the tongue, or other sensomotor phenomena of the orofacial region and the regression is not verbal auditory agnosia. Their seizures can also, less frequently, be complex-partial, generalized, or partial with secondary generalization. Most patients keep a normal global intellectual efficiency and good long-term outcome. However, some, as mentioned above, may suffer from oromotor dysfunction, neuropsychological deficits, or attention deficits with learning disorders. This may be more common in BECTS with atypical features. Atypical BECTS features include leg jerking, unilateral body sensations, ictal blindness, epigastric pain, lateral body torsion, diurnal seizures only, status epilepticus, developmental delay, and attention-deficit disorder (ADD). Some patients can have additional neuropsychological problems such as cognitive dysfunction, auditory–verbal and visuospatial memory problems, and behavioral disturbances. BECTS can be aggravated by the administration of some antiepileptic drugs (AEDS) leading to the development of CSWS. Some patients presenting with typical BECTS may present later with atypical features. Moreover, a small number of patients with BECTS develop LKS years later. Landau–Kleffner variant patients manifest LKS features, but also have the behavioral problems of autism (see the earlier section 'Clinical manifestations'). These patients are managed predominantly as LKS patients and less as autism patients.

Therapy

General Principles

Therapy of LKS still has many controversial issues due to the complete lack of controlled clinical trials in

this syndrome. The therapeutic approaches that are used are based on open label data usually collected from case reports involving small numbers of patients. There are essentially two main lines of therapy. Pharmacological treatment which includes anticonvulsants, corticosteroids, adrenocorticotropic hormone (ACTH), or intravenous immunoglobulin (IVIG), and surgical treatment. Surgical therapy is still largely controversial. In addition, while treating LKS, one should treat all its aspects including seizures, speech, and behavior. Controlling seizures is usually easy. Sometimes, improvement in EEG is associated with language restoration. Due to the fluctuating profile of LKS and its unpredicted remissions, it is difficult to assess the effectiveness of any treatment.

Antiepileptic Drugs

The conventional AEDs are effective in controlling seizures of LKS patients; however, their effect on aphasia is not consistent. Valproate (10–50 mg $kg^{-1}day^{-1}$), clobazam (1–1.6 mg $kg^{-1}day^{-1}$), and ethosuximide (20 mg $kg^{-1}day^{-1}$) were reported to be efficient in stopping seizures and in reducing language problems in about half of the patients that received them. Valproate at a dose of 20 mg $kg^{-1}day^{-1}$ was reported to be effective in ameliorating the speech problems, especially in patients whose aphasia started earlier. High dosage of oral diazepam (0.75 mg $kg^{-1}day^{-1}$ with a blood level of 100–400 ng ml^{-1}) given for six short cycles (3–4 weeks) was successful in treating ESES patients including one with LKS. Riviello from Children's Hospital, Boston, reported administering 1 mg kg^{-1}per dose of rectal or oral diazepam (up to a maximum of 40 mg) for patients with LKS or ESES on two consecutive nights with continuous video EEG monitoring for 2 days and pulse oximeter monitors for 4 h after the diazepam dose. The patients were administered 0.5 mg $kg^{-1}day^{-1}$ (up to a maximum of 20 mg) for 3–4 weeks, then the dose was tapered (2.5 mg per month). Eleven out of 13 patients responded to treatment and they had their language and EEG improved within a month. The use of ethosuximide (20 mg $kg^{-1}day^{-1}$) combined with either valproic acid, phenobarbital, carbamazepine, or a benzodiazepine, and that of vigabatrin combined with ethosuximide, carbamazepine, or clonazepam were also reported to be effective in some LKS patients. Also, lamotrigine monotherapy improved cognitive functions and language in some patients. Nicardipine (0.5–2 mg $kg^{-1}day^{-1}$) combined with conventional AEDs was reported to be effective in LKS patients. Felbamate given as polytherapy or monotherapy (60 mg $kg^{-1}day^{-1}$) was able to dramatically improve language skills and EEG in some patients. Sulthiame and levetiracetam (35–60 mg $kg^{-1}day^{-1}$) were reported to be effective in helping some LKS patients. The 4:1 ketogenic diet (4 parts saturated fats, 1 part proteins and carbohydrates) was reported to be an effective treatment of acquired epileptic aphasia patients. It is not recommended to use carbamazepine for LKS patients since it was found to often worsen this condition.

Corticosteroids and Adrenocorticotrophin Hormone

Corticosteroids and ACTH (50–80 μg day^{-1} given for 4 weeks to 3 months) have been reported to be useful in improving language, cognitive, and behavioral problems. It is not known which regimen is the most effective. Prednisone is given orally (2–3 mg kg^{-1} for 1–2 months, then tapered according to response) or given at a dose that ranges from 30 to 60 mg day^{-1}. Dexamethasone is often given in a dose of 4 mg day^{-1} for 2 weeks before being tapered off. Intravenous methyl prednisolone given as 20 mg kg^{-1} daily for 3 days, then every 4 days, followed by oral prednisolone 2 mg kg^{-1} daily for 1 month or given as 500 mg over 3 h daily for 5 days, followed by 250 mg infusion over 2 h once per month was reported to be effective in a few patients resistant to other therapies. Treatment with steroids usually needs to be prolonged for several months or for more than a year to avoid relapses. This prolonged steroid therapy may cause several side effects such as avascular necrosis of hip, hypertension, behavioral abnormalities, gastric ulcers, hyperglycemia, and immunosuppression. These side effects can be minimized by using every other day dosing or weekend pulse dosing.

Intravenous Immunoglobulin

Occasional cases of LKS respond to IVIG; usually given as 2 g kg^{-1} for 4–5 days). The response to IVIG, in those that do respond (and those are a small minority), is usually quick and dramatic starting by day 3 of the infusion and resulting in remission of speech within a few days to a few weeks. However, this at times usually lasts for 3–4 months, after which a second or a third course is needed to result in similar responses and eventually in a permanent remission.

Surgery

The surgical procedure that has been reported to improve the profile of LKS patients that have failed medical therapy is multiple subpial transection (MST). This procedure involves sectioning the horizontal interneurons, which is hoped to lead to a decrease in epileptic discharges originating from the speech cortex while maintaining the area's physiologic functions. MST was reported to result in eradication of continuous spike waves during sleep and in

ameliorating the language functions in many of the above patients. However, MST is still viewed as an experimental procedure by many experts. In fact, it is usually required, before proceeding with MST, that the patient should fulfill several restrictive criteria. These include having normal developmental history of language and cognition before starting to lose language skills, normal nonverbal cognitive functions, unilateral intra- and perisylvian epileptogenic zone, and CSWS duration of less than 3 years. Vagal nerve stimulation (VNS) has recently been introduced as another type of surgery that could help LKS patients with refractory seizure. It was reported to decrease seizure frequency by half in three out of six children. Its effects on the aphasia still need to be determined.

Speech Therapy and Behavioral Intervention

Several interventions such as sign language, a daily diary of consecutive pictures of the classroom routine, auditory training starting with environmental sounds, and 'graphic conversation' technique have been found to be useful intervention tools. Even after LKS patients receive the appropriate treatment, they can still have problems related to processing oral language. Thus, even if they respond, there is still a need to use different methods to bypass the auditory channel of processing. Moreover, patients may need some behavioral modification techniques to deal with their behavioral problems. Medications to control hyperactivity and behavioral problems are also occasionally needed.

Evolution and Prognosis

Children with LKS may show complete recovery from aphasia months or years after the onset of LKS. Others, constituting around 50% of the patients, recover partially or suffer from permanent aphasia. Some cases are characterized by periods of fluctuating remissions and relapses. Prognosis is affected by many factors such as age of onset, type of language deficit, frequency and topography of EEG abnormalities, epilepsy duration, and effectiveness and side effects of AEDs. Outcomes can be as good as complete recovery and as bad as permanent aphasia. Most patients improve with some remaining moderate language problems. Moreover, one can expect better prognosis when the onset of language deficits takes place after the age of 6, and when speech therapy is started at an early stage. The relationship between the age of onset of language deficits and the end result is contrary to the situation of childhood aphasia after structural injury. The older the LKS patients, the greater their chance of language improvement. In aphasia that is associated with a lesion, the older the age at which the lesion appears, the lower the probability of language improvement. Improvement in EEG often precedes improvement in LKS-associated language disorders, but it does not guarantee the most favorable end point. In fact, some have found a strong relationship between the manifestation of language problems and EEG changes whereby improvement in both speech and EEG occurred at the same time, whereas others did not find such a relationship.

See also: Agnosia; Aphasia: Sudden and Progressive; Epilepsy; Epilepsy: Channelopathies; Language: Aphasia.

Further Reading

Hirsch E, Marescaux C, Maquet P, et al. (1990) Landau–Kleffner syndrome: A clinical and EEG study of five cases. *Epilepsia* 31: 756–767.

Landau WM and Kleffner RF (1957) Syndrome of acquired aphasia with convulsive disorder in children. *Neurology* 7(8): 523–530.

Mikati MA and Shamsuddine AN (2005) Management of Landau–Kleffner syndrome. *Paediatric Drugs* 7(6): 377–389.

Pearl PL, Carrazana EJ, and Holmes GL (2001) The Landau–Kleffner syndrome. *Epilepsy Currents* 1(2): 39–45.

Shouse MN, Scordato JC, and Farber PR (2004) Sleep and arousal mechanisms in experimental epilepsy: Epileptic components of NREM and antiepileptic components of REM sleep. *Mental Retardation and Developmental Disabilities Research Reviews* 10(2): 117–121.

Steriade M (2005) Sleep, epilepsy and thalamic reticular inhibitory neurons. *Trends in Neuroscience* 28(6): 317–324.

Language Development

T H Mintz, University of Southern California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The ability of human children to acquire a language is one of the hallmarks of the species. Within 24 h of being born, infants already show evidence of having learned aspects of the broad rhythmic structure of their mother tongue, most likely from hearing speech *in utero* during gestation. Infants continue to attend to the patterns in their language and start to learn a considerable amount about its structure well before they start combining words in their own speech. After just a few years, children will have mastered many of the complex grammatical structures in their language, and persistent systematic errors become rare as children approach their tenth birthday. Children exposed early on to additional languages will learn them without any significant added difficulty compared with monolingual acquisition, provided they have sufficient exposure to each language.

Language development in children is remarkable for its regularity across individuals and different languages, even when the languages are quite diverse. Children universally progress through many of the same stages, mastering certain components of the grammar before others and making characteristic errors in production and comprehension at each stage. Despite individual differences in intelligence and other abilities, and despite normal differences in language exposure, all children end up able to speak and understand their native language(s) fluently. This is in marked contrast to adult learners, who show much greater variability in learning processes and ultimate outcomes. Indeed, differences between children and adult learners in ultimate outcome, in the kinds of errors made, and in the stages of acquisition have led many researchers to conclude that there is a critical or sensitive period for language development. As in other developmental domains with critical periods, learning that takes place outside this period is more variable and less robust, leading to a greater range of outcomes and more idiosyncratic patterns of acquisition.

In addition to the uniform patterns of acquisition across languages, there are some aspects of language development that are tied more closely to structural properties of the particular language being learned. Comparing components that are more or less language dependent can lead to insights about the learning mechanisms involved in language development and the properties of the linguistic input to which learners attend.

This article gives an overview of early language development, discussing some key phenomena in acquisition and some major theoretical issues. Other articles offer more-detailed discussions of specific areas within language development.

Overview of Early Language Development

First Words

Linguistic structure When acquiring a vocabulary, learners associate word forms with concepts related to their meanings. In addition, they store information about how words combine with other words to makes sentences. The combinatorial patterns in language are governed primarily by the categories of words rather than by words themselves; thus learners must represent category information. While all languages have categories for nouns and verbs, languages vary as to what other categories are differentiated grammatically, such as adjectives, adverbs, determiners, and so forth. For example, many languages have classifiers – words or morphemes (see the section titled 'Morphology') that occur obligatorily and specify general properties of objects (e.g., animate, long thin object) – but English does not have a classifier system. The presence or absence of a grammatical class in a language should not be confused with the language's ability to express related concepts. For example, in English one can talk about attributes generally designated by classifiers in other languages, but English uses phrases involving adjectives and nouns instead of a unique class of markers.

Production Children produce their first words at approximately 1 year of age, although there is considerable variability in the onset of lexical production. These words are distinct from babbles in that they have referential content, although phonetically they may only approximate the target word form. Semantically as well, a word may only approximate the adult target. Thus, [dɔ] ('daw'), might be a 1-year-old's word for dog, but also for horses and other four-legged animals.

Children's productive vocabulary increases relatively slowly as the first 50–100 words are acquired. The initial cohort of words is primarily nouns. The early dominance of nouns appears to be a universal phenomenon and is attributed to the relative ease

of determining a noun's referent (often a concrete object) versus a verb's referent (complex relationships between objects, internal mental states, etc.).

Perception Although children generally do not produce their first words until they are approximately 1 year old, they segment the stream of continuous speech sounds into wordlike units much earlier. By 4.5 months, infants can recognize their own names embedded in fluent speech, and by approximately 7.5 months, their ability to segment words is much more fully developed. Thus, many months before producing their first words, infants become aware of the sound sequences that correspond to words in their language.

Morphology

Many words are made from smaller parts – sublexical meaning-bearing units called bound morphemes. Bound morphemes cannot exist alone as words but combine in systematic ways with stems to create words. For example, in English the morpheme *-ed* signifies past tense and affixes to the ends of verbal stems to produce past tense forms: *talk* + *-ed* → *talked*; the morpheme *–er* affixes to the end of verbal stems to produce related nouns: *talk* + *-er* → *talker*.

Although there is a relatively restricted range of functions or meanings associated with bound morphemes, individual languages differ as to what grammatical functions are marked morphologically. For example, many languages mark grammatical case – subject, object, possessor, and so forth – on nouns morphologically, but in English, case marking appears only on pronouns (e.g., *he* is in nominative case for subjects, *him* is in accusative case for objects), and case roles are determined structurally, using word order. Thus, in learning a language, children must learn what the bound morphemes are in their language, as well as the way in which these morphemes combine.

Production The timing of children's first productive uses of grammatical morphemes depends on the language being learned. For example, while a typical 18-month-old learning English is unlikely to systematically vary tense or agreement marking on verbs – such as always saying *walk* instead of *walked*, *walks*, or *will walk*, and so forth – a Turkish-learning child will use past tense morphology by as young as 15 months. Cross-linguistic differences in children's mastery of grammatical components is unlikely to arise from differences in children's conceptual development or general manner of conceiving of events and relations across language. Rather, many such differences can be explained by formal differences in the morphological systems. Turkish has a richer inflectional system that morphologically marks case and number on nouns and tense, mood, aspect, and agreement on verbs. Moreover, in Turkish each marker is a unique morpheme, unlike in English. For example, the English morpheme *-s* marks both third person and singular, whereas in Turkish, one morpheme marks third person and another marks singular. Overall, then, there is a greater prevalence of bound morphemes in Turkish than in English. In addition, inflections in Turkish are phonetically more salient than in English: they are syllabic, containing unreduced vowels, and usually receiving stress. Finally, the near-perfect one-to-one mapping of morpheme to grammatical function in Turkish makes the morphological system less ambiguous than in languages like English (where, for example, *-s* marks both third-person singular on verbs and plural on nouns), which could make learning faster. Thus, language-specific factors such as the regularity of form, the uniqueness of form-to-meaning mappings, and the acoustic/phonetic salience of morphemes all influence the time course of morphological development.

Perception Before children use morphology productively, they must identify the phonological forms that correspond to bound morphemes in their language. As discussed above, in languages in which grammatical morphemes are acoustically salient, regular, and unambiguous, detection of the forms and mastery of the form–meaning mappings proceed more rapidly than in languages that do not have these properties. Nevertheless, perception of morphological forms and dependencies of forms occurs well before the forms are used productively. In English, 18-month-old infants have been shown to be aware of the dependency between the auxiliary verb *is* and the main-verb ending,*-ing*, in present progressive sentences: they differentiate grammatical sentences such as *The baker is baking bread* from ungrammatical sentences such as *The baker can baking bread*, in which *can* and *-ing* are incompatible.

First Word Combinations

During the period of early vocabulary growth, children's utterances consist almost entirely of isolated words. Children between 18 and 24 months of age start to combine words, producing both one- and two-word utterances. During that time, children's vocabulary growth rate increases as well, with an increase in the ratio of verbs to nouns compared with prior periods. Early word combinations often combine nouns with verbs, as well as nouns with

other nouns and sometimes adjectives. These multiword combinations show consistent ordering patterns that generally reflect the basic word order patterns of the exposure language. For example, the basic word order pattern in English transitive sentences is subject–verb–object (SVO). At the stage when English-learning children combine at maximum two words per utterance, they sequence verbs before nouns to express action–patient, action–theme, action–goal, and action–location relations, and they sequence nouns before verbs to express agent–action relations. For example, a child in this stage might produce 'kick ball' to indicate an action performed on a ball (action–patient), and 'baby kick' to refer to an agent performing an action. Noun–noun combinations also respect the basic word-order relations, so that in the example above, expressing both agent and patient would be accomplished by the utterance 'baby ball', in which the correct ordering relation between the two arguments is maintained but the verb is missing. The last example points to potential ambiguity in interpreting early child productions as there are also situations in which the same sequence of words ('baby ball') might express a possession relationship, that is, 'baby's ball,' without the possessive morpheme. Many analyses of children's speech thus consider the context of an utterance in determining its meaning and underlying structure. The researcher Lois Bloom pioneered this method of analysis, coined 'rich interpretation,' in the early 1970s. After carefully analyzing the situational contexts under which children's early utterances are produced, Bloom argued that children's two-word utterances express the same kinds of basic semantic and syntactic relations as adult sentences express – possessor, thing possessed, subject, direct object, indirect object, and so forth. While there was (and still is) disagreement about the particular categories and grammatical structure underlying children's early utterances, many researchers followed Bloom's lead in using context to infer structuring principles that would not be evident from analyzing the sequence of words alone.

Even languages that place few syntactic constraints on word order (marking subject, object, and direct object by case marking on nouns) have canonical word-order patterns that are governed by pragmatic and discourse constraints. For example, Turkish allows all possible orders of subject, verb, and object in transitive sentences; however, there is a statistical predominance of subject–object–verb (SOV) sequences, with SVO sequences being the second most frequent (and, indeed, Turkish is usually considered an SOV language, the most common type cross-linguistically). Children learning Turkish reflect these patterns in their own multiword utterances. Thus, word-order patterns are one of the first structural properties that young children attend to and master. Their representation of these patterns includes knowledge of how structural positions correspond to semantic roles, such as agent, patient, theme, goal, and location.

Gradual Development of Grammatical Constructions Involving Phrases

Most grammatical formalisms include categories for word sequences, or phrases, as well as for individual words, and many grammatical operations involve phrase-level constituents. For example, *the red dog* is a noun phrase (NP) in English, and it takes its category type from the phrase head (the noun, *dog*). Grammatical operations, such as question formation, operate on elements at this level: for example, the declarative *John sees the red dog* becomes *Who does John see?* in the interrogative, with the entire NP replaced by *who* (which is subsequently moved).

While phrases are an important structuring component in language, they are abstract entities, in that phrases – the boundaries between phrases, the category of phrases, and the hierarchical organization of phrases – are not explicitly marked in speech. Thus, part of what children must learn in acquiring a language is the internal organization of phrases and the sequencing of phrases within sentences. Children must also learn how grammatical operations manipulate phrases. In some cases, mastery develops in stages and suggests processes of reorganization in language development. Several examples are given in the next sections.

Question Formation

English-learning children go through several distinct stages in forming interrogative sentences, producing ungrammatical questions early on. For example, they produce sentence-initial *wh*-words without required auxiliaries: "what doing?" or "what you doing?" Next, they produce the auxiliary verbs but without the required inversion with the subject: "what you are doing?" Finally, they perform the required inversion of subject and auxiliary verb: "what are you doing?" Although there is individual variation across children in this general pattern, it is striking how predictable the development of question formation in English is. The stable pattern of errors has been used to argue for highly constrained mechanisms guiding learning, but the facts are also consistent with theories of language development that posit more general mechanisms that are sensitive to the statistical properties of the input. The acquisition of questions is also an example of the incremental development of linguistic capabilities.

In Chinese and Japanese (among other languages), forming questions does not involve moving words or inserting auxiliary verbs like *do*. Children learning such languages consistently produce well-formed questions before English-learners do. Comparing the course of development for similar syntactic functions across typologically different languages can give insights into the nature of the mechanisms and representations involved in acquisition because it provides evidence about the ease or difficulty with which different surface features are mastered to achieve similar linguistic functions.

Verb Complement Clauses

A phrase can have another phrase embedded within it. Examples are sentences with sentential complements, for example, *Mary saw (that) [Sam was sad]*, in which the embedded verb is tensed, and *Mary wanted [Sam to be happy]*, in which the embedded verb is infinitival. The function words *that* and *to* serve to mark embedded clauses, and their consistent use by children would be suggestive evidence that they have the adult complement structures in place. However, early productions of tensed complements occur without *that*, for example, *I know he sits here*. Although this appears to be a grammatical complement construction, some accounts of these early constructions propose that they involve two hierarchically adjacent phrases rather than one phrase embedded within a larger sentence. The structure of the last example would be similar to *He sits here, and I know this*. On that account, early utterances like those do not involve embedded clauses.

Similarly, children initially produce utterances such as "I want open it" or "I gonna get it" without the obligatory infinitival marker, *to*. The marker emerges gradually, first restricted only to a small set of matrix verbs (e.g., ... *want to* ...). But even in those constructions, it is rare for subjects to intervene between the main verb and *to*, with the result that the subject of the embedded clause is identical to the subject of the main clause (e.g., *I want (to) open it* vs. *I want Sam (to) open it*). These restricted aspects have led some researchers to claim that these utterances do not contain embedded complement clauses at all; they analyze the main verb as a modal verb that modifies the second verb (which would be analyzed as the main verb). In this view, *I want open it* is parallel structurally to *I can open it*. These accounts of early complement-like constructions claim that early utterances do not, in fact, involve embedded clauses.

In summary, mastery of complement phrases appears to be gradual and depends, in part, on the identity of the main verb. Even when children correctly produce and comprehend most complement constructions, some still cause difficulty. In particular, some main verbs are special, in that the embedded subject is not the subject of the embedded clause. Compare *Sam wanted Mary to set the table* with *Sam promised Mary to set the table*. In the first sentence, the embedded subject is the subject of the embedded complement, and this is the normal situation. However, in the second sentence, the matrix subject *Sam* is also the subject of the embedded clause. Children routinely interpret such sentences as though *Mary* is setting the table, and these errors can persist until about age 9.

Relative Clauses

Relative clauses are another kind of embedded clause that adds complexity to basic sentences, here within NPs. For example, the material after *that* in *the man that I know*, is a sentence embedded within an NP; the relativizers (*that*, *who*, etc.) are often optional. As with complement clauses, development of these structures in production is gradual. At first, children omit relativizers and generally modify only the last noun in a sentence, resulting in the modifying material occurring at the end of the utterance, which modifies the sentential object: "I see the car has the stripes." Later, they also produce relative clauses in subject position and are more likely to include *that* or other relative pronouns (e.g., *who*): "Man that I saw is leaving." Thus, as with complement clauses, the development of relative clauses in production is incremental.

Overregularization

There are regularities in the way a given language carries out grammatical functions. For example, in English, the past tense is formed by suffixing *-ed* onto a verb stem: *walk* → *walked*. In addition, there are exceptions to the regular patterns: the past tense of *go* is not *goed* but the irregular *went*. A universal phenomenon in language development is the overregularization of irregulars by children. When children initially produce an irregular, they may do so correctly, but as they start to learn more and more regular forms, and learn the generalized pattern, they sometimes apply the pattern to irregular forms – either to the stem (e.g., *goed*) or the irregular form itself (e.g., *wented*). These overgeneralization errors can persist in individuals for several years.

The phenomenon of overregularization has played an important role in the theorizing about language acquisition. Patterns of overregularization have been used as evidence in support of a variety

of different architectures for lexical memory and related storage and access mechanisms, from general pattern associators that treat all stems and derivations equally, to specialized mechanisms that explicitly distinguish regular and irregular forms.

Nativism and Domain Specificity

One of the core questions in language acquisition research is the nature of the mechanisms and representations that make learning a first language possible. Central issues are nativism – the degree to which acquisition relies on innate mechanisms and representations – and domain specificity – the degree to which learning mechanisms are specific to language development. Although these issues are often conflated, they are two distinct theoretical claims.

It is clear that human language involves, to some degree, innate or biologically determined components that are present in humans but absent in other species. Inherent differences in cognitive architecture across species, in part, make human language qualitatively different from the natural communication systems of other species. (Clearly there are nonneurological biological differences between species that are relevant for language learning and use. Spoken language as we know it is not possible without a vocal tract capable of producing the necessary range of phonetic material, and a species without hands or the appropriate visual system would be unable to communicate using a signed language. But the crucial formal characteristics of human language are modality independent and in principle could be implemented in a variety of physical systems. It is these properties that arguably have an innate component in the cognitive architecture of humans.)

These innate components contribute to the highly systematic and uniform pattern of development cross-linguistically and to the inability of nonhuman species to achieve the same degree of mastery. What is generally debated is the content of the innate components: is there specialized functional machinery that is dedicated to the acquisition and/or use of language, or does language have the structure it does, and do children learn language the way they do, because of more-general properties of human cognition and/or human social interaction?

One of the most influential theories about the innate component was proposed in the 1960s by Noam Chomsky. Chomsky proposed that humans possess an innate, universal grammar (UG). In its standard conception, UG specifies the structural properties that all human languages share (and thus that putatively do not have to be learned), for example, that words are organized into phrases and phrases are organized into sentences, the kind of syntactic operations that words in certain phrasal positions can undergo, and so forth. UG also contains specialized mechanisms for learning the nonuniversal idiosyncrasies of the language to which the child is exposed (e.g., whether question formation involves inversion of the subject and auxiliary verb). The motivation for the theory is often referred to as the 'poverty of the stimulus' (POS) argument. The core of the argument states that the data from which children learn language – the utterances that they hear – are too variable and too underspecified to guarantee that all learners would converge on the correct grammar without specialized constraints and linguistic knowledge guiding learning. UG provided a description of these constraints.

Within this general theoretical framework, different theories were proposed for learning mechanisms that could determine the language-specific properties that were specified in (but also highly constrained by) UG. The content of UG and the associated learning mechanisms were described in the vocabulary of formal linguistic theory and were traditionally viewed as specific to language. But in a broader sense, UG could be construed as a set of constraints on possible inferences that can be made from linguistic input. Viewed in this way, it becomes reasonable to investigate alternative sources or descriptions of these constraints. For example, some constraints on learning and generalization could emerge from constraints in substrate learning components such as working memory, sequential processing, and so forth, and constraints governing their interaction. Some universal constraints on linguistic structure could arise from social and cultural phenomena, as opposed to specific cognitive structures for language. The computational problem of linking linguistic forms to meanings, sequencing these forms when producing utterances, and comprehending these sequences when listening to speech could lead to specific kinds of structures, in order to be carried out in real time by a human brain.

Nevertheless, many constraints on linguistic structure seem difficult to motivate from these other sources. For example, from sentence (2a), question (2b) is possible, but (2c) is ungrammatical. What could account for this difference?

(2a) Sam knows a man who likes wine.
(2b) Who does Sam know who likes wine?
(2c) What does Sam know a man who likes?

Grammatical theories account for this difference by positing constraints on grammatical operations and defined on the basis of syntactic structures; in other words, domain-specific constraints. It is not immediately obvious how the alternative sources of

constraints just discussed could account for such phenomena, but a number of research programs are currently addressing just these questions.

Recovery from Errors and Negative Evidence

Part of the POS argument is the observation that a learner who was not adhering to the right structural constraints could posit grammars that overgenerate sentences. That is, the grammars would account for the utterances the child heard, but also ungrammatical utterances. This is not, in principle, an insurmountable problem. Having such a grammar, a child would be expected to make errors and produce ungrammatical sentences. But if the child received a corrective signal that the utterance was ill-formed – negative evidence – then the learner could update the grammar to rule out the ungrammatical utterance. This alternative has two essential components: (1) there must be reliable sources of feedback available to children when they make errors, and (2) the feedback should change the pattern of errors a child makes. If either of these points is not satisfied, then an alternative to a highly constrained learning mechanism becomes more difficult to maintain.

In fact, there is considerable evidence that none of these conditions are satisfied by the empirical data. A number of possible sources of negative evidence have been proposed, primarily involving the types of replies adults give in response to children's ungrammatical utterances (e.g., repeating the utterance with errors repaired, expanding the utterance, etc.), but caretakers vary widely in the kinds of responses they give to children's ungrammatical utterances, and so far no type of feedback has been shown to be reliable enough to be universally informative. Moreover, even in cases when children's errors are explicitly corrected, children very rarely alter their behavior, and the errors persist. When errors do cease, they do so without corrective feedback from adults.

Taken together, the reliability across children in the kinds of errors they make and their recovery from errors in the absence of negative evidence points to learning mechanisms that follow an internally constrained program. However, these facts are neutral as to the nature of the constraints and whether they are special to language or derive from more-general principles and properties of learning.

Cues to Structure

An important aspect of the POS argument is that utterances do not come marked with their structural description. For example, words are not inherently marked as noun, verb, adjective, and so forth, and phrasal constituency in utterances is not explicitly marked. If structure were transparent in the input, then the nature of the learning problem would be very different, as learners could simply 'read off' the structure from the utterance.

Researchers are examining the degree to which cues that correlate with important structural distinctions are available to infants and children acquiring a language. In many languages, there are bound morphemes that correlate strongly with lexical category (noun, verb, adjective, etc.). For example, case markers occur on nouns but not on verbs, and tense markers appear on verbs but not on nouns. Even though the correspondence between overt marker and lexical category may be many-to-one (e.g., because there are multiple nominal cases), as opposed to one-to-one (the simplest kind of cue), morphological markers could nevertheless be extremely useful cues to category membership.

However, not all languages have robust marking; as discussed earlier, languages like English have case marking only on pronouns, and verbal morphology is relatively impoverished. In addition, in some languages, morphemes that occur with different lexical categories are homophonous, and therefore ambiguous cues; in English, for example, *-s* is a plural marker on nouns (e.g., *cats*) and a third-person-singular agreement marker on verbs (e.g., *grows*). Therefore, while bound morphemes might provide useful category cues in some languages (e.g., Turkish), their usefulness might be limited in others. However, other cues, such as the co-occurrence patterns of words and phonological cues, have been shown to be informative for determining lexical categories.

Studies have also investigated whether prosodic information – information involving pitch, syllable duration, pause duration between words, and amplitude – could cue major phrase boundaries, such as boundaries separating subject NPs from the verb phrase. There is cross-linguistic evidence that combinations of cues involved in prosody are probabilistic indicators of phrase boundaries and could thus be used by children to aid in assigning structure to an utterance.

Another cue to phrase-level constituency that is available in some languages is inflectional marking, in particular, case marking within NPs. In German, for example, case is marked on articles and adjectives (both within an NP), and the case morpheme often (although not always) takes the same form on the article and adjective. For example, in the masculine declension, the definite article is *der* in the nominative case, and the corresponding ending on a following adjective would be *-er*; however, in the accusative case, the definite article is *den*, and the corresponding case marker on the adjective is *-en*. The phonological

similarity of endings within a phrase (the article and adjective) could be used by learners as a cue to identify phrasal constituency. Even when within-constituent endings are not as similar phonologically, the endings covary within a phrasal constituent to a significant degree, and the patterns of dependency within phrases could be a more general cue to phrasal constituents.

Investigations into possible cues to structure are uncovering a variety of probabilistic cues inherent in the structure of utterances and in statistical patterns across utterances that learners could use to recover some aspects of the structural description of an utterance. To learn successfully using these cues, learners must attend to the relevant perceptual dimensions and process their input in specific ways. Experiments with infants and young children demonstrate that learners are indeed sensitive to many of the requisite properties and perform the necessary types of analysis on the basis of those properties. These lines of research are helping to specify the nature of the innate mechanisms involved in language learning. In some cases, the sensitivities and constraints are not clearly specific to language but rather could be viewed as more-general constraints on sequential processing, auditory perception, and categorization.

See also: Bilingualism; Connectionist Models of Language Processing; Dementia and Language; Language Evolution; Language in Aged Persons; Language: Auditory Processes; Language: Nonhuman Animals; Language: Cortical Processes; Language: Aphasia; Reading; Second Language Acquisition; Sentence Production; Sentence Comprehension; Statistical Learning of Language; Word Learning.

Further Reading

Bates E and MacWhinney B (1982) Functionalist approaches to grammar. In: Wanner E and Gleitman LR (eds.) *Language Acquisition: The State of the Art*. Cambridge: Cambridge University Press.

Bloom L (1970) *Language Development: Form and Function in Emerging Grammar*. Cambridge, MA: MIT Press.

Crain S (1991) Language acquisition in the absence of experience. *Behavioral and Brain Sciences* 14: 597–650.

Elman JL, Bates EA, Johnson MH, Karmiloff-Smith A, Parisi D, and Plunkett K (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Jackendoff R (2002) *Foundations of Language*. New York: Oxford University Press.

Marcus GF (1993) Negative evidence in language-acquisition. *Cognition* 46(1): 53–85.

McClelland JL and Patterson K (2002) Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences* 6(11): 465–472.

Newport EL, Bavelier D, and Neville HJ (2001) Critical thinking about critical periods: Perspectives on a critical period for language acquisition. In: Dupoux E and Mehler J (eds.) *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*. Cambridge, MA: MIT Press.

Pinker S and Ullman MT (2002) The past and future of the past tense. *Trends in Cognitive Sciences* 6(11): 456–463.

Slobin DI (ed.) (1985) *The Cross-Linguistic Study of Language Acquisition, Vol. 1: The Data*. Hillsdale, NJ: LEA.

Language Evolution

M H Christiansen, Cornell University, Ithaca, NY, USA; and Santa Fe Institute, Santa Fe, NM, USA
S Kirby, University of Edinburgh, Edinburgh, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Language is often referred to as one of the hallmarks of our species. However, the emergence of language may be more than simply an important event in human evolution; it may be one of only eight 'major transitions' in the evolution of life, as listed by the two eminent evolutionary biologists John Maynard Smith and Eörs Szathmáry. Each major transition denotes a key event in the history of our planet, signaling radical changes in the way evolution works, starting with a change in the way molecules replicate in the very earliest stages of the origins of life, through the emergence of DNA, to larger scale subsequent phenomena such as the evolution of colonies where once there were only solitary individuals and including, as the most recent evolutionary transition, the emergence of language (Figure 1).

Why is the emergence of language such a significant event? What does it have in common with the other major evolutionary transitions? One of Maynard Smith and Szathmáry's interesting observations is that, despite their diversity, these transitions have some features in common. In particular, many of the transitions give rise to a new mechanism for the transmission of information. Language, they argue, provides just such a novel mechanism – essentially enabling a system of cultural transmission with unlimited heredity. As such, language enables the transmission and storage of very complex cultural information. But how did we, as human beings, evolve this powerful ability? Before discussing the possible answers to this question, we briefly survey the structural features of language that support this powerful information transmission system.

The Structure of Language

One way of thinking about language (although by no means the only way) is as a coding system that maps between two spaces: the space of concepts and intentions, on the one hand, and the space of articulation and perception, on the other hand (Figure 2). Traditionally, the study of the structure of language has been divided into a number of subdisciplines, each of which concerns a different aspect of this mapping system:

- *Phonetics*. The production and perception of sounds/manual gestures.
- *Phonology*. The systematic behavior of the sounds of language.
- *Morphosyntax*. The system for combining the basic meaningful units of language into words and sentences.
- *Semantics*. The meaning of words and sentences in isolation.
- *Pragmatics*. The system for relating word/sentence meaning to communicative intention in the context of communication.

The first and last two subdisciplines on this list deal mainly with the two ends of the mapping in Figure 2, whereas morphosyntax is most clearly the study of the aspects of language that govern how these two are connected, using words recruited from a mental lexicon.

What is extraordinary about this system, and what makes it particularly important for Maynard Smith and Szathmáry, is that it is constructed in such a way as to allow unbounded yet faithful transmission of information (sometimes termed 'digital infinity'). This combination of an infinite range of messages with a high-fidelity mechanism for transmitting those messages is almost unique in nature. Arguably, the only other example is the genetic code.

It is easy to understand why human language is in principle unbounded. If we were to try to find the longest sentence of English, we would fail. This is because the syntactic system delivers mechanisms that will allow us to elaborate on sentences in an unlimited fashion (e.g., by adding subordinate clauses, adverbial phrases, and prepositional phrases). This kind of infinity is 'digital' because it does not rely on continuous changes in the signal to convey changes in meaning but, rather, the addition of discrete elements. In contrast, we could imagine a different signaling system in which the pitch of a signal conveyed differences in meaning (e.g., the severity of a particular threat). This system would be infinite because there are infinitely many different pitches, but it would not be digital.

Another unusual aspect of human language is that the lexicon is flexible. New words can be added, and the meanings of words can change. Although this feature of language is not discussed as much as digital infinity, it is actually the combination of these two that really sets human language apart as a uniquely powerful tool for the unbounded transmission of cultural information. In summary, language structure allows high-fidelity, unbounded, and flexible communication.

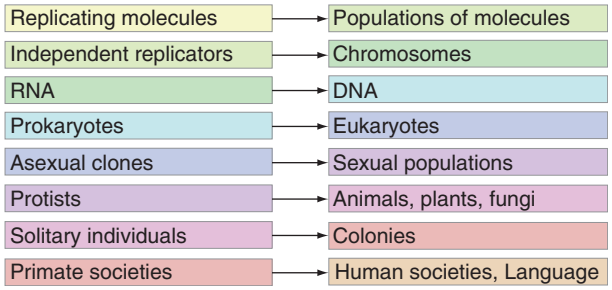

Figure 1 Maynard Smith and Szathmáry's eight major transitions in the evolution of life.

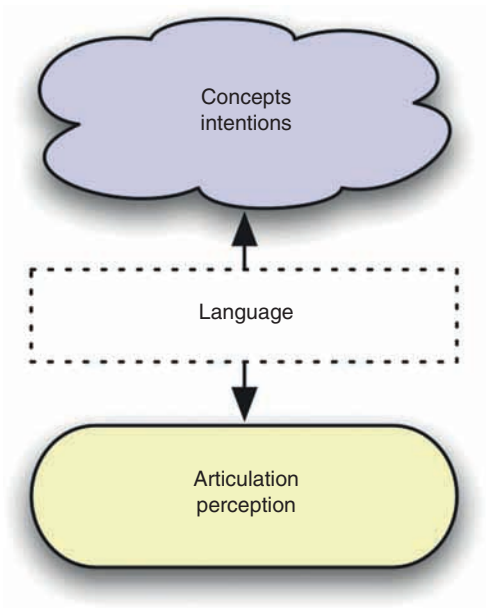

Figure 2 Language can be viewed as a system that maps between two different domains: concepts and intentions, on the one hand, and mechanisms for articulation and perception of speech (or gesture), on the other hand.

Questions surrounding the origins and evolution of language have, since the early 1990s, seen a major increase in interest in the scientific community, across a very wide range of disciplines. In the remainder of this article, we discuss major areas of consensus and current controversies. It is important to realize, however, that this is far from an exhaustive summary of a subject that draws on evidence from archeology to computer science, from genetics to philosophy.

Major Points of Consensus

Understanding language evolution poses many challenges for contemporary science, and the picture that is emerging is highly complex. Nonetheless, considerable progress has been made in the field of language evolution since it emerged as a legitimate area of scientific enquiry during the 1990s. Here, we first provide an overview of the major areas of consensus before discussing some of the current controversies.

Possible Preadaptations for Language

There is a general consensus that to understand language evolution, we need a good understanding of what language is. However, the field is divided over what the exact characterization of language should be, and in what terms it should be defined. Nonetheless, agreement appears to be in sight regarding some of the necessary steps toward language. Specifically, there seems to be agreement that prior to the emergence of language, some preadaptations occurred in the hominid lineage. There is less agreement about what these may have been, but one candidate that has been proposed by many is the ability to use symbols. In this context, symbol use is typically construed as a capacity for linking sounds or gestures arbitrarily to specific concepts and/or percepts – particularly for the purpose of communication. In addition, it has been suggested that the ability to relate these symbols to each other was a further necessary preadaptation for language. Although there is evidence that nonhuman primates have some capacity, albeit limited, for using sequences of arbitrary symbols in captivity, there is considerable debate regarding whether they use these symbols to refer things in nature. For example, the use of manual gestures for symbolic communication in the wild has been called into question. Thus, the use of complex sequences of symbols referring to objects and situations may be a uniquely human ability.

Several other possible candidates for language preadaptations have been put forward, of which we mention a few relating to changes in social or cognitive abilities. Joint attention – that is, the capacity to follow eye-gaze direction or direct the attention of another to a specific object – is important for successful communication, and it may have been a social precondition for language. Another potential social preadaptation for language is the capability of modern humans for sophisticated imitation of action sequences for the purpose of communication. Our ability to represent others as intentional beings with their own beliefs and desires, which can be manipulated by our actions, may also be a social prerequisite for language. At the cognitive level, an increase in the capacity for representing complex concepts and combinations thereof may have predated the emergence of language. Additional cognitive preadaptations that may have paved the way for language include the ability for complex hierarchical learning of sequentially presented information and increases in memory for sound sequences, both of which are important for

learning and processing of language. It should be noted, however, that many of the preadaptations mentioned previously are shared with other species, particularly other primates, and that differences in these skills may be more quantitative in nature than qualitative.

The Necessity of Interdisciplinary Collaborations

Perhaps the strongest point of consensus among language evolution researchers is a methodological one: to fully understand language evolution, it must be approached simultaneously from many disciplines. We must understand how our brains work; how language is structured and what it is used for; how early language and modern language differ from each other and from other communication systems; in what ways the biology of hominids has changed; how we manage to acquire language during development; and how learning, culture, and evolution interact.

Thus, language evolution research must necessarily be cross-disciplinary in order to provide sufficient constraints on theorizing to make it a legitimate scientific enquiry. Nevertheless, most researchers in language evolution only cover parts of the relevant data, perhaps for the reason that it is nearly impossible to be a specialist in all the relevant fields. Still, as a whole, the field appears to be moving in the direction of becoming more interdisciplinary. Collaborations between researchers in different fields with a stake in language evolution are likely to become increasingly more important.

The Importance of Modeling

Another emergent area of consensus is the growing interest in using computational modeling to explore issues relevant for understanding the origin and evolution of language. Many researchers across a variety of different disciplines now either conduct language evolution simulations or refer to such work as evidence for particular theoretical perspectives. For example, modeling work has been used to inform high-level theories about biological adaptations for grammar or the emergence of language structure through cultural transmission but also at a more detailed level, such as the evolution of phonetic gesture systems or a neural basis for grasping as a precondition language based on manual gesture. Models are useful because they allow researchers to test particular theories about the mechanisms underlying the evolution of language. Given the number of different factors that may potentially influence language evolution, our intuitions about their complex interactions are often limited. It is exactly in these circumstances, when multiple processes have to be considered together, that modeling becomes a useful, and perhaps even necessary, tool.

The role of computational modeling in language evolution research can be divided into three rough categories.

1. Evaluation: Computational models, like mathematical models, have the virtue that they enforce explicitness in the formulation of an explanation. As such, they can act as a rigorous check that a particular explanandum actually does follow from a particular explanans. In other words, they can help researchers to identify hidden problems. In some sense, they allow us to create novel experiments to test under which conditions language evolves.
2. Exploration: Computational simulations can be used (with caution) to explore the general ways in which explanatory mechanisms or theoretical constructs interact. In this mode, simulations can help direct us to new theories.
3. Exemplification: Computational simulations can be a valuable tool for demonstrating how an explanation works. They can augment verbal and mathematical theorizing to provide working models for pedagogical purposes.

Computational modeling thus provides a powerful new tool for the study of language evolution. However, it cannot stand on its own. It must take its place alongside theoretical considerations, mathematical modeling, experimentation, and data collection (e.g., linguistic and archaeological). For example, some computational models may eventually lead to mathematical models, or vice versa. Computational models may suggest novel psychological experiments and so on. We envisage that the interest in computational modeling is likely to increase, especially as it becomes more sophisticated in terms of both psychological mechanisms and linguistic complexity.

The Usefulness of Comparative Approaches

An additional type of evidence that is becoming increasingly important is data from studies that directly compare the learning and processing abilities of nonhuman primates with those of humans (either adults or children) using the same experimental paradigms. For example, comparisons of 8-month-old human infants and cotton-top tamarins on a simple artificial language learning task using the same preferential head-turn methodology indicate that both species may have similar abilities for basic statistical learning. Such work may allow us to better determine which components of language may be unique to humans and which may be shared with other species. As a case in point, comparative evidence regarding sequential learning suggests that an important

difference between human and nonhuman primates is our superior ability for learning and processing hierarchically organized temporal sequences. When combined with further corroborating evidence from neuropsychology and neurophysiology, computational simulations, and linguistic considerations, this human ability becomes a compelling candidate for a possible hominid biological adaptation that may eventually have led to the evolution of complex language. Future comparative research may reveal additional differences that can inform our understanding of language evolution.

Insights from Genetics

Advances in genetics are now figuring prominently in discussions of language evolution. As we obtain a better understanding of the genetic bases of language and cognition, as well as their interaction with the environment during development, genetic information promises to provide new constraints on language evolution theories, particularly with respect to issues related to the origin of language. However, the relationship between language and genes is extremely complex, and the relationship between genes, language, and evolution perhaps even more so. Consequently, current evidence provides few constraints on evolutionary theorizing. For example, data regarding the *FOXP2* gene have been cited in support for very different theories of language evolution, ranging from a gesture-based perspective to a speech-based perspective, from accounts involving large endowments of innate linguistic knowledge to accounts eschewing such innate knowledge. Nonetheless, there seems to be agreement that the FOXP2 data may suggest a late evolution of speech. In this way, the genetic data may be particularly useful in informing our understanding of the timeline for language evolution.

Current Controversies

Although much progress has been made and several areas of consensus have emerged, a number of major points of disagreement remain, of which we highlight the more current ones here.

Biological versus Cultural Evolution

Although there is considerable agreement about a possible symbolic preadaptation among our hominid ancestors prior to the emergence of language, opinions differ considerably about the subsequent evolution of grammatical structure.

One line of theorizing suggests that grammatical structure is a consequence of an evolved innate grammar. There are several different proposals regarding how and why a biological adaptation for grammar may have evolved in the hominid lineage by way of natural selection. One suggestion is that language evolved gradually as an innate specialization to code increasingly complex propositional information (e.g., who did what to whom, when, where, and why). This may have been for the purpose of social information gathering and exchange within a distinct 'cognitive niche' or for a kind of social 'grooming' at a distance in groups of hominids too large for establishing social cohesion through physical grooming. Another perspective suggests that grammar emerged more rapidly with the speciation event that produced modern humans approximately 120 000 years ago. Common to most of these proposals is the suggestion that language syntax shows evidence of complex design – similar to, for example, our visual system – and that biological adaptation is the only way to explain the appearance of such design.

A different line of theorizing views grammatical structure not as a product of biological adaptation but, rather, as emerging through cultural transmission of language across hundreds (or perhaps thousands) of generations of learners. Essentially, in this view language is an evolutionary system in its own right – but one driven by cultural rather than genetic evolution. This has interesting implications for Maynard Smith and Szathmáry's view of evolutionary transitions. They focus on language as a mechanism through which the transmission of cultural information is possible. The cultural approach to language evolution highlights the additional fact that knowledge of language itself is passed on culturally through the process of language learning by children.

Aspects of linguistic structure might therefore be viewed as adaptations by language to the cultural evolutionary 'problem' of being transmitted faithfully from generation to generation of language learners. In this approach, language is seen as adapting to the narrow transmission bottleneck imposed by children's learning mechanisms. Evidence in favor of this perspective has come from computer modeling of cultural transmission, the development of indigenous sign languages, and the archeological record of artifacts. Many proponents of this viewpoint on language evolution argue for a 'culture-first' perspective in which language evolved only after basic competences for relatively complex social culture had emerged in the hominid lineage. However, additional constraints would seem to be needed if the appearance of design in language is to be explained. Such constraints may be found in the limitations on our ability for sequential learning of hierarchical structure, in the learning bottleneck created by forcing languages through the limited channel of children's

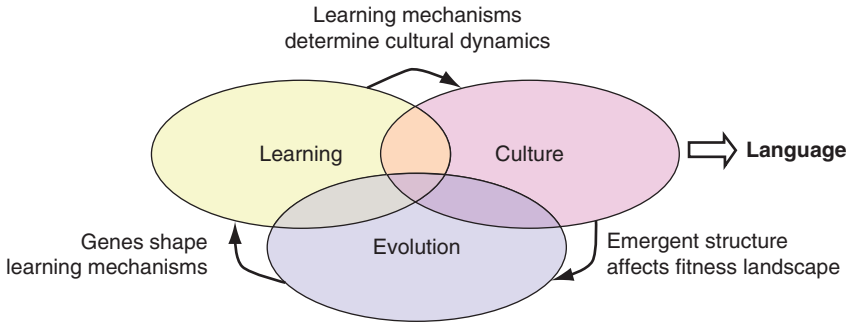

Figure 3 The complex adaptive systems view of language. Language arises from the interactions of three dynamical adaptive systems that operate on differing timescales: Biological evolution through natural selection shapes our mechanisms for learning language. Language emerges from the interaction of many individual learners through a process of cultural transmission and evolution. The structure of the language that emerges will partly determine the biological fitness of the possessors of that language, thus completing the cycle of interactions.

learning mechanisms, or in the complexities of our conceptual apparatus. Alone or in combination, these constraints have been suggested to explain the elements of language that give the appearance of design, such as linguistic universals.

It is likely, however, that biological adaptation and cultural transmission may have interacted in the evolution of language. Language arises from the interaction of three different adaptive systems: individual learning during development, cultural transmission across generations of learners, and biological evolution of the learners (Figure 3). Our understanding of such interaction is complicated by the fact that the three adaptive systems interact on three different timescales: the lifetime of the individual (tens of years), the language (thousands of years), and the species (hundreds of thousands of years). Determining the exact weighting of these three components with respect to each other and the nature of their contribution is thus an important issue for future research in the evolution of language.

Vocal versus Manual Origin of Language

Another strongly debated issue in language evolution research is whether language originated as a system of manual gestures or evolved exclusively in the vocal domain. On the one hand, it has been proposed that because vocal communication in primates is largely affective in nature, with little voluntary control, language is likely to have emerged from manual gestures rather than primate calls. In some versions of this account, the emergence of gestural language was predated by the evolution of a unique human ability for complex imitation. The subsequent change from a gestural to a primarily vocal language has been argued to be due to either increased tool use coming into conflict with the use of the hands for linguistic gestures or the 'recruitment' of vocalization through associations between gesture and sound. The close relationship between manual gesture and a subsequently evolved sophisticated ability for vocalization (in the form of speech) is furthermore suggested to have left us with the uniquely human characteristic of right-handedness.

One the other hand, critics of the gestural theory of language origin have argued that manual gestures suffer from two major disadvantages in comparison with spoken language: they require direct line of sight and they cannot be used at night. Instead, several proposals have been put forward to support the possible origin of language in the vocal domain. One suggestion is that the basic structures of syllables derive from the succession of constrictions and openings of the mouth involved in chewing, sucking, and swallowing – eventually evolving into phonetic gestures. It has also been contended that this evolutionary process may subsequently have resulted in the major syntactic distinctions between noun phrases and sentences. An alternative perspective suggests that natural selection for brain structures necessary for the motor activities involved in walking on two legs may have laid the groundwork for the evolution of the neural substrate necessary for speech production and perception, which in turn provided the basis for the emergence of syntax.

Although mathematical and computational modeling may help inform the discussions about the relationship between biological adaptation and cultural transmission in language evolution, such modeling is less likely to be able to address issues related to the origins of language. However, evidence from other disciplines, such as archeology, comparative neuroanatomy, primatology, psycholinguistics, and cognitive neuroscience, may provide clues to an answer, although it is unclear whether this debate can ever be settled completely.

The Nature of Protolanguage

Setting aside the possibility of sharing some or many of the systems underpinning language with other species, it is clear that there is a huge gulf in complexity between the language spoken by any human and the communication used by our nearest primate relatives. For many researchers, this gulf is a problem for any evolutionary account of language. It leads us naturally to wonder whether it is possible to have something that was less than a full human language but more than an animal communication system – an evolutionary 'protolanguage.'

Broadly speaking, there are two different views of what the nature of protolanguage may have been. One approach suggests that protolanguage was 'compositional,' involving simple combinations of meaningful words but with little or no syntactic structure. Alternatively, it has been proposed that protolanguage was 'holistic' in nature, with unanalyzed wholes expressing complex meanings that we today would express using multiword combinations. In both cases, however, it has been suggested that there exist a number of 'living fossils' of protolanguage that we can observe today.

Evidence in favor of compositional protolanguage comes from pidgin communication (the communication system sometimes formed in communities of adults that lack a shared language), child language at particular stages of development, and the language of trained apes. Analyses of these diverse forms of behavior suggest the possibility of an evolutionarily prior form of language that may have shared some features of fully modern language but lacked others. In particular, whereas all three use individual meaningful words, their combination to form utterance meanings involves little or no syntactic structure (e.g., there is no recursive embedding, obligatory propositional structure, or purely structural grammatical elements).

Support for holistic protolanguage can be found in the form of formulaic language, which consists of utterances or parts of utterances in normal language use that appear to be processed completely holistically as unanalyzed chunks rather than syntactic combinations of individual meaningful words. These formulas include idiomatic expressions, such as 'bought the farm' (whose meaning, 'died,' appears to be completely unrelated to its form); certain standard situational utterances, such as 'was there anything else'; and other expressions, such as 'by and large,' whose semantic opacity suggests they are holistically processed. (It is important to note that although the examples of formulas appear to be strings of words, this is simply because they exist within a language that is already syntactic.) Such formulaic use appears to be rampant in normal language use and may thus be seen as a potential key element of protolanguage.

Further differences between these two perspectives on protolanguage derive from different ideas about how full human language may have evolved from its proto-form. Whereas the compositional approach suggests that the change was a biological one, the holistic approach suggests it was driven by cultural transmission. The plausibility of the latter claim has been given some support by computational models that have demonstrated the spontaneous emergence of syntactic structure in populations initially speaking a holistic protolanguage.

Independent of the outcome of this debate, the protolanguage evidence demonstrates the logical possibility of a system of communication that is like language but with some components missing, and it furthermore points to real instances of such a system in the world that are at least partially functional. Theoretically, this points to a possible smooth adaptive pathway from an alingual state to a linguistic one. Moreover, the notion of protolanguage is interesting because it opens up the possibility of 'major transitions' within the evolutionary trajectory of language. It may thus be useful to return to Maynard Smith and Szathmáry's work and seek parallels with other evolutionary transitions to help illuminate the debate.

Conclusion

Since the 1990s, research on the evolution of language has seen rapid growth. This increased research activity is warranted given that an understanding of language evolution may shed much light on human evolutionary history. However, if Maynard Smith and Szathmáry are correct, then language evolution research may also provide important insights into the evolution of life. Thus, the evolution of language constitutes an important challenge for contemporary science. In this article, we provided a brief survey of some of the work being undertaken to answer this challenge, focusing on current trends and controversies. It is conceivable that there will be questions to which we may never find definitive answers. However, it is clear that we will only be able to address the many questions surrounding the evolution of language by taking into account the various systems that underlie it, making this scientific endeavor an interdisciplinary enterprise by necessity.

See also: Language Development; Language: Auditory Processes; Language: Nonhuman Animals; Language: Cortical Processes; Primate Communication: Evolution; Psycholinguistics; Speech Production: Development; Speech Perception: Cortical Processing.

Further Reading

Bickerton D (1995) *Language and Human Behavior.* Seattle: University of Washington Press.

Cangelosi A and Parisi D (eds.) (2002) *Simulating the Evolution of Language.* London: Springer.

Cangelosi A, Smith AD, and Smith K (eds.) (2006) *Proceedings of the Sixth International Conference on the Evolution of Language.* London: World Scientific.

Christiansen MH and Kirby S (eds.) (2003) *Language Evolution.* Oxford, UK: Oxford University Press.

Christiansen MH and Kirby S (2003) Language evolution: Consensus and controversies. *Trends in Cognitive Sciences* 7: 300–307

Deacon TW (1997) *The Symbolic Species: The Co-Evolution of Language and the Brain.* New York: Norton.

Fisher SE and Marcus GF (2006) The eloquent ape: Genes, brains and the evolution of language. *Nature Reviews Genetics* 7: 9–20.

Givón T and Malle BF (eds.) (2002) *The Evolution of Language out of Pre-Language.* Amsterdam: Benjamins.

Hauser MD (1996) *The Evolution of Communication.* Cambridge, MA: MIT Press.

Hurford J, Studdert-Kennedy M, and Knight C (eds.) (1998) *Approaches to the Evolution of Language: Social and Cognitive Bases.* Cambridge, UK: Cambridge University Press.

Kirby S (2007) The evolution of language. In: Durbar R and Barrett L (eds.) *Oxford Handbook of Evolutionary Psychology,* pp. 669–681. New York: Oxford University Press.

Knight C, Hurford J, and Studdert-Kennedy M (eds.) (2000) *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form.* Cambridge, UK: Cambridge University Press.

Maynard Smith J and Szathmáry E (1997) *The Major Transitions in Evolution.* New York: Oxford University Press.

Pinker S and Bloom P (1990) Natural language and natural selection. *Behavioral and Brain Sciences* 13: 707–784.

Tallerman M (ed.) (2005) *Language Origins: Perspectives on Evolution.* Oxford: Oxford University Press.

Wray A (ed.) (2002) *The Transition to Language.* Oxford: Oxford University Press.

Language Following Congenital Disorders (not SLI)

R Leech, University of London, London, UK
F Dick, University of London, London, UK; and University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Congenital developmental disorders can result in a wide range of language outcomes. At one extreme, children with some disorders never master complex language use. At the other extreme, after an initial period of delay, some children may achieve a language outcome indistinguishable from that of typically developing peers. Congenital disorders affecting language also vary considerably in terms of the underlying etiology; indeed, etiology of a congenital disorder is one of the most important determinants of the course of language development and eventual outcome. However, the relationship between the degree and location of frank neurological abnormality and the degree of corresponding language deficit is anything but straightforward.

In this article, we contrast the effects of two general classes of congenital disorder on language development: 'neurodevelopmental' disorders, such as Williams and Down syndromes and autistic spectrum disorder, and 'acquired' disorders, such as focal brain lesions (FLs) acquired in the pre- or perinatal period. The neurological sequelae of neurodevelopmental disorders often tend to be subtle, relatively widespread, and heterogeneous, at least on a gross anatomical level. For instance, people with autistic spectrum disorders (ASD) tend to have reduced volume in the cerebellar vermis but increased volume of the cerebral and possibly cerebellar white and gray matter. People with Williams syndrome (WS) appear to have particular cortical abnormalities in occipital and parietal regions but relatively preserved cortical structure in the superior temporal lobe, whereas people with Down syndrome (DS) appear to have dysmorphology in the medial temporal lobe. In addition, a number of postmortem studies suggest that ASD, WS, and DS may be associated with aberrant neuronal morphology, packing density, and columnar architecture, albeit in somewhat different forms. In contrast, children with perinatal FLs are assumed to have a typical neurodevelopmental trajectory and cerebral architecture except for the lesion immediately affected by the infarct (with potential longer-term quantitative changes in connecting white matter tracts and homologous gray matter regions).

A comparison of these two types of developmental disorders allows for an examination of the relationship between neural development, cerebral functional differentiation, and language acquisition across the first two decades of life. Importantly, the relative strengths and weaknesses that characterize the language profiles tend to change from infancy through toddlerhood and into late childhood, and they differ according to the etiology of the language disorder. In this article, we first trace the development (and possible neural and social underpinnings) of the linguistic strengths and weaknesses associated with three neurodevelopmental disorders (ASD, WS, and DS). We then discuss language acquisition in children with FLs, focusing not only on eventual outcome but also on the transient relationships between lesion side and site, and linguistic and nonlinguistic abilities over development.

Neurodevelopmental Disorders

As noted previously, the changes to brain structure and function that result from neurodevelopmental disorders tend to begin very early in development and to become more pervasive and more enigmatic. For instance, unlike FLs, it is not possible to diagnose developmental disorders from brain imaging because of high interindividual variability in the neural manifestation of the disorders. Children with neurodevelopmental disorders all tend to show atypical performance in many different social and cognitive skills including language. However, what makes the study of developmental disorders particularly informative is that different disorders have divergent profiles of strengths and weaknesses, not only in language processing but also in aspects of social and cognitive development that appear to be necessary for the early stages of language development. We highlight some of these early appearing social and cognitive deficits, and link them to possible linguistic deficits further 'downstream' in childhood. The following sections are roughly chronological, highlighting the diverging developmental paths in WS, DS, and ASD first in infancy and toddlerhood and later in school-age years and adolescence.

Language Learning in Infancy and Toddlerhood

Language is heavily dependent on social cognition in order to start developing. Early social skills, such as the ability of the infant to share attention with caregivers, are thought to play a substantial role in developing language and nonlinguistic communicative skills. Even in the first months of life, important social precursors to language already distinguish

between infants with neurodevelopmental disorders and typically developing (TD) infants. For example, TD infants show a preference for their mother's speech, whereas infants with ASD show no such preference and seem to have very little interest in people and social interaction. These infants show poor levels of eye contact and gaze following, and they may even attempt to avoid the gaze of a caregiver. This early lack of social motivation and attention sharing, and later lack of social conventions such as turn-taking, provides a plausible explanation for the inability of approximately half of children with ASD to develop any kind of complex language. There is some speculation that low-level attentional switching deficits in ASD may contribute to the higher-level social deficits noted here, and that these may be linked to the cerebellar abnormalities characteristic of this syndrome.

In contrast, infants with WS are typified by their strong desire for social interaction and keen interest in faces. However, this preference also leads to poor gaze-following behavior because infants with WS prefer to look at the face of the caregiver as opposed to following the direction of gaze. Such atypical use of referential information by children with WS may in part explain their delays, relative to TD children, in early language comprehension and production. Infants with DS also enjoy interacting socially, although there is a delay in mutual eye contact, which may have an impact on subsequent dyadic interactions between infant and caregiver. Moreover, once these initial problems resolve, infants with DS seem to prefer to focus on the eyes as opposed to visually exploring the facial features of the caregiver, as TD infants do. Deviations in early interactive behavior can also be seen in more complicated triadic interactions between infant, caregiver, and object. For example, infants with WS struggle to switch their attention from the caregiver to an object. This problem with triadic interactions is thought to hinder the acquisition of conceptual knowledge of objects, contributing to a subsequent delay in vocabulary development. In general, ASD, WS, and DS are all characterized to varying degrees by some form of deficit in social skills, which are very important for structuring the early word and grammar learning of infants and young children.

Precursors of Expressive Language in Developmental Disorders

As for social cognition, infants with neurodevelopmental disorders diverge from TD infants in terms of their prelinguistic 'productive' abilities, such as babbling and gesturing. As before, the exact patterns and the mechanisms underlying these delays tend to vary with the type of disorder. For instance, for infants with DS, there is an average delay in the onset of babbling of 2 months, with subsequent delays in phonological development. These may be due to the articulatory difficulties, which frequently occur in conjunction with the disorder (because of differences in palate shape and tongue position), as well as to DS children's frequent problems with hearing loss due to chronic and severe otitis media (probably due to deformations of the eustachian tube).

Unlike their profile with babbling, young children with DS produce far more communicative gestures than those seen in TD, a behavior that may serve as a compensatory mechanism for delayed and error-prone phonological production. In contrast, young children with ASD and WS gesture very little. Use of referential pointing in WS does not appear until after the naming explosion, demonstrating that children with WS do not follow the typical language development trajectory and may be using atypical mechanisms for early referential language use.

Despite the differences between DS and WS, children with DS and WS show a similar delay in the onset of babbling. Furthermore, children with WS show early problems with speech segmentation, which is surprising given the relatively strong phonological abilities of older children with WS. This highlights the fact that the pattern of deficits and relative strengths in the adult does not imply that these same deficits and strengths are present earlier in development.

Word Comprehension and Production and Early Grammatical Development

Just as prelinguistic expressive skills in developmental disorders are delayed, so is the onset of first word comprehension and production. Both WS and DS children are severely delayed in word production, obtaining a vocabulary size of approximately 50 words approximately 2 years later than a TD child, who would produce this many words at approximately 16 months. After reaching this critical vocabulary size, the language profiles of DS and WS diverge. TD children demonstrate an almost lawful relationship between the size of their vocabularies and their productive grammatical complexity, implying that a certain level of vocabulary is necessary for complex language to get started. Similarly, children with WS follow the same relationship between vocabulary size and grammatical complexity as TD children (Figure 1), albeit one that is considerably delayed (i.e., young children

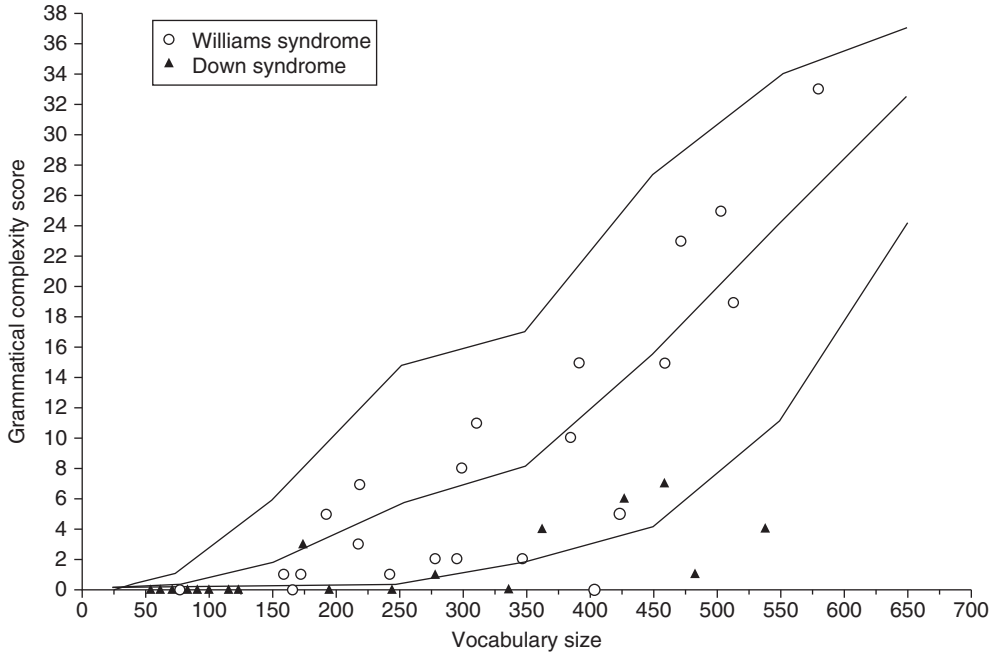

Figure 1 Vocabulary size vs. grammatical complexity. Children with WS show the same developmental trajectory as TD children (albeit one that is markedly delayed), whereas children with DS diverge from the typical trajectory, with less grammatical complexity thanwould be expected given their vocabulary size. The lines mark the 50th, 10th, and 90th percentiles for TD children. Redrawn from Bates and Goodman (1997).

with WS demonstrate similar complex language abilities as younger TD children with the same vocabulary level). In contrast, children with DS show an atypical relationship, with their grammatical complexity worse than would be predicted by their vocabulary size (Figure 1). For children with DS, syntactic development is slow and does not generally result in these children consistently producing and understanding more complex forms. However, despite the differences in the rate of grammatical development between disorders, it has been suggested that, in general, children with developmental disorders acquire grammatical knowledge in approximately the same sequence as TD children.

Perhaps not surprisingly, there are profound differences in the ways that children with these neurodevelopmental disorders use language to communicate. For example, both children with DS and children with WS generally enjoy the opportunity to interact within a social environment, and their use of language is motivated through social exchanges in conversation. By contrast, language in many children with ASD, even when fluent and social, appears to lack a sense of reciprocity in conversational norms. Such a language profile is often linked to a significant discrepancy between verbal and nonverbal IQ. Overall, early communicative styles differ as substantially between disorders as do their social and grammatical abilities.

Later Language Learning

The profiles of strengths and weaknesses seen in different developmental disorders are not constant throughout the child's life. Instead, the relative strengths and weaknesses of early and late language abilities can be quite different as the multiple interacting linguistic and social factors underlying language acquisition unfold.

Williams Syndrome

The main feature of later language development in WS is delay. Despite this, relative to mental age-matched (MA) controls, school-age and adolescent children with WS may develop good semantic knowledge, an extensive vocabulary, and complex syntax, although their vocabulary skills generally exceed their syntactic ability relative to MA controls. In this regard, when narrating a picture book story, children with WS show a very similar profile of slowed morphosyntactic development as children with a seemingly very different developmental disorder, language impairment (LI). However, by age 10–12 years, children with WS (unlike those with LI) produce the same proportion of complex syntactic structures as typically developing controls, although there are subtle differences from TD children in their discourse structure. For instance, children with WS fail to establish ties between the protagonist's goals

and his or her actions. Finally, as might be expected given their hypersocial nature, children with WS far surpass even typically developing children in their use of socially engaging cues when telling a story. Thus, part of the language profile of children with WS might be accounted for by nonlinguistic characteristics of the disorder, such as a strong desire for social interaction.

Down Syndrome

In general, older children with DS show the same sequence of developmental 'events' as do TD children but are generally very delayed. For example, children with DS usually produce the same pattern of phonological errors produced by younger TD children. However, when WS and DS children are compared across early and later childhood, there is a divergence of DS and WS developmental trajectories in different language domains. For instance, in early language development, children with WS and DS have a similar level of vocabulary, but by adolescence the vocabulary of children with WS exceeds that of those with DS. Thus, the performance of children with DS appears to asymptote at a lower level.

Autistic Spectrum Disorders

In contrast to DS, children with ASD exhibit severe pragmatic difficulties in comparison to MA controls, although there is major individual variability in language outcome. Whereas children with ASD will use language to manipulate their environment, their use of language to engage in conversation or express thoughts and emotional states is often limited. It has been suggested that autistic children may actually use their language knowledge to bootstrap their understanding of others' mental states. In terms of pragmatic ability, children with ASD struggle to understand figurative speech, such as 'It's raining cats and dogs', taking literal meanings from such sentences. Other characteristics of autistic speech include a flat pattern of intonation and echolalia (in which children repeat the verbal utterances of others). These may occur when the child is required to respond to a question or in the form of preconstructed speech during conversation.

Neurodevelopmental Disorders versus Acquired Disorders

Whereas the different neurodevelopmental disorders vary considerably in the profiles of relative strengths and weaknesses of language skills, almost all affected children fail to achieve the functional language outcome of TD children. This contrasts strikingly with children with pre- and perinatal FLs who generally fall within the normal range of language use despite the severity of their neurological damage. However, as with all such developmental disorders, a 'normal' language outcome may mask previous language delays and atypical trajectories in different aspects of language development. Thus, in the following sections we consider first the general language outcome of older children with FLs before discussing the more subtle lesion site-specific language problems observed during infancy and early childhood.

Acquired Language Disorders: Learning Language after Early Brain Injury

Pre- and perinatal focal lesions are typically the result of vascular insult (e.g., congenital arteriovenous malformation or pre- or perinatal stroke) and are exceedingly rare events, observed in approximately 1 or 2 of every 10 000 births. Of particular neuroscientific and clinical interest are cases of children with unilateral focal lesions because they generally occur without concomitant global neurological malformation (e.g., cortical dysplasias) and thus allow direct comparisons with results from adult neuropsychological patients with similar lesions.

Comparisons of children and adults with FLs have provided a unique perspective on the fundamental questions of neural development, plasticity, and domain specificity, particularly how these play out in the establishment of language function in the brain. Although the inherent variability in lesion site and size makes children with FLs far from a homogeneous group, results from 30 years of research paint a fairly consistent picture. In particular, the impressive but delayed language outcomes of children with congenital focal lesions suggest that many of the neural biases needed for language to start developing are very coarse indeed.

Lesion Location and Language Outcome

Based on the adult model, the language outcome of children with FLs should depend to a large degree on whether the injury is located in the left or the right hemisphere because left hemisphere injury in adults frequently leads to severe and lasting language problems. However, the overall picture from the literature on children with early focal lesions tends not to show consistent differences between early left and right hemisphere in terms of language processing. When sample sizes are large, and direct statistical comparisons between lesion groups are made, (i.e., right hemisphere damage (RHD) vs. left hemisphere damage (LHD), controlling for seizure history

and lesion onset time), lesion side and site generally do not predict language proficiency. If differences are observed, they are noted only for a few language subscales and are considerably smaller than those observed in adults with commensurate injuries.

For instance, language studies on adult patients with focal lesions reveal a classic double dissociation, with RHD patients showing impaired comprehension of familiar phrases and idioms (e.g., 'It's like talking to a brick wall') with relatively preserved comprehension of novel literal phrases (e.g., 'She's looking down at her black cat'), whereas LHD patients show the opposite pattern of deficits. In stark contrast, not only is there no difference between children with LHD and children with RHD on either measure but also both groups perform within the low-normal range for their age-matched controls, whereas their adult aphasic counterparts are massively impaired relative to healthy age-matched controls (Figure 2). This striking difference has been observed across a range of language tasks, such as narrating brief biographical sketches and online complex syntactic comprehension. As would be expected, adults with LHD are significantly more impaired than their RHD counterparts. However, comparisons between children with RHD and those with LHD reveal no differences. Indeed, in narration tasks, children with FLs differ very little from typically developing children, performing in the normal to low-normal range on a variety of measures. The only significant (but small) delays observed in children with FLs involved the production of somewhat fewer words and more omission errors than age-matched controls. Similarly, the syntactic comprehension of children with FLs to either hemisphere appears to be somewhat delayed but not deviant from the language trajectory of typically developing children.

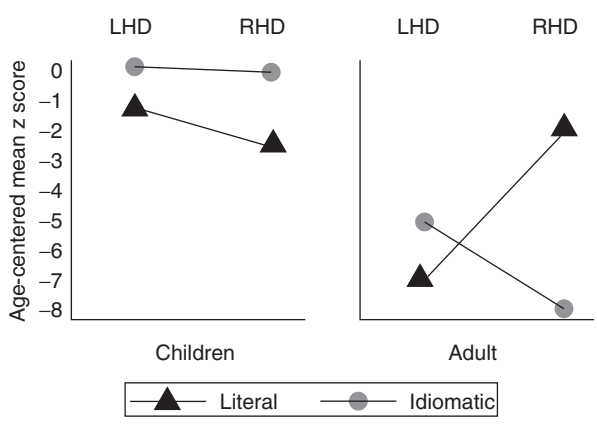

Figure 2 Accuracy in interpreting novel literal and idiomatic sentences, contrasting children and adults with left hemisphere damage (LHD) vs. right hemisphere damage (RHD). Redrawn from Kempler et al. (1999).

Lesion Location and Language Acquisition in Infants

Although there are few, if any, apparent differences between language outcomes for children with LHD and those with RHD by the time they begin primary school, it is possible that younger children and toddlers with FLs may demonstrate more subtle and transient language difficulties associated with hemisphere- and site-specific damage. Beginning at the start of language development, prospective cross-sectional and cross-linguistic studies of children with FLs allow a direct contrast between various 'stages' of brain organization for language. Several studies have investigated the early language abilities of infants from 10 to 40 months old, from first word comprehension to production, multiword utterances, and grammatical development. The results of these studies serve as *prima facie* evidence against the theoretical notion that the developing brain is a conglomeration of inexorably maturing modules devoted to different processing needs.

Early in the development of language skills (10–18 months), some children with perinatal focal lesions show a delay in language comprehension as well as gestural communication. However, this delay is only observed in children with right hemisphere injury; those with LHD (even with left temporal damage) are within the normal range on word comprehension. Not only is this the opposite of what one would expect from an adult model of language comprehension, in which left temporal damage tends to correlate with comprehension disorders, but also this finding goes against the usual lesion correlate of deficits in gesture production and comprehension, which in adults are strongly associated with LHD.

For early single-word production in the same cohort of children (10–18 months), lesions to either hemisphere provoke delays in development but are particularly severe in children with left temporal damage. This again contrasts with the adult model, in which language production deficits tend to correlate with left frontal damage. Similarly, results from multiple studies show that left temporal damage continues to be a predictor of delayed language production somewhat later in development (e.g., 19–31 months). Toddlers with left temporal damage show impaired lexical production, with commensurate delays in grammatical development. In contrast to the left temporal findings, frontal damage to either hemisphere was also implicated in production delays – again unlike the adult

model of left-lateralized frontal damage leading to production difficulties.

Neural Plasticity in Response to Focal Lesions

The low-normal language outcome of children with early onset FLs as well as the changing and distinctly non-adult-like profile of lesion site-to-language mappings over early development provide compelling evidence for the plasticity of the developing brain. These results also suggest that the usual pattern of brain organization for language (i.e., left hemisphere dominance) is neither inevitable nor even necessary for successful language processing. These findings suggest that there is considerable change in how the brain processes language, even well into the school-age years.

Clearly, these children's brains are organizing themselves differently than they would in typical development. Does this indicate that the brain is equipotential for language (i.e., any brain region can be equally co-opted into language processing), or are there early forming connections and architectures that make some developing brain regions more hospitable to language than others? More concretely, if left frontal and temporal regions are damaged early in development, will the brain regions enabling language function emerge throughout the brain (as one might predict given strict equipotentiality), or will contralateral regions homologous to those characteristically associated with language be recruited for language tasks?

This question has been investigated using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) on patients with early onset lesions. In a PET study, activation profiles for children with perinatal left and right focal lesions were compared on sentence listening and repetition tasks. Children and adolescents with early left hemisphere lesions showed extensive right lateralization of activation in both traditional perisylvian 'language' areas as well as much of the right temporal lobe (particularly the anterior temporal lobe), the angular gyrus, the precuneus, and the anterior cingulate. Roughly the opposite pattern was seen for the group with right hemisphere lesions. An fMRI study of five children with left periventricular lesions (causing hemiplegia but no epilepsy) showed a similar mirroring of typical left lateralized activation in a covert verb generation task.

Similarly, studies have directly compared typical adult language-related activation in children and adults with early and late-onset left hemisphere focal lesions, respectively. The expectation was that earlier injuries would allow greater potential for reorganization of function. Neurologically intact adults showed the usual left-lateralized perisylvian activation profile for sentence comprehension, whereas patients with early onset focal lesions showed a mirrored and right-lateralized profile, and late-onset patients show highly symmetrical perisylvian activation. The dramatic lateralization difference between the early and late-lesion groups suggests that at least some of the remarkable capacity of the early lesion group to acquire normal language may be due to development of these (or closely related) functions in the right hemisphere. These findings argue against the notion that language functions are inexorably hard-wired within specific left hemisphere regions; rather, they indicate that the brain has some flexibility to reorganize such cognitive functions, if necessary.

Interestingly, this early lesion-induced shift in interhemispheric activation is not seen in all domains, even in the face of early cerebral insult. One study comparing sentence listening with a basic motor task (finger tapping) in patients with relatively early onset left hemisphere lesions found strong rightward activation asymmetries for the language task (a pattern opposite that of healthy controls). However, the same participants showed strong leftward activation asymmetries in the motor task – a pattern very similar to controls. These results suggest that the development and commitment of neural resources for language is a protracted process – one that might be more flexible than the development of more developmentally and evolutionarily entrenched capacities such as the motor system. However, it is worth emphasizing that lesion-induced reorganization of language function appears to recruit roughly homologous regions in the right hemisphere rather than, for instance, co-opting left hemisphere visual and somatosensory regions to accomplish language tasks. This suggests that regions with certain patterns of cortical and subcortical connections may be more amenable than others to co-option or 'shared use' by language.

Conclusions

There are clear differences between 'acquired' and 'neurodevelopmental' congenital disorders, as well as considerable differences between the different neurodevelopmental disorders. The impressive language outcome in FLs is testament to remarkable abilities of an otherwise typically developing brain to adapt in the face of substantial neurological insult. This stands in stark contrast to the more complex and more pervasive language problems associated with children with ASD, WS, and DS, whose brains are following multiple and distinct atypical developmental pathways.

See also: Autism; Down Syndrome; Down Syndrome: A Disorder of Histogenesis; Language Development; Language: Cortical Processes; Language: Learning Impairments; Reading; Sentence Production; Sentence Comprehension; Word Production; Word Recognition; Word Learning.

Further Reading

Bates E and Roe K (2001) Language development in children with unilateral brain injury. In: Nelson CA and Luciana M (eds.) *Handbook of Developmental Cognitive Neuroscience*, pp. 281–307. Cambridge, MA: MIT Press.

Karmiloff-Smith A, Grant J, Berthoud I, Davies M, Howlin P, and Udwin O (1997) Language and Williams syndrome: How intact is 'intact'? *Child Development* 68: 246–262.

Rice ML, Warren SF, and Betz SK (2005) Language symptoms of developmental language disorders: An overview of autism, Down syndrome, fragile X, specific language impairment, and Williams syndrome. *Applied Psycholinguistics* 26: 7–27.

Singer Harris N, Bellugi U, Bates E, Jones W, and Rossen M (1997) Contrasting profiles of language development in children with Williams and Down syndromes delay using parental report. *Developmental Neuropsychology* 13: 345–370.

Language in Aged Persons

E A L Stine-Morrow and M C Shake, University of Illinois at Urbana–Champaign, Champaign, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Communicative competence is essential for effective functioning through adulthood. The ability to understand and remember language depends on a coordinated array of processing components that translate an orthographic or acoustic signal (i.e., printed symbols and speech sounds, respectively) into a representation of the meaning of written text or speech that changes dynamically as the signal unfolds in real time. Language production can be thought of as a reverse process, requiring message formulation from which a surface form is generated. Aging brings both growth (e.g., knowledge) and decline (e.g., speed of processing). As a result, there are a variety of changes in how language is processed, both normatively and in what is required for successful performance. In this article, we consider the nature of these changes. We first review prominent theories of cognitive aging and then present a conceptual model of language processing. We then consider how language comprehension, memory, and production change with age, taking into account data from behavioral paradigms, event-related potentials, and imaging.

Theories of Cognitive Aging

Although there are a number of competing theories for cognitive and brain mechanisms that underlie age-related changes in cognition, there is broad agreement that at a coarse level, age effects on cognition can be characterized as a result of two competing forces. On the one hand, senescence drives a decline in mental mechanics, the speed and accuracy with which elementary information processing components can be completed. On the other hand, there is accumulating evidence that the brain has immense potential for plasticity into late life so that experience-based growth in knowledge systems, skill, and expertise offers potential for growth. Language processing can potentially depend on both of these systems.

There are a variety of cognitive theories characterizing the specific nature of declines in mechanics with age, all of which have influenced both behavioral and neuroscience approaches to the study of language. The slowing hypothesis suggests that aging brings a systematic decrease in the speed with which mental operations can be performed. The working memory (WM) hypothesis posits that aging is associated with a decline in the capacity to perform basic processing operations and store their products (as might be required, for example, as one listens to a lecture and tries to construct an understanding of what is currently being discussed and integrate it with what has come before). The inhibition deficit (ID) hypothesis focuses on the important role that inhibitory function plays in effective language processing by suppressing irrelevant environmental input, inappropriate (or no longer appropriate) features of meaning, and incorrect interpretations. According to the ID account, inhibitory function is reduced with age, thus compromising comprehension by allowing irrelevant (or incorrect) information to be incorporated into the language representation – and consuming working memory capacity that might otherwise be used for effective processing. Finally, drawing on the WM hypothesis, the effortfulness hypothesis identifies sensory loss as critical to understanding the aging language system, not simply as a direct influence on the quality of signal, but as an influence on central resources that are strained by the attempt to interpret the muddy signal.

At the same time, aging may bring growth in vocabulary and some aspects of verbal ability, with evidence suggesting that such growth depends on habitual engagement with literacy-based activities. Particularized knowledge systems and expertise develop differentially as a consequence of occupational and leisure pursuits. Well-developed procedural skills, such as reading, can be well maintained into late life. Evidence suggests that language processing may impact the trajectory of mental mechanics as well, with several demonstrations of enhanced executive function among fluent bilinguals, who habitually manage two language systems.

Patterns of growth and decline are also evident in the aging brain, with selective changes in both brain structure and function. Consistent with the notion of age-related declines in controlled attention, as suggested by WM and ID accounts, the prefrontal cortex is the brain region most vulnerable to deterioration. Both evoked potential and imaging data very often show more diffuse activation in the aging brain, a pattern that has been explained alternately in terms of dedifferentiation (effectively, inhibition failure) and as compensatory recruitment.

The Nature of Language Processing

In both reading and speech understanding, the signal unfolds linearly in time. However, the structure of thought arising from this process is not linear and certainly represents information beyond the verbatim input. The task faced by theories of language comprehension, then, is to explain how this mental representation is constructed and updated over time during comprehension, as well as later when certain tasks require retrieval or use of the text representation. Contemporary models of language comprehension distinguish among distinct processes that operate in concert to construct different facets of the language representation. At the surface level, individual lexical items (words) are encoded from the orthographic or acoustic signal, and their meanings are activated. The stream is parsed into syntactically coherent segments (e.g., clauses and sentences) that establish the thematic roles (functions such as agent and object) for the constituents of the sentence. The semantic representation can be described in terms of integrated ideas (or propositions) that establish relationships among concepts given by the text, a representation called the textbase. Knowledge plays a role in facilitating integration, enabling elaborative inference, and evoking a simulation of the situation suggested by the text. Production can be described as the reverse of this process. To express a thought, a syntactic structure has to be formulated, particular lexical items must be selected to fill thematic roles (a process called lemma selection), and the phonology must be assembled (phonological encoding).

Age-Related Change in Language Processing

Language Comprehension and Memory

Consistent with the divergent age trajectories of mechanics and knowledge-based processes, aging appears to have the strongest impact on resource-consuming aspects of language processing.

Word processing Vocabulary often shows an increase with age, particularly among those who read regularly. Word recognition appears to be highly resilient through late life. For example, word frequency effects (i.e., faster processing for more familiar words) in reading and word naming are typically at least as large for older adults as they are for young. Sublexical features (e.g., neighborhood density), however, may have a smaller effect on processing time on older readers, suggesting that accumulated experience with literacy may increase the efficiency of orthographic decoding.

By contrast, declines in auditory processing can make spoken word recognition more demanding so that more acoustic information is needed to isolate the lexical item. Such effects may not merely disrupt encoding of the surface form but also tax working memory resources that would otherwise be used to construct a representation of the text's meaning. For example, when elders with normal or impaired hearing listen to a word list and are interrupted periodically to report the last word heard, they may show negligible differences. However, if asked to report the last three words, hearing-impaired elders will likely show deficits. The explanation for such a provocative finding is that the hearing-impaired elders overcome a sensory loss at some attentional cost so as to exert a toll on semantic and elaborative processes that enhance memory. Presumably, the same mechanisms would operate in ordinary language processing.

At the same time, there is evidence that older adults can take differential advantage of context in the recognition of both spoken and written words, especially in noisy environments. Semantic processes at the lexical level also appear to be largely preserved. Semantic priming effects (i.e., facilitation in word processing by prior exposure to a related word) are typically at least as large among older adults as among the young. Also, in the arena of neurocognitive function, evoked potentials show similar lexical effects for young and old – a reduced N400 component for related words relative to unrelated controls. One area of difficulty that older adults may have in word processing is in deriving the meaning of novel lexical items from context, with research showing that older adults are likely to infer more generalized and imprecise meanings relative to the young – a difference that can be largely accounted for in terms of working memory deficits.

Syntactic processing Syntax, or the set of rules directing appropriate formation of grammatical sentences, guides the near instantaneous and incremental processing of words into coherent chunks as we encounter them, a phenomenon often termed 'parsing.' Paradigms in neuroscience have shed some light on the influence of aging on brain activity during syntactic parsing. Most of this research indicates that initial parsing processes are fundamentally preserved with aging. Studies using functional magnetic resonance imaging (fMRI), for example, show that for simple sentences, the left peri-sylvian areas responsible for sentence processing produce similar patterns of activation for young and old. Also, attempts to understand ungrammatical sentences produce similar effects among young and old in evoked potentials – for example, exaggerated left anterior negativity to number violations or an exaggerated P600 to phrase

structure violations. Other research with evoked potentials suggests a qualitative shift in brain region recruitment, with brain potentials to syntactic violations becoming less asymmetric and more frontal with age. Despite substantial preservation of certain aspects of syntactic processing, complex syntax can especially compromise the older adult's ability to interpret a sentence's meaning and remember it, a finding often attributed to processes downstream from the initial parse. Such effects may be due to cognitive constraints imposed by age-related declines in the working memory resources needed to retain the products of parsing in memory. For example, behavioral and imaging studies comparing the effects of syntactically simpler subject-relative sentences (e.g., 'The dog that chased the boy is friendly') to more complex object-relative sentences (e.g., 'The dog that the boy chased is friendly') suggest that age differences emerge primarily for more complex constructions. As syntactic structure becomes more complex, older adults are more likely to show comprehension errors (e.g., was it the boy who chased the dog or vice versa?) than younger adults and may show relatively less activation in the left peri-sylvian sentence processing region. However, variability among older adults is high, so there is typically a subset of elders who perform at least as well as the young. These high-comprehension older adults appear to achieve good performance in part by upregulating additional brain regions associated with verbal working memory (e.g., left dorsal portions of the inferior frontal cortex and right temporal–parietal regions).

Textbase processing Older adults typically show poorer memory for the content expressed directly by the text. Online measures of reading (e.g., reading time and probed recognition) suggest that the fundamental mechanisms of the system used to construct an integrated representation of ideas are preserved. However, processes used to construct the textbase (e.g., to instantiate and integrate concepts in the text) are among the most resource-consuming of those required in language understanding and are hence the most vulnerable to aging. When reading is self-paced, older adults require more time for effective propositional encoding (e.g., as indexed by effective reading time, the time allocated per idea unit recalled). In listening, when the pace is controlled by the speaker, older adults may have particular difficulty in understanding and retaining the information, especially as propositional density (ideas expressed per word) is increased or in noisy environments. Older adults appear to have no difficulty drawing anaphoric inference (i.e., correctly identifying the referent when the pronoun is used to refer to a noun that was introduced earlier) over short distances, but they may have difficulty when the pronoun and referent are separated by intervening text so that the referent must be retained in working memory. Similarly, older adults may have difficulty in reactivating nouns during sentence comprehension in the way that younger adults do, which may compromise the ability to represent a coherent meaning. For example, younger adults trying to understand an object-relative construction, as described previously (e.g., 'The dog_i that the boy chased t_i is friendly') will reactivate the object (e.g., dog) of the matrix clause following the verb (i.e., at t_i, the gap in the phrase structure that leaves a trace of the constituent moved in the relativized construction) and actually show facilitated processing of the object at the gap site. Older adults are less likely to reactivate the object as the noun-gap distance is increased (e.g., 'The dog_i wearing the blue bandana that the boy chased t_i is friendly').

The general impression from the behavioral data that the semantic (textbase) representation is more fragmented and less distinctive among older adults is reinforced by evidence from evoked potentials. Semantic anomaly (e.g., 'The dog wearing the blue bandana is a tree') exaggerates the N400 component of the evoked potential for the anomalous word (in this case, 'tree') for both younger and older adults; however, the brain response to this anomaly is less pronounced among older adults. Such findings suggest that the encoding of the sentence context up to the anomalous word was not as distinctive among the older adults so that the perception of anomaly was not as strong when the implausible word was encountered. Interestingly, there is no evidence of a delayed response to the anomaly with age; rather, the data are more consistent with a reduced level of response. Contrary to early claims in the literature that aging brings no semantic deficits, it appears that difficulty in integrating concepts can lead to an impoverished semantic representation.

Ironically, older adults may ordinarily allocate less attention to textbase processing. This is demonstrated in reading time for naturalistic text, in which regression analysis is used to isolate the effects of textbase processes. For example, with lexical properties such as length and frequency controlled, reading time increases when new concepts are introduced or at the ends of syntactic constituents, which are sites at which readers pause briefly to integrate concepts into a coherent meaning. These data show that older readers may allocate less time to these processes. However, analogous to fMRI findings with sentence processing, older adults who show high levels of language memory performance often overallocate attention to these processing components or, as

described later, may shift allocation to process more holistic features of the discourse. Some older adults do show very good language comprehension and memory, with data from behavioral studies and neuroimaging suggesting that compensatory recruitment of additional resources underlies their successful performance.

Less is known about the factors that engender this recruitment. However, age effects on language performance are often moderated by verbal ability: more verbal older adults are more likely to show event-related potential patterns similar to young, allocate resources to construct the textbase, and show relatively good textbase memory. To the extent that literacy activities are practiced across the life span, growth in knowledge-based systems may support language processing at different levels.

Situation model Aside from deriving ideas directly from the text, language understanding also involves elaboration on these ideas based on existing knowledge. Some theories focus on the perceptual quality of this level of representation, which gives rise to a perceptual simulation of the events described by the text. For example, in narrative understanding, readers track the goals and emotional reactions of characters, as well as their movement through space and time. Behavioral methods to study this level of representation include probe recognition for objects in the narrative as well as reading time, both of which show subtle effects of situation model processing. Readers are slower to verify the existence (in the narrative world) of objects that are spatially distant from the protagonist relative to those that are nearby. Readers also slow down when new characters are introduced or when there is a spatial discontinuity (e.g., the locus of narrative events shifts from the village to the castle) or a temporal discontinuity (e.g., 'The next day . . .'). When the text describes a goal to be achieved (e.g., Susan intends to buy her mother a purse for her birthday), the goal is activated in memory until it is achieved so that concepts related to the goal are more quickly verified as long as the goal is open (e.g., the purse will be more quickly verified if Susan cannot find the purse when she goes to the store relative to a condition in which she succeeds in securing the gift). To the extent that these paradigms have been used to explore adult age differences in situation model processing, there has been very little evidence of developmental differences in situation construction and updating.

Outside of narrative comprehension, it appears that older adults also show the perceptual symbols effect, suggesting that fundamental processes of perceptual simulation are preserved. For example, adults are faster to verify that a pictured object was described in a sentence if the picture of the object represents its shape implied from the sentence (e.g., in encountering, 'She watched the egg frying in the pan,' readers are faster at verifying that the sentence described an egg if the drawing depicts a cracked egg with yolk and white relative to a picture of an egg in its shell, with the reverse effect for 'She watched the egg boiling in the pot'). According to perceptual symbol theory, response time is faster when the picture matches the shape implied by the sentence because sentence processing evokes a perceptual simulation of the events described so that the implied shape is more available when the picture is presented. This perceptual symbols effect is at least as large in the elderly as it is among the young.

Older adults may particularly rely on the situation model to support textbase processing. For example, in ambiguous text (e.g., 'The strength and flexibility of this equipment is remarkable. Not everyone is capable of using it even though most try at one point or another . . .'), older readers take differential advantage of titles (e.g., Driving a Car) that disambiguate the meaning to facilitate processing. Because the title renders the situation instantly transparent, both younger and older adults are more efficient in reading when it is available; however, older adults show this effect to a larger degree.

To the extent that the hallmark of situation model processing is an integration of textbase content with knowledge, one might expect that older adults would be particularly adept at inferential processing; however, this is not always the case. Although older adults are more likely to draw elaborative inference (e.g., in recall, to annotate their recollections with personal experiences or related information learned in another context), if inference is constrained so that it requires retrieval of textbase content, age deficits are the norm.

Discourse structures and context Beyond sentence processing, different genres of text have characteristic forms. For example, narratives typically begin by introducing a setting and characters and proceed to describe a series of episodes in which goals or problems are introduced to be resolved, and so on. Expository texts have certain characteristic forms of argumentation (e.g., problem–solution and thesis–evidence). Older readers generally appear to track these larger discourse structures in the same way as the young. For example, scrambling a narrative so that the canonical form is disrupted will impair recall similarly for young and old. However, other research suggests that older readers may have more difficulty in remembering narratives relative to matched expository text, presumably because of the

demands of establishing and tracking complex plot structures. For example, older adults are more likely to have difficulty when two plots are interwoven.

Recap Collectively, these findings are generally consistent with what would be expected given declines in mechanics accompanied by experience-based growth in knowledge. The most resource-intensive aspects of language processing requiring the associationist connection of concepts in propositional encoding and storage in working memory show the largest age deficits. At the same time, sensory deficits may indirectly tax working memory by diverting resources to interpret muddy input. However, knowledge-based processes are largely preserved and may support these other processes.

Language Production

In contrast to comprehension, production shows more pronounced differences as a function of age at almost every level of analysis. In this domain, methodological approaches are almost exclusively behavioral, and even among behavioral measures, the focus has been on word production and on comparing global measures of spontaneous production (e.g., diaries) because of the problems associated with constraining production in a way that can be measured (although there are interesting recent innovations).

Word production stands in sharp contrast to receptive word processes in showing age differences. Picture naming is slower and likely to produce more errors among older adults. As do the young, older adults show effects of name agreement and word frequency on naming latency, suggesting that processes of lemma selection and phonological encoding are intact. Older adults may show relatively faster response time when name agreement is high, suggesting perhaps enhanced effects of knowledge on lemma selection. Aging is also likely to bring an increase in tip-of-the-tongue experiences, that feeling of knowing the word one wants to say (e.g., this experience is evoked among some people by asking them the name of a person who collects stamps) but not quite being able to come up with it (e.g., philatelist). Interestingly, spelling errors also increase with age. This asymmetry between comprehension and production has been explained in terms of transmission deficits between the phonological and semantic networks. It is assumed that connections among lexical and phonological units generally decrease with increased age and disuse. Although the semantic representation is rich in connections, phonological units are constrained (e.g., there are rich and variable ways to encode the concept of chocolate, but if one cannot activate /ch/, production is necessarily disrupted).

Cross-sectional studies of spontaneous production show that older adults generally produce shorter sentences with fewer complex syntactic constructions (e.g., embeddings and left-branching constructions). Diary studies have shown longitudinal change toward simpler syntactic forms and reduced information content as a function of normal aging, with declines exaggerated among those later diagnosed with Alzheimer's disease.

In constrained production (e.g., construct a novel sentence from a given set of target words), age differences may be negligible when the number of target words to be accommodated in the production is small; however, as the number of targets is increased, older adults are more likely to become dysfluent, presumably because of the greater demands on syntactic construction and proposition assembly. Another form of constrained production that has been explored with aging is to create a sentence to describe a picture. In this paradigm, too, older speakers are likely to produce dysfluencies. The monitoring of eye movements has been used to examine the extent to which older adults might plan their utterances ahead to enhance fluency, but evidence on this issue is not yet clear.

Finally, there have been several demonstrations in the contexts of laboratory tasks (e.g., text recall) that older speakers may be more likely to produce off-target speech. Whereas some have shown this to be related to independent estimates of inhibition and have labeled this phenomenon verbosity, other research has demonstrated that such age differences in such elaboration can be exaggerated for personal topics, suggesting that such patterns may be related to social demands rather than to individual differences in cognitive function. This issue remains to be resolved.

Conclusion

There appears to be a two-way street between the multidirectional change in adult cognition and language processing. Age-related declines in sensory processes and cognitive capacity impact the ability to derive meaning from text and spoken language, especially as syntactic structure becomes more complex or speech rate is increased. Age-related declines also impact name retrieval and the syntactic complexity of production. Knowledge-driven processes can offset these declines to a large degree so that in many arenas, aging individuals will show better performance than their younger counterparts. At the same time, language processing (e.g., habitual reading, or learning a second language) appears to exercise the cognitive system so as to contribute to cognitive vitality.

See also: Aging of the Brain; Cognition in Aging and Age-Related Disease; Dementia and Language; Language Development; Language: Auditory Processes; Language: Cortical Processes; Reading; Sentence Production; Sentence Comprehension; Word Production; Word Recognition.

Further Reading

Bialystok E, Craik FIM, and Freedman M (2007) Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsycholgia* 45: 459–464.

Federmeier KD, McLennan DB, De Ochoa E, and Kutas M (2002) The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology* 39: 133–146.

Federmeier KD, Van Petten C, Schwarz TJ, and Kutas M (2003) Sounds, words, sentences: Age-related changes across levels of language processing. *Psychology and Aging* 18: 858–872.

Griffin ZM and Spieler DH (2006) Observing the what and when of language production for different age groups by monitoring speakers' eye movements. *Brain and Language* 99: 272–288.

Johnson RE (2003) Aging and the remembering of text. *Developmental Review* 23: 261–346.

Kemper S and Mitzner TL (2001) Language production and comprehension. In: Birren JE and Schaie KW (eds.) *Handbook of the Psychology of Aging,* 5th edn., pp. 378–398. New York: Academic Press.

Kemper S, Thompson M, and Marquis J (2001) Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging* 16: 600–614.

Meyer BJF and Pollard CK (2006) Applied learning and aging: A closer look at reading. In: Birren JE and Schaie KW (eds.) *Handbook of the Psychology of Aging,* 6th edn. New York: Elsevier.

Radvansky GA, Zwaan RA, Curiel JM, and Copeland DE (2001) Situation models and aging. *Psychology and Aging* 16: 145–160.

Stine-Morrow EAL (2007) The Dumbledore hypothesis of cognitive aging. *Current Directions in Psychological Science* 16: 289–293.

Stine-Morrow EAL, Miller LMS, Gagne DD, and Hertzog C (in press) Self-regulated reading in adulthood. *Psychology and Aging*.

Stine-Morrow EAL, Miller LMS, and Hertzog C (2006) Aging and self-regulated language processing. *Psychological Bulletin* 132: 582–606.

Thornton R and Light LL (2005) Language comprehension and production in normal aging. In: Birren JE and Schaie KW (eds.) *Handbook of the Psychology of Aging,* 6th edn. New York: Academic Press.

Wingfield A and Grossman M (2006) Language and the aging brain: Patterns of neural compensation revealed by functional brain imaging. *Journal of Neurophysiology* 96: 2830–2839.

Wingfield A and Stine-Morrow EAL (2000) Language and speech. In: Craik FIM and Salthouse TA (eds.) *The Handbook of Aging and Cognition,* 2nd edn., pp. 359–416. Mahwah, NJ: Erlbaum.

Wingfield A, Tun PA, and McCoy SL (2005) Hearing loss in older adults: What it is and how it interacts with cognitive performance. *Current Directions in Psychological Science* 14: 144–148.

Language: Aphasia

N F Dronkers, VA Northern California Health Care System, Martinez, CA, USA; University of California at Davis, Davis, CA, USA; and University of California at San Diego, La Jolla, CA, USA
J V Baldo, VA Northern California Health Care System, Martinez, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Aphasia is an acquired impairment in language due to a brain injury. To be considered aphasic, the impairment must affect a range of language functions, including speech, comprehension, reading, and writing. Aphasia usually follows from injury to the left hemisphere of the brain. In stroke, it is usually due to a disruption of blood flow in the middle cerebral artery, which provides blood supply to the lateral surface of the cortex. Aphasia can also occur following a traumatic brain injury or may emerge gradually due to progressive, degenerative brain disease. Traditionally, aphasia has been classified using subtypes that are characterized by specific patterns of impairment in spoken fluency, comprehension, naming, and repetition. These aphasia subtypes include Broca's aphasia, Wernicke's aphasia, conduction aphasia, global aphasia, anomic aphasia, and transcortical aphasia. This article reviews the behavioral characterization of these aphasia subtypes as well as discusses their anatomical basis.

Broca's Aphasia

Broca's aphasia was first described by the French neurologist, Paul Broca, in the 1860s. He described a patient who could say very few words and repeatedly uttered the sound, 'tan tan tan,' when trying to speak. Despite his poor speech output, this patient understood much of what was said to him. A second patient soon followed who showed a similar dissociation of reduced speech fluency and relatively preserved comprehension. This form of aphasia came to be known as Broca's aphasia, and it is sometimes referred to as motor aphasia or nonfluent aphasia.

The following is a speech sample of a patient with Broca's aphasia as he attempts to describe a drawing of a lakeside picnic from the Western Aphasia Battery (WAB):

> I see dog, boy, boy, sand, see book, people, boats, on boat . . . girl and man reading, man fishing . . .

Such telegraphic speech is typical of patients with Broca's aphasia. Despite its halting nature, such speech often contains important content words so that the main ideas are conveyed, but smaller functor words (e.g., 'a' and 'the') are omitted, resulting in agrammatic speech (e.g., "I see dog"). There is debate as to whether such agrammatic speech is due to the inability of these patients to use grammatical forms or whether it simply reflects a strategic choice due to a reduced processing capacity. One way in which this issue has been addressed has been to study agrammatic patients who speak different languages. In German, for example, it has been shown that patients with Broca's aphasia do not commonly omit functor words such as 'the' (as American Broca's aphasics might) because these words are critical to the meaning of the noun (whereas in English they are not). This cross-cultural difference would suggest that patients with Broca's aphasia can produce grammatical forms if necessary to communicate content but simply omit what is not essential in order to simplify production demands.

The most severe form of Broca's aphasia manifests as recurring utterances, where a patient repeatedly utters a single word or phrase (such as Broca's original patient). Interestingly, such patients can effectively use intonation to vary the pitch of such utterances and thus communicate in a limited fashion. This behavior illustrates the fact that patients with Broca's aphasia, despite their severely reduced speech output, have relatively good awareness of their communication deficit.

Comprehension in patients with Broca's aphasia is relatively preserved, as long as sentences are not complex. For example, sentences such as 'The boy was kissed by the girl' may be misunderstood as 'The boy kissed the girl,' the latter being the more canonical construction in English. Repetition and naming are also disrupted in Broca's aphasia due to the speech output disturbance. Similarly, reading and writing are affected; for example, patients often are unable to write complete phrases or sentences (parallel to their spoken output).

Commonly accompanying Broca's aphasia is apraxia of speech, a motor speech disorder characterized by a reduced ability to coordinate the articulators (lips, tongue, etc.) that results in groping and misarticulated speech. Another commonly coexisting motor speech disorder is dysarthria, which results in slurred and distorted speech due to muscular weakness. Although these motor speech disorders often accompany Broca's aphasia, they can also occur in the other aphasias described later.

Textbook descriptions of the neural correlates of Broca's aphasia most commonly associate it with Broca's area in the posterior, inferior frontal gyrus, as originally described by Broca. Current research

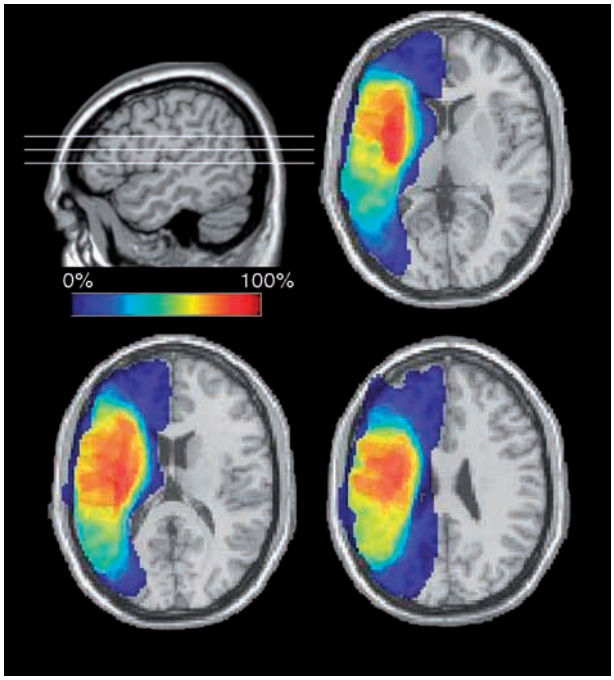

Figure 1 Average lesion map of 36 stroke patients with chronic Broca's aphasia showing the region of common overlap in the insula, rather than Broca's area.

data, however, suggest that for a chronic, persisting Broca's aphasia, the lesion is generally much larger, encompassing regions of inferior frontal and insular cortex, as well as deeper, white matter. Indeed, a recent three-dimensional magnetic resonance imaging (MRI) imaging study of Broca's two original patients' brains (Tan and Leborgne) revealed that the lesions of these historic cases actually extended quite medially, much deeper than Broca could have seen. Figure 1 shows an average lesion map of 36 stroke patients with chronic Broca's aphasia (poststroke > 12 months) from our research database. The color bar indicates the degree of lesion overlap across patients, from 0% (dark blue) to 100% overlap (dark red). As can be seen, the region of common overlap in patients with Broca's aphasia is in the insula (deep red), a region of neocortex that is buried beneath frontal and temporal cortices. We have previously shown this region to be critical for coordinating articulation, which is commonly impaired in patients with Broca's aphasia.

Wernicke's Aphasia

Karl Wernicke was a German neurologist and psychiatrist born in 1848. In 1874, at the age of just 26 years, he published a major work on aphasia that described patients with an inverse pattern of deficits from that observed by Broca, namely patients with fluent speech and a marked comprehension impairment. This form of aphasia came to be called Wernicke's aphasia but is also referred to as sensory aphasia or fluent aphasia. Unlike Broca's aphasia, Wernicke's aphasia is considered a fluent aphasia because patients with Wernicke's aphasia speak spontaneously with a normal to fast rate of speech, and their speech rhythm and prosody (melody) are normal. The following is a sample from a patient with Wernicke's aphasia describing the picnic scene:

> The/goll//dz/George. It isn't correct. He guess – he/hamilz/in the the one guy who/stamilz/in the – oh I can't say it. This girl who has her own rain. She can see this/goll/who has a ride. A picnic – he is/kain/a you see the mai tai.

Although their speech may be fluent, patients with Wernicke's aphasia are often unintelligible. This is in part due to the presence of numerous speech errors. Such errors sometimes take the form of literal paraphasias, which involve the swapping of sounds or syllables, or substituting one sound for another (e.g., saying 'pyramos' for 'pyramid'). Sometimes, these errors are so distorted that the target word is unclear (e.g., 'tiromon'). Such errors are referred to as neologisms. Patients with Wernicke's aphasia also make semantic errors, substituting a related word for an intended target (e.g., 'son' for 'brother'). Paraphasic errors occur during repetition and naming as well as spontaneous speech. Interestingly, despite the abundance of paraphasic errors in Wernicke's speech, many aspects of syntactic structure remain intact.

In addition to their paraphasic speech, patients with Wernicke's aphasia also have significantly impaired comprehension. This deficit is apparent in all modalities (i.e., whether spoken or written language), although patients may show slightly improved comprehension with written material. Unlike patients with Broca's aphasia who only have difficulty with more complex items, patients with Wernicke's aphasia fail on a range of comprehension tasks from single word comprehension to sentence comprehension.

Relative to patients with Broca's aphasia, patients with Wernicke's aphasia appear less aware of the severity of their language deficits. Patients with Wernicke's aphasia are less likely to self-correct errors in their speech production. This may be due in part to their impaired comprehension, which extends to monitoring their own speech output.

Wernicke's aphasia has traditionally been associated with lesions in Wernicke's area, which generally is defined as the left posterior, superior temporal gyrus. However, the exact boundaries of this area vary from textbook to textbook, presumably because the boundaries were never clearly defined. An overlay of 11 patients' lesions from our center reveals that a

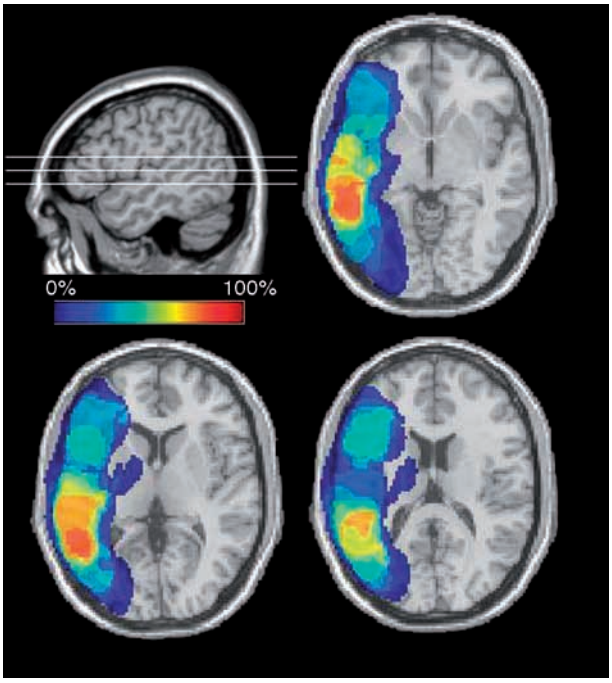

Figure 2 Average lesion map of 11 stroke patients with chronic Wernicke's aphasia. Lesions are centered in the middle temporal gyrus, rather than Wernicke's area.

lesion centered in the middle temporal gyrus is critical for a chronic, persisting Wernicke's aphasia (Figure 2). Other areas that are commonly affected include the posterior, superior temporal gyrus and inferior parietal cortex. These regions have been associated with auditory processing, lexical semantics, and verbal working memory.

Conduction Aphasia

Conduction aphasia was first described by Wernicke to characterize patients with relatively preserved comprehension and paraphasic speech. Later, Wernicke added to the list of deficits a disorder of repetition. Clinically, patients with conduction aphasia present with fluent speech and relatively good comprehension, but their speech is somewhat paraphasic and their ability to repeat is greatly reduced. Reading and writing, as well as naming, are also affected to a moderate degree. A clinical sample of such a patient's fluent spoken output describing the picnic scene follows:

> I see a bunch of people. There's a tree and a car and a house and water and the pier. Guy's fishing, and a couple o' guys that are saying hi to the sail's person. And a guy washing something next to a/hail/of sand and grass. A flag – is that a flag? And the wind and the pole and some more grass.

Despite such relatively fluent speech, patients with conduction aphasia exhibit significantly impaired repetition. Patients may get the gist or general meaning of what is said (as evidenced by their relatively good comprehension), but they are impaired at repeating material verbatim. For example, in an attempt to repeat the sentence 'The pastry cook was elated,' one patient responded "Something about a happy baker." This phenomenon suggests that patients with conduction aphasia are relying on a nonverbatim, semantic route since the phonological trace is no longer available. Dependence on this semantic route is also revealed by the fact that repetition is severely affected for nonsense words (e.g., 'meppen'). Such words have no meaning and thus cannot give rise to a paraphrase or synonym, let alone reproduction of the same word.

Although controversial, some researchers divide patients with conduction aphasia into two groups – reproduction and repetition conduction aphasia. These two groups reflect different parts of the complex repetition process. Repetition conduction aphasics are thought to primarily have a problem at the level of holding a memory trace – in other words, a deficit in verbal working memory. Reproduction conduction aphasics are thought to have their primary problem with the processing and production of phonological information. Researchers who maintain this division acknowledge that some patients diagnosed with conduction aphasia fall into both groups.

Paraphasic errors in patients with conduction aphasia may take the form of phonemic substitutions (e.g., 'free' for 'tree') or semantic substitutions (e.g., 'father' for 'brother'). However, they are more likely (than patients with Wernicke's aphasia) to attempt to correct their errors. For instance, a patient with conduction aphasia was asked to name a line drawing of a cactus and responded by saying "gaktus, no, gaksus, gackone, no, cackus, little screwed up there, cack, cactus! cactus!"

Traditional models of aphasia originally suggested that conduction aphasia should arise from lesions to the arcuate fasciculus, the white matter tract connecting Wernicke's and Broca's areas. Despite the fact that no data exist to support this model, most textbooks in psychology and neuroscience perpetuate this notion. Instead, patients with conduction aphasia most commonly have lesions centered in left posterior neocortex, most notably in inferior parietal cortex, not the arcuate fasciculus. Figure 3 shows a lesion overlay map of 13 patients with chronic conduction aphasia. As can be seen, the region of common overlap is centered in inferior parietal cortex. This region has been suggested to be critical for verbal working memory, which is consistent with the pattern of repetition deficits observed in this patient group.

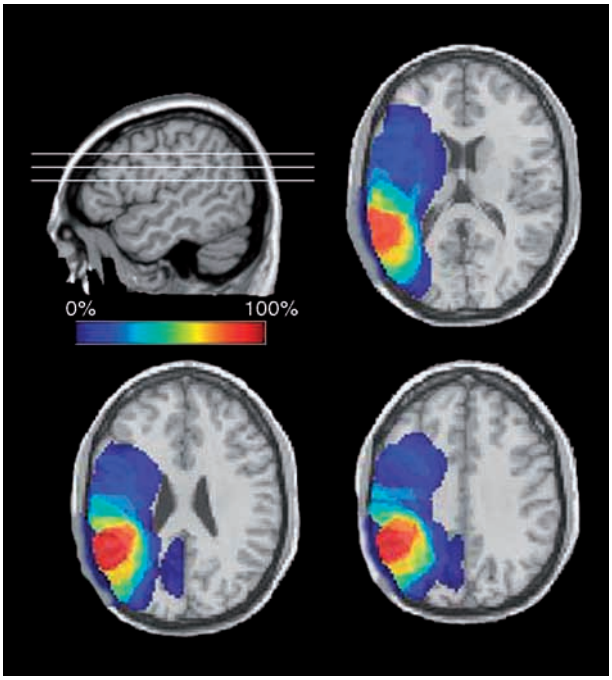

Figure 3 Average lesion map of 13 stroke patients with chronic conduction aphasia, showing a region of common overlap in the inferior parietal cortex.

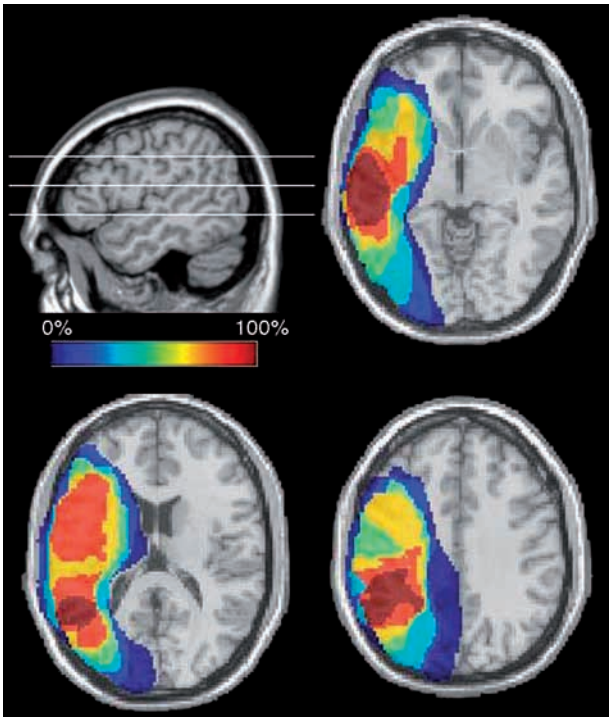

Figure 4 Average lesion map of seven patients with chronic global aphasia; showing that lesion typically involves a large portion of the left hemisphere.

Global Aphasia

Global aphasia is the most severe of all aphasia subtypes, with significant impairments across all aspects of language, namely impaired speech, comprehension, repetition, naming, reading, and writing. Patients with global aphasia may be able to utter automatic or stereotypic responses (e.g., "yes" and "no") but do so unreliably. For example, one patient with global aphasia when asked to describe the picnic scene from the WAB was only able to utter "no…no… because" on one occasion and "yeah" on another occasion. Other patients with global aphasia are only able to produce overlearned or automatic phrases (e.g., "How are you?"). Patients with global aphasia also have significantly impaired comprehension. They often have difficulty responding to very simple yes/no questions (e.g., "Are you a man?").

Although the severe loss of speech and language makes it very difficult for patients with global aphasia to communicate, they are sometimes able to convey information by varying the intonation in their voice or by using simple gestures. Importantly, patients with global aphasia can be shown to perform normally on nonverbal tasks (e.g., picture matching), demonstrating that they are not suffering from confusion or dementia.

As described previously, some patients with global aphasia have a preserved ability to utter automatic phrases or repetitive utterances. What is the brain basis of these abilities, if much of the language network is damaged? There is evidence that these automatic and overlearned phrases are generated by regions within the intact right hemisphere of the brain. The right hemisphere is also capable of supporting intonation in speech, musical ability, and emotional expression. Again, these are abilities that are sometimes preserved in patients with global aphasia.

As with the other aphasias, global aphasia is most commonly the result of a stroke in the middle cerebral artery that supplies blood to the lateral surface of the left hemisphere of the brain. Not surprisingly, lesions necessary for a persisting, chronic global aphasia are generally quite large and encompass large portions of the left peri-Sylvian region. Lesion overlays from seven patients with global aphasia are shown in Figure 4. This large lesion is responsible for the gross impairment in most aspects of speech and language since the entire language network is affected. Whereas other patients with smaller lesions may find ways of compensating for their language disabilities, patients with global aphasia have less brain reserve with which to do so.

Anomic Aphasia

Anomic aphasia is the mildest of the aphasias, with relatively preserved speech and comprehension but difficulty in word finding. The persistent inability to

find the correct word is known as anomia (literally, 'without names'). Anomia is actually a symptom of all forms of aphasia, but patients whose primary language problem is word retrieval are diagnosed as having anomic aphasia. In other words, patients with anomic aphasia have relatively preserved speech fluency, repetition, comprehension, and grammatical speech but sometimes have difficulty with word finding. The following is a sample from a patient with anomic aphasia describing the picnic scene:

> I see a boy who is flying a kite. I see a girl who pouring . . . um . . . the soda an' I see a girl who is listening her radios. I see a man who is reading a book. I see a man who is fishin' an' I see a dog who is waggin' his tail. I see a boy waiting . . . wading in the water.

Unlike the previous aphasia samples, this patient describes the scene using complete sentences and provides a relatively good level of detail. There are a few minor errors and some groping for words (e.g., 'soda'), but otherwise her speech is quite clear and communicative, typical of anomic aphasia.

When searching for a word, some patients with anomic aphasia paraphrase using words that they can easily retrieve. For instance, a patient shown a drawing of a pair of tongs said "You pick up things with it." Such circumlocutions demonstrate that patients with anomic aphasia have lost neither conceptual understanding nor the ability to build coherent sentences and phrases.

Whereas circumlocutions may be a characteristic feature of the speech of some patients with anomic aphasia, the main characteristic displayed by other patients may be periods of slow and halting speech as they search for the correct word. For instance, when shown a picture, a patient might say, "It's a, ummm, uhhh, a, a, a . . . I know what it is, it's a . . . Aww hell, it's there but I just can't get it." Such hesitancy may be so frequent that a patient may be incorrectly classified as having Broca's aphasia. Not surprisingly, patients with anomic aphasia benefit greatly from cueing, such as providing the first sound of the target word.

Low-frequency words are typically more difficult for a patient with anomic aphasia to retrieve and produce than frequently used words. It is very common for proper names to be more difficult to retrieve than common nouns, and common nouns are often more difficult to retrieve than adjectives or verbs. Thus, both the grammatical function of a word and its frequency can affect word retrieval in patients with anomic aphasia.

Lesions associated with anomic aphasia are the least well defined because word finding is sensitive to damage throughout much of the peri-Sylvian region.

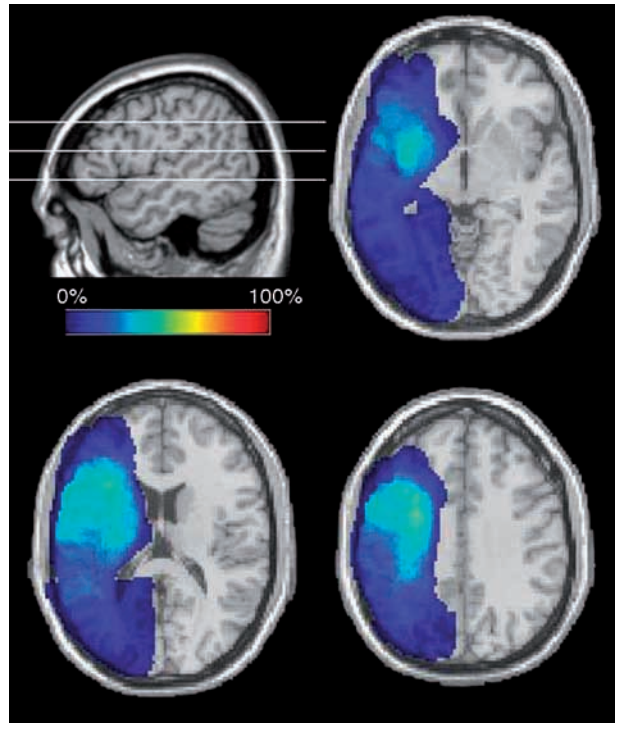

Figure 5 Average lesion map of 37 patients with chronic anomic aphasia, in which there was no single, common region of overlap.

As shown in Figure 5, there is no single, common region of overlap in this group of 37 patients with anomic aphasia but, rather, the lesions span a wide range of areas in the left hemisphere.

Transcortical Aphasias

The transcortical aphasias are a group of aphasia syndromes that all have in common a relatively preserved ability to repeat, despite other significant language deficits. Transcortical motor aphasia is similar to Broca's aphasia; that is, patients exhibit nonfluent speech and relatively good comprehension, but the ability to repeat is relatively preserved. Transcortical sensory aphasia is similar to Wernicke's aphasia, with patients exhibiting fluent speech and poor comprehension and also a relatively preserved ability to repeat. Finally, mixed transcortical aphasia is similar to global aphasia but with relatively preserved repetition. Compared to the other aphasias, the transcortical aphasias are relatively rare.

The following is a speech sample from a patient with transcortical motor aphasia describing the picnic scene:

> The girl is.. the girl is .. drinking the soda. The girl .. no .. the guy is .. no .. yeah. The guy is playing the radio. The people who . . . OK .. flying a kite. The dog is playing around the people. The uh .. the .. the . . . the . . . dog is .. no .. the fly a kite. And uh, the people who rented a boat was my friends.

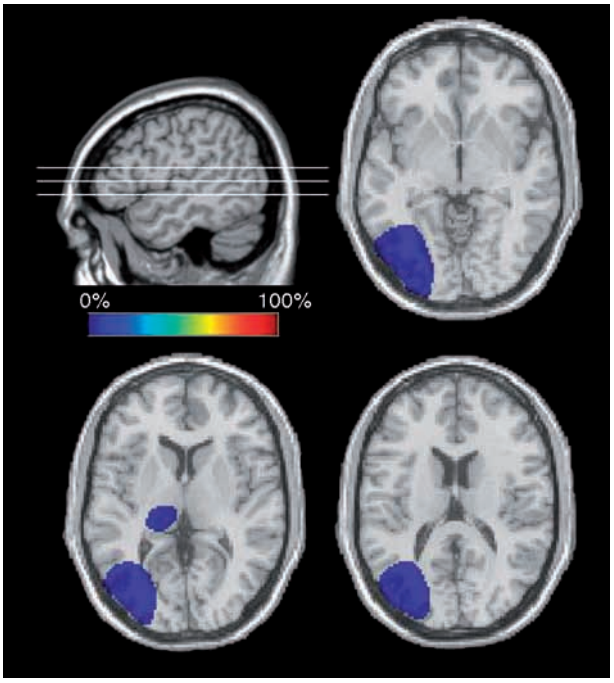

Figure 6 Lesion locations in two patients with chronic transcortical sensory aphasia.

This patient with transcortical motor aphasia exhibits hesitation, word-finding problems, and some grammatical errors – all symptoms reminiscent of a person with Broca's aphasia. However, when asked to repeat 'The pastry cook was elated,' the same patient does so flawlessly. Such intact repetition is quite remarkable given the level of impairment in spontaneous speech, and this preserved ability is what distinguishes the transcortical aphasias.

Lesion correlates of the transcortical aphasias have not been as clearly defined as the other aphasias, in part due to the fact that these aphasias are less commonly observed. Figure 6 shows the location of lesions in two different patients with transcortical sensory aphasia, the form of aphasia reminiscent of Wernicke's aphasia but with preserved repetition. As can be seen, one patient suffered a subcortical stroke, whereas the other suffered a posterior stroke that affected temporal and parietal cortex. Both patients' lesions are quite distinct from lesions typically observed in Wernicke's aphasia. Also, the lesions spare regions associated with repetition ability, such as those affected in conduction aphasia.

Conclusion

For more than 100 years, scientists have studied and characterized speech and language impairments in patients with brain injury. Such efforts have provided an important framework for our understanding of both the cognitive linguistic aspects of language and the localization of these processes in the brain. One difficulty has been that the aphasia syndromes involve a constellation of speech and language symptoms that significantly overlap across subtypes, as is apparent from the previous discussion. For this reason, newer approaches to the study of language focus on the neural correlates of discrete speech and language symptoms rather than the entire aphasic syndrome. New neuroimaging techniques such as functional MRI and voxel-based lesion symptom mapping facilitate the visualization of discrete neural regions associated with performance on specific language behaviors. These techniques will allow us to further our understanding of aphasia as well as the neural correlates of speech and language in the normal brain.

See also: Aphasia: Sudden and Progressive; Brain Trauma; Broca's Area: Evolution; Language Development; Language: Cortical Processes; Lexical Impairments Following Brain Injury; Sentence Production; Sentence Comprehension; Word Production; Word Recognition.

Further Reading

Bates E, Wilson SM, Saygin AP, et al. (2003) Voxel-based lesion-symptom mapping. *Nature Neuroscience* 6(5): 448–450.

Benson DF and Ardila A (1996) *Aphasia*. New York: Oxford University Press.

Broca P (1861) Nouvelle observation d'aphémie produite par une lésion de la troisième circonvolution frontale. *Bulletins de la Société d'anatomie (Paris) 2e serie* 6: 398–407.

Broca P (1861) Perte de la parole: Ramollissement chronique et destruction partielle du lobe anterieur gauche du cerveau. *Bulletins de la Societe d'anthropologie 1re serie* 2: 235–238.

Broca P (1861) Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletins de la Société d'anatomie (Paris) 2e serie* 6: 330–357.

Dronkers NF (1996) A new brain region for coordinating speech articulation. *Nature* 384(6605): 159–161.

Dronkers NF (2000) The pursuit of brain–language relationships. *Brain and Language* 71(1): 59–61.

Dronkers NF, Plaisant O, Iba-Zizen MT, and Cabanis EA (2007) Paul Broca's historic cases: High resolution MR imaging of the brains of Leborgne and Lelong. *Brain* 130: 1432–1441.

Eggert GH (1977) *Wernicke's Works on Aphasia: A Sourcebook and Review*. The Hague, The Netherlands: Mouton.

Goodglass H (1993) *Understanding Aphasia*. San Diego: Academic Press.

Kertesz A (1979) *Aphasia and Associated Disorders: Taxonomy, Localization, and Recovery*. New York: Grune & Stratton.

Love RJ and Webb W (2001) *Neurology for the Speech–Language Pathologist*, 4th edn., Boston: Butterworth–Heinemann.

Nadeau SE, Gonzalez Rothi LJ, and Crosson B (eds.) (2000) *Aphasia and Language: Theory to Practice*. New York: Guilford.

Sarno MT (1998) *Acquired Aphasia*, 3rd edn., New York: Academic Press.

Language: Auditory Processes

D Gow and D Caplan, Massachusetts General Hospital, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The auditory stimulus in speech perception is shaped and constrained by the acoustical properties of the human vocal tract. Speech is produced through the coordinated movement of the glottis, tongue, velum, lips, and jaw. The form of this signal is often understood in terms of a simple source-filter model in which acoustic energy produced by subglottal pressure and the vibration of the vocal folds, or turbulence produced when airflow is directed against a barrier, is given spectral shape by the natural resonances of coupled natural cavities created by different vocal tract configurations. The component sounds produced by articulation consist of simultaneous frequency-limited resonances with time-varying frequencies called formants, as well as silences, pops, clicks, and brief bursts of frequency-limited noise. Movement of articulators changes the shape of resonators, which in turn changes the frequencies of formants over time. Additionally, release of air when constrictions are released, or noise produced when constrictions are narrow but incomplete, produces noise. These sounds overlay an overall periodic structure caused by the opening and closing of the vocal folds during much of speech and a quasi-periodic structure caused by changes in subvocal pressure and vocal fold vibration.

While the specific acoustic form of spoken language reflects the microstructure of articulation defined in both time and space, this relation is complex. For example, the same [r] sound can be produced by radically different tongue movements. Conversely, the vocal tract configurations that produce [i] and [u] sounds in some contexts produce [j] and [w] sounds in others. This relationship is simplified somewhat by the fact that there is often a nonlinear or quantal relation between individual articulations and their acoustic consequences that maximizes the acoustic contrast between speech sounds, while providing relatively stable acoustic properties within them. For example, when producing a segment with the feature 'nasal,' such as [n], [m], or [ng], speakers move the velum to produce an opening between the oral and nasal cavities. While the act of opening involves a continuous articulation, there is a critical point at which the opening is large enough to allow acoustic coupling, which is marked by the appearance of a new spectral prominence, the nasal pole, which marks the segment as a nasal.

The primary computational problem posed by the perception of spoken language is created by the variability of the speech signal relative to the linguistic categories that are perceived by the listener. Spoken language is perceived as a sequence of more or less discrete events such as words, syllables, or component speech sounds. However, coarticulation, the overlapping articulation of adjacent speech sounds, often makes it impossible to draw a one-to-one mapping between a single speech sound and any interval of sound. Moreover, the acoustic realization of any unit of spoken language is highly variable, and may be influenced by a variety of factors including speaking rate, individual differences in vocal tract anatomy, language or dialect, articulatory style, the emotional state of the speaker, or social context.

Each level of linguistic representation is related to somewhat different properties of the speech signal, although these properties overlap. Word recognition relies on the perception of 'segmental' properties. These tend to be collections of relatively brief acoustic events. At typical natural speaking rates, speakers produce roughly 15 segments per second, with wide variability in the duration of different segments and temporal overlap in the distribution of properties associated with any one segment. At the level of sentence or discourse structure, 'suprasegmental' or intonational properties may provide information about syntactic constituent structure or the relative importance of individual words or phrases. These properties are typically defined over larger temporal regimens than segmental properties, and are encoded primarily in relative changes in pitch, intensity, and timing between different parts of an utterance.

These linguistic entities are perceived despite the enormous variance in how they are realized in the acoustic signal. Different research groups are currently pursuing four contrasting approaches to the issue of variance. One approach is to identify relatively invariant acoustic properties of speech sounds. While early research found an invariant correlate of one characteristic of a subset of syllable initial consonants, subsequent work has yielded few successes. A second approach focuses on possible normalization processes that may play a role in categorization or recognition. Work in this vein has demonstrated a variety of normalization processes, most notably processes of rate normalization in the classification of duration-dependent feature cues. A third approach is built on the premise that speech gestures are less

variable than the acoustic forms produced by their combination. Within this approach, researchers are concerned with trying to determine whether speech perception is mediated by articulatory representations. Finally, a fourth approach suggests that words or other linguistic units are stored for the purpose of recognition in the form of multiple exemplars that present relatively veridical representations of individual tokens encountered in the past (similar to machine speech recognition). There is evidence that listeners have access to memory of specific speech tokens. However, it is not clear whether this type of memory plays a role in the online perception of spoken language. The ultimate solution to the variability problem may involve some combination of these approaches.

Unique Neural Mechanisms for Processing Language Acoustically?

At the earliest stages of processing, both speech and nonspeech stimuli are processed using the same pathways, and are subject to the same processes. Transduction of acoustic energy by cochlear hair cells with frequency-selective tuning curves breaks acoustic input down into a representation of the intensity of acoustic energy at different frequency bands. An unresolved question is whether complex stimuli like speech are represented entirely by the spatial pattern of activation across tonotopically organized cortex, the temporal pattern of neural activity, or both. At some point, the processing of acoustic signals as linguistic entities diverges from processing acoustic signals as other types of stimuli. But whether that point occurs at the stage of processing the acoustic signal itself, or only after linguistic categories have been activated, remains a debated issue.

For some time, a considerable number of researchers argued that listeners applied different analytic processes to sounds that were and were not speech. The existence of what were thought to be unique aspects of speech perception – such as the inability to discriminate synthesized acoustic signals that did not fall into different phonemic categories while being able to discriminate signals with equivalent acoustic differences when they were perceived as being linguistically different (so-called categorical perception) – was considered strong evidence for this position. Confidence in the implications of such findings eroded when it was discovered that nonhumans show comparable phonetic discrimination effects. Similarly, psychophysical evidence suggests that humans are able to resolve temporal order of linguistic events with substantially more precision than that of nonlinguistic events. This again suggests a relatively low-level divergence between linguistic and general auditory processing streams. However, it has been suggested that this difference may be related to differences in the ability of linguistic versus nonlinguistic components to be represented at more global levels.

More recently, the question has been addressed on the basis of the location of neural responses to processing acoustic properties of linguistic and nonlinguistic stimuli. A challenge facing this approach is to equate the complexity of the speech and nonspeech stimuli while at the same time eliminating speechlike properties from the nonspeech stimuli that could cause activity in areas that are in fact uniquely devoted to processing acoustic stimuli as speech. Different research groups have taken different positions on the appropriate control stimuli. Some results suggest that perception of segmental acoustic properties as meaningful speech uniquely involves a region in the left superior temporal lobe anterior to the primary auditory cortex, but the issue remains open. Even fewer studies have examined the question of whether processing of acoustic signals to yield intonation contours utilizes the same brain regions as processing of acoustic signals to yield nonlinguistic representations, such as musical structures. One study reported that one patient whose abilities to discriminate melodies was impaired had comparable difficulties with intonation contours, but whether the same parts of the complex lesion in this patient were responsible for both functional impairments is unclear; to our knowledge, no functional neuroimaging studies have addressed this question.

Segmental Processing

The targets of segmental perceptual processes are abstract representations of the sounds of words. Phonological theories are essentially unanimous in maintaining that the sounds of words consist of a number of structures. In many models, the critical structures are 'phonemes' – abstract segments representing idealized units of sound whose presence is required for a word to be individuated. For instance, the word 'cat' consists of three phonemes, [k], [æ], and [t], in that order. A change in one of these segments to another phoneme (e.g., of [k] to [s]) turns 'cat' into a different word or into a nonword. Phonemes, too, are conceived of as complexes of 'features' – characterizations of positions and trajectories of articulatory organs and qualities of the sounds thereby produced. Examples of such features are 'stopped' (indicating full closure of the vocal tract), 'high' (referring to the position of the body of the tongue), and 'voiced' (indicating vibration of the vocal cords). Like phonemes, phonological features are abstract entities,

identified on the basis of many aspects of the sound patterns of words and other phenomena in a language. For instance, most models of features utilize a binary system, despite the continuous nature of articulatory postures, because generalizations that arise with respect to high-level phonological properties of words (e.g., effects of word formation on phonemes) can be best captured in such a system. Words also have phonological structure above the phoneme level, such as the level of syllables. In all languages, syllables in multisyllabic words have different degrees of prominence, represented as either lexical stress or tone patterns. English is a stress language (e.g., the difference between torment' [noun] and tor'ment [verb]); Chinese is a tone language (e.g., the word 'ma' can be produced with four different tones to mean 'mother,' 'hemp,' 'horse,' or 'scold').

The dominant view in psychology is that words are recognized by a process whose front end maps properties of the acoustic waveform onto phonological features and phonemes. Two literatures converge to support the significance of features in the representation of speech. A large body of work in acoustic-phonetics has identified acoustic characteristics that correspond to proposed features, and has demonstrated that systematic manipulation of these characteristics or cues leads to predictable changes in the perception of speech sounds or words. Similarly, linguists have shown that systematic variations in the sound patterns of the world's natural languages are best predicted and understood at a level of featural representation.

Based on the importance of features and feature cues in theory, the existence of phonetic feature detectors has been a subject of great interest. The strongest evidence for feature detectors comes from studies of selective adaptation effects. In selective adaptation experiments, listeners hear a series of words or syllables that share a common feature, and are then asked to make phonetic judgments about a continuum of speech sounds that differ in terms of that feature. For example, they may hear a series of syllables that begin with voiced consonants (e.g., [bu], [do], [ge]), and are then asked to categorize a series of syllables in which voicing cues are manipulated to create a continuum from [ba] to [pa]. Results show that categorization is shifted after adaptation, with listeners showing a greater likelihood of classifying ambiguous segments as being unvoiced ([pa]) (Figure 1). This result is interpreted as the result of fatigue on the part of voicing detectors. Some observers have challenged this interpretation, noting that fatigue could be occurring at many levels, including the level of raw auditory representation, feature cue, feature, syllable, or, in some cases, lexical representation.

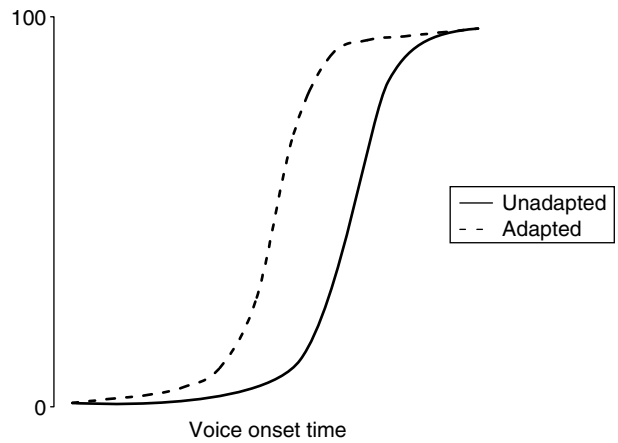

Figure 1 Schematic results showing categorical perception and selective adaptation effects in phoneme identification. Listeners are asked to identify tokens taken from a synthetic /ba/–/pa/ continuum created by increasing voice onset time. In the adaptation condition, listeners may first hear a series of consonant–vowel syllables with voiced onsets (e.g., *bu*, *do*, *ge*). Listeners in both conditions show identification curves consistent with categorical perception, but the function is shifted after adaptation as listeners show an increased bias toward perceiving segments as voiceless.

While there is little work directly examining neural responses to specific segmental features in humans, there are several studies that have looked at neural responses to phonetic feature cues in animals. In studies of the macaque, it has been shown that short voice onset time (a characteristic of segments like [b], [d], and [g]) produces a single onset transient in A1, while longer voice onset times (characteristic of voiceless segments such as [p], [t], and [k]) produce two transients. More recently, it has been found that voice onset time and aspiration (another acoustic characteristic associated with voicing) are represented in discrete tonotopic regions in A1 by synchronous activation patterns involving large populations of neurons.

Evidence from humans with normal and impaired language processing suggests that these temporospectral cues to feature identity appear to be integrated in unimodal auditory association cortex lying along the superior temporal sulcus immediately adjacent to the primary auditory koniocortex. Many language processes are lateralized, with the left hemisphere being the sole site in which they occur, but whether this is true of these early acoustic–phonetic mappings is an unsettled issue. Some researchers have suggested that the unconscious, automatic activation of features and phonemes, such as occurs as a stage in word recognition under normal conditions, occurs bilaterally, and that the dominant hemisphere is the sole site of phonemic processing that is associated with

controlled processes (such as subvocal rehearsal) and conscious processes (such as explicit phoneme discrimination and identification, making judgments about rhyme, and other similar functions). Other researchers believe that even the unconscious processes that map acoustic properties onto phonological features and phonemes in the process of normal word recognition are lateralized. The evidence for the involvement of both hemispheres in these processes is stronger for vowels than for consonants. It has been suggested that this is because the left hemisphere is specialized for perceptual categorization of acoustic signals that is based upon rapidly changing acoustic properties, though a recent authoritative review downplays this possibility on the grounds that a left hemisphere advantage is only seen for linguistic stimuli and not nonlinguistic stimuli with such temporal characteristics.

The stage of processing that follows acoustic-phonemic processing is lexical access. Most models maintain that a language user has a store of the phonological properties of words in a long-term memory structure (the 'phonological input lexicon') and that acoustic-phonemic mapping identifies features and phonemes that in turn activate these representations. Other models conceive of the long-term representations of the forms of words as distributed sets of features without discernible listings of individual words. In these models, activation of the units over which features are distributed settles into stable patterns that suffice to activate semantic and syntactic features associated with individual words. Activation of items in a phonological input lexicon is not an entirely bottom-up process, however. Prevailing models maintain that bottom-up processes are supplemented top-down in important ways by information about words and context. Indeed, listeners fill in masked gaps in the middle of words so well that they are unable to discriminate masked gaps from masked segments (the 'phoneme restoration effect').

Based on functional neuroimaging results, activation of the long-term representations of the sound patterns of words is thought to occur in the left superior temporal gyrus. Scott and her colleagues have argued that there is a pathway along this gyrus and the corresponding left superior temporal sulcus such that word recognition occurs in a region anterior and inferior to primary auditory cortex, roughly corresponding to the parabelt region of auditory cortex in the monkey, and word meanings are activated yet further along this pathway in anterior inferior temporal lobe bilaterally (Figure 2). This pathway constitutes the auditory counterpart to the visual 'what' pathway in the inferior occipital-temporal lobe, in which cells respond to features of objects

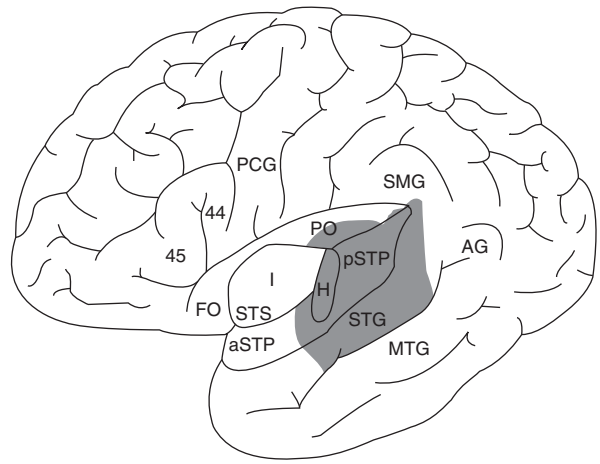

Figure 2 Depiction of the areas in the left superior temporal lobe that are involved in processing sounds and spoken words. PCG, precentral gyrus; SMG, supramarginal gyrus; AG, angular gyrus; MTG, middle temporal gyrus; aSTP and pSTP, anterior and posterior supratemporal planes; STS, superior temporal sulcus; FO, frontal operculum; PO, parietal operculum; H, transverse temporal gyrus of Heschi; STG, superior temporal gyrus. Reproduced from Hickok G and Poeppel D (2004) Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition* 92: 67–99, with permission from Elsevier.

and objects themselves. In both the auditory and visual modalities, these temporal lobe pathways extract categories that are associated with meaning, with the difference being that the forms that are extracted from auditory stimuli are words, which are associatively linked to meaning, while those extracted from visual stimuli are structural descriptions of objects, which are partially constitutive of meaning.

The combination of bottom-up and top-down processing streams in language perception is one of the factors that makes it difficult to delineate the auditory component of language processing. Speech sounds map onto words, which in turn map onto representations of meaning, syntax, discourse, and world knowledge. To the extent that activation flows both ways between these representations, higher level representations may influence activation in auditory areas. Any understanding of auditory processing in language must involve an appreciation of these connections.

Speech perception is connected to speech production, especially during language acquisition, when imitation is crucial for the development of the child's sound inventory and lexicon. On the basis of lesions in patients with repetition disorders known as 'conduction aphasia,' the neural substrate for this connection has been thought to consist of the arcuate fibers of the inferior longitudinal fasciculus, which

connect auditory association cortex (Wernicke's area in the posterior part of the superior temporal gyrus) to motor association cortex (Broca's area in the posterior part of the inferior frontal gyrus). Recent functional neuroimaging studies and neural models have partially confirmed these ideas, providing evidence that integrated perceptual-motor processing of speech sounds and words makes use of a 'dorsal' pathway separate from that involved in word recognition. Again, the parallel with perception of visually presented objects has been emphasized, with the dorsal speech pathway corresponding to the parietal-based pathway that links object recognition with object use.

Once items are identified in the phonological input lexicon, it is possible, in principle, for a great deal of subsequent language processing to be carried out on syntactic and semantic properties of words and to be far removed from what would ordinarily be considered 'auditory' processing. However, the processes that construct the forms and meanings of sentences and discourse are not as independent of auditory processing as one might imagine. The products of segmental auditory processing, believed to be based in the dorsal pathway, play an important role with respect to these levels of language through subvocal rehearsal. The exact role of subvocal rehearsal in language comprehension is debated, but there is wide agreement that it is used as at least an ancillary mechanism to keep words in an active available form, to be reconsidered by comprehension mechanisms when structures are hard to build or have to be revised.

Suprasegmental (Intonational) Processing

Intonation (or prosody) refers to the pitch contours of words and phrases. Like lexical structures, intonational structures have been characterized in abstract representational terms, of which the Tones and Boundary Indices (ToBI) system is the most widely used. The ToBI assigns various levels of height to 'tones' that indicate intonational contours, and numerical levels to types of intonational breaks between words. For instance, the sentence "I like football, hockey, and soccer" might be spoken as one intonational phrase (IP), with several smaller intermediate phrases (ips), as in $[_{IP}$ $[_{ip}$I like football], $[_{ip}$hockey], $[_{ip}$and soccer]].

Like phonemes and their features, these tones and breaks are abstract characterizations of sound, and an active field of study investigates the mapping between them and acoustic properties. Many acoustic features of ToBI notational elements have begun to be identified. Some of these features are local. For instance, longer word nucleus duration and lower fundamental frequencies tend to occur at intonation phrase boundaries compared to intermediate phrase boundaries. The higher the pitch, or the bigger the pitch rise on a word, the more intonationally prominent the word is. Sentence-final rising intonation is associated with questions rather than statements. However, many ToBI notational elements are determined by relational properties of discontinuous portions of the acoustic waveform and features. For instance, whether a boundary is an intonational or an intermediate phrase boundary is determined in part by pitch, duration, and pause features adjacent to that boundary relative to these features at other boundaries.

Listeners use intonation to obtain information about syntactic and discourse structure. At the level of syntax, intonational boundaries are relevant to how phrases are attached to one another. For instance, the sentence "When the woman washed the clothes fell off the shelf," is usually initially misanalyzed, with listeners (and readers) taking 'the clothes' to be the object of 'washed,' and then are left with an apparently ungrammatical end to the sentence. A pause (or, in writing, a comma) after 'washed' indicates the clause boundary and makes the sentence understandable. At the level of discourse, intonation provides a great deal of information, including the focus of the discourse, the illocutionary force of a sentence (statement, question), and other semantic features. For instance, the sentence "Senator Eastland doesn't grow cotton to make money" is ambiguous, conveying either the sense that Senator Eastland grows cotton for some reason other than to make money or that Senator Eastland makes money by not growing cotton. An intonational contour that places a rising and then falling tone on 'cotton' followed by an intonational phrase break after 'cotton' leads to 'cotton' being the focus of the sentence. This leads to the usually unpreferred sense that the senator makes money by not growing cotton.

The mapping between intonation and other levels of linguistic structure is complex; intonational structures tend to correlate with, but not be isomorphic to, other structures. The sentence "The man who walked the dog that chased the squirrel tripped" consists of a series of relative clauses attached to noun phrases, and its syntactic structure can be roughly sketched as $[_{S}$ $[_{NP}$ The man who walked $[_{NP}$the dog that chased the squirrel]] tripped]. Its intonational contour, however, is different, grouping the objects of each clause with their verbs ([The man who walked the dog] [that chased the squirrel] tripped]). Intonation thus gives some strong and relatively direct cues regarding higher level linguistic representations (as in the case of focus) and some less direct cues as to these structures.

The neural basis of perceiving intonation has been less intensively studied than that of word-level phonology. These studies follow the same lines as those of segmental aspects of spoken language, dividing into those that that focus on the perception of the elements that determine intonational percepts and those that focus on the mapping of intonation onto higher linguistic categories.

One topic in the first set is pitch perception. Pitch is a derived percept that stands in a complicated relationship with its acoustic counterpart, fundamental frequency, or F0. The perceived pitch that creates intonation patterns is the product of vibration of the vocal cords, which pulse energy through the vocal tract as they open and close. Pitch information may be present throughout a wide frequency range due to the presence of harmonics – bands of energy found at integer multiples of the fundamental frequency. Behavioral evidence suggests that the perception of pitch in complex tone stimuli (including language) is a constructive process involving the full harmonic structure of a signal. Evidence for this view comes from the missing fundamental phenomenon in which listeners perceive pitch even when the fundamental is removed through filtering. Interestingly, pitch perception involves anatomically dissociable mechanisms in pitch-present versus missing fundamental stimuli, suggesting the existence of multiple cortical pitch extraction mechanisms. Evidence from imaging and lesion studies in humans and animals typically points to bilateral involvement of primary auditory cortex, with a dominant role in pitch perception played by the right hemisphere. So, for example, neurons in both left and right primary auditory cortex have been found that are sensitive to pitch change and pitch direction, but behavioral tasks involving the pitch direction discrimination appear to depend on intact right hemisphere primary auditory cortex. One of the great challenges in the study of the neural correlates of pitch perception has to do with the problem of encoding. While both temporal and place encoding models of cortical pitch representation has been proposed, it is clear that neither class of models adequately relates physiological data to psychophysical data across the relevant frequency span.

Turning to the mapping of intonational percepts onto other linguistic representations, studies are scarce and results are inconsistent. One study from our lab examined the processing of intonation in the context of linguistic processing. We found that syntactic statements pronounced with question intonational contours led to increased blood oxygen level-dependent (BOLD) activity in bilateral inferior frontal and temporal regions, compared to the corresponding statements with statement intonation. Other studies, however, have found only right hemisphere activation related to intonation signaling questions, albeit in different contrasts (contrasting sentences with intonation signaling a statement or question, against sentences with the same or different tone on the final word, in Chinese). One preliminary finding of interest is that different intonational features may activate different brain regions. For instance, in one study, the areas activated by Chinese listeners making judgments about the location of contrastive stress (relevant to discourse focus) were more left lateralized than the areas activated by the same listeners making judgments about whether an intonational contour expressed a question or a statement. Many aspects of design and analysis will need to be addressed before a clear picture is likely to emerge regarding this question.

Future Directions

This brief overview indicates that many basic questions about the processing of acoustic information to yield linguistic representations remain unanswered. The application to these questions of modern investigational methods in psychology, functional neuroimaging, and neurophysiology, though in its infancy, has already added enormously to the information we have about these processes and their neural bases. Beyond these areas lie questions about closely related matters, such as the processing of indexical information, emotional state, and other percepts as separate or related processing streams. Models of the neural mechanisms that are responsible for processing the acoustic stimulus in linguistic terms are guaranteed to deepen and change as these topics are further explored.

See also: Immune System–Neuroendocrine Interactions; Magnocellular Neurosecretory System: Organization, Plasticity, Model Peptidergic Neurons; Neuroendocrine Control: Maternal Behavior; Speech Production: Adult; Speech Perception: Adult; Speech Perception: Development; Speech Perception: Neural Encoding; Speech Perception: Cortical Processing.

Further Reading

Beckman ME, Hirschberg J, and Shattuck-Hufnagel S (2005) The original ToBI system and the evolution of the ToBI framework. In: Jun S-A (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*, pp. 9–54. Oxford: Oxford University Press.

Binder JR, Frost JA, Hammeke TA, et al. (2000) Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex* 10: 512–528.

Doherty CP, West WC, Dilley LC, et al. (2004) Question/statement judgments: An fMRI study of intonation processing. *Human Brain Mapping* 23: 85–98.

Gandour J, Xu Y, and Wong D (2003) Neural correlates of segmental and tonal information in speech perception. *Human Brain Mapping* 20: 185–200.

Guenther FH and Perkell JS (2004) A neural model of speech production and its application to studies of the role of auditory feedback in speech. In: Maassen B, Kent R, Peters H, et al. (eds.) *Speech Motor Control in Normal and Disordered Speech*, pp. 29–49. Oxford: Oxford University Press.

Hickok G and Poeppel D (2004) Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition* 92: 67–99.

Johnson K (2003) *Acoustics and Auditory Phonetics,* 2nd edn. Oxford, UK: Blackwell.

Okada K, Smith KR, Humphries C, et al. (2003) Word length modulates neural activity in auditory cortex during covert object naming. *NeuroReport* 14: 2323–2326.

Patel AD (2005) The relationship of music to the melody of speech and to syntactic processing disorders in aphasia. *Annals of the New York Academy of Sciences* 1060: 59–70.

Perkell JS and Klatt D (eds.) (1986) *Invariance and Variability of Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Perkell JS, Matthies ML, Svirsky MA, et al. (1993) Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot motor equivalence study. *Journal of the Acoustical Society of America* 93: 2948–2961.

Poeppel D and Hickok G (2004) Towards a new functional anatomy of language. *Cognition* 92: 1–12.

Scott SK (2003) How might we best conceptualise speech perception? The view from neurobiology. *Journal of Phonetics* 31(3–4): 417–422.

Scott SK, Blank SC, Rosen S, et al. (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123: 2400–2406.

Scott SK and Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. *Trends in Neuroscience* 26(2): 100–107.

Scott SK and Wise RJS (2004) The functional neuroanatomy of prelexical processing of speech. *Cognition* 92: 13–45.

Stevens KN (2000) *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Tramo MJ, Cariani PA, Koh CE, et al. (2005) Neurophysiology and neuroanatomy of pitch perception: Auditory cortex. *Annals of the New York Academy of Sciences* 1060: 148–174.

Zatorre RJ, Evans AC, Meyer E, et al. (1992) Lateralization of phonetic and pitch processing in speech perception. *Science* 256: 846–849.

Zatorre RJ, Evans AC, and Meyer E (1994) Neural mechanisms underlying melodic perception and memory for pitch. *Journal of Neuroscience* 14: 1908–1919.

Language: Cortical Processes

D Caplan, Massachusetts General Hospital, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

The study of language–brain relationships is almost exclusively focused on the neural basis of the largely unconscious normal processes of speaking, understanding spoken language, reading, and writing. Data bearing these relations come from a variety of sources.

Correlating lesions with deficits provides evidence regarding brain areas that are necessary for a function. The original approach of postmortem analysis has largely been supplanted by magnetic resonance imaging (MRI), positron emission tomography (PET), and other imaging techniques. Data are also obtained by intraoperative stimulation, subdural electrode placement, and transcranial magnetic stimulation focally in normal individuals, which interfere with functions in most cases. The second approach to the study of language–brain relations is to record physiological and vascular responses to language processing in normal individuals using event-related potentials (ERPs) and magnetoencephalography (MEG), intraoperative and subdural recordings, optical imaging, and, most commonly, PET and functional MRI. Regions in which differences in neurovascular activity are associated with differences in performance on two tasks support the operations that differ in the two tasks. Functional neuroimaging in patients can reveal brain areas that are involved in an operation only postinjury.

Neural Structures Involved in Language

Beginning in the late nineteenth century, the application of deficit–lesion correlations based on autopsy to the problem of the regional specialization of the brain for language yielded important findings that remain undisputed. Chief among these is the fact that human language requires parts of the association cortex in the lateral portion of one cerebral hemisphere, usually the left in right-handed individuals. This cortex surrounds the Sylvian fissure and runs from the pars triangularis and opercularis of the inferior frontal gyrus (Brodman's area (BA) 45 and 44; Broca's area) through the angular and supramarginal gyri (BA 39 and 40) and into the superior temporal gyrus (BA 22; Wernicke's area) in the dominant hemisphere (Figure 1).

Data from other sources – deficit–lesion correlations based on antemortem neuroimaging and functional neuroimaging – have provided evidence that regions outside the peri-Sylvian association cortex also support language processing. Written language appears to involve cortex closer to the visual areas of the brain and sign language brain regions closer to those involved in movements of the hands than movements of the oral cavity. Some ERP components related to processing improbable or ill-formed language are maximal over high parietal and central scalp electrodes, suggesting that these regions may be involved in language processing. Both lesion studies in stroke patients and functional neuroimaging studies suggest that the inferior and anterior temporal lobe is involved in aspects of language processing. Injury to the supplementary motor cortex along the medial surface of the frontal lobe can lead to speech initiation disturbances; this region may be important in activating the language processing system, at least in production tasks. Activation studies have shown increased regional cerebral blood flow (rCBF) and blood oxygen level-dependent signal in the cingulate gyrus in association with many language tasks, although this activation appears to be nonspecific because it occurs in many nonlinguistic tasks as well. It has been suggested that it is due to increased arousal and deployment of attention associated with more complex tasks.

The same sources of data provide evidence that subcortical structures are also involved in language processing. Several studies report aphasic disturbances following strokes in the deep gray matter nuclei (the caudate, putamen, and parts of the thalamus). It has been suggested that subcortical structures involved in laying down procedural memories for motor functions, particularly the basal ganglia, are involved in 'rule-based' processing in language, such as regular aspects of word formation, as opposed to the long-term maintenance of information in memory, as occurs with simple words and irregularly formed words. The thalamus may play a role in processing the meanings of words. In general, subcortical lesions cause language impairments when the overlying cortex is abnormal (often, the abnormality can be seen only with metabolic scanning techniques) and the degree of language impairment is better correlated with measures of cortical than subcortical hypometabolism. It may be that subcortical structures serve to activate a cortically based language processing system but do not process language. In some activation studies involving language, the cerebellum also has increased rCBF. This may be a result of the role of this part of the brain in processes involved in timing and temporal ordering of events or because it is directly involved in language and other cognitive functions.

Ongoing studies have also clarified data regarding hemispheric dominance for language. In approximately

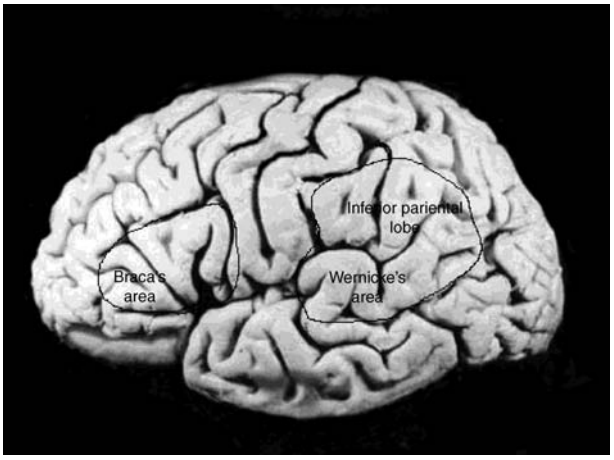

Figure 1 Depiction of the left hemisphere of the brain showing the main language areas.

98% of right-handed individuals, the left hemisphere is dominant. Approximately 60–65% of non-right-handed individuals are left-hemisphere dominant, approximately 15–20% are right-hemisphere dominant, and the remainder appear to use both hemispheres for language processing. The relationship of dominance for language to handedness suggests a common determination of both, probably in large part genetic. The neural basis for lateralization was first suggested by Geschwind and Levitsky in 1968, who discovered that part of the language zone (the planum temporale, a portion of the superior temporal) was larger in the left than in the right hemisphere. Subsequent studies have confirmed this finding and identified specific cytoarchitectonically defined regions in this posterior language area that show this asymmetry. Several other asymmetries that may be related to lateralization have been identified. The exact relationship between size and function is not known. In general, however, relative size is a good predictor of lateralization and an example of the 'bigger is better' principle that applies elsewhere to communication systems (e.g., to song nuclei in birds).

An important finding is that the 'nondominant' hemisphere is involved in many language operations. Evidence from the effects of lesions and split brain studies, experiments using presentation of stimuli to one or the other hemisphere in normal subjects, and functional neuroimaging studies indicate that the nondominant hemisphere understands many words, perhaps representing them more generally than does the left hemisphere, and suggest that it is involved in other aspects of language processing as well. Some language operations may be carried out primarily in the right hemisphere. The best candidates for these operations are ones that pertain to processing the discourse level of language, especially revising inferences, interpreting nonliteral language, and appreciating humor.

Models of Organization of the Brain for Language Processing

Two general models of the relationship of areas of the brain to components of the language processing system have been developed. Localizationist theories maintain that language processing components are localized in specific parts of the brain. 'Holist' theories maintain that linguistic representations and processes require broad areas of the brain. Five basic models, which capture the set of logically possible relations of brain areas to language processes, can be extracted from these two conceptualizations: invariant localization, variable localization, even distribution, invariant uneven distribution, and variable uneven distribution.

Invariant localization hypothesizes that only a small area of the brain supports a function. Variable localization hypothesizes that different small areas of the brain support a function in different individuals. Distribution hypothesizes that a large region of the brain supports a function. Traditional distributed models assumed an even distribution of distributed functions: All parts of the region contributed equally to the function. If a function is evenly distributed throughout a region, there can be no individual variability in its neural basis. If a function is unevenly distributed throughout a region, it may be distributed the same way in everyone (invariant uneven distribution) or differently in different individuals (variable uneven distribution). Other models are extensions of these basic five. Degeneracy is a variant of localization in which more than one structure independently supports a function; degeneracy can either be invariant (the same areas independently support the function in everyone) or variable (different areas independently support the function in different people). Variable localization could be constrained so that a function is localized more often in one area than another.

It is not possible to review all the areas of language whose neurological basis has been studied. Here, syntactic processing is discussed.

Syntactic structures determine the relationships between words that allow sentences to convey propositional information – information about thematic roles (who is initiating an action, who is receiving it, etc.), attribution of modification (which adjectives are assigned to which nouns), scope of quantification, co-reference, and other relations between words. The propositional content of a sentence conveys a

great deal of information beyond what is conveyed by words alone and is crucial to many human intellectual functions. Propositions are the source of much of the information stored in semantic memory. Because propositions can be true or false, they can be used in thinking logically. They serve the purpose of planning actions. They are the basic building blocks of much of what is conveyed in a discourse.

Unlike models of the neural basis for lexical access and lexical semantic processes, a variety of models have been proposed regarding the neural basis for syntactic processing. Of the five theories of functional neuroanatomy reviewed previously, four have been proposed regarding syntactic processing on the basis of deficit–lesion correlations: localizationist models that claim that Broca's area plays a critical role in connecting moved constituents and their antecedents; variable localization models; traditional, evenly distributed models; and invariant unevenly distributed models.

Evidence supporting these models is based on correlating deficits in syntactic comprehension with lesions. There is considerable disagreement about how to characterize these deficits. Syntactic comprehension deficits are typically established by showing that patients can understand sentences with simple syntactic structures (e.g., 'The boy chased the girl') and sentences with complex syntactic structures in which the relationships between the nouns and verbs can be inferred from knowledge of real-world events (e.g., 'The apple the boy ate was red') but not sentences with complex structures in which the relationships between the nouns and verbs depend on a syntactic structure that needs to be constructed (e.g., 'The boy the girl pushed was tall'). However, there are two issues that must be dealt with. First, it is not clear whether a patient with a performance of this type has a deficit in a particular parsing operation or has what might be called a reduction in the resources available to accomplish syntactically based comprehension. Second, there is virtually no consistency of individual patients' performances across tasks, raising questions about whether it is correct to say that a patient who fails on a particular structure has a parsing deficit.

Assuming, however, that performances such as those outlined previously on a single task are properly analyzed as deficits in particular parsing operations, the empirical basis that relates these deficits to lesions is so limited that none of the models that have been championed have been tested rigorously. The vast majority of claims about the neural basis for syntactic processing are based on patient data and do not examine lesions quantitatively. Many are based on the assumption that Broca's or nonfluent patients have 'anterior' lesions and Wernicke's, fluent, conduction and anomic patients have 'posterior' lesions, whereas the reality is far more complicated. We recently reported the most detailed study of patients with lesions whose syntactic comprehension has been assessed. Forty-two patients with aphasia secondary to left hemisphere strokes and 25 control subjects were studied for the ability to assign and interpret three syntactic structures (passives, object-extracted relative clauses, and reflexive pronouns) in enactment, sentence–picture matching, and grammaticality judgment tasks. Neither the most direct prediction made by distributed models (that lesion size correlates with performance level) nor the most commonly made claim about localization of syntactic operations (that the insertion of traces into syntactic structures occurs in Broca's area) have been tested quantitatively on the basis of radiological data. We recently began to address this question. We measured accuracy, reaction time, and self-paced listening times in picture matching and grammaticality judgment. We obtained MRI and 5-deoxyglucose PET data on 31 patients and 12 controls. The percentage of selected regions of interest that was lesioned on MRI and the mean normalized PET counts per voxel in regions of interest were calculated. In regression analyses, lesion measures in both peri-Sylvian and non-peri-Sylvian regions of interest predicted performance. Patients who performed at similar levels behaviorally had lesions of very different sizes, and patients with equivalent lesion sizes varied greatly in their level of performance. The data are consistent with a model in which the neural tissue that is responsible for the operations underlying sentence comprehension and syntactic processing is localized in different neural regions in different individuals.

Functional neuroimaging has supplanted deficit–lesion correlations as the principle source of information regarding the localization of syntactic processing in sentence comprehension. These studies have led to localizationist models. Many researchers advance models in which some aspect of parsing or sentence interpretation is localized in Broca's area or in portions of this region. Some authors have argued that the innate capacity that underlies the ability to acquire the syntax of natural language is localized in this region. However, despite the numerous advocates for localization of syntactic operations, most neuroimaging studies show that multiple cortical areas are activated in tasks that involve syntactic processing, and different areas have been activated in different tasks. Overall, these data suggest variation in the localization of the areas that are sufficient to support syntactic processing within the language area across the adult population, with perhaps some constraint on the areas

in which processing is localized as a function of the proficiency of individuals to assign syntactic structure and determine the meaning of sentences.

Conclusion

Human language is a unique representational system that relates aspects of meaning to many types of forms (e.g., phonemes, lexical items, and syntax), each with its own complex structure. Deficit–lesion correlations and neuroimaging studies are beginning to provide data about the neural structures involved in human language. It appears that many areas of the brain are either necessary or sufficient in representing and processing language, with the left peri-Sylvian association cortex being the most important such region. How these areas act to support particular language operations is not understood. There is evidence for both localization of some functions in specific regions and either multifocal or distributed involvement of brain areas in other language functions. It may be that some higher level principles are operative in this domain. For instance, content-addressable activation and associative operations such as those that underlie phoneme recognition, lexical access, and lexical semantic activation may be invariantly localized, whereas combinatorial computational operations such as those that constitute the syntax of natural language may not. However, many aspects of these topics remain to be studied with tools of modern cognitive neuroscience.

See also: Broca's Area: Evolution; Language Development; Language in Aged Persons; Language: Auditory Processes; Language: Aphasia; Reading; Sentence Comprehension; Speech Perception: Cortical Processing.

Further Reading

Annett M (1985) *Left, Right, Hand and Brain: The Right Shift Theory.* London: Erlbaum.

Beeman M, Friedman RB, Grafman J, Perez E, Diamond S, and Lindsay MB (1994) Summation priming and coarse semantic coding in the right hemisphere. *Journal of Cognitive Neuroscience* 6: 26–45.

Caplan D, Hildebrandt N, and Makris N (1996) Location of lesions in stroke patients with deficits in syntactic processing in sentence comprehension. *Brain* 119: 933–949.

Caplan D, Waters G, Kennedy D, et al. (2007) A study of syntactic processing in aphasia II: Neurological aspects. *Brain and Language* 101(2): 151–177.

Caramazza A and Zurif EB (1976) Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language* 3: 572–582.

Damasio AR and Damasio H (1992) Brain and language. *Scientific American* 1992: 89–95.

Friederici A (2002) Towards a neural basis of auditory sentence processing. *Trends in Cognitive Science* 6: 78–84.

Geschwind N (1965) Disconnection syndromes in animals and man. *Brain* 88: 237–294, 585,644.

Geschwind N and Galaburda AM (1987) *Cerebral Lateralization: Biological Mechanisms, Associations and Pathology.* Cambridge, MA: MIT Press.

Grodzinsky Y (2000) The neurology of syntax: Language use without Broca'a area. *Behavioral and Brain Sciences* 23: 47–117.

Kaan E and Swaab T (2002) The brain circuitry of syntactic comprehension. *Trends in Cognitive Sciences* 6(8): 350–356.

Kutas M and Hillyard SA (1980) Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207: 203–204.

Lashley KS (1950) In search of the engram. *Society of Experimental Biology Symposium* 4: 454–482.

Mesulam M-M (1990) Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology* 28(5): 597–613.

Mohr JP, Pessin MS, Finkelstein S, Funkenstein HH, Duncan GW, and Davis KR (1978) Broca aphasia: Pathologic and clinical. *Neurology* 28: 311–324.

Ojemann G (1983) Brain organization for language from the perspective of electrical stimulation mapping. *Behavioural and Brain Sciences* 6: 189–230.

Ojemann G and Schoenfield-McNeill J (1998) Neurons in human temporal cortex active with verbal associative learning. *Brain and Language* 64: 317–327.

Poizner H, Klima ES, and Bellugi U (1989) *What the Hands Reveal about the Brain.* Cambridge, MA: MIT Press.

Rugg MD and Coles M (eds.) (1995) *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition.* Oxford Psychology Series. Oxford: Oxford University Press.

Stromswold K, Caplan D, Alpert N, and Rauch S (1996) Localization of syntactic comprehension by positron emission tomography. *Brain and Language* 52: 452–473.

Tettamanti M, Alkadhi H, Moro A, Perani D, Kollias S, and Weiniger D (2002) Neural correlates for the acquisition of natural language syntax. *NeuroImage* 17: 700–709.

Language: Learning Impairments

J L Evans, San Diego State University, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

In this article, the characteristics of a group of children with language learning impairments are described. These children do not have problems that are usually associated with language-learning impairments, such as hearing impairment, intellectual impairments, and/or neurological damage. In the research literature, 'specific language impairment' (SLI) is the most widely used term for this group of children, and it is the one that is used here.

The criteria for SLI are quite precise. They include Performance IQ in the normal range; normal hearing; no recent episodes of otitis media with effusion; no evidence of seizure disorders, cerebral palsy, or brain lesions; no medication for control of seizures; no oral–motor structural abnormalities; and no symptoms of impaired reciprocal social interactions (e.g., autism). Typically, nonverbal IQs of these children are on average 85–100, whereas standardized language assessment scores are 1.25 standard deviations or more below one's expectations. These children pass pure-tone hearing screenings at 20 dB in both ears at frequencies of 500, 1000, 2000, and 4000 Hz. They have no focal brain lesions, traumatic brain injury, cerebral palsy, or seizures. Important is the exclusion of children who have problems with volitional control of oral–motor structures, which does not rule out mild neuromaturational delays. Interestingly, some children with SLI may be more clumsy, have slower motor responses, and have poorer fine motor skills than peers.

A final exclusion criteria is the absence of all of the symptoms listed for pervasive developmental disorder 'not otherwise specified' in the *Diagnostic and Statistical Manual of Mental Disorders*. Hence, no problems using gaze and gestures to regulate social interactions, or stereotypic, repetitive motor movements characteristic of children with autism, are seen in these children. This is not to say that they do not experience substantial difficulties in social interactions as a result of their language disorders.

The prevalence of SLI is approximately 7% of school-aged children. Although it was commonly believed that the incidence of SLI is higher in males than females, large-scale epidemiological studies now report a ratio of SLI in males to females of 1.1:1. Thus, the 'often reported' higher rate of males to females with SLI may instead be a reflection of referral bias in the public schools, the overidentification of males with SLI, and the underidentification of female children with SLI. Approximately 70% of very young children (2-year-olds) who experience delays in language production may catch up to their peers by the time they are 3 or 4 years of age; however, if a child is still classified as having language difficulties by the time he or she is in kindergarten, it is likely that he or she will continue to have language impairments throughout adolescent and adulthood. This suggests that a 'wait-and-see' approach for children beyond the age of 4 years puts them at extreme risk for pervasive and continued language impairments and may result in significant costs to health care and educational systems.

Language Abilities of Children with SLI

The work reviewed here includes only studies investigating American- and British-English-speaking children with SLI (although cross-linguistic work suggests similar findings). Children with SLI have difficulty comprehending and producing new words. Vocabulary growth is slow, and expressive vocabulary, as measured by standardized assessments, is smaller than that of age-matched peers. First words appear at older ages in children with SLI, with the average age of first word being 23 months in contrast to 12 months for typically developing children. The types of words used by children with SLI, however, mirror those of typically developing children, with names of objects/animals/substances comprising approximately 55% of their first words, words referring to actions comprising 12% of first words, and words referring to properties of objects also comprising 12%. Moreover, verb stem errors are extremely rare in children with SLI, reportedly occurring in only 7:2998 verb tokens in the speech of preschool children with SLI.

Experimental word-learning studies show that children with SLI need, on average, at least twice the number of exposures to a novel word compared to their peers before they can successfully comprehend and/or produce it. The trajectory of word learning is also qualitatively different. Children with SLI exhibit a concave word learning trajectory, with initial performance being poor but improving rapidly with increased exposure, compared to typical children, whose word learning trajectory is convex, with rapid improvement in comprehension and production accuracy after initial exposure to novel words, followed by a leveling off and stabilizing of performance. Additional semantic cues presented with a novel word increase comprehension accuracy for children with SLI, whereas phonological cues increase production

accuracy in these children but not in typical peers. Errors made by children with SLI in extending a newly learned word to a novel object mirror those of younger language-matched typical children, with the referent receiving the incorrect label when it is both perceptually and functionally similar to the target object. Studies of lexical abilities in school-aged children with SLI parallel the findings when these children were younger.

Difficulty acquiring and using grammatical morphology, including inflectional affixes such as past tense *-ed* and plural *s*, characterizes the difficulties seen in children with SLI and has been proposed as a possible phenotype marker for this disability. Typically developing children have acquired these morphological forms by the time they are approximately 4 years of age, whereas learning these forms is difficult for children with SLI, who are well into school-aged years before they learn them. On tasks designed to probe comprehension of various syntactic forms, young children with SLI are unable to identify syntax errors in sentences, such as tense marking and subject–verb agreements; have difficulty comprehending passive constructions; and rely heavily on comprehension strategies in the absence of full comprehension. Not surprising, children with SLI have more difficulty identifying syntactic violations despite easily identifying semantic violations.

In a given language, the use of grammatical morphology is highly predictable. Because of this regularity, its use is often viewed as being 'rule governed,' and children's use of these forms is an indication that they have inferred the 'rules' governing the generation of a given linguistic form. Moreover, in typically developing children, this rule can be extended to new and novel linguistic contexts, with such extensions providing critical evidence that the child has completely extracted the rule and is able to use the rule generatively (e.g.. affixing plural *s* to 'wug'). In discrimination-learning studies, in which a child is asked to extract 'regularities' from recurring stimulus presentations and then extend them to new stimulus items consistently, children with SLI are unable to extend linguistic regularities from trained to untrained forms.

The use of argument structure by children with SLI in spontaneous and elicited utterances suggests that underlying knowledge of these forms in these children may be learned through exposure to specific verbs in individual sentence constructions. Although most argument structures are evident in the speech of children with SLI, they do not consistently produce these forms. For example, children with SLI produce elements that add substantially to the information content of an utterance but which are not obligatory. Although 7- to 9-year-old children with SLI resemble typical 5-year-olds in the rate at which they use adverbial forms to express temporal relationships, they are unable to provide additional contextual information when necessary (e.g., 'the dog left' is not sufficient).

Theoretical Accounts

A critical question with regard to this population is whether SLI is a primary deficit in an independent grammar module or whether more general-purpose processing mechanisms underlie the language impairments seen in these children. Several processing mechanisms have been proposed and include but are not limited to auditory processing deficits, limited verbal working memory, slower speed of processing, and generalized limited processing capacity.

Processing capacity and speed of processing accounts view processing as trade-offs between task demands and cognitive resources. Limited capacity accounts of SLI share the tenet that children have a fixed pool of operational resources available to perform computations, and when the demands of a task exceed available resources, both processing and storage of information will be degraded. A consistent finding across both verbal and nonverbal tasks that are designed to simultaneously tax the processing and storage of information in children with SLI is that when they are required to allocate more of their working memory resources to processing information, they are left with substantially fewer resources for the storage of information compared to their unimpaired peers. Processing capacity tasks have high specificity and sensitivity in the identification of children with SLI and account for a significantly greater proportion of the variance in language abilities in SLI than standardized measures of grammatical knowledge.

Speed of processing is often viewed as a processing resource in the sense that it is finite and the faster/slower processing is performed, the better/worse the resulting level of cognitive performance will be. Children with SLI are consistently slower than their peers across a myriad of verbal and nonverbal tasks. Both large-scale meta-analysis studies and direct comparisons of reaction time (RT) data for children with SLI show that their speed of processing information is between 20 and 30% slower than that of age-matched controls. One problem in invoking a general processing mechanism account of SLI is that it is not clear whether limited capacity and/or speed of processing reflect a fundamental limitation of the underlying cognitive system or if they are themselves emergent

properties of the system, arising naturally as a consequence of interactions within the cognitive system. Limited capacity and speed of processing have been treated as separate constructs in the study of SLI.

Historically, slower reaction times in children with SLI have been interpreted as the processing of each successive step in a given task being slower by a constant proportion. By this account, the more complex the task (e.g., the greater the number of processing steps), the progressively slower reaction times will be for children with SLI. Importantly, RT data alone cannot rule out alternative slowing accounts, such as cognitive processes being slowed but speed of sensorimotor processes not being slowed, or slowing increasing proportionately with each successive processing step. Event-related potentials (ERPs) provide a means of examining speed of processing with and without higher cognitive processing. Comparisons of behavioral RTs and P300 latencies of children with SLI indicate that P300 ERP latencies are no slower than those of age-matched peers despite RTs being significantly slower. In contrast, N400 latencies, a measure of higher cognitive processing, are significantly longer in children with SLI compared to peers. Taken together, ERP data suggest that slowing in SLI may not be the primary deficit but an artifact of poorly specified and/or degraded extant knowledge in these children, requiring greater processing resources to maintain in active memory than for typically developing children.

Another hypothesized cause of SLI is impaired auditory processing. The nature of auditory processing difficulties in SLI remains open to debate, however. Some researchers have championed the idea that temporal processing deficits are the cause of SLI. Others have proposed reduced storage for phonological material in working memory or poorly specified long-term phonological representations. Studies of auditory processing accounts in children with SLI use several experimental methods, including temporal order judgment tasks, just noticeable difference tasks, phoneme identification tasks, categorical perception tasks, and mismatch negativity (MMN).

Some of the earliest studies of temporal order judgments in SLI showed that children with SLI did not differ from their peers in the interstimulus interval (ISI) needed to detect two separate tones; however, they needed much longer interstimuli durations to correctly report the order in which the tones appeared. Follow-up work has shown that children with SLI are able to successfully report the order of the steady-state vowels but not the order of the CV syllables. These data have been interpreted as a temporal processing deficit, specifically that children with SLI are unable to process rapidly changing acoustic features of speech stimuli.

In just noticeable differences tasks, children with SLI listen to pairs of stimuli and decide if they are the same or different. Typically, one of the pair is a clear exemplar of a given target stimuli, whereas the second is designed to contrast an acoustic dimension of interest. In these tasks, children with SLI require greater acoustic difference between stimulus pairs to discriminate stimuli (e.g., /ba/-/pa/ or /ba-da-ga/) compared to peers. Findings of phoneme-identification studies show the same pattern, in which children with SLI are able to easily distinguish phonemes that are spectrally dissimilar vowels [i] and [a] at both long and short durations, but they have difficulty distinguishing phonemes with greater spectral similarity (e.g., [i]-[I]). (An important finding from auditory processing work is that children with SLI who have both speech and language impairments (articulation impairments as well as language impairments) have greater difficulty with both identification and temporal order judgments.)

The extent to which children with SLI perceive speech sounds categorically has been studies extensively. What is interesting about studies of categorical perception in children with SLI is that, given the ubiquitous nature of categorical perception across species, a deficit in categorical perception in children with SLI would indicate a profound neurocognitive impairment in these children – a finding not in keeping with the perceptual and cognitive abilities seen in these children. Children with SLI consistently exhibit similar crossover points (CPs) as their peers in identification but show less-consistent labeling of unambiguous endpoint tokens and reduced discrimination. Historically, however, these studies of CP in children with SLI used synthetic speech stimuli. A series of CP studies comparing natural and synthesized speech stimuli revealed that poor speech discrimination is evident only for synthetic speech stimuli. The consistent finding of poor auditory processing under synthetic speech conditions suggests that the auditory 'perceptual' deficits in children with SLI are not the result of deficits in auditory processing *per se* but, rather, from increased processing demands inherent in synthesized speech.

MMN, a component of the auditory ERP that reflects the outcome of an automatic comparison processing between different acoustic stimuli, is a measure of sensory memory. Typically, it is elicited when a deviate stimuli occurs in a series of identical stimuli. Both pure tones and speech stimuli have been used in investigations of central auditory

processing in children with SLI. Interestingly, children with SLI have attenuated amplitudes for MMNs to speech stimuli (e.g., /ba/ /da/ /ga/) compared to age-matched peers but not for simple tone stimuli differing in either frequency (1000 vs. 12 000) or duration (175 vs. 100 ms).

Classification and Labels

The study of SLI dates back to the first half of the nineteenth century with the introduction of the term congenital aphasia (the term 'hearing mutism' was used in Germany during this same time period). Other terms used in the late 1800s and early 1900s included delayed speech development, congenital auditory imperception, and congenital verbal auditory agnosia. Interestingly, because clear neurological damage was not evident in these children, as early as the 1880s it was suggested that attention and memory were possible underlying causes of the language disorder. At the turn of the century, English and French researchers began to use the term 'congenital aphasia.' In the 1960s, terms such as developmental aphasia and congenital aphasia began to disappear, and developmental dysphasia began to be used. This shift to *dys-* reflected the differentiation of children with neurological insults and 'acquired aphasia' from children who had language impairments in the absence of brain injury. The shift also reflected the emphasis in the 1960s on detailed description of the linguistic deficit profiles of these children as well. Currently, the terms 'language/learning disabled' or 'language/learning impaired' are often used in studying the academic skills of these children, reflecting the fact that having a label of 'language impaired' does not exclude these children from being labeled as learning disabled in schools. The term 'developmental language disorder with expressive and expressive–receptive subtypes' is also used in the *Diagnostic and Statistical Manual of Mental Disorders*.

Three different approaches characterize the grouping of children with SLI: clinical, psychometric, and linguistic classifications. Clinical taxonomies of language disorders, such as that proposed by Rapin and Allen, closely parallel acquired language disorders in adults with aphasia and include phonological, syntactic, semantic, and pragmatic skills. Alternatively, Aram and Nation classify children with SLI using standardized psychometric tests assessing phonological, syntactic, and semantic aspects of language. In contrast, Crystal, Fletcher, and Garman classify children with SLI based on linguistic measures that include assessment of semantic referencing, phonology, grammar, and structure building. These different approaches have yielded not only different labels for SLI but also different theoretical accounts of the disorders, as well as different subgroupings within SLI.

Neurobiological Correlates

A common but largely unsubstantiated hypothesis in the 1950s was that left hemisphere damage was the cause of SLI. The development of computed tomography scan technology allowed for direct investigation of the hypothesis that SLI was due to left hemisphere damage. However, these studies revealed a remarkable absence of neuroradiological abnormalities in children with SLI. Because research failed to demonstrate that the language disorders seen in children with SLI stemmed from focal brain damage, researchers compared the language abilities in children with known lateralized brain lesions to those of children with SLI. Interestingly, children with unilateral brain lesions manifest profiles of language disorders that are very similar to characteristics of children with SLI. One outcome of this research was a shift in focus from models of SLI that involved left hemisphere damage to models of SLI that involved bilateral or more diffuse areas of the brain.

In unimpaired individuals with normal language abilities, the planum temporale and pars triangularis are larger in the left hemisphere than the right. Children with SLI have abnormal asymmetry of the prefrontal region (right greater than left) and abnormal symmetry in the parietal region. Moreover, the area containing the left planum is significantly smaller in children with SLI, or reverse asymmetry is seen in the region containing the planum temporale in children with SLI. Quantitative anatomical differences in the brains of children with SLI and age-matched controls shows three majors differences: Children with SLI have significantly narrower right hemispheres, have significantly smaller pars triangularis in the left hemisphere, and do not have leftward asymmetry of the pars triangularis, planum temporale, and the planum compared to their age-matched peers. Whereas the left pars triangularis is significantly smaller in children with SLI, the right pars triangularis does not differ in children with SLI and their peers.

Cerebral blood flow measures (single-photon emission computed tomography) in phoneme discrimination indicate that children with SLI do not show predominant activation of the left hemisphere. Moreover, functional magnetic resonance imaging and event-related potential (ERP) studies of toddlers at risk for SLI (e.g., late talkers), children with SLI, and even parents of children with SLI exhibit different patterns of coordinating activation among brain regions relative to controls, suggesting the possibility

of qualitatively different brain organization for language in this population.

Genetics and Heritability

There is a familial component to SLI. The concordance rate for monozygotic twins with SLI is almost 100%, and it is 50–70% for dizygotic twins. Molecular genetic studies implicate: (1) the linkage of a severe autosomal dominant speech and language disorder to the region of 7q31 (*SPCH1*) and (2) translocation in the *SPCH1* region and evidence of linkage on both chromosome 16q and 19q. Furthermore, work indicates that 16q and 19q may represent universally important loci in SLI and, more broadly, a risk factor for developmental language impairments. Importantly, in the SLI consortium study, there is no evidence for linkage to 7q near FOXP2, a gene that has been implicated in severe speech impairment and which is located in a region that may also include a major locus for autism. A study completing a genome scan for SLI used both dominant and recessive models for the three phenotype classifications: (1) clinical diagnosis, language impairment; (2) reading discrepancy; and (3) evidence for linkage between 13q21 using the recessive reading discrepancy model. Results from this study also suggest that additional loci, on 2p22 and 17q23, may comprise part of the phenotype of SLI.

SLI or Dyslexia?

One question is the extent to which SLI and dyslexia are distinct or overlapping disorders. On the one hand, SLI and dyslexia, by diagnostic criteria, are two distinct developmental language disorders, with SLI being characterized by semantics, syntax, and discourse processing difficulties, and dyslexia characterized by phonological processing deficits and problems in word reading. However, children with dyslexia exhibit deficits in semantics and syntax in addition to deficits in phoneme processing. Alternatively, children with SLI, in addition to language impairments, also exhibit substantial deficits in phonological processing. Reviews of behavioral, neurological, and genetic evidence, as well as large-scale epidemiological studies of children with SLI and dyslexia selected from a population-based sample, suggest that these two clinical populations represent two different but overlapping developmental disorders. In a large-scale population-based study, where the participant sample from which the children in the study were drawn included 527 school-aged children, the prevalence of dyslexia was examined in second, fourth, and eighth grade among children who were classified with SLI in kindergarten. A statistically significant, but limited, overlap between SLI and dyslexia existed, but the majority of the children with SLI did not have dyslexia during the school years.

An Implicit Learning Impairment?

Learning is a gradual, often unconscious process that initially is slow and inefficient, but with sufficient exposure it becomes automatic and efficient. Recent studies suggest that language impairments seen in SLI may be secondary to deficits in implicit learning. Procedural learning is the system involved in the learning of new cognitive and/or motor sequences. Grammar is also a regular pattern of sequences. Based on the idea that the lexical declarative memory system is relatively spared in children with SLI, it has been argued that the grammatical systems in these children are impaired, specifically due to neurological damage in the corticostriatum and association areas.

Research has begun to examine this directly using both statistical word learning (SWL) paradigms and a classic task of procedural learning of sequential pattern – the serial reaction time task (SRT). In SRT tasks, a child presses a button corresponding to a spatial location of a visual stimulus, trial by trial. Embedded within a random order of stimuli are patterns of locations. The learning of these embedded patterns within the sequence is evidenced by shorter reaction times of button pressing of the patterned sequence in contrast to the random sequences. Adolescent children with SLI are significantly impaired in their ability to learn the embedded sequence. Moreover, performance on the SRT task accounts for a significant amount of the variance in grammatical knowledge in these children.

Statistical word-learning paradigms also show that children with SLI are significantly impaired in the ability to implicitly discover word boundaries within continuous speech. Whereas school-aged children are able to implicitly learn novel words after only 3 min of exposure, performance of children with SLI after 20 min of exposure is at chance. Importantly, after double the exposure (40 min) to the same speech stream, the performance of children with SLI is greater than chance and no different from the performance of their peers. It is argued that implicit learning is such a primitive cognitive construct that there should be no differences across an individual's performance. However, ERP studies using the same SWL paradigm show that high word learners evidence rapid and early sensitivity to word boundaries seen in the N100 waveform

prior to demonstrating this knowledge behaviorally in posttesting, whereas poor word learners do not. Clearly, a key aspect of learning language is the tracking of the regularities in the language input. One can see how impairment in the mechanisms involved in implicitly tracking the regularities in the language input could have profound implications for language learning and suggest a single mechanism account of the pattern of deficits seen both across language domains and developmentally over time in children with SLI.

See also: Dyslexia: Neurodevelopmental Basis; Language Development; Language Following Congenital Disorders (not SLI); Language: Cortical Processes; Reading; Sentence Production; Sentence Comprehension; Speech Production: Development; Statistical Learning of Language; Word Production; Word Recognition; Word Learning.

Further Reading

Barlett CW, Flax JF, Logue MW, et al. (2002) A major susceptibility locus for specific language impairment is located on 13q21. *American Journal of Human Genetics* 71: 45–55.

Bishop DVM, Bright P, James C, Bishop SJ, and van der Lely HKJ (2000) Grammatical SLI: A distinct subtype of developmental language impairment? *Applied Psycholinguistics* 21: 159–181.

Bishop DVM and Leonard LB (eds.) (2000) *Speech and Language Impairments in Children: Causes, Characteristics, Intervention and Outcome*. Hove: Psychology Press.

Catts HW, Adlof SM, Hogan TP, and Ellis Weismer S (2005) Are specific language impairment and dyslexia distinct disorders? *Journal of Speech, Language, and Hearing Research* 48: 1378–1396.

Coady JA and Evans JL (2007) The uses and interpretations of nonword repetition tasks in children with and without specific language impairment. *International Journal of Communication Disorders* 43: 1–40.

Coady JA, Evans JL, Mainela-Arnold E, and Kluender KR (2007) Children with specific language impairment perceive speech most categorically when tokens are natural and meaningful. *Journal of Speech, Language, and Hearing Research* 50: 41–57.

Ellis Weismer S, Evans J, and Hesketh L (1999) An examination of verbal working memory capacity in children with specific language impairment. *Journal of Speech, Language, and Hearing Research* 42: 1249–1260.

Evans JL, Viele K, Kass RE, and Tang F (2002) Grammatical morphology and perception of synthetic and natural speech stimuli in children with specific language impairments. *Journal of Speech, Language, and Hearing Research* 45: 494–504.

Gauger LM, Lambardino LJ, and Leonard CM (1997) Brain morphology in children with specific language impairment. *Journal of Speech, Language, and Hearing Research* 40: 1272–1284.

Hugdahl K, Gundersen H, Brekke C, Thomsen T, Rimol LM, and Ersland L (2004) fMRI brain activation in a Finnish family with specific language impairment compared with a normal control group. *Journal of Speech, Language, and Hearing Research* 47: 162–172.

Leonard L (1998) *Children with Specific Language Impairment*. Cambridge, MA: MIT Press.

Leonard L, Ellis Weisner S, Miller C, Francis D, Tomblin B, and Kail R (2007) Speed of processing, working memory, and language impairment in children. *Journal of Speech, Language, and Hearing Research* 50: 408–428.

O'Brien EK, Zhang S, Nishimura J, Tomblin B, and Murray C (2003) Association of specific language impairment (SLI) to the region of 7q31. *American Journal of Human Genetics* 74: 1536–1543.

Plante E, Swisher L, Vance R, and Rapcsak S (1991) MRI findings in boys with specific language impairment. *Brain and Language* 41: 52–66.

Rice ML and Warren SF (eds.) (2004) *Developmental Language Disorders: From Phenotypes to Etiologies*. Mahwah, NJ: Erlbaum.

SLI Consortium(2002) A genomewide scan identified two novel loci involved in specific language impairment. *American Journal of Human Genetics* 70: 384–398.

Tomblin JB, Smith E, and Zhang X (1997) Epidemiology of specific language impairments: Prenatal and perinatal risk factors. *Journal of Communication Disorders* 30: 325–343.

Tomblin JB, Zhang X, Buckwalter P, and O'Brien M (2003) The stability of primary language disorder: Four years after kindergarten diagnosis. *Journal of Speech, Language, and Hearing Research* 46: 1283–1296.

Language: Nonhuman Animals

D M Rumbaugh and E S Savage-Rumbaugh, Great Ape Trust of Iowa, Des Moines, IA, USA
J P Taglialatela, Yerkes National Primate Research Center, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Traditionally, language is considered to be a uniquely human characteristic. However, in recent years a number of significant research reports strongly suggest basic language functions in the communication systems of a variety of mammalian and avian species. These reports justify reconsideration of language as being unique to humans and continued study of its foundations in learning and culture. At the core of these communicative systems, including human language, is the production and transmission of signals, as well as the perception, processing, and appropriate response to the signals of others.

All human languages are said to be made up of a finite set of arbitrary, discrete, and largely meaningless elements that, when combined, refer to external events or entities not necessarily present or tangible. Novel meaning can be constructed by recombining these elements based on rules, and similar or identical combinations of elements can maintain single or multiple meanings. Messages need not be honest, accurate, or meaningful. Users of a language learn their languages culturally from more experienced counterparts and retain the potential to learn new meanings or entirely different systems throughout their life span.

These features of human language have been used and misused in an attempt to evaluate the linguistic competencies of various nonhuman species. Still, the features do accurately, if not incompletely, characterize many human languages. It should not be surprising, therefore (at least in an evolutionary sense), that the features describe a number of nonhuman animal communication systems as well. Thus, the central question underlying much of the animal language research during the past century is whether or not human language is fundamentally and comprehensively different from other communicative systems.

The study of the capacity for language in nonhuman animals, particularly the great apes, has contributed richly to our understanding of human language as well as its phylogeny and evolution. One compelling perspective that has emerged as the result of ape language research is that the foundations of language are in the functional exchange of relatively complex information between two or more interactants. Therefore, rather than an 'all or none' phenomenon, language is more accurately viewed as a suite of capacities that evolved as an accretion and integration of competencies for communication and cognition.

In the early twentieth century, a number of studies were initiated that sought to determine whether or not apes were capable of acquiring and speaking human language. However, none succeeded, in large part due to anatomical constraints of their vocal tracts. In the mid-1960s, Allen and Beatrix T Gardner renewed interest in the field with their Project Washoe, aimed at establishing two-way communication between humans and a chimpanzee (*Pan troglodytes*). Because they were familiar with past failures with regard to vocal production on the part of the ape subjects, the Gardners chose American Sign Language (viewed widely as a natural language system) as a means to communicate with their female subject, Washoe. Although Washoe did not acquire signs solely through observation, she was able to learn a relatively large number of signs through training. Researchers would shape Washoe's hands into the correct positions and reward her for successive improvements. This method proved to be a great success. By 3 years of age, Washoe reliably produced 85 signs in appropriate contexts, used signs to communicate novel information in new logically relevant contexts, and combined various signs together. Washoe and other apes with her seemed to master the meaning of the gestures, and they invented ways to express themselves through both the combination of signs and the invention of novel signs.

At approximately the same time, Project Sarah was initiated by David Premack, who used tokens of varying shapes and colors to function as words in an attempt to conduct an experimental analysis of language functions. Sarah, a chimpanzee that was 5 years old at the study's onset, was trained to associate the arbitrary tokens with real-world referents and then to arrange the magnetized pieces on a metal board. Sarah was able to learn these associations, but the work was criticized for its noncommunicative nature. Sarah's behavior, rather than indicative of communicative linguistic ability, could be described adequately as complex problem solving or matching-to-sample skills.

A second project, known as the Language Analogue (LANA) Project, was initiated by Rumbaugh and colleagues in the early 1970s. A computer-based research system provided Lana, a young female chimpanzee, with a keyboard that held 125 keys, each embossed with a distinctive geometric symbol called a

lexigram. Each lexigram functioned as a word, and lexigram use was monitored by a computer. In response to various strings of lexigrams, the computer was programmed to dispense foods or drinks or to activate a number of devices that might present movies or slides on a screen, play music, open a window, invite persons into Lana's area, and so on. Lana readily learned to employ her lexigrams to request items and to respond to experimenter queries for the names and/or colors of 36 test objects, composed of six items in six different colors. Lana also initiated goal-directed conversations and innovatively modified stock sentences that she had been taught in order to facilitate her conversations and to achieve goals other than the specific ones for which the lexigrams were designated. For example, Lana used them to attract attention to malfunctioning food vendors and other devices. Comprehensive analyses indicated that even Lana's innovative use of lexigrams in novel strings gave strong evidence of the grammar demanded by the system's computer. Her use of lexigrams could not be attributed satisfactorily to simple learned sequences or to imitation.

Similar research with other great ape species has demonstrated that gorillas and orangutans can acquire and effectively use derivatives of human signed languages. The fact that it has been shown that both gorillas and orangutans have the potential for human linguistic competency may suggest that some basic competencies were present in the common ancestor of all apes and humans.

All of the research on the linguistic competencies of great apes had focused on the subjects' abilities to produce communicatively appropriate utterances (e.g., gestures, sounds, and lexigrams). However, relatively little attention was paid to whether or not (or the degree to which) the apes comprehended the symbols. In other words, it was still unknown, or at best relatively unclear, if the apes truly understood that the symbols they were so adeptly producing, and in some cases combining, were meaningful and with semantic foundations. Did they function as valid symbolic references for actual items and events in the real world?

In order to address this issue directly, Savage-Rumbaugh began work with two co-reared male chimpanzees named Sherman and Austin. Using a variation of Lana's keyboard, the focus was on symbol-based communication and cooperation between the apes and not necessarily between ape and experimenter.

The categorization skills of Sherman and Austin were examined further in an experiment designed to determine if they demonstrated the capacity to classify correctly 17 different lexigrams into the categories of 'food' or 'tool.' (The lexigrams were geometric symbols that identified various foods and tools, such as banana, apple, wrench, key, and straw.) The chimpanzee subjects could not look at the actual foods or tools when making their decision but only at each lexigram, in turn. Sherman and Austin were taught on only three exemplars each of food and tool. They were not taught how the other 17 test word-lexigrams would be classified. They could make this determination only by looking at each word in turn, recalling the actual item represented by that word, and then determining the function of that item – whether they ate it or used it to open puzzle boxes. Their performance was perfect on these controlled test trials with but one error. Clearly, their symbols had solid semantic foundations.

Sherman and Austin were taught all aspects of their language – how to request food/drinks, how to name/label items without ever receiving them, and even how to comprehend what the symbols represented (e.g., the lexigrams were embossed on the keys of their computer-monitored keyboards). After learning how to request food/drinks, they required additional training to learn how to label those items (e.g., to name them and not receive them).

It is of note that through extensive experience on tasks in which they had to coordinate their behaviors through the use of symbols, they began to employ symbols much as people do – that is, to announce their intended actions. For example, after viewing a random array of food and drinks, they would make a choice, state the name of it on the keyboard, and then give it to the experimenter – as though the experimenter had presented the object to them to be named.

This was a major milestone in the history of ape language research. Not only had strong scientific data been provided indicating nonhuman animals were able to comprehend human symbols but also the work made two additional significant contributions. First, because the apes were using a lexigram keyboard (composed of discrete visual symbols), the subjects' messages were unequivocal. Second, the subjects were (in experimental settings) communicating with one another, not human researchers. Therefore, criticisms regarding imitation were not relevant to this work.

Subsequent research with other apes built on these insights as well as the belief that the fundamental component of language was comprehension, not necessarily production. Accordingly, instead of attempting to train apes to use language, methods focused on cultivation of their comprehension.

Research by Savage-Rumbaugh and colleagues with the bonobo (*Pan paniscus*) Kanzi was the

first to demonstrate a great ape's ability to learn the meanings of symbols spontaneously. All previous ape subjects mastered their vocabularies (e.g., manual gestures, tokens, and lexigrams) as a result of specific training procedures. In contrast, Kanzi acquired lexigram competencies (and spoken English comprehension) observationally – at first, while observing his mother's training sessions and then through interactions with caregivers and researchers. Researchers working with Kanzi would use spoken language in conjunction with the lexigram keyboard during communicative interactions with Kanzi as well as with one another in Kanzi's presence. They would discuss individuals, items, and events in their environment. Kanzi was also given access to a lexigram keyboard, and in time he began to use the keyboard in conjunction with his other communicative modalities to participate in conversations with his human counterparts. In this way, Kanzi's language acquisition more closely resembled that of a typical human child – less through formal teaching and much more through observation and interaction.

Kanzi's comprehension of spoken English was evaluated experimentally to determine how thoroughly Kanzi understood the requests that were made of him. To ensure that Kanzi was responding to the words he heard, and not to context clues or cues from experimenters, requests were spoken to Kanzi by an experimenter located in another room. Kanzi responded appropriately to approximately three-fourths of more than 600 novel requests presented to him verbally in rigidly controlled experimental conditions. His performance was comparable to that of a 2½-year-old child who was included in the study for comparison. Neither Kanzi nor the human child had the benefit of people modeling the requests, nor had they been trained to do what was requested of them.

The work with Kanzi has been replicated with both a female bonobo and a female chimpanzee, who similarly acquired comprehension of spoken English and lexigrams as well as the production of lexigram utterances observationally. This collection of research is fundamentally different from previous studies in that the focus was on language comprehension and not production. Accordingly, rather than training apes to use language, methods focused on cultivation of their comprehension of speech and lexigram utterances. Caretakers and researchers spoke to and with the apes about events and things in their environment, and they concomitantly utilized the lexigram keyboard to supplement their spoken words. This integration of speech with lexigram use by researchers has resulted in the apes similarly supplementing their use of lexigrams on their keyboard with other communicative modalities, such as vocalizations and gestures, to participate in dialogues with their human counterparts.

This approach revealed the following:

- Apes possibly have a sensitive period in early infancy during which they can come to comprehend hundreds of spoken words uttered by their caretakers.
- The apes' speech understanding is not limited to single words but, rather, can include the syntax of novel spoken sentences of request at levels that compare favorably with those of a 2½-year-old child (typical sentences from the set of more than 600 included 'Go get the lettuce that's in the microwave,' 'Give the doggie some carrots,' 'Take the mushrooms outdoors,' 'Now, could you wash your ball?' and 'Put the tomato on the blanket.').
- In this type of setting, apes will spontaneously begin to produce rule-based utterances by using the word-lexigram keyboard system and gestures. Such productions are novel, incorporate a simple grammar, and are spontaneous.
- Apes can also comprehend 'reversed sentences' (e.g., 'Make the doggie bite the snake' vs. 'Make the snake bite the doggie'), indicating that they can acquire the basic principles of syntactically based word order. In addition, comprehension of sentences referring to items not immediately present was demonstrated in response to sentences such as 'Go outside and get some flowers' or 'Go outdoors and find the carrot.' Retrieval of nonpresent items was comprehended accurately even when a similar referent was present (e.g., retrieval of the carrot that was outside when there was also a carrot indoors).
- The apes participate in communicative interactions with humans, and analysis of these exchanges reveals patterns and regularities in the apes' utterances. These interactions share characteristics with human conversation because the apes participate in turn-taking and respond to and produce requests.
- In these situations, the apes employ vocalizations as a regular part of their communicative repertoire and produce structurally distinct vocal utterances that vary predictably with the semantic context in which they are produced.
- During these complex communicative interactions, the apes are able to map new words onto novel items after only a few exposures to the items, and they do so without visual contact with the items.

The initial question in the field of ape language research regarding 'the ability that apes have for language' in terms of the human standard of speech (i.e., phrase and sentence construction) led to the misguided, if not basically incorrect, conclusion that

apes do not and cannot have language, on the one hand, and that language is a uniquely human attribute, on the other hand. Alternatively, contemporary language research with apes, focused on the parameters of symbol acquisition and rearing methods in relation to their competent use, has served to

- elucidate the evolutionary and ontogenetic roots of language;
- provide training materials and techniques which greatly benefit children who, because of developmental disabilities, have language deficiencies; and
- reveal that if there are absolute 'basics' to language competence, they are probably (1) the ability to symbolize – that is, to come to use symbols to represent things not necessarily present in time/space – and (2) the ability to use learned symbols to communicate information that otherwise could not be exchanged via the unlearned modes of communication.

Early Environment and the Importance of Logic-Structures

Early environmental stimulation can have general facilitating effects on development, but it is in the logic structure of early environment that the specific basic dimensions and interests of cognition and competence are defined. Logic structure refers to recurring patterns of communication, language use, music, movement patterns as in dancing, acrobatics, etc. A corollary of this important principle is that the specific effects of the logic structure probably depend on brain size and complexity: early environment is probably much more critical to the cognitive development of children and apes than it is to monkeys and prosimians. Segerdahl, Fields, and Savage-Rumbaugh advance the significance of culture and context in both the acquisition and the comprehension of language and that language is to be understood adequately with the culture past and present.

Principles of Continuity

As a result of the research with bonobos, particularly with Kanzi, a noted comparative psychologist, Michael Domjan, concluded that the psychological, as well as the biological, continuity postulated between animals and humans has been confirmed. The child linguists Elizabeth Bates, Donna Thal, and Virginia Marchman also support the continuity perspective of language. The shattering of the "wall that separates man from chimpanzee" presents a challenge to our thinking, "but [we] believe that the world will be the better for it."

Neuroanatomical Continuities

One salient feature of language processing in the human brain is lateralization of function to one hemisphere compared to the other. According to the classical model popularized in the second half of the nineteenth century, two cerebral cortical areas in the left hemisphere are most commonly associated with language functions: a receptive region for the processing and integration of auditory sensory information commonly referred to as Wernicke's area and a productive region concerned with the encoding of vocal signals into meaningful words and sentences referred to as Broca's area. In other words, Broca's area functions primarily in the planning and execution of speech, whereas Wernicke's area functions to 'make sense' of the speech that a listener perceives. Humanlike left hemisphere neuroanatomical asymmetries have also been identified in both posterior temporal lobe and inferior frontal regions of the chimpanzee brain – regions considered homologous to Broca's and Wernicke's areas, respectively. Furthermore, asymmetry of the chimpanzee inferior frontal gyrus (the location of the Broca's area homolog) is associated with hand use during gestural communication. These results suggest that both a biological and a behavioral continuity exists between human language and the communicative systems of great apes.

Conclusion

Intensive research during the past 30 years has documented that apes

- have the ability to learn and use symbols to represent things/events not necessarily present, which is the basis of semantics;
- have the ability to use learned symbols among themselves and/or with humans to solve problems via the exchange of information;
- readily organize their learned symbols into conceptual categories (e.g., foods, tools, people, and conspecifics);
- acquire language optimally not by formal training but, rather, through daily experiences garnered during infancy; and
- comply with basic rules of grammar and comprehend novel sentences that they hear – sentences that have their meanings syntactically embedded.

Language research with animals has made fundamental contributions to knowledge and has caused us to reassess long-standing assertions and questions that pertain to claims regarding language and human uniqueness. Are humans unique? Yes, but probably

no more so than most other species of animal life. Are humans a 'creature apart' from other species with regard to advanced cognitive competencies? The answer clearly is "Perhaps more in degree than in kind."

See also: Broca's Area: Evolution; Language Evolution; Language Development; Language: Cortical Processes; Primate Communication: Evolution.

Further Reading

Benson J, Fries P, Greave W, Iwamoto K, Savage-Rumbaugh S, and Taglialatela JP (2002) Confrontation and support in bonobo–human discourse. *Functions of Language* 9(1): 1–33.

Brakke KE and Savage-Rumbaugh ES (1995) The development of language skills in bonobo and chimpanzee – I. Comprehension. *Language and Communication* 15: 121–148.

Brakke KE and Savage-Rumbaugh ES (1995) The development of language skills in Pan – II. Production. *Language and Communication* 16: 361–380.

Greenfield P and Savage-Rumbaugh ES (1993) Comparing communicative competence in child and chimp: The pragmatics of repetition. *Journal of Child Language* 20: 1–26.

Hillix WA and Rumbaugh DM (2004) *Animal Bodies, Human Minds: Ape, Dolphin, and Parrot Language Skills*. New York: Kluwer/Plenum.

Hurley S and Nudds M (2006) *Rational Animals?* Oxford: Oxford University Press.

Pate JL and Rumbaugh DM (1983) The language-like behavior of Lana chimpanzee: Is it merely discrimination and paired-associate learning? *Animal Learning and Behavior* 11(1): 134–138.

Rumbaugh DM (ed.) (1977) *Language Learning by a Chimpanzee: The LANA Project*. New York: Academic Press.

Rumbaugh DM, King JL, Beran MJ, Washburn DA, and Gould K (2007) A salience theory of learning and behavior. *International Journal of Primatology* 28: 973–996.

Rumbaugh DM and Savage-Rumbaugh ES (1994) Language in comparative perspective. In: Mackintosh NJ (ed.) *Animal Cognition*, pp. 307–328. San Diego: Academic Press.

Rumbaugh DM and Washburn DA (2003) *The Intelligence of Apes and Other Rational Beings*. New Haven, CT: Yale University Press.

Savage-Rumbaugh ES (1986) *Ape Language: From Conditioned Response to Symbol*. New York: Columbia University Press.

Savage-Rumbaugh ES, Murphy J, Sevcik RA, Brakke KE, Williams SL, and Rumbaugh DM (1993) Language comprehension in ape and child. *Monographs of the Society for Research in Child Development* 58: 1–242.

Taglialatela JP, Cantalupo C, and Hopkins WD (2006) Gesture handedness predicts asymmetry in the chimpanzee inferior frontal gyrus. *NeuroReport* 17(9): 923–927.

Taglialatela JP, Savage-Rumbaugh S, and Baker LA (2003) Vocal production by a language-competent bonobo, *Pan paniscus*. *International Journal of Primatology* 24(1): 1–17.

Large Conductance Calcium-Activated Potassium Channels

A Alioua, P Kundu, R Lu, Y Kumar, J Ou, E Stefani, and L Toro, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Large conductance voltage- and calcium-activated potassium channels are tetramers of the pore-forming α subunit, $K_{Ca1.1}$. This class of channels is highly selective for K^+, has large single channel conductance (~120 pS in physiological solutions), and is widely distributed throughout the nervous system. Membrane depolarization and increases in intracellular Ca^{2+} concentration – two parameters which are under dynamic change during a neuronal action potential – allosterically activate (open) $K_{Ca1.1}$ channels (also known as MaxiK and BK), leading to their participation in the control of neuronal excitability. $K_{Ca1.1}$ channels also play important physiological roles in organs containing smooth muscle and contribute to the control of blood pressure, uresis, and erectile function, among others.

In the tetrameric channel, each $K_{Ca}1.1$ subunit is encoded by a single gene (KCNMA1) which is capable of generating multiple splice variants (i.e., isoforms) yielding distinct phenotypes. Also, $K_{Ca}1.1$ channel phenotype can be modified by regulatory β subunits (β_1–β_4), with β_4 being abundant in brain. For efficient signaling, native $K_{Ca}1.1$ channel has been shown to form complexes by tight association with a variety of partner proteins.

The physiological implication of the $K_{Ca}1.1$ channel in the nervous system has been further appreciated by studying mice knockout models of the α ($K_{Ca}1.1^{-/-}$) or β subunit ($\beta^{-/-}$) genes. Furthermore, some types of neuronal disorders seem to be associated with inherited mutations in the KCNMA1 gene or decreased expression of $K_{Ca}1.1$ channel protein, as in schizophrenia and autism.

This article discusses advances with regard to $K_{Ca}1.1$ channel structure, distribution, and *in vivo* physiological roles, with emphasis on the nervous system.

Structural Correlates of $K_{Ca}1.1$ Channel Function

Native $K_{Ca}1.1$ currents were first recorded in 1981 from chromaffin and muscle cells. Since then, these channels have been found in a wide variety of cell types, including neurons, microglia, and astrocytes. As mentioned previously, the minimal functional arrangement of $K_{Ca}1.1$ channels allowing K^+ flow consists of four α subunits, which may not be identical due to splice variation of the same gene, thus resulting in a large number of channel isoforms with potentially distinct functions. In addition, association with regulatory β subunits, with signaling partners, and/or posttranslation modifications can increase channel diversity and function (Figure 1).

$K_{Ca}1.1$ Subunit

$K_{Ca}1.1$ was first cloned from *Drosophila* (d*Slo*) and subsequently from a number of other species, including human, mouse, and rat. Human $K_{Ca}1.1$ protein is encoded by the KCNMA1 gene, which contains 27 constitutive exons and multiple alternative exons spanning approximately 769 kb of human chromosome 10 (10q22.3) (Figure 1(a)). The translation product of constitutive exons is highly conserved among mammals, with approximately 98% amino acid sequence identity encoding for a protein of approximately 125 kDa (although splice variation can give rise to numerous isoforms). $K_{Ca}1.1$ primary structure exhibits significant similarity to signature regions of other α subunits of the voltage-dependent potassium channel superfamily, Kv. However, unlike the α subunit of other voltage-dependent potassium channels, $K_{Ca}1.1$ monomer has seven (S0–S6) instead of six (S1–S6) transmembrane domains, with four additional hydrophobic (S7–S10) segments located in its long intracellular C-terminus. Based on structure–function studies, $K_{Ca}1.1$ can be visualized as comprising four modules contributing to distinct functions in the tetrameric channel: the extracellular N-terminus and S0 to β subunit coupling, S1–S4 to voltage sensing, S5–S6 to the potassium selective pore, and the intracellular C-terminus to Ca^{2+}-sensing/signal transducing (Figure 1(b)).

N-Terminus and S0 The N-terminus of the $K_{Ca}1.1$ protein is exposed extracellularly followed by the S0 transmembrane segment. Both domains – the extracellular N-terminus and S0 – are essential for the $K_{Ca}1.1$ increase in apparent Ca^{2+} sensitivity by the β_1 subunit. This property can be modified by the length of the NH_2 terminus and the presence of splice inserts in the C-terminus of $K_{Ca}1.1$. Whether these structures are also involved in $K_{Ca}1.1$ functional association with the neuronal β_4 subunit is unknown. The length of the NH_2 terminus can vary as $K_{Ca}1.1$ mRNAs predict translation products from any of four potential start codons in exon 1 (corresponding to the protein's methionines 1–4, M1–M4; Figure 1(a)). It is unknown whether native proteins are translated from the first initiation codon (ectopic expression has been

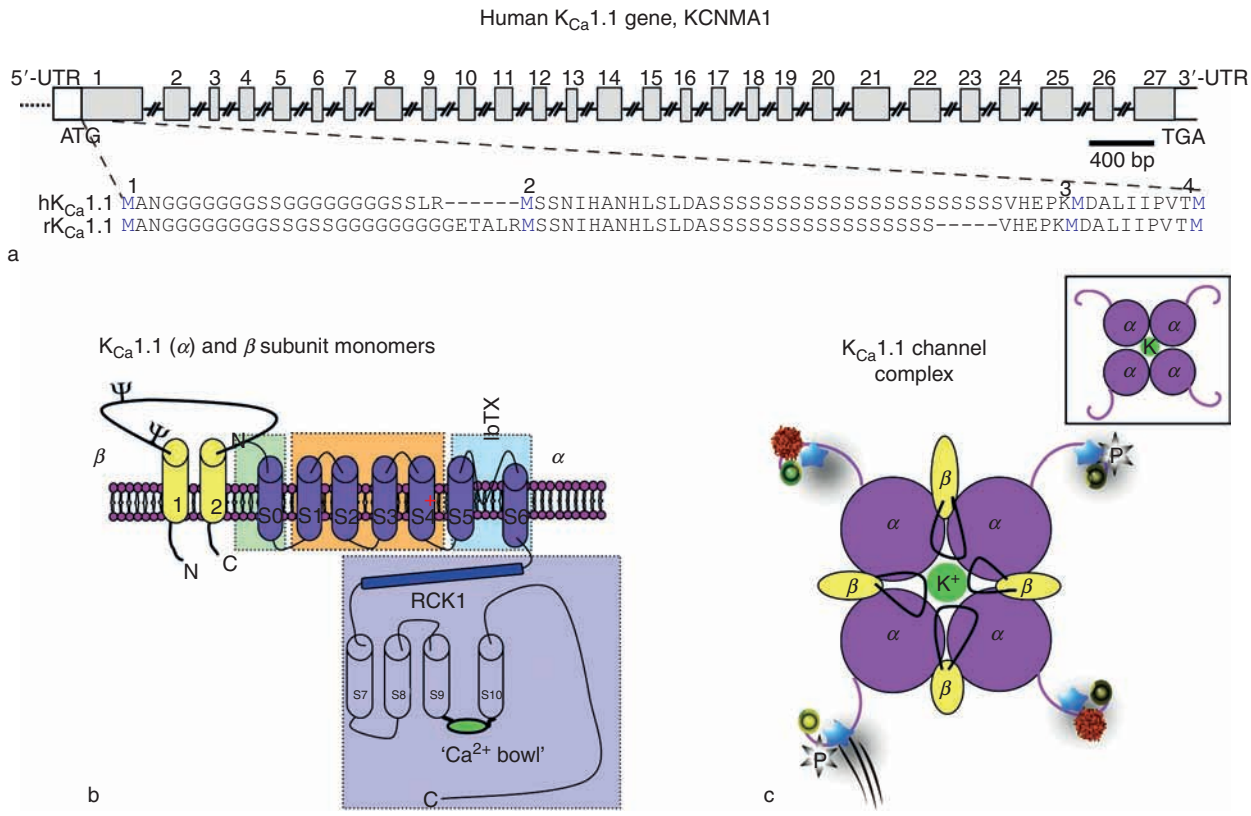

Figure 1 Gene structure and proposed membrane topology of K_{Ca}1.1 and associated β subunits. (a) Human gene KCNMA1, encoding for K_{Ca}1.1 protein, is composed of 27 constitutive exons. Three possible translation start methionines are shown (M1–M4). Gray boxes, coding sequences. (b) Topology of K_{Ca}1.1 (α subunit) and β subunits. Functional modules of K_{Ca}1.1 protein showing the N-terminal β subunit interactive site (N-terminus and S0), the voltage-sensing domain (S1–S4), the potassium selective pore (S5–S6), and the Ca^{2+}-sensing/signal transducing C-terminus. Iberiotoxin (IbTX) binds to the pore region; RCK1 (regulator of K^+ conductance 1) and Ca^{2+} bowl domains are involved in Ca^{2+} sensing. (c) A functional K_{Ca}1.1 channel is composed by the assembly of four (α) subunits (inset) able to associate with regulatory (β) subunits. The extracellular loop of β subunits projects toward the conducting pore of the channel. K_{Ca}1.1 partners associate with the channel mainly via its C-terminus (top view).

reported from the second and third but not the first initiation codon; a GC-rich sequence may hinder translation from the first initiation codon). Proteomic analysis of immunopurified K_{Ca}1.1 has provided experimental evidence that in rat brain, the native protein can start from M2.

Voltage-sensing domain In K_{Ca}1.1, the S1–S4 transmembrane segments form the voltage-sensing domain as in other voltage-gated ion channels, and charged residues in the S4 transmembrane segment together with residues in S2 and S3 serve as the gating charges. The K_{Ca}1.1 voltage sensor can couple membrane depolarization to pore opening without the need for Ca^{2+}; likewise, at negative membrane potentials where voltage sensor activation is minimized, the channel open probability increases in response to increasing intracellular Ca^{2+} concentrations, indicating that Ca^{2+} and voltage activation are two separate but interdependent processes.

Pore region In analogy to crystal structures of K^+ channels, the ion conduction pore region of K_{Ca}1.1 is formed by S5, S5–S6 linker (pore helix), and S6 α helices arranged in a tetrameric structure. The pore helix contains the conserved amino acids GYG responsible for K^+ selectivity, and a ring of aspartic acids in both inner and outer vestibules of the pore facilitates K^+ flux underlying its large conductance. The pore helix also contains the receptor for binding of the specific blocker iberiotoxin. The specificity of iberiotoxin binding toward the K_{Ca}1.1 channel has made this toxin an invaluable pharmacological tool to decipher the channel tissue distribution and to determine its physiological roles.

C-Terminus The intracellular C-terminus of K_{Ca}1.1 carries an extra four hydrophobic segments, S7–S10. This domain is approximately two-thirds of the total protein and contains multiple regulatory sites: the 'Ca^{2+} bowl' and regulator of conductance for K^+,

'RCK,' domains, which carry Ca^{2+}-sensing residues (Figure 1(b)); a Mg^{2+} sensor within the RCK domain; and multiple phosphorylation sites for cAMP- and cGMP-dependent protein kinases, for protein kinase C, and for tyrosine kinases. At least two leucine zipper domains and several protein–protein binding motifs allowing $K_{Ca}1.1$ to interact with a growing list of surrounding signaling partners have been identified. In addition, plasma membrane sorting sequences are also present in the $K_{Ca}1.1$ long carboxyl domain, indicating that the $K_{Ca}1.1$ channel possesses intrinsic signals that define its targeting.

Regulatory β Subunits

$K_{Ca}1.1$ channel regulatory β subunits (β_1–β_4) are encoded by distinct genes (KCNMB1–4) which may undergo splicing. All β subunits share a common topological structure: two transmembrane domains with both an N- and a C-terminus that are intracellular, a large extracellular loop, two or three N-linked glycosylation sites, and four conserved extracellular cysteine residues which may form disulfide bridges (Figure 1b). Functionally, co-assembly of the $K_{Ca}1.1$ channel with β subunits affects the channel biophysical and pharmacological properties. All four subunits alter $K_{Ca}1.1$ channel scorpion toxin blockade (i.e., charybdotoxin and iberiotoxin) to different degrees, with β_4 conferring high resistance to toxin block. β_1 and β_2 greatly increase $K_{Ca}1.1$ channel apparent Ca^{2+} sensitivity, β_2 and β_3 variants produce channel inactivation, and β_4 may modify $K_{Ca}1.1$ channel function in a phosphorylation-dependent manner. β_1 and β_2 subunits can also downregulate channel surface expression by favoring $K_{Ca}1.1$ endocytosis.

In contrast to $K_{Ca}1.1$, which displays a ubiquitous expression, β subunits seem to be expressed in a tissue-specific manner. In this respect, β_2 and, in particular, β_4 are enriched in the nervous system, showing differential distribution of transcript expression. At the protein level, β_2 (hippocampal neurons) and β_4 (spinal cord and astrocytes) have been detected in the central nervous system as measured by $K_{Ca}1.1$ channel electrophysiological/pharmacological properties and by proteomic analysis of coimmunopurified α and β subunit complexes from whole brain. In addition, molecular biology methods (expression of green fluorescent protein under the control of β_4 promoter in the $\beta_4^{-/-}$ mouse) predict the highest expression of β_4 protein in the dentate gyrus of the hippocampus. Thus, at least in certain regions of the brain, $K_{Ca}1.1$ channels are formed by a combination of $4K_{Ca}1.1(4\alpha)/\beta_2$ and/or $4K_{Ca}1.1(4\alpha)/\beta_4$ subunits, although the precise α:β stoichiometry or localization of subunits is still a matter of research. The following sections discuss particular characteristics of $K_{Ca}1.1$ channels (four α subunits) in association with β_2, β_3, or β_4 subunits.

β_2 (KCNMB2) In general, $K_{Ca}1.1$ currents are non-inactivating, as observed in smooth muscle and most types of neurons. However, in some cells, such as hippocampal CA1 neurons, fast-inactivating (A-type) $K_{Ca}1.1$ currents are present. This inhibition is Ca^{2+} and voltage dependent, and it can be removed by trypsin applied to the intracellular part of the channel. The molecular basis of this inactivating phenotype is the association of the $K_{Ca}1.1$ channel with the β_2 subunit(s). Similar to native channel properties, heterologous co-expression of both subunits reproduces channel fast-inactivation properties due to a depolarization-induced block of the channel by the first 19 N-terminal amino acids of the β_2 subunit. In addition to inducing fast inactivation of the $K_{Ca}1.1$ channel and similar to the β_1 subunit, the β_2 subunit enhances apparent Ca^{2+} sensitivity (induces a negative shift in the voltage activation curves), induces slower activation and deactivation kinetics, reduces the sensitivity toward charybdotoxin blockade, and catalyzes endocytosis of the $K_{Ca}1.1$ channel protein. β_2 promotes $K_{Ca}1.1$ endocytosis via two leucine-based signals in its C-terminus, LL and ΨL, where Ψ is a bulky hydrophobic residue.

β_3 (KCNMB3) Four N-terminal splice variants of different lengths termed β_3a–d can be generated from a single gene. β_3a–d mRNAs have been detected in spleen, liver, kidney, pancreas, adrenal glands, stomach, and testis; however, in brain, their transcripts are barely detectable. Isoforms β_3a–c induce inactivation of $K_{Ca}1.1$ to different degrees, with β_3b producing inactivation that is so rapid that currents appear to activate fast and rectify at low intracellular calcium. β_3a and -c, but not β_3b or -d, induce a positive shift in the voltage activation of the channel, which is opposite the effect of β_2. Remarkably, none of the β_3 isoforms increase the apparent Ca^{2+} sensitivity of the channel, and β_3d does not seem to affect any biophysical properties of $K_{Ca}1.1$. Thus, its functional role remains to be determined. Interestingly, a single base pair deletion mutant that yields a β_3b truncated protein with impaired inactivation has been proposed to associate with an increased risk of idiopathic generalized epilepsy.

β_4 (KCNMB4) Because of its striking transcript abundance in the brain, the β_4 subunit has been called the 'neuronal' β subunit. The β_4 subunit confers a key pharmacological phenotype to $K_{Ca}1.1$ channels – a profound resistance to blockade by scorpion toxins

iberiotoxin and charybdotoxin, compared to the reduced sensitivity conferred by β_1 and β_2 subunits. Apparently, the long extracellular loop of β_4 is responsible for greatly slowing charybdotoxin and iberiotoxin associations with its receptor in the $K_{Ca}1.1$ channel pore. This hindrance can be experimentally overcome by increasing toxin concentration (from the nanomolar to the micromolar range) and exposure times (from a few minutes to several tens of minutes) because the toxin association rate is dependent on its concentration. In addition to toxin-resistant channel phenotypes, β_4 produces $K_{Ca}1.1$ channels with slower activation and deactivation kinetics and also has a dual effect on apparent Ca^{2+} sensitivity: decreasing Ca^{2+} sensitivity at intracellular Ca^{2+}concentrations ($[Ca^{2+}]_i$) $<10\,\mu M$ and increasing it at $[Ca^{2+}]_i > 10\,\mu M$.

In summary, the neuronal $K_{Ca}1.1$ channel tetrameric structure can be composed of a combinatorial assembly of $K_{Ca}1.1$ subunit isoforms resulting from alternative splicing; this diversity is enhanced by potential heteromultimeric co-assembly with regulatory β subunits (β_2 and β_4) and association with other protein partners that modify channel function. A major challenge is to dissect the complex structural combinations present in different types of neurons that underlie the distinct physiological roles of the $K_{Ca}1.1$ channel in regulating neuronal excitability.

$K_{Ca}1.1$ Channel and Associating Partners in Mammalian Brain

Functional as well as immunochemical studies have reported a tight coupling/proximity between diverse signaling molecules (e.g., protein kinases and phosphatases) and the neuronal $K_{Ca}1.1$ channel. The long carboxyl domain of its α subunit, $K_{Ca}1.1$, which encompasses two-thirds of the protein, contains motifs for protein–protein interactions, and therefore the C-terminus of the channel may serve as a recognition module. Indeed, the list of proteins that interact with the mammalian $K_{Ca}1.1$ C-terminus is growing, although the mechanisms of association (direct or indirect or via scaffolding/adapter proteins) are only beginning to emerge.

Association with Ca^{2+} Channels

$K_{Ca}1.1$ from rat brain plasma membranes or microsomes physically interacts with voltage-gated calcium channels (Cav) of the L (Cav1.2; α1C), P/Q (Cav2.1; α1A), N (Cav2.2; α1B), and T (Cav3.2; α1H) types and modulatory Cavβ subunits (Cavβ1b, Cavβ2, and Cavβ3). The $K_{Ca}1.1$ L-type Ca^{2+} (Cav1.2; α1C) channel association may involve the presence of two other partners: β_2-adrenergic receptor and A-kinase-anchoring protein (AKAP150). The close association between $K_{Ca}1.1$ and Cav detected by biochemical means is also supported by electrophysiological measurements which estimate a separation of approximately 10 nm between these two types of channels. This proximity is critical because it enables $K_{Ca}1.1$ channels to be near Ca^{2+} sources that increase local Ca^{2+} to levels required for its activation. In other words, Ca^{2+} influx during an action potential via Cav channels can provide the $K_{Ca}1.1$ channel with micromolar levels of intracellular Ca^{2+} for its robust activation in restricted subplasmalemmal microdomains. This mechanism can act as a negative feedback to restrict neurotransmitter release and serve to avoid unnecessary increases in global intracellular Ca^{2+}, which would otherwise be lethal for the cells. The precise sites in $K_{Ca}1.1$ mediating association with Cavα or Cavβ subunits or the kinds of associations in different types of neurons remain to be identified.

Association with Presynaptic and Synaptic Proteins

$K_{Ca}1.1$ channels are abundantly expressed at both presynaptic and postsynaptic terminals. In these neuronal compartments, $K_{Ca}1.1$ can be visualized in close proximity with established markers of pre- and postsynaptic structures such as syntaxin-1A and PSD95, respectively. Also, in biochemical studies, $K_{Ca}1.1$ has been found associated with both proteins in brain. In particular, the synaptic membrane protein syntaxin-1 interacts with $K_{Ca}1.1$ through its C-terminus and S0–S1 loop. The interaction, which is Ca^{2+} independent, has functional consequences because it leads to reduced activation of the Ca^{2+} channel by depolarization. Interestingly, another partner of $K_{Ca}1.1$, the N-type Ca^{2+} channel (Cav2.2; α1B), also associates with syntaxin-1A by means of the intracellular loop connecting domains II and III of the channel but in a Ca^{2+}-dependent manner, being maximum near the threshold of neurotransmitter release. However, the Cav2.2 II–III loop/syntaxin-1A/$K_{Ca}1.1$ trimeric complex may not exist because syntaxin-1 was unable to simultaneously associate with the Cav2.2 II–III loop and $K_{Ca}1.1$ (although they associated independently) as if they competed for the same domain in syntaxin-1. Thus, it is conceivable that $K_{Ca}1.1$ association with syntaxin-1 may result in reduced neurotransmitter release by displacing the Cav2.2 II–III loop from syntaxin-1 and in increased excitability by reducing channel activity. The distinct Ca^{2+} dependency of syntaxin-1A association with $K_{Ca}1.1$ and N-type Ca^{2+} channels may further fine-tune the regulation of transmitter release at presynaptic terminals.

Association with Other Proteins

Ankyrin-repeat family protein (ANKRA) ANKRA interacts directly with $K_{Ca}1.1$ *in vitro*, in rat cerebral cortex, and in cultured hippocampal neurons. The ANKRA–$K_{Ca}1.1$ interaction, which has functional consequences, involves the ANKRA N-terminus with the C-terminus of $K_{Ca}1.1$. ANKRA modifies activation and deactivation kinetics of $K_{Ca}1.1$ currents in a Ca^{2+}-dependent manner; however, steady-state voltage- and Ca^{2+}-dependent activation are unaltered. At submicromolar $[Ca^{2+}]_i$ (resembling resting conditions), ANKRA accelerates channel closing within milliseconds, whereas at micromolar $[Ca^{2+}]_i$ (resembling local Ca^{2+} during bursts of action potentials), ANKRA slows down channel closing and accelerates channel activation supporting channel activity. Thus, from a physiological standpoint, ANKRA–$K_{Ca}1.1$ interaction seems well suited to efficiently control neuronal excitability. One can expect that by promoting channel opening at high local Ca^{2+}, ANKRA potentiates the hyperpolarizing effect mediated by $K_{Ca}1.1$; likewise, by favoring channel closing at low local Ca^{2+}, ANKRA keeps the channel silent at rest.

Cereblon Cereblon (CRBN; a member of the ATP-dependent Lon protease) is an enzyme implicated in memory and learning. Interestingly, CRBN and $K_{Ca}1.1$ are both highly expressed in the same areas of the brain, including the hippocampus, which is known for its involvement in memory and learning. These two proteins bind directly to each other via the C-terminus of $K_{Ca}1.1$ *in vitro*, and their proximity in brain has been confirmed biochemically and by visualization of immunolabeled complexes in cultured hippocampal neurons. Ectopically expressed CRBN suppresses $K_{Ca}1.1$ surface expression by disrupting the formation of tetramers and therefore functional channels. The physiological consequences of the interaction between CRBN and $K_{Ca}1.1$ remain to be explored, although inherited mutations of the human cereblon gene (CRBN) or $K_{Ca}1.1$ lead to mental retardation disorders.

Kinases and phosphatases Protein kinase A (PKA), protein phosphatase 2A (PP2A), and the Ca^{2+}/calmodulin phosphatase, calcineurin, can associate with $K_{Ca}1.1$. The PKA catalytic subunit associates with $K_{Ca}1.1$ via a leucine zipper motif located downstream the RCK domain that facilitates complex assembly. PP2A association with $K_{Ca}1.1$ is variant specific (associates with the C-terminal variant $K_{Ca}1.1$_DEC but not with C-terminal variants $K_{Ca}1.1$_ERL and $K_{Ca}1.1$_VYR), and direct protein–protein interaction with calcineurin's β subunit occurs within sequences flanked by the S7–S9 hydrophobic domains of the $K_{Ca}1.1$ C-terminus. Because phosphorylation/dephosphorylation can affect $K_{Ca}1.1$ channel activity in a variant-specific way, it is critical to determine cell-specific $K_{Ca}1.1$ variant expression patterns and their fine-tuning by specific kinases/phosphatases.

Cytoskeleton $K_{Ca}1.1$ channel activity is increased by disruption of the actin cytoskeleton, underscoring their physical/functional coupling. In fact, $K_{Ca}1.1$ channels associate directly with proteins that form complexes with the cytoskeleton, such as β-catenin, microtubule-associated protein 1A, and cortactin. In particular, cortactin binds to actin and is considered to play a central role in connecting signaling pathways with cytoskeleton dynamics. Cortactin binds directly to a noncanonical SH3 binding motif present in the $K_{Ca}1.1$ C-terminus and serves as a bridge to couple $K_{Ca}1.1$ channel activity to the actin cystoskeleton since $K_{Ca}1.1$ with an SH3 binding domain mutated is not activated by disruption of actin cytoskeleton. In brain, a protein complex assembled by actin/cortactin/$K_{Ca}1.1$ has been detected biochemically, and in proximal dendrites of cultured hippocampal neurons cortactin and $K_{Ca}1.1$ have been visualized in close proximity.

It is evident that the $K_{Ca}1.1$ channel has a multiplicity of domains that allow associations with different kinds of signaling/adapter partners (Figure 1c). A major goal is to dissect the localization and composition of molecular complexes as well as their dynamics in different brain cells to explain specific functional roles of the $K_{Ca}1.1$ channel in physiology and pathophysiological settings.

$K_{Ca}1.1$ Distribution in the Brain

$K_{Ca}1.1$ is expressed throughout the central nervous system in a differential way depending on the region. For example, using radiolabeled iberiotoxin – a high-affinity and specific $K_{Ca}1.1$ channel blocker – as a ligand, synaptic membranes from frontal cortex exhibit an approximately sevenfold higher channel density (190 fmol/mg protein) than the hypothalamus (29 fmol/mg protein). In brain slices, both mRNA and protein levels (evaluated with antibodies) also support regional differences in $K_{Ca}1.1$ distribution, with very high to high overall protein expression in the cerebellum, hippocampus, and cerebral cortex and low expression in the hypothalamus and midbrain (except in the interpeduncular nucleus, in which there is very high expression). In finer detail, confocal

and electron microscopy imaging have shown $K_{Ca}1.1$ protein in somatodendritic areas, axons, and presynaptic and postsynaptic terminals. This versatile subcellular distribution predicts $K_{Ca}1.1$ diverse functional roles in different types of neurons and in different regions of the same cell.

Physiological Roles of $K_{Ca}1.1$ Channels in the Nervous System

$K_{Ca}1.1$ channels are known to participate in the control of neurotransmitter release at prejunctional sites in mammalian tracheal smooth muscle and at the neuromusculoskeletal junction and rod synapses of amphibians. In mammalian central neurons, $K_{Ca}1.1$ channels are involved in action potential repolarization and generation of fast afterhyperpolarization (fAHP); their role in neurotransmitter release is supported by the visualization of the α subunit, $K_{Ca}1.1$, in synaptic terminals and by functional studies in hippocampus organ culture. Much of our knowledge of the previously mentioned physiological roles of $K_{Ca}1.1$ channels derives from the usage of iberiotoxin and paxilline, two specific inhibitors of $K_{Ca}1.1$ channels, and their induced modifications in the corresponding electrical responses. These drugs can also provide an indication of the molecular composition of the channels. Iberiotoxin at nanomolar concentrations and paxilline at micromolar concentrations readily inhibit channels formed by $4K_{Ca}1.1$ or $4K_{Ca}1.1/\beta_2$ subunits within seconds. However, unlike paxilline, 100 nM iberiotoxin is unable to inhibit $K_{Ca}1.1$ channels assembled by $4K_{Ca}1.1/\beta_4$ subunits within the same experimental time frame. Iberiotoxin can only inhibit $4K_{Ca}1.1/\beta_4$ if used in the micromolar range and for long periods of time. This type of pharmacology has been found in dorsal horn neurons in which activation of NMDA receptors coupled to 'iberiotoxin-resistant' $K_{Ca}1.1$ channels inhibits opioid release possibly via a mechanism whereby Ca^{2+} influx through NMDA receptors opens adjacent $K_{Ca}1.1$, leading to hyperpolarization and inhibition of opioid release.

On the other hand, $K_{Ca}1.1$ currents sensitive to nanomolar concentrations of iberiotoxin are found in CA1 and CA3 pyramidal neurons of the hippocampus, predicting a $4K_{Ca}1.1$ or $4K_{Ca}1.1/\beta_2/\beta$? multimeric assembly. In CA1 pyramidal neurons, $K_{Ca}1.1$ current is transient (likely assembled by $4K_{Ca}1.1/\beta_2$ subunits) and its inactivation contributes to frequency-dependent spike broadening and decreased fAHP with high-frequency discharge. In agreement, iberiotoxin broadens the action potential of the first two spikes and eliminates the fAHP of the first spike. However, a role in neurotransmitter release is less clear in this cell type because iberiotoxin is unable to increase the probability of transmitter release, as evident by its failure to decrease the paired-pulse facilitation ratio of presynaptic fibers in stratum radiatum. This is surprising because $K_{Ca}1.1$ protein is visualized near the release sites. In CA3 pyramidal neurons, $K_{Ca}1.1$ channels contribute to repolarization and fAHP of their soma action potential but have no apparent functional role in action potentials from their presynaptic axon terminals (as assessed by iberiotoxin effects or lack thereof). These properties highlight different roles of $K_{Ca}1.1$ channels in distinct regions of the same cell (Figure 2).

Null Mice Models

α subunit null mice, $K_{Ca}1.1^{-/-}$ Three neuron-related physiological roles have been attributed to $K_{Ca}1.1$ using null mice: cerebellar function and motor coordination, hearing, and control of circadian output.

Motor coordination $K_{Ca}1.1^{-/-}$ mice display a cerebellar ataxic phenotype with deficit in locomotion, motor coordination, and eye-blink reflex which is in agreement with the large amount of $K_{Ca}1.1$ in the cerebellum of wild-type animals. At the cellular level, $K_{Ca}1.1^{-/-}$ induced severe suppression of the intrinsic electrical properties of Purkinje cells: a reduction of fAHP (Figure 2, inset), a depolarized membrane potential (approximately −45 mV), and firing rates approximately 6 times lower than those of wild-type cells. However, the repolarization phase of action potential in Purkinje cells was marginally affected in $K_{Ca}1.1^{-/-}$ (although it was affected by channel blockade in wild-type cells; Figure 2, inset), suggesting that compensatory mechanisms restore this function in the $K_{Ca}1.1^{-/-}$ cerebellum. The reduced firing rates in $K_{Ca}1.1^{-/-}$ could be explained by a tonic inactivation of Na^+ channels due to the depolarized membrane potential. In general, these findings underscore a critical role of $K_{Ca}1.1$ channels in fAHP in Purkinje cells. Consistent with this idea, immunocytochemistry shows that $K_{Ca}1.1$ protein is highly expressed in soma and dendrites of this class of cells in wild-type animals.

Hearing In mammals, sound waves reaching the ear induce a mechanical vibration that excites hair cells in the cochlea. The hair cells transduce the mechanical vibrations into electrical signals and release neurotransmitter onto the afferent nerve terminals, which then reach the brain. There are two distinct types of hair cells – outer hair cells (OHCs) and inner hair cells (IHCs). $K_{Ca}1.1$ localizes in the supranuclear part of IHCs and in efferents to the OHCs, and its specific role in mammalian hearing is under debate. Two

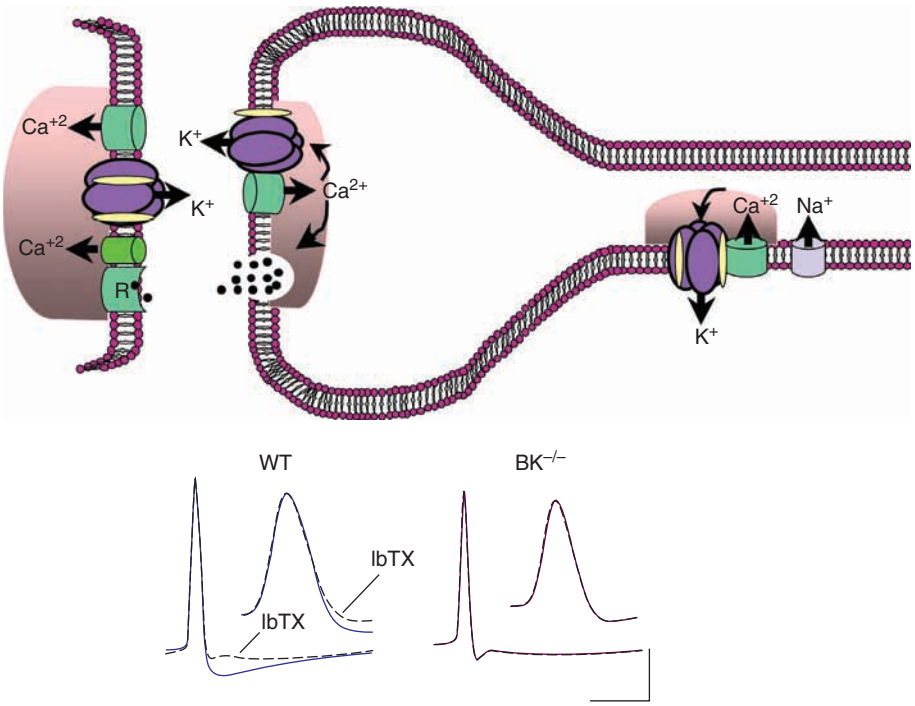

Figure 2 Schematic representation of neuronal pre- and postsynaptic terminals. Coupling of the $K_{Ca}1.1$ channel with Na^+ and Ca^{2+} channels at axonal, presynaptic, or postsynaptic localizations will have functional consequences on action potentials and neurotransmitter release. (Traces) Role of the $K_{Ca}1.1$ channel in repolarization of the action potential (AP) and fast afterhyperpolarization (fAHP, inserts) as observed by altering these components with iberiotoxin (IbTX) and by $K_{Ca}1.1$ gene ablation ($BK^{-/-}$). Scale bar = 2 ms/20 mV (fAHPs) and 0.5 ms/30 mV (APs). Reproduced with permission from Sausbier M, Hu H, Arntz C, *et al.* (2004) Cerebellar ataxia and Purkinje cell dysfunction caused by Ca^{2+}-activated K^+ channel deficiency. *Proceedings of the National Academy of Sciences of the United States of America* 101: 9474–9478. Copyright (2004) The National Academy of Sciences.

independent models of $K_{Ca}1.1^{-/-}$ have been utilized with IHCs that display normal morphology and synaptic innervation. However, in one of these models, hearing loss occurred progressively with age (>8 weeks) at high but not low frequencies and was linked to OHC degeneration and KCNQ4 ion channel reduction. A second $K_{Ca}1.1^{-/-}$ model did not display hearing loss at low or high frequencies at 12 weeks of age but showed increased resistance to noise-induced hearing loss. Nevertheless, both $K_{Ca}1.1^{-/-}$ models support a role of $K_{Ca}1.1$ in the primary afferent neuron because there was an increase in the first peak latency in the auditory brain stem responses (representing transmission through the IHC–afferent dendrite synapse), a slowed IHC membrane time constant, and spike rates of afferent auditory nerve fibers were reduced. It is possible that $K_{Ca}1.1$ has several roles in the hearing process that are underscored differentially in these knockout models for unknown reasons. Interestingly, neither β_1 nor β_4 ablation induced any changes in hearing, in contrast to the role claimed for the β_1 subunit in nonmammalian cochlea.

Circadian output In mammals, circadian behavior is controlled by the activity of suprachiasmatic nucleus (SCN) neurons of the hypothalamus. These neurons have intrinsically timed, approximately 24-h circadian electrical discharges, which are important for defining the circadian output. During daylight, rat, mouse, and hamster SCN neurons have high action potential frequency and are depolarized, whereas during the night their activity is drastically decreased. Consistent with a role of $K_{Ca}1.1$ in contributing to spike frequency rhythms of mouse SCN, its transcripts and protein levels oscillate with approximately twice the expression during dark compared to light cycles, and spike frequency rhythms are practically abolished with iberiotoxin. In addition, $K_{Ca}1.1^{+/+}$ but not $K_{Ca}1.1^{-/-}$ null mice follow the bimodal electrical states of SCN. $K_{Ca}1.1^{-/-}$ shows a clear increase in SCN firing rates during the night when electrical activity of these neurons is practically null in wild-type animals. Because mice are nocturnal, the role of $K_{Ca}1.1$ in human SCN circadian activity may be different.

β_4 subunit null mice, $\beta_4^{-/-}$ Despite the high expression of β_4 transcripts in brain, deletion of this subunit does not cause severe damage to brain function. Nevertheless, consistent with high expression of β_4 in the

dentate gyrus of the hippocampus – a structure that prevents abnormal high-frequency input from neocortex to the hippocampus – mice lacking the β_4 subunit ($\beta_4^{-/-}$) display temporal lobe seizures. Dentate gyrus granule cells of $\beta 4^{-/-}$ mice display a gain-of-function phenotype as $K_{Ca}1.1$ acquires fast gating activity and the density of functional channels per membrane patch increases. These characteristics coincide with increased firing frequencies and seizures in this animal model, whereas the slow gating activity of the wild-type $K_{Ca}1.1/\beta_4$ channel complex correlates with prolonged burst durations which would protect against seizures in normal animals. Thus, the β_4 subunit can be regarded as protective from temporal lobe seizures. These findings led to the hypothesis that pathological gain of function of $K_{Ca}1.1$ channels would favor epileptic bursting behavior. In line with this view, in cortical neurons of the epileptic mouse, fast-activating $K_{Ca}1.1$ channels are responsible for epileptic bursting behavior, and in humans a missense mutation that increases $K_{Ca}1.1$ channel activity leads to a syndrome of generalized epilepsy.

Neuroprotective Role

Results from gene silencing models suggest that $K_{Ca}1.1$ and β_4 subunits play moderate roles in the nervous system. Studies suggest that the roles of the neuronal $K_{Ca}1.1$ channel may be more prominent under pathological situations such as ischemia to improve cell survival. In brain cortical slices, $K_{Ca}1.1$ channel activity was silent under normal conditions and appeared only when neocortical pyramidal neurons started to deteriorate. Also, in an *in vitro* model of ischemia, blockade of the $K_{Ca}1.1$ channel by iberiotoxin deprived the cells from a neuroprotective mechanism, and they tended to die as a consequence. Therefore, it has been speculated that $K_{Ca}1.1$ channels may function under pathological states to provide protection to neurons.

Neurovascular Coupling

It is well known that blood flow, carrying nutrients and oxygen, increases to active brain regions due to microvascular dilation. The cellular and molecular mechanisms underlying this important physiological phenomenon are attributed to the coupling between neurons, astrocytes, and smooth muscle of the microvascular cerebral arteries. Active cortical neurons release the neurotransmitter glutamate, which in turn activates metabotropic glutamate receptors (mGluRs) in astrocytes. Activated mGluRs increase intracellular Ca^{2+} and initiate waves of Ca^{2+}, which propagate to the end-feet of astrocytes, activating astrocytic $K_{Ca}1.1$ channels and inducing K^+ efflux via these channels. Local increases in K^+ concentration in the perivascular space activate inward rectifier K^+ channels in the smooth muscle cells, producing hyperpolarization and dilation of cerebrovascular vessels. The mechanisms involved in different neurovascular beds may be distinct because in the retina light-evoked vasodilations seem not to be related to efflux of K^+ from glia with a subsequent increase in inward rectifying currents.

Channelopathies

Alteration in the function and expression of $K_{Ca}1.1$ channels may lead to human disease, including neuropsychiatric and neurodegenerative disorders. A hereditary gain-of-function mutant correlates with coexistent generalized epilepsy and paroxysmal dyskinesia, a loss-of-function mutant is linked to autism, and decreased expression in the brain has been found in schizophrenic patients.

Conclusion

Since the cloning of the first genes encoding for the mammalian $K_{Ca}1.1$ channel in the early 1990s, major advances have been realized with regard to the molecular structure, distribution, and physiological function of $K_{Ca}1.1$ channels in the nervous system. A major challenge is to elucidate the subproteome of $K_{Ca}1.1$ channels in distinct types of neural cells and visualize their subcellular distributions and dynamics as they contribute to specific functions. Identification of neuronal disorders associated with the $K_{Ca}1.1$ channel due to inherited mutations in the $K_{Ca}1.1$ protein and the possible role of the $K_{Ca}1.1$ channel as a neuroprotective mechanism make this channel an excellent candidate in the development of drugs to prevent/ameliorate neurological disorders.

See also: Calcium Channel and Calcium-Activated Potassium Channel Coupling; Inwardly Rectifying Potassium Channels; Ion Channel Localization in Cell Bodies and Dendrites; Potassium Channel Regulation; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Brenner R, Chen QH, Vilaythong A, Toney GM, Noebels JL, and Aldrich RW (2005) BK channel beta4 subunit reduces dentate gyrus excitability and protects against temporal lobe seizures. *Nature Neuroscience* 8: 1752–1759.

Du W, Bautista JF, Yang H, et al. (2005) Calcium-sensitive potassium channelopathy in human epilepsy and paroxysmal movement disorder. *Nature Genetics* 37: 733–738.

Filosa JA, Bonev AD, Straub SV, et al. (2006) Local potassium signaling couples neuronal activity to vasodilation in the brain. *Nature Neuroscience* 9: 1397–1403.

Gribkoff VK, Starrett JE Jr., and Dworetzky SI (2001) Maxi-K potassium channels: Form, function, and modulation of a class of endogenous regulators of intracellular calcium. *Neuroscientist* 7: 166–177.

Gu N, Vervaeke K, and Storm JF (2007) BK potassium channels facilitate high-frequency firing and cause early spike frequency adaptation in rat CA1 hippocampal pyramidal cells. *Journal of Physiology* 580: 859–882.

Lu R, Alioua A, Kumar Y, Eghbali M, Stefani E, and Toro L (2006) KCa1.1 channel partners: Physiological impact. *Journal of Physiology* 570: 65–72.

Meera P, Wallner M, and Toro L (2000) A neuronal beta subunit (KCNMB4) makes the large conductance, voltage- and Ca^{2+}-activated K^+ channel resistant to charybdotoxin and iberiotoxin. *Proceedings of the National Academy of Sciences of the United States of America* 97: 5562–5567.

Meredith AL, Wiler SW, Miller BH, et al. (2006) BK calcium-activated potassium channels regulate circadian behavioral rhythms and pacemaker output. *Nature Neuroscience* 9: 1041–1049.

Orio P, Rojas P, Ferreira G, and Latorre R (2002) New disguises for an old channel: KCa1.1 channel beta-subunits. *News in Physiological Sciences* 17: 156–161.

Runden-Pran E, Haug FM, Storm JF, and Ottersen OP (2002) BK channel activity determines the extent of cell degeneration after oxygen and glucose deprivation: A study in organotypical hippocampal slice cultures. *Neuroscience* 112: 277–288.

Ruttiger L, Sausbier M, Zimmermann U, et al. (2004) Deletion of the Ca^{2+}-activated potassium (BK) alpha-subunit but not the BKbeta1-subunit leads to progressive hearing loss. *Proceedings of the National Academy of Sciences of the United States of America* 101: 12922–12927.

Sausbier M, Hu H, Arntz C, et al. (2004) Cerebellar ataxia and Purkinje cell dysfunction caused by Ca^{2+}-activated K^+ channel deficiency. *Proceedings of the National Academy of Sciences of the United States of America* 101: 9474–9478.

Sausbier U, Sausbier M, Sailer CA, et al. (2006) Ca^{2+}-activated K^+ channels of the BK-type in the mouse brain. *Histochemistry and Cell Biology* 125: 725–741.

Wallner M, Meera P, and Toro L (1999) Molecular basis of fast inactivation in voltage and Ca^{2+}-activated K^+ channels: A transmembrane β-subunit homolog. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4137–4142.

Zarei MM, Song M, Wilson RJ, et al. (2007) Endocytic trafficking signals in KCNMB2 regulate surface expression of a large conductance voltage and Ca^{2+}-activated K^+ channel. *Neuroscience* 147: 80–89.

Large Dense Core Vesicles (LDCVs)

T Kögel and H-H Gerdes, University of Bergen, Bergen, Norway

© 2009 Published by Elsevier Ltd.

Introduction

Large dense core vesicles (LDCVs), also known as secretory granules (SGs), store and release neuropeptides and hormones such as neuropeptide Y, enkephalins, the adrenocorticotrophic hormone, or insulin in response to various stimuli. The release involves a Ca^{2+}-dependent fusion of LDCVs with the plasma membrane (PM) and is termed regulated exocytosis. LDCVs were discovered by electron microscopy more than 50 years ago as membrane-enclosed spherical organelles of up to several hundred nanometers in diameter characterized by an electron-dense core. Presynaptic terminals of neurons display only few LDCVs (in contrast to the numerous translucent small synaptic vesicles, which contain classical neurotransmitters), whereas endocrine cells contain high numbers of LDCVs (Figure 1). For this reason, LDCVs have been studied mainly in endocrine cells and, accordingly, the data summarized here primarily refer to insights obtained from these cells. However, it is thought that the major physicochemical and functional properties of LDCVs of endocrine cells also apply to their neuronal counterpart. Numerous studies during the past few decades have revealed that SGs originate from the *trans*-Golgi network (TGN) as short-lived vesicular intermediates. These intermediates, termed immature SGs (ISGs), go through a complex maturation process resulting in mature SGs (MSGs) (Figure 2). The process includes propeptide and prohormone processing, condensation of cargo, as well as the removal of proteins and membrane not destined for MSGs. Furthermore, MSGs undergo priming and docking to achieve competence for exocytosis. Upon appropriate stimulation of the cells, SGs fuse with the PM to release their content (Figure 2(a)).

During the past decade, the exploitation of fluorescent proteins, advanced *in vitro* systems, and mouse models has led to new insights in the biogenesis, transport, and exocytosis of LDCVs. This article summarizes both basic knowledge and new developments and concepts in the field of LDCV research.

Markers and Diversity of LDCVs

LDCVs are found in all neuroendocrine, endocrine, and exocrine tissues. They comprise spherical vesicles consisting of an acidic lumen with electron dense material surrounded by a limiting membrane (Figure 2(b)). The dense core content is composed of tightly packed aggregates of soluble proteins, whose composition varies with the respective type of tissue, and other cargo molecules including catecholamines, monoamines, and ATP. The limiting membrane contains proteins required for secretory granule acidification, uptake of nonproteinaceous molecules, transport, targeting, and fusion. SGs are typically distinguished by their morphological appearance and immunoreactivity for cell type-specific neuropeptides, hormones, or other cargo molecules including biogenic amines. Widespread markers for LDCVs are acidic proteins belonging to the granin family. These are chromogranin A (CgA), chromogranin B (CgB; also known as secretogranin I), secretogranin II (SgII; also known as chromogranin C), secretogranin III (SgIII), secretogranin IV (SgIV), secretogranin V (SgV; also known as 7B2), and secretogranin VI (SgVI). Another group of widely distributed marker proteins for SGs are the proneuropeptide and prohormone processing enzymes that convert their substrates into bioactive peptides. These enzymes include prohormone convertases 1/3 (PC1/3) and 2 (PC2) as well as carboxypeptidase E/H (CPE).

Besides the cell-type-specific diversity of SGs, the presence of different SGs in the same cell was also demonstrated in the same cell. Examples for SG diversity at the cellular level are LDCVs of gonadotrophs containing either luteinizing or follicle-stimulating hormone, whose release is independently controlled. Likewise, also in chromaffin cells, two morphologically distinct types of LDCVs containing epinephrine and norepinephrine, respectively, have been documented.

Biogenesis of LDCVs

Sorting of Proteins to LDCVs

Regulated secretory proteins (RSPs) and constitutive secretory proteins are synthesized in the rough endoplasmic reticulum and take the same intracellular route to and through the Golgi apparatus. In the most distal part of the Golgi complex, the TGN, their transport routes diverge. Constitutively secreted proteins enter constitutive secretory vesicles (CVs), which fuse with the plasma membrane without a stimulus (Figure 2(a)). RSPs are sorted into ISGs, which undergo a complex maturation process that includes budding of ISG-derived vesicles (IDVs) to remove cargo not destined for MSGs (Figure 2(a)). Two key mechanisms involved in the sorting of RSPs were elucidated: the selective (co)aggregation of RSPs and 'signal'-mediated targeting of RSPs to SGs (Figure 3).

Aggregation Morphological evidence demonstrating the presence of aggregates of RSPs in the TGN triggered the idea that aggregation is involved in the targeting of RSPs to SGs. Further evidence for this concept was obtained by biochemical studies showing that RSPs of the granin protein family, namely CgA, CgB, and SgII, aggregate at millimolar Ca^{2+} concentrations and at a pH below 6.5 – conditions which mimic the milieu of the TGN. Protons and divalent cations are thought to neutralize the repulsive negative charge of RSPs, thus inducing their dense packaging. Furthermore, constitutive markers were excluded from the formed aggregates (Figure 3). Similar conditions were found for the aggregation of other RSPs, such as pro-opiomelanocortin (POMC), von Willebrand factor, proinsulin, prolactin, CPE, and PC2. Further supporting this hypothesis is evidence that not only can constitutive proteins be separated from RSPs by aggregation of the latter but also different populations of RSPs can be segregated. This was demonstrated for the egg-laying hormone in type I neurons of *Aplysia californica*, which is processed in the TGN and results in the different sorting of the two processing products to different types of SGs. Furthermore, abundantly expressed RSPs may function as sorting helpers by providing aggregation

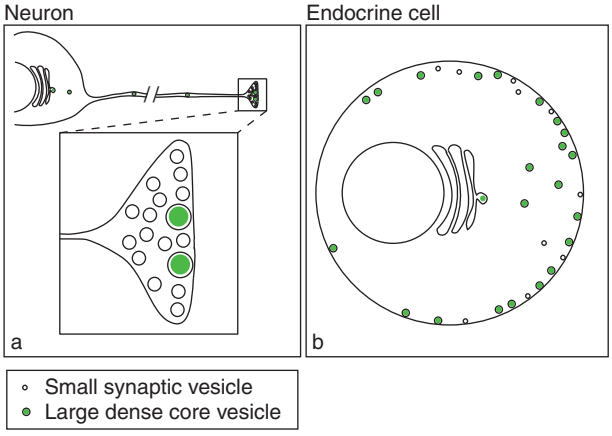

Figure 1 Large dense core vesicles (LDCVs) in neurons versus endocrine cells. The presynaptic terminal of neurons (a) contains only a few LDCVs, whereas the endocrine cells (b) comprise numerous LDCVs.

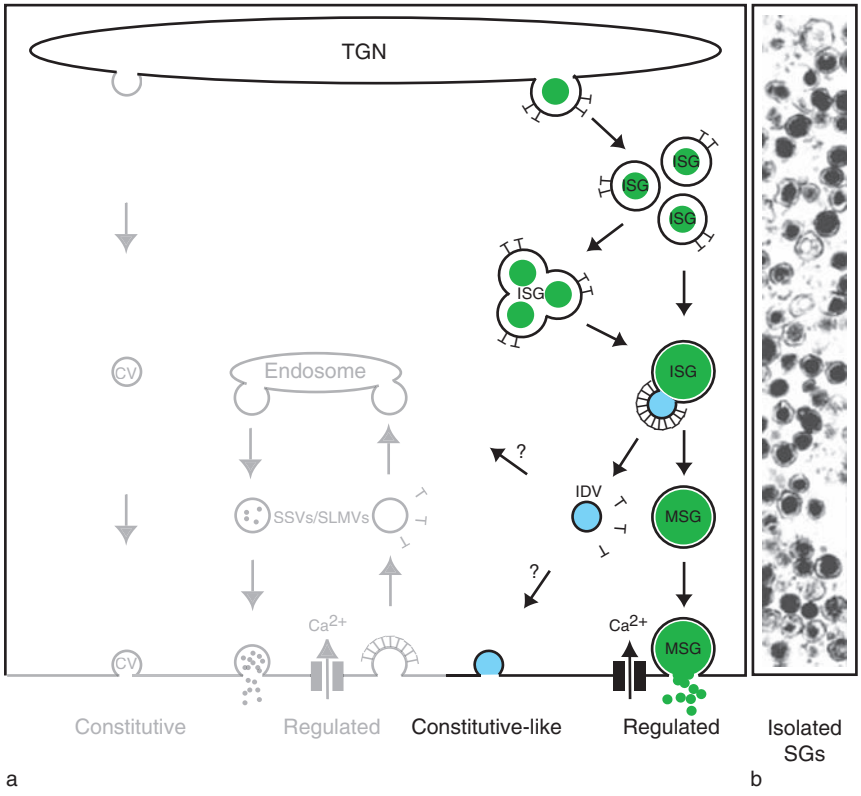

Figure 2 Trafficking of different vesicle types undergoing exocytosis. (a) Immature SGs (ISGs) bud from the TGN and may fuse homotypically. Clathrin-coated ISG-derived vesicles (IDVs) bud from maturing SGs and fuse with the PM/endosome without a stimulus. Mature SGs (MSGs) fuse with the PM in a stimulated manner upon Ca^{2+} influx. Constitutive secretory vesicles (CVs) bud from the TGN and fuse with the PM without a stimulus. Small synaptic vesicles (SSVs)/small synaptic-like microvesicles (SLMVs) bud from the endosome, are subsequently loaded with neurotransmitters, and fuse with the PM in a stimulated manner upon Ca^{2+} influx. After exocytosis, the SSV/SLMV-specific membrane patch is recycled from the PM by clathrin-mediated endocytosis. (b) Electron microscopic picture of isolated SGs from PC12 cells. Note that the trafficking of SGs, which is the focus of this article, is shown in black/color.

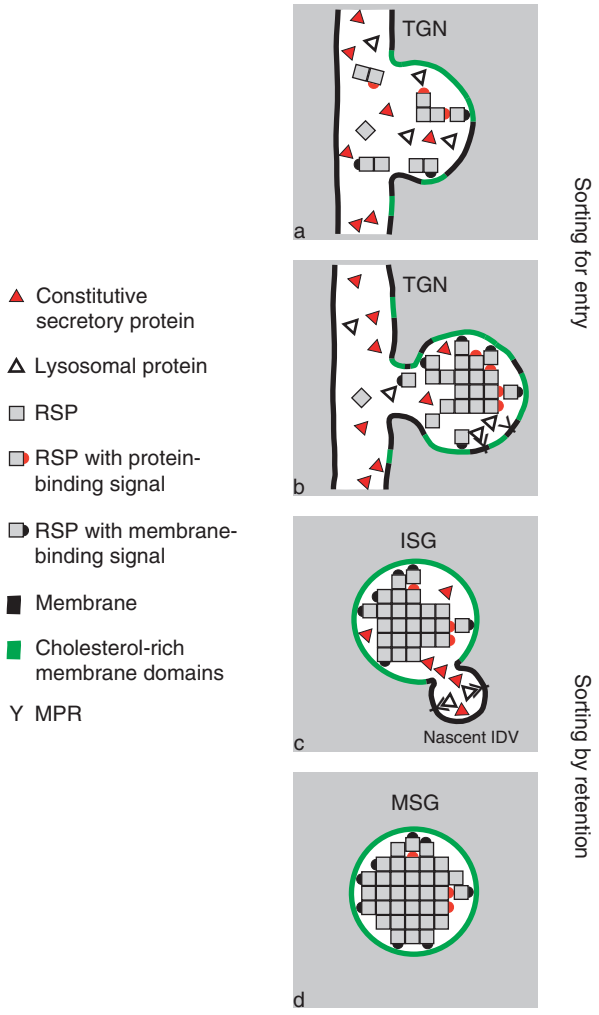

Figure 3 Sorting of regulated secretory proteins. (a) A mixture of constitutive (red triangles) and regulated secretory proteins (RSPs; gray squares) as well as lysosomal proteins (open triangles) arrives at the TGN. (b) RSPs with a protein- or membrane-binding signal (gray squares with red or black half-circles, respectively) aggregate in a Ca^{2+}/pH-dependent manner. RSPs with membrane-binding signals interact with specific, cholesterol-rich membrane patches (green lines) in the TGN. RSPs with protein-binding signals function as adaptors to link aggregated cargo to membrane-bound RSPs. (c) Upon progressive condensation of RSPs in ISGs, lysosomal and constitutive secretory proteins are excluded from the aggregates and are concentrated in nascent IDVs (lysosomal proteins bind to MPRs). (d) Mature SGs are densely packaged with aggregated RSPs.

seeds for those RSPs, which are present at a very low concentration and thus incompetent for sufficient aggregation on their own. Such a helper function was shown for CgB in AtT20 cells, in which its overexpression promoted the efficient sorting of a 23 kDa POMC fragment into SGs. Also, RSPs which lack specific targeting sequences but are competent for aggregation can be sorted via coaggregation with those RSPs possessing the collective sorting information for selective packaging in SGs (aggregative properties and specific targeting sequences) (Figure 3). This was demonstrated by using the vaccinia expression system, which shuts off the host cell protein synthesis and thereby allows the study of selected RSPs in a null background of endogenous RSPs. In this manner, a deletion mutant of CgB that lacked specific targeting sequences was expressed by a recombinant vaccinia virus and the protein and was found exclusively in CVs. However, coexpression of this mutant with non-mutated CgB by double infection rescued the sorting of the CgB mutant, thus suggesting its sorting via coaggregation. Taken together, compelling evidence exits that aggregation is a key mechanism for targeting of RSPs to SGs.

Sorting signals The search for specific sorting signals that direct proteins to SGs was mainly concerned with soluble cargo (RSPs) and, to a much lesser extent, membrane proteins of SGs. The latter include the type 1 membrane-spanning proteins peptidyl-α-amidating monooxygenase, phogrin, and muclin from exocrine SGs. In the case of phogrin, the granule sorting domain is located in the cytoplasmic tail of the protein, and it was shown that this domain binds the clathrin adaptor protein complexes AP-1 and -2. In order to find short sequences for SG-specific targeting in soluble cargo proteins, a variety of RSPs have been intensively analyzed using recombinant DNA technology for more than 30 years. Although short sequences of RSPs with targeting information were found, no consensus sequence for a sorting signal to SGs emerged. One of the identified signals consisted of an N-terminal disulfide-constrained loop structure. This loop is present in CgA and CgB, and their sequences are highly homologous. The loop of CgB was found to be sufficient to direct secretory cargo to SGs. It is of note that a constitutive reporter molecule fused with two CgB loops, one at each terminus, was sorted more efficiently compared to the fusion proteins possessing only one CgB loop. A similar loop, which is composed of a different sequence, was found in POMC. This loop was also found to be essential and sufficient for directing secretory cargo to SGs. As a second signal providing sorting information to SGs, short α-helical domains present in several RSPs were identified. Such domains were found in CPE, PC1/3 and PC2, prosomatostatin, CgA, and VGF. As a third motif, paired basic amino acids have been reported to accomplish SG-specific sorting in some proteins. These include prorenin, proneurotensin, prothyrotropin-releasing hormone, and progastrin. It appears that these paired basic amino acids constitute cleavage sites for one of the granule-resident prohormone convertases PC1/3 or PC2 because exchange of the

cleavage site to one recognized by furin, another protease that cleaves its substrates in the TGN, results in constitutive secretion of the respective products.

Thus, it appears that instead of a single sorting signal as found for the trafficking of proteins to other cellular destinations, a variety of targeting sequences emerged during evolution to accomplish sorting of RSPs to SGs. These multiple signals may function in a synergistic manner and/or may reflect the functional differences of SGs in various types of tissues or even in the same cell.

Cargo–membrane interaction The specific membrane composition of SGs implies that during the biogenesis of SGs a selective interaction of cargo molecules with specific membrane components in the TGN must occur. In the search for specific interaction sites, distinct structures of RSPs were found to be responsible for tethering RSPs to membranes specific for SGs. These structures were found to be identical to some of the aforementioned sorting signals. For CgB, it was shown that the disulfide-constrained loop mediates membrane binding in the TGN of rat pheochromocytoma (PC12) cells. In analogy to the sorting efficiency, two loops had a synergistic effect on the membrane binding efficiency compared to one loop. CPE, PC1/3, and PC2, were shown to interact with membranes through their short α-helical domains. The α-helices at the C-terminus of PC1/3 may be masked until the enzyme undergoes a conformational change in the TGN in which the tail has been proposed to unfold and render the enzyme fully active. Results suggest that these α-helices possess hydrophobic patches segregated from charged patches of amino acids to allow a shallow membrane interaction. With regard to the existing evidence that these structural motifs direct the sorting of the respective RSPs to SGs, their binding to membranes has to occur in the TGN and/or ISGs, as was shown in the case of the loop structure of CgB. However, it appears that cargo–membrane interaction is not mediated through binding to a classical sorting receptor encompassing a transmembrane domain with a cytoplasmic tail for coat assembly. In this respect, the hypothesis that CPE functions as a classical sorting receptor is controversial.

What are the requirements for the membrane binding of the aforementioned targeting sequences? A first clue about the binding site was provided by the finding that in AtT20 cells cholesterol depletion inhibited the formation of SGs, suggesting the involvement of cholesterol-rich rafts in this process. Subsequent studies provided evidence that PC2, CPE, and SgIII specifically bind to cholesterol-rich membrane patches (Figure 3). Note that in contrast to PC2 and CPE, SgIII was not found in a Triton X-100 insoluble membrane fraction. Therefore, distinct membrane patches might recruit different RSPs with specific sorting signals.

Other RSPs were shown to achieve SG-specific targeting through the interaction with membrane-tethered RSPs (Figure 3). One example is the binding of POMC to CPE via its N-terminal loop domain. This interaction was displaced by several other RSPs but not by constitutive secretory proteins. Furthermore, proinsulin, proenkephalin, and brain-derived neurotrophic factor have been proposed to interact with CPE to promote their retention in SGs. Similar to the POMC–CPE interaction, it was found that CgA was linked to SG membranes via SgIII. The finding that dibasic cleavage sites are targeting sequences for SGs suggests that PC1/3 and PC2 can act as sorting chaperones for their substrates in addition to being processing proteases.

Taken together, these data indicate that RSPs are not recruited to SGs through a single sorting motif or mechanism but, rather, through a wide variety of strategies that still have to be investigated in detail. Often, but not always, several sorting signals are combined in one protein. This might be the underlying reason for the observed phenomenon that the same protein can be sorted differently in different cell types.

Maturation of LDCVs

During maturation of SGs, progressive acidification of the lumen along with processing and packaging of soluble cargo leads to further condensation of the matrix and dense core formation. Notably, ISGs are the predominant site of proneuropeptide and prohormone conversion into bioactive peptides. The processing is accomplished by PCs at dibasic amino acid cleavage sites of RSPs. Examples of this are the conversion of POMC to adrenocorticotrophic hormone and proinsulin to insulin.

Maturation of ISGs also includes refinement of the content and remodeling of the membrane of newly formed ISGs. This process involves the homotypic fusion of ISGs and budding of IDVs from maturing SGs (Figure 2(a)). In PC12 cells, between three and five ISGs appear to fuse with the help of the soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) protein syntaxin 6 and synaptotagmin IV to form an MSG. Since MSGs are typically smaller than ISGs, homotopic fusion seems not to be obligatory for other types of neuroendocrine cells.

Membrane remodeling also consists of sorting away a specific set of proteins such as mannose-6-phosphate receptor (MPR), furin, carboxypeptidase D, and the SNAREs syntaxin 6 and VAMP4 in IDVs. This sorting depends on the ADP-ribosylation-factor (ARF)-1-dependent recruitment of AP-1 to MPR,

furin, and VAMP4 in a phosphorylation-dependent manner involving casein kinase II. In conjunction with the presence of clathrin coats on ISGs but not on MSGs, AP-1 has been proposed to recruit clathrin to ISGs and thereby promote the budding of IDVs during the maturation process. GGA (Golgi-associated γ-ear-containing ARF-binding protein) has been shown to be implicated as an essential factor in the recruitment of clathrin and thus IDV budding.

In addition to sorting away selected membrane proteins from ISGs, it is believed that soluble proteins not prone to aggregation (e.g., C-peptide, a processing product of proinsulin) are removed from maturing SGs in IDVs in a stimulus-independent manner – a process termed constitutive-like secretion. Likewise, excess membrane released by the progressively condensing core of maturing SGs is likely to be removed by the same pathway. It is an open question whether IDVs fuse with the endosomal system and/or directly with the PM. With respect to the variety of cargo molecules removed from ISGs, it is tempting to speculate that distinct subsets of these proteins may be removed in distinct IDVs.

Two levels of sorting to the regulated pathway of protein secretion can be distinguished: sorting in the TGN during formation of ISGs (termed sorting for entry; Figures 3(a) and 3(b)) and sorting in ISGs during their maturation (termed sorting by retention; Figures 3(c) and 3(d)). The relevance of sorting at these two levels appears to vary between cell types. In neurons, membrane traffic to the somatodendritic or axonal cell surface diverges in the TGN, which is located at the neuronal cell body. Thus, in this case, the TGN is likely to be the major site of sorting of RSPs and constitutive secretory proteins. This may be different for certain exocrine cells and endocrine cells such as pancreatic islet β cells, in which a considerable portion of sorting has been shown to take place at the level of ISGs.

Is There a Master Gene for LDCV Biogenesis?

CgA has been proposed to act as a master gene for the 'on/off' switch of SG biogenesis. This hypothesis was based on two lines of observations. One was that silencing of CgA expression in PC12 cells resulted in a loss of visible SGs. The other was that the expression of CgA in cells which lack a regulated secretory pathway induced the formation of vesicles with an electron dense core. However, this has raised controversy because several other findings argue against such a master gene concept. First, deletion of single SG genes in mice, including that of CgA, precludes neither the expression of other SG genes nor the biogenesis of SGs. Second, the absence of CgA in mice results in a compensatory overexpression of other granins. Third, the expression of other RSPs in cells that lack the regulated secretory pathway is also capable of inducing the formation of vesicles with an electron dense core. Taken together, these data suggest that CgA is replaceable by other SG-specific cargo molecules. It therefore seems plausible that the induction of granule-like dense core vesicles by the expression of RSPs in cells, which do not normally make granules, merely reflects the self-aggregating properties of RSPs and the general ability of the TGN to form SGs. In this respect, it is of note that the induced SGs were mainly characterized by their morphology, and it remains to be shown if such SG-like vesicles have all the key properties of *bona fide* SGs. Further arguments against a master gene function of CgA are supported by data demonstrating the convergent evolution of SGs in different species (e.g., vertebrates and ciliates), which infers that granules arose independently in different lineages.

Transport of LDCVs

During the past decade, fluorescent protein-based live cell imaging has made it possible to monitor the transport dynamics of SGs in time and space. Green fluorescent protein (GFP)-fusion proteins targeted to the lumen of SGs were shown to fluoresce despite the aggregative and acidic milieu. The imaging of newly formed ISGs on their way to the cell cortex is challenging because a neuroendocrine cell contains thousands of stored SGs with a half-life of several days. ISGs represent only a small percentage of the total SG population (Figure 4(c)). Therefore, selective labeling of ISGs is necessary to follow their route. This was achieved by employing a pulse/chase-like system based on the expression of human CgB, tagged with a temperature-sensitive GFP mutant (CgB)–GFP(S65T) in PC12 cells. The temperature-sensitive GFP mutant folds to its fluorescent form at 20 °C but not at 37 °C. Once properly folded at 20 °C, the fusion protein remains fluorescent at 37 °C. When cells were incubated at 20 °C, a temperature at which the formation of secretory vesicles is blocked, an accumulation of the fluorescent protein, selectively in the TGN, was observed (Figure 4(a)). Upon reversal of this block at 37 °C, hCgB–GFP(S65T) fluoresced in newly formed ISGs (Figure 4(b)). With this protocol, pools of SGs of a defined age were generated and followed through the cell. This revealed that ISGs move straight and in a microtubule-dependent manner from the TGN to the F-actin-rich cortex (Figure 4(c)). Similarly, in insulinoma cells, a fast and unidirectional movement of SGs was documented upon glucose stimulation. Biochemical studies showed that such movement

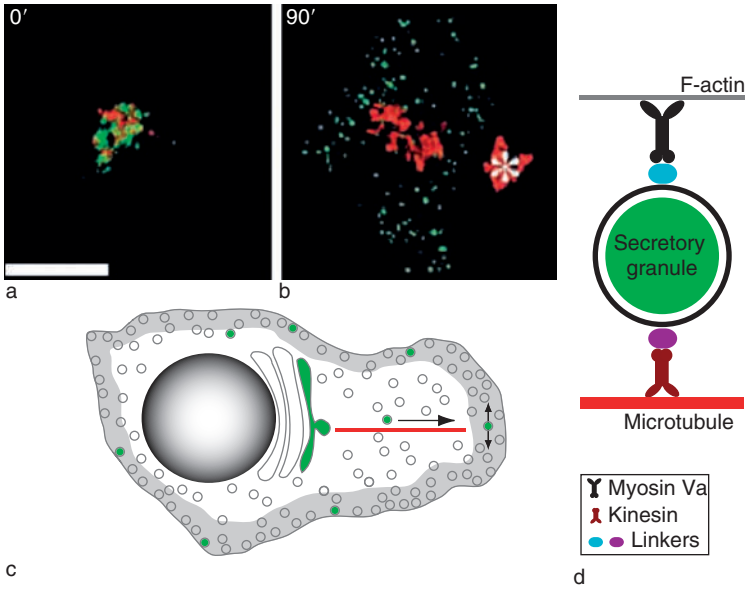

Figure 4 Transport of SGs. (a, b) GFP-based pulse/chase-like staining. (a) GFP-labeled secretory cargo (green) is concentrated selectively in the TGN of a PC12 cell during 20 °C incubation. (b) Green fluorescent SGs synthesized during 90 min after the release of the 20 °C block. Red, immunostained TGN38; asterisk, TGN of an untransfected cell. (c) Scheme illustrating the transport of newly formed ISGs as observed by the pulse/chase-like system. After their synthesis at the TGN, fluorescent ISGs (green solid circles) are transported outward along microtubules (red line, arrow) to the cortical area (gray), where they move along F-actin (double-arrow). Note that open circles indicate MSGs synthesized before application of the 20 °C block. (d) The transport of SGs along microtubules is kinesin dependent, whereas the transport along F-actin is myosin Va dependent. The respective motor proteins bind to SGs via linker proteins. (a, b) From Rudolf R, Kögel T, Kuznetsov SD, et al. (2003) Myosin Va facilitates the distribution of secretory granules in the F-actin rich cortex of PC12 cells. *Journal of Cell Science* 116: 1339–1348.

depends on the ATP-dependent conventional kinesin (Figure 4(d)). Whether this motor protein is common for other types of ISGs remains to be investigated. During their maturation, ISGs were found to move in a myosin Va-dependent manner in the F-actin-rich cortex (Figures 4(c), 4(d), and 5(a)). The role of myosin Va was investigated by the expression of a dominant-negative mutant of this motor protein. This revealed that myosin Va captures SGs and distributes them in the F-actin-rich cortex of PC12 cells. Based on these data, it has been proposed that newly formed SGs move in a microtubule-dependent manner to the cell cortex and in a myosin Va-dependent manner in the F-actin-rich cortex. Further studies under conditions in which myosin Va activity was blocked by a dominant-negative mutant of myosin Va or by RNA silencing implicated myosin Va in the exocytosis of SGs of PC12 cells and pancreatic MIN6-β cells.

In search of linker proteins recruiting myosin Va to SGs, the myosin Va-dependent transport of melanosomes served as a blueprint. According to this model, myosin Va is associated with melanosomes through a 'receptor complex' consisting of the small GTPase Rab27a and melanophilin/Slac2-a (synaptotagmin-like protein (Slp) lacking C2 domains-a). In agreement with this, it was shown that Rab27a associates with SGs in both chromaffin and PC12 cells and with insulin-containing SGs in pancreatic β cells. Furthermore, the interaction partners of Rab27a on SGs were also found to be similar to melanophilin, namely Slac2-c/MyRIP in chromaffin and PC12 cells and Slp4/granulophilin in pancreatic and MIN6 β cells. Slac2-c/MyRIP was shown to interact with Rab27a/b, myosin Va, and actin *in vitro*. However, a functional link between the 'receptor' complex and myosin Va remains to be shown *in vivo*. In all studied cases, myosin Va is bound to SGs via Rab27 and a rabphilin homolog.

The Last Steps toward Exocytosis

Before fusion with the PM, the SGs eventually undergo priming and docking in the F-actin cortex, which leads to segregation into different vesicle pools. In particular, a readily releasable pool undergoing immediate exocytosis upon stimulation and a slower reacting pool(s) that facilitates long-lasting exocytosis under sustained stimulation have been described. Key regulators for the transit of SGs from the reserve to the readily releasable pool are Ca^{2+}, cortical F-actin together with its interaction partners, and the SNARE complex at the PM. In numerous cortical areas, F-actin was shown to be

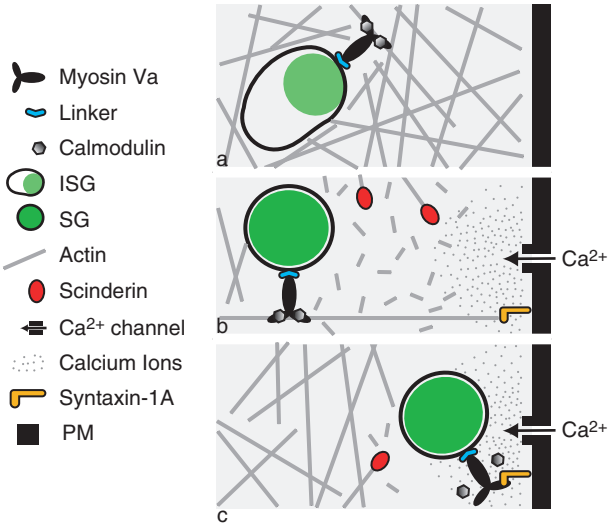

Figure 5 Scheme illustrating the involvement of myosin Va in mobilizing SGs from reserve pools toward exocytotic sites. (a) Myosin Va is associated with ISGs and mediates their F-actin-dependent movement. (b) Upon Ca^{2+} influx, scinderin is activated and disassembles F-actin. Myosin Va transports SGs from the reserve pool toward the PM along remaining F-actin tracks. (c) Further increase in Ca^{2+} concentration results in the dissociation of calmodulin from myosin Va and the binding of the latter to syntaxin 1A. Note that through these consecutive steps SGs are confined to the PM. From Eichler TW, Kögel T, Bukoreshtliev NV, and Gerdes H-H (2006) The role of myosin Va in secretory granule trafficking and exocytosis. *Biochemical Society Transactions* 34 (Pt. 5): 671–674.

disassembled upon elevated cytosolic Ca^{2+} levels during stimulation. A large body of evidence exists indicating that this process is delicately balanced by the action of many regulating factors. These include the F-actin severing proteins scinderin (Figure 5(b)) and gelsolin as well as proteins promoting F-actin assembly, such as the Rho GTPase Cdc42 and the guanine nucleotide exchange factor intersectin-1L.

Notably, in the cortical areas depleted of F-actin, fine filaments remain and are thought to guide SGs from the reserve pool to the exocytotic sites (Figure 5(b)). This, in conjunction with the finding that priming of SGs is ATP dependent, suggested the involvement of an F-actin-dependent motor protein, which was subsequently identified as myosin Va (Figures 4(d) and 5). After its arrival at the PM, myosin Va interacts with components of the docking complex as suggested by *in vitro* experiments. It was shown that micromolar Ca^{2+} concentrations lead to a replacement of calmodulin (CaM) light chains at an IQ motif of myosin Va by the t-SNARE syntaxin 1A (Figure 5(c)). Syntaxin 1A confines SGs to the exocytotic site and is part of the SNARE complex that carries out the actual fusion. However, myosin Va does not seem to play a role in the final exocytosis event since its interaction with the fusion machinery constituents *N*-ethylmaleimide-sensitive factor/α-soluble *N*-ethylmaleimide-sensitive factor attachment protein (α-SNAP) was shown to release it from the SNARE complex *in vitro* and cell depolarization induced the dissociation of myosin Va from chromaffin vesicles *in vivo*. In this context, rabphilin has been discussed as a possible bridging protein between myosin Va and SG-associated Rab27.

During recent years, the sequence of actions of the components of the SNARE complex during tethering and fusion has been dissected. Vesicle delivery and weak tethering depend on the Sec1/Munc18-like protein Munc18-1, which binds to the t-SNARE syntaxin. Syntaxin and SNAP25 are proposed to form the fusion complex in concert with the v-SNARE synaptobrevin. Thus, the demonstrated interaction of myosin Va and rabphilin as two likely components of the motor complex of SGs with the components of the fusion complex provides insights into how the transition of SGs from tethering in the F-actin cortex to docking at the PM is accomplished at the molecular level. The fusion complex initiates exocytosis only when the inhibitory effect of synaptotagmin is removed upon elevated levels of Ca^{2+}. Nevertheless, there is evidence that not all SGs are consumed by exocytosis. The fate of nonexocytosed SGs has been elucidated in chromaffin cells by the expression of atrial natriuretic factor fused to a fluorescent protein, which changes its color over time. Newly synthesized SGs were shown to be immobilized below the PM and were exocytosed preferentially, whereas older SGs seemed to be retrieved from the actin cortex to the inner cell. This elegant study corroborated previous indications that newly formed SGs are secreted foremost.

See also: Exocytosis: Ca^{2+}-Sensitivity; Presynaptic Receptor Signaling; RNA Granules: Functions Within Presynaptic Terminals and Postsynaptic Spines; Synaptic Vesicles.

Further Reading

Arvan P and Castle D (1998) Sorting and storage during secretory granule biogenesis: Looking backward and looking forward. *Biochemical Journal* 332: 593–610.

Borgonovo B, Ouwendijk J, and Solimena M (2006) Biogenesis of secretory granules. *Current Opinion in Cell Biology* 18: 365–370.

Burgoyne RD and Morgan A (2003) Secretory granule exocytosis. *Physiological Reviews* 83: 581–632.

Dikeakos JD and Reudelhuber TL (2007) Sending proteins to dense core secretory granules: Still a lot to sort out. *Journal of Cell Biology* 177: 191–196.

Eichler TW, Kögel T, Bukoreshtliev NV, and Gerdes H-H (2006) The role of myosin Va in secretory granule trafficking and exocytosis. *Biochemical Society Transactions* 34: 671–674.

Elde NC, Long M, and Turkewitz AP (2007) A role for convergent evolution in the secretory life of cells. *Trends in Cell Biology* 17: 157–164.

Glombik MM and Gerdes HH (2000) Signal-mediated sorting of neuropeptides and prohormones: Secretory granule biogenesis revisited. *Biochimie* 82: 315–326.

Gorr SU, Jain RK, Kuehn U, Joyce PB, and Cowley DJ (2001) Comparative sorting of neuroendocrine secretory proteins: A search for common ground in a mosaic of sorting models and mechanisms. *Molecular and Cellular Endocrinology* 172: 1–6.

Kim T, Gondre-Lewis MC, Arnaoutova I, and Loh YP (2006) Dense-core secretory granule biogenesis. *Physiology* 21: 124–133.

Meldolesi J, Chieregatti E, and LuisaMalosio M (2004) Requirements for the identification of dense-core granules. *Trends in Cell Biology* 14: 13–19.

Rudolf R, Kögel T, Kuznetsov SA, et al. (2003) Myosin Va facilitates the distribution of secretory granules in the F-actin rich cortex of PC12 cells. *Journal of Cell Science* 116: 1339–1348.

Solimena M and Gerdes HH (2003) Secretory granules: And the last shall be first. *Trends in Cell Biology* 13: 399–402.

Thiele C, Gerdes HH, and Huttner WB (1997) Protein secretion: Puzzling receptors. *Current Biology* 7: R496–R500.

Tooze SA (1998) Biogenesis of secretory granules in the *trans*-Golgi network of neuroendocrine and endocrine cells. *Biochimica et Biophysica Acta* 1404: 231–244.

Latrotoxin

M V Khvotchev, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

α-Latrotoxin is a large, approximately 130 kDa protein neurotoxin found in the venom of widow spiders (genus *Latrodectus*). α-Latrotoxin mediates all toxic effects of the spider envenomation in vertebrates, whereas closely related latroinsectotoxins and latrocrustatoxin also present in the venom target insects, crustaceans, and other invertebrates. A single bite of the highly poisonous black widow spider (*Latrodectus mactans*) delivers only minute quantities of α-latrotoxin and is rarely life-threatening for humans but can readily kill a domestic cat or other small mammal. α-Latrotoxin poisoning (latrodectism) produces a number of local and systemic symptoms, such as pain, muscle cramps, anxiety, headache, nausea, excessive salivation, lacrimation, and sweating, that can persist for several days. The principal sites of α-latrotoxin action in the body are synaptic junctions (particularly neuromuscular junctions) and several types of secretory cells. In all affected terminals and cells, α-latrotoxin induces massive and uncontrolled release of neurotransmitters and mediators via exocytosis of the carrier vesicles. Under experimental conditions, the toxin stimulates exocytosis of all known neurotransmitters in a wide range of synapses and cells (e.g., acetylcholine from frog neuromuscular junctions, glutamate from mammalian excitatory synapses in hippocampus, and catecholamines from chromaffin cells in adrenal glands); thus, α-latrotoxin acts on a basic conserved feature of exocytic machinery. The effects of the toxin are chronic and in most cases irreversible; afflicted nerve terminals often degenerate. The molecular mechanism of α-latrotoxin action is complex and not completely understood. It involves binding to specific cell surface receptors, insertion into plasma membrane, and unknown downstream activity that induces massive exocytosis of vesicles. Intensive study of α-latrotoxin for more than 30 years has provided important insights into mechanisms of exo- and endocytosis, neurotransmitter release, and synapse function, exemplifying the essential role of natural toxins for the advancement of neuroscience.

Synthesis and Molecular Organization of α-Latrotoxin

α-Latrotoxin is synthesized by glandular cells of the widow spiders as an inactive, approximately 160 kDa precursor protein (Figure 1). The precursor protein initially accumulates in the cytoplasm and is secreted into the lumen of the venom gland by a holocrine mechanism which involves rupture of the whole cells. Maturation of α-latrotoxin consists of a two-step proteolytic processing (removal of 20 amino acid residues at the N-terminus and approximately 200 residues at the C-terminus) and formation of intramolecular disulfide bonds and leads to activation of the toxin. Mature α-latrotoxin is divided into two structural domains (Figure 1). The N-terminal domain is unique and only found in latrotoxins. It contains three essential conserved cysteine residues and two hydrophobic regions that can potentially span the plasma membrane. The C-terminal domain of α-latrotoxin is made of 22 ankyrin repeats (repeats 1 and 16 are not perfect). Ankyrin repeats are found in many intracellular proteins and are thought to mediate stable protein–protein interactions. The high-resolution three-dimensional (3-D) structure of α-latrotoxin is not available. Lower-resolution cryoelectron microscopy has revealed that the toxin forms stable dimers in solution that further assemble into tetramers or higher-order oligomers in the presence of divalent cations. Each toxin monomer consists of three compact 3-D domains (Figure 1) called 'wing' (which contains most of the N-terminal domain), 'body' (which contains the rest of the N-terminal domain and the first six ankyrin repeats), and 'head' (which contains the last six ankyrin repeats). Functional α-latrotoxin can be purified from the venom or venom glands of widow spiders or produced and purified *in vitro* using various heterologous expression systems.

Receptors of α-Latrotoxin

High target selectivity of α-latrotoxin is mediated by specific plasma membrane receptors that recruit the toxin to the sites of action. Members of the three protein families have been identified as functional receptors for α-latrotoxin: neurexins, CIRLs/latrophilin/CLs, and receptor-like protein tyrosine phosphatase-σ (PTP-2; Figure 2). Surprisingly, these receptors have no obvious sequence similarity but are capable of binding α-latrotoxin with nanomolar affinities. In all three receptor families, the toxin binding site is situated in close proximity to the plasma membrane. The best-studied α-latrotoxin receptors, neurexins, are encoded by three genes (neurexins 1–3) with alternative use of two sets of promoters (long α and short β isoforms) and multiple splice sites that give rise to thousands of potential protein isoforms in mammals. The neurexin 1α isoform is the most abundant α-latrotoxin receptor in

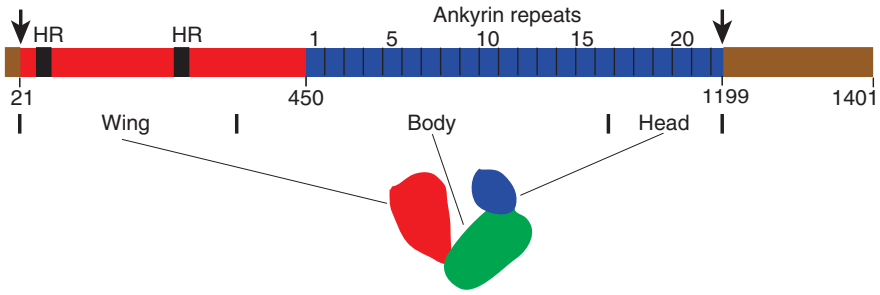

Figure 1 Structure of α-latrotoxin. (Top) Schematic diagram of primary structure of α-latrotoxin precursor. Arrows indicate the sites of proteolytic cleavage during maturation of α-latrotoxin. HR, hydrophobic region; ankyrin repeats are numbered on top and amino acid residues are numbered below the diagram. (Bottom) Three-dimensional domains ('wing,' 'body,' and 'head') and their boundaries.

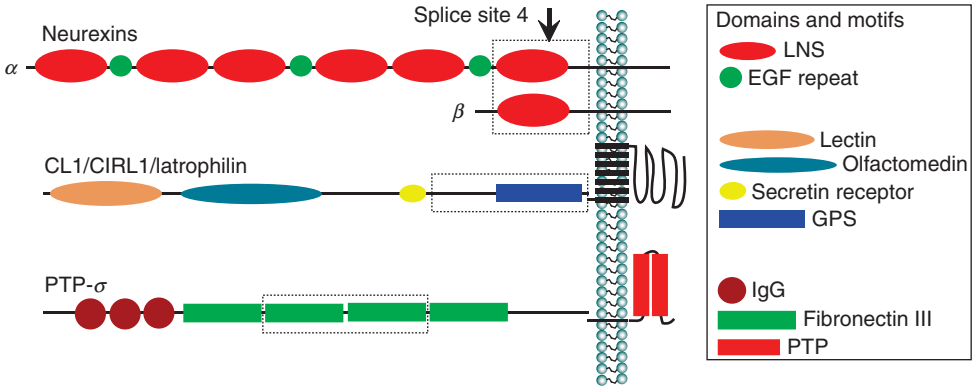

Figure 2 Receptors of α-latrotoxin. (Left) Schematic structure of α- and β-neurexins, CL1 (also called CIRL1 and latrophilin), and receptor-like protein tyrosine phosphatase σ (PTP-σ). Dotted rectangle depicts the area essential for α-latrotoxin binding. In neurexins, binding of α-latrotoxin is negatively regulated by the presence of an insert at site 4. (Right) List of major domains and motifs found in α-latrotoxin receptors.

the brain; however, other neurexins also bind the toxin. Different from CLs and PTP-σ, neurexins bind α-latrotoxin only in the presence of calcium. The interaction is also regulated by alternative splicing. Only neurexins lacking insert at the splice site 4 are functional receptors (Figure 2). Neurexins and their endogenous ligands, neuroligins, are implicated in normal synapse formation and maintenance and in pathological dysregulation of synaptic function in various forms of familial autism. CLs are a family of three orphan G-protein-coupled receptors with unknown function. CL1 and CL2 isoforms bind α-latrotoxin with high affinity, whereas CL3 does not. Extensive biochemical, physiological, and genetic studies suggest that neurexins and CLs recruit α-latrotoxin to the target plasma membrane *in vivo* but do not transduce a signal that initiates massive exocytosis. Much less is known about the physiological significance of PTP-σ as an α-latrotoxin receptor.

Membrane Active Properties of α-Latrotoxin

α-Latrotoxin forms nonselective cation channels and pores in membranes. This activity was initially observed in artificial lipid bilayers that contain no α-latrotoxin receptors. However, α-latrotoxin only forms channels and pores in the native membranes that contain the toxin receptors. The effects of α-latrotoxin on the membranes *in vivo* are strictly concentration dependent. At low, subnanomolar concentrations, α-latrotoxin induces individual, low-conductance channels permeable to sodium, calcium, magnesium, and other cations. At higher concentrations (in the low nanomolar range), the toxin also forms pores permeable to low-molecular-weight substances. Finally, at high concentrations α-latrotoxin causes detergent-like lysis of membranes and the release of soluble cytosolic markers.

Effects of α-Latrotoxin in the Presence or Absence of Extracellular Calcium

α-Latrotoxin induces massive release of acetylcholine at the neuromuscular junctions independent of external calcium. After application of low concentrations of the toxin, exocytosis of acetylcholine-filled small synaptic vesicles persists for a long time without major morphological changes at the presynaptic

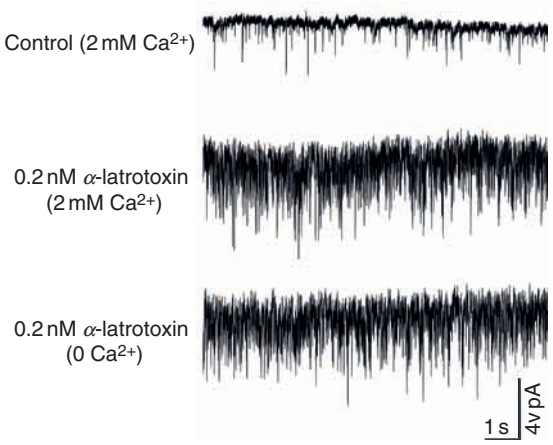

Figure 3 α-Latrotoxin stimulates spontaneous glutamate release at hippocampal synapses independent of external calcium. Representative traces illustrating excitatory miniature postsynaptic currents recorded from high-density hippocampal neuronal cultures under basal conditions (top) and after application of α-latrotoxin in the presence (middle) or absence (bottom) of external calcium.

terminals. In the absence of calcium, prolonged exposure of neuromuscular junctions to α-latrotoxin eventually leads to depletion of the synaptic vesicles at the presynapse and significant bloating of the plasma membrane, probably due to the failure of compensatory endocytosis. α-Latrotoxin also causes sustained release of glutamate and GABA at the central nervous system synapses independent of extracellular calcium (Figure 3). Higher concentrations of α-latrotoxin overwhelm the release machinery at these terminals and induce depletion of the vesicle cluster, mitochondria swelling, plasma membrane extension, ATP loss, and, finally, disintegration of the terminal. In sharp contrast, α-latrotoxin stimulates exocytosis of large dense core vesicles (e.g., neuropeptide-filled vesicles at the synapse or catecholamine-filled granules in chromaffin cells and PC12 cells) only in the presence of external calcium. α-Latrotoxin-induced release of norepinephrine and dopamine from brain synaptosomes is partially calcium dependent, probably due to the intermediate nature of the carrier vesicles for these neurotransmitters that share properties of small synaptic vesicle and large dense core vesicles. The requirement for external calcium in α-latrotoxin-induced exocytosis is an intrinsic property of the release machinery and is not mediated by differential involvement of calcium-dependent (neurexins) or calcium-independent (CLs and PTP-σ) receptors.

Mechanism of α-Latrotoxin Action

α-Latrotoxin is a potent, selective, and complex neurotoxin that exerts multiple effects on the secretion apparatus of synapses and secretory cells. The specificity of the toxin action is achieved by interaction with multiple high-affinity receptors. The stimulation of exocytosis by α-latrotoxin occurs downstream of the receptor binding by a receptor-independent action that involves insertion of the toxin into the plasma membrane. To induce exocytosis, α-latrotoxin requires basic fusion machinery (e.g., SNARE and SM proteins) but not the calcium sensing and priming modules involved in physiological membrane fusion (e.g., synaptotagmins and Munc13). Under physiological conditions, α-latrotoxin can stimulate exocytosis by a simple ionophoric mechanism with calcium ions entering through nonselective cation channels formed by the toxin. However, the ability of α-latrotoxin to induce massive exocytosis independently of external calcium points to an unknown direct activity of the toxin on the fusion machinery that overrides normal calcium control of the exocytosis. Future studies will address the exact molecular mechanism employed by this fascinating neurotoxin to hijack the tightly controlled release machinery of neurons and secretory cells.

See also: Endocytosis and Presynaptic Scaffolds; Exocytosis: Ca^{2+}-Sensitivity; Neuromuscular Junction (NMJ): A Target for Natural and Environmental Toxins in Humans; Neurotoxins; Neurotoxins and their Neurotoxicology.

Further Reading

Ceccarelli B and Hurlbut WP (1980) Ca^{2+}-dependent recycling of synaptic vesicles at the frog neuromuscular junction. *Journal of Cell Biology* 87(1): 297–303.

Davletov BA, Meunier FA, Ashton AC, et al. (1998) Vesicle exocytosis stimulated by alpha-latrotoxin is mediated by latrophilin and requires both external and stored Ca^{2+}. *EMBO Journal* 17(14): 3909–3920.

Grishin EV (1998) Black widow spider toxins: The present and the future. *Toxicon* 36(11): 1693–1701.

Henkel AW and Betz WJ (1995) Monitoring of black widow spider venom (BWSV) induced exo- and endocytosis in living frog motor nerve terminals with FM1-43. *Neuropharmacology* 34(11): 1397–1406.

Hurlbut WP, Chieregatti E, Valtorta F, and Haimann C (1994) Alpha-latrotoxin channels in neuroblastoma cells. *Journal of Membrane Biology* 138(1): 91–102.

Hurlbut WP, Iezzi N, Fesce R, and Ceccarelli B (1990) Correlation between quantal secretion and vesicle loss at the frog neuromuscular junction. *Journal of Physiology* 425: 501–526.

Ichtchenko K, Khvotchev M, Kiyatkin N, Simpson L, Sugita S, and Südhof TC (1998) Alpha-latrotoxin action probed with recombinant toxin: Receptors recruit alpha-latrotoxin but do not transduce an exocytotic signal. *EMBO Journal* 17(21): 6188–6199.

Khvotchev M, Lonart G, and Südhof TC (2000) Role of calcium in neurotransmitter release evoked by alpha-latrotoxin or hypertonic sucrose. *Neuroscience* 101(3): 793–802.

Khvotchev M and Sudhof TC (2000) Alpha-latrotoxin triggers transmitter release via direct insertion into the presynaptic plasma membrane. *EMBO Journal* 19(13): 3250–3262.

Krasnoperov V, Bittner MA, Mo W, et al. (2002) Protein-tyrosine phosphatase-sigma is a novel member of the functional family of alpha-latrotoxin receptors. *Journal of Biological Chemistry* 277(39): 35887–35895.

Li G, Lee D, Wang L, et al. (2005) N-terminal insertion and C-terminal ankyrin-like repeats of alpha-latrotoxin are critical for Ca^{2+}-dependent exocytosis. *Journal of Neuroscience* 25(44): 10188–10197.

Liu J and Misler S (1998) Alpha-latrotoxin alters spontaneous and depolarization-evoked quantal release from rat adrenal chromaffin cells: Evidence for multiple modes of action. *Journal of Neuroscience* 18(16): 6113–6125.

Matteoli M, Haimann C, Torri-Tarelli F, Polak JM, Ceccarelli C, and De Camilli P (1988) Differential effect of alpha-latrotoxin on exocytosis from small synaptic vesicles and from large dense-core vesicles containing calcitonin gene-related peptide at the frog neuromuscular junction. *Proceedings of the National Academy of Sciences of the United States of America* 85(19): 7366–7370.

Michelena P, de la Fuente MT, Vega T, et al. (1997) Drastic facilitation by alpha-latrotoxin of bovine chromaffin cell exocytosis without measurable enhancement of Ca^{2+} entry or $[Ca^{2+}]i$. *Journal of Physiology* 502(Pt. 3): 481–496.

Misler S and Falke LC (1987) Dependence on multivalent cations of quantal release of transmitter induced by black widow spider venom. *American Journal of Physiology* 253(3 Pt. 1): C469–C476.

Orlova EV, Rahman MA, Gowen B, et al. (2000) Structure of alpha-latrotoxin oligomers reveals that divalent cation-dependent tetramers form membrane pores. *Nature Structural Biology* 7(1): 48–53.

Schiavo G, Matteoli M, and Montecucco C (2000) Neurotoxins affecting neuroexocytosis. *Physiological Reviews* 80(2): 717–766.

Silva AM, Liu-Gentry J, Dickey AS, Barnett DW, and Misler S (2005) Alpha-latrotoxin increases spontaneous and depolarization-evoked exocytosis from pancreatic islet beta-cells. *Journal of Physiology* 565(Pt. 3): 783–799.

Sudhof TC (2001) Alpha-latrotoxin and its receptors: Neurexins and CIRL/latrophilins. *Annual Review of Neuroscience* 24: 933–962.

Tobaben S, Sudhof TC, and Stahl B (2002) Genetic analysis of alpha-latrotoxin receptors reveals functional interdependence of CIRL/latrophilin 1 and neurexin 1 alpha. *Journal of Biological Chemistry* 277(8): 6359–6365.

Ushkaryov Y (2002) Alpha-latrotoxin: From structure to some functions. *Toxicon* 40(1): 1–5.

Ushkaryov YA, Volynski KE, and Ashton AC (2004) The multiple actions of black widow spider toxins and their selective use in neurosecretion studies. *Toxicon* 43(5): 527–542.

Willson J, Amliwala K, Davis A, et al. (2004) Latrotoxin receptor signaling engages the UNC-13-dependent vesicle-priming pathway in *C. elegans*. *Current Biology* 14(15): 1374–1379.

Learned Flavor Aversions and Preferences

K Touzani and A Sclafani, Brooklyn College of the City University of New York, Brooklyn, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Determinants of Food Selection

A biological imperative for animals, including humans, is the identification, selection, and ingestion of nutritious foods and the rejection of toxic substances that can cause harm to the body. Deciding which potential food items to select and which to avoid typically involves the evaluation of flavor stimuli (taste, odor, texture). Many species have innate predispositions to accept some flavors (i.e., sweet taste) and to reject others (i.e., bitter taste), presumably because the flavors are predictive of naturally occurring nutritious and toxic materials, respectively. These inborn determinants of food choice, however, are very limited in number, particularly for omnivorous species, which eat a wide range of foods. However, animals readily acquire new flavor aversions and preferences through associative learning processes. In particular, if an animal consumes a novel flavored food or fluid and subsequently experiences visceral malaise, it will avoid that flavored substance in the future. This form of learning is referred to as taste aversion conditioning when the novel flavor is a pure taste (sweet, salty, sour, bitter) or flavor aversion conditioning when complex taste, odor, and/or texture stimuli are involved. Animals also learn to prefer the flavors of foods and fluids that provide positive nutritive consequences or recovery from nutritional deficiencies. Food selection is also guided by observational learning, social interactions and, in humans, cultural factors, although these are not discussed here.

Flavor Aversion Learning

Flavor aversion learning represents a powerful adaptive defense mechanism that minimizes the ingestion of toxic foods and fluids. It has been observed under natural conditions and demonstrated experimentally in many species of invertebrates and vertebrates. John Garcia was the first to introduce flavor aversion learning in the laboratory to investigate the behavioral effects of exposure to irradiation. He demonstrated that rats exposed to gamma radiation following ingestion of a novel saccharin solution readily learned to avoid this solution subsequently. Since then, flavor aversion learning has attracted the interest of many psychobiologists and neuroscientists. It is extensively used to study food selection processes and to screen drugs for potential aversive effects and, more generally, is used as a model to study neural mechanisms of learning and memory.

Behavioral Characteristics

Flavor aversion learning is a form of Pavlovian conditioning in which the flavor cue (conditioned stimulus; CS) is associated with exposure to a toxic agent (unconditioned stimulus; US) that produces nausea, gastrointestinal distress, or some other aversive state. Flavor aversion learning has special characteristics that distinguish it from traditional forms of Pavlovian conditioning:

1. Cue-consequence specificity. For many species, it is the flavor, and particularly the taste, of foods rather than other sensory properties (e.g., color, shape, location) that is associated with illness.
2. Long-delay learning. Unlike conventional forms of defense conditioning, flavor aversion learning can be established after delays of several hours between exposure to the CS and delivery of the US in a single learning trial.
3. CS–US unconscious association. Administration of deep anesthesia with pentobarbital or other general anesthetics, applied after the consumption of the CS flavor but before the administration of the toxic US, does not prevent flavor aversion learning.

Experimental Assessment

In the laboratory, flavor aversion learning is usually performed with highly motivated animals (food or water restricted) presented with a novel flavored food or fluid. Flavored solutions and thirsty animals are most commonly used for ease of measurement and preparation. Many different procedures are available to expose the animal to the flavor CS and US agent during conditioning.

Bottle technique In this standard procedure, thirsty animals are given limited access (usually 15 min) to one bottle of water for several days. When water intakes are stable, the animals are offered a novel flavored solution (the CS; e.g., saccharin solution) instead of water, the consumption of which is followed after some delay by the administration of a toxic US, such as an intraperitoneal injection of lithium chloride (LiCl). Control animals receive an injection of isotonic sodium chloride (NaCl). After 1 or 2 days of recovery, during which only water is

available, the animals may be given additional training trials with the CS and US or may be presented with the CS alone. The learned flavor aversion is revealed in the LiCl-treated group by a marked decrease in the intake of the CS solution compared to its first presentation and compared to its intake by the control group. In a common variation of this procedure, a two-bottle test (e.g., saccharin vs. water) is used to assess the strength of the conditioned flavor aversions. Two-bottle tests are generally more sensitive than are one-bottle tests in revealing the flavor aversion, because they allow the animal to satisfy its thirst while avoiding the CS solution.

Intraoral infusion method With the intraoral infusion method, animals are fitted with an intraoral (IO) catheter, which allows for the automatic infusion of the flavor stimulus into the oral cavity. Typically, thirsty animals are adapted to consume IO-infused water. Then a CS solution is infused, followed by administration of the toxic US. In a subsequent test session the learned aversion is revealed when the animal rejects the flavor by letting the CS solution drip out of its mouth. One advantage of the IO technique is that it can be used with neonatal animals or neurologically compromised adult animals that do not voluntarily consume the CS solution. The technique has also been used to determine the temporal onset of the flavor aversion by infusing the CS solution every 5 min after the US is administered.

Taste reactivity test Reduced consumption is typically taken as evidence for a conditioned flavor aversion. This measure, however, does not necessarily reflect changes in the hedonic value (or palatability) of the CS flavor. The taste reactivity test developed by Grill and Norgren in 1978 is thought to provide a more direct measure of flavor palatability. In this test, orofacial and other motor responses to IO infusions of flavored solutions are videotaped for off-line analysis of ingestion and rejection responses. Naive rats infused with inherently preferred solutions (e.g., sweet sugar solutions) display a sequence of ingestive reactions (tongue protrusions, paw licks, rhythmical mouth movements) that are distinct from aversive reactions (gapes, chin rubs, head shakes, paw treads) elicited by infusions of unpreferred solutions (e.g., bitter quinine solutions). Rats conditioned to avoid a sucrose solution by pairing its intake with LiCl or apomorphine subsequently display quinine-like aversive reactions when reexposed to the sucrose solution. However, other unconditioned stimuli, such as lactose or amphetamine, which suppress sugar intake in bottle tests, do not elicit aversive motor responses to the CS solution in taste reactivity tests. These findings suggest that some but not all US treatments condition a decrease in the hedonic evaluation of the CS flavor. Conditioned reductions in CS intake are sometimes referred to as a 'flavor aversion' when they are associated with decreased hedonic reactivity, and a 'flavor avoidance' when hedonic evaluation is not changed.

Brain Mechanisms

Flavor aversion learning involves the neural integration of sensory stimuli from the oral cavity (taste, odor, texture) and sensory stimuli from the gastrointestinal tract or generated by circulating toxins and the development of long-term flavor memories that can guide food selection. Electrophysiological and neuroanatomical studies have demonstrated convergence of gustatory and visceral information in several brain structures and provide strong clues to investigate the neural substrate of integration of oral–postingestive sensory information underlying the formation of taste aversion learning.

Lesion studies Brain areas involved in flavor aversion learning have been extensively investigated by analyzing the effects of electrolytic and excitotoxic lesions in animals.

Nucleus tractus solitarius. The nucleus tractus solitarius (NTS) is the first brain relay nucleus for gustatory and visceral sensory information. Taste aversion conditioning alters the electrophysiological response of NTS neurons to tastants, but rostral NTS lesions are reported not to block the acquisition or retention of taste aversion produced by LiCl. These lesions may have spared more caudal taste-responsive neurons that are sufficient to process the taste CS.

Area postrema. Chemosensitive neurons in the area postrema (AP) can detect blood-borne toxins, and AP lesions block or attenuate flavor aversions produced by a variety of toxic agents, including LiCl and copper sulfate. AP lesions do not, however, prevent other nausea-inducing drugs (apomorphine) from conditioning flavor aversions.

Parabrachial nucleus. In rodents and rabbits the medial parabrachial nucleus (PBN) relays gustatory information from the NTS to the forebrain, while the lateral PBN relays viscerosensory information. PBN lesions produce the most consistent and profound deficits in taste aversion learning. Lesions of medial PBN, lateral PBN, or combined lesions completely block LiCl-induced taste aversion learning. This deficit does not result from an inability to detect the taste stimulus or a failure to experience the malaise induced by LiCl. Preoperatively acquired taste aversions remain intact, indicating that the medial PBN is critical for the acquisition, but not the expression, of

conditioned taste aversions. Medial PBN lesions are also reported to block odor aversion learning, to partially attenuate aversion conditioning to complex flavors, and to spare aversion conditioning to a trigeminal stimulus (capsaicin).

Although an intact PBN is necessary for taste aversion learning, the PBN along with other hindbrain nuclei are not sufficient to mediate such conditioning. Chronic decerebrate rats, which display relatively normal acceptance and rejection responses to IO infusions of sucrose and quinine, respectively, do not learn to reject a sucrose solution when it is paired with LiCl injections. These results indicate that the forebrain is essential for taste aversion learning.

Thalamus. The parvocellular subdivision of the ventroposteromedial nucleus of the thalamus (VPMpc) relays gustatory information to the insular cortex (IC) for further processing. Excitotoxic lesions of VPMpc have little effect on taste aversion learning but accelerate its extinction. These disruptive effects are not produced when electrophysiologically guided small lesions confined to the VPMpc are performed, indicating that this thalamic nucleus is not critical for acquisition of taste aversion learning.

Insular cortex. Early studies indicated that lesions of the insular cortex impair both the acquisition and retention of taste aversion learning, leaving the taste detection and categorization functions intact. Aspiration of IC did not disrupt the development of LiCl-induced taste aversion learning as measured with the conventional bottle method, but dramatically blocked taste aversion learning as evidenced by the failure of the lesions to produce a palatability shift of the CS taste. More recent studies demonstrate that excitotoxic lesions of IC before conditioning only attenuate taste aversion learning but accelerate the extinction of the conditioned aversion when the CS flavor is presented on several days after conditioning. Furthermore, IC lesions made after the acquisition of taste aversion disrupt the retrieval of the learned aversion, indicating that the integrity of IC is required for the retention and retrieval, rather than the acquisition of taste aversion learning. Combined lesions of VPMpc and the gustatory part of the insular cortex, however, almost completely block the establishment of conditioned taste aversion, indicating that sensory inputs to these areas are essential for taste aversion learning.

Amygdala. Large electrolytic lesions of amygdala (AMY) produce deficits in flavor aversion learning whereas excitotoxic lesions, which spare fibers of passage, do not block conditioned taste aversion. More recently, excitotoxic AMY lesions have been observed to impair taste aversion learning when the CS solution is presented by IO infusions, but not when available in bottles. Inconsistent results have also been reported with selective lesions of individual AMY nuclei. These findings, along with neuroimaging data, indicate AMY involvement in taste aversion learning, but this remains to be clarified.

Hypothalamus. Early studies reported that electrolytic lesions of the lateral hypothalamus (LH) disrupted flavor aversion conditioning but that excitotoxic lesions spared taste aversion learning. Recently, excitotoxic LH lesions have been found to block flavor aversion conditioning when there is a long delay (2 h) between the CS and US, but not with shorter or no delays. These finding suggest that LH lesions may accelerate the decay of the flavor memory trace.

c-Fos neuroimaging studies The expression of immediate-early genes, such as *c-fos*, has been used in rodents to map neural pathways involved in flavor aversion learning. Fos-like immunoreactivity (FLI) is induced in the medial intermediate division of the NTS by presentation of a CS taste (sucrose, saccharin) that was previously paired with LiCl; the response to the CS is similar to that induced by LiCl alone. Lesions of either the IC or AMY or hemidecerebration abolish the conditioned expression of FLI in the NTS to the CS taste, but not the unconditioned response to LiCl. This indicates that the increased *c-fos* expression in the NTS is modulated by forebrain structures. With repeated CS–US pairings, FLI expression was obtained in other structures involved in taste aversion learning, such as the central nucleus of AMY and PBN.

Neuropharmacological studies Glutamatergic and cholinergic neural activities are implicated in taste aversion learning. Administration of the muscarinic receptor antagonist, scopolamine, in the IC before presentation of a CS taste impairs the acquisition and consolidation of the taste aversion, whereas IC injection of the glutamatergic antagonist, AP-5, impairs only the consolidation. Recent studies stress the importance of IC cholinergic activity in processing taste information, and glutamatergic activity in the basolateral nucleus of AMY in processing visceral information generated by gastrointestinal malaise. Furthermore, these studies show that AMY glutamatergic efferents to IC are crucial in the association of taste and visceral stimuli during taste aversion learning.

Other neurotransmitters, such as dopamine, are involved in taste aversion learning. Dopamine transmission within the nucleus accumbens, a structure located in the ventral striatum and involved in

food-related reward processes, has been found to be critical for the strengthening of the memory trace of the taste cue, a process indispensable for long-delay taste aversion learning. LH dopamine activity is also implicated in taste aversion conditioning.

Conditioned Flavor Preferences

Flavor preference conditioning by nutrients is well documented in rodents, ruminants, and insects, although only a few human studies have been published. Flavor preferences may be reinforced by both the orosensory and postingestive properties of nutrients, and once acquired can be long lasting. The most common paradigm used to study conditioned flavor preferences is to pair one flavor (CS+) with the nutrient US and a different flavor (CS−) with water on alternate days, and then assess the conditioned preference by presenting the CS+ and CS− flavors in a two-choice test. There are two types of learned flavor preferences, flavor–flavor conditioning and flavor–postingestive nutrient conditioning.

Flavor–Flavor Conditioning

In flavor–flavor learning, the animal develops a preference for a neutral flavor that is paired with an already preferred flavor (e.g., sweet taste). For example, rats that drink an almond-flavored saccharin solution on some days and vanilla-flavored water on other days show a preference for almond over vanilla in a two-bottle test, with both flavors presented in water. Similar preferences are conditioned for flavors added to a sucrose solution, although in this case the postingestive nutrient effects of the sugar may contribute to the preference. Humans learn to prefer odors added to sweet solutions, if they rate the sweet solution as being pleasant, and they also report that the odor has a sweet taste in the absence of the sugar.

Flavor–Postingestive Nutrient Conditioning

To demonstrate flavor–postingestive nutrient learning, the nutrient is typically infused directly into the stomach, thereby eliminating the flavor of the nutrient as a potential US. Holman, in 1968, was the first to report a conditioned flavor preference in rats by pairing the intake of a flavored solution with intragastric (IG) infusions of a liquid diet. Flavor preferences can be conditioned by IG infusions of complete diets or individual macronutrients (glucose, sucrose, maltodextrin, corn oil, casein), as well as of ethanol. In rodents, nutrient-based learning is very robust, can occur over CS–US delays, can convert the innate aversion to bitter and sour tastes into strong preferences, and are long lasting and resistant to extinction. In some but not all cases the preference is associated with an increased hedonic evaluation of the CS+ flavor, as revealed by taste reactivity tests. Also, in some but not all cases IG nutrient infusions may condition an increase in flavor acceptance; animals may consume substantially more of the CS+ flavor paired with the IG nutrient infusion than they do of a CS− flavor paired with IG water infusions.

The nature of the conditioning signal or signals generated by the postingestive actions of nutrients is not well understood, but these signals appear to be distinct from nutrient-produced satiety signals. Evidence suggests that the nutrient US originates in the intestine and/or liver, but not in the stomach. The US does not appear to be transmitted to the brain via vagal afferents, thus implicating nonvagal or humoral pathways. Nutrients differ substantially in their preference-conditioning effects (e.g., glucose vs. fructose), which indicates that there are nutrient-specific unconditioned stimuli, although a more general signal related to energy metabolism may also exist. IG nutrient infusions condition flavor preferences in food-restricted, as well as in freely fed, animals, so energy repletion is not an essential aspect of flavor preference learning. However, flavor preferences can be conditioned by stimuli related to restoration of specific nutrient deficiencies (amino acid imbalance, calcium deficiency).

Brain Mechanisms

Only recently have studies investigated brain mechanisms involved in flavor preference learning and therefore the available information is limited. Initial studies have focused on brain structures implicated in flavor aversion learning.

Area postrema The area postrema is implicated in taste aversion learning and also in detection of circulating nutrients. However, AP lesions do not block flavor preference conditioning produced by IG nutrient infusions.

Parabrachial nucleus Lesions of the medial parabrachial nucleus prevent animals from learning to prefer a CS+ taste (sour, bitter) paired with IG nutrient infusions. The same lesions only attenuate preference conditioning when the CS+ is a complex flavor that has a strong odor component. The preference learning deficit appears to represent a problem with forming a CS–US (taste–nutrient) association rather than a problem in detecting the CS taste or nutrient US, although this issue requires further study.

Amygdala Lesions of the basolateral amygdala attenuate, but do not block, flavor preference conditioning by IG nutrient infusions. Flavor preference

conditioning is completely blocked, however, by large lesions that include all AMY nuclei. The same lesions impair, but do not prevent, taste preference conditioning by IG nutrients. Thus, the sensory learning deficits produced by AMY lesions are opposite to those produced by PBN lesions.

Lateral hypothalamus Lateral hypothalamus lesions do not prevent rats from acquiring preferences for flavors paired with IG nutrient infusions, although the preferences are weaker than those displayed by intact rats. The lesions, however, totally block preference learning when the IG nutrient infusions are not concurrent with consumption of the CS+ flavor but instead occur after a short delay (15 min). It is possible that this learning deficit reflects a lesion-induced decay in the CS+ memory trace, but where and how this trace is maintained are unknown.

Insular cortex The insular cortex receives gustatory, olfactory, and visceral information, suggesting a role in flavor perception and modification by postingestive feedback. However, the IC does not seem to be critical for flavor preference learning, since large lesions, including gustatory and visceral regions, do not prevent rats from learning to prefer a CS+ flavor paired with IG nutrient infusions.

Neuropharmacological Studies

Evidence implicates brain dopamine systems in flavor preference conditioning by nutrients. Rats trained to prefer a bitter solution by pairing its intake with IG nutrient infusions subsequently showed enhanced dopamine (DA) release in the nucleus accumbens when drinking the solution, compared to unconditioned animals offered the same solution. Flavor preference conditioning by IG nutrient infusions is blocked by systemic treatment with the D1 receptor antagonist SCH23390 during training trials. However, the same drug does not prevent the expression of a previously conditioned flavor preference when administered at the time of two-bottle testing. Treatment with a D2 receptor antagonist (raclopride) has little effect on the acquisition or expression of preference conditioning by IG nutrient infusions. In contrast, both D1 and D2 receptor antagonists prevent flavor preference conditioning by the sweet taste of fructose, and SCH23390 also attenuates the expression of the sweet-taste-based preference. These data indicate differential involvement of dopamine receptors in flavor–flavor and flavor–postingestive nutrient conditioning.

To date, systemic treatment with the opioid receptor antagonist naltrexone has not interfered with the acquisition or expression of flavor preferences conditioned by the sweet taste or postingestive actions of sugar. This contrasts with the well-documented effect of opioid antagonists on the intake of sweet solutions. The influence of other neurotransmitter systems on flavor preference learning remains to be investigated.

See also: Central Gustatory System and Ingestive Behavior; Conditioned Taste Aversion; Cortical Processing of the Reward Value of Food; Flavor Physiology; Nutrient Sensing: Cellular Metabolism; Nutrient Sensing: Carbohydrates; Nutrition; Sensory Aging: Chemical Senses; Taste: Vertebrate Taste Bud Physiology; Taste: Vertebrate Central Pathways.

Further Reading

Bermudez-Rattoni F, Ramirez-Lugo L, Gutierrez R, et al. (2004) Molecular signals into the insular cortex and amygdala during aversive gustatory memory formation. *Cellular and Molecular Neurobiology* 24: 25–36.

Brunstrom JM (2005) Dietary learning in humans: Directions for future research. *Physiology & Behavior* 85: 57–65.

Bures J, Bermúdez-Rattoni F, and Yamamoto T (1998) *Conditioned Taste Aversion: Memory of a Special Kind*. Oxford: Oxford University Press.

Capaldi ED (1996) *Why We Eat What We Eat*. Washington, DC: American Psychological Association.

Garcia J, Kimeldorf DJ, and Koelling RA (1955) Conditioned aversion to saccharin resulting from exposure to gamma radiation. *Science* 122: 157–158.

Grill HJ and Norgren R (1978) Chronically decerebrate rats demonstrate satiation but not bait shyness. *Science* 201: 267–269.

Holman GL (1968) Intragastric reinforcement effect. *Journal of Comparative Physiology and Psychology* 69: 432–441.

Houpt TA (2000) Molecular neurobiology of ingestive behavior. *Nutrition* 16: 827–836.

Parker LA (2003) Taste avoidance and taste aversion: Evidence for two different processes. *Learning & Behavior* 31: 165–172.

Reilly S (1998) The role of the gustatory thalamus in taste-guided behavior. *Neuroscience & Biobehavioral Reviews* 22: 883–901.

Reilly S (1999) The parabrachial nucleus and conditioned taste aversion. *Brain Research Bulletin* 48: 239–254.

Reilly S and Bornovalova MA (2005) Conditioned taste aversion and amygdala lesions in the rat: A critical review. *Neuroscience & Biobehavioral Reviews* 29: 1067–1088.

Sclafani A (1999) Macronutrient-conditioned flavor preferences. In: Berthoud H-R and Seeley R (eds.) *Neural and Metabolic Control of Macronutrient Intake*, pp. 93–107. Boca Raton, FL: CRC Press.

Sclafani A (2004) Oral and postoral determinants of food reward. *Physiology & Behavior* 81: 773–779.

Touzani K and Sclafani A (2005) Critical role of amygdala in flavor but not taste preference learning in rats. *European Journal of Neuroscience* 22: 1767–1774.

Yeomans MR (2006) Olfactory influences on appetite and satiety in humans. *Physiology & Behavior* 87: 800–804.

Relevant Website

http://ctalearning.com – Conditioned taste aversion: An annotated bibliography.

Learning and Memory in Invertebrate Models: *Tritonia*

W N Frost and E V Megalou, Rosalind Franklin University of Medicine and Science, North Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Why Study Learning in Invertebrates?

Deciphering the neuronal mechanisms by which we acquire, store, and recall memories represents one of the major challenges for modern neuroscience. This topic is being pursued simultaneously in both vertebrate and invertebrate preparations. Given the indisputable relevance of studying learning mechanisms in vertebrates, it seems appropriate to consider why many investigators choose to work on invertebrates. First, invertebrates display many of the same types of learning enjoyed by vertebrates, including habituation, dishabituation, sensitization, classical conditioning, second order conditioning, sensory preconditioning, latent inhibition, overshadowing, blocking, context conditioning and, operant conditioning. Second, the synaptic plasticity mechanisms that appear to encode learning turn out to be highly conserved. Thus, invertebrates have been shown to have synapses that undergo the same types of experience-dependent plasticity studied in vertebrates, including paired-pulse facilitation, homosynaptic depression, presynaptic inhibition, presynaptic facilitation, posttetanic potentiation, long-term potentiation, and long-term depression. Third, molecular mechanisms also appear to be conserved, including the involvement of particular second messenger systems in short-term memory and the requirement for gene expression for the formation of stable, long-term memories. Finally, a key reason for studying learning mechanisms in invertebrates is their seemingly made-for-science nervous systems. For example, invertebrate neurons are larger (up to 1 mm in diameter) and far fewer in number than those of vertebrates and are conveniently located on the outside of their ganglia, where they can be seen and penetrated with electrodes or observed with optical recording techniques. Many of these neurons can be recognized as unique individuals, making it possible to map out discrete circuits and to return to the same neurons and synapses in successive preparations in order to characterize their role in information storage. Furthermore, because these neurons and circuits can often be studied *in situ*, the specific behavioral function of these synapses, and thus the information they encode, can often be determined, a feat seldom possible in vertebrates.

An Invertebrate Model System: The Marine Mollusk *Tritonia*

On contact with the tube feet of its sea star predators, the nudibranch mollusk *Tritonia diomedea* launches a stereotyped escape response consisting of a rapid withdrawal of gills, rhinophores, and the affected body site, followed by a vigorous escape swim (Figure 1). This rhythmic behavior consists of a series of alternating ventral and dorsal whole-body flexions that serve to get the animal up off the substrate and into the current, where it is wafted away to safety. The escape swim has a duration of 2 to 20 flexion cycles that take up to 2 min to play out. As the swim ceases, the animal then settles to the substrate and typically crawls for several minutes.

Work by several laboratories has defined the major components of the neural network mediating the *Tritonia* escape swim. Aversive skin stimuli are detected by the S-cells, a group of central afferent neurons that send an axon out to the skin. The polymodal S-cells respond to both tactile and chemical skin stimuli, including a substance present in the tube feet of predatory sea stars and many other echinoderms. The S-cells in turn make monosynaptic connections onto the Tr1 and dorsal ramp interneuron (DRI) command neurons. DRI is a paradigm example of a gating command neuron: it is activated by sensory stimuli, and its own firing is both necessary and sufficient for sensory input to trigger the swim motor program. DRI is able to exert such impact because sensory input is funneled to DRI, and it in turn makes the strongest excitatory synaptic connections yet recorded in *Tritonia*, onto the dorsal swim interneurons (DSIs) of the swim central pattern generator (CPG). Action potentials in DRI elicit nearly one-for-one firing of the six DSI neurons of the CPG. The swim CPG was described 20 years ago as consisting of three neuron groups, the DSIs, ventral swim interneurons (VSIs), and cerebral interneuron 2 (C2). This group of neurons was suggested to oscillate by virtue of their synaptic connections to one another, with no neurons having intrinsic bursting properties of their own. The DSIs have also been found to utilize the transmitter serotonin to rapidly modulate the excitability and synaptic strength of the other CPG neurons, a process believed to play a central role in reconfiguring the resting crawling network into a temporary swim CPG. This represents an example of intrinsic neuromodulation, in which modulatory neurons are embedded in the network, causing it to modulate itself every time it is activated. This is quite different from the more familiar extrinsic neuromodulation, in which the

Figure 1 *Tritonia* escape behavior and its underlying swim motor program. The sea slug *Tritonia diomedea* is shown in mid-escape from its sea star predator, *Pycnopodia helianthoides.* Also shown are three simultaneous intracellular recordings from the interneurons of the escape swim central pattern generator: dorsal swim interneuron (DSI), ventral swim interneuron (VSI), and cerebral interneuron 2 (C2). Each burst of action potentials drives one full dorsal or ventral body flexion of the rhythmic escape swim. Record duration: 25 s.

modulation is optional because the modulatory neurons are located outside the network.

The Escape Swim Is Produced by a Multifunctional Neural Network

A remarkable feature of *Tritonia*'s swim network is its degree of multifunctionality; it has been suggested to play a role in at least three different behaviors: reflexive withdrawal, swimming, and crawling. Multifunctional networks are well documented in invertebrates, in which they have typically been found to mediate structurally similar motor programs, such as those mediating different rhythmic aspects of feeding. *Tritonia*'s network is notable with respect to how dissimilar the behaviors it mediates are to one another. Reflexive withdrawal is a brief muscle-mediated retraction of body areas to skin stimulation. Swimming is a 1–2 min rhythmic behavior involving a series of alternating whole-body dorsal and ventral contractions mediated by body wall muscles. Crawling is a nonrhythmic, nonmuscular behavior mediated by the beating of millions of microscopic cilia on the animal's foot. After the swim, the same DSIs that fired rhythmically to mediate the swim then continue to fire tonically for up to an hour, when, via excitatory connections to identified foot cilia neurons, they mediate crawling. While multifunctional networks have been described in many preparations, *Tritonia*'s swim network is unusual in that it involves the most dissimilar behaviors yet found to be mediated by a single neural network, cilia-mediated, nonrhythmic crawling and muscle-mediated, rhythmic swimming.

Experience-Dependent Modifications of the Swim

Tritonia is being used to study how experience alters behavior. These studies have so far focused on prepulse inhibition and on two forms of nonassociative learning: sensitization and habituation.

Prepulse Inhibition

Most animals experiencing an unexpected strong stimulus respond with a startle response – in *Tritonia*'s case, its high-threshold escape swim. Given their robust, protective nature, it can be surprising to discover that startle responses are often significantly inhibited if a weak stimulus of almost any modality is experienced just beforehand. This experimental paradigm, termed prepulse inhibition (PPI), exposes an inhibitory sensorimotor gating mechanism whose underlying cellular mechanisms have attracted increasing interest ever since PPI was found to be defective in schizophrenia. It has been widely speculated that reduced PPI may contribute to the distractibility and sensory flooding experienced by sufferers of this disease. Could invertebrate preparations be used to gain insights into the nervous system mechanisms mediating the clinically relevant phenomenon of PPI? Pursuing this line of thinking, *Tritonia* was recently tested and shown to have PPI that is parametrically similar to that of vertebrates. In this paradigm, a brief vibrotactile stimulus delivered to the animal's midbody beginning just before a tail shock startle stimulus prevents the startle response (escape swim) that normally occurs to the shock alone. PPI does not involve learning; it is a brief, inhibitory sensorimotor gating mechanism that works, in humans and seaslugs, the very first time the prepulse is presented.

Further research has revealed two related neuronal mechanisms for PPI in *Tritonia*. First, the inhibitory interneurons that mediate PPI were identified. These were found to do so, in part, via a well-known mechanism, presynaptic inhibition of transmitter release, acting on the startle pathway

afferent neurons, reducing their ability to excite the swim circuitry. These neurons also produce conduction block of incoming action potentials in the startle pathway, causing many to fail before reaching their release sites. Because afferent pathways as well as many or most central interneuronal synapses in vertebrates are subject to presynaptic inhibition, these basic mechanisms described in *Tritonia* seem suitable to mediate PPI in vertebrates as well.

Nonassociative Learning: Sensitization

Sensitization is a form of nonassociative learning (experienced by most or all animals) in which the memory of a startling or aversive stimulus is encoded as a period of enhanced general defensive responsiveness. The neural mechanisms mediating sensitization have been studied extensively in the marine mollusk *Aplysia*, in which a single tail shock enhances the animal's tail- and siphon-elicited withdrawal responses for about an hour.

Tritonia is also being used to explore the cellular and network mechanisms mediating sensitization. Following a single aversive swim-eliciting stimulus, *Tritonia* becomes sensitized, as revealed by an hour-long period of enhanced defensive responsiveness. During this time, the animal's gill and rhinophore withdrawal reflexes display a shorter onset latency, and test swims have a lower threshold, faster onset latency, and higher cycle number.

Two nervous system alterations have been identified so far that appear to mediate these learning-related behavioral modifications: the functional strength of the synapses made by CPG neuron C2 become markedly enhanced, and the DSI neurons undergo an hour-long period of elevated spontaneous firing. Since, as discussed above, the DSIs produce a direct, modulatory enhancement of the synapses made by C2 and other CPG neurons, this prolonged DSI firing may represent a primary locus of the memory for sensitization, actively maintaining C2 synapses in their enhanced state. This is supported by the additional finding that forcing the DSIs to fire at an elevated tonic rate causes the same shortening of swim onset latency as seen in sensitization. The *Tritonia* preparation thus has allowed us to determine the information content, that is, behavioral impact, of a specific memory storage locus by selectively implanting it into the naive brain and determining its contribution to the motor program changes seen in learning.

Nonassociative Learning: Habituation

Habituation is a universal form of nonassociative learning. In habituation, behavioral responsiveness to a test stimulus decreases with repetition. It has the important function of enabling us to ignore repetitive, irrelevant stimuli so that we can remain responsive to sporadic stimuli, typically of greater significance. The widespread occurrence of habituation across phyla supports its importance, and the strong similarity in its parametric features across species is consistent with the possibility that common mechanisms mediate it. As with sensitization, the memory for habituation can be short term, lasting minutes to hours, or long term, lasting days. The neural mechanisms underlying habituation are of clinical interest. For example, long-term habituation is deliberately induced in therapeutic procedures used to treat chronic anxiety and posttraumatic stress disorder. In part because these conditions may involve an impaired ability to habituate to stressful stimuli, studies have begun using effects on habituation as a screening test for developing and comparing anxiolytic drugs.

Based on observations in the open ocean, *Tritonia*'s escape swim habituation may help prevent animals from unnecessary swimming, which would tend to carry them away in water currents and thus away from their spatially restricted food supply. In the laboratory, repeated swim stimuli lead to the progressive alteration of several behavioral features, including a lengthening of gill and rhinophore withdrawal latency, lengthening of swim onset latency, and a decrease in swim cycle number and swim flexion speed. Habituation of the *Tritonia* swim shows several of the parametric features of vertebrate habituation, including response decrement with a negative exponential time course, spontaneous recovery, faster decrement with shorter intertrial intervals, dishabituation, increased threshold, and stimulus site generalization. These similarities are consistent with the likelihood that habituation in *Tritonia* may share cellular mechanisms with those mediating habituation in vertebrates.

As mentioned above, the motor program mediating the swim can be elicited and studied in isolated brain preparations, in which cellular mechanisms can be more readily studied. Repeatedly elicited isolated brain swim motor programs habituate with the same negative exponential time course as the behavior, indicating that the key sites of plasticity are located in the central nervous system. Swim habituation thus cannot be explained by peripheral effects such as receptor adaptation or muscle fatigue.

The cellular loci mediating habituation in *Tritonia* are not yet well characterized. As in many organisms, the strength of the sensory neuron synapses mediating the test response decreases profoundly when the sensory neurons are made to fire single action potentials separated by several seconds. Recently, the probable transmitter of the sensory neurons was identified to

be glutamate, opening the way for studies of whether the activity-dependent synaptic decrement at these synapses is mediated by pre- or postsynaptic processes. In addition to this possible sensory neuron memory locus, the demonstration that habituation produced by stimulation of one sensory population generalizes to unstimulated sensory neurons suggests that the memory for habituation may involve interneuronal sites of plasticity as well.

Studies of Other *Tritonia* Behaviors

Recent observations of *Tritonia* in its natural habitat have documented several behaviors in their natural context that have long been observed in the laboratory, including feeding, mating, rheotaxis, and escape swimming following contact with the tube feet of predatory sea stars. Olfaction appears to play an important role in the natural habitat. Time lapse video studies have shown that *Tritonia* crawls upstream toward food and mates, and downstream away from predators. As described above, crawling occurs via the beating of cilia on the bottom of the foot. These are under neuronal control, and their rate of beating, and thus the speed of locomotion, is modulated by the transmitter dopamine as well as by peptide neurotransmitters. An important crawling behavior displayed by *Tritonia* is rheotaxis – orientation and crawling toward the direction of water flow. This ability is mediated via neurons having axons in the lateral branch of cerebral nerve 2, as well as by a number of identified pedal ganglion motor neurons. One neuron that controls turning while crawling, pedal neuron 3, acts by lifting the foot margin on the side ipsilateral to the direction of the intended turn, removing those cilia from contact with the substrate.

Tritonia has also been shown capable of orienting to the earth's magnetic field, which has been speculated to help it navigate with respect to potential food sources. Pedal ganglion locomotion neurons have been identified that alter their firing in response to artificially induced changes in the orientation of the magnetic field surrounding the animal.

In summary, *Tritonia* holds promise for investigations of the neural basis of many interesting aspects of behavior present in higher animals. These include prepulse inhibition, sensitization, and habituation, as well as specific behaviors such as swimming, locomotion, and magnetic field orientation.

See also: Conditioning: Theories; Cyclic AMP (cAMP) Role in Learning and Memory; Learning and Memory in Invertebrates: Mollusks; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Olfactory System: Circuit Dynamics and Neural Coding in the Locust; Operant Conditioning of Reflexes; Procedural Learning: Classical Conditioning.

Further Reading

Frost WN, Brandon CL, and Mongeluzi DL (1998) Sensitization of the *Tritonia* escape swim. *Neurobiology of Learning and Memory* 69: 126–135.

Frost WN, Brandon CL, and Van Zyl C (2006) Long-term habituation in the marine mollusc *Tritonia diomedea*. *Biological Bulletin* 210: 230–237.

Frost WN, Brown G, and Getting PA (1996) Parametric features of habituation of swim cycle number in the marine mollusc *Tritonia diomedea*. *Neurobiology of Learning and Memory* 65: 125–134.

Frost WN, Hoppe TA, Wang J, and Tian L-M (2001) Swim initiation neurons in *Tritonia diomedea*. *American Zoologist* 41: 952–961.

Frost WN, Tian L-M, Hoppe TA, Mongeluzi DL, and Wang J (2003) A cellular mechanism for prepulse inhibition. *Neuron* 40: 991–1001.

Getting PA (1983) Neural control of swimming in *Tritonia*. *Symposia of the Society for Experimental Biology* 37: 89–128.

Katz PS, Getting PA, and Frost WN (1994) Dynamic neuromodulation of synaptic strength intrinsic to a central pattern generator circuit. *Nature (London)* 367: 729–731.

Mongeluzi DL and Frost WN (2000) Dishabituation of the *Tritonia* escape swim. *Learning and Memory* 7: 43–47.

Murray JA and Willows AO (1996) Function of identified nerves in orientation to water flow in *Tritonia diomedea*. *Journal of Comparative Physiology. A, Sensory, Neural, and Behavioral Physiology* 178: 201–209.

Popescu IR and Frost WN (2002) Highly dissimilar behaviors mediated by a multifunctional network in the marine mollusk *Tritonia diomedea*. *Journal of Neuroscience* 22: 1985–1993.

Redondo RL and Murray JA (2005) Pedal neuron 3 serves a significant role in effecting turning during crawling by the marine slug *Tritonia diomedea* (Bergh). *Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology* 191: 435–444.

Sakurai A and Katz PS (2003) Spike timing-dependent serotonergic neuromodulation of synaptic strength intrinsic to a central pattern generator circuit. *Journal of Neuroscience* 23: 10745–10755.

Wang JH, Cain SD, and Lohmann KJ (2003) Identification of magnetically responsive neurons in the marine mollusc *Tritonia diomedea*. *Journal of Experimental Biology* 206: 381–388.

Woodward OM and Willows AO (2006) Dopamine modulation of Ca(2+) dependent Cl(−) current regulates ciliary beat frequency controlling locomotion in *Tritonia diomedea*. *Journal of Experimental Biology* 209: 2749–2764.

Wyeth RC and Willows AO (2006) Field behavior of the nudibranch mollusc *Tritonia diomedea*. *Biological Bulletin* 210: 81–96.

Learning and Memory in Invertebrates: *Aplysia*

J H Byrne, E G Antzoulatos, and D Fioravante, University of Texas Medical School at Houston, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The survival of all animals is dependent on their capacity to adapt to their environment by modifying their behavior. The experience-induced modification of behavior is a manifestation of learning, whereas memory is the retention of a learned behavior over time. A conceptual scheme that has driven the investigation of learning rests on the distinction between associative and nonassociative forms of learning. Nonassociative forms of learning include habituation and sensitization. Habituation is defined as the gradual waning of a behavioral response to a weak or moderate stimulus that is presented repeatedly. Following habituation, the response may be restored to its initial state, either passively with time (i.e., spontaneous recovery) or with the presentation of a novel stimulus (i.e., dishabituation). Sensitization is defined as the enhancement of a behavioral response elicited by a weak stimulus following another, usually noxious, stimulus. Sensitization can also develop in response to a moderate stimulus that is presented repeatedly at relatively short intervals.

Associative learning refers to the formation of an association either between two stimuli (i.e., classical conditioning) or between a behavior and a stimulus (i.e., operant conditioning). In classical conditioning, a novel or weak stimulus (conditional stimulus (CS)) is paired with a stimulus that generally elicits a reflexive response (unconditional stimulus (US) and unconditional response(UR)). After sufficient training with contingent CS–US presentations (which may be a single trial), the CS comes to elicit a learned response (conditional response (CR)), which often resembles the UR (or some aspect of it). The emergence of the anticipatory CR reveals that the CS has become a reliable predictor of the US. Operant conditioning is an experimental procedure in which the behavior of an animal may be followed by either a desirable or an aversive stimulus, arranged by the experimenter. The desirable stimulus (e.g., food) typically increases the future occurrence of this behavior (a process called positive reinforcement). An aversive stimulus (e.g., a noxious electric shock) tends to decrease the future probability of this behavior (a process called punishment). A behavior can also be reinforced when it becomes contingent with the removal of an aversive stimulus from the animal's environment (i.e., negative reinforcement). Thus, through the processes of operant conditioning an animal learns the outcomes of its behavior.

Humans and other animals are capable of displaying more complex forms of learning than the four types just described. However, these four types are likely to constitute the building blocks for more complex forms of learning. Thus, a major goal of neurobiologists is to explain the anatomical, biophysical, and molecular processes of the nervous system that underlie simple forms of learning and memory. Specifically, what parts of the nervous system are critical for learning? How is information about a learned event acquired and encoded in neuronal terms? How is the information stored, and, once stored, how is it retrieved? Most neuroscientists believe that the answers to these questions lie in understanding the ways in which the properties of individual nerve cells in general, and synaptic connections, in particular, change when learning occurs.

Emerging Principles

As a result of research on several vertebrate and invertebrate model systems, some general principles have emerged: (1) at least short-term forms of learning and memory require changes in existing neural circuits; (2) these changes may involve multiple cellular mechanisms within single neurons; (3) second-messenger systems appear to play a role in mediating cellular changes; (4) changes in the properties of membrane channels are commonly correlated with learning and memory; (5) changes in synaptic efficacy are correlated with short- and long-term memory; and (6) long-term memory requires new protein synthesis and growth, whereas short-term memory does not.

One animal that is well suited for the examination of the molecular, cellular, morphological, and network mechanisms underlying neuronal plasticity and learning and memory is the marine mollusc *Aplysia*. This animal has a relatively simple nervous system with large identifiable neurons that are accessible for detailed anatomical, biophysical, and biochemical studies. Neurons and neural circuits that mediate many behaviors in *Aplysia* have been identified. In several cases, these behaviors have been shown to be modified by learning. Moreover, specific loci within neural circuits at which modifications occur during learning have been identified and aspects of the cellular mechanisms underlying these modifications have been analyzed and modeled.

Behaviors and Neural Circuits

The siphon-gill and tail-siphon withdrawal reflexes of *Aplysia* have been used extensively to analyze the neuronal mechanisms contributing to nonassociative and associative learning. Feeding behavior has also emerged as a useful model system to study associative learning. It exhibits both classical and operant conditioning and neural correlates and mechanisms of both types of associative learning have been identified. In this article, we first focus on the analyses of habituation, sensitization, and classical conditioning of defensive reflexes; later, we describe the analysis of operant and classical conditioning of feeding behavior.

Withdrawal Reflexes

Within the mantle cavity is the respiratory organ of *Aplysia*, the gill, and protruding from the mantle cavity is the siphon (Figure 1). The siphon-gill withdrawal reflex is elicited when a tactile or electrical stimulus is delivered to the siphon and results in the withdrawal of the siphon and gill (Figure 1(a)). A second behavior that has been examined extensively is the tail-siphon withdrawal reflex. Tactile or electrical stimulation of the tail elicits a coordinated set of defensive responses, two components of which are a reflex withdrawal of the tail and the siphon (Figure 1(b)).

A prerequisite for the analysis of the neural and molecular basis of learning is an understanding of the neural circuit that controls the behavior. The afferent limb of the siphon-gill withdrawal reflex consists of sensory neurons with somata in the abdominal ganglion. The siphon sensory neurons (SN) monosynaptically excite the gill and siphon motor neurons (MN) that are also located in the abdominal ganglion (Figure 2(a)). Activation of the gill and siphon motor neurons leads to the contraction of the gill and siphon. Excitatory, inhibitory, and modulatory interneurons (INs) in the withdrawal circuit have also been identified, although only excitatory INs are illustrated in Figure 2. The afferent limb of the tail-siphon withdrawal reflex consists of a bilaterally symmetric cluster of sensory neurons that are located in the left and right pleural ganglia. These sensory neurons make monosynaptic excitatory connections with motor neurons in the adjacent pedal ganglion, which produce the withdrawal of the tail (Figure 2(b)). In addition, the tail sensory neurons form synapses with various identified excitatory and inhibitory INs. Some of these INs activate MNs in the abdominal ganglion, which control the reflex withdrawal of the siphon. Moreover, several additional neurons modulate the tail-siphon withdrawal reflex (see Figure 3(a)).

The sensory neurons for both the siphon-gill and tail-siphon withdrawal reflexes are similar and appear to be important plastic elements in the neural circuits. Changes in their membrane properties and the strength of their synaptic connections (synaptic efficacy) are associated with learning and memory. Moreover, the properties of these neurons are modulated by *in vitro* analogs of behavioral training.

Cellular and Molecular Mechanisms in Sensory Neurons Contributing to Habituation and Sensitization in *Aplysia*

Short-Term Memory for Habituation

Upon brief repetitive stimulation, the withdrawal reflexes of *Aplysia* undergo short-term habituation,

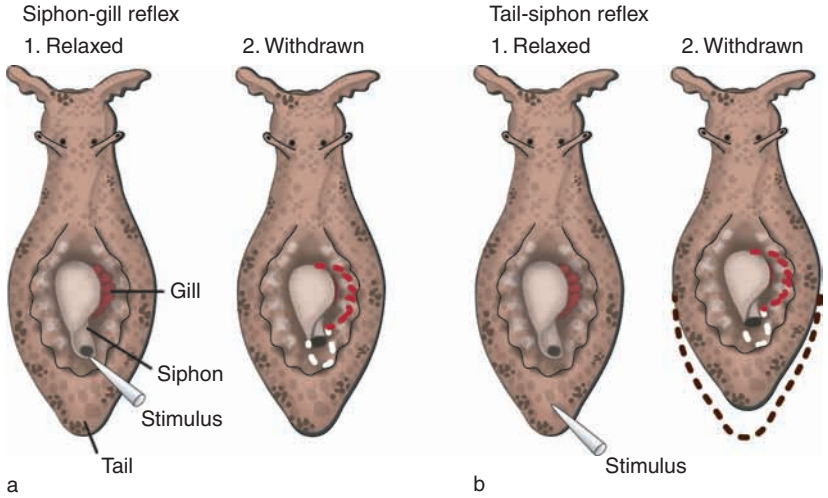

Figure 1 Siphon-gill and tail-siphon withdrawal reflexes of *Aplysia:* (a) siphon-gill withdrawal; (b) tail-siphon withdrawal reflex. (a) Shows a dorsal view of *Aplysia* (1) in a relaxed position and (2) after a stimulus (e.g., a water jet, brief touch, or weak electric shock) was applied to the siphon, causing the siphon and the gill to withdraw into the mantle cavity. (b) Shows *Aplysia* (1) in a relaxed position and (2) after a stimulus was applied to the tail, eliciting a reflex withdrawal of the tail and siphon.

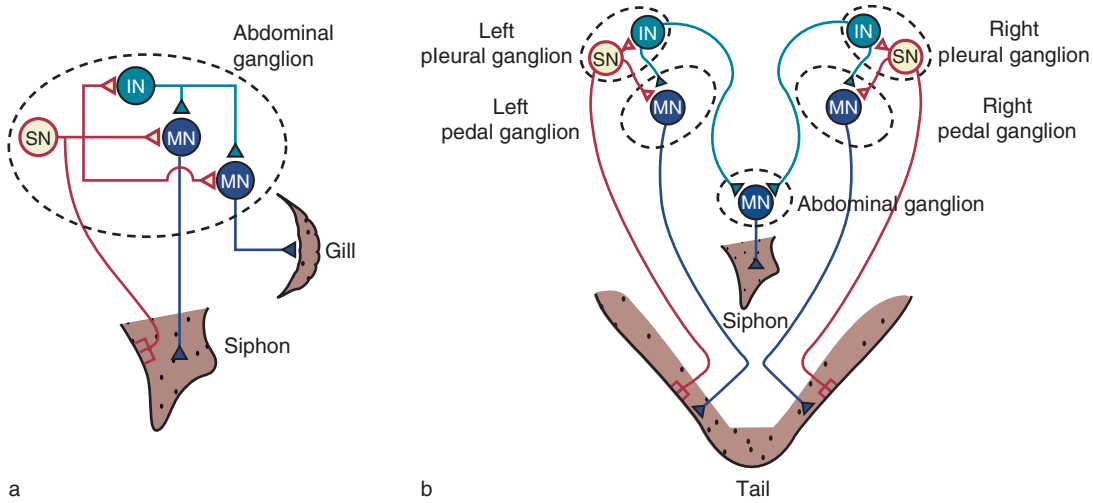

Figure 2 Simplified circuit diagrams: (a) the siphon-gill reflex; (b) tail-siphon withdrawal reflex. Stimuli activate the afferent terminals of mechanoreceptor sensory neurons (SN) whose somata are located in central ganglia. The sensory neurons make excitatory synaptic connections (triangles) with interneurons (IN) and motor neurons (MN). The excitatory interneurons provide a parallel pathway for excitation of the motor neurons. Action potentials elicited in the motor neurons, triggered by the combined input from the SNs and INs, propagate out peripheral nerves to activate muscle cells and produce the subsequent reflex withdrawal of the organs. Modulatory neurons (not shown here, but see **Figure 3(a)**), such as those containing serotonin, regulate the properties of the circuit elements and, consequently, the strength of the behavioral responses.

which has been attributed primarily to activity-dependent depression of the sensorimotor synapse. This depression, which is termed homosynaptic, leads to a decreased recruitment of MNs and decreased magnitude of the withdrawal reflexes. A proposed mechanism of short-term homosynaptic depression involves presynaptic changes in transmitter release. Experimental and computational studies of the sensorimotor synapse have suggested that short-term activity-dependent changes in transmitter release are in part determined by the dynamics of vesicle mobilization in presynaptic terminals and are limited by the gradual depletion of available vesicles.

In addition to presynaptic mechanisms of depression, postsynaptic mechanisms have also been implicated. For example, challenging the sensorimotor synapse with physiological bursts of activity results in its depression, partly due to the desensitization of postsynaptic receptors. In addition, more prolonged depression of the sensorimotor synapse (i.e., around 2 h) relies on the activation of postsynaptic *N*-methyl-D-aspartate (NMDA)-like receptors and a rise in postsynaptic Ca^{2+}. Thus, despite the apparent operational simplicity of habituation and synaptic depression, the underlying mechanisms appear to be complex and to involve multiple targets.

Long-Term Memory for Habituation

Both short- and long-term habituation share aspects of a common mechanism, synaptic depression. However, whereas short-term synaptic depression arises primarily from transient changes in release, long-term depression has been attributed to persistent structural changes in SNs. Extensive morphological analyses of SNs from habituated animals have revealed that the number of synaptic contacts is reduced compared to controls. Moreover, the structure of presynaptic terminals is affected, with fewer synaptic vesicles and reduced size of active zones (the sites of transmitter release).

Short-Term Memory for Sensitization

Short-term sensitization is induced when a single brief train of shocks to the body wall results in the release of modulatory transmitters, such as serotonin (5-HT), from a separate class of INs, referred to as facilitatory neurons (**Figure 3(a)**). These facilitatory neurons regulate the properties of the SNs and the strength of their connections with postsynaptic INs and MNs, a process called heterosynaptic facilitation (**Figures 3(b) and 3(c)**). The molecular mechanisms contributing to heterosynaptic facilitation are illustrated in **Figure 3(d)**. The binding of 5-HT to one class of receptors on the outer surface of the membrane of the SNs leads to the activation of adenylyl cyclase, which, in turn, leads to an elevation of the intracellular level of the second-messenger adenosine-3′,5′-monophosphate (cAMP) in SNs. When cAMP binds to the regulatory subunit of cAMP-dependent protein kinase, protein kinase A (PKA), the catalytic

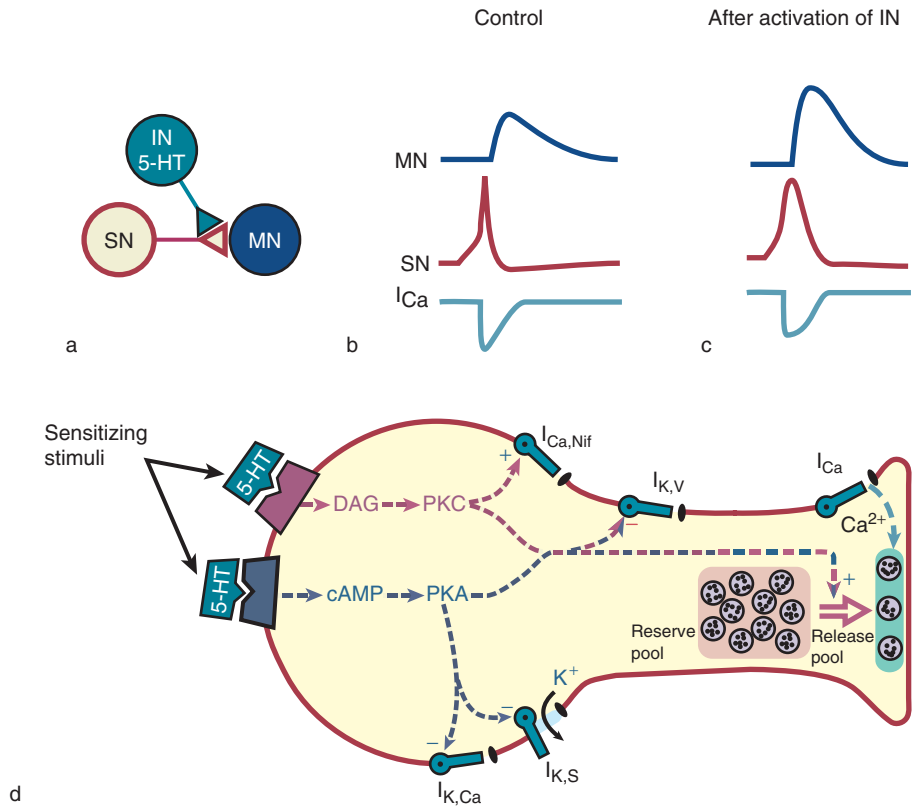

Figure 3 Model of heterosynaptic facilitation of the sensorimotor connection that contributes to short-term sensitization in *Aplysia:* (a) sensitizing stimuli activate facilitatory interneurons (IN) that release modulatory transmitters, one of which is serotonin (5-HT); (b) action potential in a sensory neuron (SN) in the control; (c) action potential in a SN after the sensitizing stimulus; (d) molecular events in the sensory neuron. In (a), the modulator leads to an alteration of the properties of the SN. In (c), the sensitizing stimulus results in greater transmitter release and, hence, a larger postsynaptic potential in the motor neuron (MN) than an action potential prior to the sensitizing stimulus (shown in (b)). For short-term sensitization, the enhancement of transmitter release is due, at least in part, to the broadening of the action potential and an enhanced flow of Ca^{2+} into the SN. In (d), 5-HT released from the facilitatory interneuron (a) binds to at least two distinct classes of receptors on the outer surface of the membrane of the SN, which leads to the transient activation of two intracellular second messengers: DAG and cAMP. These second messengers, acting though their respective protein kinases, affect multiple cellular processes, the combined effects of which lead to enhanced transmitter release when a subsequent action potential is fired in the SN. cAMP, cyclic adenosine-3′,5′-monophosphate; DAG, diacylglycerol; I_{Ca}, calcium current; $I_{Ca,Nif}$, nifedipine-sensitive Ca^{2+} current; $I_{K,S}$, S-K^+ current; $I_{K,V}$, delayed K^+ current; PKA, protein kinase A; PKC, protein kinase C.

subunit is freed and can now add phosphate groups to specific substrate proteins and, hence, alter their functional properties.

One consequence of this protein phosphorylation is an alteration of the properties of membrane channels. Specifically, the increased levels of cAMP lead to a decrease in the S-K^+ current ($I_{K,S}$), a component of the calcium-activated K^+ current ($I_{K,Ca}$) and the delayed K^+ current ($I_{K,V}$). These changes in membrane currents lead to the depolarization of the membrane potential, enhanced excitability, and an increase in the duration of the action potential.

Through the activation of PKA, cAMP also appears to mediate a membrane potential- and spike-duration-independent process of facilitation, represented in Figure 3(d) (large open arrow) as the mobilization of transmitter-containing vesicles from a storage pool to a releasable pool, making more transmitter available for release with subsequent action potentials in the SN. The phosphoprotein synapsin may be involved in the regulation of synaptic vesicle availability. Under basal conditions, synapsin associates with vesicles as well as the cytoskeleton, forming a proteinaceous meshwork that limits vesicle mobilization. Upon phosphorylation by PKA, synapsin dissociates from vesicles, thus rendering them available to be mobilized to release sites. Another molecule recently implicated in the regulation of vesicle mobilization is the Ca^{2+}-activated enzyme calpain. The combined effects of spike broadening and vesicle mobilization appear to contribute to the short-term cAMP-dependent enhancement of transmitter release.

5-HT also appears to act through another class of receptors to increase the level of the second

messenger diacylglycerol (DAG). DAG activates protein kinase C (PKC). PKC, like PKA, is involved in the spike-duration-independent process of facilitation. In addition, a nifedipine-sensitive Ca^{2+} current ($I_{Ca,Nif}$) and $I_{K,V}$ are regulated by PKC. Thus, $I_{K,V}$ is dually regulated by PKC and PKA. The modulation of $I_{K,V}$ contributes significantly to the increase in the duration of the action potential (Figure 3(c)). Because of its small magnitude, the modulation of $I_{Ca,Nif}$ appears to play a minor role in the facilitatory process.

The consequences of the activation of these multiple second-messenger systems and the modulation of multiple cellular processes become apparent on test stimulation of SNs at various times after the sensitizing stimuli (Figure 3(c)). More transmitter is available for release as a result of the mobilization process, and each action potential is broader, allowing a greater influx of Ca^{2+} to trigger the release of the available transmitter. The combined effects of mobilization and spike broadening lead to the enhanced release of transmitter from the SN and consequently a larger postsynaptic potential in the MN. Larger postsynaptic potentials lead to an enhanced activation of INs and MNs and thus an enhanced behavioral response (i.e., sensitization). Detailed mathematical models have been developed, which capture key aspects of the modulation of membrane channels and the dynamics of second-messenger systems, calcium regulation, and transmitter storage and release.

Long-Term Memory for Sensitization

Repeating the sensitizing stimuli over a 1.5 h period leads to the induction of long-term sensitization, the memory of which can persist for days to weeks. Repeated training leads to more prolonged phosphorylation and activation of nuclear regulatory proteins by PKA. Such proteins affect the regulatory regions of DNA and lead to the increased transcription of RNA and, hence, the increased synthesis of specific proteins. Some of the resulting proteins may be transcription factors that can activate other genes, some of which may be able to maintain their own activation. One of the newly synthesized proteins initiates the internalization and degradation of neuronal cell adhesion molecules (NCAMs), altering the interaction of the neuron with other cells and allowing the restructuring of the axon arbor. The sensory neuron could then form additional connections with the same postsynaptic target or make new connections with other cells. Other newly synthesized proteins such as intermediate filament proteins (IFP) are also likely to contribute to the new growth. Increased synthesis of calmodulin (CaM) also occurs, but the functional significance of this effect has not been determined.

Prolonged stimulation and increased cAMP also lead to the induction and subsequent expression of a gene encoding the protein ubiquitin C-terminal hydrolase (Uch). This neuron-specific enzyme enhances the degradation of certain proteins, including the regulatory subunits of PKA. With fewer regulatory subunits of PKA to bind to catalytic subunits, the catalytic subunits are persistently active and they may contribute to the long-term facilitation of transmitter release through the same cAMP-dependent processes seen in the short term. Some of these cAMP- and PKA-induced changes include a decrease in $I_{K,S}$ and enhanced excitability, as well as, perhaps, a prolongation or amplification of their effects on nuclear regulatory proteins. As with short-term sensitization, the enhanced release of transmitter from existing contacts of SNs on to MNs and INs underlies the long-term enhanced responses of the animal to test stimuli. However, unique to long-term sensitization, increases in axonal arborization and synaptic contacts may contribute to the enhanced activation of follower INs and MNs.

In addition to the cAMP–PKA cascade, another kinase cascade involving the mitogen-activated protein kinase (MAPK) has been identified as essential for long-term synaptic facilitation. Prolonged application of 5-HT results in the persistent phosphorylation (and subsequent activation) of MAPK. Activated MAPK translocates to the nucleus, where it may regulate gene transcription, possibly through the inhibition of cAMP response element binding protein (CREB) 2. Because under basal conditions CREB2 acts as a repressor of gene transcription, its inhibition leads to derepression and net gene expression in concert with CREB1 (see Figure 4). The involvement of additional transcription factors such as *Aplysia* activating factor (ApAF) and *Aplysia* LAPS18-like protein (ApLLP) in learning-induced gene expression is currently being investigated.

One of the genes that is expressed in response to long-term sensitization treatment encodes the *Aplysia* tolloid/bone morphogenetic protein-like (ApTBL) protein. ApTBL is a metalloprotease that may activate latent forms of transforming growth factor (TGF)-β. Similar to 5-HT, TGF-β also produces long-term increases in synaptic strength and excitability of the sensory neurons. Interestingly, TGF-β also activates MAPK and induces its translocation to the nucleus. Because TGF-β plays key roles in development, its involvement in sensitization may provide insights into the mechanistic relationship between development and learning. Another molecule that activates MAPK and is necessary for long-term synaptic facilitation is the peptide sensorin. Sensorin is secreted by SNs and regulates axonal growth and the formation of

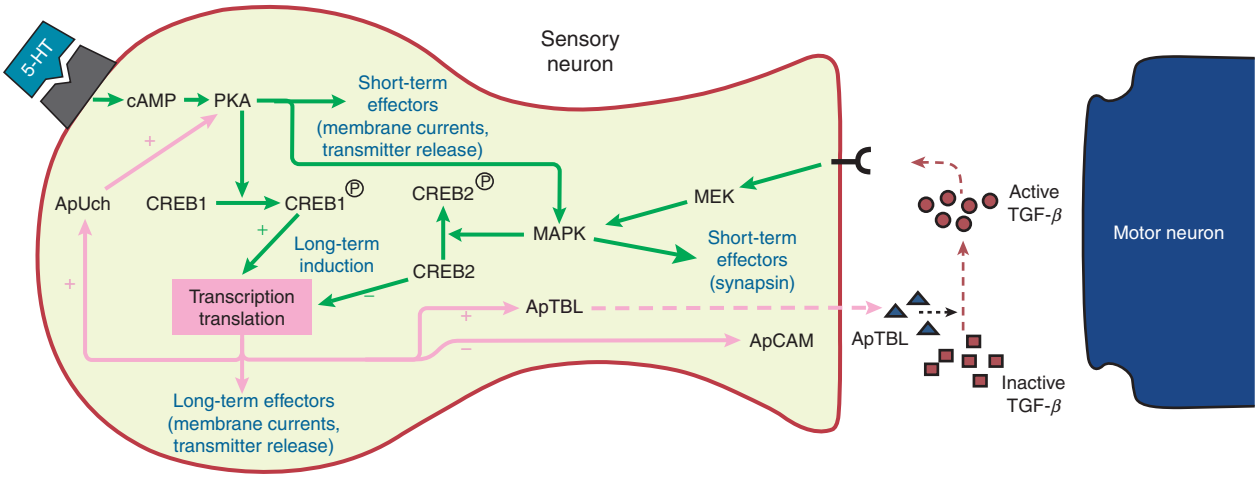

Figure 4 Simplified scheme of the mechanisms in sensory neurons that contribute to long-term sensitization and some aspects of short-term sensitization. Sensitization training leads to cAMP-dependent regulation of short-term effectors (see Figure 3(d)) and CREB1. 5-HT also activate MAPK, either through a cAMP-dependent, or through a seperate, cAMP-independent pathway. MAPK, in turn, regulates CREB2. Whereas CREB1 acts as an initiator of gene transcription, CREB2 acts as a repressor of gene transcription. The combined effects of activation of CREB1 and derepression of CREB2 lead to the regulation of the synthesis of at least ten proteins, only three of which (ApTBL, ApCAM, and ApUch) are shown. ApTBL is believed to activate latent forms of TGF-β, which can then bind to receptors on the sensory neuron. TGF-β activates MAPK, which can have both acute and long-term actions. One of its acute effects is the regulation of transmitter release. MAPK may also act by initiating a second round of gene regulation by affecting CREB2-dependent pathways. 5-HT, serotonin; ApUch, *Aplysia* ubiquitin C-terminal hydrolase; ApTBL, *Aplysia* tolloid/bone morphogenetic protein-like protein; ApCAM, *Aplysia* cell adhesion molecules; cAMP, cyclic adenosine-3′,5′-monophosphate; CREB, cAMP response element binding protein; MAPK, mitogen-activated protein kinase; MEK, mitogen-activated protein kinase kinase; PKA, protein kinase A; TGF, transforming growth factor.

new presynaptic varicosities associated with expression and maintenance of long-term memory.

Intermediate-Term Memory for Sensitization

It has recently been suggested that memory for sensitization comprises three physiologically and molecularly distinct phases: two of them, short- and long-term, have already been discussed; the third (and newly discovered) phase is intermediate-term memory (ITM), which lasts up to 3 h following sensitization training. Induction of ITM is protein synthesis-dependent but RNA synthesis-independent. In contrast, short-term memory requires neither protein nor RNA synthesis, whereas long-term memory requires both. Moreover, ITM expression requires persistent PKA activity.

An intermediate-term facilitation (ITF) of the sensorimotor synapse, which is produced by the application of five pulses of 5-HT (an analog of sensitization training), corresponds to the ITM as it displays similar temporal dynamics and requires protein synthesis but not RNA synthesis. This latter feature of ITF distinguishes it from short-term facilitation, which requires neither protein nor RNA synthesis, and long-term facilitation, which requires both. Depending on the induction protocol, ITF may require intermediate-term activation of PKA or PKC. These kinases can be activated for hours following prolonged treatment with 5-HT. Finally, the activation of previously silent release sites has also been implicated in ITF and could be a mechanism for memory consolidation.

Associative Learning in *Aplysia*

Classical Conditioning of Withdrawal Reflexes

The withdrawal reflexes of *Aplysia* can display classical and operant conditioning. Studies using a neuronal analog of differential classical conditioning indicate that a cellular mechanism called activity-dependent neuromodulation contributes to short-term classical conditioning observed at the behavioral level. Delivering a US, such as an electric shock to the tail or a peripheral nerve, activates a modulatory system, such as 5-HT, that nonspecifically enhances transmitter release from the SNs. This nonspecific enhancement contributes to short-term sensitization. The temporal specificity characteristic of associative learning results from the pairing of a CS (e.g., spike activity in one sensory neuron) with the US, which causes a selective amplification of the modulatory effects of the US in that specific SN. Unpaired activity does not amplify the effects of the US. The amplification of the modulatory effects in the paired SN leads to a pairing-specific enhancement of transmitter release from the SN. The proposed mechanism suggests that increased Ca^{2+} levels resulting from spike activity alter adenylyl cyclase via CaM and increase

the cAMP level produced by 5-HT. Thus, Ca^{2+} and CaM appear to play a role in the activity-dependent neuromodulation that may underlie associative conditioning of the tail and gill withdrawal reflexes.

Apart from activity-dependent presynaptic neuromodulation, another mechanism that has been implicated in classical conditioning is long-term potentiation (LTP). LTP of *Aplysia* sensorimotor synapses requires coincident presynaptic release of glutamate from SNs and postsynaptic depolarization of MNs. Glutamate binds to NMDA-type postsynaptic receptors, which leads to increased postsynaptic calcium concentration. Under resting conditions, the postsynaptic NMDA receptors are blocked by Mg^{2+}, and the binding of glutamate fails to activate them. However, when the postsynaptic MN is depolarized, the NMDA receptors are relieved of the Mg^{2+} block and respond to presynaptically released glutamate, thus elevating postsynaptic calcium. Recent results indicate that both LTP and activity-dependent neuromodulation are necessary for classical conditioning of withdrawal reflexes.

An important conclusion is that an aspect of a mechanism for short-term associative learning is an elaboration of the cAMP-dependent mechanisms that contribute to a simpler form of learning, sensitization. This finding raises the interesting possibility that even more complex forms of learning may be achieved by using these simpler forms as building blocks. Indeed, a mathematical description of the learning rule (activity-dependent neuromodulation) for simple classical conditioning, when incorporated into appropriate neural circuits, has the capability to simulate higher-order features of classical conditioning such as second-order conditioning and blocking, as well as features of operant conditioning.

Operant and Classical Conditioning of Feeding Behavior

Although this article focuses on the use of withdrawal reflexes of *Aplysia* as a model system to study learning and memory, the study of the feeding behavior of *Aplysia* has also recently revealed some of the mechanisms underlying operant and classical conditioning. In an operant conditioning paradigm, the delivery of a reinforcing stimulus is contingent on the expression of a designated behavior. In the case of feeding behavior, the operant behavior is ingestive or biting movements and the reinforcement is a stimulus to the esophageal nerve (En), which is enriched in dopamine processes and signals presence of food in the mouth (Figure 5). Stimulation of the En each time the animal performs a spontaneous biting movement (contingent reinforcement) results in an increase in the number of

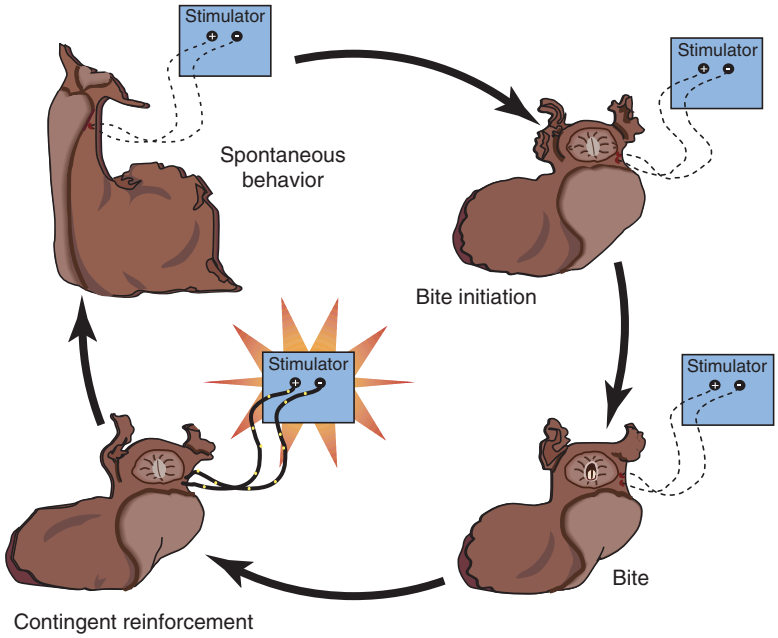

Figure 5 Operant conditioning of feeding behavior in *Aplysia*. While unrestrained in the experimental chamber, an animal moves about freely and engages in spontaneous behaviors (top left). Throughout the experiment, the animal is observed and all bites are recorded. A bite is a stereotypic movement, defined as opening of the jaws (top right) and protraction of the radula (a tongue-like organ; bottom right). Upon completion of a bite, electrical stimulation is delivered to the esophageal nerve (En; bottom left). The electrical activity induced by En stimulation resembles the pattern of activity induced by the presence of food in the esophagus. Therefore, this activity serves as a positive reinforcer of biting. Operant conditioning is manifested as an increase in the probability of spontaneous bites. Animals that receive En stimulation independently of biting (yoke control) do not display operant conditioning.

spontaneous bites during a test period. In classical conditioning of feeding behavior, touching the animal's lips with a brush can serve as the CS, which comes to elicit biting movements when it predicts stimulation of En (i.e., the US). Both classical and operant conditioning of feeding behavior appear to rely on the actions of the neuromodulator dopamine.

The development of behavioral protocols for operant and classical conditioning, as well as *in vitro* analogs, allowed for further investigations into the mechanisms underlying associative learning. Neural correlates of associative learning were investigated through the study of identified key cells in the buccal ganglia, where the central pattern generator that gives rise to the feeding behavior is located. One of these cells, B51, is a key element in the generation of ingestive-like buccal motor programs, which are patterns of activity generated by the central pattern generator (CPG) and that underlie biting movements. Investigation of the cellular properties of B51 in ganglia from operantly conditioned and control animals revealed that operant conditioning was correlated with significantly higher B51 input resistance but significantly lower burst threshold. These changes serve to increase the probability of B51 activation and would therefore facilitate the generation of the neural activity underlying biting movements in the trained animals. In contrast, classical conditioning was correlated with a significant increase in burst threshold, without a change in input resistance. The increase in burst threshold decreases the probability of B51 activation, thus placing its activity under the control of very specific input pathways, presumably the ones signaling the CS. These data suggest that B51 is an important locus of plasticity in both operant and classical conditioning of feeding behavior. They also suggest that intrinsic cellwide plasticity may be one important mechanism underlying operant conditioning, whereas synapse-specific plasticity may underlie classical conditioning. Interestingly, pairing spike activity with a puff of exogenous dopamine leads to the same biophysical changes in B51 which are produced by operant conditioning *in vivo*, indicating that spike activity in B51 may enhance the modulatory effects of dopamine. This mechanism is similar to the activity-dependent enhancement of the 5-HT effects that was proposed as a model of classical conditioning of withdrawal reflexes. Thus, activity-dependent neuromodulation may be a common mechanism for both operant and classical conditioning of various *Aplysia* behaviors.

See also: Conditioning: Theories; Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Operant Conditioning of Reflexes; Procedural Learning: Classical Conditioning.

Further Reading

Alberini CM (1999) Genes to remember. *Journal of Experimental Biology* 202: 2887–2891.

Antzoulatos EA and Byrne JH (2004) Learning insights transmitted by glutamate. *Trends in Neuroscience* 27: 555–560.

Bailey CH, Kandel ER, and Si K (2004) The persistence of long-term memory: A molecular approach to self-sustaining changes in learning-induced synaptic growth. *Neuron* 44(1): 49–57.

Baxter DA and Byrne JH (1993) Learning rules from neurobiology. In: Gardner D (ed.) *The Neurobiology of Neural Networks*, pp. 71–105. Cambridge, MA: MIT Press.

Brembs B, Lorenzetti FD, Reyes FD, Baxter DA, and Byrne JH (2002) Operant reward learning in *Aplysia*: Neuronal correlates and mechanisms. *Science* 296(5573): 1706–1709.

Byrne JH (1987) Cellular analysis of associative learning. *Physiological Reviews* 67: 329–439.

Byrne JH (ed.) (2003) *Learning and Memory,* 2nd edn. New York: Macmillan.

Byrne JH and Kandel ER (1996) Presynaptic facilitation revisited: State- and time-dependence. *Journal of Neuroscience* 16: 425–435.

Hawkins RD, Kandel ER, and Siegelbaum S (1993) Learning to modulate transmitter release: Themes and variations in synaptic plasticity. *Annual Review of Neuroscience* 16: 625–665.

Kandel ER (2001) The molecular biology of memory storage: A dialogue between genes and synapses. *Science* 294(5544): 1030–1038.

Kandel ER, Kupfermann I, and Iversen S (2000) Learning and memory. In: Kandel ER, Schwartz JH, and Jessel TM (eds.) *Principles of Neural Science,* 4th edn., pp. 1227–1245. New York: McGraw-Hill.

Purcell AL and Carew TJ (2003) Tyrosine kinases, synaptic plasticity and memory: Insights from vertebrates and invertebrates. *Trends in Neuroscience* 26: 625–630.

Reissner KJ, Shobe JL, and Carew TJ (2006) Molecular nodes in memory processing: Insights from *Aplysia*. *Cell and Molecular Life Sciences* 63: 963–974.

Sharma SK and Carew TJ (2004) The roles of MAPK cascades in synaptic plasticity and memory in *Aplysia*: Facilitatory effects and inhibitory constraints. *Learning & Memory* 11: 373–378.

Sutton MA and Carew TJ (2000) Parallel molecular pathways mediate expression of distinct forms of intermediate-term facilitation at tail sensory-motor synapses in *Aplysia*. *Neuron* 26(1): 219–231.

Learning and Memory in Invertebrates: *C. Elegans*

T A Timbers and C H Rankin, The University of British Columbia, Vancouver, BC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the 1960s Sydney Brenner and many colleagues shared the view that most of the 'interesting' problems of molecular biology had been, or were close to being, solved. It was their opinion that the nervous system was the next frontier of science, as it appeared to pose the most challenging questions. As a result, the search began for an organism that would allow for a global approach that focused on investigating the nervous system by means of genetic dissection. Brenner, after searching through many zoology texts, and examining a number of animals, settled on our organism of discussion, *Caenorhabditis elegans*. The worm, as it is often affectionately called, fit his search criteria for the ideal model organism: (1) it has a rapid lifecycle allowing for growth of large populations in a short period of time (*C. elegans* begins laying eggs at 3 days of age), (2) it has a simple reproductive cycle and genome (hermaphrodites can reproduce clonally, reducing the problem of differing genetic background between individuals), and (3) it is small in size and can be successfully frozen and thawed alive, making for easy storage in the lab.

Since the beginning of its development as a model organism a great wealth of knowledge has accumulated, including a fully sequenced genome, a complete wiring diagram of the nervous system, the developmental lineage for every cell present in the adult worm, and thousands of genetic mutants. The worm is amendable to *in vivo* whole-organism molecular analyses through the combination of microscopy and genetic-based fluorescence imaging, allowing for insight into protein localization and function. Past research has shown that to a great extent the cellular machinery and processes used by the worm are homologous to those found in the other organisms, including mammals. And of interest to the neuroscientist, *C. elegans* exhibits many basic types of behaviors which can be analyzed at behavioral, neural circuit, and genetic levels.

Although the tiny worm has the smallest nervous system of all multicellular learning and memory model systems, with only 302 neurons at its disposal, it shows a remarkable ability to learn and remember. It has been shown capable of both associative and nonassociative learning and has recently been developed as a model in which to study cellular mechanisms of learning and memory. This article discusses the behavioral paradigms used to study learning and memory in *C. elegans* as well as the proposed cellular and molecular mechanisms. The different paradigms are divided and discussed based upon their sensory modalities.

Mechanosensory Learning and Memory

The first type of learning described for *C. elegans* was habituation to a mechanosensory stimulus. In this learning paradigm the worm receives a mechanical stimulus, a tap, applied to the side of the petri plate within which it resides. The tap results in the worm changing from a forward swimming motion to a backward one for a brief period of time, after which it swims forward again. This reversal response is termed the 'tap withdrawal' response. This is indeed habituation to the mechanical stimulus and not sensory adaptation or fatigue, as the worm's response to tap can be dishabituated. After habituation training the application of a more intense stimulus, such as an electrical shock administered through the agar, causes dishabituation such that worms will respond to the subsequent presentation of the tap stimulus as if it were novel.

The neural circuit that mediates this response has been mapped and consists of mechanosensory neurons, command interneurons, and motor neurons (Figure 1). The pair of ALM cells and the AVM cells are the sensory neurons that sense head touch, while the pair of PLM cells are the sensory neurons responsible for sensing tail touch. The simultaneous stimulation of these neurons results in the reversal response seen following tap. These touch sensory neurons synapse onto the anterior ventral command interneurons (AVD, AVB, and AVA) and the posterior ventral command interneuron (PVC), which process information and project to appropriate motor neurons. The site of plasticity within the neural circuit for the response to tap is proposed to be at the glutamatergic synapses between the mechanosensory neurons and the command interneurons. Upon repeated administration of tap, worms habituate as measured by the decrease in their reversal lengths after each subsequent tap.

Short-Term Memory for Habituation

C. elegans can demonstrate both short-term memory and long-term memory (LTM) for habituation to tap. Short-term habituation is induced by administering 30 stimuli at a 10 or 60 s interstimulus interval (ISI), followed by testing for spontaneous recovery

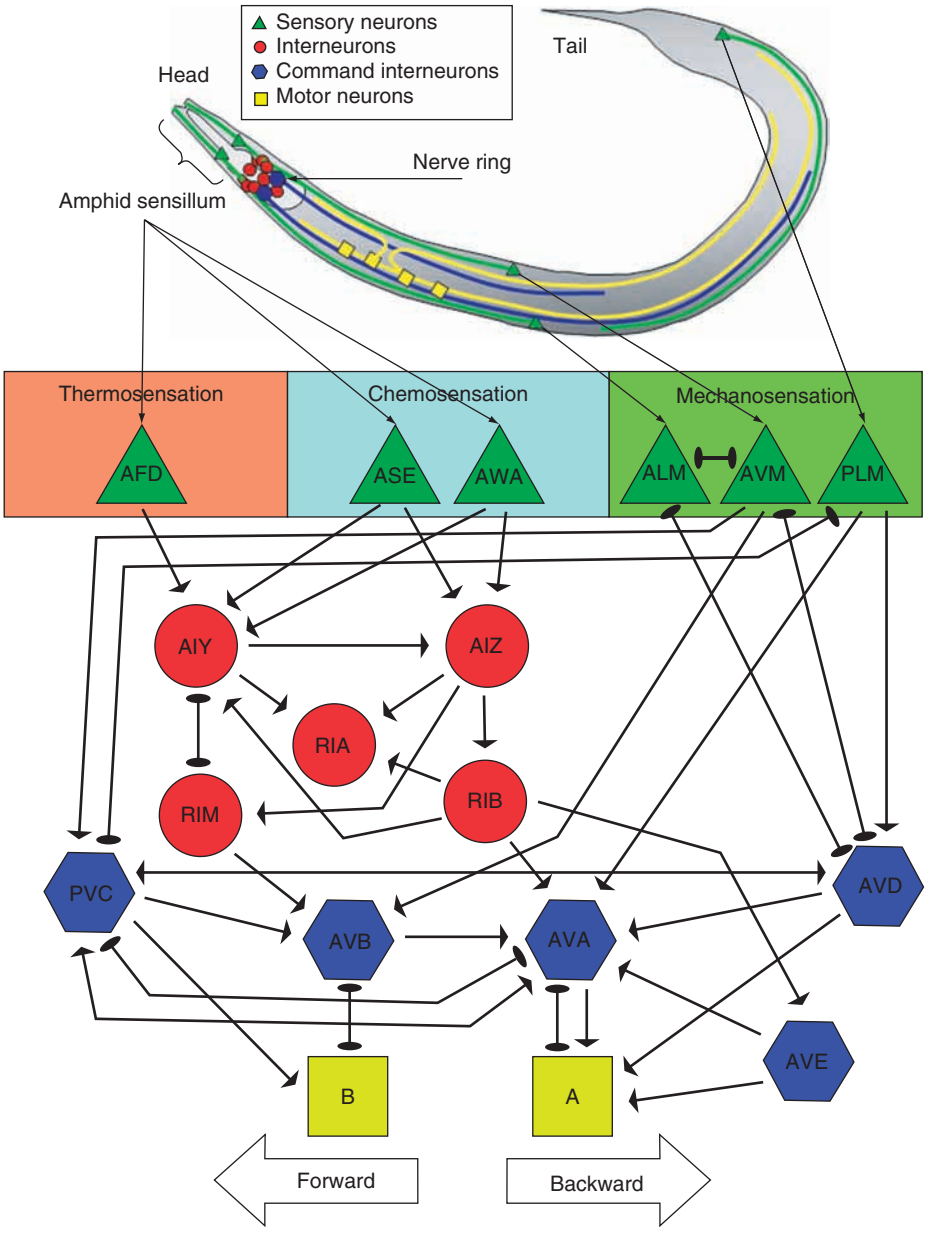

Figure 1 Schematic diagram of the main components of the neural circuits that mediate mechanosensory, chemosensory, and thermosensory behaviors that have been studied in learning and memory paradigms in *C. elegans*. Chemical synaptic connections are represented by lines with arrowheads, while electrical gap junctions are represented by lines terminating in ovals. From Giles AC, Rose JK, and Rankin CH et al. (2006) Investigations of learning and memory in *Caenorhabditis elegans*. Neurobiology of *C. elegans*. *International Review of Neurobiology* 69: 37–71.

at 30 s, 5 min, and 10 min after delivery of the last tap. Short-term habituation differs depending upon the ISI used (Figure 2). Short ISI produce deep levels of habituation that show rapid spontaneous recovery, whereas long ISI produce intermediate levels of habituation that show slower spontaneous recovery. These behaviorally different forms of short-term habituation are hypothesized to correlate with multiple molecular forms of short-term memory.

Genes involved in short-term habituation Thus far, four genes have been shown to be important for short-term habituation to tap: *eat-4*, encoding a homolog of the mammalian glutamate vesicular transporter, VGlut1; *cat-2*, encoding a homolog of the mammalian tyrosine hydroxylase; *dop-1*, encoding a homolog of the mammalian type 1 dopamine receptor; and *hab-1*, whose gene product has yet to be identified. The *eat-4* mutant worms show normal response to tap

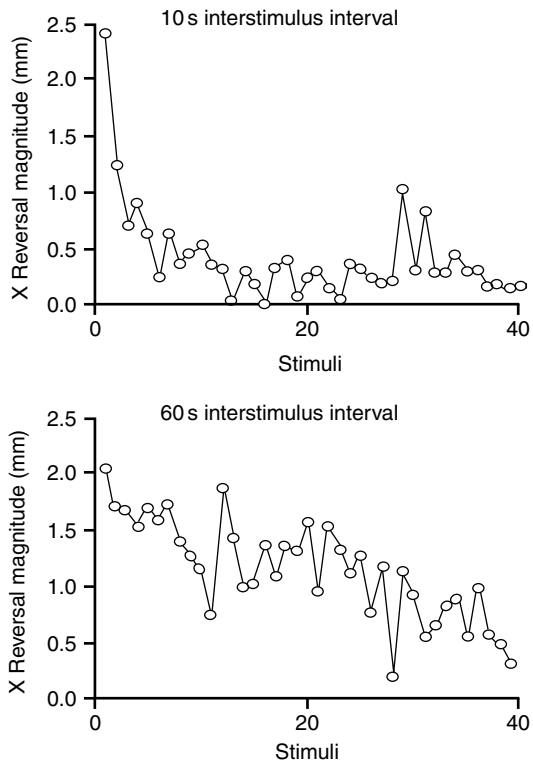

Figure 2 Habituation of mean response amplitude (reversal length in millimeters) to 40 tap stimuli delivered at a 10 s (top) and a 60 s (bottom) interstimulus interval (ISI). Habituation to stimuli presented at short ISS (10 s) produces greater levels of habituation and a more rapid decrement of the response than does habituation to stimuli presented at longer ISS (60 s).

but habituate in a more rapid manner and to a deeper level than do wild-type worms. They also show slower spontaneous recovery and interestingly do not show dishabituation. This suggests that presynaptic release of glutamate from the mechanosensory neurons is vital for wild-type short-term habituation. Both *cat-2* and *dop-1* mutant worms habituate more quickly and deeply than do wild-type worms when response is measured by the number of animals reversing to tap, but not when measured by reversal length. This may suggest that dopamine modulates the balance between the head touch circuit and the tail touch circuit and perhaps is not directly involved in the mechanism responsible for habituation. Worms with mutations in HAB-1 are reported to habituate more slowly than do wild-type worms but show normal response to tap and dishabituation. This mutation is interesting, as it appears to have the opposite phenotype from all the other known mutations affecting short-term habituation to date.

LTM for Habituation

LTM for habituation in *C. elegans* can last for as long as 48 h, is protein synthesis dependent, and is induced by spaced training. The protocol for spaced training in *C. elegans* that induces LTM consists of four blocks of 20 taps administered at a 60 s ISI with a 1 h rest period between each training block (Figure 3(a)). Worms that receive this spaced habituation training show significantly smaller mean reversal lengths compared to untrained control worms 24–48 h after the last training session. Massed training (the same number of stimuli as in spaced training, but no rest is given between blocks), on the other hand, does not induce LTM for habituation.

Genes involved in LTM for habituation Several genes involved in glutamate transmission that are found at the proposed site of plasticity have been investigated for their role in LTM in this system: *eat-4*, encoding a glutamate vesicular transporter homologous to mammalian VGlut1; *glr-1*, encoding a homolog to a mammalian kainate/α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type glutamate receptor subunit, and *nrm-1*, encoding a homolog to the mammalian *N*-methyl-D-aspartate (NMDA)-type glutamate receptor subunit. The *eat-4* and *glr-1* mutant worms are deficient for forming LTM of habituation, but, interestingly, worms do not appear to require NMDA-type glutamate receptors for memory formation, as *nmr-1* mutant worms show normal LTM (Figure 3(b)). Confocal imaging of transgenic worms expressing GLR-1 fused to a green fluorescent protein (GFP) and quantitative reverse transcriptase polymerase chain reaction (RT-PCR) have shed some light on the functional importance of GLR-1. When measured 24 h after spaced training, GLR-1 expression levels are observed to decrease; suggesting that LTM for habituation is molecularly mediated by a down-regulation of GLR-1 expression (Figure 3(c)).

Context Associations with Habituation to Tap

C. elegans is also capable of forming chemosensory context associations with habituation to tap. When worms plated on agar treated with Na^+ (taste worms are attracted to) are habituated with 30 taps at a 10 or a 60 s ISI, then are transferred onto plain agar for a 1 h rest period and, finally, rehabituated on Na^+-treated agar, they show greater retention of the earlier training (demonstrated by smaller reversal lengths in the second round of habituation) than do worms that were trained and tested on plates without the Na^+ taste. This form of associative conditioning has been shown sensitive to extinction (you do not see this effect if you leave the worms in the same context that they are trained and tested in for the 1 h rest period) and latent inhibition (preexposure to the context for 1 h before training diminishes the context effect).

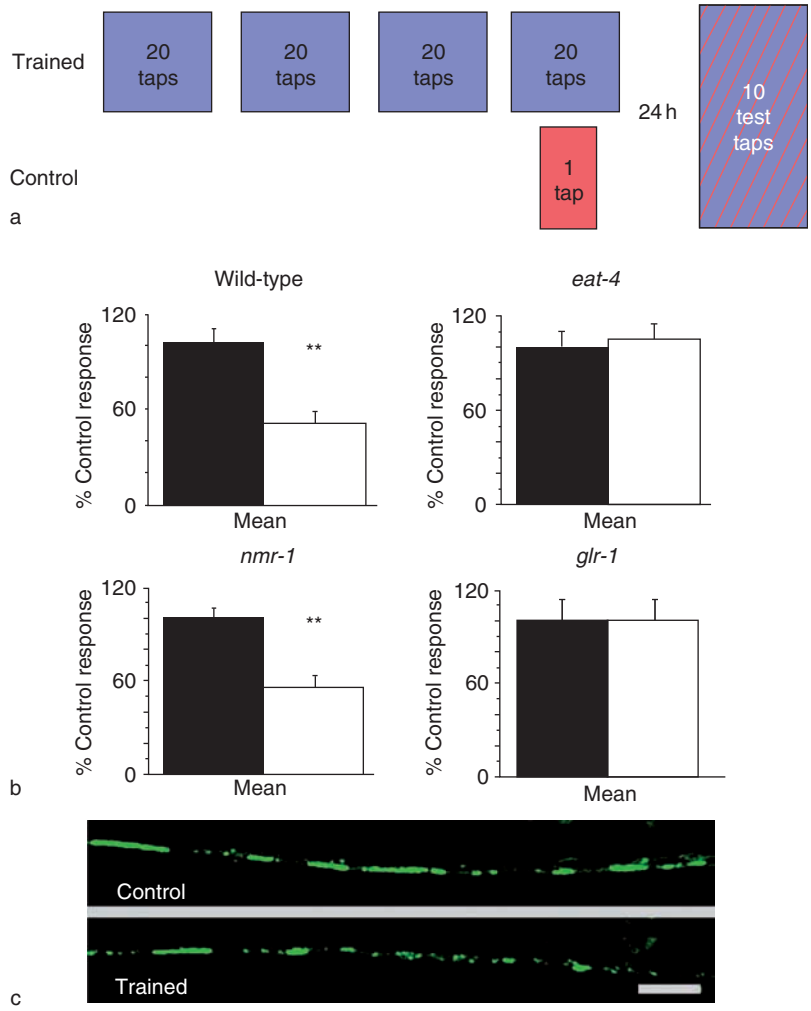

Figure 3 Long-term memory (LTM) paradigm and corresponding genetic analysis. (a) Protocol for spaced training in *C. elegans* that induces LTM. Trained worms receive four blocks of 20 taps administered at a 60 s ISI with a 1 h rest period between each training block, while control worms receive one tap. After 24 h, ten taps are administered to both control and trained worms to test for memory. Memory is seen when trained worms show significantly smaller responses than do control worms. (b) Mean reversal lengths of wild-type worms and of *eat-4, nmr-1*, and *glr-1* mutants during the ten test taps, 24 h after training, measured as a percentage of control reversal lengths for trained and control worms. Wild-type and *nmr-1* mutants show normal LTM, whereas *eat-4* and *glr-1* do not. (c) Fluorescent imaging of a GLR-1–GFP fusion protein 24 h after LTM training in trained and untrained control worms. GLR-1 expression levels are observed to decrease for trained worms. Scale bar = 20 μm.

Chemosensory Learning and Memory

Worms are attracted and migrate toward cations, anions, cyclic nucleotides, and volatile chemicals; this behavior is termed chemotaxis (Figure 4). Chemotaxis is usually measured through either one of two common assays: the first is orientation, quantified by tracking the path of individual worms as they migrate toward or away from a chemical, and the second is accumulation, quantified by counting the number of worms that collect at the location of the chemical. This behavior has been found subject to several forms of learning in *C. elegans*: habituation to low levels of odorants, state-dependent adaptation to high levels of odorants, classical conditioning, and aversive learning.

The neural circuit that mediates chemotaxis has been mapped (Figure 1) and consists of many more sensory neurons than do the other behaviors discussed in this article. In total, 11 pairs of sensory neurons in the head region of the worm are responsible for chemotaxis. Eight pairs have sensory endings that are in contact with the external environment and for the most part detect water-soluble attractants. The additional three pairs are enclosed in the amphid sheath and their main role is to detect volatile chemicals. Because these 11 pairs of neurons respond to such a large number of different chemical substances, it is thought that single chemosensory neurons are capable of detecting multiple chemicals, and, further, this

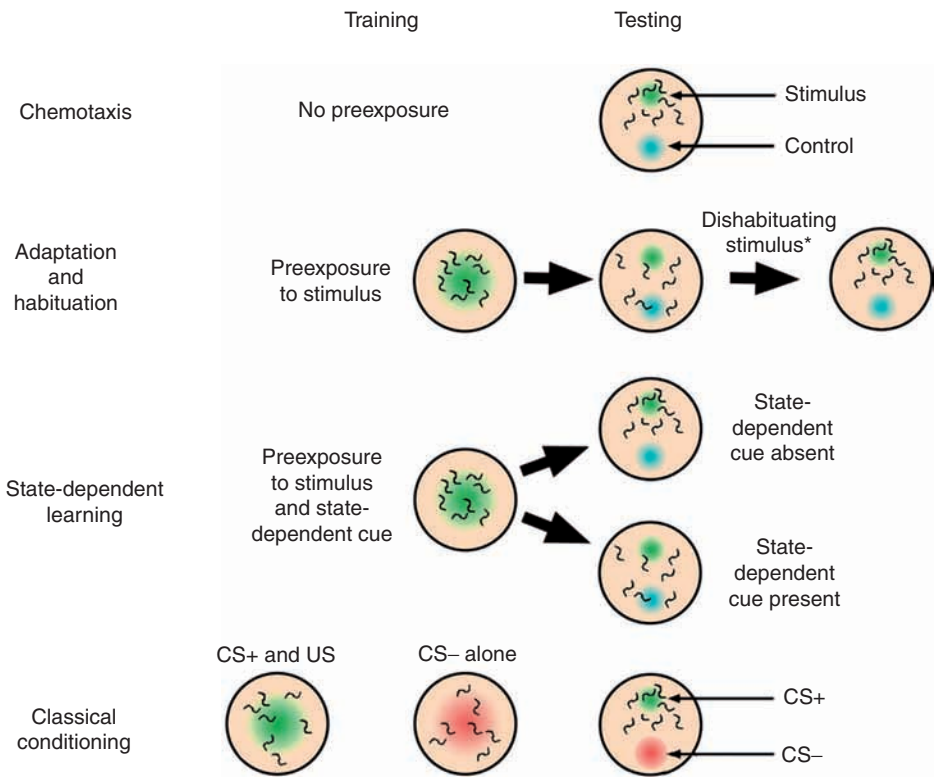

Figure 4 Schematic diagram of chemotaxis and the adaptation, habituation, state-dependent learning, and classical conditioning paradigms. In chemotaxis, worms migrate toward an attractive odor (green) such as diacetyl, rather than to a control stimulus (blue) such as water. Adaptation and habituation are produced by preexposing worms to an attractive chemical stimulus for an extended period of time. When worms are later tested for chemotaxis they no longer show a preference for that chemical over a control chemical. Adaptation is induced by exposing worms to high concentrations of the stimulus, whereas habituation to Na^+ or diacetyl is induced by exposure low concentrations. Once worms are habituated to Na^+ or diacetyl they can be dishabituated* by exposing them briefly to a high concentration of that chemical; this does not occur for adaptation. In the paradigm for state-dependent learning starved or inebriated worms are preexposed to an odor and learn to associate this odor with their physiological state. Thus, when worms are later exposed to that odor they only show adaptation when they are in the same physiological state as when they were preexposed. In the differential classical conditioning paradigm worms are first preexposed to an odor (conditioned stimulus, CS+; green) in the presence of food (unconditioned stimulus, US) and then preexposed to another odor without food (CS−; red). When worms are later placed in the presence of both odors they show a preference for the odor that was previously paired with food (CS+).

suggests that one site of plasticity for chemotaxis behavior may be intraneuronal.

Adaptation

Adaptation, a decrease in responding after lengthy exposure to a stimulus (i.e., no longer noticing an odor after a long exposure to it), itself is not considered learning; however, it can have associations formed so that it plays a role in associative learning in experiments of state dependency. In the adaptation paradigm worms are exposed to a high concentration of a particular odorant for a determined period, called the adaptation period. After the worms are moved to a chemotaxis assay plate (test odorant on one side and a control substance on the other) (Figure 4), the chemotaxis index ((number migrated to test − number migrated to control)/total number) is measured.

State-dependent learning *C. elegans* can show state-dependent learning – that is to say, they can make associations between their physiological state and stimuli in their environment (Figure 4). Worms can associate a feeding state or an inebriated state, induced by ethanol, with adaptation to an odor. When starved worms are preexposed to odor, in this case benzaldehyde, they show greater adaptation to that odor than do unstarved control worms. When worms are preexposed to the odor in the presence of an excess of food, they show less adaptation than do control worms. The neurotransmitter serotonin, which is known to be important in other feeding-state-dependent behaviors, has been shown to mediate this feeding-state-dependent association. Similar to the feeding-state experiments, inebriated worms preexposed to an odor only show adaptation to that odor

when they are tested in the presence of inebriating doses of ethanol.

Habituation

Habituation to some chemosensory stimuli can also be observed in *C. elegans*. Preexposure to weak concentrations of either Na^+ or diacetyl causes the worms to habituate to that chemical; this can be observed as a decreased percentage of worms migrating in the direction of the chemical stimulus. This habituation can be differentiated from adaptation by the application of a dishabituating stimulus, a short exposure to a higher concentration of the chemical. This brief exposure immediately reverses the behavior such that an increase in the percentage of worms migrating in the direction of the weak concentration of the chemical stimulus is seen (Figure 4). This does not occur for adaptation. Whether worms habituate to either Na^+ or diacetyl or adapt to these chemical stimuli is determined by the concentration of the chemical. Habituation is seen at low concentrations of these chemicals, whereas adaptation is seen at higher concentrations.

Classical Conditioning

Differential classical conditioning, a form of associative learning, has been shown to exist in *C. elegans*. The first paradigm used to investigate this type of learning consisted of preexposing worms to two odors and pairing one odor with (conditioned stimulus; CS+) and the other odor without (CS−) food. After training, worms placed on a test plate that contain both the CS+ and CS− odors significantly preferred the odor which had been paired with food (Figure 4). Differential classical conditioning can also be demonstrated by pairing an aversive stimulus, such as garlic, with the CS+ odor. In this paradigm worms significantly prefer the CS− odor that was not paired with the aversive stimulus. Since the development of these paradigms, several others have been intoduced. These include pairing odors with feeding states and tastes. Thus, these studies illustrate that the small worm with its tiny nervous system is capable of complex chemosensory learning.

Genes involved in classical conditioning Several genes identified by means of genetic screens and behavioral investigations are important for differential classical conditioning. Two as yet unidentified genes, *lrn-1* and *lrn-2*, show normal chemotaxis and chemotactic habituation but are deficient for conditioning. As discussed earlier, the glutamate receptor subunit, GLR-1, is important for LTM for habituation to mechanosensory stimulus; the gene for this receptor subunit has also been shown to be an important gene for chemosensory learning: *glr-1* mutants show deficiencies for both chemotactic habituation and differential classical conditioning. Last, mutations in the gene *hen-1*, whose gene product is a low-density lipoprotein (LDL) receptor-like secretory protein, lead to abnormal sensory integration of chemosensory signals and no observed learning.

Aversive Learning

One of the most recent forms of learning discovered in *C. elegans* is the capacity to modify its chemosensory preferences after being exposed to pathogenic bacteria. This means that *C. elegans* can learn to avoid odors or tastes associated with familiar pathogenic bacteria and in turn increase its preference for odors or tastes associated with familiar nonpathogenic bacteria. In this paradigm, worms were placed on pathogenic bacteria for 4 h and subsequently for 4 h on nonpathogenic bacteria. This was followed by a choice test whereby half of the agar plate contained the pathogenic bacteria and the other half contained the nonpathogenic bacteria. Worms were observed to avoid the pathogenic bacteria and were attracted to the nonpathogenic bacteria. This study also identified serotonin and the gene *mod-1*, the gene product of which is a serotonin-gated chloride channel, as important components to this behavior.

Thermotactic Learning and Memory

Behavioral plasticity has also been shown in *C. elegans* for thermotactic behavior. Thermotaxis is defined as movement toward or away from a thermal stimulus. Worms can learn to associate a particular temperature with the presence or absence of food. This can be measured by placing the worms on a temperature gradient; if the worms have recently been well fed in an environment at a constant temperature they will migrate to this temperature when placed on the gradient (Figure 5). Conversely, if the worms have recently experienced starvation at a constant temperature they will avoid this temperature when placed on the gradient. This behavior is termed isothermal tracking.

The circuit that mediates thermotaxis has been mapped and consists of one pair of anterior sensory neurons (AFD) in the head region that project to several anterior interneurons (AIY, AIZ, and RIA) which compute the information (Figure 1). Through laser ablation experiments it has been demonstrated that AIY may signal for thermotaxis to warmer temperatures, while AIZ may signal for thermotaxis to cooler temperatures. These two interneurons converge onto RIA, which is proposed to be a site of thermotactic integration.

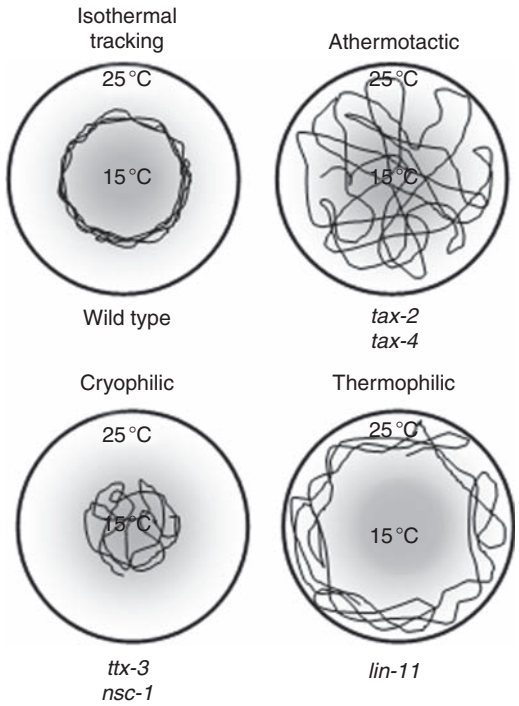

Figure 5 Schematic representation showing wild-type thermotactic behavior (isothermal tracking) and the various mutant phenotypes observed when mutations affect thermotactic behavior. Athermotactic phenotypes show no preference for any temperature. Cryophilic phenotypes show a preference for cooler temperatures than the temperature that was previously paired with food. Thermophilic phenotypes show a preference for warmer temperatures than the temperature that was previously paired with food. Adapted from Giles AC, Rose JK, and Rankin CH (2006) Investigations of learning and memory in *Caenorhabditis elegans.* Neurobiology of *C. elegans*. *International Review of Neurobiology* 69: 37–71.

The site of plasticity within the neural circuit for thermotaxis is proposed to lie within the AFD sensory neurons and the AIY interneuron.

Parsing Thermotactic Learning into Its Components

After worms learn the association of a temperature with a presence or absence of food, this memory is not resistant to future modifications, as the associated temperature can be 'reset.' Worms raised at a particular cultivation temperature with food show a preference for that temperature when placed on a temperature gradient, as stated earlier, but this association with food can be transferred to another temperature if the cultivation temperature is shifted to the new temperature in the presence of food.

This isothermal tracking can be parsed into two components: the response to a temperature shift and the change in behavior due to the presence of food. This is accomplished by observing how long it takes worms to learn a new temperature or feeding state after having already committed this type of association to memory. When worms are switched from their cultivation temperature to a new temperature, it takes them 2–3 h to learn this new association, and this learning is temperature independent. When worms are transferred to a different feeding state under the same temperature, they learn this new association in a temperature-dependent fashion. When they are transferred from a nutrient-rich to a nutrient-poor environment at cooler temperatures (17 °C), they are capable of learning the new association after 3 h, whereas at warmer temperatures (25 °C) it takes them only 30 min to make this new association. When they are transferred from a nutrient-poor to a nutrient-rich environment, they are able to learn this association in under 10 min. Similar results are seen when a double switch of temperature and feeding state are performed. This suggests that the environmental temperature is encoded through a relatively long process, while a relatively short process encodes the pairing of the feeding state to the stored temperature.

The roles that different neurotransmitters play in thermotaxis have been investigated to further dissect the two components of this behavior. The neurotransmitter serotonin is proposed to mimic a well-fed state in *C. elegans*, and an exogenous application of this neurotransmitter to starved worms during conditioning to a temperature mimics the effect of a food-rich environment, meaning that when they are placed on a temperature gradient they thermotax to the temperature at which serotonin was applied. The neurotransmitter octopamine is thought to mimic a starved state in *C. elegans*, and when exogenous octopamine is applied to well-fed worms during cultivation, worms later avoid the cultivation temperature when placed on a temperature gradient. This evidence supports the hypothesis that these neurotransmitters play a role in the associative learning for thermotaxis.

Genes Involved in Thermotactic Learning and Memory

Investigations into this behavioral plasticity have begun to identify some of the genes involved in thermotactic behaviors. The original mutant screens developed to identify genes important for thermotaxis grouped positive mutants into one of three phenotypic categories: thermophilic, worms which migrate to temperatures that are higher than their cultivation temperatures; cryophilic, worms which migrate to temperatures that are lower than their cultivation temperatures; and athermotactic, worms which show no preference for any temperature. This screen identified a plethora of mutants with thermotactic deficits,

such as *tax-2, tax-4, ttx-3*, and *lin-11*, which have mutations in genes that encode cyclic nucleotide-gated channels and transcription factors (Figure 5).

To investigate mutants defective for forming associations between temperature and food a genetic screen was done that looked for worms abnormal for updating their isothermal tracking. From this, three mutants, *aho-1, aho-2,* and *aho-3,* were identified. These mutants show normal thermotaxis when they are cultivated in a well-fed state, but they cannot learn to avoid the cultivation temperature if they are conditioned in a starved state. The deficit exhibited by these mutants implies that these mutations interfere with associating the encoded temperature with the feeding state, but does not interfere with the encoding of the temperature itself. Thus, on top of identifying genes involved in this learning paradigm, this screen also illustrates that the two aspects of thermotaxis exist (the thermal memory and the associative learning).

Several other genes implicated in thermotaxis have been also been identified. One such gene encodes neuronal calcium sensor-1 (NCS-1); *ncs-1* mutants perform cryophilically on the thermotactic learning assay. Selective rescue of NCS-1 in the AIY interneuron led to a full behavioral recovery. Rescue of NCS-1 in AIY with a transgene containing a point mutation in the calcium-binding region did not result in a behavioral recovery, suggesting that intracellular calcium signaling in the AIY interneuron may be important for thermotactic learning. Lastly, *hen-1* (shown to play a role in chemosensory learning) is also involved in thermotaxis learning. Mutants deficient for this gene will migrate to their cultivation temperature independent of the feeding state with which it was previously paired. The involvement of HEN-1 in both chemosensory and thermotactic learning provides support for a role of this protein in general sensory integration.

Conclusion

Investigations into behavioral plasticity in *C. elegans* have demonstrated that this simple worm is capable of many forms of learning and memory in a variety of sensory modalities. Simple forms of learning include chemosensory and mechanosensory habituation and dishabituation. The complex learning behaviors exhibited by the worm include classical conditioning, context conditioning, state-dependent learning, and aversive learning. Several important genes for these behaviors have already been identified and investigations into identifying other genes continue, as do studies attempting to link the molecular mechanisms responsible for learning and memory.

This rich learning and memory repertoire exhibited by *C. elegans* along with its mapped nervous system and fully sequenced genome make it an ideal invertebrate model organism to study learning and memory. Other advantages to using *C. elegans* for this are the plethora of mutants available for analysis, the increasing number of fluorescent fusion proteins for studying protein localization and function, and homology between worms and higher organisms. These all suggest that important findings on learning and memory of general interest are possible and very likely using the worm as a model system.

See also: Conditioning: Theories; Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Operant Conditioning of Reflexes; Procedural Learning: Classical Conditioning.

Further Reading

De Bono M and Maricq AV (2005) Neuronal substrates of complex behaviors in *C. elegans*. *Annual Review of Neuroscience* 28: 451–501.

Giles AC, Rose JK, and Rankin CH (2006) Investigations of learning and memory in *Caenorhabditis elegans*. Neurobiology of *C. elegans*. *International Review of Neurobiology* 69: 37–71.

Hobart O (2003) Behavioral plasticity in *C. elegans*: Paradigms, circuits, genes. *Journal of Neurobiology* 54: 203–223.

Mohri A, Kodama E, Kimura KD, et al. (2005) Genetic control of temperature preference in the nematode *Caenorhabditis elegans*. *Genetics* 169: 1437–1450.

Mori I (1999) Genetics of chemotaxis and thermotaxis in the nematode *Caenorhabditis elegans*. *Annual Review of Genetics* 33: 399–422.

Rankin CH (2002) From gene to identified neuron to behavior in *Caenorhabditis elegans*. *Nature Reviews Genetics* 3: 622–630.

Rankin CH, Beck DO, and Chiba CM (1990) *Caenorhabditis elegans:* A new model system for the study of learning and memory. *Behavioural Brain Research* 37: 89–92.

Riddle DL, Blumenthal T, Meyer BJ, et al. (1997) *C. elegans*. II. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Rose JK and Rankin CH (2001) Analyses of habituation in *Caenorhabditis elegans*. *Learning & Memory* 8: 63–69.

Steidl S, Rose JK, and Rankin CH (2003) Stages of memory in the nematode *Caenorhabditis elegans*. *Behavioral and Cognitive Neurosciences Review* 2: 3–14.

Wood WB (1988) *The Nematode Caenorhabditis elegans*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Zhang Y, Lu H, and Bargmann CI (2005) Pathogenic bacteria induce aversive olfactory learning in *Caenorhabditis elegans*. *Nature* 438: 179–184.

Relevant Websites

http://elegans.swmed.edu – *Caenorhabditis elegans* World Wide Web server (University of Texas, Southwestern Medical Center).

http://www.wormbase.org – WormBase (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).

Learning and Memory in Invertebrates: *Drosophila*

D-B G Akalal and R L Davis, Baylor College of Medicine, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Much progress has been made in the understanding of learning and memory from the neurobiological investigations in invertebrates. The fruit fly *Drosophila melanogaster*, in particular, offers numerous advantages compared to other model systems for studying learning and memory. Fruit flies are small and relatively inexpensive to culture. These aspects offer the capability of performing behavioral experiments on hundreds or thousands of subjects. The fly has a robust repertoire of behaviors that can be studied. For instances, flies can learn about olfactory, visual, taste, and tactile cues. In addition, the cellular and molecular events underlying learning can be easily related to observable and quantifiable learned behaviors using mutants and transgenic animals. Further genomic studies are now facilitated with knowledge of the complete DNA sequence of the organism. Moreover, recently developed optical imaging techniques have offered a way to peer into the brain of the animal as it learns and to observe changes in the neural activity in populations of neurons due to learning. Finally, to date, studies on genes involved in learning have revealed a remarkable structural and functional conservation of these genes with mammalian counterparts in human and mouse, offering confidence that the molecular principles underlying learning as deduced from the fly are of general importance. In this article, we describe some of the important *Drosophila* learning assays that have been developed and what they have revealed, in conjunction with molecular, cellular, and neuroanatomical tools, about the molecular and cellular biology of learning.

Olfactory Conditioning

Aversive Olfactory Conditioning

Mild electric shock has been generally employed as the aversive stimulus for two general types of aversive olfactory conditioning: operant olfactory conditioning and classical conditioning. The earliest studies employed operant olfactory conditioning in which flies learned to avoid walking into a tube containing a certain odor, after a training session that paired an odor with mild electrical shock as a negative reinforcer. Flies were allowed to run alternatively into one of two plastic tubes that contained different odorants and were shocked in only one of the two tubes and in the presence of only one of the two odors (Figure 1). After training, the flies' memory of the odor with shock versus the control odor was measured by observing their avoidance of walking into fresh tubes containing the same two odors but without shock. This measurement provided an index of selective avoidance, named the performance index, after training to an odor linked with electric shock, and therefore represented an index of olfactory learning. Since the flies were required to move into the chamber in order to experience the odor and shock, this type of conditioning has been classified as operant.

Classical olfactory conditioning is similar but utilizes an odor carried in an airstream moving past the flies, which are contained in a chamber whose floor is composed of an electrifiable copper grid (Figure 1). Using this setup, the odor and shock cues are delivered to the flies rather than having the flies move to the cues. The flies are then transferred to a T-maze where they make a binary choice between the odor associated with electric shock and a control odor. The performance index after one cycle of classical conditioning is much higher than that obtained with olfactory operant conditioning, for reasons that remain unclear. The memory of olfactory classical conditioning after a single cycle of training is generally about a day, but this can be extended to several days by conditioning flies with 5–10 cycles, one right after each other (massed training). Memory is even more prolonged if a rest of 15 min is allowed between each cycle (spaced training).

Numerous single gene mutants have been isolated that have impaired performance after olfactory operant and/or olfactory classical and aversive conditioning. Knowledge of the products of these genes has contributed to three concepts for the physiological basis for olfactory memory formation. First, seven different gene products are elements of the cyclic adenosine monophosphate (cAMP) signaling pathway. The *dunce* gene encodes cAMP phosphodiesterase; *rutabaga*, an adenylyl cyclase; *neurofibromin*, an activator of adenylyl cyclase; *DCO*, the catalytic subunit of protein kinase A; *PKA-R1*, a repressor of protein kinase A; *amnesiac*, a putative neuropeptide activator of adenylyl cyclase similar in sequence to mammalian pituitary adenylyl cyclase-activating peptide (PACAP); and *dCreb*, a transcription factor known as the cyclic AMP response element binding protein that is responsive to alterations in intracellular cAMP levels. Mutants in the first five genes impair acquisition or the formation of short-term memory. The *amnesiac*

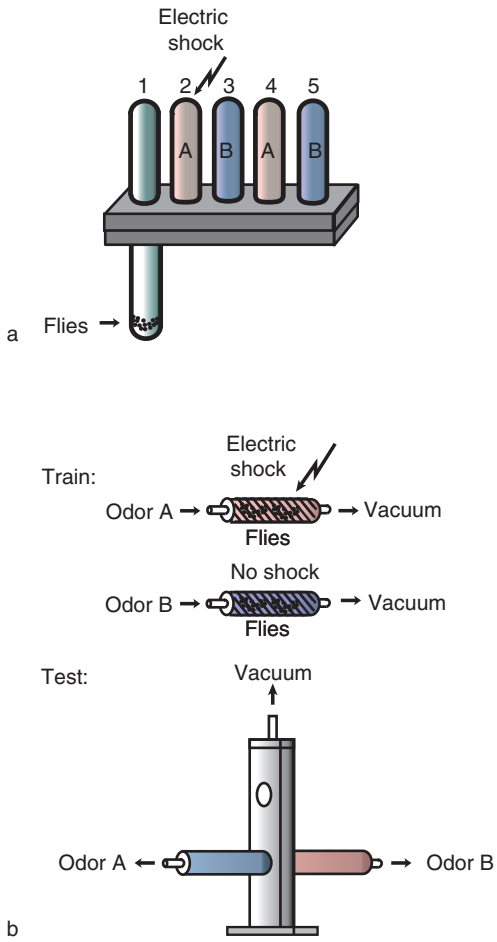

Figure 1 Aversive olfactory learning in *Drosophila*. (a) Schematic diagram of the apparatus used for operant olfactory learning. Flies are placed in a start tube and walk up the tube due to negative gravitaxis and positive phototaxis. The five tubes on the top of the apparatus can be slid horizontally so that any one of them is in register with the start tube. Flies are first allowed to walk into the tube containing odor A and electric shock for a period of 30 s. They are then tapped back into the start tube and allowed to walk into the tube containing odor B for 30 s. The animals are then tested by allowing them to walk into a fresh tube containing odor A without shock (tube 4), and the fraction of flies that avoid the odor is recorded. The flies are then tested with odor B in the last tube. Selective avoidance, after training, of the tube containing odor A but not the tube containing odor B is taken as an index of performance. (b) A schematic diagram of the apparatuses used for olfactory aversive classical conditioning. Flies are placed in a tube and odors are drawn through the tube in an air current. Electric shock is delivered when only one of the two odors is passed through the tube. After training, the animals are tested in a T-maze in which they must choose to avoid odor A or odor B, the odors being carried to the choice point in air currents. Modified with permission from John Wiley & Sons, Inc: Roman G and Davis RL (2001) Olfactory learning in *Drosophila*. *BioEssays* 23(7): 571–581. Copyright 2001 by John Wiley and Sons, Inc.

gene is thought to be involved in medium-term memory, and *dCreb*, in long-term memory. The involvement of these gene products provides compelling evidence that olfactory learning occurs, in part, through signaling by the cAMP system, although how specifically this signaling alters the physiology of the neurons involved in learning remains unknown.

The second concept for a physiological basis for olfactory memory formation is that other genes involved in the formation of short-term olfactory learning include *Volado*, which encodes a cell adhesion receptor of the integrin family, and *fasII*, which encodes a cell adhesion receptor of the immunoglobulin superfamily. These genes indicate that cell adhesion and/or signaling through cell adhesion receptors, perhaps at the synapse, are critical for acquisition or the formation of short-term memory. This observation, in retrospect, is not surprising since cell adhesion receptors represent some of the elements needed for proper communication between neurons, but nevertheless, *Drosophila* genetic studies have revealed the specific identify of the molecular players.

Third, the products of several genes required for normal long-term memory are involved in directed transport of ribonucleoproteins and local translation. The *pumilio* gene product participates in microtubule-dependent, subcellular targeting and localized mRNA translation. Protein products of *oskar*, *staufen*, and *eIF-5C* are involved in the translocation of ribonucleoprotein particles to distinct cellular compartments and their local translation. Thus, long-term olfactory memory in *Drosophila* may be served by local translation near synapses using the functions of this set of gene products.

Where in the *Drosophila* brain are olfactory memories formed and stored? The majority of the evidence based on the expression patterns of genes involved in olfactory learning, ablation experiments, neuron-specific disruption experiments, and functional imaging indicates that the mushroom body neurons are principal sites for olfactory memory formation. Blocking the synaptic output of these neurons blocks retrieval, suggesting that the functional changes underlying learning occur upstream of the output synapses of mushroom body neurons. Mutants that impair the formation of one axon branch of the mushroom body neurons, known as the vertical branch, impair long-term but not short-term memory. Thus, long-term memory may form within a specific axon branch of these neurons, or the output of these neurons, to guide behavior from long-term memory, is through this branch. However, other evidence indicates that there are at least two other areas of the brain in which olfactory memories form. Functional imaging studies have revealed that a change, or memory trace, in synaptic transmission after training occurs in the antennal lobe, although this memory lasts for only a few minutes after conditioning. In addition, functional imaging and mutant studies suggest that

the dorsal posterior medial neurons may participate in medium-term memory. Importantly, both the antennal lobe and the dorsal posterior medial neurons innervate the mushroom bodies, consistent with the possibility that all memory retrieval must pass through the mushroom body neurons.

An important recent advance that highlights the utility of *Drosophila* for studying the neurobiology of learning and memory is the development of techniques to control the expression of transgenes in the dimensions of both time and space. It is now possible to express a normal transgene in selected neurons of the nervous system and turn the expression on or off with temperature or by feeding flies the antiprogestin, RU486. This provides the investigator the ability to test whether expression of a wild-type transgene in specific neurons, and only at certain developmental times, is sufficient to rescue the learning impairment associated with mutants. These technologies have been used to show that a wild-type *rutabaga* transgene rescues the learning impairment of *rutabaga* mutants if the transgene is expressed in the adult mushroom bodies. These important experiments prove that the normal *rutabaga* gene functions in a physiological sense in the adult, rather than being involved in some aspect of brain development that impairs learning.

Appetitive Olfactory Conditioning

Drosophila can also learn about olfactory stimuli through their associations with appetitive cues. Like in operant olfactory aversive conditioning, *Drosophila* can be trained by allowing them to run alternatively into one of two plastic tubes that contain different odorants, with sucrose solution in only one of the two tubes, and in the presence of only one of the two odors. After training, the flies' attraction to the odor with sucrose versus the control odor is measured in fresh tubes containing the same two odors but without shock. Classical appetitive olfactory conditioning has also been developed in which the animals receive the olfactory cue just before or simultaneously with sucrose. The latter paradigm has been used to investigate neuromodulatory influences required for aversive versus appetitive conditioning. From these studies, the general conclusion has been formed that dopaminergic innervation of the mushroom bodies is required for aversive conditioning and octopaminergic activity is required for appetitive conditioning.

Visual Learning in the Flight Simulator

Visual conditioning in *Drosophila* has been extensively studied using an experimental setup known as a flight simulator (Figure 2). In this paradigm, a single fly is tethered in mid-air by a wire that is glued to its thorax and/or head. The wire feeds into a torque meter and a negative feedback is established between the fly's yaw torque and the angular velocity of a cylindrical visual panorama that surrounds the animal. As the fly attempts to turn left or right in flight, the negative feedback is translated into rotation of the surrounding visual panorama so that the fly detects environmental movement that simulates appropriate directional changes. A common type of experiment in the flight simulator involves T-shaped patterns that are placed on the cylindrical panorama in either an upright or inverted position. A heat beam from a lamp can be used as a negative reinforcer when the fly heads toward upright Ts and this heat can be turned off when the fly adjusts its flight towards an inverted T pattern. Memory in this paradigm is

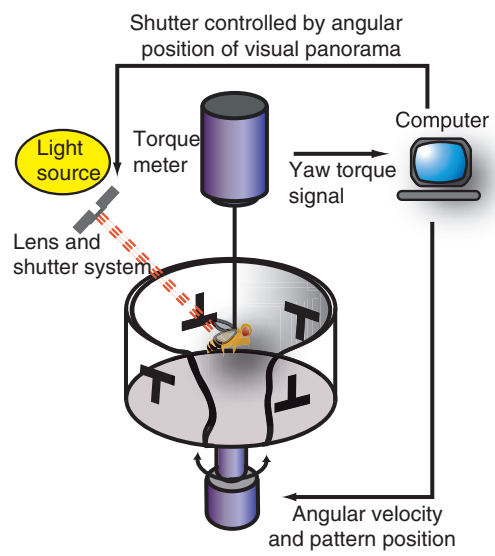

Figure 2 *Drosophila* flight simulator. The fly is attached to a torque meter by a thin wire that is glued to its head and/or thorax. The visual panorama surrounding the fly contains a visual cue (either an upright or an inverted T in this diagram) in every one of its four quadrants. Prior to training, flies exhibit no preference for either of the two cues as they fly in different directions. The fly's yaw torque is detected by an attached computer, which controls the angular velocity of the surrounding panorama in a negatively proportional manner. Thus, the panorama rotates and simulates changes in the fly's visual field that would correspond to movements in free flight. During training, a heat beam is directed to the fly's thorax and abdomen. The beam is turned on or off at the boundaries between quadrants that contain specific visual patterns using a shutter system that opens and closes based on the angular position of the panorama. Flies can be trained to associate an upright T, for example, with an aversive heat stimulus. After training, even with the heat turned off, flies will avoid the heat-reinforced quadrants and exhibit a preference for those quadrants that contain the visual cue that was not associated with heat. Reprinted by permission from Macmillan Publishers Ltd: [Nature] Liu L, Wolf R, Ernst R, et al. (1999) Context generalization in *Drosophila* visual learning requires the mushroom bodies. *Nature* 400: 753–756, copyright [1999].

measured as the tendency for the fly to avoid the heat-reinforced visual pattern after discontinuing the heat contingency.

Drosophila learning in the flight simulator is thought to be a combination of both classical and operant conditioning. The learning is operant in the sense that the fly controls heat reinforcement by its motor behaviors and classical in that the visual cues function as conditioned stimuli, with heat as the aversive unconditioned stimulus. Experiments to dissect the contributions of each type of learning have been performed. In a flight simulator without visual cues, flies can be punished with heat when they make either right- or left-hand turns. Their continued avoidance to make the turn associated with the reinforcer illustrates memory of this operant task. If the flight simulator is set so that the fly's movements are disconnected from controlling the environment, the flies can be trained to associate the visual cues of T patterns with heat, independent of their motor actions. In essence, the operator controls what the fly sees directly ahead and whether heat is provided or not. During testing, the torque meter is reconnected and the flies exhibit avoidance of the heat-reinforced patterns, thus exhibiting classical conditioning. Similar to olfactory associative conditioning, repeated training using a spaced protocol yields a long-term memory that is protein-synthesis dependent. Likewise, flies can exhibit an anesthesia-sensitive short-term memory lasting up to 20 min and an anesthesia-resistant memory from 20 min to 3 h.

Studies of visual learning in the flight simulator have provided several interesting findings regarding *Drosophila* visual learning and memory. The role of context, which is composed of stimuli of different modalities, has been investigated for this paradigm. For instance, changing the background illumination between the training and testing phases – from white to a monochromatic color, say blue or green, or from unpatterned white light to white light interspersed with dark stripes – produces no discernible effect on the performance of the flies. These experiments show that flies are capable of learning discrete pattern elements in different contexts. In other words, they exhibit context generalization. However, removing the mushroom bodies impairs the flies' ability to generalize such that they display memory only when tested in the same context as the initial training. Thus, mushroom bodies are required for context generalization. The importance of mushroom bodies has also been studied in *Drosophila* choice behavior. Flies can use both color and shape cues simultaneously in visual learning. However, when faced with a color/shape dilemma, for example, being trained to associate punishment with an upright blue T but having to choose between an upright green T or an inverted blue T during the test phase, the fly makes a discrete and firm choice based on the saliency of the visual cues. At low color intensities, the fly uses shape as the cue, whereas at high color intensity its decision is based on color. Mutant flies with absent or malformed mushroom bodies are unable to make a discrete choice for the more salient stimulus. Thus, while the mushroom bodies are dispensable for simple forms of visual learning using the flight simulator, they have a role in more complex tasks that require flies to evaluate the relative importance of different features of a complex visual environment.

The roles for other brain structures in visual learning have also been studied. Flies in the flight simulator analyze visual cues based on at least five different pattern parameters: size, color, elevation in the panorama, vertical compactness, and contour orientation. It has been demonstrated that a small group of neurons in the fan-shaped body of the central complex of *Drosophila* is important in short-term memory for the pattern parameters of 'elevation' and 'contour orientation.' The adenylyl cyclase encoded by the *rutabaga* gene is required in this brain region for flies to learn specific visual patterns. Specifically, restoring *rutabaga* function in fan-shaped body neurons that extend into layer 5 of *rutabaga* mutant flies rescues the memory defect for pattern elevation. Likewise, expression of a constitutively activated form of $G\alpha_s$ protein subunit in this brain region is sufficient to abolish the fly's memory for pattern elevation. Using a similar approach, memory for the pattern parameter of 'contour orientation' is restored in *rutabaga* mutants when a wild-type *rutabaga* transgene is expressed in neurons that extend into layer 1 of the fan-shaped body. These discoveries using the flight simulator provide encouraging initial clues toward understanding the circuitry that underlies visual pattern recognition.

Most learning paradigms used in the laboratory are used to deconstruct and reduce the learning that takes place into separable cues. However, this does not occur in nature. Another interesting discovery using the flight simulator involves crossmodal interactions in *Drosophila* learning. By adding olfactory cues during visual learning in the flight simulator it has been shown that concurrent visual and olfactory cues (bimodal conditioning) reduce the threshold for either visual or olfactory cues during memory retrieval. Furthermore, flies exhibit crossmodal memory transfer wherein sensory preconditioning of flies by simultaneous exposure to color and shape cues without reinforcement, followed by training using only a

single cue, leads to transfer of memory of one cue for the other. For example, flies simultaneously preexposed to odor and a visual cue, followed by conditioning by pairing heat punishment with the visual cue, can learn to avoid the olfactory cue that was associated with the visual cue in the preconditioning step. Sensory interactions like these may be important for fly learning in a natural environment.

Courtship Conditioning

Another type of learning that occurs within the normal behaviors of sexual courtship is courtship conditioning. Courting males perform a characteristic sequence of behaviors that includes orienting toward and following a female, tapping her with his forelegs, vibrating one wing, licking her genitalia, and mounting. Virgin females are generally receptive and quickly accept these advances of a male. If a male *Drosophila* is paired with a female that has already been courted and fertilized by another male, however, the female will reject the advances of the new male. This experience decreases the subsequent courtship activity of the rejected male so that the rejected male will not court other young virgin females for time periods that depend upon the strength of the initial courtship activity that led to rejection. If the male is paired with the fertilized female for a period of 1 h, courtship suppression lasts for 2–3 h afterwards. If the male experiences the fertilized female for 5 h, courtship suppression can persist for up to 1 week. Experience-dependent courtship suppression is an associative form of learning because it requires the presence of the trainer fly (usually a fertilized female) and associated aversive pheromones emitted by the mated female.

Mushroom bodies are also involved in the memory of courtship conditioning. Male flies missing their mushroom bodies due to chemical ablation suppress their courtship like control males do immediately after training with a fertilized female, but begin courting again about 30 min after the experience. Thus, mushroom bodies are required for memory after 30 min when trained with a fertilized female for a period of 1 h. Flies without mushroom bodies that receive extended training of 5 h exhibit no memory when tested 1 day (or more) later, suggesting that mushroom bodies are required for long-term memory after extended conditioning.

However, these specific effects of mushroom body ablation on later memory are observed only when the animals are trained in the light. If training is performed in dim red light, which flies do not detect very well, immediate memory tested within a few minutes after training is also impaired. These results are compatible with the model that courtship conditioning employs both visual and olfactory pathways, and in the absence of the visual pathway (dim red light), mushroom body ablation disrupts the major sensory pathway necessary for memory. In the presence of light, mushroom body-less flies are less amnestic, because memory traces presumably exist in the visual system and partially compensate for the impairment in the olfactory pathway. Studies of mutants and flies expressing dominant negative transgenes have revealed a requirement for *dunce*, *rutabaga*, *amnesiac*, and calcium/calmodulin-dependent kinase II for normal courtship conditioning.

Heat-Box Learning

Insects do not internally control body temperature and therefore have body temperatures that closely approximate the ambient temperature of their environment. Rapid modifications of behavior are therefore critical to avoid extreme fluctuations in body temperature that could produce death. *Drosophila* are able to sense local differences in temperature, preferring 24 °C and avoiding higher and lower temperatures when tested in a heat gradient. They are also very sensitive to temperature fluctuations and can adjust their behavior in less than a minute to avoid temperature fluctuations of more than a few degrees celsius. Using an apparatus called a heat box, *Drosophila* researchers have shown that flies are able to learn place or position.

The heat box (**Figure 3**) is a laterally transparent rectangular chamber with Peltier elements comprising the upper and lower surfaces. The Peltier elements are kept at a permissive temperature (25 ± 2 °C) or are heated to a higher restrictive temperature (usually 37 ± 2 °C) using a computerized control circuit. The chamber is divided into two halves using a directionally selective infrared light gate, invisible to the fly, which monitors whether the fly is in one or the other half. The clear plastic sides allow the transmission and detection of infrared light, and when the fly walks the length of the chamber it casts a shadow on the light gate array and a computer detects this signal.

For heat-box training, a single fly is allowed to walk freely between the two halves in complete darkness. When the fly enters one side of the chamber, the whole chamber is heated to the elevated temperature, and when the fly runs back to the other side, the chamber is cooled to the lower temperature. Flies learn from this experience, within minutes, to avoid the side of the chamber that elevates temperature. Memory of this experience is displayed by continued

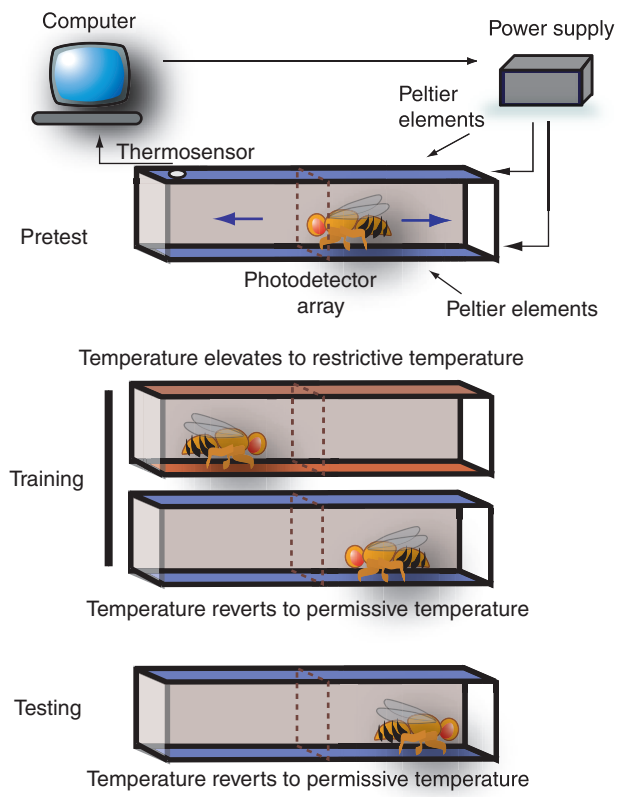

Figure 3 *Drosophila* place learning paradigm in the heat-box. Schematic of the heat-box learning assay. During the pretest, single flies are allowed to walk freely in the chamber. During training, the whole chamber heats to the restrictive temperature whenever the fly walks into a predesignated 'punished' half (the left half in this diagram). A photodetector at the middle of the chamber sends a signal to a computer whenever the fly moves from one side to the other, which subsequently controls the Peltier elements that control chamber temperature. The whole chamber cools to the permissive temperature when the fly returns to the unreinforced side. Memory is tested following training in the absence of reinforcement, during which normal flies exhibit a preference for the unreinforced side. Adapted with permission from Cold Spring Harbor Laboratory Press: Putz G and Heisenberg M (2006) Memories in *Drosophila* heat-box learning. *Learning & Memory* 9: 349–359. Copyright 2006 by Cold Spring Harbor Laboratory Press.

avoidance of the punished side even in the absence of the heat contingency.

This operant conditioning paradigm yields significant place memory for up to 2 h after training, which is dependent on tactile and self-derived idiothetic cues, probably based on walking steps and turns that allow the flies to orient themselves in the chamber. The place memory is independent of visual and olfactory stimuli. Flies lacking mushroom bodies have normal place memory. However, *rutabaga* mutants are impaired in the task, indicating a requirement for adenylyl cyclase activity outside of the mushroom bodies.

See also: Conditioning: Theories; Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Olfaction in Invertebrates: *Drosophila*; Olfactory System: Circuit Dynamics and Neural Coding in the Locust; Operant Conditioning of Reflexes; Procedural Learning: Classical Conditioning.

Further Reading

Brembs B and Heisenberg M (2000) The operant and the classical in conditioned orientation of *Drosophila melanogaster* at the flight simulator. *Learning & Memory* 7: 104–115.

Davis RL (2005) Olfactory memory formation in *Drosophila*: From molecular to systems neuroscience. *Annual Review of Neuroscience* 28: 275–302.

Guo J and Guo A (2005) Crossmodal interactions between olfactory and visual learning in *Drosophila*. *Science* 309: 307–310.

Joiner MA and Griffith LC (2000) Visual input regulates circuit configuration in courtship conditioning of *Drosophila melanogaster*. *Learning & Memory* 7: 32–42.

Liu G, Seiler H, Wen A, et al. (2006) Distinct memory traces for two visual features in the *Drosophila* brain. *Nature* 439: 551–556.

Liu L, Wolf R, Ernst R, et al. (1999) Context generalization in *Drosophila* visual learning requires the mushroom bodies. *Nature* 400: 753–756.

Mao Z, Roman G, Zong L, et al. (2004) Pharmacogenetic rescue in time and space of the rutabaga memory impairment using Gene-Switch. *Proceedings of the National Academy of Sciences of the United States of America* 101: 198–203.

McBride SM, Giuliani G, Choi C, et al. (1999) Mushroom body ablation impairs short-term memory and long-term memory of courtship conditioning in *Drosophila melanogaster*. *Neuron* 24: 967–977.

McGuire SE, Le PT, Osborn AJ, et al. (2003) Spatio-temporal rescue of memory dysfunction in *Drosophila*. *Science* 302: 1765–1768.

Putz G and Heisenberg M (2006) Memories in *Drosophila* heat-box learning. *Learning & Memory* 9: 349–359.

Schwaerzel M, Monastirioti M, Scholz H, et al. (2003) Dopamine and octopamine differentiate between aversive and appetitive olfactory memories in *Drosophila*. *Journal of Neuroscience* 23: 10495–10502.

Siwicki KK and Ladewski L (2003) Associative learning and memory in *Drosophila*: Beyond olfactory conditioning. *Behavioral Processes* 64: 225–238.

Yu D, Keene AC, Srivatsan A, et al. (2005) *Drosophila* DPM neurons form a delayed and branch-specific memory trace after olfactory classical conditioning. *Cell* 123: 945–957.

Yu D, Ponomarev A, and Davis RL (2004) Altered representation of the spatial code for odors after olfactory classical conditioning: Memory trace formation by synaptic recruitment. *Neuron* 42: 437–449.

Learning and Memory in Invertebrates: *Hermissenda*

T Crow, J-J Xue-Bian, and L-M Tian, University of Texas Medical School, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

An understanding of fundamental mechanisms of learning and memory is a major challenge for neuroscience. Well-documented descriptions of how learning is expressed in the generation of behavior have been especially difficult to obtain from studies of complex nervous systems. A strategy that has been productive in addressing these issues is to examine learning at both the behavioral and the cellular/molecular levels in animals with relatively less complex nervous systems. Using Pavlovian conditioning of invertebrates as a model for associative learning has led to the identification of cellular and synaptic mechanisms underlying the formation of basic associations. However, an understanding of associative processes underlying Pavlovian conditioning requires an explanation of not only the mechanisms of temporal contiguity and/or predictability between the conditioned stimulus (CS) and unconditioned stimulus (US), but also how conditioning-dependent changes in the relevant neural circuitry of the nervous system are expressed in the generation of the conditioned response (CR). Studies with invertebrates have provided the opportunity to examine how mechanisms of associative learning are expressed in the neural circuitry supporting the generation of learned behavior.

The nudibranch mollusk *Hermissenda crassicornis* has contributed to an understanding of Pavlovian conditioning at the cellular, molecular, and systems level. The *Hermissenda* central nervous system is relatively simple, which provides for the opportunity of studying modifications in identifiable neurons in the neural circuitry that supports conditioning. Identified neurons in the CS pathway have been studied in detail using biochemical, biophysical, and molecular techniques. The two sensory structures mediating the CS and US are central and thus their synaptic projections remain totally intact after surgical isolation of the nervous system or in experiments using semi-intact preparations. Studies of semi-intact preparations provide for the opportunity to examine the development of CS–US associations at identified sites in the nervous system and their contributions to the expression or generation of behavior. Mechanisms of CS–US contiguity have been identified and have been the focus of biophysical, biochemical, and molecular analyses. Recent studies have led to the identification of neurons that contribute to the neural circuitry supporting both the unconditioned response (UR) and the CR. Since conditioning can be studied in semi-intact preparations, an understanding of how Pavlovian conditioning is expressed in the generation of the CR is now feasible.

Pavlovian Conditioning

Pavlovian conditioning of *Hermissenda* involves stimulation of visual and graviceptive sensory pathways. The conditioning procedure consists of pairing light, the CS, with rotation or orbital shaking, the US. Rotation of the statocyst, the gravity-detecting organ, is an adequate stimulus for evoking depolarizing generator potentials and an increase in spike activity in the primary graviceptive sensory neurons, the hair cells. Two unconditioned responses, a reduced rate of forward locomotion and reflexive foot-shortening are elicited by rotation. Therefore, Pavlovian conditioning in *Hermissenda* results in the acquisition of two different CRs: inhibition of light-elicited locomotion and CS-elicited foot-shortening (see Figure 1). Dark-adapted *Hermissenda* exhibit a well-documented positive phototaxic response to illumination. Conditioning produces both light-elicited inhibition of the normal positive phototaxis and CS-elicited foot-shortening. Inhibition of phototaxis produced by conditioning is expressed by a light-dependent inhibition in the initiation of locomotion and a reduced rate of forward locomotion in light (see Figure 1). The two CRs are proposed to develop independently, which is in agreement with recent cellular studies showing that the URs involve different components of the neural circuitry responsible for foot contraction and ciliary locomotion. In this example of associative learning, both conditioning-dependent modifications of synaptic function and intrinsic cellular excitability in identified components of the neural circuit supporting ciliary locomotion result in the generation of the CR, phototaxic inhibition.

Conditioning in the two different behavioral response systems supporting the two CRs is sensitive to both CS–US contiguity and forward interstimulus-interval manipulations. Moreover, both foot contraction and inhibition of phototaxis produced by conditioning involve the development or emergence of a new response to the CS, not the potentiation (through US presentations) of an already existing response to the CS, referred to as reflex potentiation

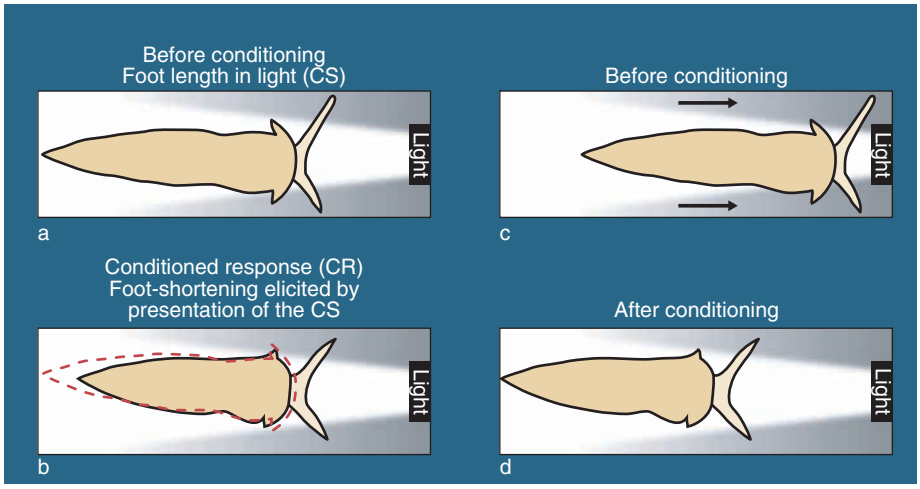

Figure 1 Pavlovian conditioning of foot-shortening and phototaxic inhibition in *Hermissenda*. (a) Foot length in light (conditioned stimulus; CS) before conditioning. (b) Conditioned response, foot-shortening elicited by the CS. Red outline indicates foot length in light before conditioning. (c) Light-elicited locomotion toward a light source (phototaxis) assessed before conditioning. (d) Inhibition of light-elicited locomotion detected after Pavlovian conditioning. Random or pseudorandom presentations of the CS and the unconditioned stimulus produce neither inhibition of phototaxis nor CS-elicited foot-shortening. Adapted with permission from Crow T (2004) Pavlovian conditioning of *Hermissenda*: Current cellular, molecular and circuit perspectives. *Learning and Memory* 11: 229–238. Copyright 2004 by Cold Spring Harbor Laboratory Press.

or alpha conditioning. In both of the CRs there is a transfer of functional aspects of the response-evoking properties of the US to the CS. This important characteristic of Pavlovian conditioning may be expressed in *Hermissenda* due to the increased complexity of the circuit supporting the CS and US, the multiple sites of convergence in the CS–US pathways, and multiple synaptic interactions within the neural network supporting the generation of behavior.

Primary Sensory Neurons of the CS and US Pathways

The morphology and physiology of the two sensory structures that are stimulated by the CS and US have been described in detail. In addition, the convergence sites providing for synaptic interactions between the CS and US pathways have been identified and examined using intracellular recording techniques.

Photoreceptors

Each eye of *Hermissenda* contains five photoreceptors: three classified as type B and two as type A. The general classification of photoreceptors can be identified further based upon their location within the eye. There are medial and lateral A and B photoreceptors and one central B photoreceptor. The synaptic connections between type B photoreceptors and between type B and type A photoreceptors are in the neuropil of the cerebropleural ganglion, and the synaptic interactions between type B photoreceptors are mutually inhibitory. The photoreceptors of *Hermissenda*, like many invertebrates, are both sensory transducers and spike generators. Light produces a depolarizing generator potential and an increase in spike activity in both type A and type B photoreceptors.

Hair Cells

The sensory structures stimulated by the US are the two central gravity-detecting statocysts. Each statocyst contains 13 hair cells whose cell bodies are located around the perimeter of the statocyst. Hair cells that are located in opposite positions in the statocyst are mutually inhibitory. Statocyst hair cells contact calcium carbonate particles, referred to as statoconia, by interactions with motile cilia that project into the lumen of the statocyst from the apical region of the somas. Rotation or gravity causes the statoconia to press against the motile cilia of hair cells in front of the centrifugal or gravitational force vector, resulting in a depolarizing generator potential and an increase in spike activity. Like the photoreceptors, statocyst hair cells are sensory transducers and support the generation of action potentials.

Interneurons in the UR Pathway

Statocyst hair cells project to photoreceptors and three identified types of interneurons located in the cerebropleural ganglia. Hair cells form monosynaptic connections with both photoreceptors and type I_e and type I_i interneurons. In addition, hair cells project polysynaptically to type I_b interneurons. Type I_e and

type I_i interneurons project polysynaptically to type III_i inhibitory interneurons, and type III_i interneurons inhibit ciliary activating motor neurons. In summary, rotation or orbital shaking, the US, depolarizes statocyst hair cells, resulting in excitation of type I_e interneurons, which produces excitation of type III_i inhibitory interneurons that inhibit or decrease the spike activity of ciliary motor neurons. Therefore, an increase in the spike activity of type III_i interneurons results in inhibition of ciliary locomotion. Activation of the identified components of the circuit supporting the US provides an explanation of the effect of rotation on ciliary locomotion.

Convergence of the CS and US Pathways

There are multiple sites of synaptic interactions between cells in the CS and US pathways. The initial site of convergence between the CS and US is at the primary sensory neurons of the visual and graviceptive pathways. The synaptic projections from statocyst hair cells to the photoreceptors are both monosynaptic and polysynaptic. Hair cells and photoreceptors form reciprocal monosynaptic inhibitory connections. Caudal hair cells inhibit photoreceptors, and cephalic hair cells are inhibited by type B photoreceptors. The second site of convergence between the CS and US pathways involves aggregates of identified interneurons in the cerebropleural ganglia. The synaptic organization of the interneurons that make up the convergence site between the visual and graviceptive pathways has now been characterized. Photoreceptors and hair cells project to aggregates of 'on' and 'off' cells designated as type I_e and type I_i cerebropleural interneurons (see Figure 2). An additional site of convergence between the CS and US pathways is the type I_b interneurons that form monosynaptic connections with foot contractile motor neurons and ciliary-activating motor neurons.

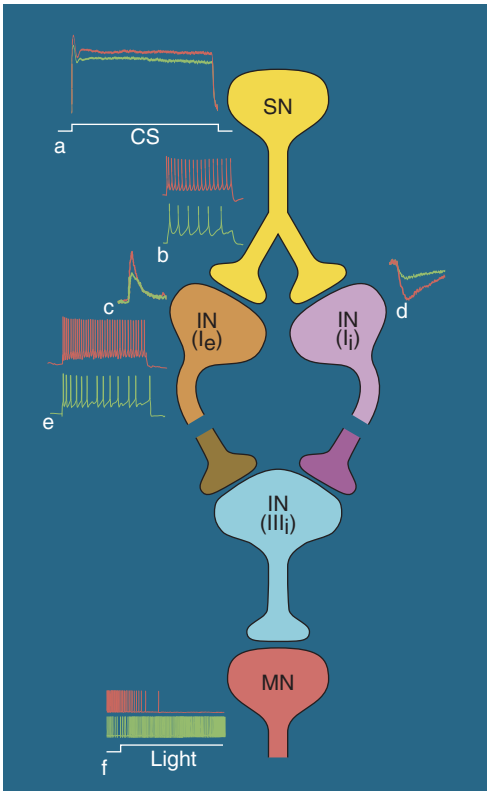

Figure 2 Intrinsic changes in synaptic efficacy and excitability in the components of the conditioned stimulus–conditioned response (CS–CR) pathway supporting Pavlovian conditioning of *Hermissenda*. Changes in both synaptic efficacy and cellular excitability contribute to CS-elicited inhibition of ciliary locomotion. Inserts a–f: Neural activity from a conditioned preparation (red), and a pseudorandom control (green). (a) The CS evokes a larger generator potential in sensory neurons (SN) (type B-photoreceptors) of conditioned animals as compared to pseudorandom controls. (b) Conditioning results in intrinsic enhanced excitability of SNs. (c) The monosynaptic excitatory postsynaptic potential in type I_e interneurons (IN) is facilitated in conditioned animals. (d) The monosynaptic inhibitory postsynaptic potential in type I_i interneurons is facilitated in conditioned animals. (e) Conditioning results in intrinsic enhanced excitability of type I_e interneurons. The net effect of the changes in synaptic efficacy and intrinsic enhanced excitability results in an increase in CS-elicited spike activity of type III_i interneurons, which produces an inhibition of ciliary motor neuron activity. (f) The CS inhibits spike activity in ciliary motor neurons (MN) and produces an increase in spike activity recorded from motor neurons of pseudorandom controls. The inhibition of ciliary motor neurons produces phototaxic inhibition in conditioned animals. Note: the broken connections between type I interneurons and the type III_i interneurons indicate an established polysynaptic projection. A monosynaptic connection between type I and type II interneurons, and type II and type III interneurons, has not been established. Adapted with permission from Crow T (2004) Pavlovian conditioning of *Hermissenda*: Current cellular, molecular and circuit perspectives. *Learning and Memory* 11: 229–238. Copyright 2004 by Cold Spring Harbor Laboratory Press.

Cellular and Molecular Mechanisms of Conditioning in Sensory Neurons and Interneurons Contributing to Pavlovian Conditioning in *Hermissenda*

The primary sensory neurons of the CS pathway (photoreceptors) and type I interneurons are identified sites of plasticity involving both intrinsic enhanced excitability and synaptic facilitation. Cellular correlates of conditioning in type B photoreceptors are expressed by enhancement of the CS-elicited generator potential and increased CS-evoked spike frequency, increased excitability to extrinsic current, modification of light-dependent, Ca^{2+}-dependent, and voltage-dependent membrane currents, and increased phosphorylation of several identified proteins. In addition to changes in primary sensory neurons, facilitation of monosynaptic and complex postsynaptic potentials (PSPs) in type I_e and type I_i interneurons is

produced by conditioning. The increased amplitude of the CS-elicited generator potential and the increase in the number of action potentials elicited by the CS are major contributors to the increased duration and amplitude of CS-elicited complex PSPs and modifications in spike activity in type I interneurons of conditioned preparations. However, the facilitation of the amplitude of monosynaptic PSPs between type B photoreceptors and type I_e and type I_i interneurons may involve both pre- and postsynaptic mechanisms (Figure 2). In addition to synaptic facilitation of the connections between B photoreceptors and type I interneurons, type I_e interneurons also express intrinsic enhanced excitability (Figure 2).

Signal Transduction Pathways

Signal transduction pathways contributing to the modifications of excitability and synaptic efficacy in type B photoreceptors of conditioned animals involve protein kinase C (PKC), extracellular signal-regulated protein kinase (ERK), cyclin-dependent kinase 5 (Cdk5), and Rho GTPase activity and its downstream target, ROCK.

Light (CS) and rotation (US) have spatially separate physiological consequences in type B photoreceptors. However both the CS and US increase cytosolic Ca^{2+} levels. Light activates phospholipase C (PLC) to produce inositol trisphosphate (IP_3) and diacylglycerol (DAG). IP_3 opens rhabdomeric Na^+ and Ca^{2+} channels, which results in a depolarizing generator potential and Ca^{2+} influx into the photoreceptors. IP_3 can also bind to its receptor (IP_3R), which triggers Ca^{2+} release from the smooth endoplasmic reticulum (ER). The Ca^{2+} influx from the rhabdomeric and the IP_3R-gated storage compartment can cause Ca^{2+} release from the ryanodine receptor-gated compartment. Rotation (US) elicits a monosynaptic GABAergic inhibitory postsynaptic potential (IPSP) in photoreceptors. The US is also proposed to activate a serotonergic polysynaptic pathway that enhances type B photoreceptor excitability. The GABAergic IPSPs in photoreceptors activate phospholipase A_2 (PLA_2) to liberate arachidonic acid (AA) and create a back-propagating wave of Ca^{2+} released from intracellular stores. When the CS and US are repeatedly paired, the Ca^{2+} influx, due to light, IP_3R stores, RyR stores, and voltage-gated Ca^{2+} channel activation sum together. The large increase in cytosolic Ca^{2+} combined with DAG and AA act to synergistically activate PKC by translocation of PKC to the membrane. Each pairing of the CS and US may thus incrementally increase the proportion of PKC translocated to the membrane. Activation of PKC is necessary for the induction of plasticity in *Hermissenda* and contributes to the phosphorylation of proteins in sensory neurons associated with conditioning. Downregulation of PKC or pretreatment with kinase inhibitors blocks the induction of short-term excitability, but not long-term excitability, suggesting the involvement of other signaling pathways. ERK, Cdk5, and the Rho/ROCK signaling pathways have been investigated in conditioned *Hermissenda*. One-trial and multitrial Pavlovian conditioning results in the activation and phosphorylation of ERK. The increased phosphorylation of ERK is blocked by pretreatment with an inhibitor of ERK-activating kinase. Serotonin, a proposed transmitter in the US pathway, also activates ERK through both a Ca^{2+}-dependent PKC pathway and a PKC-independent pathway. Rho GTPase activity and its downstream target, ROCK, contribute to the phosphorylation of a protein (cytoskeleton-related protein 24; Csp24) contributing to intermediate-term memory (see next section). Cyclin-dependent kinase 5 (Cdk5) is also involved in Csp24 phosphorylation. Inhibition of ROCK by the ROCK inhibitor Y-27632 significantly increases Csp24 phosphorylation. The Rho activator lysophosphatidic acid (LPA) or the Cdk5 inhibitor butyrolactone decreases Csp24 phosphorylation. Treatment of the nervous system with LPA activates Rho, while exposure to serotonin decreases Rho activity. The relationship between Rho GTPase activity and Cdk5 activity and enhanced excitability has not been established.

Proteins Regulated by Pavlovian Conditioning

Two proteins that are associated with conditioning in *Hermissenda* are regulated by conditioning-dependent phosphorylation. Calexcitin (CE) is a GTP- and Ca^{2+}-binding protein that can decrease K^+ currents in type B photoreceptors and may increase cytosolic Ca^{2+} concentration. CE is phosphorylated by PKC, and increased CE levels in B photoreceptors of conditioned animals results in increased excitability via K^+ channel inactivation and Ca^{2+} release from the ER due to increased CE binding to ryanodine receptors. A second protein regulated by conditioning is cytoskeleton-related protein 24, a member of the family of β-thymosin repeat proteins (i.e. proteins that contain multiple actin-binding domains). Csp24 is found throughout the circumesophageal nervous system and in the soma, axon, and terminal processes of type B photoreceptors. Actin is bound by Csp24, and actin binding is regulated in part by PKC phosphorylation of Csp24. The Rho/Rock and Cdk5 signaling pathways also contribute to the regulation of Csp24 phosphorylation. Csp24 is phosphorylated by one-trial and multitrial Pavlovian conditioning, and downregulation of Csp24 expression by an antisense oligonucleotide blocks the development of intermediate-term enhanced excitability associated with one-trial conditioning in identified type B photoreceptors. In isolated

photoreceptors, one-trial *in vitro* conditioning results in a reduction in the amplitude of the A-type transient K^+ current (I_A) and a depolarized shift in the steady-state activation curve of I_A without altering I_A inactivation. The conditioning-dependent reduction in I_A can be blocked by preincubation of isolated photoreceptors with a *Csp* antisense oligonucleotide. The regulation of Csp24 by conditioning may influence K^+ channel activity by the spatial and temporal control of actin dynamics.

Circuitry Supporting the Generation of the CR

Many marine mollusks exhibit a gliding form of locomotion produced by the movement of motile cilia on the surface of the foot. Animals are propelled forward by the movement of the cilia between a mucus film secreted from the foot and the underlying substrate. Studies in semi-intact *Hermissenda* preparations have shown that statocyst hair cell-mediated foot contraction and graviceptive ciliary locomotion involve different interneuronal circuit components from the circuit that supports light-elicited ciliary locomotion. The major components of the neural circuit underlying light-elicited ciliary locomotion in *Hermissenda* are shown in the diagram in Figure 2. Only one of the five photoreceptors in each eye, the lateral type B, is shown in the circuit. However, each identified type B and type A photoreceptor forms monosynaptic connections with different aggregates of 'on–off' cells; the type I_e and type I_i interneurons. Each photoreceptor projects to its own aggregate of at least two type I_e interneurons, and two type I_i interneurons. The synaptic connections between photoreceptors and type I interneurons are monosynaptic. The next component in the CR circuit involves converging polysynaptic projections from the type I interneurons to a type III_i inhibitory interneuron (see Figure 2). A direct monosynaptic connection between type I and type II interneurons, or type II interneurons and type III_i interneurons, has not yet been established. However, extrinsic current stimulation of type I interneurons evoke PSPs recorded in type III_i interneurons and ciliary motor neurons. Therefore, synaptic convergence from both I_e and I_i interneurons is found at the level of type III_i interneurons. Type III_i interneurons form monosynaptic inhibitory connections with identified ciliary motor neurons.

How does illumination modulate the spike activity of ciliary motor neurons? Illumination of the photoreceptors decreases the frequency of IPSPs recorded from identified ciliary motor neurons. The CS, light, regulates the activity of ciliary motor neurons by hyperpolarizing type I_i interneurons ('off' cells) and depolarizing type I_e interneurons ('on' cells). The integration of synaptic input from type I interneurons results in a net decrease in the spike activity of type III_i inhibitory interneurons during light. The decreased spike activity of type III_i inhibitory interneurons during the CS decreases the inhibition of ciliary motor neurons by type III_i interneurons, resulting in an increase in motor neuron spike activity and an increase in ciliary activity on the foot. In an untrained or normal animal, the light-induced hyperpolarization of type I_i interneurons is more effective in decreasing the frequency of IPSPs in ciliary motor neurons than is the concomitant excitation of type III_i interneurons provided by light-induced excitatory input from depolarized type I_e interneurons. Therefore, the effect of light on the activity of ciliary motor neurons is an increase in spikes and increased ciliary movement.

Conditioning-dependent reconfiguration of the neural circuitry underlying light-elicited ciliary locomotion has now been analyzed. Pavlovian conditioning produces changes in both intrinsic excitability and synaptic function in identified photoreceptors. Conditioning results in a significant increase in CS-elicited spike activity in type I_e interneurons as compared to type I_e spike activity of controls that receive independent random presentations of the CS and US with the restriction that the two stimuli could not overlap in time (i.e., pseudorandom). CS-evoked complex EPSPs in type I_e interneurons following conditioning are larger in amplitude and exhibit a greater frequency of smaller EPSPs in conditioned preparations as compared with type I_e complex EPSPs recorded from pseudorandom controls. The monosynaptic EPSP in type I_e interneurons elicited by a single spike in a lateral B photoreceptor is facilitated in conditioned animals as compared to pseudorandom controls. Similar synaptic changes (facilitation) occur in type I_i interneurons in conditioned animals as compared to pseudorandom controls. The CS elicits a significantly larger amplitude complex IPSP recorded in type I_i interneurons of conditioned animals as compared to pseudorandom controls. The greater inhibition of type I_i interneurons evoked by the CS produces a significant decrease in spike activity of type I_i interneurons relative to pseudorandom controls. Therefore, the effect of light on the spike activity of type I_i interneurons results in a greater decreased excitation of type III_i interneurons in conditioned animals. Taken collectively, the conditioning correlates detected in type I interneurons can be explained by well-documented intrinsic changes in the type B photoreceptors. The increase in the amplitude of the complex PSPs in type I interneurons may be due to both the light-evoked enhanced excitability of the

photoreceptors produced by conditioning and the facilitation of the monosynaptic PSPs between lateral B photoreceptors and type I interneurons. However, postsynaptic changes intrinsic to type I interneurons could also contribute to the modifications of type I interneurons detected in conditioned animals. Intrinsic excitability of type I_e interneurons is enhanced in conditioned animals as compared to pseudorandom controls. The intrinsic enhanced excitability potential of type I_e interneurons detected in conditioned animals increases the frequency of IPSPs recorded in type III_i interneurons produced by light-evoked depolarization of the type I_e interneurons. In contrast to the effects of enhanced excitability in I_e interneurons with conditioning, the light-evoked hyperpolarization of type I_i interneurons observed before conditioning decreases spike activity of type III_i interneurons and produces a decrease in IPSP frequency of ciliary motor neurons. Extrinsic current stimulation of I_e interneurons in conditioned animals results in more IPSPs recorded in ciliary motor neurons as compared to pseudorandom controls. The increase in the number of IPSPs detected in ciliary motor neurons can be accounted for by the increase in the number of spikes in type III_i interneurons produced by the increased excitation of type I_e interneurons of conditioned animals as compared to controls. Extrinsic current stimulation of I_e interneurons produces a significant increase in IPSP frequency in ciliary motor neurons of conditioned animals relative to controls. Therefore, the presentation of the CS in conditioned animals results in a decrease in ciliary motor neuron spike activity and an inhibition of firing during the light (**Figure 2**). In contrast, pseudorandom controls exhibit an increase in spike activity recorded in ciliary motor neurons during the light.

In summary, a combination of intrinsic excitability changes and modifications in synaptic function at specific loci in the circuit responsible for light-dependent ciliary locomotion is involved in the generation or expression of the CR elicited by the CS. Before conditioning, the light-dependent inhibition of type I_i interneurons is very effective in regulating IPSP frequency of ciliary motor neurons and their subsequent firing. Indeed, the net effect of light-dependent excitation of type I_e interneurons and light-dependent inhibition of type I_i interneurons is a decrease in IPSP frequency of ciliary motor neurons, an increase in their spike activity, and increased ciliary movement. The induction of intrinsic enhanced excitability of type I_e interneurons produced by conditioning reconfigures the neural circuit such that light, the CS, now increases IPSP frequency in ciliary motor neurons, which leads to a decrease in their spike activity during illumination and inhibition of phototaxis.

Conclusions and Discussion

The research focused on how Pavlovian conditioning is expressed in the generation of phototactic behavior in *Hermissenda* involves the identification of the neural circuit controlling ciliary locomotion and its modulation by light (CS) and graviceptive input (US). The analysis of Pavlovian conditioning in the neural circuit generating ciliary locomotion indicates that both enhanced cellular excitability and synaptic facilitation are expressed in identified circuit components at different loci within the network. The distributed nature of cellular and synaptic plasticity associated with this example of Pavlovian conditioning suggests that an adequate explanation of conditioned behavior requires both an analysis of neural circuits and the identification of mechanisms of CS–US contiguity at convergence sites between the CS and US pathways. Consistent with the view that learning may initially involve changes in preexisting synaptic connections, inhibition of phototaxic behavior produced by conditioning involves modifications of existing synaptic connections between photoreceptors and identified type I interneurons. However, the potential formation with conditioning, of new connections between neurons in the neural circuit modulating ciliary locomotion or CS-elicited foot contraction cannot be dismissed.

The analysis of conditioning correlates has revealed that the first site of intrinsic cellular and synaptic plasticity is at the initial site of convergence between the CS and US pathways, the primary sensory neurons of the CS pathway. The mechanisms of temporal contiguity between the CS and US involve both enhanced cellular excitability and enhanced synaptic strength. The changes in photoreceptor excitability produced by conditioning involve reductions in several well-characterized K^+ conductances in type B photoreceptors. The second site of intrinsic enhanced excitability is the type I_e interneurons. An analysis of membrane conductances underlying enhanced excitability intrinsic to type I_e interneurons has not yet been undertaken. The cellular and synaptic changes identified following conditioning are distributed at several loci within the network and therefore are not localized to a single synaptic site or neuron. The distributed nature of learning-dependent changes in the neural circuit responsible for the expression of the CR may account for the complexity of Pavlovian conditioning in *Hermissenda*, specifically the emergence of a new response to the CS following conditioning.

See also: Conditioning: Theories; Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Operant Conditioning of Reflexes; Plasticity of Intrinsic Excitability; Procedural Learning: Classical Conditioning.

Further Reading

Akaike T and Alkon DL (1980) Sensory convergence on central visual neurons in *Hermissenda*. *Journal of Neurophysiology* 44: 501–513.

Beggs TM, Brown TH, Byrne JH, et al. (1999) Learning and memory: Basic mechanisms and systems. In: Bloom F, Landis S, Roberts L, et al. (eds.) *Fundamental Neuroscience*, pp. 1411–1454. San Diego: Academic Press.

Crow T (2004) Pavlovian conditioning of *Hermissenda*: Current cellular, molecular and circuit perspectives. *Learning and Memory* 11: 229–238.

Crow TJ and Alkon DL (1978) Retention of an associative behavioral change in *Hermissenda*. *Science* 201: 1239–1241.

Crow T and Alkon DL (1980) Associative behavioral modification in *Hermissenda*: Cellular correlates. *Science* 209: 412–414.

Crow T and Tian L-M (2002) Facilitation of monosynaptic and complex PSPs in type I interneurons of conditioned *Hermissenda*. *Journal of Neuroscience* 22: 7818–7824.

Crow T and Tian L-M (2002) Morphological characteristics and central projections of two types of interneurons in the visual pathway of *Hermissenda*. *Journal of Neurophysiology* 87: 322–332.

Crow T and Tian L-M (2003) Interneuronal projections to identified cilia-activating pedal neurons in *Hermissenda*. *Journal of Neurophysiology* 89: 2420–2429.

Crow T and Tian L-M (2003) Neural correlates of Pavlovian conditioning in components of the neural network supporting ciliary locomotion in *Hermissenda*. *Learning & Memory* 10: 209–216.

Crow T and Tian L-M (2004) Statocyst hair cell activation of identified interneurons and foot contraction motor neurons in *Hermissenda*. *Journal of Neurophysiology* 92: 2874–2883.

Detwiler PB and Fuortes MG (1973) Responses of hair cells in the statocyst of *Hermissenda*. *Journal of Physiology* 251: 107–129.

Frysztak RJ and Crow T (1994) Enhancement of type B and A photoreceptor inhibitory synaptic connections in conditioned *Hermissenda*. *Journal of Neuroscience* 14: 1245–1250.

Gandhi CC and Matzel LD (2000) Modulation of presynaptic action potential kinetics underlies synaptic facilitation of type B photoreceptors after associative conditioning in *Hermissenda*. *Journal of Neuroscience* 20: 2022–2035.

Lederhendler II, Gart S, and Alkon DL (1986) Classical conditioning of *Hermissenda*: Origin of a new response. *Journal of Neuroscience* 6: 1325–1331.

Sahley CL and Crow T (1998) Invertebrate learning: Current perspectives. In: Martinez JL Jr. and Kesner RP (eds.) *Learning and Memory*, pp. 197–209. New York: Academic Press.

Learning and Memory in Invertebrates: Honey Bee

R Menzel, Freie Universität Berlin, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

The social life of the honeybee colony forms the evolutionary framework for the individual animal's behavior and is crucial for the survival of each bee. Social bees are exposed to a large and unpredictable range of environmental conditions, and thus learning plays an important role both in the social context and in individual adaptations. The location of the nest site needs to be learned because individual bees cannot survive. Changes to the society's conditions have to be recognized and used to control individual behavior. Communication about potential nest sites and food, water, and resin resources involves a sophisticated behavioral routine – the waggle dance – during which receivers of the information learn about the location and the quality of the respective resource.

Most studies of learning in bees have focused on operant learning in the context of food collection and on appetitive olfactory conditioning of a response, the proboscis extension response (PER). When a searching bee discovers a source of sucrose solution (under natural conditions, nectar), it quickly learns to associate the surrounding visual and olfactory signals with the reward. Acquisition functions indicate that olfactory stimuli (e.g., floral odorants) are learned within one or two learning trials, colors within one to five learning trials, and black-and-white patterns from the fifth learning trial on. Most of this learning is forward conditioning because signals perceived before the reward are associated, whereas those perceived during the reward or departure flight are associated less effectively or not at all. However, stimuli in the further vicinity of the reward localization are also learned after feeding.

Reversal learning, overlearning, and multireversal learning are effective learning phenomena in the bee, as they are for other animals. The experimental procedures that test the ability of an animal to adapt its performance to reward schedules are continuous reinforcement (CR), fixed-ratio reinforcement (FR), and fixed-interval reinforcement (FI) schedules. Bees develop a higher resistance to extinction in an FR than in a CR schedule. In an FI schedule of reinforcement, bees exhibit lower response rates than in an FR schedule. FR schedules of up to one reinforcement out of 30 trials, and FI schedules with up to 90 s between reinforcements are reached after several days of training. Additional operant learning phenomena in honeybees have been studied in the tradition of experimental psychology.

Visual Stimulus Categorization

Categorization refers to the classification of perceptual input into defined functional groups. Categorization has been studied in free-flying honeybees trained to different visual stimuli (e.g., pattern orientation or bilateral symmetry). In the tests, bees were presented with novel patterns which belonged to the same category with regard to pattern orientation or symmetry. In such transfer tests, bees chose the appropriate patterns despite the novelty of the structural details. Interestingly, bees trained for symmetry chose the novel symmetric stimuli more frequently and came closer to and hovered in front of them longer than bees trained for asymmetry did for the novel asymmetric stimuli. It was thus suggested that bees have a predisposition for learning and generalizing toward symmetry. Such a predisposition can either be innate (and could facilitate a better and faster learning about stimuli that are biologically relevant) or be based on the transfer of past experience from predominantly symmetric flowers in the field.

The cases of categorical learning studied so far (orientation and bilateral symmetry of patterns) have been interpreted to demonstrate that bees focus their attention on a single feature at a time (orientation or bilateral symmetry) to solve the problem. In fact, it has been repeatedly argued that due to limited cognitive capabilities, bees are able to focus only on a single isolated feature at a time and cannot therefore attain levels of stimulus classification such as configural categorization as exhibited by humans. Recently, it was shown that a further level may exist in honeybee visual categorization. In addition to focusing on a single feature, honeybees were shown to assemble different features to build a generic pattern representation, which could be used to respond appropriately to novel stimuli sharing such a basic layout. Honeybees trained with a series of complex patterns sharing a common layout comprising four edge orientations remembered these orientations simultaneously in their appropriate positions and transferred their response to novel stimuli that preserved the trained layout. Honeybees also transferred their responses to patterns with fewer correct orientations, depending on their match with the trained layout. These results show that honeybees extract regularities in their visual environment and establish correspondences among

correlated features such that they generate a large set of object descriptions from a finite set of elements. In a related study, it was reported that honeybees have the ability to group similar natural images together. The authors suggested that configurational properties of the figures, in which specific features such as circular symmetry, angular periodicity, bilateral symmetry, and the presence of a horizontal high-contrast edge, the horizon, were integrated, could help the bees to classify the different stimuli. Thus, we can conclude that honeybees perform visual stimulus categorization.

Rule Learning

Rule learning presupposes the positive transfer of an appropriate response from a known set to a novel set of stimuli. One example is the detection of sameness and of difference. These rules are uncovered through delayed-matching-to-sample (DMTS) and the delayed-nonmatching-to-sample (DNMTS) experiments, respectively. Animals are exposed to regularly changing samples and then to a set of stimuli, one of which is either identical to the sample (DMTS) or different (DNMTS) and which is then reinforced. Honeybees foraging in a Y maze learn both rules when trained to either pairs of colors or patterns. Bees trained with colors and presented in transfer tests with gratings that they have not experienced before solved the problem and chose the grating identical to the sample at the entrance to the maze. Similarly, bees trained with the gratings and tested with colors in transfer tests also solved the problem and chose the novel colors corresponding to those of the sample grating at the maze entrance. Bees also transferred between olfactory and visual stimuli. Bees' ability to solve DMTS tasks has recently been verified. It was found that the working memory for the sample underlying solving a DMTS task is around 5 s. Moreover, bees trained in a DMTS task can learn to pay attention to one of two different samples presented successively in a flight tunnel (either to the first or to the second) and can transfer the learning of this sequence weight to novel samples.

Another case of rule learning, the transitive inference problem, was tested, and the result was that bees cannot solve it. In this problem, animals have to learn a transitive rule, that is, $A > B$, $B > C$, then $A > C$. Preference for A over C in this context can be explained by two strategies: (1) deductive reasoning in which the experimental subjects construct and manipulate a unitary and linear representation of the implicit hierarchy, $A > B > C$; or (2) responding as a function of reinforced and nonreinforced experiences, in which case animals choose among stimuli based on their associative strengths, that is, on the effective number of reinforced and nonreinforced experiences with the stimuli. Bees were trained using five different visual stimuli A, B, C, D, and E in a multiple discrimination task A+ versus B−, B+ versus C−, C+ versus D−, D+ versus E−. Training therefore involved overlapping of adjacent premise pairs ($A > B$, $B > C$, $C > D$, $D > E$), with an underlying linear hierarchy $A > B > C > D > E$. After training, bees were tested with B versus D, a nonadjacent pair of stimuli that were never explicitly trained together. Honeybees learned the premise pairs as long as these were trained as uninterrupted, consecutive blocks of trials, but if shorter and interspersed blocks of trials were used, such that bees had to master all pairs practically simultaneously, performance collapsed and the bees did not learn the premise pairs. It was concluded that bees do not establish transitive inferences between stimuli but rather guide their choices by the joint action of a recency effect (preference for the last rewarded stimulus, D) and by an evaluation of the associative strength of the stimuli (in which case, the preference for B should be evident).

Conditional Discriminations: Occasion Setting and Contextual Learning

Contextual learning is a term widely used for describing conditional discrimination, which can be subsumed in several cases of the occasion-setting problem. In these conditions, a given stimulus, the occasion setter, informs the animal about the outcome of its choice. In one situation, bees were trained to fly to two feeding stations, one in the morning and one in the afternoon. Thus, depending on the time of day, bees flew in one direction or the other, thus showing that temporal information can act as the occasion setter for a sensory–motor routine. When trained bees were captured after arriving at the hive and then released at the wrong site (i.e., at the afternoon site in the morning or at the morning site in the afternoon), they flew straight back to the hive. To fly homeward, the bees retrieved the memory for the correct flight path from long-term memory stores that were established during training and that related local landmarks to the flight path to the hive.

Olfactory cues can also act as occasion setters for the choice of appropriate navigation vectors. It was shown that free-flying bees trained to forage at two distinct feeders, placed at different outdoor locations and carrying different scents or colors, recalled the locations (or colors) of these food sites and flew to them when the training scents were blown into the hive and the scents and food at the feeders were removed. These results indicate that familiar scents can trigger navigational and visual memories in experienced

bees to define appropriate context-dependent responses. These observations resemble earlier reports in which bees were successfully stimulated to dance at night. It was found that they indicated the afternoon feeding place before midnight and the morning feeding place after midnight. Most interestingly, around midnight bees performed either unoriented dances or indicated both directions in successive runs.

Choice Strategies

A bee foraging in a patch of four artificial flowers learns continuously about the reward conditions of each flower. Both excitatory and inhibitory learning contribute to memory formation, but they do so asymmetrically. The bees' choice behavior was modeled by applying the Rescorla–Wagner version of the difference rule in learning, and it was found a best fit if a positive learning rate of 0.8 and a negative learning rate of 0.08 were assumed. From these results, we would expect risk-prone behavior in honeybees in a situation of flowers (artificial feeders) with variable amounts of reward because the occasionally experienced higher rewards should lead to better learning than the equally frequently experienced lower rewards. However, risk-averse behavior was found first both in bees and bumblebees foraging on artificial flowers offering variable or constant rewards. These data were interpreted as showing that bees optimize the mean rate of reward per flower. In later studies, no indication of risk-averse behavior was found in honeybees confronted with variance in sucrose solution concentration or in bumblebees confronted with variance in either sucrose concentration or volume. These latter studies on risk sensitivity by bees were hampered by the fact that the variance in reward introduced by the bees' choice behavior itself did not take into account unrewarded immediate revisits to the just-depleted feeders. A follow-up study, in which this parameter was carefully monitored and in which variance of reward was the only parameter systematically changed, found that bees distributed their choices with respect to the amount of reward and variance but that their choice behavior did not change with the amount of variance, consistent with a risk-indifferent strategy. Therefore, bees ignore the variance in reward, not because they cannot detect and learn the differences in the amount of rewards but, rather, because they appear to eliminate the parameter reward variance when they choose a particular feeder and refer only to the mean amount of reward rate.

Response Conditioning

Classical conditioning of innate responses is a most convenient way to study the behavioral and neural mechanisms of associative learning. In hungry honeybees, a sucrose stimulus at the antennae reliably releases the extension of its proboscis (the PER). Conditioning to olfactory or mechanical stimuli as a conditioned stimuli (CS) is quite effective, but colors can also be used and were indeed the stimuli used first to study associative learning in bees harnessed in a tube. The associative nature of PER conditioning was established by demonstrating that only forward pairing of CS–unconditioned stimulus (US) sequences is effective, by various control groups (unpaired CS and US, CS- or US-only presentations), and by differential conditioning of the two olfactory stimuli. The predictive value of the CS depends on the reliability with which it is causally related to the US. In differential conditioning, the reversal to the initially unpaired stimulus is slower after more frequent unreinforced preexposures than after a lesser number of preexposures. The same applies for US-only preexposures in an otherwise reinforced context. Partial reinforcement schedules have little effect on the acquisition function because extinction trials do not alter the CR probability but, rather, increase the resistance to extinction and other measures of learning.

Overshadowing and blocking experiments are used to characterize the informational content of the CS. Blocking has been found in some experiments and not in others, depending on which odors were used and how the results were related to control groups. It was concluded that attentiveness is not a limiting factor and that the salience or associability of a stimulus is indeed an important factor. Second-order conditioning is a procedure that tests whether a CS can acquire the potential of an US. Although the earlier studies lacked control groups, recent studies clearly document second-order conditioning.

Nonelemental Forms of Conditioning

In studies of nonelemental learning, paradigms have been developed for which the associative strength of a specific event, stimulus or reaction, is ambiguous and therefore cannot be predictive for solving a problem. In the honeybee, several recent studies have addressed the issue of elemental versus nonelemental learning using visual training of free-flying animals or olfactory conditioning of harnessed animals. In both modalities, bees are able to solve a biconditional discrimination (AB+, CD+, AC−, BD−). This capacity demonstrates that both visual and olfactory compounds are learned under certain circumstances as entities different from the simple sum of their elements. Similarly, negative patterning experiments (A+, B+, AB−) have also been performed in both the visual and the olfactory modality. Such discrimination can be

explained only if the compound AB is treated as being different from the simple sum of its elements. The results are interpreted according to a modified unique cue hypothesis. A unique cue is generated when bees receive olfactory input to both brain sides because bilateral olfactory input is required to solve a negative patterning task. A form of context-dependent conditioning can be established when bees receive differential conditioning of A+, B− via one antenna, and A−, B+ via the other. Bees learn to respond appropriately to the rewarded odor and inhibit their response to the nonrewarded odor on each side. Side-specific input was used to test the contribution of the mushroom body (MB) to elemental and nonelemental forms of conditioning. Bees with unilateral lesions of the mushroom bodies generated by larval treatment with hydroxyurea were trained under different olfactory conditioning designs. When odorants were delivered in a side-specific manner, bees with unilateral mushroom body lesions could not solve an unambiguous double discrimination (paradigm 1: A+, B− on one antenna, C+, D− on the other; A+B− / C+D−), whereas they could solve at least one of both discriminations of an ambiguous problem (paradigm 2: A+, B− on one antenna, A−, B+ on the other; A+B− / A−B+). In the latter case, they solved the discrimination proposed to their intact brain side. Non-ablated bees could learn both side-specific discriminations. When odorants were delivered simultaneously to both antennae (paradigm 3: A+B-C+D−), ablated bees learned slower than normal bees. Thus, in all three cases, the unilateral loss of a median calyx affected olfactory learning. It was proposed that MBs are required for solving elemental olfactory tasks whose complexity is enhanced by virtue of the number of stimuli involved (paradigms 1 and 3: four stimuli) and that MB ablations could have an effect on the inhibition of information exchange between brain hemispheres.

Operant Conditioning of Restrained Bees

The honeybee antennae can be operantly conditioned to assume a certain posture. In a more reduced preparation, the activity of the fast flagellum flexor muscle is recorded extracellularly from the scapus of the antenna. Whenever the muscle activity exceeds a defined reward threshold, the animal is rewarded with a drop of sucrose solution. After several US deliveries, the frequency of the muscle potentials increases significantly over the spontaneous frequency. The conditioned changes of frequency can be observed for 30 min after conditioning and do not occur in yoked controls, which were rewarded independent of nerve activity.

Sequential Memory Processing

A stable, lifelong memory is formed after even only a few learning trials in free-flying bees. The retention curve after a single learning trial follows a biphasic time course, which indicates that memory and retrieval are at first very high, fade to a minimum 2–3 min after learning, rise again within the following 20 min, and stay high for a number of hours; this is true both in PER conditioning and for free-flying bees. The early memory phase in the minutes range is particularly susceptible to both extinction and reversal learning, whereas the consolidated memory is much more resistant. The sensitizing effect of the appetitive stimulus overlaps with the early high response level, indicating that, at least in PER conditioning, the nonassociative memory initiated by the US contributes substantially to the response probability immediately after the single learning trial. Amnestic treatments (cooling, narcosis, and weak electroconvulsive brain stimulation), if applied within 2–3 min after the single-trial learning, erase the memory trace, indicating a susceptible short-term memory (STM). The memory initiated by a single conditioning trial consolidates within a few minutes so that it becomes resistant to retrograde amnestic treatments. The time course of retrograde amnestic sensitivity is independent of the kind of learning (operant color learning or olfactory PER conditioning), but depends on the particular structure of the brain where the treatment is applied. Several conditioning trials within STM make the memory trace immune to amnestic treatment.

Several forms of longer-lasting memory have been discerned on the basis of behavioral, biochemical, and pharmacological experiments performed with olfactory PER conditioning. The respective time courses and transitions between five forms of memory have been related to the temporal sequence during the foraging cycle:

1. Associative induction and early short-term memory (eSTM). eSTM covers the time window during which a foraging bee can expect to be exposed to the same stimuli because flowers most frequently occur in patches. No specific choices need to be performed at this time, and general arousal (depending on the strength of the US) will suffice to control whether the animal stays in the patch or postpones choices for a later time.

2. Late short-term memory (lSTM). The transition to the selective associative memory trace during lSTM is a rather slow process after a single learning trial lasting up to several minutes, and it happens very quickly after multiple learning trials. Thus, consolidation is both

time- and event-dependent, where events have to be associative experiences and not just CS or US repetitions. The behavioral relevance of these findings for foraging behavior has been related to the temporal separation between intra- and interpatch visits.

3. Mid-term memory (MTM). At the beginning of MTM, behavior is controlled by consolidated, highly specific memory. At this stage, memory is more resistant to extinction, conflicting information, and elapsing time, and some information about the context dependencies may have already been stored. Under natural conditions, bees have usually returned to the hive and departed on a new foraging bout within the time window of MTM.

4–5. Long-term memory (LTM), early and late. LTM requires multiple learning trials, indicating that specific information which can be extracted only from multiple experiences (signal reliability and context dependence) controls the transfer to LTM. Structural changes in the connectivity between neurons have been proposed as the substrates for LTM in vertebrates and invertebrates. In bees, direct evidence for LTM-related structural changes are lacking, but measurements of the volume of the subcompartments of the MBs indicated that more experienced bees have bigger dendrite volumes and more elaborate dendritic branches. The biological circumstances of two forms of LTM may be related to the distinction between those forms of learning which usually lead to lifelong memories (e.g., visual and olfactory cues characterizing the home colony) and those which are stable but need updating on a regular basis (e.g., visual and olfactory cues of feeding places).

See also: Communication in the Honeybee; Conditioning: Theories; Conditioning: Simple Neural Circuits in the Honey Bee; Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Olfaction in Invertebrates: Honeybee; Operant Conditioning of Reflexes; Procedural Learning: Classical Conditioning.

Further Reading

Bitterman ME, Menzel R, Fietz A, and Schäfer S (1983) Classical conditioning of proboscis extension in honeybees (*Apis mellifera*). *Journal of Comparative Psychology* 97: 107–119.

Giurfa M, Eichmann B, and Menzel R (1996) Symmetry perception in an insect. *Nature* 382: 458–461.

Giurfa M, Zhang SW, Jenett A, Menzel R, and Srinivasan MV (2001) The concepts of "sameness" and "difference" in an insect. *Nature* 410: 930–933.

Greggers U and Menzel R (1993) Memory dynamics and foraging strategies of honeybees. *Behavioral Ecology and Sociobiology* 32: 17–29.

Lachnit H, Giurfa M, and Menzel R (2004) Odor processing in honeybees: Is the whole equal to, more than, or different from the sum of its parts? In: Slater PJG, Rosenblatt JS, Roper TJ, Snowdon CT, Brockmann HJ, and Naguib M (eds.) *Advances in the Study of Behavior,* vol. 34, pp. 241–264. Amsterdam: Elsevier.

Menzel R (1990) Learning, memory, and 'cognition' in honey bees. In: Kesner RP and Olton DS (eds.) *Neurobiology of Comparative Cognition*, pp. 237–292. Hillsdale, NJ: Lawrence Erlbaum.

Menzel R (1999) Memory dynamics in the honeybee. *Journal of Comparative Physiology A* 185: 323–340.

Stach S, Benard J, and Giurfa M (2004) Local-feature assembling in visual pattern recognition and generalization in honeybees. *Nature* 429: 758–761.

Zhang SW, Srinivasan MV, Zhu H, and Wong J (2004) Grouping of visual objects by honeybees. *Journal of Experimental Biology* 207: 3289–3298.

Learning and Memory in Invertebrates: *Limax*

A Gelperin, Monell Chemical Senses Center, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

> Perhaps these insects are little machines in a deep sleep, but looking at their rigidly armored bodies, their staring eyes, and their mute performances, one cannot help at times wondering if there is anyone inside. (Dethier, 1964)

This sentence concludes a remarkable paper published in 1964 by Vincent Dethier, in which he considers the evidence that the cognitive abilities of insects are equivalent to those of mammals. Since that time the comparative study of cognition has expanded considerably, particularly in the area of linguistics, tool use and social organization. For example, arousal and cognition are now studied in flies and bees. Reprinted with permission from Dethier VG (1964) Microscopic brains. *Science* 143: 1138–1145. Copyright [1964] [AAAS].

Contemporary students of molluskan behavior in general have not used concepts from comparative cognition to describe even complex and plastic behavioral responses. Students of learning and memory in octopus, a mollusk with brain regions equivalent in complexity to the mammalian cerebellum and with extremely flexible behavioral strategies, have not yet ventured into the realm of comparative cognition to find a framework for understanding the behavior of this remarkable animal. By contrast, our studies of learning and memory in *Limax maximus* (Limax) have been guided by the notion that, within the realm of odor learning, whatever a rodent can do, Limax can do. We have found this approach very productive in revealing the complexity of Limax odor learning, as detailed below.

Although mollusks and mammals diverged from their common ancestor some 600 million years ago, there is ample evidence for conserved genes, cellular signaling processes, mechanisms of neural integration, mechanisms of learning, and forms of learning used to solve analogous problems between the two taxa. These observations suggest that prospecting for mammalian memory mechanisms in the molluscan nervous system is a productive strategy, as amply demonstrated by dissection of short- and long-term memory storage mechanisms in *Aplysia*.

Commonalities between mollusks and mammals have emerged in a particularly clear way from analyses of the molecular, cellular, and network dynamics associated with associative learning in Limax, specifically in the realm of olfactory information processing and odor learning. Vertebrate and invertebrate olfactory processing systems share a number of basic design features as do mammals and mollusks. Within the gastropod mollusks, exemplified by the land snails and slugs, a common plan of olfactory system organization is evident as is a highly developed capacity for olfactory learning. The following will outline some of the universal design principles displayed by the olfactory information processing system of Limax.

Higher-Order Conditioning

In designing experiments to probe the complexity of Limax odor learning, it has proved to be remarkably productive to regard Limax as possessed of mammalian learning skills in the domain of associative learning about odorants as conditioned stimuli and aversive tastes as unconditioned stimuli. In particular, the Kamin blocking effect was an early striking example of higher-order conditioning shown robustly in Limax odorant conditioning. We have recently revisited the question as to how the blocking phenomenon, which requires that conditioned stimuli acquire predictive relationships with unconditioned stimuli, can be implemented logically and with plausible neurophysiological mechanisms, a subject of continuing interest to students of the mammalian and insect central nervous system (CNS).

The blocking phenomenon involves two phases of conditioning. In the first phase of conditioning a conditioned stimulus (CS1) is paired with the unconditioned stimulus (US). In the second phase of conditioning a compound stimulus consisting of CS1 and a second conditioned stimulus (CS2) is presented followed by the US. Blocking is evidenced by lack of association between CS2 and the US, as found in other invertebrates and mammals.

Our circuit solution to the blocking problem in Limax requires a facilitator neuron (FN) interposed between sensory inputs from CS1, CS2, and US on the one hand and the motorneurons executing the conditioned response. It also requires reciprocal inhibition between the CS1, CS2, and US inputs, which can be amplified by learning via activation of the FN. These interactions were explored in simulations based on an integrate-and-fire model used previously for spiking cortical neurons and synaptic connections that can show potentiation. There are several cellular loci in Limax that can potentially provide tests of the predictions of the proposed mechanism for blocking, including the serotonergic metacerebral giant cell, the Limax feeding command neurons, and the recently identified odor-responsive parietal neuron v-PN. The

recent demonstration of odor conditioning by the isolated Limax CNS greatly augments the feasibility of these experiments.

Serotonergic modulation of sensory neurons is a central element in plasticity of the *Aplysia* gill- and siphon-withdrawl circuit while depletion of serotonin impairs short-term but not long-term memory in Limax. While slugs will quickly learn to avoid a diet devoid of the essential amino acid tryptophan, if a tryptophan-deficient diet is the only food available, slugs will maintain normal levels of brain serotonin and function of the serotonergic metacerebral cell in spite of this dietary challenge. In mammals, cells in the anterior piriform cortex are sensitive to deficiencies in essential amino acids, based on uncharged transfer RNA inducing phosphorylation of eukaryotic initiation factor 2 via a nonderepressing 2 kinase. These findings in the mammalian system provide guidance to the search for neurons in the Limax CNS responding to diets devoid of essential amino acids and directing selective food avoidance.

Flexible Storage of Compound Odor Stimuli

The issue of how compound odor stimuli are stored in the CNS can be explored by training animals to associate an aversive US with a compound CS under conditions where they experience the compound of CS1 and CS2 as a spatially and temporally uniform mixture or as a spatially and temporally heterogeneous mixture. The heterogeneous mixture of CS1 and CS2 provides information on the components of the mixture as well as the mixture itself. The uniform mixture is more likely to be treated as a single odor object. A difference in storage mechanism is revealed by measuring the aversion to CS1 and CS2 presented alone after aversive compound conditioning to the mixture. Limax appears to store a uniform mixture of CS1 and CS2 as a unique odor object such that strong aversive responses are shown to the CS1 + CS2 compound but clear attraction to CS1 alone and CS2 alone remains. Conversely, pairing the US with a heterogeneous mixture of CS1 and CS2 leads to a different form of sensory storage because after conditioning slugs show strong aversion to CS1 presented alone and CS2 presented alone as well as strong aversion to the compound stimulus. Similar results were obtained using cooling-induced retrograde amnesia to probe the nature of CS storage after compound conditioning. Using cooling to alter stimulus representations, slugs appeared to learn a binary odor mixture as a unitary object when they had no exposure to the separate elements of the mixture. These experiments emphasize the synthetic role of olfactory processing and the critical role of learning in stimulus processing.

Evolution of the Form of the Memory Trace

Mollusks and mammals also share the use of oscillatory dynamics in early olfactory centers (procerebral (PC) lobe in terrestrial mollusks and olfactory bulb (OB) in mammals). The rosy wolfsnail, *Euglandina rosea,* uses chemosensation for hunting its prey and has a PC lobe with oscillating local field potentials (LFPs), as in Limax and *Helix*. Hunting dogs use olfaction for trail following and use active odor sampling, or sniffing as an essential component of their odor sampling strategy. Analogously, snails show a variety of 'rapid' tentacle (nose) movements associated with odor localization that may serve some of the same functions as sniffing. Sniffing may be associated with taking olfactory snapshots of odorants inputs when rapid sampling and decision making are required.

Providing Motivation for Learning Odor Associations

We have found that water deprivation is a potent motivational method for obtaining conditioned odor aversions in Limax. Slugs are weighed and then dehydrated by placing them in mesh cages in a container with dessicant and a fan (Figure 1). Dehydration continues until the slugs lose 35–40% of their initial fully hydrated body weight. Rapid dehydration produces maximally motivated slugs. Dehydrated slugs actively search for a source of water for contact rehydration, which occurs when the slug sits on a moist surface and absorbs water through its foot via a paracellular pathway. During training, dehydrated slugs are first offered saltwater containing quinine and NaCl for rehydration, paired with odor A. Contact of the slug with the saltwater elicits rapid withdrawal. Slugs are then offered safe pure water paired with odor B. Odor sources are placed under the mesh floor of the training chamber with the water source located directly above the odor source. The slug is given 30 min to rehydrate on the moist surface above odor B. The next day, slugs are dehydrated again and positioned between sources of odor A and odor B located under the mesh floor of the testing chamber. Latency to approach an odor source and time spent within a 9 cm circle centered on each odor source are recorded. In three experiments with six different odorants in which 31 slugs were trained with odor A signaling safe water and odor B signaling saltwater, 90% of trained slugs chose odor A. Since slugs can be

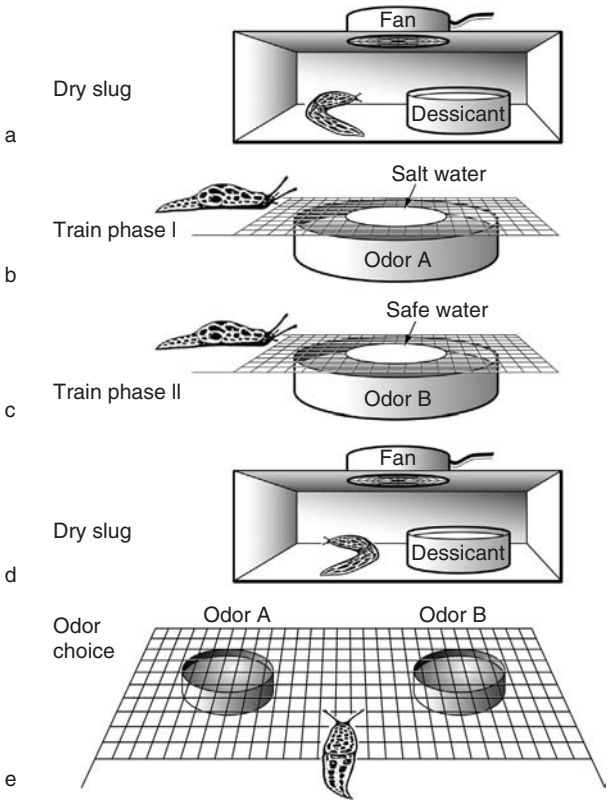

Figure 1 Method for appetitive conditioning of odor preferences with water reward: (a) slug is dehydrated to 60–65% of its fully hydrated weight within 30 min; (b) slug is offered dangerous water with odor A; (c) slug is offered safe water with odor B; (d) slug is again dehydrated; (e) dehydrated slug is offered a choice between spending time over odor A or odor B.

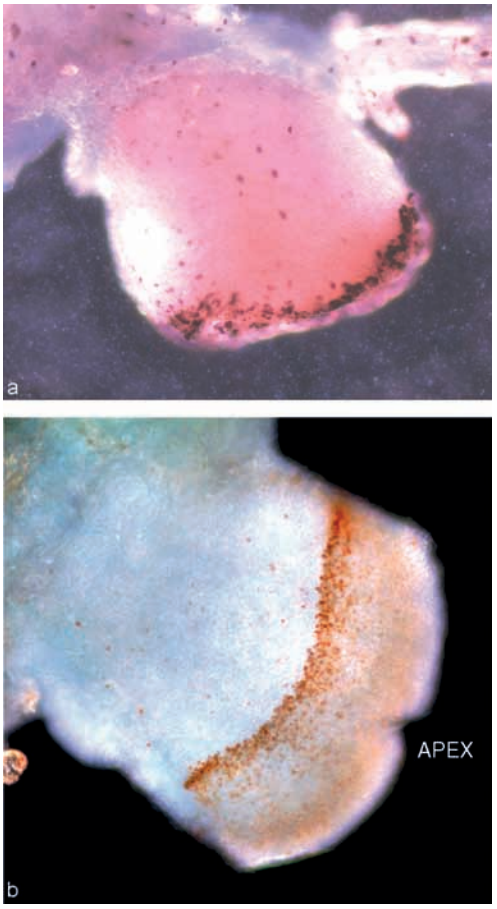

Figure 2 Zones of neurogenesis in the procerebral lobe are revealed by BrdU labeling: (a) PC lobe of a slug exposed to BrdU at 3 days posthatching and BrdU staining developed 1 day later; (b) PC lobe of a slug exposed to BrdU at 3 days posthatching and BrdU staining developed 2 months later. The apical region of the PC lobe is a site of continued neurogenesis over this period.

dehydrated in 40–60 min and survive repeated bouts of dehydration after contact rehydration, the method of water–odor pairing during contact rehydration is much more convenient than food deprivation, which can require many days to re-establish appetitive odor responses after feeding to satiation. It remains a mystery as to how an animal with a completely moist surface can show directed orientation to a localized source of water, as dehydrated slugs do.

Neurogenesis in Olfactory Pathways

Another design principle of olfactory information processing shared between mammals and mollusks is the addition of new circuit elements posthatching or postparturition to both the sensory periphery and the central processing structures – PC lobe and OB. To investigate the locus and extent of neurogenesis in the Limax PC lobe, newly hatched slugs (30–40 mg) were immersed in a saline solution containing bromodeoxyuridine (BrdU) for 90 min. After BrdU treatment slugs were rinsed with Limax saline and placed in containers with a moist substrate and their maintenance diet, rat chow. At various periods after BrdU treatment brains were harvested and processed as whole mounts for immunocytochemical demonstration of BrdU incorporation into DNA of dividing neurons. The CNS was processed with a protocol supplied with the BrdU *in situ* detection kit. Briefly, tissue was fixed, treated to open DNA strands, and then treated with biotinylated anti-BrdU antibody and streptavidin horseradish, peroxidase (HRP) secondary antibody. Preparations were developed with diaminobenzidine as the chromogen and mounted in glycergel prior to digital photography. Figure 2(a) shows the PC lobe from a slug given a BrdU bath 3 days posthatching and the PC lobe harvested 1 day later and developed to reveal BrdU labeled neurons. The zone of neurogenesis is clearly visible at the apical pole of the PC lobe, which is also the site of initiation of activity waves that traverse the apical–basal axis at 0.7 Hz in isolated preparations.

Figure 2(b) shows the band of neurogenesis revealed in the PC lobe treated to reveal neuronal BrdU incorporation 2 months after BrdU exposure of the hatchling slug. As the apical neurons in the zone of neurogenesis shown in Figure 2(a) continue to divide, the more apical of the daughter neurons continue to divide so that the original BrdU labeled neurons are left behind and the apical region of the PC lobe extends. Since this active neurogenesis occurs during natural olfactory stimulation, it is possible that the diversity and intensity of olfactory experience, and perhaps odor learning, modulate the rate and extent of neurogenesis in the PC lobe or neuronal survival. Enhanced survival of new neurons in the mammalian olfactory bulb due to enriched odor stimulation has been documented, with a concomitant enhancement of short-term memory.

Procerebrum as an Olfactory Learning Center

The procerebrum or PC lobe is a division of the cerebral ganglion unique to terrestrial slugs and snails which is specialized for the processing of olfactory information. The PC lobe contains several tens of thousands of small (7–9 μm soma diameter) local interneurons comprised of at least two subpopulations, bursting (B) and nonbursting (NB) neurons. B neurons are a few percent of the total neuronal population and are coupled with each other by chemical and electrical synapses. B neurons have extensive synaptic connections with NB neurons. Stimulation of the superior tentacle with odorants results in inhibitory responses in 45% of NB neurons, while 11% of NB neurons show an excitatory response. The periodic membrane potential trajectories of B and NB cells lead to a phase-dependent response to sensory inputs. Both B and NB cells contain the mRNA for nitric oxide synthase and soluble guanylate cyclase, critical to the control of oscillation frequency in the PC lobe (see below). The B cells located in the apical region of the PC lobe are more excitable (shorter interburst interval) than the B cells located in the basal region of the PC lobe. Chloride-dependent periodic depolarizations in apical B cells have a higher amplitude than the chloride-dependent depolarizations recorded from B cells in the basal region of the PC lobe. This gradient of excitability and pattern of connectivity give rise to activity waves in the PC lobe traveling from apex to base at 1.1 mm s^{-1} at a repetition rate of 0.7 Hz *in vitro*. These traveling waves have also been found in mammalian cortex although their computational role is still unclear. Evidence for traveling waves in the PC lobe of intact behaving slugs is suggestive but not definitive, based on *in vivo* recordings of neuronal dynamics in the PC lobes of intact, awake, behaving slugs. There are direct interactions between the oscillatory circuitry in the PC lobe and motor neurons controlling nose positioning and between PC lobe circuitry and remote neurons in the pedal and buccal ganglia. A behavioral interaction between olfactory inputs and feeding motor responses has been documented. Other regions of the cerebral ganglion adjacent to the PC lobe may also show oscillatory local field potentials.

Increasing evidence indicates that the PC lobe is involved in odor learning in Limax. Optical studies of odor-elicited activity in the PC lobe showed changes due to prior conditioning of responses to an odorant. Wavelet-based analysis may help to characterize more completely the complex dynamics of local field potential activity as modified by odor input in the PC lobe. A small set of genes is specifically activated in PC lobe neurons due to odor learning. Antibodies to the gene product for one of the learning-activated genes have been made and used to show the enhanced levels of expression of the gene after odor learning.

A direct test of the role of the PC lobe in odor learning in Limax is based on studies of the effects of lesions of the PC lobe on subsequent odor learning. Ablation of the PC lobe 7 days prior to odor conditioning produced clear deficits in odor conditioning not evident in slugs given control operations. PC lobe ablation after odor conditioning also reduced avoidance of the aversively conditioned odor. Control experiments showed that PC lesioned slugs were not anosmic. These results strongly support evidence from optical, molecular, and activity-dependent dye uptake studies also indicating a selective role of the PC lobe in odor learning.

Other genes have been identified that are activated by noxious stimuli like those used as the US in odor conditioning. One kruppel-like factor (KLF) family transcription factor gene (Limax KLF; limKLF) is upregulated 30 min after depolarization of the isolated CNS with high potassium saline. An immediate early gene, CCAAT enhancer binding protein (C/EBP) was cloned by reverse transcription-polymerase chain reaction (RT-PCR). C/EBP was also activated by noxious electrical stimuli applied to the Limax body wall. It will be interesting to see if the aversive taste of quinine, used as the US in many odor conditioning experiments with Limax, also activates C/EBP.

The tentacular ganglion located just behind the nose in the superior tentacles also shows oscillatory local field potentials. Odor-elicited uptake of a fluorescent derivative of glucose has been demonstrated in the tentacular ganglion. The oscillatory activity of the tentacular ganglion is modulated by

ACh, 5HT, and GABA as identified immunocytochemically. Odor-elicited spike activity can be recorded in the tentacular nerve and correlated with oscillations in LFP. Recent work shows that the summed receptor potentials of the olfactory receptors, recorded as the electro-olfactogram (EOG), can also show oscillations, as previously found is several vertebrate olfactory epithelia. How the interactions of the peripheral EOG oscillation and the central LFP oscillation enhance odorant detection remains to be determined.

A variety of other cells and neurotransmitters have been shown to play some role in odor aversion learning by Limax. The v-PN neuron was first identified as an output neuron whose activity could be used to provide an index of aversive evaluation of odor input. Monitoring activity in v-PN was crucial to the demonstration of *in vitro* odor learning by the isolated Limax CNS. Among the 21 neurotransmitters known to play a role in Limax odor processing, serotonin has been shown to play an essential role in odor learning. The peptide FMRFamide may play a regulatory role in stimulating NO production by the neuronal nitric oxide synthase (nNOS) containing neurons in the PC lobe. Odor discrimination can be blocked by pharmacological block of nNOS. Nitric oxide plays a major role in controlling network dynamics in the Limax olfactory system both centrally and peripherally. NO is also a modulator of chemosensory processing in *Aplysia*. Serotonin and NO have complementary effects on elements of the olfactory processing network in Limax.

See also: Conditioning: Theories; Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Olfactory System: Circuit Dynamics and Neural Coding in the Locust; Operant Conditioning of Reflexes; Procedural Learning: Classical Conditioning.

Further Reading

Cooke IRC and Gelperin A (2001) *In vivo* recordings of spontaneous and odor-modulated dynamics in the Limax olfactory lobe. *Journal of Neurobiology* 46: 126–141.

Delaney K and Gelperin A (1986) Post-ingestive food-aversion learning to amino acid deficient diets by the terrestrial slug *Limax maximus*. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 159: 281–295.

Dethier VG (1964) Microscopic brains. *Science* 143: 1138–1145.

Ermentrout B, Wang JW, Flores J, and Gelperin A (2004) Model for transition from waves to synchrony in the olfactory lobe of Limax. *Journal of Computational Neuroscience* 17: 365–383.

Fujie S, Yamamoto T, Murakami J, et al. (2005) Nitric oxide synthase and soluble guanylyl cyclase underlying the modulation of electrical oscillations in a central olfactory organ. *Journal of Neurobiology* 62: 14–30.

Fukunaga S, Matsuo R, Hoshino S, and Kirino Y (2006) Novel kruppel-like factor is induced by neuronal activity and by sensory input in the central nervous system of the terrestrial slug Limax valentianus. *Journal of Neurobiology* 66: 169–181.

Gelperin A (1999) Oscillatory dynamics and information processing in olfactory systems. *The Journal of Experimental Biology* 202: 1855–1864.

Gelperin A (2006) Olfactory computations and network oscillations. *Journal of Neuroscience* 26: 1663–1668.

Goel P and Gelperin A (2006) A neuronal network for the logic of Limax learning. *Journal of Computational Neuroscience* 21: 259–270.

Inoue T, Murakami M, Watanabe S, Inokuma Y, and Kirino Y (2006) *In vitro* odor-aversion conditioning in a terrestrial mollusk. *Journal of Neurophysiology* 95: 3898–3903.

Ito I, Watanabe S, and Kirino Y (2006) Air movement evokes electro-olfactogram oscillations in the olfactory epithelium and modulates olfactory processing in a slug. *Journal of Neurophysiology* 96: 1939–1948.

Kasai Y, Watanabe S, Kirino Y, and Matsuo R (2006) The procerebrum is necessary for odor-aversion learning in the terrestrial slug Limax valentianus. *Learning and Memory* 13: 482–488.

Prior DJ (1985) Water-regulatory behaviour in terrestrial gastropods. *Biological Reviews of the Cambridge Philosophical Society* 60: 403–424.

Shirahata T, Tsunoda M, Santa T, Kirino Y, and Watanabe S (2006) Depletion of serotonin selectively impairs short-term memory without affecting long-term memory in odor learning in the terrestrial slug Limax valentianus. *Learning and Memory* 13: 267–270.

Wang JW, Denk W, Flores J, and Gelperin A (2001) Initiation and propagation of calcium-dependent action potentials in a coupled network of olfactory interneurons. *Journal of Neurophysiology* 85: 977–985.

Yasui K, Matsuo R, and Kirino Y (2004) Onset of amnesia is delayed with a decrease in inhibition of protein synthesis during odor-taste associative learning in the terrestrial slug Limax valentianus. *Zoological Science* 21: 1163–1166.

Learning and Memory in Invertebrates: Mollusks

C M Sherff and T J Carew, University of California, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Over the past several decades, it has become evident that fundamental mechanisms of learning and memory are remarkably conserved among animals, ranging from humans and other higher mammals to lower vertebrates and invertebrates, such as fruit flies and mollusks. Molluskan species in particular have proven quite useful for cellular and molecular studies; they exhibit several types of learning, including nonassociative forms (sensitization, habituation, etc.) and associative forms (classical conditioning and operant conditioning). Because of the relative simplicity of their nervous systems, mollusks provide powerful models for examining the mechanisms of the induction and maintenance of memory. One of the main strengths of the molluskan systems is the ability to make direct links between the animal's behavioral plasticity and its underlying cellular and molecular mechanisms. Because of the space limitations of this review, we are unable to discuss the contributions of the numerous molluskan model systems (such as the terrestrial snail, *Helix*; the marine mollusk, *Tritonia*; and the terrestrial snail, *Limax*), each of which exhibits similar forms of learning to those we will describe; nor will we discuss the cephalopods (octopus, squid, and cuttlefish), which have complex nervous systems and are capable of more advanced forms of training, including spatial learning and observational learning. Rather, we will highlight three molluskan species that have made significant contributions to our understanding of the cellular and molecular mechanisms of learning and memory: the marine mollusk, *Aplysia californica*; the nudibranch, *Hermissenda crassicornis*; and the pond snail, *Lymnaea stagnalis*.

Aplysia

Of the molluks, *Aplysia* is perhaps the most extensively studied in the field of learning and memory. Because of the relative simplicity of its nervous system and the animal's large size, *Aplysia* is particularly amenable to combined behavioral, circuit, and cellular analyses. *Aplysia* exhibits several forms of memory, including both associative and nonassociative forms. Two behavioral systems have been the main focus of research in *Aplysia*: (1) the defensive withdrawal reflexes of the tail, siphon, and gill and (2) the feeding system. Multiple forms of memory have been studied in each. The withdrawal reflexes, which have been the most extensively studied, exhibit nonassociative forms of memory, including sensitization and habituation, as well as classical conditioning. The feeding system can also be conditioned using both operant and classical procedures. Very recently, it has also been shown that the gill reflex can also be operantly conditioned. The fortuitous finding that multiple forms of memory can be induced by a single simple circuit offers a unique opportunity to compare the cellular and molecular mechanisms that mediate different forms of memory in the same system.

Sensitization

The defensive withdrawal reflexes of the *Aplysia* tail, siphon, and gill show both sensitization, an increase in behavioral responsiveness following a noxious stimulus, and habituation, a decrement in a behavioral responsiveness following repeated, weak stimulation. In response to a weak tactile stimulation to the body wall or tail, *Aplysia* reflexively withdraws its gill, siphon, and tail. A strong, noxious stimulation, such as shock (e.g., to the tail or head), sensitizes these reflexes, producing an increase in their magnitude and duration in response to subsequent weak, tactile stimuli. Conversely, weak low-frequency stimulation habituates these reflexes. Although important research on habituation is currently being conducted, we will restrict the discussion here to sensitization because its mechanisms are better understood.

The experimental preparation often used is the tail-elicited siphon withdrawal (T-SWR). The amount and pattern of training trials is important in determining the type and the duration of the memory induced in many systems, ranging from insects to mammals. The efficacy of several training patterns has been explored in *Aplysia*. For example, a single tail shock induces short-term memory (STM) for sensitization, which lasts <30 min. Multiple (four or five) trials spaced by 15 min induces two long-lasting forms of memory: intermediate-term memory (ITM) that lasts 90 min–3 h, and long-term memory (LTM) that begins to be expressed about 10 h after training and lasts at least 2 days. Massed training, in which the same total amount of training is given but without any intervening intertrial intervals (ITIs), does not induce LTM, and induces only an early decaying form of ITM. Finally, multiple trials spaced over several days (one trial per day for 4 days) induce LTM that lasts several weeks.

Another training parameter that affects memory induction is neuronal activity. In the sensitization experiments described above, one site on the tail is used to test the reflex, and shock is applied to another site; thus the sensory neurons (SNs) at the test site do not fire during the tail shock. This form of sensitization is termed 'repeated-trial' (RT) sensitization. However, the shock site itself also shows sensitization, and this 'site-specific' sensitization differs from RT sensitization in that ITM induction requires only a single shock.

Related to the relationship between trial number and pattern is the question of how a prior learning experience influences later training. Recently, two forms of latent learning have been described in the T-SWR. Latent learning is a phenomenon in which learning-related changes occur, but these changes are subthreshold for detection by memory tests. The presence of latent learning is typically revealed as a facilitation of subsequent memory formation, often called savings. In one study, LTM was induced and then allowed to decay. After the T-SWR returned to baseline (the memory was 'forgotten'), animals were given stimulation that typically produces no learning. In previously trained animals, this stimulus pattern induced both ITM and LTM. In the second study, LTM was induced ipsilaterally, after which, animals were given a brief shock that is below threshold for induction of sensitization in otherwise untrained animals. This stimulus, given to the contralateral side, induced sensitization in previously trained animals.

Cellular and molecular mechanisms A major focus of recent research in *Aplysia* has been to relate the pattern of sensitization training to the molecular mechanisms involved in the induction and expression of memory. These experiments have been carried out using *in vitro* and semi-intact preparations. RT memory and site-specific memory have different requirements for protein kinase A (PKA) and protein kinase C (PKC) as well as different requirements for translation and transcription. RT-ITM is dependent on PKA and requires protein synthesis, and RT-LTM requires both translation and transcription, plus mitogen-activated protein kinase (MAPK) activation and tyrosine kinase activity. Site-specific ITM differs from RT-ITM in that it is dependent on PKC (PKM) and does not require protein synthesis. However, like RT-ITM, it requires MAPK activation.

A more thorough investigation of the molecular mechanisms of ITM and LTM has been achieved using a cellular model of sensitization, facilitation of tail SN–motor neuron (MN) synapses. Exogenous application of serotonin (5-HT), a neurotransmitter released by tail shock or tail nerve shock induces facilitation of SN–MN synapses. A single pulse of 5-HT induces short-term facilitation (STF) that lasts <30 min and requires activation of PKA (or PKC if the synapse has been previously depressed). Five pulses of 5-HT induce both intermediate-term facilitation (ITF), that lasts about 90 min and requires synthesis of new protein, and long-term facilitation (LTF) that lasts >24 h and requires both translation and gene transcription. Morphological changes in SN structure, including an increase in the number of varicosities, accompany LTF induction and are required for the maintenance of LTF from 24 to 72 h. The mechanisms of LTF have been widely studied and are described in depth elsewhere. Briefly, activation of PKA activates MAPK, ultimately resulting in cyclic AMP-responsive element binding protein (CREB)-mediated transcription that synthesizes immediate early genes (IEGs). Examples of IEGs that are important for the induction of LTF include CCAAT enhancer-binding protein (C/EBP) and ubiquitin hydrolase (Ap-Uch). Epigenetic mechanisms such as direct modification of DNA or histones have recently been found to be important for long-lasting synaptic modifications, both in vertebrate and in invertebrate models of learning and memory. In *Aplysia*, increasing histone acetylation facilitates the induction of LTF. Finally, secreted factors also appear to have a role in the induction of LTF. Both transforming growth factor-beta (TGF-β) and a TrkB ligand are involved, as well as an SN-specific peptide, sensorin. Not much is known about ITF, but this midrange form of memory seems to utilize some of the same players as LTF. For example, induction of ITF also requires both PKA and MAPK activation.

Synaptic correlates have also been found for site-specific memory. In site-specific memory, the test site is activated during training; thus, the cellular correlate for site-specific memory involves pairing a pulse of 5-HT with direct SN activation with current injection. This correlate is called activity-dependent (AD) facilitation, and a single pairing induces both ITF and LTF. As with the two different forms of memory, the two corresponding forms of facilitation have different molecular requirements. AD-ITF requires PKC, not PKA, for induction, and it does not require protein synthesis.

Associative Forms of Learning

Associative forms of memory include both classical, or Pavlovian, conditioning and operant conditioning. Classical conditioning is a form of learning in which a nominally neutral stimulus (conditioning stimulus, CS), when paired with a reinforcing stimulus (unconditioned stimulus, US) that causes an unconditioned reflexive response, gains the ability to cause the

response itself. When the CS and US are paired, the animal learns to associate the CS with the US such that, after training, the animal then responds to the CS alone. Operant conditioning is similar except that the animal learns to associate its own behavior with a reinforcer. Both classical and operant conditioning can be induced in the feeding system in *Aplysia*. The fact that both of these forms of learning impact the same circuit presents an opportunity to compare the cellular and molecular mechanisms underlying them.

Classical Conditioning

Classical conditioning in *Aplysia* was first described for the T-SWR, and its mechanisms were found to involve AD synaptic facilitation. More recently, classical conditioning has been investigated in the feeding system, in which tactile stimulation of the lips (CS) is paired with ingestion of food (US). Pairing of the tactile stimulation with seaweed leads to an increase in the number of bites the *Aplysia* makes in response to presentation of the CS alone. Ten trials spaced by 4 min induces memory in the form of an increase in the number of bites measured 1 h and 24 h after training.

Cellular mechanisms The cellular mechanisms of conditioning can be investigated in *in vitro* preparations in which the CS and US are substituted by nerve stimulation: stimulation of the antennal nerve, AT4 substitutes for the CS, and stimulation of a branch of the esophageal nerve, En2, substitutes for the US. The responses of central pattern generator (CPG) neurons to the CS are altered after paired training: B4/5 (increased number of ingestion-like buccal motor patterns (BMPs)), B31/32 (increase in the peak amplitude and net size of complex excitatory postsynaptic potentials (EPSPs)), and CB1–2 (increase in the number of action potentials).

Operant Conditioning

Operant conditioning in *Aplysia* was first described in the head-waving system. Three other forms of operant conditioning have since been described in *Aplysia*: aversive and appetitive feeding responses, and more recently, it has been found that the gill-withdrawal reflex can also be operantly conditioned. For appetitive operant conditioning, seaweed is presented to *Aplysia*, either with forceps or in a plastic net with a hole in it. The animal bites the food, pulls it into its buccal cavity, and swallows it. Since swallowing is the reinforcing stimulus, food is provided to the animal as long as it continues to feed.

In aversive conditioning, no holes are made in the net, so the food cannot be swallowed. After several failed swallowing attempts, the food is rejected. Over the training period, the animal decreases the frequency of its swallowing attempts. A single 5 min training session induces very short-term memory that lasts only 15 min and LTM that is expressed 12–24 h after training, a longer training session that lasts until cessation of responsiveness induces STM that lasts 30–60 min, and, finally, multiple training sessions (3×5 min, 30 min ITI) induce ITM that lasts up to 4 h. These results demonstrate that different phases of memory can be induced independently as one 5 min training session induces very short-term memory and LTF in the absence of STF and ITF. A requirement for protein synthesis for the induction of LTF is suggested by the finding that of the four described memory phases, only LTM was blocked by cooling the animal for 15 min immediately after training.

Aplysia has also been a useful model system in which to examine the modulatory effects of circadian rhythms on induction of memory. *Aplysia californica*, a diurnal species, shows greater memory for operant conditioning (as well as sensitization) when trained and tested during the day as opposed to at night; conversely, the nocturnal species, *Aplysia fasciata* exhibited improved memory when trained and tested at night.

Cellular mechanisms *In vitro* operant conditioning is induced by En2 stimulation (reinforcer) given, contingent upon spontaneous production of either ingestion-like behavior (appetitive conditioning) or egestion-like behavior (aversive). Thus, activity in a specific nerve, En2, acts as a general reinforcer during conditioning of feeding responses.

One CPG neuron, B51, has been found to be a site of modulation during operant learning. Activity in B51, a member of the CPG that produces BMPs, correlates with the occurrence of the ingestion-like motor patterns that can be induced by stimulation of nerve n2,3. B51's that have received contingent reinforcement show greater input resistance, an increase in the occurrence of plateau potentials, and decreased burst thresholds, but showed no change in V_m relative to either yoked or untrained controls.

Convergence of Different Forms of Memory Processing

Now that the mechanisms of various associative and nonassociative forms of memory have been described at the behavioral and cellular levels, it is possible to compare them on a mechanistic level. The first analysis of this sort involved comparing the cellular effects classical conditioning on interneuron B51 to what was already known for operant conditioning (see above). As in operant conditioning, classical

conditioning led to an increase in the number of CS-evoked plateau potentials; however, in contrast to operant conditioning, classical conditioning caused no significant change in input resistance, and in fact there was an increase in plateau potential threshold. Thus, although both types of conditioning result in an increase in the number of plateau potentials in B51, the mechanisms are not the same.

Another site of convergence of operant and classical conditioning is the requirement for dopamine (DA). DA plays an essential role in reward learning, both in vertebrates and in invertebrates. DA was found to be required for *in vitro* models of both operant and classical conditioning.

Over the last several decades, a wealth of information has been learned about several types of memory formation in *Aplysia*. The mechanisms of both associative and nonassociative forms of memory have been studied at the behavioral, synaptic, and molecular levels. As researchers continue to learn more about these mechanisms, they can also now begin to compare the signaling pathways involved between different types of associative memory, for example, between operant and classical conditioning in the feeding behavior (see above), and also between associative and nonassociative forms; for example, it has recently been shown that the gill-withdrawal reflex, long known to exhibit sensitization and habituation, also demonstrates operant conditioning.

Hermissenda

Rotation-induced inhibition of phototaxis in the nudibranch mollusk *Hermissenda* has been a powerful model for the study of classical conditioning. The circuitry involved in the behavior is more complex than either of the *Aplysia* circuits; however, the easily accessible sensory neurons (type B photoreceptors) are a site of convergence for the CS (light) and US (rotation) pathways, providing an excellent opportunity to examine the cellular and molecular mechanisms underlying the interaction of these two signals. Prior to conditioning, *Hermissenda* display a phototactic response by locomoting to a light source (CS). Conversely, when shaken or rotated vigorously (US), the animal contracts its foot and ceases locomotion. After CS–US pairing, *Hermissenda* exhibit two conditioned responses to the CS alone: (1) inhibition of light-induced locomotion and (2) foot shortening. This behavior has been extensively studied both experimentally and with computational models, and it has been found to share many of the characteristics of classical conditioning seen in mammals, including extinction and savings.

Recent studies have focused on two mechanistic levels. First, ongoing examination of the primary sensory neurons continues to further elucidate the cellular and molecular mechanisms of the convergence of the CS and US signals at the type B photoreceptors. Second, progress is being made to identify interneurons and motor neurons downstream of the primary SNs, and to determine (1) how their circuit and cellular properties are altered by conditioning, and (2) how these changes contribute to changes in conditioned behavior.

Memory for light-induced inhibition of phototaxis can exist in several temporal domains. One or two pairings induce STM, which lasts <10 min and requires activation of PKC. Multiple pairings induce longer-lasting forms of memory: ITM, which lasts about 2.5 h and requires protein synthesis, and LTM, which lasts ≥ 1 day and requires both translation and transcription. Recently, two forms of LTM have been differentiated mechanistically; consolidated LTM (CLTM) (which lasts up to 3 days) can be distinguished from LTM (which lasts only about 24 h) by its requirement for a second round of protein and RNA synthesis, and by its disruption by the arginyl-glycyl-aspartate (RGD) peptide, a competitive inhibitor of integrins.

One characteristic of LTM that is currently a subject of much investigation in mammalian systems, as well as in invertebrates, is its susceptibility to disruption by protein synthesis blockers during reactivation by the CS alone, a process called reconsolidation. Evidence for reconsolidation has been demonstrated in LTM in *Hermissenda*. Animals were given LTM training, followed a few hours later by a single retrieval test (CS alone) to reactivate the memory. If this test was followed by treatment with a protein synthesis blocker (anisomycin) within 10–30 min, LTM was not induced. Similar results were obtained if transcription was blocked with actinomycin or if bond formation between cell adhesion molecules (CAMs) was blocked with the RGD peptide.

Cellular and Molecular Mechanisms

Light activates rhodopsin in the cell membranes of two types of photoreceptors of the *Hermissenda* eye: type A and type B. The US, shaking or rotation of the animal, excites hair cells (HCs) in the statocysts. As mentioned earlier, the first site of convergence of the CS and US is at the primary SNs, namely the B-type photoreceptors. The HCs synapse onto the type B photoreceptors both directly through GABA-ergic synapses and indirectly via as yet unidentified serotonergic interneurons. After CS–US pairing, photoreceptors exhibit increased excitability, increased input resistance, increased firing frequency, and increased generator potentials in response to light. These cellular

changes are also observed in *in vitro* preparations using a single trial conditioning procedure in which light (CS) is paired with 5-HT application (US). We will focus here on the molecular mechanisms of convergence at the type B photoreceptors, though it should be kept in mind that there are further sites of CS–US interaction, including the first level of interneurons.

In the type B photoreceptors, the main effect of the combined CS and US signals is considered to be an enhanced increase in internal Ca^{2+} via activation of two different signaling pathways, which induces a persistent activation of PKC. Light activates rhodopsin on the B photoreceptor membrane, which, via G-protein interactions, activates a signaling cascade through PLC, DAG, IP3, and the IP3 receptor to release Ca^{2+} into the cytoplasm from endoplasmic reticulum (ER) stores. Activation of the HCs by shaking or rotation causes release of GABA from the HCs, which binds metabotropic GABA receptors on the photoreceptor and leads to Ca^{2+} release from intracellular stores through mechanisms that are not entirely known but are thought to involve phospholipase C (PLC) activation. In addition, HCs also make excitatory serotonergic connections onto the type B photoreceptors, probably through a polysynaptic pathway. 5-HT activates MAPK, through both PKC-dependent and PKC-independent pathways.

Accompanying the long-term physiological changes is a change in the morphology of the type B photoreceptors. Multiple paired training trials induce contraction of the terminal branches of the type B photoreceptors. This morphological response is sensitive to ryanodine receptor (Ryr) blockers and therefore is dependent on Ca^{2+} release from internal stores. The functional significance of the reduction in arborization is not yet clear.

Downstream Effectors of Transcription and Translation

Much of the focus of the molecular mechanisms of ITM and LTM in *Hermissenda* has been on the events occurring upstream of transcription. However, the protein products of conditioning-induced translation and transcription are currently being investigated. One class of molecules that appear to be necessary for intermediate- and long-term changes in synaptic efficacy is involved in actin stability. Cytoskeleton-related proteins play important roles in cell morphology and in the composition of the cell membrane, and the activity of several of these have been found to be altered by conditioning in *Hermissenda*. Conditioned stimulus pathway phosphoprotein 24 (Csp24) is a cytoskeletal-related protein with several actin-binding domains that are upregulated and phosphorylated by procedures that induce intermediate-term and long-term increases in excitability, and it is associated with a decrease in a specific potassium current, I_A. In single trial conditioning, 5-HT mediates phosphorylation of Csp24 through at least two pathways: (1) by inhibition of the Rho/ROCK pathway, and (2) by an increase in activity of the cyclin-dependent kinase 5 (cdk 5) pathway. Another protein that is upregulated by paired CS–US training is calexcitin, a GTP- and Ca^{2+}-binding protein. Calexcitin is activated by PKC and has the dual role of inactivating the K^+ currents I_A and I_{Ca-K} and binding to the Ryr receptor to release Ca^{2+} from internal stores.

Hermissenda is a model system that has provided a wealth of information about the cellular and molecular mechanisms underlying the convergence of CS and US pathways. What is missing now are many of the links between the various players, for example, the specific role of MAPK in the induction of LTM, the interaction between PKC activation, the role of 5-HT, and the function of the morphological changes for the expression of LTM. Recent work is now focusing on the role of the next order of cells, the effects of classical conditioning in the first level of interneurons. It will be of great interest to learn how the CS and US interact at these levels.

Lymnaea

Several forms of associative conditioning have been described in *Lymnaea*. These include operant conditioning of the respiratory response and classical conditioning of feeding behavior. The respiratory system in particular has the advantage of having a simple circuit, providing the opportunity to examine the specific roles of individual, identified neurons in the formation of memory. Specifically, the CPG neuron RPed1 has been shown to be a site of memory formation and storage. Similarly, feeding behavior, although it is composed of a slightly more complex circuit, has also provided an opportunity to study the role of a specific neuron, the cerebral giant cell (CGC) in memory formation. The feeding behavior is of further interest because this simple circuit is capable of rather complex conditioning. The emphasis of research in *Lymnaea* conditioning has been on the training parameters required for memory induction and on cell and circuit properties. Nevertheless, progress is also being made in the molecular mechanisms underlying cellular changes following conditioning.

Classical Conditioning

The *Lymnaea* feeding response can be classically conditioned, and both appetitive conditioning and conditioned taste aversion (CTA) have been demonstrated. By pairing either tactile stimulation (lip touch) or chemosensory stimulation (amyl acetate)

as the CS with sucrose (US), the CS gains the ability to elicit a feeding response. Appetitive conditioning can be induced with either multiple pairings or with a single pairing in animals that have been deprived of food for several days. To induce CTA, sucrose (CS) is paired with KCl (US). KCl alone decreases the feeding response and elicits a withdrawal response, and after pairing with sucrose, sucrose alone evokes a reduced feeding response that can last up to 1 month.

Classical conditioning in *Lymnaea* shares several features with classical conditioning observed in mammals, including taste discrimination and second-order conditioning. To test taste discrimination, animals were first tested separately with two CS stimuli, sucrose and carrot juice. CTA was established by ten pairings of CS1 (either sucrose or carrot juice) and the US (KCl). After conditioning, animals showed CTA to CS1, but not to the other unpaired CS. Thus, the animals can discriminate between different appetitive stimuli. To test second-order conditioning, these same animals were then given CS2 paired 10 times with CS1. When the animals were then tested with the presentation of CS2 alone, they showed a reduction in their feeding response, demonstrating that they are capable of second-order conditioning.

Cellular and molecular mechanisms In contrast to the extensive investigation into the behavioral aspects of learning and memory in the *Lymnaea* feeding system, examination of the molecular mechanisms required for memory induction is only beginning to be explored. As is typical of most memory systems, LTM of appetitive conditioning requires both transcription and translation, and, as was shown in *Aplysia*, the signaling pathways include MAPK, CREB, and C/EBP. At the cellular level, initial studies have found that injecting cAMP into a CPG neuron involved in the gating of the feeding response (CGC) induces LTF of synapses between CGC and feeding MNs, and this facilitation is blocked by injection of CRE oligonucleotides. The IEG C/EBP has been cloned in *Lymnaea*, and the level of C/EBP is modulated in MN B2 by CTA training. Training decreased LymC/EBP mRNA levels but increased translation and phosphorylation of LymC/EBP protein. Consolidation of LTM in *Lymnaea* appears to require nitric oxide (NO) signaling. Two different neuronal NO synthase genes have been cloned from *Lymnaea*, one of which shows a transient increase in mRNA levels after conditioning.

Operant Conditioning

Lymnaea have the capability to breathe through two different mechanisms. They can either absorb oxygen from the water directly through their skin by cutaneaous respiration, or, under anoxic conditions, they will come to the surface and breathe air into their lungs by opening a respiratory structure called a pneumostome. In the laboratory, anoxic conditions can be introduced by bubbling N_2 into a beaker of water. Snails can be operantly conditioned to reduce the number of times they open their pneumostomes for respiration by applying an aversive tactile stimulus to the lip area of the pneumostome whenever it opens.

The pattern of training trials (trial number and spacing) affects the duration of the memory induced. In general, two training trials (10–45 min each) at an ITI of 10 min to 1 h induce a protein synthesis-dependent form of memory (ITM) that lasts up to 2–3 h. LTM, defined by a translational and transcriptional dependence, is induced by increasing the number of training trials, the ITI, or both. Depending on the precise pattern of induction, LTM can last from up to 24 h to several weeks.

Maintenance of LTM in *Lymnaea* can be manipulated by extinction training, and by disruption of protein synthesis during reactivation of the memory. Three 45 min training trials given over 2 days induce LTM that lasts up to 5 days. Extinction trials (three sessions over 2 days, without reinforcement), begun 1 h after the original training, lead to an increase in pneumostome openings within 2 h. When tested 24 h after extinction training, spontaneous recovery of the operant behavior is observed.

As in *Hermissenda*, memory in *Lymnaea* also exhibits a protein synthesis dependence during memory reactivation. In *Lymnaea* reconsolidation experiments, LTM lasting 7 days was induced in one group of animals using training trials spaced over 2 days. A second group of animals served as yoked controls. On day 7, a single training session was given to all animals. Four hours later, animals that received contingent reinforcement showed memory, whereas yoked controls did not. Reactivation-induced memory was blocked if animals were cooled for 1 h immediately after reactivation or were injected with actinomycin to block transcription.

Cellular and molecular mechanisms Much of the recent research in *Lymnaea* operant conditioning has focused on the role of an identified neuron of the respiratory CPG, RPeD1. The CPG circuit responsible for respiratory pumping consists of a three interneurons (RPeD1, VD4, and IP3I) that are connected by reciprocal excitatory and inhibitory synapses. These neurons and their connections can be studied in semi-intact preparations as well as in isolated cell culture. A decrease in activity in RPeD1 is important for the induction of LTM. The intrinsic membrane

properties of RPeD1 are not altered by conditioning; however, RPeD1 neurons were more often quiescent in ganglia from trained animals, and their synapses onto IP3 were weaker. Operant conditioning was also found to affect the synaptic connections between CPG neurons and between RPeD1 and the pneumostome muscles. Operant conditioning reduced the number of IP3-induced RPed1 bursts and reduced the percentage of cases in which RPed1 stimulation induced pneumostome closure. In addition, in a gain-of-function experiment, it was found that although 2×20 min training sessions do not normally induce 18 h LTM, this training regimen paired with hyperpolarization of RPeD1 during the intervening ITIs did induce LTM. A more direct role for RPeD1 was found by Scheibenstock et al. When these authors ablated the soma of RPeD1 by intracellular injection of pronase, (1) the animal still showed normal respiratory behavior 2 days later, (2) RPed1 maintained appropriate synapses with the other CPG cells, and (3) conditioning still induced ITM; however, LTM was not induced. If ablation of RPed1 was delayed until 1 h after training, LTM was induced, indicating that RPeD1 is a site of induction for long-term operant conditioning in *Lymnaea*. The role of RPeD1 has also been studied in the induction of extinction and reconsolidation. Ablation of RPeD1 either 1 h after operant conditioning or 2 days before extinction training blocks extinction, and ablation immediately after reactivation blocks reconsolidation.

The signaling pathways involved in operant conditioning in *Lymnaea* have not been fully examined. CREB1 and CREB2 have been cloned in *Lymnaea*. *In situ* hybridization with a CREB2 antisense reveals the presence of CREB2 throughout the central nervous system (CNS), while similar labeling with a CREB1 antisense indicates that a much smaller population of cells expresses this mRNA. Neurons that express CREB1 include the CGC which is involved in the feeding system and RPeD1. Thus, the stage is now set for the investigation of these transcription factors, which have been clearly implicated in LTM in other systems.

Lymnaea provides an important system for the study of learning and memory because both the feeding and respiratory systems exhibit rather complex learning, especially considering the simplicity of the CPG circuits involved in the behaviors. It will be of considerable interest to map the signaling pathways and the synaptic changes responsible for these forms of memory.

Concluding Remarks

In conclusion, several molluskan preparations have contributed importantly to our understanding of the cellular and molecular mechanisms of learning and memory. Each animal we have discussed exhibits very different types of memory in a variety of different response systems. Nonetheless, several striking similarities emerge. First, each system exhibits memories in three mechanistically distinct temporal domains. Second, many similar signaling cascades are implicated in memory formation, especially, PKA, PKC, and MAPK. Finally, LTM in each animal involves gene activation. Even more striking is the fact that these same temporal phases, signaling cascades, and transcriptional mechanisms are involved in a variety of forms of memory in higher mammalian species. Taken collectively, it appears that evolution has used a common tool kit to implement similar memory mechanisms across the animal kingdom.

See also: Invertebrate Models to Study Learning and Memory: Lymnaea; Learning and Memory in Invertebrates: *Aplysia*; Learning and Memory in Invertebrates: *Hermissenda*; Neuroplasticity: Computational Approaches; Synaptic Plasticity: Short-Term Mechanisms.

Further Reading

Benjamin PR, Staras K, and Kemenes G (2000) A systems approach to the cellular analysis of associative learning in the pond snail *Lymnaea*. *Learning and Memory* 7: 124–131.

Blackwell KT (2006) Subcellular, cellular, and circuit mechanisms underlying classical conditioning in *Hermissenda crassicornis*. *Anatomical Record Part B, New Anatomist* 289: 25–37.

Botzer D, Markovich S, and Susswein AJ (1998) Multiple memory processes following training that a food is inedible in *Aplysia*. *Learning and Memory* 5: 204–219.

Crow T (2004) Pavlovian conditioning of *Hermissenda*: Current cellular, molecular, and circuit perspectives. *Learning and Memory* 11: 229–238.

Crow T and Tian LM (2006) Pavlovian conditioning in *Hermissenda*: A circuit analysis. *Biological Bulletin* 210: 289–297.

Guan Z, Giustetto M, Lomvardas S, et al. (2002) Integration of long-term-memory-related synaptic plasticity involves bidirectional regulation of gene expression and chromatin structure. *Cell* 111: 483–493.

Lechner HA, Baxter DA, and Byrne JH (2000) Classical conditioning of feeding in *Aplysia*. I: Behavioral analysis. *Journal of Neuroscience* 20: 3369–3376.

Lechner HA, Baxter DA, and Byrne JH (2000) Classical conditioning of feeding in *Aplysia*. II: Neurophysiological correlates. *Journal of Neuroscience* 2: 3377–3386.

Lorenzetti FD, Mozzachiodi R, Baxter DA, and Byrne JH (2006) Classical and operant conditioning differentially modify the intrinsic properties of an identified neuron. *Nature Neuroscience* 9: 17–19.

Lukowiak K, Sangha S, Scheibenstock A, et al. (2003) A molluscan model system in the search for the engram. *Journal de physiologie-Paris* 97: 69–76.

McComb C, Rosenegger D, Varshney N, Kwok HY, and Lukowiak K (2005) Operant conditioning of an *in vitro*

CNS-pneumostome preparation of *Lymnaea*. *Neurobiology of Learning and Memory* 84: 9–24.

Reissner KJ, Shobe JL, and Carew TJ (2006) Molecular nodes in memory processing: Insights from *Aplysia*. *Cellular and Molecular Life Sciences* 63: 963–974.

Ribeiro MJ, Schofield MG, Kemenes I, et al. (2005) Activation of MAPK is necessary for long-term memory consolidation following food-reward conditioning. *Learning and Memory* 12: 538–545.

Sugai R, Shiga H, Azami S, et al. (2006) Taste discrimination in conditioned taste aversion of the pond snail *Lymnaea stagnalis*. *Journal of Experimental Biology* 209: 826–833.

Susswein AJ, Schwarz M, and Feldman E (1986) Learned changes of feeding behavior in *Aplysia* in response to edible and inedible foods. *Journal of Neuroscience* 6: 1513–1527.

Sutton MA, Masters SE, Bagnall MW, and Carew TJ (2001) Molecular mechanisms underlying a unique intermediate phase of memory in *Aplysia*. *Neuron* 31: 143–154.

Learning, Action, Inference and Neuromodulation

P Dayan, University College London, London, UK
N D Daw, New York University, New York, NY, USA
Y Niv, Princeton University, Princeton, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuromodulators, including dopamine, norepinephrine, acetylcholine, and serotonin, exert a wide and diverse collection of effects on a broad range of cortical and subcortical target structures. They are also critically implicated in many aspects of normal behavior and both neurological and psychiatric conditions. With the exception of acetylcholine, they are synthesized by neurons whose somas lie in relatively small subcortical regions such as the ventral tegmental area and substantia nigra pars compacta (for dopamine), the locus coeruleus (for norepinephrine), and the dorsal and median raphe nuclei (for serotonin). However, these neurons can have massive axonal arbors, innervating large areas. The activity of the neuromodulatory neurons and the release of their associated neuromodulators (which is subject to many factors other than the momentary spiking rates) onto the bewildering diversity of their pre- and postsynaptic receptors exhibit a large number of incompletely understood temporal and spatial regularities.

There are obviously huge gaps in our knowledge about neuromodulators, and there are doubtless many as yet undiscovered complexities. However, in some key respects, they appear to admit relatively simple theoretical characterizations associated with learning predictions of affectively important outcomes such as rewards and punishments, learning to choose actions that optimize those outcomes, and regulating and controlling certain general aspects of information processing. These are the most basic functions associated with the survival of mobile, decision-making organisms, and it turns out that a very helpful normative underpinning for these theories can be derived from statistics, operations research, and engineering. It should be acknowledged at the outset, however, that these theories ignore many important functions of the neuromodulators, such as their role in regulating sleep–wake cycles. Nonetheless, they are influential in organizing large bodies of experimental results and also in compelling new empirical approaches.

Under these theoretical views, neuromodulators report on two main quantities: (1) information associated with positive and negative outcomes and (2) information associated with uncertainty. It seems, at least to a crude approximation, that dopamine and serotonin play a special role in positive and negative affect, with norepinephrine and acetylcholine being particularly involved with uncertainty. Further, orthogonal to this distinction, the neuromodulators also have two main classes of effects on neural processing: (1) influencing neural plasticity and learning and (2) regulating different sources of input to neurons and controlling competition between networks of neurons.

This article first provides a somewhat general characterization of issues surrounding learning and then briefly discusses immediate effects of the neuromodulators on inference and regulation.

Learning

Long-term synaptic plasticity, that is, persistent facilitation and depression of the efficacies of synapses, can be characterized at many different levels of biological realism and according to many different theoretical abstractions. Unfortunately, experimental difficulties have hindered efforts to understand completely the range of effects of neuromodulators on plasticity; therefore this article proceeds at a more abstract level appropriate to the relevant normative theories.

The central concept for the learning associated with neuromodulators is prediction. Consider a very simple experiment comprising many trials, in each of which a particular image may or may not be presented and a drop of pleasant-tasting fruit juice may or may not be delivered (regardless of what the subjects do). If the juice is provided more often with than without the image, then subjects learn to expect or predict the delivery of the juice when the image is presented. Such predictions can be measured via behaviors such as anticipatory salivation or licking, which are thought to be triggered more or less directly and automatically by the prediction of the reward. This is a very simple instance of what is known as classical or Pavlovian conditioning. The image is called a conditioned stimulus, the juice is known as an unconditioned stimulus, and the predictive behaviors are termed conditioned responses.

Trial-Based Prediction

Reinforcement learning theories formalize this prediction learning by positing that the presence of the image on trial n is coded by the activity of a unit $x^n = 1$ (with $x^n = 0$ if the image was not presented). This unit makes a connection, via an abstract synapse with strength w^n, to a neuron. This neuron also receives information about the delivery ($r^n = 1$) or nondelivery ($r^n = 0$) of juice on trial n. The value r^n would be scaled

on the basis of the appetitive utility of the juice. The idea is that the synaptic connection should be modified according to a learning rule, such that the presence of the stimulus will predict the occurrence of the reward.

One rather general form of such a learning rule is

$$w^{n+1} = w^n + \alpha^n \delta^n x^n \qquad [1]$$

which indicates how w^n changes from one trial to the next. This update is a function of three quantities:

α^n (learning rate): scales the size of the change to the synapse.

δ^n (prediction error): incorporates information about unpredicted or incorrectly predicted outcomes and typically controls the direction of change

x^n (synaptic eligibility): here, limits synaptic change to those trials in which the image is presented ($x^n = 1$).

Different reinforcement learning theories posit different forms for these terms. Many early ideas in this area came from the field of mathematical psychology, where they were invented to describe the development of behavioral responding in conditioning experiments. Their relationship both with the neural data on the actual activities and concentrations of the neuromodulators and with normative computational ideas came substantially later.

One of the most important psychological learning rules is called the Rescorla–Wagner rule. This is based on the idea that the net prediction of juice on trial n is just the total presynaptic input, here $v^n = x^n \times w^n$, and derives learning from the prediction error, the difference between the juice that is provided and the juice that is predicted:

$$\delta^n = r^n - w^n \times x^n \qquad [2]$$

$$= r^n - v^n \qquad [3]$$

In its simplest form, the Rescorla–Wagner rule also suggests that the learning rate $\alpha^n = \alpha$ is constant. Figure 1(a) shows the operation of this rule, indicating how the values of r^n and x^n specify w^n over a whole set of trials.

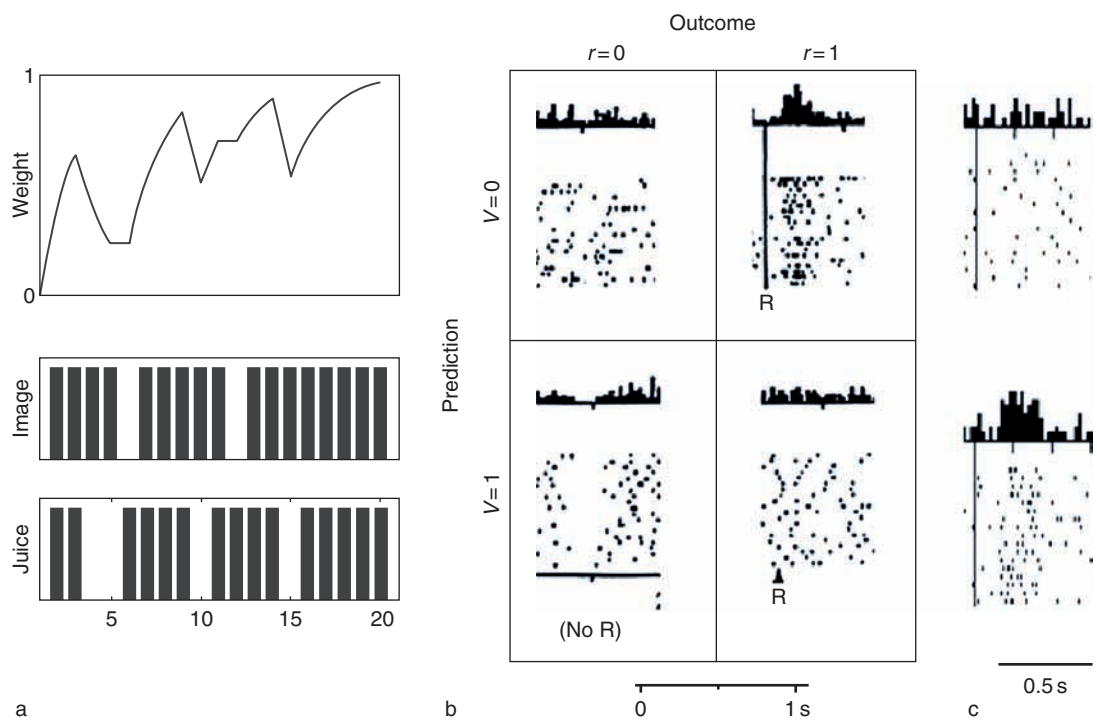

Figure 1 Dopamine and learning rules. (a) The course of learning using the Rescorla–Wagner rule. The top plot shows the evolution of the weight w^n over the course of the 20 trials with an image and juice being provided as in the bottom plot. The learning rate $\alpha = 0.4$. The stochastic nature of the learning to a level around the asymptotic value of 0.8 is apparent. (b) Evidence for a prediction error at the time of reward expectancy is shown in the activity of dopamine cells. The four boxes show rasters and accumulated histograms of the activity of identified dopamine neurons in macaque monkeys (during 20 trials), where the reward is a small drop of juice (provided at the time marked by R). The plots are in accord with the prediction error of eqn [3]. (c) The activity of dopamine neurons in response to the presentation (marked by vertical lines) of a conditioned stimulus which either does not (upper) or does (lower) predict the future delivery of juice. These are consistent only with temporal difference learning and not the Rescorla–Wagner rule. (b) Adapted from Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599. (c) Reprinted, with permission, from the Annual Review of Physiology, Volume 57 © 2006 by Annual Reviews www.annualreviews.org.

Although extremely simple, this rule has quite a number of desirable computational properties and an excellent normative basis in engineering (where it is more usually known as the delta rule). Most important, it leads to correct predictions, at least on average. In particular, if the juice is always provided with the image, then the weight w^n will come to take the value 1; if the juice is provided on half the trials with the image, w^n will take the value of 0.5, and if the juice is never provided, then w^n will become 0. The rule also has a decent evidentiary trail in psychology, capturing many empirical phenomena surrounding the acquisition and extinction of conditioned behaviors in variants of the simple Pavlovian conditioning experiment discussed here.

Somewhat remarkably, the error signal δ^n in this form of the rule also appears to capture well some, but crucially not all, of the characteristics of the phasic spiking activity of dopamine cells as recorded electrophysiologically in the midbrain during a learning task like this. Figure 1(b) shows an example of this match. The figure shows a raster plot of the activity of a dopamine neuron in an awake, behaving macaque monkey at the time of the delivery or nondelivery of an unexpected or expected reward. If, indeed, the prediction error occasioned by delivery or nondelivery of juice is reflected by phasic bursts of activity above or below the slow, irregular baseline of neural firing, then from eqn [3], one should expect the following: suprabaseline activity for the delivery of the unexpected reward ($\delta^n = 1-0 = 1$; for instance, at the outset of learning, when the participant cannot make an accurate prediction); infrabaseline activity for the nondelivery of the expected reward ($\delta^n = 0-1 = {}^{-}1$); and, crucially, baseline activity when an expected reward is provided ($\delta^n = 1-1 = 0$). We would also expect baseline activity for the nondelivery of a nonexpected reward ($\delta^n = 0-0 = 0$). The figure, although not showing data collected in exactly this manner, shows just these expectations to be fulfilled. Note, though, that dopamine activity does not offer a completely linear code for δ^n. The potential fidelity of the coding of negative prediction errors is limited by the fairly low baseline activity of the neurons; there is also some evidence that the coding is adaptive, moulding itself to the appropriate statistical contingencies.

In sum, at the time at which reward delivery is expected, dopamine neurons appear to convey a form of prediction error for the delivery of reward, a signal that is exactly appropriate for training predictions (as in eqn [1]). This finding leads to a large range of questions about what happens when the temporal relationship between the image and the juice is manipulated (the data here showing that the Rescorla–Wagner rule is critically flawed) or when there are multiple predictive images or stimuli, how prediction learning and action learning are coupled, and how punishments (e.g., small electric shocks) are processed. There are also some data about the neural structures involved in the various components of this learning rule (i.e., x^n and w^n), notably implicating nuclei in the amygdala and the orbitofrontal cortex, together with their connections to the nucleus accumbens, all structures that receive a strong dopamine projection.

Temporal Prediction

The use of the word prediction hints at the critical role that time plays in conditioning experiments. Broadly speaking, to be a good predictor, it is important that the image precede (rather than coincide with or follow) the juice. This leads to both conceptual and psychological/neural problems. The conceptual problem is that it is necessary to change the goal of prediction – there is no point in predicting the amount of juice that will be provided at the same time as the image, since they are not delivered at the same time; rather, it is necessary to predict the amount that will come in the future. An important step in the field was the suggestion by Richard Sutton and Andrew Barto that participants should try to predict a long-run measure of the reward, such as its sum over a whole trial. Achieving this requires a different prediction error from that in eqn [3]. It also turns out that this new prediction error accounts for more features of the activity of dopamine neurons, specifically their firing patterns at the time of the image presentation.

To understand this new prediction error, one needs a new variable t to count time within each trial. For instance, the image might be presented at time $t = 2$, the juice at time $t = 5$, with the trial ending at $t = T$. Now the reward on trial n is indexed by time $r^n(t)$ with, for instance, $r^n(5) = 1$ on trials on which the juice is given. The new task of the participants is to predict $v^n(t)$ at time t within a trial as the sum of all the future rewards that will be provided on average during that trial after t. Assuming for the moment that rewards are delivered deterministically, this means that the following relationship should hold (for which is used the symbol $\triangleq$):

$$v^n(t) \triangleq r^n(t) + r^n(t+1) + r^n(t+2) + \ldots + r^n(T) \qquad [4]$$

The obvious trouble with this definition as a goal for prediction is that the right-hand side, measuring the actual delivery of juice, cannot be assessed until the end of the trial. The solution to this problem, which underlies the new prediction error, is to observe that the sum can be split into its first and

subsequent terms:

$$v^n(t) \triangleq r^n(t) + [r^n(t+1) + r^n(t+2) + \ldots + r^n(T)] \quad [5]$$

so if the prediction at time $t+1$, $v^n(t+1)$, is, as it should be, equal to $r^n(t+1) + r^n(t+2) + \ldots r^n(T)$, then,

$$v^n(t) \triangleq r^n(t) + v^n(t+1) \quad [6]$$

The new prediction error becomes the discrepancy between the left- and right-hand sides and can be computed after waiting just one timestep, without waiting for the end of the trial:

$$\delta^n(t) = r^n(t) + v^n(t+1) - v^n(t) \quad [7]$$

This is called the temporal difference prediction error because of the key role played by the change in the prediction over consecutive timesteps ($v^n(t+1) - v^n(t)$). Even though, during learning, it cannot be guaranteed that $v^n(t+1) = r^n(t+1) + r^n(t+2) + \ldots + r^n(T)$, it turns out that a learning rule like eqn [1], using $\delta^n(t)$ to update the weights associated with the prediction made at time t will also ultimately arrive at the correct values. Learning of this sort is said to involve bootstrapping because the predictions at one time are learned partly on the basis of predictions at the subsequent time.

The most important practical difference between the previous and present prediction errors (i.e., between eqns [3] and [7]) is the value of the new prediction error associated with the first presentation of the image (in this case, $\delta^n(1) = r^n(1) + v^n(2) - v^n(1) = 0 + v^n(2) - 0 = v^n(2)$). If the image is always followed by juice, then, once learning is complete, the prediction changes from $v^n(1) = 0$ to $v^n(2) = 1$ when the image is provided, resulting in a transient prediction error at time $t = 1$, $\delta^n(1) = 1$. The prediction then remains at 1 until the reward is actually delivered, so there is no difference between $v^n(t)$ and $v^n(t+1)$, and therefore $\delta^n(t) = 0$ for all t preceding the time of reward. At the reward time, (here $t = 5$), no further reward is expected in the trial, so $v^n(6) = 0$, and the error from eqn [7] is $\delta^n(5) = r^n(5) - v^n(5)$. Hence, this prediction error behaves similarly to that in the Rescorla–Wagner rule for all the situations shown in Figure 1(b).

Figure 1(c) shows that the activity of dopamine neurons at the time of the image follows exactly the pattern expected from eqn [7]. It is as if the activity that at the beginning of training happens at the time of the unexpected reward moves backward in time across the trial so that it is instead initiated by the earlier reliable predictor of that reward, namely, the image. The transient activity at the time of the image can play a number of important functions, notably, allowing a predictor of reward to act like a surrogate reward itself, an effect known in psychology as conditioned reinforcement.

The neural problem associated with temporal prediction is exactly that the predictors need to be able to count time; in this case, the time since the image was presented. The well-timed depression in the activity of the dopamine neurons when an expected reward is not delivered is clear evidence that exactly this happens (and incidentally argues strongly against certain other construals of the dopamine activity, for instance that it flags all 'salient' events without regard to their appetitive or aversive valence). However, although there are various suggestions about the involvement of the cerebellum, the hippocampus, the striatum, and the parietal and frontal cortices in counting time, the exact nature of this signal is not well understood.

Multiple Predictive Stimuli

So far only the case of just one potentially predictive image has been considered. Now consider the case in which there are multiple predictive stimuli. It turns out that only the temporally simple case (associated with the Rescorla–Wagner rule) has to be treated, since the issues about time t within a trial are orthogonal to those about the multiple stimuli.

The Rescorla–Wagner version of the learning rule (eqn [1]) was designed to address psychological phenomena that arise when there are multiple predictive stimuli. The main idea is to elaborate it to the following:

$$w_i^{n+1} = w_i^n + \alpha_i^n \delta^n x_i^n \quad [8]$$

where there are now multiple stimuli x_i^n, each with its own prediction weight w_i^n, but, critically, still with only one global prediction error δ^n.

The Rescorla–Wagner rule involves the net prediction v^n of juice, defined as the sum of the predictions associated with each stimulus:

$$v^n = x_1^n \times w_1^n + x_2^n \times w_2^n + \ldots \quad [9]$$

If, perhaps as a result of prior learning, one of the stimuli present on a trial already predicts the reward correctly, then the prediction error (of eqn [3]) will be 0 ($\delta^n = 0$), and therefore the synapses associated with any other stimuli which are also present, including those that do not yet predict the reward, will not change. This effect is known in the psychological literature as blocking. The ample behavioral evidence for blocking was taken as an early sign that learning is indeed driven by prediction errors rather than, for instance, simply correlations. There is also some evidence for blocking in the activities of dopamine neurons.

Stimulus multiplicity could also affect the other components of the learning rule. In particular, a number of competing learning rules that have been suggested on the basis of psychological data suggest that there are stimulus-specific predictions (and prediction errors) and invoke stimulus-specific (and changing) learning rates α_i^n. These rules treat the learning rates as associabilities, reflecting a measure of the attention that the learners lavish on the stimuli. The most celebrated of these accounts, due to John Pearce and Geoffrey Hall, suggests that stimuli should be associable to the degree that they have had surprising consequences. In this view, multiple stimuli can interact with one another to produce learning phenomena such as blocking at least partly through competition for these associabilities. This account has attracted many experimental and theoretical studies.

Experimentally, Peter Holland, Michela Gallagher, and their colleagues have collected evidence from lesion, pharmacological, and behavioral studies suggesting that associabilities are realized (at least in rats) through a pathway involving the central nucleus of the amygdala, attentional and control regions of the cortex, and, notably for the topic of this article, various parts of the acetylcholine system. Oddly, there appear to be different anatomical substrates for increases and decreases in stimulus associability.

From a theoretical perspective, the sort of surprise inherent in Pearce and Hall's suggestion can be formulated as a particular form of uncertainty (a fact that, together with Holland and Gallagher's six results and others, led to the suggestion that such uncertainty might be reported by acetylcholine). Normative theories of prediction learning arising in the fields of statistics and engineering actually suggest learning rules rather like eqn [8] and provide a precise characterization of what factors should control α_i^n. The best known and simplest of these theories, called the Kalman filter, suggests that the learning rates act to divide up the impact of the prediction error δ^n on the weights w_i^n for each of the stimuli such that the stimuli whose weights are least well known (i.e., most uncertain) are most susceptible to learning. That is, the more uncertain the prediction made by one stimulus (for instance, because it has been observed only a few times), the more it is deemed responsible for any nonzero prediction error δ^n, and so the more its prediction w_i^n changes. Variants of this theory also link a stimulus's uncertainty to the degree of surprise that has accompanied it, as for associability in the Pearce–Hall rule. Additionally, in noisy environments in which stimuli and rewards are inherently unreliable, the learner should expect substantial prediction errors even if its predictions are as good as can be. This should have the effect of reducing all the learning rates.

Although the Kalman filter certainly captures something of the flavor of the Pearce–Hall learning rule, its detailed predictions and putative cholinergic substrates have yet to be fully tested, either behaviorally or neurally. Norepinephrine also seems to exert rather general effects on the overall speed of learning, particularly in circumstances in which there are gross changes to an environment. One theoretical account of this suggests that norepinephrine reports a different form of uncertainty, that associated with the whole context. This is coupled to learning since contextual change typically induces a need for gross revision or fresh acquisition of predictions. There are two final issues associated with multiple stimuli. First, the task of combining the predictions of reward made by multiple stimuli can be considered an example of the more general statistical problem of combining multiple sources of evidence. There is ample psychological data suggesting that, as in most statistical accounts, this sort of combination is typically not additive (as in eqn [9]) but rather is competitive, with the predictions made by different stimuli being averaged together and the prediction of the most reliable (least uncertain) predictor being weighted most heavily. How important this is for prediction learning, and indeed what its neural substrate is, are presently unclear.

Second, the notion that each stimulus has its own independent, binary input feature x_i is too simple, particularly if the stimulus representations associated with prediction learning are those in refined areas such as prefrontal cortex. Rather, sophisticated cortical (and perhaps hippocampal) mechanisms will play a critical role in creating stimulus representations that are appropriately tailored to sensory experience.

Action Learning

So far, this article has considered prediction learning but not the selection of actions based on such predictions. Although the prediction of upcoming reward can in itself directly trigger conditioned behavioral responses, such as approach, or the salivation of Pavlov's dogs, these represent a fairly limited portion of the behavioral repertoire. Animals can also learn to take arbitrary actions (such as pressing a lever) in order to obtain rewards that are contingent on such actions. This learning is known as instrumental conditioning, and its neural substrates have actually been more extensively investigated than those of Pavlovian conditioning. This area has revealed some important complexities; for instance, in rats, an instrumentally conditioned behavior such as a lever press can apparently arise through either of at least two neurally and behaviorally dissociable pathways. One of these involves at least the prelimbic prefrontal

cortex and the dorsomedial striatum, may be somewhat independent of dopaminergic neuromodulation, and produces actions that can behaviorally be shown to be sensitive to the particular reward (such as food) expected. Such actions are therefore known as 'goal-directed.' The other pathway involves the dorsolateral striatum and the infralimbic prefrontal cortex, is strongly dependent on dopamine, and produces so-called 'habitual' actions. These are behaviorally characterized by their insensitivity to the particular reward expected; for instance, a rat may habitually press a food delivery lever even if it is not hungry.

There is both behavioral and neural evidence that the habitual action system is based on prediction learning of almost exactly the sort described above. This is consonant with a much larger body of evidence that dopamine is involved in action choice; besides habitual instrumental action, the neuromodulator is associated with addictive drugs, with self-stimulation experiments in which animals will work to receive electrical stimulation in some brain areas, and with motor pathologies such as Parkinson's disease. One particular view of action learning has recently received strong support from data on the activity of dopamine neurons in a task which involves monkeys' making choices between alternatives associated with different probabilities of reward. Under this scheme, the predictions should be not of the reward that will accrue after time t (written as $v^n(t)$ in eqn [4]), but rather of those that will accrue after time t if the learner chooses a particular action (say, a (t)) at time t. These so-called Q-values (written as $Q(t,a(t))$ were originally suggested by Christopher Watkins. If such predictions were known perfectly, action selection could ensue simply by choosing the action whose value is greatest, that is, the one predictive of the most reward. For a number of reasons, it can be prudent to choose somewhat randomly, assigning a greater chance to actions with greater Q-values:

$$P(a(t) = a) = \frac{e^{\beta Q(t,a)}}{\sum_b e^{\beta Q(t,b)}} \qquad [10]$$

Here, $\beta > 0$ controls how fierce the competition is; the larger β, the more likely the action with the largest Q-value will be chosen. One reason to employ such a probabilistic rule is to ensure exploration of unfamiliar alternatives and to avoid behavioral rigidity. It has been suggested that norepinephrine regulates exploration by controlling β, coupling the long-term success of a behavioral policy to reduced exploration and long-term failure to enhanced exploration. There is also starting to be some evidence (from functional imaging in humans) about the involvement of neural structures such as the frontal pole and the intraparietal sulcus in regulating exploration on a trial-by-trial basis.

The recent evidence from the activity of dopamine neurons during choice between several possibly rewarding actions favors a particular temporal-difference rule for learning the Q-values. Just as the prediction error $\delta^n(t)$ in eqn [7] is based on the difference between two successive predictions $v^n(t+1)$ and $v^n(t)$, this new prediction error is based on the difference between two successive predictions, $Q^n(t+1,a(t+1))$ and $Q^n(t, a(t))$, depending on the actions that are actually executed at times t and $t+1$:

$$\delta^n(t) = r^n(t) + Q^n(t+1, a(t+1)) - Q^n(t, a(t)) \qquad [11]$$

This prediction error can then be used at the heart of a learning rule just like that in eqn [1], with the prediction $Q^n(t,a(t))$ being updated on the basis of this prediction error.

Appetitive and Aversive Predictions

This article has discussed appetitive predictions and the role of dopamine. However, in many circumstances, it is actually even more important to predict aversive outcomes. Further, animals can certainly learn to execute particular actions to avoid punishments and can titrate their actions to balance the benefits and costs of rewards and punishments associated with particular actions. The difference between appetitive and aversive predictions and outcomes is completely transparent to the normative theories, which happily consider them to live at positive and negative segments of a single dimension of value. However, there is a huge body of psychological and neural data showing that animals treat them very differently. There are also many open controversies in the literature, including a very active debate about the role of dopamine in aversively motivated behaviors. Aversive predictions play a key role in psychiatric conditions such as anxiety disorders (whose basis may be nonfactual predictions of threat) and paranoid schizophrenia.

These various data have yet to be encompassed in the sort of integrated computational, psychological, and neurobiological account outlined here for appetitive prediction learning and action choice. A popular psychological idea is that separate, but interacting, opponent systems are involved in training appetitive and aversive predictions. One possibility, suggested by William Deakin and Frederico Graeff and still somewhat speculative, is that one part of the serotonin system acts as the aversive opponent to dopamine. The data supporting this idea mainly originate in studies of the behavior of animals under

pharmacological manipulation of serotonergic and dopaminergic function. Full tests of this hypothesis using electrophysiological recordings of serotonergic neurons are only just beginning.

Immediate Effects of Neuromodulators

This article has emphasized the effects of neuromodulators on learning. However, a more traditional view of neuromodulation, established most firmly in studies of central pattern generation in invertebrates, is that it regulates connectivity and excitability over the short term, allowing multiple functional networks to share a single anatomical substrate. Even in the context of the sorts of conditioning functions discussed here, there is indeed evidence that the neuromodulators exert short-term effects not mediated through learning. A few, in particular, have been most extensively investigated, including the impact of dopamine on working memory and on the vigor of responding, and the role of acetylcholine and norepinephrine in regulating different sources of information associated with making predictions.

There is quite some evidence that dopamine, like other neuromodulators, can immediately influence the excitability of neurons, often characterized by the gain of their input–output functions. If the neurons are connected in competitive networks (for instance, subserving the choice between multiple possible actions), then changing the gain (a bit like changing β in eqn [10]) can change the dynamics and stability of the competition. The effect of this on action choice has been studied in some detail by Jonathan Cohen and his colleagues. In prefrontal cortex, a similar effect of dopamine has been suggested as being important for gating the short-term storage of information (such as the fact that an image predicting reward was recently presented). As shown, this sort of storage plays a critical role in making appropriate predictions and therefore also in choosing appropriate actions.

Dopamine also appears to control the vigor or the force of behavioral responses. For instance, the overt effects of dopaminergic agonists such as amphetamine are hyperactivity; antagonists produce lethargy and affect the propensity to choose actions that have high costs. These effects also appear to be immediate, in that learning is not required for them to be exhibited. The reinforcement learning theories so far described do not encompass such phenomena, since they are designed to address choice among a discrete set of candidate actions and lack any analogue of a response vigor. A recent extension of the standard reinforcement learning picture accommodates free-operant tasks, in which subjects choose the latency of each of their responses, balancing the costs and benefits of vigorous responding. The model argues that tonic levels of dopamine report the average rate of reward and act as a form of opportunity cost which determines the optimal response vigor. The greater the opportunity cost, the greater the expectation of reward per unit time, and so the faster the participant must act to achieve this. This also implicates tonic dopamine in mediating the effects of motivation on response vigor. For instance, satiety, reducing the value of each reward, would decrease the net average rate of reward and thus result in lower tonic dopamine levels and more-sluggish responding.

Finally, if acetylcholine reports on the learner's uncertainty about aspects of its environment, then from a statistical viewpoint, it should influence not only learning but also the interpretation of sensory input – in statistical terms, inference. In particular, if uncertainty is high, then 'top-down' information about prior beliefs and expectations should be less trusted than 'bottom-up' information from sensation. This involves just the sort of regulation of connectivity that is implied by invertebrate neuromodulation. There is indeed substantial evidence that acetylcholine can boost the impact of bottom-up information compared with top-down information. Concomitantly, if norepinephrine reports on contextual change, then it should also have the effect of regulating the use of top-down information.

Summary

Neuromodulators are involved in some of the most basic adaptive functions of an organism, notably how it behaves and learns to behave appropriately in light of rewards, punishments, and other information. Dopamine and, putatively, serotonin appear to report prediction errors for future reward and punishment, to influence the acquisition of predictions and appropriate habitual actions, and to effect motivationally appropriate responding. Acetylcholine and norepinephrine seem to report on forms of uncertainty to control the allocation of learning among different possible predictors, the overall speed of learning, and the way that information from sensation is integrated with information based on past experience in an environment.

Accounts adapted from statistics, operations research, computer science, and engineering offer a normative underpinning for both the behavioral data and at least some key aspects of the neural substrates.

See also: Acetylcholine Neurotransmission in CNS; Basal Ganglia: Acetylcholine Interactions and Behavior; Basal Ganglia: Habit; Conditioning: Theories; Delayed Reinforcement: Neuroscience; Dopamine; Dopamine – CNS Pathways and Neurophysiology; Neuroendocrine

Control of Energy Balance (Central Circuits/ Mechanisms); Neuromodulation; Neutrotransmission and Neuromodulation: Acetylcholine; Operant Conditioning of Reflexes; Prediction Errors in Neural Processing: Imaging in Humans; Procedural Learning: Classical Conditioning; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Sleep and Sleep States: Cytokines and Neuromodulation.

Further Reading

Aston-Jones G and Cohen JD (2005) An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Reviews of Neuroscience* 28: 403–450.

Braver TS, Barch DM, and Cohen JD (1999) Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry* 46: 312–328.

Daw ND, Kakade S, and Dayan P (2002) Opponent interactions between serotonin and dopamine. *Neural Networks* 15: 603–616.

Dayan P, Kakade S, and Montague PR (2000) Learning and selective attention. *Nature Neuroscience* 3: 1218–1223.

Deakin JFW and Graeff FG (1991) 5-HT and mechanisms of defence. *Journal of Psychopharmacology* 5: 305–316.

Dickinson A (1980) *Contemporary Animal Learning Theory*. Cambridge, UK: Cambridge University Press.

Doya K (2000) Meta-learning, neuromodulation and emotion. In: Hatano G, Okada N, and Tanabe H (eds.) *Affective Minds*, pp. 101–104. Amsterdam: Elsevier Science.

Holland PC and Gallagher M (1999) Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences* 3: 65–73.

Pearce JM and Hall G (1980) A model for Pavlovian learning: Variation in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review* 87: 532–552.

Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

Sutton RS and Barto AG (1998) *Reinforcement Learning*. Cambridge, MA: MIT Press.

Yu AJ and Dayan P (2005) Uncertainty, neuromodulation, and attention. *Neuron* 46: 481–492.

Lexical Impairments Following Brain Injury

R Casas and D Tranel, University of Iowa, Iowa City, IA, USA

© 2009 Elsevier Ltd. All rights reserved.

Lexical retrieval, which refers to the ability to retrieve words, is an essential component of language communication. Production of an appropriate word for a given stimulus involves activation of multiple processes in the brain. Damage to brain areas associated with the processes involved in lexical retrieval can lead to naming deficits that disrupt fluent speech, which can greatly encumber an individual's ability to communicate. Such impairments are common sequelae of brain injury and thus are an important topic of clinical and scientific inquiry. The types of brain injuries that commonly lead to impairments in lexical retrieval include stroke, head injury, surgical resection to treat pharmacoresistant epilepsy (left temporal lobectomy), and degenerative diseases that have a proclivity for left hemisphere regions important for lexical retrieval (including Pick's disease and Alzheimer's disease). This article summarizes the nature and types of lexical impairments that can occur as a result of damage to specific brain areas.

Deficits in lexical retrieval refer to the inability to provide names for things such as concrete entities (typically named by nouns), actions (typically named by verbs), or spatial relationships (typically named by prepositions). Defined as 'anomia' in the medical and scientific literature, lexical retrieval impairments are a frequent part of the symptom complex that characterizes patients with aphasia. In general, aphasia can be defined as an acquired disturbance of the comprehension and formulation of verbal messages following damage to language-related brain structures (typically in the left hemisphere). Lexical retrieval, which requires the capacity to retrieve a particular word to designate an entity or event at the appropriate level of specificity, is different from the ability to recognize the entity or event. Whereas 'naming' refers specifically to knowing what something is called, 'recognition' involves semantic understanding of the entity or event and, hence, the ability to retrieve conceptual knowledge for it. Patients with anomia have lost the ability to retrieve names of things, but they have not lost the ability to recognize what things are. Hence, even when they cannot name things, anomic patients can usually produce accurate descriptions of those things or indicate by gestures and other nonverbal forms of communication that they have normal knowledge of things. For example, when shown a picture of a camel, the patient may say, "That is an animal that has humps on its back, lives in the desert, and can go for a long time without water." When shown a picture of Bill Clinton, the patient may say, "That guy was a president; had an affair, had a southern accent." Similar 'subclinical' forms of lexical retrieval problems occur fairly frequently in the realm of normal experience, particularly under conditions of fatigue, distraction, or in connection with normal aging. A related phenomenon is 'tip-of-the-tongue' state, which occurs when normal individuals experience the inability to retrieve a particular name (especially proper names) even though they know perfectly well what it is that they are attempting to name.

Neuropsychological studies of naming and recognition suggest that these two processes may be subserved by partially distinct brain networks. Some evidence for the neural separation of the processes involved in lexical retrieval and recognition comes from category-related deficits that are limited to the verbal modality, such as the famous case of patient 'CP.' CP was selectively impaired in the naming of artifacts (man-made objects such as tools and utensils). Although CP was able to recognize objects, he could not name them, indicating a lexical rather than conceptual deficit. Interestingly, CP could recognize and name objects from other stimulus categories (e.g., familiar faces and animals). Additional evidence compatible with the neural separation between conceptual and lexical retrieval processes comes from research with patients suffering from semantic dementia, who often have bilateral temporal atrophy. Studies have shown that these patients are impaired in both naming and recognition, whereas patients with damage limited to the temporal region tend to be impaired in naming but not in recognition. Additional evidence for the neural separation between the systems subserving lexical retrieval and recognition comes from studies which examined lesion overlap in patients who showed only naming problems and patients who showed both naming and recognition problems.

Relative to the main types of aphasia syndromes, pure anomia in the absence of additional speech or linguistic deficits is fairly uncommon. It occurs far less frequently than most of the 'classic' aphasia syndromes, such as Broca's aphasia and Wernicke's aphasia. For those cases for which lexical impairments occur as an isolated manifestation of acquired brain dysfunction, the designation of 'anomic aphasia'

applies. Studies of patients with anomic aphasia have identified distinct brain structures that are specialized for different categories of concrete entities (faces, landmarks, animals, tools, and fruits and vegetables) as well as for different types of words (i.e., words referring to concrete objects, actions, and spatial relationships). The classic aphasia syndromes are associated with brain damage in the vicinity of the Sylvian fissure in the left hemisphere (the left hemisphere is dominant for language in the vast majority (approximately 98%) of right-handed individuals and in the majority (approximately 70%) of left-handed individuals). Isolated defects in naming, however, which define anomic aphasia, are associated primarily with damage to structures in the left hemisphere outside the classic language regions. Specifically, anomic aphasia is most often caused by damage to the left anterior temporal lobe, the inferior and lateral aspect of the left temporal lobe, or the left occipitotemporal junction. Scientific investigations of patients with anomic aphasia, using modern neuroanatomical and neuropsychological techniques, have revealed a number of intriguing associations between specific brain structures and specific types of naming abilities. Studies in normal subjects, using functional neuroimaging procedures (positron emission tomography and functional magnetic resonance imaging), have corroborated several of these findings.

Category-specific lexical impairments following acquired brain injury have been observed and investigated by several researchers. With regard to the retrieval of words for various categories of concrete objects, work conducted in our laboratory has led to several general conclusions. First, there is strong left hemisphere specialization, as would be expected for word retrieval processes (although category-related word retrieval deficits have been reported in connection with right hemisphere lesions as well). Second, there is consistent evidence for category relatedness in the arrangement of systems that support word retrieval for different categories of concrete entities. Lesion studies have shown that defective naming of unique persons and unique landmarks is associated with lesions in the left temporal pole. Defective naming of animals is associated with lesions in left anterior inferior temporal (IT) areas, anterior insula, and the dorsal temporo-occipito-parietal junction. Defective naming of tools is associated primarily with lesions in the left posterior lateral temporo-occipito-parietal junction and with lesions in the inferior pre- and post-central gyri and the insula. Defective naming of fruits and vegetables is associated with lesions in left inferior pre- and post-central gyri and anterior insula. Defective naming of musical instruments is associated with lesions in the left temporal polar region and anterior IT, the posterolateral temporal region (close to the motion-related area known as MT), the insula, and the inferior pre- and post-central gyri. The degree of segregation for different categories varies, and is more or less defined, depending on the categories being contrasted. For example, the association between retrieval of proper names (persons and landmarks) and the left temporal polar region has been particularly strong, as has the association between retrieval of tool names and the left posterior lateral temporo-occipito-parietal junction. However, categories such as musical instruments and fruits and vegetables have not revealed highly localized and reliable neural correlates. Convergent findings have been reported by a number of other laboratories.

In addition to the category-specific distinctions for concrete entities, there are also intriguing distinctions between different types of words. For example, the brain regions that are important for retrieving words for concrete objects (nouns) are partially separate from those that are important for retrieving words for actions (verbs). As noted previously, the retrieval of words for actions is associated with neural structures in the frontal operculum (in front of the Rolandic sulcus) of the left frontal lobe. Researchers in our laboratory have conducted several studies focused on the neural correlates of retrieving words for actions. Specifically, these studies attempted to determine if the neural systems subserving lexical retrieval for objects and actions were segregated. They also investigated whether lexical retrieval for actions involved not only structures of the left inferior frontal gyrus, particularly the frontal operculum, but also intermediary structures located in left premotor/prefrontal areas. One such study used the lesion method and involved 75 subjects with focal, stable lesions in the left or right hemisphere. The experimental tasks were standardized procedures for measuring action and object naming. Results from the study indicated that lesions of the left anterior temporal and inferotemporal regions, which produce impairments in naming of concrete entities, did not cause action-naming deficits. Additionally, lesions associated with impaired action naming overlapped maximally in the left frontal operculum and in the underlying white matter and anterior insula. A follow-up analysis indicated that action-naming impairments, especially when they were disproportionate relative to concrete entity naming impairments, were associated with not only premotor/prefrontal lesions but also lesions of the left mesial occipital cortex and those of the paraventricular white matter underneath the supramarginal and posterior temporal regions.

Interestingly, some studies have suggested that noun–verb homophones (e.g., words such as 'hammer'

or 'duck', which are used frequently as either nouns or verbs) are retrieved by the brain system that fits the context in which the word is being used. For example, if 'hammer' is being used as a noun, the temporal lobe system will be used, but if 'hammer' is being used as a verb, the frontal lobe system will be used. Such dissociations may appear rather curious on the surface, but there are compelling explanations and theories of why the brain has organized knowledge in different regions to subserve words from different grammatical categories. For example, factors such as whether an entity is unique (e.g., 'Tom Hanks') or nonunique (e.g., a screwdriver), whether it is living (e.g., pig) or nonliving (e.g., a hammer), whether it is manipulable (e.g., a wrench) or nonmanipulable (e.g., a giraffe), or whether it makes a distinctive sound (e.g., a rooster) or not (e.g., a thimble) are important in determining which neural structures will be used in the mapping and retrieval of knowledge for entities, including their names. The dissociation between deficits in naming objects and deficits in naming actions is also confirmed in studies of patients with degenerative conditions, such as variants of primary progressive aphasia. Also, patients with semantic dementia (SD) have been shown to be impaired in their semantic knowledge concerning objects. The speech of patients with SD is fluent but dramatically marked by anomia. Interestingly, SD is characterized by prominent atrophy in anterolateral temporal cortex (one of the brain structures that, when damaged, has been associated with category-specific lexical impairments for concrete entities) of either the left or both the left and the right hemispheres. Patients with progressive nonfluent aphasia (PNFA), on the other hand, tend to be more impaired in action rather than object naming. Their speech is nonfluent and effortful, with grammatical and phonological errors. PNFA is also characterized by neuronal loss most prominent in inferior frontal and superior temporal regions. These findings are in general agreement with the findings from focal lesion patients, as noted previously.

Very little research has explored which neural systems may be important for retrieving words for spatial relationships (locative propositions such as 'in,' 'on,' and 'around'). Our laboratory has carried out exploratory work on this domain. In one lesion study, the hypothesis that processing the meanings of locative prepositions depends on neural structures in the left inferior prefrontal cortex and left inferior parietal cortex was tested. Seventy-eight subjects with focal, stable lesions to various parts of the telencephalon and a comparison group of 60 normal participants performed tasks that required production, comprehension, and semantic analysis of locative prepositions. In support of the hypothesis, it was found that in subjects with impaired knowledge of locative prepositions, the highest region of lesion overlap was in the left frontal operculum and the left supramarginal gyrus, and in the white matter subjacent to these two areas.

A second study on six patients who had pervasive defects for locative prepositions confirmed that such defects were associated specifically with damage to the posterior left frontal operculum, white matter subjacent to this region, and white matter underneath the inferior parietal operculum. These patients did not have basic impairments in spatial processing or working memory, and they had relatively well-preserved conceptual knowledge for actions and various categories of concrete entities (e.g., persons, animals, and tools). All six patients, however, had defects in naming actions, and some also had defective naming of some categories of concrete entities. Overall, the findings converge well with results from functional imaging, and with results from classic cases in the aphasia-based literature, and suggest that the left inferior prefrontal and left inferior parietal regions have crucial, albeit not exclusive, roles in processing knowledge associated with locative prepositions.

The anomic deficits described throughout this article can occur in connection with any sensory modality – for example, when a person is attempting to name a picture of something, a sound, a smell, or something that is felt by the hand. Anomia can occur in the course of verbal discourse, such as when one is speaking and suddenly cannot retrieve the name for a particular concept that is part of the intended utterance. The majority of scientific inquiries into the phenomenon of anomia, however, have focused on the visual modality and have used paradigms in which subjects are presented pictures (or actual objects) and asked to name them. This format, known as 'visual confrontation naming,' is also the standard paradigm for assessing naming in patients with brain injuries. As a consequence, most of our current knowledge regarding the brain underpinnings of word retrieval and most theoretical accounts of this process are heavily tied to the visual modality. However, evidence suggests that the modality in which a stimulus is perceived may not make much difference. For example, retrieval of the name 'rooster' when confronted with a picture of a rooster, or when confronted with the characteristic sound of a rooster, appears to depend on the same left temporal lobe region.

For the majority of individuals, those who have not acquired a brain injury or do not suffer from a neurodegenerative disorder, lexical retrieval comes naturally and effortlessly. However, the relative ease with which most of us are able to provide names for various objects, actions, and entities is somewhat

deceptive. There are a number of neural systems involved in the recognition and retrieval of appropriate word forms that, when disrupted by brain injury, can produce a variety of deficits. Lesion studies of patients with anomic aphasia have provided a unique opportunity to learn how the brain operates the processes associated with word retrieval. Results from these studies have identified different brain structures that are specialized for different types of words and different categories of entities. Future research on patients with anomic aphasia will continue to shed light on how the human brain operates language processes and, it is hoped, will help to inform rehabilitation efforts aimed at patients with acquired disturbances of naming.

See also: Aphasia: Sudden and Progressive; Brain Trauma; Broca's Area: Evolution; Dementia and Language; Language Development; Language Following Congenital Disorders (not SLI); Language: Cortical Processes; Language: Aphasia; Language: Learning Impairments; Word Recognition; Word Learning.

Further Reading

Burke DM, MacKay DG, Worthley JS, and Wade E (1991) On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language* 30: 542–579.

Carammaza A and Hillis A (1991) Lexical organization of nouns and verbs in the brain. *Nature* 349: 788–790.

Damasio H, Grabowski TJ, Tranel D, Hichwa R, and Damasio A (1996) A neural basis for lexical retrieval. *Nature* 380: 499–505.

Dell GS, Schwartz MF, Martin N, Saffran EM, and Gagnon DA (1997) Lexical access in aphasic and nonaphasic speakers. *Psychological Review* 104: 801–838.

Hillis AE and Caramazza A (1995) Representations of grammatical categories of words in the brain. *Journal of Cognitive Neuroscience* 7: 396–407.

Kemmerer D (2005) The spatial and temporal meanings of English prepositions can be independently impaired. *Neuropsychologia* 43: 797–806.

Kemmerer D and Tranel D (2003) A double dissociation between the meanings of action verbs and locative prepositions. *Neurocase* 9: 421–435.

Knibb JA, Xuereb JH, Patterson K, et al. (2006) Clinical and pathological characterization of progressive aphasia. *Annals of Neurology* 59: 156–165.

Levelt WJM, Roelofs A, and Meyer AS (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22: 1–75.

Martin A, Wiggs CL, Ungerleider LG, and Haxby JV (1996) Neural correlates of category-specific knowledge. *Nature* 379: 649–652.

Pulvermüller F (1999) Words in the brain's language. *Behavioral and Brain Sciences* 22: 253–336.

Tranel D (2006) Impaired naming of unique landmarks is associated with left temporal polar damage. *Neuropsychology* 20: 1–10.

Tranel D, Adolphs R, Damasio H, and Damasio AR (2001) A neural basis for the retrieval of words for actions. *Cognitive Neuropsychology* 18: 655–670.

Tranel D, Damasio H, and Damasio AR (1997) A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia* 35: 1319–1327.

Tranel D, Grabowski TJ, Lyon J, and Damasio H (2005) Naming the same entities from visual or from auditory stimulation engages similar regions of left inferotemporal cortices. *Journal of Cognitive Neuroscience* 17: 1293–1305.

Tranel D and Kemmerer D (2004) Neuroanatomical correlates of locative prepositions. *Cognitive Neuropsychology* 21: 719–749.

LIM Kinase and Actin Regulation of Spines

Z Jia, Z Todorovski, Y Meng, S Asrar, and L-Y Wang, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The dendritic spine is the major postsynaptic site of glutamatergic synapses and is thought to be important for synaptogenesis, synaptic regulation, and cognition. Spine changes and deficits are associated with a wide range of physiological and pathological conditions, including synaptic plasticity and mental disorders. For the last decade, a variety of techniques, including imaging, molecular perturbations, and genetic analyses, have been employed to investigate spine properties and their relation to synaptic regulation. As a result, a large number of molecules have been identified to regulate spine formation and morphology. Significantly, despite the diversity of these molecules, they appear to converge to affect signaling processes involved in the regulation of the actin cytoskeleton, the key structural component of the spine (Figure 1). In this article, we discuss the role of the LIM (Lin-11, Isl-1, and Mec-3) domain containing protein kinase (LIMK) family in the regulation of the neuronal actin network and structures. Although the members of the LIMK family are also present and important for presynaptic terminals, we focus our discussion on the postsynaptic specialization and functional significance, including spine morphogenesis, postsynaptic structure, synaptic plasticity, and cognition.

The Dendritic Spine, Synaptic Regulation, and Mental Diseases

In the mammalian central nervous system (CNS), glutamate is the major excitatory neurotransmitter that exerts its effects by activating various glutamate receptors, including *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), and metabotropic G-protein-coupled receptors (mGluRs). The vast majority of excitatory synapses are formed on the head of the dendritic spine, a highly specialized dendritic protrusion where the glutamate receptors and associated molecules are localized. Therefore, it is not surprising that spine properties (e.g., spine density, morphology, and plasticity) have profound effects on synaptic structures and functions. For example, the size and geometry of the spine are linked to the numbers of synaptic glutamate receptors, Ca^{2+} dynamics, and the ability to undergo bidirectional changes in the efficacy of synaptic transmission. This includes hippocampal long-term potentiation (LTP) and long-term depression (LTD), the most extensively studied forms of long-lasting synaptic plasticity thought to be critical for proper neuronal connections and memory formation. Conversely, synaptic activities and learning experience are also found to regulate spine density, shape, and motility. These observations suggest that spines are essential and integral components of synaptic regulation and cognitive processes.

Consistent with their role in synaptic and cognitive regulation, both human and animal studies have suggested that spines may be a structural target of various mental disorders and neurological diseases. For example, morphological studies on patients with mental retardation (MR) revealed profound deficits in spine properties (e.g., reduction in spine density and abnormal spine morphology) in the absence of gross changes in the CNS anatomy. In addition, animal models of MR and drug addiction exhibit changes in spine development, maturation, and morphogenesis. Although it is not clear whether spine deficits are a primary cause or a secondary consequence of these mental disorders, human molecular genetic studies support the hypothesis that spines are among the main targets of genetic abnormalities. For example, an increasing number of MR genes, including the LIMK-1 gene, have been identified as encoding putative proteins potentially involved in the signal transduction pathways. These are mediated by the Rho family small GTPases, which are central mediators of actin dynamics and cell morphology. Given the predominant role of the actin cytoskeleton in the structural properties of the spine, it is likely that MR gene mutations result in abnormal Rho signaling that directly targets spines.

LIM Kinase 1 and Actin

The LIMK family is a unique group of serine/threonine/tyrosine protein kinases with two N-terminal LIM domains, a PDZ (PSD-95, Dlg, ZO-1) domain, and a C-terminal kinase catalytic domain. Both LIM and PDZ domains have been found in various proteins that are believed to mediate protein–protein interactions, but the role of these protein modules associated with LIMK has not been clearly defined. Although two members of the LIMK family (LIMK-1 and LIMK-2) have been identified and found to be expressed in the mammalian CNS, most studies related to neuronal structures and functions have focused on LIMK-1. However, because of their

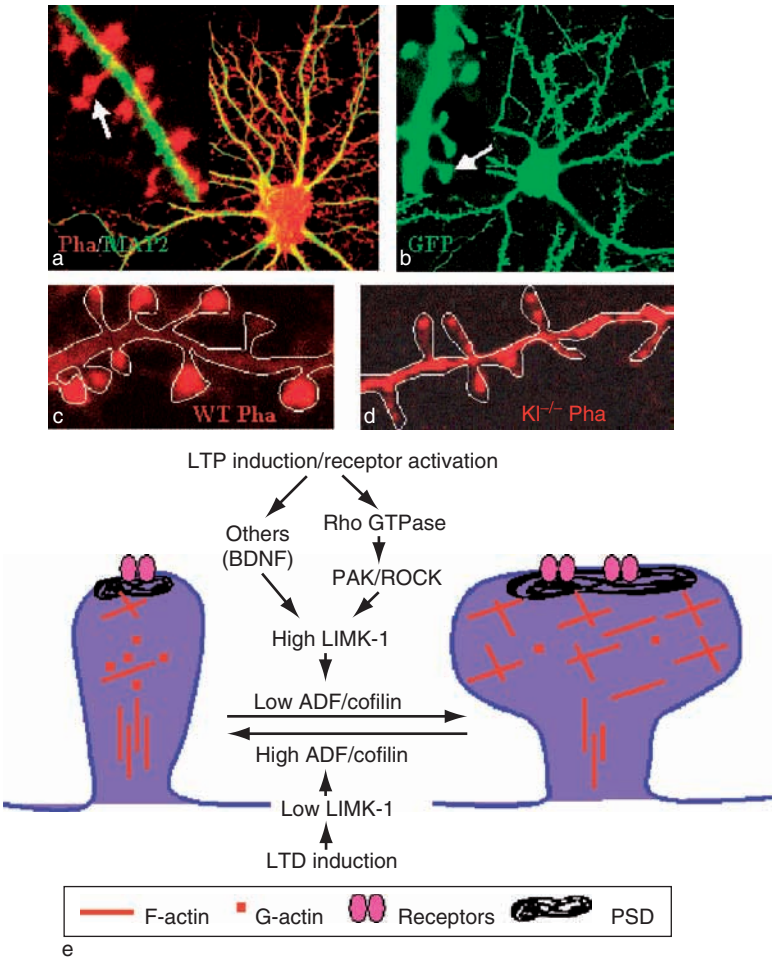

Figure 1 Regulation of spines and synaptic function by LIMK-1: (a) cultured hippocampal neurons stained with the F-actin dye phalloidin (red) and an antibody against microtubule-associated protein (MAP)2 (green), showing mushroom-shaped spines (arrows) with intense F-actin presence but a lack of microtubules; (b) cultured hippocampal neurons transfected with green fluorescent protein (GFP), showing similar spine morphology as revealed in (a) by phalloidin staining; (c) representative dendritic branch of cultured hippocampal neurons prepared from wild-type (WT) mice, showing distinct spine heads with significantly higher F-actin staining than adjacent dendritic shafts; (d) representative dendritic branch of cultured hippocampal neurons prepared from LIMK-1 knockout ($KI^{-/-}$) littermates, showing a lack of distinct spine heads and abnormal accumulations of F-actin in the dendritic shaft; (e) a hypothetical model for LIMK-1 regulation of spine morphology and synaptic function. In this model, the protein level or the activity of LIMK-1 can be regulated by various signals, including synaptic activity. For example, LIMK-1 could be reciprocally regulated by LTP and LTD induction. High LIMK-1 during LTP results in spine F-actin accumulations and spine enlargement, whereas low LIMK-1 causes actin depolymerization and spine shrinkage. ADF, actin depolymerization factors; BDNF, brain-derived neurotrophic factor; GTPase, guanosine-5′-triphosphate binding protein; LIMK, LIM kinase; LTP, long-term potentiation; PAK, p-21 activated kinase; PSD, postsynaptic density; ROCK, Rho kinase.

similar biochemical properties and regulation, it is expected that at least some of the known neuronal functions found in LIMK-1 studies can also be extended to LIMK-2.

Consistent with cell line studies, LIMK-1 is a potent regulator of the actin cytoskeleton in the mouse CNS. Thus, the disruption of LIMK-1 causes profound deficits in actin properties. In mature pyramidal neurons, actin is the predominant cytoskeletal protein in the dendritic spine. Therefore, cycling between polymerized actin filaments (F-actin) and monomeric actin (G-actin) is the key determinant of a stable spine structure as well as its dynamic changes, including spine motility and plasticity. As shown in Figure 1(a), the F-actin dye phalloidin staining (red) is extremely intense in the spine, which essentially lacks microtubules (green) that are abundant in the dendritic shaft. The importance of the actin cytoskeleton in neuronal structure and function has been demonstrated in numerous studies. Direct perturbations of actin dynamics using pharmacological reagents significantly affect a wide range of neuronal properties, including spine formation, synaptogenesis, spine motility, and synaptic plasticity. Although not clearly defined in molecular organization and functional significance, spine F-actin appears to exist in

at least three forms, each apparently associated with distinct spine compartments or organelles. First, a network of actin filaments is present throughout the spine head, some of which are in contact with a distinct electron-dense region called the postsynaptic density (PSD). This fraction of actin filaments may serve as a major structural component of the spine and is probably critical for proper spine morphology. Second, a pool of F-actin is physically contained in the PSD, possibly serving as anchoring proteins for the PSD components, including glutamate receptors. For example, NMDA receptors can directly interact with the actin-binding protein α-actinin, which may stabilize the former by linking them to the actin cytoskeleton. In addition, this pool of actin may directly participate in the regulation of the channel properties of the receptors and their signaling to the spine cytosol. Finally, a more stable class of actin filaments are found along the neck of the spine. LIMK-1 appears to be particularly important for the first two pools of F-actin because LIMK-1 knockout neurons display a significant reduction in the amount of F-actin at the spine head but not at the neck of the spine. In fact, abnormal accumulations of F-actin clusters in knockout mice are significantly enhanced, both at the base of the spine neck and in adjacent dendritic shafts. It is possible that separate signaling processes independent of LIMK-1 are responsible for maintaining dendritic actin filaments. These results suggest that LIMK-1 is of special significance to the regulation of actin reorganization in response to spine and synaptic activities.

LIM Kinase 1 and Spine Morphogenesis

In accordance with its regulation of spine actin, perhaps the most clearly demonstrated role of LIMK-1 is the regulation of spine structure, specifically the morphology of the spine. Although spines are very heterogeneous in size and shape, as well as being highly dynamic (they can be easily altered by additional factors, including experimental conditions and the age of spine development), most of them have an enlarged head and a thin constrictive neck (i.e., mushroom-shaped; **Figure 1(a)**). This shape is well conserved both in cultured neurons and in the *in vivo* brain and appears to require functional LIMK-1. Thus, hippocampal pyramidal neurons from LIMK-1 knockout mice exhibit filopodium-like spine morphology without a distinct spine head. These types of spines are also characteristic of developing or immature neurons, suggesting that LIMK-1 is required for spine maturation or maintenance. Conversely, increasing LIMK-1 expression leads to a significant enlargement of the spine heads. Therefore, LIMK-1 appears to be specific and sufficient to govern bidirectional regulation in spine morphogenesis or spine maturation. Consistent with its role in spine regulation, LIMK-1 potently affects the postsynaptic structure. Mammalian asymmetric excitatory synapses are characterized by the presence the PSD and an opposing presynaptic terminal containing numerous vesicles. Although these structures are clearly discernible in LIMK-1 knockout mice, the size of PSD is significantly reduced compared to that of the wild-type animals. The correlated reductions in spine and synapse sizes are consistent with the idea that proper spine morphology or maturation is critical for normal synaptic structure. These structural deficits probably contribute to the deficits in synaptic transmission and plasticity associated with the LIMK-1 knockout mice.

LIM Kinase 1 and Synaptic Regulation

LIMK-1 appears to be important for both basal synaptic transmission and synaptic plasticity. For example, LIMK-1 knockout mice exhibit elevated levels of basal excitatory synaptic responses in the CA1 region of the hippocampus. These changes appear to be inconsistent with structural alterations that predict a reduction in synaptic response. However, it remains to be investigated if the molecular compositions of the spine and synapse (such as synaptic glutamate receptors) are affected in the knockout mice. Other factors, including presynaptic neurotransmitter release, may also contribute to the changes observed in the basal synaptic strength of knockout mice. In support of this possibility, several important studies showed that perturbations of actin dynamics affect the release of both the readily releasable and more distal reserve pool of synaptic vesicles. The effect of LIMK-1 on long-lasting synaptic plasticity is also complex. In the CA1 region of the hippocampus, brief high-frequency tetanic stimulation (e.g., 5–100 Hz) induces a persistent increase in the excitatory synaptic transmission or LTP. CA1-LTP is of particular interest and is functionally significant compared to other forms of synaptic plasticity because it shares some of the important properties with learning and memory, including associativity and specificity, therefore providing a potential mechanism for information processing and storage. Although much remains to be investigated regarding the molecular mechanisms underlying CA1-LTP, it is generally agreed that the tetanic stimulations (or other conditioning protocols) transiently activate NMDA receptors, resulting in a significant increase in Ca^{2+} ions in the postsynaptic spine. A series of signaling processes are then initiated, which eventually lead to enduring enhancements of synaptic responses

predominantly mediated by AMPA receptors. In addition, at least two distinct phases of LTP can be generated depending on the induction protocols: an early phase (E-LTP) that is induced by a relatively short and weak induction protocol and thought to be accomplished by modifications of existing synapses, such as increased numbers and/or channel conductance of synaptic AMPA receptors, and a late phase (L-LTP) that is produced by a more sustained conditioning (e.g., multiple trains of 100 Hz stimulation) and believed to require new gene expression, protein synthesis, and, ultimately, the formation of new spines and synapses. E-LTP and L-LTP are functionally important because they are thought to underlie short- and long-term memory, respectively. LIMK-1 regulation of CA1-LTP appears to be both phase- and frequency-dependent. For example, applications of 5–10 Hz stimulations are able to induce a form of E-LTP in wild type but not in LIMK-1 knockout mice. On the other hand, E-LTP induced by a stimulation protocol of one train of 100 Hz is dramatically enhanced in the knockout mice compared to wild-type controls. Interestingly, LTP induced by multiple trains of 100 Hz stimulation exhibits a faster decay in the knockout mice, suggesting a deficit in L-LTP. Although these results indicate that LIMK-1 regulates multiple forms of LTP, the underlying mechanisms may be completely related to actin and spine properties. It is possible that changes in spine morphology may have nonspecific effects on the molecular composition (e.g., glutamate receptors) and the signaling processes of the spine that can, in turn, influence LTP. For example, spines with a small head may reach critical Ca^{2+} concentrations more easily, and therefore, LTP could be induced more readily in these spines. It has been shown that spine Ca^{2+} compartmentalization and diffusion to the dendritic shafts are related to the spine shape. It is therefore important to study basal and activity-dependent Ca^{2+} dynamics in LIMK-1 knockout neurons. In addition, changes in the amount of F-actin may directly affect LTP processes independent of spine morphology. For example, protein trafficking at the synapse, a key mechanism thought to be responsible for changing the number of synaptic glutamate receptors during, E-LTP, is regulated by actin dynamics. Finally, deficits in actin dynamics may prevent structural reorganization, including the formation of new spines and synapses thought to be critical for L-LTP.

LIM Kinase 1 and Cognition

The significant role of LIMK-1 in cognitive processes is evident in that LIMK-1 knockout mice exhibit altered performances in several learning and memory tasks. For example, in a fear-conditioning paradigm based on electric foot shocks (which provides a robust associative learning test in mice), the LIMK-1 knockout mice show elevated and more persistent freezing responses in both cued and contextual tests 24 h following training, suggesting enhanced performances in these tasks. However, in a water maze test designed to evaluate spatial learning and memory in rodents, although LIMK-1 knockout mice were comparable to controls in initial learning acquisitions, they were significantly impaired in locating the platform in a reverse learning paradigm (a further test following the initial learning that alternates the locations of the platform during training). Although these results clearly indicate abnormalities in at least certain aspects of cognitive processes in LIMK-1 knockout mice, further behavioral analyses are needed to determine what specific forms of learning and memory are affected. For example, although LIMK-1 knockout mice display an increased amount of freezing 24 h following training, it is not known whether such an enhancement occurs immediately after training and how long it persists. Given the deficits in actin properties and E-LTP, it is possible that these mice have impairments in longer-term memory in which structural changes are considered to be critical. Further assessment of other behavioral responses in LIMK-1 knockout mice is therefore important.

It is known that one copy of the LIMK-1 gene, along with several others, is deleted in patients with Williams syndrome, a developmental condition with an intriguing cognitive profile, including severe deficits in visual-spatial cognition in the presence of well-developed language and social skills. The role of LIMK-1 in various aspects of this mental disorder is unclear, although the results from the LIMK-1 knockout mice are consistent with the hypothesis that dysfunctional LIMK-1 contributes to the cognitive deficits of this disease.

Molecular Mechanisms Underlying LIM Kinase 1 Actions

The most extensively studied molecular targets of the LIMK family are members of the actin depolymerization factors (ADF)/cofilin family. These are actin-binding proteins whose functions include severing actin filaments and promoting actin depolymerization by binding to the pointed end of F-actin. The activity of ADF/cofilin is stringently regulated by protein phosphorylation on serine 3. Thus, in vertebrates, serine 3 phosphorylation inactivates ADF/cofilin proteins and completely abolishes their ability to bind to and disassemble actin filaments. Only

recently, the specific protein kinases that are responsible for ADF/cofilin phosphorylation have been revealed to be LIMK-1. Although additional proteins have since been identified that are also capable of phosphorylating ADF/cofilin, LIMK-1 is perhaps the predominant protein in phosphorylating and regulating ADF/cofilin activity in the mammalian CNS. Thus, the phosphorylated forms of ADF/cofilin are diminished in LIMK-1 knockout brain tissues and neurons. Furthermore, LIMK-1 appears to be a key mediator of activity-dependent changes in ADF/cofilin phosphorylation and dephosphorylation cycling. In brain slices and neurons, the amount of phosphorylated ADF/cofilin undergoes rapid alterations in response to a wide range of external stimuli, including neurotrophic factors and the activation of glutamate receptors. In the absence of LIMK-1, these stimulus-induced changes in phosphorylated ADF/cofilin are no longer present. For example, glutamate treatment of hippocampal slices or cultured hippocampal neurons causes a rapid reduction in the amount of phosphorylated ADF/cofilin without changes in total ADF/cofilin protein level, and this reduction is abolished in LIMK-1 knockout mice. Given the importance of ADF/cofilin in regulating actin dynamics, it is likely that the effects of LIMK-1 on the actin cytoskeleton and structural properties are mediated by changes in ADF/cofilin activity. In agreement with this possibility, several recent studies have shown that direct manipulations of ADF/cofilin perturb both basal and activity-dependent bidirectional changes in spine morphology as well as synaptic plasticity.

In addition to ADF/cofilin, a number of additional synaptic proteins have been shown to directly interact with LIMK-1, and therefore potentially mediate the effects of LIMK-1. These include protein kinase C (PKC), neuregulins, and cyclic adenosine-3′, 5′-monophosphate (cAMP) responsive element-binding protein (CREB). Although all these proteins are known to directly regulate various aspects of synaptic transmission and plasticity, the functional significance of their interactions with LIMK-1 remains to be elucidated. For example, are the activities of these proteins regulated by LIMK-1? It is important to emphasize that LIMK-1 knockout mice are significantly impaired in L-LTP, where the transcription factor CREB is thought to be critical in promoting activity-dependent gene expression. The activity of CREB is regulated by protein phosphorylation at serine 133 by multiple protein kinases, including LIMK-1. Therefore, it is possible that LIMK-1 regulates L-LTP through phosphorylating and activating CREB. It is of considerable interest to determine whether CREB activity or distribution is altered in LIMK-1 knockout mice.

Significance of LIM Kinase Signaling

Among synaptic proteins, the Rho family GTPases (RhoA, Rac1, and Cdc42) are regarded as the central regulators of actin dynamics in response to various stimuli, including neuronal activities. They are small guanine nucleotide-binding proteins that function as molecular switches cycling between an active GTP-bound state and an inactive GDP-bound state. The relative ratios of these two states dictate the activity of the Rho GTPases, and they are regulated by multiple proteins, including GAPs (GTPase activating protein) and GEFs (guanine nucleotide exchange factors), many of which have been found to interact with and to be regulated by various receptor proteins. The critical role of Rho signaling in spine and synaptic regulation is supported by many recent studies in which manipulations of Rho proteins or their regulators drastically affected spine development and morphogenesis, synaptic plasticity, and memory formation. Although the effects of Rho proteins (including those on spine properties) have been shown to be diverse and complex, it is generally thought that these effects are achieved through changes in the actin cytoskeleton and that LIMK-1 provides an important molecular link between Rho activation and actin. At least two families of protein kinases, p-21 activated kinases (PAKs) and Rho kinases (ROCKs), have been identified and shown to interact with active forms of Rho GTPases. Significantly, both PAKs and ROCKs can directly phosphorylate and stimulate the kinase activity of the LIMK family. Therefore, a sequential protein phosphorylation pathway consisting of PAK/ROCK-LIMK-ADF/cofilin has been proposed as a key signaling mechanism between the Rho GTPases and actin dynamics (Figure 1(e)). Consistent with this model, manipulations of LIMK-1 have shown to be able to block various effects on actin and structural properties induced by Rho activation in neurons and other cell types. Because Rho proteins can be regulated by many channels and receptors, including glutamate receptors, LIMK-ADF/cofilin may serve as a common molecular target for a variety of stimuli to regulate the actin network.

Because LIMK-1 is also a potential target of other synaptic proteins that have no apparent involvement in Rho signaling, it may also be responsible for Rho-independent actin reorganization at the synapse. As discussed earlier, both PKC and neuregulins interact with LIMK-1, and therefore some of their synaptic functions, particularly those related to the actin cytoskeleton, could be mediated by LIMK-1. Consistent with this possibility, the effect of PKC activators on

synaptic responses is dramatically reduced in LIMK-1 knockout mice, indicating that LIMK-1 is required for at least some of the actions of PKC. Further dissection of the effects of PKC mediated by LIMK-1 is required.

Finally, changes in LIMK-1 protein level may be responsible for a wide array of activity- or experience-dependent structural changes. For instance, a recent study demonstrated that a brain-specific microRNA (miR-134) negatively regulates the size of dendritic spines and this regulation is mediated by inhibiting the dendritic translation of LIMK-1 mRNA. Thus, overexpression of miR-134 results in a reduction of both the amount of the LIMK-1 protein and spine volume. Conversely, the inhibition of endogenous miR-134 leads to a significant increase in both the LIMK-1 protein and spine size. Moreover, this miR-134-regulated LIMK-1 protein synthesis appears to be responsible for the effect of brain-derived neurotrophic factor on spine development and maturation. Changes in LIMK-1 protein levels have also been found to be associated with other conditions, including drug administration and learning experience.

Concluding Remarks

Accumulating evidence indicates that LIMK-1 is a key synaptic molecule in regulating spine and synaptic properties. Although LIMK-1 appears to mainly target the actin cytoskeleton through directly phosphorylating and inactivating ADF/cofilin, a number of other synaptic proteins have been shown to interact with LIMK-1. It is important to investigate the significance and underlying mechanisms of these protein interactions and to determine their relative contributions to spine and synaptic regulation. Studying the molecular details of LIMK-1 regulation is expected to greatly contribute to the better understanding of the fundamental mechanisms responsible for synapse formation and function.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Cytoskeleton in Plasticity; Dendritic Spine History; Postsynaptic Specialization Assembly; Rho GTPases and Spines; Spines and Mental Disorders.

Further Readings

Chernoff J (1999) Close encounters of the LIM-kinase. *Nature Cell Biology* 1: E115–E117.

Dillon C and Goda Y (2005) The actin cytoskeleton: Integrating form and function at the synapse. *Annual Review of Neuroscience* 28: 25–55.

Lisman J (2003) Actin's action in LTP-induced synapse growth. *Neuron* 38: 361–365.

Matus A (2000) Actin-based plasticity in dendritic spines. *Science* 290: 754–778.

Meng JS, Meng YH, Hana A, Janus C, and Jia ZP (2005) Abnormal long-lasting synaptic plasticity and cognition in mice lacking the mental retardation gene Pak3. *Journal of Neuroscience* 25: 6641–6650.

Meng YH, Zhang Y, Tregoubov V, et al. (2002) Abnormal spine morphology and enhanced LTP in LIMK-1 knockout mice. *Neuron* 35: 121–133.

Meng YH, Zhang Y, Tregoubov V, Falls DL, and Jia ZP (2003) Regulation of spine morphology and synaptic function by LIMK and the actin cytoskeleton. *Reviews in the Neurosciences* 14: 233–240.

Sarmiere PD and Bamburg JR (2002) Head, neck, and spines: A role for LIMK-1 in the hippocampus. *Neuron* 35: 3–5.

Lipids and Membranes in Brain Aging

W G Wood, University of Minnesota School of Medicine, Minneapolis, MN, USA
G Y Sun, University of Missouri, Columbia, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Biological membranes have multiple roles in cells. At a macro level, the membrane acts as a barrier between the cell interior and extracellular environment and between the cytoplasm and intracellular organelles. Viewed at the molecular level, membranes are composed of lipids and proteins and in some instances are in dynamic interaction with each other. How membranes are organized has evolved from the early notion in the 1850s describing the membrane as a relatively thin ion-impermeable layer covering cells and subsequent work showing that such a membrane contains lipids based on lipid solubility experiments. Later evidence in the 1920s suggested a bilayer of lipids with a hydrophobic interior and hydrophilic part of the lipid in contact with the extracellular and intracellular aqueous environment. The idea that proteins may be associated with the membrane was introduced in the 1930s, and later work introduced the concept that the membrane was a single lipid layer with proteins or what was referred to as the unit membrane model. In 1972, Singer and Nicolson proposed the fluid mosaic model of membranes, with lipids and proteins randomly distributed with lateral and rotational diffusion. Being a major component of the membrane, lipids were viewed as being important in maintaining the fluidity of the bilayer. An optimal fluidity was thought to be required for numerous membrane functions, such as transport, signal transduction, exocytosis, and endocytosis. Changes in lipid composition, such as an increase in cholesterol content, would reduce fluidity and alter activity of membrane function, which could be detrimental to the cell. That the state of the membrane was important to cell function has been proposed to be a factor in several different pathophysiological conditions ranging from alcoholism to Zellweger's syndrome. Changes in membrane fluidity and lipid composition have previously been proposed to be major contributors to cellular dysfunction observed with increasing age. This article critically examines aging and brain membranes from the perspectives of changes in fluidity, lipid composition, and membrane domains and identifies areas in which further study is needed.

Brain Membrane Fluidity

Fluidity is a broad term used to describe the lateral motion of lipids in membranes. Several studies have reported a reduction in fluidity of brain membranes with increasing age. However, reduced membrane fluidity in aged membranes has not been consistently observed. The lack of consistency in membrane fluidity and aging studies could possibly be attributable to different techniques (electron spin resonance, fluorescence spectroscopy, and nuclear magnetic resonance), probes residing in different membrane locations (DPH, pyrene, 5-doxyl stearic acid), purity of the membrane preparation (whole brain homogenate, synaptosomes, synaptic plasma membranes in different brain regions), and different species used. Regardless of whether there are aged differences in membrane fluidity, one conclusion regarding the extant studies on fluidity of brain membranes and aging is that most have been limited in their technological sophistication and precision. For example, fluidity actually consists of two components which have been characterized as rate of motion, a dynamic component, and the extent of motion, a static component. Parameters such as limiting anisotropy and order parameter refer to static components, whereas rotational correlation time and rotational rate denote dynamic components. Studies examining brain membrane fluidity in aged organisms have not distinguished between dynamic and static components. Furthermore, the majority of studies on fluidity of brain membranes of aged organisms have used techniques which provide an estimate of the overall average fluidity of the membrane in contrast to more circumscribed areas of the membrane. There may be specific differences in fluidity of the aged brain membrane which would not be detected by conventional techniques, and this issue is discussed later. Another limitation of studies on membrane fluidity and aging is that changes in fluidity on membrane and cell functions in general are not well understood.

Brain Membrane Cholesterol Levels

Lipids are a major contributor to the fluidity of membranes, and this has been demonstrated experimentally in both biological and model membranes. Cholesterol, which accounts for more than 25% of the total lipid of synaptic plasma membranes (SPMs), is a key factor in determining membrane fluidity. Above the phase transition of the lipid environment, reducing membrane cholesterol increases fluidity, whereas the opposite

effect is observed when membrane cholesterol levels are increased. Age differences in brain membrane cholesterol levels have not been consistently observed. This observation applies to both human and animal studies. Even in those studies reporting an increase in cholesterol levels with increasing age, the mechanism for such an increase has not been forthcoming. The biosynthesis of cholesterol requires more than 30 enzymes and several cofactors. To date, there has been very little work on the brain cholesterol synthetic pathway in aging organisms.

Cholesterol is derived from mevalonate, which is also the precursor of various compounds including isoprenoids. There is great interest in the role of isoprenoids such as geranylgeranyl pyrophosphate (GGPP) and farnesyl pyrophosphate (FPP), which prenylate various small GTPases, such as Rho, Rac, and Ras. These proteins are involved in a variety of cell signaling functions. GGPP and FPP levels have not been reported in brain, but an isoprenoid whose abundance has been reported is dolichol. Dolichols are long-chain polyisoprenoids derived from FPP. These compounds are ubiquitous among body organs, with the highest levels observed in testis, kidney, brain, and various exocrine glands. Dolichol can be in a free form or phosphorylated. Dolichol phosphates are intermediates in the synthesis of oligosaccharide groups of glycoproteins. The cellular function of free dolichols is not known. Free dolichol increased fluidity of SPMs, and it has been detected in membranes in which it might function to regulate membrane fluidity. There is a marked increase of free dolichol in human brain tissue with aging and in SPMs of aged mice compared to younger mice. Increased free dolichol content has also been observed in brain tissue from individuals with Alzheimer's disease and individuals with neuronal lipofuscinosis. The precise function of free dolichol and the consequences of elevated dolichol levels observed with increasing age and certain neurodegenerative diseases have not been established.

Brain Membrane Phospholipids and Fatty Acids

In order of abundance, the membrane phospholipids in brain are phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylserine (PS), phosphatidylinositol (PI), and cardiolipin (CL). Neural membranes are also enriched in ether phospholipids (i.e., phospholipids with alkyl or alkenyl linkage in the *sn*-1 position). Whereas most diacyl-type phospholipids are synthesized in the endoplasmic reticulum, ether phospholipids are synthesized in peroxisomes and changes in levels of ether lipids are associated with peroxisomal disorders. Cardiolipin, an unusual phospholipid with two glycerol backbones and four fatty acids, is exclusively present in the mitochondrial membrane. Cardiolipin plays an important role in regulating respiratory chain activity in mitochondria. Phospholipids in brain membranes are enriched in plasmalogens of the ethanolamine type (alkenyl-acyl-gly cerophosphoethanolamine(PEpl)). PEpl is especially enriched in myelin, comprising more than 60% of total ethanolamine phospholipids. Plasmalogens may act as endogenous antioxidants. Although phosphatidylinositol (PI) comprises only 4% of total phospholipid, it is metabolically active and is converted to phosphatidylinositol-4-phosphate (PIP) and phosphatidylinositol-4,5-bisphosphates (PIP2) by kinases. PIP2 plays an important role in the stimulus–signal coupling reactions that lead to generation of second messengers. Stimulation of many G-protein-coupled receptors (GPCRs) leads to activation of phospholipase C (PLC), which hydrolyzes PIP2 to form diacylglycerol (DG), a second messenger for activation of protein kinase C (PKC) and inositol trisphosphate (IP3), a second messenger for mobilization of intracellular Ca^{2+} stores. Another PI kinase, PI3K, adds a phosphate group to the third position of PIP2 to form PIP3. Activation of PI3K by tyrosine kinase and/or kinases associated with GPCRs has been implicated in signaling pathways associated with cell growth, proliferation, survival, and plasticity.

Fatty Acids of Brain Phospholipids

Phospholipids in neural membranes are enriched in polyunsaturated fatty acids such as 20:4n-6 (AA) and 22:6n-3 (DHA). Each type of phospholipid is composed of unique patterns of acyl or alkenyl groups, and these patterns may vary depending on their subcellular localization. For example, myelin is known for its high content of PEpl, and this phospholipid is enriched with monounsaturated fatty acids such as 18:1 and 20:1. In fact, 20:1 is an unusual fatty acid almost exclusively found in the myelin membrane. On the other hand, fatty acids of PEpl in the nerve ending fraction (synaptosome) are enriched in 20:4n-6 and 22:6n-3. Earlier studies showed an age-dependent increase in monounsaturated fatty acids (e.g., 18:1 n-9 and 20:1 n-9 (~43%)) and a decrease in polyunsaturated fatty acids (e.g., 22:6 n-3 (~14%)) in mouse brain PE between 3 and 26 months of age. Similar changes are found in PE in the brain of rhesus monkey. Studies on the glutathione redox state in mouse brain showed an age-dependent decrease in the ratio of reduced to oxidized glutathione (GSH/GSSG) between 3 and 21 months of age. Whether the age-dependent decrease in polyunsaturated fatty acids is

associated with the decrease in GSH remains to be investigated. Since levels of PEpl and 22:6n-3 decrease with normal aging and with memory impairment, considerable interest has focused on whether dietary supplementation of fish oil which is enriched in n-3 fatty acids may help to reverse the age-related deficits and improve memory and cognitive functions.

The fatty acid profile of PS is enriched in 22:6 n-3, which is predominantly coupled with 16:0 in the *sn*-1 position. On the other hand, fatty acids in PI, as well as poly-PI are enriched in 20:4 n-6, which is linked to 18:0 in the *sn*-1 position. Consequently, when PIP2 is hydrolyzed by PLC to form DG and IP3, the fatty acid of this DG pool is enriched in 18:0/20:4 and is different from the DG pools associated with PL biosynthesis. Unlike fatty acids of phospholipids in other organs, brain fatty acids have low levels of 18:2 n-6, although this fatty acid is enriched in CL. The unique presence of CL in mitochondria suggests its role in mitochondrial membrane function. There is evidence that in cerebral ischemia, CL in mitochondria decreases together with release of cytochrome c. Because these anionic phospholipids are present in minute amounts in the brain, little is known about age-dependent changes of these phospholipids.

Enzymes Regulating Membrane Phospholipids

Fatty acids in the *sn*-2 position of phospholipids are maintained in a dynamic equilibrium, regulated by phospholipases A_2 (PLA_2) and lysophospholipid acyltransferases, constituting the deacylation–reacylation cycle. It is important to note that whereas hydrolysis of fatty acids by PLA_2 releases fatty acids from phospholipids, the reacylation process involves ATP-dependent conversion of a fatty acid to its acyl-CoA and transfer of acyl-CoA to lysophospholipids. Although the number of studies on the structure and function of different types of PLA_2 has increased considerably in the past decade, studies regarding lysophospholipid acyltransferases are still lacking. Besides PLC, phospholipids are also hydrolyzed by phospholipase D (PLD), resulting in formation of phosphatidic acid (PA). In the presence of ethanol, PLD may mediate transphosphatidylation of PC to form phosphatidylethanol, which is a marker of ethanol metabolism.

Turnover of fatty acids among different phospholipids is dependent not only on the type of phospholipids but also on the types of membranes in different cell types. Thus, it is conceivable that fatty acids in myelin phospholipids undergo slower turnover compared to synaptic membranes. Age-related decrease in turnover of fatty acids in SPM phospholipids may provide an explanation for the decrease in synaptic functions during aging.

Phospholipase(s) A_2

PLA_2s are enzymes for cleaving fatty acids in the *sn*-2 position of phospholipids. These enzymes are ubiquitous in mammalian cells. Besides their role in maintenance of cell membrane phospholipids, they also play an important role in providing lipid mediators such as arachidonic acid (AA), which serves as the precursor for synthesis of a number of eicosanoids. Hydrolysis of phospholipids by PLA_2s also produces lysophospholipids which possess detergent-like properties and can alter membrane micro domains. Among more than 20 different types of PLA_2s, studies have focused on three groups, namely the calcium- and kinase-dependent group IV cytosolic PLA_2 ($cPLA_2$), the Ca^{2+}-independent group VI PLA_2 ($iPLA_2$), and the Ca^{2+}-dependent secretory PLA_2 ($sPLA_2$).

$cPLA_2$ is composed of three isoforms: α, β, and γ. In brain, most studies have focused on $cPLA_2$-α, an 85-kDa protein comprising a Ca^{2+} binding C2 domain and a catalytic domain with several putative phosphorylation sites for mitogen-activated protein kinase (MAPK), PKC, and Ca^{2+}/calmodulin-dependent protein kinase II. Upon binding with Ca^{2+} and phosphorylation by protein kinases, $cPLA_2$-α is translocated to the membrane and targets phospholipids, especially those containing AA. Consequently, $cPLA_2$-α is activated by metabotropic and ionotropic receptors and releases AA for biosynthesis of eicosanoids. However, the mechanism for the coupling between $cPLA_2$ and cyclooxygenases (COX) remains a subject for further investigation. $cPLA_2$s have been shown to target a number of intracellular membranes, including the Golgi complex, plasma membranes, and nuclear membranes. The putative pleckstrin homology (PH) domain in $cPLA_2$ suggests possible interaction with acidic phospholipids such as PIP2 which are localized in plasma membranes. There is evidence that the translocation process may involve binding to cytoskeletal proteins and intermediate filaments such as vimentin and actin.

Secretory PLA_2s are low-molecular-weight proteins ($\sim$14 kDa). More than ten subtypes of $sPLA_2$s are present in the mammalian system, including the brain. In recent years, considerable attention has been given to the $sPLA_2$-IIA in the cardiovascular system because of its association with inflammatory responses in diseases such as arthritis, atherosclerosis, and sepsis. Indeed, $sPLA_2$ is an independent risk factor for cardiovascular diseases. Although studies have implicated the increase in $sPLA_2$-IIA in reactive astrocytes in response to cerebral ischemia, studies demonstrating $sPLA_2$ in human brain and in relation with other neurodegenerative diseases have not been forthcoming. In astrocytes, mRNA expression for $sPLA_2$-IIA along with COX-2 and inducible nitric oxide synthase are stimulated by

proinflammatory cytokines (tumor necrosis factor-α, interleukin-1β, and interferon-γ) and lipopolysaccharides, and the increase in inflammatory response is coupled with production of prostaglandin E_2 and nitric oxide.

Since $sPLA_2$-IIA is induced by proinflammatory cytokines, it needs to be determined whether its release from glial cells can cause neuronal damage. Studies using extracellular forms of $sPLA_2$ have clearly demonstrated the ability of $sPLA_2$-IIA to cause neuronal excitotoxicity and apoptosis by enhancing the ionotropic receptor and Ca^{2+} channel activities. Other studies have shown that secreted $sPLA_2$-IIA can perturb cellular membranes, especially those undergoing apoptosis. In PC12 cells, lysophospholipids produced by $sPLA_2$-IIA were shown to enhance neurite outgrowth, further implicating an important role of this enzyme in neurodegenerative diseases.

$iPLA_2$ consists of two major subtypes, VIA and VIB. $iPLA_2$-VIA comprises 85–88 kDa proteins with three splice variants – VIA-1, VIA-2, and VIA-3 – all containing multiple ankyrin domains, suggesting its ability to interact with other proteins. $iPLA_2$-VIA shows no apparent specificity for the fatty acids in the *sn*-2 position or the type of phospholipid headgroups. This enzyme can efficiently hydrolyze oxidized phospholipids. In agreement with its major role in maintenance of membrane phospholipid homeostasis, $iPLA_2$ activity is frequently proportional with the level of lysoPC upon inhibition by BEL, a hydrophobic serine-reactive inhibitor specific for $iPLA_2$. Nevertheless, there is increasing evidence that this enzyme may also play a role in stimulus–response coupling activity and in the release of AA for prostaglandin biosynthesis.

Oxidative Stress, Aging, and Membrane Phospholipids

Due to the high oxygen consumption, the enrichment of polyunsaturated fatty acids, and the relatively low activity of antioxidative defense enzymes, cells in brain are highly susceptible to oxidative insults, which are important underlying factors for the pathogenesis of a number of age-related neurodegenerative diseases. Several cellular enzymes, including enzymes in the respiratory chain in mitochondria, xanthine/xanthine oxidase, myloperoxidase, cytochrome P450 in cell cytoplasm, monoamine oxidases, cyclooxygenases, and NADPH oxidase, can catalyze molecular oxygen to produce reactive oxygen species (ROS) such as superoxide anion (O_2^-) and H_2O_2. In the presence of metal ions, Fe^{2+}, and Cu^{2+}, H_2O_2 can be further converted to hydroxyl radical (OH) through the Fenton reaction. Hydroxyl radicals can react with polyunsaturated fatty acids in membrane phospholipids forming peroxyl radical (ROO) and propagate the chain reaction of lipid peroxidation. External factors such as radiation, drugs, and disease states can upset the oxidative–antioxidative equilibrium, rendering cells to become pro-oxidative. More studies should be directed at examining the mechanism(s) leading to the increase in oxidized lipids in brain, especially under pathological conditions.

Among the phospholipids, plasmalogens appear to be most susceptible to oxidative stress. There is considerable evidence that production of ROS and lipid peroxidative products in brain increases with age and that the age-related increase in oxidative stress may enhance neuroinflammatory responses in the aged population. An obvious decrease in plasmalogens was observed in Alzheimer's brain and the decrease could be detected even at early stages of the disease. Lipid peroxidation products such as 4-hydroxynonenal (HNE) can form adducts with proteins and cause neurotoxicity. Thus, the increase in HNE has been regarded as an important contributor to the pathogenesis of many neurodegenerative diseases, including Alzheimer's disease and stroke.

Brain Lipid Domains

The preceding discussion examined membranes from the perspective of broad changes in fluidity and lipid composition in contrast to specific areas or domains of membranes. Bretscher and others in the 1970s postulated and demonstrated that the two leaflets which form the bilayer are asymmetric in their lipid distribution. Figure 1(a) is a model of SPM and it can be seen that the outer or exofacial leaflet is enriched in phosphatidylcholine and sphingomyelin. In contrast, the inner or cytofacial leaflet is enriched in phosphatidylethanolamine, phosphatidylserine, phosphatidylinositol, and cholesterol. This lipid asymmetry confers differences in the charge properties and fluidity of the two leaflets. The exofacial leaflet is considered to be neutral or zwitterionic, whereas the cytofacial leaflet is more negatively charged due to the enrichment of the anionic phospholipids phosphatidylserine and phosphatidylinositol. Mechanisms involved in the maintenance of phospholipid asymmetry are not completely understood. There is evidence that at least three different proteins may be involved in promoting phospholipid asymmetry: an ATP-dependent transporter (flippases) directed at the cytofacial leaflet, an ATP-dependent transporter (floppase) directed at the exofacial leaflet, and an ATP-independent transporter (scramblase) interacting with both leaflets. Phospholipid asymmetry is thought to be important in such cell signaling functions as Ca^{2+} homeostasis and receptor-mediated signal transduction. A loss of phospholipid asymmetry has been observed in sickle cell anemia,

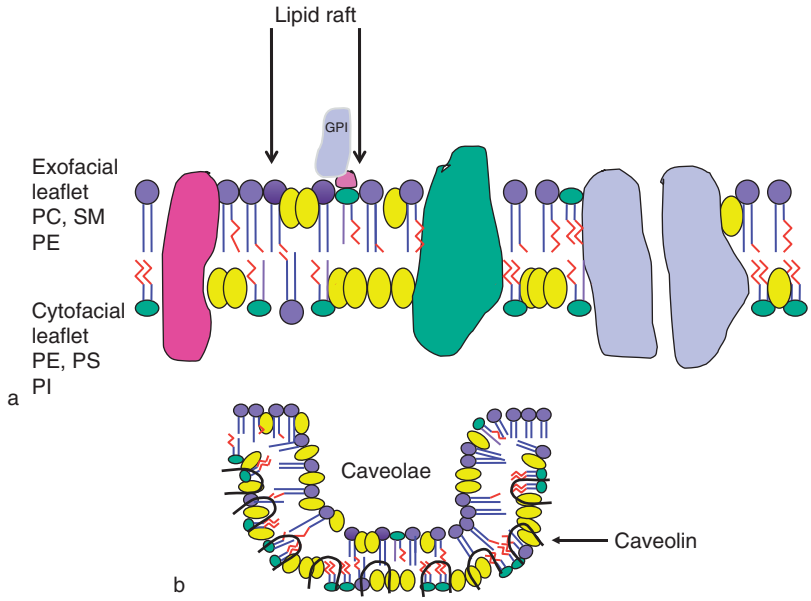

Figure 1 Plasma membrane domains. (a) The two leaflets of the plasma membrane showing the asymmetric distribution of cholesterol and phospholipids. Large globular structures represent proteins. Cholesterol (denoted by yellow structure), phosphatidylinositol (PI), phosphatidylethanolamine (PE), and phosphatidylserine (PS) are enriched in the cytofacial leaflet. Phosphatidylcholine (PC) and sphingomyelin (SM) are in abundance in the exofacial leaflet, and there is some PC in the cytofacial leaflet. A lipid raft is denoted in the exofacial leaflet (a) containing a glycosylphosphatidyl inositol-anchored protein (GPI). (b) A model of caveolae enriched in cholesterol, sphingolipids, and caveolin. Reproduced from Wood WG, Igbavboa U, Eckert GP, and Müller WE (2007) Cholesterol – a Janus faced molecule in the central nervous system. In: Lajtha A (ed.) *Handbook of Neurochemistry and Molecular Neurobiology*, 3rd edn., pp. 151–170. Heidelberg, Germany: Springer. With kind permission of Springer Science and Business Media.

sperm cell maturation, and apoptosis. In human aging erythrocytes whose life span in the body is approximately 120 days, phospholipid asymmetry is reduced. There have been a few reports on phospholipid asymmetry in brain membranes whose distribution is similar to that of nonbrain membranes, but there have been no reported studies on phospholipid asymmetry in brain membranes of aging organisms.

Exofacial and Cytofacial Leaflet Fluidity

There is a major difference in fluidity of the exofacial leaflet compared to the cytofacial leaflet. The exofacial leaflet is markedly more fluid, as determined by limiting anisotropy and rotational relaxation time of DPH in SPM, and is more easily perturbed than is the cytofacial leaflet. For example, 25 mM ethanol significantly increased fluidity of the exofacial leaflet but had no effect on the cytofacial leaflet even with an ethanol concentration as high as 400 mM. Chronic ethanol consumption reduced fluidity of the SPM exofacial leaflet producing resistance to ethanol perturbation, demonstrating cellular membrane tolerance. The asymmetry in fluidity of the two leaflets is diminished with increasing age in mouse SPM. In mice 3 or 4 months old, the exofacial leaflet was 21% more fluid than the cytofacial leaflet; in mice 14 or 15 months old, it was 9% more fluid; and in mice 24 or 25 months old, differences between the two leaflets in fluidity were indistinguishable.

Transbilayer Cholesterol Distribution

The major difference in fluidity between SPM exofacial and cytofacial leaflets is associated with the transbilayer cholesterol distribution (TCD). Figure 1(a) shows that cholesterol is asymmetrically distributed between the two leaflets. In young mice, approximately 15% of SPM cholesterol resides in the exofacial leaflet and the remaining 85% in the cytofacial leaflet. TCD is not static but, rather, can be altered by different treatment conditions. Chronic ethanol consumption, apolipoprotein E (APOE) and its isoforms, and HMG-CoA reductase inhibitors (statins) all have been shown to alter TCD. Increasing age had a striking effect on SPM TCD. The exofacial leaflet of aged mice (24 or 25 months old) showed a twofold increase in cholesterol compared to that of younger mice. Mice 14 or 15 months of age were intermediate in their cholesterol distribution with respect to the young and aged mice. Interestingly, the increase in exofacial leaflet cholesterol of the aged mice was not the result of an increase in total SPM cholesterol. There were no significant differences in total SPM

cholesterol among the three age groups. A valuable conclusion which can be drawn from this study is that striking changes in neuronal membrane cholesterol homeostasis can occur in the absence of changes in total membrane cholesterol levels.

The mechanism for maintaining TCD is not well understood. Several potential candidates have been proposed, including APOE, the low-density lipoprotein receptor (LDLR), sterol carrier protein-2 (SCP-2), polyunsaturated fatty acids, fatty acid-binding proteins (FABPs), and caveolin. However, it is worth mentioning that even when some of these potential mechanisms are experimentally altered, the cytofacial leaflet continues to maintain the greater amount of cholesterol compared with the exofacial leaflet. From these results, one conclusion is that there may be multiple mechanisms which are involved in the regulation of TCD. These mechanisms might involve multiple sites, such as the surface of the exofacial leaflet (APOE, LDLR, and caveolin), within the plasma membrane (fatty acid composition) and cytofacial leaflet interaction with intracellular proteins (SCP-2 and FABPs). A question which has not been addressed is the functionality of changes in TCD. In the case of chronic ethanol consumption, the redistribution of cholesterol may be adaptive and diminish partitioning of ethanol into the cell, in turn altering cell function. With respect to the redistribution of cholesterol in aged SPM, it is not known if such changes are adaptive or have pathophysiological consequences.

Lipid Rafts and Caveolae

Two other lipid domains which have received substantial attention are lipid rafts and caveolae. In this discussion, lipid rafts refer to structures enriched in cholesterol, glycosphingolipids, and glycosylphosphatidylinositol-anchored proteins and not containing the caveolae protein, caveolin, as shown in Figure 1(a). Other names for lipid rafts include detergent-insoluble glycolipid-enriched membranes and detergent-resistant membranes. Caveolae as seen in Figure 1(b) are defined as membrane invaginations containing the protein caveolin and are enriched in cholesterol and glycosphingolipids. There has been considerable interest in these lipid domains, and several reviews have been published. In contrast, there has been very little advancement in the aging field regarding brain lipid rafts and caveolae. Levels of the lipid raft markers flotillin-1 and flotillin-2 did not differ in brain lipid rafts prepared from a whole brain homogenate of young and aged rats. In contrast, a reduction of flotillin-1 levels as well as alkaline phosphatase activity was observed with increasing age in synaptosomal lipid rafts of aged mice expressing human apoE3 and apoE4 compared with young apoE3 mice. Lipid rafts of young apoE4 mice were more similar to those of older mice compared to young apoE3 mice for flotillin-1 levels and alkaline phosphatase activity. Cholesterol and sphingomyelin levels of lipid rafts did not differ between the young apoE3 and apoE 4 mice, but levels of those lipids did increase with age. Both increasing age and the apoE4 genotype are risk factors for Alzheimer's disease, and both have similar effects on brain lipid raft protein markers. There is evidence that lipid rafts may play a role in the processing of amyloid β-protein, which is thought to be a causal factor in the pathophysiology of Alzheimer's disease.

Superoxide dismutase-1, which is associated with membrane caveolae, was found to be reduced in aortic tissue of aged rats compared to younger animals. Caveolin-1 is a major protein of caveolae, and it has been demonstrated that the life span of caveolin-1 null mice is reduced by approximately 50% compared with wild-type mice and caveolin-1 heterozygous (+/−) mice. Brain caveolin-1 levels were found to be decreased in aged rats compared to younger rats.

The concept of lipid rafts, like that of the membrane, has been evolving. At a Keystone Symposium on Lipid Rafts and Cell Function in 2006, a consensus was reached that the term 'lipid rafts' should be changed to 'membrane rafts' recognizing the role of raft proteins, and that caveolae should be considered a membrane raft. Another issue which was discussed at that meeting and that has been proposed by others was that some membrane rafts may be transient with a lifetime of milliseconds or less, in contrast to a static structure. Such a short lifetime could pose problems for isolation of rafts as commonly done using density gradient centrifugation. An understanding of membrane rafts in aging organisms will require conceptual and technological approaches which are state of the art in the membrane raft field.

Conclusions

A broad conclusion which can be reached is that studies on age differences in membrane organization have mostly been approached from the standpoint of the Singer–Nicolson fluid mosaic membrane model. However, modification of this model has been proposed by different groups suggesting a more mosaic-like structure which is heterogeneous, with lipid–lipid, protein–protein, and lipid–protein interactions, in contrast to a fluid membrane in which proteins are unrestricted in their lateral movement. The majority of studies on age differences in membranes, whether fluidity or lipid composition, have examined bulk changes in the membrane. Although this approach provides an

overview of the membrane environment, it does not address the vast heterogeneity of the membrane. There is a need to study both macro and micro membrane domains in aging organisms and the functional consequences of membrane domains on cellular function.

See also: Aging of the Brain; Brain Composition: Age-Related Changes; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Metal Accumulation During Aging; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Plasticity: Neuronal Sprouting.

Further Reading

Datta DP (1987) *A Comprehensive Introduction to Membrane Biochemistry.* Madison, WI: Floral.

Edidin M (2003) The state of lipid rafts: From model membranes to cells. *Annual Review of Biophysics and Biomolecular Structure* 32: 257–283.

Jacobson K, Sheets ED, and Simson R (1995) Revisiting the fluid mosaic model of membranes. *Science* 268: 1441–1442.

Lommerse PHM, Spaink HP, and Schmidt T (2004) *In vivo* plasma membrane organization: Results of biophysical approaches. *Biochimica et Biophysica Acta* 1664: 119–131.

Packer L, Deamer DW, and Heath RL (1967) Regulation and deterioration of structure in membranes. *Advances in Gerontology Research* 2: 77–120.

Pike LJ (2006) Rafts defined: A report on the Keystone Symposium on Lipid Rafts and Cell Function. *Journal of Lipid Research* 47: 1597–1598.

Schroeder F (1984) Role of membrane lipid asymmetry in aging. *Neurobiology of Aging* 5: 323–333.

Sun AY and Sun GY (1979) Neurochemical aspects of the membrane hypothesis of aging. *Interdisciplinary Topics in Gerontology* 15: 34–53.

Sun GY, Xu J, Jensen MD, and Simonyi A (2004) Phospholipase A2 in the central nervous system: Implications for neurodegenerative diseases. *Journal of Lipid Research* 45: 205–213.

Vereb G, Szöllösi J, Matkó J, et al. (2003) Dynamic, yet structured: The cell membrane three decades after the Singer–Nicolson model. *Proceedings of the National Academy of Sciences of the United States of America* 100: 8053–8058.

Verkleij AJ and Post JA (2000) Membrane phospholipid asymmetry and signal transduction. *Journal of Membrane Biology* 178: 1–10.

Wood WG, Eckert GP, Igbavboa U, and Müller WE (2003) Amyloid beta-peptide interactions with membranes and cholesterol: Causes or casualties of Alzheimer's disease. *Biochimica et Biophysica Acta* 1610: 281–290.

Wood WG, Igbavboa U, Eckert GP, and Müller WE (2007) Cholesterol – a Janus faced molecule in the central nervous system. In: Lajtha A (ed.) *Handbook of Neurochemistry and Molecular Neurobiology,* 3rd edn., pp. 151–170. Heidelberg, Germany: Springer.

Wood WG, Schroeder F, Igbavboa U, Avdulov NA, and Chochina SV (2002) Brain membrane cholesterol domains, aging, and amyloid beta-peptides. *Neurobiology of Aging* 23: 685–694.

Zs-Nagy I (1979) The role of membrane structure and function in cellular aging: A review. *Mechanisms of Ageing and Development* 9: 237–246.

Lipofuscin and Lipofuscinosis

E Marani, University Twente, Enschede, The Netherlands, Leiden Medical University Centre, Leiden, The Netherlands
K G Usunoff, Medical University, Sofia, Bulgaria
H K P Feirabend, Leiden Medical University Centre, Leiden, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Lipofuscin is a pigmented oxidized protein–lipid compound, stored in a granulum or residual body in the cells of the nervous system, also called the 'wear-and-tear' pigment. Its chemical composition is ill-defined, since it is cell-type dependent, but the oxidation of its polyunsaturated lipids is well understood. Lipofuscin's presence starts around the first year, is already substantial around the sixth year of life, is further increasing in quantity during lifetime, and is abundantly present in neurons during aging. Remarkably, lipofuscin is found in autonomic ganglia during the fetal period.

Lipofuscin granula are concentrated in special compartments in most types of neurons (Figure 1), and its presence is accentuated by various stainings. A lipofuscin granule has a yellow reddish/brownish appearance, is capable of auto-fluorescence, (Figure 2), and contains several substances, for example, fat, carbohydrates (sometimes disputed), and proteins.

Lipofuscin is also present in other parts of the body (e.g., liver, digestive tract, and heart), but this is beyond the scope of this article.

History

From the earliest moment that fixed nervous tissue could be studied, lipofuscin was recognized and distinguished from the black pigment neuromelanin (for a review see Usunoff et al. (2002)), although both substances were placed in the group of fuscins.

The first description of pigments is by Purkyně (1837). Within the dorsal root ganglion cells he described a pigment, now known as lipofuscin and in the substantia nigra the black pigment neuromelanin (Purkyně's figure (Figure 1) shows the lipofuscin pigments, however, in multipolar cells). One year later, Remak (1838) published a book on sympathic ganglia and also indicated a pigment in these cells. Presumably the pigments could be detected due to the achromatic microscope that was first used by both authors. Eventually, the human, autogenous pigments were divided by Sachs (1943) into fuscin and melanin. Fuscins could fluoresce in ultraviolet light (Figure 2); neuronal melanin required oxidation before fluorescence occurred. Around the previous centennial, within a decade (1895–1905), a series of publications appeared studying lipofuscin, proposing that lipofuscin is the expression of a degenerative process, that starts early in life, but this degeneration can be arrested due to the compensatory capacity of the neuronal cytoplasm. Reduction of this compensatory capacity slowly leads to death of the neuron (Mühlmann, 1901). So, lipofuscin is considered a 'sign or marker', in the form of a storage (end) product of a 'cellular process' (degeneration, nowadays aging). This basic view has not been changed up till now.

Relation to Homologous Substances

Since lipofuscins, also called chromolipoids, are mainly or partially derived from lipid precursors, they are easily related to other lipid pigments: haemofuscins, cytolipochrome, ceroids, pseudomelanosis pigment, iron-containing lipid pigment, and lipochrome. The original lipid is colorless until it contains dissolved carotenoids and is oxidized to an insoluble, colored product. Strong variability exists as to the type of cell, the amount of protein involved, and whether the constituting amino acids are either oxidized or not (Pearse, 1985).

The degree of autoxidation as studied by different coloring techniques shows increasing oxidation for different types of polyunsaturated lipids: early oxidation as in apocrine globules and macrophage lipids can be distinguished from ceroid pigments. Ceroids are in a further stage of oxidation, and in turn lipofuscins can be noted by an even stronger oxidation. The melanin, hemosiderin, heamatoidin, and chromaffin granules are easily recognized by their black color and/or iron-containing properties (Pearse, 1985).

Enzymes

A few enzymes are localized in lipofuscins: nonspecific esterases, acid phosphatases, and proteases. Due to the increase in volume of the amount of granula, they displace cytoplasm that contains oxidative enzyme activity, resulting in empty zones for these enzymes in the cells. Some enzymes are noted near lipofuscin granula. A 'striking' increase was noted for glucose-6-phosphate dehydrogenase in the sites of lipofuscin accumulation. In general, nerve cells with a highly active respiratory apparatus seem to accumulate more lipofuscin than cells with a low oxidative metabolism (Friede, 1966).

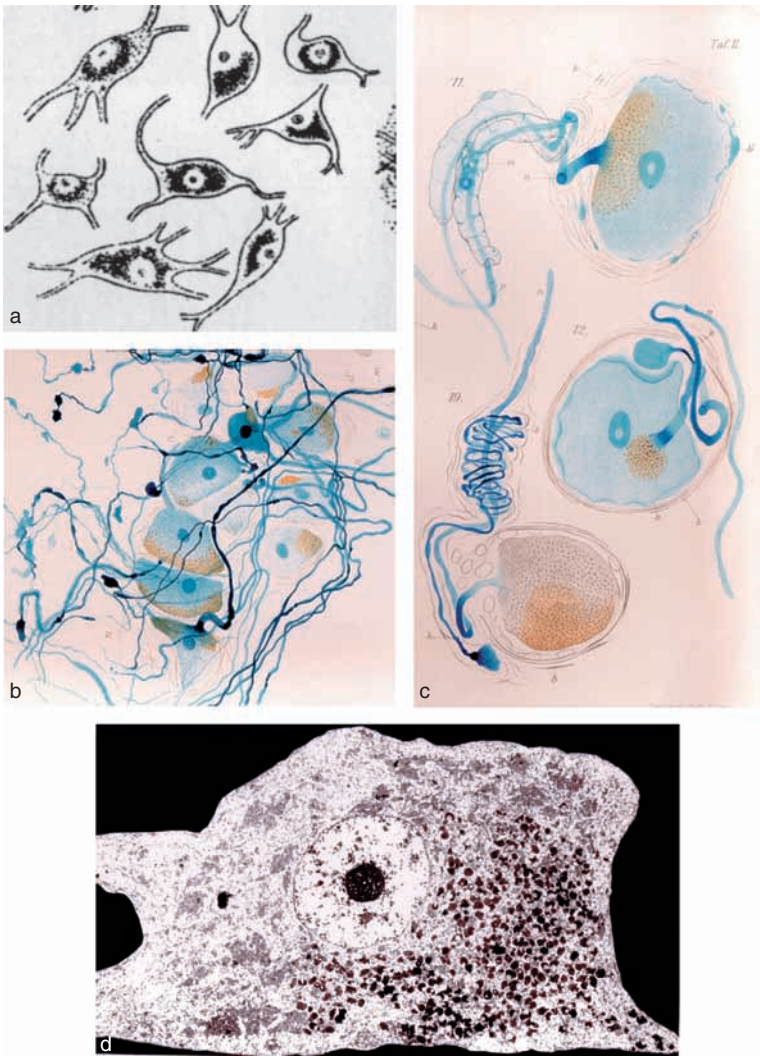

Figure 1 (a) Original picture of Purkyně (1837), indicating pigment in multipolar neurons. (b) Area taken from a horse dorsal root ganglion colored with methylene blue. Pigment is in brown color. Note that all neurons accumulated lipofuscin. (c) Three different horse dorsal root ganglion neurons (type I and II; upper one shows type I, lower figures show type II) colored with methylene blue. Pigment is restricted to certain compartments of the cytoplasm. A section through such a cytoplasmatic area shows the whole soma loaded with pigment granula. Pigment accumulation is cell-type independent in dorsal root ganglia. (d) Median section through a human neuron of Onuf's nucleus. Compartmentalization and strong accumulation of lipofuscin granula can be noted. Cytoplasm area is $\pm 1500\,\mu m^2$. (a) From Heldoorn M (2003) Modelling Onuf's Nucleus: Single Cell and Network Simulations. PhD Thesis, Leiden University. (b, c) From Dogiel AS (1908) Der Bau der Spinalganglien des Menschen und der Säugetiere. Gustav Fischer Verlag: Jena.

Staining

Lipofuscin is stained by several lipid-staining methods: Sudan III and oil red. It gives acid-fast coloration with carbol fuchsin. It stains with ferric ferricyanide (Schmorl method I), methyl green, and the periodic acid-Schiff (PAS) reaction.

The reason for the PAS-positive reaction in only the intermediate stages of oxidation of lipids is unclear. "It certainly does not mean, as many authors have concluded, that the lipid concerned belongs to the carbohydrate containing glycolipid (cerebroside) group. Pearse says, Aldehydes can be produced from non-carbohydrate-containing unsaturated phosphatides by the action of periodic acid. However, "It is more likely that 1,2-glycols are formed by the hydroxylation of epoxides engendered in the earliest stages of lipofuscin formation and be converted to Schiff-positive dialdehydes by oxidation with periodate". The J-band effect stain for lipofuscins is recommended by Pearse, and in his *Histochemistry*, volume 2, other staining methods can be consulted.

Nile blue staining is used for the differentiation of melanin and lipofuscin, and silver nitrate–gold

chloride demonstrates lipofuscin granules as small distinct dots, surrounded by a halo (Gurr, 1962).

The Masson–Fontana ammoniacal silver technique, previously thought to be specific for neuromelanins, because it is a strong argentaffin reactive pigment, also stains lipofuscin (Usunoff et al., 2002).

Based on a series of studies on the common histochemical properties of neuromelanin and lipofuscin, Barden (1969) insisted that neuromelanin is melanized lipofuscin.

Braak et al. (1988) developed a silver impregnation method for neurofibrils in which counterstaining with crotonaldehyde fuchsin followed by performic acid oxidation demonstrates lipofuscin in human pathological brains. This technique is highly recommended for human lipofuscin brain studies.

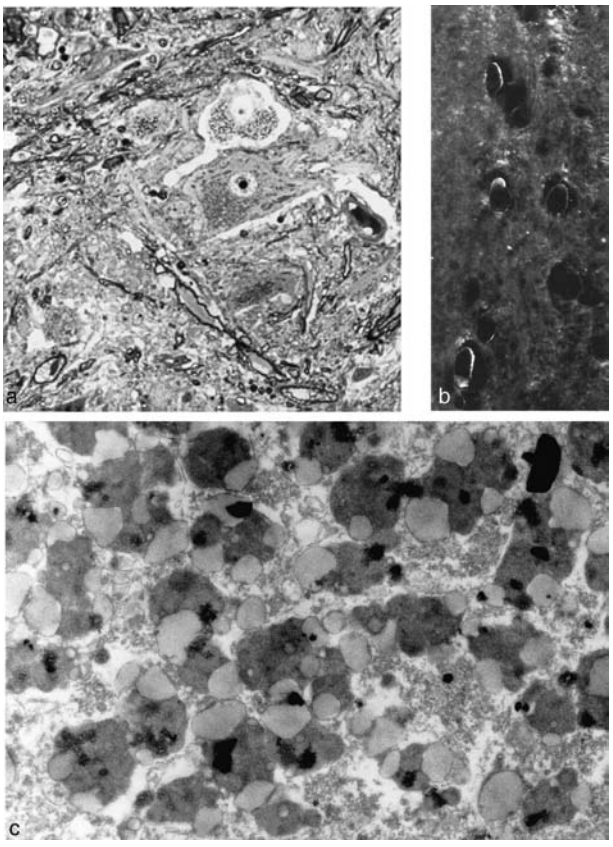

Figure 2 (a) Lipofuscin granula are present in both light and dark cells of human neurons of Onuf's nucleus (±500×). (b) Auto fluorescence of trigeminal mesencephalic nucleus neurons. A clear fluorescent halo is present. Long diameter of neurons is 45–60 μm. (c) Enlargement of lipofuscin granula in a neuron of the human nucleus of Onuf (±30 000×). (a) Heldoorn M (2003) Modelling Onuf's Nucleus: Single Cell and Network Simulations. PhD Thesis, Leiden University. (b) Usunoff KG, Marani E, Schoen JHR (1997) The trigeminal system in man. *Advances in Anatomy, Embryology, and Cell Biology*, 136: 1–125. (c) Courtesy of Dr. Feirabend.

Biochemistry

X-ray absorptions show 56% by weight of dry substance in lipofuscin, indicating that a high content of solid material is present. Ultraviolet absorption indicates the presence of pterins or aromatic amino acids. More recent studies show that lipofuscin granula from the eye showed a 44% protein and 50% lipid content (Warburton et al., 2005). Results from other areas can range from 30% to 70% protein content. The same, but inverse to protein, variability is noted for lipids. Using lectins, mannose was detected to be the dominant sugar residue, but also acetylglucosamine, acetylgalactosamine, sialic acid, galactose, and fucose are detected in human neuronal lipofuscin (Monserrt et al., 1995). Lipofuscins are also enriched in Al, Cu, and Fe, and in dolichols, the essential cofactor for the *N*-glycosylation of membrane proteins (Harman, 1989).

Lipofuscins originate from oxidation of polyunsaturated lipids. However, cellular protein catabolism is also proposed. The transition from propigment (ceroids) into lipofuscin is hard to indicate.

Electron Microscopy

Ultrastructure of the lipofuscin granula is demonstrated in this part by human material of a 75-year-old man, in which Onuf's nucleus was studied (Figure 2) and rat aged striatal neuron (Figure 3). Lipofuscin granula (diameter can go up to 2.5–3 μm) contain one or more translucent vacuoles. Around them, an electron denser substance or a matrix adheres, in which electron black spongy or black granular material is present. Trilinear

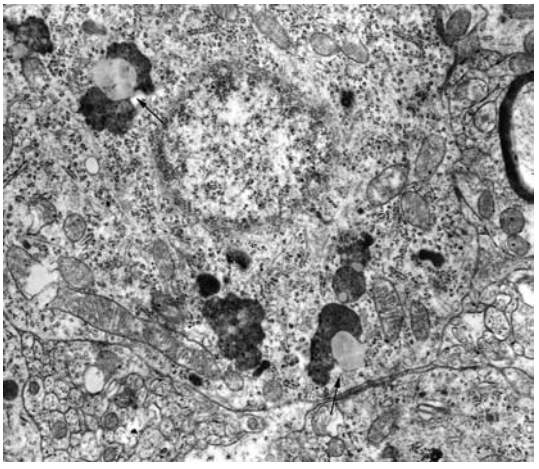

Figure 3 Striatal neuron from an aged rat (2.1 year). Arrows indicate lipofuscin granula. Note the heterogeneous matrix of the lipofuscin granula as compared to human ones in Figure 2(c) (±30 000×). From Itzev D, Lolova I, Lolov S, and Usunoff KG (2001) Age-related changes in the synapses of the rat's neostriatum. *Archives of Physiology and Biochemistry* 109: 80–89.

(Man) and/or pentalinear (other mammals) structures can be discerned in the pigment matrix (Boellard et al., 2004). The whole granulum is surrounded by a unit membrane. Translucent areas between the membrane, the vacuoles, and dense substance are found (Figure 2).

The lipofuscin granula are different in the hippocampal CA1 and CA3/4 areas in man. The hippocampal lipofuscin granula are totally different from those in the neocortex of man. Moreover, species-specific differences are found electron microscopically (cf. Figures 2(c) and 3).

Lysosomal Origin

Lipofuscin granula are thought to develop from lysosomes (Essner and Novikoff, 1960). Its origin therefore is intracellular. In short, autophagosomes fuse with a so-called dense body or lysosome to produce a secondary lysosome or residual body. Extra lipid and/or tyrosine phenol uptake will result into a residual body containing lipofuscin. The accumulating heme compounds contribute to the oxidation process and this theory also explains the presence of the type of enzymes in lipofuscin granula; for example, acid phosphatase is the marker enzyme for lysosomes.

The other studies state that the lipid pigments are derived from mitochondria (Duncan et al., 1960), while the bridge between both theories is that lipofuscin granula occur due to the incomplete lysosomal degradation of damaged mitochondria (Gray and Woulfe, 2005).

Peroxisomes are also involved in lipid breakdown; however, their granulum volume hardly increases, as occurs in lipofuscin lysosomes, and their marker enzymes are catalase and urate oxidase. Using differential centrifugation peroxisomes can be distinguished from lysosomes.

Manipulation of Lipofuscin Content

Vitamin D deficiency is responsible for an abundant increase in lipofuscin granula (Menken et al., 1986). Administration of vitamin E (α-tocopherol) diminishes the proliferation of lipofuscin granula. Vitamin A-deficient rats contained less lipofuscin in their retinal pigment epithelium.

Carotenoid pigments are related to vitamin A. Beta-carotene by the uptake of water produces two molecules of vitamin A. Moreover, rhodopsin is a carotenoid complex. Carotenoids are extremely soluble in fat, but once dissolved they are insoluble in water, glycerol, or formalin, and then held responsible for the lipofuscin color. Increase in lipofuscin granula in the pigment layer of the eye is supposed to be related to vitamin A autoxidation. Reduction of this autoxidation by vitamin E should explain the reduction of lipofuscin content. This is a debated issue.

Topography

Within the 'normal' nervous system, not all neurons develop lipofuscin granula. Good examples are the hypothalamic nuclei in man, especially the supraoptic and paraventricular nuclei, and also the type III cells in the claustrum (Braak and Braak, 1982). On the other hand, various types of neurons accumulate severely lipofuscin granula, called lipophilic, and others only few, called lipophobic (Obersteiner, 1903). Brain regions can also be called lipophobic (e.g., hippocampal CA1) or lipophilic (e.g., hippocampal CA3/4).

The distribution and amount of granula is different in nervous system areas; for example, the distribution of lipofuscin in the magnocellular nuclei in the basal forebrain shows for large multipolar neurons: 10% of these cells are lipofuscin-loaded in the medial septal and vertical limb nucleus of the diagonal band of Broca, 36% in the horizontal limb of the Broca band and anteromedial subnucleus of the basal nucleus, and 54% in the posterolateral subnucleus of the basal nucleus. The number of these multipolar lipofuscin-loaded neurons is more pronounced in Parkinson's disease (Ulfig et al., 1981).

Lipofuscin granula are also found in glia, epithelia of the choroids plexus, meninges, phagocytes, perivascular spaces, and the pigmented epithelium of the eye.

It is only possible to give short remarks on lipofuscin brain topography here. Lipophobic nuclei are nucleus tractus solitarius, the colliculus inferior, and the nucleus interpeduncularis. Lipophilic nuclei are all the cranial and spinal motor nuclei, large cells of the reticular formation, Deiter's neurons, neurons of the mesencephalic trigeminal nucleus, red nucleus, many thalamic nuclei, and cortical pyramidal cells. Good overviews can be found by Obersteiner, Altschul (1938, 1943), and Wolf and Pappenheimer (1945). One should especially consult the topographic articles of Braak and Braak, since a systematic atlas of their findings is missing.

Theories

Several theories are developed concerning lipofuscin granula. An incomplete enumeration follows here:

1. *The degeneration or aging theory.* See start of this article and Porta (2002) for an overview on aging.
2. *The proteolysis theory.* Nonenzymatic oxidation of proteins of the outer-segment components in the eye is involved in lipofuscin production (Katz, 1989).

3. *Removal theory.* Lipofuscin is produced in order to package and remove it from the neurons into macrophages and blood vessels (Gless et al., 1986).
4. *Myelination theory.* The greater the degree of myelinization and the less intense the pigmentation are, the more resistant a given projection neuron is to oxidative stress as well as to the development of Alzheimer's disease-related neurofibrillary changes and vice versa (Braak and Braak, 1999, 2000).
5. *Scavenger receptor theory.* A series of scavenger receptors is capable to recognize and internalize modified lipoproteins, making them susceptible to degeneration. The theory tries to explain several diseases (Zingg et al., 2000).
6. *The oxidation theory.* There is an increasing oxidation level from polyunsaturated lipids, via ceroids, lipofuscins, and perhaps even to the neuromelanins.
7. *The respiration chain theory.* High respiration activity is related to high lipofuscin accumulation.

Lipofuscinosis

The disease (also called Batten's disease and grouped under the neural lipidoses) was ill-defined and different illnesses on first sight were grouped to it. This was because the sign (accumulation of lipofuscin) was considered and not the deterioration of the cellular process underlying the occurrence of high concentrations of lipofuscin granula.

Modern genetic and molecular biology changed that view. Recently the neuronal (ceroid) lipofuscinoses are all considered an inherited neurodegenerative disorder (Goebbles et al., 1999; Haltia, 2003; Santavuori, 1988). Mutations in the *CLN* genes (locus name for the neuronal ceroid lipofuscinoses) are held responsible for this disease. Several of the genes are coded for and the modern nomenclature now is: *CLN 1*, infantile precocious form (called still Santavuori–Haltia disease); *CLN 2*, retarded infantile form (Janski– Bielschowsky disease); *CLN 3*, juvenile form, called the Spielmeyer–Vogt or the infantile Batten's disease; *CLN 4*, adult form, known as Kufs disease; *CLN 5*, variable form of CLN 2, also called Lake and Cavanagh disease or juvenile precocious form Finnish variant; and *CLN 8*, recently described form Turkish variant, starting early in infancy (Ranta et al., 1999).

The basic disturbed proteins and gene codes are known for most *CLN* types:

1. *CLN 1.* Localized at 1p32, coding for a glycoprotein of 37 kDa called palmyltoyl thioesterase. This is a lysosomal enzyme.
2. *CLN 2.* Localized at 11p15, coding for a lysosomal peptidase.
3. *CLN 3.* Localized at 16p, coding for a lysosomal transmembrane protein.
4. *CLN 4.* Gene not localized.
5. *CLN 5.* Localized at 13q22, coding for a transmembrane protein with unknown function.
6. *CLN 8.* Localized at 8p23, coding for a transmembrane protein, present in the endoplasmic reticulum.

One should note that the substances accumulated are different for the various lipofuscin diseases due to the different mutated enzymes. The disease can be related to a severe astrocyte gliosis coinciding with neuronal atrophy, for example, in the cortex and cerebellum. Especially astrocyte surcharging of substances is found in the hippocampus, the brain structure related to epileptic seizures.

Symptoms of the disease are blindness or visual disturbances, ataxia, mental retardation, dementia, myoclonus, choreoathetosis, and (epileptic) seizures.

A family history of lipofuscinosis indicates inherited autosomal recessive traits.

Based on all the evidence available nowadays, the neurolipofuscinoses are grouped under the heading hereditary lysosomal storage diseases, as first proposed by Koenig (1969).

See also: Autonomic Neuroplasticity and Aging; Brain Composition: Age-Related Changes; Lipids and Membranes in Brain Aging; Lysosome and Endosome Organization and Transport in Neurons.

Further Reading

Altschul R (1938) Ueber das sogenannte 'Alterspigment' der Nervenzellen. *Virchows Archiv Pathology Anatomy* 301: 273–286.

Altschul R (1943) Lipofuscin distribution in the basal ganglia. *Journal of Comparative Neurology* 78: 45–57.

Boellaard JW, Schlote W, and Hofer W (2004) Species-specific ultrastructure of neuronal lipofuscin in hippocampus and neocortex of subhuman mammals and humans. *Ultrastructural Pathology* 28: 341–351.

Braak H and Braak E (1982) Neuronal types in the claustrum of man. *Anatomy and Embryology* 163: 447–460.

Braak H, Braak E, Ohm T, and Bohl J (1988) Silver impregnation of Alzheimer's neurofibrillary changes counterstained for basophilic material and lipofuscin. *Stain Technology* 63: 197–200.

Braak H, DelTredici K, Schultz C, and Braak E (2000) Vulnerability of select neuronal types to Alzheimer's disease. *Annals of the New York Academy of Sciences* 924: 53–61.

Braak E, Griffing K, Arai K, Bohl J, Bratzke H, and Braak H (1999) Neuropathology of Alzheimer's disease: What is new since A. Alzheimer? *European Archives of Psychiatry and Clinical Neuroscience* 249(supplement 3): 14–22.

Barden H (1969) The histochemical relationship of neuromelanin and lipofuscin. *Journal of Neuropathology and Experimental Neurology* 28: 419–441.

Duncan D, Nall D, and Morales L (1960) Observations on the fine structure of old age pigment. *Journals of Gerontology* 15: 366–372.

Essner E and Novikoff AB (1960) Human hepatocellular pigments and lysosomes. *Journal of Ultrastructural Research* 3: 374–391.

Friede RL (1966) *Topographic Brain Chemistry.* New York: Academic Press.

Glees P, Aloj Totaro E, and Pisanti F (1986) Possible removal route of osmiophilic material (lipofuscin) from spinal ganglia of *Torpedo marmorata*. *Journal Fur Hirnforschung* 27: 79–86.

Goebbels HH, Mole SE, and Lake BD (1999) *The Neuronal Ceroid Lipofuscinoses (Batten Disease)*. Amsterdam: IOS Press.

Gray DA and Woulfe J (2005) Lipofuscin and aging: A matter of toxic waste. *Science of Aging Knowledge Environment* 2: re 1.

Gurr E (1962) *Staining Animal Tissues: Practical and Theoretical.* London: Leonard Hill.

Haltia M (2003) The neuronal ceroid-lipofuscinoses. *Journal of Neuropathology and Experimental Neurology* 62: 1–13.

Harman D (1989) Lipofuscin and ceroid formation: The cellular recycling system. *Advances in Experimental Medicine and Biology* 266: 3–15.

Katz M (1989) Incomplete proteolysis may contribute to lipofuscin accumulation in the retinal pigment epithelium. *Advances in Experimental Medicine and Biology* 266: 109–116.

Koenig H (1969) Lysosomes in the nervous system. In: Dingle JT and Fell HB (eds.) *Lysosomes in Biology and Pathology,* vol. 2, pp. 111–162. Amsterdam: North Holland.

Menken BZ, Su LC, Ayaz KL, and Csallany AS (1986) Organic solvent-soluble lipofuscin pigments and gluthathione peroxidase in mouse brain and heart: Effects of age and vitamin. *European Journal of Nutrition* 116: 350–355.

Moserrat AJ, Benavides SH, Berra A, et al. (1995) Lectin histochemistry of lipofuscin and certain ceroid pigments. *Histochemistry and Cell Biology* 103: 435–445.

Mühlmann M (1901) Weitere Untersuchungen ueber die Veränderungen der Nervenzellen in verschiedenem Alter. *Archiv fuer mikroskopischen Anatomie* 58: 231–247.

Obersteiner H (1903) Ueber das hellgelbe Pigment in den Nervenzellen und das Vorkommen weiterer fettähnlicher Körper im Centralnervensystem. *Arbeiten aus dem Neurologischen Instituut Universitaet Wien* 10: 245–274.

Pearse AGE *Histochemistry: Theoretical and Applied,* vol. 2, pp. 899–928. Edinburgh London: Churchill Livingstone.

Porta EA (2002) Pigments in aging: An overview. *Annals of the New York Academy of Sciences* 959: 57–65.

Purkyně JE (1837) Bericht ueber die Versammlung deutscher Naturforscher und Ärtze in Prag im September 1837. Pt. 3, sec. 5, A Anatomische–physiologische Verhandlungen. University Prague, 177–180.

Ranta S, Zhang Y, Ross B, et al. (1999) The neuronal ceroid lipofuscinoses in human EPMR and mnd mutant mice are associated with mutations in CLN8. *Nature Genetics* 23: 233–236.

Remak R (1838) Obsevationes anatomicae et microscopicae de systematis nervosi structura. Berlin: Reimer.

Sachs HW (1943) Ueber die autogene Pigmente, besonders Lipofuscin und seine Abgrenzung von Melanin. *Beitraege der Pathology und Anatomie* 108: 272–314.

Santavuori P (1988) Neuronal ceroid-lipofuscinoses in childhood. *Brain Development* 10: 80–83.

Ulfig N, Braak E, and Braak H (1989) Changes within the basal nucleus in Parkinson's disease. *Progress in Clinical and Biological Research* 317: 493–500.

Usunoff KG, Itzev DE, Ovtscharoff WA, and Marani E (2002) Neuromelanin in the human brain: A review and atlas of pigmented cells in the substantia nigra. *Archives of Physiology and Biochemistry* 110: 257–369.

Warburton S, Southwick K, Hardman RM, Secrest AM, et al. (2005) Examining the proteins of functional retinal lipofuscin using proteomic analysis as a guide for understanding its origin. *Molecular Vision* 11: 1122–1134.

Wolf A and Pappenheimer AM (1945) Occurrence and distribution of acid-fast pigment in the central nervous system. *Journal of Neuropathology and Experimental Neurology* 4: 402–406.

Zingg JM, Ricciarelli R, and Azzi A (2000) Scavenger receptors and modified lipoproteins: Fatal attraction? *IUBMB Life* 49: 379–403.

Liprins, ELKS, and RIM-BP Proteins

S Schoch, T Mittelstaedt, and M Zürner, University of Bonn, Bonn, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Exocytosis of neurotransmitter-filled synaptic vesicles is temporally and spatially highly coordinated. The docking, priming, and fusion of vesicles takes place at a specialized area of the presynaptic plasma membrane, called the active zone (AZ). A proteinaceous network, the cytomatrix at the active zone (CAZ), is involved in the different steps of the synaptic vesicle cycle. Several large multidomain proteins that do not contain transmembrane domains have been identified as being enriched at the active zone and interacting with one another in various combinations. Three of these protein families, ELKS, liprins, and Rab3-interacting molecule (RIM) binding proteins (RIM-BPs), are the focus of this article.

ELKS/CAST/ERC Proteins

ELKS/CAZ-associated structural protein (CAST)/ERC proteins were identified in several independent screens and therefore are known under a variety of names. They were first described as part of a gene rearrangement with the receptor tyrosine kinase RET and named ELKS due to their high content in the amino acids glutamate (E), leucine (L), lysine (K), and serine (S). Oligomerization of the chimeric ELKS-RET protein via the N-terminus has been shown to lead to its transformation into an oncoprotein and cause papillary thyroid carcinomas. Furthermore, ELKS has been described as being an essential regulatory subunit of the IκB kinase (IKK) complex in the nuclear factor (NF)-κB signaling cascade. It was also isolated as Rab6-interacting protein 2 (Rab6IP2B) and implicated as acting together with the small guanosine triphosphatase (GTPase) Rab6 in Golgi transport. ELKS were identified as components of the presynaptic cytomatrix (CAZ) independently in a biochemical screen for proteins enriched in the postsynaptic density fraction as CAST (CAZ-associated structural protein) and in a yeast two-hybrid screen for binding partners of the active zone protein RIM as ERC (ELKS/Rab6-interacting protein/CAST).

Gene Structure

Mammals contain two inordinately large *ELKS* genes, *ELKS1* (= ERC1/CAST2) and *ELKS2* (= ERC2/CAST), in which exons are separated by introns of more than 100 kb. The ELKS gene structure exhibits an unusual organization in that most exons are in-frame, allowing for alternative splicing. Whereas one short homologous sequence (SLTS) is variably inserted in ELKS1 and ELKS2, two separate interior sequences are alternatively spliced in either ELKS1 or ELKS2. Most strikingly, alternative splicing of the ELKS1 (but not the ELKS2) C-terminus results in two isoforms with diverging tissue distributions and biochemical properties. The two ELKS1 C-termini are coded for by distinct exons that are separated by a large intron. Whereas ELKS1A encompasses a longer sequence with no homologies to other known proteins, the shorter C-terminus in ELKS1B is almost identical to ELKS2 (Figure 1).

Expression and Localization

ELKS1A is expressed exclusively outside the brain beginning early during development. By contrast, the expression of ELKS1B and ELKS2 is restricted to the brain and becomes detectable at embryonic day (E) 14.5, when most synaptic proteins are synthesized. During the first postnatal days, the mRNA of both isoforms is widely expressed in the brain; however, in the adult brain ELKS1B mRNA is mainly detected in the hippocampus and cerebellum, whereas ELKS2 is abundantly expressed in the forebrain, including the hippocampus, cortex, olfactory bulb, and thalamus. Of the three ELKS proteins, only ELKS2 is insoluble, enriched in the postsynaptic density fraction, and exclusively localized at the presynaptic AZ. By contrast, ELKS1B is enriched in the synaptic plasma membrane but is also soluble and present at the AZ as well as in the cytoplasm. ELKS1A is largely cytosolic. In immunoelectron microscopy, ELKS2 can be detected only in the presynaptic nerve terminal, with symmetrical synapses exhibiting weaker labeling than asymmetrical synapses. In the ribbon synapses of the retina, ELKS2 was shown to localize to the base of ribbons, whereas ELKS1B was detected around the ribbons.

Protein Interactions

The amino acid sequences of ELKS1 and ELKS2 are highly homologous to one another and conserved in human, mouse, and rat. ELKS1B and -2 are 120 kDa proteins that are composed of four coiled-coil domains with no homology to other proteins and a C-terminal consensus-binding motif (IWA) for binding the PDZ (Psd-95, Dlg, and ZO1) domains of RIM1α, RIM2α/β, and syntenins (Figures 1 and 2(a)). Whereas the interaction of the ELKS1B/2 C-terminus with the RIM–PDZ domain is highly specific in that it is

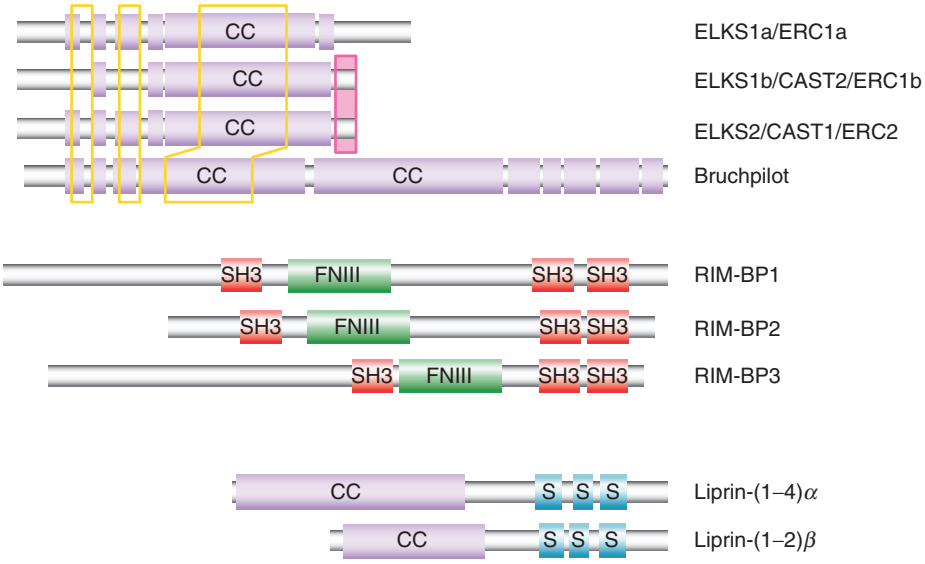

Figure 1 Domain structure of ELKS, RIM binding protein (RIM-BP), and liprin protein families. ELKS proteins are composed of highly homologous coiled-coil domains (homology regions are highlighted in yellow boxes) and an isoform-specific C-terminus, which in ELKS1B and ELKS2 contains the RIM-PDZ-binding motif (pink box). RIM-BPs consist of three homologous SH3 domains and three contiguous fibronectin type III domains. Liprins contain N-terminal coiled-coil domains and three C-terminal SAM domains. Within the liprin subfamilies, isoforms exhibit a high degree of homology. CAST, CAZ-associated structural protein; CAZ, cytomatrix at the active zone; ERC, ELKS/Rab6-interacting protein/CAST; PDZ, Psd-95, Dlg, and ZO1; RIM, Rab3-interacting molecule; SAM, sterile alpha motif; SH3, Src homology 3.

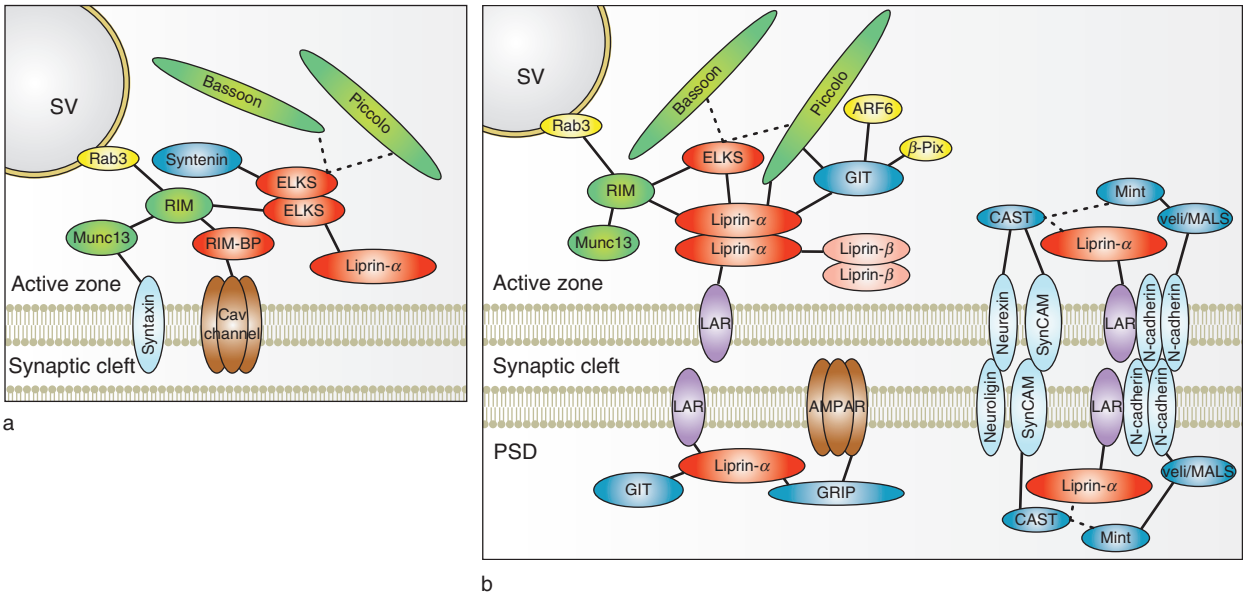

Figure 2 Schematic diagram of protein interactions: (a) of ELKS and RIM-BPs; (b) of ELKS and liprins. Proteins specifically enriched at the active zone are in green. Protein interactions that are mutually exclusive are depicted by dotted lines. ADP, adenosine diphosphate; AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; ARF, ADP ribosylation factor; CAST, CAZ-associated structural protein; CAZ, cytomatrix at the active zone; GIT, G-protein-coupled receptor kinase-interacting protein; GRIP, glutamate receptor interacting protein; LAR, leukocyte common antigen related; PAK, p21-activated protein kinase; Pix, PAK-interacting exchange protein; PSD, postsynaptic density; RIM, Rab3-interacting molecule; RIM-BP, RIM binding protein; SV, synaptic vesicle; veli/MALS, mammalian LIN-7.

abolished even by conservative substitutions, the binding to syntenin is affected only by one mutation in the last four residues of ELKS (GIWA, -B, -C, and -D) out of five mutations tested. ELKS proteins can form hetero- and homooligomers through their N- and C-terminal domains. The central region of ELKS interacts with the C-terminal region of Bassoon and Piccolo. Although the binding of ELKS with Bassoon

or Piccolo is mutually exclusive, ELKS can interact simultaneously with Bassoon/Piccolo and RIM. Moreover, ELKS proteins interact with all four liprin-α isoforms, and ELKS1A/B bind to the small GTPase Rab6. In cultured neurons, it was shown that ELKS2 is localized to synapses through its N-terminal half (including the 1 and 2 coiled-coil domains) independently of its interaction with RIM. In the same system, overexpression of ELKS2 promotes the synaptic clustering of liprins-α and syntenin-1 and is involved in, but not essential for, the localization of RIM at the AZ. However, ELKS is not required to target Bassoon to synapses.

Function

To study the role of ELKS in the regulation of synaptic vesicle fusion, superior cervical ganglion neurons (SCG) were microinjected with peptides of fusion proteins that interfered with the binding of ELKS to either RIM or Bassoon. The inhibition of both interactions resulted independently in an impairment of synaptic transmission 70 min after the microinjection of the peptides.

ELKS function has also been studied outside the nervous system. ELKS1 is expressed in pancreatic β-cells and was shown to play a role in insulin exocytosis from these cells. Introduction of the Bassoon-binding region of ELKS1 into pancreatic β-cells led to a reduction in the docking and fusion of insulin granules. Furthermore, knockdown of ELKS1 by siRNA resulted in a decrease of docked granules and glucose-evoked insulin release. PC12 cells also express ELKS1, and overexpression of ELKS1 enhanced stimulated exocytosis of human growth hormone (hGH) from PC12 cells. The binding of ELKS1 to RIM2 as well as to Bassoon was necessary to observe this increase.

Caenorhabditis elegans ELKS

A single ELKS homolog was detected in *C. elegans* that resembles mammalian ELKS weakly throughout the protein but exhibits patches of strong similarity mainly in the C-terminus, which is identical to ELKS1B/2. It was shown that the direct interaction between RIM and ELKS is conserved for the *C. elegans* proteins. RIM and ELKS are co-localized in the *C. elegans* nervous system, and the localization of both proteins is independent of the kinesin motor protein unc-104. Moreover, animals devoid of ELKS have a normal RIM staining pattern and vice versa, suggesting that in *C. elegans* this interaction is not necessary to localize RIM or ELKS to the synapse. However, expression of the soluble PDZ domain of RIM interferes with ELKS AZ targeting and truncated RIM (consisting of the PDZ and C2A domain) targets release sites in an ELKS-dependent manner, suggesting a tight association of the two proteins *in vivo*. In addition, worms deficient for ELKS display normal behaviors and are not resistant to the acetylcholine esterase (AChE) inhibitor Aldicarb. Accordingly, an electrophysiological analysis did not reveal any synaptic phenotype. Therefore, ELKS is a component of the *C. elegans* AZ, but is not a critical part of the release machinery.

Drosophila Bruchpilot

Interestingly, the recently identified *Drosophila* protein Bruchpilot (BRP) contains in its N-terminus short stretches of up to 67% identity with mammalian and *C. elegans* ELKS. However, the C-terminal coiled-coil domain has no direct vertebrate homolog and does not contain the PDZ-binding motif (IWA) of ELKS1B/2 and *C. elegans* ELKS (Figure 1). Instead, this sequence exhibits homology to large cytoskeletal proteins such as plectin and myosin heavy chain and encompasses two leucine zipper domains and a glutamine-rich C-terminus. The expression of BRP is restricted to the nervous system, and there is no nonneuronal isoform. An elegant analysis using stimulated emission depletion fluorescence microscopy (STED) showed that BRP is localized in donut-shaped structures centered at AZs of neuromuscular synapses.

An analysis of mutant flies in which nearly the entire open reading frame was deleted revealed that these animals exhibited slightly smaller neuromuscular junctions (NMJs), with fewer individual synapses, whereas the individual receptor fields were enlarged. Strikingly, the dense-core projections at the AZs (T-bars) were completely missing. At the NMJ of BRP mutant larvae, the amplitudes of excitatory junctional currents (eEJC) were drastically decreased, the nerve-evoked responses were less synchronized, and the initial release probability was reduced. Furthermore, the density of Ca^{2+}-channel clusters was severely decreased over the entire NMJ.

Liprins

Liprins are a family of multidomain scaffolding proteins that are highly evolutionarily conserved from *C. elegans* to humans. Liprins-α were first described as cytoplasmic-binding partners for the leukocyte common antigen-related (LAR) receptor protein tyrosine phosphatase (RPTPase). Several recent studies have identified liprins-α as components of the presynaptic CAZ and as part of the proteinaceous network present in dendritic spines. Furthermore, there is accumulating evidence for an integral role of liprins-α in diverse neuronal processes.

Gene Structure

The liprin protein family is composed of two subfamilies, named liprins-α and liprins-β (Figure 1). Whereas in *C. elegans* and *Drosophila* a single liprin-α gene is found, syd-2 and Dliprin, respectively, mammalian genomes contain four liprin-α isoforms (liprin-α1, -α2, -α3, and -α4). The liprin-β family consists of two genes in mammals, liprin-β1 and -β2, and one in invertebrates, celiprin-β in *C. elegans* and Dliprin-β in *Drosophila*. Liprins are composed of several N-terminal coiled-coil domains and three C-terminal sterile alpha motif (SAM) domains (Figure 1). Only liprins-α1–3 have an additional PDZ-domain binding motif at their C-terminus. The domain structure is highly homologous among the individual subfamilies. In particular, the individual domains exhibit a high degree of identity between the invertebrate and mammalian isoforms.

Expression and Localization

The different members of the liprin gene family show diverging expression patterns. Of the four mammalian liprins-α, the mRNA expression of liprin-α2 and liprin-α3 is restricted to the nervous system. By contrast, liprin-α1 mRNA is found in most tissues and liprin-α4 mRNA can be detected in the brain, heart, and muscle, although at very low levels. Whereas liprin-β2 expression is ubiquitous, liprin-β1 mRNA is mainly expressed in the heart and to a much lesser extent in the brain, liver, placenta, lung, muscle, kidney, and pancreas.

Studies in primary neuronal cultures showed a punctate staining pattern for liprins-α that co-localize extensively but not exclusively with markers for the presynaptic active zone and the postsynaptic density (PSD). Furthermore, liprin-α labeling could be detected at the excitatory and inhibitory synapses. However, liprins-α are also distributed to various nonsynaptic structures in the axons, dendrites, and soma. Ultrastructural analysis by electron microscopy has revealed a pre- and postsynaptic localization of liprins-α. Moreover, liprins-α were preferentially present in the proximity of the pre- and postsynaptic membrane, whereby the lateral distribution along the length of the AZ and the PSD appeared uniform. In subcellular fractionation analyses, liprins-α were tightly associated with the PSD fraction, lending further support to an enrichment of liprins-α in the proteinaceous network at the AZ and the PSD. In *Drosophila,* Dliprin expression has been shown to be restricted to the central nervous system and muscle. However, at *C. elegans* synaptic junctions, syd-2 is strictly localized at presynaptic terminals and tightly associated with the electron-dense specializations within the active zone.

Protein Interactions

The domain structure of liprins-α suggests that they can engage in multiple protein–protein interactions and, fittingly, a large number of binding partners for liprins-α have been isolated in recent years (Figure 2(b)). Within the N-terminal coiled-coil regions of liprins-α, separate binding domains for the presynaptic AZ proteins RIMs and ELKS, as well as for G-protein-coupled receptor kinase-interacting protein (GIT), a GTPase activating protein for the family of adenosine diphosphate (ADP)-ribosylation factor GTPases, have been characterized. Liprins-α bind to the C-terminal C2B domain of RIMs, to the N-terminal coiled-coil domain of ELKS, and to the C-terminal G-protein-coupled receptor kinase-binding (GRKBD) domain of GIT. Furthermore, liprins-α interact indirectly via GIT with another component of the presynaptic CAZ, named Piccolo. The liprin-α region, comprising the interaction sites for ELKS and GIT, mediates binding to the motor protein kinesin family member (KIF)1A. Several binding partners have also been identified for the C-terminal three SAM domains. Liprins-α (but not liprins-β) interact via the SAM domains with the catalytically inactive D2 domain of the receptor protein tyrosine phosphatases LAR, PTPδ, and PTPσ. A recent report described the binding of the first SAM domain of liprin-α2 to the N-terminus of calcium/calmodulin-dependent serine protein kinase (CASK), thereby linking liprins-α to the evolutionarily conserved tripartite complex consisting of CASK, Mint-1, and veli/MALS (lin-7–lin-10–lin-2) and via CASK to the cell-adhesion molecules neurexins and SynCAMs. However, the interaction of liprins-α and Mint-1 with CASK seems to be mutually exclusive. So far, only the multi-PDZ domain-containing protein glutamate receptor interacting protein (GRIP) has been shown to bind with PDZ6 to the C-terminal PDZ-binding motif of liprins-α. Interestingly, homo- and heterodimerization has been observed within the liprin protein family. The α- and β-subfamilies form homodimers via their N-terminal coiled-coil domains and heterodimers through the SAM domains. Considering all the protein interactions that have been reported for the various domains of liprins-α, it seems very likely that they are part of various multiprotein complexes; for example, liprins-α were found in one complex with GRIP, glutamate receptor (GluR)2/3, and GIT and in another with ELKS, Piccolo, and GIT.

Function

The first investigations into the function of liprins-α were carried out in the nematode *C. elegans*. The liprin-α homolog syd-2 was identified in a screen for

mutants affecting the localization of the synaptic vesicle protein synaptobrevin. Mutant animals expressing low levels of a short truncated syd-2 protein suffered from reduced locomotion and defective egg laying and exhibited a degree of resistance to aldicarb comparable to worms deficient for rab3 and aex-3 and indicative of a defect in synaptic transmission. The overall morphology of motor neurons was unaltered; however, presynaptic proteins such as syntaxin, rab3, and synaptotagmin were diffusely localized. An ultrastructural analysis of synapses at the NMJ revealed that the length of the presynaptic active zones was increased. The phenotype of the mutant animals could be fully rescued by the presynaptic overexpression of full-length syd-2. The deletion of the SAM domains reduced the rescuing ability of overexpressed syd-2. Taken together, these results provided the first evidence for a role of liprins-α in the organization of presynaptic active zone ultrastructure.

This hypothesis was further substantiated by studies at the NMJ junction of Dliprin-deficient flies. In mutant flies, the NMJ exhibited a decrease in branch complexity and a reduction in the number of boutons formed per NMJ. Electron micrography of individual synapses revealed that the size and shape of the AZ was dramatically altered in Dliprin-deficient flies compared to wild-type flies. On average, the total AZ area was increased by 2.4-fold. Furthermore, synaptic transmission at the NMJ was impaired because evoked excitatory junctional potentials were reduced by 36%. These results strongly imply that Dliprin is a requisite structural component of the AZ. Interestingly, the phenotype of animals deficient for Dlar, the *Drosophila* homolog of the LAR RPTP, with regard to synaptic ultrastructure is very similar to the one observed in the Dliprin mutant flies, indicating that Dliprin and Dlar act in the same pathway. However, Dlar-deficient flies show an additional defect in axon guidance at the NMJ that is not replicated in Dliprin mutants. Recent data suggest that Dliprin, together with Dlar and N-cadherin, is required for photoreceptor target selection in *Drosophila*. Furthermore, Dliprin has been described as a downstream effector of a ubiquitin ligase named the anaphase-promoting complex/cyclosome (APC/C). It has been suggested that APC/C regulates synaptic size in neurons and synaptic transmission in muscle via control of local ubiquitin-mediated protein degradation.

Dliprin has further been implicated in the trafficking of synaptic vesicles (Figure 3). Dliprin mutants showed abnormal punctate accumulations of the synaptic vesicle protein synaptotagmin along the axon and a lower density of synaptic proteins at the NMJ. Ultrastructural analysis revealed an increase in the number of clear-core vesicles in mutant axons in the absence of ectopic presynaptic specializations. Even though no change in the total flux of synaptic vesicle precursors (SVPs) was observed, the loss of Dliprin resulted in a significant decrease of the transport of SVPs in the anterograde direction with a concomitant increase in retrograde flux. It was hypothesized that in the absence of Dliprin, anterograde traffic is disrupted by decreasing the processivity of transport and increasing the initiation of retrograde movement. Dliprin interacts with the *Drosophila* kinesin-1 (khc-1), in analogy to the binding of mammalian liprin-α to the motor protein KIF1A that mediates long-range transport of AZ components in neurons along microtubuli. Studies in primary neuronal cultures showed liprins-α and KIF1A to be co-localized in various subcellular regions. Therefore, Dliprin/liprins-α could potentially function as a linker between KIF1A and its cargo or directly promote kinesin activity (Figure 3). Liprins-α have also been linked to postsynaptic transport processes. The ternary complex of GluR2/3–GRIP–liprin-α–GIT has been shown to be involved in the dendritic targeting

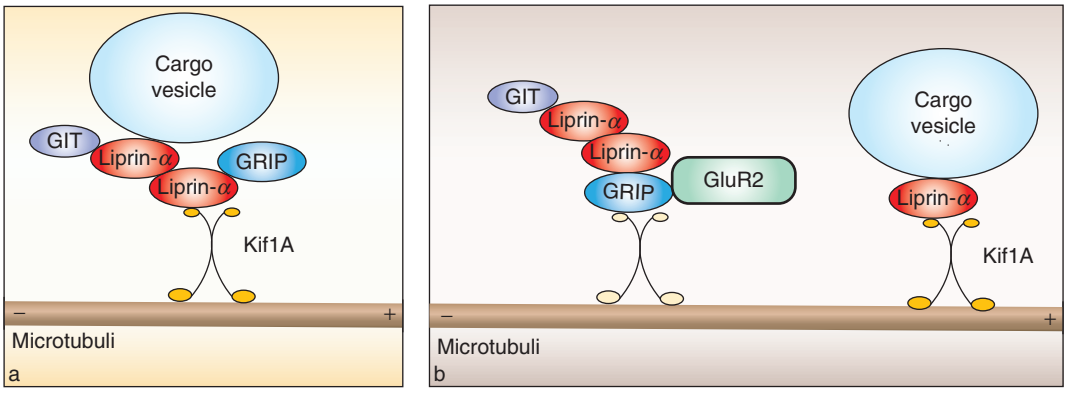

Figure 3 Role of liprins-α in neuronal transport processes: (a) in axons; (b) in dendrites. Kinesin motor proteins traveling along microtubuli toward the synapse are in yellow. GIT, G-protein-coupled receptor kinase-interacting protein; GluR, glutamate receptor; GRIP, glutamate receptor interacting protein; Kif, kinesin family member.

of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) . Both inhibition of the interaction between GRIP and liprins-α and the one between GIT and liprins-α results in a decrease of AMPAR clustering and surface expression. In addition, overexpression of liprin-α1 that is unable to bind to GRIP reduces the density of dendritic spines. It was proposed that AMPARs are transported to synapses and dendritic spines by a mechanism that requires liprins-α and their interaction with GRIP, GIT, and LAR-RPTPs. Furthermore, the binding between liprin-α and GIT may either facilitate the physical retention of AMPARs at the surface membrane or might actively inhibit the endocytosis of AMPARs. Liprins-α have also been shown to participate in localizing the scaffolding protein veli/MALS to synapses. Synaptic localization of liprins-α, in turn, seems to be partially dependent on their interaction with the AZ protein ELKS.

In summary, liprins-α engage in multiple protein–protein interactions and thereby are components of several large protein complexes/networks. There is accumulating evidence that they are critically involved in the organization and maintenance of the synaptic ultrastructure, the transport of pre- and postsynaptic cargo, and the target selection of specific axons.

RIM Binding Proteins

RIM-BP1/2 were isolated in a yeast two-hybrid screen with the C-terminus of the RIM. The human isoform of RIM-BP1 had been previously described as a protein binding to the peripheral benzodiazepine receptor and named peripheral benzodiazepine receptor-associated protein 1 (PRAX-1/bzrap1).

Gene Structure

The mammalian genome contains three RIM-BP genes, whereas a single gene has been identified in *Drosophila* and *C. elegans*. RIM-BPs are large multimodular proteins and are composed of an N-terminal Src homology 3 (SH3) domain, three central contiguous fibronectin type III (FNIII) repeats, and two tandem SH3 domains at the C-terminus (Figure 1). The amino acid sequences of the SH3 domains as well as the FNIII repeats are rather untypical but highly conserved between the individual RIM-BPs. The flanking regions are more divergent among the different isoforms and are rich in arginine, lysine, glutamic acid, and leucine. RIM-BP1 contains three proline-rich regions with a proline content around 30% and strings of charged residues with up to 13 consecutive glutamic acid residues. However, RIM-BP2 and-BP3 share only one proline-rich region with RIM-BP1, which is located downstream of the FNIII repeats. The long strings of glutamic acid residues are not conserved in RIM-BP2 and -BP3.

Expression and Localization

Expression of RIM-BP1 and RIM-BP2 is mainly restricted to the brain. A pan-RIM-BP antibody raised against the C-terminal SH3 domains of chicken RIM-BP2 resulted in an extensive and almost uniform neuronal labeling in the chicken brain. The protein could be detected pre- and postsynaptically. In cultured dorsal root ganglion neurons from chicken, a punctate distribution was observed that partially overlapped with the labeling of the synaptic vesicle protein SV2A but was also extrasynaptic. Subcellular fractionation revealed that RIM-BPs co-purify with components of the presynaptic cytomatrix and postsynaptic density proteins. Furthermore, it was shown that RIM-BPs are localized at active zones of hair cells and ribbon synapses. Interestingly, chicken RIM-BP2 expressed in human embryonic kidney (HEK) cells is diffusely distributed in the cytoplasm. However, co-expression with the Ca^{2+} channel Cav1.3 results in a clustered distribution. Immunocytochemical staining of rat brain sections with anti-RIM-BP1 antibody (raised against a peptide localized in the C-terminus of human RIM-BP1) revealed neuron-specific labeling that was intense in the cell cytoplasm and in the proximal axonal segments.

Protein Interactions

So far, protein interactions have been described only for the RIM-BP SH3 domains. The second SH3 domain of RIM-BPs binds to a proline-rich sequence in the C-terminus of RIMs. Furthermore, all SH3 domains of RIM-BP2 have been shown to interact with L-type Ca^{2+} channel Cav1.3 and N-type Ca^{2+} channel Cav2.2 (Figure 2(a)). RIM-BP SH3 domains bind to class I ligand consensus sites, +XϕPXϕP (in which + is a basic residue, usually arginine; X is any amino acid; and ϕ is a hydrophobic residue). The Cav1.3 interaction motif, RLLPPTP, fits the consensus sequence for class I ligands, differing only by a nonhydrophobic threonine residue in the sixth position. This threonine is conserved in the PXXP motifs of Cav2.2, Cav2.1 (RQLPQTP), and Cav1.4 (RLLPPTP). Therefore, RIM-BPs could interact with various Ca^{2+} channel α subunits in the brain. Interestingly, RIM-BPs can bind to Ca^{2+} channels and RIMs simultaneously. Also, the binding to the last 14 amino acids of the peripheral benzodiazepine receptor is mediated by the RIM-BP C-terminus, containing the last two SH3 domains. However, to obtain a high level of binding, it was necessary to use three copies of the 14-amino-acid sequence separated by glycines.

Function

So far, the role of RIM-BPs in the regulation of secretion has been studied only in PC12 cells. In this system, depolarization-induced secretion was enhanced by the overexpression of fusion proteins that interfered with the interactions of RIM-BP with either RIMs or Ca^{2+} channels. These results led to the hypothesis that RIM-BPs repress RIMs and Ca^{2+} channels. However, no effect on secretion was observed by overexpressing the full-length RIM-BP2. This might be due to the fact that all endogenous binding sites in RIMs and Ca^{2+} channels are already occupied. However, the ability of RIM-BPs to bind to RIMs and Ca^{2+}channels simultaneously, as well as their potential to interact with additional proteins, points to a role as a linker among the site of Ca^{2+} influx, the priming machinery, and synaptic vesicles.

See also: Active Zone; AMPA Receptor Cell Biology/Trafficking; Calcium Channels; Development of Drosophila Neuromuscular Junctions; Presynaptic Development and Active Zones; Presynaptic Regulation by Liprins; Receptor Trafficking; Rab3A Interacting Molecules (RIMs); Voltage-Gated Calcium Channels.

Further Reading

Choe KM, Prakash S, Bright A, and Clandinin TR (2006) Liprin-alpha is required for photoreceptor target selection in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* 103: 11601–11606.

Galiegue S, Jbilo O, Combes T, et al. (1999) Cloning and characterization of PRAX-1: A new protein that specifically interacts with the peripheral benzodiazepine receptor. *Journal of Biological Chemistry* 274: 2938–2952.

Hibino H, Pironkova R, Onwumere O, Vologodskaia M, Hudspeth AJ, and Lesage F (2002) RIM binding proteins (RBPs) couple Rab3-interacting molecules (RIMs) to voltage-gated Ca(2+) channels. *Neuron* 34: 411–423.

Kaufmann N, DeProto J, Ranjan R, Wan H, and Van Vactor D (2002) Drosophila liprin-alpha and the receptor phosphatase Dlar control synapse morphogenesis. *Neuron* 34: 27–38.

Kittel RJ, Wichmann C, Rasse TM, et al. (2006) Bruchpilot promotes active zone assembly, Ca^{2+}-channel clustering, and vesicle release. *Science* 312: 1051–1054.

Ohtsuka T, Takao-Rikitsu E, Inoue E, et al. (2002) Cast: A novel protein of the cytomatrix at the active zone of synapses that forms a ternary complex with RIM1 and munc13–1. *Journal of Cell Biology* 158: 577–590.

Schoch S and Gundelfinger ED (2006) Molecular organization of the presynaptic active zone. *Cell Tissue Research* 326: 379–391.

Serra-Pages C, Kedersha NL, Fazikas L, Medley Q, Debant A, and Streuli M (1995) The LAR transmembrane protein tyrosine phosphatase and a coiled-coil LAR-interacting protein co-localize at focal adhesions. *EMBO Journal* 14: 2827–2838.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Teng FY and Tang BL (2005) APC/C regulation of axonal growth and synaptic functions in postmitotic neurons: The liprin-alpha connection. *Cellular and Molecular Life Sciences* 62: 1571–1578.

Wang Y, Liu X, Biederer T, and Sudhof TC (2002) A family of RIM-binding proteins regulated by alternative splicing: Implications for the genesis of synaptic active zones. *Proceedings of the National Academy of Sciences of the United States of America* 99: 14464–14469.

Wang Y, Sugita S, and Sudhof TC (2000) The RIM/NIM family of neuronal C2 domain proteins: Interactions with Rab3 and a new class of Src homology 3 domain proteins. *Journal of Biological Chemistry* 275: 20033–20044.

Zhai RG and Bellen HJ (2004) The architecture of the active zone in the presynaptic nerve terminal. *Physiology* (Bethesda, MD) 19: 262–270.

Zhen M and Jin Y (1999) The liprin protein SYD-2 regulates the differentiation of presynaptic termini in *C. elegans*. *Nature* 401: 371–375.

Zhen M and Jin Y (2004) Presynaptic terminal differentiation: Transport and assembly. *Current Opinion in Neurobiology* 14: 280–287.

Lissencephaly

T Pramparo and T Wynshaw-Boris, University of California San Diego School of Medicine, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The pattern of development of the human cerebral cortex is complex and leads to the unique diversity of cognitive functions that makes us human. Many advances have been made by studying the variations from the basic pattern of forebrain organization in different vertebrates. The mammalian cerebral cortex, a patterned structure, is divided into anatomically distinct and functionally specialized areas leading to a species-specific map. Mature cortical areas differ by their location within the cortex, molecular properties, histological organization, patterns of connectivity, and function. Within the neocortex, rostral regions regulate motor and executive functions whereas caudal regions process somatosensory, auditory, and visual inputs. These different cortical areas have a precise connectivity, particularly with nuclei within the dorsal thalamus, which provides some of the principal inputs to the cerebral cortex. Although many of these areas can be related to each other during vertebrate evolution and share common features of cortical development, many other aspects remain poorly understood. Proliferation, migration, differentiation, and survival of neuronal precursors have been modified during brain evolution to allow for the development of the highly organized primate neocortex. Whenever one of these mechanisms, and in particular neuronal migration, is disrupted or its delicate balance is perturbed, neocortical disorders may arise and lead to different forms of mental retardation or cognitive disabilities and severe epilepsy.

Lissencephaly (LIS) is one of the best characterized and studied human genetic malformations of the cerebral cortex (MCCs). The primary defect underlying the smoothening of the brain of lissencephalic patients (Figure 1) is defective neuronal migration. The inability of postmitotic neurons to reach their final destination and correctly populate the cortical plate of the cerebral cortex consequently leads to abnormal cortical thickness and reduced or absent gyri and sulci of its surface. Neuronal migration is driven by complex mechanisms that are orchestrated by many proteins that interact mainly, but not exclusively, to promote the recruitment, organization, stability, movement and function of microtubules. Scientific understanding of the molecular mechanism and the genetic interactions that form the basis of neuronal migration has been aided by the analysis of spontaneous and engineered mouse models, and the gene function of some of the key regulatory proteins has been elucidated. The ability to study the migration features of such mouse models has been greatly facilitated by the development of *in vivo* time-lapse microscopy methods for the imaging of migrating cells. Although our knowledge of neural migration and neocortical development has been partially clarified, the mechanisms and pathways controlling these processes are far from being fully elucidated, and many questions remain unanswered.

Human Genetic Malformations

The development of the cerebral cortex is characterized by three main steps: (1) proliferation and differentiation of the neuronal stem cells into neuroblasts and glia cells, (2) migration of neuronal precursors toward the cortical plate by either radial or tangential movements, and (3) cortical organization into six layers associated with synaptogenesis and apoptosis. These processes are dynamic and occur simultaneously during the different stages of development, and perturbation of these steps leads to MCCs and malfunction. Defects occurring during the early phases of proliferation are usually associated with altered differentiation of both neuroblasts and glia cell precursors, while defects affecting neuronal migration are characterized by malpositioning of neurons along the six layers of the developing cortex. The genetic study of cerebral cortical malformations in humans has identified several genes playing crucial roles during neocortical development. Malformations due to defective neuronal migration include periventricular nodular heterotopia, classical LIS, subcortical band heterotopia (SBH), and Zellweger syndrome. Malformations due to abnormal cortical organization, also called proliferation and patterning disorders, include polymicrogyria and schizencephaly. More-detailed information regarding MCCs may be found in recent reviews. Migration is not the only process responsible for MCCs; proliferation, maturation, and survival of neurons may be critical as well for a proper cortical development. Abnormalities of the brain associated with an impaired proliferation capacity of multipotent stem cells and neuronal precursors may be responsible for microcephaly, a condition characterized by a smaller size of the brain. Compound brain phenotypes may also occur, such as the case of microlissencephaly or the cerebral

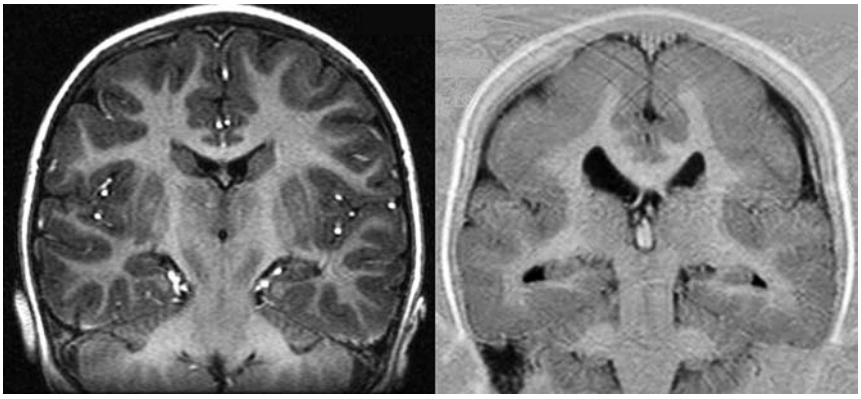

Figure 1 Brain MRI scan, coronal section passing through the hippocampus. The lissencephalic brain (right) shows diffuse cortical thickening (about 1 cm vs. 4 mm of the normal cortex, left) with simplified gyral pattern, especially in the brain convexity. The cortical sulci are shallow, and the number of convolutions is reduced. Courtesy of Dr. Renzo Guerrini, Department of Pediatric Neurosciences, Pediatric Hospital A Meyer and University of Florence, Firenze, Italy.

abnormalities associated with Neu–Laxova syndrome, in which cell survival seem to play and important role. MCCs may be associated with epilepsy that typically tends to be severe, although its incidence and type vary in different malformations; it has been estimated that $\leq$40% of children with drug-resistant epilepsy have a cortical malformation. Some of the causative genes for MCCs have been identified, but most remain unknown; only a dozen or so genes have been found causing epilepsy, and about 400 unbalanced chromosome rearrangements are associated with epilepsy syndromes, thus suggesting that a large number of dosage sensitive genes remain to be discovered.

Neuronal Migration

The mammalian neocortex is organized into six distinct cortical layers, each with distinct neuronal morphology and functions. The distribution of neuronal cells during cortical development occurs mainly by radial and tangential movements. In the neocortex, most neurons and glia arise from radial glia progenitor cells. Radial migration represents the predominant direction and type of movement for the positioning of most neurons in the neocortex departing from the ventricular zone (VZ), where postmitotic neuroblasts arise from radial glia progenitor cells and migrate toward the pial surface. This type of migration is characteristic of the dorsal telencephalon giving rise to the excitatory glutamatergic projection neurons of the cerebral cortex, which sequentially reach the different layers of the cortex. Tangential migration, on the other hand, occurs mainly from progenitors arising in the ventral telencephalon, generating γ-aminobutyric acid (GABA)ergic inhibitory neurons, including basal ganglia neurons, as well as interneurons that migrate tangentially to contribute to the formation of the cortex and to reach the olfactory bulb by following the rostral migratory stream. Cortical layering occurs in the mouse between days 11 and 18 during embryogenesis and in the human between 6 and 20–24 weeks. The generation of the six layers occurs in an inside-out fashion: early migrating neurons will populate the deeper level of the cortex, and later migrating neurons the more superficial ones. The first postmitotic neurons produced in the VZ migrate to form a subpial preplate or primitive plexiform zone. Subsequently, later migrating neurons will form the cortical plate by migrating into the preplate to split it into the superficial molecular layer (or layer I or marginal zone containing Cajal–Retzius neurons) and the deep subplate. The different waves of migrating neurons will pass the subplate and end their migration below layer I, forming cortical layers VI, V, IV, III, and II in an inside-out fashion. Radial migration occurs either by locomotion, in which neurons migrate along the processes of radial glia cells (progenitors guiding migration), or by nuclear translocation, in which migrating neurons extend their leading process toward outer levels of the cortex, to be followed by centrosome–nuclear movement. The two types of radial migration are dynamically different and occur at different times during neocortical development; moreover, recent observations suggest that neuronal migration displays different and complex patterns characterized by spatially and temporally distinct changes in the direction of movement and speed of migration for both pyramidal cells and interneurons.

Time lapse imaging studies using cortical slice cultures have demonstrated that neurons generated in the VZ at later stages of development pass through

a series of distinct stages of migration characterized by changes in cell shape, direction of movement, and speed of migration as they move to the cortical plate. According to these observations, radial migration does not occur directly to the cortical plate but is divided into four different phases. During phase 1, newly generated neurons at the ventricular surface move radially from the VZ to the subventricular zone (SVZ). At this point, phase 2, they pause at the border of the intermediate zone (IZ) with the SVZ and acquire a multipolar stage. This morphological change provides cells with a dynamic capability to change polarity, extending and retracting processes to move within the SVZ. Consequently, they do not appear to depend on radial glia fibers for their movement, as demonstrated by their ability to move tangentially. Most of the neurons go through a third phase characterized by a bipolar state, in which they extend processes toward the VZ, followed by retrograde translocation of the cell body. In the fourth phase, neurons reverse polarity and extend a pial-directed leading process, take on the bipolar morphology of migrating neurons, and begin radial migration to the cortical plate. Some neurons, however, after pausing in the SVZ, do not proceed through phase 3 with a retrograde motion toward the ventricle but instead progress directly from phase 2 to phase 4.

It is interesting that these four distinct phases of neuronal migration may be relevant to partially explain the generation of human migration disorders. For instance, in SBH a band of cortical neurons fails to migrate fully into the cortex and collects in the subcortical white matter. It is possible that these neurons fail to migrate because they are not able to make the transition from phase 2 to phase 3 of migration and therefore remain as immature multipolar neurons in the subcortical white matter. Similarly, periventricular nodular heterotopia may reflect a failure of neurons to transit from phase 3 to phase 4 of migration or inappropriate migration backward toward the ventricular surface. In this regard, *in utero* electroporation experiments indicate a cellular role of filamin A in exit from the multipolar stage; indeed, excess expression of the rodent filamin A or RNAi of Filip1 (which encodes a filamin A-binding protein that targets filamin for degradation) reduces the number of multipolar cells in the SVZ and IZ. By contrast, overexpression of Filip1 and the resulting decrease in Flna expression causes accumulation of cells in the multipolar stage. Filamin A is thus required for migration within the multipolar stage and for the transition into bipolar modes of migration. LIS1 knockdown (see the section titled 'LIS: clinical aspects and molecular genetics') causes accumulation in the VZ and SVZ of multipolar cells that have reduced capacity for migration. Doublecortin (DCX) knockdown in rats (see below) causes cells to accumulate in the multipolar stage in the IZ. Overexpression of DCX, by contrast, causes cells to adopt a highly bipolar morphology. Therefore, similar to FLNA, DCX loss of function and gain of function cause increases and decreases, respectively, in the number of cells in the multipolar stage. Overall, these observations suggest that the multipolar stage is a point of vulnerability to disruption that will require further investigation.

Molecular Pathways of Neuronal Migration

During the past decade, several molecules have been identified that participate in the control of neuronal migration, and several of the mechanisms driving neuronal movement have been partially clarified. The cloning of human disease genes and the study of genetically modified mouse models have helped in the understanding of the function of these gene products. Four main categories can be distinguished in triggering initiation, progression, modulation, and termination of neuronal migration. Filamin-A (an actin-binding protein) and Arfgef2 (or adenosine diphosphate–rybosylation factor GEF2 involved in vesicle trafficking) seem to play an important role in the initiation of migration, whereas DCX, a microtubule-associated protein (MAP), and LIS1, a MAP and dynein regulator, are involved in the regulation of its progression, and mutation of these genes results in slow or delayed migration. Other molecules (MAP1B, MAP2, and Tau) have been described and associated with migration defects, although not yet with human disorders. A second category includes proteins that play a role in lamination, such as reelin (a glycoprotein of the extracellular matrix), Dab1 (an adaptor protein), two reelin receptors (apolipoprotein E receptor 2 (ApoER2), and very low-density lipoprotein receptor (VLDLR)), p35 (an activator of cyclin-dependent kinase 5 (Cdk5), Cdk5 (a serine-serine-threonine kinase), and Brn1/Brn2 (transcriptional activators of Cdk5 and Dab1). Disruption of each of these proteins results in cortical layer inversion. A third category of molecules includes modulators of glycosylation that likely provide stop signals for migrating neurons. These molecules include protein O-mannosyltransferase, protein O-mannose b-1, 2-*N*-acetylglucosaminyltransferase, fukutin (a putative glycosytransferase), and focal adhesion kinase. The last group of molecules acts as modulators of neuronal migration and may be trophic factors such as brain-derived neurotrophic factor and thyroid hormones or neurotransmitters like glutamate and γ-aminobutyric acid.

One molecular pathway important for the progression of neuronal migration ultimately leads to the regulation of dynein motor function via LIS1. A plausible simplified model for this pathway has been proposed regulating nucleokinesis during neuronal migration through the regulation of dynein motor function (Figure 2). RELN binds to the VLDLR–ApoER2 complex to activate mDab1. Activated mDab1 associates with cAbl and activates cables. Cables has been identified as a potential link between the RELN pathway and Cdk5. Cables activates Cdk5 through interaction with cAbl and mDab1, bridging the RELN and Cdk5 pathways. Activated Cdk5 then phosphorylates NDEL1 (and perhaps NDE1), and P-NUDEL binds to 14-3-3ε to protect it from phosphatase attack. A LIS1–NudE homolog complex then positively regulates dynein motor function through direct association with CDHC and dynein LC. It appears that Cdk5, LIS1, NDEL1, NDE1, and 14-3-3ε regulate dynein motor function in a positive fashion, since decreasing the activity, decreasing the amount, or eliminating these proteins completely results in decreased dynein motor function and mislocalization of dynein components and centrosomal proteins. Thus, the apparent function of the LIS1 pathway is to activate dynein motors and place them in the proper cellular location for function. This model is undoubtedly an oversimplification of what is likely to be a process that is regulated in a highly ordered and complicated fashion. There may be many branch points to and from this basic backbone, and much of the cross-talk between pathways remains to be identified.

LIS: Clinical Aspects and Molecular Genetics

Classical LIS, or 'smooth brain,' represents a pathological condition of the brain characterized by a simplified pattern or the absence of the normal pattern of

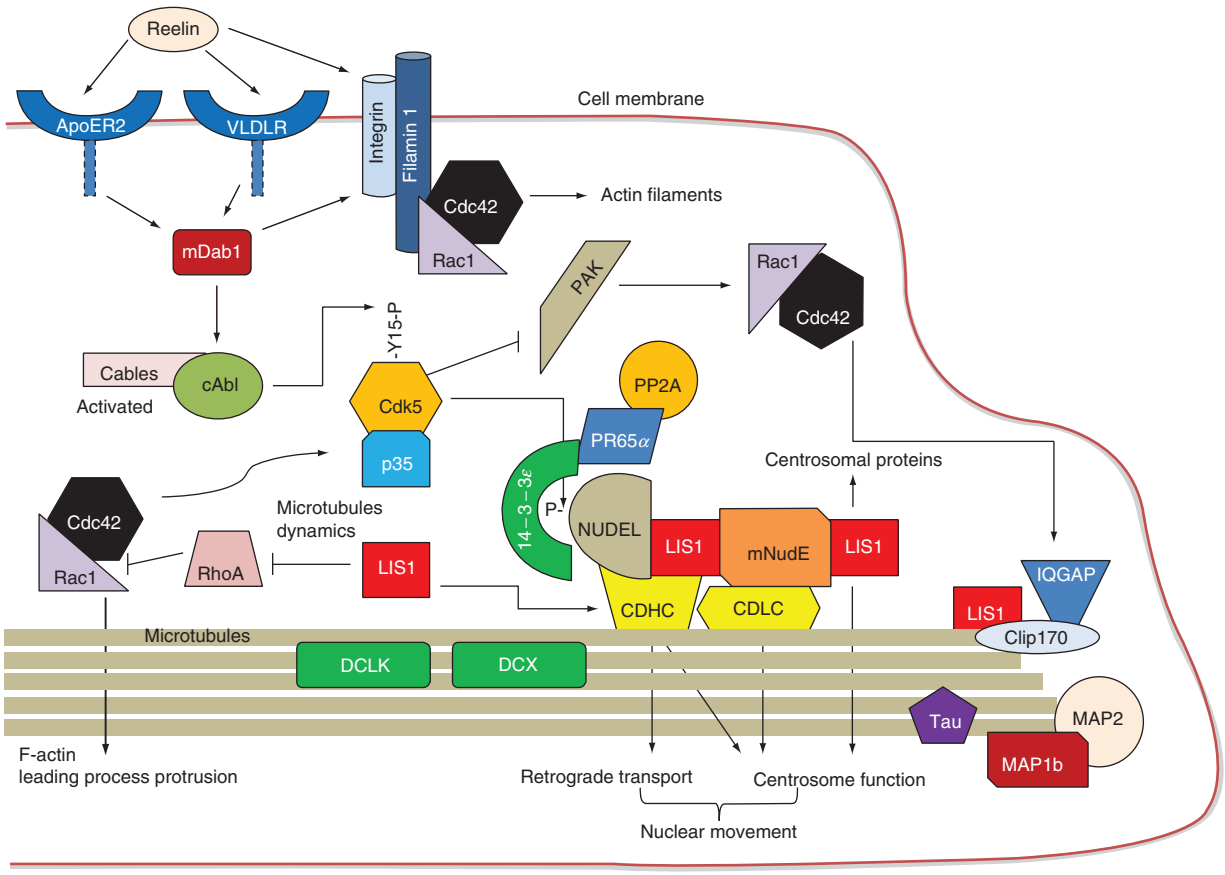

Figure 2 A simplified molecular pathway of the regulation of neuronal migration. The disruption of both actin filaments and microtubule dynamics can inhibit neuronal migration. The molecular cascade is initiated by the binding of reelin to its receptors. The stabilization of microtubules via mitogen-activated proteins (MAPs) and the activation of the dynein motor complex play a critical role in the nucleus–centrosome coupling and nucleokinesis. See text for details. ApoER2, apolipoprotein E receptor 2; Cables, a potential link between the RELN pathway and Cdk5; Cdk5, cyclin-dependent kinase 5; Dab1, an adaptor protein; DCX, doublecortin; CDHC, cytoplasmic dynein heavy chain; CDLC, cytoplasmic dynein light chain; Cdc42, cell division cycle 42; Rho, Rho family GTPase; PP2A, protein phosphatase 2A; VLDLR, very low-density lipoprotein receptor.

gyri and sulci, resulting in a reduction of the cortical surface with abnormal thickness and layer organization (**Figure 1**). This disorder affects about one in 40 000 individuals and is inherited as either an autosomal or an X-linked dominant disorder. Typically, LIS, diagnosed through magnetic resonance imaging, is associated with severe mental retardation and intractable epilepsy that consequently lead to a shortened life span. Syndromes related to LIS include Miller–Dieker syndrome (MDS; a severe form of LIS); double cortex syndrome, or SBH (a less severe condition); and LIS with cerebellar hypoplasia, polymicrogyria, and cobblestone with associated congenital muscular dystrophy and eye disease. In LIS, neurons migrate only partially toward their proper cortical destination so that in the mature cortex, gyri and sulci fail to form. SBH is a disorder in which bilateral bands of grey matter are interposed in the white matter between the cortex and the lateral ventricles. Two major forms of LIS have been differentiated, classical LIS (type I) and cobblestone (type II), although careful delineation of syndromes has revealed multiple forms of LIS.

Two genes have been associated with classical LIS and SBH. The first gene, LIS1 (Online Mendelian Inheritance in Man (OMIM) database number 601545), on chromosome 17p13.3, is responsible for the autosomal form of LIS, while the second gene, DCX (or XLIS, OMIM 300067), is X-linked. Although mutations in either gene can result in either LIS or SBH, the majority of cases of classical LIS are due to deletions or mutations of the LIS1 gene, whereas the majority of cases of SBH are due to mutations of the DCX gene. The most important characteristics of LIS in patients with LIS1 mutations are the very thick (10–20 mm) cortex, gyral malformations that are more severe in posterior than anterior brain regions, and a prominent cell-sparse zone in the cortex, whereas DCX mutations result in LIS more severe in anterior brain regions.

Mouse knockouts for *Lis1* with graded reduction of the protein level displayed a severe developmental brain disorder as the result of an abnormal neuronal migration. Mice with 50% of LIS1 protein displayed cortical, hippocampal, and olfactory bulb disorganization resulting from delayed neuronal migration. Mice with further reduction (35% of LIS1 protein) displayed a more severe brain disorganization, as well as cerebellar defects. These results suggested that *Lis1* has an essential and dosage-sensitive role in neuronal migration during brain development. Reduction of *Lis1* levels has been shown to have a pleiotrophic effect involving interkinetic nuclear migration, neuroblast proliferation, and survival.

Patients with LIS and deletions of 17p13.3 (MDS) have more-severe LIS than do patients with isolated LIS sequence (ILS) in addition to facial abnormalities, thus suggesting that at least a second gene in the deleted region may contribute to the migration defects during cortical development. MDS is always associated with LIS grade 1 (essentially complete agyria), while ILS has been associated with a more variable LIS phenotype ranging from grade 2 to grade 4, in which pachygyria is most frequently observed. The definition of the deleted region in MDS patients (about 400 kb) suggested that 14-3-3ε was the best candidate gene contributing to the severity of phenotype. Indeed, mouse knockouts for *14-3-3ε* displayed abnormal cortical development due to a reduced capacity of neurons to migrate and populate the upper level of the cortex. In addition, double heterozygote mouse mutants for *Lis1* and *14-3-3ε* displayed increased severity of cortical migration defects and hippocampal disorganization compared with the single heterozygotes. The finding of slower migration in the cortex and hippocampus of double heterozygotes, compared with single heterozygotes and wild-type mice, strongly suggested a genetic interaction between *Lis1* and *14-3-3ε* during brain development and neuronal migration, thus partially clarifying the phenotype differences between MDS and ILS patients.

LIS1 encodes a protein that is similar to the β subunit of heterotrimeric G-proteins and has two functions: a regulatory subunit of platelet-activating factor acetylhydrolase (PAF-AH), an enzyme that degrades the bioactive lipid PAF, and a dynein regulator. The importance of LIS1 as part of PAF-AH in neuronal migration is unclear. LIS1 protein has been shown to co-localize with microtubules and to promote their stabilization, and an ortholog of LIS1 in *Aspergillus nidulans*, nudF, mediates nuclear translocation by interacting with microtubules and dynein. Thus, LIS1 exerts its effects on migration through regulating dynein function and microtubule stability. The gene is expressed in neural progenitors within the VZ as well as in differentiated neurons, suggesting a function during multiple stages of neural development. The interaction between LIS1 and dynein is critical for the binding of microtubules to the cell cortex and generation of cellular machinery for neuronal division. LIS1 binds cytoplasmic dynein and dynactin and localizes to the cell cortex and to mitotic kinetochores. It has been demonstrated that the perturbation of LIS1 in cultured cells interferes with mitotic progression and leads to spindle disorientation. The use of *in utero* electroporation of *Lis1* small interfering RNA and dominant-negative forms of LIS1 or dynactin allowed the first real-time imaging of cortical cell progenitors lacking LIS1 or dynactin function. These experiments resulted in the perturbation of the progression of neural progenitors through

the cell cycle in the VZ, causing a reduction of proliferation and the block of postmitotic precursors through the SVZ, leading to the accumulation of multipolar progenitor cells within the SVZ. Additional insights into the role of LIS1 in cortical development are coming from the analysis of human neural precursors isolated from a 33-week postmortem fetus with MDS. This study provides evidence of disruption in cell proliferation, in addition to migration, and suggests that this process might have an important role in the development of LIS.

DCX is a MAP that stabilizes actively polymerized microtubules and is expressed in newly postmitotic migrating neurons in both the central and the peripheral nervous systems. *Dcx* heterozygote and hemizygote knockout mice showed remarkably normal overall brain morphology, with six layers of the cerebral cortex. In contrast to the normal cerebral cortex, however, the hippocampal formation was malformed, affecting the CA fields in both genotypes. The anatomical abnormalities of the hippocampus were shown to be associated with learning and memory deficits. The central role of DCX in neuronal migration, through the organization and stability of microtubules, and the previous finding of LIS1 effects on microtubules, suggest that microtubule regulation is a key component of neuronal migration in the cerebral cortex. In fact, manipulation studies of LIS1, DCX, dynein, and microtubules have demonstrated that LIS1 and DCX mediate nucleus-centrosome coupling in migrating neurons through dynein, and overexpression of either one enhances migration and rescues the reduced nucleus-centrosome distances. Recent data have shown that DCX is also expressed in tangentially migrating neuroblasts in the rostral migratory stream and that it is an important regulator of both radial and tangential migration. Moreover, DCX hemizygosity was responsible for a branched phenotype in neurons that failed to migrate, thus suggesting an additional role of DCX in maintaining the bipolar morphology of migrating neurons, together with its involvement in nuclear translocation.

Conclusion

In the past 10 years, major progress has been made in the diagnostic recognition of MCCs, especially through the use of magnetic resonance imaging. Variations in distribution and depth of cortical sulci, cortical thickness, boundaries between grey and white matter, and signal intensity allow recognition of different malformation patterns. Sophisticated brain imaging techniques in the emerging field of behavioral neurogenetics allow the precise identification of these malformative patterns, greatly contributing to the selection of patients with similar or identical phenotypes. In turn, the improved selection of patients facilitates the analysis and the discovery of the genetic causes for these conditions. The study of patients with LIS has provided an entry point to the molecular pathways governing neocortical development and neuronal migration in particular. Although the basic mechanisms driving the movement of neurons are becoming clearer, many questions remain to be answered, and most likely other pathways remain to be uncovered. Migration is obviously the driving major force for cortical organization, but the importance of proliferation of neuronal precursors and the survival of postmitotic neurons are events that need to be further investigated. One possibility for improving the body of knowledge is to further investigate the molecules or pathways and binding partners that have already been discovered by means of either transgenic mice or sophisticated *in vivo* technologies. An alternative approach would be to identify new candidate genes with positional cloning methods by screening patients with cortical abnormalities and testing these genes with functional tools. Techniques such as RNA interference and the *in vivo* imaging of neurons using slice cultures will be useful for testing new genes and potentially detecting novel players on the migration pathways.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Cell Replacement Therapy for Huntington's Disease; Cerebral Cortex; Developmental Disability and Fragile X Syndrome: Clinical Overview; Frontal Cortex Evolution in Primates; Lissencephaly Type I and Periventricular Heterotopia; Memory Consolidation: Cerebral Cortex; Neocortex: Origins; Neural Patterning: Arealization of the Cortex; Neurulation.

Further Reading

Bai J, Ramos RL, Ackman JB, et al. (2003) RNAi reveals doublecortin is required for radial migration in rat neocortex. *Nature Neuroscience* 6: 1277–1283.

Barkovich AJ, Kuzniecky RI, Jackson GD, Guerrini R, and Dobyns WB (2005) A developmental and genetic classification for malformations of cortical development. *Neurology* 65: 1873–1887.

Gambello MJ, Darling DL, Yingling J, et al. (1999) Doublecortin is a microtubule-associated protein and is expressed widely by migrating neurons. *Neuron* 23: 257–271.

Grove EA and Fukuchi-Shimogori T (2003) Generating the cerebral cortical area map. *Annual Review of Neuroscience* 26: 355–380.

Guerrini R and Marini C (2006) Genetic malformations of cortical development. *Experimental Brain Research* 173: 322–333.

Gupta A, Tsai LH, and Wynshaw-Boris A (2002) Life is a journey: A genetic look at neocortical development. *Nature Reviews Genetics* 3: 342–355.

Hirotsune S, Fleck MW, Gambello MJ, et al. (1998) Graded reduction of Pafah1b1 (Lis1) activity results in neuronal migration defects and early embryonic lethality. *Nature Genetics* 19: 333–339.

Kato M and Dobyns WB (2003) Lissencephaly and the molecular basis of neuronal migration. *Human Molecular Genetics* 12 (Review Issue 1): R89–96.

Koizumi H, Higginbotham H, Poon T, et al. (2006) Doublecortin maintains bipolar shape and nuclear translocation during migration in the adult forebrain. *Nature Neuroscience* 9: 779–786.

Noctor SC, Martinez-Cerdeno V, Ivic L, and Kriegstein AR (2004) Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. *Nature Neuroscience* 7: 136–144.

Sheen VL, Ferland RJ, Harney M, et al. (2006) Impaired proliferation and migration in human Miller–Dieker neural precursors. *Annual Neurology* 60: 137–144.

Sur M and Rubenstein JL (2005) Patterning and plasticity of the cerebral cortex. *Science* 310: 805–810.

Tanaka T, Serneo FF, Higgins C, et al. (2004) Lis1 and doublecortin function with dynein to mediate coupling of the nucleus to the centrosome in neuronal migration. *Journal of Cell Biology* 165: 709–721.

Toyo-oka K, Shionoya A, Gambello MJ, et al. (2003) 14-3-3epsilon is important for neuronal migration by binding to NUDEL: A molecular explanation for Miller–Dieker syndrome. *Nature Genetics* 34: 274–285.

Relevant Website

http://www.ncbi.nlm.nih.gov/ – OMIM – Online Mendelian Inheritance in Man.

Lissencephaly Type I and Periventricular Heterotopia

E Parrini and R Guerrini, Children's Hospital A. Meyer–University of Florence, Florence, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The cortex, or pallium, comprises several subdivisions. In mammals, these functionally distinct regions include the hippocampus, parahippocampal areas, neocortex, entorhinal cortex, olfactory cortex, claustrum, and olfactory bulb. The remarkable conservation of the pattern of areal divisions within a given species suggests a highly conserved and rigidly regulated regional specification program that controls their development. Neocortical regional specification depends on interactions between intrinsic properties of cortical cells with input from subcortical structures.

The development of the human cerebral cortex is a dynamic process that can be divided into partially overlapping stages that occur during several gestational weeks. During the first stage, stem cells proliferate and differentiate into young neurons or glial cells deep in the forebrain, in the ventricular and subventricular zones lining the cerebral cavity. After their final mitotic division, cortical neurons migrate away from their place of origin in a radial fashion along the radial glial fibers from the periventricular region or tangentially, primarily from the ganglionic eminences, toward the pial surface, where each successive generation passes one another and settles in an inside-out pattern within the cortical plate. In a third phase, cortical organization into six layers occurs, associated with synaptogenesis and apoptosis (Figure 1). In humans, the proliferation stage ranges from prenatal weeks 5–6 to weeks 16–20, migration from prenatal weeks 6–7 to weeks 20–24, and organization from week 16 until well into postnatal life. When migration is complete, the cortex is a six-layered structure, with each layer comprising different types of neurons that form discrete connections within the central nervous system (CNS) and performing distinct functions. When neurons reach their destination, they stop migrating and order themselves into specific architectonic patterns. Mechanisms by which cell migration into the cortical plate stops at the appropriate location have been elucidated through the characterization of the *Reeler* mutant mouse. In this animal model, the cortical pattern is the opposite of the normal inside-to-outside development of the cerebral cortex. These data suggest *Reln* is required for the normal inside-to-outside positioning of cells as they migrate from the ventricular zone, a first component of a signaling pathway guiding cells to the correct location in the cerebral cortex.

Genetic studies of cerebral cortex development and neuronal migration have been remarkably successful over the past few years, uncovering several genes that, when mutated, cause disorders of neuronal migration in mice and in humans. Cell migration has been implicated in many disorders; however, lissencephaly (LIS) and periventricular heterotopia (PVH) are the only disorders that are clearly pathogenetically proven to be cell-migration defects.

Lissencephaly Type 1

Definition and Histopathology

Lissencephaly (LIS) (derived from the Greek words *lissos,* meaning smooth, and *enkephalos,* meaning brain) is a neuronal migration disorder characterized by absent (agyria) or decreased (pachygyria) convolutions, producing a smooth cerebral surface. Several different types of LIS have been recognized. The most common type, known as classical LIS or type 1 LIS, features a very thick 10–20 mm cortex and no other major brain malformations.

The cytoarchitecture consists of four primitive layers, including an outer marginal layer, which contains Cajal-Retzius neurons (layer 1); a superficial cellular layer, which contains numerous large and disorganized pyramidal neurons (layer 2) corresponding to the true cortex; a variable cell sparse layer (layer 3); and a deep cellular layer composed of medium and small neurons, which extends to more than half the width of the mantle (layer 4) (Figure 2). The white matter, which is severely reduced in volume, occasionally contains individual neurons or a collection of neurons forming the heterotopia.

Subcortical band heterotopia (SBH) is a related disorder in which there are bilateral bands of gray matter found interposed in the white matter between the cortex and the lateral ventricles. SBH may appear as a solid band of heterotopic tissue or as numerous islands of radially oriented gray matter separated by white matter. Pyramidal cells in the heterotopic band may appear smaller than normal (i.e., those in the overlying cortex, which is usually structurally normal with the exception of shallow sulci).

Genetics

Two genes have been associated with classical LIS and SBH. The *LIS1* gene (Online Mendelian Inheritance in Man (OMIM): *601545) on chromosome 17p13.3

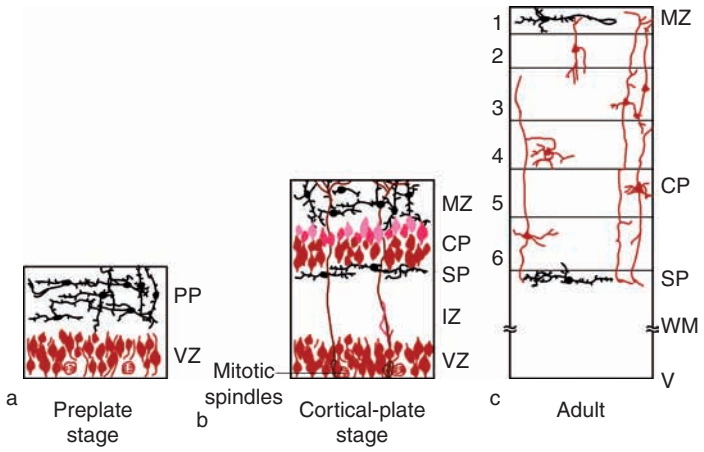

Figure 1 Normal cortical development: (a) preplate stage; (b) cortical-plate stage; (c) adult stage. The preplate stage (a) is marked by the formation of the preplate (PP), a plexiform layer that forms around embryonic day (E)10–12 in mice and embryonic week 8–9 in humans. The cortical-plate stage (b) is marked by the proliferation of neuroblasts in the ventricular zone (VZ) (indicated by mitotic spindles) and the migration of these cells along radial glial fibers (brown processes). Migrating neurons penetrate the preplate, split the preplate into the subplate (SP) and the marginal zone (MZ), and deposit in the cortical plate (CP). Neurons generated later must pass through the SP and pass previously deposited neurons to reach the outer limits of the developing CP (indicated by the progressively lighter shades of green in more-superficial CP neurons). The cortical-plate stage lasts from E10 to E17 in mouse and embryonic week 10–18 in humans. The adult stage (c) is marked by the lamination of the CP into a six-layered structure. IZ, intermediate zone; V, ventricle; WM, white matter. Adapted from Gleeson JG and Walsh CA (2000) Neuronal migration disorders: From genetic diseases to developmental mechanisms. *Trends in Neuroscience* 23(8): 352–359, with permission from Elsevier.

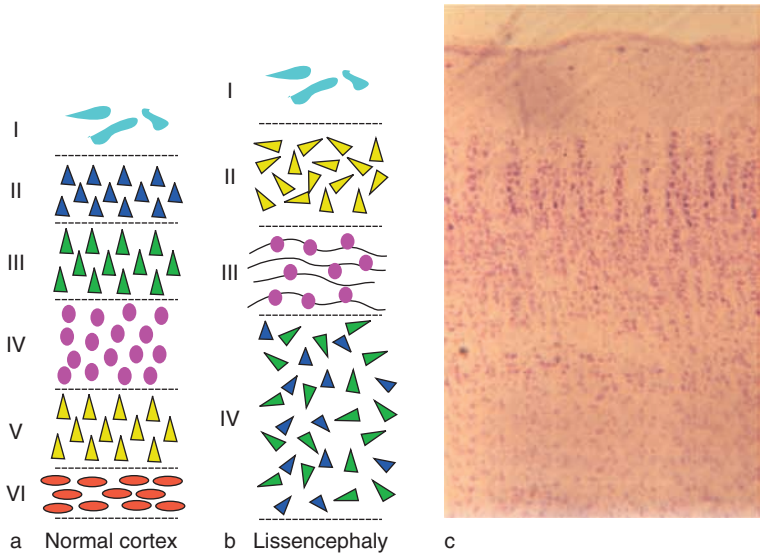

Figure 2 Coritcal layering: (a) normal cortical laminar organization (six layers); (b) cortical layering in classical lissencephaly (four nonorganized layers); (c) histopathological section from a patient with lissencephaly. Adapted from Golden JA and Harding BN (2004) *Pathology & Genetics: Developmental Neuropathology*. Basel: ISN Neuropath Press, with permission from Blackwell Publishing.

is responsible for the autosomal form of LIS, whereas the double-cortin gene (*DCX* or *XLIS*) (OMIM: *300121) is X-linked. Although mutations in either gene can result in either LIS or SBH, most cases of classic LIS are due to deletions or mutations of *LIS1*, whereas most cases of SBH are due to mutations of *DCX*. *LIS1* mutations result in more severe LIS in the posterior brain regions (posterior (p) > anterior (a) gradient), whereas *DCX* mutations result in more severe LIS in the anterior brain regions (a > p gradient) (Figure 3). Among all the patients with isolated LIS, 40% exhibit a deletion involving the entire *LIS1* gene, and 25% show an intragenic mutation (4% gross rearrangement, 17% deletion/truncating mutations, and 4% missense mutations). Patients with missense mutations generally have less severe malformations and may accordingly present with milder motor and cognitive impairment. Severe truncating

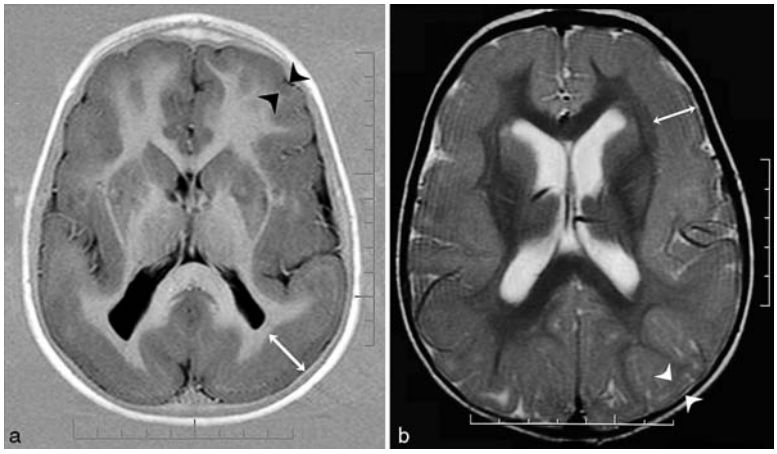

Figure 3 Brain MRI scan; axial section: (a) classical lissencephaly in a boy with *LIS1* gene mutation; (b) lissencephaly in a girl with *DCX* mutation. In (a), there is a typical posterior > anterior malformative gradient with relative preservation of the gyral pattern and cortical thickness in the anterior brain; cortical thickness is around 6 mm in the frontal lobes (two black arrowheads; normal cortical thickness = 4 mm) and around 3 cm in the posterior brain (white arrow). In (b), there is a typical anterior > posterior malformative pattern; cortical thickness is around 2 cm in the frontal lobes (white arrow) and around 4 mm in the posterior brain (two white arrowheads).

mutations cause severe LIS, whereas milder mutations, usually missense mutations, cause pachygyria and rare cases of SBH. Mosaic mutations of *LIS1* also cause SBH in the posterior brain.

DCX mutations classically cause the striking SBH phenotype, in which a second band of cortical neurons exists within the white matter, below the true cortex (Figure 4). Women with *DCX* mutations have an anteriorly predominant band/pachygyria. However, rare carrier women have been described who harbor missense mutations but have normal brain magnetic resonance imaging (MRI), due to either favorable X-inactivation skewing or to mutations having mild functional consequences. Mutations of the coding region of *DCX* have been found in all reported pedigrees, including families in which females have SBH and males have LIS, and in approximately 80% of sporadic females and 25% of sporadic males with SBH. Genomic deletions of the *DCX* gene have been identified in females with sporadic SBH and in males with X-linked lissencephaly. Maternal germ-line or mosaic *DCX* mutations may occur in approximately 10% of cases of either SBH or XLIS. Hemizygous males with *DCX* mutations have classic lissencephaly, but rare boys with missense *DCX* mutations with an anteriorly predominant SBH have also been described.

Recently, it has been demonstrated that a mutation in the guanosine triphosphate (GTP) binding pocket of a-1 tubulin (*Tuba1*) generates a mutant mouse with abnormalities in the laminar architecture of the hippocampus and cortex, accompanied by impaired neuronal migration. Mutations in the human homolog of *Tuba1*, *TUBA3*, were subsequently identified in two patients with LIS. The first patient presented classic LIS with a thick disorganized cortex, epilepsy, severe mental retardation, and motor impairment. The second patient showed a less severe phenotype, with pachygyria in the temporal regions and around the Rolandic sulcus, accompanied by abnormal organization of the hippocampus.

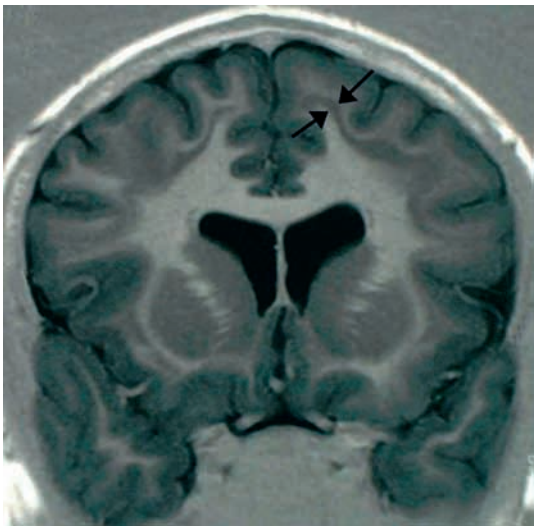

Figure 4 Brain MRI scan; coronal section of a young woman with a missense mutation of the *DCX* gene. The gyral pattern of the cortex looks normal, but beneath the cortex, and separated from it by a thin layer of white matter, is present a subcortical laminar (band) heterotopia (black arrows) showing the same signal intensity as the cortex. The thickness of the heterotopic band is variable from approximately 3 mm to 1 cm.

Experimental Models and Pathogenesis

LIS1 encodes a 45 kDa protein (PAFAH1B1), which is similar to the α subunit of heterotrimeric G-proteins and functions as a regulatory subunit of platelet-activating factor acetylhydrolase (PAF-AH), an enzyme that degrades the bioactive lipid PAF. PAFAH1B1 homozygous mutant mice die at the trophoblast stage, whereas heterozygous mice are grossly normal but show a dose-dependent histopathological disorganization of cortical lamination, as well as hippocampal and cerebellar cortical defects. A reduction in the migration of *in vitro* cerebellar granule cells occurred on treatment with a PAF agonist, implying that the LIS1 protein function in PAF-AH is related to its essential role in migration. The LIS1 protein has been shown to co-localize with microtubules and promote their stabilization. Proper nuclear migration in the *Aspergillus nidulans* depends on NudF, an ortholog of LIS1. The human homolog of NudC, which associates with NudF in yeast, in conjuction with LIS1 regulates dynein activity during mitosis in vertebrate cells, suggesting that classic LIS is caused by a loss of dynein function, which is required for nuclear translocation. This hypothesis is supported by the demonstration that LIS1 also interacts with NUDEL and mNudE to regulate cytoplasmic dynein motor function and location within the cell. Loss of function of either NUDEL, LIS1, or of dynein impact on the microtubule network that couples the centrosome and nucleus, leads to a defect in nuclear translocation, cell migration, and positioning during cortical development. The LIS1–NUDEL pathway has also been linked to the Reelin–/Cdk5–p35 pathway. The phosphorylation of NUDEL by Cdk5 appears to control its cellular localization and probably influences dynein motor function (Figure 5). Thus, the LIS1 protein probably exerts its effects on migration through microtubules.

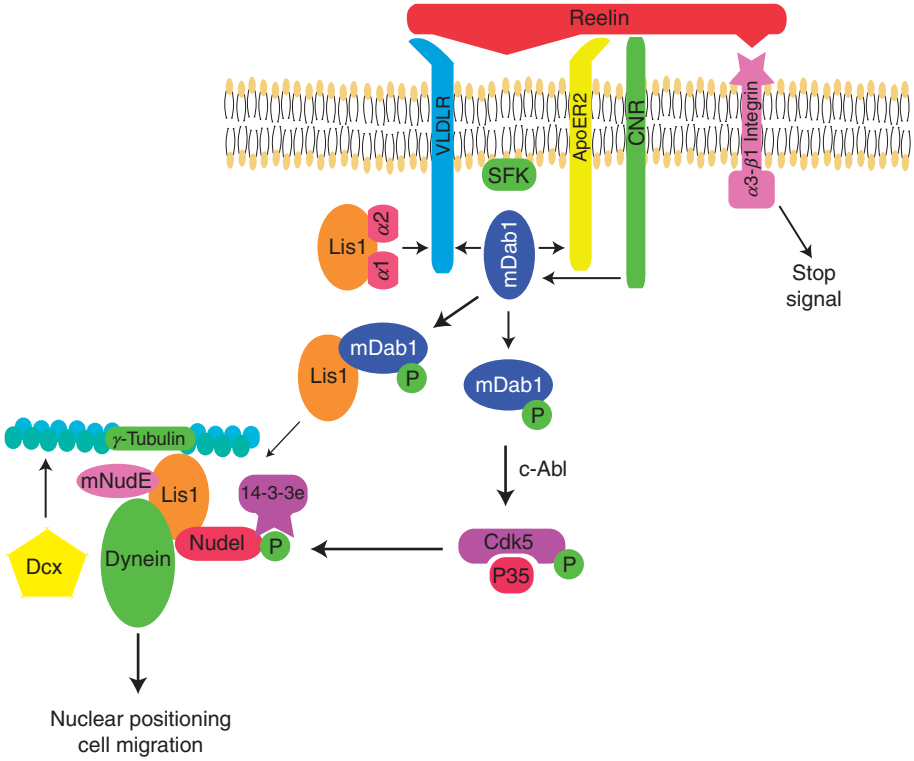

Figure 5 Lis1 and Reln signaling pathways for neuronal migration. Extracellular Reln binds to one of three receptors complexes: cadherin-related neuronal receptor (CNR), VLDLR/ApoER2 or 3β-integrin. This binding activates mDab1, which can activate Cdk5/p35. Cdk5 can phosphorylate many intracellular targets, including Nudel. Phosphorilated Nudel forms a complex with Lis1, mNudE, dyneins, and microtubules. This complex is necessary for cell migration and nuclear positioning. Dcx is also believed to interact with Lis1 and modulate microtubule function. In addition, RELN binding to the VLDLR and ApoER2 causes src-family kinase (SFK) activation and Dab1 phosphorylation. Upon Reelin stimulation, phosphorylated Dab1 interacts with Lis1. Lis1 also binds the catalytic subunits of the Pafah1b complex (α1 and α2) as well as components of the cytoplasmic dynein complex. The α1 and α2 subunits also bind to VLDLR and compete with Dab1 for receptor occupancy. The binding of catalytic Pafah1b subunits to VLDLR may displace P-Dab1 and promote its interaction with Lis1. ApoER2, apolipoprotein E receptor 2; VLDLR, very low density lipoprotein receptor. Adapted from Golden JA and Harding BN (2004) *Pathology and Genetics: Developmental Neuropathology*. Basel: ISN Neuropath Press, with permission from Blackwell Publishing.

DCX encodes for a 40 kDa microtubule-associated protein (DCX) that is expressed in migrating neuroblasts exclusively in the fetal brain, including the forebrain and cerebellum. The DCX protein contains two tandem evolutionarily conserved repeats (doublecortin domain) that form a b-grasp superfold. Each of the repeats binds to tubulin but not to microtubules; so both repeats are necessary for microtubule polymerization and stabilization. DCX was disrupted in mice using targeted mutagenesis. However, these mice did not have a cortical defect. More recently, using the RNA interference method, Dcx function was reduced in a rat model that resulted in SBH resembling that seen in humans. This model made it possible to directly examine the role of DCX in radial migration and associated cellular morphogenesis through cortical development. In particular, DCX is required for all radial migrating neocortical neurons to enter the cortical plate from the interventricular zone.

Lis1 and Dcx display overlapping localization and may interact. Dcx and Lis1 co-immunoprecipitate from brain lysate, and purified Dcx and Lis1 bind cooperatively to microtubules. Studies on mouse cerebellar granule neurons demonstrated that Lis1 and Dcx interact with dynein to mediate nucleus–centrosome coupling in neuronal migration, suggesting that proper coupling may be critical in neuronal migration. In addition, LIS1 and DCX interact physically, both *in vivo* and *in vitro*, suggesting that the addition of DCX and LIS1 to tubulin enhances polymerization in an additive fashion. In *in vitro* competition assays, when LIS1 is added first, DCX competes with LIS1 in its binding to microtubules, but when DCX is added prior to the addition of LIS1 it enhances the binding of LIS1 to microtubules. LIS1 and DCX cross-talk is important to microtubule function in the developing cerebral cortex. The interaction of both DCX and LIS1 with microtubules may explain the striking similarities between the lissencephalic phenotypes that result from mutations in either gene.

Epidemiology and Clinical Features

Classical lissencephaly is rare with a prevalence of 11.7 per million births. All patients have early developmental delay, early diffuse hypotonia, later spastic quadriplegia opisthotonus, and eventual severe or profound mental retardation. Rare patients with pachygyria may have moderate mental and motor impairment. Some children with LIS live more than 20 years, although anecdotal experience suggests that the life span is less than 10 years in most patients. Seizures occur in over 90% of children, with onset before 6 months in approximately 75%. Approximately 80% of children have infantile spasms, although the electroencephalogram (EEG) does not show typical hypsarrhythmia. Later, most children have mixed seizure disorders. Because most clinical and neurophysiological studies on children with LIS were conducted before a genetic distinction was made between XLIS and LIS1, it is unknown whether these two forms have distinctive electroclinical patterns.

Genetic Counseling

All reported *LIS1* mutations are *de novo*. Nevertheless, if an *LIS1* mutation is found, it is correct to perform mutation analysis on both parents. Given the theoretical risk of germ-line mosaicism in either parent (which has never been demonstrated for *LIS1*), a couple with a child with LIS is usually given a 1% recurrence risk.

When a mutation in the *DCX* gene is found in a boy with LIS, the mutation analysis of *DCX* should be extended to the mother, even if her brain MRI is normal. If the mother is a mutation carrier, the mutation will be transmitted according to Mendelian inheritance. If the mother is not a carrier, she can still be at risk for germ-line mosaicism; the risk of transmitting the mutation might be roughly estimated at around 5%. For this reason, a prenatal diagnosis is indicated in every pregnancy of a woman who has a child with features of X-linked LIS with a *DCX* mutation.

Miller–Dieker Syndrome

Miller–Dieker syndrome (MDS) is caused by a contiguous gene deletion. Classic LIS is accompanied by distinct dysmorphic facial features, including a prominent forehead, bitemporal hollowing, flattened ear helices, mild hypertelorism, epicanthic folds, a short nose and anteverted nares, prominent lateral nasal folds, and a round philtrum. Additional malformations can be observed. Affected children have severe delay, epilepsy, and feeding problems. Deletions of 17p13.3, including the *LIS1* gene, are found in almost 100% of patients. The deletion of two additional genes, *CRK* and *14-3-3ε* telomeric to LIS1, may contribute to the most severe LIS grade and dysmorphic features observed in MDS.

X-Linked Lissencephaly with Corpus Callosum Agenesis and Ambiguous Genitalia

X-linked lissencephaly with absent corpus callosum and ambiguous genitalia (XLAG) is a severe malformation syndrome that is observed only in boys. The anatomoclinical spectrum includes LIS with a posterior-to-anterior gradient and an only moderate increase of the cortical thickness (only 6–7 mm in XLAG, vs. 15–20 mm in classic LIS) (Figure 6), an absent corpus callosum; poorly delineated and cavitated basal

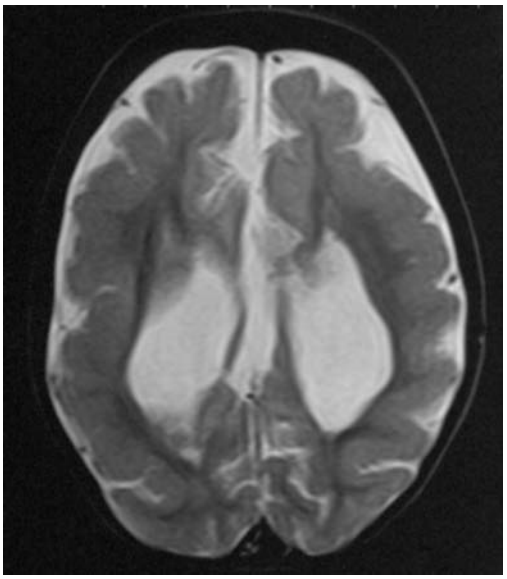

Figure 6 Brain MRI scan; axial section of a 1-year-old boy with X-linked lissencephaly, with corpus callosum agenesis and ambiguous genitalia due to the mutation of the ARX gene. Note the absence of the corpus callosum with ventriculomegaly and lissencephaly.

ganglia; postnatal microcephaly; neonatal-onset epilepsy; hypothalamic dysfunction, including deficient temperature regulation; chronic diarrhea; and ambiguous genitalia with micropenis and cryptorchidism. Early death is not uncommon. Brain neuropathology reveals an abnormally laminated cortex containing exclusively pyramidal neurons, with a pattern suggesting the disruption of both tangential and radial migration, dysplastic basal ganglia, hypoplastic olfactory bulbs and optic nerves, abnormal gliotic white matter containing numerous heterotopic neurons, and complete agenesis of the corpus callosum without Probst bundles.

Mutations of the X-linked *aristaless*-related homeobox gene (*ARX*) (OMIM: *300382) were identified in individuals with XLAG and in some female relatives. Females carrying *ARX* mutations usually have normal cognition and may either have a normal brain MRI scan or show partial or complete agenesis of the corpus callosum. However, mild mental retardation and epilepsy have been reported in rare carriers. Mouse *Arx* and human *ARX* are expressed at high levels in both dorsal and ventral telencephalon, including the neocortical ventricular zone and germinal zone of the ganglionic eminence, with less intense signals in the subventricular zone, cortical plate, hippocampus, basal ganglia, and ventral thalamus. The murine *Arx* gene is specifically expressed in the γ-aminobutyric acid (GABA)ergic neural lineages both during development and in adulthood. *Arx*-deficient mice show deficient tangential migration and abnormal differentiation of interneurons containing GABA(GABAergic interneurons) in the ganglionic eminence and neocortex. These findings suggest that the different *ARX* mutations may affect in different grades the overall GABAergic system functionality, leading to the clinical features of XLAG in humans, and might account for the severe neonatal epileptic encephalopathy with suppression-burst EEG that is often observed in affected boys. The mutations of the *ARX* gene in XLAG patients are predominantly premature termination mutations. Missense mutations are less common and are located essentially in the homeobox domain. Patients carrying nonconservative missense mutations within the homeodomain showed less severe XLAG, whereas conservative substitution in the homeodomain causes Proud syndrome (agenesis of the corpus callosum (ACC) with abnormal genitalia). A nonconservative missense mutation near the C-terminal aristaless domain causes unusually severe XLAG with microcephaly and mild cerebellar hypoplasia.

Autosomal Recessive Lissencephaly with Cerebellar Hypoplasia

An autosomal recessive form of LIS associated with severe abnormalities of the cerebellum, hippocampus, and brain stem was mapped to 7q22, and mutations were identified in the *RELN* gene (OMIM: *600514). Patients presented dysmorphic facial features, including bitemporal hollowing, sloping forehead, widely set eyes, and prominent nasal bridge. A similar phenotype had previously been described as the Norman–Roberts syndrome. *RELN* encodes a large (388 kDa) extracellular matrix protein that acts on migrating cortical neurons by binding to the very low-density lipoprotein receptor (VLDLR), the apolipoprotein E receptor 2 (ApoER2), α3-β1 integrin, and cadherins-related receptors (CNRs). In mice, *Reln* mutations cause cerebellar hypoplasia, abnormal cerebral cortical neuronal migration, and abnormal axonal connectivity. Neurons in affected mice fail to reach their correct location in the developing brain, disrupting the organization of the cerebellar and cerebral cortices and other laminated regions. In this animal model, the cortical layering appears inverted. Thus, Reln is thought to control cell–cell interactions critical for cell positioning in the brain. Binding of Reln to the α3-β1 functions as a stop signal after neuronal migration; however, the downstream components within the cells that regulate the migration stop are not known. Interactions between the Reln and Lis1 signaling pathways have been demonstrated. The interaction of Reln with the CNRs phosphorylates the cytoplasmic messenger mDab1, possibly through the

VLDLR. Phosphorylated mDab1 can interact with a variety of proteins, including Src, which is involved in the cytoskeletal remodeling, or the protooncogene c-Abl. Once activated, c-Abl phosphorylates and activates Cdk5. Cdk5 and p35 (another activator of Cdk5) have also been implicated in directing neurons to the appropriate location within the cerebral cortex. The interaction between Lis1 and Dab1 has also been demonstrated. This interaction takes place when Dab1 is phosphorylated on tyrosine residues in response to Reln. In addition, it has been demonstrated that both the α subunits of the Pafah1b bind the VLDLR, suggesting that the Pafah1b complex modulates the Reln pathway. The binding of catalytic Pafah1b subunits to the VLDLR may displace phosphorylated Dab1 and promote its interaction with Lis1 (Figure 5).

Nine additional phenotypes have been also associated with the LIS–SBH spectrum, although they do not belong to the subgroup of the type 1 LIS. Clinical features and associated genes or loci, if identified, are reported in Table 1.

Periventricular Nodular Heterotopia

Definition and Histopathology

PNH, most often bilateral, consists of nodules of gray matter located along the lateral ventricles with a total failure of migration of some neurons. PNH can be highly variable in extent, ranging from isolated single nodules to confluent bilateral lesions. PNH is most common near the angle of the lateral ventricles and around the temporal horns. The cortex overlying PNH often shows macroscopic abnormalities, most commonly polymicrogyria. Nodules of PNH may sometimes show some rudimentary lamination, and its neurons show changes in size consistent with developmental age. A variety of neuronal types are present, including small neurons that express parvalbumin, calretinin, or neuropeptide Y (NPY). Decreased expression of the *N*-methyl-D-aspartate (NMDA) receptor and its signaling pathway has been reported in PNH nodules and overlying dysplastic cortex.

Early fluorodeoxyglucose-positron emission tomography studies showed that heterotopia has the same

Table 1 Additional phenotypes associated with the lissencephaly–SBH spectrum but not belonging to the subgroup of the type 1 lissencephalies

Syndrome	*Clinical features associated with lissencephaly/pachigyria*	*Locus (gene)*	*OMIM accession number*
Baraitser–Winter syndrome	Eye coloboma, ptosis, hypertelorism, epicanthic folds, broad nasal bridge, long philtrum, thin upper lip, telechantus, short stature	Possibly autosomal recessive	
Walker–Warburg syndrome (HARD +/− E syndrome)	Congenital muscular dystrophy, congenital retinal nonattachment with or without microophthalmia and persistent hyperplastic primary vitreous	9q34.1 (*POMT*1), 14q24.3 (*POMT2*)	236670
Muscle–eye-brain disease (MEB)	Congenital muscular dystrophy, congenital hypotonia and muscle weakness, severe visual failure, uncontrolled eye movements, myopia	19q13.3 (*FKRP*), 1p34–p33 (*POMGNT1*)	253280
Fukuyama congenital muscular dystrophy (FCMD)	Generalized muscle weakness, hypotonia, speech defect	9q31 [*fukutin* (*FCMD*)]	253800
Lissencephaly and bone dysplasia	Craniofacial edema and arthrogryposis; epiphyseal stippling of cervical vertebrae, feet, and sacrum; shortened metacarpal bones and hypoplastic phalanges	Supposed autosomal recessive	601160
Neu-Laxova syndrome (NLS)	Severe subcutaneous edema, atrophic muscles, camptodactyly, syndactyly of toes and fingers, hypoplastic genitalia, hypertelorism, protruding eyes	Unknown	256520
Lissencephaly with cleft palate and cerebellar hypoplasia	Cleft palate	Unknown	604382
Craniotelencephalic dysplasia	Frontal bone protrusion, encephalocele, craniosynostosis, septooptic dysplasia	Unknown	218670
Muscular dystrophy with severe central nervous system atrophy and absence of large myelinated fibers	Hypertelorism, telecanthus, small and posteriorly angulated ears, micrognathia and high-arched palate, severe hypotonia, muscle weakness	Unknown	601170

OMIM, Online Mendelian Inheritance in Man; SBH, subcortical band heterotopia.

metabolic activity as normal gray matter. Functional MRI studies indicated that PNH can also be functionally integrated in the motor circuits, suggesting that neurons that have failed to migrate have maintained the information that allows them to assemble in functionally active aggregates and to participate in integrated networks.

Although most patients with PNH come to medical attention because they have epileptic seizures, there is a wide spectrum of clinical presentations with some correlation between the size of PNH and the likelihood of concomitant structural abnormality of the cortex and clinical severity.

Genetics

PNH occurs much more frequently in women as an X-linked trait (Figure 7), which is associated with high rates of prenatal lethality in male fetuses and a 50% recurrence risk in the female offspring of affected women. Almost 100% of families with X-linked bilateral PNH and approximately 26% of sporadic patients harbor mutations of the *filamin A* gene (*FLNA*) (OMIM: +300017). The low percentage of *FLNA* mutations in sporadic cases might be explained by somatic mosaicism. Heterozygous females have normal to borderline intelligence and epilepsy. Coagulopathy and cardiovascular abnormalities have been observed in some patients. A few living male patients with bilateral PNH due to *FLNA* mutations are on record. Mild missense mutations or mosaic mutations, probably causing a limited functional defect of the *FLNA* protein, account for the survival of affected males, who can, in turn, transmit their genetic defect to their daughters. Rare patients of both genders with *FLNA* mutations had unilateral PNH or classic bilateral PNH associated with Ehlers–Danlos syndrome.

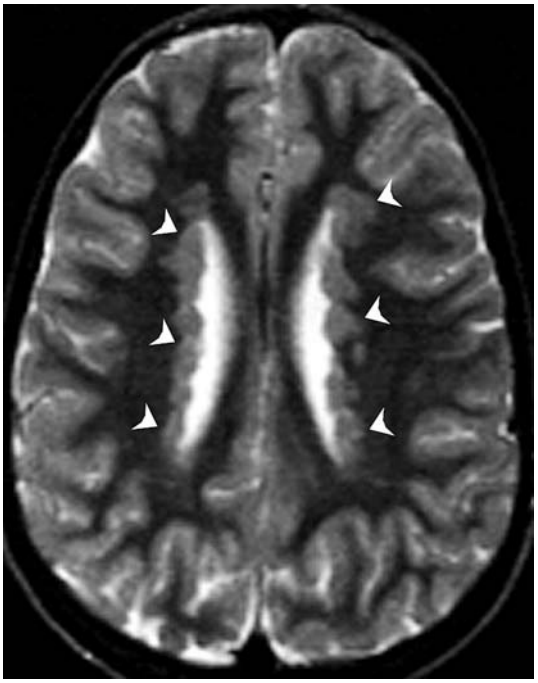

Figure 7 Brain MRI scan; axial section of a typical, classic bilateral periventricular nodular heterotopia in a woman with a missense mutation of the *FLNA* gene. The bilateral nodules of subependymal heterotopia are contiguous and rather symmetric, extensively lining the ventricular walls (white arrows). Patients with periventricular heterotopia different from the kind reproduced in this picture have very low chances of carrying a mutation on the *FLNA* gene.

Mutations in other genes can cause PNH. A rare recessive form due to mutations of the adenosine diphosphate-ribosylation factor guanine nucleotide exchange factor 2 (*ARFGEF2*) (OMIM: *605371) has been reported in two consanguineous pedigrees. This gene encodes for the protein brefeldin A-inhibited GEF2 (BIG2), which is required for vesicle and membrane trafficking from the *trans*-Golgi network. Impaired vesicle trafficking prevents transport to the cell surface of polarized molecules, such as E-cadherin and β-catenin, thereby disrupting proliferation and migration during cortical development. Affected children have microcephaly, severe delay, and early-onset seizures, including infantile spasms. PNH may also be observed in patients with chromosomal rearrangements, such as duplication of 5p15.1 or 5p15.33. PNH has also been associated with hydrocephalus, frontonasal malformation, severe mental retardation and syndactyly, Ehlers–Danlos syndrome, fragile X syndrome, and nephrosis. Whether each of these variants represents a truly distinct compound disorder with a specific genetic mechanism remains to be determined. In some such syndromes, the malformation can result from small chromosomal rearrangements involving the *FLNA* gene and other unknown genes.

FLNA mutations have also been associated with four additional multiorgan syndromes: otopalatodigital syndrome type 1 and 2 (OPD1 and OPD2), frontometaphyseal dysplasia (FMD) and Melnick–Needles syndrome. In contrast to the loss-of function *FLNA* mutations seen in PNH patients, these four syndromes appear to involve gain-of-function *FLNA* mutations.

Experimental Models and Pathogenesis

FLNA maps to Xq28 and codes for the F-actin-binding cytoplasmic cross-linking phosphoprotein Filamin A (FLNA), composed by three major functional domains: (1) a tandem N-terminal calponin-homology domain (CHD1 and CHD2), conferring F-actin binding properties; (2) 15 + 8 internally homologous Ig-like

repeats separated by a short run with an unique sequence (hinge 1), important for flexibility; and (3) a second short run (hinge 2) followed by the C-terminal repeat 24, which is important for binding to a wide range of proteins and for dimerization. The N-terminal actin-binding domain displays strong structural and functional similarity to the N-terminal domains of dystrophin, α-actin, and β-spectrin. The 24th repeat is truncated, allowing dimerization of two *FLNA* molecules at the C-terminus. FLNA dimers bind membrane-associated proteins such as b1 and b2 integrins, tissue factor, and presenilin1. FLNA can also bind other membrane-associated molecules, such as glycoprotein Ib-α, through repeats further from the C-terminus. The specific roles of FLNA and its association with pathological conditions are yet to be fully understood. Additional functional complexity is conferred by the existence and by the time/tissue-specific expression of the two human paralogous genes, *Filamin B* and *Filamin C*. The Y-shaped homodimerized FLNA promotes the organization of actin filaments in orthogonal networks of the cytoskeleton. A similar mechanism might also be hypothesized for Filamin B and C, even though their heterodimerization has been demonstrated only *in vitro*. In addition, Filamin A and Filamin B co-localize within neuronal precursors, and the heterodimerization of these two isoforms has been hypothesized. Interactions of the dimer with integrin receptors, transmembrane proteins, and several signaling proteins are believed to integrate extracellular and cytoplasmic signals with cellular cytoskeleton rearrangements and membranes reshaping. Filamin homologs have been shown to be implicated in the regulation of cell stability, protrusion, and motility across various biological systems. The mouse *Filamin A* ortholog was abundantly expressed in the cell soma and leading processes of migratory neurons, and it is widely distributed across the cortical mantle, reaching the highest levels in the ventricular zone within the cortex during neurogenesis. FLNA probably influences neuroblast migration during cortical development in vertebrates, and PNH in humans probably results from the disruption of this process. A model of *FLNA* function postulates that, in *FLNA*-deficient neurons, part of the essential migratory motor is defective and that these defective neurons are incapable of migration. An alternative model proposes that FLNA acts earlier in development, as a component of a switch that is required for a neuron to become competent for subsequent migration. In the latter model, FLNA still functions by structuring actin networks at the cell periphery but, instead of acting at the leading edge of the migrating neuron, FLNA maintains a static focal contact between a stationary neuron and a neighboring radial glial cell. FLNA also promotes orthogonal branching of actin filaments and is important for coagulation and vascular development. These coagulation and vascular-related functions of FLNA might account for the prenatal male lethality observed in most pedigrees, a hypothesis supported by the birth of a male infant to a mother with PNH, who shared her affected haplotype and who died postnatally of severe widespread hemorrhage. Cardiovascular or gut malformations might also account for prenatal male lethality. The complete loss of Flna in mice results in embryonic lethality with severe cardiac structural defects involving the ventricles, atria, and outflow tracts, as well as widespread aberrant vascular patterning. Despite these widespread developmental defects, the migration and motility of many cell types, including Flna-deficient neural crest cells, do not appear to be affected. Moreover, Flna-null vascular endothelial cells display defects in cell–cell contacts, suggesting cell motility-independent functions of FLNA in cell–cell contacts and adherens junctions during the development of many organs.

Genetic Counseling

Genetic counseling is relatively easy in familial cases with a clear X-linked pattern of inheritance. Classic PNH with cerebellar hypoplasia and no dysmor phic features is much more frequent in women and more likely to be due to *FLNA* mutations than in atypical cases. Among carrier women, approximately half have *de novo* mutations of *FLNA*, whereas the remaining half have inherited mutations. Although maternal transmission is much more likely, father-to-daughter transmission is possible, implying that either parent can transmit the mutation to a female proband. If neither parent has epilepsy or cognitive impairment, the proband's mother should be studied first to confirm the mutation or the brain abnormality. If the mother is mutation-negative and the proband is a female, the father should also be studied. Given that the germ-line mosaicism of *FLNA* has never been reported, the recurrence risk (for other children) seems to be very low when a mutation is found in the proband but neither parent is a carrier.

See also: Cerebral Cortex; Corpus Callosum: Agenesis; Double Cortex; Lissencephaly; Microtubules: Organization and Function in Neurons.

Further Reading

Barkovich AJ, Kuzniecky RI, Jackson GD, Guerrini R, and Dobyns WB (2005) A developmental and genetic classification for malformations of cortical development. *Neurology* 65: 1873–1887.

Cardoso C, Leventer RJ, Dowling JJ, et al. (2002) Clinical and molecular basis of classical lissencephaly: Mutations in the *LIS1* gene (PAFAH1B1). *Human Mutation* 19(1): 4–15.

des Portes V, Pinard JM, Billuart P, et al. (1998) A novel CNS gene required for neuronal migration and involved in X-linked subcortical laminar heterotopia and lissencephaly syndrome. *Cell* 92: 51–61.

Dobyns WB, Truwit CL, Ross ME, et al. (1999) Differences in the gyral pattern distinguish chromosome 17-linked and X-linked lissencephaly. *Neurology* 53(2): 270–277.

Fox JW, Lamperti ED, Eksioglu YZ, et al. (1998) Mutations in filamin 1 prevent migration of cerebral cortical neurons in human periventricular heterotopia. *Neuron* 21: 1315–1325.

Gleeson JG and Walsh CA (2000) Neuronal migration disorders: from genetic diseases to developmental mechanisms. *Trends in Neuroscience* 23(8): 352–359.

Golden JA and Harding BN (2004) *Pathology & Genetics: Developmental Neuropathology.* Basel: ISN Neuropath Press.

Guerrini R and Marini C (2006) Genetic malformations of cortical development. *Experimental Brain Research* 173(2): 322–333.

Guerrini R, Mei D, Sisodiya S, et al. (2004) Germline and mosaic mutations of FLN1 in men with periventricular heterotopia. *Neurology* 63(1): 51–56.

Kato M and Dobyns WB (2003) Lissencephaly and the molecular basis of neuronal migration. *Human Molecular Genetics* 12(R1): R89–R96.

Parrini E, Ramazzotti A, Dobyns WB, et al. (2006) Periventricular heterotopia: Phenotypic heterogeneity and correlation with Filamin A mutations. *Brain* 129: 1892–1906.

Rakic P (2006) A century of progress in corticoneurogenesis: From silver impregnation to genetic engineering. *Cerebral Cortex* 16(supplement 1): i3–i17.

Sheen VL, Dixon PH, Fox JW, et al. (2001) Free mutations in the X-linked filamin 1 gene cause periventricular nodular heterotopia in males as well as in females. *Human Molecular Genetics* 10(17): 1775–1783.

Sheen VL, Ganesh VS, Topcu M, et al. (2004) Mutations in ARFGEF2 implicate vesicle trafficking in neural progenitor proliferation and migration in the human cerebral cortex. *Nature Genetics* 36(1): 69–76.

Wynshaw-Boris A and Gambello MJ (2001) LIS1 and dynein motor function in neuronal migration and development. *Genes & Development* 15(6): 639–651.

Zhang G, Assadi AH, McNeil RS, et al. (2007) The Pafah1b complex interacts with the Reelin receptor VLDLR. *PLoS ONE* 28: e252 [online].

Relevant Website

http://www.ncbi.nlm.nih.gov – National Institutes of Health, Online Mendelian Inheritance in Man.

Localizing Signal Sources

G M Klump, Universität Oldenburg, Garching, Germany

© 2009 Published by Elsevier Ltd.

In animal communication, it is advantageous for the receiver to locate the source of a signal (i.e., the sender) that it has detected and possibly recognized. For example, if the sender is a territorial competitor, a territory holder as the receiver would benefit by being able to pinpoint the competitor's location and to evaluate whether this competitor already intruded into the territory. In many behavioral contexts, such as in the case of searching for a mate or searching for food, the response of the receiver will be an approach toward the source that also requires localization (a response that is composed of orienting toward the signal source and the approach movement is termed 'taxis'). Even in the case that fleeing rather than approaching the source of a signal is the adaptive response, a receiver may benefit by choosing a direction of the escape route away from the signal source that minimizes its risk.

Localizing a signal source in three dimensions in general requires being able to determine the direction from which a signal emanated. Directional information can be separated into azimuth (angle on the horizontal plane) and elevation (angle on the vertical plane). If the location of the origin of a signal has to be determined on the spot, it is also necessary to obtain a measure of the distance between the receiver and the signal source. Alternatively, a receiver may combine the information on the direction with an approach to the signal source that finally leads it to its location. Which strategy for localizing the signal source will be more advantageous will depend on (among other factors) the modality of the signal, the physical effects of the environment on the signal when being transmitted from the sender to the receiver, and the representation of the directional cues in the sensory organs and the neural system of the receiver. Furthermore, senders of communication signals are not necessarily stationary but can move while signaling. The trajectories of a moving sender should also be evaluated when locating the signal source.

This article focuses on the localization of sound signals and only briefly discusses the localization of optical and chemical signals. Some aspects (e.g., the short-distance localization of signal sources in the perception of electrical signals by fish) will not be discussed.

Evaluating Cues for Localizing the Signal Source

Visual Signals

The physical characteristics of the signals and the receiver's sensory organs will determine which cues can be used for localization of communication signals. Eyes often can provide information on azimuth and elevation simply as a result of the physical projection of the image on a surface of receptors that forms a retina. Light from different locations in the environment surrounding the receiver will excite different spots on the retina (or the compound eye of an insect), and the position of the source of a visual signal in the environment can be determined by the place of excitation on the retina. Determining the distance of a signal source for an optical signal requires the evaluation of additional information. Stereo vision is one mechanism for distance (i.e., depth) perception exploiting the disparities of the projection of the images on the retinae of the two eyes. Another possibility to infer the distance of a stationary visual object is to make use of the information of motion parallax and the visual flow field that is generated by movement of the receiver. Objects close to the receiver will elicit a larger shift of the image on the retina and a faster optical flow than objects at a greater distance. This cue can also be used in monocular vision. Finally, previously acquired knowledge about the object can be used to infer the distance. For example, if the size of the object is known, then the distance can be estimated based on the size of the projection on the retina.

Acoustic Signals

Unlike for localization of visual targets, there is no direct spatial representation of the environment in each receiving sensory organ that can be exploited to localize the source of acoustic communication signals. Characteristics of acoustic signals, such as the overall intensity, spectral composition, time of arrival, or other features, will vary depending on the azimuth and elevation of the sound source or its distance from the receiver. In general, the auditory system infers azimuth and elevation of a signal source by comparing the input to two receiving sensory organs – that is, the ears. However, some sound-receiving organs can also encode directional information about the sound source directly without requiring a comparison between two sound-receiving organs.

Spectral composition of the signals at the two ears is one cue that may be exploited to infer the source location. Due to the physical characteristics of the

outer ear and the immediate surround created by the head, body, etc., differences between the ears in signal amplitude at the various frequencies comprising a specific signal (i.e., the amplitude spectrum) can change systematically with sound source azimuth or elevation. The amount of amplitude change that occurs for signals with a different frequency composition impinging on the two acoustically separated ears will depend on the wavelength of the signal and the dimensions of the structure carrying the ears (mostly the head). If the wavelength is similar to the size of the head (or other structure), it can provide a sound shadow that generates interaural intensity differences that change systematically in relation to the azimuth and elevation of the sound source. Many mammals use ultrasonic communication signals with frequencies above 20 kHz (and wavelengths below 2 cm) that will generate interaural intensity differences that are sufficiently large for localization. If the wavelength of the sound is much larger than the diameter of the head (e.g., 10 times larger), interaural intensity differences are rather small (not more than 1 or 2 dB) and accurate sound localization based on intensity differences is compromised. A minimum interaural intensity difference of 1 dB is needed to perceive a change in the angle of sound incidence. However, in many land vertebrates with small heads and a small interaural distance that use low-frequency communication signals with a wavelength that is much larger than their head diameter, an acoustic coupling between the middle ear cavities or an acoustic connection of other inputs to the middle ear cavity (e.g., in many frogs) can be observed. This acoustic coupling allows interference between inputs reaching both sides of the tympanum of each ear, thus producing a pressure gradient that can enlarge interaural intensity differences. A similar mechanism based on acoustic interference of sound impinging on the inside and outside of a tympanum is also found in some insects and can create a large variation of the sound intensity experienced at the ears when the azimuth angle of sound incidence changes.

So far, interaural intensity differences have been discussed without a specific reference to the overall spectrum of the sound. However, the overall shape of the spectrum of a broadband sound also provides important cues for localization. Changes in the shape of the spectrum of a sound may be created by the pinna or other structures of the ear (e.g., the coupled middle ear cavities) that will enhance or reduce the amplitude of frequencies in the spectrum of an incoming sound. In mammals and other vertebrates, the physical filter characteristics of the ear will lead to specific notches in the spectrum of a sound. These notches play an important role in sound source localization. They aid the auditory system in computing the elevation of a sound source or resolve front/back errors in localization (there are multiple sites in space that will produce similar interaural time and intensity differences at the two ears).

The second cue used by the auditory system for localizing signal azimuth is the interaural time difference that is due to the different travel times of the sound from the source to the two ears if the source is located somewhat to the side of the head. Interaural time differences (ITDs) relate to the angle of sound incidence θ as given by

$$\mathrm{ITD} = \frac{ar}{c}\sin(\theta) \qquad [1]$$

where a is a constant factor, r is the head radius, and c is the speed of sound.

According to Kuhn, at low frequencies (i.e., if the product of wave number and the head radius is smaller than approximately 2), it is more appropriate to calculate ITDs with a value of 3 for the factor a, and at higher frequencies it is better to set the factor a to the value of 2. The minimum ITD that can be perceived by a land vertebrate auditory system is on the order of 10–50 μs. In an absolute localization task involving low-frequency sounds (e.g., 1 kHz), this ITD threshold translates to an accuracy of between 4° and 27° for a head diameter of 2.5 cm (e.g., a gerbil) and between 1° and 7° for a head diameter of 15 cm (e.g., a human). Psychophysical measurements of the localization accuracy are consistent with the results of these calculations.

The ITD may be weighted differently at various times after the start of the signal. If the signal has sharp onset transients, these transients arriving at the ear first may be, at least in mammals, quite prominent in determining the perceived location of the signal source. Transients in signals that are due to reflections on surfaces that arrive 10 ms or more later than the primary signal traveling on the direct path from the source to the ears do not strongly contribute to the perceived location of the source (i.e., the primary signal takes precedence). If no strong transients are provided by the signal or if the signal is relatively long, ongoing time differences are also important. For signals composed of low frequencies (i.e., below 2 or 3 kHz in general) that allow the encoding of signal phase by the auditory system (the barn owl may be an exception, with excellent phase locking up to 9 kHz), ITDs may also be derived from the ongoing time differences and not predominantly from transients.

It is obvious from the previous discussion that signals with wavelengths that are much larger than the dimensions of the head (or other parts of an

animal body) separating the ears will only provide minute interaural time and intensity difference cues. A typical example is the acoustic signals used for mate attraction in the green tree frog (*Hyla cinerea*). Female green tree frogs approach males for mating that are advertising their location by high-level calls with peak frequencies of 900 and 3000 Hz (with corresponding wavelengths of 38 and 11 cm, respectively). If the ears of a female with a typical head diameter of 1.2 cm would operate as two separate pressure receivers, the maximum ITD given by eqn [1] is only 52 μs. The available ITD between the ears corresponding to the female green tree frogs' localization accuracy of approximately 20° is 18 μs. However, the related time difference observed in the firing of action potentials of the auditory nerve of the green tree frog for the same change in the frontal angle of sound incidence is approximately a factor of 10 larger (and can be as large as 1.3 ms when the sound is coming from the side). This discrepancy suggests that in this species the ears are not completely decoupled but, rather, may form a pressure gradient receiver system that can provide much larger interaural time and intensity differences than those found in the sounds impinging on the ears. Physical measurements of the acoustic directionality of the frog ear suggest that a number of different sound inputs contribute to the pressure gradient system in the green tree frog and other species.

A small parasitoid fly, *Ormia ochracea*, with two sound-receiving membranes separated by a mere 0.5 mm has evolved another solution to localizing calling crickets that produce signals in the frequency range of 5 kHz (i.e., with a wavelength that is two orders of magnitude larger than the distance between the fly's 'ears'). The localization accuracy in the fly has been described as being so high that it would require a detection of ITDs on the order of 50 ns (i.e., three orders of magnitude higher than that discussed previously). Also in this small fly, the two eardrums are functionally coupled, and the specific anatomy underlying the coupling results in the generation of receptor spike time differences that vary by 3.5 μs per degree change in the angle of sound incidence. The fly's localization accuracy suggests that it can distinguish differences in the timing of the action potentials as small as 7 μs. This task probably requires combining the information from many receptors since the jitter in the timing of the action potentials of individual receptors is an order of magnitude larger than the threshold deduced from the fly's behavioral performance.

So far, we have only considered localization of signal sources using sound pressure or pressure differences. Another cue is provided by the movement of particles that are oscillating back and forth around their resting position at the frequency of the sound driving the movement. Velocity receptors such as the tiny hairs in an arthropod cuticula or the hair cells in the sacculus of the inner ear of a fish can register the particle movement, which is directional. It occurs in the direction of the pressure wave originating from the sound source and opposite this direction (i.e., the particles move on a line pointing to the source). If velocity receptors can respond differentially to the particle movement related to the various angles of sound incidence, this feature can be used to localize a sound source. Afferent neurons innervating the hair cells in the sacculus of a fish, for example, have a distinct polarity that results in a large response along one directional axis and a small response in the direction perpendicular to this axis. Using a population of such directionally selective afferent neurons from the right and left sacculus, the brain can evaluate the location of the source in relation to the body axis of the fish. It has been suggested that this may explain the behavioral observation that fish exhibit differential responses in relation to underwater sound sources placed in different directions.

Chemical Signals

As in the case of auditory localization, localizing sources of chemical signals over a distance by the sense of smell also requires computations by the neural system since the receptor organ does not provide immediate directional cues. Due to the nature of the olfactory signals in the environment of an animal, however, these computations may involve the input from other sensory modalities or require considerable integration of information from the sensory input over extended time periods. Starting at the source of an odor, the turbulent flow of the medium transporting the odor signal (air or water) leads to a breaking up of the odor plume into a patchy landscape of odor eddies. These provide intermittent signals to the receptor organ that are analyzed to localize the odor source. Various mechanisms have been suggested for localizing the source, and some have been tested by implementing these mechanisms in a robot and comparing the robot's behavior to that of an animal such as a lobster. In animals with spatially separated bilateral odor receptor organs, such as the lobster with its lateral antennular flagellum on each side, bilateral comparisons of odor concentrations can be exploited in source localization. Observations in the lobster suggest that its path is continuously affected by the bilateral odor pulse input resulting in eddy chemotaxis. Information about the flow of the medium from the mechanoreceptors is frequently integrated with information on the concentration of the odor:

The animal moves against the current while responding to the odor plume (i.e., it performs an eddy rheo-chemotaxis). A number of variants of these mechanisms have been proposed, and most animals that rely on olfactory cues for their orientation have developed specific behavioral strategies for tracking the source of an odor.

Strategies for Localizing Sources

Two major strategies have been identified for localizing the source of a communication signal. The first, the open-loop strategy, makes use of the input to the receptor organs to calculate the corresponding direction of the signal source instantaneously. Localization ability must be absolute – that is, the receiver must be able to infer the source location in space and it is not sufficient for localizing the signal source if the receiver can only perceive a change in source location (which is termed 'relative localization'). To be useful, this strategy requires the ability for a rather accurate localization based on the available cues since the estimate of the location of the signal source relies on a single brief measurement. The sound localization of barn owls (*Tyto alba*) responding to a brief transient signal such as a short rustling noise created by a mouse moving through leaf litter is an example of such open-loop localization. Upon hearing the stimulus, the owl will initiate a saccadic head movement that continues past the end of the stimulus and positions the head so that it faces the direction from which the sound emanated. The dynamic movement of the head during the saccade follows a stereotyped pattern. When the endpoint of the movement is reached after the stimulus has ended, no further corrective movements are made (if the stimulus continues, another saccade may be initiated). The advantage of the open-loop localization strategy is that it allows a fast response since no integration of information takes place over an extended time period.

The second strategy, called a closed-loop strategy, acquires information about the location of a source by making sequential comparisons. Applying this strategy, the animal integrates the information on the location of the signal source during its own movements. The ability to lateralize (i.e., to perceive whether the sound source is left or right of the body axis) is sufficient for acoustic closed-loop localization. If, for example, an animal first perceives a sound from the right and then turns rightward until the sound is being perceived from the left, it will have crossed the line pointing to the sound source. Many animals appear to exploit such a strategy to localize a signaling mating partner during a phonotactic approach. The green tree frog female, for example, follows a zigzag path when approaching a calling male. If she points to the left of the male, she will turn and jump to the right and vice versa. The advantage of this strategy is that she can reliably find the calling male in a chorus based only on acoustic cues, although her localization accuracy at best is only approximately 20 °. Closed-loop strategies with an integration of information over an extended time period are especially suitable to locate a signal source if the relevant cues vary considerably over time and a single estimate of a source location from one isolated measurement is unreliable. A typical example of coping with such a variable cue is the localization of odor sources. The odor eddies that are generated by turbulences in the medium transporting the odor stimulus lead to a highly variable stimulation of the receptors, making each single estimate of the source direction unreliable.

See also: Auditory Localization; Communication Networks and Eavesdropping in Animals; Endocrinology of Animal Communication: Behavioral; Multimodal Signaling in Animals; Pheromones and other Chemical Communication in Animals; Signal Identification: Peripheral and Central Mechanisms; Signal Transmission in Natural Environments; Signal Design Rules in Animal Communication; Sound Localization: Neural Mechanisms; Visual Signaling in Animals.

Further Reading

Atema J (1996) Eddy chemotaxis and odor landscapes. Exploration of nature with animal sensors. *Biology Bulletin* 191: 129–138.

Babineau D, Lewis JE, and Longtin A (2007) Spatial acuity and prey detection in weakly electric fish. *PLoS Computational Biology* 3(3): e38 (doi:10.1371/journal.pcbi.0030038).

Blauert J (1997) *Spatial Hearing: The Psychophysics of Human Sound Localization*, rev. edn. Cambridge, MA: MIT Press.

Bradbury JW and Vehrenkamp SL (1998) *Principles of Animal Communication*. Sunderland, MA: Sinauer.

Carlile S, Martin R, and McAnally K (2005) Spectral information in sound localization. *International Review of Neurobiology* 70: 399–434.

Fernald RD (2000) Evolution of eyes. *Current Opinion in Neurobiology* 10: 444–450.

Grasso FW and Basil JA (2002) How lobsters, crayfishes, and crabs locate sources of odor: Current perspectives and future directions. *Current Opinion in Neurobiology* 12: 721–727.

Klump GM (2000) Sound localization in birds. In: Dooling RJ, Popper AN, and Fay RR (eds.) *Comparative Hearing: Birds and Reptiles*, pp. 247–305. New York: Springer.

Klump GM, Benedix JH, Gerhardt HC, and Narins PM (2004) AM representation in green treefrog auditory nerve fibers: Neuroethological implications for pattern recognition and sound localization. *Journal of Comparative Physiology* 190: 1011–1021.

Koehl MAR (2006) The fluid mechanics of arthropod sniffing in turbulent odor plumes. *Chemical Senses* 31: 93–105.

Konishi M (2003) Coding of auditory space. *Annual Review of Neuroscience* 26: 31–55.

Kral K (2003) Behavioral–analytical studies of the role of head movements in depth perception in insects, birds and mammals. *Behavioral Processes* 64: 1–12.

Kuhn GF (1977) Model for the interaural time differences in the azimuthal plane. *Journal of the Acoustical Society of America* 62: 157–167.

Popper AN and Fay RR (2005) *Sound Source Localization*. New York: Springer.

Schluger JH and Hopkins CD (1987) Electric fish approach stationary signal sources by following electric current lines. *Journal of Experimental Biology* 130: 359–367.

Vickers NJ (2006) Winging it: Moth flight behaviour and responses of olfactory neurons are shaped by pheromone plume dynamics. *Chemical Senses* 31: 155–166.

Wagner H (2004) A comparison of neural computations underlying stereo vision and sound localization. *Journal of Physiology – Paris* 98: 135–145.

Yost WA (1977) Lateralization of pulsed sinusoids based on interaural onset, ongoing and offset temporal differences. *Journal of the Acoustical Society of America* 79: 426–442.

Yost WA and Dye RH (1988) Discrimination of interaural differences of level as a function of frequency. *Journal of the Acoustical Society of America* 83: 1846–1851.

Long-Term Depression (LTD): Endocannabinoids and Cerebellar LTD

D M Lovinger, National Institutes of Health, Rockville, MD, USA

Published by Elsevier Ltd.

Endocannabinoid-Dependent Long-Term Depression Is a Form of Synaptic Plasticity

Glutamatergic synapses are capable of long-lasting changes in their ability to communicate. These enduring changes in efficacy are often referred to as long-term plasticity. Long-term depression (LTD) a long-lasting decrease in synaptic efficacy, occurs at glutamatergic synapses throughout the brain. Subtypes of LTD with different mechanisms have been identified. One recently discovered form of LTD involves endogenous fatty acid derivates that are agonists for cannabinoid receptors (called endocannabinoids). Endocannabinoid-dependent LTD (eLTD) is widespread in the forebrain. This form of plasticity has been observed at the following glutamatergic forebrain synapses:

- Glutamatergic synapses between deep-layer cortical projection neurons and striatal medium spiny projection neurons.
- Glutamatergic synapses on medium spiny projection neurons in the nucleus accumbens.
- Glutamatergic synapses between pairs of layer 5 thick-tufted neurons in the primary visual cortex.
- Glutamatergic synapses between layer 4 neurons and layer 2/3 neurons in the whisker barrel field sensory cortex.

A form of eLTD induced by the application of the addictive drug amphetamine has been observed at glutamatergic inputs to neurons in the lateral amygdala. The synapses that exhibit LTD in this case arise from extrinsic cortical inputs to the amygdala. In addition, eLTD has been described at γ-aminobutyric acid (GABA)ergic inhibitory synapses in other brain regions, including the basolateral amygdala and hippocampus. It is likely that this form of plasticity will be found elsewhere in the nervous system in the coming years. A form of LTD found within the cerebellum also has mechanistic similarities to eLTD (see the section titled 'Cerebellar long-term depression').

Endocannabinoids

Endocannabinoids are fatty acid analogs that are metabolites of membrane lipids. These compounds are formed inside of neurons as a result of lipid breakdown catalyzed by phospholipases. There are two major endocannabinoids in the body: arachidonylethanolamide (AEA, also known as anandamide) and 2-arachindonyl glycerol (2-AG). Increases in intracellular calcium play a prominent role in endocannabinoid formation because they stimulate the enzymes necessary for biosynthesis of the compounds.

The biological actions of endocannabinoids are produced via interactions with several target proteins, the most prominent of which are the cannabinoid receptors. Cannabinoid receptors were originally discovered as targets for Δ9-tetrahydrocannabinol (Δ9-THC), the psychoactive ingredient in marijuana and other drugs derived from the *Cannabis sativa* plant. These receptors couple to G-proteins and activate biochemical signals inside cells. In the brain, the CB1 type of cannabinoid receptor (CB1R) is the major endocannabinoid target. AEA and 2-AG act as agonists at the CB1R. Most CB1Rs in the brain are located on presynaptic terminals of GABAergic and glutamatergic synapses. The CB1Rs generally couple to $G_{i/o}$-type G-proteins. They usually inhibit neurotransmitter release.

Induction and Expression Mechanisms of eLTD

eLTD is initiated by the coactivation of presynaptic glutamatergic afferents and the postsynaptic neuron. Experiments designed to elicit eLTD usually involve electrical stimulation of afferent fibers and electrophysiological recording from inside the postsynaptic neuron in brain-slice preparations. However, eLTD can also be induced in dual recordings from a single presynaptic neuron that forms a glutamatergic synaptic connection with a single postsynaptic neuron. The patterns of activation necessary for eLTD induction differ at different synapses. Induction protocols range from one-to-one pairing of pre- and postsynaptic action potentials (called spike timing-dependent synaptic plasticity (STDP)) at low frequencies of activation, to high-frequency afferent activation paired with postsynaptic depolarization.

There are several general features of eLTD at different synapses and variants that are specific to a given synapse. The induction of eLTD involves a sequence of events that starts with presynaptic activation. Repetitive activation of excitatory glutamatergic synapses stimulates postsynaptic signals that are involved in the induction of most forms of eLTD. The sequence of events involved in eLTD induction is:

1. Presynaptic glutamate release.
2. Depolarization leading to the activation of voltage-gated calcium channels and a rise in

postsynaptic calcium concentration (these steps are not required for eLTD of inhibitory synapses in the basolateral amygdala and hippocampus).
3. Activation of metabotropic glutamate receptors (mGluRs), leading to phospholipase activation.
4. Formation and release of endocannabinoids from the postsynaptic neuron.
5. Activation of presynaptic CB1Rs.

The initial postsynaptic depolarization necessary for most forms of eLTD arises from the activation of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type glutamate receptors subsequent to the presynaptic release of glutamate. The rise in postsynaptic calcium can come from several sources. In most forms of eLTD a voltage-gated calcium channel is activated by depolarization, and this allows calcium influx to the postsynaptic neuron. In the case of eLTD in striatum, an L-type calcium channel plays this role. A T-type channel appears to be the predominant calcium channel involved in eLTD induction in the whisker barrel cortex, with L-type channels playing a lesser role at this synapse. Activation of mGluRs can also contribute to the rise in postsynaptic calcium. These receptors activate the enzyme phospholipase C, which catalyzes the breakdown of phosphatidylinositol into diacylglycerol (DAG) and inositol trisphosphate (IP3). The IP3 generated by this reaction acts on receptors that release calcium from intracellular stores, and this is another source of calcium that can contribute to endocannabinoid production and eLTD induction. The DAG formed by mGluR activation of phospholipase C can also contribute to endocannabinoid formation. The enzyme known as DAG lipase catalyzes the formation of 2-AG. Inhibitors of DAG lipase prevent eLTD at several synapses, suggesting that 2-AG is a retrograde signal in this form of plasticity. The key role of mGluRs in eLTD is underscored by the fact that the application of mGluR agonists can produce such plasticity at several synapses.

In contrast to several other forms of long-term synaptic plasticity, eLTD usually does not require activation of *N*-methyl-D-aspartate (NMDA)-type glutamate receptors. The exception is eLTD in the cortex (Figure 1).

Endocannabinoids formed in the postsynaptic neuron traverse the synapse in a retrograde direction (post-to-presynaptic) and activate CB1Rs. The retrograde signal appears to be restricted to presynaptic terminals impinging on the postsynaptic neurons that produce the endocannabinoids. This finding is consistent with the idea that endocannabinoids do not diffuse more than 10–20 μm from their site of release. The CB1Rs activated by endocannabinoids then initiate molecular events leading to a long-lasting decrease in the probability of neurotransmitter release. Synergistic effects of CB1Rs and other receptors may be required for eLTD induction at many synapses. For example, eLTD at synapses in visual cortex requires the activation of NMDA-type glutamate receptors, most likely presynaptic receptors that contain the NR2B subunit. Activation of a presynaptic NMDA receptor has also been implicated in eLTD at synapses in the whisker barrel cortex. Expression of eLTD becomes independent of CB1R activation within minutes after the induction paradigm (Figure 2). Thus, the expression mechanism is set into motion by CB1R activation, but eLTD persists well beyond the brief initial period of receptor activation. It is not yet clear which endocannabinoid acts as the retrograde signal, but studies in the hippocampus and sensory cortex suggest a role for 2-AG based on the fact that an inhibitor of the DAG lipase enzyme prevents eLTD induction. The molecular mechanisms underlying sustained eLTD expression are poorly understood at present.

The induction of eLTD can also be modulated by neurotransmitters other than those directly involved in induction or expression. For example, in dorsal striatum dopamine acts on D2-type receptors to increase the likelihood that LTD will be initiated. However, recent evidence indicates that the D2 receptor is not strictly necessary for eLTD induction or expression because the receptor can be bypassed, for example, by combined depolarization and activation of postsynaptic mGluRs. A similar role for dopamine is probably involved in amphetamine-induced eLTD in the amygdala.

As mentioned previously, the expression of eLTD has been observed at both excitatory glutamatergic and inhibitory GABAergic synapses in different brain regions. However, even in cases in which the depression is expressed as a decrease in GABAergic transmission, the initiation process usually involves glutamatergic transmission that activates mGluRs.

Cerebellar LTD

The LTD observed at parallel fiber (PF) inputs to cerebellar Purkinje cells (PCs) was the first form of LTD characterized in the brain. Theoretical models of cerebellar function predicted that LTD at this synapse might have a crucial role in motor learning and that the coactivation of PFs and climbing fibers (CFs) would activate this type of plasticity. Consistent with these predictions, it has been found that paired PF and CF activation at low frequencies does indeed produce a long-lasting decrease in the efficacy of synaptic transmission at the PF–PC synapse. This form of LTD has been shown to occur *in vivo* and has also been found in brain slices and in single Purkinje

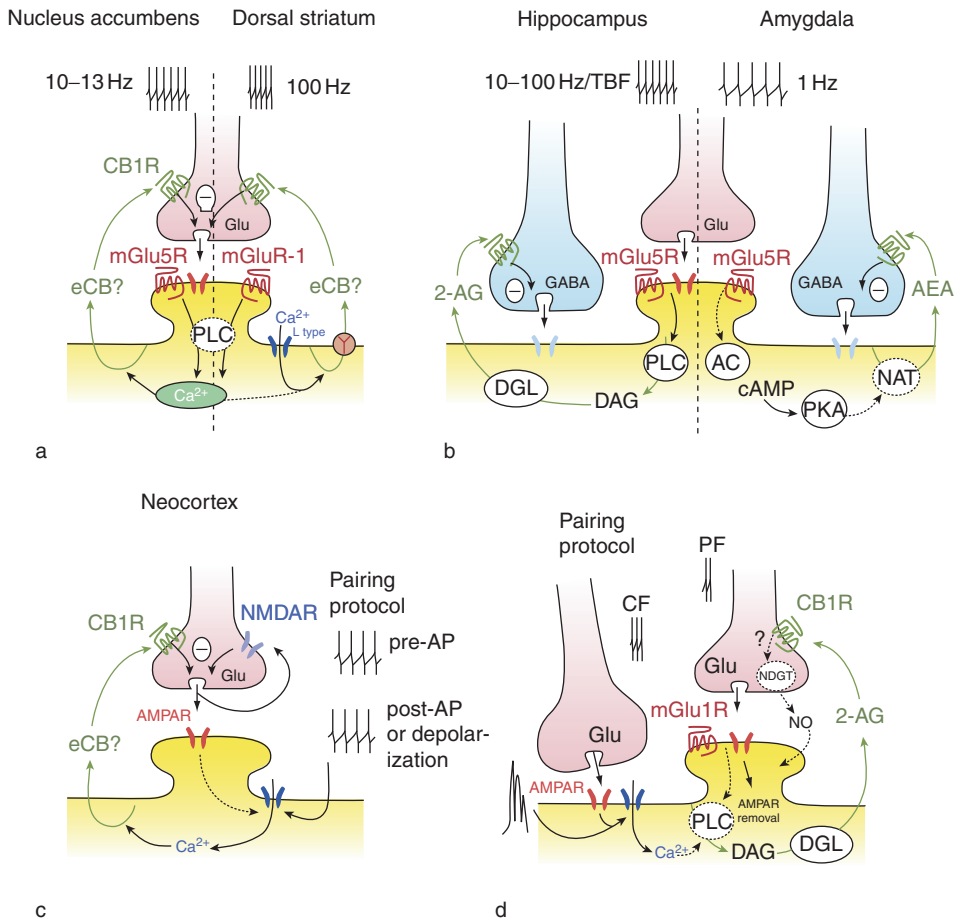

Figure 1 Induction mechanisms of four types of eLTD: (a) homosynaptic LTD; (b) heterosynaptic LTD; (c) spike timing LTD; (d) cerebellar LTD. The diagrams show the molecular mechanisms proposed to participate in initiation of eLTD at the striatal neocortical and cerebellar glutamatergic synapses (a, c), and amygdala and hippocampal GABAergic synapses (b). Note that eLTD expression is probably presynaptic in the amygdala, hippocampus, neocortex, and striatum but postsynaptic in the cerebellum. eLTD in the neocortex and striatum is homosynaptic, whereas in the amygdala and hippocampus it is heterosynaptic. Cerebellar LTD requires pairing of afferent inputs. 2-AG, 2-arachindonyl glycerol; AC, amygdala complex; AEA, arachindonylethanolamide; AMPAR, alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; AP, action potential; cAMP, cyclic adenosine-3′,5′-monophosphate; CB1R, CB1 receptor; CF, climbing fiber; DAG, diacylglycerol; DGL, lower layer of the dentate gyrus; eCB, endocannabinoid; eLTD, endocannabinoid-dependent long-term depression; GABA, γ-aminobutyric acid; Glu, glutamate; LTD, long-term depression; mGluR, metabotropic glutamate receptor; NAT, *N*-acetyltransferase; NMDAR, *N*-methyl-D-aspartate receptor; NOS, nitric oxide synthase; PF, parallel fiber; PKA, protein kinase A; PLC, phospholipase C; TBF, theta-burst frequency. Reproduced, with permission, from the *Annual Review of Neuroscience*, Volume 29, ©2006 by Annual Reviews. Chevaleyre V, Takahashi KA, and Castillo PE (2006) Endocannabinoid-mediated synaptic plasticity in the CNS. *Annual Review of Neuroscience* 29: 37–76.

neurons grown in cell culture. In the cell culture preparation, the activation of afferent inputs can be replaced by the application of receptor antagonists to induce LTD and examine mechanisms of expression.

A large array of molecules has been implicated in cerebellar LTD, including many that also participate in eLTD. There is consensus about the involvement of the following mechanisms:

- Presynaptic release of amino acid transmitters from CFs and PFs.
- Postsynaptic depolarization and activation of voltage-gated channels, including calcium channels.
- Activation of AMPA-type glutamate receptors and receptors containing the glutamate $\delta 2$ subunit.
- Activation of type I mGluRs.
- Increased postsynaptic intracellular calcium and sodium.
- Activation of G-proteins and phospholipases.
- Protein phosphorylation/dephoshorylation, involving the activation of enzymes such as protein kinase C (PKC), cyclic guanosine monophosphate (cGMP)-dependent protein kinase and protein phosphatases PP1 and PP2A. The AMPA receptor itself is an important target for phosphorylation.

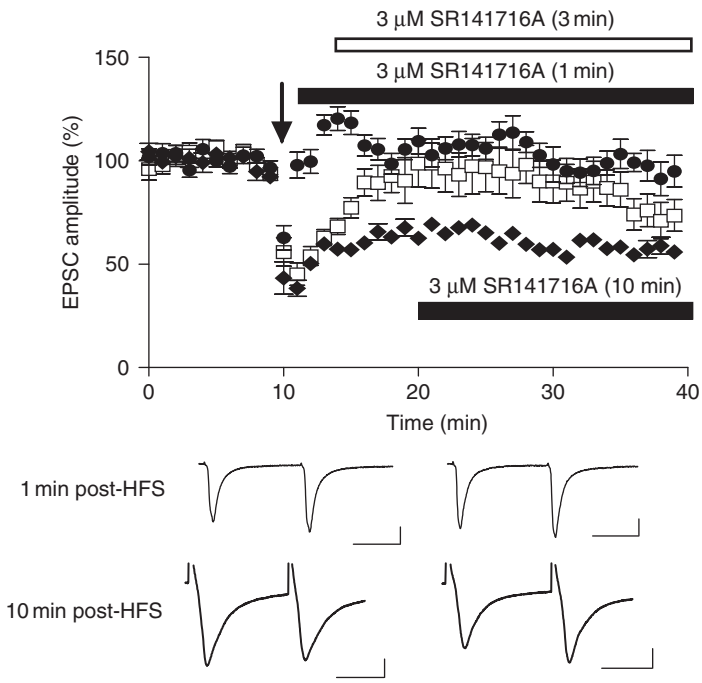

Figure 2 Time course of endocannabinoid release and CB1 activation during striatal LTD. Expression of LTD is prevented when CB1 antagonist is applied 1 min after high-frequency stimulation (HFS) but not when application begins 10 min after induction. Results with application at 3 min are intermediate. Calibration = 100 pA, 25 ms. EPSC, excitatory postsynaptic current; LTD, long-term depression. Reproduced from Ronesi J, Gerdeman GL, and Lovinger DM (2004) Disruption of endocannabinoid release and striatal long-term depression by postsynaptic blockade of endocannabinoid membrane transport. *Journal of Neuroscience* 24: 1673–1679, copyright 2004 by the Society for Neuroscience, with permission.

- Internalization of GluR2-containing AMPA receptors.
- Transcription and translation (in late-phase LTD).

Other molecular mechanisms have been implicated in the induction of cerebellar LTD, but they may play modulatory roles or may represent alternative pathways for activating the molecular mechanisms involved in LTD induction. For example, activation of nitric oxide synthase, leading to formation of the gaseous messenger nitric oxide (NO), has been implicated in cerebellar LTD induction. However, NO is not necessary for LTD in the cell culture preparation. This is most likely due to the fact that LTD in cell culture results from the strong activation of phosphorylation by the PKC pathway. In the slice preparation, NO released from stellate-type interneurons in response to glutamate activation of NMDA receptors appears to participate in LTD induction by the activation of guanylate cyclase, leading to the formation of cGMP. This intracellular second messenger, in turn, stimulates cGMP-dependent PKC to phosphorylate a substrate that inhibits the activity of the protein phosphatases PP1 and PP2A. Phosphatase inhibition presumably stabilizes GluR2 in the phosphorylated state and promotes AMPA receptor internalization.

Expression of cerebellar LTD almost undoubtedly involves postsynaptic mechanisms. Decreased AMPA receptor-mediated responses are observed that are tightly coupled to LTD expression. No reliable evidence of a presynaptic change associated with cerebellar LTD has yet been observed. The predominant cellular mechanism believed to be important for the expression of this form of LTD is the internalization of AMPA receptors at PF–PC synapses. Phosphorylation of the AMPA receptor itself, most likely via PKC-mediated transfer of phosphate to serine 880 on the GluR2 subunit, is a key step in this process. Phosphorylation at this residue enhances protein interacting with C kinase 1 (PICK1) binding to GluR2, and this protein–protein interaction results in the subsequent internalization of GluR2-containing AMPA receptors by a process of endocytosis involving the coating of the vesicles by the molecule clathrin. Reduction in the number of synaptic AMPA receptors leads to the reduction in the amplitude of the PF synaptic responses.

It has also been suggested that the activation of CB1Rs plays a role in cerebellar LTD. Postsynaptic mGluR activation and increases in intracellular calcium stimulate endocannabinoid production in Purkinje neurons and the subsequent activation of

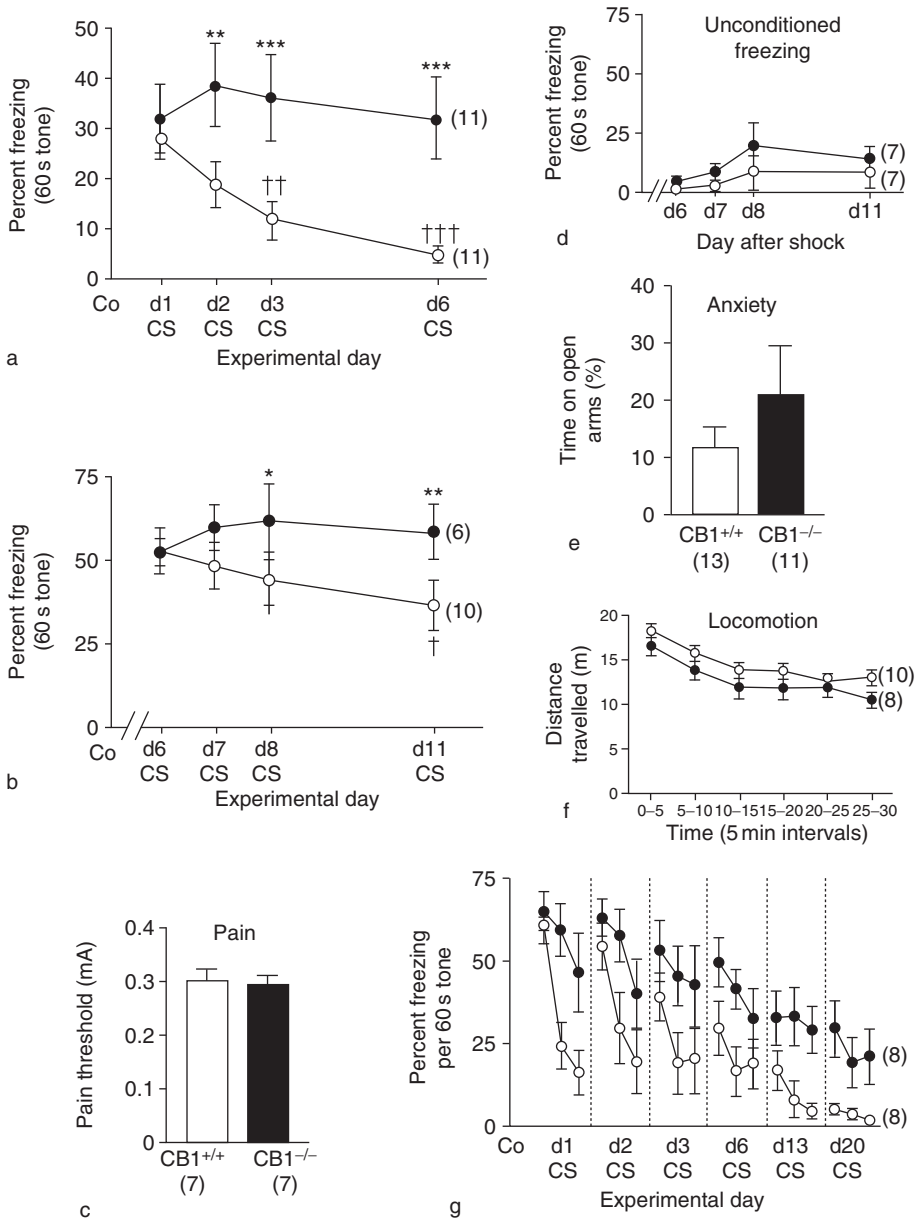

Figure 3 Extinction of aversive memories is impaired in the CB1 receptor knockout mouse: (a) after 1 day of conditioned freezing; (b) after 6 days of conditioned freezing; (c) pain thresholds; (d) unconditioned freezing; (e) open arm entries in the elevated plus maze; (f) horizontal open-field locomotion; (g) stronger conditioning regimen. Conditioned freezing extinguishes over several trials in $CB1^{+/+}$ but not in $CB1^{-/-}$ mice after 1 day (a) or 6 days (b) of conditioning. Pain thresholds, unconditioned freezing, open arm entries in the elevated plus maze, and horizontal open-field locomotion do not differ across the different mouse lines. In (g), some extinction is observed in $CB1^{-/-}$ mice with a stronger conditioning regimen, but extinction was still slower and less complete in comparison to $CB1^{+/+}$ mice. Black symbols indicate $CB1^{-/-}$ mice; white symbols indicate $CB1^{+/+}$ mice. CS, conditioned stimulus. Reprinted by permission from Macmillan Publishers Ltd: *Nature*, Marsicano G, Wotjak CT, Azad SC, et al. (2002) The endogenous cannabinoid system controls extinction of aversive memories. *Nature* 418(6897): 530–534, copyright 2002.

presynaptic CB1Rs. A recent study suggested that these molecular events take place during the induction of cerebellar LTD. The presynaptic CB1 activation leads to the release of another factor, possibly NO, that then acts on the postsynaptic neuron to induce LTD.

Other forms of synaptic plasticity occur in the cerebellum that can interact with PF LTD. The CF synapses are capable of expressing LTD when activated by sustained 5 Hz electrical stimulation in cerebellar slices. CF LTD may also involve a postsynaptic expression mechanism and may be involved in pruning back excessive innervation by this powerful synaptic input. Recent studies have characterized a form of long-term potentiation (LTP) that reverses PF LTD. This form of LTP is activated by sustained low-frequency PF input in

the absence of CF coactivation. It is expressed postsynaptically and involves phosphatase activation.

Functional Roles of eLTD

The roles played by eLTD in the function of neural circuits and in behavior are not yet fully understood. The suppression of inhibitory transmission produced by eLTD at GABAergic synapses produces a long-lasting increase in the potential for excitation of the postsynaptic neuron by glutamatergic synaptic inputs. In the hippocampus, this can lead to the priming of LTP at glutamatergic synapses to CA1 pyramidal neurons. The suppression of inhibition allows for greater excitation of the postsynaptic neuron in response to the activation of glutamatergic inputs. This increases the likelihood of NMDA receptor activation and thus enhances the probability of LTP induction.

Roles for eLTD in learning and memory have been postulated, as is the case for other forms of persistent synaptic plasticity. The only example to date in which eLTD has been linked to a specific form of learning is in the extinction of fear conditioning. The basolateral amygdala (BLA) is one site of information storage important for forming conditioned fear responses to previously neutral stimuli by pairing them with aversive stimuli. Endocannabinoids in the BLA appear to play a role in this extinction, and eLTD at GABAergic synapses within this brain region is a candidate mechanism for the decrease in efficacy of inhibitory transmission thought to underlie extinction (Figure 3).

The discovery that eLTD accounts for spike timing-dependent, long-lasting depression of excitatory transmission in the visual cortex suggests a role for this plasticity in cortical development and information storage. Plasticity at cortical glutamatergic synapses is thought to shape patterns of information flow through cortical circuitry during development by setting relative efficacies of thalamocortical and cortico-cortical synapses. These efficacies are shaped by sensory experience. For example, ocular dominance plasticity is a well-characterized change in cortical neuron responses to visual input. Visual experience, including deprivation, alters the strength of neuronal responses from the different eyes, and these experience-dependent changes may involve eLTD and other forms of synaptic plasticity. Within the rodent whisker barrel field sensory cortex, eLTD may also play roles in information storage related to tactile stimuli, as well as in synaptic development. A form of LTD that shares expression mechanisms with eLTD occurs during deprivation of sensory input from facial whiskers, although it is still to be determined if this LTD requires endocannabinoids or the CB1R.

Endocannabinoids and cannabinoid receptors are implicated in neural responses to drugs of abuse, including not only cannabis-based drugs but other addictive substances. Reducing CB1R activity with either antagonists or gene-targeted knockout reduces the self-administration of a variety of drugs of abuse. It has been suggested that eLTD might play a role in neuroadaptations to drugs and drug-related conditioning and that these processes might contribute to addiction. In the nucleus accumbens, eLTD is lost following chronic exposure to Δ9-THC. The loss of plasticity in this case is most likely a consequence of a drug-induced decrease in the number of CB1Rs (receptor downregulation). The loss of eLTD may contribute to behavioral tolerance to cannabinoid drugs. The eLTD induced by amphetamine in the lateral amygdala could play a role in drug-related conditioning, such as conditioned place preference.

Cerebellar LTD is thought to be part of an error-adaptation system involved in motor learning in which PF synapses activated during incorrect performance are depressed based on feedback from the CF inputs. This can come into play during fine-tuning of the vestibulo-ocular reflex. Cerebellar LTD may also play roles in Pavlovian eyeblink conditioning and locomotor adaptation.

See also: Cerebellar Deep Nuclei; Cerebellum and Oculomotor Control; Cerebellum: Models; Cognition: Cerebellum Role; Endocannabinoid Role in Synaptic Plasticity and Learning; Energy Homeostasis: Endocannabinoid System; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Hebbian Plasticity; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Metabotropic Glutamate Receptors (mGluRs): Functions; Procedural Learning: Cerebellum Models; Spike-Timing Dependent Plasticity (STDP); Spike-Timing-Dependent Plasticity Models.

Further Reading

Alger BE (2002) Retrograde signaling in the regulation of synaptic transmission: Focus on endocannabinoids. *Progress in Neurobiology* 68: 247–286.

Belmeguenai A and Hansel C (2005) A role for protein phosphatases 1, 2A, and 2B in cerebellar long-term potentiation. *Journal of Neuroscience* 25(46): 10768–10772.

Bender VA, Bender KJ, Brasier DJ, and Feldman DE (2006) Two coincidence detectors for spike timing-dependent plasticity in somatosensory cortex. *Journal of Neuroscience* 26: 4166–4177.

Chevaleyre V and Castillo PE (2003) Heterosynaptic LTD of hippocampal GABAergic synapses: A novel role of endocannabinoids in regulating excitability. *Neuron* 38(3): 461–472.

Chevaleyre V, Takahashi KA, and Castillo PE (2006) Endocannabinoid-mediated synaptic plasticity in the CNS. *Annual Review of Neuroscience* 29: 37–76.

DeZeeuw CI, Hansel C, Bian F, et al. (1998) Expression of a protein kinase C inhibitor in Purkinje cells blocks cerebellar LTD and adaptation of the vestibulo-ocular reflex. *Neuron* 20: 495–508.

Gerdeman GL, Ronesi J, and Lovinger DM (2002) Postsynaptic endocannabinoid release is necessary for long-term depression in the striatum. *Nature Neuroscience* 5(5): 446–451.

Gerdeman GL and Lovinger DM (2003) Emerging roles for endocannabinoids in long-term synaptic plasticity. *British Journal of Pharmacology* 140: 781–789.

Hansel C and Linden DJ (2000) Long-term depression of the cerebellar climbing fiber-Purkinje neuron synapse. *Neuron* 26: 473–482.

Hansel C, Linden DJ, and D'Angelo E (2001) Beyond parallel fiber LTD: The diversity of synaptic and non-synaptic plasticity in the cerebellum. *Nature Neuroscience* 4(5): 467–475.

Hoffman AF, Oz M, Caulder T, and Lupica CR (2003) Functional tolerance and blockade of long-term depression at synapses in the nucleus accumbens after chronic cannabinoid exposure. *Journal of Neuroscience* 23(12): 4815–4820.

Huang YC, Wang SJ, Chiou LC, and Gean PW (2003) Mediation of amphetamine-induced long-term depression of synaptic transmission by CB1 cannabinoid receptors in the rat amygdala. *Journal of Neuroscience* 23(32): 10311–10320.

Ito M (2001) Cerebellar long-term depression: Characterization, signal transduction, and functional roles. *Physiological Reviews* 81(3): 1143–1195.

Ito M and Kano M (1982) Long-lasting depression of parallel fiber-Purkinje cell transmission induced by conjunctive stimulation of parallel fibers and climbing fibers in the cerebellar cortex. *Neuroscience Letters* 33: 253–258.

Kreitzer AC and Malenka RC (2005) Dopamine modulation of state-dependent endocannabinoid release and long-term depression in the striatum. *Journal of Neuroscience* 25(45): 10537–10545.

Maldonado R, Valverde O, and Berrendero F (2006) Involvement of the endocannabinoid system in drug addiction. *Trends in Neurosciences* 29: 225–232.

Marsicano G, Wotjak CT, Azad SC, et al. (2002) The endogenous cannabinoid system controls extinction of aversive memories. *Nature* 418(6897): 530–534.

Robbe D, Kopf M, Remaury A, Bockaert J, and Manzoni OJ (2002) Endogenous cannabinoids mediate long-term synaptic depression in the nucleus accumbens. *Proceedings of the National Academy of Sciences of the United States of America* 99(12): 8384–8388.

Ronesi J, Gerdeman GL, and Lovinger DM (2004) Disruption of endocannabinoid release and striatal long-term depression by postsynaptic blockade of endocannabinoid membrane transport. *Journal of Neuroscience* 24: 1673–1679.

Safo PK and Regehr WG (2005) Endocannabinoids control the induction of cerebellar LTD. *Neuron* 48: 647–659.

Sjostrom PJ, Turrigiano GG, and Nelson SB (2003) Neocortical LTD via coincident activation of presynaptic NMDA and cannabinoid receptors. *Neuron* 39(4): 641–654.

Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms

C Lüscher, University of Geneva, Geneva, Switzerland
M Frerking, Oregon Health and Science University, Beaverton, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

How Is LTD Induced?

Although *N*-methyl-D-aspartate receptor (NMDAR-) and metabotropic glutamate receptor (mGluR-) long-term depression (LTD) can both be elicited *in vivo*, the mechanisms of LTD have been best studied in the acute brain-slice preparation, where it can be induced with a number of protocols. At most synapses, NMDAR-LTD is induced by a modest activation of NMDARs, whereas stronger activation leads to long-term potentiation (LTP). Exceptions to this rule may apply to the parallel fiber (PF)–Purkinje cell synapse in the cerebellum.

Activity-Dependent Induction

For NMDAR-LTD, prolonged low-frequency afferent stimulation (e.g., 900 stimuli at 1–5 Hz) has proven to be very efficient. Brief low-frequency afferent stimulation coupled to weak postsynaptic depolarization to –40 mV (to promote a modest relief of the Mg block of the NMDAR) can also efficiently induce NMDAR-LTD. In addition, NMDAR-LTD can be induced by the correlated activation of the pre- and postsynaptic neuron in many systems in which the firing of the postsynaptic neurons precedes the presynaptic action potential (AP) that releases the glutamate. The proposed mechanism to explain this spike timing-dependent LTD is that the AP in the postsynaptic cell back-propagates to the synapse, relieving the Mg block of the NMDAR. If the back-propagating AP precedes the release of glutamate, a low level of NMDAR activation is generated that leads to LTD.

To induce mGluR-LTD, a wide variety of stimulation protocols have been used. Typically mGluR-LTD requires a higher stimulation frequency than NMDAR-LTD (up to 300 Hz) or burst firing (e.g., several repetitions of five stimuli at 66 Hz), most likely due to the extrasynaptic location of mGluRs. In the hippocampus, a typical induction protocols consists of stimulation at 5 Hz for 3 min or the delivery of pairs of presynaptic APs in rapid succession.

Chemical Induction

Because activity-dependent LTD is generally specific to the small fraction of synapses that receive that activity, there has also been interest in chemical induction of LTD with receptor-selective agonists. This approach is particularly useful for studying the mechanisms underlying both NMDAR-LTD and mGluR-LTD because such treatment uniformly changes transmission in the majority of synapses.

NMDAR-LTD can be chemically induced by the exposure of a slice with NMDA for a few minutes. This chemically induced NMDAR-LTD occludes with NMDAR-LTD that is induced by synaptic stimulation, suggesting that the two overlap substantially. Similarly, mGluR-LTD can be induced by bath application of the drug dihydroxyphenylglycol (DHPG), a selective agonist of group I mGluRs. As with NMDAR-LTD, mGluR-LTD induced chemically occludes with mGluR-LTD induced by synaptic activity.

What Receptors Are Involved in the Induction of LTD?

An interesting idea, introduced about 6 years ago, is that the subunit composition of activated NMDARs determines whether LTD or LTP is elicited. Based on the use of subtype-selective NMDAR antagonists, it has been reported that the NR2A-containing NMDARs selectively elicit LTP but not LTD, whereas NR2B-containing receptors selectively elicit LTD but not LTP in both the hippocampus and cortex. However, results from genetic manipulations suggest that NR2B can contribute to LTP. Several studies have also failed to replicate the reported NMDAR-subtype dependence of LTP or LTD in the hippocampus. LTD is, similarly, not obviously subunit-dependent in the anterior cingulate cortex.

mGluR-LTD requires the activation of mGluRs, but there are several mGluR subtypes (mGluR1–8) and the particular mGluR leading to LTD appears to depend on the preparation in which it is induced. DHPG, a selective agonist of group I mGluRs (mGluR1 and mGluR5) elicits mGluR-LTD, which occludes with synaptically induced plasticity in many different systems, suggesting that in most cases mGluR-LTD is mediated by mGluR1, mGluR5, or both. The selective mGluR1 antagonist CPCOOEt blocks synaptically induced mGluR-LTD in the ventral tegemental area (VTA), and mGluR1 is also implicated in LTD in the cerebellum and neostriatum. mGluR5, on the other hand, has been shown to be the principal receptor responsible for mGluR-LTD in the hippocampus, cortex, and nucleus accumbens, although recent data in knockout mice indicate that the two

receptors may work in synergy. The activation of group II mGluRs (mGluR2 and mGluR3) has also been implicated in the CA3 and dentate gyrus of the hippocampus, as well as some regions of the cortex and the amygdala.

Intracellular Events Triggered by the Induction of LTD

The crucial event triggered by NMDAR activation during NMDAR-LTD is calcium influx (see Figure 1). The requirement for modest, but not strong, NMDAR activation suggests that a modest increase in calcium can lead to NMDAR-LTD, whereas strong increases give rise to NMDAR-LTP. This notion is supported by manipulations of intracellular calcium in which weak and prolonged calcium stimuli lead to LTD, whereas rapid and robust calcium stimuli lead to LTP. The exact determinants of the time course and spatial extent of the calcium signal still have to be worked out, and it remains unclear whether [Ca] levels alone are sufficient to determine whether LTP or LTD is induced. However, an attractive explanation for the bidirectional effects of calcium is that the calcium-dependent effectors that are required for NMDAR-LTD are very sensitive to calcium, whereas the effectors for NMDAR-LTP are less so.

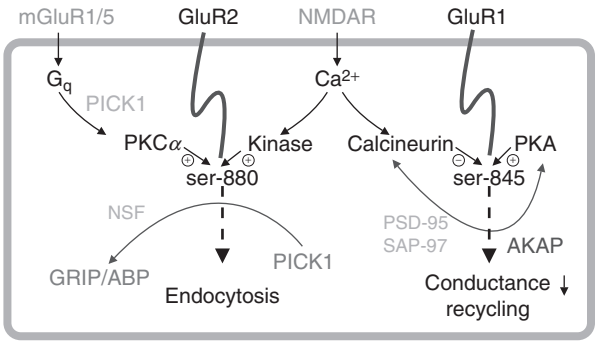

Figure 1 Schematic of intracellular signaling pathways involved in postsynaptic NMDAR-LTD and mGluR-LTD. Both forms of LTD involve the phosphorylation of ser-880 on the C-terminal tail of GluR2, which eventually increases endocytosis of mobile AMPARs. Mobility is conferred by exchanging GRIP/ABP for PICK1. NMDAR-LTD is, in addition, associated with a dephosphorylation of ser-845 on the C-terminal tail of GluR1, leading to decreased conductance and recycling. AKAP regulates the balance between constitutive PKA phosphorylation and activity-driven dephosphorylation of ser-845. AKAP, A-kinase associating protein; GRIP/ABP, glutamate receptor interacting protein/AMPA receptor binding protein; mGluR, metabotropic glutamate receptor; NMDAR, *N*-methyl-D-aspartate; receptor; NSF, *N*-ethylmaleimide sensitive factor; PICK1, protein interacting with C kinase 1; PKA, protein kinase A; PKC, protein kinase C; PSD, postsynaptic density protein; SAP, synapse-associated protein; ser, serine.

Consistent with this line of reasoning, NMDAR-LTD requires the calcium/calmodulin-dependent phosphatase calcineurin, which has a very high affinity for calcium-bound calmodulin (~100 pM). The protein phophatase 1 (PP1) is also required for at least some forms of NMDAR-LTD. NMDAR-LTD leads to a dephosphorylation of ser-845 of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor (AMPAR) subunit GluR1. Thus, NMDAR activation leads to the activation of PPs, which leads to AMPA receptor dephosphorylation. A caveat, however, is that there is much yet to be understood regarding how phosphatase activity is required for NMDAR-LTD; activity-dependent NMDAR-LTD requires both calcineurin and PP1, but chemically induced NMDAR-LTD does not appear to require PP1. Moreover, the phosphatase that actually directly targets GluR1 has not yet been identified.

mGluR-LTD depends on the activation of G-proteins of the pertussis-toxin-insensitive G_q and $G_{\alpha 11}$ family (see Figure 1). Group I mGluRs couple to phosphoinositide signaling via G_q. G_q seems to predominate in mGluR-LTD at most synapses, including in area CA1 of the hippocampus and at cerebellar PF synapses, where LTD was strongly reduced in $G_q^{-/-}$ mice, whereas in $G_{\alpha 11}^{-/-}$ mice the reduction was much smaller.

The events that happen downstream of G_q activation, however, are less clear. G_q is coupled to phospholipase C (PLC), which triggers the synthesis of inositol triphosphate (IP_3) and protein kinase C (PKC). This ultimately leads to release of calcium from intracellular stores. PKC is clearly involved in mGluR-LTD in the cerebellum and the VTA, but in the hippocampus mGluR-LTD seems not to require PKC or even PLC. The involvement of several other signaling pathways has been suggested, including p38 mitogen-activated protein (MAP) kinase, the extracellular signal-regulated kinase, tyrosine phosphatases, and phosphoinositide-3-kinase. Further study to define the intracellular cascades activated by mGluRs in the hippocampus will be of interest.

Expression Mechanisms Underlying NMDAR- and mGluR-Dependent LTD

There is a general agreement that at least some forms of NMDAR-LTD and mGluR-LTD are expressed as a postsynaptic reduction in AMPAR function, and evidence suggests that both forms of LTD act primarily via the clathrin-mediated endocytosis of AMPARs. However, there are several differences between these two types of LTD, and a number of significant questions remain to be addressed.

AMPAR Trafficking

The first evidence that NMDAR-LTD was due to a withdrawal of AMPARs from the cell surface came from a series of studies demonstrating that NMDAR activation could lead to the internalization of AMPARs in cultured hippocampal neurons and that NMDAR-LTD in hippocampal slices was blocked by disrupting interactions between key components of the endocytic machinery (the proteins amphiphysin and dynamin). LTD is abolished by blocking the interactions between AMPAR subunits and the AP-2 complex, which links to clathrin and is involved in the formation of endocytic-coated pits. In addition, LTD is impaired by viral transfections of the dominant-negative form of the small GTPase Rab5, which is thought to be involved in the trafficking of vesicles between the plasma membrane and endosomal compartments. Thus, NMDAR-LTD requires several components of endocytotic machinery at different stages of the endocytic pathway, and NMDAR activation leads to the withdrawal of AMPARs from the cell surface.

Some evidence also implicates receptor endocytosis in mGluR-LTD, although the issue has received less direct experimental attention. In the cerebellum, the postsynaptic infusion of peptides that interfere with the amphphysin–dynamin interaction blocks mGluR-dependent LTD. Activation of group I mGluRs has also been found to cause a reduction in AMPAR clustering at synapses in hippocampal cell cultures that is, again, blocked by interfering with the amphiphysin – dynamin interaction. Thus, mGluR-LTD appears to lead to loss of synaptic AMPARs via receptor internalization. However, not all data are readily compatible with this model. NMDAR-LTD leads to a reduction in postsynaptic responsiveness to exogenous agonists, as does mGluR-LTD in the cerebellum. Surprisingly, however, mGluR-LTD induced by DHPG application in the hippocampus does not. One possible explanation is that in the hippocampus, mGluR-LTD leads to a selective internalization or dispersion of synaptic AMPARs while leaving extrasynaptic AMPARs unaffected. In fact, with both types of LTD the time course and location of endo- and exocytosis remains to be fully identified. In hippocampal neuronal cultures, NMDAR activation induces a rapid lateral movement that precedes internalization at extrasynaptic sites.

Several studies also implicate a presynaptic reduction of transmitter release probability in mGluR-LTD, particularly in hippocampus. It remains unclear whether mGluR-LTD has both presynaptic and postsynaptic components that operate in tandem, or whether AMPAR loss preferentially affects synapses which have a high release probability.

AMPAR Subunits and LTD

Because AMPARs are heteromultimeric complexes that can be composed of any of four different subunits, GluR1–4, there has been considerable emphasis on the possible differential involvement of specific AMPAR subunits in the expression of LTD.

GluR2 The extreme C-terminus of GluR2 and GluR3 contains a postsynaptic density protein (PSD-) 95, Dlg, and ZO1 (PDZ) ligand sequence that is recognized by several proteins with PDZ domains (the glutamate receptor interacting protein (GRIP), the AMPA receptor binding protein (ABP), and the protein interacting with C kinase 1 (PICK1)). The binding of GluR2 to GRIP/ABP appears to be important in stabilizing AMPARs at the synapse and may also be involved in retaining AMPARs in intracellular pools. Blocking the interaction between GluR2 and PICK1 blocks the removal of GluR2 from the plasma membrane and recent data suggest a model in which the binding of PICK1 allows the mobility of AMPARs. The GluR2/PICK1 complex can be disassembled by ATPase activity of the *N*-ethylmaleimide sensitive factor (NSF), which binds to GluR2 (but not GluR3) at a site upstream of the PDZ ligand.

In addition to binding with GluR2, PICK1 also binds to the α isoform of PKC and to GRIP/ABP. This complex of PICK1, PKC, and GRIP/ABP recruits PKCα to synapses. GluR2 has a PKC phosphorylation site (serine 880 (ser-880)) that is near the PDZ ligand sequence, and the phosphorylation of GluR2 at ser-880 disrupts the interaction between GluR2 and GRIP/ABP but not the interaction between GluR2 and PICK1. Thus, a simple model is one in which LTD induction activates PKCα, leading to the PICK1-mediated targeting of PKCα to synapses. This would promote phophorylation of ser-880 on GluR2, causing GluR2 to release GRIP/ABP and bind to PICK1, promoting lateral diffusion and ultimately internalization.

An elegant series of studies suggest that just such a mechanism accounts for the mGluR-dependent LTD in the cerebellum. Cerebellar LTD requires PKC, and in particular PKCα. LTD induction leads to the phosphorylation of ser-880, and genetic manipulations that disrupt the PKC consensus site on GluR2, as do disruptions of the PDZ ligand site on PKCα that mediates its interaction with PICK1.

PDZ domain-mediated interactions among GluR2, PICK1, and PKCα also appear to be required for cerebellar LTD. Peptides that disrupt the PDZ interactions between GluR2 and PICK1 attenuate cerebellar LTD without affecting basal transmission. Cerebellar LTD is also abolished in knockout mice lacking either GluR2

or PICK1. More specific evidence of the necessity for GluR2–PICK1 interaction is provided by the observation that cerebellar LTD is also impaired by genetic manipulations that disrupt the PDZ ligand on GluR2 or the PDZ domain on PICK1.

Finally, interactions between NSF and GluR2 appear to be essential for the synaptic incorporation of AMPARs that are competent to undergo cerebellar LTD. Transfection of GluR3 into Purkinje cells from GluR2 knockout mice does not rescue cerebellar LTD, nor is the transfected GluR3 incorporated into synapses. However, a mutated version of GluR3 that contains the NSF binding site of GluR2 is incorporated into Purkinje cell synapses from GluR2 knockout mice, and these synaptically incorporated AMPARs are competent to undergo cerebellar LTD.

Do similar mechanisms account for mGluR-LTD observed at other synapses or for NMDAR-LTD? It is clear that at least some synapses express a form of mGluR-LTD that cannot occur via a loss of GluR2-containing AMPARs – at excitatory synapses on to dopamine cells in the VTA, for example, mGluR-LTD causes a shift in the biophysical properties of the AMPAR excitatory postsynaptic current (EPSC) that suggests a selective loss of AMPARs that lack the GluR2 subunit.

At first glance, NMDAR-LTD also seems unlikely to act via similar mechanisms because NMDAR-LTD does not require PKC activity. However, NMDAR-LTD can lead to PKC-independent phosphorylation of ser-880 through an as yet unidentified kinase, raising the possibility that a similar mechanism can be engaged by the same phosphorylation mediated by a different kinase. Moreover, peptides that interfere with the interaction between GluR2 and PDZ domain-containing proteins do impair LTD, although there is disagreement about whether GRIP/ABP or PICK1 is the essential mediator of these effects. Peptides that disrupt the interaction between GluR2 and NSF have also been observed to cause a rundown of synaptic transmission and a concomitant impairment of LTD.

However, a few notes of caution are in order. The NSF binding site on GluR2 overlaps substantially with a binding site for the AP-2 adaptor complex that links to clathrin and is involved in the formation of endocytic-coated pits. This AP-2 binding site is also present in GluR1 and GluR3, and the peptides used in initial studies of the NSF–GluR2 interaction also disrupt the interaction between AMPARs and AP-2. Experiments with more selective peptides suggest that AP-2 interactions with AMPARs are required for NMDAR-LTD, whereas NSF–GluR2 interactions are required for the rundown of basal transmission. Mutated versions of GluR2 in which the interactions with GRIP, ABP, and PICK1 are abolished can still undergo normal NMDA-induced internalization in hippocampal cultures, and a mutated version of GluR2 in which the phosphorylation of ser-880 is disrupted is still competent to undergo LTD in hippocampal slices, albeit less efficiently than wild-type GluR2.

More decisively, NMDAR-LTD is unaffected in GluR2 knockout mice, and it is actually enhanced in GluR2–GluR3 double-knockout mice. This makes it difficult to argue that either GluR2 or GluR3 is a critical component of NMDAR-LTD. One possible explanation is that the peptides used to selectively interfere with GluR2 function are not as specific as previously thought. This possibility is supported by the difficulties in distinguishing between NSF and AP-2 interactions with GluR2 and also by the observation that the peptides that interfere with GluR2–PDZ domain interactions can also impair PDZ domain interactions with distantly related kainate receptor subunits.

GluR1 An alternative subunit-selective mechanism for NMDAR-LTD is suggested by the observations that the ser-845 residue on GluR1 is dephosphorylated by exogenous NMDA in cell culture or by more conventional LTD induction in hippocampal slices. Ser-845 is constitutively phosphorylated by protein kinase A (PKA), and PKA activation has little effect on AMPAR function under basal conditions, presumably because of basal phosphorylation. However, the inhibition of PKA can depress AMPAR-mediated synaptic transmission and occlude LTD. PKA and calcineurin are colocalized by the A-kinase associating proteins (AKAPs), which are targeted to synapses by interactions with the scaffolding proteins PSD-95 and synapse-associated protein (SAP-)97. An accumulating body of evidence suggests indicates that AKAPs are dynamically regulated during LTD to shift the equilibrium between PKA and calcineurin localization at synapses. This determines the phosphorylation state of ser-845, and manipulations that interfere with PKA binding to AKAPs cause a depression of AMPAR-mediated transmission that occludes LTD. Further experiments to selectively examine the role of ser-845 in NMDAR-LTD will be of interest.

How might ser-845 dephosphorylation be involved in LTD? Phosphorylation of ser-845 can enhance AMPAR function by increasing the peak open probability of the channel, so calcineurin-mediated dephosphorylation of GluR1 could directly contribute to LTD expression even in the absence of changes in receptor trafficking. However, dephosphorylation of ser-845 appears to be downstream of receptor internalization and prevents the recycling of AMPARs after they have been internalized. Thus, current evidence suggests that the dephosphorylation of GluR1 is a

fairly late event in the mechanisms underlying LTD that may have more to do with preventing AMPAR reinsertion than with the withdrawal of AMPARs from the synapse.

Maintenance of LTD

The previous section focused on the mechanisms by which AMPAR function is depressed during LTD. We now turn our attention to the mechanisms that allow the depression to persist long after the induction event is over. The answers to this question are still largely unknown, although the available data are sufficient in some cases to draw some surprising conclusions that deserve mention.

In the case of NMDAR-LTD, some evidence is consistent with the idea that LTD is maintained by the change in phosphorylation of the AMPAR itself. As discussed previously, dephosphorylation at ser-845 on GluR1 prevents internalized AMPARs from recycling back to the synapse, suggesting that the phosphorylation state of this site may determine whether LTD persists; consistent with this idea, PKA activation can lead to a reversal of LTD following the stable induction of LTD, as can synaptic activity. Synaptically induced de-depression can be reduced by the inhibition of PKA, providing further evidence for dephosphorylation of PKA targets as a mechanism that is responsible for the maintenance of NMDAR-LTD.

However, it need not be the case that NMDAR-LTD is maintained by the duration of modifications to AMPARs. An alternative is that the number of AMPARs at the synapse is limited by proteins that tether the AMPAR at the postsynaptic density; a decrease in the amount of this tether would preclude AMPAR reinsertion at synapses following LTD induction even if the AMPARs were recycled to the cell surface because there would be no mechanism to hold them at the synapse. The protein PSD-95 has many properties that are consistent with such a tether. Manipulations that increase or decrease PSD-95 expression cause parallel changes in the synaptic AMPAR-mediated response. This may be relevant to the maintenance of NMDAR-LTD, because NMDA applied to hippocampal neurons in culture leads to the ubiquitination and subsequent proteosomal degradation of PSD-95. This loss of PSD-95 is dependent on calcineurin and blocked by PKA activation, similar to LTD; moreover, the inhibition of proteosome function prevents the NMDA-induced reduction in surface expression of GluR1 and GluR2. NMDA also actively recruits proteosomes to dendritic spines. Further characterization of this potential mechanism for LTD maintenance will be of interest.

In the case of mGluR-LTD, it has long been known that the LTD induced by DHPG can be reversed long after stabilization by the application of group I mGluR antagonists. This indicates, surprisingly, that the expression mechanisms underlying mGluR-LTD are maintained by constitutive activation of an mGluR following LTD induction; in the absence of mGluR activity, the depression is quickly reversed.

An alternative model that has some experimental support is that mGluR-LTD might be maintained by the synthesis of new proteins that stabilize LTD expression. Both LTD and the internalization of AMPAR subunits are blocked by inhibitors of mRNA translation. This potential link between translation of new proteins and mGluR-LTD became a source of considerable excitement when it was found that mice lacking the fragile X mental retardation protein (FMRP) have enhanced mGluR-LTD in both the hippocampus and cerebellum.

The loss of function of FMRP is responsible for the fragile X syndrome in humans, an X-linked form of mental retardation. FMRP is an RNA-binding protein, although it remains a topic of debate whether FMRP represses or stimulates translation. Regardless of the precise action of FMRP on translation, the observation that FMRP enhances mGluR-LTD seemed at first to clearly indicate a role for changes in protein synthesis as an essential part of mGluR-LTD. Surprisingly, however, subsequent analysis of FMRP knockout mice indicated that mGluR-LTD in these mice is not only enhanced but also independent of protein synthesis. Similarly, the inhibition of proteosome function has recently been found to block mGluR-LTD in wild-type mice but not in FMRP knockout mice. These results are difficult to reconcile with an integral role for protein synthesis or degradation in mGluR-LTD, and further study will be of interest.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Depression: Cerebellum; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Metabotropic Glutamate Receptors (mGluRs): Functions; Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology; NMDA Receptors and Development; NMDA Receptors and Disease + C464; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Depression.

Further Reading

Anwyl R (2006) Induction and expression mechanisms of postsynaptic NMDA receptor-independent homosynaptic long-term depression. *Progress in Neurobiology* 78: 17–37.

Bear MF (2003) Bidirectional synaptic plasticity: From theory to reality. *Philosophical Transactions of the Royal Society of London. B. Biological Sciences* 358: 649–655.

Bear MF, Huber KM, and Warren ST (2004) The mGluR theory of fragile X mental retardation. *Trends in Neuroscience* 27: 370–377.

Bellone C and Lüscher C (2005) mGluRs induce a long-term depression in the ventral tegmental area that involves a switch of the subunit composition of AMPA receptors. *European Journal of Neuroscience* 21: 1280–1288.

Carroll RC, Beattie EC, vonZastrow M, and Malenka RC (2001) Role of AMPA receptor endocytosis in synaptic plasticity. *Nature Reviews Neuroscience* 2: 315–324.

Cull-Candy S, Kelly L, and Farrant M (2006) Regulation of Ca^{2+}-permeable AMPA receptors: Synaptic plasticity and beyond. *Current Opinion in Neurobiology* 16: 288–297.

Dan Y and Poo MM (2006) Spike timing-dependent plasticity: From synapse to perception. *Physiological Reviews* 86: 1033–1048.

Hanley JG (2006) Molecular mechanisms for regulation of AMPAR trafficking by PICK1. *Biochemical Society Transactions* 34: 931–935.

Jorntell H and Hansel C (2006) Synaptic memories upside down: Bidirectional plasticity at cerebellar parallel fiber-Purkinje cell synapses. *Neuron* 52(2): 227–238.

Lüscher C, Nicoll RA, Malenka RC, and Muller D (2000) Synaptic plasticity and dynamic modulation of the postsynaptic membrane. *Nature Neuroscience* 3: 545–550.

Malenka RC and Bear MF (2004) LTP and LTD: An embarrassment of riches. *Neuron* 44: 5–21.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* 25: 103–126.

Meng Y, Zhang Y, and Jia Z (2003) Synaptic transmission and plasticity in the absence of AMPA glutamate receptor GluR2 and GluR3. *Neuron* 39(1): 163–176.

Oliet SH, Malenka RC, and Nicoll RA (1997) Two distinct forms of long-term depression coexist in CA1 hippocampal pyramidal cells. *Neuron* 18: 969–982.

Segal M (2005) Dendritic spines and long-term plasticity. *Nature Reviews Neuroscience* 6(4): 277–284.

Long-Term Depression: Cerebellum

M Ito, RIKEN Brain Science Institute, Wako, Japan

© 2009 Elsevier Ltd. All rights reserved.

Long-term depression (LTD) is a type of synaptic plasticity in which the efficacy of signal transmission across a synapse continuously decreases after a certain triggering activity. LTD in the cerebellum was proposed as a theoretical possibility around 1970 and was discovered a decade later. Thus far, several subtypes of LTD, varying in cellular and molecular mechanisms, have been found in the cerebellum, hippocampus, neocortex, striatum, and amygdala in vertebrate brain and also in molluscan ganglia. LTD may occur in synapses that have received a triggering activity (homosynaptic LTD) or in those that have not received a triggering activity (heterosynaptic LTD). Some temporal phases of LTD have been distinguished (short, long, and late phases). LTD occurs not only in excitatory synapses, but also in inhibitory synapses. LTD may weaken or functionally interrupt useless or erroneous synaptic connections between neurons, providing a mechanism synergistic with long-term potentiation (LTP) in various forms of learning and memory.

Induction and Observation of LTD

In Cerebellum and in a Cerebellum-Like Structure

In the cerebellar cortex, each Purkinje cell forms two distinct types of excitatory synapses, one from parallel fibers and the other from a climbing fiber. LTD occurs when these two types of synapses are activated repeatedly in approximate synchrony, leading to a continuous decrease in the synaptic strength of parallel fibers. In *in vitro* cerebellar slices, LTD is typically induced by the conjunctive stimulation of parallel and climbing fibers at 1 Hz for 5 min (300 stimuli), which decreases transmission efficacy by 30%. A repetitive stimulation of parallel fibers followed by the stimulation of climbing fibers (a few times) is also effective in inducing LTD. Such LTD can be observed by electrical recording from individual Purkinje cells, and also by optical imaging of the surface of the cerebellum using voltage-sensitive dyes. LTD is robust in the sense that its induction does not critically depend on the timing between climbing fiber stimulation and parallel fiber stimulation. In inducing LTD, climbing fiber stimulation is often replaced by current-induced membrane depolarization that enhances the influx of Ca^{2+} into Purkinje cell dendrites, as in the case of climbing fiber stimulation.

LTD induced by the conjunctive stimulation of parallel fibers and a climbing fiber is represented by a decrease in parallel-fiber-evoked excitatory postsynaptic potential (EPSP) or current (EPSC), which usually fully develops gradually over 30–60 min. In contrast, LTD induced by parallel fiber stimulation in combination with depolarizing pulses develops faster (in 5–10 min). The difference between these two types of LTD indicates the fast phase of LTD that follows parallel fiber stimulation–depolarization conjunction, but not parallel fiber–climbing fiber conjunction. LTD can be observed to continue for 1–3 h without any signs of recovery. Under the influences of protein synthesis inhibitors, previously developed LTD declines over 1 h, revealing the protein-synthesis-dependent slow phase of LTD.

Strong stimulation of parallel fibers causes homosynaptic LTD in Purkinje cells. This is because a sufficiently large influx of Ca^{2+} occurs in strongly activated Purkinje cell dendrites. Conjunctively induced LTD is accompanied by LTD in neighboring synapses that are located within a distance of 100 μm and are not involved in conjunctive stimulation. Nevertheless, such LTD occurs in only neighboring synapses that are active even at a very low rate of discharge (~0.2 Hz), indicating a weak temporal relationship with climbing fiber activity. Thus, the input specificity of LTD still holds in a wider sense. An input-non-specific LTD occurs in inhibitory synapses that Purkinje cells supply to cerebellar nuclear neurons. In this case, repetitive stimulation of a portion of inhibitory synapses leads to LTD in this portion as well as in other, unstimulated portions.

The brains of certain fish species contain a cerebellum-like structure that lacks climbing fibers. LTD occurs in Purkinje cell-like neurons in parallel-fiber-evoked EPSPs when EPSPs repeatedly precede postsynaptic spikes by 60 ms, at the most.

In Hippocampus and Neocortex

In pyramidal cells of the hippocampal CA1 region, a low-frequency stimulation (LFS) of a bundle of presynaptic fibers (1 Hz for 5–15 min) typically induces LTD, whereas a high-frequency stimulation (five stimuli at 100 Hz repeated at 200 ms intervals for 2 s) induces LTP. LFS-induced LTD also occurs in layer III neurons of the visual cortex following the stimulation of the white matter or layer IV. LTD has also been observed in the sensorimotor and prefrontal cortex.

An associative stimulation of presynaptic fibers with postsynaptic membrane depolarization (5 Hz for 16 s) also induces LTD or LTP, depending on the time from pre- to postsynaptic activities. Repetitive postsynaptic spiking within a time window of 20 ms after presynaptic activation results in LTP, whereas

postsynaptic spiking within a window of 20 ms before a repetitive presynaptic activation leads to LTD. A significant LTP occurs only at synapses with a low initial strength, whereas the extent of LTD does not show obvious dependence on the initial synaptic strength. LTD occurs in stimulated (homosynaptic LTD) and unstimulated (heterosynaptic LTD) synapses, whereas LTP occurs only homosynaptically. As revealed in a triple chain of cultured hippocampal neurons, the induction of associative LTD in a synapse is accompanied by a back-propagated induction of LTD in a synapse of a presynaptic cell. Through this induction, temporal information coded in the timing of individual spikes may be converted into spatially distributed patterns of continuous synaptic modifications in a neural network. In fact, it has been shown that the receptive fields of superior colliculus neurons are modified through a temporary exposure to patterns of visual stimuli.

Associative LTP or LTD is also induced in the hippocampus and neocortex, when presynaptic stimulation is paired with membrane depolarization. A strong membrane depolarization induces LTP, whereas a weak membrane depolarization causes LTD. This dichotomy depends on the amount of Ca^{2+} influx evoked by the depolarization.

In Amygdala, Striatum, and Molluscan Ganglia

LTD occurs in the basolateral amygdala, following a LFS (1 Hz/900 pulses) of the mouse lateral amygdala of an *in vitro* slice preparation. In slices from 2-week-old mice, interneurons express pronounced LTD, in contrast to those in slices from adult mice (2–4 months old). Principal neurons show only a weak LTD. In the basolateral nucleus of the amygdala, γ-aminobutyric acid (GABA)-ergic interneurons and serotonergic afferents appear to play a pivotal role in the induction of LTD.

The corticostriatal glutamatergic synapses in the striatum exhibit LTD in a high percentage of medium spiny neurons. D2 receptors mediate this LTD induction. However, the medium spiny neurons form two groups. Only one group, striatopallidal neurons, expresses D2 receptors, whereas the other group, striatonigral neurons, expresses D1 receptors. A recent study suggests that the D2 receptor-mediated dopaminergic control of LTD induction is not directly exerted on the corticostriatal synapses, but that it is mediated by cholinergic interneurons that express D2 receptors; D2 receptor-mediated LTD in these interneurons leads to reduction of M1 receptor tone at corticostriatal synapses.

Invertebrate central synapses also possess the capacity for LTD. For example, activation of sensory neurons at 2 Hz for 15 min induces LTD of isolated *Aplysia* sensorimotor synapses in cell culture.

Signal Transduction Underlying LTD

Conjunctively induced LTD is a purely postsynaptic event, whereas other LTD subtypes include presynaptic factors. Postsynaptic signal transduction processes have been analyzed in synaptically induced LTD and in various reduced forms of LTD induced with a chemical or by an electrical stimulation of neurons in place of synaptic activation. The analytical methods have been improved using various pharmacologically specific antagonists and genetic manipulation. More than 30 molecules have been identified in the signal flow map underlying conjunctively induced LTD (Figure 1).

Receptors

LTD in excitatory synapses is, at least in part, caused by a reduction in the number of functional α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor molecules in the postsynaptic membrane. Other types of receptors play roles in eventually inducing this change. The activation of *N*-methyl-D-aspartate (NMDA) receptors is required for LTD induction in excitatory synapses but not in inhibitory synapses of hippocampal neurons. Even though NMDA receptors are not functional in adult Purkinje cells, the activation of NMDA receptors present in other cells, probably in stellate cells, is required for the induction of conjunctive LTD in Purkinje cells. Parallel fiber synapses on Purkinje cells contain, besides AMPA receptors, type 1 metabotropic glutamate receptors (mGluR1s) and δ2 glutamate receptors, whereas climbing fiber synapses contain AMPA receptors, type 1 corticotropin-releasing factor receptors (CRFR1s), and type 1 insulin-like growth factor receptors (IGF1Rs). The δ2 receptor is an orphan receptor with a still unknown function. The mGluR1s, CRFR1s, and IGF1Rs are associated with G-proteins which are coupled with certain second messenger processes. Because the inhibition of any of these receptors results in the blockage of LTD, a concerted activation of all these receptors leads to the conjunctive induction of LTD (Figure 1).

Calcium Entry and Release from Intracellular Stores

The conjunctive induction of LTD requires an increase in $[Ca^{2+}]_i$, because LTD is blocked by the injection of a Ca^{2+} chelator (ethylene glycol tetraacetic acid; EGTA) into Purkinje cells. Climbing fiber impulses evoke Ca^{2+} influx into dendrites of Purkinje cells through voltage-sensitive Ca^{2+} channels. As in underlying homosynaptic LTD, Ca^{2+} influx also occurs in association with parallel-fiber-induced EPSPs, when the EPSPs are sufficiently large to

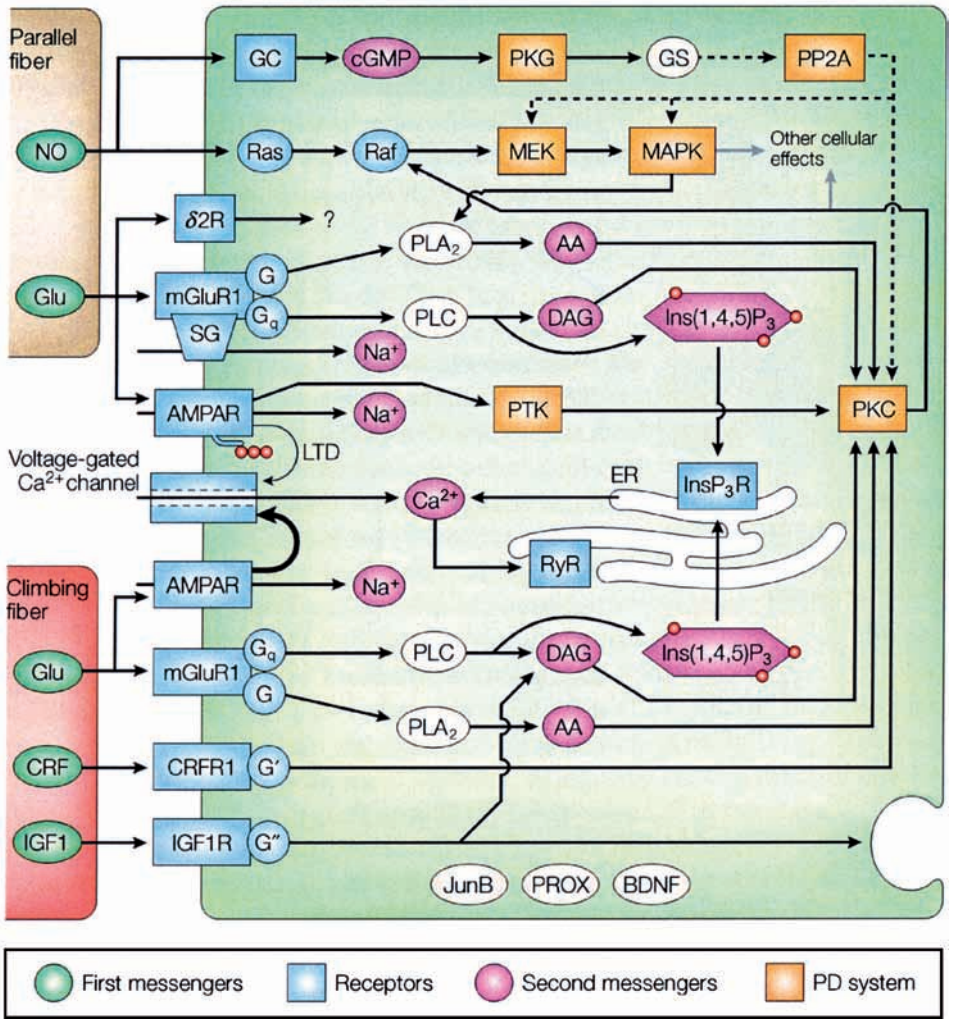

Figure 1 Chart of signal transduction pathways following stimulation of a parallel fiber and a climbing fiber. GC, guanylyl cyclase; MEK, mitogen-activated protein kinase (MAPK) kinase; AA, arachidonic acid; $InsP_3R$, $Ins(1,4,5)P_3$ receptor; SG, slow EPSP generator associated with mGluR1; RyR, ryanodine receptor; *JunB*, an immediate-early gene expressed during LTD; PROX, a rapidly turned over protein that plays a role in the initial phase of LTD induction; BDNF, brain-derived neurotrophic factor; PLA_2 phospholipase A_2; PKC, protein kinase C; DAG, diacylglycerol; ER, endoplasmic reticulum; PLC, phospholipase C; cGMP, cyclic guanosine monophosphate; PKG, a cGMP-dependent protein kinase; PTK, protein tyrosine kinase; PP2A, protein phosphatase type 2A; AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; mGluR1, type 1 metabotropic glutamate receptor; CRFR1, type 1 corticotropin-releasing factor receptor; IGF1R, type 1 insulin-like growth factor receptor. Colors: green, first messengers; blue, receptors, channels, and receptive enzyme; yellow, protein kinases; brown, protein phosphatase. Solid arrows indicate activating effects; dashed arrows indicate inactivating effects. Reproduced from Ito M (2002) The molecular organization of cerebellar long-term depression. *Nature Reviews Neuroscience* 3: 896–902, with permission.

activate voltage-sensitive Ca^{2+} channels. Ca^{2+} is also released from intracellular stores in the endoplasmic reticulum, when inositol trisphosphate (IP_3; also $Ins(1,4,5)P_3$) or ryanodine receptors on the reticulum are activated. The inhibition of IP_3 receptors, the depletion of Ca^{2+} stores, or the genetically induced loss of the endoplasmic reticulum in dendritic spines results in the blockade of conjunctively induced LTD. A double-shock stimulation of parallel fibers at 50–60 ms intervals induces an optimal activation of Ca^{2+} release from the endoplasmic reticulum, and therefore induces LTD effectively.

In hippocampal and neocortical neurons, LTD induction depends on membrane potential (see earlier), which determines the influx of Ca^{2+} ions into neurons through cation channels associated with NMDA receptors. These channels are normally blocked by Mg^{2+} that is removed at a depolarized membrane potential. Modest and strong activations of NMDA receptors lead to LTD and LTP, respectively. Inhibition of postsynaptic Ca^{2+} influx by a partial blockade of NMDA receptors shifts LTP to LTD. LTP is also shifted to LTD by injecting EGTA into neocortical neurons. Induction of homosynaptic and heterosynaptic

LTDs in the hippocampus requires functional ryanodine and IP_3 receptors, respectively. A functional blockade or the genetic deletion of type 1 IP_3 receptors leads to a shift of LTD to LTP and the inhibition of heterosynaptic LTD, whereas the blockade of ryanodine receptors inhibits only homosynaptic LTD. In mouse amygdala slices, Ca^{2+} influx via NMDA receptors is essential in triggering LTD; loading single neurons with the calcium chelator 1,2-bis(*o*-aminophenoxy)ethane-*N*,*N*,*N*′,*N*′-tetraacetic acid (BAPTA) (30 mM) does not decrease the EPSC following LFS, and the induction of LTD is suppressed by a competitive NMDA receptor antagonist. Calmodulin kinase II (CamKII) plays a role in amygdala LTD because the application of a specific CamKII antagonist, KN-62, blocks LFS-induced LTD.

Metabotropic Glutamate Receptor-Driven Processes

The activation of mGluR1s in Purkinje cells results in the activation of phospholipase C (PLC), which generates IP_3 (see earlier) and diacylglycerol (DAG) from membrane phospholipids. DAG activates protein kinase C (PKC). The mGluR1s can also be coupled to phospholipase A_2 (PLA_2), which produces arachidonic acid and oleic acid from membrane phospholipids. The inhibition of mGluR1, PKC, PLA_2, or IP_3 receptors prevents LTD induction, indicating that these molecules are involved in LTD induction. The application of an mGluR5 agonist, dihydroxy propoxy methyl guanine (DHPG), effectively induces a continuous decrease in EPSPs in hippocampal CA1 neurons, which has been studied as a reduced form of hippocampal LTD. In amygdala slices, LTD is blocked by a metabotropic glutamate receptor antagonist and by 5-hydroxytryptamine (5-HT) via the activation of 5-HT(1A) receptors. Moreover, blocking $GABA_A$ receptor-mediated synaptic transmission with bicuculline or picrotoxin inhibits LTD induction. GABAergic interneurons and serotonergic afferents may play roles in amygdala LTD induction.

Nitric Oxide and cGMP

Cerebellar granule cells contain a neuronal-type nitric oxide synthase (nNOS), and parallel fiber terminals release NO upon stimulation. NO then diffuses into Purkinje cells and activates guanylate cyclase to increase the level of cyclic guanosine monophosphate (cGMP). cGMP in turn activates a cGMP-dependent protein kinase (PKG). The role of NO in the conjunctive induction of LTD is evident because such induction is blocked when an inhibitor of NO synthase, N^{ω}-monomethyl-L-arginine (L-NMMA), or hemoglobin, which absorbs NO, is applied to cerebellar slices. Furthermore, sodium nitroprusside, which releases NO, or a membrane-soluble derivative of cGMP induces an LTD-like phenomenon when applied to cerebellar slices together with AMPA.

Protein Kinases and Phosphatases

The conjunctive induction of cerebellar LTD requires the activities of PKC, PKG, protein tyrosine kinase (PTK), and mitogen-activated protein kinase (MAPK), but not that of PKA. PKC phosphorylates serine 880 of AMPA receptors, and PTK interacts with PKC. PKG acts on a PKG-specific substrate, G-substance, which is abundant in Purkinje cells. The phosphorylated G-substrate becomes a potent inhibitor of protein phosphatases, presumably type 2A (PP2A). Specific inhibitors of PP2A block LTD induction.

PKC and PTK, but not PKA, are involved in mGluR-induced LTD in neurons of the hippocampal dentate gyrus. A Ca^{2+}/calmodulin-dependent protein phosphatase, calcineurin (or PP2B), mediates LTD induction in the hippocampus and visual cortex because a calcineurin inhibitor, FK506, blocks LFS-induced LTD. Mice deficient in the calcineurin catalytic unit B1 exhibit significantly suppressed LFS-induced LTD in the hippocampus. This is in contrast to the finding in Purkinje cells in which the inhibition of PP2A causes LTD. Calcineurin is also involved in LTD induction, downregulating $GABA_A$ receptors in inhibitory synapses of the hippocampus.

Protein Synthesis

Protein synthesis plays roles in hippocampal neuronal processes because a translational inhibitor blocks the late phase of LFS-induced LTD as well as the entire course of DHPG-induced LTD. Translational inhibitors suppress the late phase of LTD in cultured Purkinje cells, but they entirely block LTD in cerebellar slices. Studies using 5 min pulses of translational inhibitors revealed that the synthesis of a protein with a rapid turnover is required for LTD induction only during the 15 min period from the onset of conjunctive stimulation. A 5 min pulse application of transcriptional inhibitors also effectively blocks LTD induction with a delay of 30 min.

Mutant mice lacking the fragile X mental retardation type 1 *FMR* gene show abnormally slender dendritic spines and exhibit an abnormally enhanced LTD in the visual cortex and cerebellum. *FMR1* encodes an RNA-binding protein that acts as a negative translational regulator. Studies of *FMR1* demonstrated the importance of local protein synthesis within a neuron as a determinant of normal synaptogenesis and development of cognitive abilities.

Phosphorylation and Internalization of AMPA Receptors

Of the subunits constituting AMPA receptors, GluR2 subunits are predominant in parallel fiber–Purkinje cell synapses. Evidence indicates that at the late phase of the conjunctive induction of cerebellar LTD, GluR2 is phosphorylated at its serine 880 anchorage to glutamate receptor-interacting protein (GRIP), and free AMPA receptors are removed from the postsynaptic membrane by internalization via endocytosis. A similar mechanism is considered applicable to associative LTD in hippocampal neurons.

Cannabinoid

It has recently been revealed that transmitter release at parallel fiber–Purkinje cell synapses is regulated through the activation of presynaptic cannabinoid receptors by endogenous cannabinoids that are released as retrograde messengers from Purkinje cells. This retrograde regulation occurs in response to the activation of postsynaptic mGluR1s. It accounts for the short-term depression of parallel fiber–Purkinje cell transmission that follows a brief tetanic stimulation of parallel fibers.

Molecular Organization of LTD

Of the more than 30 molecules involved in LTD induction, some may serve as mediators that are indispensable for LTD induction, whereas others may serve as modulators of the functions of the mediators. The lack of a mediator should always lead to the blockade of LTD, but that of a modulator would not necessarily block LTD induction. For example, the NO–cGMP pathway is required for LTD induction in cerebellar slices, but not for the induction of the reduced form of LTD in isolated Purkinje cells. Because climbing fiber activation can be replaced by current-evoked depolarization in LTD induction, AMPA receptors activated together with the activation of voltage-gated Ca^{2+} channels should normally be a mediator, whereas mGluR1s, CRFR1s, and IGF1Rs at climbing fiber synapses should be modulators. This view is also supported by the observation that LTD can be induced in tissue-cultured Purkinje cells lacking climbing fibers, using depolarization pulses. Thus far, parallel fiber–AMPA receptors coupled with voltage-gated Ca^{2+} influx, and parallel fiber–mGluR1 receptors coupled with intracellular Ca^{2+} release, are thought to constitute duplicated mediators; either one can mediate LTD.

The signal flow map shown in Figure 1 should also include a coincidence-detector mechanism by which multiple signals that coincide within a short time window induce a supralinear effect, that is, LTD. The first candidate for such a coincidence detector is IP_3 receptors on the endoplasmic reticulum, activated via mGluR1s in parallel fiber synapses. The activation of climbing fiber synapses evokes the influx of Ca^{2+}, which sensitizes IP_3 receptors. The coincidence of these two events leads to increases 8.2 times larger than the linear sum of the increases induced by a separate stimulation of parallel fibers and a climbing fiber.

The second candidate for the coincidence detector is voltage-gated Ca^{2+} channels in dendritic spines. These channels are activated superlinearly by parallel-fiber-evoked EPSPs and climbing-fiber-evoked EPSPs; the increase in $[Ca^{2+}]_i$ so induced is 4.5 times larger than the linear sum of the increases induced by individual stimulation.

The high $[Ca^{2+}]_i$ that results from coincidence detection in either or both of the two mediator pathways leads to the next stage of signal transduction on the way to LTD. This is the phosphorylation and internalization of AMPA receptors, and this is carried out by the phosphorylation–dephosphorylation system (see preceding sections, 'Phosphorylation and internalization of AMPA receptors' and 'Cannabinoid'). The possible operation of the entire multiple molecular signal transduction underlying LTD has been reproduced by kinetic simulation.

Functional Roles of LTD

In the cerebellum-like structure in fish, LTD plays a distinct role in attenuating the effects of sensory inputs coinciding with command signals to allow only unpredicted sensory inputs to predominate, thus forming an adaptive sensory processor. The involvement of LTD in certain forms of cerebellar learning has been demonstrated using pharmacological or genetic manipulations that impair LTD. Applying an NO scavenger or transfecting a pseudoinhibitor of PKC to the cerebellum abolishes the adaptive change in the vestibuloocular reflex (VOR). An NO scavenger or an NOS inhibitor also blocks the adaptive change that occurs in the smooth pursuit eye movement during repeated sudden increases in the velocity of a moving spotlight. Mice deficient in the glial fibrillary acidic protein or phospholipase C β4 lack LTD and do not exhibit the classical conditioning of eye blinking. A walking cat or mouse can normally adapt to a sudden change in running belt conditions, but cannot under the influence of an NO scavenger or an NOS inhibitor or when it has mGluR1 deficiency, which blocks LTD. Mice in which hippocampal LTD is suppressed because of a

conditional knockout of calcineurin are specifically impaired in working memory and episodic-like memory tasks, including a delayed matching-to-place task and a radial maze task.

LTD as Memory Trace

LTD can be observed continuously for 1–3 h, after which long-term depressed synapses may disappear, because, after repeated VOR adaptations, a decrease in the number of synapses occurs in parallel fiber–Purkinje cell synapses, or the long-term depressed synapses may be maintained as silent synapses, structurally unchanged but functionally silent. This unexpected possibility has recently been indicated by three observations. First, up to 80% of parallel fiber–Purkinje cell synapses are silent or have a very low transmission efficacy. Second, in a sagittal slice of the rat cerebellum, parallel fiber stimulation excites an area of 380 μm^2 covering 1140–2660 parallel fiber–Purkinje cell synapses, but evokes EPSCs 5- to 30-fold larger than those evoked by a unitary parallel fiber synaptic current (12 pA). Hence, the electric stimulation of parallel fibers in a small dendritic area of a Purkinje cell activates at most 2.6% (30/1140) of the parallel fiber–Purkinje cell synapses located within the area. Third, in the forelimb area of the C3 zone, each Purkinje cell responds, via parallel fibers, to stimulation in only one of the 30–40 types of receptive fields that cover the forelimb skin in its entirety. The receptive fields of Purkinje cells vary even between neighboring Purkinje cells located along the same parallel fiber bundle. Thus, effective parallel fiber inputs to a Purkinje cell can be only about 3% of the total number of cutaneously activated parallel fibers forming synaptic contacts with the cell.

Surprisingly, an electric burst stimulation of parallel fibers (100 Hz burst of 15 pulses every 3 s for 5 min), unpaired with climbing fiber activity, induces a very large, long-lasting increase in the receptive field size of Purkinje cells over 2–6 h. On the other hand, the stimulation of parallel fibers, paired with climbing fiber activity, reverses the changes in the size of the receptive field. On the basis of these findings, Ekerot and Jörntell suggest that silent synapses are produced functionally by LTD, and that learning converts silent synapses to active synapses, or vice versa, by LTP or LTD induction. The finding on silent synapses opens a new aspect of the cerebellar microcomplex mechanism. The relationships of silent synapses with synaptic plasticity and their significance in information processing should be investigated further.

Another intriguing issue concerning LTD as a memory process is whether the cerebellar cortical network permanently stores memory via LTD, or whether memory is transferred to cerebellar or vestibular nuclei to be consolidated. Relevant to this issue, two phases, fast and slow, of VOR adaptation have recently been distinguished. LTD in floccular Purkinje cells accounts for the memory of fast-phase VOR or optokinetic response (OKR) adaptation, which develops during continuous rotation for about 1 h. This was demonstrated by blocking such adaptation by injecting a localized anesthetic, lidocaine, into the flocculus. In contrast, the slow-phase VOR or OKR adaptation, established by repeated trials of screen rotation for 1 week, continues after the flocculus is injected with an AMPA/kainate receptor blocker, 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) or lidocaine. Nevertheless, the slow-phase OKR adaptation is suppressed by an NOS inhibitor that blocks LTD, or by a lesion in the inferior olive. Therefore, LTD in the flocculus is required for the development of the fast and slow phases of OKR adaptation, but not for the maintenance of the previously formed, slow-phase adaptation. Underlying the slow-phase OKR adaptation, an increase in vestibular nerve-induced activity occurs in vestibular relay neurons. This indicates that non-input-specific plasticity in vestibular nuclei accounts for the slow-phase OKR adaptation. In the eye blink conditioning, information on the magnitude of the eye blink reflex has been assumed to be stored in the anterior interpositus nucleus, whereas information on the timing of the eye blink reflex is stored in the cerebellar cortex. In this case as well, plasticity changes occur in cellular excitability that lacks input specificity. It is important to define the capacity of memory storage in nuclei differently from that in the cerebellar cortex, in light of the elaborateness of the neuronal network structures, and, accordingly, of the capabilities of processing and retaining spatiotemporal-patterned information.

Other issues to be addressed are whether information acquired through synaptic plasticity is eventually stored at the level of modified molecules, such as altered genes, and whether new proteins produced by these genes are carriers of memory traces. The involvement of the immediate-early genes *JunB* and c-*Fos* in LTD induction and the involvement of CamKIV and cAMP response element-binding (CREB) protein in the late phase of LTD have been suggested, but these experiments lasted no more than 2 h.

Summary

Investigations during the past 25 years have established LTD as a major type of synaptic plasticity that plays roles in various forms of learning and memory. Complex molecular mechanisms underlying

LTD induction have been revealed. It will be a future task to confirm whether LTD is converted to a permanent memory, and, if so, by what mechanism. We should also determine the specific functional roles of LTD in individual forms of learning and memory.

See also: Calcium Channels; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Glutamate Receptor Organization: Ultrastructural Insights; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Depression (LTD): Endocannabinoids and Cerebellar LTD; Long-Term Potentiation (LTP); Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Depression; Synaptic Mechanisms of Learning.

Further Reading

Bell C, Bodznick D, Montgomery J, et al. (1997) The generation and subtraction of sensory expectations within cerebellum-like structures. *Brain, Behavior and Evolution* 50(supplement 1): 17–31.

Bi G-Q and Poo M-M (1999) Distributed synaptic modification in neural networks induced by patterned stimulation. *Nature* 401: 792–796.

Hansel C and Linden DJ (2000) Long-term depression of the cerebellar climbing fiber–Purkinje neuron synapse. *Neuron* 26: 473–482.

Huber KM, Kayser MS, and Bear MF (2000) Role for rapid dendritic protein synthesis in hippocampal mGluR-dependent long-term depression. *Science* 288: 1254–1256.

Ito M (2001) Long-term depression – characterization, signal transduction and functional roles. *Physiological Reviews* 81: 1143–1195.

Ito M (2002) The molecular organization of cerebellar long-term depression. *Nature Reviews Neuroscience* 3: 896–902.

Ito M (2006) Cerebellar circuitry as a neuronal machine. *Progress in Neurobiology* 78: 272–303.

Kelleher RJ III, Govindarajan A, and Tonegawa S (2004) Translational regulatory mechanisms in persistent forms of synaptic plasticity. *Neuron* 44: 59–73.

Kuroda S, Schweighofer N, and Kawato M (2001) Exploration of signal transduction pathways in cerebellar long-term depression by kinetic simulation. *Journal of Neuroscience* 21: 5693–5702.

Luscher C, Nicoll RA, Malenka RC, et al. (2000) Synaptic plasticity and dynamic modulation of the postsynaptic membrane. *Nature Neuroscience* 3: 545–550.

Malenka RC and Bear MF (2004) LTP and LTD: An embarrassment of riches. *Neuron* 44: 5–21.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* 25: 103–126.

Sanes JR and Lichtman JW (1999) Can molecules explain long-term potentiation? *Nature Neuroscience* 2: 597–604.

Wang SSH, Denk W, and Hausser M (2000) Coincidence detection in single dendritic spines mediated by calcium release. *Nature Neuroscience* 3: 1266–1273.

Wang Z, Kai L, Day M, et al. (2006) Dopaminergic control of corticostriatal long-term synaptic depression in medium spiny neurons is mediated by cholinergic interneurons. *Neuron* 50: 443–452.

Long-Term Potentiation (LTP)

J D Sweatt, University of Alabama at Birmingham, Birmingham, AL, USA

© 2009 Elsevier Ltd. All rights reserved.

Synaptic Plasticity and Memory

The particular circuits and neuronal connections that underlie most forms of mammalian learning and memory are mysterious at present. There is little understanding of the means by which complex memories are stored and recalled at the neural circuit level – this is a very important avenue of modern research.

This article focuses on long-term potentiation (LTP), a phenomenon that is currently the best candidate for a cellular mechanism contributing to learning and memory in the mammalian central nervous system (CNS). LTP has been extensively studied and is the best-understood form of synaptic plasticity at the molecular level. Moreover, it is a specific candidate cellular mechanism for mediating certain forms of associative learning, spatial learning, and adaptive change in the CNS, in particular in the amygdala, hippocampus, and cerebral cortex, respectively.

Learning and memory manifest themselves as a change in an animal's behavior, and scientists capitalize on this to study these phenomena by observing and measuring changes in an animal's behavior in the wild or in experimental situations. However, all the behavior exhibited by an animal is a result of activity in the animal's nervous system. The nervous system comprises many kinds of cells, but the primary functional units of the nervous system are neurons. Because neurons are cells, all of an animal's behavioral repertoire is a manifestation of an underlying cellular phenomenon. By extension, changes in an animal's behavior, such as those that occur with learning, must also be subserved by an underlying cellular change.

By and large, the vast majority of the communication between neurons in the nervous system occurs at synapses. Because synapses mediate the neuron–neuron communication that underlies an animal's behavior, changes in behavior are ultimately subserved by alterations in the nature, strength, or number of interneuronal synaptic contacts in the animal's nervous system. The capacity for alterations of synaptic connections between neurons is referred to as synaptic plasticity, and one of the great unifying theories to emerge from neuroscience research in the last century was that synaptic plasticity subserves learning and memory. LTP (of some sort, at least) is the specific form of synaptic plasticity that is the leading candidate for the mechanism subserving behavior-modifying changes in synaptic strength that mediate higher-order learning and memory in mammals.

One of the pioneers who advanced the idea that changes in neuronal connectivity are a mechanism for memory was the Canadian psychologist Donald Hebb, who published his seminal formulation as what is now generally known as Hebb's postulate:

> When an axon of cell A . . . excites cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells so that A's efficiency as one of the cells firing B is increased. (Hebb, 1949; p. 62)

A Breakthrough Discovery – LTP in the Hippocampus

LTP, a cellular phenomenon with the attributes predicted by Hebb's postulate, has been most widely studied in the hippocampus, an area of the CNS widely documented as being important for memory information. As a young postdoctoral researcher, Tim Bliss set out to find a long-lasting form of synaptic plasticity in the hippocampus, and after teaming up with Terje Lomo in Per Anderson's laboratory in Oslo, he did just that. The seminal 1973 report by Bliss and Lomo describing a phenomenon they termed "long-term potentiation" of synaptic transmission set the stage for what is now over three decades of progress in understanding the basics of long-term synaptic alteration in the CNS.

In their experiments, Bliss and Lomo recorded synaptic responses in the dentate gyrus, stimulating the perforant path inputs from the entorhinal cortex. They used extracellular stimulating and recording electrodes implanted in the animal, and the basic experiment was begun by recording the baseline synaptic transmission in this pathway. They discovered that a brief period of high-frequency (100 Hz tetanic) stimulation led to a robust increase in the strength of synaptic connections between the perforant path inputs from the entorhinal cortex to the dentate granule neurons in the dentate gyrus (Figure 1). They also observed an increased likelihood of the cells firing action potentials in response to a constant synaptic input, a phenomenon they termed E-S (excitatory postsynaptic potential (EPSP)-to-spike) potentiation.

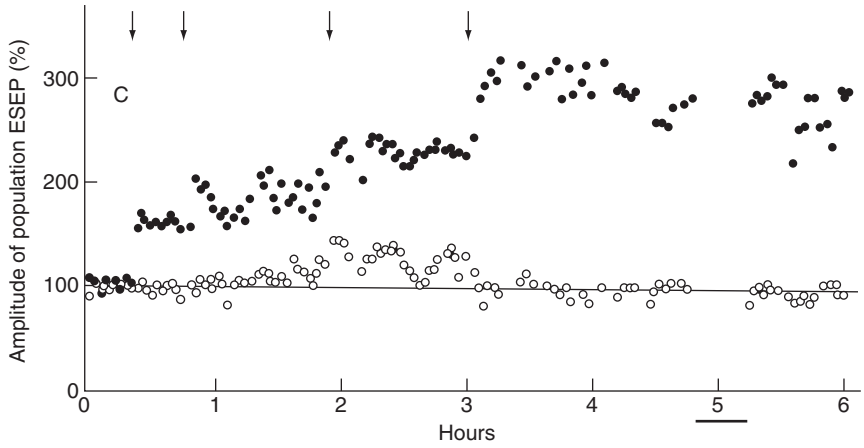

Figure 1 Bliss and Lomo's first published long-term potentiation (LTP) experiment, in which Tim Bliss and Terje Lomo demonstrated the LTP of synaptic transmission at perforant path inputs into the dentate gyrus (see Figure 4). Arrows indicate the delivery of high-frequency synaptic stimulation, resulting in LTP. Filled circles are responses from the tetanized pathways; open circles are a control pathway that did not receive tetanic stimulation. The bar, where no data points are available, indicates a period of time when Bliss fell asleep. Data acquisition in this era involved the investigator directly measuring by hand synaptic responses from an oscilloscope screen. Moreover, it was not unusual for experiments to extend overnight due to the long time involved in preparing the rabbit for the experiment, implanting the electrodes into the brain, and establishing a stable recording configuration. EPSP, excitatory postsynaptic potential. Reproduced from Bliss TV and Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* 232: 331–356, with permission from Elseiver.

These two phenomena together were termed LTP. LTP lasted many, many hours in this intact rabbit preparation. The appeal of LTP as an analog of memory was immediately apparent – it is a long-lasting change in neuronal function that is produced by a brief period of unique stimulus, exactly the sort of mechanism that had long been postulated to be involved in memory formation. This pioneering work of Bliss and Lomo set in motion a several-decades-long pursuit by numerous investigators geared toward understanding the attributes and mechanisms of LTP.

The Hippocampal Circuit and Measuring Synaptic Transmission in the Hippocampal Slice

Bliss and Lomo did their experiment using the intact rabbit, stimulating and recording in the anesthetized animal using implanted electrodes. In recent times, this preparation has been largely supplanted by the use of recordings from hippocampal slices maintained *in vitro* (see Figures 2 and 3). Most experiments on the basic attributes and mechanisms of LTP have been studies of the synaptic connections between axons from area CA3 pyramidal neurons that extend into area CA1. These are the synapses onto CA1 pyramidal neurons that are known as the Schaffer collateral inputs.

In a popular variation of the basic LTP experiment, extracellular field potential recordings in the dendritic regions of area CA1 are used to monitor synaptic transmission at Schaffer collateral synapses (see Figures 3 and 4). A bipolar stimulating electrode is placed in the stratum radiatum subfield of area CA1, and electrical stimuli are delivered. Stimuli delivered in this fashion stimulate the output axons of CA3 neurons that pass nearby, causing action potentials to propagate down these axons. Cellular responses to this stimulation are recorded using extracellular or intracellular electrophysiological recording techniques.

The typical waveform in an extracellular recording consists of a fiber volley, which is an indication of the presynaptic action potential arriving at the recording site, and the EPSP itself. The EPSP responses are a manifestation of synaptic activation (depolarization) in the CA1 pyramidal neurons. For measuring field (i.e., extracellularly recorded EPSPs), the parameter typically measured is the initial slope of the EPSP waveform (see Figure 3). Extracellular field recordings measure responses from a population of neurons, so EPSPs recorded in this fashion are referred to as population EPSPs (pEPSPs).

As a prelude to starting an LTP experiment input–output (I/O) functions for stimulus intensity versus EPSP magnitude are recorded in response to increasing intensities of stimulation (see Figure 4). Baseline synaptic transmission at this constant test-stimulus intensity is usually monitored for a period of 15–20 min to assure a stable response.

Once the health of the hippocampal slice is confirmed, as indicated by a stable baseline synaptic response, LTP can be induced using any one of a wide variety of different LTP induction protocols. Many

popular variations include a single or repeated period of 1 s 100 Hz stimulation (with the delivery of the 100 Hz trains separated by 20 s or more) in which stimulus intensity is at a level necessary for approximately half-maximal stimulation (see Figure 4). This protocol gives robust LTP that lasts for essentially as long as the researcher can keep the hippocampal slice alive.

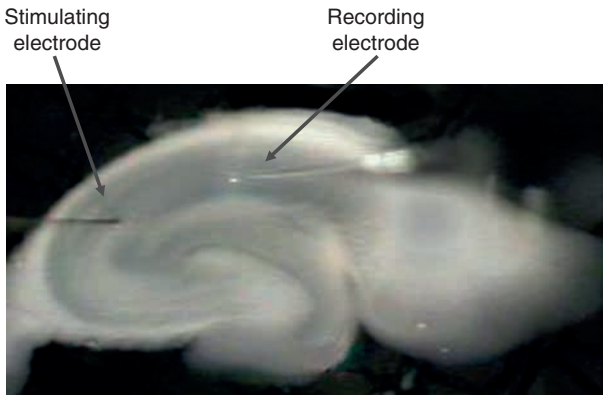

Figure 2 Electrodes in a living mouse hippocampal slice maintained in a recording chamber. Responses in area CA1 are recorded using a saline-filled glass micropipette electrode and a bipolar platinum stimulating electrode (see Figure 4). Reproduced from Sweatt D (2003) *Mechanisms of Memory*. London: Elsevier, with permission from Elsevier.

N-Methyl-D-Aspartate Receptor-Dependence of LTP

In 1983, Graham Collingridge made the breakthrough discovery that induction of these tetanus-induced forms of LTP are blocked by a blockade of a specific subtype of glutamate receptor, the *N*-methyl-D-aspartate (NMDA) receptor. Collingridge's fascinating discovery was that the glutamate analog amino-phosphono-valeric acid (APV), an agent that selectively blocks the NMDA subtype of glutamate receptor, can block LTP induction while leaving baseline synaptic transmission entirely intact.

This was the first experiment to provide a specific molecular insight into the mechanisms of LTP induction. The properties of the NMDA receptor that allow it to function in this unique role as a trigger of LTP are important (see later in this article). For our purposes right now, suffice it to say that pharmacological blockers of NMDA receptor function have allowed the definition of different types of LTP that can be selectively induced with various physiological stimulation protocols. For example, subsequent work has shown that an NMDA receptor-independent type of LTP can be induced in area CA1 and elsewhere in the hippocampus (mossy fibers to be precise), as well as other parts of the CNS.

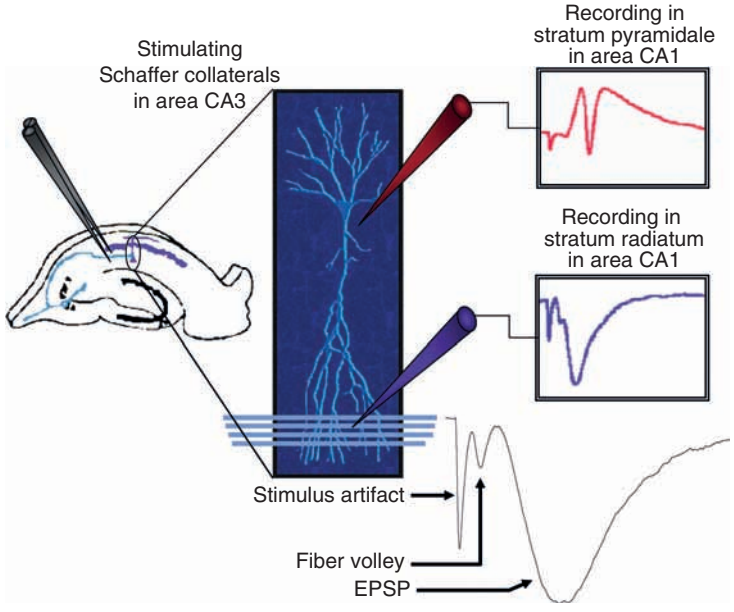

Figure 3 Recording configuration and typical physiological responses in a hippocampal slice recording experiment. Electrode placements and responses from the stratum pyramidale (cell body layer) and stratum oriens (dendritic regions) are shown. In addition, the typical waveform of a population EPSP is illustrated, showing the stimulus artifact, fiber volley, and population EPSP. EPSP, excitatory postsynaptic potential. Figure and data by Joel Selcher. Reproduced from Sweatt D (2003) *Mechanisms of Memory*. London: Elsevier, with permission from Elsevier.

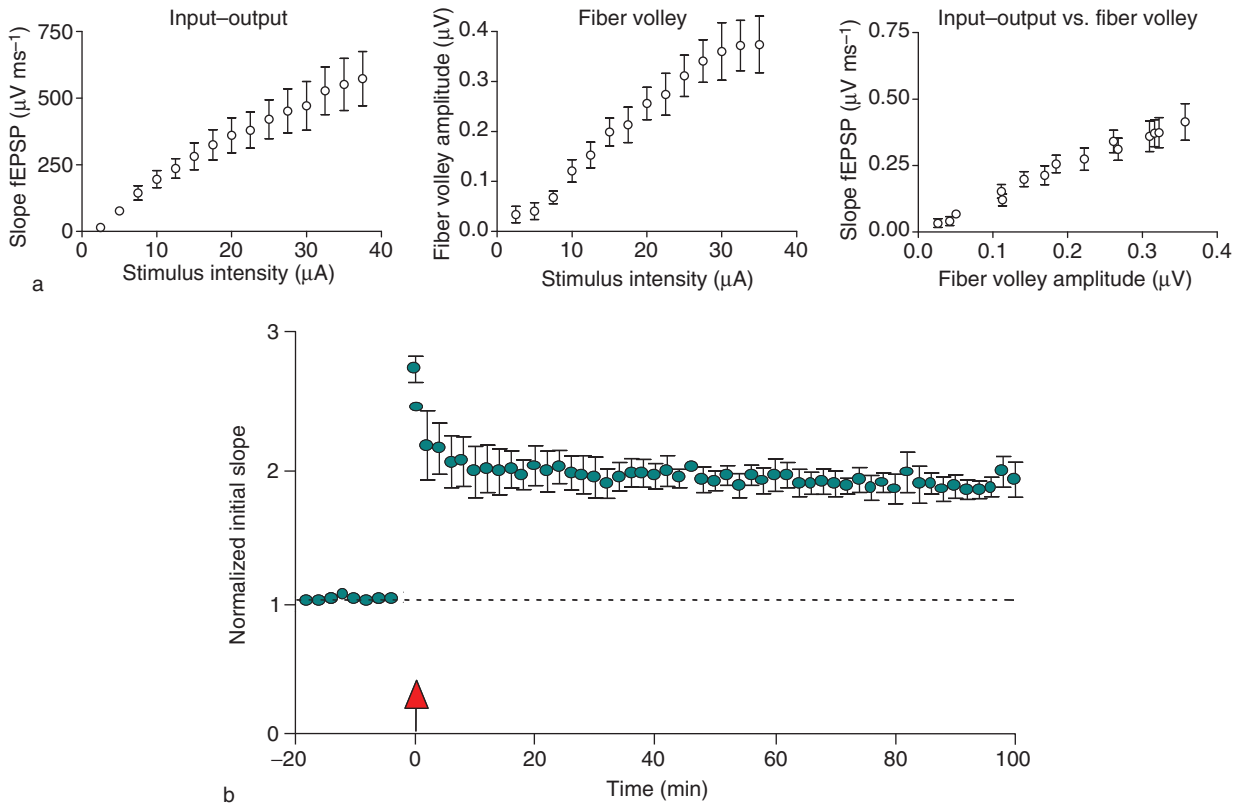

Figure 4 An input–output curve and typical LTP experiment: (a) the relationship of EPSP magnitude versus stimulus intensity (in microampheres) and the same data converted to an input–output relationship for EPSP versus fiber volley magnitude to allow an evaluation of postsynaptic response versus presynaptic response in the same hippocampal slice; (b) a typical high-frequency stimulation-induced potentiation of synaptic transmission in area CA1 of a rat hippocampal slice *in vitro*. The arrow in (b) indicates the delivery of 100 Hz (100 pulses s^{-1}) synaptic stimulation. EPSP, excitatory postsynaptic potential. Data courtesy of Ed Weeber and Coleen Atkins. Reproduced from Sweatt D (2003) *Mechanisms of Memory*. London: Elsevier, with permission from Elsevier.

Early studies of LTP used mostly high-frequency (100 Hz) stimulation in repeated 1 s-long trains as the LTP-inducing stimulation protocol. Although these protocols are still widely used to good effect, it is clear that such prolonged periods of high-frequency firing do not occur physiologically in the behaving animal. LTP can also be induced by stimulation protocols that are much more like naturally occurring neuronal firing patterns in the hippocampus. To date, the forms of LTP induced by these types of stimulation have all been found to be NMDA receptor-dependent in area CA1. Two popular variations of these protocols are based on the natural occurrence of an increased rate of hippocampal pyramidal neuron firing while a rat or mouse is exploring and learning about a new environment. Under these circumstances, the hippocampal pyramidal neurons fire bursts of action potentials at approximately 5 bursts per second, or 5 Hz. This is the hippocampal theta rhythm that has been described in the literature. One variation of LTP-inducing stimulation that mimics this pattern of firing is referred to as theta-frequency stimulation (TFS) and consists of 30 s of a single stimulus delivered at 5 Hz. Another variation, theta-burst stimulation (TBS) consists of three trains of stimuli delivered at 20 s intervals, each train composed of 10 stimulus bursts delivered at 5 Hz, with each burst consisting of four pulses at 100 Hz. It is worth noting that these patterns of stimulation, which are based on naturally occurring firing patterns *in vivo*, lead to LTP in hippocampal slice preparations as well.

Pairing LTP

Of course, there are much more sophisticated electrophysiological techniques than extracellular recording that we can use to monitor synaptic function. Intracellular recording and patch clamp techniques that measure electrophysiological responses in single neurons have also been used widely in studies of LTP. In these recording configurations, we can induce synaptic potentiation using tetanic stimulation or

theta-pattern stimulation, and we can measure LTP as an increase in postsynaptic currents through glutamate-gated ion channels or as an increase in postsynaptic depolarization when monitoring the membrane potential.

Control of the postsynaptic neuron's membrane potential with cellular recording techniques also allows for some sophisticated variations of the LTP induction paradigm. In one particularly important series of experiments, it was discovered that LTP can be induced by pairing repeated single presynaptic stimuli with postsynaptic membrane depolarization, called pairing LTP.

The basis for pairing LTP comes from one of the fundamental properties of the NMDA receptor. The NMDA receptor is both a glutamate-gated channel and a voltage-dependent one. The simultaneous presence of glutamate and a depolarized membrane is necessary and sufficient (when the co-agonist glycine is present) to gate the channel. Pairing synaptic stimulation with membrane depolarization provided via the recording electrode (plus the low levels of glycine always normally present) opens the NMDA receptor channel and leads to the induction of LTP.

How is it that the NMDA receptor triggers LTP? The NMDA receptor is a Ca^{2+} channel, and its gating leads to elevated intracellular Ca^{2+} in the postsynaptic neuron. It is this Ca^{2+} influx that triggers LTP. The molecular processes triggered by this Ca^{2+} influx are quite complex and involve many of the basic molecular processes underlying synaptic transmission, including presynaptic functions such as glutamate release and postsynaptic functions such as neurotransmitter receptor regulation. A brief listing of some of the molecular sites subject to modification in LTP is given in Figure 5.

The gating of the NMDA receptor/channel involves a voltage-dependent Mg^{2+} block of the channel pore. Depolarization of the membrane in which the NMDA receptor resides is necessary to drive the divalent Mg^{2+}cation out of the pore, which then allows Ca^{2+} ions to flow through. Thus, the simultaneous occurrence of both glutamate in the synapse and a depolarized postsynaptic membrane is necessary to open the channel and allow LTP-triggering Ca^{2+} into the postsynaptic cell.

These properties, glutamate dependence and voltage dependence, of the NMDA receptor allow it to function as a coincidence detector. This is a critical aspect of NMDA receptor regulation, and this allows for a unique contribution of the NMDA receptor to information processing at the molecular level. Using the NMDA receptor, the neuron can trigger a unique event, Ca^{2+} influx, specifically when a particular synapse is both active presynaptically (glutamate is present in the synapse) and postsynaptically (when the membrane is depolarized).

This confers a computational property of associativity on the synapse. This attribute is nicely illustrated by pairing LTP, as already described, when low-frequency synaptic activity paired with postsynaptic depolarization can lead to LTP. The associative property of the NMDA receptor allows for many other types of sophisticated information processing as well, however. For example, activation of a weak input to a neuron can induce potentiation, provided a strong input to the same neuron is activated at the same time. These particular features of LTP induction have stimulated a great deal of interest because they are reminiscent of classical conditioning with depolarization and synaptic input roughly corresponding to unconditioned and conditioned stimuli, respectively.

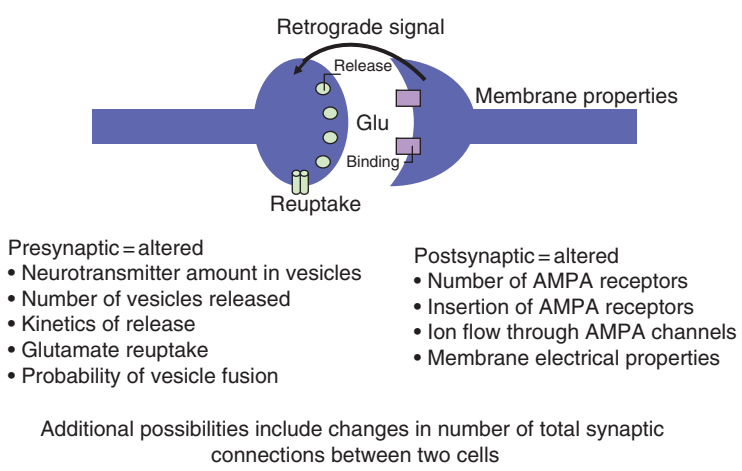

Figure 5 Potential sites of synaptic modification in LTP. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; Glu, glutamate; LTP, long-term potentiation. Figure by J. David Sweatt and Sarah E. Brown. Reproduced from Sweatt D (2003) *Mechanisms of Memory*. London: Elsevier, with permission from Elsevier.

LTP Can Be Divided into Phases

Contemporary models divide very long-lasting LTP (i.e., LTP lasting in the range of 5–6 h) into at least two phases. LTP comprising both phases can be induced with repeated trains of high-frequency stimulation in area CA1, and the phases are expressed sequentially over time to constitute what we call LTP. Late LTP (L-LTP) is hypothesized to be dependent for its induction on changes in gene expression, and this phase of LTP lasts many hours. Early LTP (E-LTP) is probably subserved by persistently activated protein kinases; it starts at around 15 min or less posttetanus and is over after approximately 2–3 h.

E-LTP and L-LTP refer to different temporal phases of LTP. These two phases are subserved by different maintenance mechanisms of different time courses and durations. Maintenance, or more specifically a maintenance mechanism, refers to the persisting biochemical signal that lasts in the cell. This persisting biochemical signal acts on an effector, for example, a glutamate receptor or the presynaptic release machinery, resulting in the expression of LTP. In contrast, induction refers to the transient events serving to trigger the formation of LTP.

It is important to keep in mind that, depending on the design of the experiment, induction, maintenance, and expression can be differentially inhibited (see Figure 6). The simplest type of experiment does not do this. For example, if we apply an enzyme inhibitor (or knock out a gene) before, during, and after the period of LTP-inducing high-frequency stimulation, this manipulation may block LTP; however, this does not distinguish whether the missing activity is required for the induction, the expression, or the maintenance of LTP. To distinguish among these possibilities, imagine instead that we apply the inhibitor selectively at different times during the experiment. If inhibitor is applied only during the tetanus and then is washed out and it blocks the generation of LTP, we can conclude that the enzyme is necessary for LTP induction. If the inhibitor is applied after the tetanus and it reverses the potentiation, it may be blocking either the maintenance or expression of LTP. Transient application of an inhibitor after tetanus may block synaptic potentiation, but the potentiation recovers after removal of the inhibitor. This is a blockade of LTP expression. However, if the inhibitor causes the potentiation to be lost irreversibly, the inhibitor has then by definition blocked the maintenance of LTP.

LTP outside the Hippocampus

The abundance of literature dedicated to studying LTP in the hippocampus might lead a newcomer to the field to suppose that LTP is somehow restricted to these synapses. However, plasticity of synaptic function, including phenomena such as LTP and long-term depression (LTD), is the rule rather than the exception for most forebrain synapses. LTP outside the hippocampus has been studied mostly in the cerebral cortex and the amygdala. The likely functional roles for LTP at these other sites are quite diverse, but two specific examples are worth highlighting. LTP-like processes in the cerebral cortex play a role in the activity-dependent development of the visual system and other sensory systems. LTP in the amygdala has received prominent attention as a mechanism contributing to cued fear conditioning.

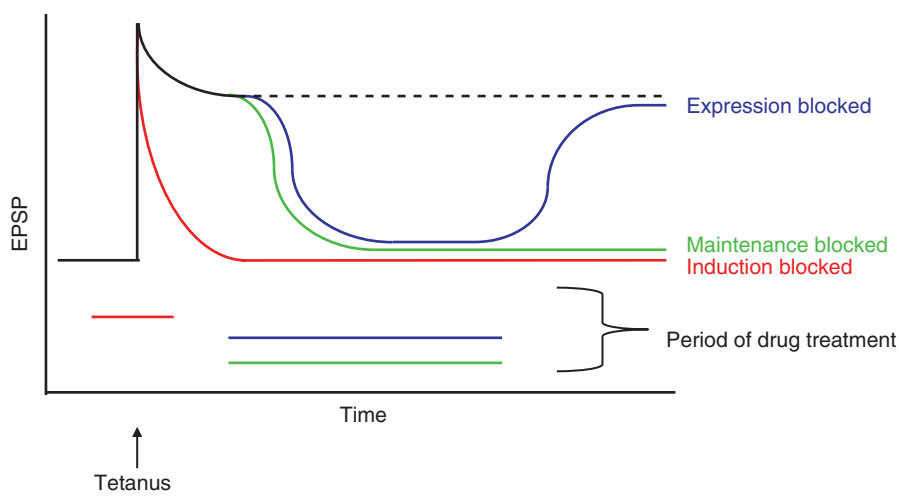

Figure 6 The different experimental approaches to dissecting the effects on the biochemical mechanisms subserving LTP induction, maintenance, or expression. EPSP, excitatory postsynaptic potential; LTP, long-term potentiation. Reproduced from Sweatt D (2003) *Mechanisms of Memory*. London: Elsevier, with permission from Elsevier.

The role of LTP in amygdala-dependent fear conditioning, in fact, is the area for which the strongest case can be made for a direct demonstration of a behavioral role for LTP. It is important to bear in mind that cortical LTP and amygdalar LTP probably exhibit some mechanistic differences from the NMDA receptor-dependent LTP that we are focusing on.

Depotentiation and LTD

If synapses can be potentiated and this potentiation is very long-lasting, over time the synapses will be driven to their maximum synaptic strength. In this condition, there is no longer synaptic plasticity and no further capacity for that synapse to participate in synaptic-plasticity-dependent processes. Worse, over the lifetime of an animal synapses will by random chance experience LTP-inducing conditions (presynaptic activity coincident with a postsynaptic action potential, for example) numerous times. If LTP is irreversible, ultimately every synapse will be maximally potentiated – obviously not a desirable condition vis-à-vis memory storage.

Consideration of this conundrum raises two implications. First, synapses that are involved in lifelong memory storage must be rendered essentially aplastic. In order to have good fidelity of memory storage over the lifetime of an animal, a synapse involved in permanent memory storage must be rendered immutable to a change in synaptic strength due to the random occurrence of what would normally be LTP-inducing stimulation.

But what about synapses such as those in the hippocampus that are not sites of memory storage but, rather, whose plasticity is part of the active processing of forming new long-term memories? In order to retain their plasticity and, hence, their capacity to contribute to memory formation, their potentiation must be reversible. Schaffer collateral synapses can undergo activity-dependent reversal of LTP, a phenomenon termed depotentiation (see Figure 7). Another activity-dependent way to decrease synaptic strength is LTD, the mirror image of LTP. LTD is a long-lasting decrease of synaptic strength below the baseline. Using logic similar to that of the first paragraph of this section, the phenomenon of dedepression of synaptic transmission is implied, although this has not been widely studied at this point.

As a practical matter, it is often difficult to separate depotentiation from LTD experimentally. For example, a baseline response in hippocampal slices or *in vivo* probably is a mixture of basal synaptic activity and activity at previously potentiated synapses. Moreover, for the most part the stimulation protocols used to induce depotentiation are variations of the protocols used to induce LTD. Nevertheless, mechanistic investigations have made clear that depotentiation and LTD use different mechanisms and thus must be considered distinct processes.

Physiological LTD (and depotentiation) induction protocols generally involve variations of repetitive 1-Hz stimulation. A common protocol is to deliver 900 stimuli at 1 Hz, but there are also LTD protocols that use random small variations in frequency in the 1-Hz region and variations that use paired-pulse stimuli delivered at 1 Hz. Synaptic depression appears to be fairly robust *in vivo*, but it is quite difficult to induce in hippocampal slices from adult animals. LTD *in vitro* is almost always studied using slices from immature animals or cultured immature neurons, and it is possible that LTD as it is currently studied *in vitro* is largely a manifestation of what is normally a developmental mechanism.

One ironic aspect of the LTP/LTD story is that both phenomena at Schaffer collateral synapses can be blocked by NMDA receptor antagonists. This suggests that Ca^{2+} influx triggers both processes, and indeed current models of LTD induction hypothesize that LTD is caused by an influx of Ca^{2+} that achieves a lower level than that needed for LTP induction. This lower level of Ca^{2+} is hypothesized to selectively activate protein phosphatases and, by this mechanism, lower synaptic efficacy.

Another very different type of LTD is cerebellar LTD. Cerebellar LTD occurs at synapses on Purkinje neurons in the cerebellar cortex. Cerebellar LTD is a very interesting phenomenon because its behavioral role is much better understood than the hippocampal plasticity phenomena. Among other things, cerebellar LTD is involved in associative eyeblink conditioning, a cerebellum-dependent classical conditioning paradigm. Considerable progress has been made in investigating the roles and mechanisms of cerebellar LTD as well.

Summary

Like learning, LTP can be defined as a long-lasting change in output in response to a transient input. The persistence of this effect has been demonstrated to extend many hours *in vitro* and several weeks *in vivo*. We do not know how LTP relates to memory, and there is evidence for and against the hypothesis that hippocampal LTP is involved in memory. Regardless, it is the best-understood example of long-lasting synaptic plasticity in the mammalian CNS, and it is a model for how long-lasting memory-associated changes are likely to occur in the CNS. Studying LTP has yielded valuable insights into the mechanisms of plasticity that underlie learning and memory in

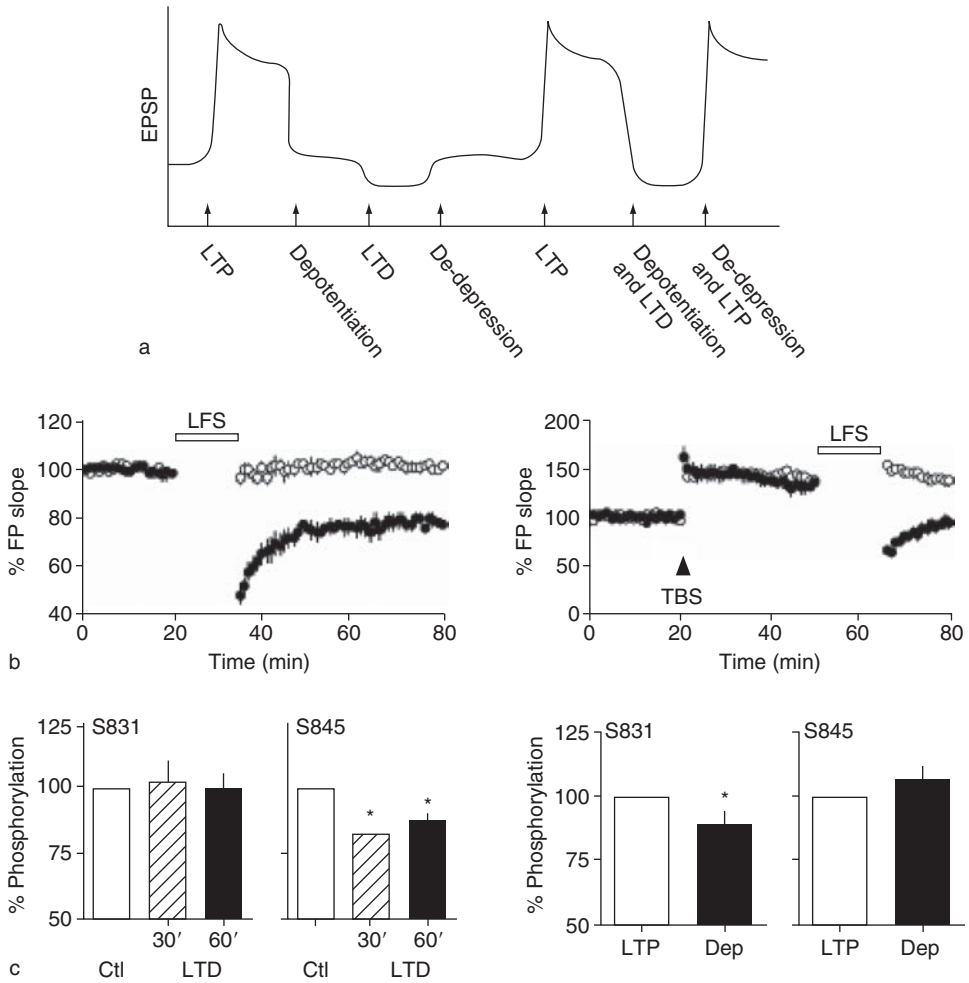

Figure 7 Depotentiation and LTD: (a) LTP, depotentiation, LTD, de-depression, and combinations of these; (b) LTD and depotentiation in hippocampal neurons, shown in simultaneous recording of slices receiving baseline stimulation (control, open circles) and 1-Hz stimulation (closed circles); (c) LTD, depotentiation, and dephosphorylation. In (b), regulation of distinct AMPA receptor phosphorylation sites during bidirectional synaptic plasticity is shown. In (c), homosynaptic LTD in CA1 is associated with dephosphorylation of GluR1 at a PKA site (ser (S)845); depotentiation gives dephosphorylation at a CaMKII/PKC site (S831). AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; CaMKII, calmodulin-dependent protein kinase II; Ctl, control; Dep, depotentiation; EPSP, excitatory postsynaptic potential; FP, field potential; GluR, glutamate receptor; LFS, low-frequency stimulation; LTD, long-term depression; LTP, long-term potentiation; PKA, protein kinase A; PKC, protein kinase C; TBS, theta-burst stimulation. Reproduced from Sweatt D (2003) *Mechanisms of Memory*. London: Elsevier, with permission from Elsevier.

the brain. The bona fide changes in neuronal connections that occur *in vivo* may or may not be identical to LTP as it is presently studied in the laboratory, but this does not diminish its utility as a cellular model system for studying lasting neuronal change in the mammalian CNS.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Hippocampus; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Transmission: Models.

Further Reading

Andersen P, Sundberg SH, Sveen O, and Wigstrom H (1977) Specific long-lasting potentiation of synaptic transmission in hippocampal slices. *Nature* 266: 736–737.

Barrionuevo G and Brown TH (1983) Associative long-term potentiation in hippocampal slices. *Proceedings of the National Academy of Sciences of the United States of America* 80: 7347–7351.

Bliss TV and Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* 232: 331–356.

Collingridge GL, Kehl SJ, and McLennan H (1983) Excitatory amino acids in synaptic transmission in the Schaffer collateral-commissural pathway of the rat hippocampus. *Journal of Physiology* 334: 33–46.

Hebb DO (1949) *The Organization of Behavior: A Neuropsychological Theory.* Hillsdale, NJ: Lawrence Erlbaum.

Johnston D and Amaral DG (1998) Hippocampus. In: Shepherd GM (ed.) *The Synaptic Organization of the Brain,* 4th edn., pp. 417–458. New York: Oxford University Press.

Johnston D and Wu SM (1995) *Foundations of Cellular Neurophysiology.* Cambridge, MA: MIT Press.

Lee HK, Barbarosie M, Kameyama K, Bear MF, and Huganir RL (2000) Regulation of distinct AMPA receptor phosphorylation sites during bidirectional synaptic plasticity. *Nature* 405: 955–959.

Linden DJ (1999) The return of the spike: Postsynaptic action potentials and the induction of LTP and LTD. *Neuron* 22: 661–666.

Magee JC and Johnston D (1997) A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. *Science* 275: 209–213.

Nicoll RA and Malenka RC (1999) Expression mechanisms underlying NMDA receptor-dependent long-term potentiation. *Annals of the New York Academy of Sciences* 868: 515–525.

Sweatt JD (2003) Mechanisms of Memory. London: Elsevier.

Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP

A Contractor, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Mossy fiber synapses in the hippocampus are formed by the axons of the granule cells and their postsynaptic targets in the hilar and CA3 regions of the hippocampus. The excitatory synapses formed with the principal pyramidal neurons exhibit a number of unique structural and functional properties, including a distinctive form of long-term potentiation (LTP) that is expressed in the presynaptic terminal and is dependent on cyclic adenosine monophosphate (cAMP) signaling. Here the current state of our understanding of this form of synaptic plasticity is reviewed.

The Mossy Fiber Synapse

The hippocampus is a widely studied structure because of its well-described involvement in memory processing and formation. The circuitry of excitatory synaptic connections in this structure is relatively simple with an almost linear progression in connectivity, and therefore information flow, between the granule neurons in the dentate gyrus, which receive the largest proportion of extrinsic connections from the cortex, through the CA3 and CA1 regions and outputting to the subiculum and cortex. However, this simplified scheme in reality is somewhat more complex; in the CA3 region in particular, extensive axon collaterals of the pyramidal neurons make synapses onto other pyramidal neurons and local circuit interneurons. Neurons in the CA3 also receive a direct perforant path connection from the cortex. While the absolute number of collateral synapses per pyramidal neuron is much greater, the connection between the granule cell axons, termed the mossy fibers, and the CA3 pyramids has the greatest impact on CA3 pyramidal excitability. The mossy fiber axons extend from the dentate gyrus traversing through the hilus and projecting through the stratum lucidum. This long axonal projection (upwards of 3 mm in the rat) forms en passant connections in the CA3 with the proximal dendrites of the pyramidal neurons and the local circuit interneurons. In a stunning example of target cell-specific synaptic segregation, the connections to the pyramidal neuron and interneuron have completely divergent functional properties. Connections onto the pyramidal neurons are formed by large mossy fiber boutons that are typically 2–8 μm in diameter. From these large synaptic terminals, filopodial extensions reach out tens of micrometers to form synaptic contact exclusively with interneurons. The main mossy fiber bouton apposes complex, multi-headed spines, termed thorny excresences, which extend from the proximal dendrites of the CA3 pyramidal neurons; each terminal contains multiple release sites that are apposed to multiple postsynaptic densities on the same spine. This article focuses almost exclusively on the mossy fiber connection to the principal CA3 pyramidal neurons.

Expression Mechanisms of Mossy Fiber LTP

Long-term potentiation (LTP) was first described in the hippocampus in pioneering work from Tim Bliss and Terje Lømo in the early 1970s. These seminal studies quickly led to a large and sustained effort to describe a synaptic phenomenon that is now widely believed to be a cellular correlate of memory formation. LTP at the mossy fiber to CA3 pyramidal neuron synapse was described a decade after these initial findings and soon proved to be mechanistically distinct from LTP in the CA1 region of the hippocampus and the dentate gyrus. Several important findings about the induction of mossy fiber LTP were described in these early studies. For instance, several groups found that *N*-methyl-D-aspartate (NMDA) receptor activation, a defining characteristic of LTP at most other synapses, was not required for the induction of mossy fiber LTP. Whereas some aspects of mossy fiber LTP remain controversial, in particular the requirement of postsynaptic Ca^{2+} signaling for induction, other aspects, such as the expression mechanisms, are more firmly established. Several coincident lines of evidence have confirmed that mossy fiber to CA3 pyramidal neuron LTP is expressed in the presynaptic terminal as a long-lasting increase in glutamate release probability. Among these analyses are clear reductions in paired pulse ratio, increased rate of open channel block of NMDA receptors, and quantal variance analysis, which strongly support an increase in transmitter release probability following induction of LTP.

Involvement of cAMP Signaling in Mossy Fiber LTP

A complete description that incorporates all the published findings for the expression mechanisms of

mossy fiber LTP has not been established, and it seems likely that a consensus on a simple unitary model is likely to be elusive because several parallel pathways may underlie mossy fiber LTP. Despite this, the involvement of several molecules has been firmly established using molecular/genetic techniques and more acute pharmacological manipulation. A major signaling pathway that has overwhelming evidence supporting its involvement in the expression of mossy fiber LTP is the cAMP cascade resulting in the activation of the cAMP-dependent protein kinase PKA. The earliest evidence that cAMP might be involved in mossy fiber LTP came from pharmacological manipulations. Application of the diterpene forskolin, which directly activates adenylyl cyclase to elevate cAMP, produces a long-lasting enhancement of mossy fiber transmission. Similarly, analogs of cAMP also potentiate mossy fiber transmission. Forskolin potentiation and activity-induced LTP mutually occlude one another, suggesting, but not implicitly confirming, that elevation of cAMP is involved in mossy fiber LTP. Further support for a role of cAMP elevation in mossy fiber LTP comes from studies in knockout mice. Evidence suggests that elevations in presynaptic Ca^{2+} are required for the LTP cascade, specifically through the activation of presynaptic R-type voltage-dependent Ca^{2+} channels. Therefore, the focus has been on Ca^{2+}-activated forms of adenylyl cyclase as the effectors of this rise in cAMP. There are two Ca^{2+}-sensitive adenylyl cyclases expressed in the central nervous system (CNS): adenylyl cyclase 1 (AC1) and adenylyl cyclase 8 (AC8). Knockout mice for either one of these proteins show impaired mossy fiber LTP, further supporting the role of cAMP signaling at this synapse.

Downstream Targets of cAMP

Elevations in cAMP may have several potential downstream targets that could mediate changes in release probability. At the mossy fiber synapse, the emphasis has been on the cAMP-dependent protein kinase PKA; however, it is also worth mentioning the other potential effectors that might be involved. At the neuromuscular junction of both crayfish and *Drosophila*, cAMP elevation potentiates transmitter release by acting directly on presynaptic hyperpolarization-activated cyclic nucleotide-gated channels that underlie the hyperpolarization-activated cation channel (*I*h). A similar role for presynaptic *I*h in mossy fiber LTP is controversial. Incubation of ZD7288, an *I*h channel blocker, reduces LTP and forskolin-mediated potentiation of mossy fiber transmission. However, the interpretation of this finding is complicated, as ZD7288 was found to have nonspecific actions on synapses and by itself causes a robust depression of mossy fiber basal synaptic transmission. Presynaptic *I*h therefore is unlikely to play a substantive role in mossy fiber LTP.

Recently there has been considerable interest in another downstream target of cAMP: the cAMP-sensitive guanine nucleotide exchanging factor (GEF) Epac. This GEF has been implicated in synaptic potentiation at the medial nucleus of the trapezoid body (MNTB) neurons in the brain stem and also in potentiation of transmission at the *Drosophila* neuromuscular junction. Epac is expressed at high levels in the hippocampus; however, thus far there has not been any suggestion that it plays a role in transmitter release or potentiation at any hippocampal synapse, including at mossy fibers.

The more commonly accepted downstream target of cAMP is PKA, and, as alluded to earlier, there is substantial evidence that this kinase plays a central role in mossy fiber LTP. Inhibitors of PKA block mossy fiber LTP, and targeted mutation of either regulatory or catalytic subunits of PKA has specific effects on mossy fiber LTP. While it is commonly assumed that PKA in the presynaptic terminal plays a critical role, it is worth noting that inclusion of PKA inhibitory peptide in the postsynaptic cell can also block the establishment of mossy fiber LTP, possibly suggesting a more complex scheme in which both pre- and postsynaptic PKA activity is required for the full expression of mossy fiber LTP.

Potential Substrates of PKA Involved in Elevation of Transmitter Release

All available evidence seems to suggest that PKA is central to mossy fiber LTP; therefore, there has been a concerted effort to uncover the molecular substrates of this enzyme that could underlie the sustained increase in transmitter release. Two molecules that have received particular scrutiny for their role in mossy fiber LTP are Rab3a and RIM1α. Rab3a is a vesicle-associated GTPase which has been implicated in vesicle fusion. Rab3a is not itself a substrate of PKA, at least in *in vitro* assays, but it interacts with other synapse-associated proteins that are. RIM1α is an active zone protein which is thought to be a presynaptic scaffold. The involvement of both these molecules in mossy fiber LTP has been determined using examination of knockout mice. The model that has emerged from these studies suggests that phosphorylation of RIM1α and its interaction with Rab3a are required for the long-lasting increase in neurotransmitter release. While the exact roles of

these proteins are still unknown, the clear impairment in mossy fiber LTP in knockout mice put these two molecules at the center of efforts to uncover the molecular basis for plasticity at this synapse. However, several observations suggest that this model may be somewhat simplistic. One assumption from this model might be that if PKA and Rab3a converge upon RIM1α, raising cAMP with forskolin would abrogate the synaptic potentiation observed with this diterpene in RIM1α knockout mice. However, normal forskolin-mediated potentiation is observed, suggesting that other parallel mechanisms may in fact mediate cAMP signaling. While it remains unclear what the precise nature of this mechanism might be, several potential PKA substrates have been excluded as mediators of mossy fiber plasticity. Knockout mice in which synapsin I and II or rabphilin are ablated express normal mossy fiber LTP, suggesting that these three PKA substrates play no role. Therefore, while there is evidence that activation of PKA phosphorylates RIM1α and possibly facilitates an interaction with the vesicle-associated protein Rab3a resulting in a sustained increase in transmitter release, this model does not integrate all the available data from mechanistic studies and is likely to undergo further revision as additional studies uncover more information about how PKA exerts its actions in mossy fiber LTP.

State-Dependent cAMP Signaling at Mossy Fiber to Interneuron Synapses

As noted previously, the mossy fiber axons make contact onto both pyramidal neurons and stratum lucidum interneurons in the CA3 region. These two synaptic connections demonstrate a remarkable segregation in their functional properties. High-frequency stimulation of mossy fibers, which causes LTP at pyramidal cell synapses, actually depresses mossy fiber interneuron synapses through activation of presynaptic metabotropic glutamate receptor 7 (mGluR7) and protein kinase C (PKC) signaling. Naive interneuron synapses are not potentiated by elevating cAMP by application of forskolin, suggesting that these synapses are under regulation by fundamentally different mechanisms despite being formed by the same axons and being closely spatially arranged (filopodial synapses onto interneurons emanate from the main mossy fiber boutons that form the synapses onto the pyramidal neuron). However, one interesting recent report demonstrated that activation of mGluR7 and depression of the mossy fiber interneuron synapse uncovers a subsequent sensitivity to cAMP-dependent potentiation. This state-dependent responsiveness to cAMP suggests that this signaling pathway is more ubiquitous in mossy fibers than was originally thought.

Mossy Fiber LTD cAMP Dependency

The functional inverse of LTP is long-term depression (LTD) of synaptic transmission. A simplistic view of LTD is that it should also engage the same molecular machinery as LTP, but in reverse. While it turns out that things are usually more complicated, there are many examples of molecules that have central roles in both LTD mechanisms and LTP. At the mossy fiber to pyramidal cell synapse, LTD involves cAMP signaling and therefore may engage at least some of the same machinery as LTP. Mossy fiber LTD can be induced by long trains of low-frequency stimulation or by the application of group II mGluR agonist combined with activity. LTD at this synapse is blocked by inhibitors of PKA. In initial studies it was also reported that LTD was absent in Rab3a knockout mice. These results had established a model that seemed to involve the same molecular players as mossy fiber LTP. However, later studies by the same group could not replicate this finding in Rab3a knockout mice, bringing this model into some doubt. Examination of knockout mice for the other prominent player in mossy fiber LTP, RIM1α, is also consistent with a separate molecular pathway for LTD. In fact, the magnitude of mossy fiber LTD in RIM1α knockout mice is enhanced, possibly due to a masked potentiation occurring in parallel with LTD with low-frequency stimulation at some synapses. Therefore, it seems that LTD of mossy fiber to pyramidal cell synapses, while involving cAMP signaling, is unlikely to be an inverse process of LTP, but instead involves an alternate pathway whose components are still unknown.

Comparison with Other cAMP-Dependent Forms of LTP

Three other synapses have been demonstrated to have forms of LTP similar to that at the mossy fiber pyramidal cell synapse. Corticothalamic fibers originating in layer VI of the neocortex form synapses onto neurons in the ventrobasal thalamus. LTP at corticothalamic synapses is NMDA receptor independent and expressed presynaptically. These synapses are also subject to forskolin potentiation, and LTP can be blocked by PKA inhibitors. Beyond these initial finding it is not clear whether corticothalamic synapses utilize the same molecular machinery as mossy fibers for the expression of LTP.

In the cerebellum, parallel fiber axons of the granule cells form synapses onto Purkinje cell dendrites in

the molecular layer. Parallel fiber transmission is potentiated by forskolin, which occludes subsequent LTP. Further, parallel fiber LTP induction is blocked by inhibitors of PKA and is almost absent in RIM1α knockout mice, further suggesting that parallel fibers and mossy fibers utilize similar LTP mechanisms. In a reduced co-culture system, reintroduction of RIM1α rescued parallel fiber LTP, but expression of a mutant of a putative phosphorylation site of RIM1α (serine 413) did not rescue and itself acted as a dominant negative to block LTP in wild-type cultures. These experiments have further reinforced the central role of phosphorylation of RIM1α in the expression of presynaptic cAMP-dependent forms of LTP.

A final example of presynaptic LTP that is dependent on cAMP signaling is at the corticoamygdala synapse. Although the mechanisms underlying this form of LTP are likely to be different than those at mossy fiber synapses (for instance, NMDA receptor activation is required for induction), some similarities exist. Synaptic transmission is potentiated by forskolin, and this occludes subsequent LTP induction, and PKA inhibitors also block LTP. Interestingly, corticoamygdala LTP is impaired in Rab3a knockout mice but is not significantly impaired in RIM1α knockouts, perhaps suggesting a different signaling pathway through a different, as yet unidentified, PKA substrate.

Outstanding Questions

While evidence from numerous studies has begun to dissect the molecular pathways that underlie LTP at mossy fiber synapses, it seems we are some way from consolidating all the available evidence into a single model. There is strong evidence that cAMP signaling plays a central role in the expression mechanisms; however, it is possible that it exerts its effects through multiple pathways. Evidence from knockout studies has demonstrated the importance of Rab3a and RIM1α, the putative downstream substrate of PKA, in mossy fiber LTP; however, unpredictably, forskolin potentiation is still observed in these mice. The role of presynaptic *I*h has been disputed at mossy fiber synapses, but at least at other synapses it does play a role in regulating transmitter release. Similarly, many of the early studies have relied upon inhibitors whose specificity is somewhat unreliable. There are numerous inconsistencies in the literature regarding the induction and expression of mossy fiber LTP. Among these is the finding that postsynaptic mechanisms may play a role in the induction process. Recently there has been evidence that ephrin signaling is also required for mossy fiber LTP. It remains to be determined whether a consensus can be established and whether in fact mossy fiber LTP might turn out to be more complex than first thought.

See also: Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); NMDA Receptors and Development; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Castillo PE, Janz R, Sudhof TC, Tzounopoulos T, Malenka RC, and Nicoll RA (1997) Rab3A is essential for mossy fibre long-term potentiation in the hippocampus. *Nature* 388: 590–593.

Castillo PE, Schoch S, Schmitz F, Sudhof TC, and Malenka RC (2002) RIM1alpha is required for presynaptic long-term potentiation. *Nature* 415: 327–330.

Henze DA, Urban NN, and Barrionuevo G (2000) The multifarious hippocampal mossy fiber pathway: A review. *Neuroscience* 98: 407–427.

Kaeser PS and Sudhof TC (2005) RIM function in short- and long-term synaptic plasticity. *Biochemistry Society Transactions* 33: 1345–1349.

Malenka RC and Bear MF (2004) LTP and LTD: An embarrassment of riches. *Neuron* 44: 5–21.

Nicoll RA and Schmitz D (2005) Synaptic plasticity at hippocampal mossy fibre synapses. *Nature Reviews Neuroscience* 6: 863–876.

Weisskopf MG, Castillo PE, Zalutsky RA, and Nicoll RA (1994) Mediation of hippocampal mossy fiber long-term potentiation by cyclic AMP. *Science* 265: 1878–1882.

Long-Term Potentiation (LTP): NMDA Receptor Role

A J Doherty, S M Fitzjohn, and G L Collingridge, University of Bristol, Bristol, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The subtype of glutamate receptor known as the *N*-methyl-D-aspartate (NMDA) receptor is the trigger for both long-term potentiation (LTP) and long-term depression (LTD) at the majority of synapses in the central nervous system (CNS), where long-lasting synaptic plasticity has been observed. The role of NMDA receptors (NMDARs) in synaptic plasticity was first identified, and has been most extensively studied at, the synapses made between CA3 and CA1 pyramidal neurons in the hippocampus, the Schaffer collateral–commissural pathway. While most is known about LTP at these CA1 synapses, it seems that the principles discovered here apply elsewhere in the brain. The following discussions pertain to LTP at CA1 synapses.

Glutamate Receptors and Synaptic Transmission

It is commonly considered that α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) mediate synaptic transmission and NMDARs trigger alterations in synaptic efficiency. However, the reality is more complex. AMPARs do indeed mediate the majority of the response to low-frequency stimulation. However, NMDARs contribute significantly to high-frequency synaptic transmission. Given that in the conscious animal CA3 neurons often discharge in high-frequency bursts, then NMDARs are likely to play a major role in the transmission of synaptic information within the CA1 area. It is also sometimes mistakenly considered that only AMPARs mediate the response that is modified during LTP. While it is true that it is often the AMPAR-mediated response that is studied, it is equally true that the NMDAR-mediated component of synaptic transmission has the capacity to undergo LTP.

NMDAR Properties Govern Their Role in Synaptic Transmission

The NMDAR has unique properties that govern its role in synaptic processing. First, it is strongly antagonized by micromolar concentrations of Mg^{2+}. Second, this block is highly voltage dependent. Third, the NMDAR-mediated conductance has slow kinetics, compared with the AMPAR-mediated response. Fourth, the NMDAR not only passes monovalent cations but also has a significant permeability to Ca^{2+}.

During low-frequency transmission there is rapid activation of AMPARs. However, while L-glutamate binds to NMDARs, they are largely blocked by Mg^{2+}, which is present in the synaptic cleft at around 1 mM (Figure 1). This prevents NMDARs from contributing appreciably to the synaptic response. During high-frequency transmission there is a sustained depolarization, due in part to the temporal summation of AMPAR-mediated excitatory postsynaptic potentials (EPSPs), and this alleviates the voltage-dependent Mg^{2+} block to enable NMDARs to contribute to the synaptic response. Since they have a slow decay phase, the NMDAR-mediated EPSPs summate very effectively.

Induction of LTP

Induction by High-Frequency Stimulation

Conditions that enable the activation of NMDARs can result in the induction of LTP. The most commonly used induction stimulus is a continuous high-frequency train (tetanus), sometimes delivered more than once. This provides the depolarization to transiently relieve the Mg^{2+} block of NMDARs. The Ca^{2+} that then enters the neuron through the NMDARs provides the initial trigger for the plastic change. Although a tetanus typically comprises 100 stimuli or more (e.g., 100 Hz, 1 s), far fewer stimuli can induce LTP, provided that they are appropriately timed. One such protocol is the delivery of brief high-frequency bursts (e.g., four shocks at 100 Hz) with an interburst frequency of around 5 Hz (i.e., an interburst interval of 200 ms). This pattern of activity mimics the physiological activity that occurs during the theta rhythm, a frequency of activation observed in the electroencephalogram (EEG) when animals are exploring their environment. The minimal induction protocol is the delivery of a single priming stimulus followed approximately 200 ms later by a single burst (which could comprise as few as two stimuli, though typically four are used). Thus, under optimal conditions LTP can be induced by three to five stimuli.

Induction by Pairing

LTP can also be induced by pairing. This is achieved by maintaining low-frequency stimulation but depolarizing the neuron artificially by injecting current through

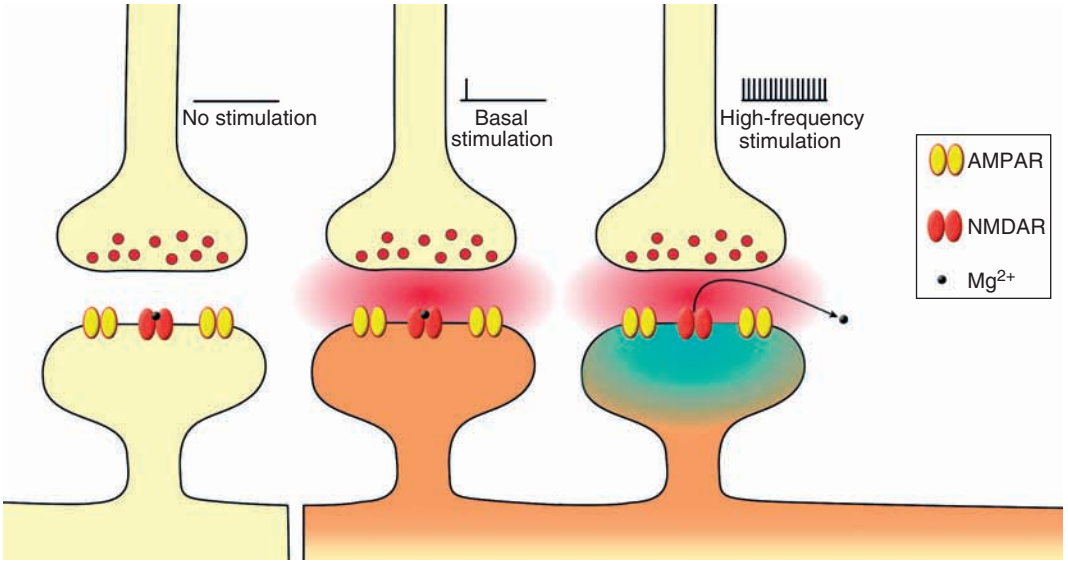

Figure 1 *N*-Methyl-D-aspartate receptors (NMDARs) are activated by high-frequency stimulation. Under basal conditions, NMDAR activity is blocked due to voltage-dependent binding of Mg^{2+} ions. A single stimulus results in the release of L-glutamate and activation of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) but does not provide sufficient depolarization to repel the Mg^{2+} ions. High-frequency stimulation results in a more sustained depolarization of the postsynaptic terminal, leading to the release of the Mg^{2+} block and an increase in cytosolic Ca^{2+} ions. The rise in Ca^{2+} concentration may be localized to the activated synapse, ensuring that only the activated synapse undergoes long-term potentiation.

the recording electrode. Typically, approximately 50 pairings of low-frequency stimulation and depolarization are employed. LTP can also be induced by the pairing of appropriately timed presynaptic and postsynaptic action potentials. The common feature of all of these induction protocols is that they enable sufficient depolarization to transiently relieve the Mg^{2+} block of NMDARs. Once this occurs, it enables Ca^{2+} to enter the synapse through the NMDARs in sufficient amounts to trigger the biochemical cascades that result in enhanced synaptic transmission.

NMDAR Activation without LTP

However, activation of NMDARs does not necessarily induce LTP. Under conditions whereby the LTP process has been saturated (by prior periods of LTP induction), protocols that normally induce LTP result in no change in synaptic strength. In addition, certain patterns of activation can activate NMDARs but induce long-term depression rather than LTP. If a pathway has already undergone LTP, then these same patterns of activation may induce depotentiation (DP) of the potentiated response. Typically, NMDAR-dependent LTD and DP are induced by prolonged periods of low-frequency stimulation (1–3 Hz). Finally, an intermediate frequency (e.g., 5 Hz) will neither induce LTP nor LTD, while still activating NMDARs.

Features of NMDAR-Dependent LTP

Input Specificity

NMDAR-dependent LTP has several hallmark features. Input specificity refers to the property that if two inputs are activated and converge on a population of neurons, or indeed a single neuron, but an induction protocol is delivered to just one of the inputs, then only that input will undergo LTP. The other input typically will not change its efficiency of transmission, or under certain circumstances may be depressed via a process known as heterosynaptic depression. Input specificity is an important property. It means that the unit of modification is the activated input, not the entire neuron. This implies that the unit of information storage can be as small as a single synapse. Input specificity is due to the requirement for L-glutamate to activate NMDARs and the fact that the Ca^{2+} that permeates NMDARs does not spread very far. Therefore, when a neuron is sufficiently depolarized, those synapses that are activated by synaptically released L-glutamate can undergo LTP. Nearby neighbors may also undergo LTP if either the synaptically released L-glutamate 'spills over' in sufficient amounts to activate enough NMDARs or if the Ca^{2+} spreads beyond the activated spine. In its purest form, LTP is restricted to the activated inputs.

Cooperativity

Cooperativity refers to the need to stimulate multiple afferent fibers to induce LTP. Experimentally this is observed as the failure of a 'weak tetanus' to induce LTP, while a 'strong tetanus' that activates more fibers is able to induce LTP. Cooperativity can be explained by the need for multiple inputs to be activated around the same time to provide sufficient depolarization to remove enough of the Mg^{2+} block to enable the induction of LTP. This property provides a threshold of activity that has to be exceeded for synaptic plasticity to occur. CA3 neurons typically fire in brief, synchronous bursts and this is presumably to exceed the cooperativity threshold to enable LTP to occur at CA1 synapses.

Associativity

Associativity is an extension of cooperativity, but whereby the 'strong tetanus' is delivered to an independent input. There is a critical timing window in which the strong input can be delivered after the weak input and still be effective. Associativity is explained by the weak input providing the necessary L-glutamate but insufficient depolarization, whereas the 'strong input' provides the (additional) depolarization to sufficiently alleviate the Mg^{2+} block. The associative depolarization could be provided by a second glutamatergic input or by a different neurotransmitter system that provides or enables depolarization at the critical time. Experimentally, the depolarization can be provided by the recording electrode (in a pairing experiment) and this can permit a single input onto a single neuron to undergo LTP. Thus the requirement for synchronous activity to induce LTP is to provide sufficient depolarization. In conclusion, the Mg^{2+} block and Ca^{2+} dependence of the NMDAR explain the properties of input specificity, cooperativity, and associativity.

Role of γ-Aminobutyric Acid Inhibition in LTP

Role of Postsynaptic γ-Aminobutyric Acid Receptors

γ-Aminobutyric acid (GABA) inhibition plays a pivotal role in LTP. When several afferents are activated at a low frequency the EPSP is mediated essentially exclusively by AMPARs. This is only because of synaptic inhibition, which is also activated via the excitation of feed-forward interneurons. If GABA receptors are blocked pharmacologically, then the AMPA-mediated EPSP is prolonged and NMDARs contribute much more substantially to the later part of a composite EPSP. Thus, it is the hyperpolarization by GABA to maintain, and indeed intensify, the Mg^{2+} block that prevents NMDARs contributing to any appreciable extent during low-frequency transmission. Thus, blocking GABA inhibition can facilitate the induction of LTP. However, the inhibition of GABA is not just some pharmacological curiosity. During the more physiological patterns of induction (priming and theta) there is an activity-dependent depression of GABA inhibition that is required to enable sufficient activation of NMDARs. This activity-dependent depression is mediated largely by the activation of a presynaptic $GABA_B$ autoreceptor.

Role of $GABA_B$ Autoreceptors

During low-frequency stimulation, GABA is released and evokes a biphasic inhibitory postsynaptic potential (IPSP), mediated by $GABA_A$ and $GABA_B$ receptors, to restrict the synaptic activation of NMDARs (Figure 2). Some of the GABA that is released activates a presynaptic $GABA_B$ autoreceptor, which inhibits subsequent GABA release. During low-frequency transmission this has no effect on GABA transmission, since the effect fully recovers within a second or so. But during high-frequency transmission the autoreceptor mechanism leads to a depression in GABA inhibition. The maximum effect occurs after around 200 ms, which explains the effectiveness of this interval for theta and priming-induced LTP. The $GABA_B$ autoreceptor mechanism is essential for the induction of LTP under these conditions, since antagonizing this receptor completely blocks priming-induced LTP. However, the depression of synaptic inhibition only occurs for a few stimuli as the synapse undergoes an activity-dependent facilitation, and, when sufficient stimuli are delivered, a depolarizing potential is generated. Therefore when more artificial induction protocols are used, such as tetanus, blocking $GABA_B$ autoreceptors has no effect on the induction of LTP. Given that in the exploring animal there are brief periods of synchronized high-frequency activity occurring at the theta frequency, it is likely that $GABA_B$ autoreceptors are critically involved in the induction of LTP *in situ*.

Signaling Mechanisms Involved in LTP

The induction of LTP requires an elevation of postsynaptic Ca^{2+}. This Ca^{2+} enters the synapse via the activated NMDARs, where the Ca^{2+} signal may be greatly magnified by Ca^{2+}-induced Ca^{2+} release from intracellular stores. It then activates various kinases. Several protein kinases have been implicated in the induction of LTP. These include Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII), protein kinases A and C (PKA, PKC), protein tyrosine kinase (PTK), mitogen-activated protein kinase (MAPK), and

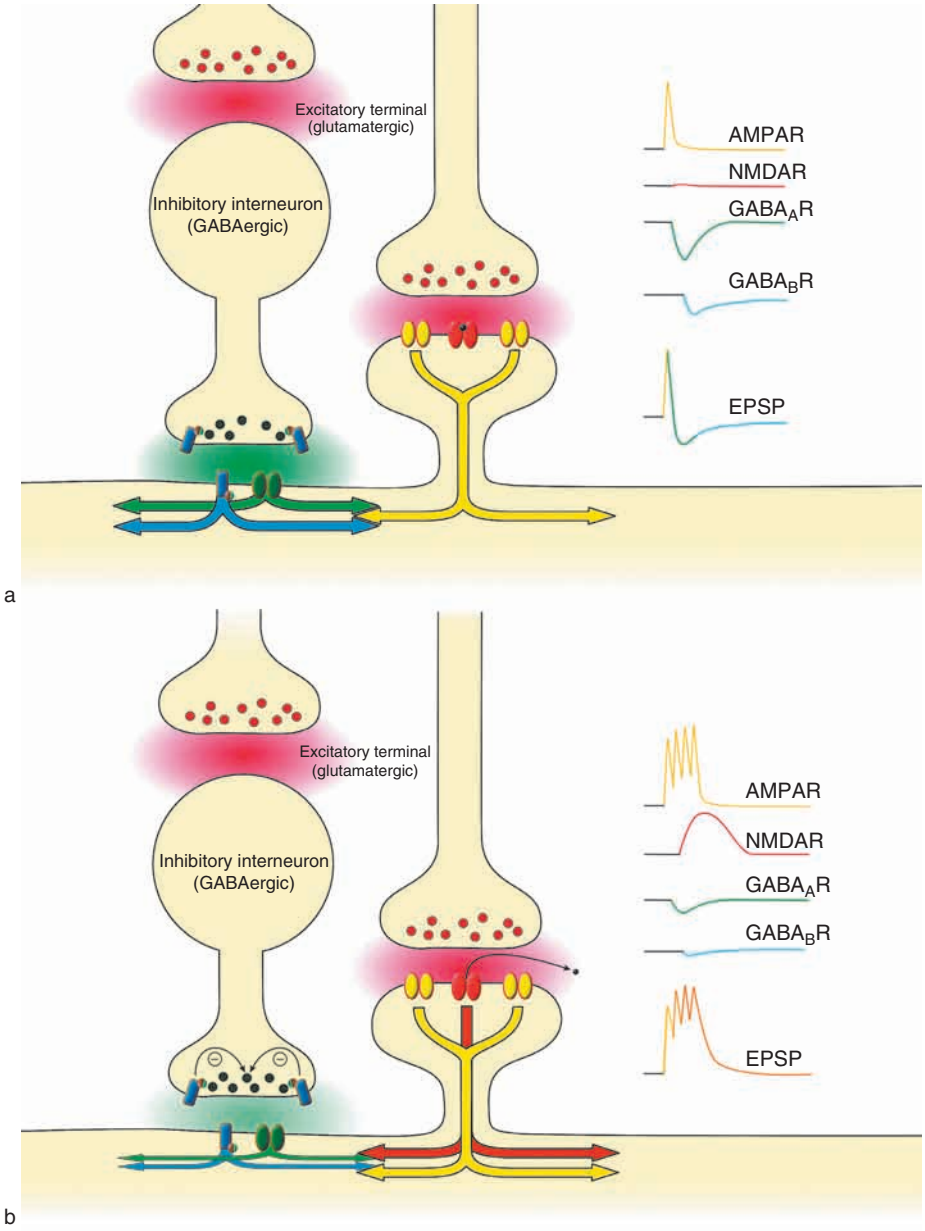

Figure 2 Role of γ-aminobutyric acid (GABA ergic) inhibition in the induction of long-term potentiation. (a) Activation of glutamatergic inputs onto CA1 hippocampal neurons is accompanied by the activation of inhibitory GABAergic interneurons. These in turn provide an inhibitory input onto the pyramidal cells. GABA release is delayed with respect to glutamate release, as there are two synapses for the signal to pass through, rather than one. This results in a rapid depolarization due to α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor (AMPAR) activation followed by a rapid inhibition due to GABA$_A$ receptor (GABA$_A$R) activation, ensuring the maintenance of the Mg^{2+} block. Sustained inhibition is then maintained by activation of postsynaptic GABA$_B$Rs. Simultaneously, activation of presynaptic GABA$_B$ autoreceptors results in a temporary inhibition of GABA release (EPSP, excitatory postsynaptic potential). (b) A burst of high-frequency stimulation while GABA release is reduced (200 ms after the priming stimulus) results in sustained postsynaptic depolarization, overcoming the weakened inhibitory inputs. This sustained depolarization results in the release of Mg^{2+} from the *N*-methyl-D-aspartate receptor (NMDAR) and the influx of Ca^{2+} ions, leading to the induction of long-term potentiation.

phosphatidylinositol 3-kinase (PI3K). Why the need for multiple kinases? It seems that different kinases may be important at different developmental stages and for different phases of LTP. A full understanding of how kinases lead to persistent alterations in synaptic strength during LTP is far from complete.

Expression Mechanisms Involved in LTP

Phases of LTP

Numerous studies have attempted to determine how AMPAR-mediated synaptic transmission is enhanced in LTP, and multiple mechanisms have been identified.

It is clear that different mechanisms occur at different stages of the LTP process. For example, the earliest phase of LTP, often referred to as short-term potentiation (STP) and most prevalent when a high-frequency tetanus is used to induce the process, probably involves an increase in L-glutamate release. The next phase, commonly referred to as early-phase LTP (E-LTP), most probably involves an alteration in AMPAR number or properties. A similar alteration in AMPAR function may also account for the next, protein synthesis-dependent phase of LTP, known as late-phase LTP (L-LTP). However, the expression mechanisms of LTP seem also to be developmentally regulated, with presynaptic mechanisms appearing to play a more important role in E-LTP within the first week of life.

Mechanisms of LTP Expression

Two ways in which AMPARs function at synapses may be directly altered during LTP have been identified. There can be an increase in the number of receptors or an increase of the efficiency of the receptors (Figure 3). In the latter case, the mechanism involves an increase in the single-channel conductance of the receptors, so that they pass more current in unit time. The extent to which this results from a modification of receptors already inserted in the synapse – for example, by phosphorylation, or by the exchange of high- for low-conductance receptors – is not yet known. One idea is that AMPARs lacking the GluR2 subunit can be transiently inserted during LTP. This has the effect of both increasing single-channel conductance and providing an activity-dependent Ca^{2+} signal, by permeation through the now Ca^{2+}-permeable AMPARs. This Ca^{2+} signal might then drive the insertion of additional Ca^{2+}-impermeable AMPARs.

A number of proteins that interact directly with AMPARs have been found to affect their targeting and stabilization at synapses, or their functional properties. These include two membrane-associated guanylate kinases (MAGUKs) – synapse-associated protein 97 (SAP97) and postsynaptic density-95 protein (PSD95), the latter of which binds to an accessory AMPAR protein known as a transmembrane AMPA receptor protein (TARP). Another protein that binds to AMPARs and may be involved in LTP is PICK1 ('protein interacting with C kinase 1'). Elucidating how these and other interacting proteins, together with small guanosine triphosphatases (GTPases) and protein kinases, function to regulate AMPARs during LTP is the focus of much current interest.

Metaplasticity

Metaplasticity is a term that is used to refer to the plasticity of synaptic plasticity. With respect to NMDAR-dependent LTP this can be manifest in several ways. For example, a stimulus protocol (e.g., continuous low-frequency stimulation) that has no effect on baseline transmission can cause LTD of synaptic transmission following the induction of LTP. When the depression is conditional on prior LTP it is referred to as depotentiation. Two forms of depotentiation have been identified; one is dependent on the activation of NMDARs and the other is dependent on activation of metabotropic glutamate receptors (mGluRs). Similarly, two forms of LTD of baseline transmission (sometimes referred to as *de novo* LTD)

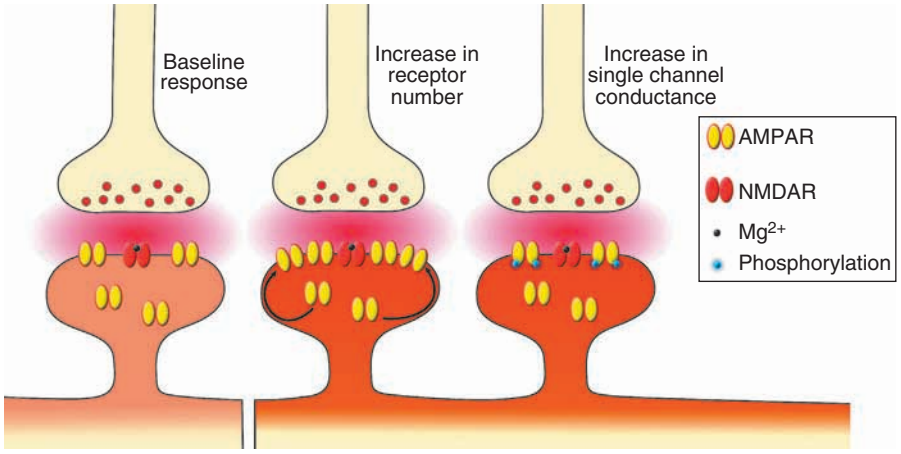

Figure 3 Expression of long-term potentiation. There are multiple mechanisms that result in the long-term increase in synaptic strength that come under the umbrella of long-term potentiation. The number of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) present in the postsynaptic membrane may be increased by exocytosis from an intracellular pool, or the activity of those AMPARs already present may be modified by, for example, phosphorylation. Additional mechanisms include the increase in glutamate release (NMDAR, *N*-methyl-D-aspartate receptor).

have been identified that are dependent on activation of NMDARs or mGluRs. The induction of LTP, while enabling depotentiation, can also inhibit the induction of NMDAR-dependent *de novo* LTD. The induction of NMDAR-dependent LTP can be affected by the prior history of synaptic activity. For example, prior activity at NMDARs can result in inhibition of subsequent LTP, while prior activity of mGluRs can result in facilitation of LTP. Finally, it should be noted that the NMDAR-mediated component of synaptic transmission is itself plastic – both LTP and LTD have been demonstrated.

See also: AMPA Receptors: Molecular Biology and Pharmacology; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Metabotropic Glutamate Receptors (mGluRs): Functions; Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology; NMDA Receptors and Development; NMDA Receptors and Disease + C464; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Depression.

Further Reading

Alford S, Frenguelli BG, Schofield JG, et al. (1993) Characterization of Ca^{2+} signals induced in hippocampal CA1 neurones by the synaptic activation of NMDA receptors. *Journal of Physiology* 469: 693.

Ault B, Evans RH, Francis AA, et al. (1980) Selective depression of excitatory amino acid induced depolarizations by magnesium ions in isolated spinal cord preparations. *Journal of Physiology* 307: 413–428.

Bliss TV and Collingridge GL (1993) A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature* 361: 31–39.

Bliss TVP, Collingridge GL, and Morris RGM (2007) Synaptic plasticity in the hippocampus. In: Andersen P, Morris RGM, Amaral DG, et al. (eds.) *The Hippocampus*, pp. 343–474. Oxford, UK: Oxford University Press.

Collingridge GL, Kehl SJ, and McLennan H (1983) Excitatory amino acids in synaptic transmission in the Schaffer collateral-commissural pathway of the rat hippocampus. *Journal of Physiology* 334: 33–46.

Collingridge GL, Isaac JT, and Wang YT (2004) Receptor trafficking and synaptic plasticity. *Nature Reviews Neuroscience* 5: 952–962.

Dale N and Roberts A (1985) Dual-component amino-acid-mediated synaptic potentials: Excitatory drive for swimming in *Xenopus* embryos. *Journal of Physiology* 363: 35–59.

Davies CH, Starkey SJ, Pozza MF, et al. (1991) GABA autoreceptors regulate the induction of LTP. *Nature* 349: 609.

Derkach VA, Oh MC, Guire ES, et al. (2007) Regulatory mechanisms of AMPA receptors in synaptic plasticity. *Nature Reviews Neuroscience* 8(2): 101–113.

Herron CE, Lester RA, Coan EJ, et al. (1986) Frequency-dependent involvement of NMDA receptors in the hippocampus: A novel synaptic mechanism. *Nature* 322: 265–268.

MacDermott AB, Mayer ML, Westbrook GL, et al. (1986) NMDA-receptor activation increases cytoplasmic calcium concentration in cultured spinal cord neurones. *Nature* 321: 519–522.

Nowak L, Bregestovski P, Ascher P, et al. (1984) Magnesium gates glutamate-activated channels in mouse central neurones. *Nature* 307: 462–465.

Relevant Website

http://www.bris.ac.uk – University of Bristol, Medical Research Council Centre for Synaptic Plasticity.

Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity

M C Crair, Yale University, New Haven, CT, USA
R D Shah, Yale University, New Haven, CT; and Baylor College of Medicine, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

It has been more than 50 years since Donald Hebb proposed that associative learning might be accomplished by increasing the strength of synaptic connections between neurons whose activity is correlated. Since the discovery of long-term potentiation (LTP) and long-term depression (LTD) more than 30 years ago, a great deal of experimental and theoretical evidence has accumulated lending support to the hypothesis that these forms of Hebbian synaptic plasticity might underlie numerous forms of experience-dependent plasticity, including learning, memory, and activity-dependent development. This article focuses on the phenomenon of experience-dependent development and plasticity of sensory systems and addresses how LTP and LTD mechanisms at excitatory synapses might be used to mediate such changes.

Mechanisms of LTP and LTD

LTP and LTD are nearly ubiquitous phenomena, having been observed at almost every excitatory synapse studied. Their mechanisms of induction and expression and their functions, however, are subjects of much debate and likely depend on which synapses, which neural circuits, and what stage of development are involved. Much of the characterization of how LTP and LTD are induced and expressed has come from studies examining excitatory synapses in hippocampal slice preparations. These synapses, like most excitatory synapses in the mammalian brain, communicate through the release of glutamate onto postsynaptic α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs), whose strength and number determine the efficacy of synaptic transmission. Although there is still debate over the precise molecules and pathways involved in synaptic plasticity, one of the most common forms depends on postsynaptic activation of the *N*-methyl-D-aspartate type of ionotropic glutamate receptors (NMDARs). NMDARs serve as molecular coincidence detectors in that they are active only when the synaptic membrane is depolarized and glutamate is bound to the receptor, thereby satisfying Hebb's rule of correlated presynaptic and postsynaptic activity. In addition, NMDARs conduct calcium ions, which can activate a variety of second messenger pathways. In the simplest models, a large calcium influx through NMDARs triggered by strong synaptic activation can activate kinases such as calcium/calmodulin-dependent protein kinase II (CaMKII) and protein kinase A (PKA). These kinases then lead to LTP either by trafficking new AMPARs to activated synapses or by acting on the biophysical properties of existing AMPARs via posttranslational modifications. Conversely, a smaller calcium influx through NMDARs triggered by weak synaptic activation will recruit phosphatases that can lead to LTD via the opposite mechanisms.

Detailed mechanistic studies of hippocampal LTP and LTD have shown how the history and pattern of synaptic activity necessary for a certain level of NMDAR activation can regulate LTP and LTD and also how modulating specific biochemical signaling pathways can influence their expression. Many of these mechanisms are known to be shared by excitatory synapses outside the hippocampus, including those in the sensory neocortex.

Sensory Maps, Critical Periods, and Experience-Dependent Plasticity

Sensory systems are typically organized in a precise topographic manner, such that neurons that respond to nearby locations in the sensory periphery are also physically located closest to each other in the brain. Examples of such topography are the retinotopic map and the map of eye preference (ocular dominance) in the primary visual cortex (V1), which simultaneously preserves visuotopic order while separating the contralateral and ipsilateral eye inputs originating from the two eyes. In a qualitatively similar manner, the somatosensory cortex of rodents contains individual 'barrels' that represent input from single facial whiskers on the snout while preserving somatotopic order across the entire whisker field. Analogously, in the auditory cortex, there is a tonotopic organization of neurons that respond best to different audible frequencies.

One of the salient features of these maps is that they are not entirely static in adults but undergo plastic changes in response to sensory manipulations and behaviorally relevant experience throughout life. In the somatosensory cortex of adult monkeys, for example, there is a topographic representation of the body surface. This somatotopic map can be manipulated in an experience-dependent manner, as amputation of

digits from the hand results in the loss of cortical responsiveness to the amputated digits, accompanied by an increase in cortical representation to the neighboring spared digits.

Early in life, at a time when the development and refinement of these maps is ongoing, plasticity is particularly dramatic during a 'critical period,' a narrow time window in which changes in sensory activity can have permanent anatomical and physiological effects. For example, in the rat auditory system, the development of a precise tonotopic map involves a sharpening of tone-evoked receptive fields such that cortical neurons progressively respond to a more and more narrow range of audible frequencies. This refinement is dependent on auditory experience as presentation of pulsed monotone frequencies during the critical period results in an overrepresentation of the pulsed frequency at the expense of other frequencies. When rats are presented with pulsed white noise during the critical period, which serves to synchronize all the auditory inputs to the cortex, there is also a severe impairment of the developmental refinement of auditory receptive fields and tonotopicity. These results suggest that the temporal pattern of activity provides an instructive signal for sensory map refinement and that activity-dependent competition between inputs is required for experience-dependent plasticity.

At the cellular level, sensory experience can also be used to shape the physiological responses of single neurons in a Hebbian manner. In another example from the auditory system, pairing tones with either depolarizing or hyperpolarizing current injected into auditory cortex neurons results in responses to tones paired with depolarized cells being significantly increased relative to responses to tones paired with hyperpolarizing current. These results support a classical Hebbian model of experience-dependent plasticity in the auditory system, and indeed, associative LTP has been observed at auditory cortical synapses *in vitro*.

A definitive causal or mechanistic link between Hebbian synaptic plasticity (LTP and LTD) and experience-dependent map plasticity is difficult to demonstrate empirically. There is, however, a set of criteria to support this causal relationship. It should be noted that even these criteria do not prove causality but serve only as the essential correlations necessary to examine further whether a causal relationship does indeed exist:

1. Synaptic plasticity (LTP and LTD) must exist at the same connections that display experience-dependent plasticity.
2. Manipulations that block synaptic plasticity should impair map plasticity.
3. Experience-dependent plasticity should induce the same cellular and molecular changes observed during synaptic plasticity.
4. Experience-dependent plasticity should occlude further synaptic plasticity.

Although map plasticity has been reasonably well characterized in the auditory cortex, the underlying synaptic or circuit mechanisms in this system remain unclear. In contrast, a more convincing link between experience-dependent plasticity and synaptic plasticity exists in the visual and somatosensory systems.

Synaptic Plasticity during Experience-Dependent Plasticity in the Visual System

One of the most informative model systems for studying experience-dependent plasticity has been the system of ocular dominance columns in V1. In monkeys, the thalamic projections representing inputs from the two eyes remain segregated in the cortex and, when visualized with anatomical techniques, form an array of alternating stripes or ocular dominance columns. In a series of classic studies, David Hubel and Torsten Wiesel discovered that in response to one eye's deprivation of visual input (monocular deprivation) during a critical period of development, the ocular dominance columns of the deprived eye shrink, while those of the spared eye expand (Figure 1). Physiologically, Hubel and Wiesel also showed that there is a functional depression of responses in cortical neurons to stimulation of the deprived eye and a commensurate increase in responses to the spared eye following monocular deprivation. It is important to note that these functional changes occur only during an early critical period of development.

These studies laid the groundwork for examining the role of neural activity in shaping functional connectivity in the developing and adult visual system. A large body of work across multiple model systems has confirmed the requirement for retinal activity during ocular dominance plasticity. In the cat visual cortex, ocular dominance plasticity is associated with distinct anatomical changes, as geniculocortical axons rapidly retract their branches in response to either brief or prolonged monocular deprivation, while an expansion of the axonal projections from the nondeprived eye occurs only after prolonged deprivation. Subsequent experiments in the mouse visual cortex showed that physiological changes involve both a rapid depression of the deprived eye response and a slower potentiation of the spared eye response, consistent with the distinct anatomical signatures of monocular deprivation found in the cat.

Mechanisms of Hebbian plasticity have been widely hypothesized to mediate this global plasticity

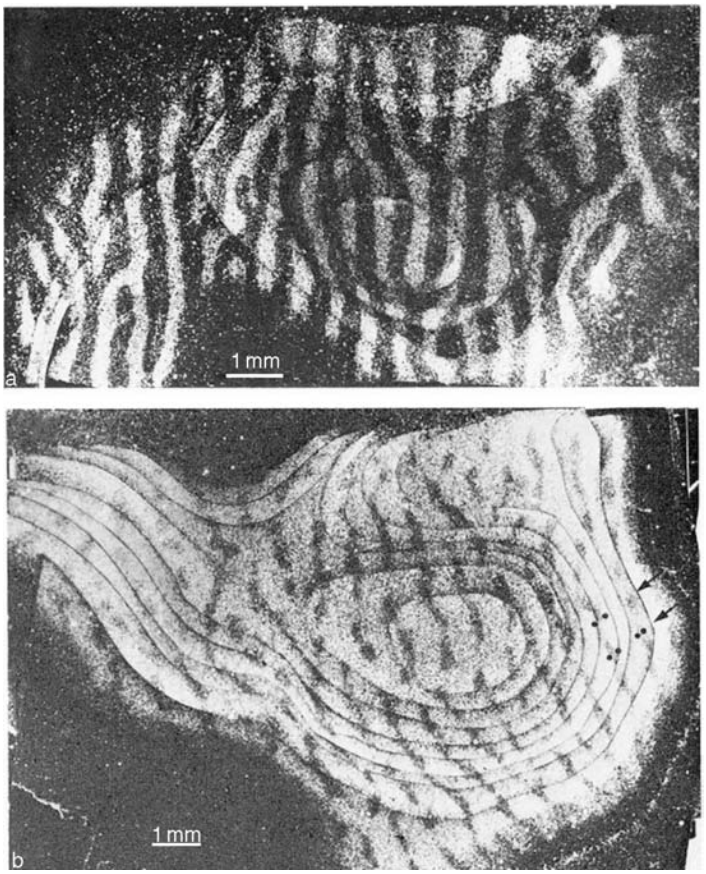

Figure 1 Ocular dominance columns and the effects of monocular deprivation in macaque monkey. Using a radioactive marker injected into one eye, sections of primary visual cortex can be subsequently visualized for the marker using a technique called autoradiography. (a) An example of such an image, with white areas corresponding to the injected eye and dark areas corresponding to the noninjected eye. In normal animals such as these, the projections from the two eyes segregate in layer 4 of the cortex into stripes, or 'ocular dominance columns,' which occupy a similar area of visual cortex. (b) An example from a monocularly deprived animal, in which the marker was injected into the spared, open eye. As a result of deprivation, the area occupied by the spared eye increases, while the area devoted to the deprived eye shrinks dramatically. The dots represent the position of six lesions made through an electrode during electrical recordings from this animal. The arrows show the location of the two penetrations made with the electrodes. Reproduced from Hubel DH and Wiesel TN (1977) Ferrier lecture: Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London, Series B: Containing Papers of a Biological Character* 198: 1–59, with permission from The Royal Society.

since they provide a natural mechanism to mediate competition among inputs. Modeling studies support this hypothesis by showing how a greater amount of coincident activity between spared inputs and cortical neurons can result in a greater functional representation of spared inputs relative to deprived inputs via a Hebbian synaptic learning rule (Figure 2).

During the past 20 years, a great deal of experimental evidence has accumulated to support these theories. For example, binocular deprivation during the critical period does not induce ocular dominance plasticity, and reversing a preexisting deprivation produces functional and anatomical recovery. In addition, alternating monocular deprivation between the two eyes results in a decrease in the number of binocularly responsive neurons. Finally, when cortical activity is suppressed pharmacologically, not only is ocular dominance plasticity prevented in response to monocular deprivation, but there is actually a shift in favor of deprived-eye responses when measured physiologically and anatomically. This 'reverse plasticity' results in a greater representation of deprived eye columns over spared eye columns. These results confirm the role of coincident pre- and postsynaptic activity during experience-dependent plasticity, as predicted by Hebbian learning rules, and also strongly suggest a role for competitive mechanisms of synaptic plasticity between inputs from the two eyes as a substrate for experience-dependent plasticity (Figure 3).

Direct tests of Hebbian mechanisms in V1 come from studies recording intracellularly from cortical neurons *in vivo*. Pairing depolarizing, excitatory current injections into cortical neurons with visual stimuli from one eye (positive pairing) results in a long-lasting increase in visual response, while pairing hyperpolarizing current with visual stimuli from the

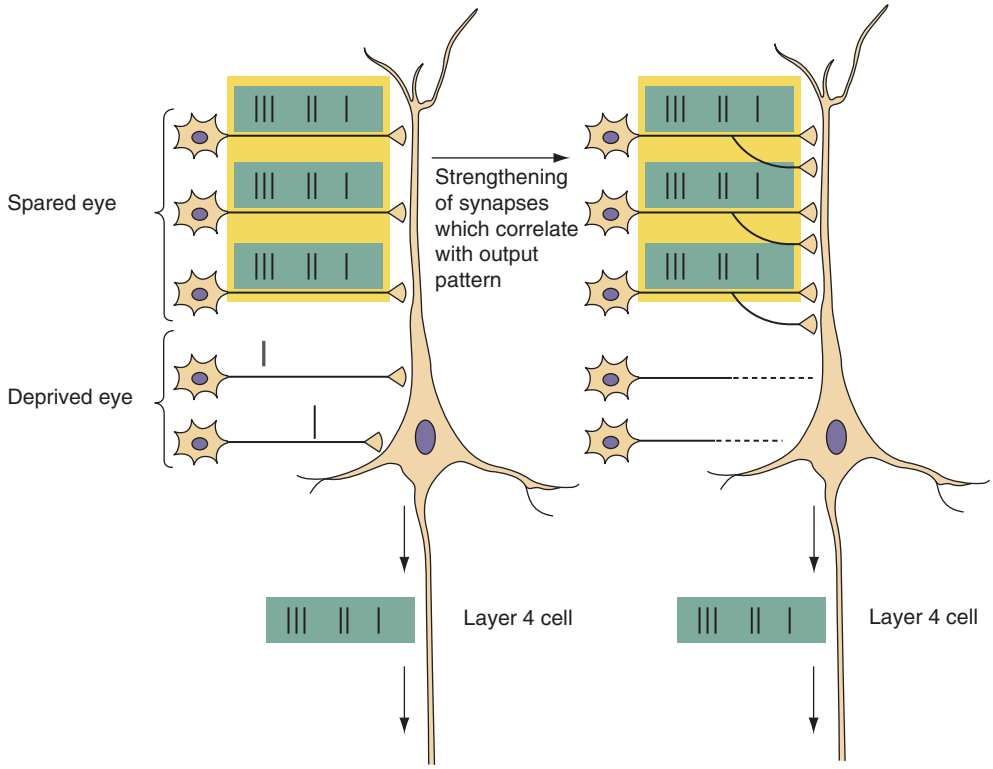

Figure 2 Diagram depicting how Hebbian learning rules could underlie ocular dominance plasticity. The cells represent groups of layer 4 neurons in the visual cortex, which receive thalamic input representing each eye. During monocular deprivation, the inputs from the spared eye have much greater activity than do those from the deprived eye and are therefore better able to drive activity in the postsynaptic neurons. Based on mechanisms of Hebbian synaptic plasticity, these activity correlations would result in the strengthening of synapses associated with the spared eye. The inputs from the deprived eye, meanwhile, have greatly reduced activity and are therefore unable to influence postsynaptic activity. These synapses are therefore gradually weakened and eventually eliminated. Reproduced from Purves DA, Augustine GJ, Fitzpatrick D, Katz LC, LaMantia AS, and McNamara JO (eds.) (1997) *Neuroscience*. Sunderland, MA: Sinauer, with permission from Sinauer Associates, Inc.

other eye (negative pairing) results in a long-lasting decrease in visual response. Similar reinforcement studies were used to test plasticity of orientation preference, and positive pairings consistently resulted in a potentiation of responses to the paired orientation relative to negative pairings. These results were most robust during the critical period for experience-dependent plasticity, suggesting a correlation between the cellular mechanisms of response potentiation and map plasticity.

More direct evidence of a role for synaptic plasticity during map plasticity comes from *in vitro* experiments in slices of rodent visual cortex. In accordance with the stipulations described above, both LTP and LTD of visual cortical neurons were observed during the critical periods. Both high-frequency stimulation, which should provide a robust postsynaptic calcium profile, and pairing presynaptic stimulation with postsynaptic current injection reliably induced LTP, satisfying Hebb's rule for associative potentiation. It is interesting that the capacity for synaptic plasticity is greatest between postnatal days (P) 14 and 21 and subsequently decreases with age, therefore demonstrating a critical period similar to that of experience-dependent map plasticity (Figure 4(c)). This synaptic plasticity is also regulated by visual experience, as dark-reared animals show increased LTP and reduced LTD (Figure 4(d)). These results suggest that synaptic plasticity mechanisms themselves are regulated by sensory experience through a 'sliding threshold' for synaptic plasticity, or a 'metaplasticity' that could underlie activity-dependent modifications in the visual system. This 'Bienenstock, Cooper, and Munro theory' of synaptic plasticity proposes that the history of activity at any given synapse can regulate the capacity for further plasticity at that synapse. Synapses would have a variable modification threshold, and if their responses are previously potentiated, then their capacity for further LTP is decreased, and vice versa. These theories also predict that sensory experience can regulate plasticity in a bidirectional manner, and therefore that homosynaptic LTD might underlie the loss of visual responsiveness due to monocular deprivation. Homosynaptic LTD is in fact observed at visual

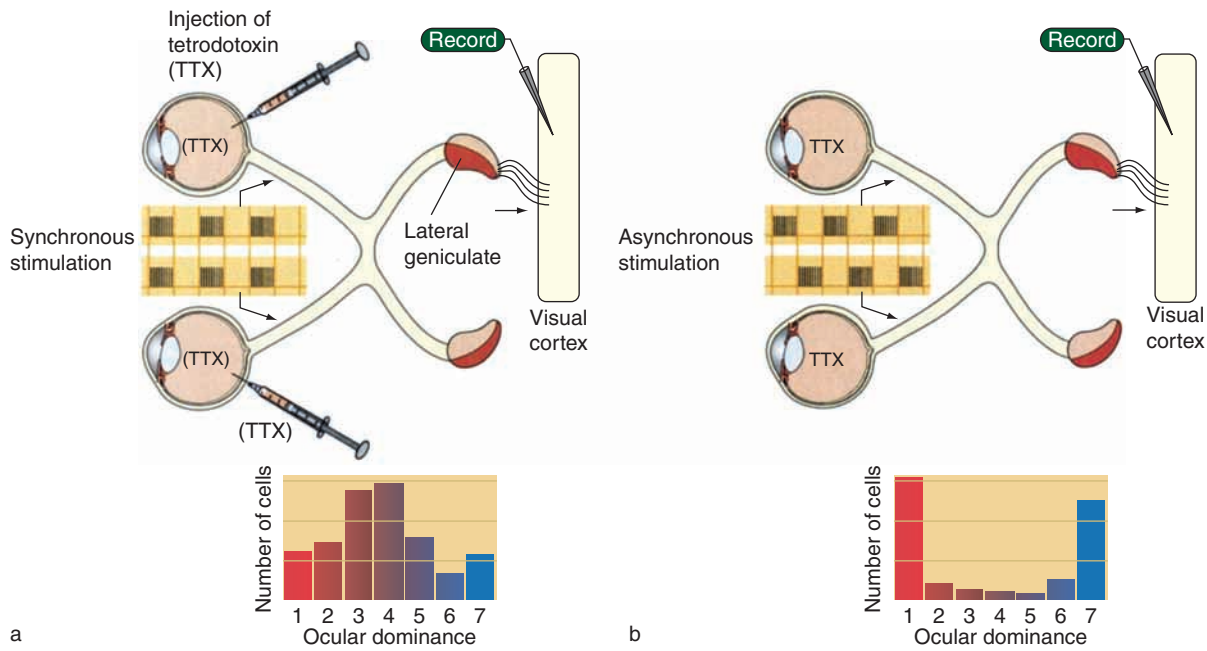

Figure 3 Competition between afferents serving the two eyes drives ocular dominance. In these experiments, all activity in the eyes was blocked by chronic injections of tetrodotoxin, a sodium channel blocker that prevents action potential firing. Recordings were subsequently made in the visual cortex to determine the ocular dominance index of each cell, in which a score of 1 indicates that the cell responds exclusively to stimulation of the ipsilateral eye and a score of 7 indicates response only to stimulation of the contralateral eye. (a) The activity patterns of each eye were synchronized by direct electrical stimulation of the optic nerves. The result of this manipulation is that most visual cortical neurons come to respond to stimulation of either eye. (b) In contrast, asynchronous stimulation was applied to the nerves. Under these conditions, the ocular dominance profile is highly segregated, indicating that most cortical neurons respond only to stimulation of one or the other eye. As described in the article, alternating monocular deprivation results in an ocular dominance profile similar to that shown in (b). These results support the hypothesis that activity-dependent competition between the thalamic inputs of each eye is responsible for ocular dominance plasticity in the visual cortex. Reproduced from Purves DA, Augustine GJ, Fitzpatrick D, Katz LC, LaMantia AS, and McNamara JO (eds.) (1997) *Neuroscience*. Sunderland, MA: Sinauer, with permission from Sinauer Associates, Inc.

cortical synapses *in vitro*, and monocular deprivation serves to induce homosynaptic LTD of deprived eye responses *in vivo*. Furthermore, this depression is not observed after monocular inactivation due to tetrodotoxin injections, confirming a necessity for presynaptic activity for experience-dependent LTD and strengthening a metaplastic model for visual cortical plasticity.

If synaptic plasticity mechanisms mediate experience-dependent plasticity, manipulations that block it should also block map plasticity. Numerous studies have shown that ocular dominance plasticity is blocked by applying NMDAR antagonists to the visual cortex, a manipulation known to prevent LTP *in vitro*. Repeated presentation of a single orientation also results in the selective potentiation of responses to that orientation in mouse and cat V1. This stimulus-selective response potentiation is blocked by both NMDAR antagonists and the injection of peptides that interfere with AMPAR trafficking, both of which are known to be required for cortical LTP.

The molecular and functional changes that are associated with homosynaptic LTD are also observed after monocular deprivation. Decreases in AMPAR subunit phosphorylation and AMPAR surface expression are observed after deprivation, in addition to a decrease in synaptic transmission. Prior monocular deprivation also occludes subsequent induction of LTD, offering further proof that the two are mechanistically related. In combination, these results from the visual system satisfy all the stipulations outlined above for consistency with a causal relationship between Hebbian LTP, LTD, and experience-dependent plasticity.

Synaptic Plasticity during Experience-Dependent Map Plasticity in the Somatosensory Cortex

Even more compelling evidence for a role of LTP and LTD in experience-dependent plasticity comes from studies in the barrel cortex of the rodent somatosensory system. Barrels are clusters of cells located in layer 4 (L4) of the cortex that receive topographic input from individual vibrissae located on the snout. This input is relayed through the ventrobasal thalamus. Excitatory neurons in each barrel then project up to L2/3 neurons in the same column, maintaining

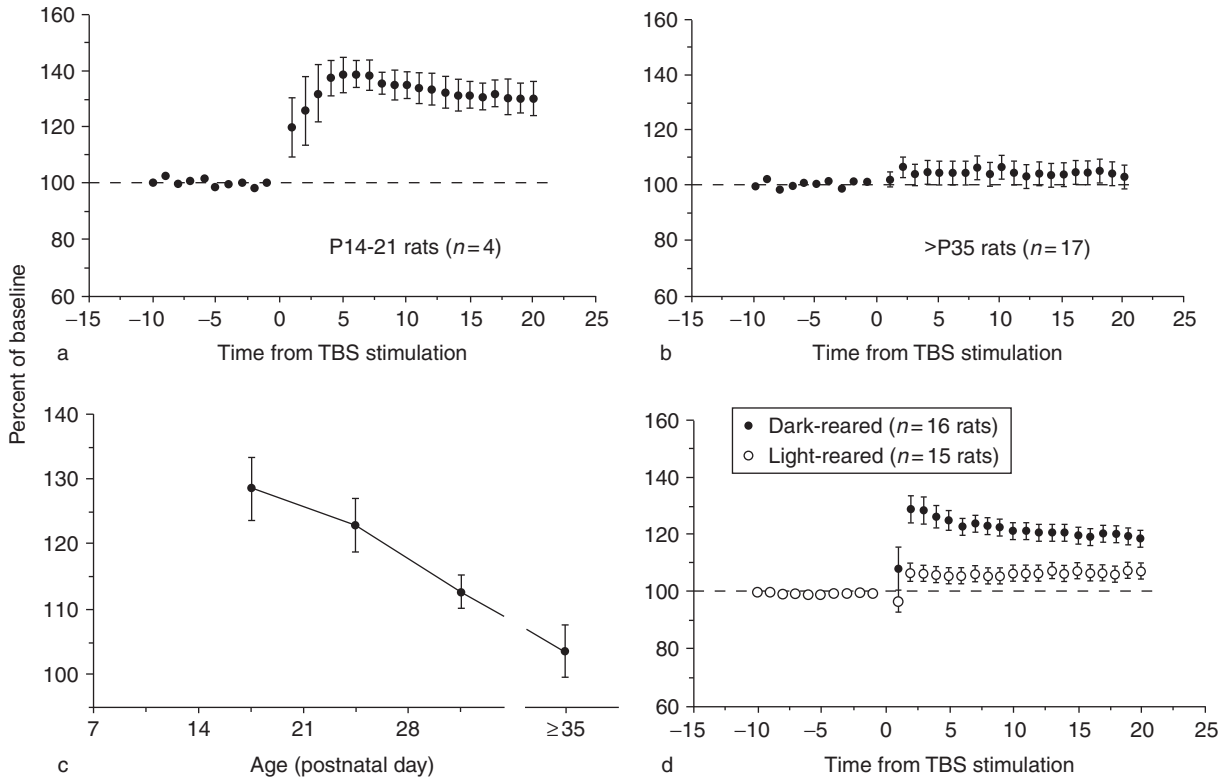

Figure 4 Synaptic plasticity in the visual cortex is regulated by age and sensory experience. The data shown here depict the effect of theta-burst stimulation (TBS) in white matter on the amplitude of extracellularly recorded field potentials in L2/3 of rat visual cortex slices. (a) TBS stimulation induces a robust LTP during the critical period for visual cortical plasticity. (b) After the critical period, however, TBS fails to induce LTP. (c) The developmental regulation of TBS-induced LTP in visual cortical slices. (d) In addition to age, LTP is regulated by sensory experience. In dark-reared animals, the capacity for LTP is much greater than in normal, light-reared animals. Field potential amplitudes were normalized to the time from TBS onset, and each time point was averaged across animals and plotted as the mean ± SEM. These results suggest that visual experience might induce LTP *in vivo*, and therefore that the induction of further LTP is occluded by a 'sliding threshold' model (see the section titled 'Synaptic plasticity during experience-dependent plasticity in the visual system' for explanation). Reprinted by permission from Macmillan Publishers Ltd: [Nature] Kirkwood A, Lee HK, and Bear MF (1995) Co-regulation of long-term potentiation and experience-dependent synaptic plasticity in visual cortex by age and experience. *Nature* 375: 328–331, copyright (1995).

strong feed-forward excitation from the 'principal whisker.' Experience-dependent plasticity is robust in this system and has been observed at multiple points along the sensory pathway. Cortical organization at the level of the thalamocortical barrel map is particularly sensitive to early experience, as manipulations in whisker activity can induce rapid and long-lasting anatomical map plasticity. In response to whisker trimming or deprivation, the cortical area devoted to the deprived whisker will shrink, and the areas representing neighboring whiskers will expand into the deprived territory (Figure 5(b)). In L4 of rodents, this plasticity obeys a strict critical period during development as barrel map cytoarchitecture remains nearly static after the first postnatal week (Figure 5(c)). Physiologically as well, responses to deprived whiskers are reduced in the corresponding barrel, while thalamocortical input from neighboring spared whiskers is functionally increased. Outside L4 in the cortex (L2/3, 5, and 6), anatomical and functional plasticity decreases with age, but significant residual plasticity remains in the adult.

LTP- and LTD-like phenomena are likely to contribute at least in part to this experience-dependent plasticity. After whisker deprivation, the resulting anatomical rearrangements in barrels are prevented with selective administration of glutamate receptor antagonists to the cortex. Functionally as well, cortical activity blockade serves to prevent the potentiation of spared whisker responses and the depression of deprived whisker responses. These results support a model in which postsynaptic activity is required for map plasticity, consistent with Hebbian synaptic plasticity.

The initial development of the barrel map is also thought to rely on activity-dependent mechanisms like LTP and LTD. Coinciding with the critical period for L4 map plasticity, there is a critical period for synaptic plasticity of thalamocortical afferents to L4; not only does thalamocortical synaptic plasticity exist during the periods of map formation, but the

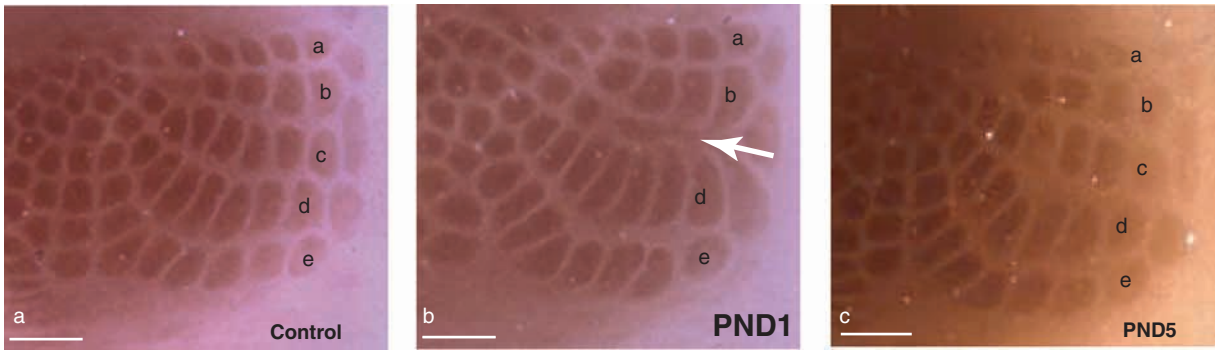

Figure 5 There is a critical period for barrel map plasticity in the rodent somatosensory cortex. Staining for cytochrome oxidase (CO) activity in thin brain sections can reveal the pattern of thalamic afferent clustering into barrels in the somatosensory cortex of rodents. In response to cauterization, or removal, of the entire 'c' row of whiskers from one side of the snout early in development, the contralateral hemisphere shows robust map plasticity. This involves both a shrinkage of area devoted to the deprived whiskers and an expansion of the neighboring barrels, as shown in (b). (a) is the control, ipsilateral hemisphere. This map plasticity has a strict critical period, and if cauterization is performed at postnatal day (PND) 5, no thalamic rearrangements are observed, as shown in (c). Scale bar = 500 μm.

ability to induce LTP decreases during the same period that the ability to induce experience-dependent plasticity decreases (Figure 6). Moreover, blocking or disrupting NMDAR activity prevents the proper formation of the barrel map, fulfilling the first two stipulations described above.

Perhaps the most direct evidence for the involvement of LTP in barrel cortex plasticity comes from experiments showing that whisker experience can drive AMPARs into synapses, a process that mimics the increase in synaptic strength that occurs during LTP *in vitro*. Furthermore, manipulations that block AMPAR trafficking into synapses also prevent the synaptic potentiation that is normally observed following experience-dependent plasticity. These observations provide direct support for a causal relationship between LTP and experience-dependent plasticity in the somatosensory barrel map system.

Since experience-dependent plasticity in the barrel cortex involves both a selective increase in response to salient inputs and a decrease in response to deprived or unused inputs, the expression of thalamocortical LTD would also theoretically be necessary. Indeed, NMDAR-dependent LTD has been observed at thalamocortical synapses in L4, and it is important to note that this LTD is also developmentally regulated. The magnitude of thalamocortical LTD is greatest during the critical period for map plasticity, suggesting a mechanistic link.

Layer 2/3 of barrel cortex is also susceptible to map plasticity in response to whisker deprivation, and these same sensory manipulations induce LTD of layer 2/3 synapses *in vitro*. In addition, this whisker deprivation-induced LTD occludes the induction of further LTD *in vitro* when compared with nondeprived synapses. These results suggest that map plasticity shares the same cellular mechanisms as synaptic plasticity and strengthens the causal link between the two phenomena.

Spike Timing-Dependent Plasticity during Experience-Dependent Plasticity

In addition to the learning rules already described, another form of Hebbian synaptic plasticity that has received a great deal of attention over the past two decades depends on the temporal order of activity between pre- and postsynaptic cells. Referred to as spike timing-dependent plasticity (STDP), this learning rule has proved to be an effective means of inducing experience-dependent plasticity in neural circuits. The typical STDP rule predicts that presynaptic activity preceding postsynaptic spiking within approximately 10 ms will lead to synaptic strengthening (LTP), while the reverse spiking order will lead to weakening (LTD). Initially described in cerebellar-like structures, STDP has been observed in numerous brain areas across many different species, underscoring its potential importance in mediating normal development and learning.

STDP learning rules are naturally associated with experience-dependent plasticity and have been implicated in sensory map plasticity in many different systems. In the barrel cortex, STDP is observed at L4 to L2/3 synapses *in vitro*, and as described above, whisker deprivation produces a long-lasting depression at these same synapses. It is interesting that L4 cells consistently fire before L2/3 cells in normal animals when whisker-evoked spiking responses are measured *in vivo*. After whisker deprivation, however, this firing order is reversed, suggesting that a spike timing-dependent learning rule mediates synaptic depression at this synapse.

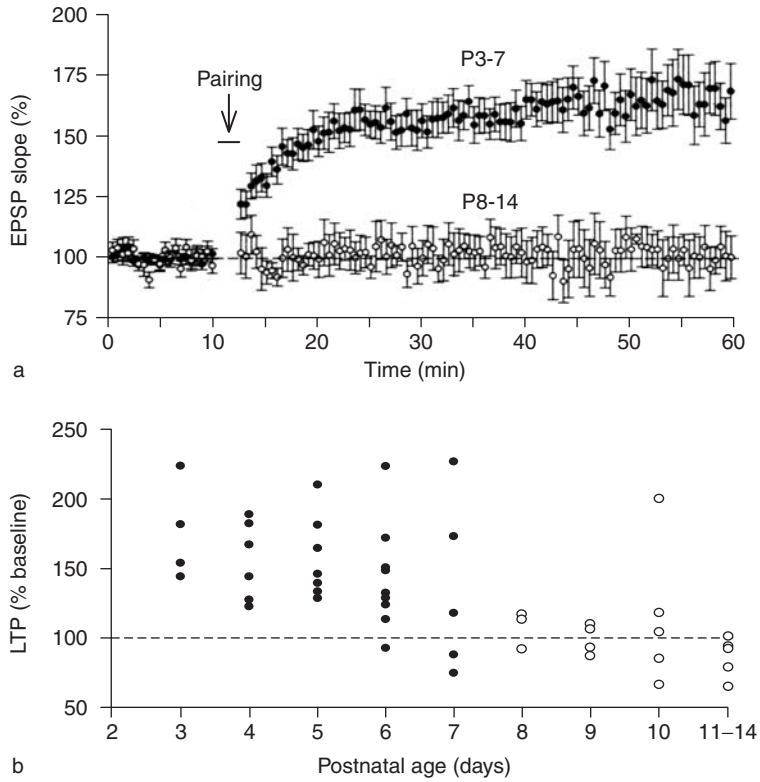

Figure 6 Thalamocortical long-term potentiation (LTP) has a critical period similar to that of barrel map plasticity. Date collected from P 3-7 animals are shown as black dots, and date from p 8-14 animals are shown as white dots. (a) LTP and long-term depression (LTD) at thalamocortical synapses in the barrel cortex are easily induced during the first week after birth. Here, LTP in an *in vitro* brain slice preparation is produced by pairing postsynaptic depolarization with presynaptic stimulation, conditions that mimic Hebbian plasticity. The same pairing protocol is incapable of inducing plasticity in animals more than 1 week old. (b) The amount of LTP decreases gradually with age over the first postnatal week at thalamocortical synapses in rodent barrel cortex, paralleling the decline in *in vivo* barrel map plasticity. EPSP slopes were normalized to the time from pairing onset, and each time point was averaged across animals and plotted as the mean ± SEM. Date collected from P3-7 animals are shown as black dots, and data from P8-14 animals are shown as white dots. Reprinted by permission from Macmillan Publishers: Nature Crair MC and Malenka RC (1995) A critical period for long-term potentiation at thalamocortical synapses. *Nature* 375: 325–328, copyright (1995).

In the visual system, STDP has been implicated in several different experience-dependent modifications. Receptive fields of neurons in the visual cortex of the rat, for example, can be modified by pairing visual stimulation with spiking in the cortical neuron. If the visual stimulation precedes the spiking, the response is enhanced to stimulation from that area, and the converse is observed when the order of activation is reversed. In the developing tadpole optic tectum, STDP learning rules serve to rapidly shift the visual receptive fields either toward or away from any given area in the visual field, depending on the relative timing of postsynaptic spiking.

The development of direction sensitivity in the developing tadpole tectum has also been attributed to a form of STDP learning rule. By presenting bars at specific speeds, tectal neurons can be trained to become selective for the direction of movement when allowed to spike. Further studies confirm that the speeds that are effective for this induction produce temporal patterns of synaptic activation that are consistent with a spike-timing learning rule and indeed are able to induce the required synaptic modifications.

Finally, the development of orientation selectivity in the visual system is also modulated by sensory experience and regulated by an STDP learning rule. For example, orientation maps in the visual cortex of cats rapidly degrade during the critical period when deprived of visual experience. If the animal is reared under conditions that limit its exposure to just one orientation during the critical period, however, the cortex comes to respond preferentially to that orientation and not to others. At the same time, pairing visual stimuli of specific orientations with spikes in cortical neurons can lead to a shift in orientation preference toward the paired orientation if the stimuli preceded the spike. Consistent with an STDP learning rule, the response to the presented orientation is diminished when it is presented after the spiking. Together, these results highlight the importance of

spike timing-dependent synaptic plasticity in mediating the experience-dependent plasticity of sensory systems, during both development of cortical maps and adult learning.

Synaptic Homeostasis during Experience-Dependent Plasticity

It is now clear that neural activity can play an instructive role in regulating the functional connectivity of neural circuits through various mechanisms of LTP and LTD induction. These Hebbian changes, however, are inherently unstable and can lead to the runaway excitation or depression of a subset of synapses when left unchecked. For example, if LTP is based on the ability of a presynaptic neuron to fire a postsynaptic neuron effectively, then the resulting potentiation will subsequently increase the capacity of that connection to induce firing again in the future, thereby resulting in even greater LTP. This cycle will continue again and again and eventually can cause a destabilizing effect on the circuit. For this reason, Hebb-based models for experience-dependent development and plasticity of neural circuits typically require some type of homeostatic rules to constrain synaptic strength within certain physiological limits.

In the last decade, a number of elegant studies have been performed to investigate homeostatic mechanisms that can permit selective changes at appropriate synapses without degrading the function of the entire neural circuit. In cortical and hippocampal cultures, for example, individual neurons dynamically adjust the strength of all their excitatory synapses during different levels of input activity in order to maintain a target firing rate. In response to a global reduction in afferent activity, excitatory synapses will increase their strength, whereas a global increase in activity will induce an overall reduction in synaptic strength. These results suggest there are cell-autonomous mechanisms that not only can detect the overall level of activity in individual neurons but that are also capable of modulating the strength of all the synapses onto that neuron when firing rates fall out of specific target limits.

One of the mechanisms that may mediate this type of homeostasis is termed 'synaptic scaling.' In synaptic scaling, the distribution of amplitudes of excitatory currents at all the synapses onto a neuron (the miniature excitatory postsynaptic current (mEPSC)) increases in response to reduced activity and decreases in response to increased activity. These distributions are scaled up or down in a proportional manner, thereby preserving the relative differences in synaptic weights among all the synapses onto any given neuron. This synaptic scaling allows for competition among inputs that undergo Hebbian synaptic plasticity during experience-dependent modifications. If a subset of inputs has been recruited for potentiation, then the overall level of activity will increase in that neuron, and scaling mechanisms will result in a proportional decrease in strength of all synapses. In this case, the weakest inputs will be selectively eliminated, and the strongest (and presumably most salient) inputs will be selectively maintained.

It is worth noting that synaptic scaling has been observed in the developing visual system. As described above, monocular deprivation results in a number of physiological and anatomical changes in the primary visual cortex. In the rodent, 2 days of monocular deprivation results in an increase in the amplitude of mEPSCs onto pyramidal neurons in the deprived hemisphere relative to the control hemisphere. These changes are consistent with the observations of synaptic scaling described in culture and are reversed with subsequent visual experiences. This suggests that synaptic scaling mechanisms are recruited by changes in sensory experience, and that synaptic homeostasis, in addition to Hebbian LTP and LTD, is important for regulating the experience-dependent plasticity and development of neural circuits.

Alternative Mechanisms of Experience-Dependent Plasticity

Although the evidence presented above is compelling for a role of Hebbian LTP and LTD in various forms of experience-dependent plasticity, the criteria listed in the section titled 'Sensory maps, critical periods, and experience-dependent plasticity' are by nature correlative and do not prove causality. Indeed, there is evidence that suggests manipulations that block LTP and LTD do not always interfere with experience-dependent plasticity. For example, a number of signaling molecules downstream of NMDAR activation are required for *in vitro* LTP, including PKA and CaMKII. In slices of visual cortex, genetically disrupting PKA signaling prevents the expression of pairing-induced LTP and LTD in some cortical neurons. These same genetic manipulations, however, do not prevent the emergence of ocular dominance plasticity in response to monocular deprivation, suggesting a dissociation between experimentally observed forms of homosynaptic plasticity and *in vivo* circuit changes. This dissociation can be explained on a number of levels, including cell-autonomous molecular and genetic compensation in disrupted cortical neurons, or circuit-level compensation by separate neural circuits whose capacity for plasticity remains intact.

What alternative mechanisms exist that could underlie the changes induced by sensory experience?

Instead of long-term changes in synaptic efficacy, changes in short-term plasticity and short-term synaptic dynamics could play a role. Indeed, it might be more efficient for neural circuits to take advantage of subtle and cumulative changes to adapt their functions rather than employing energetically costly forms of large changes in synaptic efficacy.

In addition, the balance of excitation and inhibition could be regulated to mediate the changes observed during experience-dependent plasticity. The critical period for ocular dominance plasticity, for example, has been shown to depend on the onset of inhibition within the primary visual cortex. In mice that have been genetically modified to lack functional γ-aminobutyric acid (GABA) release, ocular dominance plasticity is prevented but can be rescued by pharmacologically restoring GABA function. It is interesting that this rescue can occur at any age, suggesting that the emergence of inhibitory circuits is an important determining factor in critical period onset.

Concluding Remarks

Over the past 30 years, a great deal of experimental evidence has emerged supporting a role for Hebbian LTP and LTD in mediating a large range of functions, including activity-dependent development, sensory map plasticity, and learning and memory. A combination of *in vivo* and *in vitro* studies from the visual and somatosensory systems supports the hypothesis that mechanisms of synaptic plasticity contribute at least in part to various types of experience-dependent plasticity. These studies have shown that sensory experience can both induce and occlude synaptic plasticity, and those manipulations that block synaptic plasticity *in vitro* can also block experience-dependent plasticity *in vivo*. Although no direct causal link has emerged, and may never emerge, a continued examination of the mechanisms responsible for synaptic learning rules and synaptic homeostasis during the formation and plasticity of neural circuits will prove to be an effective means of examining how our brains are capable of adapting to changes in the myriad forms of information we experience during our lifetimes.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Hebbian Plasticity; Long-Term Depression: Cerebellum; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Depression (LTD): Endocannabinoids and Cerebellar LTD; Long-Term Potentiation (LTP); Spike-Timing Dependent Plasticity (STDP); Spike-Timing-Dependent Plasticity Models; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Allen CB, Celikel T, and Feldman DE (2003) Long-term depression induced by sensory deprivation during cortical map plasticity in vivo. *Nature Neuroscience* 6: 291–299.

Buonomano DV and Merzenich MM (1998) Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* 21: 149–186.

Crair MC and Malenka RC (1995) A critical period for long-term potentiation at thalamocortical synapses. *Nature* 375: 325–328.

Dan Y and Poo MM (2006) Spike timing-dependent plasticity: From synapse to perception. *Physiological Reviews* 86: 1033–1048.

Feldman DE and Brecht M (2005) Map plasticity in somatosensory cortex. *Science* 310: 810–815.

Fox K and Wong RO (2005) A comparison of experience-dependent plasticity in the visual and somatosensory systems. *Neuron* 48: 465–477.

Frenkel MY, Sawtell NB, Diogo AC, Yoon B, Neve RL, and Bear MF (2006) Instructive effect of visual experience in mouse visual cortex. *Neuron* 51: 339–349.

Hensch TK (2005) Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience* 6: 877–888.

Heynen AJ, Yoon BJ, Liu CH, Chung HJ, Huganir RL, and Bear MF (2003) Molecular mechanism for loss of visual cortical responsiveness following brief monocular deprivation. *Nature Neuroscience* 6: 854–862.

Hubel DH and Wiesel TN (1977) Ferrier lecture: Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London, Series B: Containing Papers of a Biological Character* 198: 1–59.

Kirkwood A, Lee HK, and Bear MF (1995) Co-regulation of long-term potentiation and experience-dependent synaptic plasticity in visual cortex by age and experience. *Nature* 375: 328–331.

Malenka RC and Bear MF (2004) LTP and LTD: An embarrassment of riches. *Neuron* 44: 5–21.

Purves DA, Augustine GJ, Fitzpatrick D, Katz LC, LaMantia AS, and McNamara JO (eds.) (1997) *Neuroscience*. Sunderland, MA: Sinauer.

Takahashi T, Svoboda K, and Malinow R (2003) Experience strengthening transmission by driving AMPA receptors into synapses. *Science* 299: 1585–1588.

Turrigiano GG and Nelson SB (2004) Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience* 5: 97–107.

Lysosomal System

G M Pastores, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The lysosomal storage diseases (LSDs) are a clinically heterogeneous group of inherited disorders associated with the accumulation of incompletely degraded substrates within membrane-bound bodies (the lysosome), usually due to deficiency in the activity of a specific enzyme/hydrolase (Figure 1). The clinical manifestations associated with distinct LSDs reflect the cellular sites of substrate storage and resultant organ dysfunction. In contrast to peroxisomal disorders, which are often associated with severe developmental malformations in the early-onset forms, the LSDs involve tissues and organs that develop normally but subsequently malfunction. Table 1 provides a summary of the features that may be helpful in distinguishing the lysosomal storage group of disorders from defects involving other cellular organelles, such as the peroxisome or mitochondria.

In contrast to defects of small molecules that involve amino acids, organic acids, or simple sugars (e.g., phenylketonuria, methylmalonic acidemia, and galactosemia, respectively), analysis of metabolites that accumulate in patients with LSD is not practical or readily available, except for examination of the pattern of urinary mucopolysaccharide (glycosaminoglycans)/oligosaccharide excretion in the mucopolysaccharidoses (MPSs) or oligosaccharidoses, respectively. Even in the latter cases, it is necessary to obtain biochemical confirmation of the diagnosis in every patient for accurate prognostication, management, and genetic counseling.

Except for Fabry disease, Hunter syndrome (MPS-II), and Danon disease, which are inherited in an X-linked recessive fashion, the LSDs are transmitted as autosomal recessive traits. Prenatal diagnosis is possible for almost all LSDs, based on analysis of enzymatic activity, molecular genetic testing, and/or ultrastructural examination for inclusion bodies in cells from cultured chorionic villi or amniocytes.

The Lysosome and Its Contents

The lysosome is a membrane-bound intracytoplasmic vacuole that contains several enzymes required for the degradation of complex lipids, protein, and nucleotides. It is an organelle that maintains an acidic milieu (pH $\sim$ 5.4), necessary for the optimal activity of the enclosed enzymes and their cofactors/activators. There are over 40 different lysosomal diseases (with a combined incidence of $\sim$1:5000 births), classified according to the metabolites that are stored within the various organs/tissues involved (Table 2) or, more recently, by the nature of the molecular defect (see later, section 'Pathophysiology and diagnosis').

The trafficking of most lysosomal hydrolases is mediated by a mannose-6-phosphate (M-6-P)-dependent pathway, in which asparagine-linked oligosaccharides on newly synthesized lysosomal hydrolases are modified in a two-step process to contain a terminal M-6-P moiety. The initial step in posttranslational M-6-P modification is catalyzed by the enzyme uridine diphosphate (UDP)-*N*-acetylglucosamine-1-phosphotransferase, and involves the transfer of *N*-acetylglucosamine (GlcNAc)-1-phosphate from UDP-GlcNAc to specific 1,2-linked mannoses on lysosomal hydrolases. The second step involves the enzyme *N* -acetylglucosamine-1-phosphodiester-*N*-acetylglucosaminidase, which facilitates the removal of the covering GlcNAc and generates the terminal M-6-P. Lysosomal enzymes bearing the M-6-P modification then bind to one of two M-6-P-receptors in the *trans*-Golgi and are transferred to the lysosome. Failure to make these modifications results in the enzymes entering the secretory pathway followed by inappropriate secretion from the cells, and the accumulation of various incompletely metabolized substances within the lysosome as a consequence of the intracellular deficiencies.

The enzyme *N*-acetylglucosamine-1-phosphotransferase has been found to be composed of three subunits ($\alpha_2\beta_2\gamma_2$) coded for by two different genes (α and β by one gene on chromosome 4q21-q23, and γ by another gene on chromosome 16p). Mucolipidosis types II (I-cell disease) and IIIA (pseudo-Hurler polydystrophy) are due to allelic defects in the $\alpha\beta$ gene, whereas mucolipidosis IIIC results from a defect specifically in the gene coding for the γ subunit.

Patients with I-cell (Leroy) disease present in early infancy with coarse facial features, and develop severe psychomotor retardation and joint contractures. Pseudo-Hurler polydystrophy can be distinguished from the 'classic' Hurler syndrome (α-L-iduronidase deficiency) because of its slower progression, the absence of organomegaly, and absence of mucopolysaccharidurie. Typical clinical findings include stiffness of the hands and shoulders, claw-hand deformity, scoliosis, short stature, coarse facies, and mild mental retardation. Radiographically, severe dysostosis multiplex of the hip is characteristic and frequently disabling. The clinical diagnosis of I-cell disease and pseudo-Hurler

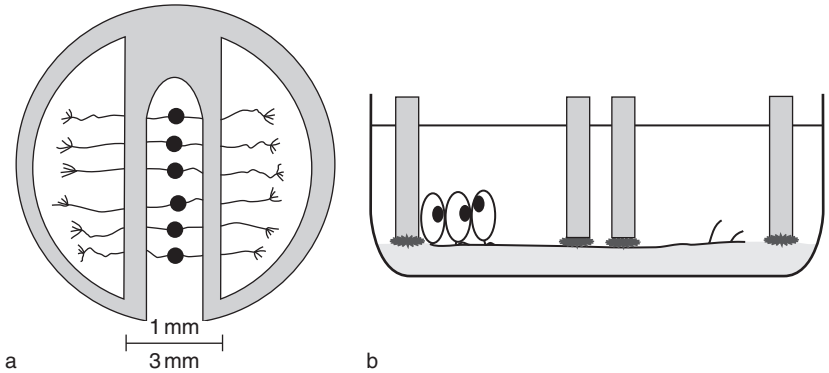

Figure 1 Lysosomal storage disease (LSD) mechanisms of pathogenicity. Disorders of lysosomal storage cause acute or chronic progressive diseases, characterized at the cellular level by a metabolic gridlock usually resulting from either a deficiency of a catabolic enzyme or a defect in vesicular transport.

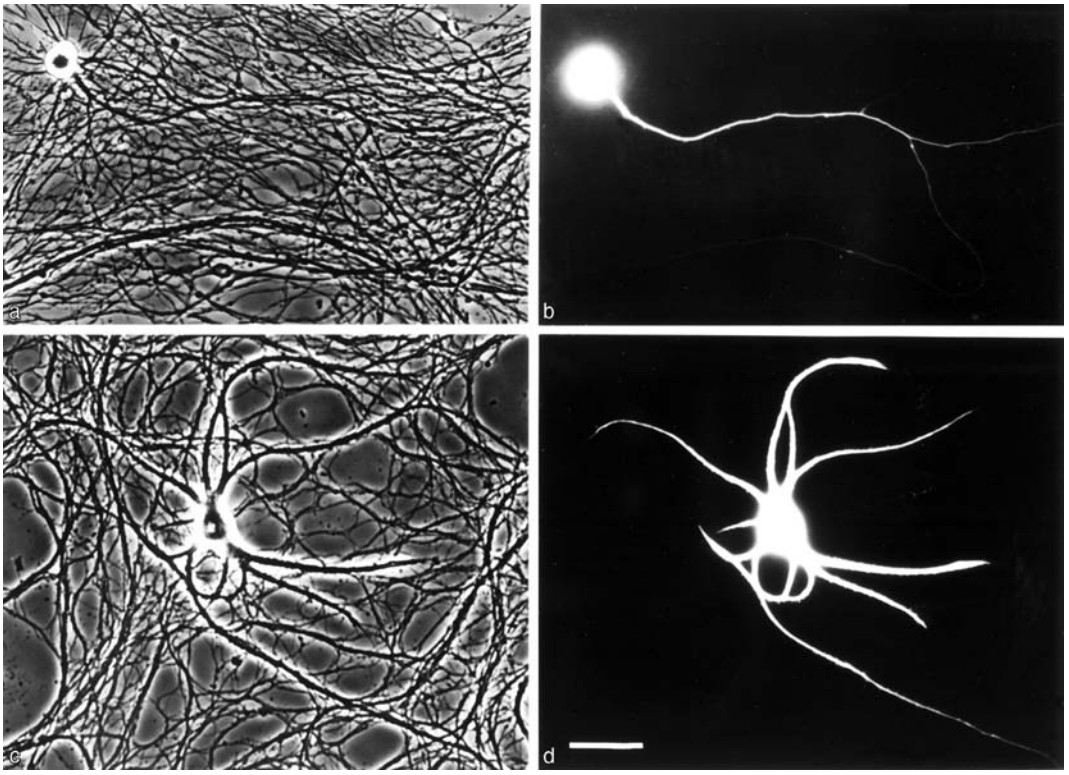

Figure 2

polydystrophy can be confirmed indirectly by finding elevated serum lysosomal enzyme levels and/or decreased lysosomal enzyme levels in cultured fibroblasts. Alternatively, GlcNAc-phosphotransferase can be directly assayed in cultured fibroblasts, or the responsible gene can be screened for mutations.

Unlike the related diseases, mucolipidosis II and IIIA, the enzyme affected in mucolipidosis IIIC (*N*-acetylglucosamine-1-phosphotransferase), also known as variant pseudo-Hurler polydystrophy, retains full transferase activity on synthetic substrates but lacks activity on lysosomal hydrolases. The defect may lie in failure of lysosomal hydrolase recognition.

Pathophysiology and Diagnosis

The progressive intralysosomal accumulation of the incompletely degraded, complex substrates represents the initial and possibly major insult to the cells. Lysosomal storage diseases can occur through various means, including (1) a primary deficiency of a specific hydrolase (e.g., Tay–Sachs, Gaucher, and

Fabry diseases); (2) deficiency of either a 'protective protein' that aids in the lysosomal targeting and prevention of premature degradation of the enzyme (e.g., galactosialidosis), or an 'activator protein' necessary for enzyme–substrate interaction and degradation (GM2 gangliosidosis, AB variant); (3) abnormal enzyme processing by the required enzymes in the endoplasmic reticulum (e.g., multiple sulfatase deficiency); and (4) failure to attach the appropriate targeting signals (M-6-P) in the Golgi apparatus (e.g., I-cell disease). More recently, it has been shown that LSDs can also result from defects in the removal or transport of the substrate from lysosomes (e.g., Niemann–Pick type C and sialic acid storage diseases) or defects in an integral membrane protein of endosomal–lysosomal membranes, leading to abnormalities in endocytosis, vesicle fusion, or the internal endosomal or lysosomal environment (e.g., Danon disease).

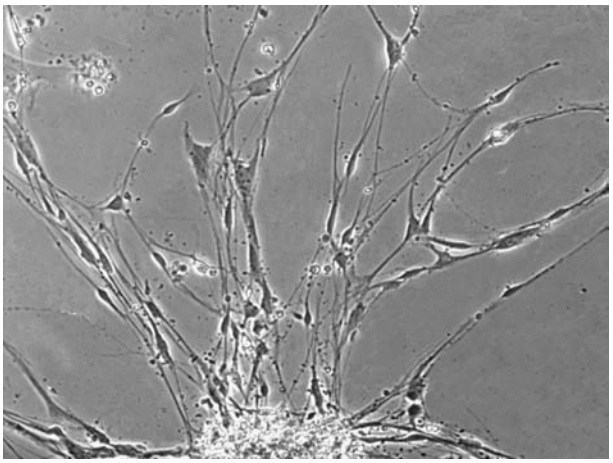

Figure 3

As noted, besides the single enzyme deficiency diseases, LSDs can also arise secondary to multiple enzyme deficiencies. Unlike the peroxisomal disorders, wherein multiple enzyme deficiencies result from a defect of peroxisomal membrane assembly and/or targeting, this subgroup of LSDs arise from either (1) a defect of posttranslational modification involving the glycosylation (mucolipidosis II and IIIA) or sulfation (multiple sulfatase deficiency) of the enzyme/protein product or (2) absence of a protective protein (galactosialidosis I and II).

Multiple sulfatase deficiency (Austin disease) is a rare disease characterized by coarse facial features, hepatosplenomegaly, and skeletal changes in association with progressive neurodevelopmental deterioration and peripheral neuropathy. The disease results from deficiencies of several enzymes (i.e., arylsulfatases A, B, and C, and mucopolysaccharide sulfatase) that normally require the conversion of a specific cysteine thiol residue to an aldehyde to function properly. The defect in the posttranslational modification of these enzymes leads to loss of their catalytic activity, and the accumulation of mucopolysaccharides and steroid sulfates.

Galactosialidosis results from a deficiency of a protective protein/cathepsin A that prevents the intralysosomal degradation of β-galactosidase and sialidase, leading to a combined deficiency of these two enzymes. Galactosialidosis is characterized by coarse facial features, a cherry-red retinal spot, myoclonus, and cerebellar ataxia.

Rare variants of certain sphingolipid storage disorders result from a defect in the enzyme cofactor/activator rather than a primary hydrolase defect.

Table 1 Features helpful in distinguishing lysosomal storage disorders from other organelle diseases

Lysosomal	*Peroxisomal*	*Mitochondrial*
Oligosystemic (RES, CNS)	Multisystemic	Multisystemic
Acute and chronic course	Hypotonia	Lethargy
Behavioral changes	Spatic paraparesis, leukodystrophy (XL-ALD)	Failure to thrive
Seizures (late)		Cerebral dysgenesis
Leukodystrophy		Sensorineural hearing loss
Cerebellar atrophy	Neuronal migration defects (ZS)	Ophthalmoplegia
Macular cherry-red spot		Retinitis pigmentosa
Lung involvement (interstitial disease)	Developmental delay	Cardiomyopathy
	Failure to thrive	Myopathy
Hepatosplenomegaly	Abnormal craniofacies	Diabetes mellitus
Skin lesions (angiokeratoma)	Rhizomelic short stature	
Bone disease (dysostosis multiplex)	Cerebral dysgenesis	
	Sensorineural hearing loss	
	Hepatocellular dysfunction	
	Adrenal dysfunction	

RES, reticuloendothelial system; CNS, central nervous system; XL-ALD, X-linked adrenoleukodystrophy; ZS, Zellweger syndrome.

Table 2 Lysosomal storage disease substrates, enzyme deficiencies, and gene loci

Stored substrate	*Disease*	*Enzyme deficiency*	*Gene locus*
A. Sphingolipids			
GM2 gangliosides, glycolipids, globoside oligosaccharides	Tay–Sachs, GM2 gangliosidosis (three types)	α Subunit β-hexosaminidase	15q23-q24
GM2 gangliosides, oligosaccharides	Sandhoff disease, GM2 gangliosidosis	β Subunit β-hexosaminidase	5q13
GM2 gangliosides, glycolipids	GM2 gangliosidosis	GM2 activator	5q32–33
GM1 gangliosides, oligosaccharides, keratan sulfate, glycolipids	GM1 gangliosidosis (three types)[a]	β-Galactosidase	3p21–3pter
Sulfatides	Metachromatic leukodystrophy (MLD)	Arylsulfatase A (galactose-3-sulfatase)	22q13.31-qter
GM1 gangliosides, sphingomyelin, glycolipids, sulfatide	MLD variant	Saposin B activator	10q21
Galactosylceramides	Krabbe disease	Galactocerebrosidase	14q31
α-Galactosyl-sphingolipids, oligosaccharides	Fabry disease	α-Galactosidase A	Xq22
Glucosylceramide, globosides	Gaucher disease (GD) (three types)[a]	β-Glucosidase	1q2l
Glucosylceramide, globosides	GD (variant)	Saposin C	10q2l
Ceramide	Farber disease (seven types)	Acid ceramidase	8p22-p21.2
Sphingomyelin	Niemann–Pick disease, types A and B	Sphingomyelinase	11p15.1-p15.4
B. Mucopolysaccharidoses (glycosaminoglycans)			
Dermatan sulfate (DS) and heparan sulfate (HS)	Mucopolysaccharidose (MPS) I, Hurler, Scheie	α-L-Iduronidase	4p16.3
DS and HS	MPS II, Hunter	Iduronate-2-sulfatase	Xq27.3-q28
HS	MPS IIIA, Sanfilippo A	Sulfamidase	17q25.3
HS	MPS IIIB, Sanfilippo B	α-*N*-Acetylglucosaminidase	17q21.1
HS	MPS IIIC, Sanfilippo C	Acetyl-CoA:α-glucosaminide-*N*-acetyltransferase	–
HS	MPS IIID, Sanfilippo D	*N*-Acetylglucosamine-6-sulfatase	12q14
Keratan sulfate (KS)	MPS IVA, Morquio A	Galactosamine-6-sulfatase	16q24.3
KS	MPS IVB, Morquio B	β-D-Galactosidase	3p21.33
DS	MPS VI, Maroteaux-Lamy	*N*-Acetylgalactosamine-4-sulfatase	5q13-q14
DS and HS	MPS VII, Sly	β-D-Glucuronidase	7q21.1-q22
Hyaluronan	MPS IX	Hyaluronidase	3p21.3
C. Glycogen			
Glycogen	Pompe, glycogen storage disease (GSD) IIA	α-D-Glucosidase	17q23
Glycogen	Danon disease	Lysosomal-associated membrane protein 2 (LAMP-2)	Xq24–25
D. Oligosaccharides/glycopeptides			
α-Mannoside	α-Mannosidosis	α-mannosidase	19p13.2-q12
β-Mannoside	β-Mannosidosis	β-Mannosidase	4q22-q25
α-Fucosides, glycolipids	α-Fucosidosis	α-Fucosidase	1p34.1-p36.1
α-*N*-Acetylgalactosaminide	Schindler disease	α-*N*-Acetylgalactosaminidase	22q13.1-q13.2
Sialyloligosaccharides	Sialidosis	α-Neuraminidase	6p21.3
Aspartylglucosamine	Aspartylglucosaminuria	Aspartylglucosaminidase	4q34-q35
E. Multiple enzyme deficiencies			
Glycolipids, oligosaccharides	Mucolipidosis II (I-cell disease)	*N*-Acetylglucosamine-1-phosphotransferase	4q21-q23
Sulfatides, glycolipids, glycosaminoglycans	Multiple sulfatases	SUMF-1	3p26
F. Lipids			
Cholesterol esters	Wolman disease, cholesteryl ester storage disease (CESD)	Acid lipase	10q23.2-q23.3

Continued

Table 2 Continued

Stored substrate	*Disease*	*Enzyme deficiency*	*Gene locus*
Cholesterol, sphingomyelin	Niemann–Pick disease type C	NPC1; HE1	18q11-q12; 14q24.3
G. Monosaccharides/amino acid/monomers			
Sialic acid, glucuronic acid	Salla, infantile sialic acid storage disease (ISSD)	Sialin	6q14-q15
Cystine	Cystinosis	Cystinosis	17p12
H. Peptides			
Bone proteins	Ycnodysostosis	Cathepsin K	1q21
I. S-Acylated proteins			
Palmitoylated proteins	Infantile neuronal ceroid lipofuscinosis (NCL)	Palmitoyl-protein thioesterase	1p32
Pepstatin-insensitive lysosomal peptidase	Late-infantile NCL	Pepstatin-insensitive lysosomal peptidase	11p15

[a]Three types imply infantile, childhood, and adulthood presentations.

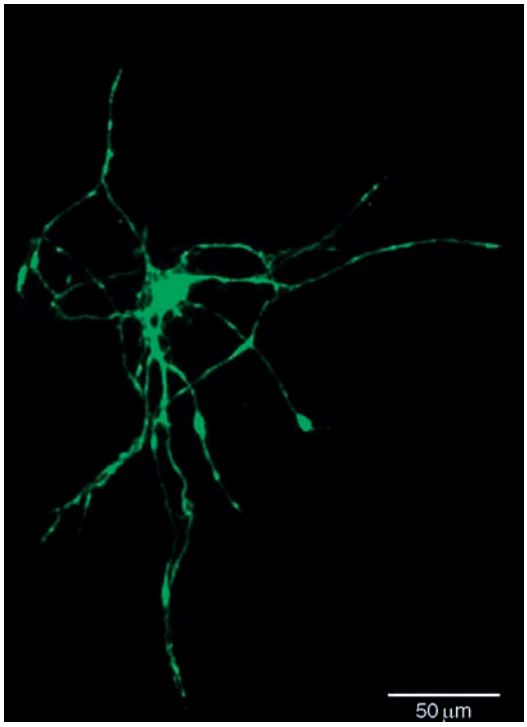

Figure 4

There are two categories of sphingolipid activators, one representing the GM2 activator and a second group of four molecules (saposins A, B, C, and D) derived by proteolytic cleavage of a common precursor, prosaposin, encoded by a gene localized to chromosome 10. Deficiency of the GM2 activator results in the so-called AB variant of GM2 gangliosidosis. Saposin B activates arylsulfatase A; its deficiency gives rise to a variant of metachromatic leukodystrophy (MLD). Saposin C activates glucocerebrosidase and β-galactocerebrosidase; its deficiency leads to a clinical picture often referred to as an atypical form of Gaucher disease (GD; primarily due to glucocerebrosidase deficiency) because of overlap in the clinical features observed in GD patients with the type 3 variant (subacute neuropathic GD). Disorders resulting from cofactor deficiencies are characterized by normal enzymatic activities *in vitro* when using the synthetic (artificial) substrate; thus, the diagnosis can be missed by routine biochemical testing unless the study is performed using the natural glycosphingolipid substrate. Alternatively, molecular analysis may reveal the presence of mutations in the relevant encoding genes.

A few LSDs result from transport defects, including Salla disease and the severe infantile form (ISSD), cystinosis. Salla disease and ISSD are allelic autosomal recessive disorders resulting from a defective gene *SLC17A5* (sialin), an anion/cation symporter, mapped to chromosome 6q14-q15. Both conditions are neurodegenerative disorders characterized by hypotonia, cerebellar ataxia, and mental retardation, with large amounts of sialic acid excreted in the urine. ISSD is further characterized by the presence of visceromegaly and coarse facial features.

Cystinosis is characterized by the accumulation of cystine crystals in lysosomes and principally affects the kidneys. It has been attributed to defects in cystinosin (mapped to chromosome 17p12), an integral lysosomal membrane protein with structural similarities to other membrane transporters. This autosomal recessive disorder comprises three allelic clinical forms, varying in severity and age of onset. The infantile form

generally appears at 6–8 months of age with a proximal renal tubulopathy which, in the absence of renal transplantation, can lead to death by 10 years of age due to renal failure. Other clinical signs include retinal blindness, hypothyroidism, diabetes mellitus, swallowing difficulties, and neurological deterioration. The juvenile form is characterized by glomerular renal damage, which manifests at around 10–12 years of age and slowly progresses to glomerular insufficiency and photophobia due to corneal cystine-crystal deposits. Finally, the ocular nonnephropathic form is solely characterized by a mild photophobia but no renal anomalies.

Unlike the small-molecule diseases (i.e., defects of organic and amino acid metabolism) that lead to acute neurotoxicity, LSDs are often associated with a clinical picture consistent with chronic or subacute encephalopathy. Myoclonic seizures are seen in the following disorders: GM2 gangliosidosis, sialidosis type 1, Schindler disease (α-*N*-acetylgalactosaminidase deficiency), Gaucher disease types 2 and 3, and fucosidosis. However, it should be noted that there are certain subtypes – for example, Gaucher disease type 1, Scheie syndrome (MPS-IS), mild Hunter syndrome (MPS-II), Morquio syndromes A (MPS-IVA) and B (MPS-IVB), Maroteaux–Lamy (MPS-VI), and cystinosis – that are not associated with primary central nervous system involvement.

As a group, the various LSDs are characterized by multisystemic involvement and wide heterogeneity in clinical presentation. Although visceromegaly (e.g., hepatosplenomegaly) is a typical manifestation, there are several exceptions (Table 3). Varying disease severity has been attributed to different mutations that produce a protein with differing residual enzyme activities and potential patient-specific differences in substrate flux (i.e., cellular toxicity is hypothesized to occur when the substrate burden has breached a certain threshold).

A number of causal mutations have been associated with each of the LSDs. In certain disorders, such as Tay–Sachs disease, Gaucher disease, and Niemann–Pick A and B, which are prevalent among individuals of Ashkenazi Jewish ancestry, there are common disease alleles. This fact permits carrier screening prior to marriage introductions among the Orthodox community, and the consideration of prenatal diagnosis in carrier couples at risk. For most of the other LSDs, certain mutations may be common in particular population groups (e.g., Caucasians), although the frequency is not high enough to justify general screening. In most cases, individual families tend to have unique mutations ('private alleles') and mutation analysis is available on a research basis only. Studies have shown genotype–phenotype correlations often to be imperfect.

Table 3 Disorders with CNS involvement but no significant somatic signs

- GM2 gangliosidosis
- α-*N*-Acetylgalactosaminidase (Schindler disease)
- Krabbe disease (globoid cell leukodystrophy)
- Metachromatic leukodystrophy
- Sialidosis type 1
- Mucolipidosis IV
- Neuronal ceroid lipofuscinosis

Microscopic examination of tissue from certain patients suspected to have an LSD may reveal residual bodies within the lysosome – for example, particulate glycogen in glycogenosis type II (Pompe disease) and Danon disease, and lamellar profiles in the sphingolipidoses (i.e., membranous cytoplasmic bodies in Tay–Sachs disease, zebra bodies in the mucopolysaccharidoses, and prismatic inclusions in MLD). Analysis of peripheral smear and skin biopsy may also provide a clue to the diagnosis and help to focus testing and avoid random or sequential screening for a battery of lysosomal enzyme activities. Skin biopsy examination for the presence of characteristic storage bodies is specific in the diagnosis of two disorders, neuronal ceroid lipofuscinosis and mucolipidosis IV. Identification of the underlying defects in mucolipidosis IV (ML IV) and certain forms of neuronal ceroid lipofuscinosis (NCL; see later) now enables mutation analysis as a means of diagnosis for these autosomal recessive disorders in clinically suspected cases. These advances obviate the need for invasive diagnostic procedures and for the required expertise, which is often not readily available, for accurate interpretation of biopsy findings.

When a skin biopsy is performed for ultrastructural studies, it is usually preferable to obtain a second sample for culture; the cultured sample can then be used to analyze specific enzyme activity and also can serve as a source of DNA for subsequent mutation analysis. Moreover, in patients suspected to have Niemann–Pick type C (NPC), analysis of cholesterol esterification and staining for filipin may be performed on skin fibroblasts. This method is also the basis of prenatal diagnosis for NPC, using either cultured villi (CVS; chorionic villus sample) or amniocytes, when the causal mutations are unknown.

It is likely that a host of other factors, including mechanical and vascular changes and immune-mediated events, contributes to the development of various disease-related complications. For instance, a marked increase and abnormal distribution of endothelin-1 have been noted in the neurons and glial cells of a patient with galactosialidosis. In addition, evidence of loss of CD31 immunoreactivity and

the breakdown of vascular cell adhesion molecules in vacuolated endothelial cells has also been reported. These findings are felt to promote development of brain infarctions and the other pathological changes seen in patients with galactosialidosis. In Fabry disease, patients show increased serum levels of various endothelial prothrombotic factors and leukocyte adhesion-molecule expression, changes that are felt to influence the likelihood of stroke. In Krabbe disease (KD, or globoid cell leukodystrophy), the death of oligodendrocytes and loss of myelin have been attributed to psychosine (galactosylsphingosine) accumulation. Recent studies involving the twitcher mouse (a model for KD) brain revealed progressive loss of peroxisomal proteins/functions and induction of expression of inflammatory cytokine tumor necrosis factor-α (TNF-α), which may represent one of the pathways leading to cell death.

Increased knowledge about LSDs and the availability of treatment, including bone marrow transplantation (BMT), enzyme replacement therapy, and substrate reduction therapy, in some forms has stimulated interest in population-based screening. In the absence of specific and effective therapy for most other clinical types, an advantage to early identification of cases includes the prevention of recurrence in families at risk. Furthermore, early diagnosis will allow clinicians to take advantage of the opportunity presented by the naturally suppressed immune system of the neonate, potentially improving outcome following BMT/hematopoietic stem cell transplant (HSCT). Screening strategies that are under development or validation have used blood spots on filter paper (Guthrie cards) as the means of sampling, because of ease in shipping and handling and the potential for large-scale automation. Methods of detection have included a combination of immune-quantification assay or enzyme microanalytic methods, to measure by-products of lysosomal metabolism (lysosomal-associated membrane protein 1 and saposin C) and/or enzyme protein concentration or levels of activity. These approaches may address the limitations of a DNA-based strategy, which can be labor intensive because of the large number of causal mutations associated with most LSDs and low mutant-allele frequencies (i.e., absence of a common mutation).

Several animals (including dogs and cats) have been identified as having a LSD. Additionally, homologous recombination and embryonic stem cell technology have enabled development of authentic animal (usually mouse) models for several entities, including all the sphingolipidoses, two mucopolysaccharidoses, and aspartylglucosaminuria.

Recent Advances

This section highlights recent findings in a few selected disorders, including the neuronal ceroid lipofuscinosis, pycnodysostosis, ML IV, NPC, and Danon diseases. The NCLs are a group of neurodegenerative disorders (outlined in Table 4) characterized by the accumulation of fluorescent storage material in neurons and other cell types. The unifying clinical features include visual impairment, progressive myoclonic epilepsy,

Table 4 The neuronal ceroid lipofuscinoses (NCLs)

Disease	*Gene map*	*Defect*	*Ultrastructure*
Infantile NCL (Haltia–Santavuori disease), allelic with juvenile-onset NCL with granular osmiophilic deposits (GRODs)	1p32, *CLN1*	Lysosomal palmitoyl-protein thioesterase (PPT)	GRODs, saposins A and D
Classic late-infantile NCL (Jansky–Bielschowsky disease)	11p15, *CLN2*	Pepstatin-insensitive lysosomal tripeptidyl peptidase	Curvilinear (CV) bodies; fingerprint (FP) bodies, subunit C of ATP synthase
Juvenile-onset NCL (JNCL; Batten disease)	16p12, *CLN3*	Gene encoding a 438-amino-acid (AA) membrane protein with unknown function	FP bodies, subunit C of ATP synthase
Adult NCL	Unassigned, *CLN4*	438-AA membrane protein, unknown function	Mixed, subunit C of ATP synthase
Finnish variant late-infantile NCL	13q22, *CLN5*	407-AA membrane protein, unknown function	CV and FP bodies, saposins A and D
Variant late-infantile NCL	15q21-q23, *CLN6*	311-AA membrane protein, unknown function	CV and FP bodies, subunit C of ATP synthase
Turkish variant late-infantile NCL (LINCL)	8p23, *CLN7*	Unknown	FP bodies or mixed
Northern epilepsy	8p23, *CLN8*	Unknown	Curvilinear or GROD-like bodies, subunit C of ATP synthase
Variant JNCL (vJNCL)	Unknown, *CLN9*	Unknown	CV and FP bodies, GROD, subunit C of ATP synthase

and cognitive decline. There are several NCL subtypes, delineated on the basis of age of onset, clinical course, and ultrastructural biopsy findings (e.g., granular osmiophilic bodies; curvilinear or fingerprint pattern). The underlying molecular genetic defects for certain subtypes are known (Table 4). Biochemical testing and DNA analysis have recently become available for some forms. The specific roles of the various gene products and the mechanisms involved in neuronal cell degeneration remain to be established. The *CLN3* gene product (the protein CLN3P, or battenin) is a transmembrane protein found in endosomal or lysosomal membranes and may function as a regulator of intravesicular pH. An abnormality of lysosomal pH could lead to faulty protein degradation and explain the presence of storage material found in defective cells. More recently, the neuronal loss in Batten disease has been shown to be due to a combination of apoptosis and autophagy, suggesting that CLN3P, the defective protein, may have an antineuronal death function.

In addition to the NCLs, pycnodysostosis has recently been added to the list of LSDs. Pycnodysostosis is an autosomal recessive skeletal dysplasia characterized by osteosclerosis, bone fragility, and short stature. The gene defect has been mapped to chromosome 1q21. Its protein product, cathepsin K, is a lysosomal cysteine protease that is highly expressed in osteoclasts and is involved in bone remodeling.

Mucolipidosis IV is a progressive disorder associated with corneal clouding, severely delayed developmental milestones, and mental retardation. Neuronal loss and astrocytosis can be found in the cerebral cortex, basal ganglia, deep cerebellar nuclei, and brain stem nuclei. The defective gene (*MCOL-1*) has been mapped to chromosome 19p13.2-p13.3. The gene product (mucolipin-1), a nonspecific cation channel, is speculated to play a role in the late-endocytic pathway and final stages of parietal cell activation. Unlike the other types of mucolipidosis, patients with ML IV do not show skeletal changes or organomegaly. Diagnosis is based on demonstration of inclusion bodies in the lysosomes of different cells types (e.g., skin or conjunctiva). Ultrastructural studies reveal heterogeneous lysosomal storage of laminated membranous material and granular, amorphous vacuoles. Prenatal diagnosis is possible through specific mutation analysis, or linkage studies in cases where the causal mutation is unknown and with electron microscopic examination of amniocytes in search of the inclusion bodies.

Niemann–Pick type C is a neurovisceral disorder characterized by a wide spectrum of manifestations, ranging from fatal neonatal liver disease to isolated splenomegaly in late childhood. Neurological findings include ataxia, vertical supranuclear ophthalmoplegia, dystonia, and dementia with progressive deterioration. Neurofibrillary tangles have been found in the brain of patients with a slowly progressive course. Biochemically, there is evidence of lysosomal sequestration of endocytosed low-density lipoprotein (LDL)-derived cholesterol, premature and abnormal enrichment of cholesterol in *trans*-Golgi cisternae, and accompanying anomalies in intracellular sterol trafficking. Genetic heterogeneity, established by cell hybridization and linkage studies, has indicated the presence of two complementation groups that are indistinguishable from each other from a clinical, cellular, or biochemical perspective. Molecular studies have found that NPC can arise from a defect of a major locus, *NPC1* (~95% of cases) mapped to chromosome 18q11, or a minor locus (the *NPC2*, or *HE1*, gene) on chromosome 14q24.3. The *NPC1* gene encodes a multidomain membrane protein that may play a role in retrograde transport of several types of lysosomal cargo and/or may function as a transmembrane molecular pump. The *NPC2* (*HE1*) gene encodes a soluble lysosomal protein with cholesterol-binding properties. The accumulation of various glycosphingolipids, including GM2 ganglioside, in NPC is believed to contribute to the neurodegenerative changes encountered in this condition.

Studies have shown that Niemann–Pick type D (NPD), which is relatively common in southwestern Nova Scotia due to a founder effect (i.e., a high incidence within a population derived from a common ancestor), also arises from mutation of *NPC1*. Thus, Niemann–Pick types C and D are allelic disorders – that is, they are disorders which result from different mutations in the same gene. A point mutation within the *NPC1* gene (G3097→T; Gly992→Trp) has been demonstrated in NPD patients.

Danon disease (lysosomal glycogen storage disease with normal acid maltase) is a rare disorder characterized by mental retardation, myopathy, and cardiomyopathy. It is inherited in an X-linked inheritance fashion. As in Pompe disease (an acid maltase, i.e., β-glucosidase, deficiency), muscle cells exhibit intracytoplasmic vacuoles containing glycogen and autophagocytic material. Defects in the gene encoding lysosomal-associated membrane protein 2 (LAMP-2) have been identified as the cause. It is suspected that storage in Danon disease occurs secondary to a failure of fusion of autophagic vacuoles with endosomes/lysosomes or the maturation of autophagolysosomes into digestive organelles.

Additional very rare LSDs have been described. Glutamyl ribose-5-phosphate storage disease is a newly delineated lysosomal storage disorder associated with progressive neurological deterioration (developmental delay, seizures) and renal failure. In this disease the

deficiency of adenosine diphosphate (ADP)-ribosyl protein lyase results in a defect of posttranslational modification (ADP-ribosylation) of proteins. Patients demonstrate neuronal storage of periodic acid–Schiff (PAS)-positive material. Recently, a deficiency of serum hyaluronidase (designated as mucopolysaccharidosis IX) was reported associated with a mild clinical phenotype characterized by periarticular soft tissue masses, mild short stature, and absence of neurological or visceral involvement. The patient showed histological and ultrastructural evidence of a lysosomal storage disease. A mutation in hyaluronidase 1 (chromosome 3p21.3) was identified. In addition, recent examination of the biochemical and molecular data of patients with α-*N*-acetylgalactosaminidase (α-NAGA) deficiency, which is thought to lead to a form of neuroaxonal dystrophy (Schindler disease), strongly suggests that factors other than α-NAGA may be relevant in determining disease expression.

Management

The mainstay of treatment for most of the LSDs has been primarily palliative or supportive care. Early complementation studies have shown that when two cells lines from patients with different LSDs, such as Hurler (MPS I) and Hunter (MPS II) syndromes, are co-cultured, there is often cross-correction of the metabolic defect. This observation provided the rationale for consideration of cellular (bone marrow) transplantation, enzyme therapy, and gene therapy.

Bone marrow therapy and hematopoietic stem cell transplantation have been attempted for a variety of LSDs, with varying success. The differences in response appear to be dependent on the age of onset and rate of disease progression. Treatment is directed at prevention or amelioration of the inexorable neurological deterioration that is the major pathophysiological event. Recent reports on BMT include description of long-term engraftment in a single patient with Wolman disease; this led to normalization of peripheral leukocyte lysosomal acid lipase enzyme activity, associated with gains in developmental milestones, normal liver function, and normal levels of cholesterol and triglyceride. In a patient (a 9.6-month-old female) with Farber disease, BMT was reported to result in resolution of several features, including subcutaneous lipogranulomata, pain on joint motion, and hoarseness, 2 months posttransplant. Despite modest gains in cognitive and language development, hypotonia and delayed motor skills persisted and there was progressive decline with loss of engraftment, and death at 28 months after BMT (age 37.5 months). In four patients with late-onset Krabbe disease (globoid cell leukodystrophy), the central nervous system deterioration was reversed. Brain magnetic resonance imaging (MRI) showed a decrease in signal intensity in the three patients with late-onset disease who were assessed both before and after transplantation, and abnormalities in cerebrospinal fluid total protein levels were corrected in three of these patients. Long-term data on a greater number of patients for whom BMT has been attempted are required before consideration of its more generalized application.

The high morbidity and mortality associated with BMT has prompted consideration of other therapeutic approaches. Enzyme therapy for Gaucher disease was introduced in 1991, initially using the purified human placental-derived glucocerebrosidase (acid β-glucosidase). Subsequently, the placental product (Ceredase) was replaced with a recombinant enzyme (Cerezyme), isolated from Chinese hamster ovary (CHO) cells transduced with the human gene sequence and subsequently modified to expose the α-mannosyl residues of the protein product for enhanced uptake by the macrophage. The use of these enzyme formulations in symptomatic patients has been shown to be safe and effective in reversing the clinical manifestations of Gaucher disease. Other disorders in which enzyme therapy is available include Fabry disease, Pompe disease, and certain subtypes of MPS (I, II, and VI). These diseases are associated with well-characterized phenotypes and defined single-enzyme deficiencies. In contrast to the recombinant enzyme administered to patients with Gaucher disease, the enzyme formulations for the latter LSDs rely primarily on targeted delivery of the relevant enzyme through the M-6-P receptors. For Fabry disease, two enzyme formulations are currently available in Europe, distinguished based on the cellular source of the recombinant enzyme. The recombinant α-galactosidase A is generated using either human (Replagal) or CHO (Fabrazyme) cells; the latter formulation is currently the only form available in the United States.

Substrate synthesis inhibition (also referred to as substrate reduction therapy), using miglustat (an imino sugar), has recently been approved for the treatment of Gaucher disease. The same molecule is being explored as a therapeutic option for other types of sphingolipidosis (GM2 gangliosidosis, including Sandhoff disease and Fabry disease) and for Niemann–Pick type C. This approach relies on maintaining substrate flux by limiting the amount of precursor substrate synthesized and ultimately presented to an impaired enzyme that has low but some degree of residual activity. An interesting study recently reported on the potential for the use of phosphocysteamine in patients with neuronal ceroid lipofuscinosis. Phosphocysteamine has been administered to patients with nephropathic

cystinosis as a means of suppressing lysosomal cystine storage. Phosphocysteamine also shows the ability to disrupt thioester linkages in palmitoylated protein and, thus, may mediate the depletion of lysosomal ceroids in the NCLs. Gene therapy is also being considered, although most trials have been restricted to animal models of disease. A limited trial performed on humans with Gaucher disease has not been shown to lead to sustained benefit because enzyme expression was not persistent in the virally transduced cells. More recent investigational strategies include chaperone-mediated enzyme enhancement and stop-codon readthrough.

See also: Lysosome and Endosome Organization and Transport in Neurons; Niemann–Pick Disease.

Further Reading

Beutler E (2006) Lysosomal storage diseases: Natural history and ethical and economic aspects. *Molecular Genetic and Metabolism* 88: 208–215.

Brady RO (2006) Enzyme replacement for lysosomal diseases. *Annual Review of Medicine* 57: 283–296.

Fukuda T, Ewan L, Bauer M, et al. (2006) Dysfunction of endocytic and autophagic pathways in a lysosomal storage disease. *Annals of Neurology* 59(4): 700–708.

Haley SM, Fragala-Pinkham M, Latham NK, et al. (2005) The emerging role of the pediatric physical therapist in evaluation and intervention for individuals with lysosomal storage diseases. *Pediatric Physical Therapy* 17(2): 128–139.

Hodges BL and Cheng SH (2006) Cell and gene-based therapies for the lysosomal storage diseases. *Current Gene Therapy* (2): 227–241.

Hoffmann B and Mayatepek E (2005) Neurological manifestations in lysosomal storage disorders – from pathology to first therapeutic possibilities. *Neuropediatrics* 36(5): 285–289.

Lachmann RH (2006) Miglustat: Substrate reduction therapy for glycosphingolipid lysosomal storage disorders. *Drugs of Today (Barcelona)* 42(1): 29–38.

Martin PL, Carter SL, Kernan NA, et al. (2006) Results of the cord blood transplantation study (COBLT): Outcomes of unrelated donor umbilical cord blood transplantation in pediatric patients with lysosomal and peroxisomal storage diseases. *Biology of Blood and Marrow Transplantation* 12(2): 184–194.

Meikle PJ, Grasby DJ, Dean CJ, et al. (2006) Newborn screening for lysosomal storage disorders. *Molecular Genetics and Metabolism* 88: 307–314.

Pastores GM (2003) Enzyme therapy for the lysosomal storage disorders: Principles, patents, practice and prospects. *Expert Opinion on Therapeutic Patents* 13(8): 1157–1172.

Pastores GM (2006) Miglustat: Substrate reduction therapy for lysosomal storage disorders associated with primary central nervous system involvement. *Recent Patents on CNS Drug Discovery* 1(1): 77–82.

Pastores GM and Barnett NL (2005) Current and emerging therapies for the lysosomal storage disorders. *Expert Opinion on Emerging Drugs* 10(4): 891–902.

Sands MS and Davidson BL (2006) Gene therapy for lysosomal storage diseases. *Molecular Therapy* 13: 839–849.

Vellodi A (2005) Lysosomal storage disorders. *British Journal of Haematology* 128(4): 413–431.

Lysosome and Endosome Organization and Transport in Neurons

J G Gindhart and K P Weber, University of Richmond, Richmond, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The cellular compartments interposed between the trans-Golgi and plasma membrane are essential for neuron development and function. Neuronal endosomal and lysosomal pathways are similar to other cell types in the response to extracellular signals, dissociation of signal from receptor, receptor down-regulation, and killing of foreign invaders. Neurons also use endosomal pathways to recycle synaptic vesicles and reload them with neurotransmitter. A key question is how compartments of the endosome–lysosome pathway are related spatially in the neuronal cytoplasm, and how the often great distances between them are traversed in light of temporal parameters of cellular processes such as cell death, differentiation, and intercellular communication. The actin and microtubule cytoskeletal elements and their associated motor proteins effect the movement of endosomal vesicles in neurons. Membrane budding and fusion events are tightly regulated in response to changing conditions within and outside the neuron, and vesicle budding and fusion events are coupled to the recruitment of specific motor proteins through the activity of the Rab superfamily of small guanosine triphosphatase (GTPase) proteins. Disruption of endosomal transport pathways through impairment of motor protein function can cause neurodegenerative disease, whereas illnesses related to disruption of cholesterol homeostasis and glycolipid metabolism are caused by loss of function or transport of specific lysosomal enzymes. Here we review the importance of endosomal and lysosomal transport pathways in neurons, the role and regulation of motor proteins involved in those pathways, and the consequences of disrupting their normal function.

Endosome to Lysosome, with Stops along the Way

Endosome and lysosome compartments are characterized by their morphological and functional properties. Endosomes are typically involved with the sorting and delivery of lipid vesicles and their contents to and from the plasma membrane. 'Early' endosomes are positioned close to sites of active endocytosis, where they can act as a recycling compartment for vesicles budding from the plasma membrane. In this way, early endosomes provide a sorting environment for lipid, protein, and extracellular molecules contained within endocytic vesicles. Early endosomes are sometimes formed by the fusion of several endocytic vesicles, but they often exist as stable structures near the presynaptic and postsynaptic membranes. Sorting signals on transmembrane and soluble proteins provide instructional cues regarding whether a protein will return to the plasma membrane or be packaged into membrane-bound organelles destined for the 'late' endosome. For example, transmembrane receptors essential for neurotransmission may be returned to the plasma membrane, but components of active signaling pathways may be transported to the cell body, where those cues are amplified and implemented by downstream signaling events. In some neurons, a distinct recycling endosome is positioned between the early endosome and plasma membrane in the recycling pathway.

Membrane-bound organelles destined for the late endosome bud from the early endosome and are transported in a retrograde fashion toward the cell interior. In some neurons, a perinuclear endocytic recycling compartment serves as an intermediate stopping point between early and late endosomes. The transport 'carriers' between early and late endosomes can be individual vesicles, elongated tubules, or multivesicular bodies (MVBs) formed by fusion and subsequent modification of early endosome vesicles. There is some evidence that specialized signaling endosomes, having characteristics of early endosomes, transport growth factor signals along the axon, providing rapid signal propagation and spatial-temporal regulation of signaling cues.

The late endosome is a key sorting point for membrane vesicles and their contents. Proteins destined for the lysosome arrive at the late endosome from the trans-Golgi and early endosome compartments; vesicles bud from the late endosome to travel to the trans-Golgi or lysosome. Signal motifs and post-translational modifications such as glycosylation determine whether a protein in the late endosome will be incorporated into vesicles destined for the trans-Golgi or lysosome. For example, transmembrane receptors transported to the late endosome are usually transported to lysosomes, but phagocytosed virus particles and active signal transduction pathway components are transported to the trans-Golgi. In addition, cytosolic cysteine residues of many membrane-associated proteins can be targeted with ubiquitin to signal their transport to multivesicular bodies from the endosomal pathway and subsequent degradation within lysosomes.

Lysosomes are acidic organelles that facilitate the digestion of proteins, organelles, and extracellular debris through the action of acid hydrolases. In neurons, lysosomes have an important role in receptor downregulation of membrane receptors to attenuate response to extracellular ligands. Lysosomes are also essential for degrading G-protein aggregates associated with neuropathies such as Huntington's disease (HD) and are important for regulating cholesterol homeostasis. Receptor proteins on the plasma membrane are delivered directly to lysosomes through the formation of MVBs that fuse with the lysosome. Cytoplasmic structures such as organelles, ribosomes, and protein aggregates are delivered to lysosomes through the formation of autophagic vacuoles by membrane cisterna that surround the cytoplasmic component to be degraded and then fuse with the lysosome. Lysosomes are also formed by the maturation and acidification of late endosomes, thereby degrading endocytosed material not targeted to the trans-Golgi for recycling.

Positioning of Endosomes and Lysosomes in Neurons

A stereotypical neuron has morphology analogous to that of a polarized epithelial cell. The axon is the apical surface; organelles are actively targeted to the axon using cargo-specific information and cytoskeletal modifications; organelles and proteins lacking an axon targeting signal are excluded from axons at the axon hillock and thus reside in the cell body. Early and recycling endosomes are located near the presynaptic and postsynaptic membranes and in the cell body, but lysosomes are typically juxtanuclear, near the microtubule organizing center (MTOC). Late endosomes are usually found in the cell body near the Golgi complex. To ensure normal neuron function, organelles of the endosome pathway must be located at their appropriate locations within the cell, and protein–protein and membrane–membrane interactions must occur specifically and efficiently. Small GTPases act as master regulators of organelle positioning and membrane dynamics, whereas motor proteins and the microtubule and actin cytoskeleton facilitate interorganelle transport and maintenance of organelle distribution.

A Large Family of Small GTPases Controls Transport and Fusion Events

Fusion of donor and target membranes in the endosome pathway must be controlled precisely to maintain the distinctiveness of each compartment and deliver macromolecules to the appropriate destination. Membrane interactions among cell compartments are controlled by small GTPases that cycle between guanosine diphosphate (GDP)-bound cytosolic and GTP-bound membrane-associated states. The Rab superfamily of small GTPases regulates membrane budding, transport, and fusion events within the endosome pathway. There are many different Rab family members, each of which interacts with a different membrane type to facilitate homotypic or heterotypic membrane fusion and provide directionality to membrane transport. Rab proteins do not perform membrane budding, transport, and fusion; rather, GTP-bound Rabs recruit specific proteins to donor and target membranes to facilitate these events. For example, early endosome membranes bind Rab5-GTP through the activity of Rab5-binding protein Rabex5 and Rabaptin5. Next, Rab5 recruits phosphatidylinositol-3-OH kinase (PI(3)K) effectors that add phosphate groups to phosphatidylinositol to generate phosphatidylinositol-3-phosphate (PtdIns(3)P). Proteins containing the Fab1p, YOTB, Vac1p, and EEA1 (FYVE) finger structural motif bind to PtdIns(3)P and are recruited to microdomains of PtdIns(3)P on early endosomes. FYVE finger proteins act as Rab5 effectors to facilitate vesicle budding, transport, and fusion events away from early endosomes, congression of vesicles during early endosome formation, and tethering of early endosome membranes in close apposition to donor compartments such as the presynaptic membrane. These findings and others support a model in which different Rabs are recruited to specific membrane compartments and the binding partners and effectors of these Rabs impart each membrane compartment with its specific functional properties. Therefore, while the detailed mechanisms controlling effector protein recruitment and function differ for individual Rabs, the general theme of Rabs as the master controllers of endosome pathway transport is shared within the endomembrane system.

Microtubule Motors Involved in Endosome and Lysosome Transport

Microtubule motor proteins convert the energy of ATP hydrolysis into processive movement along microtubules. There are two major classes of microtubule motor proteins, kinesins and dyneins. Kinesins typically walk toward the plus end of microtubules, whereas dyneins walk toward the minus end. Microtubules are usually oriented so that the plus end is located at the cell periphery and the minus end is closer to the cell body. In neurons, microtubule plus ends point toward the presynaptic membrane, but dendritic microtubules are of mixed polarity. The orientation of microtubules is important for developing hypotheses for which types of motor

protein can drive organelle translocation and positioning in the endosome–lysosome system. For example, plus end-directed motors such as kinesin-1 could be involved in delivering receptors from recycling endosomes to the plasma membrane, whereas dynein could be responsible for delivering endosome-derived tubules and multivesicular bodies from distal locations to late endosomes in the cell body. Several microtubule motors have been shown to mediate membrane transport in the endosome–lysosome pathway. Although some studies have been conducted in nonneuronal systems, such as cultured mammalian cells, the motors described below are expressed in the nervous system; a reasonable prediction is that these proteins display similar transport properties in neurons.

KIF13A

The kinesin-3 family member KIF13A transports vesicles from the trans-Golgi to endosomes, lysosomes, and plasma membrane through direct binding to β1-adaptin component of the adaptor protein 1 (AP-1) adaptor complex. KIF13A messenger RNA (mRNA) is expressed ubiquitously in the developing and adult mouse and is found in membrane and cytosolic fractions from mouse brain lysates, and immunostaining shows that KIF13A is associated with the trans-Golgi and AP-1-containing cytoplasmic vesicles. AP-1 associates with membranes through its interaction with the transmembrane mannose-6-phosphate receptor (M6PR) in the trans-Golgi, providing a cue to build a clathrin-coated vesicle at the site of AP-1 binding and tether a motor to the nascent vesicle. The M6PR directs proteins from the trans-Golgi to late endosomes and lysosomes through cytoplasmic binding of AP-1 and lumenal binding of glycoproteins modified with M6P. The phenotype of KIF13A knockouts is not known; overexpression of KIF13A in cultured cells causes redistribution of AP-1 and M6PR from the Golgi to the tips of neuritelike processes, but the overall distribution of endosomes and Golgi membranes is not affected. M6PR at the plasma membrane binds to soluble ligands such as insulin-like growth factor II, mediating their endocytosis and degradation within the lysosome. This suggests a model in which KIF13A might have an indirect role in nervous system development by delivering M6PR to the plasma membrane. M6PR could then attenuate the extracellular concentration of specific growth factors through their endocytosis and lysosomal degradation.

KIF16B

The kinesin-3 family member KIF16B has a Phox homology (PX) lipid-binding domain that facilitates interaction with membranes containing PtdIns(3)P-lipids. Rab5 recruits KIF16B to early endosomes through Rab5 effector hVPS34/p150, which catalyzes formation of the phospholipid PtdIns(3)P on early endosome membranes. In cultured cells, KIF16B is localized to early endosomes in a PtdIns(3)P- and Rab5-dependent manner. Overexpression of KIF16B causes early endosomes to redistribute to the cell periphery, and loss of KIF16B function causes perinuclear endosome clustering. Knockout experiments in cultured cells implicate KIF16B in early endosome recycling events, such as recycling of transferrin and epidermal growth factor receptor (EGFR) to the plasma membrane. Reduction of KIF16B function causes delayed recycling of transferrin receptor back to the plasma membrane and prevents the entry of endocytosed EGFR into the lysosomal degradation pathway. This leads to an accumulation of EGFR in early endosomes of cells that lack KIF16B. Because epidermal growth factor and its receptors have important roles in the development and differentiation of pre- and postmitotic neurons, KIF16B, like KIF13A, could have an important but indirect role in neural cell signaling pathways during and after embryonic development.

Kinesin-2

Kinesin-2 is important for the transport and distribution of membrane-bound organelles in neuronal and nonneuronal cells. Loss of kinesin-2 function causes defects in axonal transport in *Drosophila* sensory neurons, failure to transport pigment granules in vertebrate photoreceptor and pigmented epithelial cells, and loss of cilia required for sense organ and kidney function. Recent experiments in cultured cells have demonstrated that loss of kinesin-2 function causes late endosomes and lysosomes to become concentrated in the cell interior near the MTOC. This effect is caused by a reduction in outward plus end movement of late endosomes and lysosomes away from the MTOC. While the transport of endocytosed probes from early to late endosomes is not affected in cells lacking kinesin-2, it is plausible that recycling of endocytosed receptors back to the plasma membrane could require kinesin-2 function. Kinesin-2 interacts directly and indirectly with membrane cargoes through protein–protein interactions between the kinesin-2 tail domain and proteins on the cargo surface. Regulation of membrane–kinesin-2 interactions can occur directly, as endosomal Rab4 binds kinesin-2 in coimmunoprecipitation experiments, or indirectly, as kinesin-2 binds cytoplasmic dynein, which can also bind membrane cargoes (described below). Kinesin-2 is expressed ubiquitously but accumulates at highest levels in the nervous system; future work will explore further the potential role of kinesin-2 in lysosome and late endosome transport *in vivo*.

Kinesin-1

Kinesin-1 is the major anterograde motor for transport along microtubules. Loss of kinesin-1 function results in lethality and defects in intracellular transport and nervous system development. Kinesin-1 is essential for the transport of mitochondria, endoplasmic reticulum- and Golgi-derived vesicles, messenger RNAs, and tubulin and intermediate filament subunits, but kinesin-1 family members such as mouse KIF5 are also directly involved in endosome and lysosome transport in neurons. Mutations that block KIF5 motor activity block centrifugal transport of lysosomes. Earlier studies also show that kinesin-1 is involved in the movement of endosome-derived membranes and dispersion of vesicles from the endosome recycling compartment. Another important function of kinesin-1 is the transport of minus end-directed microtubule motors such as dynein and KIFC2 to the cell periphery, where they can bind to cargoes for centripetal transport toward the cell interior. Like kinesin-2, specific protein–protein interactions mediate cargo–kinesin-1 interactions; for example, the adapter protein glutamate receptor interacting protein 1 links kinesin-1 to dendritic-recycling endosomes containing the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor. Although plus end motors such as kinesin-2, KIF16B, and KIF13A are required for specific endosomal transport events, it is likely that there is some functional redundancy among these motors and kinesin-1.

KIFC2

The minus end-directed kinesin-14 family member KIFC2 carries multivesicular bodies to dendrites and is essential for the fission of vesicles destined for late endosomes from early endosomes. KIFC2 is expressed in the developing and adult mouse brain and is found associated with membrane-bound organelles such as endosomes and MVBs in dendrites, axons, and cell body. In mammalian cells, KIFC2 and kinesin-1 are associated with early endosomes, but dynein, KIFC2, and kinesin-2 are found on late endosomes. Rab4 regulates KIFC2 activity, but in an unexpected way. Rab4-GDP enhances KIFC2 activity, in contrast to kinesin-2, whose activity is enhanced by Rab4-GTP. This observation suggests that the nucleotide-bound state of Rab4 may regulate whether early endosomes move toward the plasma membrane (Rab4-GTP, kinesin-2 active) or toward the late endosome (Rab4-GDP, KIFC2 active). There are some questions about the *in vivo* importance of KIFC2-dependent transport, as mice lacking KIFC2 are normal; perhaps this reflects functional redundancy with other minus end microtubule motors such as dynein. Therefore, while KIFC2 clearly functions in the retrograde transport of early endosomes, further work is required to better understand the relevance of KIFC2-mediated transport *in vivo*.

Cytoplasmic Dynein

The minus end-directed microtubule motor cytoplasmic dynein transports organelles toward the cell interior. It is composed of many subunits, some of which hydrolyze ATP to generate force and others, such as dynactin, regulate dynein-based transport and link motor activity to cargo transport. Dynein is important for transporting retrograde signaling endosomes inward from the synapse and delivering autophagosome contents to lysosomes, and it has an essential role in the formation of aggresomes, which are clusters of cytoplasmic protein aggregates transported toward the MTOC. Aggresomes are thought to have a neuroprotective function by concentrating protein aggregates in one location, thereby helping to facilitate their removal by autophagy.

A current model for dynein-dependent membrane transport is that dynein and its associated proteins are transported by kinesins to the plus end of microtubules at the cell periphery, where dynactin and membrane cargoes bind to the microtubule tip in a CLIP170-dependent manner. Next, dynactin recruits dynein, and the motor cargo complex proceeds along microtubules toward the cell interior. An alternative scenario is that stable microtubules located near cortical actin filaments can have cargoes delivered to their plus ends by actin–myosin-dependent pathways. Subunits of the dynein–dynactin complex interact with membranes through protein–protein and protein–lipid interactions. For example, dynein light chain and dynactin are Rab4 and Rab5 effectors on early endosomes, and the Rab7 effector protein called Rab-interacting lysosomal protein recruits dynein to late endosomes.

Role of Actin and Myosin in Endosome–Lysosome Transport

The actin cytoskeleton has an important role in membrane transport within the endosome pathway. Like microtubules, actin filaments have an intrinsic polarity. The plus (barbed) end of actin filaments is typically oriented toward the plasma membrane, so myosin motors that move toward the barbed end have been proposed to be involved in moving organelles toward the plasma membrane, whereas minus (pointed) end-directed myosins may be involved in centripetal transport events. While actin filaments can serve as a substrate for myosin-based transport events, actin filament polymerization near membranes can provide

mechanical forces sufficient for vesicle movement. The Rho family of small GTPases, such as RhoA, Rac1, and Cdc42, can bind membranes and recruit proteins that nucleate actin filament growth. Rho GTPases can also affect actin dynamics by stimulating actin filament severing and depolymerization. This facilitates remodeling of the actin cytoskeleton near the cell surface, which is often required for endocytosis, exocytosis, and phagocytosis to occur. In the endosome–lysosome pathway, Cdc42-like protein TCL is required for delivery of clathrin-coated vesicles to early endosomes, and RhoB and RhoD affect endosome dynamics through recruitment of actin filaments to early endosomes. Therefore, it appears that actin dynamics have an essential role in intracellular membrane dynamics, as has been observed for cell motility phenomena.

Myosin-V

The plus end-directed myosin-V translocates vesicles processively along actin filaments, thus enabling this motor to transport cargoes over relatively long distances, but it also plays an important role in tethering organelles near the plasma membrane. Myosin-V has been implicated in the transport of specialized lysosomal organelles called melanophores in pigmented epithelial cells, in addition to phagosomes and vesicles associated with early endosomes. Regulated attachment of myosin-Va to melanophores is mediated by melanophilin, which links myosin-Va to Rab27 on the melanosome surface. Loss of myosin-Va, Rab27, or melanophilin function in mice results in skin pigmentation defects because of dynein-dependent transport of melanophores away from the plasma membrane. Myosin-Vb binds to Rab11 family of interacting protein 2 (Rab11-FIP2), an effector of Rab11 on recycling endosomes, suggesting that FIP2 recruits myosin-Vb to recycling endosomes, perhaps for delivery of vesicles along actin filaments to the plasma membrane. Interactions between myosin-V, dynein, and kinesin have been observed, suggesting that these motors may link actin- and microtubule-dependent transport.

Myosin-VI

The nonprocessive minus end-directed motor myosin-VI has an important role in diverse membrane transport events. It is found associated with clathrin-coated endocytic vesicles, Golgi membranes, and secretory vesicles. Because myosin-VI is a minus end motor, a model for its function is the delivery of endocytic vesicles to early endosomes, as well as the maintenance of Golgi morphology. Loss of myosin-VI function in *Snell's waltzer* mice results in deafness, defects in endocytosis, and a small Golgi phenotype. Myosin-VI binds to trans-Golgi membranes through its interaction with optineurin, which is an effector of Rab8. Similar interactions are observed between myosin-VI and Rab11-FIP2, an effector of Rab11 on recycling endosomes, much like the interaction observed between myosin-Vb and Rab11-FIP2. During endocytosis, myosin-VI interacts with Dab2, which binds to the low-density lipoprotein (LDL) receptor and clathrin adapter protein AP-2. Therefore, it appears that Dab2 links centripetal actin-dependent transport by myosin-VI with formation of endocytic vesicles containing the LDL receptor.

Myosin-VII

Myosin-VII is a plus end-directed actin motor required for melanosome and lysosome transport. Loss of myosin-VIIa function is responsible for the *shaker-1* mouse phenotype and causes human Usher syndrome type 1B (USH1B). The symptoms of USH1B include deafness, blindness, and vestibular dysfunction. The blindness observed in USH1B patients is caused by a failure of melanosome transport in retinal pigmented epithelial cells, as well as defective phagocytosis of photoreceptor disks, a process essential for the replacement of aging rhodopsin-containing membranes. Myosin-VIIa binds to membranes using a mechanism similar to myosin-Va, using the Rab27 effector melanophilin to bind membranes in a Rab27-dependent manner.

Myosin-X

The plus end-directed myosin-X is essential for phagocytosis in mammalian cells. Like KIF16B, myosin-X binds to phospholipids modified by PI(3)K. Myosin-X is important for filipodia extension in motile cells; by analogy, perhaps it provides motile forces for pseudopod extension at the plus end of actin filaments. This observation leads to the appealing model that small GTPases recruit PI(3)K to sites where phagocytosis will occur, and then the generation of PtdIns (3)P recruits myosin-X, resulting in the formation of a phagocytotic cup around the particle to be engulfed. At this time, the identity of the proteins required for recruitment of myosin-X to membranes is unknown.

Involvement of Endosome–Lysosome Transport in Neurological Disease

Defects in the endosomal pathway have been implicated in the pathogenesis of various neuropathies. Disruption of nerve growth factor (NGF) signaling, which leads to a decreased rate of neurite outgrowth, has been implicated in a number of neurodegenerative diseases, including Down syndrome, Alzheimer's disease (AD) and HD. Binding of NGF to the TrkA receptor at the synapse triggers the NGF signaling pathway,

which induces receptor dimerization, phosphorylation, and the formation of early endosomes transported by dynein to late endosomes. There, TrkA dissociates under increasingly acidic conditions and is usually recycled back to the plasma membrane, whereas NGF translocates to the nucleus to regulate transcription of select genes. NGF signaling is disrupted in the mouse model for Down syndrome, which displays similar pathology to AD, and these defects are similar to those observed in NGF or TrkA knockout mutants.

TrkA expression is downregulated and neurite outgrowth is impaired in cells in which huntingtin-associated protein-1 (HAP1) is mutated. HAP1 was originally identified in a yeast two-hybrid screen for proteins that bind huntingtin, which contains polyglutamine expansion mutations that form aggregates in HD. HAP1 also binds the kinesin light chain and the p150Glued subunit of dynactin, indicating that it may be involved in motor protein-mediated trafficking. Neurons in which HAP1 is expressed display less-serious degenerative phenotypes, and it is interesting that HAP1 also binds vesicles containing amyloid precursor protein, providing a possible connection to the defects in NGF signaling displayed in AD.

A second huntingtin-associated protein, HAP40, mediates the recruitment of huntingtin to early endosomes in a Rab5-dependent manner and may help early endosomes move from actin-dependent transport near the plasma membrane to microtubules. Upregulation of HAP40 is observed in fibroblasts in HD patients, resulting in reduced movement of early endosomes along microtubules and preferential localization of HAP40 to cortical actin filaments.

The clearance of aggregate-prone proteins, including huntingtin, appears to be mediated by dynein, which facilitates autophagosome–lysosome fusion. Autophagy is important in the clearance of the small, oligomeric precursors into larger protein aggregates that appear to be particularly harmful in neurodegenerative disease. Dynein, as previously mentioned, also mediates the formation of larger protein aggregates, known as aggresomes, which it transports toward the MTOC. In contrast to smaller aggregate precursors, aggresomes appear to play a protective role in neurodegenerative disease.

Disruption of the function or regulation of individual Rabs may contribute to some neurological diseases. Charcot–Marie–Tooth syndrome (CMTS) and Niemann–Pick disease involve inhibition of the bidirectional transport and elevated cholesterol levels of late endosomes. Symptoms of these diseases include impaired motor skills, loss of nerve sensitivity, and paralysis. Cholesterol upregulates the Rab7 cycle, in which Rab7 induces the recruitment of dynein–dynactin to late endosomes. Forms of CMTS are caused by mutations in either Rab7 or the plus end-directed KIF1B motor protein that impairs anterograde transport of cholesterol-containing endosomes, leading to defects in cholesterol metabolism. Mutations in ALS2 can cause amyotrophic lateral sclerosis (ALS), and ALS2 binds Rab5 *in vitro* and functions as an Rab5 guanine exchange factor. In addition, ALS2 associates with endocytic membranes, further supporting a role for ALS2 in endocytic vesicle formation and trafficking and suggesting that ALS symptoms could partially result from a disruption of Rab5-dependent endosome transport.

In addition to its involvement in disease, dynein-mediated retrograde endosomal signaling is important in the response to nerve injury. Dynein binds proteins such as importins that link it to a variety of signals involved in nerve injury response pathways. Importins bind proteins with nuclear localization signals, including transcription factors and members of the microtubule-associated protein (MAP) kinase family, transporting them to the cell body and toward the nucleus through their association with dynein. In addition, scaffold proteins involved in the endosomal pathway, such as vimentin and JIP3, are also implicated in the connection between importins and signaling molecules. Vimentin binds both the MAP kinase Erk and importin-α, and JIP3 and its activated binding partner, Jnk, interact with both dynein and kinesin-1 and are associated with vesicles in axons following injury. The transport of this injury-signaling complex results in various cellular responses, including the downregulation of neurotrophic signaling.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axonal Transport Disorders; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and Neurodegenerative Diseases; Microtubules: Organization and Function in Neurons; Myosin Transport and Neuronal Function; Nerve Growth Factor; Presynaptic Endosomes.

Further Reading

Alberts B, Johnson A, Lewis J, et al. (2002) *Molecular Biology of the Cell*. New York: Garland.

Brown CL, Maier KC, Stauber T, et al. (2005) Kinesin-2 is a motor for late endosomes and lysosomes. *Traffic* 6: 1114–1124.

Campenot RB and MacInnis BL (2004) Retrograde transport of neurotrophins: Fact and function. *Journal of Neurobiology* 58: 217–229.

Caviston JP and Holzbaur ELF (2006) Microtubule motors at the intersection of trafficking and transport. *Trends in Cell Biology* 16: 530–537.

Grosshans BL, Ortiz D, and Novick P (2006) Rabs and their effectors: Achieving specificity in membrane traffic. *Proceedings of the National Academy of Sciences of the United States of America* 103: 11821–11827.

Gunawardena S and Goldstein LS (2004) Cargo-carrying motor vehicles on the neuronal highway: Transport pathways and neurodegenerative disease. *Journal of Neurobiology* 58: 258–271.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews Neuroscience* 6: 201–214.

Hoepfner S, Severin F, Gillooly D, Stenmark H, and Zerial M (2005) Modulation of receptor recycling and degradation by the endosomal kinesin KIF16B. *Cell* 121: 437–450.

Krendel M and Mooseker MS (2005) Myosins: Tails (and heads) of functional diversity. *Physiology* 20: 239–251.

Pollard TD and Earnshaw WC (2004) *Cell Biology.* Philadelphia: Saunders.

Ridley AJ (2006) Rho GTPases and actin dynamics in membrane protrusions and vesicle trafficking. *Trends in Cell Biology* 16: 522–529.

Russell MR, Nickerson DP, and Odorizzi G (2006) Molecular mechanisms of late endosome morphology, identity and sorting. *Current Opinion in Cell Biology* 18: 422–428.

Salehi A, Delcroix J, and Mobley WC (2003) Traffic at the intersection of neurotrophic factor signaling and neurodegeneration. *Trends in Neuroscience* 26: 73–80.

Vaughan KT (2005) Microtubule plus ends, motors, and traffic of Golgi membranes. *Biochimica et Biophysica Acta* 1744: 316–324.

Zerial M and McBride H (2001) Rab proteins as membrane organizers. *Nature Reviews Molecular Cell Biology* 2: 107–117.

Machine Haptics

M A Srinivasan, Massachusetts Institute of Technology, Cambridge, MA, USA
R Zimmer, University of London, London, UK

© 2009 Published by Elsevier Ltd.

Introduction

Modern haptics can be defined as the science, technology, and applications associated with information acquisition and object manipulation through touch. Within this broad multidisciplinary field, human haptics is concerned with manual exploration and manipulation by humans, whereas machine haptics is concerned with touch-based interactions mediated by electromechanical and computational devices. Haptic interfaces enable users to engage manually with things and people that are either virtual or distant. This may seem like science fiction, but it is in many ways not dissimilar from visual interfaces, which are now so commonplace as to seem completely natural. In visual interfaces, an image of a real or imagined object or medium is rendered graphically and the digital representation is available to be manipulated by a computer, sent from computer to computer, or displayed on screens; with haptic interfaces, a digital representation of the feel – the three-dimensional contour, hardness, and smoothness – of an object is conveyed to the user, which can be manipulated, transported, and available to different kinds of displays that are described later. Haptic interfaces seem less natural to us than visual interfaces partially because they are not yet prevalent and partially because haptic interfaces offer real differences from our accustomed ways of interacting with computers, some of which follow from the unique properties of the sense of touch.

Touch is the most immediate of the primary senses (touch, sight, and audition) with which we apprehend the world. Unlike the things we see and the things we hear, which, even when not apprehended through machines, are always at some remove from us, the things we feel are things that are in contact with us, things that are touching us as we touch them. Once immediate senses, such as touch and taste, are added to the engagement with computers, the experience becomes manifestly more immediate, more participatory, more part of a real world, and more immersive. Moreover, touch is the most consciously active, and most symmetric, of our senses. The things we feel are, often, things that we are actively exploring through our hands and are always things that can be affected by that exploration. Therefore, machine haptics, to a much larger extent than visual or aural interaction with computers, has this dual nature, consisting of the following:

- *Haptic displays*. Devices that provide simulations of the feel of objects during sensory-dominant manual exploration and motor-dominant manipulation. The use of these devices is parallel to that of a computer screen or an audio speaker.
- *Haptic sensors*. Devices that respond to information gained from the motions and/or forces of users' interactions. These are used to gather information about a user's manipulations similar to the computer mouse and in some ways are parallel to scanners and cameras that gather visual information about objects, or microphones that gather sound. However, haptic sensors are used to infer user's intention in interacting with objects and models of objects and interactions are built to take advantage of this information.

A full machine haptic system would incorporate both of these aspects and their interaction. For example, imagine a physical interface that worked as a kind of virtual plastic modeling material, such as plasticene or clay whose material behavior can be controlled through a computer program. Suppose its shape, texture, and hardness could be continually updated using a computer model based on the forces applied by the user's hands. That recomputed model would then affect what the user feels through the haptic display, thus enabling different object feels through the same physical interface.This scenario is beyond the scope of current technology but will eventually be realizable. It is important to note how this kind of system differs from vision. A full visual system may have a camera to take visual input and a screen to give graphical output, but the very act of the user looking at the screen does not provide much information to the system. The haptic interface is more similar to a head-mounted visual display in which the position and orientation of the user's head drives the graphical image.

Machine haptics can involve both the stimulation and the simulation of the human haptic system, as represented by two broad categories of systems in which haptic machines are used: virtual environments (VEs), in which the human interacts with a computer-based synthetic environment, and teleoperation (TO), in which the human can operate on the real environment at a distance through interface devices.

In the case of VE systems, the computer represents a VE and there are essentially two actors in the system: the human and the computer. Figure 1 shows the

relationship between the machine and the human in a haptics-enabled VE. The central box contains the tip of a haptic interface device which the user grasps. The left-hand loop of the picture shows the human interaction with the device. The human senses, through his/her contact with the device, forces to which he/she responds by providing motor signals that move his/her hand, which in turn affects the device. As shown in the right-hand loop, the computer attached to the device senses the movement of the device and responds by sending commands that create the forces the device applies on the human. Both the human and the computer systems have sensors (receptors and encoders), processors (brain and computer), and actuators (muscles and motors). That is, the machine haptic system mirrors and extends the human haptic system.

In the case of TO systems, the haptic interface is connected to a computer as previously discussed, but that computer, in turn, is connected to another computer, either directly or over a network such as the Internet, and haptic signals are sent back and forth between the two. This allows a user of the first computer to interact with a real environment through an interface device such as a robot that is connected to the second computer. This kind of TO system is summarized in Figure 2.

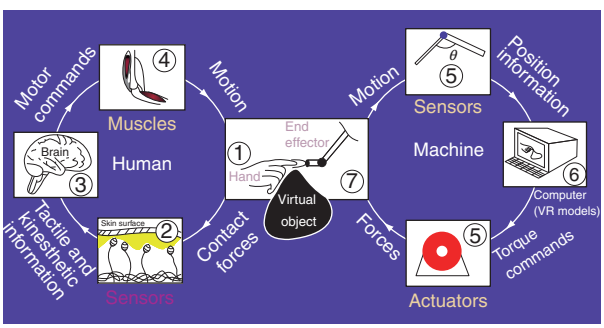

Figure 1 Human–machine haptic interaction in virtual environments.

More complex haptic interactions are possible with a system that combines VEs with TO to enable two-way interactions across a network. One case of this was in 2002 when people at the MIT Touch Lab shook hands, transatlantically, with people at University College, London. In the future, haptic interactions at a distance will be prevalent, powerful, and important. These systems bring with them several new technical challenges (e.g., dealing with the latency of the transmitted signals) as well as new kinds of applications. However, for the sake of ease of explication, this article is confined to the more simply explained VE paradigm until the applications section, in which emerging practical uses of VE and TO systems are discussed together. The section above brings out at least two reasons why a neuroscientist might be interested in machine haptics. First, machine haptics provides a particular way of doing neuroscience: having unprecedented control over touch stimuli displayed through haptic interfaces allows experimentation with the human haptic sensory system, the human motor system, and interactions between the two, as well as with other sensory systems such as vision and hearing. Second, machine haptics provides an application area for neuroscience: bringing neuroscientific knowledge of the human haptic system to engineering design will help create improved haptic interfaces in the future. The development of engaging, compelling haptic interfaces will require an understanding of the roles played by the mechanical, sensory, motor, and cognitive subsystems of the human haptic system. Therefore, machine haptics, both as a research topic and as a branch of engineering, in its search to provide natural touch interfaces for computers, requires an understanding of human haptics and an understanding of the psychology of human–machine and human–human interactions in VE and TO systems, as well as the technological questions of hardware and software.

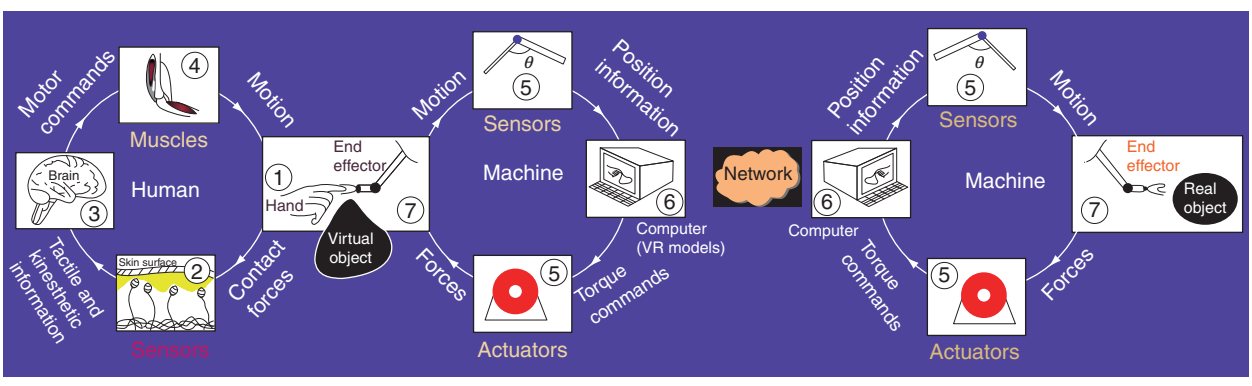

Figure 2 Human–machine haptic interaction in teleoperated environments.

However, because human haptics is treated in another article in this encyclopedia, we concentrate here on the technological aspects, providing discussions of how far the technology (both hardware and software) has already come in providing effective interfaces, what will make more engaging interfaces in the future, and current and potential application areas for these technologies.

What Needs to Be Simulated?

Although haptics includes feeling by other parts of the body as well, including tongues and backs, the vast majority of haptic machines, to date, concerns interfacing with hands. The goal of this work is to create systems that will respond to a user's interactions in ways that are similar to, and ultimately indistinguishable from, real objects when they are explored by a user's hands, either directly or through the use of a tool. The properties of an object that need to be modeled and simulated, therefore, are exactly those properties that people explore with their hands. These include shape, size, and possibly weight, but also surface properties of the material, such as roughness and compliance, as well as thermal properties.

Moreover, the interface, if it is to get input from the user's interactions, must also model the user. That is, there must be pressure exerted by the user on the object and even the user's movements when no contact is made. Imagine the situation of picking up an object and moving it. To be able to sense this, the system must be able to track (1) free motion, in which no physical contact is made with objects in the environment; (2) contact involving unbalanced resultant forces, such as pressing an object with a finger pad; and (3) contact involving self-equilibrating forces, such as squeezing an object in a pinch grasp. The design of a haptic interface depends fundamentally on which of these two last elements you are most interested in simulating.

Haptic Interfaces: How Can Touch Be Simulated?

To give a sense of what machine haptics is about, we describe the basic operations of two kinds of haptic displays, and one idea for a haptic input device, before generalizing to map out a taxonomy of machine haptics.

Tool-Based, Force Feedback Haptic Displays

This kind of display aims to simulate the experience of touching an object with a penlike tool, or stylus. The user holds the stylus and moves it, while the movement is impeded in a way that simulates probing the displayed object with a penlike object in the real world. Figure 3 show one such device, SensAble Technologies' Phantom.

Figure 3 The Phantom haptic interface.

The stylus is attached to the tip of what is essentially a small robot arm through three rotary joints. The robot arm has three additional rotary joints, each controlled by an independent motor. The device exerts forces on the hand through the stylus by sending appropriate currents to the motors, controlling how easily each of the joints can rotate. In this way, force on the user can be approxiate currents to in any direction. The computer that is connected to the force-feedback interface takes readings, roughly once every millisecond, from all the six joints. Using this information, the computer can work out the position and orientation of the stylus to determine where the stylus maps onto the virtual world and its proximity to the virtual object being touched.

When the virtual position of the stylus collides with one of the objects in the virtual world, one or more of the motors stiffen a joint or several joints to exert a force on the user's hands that simulates the force that the modeled object would have exerted had the user collided with it in the real world. By moving the stylus around the space and probing until a force is felt, a user is essentially exploring the space, getting a sense of the contour, hardness, and texture of the objects. The user is prevented from penetrating the virtual object in a way that mimics the solidity of the physical object that is represented, unless penetration is desired, as in the case of virtual injections or surgery.

Tactile Display Arrays

Whereas the device described previously simulates the effect of exploring a space or an object with a tool, tactile arrays re-create the sense of exploring with fingers. This clearly calls for much finer distinctions and more delicate ways of imparting information. There are various technologies for trying to achieve this, including shape-changing displays, which convey the local shape of contact by controlling the deformation or forces distributed on the skin; vibrotactile displays,

which deliver mechanical energy through an array of vibrating pins placed against the skin; and electrotactile displays, which stimulate the skin through surface electrodes. Figure 4 shows one tactile display device developed at the MIT Touch Lab.

The display shown in Figure 4 has 30 pins, separated by about 2 mm. Each pin is capable of providing flexible mechanical stimulation in a wide range (static to ~325 Hz) with a normal displacement (~700 μm), large enough to feel even at low frequencies. This kind of device requires a finer modeling than tool-based systems. Here, small mechanical signals are imparted to the fingertips in an attempt to stimulate the same neural signals from the fingers to the brain as would be sent if the finger had been in contact with an object with particular material properties. This device has been shown to successfully display a wide variety of textures that are perceivable by human subject, although the current technology is limited. This technology is not yet mature and it does not permit spatial resolution that is as fine as that of the human fingertip, where mechanoreceptors are densely packed. When a user is attending closely to the display, he or she can still feel the individual pins, but this is a technical limitation that is expected to be overcome in the near future.

Haptic Sensing: Tactile Sensor Array

The two already-described devices are both displays providing outputs to the user so that he or she can feel an object. There are also haptic interactions in the other direction in which the computer gets information from the user's touch. A computer mouse is a haptic sensing device. The user touches the mouse, pressing buttons and/or moving the mouse's position, and the computer responds to those actions. Compared to the hand, the mouse is a crude device which only responds in a binary way to button pressing. There are much more sophisticated ways of reading the touch of a hand. One such is an array of pressure sensors from pressure profile systems, as shown in Figure 5.

The array contains a soft middle layer, surrounded by a layer of electrodes on either side. Two more soft layers enclose the electrodes to form the outer skin at the top and bottom. Pressure exerted on any part of the array alters the capacitance in the local area by decreasing the distance between the two electrode layers. In this way, a profile of how much a user is pressing each part of the material can be computed. If, for example, a material containing this kind of array was used to cover a joystick, then the system could gather accurate information about what parts of the joystick were being squeezed, and with what force, at any time, allowing for a quite intricate form of computer control. Tactiles sensor arrays are also used in robotics, when a robot is fitted with an array as a way of gathering information about the world it is exploring.

The two display and one haptic input technologies we have described are just three of many. In the next section, we put these in context by creating a general taxonomy.

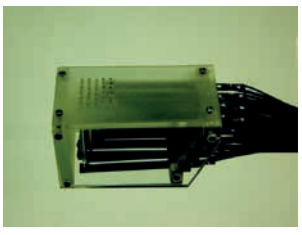

Figure 4 Tactile display device.

Classification of Haptic Technologies

We have already discussed the fact that haptic interfaces come in several varieties. In this section, we enumerate some of the properties that differentiate haptic interfaces from one another and give an account of what kind of devices satisfy some of the combinations of attributes.

Figure 5 Array of pressure sensors.

Displays or Sensors

As already discussed, some devices are primarily displays, giving haptic information to the user, whereas others are sensing devices that take readings from the user's haptic interactions and interpret them as input data for the computer. In mechanical engineering terms, a system that puts out motions and forces is called 'active' and a system that only measures motions and forces is said to be 'passive.' In these terms, pure input devices are passive and haptic displays are necessarily active. Previously, we described the hypothetical medium of virtual clay, which we might think of as an active device.

Tool-Based or Direct Touch

Some haptic interface systems work by direct contact and others by simulating the use of tools. The directly felt interfaces could be meant for hands or fingers, but they also may be wearable devices that give signals directly to (and/or receive signals directly from) other parts of the body. A perfectly realistic interface would allow both tool-based and direct simulation.

Ground-Based or Body-Based Interface Device

Haptic interfaces may be stand-alone devices resting on a surface external to the user or they may be worn by the user. Since the forces applied by the user on the device in manipulating it need to be balanced by equal and opposite ground forces, the path of how these forces are eventually transmitted to the ground – whether directly through the device or through the user's body – strongly influences the user's haptic experience.

As each of these attributes takes on one of two values, their set leads to a taxonomy of eight classes, all of which are theoretically possible. Each of the previously discussed attributes is described with the more easily achieved value first, so the easiest kind of interface to develop is one that is passive (i.e., an input device), tool-based, and ground-based. A computer mouse is such a device: it is an input device, it responds to your touch, responding as a tool would, and it is based on the desktop. So is a keyboard and the simplest joysticks. However, the case of joysticks can be more complicated because even the earliest of these, developed to control aircraft and connected by mechanical wires to the flight surfaces of the aircraft, unwittingly presented force information about loads on the flight surfaces to the pilot, so they were a sort of active interface device. Some joysticks currently available for games purposely give force-feedback information to the user. The PHANTOM, for example, is a ground-based active device capable of simulating tool-based touch interactions.

In terms of body-based devices, a wearable one with sensors that track your body's interactions with real or imagined surfaces but do not send you signals is a passive, body-based interface. For example, there are wearable haptic gloves, such as the CyberGlove from Immersion Corporation, which have joint angle sensors that can track hand posture. A glove like this, when coupled with some kind of spatial tracking of the whole hand, provides an interface that keeps track of the user's hand position in free space as well as his/her finger positions during interaction with virtual objects. Some of the earliest body-based systems developed for teleoperation, however, were already active. These were exoskeletal devices that fit over and move with the limbs or fingers of the user. An example is the CyberGrasp from Immersion Corporation, which uses a CyberGlove to track hand and finger posture and the exoskeletal structure to transmit forces to the fingertips (Figure 6).

Because body-based devices can move as much as arms and hands do, they can typically monitor and stimulate over a much wider range of unrestricted user motion than can be provided by ground-based devices. However, they are awkward to wear, and providing high-quality force feedback with such devices that is commensurate with human resolution is proved to be difficult and places great demands on actuator size minimization and control bandwidth.

On the other hand, with a tactile display embedded in an item of clothing or glove, people can interact with computers by receiving information through feeling. This can provide a much more compelling sensation of being immersed in a virtual environment than sound and vision alone. With this kind of wearable display, for example, you could virtually bump into objects and people, or feel the wind brush past you. While some limited demonstrations have been achieved, this type of technology is not yet ready for such highly realistic and immersive sensation, but its

Figure 6 The CyberGrasp.

potential is vast. Human tactile perception is robust and our skin provides approximately 20 square feet of touch receptors. There are already prototype systems based on a variety of tactile stimulation and sensing technologies and these can be very valuable. Even without the difficult application of simulating bodily engagement with a whole world, wearable tactile displays provide a channel for information gathering that is complementary to audio and visual signals and can be unintrusive and natural. They are discrete and seamlessly integrate with most human activity. Tactile displays will neither conflict with nor replace audio and visual displays but, rather, support information on these other displays and fill in the gaps where necessary.

In fact, the perfect haptic interface would be both input and output, modeling the user's actions as it provides information to him/her, and also simulate both touching with hands directly and interacting with tools. Such an interface would seamlessly measure the position and posture of the user's hand and send forces to the hand. However, the large number of degrees of freedom of the hand, the extreme sensitivities of cutaneous receptors, together with the presence of mass, friction, and limitations of sensors and actuators in the interface, make such an ideal impossible to achieve with current technology. The technology will eventually mature so that such a system is possible.

Haptic interfaces will always only offer approximations of real interactions, just as do any of the current visual and auditory interfaces. However, that does not mean that haptic interfaces will always feel unconvincing. Much thought has gone into the way in which a sequence of discrete images can be made to appear visually continuous, relying on the persistence of vision (which can be thought of as a limitation on the acuity of human vision). We expect that there are similar approximations that will take advantage of the limits of human haptic acuity to generate synthetic haptic experiences that are indistinguishable from real lived experience. To develop these approximations will require quantitative human studies. Basic understanding of the biomechanical, sensorimotor, and cognitive abilities of the human haptic system is critical for proper design specification of the hardware and software of haptic interfaces.

In the next section, we discuss applications of the present and future technologies.

Applications

There are two main modalities of applications of haptic interfaces: one concerns interfaces to VEs that only exist with a computer, and the other concerns TO, taking haptic and other information from and/or making its own distant presence felt on a nearby or far-away real environment. Many of the application areas have instances of both as we shall see. A few of the many potential applications are described in the following.

Medicine

One of the very active areas for haptic applications is in the development of virtual reality-based systems for training medical personnel. Here, the computer acts as a virtual human body, both graphically and haptically. For example, the trainee surgeon sees a screen containing a picture that might resemble a televised picture of a human body being operated upon while the surgeon moves his/her scalpel and feels, using a force-feedback output device, a simulation of probing and cutting the body. This is not necessarily an inferior replacement for the cadavers, which are currently relatively rare in medical schools, but it could in principle provide a more accurate feel than either a deceased human or a different live animal. A TO version of this application could not only train surgeons at a distance but also improve the interface for real surgery performed from a distance. Such a system could filter out the natural hand tremors of a surgeon and the system could be designed such that the macroscopic motions of the surgeon are translated into microscopic motions of the robot performing surgery on a patient. Similar medical applications include micro and macro robots for minimally invasive surgery and remote diagnosis of diseases.

Communication across the Internet

The idea of seeing people over the Internet, through the use of webcams as sensory devices and computer screens as display devices, is now commonplace. Much the same thing could be done with the kind of input and display devices we have described above, as was done during the transatlantic touch demonstration described previously. This could be the harbinger of a future version of the Internet, in which remote physical contact plays a large part.

Psychological Studies of Perception

Having a virtual world in which the haptic aspects are under the control of the experimenter could be as beneficial to studies of haptic perception, and perception in general, as virtual worlds have been in studies of visual perception. The TO version of this could enable studies to be done collaboratively by researchers in different laboratories, potentially in very different parts of the world.

Entertainment

Haptic interfaces can make video games and simulators more engaging by adding another dimension to their interaction. There are already force-feedback joysticks in use in various games, allowing players to feel as if they are really shooting guns or flying planes. These will be replaced by much more accurate, more sensitive, and more engaging force and tactile feedback interfaces.

Education

Haptic interfaces can help us feel otherwise intangible worlds, such as nano-, macro-, or astronomical-scale worlds. For example, a project in the United States is exploring how the addition of haptic feedback to computer-generated 3D virtual models of an animal cell influences students' understandings of cell concepts. The Haptic Cell Exploration instructional program contains a virtual model that depicts the 3D nature and spatial arrangement of an animal cell, including its typical parts (organelles), and allows students to feel the various parts of the cell.

Training

Tactile sensory and display arrays could be used to train people in a very wide variety of tasks, such as how to hold tools and to operate equipment correctly. The arrays could send the information to a computer that would be able to inform the user how well he or she is doing.

Art and Design

Haptic interfaces could be integrated into CAD systems so that designers could manipulate their designs. Similarly, sculptures could be created by capturing the results of virtual manipulations and sending the results to three-dimensional printers. In a TO mode, this could lead to cooperative virtual sculpting across the Internet.

Access to Art and Artifacts

Several experimental museum and gallery exhibitions in which visitors used technology to feel simulations of a museum's collections have occurred already. In the long term, it may be possible to access the feel of a museum's collection over the Internet as easily as we can now access visual images. Moreover, there can be virtual art exhibits, concert rooms, and museums in which the user can log in remotely to, for example, play the musical instruments as part of the concert.

Aids to Overcome Blindness and Other Disabilities

Tactile vibrating devices have long been used to convey speech information to people who are deaf. New haptic displays could be used to provide information from the Internet in ways that are accessible to people for whom the Internet or museum collections are inaccessible, for example. They have even been used as training systems in the rehabilitation of individuals with motor impairment.

See also: Human Haptics; Proprioception; Tactile Texture.

Macroglial Lineages

A M Butt, University of Portsmouth, Portsmouth, UK

© 2009 Published by Elsevier Ltd.

Introduction

Macroglia together with neurons make up the two classes of neural cells in the central nervous system (CNS); both macroglia and neurons have a common neuroectodermal embryonic origin. The three major subtypes of macroglia are astrocytes, oligodendrocytes, and the ependymal cells. The fourth major glial cell type in the CNS is the microglia, which are nonneuronal cells that have a mesodermal origin; microglia originate from macrophages that invade the brain during early development. There are also many glia in the peripheral nervous system (PNS), mainly Schwann cells, together with satellite cells of sensory and sympathetic ganglia and glial cells of the enteric nervous system of the gastrointestinal tract. CNS and PNS neural cells have clearly distinct lineages from an early stage of embryonic development.

Astroglia and oligodendroglia are the most numerous and functionally important macroglial cell types; ependymal cells are simple cuboidal cells that line the ventricles in the brain and the central canal in the spinal cord. Astrocytes are morphologically and functionally diverse and have a widespread distribution throughout the brain and spinal cord; astrocytes have the archetypal feature of expressing glial fibrillary acidic protein (GFAP). The two major classes of astrocytes are protoplasmic astrocytes and fibrous astrocytes, which are found in the gray and white matter, respectively. In addition, there are specialized astroglial cells throughout the CNS; for example, radial glia are bipolar astroglial cells that predominate in the developing brain, and Müller glia and Bergmann glia are specialized astroglia restricted to the retina and cerebellum, respectively. Unlike astroglia, oligodendroglia are highly specialized cells and consequently have a more restricted distribution; oligodendrocytes form the myelin sheaths that insulate axons in the CNS and are most numerous in white matter. Until recently, the accepted model was that macroglia and neurons were derived from separate glial and neural progenitors. However, glial and neuronal lineages are much more closely related than was previously thought. Astrocytes and oligodendrocytes develop from multipotent neural stem cells that also generate neurons. Indeed, astroglial stem cells and oligodendrocyte progenitor cells (OPCs) persist in the mature brain and are potential sources of new neurons in the adult.

Neuroepithelium

Neurons and macroglia derive from the neuroepithelium, which at the earliest stages of development forms the neural plate. As the neural plate thickens, the neural groove forms and the neural plate closes dorsally to form the neural tube, the lumen of which becomes the ventricular system. The cells lining the ventricles of the neural tube form a primary proliferative zone termed the ventricular zone (VZ), and a secondary proliferative zone called the subventricular zone (SVZ) emerges later from the VZ. The VZ ultimately disappears, but the SFV persists into adulthood within the forebrain. As the neural tube expands, the rostral area delineates into the forebrain, midbrain, and hindbrain, and the caudal neural tube forms the spinal cord. During this regionalization, neural crest cells differentiate from the neural tube and migrate to specific regions to form the neurons and glia of the PNS. Thus, CNS and PNS neural cells have clearly distinct lineages from an early stage.

Radial Glia Are Neural Stem Cells

The neuroepithelial cells of the VZ can be considered multipotent neural stem cells in the sense that their progeny gives rise to all of the neurons and macroglia in the CNS. The first cells that emerge from the neuroepithelium are the radial glia. These are distinguished from neuroepithelial cells by the expression of a number of antigens, including the astroglial markers GFAP, vimentin, and the calcium-binding protein S-100β, as well as the glutamate transporter GLAST, nestin, and tenascin-C. The somatas of radial glial cells are located in the VZ and their processes extend to the opposite wall of the neural tube or the pial surface. Radial glia are the main neural stem cells during development and are responsible for most of the subsequent neurogenesis, giving rise first to neurons and later to astrocytes and some oligodendrocytes by a process of 'transdifferentiation.' The majority of oligodendrocytes, however, originate at an early stage of development from glial precursors that are generated in specific sites in the brain and spinal cord. Astrocytes are generated later in development, both from radial glia and from glial precursors that also give rise to oligodendrocytes; the proportion of the final population of astrocytes derived from radial glia and glial precursors depends on the region of the CNS. Radial glia not only produce neurons and glia but also form a scaffold along which newborn progenitors migrate from the SVZ. Moreover, the radial glial cells and astrocytes that differentiate from them retain the function of stem cells in the adult brain.

Glial Restricted Precursors

Multipotent neural stem cells (neuroepithelium) do not differentiate into astrocytes and oligodendrocytes directly but, rather, through intermediate lineage-restricted stages. The first of these to develop are the glial-restricted precursors, which are distinguished from neuroepithelial cells by expression of A2B5 and nestin and are localized to the ventral neural tube. Several stages in the development of astrocytes and oligodendrocytes have been distinguished, with the main ones being oligodendrocyte and type-2 astrocyte precursors (O2A cells), oligodendrocyte precursors, astrocyte precursors, and NG2-expressing glia. The first stage in the differentiation of oligodendrocytes and probably astrocytes is the bipotential O2A glial precursor, which *in vitro* can generate either type of glial cell; O2A cells are distinguished by expression of platelet-derived growth factor-α receptors (PDGF-αRs) and NG2, in addition to A2B5. It is uncertain to what extent O2A cells are bipotential *in vivo*, where they may serve predominantly as OPCs which generate only oligodendrocytes. OPCs express the early oligodendrocyte lineage markers PDGF-αR and NG2 and later the homeobox transcription factor Nkx2.2 and the oligodendrocyte lineage gene Olig2 (basic helix–loop–helix factor). Astrocyte precursors have been isolated from numerous regions of the embryonic CNS; they express A2B5 but not GFAP until a relatively late stage of differentiation, and they are therefore difficult to distinguish from glial precursors and O2A cells. Expression of the extracellular matrix transmembrane protein CD44 may serve as an astrocyte precursor marker. Early glial precursors are small cells, with one or more processes, and are highly mobile, migrating from multiple sites to colonize the entire CNS white and gray matter. They do not express early glial antigens such as PDGF-αR and NG2 until they exit the SVZ, and as they migrate, they begin to acquire markers of OPCs and astrocytes. The relative contributions of O2A cells, OPCs, and astrocyte precursors to gliogenesis *in vivo* are not clear; they display many similarities and may represent different stages in glial precursor differentiation or distinct lineages that arise separately from neural precursors. Indeed, at least *in vitro*, glial precursors are not truly lineage restricted in that they can generate neurons, astrocytes, and oligodendrocytes under appropriate culture conditions. In addition, after generating oligodendrocytes during development, a significant population of NG2 glia persists throughout the white and gray matter of the adult brain; the functions of adult NG2 glia are unresolved, but they serve in part as a pool of adult OPCs and they may even be multipotent adult neural stem cells. There are also GFAP-positive astrocyte-like neural stem cells in specific regions of the adult CNS, and these are multipotent and self-renewing.

Regional Variations

In the forebrain, glial precursors in the SVZ migrate along radial glia to their final sites in the white matter and cortex to generate oligodendrocytes and some astrocytes, in addition to adult NG2 glia. Astrocytes are derived from migratory glial precursors and from radial glia. In early forebrain development, oligodendrocyte precursors are generated in localized sites ventrally and migrate dorsally. Later, glial precursors that originate from a large dorsolateral SVZ near the lateral ventricle give rise to oligodendrocytes and some astrocytes. Oligodendrocytes and the diverse astrocyte cell types in the cerebellum are derived predominantly from migratory glial precursors from an area dorsal to the fourth ventricle; Bergmann glia arise from embryonic radial glia and share a common lineage with Purkinje neurons. In the embryonic retina, Muller glia arise from a common precursor with neurons, and astrocytes develop at a later stage from glial precursors that migrate into the retina via the optic nerve; oligodendrocytes are absent from the retina of most species. Most astrocytes in the optic nerve are derived from intrinsic radial-like astroglia, whereas oligodendrocyte precursors migrate into the nerve during the perinatal period. In the spinal cord, oligodendrocytes originate from neural stem cells in the ventral neuroepithelium, which first generate neurons and then oligodendrocyte precursors. Later, oligodendrocyte precursors also originate from dorsal areas. Spinal cord astrocytes appear to be derived from separate astrocyte precursors after neurons and oligodendrocytes; as in other areas of the CNS, astrocytes most likely also arise from radial glia.

Gliogenesis in the Adult Central Nervous System

Gliogenesis occurs at a slow rate throughout the adult CNS. New glial cells are born locally, predominantly from parenchymal multipotent glial precursor cells and also, in the case of astrocytes, from division of mature cells. The type of cell produced is determined by locality of the precursors; they mostly produce oligodendrocytes in white matter, whereas astrocytes and oligodendrocytes are produced in the same quantities elsewhere. These glial precursors do not express GFAP, but they do express nestin and NG2. It is not clear that all NG2-expressing glia in the adult CNS are glial precursors. NG2 glia are slowly proliferating

cells that can generate oligodendrocytes *in vivo* and can also generate astrocytes and neurons *in vitro*. Nonetheless, the substantial majority of NG2 glia in the mature CNS appear to be nonmitotic, but like astrocytes, they may retain the function of stem cells in the brain throughout maturation and adulthood. In addition, self-renewing and multipotent neural stem cells are widespread in the ventricle walls of the adult CNS. These neural stem cells have morphological, physiological, and biochemical/immunological characteristics of astrocytes; those located in the hippocampus and olfactory subendyma generate neurons throughout life. The relative contributions of parenchymal NG2 glia and 'stem' astrocytes to gliogenesis in the adult remain to be resolved.

See also: Astrocyte: Identification Methods; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Microglia Properties; Neural Stem Cells and CNS Diseases; Neural Stem Cells: Adult Neurogenesis; Oligodendrocyte Morphology; Oligodendrocyte and Schwann Cell Identification Methods; Radial Glial Cells: Brain Functions.

Further Reading

Alvarez-Buylla A, Garcia-Verdugo JM, and Tramontin AD (2001) A unified hypothesis on the lineage of neural stem cells. *Nature Reviews Neuroscience* 2: 287–293.

Butt AM (2005) Synantocytes: The fifth element. *Journal of Anatomy* 207: 695–706.

Doetsch F (2003) The glial identity of neural stem cells. *Nature Neuroscience* 6: 1127–1134.

Goldman JE (2005) Lineages of astrocytes and oligodendrocytes. In: Kettenmann H and Ransom BR (eds.) *Neuroglia*, 2nd edn., pp. 72–84. New York: Oxford University Press.

Magnetic Resonance Spectroscopy

J R Alger, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Magnetic resonance spectroscopy (MRS) is an analytical tool that detects radio frequency electromagnetic signals that are produced by the atomic nuclei within molecules. It can be used to obtain *in situ* concentration measures for certain chemicals in complex samples, such as the living brain. In certain specific circumstances, MRS also provides means of performing kinetic analyses of dynamic neurochemical processes.

MRS is essentially identical with nuclear magnetic resonance (NMR) spectroscopy, which has been a part of chemistry and physics for the past half century. MRS is the specific term that is used when an NMR spectroscopy procedure is applied in a biological or medical context. Terminology that eliminates 'nuclear' was adopted for medical and biological NMR procedures to avoid the implication that nuclear radioactivity is somehow involved in the procedure.

MRS overlaps to a very great extent with magnetic resonance imaging (MRI). MRS studies usually involve the detection of NMR signals produced by chemical compounds other than water, with the goal generally being to evaluate *in situ* biochemistry, whereas MRI usually detects NMR signals produced by the tissue water, with the goal being to make images that depict the macroscopic anatomy of soft tissues such as the brain.

MRS as a Neuroscience Tool

The development of MRS as a neuroscience tool grew from several key advances in science and technology that have occurred over the past few centuries. These include Fourier's eighteenth-century techniques for working with complicated periodic mathematical functions, nineteenth-century descriptions of electricity and magnetism by Faraday, Ampere, and Maxwell, the early twentieth-century discovery of quantum mechanics by Planck, Heisenberg, and Born, the description of the basic magnetic properties of atomic nuclei by Stern, Rabi, Bloch, and Purcell in the 1930s and 1940s, and the elucidation of intermediary metabolism beginning with the work of Krebs and Lippman in the 1940s. Moreover, MRS would not be possible without twentieth-century technological advances in radio communication, digital computing, and superconducting magnets. In 1991, Richard Ernst was awarded the Nobel Prize in Chemistry for work done in the 1960s and 1970s that directly led to the development of MRS techniques that are used today for neuroscience research and for routine evaluation of patients. In 2003, Sir Peter Mansfield and Paul Lauterbur were jointly awarded the Nobel Prize in Physiology or Medicine for discoveries related to the imaging procedures that are now used in MRI and MRS.

This convergence of science and technology advances has made it possible to perform MRS studies in the brains of animals and humans since the early 1980s. Other parts of the body can also be studied by MRS, but the brain has been the most well-explored organ system because it has fortuitous chemical, physical, and structural properties that make it readily amenable to MRS study. During the 1980s and 1990s many studies established MRS as a valid neuroscience tool. This validation work is still proceeding in concert with further technological development of MRS. It is now possible to routinely perform MRS examinations of clinical brain disorders in human patients. Although the preponderance of recent MRS developmental emphasis has been on clinical applications, MRS also has value for addressing fundamental neuroscience questions. MRS studies of animals, such as genetically altered rodents, are feasible as are studies of cultured cells, histological tissue samples, and chemical extracts of neurologically significant tissues. The section titled 'Further reading' at the end of this article provides an introduction to full scope of neuroscientific MRS applications, including clinical and basic science applications.

Physical Principles of MRS

MRS utilizes the fact that certain atomic nuclei behave as if they were spinning bar magnets. The nuclear magnetic fields interact with a magnetic field that is produced by a strong magnet in which is placed the sample that is being examined. Through this magnetic interaction, energy is exchanged between the nuclear magnetic fields and the externally applied field. Energy exchange of this sort is known in physics and engineering as 'resonance.' In physics and chemistry, 'spectroscopy' is used to describe the study of the specific frequency (or wavelength) at which resonance occurs. In the typical MRS study, the magnetic resonance has a time dependence that is analogous with the electromagnetic oscillation of the radio waves that are used in radio and television

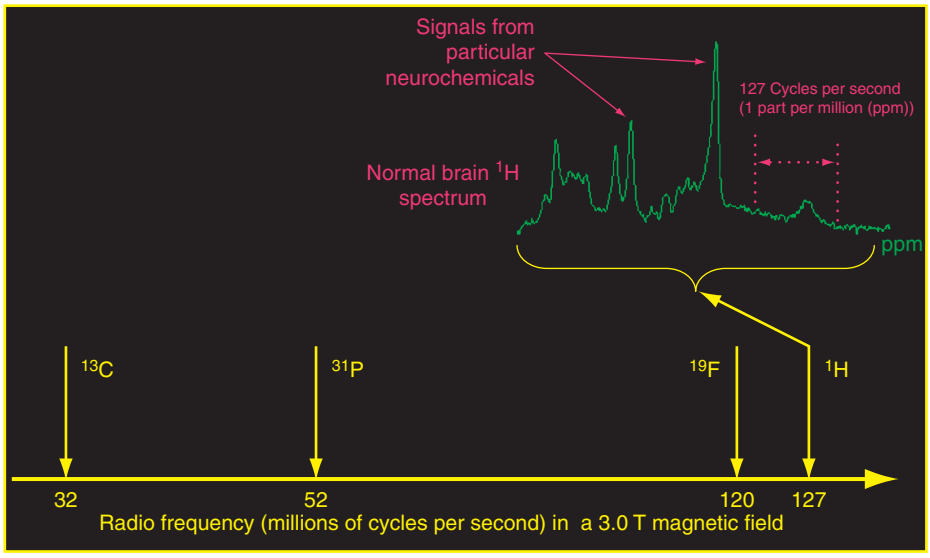

Figure 1 Signal frequency measurements in magnetic resonance spectroscopy (MRST). MRS signals from different atomic nuclei (e.g., ^{13}C, ^{31}P, ^{19}F, and ^{1}H) are distinguished by their unique frequencies at a particular field strength. The characteristic frequencies for some isotopes commonly used in brain MRS are shown. The spectrum in the upper right-hand corner shows localized ^{1}H-MRS data from a normal human brain. It illustrates that particular ^{1}H nuclei located on different molecules, or in different parts of the same molecule, are shifted in frequency by a small amount relative to the characteristic frequency for ^{1}H. The shifts are conveniently measured in parts per million (ppm) relative to the characteristic frequency for the nucleus.

communication. Accordingly, MRS is sometimes referred to as a type of radio frequency spectroscopy.

The typical MRS hardware designed for the study of humans uses superconducting magnet technology to immerse the sample, organism, or organ being studied in a static magnetic field of 1.5 T (about 30 000 times stronger than the Earth's magnetic field). MRS examinations of animals and *in vitro* MRS studies are usually done using magnetic fields that are two to ten times stronger. MRS signal strength generally improves as the static magnetic field strength is increased, and it is desirable to use the strongest magnetic field available. However, there are practical limitations. Magnets having a field strength greater than 3.0 T that are sufficiently large to accommodate the human body are very expensive and are presently used only in research studies.

A second externally applied magnetic field, the strength of which has a temporal periodicity, is also necessary for MRS. It is through this time-varying magnetic field that the specific resonance interaction can be detected (i.e., the spectroscopy). Because the applied time-varying magnetic field must have temporal characteristics similar to those found in the radio waves used in radio and television, the term 'radio frequency (rf) field' is often used to denote this second obligatory magnetic field. The device that produces the rf field is the rf coil, which must be placed around or very near to the sample or body part being examined.

The basic procedure for acquiring MRS data is to place the item of interest in the strong static magnetic field and also within an rf coil. The rf coil is then turned on (i.e., pulsed) very briefly under computer control to create an rf field that interacts with a broad range of different types of atomic nuclei within the sample. Once the rf pulse is turned off, the magnetic interaction between the nuclear magnetism and the static magnetic field produces distinct electromagnetic signals in the rf coil at distinct frequencies. These signals are subjected to analog and digital processing to result in a magnetic resonance spectrum. A spectrum is a two-dimensional plot of frequency on the horizontal axis and intensity of resonance interaction on the vertical axis (**Figures 1** and **2**). Many variants of this basic procedure are available to meet specific goals. The rf pulse is usually turned on more than once during the procedure, and the term 'pulse sequence' is used to convey the idea that the data are generated by a complicated series of pulses designed to meet specific experimental needs.

MRS signals are generally quite weak compared with other types of spectroscopy used in biology and medicine. It usually is necessary to repeat the signal detection multiple times and to average the results to acquire a sufficiently strong MRS signal to identify the MRS signals above the noise that is inherently detected as part of the signal detection procedure. Signal averaging is productive because the desired

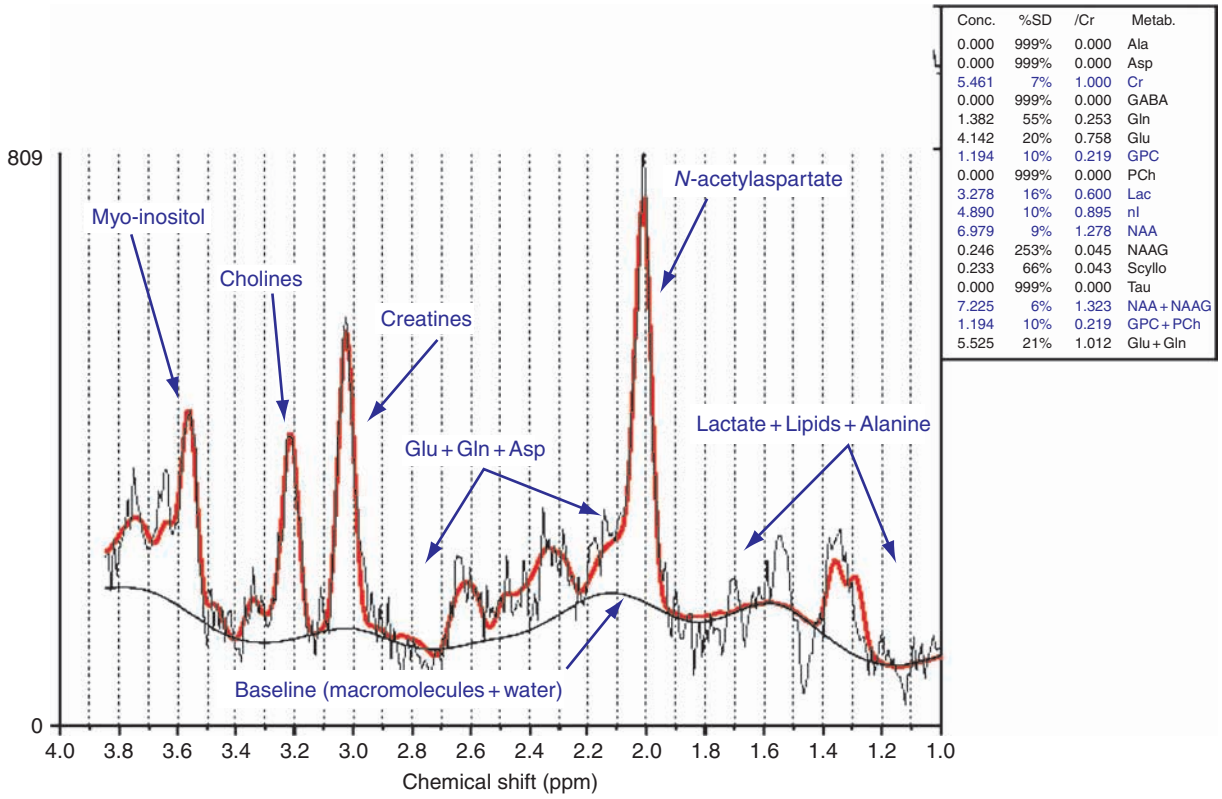

Figure 2 Normal brain magnetic resonance spectroscopy (MRS), showing ^{1}H-MRS with fitting and quantitation. Localized single-volume ^{1}H-MRS of normal human brain is depicted by the black tracing. Note the presence of noise in the black tracing, illustrating an inherent limitation in the detection of certain weaker signals. Some of the key signals are labeled in blue. The baseline arises primarily from low molecular weight macromolecules present in the tissue. The red line shows the best fit to a spectrum model containing contributions from the molecules that are likely to be present in brain tissue. From the fitting process the relative amplitudes (signal area) of each of the contributors can be determined. When these areas are related to the water signal area (not shown), the tissue concentration can be determined because the water concentration is known. The resultant concentration determinations for different neurochemicals are provided in tabular format in the upper right. Fitting and quantitation were performed using a commercially available software package (LCModel, available online at http://S-provencher.com/pages/lcmodel.shtml).

signal is identical on each repeat, whereas the noise is random and differs between repeats.

In Vitro and *In Vivo* MRS

In vitro MRS studies frequently use some form of sample preparation. Typically an acid extraction procedure is used to separate the water-soluble low molecular weight molecules from a tissue sample. The tissue extract is then subjected to the same NMR spectroscopy procedures that would be used in an analytical chemistry NMR laboratory. Extraction has the advantage of providing a magnetically homogeneous and dynamically stable sample that can be examined using a very high magnetic field strength, leading to very high-quality spectra. Extract spectra of the type shown in Figure 3 provide an MRS fingerprint of all chemicals that are present with a far greater sensitivity and specificity than can be obtained with *in vivo* MRS observations.

For *in vivo* MRS there is the need to examine only a small well-defined portion of the brain. Unless special localization measures are taken, one tends to obtain signals from the entire brain and overlying tissue, rather than just part of it. The inherent sensitivity of the MRS technique places a lower theoretical limit on the volume of tissue that can provide adequate MRS signal strength within a limited time. In human brain studies, this practical lower limit is approximately 1 cm^3. A typical human brain MRS region of interest is shown in Figure 4. For animal studies, volumes in the range of 10–50 mm^3 can be examined with the appropriate equipment. Two general MRS localization procedures are available. These are known as localized single-volume MRS (SV-MRS) and magnetic resonance spectroscopic imaging (MRSI). In SV-MRS, one uses a set of magnetic resonance images to define the location of tissue to be examined, and only the MRS signals from this location are acquired (Figure 4). In MRSI, one uses MRI to define a larger

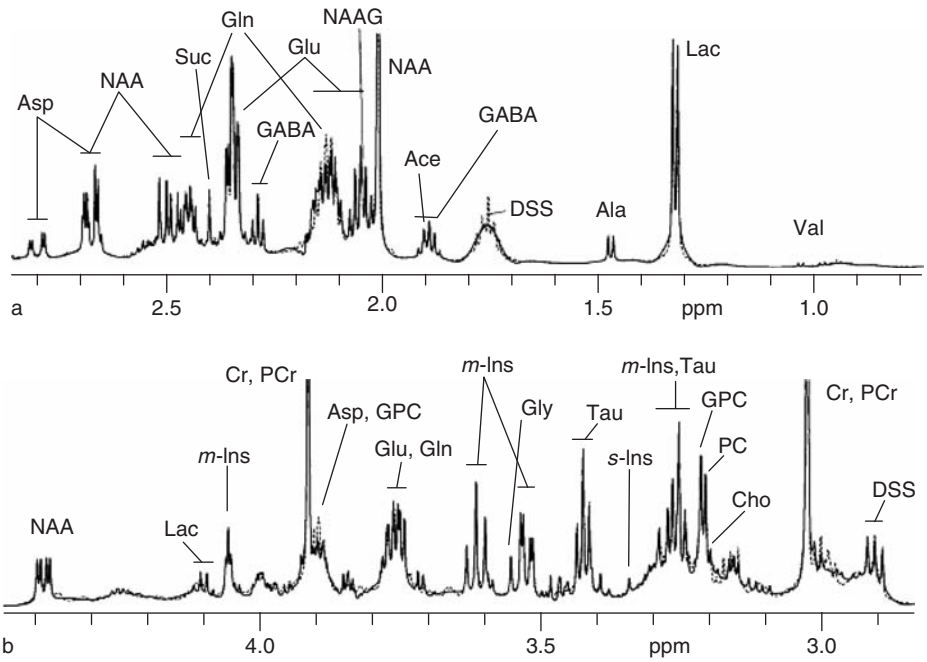

Figure 3 Brain magnetic resonance spectroscopy (MRS), showing brain extract ^{1}H-MRS. ^{1}H-MRS of an acid extract of rat brain tissue is provided in two frames, showing the spectrum's 0.5–3.0 ppm portion (a) and the 2.8–4.5 ppm portion (b). As can be appreciated by comparing this spectrum with those shown in **Figures 1** and **2**, extract spectra provide greater detail than do *in vivo* spectra. Many more signals can be uniquely identified with better sensitivity. The tissue molecules that generated many of the signals are identified by abbreviations in the figure. A discussion of the individual signals and the molecules responsible for them is beyond the scope of this article. For these details the reader should consult the original source. Reproduced from Govindaraju V, Young K, and Maudsley AA (2000) Proton NMR chemical shifts and coupling constants for brain metabolites. *NMR in Biomedicine* 13: 129–153.

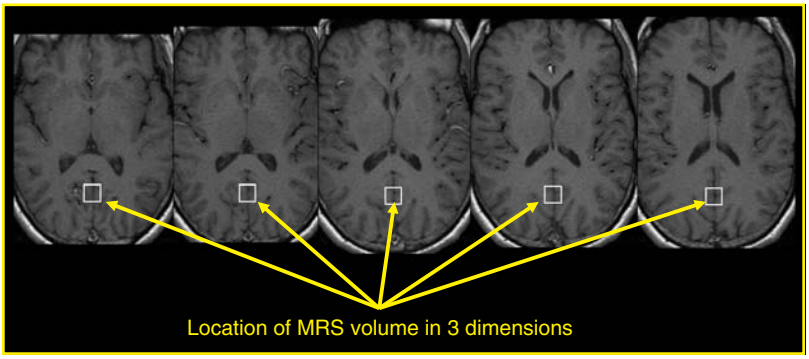

Figure 4 Localized magnetic resonance spectroscopy (MRS) volume selection. In localized MRS studies of brain, a magnetic resonance image (MRI) study is first performed. One then uses the anatomy visualization provided by MRI to define which single volume of brain tissue to examine with MRS. The white squares in the figure illustrate the size of the typical volume used in single-volume MRS. The figure also illustrates that the MRS volume has thickness. It can intersect with more than one of the MRI slices. The figure shows all of the MRI slices that overlap with the chosen MRS volume. Examples of spectra that result from such localization procedures appear in **Figures 1** and **2.**

region of interest (Figure 5). MRSI produces an array of spectra from all the volume elements within the region of interest. Using these data, it is possible to create images that depict how individual signals vary through the array or to evaluate the spectra produced by specific locations.

The time required for data acquisition is an important factor for *in vivo* MRS studies. The individual must remain essentially motionless over the 10 or more minutes required to perform the MRS data acquisition. Although this is tolerable for many humans, some ill patients and younger children find this difficult. For these individuals, sedation or anesthesia is required. For animal studies, sedation, anesthesia, or paralysis is almost always required to prevent movement during the MRS data acquisition.

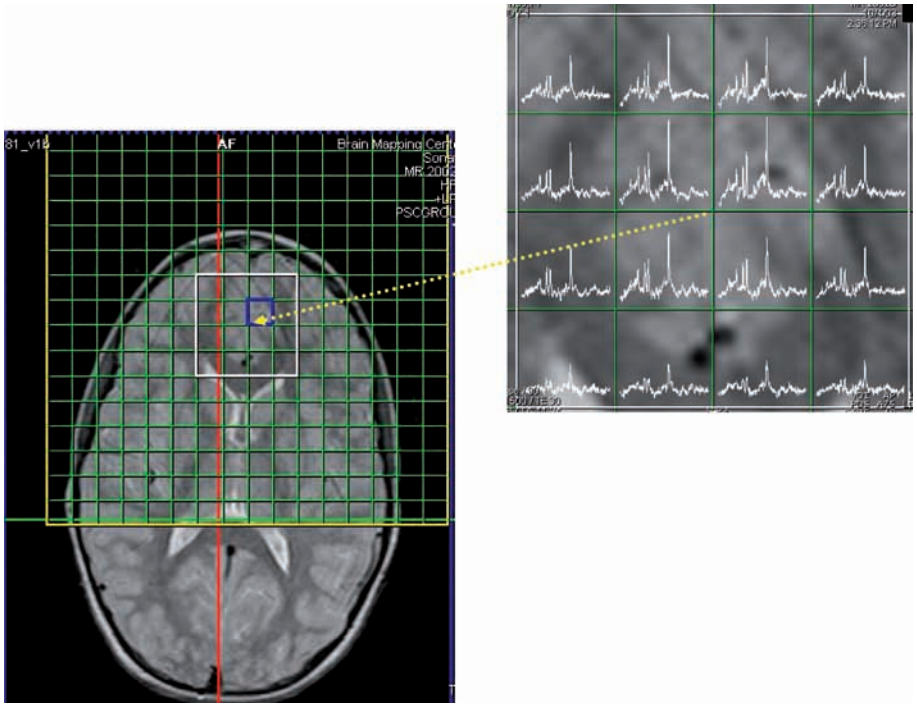

Figure 5 Magnetic resonance spectroscopic imaging (MRSI) procedure. An MRI study is performed to determine the slice of interest (shown as a white square overlaid onto the grayscale MRI), after which MRS procedures are used to obtain a spectrum from each of the volume elements shown in the green grid (array of spectra shown to the right).

Detectable Neurochemicals

Only certain atomic nuclei (isotopes) behave as spinning bar magnets and are capable of producing MRS signals. The nuclei relevant to medical or biological MRS studies are ^{1}H, ^{31}P, ^{13}C, ^{7}Li, and ^{19}F. MRS signals produced by these unique isotopes are easily distinguished by their very different characteristic frequencies (Figure 1). For instance, at a static magnetic field strength of 3.0 T, all ^{1}H nuclei produce signals in the immediate vicinity of 127 million cycles per second, whereas all ^{31}P nuclei produce signals in the immediate vicinity of 52 million cycles per second. These between-nuclei differences are so great, that different rf coils are usually needed for different nuclei. The relevant nuclei have different attributes and practical limitations with respect to MRS signal production. An equal number of each type of nucleus produces differing inherent signal strengths. Furthermore, all MRS nuclei do not have the same abundance in nature. ^{13}C, for instance has a natural abundance of only 1%. Of the aforementioned nuclei, ^{1}H produces the strongest MRS signal because it has a natural abundance of almost 100% and it has the highest inherent sensitivity of all the magnetic nuclei. For this reason it is the most frequently exploited nucleus in routine clinical MRS implementations. Moreover, ^{1}H-MRS is somewhat easier to perform than are other types of MRS, because it uses the same equipment as used in conventional MRI.

The ability to detect unique signals from different chemicals with MRS results from the fact that the frequency of the magnetic interaction is directly proportional to the static magnetic field strength at the nucleus. The electron clouds which surround the atomic nuclei within molecules circulate in ways that tends to alter the actual magnetic field at the atomic nucleus, and this causes small alterations in the MRS signal frequency. In this manner, the electrons that make up the unique chemical structure surrounding each atomic nucleus cause different nuclei within the same molecule, and the nuclei within different molecules, to produce unique and often distinguishable MRS frequencies. The difference in frequency resulting from chemical structure is rather small (Figure 1). Important MRS signals may be separated from each other in frequency by only a few cycles per second, with each of their signals having a frequency of millions of cycles per second. For this reason it is common practice to specify the frequency of a particular signal by stating how much its frequency differs from some that are produced by an accepted standard nucleus. This practice leads to the specification of signal frequencies in parts per million (ppm).

There are two major factors that practically limit the number of neurochemicals that can be detected with MRS. First, the signal-producing molecules must be relatively free to move from place to place and tumble around within the various aqueous solutions that make up biological tissue. Macromolecules having molecular masses greater than about 20 000 daltons fail to move with sufficient freedom, and signals arising from the atomic nuclei within such structures cannot be practically detected. Second, overcoming the inherent noise in the signal-detection process is challenging. There must be sufficient atomic nuclei of a specific type to generate a recognizable signal above the noise level. The necessary amount of material varies somewhat depending on the type of MRS study and the nucleus. A good rule of thumb for the ^{1}H-MRS is that there must be at least about 10^{18} molecules of interest (1 μmol) in the volume of interest for signal recognition. Frequently the sample or tissue volume that is examined is adjusted to achieve this requirement. This leads to volume elements of the sizes illustrated in Figures 4 and 5.

Only the most heavily concentrated neurochemicals can be detected *in situ* with MRS because of its sensitivity requirements. In ^{1}H-MRS and ^{13}C-MRS, signals are readily detected from a variety of intermediary metabolites, amino acids, and some neurotransmitters (Figures 2 and 3). Certain interesting neurotransmitters (e.g., dopamine, serotonin) cannot be detected because their levels are too low. Hence ^{1}H-MRS and ^{13}C-MRS can be used to evaluate metabolism and its derangements in brain tissue. Because neurons and glia have differing metabolite profiles, alterations in the signal patterns detected with MRS in disease conditions are frequently interpreted in terms of altered densities of neurons and glia in the tissue. ^{31}P-MRS (Figure 6) is used to detect signals from metabolites involved in energy metabolism, including adenosine triphosphate, phosphocreatine, and inorganic phosphate. Magnesium ion concentration and pH can be determined from measures of frequency differences between certain metabolite signals. Also detectible with ^{31}P-MRS are certain metabolites associated with phospholipid and membrane metabolism. Studies with ^{19}F-MRS have demonstrated that the brain levels of certain fluorinated drugs can be measured. Similarly ^{7}Li-MRS can be used to detect and measure the lithium levels in the brains of patients being treated for depressive conditions.

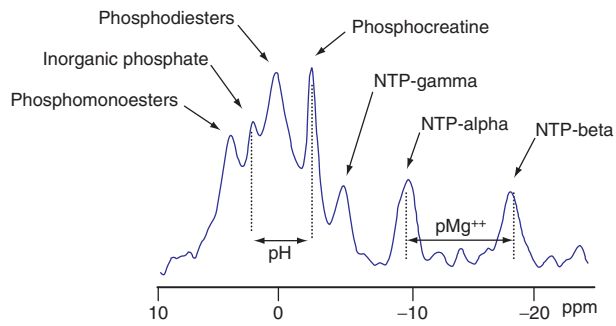

Figure 6 Brain magnetic resonance spectroscopy (MRS), showing ^{31}P-MRS of human brain. This spectrum was obtained from a 20 mm-thick axial brain volume that mostly encompassed the frontal lobes. The key signals are shown. Adenosine triphosphate (ATP) is the predominant contributor to the nucleotide triphosphate (NTP) signals. The phosphodiesters and phosphomonoesters are produced by metabolites involved in membrane metabolism. These signals are frequently of interest in conditions in which there is an acceleration of membrane synthesis (i.e., cancer) or a degeneration of membranes (i.e., neurodegenerative conditions). The intracellular concentrations of hydrogen ion (pH) and magnesium ion (pMg^{2+}) can be inferred from measurements of certain signal frequency differences.

Quantification of Neurochemical Levels

Magnetic resonance instruments can provide a highly accurate measure of signal frequency but are not designed to produce such exact measures of signal amplitude (the area under the signal) that are calibrated against a defined universal standard. Hence MRS permits ready conclusions about whether a particular neurochemical signal is present above the noise level at a specific frequency, but it is somewhat more difficult to determine exactly what is the tissue concentration of the signal-producing neurochemical. Measures of signal intensity must be quoted with reference to some calibration signal that is also present in the spectrum being analyzed or can be acquired from some other MRS or MRI measurement of the same tissue region in the same individual. For this reason, MRS results are often reported as the ratio of two signals (e.g., the ratio of the amplitudes of the *N*-acetylaspartate and creatine signals). Studies performed during the past few years have demonstrated that the tissue water can provide a stable reproducible calibration signal. Because the concentration of water is known and reasonably uniform within soft tissue, it is possible to estimate the absolute concentrations of other signal-producing neurochemicals through measurement of their signal amplitudes relative to that of water. The absolute concentrations derived in this way have varying precision (Figure 2). Amplitude measures of the smaller signals are influenced to a greater proportional extent by noise compared with measures of the larger signals. Therefore, the concentration estimates for the weaker signals are correspondingly less precise. Furthermore, it is frequently the case that one MRS signal overlaps with another

MRS signal, and errors in concentration estimation can arise from the inability to correctly attribute signal amplitude to different signal-generating molecules (Figure 2).

Metabolic Flux Measurements

MRS is primarily a method for detecting and measuring the concentrations of particular signal-generating neurochemicals. It does not inherently measure metabolic flux except in special study designs. In this regard, MRS is quite distinct from other metabolic imaging techniques, such as positron emission tomography, that use radiotracers to measure metabolic flux.

There are several MRS study designs that can provide metabolic flux measures. One may introduce a physiological perturbation and then repetitively collect MRS data over time. Such data provide measures of neurochemical concentration versus time, from which metabolic flux measures can be derived. ^{31}P-MRS studies of the influence of ischemia on brain energy metabolites have been performed in this fashion. Another method involves labeling certain metabolic pools with stable isotopes that can be detected with MRS. ^{13}C is an example of one such isotope. It has a low natural abundance and produces almost no background MRS signal. It is possible to purchase certain precursor metabolites (e.g., glucose) enriched with ^{13}C at specific carbon positions. Such labeled molecules can be introduced into the bloodstream and their metabolism followed through time course measures of MRS signals produced by the labeled precursor and its metabolites. In the case of some relatively fast metabolic reactions, it is also possible to use MRS to follow the flux of magnetic labels (which are noninvasively introduced by the scanner hardware) between products and reactants. Such methods have been most frequently used to measure creatine kinase flux with ^{31}P-MRS.

Summary

MRS provides a means of measuring the concentrations of the more concentrated neurochemicals in the brain or in tissue extracts using magnetic resonance signals produced by atomic nuclei. Brain MRS measurements can be made with regional specificity and without exposure to ionizing radiation. The technique requires only exposure to magnetic fields. MRS is now becoming routinely used in clinical settings to evaluate a number of brain diseases. It can also be useful in basic neuroscientific investigations.

See also: Brain Injury: Magnetic Resonance Studies of Metabolic Aspects; Cognitive Neuroscience: An Overview; Event-Related Potentials (ERPs) and Cognitive Processing; fMRI: BOLD Contrast; Neuroimaging; Perfusion MRI; Positron Emission Tomography (PET).

Further Reading

Burlina AP, Aureli T, Bracco F, et al. (2000) MR spectroscopy: A powerful tool for investigating brain function and neurological diseases. *Neurochemical Research* 25: 1365–1372.

Gadian DG (1995) *NMR and Its Applications to Living Systems*, 2nd edn. New York: Oxford University Press.

Govindaraju V, Young K, and Maudsley AA (2000) Proton NMR chemical shifts and coupling constants for brain metabolites. *NMR in Biomedicine* 13: 129–153.

Hsu YY, Du AT, Schuff N, et al. (2001) Magnetic resonance imaging and magnetic resonance spectroscopy in dementias. *Journal of Geriatric Psychiatry and Neurology* 14: 145–166.

Lee PL and Gonzalez RG (2000) Magnetic resonance spectroscopy of brain tumors. *Current Opinion in Oncology* 12: 199–204.

Lyoo IK and Renshaw PF (2002) Magnetic resonance spectroscopy: Current and future applications in psychiatric research. *Biological Psychiatry* 51: 195–207.

Kwok L, Smith JK, Castillo M, et al. (2006) Clinical role of proton magnetic resonance spectroscopy in oncology: Brain, breast and prostate cancer. *Lancet Oncology* 7: 859–868.

Morris PG (1999) Magnetic resonance imaging and magnetic resonance spectroscopy assessment of brain function in experimental animals and man. *Journal of Psychopharmacology* 13: 330–336.

Novotny EJ, Fulbright RK, Perl PL, et al. (2003) Magnetic resonance spectroscopy of neurotransmitters in human brain. *Annals of Neurology* Supplement 6: S25–S31.

Ross B and Bluml S (2001) Magnetic resonance spectroscopy of the human brain. *Anatomical Record* 265: 54–84.

Pirko I, Fricke T, Johnson AJ, et al. (2005) Magnetic resonance imaging, microscopy and spectroscopy of central nervous system in experimental animals. *NeuroRx* 2: 250–264.

Ruggieri PM and Najm IM (2001) MR imaging in epilepsy. *Neurologic Clinics* 19: 477–489.

Saunders DE (2000) MR spectroscopy in stroke. *British Medical Bulletin* 56: 334–345.

Ugurbil K, Adriany G, Andersen P, et al. (2000) Magnetic resonance studies of brain function and neurochemistry. *Annual Review of Biomedical Engineering* 2: 633–660.

Relevant Website

http://S-provencher.com – Software resource for fitting and quantitation of magnetic resonance spectroscopy see LC Model.

Magnetic Sense in Animal Navigation

T Ritz, University of California at Irvine, Irvine, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

The geomagnetic field provides an omnipresent clue that can be used, in principle, to obtain information about direction and position, in particular geographic latitudes. Although migratory birds present perhaps a particularly spectacular example, magnetic sensory capabilities are widespread and have been demonstrated in animals as different as bacteria, honey bees, newts, mollusks, salmon, sea turtles, migratory and nonmigratory birds (pigeons and chicks), and mole rats, to name a few. It is likely that magnetic sensing presents an evolutionary advantage to many territorial animals, for example, for finding the closest hiding place or food storage location from the current position of an animal. Migratory animals may have developed this useful ability further to aid them in their long-distance orientation tasks.

Unlike other sensory modalities, such as vision, hearing, touch, and olfaction, very little is known about the receptors, signal transduction, and neural processes underlying magnetic sensing. Most of the available information about magnetic sensing today is based on sometimes variable and contradictory behavioral observations and only very few electrophysiological and anatomical studies. The reasons for the comparative lack of knowledge about magnetoreceptors may be rooted in the unusual features of magnetic fields as stimuli. Magnetic fields pass relatively freely through tissue, making it less obvious where to pinpoint loci for potential magnetosensory structures. Unlike other stimuli, the geomagnetic field is not an on/off stimulus but is always present in similar intensity, suggesting that standard test approaches such as measuring dose–response curves or action spectra in behavioral tests may be less meaningful than for other sensory modalities. On the other hand, the geomagnetic field is sufficiently weak, which limits the possible physical mechanisms of magnetic field detection to less than a handful of plausible hypotheses. Theoretical considerations thus allow the design of diagnostic experiments to identify the underlying mechanism for a particular response. Two hypotheses in particular – magnetoreception based on magnetite (a biogenic magnetic material) and magnetoreception based on magnetically sensitive photoreceptor-based reactions – have received much attention.

In this article, we first briefly review the various types of well-documented magnetic responses and then present both the magnetite and photoreceptor hypotheses of magnetic sensing. In recent years, much progress has been made toward accumulating evidence for both hypotheses, raising the prospect of solving the long-standing mystery of magnetic sensing in the not-too-distant future.

Examples of Magnetic Sensing in Animal Navigation

Magnetic Compasses

The most convincing type of experiment that has been used to study magnetic sensing measures magnetic compass responses. In magnetic compass experiments, an animal is placed in an arena where it is motivated to move in a particular direction. For example, a migratory bird in the Northern Hemisphere is motivated to move north in the spring and will do so even in a small funnel cage. As the direction of the magnetic field in the arena is changed in repeated tests, we find that the headings of the animals change accordingly. Such magnetic compass abilities have been detected and reproduced for many species of migratory birds as well as for many other animals, including chicks, sea turtles, honey bees, mole rats, and newts. In migratory birds, the magnetic compass responses have been characterized in the most detail. Migratory birds are sensitive to the dip or inclination angle; reversing the vertical component of the magnetic field results in a reversal of directions, whereas a change in the polarity of the magnetic field has no effect on their orientation. Ambient light conditions affect the magnetic responses significantly. Under full-spectrum (white) light, birds respond by moving in the appropriate migratory direction. However, under monochromatic light, birds may either orient in the appropriate migratory direction, be disoriented, or move along a direction or axis that is unrelated to the migratory direction (called fixed-direction or fixed-axis responses). It appears that, under monochromatic light conditions, normal migratory behavior is observed mostly under short-wavelength blue and green light of low intensity, corresponding to light approximately 45 min after sunset. Disoriented responses are more frequent under long-wavelength yellow and red light of any intensity, whereas fixed-direction or fixed-axis responses appear mostly for higher-intensity short-wavelength light. No clear picture has emerged yet as to which photoreceptors might be involved in these responses.

Signposts or Markers

Birds as well as sea turtles follow often complex migratory routes; for example, sea turtles follow large currents such as the North Atlantic gyre. At certain points along those routes, crucial orientation decisions need to be made. For example, sea turtles need to decide to stay in the Atlantic gyre. When exposed in a lab setting to magnetic field conditions that they encountered at these crucial points, sea turtles responded by swimming in a direction that would keep them within the gyre. Likewise, some birds in Europe plot a migratory route that leads them through Spain, thereby crossing the Mediterranean Sea at its narrowest point on their way to Central Africa. Exposure of these birds to the magnetic field conditions in Spain plays a role in triggering the necessary route change from southwest to southeast. Specific magnetic field conditions also trigger the fattening of birds prior to crossing large deserts. In this way, magnetic fields can act as markers or signposts. A particularly intriguing aspect of these signpost responses is that they are observed even in freshly hatched sea turtles, suggesting that they are either genetically coded or determined in the embryonic stages.

Conditioning

For a long time, conditioning animals to magnetic stimuli was unsuccessful, but in recent years the number of studies reporting successful conditioning has grown significantly. Operant conditioning to magnetic stimuli has been reported for honey bees, different species of fish including the genetic model zebra fish, and, very recently, pigeons. The last study is particularly noteworthy because more than a dozen previous attempts failed to achieve magnetic conditioning in pigeons. In the study, pigeons learned to discriminate between the presence and absence of a magnetic anomaly created in the center of a tunnel-shaped arena. The discrimination performance is relatively low, suggesting that discrimination of the magnetic stimulus is a difficult task. Conditioning experiments may open the way to better identifying what information animals derive from the geomagnetic field by identifying the thresholds of intensity and/or spatial inhomogeneities required for successful conditioning. Conditioning to magnetic directions has very recently been reported for chicks.

Magnetite-Based Magnetoreception

Magnetite (γ-Fe_3O_4) is a widespread iron oxide that can be produced in biological cells. Its magnetic properties depend on the size and geometry of the magnetite crystals. In very small crystals, thermal fluctuations prevent them from carrying a stable magnetization. However, an external magnetic field induces magnetization in a collection of such particles, which is lost once the field is switched off. Such a collective behavior is called superparamagnetism (with the prefix super referring to the fact that such a collective has a surprisingly high susceptibility, that is, even a field of low intensity can induce a large magnetization). Larger crystals become ferromagnetic, that is, they possess a permanent magnetic dipole moment at room temperature. Crystals of this size are often labeled single-domain crystals to distinguish them from even larger crystals with multiple magnetic domains, each with a permanent magnetic dipole.

For ferromagnetic materials and certain geometries of superparamagnetic material, the geomagnetic field exerts a torque on the magnetite, thereby providing a force that rotates it into the direction of the magnetic field. Such a torque mechanism is, of course, the basis of the mariner's compass needle and is theoretically well understood. A torque produces mechanical deformation, which in combination with a mechanoreceptor could create a receptor potential and, hence, trigger a nerve signal. Moreover, there is biological precedence – a torque mechanism is realized in magnetic bacteria that contain chains of magnetite and can be considered microscopic compass needles. Due to the torque of the magnetic field, the bacteria become aligned with the geomagnetic field lines and use this to swim in upward or downward directions. It is less clear whether the detection of the magnetic field by bacteria amounts to true sensing or whether it is merely a passive effect because the torque of the magnetic field will rotate living and dead bacteria alike.

Evidence for Magnetite-Based Magnetoreception

Simple detection of magnetite in organisms can be taken only as a first step toward identifying possible magnetoreceptive structures because magnetite can also be produced as a by-product of iron metabolism and it is used for physiological functions independent of magnetic field detection (e.g., for the stabilization of bone tissue because it is hard and within vestibular systems because of its high density). Characterization of the structure and physical properties of the detected magnetic material is needed as a further step to determining that this material is indeed involved in magnetic field detection.

Ferromagnetic crystals of magnetite have been found in animal tissue and closely resemble bacterial magnetite crystals. For example, magnetite crystals have been extracted from ethmoid tissue of the sockeye salmon,

with a small range of grain sizes, as is typical of biologically controlled mineralization processes. Due to the extraction procedure, however, there was no information left about the *in situ* disposition of crystals and their histological context. Strong evidence for a magnetite-based magnetoreceptor cell has been found in another fish, the trout. Through tracing from magnetically sensitive neurons, candidate magnetoreceptor cells were identified in the olfactory lamellae. Within these cells, iron-rich crystals were identified, and it was later shown that these crystals have a permanent magnetic moment. The exact physical arrangements of the crystals and their connections to the nerve cells still remain to be determined.

In birds, ultrastructural analysis reveals an intriguing magnetite-containing structure in the beak with rather different properties from the structures observed in fish. The bird structure contains nanometer-size magnetic particles that fill corpuscles of approximately 2 μm in diameter, as well as bands of approximately 15–25 magnetic platelets, each approximately $1 \times 1 \times 0.1$ μm in size. Furthermore, vesicles of unknown material of approximately 5 μm in diameter can be seen that are encased by magnetite platelets. From the size of the platelets and particles, we expect that all magnetic material is superparamagnetic. Overall, six such structures are located in the upper beak of birds, three on each side of the beak. The magnetite-containing structures are located in the vicinity of nerve endings that are likely to be sensitive to mechanical, but not surface, stimulation and could be well suited for the detection of mechanical strains. It remains to be seen how strong a magnetic field needs to be to produce mechanical deformations that trigger nerve signals. One interesting aspect of the arrangement of the bands of platelets is that their axes appear to be oriented in different, almost perpendicular directions in the three structures in each side of the beak. Thus, each of the three structures could, in principle, measure a different component of the geomagnetic field vector, allowing a one-shot measurement of the magnetic field intensity comparable to a three-axis magnetometer.

Beyond the suggestive arrangement of magnetite structures, evidence for magnetite-based magnetoreception derives from behavioral experiments in which animals were exposed to strong (>0.5 T) but short (several milliseconds) magnetic pulses prior to magnetic orientation testing. Such strong magnetic pulses should remagnetize ferromagnetic magnetic material and we therefore expect disoriented responses if magnetoreceptors are based on such material. In adult migratory birds and homing pigeons, the application of pulse magnetization resulted in a significant deflection of oriented responses and sometimes in bimodal (as opposed to unimodal) responses. The effects of pulse magnetization usually vanished after several days, and all birds returned to normal migratory orientation after 1 week. The disruptive effect of pulse magnetization in adult birds suggests that magnetite is involved in magnetic field detection, although we cannot conclusively exclude nonspecific effects on other substrates because of the strength of the pulse. On the other hand, the lack of effects on magnetic orientation in juvenile birds suggests that all magnetoreception in birds is not based on magnetite. The difference between adult and juvenile birds is commonly rationalized by assuming that adult or experienced birds (i.e., birds that have completed migratory or homing routes) have developed a magnetic map with the magnetite-based sensor providing map information through intensity measurements. On the other hand, juvenile birds are assumed to only use the magnetic compass, presumably not based on magnetite. Pulse magnetization would then disrupt only the map sensor but not the compass sensor. For adult birds, the effects of pulse magnetization vanished when the ophthalmic nerve was anesthetized, suggesting that the structure affected by the pulse is located in the beak or nose, which is consistent with the structural observations. Juvenile birds continued to show normal magnetic orientation even when the ophthalmic nerve was anesthetized. Pigeons lost the ability to discriminate a strong magnetic anomaly in conditioning experiments when the ophthalmic nerve was cut, suggesting that the magnetite-based beak structure is also involved in detecting such anomalies.

Chemical- (or Photoreceptor-) Based Sensing

Magnetic fields can affect the electronic states of atoms and molecules, in particular their spin states. The magnetic fields required to overcome thermal fluctuations are usually much stronger than the geomagnetic field. However, in chemical reactions that involve an intermediate step with unpaired electron spins (called radical-pair intermediates) much weaker magnetic fields can change spin states and thus change the yields and rates of spin-dependent products. The size of magnetic field effects depends on the internal magnetic fields within the molecules (called hyperfine interactions) and on the lifetime of the radical-pair intermediates with lifetimes of at least several hundred nanoseconds required for the effects of Earth-strength magnetic fields to develop. The size of magnetic field effects can be calculated if the molecular determinants are known and the effects of Earth-strength magnetic fields on radical-pair reactions have indeed been measured *in vitro*.

Because hyperfine interactions of many nuclei are anisotropic, the effects of an external magnetic field on a single radical-pair reaction depend on the strength as well as on the direction of the external magnetic field with respect to the molecular frame of reference.

The formation of radical pairs is quite common in biological systems, in particular for reactions involving photopigments. Photopigments can respond to light in fundamentally different ways. Photopigments such as retinal in the opsin family of photoreceptors undergo a conformational change upon light absorption, triggering the reaction cascade underlying normal visual photoreception. Other photopigments, such as chlorophylls and flavins, use light energy to transfer electrons to nearby molecules, thereby generating radical pairs. The angular dependence of magnetic field effects on radical-pair reaction yields can, in principle, form the basis for a physiological magnetic compass, provided that:

1. Photopigments that can form long-lived radical pairs exist in birds.
2. A sufficient number of photopigments are fixed in the same orientation to provide directional information.
3. A link of photopigments to a signal transduction system (e.g., the visual system) exists.

Unlike other sensory receptor mechanisms, in which a stimulus creates a signal, in the radical-pair mechanism, a magnetic stimulus modulates an already existing signal by changing the balance between two intermediate reaction products. Magnetic fields can, in this way, be detected by their indirect influence on another signal transduction pathway. Considering photopigments as magnetic field effect mediators, it appears most plausible that magnetic fields affect photoreception. The absorption of light by a photopigment with appropriate photochemistry triggers a radical-pair reaction, thereby modulating the spin states of the photopigment. If, for example, only the singlet state (but not the triplet state) of the photopigment can initiate subsequent signal transduction steps, the change of singlet yields by the magnetic field results in proportional changes in the photoreceptor signal intensity. Assuming a fixed orientation of the radical-forming photopigments inside the retinal cells of the eye, the magnetic field modulates radical-pair reactions differently in different parts of the retina, leading to the perception of the magnetic field as visual modulation patterns.

Evidence for Chemical- (or Photoreceptor-) Based Sensing

Initial support for the radical-pair hypothesis was derived from studies demonstrating that the compass of birds is affected by the availability of light with specific wavelengths. Furthermore, birds with their pineal gland removed can still orient magnetically, suggesting that the eyes are the most likely site of processing light information in relation to the magnetic compass. Recently, it was shown that magnetic compass responses of European robins appear to be strongly lateralized and possibly located in the right eye because covering up the right eye of migratory birds seemed to destroy their magnetic compass-orientation capabilities. However, none of these studies rules out the possibility that a change in light conditions affects behavioral responses due to a change in motivation or other reasons unrelated to the magnetic-sensing mechanism. Thus, the effects of ambient light alone cannot be taken as evidence for the existence of a radical-pair mechanism.

In principle, spectroscopy with oscillating magnetic fields can reveal the chemical nature of the molecules involved in a radical-pair reaction. Such electron spin resonance spectroscopy is only possible *in vitro* because fluorescence of the product yields is too small to be detectable *in vivo*. However, a form of *in vivo* magnetic resonance spectroscopy can be attempted by detecting the effect of high-frequency oscillating magnetic fields on the behavior of the animal. Oscillating fields in the 1–100 MHz range are expected to interfere with detection of the geomagnetic field in a radical-pair mechanism, and sufficiently strong oscillating fields should therefore result in disorientation in a behavioral test. The application of oscillating magnetic fields, indeed, resulted in bird responses consistent with expectations from the radical-pair mechanism. Fields of 1.3 and 7 MHz at a strength of 470 nT resulted in disorientation in magnetic compass responses unless they were presented parallel to the static (i.e., geomagnetic field), in which case no effect of the oscillating fields was discernible, as expected for a resonance effect. Other orientation responses of birds and responses of mole rats, with a compass presumably based on magnetite, were unaffected by the oscillating fields. These tests together provide strong indirect evidence that the magnetic compass of bird involves a radical-pair mechanism, but provide no information about the chemical nature of the radical pair.

Cryptochromes have been suggested as possible magnetoreceptor candidates because they have the ability to form long-lived radical pairs and because they are found in the eye of birds. In plants, cryptochromes mediate several blue-light-dependent responses, for example, growth inhibition and flowering time. Recently, it was shown that cryptochrome-dependent responses are enhanced through a change in the intensity of the local magnetic field from 33–44 to 500 μT. Control

responses mediated by other photororeceptors or in mutants lacking cryptochromes showed no magnetic field effects.

These findings demonstrate a link between light sensing through cryptochromes and magnetic field effects, and they pave the way toward the identification of the magnetic signal transduction mechanism through molecular biological and genetic studies in the *Arabidopsis* model system.

Other Mechanisms

In principle, magnetic fields could be detected through induction by sufficiently sensitive electric sensors, such as the Lorenzini ampullae in fish. However, because air is a much less conductive medium than water, the consensus view is that the electric sensory structures in land-based animals or birds have to be even larger. That is, they would be of such size that they could hardly be overlooked by standard anatomical characterizations. The lack of a candidate anatomical structure suggests that induction-based magnetoreception may be an issue of interest only for animals in aqueous environments. There exist a plethora of suggested possible alternative mechanisms for detecting the geomagnetic field in biological systems. However, though many of these suggestions are plausible for the detection of magnetic fields ten or hundred times stronger than the geomagnetic field, none of these suggestions has provided a convincing physical mechanism by which magnetic fields of geomagnetic field strength (30–70 μT) can be detected against the background of thermal and molecular shot noise. To date, magnetite-based and radical-pair reaction-based magnetoreception are the only two mechanisms for which *in vitro* experimental evidence demonstrates the effects of Earth-strength magnetic fields. Although some behavioral data suggest that much smaller intensity changes (down to less than 20 nT) may be detected in animals, no plausible physical mechanism exists yet that can explain how biological sensors can detect such tiny changes.

Neurophysiological Approaches

Early electrophysiological recordings suggested several centers in birds where magnetoreception might be processed. The firing rates of neurons in the nucleus of the basal optic root (nBOR), a center that combines visual with vestibular information, changed in response to a magnetic field stimulus with slowly changing direction. A comparable stimulus also resulted in changes of neuron responses in the optic tectum, with either response disappearing when the optic nerves were cut. Increased neural responses were furthermore recorded in trigeminal ganglion cells when the vertical magnetic field intensity was increased by 5 mT or more beyond the local magnetic field intensity, resulting in a change both of total intensity and inclination of the magnetic field. Responses of neurons in the trigeminal nerve were also observed in trout, in which single neurons were identified in the superficial ophthalmic branch (ros V) of the trigeminal nerve that responded to changes in the intensity but not the direction of an imposed magnetic field. The discrimination of a magnetic anomaly by pigeons in conditioning experiments was disrupted when the ophthalmic branch of the trigeminal nerve was cut. The responses in the nBOR and optic tectum are consistent with a magnetic sensor in the eye, whereas the responses in the ophthalmic branch of the trigeminal nerve are consistent with a magnetic sensor in the beak or nose. However, many of the electrophysiological studies date back to more than 20 years ago, and later attempts to reproduce the results of the bird studies have sometimes met with difficulties.

In recent years, several neurophysiological studies have focused on behavioral molecular mapping techniques. Immediate early genes such as ZENK and c-fos are expressed only in cells showing neuronal activity and can therefore be used to identify brain areas active during a particular behavioral or sensory task. In mole rats (which are subterranean mammals), neuronal activity levels in neurons in the suprachiasmatic nucleus (SCN) differed between control animals and those placed in an unfamiliar environment with a changed magnetic field. In birds, a brain area named cluster N was identified in the forebrain; it is active at night during magnetic orientation and loses activity when the eyes are covered. However, no changes of neuronal activity could be observed for different magnetic field conditions. This could mean that, instead of cluster N, other brain areas are involved in magnetic processing that do not express the genetic markers such as is the case for the nBOR and the thalamus of birds. Alternatively, mapping with genetic markers that integrate over neuronal activity may not be sufficiently sensitive to detect changes in activity patterns due to magnetic fields in birds. Imaging techniques with better time resolution, such as Ca imaging may be able to resolve the magnetically sensitive centers in the brain.

Outlook

Much of the debate in the past on the mechanistic aspects of magnetic sensing centered on the question of whether a magnetite or a radical-pair-based mechanism is the exclusive basis of magnetic sensing. This somewhat myopic view is now gradually changing to a recognition that both mechanisms may operate,

possibly even in the same animal. In hindsight, such multiplicity does not appear surprising because sensory systems often show the specialization of receptors for different tasks.

In fish, there is strong evidence for a ferromagnetic magnetite-based structure providing magnetic information and no hint of additional magnetic sensory structures. Neurons in the trigeminal ganglion are prime candidates for electrophysiological studies revealing the sensitivity of the structure to intensity and directional stimuli.

In birds, an interesting magnetite-containing structure has been identified in the beak of birds. Magnetic information is probably transmitted from this structure through the ophthalmic branch of the trigeminal nerve in adult (but not juvenile) birds, but the biological function of this information is not yet entirely clear. Behavioral evidence points to the existence of a second magnetic sensory structure responsible for normal magnetic compass orientation that operates on a radical-pair mechanism; it is probably linked to photoreception, possibly in the retina. The location and structure of this sensor and the pathway of neural information remain unknown. Visual centers and ganglion cells appear the best candidates for electrophysiological studies. The confluence of results of modeling with increased predictive power, novel behavioral studies with more diagnostic conditions, and neurobiological mapping studies suggests that the field is nearing the unequivocal identification of the magnetoreceptors.

See also: Bird Brain: Evolution; Cone Photopigment Evolution; Optic Nerve, Optic Chiasm and Optic Tracts; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vision: Mechanisms of Orientation, Direction and Depth.

Further Reading

Fleissner G, Holtkamp-Rotzler E, Hanzlik M, et al. (2003) Ultrastructural analysis of a putative magnetoreceptor in the beak of homing pigeons. *Journal of Comparative Neurology* 458: 350–360.

Freire R, Munro UH, Rogers LJ, Wiltschko R, and Wiltschko W (2005) Chickens orient using a magnetic compass. *Current Biology* 15: R620–R621.

Johnsen S and Lohmann KJ (2005) The physics and neurobiology of magnetoreception. *Nature Reviews Neuroscience* 6: 703–712.

Kirschvink JL and Gould JL (1981) Biogenic magnetite as a basis for magnetic field detection in animals. *BioSystems* 13: 181–201.

Mora CV, Davison M, Wild JM, and Walker MM (2004) Magnetoreception and its trigeminal mediation in the homing pigeon. *Nature* 432: 508–511.

Mouritsen H and Ritz T (2005) Magnetoreception and its use in bird navigation. *Current Opinion in Neurobiology* 15: 406–414.

Nemec P, Burda H, and Oelschlager HH (2005) Towards the neural basis of magnetoreception: A neuroanatomical approach. *Naturwissenschaften* 92: 151–157.

Ritz T, Adem S, and Schulten K (2000) A model for photoreceptor-based magnetoreception in birds. *Biophysical Journal* 78: 707–718.

Ritz T, Thalau P, Phillips JB, Wiltschko R, and Wiltschko W (2004) Resonance effects indicate a radical-pair mechanism for avian magnetic compass. *Nature* 429: 177–180.

Walker MM, Diebel CE, Haugh CV, Pankhurst PM, Montgomery JC, and Green CR (1997) Structure and function of the vertebrate magnetic sense. *Nature* 390: 371–376.

Wiltschko R and Wiltschko W (1995) *Magnetic Orientation in Animals*. Berlin: Springer.

Wiltschko W and Wiltschko R (2005) Magnetic orientation and magnetoreception in birds and other animals. *Journal of Comparative Physiology A* 191: 675–693.

Magnetoencephalography

D Cohen, Massachusetts General Hospital, Charlestown, MA, USA; and Massachusetts Institute of Technology, Cambridge, MA, USA
E Halgren, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The same fluctuating electrical sources in the brain that produce the scalp potentials of the electroencephalogram (EEG) also produce a magnetic field over the head. A measurement of this field is called a magnetoencephalogram (MEG). Because the MEG and EEG are produced by the same sources, MEG recordings can resemble those of EEG. For example, a single MEG trace (a recording of the fluctuating magnetic field vs. time at one location on the head) can resemble an EEG trace recorded at a related location. However, some MEG recordings look different from equivalent EEG recordings. Thus, an MEG spatial map over the head is quite different from the equivalent EEG map, where each map shows the magnetic field or scalp potential over a large region of the head, at one particular instant in time (sometimes called a snapshot). They are different because the MEG spatially samples the electrical sources differently. This allows the MEG map to provide some different information about the sources in the brain than does the EEG map. In that sense, the MEG is complementary to the EEG, and both are necessary to obtain the maximum amount of information about these electrical sources.

Among the noninvasive techniques for functional imaging of the human brain, the MEG and EEG are unique in that they are directly and instantaneously tied to neuronal currents, the brain's information-processing mechanism. In particular, the MEG and EEG signals are mostly produced by the postsynaptic currents in the pyramidal cells of layers III and V of the cortex. In contrast, functional magnetic resonance imaging (fMRI), positron emission tomography, and single-photon emission tomography have an indirect and delayed relationship to synaptic events through hemodynamics or metabolism. However, the spatial localization of MEG and EEG sources is uncertain, in contrast to definitive localization of the three tomographic methodologies. Other forms of brain imaging, such as MRI and computer assisted tomography, only look at anatomic structures, not functions. Use of the MEG is sometimes called magnetic source imaging.

The brain's fluctuating magnetic field is very weak, commensurate with the weak potentials measured by the EEG. MEG signal levels are typically below 10^{-12} T (tesla), or 10^{-8} G (gauss) in older units; this is much weaker than the urban fluctuating magnetic background of approximately 10^{-7} T or 10^{-3} G and weaker than the earth's steady field of approximately 0.5×10^{-4} T or 0.5 G. Therefore, the two main requirements for measuring the MEG are a magnetic detector of high sensitivity and the suppression of the fluctuating magnetic background (the steady background usually presents no problem).

The sensitive detector used is the SQUID (Superconducting Quantum Interference Device), which operates at 4 K and is therefore contained in a liquid helium dewar. Modern MEG dewars are helmet shaped and enclose most of the head. The dewar typically contains several hundred SQUIDs, where each SQUID is fed by a magnetic sensing coil, and most of these coils are arrayed on a spherical section over the head at grid points 2 or 3 cm apart. This spherical section is at an average distance of approximately 2 cm from the adult scalp, where some of the separation is due to the vacuum space in the dewar. The many sensing coils allow simultaneous magnetic field measurements at the many coil locations over the head, resulting in a continuous series of 'snapshot' maps over the head. The most fundamental choice of sensing coil is a simple wire loop, which measures the component of the magnetic field vector normal to the loop, B_z. One loop connected to its SQUID is called a magnetometer, and if all sensors in a dewar were chosen to be magnetometers, then each snapshot is a map of the magnetic component B_z over the head.

The fluctuating background is suppressed mostly in two ways. First, the measurements are made in a magnetically shielded room, which can exclude almost all of the fluctuating external fields. An early example is shown in Figure 1. Second, the gradient of B_z can be measured instead of B_z itself, where the gradient is the difference in B_z at two locations over the head. These can be neighboring locations along the spherical surface, parallel to the skull, or they can be radially separated locations, perpendicular to the skull. The measurement in each case is said to be due, respectively, to a planar or a radial gradiometer. The planar gradiometer coil consists of two side-by-side loops magnetometer connected in opposition, whereas the radial gradiometer coil consists of two coaxial loops connected in opposition, separated radially. The background is suppressed because the gradient of the background is much reduced compared with that of the brain. In other words, a background source which is distant gives almost equal signals to the two gradiometer coils and hence is largely canceled,

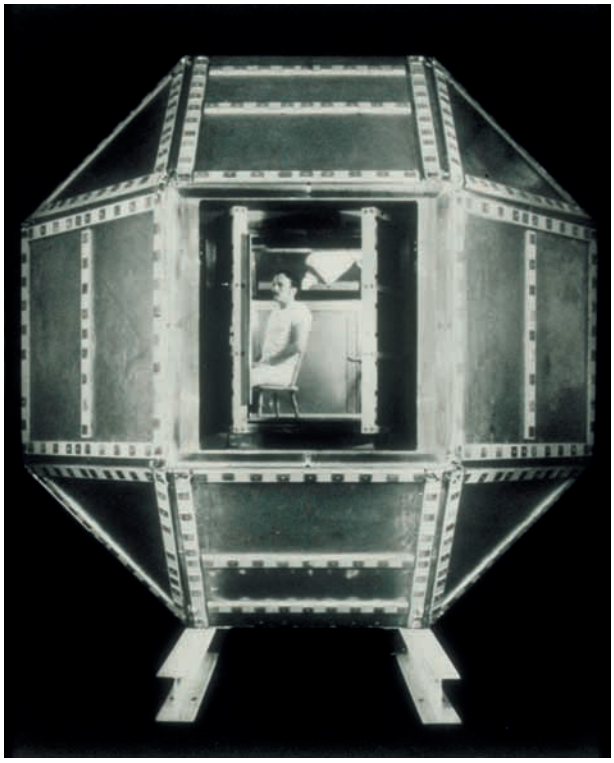

Figure 1 The magnetically shielded room at the Massachusetts Institute of Technology, where the first MEG measurements with the SQUID were made by one of the authors (Cohen) in 1972. This illustrates heavy, effective magnetic shielding. The walls contain five layers of shielding: three are of the expensive high-mu shielding material called moly permalloy, and two are of pure aluminum, for eddy current shielding. The room is roughly spherical because of the efficiency of this shape. This is an example of an effective room that removes almost all urban background, but it is expensive to build. Newer rooms are larger, to accommodate the large helmet dewars, but are box shaped and usually have fewer layers of moly permalloy and hence are basically less expensive. Also illustrated here is the basic setup for measuring the MEG. The subject is shown positioned for recording his MEG and wears clothes that contain no magnetic materials (no zippers or shoes), which could cause relatively large magnetic disturbances. The large white cylinder near the subject's head is the liquid helium dewar, and this early dewar contains only a single SQUID detector in the tail. At that time, before whole-head dewars were developed, maps were made by tedious sequential measurements over the head.

whereas the field from the nearby brain is different in each and hence is not canceled. A good, multilayer room can eliminate the background almost completely but adds expense, whereas the gradiometer method, although less expensive, cannot eliminate background as well. Each MEG system is a compromise among these various factors and can contain magnetometers, gradiometers, or a combination of these, housed in a particular choice of shielded room. The background can also be reduced in a third way – via software processing of the MEG signals, where the software is able to recognize signals from an outside source and reduce or eliminate them. The cost of an entire MEG system, including shielded room, SQUID system, hardware, and software, is approximately $3 million. For example, see Figure 2.

Using the MEG

MEG measurements are made in the following way. First, all magnetic material on the subject's body is removed or demagnetized, and the subject changes into magnetic-free clothes. Fiducial locations on the head are accurately noted; in one system, three small coils are attached, which will be continuously sensed by the MEG for accurate head location. If the EEG is to be recorded simultaneously with the MEG for complementary information, EEG electrodes are now applied to the scalp. Next, the subject, seated or in the supine position, is placed with his or her head in the helmet dewar, as shown in Figure 2. Next, the raw recordings are made, which consist of MEG traces on all channels, due either to evoked neural activity or to spontaneous activity, depending on the purpose of the session.

Then, in most cases, steps are taken to extract the signal from the noise, where the noise is often unwanted MEG signal from areas of the head not being measured. Traditionally, this entails averaging the signal with respect to a repeated event (e.g., sensory stimulus onset) to produce an 'event-related field' completely analogous to the 'event-related potential' derived by averaging EEG. Increasingly, such 'time-domain' processing is supplemented by 'frequency domain' analyses in which the power in different frequency bands is estimated at different latencies after the stimulus and then averaged across trials. Such processing permits activity that is not strictly phase-locked to the stimulus to be studied. Finally, a sequence of spatial maps (snapshots) are extracted from the trace maps, each for a different instant in time, and used to estimate the neural sources in the brain that are generating the signals. This estimate is termed the 'inverse solution' and is usually the purpose of the measurements.

Unfortunately, inverse solution MEG (and EEG) maps have no unique solution and are handled as follows. All inverse solutions require, in addition to the MEG field map, two elements. First, they require a biophysical model for calculating how activity in any particular brain region would appear at the sensors. Such 'forward calculations' vary in regard to how accurately they model the shape of the head and the conductance in its various compartments. All inverse solutions also require *a priori* assumptions, allowing them to choose between the infinite

Figure 2 An example of a state-of-the-art MEG system. Although not apparent, this system is housed in a large shielded room. The entire structure at the subject's head is called a gantry, and it is set to perform an MEG scan of a subject in the supine position. The gantry can also be swung upward when needed to accommodate a subject in the sitting position. As in **Figure 1**, the subject wears no magnetic material. The subject's head is shown positioned in a helmet-shaped liquid helium dewar, as part of the gantry. This dewar contains 306 sensors surrounding the head, illustrated in the inset at left, oriented as if the patient was in the sitting position (head vertical). The sensors consist of 306 'pickup coils' centered at 102 locations over the head, with three different coils at each location. One coil is a simple outside square and hence a magnetometer; the other two are rectangular planar gradiometers, oriented at right angles to each other. Three SQUIDS are part of each square. Thus, the entire system consists of 102 magnetometers and 204 planar gradiometers for a total of 306 MEG channels. This MEG system, called Vector View 306, is made by Elekta AB (Finland). Other manufacturers supply similar systems, each with its own distinctive features.

number of current sources that can produce any given MEG (or EEG) map. The simplest inverse solutions assume that the field is generated by a single focal source, called a current dipole, that is defined by its position, orientation, and strength. Such assumptions are strictly accurate for only the first few milliseconds after the initial arrival of activity in sensory cortex – for example, prior to 20 ms in somatosensory cortex. However, in many cases the MEG field can be well approximated by an equivalent current dipole even when it is known to be generated in multiple locations of varying extent. Other *a priori* assumptions may concern how many brain areas are contributing to the solution, their focality, and/or their correlation with one another.

There exists extensive literature on MEG inverse methods. Some methods estimate the activity at all brain locations, whereas others restrict possible sources to dipoles that are in the cortex and perpendicular to its local surface, as determined from the MRI in the same subject. This assumption is reasonable because the MEG is mainly generated by currents within the apical dendrites of pyramidal neurons, which are oriented perpendicularly. These currents may be from ligand-gated channels at excitatory or inhibitory synapses or from voltage-gated active conductances. Action-potentials do not contribute significantly to the MEG due to the random (canceling) spatial arrangement of their generating currents and their lack of synchrony relative to their short duration; further, their 'ends' cancel internally. Note that although MEG and EEG are generated by current sources, most sources do not contribute to either one because of cancellation at many spatial levels within the same dendrite, cell, or cortical column

or between cortical areas. Thus, even if the inverse estimate were accurate, it would show only the uncanceled sources.

Inverse methods often use additional assumptions to localize activity, such as the assumption that the brain minimizes the total power across its surface or total current. Such assumptions have little, if any, biological justification. Consequently, it is important to seek validation of proposed solutions using more certain methods. The most direct validation can be obtained from intracranial EEG recordings during the same tasks as those recorded with MEG. Although certain localization can be obtained with intracranial EEG, such recordings are only obtained in the context of severe epilepsy, and only in clinically indicated locations, so issues of pathology and sampling limit their availability and interpretation. In healthy subjects, the reasonableness of MEG inverse estimates can be judged against fMRI results in the same task. However, there are many reasons why MEG and fMRI signals may be generated in different locations, so the lack of correspondence is not definitive. Conversely, the locations of fMRI activations may be used to bias MEG inverse estimates, provided that the method allows for mismatches between the underlying generators.

Differences between MEG and EEG

There are three differences between MEG and EEG instantaneous spatial maps that allow the MEG to give different information from the EEG. These differences are calculated when the head is idealized as a system of concentric spherical shells representing the scalp and skull, etc. Also, the current source generating the MEG and EEG is idealized to be a dipole, the simplest source possible (a more complex neural source can always be synthesized as a sum of dipoles). In the following, we discuss the calculated MEG and EEG maps and compare them with regard to their ability to determine the location and angle of the dipole which generated them – that is, to localize the originating dipole.

The first difference is due only to the orientation of the dipole. For a dipole oriented radially to the spherical surface (skull), it can be shown, surprisingly, that the external magnetic field is everywhere zero; hence, there is no MEG signal. This is due to the symmetry cancellations in perfectly spherical conductors. Therefore, there is only an MEG signal for a dipole oriented tangentially, where there is no perfect symmetry cancellation. Because the MEG/EEG sources are mostly pyramidal cells in the cortex, and these have a net direction perpendicular to the cortical surface, it follows that sources radial to the skull are mostly located on the gyri and tangential sources mostly in the sulci, as shown in Figure 3. Thus, the MEG is largely due to the sources in the sulci; the EEG is due to both. Therefore, one consequence of this difference is that the MEG should see those tangential sources that on the EEG may be dominated or masked by strong radial sources. Indeed, radial sources tend to dominate the EEG because on the gyri they are closer to the electrodes on the scalp. The fact that radial sources produce zero external field has an important consequence for sources deep in the head. Because a dipole at the center of a sphere is always a radial source and hence produces zero external field, it follows that the MEG due to deep sources is suppressed compared to the EEG; the deeper the source, the more the suppression. Because of this radial–dipole difference, the MEG map is due to a subgroup of the sources seen on the EEG. That is, the MEG sees less than does the EEG.

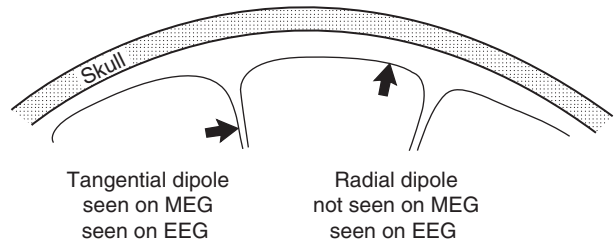

Figure 3 Two electrical sources in the cortex; one is oriented tangentially to the spherical skull and the other radially. Each is represented as a current dipole (arrows). Because sources in the cortex are oriented perpendicularly to the cortical surface, tangential sources are located in the sulci and radial sources are in the gyri. Although the EEG sees both sources, it is dominated by the radial sources in the gyri. If the radial dipole were shifted, for example, to the right, slightly into the bend, it would then have a small tangential component and hence would now be visible on the MEG. Thus, the gyral area of invisibility is less than is apparent. Also, in cases in which the lower ends of sulci bend around, there are radial–tangential complications.

The other two differences in MEG and EEG maps involve actual map patterns due to a tangential dipole. These are illustrated in Figure 4. First, the MEG pattern is seen to be oriented perpendicularly to the EEG pattern due to a basic orthogonality between magnetic and electrical fields. Consequently, MEG localizes a source better in the y direction, whereas the EEG localizes better in the x direction. Second, the MEG pattern is somewhat tighter, largely due to the smearing by the high-resistivity skull of the surface potentials measured by the EEG. In contrast, for perfectly spherical surfaces, the MEG is unaffected by the resistivities of the different layers. The consequence of this size difference is that the MEG localizes a tangential dipole somewhat better in its best direction (y), in comparison to the EEG in its best direction (x), if all other factors are equal, such as

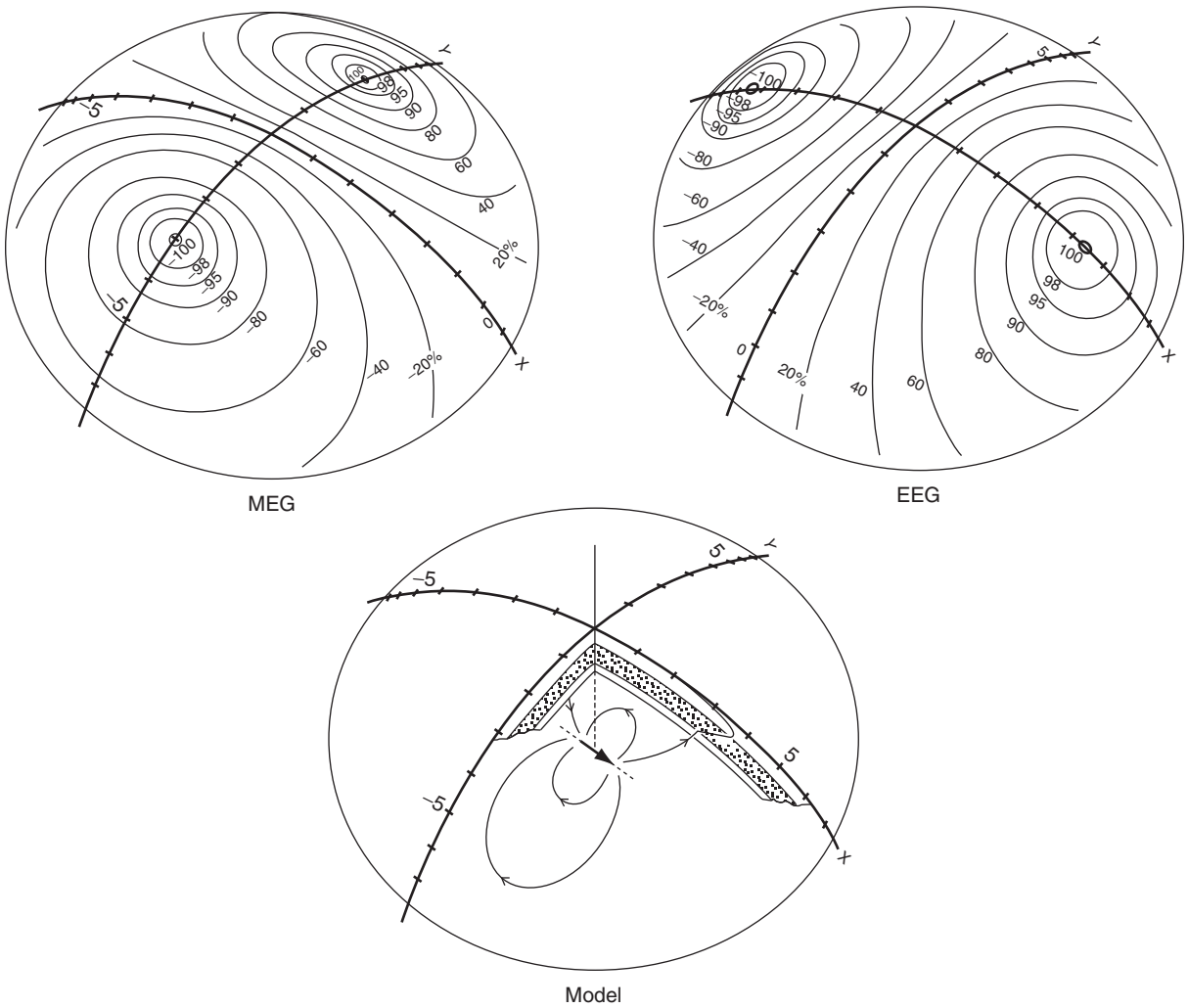

Figure 4 Difference between the MEG and EEG maps, computed by using an ideal spherical model of the head containing a tangential dipole source (arrow) at a depth of 2.7 cm. The model cutaway shows four spherical boundaries representing scalp, skull, cerebrospinal fluid, and brain. The continuous spreading currents generated by the dipole are seen as thin lines. The maps are shown on a spherical 'cap,' which has *x* and *y* coordinates with ticks at each centimeter, with contour lines every 20%. The MEG map is of the component of magnetic field vector normal to the surface (B_z), whereas the EEG map is of the surface potential *V*. The MEG map is different from the EEG map in that it is oriented by 90° and is tighter by approximately 30%. It can therefore localize somewhat better in its 'best' direction (*y*) compared with the best direction (*x*) in the EEG, according to this ideal model.

modeling errors – for example, errors in knowing the conductivities of all the layers. However, such modeling errors are considered worse for the EEG, which thus favors the MEG.

Thus, all in all, MEG advantage in localizing a dipole is in the range of about 30% better (dipole shown here) to perhaps a factor of 2 better, depending for example, on resistivity errors in the EEG modeling. Also, in the actual nonspherical head, the difference between radical and tangential dipoles becomes diffused. Although there have been recurrent claims that the MEG can localize a dipole source far better than can the EEG – for example, to within 2 mm (because of reduced sensitivity to conductivity errors) – there is no experimental evidence for this claim in the human head. The one reported experiment under these conditions, using implanted current sources in the brain, has shown only a minor advantage, yielding 8 mm accuracy for the MEG versus 10 mm for the EEG. Also, several advanced computer modeling studies have shown no significant MEG localization advantage. Finally, because MEG calculated sources are laid onto MRIs of the subjects, there is a co-registration error from MRI to MEG of approximately 3 mm. In brief, the 2 mm claim is unproven and unreasonable. However, the MEG has contributed to EEG source localization in an unexpected way. Because the MEG was developed in physics labs, advanced techniques were used for inverse solutions on the maps to find the sources. EEG data had usually been interpreted

by visual inspection, occasionally with large errors. However, EEG manufacturers began duplicating the MEG inverse solutions, and now include such software in EEG systems, so that EEG localization is now more accurate.

Based on the three differences discussed above, it can simply be stated as "the MEG usually sees less than the EEG sees, but sees it somewhat more clearly".

There are other MEG–EEG differences that are of a more practical nature. One difference, favorable to the EEG, is the MEG nuisance of embedded magnetic material in the subject, such as magnetic dental work, which creates large interfering magnetic signals. This material can often be demagnetized with a handheld magnetic eraser. However, there are differences that are favorable to the MEG. One is that the MEG does not require electrodes to be pasted on the head, as does the EEG. Perhaps a more important difference concerns DC; galvanic skin (scalp) potentials severely hamper EEG DC measurements, but the MEG does not see these obstacles. In this regard, the MEG has a distinct advantage that has yet to be fully exploited.

Current Status and Future Prospects

There are approximately 120 whole-head MEG systems in use or on order throughout the world, some with more than 300 SQUIDs per helmet. Many of these systems are located in Germany, Japan, the United States, Finland, and Canada. Thus, an increasing number of MEG maps are being recorded, often in combination with EEG maps, due both to spontaneous signals and to event-related signals. The overall purpose is to determine if the selectivity of the MEG can clarify the sources of the signals. The efforts are divided into two broad areas: research on the workings of the normal brain and the search for clinical applications of the MEG.

In research on the normal brain, MEGs are being studied for responses evoked by stimulation of each of the five senses. To date, MEG has produced evidence supporting locations of sources in the primary sensory cortices. One early example is the well-studied 20 ms somatosensory signal from the human brain in response to peripheral nerve (wrist) stimulation. From the EEG alone, the source of this signal had been ambiguous, with a choice between two radial or one tangential sources. However, the combination of MEG and EEG maps showed the source to be mostly a tangential dipole. In another example involving somatosensory cortex, MEG and EEG were combined to study the finger representation in the cortex. It was found that MEG and EEG combined gave more localization accuracy than each separately. Some of the studies used MEG without EEG; one such study reported increased strength of cortical sources of left fingers of string players. Without EEG maps, however, there can be doubt that the useful information obtained was indeed due to MEG selectivity.

Overall, MEG has demonstrated internal organization in some of the primary areas of the human brain – organization that was previously seen only in the animal brain. MEG has also shown the separation of activity between primary and secondary areas, for example, in the somatosensory cortex and in the auditory cortex. Interactions between successive cortical areas are increasingly being inferred from the phase-locking, coherence, or Granger causality, calculated on individual trials between waveforms recorded at different locations.

MEGs are being studied with regard to the later components as well. After processing in these initial sensory areas of the brain, there are further processors that decode certain classes of stimuli, such as faces. MEG has been able to localize the face area to the same location identified with fMRI and intracranial EEG and has been used to characterize its cognitive responses to a range of face-like stimuli (**Figure 5**). The situation becomes more complicated when moving further downstream to neural activity underlying higher cognitive functions, such as understanding sentences. The fact that neural activity can spread to distant cortical areas in approximately 20 ms suggests that multiple widespread cortical areas may be involved in the response to words, which peaks at approximately 400 ms. Modeling in such situations with a distributed source instead of a dipole (**Figure 6**) localizes MEG activity to the same locations that have been found with hemodynamic measures and with direct intracranial EEG recordings in epileptic patients. Data suggest that in situations such as these, efforts to localize distributed generators appear to have been more successful with MEG than with EEG. If true, this may be due to the fact that, as previously mentioned, the MEG sees fewer sources with a somewhat tighter field than EEG and thus is better able to disentangle multiple distributed sources. However, it is also possible that with better forward models and inverse methods, EEG will be equally successful in localizing complex distributed generators. In any case, modeling studies and some experiments clearly suggest that the best localization will only be obtained by combining MEG and EEG.

Concerning clinical application of MEG, efforts are being made to evaluate MEG usefulness in various areas, including presurgical functional mapping, head injury, and epileptic spike localization. The purpose of presurgical mapping is to locate the regions of the sensorimotor strip using evoked response in order to

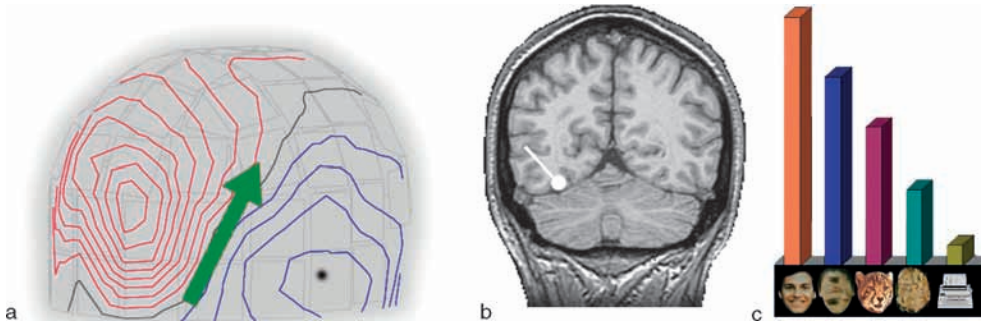

Figure 5 Use of MEG in face recognition research. (a) MEG contour map (50 fT step/line) drawn on the head of the subject and sampled at 165 ms after a photo has been viewed. The head is viewed from the right rear; the dark spot marks the right ear. The pattern is seen to be roughly dipolar; hence, a single dipole was assumed for the source. An inverse solution located the dipole(green arrow), where the depth is not seen here. (b) The same dipole source superimposed on the subject's MRI, seen from the front. Here, the dipole is shown as a white dot, for accurate location, in the right fusiform gyrus. The attached rod gives the dipole direction and amplitude. (c) The dipole amplitudes after viewing any of five face photographs and controls, averaged over ten subjects. The normal human face evokes the largest response. The MEG can thus be used to estimate the location, timing, and strength of a neural process associated with face encoding (Halgren et al.).

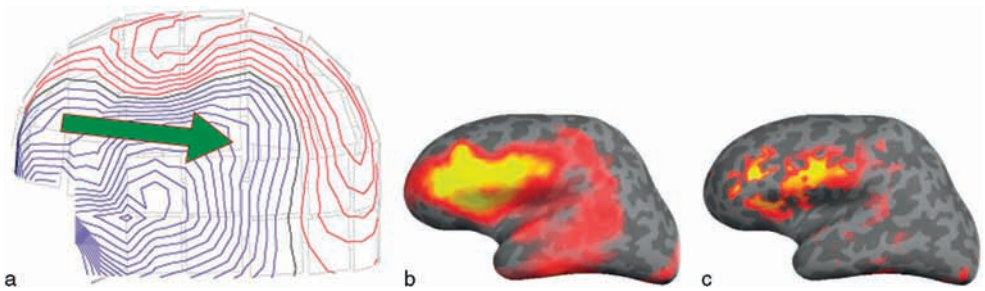

Figure 6 Use of the MEG with multimodal integration in reading research. Subjects read words and made a semantic judgment (pressing a key if the word referred to an object or animal more than 1 foot in length). MEG maps were produced by subtracting the signals evoked by novel words from those evoked by words that had been read previously in the same task. (a) MEG contour map, viewed from the subject's left, at 540 ms after stimulus onset. Again, the pattern is roughly dipolar, and the green arrow shows the computed source dipole. It appears 'off-center' in this case because of the angle of view. (b) A determination of the sources, from the same contour map, using a nondipolar advanced inverse method allowing a continuous distribution, but sources are now constrained to be on the cortical surface (Dale et al.). The 'inflated' left cortical surface is shown in the same view as in (a). (c) The same data as b, except the source has been further biased by the fMRI location of activation in the same subject and task. Note the increased spatial detail revealed by the integration of the fMRI information in the same task. The MEG data thus reflect the repetition priming effect on semantic processing.

minimize neurological deficit due to surgery. MEG localizes well in this task, as does fMRI. In some cases, the hemodynamic response near a tumor or arteriovenous malformation may be abnormal, and MEG may thus be preferable. In head injury, MEG appears to show focal slow-wave abnormalities more reliably than does EEG; hence, it shows more promise of providing localizing information. In the measurement of epileptic spikes, MEG, when used with EEG, is able to clarify or localize some epileptic foci better than can EEG alone. MEG has also shown spikes that appear to be masked on EEG by other spontaneous brain activity, seen less on MEG. Definitive localization of seizure foci, adequate to guide their surgical removal, often requires the placement of intracerebral electrodes. The hope is that the noninvasive combination of MEG and EEG may allow such electrodes to be more successfully targeted or, in some cases, to be avoided altogether. MEG also appears useful when a skull defect from a prior operation severely distorts the propagation of EEG to the scalp. In this regard, MEG of neonates, which have begun to be measured with smaller helmets, has an advantage in source localization because the fontanelles in the infant skull cause larger errors in EEG. Also, MEG research studies on neuropsychiatric disorders such as schizophrenia, autism, and pain often show promise because of MEG's greater specificity compared with EEG. These studies can provide insight into the neural bases of severe dysfunction in the absence of any structural abnormality.

Expansion of clinical MEG use from regional referral centers to community hospitals will require further evaluation.

Concerning future possibilities, work measuring low-field MRIs using SQUID detectors shows promise for simultaneous measurement of MEG and MRI, eliminating the coregistration error (~3 mm). The future possibilities of MEG depend on how well its advantages will balance against its practical problems and high cost. The practical problems include not only the reduction of magnetic background but also the effort in maintaining a cryogenic system that needs refilling with liquid helium once or twice per week. A cost reduction would make MEG more attractive for clinical diagnosis. In any case, MEG will be used well into the future, both because the large MEG systems now in use or coming online will generate many years of investigations and because there is a general, long-term need for noninvasive measurements of the brain.

See also: Cognition: An Overview of Neuroimaging Techniques; Electroencephalography (EEG); Electrophysiology: EEG and ERP Analysis; Event-Related Potentials (ERPs) and Cognitive Processing; Evoked Potentials: Recording Methods; Human Depth Electrodes; Neuroimaging.

Further Reading

Barkley GL and Baumgartner C (2003) MEG and EEG in epilepsy. *Journal of Clinical Neurophysiology* 20: 163–178.

Cohen D (1972) Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer. *Science* 175: 664–666.

Cohen D and Cuffin BN (1983) Demonstration of useful differences between magnetoencephalogram and electroencephalogram. *Electroencephalography and Clinical Neurophysiology* 56: 38–51.

Cohen D, Cuffin BN, Yunokuchi K, et al. (1990) MEG versus EEG localization test using implanted sources in the human brain. *Annals of Neurology* 28: 811–817.

Dale AM and Halgren E (2001) Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. *Current Opinion in Neurobiology* 11: 202–208.

Dale AM, Liu AK, Fischl B, et al. (2000) Dynamic statistical parameter mapping: Combining fMRI and MEG to produce high resolution imaging of cortical activity. *Neuron* 26: 55–67.

Haddad N, Shihabuddin B, Lowery CL, Holst M, and Eswaran H (2006) Magnetoencephalography in healthy neonates. *Clinical Neurophysiology* 117: 289–294.

Halgren E, Raij T, Marinkovic K, Jousmäki V, and Hari R (2000) Cognitive response profile of the human fusiform face area as determined by MEG. *Cerebral Cortex* 10: 69–81.

Hamalainen M and Hari R (2002) Magnetoencephalographic characterization of dynamic brain activation: Basic principles and methods of data collection and source analysis. In: Toga AW and Mazziotta JC (eds.) *Brain Mapping, the Methods*, pp. 227–253. San Diego, CA: Academic Press.

Hillebrand A and Barnes GR (2002) A quantitative assessment of the sensitivity of whole-head MEG to activity in the adult human cortex. *NeuroImage* 16: 638–650.

Knake S, Halgren E, Shiraishi H, et al. (2006) The value of multichannel MEG and EEG in the presurgical evaluation of 70 epilepsy patients. *Epilepsy Research* 69: 80–86.

Lopes da Silva D (2004) Functional localization of brain sources using EEG and/or MEG data: Volume conductor and source models. *Magnetic Resonance Imaging* 22: 1533–1538.

Pantev C and Lutkenhoner B (2000) Magnetoencephalographic studies of functional organization and plasticity of the human auditory cortex. *Journal of Clinical Neurophysiology* 17: 130–142.

Sharon D, Hämäläinen MS, Tootell R, Halgren E, and Belliveau JW (2007) The advantage of combining MEG and EEG: Comparison to fMRI in focally-stimulated visual cortex. *NeuroImage* 36: 1225–1235.

Vrba J and Robinson SE (2001) Signal processing in magnetoencephalography. *Methods* 25: 249–271.

Relevant Websites

http://www.nmr.mgh.harvard.edu – Athinoula A. Martinos Center for Biomedical Imaging.

http://www.biomag.hus.fi – BioMag Laboratory at Helsinki University Central Hospital.

http://www.nips.ac.jp – Bioorg Center at Okazaki, Japan.

http://jenameg10.meg.uni-jena.de – Biomagnetic Center at Jena.

http://www.elekta.com – Elekta AB Company.

http://www.4dneuroimaging.com – 4-D Neuroimaging Company.

Magnocellular Neurosecretory System: Organization, Plasticity, Model Peptidergic Neurons

G I Hatton, University of California, Riverside, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Historical Perspective on Neurosecretion

Studies of the magnocellular hypothalamo-neurohypophysial system (mHNS) began before anyone was aware that there was such a system. Cells of the pituitary gland were firmly believed to be the sources of all factors controlling the functions now known to be carried out by peptides that originate in the brain. One of these functions, the 'oxytocic principle,' as he called it, was discovered quite serendipitously in 1909 by Henry Dale, widely known as the propounder of Dale's principle of neurotransmitter release. Dale found that smooth muscle contractions resulted when posterior pituitary extracts were placed on strips of virgin guinea pig uterus. Similar pituitary extracts were shown by others to cause antidiuresis and pressor effects. Not until du Vigneaud and colleagues in the 1950s sequenced the peptides oxytocin (OT) and vasopressin (VP) was it determined that two separate peptides were responsible for the different actions of neurohypophysial extracts. This ambiguity had endured because, at the high concentrations that were typically used, each of these peptides has some action at the other peptide's receptors.

At about this same time, an important histological advance, the eponymous Gomori stain (chrome alum hematoxylin) was being used to investigate brain areas that are rich in disulfide bond-containing compounds, such as those in the neurohypophysis. This work was done by the team of Wolfgang Bargmann and Ernst and Berta Scharrer. To the disbelief of the entire field of endocrinology, they showed that the active substances in the posterior pituitary actually originated in the hypothalamus, in cells that were best described as neurons. Difficult to accept in those early days was the idea that substances were made in neurons up in the brain and transported down the axons to remote release sites in the neurohypophysis. Furthermore, these neurons had axons that terminated, not in the usual synapses, but nonsynaptically near blood vessels, and these neural terminals in the posterior pituitary were the sites of peptide hormone storage. Thus was born the concept of neurosecretion. Ultimate acceptance of this concept would come only after much debate and the completion of many functional studies. Even then, it was hard for some to believe that the electrical activity of neurons had anything to do with hormonal responses to physiological challenges. As late as the early1970s, those neuroendocrinologists having an electrophysiological bent always felt compelled to point out to their still skeptical colleagues evidence that confirmed the close positive relationship between neuronal electrical activity and peptide release. This, of course, is widely accepted today.

These neurosecretory cells and their main release products, OT and VP, attracted the attention of a relatively small but active cadre of scientists around the world, whose collective work over the years became a large database that supported increasingly sophisticated analyses of this system. Ultimately, this accumulated database facilitated meaningful answers to questions about the mHNS that were not so easily extractable from others, and the mHNS became a model system for the study of neurosecretion and of the brain's peptidergic neurons.

Ubiquitous in the brains of mammalian species, the mHNS and its functions in peripheral tissues have been most thoroughly characterized in the rat. Much of what has been discovered and described in the rat, however, has been generalizable to the other species that have been investigated. Anatomically, this includes the types of neurons and glial cells in the mHNS, as well as their relationships to one another, and the relationships of the various cellular compartments comprising the mHNS to the brain vasculature. Capillary density is generally found to be high in peptidergic systems, owing largely to their elevated metabolic rates. Capillary density in the brain reaches its zenith in the mHNS.

Basic Organization of the mHNS

Shown in Figure 1 is a three-dimensional representation of a rat brain cut away at the level of the anterior hypothalamus to reveal two prominent mHNS nuclei: the paraventricular nuclei (PVN), positioned on either side of the dorsal third ventricle, and the supraoptic nuclei (SON), lateral to the optic tracts. In the inset at lower left, a Nissl-stained section reveals the dense packing of the magnocellular cell bodies in these two nuclei. Projection neurons of the SON are all magnocellular and are either OTergic or VPergic. The PVN consists of groups of both magnocellular (laterally) and parvocellular (medially) neurons (see inset in Figure 1). OT and VP neurons of the former groups project axons laterally and ventrally to join those of the dorsomedially projecting SON neurons.

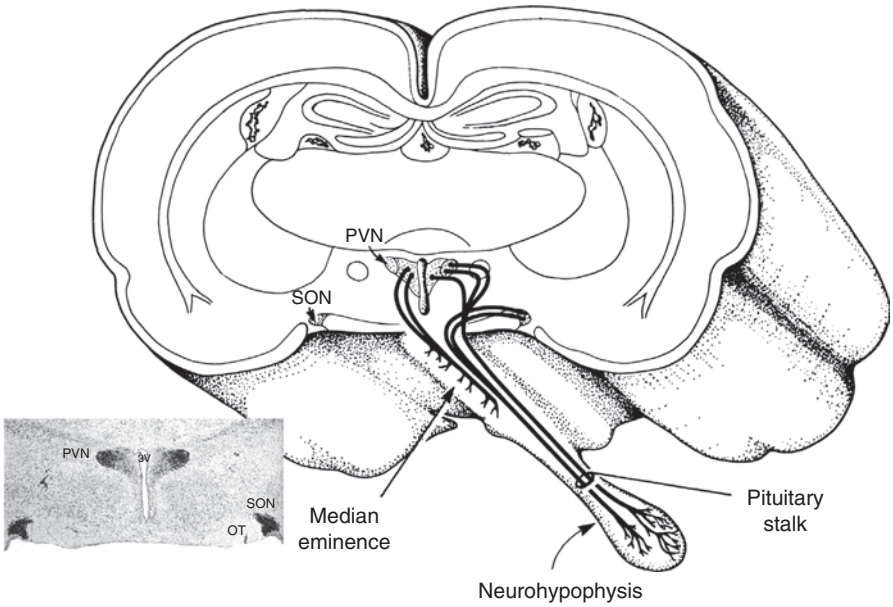

Figure 1 Basic organization of the hypothalamo-neurohypophysial system. Three-dimensional representation of rat brain at the level of paraventricular (PVN) and supraoptic nuclei (SON) in the anterior hypothalamus. Major axonal projections of the magnocellular neurons via the pituitary stalk to the neurohypophysis are shown. PVN parvocellular axonal projections to the median eminence are also diagrammed. Inset: Nissl-stained section taken at the same level as shown in the diagram. Magnification 60×. 3V, third ventricle; OT, optic tract.

Together, all these axons then project posteromedially and ventrally in the internal zone of the median eminence to travel in the pituitary stalk. Principal axonal outputs of the magnocellular neurons in these two nuclei are shown in the diagram projecting via the pituitary stalk to terminate in the neurohypophysis. Also shown are projections from the parvocellular neurons of the PVN (not part of the mHNS) to the external zone of the median eminence, where, along with projections of peptidergic neurons from other hypothalamic cell groups, they influence the anterior pituitary.

The focus in this article is on the neurons, glia, and other important structural elements of the SON and neurohypophysis since most of the more enlightening recent developments relative to the topic being covered have involved these portions of the mHNS. It is important to note that the presence of readily identifiable peptides, either OT or VP, makes it possible to study the dendrites and somata in the hypothalamus of unequivocally the same neurons, the axons and terminals of which can be investigated in the neural stalk and neurohypophysis. This advantage is found in few other places in the mammalian central nervous system.

Organization of the SON

Among hypothalamic nuclei, the SON is one of the most highly organized, another factor contributing to its model status. As alluded to earlier, SON neurons are magnocellular and normally synthesize either OT or VP. There are, however, some special circumstances under which a subset of SON neurons synthesizes both peptides. Figure 2(a) is a diagrammatic representation of some of the prominent features of SON organization. The right SON is shown in Figure 2(a); thus the nucleus lies to the right of the optic tract. The nucleus has been partitioned into three zones. Dorsalmost is the somatic zone (SZ). Magnocellular neurons here are shown to be tightly packed, generally bipolar cells with primary dendrites tending to project dorsolaterally or ventromedially. Axons arise from the dorsal primary dendrites and exit the nucleus dorsally or dorsolaterally, as shown in the immunocytochemical stain for both OT and VP in Figure 2(b). Returning to Figure 2(a), ventral to the SZ, the dendritic zone (DZ) consists primarily of the ventrally projecting dendrites that have turned to run parallel to the brain surface and are, therefore, shown here in cross section. Not shown in the simplified diagram, but quite salient in the immunostained image of Figure 2(b), is a spur of SON dendrites that projects ventrolaterally along the pial surface of the periamygdaloid cortex. These dendrites receive direct synaptic input from the mitral cells of the main and accessory olfactory bulbs. Subjacent to the DZ lies a line of astrocyte cell bodies, shown in the diagram as filled ovoids. If projected in three dimensions, this line of astrocytes becomes a sheet that underlies the entire nucleus. Astrocytic processes from these glia fill the spaces between the dendrites, between the somata, and between the neural compartment and the basal

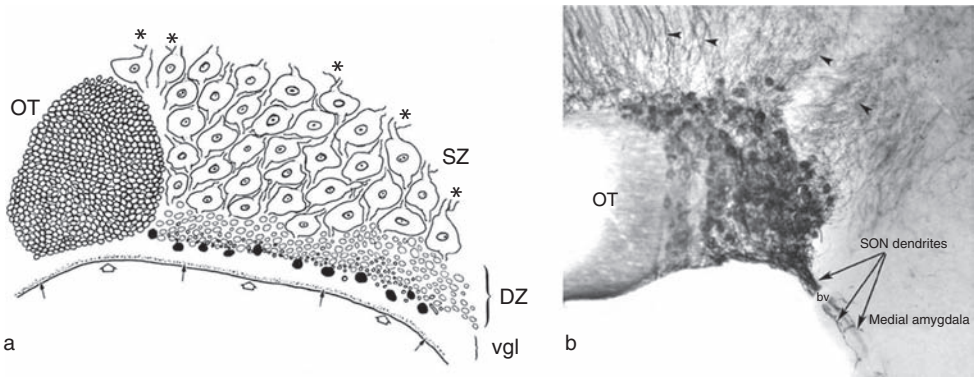

Figure 2 Basic organization of the supraoptic nuclei (SON). (a) Semidiagrammatic representation of the right SON with the myelinated fibers of the optic tract (OT) shown on the left. Large somata lie dorsally, their dendrites truncated, in the somatic zone (SZ). Primary dorsal dendrites are shown giving rise the axonal processes (*). Ventrally projecting dendrites turn to course anteroposteriorly in the dendritic zone (DZ), parallel to the pial surface (open arrowheads). Subjacent to these dendrites lie the cell bodies (filled ovoids) of the astrocytes that constitute the ventral glial lamina (vgl). Processes of these astrocytes fill the clear space between the black ovoids and the basal lamina (small arrows). (b) SON cells stained immunocytochemically for oxytocin and vasopressin. Individual axons leaving the SON are indicated by arrowheads. A special group of SON dendrites projects along the periamygdaloid cortex. These dendrites receive olfactory inputs. Magnification 120×. bv, blood vessel.

lamina lining the pial surface of the brain. Important details are shown in Figure 3. Although there are conventional, star-shaped astrocytes in the more dorsal regions of the SON (not shown in the diagram), it is these ventrally positioned astrocytes in the ventral glial lamina (vgl) whose processes are most involved in the plasticity and remodeling of the nucleus in response to physiological stimulation (discussed in the section titled 'Morphological plasticity: Function-related remodeling in the mHNS').

Figures 3(a) and 3(b) presents two different perspectives on the ventral portion of the SON. On the left, an electron micrograph shows the astrocytic membrane filling the space of the vgl and surrounding a group of ventral dendrites (seen in cross section). Note that the folds of astrocytic membrane are oriented in a predominantly vertical (dorsoventral) direction. Dual fluorescence immunostaining for OT (VP neurons appear as black holes) and for the astrocyte marker glial fibrillary acidic protein (GFAP) are shown in the confocal micrograph on the right. Since staining of the intermediate filament, GFAP, reveals only about 15% of the astrocyte volume, the space-filling capacity of these astrocytes is not well represented in Figure 3(b) (cf. Figure 3(a)). Well shown in Figure 3(b), however, are the astrocytic processes projecting dorsally to separate adjacent OT dendrites and somata. Individual processes from these ventrally positioned astrocytes usually project all the way to the dorsal aspects of the nucleus, that is, they span the entire dorsoventral extent of the SON neuronal compartment.

Organization of the Neurohypophysis

Dominating the neurohypophysial landscape are the pituicytes, specialized astrocytes that constitute the only resident cell population there. Although they can be identified by their expression of GFAP, pituicytes are also morphologically distinctive in having large lipid inclusions (see Figure 4). As is true of astrocytes in the brain, the pituicytes tend to occupy the basal lamina that separates neural from non-neural tissue, such as the neural from the vascular compartment. Unlike the SON, in the posterior pituitary, the vascular compartment contains fenestrated capillaries. Both the pituicytes and the neuronal elements found there are, therefore, located outside the blood–brain barrier. Neurohypophysis-bound axons of mHNS neurons arborize on entering the organ. Individual terminal branches and axonal endings can be found either within a pituicyte (Figures 4 and 5), among pituicytes, or adjacent to the basal lamina along the perivascular space. It was most likely these neurosecretory granule-containing axons, completely engulfed by pituicytes, that misled earlier investigators into the belief that the pituicytes were the secretory cells of the neurohypophysis. Various types of neurosecretory terminal endings, showing their relationships with the other structures, are depicted in Figures 4 and 5.

Morphological Plasticity: Function-Related Remodeling in the mHNS

Early work had established that the dehydration attendant on water deprivation or saline drinking for several days resulted in substantial increases in the somatic volumes of SON neurons. Not unexpectedly, rehydration after a period of dehydration reversed the changes wrought by the dehydrating

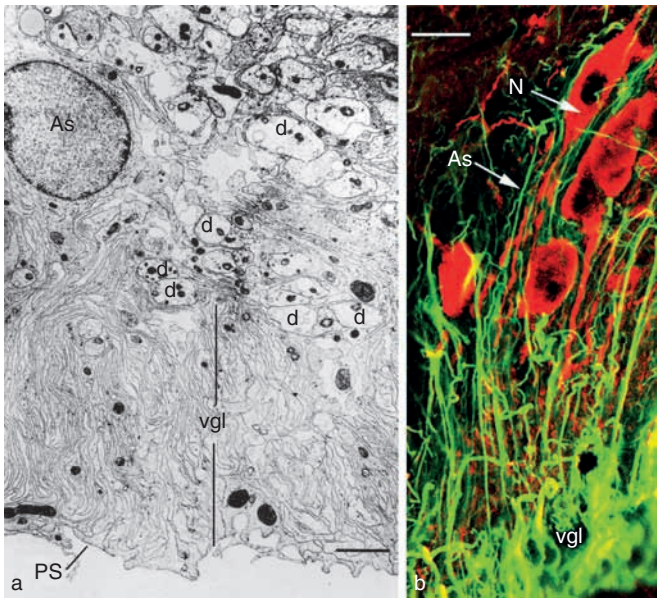

Figure 3 Astrocyte occupation of the ventral supraoptic nuclei (SON). (a) Electron micrograph of the ventral glial lamina (vgl) showing astrocytic membrane filling the space between the ventralmost distal dendrites (d), appearing in cross section, and the pial surface (PS) of the brain. As, astrocyte nucleus. (b) Confocal image of the SON, including the ventral somatic zone (SZ), the dendritic zone (DZ), and the vgl areas diagrammed in (a). Immunostained are optic tract (OT) neurons (N) in red and glial fibrillary acidic protein in green. Astrocytic processes (As) projecting toward the dorsal SON separate adjacent OT neuronal dendrites and somata. Scale bar = 2 μm (a), 20 μm (b).

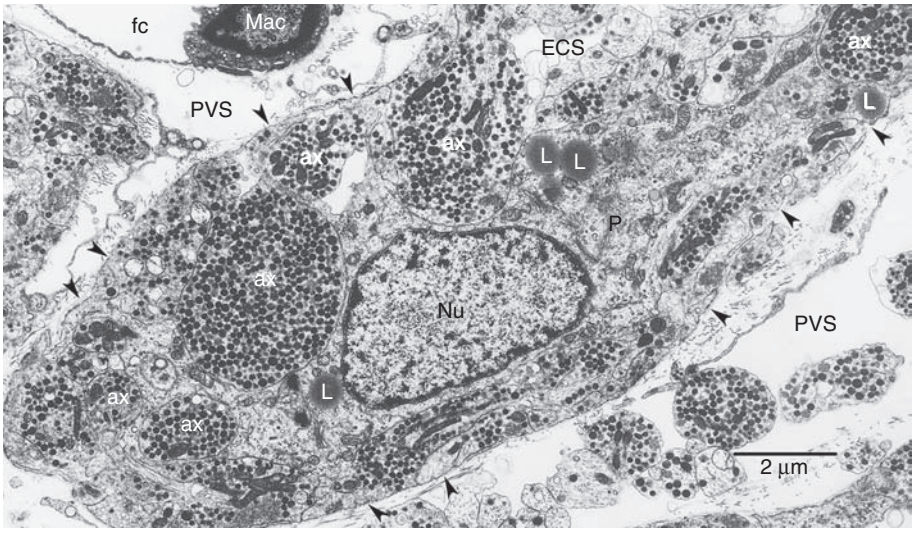

Figure 4 Under basal conditions, a neurohypophysial astrocyte, pituicyte (P), lies adjacent to the perivascular space (PVS) near a fenestrated capillary (fc) and occupies a large percentage of the local basal lamina (arrowheads). Several neurosecretory axons (ax) lie surrounded by pituicyte cytoplasm. Small amounts of extracellular space (ECS) are visible. L, lipid bodies; Mac, perivascular macrophage; Nu, pituicyte nucleus.

stimuli. Such morphological alterations paralleled the changes in electrical activity that were observed under similar dehydration and/or rehydration conditions. It had long been known that such activating stimuli caused depletion of neurosecretory material from the posterior pituitary and that rehydration resulted in the replenishment of the peptide stores there. Thus, stimuli that activated the mHNS reliably produced modifications in both SON and neurohypophysial morphology, and these changes were readily detectable when comparisons were made between activated and basal conditions. Electron microscopic studies of the

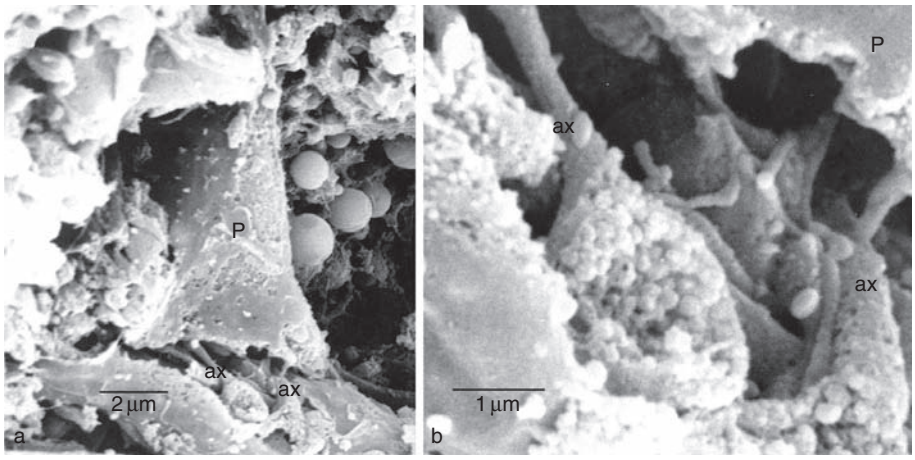

Figure 5 Scanning electron micrographs of neurohypophysis. (a) Pituicyte (P) shown at low magnification, torn open to reveal neurosecretory axons (ax) inside or penetrating its cytoplasm. (b) Higher magnification of the axons shown in (a). The membrane has been partially torn off in the preparation, revealing the densely packed neurosecretory granules.

mHNS comparing basal and dehydrated conditions revealed that, in addition to somatic size increases, rather profound morphological changes took place in cell–cell juxtapositions (involving both neurons and astrocytes) of the SON.

Basal Conditions

Animals in a temperature- and humidity- controlled environment, with free access to food and water and not in any special reproductive state (e.g., pregnant or lactating), are generally considered to be in a basal state with regard to the functions of the mHNS. Under basal conditions, neuronal excitability and firing rates are low.

Morphology: SON Despite being densely packed, the neurons of the SON under basal conditions are usually separated from their closest neighbors by astrocytic processes, as seen at low magnification in Figure 3(b). A deeper appreciation for the glial–neuronal juxtapositions is gained at the ultrastructural level. Figure 6 shows the typical relationships observed between SON astrocytes and their associated neuronal somata (Figure 6(a)) and dendrites (Figure 6(b)) under basal conditions. One or more fine astrocytic processes insinuate themselves between adjacent neuronal somata, also enclosing the terminal boutons making axosomatic synapses (not shown). It is known that these astrocytic processes, as elsewhere in the brain, are vested with the responsibility of (1) taking up extracellular potassium concentration (K_o^+), which rises when the neurons increase their firing of action potentials; (2) taking up synaptically released transmitters, such as glutamate or γ-aminobutyric acid (GABA); and (3) control of the extracellular space

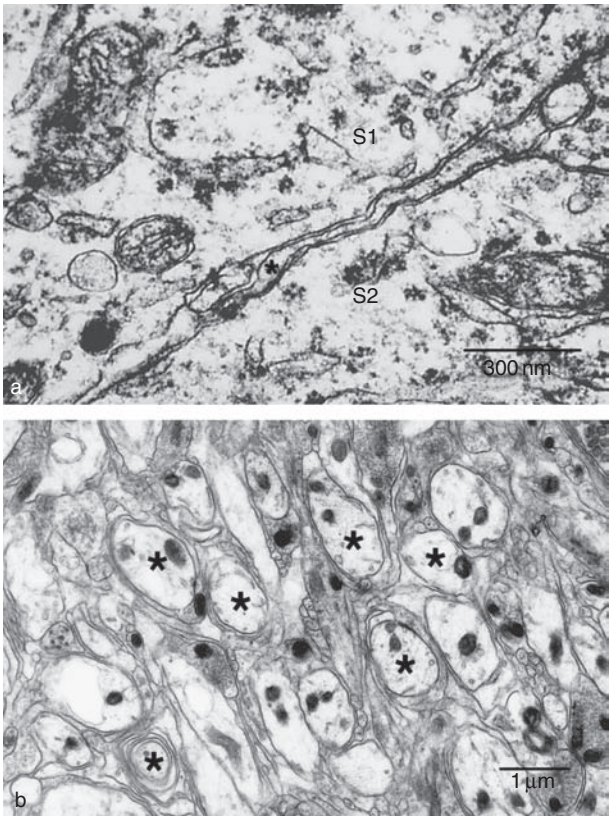

Figure 6 Typical supraoptic nuclei (SON) ultrastructure under basal conditions. (a) Two neurosecretory somata (S1 and S2) are separated by a fine astrocytic process (*) and extracellular space (lower left). (b) Dendrites (*) tend to be wrapped by astrocytic processes and are generally not in close apposition with neighboring dendrites.

volume. Overwhelmingly, the type of synapse seen under basal conditions is the conventional 'single' synapse, in which one terminal bouton forms a synapse with only one postsynaptic element, in this case a

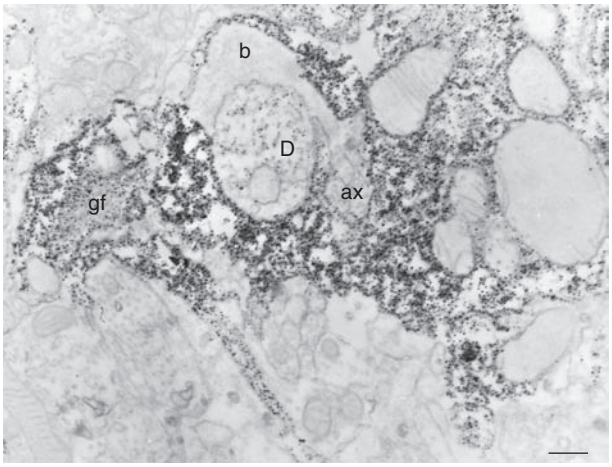

Figure 7 Supraoptic nuclei (SON) astrocyte surrounding a dendrite (D) and part of an axon (ax) whose synaptic bouton (b) is making contact with the dendrite. Glial filaments (gf) can be seen in this astrocyte, which is immunostained for taurine (basal conditions). Scale bar = 290 nm.

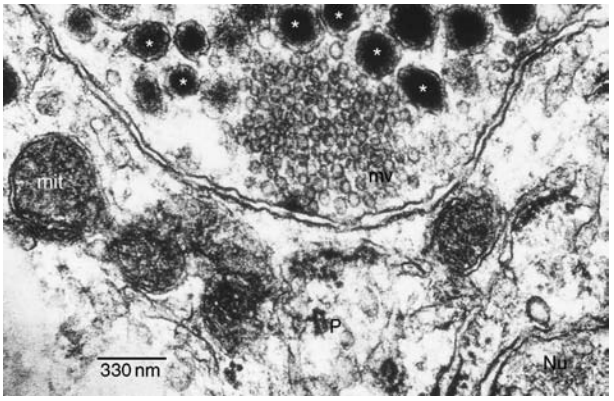

Figure 8 Neurosecretory axon terminal containing dense core vesicles (*) and microvesicles (mv) making a synaptoid contact with a pituicyte (P). Note that the microvesicles form an active zone similar to a synapse and that the dense core vesicles are away from this zone. There is no membrane specialization of the pituicyte side of this contact zone. Mit: mitochondrion; Nu: nucleus.

soma. This has been determined quantitatively by serial section analysis at the ultrastructural level.

SON dendrites, too, tend to be separated from their immediate neighbors by astrocytic processes (Figure 6(b)). Axo-dendritic synapses also tend to be enclosed by astrocytes (see Figure 7), and these synapses are predominantly of the single type. Although these interposed astrocytic processes limit dendritic contact under based conditions, a low, but detectable, level of dendro-dendritic gap junctional communication appears to exist at this time. SON gap junctions have been inferred largely from dye and/or tracer coupling experiments. In such studies the fluorescent dye Lucifer yellow or the tracer neurobiotin or both together have been injected into single SON neurons in slices. Transfer of the injectant to other cells is taken as evidence of gap junctions. The coupling observed is between the dendrites in relative close proximity to one another and is homotypic with regard to cell type; that is, there is no cross-coupling between OT and VP neurons. Finally, as illustrated in Figure 3(a), there is an overall vertical orientation of the astrocytic processes filling the vgl under basal conditions.

Chemistry: SON Although OT and VP neurons are indistinguishable solely on the basis of morphology, they have been shown to differ in their expression of proteins other than OT and VP. Two of these are the calcium-binding proteins calbindin-D_{28k} and calretinin, expression of which appears to be related to the patterns of firing displayed by the two types of neurons and to the changes that occur when the mHNS is activated. Under basal conditions, OT neurons more or less uniformly express relatively large amounts of calbindin-D_{28k} and lesser amounts of calretinin. Calbindin-D_{28k} is expressed in either relatively small amounts or is unexpressed in individual VP neurons, and calretinin is virtually absent. SON astrocytes synthesize and release the GABA-like amino acid taurine (Figure 7) under basal conditions, that is, low extracellular fluid osmotic pressure, contributing to tonic inhibition of VP neurons at the hypothalamic level, via glycine receptors.

Basal Conditions: Neurohypophysis

Morphology Pituicytes, the resident astrocytes of the posterior pituitary, occupy approximately 30% of the total neurohypophysial volume and an even higher percentage (~70%) of the perivascular basal lamina under basal physiological conditions. To a large extent, the neurosecretory axons and axonal endings are surrounded or completely engulfed by pituicytes at this time (see Figures 4 and 5). Moreover, each axon or axonal ending tends to be completely filled with dense core vesicles. Specialized contacts exist between some of the neurosecretory axons and pituicytes. These 'synaptoid' contacts resemble synapses; however, unlike synapses, they have only presynaptoid specializations, with no postsynaptoid densities and no obvious extracellular matrix material in the synaptoid cleft (Figure 8). Therefore, they are called synaptoid contacts and are relatively small in number under basal conditions. Synaptoid contacts are thought to be conduits for glial–neuronal signaling.

Chemistry As would be expected, the distribution of calbindin-D_{28k} and calretinin in the neurohypophysial processes is similar to what exists in the

SON neurons where the major protein synthesis takes place, with the OT axons expressing both proteins and the VP neurons expressing little, if any, of the former and none of the latter. Like the SON astrocytes, the pituicytes are rich in and release taurine under basal conditions. Thus, taurine, via glycine receptors, appears to exert a controlling influence at both the somatodendritic and axon terminal portions of the mHNS.

Plastic Changes with Activation: The Stimuli

Dehydration stimuli are not uniquely capable of activating the mHNS, such as by elevating neuronal firing rates and excitability or increasing synthesis and release of OT and VP. Dehydration is a popular experimental tool because it is convenient to use, its onset can be relatively accurately estimated, and it is powerful in that it produces effects that are both reliable and robust. Furthermore, acute versus chronic effects of dehydration can be readily assessed. Hemorrhage and other stressors, such as physical restraint, also activate the system. These stimuli, however, are not entirely physiological, a criticism that has also been leveled at some methods of producing dehydration. Events that occur during the natural course of pregnancy, parturition, and lactation and suckling of the young also activate the mHNS and are considered unqualifiedly physiological. One or another of these conditions, in particular lactation and suckling, has often been used in studies of morphological plasticity and remodeling of the mHNS. The effects produced by this stimulus complex are every bit as robust as those of dehydration, if not more so. Robust effects are especially observable in the OT cell portion of the mHNS. Furthermore, weaning of the young, not unlike rehydration after dehydration, has been effective in reversing the changes wrought by pregnancy, parturition, and lactation. Except for hemorrhage, which has not been tested, activation of the mHNS by any of these methods produces quite similar morphological changes.

Plastic Changes with Activation: SON

Morphology As measured from the animal's last voluntary drink, as little as 12 h (with a trend already established by 4 h) of water deprivation physiologically activates SON neurons, producing observable changes in endoplasmic reticulum and other subcellular organelles. Presumably, this detection of increased extracellular fluid concentration and/or decreased volume by mHNS cells (neurons and glia), and by other osmosensitive neurons that are synaptically connected to SON cells, also leads to the enhanced neuronal firing that results in OT and VP release. If intake of exogenous water does not correct the extracellular fluid imbalance, the SON undergoes a series of functional changes and rather remarkable morphological remodeling. First, even slight increases in extracellular fluid osmotic pressure curtail release of taurine in most cells tested. Thus, that source of tonic inhibition by the local astrocytes is eliminated. Increased action potential production due to this disinhibition and to increased synaptic drive from amino acid and catecholaminergic inputs apparently acts on both the SON neurons and their associated astrocytes via combinations of α-amino-3-hyroxy-5-methyl-4-isoxazole-propionic acid, *N*-methyl-D-aspartate, GABA, and α- and/or β-adrenergic receptors. Collectively, the results of the extracellular microenvironmental changes and receptor activation produce retraction of the fine glial processes separating adjacent SON dendrites and somata from their respective neighbors. Astrocyte membrane filling the vgl also undergoes reorganization from its usual predominantly vertical orientation to a more horizontal one. One effect of this reorganization is to reduce vgl thickness, thereby reducing the distance between the cerebrospinal fluid and the SON neuronal compartments.

In the SZ, astrocytic process retraction results in increased soma-somatic appositions (**Figure 9(a)**); that is, somatic membrane that was formerly apposed (across a $\sim$10 nm extracellular space) to the membrane of an astrocytic process now faces another neuronal membrane. Axo-somatic synapses that were formerly covered by astrocytes that would take up excess transmitter and K_o^+ are now also apposed to neuronal membrane. The functional implications of these changes are exactly what would be expected from the increases observed in neuronal excitation under these conditions. Less of the K_o^+ generated by the now more frequently occurring action potentials is taken up due to the relative paucity of astrocytic processes separating somata, leading to increased depolarization. Amino acid transmitters can now more readily spill over and remain longer in the vicinity after glial retraction. Consequences too complex to discuss in depth here accrue from reduced transmitter uptake due to retracted glial processes. These include enhanced potential activation of GABA receptors (usually more frequent on the soma); of glutamate receptors, both ionotropic and metabotropic; and reduced breakdown of adenosine triphosphate (ATP) to adenosine, since ecto-5′-nucleotidase, an important enzyme in ATP metabolism, is of astrocytic origin. This latter reduction in glial participation means that the generally excitatory effects of ATP on SON neurons last longer because the adenosine-induced inhibitory effects, via adenosine A_1 receptors, are delayed.

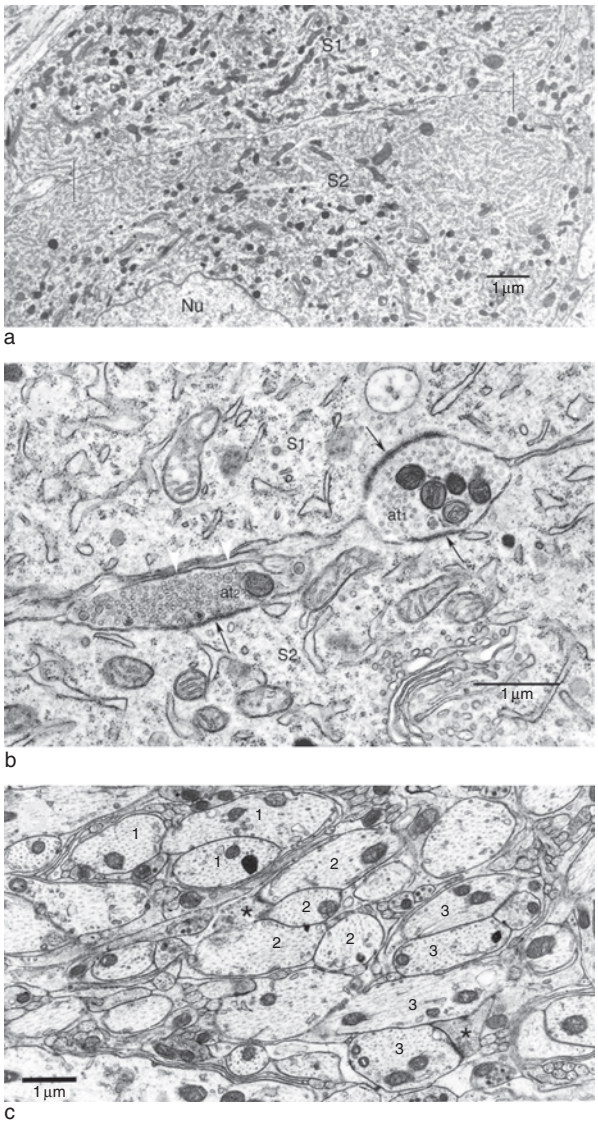

Figure 9 Ultrastructural features of the supraoptic nuclei (SON) under activated conditions. (a) Two somata (S1 and S2) in close apposition (i.e., without an intervening astrocytic process). This apposition (between vertical lines) is more than 10 µm in length. Nu: nucleus. (b) Axosomatic multiple synapse made by axon terminal 1 (at1) with the two somata (arrows). A conventional single synapse is made by at2 with S2 (arrow). A fine glial process separates at2 from S1 (white arrowheads). (c) Dendro-dendritic close appositions (bundling) and multiple synapses. Three dendritic bundles (i.e., two or more dendrites without intervening glial processes) are identified with the same numbers (1, 2, or 3). Two multiple synapses (*) are evident at bundles 2 and 3.

One of the more remarkable occurrences in response to physiological activation of the mHNS is the appearance of new synapses. Unlike the conventional ones seen under basal conditions, in the new synapses one synaptic bouton contacts two or more (up to seven have been documented) postsynaptic neurons (Figure 9(b)). These so-called multiple or shared synapses are thought to be formed from the conventional single synapses shortly after astrocytic process withdrawal exposes the synaptic terminal membrane, which had previously been covered by glia, to another neuronal membrane. Several pieces of evidence support this hypothesis. First, the appearance of multiple synapses is too rapid (minutes to a few hours) to be due to axonal sprouting. Second, no growth cones have ever been observed during this time. Third, and perhaps most convincing, is the observation that the increase in multiple synapses is accompanied by a decrease in the percentage of single synapses. Since glutamatergic and GABAergic multiple synapses both increase with prolonged activation of the system, current thinking is that one effect of increased multiple synapses is to enhance the coordination of incoming signals to influence excitability and, under certain circumstances, even synchrony. It well may be that this increase in the number of synapses also compensates for the loss of overall synaptic efficacy that would occur as the SON somata enlarge under chronic stimulation.

Remodeling similar to that described for the SZ occurs in the DZ. Astrocytic process retraction here allows dendro-dendritic close appositions, so several neighboring dendrites are more likely to be apposed to each other than to glial processes (Figure 9(c)). This has been termed 'dendritic bundling.' DZ multiple synapses also increase in frequency as the single synapses are exposed to other neuronal membrane. Often these new synapses appear to make contact with two or more dendritic bundles. Demonstrations that SON neurons can release peptides, and perhaps glutamate, from their dendrites as well as from their axon terminals add considerable complexity to the potential effects of astrocytic process withdrawal. Functional implications of this are still under investigation, but the effects on neuronal activity of dendritic VP versus OT release differ in important ways from one another. Glial process withdrawal with activation of the system is coincident with an increased frequency of neuronal dye coupling, unique to the dendritic portion of these cells. It is likely that the multiple synapses in the DZ coordinate mostly excitatory input. Increased junctional communication would tend to aid in this coordination by bringing the membrane potentials of local groups of neurons into register. Thus, the SON could be said to be 'rewired' by physiological stimulation of any significant duration.

Chemistry The plastic changes that accompany dehydration are cotemporal with downregulation of tenascin-C, an extracellular matrix molecule expressed by astrocytes and identified in the stabilization of synapses. Activation by dehydration causes an upregulation of calretinin protein expression in the OT

neurons, with no change in expression of calbindin-D_{28k} in either OT or VP neurons. It may be of some functional importance that dehydration-activated OT neurons fire many more action potentials per unit time than do VP cells, which fire phasic bursts interrupted by significant silent periods. Since the calretinin tends to be membrane-bound, whereas calbindin-D_{28k} is cytosolic, upregulation of calretinin may aid the OT cell in dealing with increased calcium loads via voltage-gated calcium channels during times of heightened electrical activity and secretion. Molecular studies have determined that many genes, in addition to those for OT and VP, are upregulated during dehydration, but that a nearly equal number are downregulated. The formidable task of sorting these out is ongoing.

Plastic Changes with Activation: Neurohypophysis

Morphology Alterations occur in the neurohypophysis that parallel those in the SON. Under the influence of catecholaminergic input from both brain and blood, acting on β-adrenergic receptors, pituicytes become process-bearing and release the axons that they engulf under basal conditions (Figure 10). Pituicytes also withdraw from their prominent occupation of the basal lamina separating the neural compartment from the perivascular space surrounding the fenestrated capillaries. This allows direct basal laminar contact by more neurosecretory terminals. For example, in the hour or so that it takes to deliver a litter of pups, the percentage of the basal lamina occupied by neural versus astrocyte membrane can rise from 30% to 70%. Another consequence of the pituicyte shape changes is an expansion of the extracellular space within the neurohypophysis (cf. Figure 10). Dehydration stimuli also have the capacity to increase the number of neurohypophysial synaptoid contacts. Except for the latter change, which is still not clearly understood, these modifications all work to facilitate the entry of secreted peptide into the blood.

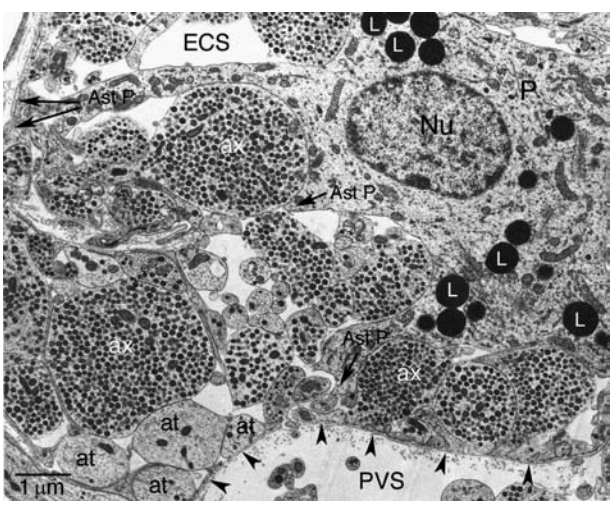

Figure 10 Under activated conditions (immediately following parturition), pituicytes (P) are process bearing (Ast P) and engulf no neurosecretory axons (ax). Pituicytes occupy relatively much less than previously of the basal lamina (arrowheads) along the perivascular space (PVS). This is now occupied by neuronal membrane, such as that of the axon terminals (at) that have released their contents and appear devoid of dense core vesicles. The extracellular space (ECS) is also expanded by activation. L, lipid bodies; Nu, pituicyte nucleus.

Chemistry As with the morphology, the activation-associated changes in chemistry parallel those in the SON. Calretinin is selectively upregulated in OT axon terminals, where significant peptide release occurs. Meanwhile, tenascin-C in the extracellular space is downregulated.

Perhaps it bears reiterating here that all these plastic changes that occur in response to activation of the mHNS are reversible. That is, with rehydration or weaning of the pups, or with relief from a prior restraint stress, the mHNS returns to the morphology and the chemistry of the status quo ante. It is also worth mentioning that reversal of the changes occurs over a longer time than the induced remodeling.

Model Peptidergic Neurons: Older versus Current Ideas

Rather than endocrine cells that have taken on a few neuronal characteristics, peptidergic neurons are primarily neurons, albeit ones that have, in the words of Berta Scharrer, "developed secretory activity to the point where it takes precedence over all the other functions of the cell." Secretory activity taking precedence, it has since been learned, does not mean that these cells do not behave like neurons. Instead they have added a dimension to their complexity. Those seeking to emphasize one or another aspect of their genomic, biosynthetic, or neurochemical nature have often missed or ignored the complexity of peptidergic neurons. Perhaps a more complete description of the model neurons from the mHNS will help correct some long-surviving misconceptions (see Figure 11).

Outdated Neurosecretory Models

Peptidergic neurons are often diagrammed, even in recent publications, as a smooth bag of synthetic machinery with a stubby terminal end-foot landing directly on a blood vessel. Such portrayals could easily mislead one into concluding that peptidergic neurons are entirely different from neurons in general. All peptides secreted from those conceptualized neurons are pictured as being liberated from the terminal

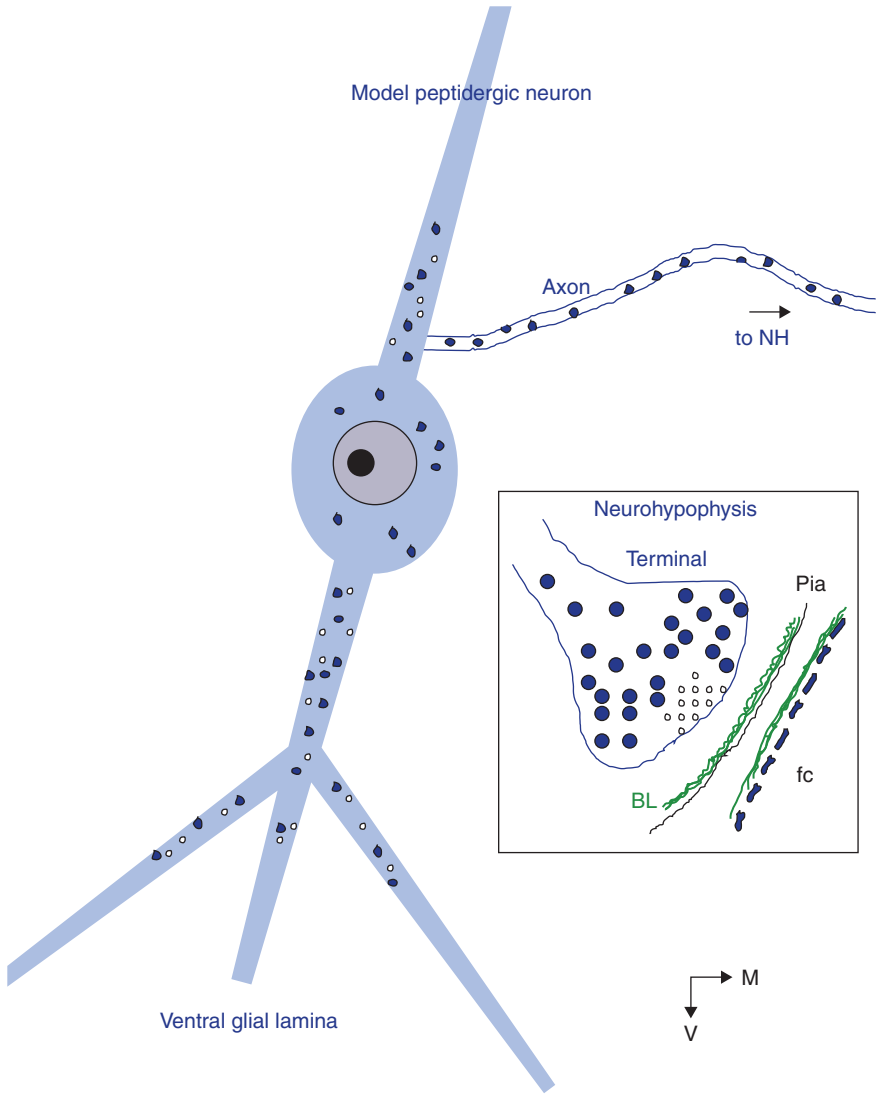

Figure 11 Updated model of a peptidergic neuron based on accumulated knowledge of the hypothalamo-neurohypophysial system. Peptidergic neurons have dendrites that play important functional roles. Peptides can be selectively released from sites other than the axon terminals, such as dendrites and somata. In addition to dense core vesicles (dark blue), the dendrites contain and may release neurotransmitters such as glutamate from microvesicles (clear circles). Both are shown here. NH, neurohypophysis. Arrows at bottom right indicate medial (M) and ventral (V). The inset depicts a neurosecretory axon terminal that also contains dense core vesicles and microvesicles in the arrangement found in actual terminals, that is, with the microvesicles forming an active zone and with the dense core vesicles peripherally located. This terminal is juxtaposed to, not a blood vessel, but the basal lamina (BL, in green) separating the neural compartment from the pia mater (Pia), which is itself separated from the fenestrated capillary (fc) by perivascular space and another BL. See text for comparison with earlier models.

end-foot, as though into a synaptic cleft. These depictions have also implied that all the observed intracellular vesicles in these neurons are somehow related to the neurosecretory process (e.g., either secretory vesicles or recycled membrane), an unwarranted inference. Conceptualizing the neuron's cell body as a smooth bag should be taken, not so much as an indicator of ignorance that these neurons possess dendrites, but instead as indicating the belief that the dendrites are an unimportant feature of neurosecretory cells, a belief in need of correction.

An Updated Conceptualization

Figure 11 is a semidiagrammatic presentation of a model neuron from the SON, complete with dendrites that contain dense core vesicles. This feature is crucial since we know that there is dendritic release of peptide during certain types of activation. Microvesicles are also depicted in the dendrites, as recent evidence shows that SON neurons express vesicular glutamate transporters in dendrites and terminals. It is of some importance that the axon arises from a primary

dendrite, since this is consistent with the cell's ability to selectively release peptides from either the dendrites or the axon or both. In the inset, the terminal is shown to contain both dense core vesicles and clear microvesicles. These are arranged as they are in most neurons that release both peptides and classical transmitters, with the microvesicles clustered at an 'active zone' and the neurosecretory vesicles peripheral to that zone (cf. Figure 8). The terminal membrane abuts the basal lamina separating the neural compartment from the pia mater, which in turn is separated from the fenestrated capillary by perivascular space and another basal lamina.

Concluding Remarks

Our understanding of the peptidergic neuron and its relationship to the structures with which it interacts, both neural and nonneural, has changed in rather important ways in recent years. Investigators better understand the complexities of this neuron type and have a fuller appreciation for its neuronal characteristics, in addition to its secretory role. As well researched as is the remarkable morphological plasticity of the mHNS, mechanisms underlying this rewiring of a brain area in response to physiological stimulation are still hot topics of study. Many other brain areas appear to share at least some of these plastic capabilities, making future work in this field more exciting.

See also: Glial Plasticity and Neuroendocrine Regulation; Neuroendocrinology; Neurohypophyseal System; Neurosecretion (Regulated Exocytosis in Neuroendocrine Cells); Oxytocin (Peripheral/Central Actions and their Regulation); Peptidergic Receptors; Pituitary Gland (Cell Types, Mediators, Development); Vasopressin/Oxytocin and Receptors.

Further Reading

Burbach JPH, Luckman SM, Murphy D, and Gainer H (2001) Gene regulation in the magnocellular hypothalamo-neurohypophysial system. *Physiological Reviews* 81: 1197–1267.

Hatton GI (1990) Emerging concepts of structure-function dynamics in adult brain: The hypothalamo-neurohypophysial system. *Progress in Neurobiology* 34: 437–504.

Hatton GI (1997) Function-related plasticity in hypothalamus. *Annual Review of Neuroscience* 20: 375–397.

Hatton GI (1999) Astroglial modulation of neurotransmitter/peptide release from the neurohypophysis: Present status. *Journal of Chemical Neuroanatomy* 16: 203–222.

Hatton GI (2004) Dynamic neuronal-glial interactions: An overview 20 years later. *Peptides* 25: 403–411.

Hatton GI (2004) Morphological plasticity of astroglial/neuronal interactions: Functional implications. In: Hatton GI and Parpura V (eds.) *Glial Neuronal Signaling*, pp. 99–124. Amsterdam: Kluwer.

Hatton GI and Parpura V (eds.) (2004) *Glial Neuronal Signaling*. Amsterdam: Kluwer.

Miyata S and Hatton GI (2002) Activity-related, dynamic neuron-glial interactions in the hypothalamo-neurohypophysial system. *Microscopy Research and Technique* 56: 143–157.

Poulain DA, Oliet SHR, and Theodosis DT (eds.) (2002) *Vasopressin and Oxytocin: From Genes to Clinical Applications. Progress in Brain Research*, vol. 139. Amsterdam: Elsevier.

Theodosis DT (2002) Oxytocin-secreting neurons: A physiological model of morphological neuronal and glial plasticity in the adult hypothalamus. *Frontiers in Neuroendocrinology* 23: 101–135.

Mammalian Neuropeptide Families

J P H Burbach, University Medical Center Utrecht, Utrecht, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Discovery of Neuropeptides

As early as the 1890s, it was noted that extracts of tissues and organs contained biologically active substances which could be purified in later years. Isolation has been guided by bioassays from the beginning up to today; first *in vivo* effects or organ preparations were used, but now sophisticated high-throughput cell systems with engineered receptors and proteomics lead to the discovery of novel neuropeptides.

In the 1950s, it became technically possible to determine the amino acid sequence of peptides, leading to the elucidation of the chemical structure and chemical synthesis of neuropeptides, such as oxytocin (OXT) and vasopressin (AVP) by Du Vignaud. Such peptides were mostly studied as hormones, but in the 1960s it was recognized that hormones could affect brain functions as monitored by behavioral assay. The structural requirements for central activity were often distinct from the endocrine effects, pointing to a multiplicity of receptors. De Wied called these as neuroactive fragments or neuropeptides and developed the neuropeptide concept. The neuropeptide field fully bloomed after the discovery of the enkephalins by Terenius and Kosterlitz in 1975.

Toward a Definition of Neuropeptides

With the growing number of peptides and small proteins affecting neural cells in some way or the other, a definition pinpointing the essence of a neuropeptide has lagged behind. The hallmarks of neuropeptides are (1) gene expression and biosynthesis by neural cells, (2) regulated release, and (3) the ability to modulate or mediate neural functioning. Each of these elements alone may not be exclusive to what we generally call neuropeptides.

Synthesis by Neural Cell

Neuropeptide gene expression and biosynthesis are associated *a priori* with neurons. However, glial cells, in particular microglia, synthesize and secrete numerous soluble factors, such as chemokines, cytokines, neurotrophic factors, hormones, and complement proteins, which have biological activities in the nervous system. This raises the issue of whether these can be considered neuropeptides.

Regulated Release

Controlled secretion is a basis for the regulation of chemical communication. Neurons expressing classical neuropeptides use the regulated secretory pathway, allowing the storage of synthesized peptides in large dense-cored vesicles and controlled release after a stimulus. The neuropeptide precursor contains the structural prerequisites to be sorted into the regulated secretory pathway and to be processed into active peptides. A signal sequence is key to entry into the regulated secretory pathway. Thus, neuropeptide precursors contain distinct structural characteristics. However, other proteins not classically considered neuropeptides can also be funneled through the regulated secretory pathway due to a signal sequence and can undergo processing. Still other proteineous factors are secreted through the constitute pathway, but achieve a regulated mode by strictly controlled gene transcription.

Architecture of Neuropeptide Precursors

Neuropeptide precursors, or preproneuropeptides, are the initial gene products and are structured to undergo appropriate biosynthesis in the regulated secretory pathway of a neuron. Both proteolytic and derivatizing posttranslational modifications are involved in biosynthesis. The salient structural characteristics of these modifications, and hence of neuropeptide precursors, are:

1. *The signal peptide*. This 20–25-amino-acid long N-terminal portion is required for the entry of the newly synthesized gene product into the lumen of the endoplasmatic reticulum (ER). It is cleaved off during passage of the ER membrane, leaving the proneuropeptide in the ER-Golgi for further sorting into the regulated secretory pathway.
2. *The basic motifs*. Pairs of the basic amino acids lysine (Lys, K) and arginine (Arg, R), or more rarely a single Arg, in the appropriate structural environment of the precursor serve as recognition sites and substrates of prohormone convertases (PCs). These PCs belong to a small family of subtilisin proteases with cell-specific expression. Cleavage leads to the generation of specific peptides and to cell-specific differences in PCs to different sets of peptides from the same precursor.
3. *C-terminal amidation*. PC-generated peptides can be subject to further modification by peptidylaminotransferase (PAM). This enzyme uses a C-terminal glycine (Gly, G) as the amide donor for the preceding amino acid. This results in a peptide with a C-terminal amide. This amide

occurs in many neuropeptides and is often essential for biological activity. This structural characteristic was elegantly exploited by Mutt to isolated novel neuropeptides and peptide hormones on the bases of a chemical assay for C-terminal amides, such as neuropeptide Y, peptide YY, and peptide HI (all named after the C-terminal-amidated amino acid).

Modulation of Neural Functions

Neuropeptides have neural bioactivities that are observable at almost all levels of neural functioning. Effects of neuropeptides have been established at the genetic, biochemical, cellular, organism, and behavioral levels. In this respect, the biological activities of many neuropeptides have been indicated to be pleiotropic. Such effects are not always distinguishable from that of cytokines, or growth factors. Typical for the cellular action of neuropeptides are the relatively slow response and action compared to excitatory amino acids and catecholamines. The relatively slow response of neuropeptide secretion is due to storage in dense-cored vesicles that are not docked at the cellular release site (e.g., the synaps or synaptic button), requiring recruitment first. This happens at elevated Ca^{2+} levels, which require enhanced or repeated stimulation unlike the fast transmitters, which are ready for release.

The relatively slow action is further due to the type of receptor that almost all neuropeptides act on – G-protein-coupled receptors (GPCRs). These receptors trigger an intracellular cascade of molecular events that results in cellular responses. The time span over which such a response is in effect (seconds or longer) is considerably longer than that of neurotransmitters that modulate ion fluxes directly through action on ion channels (milliseconds).

Definition of Neuropeptides

Based on these considerations, here is a conservative definition of a neuropeptide: a small proteineous substance produced by the processing of a precursor, released in a regulated fashion by neurons and acting on neural substrates (e.g., neurons), glial cells, or neuronal target cells (e.g., a gland or muscle). Approximately 90 genes encoding for such neuropeptides have been recognized (Table 1). A more liberal definition includes all neural secretory peptides and proteins of the brain, defined as proteineous substance produced and released by neural cells and acting on any neural substrate or target. This includes peptides that are known best as peptide hormones of the endocrine system, chemokines of the immune system, and growth factors if synthesized by neurons. This definition still excludes hormones of nonneuronal peripheral origin which signal to the brain as being neuropeptides.

Table 1 lists the peptides that have been recognized as neuropeptides according to the conservative definition. To appreciate peptide hormones, chemokines, and growth factors as candidate neuropeptides, these classes of proteineous signaling molecules are briefly addressed next.

Candidate Neuropeptides

Peptide Hormones

De Wied discovered neuropeptides by studying the behavioral effects of the peptide hormones adrenocorticotropin (ACTH) and AVP. A hormone is a substance that is secreted by one organ and acts on another after transport by the bloodstream. In this way, cells at different locations in the organism communicate in an endocrine fashion. Many peptide hormones are also synthesized by neurons and indisputably belong to the classical neuropeptides according the conservative definition, like those derived from proopiomelanocortin (POMC). Sometimes the forms in which related peptides act are different between the nervous system and the endocrine system. For example, ACTH and β-lipotropin are the prime peptide hormones derived from POMC in the periphery, whereas α-melanocyte-stimulating hormone (α-MSH) and β-endorphin are the major POMC neuropeptides in the brain. This is due to differences in processing between adrenocorticotrophes of the anterior pituitary gland, secreting ACTH for peripheral action, and the hypothalamic POMC neurons, secreting α-MSH and β-endorphin from neuronal processes. The different sets of biologically active peptides are accommodated by differences between the peripheral and central receptors. In addition, differential expression of two paralogous genes and differential splicing of one gene transcript can result in a nervous system-specific population of neuropeptides distinct from the related peripheral peptide hormones. Examples are found in the calcitonin/calcitonin gene-related peptide (CGRP) family of genes and peptides.

Some classical peptide hormones, such as prolactin, appear to be expressed in a restricted manner in the nervous system, together with its receptor. Other peptide hormones have not been found to be expressed in the brain or have not been addressed for this issue extensively, for instance growth hormone. The latter cannot be considered neuropeptides and have not been included in Table 1. However, with ongoing examination of brain-expressed genes, we may find that several of these may be synthesized in the nervous

Table 1 Portal to neuropeptide gene families hyperlinked to bioinformatics data sources

Gene (abbreviation)[a]	*Human chromosomal localization*[a]	*Precursor*[b]	*Active peptide(s)*[c]
Opioid gene family			
Proenkephalin gene (PENK)	chr8 (q12.1)	Preproenkephalin	Leu-enkephalin, Met-enkephalin, amidorphin, adrenorphin, peptide B, peptide E, peptide F, BAM22P
Proopiomelanocortin gene (POMC)	chr2 (q12.2)	POMC	α-Melanocyte-stimulating hormone (α-MSH), γ-melanocyte-stimulating hormone (γ-MSH), β-melanocyte-stimulating hormone (β-MSH), adrenocorticotropic hormone (ACTH), β-endorphin, α-endorphin, γ-endorphin, β-lipoprotein (β-LPH), γ-lipoprotein (γ-LPH), corticotropin-like intermediate peptide (CLIP)
Prodynorphin gene (PDYN)	chr20 (p13)	Preprodynorphin	Dynorphin A, dynorphin B, α-neo-endorphin, β-neo-endorphin, dynorphin-32, leu-morphin
Orphanin gene, pre-pronociceptin gene (PNOC)	chr8 (p21.1)	Prepronociceptin, preproorphanin	Nociceptin (orphanin FQ), neuropeptide 1, neuropeptide 2
Vasopressin/oxytocin gene family			
Vasopressin gene (AVP)	chr20 (p13)	Preprovasopressin-neurophysin Ii	Vasopressin (VP), neurophysin II (NP II), C-terminal glycopeptide CPP
Oxytocin gene (OXT)	chr20 (p13)	Preprooxytocin-neurophysin I	Oxytocin (OT), neurophysin I (NP 1)

Continued

Table 1 Continued

Gene (abbreviation)[a]	*Human chromosomal localization*[a]	*Precursor*[b]	*Active peptide(s)*[c]
CCK/gastrin gene family			
Gastrin gene (GAST)	chr17 (q21.2)	Preprogastrin	Gastrin-34, gastrin-17, gastrin-4
Cholecystokinin gene (CCK)	chr3 (p22.1)	Prepro-CCK	CCK-8, CCK-33, CCK-58
Somastostatin gene family			
Somastostatin gene (SST)	chr3 (q27.3)	Preprosomatostatin	SS-12, SS-14, SS-28, antrin
Cortistatin gene (CST)	chr1 (p36.22)	Preprocortistatin	Cortistatin-29, cortistatin-17
F- and Y-amide gene family			
Gonadotropin inhibitory hormone gene, RF-amide related peptide gene (RFRP)	chr7 (p15.2)	Prepro-NPRF	QRF-amide (neuropeptide RF-amide, GnIH, p518, RF-related peptide-2), RF-related peptide-1, RF-related peptide-3
Neuropeptide FF gene (NPFF),	chr12 (q13.13)	Prepro-NPFF	Neuropeptide FF, neuropeptide AF, neuropeptide SF
Neuropeptide Y gene (NPY)	chr7 (p15.3)	Prepro-NPY	NPY, C-flanking peptide CPON
Pancreatic polypeptide gene (PPY)	chr17 (q21.31)	Prepro-PPY	PPY
Peptide YY gene (PYY)	chr17 (q21.31)	Prepro-PYY	PYY, PYY-(3–36)
Prolactin-releasing peptide (PRLH)	chr2 (q37.3)	Prepro-PrRP	PrRP-31, PrRP-20
Calcitonin gene family			
Calcitonin I gene (CALCA)	chr11 (p15.2)	Prepro-CALC Prepro-CGRP-alpha	Calcitonin, katacalcin Calcitonin gene related peptide I (α-CGRP)
Calcitonin II gene (CALCB)	chr11 (p15.2)	Prepro-CGRP-beta	Calcitonin gene related peptide II (β-CGRP)
Islet amyloid polypeptide gene (IAPP), Amylin gene	chr12 (p12.1)	Prepro-IAPP	IAPP (amylin, amyloid polypeptide)
Adrenomedullin gene (ADM)	chr11 (p15.4)	Preproadrenomedullin	Adrenomedullin, AM, PAMP

Adrenomedullin-2 gene (ADM2)	chr22 (p13.33)	P13 P12 11.2 11.21 12.1 22q12.3 13.1 13.2 13.31	Prepro ardenomedullin-2	Adrenomedullin-2, intermedin-long (IMDL), intermedin-short (IMDS)
Natriuretic factor gene family				
Atrial natriuretic factor gene (NPPA)	chr1 (p36.22)	31.1 1q12 41 43	Prepro-ANP	Atrial natriuretic factor (natriuretic peptide A, ANF, ANP, natriodilatine, cardiodilatine-related peptide)
Brain natriuretic factor gene (NPPB)	chr1 (p36.22)	31.1 1q12 41 43	Prepro-BNP	Brain natriuretic factor (natriuretic peptide B, BNF, BNP)
Natriuretic peptide precursor C gene (NPPC)	chr2 (q37.1)	12	Prepro-CNP	C-type natriuretic peptide (CNP-23), CNP-29, CNP-53
Bombesin-like peptide gene family				
Gastrin releasing peptide gene (GRP)	chr18 (q21.32)	q11.2 q12.1 12.2 q12.3 21.1 21.2 22.1 q23	Prepro-GRP-1	GRP-27, GRP-14, GRP-10 (neuromedin C)
			Prepro-GRP-2	GRP-27, GRP-14, GRP-10 (neuromedin C)
			Prepro-GRP-3	GRP-27, GRP-14, GRP-10 (neuromedin C)
Neuromedin B gene (NMB)	chr15 (q25.2–q25.3)	13 12 q14 q23	Preproneuromedin B1	Neuromedin B (Ranatensin-like peptide, RLP)
			Preproneuromedin B2	Neuromedin B (Ranatensin-like peptide, RLP)
Endothelin gene family				
Endothelin 1 gene (EDN1)	chr6 (q24.1)	12 13 q21 27	Preproendothelin 1 (PPET1)	Endothelin 1 (ET-1)
Endothelin 2 gene (EDN2)	chr1 (p34.2)	31.1 1q12 41 43	Preproendothelin 2 (PPET2)2	Endothelin 2 (ET-2)
Endothelin 3 gene (EDN3)	chr20 (q13.32)	20p13 12.3 p12.1 q12 q13.2	Preproendothelin 3 (PPET3)	Endothelin 3 (ET-3)

Continued

Table 1 Continued

Gene (abbreviation)[a]	*Human chromosomal localization*[a]	*Precursor*[b]	*Active peptide(s)*[c]
Glucagon/secretin gene family			
Glucagon gene (CGC)	chr2 (q24.2)	Preproglucagon	Glicentin; glicentin-related polypeptide (GRPP); oxyntomodulin (OXY) (OXM); glucagon; glucagon-like peptide 1 (GLP-1); glucagon-like peptide 1(7-37) (GLP-1(7-37)); glucagon-like peptide 1(7-36) (GLP-1(7-36)); glucagon-like peptide 2 (GLP-2)]
Secretin gene (SCT)	chr11 (p15.5)	Preprosecretin	Secretin
Vasoactive intestinal peptide gene (VIP)	chr6 (q25.2)	Prepro-VIP-1	VIP, PHM-27/PHI-27, PHV-42
		Prepro-VIP-2	VIP, PHM-27/PHI-27, PHV-42
Pituitary adenylcyclase-activated peptide gene (ADCYAP1)	chr18 (p11.32)	Prepro-PACAP	PACAP-38, PACAP-27, PRP-48
Growth hormone releasing hormone gene (GHRH)	chr20 (q11.23)	Prepro-GHRH	GHRH (somatoliberin, GRF, somatocrinin, somatorelin, sermorelin)
Gastric inhibitory peptide gene (GIP)	chr17 (q21.32)	Prepro-GIP	GIP (gastric inhibitory peptide, glucose-dependent insulinotropic polypeptide)
CRH-related gene family			
Corticotropin-releasing hormone gene (CRH)	chr8 (q13.1)	Prepro-CRH	CRH
Urocortin gene (UCN)	chr2 (p23.3)	Prepro-UNC I	UNC I
Urocortin II gene (UCN2)	chr3 (p21.31)	Prepro-UNC II	UNC II, stresscopin-related peptide
Urocortin III gene (UCN3)	chr10 (p15.1)	Prepro-UNC III	UNC III, stresscopin

Urotensin-II (UTS2)	chr1 (p36.23)	Preprourotensin-2, isoform a	Urotensin-2
		Preprourotensin-2, isofrom b	Urotensin-2
Kinin and tensin gene family			
Preprotachykinin A gene (TAC1)	chr7 (q21.3)	α-PPTA	Substance P, neurokinin A (NKA, substance K, neuromedin L), neuropeptide K, neuropeptide gamma
		β-PPTA	Substance P, neuropeptide K, neurokinin A
		γ-PPTA	Substance P, neurokinin A, neuropeptide γ
		δ-PPTA	Substance P, neuropeptide K, neurokinin A
Preprotachykinin B gene (TAC3)	chr12 (q13.3)	PPTB, isoform 1	Neuromedin K, neurokinin B
		PPTB, isoform 2	Neuromedin K, neurokinin B
Neuromedins			
Neuromedin S gene (NMS)	chr2 (q11.2)	Preproneuromedin S	Neuromedin S
Neuromedin U gene (NMU)	chr4 (q12)	Preproneuromedin U, multiple isoforms	Neuromedin U
Kininogen-1 gene (KNG1)	chr3 (q27.3)	Kininogen-1 precursor, multiple isoforms	Bradykinin, kallidin, LMW-K-kinin, HMW-K-kinin
Angiotensinogen gene (AGT)	chr1 (q42.2)	Angiotensinogen preprotein	Angiotensin I, angiotensin II, angiotensin-(1–7)
Neurotensin gene (NTS)	chr12 (q21.31)	Preproneurotensin	Neurotensin (NT), neuromedin N
Granins			
Chromogranin A gene (CHGA)	chr14 (q32.12)	Chromogranin A precursor	Chromogranin A, beta-granin, vasostatin
Chromogranin B gene (CHGB)	chr20 (p12.3)	Chromogranin B precursor	Chromogranin B (secretogranin I), CCB peptide, GAWK peptide
		Chromogranin B precursor variant	Chromogranin B (secretogranin I)

Continued

Table 1 Continued

Gene (abbreviation)[a]	*Human chromosomal localization*[a]	*Precursor*[b]	*Active peptide(s)*[c]
Secretogranin II gene (SCG2)	chr2 (q36.1)	Secretogranin II precursor , chromogranin C precurosor	Secretogranin II (chromogranin C), EM66, secretoneurin
Secretogranin III gene (SCG3)	chr15 (q21.2)	Secretogranin III precursor	Secretogranin III
Secretory granule neuroendocrine protein 1, 7B2 gene (SGNE1)	chr15 (q13.3)	Secretory granule neuroendocrine precursor	Secretory granule neuroendocrine protein-1 (7B2)
VGF nerve growth factor inducible protein (VGF)	chr7 (q22.1)	VGF-precursor	VGF (NGF-inducible protein, neurosecretory protein), TLPQ-62, TLPQ-21, AQEE-30, LQEQ-19
Motilin family			
Motilin gene (MLN)	chr6 (p21.31)	Prepromotilin isoform 1	Motilin
		Prepromotilin isofrom 2	Motilin, motillin-associated peptide
Ghrelin gene (GHRL)	chr3 (p25.3)	Preproghrelin	Ghrelin, obestatin
Galanin family			
Galanin gene (GAL)	chr11 (q13.2)	Preprogalanin	Galanin, galanin message associated peptide (GMAP)
Galanin-like peptide precursor gene (GALP)	chr19 (q13.42)	Galanin-like peptide precursor	Galanin-like peptide (GALP)
GnRH family			
Gonadotropin-releasing hormone gene (GnRH1)	chr8 (p21.2)	Prepro-GnRH1	GnRH (LHRH, gonadoliberin)
Gonadotropin-releasing hormone gene (GnRH2)	chr20 (p13)	Prepro-GNRH2, isoform a	GnRH2 (LHRH II, gonadoliberin II)
Neuropeptide B/W family			
Neuropeptide B gene (NPB)	chr17 (q25.3)	Preproneuropeptide B, PPL7	Neuropeptide B-23 (peptide L7), neuropeptide B-29

Neuropeptide W gene (NPW)	chr16 (p13.3)	Preproneuropeptide W, PPL8	Neuropeptide W-23 (peptide L8), neuropeptide W-30
Neuropeptides S gene (NPS)	chr10 (q26.2)	Preproneuropeptide S	Neuropeptide S
Neurexophilins			
Neurexophilin-1 (NXPH1)	chr7 (p21.3)	Neurexophilin-1 precursor	
Neurexophilin-2 (NXPH2)	chr2 (q22.1)	Neurexophilin-2 precursor	
Neurexophilin-3 (NXPH3)	chr17 (q21.33)	Neurexophilin-3 precursor	
Neurexophilin-4 (NXPH4)	chr12 (q13.3)	Neurexophilin-4 precursor	
Insulin family			
Insulin gene (INS)	chr11 (p15.5)	Preproinsulin	Insulin
IGF-1 gene (IGF1)	chr12 (q23.2)	IGF-1 precursor	IGF-1 (somatomedin C)
IGF-2 gene (IGF2)	chr11 (p15.5)	IGF-2 precursor	IGF-2 (somatomedin A)
Relaxin-1 gene (RLN1)	chr9 (p24.1)	Preprorelaxin-1	Relaxin-1
Relaxin-2 gene (RLN2)	chr9 (p24.1)	Preprorelaxin-2, isoform 1 Preprorelaxin-2, isoform 2Relaxin-2	Relaxin-2
Relaxin-3 gene (RLN3)	chr19 (p13.12)	Preprorelaxin-3	Relaxin-3
Prokineticins			
Prokineticin-1 gene (PROK1)	chr1 (p13.3)	Prokineticin-1 precursor	Prokineticin-1 (PK1, mambakine, endocrin gland-derived VEGF)
Prokineticin-2 gene (PROK2)	chr3 (p13)	Prokineticin-2 precursor	Prokineticin-2 (PK2, protein Bv8 homolog, PROK2 protein)
No-family neuropeptides			
Thyrotropin-releasing hormone gene (TRH)	chr3 (q21.3)	Prepro-TRH	TRH (thyroliberin)
Parathyroid hormone-like hormone gene (PTHLH)	chr12 (p11.22)	Prepro-PTH-like hormone, isoform 1	PTHrP-(1–36), PTHrP-(38–94), PTHrP-(107–139) (osteostatin)
Melanin-concentrating hormone gene (PMCH)	chr12 (q23.2)	Prepro-MCH	MCH, neuropeptide Glu-Ile (NEI), neuropeptide Gly-Glu (NGE)

Continued

Table 1 Continued

Gene (abbreviation)[a]	*Human chromosomal localization*[a]	*Precursor*[b]	*Active peptide(s)*[c]
Hypocretin gene (HCRT)	chr17 (q21.2)	Preprohypocretin	Hypocretin-1 (orexin A), hypocretin-2 (orexin B)
Cocaine- and amphetamine-regulated transcript gene (CART)	chr5 (q13.2)	Prepro-CART	CART-(1–39), CART-(42–89)
Agouti-related protein homolog gene (AGRP)	chr16 (q22.1)	AGRP precursor isoform 1	AGRP
		Prepro-AGRP isoform 2	AGRP
Prolactin (PRL)	chr6 (p22.3)	Prolactin precursor	Prolactin
Apelin gene (APLN)	chrX (q25)	Preproapelin	Apelin-13, apelin-17, apelin-36 (APJ ligand, AGTRL1 ligand)
Metastasis-suppressor KiSS (KISS1)	chr1 (q32.1)	Kiss-1	Metastin (kisspeptin-54), kisspeptin-14, kisspeptin-13, kisspeptin-10
Cerebellins			
Cerebellin-1 gene (CBLN1)	chr16 (q12.1)	Cerebellin-1 precursor	Cerebellin-1 (Cbln1)
Cerebellin-2 gene (CBLN2)	chr18 (q22.3)	Cerebellin-2-precursor	Cerebellin-2 (Cbln2)
Cerebellin-3 gene (CBLN3)	chr14 (q12)	Cerebellin-3 precursor	Cerebellin-3 (Cbln3)
Cerebellin-4 gene (CBLN4)	chr20 (q13.2)	Cerebellin-4 precursor	Cerebellin-4 (Cbln4, cerebellin-like glycoprotein-1)
Adipose neuroepeptides			
Leptin/ob gene (LEP)	chr7 (q32.1)	Preproleptin	Leptin (obesin)
Adiponectin gene (ADIPOQ)	chr3 (q27.3)	Adiponectin precursor	Adiponectin (Acpr30)

Visfatin gene (PBEF1)	chr7 (q22.2)		Visfatin precursor	Visfatin-1 (pre-B cell colony enhancing factor-1 (PBEF1), nicotinamide phosphoribosyltrans-ferase)
Resistin gene (RETN)	chr19 (p13.2)		Resistin precursor	Resistin (Cysteine-rich secreted protein FIZZ3, Adipose tissue-specific secretory factor, Cysteine-rich secreted protein A12-alpha-like 2)
			Resistin-delta2 precursor	Resistin-delta2
Resistin-like gene (RETLB)	chr3 (q13.13)		Resistin-like molecule Beta precursor (RELMbeta)	Resistin-like molecule (cysteine-rich secreted protein FIZZ2, colon- and small intestine-specific cysteine-rich protein, cysteine-rich secreted protein A12-alpha-like 1, colon carcinoma-related gene protein)
Nucleobindin-2/NEFA gene (NUCB2)	chr11 (p15.1)		Nucleobindin-2	Nesfatin-1

[a]Genes are abbreviated according to the official human gene nomenclature and linked to the corresponding locus in the human genome (March 2006 assembly) through the USCS Genome Browser. The ideograms show the human chromosomal localization. The USCS Genome Browser provides bioinformatics data on transcripts, expression, homologies to other species and gene variations. The Allen Brain Atlas provides data on *in situ* hybridization in the mouse brain for each gene.
[b]Precursors are given as prepro-peptide and are linked to homologous proteins of human and other species through the National Institutes of Health NCBI . Primary structures of precursors are found here by GI numbers. Different splice variants of the same gene are given separately.
[c]Neuropeptides are presented as the processed, biologically active products of the neuropeptide precursors (prepro-peptide) and the encoding gene. The presented neuropeptides fall under the conservative definition, and brain expression has been reported. The exception is the family of adipose neuropeptides, which have been implicated as hormones produced by adipose tissues signaling to the brain. However, increasing evidence indicates that they can also be produced by neurons within the nervous system. Neuropeptide genes are classified into families that share structural similarities or functional similarities (opioid peptides). Biological activities and cognate receptors have not been described for all mentioned neuropeptides.
This table was condensed from the online neuropeptide database http://www.neuropeptides.nl

system and should therefore be added to the class of neuropeptides.

Chemokines

Chemokines (from chemotactic cytokines) are originally known as secreted factors that mediate leukocyte trafficking. They form a family of approximately 50 proteins of 8–14 kDa and share structural similarities that classify them in the C, CC, CXC, and CXXXC subfamilies. This classification refers to the structural configuration of two conserved Cys residues. The nomenclature of individual chemokines is, for example, XCL1, CCL1, CXCL1, and CX3CL1, respectively. An extensive database is available online.

There is growing evidence for the neuronal expression of chemokines, chemokine receptor expression by neurons, and electrophysiological effects of chemokines on neurons. Neural functions of chemokines have been implicated in neuronal migration, neuroinflammation, and neuronal excitability. Chemokines, at least a number of them, may thus be considered as potential neuropeptides.

Neural Growth Factors

Growth factors are proteineous signaling molecules that are produced and secreted from cells. They provide extracellular support to the proliferation, differentiation, and maintenance of other or the same cells in a paracrine or autocrine fashion, respectively. Among the growth factors, many are expressed in the brain, where they act during embryonic development and during plastic neuronal changes and maintenance states of the adult nervous system. Some may act like transmitters, for example brain-derived neurotropic factor (BDNF); these growth factors fulfill the criteria of neuropeptides and should be considered as such. Growth factors come in many families, often with a large and diverse family tree. Examples are the neurotropins, fibroblast growth factors (FGFs), Wnts, bone morphogenic factors (BMPs), epidermal growth factor (EGF), and transforming growth factor (TGF-)α and -o, including inhibin/activins, granulins, and others. For many or most individual members, the expression in the nervous system and their participation in neural communication is not known. It is beyond the scope of this article to list all the growth factors which could potentially considered neuropeptides.

Bioinformatics of Neuropeptides

Bioinformatics provides tools for the analysis of large amounts of data which can be used in originals ways to answer biological questions. Public databases have become an important starting point for bioinformatics; for example, data on genome, gene, and transcript for all neuropeptides can be found in the Human Genome Browser. A comparison of the chromosomal localization of the neuropeptide genes (**Figure 1**) shows, first, that neuropeptide genes are localized all over the human genome and, second, that genes of neuropeptides that belong to the same families are not generally clustered in the same locus. This indicates that neuropeptides with related structural or functional features are not the result of recent gene duplication, but have an ancient evolutionary origin. Only a few hot spots can be distinguished:

- *Chromosome 1(p36)*. The CST, UTS2, NPPA, and NPPB genes.
- *Chromosome 11(p15)*. The CALCA, CALCB, ADM, NUCB2, SCT, IGF2, and INS genes.
- *Chromosome 17(q21)*. The GAS, HCRT, PPY, PYY, GIP, and HXPH3 genes.

In these three loci, closely related genes consecutive to one another exist: NPPA-NPPB on chromosome 1 (p36.22), CALCA-CALCB on chromosome 11(p15.2), INS-IGF2 on chromosome 11(p15.5), and PPY-PYY on chromosome 17(q21.31). The AVP–OXT locus on chromosome 20(p13) is another locus with two related genes. These are probably rare examples of evolutionary gene duplications. The later genes share a high degree of homology, are positioned in the opposite orientation with a very short intergenic region, are expressed by related magnocellular neurons in the hypothalamo-neurophypophysial system, and code for highly homologous precursors, but still the active peptides AVP and OXT have distinct physiological functions. The pro-melanin-concentrating hormone (PMCH) and insulin-like growth factor (IGF-)1 genes are neighbors on chromosome 12(q23.2). Notably, the INS-IGF2 gene is a read-through gene which overlaps with the insulin (INS) gene at the 5′ region and with the IGF2 gene at the 3′ region. The INS-IGF2 locus is on chromosome 11(p15.5).

Comparing structures of neuropeptide precursors between species (**Table 1**) shows that neuropeptide precursors are not very strongly conserved proteins, in contrast to, for example, transcription factors. However, the highest conservation is always found for the peptide domains that carry biological activities. With the notion that these are flanked by paired-basic PC cleavage sites and sometimes have a C-terminal Gly for amidation, bioinformatics searches can be designed for these features to discover new biologically active peptides in existing precursors or even to perhaps find novel neuropeptide precursors and genes. In this respect, knowledge about the expression in neural systems is essential. The online Allen Brain Atlas provides valuable data about *in situ* hybridization for virtually

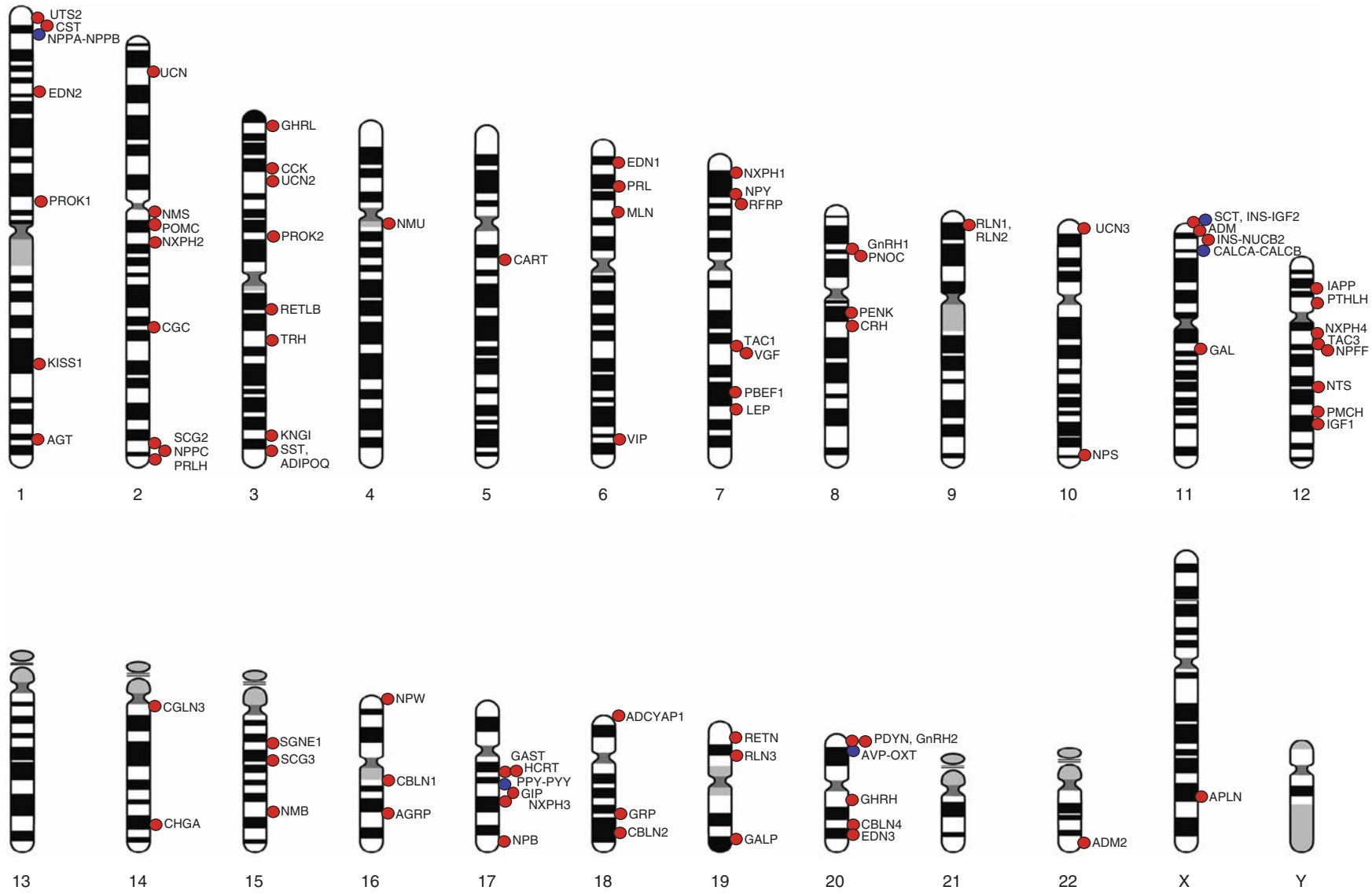

Figure 1 Ideogram of the chromosomal localization of neuropeptide genes in the human genome. Gene loci are indicated by red dots. Blue dots localize gene duplications within the same locus. Note the presence of hot spots, such as those on chromosomes 1, 11, 12, 17, and X, and the more empty chromosomes 4, 5, 9, 13, and 21.

all annotated genes in the mouse genome. These data can be accessed for all neuropeptides mentioned in Table 1.

Conclusion

Almost 90 genes exist in the mammalian genome that encode neuropeptide precursors. Clustering on structural features and on rather arbitrary functional aspects results in approximately 10 subfamilies and many no-family peptides that do not have a close structural similarity to neuropeptides. The definition of these 90 genes is largely based on a classical view on neuropeptides. A broader view may include genes that encode other signaling proteins and peptides. These and the analysis of the large number of hypothetical proteins in the genome identified by bioinformatics may lead to an expansion of the neuropeptide family, making the neuropeptides by far the largest class of signaling molecules in the brain.

See also: Invertebrate Neurohormone GPCRs; Nerve Growth Factor; Neuropeptide Signaling in Invertebrates; Neuropeptide Inactivation or Metabolism; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides Phylogeny and Evolution; Neuropeptides: Discovery; Oxytocin (Peripheral/Central Actions and their Regulation).

Further Reading

De Wied D (1987) The neuropeptide concept. *Progress in Brain Research* 72: 93–108.

Hökfelt T, Johansson O, Ljungdahl A, Lundberg JM, and Schutlzberg M (1980) Peptidergic neurons. *Nature* 284: 515–521.

Hughes J, Smith TW, Kosterlitz HW, Fothergill LA, Morgan BA, and Morris HR (1975) Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature* 258: 577–580.

Kim T, Gondre-Lewis MC, Arnaoutova I, and Loh YP (2006) Dense-core secretory granule biogenesis. *Physiology* 21: 124–133.

Klavdieva MM (1995) The history of neuropeptides I. *Frontiers in Neuroendocrinology* 16: 293–321.

Klavdieva MM (1996) The history of neuropeptides II. *Frontiers in Neuroendocrinology* 17: 126–153.

Klavdieva MM (1996) The history of neuropeptides III. *Frontiers in Neuroendocrinology* 17: 155–179.

Klavdieva MM (1996) The history of neuropeptides IV. *Frontiers in Neuroendocrinology* 17: 247–280.

Tatemoto K and Mutt V (1980) Isolation of two novel candidate hormones using a chemical method for finding naturally occurring polypeptides. *Nature* 285: 417–418.

Ubogu EE, Cossoy MB, and Ransohoff RM (2006) The expression and function of chemokines involved in CNS inflammation. *Trends in Pharmacological Sciences* 27: 48–55.

Relevant Websites

http://www.brain-map.org – Allen Brain Atlas.

http://cytokine.medic.kumamoto-u.ac.jp – Chemokine database.

http://www.ncbi.nlm.nih.gov – National Center for Biotechnology Information, National Institutes of Health. Structure of neuropeptide precursors.

http://www.neuropeptides.nl – Neuropeptides.

http://genome.ucsc.edu – University of California Santa Cruz Human Genome Browser.

Mammalian Sleep and Circadian Rhythms: Flies

J M Donlea and P J Shaw, Washington University in St. Louis, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Does the Fruit Fly Sleep?

Although the fruit fly has been effective in elucidating fundamental biological principles that are relevant for human health and disease, the observation that it lacks the neural architecture to generate mammalian-like sleep electroencephalogram (EEG) patterns has raised concerns. These concerns are based on two issues: (1) Can an organism with little structural similarity to mammals inform us about the mammalian condition? And (2) are the electrophysiological signals characteristic of mammalian sleep required to carry out sleep function? A growing amount of data indicates that, even in the absence of structural similarities, the fly can identify common elements relevant for mammalian physiology. For example, while the auditory organ, visual apparatus, olfactory systems, and core circadian pacemaker of the fly share no overt structural similarities with those of mammals, they all share common molecular frameworks. Moreover, electrophysiological activity of the cerebral cortex may not be required to carry out critical sleep functions. That is, recent data by Cirelli and colleagues suggest that the cellular processes associated with sleep occur in brain regions, such as the cerebellum, in which the signature sleep EEG patterns are either attenuated or absent. This finding suggests that cortical EEG patterns may only be important to carry out some, but not all, of sleep's functions. Together these data indicate that, in the absence of mammalian-like cortical structures and EEG patterns, *Drosophila* can be utilized to identify fundamental sleep processes.

Historically, sleep has been defined by both electrophysiological and behavioral criteria. In mammals and birds, sleep has been subdivided into two separate states, nonrapid eye movement (NREM) and rapid eye movement (REM) sleep. These states can be distinguished from each other and from waking based upon the distinct patterns of electrical potentials measured in the cortex (EEG), the muscle (electromyogram; EMG), the eye (electrooculogram; EOG), and, in some species, the deep brain structures (e.g., pontogeniculooccipital waves). No single electrophysiological signal has yet proved sufficient to unambiguously describe sleep. Similarly, behavioral definitions of sleep also require the simultaneous evaluation of several criteria. These criteria include (1) prolonged periods of quiescence, (2) reduced responsiveness to external stimuli, (3) rapid reversibility, which distinguishes sleep from hibernation or coma, and (4) homeostatic regulation – the increased need for sleep that follows sleep deprivation.

Careful observations of fly behavior using a variety of monitoring techniques have demonstrated that *Drosophila* flies exhibit consolidated periods of both activity and quiescence. During periods of quiescence lasting 5 min or longer, flies are unresponsive to mild external stimuli but can be quickly aroused with stronger stimulation. Furthermore, when flies are kept awake they respond by exhibiting large compensatory increases in the amount of quiescence the next day. These periods of quiescence are associated with increased arousal thresholds and cannot be explained by confounding variables, such as the intensity of the stimulation used to keep them awake, increased locomotor activity, or food deprivation. Indeed, homeostatic regulation has now been observed in flies that have been kept awake using at least five distinct deprivation methods. Thus, all of the behavioral requirements of sleep, originally defined for mammals, are observed in the fly.

Do quiescent periods in the fly share other important similarities with mammalian sleep? As in mammals, pharmacological agents such as caffeine, amphetamine, and cocaine increase waking in flies. In addition, pharmacological agents that promote sleep in mammals, such as antihistamines, also increase sleep and reduce the latency to sleep onset in flies. More importantly, several molecular markers that are modulated by sleep and waking in mammals are similarly modulated in *Drosophila*. Thus, with the exception of showing sleep spindles, K-complexes, and slow synchronized cortical activity in the frequency range 1–4 Hz, sleep in the fruit fly is remarkably similar to that found in mammals and birds. Indeed, behavioral, ontogenetic, pharmacological, molecular, and genetic studies indicate that quiescence in *Drosophila* shares many of the critical features of mammalian sleep.

What Can Circadian Mutants Tell Us about Sleep?

In mammals, sleep is regulated by the interaction between two processes. The first is a homeostatic process which increases during waking and dissipates during sleep. The second is a circadian process which alternatively promotes both sleep and wakefulness at specific circadian times. Thus, a wake-promoting signal, generated by the circadian clock, increases throughout the biological day to counter increasing

sleep drive and maintain appropriate levels of alertness. During the latter part of the biological night, a sleep-promoting signal, also generated by the circadian clock, increases to preserve sleep as homeostatic signals are dissipated. These signals create wake-maintenance and sleep-maintenance zones, respectively. Just as it is difficult to stay awake and vigilant during the sleep-maintenance zone, sleep is distorted when it occurs during the wake-maintenance zone. Nonetheless, elucidating the anatomy, physiology, and neural circuitry necessary for the clock to modulate sleep and waking in mammals poses a significant challenge. Fortunately, *Drosophila* chronobiologists have generated a number of genetic tools to investigate the cellular physiology of both pacemaker cells and their outputs.

The Timing of Sleep Is Influenced by the Circadian Clock in *Drosophila*

Both flies and humans spend the majority of their biological day engaged in waking activities and consolidate the majority of their sleep during the biological night (Figure 1). Interestingly, flies, like humans, display biphasic sleep patterns with a circadian-modulated elevation in sleeptime during the mid-afternoon (a 'siesta'). Moreover, flies exhibit evidence for a wake-maintenance zone at the end of the biological day as measured by increased wake time and an increase in the intensity of activity (Figures 1(a) and 1(b)). Thus, the clock influences the amount and timing of both sleep and waking during the biological day. Does the clock influence the amount and timing of sleep at night? As sleep drive increases, flies would sleep in longer bouts. Thus, the longest sleep bouts should occur at the point of the circadian day, when sleep drive is highest and circadian wake-promoting signals are low. Indeed, the longest sleep bouts tend to occur at the beginning of the biological night (Figure 1(c)). This pattern is reminiscent of that seen when evaluating electrophysiological makers of sleep intensity in mammals. Thus, as in mammals, sleep and waking are strongly modulated by the clock during both the biological day and night.

Homeostatic Sleep Drive Is Gated by the Circadian Clock in *Drosophila*

When wild-type *Canton-S* (*Cs*) flies are kept awake at night, they exhibit a compensatory increase in sleep the next day. In the morning following sleep deprivation, sleep is dramatically increased over baseline but declines during the wake-maintenance zone (Figure 2). Sleep remains at pretreatment levels during the subsequent biological night, presumably due to ceiling effects, and then increases again above baseline the second day during the noontime siesta. Sleep once again declines during the wake-maintenance zone in the late day and increases above baseline once again the following afternoon. Thus, the clock gates the timing of homeostatic discharge. This interpretation is further supported by the observation that sleep homeostasis is attenuated or blocked when flies are released into recovery during the wake-maintenance zone. As seen in Figure 2, the sleep rebound did not occur during the last 3 h or the wake-maintenance zone and was delayed ~15 h until the next morning. Together these data indicate that the clock can either facilitate or interfere with sleep at various times of the circadian day, even when sleep drive is very high.

Sleep Homeostasis Is Altered in Canonical Loss-of-Function Clock Mutants

Sleep homeostasis has been evaluated in the canonical loss-of-function clock mutants *period* (per^{01}), *timeless* (tim^{01}), *Clock* (Clk^{jrk}), and *cycle* (cyc^{01}) to investigate the interaction between homeostatic and circadian processes. When per^{01}, tim^{01}, and Clk^{jrk} flies are kept awake for 12 h, they are able to reclaim 100% of the sleep that was lost within the first 12 h of recovery (Figure 3). In comparison, wild-type flies with an intact circadian clock recover between 40% and 50% of their lost sleep within the same time frame (Figure 2(a)). These experiments succeeded in replicating the classic studies in rodents showing that sleep homeostasis was preserved in animals in which the SCN had been ablated. However, the failure to observe homeostasis in tim^{01} mutants following shorter sleep deprivation protocols suggested that the components of the clock may play additional roles beyond simply keeping time. Indeed, cyc^{01} mutants demonstrate one of the most dramatic responses to sleep deprivation yet described. Even minor disruptions in sleep in cyc^{01} mutant flies, lasting as little as 2–3 h, produce a vigorous compensatory sleep rebound that is 6 times larger than that of wild-type flies and is not complete until the cyc^{01} fly has slept an additional 3 min over baseline for each minute of sleep lost. Increased sleep homeostasis was observed even when the cyc^{01} mutation had been placed into three additional genetic backgrounds and tested with the appropriate deficiency, indicating that the observation was unlikely a result of background effects. Perhaps the most interesting finding was that the cyc^{01} flies began to die when they were kept awake for 10 h. It should be noted that death following 12 h of sleep deprivation is an extremely uncommon occurrence. For example, it requires approximately 70 h for *Cs*, per^{01}, and Clk^{jrk} mutants to succumb from sleep loss. Interestingly, the other clock mutant with altered homeostasis, tim^{01},

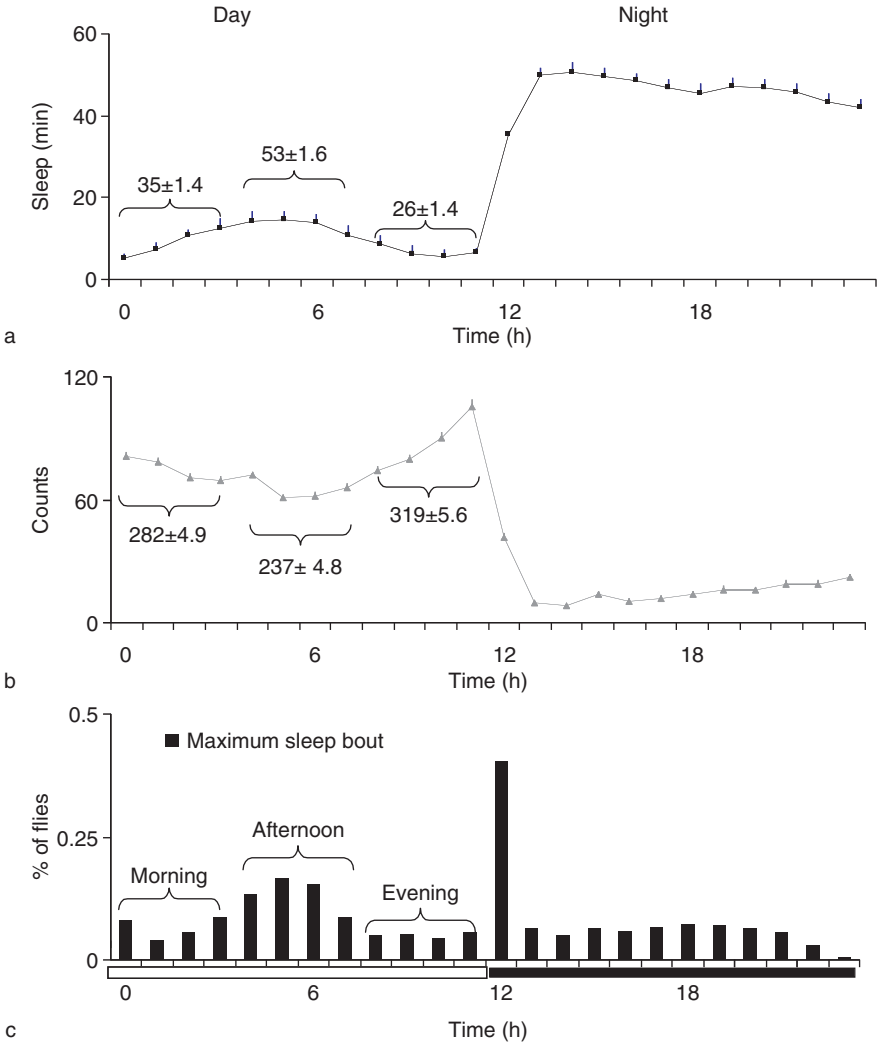

Figure 1 Circadian distribution of sleep and waking. (a) Minutes of sleep (mean ± SEM) for each hour of the 24 h day in female *Canton-S* flies maintained on a 12:12 light:dark schedule. (b) Locomotor activity as measured by infrared beam crossing in TriKinetics activity monitors. (c) Percentage of flies that exhibit their longest sleep bout of the day as a function of circadian time. Values are calculated separately for day and night to allow visualization of daytime sleep. Brackets and values indicate the sum of sleeptime activity during 4-h blocks in the morning, afternoon, and evening. Open horizontal bar indicates lights on; closed horizontal bar indicates lights off.

succumbed in 30 h, half as much time required for wild-type flies. Sleep deprivation-induced lethality was observed in cyc^{01} using three independent deprivation methods, while no lethality was observed when cyc^{01} flies were administered the same total number of stimuli but without producing 12 h of continuous wakefulness. Importantly, cyc^{01} mutants were more resistant than were other clock mutants to various stressors. These data indicate that the mortality associated with sleep deprivation was not due to hypersensitivity to nonspecific environmental stimuli.

Following sleep deprivation, cyc^{01} flies exhibited a small but consistent reduction in the expression of several stress response genes, including *hsp70*, *Hsp83*, *droj1*, and *hsc70-3*. In contrast, these genes remained near baseline or were elevated in the other clock mutants and wild-type flies, a result also observed in rats and mice. When chaperone proteins are activated in cyc^{01} flies by briefly exposing them to heat prior to sleep deprivation, the lethal effects of sleep loss are ameliorated. In addition, flies carrying a mutation for the *heat shock protein Hsp83* ($Hsp^{8308445}$) showed exaggerated homeostatic response and died after sleep deprivation. Taken together, these data indicate that the induction of certain chaperone proteins protects against the lethal effects of sleep loss and provides a first hint about the functional targets of sleep and its molecular mechanisms.

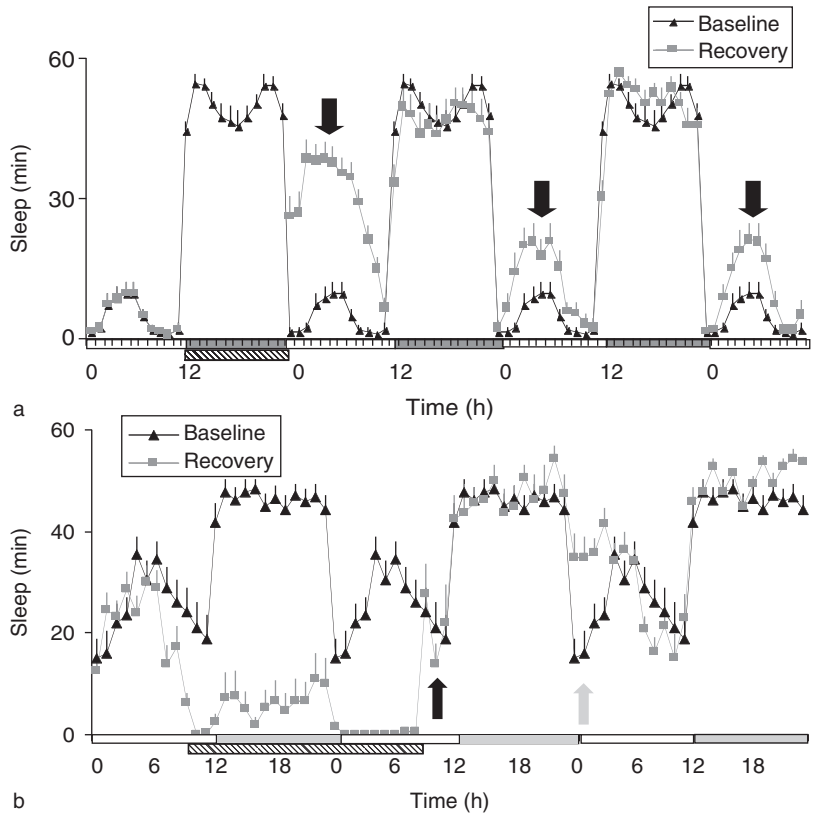

Figure 2 Sleep homeostasis is gated by the clock. (a) Following 12 h of sleep deprivation, homoeostasis occurs predominantly during the noontime siesta (black arrows). (b) Following 24 h of sleep deprivation ending during the wake-maintenance zone (black arrow), homeostasis is delayed by ~15 h (gray arrow). Minutes of sleep per hour in female flies maintained on a 12:12 light:dark schedule ($n = 32$). Sleep deprivation is indicated by hashed horizontal bars.

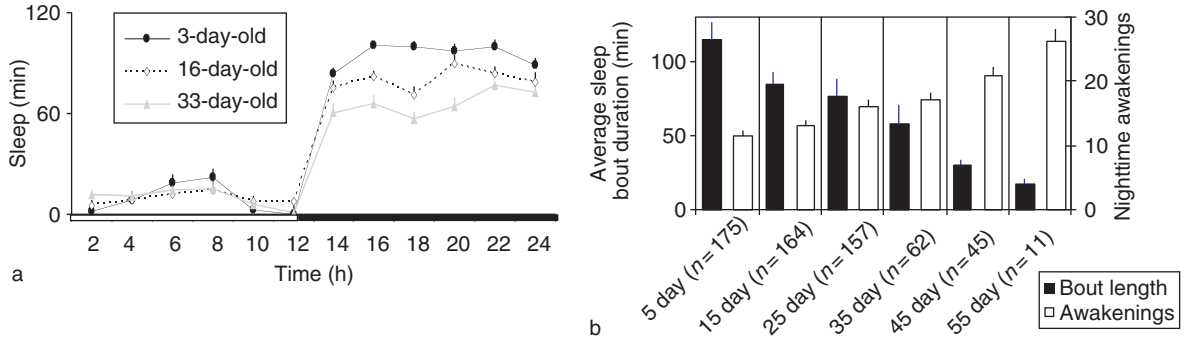

Figure 3 Sleep is disrupted in older flies. (a) Minutes of sleep per hour in young (circles), middle-aged (diamonds), and older (triangles) flies maintained on a 12:12 light:dark schedule. Bar indicates lights on and lights off. (b) Average sleep bout duration declines with age (black) while the number of nighttime awakenings increases (white).

Aging

Aging is associated with modification in both sleep and circadian systems. In humans, sleep duration declines with age and is associated with increased nighttime awakenings and reduced slow-wave sleep. Aging is also associated with modification in circadian timing as measured by reduced period and amplitude of temperature, sleep, and hormonal rhythms. These altered rhythms are associated with observable changes in the functioning of the suprachiasmatic nucleus, which is the master circadian pacemaker in mammals. One example is an age-dependent reduction in the expression of core clock genes, including *Clock* and the mammalian homolog, *cycle* (*Bmal1*). Frequently, the circadian sleep cycle is phase advanced such that older adults both go to sleep earlier and wake up

earlier than do younger individuals. Not surprisingly, the prevalence of sleep disorders such as sleep apnea, periodic limb movement disorder, restless legs syndrome, and insomnia increase with age and pose an increased risk for health problems and reduced quality of life.

Because sleep deficits both precede and follow age-related changes in physical and mental illnesses, a causal role of sleep on aging remains unproved. While it is clear that disease can interfere with sleep, data now indicate that sleep quality can, in turn, influence the progression of disease. It should be noted that no intervention study in humans has evaluated the long-term effects of sleep deficits in older adults. As a consequence, there is no consensus regarding which, if any, age-related sleep changes are normative. One barrier to understanding how sleep and aging interact is the high cost of long-term sleep recordings in humans and rodents. An additional barrier for establishing a causative role of sleep on aging is the absence of an experimental protocol that can induce sleep deficits recapitulating those found in older human adults. Virtually, all sleep disruption experiments in humans and animals are very short in duration, lasting between hours and a few days. The amount of sleep lost expressed as a percentage of lifetime sleep is quite small (~0.037%). Even lethal doses of chronic long-term sleep deprivation in rats can be viewed as an acute challenge in that only 2% of lifetime sleep is lost during a discrete portion of the animal's life. In contrast, the common sleep deficits reported by humans tend to last for several years or more. For example, it has been estimated that an individual will suffer from sleep apnea between 4 and 14 years before being diagnosed. Even after diagnosis, full-time compliance with recommended treatment options is rare. The median duration of severe insomnia is 4 years, and 44% of these individuals continue to report symptoms 10 years later. Excessive daytime sleepiness can be quite protracted in older adults, and in Parkinson's disease patients persists for 4 years or more. Not only do current sleep disruption protocols fall short of approximating the sleep loss found in clinical sleep disorders, they may not even reflect the amount of sleep lost in healthy adults due to chronic voluntary or involuntary (e.g., shift work) sleep restriction. To summarize, sleep deprivations in human and animal studies occur once during a discrete portion of the individual's life, are short in duration, and do not approximate challenges that are commonly observed in populations that experience sleep deficits. Consequently, the long-term effects of chronic sleep disruption on health and aging are largely unknown.

Given its short life span, ~60 days, *Drosophila* is an ideal model system to study the relationship between sleep and aging. As with humans, sleeptime starts to decline in middle-aged *Drosophila* and continues to fall in older flies (**Figure 3(a)**). In addition to changes in sleeptime, older *Drosophila* display other alterations in sleep architecture that resemble those found in older human adults. Thus, sleep becomes fragmented, longer sleep episodes become rare, and the animals display an increase in nighttime awakenings (**Figure 3(b)**). Sehgal and colleagues have shown that sleep fragmentation is associated with a reduction in the strength of sleep–wake cycles as measured by fast Fourier transform. As flies become very old, daytime sleep increases, presumably in an effort to compensate for poor sleep at night, such that total sleep duration is not reduced further. Aging in the fly can be easily manipulated by simply rearing the animals at different ambient temperatures. Flies reared at warm temperatures (29 °C) have a reduced life span compared to flies reared at cooler temperatures (18 °C). Interestingly, when flies were housed at cooler temperatures, sleep impairments were slowed substantially over time compared with their accelerated-aging siblings. Thus, Sehgal and colleagues demonstrated that sleep fragmentation is associated with physiological, not chronological, age. Such a conclusion would not be possible using mammalian protocols and emphasize the utility of the fly to address problems that have been intractable using traditional approaches. It should be noted that intensive research efforts have begun to separately elucidate biological principles of aging and sleep. By combining these tools, it should be possible to determine the precise relationship between sleep and aging.

Plasticity

Both sleep and circadian rhythms play an important role in synaptic plasticity. Rhythms influence plasticity by modulating both transcription and neuronal firing rates during specific times of the biological day. Similarly, sleep is associated with specific changes in transcription and neuronal firing rates that are conducive for the modulation of synaptic strength. Indeed, studies in humans, rodents, and birds report that waking neuronal activity in specific circuits is replayed during sleep, and this replay is associated with improved performance. For example, when new memories are cued with an odor during waking and that odor is presented during sleep to facilitate reactivation of the circuit, individuals exhibit improved retention of memories. Although the earliest investigations in humans yielded conflicting

results, recent human studies have clearly established a connection between memory consolidation and sleep. These studies are complemented by a large number of studies in smaller animals; these studies have begun to tell a clear story – that sleep following a novel experience aids in the consolidation of newly formed memories into longer-lasting behavioral changes and that this function of sleep is phylogenetically conserved.

Given the success of using *Drosophila* as a model organism to identify candidate genes that are involved in complex behaviors such as circadian rhythms and learning and memory, it seems likely that the fly could be a valuable tool to elucidate the genetic and neural mechanisms of sleep-dependent plasticity. Indeed, Ganguly–Fitzgerald's recent data have provided a basis for the examination of sleep-dependent behavioral plasticity in the fly. Using a behavioral paradigm in which age-matched sisters were held either in social isolation (housed individually in small food tubes) or in social enrichment (housed in groups in larger food vials), it was found that after several days of enrichment, wild-type flies exhibit a significant increase in daytime sleeptime and intensity that lasts for several days. Through this social enrichment paradigm, genetic pathways that mediate experience-dependent changes in sleep were explored; flies that were mutant for several classical learning and memory genes showed defects in experience-dependent sleep following social enrichment. Many of these mutants, including rutabaga and dunce, are involved in the protein kinase A/cyclic adenosine monophosphate (PKA/cAMP) signaling cascade. This pathway has previously been implicated by Hendricks and colleagues in the modulation of sleeptime in the fly; several mutations that resulted in a downregulation of cAMP signaling induced a decrease in total sleeptime while, conversely, mutations that increased cAMP signaling elevated total sleeptime. This work also found that expression of an activating isoform of the transcription factor cAMP response element-binding protein (CREB) increases sleeptime in flies, while expression of a repressor isoform of CREB decreases total sleeptime. Interestingly, it has also been shown that CREB

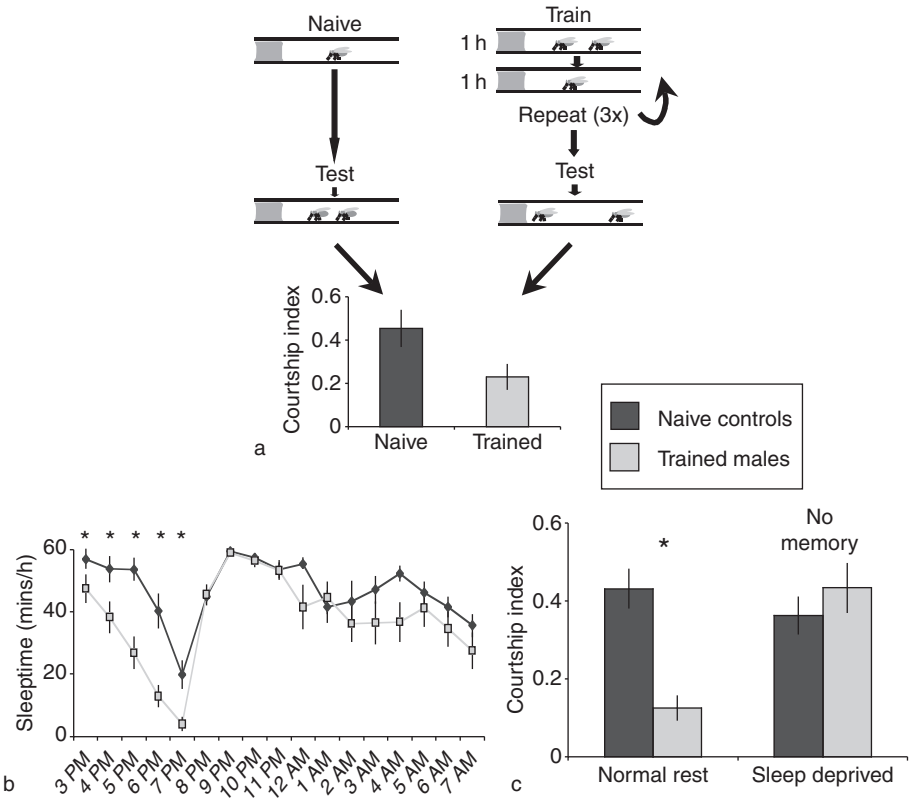

Figure 4 Sleep is necessary for memory consolidation. (a) Following a spaced training protocol consisting of three 1 h training periods with males that express aphrodisiac pheromones, a conditioned-stimulus male will display significantly reduced courtship toward a virgin female for up to 5 days. (b) Minutes of sleep per hour exhibited by either recently trained males (diamonds) or their naive brothers (squares). (c) Trained male flies demonstrate a significant reduction in Courtship Index toward virgin females for several days following training if allowed to rest *ad libitum* following training. Sleep deprivation immediately following courtship training prevents this long-lasting reduction in courtship.

activity plays an important role in the consolidation of short-term memory into longer-lasting changes in behavior; the same CREB activator that increases sleep also enhances memory consolidation while expression of the CREB repressor blocks memory consolidation.

While the neural circuits that mediate and are altered by sleep-dependent plasticity are unknown, independent studies at Northwestern University and the University of Pennsylvania have implicated the mushroom bodies (MBs) in sleep regulation. Insect MBs are higher-order sensory processing centers that have been shown to regulate olfactory learning and memory. This work demonstrated that disruption of cAMP/PKA signaling in subsets of neurons contained in the MBs was sufficient to profoundly alter sleeptime. Although these studies did not directly examine the effect of modulation of MB activity on sleep-dependent plasticity, the finding that MB neurons can play a role in sleep regulation suggests that the MB could be an important component of a circuit that relies on sleep to modulate the strength of synaptic connections.

While the molecular and anatomical components of experience-dependent changes in sleep have not been fully uncovered, it is clear that experience-dependent sleep plays a necessary role in the consolidation of transient, short-term memory into longer-lasting behavioral changes. This relationship has been studied more specifically in the fly using an associative memory paradigm called courtship conditioning (**Figure 4(a)**). In this paradigm, a male attempts to court, but is rejected by, recently mated females or males expressing aphrodisiac pheromones. As a result of this training, male flies will exhibit a dramatic long-lasting reduction in courtship toward normally attractive virgin females relative to their untrained naive brothers. During memory consolidation immediately after training, flies exhibit increased sleep (**Figure 4(b)**). Moreover, their sleep appears to be deeper during consolidation as evidenced by an increase in sleep bout duration. Importantly, depriving newly trained flies of sleep disrupts memory consolidation (**Figure 4(c)**). Thus, the increase in sleep following training appears to play an important role in the process of consolidation. These data are consistent with results obtained in humans, rodents, and birds, providing additional phylogenetic evidence that sleep plays an important role in memory consolidation. Together, all of these data suggest that this role may well be universally conserved. Future studies of sleep-dependent plasticity must take advantage of these complementary model systems to be most productive; ultimately, the fly may have much to tell us about how sleep alters the synapse.

See also: Circadian Regulation in Invertebrates; Circadian Rhythms: Influence of Light in Humans; Circadian Rhythms in Sleepiness, Alertness, and Performance; Genetic Regulation of Circadian Rhythms in Drosophila; Shift Work and Circadian Rhythms; Sleep and Waking in Drosophila; Sleep and Circadian Rhythm Disorders in Human Aging and Dementia.

Further Reading

Cirelli C, Lavaute TM, and Tononi G (2005) Sleep and wakefulness modulate gene expression in *Drosophila*. *Journal of Neurochemistry* 94: 1411–1419.

Ganguly-Fitzgerald I, Donlea J, and Shaw PJ (2006) Waking experience affects sleep need in *Drosophila*. *Science* 313: 1775–1781.

Hendricks JC, Finn SM, Panckeri KA, et al. (2000) Rest in *Drosophila* is a sleep-like state. *Neuron* 25: 129–138.

Hendricks JC, Lu S, Kume K, et al. (2003) Gender dimorphism in the role of cycle (BMAL1) in rest, rest regulation, and longevity in *Drosophila* melanogaster. *Journal of Biological Rhythms* 18: 12–25.

Hendricks JC, Williams JA, Panckeri K, et al. (2001) A non-circadian role for cAMP signaling and CREB activity in *Drosophila* rest homeostasis. *Nature Neuroscience* 4: 1108–1115.

Ji D and Wilson MA (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience* 10: 100–107.

Koh K, Evans JM, Hendricks JC, et al. (2006) A *Drosophila* model for age-associated changes in sleep:wake cycles. *Proceedings of the National Academy of Sciences of the United States of America* 103: 13843–13847.

Pitman JL, McGill JJ, Keegan KP, et al. (2006) A dynamic role for the mushroom bodies in promoting sleep in *Drosophila*. *Nature* 441: 753–756.

Rasch B, Buchel C, Gais S, et al. (2007) Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science* 315: 1426–1429.

Shaw PJ, Cirelli C, Greenspan RJ, et al. (2000) Correlates of sleep and waking in *Drosophila melanogaster*. *Science* 287: 1834–1837.

Shaw PJ, Tononi G, Greenspan RJ, et al. (2002) Stress response genes protect against lethal effects of sleep deprivation in *Drosophila*. *Nature* 417: 287–291.

Tononi G and Cirelli C (2006) Sleep function and synaptic homeostasis. *Sleep Medicine Reviews* 10: 49–62.

Walker MP and Stickgold R (2006) Sleep, memory, and plasticity. *Annual Review of Psychology* 57: 139–166.

MAP Kinase Signaling in Learning and Memory

S Blum and P K Dash, University of Texas Medical School at Houston, Houston, TX, USA

© 2009 Published by Elsevier Ltd.

Introduction

Structural and/or functional changes at synapses (synaptic plasticity) allow the nervous system to remain modifiable throughout life. These modifications in the nervous system are thought to allow learning and memory for events experienced by the organism and for adaptive responses in behavior. Two principal molecular mechanisms of synaptic plasticity are modification (e.g., phosphorylation) of existing proteins, and stimulation of gene expression and protein synthesis. While short-term memory (memories lasting for minutes to hours) involves phosphorylation of existing proteins, long-term memory (memories lasting days to weeks, and sometimes a lifetime) requires gene expression and protein synthesis. Gene expression and protein synthesis are initiated by phosphorylation of transcription factors and components of protein synthesis machinery.

The mitogen-activated protein kinases (MAPKs) are a group of related kinases which modulate the activity of intracellular proteins via phosphorylation, and can lead to changes in gene expression and protein synthesis. Three different families of MAPKs have been well characterized: the extracellular signal-regulated kinase (ERK), protein 38 (p38) MAPK, and c-jun N-terminal kinase (JNK) MAPK. These kinases are activated by growth factors, neurotransmitters, stress, and cell death signals. The basic MAPK cascade activation module applies to all three families, and consists of three kinases that are phosphorylated in a sequential manner: MAP kinase kinase kinases (MAPKKKs) are serine/threonine kinases that, when activated, phosphorylate and activate MAP kinase kinase (MAPKK). MAPKKs are dual-specificity kinases that phosphorylate both threonine and tyrosine residues within a Thr-X-Tyr (T-X-Y) motif in MAPKs that leads to MAPK activation (Figure 1). The X corresponds to the amino acid Glu in ERK, a Pro in JNK MAPK and a Gly in p38 MAPK.

Two mechanisms have been proposed to explain how specificity is achieved from an upstream MAPKKK to a specific MAPK: (1) scaffolding proteins that assemble different components of a specific MAPK cascade and (2) direct physical contact between the components of a cascade. MAPKs are serine/threonine kinases that, when activated, phosphorylate a large number of substrate proteins in the cell, including ion channels and transcription factors. The best-characterized MAPK pathway is the ERK pathway. Initially, ERK was found to be principally involved in cell responses to growth factors and other mitogens, controlling cell proliferation and differentiation. It was subsequently observed that ERK is highly expressed in mature neurons, which do not divide and are terminally differentiated. This, along with observations that ERK can be activated by glutamate and other neurotransmitters, pointed to a role for this cascade in synaptic plasticity, and learning and memory. Although roles in synaptic plasticity for the other two major MAPK cascades, JNK MAPK and p38 MAPK, are beginning to emerge, this article specifically focuses on the basic features of the ERK cascade and its role in synaptic plasticity, and learning and memory.

The Cascade

The prototypic members of the MAPK family, ERK1 (p44 MAPK) and ERK2 (p42 MAPK), are activated in response to growth factors and other neurotransmitters (Figure 1). The 'classical' mechanism for growth factor-mediated ERK activation is via receptor tyrosine kinases (RTKs) that become autophosphorylated on specific tyrosine residues following growth factor binding. When phosphorylated, these tyrosine residues create docking sites for the binding of adapter proteins such as growth factor receptor-binding protein 2 (Grb2), or intermediate factors such as Shc, FRS2, IRS-1, and IRS-2. Binding of Grb2 results in the accumulation of Grb2–son of sevenless (SOS) complexes near the cell membrane where SOS (a guanine exchange factor or GEF) activates Ras, a G-protein. Ras is activated when bound to the nucleotide GTP, an interaction facilitated by SOS. Conversely, Ras is inactivated when the GTP is hydrolyzed, a process augmented by GAPs (GTPase activating proteins). In addition to this 'classical' activation cascade, Ras can be activated following membrane depolarization or glutamatergic signaling via increased intracellular calcium. The mechanism by which elevated calcium triggers Ras activation differs slightly among cell types and various neurotransmitter stimuli, but invariably depends on GEFs and GAPs.

Immediate downstream signal transducers of Ras are the Raf family of proteins (Raf1, A-Raf, and B-Raf), which are the MAPKKKs of the ERK cascade. The major, and possibly only function of Ras in the activation of Raf, is the relocation of Raf to the cell membrane, where it can be activated by an unknown

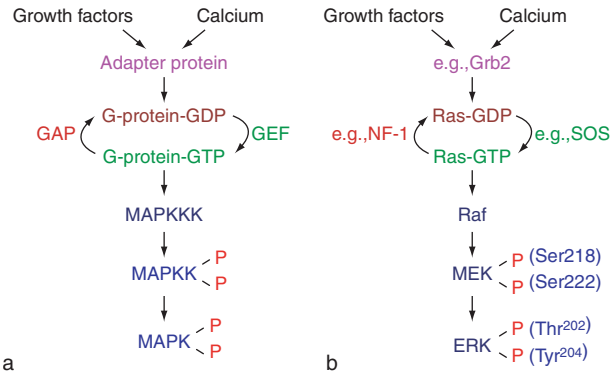

Figure 1 Activation of (a) MAPK cascade and (b) ERK cascade.

mechanism. Once activated, Raf then initiates the sequential phosphorylation and activation of MAPKKs or MEKs and ERK. Activated ERK phosphorylates a large range of substrates, including transcription factors, cytoskeletal proteins, regulatory enzymes, and other kinases.

In addition to ERK1 and 2, ERK3–8 have been identified. Of these, ERK5 (also called big ERK or BMK5) has received the most attention because of the discovery that commonly used MEK1/2 inhibitors PD98059 and U0126 also block the ERK5 pathway. Although the involvement of ERK5 in neuronal survival has been established, its role in learning and memory has not been examined.

Signal transduction for the MAPK family of proteins involves growth factor and/or neurotransmitter receptor-mediated activation of small G-proteins (e.g., Ras for the ERK pathway). In the absence of ligand or calcium rise, G-proteins are bound to GDP and are inactive. The activity of the G-protein is enhanced by ligand binding and the exchange of GDP for GTP, a process facilitated by GEFs. GTP-bound G-protein activates the MAPKKK (Raf for the ERK cascade), which, in turn, initiates the sequential phosphorylation of MAPKK and MAPK. Dual phosphorylation of MAPK on T-X-Y residues results in MAPK (or ERK) activation. MAPK can directly modulate cellular functions by phosphorylating cellular proteins, and indirectly through the activation of other kinases such as $p90^{rsk}$.

The ERK Cascade in Synaptic Plasticity

ERK is expressed at high levels in brain areas implicated in learning and memory. Pharmacological inhibitors of the ERK cascade (PD98059, SL327, UO126) became available in the mid-1990s and have intensified the investigation of the role of ERK in synaptic plasticity. These studies have utilized a variety of approaches, including cell culture, electrophysiology, and *in vitro* correlates of learning and memory, to evaluate the necessity of the ERK cascade in enhancing synaptic communication. For example, many features of learning and memory in *Aplysia californica* (a marine mollusk) can be reproduced in an *in vitro* co-culture system consisting of a sensory neuron and a motor neuron. In this preparation, a single application of the transmitter serotonin (a transmitter released during sensitization training) enhances synaptic communication between the sensory and the motor neuron that lasts for minutes, a correlate of short-term memory. Multiple applications of serotonin enhance synaptic transmission that can last for days, a correlate of long-term memory called long-term facilitation (LTF). Blockade of the ERK cascade, either by a blocking antibody against ERK (blocking or neutralizing antibodies bind to the protein of interest and inhibit its function) or by pharmacological inhibition of the upstream kinase MEK, impairs this LTF of synaptic transmission. Interestingly, MEK inhibition has no effect on short-term synaptic facilitation. Similarly, studies of ERK phosphorylation in *Hermissenda* have shown that both one-trial and multiple-trial conditioning paradigms (both paradigms induce long-lasting plasticity) are associated with increased activation of the ERK cascade.

Delineation of the role of the ERK cascade in synaptic plasticity has been furthered by electrophysiological studies. ERK signaling has been best studied in the classical LTP between CA3 and CA1 synapses of the hippocampus that requires *N*-methyl-D-aspartate (NMDA) subtype of glutamate receptors. Tetanizing stimulation that induces LTP in hippocampal CA3–CA1 synapses is associated with increased ERK phosphorylation, leading to its activation. This activation was found to occur in area CA1 and was dependent on both protein kinase C activity and stimulation of NMDA receptors. A requirement for ERK has also been demonstrated in cortical LTP. For example, ERK is required for a shift in ocular dominance following monocular deprivation. This process, thought to depend on cortical LTP, is prevented by MEK inhibitors. Similar to that seen for LTP, a requirement for ERK activity has also been observed in long-term depression (LTD), the weakening of neuronal synaptic strength. For example, cerebellar LTD (a proposed mechanism for motor learning and memory) is associated with activation of ERK in Purkinje cells and can be blocked by application of an MEK inhibitor or by intracellular infusion of an MAPK blocking antibody. ERK activity has also been shown to affect hippocampal CA1 theta frequency firing. This pattern of neuronal firing is typical when rats are learning about their environment, and appears to be

dependent on an ERK-mediated suppression of potassium channel function and an increase in membrane excitability. For example, inhibition of ERK activity in the CA1 subfield blocks complex spike bursting and the LTP that would normally be induced. One potassium channel, in particular, the voltage-dependent K^+ channel K_v 4.2, is an excellent ERK substrate and may mediate the influence of ERK on CA1 theta frequency firing.

Learning and Memory

The studies described above and others show that ERK activity is required for long-lasting synaptic plasticity. Consistent with these findings, ERK has been demonstrated to be critical for long-term memory in intact animals. The behavioral tasks in which the role of ERK has been best characterized are fear conditioning and spatial learning using rodents. Fear conditioning, a type of Pavlovian conditioning, is typically performed by training the subject to associate an otherwise nonthreatening cue such as a tone (called the conditioned stimulus or CS) with an aversive stimulus such as a mild footshock (the unconditioned stimulus or US). Upon subsequent presentation of the tone, trained animals display a characteristic period of immobility called the freezing response that can be monitored and used as a measure of fear memory. This cue-elicited fear response has been shown to depend on the function of the amygdala. In addition to this simple association, animals also learn to fear the specific environment (the context) in which the conditioning took place. Contextual fear, which is tested by placing the conditioned animal back into the training environment and monitoring freezing behavior, is dependent on both the hippocampus and the amygdala. Using phosphorylated ERK as an indicator of increased ERK activity, it has been reported that fear conditioning is associated with increased ERK activity in the hippocampus and the lateral amygdala of conditioned rats. Both systemic and targeted application of MEK inhibitors into animals prior to conditioning significantly impairs long-term fear memory. Consistent with this, transgenic mice carrying a dominant-negative MEK have decreased hippocampal ERK activity and reduced context-specific fear memory.

Spatial learning and memory, a hippocampus-dependent behavior necessary for spatial navigation, is commonly tested using the Morris water maze task. In this task, rats learn to use extra-maze cues to find the location of an escape platform hidden below the surface of the water. Training in the Morris water maze results in a significant increase in the number of hippocampal neurons, specifically those in area CA1, that contain activated ERK. *In vivo* inhibition of the ERK cascade within the hippocampus blocks the training-related activation of ERK and impairs long-term, but not short-term, memory for the location of the hidden platform. Importantly, this memory impairment only occurs when the inhibitor is administered at the time of ERK activation. Delaying the infusion beyond the time of ERK activation results in normal long-term spatial memory. In addition to fear conditioning and spatial learning, ERK activity has been shown to be necessary for long-term memory in other tasks such as taste-aversion conditioning, memory extinction, and recognition memory, demonstrating a conserved role for this kinase in numerous brain structures and associated behaviors.

Disruption of ERK-mediated plasticity in different parts of the brain may therefore provide insights into the contribution of a structure to a particular component of memory. For instance, lesion studies have shown that the prefrontal cortex is required for trace conditioning (a type of fear conditioning in which the CS and US are separated by a time gap), but not delay conditioning (a type of fear conditioning in which the CS and US co-terminate). However, it was not clear if the role of the prefrontal cortex in trace conditioning was due to memory storage or merely information processing. Using targeted infusions of MEK inhibitors, it has been shown that ERK-mediated plasticity in the prefrontal cortex is required for the storage of information related to the temporal relationship between two separated events. In contrast, ERK activation is not observed in the prefrontal cortex, but rather in the amygdala, following delay conditioning. These studies therefore confirmed the earlier findings regarding the involvement of the prefrontal cortex in trace conditioning, and extended this role to included one for memory storage rather than just information processing. In addition to differentiating the role of a structure in different behaviors, blockade of ERK-mediated plasticity can be used to identify the roles of specific structures within a behavior. For example, direct examination of the consequences of ERK inhibition in two structures required for performance in the water maze task, the hippocampus and the entorhinal cortex, has allowed for further delineation of the roles of these two structures in spatial memory. These studies revealed that ERK-mediated plasticity in the entorhinal cortex may be involved in storage of information related to the strategy needed for solving this spatial task, whereas the hippocampus appears to store information related to the actual location of the hidden platform. In concert, the memories stored in the entorhinal cortex and the hippocampus may allow for an animal to navigate from a distance to a known spatial

location. In contrast to the complimentary roles of these structures in spatial memory, inhibition of ERK-mediated plasticity in the hippocampus blunts, while its inhibition in the entorhinal cortex enhances, long-term contextual memory. These and other studies have demonstrated that targeted disruption of ERK signaling can be used as a method of specifically impairing long-lasting plasticity in order to assess the type of information stored within the structure of interest.

Downstream Effects of ERK Activation

The mechanism by which ERK-mediated signaling contributes to long-term synaptic plasticity and memory is likely to involve multiple cellular and molecular processes. As will be discussed below, these processes include enhanced synaptic transmission, changes in gene expression, protein synthesis, and morphological changes (Figure 2).

A number of mechanisms by which ERK alters neuronal properties have been identified, and include (1) synaptic insertion of AMPA receptors that can enhance synaptic communication, (2) phosphorylation of cytoskeletal proteins that can alter neuronal morphology, and (3) modulation of translational and transcriptional machineries that can cause long-lasting alterations in neuronal plasticity.

Synaptic Transmission

Several studies have indicated that an increase in number of postsynaptic AMPA subtype glutamate receptors is an important mechanism for LTP. This is thought to occur by insertion of AMPA receptors into the postsynaptic sites following LTP-inducing stimulation. LTP, as well as the synaptic insertion of AMPA receptors, requires the activity of calcium/calmodulin-dependent protein kinase II (CaMKII). Expression of an active form of Ras, which activates ERK, increases synaptic AMPA receptors and this effect is blocked by the MEK inhibitor PD98059. Furthermore, CaMKII-driven synaptic insertion of AMPA receptors also requires ERK activity (Figure 3). These studies suggest that ERK participates in synaptic plasticity by regulating insertion of AMPA receptors into postsynaptic sites, thereby increasing synaptic strength.

Gene Expression

One of the key requirements of long-term memory formation is altered expression of specific genes as a result of learning. Once activated, ERK can directly (via translocation to the nucleus as a dimer) and indirectly (via activation and subsequent translocation of downstream kinases) phosphorylate several transcription factors that alter the expression of target genes (Figure 2). These include the transcription factor CREB (cAMP response element binding protein), the ternary complex factor (TCF) family members Elk-1 and SAP-1, and the AP-1 (activator protein-1) family members c-Jun, c-Fos, and ATF-2. In support of a role for ERK in regulating memory-related gene expression, its translocation to the nucleus of hippocampal neurons has been observed following training of rodents in the Morris water maze task. Furthermore, the phosphorylation of

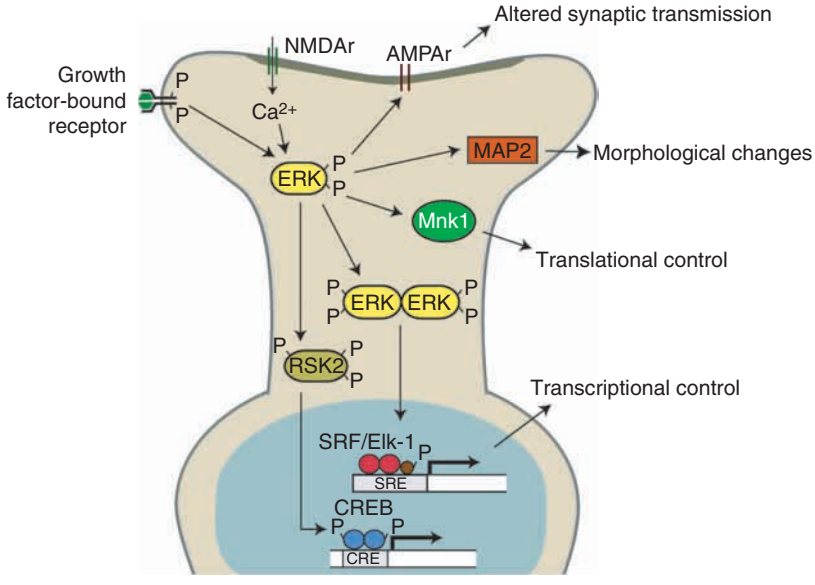

Figure 2 ERK-mediated synaptic plasticity may involve several downstream processes.

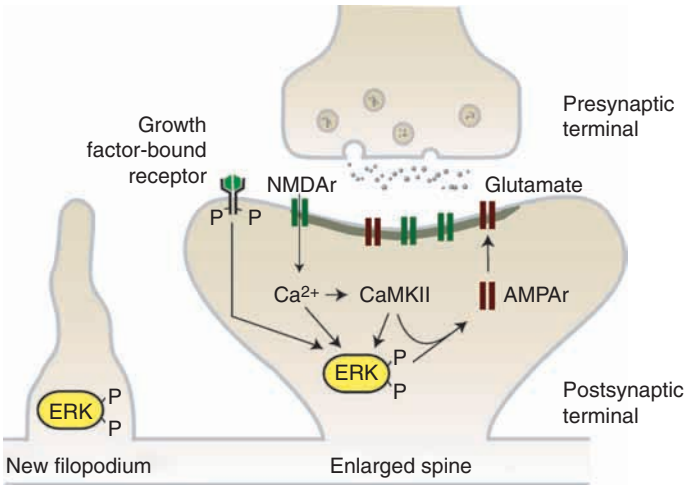

Figure 3 Synaptic plasticity requires ERK-mediated neuronal changes.

Elk-1 associated with *in vivo* LTP induction in the dentate gyrus can be blocked by a MEK inhibitor.

The phosphorylation of CREB on Ser133 increases the expression of genes containing the calcium/cAMP-response element (CRE) in their promoter regions. In response to neuronal activity, a rapid initial phosphorylation of Ser133 is carried out by CaMKIV, followed by a sustained phosphorylation mediated by ERK. Ser133 phosphorylation by ERK appears to be indirect and involves ribosomal protein S6 kinase 2 (RSK2, also called p90RSK or MAPKAP kinase-1) and/or mitogen and stress-activated kinase (MSK). CREB-mediated gene expression has been shown to be a necessary step for formation of long-term synaptic plasticity and memory in a variety of species. For example, blockade of CREB activity impairs long-term facilitation between *Aplysia* sensory–motor synapse, causes long-term memory dysfunction in the fruit fly *Drosophila melanogaster*, and impairs long-term spatial memory in mice and rats.

The expression of genes containing the serum response element (SRE) is enhanced by binding of the serum response factor (SRF)-Elk-1 complex. Elk-1 (and related TCF family members SAP-1 and SAP-2) are phosphorylated in response to ERK activation which facilitates DNA binding and complex formation with SRF. Genes whose protein products have been implicated in plasticity and memory such as zif268 and egr-1 are known to contain SRE sequences in their promoter regions, suggesting a role for SRF-Elk-1 in memory formation. Consistent with this suggestion, blockade of SRF-Elk-1 binding by direct hippocampal infusion of synthetic SRE sequence-containing decoy oligonucleotides impairs long-term spatial memory formation. Furthermore, Elk-1 phosphorylation is enhanced in the insular cortex following taste-aversion conditioning, the memory of which can be blocked by targeted infusion of a reversible MEK inhibitor.

In addition to phosphorylating specific transcription factors, ERK also phosphorylates other proteins that participate in transcription. For example, ERK-mediated histone H3 phosphorylation has been observed in the hippocampus of rats trained in a contextual fear-conditioning task, and histone acetyltransferase activity may be regulated, in part, through ERK-mediated signaling. As histone acetylation is a prerequisite for gene expression, these epigenetic mechanisms of ERK action may also play a role in regulating gene expression and long-lasting synaptic plasticity.

Protein Synthesis

The synthesis of proteins (or translation) from mRNA in general, and from mRNA localized at synapses that are activated during learning in particular, is thought to be critical for maintaining memory. Recent studies are beginning to show that ERK can regulate the translation of both somatically and dendritically localized mRNAs. The translation of mRNA requires not only ribosomes, but several additional proteins as well that participate in the initiation of translation. Of these proteins, translation initiation protein eIF4E (eukaryotic initiation factor 4), 4EBP-1 (eukaryotic initiation factor 4E-binding protein 1), and the ribosomal protein S6 have been demonstrated to be regulated by ERK-mediated phosphorylation. For example, mice expressing a forebrain-specific, dominant negative isoform of MEK (resulting in inhibited ERK activity) have impaired translation-dependent LTP that is associated with reduced phosphorylation of eIF4E, 4EBP-1, and S6. Similarly, NMDA-mediated activation of Mnk1 (the kinase

responsible for eIF4E phosphorylation) has been shown to depend on ERK activity. As ERK-dependent Mnk1 activation occurs in both the cell soma and the dendrites of NMDA stimulated neurons, this may represent a mechanism by which local and global mRNA translation are regulated for the development of long-lasting plasticity.

Morphological Changes

Morphological changes such as new synapse formation, enlarged spines, and postsynaptic density are routinely associated with long-term synaptic plasticity and memory. For example, stimulation to induce LTP and depolarization of neurons in culture have been associated with increase in the number of filopodia, a precursor of synapse formation, and the size of dendritic spines (Figure 3). Similarly, exposure of hippocampal CA1 slice cultures to brain-derived neurotrophic factor (BDNF), a potent mediator of synaptic plasticity, increases spine density in apical dendrites. Although the mechanism by which ERK participates in these processes has not been fully delineated, each of these morphological changes has been shown to be blocked by ERK inhibition. As MAPK was initially identified based on its ability to phosphorylate microtubule-associated protein (MAP), the phosphorylation of cytoskeletal proteins participating in spine morphology and synapse formation is likely to be involved. Alternatively (or in addition), enhanced translation of dendritically localized mRNA in response to ERK activity may underlie these morphological changes.

In response to training-related release of growth factors or neurotransmitters, activated ERK initiates cellular changes that alters synaptic plasticity. These include enlargement of dendritic spines, insertion of postsynaptic AMPA receptors, and the growth of new filopodium, an initial process for the formation of new synapses. Each of these processes has been shown to be blocked in response to treatments which inhibit the ERK cascade.

Ras-ERK Pathway in Human Memory

Mutation in genes encoding for proteins in the Ras-ERK cascade has been linked to cognitive and memory impairments in humans. The neurofibromatosis-1 gene (NF-1) encodes for a Ras-GAP. Mutations in the NF-1 gene prolong Ras activation leading to excessive stimulation of ERK and its downstream substrate proteins and are associated with learning and memory deficits in humans. Coffin–Lowery syndrome (CLS) is associated with dysfunctional motor learning and low IQ scores. This syndrome is caused by mutations in the RSK2 gene, a downstream effector of ERK. Tuberous sclerosis (TS) syndrome results from mutations in tuberin or hamartin (components of the tuberous sclerosis complex (TSC) that regulates protein translation) and is associated with reduced B-Raf and ERK activity. TS results in severe mental retardation, autism, seizures, and tumors of the brain, heart, and kidneys. Finally, Rubenstein–Taybi syndrome is caused by mutations in the gene encoding the CREB-binding protein (CBP – a histone acetylase) and is associated with mental retardation and poor learning and memory. These studies suggest that an optimal level of ERK activation may be required for normal memory formation with excessive or insufficient activation of the ERK cascade giving rise to cognitive dysfunction.

Further studies are needed to fully delineate the mechanisms by which ERK and its downstream effectors influence neuronal morphology and plasticity. Although several mechanisms have been suggested, the manner in which one or more of these processes is specifically initiated is a subject of continued investigation. Further examination of the interaction between the ERK cascade and other intracellular signaling pathways will shed light on the manner in which individual stimuli, or multiple stimuli in concert, give rise to plasticity-related cellular changes.

See also: Adult Cortical Plasticity; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Stress and Neuronal Plasticity.

Further Reading

Blum S, Moore AN, Adams F, and Dash PK (1999) A mitogen-activated protein kinase cascade in the CA1/CA2 subfield of the dorsal hippocampus is essential for long-term spatial memory. *Journal of Neuroscience* 19: 3535–3544.

Dash PK, Hochner B, and Kandel ER (1990) Injection of the cAMP-responsive element into the nucleus of Aplysia sensory neurons blocks long-term facilitation. *Nature* 354: 718–721.

Eichenbaum H (2002) *The Cognitive Neuroscience of Memory: An Introduction*. New York: Oxford University Press.

Kandel ER (2001) The molecular biology of memory storage: A dialogue between genes and synapses. *Science* 294: 1030–1038.

Kelleher RJ III, Govindarajan A, and Tonegawa S (2004) Translational regulatory mechanisms in persistent forms of synaptic plasticity. *Neuron* 44: 59–73.

Kornhauser JM and Greenberg ME (1997) A kinase to remember: Dual roles for MAP kinase in long-term memory. *Neuron* 18: 828–831.

Martin KC and Kandel ER (1996) Cell adhesion molecules, CREB, and the formation of new synaptic connections. *Neuron* 17: 567–570.

Squire LR and Kandel ER (1999) *Memory: From Mind to Molecules*. New York: W.H. Freeman.

Sweatt JD (2003) *Mechanisms of Memory*. San Diego, CA: Academic Press.

Weeber EJ and Sweatt JD (2002) Molecular neurobiology of human cognition. *Neuron* 14: 845–848.

Map Plasticity and Recovery from Stroke

R J Nudo and S Barbay, University of Kansas Medical Center, Kansas City, KS, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

One of the more remarkable discoveries of the late twentieth century regarding neural organization was that the brain, and in particular the cerebral cortex, is alterable throughout life. Maps of sensory and motor topography in the cortex are mutable, and functional boundaries can be altered in detail by use-dependent mechanisms. While such functional and structural plasticity had long been known during critical developmental stages, adult plasticity has heuristic appeal as it provides a framework to explain how the brain might accommodate the learning of new skills and behavioral strategies or adapt existing abilities to environmental challenges. In addition, neuroplasticity has come to be appreciated as one of the principal mechanisms for recovery of behavioral function after injury to the nervous system. While this area of research is still evolving, it appears that recovery after cortical (and perhaps subcortical) injury in adult mammals coincides with extensive functional and structural reorganization of the cerebral cortex. This article provides an overview of the role of cortical plasticity in recovery of function after a common clinical condition, ischemic stroke.

Stroke is the result of a cerebrovascular event that interrupts blood flow in the brain, depriving cells of oxygen and glucose. Necrosis can occur within minutes and, if not treated within the first few hours after onset, will likely result in some loss of function. Acute stroke therapy is still limited (e.g., thrombolysis, clot removal), and thus most stroke survivors experience some permanent loss of motor function. The motor areas of the cerebral cortex are typically affected by stroke as a result of reduced blood flow in the middle and anterior cerebral arteries that branch from the common carotid arteries supplying the brain with oxygenated blood. An occlusion in the middle cerebral artery, the most common location for clots to occur, can result in functional deficits in the upper extremities, such as the arm, hand, and face contralateral to the damaged cerebral hemisphere. Functional impairments in the lower extremities are also common after stroke.

Though residual impairments are often permanent, drastically affecting a stroke survivor's quality of life and ability to maintain independence, substantial recovery often occurs, especially in the first few weeks to months after the event. Such recovery is often said to be 'spontaneous' as it does not seem to depend on any particular intervention, such as physical or occupational therapy. Three basic theories regarding the underlying mechanisms contributing to spontaneous recovery are resolution of diaschisis, compensatory behavior, and adaptive plasticity (Figure 1). Diaschisis refers to the disruption in blood flow and metabolism in otherwise intact regions of the brain that are anatomically connected with the injured zone. Since these areas survive the acute event, metabolism and blood flow eventually are normalized, and thus at least some recovery is thought to result from the resolution of this process. Compensatory behaviors are common after stroke in humans and in animal models. As the affected limb is used less, individuals become functional by using the better limb or by employing alternative kinematic strategies (e.g., use of proximal musculature to propel a weak distal limb forward). Finally, adaptive plasticity is now thought to occur in both the peri-infarct and more remote tissue and is the focus of the remainder of this article. It has long been assumed that other, undamaged portions of the cerebral cortex must have the ability to 'take over' the function of the damaged region. A better understanding of how this process occurs, therefore, might open a path to development of novel interventions that optimize this process.

New technologies now allow us to track functional changes in the cerebral cortex after stroke. Cortical plasticity is therefore becoming a major focus for diagnostic assessment of stroke patients through various noninvasive techniques such as functional magnetic resonance imaging (fMRI) and transcranial magnetic stimulation (TMS). Interpretation of information derived from these techniques is greatly influenced by basic animal research that allows for more-precise measures of cortical function than is currently possible with humans. The consequence from bridging basic and clinical research is a better understanding of how cortical motor areas interact within various neural networks to produce adaptive motor control under normal and perturbed conditions.

Animal research has been vital in addressing basic mechanistic questions regarding normal cortical function as well as how the cortex responds to injury. The goal of current research is to understand how neural plasticity can be directed through behavioral and pharmacological therapies to optimize recovery of function and foster relearning of motor skills.

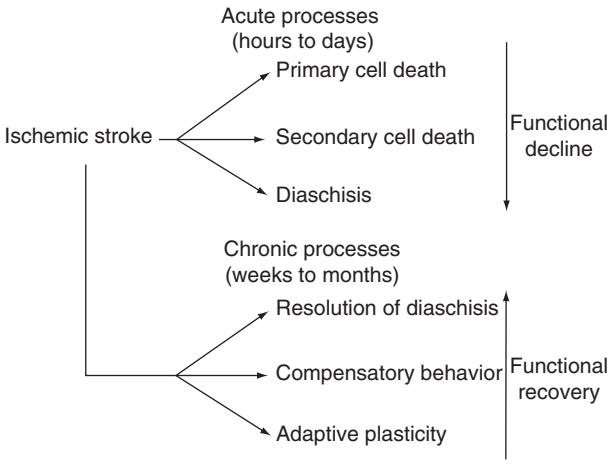

Figure 1 Acute and chronic processes contributing to initial functional deficits and subsequent spontaneous recovery after stroke. Neuronal cell loss occurs rapidly in the ischemic core. Further neuronal cell loss occurs over the course of hours to days in the peri-infarct region (penumbra). Neuronal populations remote from, but connected to, the injured core undergo a reduction in blood flow and metabolism. Spontaneous recovery (i.e., recovery in the absence of specific interventions) is thought to occur via resolution of diaschisis, employment of compensatory kinematic strategies (use of opposite limb or more-proximal muscles of the affected limb), and adaptive plasticity. Adaptive plasticity, or neuroplasticity, is a phenomenon based on several synaptic and cellular mechanisms, such as axonal and dendritic sprouting, synaptogenesis, long-term potentiation, long-term depression, and unmasking.

Organization of Primate Motor Cortex

Motor Cortex Defined

Classically, cortical motor function has been ascribed to that part of the cerebral cortex that is most excitable via electrical stimulation. That is, the primary motor cortex (M1) located in the precentral gyrus was originally defined on the basis of the ability to evoke contractions in skeletal muscles by applying electrical current either on the surface of the brain or deep within the cortical grey matter. The currents required for evoking such contractions are lowest in M1, and using electrical stimulation techniques, a complete representation of body parts can be demonstrated. Later, premotor or secondary motor areas were discovered in more anterior parts of the frontal cortex. These areas are also somatotopically arranged, but somewhat higher currents are required to elicit movements or muscle contractions.

Based on these physiological criteria, frontal cortex was classically seen as motor in function, while parietal cortex was seen as sensory in function. However, there is increasing evidence for a dynamic interaction between motor and sensory areas for learning new motor skills, performing goal-directed behaviors, or adapting old behavioral repertoires to new conditions. At least two cortical pathways have been identified for sensorimotor integration in the cortex. First, sensory information (primarily somatosensory, but also visual) originating in the parietal cortex innervates several motor areas in the frontal cortex. Sensory information also reaches the frontal cortex from the cerebellum and basal ganglia. These connections are important for fundamental behaviors such as sensorimotor integration for reaching and grasping.

Multiple Motor Areas Contributing to Motor Control in Primates

In the frontal cortex, several separate but interrelated motor areas have been identified in primates. The separation between M1 and primary somatosensory cortex (S1) marks the posterior limits of the frontal cortical area. In most primates, this area is buried in the depths of the central sulcus, that is, within the precentral gyrus. More anterior are several nonprimary motor cortices. The medial premotor cortex is referred to as the supplementary motor area (SMA), and the lateral premotor cortex is divided into the dorsal premotor area (PMd) and the ventral premotor area (PMv). In many primate species, these areas have been further subdivided into caudal and rostral components. Other motor areas in the frontal cortex include the cingulate motor areas within the medial wall of the cortex. A common characteristic of these motor areas is that each has neurons that project directly to the spinal cord (corticospinal neurons), influencing motor neuron output and hence muscle activity. Corticospinal neurons in many primate species have monosynaptic connections with motor neurons in the ventral horn of the spinal cord (corticomotor neuronal cells). Corticomotor neuronal cells contribute to hand dexterity, allowing for individual movement of the digits. This organization provides parallel influence on populations of motor neurons in the spinal cord from different motor areas. Therefore, there is not a strict serial control of movements, with M1 being the final common pathway for activation of motor neurons. All the motor areas contribute to voluntary control of movement within an integrated network from cortex to spinal cord.

There are several functional differences among these motor areas. First, M1 contains a complete somatotopic movement representation of major body parts organized from lower extremity on the medial cortical surface to upper extremity on the lateral surface of the cortex. A complete somatotopic map is not true of all motor areas. For example, PMv contains an upper extremity and face representation but no lower extremity representation. Most of the corticospinal projections originate in the frontal

cortex (~70%), and about 30–35% of these originate in M1. The rest of the corticospinal projections are distributed among the other nonprimary motor areas, as well as parietal areas. The vast majority of corticomotor neuronal cells originate in M1. These anatomical features in M1 indicate that this area has the most direct control of kinetic and kinematic aspects of movement control. Nonprimary motor areas are involved more in context-dependent organization of movements (both internal context related to self-initiated movements and environmental context related to spatial information and stimulus associations). In general, PMv mediates visual–motor and somatosensory–motor integration for motor control of the upper extremity. PMd mediates visuomotor association tasks, and SMA mediates motor sequences and bimanual motor control. Cingulate motor functions are similar to SMA and depend on self-initiated movements (i.e., they rely on motor cues).

The functional differences between M1 and nonprimary motor areas are not absolute. Functional properties of individual neurons overlap across M1 and nonprimary motor areas. It is this functional overlap between motor areas, in addition to shared intercortical connections and parallel corticospinal connections, that may allow for adaptive reorganization of motor control in the injured brain.

Divergence of Motor Cortex Outputs as a Basis for Dynamic Organization

There is also functional overlap within each motor area. Single projection neurons can influence multiple muscles. Also, clusters of cortical neurons that typically converge on the same motor neuron pool may overlap within the cortex with a cluster of neurons converging on a different motor neuron pool. Thus, stimulation within overlapping clusters of neurons produces movements of multiple joints (Figure 2). In addition to this divergent/convergent output pathway to the spinal cord, local circuits may contribute to plasticity in the cortical output neurons selected for a particular motor act. Plasticity of local, or intrinsic, circuitry is thought to be mediated by the well-known synaptic plasticity mechanisms, long-term potentiation and long-term depression.

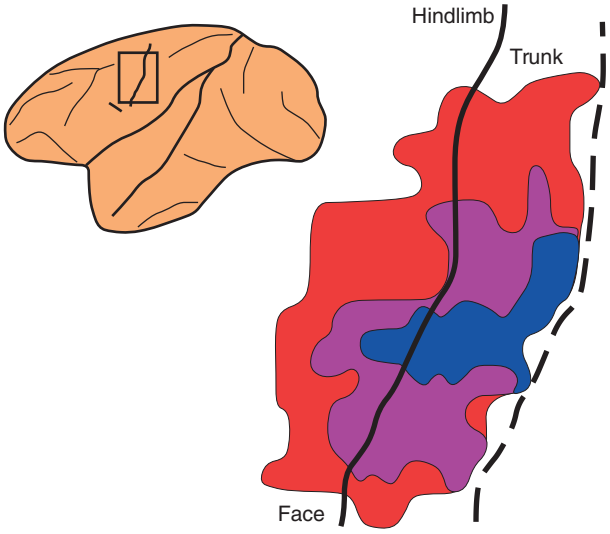

Figure 2 Overlapping of distal and proximal forelimb muscle representations in primary motor cortex of macaque monkeys. Inset depicts a lateral view of the macaque monkey cortex. The enlarged area depicts the results of a motor mapping procedure in which microelectrodes probed the depths of the precentral gyrus to determine what muscles were facilitated by electrical stimulation at each of dozens of sites. Electromyographic (EMG) activity was recorded simultaneously from 24 muscles in the proximal and distal forelimb. Typically, each corticomotor neuronal cell can influence up to four or five different motor neuron pools. The blue area denotes sites where stimulation evoked EMG activity in distal muscles only, and the red area denotes sites where stimulation evoked EMG activity in proximal muscles only. In the intervening zone (purple), stimulation evoked EMG activity in both distal and proximal muscles. Together, this data suggests that the organization of primary motor cortex is highly overlapping with respect to corticospinal outputs to motor neuron pools. Such divergence in basic descending anatomy may underlie the capacity for functional plasticity in this system during learning and after injury. Solid line indicates the convexity of the precentral gyrus, and the dotted line indicates the trough of the central sulcus. Figure redrawn from Park MC, Belhaj-Saif A, Gordon M, and Cheney PD (2001) Consistent features in the forelimb representation of primary motor cortex in rhesus macaques. *Journal of Neuroscience* 21: 2784–2792.

Physiological and Anatomical Reorganization in Motor Cortex

Microelectrode Stimulation Methods for Monitoring Motor Map Topography

The advent of electrophysiological techniques in the mid-nineteenth century provided a method to map movement representations within an intact cerebral cortex via surface stimulation. The resolution of these maps is coarse due to spread of current over-extensive areas of cortex. New electrophysiological techniques developed in the 1960s and 1970s required less current and thus provided higher-resolution motor maps. This popular approach is typically called intracortical microstimulation (ICMS) when it is applied to the cerebral cortex. Small-diameter stimulating microelectrodes are introduced into the deep layers of cortex, in the vicinity of the corticospinal neuron cell bodies. Employed in either awake or anesthetized animals, low-current pulse trains stimulate relatively small clusters of neurons, resulting in contraction of skeletal musculature (Figure 2). In addition to

stimulation-based techniques, the functional organization of motor cortex can be explored by recording a neuron's activity while an awake animal engages in a behavioral task. The relationship between neuronal firing patterns and kinetic and kinematic aspects of movement, as well as the distribution of electromyographic activity, has provided important insights into motor cortex function. These methods have contributed complementary information about functional organization of the motor cortex and are capable of detecting functional changes in the cortex as behavioral demands are varied.

Role of Postinfarct Behavioral Experience on Motor Map Plasticity after Stroke

A series of ICMS experiments beginning in the 1990s have characterized several fundamental aspects of cortical plasticity phenomena associated with recovery of function. These experiments have mainly used a nonhuman primate ischemic infarct model, but rat models have also been employed. Squirrel monkeys (genus *Saimiri*) were used because their lissencephalic cortex provides an optimal substrate for deriving highly detailed functional maps. In these studies, reach and grasp behavior was assessed for its clinical relevance. Squirrel monkeys were trained to retrieve small food pellets from various sizes of food wells drilled into a Plexiglas board (modified Klüver board). ICMS maps of the forelimb were derived to determine the spatial organization of motor cortex before and after either motor skill training or subtotal ischemic damage in the cortical motor regions. Then the effects of postinfarct training to relearn the reach and retrieval task (rehabilitative intervention) were examined.

The first studies explored cortical plasticity within M1 of normal (uninjured) monkeys and provided a basic description of fundamental properties of cortical plasticity relative to the distal and proximal forelimb representation in M1. Functional reorganization of motor maps was found to be dependent on prior behavioral experience. Cortical maps remain stable from one mapping session to the next unless the monkeys are required to learn a new motor skill. Reach and retrieval training on the Klüver board results in an increase in the representation of the digits in M1 and a relative decrease in the representation of the wrist and forearm. If the task is changed to require the monkeys to turn a handle to receive a food pellet, shifting the behavioral demand from digit action to wrist action, wrist representations expand at the expense of digit representations. Further, multijoint movement representations appear after training, reflecting joint combinations and sequences used in the actual task. Multijoint representations thus are driven by temporally correlated activation of two or more movements during training and may reflect the development of muscle and joint synergies in motor cortex. Rat studies have shown that training on similar tasks increases the number of synapses per neuron in motor cortex. These behaviorally driven changes in motor maps appear to be skill or learning dependent, as training animals on a task requiring no additional learning or skill acquisition (e.g., monkeys retrieving pellets from a large well) results in no changes in either motor maps (shown in monkeys and rats) or synaptic density (rats).

Following a subtotal ischemic infarct within M1, monkeys recover the previously learned reach and retrieval skills without any specific behavioral intervention (spontaneous recovery). However, this recovery is associated with a shift in functional borders between distal and proximal forelimb representations, resulting in a further loss of digit, wrist, and forearm representations as proximal (elbow and shoulder) representations expand. It is possible that this seemingly maladaptive expansion is due to compensatory use of more-proximal muscles to accomplish the task. For example, in human stroke survivors, the trunk is used in a compensatory manner in retrieving objects. As the distal limb has less range of motion, the proximal muscles aid in propelling the limb forward. Thus, while individuals can become functional after cortical injuries, the tasks are accomplished by very different kinematic patterns accompanied by altered cortical motor organization.

Rehabilitative training beginning within the first week after the ischemic infarct alters the pattern of cortical reorganization, resulting in maintenance of distal forelimb representations spared by the initial infarct and in some cases in an expansion of distal forelimb representations into the proximal forelimb area (Figure 3). Delaying rehabilitative training by 1 month constrains functional reorganization such that no appreciable recovery of spared distal forelimb representation is maintained, similar to what is seen after spontaneous recovery. However, there is an increase in the representation of atypical multijoint movements in the spared distal forelimb area. Most of these multijoint movements involve distal and proximal movements (such as simultaneous wrist and shoulder movements). Results from these initial studies demonstrate the advantages of rehabilitative training for postinfarct recovery of function on a cortical level, distinguishing it from recovery that may occur without directed intervention. These results also suggest that such poststroke rehabilitative training should be introduced early after the ischemic event to maximize the effect on cortical reorganization.

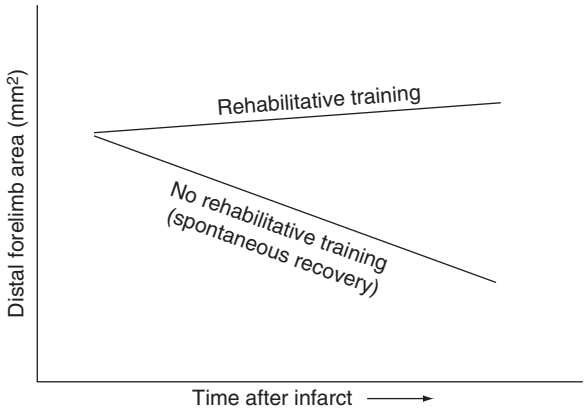

Figure 3 Effects of behavioral experience on spared distal forelimb representations after a subtotal ischemic lesion in primary motor cortex (M1). Nonhuman primate studies have demonstrated that after a small ischemic lesion in M1 affecting less than 50% of the M1 distal forelimb (hand) representation, the remaining hand representation is altered. However, if monkeys recover spontaneously, that is, without any rehabilitative training, the remaining spared distal forelimb representation in M1 shrinks further. However, if monkeys undergo repetitive training with the impaired forelimb on reach and retrieval tasks requiring increasing levels of manual skill, the spared distal forelimb representation is maintained. This result emphasizes the importance of rehabilitative training in recovery after stroke. Later studies demonstrated that the benefit of training on postinfarct maps in M1 occurs only when the training is introduced early after injury.

Physiological and Anatomical Plasticity in Remote Motor Areas after Stroke

Nonprimary motor areas have properties of cortical plasticity similar to those observed in M1. For example, in PMv, the general topography of the proximal and distal forelimb is similar: There is a fractionated map of digit, wrist, and forearm representations surrounded by proximal limb representations. After M1 infarction, PMv forelimb representations are altered, even though the forelimb escapes the direct impact of the ischemic event. More specifically, there is a positive linear relationship between expansion of the distal forelimb area in PMv and the size of M1 infarction. In all cases in which at least 50% of M1 distal forelimb area was damaged, digit, wrist, and forearm representations expanded. This may be an example of vicarious substitution in that it is possible that to some extent PMv assumed M1's role in skilled hand behaviors. It is possible that functional plasticity associated with recovery of motor function is distributed across all motor-related areas in the cortex.

Recently, anatomical evidence consistent with the vicarious substitution hypothesis regarding PMv has been shown in squirrel monkeys. In primates, M1 receives cutaneous and proprioceptive somatosensory information related to the distal forelimb from portions of S1. Normally, there are virtually no connections between S1 and PMv. However, several months after a large M1 infarct, novel projections are found between PMv and S1 areas 1 and 2. This suggests that somatosensory information from the forelimb may be reestablished within the motor network through PMv. This study represents the first evidence of a major alteration of intracortical wiring patterns between different cortical fields in any adult mammal. Earlier studies in rats demonstrated anatomical changes within the peri-infarct region after motor cortex injury, such as increase in neuropil, synaptogenesis, and local axonal sprouting over about 1–2 mm. Such anatomical reorganization may be much more extensive than previously thought.

Other evidence from rats demonstrates that unilateral cortical damage has substantial impact on the anatomy of remote areas, especially in the contralateral (uninjured) motor cortex. A time-dependent increase in dendritic branching occurs in the contralateral cortex, followed by pruning and synaptogenesis. At least some of this anatomical change may be related to increased use of the less-impaired forelimb for manipulation, locomotion, and postural support. In addition, the uninjured motor cortex sprouts novel connections with subcortical structures. Normally, the motor cortex on one side sends axons to the striatum of the same side only. However, after unilateral injury, the intact motor cortex sends projections across the midline to terminate in the contralateral striatum.

Plasticity in Cerebral Cortex Associated with Stroke in Humans

Differences between Animal Models and Clinical Stroke

Animal models provide important information regarding mechanisms underlying postinjury recovery. Their value lies in the ability to control various parameters of the stroke event, such as lesion size and location, temperature, and behavioral conditions. One can monitor the organization of brain activity before and after the ischemic insult, allowing specific hypotheses to be tested empirically. However, these distinct advantages must be considered along with some notable caveats regarding the translation of information from animal models to the understanding of human stroke. Most often, animal studies of stroke are conducted in young, healthy specimens as they are most feasible for *in vivo* analysis of physiology, behavior, and so forth. However, human stroke frequently occurs in frail, older individuals with a host of comorbid conditions, such as high blood pressure, diabetes, and so forth. In fact, stroke prevalence doubles by decade

after about age 50. Following stroke, a significant proportion of individuals experience clinical depression, which may also alter the course of recovery. Thus, while animal studies are invaluable to our understanding, they paint an incomplete picture of clinical stroke and its aftermath.

Nevertheless, certain fundamental traits are common between experiment stroke in animals and clinical stroke in humans. For example, humans undergo substantial recovery after stroke, regardless of the presence or absence of rehabilitative interventions. This phenomenon is known as spontaneous recovery and is most rapid during the first month poststroke. However, it is likely to continue for several months. Specific types of poststroke treatment can enhance recovery. One popular approach, known as constraint- induced movement therapy, constrains the good hand in a sling or mitten and encourages frequent use of the more impaired hand.

Methods for Mapping Brain Reorganization in Humans

Two approaches are commonly used for mapping brain plasticity after stroke in humans. The first, fMRI, depends on magnetic susceptibility of certain substances in brain tissue, such as hemoglobin in blood. Patients are placed in a high magnetic field, and images of relative magnetic susceptibility can be derived from thin, sequential 'slices' through the brain. Deoxygenation of hemoglobin, which occurs when neurons are physiologically active, produces a signal change in the magnetic field. Thus, the fMRI images represent a correlate of neural activity in the brain. A second noninvasive method, TMS, relies on the ability of rapidly changing magnetic fields to induce a weak electrical current in brain tissue. By the use of a magnetic field focused on a relatively small part of the brain by means of a handheld wand, a map of motor outputs from the cerebral cortex can be derived, much like those described via invasive experiments in animals. However, the spatial resolution of TMS is at best an order of magnitude more crude than the invasive procedures used in animals. Both of these techniques are relatively safe, even in clinical populations, and thus have been used in numerous human stroke studies to determine patterns of reorganization.

Cortical Reorganization in Humans after Stroke

On the basis of TMS and fMRI mapping studies after stroke, some conclusions may be drawn. First, as in animal studies, human brains undergo substantial plasticity after stroke. As recovery proceeds, enhanced activity is often seen in the spared tissue adjacent to the stroke injury. Second, additional activity is seen in remote cortical areas on the injured side of the brain, as well as on the uninjured side. The major question that still remains is whether such changes in activation patterns are adaptive, maladaptive, or epiphenomenal.

There has been considerable interest in determining whether contralesional activation after stroke might be adaptive. It is well known from developmental studies in animals that unilateral damage to the cortex results in atypical ipsilateral corticospinal projections in the adult. Presumably, early in development, exuberant projections exist such that corticospinal innervation of the cord is bilateral. During development, the ipsilateral connection is largely pruned back. The unilateral injury at a critical developmental stage may prevent the pruning. Whether unilateral damage results in significant sprouting of descending pathways in adults is still subject to some debate. However, it is important that as stroke recovery proceeds, the contralesional activity seen in fMRI appears to wane as more ipsilesional activity waxes. Further, when stroke patients are trained to use their impaired hand, improved performance is correlated with enhanced ipsilesional activity. Therefore, information to date would suggest that poststroke contralesional activity is not critical for recovery and may possibly be maladaptive. However, this issue is still somewhat unsettled, especially with regard to very large infarcts, in which ipsilesional motor areas are largely eliminated.

New Therapies for Poststroke Plasticity and Recovery

As theories regarding neuroplasticity after stroke have matured, several innovative therapeutic approaches have been developed that purportedly modulate these processes. Central to most of these therapies is the potent influence that behavioral experience has on physiological and anatomical organization in the injured brain. The effectiveness of constraint-induced movement therapy has been demonstrated in a recent phase-III clinical trial. Several adjuvant therapies that are used in conjunction with behavioral training are now being explored. There is a large literature in animal research demonstrating the interactive effect of d-amphetamine and behavioral experience after brain injury. The combined treatment appears to be more effective than either treatment alone. Clinical trials are now investigating its use in human stroke. Further, a new approach combining epidural electrical stimulation and behavioral training has shown promise in both animal models of stroke and early clinical trials in humans. While the mechanisms underlying

these adjuvant therapies are not well known, it is likely that they act by establishing a more receptive environment for behavioral interventions to have their maximal effect on neuroplasticity mechanisms.

See also: Animal Models of Stroke; Autonomic Neuroplasticity and Regeneration; Brain Injury: Functional Recovery After; Cortical Plasticity and Learning: Mechanisms and Models; Corticomotoneuronal System; Corticospinal Development; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Neuronal Plasticity after Cortical Damage; Stem Cells and CNS Repair; Stroke; Stroke: Injury Mechanisms; Synaptic Mechanisms of Learning.

Further Reading

Carmichael ST (2006) Cellular and molecular mechanisms of neural repair after stroke: Making waves. *Annals of Neurology* 59: 735–742.

Cramer SC (2004) Functional imaging in stroke recovery. *Stroke* 35: 2695–2698.

Dancause N, Barbay S, Frost SB, et al. (2005) Extensive cortical rewiring after brain injury. *Journal of Neuroscience* 25: 10167–10179.

Jones TA and Schallert T (1992) Overgrowth and pruning of dendrites in adult rats recovering from neocortical damage. *Brain Research* 581: 156–160.

Nudo RJ, Wise BM, SiFuentes F, and Milliken GW (1996) Neural substrates for the effects of rehabilitative training on motor recovery after ischemic infarct. *Science* 272: 1791–1794.

Nudo RJ, Larson D, Plautz EJ, et al. (2003) A squirrel monkey model of poststroke motor recovery. *ILAR Journal* 44: 161–174.

Park MC, Belhaj-Saif A, Gordon M, and Cheney PD (2001) Consistent features in the forelimb representation of primary motor cortex in rhesus macaques. *Journal of Neuroscience* 21: 2784–2792.

Rathelot JA and Strick PL (2006) Muscle representation in the macaque motor cortex: An anatomical perspective. *Proceedings of the National Academy of Sciences of the United States of America* 103: 8257–8262.

Ward NS (2006) The neural substrates of motor recovery after focal damage to the central nervous system. *Archives of Physical Medicine and Rehabilitation* 87: S30–S35.

Wolf SL, Winstein CJ, Miller JP, et al. (2006) Effect of constraint-induced movement therapy on upper extremity function 3 to 9 months after stroke: The EXCITE randomized clinical trial. *Journal of the American Medical Association* 296: 2095–2104.

Relevant Websites

http://www.strokeassociation.org – American Stroke Association.
http://www.ninds.nih.gov – National Institute of Neurological Orders and Stroke.
http://www.stroke.org – National Stroke Association.

Marsupial Neocortex

S J Karlen and L Krubitzer, University of California at Davis, Davis, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Although most of us are familiar with only a few species of marsupials, such as kangaroos, koalas, and opossums, there are actually hundreds of marsupial species. Like placental mammals, they are an extremely diverse group that has a number of behavioral specializations related to individual lifestyles. For example, there are marsupials that are terrestrial, arboreal, burrowing, and aquatic. Marsupials can be highly social, such as the striped possum, or solitary, such as the koala; they can be carnivorous, herbivorous, or omnivorous. This diversification is tied to the evolution of unique peripheral morphologies and highly specialized behaviors, both of which are related to alterations in overall brain size, neocortical organization, and connectivity.

In this article, we begin by examining the general lifestyle and morphological characteristics of all marsupials. We then describe the organization of somatosensory, visual, auditory, and motor cortex in marsupials. Finally, we discuss how alterations in the basic plan of organization are associated with specialized behaviors and individual lifestyles.

General Features of Marsupials

Marsupials represent one of three subclasses of mammals that include Prototherians (monotremes), Metatherians (marsupials), and Eutherians (placentals). Monotremes diverged early in evolution, and evidence from gene sequencing studies indicates that marsupials arose concurrently with placental mammals, approximately 180 Ma (Figure 1). The first marsupials are thought to have arisen in North or South America and subsequently migrated to Australia and Southeast Asia, possibly through Antarctica; however, an alternative theory suggests that marsupials may have arisen in Laurasia, somewhere in present-day Asia, and migrated to North America.

Extant marsupials have been subdivided into seven different orders, and more than 270 species have been identified (Figure 1). The exact phylogenetic relationship between marsupial orders is unresolved because different types of data result in conflicting phylogenetic relationships. In general, Didelphimorphia are considered to be the oldest order. This group is relatively homogeneous, and all species within this order have relatively small brains with few cortical fields. The orders Paucituberculata and Microbiotheria are also found in the Americas, but very little is known about the organization of their neocortex. The remaining four orders are found in Australia, Tasmania, New Guinea, and neighboring islands, and they include most of the well-known marsupials.

Regardless of phylogeny, there are a number of features that all marsupials share. The most noted is the marsupium, or pouch, from which they derive their name. The marsupium varies between species from a closed pouch to a small, primal pouch or no pouch at all. Marsupials have a short gestation, between 8 and 42 days, and are born in an immature state. When born, they must make their way to the marsupium and attach to the mother's nipple in order to survive (Figure 2). Once in the pouch, the young animals reside there for the remainder of gestation.

Marsupial brains vary dramatically in size, shape, smoothness of the neocortical surface, and encephalization (the expected size of the brain relative to body size). For example, some marsupials have small, smooth brains with a small proportion of the brain devoted to the neocortex, whereas others have relatively larger, fissured brains with a larger proportion of the brain devoted to the neocortex (Figure 3). In terms of brain organization, perhaps the most notable feature of the forebrain is that marsupials (and monotremes) lack a corpus callosum. Instead, most interhemispheric communication occurs through the anterior commissure, which has been shown to be functionally equivalent to the corpus callosum in placental mammals.

Finally, all but three species of marsupials are nocturnal or crepuscular. This aspect of lifestyle is particularly interesting since all monotremes are nocturnal, suggesting that the common mammalian ancestor was nocturnal.

Somatosensory Cortex

The organization of somatosensory cortex has been studied in several species, all of which have a primary somatosensory area (S1) and a second somatosensory area (S2) located caudolateral to S1. The functional organization of S1 has been described using electrophysiological mapping techniques. In all species examined, S1 contains a complete and inverted representation of the contralateral body surface with the tail represented most medially, followed by the representation of the hindlimb, trunk, forelimb, forepaw, face, and oral structures in a medial to lateral progression. Most neurons in S1 respond to cutaneous

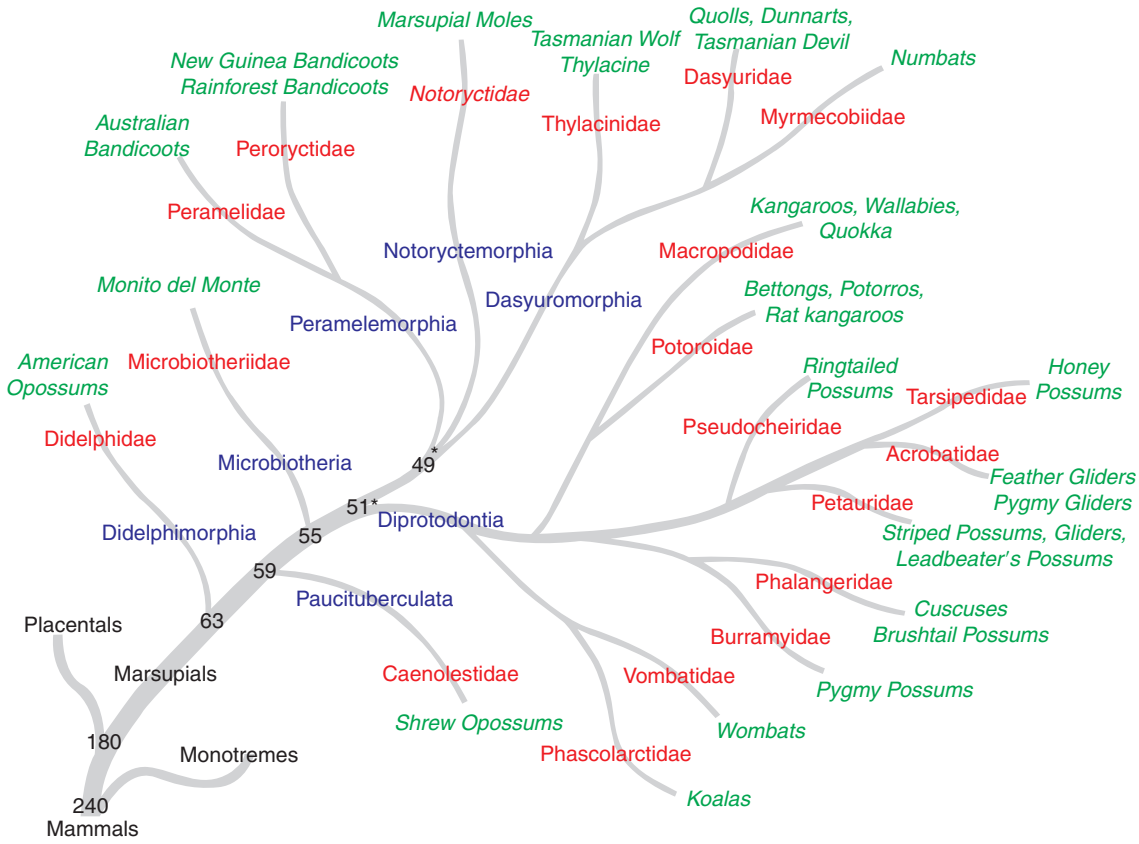

Figure 1 A phylogenetic tree illustrating the relationship between marsupials and other mammals. The seven extant marsupial orders are shown in blue, names of the families within each order are given in red, and examples of well-known species are in green. The timescale is in millions of years ago and is based on DNA sequence data from a 2004 study by Murphy et al. Asterisks indicate branch points that are ambiguous.

Figure 2 Marsupials are born at a very immature state, and they must make their way to the marsupium, or pouch, and attach to the mother's nipple in order to survive. (a) A litter of short-tailed opossums, a pouchless marsupial, on the day of birth. The immaturity of the young is evident in the embryonic developmental state of most of their body features and sensory organs (b). They are hairless, their eyes and ears are poorly developed, the hindlimbs and tail are stumps, and bone development is extremely immature. Only their olfactory system, their forelimbs, and, to a lesser extent, portions of the somatosensory system are developed at birth.

stimulation, and receptive fields for neurons are generally small, particularly on specialized morphological structures such as the vibrissae or forepaws.

In many instances, the specialized morphological structures are associated with unique behaviors. For example, the structure of the forepaw can vary dramatically between species from the spade-shaped forepaw of the marsupial mole, used for digging, to the specially elongated fourth digit of the striped possum used to extract insects while foraging. The cortical representation of these specialized body parts in marsupials is generally enlarged in S1. This is called cortical magnification and is found in placental and monotreme mammals as well.

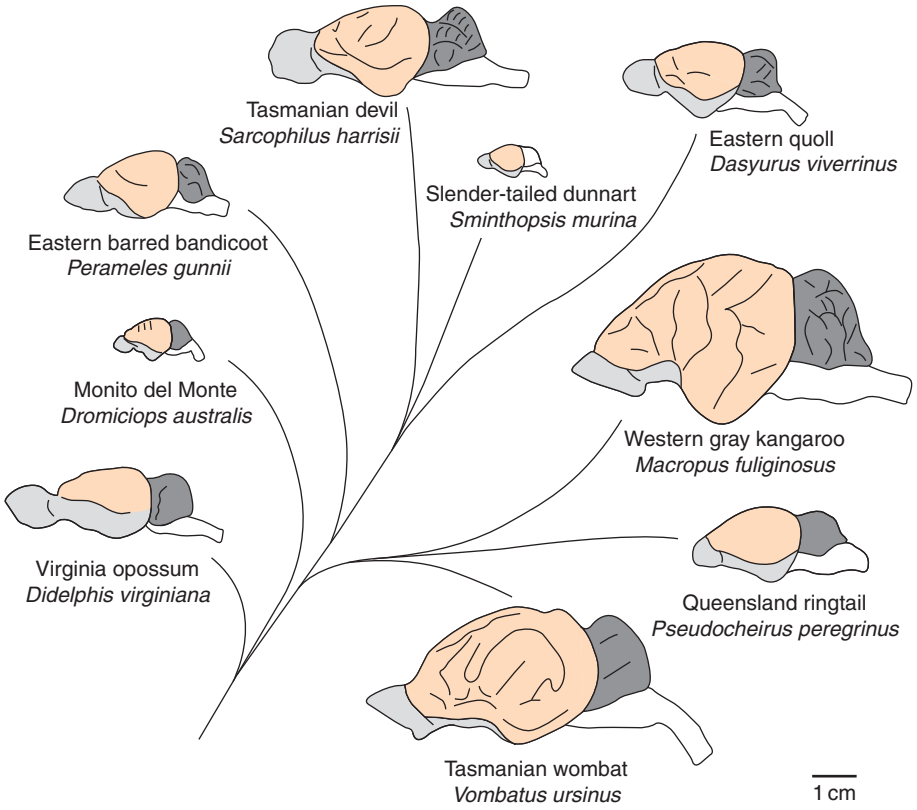

Figure 3 Lateral views of nine representative marsupial brains. Marsupial brains vary dramatically in their overall size; compare the brain of the Western gray kangaroo with that of the slender-tailed dunnart. The relative size of distinct brain areas also differs between species, such as the size of the olfactory bulb and pyriform cortex (gray), the size of the neocortex (light orange), and the size of the cerebellum (dark gray). Finally, the smoothness of the neocortical surface varies between species, such as the lissencephalic neocortex of the Virginia opossum versus the highly fissured neocortex of the Tasmanian wombat. Lines denote phylogenetic relationships between species; rostral is left and medial is up.

Caudolateral to S1 lies a second, smaller representation of the contralateral body surface called S2. This area has been electrophysiologically identified in all marsupials examined. As in placental mammals, S2 in marsupials contains a complete, noninverted representation of the contralateral body surface with the representation of the face and oral structures located adjacent to similar representations in S1. The forelimb and hindlimb representations are located progressively caudal, and the trunk and proximal body parts are represented medially. The receptive fields for neurons in S2 are larger than those for neurons in S1 and are often responsive to bilateral somatosensory stimulation and/or auditory stimulation. In many species, the S2 region actually contains two separate fields, a rostral field called the parietal ventral area (PV) and a caudal field termed S2. Both fields are somatotopically organized, and neurons in each field have large receptive fields.

Two other fields have been identified in the parietal cortex of several marsupial species; one field is rostral to S1, called the somatosensory rostral field (SR), and the other is caudal to S1, called the somatosensory caudal field (SC). Neurons in both of these fields respond to stimulation of deep receptors or high-threshold cutaneous receptors. Because of similarities in location, stimulus preference, and patterns of connectivity, we have proposed that these areas correspond to fields that have been identified in other mammals, specifically that SR is homologous to area 3a of primates and flying foxes and that SC is homologous to the posterior parietal area (PP) in squirrels and posterior parietal cortex in rats. However, this hypothesis has not been tested.

Studies of corticocortical and thalamocortical connections of somatosensory cortex in marsupials are relatively restricted to the connections of S1. Corticocortical connections have been examined in the brush-tailed possum, Virginia opossum, white-eared opossum, short-tailed opossum, striped possum, and Northern quoll. In all animals, S1 has dense corticocortical connections with S2, PV, SR, SC, cortex immediately lateral to S1, and, in the brush-tailed possum only, motor cortex. S1 has sparse connections with cortex of the medial wall. Interhemispheric connections are restricted to S1, PV, and cortex lateral

to S1. Thalamocortical connections have been examined in the Virginia opossum, the brush-tailed possum, the Northern quoll, and the Tammar wallaby. In the Virginia opossum, thalamic connections with S1 arise predominantly from the ventroposterior nucleus (VP). Moderate projections are also from the ventrolateral nucleus (VL), the central intralaminar nucleus (CIN), the central lateral nucleus (CL), and the ventromedial nucleus (VM). Similar patterns of thalamocortical connectivity have been observed after injections into parietofrontal cortex (which contains S1) in the brush-tailed possum and the Northern quoll.

Taken together, the electrophysiological recording data and architectonic analysis indicate that two unimodal somatosensory fields are present in the neocortex of marsupials, S1 and S2. In some species, PV is clearly present as well. An additional field, SR, is associated with the somatosensory system but is also likely associated with motor processing. Likewise, the field caudal to S1, SC, is associated with somatosensory processing but may be homologous to one or more fields in the posterior parietal cortex of primates that are involved in sensorimotor integration.

Visual Cortex

The organization of the primary visual area (V1) has been examined using electrophysiological recording techniques in several different marsupials, including the Virginia opossum, the white-eared opossum, the big-eared opossum, the brush-tailed possum, the Tammar wallaby, the Northern quoll, and the short-tailed opossum. As in other mammals, V1 is located at the caudal pole of the occipital lobe and contains a complete visuotopically organized representation of the contralateral visual field. The representation of the upper visual field is located caudolaterally in cortex, the lower visual field representation is located rostromedially, the horizontal meridian bisects the upper and lower visual field representations, and the vertical meridian forms the rostrolateral boundary of this field. Several studies have demonstrated that cortex immediately rostral to V1 contains neurons responsive to visual stimulation, and this area, called the second visual area (V2), has been mapped in detail in the Northern quoll. As in placental mammals, V2 contains a complete representation of the visual hemifield, with the vertical meridian represented at the caudomedial border of the field, adjacent to the vertical meridian border of V1. The horizontal meridian bisects the field and forms the rostrolateral boundary of the field. The lower visual quadrant is represented rostromedially and the upper visual quadrant is represented caudolaterally.

In studies in which electrophysiological recordings were combined with architectonic analysis, V1 has been shown to be architectonically distinct from V2, and its appearance is similar to that described for V1 in placental mammals. In Northern quolls, V2 stains lightly for both myelin and cytochrome oxidase (CO), and in favorable preparations V2 is actually composed of heterogeneous light and dark patches in CO-stained tissue.

The cortical connections of V1 are similar in all species examined. For example, in the Virginia opossum, brush-tailed possum, mouse opossum, Tammar wallaby, big-eared opossum, and short-tailed opossum, V1 is densely connected with V2 (peristriate cortex), cortex just lateral to V2 in multimodal cortex (parietal cortex and PP), the caudotemporal area (CT), posterolateral peristriate cortex, and perirhinal cortex. Furthermore, the connections of V1 are not restricted to the representation of the vertical meridian, as is the case for a number of placental mammals. Instead, connections of V1 are found throughout V2 and CT of the ipsilateral hemisphere, although the connections with V2 are patchy and appear to be related to heterogeneities identified using CO stains. Studies of the overall pattern of commissural connections also demonstrate a patchy distribution with V1 and V2 of the opposite hemisphere, indicating that V2 may be modularly organized as it is in primates and other mammals.

Thalamocortical connections of V1 in marsupials are similar to those described in a number of placental mammals. In the brush-tailed possum, the Virginia opossum, the big-eared opossum, the Tammar wallaby, and the short-tailed opossum, the primary source of input to V1 is from the dorsal division of the lateral geniculate nucleus (LGd), with moderate inputs arising from the lateral posterior nucleus (LP). In the brush-tailed possum, the primary source of input to V2 is from LP, with sparse input from LGd. In the Virginia opossum, the lateral intermediate nucleus also projects sparsely to V2.

Taken together, the data indicate that visual cortex in marsupials contains at least two cortical areas, V1 and V2. Each field is topographically organized similar to that in placental mammals. Furthermore, features of neural response properties, histochemical appearance, cortical and subcortical connections, and even the presence of a modular organization are common for visual cortex across all mammals.

Auditory Cortex

Little is known about the number of auditory areas or the detailed organization of auditory cortex across marsupials. Most of what we do know comes from

work on two species, the Northern quoll (previously called the native cat) and the brush-tailed possum. In the Northern quoll, detailed maps of the primary auditory area (A1) have been generated, and these maps are tonotopically organized, as in placental mammals. Specifically, isofrequency bands run mediolateral in auditory cortex, with high frequencies represented rostrally and low frequencies represented caudally. In these animals, there is a disproportionate representation of high frequencies, and neurons in this high-frequency representation have very low thresholds. Studies of adult and juvenile Northern quoll vocalizations indicate that most adult vocalizations occur at the lower end of the hearing range, between 0.4 and 2 kHz, but that juveniles have calls within the peak frequencies of adult hearing, between 6 and 12 kHz. Furthermore, vocalizations and other sounds made by predators and prey of this species have a peak energy between 2 and 5 kHz and greater than 10 kHz, suggesting that the cortical magnification of high frequencies in the Northern quoll is important for predator and prey localization. Another auditory field just rostral to A1 was identified in the Northern quoll, but the data on this field were relatively sparse.

In the brush-tailed possum, the orientation of A1 is different than in other mammals in that high frequencies are represented dorsally and low frequencies are represented ventrally. However, as in the Northern quoll, the lowest thresholds for neurons in A1 in the brush-tailed possum fall between 12 and 21 kHz, similar to the range of frequencies used in vocalizations of the Virginia opossum.

Thalamocortical and corticocortical connections of auditory cortex have been examined in the Virginia opossum, the brush-tailed possum, and the Northern quoll. In all of these marsupials, the primary input to A1 is from the medial geniculate nucleus (MGN). Sparse projections also arise from the suprageniculate nucleus. Since injections into different best frequency representations in the cortex produced discrete bands of label in MGN, the results indicate that projections from MGN to A1 are organized tonotopically.

Although ipsilateral cortical connections of A1 have not been described for any marsupial, dense projections from homotopic locations in A1 in the hemisphere contralateral to that injected have been described for both the brush-tailed possum and the Northern quoll. Furthermore, in the Northern quoll, A1 has been reported to connect ipsilaterally with the putamen and bilaterally with the lateral amygdala. Direct connections between the amygdala and A1 have not been reported for any placental mammal and may be a specialization of the Northern quoll or may represent a primitive pathway of mammals that was subsequently lost in placental mammals with the addition of more auditory cortical fields. Support for the latter supposition comes from studies in marmoset monkeys which indicate that higher order auditory areas, such as the rostromedial field, have ipsilateral connections with the lateral amygdala.

Motor Cortex

The state of motor cortex in marsupials is contentious. The earliest experiments used gross electrical stimulation with unipolar and bipolar electrodes and demonstrated that motor cortex is located at the rostral pole of the neocortex, usually around the orbital sulcus. These experiments found that movements of individual body areas, such as the face, vibrissa, and forepaws, are represented progressively across the cortical surface; however, only half of the studies were able to find a clear representation of body areas posterior to the forepaws (i.e., the hindpaws and tail). Thus, for many years there was a debate as to whether marsupials had a motor representation of the lower body.

As techniques improved, several studies tried to address this issue. In 1939, Golby demonstrated the existence of a complete motor representation of the body at the rostral pole of the neocortex in the brush-tailed possum. Architectonic examination of cortex in this stimulated area confirmed that it corresponded to primary motor area (M1), as described in monotremes and placental mammals. Concurrently, Abbie demonstrated that body movements could be evoked at the far rostral pole of cortex in six different species of marsupials – the long-nosed bandicoot, the Tasmanian devil, the Eastern quoll, the brush-tailed possum, the red-necked wallaby, and the agile wallaby. His results demonstrated that motor cortex was topographically organized in all six species, although in the long-nosed bandicoot, the Tasmanian devil, and the Eastern quoll, the neocortex did not include a motor representation of the lower body. He concluded that there was probably no cortical motor representation of the lower body in the orders Peramelemorphia and Dasyuromorphia, but that there was a complete representation in the order Diprotodontia. Furthermore, he observed that the excitable cortex in all six species was coextensive with a larger layer IV, indicative of sensory cortex, which suggested that motor and somatosensory cortex overlapped.

In the early 1960s, Lende used more refined methods of cortical stimulation to explore sensory and motor cortex in the Virginia opossum and the wallaby. His results confirmed the presence of hindlimb and tail motor representation on the medial wall of the neocortex in both species (orders Didelphimorphia and Diprotodontia, respectively). Unlike earlier

marsupial studies, Lende demonstrated that these animals had a complete overlap of S1 and M1, which he called a sensory–motor amalgam. At the time, the results from his study were surprising but seemed to be in agreement with earlier data. These findings spurred a number of important ideas regarding cortical field evolution in mammals. Most notably, Lende proposed that marsupials represent a primitive state of cortical organization in which S1 and M1 are completely overlapping and that the evolution of this region is marked by a progressive separation of these fields in placental mammals, ultimately into two complete and separate sensory (S1) and motor (M1) representations (**Figure 4(a)**).

Early observations by Lende have been supported by several subsequent cortical stimulation studies in marsupials, including the brush-tailed possum, the white-eared opossum, and the short-tailed opossum. However, in the short-tailed opossum only movements of the jaw and vibrissae could be evoked in this sensory–motor amalgam. The finding that no other body part movements (particularly forepaw movements) could be evoked anywhere in the motor cortex of the short-tailed opossum is surprising in light of previous studies that demonstrated that dense corticospinal projections originate in S1.

Lende's hypothesis of a complete sensory–motor amalgam is not supported by all marsupial motor studies. In particular, in some marsupials there is a region of cortex located anterior to the proposed sensory–motor amalgam that does not respond to somatosensory stimulation and does not have connections with VP. Since these marsupials are in the order Diprotodontia, Haight and Neylon proposed that this order does not have a complete overlap of sensory and motor areas. Instead, they divided mammalian sensory motor organization into three types: (1) in placental mammals, S1 and M1 are completely separate and the body representations are mirror images of one another; (2) in marsupials and placental edentates, there is a complete overlap of S1 and M1, both with the same orientation; and (3) in monotremes, there is dual motor representation with one being coextensive with S1 and possessing the same orientation and the other occupying an anterior, mirrored position (**Figure 4(b)**). The partially separated S1/M1 in Diprotodontia and the fully separated S1 and M1 in placental mammals may represent evolutionary convergence.

Like cortical stimulation studies, neuroanatomical studies indicate that S1 and M1 in marsupials are partially overlapping, but the data are equivocal with respect to the completely overlapping sensory–motor amalgam. For example, in the Virginia opossum, projections from VL (associated with the motor system), VP (associated with the somatosensory system), and CIN (associated with convergent somatic and cerebellar input) all converge on S1. Furthermore, the projections of VL are restricted to S1, supporting the idea of a completely overlapping sensory–motor amalgam. On the other hand, in the brush-tailed possum, thalamocortical projections appear to fall into three distinct zones. Cortex at the far rostral pole receives inputs from the mediodorsal

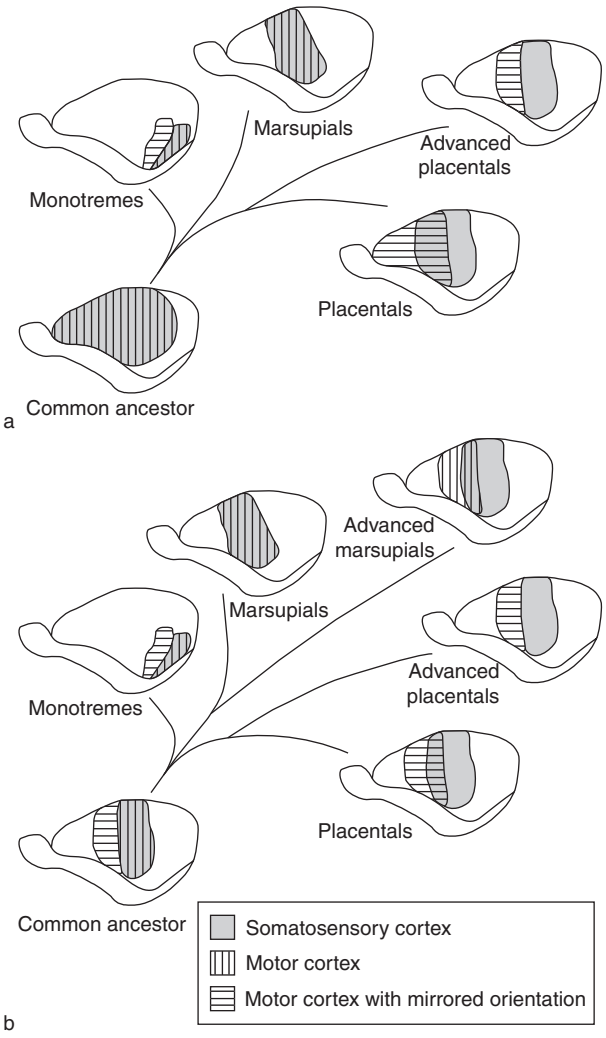

Figure 4 A diagram of the two prominent hypotheses on the evolution of sensory–motor cortex in mammals. (a) In 1969, Lende proposed that marsupials represent a primitive state of cortical organization in which S1 and M1 are completely overlapping and that the evolution of this region is marked by a progressive separation of these fields in placental mammals, ultimately into two complete and separate sensory (S1) and motor (M1) representations. (b) In 1979, Haight and Neylon proposed that there are three types of sensory motor organization. In monotremes (and potentially the common ancestor), there is dual motor representation with one being coextensive with S1 and possessing the same orientation and the other occupying an anterior, mirrored position. In placental mammals, S1 and M1 are completely separate and the body representations are mirror images of one another. In marsupials and placental edentates, there is a complete overlap of S1 and M1, both with the same orientation.

nucleus, generally associated with frontal cortex in placental mammals. Agranular cortex just caudal to this frontal region and rostral to S1, in the location of M1 in placentals mammals, receives input predominantly from VL and less dense inputs from VP. Projections to S1, as described previously, are predominantly from VP and the posterior nucleus and sparsely from VL. Furthermore, comparative studies in which retrograde tracers were injected into the spinal cord of a number of different mammals indicate that the major source of spinal input from the cortex is similar in all mammals investigated, including several marsupial species. Specifically, there are two or three corticospinal projection zones (termed A, B, and C) that all mammals have, at least in part. One of these projection zones (A) appears to correspond to somatosensory and motor cortex. Although no direct correlations were made, projection zone A in the Virginia opossum and the short-tailed opossum appears to be restricted to somatosensory cortex, whereas this same projection zone in rodents and primates has a rostral extension into the location of motor cortex.

When considered together, results from cortical stimulation and neuroanatomical studies in marsupials are somewhat conflicting. Early cortical stimulation studies in marsupials indicate that motor cortex is at most partially overlapping with somatosensory cortex, whereas later studies indicate that motor cortex is completely overlapping with somatosensory cortex, but only in certain species. These differences could be due to techniques used, anesthetic effects, and species differences. For instance, studies in which the data appear to support a complete overlap were performed in two relatively primitive species, the Virginia opossum and the short-tailed opossum, both from the order Didelphimorphia. Furthermore, the restriction of thalamic projections from motor nuclei of the thalamus to S1 was observed only in the Virginia opossum. Thus, it may be that the order Didelphimorphia represents a more primitive state, and that modern Australian radiations have evolved motor cortex organization that more closely resembles that of most placental mammals.

The Evolution of Cortical Fields in Marsupials and Other Mammals

Comparative studies of sensory and motor neocortex in a number of marsupials indicate that despite differences in lifestyle, size, and phylogeny, all marsupials share a common plan of cortical organization (Figure 5). For somatosensory cortex, S1 and S2 are present in all marsupials investigated, as well as in placental and monotreme mammals. This suggests that these fields arose very early in mammalian evolution and were retained in all lineages. Three other somatosensory fields (PV, SR3a, and SC/PP) have been identified in several marsupials, although in some species these fields are more primitive and poorly developed, such as in the Virginia opossum and the short-tailed opossum (Figures 5(a) and 5(b)). As in marsupials, area 3a in placental mammals is well developed in some species, such as primates, and poorly developed in other species, such as mice. Similarly, PP is present in placental mammals but has been greatly expanded in primates into a large area with numerous cortical fields. The presence of these fields in both marsupials and placentals, but not in monotremes (the presence of a separate S2 and PV in monotremes is unresolved), indicates that these fields arose after the monotreme radiation.

For visual cortex, V1 and V2 have been identified in all marsupial and placental mammals studied, indicating that both fields were present in the common ancestor. However, V2 is less well developed in some species than in others. V1 has been identified in monotremes, but V2 does not appear to be present in this group. This suggests that V2 arose after the monotreme radiation.

Although less work has been done on the organization of auditory cortex, A1 has been identified (electrophysiologically and/or architectonically) in all marsupials investigated, as well as in placental and monotreme mammals. The ubiquity of this field in all subclasses indicates that it was present in the earliest common ancestor.

Finally, the status of motor cortex in marsupials is contentious. However, when all data are examined it appears that two types of organization exist. The first type is represented in primitive marsupials, such as the order Didelphimorphia. In these animals, there appears to be a complete sensory–motor overlap in that S1 is coextensive with M1. In more advanced marsupials, such as the order Diprotodontia, S1 and M1 only partially overlap. The state of organization in the common ancestor of all mammals is difficult to deduce from existing data. However, when comparing data across species, we believe that the common ancestor had a complete sensory–motor overlap. In the three species of extant monotremes, which are highly derived, these fields became distinct. In early marsupials, this overlap was inherited from the common ancestor, and in some advanced marsupial lineages these fields have become more distinguished and now only partially overlap. We propose that in early placental mammals a sensory–motor overlap existed, and it still exists for some primitive groups such as insectivores. In other groups, such as rodents, these areas are partially separated, and in some groups, such as primates, these areas have completely separated.

Although similarities exist in the organization of the cortex across all mammalian species, there are several

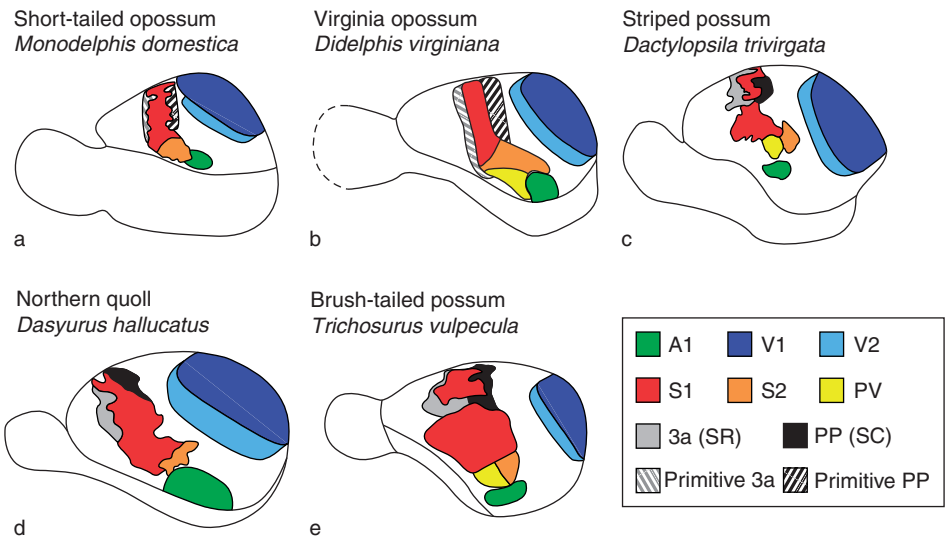

Figure 5 A summary of sensory neocortical organization in five marsupial species. Somatosensory areas S1 (red) and S2 (orange) have been identified in all marsupials (a–e) and all mammals. In some species, S2 has been subdivided into S2 and PV (yellow; (b, c, e). Two additional fields, 3a (gray) and PP (black), have also been identified (a–e), although in some species, such as the short-tailed opossum (a) and the Virginia opossum (b), these fields are more primitive (primitive 3a (gray striped) and primitive PP (black striped)). Visual areas V1 (dark blue) and V2 (light blue) have been identified in all marsupials (a–e); however V2 has only been fully described in the Northern quoll (d). Similarly, the primary auditory area A1 (green) has been identified in all marsupials (a–e) and all mammals; however, it has only been described in detail in the Northern quoll and brush-tailed possum (d, e). Rostral is to the left; medial is up. (a) Based on Huffman KJ, Nelson J, Clarey J, and Krubitzer L (1999) Organization of somatosensory cortex in three species of marsupials. *Dasyurus hallucatus, Dactylopsila trivirgata*, and *Monodelphis domestica*: Neural correlates of morphological specializations. *Journal of Comparative Neurology* 403: 5–32; Frost SB, Milliken GW, Plautz EJ, Masterton RB, and Nudo RJ (2000) Somatosensory and motor representations in cerebral cortex of a primitive mammal (*Monodelphis domestica*): A window into the early evolution of sensorimotor cortex. *Journal of Comparative Neurology* 421: 29–51; Kahn DM, Huffman KJ, and Krubitzer L (2000) Organization and connections of V1 in *Monodelphis domestica*. *Journal of Comparative Neurology* 428: 337–354. (b) Beck PD, Pospichal MW, and Kaas JH (1996) Topography, architecture, and connections of somatosensory cortex in opossums: Evidence for five somatosensory areas. *Journal of Comparative Neurology* 366: 109–133. (c) Huffman KJ, Nelson J, Clarey J, and Krubitzer L (1999) Organization of somatosensory cortex in three species of marsupials. *Dasyurus hallucatus, Dactylopsila trivirgata*, and *Monodelphis domestica*: Neural correlates of morphological specializations. *Journal of Comparative Neurology* 403: 5–32. (d) Rosa MG, Krubitzer LA, Molnar Z, and Nelson JE (1999) Organization of visual cortex in the Northern quoll, *Dasyurus hallucatus*: Evidence for a homologue of the second visual area in marsupials. *European Journal of Neuroscience* 11: 907–915. (e) Elston GN and Manger PR (1999) The organization and connections of somatosensory cortex in the brush-tailed possum (*Trichosurus vulpecula*): Evidence for multiple, topographically organized and interconnected representations in an Australian marsupial. *Somatosensory Motor Research* 16: 312–337 and Huffman KJ, Nelson J, Clarey J, and Krubitzer L (1999) Organization of somatosensory cortex in three species of marsupials. *Dasyurus hallucatus, Dactylopsila trivirgata, and Monodelphis domestica*: Neural correlates of morphological specializations. *Journal of Comparative Neurology* 403: 5–32.

differences in cortical field organization that relate to species-specific lifestyle. The first is a difference in the amount of cortex devoted to a particular sensory system. For example, in the Northern quoll (Figure 5(d)), visual cortex and auditory cortex are relatively large compared to the size of somatosensory cortex. This may be related to the Northern quoll's carnivorous lifestyle; it is an agile hunter and likely coordinates both auditory and visual information to capture prey. Conversely, the brush-tailed possum, which is a phalanger with a well-developed hand, typically eats leaves and flowers and has a larger somatosensory cortex than either visual or auditory cortex (Figure 5(e)). The second difference between species is the cortical magnification of behaviorally relevant sensory surfaces. For example, in the striped possum the representations of the fourth digit and the tongue are relatively large compared to other body part representations within S1 and compared to similar representations in S1 of other species. In the striped possum, the fourth digit and the tongue are both used in their specialized feeding behavior. In the Tammar wallaby, one of only three diurnal marsupials, an enlarged representation of central vision is present within V1, as compared to the organization of V1 in nocturnal marsupials. Finally, high-frequency representations are relatively large in the Northern quoll, and these frequencies match the vocalization frequencies of predators and prey of this species.

The types of adaptations that have been made to cortical fields in marsupials are remarkably similar to those described in placental mammals. This is not

surprising given that marsupials, like placental mammals, have undergone significant radiations during evolution and have diversified greatly, particularly in Australia. This diversification consists of a wide array of morphological, behavioral, and cortical specializations. Many of the morphological and behavioral specializations are associated with patterns of activity (diurnal vs. nocturnal), habitat selection (arboreal, terrestrial, aquatic, or subterranean), and diet (carnivorous, herbivorous, omnivorous, or frugivorous). The similarities in cortical field organization indicate that there are large constraints imposed on evolving systems and that similar solutions to environmental challenges have independently emerged.

See also: Comparative Neurobiology: History; Evolution of Vertebrate Brains.

Further Reading

Aitkin L (1995) The auditory neurobiology of marsupials: A review. *Hearing Research* 82: 257–266.

Aitkin LM, Irvine DR, Nelson JE, Merzenich MM, and Clarey JC (1986) Frequency representation in the auditory midbrain and forebrain of a marsupial, the northern native cat (*Dasyurus hallucatus*). *Brain, Behavior, and Evolution* 29: 17–28.

Beck PD, Pospichal MW, and Kaas JH (1996) Topography, architecture, and connections of somatosensory cortex in opossums: Evidence for five somatosensory areas. *Journal of Comparative Neurology* 366: 109–133.

Elston GN and Manger PR (1999) The organization and connections of somatosensory cortex in the brush-tailed possum (*Trichosurus vulpecula*): Evidence for multiple, topographically organized and interconnected representations in an Australian marsupial. *Somatosensory Motor Research* 16: 312–337.

Frost SB, Milliken GW, Plautz EJ, Masterton RB, and Nudo RJ (2000) Somatosensory and motor representations in cerebral cortex of a primitive mammal (*Monodelphis domestica*): A window into the early evolution of sensorimotor cortex. *Journal of Comparative Neurology* 421: 29–51.

Haight JR and Neylon L (1979) The organization of neocortical projections from the ventrolateral thalamic nucleus in the brush-tailed possum, *Trichosurus vulpecula*, and the problem of motor and somatic sensory convergence within the mammalian brain. *Journal of Anatomy* 129: 673–694.

Huffman KJ, Nelson J, Clarey J, and Krubitzer L (1999) Organization of somatosensory cortex in three species of marsupials. *Dasyurus hallucatus, Dactylopsila trivirgata*, and *Monodelphis domestica*: Neural correlates of morphological specializations. *Journal of Comparative Neurology* 403: 5–32.

Johnson JI (1990) Comparative development of somatic sensory cortex. In: Jones EG and Peters A (eds.) *Cerebral Cortex*, pp. 335–449. New York: Plenum.

Kahn DM, Huffman KJ, and Krubitzer L (2000) Organization and connections of V1 in *Monodelphis domestica*. *Journal of Comparative Neurology* 428: 337–354.

Lende RA (1963) Cerebral cortex: A sensorimotor amalgam in the marsupiala. *Science* 141: 730–732.

Lende RA (1969) A comparative approach to the neocortex: Localization in monotremes, marsupials, and insectivores. *Annals of the New York Academy of Sciences* 167: 262–275.

Murphy WJ, Pevzner PA, and O'Brien SJ (2004) Mammalian phylogenomics comes of age. *Trends in Genetics* 20: 631–639.

Rosa MG, Krubitzer LA, Molnar Z, and Nelson JE (1999) Organization of visual cortex in the Northern quoll, *Dasyurus hallucatus*: Evidence for a homologue of the second visual area in marsupials. *European Journal of Neuroscience* 11: 907–915.

Rowe M (1990) Organization of the cerebral cortex in monotremes and marsupials. In: Jones EG and Peters A (eds.) *Cerebral Cortex*, pp. 263–334. New York: Plenum.

Mass Spectroscopy of Proteins

E V Romanova, S P Annangudi, and J V Sweedler, University of Illinois, Urbana, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Mass spectrometry (MS) is an analytical technique for the determination of the composition of chemical substances by measuring accurate molecular masses. In MS analysis, the molecules of interest are vaporized and ionized, and the mass-to-charge (m/z) ratios of molecular ions are determined. When these gas-phase ions are broken down into characteristic fragments, the measurement of the masses of the fragments provides information on the original ion's molecular structure. Although MS has been used to characterize small molecules for almost a century, the ability to vaporize and ionize larger biopolymers such as proteins was limited, preventing the application of MS to many protein characterization tasks. However, breakthrough developments in soft ionization methods, acknowledged by the Nobel Prize to K Tanaka and JB Fenn in 2002, advanced the application of MS to protein analysis. By combining liquid-phase separations to fractionate complex samples with modern MS measurement techniques, a wide range of peptide and protein characterization efforts are now possible. In recent decades, MS has developed into a highly sensitive, accurate, and powerful tool for rapid qualitative and semi-quantitative characterization of peptides and proteins present in samples ranging from individual cells, tissue sections, and biological fluids to entire organisms. This article provides a brief overview of the capabilities of MS for protein measurements, with an emphasis on neuroscience-related samples. Throughout the article, illustrative references provide additional information.

Principles of Mass Spectrometric Analysis and Instrumentation

MS analysis consists of two essential steps: (1) a process to vaporize and ionize analytes of interest from a sample, followed by (2) a means to measure the masses of the resulting ions. The most effective ionization of proteins is achieved by soft ionization methods such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) (Figure 1). The ionization of biological samples by both ESI and MALDI imparts low internal energies on to ionized molecules, limiting their fragmentation/breakdown so that larger molecules can be studied. Briefly, in MALDI a liquid or solid biological sample is co-crystallized with a large excess of ultraviolet (UV)-absorbing matrix; illumination of the matrix-sample crystals by a UV laser results in a microexplosion that produces primarily singly charged gaseous analyte ions. In ESI, electric potential is applied to a capillary filled with a liquid sample. This generates an electrospray consisting of small liquid droplets containing the sample, which are desolvated and form multiply charged gas-phase ions. Unlike a MALDI mass spectrum, an ESI mass spectrum of an individual peptide or protein contains a number of signals corresponding to different charge states of each analyte.

The two approaches are complementary and can be used for the same samples. However, differing capabilities and figures of merit often suggest one approach may be more suited to a particular sample type than another. As an example, samples containing high levels of salts, or those that require direct measurement of compounds from a tissue, may be better suited to MALDI approaches, whereas liquid-phase purifications of biological samples may be more easily interfaced to an ESI MS system. Both ionization approaches are often combined with triple quadrupole, ion trap, Fourier transform (FT) ion cyclotron resonance (ICR) or time-of-flight (TOF) mass analyzers (see Box 1), each of which has particular advantages for protein identification. The wide range of current mass spectrometric instrumentation provides a choice of tools for a variety of applications in life sciences.

Protein Identification and Structure Determination

Identifying the protein components of specific brain regions and individual nerve cells provides a basis for understanding numerous neurobiological phenomena. MS methodologies have the capability of identifying proteins by directly deducing their primary structure. The most common approach, known as peptide mass fingerprinting or a bottom-up measurement, involves digestion of a purified protein of interest or a simple protein mixture into smaller peptides using enzymes such as trypsin. The masses of the resulting peptide ions are measured, compared to predicted masses of proteolytic peptides from a database (which often are unique for a given protein), and mapped to protein sequence stretches that match experimentally observed peptide masses (Figure 2). Peptide mass fingerprinting alleviates the need for characterizing high-molecular-weight proteins and makes protein analysis amenable

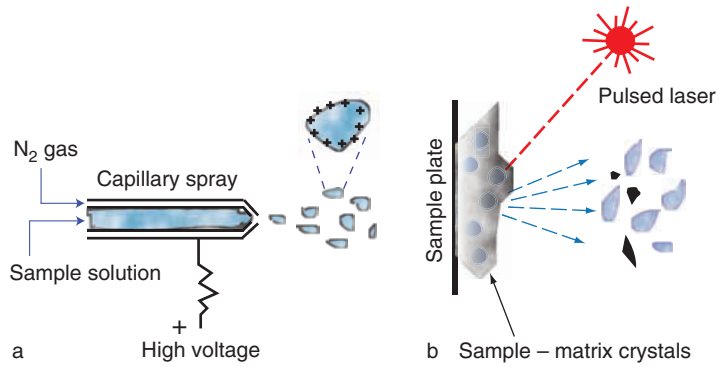

Figure 1 Two ionization techniques commonly used in biological mass spectroscopy: (a) electrospray ionization (ESI); (b) matrix-assisted laser desorption/ionization (MALDI). As shown in (a), ESI is an ionization technique that uses high voltage at a capillary tip to disperse and spray solvated analytes from the capillary. The spray produces multiply charged ions that are transferred directly into a mass analyzer. As shown in (b), MALDI is a technique that uses a laser to irradiate analytes co-crystallized with a matrix that absorbs the laser energy and forms singly charged gaseous analyte ions.

Box 1 The mass analyzers used for MS-based characterization

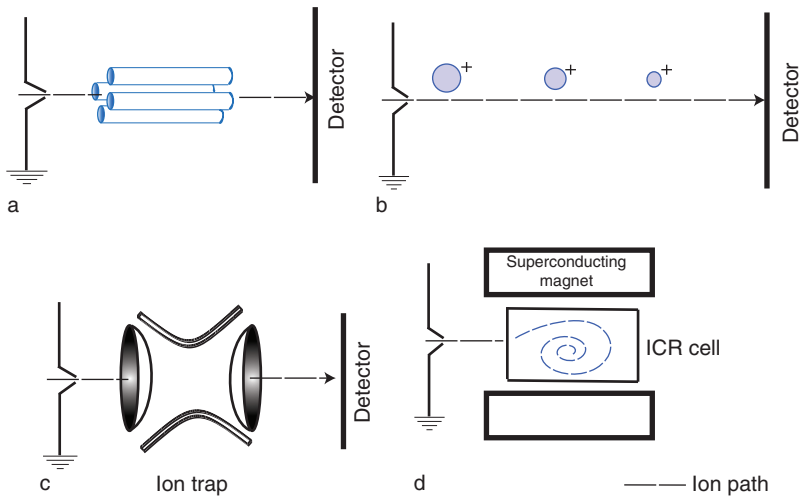

The mass analyzer is the part of the measurement process that subjects the ions to a series of electric or magnetic fields in order to allow determination of the *m/z* of the ions. As different mass analyzers offer different figures of merit (measurement speed, mass resolution, mass range, dynamic range), a brief understanding of these differences can help the user select the appropriate instruments for a specific measurement task. The following briefly describes and illustrates four common mass analyzers:

(a) In quadrupole mass analyzers, the ions generated from either of the ionization techniques (ESI or MALDI) pass through a set of four rods that have DC and alternating RF potentials applied to them that focus the ions and allow selected ions to reach the detector. In a triple quadrupole instrument, the ions pass through successive sets of quadrupoles, with each set used for ion selection, ion focusing, and fragmentation, respectively, before they reach the detector.
(b) Quadrupole ion trap instruments use hyperbolic-shaped electrodes to confine ions to stable trajectories inside a specific region, the ion trap, with the potentials applied to these electrodes set depending on the *m/z* range of the ions to be measured. The confined ions can then be either ejected into the detector or dissociated by collisions prior to remeasurement or ejection.
(c) In TOF analyzers, the ions are formed in the ionization source and accelerated by an applied potential so that all ions have the same kinetic energy before being introduced to the field-free region of the mass analyzer. Thus, ions travel at different velocities related to the kinetic energy of the ion. The flight time that each ion spends traveling in the field-free region is converted to an *m/z* value. Conceptually, the TOF instrument is a simple mass analyzer that has the ability to measure ions over a wide mass range.
(d) The operational principle of FT-ICR is to subject ions to a uniform, high magnetic field while in the ICR cell; this field causes the ions to orbit in circular orbits (e.g., cyclotron motion). Because a large number of orbits can be measured, the frequency of ion motion can be measured accurately and precisely. These frequencies are converted to *m/z* values using the Fourier transform algorithm. FT-ICR offers unparalleled mass accuracy and resolution.

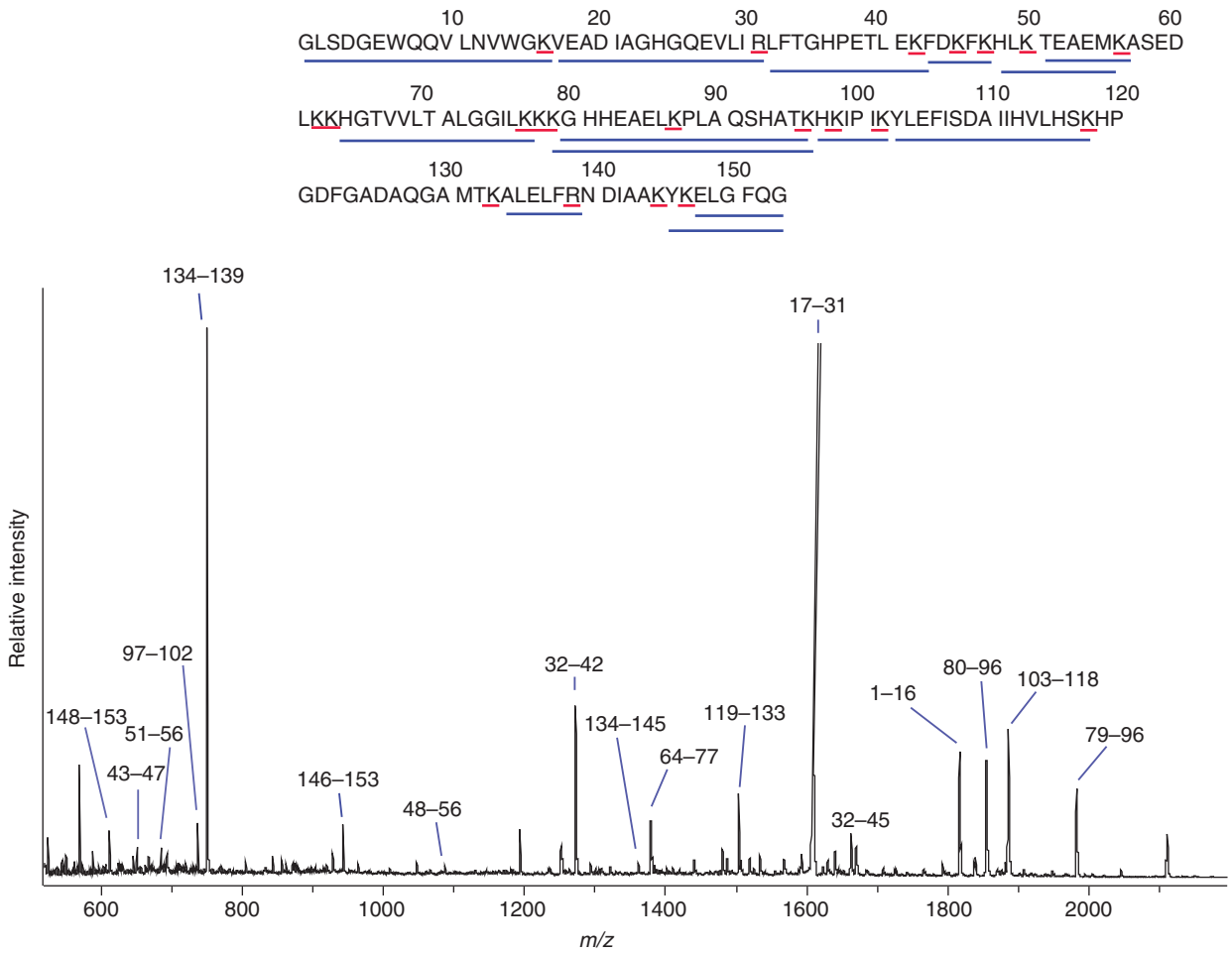

Figure 2 Peptide mass fingerprint of myoglobin by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectroscopy. Mass spectrum shows multiple peaks that represent peptides resulting from proteolytic digestion of purified horse myoglobin standard. Peptide peaks are labeled by position of the peptide amino acids in the protein sequence. Identities of measured peptides are often assigned using bioinformatics approaches such as MASCOT. This program compares the detected masses in the spectra to predicted masses of proteolytic peptides from the protein database and determines which proteins best match statistically with the experimental data. Sequence stretches underlined in blue denote the peptide identified in the spectrum, and red denotes the sites of trypsin cleavage. In this example, the resulting tryptic peptides cover 94% of the protein sequence. *m/z*, mass-to-charge ratio.

to several types of mass spectrometers. However, if a crude extract containing numerous proteins is digested, it places severe demands on the separation and analysis system; a sample that initially contained hundreds of protein compounds contains tens of thousands of proteolytic peptides after digestion.

Existing knowledge on gene expression in the biological tissue or organism under analysis allows identification of proteins by peptide fingerprinting when combined with database searches. Numerous search algorithms are available that rank the recovered protein sequences according to their probability of producing a peptide map that matches the experimentally observed peptides. As the number of organisms with known genomes increases, peptide fingerprinting is proving to be an excellent verification tool in protein expression studies.

Instrumentation platforms that couple two or more mass analyzers in succession, known as tandem mass spectrometry (MS/MS), enable identification of a protein by fragmenting the analyte to allow the determination of its partial or complete amino acid sequence. For example, when analyzing a protein digest by MS/MS, one of the proteolytic peptides can be chosen after the first MS measurement and its molecular ions selectively fragmented by a variety of physical approaches for molecular dissociation. Ideally, fragmentation produces a ladder series of fragment ions (truncated forms of the initial peptide ions) that are analyzed by a second mass analyzer. Mass differences between fragment ions provide information on the arrangement of the amino acids within the fragmented peptide. The deduced peptide sequence, in turn, can be statistically matched to a specific part of a protein

in the database. The ability to sequence peptides by fragmentation minimizes the number of proteolytic peptides needed for confident protein assignment, which extends the use of the mass fingerprinting approach to identification of proteins in more complex mixtures, as well as to proteins that produce fewer proteolytic peptides. Sequencing of peptides from unknown proteins (*de novo* sequencing) may lead to the characterization of new gene products and annotation of new genes.

With the advent of high-resolution mass spectrometers, a second, newer approach to protein identification has been developed, referred to as top-down MS. Unlike peptide mass fingerprinting, top-down analysis begins with direct MS detection and sequencing of an intact protein and lays a foundation for full structural protein characterization. The sequencing of intact proteins is achieved by producing various truncated forms of an entire protein ion by fragmentation inside the mass spectrometer (Figure 3). Thus far, FT-ICR MS appears well suited for top-down protein analysis because it has the required high resolving power of mass measurement and sensitivity. In addition, highly efficient fragmentation techniques that yield complementary information on whole protein fragment ions have been developed specifically for FT-ICR MS. For protein assignment, the data obtained by top-down MS/MS can be processed by search engines in a manner similar to that of the bottom-up approach already described. Although highly accurate and effective in protein identification, the top-down approach is challenging when routinely

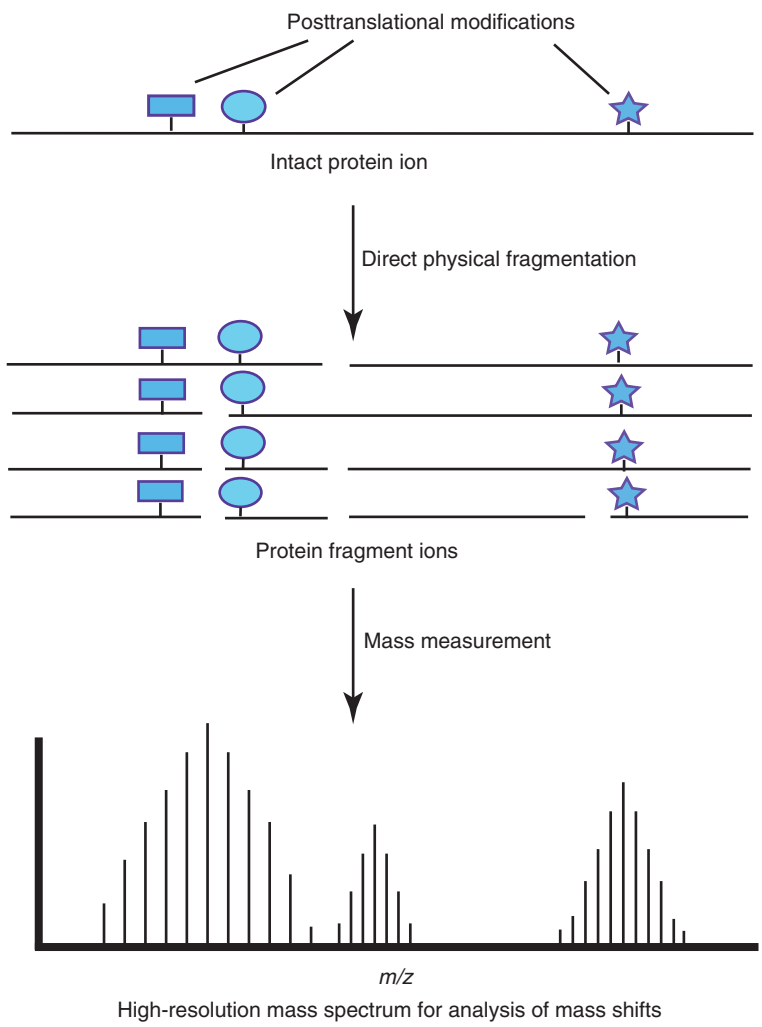

Figure 3 Schematic of the top-down approach for protein sequencing and identification. First, the mass of the whole protein ion is determined. Protein ions are then fragmented within the mass spectrometer and the accurate mass of each of the resulting fragments is measured. Patterns of mass shifts between fragment ions are indicative of the sequence in which amino acids are arranged in the original molecular ion. Analysis of fragment ion patterns allows reconstruction of the protein sequence from both or either terminus. Known variations of mass shifts in fragment ion patterns elucidate the presence of posttranslational modifications. Identification of the deduced protein sequences requires database searches. *m/z*, mass-to-charge ratio.

handling proteins >50 kDa, crude samples with >200 proteins, or complex posttranslational modifications such as glycosylation.

Characterization of Protein Posttranslational Modifications and Differential Gene Expression

According to a finding of the Human Genome Project, expressed protein forms outnumber the protein-coding genes by an order of magnitude (~20 000–25 000 genes vs. ~400 000 proteins). Protein diversity that begins with a single nucleotide polymorphism, cell- or tissue-specific alternative splicing, or editing of mRNAs encoded by one gene can be increased further by the addition or removal of various chemical groups to or from the primary protein structure after translation. These chemical changes, known as posttranslational modifications (PTMs), result in alterations of inherent molecular properties such as charge or conformation. These changes often confer unique bioactivity to the modified protein. MS is ideal for the discovery and characterization of PTMs on proteins because most PTMs result in characteristic mass changes to the original molecular ion. For example, phosphorylation of a protein molecule at serine, threonine, tyrosine, or aspartate results in a mass shift of 80 Da due to the addition of a phosphate group. In many cases, comparing calculated and measured molecular masses of intact ions can be sufficient to determine whether a protein or peptide molecule carries a PTM. With the help of the fragmentation techniques already described, the location of an anticipated PTM can be mapped to a particular amino acid residue within the protein by analyzing mass shifts among the fragment ions generated in an MS/MS spectrum.

MS is also effective for the study of protein isoforms resulting from differentially expressed genes. Minor changes in a protein's primary structure associated with alternative splicing or editing of the protein-encoding mRNA lead to changes in the molecular masses of the original protein ion. Top-down protein identification by FT MS is especially effective in this case because various direct fragmentation techniques yield complementary fragment ions that often provide 100% protein sequence coverage and localize sites of basic PTMs.

Quantitative Analysis of Protein Expression

MS approaches to quantify protein levels from two or more complex mixtures use the ability of MS to differentiate isotopically distinct but chemically identical analytes. The technique is based on labeling each sample using isotopic labels, thus conferring a mass change that is detected using MS (Figure 4).

Various labeling approaches have been developed for quantitative analysis of protein expression in biological samples, with isotopic labels, incorporated via metabolic, enzymatic, or chemical reaction, being the most widely used. In metabolic labeling, isotopes such as N^{15} or C^{13} are incorporated into cultured cells via a culture medium that is enriched with amino acids containing stable isotopes in light and heavy forms. The intensity ratio of the peptide peaks labeled with light and heavy isotopes in a mass spectrum reflects the relative abundance of the two peptide species in the sample (Figure 4(c)). Alternatively, isotopes can be introduced at particular places on protein molecules via specific chemical or enzymatic reactions with isotopically encoded reagents. The two MS-based quantitative reagents that have rapidly gained popularity are isotope-coded affinity tags and isobaric tags for relative and absolute quantification.

Strategies for MS determination of the absolute amounts of selected proteins in a sample rely on the addition of known amounts of isotopically labeled recombinant or synthetic internal standards of native proteins found in the sample. When added to a protein tryptic digest, such standards undergo trypsin treatment along with the endogenous proteins. The resulting tryptic peptides are then detected in a mass spectrometer as pairs of labeled and unlabeled identical peptides, differing by the mass of the known isotopic tag. The relative abundance of the endogenous proteins in the original sample is deduced from the ratio between the peak intensities of each peptide pair.

Applications of Mass Spectrometry in Functional Proteomics

Dynamic molecular networks, which often include supramolecular assemblies of noncovalent protein–protein, protein–DNA, and protein–ligand complexes, regulate central neurophysiological processes such as signal transduction and plasticity. Although the structural analysis of macromolecular formations presents a significant analytical challenge, MS becomes increasingly useful in mapping protein interactions. One MS-based approach for studying protein complexes uses the ability of ESI to preserve noncovalent interactions during analysis. In contrast to biophysical techniques such as X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR), MS has the ability to detect different species simultaneously, often at lower concentrations. Using MS, it is possible to measure the mass of the entire complex, as well as the individual subunits produced by molecular collisions inside a tandem mass spectrometer. MS measurements

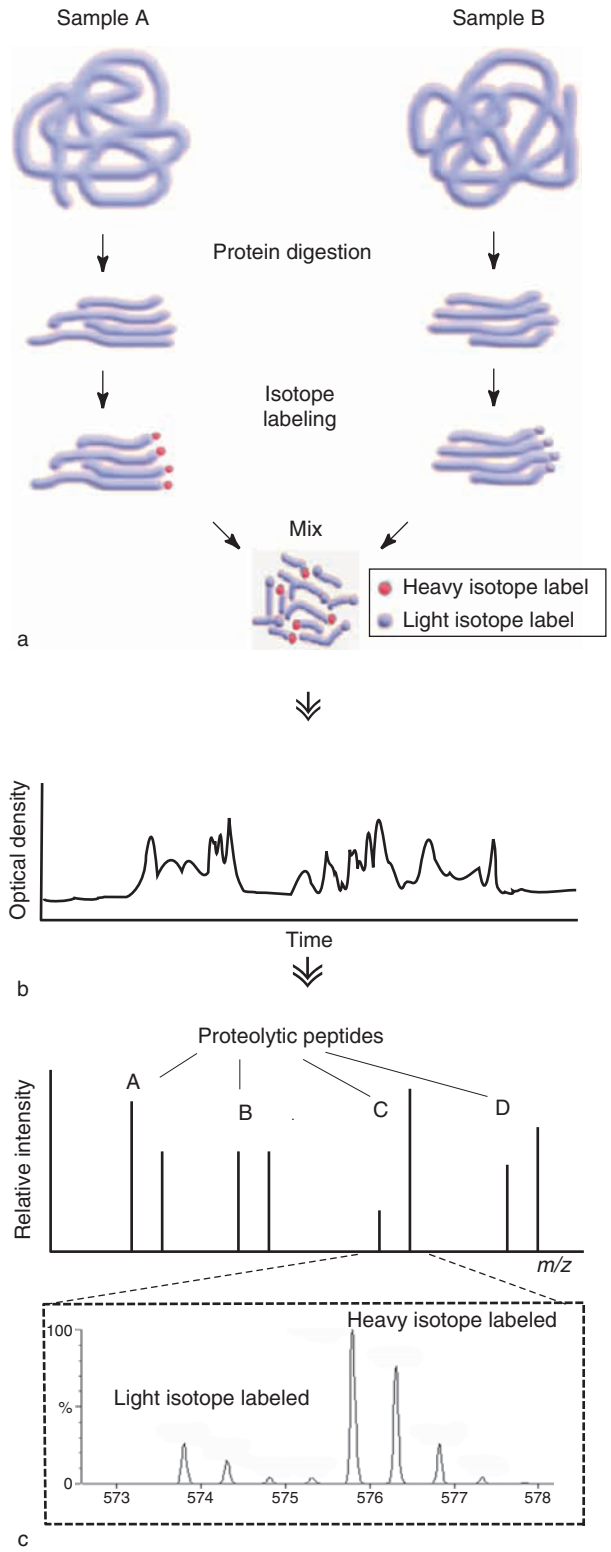

Figure 4 Protein quantitation using stable isotope labeling and liquid chromatography (LC) mass spectroscopy: (a) labeling and mixing samples A and B; (b) reverse phase liquid chromatography separation (RP-LC); (c) mass spectrometry. In (a), proteins from two different samples (A and B) are isolated and digested using a proteolytic enzyme. Resulting proteolytic peptides are labeled using either heavy (sample A) or light (sample B) isotopic modifiers and combined. In (b), this mixture is fractionated by LC. Identical peptides originating from samples A and B, now carrying heavy or light isotope labels, respectively, coelute in the chromatogram. Coeluting, identically labeled peptides differ by Δm, which equals the mass difference between heavy and light isotope labels. Therefore, identical peptides from different samples can be resolved and analyzed using a mass spectrometer (c). The differences in peak intensities of the labeled peptides reflect the differences in the peptide amounts in the original samples. The lower trace in (c) shows an example of a high-resolution zoom-in view of a typical quantitative mass spectrum. *m/z*, mass-to-charge ratio.

help deduce the stoichiometry of the components making up the complex and may reveal the presence of various metals in metalloproteins. The energy (voltage) required for dissociation of each particular complex serves as a relative measure of complex stability and interaction strength. Because MS can be applied to mapping protein interactions on a global scale, it may also be considered an improved alternative to yeast two-hybrid screening, co-immunoprecipitation, and affinity chromatography – methods used extensively in neuroscience.

Mass Spectrometry for Biomarker Discovery

The ability to measure simultaneously numerous protein components in a sample, and to differentiate between proteins and peptides with PTMs, makes MS well suited for large-scale comparisons of protein expression profiles in search of protein biomarkers that establish links to normal or pathogenic health conditions, disease progression, or therapeutic responses. The power of MS in biomarker discovery is that, in contrast to biomarker screening methods based on the use of selective probes, MS can be used without prior information on the form of the resulting biomarker. Moreover, instead of tracing changes in just a few particular proteins targeted by probes, MS screening can detect unexpected sets of diverse molecular changes that encompass a range of functional pathways in affected individuals. MS screening can be performed on a limited amount of tissue extract or body fluid and, in some cases, without extensive sample purification. Furthermore, MS profiling can be helpful in diagnosing conditions that include genetic abnormalities such as progressive mutations, as well as identifying PTMs. Current discovery efforts are focused on cancer, heart disease, Alzheimer's disease, and rheumatoid arthritis markers.

Mass Spectrometric Measurements in the Brain

This article describes the capability of MS to detect, structurally characterize, identify, compare, and

quantify compounds in biological samples at different organismal levels, facilitating investigations in many subfields of neuroscience. For instance, analysis of proteins at the system level is challenging because of the inherent complexity of the brain. Fractionation techniques such as two-dimensional gel electrophoresis, liquid chromatography, and affinity purifications are commonly used in conjunction with MS to reduce the chemical complexity of such samples and to enhance investigation at the protein level. Profiling of protein expression directly in brain tissues or brain extracts contributes to comparative proteomic analyses in normal as well as pathological conditions, including neurodegeneration, neurotrauma, psychiatric diseases, drug addiction, and nervous system tumors.

An alternative approach is to decrease anatomical complexity and perform the analysis at the level of the individual neuron or even subcellular compartments. Currently, single-cell MS is effectively being used to map the cellular localization of novel peptide transmitters and analyze neuropeptide release at identified release areas. Combined with electrophysiology and molecular biology, single-cell MS has become an indispensable tool for delineating neuronal circuits that regulate simple behaviors in model invertebrates.

A newer application of MS in neuroscience is imaging mass spectrometry. Unlike other imaging techniques, such as immunohistochemistry and autoradiography, MS imaging simultaneously detects numerous compounds without the use of specific molecular probes and labels. Imaging MS becomes increasingly practical for mapping the spatial distribution of peptides and proteins and their structural variances directly in brain tissue and even individual neurons.

Currently, MS is a mainstream neuroproteomic technology well suited for studying protein function. Progress in sampling and separation methodologies, engineering, microfabrication, and bioinformatics continues to advance MS-based approaches for the investigation of protein structure and function in the central nervous system.

See also: Fluorescent Biomarkers in Neurons; Neuroproteomics.

Further Reading

Bogdanov B and Smith RD (2005) Proteomics by FTICR mass spectrometry: Top down and bottom up. *Mass Spectrometry Reviews* 24: 168–200.

Caldwell RL and Caprioli RM (2005) Tissue profiling by mass spectrometry: A review of methodology and applications. *Molecular and Cellular Proteomics* 4: 394–401.

Caprioli RM (ed.) (2004) *The Encyclopedia of Mass Spectrometry: Biological Application, A*. New York: Elsevier Science.

Choudhary J and Grant SG (2004) Proteomics in postgenomic neuroscience: The end of the beginning. *Nature Neuroscience* 7: 440–445.

Kalia A and Gupta RP (2005) Proteomics: A paradigm shift. *Critical Reviews in Biotechnology* 25: 173–198.

Krogan NJ, Cagney G, Yu H, et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440: 637–643.

Li L, Garden RW, and Sweedler JV (2000) Single-cell MALDI: A new tool for direct peptide profiling. *Trends in Biotechnology* 18: 151–160.

Relevant Websites

http://www.asms.org – American Society for Mass Spectrometry (tutorial).

http://www.i-mass.com – International Mass Spectrometry Web Resource.

http://www.ionsource.com – Mass Spectrometry and Biotechnology resources.

Mechanoreceptors

A S French and P H Torkkeli, Dalhousie University, Halifax, NS, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Mechanoreceptors are primary sensory structures that provide information about mechanical features of the internal and external environments. They always contain at least one sensory neuron that responds to mechanical stimuli with a change in membrane potential. This graded receptor potential usually initiates or modulates action potentials that propagate along an axon to the central nervous system. Some mechanoreceptors, such as vertebrate auditory organs, separate the two functions between two different cells, with the first cell releasing a chemical transmitter to excite action potentials in a second cell. Also, some invertebrate sensory neurons are close enough to the central nervous system to transmit the graded receptor potential directly. Mechanoreceptors detect a wide range of mechanical stimuli from both the external environment (e.g., touch, vibration, and hearing) and the internal environment (e.g., muscle length and hollow organ distension).

Mechanoreception can be conceptually divided into three stages: coupling, transduction, and encoding. The mechanical stimulus is usually coupled by a physical structure to produce stress at the sensitive membrane of the sensory neuron. These physical structures are often complex, and their morphologies have been used to identify and classify different types of mechanoreceptors. Examples include the hair follicles that cover much of the mammalian body and the chordotonal organs of arthropods that provide hearing and vibration sensitivity. Transduction is the process by which mechanical stress in the cell membrane produces the receptor current, leading to the receptor potential. This essential and defining property of all mechanosensory neurons is due to the presence of mechanically activated ion channels in the cell membrane. However, mechanically activated ion channels also occur in cells that are not mechanosensory neurons. The receptor potential is usually encoded into action potentials by a process generally similar to action potential encoding in other neurons. Variations in the sensitivity and dynamic properties of encoding often contribute to the functional properties of the mechanoreceptor, such as the high-frequency response of some vibration detectors.

Mechanical Coupling to the Sensitive Membrane

Structures of both invertebrate and vertebrate mechanoreceptors vary from simple unbranched dendrite endings to complex and elaborate forms made from additional cells and extracellular materials. Many soft-bodied animals have relatively simple sensory endings just below the outer layers of the integument. For example, the roundworm, *Caenorhabditis elegans*, has single nerve cell endings that approach the skin, where they are surrounded by hypodermal cells and a dense sheath of extracellular matrix. However, mechanoreceptors of more elaborate animals, including vertebrates and arthropods, range from relatively simple to very complex forms.

Arthropods have highly specialized cuticular structures that allow external mechanical stimuli to reach the sensory endings through the hard exoskeleton. However, the wide range of external shapes and sizes of these structures can be classified into four basic types: hair, campaniform, chordotonal, and slit sensilla (Figure 1). Hair sensilla respond to disturbances of air or fluid outside the animal, campaniform sensilla detect stress in the cuticle, and chordotonal sensilla respond to a variety of stimuli that originate inside or outside the body, including vibration and sound. Slit sensilla have only been found in arachnids and serve functions similar to those of the campaniform sensilla of insects.

Arthropods also contain another type of mechanoreceptor, the multipolar receptors. These are found throughout the body, from just below the integument to deep tissues such as muscle, gut, and reproductive organs. They have highly branched nerve endings that are usually sheathed by glial cells and connective tissue, except for the extreme tips of the dendrites, where naked nerve endings are assumed to transduce mechanical displacements of the innervated tissues.

Vertebrate mechanoreceptors possess an even wider range of morphological forms, which has made classification difficult. Mammalian mechanoreceptors vary from unbranched nerve endings, apparently lacking any specialization, to complex structures such as the Pacinian corpuscles with their numerous lamellar wrappings (Figure 2). The increasing morphological complexity is correlated with several other factors, including the embryonic origin of the tissues surrounding the sensory ending, the range of animals in which the receptors are found, and the sensitivity to rapidly changing, rather than sustained, mechanical stimuli. Simpler endings are generally associated with

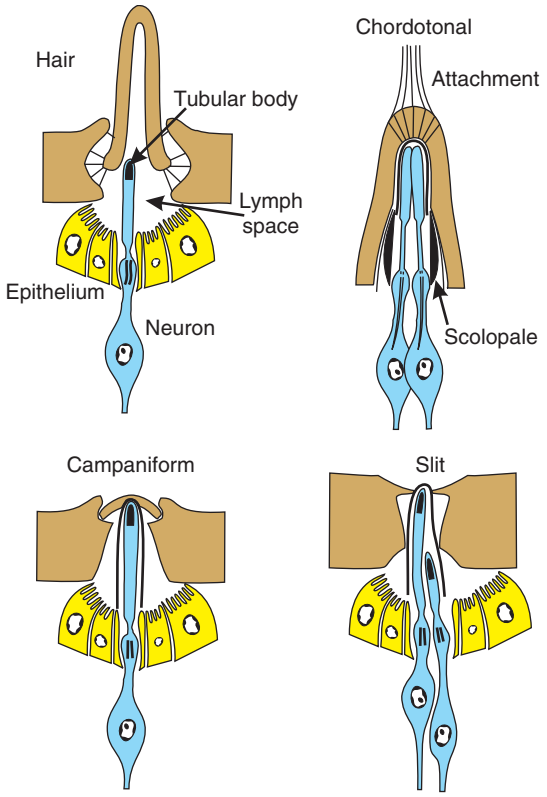

Figure 1 Examples of typical arthropod cuticular mechanoreceptors. In each case bipolar neurons (blue) send distal sensory dendrites to interact with cuticular structures. The tips of the dendrites are surrounded by a lymph space that is sealed from the interstitial space by tight junctions between epithelial cells (yellow). The lymph space contains a high concentration of the ion that carries the sensory current and is often at a positive potential relative to interstitial fluid. Hair structures have a wide range of shapes and sizes, from long, thin wind-detecting hairs to short bristles and pegs. Multiple neurons may be present below the hairs. Chordotonal sensilla are further below the cuticular surface and have a set of scolopale rods, like the staves of a barrel, surrounding the sensory dendrites. Many auditory organs are based on chordotonal receptors. Campaniform sensilla have a characteristic bell-shaped cap that squeezes the dendrite tip when the cuticle is stressed. Slit sensilla of spiders serve a function similar to that of campaniform sensilla, but have long slits that usually compress multiple pairs of dendrites.

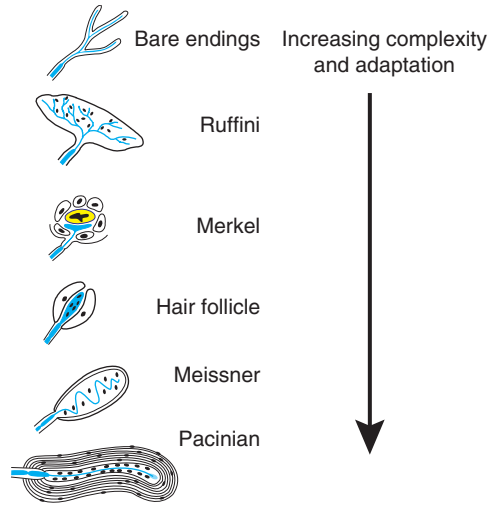

Figure 2 Some vertebrate skin mechanoreceptors. From top to bottom: Sensory neurons (blue) surrounded by increasingly complex morphological structures, from bare nerve endings to the elaborate lamellar wrappings of the Pacinian corpuscles. There is a corresponding transition from receptors with little adaptation, such as the Ruffini endings that detect steady indentation of the skin, to rapidly adapting receptors, such as Pacinian and Meissner endings that detect vibration. The function of the additional Merkel cell (yellow) is still controversial and may include mechanotransduction. Bare nerve endings and some Ruffini-type endings may be innervated by unmyelinated axons, but the more complex receptors have myelinated axons.

tissues of mesodermal origin (connective tissues and muscle) and include the Ruffini endings of joints and skin, Golgi tendon organs, and muscle spindles. These receptors are found throughout the vertebrates and they usually respond to steady deformation, such as the constant firing of joint and muscle receptors that gives kinesthesia, or awareness of body posture. At the other end of the scale are structurally complex receptors associated with tissues of endodermal or ectodermal origin such as the Meissner and Pacinian corpuscles of the skin. These preferentially respond to rapidly changing stimuli such as vibration, or movement of the skin across a rough surface. They are unknown in fish, uncommon in amphibians or reptiles, and increasingly common in birds and mammals, suggesting a relatively recent evolution of complex morphological endings in the vertebrates, possibly associated with the shift from aquatic to terrestrial environments.

Relatively little is known about the detailed functional properties of external structures surrounding mechanically sensitive nerve endings in vertebrates or invertebrates, because, in most cases, the exact location of transduction, the mechanical properties of the external components, and the component mechanical connections are unknown. The Pacinian corpuscle is a very rapidly adapting receptor, responding best to high-frequency vibration. A significant part of this adaptation is due to the numerous lamellar wrappings, because the receptor potential adapts much less when most of them are removed. Slow indentations allow the lamellae to move relative to each other but more rapid stimuli are transmitted by the viscosity of the fluids between them. However, other mechanisms are capable of producing rapid adaptation in mechanoreceptors, so it is not clear if this is the major function of the lamellae or how significant are the detailed differences between the various vertebrate receptor structures.

The functions of external coupling structures are better understood in some arthropod mechanoreceptors. Hair sensilla form levers that allow movements of external air or water to compress sensory dendrite endings, while the bell-shaped caps of campaniform sensilla convert stress in the cuticle to compression at the tips of sensory dendrites. Slit organs of spiders perform similar stress-detecting functions, although the coupling mechanisms are not so clear. Chordotonal sensilla of insects and crustaceans contain particularly elaborate structures, the functions of which are still uncertain. While all arthropod cuticular sensilla have dendrites derived from modified ciliary structures, there is evidence that the cilia of chordotonal sensilla may retain the capacity to move actively, leading to the idea of feedback loops between mechanotransduction and ciliary movement. Positive feedback could enhance sensitivity to small stimuli, or negative feedback could serve as a gain control for large stimuli. Possibly, both mechanisms operate. Interestingly, chordotonal sensilla function as auditory receptors in many insects, and similar positive or negative feedback loops have been proposed between the outer and inner hair cells of vertebrate auditory systems.

Although the general functions of many arthropod coupling structures are understood, their dynamic properties remain largely unknown; partly because the mechanical properties of the individual components are difficult to measure or estimate. The dynamic properties of spider slit organs have been studied intensively, and there is some evidence for frequency tuning by slit length, but the overall dynamic responses seem dominated by ionic events during action potential encoding. In general, the reasons for the evolution of such varying and complex coupling structures are enigmatic, particularly since arthropod multipolar receptors seem to provide many similar functions without any comparable coupling structures.

Transduction of Mechanical Stimuli

Mechanotransduction produces a receptor current through the membrane of the sensory ending. This current charges the membrane capacitance to give a membrane potential, which is depolarizing in all cases where it has been measured. It is now widely agreed that transduction is mediated by mechanically activated ion channels that open in response to mechanical deformation of the membrane. Possible alternative mechanisms include mechanically sensitive membrane receptors that activate ion channels by a second messenger cascade or, for ciliary receptors, a reversed ciliary beat in which the forced sliding of dynein arms leads to a chemical signal. However, there is no direct evidence for any kind of chemical intermediates, and the speed of mechanotransduction can approach submillisecond timescales, arguing strongly for a direct mechanism.

Although mechanically activated ion channels have been characterized in a wide range of preparations, from bacteria to human cells, direct observations of single channels that are clearly producing transduction in a mechanosensory neuron have proved elusive. Similarly, no genes or protein molecules have been identified that definitely provide sensory mechanotransduction. Differences in the ionic basis of transduction in different mechanoreceptors suggest that a range of different proteins may be involved. The receptor currents of Pacinian corpuscles and muscle spindles are carried primarily by sodium ions, although the channels are permeable to a range of cations. A similar situation prevails in crayfish stretch receptors, which are multipolar receptors. Spider slit sensilla also use sodium as the current carrier, but in this case the channels are highly selective for sodium, and lithium ions pass only with greatly reduced permeability. In contrast to these systems, both vertebrates and invertebrates have mechanoreceptors in which potassium is the main charge carrier, although some of these are relatively unselective for cations, including calcium. They include the hair cells of the vertebrate hearing and vestibular organs, as well as all of the insect cuticular sensilla that have been examined.

Another feature found in many vertebrate and invertebrate mechanoreceptors is the existence of a specialized lymph space surrounding the region of the sensory neuron where transduction is believed to occur (**Figure 1**). Lymph spaces are usually lined by epithelial cells, which produce a relatively high concentration of the ion that carries the receptor current, and often an electrical potential that tends to drive those ions into the sensory ending. These arrangements seem clearly designed to increase the electrochemical force driving the receptor current, and therefore amplify the effect of opening the mechanically activated ion channels.

Two major groups of ion channel molecules have been associated with mechanotransduction. Members of the transient receptor potential (TRP) family of channels have been implicated in a range of sensory functions of both vertebrates and invertebrates, including phototransduction, thermal transduction, mechanotransduction, pain, and osmosensation. TRP channels have been strongly linked to hearing and touch in *Drosophila*, and to touch in *C. elegans*. Homologous molecules have been found in vertebrates, but primarily are associated with osmolarity detection in a wide range of neural and nonneural tissues. Other TRP channels have been found in vertebrate pain receptors, as well as mouse, bullfrog, and

zebra fish inner ear hair receptors, appearing at the same embryonic stage as does sound sensitivity in mice. However, none of the evidence yet gives clear proof that these channels are the primary source of the receptor current.

The other channel family associated with mechanotransduction is the degenerin/acid-sensitive ion channel/epithelial sodium channel (DEG/ASIC/ENaC) family, best known for the amiloride-blockable epithelial sodium channels that conduct sodium flux through a wide range of epithelia. In *C. elegans*, two of four proteins found only in mechanoreceptor cells are DEG molecules that have been proposed to form the core of the mechanotransduction channels, and the receptor current was carried by sodium ions. A DEG gene family was also associated with mechanosensitivity in *Drosophila* larvae. In rodents, several members of the DEG family have been found in dorsal root ganglia and in fine nerve endings surrounding tactile hairs. Knockout animals for one channel, BNC1, showed reductions, but not elimination, of mechanosensation, and none of these molecules has yet been identified in well-known skin mechanoreceptors, such as Pacinian corpuscles or Ruffini endings.

Although the molecular evidence favors TRP channels in *Drosophila* mechanotransduction, indirect evidence in another arthropod preparation, spider slit sensilla, is more supportive of ASICs. Receptor current in these neurons is highly selective for sodium, blocked by amiloride, and strongly increased by low pH, which are all characteristic properties of acid-sensitive channels. Other members of this channel family have exhibited mechanosensitivity in expression systems.

A few single-channel measurements have been made from mechanoreceptor cells, with conductance estimates of about 40–100 pS. Single-channel openings of about 100 pS conductance are seen in whole-cell recordings from cochlear hair cells during mechanical stimulation. Noise analysis, which examines the statistical properties of the receptor current, has also been used in several preparations, including hair cells and spider slit sensilla, giving generally lower (5–50 pS) estimates of single-channel conductance.

Two other commonly proposed features of sensory mechanically activated channels are heteromeric construction and connections to extracellular and intracellular structural proteins. Evidence from several preparations indicates that multiple proteins are required to form functioning eukaryotic mechanically activated channels, and this may explain the difficulty of demonstrating mechanosensitivity from proteins expressed in oocytes or other systems. Mechanical connections to cytoskeletal and extracellular matrix structures have been proposed by several lines of evidence, including the amino acid sequences of proposed channel molecules. It has also been argued that lipid membrane alone could not provide enough force to open a protein channel. Microtubules are often prominent in mechanoreceptor endings, and in some cases have been suggested to form a cytoskeletal anchor, but several studies in arthropods and in *C. elegans* have shown that mechanotransduction persists after pharmacological destruction of microtubules.

Encoding Mechanosensory Information

With the exception of a few unusual cases, the depolarizing receptor potential leads to action potentials that conduct the encoded mechanical signal to the central nervous system. The usual mechanism for this encoding is the familiar combination of regenerative voltage-activated sodium channels and delayed rectifier voltage-activated potassium channels. However, a range of variations on this theme has been found. Important issues in encoding are where in the sensory cell the action potentials begin, what other ion currents are activated during the encoding, and how these various ion currents contribute to the overall sensitivity and dynamics of mechanoreception.

Initial measurements in Pacinian corpuscles suggested that the receptor potential is conducted passively from the distal nerve ending to the first node of Ranvier of the myelinated axon, which is still within the lamellar capsule. However, more recent work indicates that voltage-activated sodium channels are found in the most distal parts of the neuron, where transduction occurs, suggesting that action potentials may be initiated immediately following transduction by the same area of membrane. In muscle spindles there is also evidence for spiking activity in the most distal regions of the sensory endings, where transduction takes place.

A similar shift of opinion has occurred for the bipolar neurons of arthropod cuticular mechanoreceptors. Here, it was initially assumed that the receptor potential propagates passively through the sensory dendrite and soma to produce action potentials in the first part of the axon. However, voltage-activated sodium channels have been detected throughout these neurons, and both intracellular and extracellular recordings of several mechanosensory neurons have clearly shown that action potentials start in the most distal dendrites. Encoding of action potentials as close as possible to the transduction site must result in rapid transmission of information to the central nervous system, with obvious evolutionary advantages.

There is not much direct evidence of other voltage-activated currents in vertebrate somatic mechanosensory neurons, but voltage-activated calcium currents are well characterized in vertebrate hair cells. Calcium

plays important roles in hair cell adaptation, both by acting on motor proteins within the cells and by opening calcium-activated potassium channels. In some vertebrate hair cells the feedback loop – formed by membrane potential, calcium current, calcium-activated potassium current, and membrane potential – has been implicated in tuning the optimal frequency sensitivity of mechanotransduction. Calcium currents, calcium-activated potassium currents, and several different types of voltage-activated potassium currents have also been characterized in arthropod cuticular receptors and multipolar receptors. Different combinations of such additional currents seem to be characteristic of each particular type of mechanosensory neuron, and result in different patterns of action potential firing.

Mechanosensory neurons display a wide range of action potential encoding, from the relatively steady firing of muscle spindles and Ruffini endings to the brief, transient responses of Pacinian corpuscles as a mechanical stimulus is applied and withdrawn. A similar range is found in invertebrate receptors, even sometimes within one structure, such as the position and acceleration detectors of leg joint chordotonal organs, the two types of spider slit sensilla neurons, and crustacean stretch receptors. Although mechanical components and receptor current have often been implicated in the overall dynamics, as described earlier, voltage-activated currents also play major roles. In Pacinian corpuscles the action potential encoding is transient, even when most of the mechanical adaptation has been removed. In cockroach tactile spine neurons both voltage-activated and calcium-activated potassium currents seem to control the rapid reduction of firing following a step movement, while the rapid adaptation of spider slit sensilla neurons could be completely explained by the inactivation properties of the voltage-activated sodium channels. It seems that a wide variety of components have been used in different mechanoreceptors to control the relationship between mechanical stimulus and the resulting action potential train.

Following encoding, action potentials from mechanosensory neurons propagate into the central nervous system along nerve axons. In vertebrates, the cell bodies of these neurons are located in the dorsal root ganglia of the spinal cord, and the axons enter via the dorsal roots. Mechanosensory axons are generally in the Aα and Aβ groups of myelinated fibers with conduction velocities in the range of 40–120 m s^{-1}. Bare nerve endings are innervated by myelinated Aδ or unmyelinated C fibers, with conduction velocities in the range of 0.5–30 m s^{-1}. The major transmitter at the first central synapse in vertebrates is probably the excitatory amino acid glutamate. Arthropod cuticular mechanosensory neurons have their cell bodies in the periphery and send axons into the central nervous system via nerve roots of the segmental ganglia, which are sometimes fused into larger structures. Conduction velocities are typically 1–5 m s^{-1}. The dominant transmitter for arthropod mechanoreceptors is acetylcholine.

Efferent and Humoral Modulation of Mechanosensory Neurons

Evidence indicates that the sensory neurons of all vertebrate and invertebrate mechanoreceptors receive inhibitory efferent innervation of their centrally located axon terminals. Well-known examples are the efferent innervation of vertebrate muscle spindle and tendon organ terminals. Peripheral mechanoreceptors and pain receptors receive similar synaptic inputs onto their central projections. This type of presynaptic innervation of mechanosensory neurons is remarkably similar in all vertebrate and invertebrate species studied so far. The most common transmitter released from the efferent neurons is γ-aminobutyric acid (GABA), and its effects have been well characterized in insects, crustaceans, and vertebrates.

In addition to this inhibitory control of axon terminals, there is accumulating evidence that modulation of many mechanosensory neurons occurs at their distal sensory endings and cell bodies. The γ-innervation of muscle spindles and the inhibitory neurons of crustacean stretch receptors (Figure 3) provide vertebrate and invertebrate examples. Somatic innervation is indicated by transmitter receptors found in vertebrate dorsal root ganglion neurons. In the vertebrate periphery, unmyelinated nerve fibers have been seen associated with slowly adapting cutaneous mechanoreceptors, as well as inside the capsules of Pacinian corpuscles. Some of these endings are probably efferent and have been suggested to modulate sensory reception. There is also evidence that Merkel cells may function as secretory cells, rather than as mechanoreceptors, and modulate their associated nerve endings. However, the function of Merkel cells remains controversial. Modulation of vertebrate cutaneous mechanoreceptors and pain receptors by circulating factors also occurs. Sympathetic efferents have modulatory effects on muscle spindles, Pacinian corpuscles, several types of low-threshold skin mechanoreceptors, and pain receptors, and these neurons express adrenergic receptors. Rohon–Beard neurons, which are developmentally early amphibian and fish touch receptors, have serotonergic efferent innervation, and serotonin is suggested to alter their mechanical sensitivity by inhibiting both low- and high-voltage activated calcium currents.

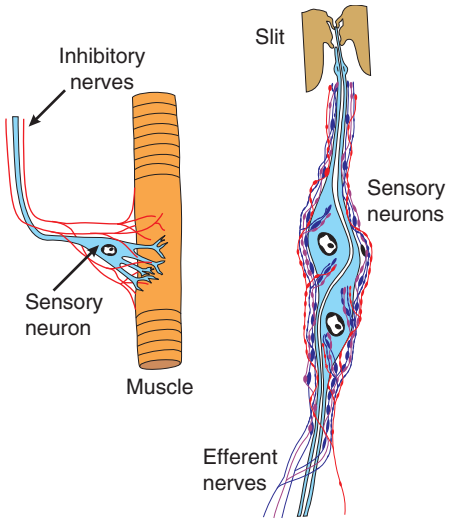

Figure 3 Invertebrate mechanoreceptors with well-characterized efferent modulation. Crustacean stretch receptors (left) are multipolar mechanoreceptors that, in this example, detect muscle stretch. Their sensory neurons (blue) are modulated by several efferent neurons that synapse onto their cell bodies and dendrites, as well as the muscle cells, releasing γ-aminobutyric acid and glutamate. Spider slit sensilla (right) contain pairs of mechanosensory neurons, the axons, cell bodies, and dendrites of which receive extensive innervation from at least four different efferent neurons. γ-Aminobutyric acid, glutamate, octopamine, and acetylcholine synaptic transmissions have been characterized.

Insect mechanosensory neurons are often modulated by biogenic amines, especially octopamine, and the locust dorsal unpaired median neurons, which secrete octopamine, have terminals in the periphery closely associated with sensory dendrites. Previous findings of octopamine effects in situations in which axon terminals were removed during recordings also suggested that octopamine receptors are located in the neurons' peripheral parts. Although octopamine modulation of insect mechanoreceptors is probably the most thoroughly studied neuromodulatory system, there are no clear conclusions about the mechanisms involved. Octopamine may increase or decrease spiking frequency, even in the same neuron.

Efferent innervation of the peripheral regions of mechanosensory neurons has been clearly demonstrated by morphological and immunocytochemical findings in arachnids and crustaceans (Figure 3). Both groups of sensory neurons have dense efferent innervation, forming several types of synapses on all parts of the sensory neurons, including their somata and dendrites. GABAergic inhibition of crustacean stretch receptor neurons is mediated by ionotropic anion receptors. Slit sensilla and other cuticular mechanoreceptors of spiders have at least four different types of synaptic structures on the efferent terminals, based on the shapes and sizes of synaptic vesicles and immunogold labeling. Several chemical transmission systems have been identified, including GABAergic inhibition via both ionotropic and metabotropic receptors, glutamatergic inhibition, octopaminergic excitation, and cholinergic inhibition. The effects of these different types of synaptic modulation on mechanoreception are not well understood, but could include control of the overall sensitivity and dynamic responses of the neurons or more selective tuning to specific signals. It seems likely that both vertebrate and invertebrate central nervous systems are able to rapidly adjust the behavior of many types of mechanosensory neurons, but the functional significance of this widespread control is not yet established.

Conclusions

While the basic steps of mechanoreception are well established, each of the three major stages (coupling, transduction, and encoding) continues to present major challenges that require further research. The functional significance of the wide range of coupling structures and the mechanical properties of their individual components remain largely unknown. We have yet to clearly identify any of the proteins involved in mechanotransduction, the channel-opening mechanisms, or their connections to other structures. Encoding is well understood in some preparations, but its wide variability and locations within complex neuronal geometries make it difficult to generalize. Finally, the efferent and humoral modulation of mechanosensory neurons in the periphery, although not completely novel, has recently proved surprisingly widespread and complex, with little information yet about its central organization or functional roles.

See also: Action Potential Initiation and Conduction in Axons; Hair Cells: Sensory Transduction; Ion Channel Localization in Axons; Mechanosensory Transduction; Presynaptic Inhibition; Tactile Coding in Peripheral Neural Populations; Transient Receptor Potential (TRP) Channels.

Further Reading

Barth FG (2001) *A Spider's World. Senses and Behavior.* New York: Springer.

Bell J, Bolanowski SJ, and Holmes MH (1994) The structure and function of Pacinian corpuscles: A review. *Progress in Neurobiology* 42: 79–128.

Fabian-Fine R, Seyfarth E-A, and Meinertzhagen IA (2002) Peripheral synaptic contacts at mechanoreceptors in arachnids and crustaceans: Morphological and immunocytochemical characteristics. *Microscopy Research and Technique* 58: 283–298.

Foelix RF (1975) Occurrence of synapses in peripheral sensory nerves of arachnids. *Nature* 254: 146–148.

French AS, Torkkeli PH, and Seyfarth E-A (2002) From stress and strain to spikes: Mechanotransduction in spider slit sensilla. *Journal of Comparative Physiology A* 188: 739–752.

Gingl E and French AS (2003) Active signal conduction through the sensory dendrite of a spider mechanoreceptor neuron. *Journal of Neuroscience* 23: 6096–6101.

Goodman MB and Schwarz EM (2003) Transducing touch in *Caenorhabditis elegans*. *Annual Review of Physiology* 65: 429–452.

Lewin GR and Moshourab R (2004) Mechanosensation and pain. *Journal of Neurobiology* 61: 30–44.

Malinovsky L (1996) Sensory nerve formations in the skin and their classification. *Microscopy Research and Technique* 34: 283–301.

Montell C (2005) *Drosophila* TRP channels. *Pflugers Archiv – European Journal of Physiology* 451: 19–28.

Sachs F (1997) Mechanical transduction by ion channels: How forces reach the channel. In: Froehner SC (ed.) *Cytoskeletal Regulation of Membrane Function*, pp. 209–218. New York: Rockefeller University Press.

Swerup C and Rydqvist B (1992) The abdominal stretch receptor organ of the crayfish. *Comparative Biochemistry and Physiology A* 103: 423–431.

Syntichaki P and Tavernarakis N (2004) Genetic models of mechanotransduction: The nematode *Caenorhabditis elegans*. *Physiological Reviews* 84: 1097–1153.

Torkkeli PH and Panek I (2002) Neuromodulation of arthropod mechanosensory neurons. *Microscopy Research and Technique* 58: 299–311.

Widmer A, Höger U, Meisner S, et al. (2005) Spider peripheral mechanosensory neurons are directly innervated and modulated by octopaminergic efferents. *Journal of Neuroscience* 25: 1588–1598.

Mechanosensory Transduction

D Grundy, The University of Sheffield, Sheffield, UK
S J H Brookes, Flinders University, Adelaide, SA, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Viscerosensory neurons are specialized to detect changes within the lumen or wall of visceral organs, and through the generation of trains of action potentials convey the coded information that keeps the central nervous system (CNS) abreast of events in internal organs, such as the bladder, bowel, or cardiovascular system. Most of this information goes unperceived but is used instead to trigger reflexes and behaviors that coordinate visceral function and maintain homeostasis. However, some sensory afferents also mediate sensations such as bladder fullness and urge to defecate. Some also mediate pain, which tends to be vague and poorly localized, often being referred to somatic sites (dermatomes) because of the way visceral and somatic sensory inputs converge in the spinal cord. This article discusses the transduction mechanisms responsible for signaling reflex and sensory function, with an emphasis on the gastrointestinal (GI) tract; because of its dual sensory innervation, the GI tract has functional subpopulations of visceral afferents that are specialized to signal low-intensity (physiological) and high-intensity (nociceptive) information.

Sensory Innervation to the GI Tract

Sensory neurons innervating the GI tract terminate within the gut wall in various layers, including muscle, mucosal epithelia, and enteric ganglia. Other endings terminate in the serosa and mesenteric attachments and form a dense network around mesenteric blood vessels and their tributaries in the gut wall. These various ending maintain a steady flow of afferent traffic to the CNS, relaying information on activity both within and outside the gut wall. This sensory information is conveyed to the CNS by separate vagal and spinal pathways to the brain stem and spinal cord, from where information is processed and projected to higher brain areas. Spinal afferents can also be subdivided into those running via the splanchnic nerves and those in the pelvic nerves.

Vagal and pelvic afferents generally process physiological information (e.g., the nature and composition of the luminal contents, the presence and amplitude of ongoing motor activity of the gut). In contrast, splanchnic afferents also process pathophysiological information, including potentially noxious mechanical or chemical stimuli arising through tissue injury, ischemia, and inflammation. However, there is a degree of overlap in the sensitivity of these two pathways, and there is also some interplay between them.

Sensory Endings

A majority of GI afferents terminate within the gut wall as bare nerve endings. Thus the differential sensitivity of afferents arises from their location in the gut wall, their relationship with other structures, and the receptors and ion channels that they express.

Mucosal Afferents

Afferent endings in the mucosa are in close association with the lamina propria adjunct to the mucosal epithelium, but are never exposed directly to the contents of the lumen. Mucosal afferents are chemosensitive, responding to the chemical nature of luminal contents. They do not respond to distension, but are exquisitely sensitive to local mechanical stimulation that deforms the mucosal epithelium.

Muscle Afferents

Vagal afferent endings within the muscle layers and associated myenteric plexus have been classified into two types: intramuscular arrays (IMAs) and intraganglionic laminar endings (IGLEs). IMAs are distributed within the muscle sheets and appear to make direct contact either with the muscle fibers or with intramuscular interstitial cells of Cajal (ICCs), which may play a role in mechanotransduction, responding to stretch or contraction of the muscle bundles. In contrast, IGLEs consist of flattened endings that branch extensively inside the connective tissue sheath surrounding myenteric ganglia. Their role had been obscure until recent studies showed that they, but not their parent axons, were sensitive to mechanical stimulation. Since IGLEs are located between the circular and longitudinal muscle layers, they are exposed to shearing forces generated during muscle stretch or contraction and have been suggested to convey mechanosensory information relevant to distension and contraction of the bowel wall. Vagal IGLEs are mainly found in the esophagus and stomach. A similar type of ending has been identified in the rectum. The latter types of endings are not derived from vagal afferents, but are likely to arise from afferents running in pelvic nerves.

Serosal and Mesenteric Afferents

Spinal afferent terminals are distributed throughout the gut wall but are also located in the serosa and mesenteric attachments, often associated with mesenteric blood vessels. Their location and response characteristics suggest that spinal afferents respond to distortion of the viscera during distension and contraction. However, spinal afferent terminals are also found in the mucosa. These spinal afferents respond to mechanical stimulation and to the chemical environment within the lamina propria, and in particular respond to the changing chemical milieu following injury, ischemia, or infection.

Stimulus–Response Function

The relationship between the intensity of stimulation and the degree of activation of the sensory afferent is known as the stimulus–response function. Afferent information conveyed by spinal, pelvic, and vagal mechanosensitive afferents is somewhat different in its sensory–response function, as revealed by direct electrophysiological recordings of afferent traffic en route to the CNS. The factors that determine sensitivity may include the location of the endings within the bowel wall, the latter being influenced by the extent to which mechanical forces are distributed and dissipated by nonneural structures in the bowel wall. Different mechanosensitive channels may also contribute to their different stimulus–response functions, but these have not been identified.

When the bowel is distended, vagal and pelvic muscle mechanoreceptors have low thresholds for activation, responding to pressure rises of just a few mmHg, and reach maximal responses within physiological levels of distension. These are referred to as low-threshold mechanoreceptors and convey information relating to normal physiological events in the bowel. These endings also respond during contraction and therefore signal to the CNS information relevant to each and every contraction that occurs anywhere along the length of the GI tract. Other spinal afferents also have low thresholds for activation but continue to respond beyond the physiological range and thus encode both physiological and noxious levels of stimulation; these are called wide-dynamic-range fibers. These spinal endings can contribute to signaling visceral pain through some intensity code that recognizes extreme levels of distension or contraction. Other spinal afferents, particularly those with endings in the serosa and mesenteric attachments, respond only to noxious levels of distension (high-threshold mechanoreceptors) and are referred to as nociceptors. These only respond when the bowel is overdistended or during powerful contractions that distort the bowel. An extreme example of mechanical nociception has been described in sensory endings that fail to respond even to extreme levels of distension. These are the so-called sleeping or silent nociceptors that can be awakened to become mechanically sensitive under conditions of injury or inflammation (see later). The latter illustrate the fact that mechanosensitivity is not fixed either in terms of threshold for activation or gain in the stimulus–response relationship, and as such the threshold can be reduced and the gain increased after injury or during inflammation. This leftward shift in the stimulus–response function means that for a given level of stimulation, there is a greater afferent barrage generated. This can give rise to altered perception such that stimuli that are normally innocuous can cause pain while the response to a painful stimulus becomes exaggerated (hyperalgesia).

Transduction by Visceral Afferents

Transduction is the process by which sensory neurons generate trains of action potentials from an adequate stimulus. The conventional understanding is that mechanisms in the neuron produce a generator potential, which in turn evokes a train of action potentials, encoding details of the stimulus such as the timing, rate of onset, amplitude, and duration. While specific transduction mechanisms probably underlie the generator potential, the final pattern of firing is also influenced by accessory mechanisms, including various voltage- and calcium-dependent ion channels. This is analogous to synaptically evoked firing in which the patterned output of action potentials is not simply a frequency code of transmitter-driven membrane potential, but is also powerfully influenced by the density of voltage-dependent sodium, calcium, and potassium channels, by calcium-dependent potassium channels, and potentially conductances from many other sources, especially transient receptor potential (TRP) channels of which TRP vanilloid 1 (TRPV1) has been best characterized in visceral afferents. In most visceral afferent neurons it is not possible to record directly the generator potential, due to the small dimensions of the mechanosensitive endings in peripheral tissues.

Mechanoreceptors at different layers of the gut wall are specialized to encode different aspects of physical stimuli imposed on the gut tissue. It is likely that different transduction mechanisms may be employed by different mechanoreceptors. For visceral mechanoreceptors, there are two major mechanisms which could underlie the generator potential evoked by local distortion of nerve endings. The first is 'indirect,' or 'chemical,' transduction in which another cell type, probably nonneuronal, is activated by the stimulus and which then releases a transmitter or

mediator onto the sensory axon, causing it to fire. Examples of such mechanisms may include Merkel receptors in the skin, hair cells in the inner ear, and glomus cells for some chemosensitive visceral afferents. In contrast, 'direct' transduction occurs when the nerve ending itself contains all of the essential elements required to convert the stimulus into depolarization and firing.

Indirect Mechanotransduction

The transduction in mucosal mechanoreceptors seems to be better explained by an indirect mechanism involving release of mediators than by a direct mechanism involving mechanosensitive ion channels on nerve endings. Mucosal mechanoreceptors do not readily respond to distension (unless there is deformation of the mucosa as a result of over distension) but yet are very sensitive to light stroking of the mucosa, suggesting the transduction site might be on the luminal border of the epithelial.

Thus the transduction process does not occur at the level of the sensory nerve ending itself, but rather in a specialized sense cell in close association with the afferent ending. In this case, the afferent neuron does not respond directly to a particular stimulus, but does so following the release of a mediator from another cell which functions as the sensory detector. This putative mechanism of mechanotransduction is attractive for GI afferents because of both the close association between afferents and other cell types (e.g., mast cells, ICCs, enteric neurons, epithelial cells, enterochromaffin cells) and the expression of a variety of transmitter receptors on the afferent terminals. In this respect particular attention has been paid to enteroendocrine cells in the mucosal epithelium. These cells have an apical tuft of microvilli exposed to the intestinal lumen; this tuft is proposed to be the site of sensory transduction, and in response to an appropriate stimulus, releases the contents of storage granules across the basolateral membrane to stimulate afferent terminals in close proximity within the lamina propria, via a paracrine mechanism. Much recent attention has focused on two such mediators, namely, ATP and 5-hydroxytryptamine (5-HT), which have been shown to play important roles in sensory signaling.

Mechanosensitive Release of Mediators

Mechanosensitive release of 5-HT from the gut wall was classically described by Bulbring in the 1950s. Since then there has been a wealth of literature describing a role for 5-HT in enteric reflex control. However, the role of 5-HT in mechanotransduction in vagal and spinal mechanoreceptors is more controversial, although data are accumulating indicating that under circumstances, such as inflammation, when EC cell hyperplasia occurs and 5-HT reuptake mechanisms are blunted, there is augmented sensitivity to 5-HT, which may impact on sensory signaling. 5-HT can act on a number of G-protein-coupled receptors as well as on the ligand-gated 5-HT_3 channel. Immunohistochemical evidence suggests that 5-HT_3 receptors are present on both dorsal root ganglion (DRG) and nodose ganglion cells that innervate the GI tract.

Burnstock has hypothesized that stretch stimuli in tubular organs may release ATP from epithelial cells and that this then activates sensory nerve terminals through interacting with purinergic receptors. ATP can act either on P2Y receptors, which are G-protein-coupled receptors, or on P2X receptors, which are ligand-gated ion channels. Immunohistochemical evidence exists for P2X receptors ($P2X_3$) on spinal afferents innervating the colon, and on vagal endings ($P2X_{2,3}$) innervating the esophagus and stomach. A role for ATP in sensory signal transduction has been best characterized for the bladder, where urethelial cells release ATP on distension; also, filling-induced bladder afferent discharge and micturition reflex can be blocked by P2 receptor antagonists or when the $P2X_3$ receptor is knocked out. Several lines of evidence also support the purinergic hypothesis for mechanotransduction in the GI tract: (1) mechanosensitive release of ATP from the gut wall has been demonstrated, (2) P2 receptors have been identified on nodose and DRG neurons, (3) exogenous ATP has been shown to activate extrinsic afferent endings, and (4) P2 receptor antagonists reduce pelvic afferent response to colorectal distension in the rat.

Direct Mechanotransduction

Direct mechanosensitivity arises because of the presence of mechanically sensitive ion channels in the nerve terminal membrane. Mechanical deformation of the nerve ending leads to the opening or closing of ion channels that allow charged molecules to pass in or out of the cell, which in turn leads to an alteration in the excitability of the nerve terminal. A variety of mechanosensitive ion channels have been identified, including a class of receptors exemplified by the epithelial amiloride-sensitive sodium channel (ENaC)/degenerin family. Recent data have implicated members of the acid-sensing ion channel (ASIC) family in the modulation of mechanosensory transduction. However, at present, the ion channels and receptors that underlie mechanosensitivity in the GI tract remain unknown.

Identifying the sites at which mechanotransduction occurs has provided some useful insights into afferent function. This has been achieved by locating transduction sites using focal mechanical stimuli (for example, fine probes or von Frey hairs) and marking them for subsequent visualization of the axonal structures associated with mechanosensory 'hotspots.' An extensive description of specialized sensory endings in the skin has been developed using this approach. In the gut, similar studies have been attempted, using anterograde labeling of sensory axons from the extrinsic nerve trunks in which they have been recorded, thus distinguishing their axons from the numerous axons of enteric, sympathetic, and parasympathetic neurons which would otherwise confound identification of sensory endings. Low-threshold vagal mechanoreceptors in the stomach and esophagus have been shown to have mechanotransduction sites corresponding to the IGLEs in myenteric ganglia. They are also exquisitely sensitive to localized distortion of the gut wall with light von Frey hairs. A similar type of ending has been shown to exist for low-threshold, slowly adapting pelvic afferents supplying the rectum. Interestingly, these types of endings are restricted to the rectum and very distal colon, where reflexes necessary for defecation are localized. IGLEs, therefore, represent a class of GI mechanosensory endings involved in transducing low-intensity information.

Transduction Mechanisms

The location of IGLEs in proximity to enteric neurons raises the possibility that they may be activated indirectly following release of chemicals into the synaptic neuropil within the myenteric plexus. To distinguish between 'indirect' and 'direct' mechanisms of transduction, the time course of transduction gives useful insight because 'indirect' mechanisms are likely to be relatively slow – a receptor cell has first to be activated, then must release a chemical mediator, which must then change the membrane potential of the sensory nerve ending to evoke firing. The transduction delays for vagal and rectal low-threshold mechanoreceptors have been measured by subtracting the conduction delay (from an identified transduction site) from the latency of the first action potential evoked by a mechanical stimulus with a sharp onset. For both vagal and rectal mechanoreceptors, the transduction delay required for IGLEs to convert mechanical deformation into an action potential was a few milliseconds. This is rather faster than might be expected for chemical transduction, unless this was mediated via very rapid release of the responsible mediator, which requires active exocytosis. Since exocytotic release of mediators is a calcium-dependent mechanism, the demonstrations that transduction is preserved in solutions with low extracellular calcium and elevated magnesium ion concentrations show that exocytotic release of mediators was not obligatory for mechanotransduction by either vagal or rectal IGLEs. Thus both types of mechanoreceptor probably express the stretch-activated ion channels in their endings and use 'direct' mechanisms of mechanotransduction, although the specific ion channels involved have not yet been identified.

Further support for direct mechanotransduction comes from pharmacological study of mechanosensitivity. A range of compounds known to block stretch-activated ion channels have been applied to both vagal IGLEs and rectal IGLEs. Thus, benzamil, a blocker of a variety of stretch-activated ion channels, reduces mechanosensitivity. Likewise, the TRP channel blocker SKF96,365 also reduces mechanosensitivity by both vagal and rectal mechanoreceptors. However, many such drugs show poor selectivity for stretch-activated ion channels and interpreting their effects is fraught with difficulty. For example, gadolinium ions (Gd^{3+}) have been widely used to block stretch-activated ion channels, although they are known to affect a variety of other ion channels, including voltage-activated sodium channels. While Gd^{3+} certainly reduces the excitability of rectal IGLEs to rectal distension, the effects are entirely accounted for by changes in muscle compliance. Indeed, responses to von Frey hairs, which distort transduction sites in a manner that is largely independent of muscle tone, were not affected by gadolinium.

Mechanonociception

Another major class of mechanoreceptors, which typically have higher activation thresholds, has been described as having endings associated with arterial branch points in the mesentery and serosa of the gut, and in similar locations for many other viscera, including spleen, pancreas, uterus, and ovaries. These mechanoreceptors are considered to be important in nociception but little is known about their mechanisms of transduction. They are well established to respond to a variety of chemical stimuli released during tissue damage (see later). In addition, they may also be activated by ATP, glutamate, and 5-HT; mediators possibly released under more physiological conditions. Despite this, there is little evidence that any of these mediators is essential for transduction, although they are all likely to be involved under pathophysiological conditions. It is well established that distension of the naive (uninflamed) gut causes pain in both humans and animals. The nature of the adequate stimulus for these mechanoreceptors is not well established: they are certainly activated by large distensions of the bowel, but

whether they detect intramural stress or strain is not currently clear. It is also possible that they detect changes in intestinal hemodynamics, responding perhaps to reductions in blood flow due to collapse of vessels when intramural tension exceeds vascular pressure. Such a mechanism may be mediated indirectly, perhaps by release of lipid mediators. Given that pain is such an important symptom of visceral disorders, the importance of these questions is hard to overrate.

Modulation of Visceral Afferent Excitability

In addition to transduction mechanisms, substantial modulation of the excitability of peripheral sensory endings is known to occur, especially in visceral afferents. Modulation is mediated by a wide variety of substances released from nearby and distant cells, including neurotransmitters, cytokines, and autacoids. While it is useful to distinguish transduction from modulation, it should be borne in mind that these mechanisms probably form a continuum. When afferents show spontaneous firing in the absence of a specific stimulus, local release of modulatory substances can directly cause an increase in firing frequency without activating transduction mechanisms.

Sensitization

There is an enormous range of chemical mediators that influence the sensitivity of visceral afferents. Such mediators are usually released under conditions of inflammation, injury, or ischemia from a plethora of cell types (e.g., platelets, leukocytes, lymphocytes, macrophages, mast cells, glia, fibroblasts, blood vessels, muscle cells, and neurons). Each of these specific cells (e.g., mast cells) may release several of these modulating agents, some of which may act directly on the sensory nerve terminal, while others may act indirectly, following release of other agents from other cells, in a series of cascades.

Electrophysiological, immunocytochemical, and molecular biological techniques have revealed the functional expression of receptors to various mediators on the cell bodies of visceral sensory neurons in the dorsal root or nodose ganglia or on their processes in the gut wall. These diverse mediators, which include amines, purines, prostanoids, protease, cytokines, etc., produce their effects on visceral afferent nerves by three distinct processes: (1) by direct activation, which ultimately involves the opening of ion channels present on the nerve terminals, (2) by sensitization, which may occur in the absence of a direct stimulation, but which usually results in afferent hyperexcitability to both chemical and mechanical modalities (this may arise following G-protein-coupled alterations in second-messenger systems, which often lead to phosphorylation of membrane receptors and ion channels that control excitability), and (3) by altering the phenotype of the afferent nerve (e.g., through alterations in the expression of mediators, channels, and receptors, or modulating the activity of these by changing the ligand-binding characteristics or coupling efficiency of other receptors). Neurotrophins, in particular nerve growth factor and glial-derived neurotrophic factor, influence different populations of visceral afferents and play an important role in adaptive responses to nerve injury and inflammation. Voltage-gated sodium channels are one example of a family of ion channels, some of which play an important role in regulating pain thresholds. Under these conditions the afferents are sensitized and this process is believed to be the underlying mechanism of the hypersensitivity observed in some clinical conditions.

Summary

In conclusion, visceral afferents make a number of distinctive, specialized endings in the periphery, each of which is probably tuned to a different afferent stimulus. The afferents have specific mechanisms for transduction of stimuli, but their excitability is also modulated by a wide variety of other chemical mediators released from a range of other cell types under physiological or pathophysiological conditions. While some of these mechanisms are reasonably understood, much remains to be determined about the key factors that control the excitability of visceral afferent nerve fibers.

See also: Adenosine Triphosphate (ATP); Adenosine Triphosphate (ATP) as a Neurotransmitter; Mechanoreceptors; Motor Control of Feeding and Drinking; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation.

Further Reading

Berthoud HR, Blackshaw LA, Brookes SJ, et al. (2004) Neuroanatomy of extrinsic afferents supplying the gastrointestinal tract. *Neurogastroenterology & Motility* 16: 28–33.

Beyak MJ, Bulmer DC, Jiang W, et al. (2006) Extrinsic sensory afferent nerves innervating the gastrointestinal tract. In: Johnson LR (ed.) *Physiology of the Gastrointestinal Tract*. Amsterdam: Elsevier.

Brierley SM, Jones RC III, Gebhart GF, et al. (2004) Splanchnic and pelvic mechanosensory afferents signal different qualities of colonic stimuli in mice. *Gastroenterology* 127: 166–178.

Gebhart GF (2000) Pathobiology of visceral pain: Molecular mechanisms and therapeutic implications. IV: Visceral afferent contributions the pathobiology of visceral pain. *American Journal of Physiology – Gastrointestinal and Liver Physiology* 278: G834–G838.

Julius D and Basbaum AI (2001) Molecular mechanisms of nociception. *Nature* 413: 203–210.

Kirkup AJ, Brunsden AM, and Grundy D (2001) Receptors and transmission in the brain–gut axis: Potential for novel therapies. I: Receptors on visceral afferents. *American Journal of Physiology – Gastrointestinal and Liver Physiology* 280: G787–G7894.

Lynn PA, Zagorodnyuk VP, Hennig GW, et al. (2005) Mechanical activation of rectal intraganglionic laminar endings of the guinea pig distal gut. *Journal of Physiology* 564: 589–601.

Page AJ, Brierley SM, Martin CM, et al. (2005) Different contributions of ASIC channels 1a, 2, and 3 in gastrointestinal mechanosensory function. *Gut* 54: 1408–1415.

Raybould HE, Cooke HJ, and Christofi FL (2004) Sensory mechanisms: Transmitters, modulators and reflexes. *Neurogastroenterology & Motility* 16: 60–63.

Rong W, Hillsley K, Davis JB, et al. (2004) Jejunal afferent nerve sensitivity in wild-type and TRPV1 knockout mice. *Journal of Physiology* 560: 867–881.

Vlaskovska M, Kasakov L, Rong W, et al. (2001) P2X3 knock-out mice reveal a major sensory role for urothelially released ATP. *Journal of Neuroscience* 21: 5670–5677.

Wood JN (2000) Pathobiology of visceral pain: Molecular mechanisms and therapeutic implications. II: Genetic approaches to pain therapy. *American Journal of Physiology – Gastrointestinal and Liver Physiology* 278: G507–G512.

Wynn G, Rong W, Xiang Z, et al. (2003) Purinergic mechanisms contribute to mechanosensory transduction in the rat colon. *Gastroenterology* 125: 1398–1409.

Zagorodnyuk VP, Chen BN, Costa M, et al. (2003) Mechanotransduction by intraganglionic laminar endings of vagal tension receptors in the guinea pig oesophagus. *Journal of Physiology (London)* 553: 575–587.

Zagorodnyuk VP, Lynn P, Costa M, et al. (2005) Mechanisms of mechanotransduction by specialized low-threshold mechanoreceptors in the guinea pig rectum. *American Journal of Physiology – Gastrointestinal and Liver Physiology* 289: G397–G406.

Medulloblastoma

D Kombogiorgas and S Sgouros, Birmingham Children's Hospital, Birmingham, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term medulloblastoma was first introduced by Bailey and Cushing in 1925. They used this term to describe a very cellular tumor apparently arising from the roof of the fourth ventricle and projecting into the center of the cerebellum and the cavity of the fourth ventricle. Terms such as neurocytoma, neuroblastoma, sarcoma, and spongioblastoma cerebelli had been used previously occasionally for this histological type. In 1973, Hart and Earle used the term primitive neuroectodermal tumor (PNET) to refer to supratentorial malignant small-cell tumors that were believed to arise from primitive embryonal cells. Ten years later, Rorke and Becker and Hinton suggested separately that all malignant small-cell tumors, including medulloblastoma, should be classified as PNETs and subdivided by location. They suggested the term medulloblastoma for the undifferentiated neuroepithelial round-cell tumors located in the cerebellum. Currently, based on World Health Organization (WHO) classification, the diagnosis of PNET is a generic term for cerebellar medulloblastomas and for neoplasms morphologically indistinguishable from medulloblastomas but located at other sites in the central nervous system (CNS).

Epidemiology

Brain and spinal cord tumors are the second most frequent type of cancer after leukemia in children under 15 years of age. Medulloblastoma is the most frequent brain tumor for both sexes combined after astrocytomas, and it constitutes approximately 20% of childhood cerebral tumors, with a reported male-to-female ratio of 1.5–2:1 and an incidence of two to six cases per million population per year in various countries. In children, medulloblastomas represent 33% of the tumors of the posterior cerebral fossa and 40% of the cerebellar tumors. Mean age at diagnosis is around 7 years of age with peaks at 3 and 7 years.

Histopathology

Classification and Grading

Embryonal (primitive) neuroepithelial neoplasms make up 4–5% of CNS neoplasms. Their features include:

- Predominance in children.
- A tendency to disseminate through cerebrospinal fluid (CSF) pathways.
- A dominant population of small undifferentiated cells.
- High mitotic indices and widespread apoptosis.
- A potential for divergent neuroepithelial differentiation.

Medulloblastoma is a poorly demarcated, pink-purple, soft, friable mass that arises in the cerebellar vermis, usually from the inferior medullary velum (see Figure 1). The tumor may extend through the fourth ventricle into the cerebral aqueduct or into the cisterna magna via the foramen of Magendie. It may invade the cerebellum, brain stem, or cerebellar peduncle. Involvement of the cerebellar hemispheres is uncommon in children, but is more frequent in adults. Brain stem infiltration occurs in 15–36% of patients. There may be foci of hemorrhage or necrosis, but cysts are unusual.

Histologically, childhood and adult medulloblastoma are identical, being highly cellular, malignant, invasive tumors corresponding to WHO malignancy grade IV. They consist of densely packed tumor cells with round to oval or carrot-shaped hyperchromatic nuclei with scanty cytoplasm, high mitotic and apoptotic rates, and usually neuroblastic rosettes in some areas. Microvascular proliferation is relatively uncommon (see Figure 2).

According to the 2002 WHO classification of tumors of the nervous system, the medulloblastoma (i.e., medulloblastoma (classic) (9470/3)) has four variants:

- Desmoplastic medulloblastoma (9471/3).
- Large-cell medulloblastoma (9474/3).
- Medullomyoblastoma (9472/3).
- Melanotic medulloblastoma (9470/3).

The most common variant is the desmoplastic medulloblastoma, which occurs particularly in older children and adults. It constitutes approximately 20% of adult and 10% of childhood medulloblastomas. It tends to be located within the cerebellar hemisphere and is often better demarcated and firmer than other variants, as a result of the greater connective tissue component. It is distinguished from classic medulloblastoma by the presence of an intercellular stromal network, consisting of reticulin and collagen that separates groups of typical medulloblastoma cells into islands. Nodular or desmoplastic medulloblastoma contains centers of dark cells surrounding pale islands, analogous to the germinal centers of a

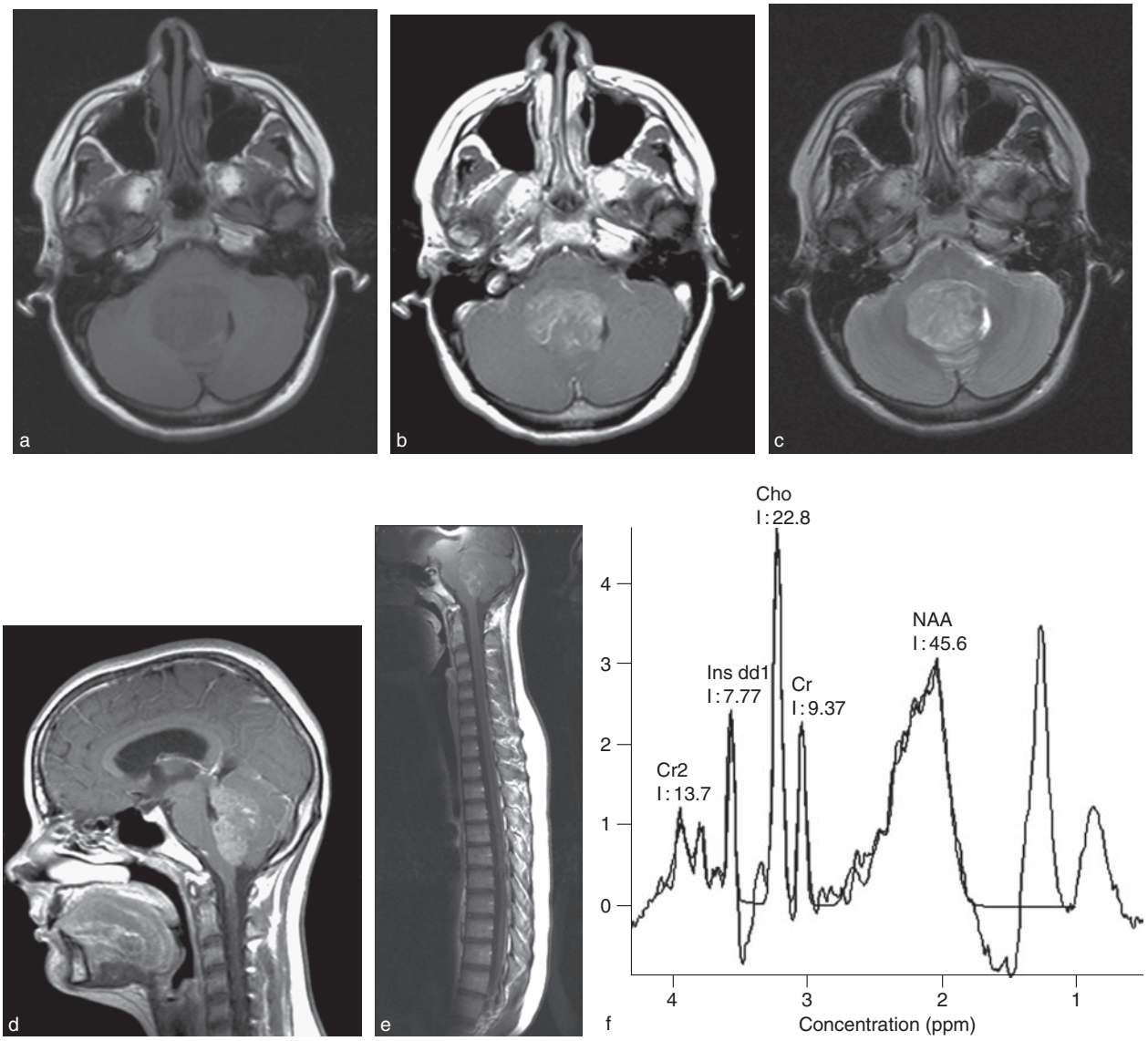

Figure 1 Magnetic resonance (MR) scans of a 12-year-old girl who presented with a 4-week history of headaches, drowsiness and vomiting and was found to have medulloblastoma: (a) brain axial T1-weighted image without contrast; (b) brain axial T1-weighted image with contrast; (c) brain axial T2 weighted image; (d) brain sagittal T1-weighted image with contrast; (e) spinal sagittal T1-weighted image with contrast; (f) brain MRS spectrum. There is a midline tumor in the posterior fossa, arising from the cerebellar vermis and projecting into the fourth ventricle cavity, which it fills almost entirely (a–d). The tumor is enhanced with gadolinium contrast intravenous injection in (b) and (d). It is well circumscribed and is not causing significance edema in the surrounding brain (b, c). The third ventricle is dilated due to obstructive hydrocephalus caused by the tumor (d). In the spinal views, there is a faint film of enhancement on the pial surface of the spinal cord (e), which denotes the presence of metastatic disease. In the spectroscopy analysis (f), there is a high Cho peak and comparatively lower NAA peak, giving a low NAA:Cho ratio. Cho, choline; Cr, creatine; Cr2, creatine ethyl ester; I:, integral; Ins dd1, inositol; MRS, magnetic resonance spectroscopy; NAA, *N*-acetyl aspartate.

lymph node. It is associated with better prognosis than the classic histological subtype in children, but the converse is true in adults.

Differentiated medulloblastomas with neuroblastic features are distinguished by Homer Wright rosettes. These rosettes are formed by cells surrounding fibrillary structures and characteristically do not encircle blood vessels. The rosettes of neuroblastic medulloblastoma are glial fibrillary acidic protein (GFAP)-negative.

Large-cell medulloblastomas have been characterized recently as a highly malignant variant and represent approximately 4% of medulloblastomas. Commonly, there are findings of large areas of necrosis, high mitotic activity, and high apoptotic rate. They have high tendency for CSF dissemination

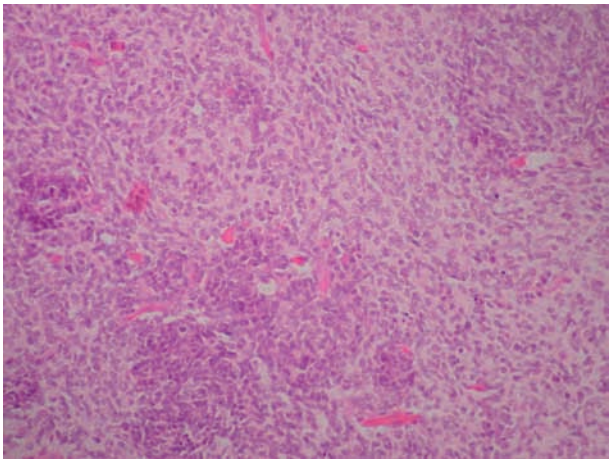

Figure 2 Histological slide of a classic medulloblastoma. There are densely packed tumor cells with round to oval-shaped hyperchromatic nuclei with scanty cytoplasm, some mitoses, and some neuroblastic rosettes. Heamatoxylin & Eosin stain, 200×. Photograph courtesy of Dr. Marie-Anne Brundler, Birmingham Children's Hospital, UK.

and also may give lymph node metastasis. A pleomorphic large-cell variant of medulloblastoma is also recognized and has a poor prognosis.

Medullomyoblastomas are rare embryonal tumors of the CNS, arise exclusively in the cerebellum and occur almost always in children younger than 10 years of age. Melanotic medulloblastomas are very rare variants of medulloblastoma with pigmented cells which contain melanin. They consist of anaplastic cells with slight neuronal differentiation, but without the fibrillary background of neuroblastomas, and epithelial islands pigmented with melanin. The latter participate in the spread through the subarachnoid space, but do not extend beyond the dura. Both focal melanin production and evidence of striated muscle differentiation can occur in medulloblastomas, creating the variants melanotic medulloblastoma and medullomyoblastoma.

Staging

Medulloblastoma has a strong propensity to metastasize within the subarachnoid space with microscopic deposits, gross nodules, or en plaque masses, due to the tendency of the tumor to penetrate the ependymal surface. Autopsy series have revealed that cerebrospinal spread occurs in up to 50% of patients. Extraneural metastasis occurs in approximately 5% of patients and most commonly involves the bone marrow.

Chang and colleagues developed an operative staging system based on the TNM (tumor, spread to lymph nodes, and metastasis)-staging designation for the patients with medulloblastoma in 1969 (see Table 1). Because there are no lymphatic channels in the brain, they omitted the class N in their staging system. This system has been in common use worldwide since its introduction.

In 1985, Laurent and colleagues reported a postoperative MAPS (metastasis, age, pathology, and surgery) classification system (see Table 2) to evaluate mainly the prognostic impact of the extent of a residual tumor after the initial definitive surgery as it is assessed by postoperative computerized tomography (CT) scan and to analyze these data in a prospective manner. However, it remains unclear why and how they selected the number 1.5 cm as cut-off for the tumor diameter in the residual tumor-staging classification.

To establish an accurate assessment of disease risk among children with medulloblastoma, Gilbertson and colleagues found that high-molecular-risk patients can be designated as those with high ErbB2 receptor expression ($\geq 50\%$ ErbB2 positive tumor cells) and/or isolated 17p loss, whereas low-molecular-risk patients can be designated as those whose tumors demonstrated neither aberration. They observed no significant relationship between patient age and any molecular abnormality.

Molecular Genetics

Genetic aberrations are believed to contribute to the pathogenesis of medulloblastomas. Approximately 10% of the medulloblastomas carry double minute chromosomes as a cytogenetic sign of amplification of the *C-myc* or *N-myc* genes. Loss of chromosome 17p is the most frequent finding in medulloblastomas (present in up to 50% of the cases), in many tumors related to the occurrence of an isochromosome 17q, suggesting the presence of one or more medulloblastoma-related tumor-suppressor genes on chromosome arm 17p. Interestingly, chromosome 17 alterations are mostly present in medulloblastomas of the classic (nonnodular) type, but are absent in most nodular/desmoplastic-type medulloblastomas. In the latter type, loss of chromosome 9q and inactivation of the Patched (Ptch) tumor-suppressor gene have been described, leading to the hypothesis that there are two distinct entities of medulloblastomas. Other chromosomal aberrations include loss of l0q (in 35% of patients).

Familial adenomatous polyposis (FAP) syndrome arises from mutations in the APC (5q) gene and is associated with medulloblastoma formation. Also, Rubinstein–Taybi syndrome, a condition characterized by short stature, beaked nose, malformed ears, angulated thumbs and toes, mental retardation, and undescended testicles, has also been found to be associated with medulloblastoma. Many of those patients have a

Table 1 Chang classification of medulloblastoma

	Description
T stage	
T1	Tumor <3 cm in diameter and limited to the classic midline position in the vermis, the roof of the fourth ventricle, and less frequently to the cerebellar hemispheres
T2	Tumor >3 cm in diameter, invading one adjacent structure or partially filling the fourth ventricle
T3a	Tumor further invading two adjacent structures or completely filling the fourth ventricle with extension into the aqueduct of Sylvius, foramen of Magendie, or foramen of Luschka, thus producing prominent internal hydrocephalus
T3b	Tumor arising from the floor of the fourth ventricle and filling the fourth ventricle
T4	Tumor further spreading through the aqueduct of Sylvius to involve the third ventricle or midbrain, or tumor extending to the upper cervical cord
M stage	
M0	No evidence of gross subarachnoid or hematogenous metastasis
M1	Microscopic tumor cells found in cerebrospinal fluid
M2	Gross nodular seedings demonstrated in the cerebellar, cerebral subarachnoid space, or in the third or lateral ventricles
M3	Gross nodular seedings in the spinal subarachnoid space
M4	Extraneuroaxonal metastasis

Table 2 MAPS classification of medulloblastoma

	Description
Metastasis	
M0	No evidence of metastasis
M1	Cells present in leptomeninges or cerebrospinal fluid
M2	Nodular seeding demonstrated in supratentorial compartment (enhanced CT scan)
M3	Nodular seeding of spinal compartment (myelogram)
M4	Combined nodular seeding of two compartments (supratentorial and/or infratentorial and/or spinal)
M5	Evidence of systemic seeding (bone marrow, various bone, and viscera organs)
Age	
A1	>3 years
A2	<3 years
Pathology	
P1	Microscopic appearance: benign, classic for tumor type, well-differentiated single cell line
P2	Microscopic appearance: anaplastic, malignant, multiple differentiated cell lines
Surgery[a]	
S0	No tumor
S1	Remaining tumor < 1.5 cm (largest diameter)
S2	Remaining tumor > 1.5 cm (largest diameter)
S3	Remaining tumor (any size) invading brain stem structures (medulla oblongata, pons, midbrain, diencephalon)
S4	Remaining tumor (any size) extending into more than one central nervous system compartment (supratentorial, infratentorial, spinal compartments)

[a]Postoperative residual tumor confirmed by enhanced CT scan of the brain.
CT, computerized tomography; MAPS, metastasis, age, pathology, and surgery.

genetic mutation which has been mapped to chromosome 16p13.3, a genomic region encoding cyclic adenosine monophosphate (cAMP) response element binding protein (CREB) binding protein (CBP).

Ptch is the human homolog of *Drosophila* patched and is a transmembrane receptor for the secreted ligand Sonic Hedgehog (Shh) signaling pathway. The Shh signaling pathway is vital during development for the formation of many vertebrate organ systems, especially the CNS. The Ptch gene has been localized to chromosome 9q, and medulloblastomas are known to have loss of heterozygosity on chromosome 9q, especially medulloblastomas of the desmoplastic subtype. Germ line mutations of Ptch located on chromosome arm 9q causes Gorlin's syndrome (nevoid basal cell carcinoma syndrome). These patients are predisposed to develop basal cell carcinomas and desmoplastic medulloblastomas.

The Shh–Ptch signaling pathway is a major mitogenic regulator of external granular layer (EGL) progenitor cells. Purkinje neurons are the main producers the glycoprotein Shh during cerebellar development. Secreted Shh binds to the receptor Ptch, which is mainly expressed on EGL precursor cells and activates the pathway by relieving the inhibition of Smoothened (Smoh). This results in the activation of target genes,

such as Ptch itself and the Gli family of transcription factors. Signaling leads to translocation of Gli proteins (intracellular effector of Shh signaling) into the nucleus, where they induce the transcription of specific target genes. Medulloblastoma might arise from a deregulation of Shh–Ptch signaling in granule-cell precursors.

The overexpression of Bmi-l and Ptch has been found in a significant fraction of human medulloblastomas, suggesting an activation of the Shh pathway in these tumors. All tumors with low Ptch expression levels did not overexpress Bmi-1, suggesting that Bmi-l is a possible downstream effector of Shh signaling and that Bmi-l overexpression is an alternative or additive mechanism to Ptch mutation in medulloblastoma pathogenesis. The function of Bmi-l downstream of Shh provides a strong link with cell-cycle control. The perturbation of cell-cycle control leads to prolonged proliferation and delayed differentiation.

The Wingless pathway is another developmental pathway, which is involved in the growth of neural progenitors. Mutations of components of this pathway, including its central components β-catenin and Axin1/Axin2, have been identified in medulloblastomas, mostly in the classic subtype. These genetic events lead to a nuclear accumulation of β-catenin and increased expression of specific target genes in medulloblastomas.

Prognostic Factors

Prognosis of patients with medulloblastoma depends on several factors, such as the presence of metastases at diagnosis and the use of radiotherapy or chemotherapy, extent of residual tumor, gender, duration of symptoms, insertion of ventricular-peritoneal shunt within 30 days of the surgical resection of tumor, and genetic profile of the medulloblastoma (high ErbB2 receptor expression and isolated 17p loss). Also, an increase in 18 fluoro-2-deoxyglucose (18 FDG) uptake in the medulloblastoma during positron emission tomography (PET) is a significant negative survival prognostic factor.

Leptomeningeal metastases are usual in children with medulloblastoma and have been reported to be as high as 46% but usually range between 10% and 35%. However, spinal intramedullary metastasis at diagnosis is very rare. Metastatic medulloblastomas have metabolite profiles indicative of increased cell growth and decreased cell death compared with localized tumors, reflecting intrinsic differences in underlying biology. Localized tumors with a magnetic resonance spectroscopy (MRS) metabolite profile similar to those with metastatic disease may be at increased risk of metastatic relapse. Approximately 5% of patients with medulloblastoma may be expected to develop systemic metastases, and the average survival of those patients is approximately 7 months after the appearance of systemic metastases.

Currently, children more than 3 years of age are divided into standard- and high-risk groups. The standard-risk group includes patients with nonmetastatic disease and less than a 1.5-cm^2 residual tumor, based on immediate postoperative gadolinium-enhanced magnetic resonance imaging (MRI) scan of the head and the spine and a negative cytology obtained from a lumbar puncture approximately 10 days after surgical resection of the tumor. The high-risk group includes patients with metastatic disease or those with more than a 1.5-cm^2 residual tumor on immediate postoperative CT or MRI scan, although recent studies (e.g., PNET IV) do not use residual disease as a high-risk feature because some studies have shown this not of prognostic value. Brain stem invasion, which was previously thought to be a high-risk feature, is no longer considered to be so. Gajjar an colleagues reported that there was no difference in progression-free survival (PFS) for patients who underwent gross total resection compared to those with any detectable residual tumor.

Patients with medulloblastoma are at a 5.4-fold excess risk for secondary cancers, including astrocytomas, oligodendrogliomas, salivary gland tumors, cervical and uterine cancer, thyroid tumors, and acute lymphoblastic leukemia, but it is difficult at times to determine the relative impact of radiation versus a possible underlying genetic proclivity to malignancy because most medulloblastoma patients receive cranio-spinal radiation.

Clinical Presentation

Usually, children with medulloblastoma present with the classic triad of symptoms of morning headache, vomiting, and lethargy, which are caused by a posterior fossa mass or associated hydrocephalus. Infants exhibit the classic triad of symptoms less often because of unfused sutures and corresponding compensation for elevated intracranial pressure. Common symptoms in infants include macrocephaly, a full anterior fontanelle, irritability, poor feeding, intermittent vomiting, and the sun-setting sign. Although papilloedema is very common in older children, it is rare in infants. Cerebellar signs include truncal, gait or limb ataxia, or/and dysmetria. Brain stem invasion by the medulloblastoma causes bulbar or facial nerve palsies. Sixth-nerve palsy is usually a result of hydrocephalus. Torticollis may be due to impacted cerebellar tonsils into the foramen magnum. Subsequently, the children tend to maintain a fixed posture, with the head tilted

and partially externally rotated to one side. Passive flexion of the patient forward is actively resisted and results in severe pain. A short duration of symptoms correlates with a higher stage of disease. Symptoms usually last from a few weeks to up to 3 months, although patients presenting with symptoms for 1 year have been reported. Other symptoms include sleepiness, reduced appetite, weight loss, blurred vision, and change in personality.

The most common neurological signs at the time of presentation include ataxia/ataxic gait, nystagmus, cranial nerve palsies, papilloedema, increased head circumference, torticollis, limitation of upward gaze, and reduced visual acquity.

Radiology

Medulloblastomas are usually seen in the midline arising from the anterior portion of the vermis and then growing into the inferior or superior velum of the fourth ventricle. Obstructive hydrocephalus is common and is caused either by filling of the fourth ventricle by the tumor or by compression or sharp angulation of the cerebral aqueduct (see Figure 1). Medulloblastomas have a 10–21% incidence of calcification. Cystic change occurs in 10–20% of children and 59–82% of adults. Hemorrhage within pediatric medulloblastomas is relatively uncommon.

On nonenhanced CT, medulloblastomas appear as homogeneous hyperdense well-circumscribed masses in the midline, associated with the vermis. Increased attenuation on unenhanced scans is felt to correlate with the high degree of cellularity, scanty cytoplasm, and desmoplasia present in these tumors. The homogeneity of density and hyperdensity on a CT scan are the best findings that suggest medulloblastoma over ependymoma.

MRI is the investigation of choice. Sagittal MR images are optimal for visualizing the origin of the medulloblastoma from the vermis. Medulloblastoma is usually hypointense on T-1-weighted and isointense on T-2-weighted images. It is typically very well defined, distorts the fourth ventricle, and does not demonstrate a large amount of associated edema. A spinal MR scan is required preoperatively to assess the presence of spinal metastases (see Figure 1).

The spectroscopic signature of a PNET can distinguish it from other posterior fossa masses in children. PNET tumors show low *N*-acetylaspartate : choline (NAA:Cho) ratio and creatine/ phosphocreatine : Cho-containing compounds (Cr:Cho) ratios compared with low-grade astrocytomas and ependymomas, which have higher NAA:Cho ratios. The Cr:Cho ratio is highest for ependymomas compared to astrocytomas and PNETs. Lactate is usually not present in PNETs, but it is often seen in pilocytic astocytomas and ependymomas (see Figure 1).

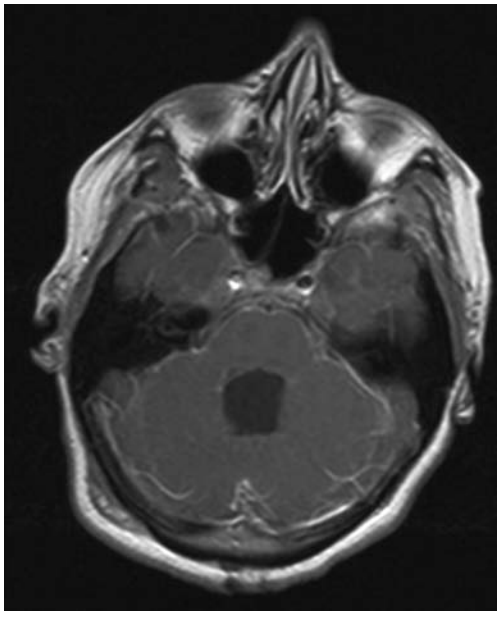

Figure 3 Postoperative brain T-1-weighted magnetic resonance (MR) scan with contrast of the patient in Figure 1, obtained 24 h postsurgery. Total surgical excision of the tumor has been achieved.

A routine postoperative cranial MRI is obtained during the immediate postoperative period (see Figure 3). Most chemotherapy protocols require imaging within the first 72 h. Early enhancement has been shown to be associated with residual tumor, which lacks a blood–brain barrier. Imaging of the patients after the third postoperative day may be misleading because dural and operative bed enhancement may simulate a residual tumor or subarachnoid and leptomingeal metastatic disease. Also, the use of hemostatic agents at the tumor bed can be misinterpreted as a residual tumor on early postoperative CT scans.

Metaiodobenzylguanidine (MIBG) is a physiological analog of norepinephrine used in scintigraphy. The 1^{23} I-MIBG single photon emission computerized tomography (SPECT) technique has been reported in the literature as a innovative approach in the diagnosis of medulloblastoma, providing valuable information about the differential diagnosis of medulloblastoma from other posterior fossa tumors and an assessment of the effects of treatment. Another radiological method for the evaluation of patients with medulloblastoma is the 18 FDG PET. High-grade brain tumors accumulate 18 FDG more avidly than normal neurons, enabling their identification amid the background FDG uptake in the brain.

Currently, surveillance MRI of children with medulloblastoma is used routinely in daily clinical

practice to assess the results of treatment and to diagnose an asymptomatic relapse of the disease. The highest risk of relapse is in the first 2 years post-surgery. For this reason, in most protocols MR scans are obtained every 3 months the first year, every 6 months after the second year, and yearly thereafter.

Management

The treatment of patients with medulloblastoma includes maximum surgical resection, followed by adjuvant radiotherapy of whole neuroaxis in patients over the age of 3 years and chemotherapy.

Surgery

Surgical treatment plays an important role in the treatment of medulloblastoma because it establishes the diagnosis, restores the patency of CSF pathways, and reduces the volume of local disease totally or nearly totally in patients with brain stem invasion. It is preferred that a thin film of tumor be left on the brain stem, rather than attempting to excise it, to avoid iatrogenic devastating brain stem injury with bilateral abducent, bilateral facial, or lower cranial nerves palsies. In addition, brain stem invasion (Chang stage T3b), previously considered an indicator of high risk, is now thought not to affect prognosis.

Surgical treatment is usually required on a semi-urgent basis because most patients present with symptoms and signs of increased intracranial pressure. Preoperative administration of corticosteroids decreases the intracranial pressure and brain stem compression, allowing the surgical resection of the tumor to be performed within the next couple of days after admission. In cases of significantly high intracranial pressure in which children present with lethargy, bradycardia, or respiratory collapse, urgent diversion of the CSF is required by the insertion of an external ventricular drain or by endoscopic third ventriculostomy. Permanent treatment of hydrocephalus by the insertion of a ventriculo-peritoneal shunt or by an endoscopic third ventriculostomy is required in 25% of patients under 3 years of age and in 10% of those over 3 years.

Surgical excision of the tumor is performed through a straight midline sagittal suboccipital incision with the patient in prone position and the head flexed. After dissection of the nuchal ligament, the excision of a portion of occipital bone and posterior arch of atlas is performed, followed by the opening of the dura in a Y fashion. The tumor is dissected from the normal cerebellum and removed piecemeal by using the cavitron ultrasonic aspirator (CUSA). It is fundamental for the surgeon to visualize the floor of the fourth ventricle and isolate it from the tumor with a patty, aiming to obtain a dissection plane between the tumor and the brain stem and avoid its injury. It is well documented that patients with total or near-total excision ($<1.5\,cm^2$ of residual tumor in its largest part) on the immediate postoperative MR scan have a better prognosis. In the case of a larger residuum, reoperation is commonly performed, aiming for total or near-total excision. Nowadays, the mortality from surgical resection of medulloblastoma is in the range 1–5%, depending on the preoperative clinical state of the patient.

Most common postoperative complications include hematoma, infection/meningitis, aseptic meningitis, hydrocephalus, pseudomeningocele, CSF leak, cerebellar mutism, cranial nerve palsy, and hemiparesis. Mutism is generally defined as complete absence of speech in a conscious patient, with intact comprehension and no evidence of oral apraxia. The essential anatomical substrate of mutism is probably extensive lesion of the vermian and/or paravermian cerebellar region, as well as the hemispheric cortex, cerebellar peduncles, fibers from the dentate-thalamo-cortical pathway, and part of deep nuclei due to postoperative arterial spasm or retraction injury and edema.

Radiotherapy

Medulloblastoma was a lethal disease for all patients in the pre-radiotherapy era. After the introduction of radiotherapy of cranio-spinal axis, the survival improved to up to 54%. For standard-risk patients, the conventional doses of radiotherapy are around 36 Gy to the cranio-spinal axis, together with a boost of 18–20 Gy to the posterior fossa. Usually, the posterior fossa irradiation of 54–56 Gy is delivered in 30 daily fractions, and the rest of cranio-spinal axis irradiation of 20 Gy is delivered in 20 daily fractions. The 5-year survival of patients receiving such doses has been reported to be between 55 and 70%.

Although radiation is a highly effective therapy for medulloblastoma, it causes severe damage to the developing brain, especially in children younger than 3 years of age. Radionecrosis, mineralizing microangiopathy and calcification of brain, hypopituitarism (growth hormone, follicle-stimulating hormone (FSH), and luteinizing hormone (LH) deficiency), gonadal dysfunction in girls, hypothyroidism, cataract, ototoxicity, reduced perceptual motor-task performance and motor dexterity, and secondary malignancies are well documented long-term complications of radiotherapy for pediatric brain tumors. To reduce the irradiation toxicity to the CNS, hyperfractionated (two daily fractions of) irradiation of cranio-spinal axis has been used in clinical trials (HIT-PNET 4) for the same dose of

radiation as used in conventional radiotherapy. Nowadays, most centers do not administer radiotherapy in patients younger than 3 years of age.

Chemotherapy

In recent years, the combination of surgery with radiotherapy and chemotherapy has improved the prognosis of children with medulloblastoma up to 83%. In the 1980s, the debilitating cognitive and endocrinological sequelae of cranio-spinal radiotherapy coupled with the poor survival for children with residual tumor following surgery led to successive randomized clinical trials by the International Society of Paediatric Oncology (SIOP), Children's Cancer Group (CCG 942), and Pediatric Oncology Group (POG 7909), which introduced chemotherapy to improve survival and reduce radiation dosage. The extent of tumor spread at diagnosis has been always the most robust predictor of outcome. Subsequently, for the patients whose preoperative MR scans do not reveal any metastatic disease, diagnostic lumbar puncture and cytological examination of the CSF are performed to establish the presence of microscopic dissemination of medulloblastoma cells in the CSF. However, the impact of only microscopic tumor cells in lumbar CSF on survival remains debatable. In both CCG 921 and the German trial HIT'91, overall survival was not significantly different in children who were categorized as having no metastasis or microscopic metastasis. Children younger than 3 years are always considered to be at high risk because of their well-established higher rates of recurrence. Chemotherapy in those children is a temporizing agent, allowing their brain time to develop and receive radiotherapy after the age of 3 years. The most commonly used drugs in various chemotherapy schemes include ifosamide, etoposide, methotraxate, cisplatin, vincristine, carboplatine, lomustine, and cyclophosphamide.

Conclusion

With modern multimodality treatment, the outcome of patients with medulloblastoma has improved dramatically. For standard-risk patients (older than 3 years of age, no metastasis, less than 1.5 cm^2 of residual disease), the 5-year progression-free survival varies between 60 and 94% in various clinical series. For high-risk patients (younger than 3 years of age, metastatic disease, more than 1.5 cm^2 of residual disease), the 5-year progression-free survival varies between 43 and 84%. Despite the improvement in survival rate, medulloblastoma remains a challenging disease because those who survive have significant cognitive and physical impairments secondary to both the disease and its treatment. Further understanding of medulloblastoma biology will ultimately lead to the improvement of current therapies and the development of new ones.

See also: Brain Injury: Magnetic Resonance Studies of Metabolic Aspects; Brainstem and Cranial Nerves; Cerebellum: Clinical Pathology; Magnetic Resonance Spectroscopy; Opsoclonus/Myoclonus and Neuroblastoma; Perfusion MRI; Positron Emission Tomography (PET); Tumors of the Brain and the Spinal Cord.

Further Reading

Albright AL, Wisoff JH, Zeltzer PM, Boyett JM, Rorke LB, and Stanley P (1996) Effects of medulloblastoma resections on outcome in children: A report from the Children's Cancer Group. *Neurosurgery* 38: 265–271.

Chang CH, Housepian EM, and Herbert C Jr. (1969) An operative staging system and a megavoltage radiotherapeutic technique for cerebellar medulloblastomas. *Radiology* 93: 1351–1359.

David KM, Casey AT, Hayward RD, Harkness WF, Phipps K, and Wade AM (1997) Medulloblastoma: Is the 5-year survival rate improving? A review of 80 cases from a single institution. *Journal of Neurosurgery* 86: 13–21.

Evans AE, Jenkin RD, Sposto R, et al. (1990) The treatment of medulloblastoma: Results of a prospective randomized trial of radiation therapy with and without CCNU, vincristine, and prednisone. *Journal of Neurosurgery* 72: 572–582.

Gilbertson R, Wickramasinghe C, Hernan R, et al. (2001) Clinical and molecular stratification of disease risk in medulloblastoma. *British Journal of Cancer* 85: 705–712.

Grill J, Sainte-Rose C, Jouvet A, et al. (2005) French Society of Paediatric Oncology: Treatment of medulloblastoma with postoperative chemotherapy alone – an SFOP prospective trial in young children. *Lancet Oncology* 6: 573–580.

Kombogiorgas D, Sgouros S, Walsh AR, et al. (2007) Outcome of children with posterior fossa medulloblastoma: A single institution experience over the decade 1994–2003. *Child's Nervous System* 23: 399–405.

Kortmann RD, Kuhl J, Timmermann B, et al. (2000) Postoperative neoadjuvant chemotherapy before radiotherapy as compared to immediate radiotherapy followed by maintenance chemotherapy in the treatment of medulloblastoma in childhood: Results of the German prospective randomized trial HIT '91. *International Journal of Radiation Oncology Biology and Physics* 46: 269–279.

Krischer JP, Ragab AH, Kun L, et al. (1991) Nitrogen mustard, vincristine, procarbazine, and prednisone as adjuvant chemotherapy in the treatment of medulloblastoma: A Pediatric Oncology Group study. *Journal of Neurosurgery* 74: 905–909.

Packer RJ, Goldwein J, Nicholson HS, et al. (1999) Treatment of children with medulloblastomas with reduced-dose craniospinal radiation therapy and adjuvant chemotherapy: A Children's Cancer Group study. *Journal of Clinical Oncology* 17: 2127–2136.

Packer RJ, Sutton LN, Elterman R, et al. (1994) Outcome for children with medulloblastoma treated with radiation and cisplatin, CCNU, and vincristine chemotherapy. *Journal of Neurosurgery* 81: 690–698.

Peet AC, Davies NP, Ridley L, et al. (2007) Magnetic resonance spectroscopy suggests key differences in the metastatic behaviour of medulloblastoma. *European Journal of Cancer* 43: 1037–1044.

Pietsch T, Taylor MD, and Rutka JT (2004) Molecular pathogenesis of childhood brain tumors. *Journal of Neurooncology* 70: 203–215.

Tait DM, Thornton-Jones H, Bloom HJ, Lemerle J, and Morris-Jones P (1990) Adjuvant chemotherapy for medulloblastoma: The first multi-centre control trial of the International Society of Paediatric Oncology (SIOP I). *European Journal of Cancer* 26: 464–469.

Zeltzer PM, Boyett JM, Finlay JL, et al. (1999) Metastasis stage, adjuvant treatment, and residual tumor are prognostic factors for medulloblastoma in children: Conclusions from the Children's Cancer Group 921 randomized phase III study. *Journal of Clinical Oncology* 17: 832–845.

Relevant Websites

http://www.cclg.org.uk – Children's Cancer and Leukaemia Group (UK).

http://www.childrensoncologygroup.org – CureSearch, Children's Oncology Group (USA).

http://www.kinderkrebsinfo.de – Gesellschaft für Pädiatrische Onkologie und Hämatologie (Germany).

http://www.la-sfop.com – Societé Française des Cancers de l' Enfant (France).

Melanocortins: Brain Effects

M Merkestein, W H Gispen, and R A H Adan, University Medical Center Utrecht, Utrecht, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

The melanocortins (MCs) are a group of proopiomelanocortin (POMC)-derived peptides, sharing a core amino acid motif. The main sites of POMC expression are the arcuate nucleus and the nucleus of the solitary tract (NTS) in the brain, and the anterior lobe of the pituitary gland. Lower amounts of POMC are expressed in peripheral tissues such as skin and certain immune cells.

POMC is cleaved by prohormone convertase 1 (PC1) and 2 (PC2) into the melanocortins, which include α-, β-, and γ-melanocyte-stimulating hormone (MSH) and adrenocorticotropin (ACTH). POMC contains the sequences of the nonMC peptides β-endorphin and lipotropic hormone as well. The MCs share the same signature core amino acid sequence His-Phe-Arg-Trp. This sequence is required for binding to one of the five MC receptor subtypes (MC1–MC5).

MC receptors are seven-transmembrane G-protein-coupled receptors which are positively coupled to adenylate cyclase and, hence, enhance cyclic adenosine monophosphate (cAMP) and protein kinase A signaling. The MC1 receptor MSH receptor, which is mainly expressed on skin melanocytes, plays an important role in the regulation of pigmentation and binds to α- and β-MSH and ACTH. The MC2 receptor (ACTH) receptor, which is expressed in the adrenal gland, binds exclusively to ACTH and controls corticosteroidgenesis in the adrenal cortex. The MC3 and MC4 receptors are the main brain MC receptors, although the MC1 and MC5 receptors have also been found in the brain but at lower expression levels. The MC3 receptor binds to all MCs, including γ-MSH, whereas the MC4 and MC5 receptors bind to all MCs except γ-MSH. MC5 receptor is specifically expressed in the exocrine glands and regulates the synthesis and secretion of exocrine gland products.

This article focuses on the effects of MCs and on the physiological roles of the brain MC system. In addition to the well-known roles of the MC system in the regulation of energy balance, we review its role in fever, the immune system, nerve regeneration, nociception, reward, grooming behavior, stress and anxiety, sexual behavior, and blood pressure and heart rate (Figure 1).

The Brain Melanocortin System

Fifty years ago, the first direct effects of MCs on the brain were described. Central injections of ACTH were shown to affect avoidance and grooming behavior in rats. In the late 1970s, the presence of immunoreactive neurons to MC antibodies was demonstrated, which was a strong indication that the MC system plays a role in the brain. Another indication of the brain MC system was the cloning of the POMC gene and its expression in the arcuate nucleus. Because of difficulties in labeling MCs, it took until the late 1980s before binding sites of MCs could be detected in the brain. The binding of MCs was shown to be restricted to specific areas of the brain. Only slightly later, MC receptors were cloned, and the MC3 and MC4 receptors were shown to be expressed in the brain. The availability of clones of MC receptors accelerated the identification of MC receptor antagonists such as SHU9119, which antagonizes the MC3 and MC4 receptors. This pharmacological tool was very helpful in establishing which effects of MCs were mediated via brain MC receptors.

A unique feature of the MC system is that there are, in addition to endogenous agonists (the MCs), endogenous inverse agonists. An inverse agonist is a ligand that binds to the same receptor-binding site as an agonist and not only antagonizes the effects of an agonist but, moreover, exerts the opposite effect by suppressing spontaneous receptor signaling (when present). A spontaneous gain of function mutation in mice in the agouti locus led to the identification of endogenous inverse agonists for the MC system. This mutation resulted in a yellow fur due to overexpression of agouti protein, which acts as an inverse agonist at the MC1 receptor. Based on its homology to agouti, agouti-related protein (AgRP) was cloned and demonstrated to be expressed in the arcuate nucleus in neurons which are situated close to POMC neurons and project to similar brain areas. AgRP was shown to act as an inverse agonist at the MC3 and MC4 receptors.

Regulation of Energy Balance

The effect of MCs on energy balance was discovered in the mid-1980s, when Bertolini and coworkers and, later, Tsuii and Bray reported that a central injection of MCs reduces ingestive behavior. This was just one of the many effects of MCs reported in that period. In the mid-1990s, the importance of its role in energy balance was underlined by the discovery that the

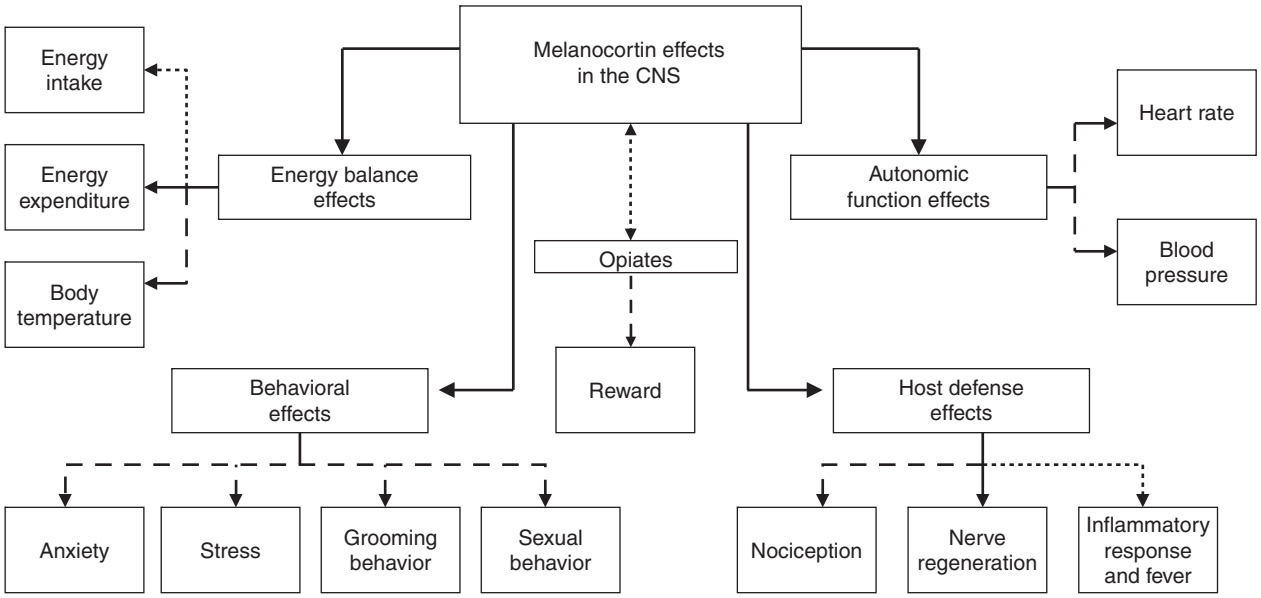

Figure 1 Overview of the effects of melanocortins in the central nervous system (CNS). The brain melanocortin system affects multiple systems: energy balance, behavior, autonomic system function, and host defense. Dotted arrows indicate a stimulatory effect of the melanocortin system; dashed arrows indicate an inhibitory effect of the melanocortin system.

central MC system acts downstream of leptin, an adipose tissue-derived hormone, released in proportion to body fat content. Physiological levels of leptin result in an increase in energy expenditure and a decrease in energy intake which involves POMC neurons, whereas orexigenic AgRP/neuropeptide Y (NPY) neurons are inhibited. This effect of leptin can be blocked by the administration of the MC receptor antagonist SHU9119. In addition to increasing the firing rate of POMC neurons, leptin also rapidly increases N-acetylation of MSH, which makes MSH more powerful in inhibiting feeding behavior.

Another indication for the role of the MC system in the brain was the observation that mice without the MC4 receptor or the POMC gene have an obese phenotype. Interestingly, humans with mutations in either the POMC or MC4 receptor gene are also susceptible to develop obesity, and nowadays MC4 receptor mutations are the most frequent cause of monogenic obesity in populations of morbidly obese children. Hence, the MC system is also essential for the regulation of energy balance in humans.

The MC system plays a physiological role in the regulation of energy balance. In a state of negative energy balance (e.g., in starvation), leptin levels drop, which results in increased mRNA levels of AgRP in the arcuate nucleus. In a state of positive energy balance (e.g., in overfed rats), POMC mRNA levels are increased, which indicates that the MC system is physiologically involved in counteracting a positive energy balance. The observations that animals with deficits in MC signaling, such as POMC or MC4 receptor knockouts, and rodents chronically infused with MC receptor antagonists develop obesity emphasize this role of the MC system. On the other hand, the administration of a MC agonist, hence activating the MC system, leads to a temporary decrease in food intake and a sustained decrease in body weight.

The MC system affects energy balance not only via the regulation of food intake but also via the regulation of energy expenditure. Rats treated with SHU9119, but pair-fed to controls, show a small increase in body weight which cannot be due to an increase in energy intake. The increase in body weight might be due to the observed decrease in locomotor activity and body temperature in these rats. Furthermore, the inhibition of the MC system suppresses uncoupling protein (UCP) 1 expression in brown adipose tissue, thus inhibiting the conversion of energy into heat production. In addition, analysis of MC4 receptor knockout mice revealed a defect in diet-induced thermogenesis. In contrast to MC4 receptor knockouts, MC3 receptor knockout mice have a normal body weight but an increased body fat percentage. There is an additive effect on body weight in MC3–MC4 receptor double knockouts, indicating that MC3 and MC4 receptors, which are expressed in partially overlapping brain circuits, have different roles in the regulation of energy balance.

Analysis of meal patterns revealed that MCs affect meal termination rather than meal initiation. MC receptor agonists reduce meal size but have no effect

on meal frequency, whereas SHU9119 increases meal size without influencing meal frequency. Furthermore, Benoit and colleagues demonstrated that MC receptor agonists are less potent in suppressing food intake when rats are trained to eat, that is, when they expect a meal or when they have learned to eat large amounts of food in a short time period. MC 4 receptors in the dorsal vagal complex in the brain stem, the brain area which is implicated in the termination of meals (satiety), could well mediate these effects of MCs. Indeed, the application of MCs in the fourth ventricle, near the brain stem, is as potent as its administration into the lateral ventricle, near the hypothalamus. This emphasizes the role of the brain stem in mediating the effect of MCs on food intake. Recently it was demonstrated that MCs increase the sensitivity for the satiety hormone cholecystokinin (CCK) at the level of the dorsal vagal complex. However, the brain stem is certainly not the only brain area where activation of MC receptors inhibits food intake. Local injection of MCs in several brain areas showed that low doses of MCs in the paraventricular (PVN) and dorsomedial (DMH) hypothalamus and the medial preoptic area inhibit food intake as well.

Brain areas which contain and integrate information on taste and palatability, such as the amygdala, NTS, and parabrachial nucleus, also express MC receptors. This provides an anatomical basis for an effect of MCs on food choice. Indeed, MC receptor agonists specifically reduce the intake of fat, but not of protein or carbohydrates, on a three-choice macronutrient diet, whereas the inverse agonist AgRP selectively increases the intake of high-fat diets in rats. Furthermore, mice with mutations in the MC system display an increase in energy intake when fed high-fat foods.

By using a gene therapy approach, it was demonstrated that the overexpression of the MC receptor inverse agonist agouti in different hypothalamic regions results in the development of different obese phenotypes. The overexpression of agouti in the PVN results in hyperphagia and obesity, in the DMH in a late-onset obesity syndrome, and in the lateral hypothalamus in obesity and exaggerated food efficiency but only when the rats are fed a high-fat diet. The exact mechanisms underlying these effects are largely unknown, but may find their origin in the interactions of MC receptors in the different brain regions with downstream effector pathways. In the PVN, the downstream effector pathway probably includes neurons that express corticotropin-releasing hormone (CRH) and thyrotropin-releasing hormone (TRH), which regulate neuroendocrine and autonomic pathways. The effects on energy balance could be similar to adaptations in these systems in response to stress, fasting, physical activity, and cold exposure. Within the DMH, signaling by MC receptors probably involves more chronic changes in energy balance, such as those required for lactation or perhaps chronic food restriction, and probably includes the activation of the orexigenic NPY-containing neuronal cell bodies. Within the lateral hypothalamus, downstream effector pathways activated by MC receptor signaling probably involve brain areas which play a role in motivation. Infusion of the MC receptor antagonist SHU9119 enhances the expression of melanin-concentrating hormone (MCH). Lateral hypothalamic neuronal cell bodies which express MCH project to a wide variety of areas, including the nucleus accumbens (Nac), a forebrain region involved in reward. MCH induces feeding behavior when injected into the Nac. As mentioned earlier, brain areas involved in taste and palatibility specifically increase the intake of fat.

Finally, it has been suggested that the effects of AgRP on food intake, particularly the long-term effects, are dependent on opioid receptor functioning and may involve forebrain regions such as the Nac as well. Whether these effects are specific for AgRP or the effects are just due to changes in MC receptor activity in general is not evident yet.

Eating Disorders

MCs are proposed to play a role in disorders of negative energy balance, such as anorexia nervosa. Indeed, the Ala67Thre AgRP single-nucleotide polymorphism, which reduces the inverse agonism activity of AgRP, occurs more frequently in patients with anorexia nervosa than in controls. Using an animal model for anorexia, the activity-based anorexia model (in which rats have a running wheel in their cage and restricted access to food), Hillebrand and colleagues demonstrated that, despite the reduction in body weight due to food restriction, POMC expression increased in the first days, suggesting increased activity of melanocortinergic neurons. Furthermore, after 1-week exposure to this model, an increased density of MC receptors in the ventromedial hypothalamus was observed. Taken together, these rats are more sensitive to weight loss because of an increased level of POMC, the MC precursor, and more MC receptors; hence, they are more sensitive for the anorectic effects of MCs. However, an enhanced endogenous AgRP expression in the arcuate nucleus was observed as well, which makes it difficult to assess what the net total result on the MC receptors is. In the activity-based anorexia model, additional central administration of AgRP decreased energy expenditure and increased food intake in these rats, which improved their condition, suggesting that increased melanocortinergic tone contributes to anorectic behavior in the

anorexia model. On the other hand, central administration of α-MSH on exposure to the model worsened their condition and accelerated the development of anorexia, in addition to increasing the activity of the hypothalamic–pituitary–adrenal (HPA) axis. Interestingly, in anorexia nervosa patients, the activity of this axis is increased as well.

Regulation of Body Temperature and Fever

The thermoregulatory actions of MCs are dependent on the physiological state. Under normal conditions, MCs modestly increase body temperature, which is probably part of the repertoire of effects by which the MC system controls energy balance. Because MCs, in particular, increase body temperature in the dark phase, when also locomotor activity is increased by MCs, at least part of this thermogenic effect is caused by increased locomotion. However, because MCs also stimulate thermogenesis by stimulating heat production in brown adipose tissue (via the stimulation of UCP-1 expression), other mechanisms contribute to the thermogenic effect of MCs.

In febrile rats, MCs decrease body temperature. Central administration of MCs inhibits a lipopolysaccharide (bacterial endotoxin)-induced febrile response, whereas central administration of SHU9119 increases this febrile response. This effect on the febrile response is mediated via the MC4 receptors in the brain because it can blocked with an MC4 receptor-selective antagonist. Increased levels of endogenous MCs in the septal region are reported during fever in rabbits, suggesting that the endogenous MC system is implicated in the febrile response. In addition, injection of MCs into the septal and preoptic regions in rabbits prevents fever. Moreover, these regions express MC4 receptors and contain thermosensitive neurons, which all suggests that these regions play a role in the effect of MCs on fever.

Regulation of the Inflammatory Response

MCs are potent modulators of inflammation. Central application of MCs inhibits inflammatory responses induced by the injection of interleukin β into the ear or hind paw of rodents, which seem to require an intact sympathetic nervous system. However, peripheral administration of MCs also inhibits inflammatory responses.

MCs have been shown to exert anti-inflammatory responses via MC receptors expressed on blood cells, cells of the immune system, and a variety of other cells. The MC1 receptor has been demonstrated on monocytes, macrophages, neutrophils, mast cells, fibroblasts, and dendritic cells. In addition to the MC1 receptor, MC3 receptors expressed in macrophages and MC5 receptors expressed in lymphocytes have been implicated in the anti-inflammatory effects of MCs. MCs inhibit the release of inflammatory mediators, such as cytokines, even in cells *in vitro*. The cytokine release is affected by MCs via inhibition of activation of the nuclear factor (NF)-κB. NF-κB is a transcription factor that regulates many genes including, cytokines and chemokines, which might explain the broad effects of MCs on mediator production.

α-MSH(11–13), which does not bind to MC receptors, has also anti-inflammatory properties. This indicates that some of the effects of MCs are mediated via a mechanism other than the MC receptors.

Nerve Regeneration

Both systemic and local application of MCs enhances recovery after peripheral nerve injury, provided that the application occurs within days after the injury. The observed neurotrophic effects of MCs are improved nerve regeneration and muscle innervation. *In vitro* studies demonstrated that MCs promote the outgrowth of neurites in cell lines. These effects are also demonstrated *in vivo;* α-MSH stimulates spinal neurite outgrowth after a spinal cord injury. The exact mechanism via which MCs improve nerve regeneration is not known yet. α-MSH affects neurons, but glial cells can also be a target for the MCs.

Because γ-MSH, which activates only the MC3 receptor, does not stimulate nerve regeneration following a sciatic nerve crush lesion, the involvement of the MC3 receptor is ruled out. MT-II, a potent MC receptor agonist that activates MC1, -3, -4, and -5 stimulated nerve regeneration in this model. The MC3/4 receptor antagonist SHU9119 was also shown to stimulate nerve regeneration. SHU9119 is an agonist for the MC1 and MC5 receptors. The acceleration of functional recovery by MCs after peripheral nerve injury is most likely mediated via the MC5 receptor because ligands that were able to activate MC1, but not MC5, receptors were inactive in stimulating nerve regeneration.

Nociception

Both MCs and the opiate β-endorphin are derived form POMC. Opiates are well known for their analgesic effects. MCs also have an effect on the perception of pain, but, in contrast to opiates, MCs are believed to increase the sensitivity to painful stimuli, as demonstrated by the induction of hyperalgesia by the central administration of α-MSH in the tail flick test and the hot plate test in rats. Treatment with

SHU9119, as well as with selective MC4 receptor antagonists, decreases allodynia and nociception in a rat model for neuropathic pain. Also, in the mouse formalin test MC agonists increased nociception, whereas MC antagonists reduced nociception. Upon neuropathic pain, induced by chronic restriction of the sciatic nerve, MC receptor binding increases in areas in the spinal cord which are involved in the transmission of pain. The increased pain perception in this model is mediated by the MC4 receptor because a MC-3-specific agonist has no effect. A MC4-selective antagonist also decreases nociception in this model, further emphasizing the involvement of the MC4 receptor in pain processing. Furthermore, the MC4 is the only MC receptor present in the spinal cord, where the processing of pain takes place. The effects of MCs on nociception can be modulated by opiates. Low doses of the opiate antagonist naloxone inhibit the antinociceptive effect of SHU9119, and when administered together, the effects of the opiate agonist morphine and SHU9119 are additive in rats with neuropathic pain. This indicates that MC receptor antagonists are potential analgesic drugs, which reduce pain perception and increase the effect of morphine. However, they might not function in the long term because the effects on nociception are only transient in the neuropathic pain model. Therefore, it remains to be clarified whether MC receptor ligands are of any use in the treatment of chronic pain.

Interaction with Opiates and the Reward System

In addition to in the modulation of pain as already described, MCs and opiates antagonize one another in several other functions, including feeding behavior. Upon stimulation of POMC neurons and processing of POMC, α-MSH and β-endorphin are stored in the same vesicles and are consequently released at the same time. Because MC receptors couple positively to adenylate cyclase, whereas opiate receptors couple negatively to this pathway, the functional antagonism between these systems is likely to be regulated at this level.

Interestingly, *N*-acetyltransferase acetylates both α-MSH and β-endorphin, making the MC more stable and potent but decreasing the affinity of β-endorphin for opioid receptors. Hence, the regulation of *N*-acetyltransferase activity (e.g., by leptin) has an effect on the balance in efficacy of MCs and β-endorphin after secretion on the activation of POMC neurons. This regulation of efficacy takes also place at the receptor level. Chronic administration of morphine, an opioid, leads to the reduced expression of MC4 receptor mRNA in the nucleus accumbens and the periaquaductal gray, brain regions involved in reward and the reinforcing properties of drugs of abuse. Interestingly, both self-administration of opiates and analgesia caused by opiates can be antagonized by MCs.

Furthermore, MCs modulate the response to addictive drugs because the administration of MC receptor antagonists in the nucleus accumbens inhibits the primary and conditioned reinforcing effects of cocaine. In addition, the MC receptor agonist MTII increases the rewarding effect of amphetamine. These addictive drugs might also affect MC signaling. Chronic cocaine administration, for example, increases the expression of MC4 receptors in the nucleus accumbens, hippocampus, and striatum.

Grooming Behavior

Grooming behavior is displayed by rats as a novelty-induced stress response. It involves activities such as washing, scratching, and licking of the body surface. It is considered that this behavior serves to reduce arousal after stress.

This novelty-induced grooming can be inhibited with MC antagonists, MC antibodies, and antisense POMC nucleotides. MC antagonists are able to inhibit the physiological relevant grooming behavior induced by a novel environment. Intracerebroventricular administration of ACTH and α-MSH induces excessive grooming behavior and stimulates the HPA axis. This excessive grooming behavior seems to be mediated via the MC4 receptor because MC4 receptor-specific antagonists inhibited α-MSH-induced grooming. Furthermore, selective MC4 receptor agonists induce grooming behavior in proportion to their ability to activate the MC4 receptor.

Brain regions which are likely to be involved in the grooming response are the PVN, substantia nigra, and periaqueductal grey because they all express the MC4 receptor and local injections of ACTH in these areas induce excessive grooming behavior.

Anxiety and Stress

MCs have an influence on the stress response because the central administration of MCs stimulates the HPA axis. This could possibly be mediated via stimulation of MC4 receptors present on vasopressinergic/corticotropic neurons in the PVN because plasma levels of stress hormones such as ACTH and corticosterone increase after an injection of α-MSH into the PVN. The increase in plasma corticosterone levels is reduced by the selective MC4 receptor antagonist HS014, supporting the role of the MC4 receptor in the activation of the stress response by MCs. The involvement of the

MC system in the physiological response to stress is demonstrated by the downregulation of AgRP expression and upregulation of POMC and MC4 receptor expression in the arcuate nucleus after footshock stress.

MCs are also thought to have anxiogenic effects, mediated by the MC4 receptor. These anxiogenic effects are demonstrated in the elevated plus maze after an injection of α-MSH into the medial preoptic area and ventromedial nucleus, and in the Vogel test after the administration of α-MSH and MTII. Other supporting evidence for the anxiogenic action of MCs is the decrease in social contact and increased aggressive behavior in the social interaction test after administration of MCs. On the other hand, MC4 receptor antagonists have anxiolytic-like activity in the mouse light–dark exploration test. These anxiolytic-like effects of the MC4 receptor antagonists are clearer in stressful than in nonstressful conditions, which supports the role of the endogenous MC system in a normal stress response. Future research will determine whether MC4 receptor antagonists are potent anxiolytic drugs.

Sexual Behavior

Early studies described the effects of intracerebroventricularly administered ACTH and α-MSH in stimulating penile erections, ejaculation, sexual posturing, and genital licking in male rats but only in the presence of testosterone. In female rats, MCs stimulate lordosis behavior, and the medial preoptic area has been proposed as site of action. In addition, in humans, the superpotent MC agonist MTII induces penile erections and increases sexual desire. Peripheral as well as central MC receptors are considered to be involved in the effects of MCs on sexual behavior. MC5 receptors expressed on preputial glands may play a role in the stimulation of sexual behavior induced by peripheral administration of MCs because MCs stimulate the release of pheromones. However, central administration of MCs affects sexual behavior, in particular penile erection. Injection of MCs into the periventricular hypothalamus, which includes the PVN, stimulates penile erection in rats. Central administration of a MC4 receptor agonist improves sexual function, but has no effect in mice which lack the MC4 receptor. These MC4 receptor-null mice demonstrate a diminished copulatory behavior. Additional evidence for the role of the MC4 receptor in sexual behavior is that, in addition to the spinal cord and brain sites putatively involved in this MC effect, the human and rat penis also express MC4 receptors. However, the effect of MCs on lordosis in female rats could not be blocked by MC4 receptor antagonists. Furthermore, other studies found no effect of MC4 receptor blockage on copulatory behavior in male rats. The MC4 receptor is assumed to play the most important role in the central regulation of sexual behavior by MCs.

Regulation of Blood Pressure and Heart Rate

The effect of MCs on blood pressure and heart rate is rather complicated. Peripheral and central administration of γ-MSH affect heart rate and blood pressure depending on the physiological state. In conscious rats γ-MSH increases blood pressure and heart rate, but in anesthetized rats the opposite effects were observed. The peripheral administration of γ-MSH has a greater effect than its central administration, whereas α-MSH decreases heart rate and blood pressure only when administered centrally. The normal-pressure effect of γ-MSH is not mediated by MC receptors because the effect cannot be blocked with MC receptor antagonists and because even smaller fragments of γ-MSH, such as γ-MSH (6–12), which do not have any affinity for MC receptors, increase blood pressure and heart rate. γ-MSH shares some sequence homology with FMRF-amides, which show a similar effect on blood pressure and heart rate. Hence, the cardiovascular effects of γ-MSH could be mediated via an interaction of γ-MSH and FMRF-amide receptors. However, there is some evidence that a small part of the pressure effect of γ-MSH is mediated via MC3 receptors in renal nerves that are involved in the natriuretic effect of γ-MSH, counteracting salt-sensitive hypertension. Chronic treatment with intracerebroventricular MTII leads to a modestly increased blood pressure and heart rate, which is mediated via the adrenergic system, possibly reflecting effects of central MC receptor activation on sympathetic outflow.

Other MCs are able to affect blood pressure and heart rate via their central receptors. Injection of MCs into the NTS or dorsal motor nucleus of the vagus nerve results in bradycardia and a lower blood pressure, which is inhibited by the co-administration of SHU9119. This depressor effect of MCs is mediated via MC4 receptors, probably located in the NTS.

In summary, MC receptors expressed in the kidney and NTS appear to be involved in the effects mediated by MC receptors on blood pressure and heart rate. FMRF-amide receptors may mediate some of the effects of γ-MSH on increasing blood pressure, heart rate, and cerebral blood flow as well.

See also: Eating Disorders; Heart Rate Variability: A Neurovisceral Integration Model; Neuropeptides: Food Intake; Neuropeptides: Mental Disease; Proopiomelanocortin; Sexual Behavior: Neuroendocrine Control.

Further Reading

Adan RA (2006) Constitutive receptor activity series: Endogenous inverse agonists and constitutive receptor activity in the melanocortin system. *Trends in Pharmacological Sciences* 27(4): 183–186.

Adan RA and van Dijk G (2006) Melanocortin receptors as drug targets for disorders of energy balance. *Central Nervous System and Neurological Disorders – Drug Targets* 5(3): 251–261.

Adan RA and Gispen WH (2000) Melanocortins and the brain: From effects via receptors to drug targets. *European Journal of Pharmacology* 405(1–3): 13–24.

Adan RA, Tiesjema B, Hillebrand JJ, la Fleur SE, Kas MJH, and de Krom M (2006) The MC4 receptor and control of appetite. *British Journal of Pharmacology* 149(7): 815–827.

Benoit S, Schwartz M, Baskin D, Woods SC, and Seeley RJ (2000) CNS melanocortin system involvement in the regulation of food intake. *Hormones and Behavior* 37(4): 299–305.

Catania A, Gatti S, Colombo G, and Lipton JM (2003) Alpha-melanocyte stimulating hormone in modulation of inflammatory reactions. *Pediatric Endocrinology Reviews* 1(2): 101–108.

Chaki S and Okuyama S (2005) Involvement of melanocortin-4 receptor in anxiety and depression. *Peptides* 27(10): 1952–1964.

Cone RD (2005) Anatomy and regulation of the central melanocortin system. *Nature Neuroscience* 8(5): 571–578.

Humphreys MH (2004) Gamma-MSH, sodium metabolism, and salt-sensitive hypertension. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 286(3): R417–R430.

Huszar D, Lynch CA, Fairchild-Hutress V, et al. (1997) Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* 88(1): 131–141.

Starowicz K and Przewlocka B (2003) The role of melanocortins and their receptors in inflammatory processes, nerve regeneration and nociception. *Life Sciences* 73(7): 823–847.

Versteeg DH, Van Bergen P, Adan RA, and De Wildt DJ (1998) Melanocortins and cardiovascular regulation. *European Journal of Pharmacology* 360(1): 1–14.

Vrinten DH, Gispen WH, Groen GJ, and Adan RA (2000) Antagonism of the melanocortin system reduces cold and mechanical allodynia in mononeuropathic rats. *Journal of Neuroscience* 20(21): 8131–8137.

Wilkinson CW (2006) Roles of acetylation and other post-translational modifications in melanocortin function and interactions with endorphins. *Peptides* 27(2): 453–471.

Williams DL and Schwartz MW (2005) The melanocortin system as a central integrator of direct and indirect controls of food intake. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 289(1): R2–R3.

Melatonin Regulation of Circadian Rhythmicity in Vertebrates

D R Weaver, University of Massachusetts Medical School, Worcester, MA, USA
S W Lockley, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

In this article, we briefly outline the mechanisms of melatonin regulation in vertebrates, review effects of endogenous and exogenous melatonin on circadian rhythms, and discuss studies to determine the melatonin receptor subtype(s) mediating circadian responses to the hormone.

Melatonin, a Neuroendocrine Signal of Darkness

Regulation of Pineal Melatonin Production by the Circadian Clock and Light

Melatonin (*N*-acetyl-5-methoxytryptamine) is produced in the pineal gland from serotonin in a two-step enzymatic process. The first step, mediated by arylalkylamine *N*-acetyltransferase (AANAT), is rate limiting in the production of melatonin. AANAT activity is regulated by the circadian system, such that melatonin synthesis is elevated at night in virtually all vertebrate species. In the apparent absence of any storage mechanism, melatonin simply diffuses from the pineal into the circulation and cerebrospinal fluid. AANAT activity is normally high at night, but light exposure during the normal dark period leads to an acute and rapid reduction in AANAT protein levels, cessation of melatonin synthesis, and a rapid decline in blood levels of the hormone. In rodents, a very large-amplitude rhythm in AANAT transcript levels drives rhythmic melatonin production, yet studies of the effects of light exposure at night reveal that there is a fascinating and major posttranscriptional component. In darkness, sympathetic input to the pineal promotes melatonin synthesis and also actively blocks proteosomal degradation of AANAT. With light exposure, proteosomal degradation is disinhibited, leading to very rapid (<10 min) destruction of AANAT despite continued elevation of transcript levels. In primate and ungulate species, the AANAT transcript rhythm is low amplitude, and regulation of proteosomal proteolysis is thought to play a more major role in rhythmic (as well as light-regulated) melatonin secretion than transcriptional regulation. The switch-like regulation of proteosomal proteolysis is achieved through phosphorylation-dependent binding of AANAT to 14-3-3 proteins.

The primary role of melatonin is to provide a biochemical measure of night. Both night-active (nocturnal) and day-active (diurnal) animals produce melatonin at night, so any direct effects of melatonin on behavior are species specific. Although sometimes called the 'hormone of darkness,' melatonin production is not stimulated by darkness, nor does it require the presence of a light–dark cycle to stimulate its 24 h rhythm. The circadian rhythm of melatonin is driven endogenously by the circadian system. Critical structures in this system include the retinae, which detect light and convey lighting information to the circadian clock in the suprachiasmatic nucleus (SCN) of the hypothalamus. From the SCN, a circuitous polysynaptic pathway projects via the spinal cord, entering the cranium as sympathetic postganglionic fibers which are the ultimate mediators of input from the circadian system to regulate melatonin production.

The duration of the nocturnal melatonin secretory episode is related to the duration of the night; therefore, melatonin provides information about changes in photoperiod (season). Changes in the duration of melatonin production result in seasonally appropriate 'photoperiodic responses' including regulation of body composition, thermoregulation, pelage (fur) density and coloration, in addition to reproduction. These photoperiodic responses, mediated by seasonal changes in the melatonin secretory profile, are detailed elsewhere in this encyclopedia.

Extrapineal Sources of Melatonin Production

The rhythms in melatonin levels detected in plasma arise from the pineal in vertebrates, but melatonin is produced (and may have effects) locally in other tissues. In particular, melatonin is also synthesized rhythmically in the retina and influences retinal physiology. In some nonmammalian vertebrates, particularly in avian species, ocular melatonin seeps into the bloodstream and influences circadian function. The gastrointestinal tract can also synthesize melatonin, and here the production of melatonin may be tied more to precursor availability than to circadian or photic regulation of AANAT. Melatonin produced in the gut is not thought to contribute substantially to circulating levels of the hormone in normal circumstances.

Melatonin and Circadian Organization in Nonmammalian Vertebrates

An endogenous circadian system allows an organism to anticipate environmental events, rather than simply reacting to them, so that physiology and behavior are appropriately timed to both internal

and external events. Conceptually, the circadian timing systems can be divided into input, oscillator, and output components. Input pathways relay information about external (and internal) conditions to the pacemaker. The pacemaker oscillates, usually at a period slightly different from 24 h, and the timing of its oscillation can be altered by appropriately timed input stimuli. The output pathways are overt neural and hormonal rhythms that regulate physiology and behavior.

There is a fundamental difference between the role of melatonin in circadian organization in mammalian and that in nonmammalian species. The mammalian circadian timing system is dominated by a strong central circadian pacemaker, located in the SCN of the anterior hypothalamus. The major input pathway is a direct, retinohypothalamic tract that conveys lighting information from the retina. Efferents from the SCN regulate circadian melatonin production and also mediate the acute effects of light to suppress melatonin production. Although the presence of melatonin receptors in the SCN of many species suggests that melatonin could serve as an input, melatonin is viewed largely as an output of the circadian timing system in mammals. The melatonin rhythm is a strongly circadian rhythm that can be measured, for example, in humans with relative ease to determine a subject's biological time (the phase of rhythmicity). In contrast, nonmammalian vertebrates have a more complex circadian organization in which circadian photoreception occurs directly in several tissues in addition to the retina; these directly photoreceptive tissues include the brain and the pineal gland. Rhythmicity in these tissues is important for maintenance of rhythmicity in hypothalamic structures that appear functionally equivalent to the SCN.

Endogenous Melatonin Is Important for Circadian Organization in Nonmammalian Vertebrates

Melatonin plays a key role in coupling oscillators within the circadian timing system of several nonmammalian species. Indeed, the circadian system in several avian species has been described as a neuroendocrine loop, in which neural sites regulate melatonin, and melatonin in turn couples these sites together into a functional pacemaker which subsequently regulates behavior.

Key experiments that demonstrate the role of melatonin in circadian organization have involved removing sources of melatonin (the pineal gland and, in some cases, also the eyes, depending on species) and observing alterations in rhythms in constant conditions. The most compelling data come from sparrows, in which pinealectomy leads to arrhythmicity, and rhythmic melatonin replacement (by injection, feeding, or even transplantation of pineal tissue) leads to restoration of rhythmicity (Figure 1). In other species, changes in the circadian cycle length (period) occur in constant conditions after pinealectomy, and administration of exogenous melatonin can alter rhythm period or phase. This variability between species likely reflects species differences in the level of dependence on melatonin to couple oscillators and the relative autonomy of these areas.

As discussed elsewhere in this encyclopedia, circadian oscillations in vertebrates are based on an intracellular, cell-autonomous, molecular feedback loop. In the absence of melatonin, this transcriptional/translational feedback loop may persist within individual cells of nonmammalian vertebrates, but the coordination among oscillators may be lost, leading to loss of overt, physiological rhythmicity. Alternatively, the absence of melatonin may disrupt the molecular feedback loop within individual cells. Methods for introducing reporter genes into tissues and assessing rhythmic transcription within individual cells have been used to study mammalian cells; application of these methods to putative hypothalamic pacemaker cells from nonmammalian vertebrates will clarify the mechanism of melatonin action on circadian rhythmicity in these species.

Exogenous Melatonin Affects Circadian Rhythmicity in Nonmammalian Vertebrates

Administration of melatonin by means which result in levels sufficiently high to swamp out the endogenous rhythm of melatonin production (thereby removing the rhythmic internal signal that melatonin normally provides) usually leads to dramatic responses in nonmammalian vertebrates. These responses include loss of rhythmicity or significant alteration in circadian cycle length (period). Administration of supraphysiological doses of melatonin at 24 h intervals (by daily infusion, injection, or feeding) to animals with free-running rhythms in constant conditions, however, can lead to synchronization of rhythms to the 24 h period of administration. Although effects of exogenous melatonin on rhythms in nonmammalian vertebrates have been described in many publications, it should be recognized that negative data from similarly designed experiments may not have been published due to the inherent selection for 'positive data' in the peer-review process.

Melatonin Effects on Circadian Rhythms in Mammals

Endogenous Melatonin Has Little Impact on Circadian Rhythmicity in Mammals

As in nonmammalian species, the impact of pineal melatonin on rhythmicity can be assessed by removing

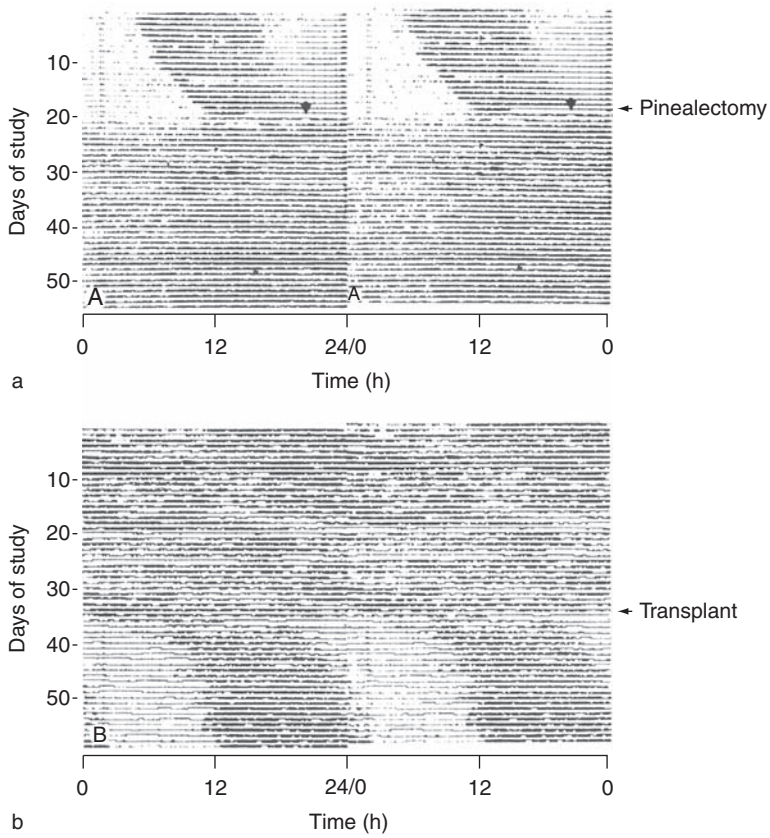

Figure 1 Pineal melatonin is important for circadian organization in house sparrows. (a, b) Perch-hopping activity of sparrows maintained in constant darkness plotted in actogram format. Each horizontal line represents 48 h of data. Data are double-plotted, meaning that each day of data occurs both to the right and below the previous day. Activity is indicated by vertical deflections from the baseline (thicker horizontal lines). (a) A sparrow with a free-running activity rhythm. Note immediate loss of circadian rhythmicity upon surgical removal of the pineal gland. (b) Rhythmicity is restored in a pinealectomized sparrow by transplantation of a pineal gland into the anterior chamber of the eye. The animal was housed in constant darkness except for a brief light pulse at the time of the transplantation surgery. (a) From Gaston S and Menaker M (1968) Pineal function: The biological clock in the sparrow. *Science* 160: 1125–1127. Copyright 1968 AAAS. Reproduced with permission of AAAS. (b) From Menaker M and Zimmerman N (1976) Role of the pineal in the circadian system of birds. *Integrative and Comparative Biology* 16: 45–55. Reproduced with permission from Oxford University Press.

the pineal gland. In contrast to nonmammalian vertebrates, however, pinealectomy has only subtle effects on behavioral circadian rhythms in mammals. In no mammalian species has pinealectomy been reported to lead to loss of rhythmicity or a major change in circadian organization. These data support a model in which the mammalian circadian timing system has a strong central pacemaker (in the SCN) and that coupling within the SCN occurs well even in the absence of melatonin. In this model, endogenous melatonin is relegated primarily to the role of an output rather than being both an output and an input. It should be recognized, however, that the pineal melatonin rhythm is important for regulating rhythmicity within a number of tissues endowed with melatonin receptors. Although these tissues do not contribute to regulation of locomotor activity rhythms, they may impact physiology substantially. For example, rhythmic gene expression in the pituitary pars tuberalis of both mice and hamsters is strongly affected by pineal melatonin. This tissue influences prolactin secretion from the pituitary gland, and in seasonally breeding species this appears to play an important role in regulating annual changes in gonadal status and pelage. Melatonin also appears to act directly in the dorsomedial hypothalamus of seasonal breeders to regulate gonadotrophin secretion. These studies emphasize the physiological importance of the circadian rhythm in pineal melatonin production for photoperiodic responses. Although the pineal melatonin rhythm may not have a striking impact on circadian organization in most mammals, pineal melatonin does have other important physiological functions in species attuned to respond to it.

Removal of melatonin may interfere with responsiveness to other environmental stimuli and therefore have an indirect effect on circadian organization. For example, pinealectomy of rats makes them more susceptible to disruption of their rhythms by constant light. Constant light alone lengthens circadian period and, at high intensities, results in loss of rhythmicity. Alterations in the rate of re-entrainment to

phase-shifted lighting cycles have been reported in pinealectomized animals. Although these results are consistent with a role for melatonin as a coupling agent between oscillators, they are also consistent with an alteration in the sensitivity to light following pinealectomy. These conflicting interpretations also affect studies examining light-induced resetting of the circadian clock, as alterations in sensitivity to light could alter the magnitude of phase shift following a light pulse.

The role of melatonin can also be assessed indirectly by examining animals that do not produce melatonin. Most inbred strains of mice lack the capacity to produce a high-amplitude rhythm in plasma melatonin levels due to defects in the synthetic pathway for melatonin. Despite this, mice have perfectly reasonable rhythms in locomotor activity, temperature, feeding, adrenal hormones, and other output markers of circadian function. Furthermore, mice with targeted disruption of MT1 and MT2 melatonin receptor subtypes have been generated, and these mice also have apparently normal behavioral rhythmicity in constant conditions. These studies indicate that melatonin is not a necessary component of the circadian timing system in mice. Similarly, inhibition of human melatonin production also does not appear to have a major impact on circadian organization. Evidence for this comes from human subjects with spinal cord injuries that disrupt the pathway between the SCN and the pineal gland and from hypertensive patients receiving β-adrenergic receptor antagonists that inhibit melatonin synthesis; in both situations, rhythmic melatonin production is lost, yet circadian rhythmicity persists (Figure 2).

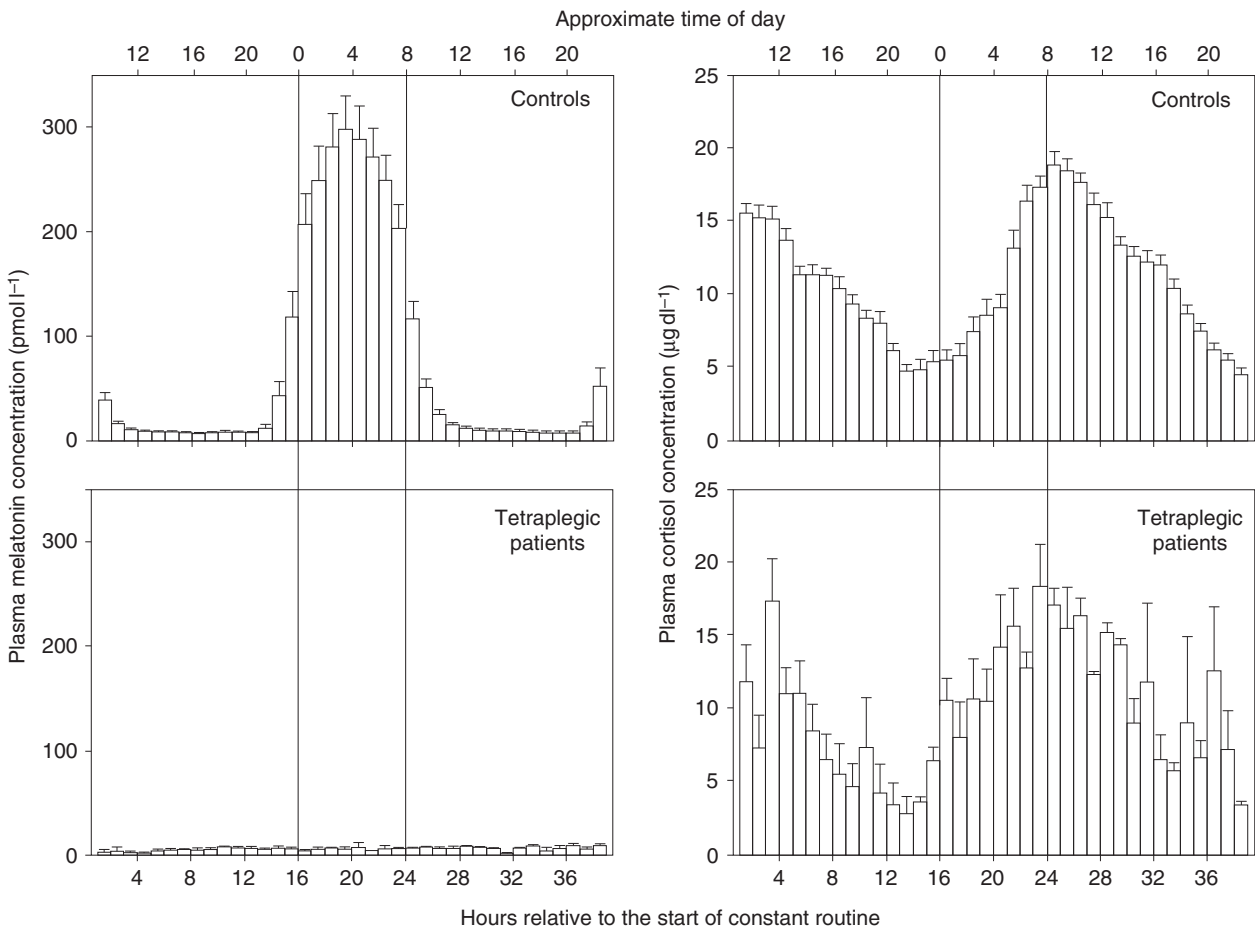

Figure 2 Circadian rhythms persist in humans despite disruption of the melatonin rhythm. The top panels show the melatonin (left) and cortisol (right) rhythms from control subjects ($n = 24$) during a constant routine procedure under dim lighting. Values represent mean ± SEM for data binned by hour for each subject. Normal circadian rhythms in the profiles for both hormones are readily apparent. Habitual sleep times (0:00–8:00 h) are represented by the open box. The lower panels show melatonin (left) and cortisol (right) data from three tetraplegic patients studied under the same conditions. Cervical transections were at C4, C6, and C6/C7. Cervical transection led to complete loss of melatonin secretion, yet the cortisol rhythm is preserved. Adapted from Zeitzer JM, Ayas NT, Shea SA, Brown R, and Czeisler CA (2000) Absence of detectable melatonin and preservation of cortisol and thyrotropin rhythms in tetraplegia. *Journal of Clinical Endocrinology and Metabolism* 85: 2189–2196.

Pharmacological Effects of Melatonin in Mammals

Although pineal melatonin does not play a necessary role in circadian organization in mammals, exogenous melatonin can nevertheless influence circadian rhythms in mammalian species including humans. With respect to melatonin effects on rhythmicity, the most extensively studied mammalian species is the rat. In this species, the free-running period in constant darkness is slightly more than 24 h (e.g., ~24.3 h). Timed, daily injections of melatonin (at 24 h intervals) are essentially without effect on free-running rhythms until the time of the injection slightly precedes the onset of daily activity. When melatonin is injected during this narrow window of sensitivity, it leads to small daily phase advances and thus to synchronization of activity onset to a 24 h period. Injection of saline vehicle does not produce this effect (Figure 3). Upon discontinuation of the daily melatonin injection (or substitution of saline), rats free-run from the time of the last injection, indicating they were indeed entrained by the daily injections (as opposed to being simply excited by the injection, resulting in increased activity every 24 h which could be misinterpreted as entrainment). Studies administering single injections of melatonin to rats free-running in constant conditions further demonstrate that the response depends on the phase of injection. Melatonin causes small (<1 h) phase advances when administered shortly before the time of activity onset. Some studies reveal a second window of responsiveness late at night.

Mice have also been used rather extensively for studies of circadian responses to melatonin, although their response magnitudes are quite small: a single injection at the time of peak sensitivity (CT10) does not cause a detectable phase shift, whereas three consecutive daily injections (at the same projected time, CT10) cause a phase advance averaging approximately 40 min (Figure 4). An additional difficulty in studying mice is that their free-running period is less that 24 h so that repeated administration of melatonin at 24 h intervals at a time that causes a phase advance is not expected to cause entrainment. Thus, additional paradigms (e.g., one to determine whether melatonin accelerates the rate of re-entrainment to a phase-shifted light–dark cycle) have been used in mice (Figure 5). Despite the small

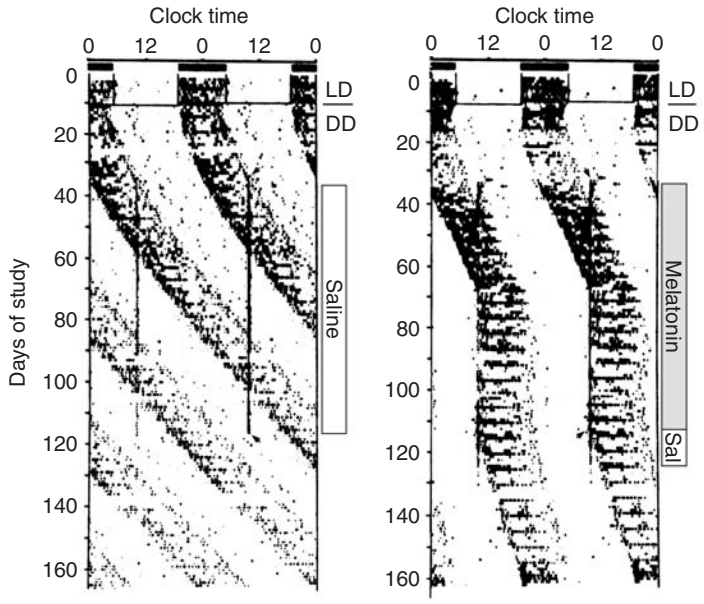

Figure 3 Entrainment of rat locomotor activity rhythms by timed daily injection of melatonin. Panels represent double-plotted actograms of wheel-running behavior. Female rats were initially housed in a 14 h light:10 h darkness (LD) lighting cycle and were transferred to constant darkness (DD) on day 11 of the study. Both rats show free-running rhythms with a cycle length of more than 24 h in DD, as is typical of rats. Beginning on day 34, each rat received one injection per day at 10:00 a.m. The animal receiving daily injections of the saline vehicle (left) continued to free-run throughout the period of study. In contrast, the animal receiving melatonin (100 $\mu g\,kg^{-1}$ subcutaneously) free-ran until the time of injection was approximately coincident with activity onset and then it 'locked on' to the melatonin injections. Entrainment to melatonin is accomplished by small phase advances in response to each daily injection. On day 115, daily injections of saline commenced; the animal lost entrainment despite continued handling and saline injection. These representative records show that melatonin can lead to phase-dependent and reversible entrainment of activity rhythms in rats. From Thomas EM and Armstrong SM (1988) Melatonin administration entrains female rat activity rhythms in constant darkness but not in constant light. *American Journal of Physiology* 255: R237–R242. Reproduced with permission of the American Physiological Society.

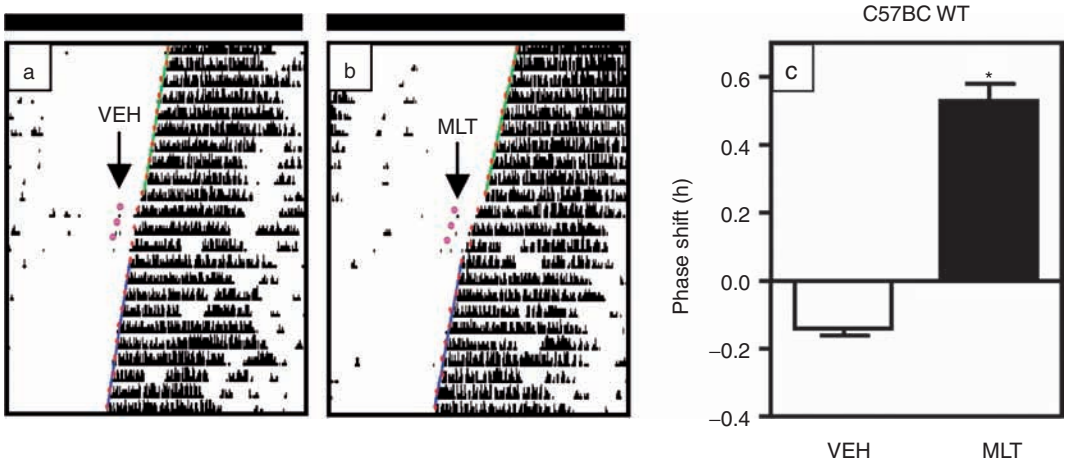

Figure 4 Melatonin can phase shift free-running circadian rhythms in mice. (a, b) Representative single-plotted actograms showing the phase-shifting effect of melatonin injections in mice. Mice were free-running in constant darkness and received injections (purple circles) of vehicle (VEH) (a) or melatonin (MLT; 90 μg per mouse) (b) on three consecutive days. Injections occurred at circadian time 10 (2 h before projected activity onset). Activity onsets (red circles) were determined for each circadian cycle, and the magnitude of phase shift for each animal was determined by comparing the time difference, on the day of treatment, between a line through activity onsets before treatment (green line) and a line through activity onsets after treatment (blue line). (c) Average responses ($n = 9$ per treatment). The asterisk indicates a significant difference between the two treatment groups ($p < 0.001$). C57BC WT refers to the strain origin of the animals; they were wild-type mice generated by intercrossing heterozygous MT1-deficient mice with a primarily C57BL/6J genetic background. From Dubocovich ML, Hudson RL, Sumaya IC, Masana MI, and Manna E (2005) Effect of MT1 melatonin receptor deletion on melatonin-mediated phase shift of circadian rhythms in the C57BL/6 mouse. *Journal of Pineal Research* 39: 113–120. Reproduced with permission from Blackwell Publishing.

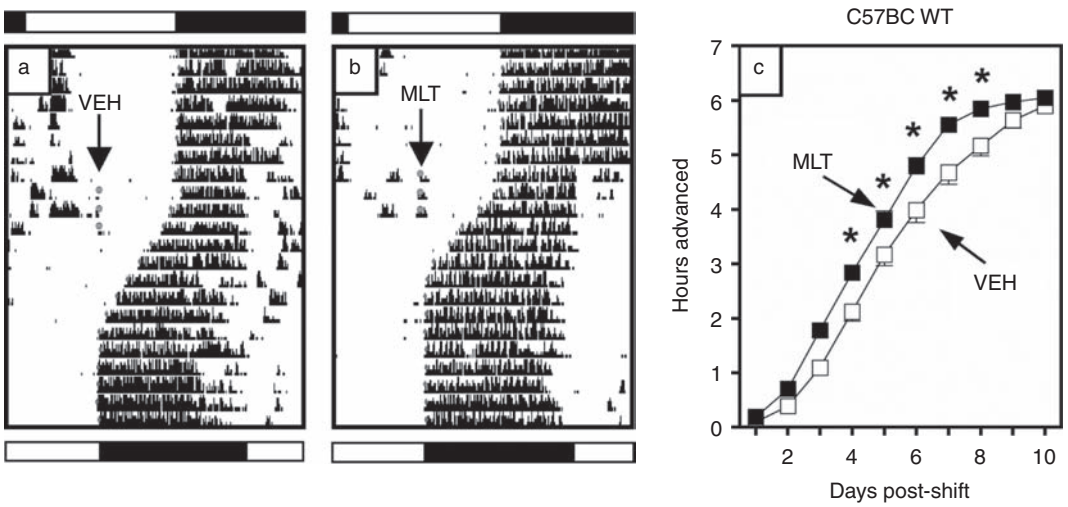

Figure 5 Melatonin accelerates re-entrainment of mice to a phase shift of the light–dark cycle. (a, b) Single-plotted actograms of representative animals receiving vehicle (VEH) (a) or melatonin (MLT; 90 μg/mouse/day) (b) in a re-entrainment paradigm. The original lighting cycle is shown at the top of the panel, whereas the shifted lighting cycle is shown at the bottom. Injections (purple circles, indicated by arrows) were administered 30 min prior to lights-out of the new lighting cycle, beginning on the first day of the advanced cycle and continuing for 3 days. Activity onsets (red) are indicated for each day, and fitted lines through onsets before and after the phase advance shift of the lighting cycle are shown. (c) Average responses. Data points are mean ± SEM hours of advanced activity onset on each day after the shift ($n = 20$ and 27 per treatment). The asterisk indicates a significant difference between the two treatment groups ($p < 0.05$). C57BC WT refers to the strain origin of the animals; they were wild-type mice generated by intercrossing heterozygous MT1-deficient mice with a primarily C57BL/6J genetic background. From Dubocovich ML, Hudson RL, Sumaya IC, Masana MI, and Manna E (2005) Effect of MT1 melatonin receptor deletion on melatonin-mediated phase shift of circadian rhythms in the C57BL/6 mouse. *Journal of Pineal Research* 39: 113–120. Reproduced with permission from Blackwell Publishing.

magnitude of circadian responses to melatonin in mice, studies in this species have been prompted by the availability of genetically modified mice with disruption of melatonin receptor subtypes.

A third nocturnal rodent species in which the effects of melatonin have been examined is the Syrian hamster. There is considerable discrepancy in the literature regarding the effects of melatonin in this

species. Studies using very large doses of melatonin report melatonin-induced alterations in circadian rhythms, whereas one study reveals that there is no greater response to melatonin than to vehicle and that the response attributed to melatonin in other work could be due to the effects of arousal related to injection.

Rhythmic administration of melatonin has also been found to entrain diurnal rodent species, including the Indian palm squirrel and the Nile grass rat. In these animals, there appear to be two windows of sensitivity so that entrainment can occur by either phase advances or phase delays. However, another diurnal species, the chipmunk, appears to be insensitive to entrainment by melatonin injections. Information on melatonin receptor subtype, distribution, and sensitivity is not available for these species, and the reasons for these large interspecies differences are unknown.

Melatonin Entrains Developing Hamsters

The most robust entrainment response to melatonin in mammals is the impact of melatonin on the developing Syrian hamster. In this species, and in several other rodents, maternal cues are able to set the phase of the circadian oscillator in the fetal SCN. This prenatal entrainment results in within-litter synchrony of circadian phase, which can be detected as synchronization of rhythms within the fetuses of a litter. Alternatively, synchronization can be detected by rearing the pups in constant conditions and then assessing the phase of rhythms during the postnatal period. Pineal *N*-acetyltransferase rhythms in 10-day-old rat pups and locomotor activity rhythms in several other species have been used to detect within-litter synchronization. Cross-fostering studies reveal that the major influence of the mother on the pups occurs during the prenatal and early postnatal periods. Destruction of the maternal SCN early in gestation leads to a litter in which the pups are not temporally coordinated, showing that the synchronization that normally occurs is not just due to the developing clocks being started or synchronized by a developmental event that the developing offspring share, such as birth.

Melatonin from the maternal circulation readily crosses the placenta, and rhythms in maternal melatonin are paralleled by rhythms in fetal melatonin levels. Furthermore, melatonin passes from mother to pup via the milk. Thus, melatonin is an attractive candidate for a hormonal cue that could mediate maternal effects on pup rhythmicity. Indeed, prenatal transfer of melatonin from mother to fetus has profound effects on pubertal development in Siberian hamsters. Studies in several species indicate that melatonin is not necessary for maternal entrainment of the fetus, however. Whereas endogenous melatonin does not appear necessary for entrainment of the developing circadian clock, exogenous melatonin has a profound entraining effect during development in Syrian hamsters. An elegant series of studies by Fred Davis and colleagues assessing entrainment of pups of SCN-lesioned mothers showed that (1) melatonin and D1-dopamine receptor activation entrain the developing circadian clock to opposite phases; (2) the minimal melatonin treatment is a single injection; and (3) there is a developmental window for the effect of melatonin, with loss of sensitivity to melatonin late in the first week of life. There is a marked downregulation in melatonin receptor density in the Syrian hamster SCN in the first week after birth, and this downregulation of receptors may limit the responsiveness of the SCN at later ages. Alternatively, there may simply be a transition of the SCN away from sensitivity to melatonin, similar to the loss of sensitivity to dopaminergic activation that occurs in rats during a comparable, early postnatal period. The efficacy of melatonin in synchronizing the pup phase in the offspring of SCN-lesioned dams indicates that melatonin is one important signal for entraining the developing circadian oscillator, but it is clearly not the only signal. The intrauterine environment is rich in other temporal cues (e.g., temperature, nutrients, and other hormones) and it appears that multiple redundant signals usually lead to entrainment of the fetus.

Experimental and Therapeutic Uses of Exogenous Melatonin in Humans

Studies demonstrating the impact of melatonin on rhythms in other species, and the presence of specific, high-affinity receptors for melatonin within the human SCN, have fuelled interest in the use of melatonin as a chronobiotic (an agent for resetting the oscillator regulating behavioral rhythms). One proposed application of melatonin is to facilitate re-entrainment after a phase shift of the light–dark cycle caused by transmeridian flight (e.g., a jet-lag pill). Of greater general importance is the proposed use of agents such as melatonin to rapidly facilitate adaptation to changes in the light–dark cycle induced by shift work. In view of the increasing prevalence and the negative health impact of shift work, this application looms very large in terms of the scope of the population that could benefit. The much smaller populations afflicted with other circadian rhythm sleep disorders, such as advanced and delayed sleep phase syndrome, and particularly totally blind people (who commonly suffer from non-24 h sleep–wake syndrome), could also benefit from a safe and effective chronobiotic.

Melatonin can indeed cause small phase shifts in humans, and phase–response curves describing both advances and delays have been published. The

maximum phase advance in response to melatonin occurs during the late afternoon to early evening, and the peak response is on the order of approximately 30 min. Phase delays occur when melatonin is administered in the early morning, with a similar magnitude of response. The mechanism of action of melatonin remains unclear, however. Melatonin has also been described as a sleep-promoting agent, and alteration in sleep regulation directly, or alteration of light–dark exposure when subjects close their eyes to sleep, may have a feedback effect on circadian clock function. Although melatonin could influence sleep by reducing the firing rate of SCN neurons, it could also do so by effects on the peripheral vasculature leading to changes in thermoregulation; thermoregulatory adjustments play an important role in sleep induction.

One situation in which even small phase shifts can be critically important is in the case of humans who are totally blind. In the absence of light information reaching the SCN, the circadian pacemaker cycles at its endogenous period, which is usually slightly more than 24 h (averaging ~24.5 h in blind subjects). Although a small minority of blind people without conscious light perception retain circadian photoreception, the majority do not and thus are insensitive to light both visually and circadianly. In these people, and in the majority of individuals lacking eyes (e.g., as a result of injury or surgical resection of cancer), the resultant lack of synchronization between the internal non-24 h circadian system and the external 24 h environment induces a chronic condition, non-24 h sleep–wake disorder. The disorder is characterized by cyclic episodes of good sleep, followed by poor nighttime sleep and excessive daytime sleepiness, as the individual free-runs and thus experiences alternating periods of his or her biological day being coincident with, and later being out of phase with, the 24 h day. In these individuals, a chronobiotic that achieves a daily phase advance equal to the difference between their free-running period and the 24 h day (e.g., 0.5 h per day, on average) can lead to stable entrainment, immediately correcting the debilitating sleep disorder. Indeed, this has been found in some studies involving daily melatonin administration (Figure 6(a)). Notably, remarkable variation between blind subjects has been noted in the response to timed daily melatonin administration (Figure 6). The initial phase of treatment, the dose of melatonin, and the individuals' circadian period all appear to influence the ability of blind subjects to entrain to daily melatonin administration. Smaller doses (0.5 mg or lower) appear to be at least or even more effective than larger doses (5 mg or higher), and at such higher doses entrainment appears to depend on initiating treatment at the correct phase. Individuals with a period closer to 24 h also may be more likely to respond positively due to the relatively smaller daily shift required to entrain to a 24 h period. Entrainment to timed daily melatonin administration remains the only treatment available to totally blind subjects with non-24 h sleep–wake disorder, and although the small number of trials conducted to date are not all consistent, melatonin treatment has been proven in principle and is a promising solution to this particular disorder.

The small magnitude of phase shift caused by melatonin would seem to be inconsistent with proposed therapeutic applications of melatonin that require a much larger amplitude of response (e.g., for achieving a 5 h phase shift after a transatlantic flight). However, melatonin and light exposure may synergize in their effects on the circadian clock in sighted individuals. The phase–response curves to melatonin and light are approximately opposite and the light responses are larger than the effects of melatonin. Melatonin has been shown to hasten re-entrainment to a phase-shifted lighting cycle in simulated (and actual) jet lag and shift work paradigms. Whether this action is due to direct effects on the SCN circadian clock and/or via changes in sleep–wake or light–dark cycles remains to be established. Given the dual effects of melatonin, its administration should be timed to enhance both its chronobiotic and its potential sleep-facilitating effects.

Melatonin Receptors

Identification of Melatonin Receptor Subtypes

Radioactively labeled 2-iodomelatonin (2-[^{125}I]iodomelatonin, ^{125}I-Mel) has been used to identify and characterize high-affinity melatonin binding sites ($K_D < 200$ pM). High-affinity melatonin receptors are coupled to G-proteins because GTP analogs inhibit high-affinity ^{125}I-Mel binding. Autoradiographic methods to localize sites of ^{125}I-Mel binding led to detection of site-specific distribution of binding sites, including a high concentration of melatonin receptor binding in the SCN of most (but not all) species (Figure 7(e)). Melatonin affects second messenger pathways, particularly G_i-mediated inhibition of forskolin-stimulated cAMP accumulation, in those tissues in which the receptor concentration is relatively high.

In addition to high-affinity sites, a lower affinity 2-[^{125}I]iodomelatonin binding site with K_D values in the 1–2 nM range was identified in hamster tissues, with a fundamentally different pharmacological profile, most notably a relatively higher affinity for *N*-acetylserotonin and prazosin, and guanine nucleotides do

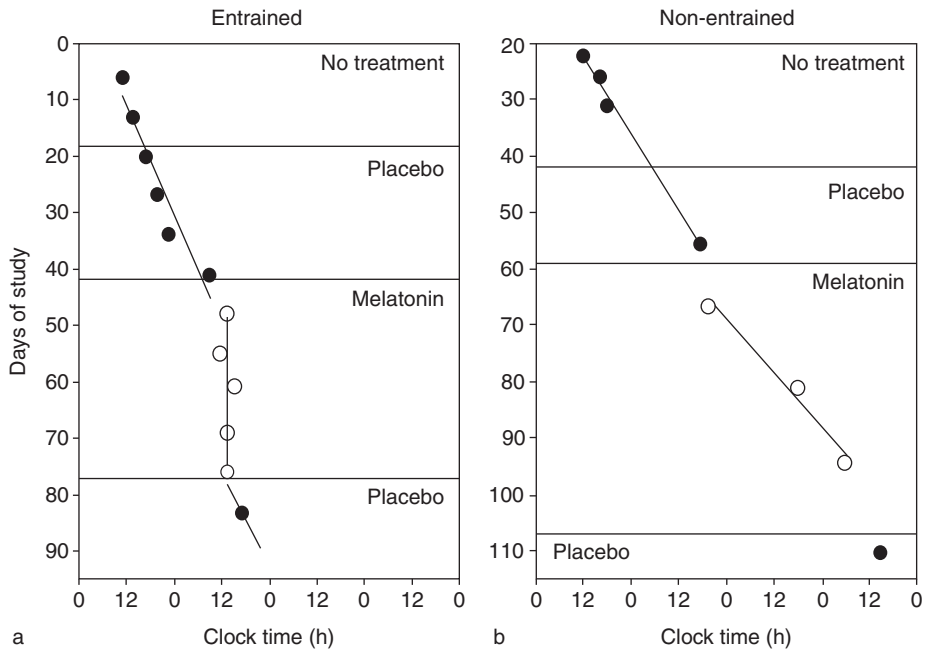

Figure 6 Variable responsiveness of blind humans to daily melatonin treatment. (a and b) Two totally blind men with free-running circadian rhythms received melatonin (5 mg melatonin *per os* at 21:00 h) (open circles) and no treatment/placebo (solid circles) for at least one full circadian cycle. Horizontal lines indicate transitions between treatments. Data points represent the peak times of the urinary cortisol rhythm, assessed by weekly collections of urine at 4 h intervals (8 h at night) for 24–48 h. Lines represent the period of the cortisol rhythm during various phases of treatment, determined using best fit linear regression. Period estimates from the cortisol rhythm were comparable to period estimates based on melatonin metabolite rhythms before treatment. (a) Subject 31 initially had a free-running rhythm with a period of 24.57 h. Melatonin treatment began at 21:00 h, at a time that corresponded to an internal circadian time (CT) of CT16, approximately 2 h after the onset of endogenous plasma melatonin production and at a phase known to cause an advance shift. Melatonin treatment entrained the cortisol rhythm to a 24 h (24.0 ± 0.2 h) period and to a normal phase (12.9 ± 1.2 h). One week after cessation of melatonin treatment (day 83), the cortisol phase occurred at 17.1 h, indicating that the subject reverted to a free-running period of more than 24 h after discontinuation of melatonin treatment. (b) Subject 45 initially had a free-running rhythm with a period of 24.87 h as determined by the urinary cortisol rhythm. Melatonin treatment began at 21:00 h, at a time corresponding to CT5 (approximately 9 h before the onset of endogenous plasma melatonin production). Melatonin treatment failed to entrain the cortisol rhythm (period during melatonin treatment, 25.22 ± 2.59 h) despite being administered for nearly two full circadian cycles. Based on data from Lockley SW, Skene DJ, James K, et al. (2000) Melatonin administration can entrain the free-running circadian system of blind subjects. *Journal of Endocrinology* 164: R1–R6.

not affect binding to these sites. The first melatonin receptor nomenclature was introduced to distinguish high-affinity sites that correlated with functional activity (ML-1) from the lower affinity, prazosin-sensitive sites (ML-2).

Work by Steve Reppert and colleagues led to the isolation and characterization of cDNAs encoding a family of high-affinity G-protein-coupled melatonin receptor subtypes, all of which fall within the pharmacologically characterized ML-1 receptor class and all of which inhibit forskolin-stimulated cAMP accumulation when expressed in cell lines. Thus, the melatonin receptors called Mel1a (now MT1) and Mel1b (now MT2), as well as the nonmammalian Mel1c receptor subtype, are all 'ML-1' type by the older nomenclature. Mice with targeted disruption of the MT1 and MT2 receptor subtypes have been generated and used to assess receptor subtypes mediating responses to melatonin.

The unique pharmacological profile of the ML-2 site allowed development of a specific radioligand and the eventual purification of the ML-2 site, leading to its identification as the enzyme, quinone reductase type 2 (QR2). In the current receptor nomenclature, the 'ML-2' binding site was renamed the MT3 melatonin binding site. That QR2 binding accounts for the MT3 binding sites is most conclusively shown by the absence of MT3 type binding in QR2 knockout mice.

A mammalian melatonin receptor-related receptor (now called GPR50) has also been identified. This receptor is structurally related to the family of G-protein-coupled melatonin receptors, having a similar amino acid sequence and an identical exon/intron structure. Human, mouse, and sheep homologs of this receptor have been identified, but none is capable of binding ^{3}H-melatonin or 2-[^{125}I]iodomelatonin; GPR50 remains an orphan receptor.

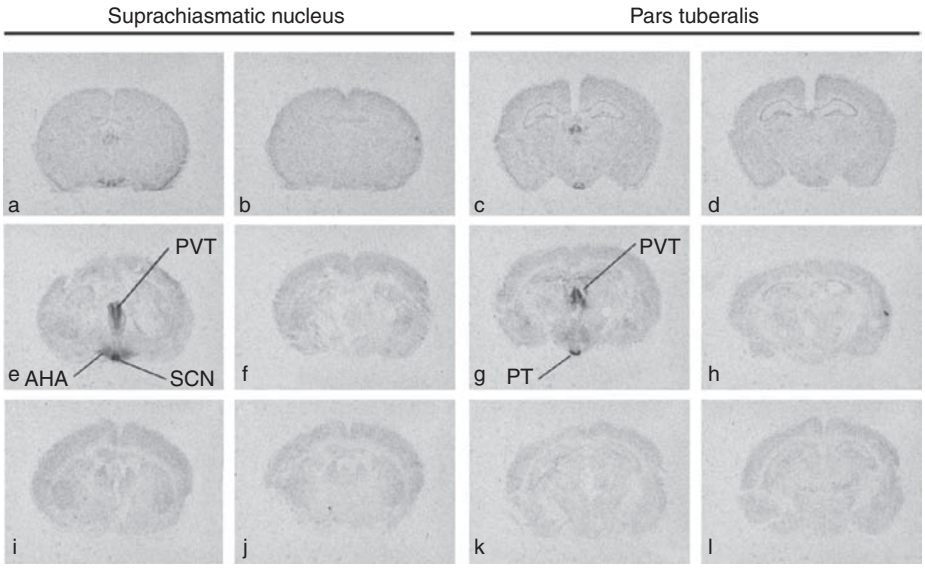

Figure 7 Melatonin receptor gene expression and receptor binding in mouse brain. (a–d) Brain sections processed for *in situ* hybridization. Panels (a) and (c) were hybridized with the antisense probe to detect MT1 melatonin receptor transcript, whereas panels (b) and (d) are adjacent sections hybridized with the sense probe (as a control). Expression of MT1 receptor mRNA is present in areas that coincide with regions containing melatonin receptor binding (see (e) and (g)). (e–l) Melatonin receptor binding assessed by incubation of brain sections with 100 pM 2-[^{125}I]iodomelatonin. Panels (e–h) are sections from a wild-type mouse, whereas panels (i–l) are from an MT1 receptor knockout mouse. Sections shown in panels f, h, j, and l were incubated with 100 pM 2-[^{125}I]iodomelatonin plus an excess (1 μM) of melatonin to determine nonspecific binding. Mice with targeted disruption of the MT1 receptor gene lack readily detectable melatonin receptor binding by autoradiographic methods. AHA, anterior hypothalamic area; PT, pars tuberalis; PVT, paraventricular nucleus of the thalamus; SCN, suprachiasmatic nucleus. Panels (a–d) are based on unpublished data of S Haddady, CM Lambert, and DR Weaver confirming findings reported in Roca AL, Godson C, Weaver DR, and Reppert SM (1996) Structure, characterization, and expression of the gene encoding the mouse Mel-1a melatonin receptor. *Endocrinology* 137: 3469–3477. Panels (e–h) are based on experiments conducted by DR Weaver and reported in Liu C, Weaver DR, Jin X, et al. (1997) Molecular dissection of two distinct actions of melatonin on the suprachiasmatic circadian clock. *Neuron* 19: 91–102.

Melatonin Receptor Subtype Expression in SCN

As noted previously, there are two high-affinity, G-protein-coupled receptors for melatonin in mammals – the MT1 and MT2 receptor subtypes. The MT1 receptor transcript is expressed in mammalian brain and pituitary (pars tuberalis) in a manner that agrees well with the distribution of detectable melatonin receptor binding sites (Figure 7). Furthermore, mice with targeted disruption of the MT1 receptor lack detectable melatonin receptor binding by autoradiographic methods (Figures 7(i) and 7(k)), whereas mice with disruption of the MT2 receptor have no detectable decrement in melatonin receptor binding. The MT2 receptor transcript appears to be broadly but weakly expressed, and the MT2 receptor protein appears to be present at rather low levels. Thus, the majority of melatonin receptor binding is due to MT1 rather than MT2 receptor expression.

Molecular Dissection of Circadian Effects of Melatonin

Four SCN-mediated responses have been examined in an effort to dissect the melatonin receptor subtype(s) affecting the SCN. One robust SCN response to melatonin is an acute inhibition of electrical activity, detected by multiunit recording from SCN slices maintained *in vitro* (Figure 8). This inhibitory response to melatonin is mediated by the MT1 receptor. The response is completely absent in MT1 receptor knockout mice and is unaffected in MT2 receptor knockout mice. A second *in vitro* SCN response to melatonin involves monitoring the phosphorylation state of calcium and cyclic AMP response element binding protein (CREB). Exposure of SCN slices to pituitary adenylate cyclase activating peptide (PACAP) late during the subjective day leads to an increase in the number of neurons in which CREB is phosphorylated. Pretreatment of SCN slices with melatonin at concentrations of 1–10 nM blocks the PACAP-induced increase in CREB phosphorylation. Mice with targeted disruption of the MT1 receptor respond normally to PACAP but do not respond to melatonin (1 nM) (Figure 9). At higher concentrations of melatonin, however, inhibition of CREB phosphorylation occurs even in MT1 receptor-deficient mice. This higher threshold response to melatonin is blocked by the MT2 receptor antagonist 4P-PDOT and is also

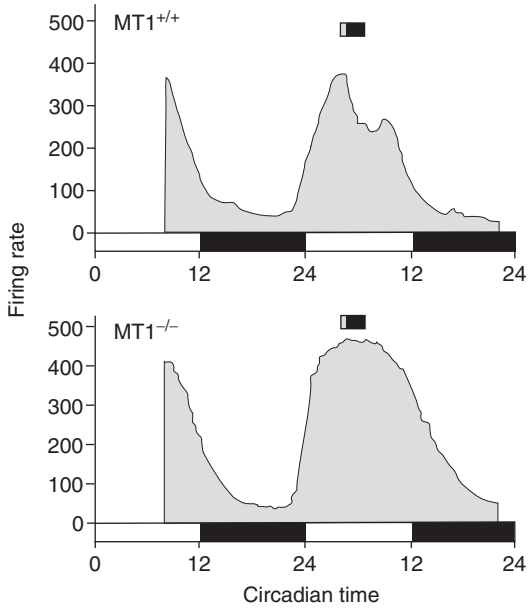

Figure 8 Melatonin inhibits SCN neuronal firing through the MT1 melatonin receptor. Schematic of representative data. Slices containing the SCN were prepared from a wild-type (MT1+/+) and an MT1 receptor-deficient mouse (MT1−/−) and were monitored by multiunit recording of electrical activity for approximately 36 h. The rate of firing (in spikes per second) is indicated by the black line with gray fill below. On the second day *in vitro*, the slices were exposed for 30 min intervals to vehicle (gray box) and then to melatonin (black box) at escalating doses (0.1, 1, 10, and 100 nM) during the circadian day. Inhibition of electrical activity occurs in the slice from the wild-type mouse but not in the slice from the MT1 receptor-deficient mouse. The lighting cycle to which the animals were exposed before slice preparation is shown at the bottom of each panel. Based on data reported in Liu C, Weaver DR, Jin X, et al. (1997) Molecular dissection of two distinct actions of melatonin on the suprachiasmatic circadian clock. *Neuron* 19: 91–102.

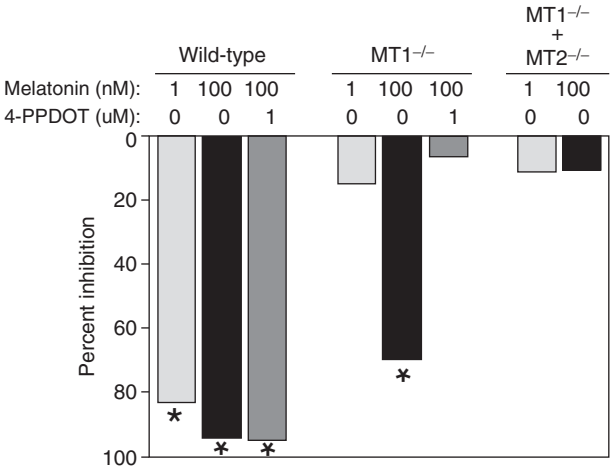

Figure 9 Melatonin inhibits PACAP-induced CREB phosphorylation through the MT1 receptor at low concentrations and through the MT2 receptor at higher concentrations. SCN slices from mice of the genotypes shown were treated with PACAP 2 h before lights-out and fixed 15 min later. Phospho-CREB immunoreactive nuclei in the SCN were assessed by immunohistochemistry. Data represent the percentage inhibition in the number of phospho-CREB-labeled neurons in the SCN. Percentage inhibition is expressed relative to slices treated with PACAP alone (0% inhibition) or the vehicle alone without PACAP (100% inhibition). Standard error values were within 10% for each experiment and are not shown for clarity. To combine data from experiments reported in two different publications, the results of the MT1−/− groups from the two papers have been averaged. Responses to low concentrations (1 nM) of melatonin require the MT1 receptor. A response to higher melatonin concentrations (100 nM) occurs in the absence of the MT1 receptor; this response is blocked by an MT2 receptor antagonist (4P-PDOT) and is absent in double-knockout mice. Asterisks indicate significant inhibition ($p < 0.05$). Based on data reported in von Gall C, Weaver DR, Kock M, et al. (2000) Melatonin limits transcriptional impact of phosphoCREB in the mouse SCN via the Mel_{1a} receptor. *NeuroReport* 11: 1803–1807; and Jin X, von Gall C, Pieschl RL, et al. (2003) Targeted disruption of the mouse Mel_{1b} melatonin receptor. *Molecular and Cellular Biology* 23: 1054–1060.

absent in MT1/MT2 double-knockout mice (Figure 9). Thus, the regulation of CREB phosphorylation state is controlled primarily by the MT1 receptor, with the MT2 receptor having an influence only at higher melatonin concentrations.

Curiously, a third *in vitro* response of the SCN to melatonin suggests a primary role of the MT2 receptor. This paradigm involves application of melatonin to slices prepared from the SCN, followed by monitoring the electrical activity of single SCN neurons over the following day to determine the time of the peak in single unit firing rate. In this assay system, melatonin produces a 3 or 4 h phase advance when applied during a sensitive window during late subjective day (at CT10). This response to melatonin is preserved in MT1 receptor knockout mice, but it is blocked by pertussis toxin, suggesting involvement of a G-protein-coupled receptor (presumably the MT2 receptor). Studies in rats suggest that this *in vitro* phase-shifting response is blocked by a competitive MT2 receptor antagonist. However, Siberian hamsters (which naturally have a defect in the MT2 gene leading to the absence of functional MT2 receptors) also respond to melatonin in this assay. A very peculiar aspect of the single unit response to melatonin is the inexplicably high potency of melatonin: 3–10 pM concentrations of melatonin will cause a phase shift in the single unit assay, yet concentrations 300- to 3000-fold higher (1–10 nM) are required for most other responses assessed in SCN slices *in vitro*. It is not clear why the response to melatonin in the single unit assay is so much larger than the *in vivo* response, or why it occurs at concentrations of melatonin that are so low. Furthermore, the receptor mechanism for this *in vitro* phase-shifting response does not seem consistent with results obtained *in vivo*. The single unit assay may monitor a subset of SCN

neurons that are very highly responsive to melatonin but that appear not to correlate well with other *in vitro* responses to the hormone.

Phase-shifting responses to melatonin have also been described in intact mice. As noted previously (Figure 4), three consecutive daily injections (at the same projected time) cause a phase advance averaging approximately 40 min in wild-type mice. Dubocovich and colleagues have shown that this response to melatonin is absent in MT1 receptor-deficient mice. Similarly, this group reports that the modest acceleration of re-entrainment to a phase-shifted lighting cycle (Figure 5) is absent in MT1 receptor-deficient mice. In contrast, an earlier study using antagonist administration indicated that the MT2 receptor was necessary for the phase advance in response to melatonin injections. As noted previously, the most profound circadian response to melatonin administration in mammals occurs in developing Syrian hamsters, a species that naturally lacks a functional MT2 receptor. Thus, the literature is contradictory with respect to the receptor subtype most important for SCN responses to melatonin. Although there are exceptions, studies aimed at identifying the receptor subtypes responsible for responses to melatonin generally indicate that the major effects of melatonin are mediated by the MT1 melatonin receptor subtype. There has been interest in melatonin and in the development of melatonin analogs for chronobiotic and sleep-promoting effects. It is not clear whether there would be any advantage of subtype-selective pharmaceuticals affecting melatonin receptors.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Organization in Non-Mammalian Vertebrates; Circadian Function and Therapeutic Potential of Melatonin in Humans; Entrainment of Circadian Rhythms by Light; Pineal Gland and Melatonin; Seasonal Timing: Neural Mechanisms.

Further Reading

Arendt J (1995) *Melatonin and the Mammalian Pineal Gland*. London: Chapman and Hall.

Cassone VM (1990) Effects of melatonin on vertebrate circadian systems. *Trends in Neuroscience* 13: 457–464.

Cassone VM and Menaker M (1984) Is the avian circadian system a neuroendocrine loop? *Journal of Experimental Zoology* 232: 539–549.

Czeisler CA and Turek FW (1997) *Special Issue: Melatonin, Sleep and Circadian Rhythms: Current Progress and Controversies*. *Journal of Biological Rhythms* 12: 485–708.

Davis FC (1997) Melatonin: Role in development. *Journal of Biological Rhythms* 12: 498–508.

Dubocovich ML, Hudson RL, Sumaya IC, Masana MI, and Manna E (2005) Effect of MT1 melatonin receptor deletion on melatonin-mediated phase shift of circadian rhythms in the C57BL/6 mouse. *Journal of Pineal Research* 39: 113–120.

Gastel JA, Roseboom PH, Rinaldi PA, Weller JL, and Klein DC (1998) Melatonin production: Proteasomal proteolysis in serotonin *N*-acetyltransferase regulation. *Science* 279: 1358–1360.

Gaston S and Menaker M (1968) Pineal function: The biological clock in the sparrow. *Science* 160: 1125–1127.

Jin X, von Gall C, Pieschl RL, et al. (2003) Targeted disruption of the mouse Mel_{1b} melatonin receptor. *Molecular and Cellular Biology* 23: 1054–1060.

Lincoln GA, Clarke IJ, Hut RA, and Hazlerigg DG (2006) Characterizing a mammalian circannual pacemaker. *Science* 314: 1941–1944.

Liu C, Weaver DR, Jin X, et al. (1997) Molecular dissection of two distinct actions of melatonin on the suprachiasmatic circadian clock. *Neuron* 19: 91–102.

Lockley SW, Skene DJ, James K, et al. (2000) Melatonin administration can entrain the free-running circadian system of blind subjects. *Journal of Endocrinology* 164: R1–R6.

Mailliet F, Ferry G, Vella F, et al. (2004) Organs from mice deleted for NRH:quinone oxidoreductase 2 are deprived of the melatonin binding site MT3. *FEBS Letters* 578: 116–120.

Menaker M and Zimmerman N (1976) Role of the pineal in the circadian system of birds. *Integrative and Comparative Biology* 16: 45–55.

Roca AL, Godson C, Weaver DR, and Reppert SM (1996) Structure, characterization, and expression of the gene encoding the mouse Mel-1a melatonin receptor. *Endocrinology* 137: 3469–3477.

Thomas EM and Armstrong SM (1988) Melatonin administration entrains female rat activity rhythms in constant darkness but not in constant light. *American Journal of Physiology* 255: R237–R242.

von Gall C, Stehle JH, and Weaver DR (2002) Mammalian melatonin receptors: Molecular biology and signal transduction. *Cell and Tissue Research* 309: 151–162.

von Gall C, Weaver DR, Kock M, et al. (2000) Melatonin limits transcriptional impact of phosphoCREB in the mouse SCN via the Mel_{1a} receptor. *NeuroReport* 11: 1803–1807.

Weaver DR (1999) Melatonin and circadian rhythmicity in vertebrates: Physiological roles and pharmacological effects. In: Turek FW and Zee PC (eds.) *Regulation of Sleep and Circadian Rhythms*. New York: Dekker.

Zeitzer JM, Ayas NT, Shea SA, Brown R, and Czeisler CA (2000) Absence of detectable melatonin and preservation of cortisol and thyrotropin rhythms in tetraplegia. *Journal of Clinical Endocrinology and Metabolism* 85: 2189–2196.

Memory Consolidation: Cerebral Cortex

B Bontempi, Université de Bordeaux 1, Talence, France
P W Frankland, Hospital for Sick Children, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Converging evidence from clinical and animal research supports the idea that our memories of everyday life – of people, places, and events – are not formed instantaneously but are gradually stabilized over time via a process called memory consolidation. Müller and Pilzecker were the first to adopt the term consolidation in 1900; just as in the definition in the dictionary, it means to make stronger or firm. However, the consolidation process must not be considered to be a unitary and passive mechanism enabling new memories to be transformed from an initially labile state to a more permanent state. On the contrary, memory consolidation is a highly dynamic process that involves large-scale reorganization at both the synaptic and entire interconnected brain system levels.

Recent studies have shed light on the dynamics of hippocampal–cortical interactions that allow recently acquired memories to be transformed into enduring memories in distributed cortical networks. Here we review these recent advances by presenting experimental evidence that forms the bases for two competing models of memory consolidation. Mechanisms underlying cortical storage are also briefly discussed, with a particular emphasis on the role of the prefrontal cortex, which seems to play a privileged role in processing remote memories.

Time Course of Memory Consolidation

Two main forms of consolidation can be distinguished on the basis of their temporal dynamics. Synaptic consolidation primarily involves morphological changes within localized neuronal circuits, such as the restructuring of existing synaptic connections or the growth of novel ones. Such changes, which occur rapidly after a learning episode (within minutes to several hours), are triggered by a cascade of different metabolic events which include activation of central receptors, release of different neurotransmitters and second messengers, activation of transcription factors and associated target genes, and ultimately the synthesis of a plethora of proteins directly responsible for the observed structural changes. Over the past decade, much has been learned about the cellular and molecular bases of this process. By contrast, system consolidation is a much slower process which can last for weeks to months, or even up to several decades, depending on the species. Consolidation in this case refers to a gradual process of reorganization and a differential involvement of the brain regions that support memory processing. For instance, the brain regions that are crucial for processing recently acquired information are not necessarily identical to those that are important for later memory recall. The mechanisms underlying this reorganization are largely unknown.

What are the brain regions critical for recent and remote memory processing? A pioneer in the field was the French psychologist Ribot, who pointed out as early as 1881 that memory loss after a brain trauma was often related to the age of the memory. He concluded, based on clinical observations, that retrieval of recent memory was typically more impaired than that of older (or more remote) memory, suggesting that memories may be reorganized in a time-dependent manner. It is only later, during the mid-twentieth century that refined anatomical approaches made it possible to identify damage to the medial temporal lobe, including the hippocampus, as being responsible for the occurrence of this typical graded amnesia. Certainly one of the most famous patients in the field of cognitive neuroscience is patient H.M., who suffered from intractable epileptic seizures whose focus was located within the medial temporal lobe. It was decided at the time to excise large amounts of his medial temporal lobe bilaterally. Although the surgery was successful in abolishing the epilepsy, one of the unexpected consequences was that H.M. exhibited severe retrograde amnesia, that is, the loss of memory for information acquired prior to the onset of amnesia. H.M.'s surgery also produced persistent anterograde amnesia, or the inability to form new memories. Particularly interesting was the fact that H.M.'s retrograde amnesia was temporally graded just as previously described by Ribot – memories acquired recently were more affected than memories acquired longer ago during his childhood. This particular pattern of memory impairment together with other case studies was interpreted as showing that the hippocampus and related structures of the medial temporal lobe have a time-limited role in memory storage and that memories, once consolidated, must be stored elsewhere. This idea forms the central tenet of most contemporary views of system consolidation; the hippocampus acts as a temporary consolidation organizing device, but storage of enduring memories depends on broadly distributed extra-hippocampal networks, presumably located within the cortical regions.

Models of Memory Consolidation

In the standard model of memory consolidation, illustrated in Figure 1, the hippocampus is believed to rapidly integrate and bind information from distributed cortical neurons that support the various features of a whole experience to form a coherent memory trace. Consolidation of this new memory trace at the cortical level then occurs slowly via reactivation of hippocampal–cortical networks, eventually leading to the increase in the strength and stability of cortical–cortical connections. Over time, as memories mature, the role of the hippocampus gradually diminishes until structured cortical networks become capable of retrieving remote memories independently.

However, it has been argued that not all the experimental evidence supports this model. First, retrograde amnesia is not always temporally graded. Flat gradients where both recent and remote memories are similarly impaired have been observed. Second, certain gradients have been reported to last for decades or up to almost the entire human life span in some amnesic patients. These observations have prompted some researchers to challenge the standard model of memory consolidation and to propose an alternative view, known as the multiple trace theory, in which the hippocampus is necessary for memory recall as long as memories are viable. According to this theory, each memory trace, which corresponds to a specific conscious experience, consists of a cohesive hippocampal–cortical ensemble. Each time that a particular event is recalled, this event is recoded in the form of multiple related memory traces dispersed over larger areas of the hippocampus. Consequently, the relative sparing of remote memories in amnesic patients is a function of the extent of hippocampal damage, with limited damage producing temporally graded retrograde amnesia and extensive lesions resulting in a flat gradient for retrograde amnesia. Another interesting feature of this model is that the form of memory being processed is thought to determine the time course of hippocampal engagement during consolidation. Whereas the standard model favors a disengagement of the hippocampus regardless of the type of declarative memory (memory allowing the conscious recollection of facts and events) being processed, the multiple trace theory predicts that complete hippocampal lesions will produce temporally graded retrograde amnesia only for semantic (general knowledge of facts) but not episodic (richly detailed recollection of specific events) memory. Some case studies, for instance, patient E.P., who has extensive bilateral lesions of the medial temporal lobe but excellent autobiographical memories from his youth, do not support this prediction.

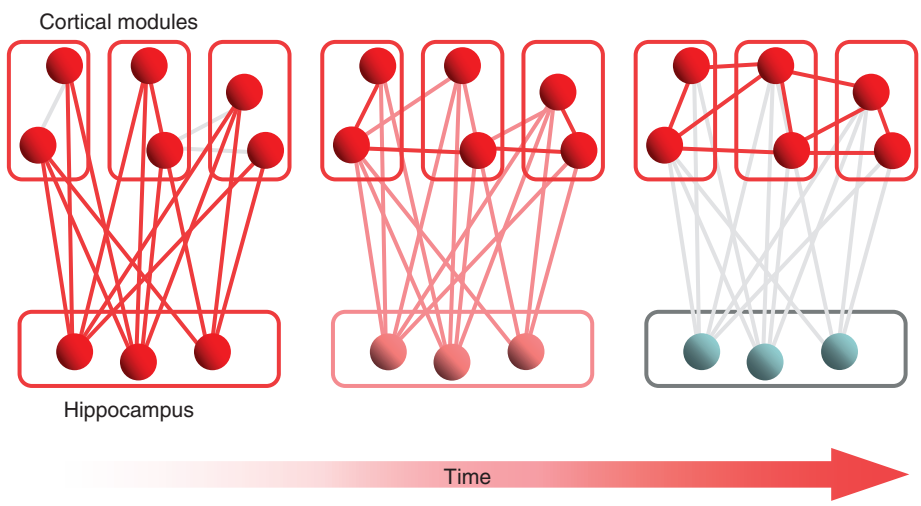

Figure 1 The standard model of memory consolidation. Perceptual, motor, and cognitive information is initially processed by a large number of specialized primary and associative cortical areas represented here by cortical modules. The medial temporal lobe, including the hippocampus, integrates the various features of an experience and binds them rapidly into a coherent memory trace. System consolidation then occurs slowly over time and involves a hippocampal–cortical dialog to gradually strengthen cortical–cortical connections. This involves the strengthening of existing cortical–cortical connections and the creation of new connections between previously unconnected neurons. As cortical memories mature and gain stability, the role of the hippocampus gradually diminishes allowing structured cortical networks to retrieve remote memories independently. A key feature of this model is that changes in the strength of hippcampal–cortical connections are fast but short-lasting whereas changes within cortical–cortical connections are slow but long-lasting. Plasticity phenomena at both hippocampal and cortical sites are likely to involve Hebbian-like mechanisms, that is increased synaptic strength between two neurons where the first neuron repeatedly stimulates the second one. From Frankland PW and Bontempi B (2005) The organization of recent and remote memories. *Nature Reviews Neuroscience* 6: 119–130.

Furthermore, there are several examples of temporal gradients in animals even after complete hippocampal lesions. A central issue is whether remote episodic memories following extensive hippocampal lesions are as detailed and vivid as those in control subjects.

Experimental Evidence

In Animals

Temporally graded retrograde amnesia has been shown across a variety of species, suggesting that time-dependent memory reorganization is an evolutionarily conserved process. Designing experiments in laboratory animals to study this phenomenon offers a numbers of advantages. First, what is learned and when it is learned can be rigorously controlled across times in a prospective manner, which is difficult to achieve in humans, in which studies rely almost exclusively on retrospective questionnaires. Perhaps even more important, it is possible to produce discrete brain lesions restricted to specific regions of interest. Close analysis of the literature reveals that the length of the retrograde amnesia gradient seems to be a function of several factors including the species, the complexity of the memory paradigm, the level of mastery of the task and the type, and the extent and location of the brain lesion.

Although lesion approaches have brought a great deal of information about the potential roles of the different regions of the medial temporal lobe in memory processing, they tell us nothing about the contribution of the hippocampus when this region was intact at time of retrieval. To circumvent this problem, functional brain imaging techniques similar to positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) in humans have been developed and applied to studying memory processes. One such technique is the (^{14}C)2-deoxyglucose autoradiographic method, which allows for regional measurements of levels of glucose consumption, and thereby neuronal activation, in the entire brain. This technique was applied to spatial discrimination learning in mice (**Figures 2(a) and 2(b)**). In accordance with the standard model of memory consolidation, lengthening the retention interval from 5 days (recent memory) to 25 days (remote memory) resulted in decreased metabolic activity in the hippocampus and a concomitant increase in several cortical areas (**Figure 2(c)**).

More recently, cellular imaging experiments have used the expression of activity-dependent genes such as c-fos and zif-268 as markers of brain activity. Using these approaches, it has been possible to track changes in the organization of spatial and contextual memories. Again, a converging and consistent pattern of neuronal reorganization emerged – the retrieval of recent memory engaged the hippocampus, whereas the retrieval of remote memories was associated with increased activity in cortical regions such as the prefrontal, anterior cingulate, and retrosplenial cortices. More refined reorganization of neuronal activity across cortical layers was also observed, for instance, in the parietal cortex. Recruitment of these different cortical regions may reflect their increasing roles in memory storage, effortful recall, performance monitoring, and use of consolidated remote memories. Consistent with these imaging data, pharmacological inactivation of different cortical sites specifically disrupts recall of remote, but not recent, memories.

In Humans

Temporal gradients of retrograde amnesia have been reported in humans, and these case studies are well documented. One striking difference from animal studies concerns the length of the gradient. Most often, amnesia extends back across a period of several years and, in some cases, for several decades. Both remote semantic memory for facts and remote episodic memory for autobiographical events appear to be spared even though some hippocampal patients have been diagnosed as having difficulties in recollecting detailed episodes from their early life, possibly because of extensive damage outside the medial temporal lobe.

Results from functional brain imaging studies have proven to be more controversial, in part because of the technical difficulties in recording activity across subjects accurately and in designing appropriate comparative baseline control tasks in order to isolate neuronal activity associated with nonspecific aspects of the testing procedure. A recent fMRI study by Takashima and colleagues, however, appears to be quite consistent with the imaging data obtained in animals. The authors were successful in studying remote memory in a prospective manner, just like the experimental designs classically used in animals. Healthy adult subjects were asked to memorize a series of photographs of various landscapes. Recognition of these photographs was then probed during a testing session that occurred either 1 day, 1 month, or 3 months later. The correspondence with the imaging results obtained in animals is striking. Not only was the recall of remotely acquired pictures independent of the hippocampus and dependent on the medial prefrontal cortex, but this dramatic reorganization occurred at a rapid pace that was of a magnitude comparable to that observed in animal's studies of retrograde amnesia.

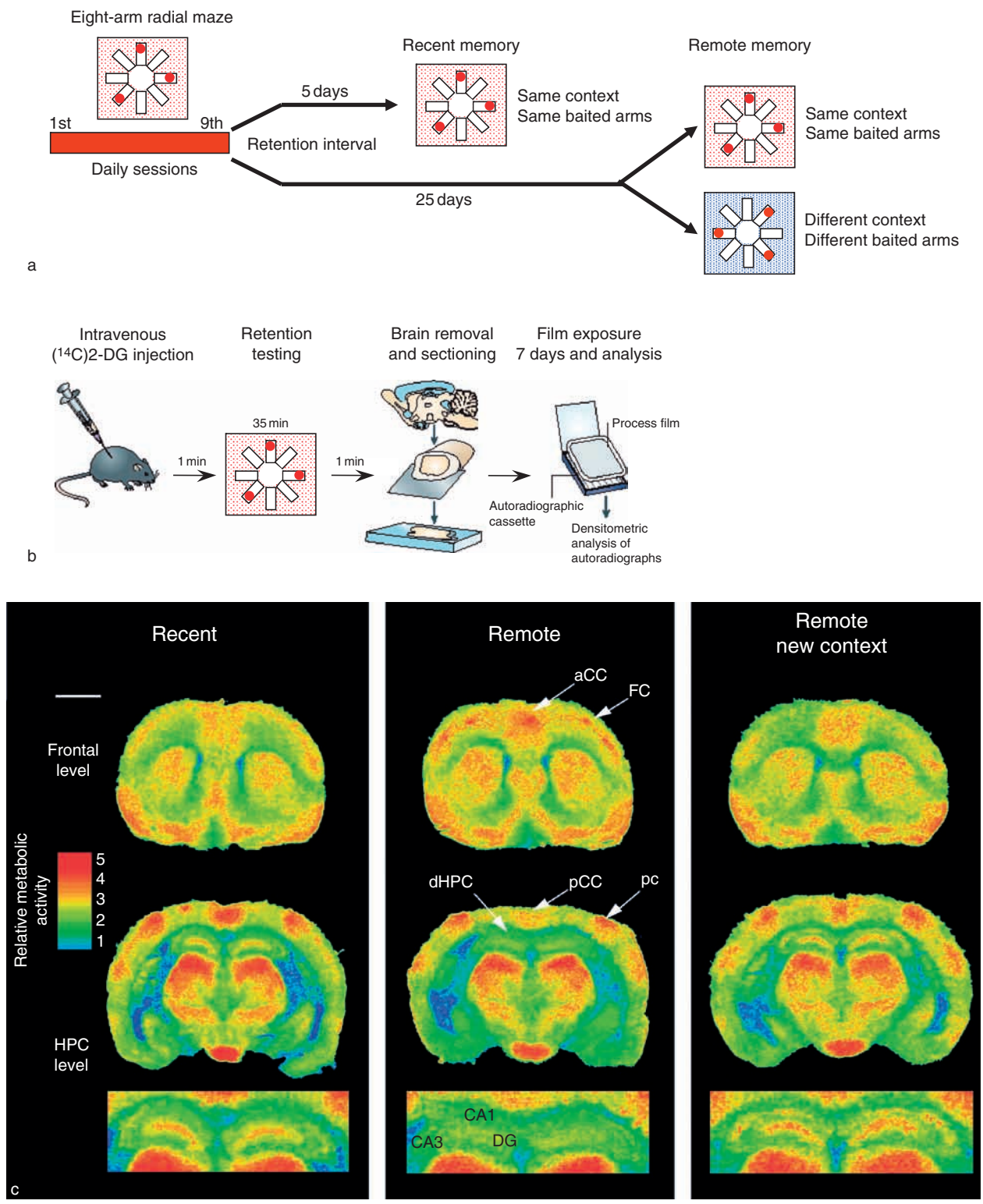

Figure 2 System consolidation as revealed by functional brain imaging in mice: (a) testing procedure; (b) imaging procedure; (c) color-coded autoradiographs of coronal sections of mice tested for recent memory (left), remote memory (center), or remote memory in the new context (right). As illustrated in (a), mice were first submitted to spatial discrimination testing in the eight-arm radial maze, wherein they were trained to locate and remember the spatial positions of three constantly baited arms. Following 9 days of training to acquisition criterion (80% correct responses), memory performance was then assessed either 5 days (recent test) or 25 days (remote test) later. An additional group of mice was tested 25 days later with a similar maze but located in a different room. As illustrated in (b), following intravenous injection of the (^{14}C)2-deoxyglucose (2DG) radioactive tracer, mice were submitted to recent or remote retention testing in the

Mechanisms Underlying System Consolidation

Over the last decade, multidisciplinary approaches in animals have contributed to the elucidation of many of the cellular and molecular mechanisms underlying memory processing in hippocampal networks during synaptic consolidation. Much less is known, however, about the dynamics of hippocampal–cortical interactions that allow these memories to be transformed into remote memories in cortical networks. Memory reactivation remains the core mechanism in most models of system consolidation. Repeated activation of hippocampal networks is thought to reinstate, in a coordinated and synchronized manner, activity in different cortical networks. Reactivation can be induced either during online states of conscious recollection (or quiet wakefulness) or during offline states such as sleep. Marr was among the first to suggest that offline replay of previous waking pattern of neural activity during sleep may drive consolidation and ensure a gradual stabilization of cortical memory traces. A role for sleep in memory consolidation is now well documented. Although exceptions exist, a gain in performance after overnight sleep has been reported. Conversely, sleep deprivation can interfere with, and even prevent, memory consolidation. At the molecular level, sleep has been associated with the upregulation of approximately 100 known genes (and ~400 unidentified ones) independent of circadian rhythm. It is reasonable to assume that at least some of these genes, either directly or indirectly via triggering of specific signaling pathways, mediate incremental changes in synaptic plasticity within cortical networks.

In the hippocampus, rapid encoding is likely to predominantly involve changes in weight between existing connections via long-term potentiation (LTP) and long-term depression (LTD) (or weight plasticity). In the cortex, however, where network reorganization occurs slowly, an additional form of plasticity may involve the creation of new synaptic connections between previously unconnected neurons. In the animal brain imaging studies already mentioned, it is this form of wiring plasticity that was observed. Long-term storage of spatial and contextual information was accompanied by the generation of novel synapses in specific areas of the prefrontal and anterior cingulated cortices, thus indicating that wiring plasticity most likely contributed to changing the organization and architecture of distributed cortical networks supporting remote memory storage (Figure 3). Such morphological changes appear to be crucial because α-calcium/calmodulin-dependent

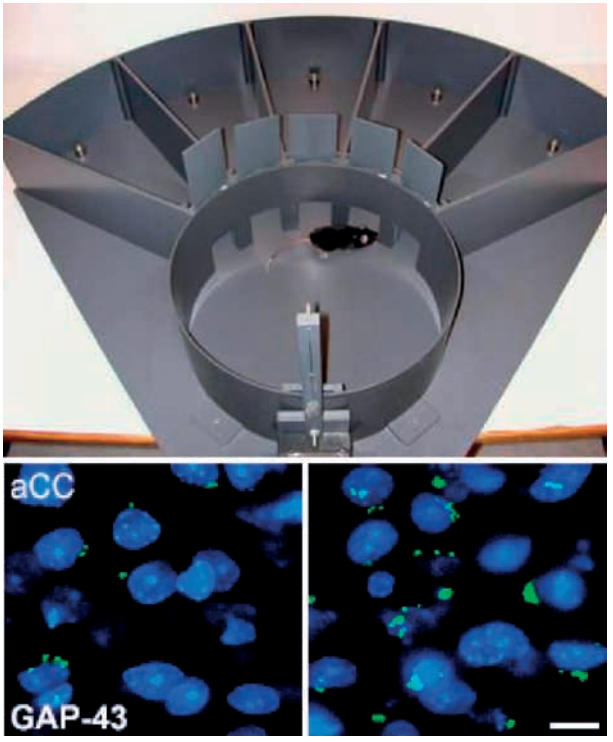

Day 1 Day 30

Figure 3 Remote memory storage in the cortex involves wiring plasticity. The expression of growth-associated protein 43 (GAP-43), a presynaptic protein that controls axon growth and sprouting and that is used as a marker of newly formed synapses, was examined in the anterior cingulate cortex (aCC) of animals tested either 1 day (recent memory) or 30 days (remote memory) following spatial discrimination in a five-arm maze (top image). Cortical consolidation of the spatial position of the baited arm of the maze was accompanied by an increase in GAP-43 staining on day 30 as compared with day 1 (bottom images). Nuclei appear in blue, GAP-43 fluorescent staining is green. This finding supports that cortical networks undergo structural changes as memories mature over time. Scale bar = 10 μm.

eight-arm radial maze. Their brains were rapidly removed on completion of testing and prepared for autoradiography, which allowed for quantitative measurement of testing-induced changes in neuronal activation in various brain regions. In (c), the lower section of each panel shows a magnified view of the dorsal hippocampus (dHPC), which includes the CA1, CA3, and dentate gyrus (DG) regions. Increasing the retention interval from 5 to 25 days resulted in a significant decrease of metabolic activity in the hippocampus and a concomitant increase of activity in several cortical regions such as the frontal (FC) and anterior cingulate (aCC) cortices. These data point to a time-limited role of the hippocampus in remote memory storage. Over time, as memories mature, distributed cortical areas become capable of mediating the retrieval of remote memories independently. However, when mice are confronted with a different context at the remote time, the hippocampus is reengaged, indicating that this brain region is required to encode novel information. HPC, hippocampus; PC, parietal cortex; pCC, posterior cingulate cortex. From Frankland PW and Bontempi B (2005) The organization of recent and remote memories. *Nature Reviews Neuroscience* 6: 119–130.

protein kinase II mutant mice exhibiting normal hippocampal plasticity (but lacking cortical plasticity) are unable to form enduring memories. The hippocampus has been conceptualized as a fast learner and the cortex as a slow learner. Differences in the relative prevalence of weight versus wiring plasticity may underlie this apparent division of labor. Slow incremental changes in cortical networks may help prevent catastrophic interference, as emphasized in current connectionist models; protect existing cortical memories from being erased; and ensure constant updating of these cortical memories.

One question that often comes to mind is how the hippocampus manages to reactivate the right network of cortical neurons related to a specific experience. Because hippocampal–cortical projections are topographically organized, several researchers have suggested that the hippocampus might store some sort of indices of all the relevant neocortical synapses that need to be strengthened as memories mature. Although still speculative at present, identification of cortical synapses could be achieved via the setting of synaptic tags at all relevant synapses that were activated at the time of initial encoding in the primary sensory as well as associative cortices.

A Crucial Role for the Prefrontal Cortex in Processing Remote Memory

There is new evidence from both animal and human studies in favor of a crucial role for the prefrontal cortex in remote memory recall. Classically, the prefrontal lobe has been implicated in mediating effortful monitoring and response selection during retrieval. Forgetting, which may reflect a partially degraded memory trace, is inevitably associated with long-term memory storage, and we are all familiar with the increased difficulty in retrieving information acquired long ago compared to recently acquired information. The prefrontal lobe may also have additional roles in retrieval processes. The prefrontal cortex consists of several highly interconnected regions, including the prelimbic, inframimbic, and anterior cingulate areas. These regions are reciprocally connected to the sensory, motor, and limbic cortices and are therefore ideally suited to integrate complex information from different sensory sources. This potential for integration has led to the hypothesis that the ability of the prefrontal cortex to process remote memories might mirror that of the hippocampus to process recent information. At the time of encoding, the hippocampus is thought to integrate information from various independent cortical modules that represent the various features of an experience and then to rapidly fuse these features into a coherent memory trace. As memories become gradually consolidated in the cortex, this integrative function may be transferred to the prefrontal cortex, which then become capable of integrating information embedded within other distributed cortical networks (Figure 4). More challenging is the recent view that the prefrontal cortex might modulate hippocampal activity during memory storage. Intriguing functional brain

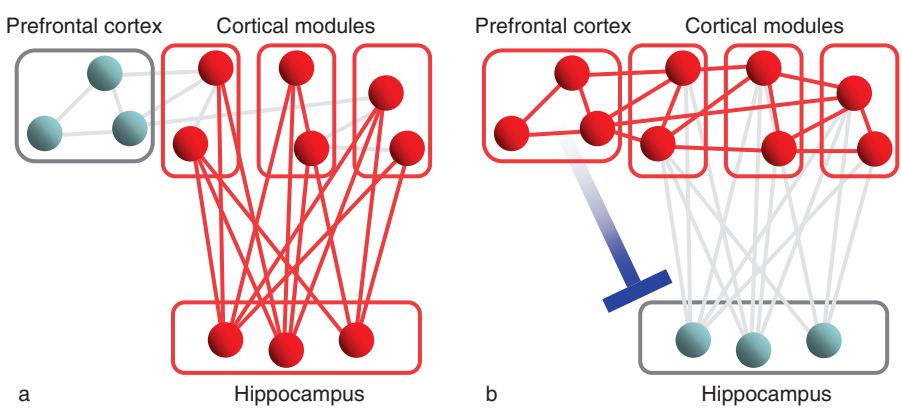

Figure 4 A privileged role of the prefrontal cortex (PFC) in systemic consolidation: (a) recent memory; (b) remote memory. The PFC is thought to have multiple roles in remote memory storage and retrieval. At the time of encoding, memories are encoded in hippocampal–cortical networks. The hippocampus is crucial in integrating information from distributed cortical modules, each representing individual components of a memory. As memories mature over time, cortical–cortical connections are strengthened, allowing cortical regions, and particularly the PFC, to ensure retrieval of remote memories independently and to assume the integrative role initially exerted by the hippocampus. Consistent with this, the inactivation of the PFC blocks the expression of remote memory, but spares recent memory, in rodent tests of contextual and spatial memory. The prefrontal cortex might also modulate hippocampal activity during remote memory recall. The hippocampus is normally active when processing the outside environment. However, when incoming information matches previously stored cortical information, the PFC inhibits hippocampal activity by direct or indirect pathways to prevent the encoding of redundant information. In the presence of a mismatch situation, no inhibition is exerted and the hippocampus is engaged as usual. From Frankland PW and Bontempi B (2005) The organization of recent and remote memories. *Nature Reviews Neuroscience* 6: 119–130.

imaging results, indeed, point to the inhibition of hippocampal activity when remote memories are successfully recalled, suggesting that the prefrontal cortex might exert a top-down inhibitory control over hippocampal activity and prevent the consolidation of existing (i.e., already corticalized, and thereby redundant) information (Figure 4). This view is consistent with the existence of match and mismatch comparator neurons in the cortex, especially in the prefrontal cortex.

Concluding Remarks

Since the first clinical observations of Ribot more than a century ago, we have made considerable progress in unraveling the neurobiological bases of memory consolidation. The consolidation process is no longer seen as a simple passive phenomenon but as a dynamic process of neuronal reorganization that affects extensive hippocampal and cortical networks over timescales much longer than initially envisioned. Functional brain imaging techniques coupled with multiple cell recording have allowed new breakthroughs in the identification of the different brains regions and neuronal network supporting recent and remote memory storage. Yet, despite dramatic advances in the field of functional genomics and proteomics, we are still far from understanding all the precise cellular and molecular mechanisms underlying memory reorganization within cortical networks.

See also: Amnesia: Declarative and Nondeclarative Memory; Animal Models of Amnesia; Cortical Plasticity and Learning: Mechanisms and Models; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Functional Amnesia; Hippocampus and Neural Representations; Long-Term Depression: Cerebellum; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Memory Representation; Memory Consolidation: Systems; Perirhinal Cortex: Neural Representations; Sleep-Dependent Memory Processing; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Chklovskii DB, Mel BW, and Svoboda K (2004) Cortical rewiring and information storage. *Nature* 431: 782–788.

Dudai Y (2004) The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology* 55: 51–86.

Eichenbaum H (2002) *The Cognitive Neuroscience of Memory: An Introduction.* New York: Oxford University Press.

Eichenbaum H (2004) Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron* 44: 109–120.

Frey U and Morris RG (1998) Synaptic tagging: Implications for late maintenance of hippocampal long-term potentiation. *Trends in Neuroscience* 21: 181–188.

Frankland PW and Bontempi B (2005) The organization of recent and remote memories. *Nature Reviews Neuroscience* 6: 119–130.

Kandel ER (2004) The molecular biology of memory storage: A dialog between genes and synapses. *Bioscience. Reports.* 24: 475–522.

Marr D (1971) Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of* London *B; Biological Science* 262: 23–81.

McClelland JL, McNaughton BL, and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex?: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102: 419–457.

McGaugh JL (2000) Memory – a century of consolidation. *Science* 287: 248–251.

Moscovitch M, Rosenbaum RS, Gilboa A, et al. (2006) Functional neuroanatomy of remote episodic, semantic and spatial memory: A unified account based on multiple trace theory. *Journal of Anatomy* 207: 35–66.

Müller GE and Pilzecker A (1900) Experimentelle Beiträge zur Lehre von Gedächtnis [Experimental contributions to the theory of memory]. *Zeitschrift für Psychologie* 1: 1–300. [Summarized by McDougall W (1901) in *Mind* 10: 388–394].

Ribeiro S and Nicolelis MAL (2004) Reverberation, storage, and postsynaptic propagation of memories during sleep. *Learning & Memory* 11: 686–696.

Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry* 20: 11–21.

Squire LR and Alvarez P (1995) Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinion in Neurobiology* 5: 169–177.

Squire LR and Kandel ER (2000) *Memory: From Mind to Molecules.* New York: W. H. Freeman.

Squire LR, Stark CEL, and Clark RE (2004) The medial temporal lobe. *Annual Review of Neuroscience* 27: 279–306.

Takashima A, Petersson KM, Rutters F, et al. (2006) Declarative memory consolidation in humans: A prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America* 103: 756–761.

Teng E and Squire LR (1999) Memory for places learned long ago is intact after hippocampal damage. *Nature* 400: 675–677.

Memory Consolidation: Systems

K A Paller, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

The Consolidation of a Memory

At the moment when we perceive an event, the journey of memory storage begins. In a sense, such an event may exist for many years before finally being remembered – where is it, what is it, and how does it manage to persist?

Whereas memory is revealed by an organism's behavior at the time of retrieval, neural substrates of memory storage assume some persistent form during the entire retention interval. The ability to remember an event is made possible by changes in neurons and in networks of neurons – neural plasticity. Yet, these changes are quite unlike when a liquid is poured into a mold and solidifies as plastic or when clay is shaped and fired to form pottery. A memory trace is not a static entity. It does not remain in a fixed state, like a file in a file drawer or on a computer. Rather, further neural changes can transpire during the time period between acquisition and retrieval as a memory is associated with other stored information and potentially becomes more stable.

What is the fate of the neural plasticity first set into motion when an event is experienced? An understanding of the concept of memory consolidation requires understanding the form of memory traces (engrams) both in neural and cognitive terms, as well as specifying how memory traces change over time.

Memory Distinctions

Memory storage and memory retrieval mechanisms are known to differ in fundamental ways according to the type of memory in question. Likewise, memory consolidation proceeds differently for different types of memory.

A long-standing distinction in research on human memory allows for separate mechanisms for short-term storage lasting seconds, perhaps minutes, and long-term storage lasting much longer. Donald Hebb's seminal proposal stated that reverberating neural circuits can keep an experience in mind for a short time, whereas structural changes in neuronal ensembles underlie storage that persists over longer periods. Mental representations of perceptual and conceptual information can be actively maintained through rehearsal. In this sense, short-term storage cannot be defined by a characteristic time span. Rehearsal can continue for an indefinite amount of time. The term immediate memory aptly draws attention to this aspect of memory and is akin to what William James termed primary memory. In contrast, secondary memory refers to information brought back to mind after earlier leaving one's awareness, rather than information kept in mind through rehearsal. The length of the retention interval, short term versus long term, is thus not a suitable way by itself to differentiate between types of memory. Immediate memory (also known as working memory) belongs in a class by itself due to its dependence on the rehearsal of information.

Aside from immediate memory, there are several distinct ways in which behavior comes under the influence of past experience. Of central relevance for major territories of human cognition, declarative memory pertains to memory for complex facts and personally experienced events, and can be distinguished from a diverse set of memory phenomena collectively known as nondeclarative memory. Nondeclarative memory includes skill learning, habit learning, simple forms of conditioning, various types of priming that can be measured in implicit memory tests, and nonassociative forms of learning like habituation and sensitization.

Some of the molecular building blocks of memory storage are common across these different types of memory. Elaborate layers of molecular machinery can require some passage of time so as to complete the process of memory formation. In many cases, neural plasticity induced during learning can be modulated by subsequent events. Release of epinephrine and cortisol, for example, can allow emotional significance to regulate memory strength. Selective gene activation and protein synthesis also play key roles in stabilizing synaptic changes.

These aspects of memory fall in a category called synaptic consolidation. Relevant molecular changes can require minutes to hours. On the other hand, the present discussion focuses on neural mechanisms of systems-level consolidation that are specific to declarative memory and that tend to occur over a much longer timescale.

Consolidating Declarative Memories

Declarative memory concerns the ability to bring back to mind factual and episodic information. Each of us maintains a colossal but changing record of facts and events to access when needed. This information can potentially stay with us for so many years that one might say that declarative memories can last a lifetime. One general principle of memory storage is that cortical networks specialized for processing specific types of information are the very same

networks involved in storing that information as fragments of declarative memories. Those fragments must be linked together in order for any particular declarative memory to exist.

Although we have learned much about the brain mechanisms responsible for declarative memory, and about ways to assess this type of memory using recall and recognition tests, pinpointing the exact neural substrates of any one specific autobiographical event or fact is beyond current technology. Indeed, a declarative memory does not reside in a single location, but rather, it depends on a dynamic network of neurons. Lively controversy surrounds questions about how consolidation operates on such a network, the precise length of time consolidation requires, and whether these memories assume a labile form under special circumstances.

Empirical evidence that can be brought to bear on theories of consolidation comes from multiple sources. Neuropsychological studies of patients with amnesia, in particular, suggest that declarative memories can change to become resistant to disruption and that this change is a by-product of consolidation. Critical brain damage in amnesia tends to include midline diencephalic or medial temporal structures, including the hippocampus and adjacent cortical regions, though the distinct memory functions presumably mediated by these different regions are currently unclear. Studies of animal models have also been designed to identify the neuroanatomical basis of this critical contribution to memory.

Evidence from Human Memory Disorders

Memory impairment tends to comprise both anterograde amnesia, a pronounced difficulty in learning new information, and retrograde amnesia, a disruption of the ability to remember information that had been learned prior to the onset of symptoms. In either case, patients experience difficulty remembering factual as well as event information (though dissociations are possible, at least in retrograde amnesia). Some amnesic patients exhibit severe anterograde amnesia and severe retrograde amnesia; some exhibit mild anterograde amnesia and mild retrograde amnesia. Given that neurological insult can produce correlated anterograde and retrograde amnesia in this manner (even though there are important exceptions), standard accounts specified a single mechanism that led to both effects. In particular, a core defect in a protracted consolidation process required for declarative memory storage could result in both types of impairment; new learning would not be normal without consolidation, and old memories would not be retrieved normally if their consolidation had not run its course.

In analyses of retrograde amnesia, it is useful to take into account the age of memories that might be retrieved (i.e., the length of the retention interval) by distinguishing between recent memory and remote memory. Indeed, a prevalent characteristic of retrograde memory impairment, described in 1882 by Theodore Ribot and known as Ribot's law, states that the dissolution of memory is inversely related to the recency of the event. Numerous cases of retrograde amnesia have been described in which the impairment was taken to be limited to recent memories, with preservation of remote memories. Patterns of temporally graded retrograde amnesia have been demonstrated using a variety of tests (**Figure 1**).

Implications of Temporally Graded Retrograde Amnesia

The notion that consolidation progresses over an extended period of time is consonant with evidence from retrograde amnesia, given the following scenario. To the extent that memories were formed recently, consolidation processes would have as yet been insufficient to produce a stabilized cortical memory. These memories become inaccessible in the absence of retrieval mechanisms that depend on hippocampus and related structures. This standard neuroanatomical account of consolidation thus characterizes recent memories as those that depend on both cortical and hippocampal networks.

In contrast, to the extent that an old memory has been subject to sufficient consolidation, it can be retrieved via cortical retrieval mechanisms. Whether hippocampal participation remains critical for some remote memories in some patients, however, is currently controversial.

This standard conceptualization of consolidation is as a process whereby declarative memories outgrow their dependence on corticohippocampal networks to become cortically self-reliant. This metamorphosis may involve increased connectivity among the components of the memory in the cortex; accordingly, the process has sometimes been referred to as cross-cortical consolidation. Cross-cortical consolidation may also involve the formation of new cortical representations that function to represent the gist or higher-order meaning of the memory while simultaneously enhancing the coherence of the set of neocortical storage sites (**Figure 2**).

Testing such speculations about how declarative memory representations change as a result of consolidation will require new sources of neurophysiological and neuroanatomical evidence. Before delving into this sort of evidence, we must examine doubts that have been raised about the evidence from retrograde amnesia.

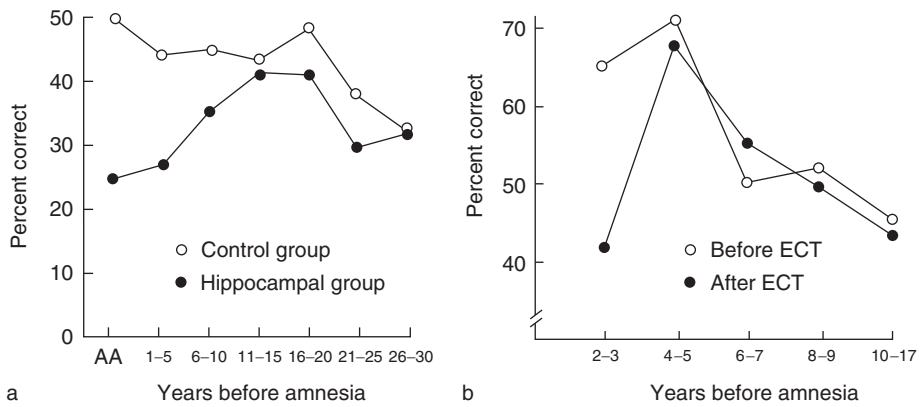

Figure 1 Forgetting curves showing temporally graded retrograde amnesia. (a) Recall was tested using questions about news events that occurred in preceding years (1950–2002) in five patients with damage limited primarily to the hippocampal region and 2–3 control subjects matched to each amnesic patient. The time period labeled AA (anterograde amnesia) represents recall deficits for information acquired after the onset of amnesia. (b) Recognition was tested for TV shows that were broadcast for only one season in preceding years (1957–72) in 16 patients who were given a course of electroconvulsive therapy (ECT) as treatment for depressive illness. Subjects were required to select actual TV show titles presented among fabricated titles, and the guessing rate was 25%. This test is advantageous because TV show titles were learned in a specific year and were unlikely to be rehearsed repeatedly afterwards. ECT was given three times per week and testing occurred before the first treatment and 1 h after the fifth treatment. These memory deficits did not result from medications or depressive illness *per se*. Memory abilities improved within a few weeks after ECT was completed. Adapted from Manns JR, Hopkins RO, and Squire LR (2003) Semantic memory and the human hippocampus. *Neuron* 38: 127–133; and from Squire LR, Slater PC, and Chace PM (1975) Retrograde amnesia: Temporal gradient in very longterm memory following electroconvulsive therapy. *Science* 187: 77–79.

Controversies and Challenges in Studies of Retrograde Amnesia

The extant literature on retrograde amnesia presents a complicated story that does not uniformly support this standard account of consolidation. Temporally graded retrograde amnesia has been observed reliably in patients with closed head injury, Korakoff's syndrome, and other disorders of memory. Observations by Scoville and Milner in 1957 of patient HM, who became amnesic following bilateral medial temporal lobectomy for medically intractable epilepsy, showed retrograde amnesia for autobiographical memory of roughly 2 years in duration. However, an insidious disease onset, as is likely in Korsakoff's syndrome, or factors like epilepsy or certain medications, could lead to apparent retrograde deficits that actually resulted from poor encoding during those years.

Conventional methods for assessing remote memory also have limitations. Whereas tests of anterograde amnesia concern information delivered systematically to the patient under well-controlled conditions, tests of retrograde amnesia tend to probe information acquired under unknown circumstances. When testing memory for historical events or celebrities, for example, it can be difficult to determine the patient's premorbid knowledge. To attempt to overcome this problem, experimenters can learn about important life events from informants or create specialized tests for specific information.

In constructing retrograde tests, it can be helpful to select test items such that various factors are equivalent across time periods. In this way, for example, difficulty might be equated such that an observed forgetting curve would reflect true forgetting rather than greater difficulty for older items. However, equating difficulty may lead to other differences across items, as the most remote items may tend to concern factual rather than episodic knowledge. Interpretations of retrograde gradients thus require careful attention to the makeup of test materials.

Another interpretive limitation results because most people show superior autobiographical retrieval for events from their early adulthood. One explanation is that these years often include people's most momentous experiences and so have a powerful influence on their life story and personality. Events from these times may also be extensively intertwined with other relevant information such that they are not easily lost.

Importantly, many reports of retrograde amnesia have demonstrated impairments not limited to a few years, but spanning one or more decades. Some patients lose essentially their whole life, in that they are unable to retrieve many autobiographical memories from any past time period. In contradistinction to the temporally graded period of retrograde amnesia shown in Figure 1, these patients show no gradient. A process of gradual consolidation does not

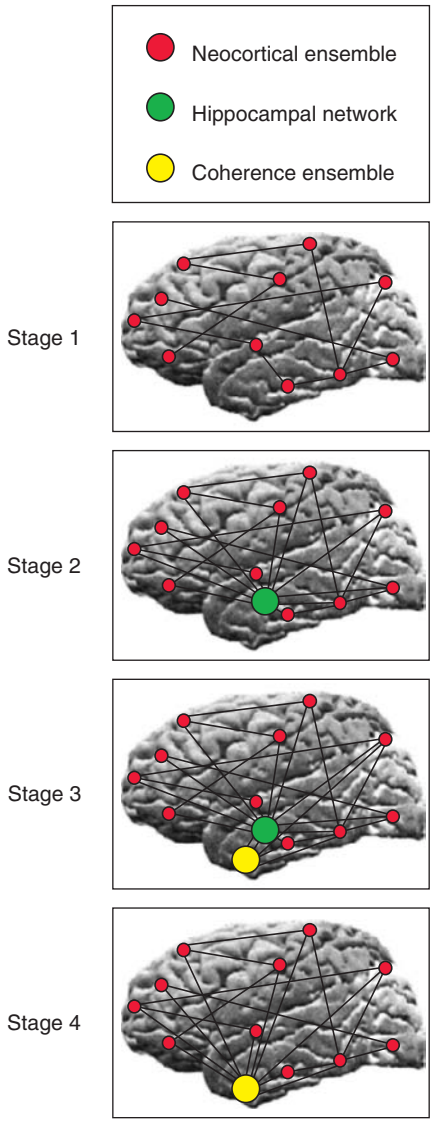

Figure 2 A speculative account of cross-cortical consolidation represented schematically. Initially (stage 1), a fact or event is encoded via representations in multiple cortical regions with binding and cognitive control provided by prefrontal cortex. A selection of neocortical ensembles is figuratively shown in red (allied subcortical networks are omitted). At the same time, high-level features in this dispersed representation involve hippocampal networks (stage 2). These hippocampal networks trigger pattern completion such that the set of dispersed cortical fragments can later be activated together. Over time (stage 3) and with the aid of hippocampal networks, coherence ensembles begin to form in various cortical regions such as medial temporal cortex or other limbic-associated cortex (e.g., entorhinal, perirhinal, parahippocampal, posterior cingulate, and medial prefrontal cortex). These coherence ensembles ultimately become part of the memory representation, taking on a superordinate role. Network connections become strengthened when the memory is retrieved and/or related to other stored information. By this account, a coherence ensemble is a cortical network that functions to maintain cohesiveness among the various parts of a declarative memory, and that ultimately (stage 4) takes over the binding role of the hippocampus. Adapted from Paller KA (2001) Neurocognitive foundations of human memory. In: Medin DL (ed.) *The Psychology of Learning and Motivation*, vol. 40, pp. 121–145. San Diego, CA: Academic Press.

seem to supply a viable explanation for such extensive memory loss. Moreover, dense retrograde impairments have been found in some patients with minimal anterograde amnesia, although this phenomenon is controversial and may be explained in some cases by other factors, such as psychogenic amnesia. Another complication is that retrograde impairments can disproportionately affect episodic versus semantic knowledge. Some patients display poor memory primarily for remote autobiographical information, some primarily for remote factual information.

One theoretical view developed to account for extensive episodic memory loss is known as multiple trace theory. By this view, the hippocampus contributes to episodic memory retrieval for all time periods. This contribution provides episodic retrieval with the contextual detail that supports the ability to approximate re-living a past event. Memories for factual information do not include this contextual detail and become hippocampal-independent after sufficient consolidation. The idea that episodic memories do not become independent of the hippocampus provides an explanation for retrograde amnesia for episodes across all time periods.

Interpretations of these intriguing patterns of retrograde impairment, however, must take into account the fact that retrograde amnesia can arise from several fundamentally different causes. Although consolidation provides a viable explanation for some instances of retrograde amnesia, memories can also be lost when pathology disrupts cortical areas responsible for the storage of this information. Whereas declarative memories are dispersed across multiple cortical regions, some of the multimodal and multidimensional knowledge critical for retrieving certain memories may depend on the integrity of special convergence zones in the brain, perhaps including medial temporal cortex (i.e., entorhinal, perirhinal, and parahippocampal), orbitofrontal cortex, medial prefrontal cortex, parietal cortex, and posterior cingulate cortex. As described above, a new representation formed in one of these regions may become part of the declarative engram by virtue of the consolidation process. Further research is needed to delineate the type of information stored in such regions. Conceivably, decades of retrograde amnesia can arise when storage sites in these critical regions are disrupted, in which case the evidence would have no bearing on specifying the time span of consolidation. Brain damage to some of these regions could conceivably result in extensive retrograde amnesia with minimal anterograde amnesia if the hippocampus and some adjacent medial temporal cortex remained intact.

Another explanation for dense retrograde amnesia must also be considered. Cross-cortical consolidation may not be the culprit when retrograde amnesia

across all time periods results from a profound defect in retrieval. Indeed, various sorts of evidence now point to a critical role of the frontal lobes in orchestrating retrieval. Full-blown memories of distinct autobiographical episodes can require extensive memory search. The planning, organization, and evaluation processes required for retrieval can become extremely challenging after frontal brain damage. Also, dissociations in the type of information that can and cannot be retrieved could arise from deficits based on either corrupted storage sites or disabled retrieval.

The interpretation of long periods of remote memory loss must also be informed by evidence for systematic relationships between the length of the retrograde loss and sites of brain damage. More evidence of this sort is needed, but results from postmortem histological analysis for several cases suggest that partial hippocampal damage produces limited retrograde impairment (1–2 years), whereas more extensive damage to the hippocampus and entorhinal cortex produces broader retrograde impairment (15–25 years).

In sum, the literature includes an inconsistent mix of reports of retrograde amnesia that either extends over decades or is limited to a recent period not longer than a few years. In cases of temporally extensive retrograde amnesia, storage and retrieval accounts must be carefully evaluated. Perhaps all old memories in healthy people are subject to steady reorganization. If so, an interesting but untested possibility is that the longer a patient has amnesia, the more dense their retrograde amnesia will be. At any rate, further research will hopefully expand our understanding of the range of retrograde deficits that can be observed and clarify when such memory disorders arise from a consolidation defect.

Problems with Specifying the Age of a Memory

A further theoretical concern haunts the evaluation of all data on remote memory in amnesic patients. How should we conceptualize an episodic memory that is 20 or more years old?

An episodic memory is defined with reference to one specific event that was experienced in a unique spatiotemporal context. If a person is capable of remembering an event from 20 years ago, it can reasonably be assumed that this memory was also retrieved multiple times in the intervening years. Nevertheless, researchers commonly refer to this as a 20-year-old memory. If intervening retrieval influences memory storage, it might be more appropriate to regard the effective retention interval as much shorter than 20 years. Although an individual may recall an event from 20 years ago, the memory itself could be largely based on modifications made more recently. Indeed, past episodes not retrieved and elaborated upon multiple times over many years may be the ones we tend to forget.

Even younger memories are subject to concerns about effects of intervening retrieval. Once we acknowledge that memories are accessed in sleep, which now seems undeniable, we cannot rule out the idea that every instance of remote memory retrieval is a function of storage events at multiple time points.

By the present account of consolidation, each time a memory for an episode is retrieved, there is a possibility that the information in question is associated with other information so as to expand the operative nature and meaning of the memory. Retrieved memories are regularly reevaluated and related to new information in mind at the moment of retrieval, including the context of the new retrieval episode. Moreover, new events can provoke the reinterpretation of past events. Thus, the memory trace should be conceived of as product of memory storage operations engaged during various retrieval experiences. Unless intervening retrieval can be ruled out, it is problematic for the age of an episodic memory to be defined simply on the basis of when the original event occurred.

This reconceptualization of remote memories provides an important point of contact between multiple trace theory and the standard model of consolidation. A recalled episode may be tantamount to a retelling of prior re-tellings of the same story, rather than a replay of an ancient story set in stone long ago.

There are thus two reasons an amnesic patient's retrograde impairment cannot be specified by a simple duration. First, consolidation for various factual and episodic memories may proceed at a different rate depending on the type of information and the ways in which it is retrieved and associated with other information. Second, as underscored by the preceding discussion, specifying a single operative age for a given memory is problematic.

Memory Transformation due to Consolidation

Connections to the cortical storage sites that comprise a declarative memory may be formed via the hippocampus at initial encoding, and then used when that memory is subsequently retrieved. In this way, memories may be subject to continual hippocampal-dependent restructuring for some time. Some remote memories may ultimately reach a state in which any further changes are relatively minimal, even though these memories may retain their episodic quality.

Whereas consolidation can produce a hippocampal-independent memory, it can also change the representational quality of a memory. In some cases of

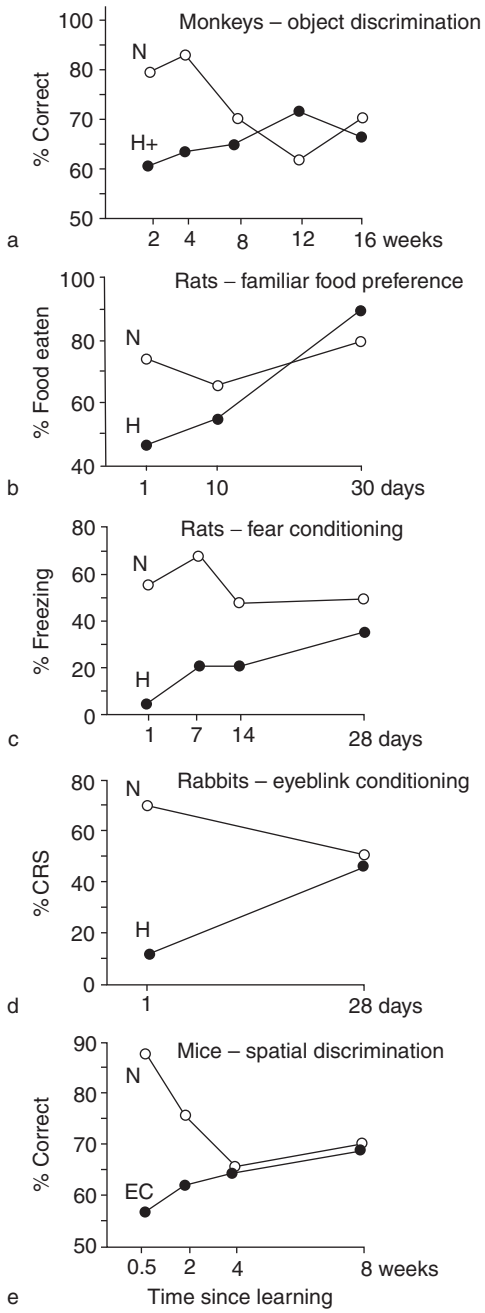

Figure 3 Several demonstrations of temporally graded retrograde amnesia following damage to the hippocampal region. Learning took place at multiple times before surgery. Two groups of animals were tested in each experiment, normal animals (N) and animals with either lesions of the hippocampus (H), lesions of hippocampus plus surrounding cortex (H+), or lesions of entorhinal cortex (EC). Adapted from Eichenbaum H (2002) *The Cognitive Neuroscience of Memory: An Introduction*. New York: Oxford University Press. (a) Retention of 100 object discrimination problems in monkeys. Data from Zola-Morgan SM and Squire LR (1990) The primate hippocampal formation: Evidence for a time-limited role in memory storage. *Science* 250: 288–290. (b) Retention of the social transmission of food preference in rats. Data from Clark RE, Broadbent NJ, Zola SM, and Squire LR (2002) Anterograde amnesia and temporally graded retrograde amnesia for a nonspatial memory task after lesions of hippocampus and subiculum. *Journal of Neuroscience* 22: 4663–4669. (c) Contextual fear conditioning in rats. Data from Kim JJ and Fanselow MS (1992) Modality-specific retrograde amnesia of fear. *Science* 256: 675–677. (d) Trace eyeblink conditioning in rabbits. CRS, chronic restraint stress. Data from Kim JJ, Clark RE, and Thompson RF (1995) Hippocampectomy impairs the memory of recently, but not remotely, acquired trace eyeblink conditioned responses. *Behavioral Neuroscience* 109: 195–203. (e) Retention of maze problems in mice. Data from Cho YH, Beracochea D, and Jaffard R (1993) Extended temporal gradient for the retrograde and anterograde amnesia produced by ibotenate entorhinal cortex lesions in mice. *Journal of Neuroscience* 13: 1759–1766.

repeated retrieval of an event memory, contextual details are lost so that the event itself comes to be represented only abstractly. That is, specific episodic information can become extraneous through a process of decontextualization. Although these memories start out as contextually rich episodic memories, vivid perceptual and contextual details gradually become unavailable – factual information about the event becomes the only part of the memory that can be recalled. In this manner, an episodic memory can be said to have changed into a semantic memory. Decontextualization may produce an enduring semantic memory in conjunction with a shift in memory storage from neocortical/hippocampal to only neocortical.

Computational models of consolidation have implemented procedures that theoretically could lead to the formation of decontextualized memories. In the Complementary Learning Systems framework, for example, hippocampal networks implement fast links among neocortical components. Memory representation in the neocortex, in contrast, changes more slowly. This scheme keeps neocortical memories safe from interference that would result if networks changed based on each new piece of episodic information. Instead, slow neocortical changes can lead gradually to the buildup of semantic knowledge. The time course of consolidation can then be conceived as a reflection of the time-consuming training of cortical networks by hippocampal networks.

Evidence from Prospective Studies, Neuroimaging, and Sleep Research

Müller and Pilzecker introduced the term consolidation (*Konsolidierung*) in 1900 and proposed that memories could change over time from a fragile, labile form to a fixed form. In 1903, Burnham took up consolidation in the context of the retrograde difficulties experienced by amnesic patients and combined the notion of retrieval and association with the idea of neural reorganization. However, we have only limited opportunities to observe these consolidation-based changes directly. Such opportunities will increase as we (1) begin to understand the physiological changes inscribed in brain networks responsible for

declarative memories and (2) develop ways to overcome the methodological and interpretative limitations discussed above. Prospective studies offer the opportunity to analyze well-controlled memory storage over a certain time period. Instead of relying on memory tests for retrospective information acquired under unknown circumstances, the circumstances of encoding can be controlled such that a range of retention intervals can be systematically examined.

In animal studies of consolidation, a variety of approaches have been used in prospective studies, with lesions created surgically to encompass specific regions (Figure 3). Many of these studies have obtained results supporting the hypothesis that medial temporal damage produces temporally graded retrograde amnesia (although there are also some exceptions). These observations of retrograde amnesia in multiple species and with multiple types of testing procedure also attest to the feasibility of future efforts to delineate the precise causes of retrograde memory impairment, and in this way clarify how these memories are consolidated.

Human neuroimaging has also been used in studies designed to examine how hippocampal contributions to retrieval may change as a function of the time since initial encoding. Interpretations must contend with many of the limitations discussed above. Hippocampal activity for recent and remote memories, in particular, can be engaged due to retrieval processing *per se* and/or due to encoding of the new recollective experiences prompted by the retrieval cues. The possibility of superimposed encoding activity can thus pose difficulties for conclusive interpretations of the role of the hippocampus in recent versus remote memory retrieval in these studies.

Prospective methods have also been used in some neuroimaging studies. In one such study, subjects viewed a series of 320 landscape photographs and were tested up to 90 days later. Brain activations for recognized pictures were examined at four retention intervals (Figure 4). Consolidation apparently led to a change in the anatomical basis of memory storage, in line with predictions based on temporally graded memory deficits due to hippocampal damage. The hippocampal contribution appeared to decrease with time while an anterior medial prefrontal contribution increased with time. Evidence from animal studies

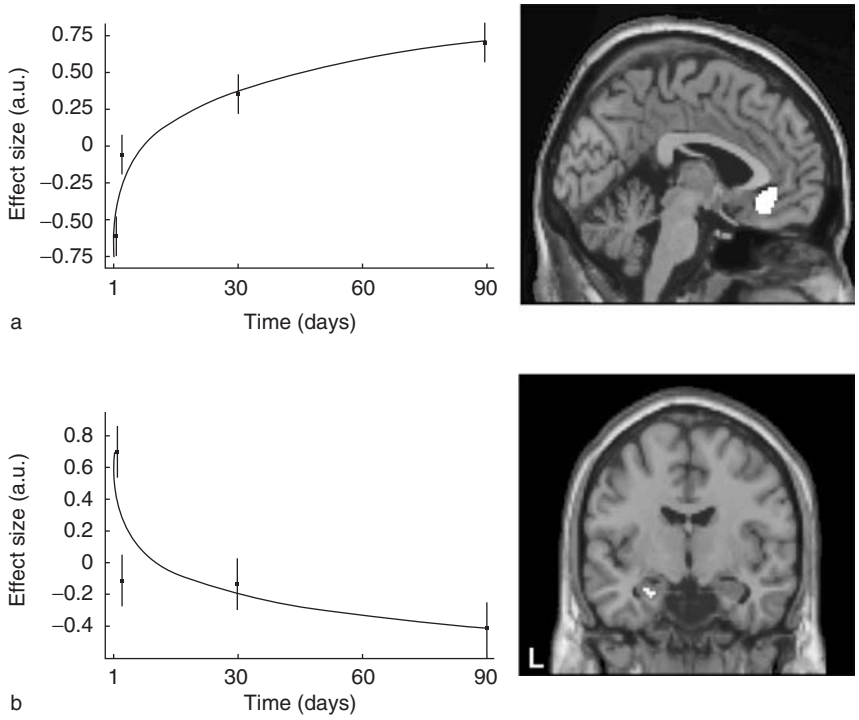

Figure 4 Results from an fMRI experiment in which brain activations for confident recognition responses were compared as a function of whether initial encoding took place 1, 2, 30, or 90 days earlier. (a) A region of ventral prefrontal cortex showed increased activity with increasing study–test delay. (b) Bilateral hippocampal regions showed decreased activity with increasing study–test delay. These findings suggest that hippocampal networks play a temporary role in memory storage, and that this role may be transferred to cortical regions such as medial prefrontal cortex. This experiment also included a rest period about 3 h after encoding, during which polysomnographic recordings were made, and those individuals who exhibited a greater amount of slow-wave sleep during this period tended to demonstrate superior memory for the photographs on later tests. Reproduced from Takashima A, Petersson KM, Rutters F, et al. (2006) Declarative memory consolidation in humans: A prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America* 103: 756–761. Copyright (2006) National Academy of Sciences, USA.

also supports the idea that this region of prefrontal cortex, which receives information from many other cortical regions, may play a key role in memory reorganization.

Evidence from multiple sources supports the idea that consolidation can occur both during waking and during sleep. Hippocampal and cortical neurons in animals show evidence of memory replay during postacquisition waking and sleep. Plasticity-related gene expression in medial temporal regions is altered systematically during sleep. Hippocampal theta during sleep may also index memory processing, perhaps involving temporal coupling across brain regions. Sleep spindles in cortical electroencephalogram (EEG) recordings at 12–15 Hz become more plentiful after periods of declarative memory encoding during waking, and together with hippocampal sharp wave/ripple EEG patterns may reflect hippocampal–neocortical dialog. Depth EEG recordings have also implicated interactions between hippocampus and medial temporal cortex in memory processing during sleep. Various physiological events that have yet to be deciphered may support memory processing during multiple stages of sleep. Many dreams occur together with eye movements, and these rapid eye movement (REM) periods have some clear ties to memory, perhaps emotional memories in particular. However, some evidence suggests that REM may not be the only time, or even the foremost time, for memory reprocessing. Another stage, slow-wave sleep (SWS), which is prevalent early in the night, may play a vital role in hippocampal–neocortical interactions that promote declarative memory consolidation. During SWS, reduced cholinergic activation and/or cortisol feedback in the hippocampus may promote hippocampal information flow to the neocortex. One possible sign of the processing of declarative memories during SWS is a buildup of slow negative potentials over frontal cortex, and correlations have been observed between SWS and later performance on declarative memory tests for information learned prior to going to sleep. Although many questions remain about memory processing during sleep, this research area will undoubtedly be a source of much new thinking on consolidation in the coming years.

See also: Amnesia: Declarative and Nondeclarative Memory; Cognitive Neuroscience: An Overview; Declarative Memory System: Anatomy; Episodic Memory; Hippocampus and Neural Representations; Hormones and Memory; Memory: Computational Models; Memory Consolidation: Cerebral Cortex; Memory Disorders; Memory Representation; Short Term and Working Memory.

Further Reading

Axmacher N, Mormann F, Fernández G, et al. (2006) Memory formation by neuronal synchronization. *Brain Research Reviews* 52: 170–182.

Burnham WH (1903) Retroactive amnesia: Illustrative cases and a tentative explanation. *American Journal of Psychology* 14: 382–396.

Buzsáki G (1998) Memory consolidation during sleep: A neurophysiological perspective. *Journal of Sleep Research* 7 (supplement 1): 17–23.

Cho YH, Beracochea D, and Jaffard R (1993) Extended temporal gradient for the retrograde and anterograde amnesia produced by ibotenate entorhinal cortex lesions in mice. *Journal of Neuroscience* 13: 1759–1766.

Clark RE, Broadbent NJ, Zola SM, and Squire LR (2002) Anterograde amnesia and temporally graded retrograde amnesia for a nonspatial memory task after lesions of hippocampus and subiculum. *Journal of Neuroscience* 22: 4663–4669.

Dudai Y (2004) The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology* 55: 51–86.

Duncan CP (1949) The retroactive effect of electroshock on learning. *Journal of Comparative Physiology and Psychology* 42: 32–42.

Eichenbaum H (2002) *The Cognitive Neuroscience of Memory: An Introduction.* New York: Oxford University Press.

Frankland PW and Bontempi B (2005) The organization of recent and remote memory. *Nature Reviews Neuroscience* 6: 119–130.

Gais S and Born J (2004) Declarative memory consolidation: Mechanisms acting during human sleep. *Learning and Memory* 11: 679–685.

Hebb DO (1949) *The Organization of Behavior.* New York: Wiley Press.

Kapur N (1993) Focal retrograde amnesia in neurological disease: A critical review. *Cortex* 29: 217–234.

Kim JJ and Fanselow MS (1992) Modality-specific retrograde amnesia of fear. *Science* 256: 675–677.

Kim JJ, Clark RE, and Thompson RF (1995) Hippocampectomy impairs the memory of recently, but not remotely, acquired trace eyeblink conditioned responses. *Behavioral Neuroscience* 109: 195–203.

Kopelman MD (in press) Retrograde memory loss. In: Miller BL and Goldenberg G (eds.) *Handbook of Clinical Neurology: Neuropsychology and Behavior.* Elsevier.

Manns JR, Hopkins RO, and Squire LR (2003) Semantic memory and the human hippocampus. *Neuron* 38: 127–133.

McGaugh JL (2000) Memory: A century of consolidation. *Science* 287: 248–251.

Meeter M and Murre JM (2004) Consolidation of long-term memory: Evidence and alternatives. *Psychology Bulletin* 130: 843–857.

Moscovitch M, Rosenbaum RS, Gilboa A, et al. (2005) Functional neuroanatomy of remote episodic, semantic and spatial memory: A unified account based on multiple trace theory. *Journal of Anatomy* 207: 35–66.

Paller KA (1997) Consolidating dispersed neocortical memories: The missing link in amnesia. *Memory* 5: 73–88.

Paller KA (2001) Neurocognitive foundations of human memory. In: Medin DL (ed.) *The Psychology of Learning and Motivation*, vol. 40, pp. 121–145. San Diego, CA: Academic Press.

Paller KA and Voss JL (2004) Memory reactivation and consolidation during sleep. *Learning and Memory* 11: 664–670.

Schacter DL and Tulving E (eds.) (1994) *Memory Systems 1994*. Cambridge, MA: MIT Press.

Squire LR (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* 99: 195–231.

Squire LR, Cohen NJ, and Nadel L (1984) The medial temporal region and memory consolidation: A new hypothesis. In: Weingartner H and Parker E (eds.) *Memory Consolidation*, pp. 185–210. Hillsdale, NJ: Lawrence Erlbaum Associates.

Squire LR, Slater PC, and Chace PM (1975) Retrograde amnesia: Temporal gradient in very long term memory following electroconvulsive therapy. *Science* 187: 77–79.

Squire LR and Schacter DL (eds.) (2002) *The Neuropsychology of Memory,* 3rd edn. New York: Guilford Press.

Takashima A, Petersson KM, Rutters F, et al. (2006) Declarative memory consolidation in humans: A prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America* 103: 756–761.

Zola-Morgan SM and Squire LR (1990) The primate hippocampal formation: Evidence for a time-limited role in memory storage. *Science* 250: 288–290.

Memory Disorders

A Fradera and M D Kopelman, King's College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Varieties of Amnesic Syndrome

The amnesic syndrome can be defined as an abnormal mental state in which memory and learning are affected out of all proportion to other cognitive functions in an otherwise alert and responsive patient. Various disorders can give rise to an amnesic syndrome, including the alcoholic Korsakoff syndrome, herpes encephalitis, severe hypoxia, certain vascular lesions, head injury, deep midline tumors, basal forebrain lesions, and occasionally the amnesic variant of mild cognitive impairment Alzheimer's dementia. These syndromes can produce both anterograde amnesia (AA), whereby memory is affected for events following the onset of the disorder, and retrograde amnesia (RA), affecting events from before the onset.

Korsakoff Syndrome

Korsakoff syndrome is the result of nutritional depletion (i.e., thiamine deficiency, due most commonly to alcohol abuse). Diagnosis typically follows an acute Wernicke encephalopathy involving confusion, ataxia, nystagmus, and ophthalmoplegia; however, onset can also be insidious, with little history of these features. Korsakoff patients will show intact short-term memory over spans of seconds but severe impairment for learning over longer periods, together with a retrograde memory loss that typically covers years or decades. Reasoning is broadly intact but patients will often exhibit some abnormal facets of cognition, including repetitive questioning and a dysexecutive profile that may alter behavior – for example, apathy or irritability – together with disorientation in time and its attribution; patients may also confabulate. Consistent with this, some degree of general cortical atrophy is usually observed, particularly in the frontal lobes.

The Korsakoff syndrome is notable in that there is a distinct neurochemical pathology, with important implications for treatment. Research on animals and on malnourished prisoners-of-war established thiamine depletion as the pathological basis for the acute Wernicke episode, and the subsequent (Korsakoff) memory impairment. The mechanisms by which thiamine depletion acts to produce the Wernicke–Korsakoff neuropathology is not fully understood, nor is the genetic contribution to acquiring the condition. Along with the mamillary bodies, the thalamus is considered the critical site of damage in producing the amnesia; early research suggested that damage to the medial dorsal nucleus was key, but more recent evidence points to the anterior principal nucleus.

Herpes Encephalitis

Herpes encephalitis (HSE) can give rise to a severe form of amnesic syndrome. The onset is typically abrupt, arising either from primary infection or reactivation of the virus from an earlier infection, and accompanied by acute fever, headache, nausea, and, in some cases, seizures and behavioral changes. Diagnosis is by the polymerase chain reaction (PCR) test or a raised titer of antibodies to the virus in the cerebrospinal fluid (CSF), but sometimes a presumptive diagnosis is made on the basis of the clinical picture as well as severe signal alteration in the temporal lobes on brain magnetic resonance imaging (MRI), associated with loss of tissue volume, hemorrhage, and edema.

The disorder will usually produce extensive temporal lobe damage that is usually bilateral, together with some general cortical atrophy and changes in orbitofrontal regions and surrounding frontal lobe. The medial temporal lobe structures are usually particularly severely affected (see Figure 1) including the hippocampi, amygdalae, and the entorhinal, perirhinal, and parahippocampal cortices – areas critical to memory formation. The disease may also implicate basal forebrain structures which give cholinergic outputs to the hippocampi.

Herpes encephalitis and the Korsakoff syndrome lead to a very similar memory disorder that reflects the fact that the hippocampi are linked to the thalami and mamillary bodies by several neural connections. These patients show similar forgetting rates. However, herpes patients typically show more severe involvement of early retrograde memories, a particular impairment for spatial memory (especially when the right hippocampus is implicated, which can result also in an impairment in face recognition) and a greater degree of semantic memory loss from concomitant damage to inferolateral temporal structures. These patients may, however, have a better insight into their memory disorder.

Severe Hypoxia

Severe hypoxia can give rise to an amnesic syndrome following carbon monoxide poisoning, cardiac and respiratory arrests, or suicide attempts by hanging or by poisoning with the exhaust pipe from a car. Drug overdoses may precipitate prolonged unconsciousness and cerebral hypoxia, and this quite commonly occurs in heroin abusers. The consequences of this can be quite variable, but typically severe hypoxia

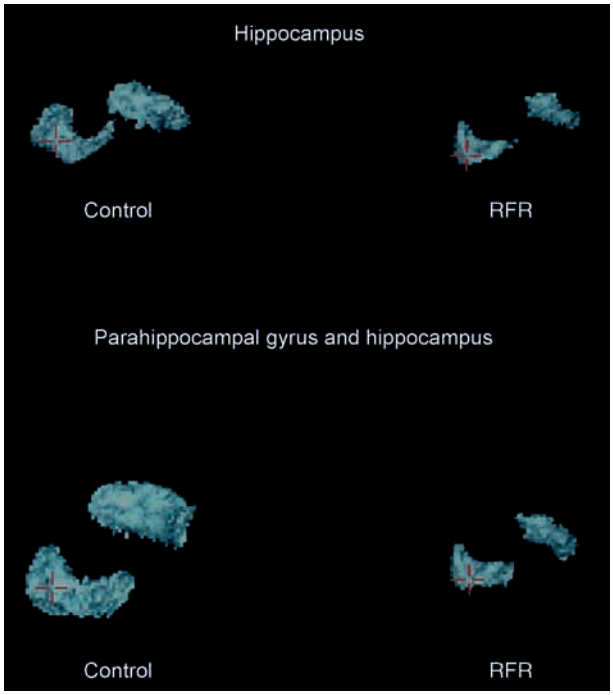

Figure 1 Three-dimensional surface renderings of the hippocampi (top) and the combined parahippocampal and hippocampal structures (bottom) in a healthy control volunteer (left) and in a herpes encephalitis patient (right). These images are taken from planimetric segmentations of the structures from three-dimensional magnetic resonance images. Reproduced from Kopelman MD, Lasserson D, Kingsley DR, et al. (2003) Retrograde amnesia and the volume of critical brain structures. *Hippocampus* 13: 879–891, with permission.

involves changes in the hippocampi, with evidence of loss of pyramidal cells in the CA1 region reported. Fluorodeoxyglucose–positron emission tomography (FDG–PET) studies have also shown reduced glucose metabolism in the thalamus.

Vascular Disorder

Vascular disorders can particularly affect memory, as opposed to general cognitive functioning, in cases of (1) thalamic, medial temporal, or retrosplenial infarction or (2) subarachnoid hemorrhage. Thalamic damage produces amnesia when there is anterior thalamic involvement, as this is where the mamillothalamic tract projects into the thalamus. The features of this disorder are an anterograde amnesia with a limited retrograde component (although when frontocortical projections are affected, a more extensive retrograde amnesia may be observed). Hippocampal damage is due to disruption of the supply from the anterior and posterior choroidal arteries, the latter being a branch of the posterior cerebral circulation; bilateral damage will produce a global amnesia whereas unilateral damage produces a more material-specific memory loss. The retrosplenium is also supplied from the posterior cerebral artery, and infarction or hemorrhage can produce amnesia by disrupting connections to the anterior thalamus, entorhinal, and parahippocampal cortices.

Subarachnoid hemorrhage following rupture of an aneurysm can result in memory impairment, whether the anterior cerebral or posterior cerebral circulation in the Circle of Willis is involved. Ruptured aneurysms in the anterior communicating artery can implicate the basal forebrain and ventromedial frontal structures. To produce a persistent amnesic syndrome it is thought critical that damage occurs to the septal nucleus of the basal forebrain, which provides cholinergic projections to the hippocampi.

Head Injury

Following head injury both transient and persisting amnesias can arise. Posttraumatic amnesia (PTA) is assumed to reflect the degree of underlying diffuse brain pathology (axonal shearing and tearing) resulting from rotational forces, but other pathological mechanisms include contrecoup damage, hypoxia, and intracranial hemorrhage. During PTA patients show accelerated forgetting of material, and following recovery from head injury memory impairments (in new learning) are the slowest cognitive function to improve. Head injury can also produce an RA which is usually fairly brief, but may be extensive in severe cases. Although most patients with head injury show a recovery over the course of 1–3 months, a proportion will show deficits that extend well beyond this. It appears that amnesia following head injury does not insulate against posttraumatic stress disorder (PTSD) for the event, as windows of experience tend to be preserved that allow the reexperiencing of emotionally harrowing elements of the event.

Transient Global Amnesia

Transient global amnesia (TGA) occurs over a period of several hours (average 4–6 h, up to 12 h), in which the sufferers (typically older males) retain their sense of personal identity but show a profound anterograde amnesia together with some degree of RA, typically more pronounced for episodes than for facts. This episode is sometimes preceded by headache or nausea, a stressful life event, a medical procedure, intense emotion, or vigorous exercise, and an episode of migraine may be a precipitating factor. TGA is currently understood to be due to transient dysfunction in limbic–hippocampal circuits, crucial to memory formation.

Transient Epileptic Amnesia

Transient epileptic amnesia refers to a small minority of TGA patients in whom epilepsy appears to be the

underlying cause. Standard electroencephalogram (EEG) and computed tomography (CT) scan findings are often normal, but abnormalities may be found on sleep EEGs. Episodes may be ictal or postictal in nature. Memory deficits between attacks are sometimes reported, both for new learning and particularly gaps in past memory. One possible explanation for these gaps is that brief runs of seizure activity at a clinically undetectable level have disrupted normal encoding. Another is that the current epilepsy somehow prevents the retrieval of old, autobiographical memories. Automatisms and postictal confusional states that can arise from epilepsy are also associated with memory disruption: automatisms involve limbic structures that are otherwise responsible for memory formation, and hence amnesia for these periods is always present and is usually complete.

Summary of Amnesic Syndrome

The preceding examples of amnesic syndrome do not, of course, comprise a comprehensive list of the disorders which give rise to transient or persistent amnesia. What they have in common is the neuroanatomical basis for the memory disruption: temporary or permanent dysfunction or damage in medial temporal/diencephalic circuitry, or damage to the basal forebrain cholinergic inputs to that circuitry. Our discussion now turns to the different components of memory that can be affected by damage to this circuitry, and interactions with frontal lobe and temporal neocortical structures. These are anterograde and retrograde episodic memory deficits (see Figure 2), semantic memory deficits, and confabulatory episodes.

Various aspects of anterograde memory can be affected in amnesic disorders. Anterograde memory encompasses working memory – which holds information for brief periods of time (a matter of seconds) and allocates resources – and the more permanent episodic memory – which is explicit and refers to specific learned events. It also covers forms of memory whereby information is decoupled from its learning context, such as semantic memory for facts and knowledge, and implicit memory in the form of classical conditioning, procedural learning of perceptuo-motor skills, and priming effects.

Episodic Memory Deficits

There has been much debate about the nature of the episodic memory impairment in amnesia. An overview of the basic processes of memory is illustrative: memories must initially be registered in some way (encoded); they must be stored in an effective fashion, and they must be able to be recovered from the memory system (retrieval). Amnesia could in principle be due to problems at any or all of these stages.

It now seems unlikely that initial encoding is failing in amnesia, as memory performance is largely unaffected by conditions that maximize encoding support for studied items; also, working memory can be preserved in the presence of profound amnesia. A more credible account is that disruption occurs later, during the transition of memories into long-term storage, termed consolidation, which may involve processes operating over hours as well as months or even years. The hippocampus has been identified as a critical site in terms of this consolidation, but the process and nature of the process are still under debate, particularly with regard to the implications of a longer term consolidation process for retrograde amnesia (discussed more fully later). Memories that are successfully consolidated into the store may be threatened by the failure of storage itself: this would be reflected by accelerated rates of forgetting. Following work with

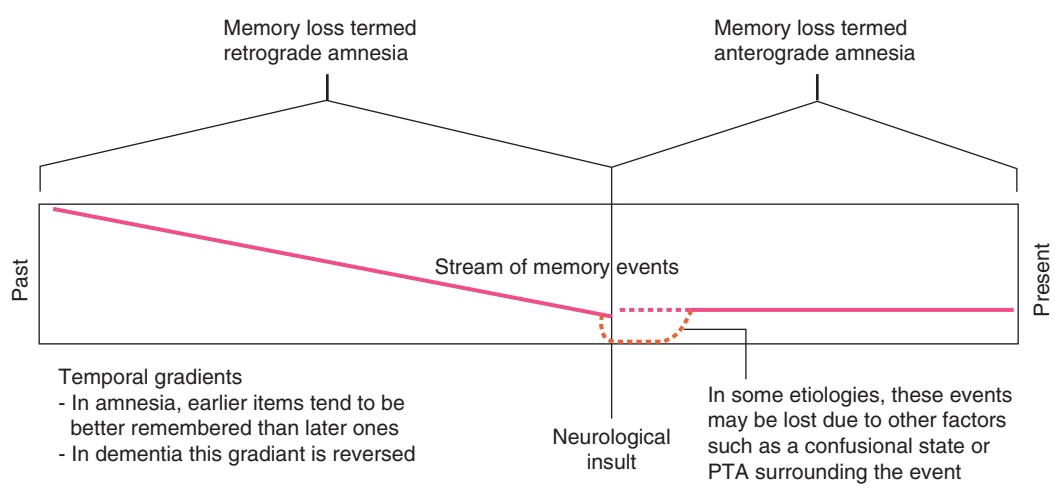

Figure 2 The taxonomy of episodic memory loss due to amnesia (PTA, posttraumatic amnesia). Lines with higher gradients denote better memory for that period.

the medial temporal lobe patient HM and others, this account was put forward to distinguish hippocampal from diencephalic amnesia, but subsequent, more stringent studies found that differences in the forgetting rate of patients from controls were restricted to recall rather than recognition memory, and that this was true across both types of pathology.

A further account is that amnesia reflects a specific deficit in the acquisition of contextual information, in terms of the spatial and temporal information tying an item to the environment in which it was encountered. Such specific deficits are indeed found in amnesics, but not in all cases, and the patterns differ substantially across groups. In terms of the hippocampal lesion sites associated with amnesia, the account has been reformulated as a deficit in binding complex associations or relations (not merely between item and position, but between very different types of items, such as pairing words and pictures); amnesia is thought to disrupt the flexible association of items, and the composition of these items with respect to one another.

Finally, the impairment may lie at the final stage of memory processing, the retrieval of items from storage. The case for the pure retrieval hypothesis is made on the basis of findings showing that amnesics clearly have memory information but produce it inappropriately – for example, producing true memory items on a task that requires another type of item, as well as evidence that retrieval cues can improve performance. However, this pattern can also be seen in the performance of healthy participants, suggesting that it may be a consequence of poor memory rather than its cause. Cueing effects may also simply reflect priming effects – that is, an advantage to memory that is purely implicit in nature.

In summary, the balance of evidence seems to favor a primary role for an acquisition/consolidation problem, some slighter impairments in retention (forgetting), detectable only on recall testing, and a possibly secondary deficit in retrieval processes. Basic research has demonstrated a biological mechanism for consolidation and its disruption, and primate work indicates that medial temporal lobe damage disrupts pair-coding neuronal responses, a mechanism likely to be critical in the support of long-term memory representations. The disruption of acquisition in amnesia may be particularly characterized by a deficit in binding different types of material, including contextual information, and this relates to various distinctions drawn within episodic memory, which is now considered.

Explicit and Implicit Memory

Relative to explicit memory, implicit memory is generally much better preserved in amnesia. Patients show preserved perceptuo-motor skills, and respond appropriately to simple forms of classical conditioning (such as the eye-blink response). They also show priming effects (a more rapid response to an item on the basis of its prior presentation, or that of a similar item), suggesting that this process does not rely on diencephalic/medial temporal brain structures, but rather upon cortical regions. Cases have been described whereby perceptual priming is impaired, even in the face of preserved explicit memory; a role for occipital lobe circuits has been implicated in these cases. However, amnesic patients do show impaired priming in certain experimental conditions (e.g., in difficult tasks when baseline responding has been controlled) or in associative learning. This implicates a contribution of medial temporal/diencephalic structures to priming in these circumstances.

Recall and Recognition Memory

The distinction in memory between recall (items are directly retrieved from memory with little support from cues) and recognition (information is provided to an individual who must simply make a judgment as to whether it has been previously experienced) is of great current interest in contemporary research. At issue is whether amnesic patients may show a disproportionate impairment for one form over the other, suggesting that there may be distinct neural loci underlying these processes. One position is that recall is subserved by a set of neural structures comprising the hippocampi, fornices, mamillary bodies, mamillothalamic tract, and anterior thalami – structures which are crucial in attributing contextual information, including time and spatial location, to memory items, thereby allowing the items to be distinguished and retrieved in the manner required for recollection. Recognition is supported by brain regions (including the perirhinal cortex) which are critical for familiarity judgments. There is some evidence that hippocampal damage disproportionately affects recall memory, as in some cases of developmental amnesia.

An opposing position is that both recall and recognition are served by medial temporal lobe structures and that dissociations across recall and recognition tasks reflect the severity of amnesia. Particular support for this position is given by studies showing that damage limited to the hippocampal formation has produced impairments in recognition memory. This is a touchstone issue which is continuing to preoccupy researchers.

Remembering and Knowing; Recollection and Familiarity

Gardiner and Richardson-Klavehn have distinguished between the subjective states of remembering and

knowing. Remembering (R) involves the reexperience of 'intensely personal experiences of the past,' whereas knowing (K) is when 'we are aware of knowledge that we possess but in a more impersonal way,' without the experience of reliving the acquisition of this knowledge. Do these types of memories arise from a single unitary process, from two independent but overlapping processes, or from entirely exclusive processes? Evidence that the same memory items can be variously experienced as R or K, depending on the passage of time, has been used to argue that the processes are not exclusive. Extensive work has examined how these concepts map onto the recall and recognition divisions of memory. It seems likely that both familiarity and recollection can contribute to recognition, with characteristics of performance varying according to the experience supporting it (e.g., judgments that rely upon recollection show greater levels of subjective confidence). Familiarity is not accompanied by the contextual detail required to spontaneously evoke a piece of memory information, whereas recall necessarily relies upon recollection. An analysis of data from several studies of amnesia (due to hippocampal or diencephalic damage) suggests that recollection is disrupted to a greater extent than is familiarity. This evidence leads to the possibility that a defining feature of amnesia is the loss of contextual/associational information, leading to a memory experience that is both impoverished and lacking the subjective experience of 'mental time travel' – the rich and fluid memorial reexperiencing of an event.

Semantic Memory Deficits

Semantic memory comprises knowledge of facts, concepts, and language, stored without corresponding information about the initial learning experience. The cognitive neuroscience of semantic memory – in particular, its relation to episodic memory – is an issue of current theoretical interest. There are several ways in which this issue has been investigated.

Semantic Dementia

Semantic dementia is a degenerative disorder that causes a progressive loss of semantic knowledge which can occur across both verbal and nonverbal domains. In the verbal domain, this leads to loss of word comprehension and naming, and increasingly degraded and empty speech, though the latter remains fluent and grammatical in output. Perceptual and reasoning abilities are also intact, and episodic memory is relatively preserved, as are some classes of knowledge, such as numerical information. The underlying pathology is generally confined to inferior and lateral temporal gyri, more commonly left-lateralized, although right-lateralized atrophy can occur, producing impairments in visual semantics.

Patients often show impaired knowledge of subordinate meaning categories relative to superordinate categories, such as loss of 'dog' but preservation of 'animal'; over time, however, superordinate knowledge may also succumb to the disease. This has been taken as evidence for a hierarchical organization of semantic memory, but it may reflect a more secure neural underpinning for the superordinates, due to their higher usage frequency or their greater degree of connectedness to other concepts. Crucial to the debate on memory systems is the fact that, when compared with amnesic disorders, there can be a classic double dissociation, in that semantic dementia results in semantic impairments with preserved episodes, whereas amnesia leads to episodic impairments with relatively preserved semantic information.

New Learning of Semantic Material in Amnesia

Individuals who suffer severe amnesia provide the opportunity to assess whether the acquisition of semantic material depends upon the episodic memory system. In principle, if new semantic knowledge can be acquired despite a very severe episodic memory deficit, it implies a separation of the neurobiological systems underlying semantic memory and episodic memory. A number of studies in the past 10 years suggested that patients who show severe episodic memory impairments may show significantly better performance on tasks of new vocabulary or recognizing public figures and events. For example, a densely amnesic individual with almost complete bilateral destruction of the hippocampi and severe atrophy of the perihippocampal tissue was tested on knowledge of vocabulary and famous individuals that had arisen since the onset of amnesia; he demonstrated access to explicit semantic knowledge in these domains, albeit at lower levels than within matched controls. However, this finding has been contested, due to some cases of medial temporal lobe amnesia apparently showing little ability to acquire new knowledge.

Cases of developmental amnesia have been studied recently, again offering the opportunity to assess current semantic memory when the acquisition of new episodic memories has been severely impaired. For example, the patient Jon, amnesic due to neonatal injury, is impaired on recall tasks and does not describe any subjective sense of episodic recollection. However, he has been able to perform relatively well in mainstream education – developing competencies in speech, reading, and writing, and scores within the normal range on measures of comprehension, world knowledge, and vocabulary that tap semantic information.

Retrograde Amnesia

In this section, we address amnesia for past events: its nature in different kinds of memory disorder, its implications for theories of memory storage, and whether it can occur in the absence of broader memory impairment.

Temporal Gradients

Investigations of remote memory loss have often striven to assess Ribot's law, expressed in 1882:

> The progressive destruction of memory follows a logical order . . . it begins at the most recent recollections which, being . . . rarely repeated and . . . having no permanent associations, represent organization in its feeblest form. (Théodule Ribot, 1839–1916)

This is reflected in a temporal gradient whereby the most recent memories are lost but more remote ones are intact. Such temporal gradients have been reported in a number of patients with amnesia of both temporal lobe and diencephalic origin, and a more gentle gradient is also observed in Alzheimer's dementia (see Figure 3).

What is the basis of these gradients? In Korsakoff patients, one possible explanation is that prolonged heavy drinking leading up to the Wernicke episode might result in a progressive anterograde memory loss over this period. However, this view is not supported by cases in which memories lost due to the syndrome are known to have been in the awareness of the individual prior to onset (e.g., memories were described in diary entries), nor by the fact that severity of drinking does not correlate with severity of retrograde deficit. Drinking may, at best, explain the relative steepness of the gradient compared with other syndromes, but not the gradient itself.

Consolidation theory suggests that the medial temporal/diencephalic system is responsible for both memory formation and consolidation of memories into permanence over a period of years, whereas older memories are permanently distributed through the neocortex. This consolidation/structural reallocation account is supported by findings of minimal RA when lesions are confined to the diencephalon or medial temporal lobe, whereas more substantial RA is underpinned by more extensive pathology, which at its extreme (when there is widespread damage of the kind observed in HSE or Alzheimer dementia) produces a flat temporal gradient. However, this consolidation process would operate over a much longer duration than that involved in transfer from short-term to long-term memory, and an extensive temporal gradient, going back 20–30 years, would imply that physiological consolidation occurs over very long durations, which is problematic for this account.

Another account is the multiple trace theory (MTT), in which the hippocampi are assumed to be continuously involved in the storage and retrieval (reactivation) of memory traces. This model proposes that activations within neocortical and other structures represent information, and that these are bound into a memory trace by hippocampal neurons, acting as an index or marker of the distributed memory information. Each reactivation of a memory trace results in the formation of multiple traces indexed by the hippocampus, and the extraction of more generalized factual information into semantic memory stores. In turn, the existence of multiple traces results

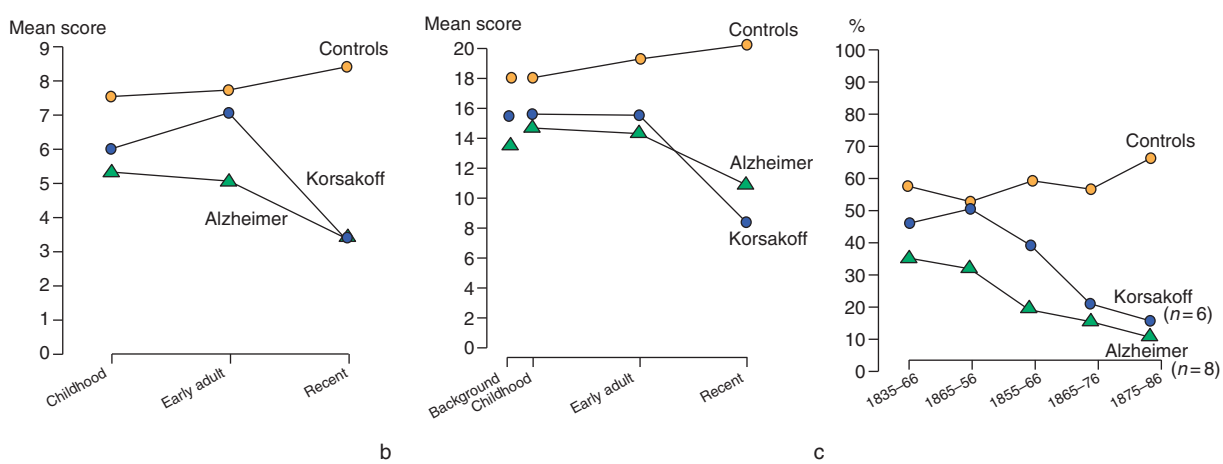

Figure 3 Temporal gradients across different tasks. Remote memory curves on different types of tasks obtained in healthy control participants, Korsakoff patients, and Alzheimer patients. (a) Autobiographical incident recall. (b) Personal semantic fact recall. (c) News event recall. Adapted from Kopelman MD (1989) Remote and autobiographical memory, temporal context memory, and frontal atrophy in Korsakoff and Alzheimer patients. *Neuropsychologia* 27: 437–460, with permission from Elsevier.

in temporal gradients with relative preservation of older (more rehearsed) memories, and the extent of RA would depend upon the size of the hippocampal lesion (with severe atrophy/damage leading to a flat gradient). However, past factual information is not coupled to the hippocampus and hence semantic memories can show a temporal gradient, even when severe hippocampal pathology results in a flat gradient for episodic memories.

A third theory is that, over time, episodic memories are recomposed in a semantic form, losing their vividness and contextual (episodic) aspects: a transformation of memories, rather than their replication. Semantic memories are better insulated from loss in brain damage, as they are presumably stored in more widespread and interconnected networks.

At the time of writing, these various positions remain as credible alternative theories of retrograde amnesia. In support of MTT are reports of patients with steep gradients for personal semantic facts together with flat autobiographical memory gradients, following medial temporal lobe damage. However, many such patients also have extensive damage elsewhere, and other patients with damage confined to the medial temporal lobe or diencephalon show a brief, graded RA – the latter is more consistent with the structural reallocation account of RA. Meanwhile, the episodic-to-semantic memory account – wherein early semanticized memories are easier to retrieve under suboptimal conditions – is consistent with evidence of a retrieval deficit in RA, and the weakness of the relationship between anterograde and retrograde deficits. However, temporal gradients can also be found within the semantic domain, such as for new word learning.

Focal Retrograde Amnesia

A number of cases in the literature have been reported to show a profound RA, with anterograde amnesia slight or absent. However, closer examination of these cases show they provide only limited evidence for the existence of a neurological focal retrograde amnesic syndrome. In a review of the literature, Kopelman found that in many cases the anterograde component was substantial, in others modest but certainly present, and that effort has not always been made to equate the difficulty of retrograde and anterograde tests, which may differentially rely on specific cognitive processes such as visual imagery. Accounts of focal RA have relied on impairments in semantic, autobiographical, or public event knowledge, and may have overemphasized isolated dips in performance in the absence of a more generalized impairment. Moreover, a thorough assessment of psychiatric state is rarely reported; this is relevant, as the retrograde impairment may reflect a psychological or simulated component (see Figure 4). The more convincing cases have included those who showed an initial anterograde amnesia which then abated, and cases of transient epileptic amnesia; however, the former may be better characterized as instances of failure to recover retrograde amnesia, and the latter could be explained as the consequence of periods of faulty encoding across the life span due to brief runs of seizure activity. Cases in which a psychological component made a substantial contribution are interesting in their own right and pose issues both for the clinician and for the researcher of memory. There is continuing debate as to whether psychogenic cases are underpinned by a neurophysiological component – such as in frontal control mechanisms – or whether their understanding is best attempted at a more cognitive or psychosocial level, such as memory suppression mechanisms deployed in response to certain stressors.

Confabulation and Memory Distortion

Confabulation was defined by Berlyne as "falsification of memory occurring in clear consciousness in association with an organically derived amnesia." He argued that it can arise either in a momentary fashion or as spontaneous outbursts from the patient. The former is far more common, and is characterized by the confabulated content being produced as a response to a demand for memory information unavailable to the subject. These provoked confabulations can be viewed as an extension of the kind of intrusion errors seen in recall memory performance by healthy individuals, as well as occasional fill-ins and guesses intended to smooth over awkwardness. Spontaneous confabulation is rarer, but when it occurs it tends to persist either over a short period of time, such as within a confusional state, or over more extended periods, in which case it is usually associated with ventromedial frontal lobe pathology.

A number of theories of spontaneous confabulation have been proffered. One is that the memory errors are the consequence of misattribution of context to memories, such that memories associated with a certain location or time are attributed to another (such as recalling a bank transaction and relocating this event to the current environment of a hospital). This idea originates from Korsakoff's descriptions of confabulations that were essentially confusions in time. Another possibility is that the fault lies not in context memory but within a broader evaluation of the origin of memories. According to this view, much of the patients' memorial information is impoverished or uncoupled from details of where or when it originated, from whom it was given, and

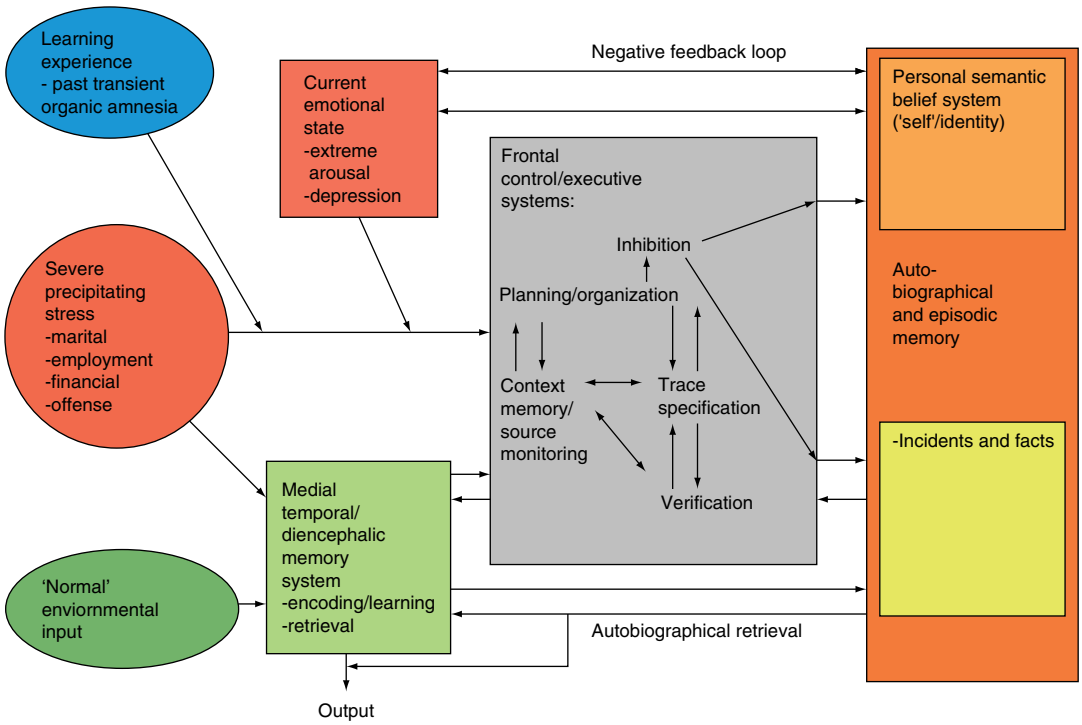

Figure 4 Social factors and brain systems influencing autobiographical memory retrieval and personal identity. The diagram shows frontal lobe processes (gray box) involved in confabulation (trace specification/verification; context memory/source monitoring), and social factors (ovals, circle) and brain systems (rectangles) involved in psychogenic amnesia: Inhibition leads to impaired retrieval of incidents and facts, and severe inhibition affects orientation in person. Adapted from Kopelman MD (2000) Focal retrograde amnesia and the attribution of causality: An exceptionally critical review. *Cognitive Neuropsychology* 17: 585–621.

whether the experience occurred at all, or was simply imagined. According to both of these views, the fault lies with the amount of memory information that is retrieved, whereby degraded information, or that uncoupled from associated information, results in confabulatory errors.

A further theory is that the stages of memory search and retrieval are disrupted such that memories are poorly selected, and then verified without adequate monitoring of their appropriateness. Initial memory search suffers either because of introduced noise or overbroad search criteria. Subsequent retrieval stages are also prone to failure, including checking that the memory accords with the initial goal of searching for it, as well as whether it is implausible or otherwise problematic. The patient thus produces erroneous memories that a healthy memory system would have halted within its verification system. Plausibility is given to this account by the fact that memory retrieval, information assessment, and judgments are all cognitive processes associated with frontal lobe function. There is also experimental evidence that confabulators make errors in memory judgments that are not dependent on context, but which depend upon accurate strategic retrieval, such as recalling details of fairy tales and bible stories.

Schnider has argued that the problem in confabulation is not in memory *per se*, but in a component of the brain's reward system that acts to suppress memory traces that do not pertain to current ongoing reality. According to this view, evoked memories are accompanied by mental associations to other more tangential memories; these are normally easily distinguished and inhibited when the individual wants to act or make a decision, but for confabulators these persist with a strong experiential sense of relevance (see Figure 5 for a depiction of this). Schnider notes that a neuroanatomical site common to spontaneously confabulating patients is the orbitofrontal cortex, part of the anterior limbic system which orchestrates communication between the neocortex and the subcortical reward system, directly via the ventral striatum. He also provides data which demonstrate that spontaneous confabulators fail an experimental test distinguishing currently relevant from irrelevant memories. Later, they pass this test once their confabulating has resolved. However, some recent evidence contests the specificity of this linkage.

There has been recent evidence that confabulations may be biased toward emotionally motivated content. Patients will produce information that is self-enhancing – that they are younger than they truly are,

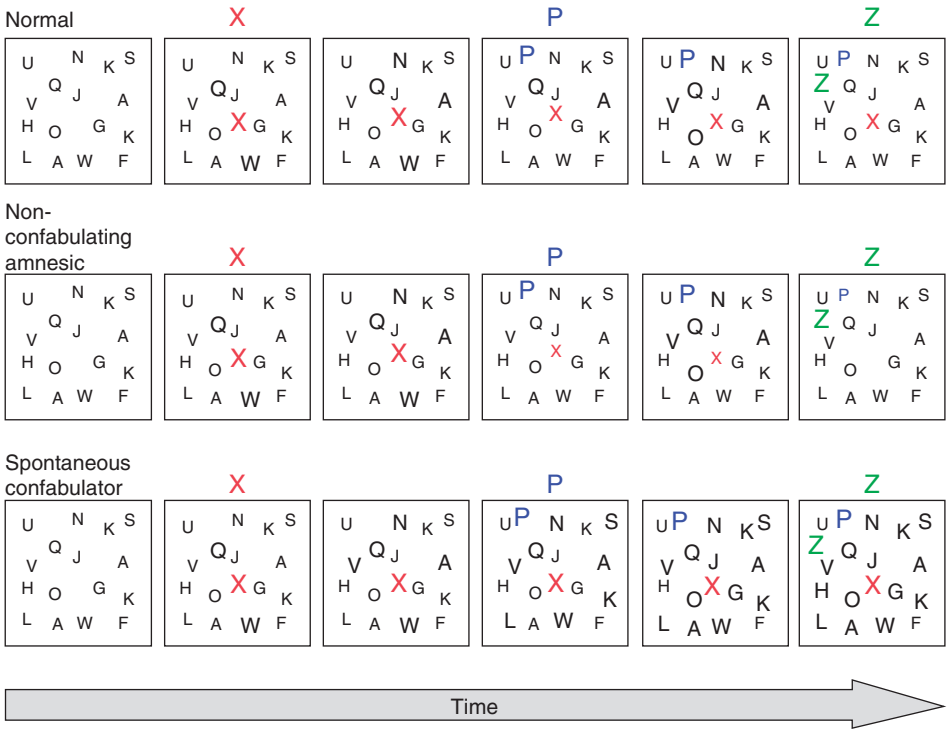

Figure 5 Model of 'now representation' in thinking. Any letter is meant to indicate an event or memory trace. The sizes of the letters indicate their saliency in thinking. Letters outside the boxes are external cues. In normal memory, traces are activated by external cues and associations with other activated cues. When the system is required to adapt to ongoing reality (a fresh cue), these items are then suppressed, with weaker (more tangential) traces fading quicker than recently seen ones; thus, the healthy brain remains in the 'now' but has access to the past. In amnesics, this process occurs, but the trace relating to the recently viewed item is lost at a rate similar to that for more tangential items, trapping amnesics in the 'now.' Spontaneous confabulators are unable to suppress the memories that are no longer currently relevant, allowing a tumult of traces to ensue.

that dead or distant relations are alive or near, or that the severity or awkwardness of their current situation is minimized. This may be because the damage to the executive system has eroded the distinctions between the idealized representation of self and the actual sense of self, such that idealizing distortions are more likely to be accepted than in healthy individuals.

Summary

The state of our understanding of memory disorders continues to develop. The natures of the various etiologies are well established and currently the focus of interest is upon the conceptual issues that underpin memory formation, loss, and confusion. The roles of the hippocampi in recollection and in recall memory and the putative hippocampal role in the reactivation or consolidation of memories are of crucial interest, as are formation of semantic memories and whether this depends upon the episodic memory store. The psychological contribution to memory disorders is likely to garner increasing attention, not just in psychogenic fugue and focal retrograde amnesia, but also in interaction with clearly neurological factors, as will the deficits that give rise to pathological false memory. Progress will be made through neuropsychological studies of patients, together with neuroscientific techniques – preferably employed in concert – resulting in better understanding of the causes of memory disorders.

See also: Amnesia: Declarative and Nondeclarative Memory; Animal Models of Amnesia; Episodic Memory; Event-Related Potentials (ERPs); Functional Amnesia; Hippocampus; Humans; Memory Consolidation: Systems; Recognition Memory; Semantic Memory; Short Term and Working Memory.

Further Reading

Aggleton JP and Brown MW (2006) Interleaving brain systems for episodic and recognition memory. *Trends in Cognitive Sciences* 10: 455–463.

Baddeley A, Kopelman MD, and Wilson B (eds.) (2002) *Handbook for Memory Disorders*, 2nd edn. Chichester: Wiley.

Cohen NJ, Poldrack RA, and Eichenbaum H (1998) Memory for items and memory for relations in the procedural/declarative memory framework. *Memory* 6: 6689–6699.

Fotopoulou A, Solms M, and Turnbull O (2004) Wishful reality distortions in confabulation: A case report. *Neuropsychologia* 42: 727–744.

Hodges J (2006) Alzheimer's centennial legacy: Origins, landmarks and the current status of knowledge concerning cognitive aspects. *Brain* 129: 2811–2822.

Kopelman MD (2000) Focal retrograde amnesia and the attribution of causality: An exceptionally critical review. *Cognitive Neuropsychology* 17: 585–621.

Kopelman MD (2002) Disorders of memory. *Brain* 125: 2151–2190.

Mayes A, Montaldi M, and Migo E (2007) Associative memory and the medial temporal lobes. *Trends in Cognitive Sciences* 11(3): 126–135.

Mishkin M, Vargha-Khadem F, and Gadian DG (1998) Amnesia and the organization of the hippocampal system. *Hippocampus* 8: 212–216.

Moscovitch M, Nadel L, Winocur G, et al. (2006) The cognitive neuroscience of remote episodic, semantic and spatial memory. *Current Opinion in Neurobiology* 16: 179–190.

Neary D, Snowden J, and Mann D (2005) Frontotemporal dementia. *Lancet Neurology* 4: 771–780.

Schacter DL and Slotnick SD (2004) The cognitive neuroscience of memory distortion. *Neuron* 44: 149–160.

Schnider A (2003) Spontaneous confabulation and the adaptation of thought to ongoing reality. *Nature Reviews Neuroscience* 4: 662–671.

Squire LR (2006) Lost forever or temporarily misplaced? The long debate about the nature of memory impairment. *Learning and Memory* 13: 522–529.

Squire LR and Schacter DL (2002) *Neuropsychology of Memory*, 3rd edn. New York: Guilford Press.

Warrington EK (1975) The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology* 27: 635–657.

Yonelinas AP (2002) The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language* 46: 441–517.

Relevant Websites

http://www.neuroskills.com – Centre for NeuroSkills, Traumatic Brain Injury Resource Guide.

http://www.gluck.edu – Memory Disorders Project, Rutgers University (Newark, NJ, USA).

Memory Representation

T Fukushima, H Kasahara, T Kamigaki, and Y Miyashita, The University of Tokyo, Tokyo, Japan

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During the course of everyday life, we are nearly continuously forming new memories – and then retrieving them on demand. Some of these memories are only short-lived, like memoranda (e.g., a telephone number when dialing); without rehearsal, they are gone from our mind within a few seconds. This type of memory is called short-term memory or, in other words, working memory, in that it is thought to be critical for thinking or higher cognitive functions. Other memories are long-lasting and accumulate in our brains. These include personal memories of past experiences (episodic memory; e.g., "you had a good dinner with your friend last weekend") and knowledge about the world (semantic memory; e.g., "a dog has four legs and wags its tail"). This type of memory is classified as long-term memory or, more precisely, long-term declarative memory, which can be consciously recollected. Both short-term memories and long-term declarative memories are thought to be represented and retained in the cerebral cortex. There are also other memory systems, including priming, procedural learning, classical conditioning, and emotional learning, which are collectively termed implicit or nondeclarative memory. These memory systems work unconsciously, and the results of their processing are expressed in behavior or performance of tasks. Although the cerebral cortex also plays a key role in nondeclarative memory systems, in this article we will focus on representation of short-term memory and long-term declarative memory, which will be described mainly based on animal studies carried out using macaque monkeys.

Short-Term Memory Representation during Working Memory Tasks

Short-term memory is thought to be represented in a network involving several cortical regions, including the prefrontal cortex and the posterior sensory cortices. Within this network, the prefrontal cortex is thought to play a pivotal role by controlling and making use of the subsidiary posterior cortices to enable retention of short-lived information that is relevant for goal-directed behavior at that moment. Our understanding of the role played by the prefrontal cortex derives from several observations. (1) Monkeys with prefrontal lesions show an impaired ability to perform delayed response tasks or delayed alternation tasks, in which they must retain a sample position or previously selected position for brief periods to get rewards. Moreover, monkeys with mid-dorsolateral prefrontal lesions show an impaired ability to perform working memory tasks, such as self-ordered tasks in which they must keep in mind their previous selections and utilize the information for subsequent correct responses. For instance, in one self-ordered task the monkey faced three food containers, each with a distinctly different shape and color. In the first trial, all three containers had a reward inside, and the monkey could select any container to get a reward. In subsequent two trials, however, the monkey would be rewarded only if he selected a container that was not selected in an earlier trial. Thus, especially in the third trial, he had to keep in mind his previous selections and refer to them to correctly avoid vacant containers. (2) During short-term memory tasks or working memory tasks, the prefrontal cortex exhibits sustained neuronal activity during delay periods in which monkeys are required to hold information in memory in order to perform their tasks. This sustained activity is maintained across intervening distractors as long as the information in short-term storage is task-relevant and necessary. (3) Neurons in the prefrontal cortex are able to encode a variety of types of information depending on the situation; they are not restricted to specific modalities. Prefrontal neuronal activity has been associated with short-term storage of information about stimulus identity, about stimulus position and about their concatenation, and even about more abstract notions, such as a task mode (which rule to adopt) and category (which category stimuli belong to). Prefrontal neurons also encode prospective information, including target identity or positions for impending responses. (4) Several prefrontal areas are further activated in imaging studies when required to manipulate the contents held in short-term memory. Moreover, a population of prefrontal neurons is involved in instructed shift of spatial short-term memory or in transformation from sensory-derived information to motor-related information. A keyword to describe the nature of prefrontal neuronal activity is thought to be 'relevance' – that is, neurons in the prefrontal cortex appear to be able to encode anything relevant to goal-directed behavior at that moment. That said, some mild tendency toward regional modal specificity is suggested in the prefrontal cortex: the dorsolateral region is associated with spatial information, the ventrolateral region with object information. This tendency is consistent with differences in corticocortical

connections between specific regions of the prefrontal cortex and the posterior sensory cortices.

The macaque inferior temporal (IT) cortex is one of the aforementioned posterior sensory cortices and plays a part in representing short-term memory of visual objects. The IT cortex is known to be the highest region in the object vision pathway, through which external objects are analyzed in a stepwise fashion from retinal visual inputs toward object identity. IT neurons respond to visual features which vary from moderately complex shapes (e.g., a spiny starlike shape or a striped triangle) to combinations of such shapes with color or texture. One early study revealed that IT neurons show sustained activity during the delay periods in a delayed matching-to-sample (DMS) task making use of colored stimuli. During the task, the monkeys were required to retain a sample color in mind for 16 to 20 s, after which they had to discriminate a matching choice color from a distractor color. Some of these delay-responsive neurons exhibited different levels of activity depending on the color of the stimuli, which suggests that this activity is a neuronal representation of color information that the monkeys retained in short-term storage while performing the task (**Figure 1**). In later studies, it was demonstrated that a population of neurons in the macaque IT cortex exhibit strong selectivity for complex computer-generated visual patterns. This selective activity was observed not only during stimulus presentation but also after offset of the stimulus, during the delay period when monkeys were required to keep the stimulus information in mind. For instance, when monkeys were trained to perform a DMS task involving a large set of fractal patterns and a 16 s delay period, it was found that individual anteroventral IT neurons commonly exhibited strong delay period activity only after presentation of certain stimuli making up a small fraction of the stimulus set (**Figure 2(a)**). Notably, this selective delay period activity was independent of the physical attributes of the sample stimuli – that is, it was unaffected by the size, orientation, color, and position of the stimuli (**Figure 2(b)**). Thus delay period activity in these IT neurons probably represents short-term information about the stimulus identity.

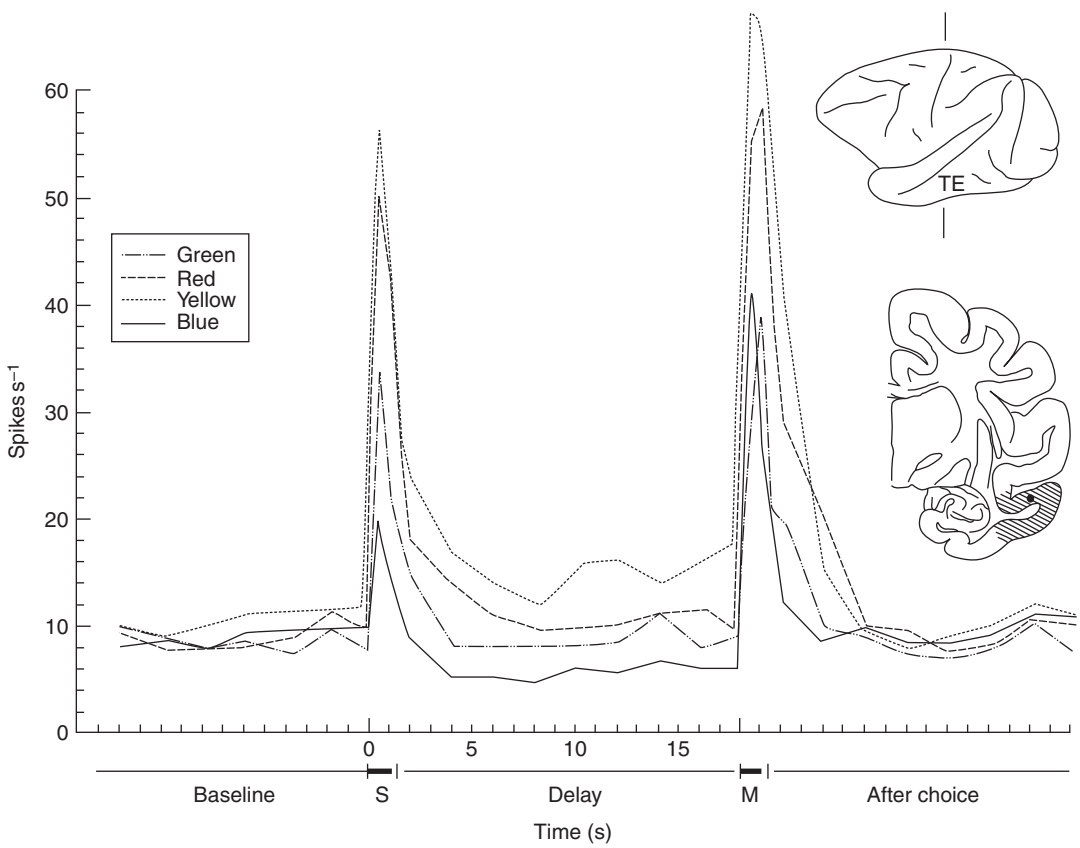

Figure 1 Activity of an IT neuron during a DMS task using four sample colors. Sample (S) and matching (M) periods are indicated by horizontal bars below the horizontal axis. Every line represents the mean firing rate during four to six trials with each of the four sample colors. The delay period activity of this neuron differed depending on the color of the sample. The neuron also exhibited color-dependent differences in activity during the sample and matching periods. The brain schemas on the right show area TE and the position of the neuron (black dot) in a cross-section of area TE (shaded). Reprinted with permission from Fuster JM and Jervey JP (1981) Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science* 212: 952–955. Copyright [1981] AAAS.

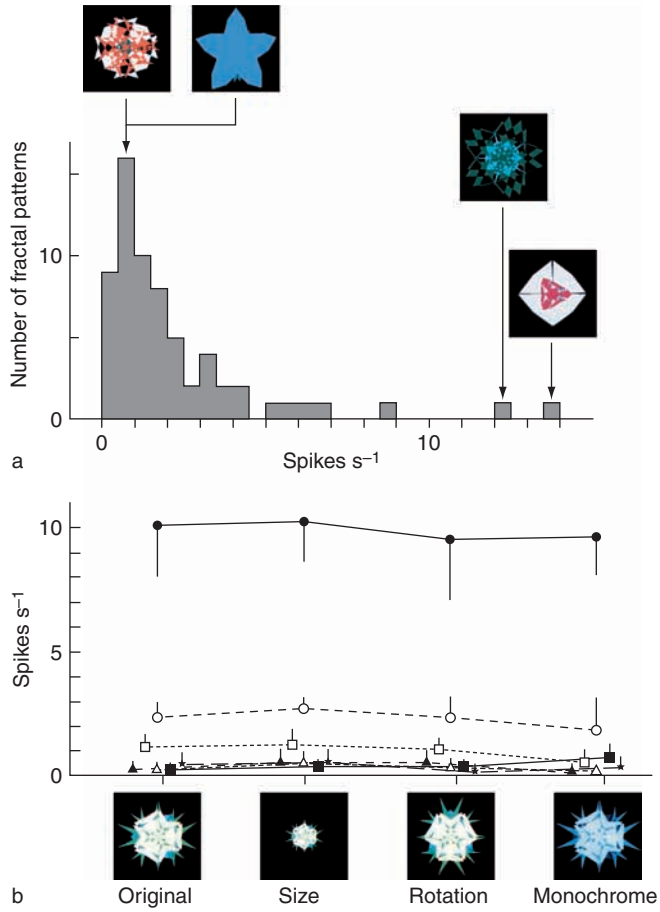

Figure 2 Response properties of macaque anteroventral IT neurons in a DMS task using a large set of colored fractal patterns. (a) Distribution of the responses of a neuron to 64 different fractal patterns. For each fractal pattern, the neuronal activity during the delay periods was averaged over more than four trials in which the fractal patterns were presented as sample stimuli. The horizontal axis indicates response strength in terms of firing rates, and the vertical axis indicates the number of fractal patterns. Examples of the fractal patterns used are also shown, with arrows indicating the neuronal responses they elicited. This neuron responded highly selectively to a few optimal stimuli (two examples are on the right-hand side) and only poorly to the others (the examples on the left-hand side). (b) Responses of another neuron. This neuron shows invariance in its responses after altering the size, orientation, or color of the stimulus: the original fractal patterns (original) were reduced in size by half (size), rotated 90° clockwise (rotation), or transformed from color to monochrome by referring to a pseudo-color lookup table (Monochrome). The averaged delay activities are plotted against stimulus transformation. This IT neuron discharged during the delay after one particular fractal pattern (closed circles) but not after others (the other markers), irrespective of pattern size, orientation, or color. Most of the neurons tested showed similar response invariance: size invariance was observed in 16 of 19 neurons, orientation invariance in 5 of 7 neurons, and color invariance in 15 of 20 neurons. With the neurons that did not show response invariance, manipulation of the most effective pattern reduced or abolished delay activity. The closed circles represent responses to the original optimal fractal pattern and its transformed patterns (shown at the bottom of the figure). Error bars indicate standard deviations for 4–15 trials. The other markers denote responses to the original suboptimal fractal patterns and their transformed patterns. Reproduced from Miyashita Y and Chang HS (1988) Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* 331: 68–70, with permission from Nature.

Long-Term Memory Representation in the Macaque IT Cortex

Emergent Stimulus Selectivity in IT Neurons

Neurons in the IT cortex show a high degree of plasticity with respect to their selectivity for visual stimuli. As described in the previous section, a substantial population of macaque IT neurons responds highly selectively to computer-generated geometric stimuli (Figure 2(a)), even though these artificial stimuli were completely unfamiliar to naive monkeys. Since it is highly unlikely that the IT cortex inherently contains neurons that are tuned to these computer-generated stimuli, their strong selectivity is thought to be newly acquired through repeated experience. Consistent with this idea, IT neurons have been shown to respond more selectively after training than before training to some types of stimuli,

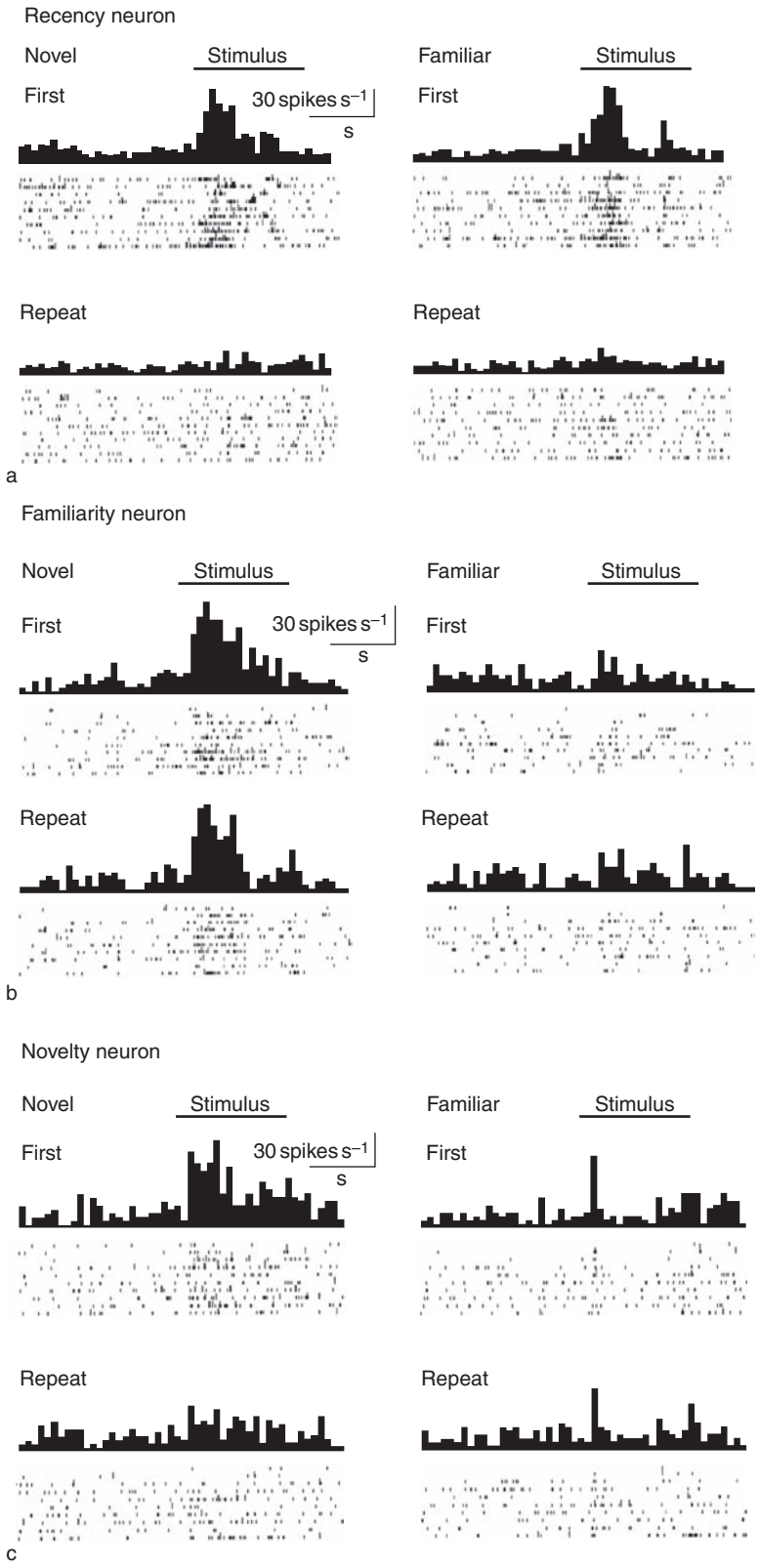

Figure 3 Responses of neurons in the perirhinal cortex and adjacent cortical areas to visual stimuli during a serial recognition task. (a) Responses of a recency neuron. The neuronal responses to the first (top row) and repeated (bottom row) presentations of novel (left) and familiar (right) stimuli are shown as peristimulus time histograms of the neuronal discharge and rastergrams denoting individual discharges in several trials. Note that the responses to the stimuli were significantly larger upon the first presentation in the session than they were upon repeated presentations, regardless of the stimulus familiarity – that is, these responses signaled whether the stimulus had

including fractal patterns, Fourier descriptors, baton-like stimuli, and animal-like stimuli. Training also increases the proportion of neurons that will selectively respond to a particular visual stimulus. Moreover, visual experience increases the similarity of the responses of nearby neurons to a particular visual stimulus. This emergent stimulus selectivity within the IT cortex appears to result from the reorganization of neural connections through experience or learning, which probably boosts perceptual processing of appropriate stimuli and leads to the development of expertise when monkeys are trained on a task. In that sense, the strong emergent stimulus selectivity of IT neurons can be said to be an expression of long-term memory representation.

Neuronal Correlates of Recognition Memory

Recognition memory, which is a type of long-term memory that includes an episodic aspect, is the ability to judge whether an item was recently encountered or not. This memory is thought to be largely dependent on the perirhinal cortex, which is situated in the anteroventral part of the IT gyrus neighboring the hippocampus, and contains neurons exhibiting activity that is modulated by the recency and familiarity of visual stimuli. For instance, when monkeys performed a serial recognition task in which they were required to discriminate novel or unfamiliar stimuli from familiar or recently seen stimuli, three types of neuronal responses were observed in the perirhinal cortex and adjacent cortical areas, including the entorhinal cortex and area TE (Figure 3). One group of neurons appeared to encode recent presentation of a visual stimulus: the neurons showed diminished responses to repeated presentation during a daily session or after 24 h, irrespective of stimulus familiarity. Another group of neurons appeared to encode the familiarity of a visual stimulus: these showed diminished responses to familiar stimuli, irrespective of the recency of presentation. A third group appeared to encode a combination of recency and familiarity: these neurons showed specific responses to the first presentation of novel stimuli. The proportions of these three neuronal types are similar in the perirhinal cortex and adjacent cortical areas, and their responses are thought to be highly important for performance of recognition memory tasks. Thus, reduced neuronal activity within the perirhinal cortex and adjacent cortical areas, related to the recency and familiarity of visual stimuli, may be neuronal correlates of visual recognition memory.

Neuronal Correlates of Semantic-Like Long-Term Memory

The stimulus selectivity of neurons in the macaque IT cortex may also support formation of long-term associative representations among disparate visual stimuli. For example, in an electrophysiological study, monkeys were trained on a DMS task making use of a large stimulus set of about 100 fractal patterns that were always presented in a predetermined order (Figure 4). After a 16 s delay period following presentation of a fractal sample, the monkeys were required to discriminate a matching choice pattern from a distractor pattern. Single-unit recordings made after overtraining on the task revealed that a population of anteroventral IT neurons exhibited correlated responses to sample patterns that had no geometrical similarity (Figure 4(a)) but were repeatedly presented in close proximity along the predetermined order of the stimulus set (Figure 4(b)). This suggests that, guided by the close temporal proximity of the presented visual stimuli, IT neurons form representations of the temporal relationships between stimuli.

Other evidence of associative representations among stimuli within the IT cortex comes from the findings of a later electrophysiological study which demonstrated that a population of anteroventral IT neurons learned to encode semantic-like relationships between pairs of visual stimuli that are geometrically unrelated and have no other relationship to one another, except that they are operationally associated in the experiments. Specifically, monkeys were trained on a pair-association (PA) task, which is a well-known neuropsychological test that is widely used to assess dysfunction of the medial temporal lobe memory system in humans (usually, the test is composed of tens of word pairs). In a visual PA task using pairs of Fourier descriptors, one member of a pair was presented as a cue stimulus at the beginning of a trial, then after a

been seen recently, but not its relative familiarity. (b) Responses of a familiarity neuron. Responses to novel stimuli were significantly larger than they were to familiar stimuli, regardless of whether the stimuli had been presented once or repeatedly during the session – that is, these responses signaled the relative familiarity of the stimulus, but not whether it had been seen recently. (c) Responses of a novelty neuron. Responses to the first presentations were significantly larger than they were to the repeated presentations of novel stimuli; maximum responses to familiar stimuli were comparable but the firing duration was greatly shortened – that is, this neuron responded best to the first presentation of novel stimuli, and its responses signaled both the relative familiarity of a stimulus and whether it had been seen recently. Reproduced from Xiang JZ and Brown MW (1998) Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology* 37: 657–676, with permission from Elsevier.

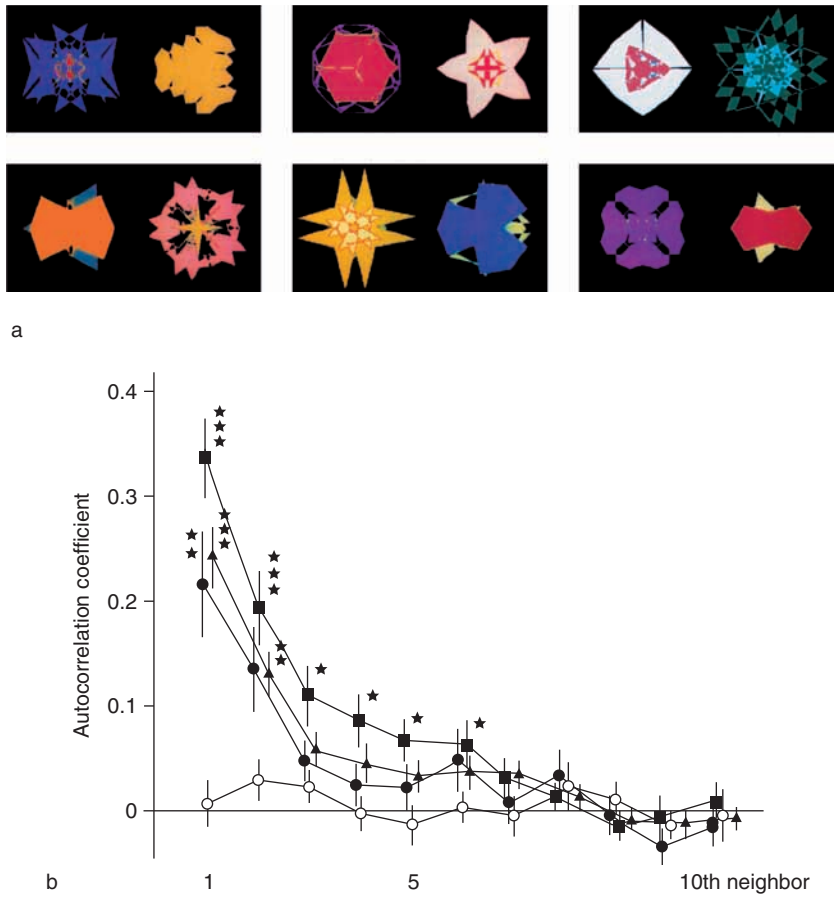

Figure 4 Correlated responses of macaque anteroventral IT neurons to disparate fractal patterns presented in close proximity along a predetermined sequence. (a) While monkeys performed a DMS task in which colored fractal patterns were presented in a predetermined order, single-unit recordings were made from the anteroventral part of the IT cortex. These six pairs of fractal patterns respectively elicited the two highest levels of delay activity in six different neurons. Note that there is no similarity between the geometric patterns of the two optimal stimuli. (b) Correlated responses were observed to disparate fractal patterns in close proximity along the serial position numbers in a learned sample set, but not in a new one. Of 206 IT neurons recorded while monkeys performed the task, 57 showed shape-selective delay activities. For each of those 57 neurons, autocorrelation coefficients for the responses to learned and new sample sets of fractal patterns were calculated and then plotted against the 'distance' between the serial position numbers used to calculate the coefficients. The correlation between neuronal responses was high for fractal patterns in close proximity within the learned set, but was nearly zero for both distantly positioned learned patterns and novel patterns. (●, ○) Average autocorrelation coefficients for the learned and new stimulus sets in 17 neurons that were successfully tested using both sets. (▲) Coefficients for the learned set in the 57 neurons. (■) Coefficients for the learned set in the 28 neurons for which the nearest-neighbor correlation for the serial position number was significant ($P < 0.05$) according to Kendall's test. Error bars are standard errors. Three vertical stars indicate $P < 0.001$; two vertical stars indicate $P < 0.01$; one star, $P < 0.05$ vs. the value for new stimulus list (Kolmogorov–Smirnov test). Reproduced from Miyashita Y (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335: 817–820, with permission from Nature.

following delay period the monkeys are required to choose the other member of the pair. Single-unit recordings from the anteroventral part of the IT cortex made while the monkeys performed the task revealed that a population of neurons showed comparable or correlated responses to two Fourier descriptors arbitrarily paired in the task (Figure 5); they clustered within 'hot spots' in the region. This correlated selectivity for paired visual stimuli develops while the monkey is learning the relationships between the stimuli. In addition, single-unit recordings also revealed that lesions in the perirhinal and entorhinal cortices, which comprise the interface between the IT cortex and the medial temporal lobe memory system, reduced correlated selectivity for trained paired stimuli in the anteroventral IT cortex (more precisely, anteroventral part of area TE; TEav), even though the neuronal responsiveness to individual stimuli used in the visual PA task was maintained. Moreover, the lesions also impaired formation of correlated selectivity for new pairs in the region. Thus, semantic-like association of one visual image with another, which can be regarded as an elemental unit for the elaborate semantic memory network, is thought

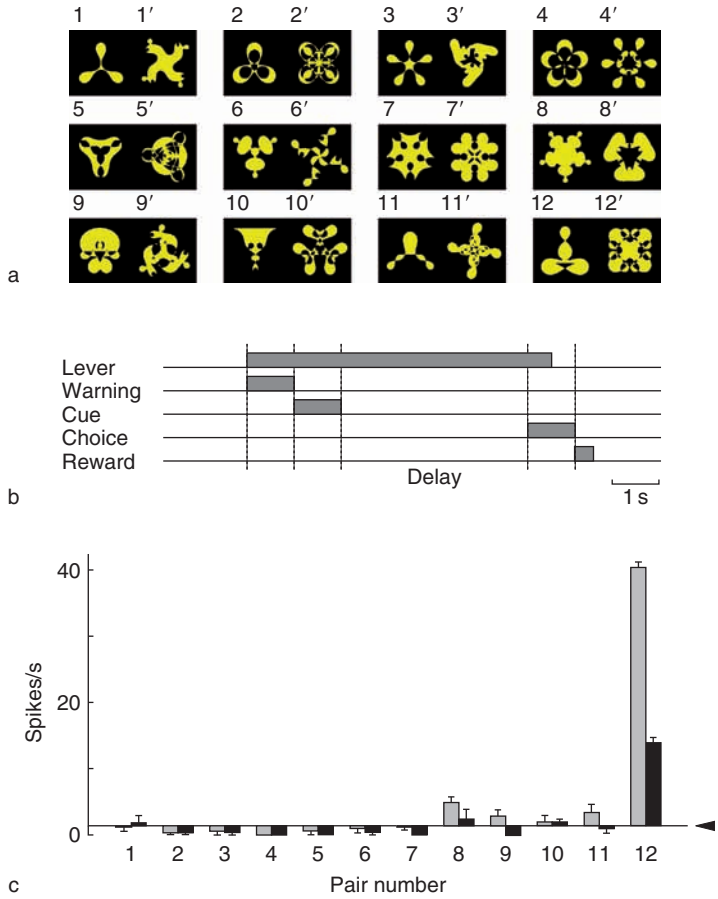

Figure 5 Associative coding of two disparate Fourier descriptors in the macaque anterior IT cortex. (a) Twelve pairs of Fourier descriptors (pictures 1 and 1′ to 12 and 12′) were used as stimuli in a visual PA task for monkeys. The monkeys learned to retrieve the other member of the pair when one member was presented as a cue stimulus. (b) Sequence of trial events in the visual PA task. Lever: lever press by the monkey to initiate a new trial. Warning: green square (1 s). Cue: one of 24 pictures serving as a cue stimulus (1 s). Delay: green square (4 s). Choice: a choice array of two stimuli, the paired associate of the cue and a distractor from a different pair. Reward: fruit juice reward for correctly selecting the associate. (c) Responses of an IT neuron to the 12 pairs of Fourier descriptors. Shown are mean discharge rates elicited by each picture during the cue presentation period relative to the spontaneous discharge rate (indicated by the arrowhead). Cue stimuli are labeled 'pair number' on the horizontal axis. For instance, in pair 1, the left bar indicates the response to picture 1, the right bar the response to picture 1′. The error bars indicate SEM. This neuron responded highly selectively to both pictures in pair 12, even though pictures 12 and 12′ had no apparent geometrical similarity (see panel (a)). Reproduced from Sakai K and Miyashita Y (1991) Neural organization for the long-term memory of paired associates. *Nature* 354: 152–155, with permission from Nature.

to be formed and represented in the neuronal responses of the IT cortex, probably supported by the medial temporal lobe memory system.

Molecular Biological Aspects of Memory Representation

The neuronal plasticity involved in formation of semantic-like associative representations in the anteroventral part of the IT cortex is thought to be supported by reorganization of the synaptic connections among neurons. Consistent with this idea are the findings of a study in which the use of tracer agents was combined with single-unit recording to reveal that divergent neuronal projections from the lateral portion of the anteroventral IT cortex (TEav) to the medial portion (area 36 in the perirhinal cortex; A36) differ depending upon the stimulus selectivity of the projecting TEav neurons. Specifically, when trained on a visual PA task, TEav neurons showing stimulus selectivity for computer-generated patterns (Fourier descriptors) have a greater tendency to project only to a hot spot of A36 neurons encoding semantic-like relationships between the paired patterns than do TEav neurons without selectivity for the trained patterns (these are probably responsive for other visual stimuli not used in the task). Such anatomical and physiological coupling – that is, less

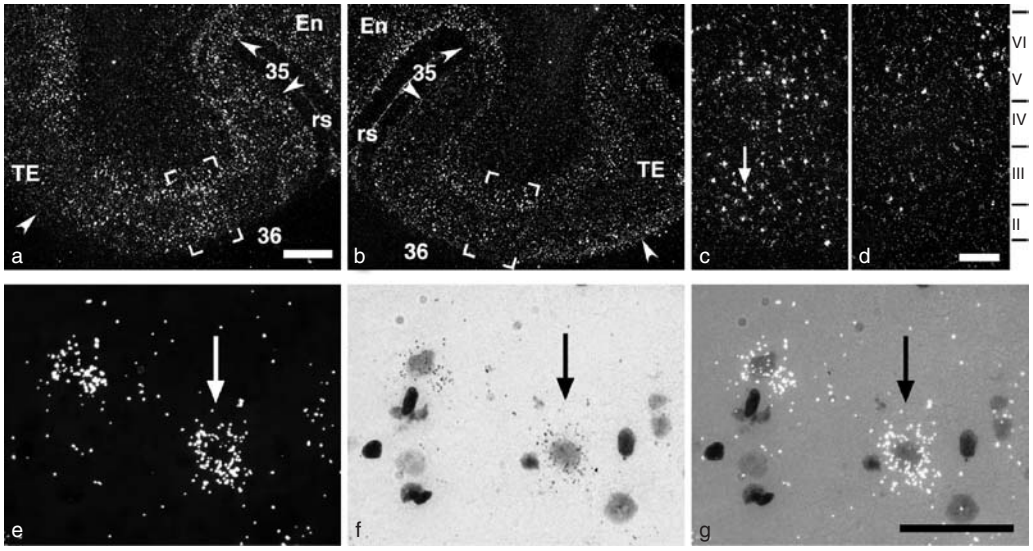

Figure 6 *In situ* hybridization of BDNF mRNA. (a–d) Distribution of BDNF mRNA in the IT gyrus of the PA and control hemispheres. BDNF mRNA accumulated in a patchlike cluster in area 36 (36) of the PA hemisphere (framed area in (a)), but not in area 36 of the control hemisphere (b). The framed areas in (a) and (b) are enlarged in (c) and (d), respectively. BDNF mRNA-positive cells were observed in layers V/VI and II/III of the PA hemisphere (c). En, entorhinal cortex; 35, area 35; TE, area TE; rs, rhinal sulcus. Arrowheads mark the boundaries between different cortical areas. (e–g) Cells positive for BDNF mRNA in layers II/III of the PA hemisphere. The cell indicated by an arrow in (c) is enlarged and shown in darkfield (e), brightfield (f), and brightfield with epi-illumination (g). Silver grains were concentrated around lightly Nissl-stained neuronal nuclei. Cortical layers in area 36 are indicated along the margin of (d). Scale bars = 1 mm (a, b), 250 μm (c, d), and 50 μm (e–g). Reproduced from Tokuyama W, Okuno H, Hashimoto T, et al. (2000) BDNF upregulation during declarative memory formation in monkey inferior temporal cortex. *Nature Neuroscience* 3: 1134–1142, with permission from Nature.

divergence coupled with stimulus selectivity – suggests that synaptic connections were pruned in the course of training on the visual PA task.

In a similar vein, several molecular biological studies demonstrated that while monkeys are learning semantic-like relationships in the PA task, expression of messenger ribonucleic acids (mRNAs) encoding proteins involved in synaptic plasticity is upregulated in the anteroventral IT cortex. In these studies, quantitation of mRNA using reverse transcription polymerase chain reaction (RT-PCR) was combined with a split-brain preparation in which the anterior commissure and the entire extent of the corpus callosum were transected. This enabled comparison of mRNA expression in a PA hemisphere and a control hemisphere within the same monkey, thereby eliminating genetic and cognitive variations between individuals. The experimental results revealed that mRNAs encoding brain-derived neurotrophic factor (BDNF) and the immediate-early gene *zif268* product (Zif268, also known as the transcriptional regulatory factor early growth response 1 (EGR1)) were upregulated in A36, where semantic-like associative memories were reflected within neuronal activity. Moreover, the results of *in situ* hybridization analyses demonstrated that cells positive for BDNF (Figure 6) and Zif268 mRNAs form a patchlike cluster in A36. These observations suggest that upregulation of BDNF and Zif268 mRNAs contributes to the neuronal reorganization that results in formation of semantic-like visual associations encoded within the neuronal activity in the anteroventral part of the macaque IT cortex. A likely scenario is that BDNF induces morphological changes in axonal and dendritic complexity, depending on neuronal activity, while Zif268 triggers a cascade of gene activation leading to synaptic reorganization.

Conclusion

The findings summarized in this article suggest that the prefrontal cortex works as a pivot or control center for short-term neuronal representations, while subsidiary domain-specific posterior cortices are employed to specify the physical features or object identity of the contents held in short-term memory. Visual recognition memory may be reflected by diminished neuronal activity within the perirhinal cortex and adjacent cortical areas, while semantic-like visual associative memories can be formed and represented within the neuronal activity of the anteroventral part of the IT cortex, probably supported by functions of the medial temporal lobe memory system. Viewed from a molecular biological standpoint, such semantic-like associations are thought to be formed through synaptic

reorganization, as mRNAs involved in synaptic plasticity are selectively upregulated in cortical areas where associations are encoded during semantic-like learning. Thus, the cerebral cortex, which processes and analyzes external sensory inputs, also represents and stores short-term and long-term memories in a domain-specific fashion – that is, memories are represented in the corresponding cerebral areas that processed the initial information from which the memories originated.

See also: BDNF in Synaptic Plasticity and Memory; Declarative Memory System: Anatomy; Goal-Directed Behavior Theories; Humans; Prefrontal Cortex: Structure and Anatomy; Recognition Memory; Semantic Memory; Short Term and Working Memory; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Desimone R (1996) Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences of the United States of America* 93: 13494–13499.

Eichenbaum HB (2003) Learning and memory: brain systems. In: Squire LR, Bloom FE, McConnell SK, et al. (eds.) *Fundamental Neuroscience,* 2nd edn., pp. 1299–1327. London: Academic Press.

Fukushima T, Hasegawa I, and Miyashita Y (2004) Prefrontal neuronal activity encodes spatial target representations sequentially updated after nonspatial target-shift cues. *Journal of Neurophysiology* 91: 1367–1380.

Fuster JM (1997) *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe,* 3rd edn. Philadelphia: Lippincott-Raven.

Fuster JM and Jervey JP (1981) Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science* 212: 952–955.

Goldman-Rakic PS (2004) Domain specificity in cognitive systems. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences,* 3rd edn., pp. 665–678. Cambridge, MA: The MIT Press.

Hasegawa I, Fukushima T, Ihara T, et al. (1998) Callosal window between prefrontal cortices: Cognitive interaction to retrieve long-term memory. *Science* 281: 814–818.

Higuchi S and Miyashita Y (1996) Formation of mnemonic neuronal responses to visual paired associates in inferotemporal cortex is impaired by perirhinal and entorhinal lesions. *Proceedings of the National Academy of Sciences of the United States of America* 93: 739–743.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24: 167–202.

Miyashita Y (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335: 817–820.

Miyashita Y (2004) Cognitive memory: Cellular and network machineries and their top-down control. *Science* 306: 435–440.

Miyashita Y and Chang HS (1981) Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* 331: 68–70.

Sakai K and Miyashita Y (1991) Neural organization for the long-term memory of paired associates. *Nature* 354: 152–155.

Squire LR and Kandel ER (1999) *Memory: From Mind to Molecules.* New York: Henry Holt and Company.

Squire LR and Zola-Morgan S (1991) The medial temporal lobe memory system. *Science* 253: 1380–1386.

Tokuyama W, Okuno H, Hashimoto T, et al. (2000) BDNF upregulation during declarative memory formation in monkey inferior temporal cortex. *Nature Neuroscience* 3: 1134–1142.

Xiang JZ and Brown MW (1998) Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology* 37: 657–676.

Memory: Computational Models

M W Howard, Syracuse University, Syracuse, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Computational models of memory are constrained from a number of directions. A physically accurate model of memory must simultaneously describe the details of behavioral performance and the mechanistic neural processes that support such performance. Such a model should simultaneously explain both changes in behavior due to manipulations of the neural substrate, such as lesions or pharmacological manipulations, and the detailed neurophysiology of the underlying structures, as measured by functional magnetic resonance imaging (fMRI), electroencephalogram/event-related potentials (EEG/ERP), and the activity of single neurons. We are far from such a comprehensive view of the neural basis of the types of memory typically studied by cognitive psychologists. However, some progress has been made toward this ideal in several domains. This work can be divided into content areas, distinguished by the types of memory and brain region supporting them, or the level of approach. This article describes the levels of approach that are possible in constructing computational models of memory and reviews progress in one content area – computational models of working memory.

Approaches to Modeling Memory

An endstate comprehensive neural model of cognition must be formulated at several levels of analysis, from the molecular to the very abstract. The question, however, is where to start in this vast endeavor. Presumably we can start from an accurate projection of the correct comprehensive model at any particular level of description and work at redescribing the model at adjacent levels of description. Appropriately, then, computational models of memory have been constructed at several levels of description.

Formal or mathematical models of memory are typically focused on describing behavioral performance in as near to complete detail as possible. Typically models in this tradition start from some simple but abstract low-level structure from which asymptotic or average behavior can be inferred. Although it is clearly desirable that this implementation be interpretable as some neural process, concerns of neural realism have not historically been considered diagnostic in this tradition. For instance, stimulus sampling theory, an early example of this approach developed in the 1950s, postulated that a stimulus was represented by a set of abstract stimulus elements. Learning a stimulus–response association consisted of binding a subset of the elements corresponding to a stimulus to a response. With repeated presentation of a stimulus, a new sample of elements is conditioned to the appropriate response, resulting in learning that gradually approaches an asymptote. An elaborated version of this simplified model was developed to describe changes in the strength of retention with different retention intervals and presentation schedules by postulating that the set of stimulus elements available to be conditioned fluctuated gradually over time. Although the abstract processing elements of stimulus sampling theory do not map simply on to any known neural process, or even brain region, closed-form analytic solutions can be derived for a number of important quantities that map on to behavioral observables. Mathematical models of memory have grown more elaborate since the 1950s, so that computer simulation is usually required to develop complete behavioral predictions.

To provide a more concrete idea of how such ideas might be actualized, Figure 1(a) illustrates some of the basic assumptions of stimulus sampling theory. A set of abstract elements, represented by filled or open circles, are available to represent a stimulus on each trial. The set of such elements used to represent the stimulus on a trial is represented by the set S. The remainder of the set of stimulus elements, which are unavailable, is referred to as S'. On the first learning trial, shown on the left, all of the elements in S are conditioned to the response. Conditioned elements are shown as large open circles, whereas unconditioned elements are shown as small filled circles. Over a delay, abstract processing elements shift from S to S', and vice versa, with transition probabilities defined by j and j'. These fluctuations lead to a forgetting of the conditioned response as conditioned elements used to describe the stimulus transition from the active set to the inactive set.

Connectionist approaches to memory modeling adopt a somewhat intermediate level of description. Rather than relying on completely abstract processing elements, connectionist models rely on sets of interconnected units that are often mapped explicitly on to specific brain regions. These units typically have a scalar activation at each time step that corresponds to the firing rate, perhaps averaged over a group of physical neurons and connections, perhaps modifiable, between units that correspond to synapses or at least to pathways between brain regions. Although we can find counterexamples to all such generalizations, connectionist models of memory focus less on a quantitative description of behavior than do

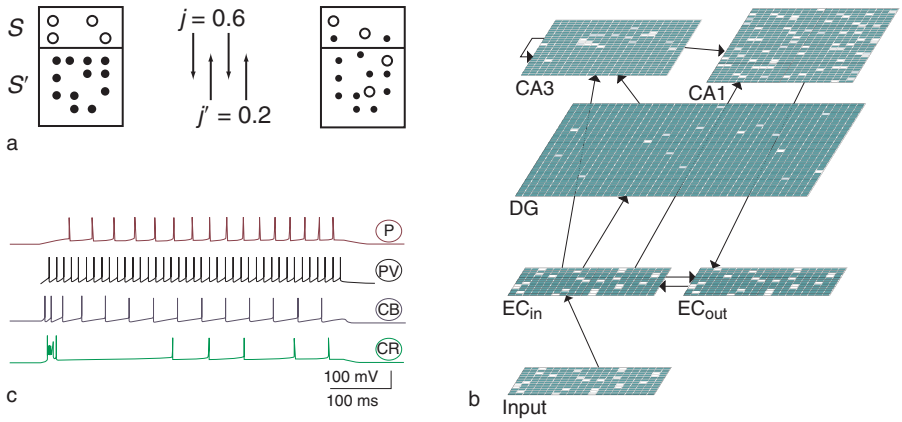

Figure 1 Approaches to computational models of memory: (a) mathematical model; (b) connectionist model; and (c) biophysical model. In (a) Estes's stimulus sampling model exemplifies mathematical models; they generate behavioral predictions from a small set of mathematical rules that need not map on to actual neural processes. In (b), Norman and O'Reilly's model of item recognition exemplifies connectionist models; they describe memory performance using a set of quasineural processing elements. In (c), Wang et al.'s model of working memory exemplifies biophysical models of memory; they attempt to provide a realistic description of the cellular processes that contribute to memory performance. CA3 and CA1, subfields of the hippocampus; DG, dentate gyrus; ECin, input layer of the entorhinal cortex; ECout, output-layer of the entorhinal cortex; P, pyramidal cells; PV, CV, CR, subpopulations of interneurons; SR, stimulus-response interaction; STS, short-term store. (b) Adapted from Norman KA and O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review* 110(4): 611–646.

mathematical models. These models are evaluated as much on the basis of their neural plausibility, given what is known about the anatomical circuits and physiology underlying the regions. These constraints can come from correctly describing the qualitative effects of physiological manipulations, such as brain lesion or pharmacological manipulations.

Figure 1(b) shows an example of a connectionist model of item recognition memory, a memory task in which the subject is given a probe item and asked whether it was presented as part of a study episode. In this model, a set of interconnected abstract processing elements are identified with brain regions. In this case, an input pattern is presented to a set of processing elements identified with input layers of the entorhinal cortex (EC). These processing elements are connected to other elements identified with other brain regions – in this case the dentate gyrus (DG) and the subfields of the hippocampus (CA3 and CA1). The model's memory for a recognition probe can be assessed by examining the degree to which the pattern in the output layer of the EC (EC_{out}) resembles the input pattern presented to the input layer (EC_{in}) when the pattern is repeated as a probe item after learning.

At a yet more detailed physical level, biophysical models of memory attempt to describe memory processes using models of neural activity that are not merely plausible but as close to physically accurate as possible. Typically model neurons in these simulations fire individual spikes, with a set of realistic conductances taken from experimental data. The goal of biophysical models of memory is typically not to model behavioral data *per se* but rather to show how a mechanism believed to relate to some cognitive mechanism supporting some type of memory performance could be physically implemented. These models are typically evaluated in terms of their ability to accurately describe the currents and firing rates observed in neurons during performance of a task believed to be related to the cognitive processes that are under study.

Figure 1(c) shows a biophysical model designed to capture aspects of working memory maintenance, a mechanism believed to be required to retain information in attention for a short time. In this simulation, there are several subtypes of simulated neurons that are interconnected to one another. Figure 1(c) shows the response of each of the types of model neurons to a transient injected current. Simulated pyramidal cells (P) are interconnected with three types of interneurons. The pyramidal cells are modeled as three-compartment neurons – that is, at the subcellular level. The pyramidal cells and interneuon populations differ in their intrinsic properties, synaptic properties, and network connectivity. In this simulation, the traces labeled PV, CB, and CR correspond to three different subpopulations of interneurons. This model is primarily evaluated on the basis of its ability to describe the detailed behavior of neurons of the different classes observed in the prefrontal cortex during the maintenance phase of a working memory task in which a monkey must remember a spatial location.

These three examples illustrate a number of differences among mathematical, connectionist, and biophysical approaches. They can be seen as lying on a

continuum in terms of several variables. The spatial scale over which these models are specified obviously varies, from the behavior of the whole organism in the case of mathematical psychological models to sets of cells identified with brain regions in the case of connectionist models to subcellular processes in the case of the biophysical model. These approaches also differ with respect to the temporal scale over which the models are specified. Whereas mathematical models often treat memory in units of stimulus presentations, which may be several seconds, connectionist models typically measure time in tenths of seconds. In contrast, biophysical models often describe somatic voltage at the submillisecond timescale.

In addition to the (often considerable) difficulty associated with getting a model to work within a particular level of description, there is an additional level of difficulty in constructing models that make contact across levels of description. We use the term memory to encompass a very broad range of phenomena, including everything from an aplysia learning to retract its gill to a person consciously recollecting the sounds and sights of a specific event given, say, a familiar odor as a cue. These phenomena, and others, are typically studied separately, with the range of phenomena under examination organized by a putative memory system, or perhaps the specific brain region (or regions) believed to be crucial in a particular task. A great many such problems that have been addressed using computational models of memory at one or more levels of investigation. Some of the more notable topics that have benefited from computational models of memory include categorization, delay and trace conditioning, cognitive control, associative learning, and hippocampal-dependent memory. A thorough survey of these diverse problems is impossible given the space constraints of this article. Instead, we restrict the content of our discussion severely and discuss computational models of working memory maintenance. This area is perhaps particularly well suited to giving the reader a flavor of the different approaches that are possible using computational models of memory. The models that have been developed to study working memory are also perhaps notable in that a number of attempts have been made to bridge multiple levels of analysis.

Models of Working Memory

To cognitive psychologists, the term working memory is used to describe the set of cognitive processes required to store information actively in attention for a short period of time and make decisions based on that information. Ideas about working memory have evolved from previous conceptions of short-term memory. These models pay differing amounts of attention to the necessity of, on the one hand, storing information for a short period of time and, on the other, controlling what information enters into short-term storage and how it might be manipulated.

The 1960s and 1970s saw the development of very successful mathematical models of short-term memory. The most prominent of these was the Atkinson and Shiffrin (AS) buffer model (**Figure 2(a)**). In this model, attentional control processes gate information into and out of a labile, but limited-capacity, short-term store (STS). Activity in short-term store is closely associated in this view with the process of rehearsal, in which subjects actively and consciously repeat to-be-remembered materials to themselves. For instance, suppose that someone tells you a phone number that you must remember for a few seconds before you can write it down. A common strategy in this situation is to repeat the digits to yourself to keep them fresh in your memory. Subjects in human verbal learning studies frequently adopt a rehearsal strategy to help them remember the study items. Formally, STS was conceived of as a small number of slots, typically two to four, into which study items could be copied. There are a small integral number of slots, and an item can either be present or absent in the buffer – no partial activation was allowed. The relatively small capacity of STS means that decisions must be made about what information to prioritize. In the AS model, these decisions are assumed to be under conscious strategic control, although the rules for deciding what information should be maintained or discarded were not explicitly modeled but rather assumed to reflect specific strategies appropriate to a particular task. Finally, STS serves as a gateway for information to be stored in long-term store (LTS). The longer an item remains active in STS, the better the chances that it will be successfully stored and later recovered from LTS.

The AS model and its successors were extraordinarily successful at describing human behavioral data in a relatively wide variety of experimental paradigms, including a continuous paired-associates task and the free recall task, in which subjects are presented with a list of words and then instructed to recall them in any order they think of them. As a consequence, ideas about short-term memory from the AS model, including the notion of a short-term store that retains an integral number of stimuli in a discrete fashion, have persisted and been implemented in a variety of biophysical and connectionist models of working memory.

Connectionist models of working memory have emphasized the importance of maintenance of information as a subprocess in the service of cognitive

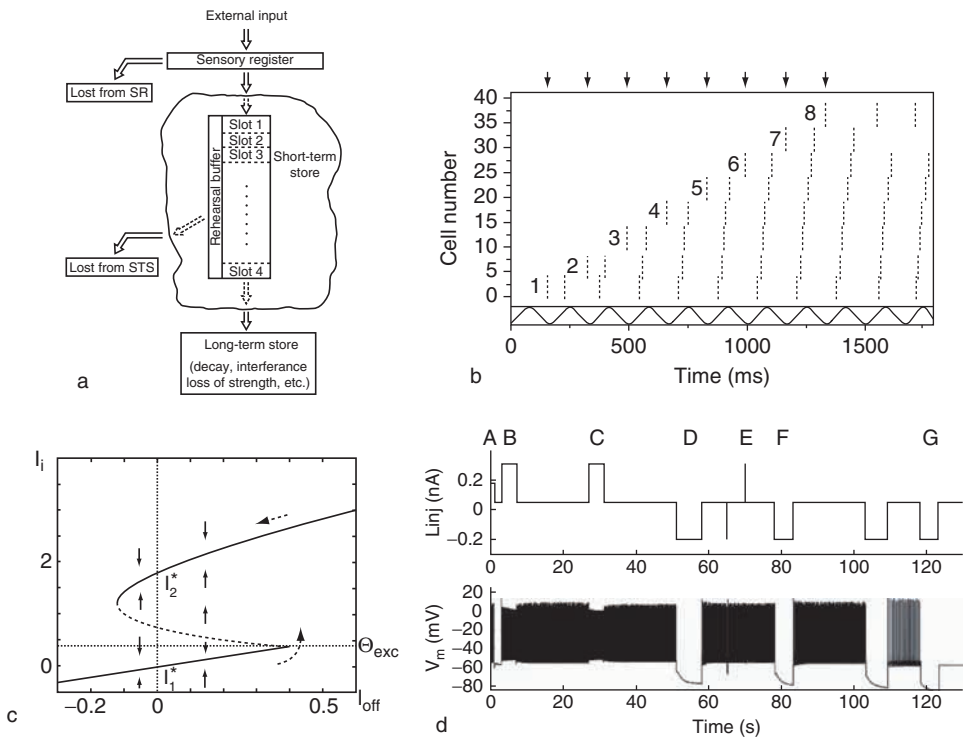

Figure 2 Approaches to models of working memory: (a) buffer model; (b) Lisman–Idiart–Jensen (LIJ) model; (c) bifurcation plot of dynamics in a multistable biophysical model; (d) biophysical simulation of multistability observed in layer V cells in the entorhinal cortex. In (a), incoming information is stored in one of a discrete number of working memory buffers; information is either present or absent in these buffer slots. In (b), the LIJ model is a biophysical instantiation of the buffer model in which neurons representing different items coactive in the buffer occupy different subcycles of an ongoing theta oscillation. A raster plot from a biophysical simulation is shown. Neurons corresponding to each of several items are presented, and patterns of activated items are either in the buffer or not. For instance, item 7 is initially present in the buffer, but then falls out when item 8 is presented. In (c), the multistable biophysical model exploits network properties to implement bistability. The network consists of cell assemblies with excitatory recurrent connections and inhibitory lateral connections. The *y* axis plots the firing rate of one of the assemblies and the *x* axis plots its external input current. The stable states of the dynamics are shown with solid lines. When the input current exceeds a threshold (approximately 0.4 in the figure), the activity of the assembly jumps from the value on the lower line to the value on the upper line (curved arrow). After this takes place, the cell remains active even after the external current is removed (dashed line). A sufficiently large inhibitory current can cause the assembly to return to rest (left side of figure). In (d), under appropriate conditions, layer V cells respond to a depolarizing input with a sustained nonzero firing rate. As additional inputs are presented, the cell adopts a new stable firing rate (B and C). This property does not depend on recurrent connections because it was observed under conditions of synaptic isolation. In the biophysical simulation shown here, multistability is a consequence of a fixed-point attractor, the location of which can be changed by changes in the number of activated channels, a process which is triggered by sufficiently large changes in Ca^{2+} concentrations. Note the long time scale.

control. For instance in the Stroop task, subjects are presented with a color name in a colored font that may be inconsistent with the named color; for instance, the word red written in blue. In this task, subjects are instructed to read the word rather than report its actual color. Subjects are much slower to read the name, and they make more errors, when the physical color is inconsistent with the written color name (e.g., the word red written in blue) than when it is consistent (e.g., the word red written in red). Neuropsychological data indicate that the ability to inhibit the tendency to say the name of the physical color depends on an intact prefrontal cortex. A number of closely related connectionist models of working memory describe the ability to control the inappropriate response by means of active attentional maintenance of a rule representation (i.e., say-the-name-of-the-word). This rule representation modulates, or gates, the appropriate behavioral response to the physical stimulus. For the rule representation to remain active, it must be actively maintained in a short-term buffer in prefrontal cortex. In these connectionist models, then, the concept of active maintenance is retained, but the emphasis is more on the maintenance of rules than on information to be committed to long-term memory.

Biophysical models of working memory have focused on the mechanisms of active maintenance rather than on either the role of this maintenance in executive processing or its potential role in transferring

information to long-term memory. A variety of biophysical mechanisms have been proposed for working memory maintenance in at least two distinct regions. These models can be roughly divided into two classes: those that rely on network properties to allow persistent firing and those that rely on intrinsic properties of individual neurons that allow them to maintain firing over long durations. A goal of many of these efforts has been to implement bistability – a population of neurons that have two stable states, one with a high sustained firing rate and one with a much lower firing rate. In this sense, these models have retained the idea of all-or-none activation present as early as the AS buffer model.

Network models rely on the idea that a population of neurons is divided into cell assemblies with strong exitatory connections within an assembly and inhibitory connections between neurons from different cell assemblies. If the parameters of the model are chosen appropriately, the network will have the property that, if an assembly is given a sufficiently large external input, the recurrent connections sustain a high level of firing that can persist even when the external input is removed. The assembly will then remain active until a sufficiently large inhibitory input, either an inhibitory external input or an excitatory input to another assembly, causes it to return to the rest state (Figure 2(c)). Network models are widely believed to describe the basis of sustained activity observed in neurons in the prefrontal cortex. The anatomical organization of the dorsolateral prefrontal cortex, with interconnected pyramidal cells and a spatially distributed network of interneurons, suggests that it may be organized into columns that serve as cell assemblies.

Cellular mechanisms for bistability that do not depend critically on recurrent connections have also been proposed. The Lisman–Idiart–Jensen (LIJ) model proposes that large afterdepolarizations observed in some cells may be sufficient to cause a cell to fire again. If an initial spike can cause a second spike, then the cell will continue firing for a macroscopic time interval. In order to allow for multiple items to be maintained simultaneously, the LIJ model proposes that brain oscillations play a role in organizing the firing of patterns of spikes corresponding to separate items that are activated in memory. In the LIJ model, theta oscillations, large 4–12 Hz oscillations in the local field potential that are especially prominent in the hippocampus, serve to provide a temporal frame for the item representations such that spikes corresponding to each of the items active in working memory fire once per theta cycle. The patterns corresponding to different items are kept separate from one another because the patterns are constrained to fire on different cycles of gamma (40–80 Hz) oscillations (see Figure 2(b)).

Using nested gamma and theta oscillations, the LIJ model has been used to model most of the features of the AS model in a biophysically plausible way. For instance, the biophysical model allows patterns of activity to drop out of the buffer in an all-or-none fashion (note the fate of pattern 7 in Figure 2(b)). Inclusion of synaptic plasticity via *N*-methyl-D-aspartate (NMDA) channels has been shown to lead to long-lasting changes in the strength of the connections between items active in short-term memory, analogous to the transfer from STS to LTS, critical in the AS model. The major application of the LIJ model to directly describing cognitive data has been the Sternberg task. In the Sternberg task, the participant studies a list consisting of no more than a handful of stimuli. At the end of the list, the participant is given a probe item and must decide if the probe was part of the list or not. The key finding is that the amount of time it takes the subject to decide if the item was on the list or not increases linearly with the number of items on the list, at least for small list lengths. The LIJ explains this linear slope because of the strict temporal ordering of the items active in memory. In addition to the fact that the slope, typically around 40 ms/item, is on the order of what we would expect if we had to wait for a gamma cycle for the next item to come up, variants of the LIJ model have been shown to describe human performance in the Sternberg task in remarkable detail. Moreover, since the LIJ was first proposed, a wealth of data implicating theta and gamma oscillations in working memory and the Sternberg task in particular have been discovered. The LIJ provides a biophysical implementation of the assumptions of the AS model and also of Sternberg's ideas about how subjects scan through a short list of items to reach a decision on the probe. As such, it provides a remarkable example of a model that bridges multiple levels of description.

More elaborate forms of persistent neural activity that do not depend on interconnections between cells have been observed in the EC. Cells in layer V of the EC have been shown to be able to produce extremely long-lasting persistent firing in the absence of synaptic input. Moreover, these cells are able to gradually change their firing to reflect the summation of their inputs. That is, a first stimulus might cause the cell to fire steadily at a given rate. A second stimulus would cause the neuron to jump to a new steady firing rate. The mechanisms underlying this remarkable ability are a topic of intense investigation. It is known that persistent firing depends on a Ca^{2+}-sensitive nonspecific cation (CAN) channel that remains open for long periods. The firing rates a cell can sustain are

apparently graded, rather than discrete, and are observed even when intracellular stores of Ca^{2+} are depleted. One theory of how this remarkable behavior is possible is that at any time the current dynamics are controlled by a fixed-point attractor that is insensitive to small perturbations. However, sufficiently large (in terms of amplitude and/or duration) changes in the internal Ca^{2+} concentration cause the location of this fixed state to gradually change. The exact intracellular mechanisms of how this is accomplished are not known, but a plausible multicompartment model implementing these ideas is able to simulate the behavior of multistable integrator cells in remarkable detail (see **Figure 2(d)**).

Intrinsic persistent activity is an extremely important property for a general working memory system to have. If the ability to maintain information in working memory requires a preexisting synaptic network, as in the recurrent networks supporting bistable cell assemblies such as those shown in **Figure 2(c)**, then, because cell assemblies must support one another, the number of patterns that can be maintained (at different times) must be less than the number of neurons. In contrast, if neurons can maintain their own firing for arbitrarily long periods of time without network connections, then the number of distinguishable patterns that can be maintained at different times can even exceed the number of neurons. This property is extremely useful for a working memory buffer in the spirit of the AS model that should be able to store at least all the words in the English language.

The other unique property of multistable cells is their ability to support a gradual decay of information rather than the precipitous all-or-none dropout rule of the AS model that is implemented in the bistable biophysical models. This gradual decay of information in working memory is a characteristic of a more recent mathematical model of working memory maintenance and storage of information into long-term memory – the temporal context model (TCM). Much like the AS model, TCM was formulated as a mathematical model of experiments in which participants learn and recall lists of words. As such, it has predicted and explained several findings that would be counterintuitive if the AS model were a literally accurate description of maintenance and storage of recently presented information. TCM differs from the AS model in two respects: (1) rather than decaying in an all-or-none fashion, information decays gracefully from the working memory store in TCM, which is referred to as a temporal context vector, allowing TCM to describe gradual forgetting over much longer time scales than the AS model can, a property that is consistent with some behavioral data; (2) in TCM the input to the working memory store caused by an item is not fixed over time, as in the AS model, but can change to reflect the temporal context in which the item was presented. This ability to bind items to the temporal contexts they are encoded in turns out to have several important implications.

Recently, proponents of TCM made a bridging hypothesis to connect the abstract mathematical model with the actual neural substrate. They proposed that the temporal context vector should correspond to the pattern of activity across cells in extra-hippocampal medial temporal lobe regions, in particular the EC. They showed that a population of multistable cells, such as those illustrated in **Figure 2(d)**, if subject to cortical gain control, could implement the equation for the change of temporal context in such a way as to allow the gradual decay of information in the temporal context vector. Continuing the mapping between TCM and the brain, the function of the hippocampus proper is to enable items to be bound to and subsequently to recover the temporal context or state of activation in the EC, that obtained when they were studied. This assumption made it possible for the model to account for the effect of hippocampal lesions on learning tasks in rats.

The approach of models exploiting the bistability property introduced by the AS model and the multistability property exploited by TCM differ considerably in their behavioral predictions and the range of potential implementations. It might be nice if one could declare one approach correct and the other incorrect. A more likely outcome is that both reflect some insight into how working memory maintenance behaves in different circumstances. One means of reconciliation is to note that the connectionist view of working memory maintenance as something that retains goal states in prefrontal cortex does not overlap particularly strongly with the view of a gradually changing representation of recent experience in extra-hippocampal medial temporal lobe cortical areas. Perhaps both are accurate descriptions of related but distinct phenomena. In this case, bistable maintenance of goal states would not typically have much to do with how we remember recent events, whereas a gradually changing state of temporal context would not be much use in performing the Stroop task.

Computational models of memory can be roughly divided into three classes that differ in their level of realism and commitment to explaining cognitive versus neural data: formal or mathematical models, connectionist models, and biophysical models. Although each of these approaches is valid in its own right, the fact that recent models are attempting to bridge these levels is a promising development. Computational models of working memory at all three levels have been reviewed, but it should be kept

in mind that similar developments are taking place in the study of many types of memory and many different brain regions.

See also: Episodic Memory; Event-Related Potentials (ERPs); Executive Function and Higher-Order Cognition: Computational Models; Hippocampus and Neural Representations; Hippocampus: Computational Models; Humans; Recognition Memory; Short Term and Working Memory.

Further Reading

Atkinson RC and Shiffrin RM (1968) Human memory: A proposed system and its control processes. In: Spence KW and Spence JT (eds.) *The Psychology of Learning and Motivation,* vol. 2, pp. 89–105. New York: Academic Press.

Botvinick MM, Braver TS, Barch DM, Carter CS, and Cohen JD (2001) Conflict monitoring and cognitive control. *Psychological Review* 108(3): 624–652.

Davelaar EJ, Goshen-Gottstein Y, Ashkenazi A, Haarmann HJ, and Usher M (2005) The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review* 112(1): 3–42.

Durstewitz D and Seamans JK (2006) Beyond bistability: Biophysics and temporal dynamics of working memory. *Neuroscience* 139(1): 119–133.

Durstewitz D, Seamans JK, and Sejnowski TJ (2000) Neurocomputational models of working memory. *Nature Neuroscience* 3: 1184–1191.

Estes WK (1955) Statistical theory of spontaneous recovery and regression. *Psychological Review* 62: 145–154.

Fransén E, Tahvildari B, Egorov AV, Hasselmo ME, and Alonso AA (2006) Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron* 49(5): 735–746.

Howard MW, Fotedar MS, Datey AV, and Hasselmo ME (2005) The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review* 112(1): 75–116.

Howard MW and Kahana MJ (2002) A distributed representation of temporal context. *Journal of Mathematical Psychology* 46(3): 269–299.

Jensen O and Lisman JE (1998) An oscillatory short-term memory buffer model can account for data on the Sternberg task. *Journal of Neuroscience* 18: 10688–10699.

Jensen O and Lisman JE (2005) Hippocampal sequence-encoding driven by a cortical multiitem working memory buffer. *Trends in Neuroscience* 28(2): 67–72.

Lisman JE and Idiart MA (1995) Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science* 267: 1512–1515.

Miller P and Wang XJ (2006) Power-law neuronal fluctuations in a recurrent network model of parametric working memory. *Journal of Neurophysiology* 95(2): 1099–1114.

Norman KA and O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review* 110(4): 611–646.

Raaijmakers JGW and Shiffrin RM (1980) SAM: A theory of probabilistic search of associative memory. In: Bower GH (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory,* vol. 14, pp. 207–262. New York: Academic Press.

Rougier NP, Noelle DC, Braver TS, Cohen JD, and O'Reilly RC (2005) Prefrontal cortex and flexible cognitive control: rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America* 102(20): 7338–7343.

Wang XJ, Tegnér J, Constantinidis C, and Goldman-Rakic PS (2004) Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proceedings of the National Academy of Sciences of the United States of America* 101(5): 1368–1373.

Memory: Genetic Approaches

J Z Tsien, Boston University, Boston, MA, USA

© 2009 Published by Elsevier Ltd.

Introduction

It is widely believed that memory is formed and stored in the brain by modification of the physiological or physical features of synapses, the specialized junctions where neurons connect with each other, a notion that can be traced back to Santiago Ramon y Cajal at the end of the nineteenth century. The foundations for exploring the molecular and genetic basis of learning and memory were further laid in 1949, when Canadian psychologist Donald O Hebb came up with what is known as Hebb's rule. In essence, he postulated that a memory is produced by coincident neural activity; when two connected nerve cells are active simultaneously, the strength of their synaptic connection increases; this confers a basis for the persistence of memory.

To understand whether the brain actually uses 'coincidence-detection' logic to modify synaptic efficacy and store information, researchers have focused on the hippocampus, a region shown to be essential for memory formation, and have found that high-frequency electrical stimulation can produce a persistent increase in synaptic efficacy, a process termed long-term potentiation (LTP). Although LTP was initially discovered in the hippocampus by Bliss and Limo, subsequent studies have shown that it can be elicited in many brain regions. As a result, LTP becomes a major experimental model for the investigation of molecular and cellular mechanisms of synaptic plasticity.

A series of investigations have revealed that the *N*-methyl-D-aspartate (NMDA) receptor is a crucial molecular switch for the induction of LTP. The activation of the NMDA receptor requires both the release of glutamate from the presynaptic neurons and the deplorization of the postsynaptic cells (**Figure 1(a)**). Thus, the NMDA receptor is an ideal to implement the Hebb's rule at the synaptic level. The opening of the NMDA receptor channel allows the influx of calcium, which then activates its downstream signaling molecules such as Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII), PKC, PKA, and MAP kinase. Through a yet less defined process, it eventually leads to long-term modifications in synaptic transmission efficacy.

Hebb's Rule, NMDA Receptor Coincidence-Detection, and Memory Formation

To examine the role of those signaling molecules in the actual memory process, researchers have used two major approaches, namely pharmacological inhibition and genetic manipulation. For example, pharmacological investigations in the late 1980s and early 1990s reported that infusion of NMDA receptor antagonists into brain ventricles resulted in animals' poor performance in the Morris water maze. At the first glimpse this seemed to provide evidence for the role of hippocampal LTP in memory formation. However, careful analyses revealed that poor performances in the water maze tests could be a result of drug-induced side effects including sensorimotor disturbances. In addition, chronic APV exposure to neurons has been shown to upregulate NMDA receptors including the NR2B subunit, thus raising the question as to what chronic APV infusion in the brain is actually doing to NMDA receptor sensitivity and function.

The complications from pharmacological approach have compelled the researchers to use gene knockout methods to investigate the role of candidate molecules in learning and memory. However, conventional gene knockout, a method which deletes a gene of interest in every cell of the body starting from embryonic day 1, has its own limitations. For example, the knockout of the NMDA receptor 1 subunit (NR1), a key subunit essential for the channel activity, led to the mice dying within 12 h of birth. In many other cases, knockout often leads to developmental abnormality and altered brain structures. Therefore, the lack of tissue and temporal specificity in conventional gene knockout has limited its application in analyzing gene function in cognition and behavior.

To overcome those limitations, conditional gene knockout techniques were developed by Tsien and his colleagues in 1996. This second-generation gene knockout employs Cre/loxP-mediated recombination system to restrict deletion of any gene of interest to specific cells, tissues, or organs in mice. By applying its powerful technique, researchers knocked out the NR1 gene specifically in the CA1 region of the hippocampus in young adult mice. Analyses revealed that the knockout of the NMDA receptor prevented all forms of NMDA receptor-mediated synaptic plasticity in the CA1 region, and those CA1-specific NR1 knockout mice exhibited profound deficits in

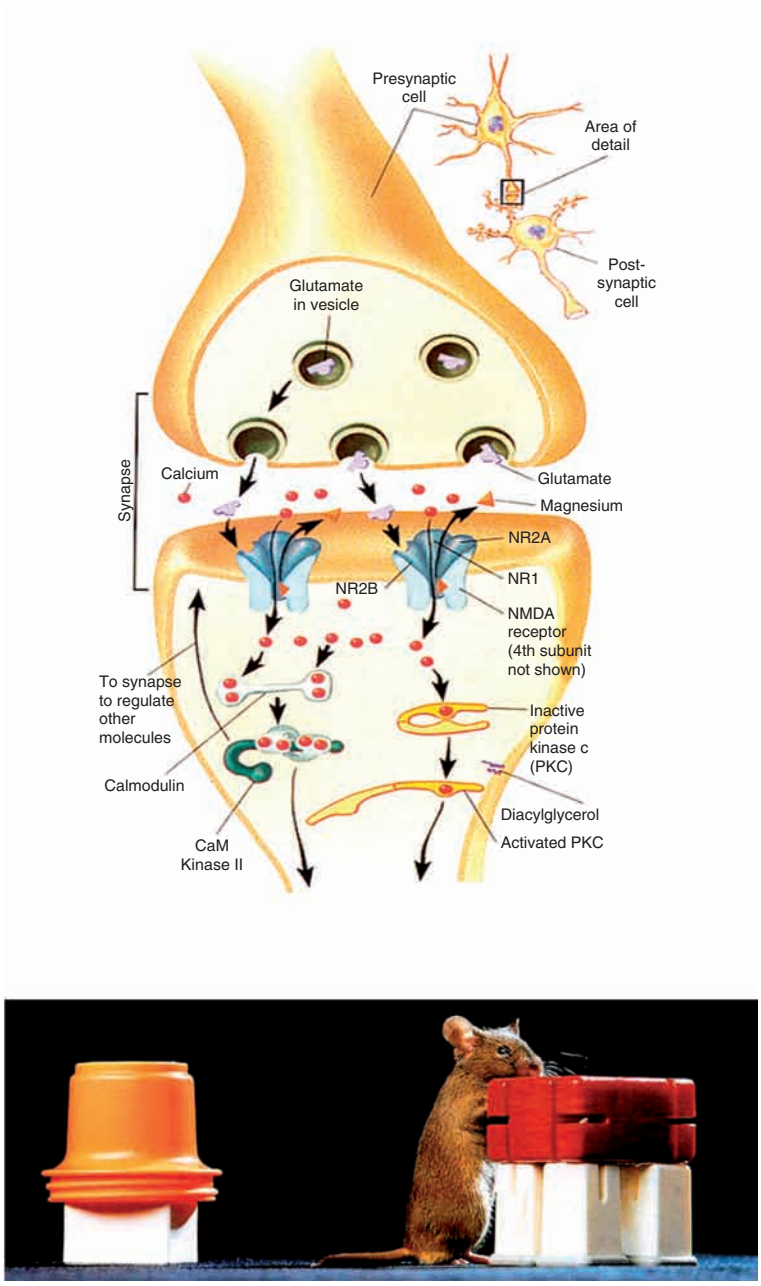

Figure 1 The NMDA receptor as a cellular coincidence-detector for memory formation. (a) Sketch of brain cells linked in vast networks, with zoom in at synapses showing the NMDA receptor, NR2B subunits, intracellular signaling molecules, etc. Activation of the NMDA receptor requires the binding of glutamate (from presynaptic terminals) and the removal of Mg from its pore (by postsynaptic depolization). (b) Creation of smart mice, Doogie, has stringently validated the Hebb's rule. Figure adapted from Tsien JZ (2000) Building a brainier mouse. *Scientific American* 282: 62–68.

a variety of spatial and nonspatial memory tests including the hidden-platform water maze, novel object recognition, contextual fear conditions and social transmission of food-preference, etc. Additional knockout of the NR1 in the CA3 region also shows that it can lead to memory deficits, although less pronounced in CA3-specific knockout than that of CA1-specific knockout. Thus, those conditional gene knockout experiments strongly suggest the important role of hippocampal NMDA receptors in controlling the formation of hippocampal-dependent learning and memory.

Interestingly, daily exposures of these CA1-specific NR1 mice to enriched environments can significantly alleviate some of those learning and memory deficits. Later studies showed that daily enrichment produces significant changes in the expression of several hundred genes, as well as an increase in dendritic

spines in the brains of those mutant mice, suggesting the beneficial effects of enriched environments. However, similar enrichment seems to be incapable of rescuing the memory deficits in the forebrain-specific NMDA receptor knockout, suggesting that the enrichment may exert its effects through a cortical NMDA receptor-mediated mechanism.

While the conditional gene knockout experiments are supportive of a role of the NMDA receptors in memory process, those experiments did not necessarily represent the most stringent evidence for linking the synaptic coincidence-detection feature of the NMDA receptor to memory formation. The knockout of the NR1 gene removes the physical channel structure from the synapses and disables all function of the NMDA receptor, thus resembling the sledge-hammer approach.

In 1999, researchers were able to fine-tune the coincidence-detection function of the NMDA receptor by manipulating the NR2B subunit of the NMDA receptor channel complex and creating the smart transgenic mice, nicknamed 'Doogie' (Figure 1(b)). Those NR2B transgenic mice have exhibited superior learn and memory functions in at least six different behavioral tests. Therefore, a series of stringent genetic experiments have unequivocally demonstrated that NMDA receptor-mediated coincidence-detection plays an essential role in memory formation. In addition, genetic manipulations of NMDA receptor-downstream molecules, such as CaMKII, PKA and MAP kinases, have also affected learning and memory performances.

Molecular Mechanisms Underlying Memory Consolidation and Storage

Memory process can be generally divided into four distinct stages: learning, consolidation, storage, and retrieval. Studies of amnesiac patients and brain lesions in experimental animals have revealed that memory consolidation requires the hippocampus to continue its engagement for an additional period of time after learning – estimated to be days and week(s) in rodents and months to year(s) in primates. The consolidation process is believed to be crucial for converting short-term labile memory into long-term stable memory in the brain.

It is generally agreed that long-term memory is ultimately expressed in the form of structural and physiological changes in synapses. Until recently, the dominant view in the field was that this structural modification of synapses is the result of a learning-triggered molecular cascade consisting of receptor activation, transient changes in protein phosphorylation levels, new protein synthesis, and new gene expression. However, a variety of emerging evidence suggests that this 'single cascade hypothesis' cannot fully account for the formation and consolidation of long-term memory in the mammalian brain. First, the timescales for a molecular cascade, which typically occurs on the timescale of minutes to hours, do not fit with the timescale of memory consolidation, which operates on the timescale of week(s) in rodents and years in humans. Second, although protein synthesis inhibitors seem to produce long-term memory deficits (often tested within a day or days of training), spontaneous recovery or reminder-induced recovery of memory over the time course of a week or weeks has been reported in animals initially thought to be amnesic. Finally, individual synaptic receptor and structural protein molecules have relatively short lifetimes, in brain as in other organs. For example, a recent study shows that synaptic NMDA receptors are degraded within 5 days in brains of freely behaving mice. Such turnover rates of synaptic receptors and their associated signaling and structural proteins would inevitably lead to deleterious drifts in synaptic efficacy, thereby destabilizing the stored memory traces, in the absence of restorative mechanisms. Thus, the observed metabolic turnovers of synaptic proteins raise a fundamental question: how can the structural changes at synapses, as envisioned by the Hebb's biochemical cascade, sustain memories over long periods of time in spite of such dynamic turnover?

An alternative mechanism to account for both the molecular and system-level consolidation of long-term memory comes from the application of third-generation gene knockout techniques. In 2000, researchers developed inducible, reversible, and CA1-region specific NR1 knockout mice (iCA1-KO) by combining the tTA and Cre/loxP recombination system in an effort to test whether the NMDA receptor activity is required during the memory consolidation stage. Surprisingly, the deletion of the CA1 NMDA receptor in the post-learning period produced severe impairment in memory consolidation. It further revealed that the first two posttraining weeks constitute the critical time-window during which the NMDA receptor reactivations are required. Since the NMDA receptor in the CA1 region is known to be not involved in basal synaptic transmission, those experiments have led the researchers to postulate that NMDA-reactivation during the postlearning period is to trigger the repeated synaptic reentry–reinforcement (SRR) process to consolidate the newly acquired memory trace at the synaptic level (Figure 2).

Consistent with the requirement of NMDA receptor-reactivation for memory consolidation, it has been found that memory consolidation is also dependent

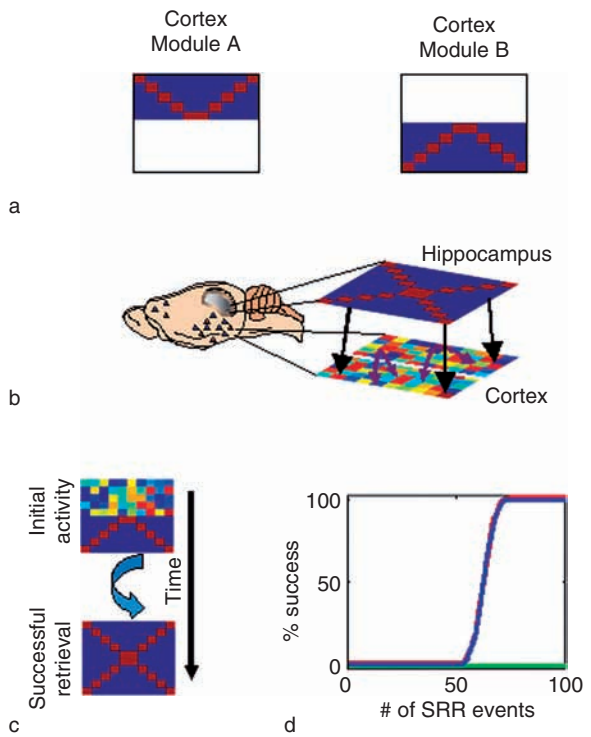

Figure 2 Synaptic reentry–reinforcement (SRR) process for memory consolidation and storage. (a) Reactivations of the NMDA receptor during the consolidation period trigger the repeated synaptic reentry–reinforcements of synaptic memory traces initiated by original learning. (b) Hippocampus also serve as the coincidence-regenerator for coactivating the cortical neurons belonging to different sensory modules, resulting in the SRR-based strengthening of cortical synaptic efficacies, primarily between cortical modules A and B, as well as within each module. Once fully consolidated, the cortically stored memories across all modules can be retrieved by partial cues. The cortical storage of remote memories also requires NMDA receptor reactivation-based SRR process to counteract the synaptic turnover-related draft in synaptic efficacy Figure adapted from Wittenberg GM and Tsien JZ (2002) An emerging molecular and cellular framework for memory processing by the hippocampus. *Trends in Neuroscience* 25: 501–505.

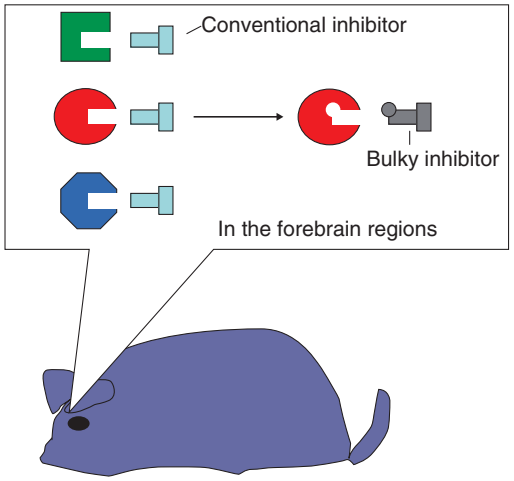

Figure 3 Bump-and-hole-based inducible protein knockout strategy. By integrating convergent protein engineering and rational inhibitor design, researchers developed an *in vivo* conditional protein knockout and/or manipulation technology. This method is based on the creation of a specific interaction interface between a modified protein domain and sensitized inhibitors. By introducing this system into genetically modified mice, we can readily manipulate the activity of a targeted protein, such as αCAMKII, on the timescale of minutes in specific brain subregions of freely behaving mice Part of the figure adapted from Wang H, Shimizu E, Tang YP, et al. (2003) Inducible protein knockout reveals temporal requirement of CaMKII reactivation for memory consolidation in the brain. *Proceedings of the National Academy of Sciences of the United States of America* 100: 4287–4292.

on reactivations of CaMKII, the key intracellular mediator for the NMDA receptor activation, during the first postlearning week. Using inducible protein knockout technique (Figure 3), researchers have found that mice with either the normal or higher level of CaMKII activity can learn and consolidate fear memories equally well. However, switching back and forth between two different levels of CaMKII activity within the first postlearning week severely disrupts synaptic consolidation of 1 month long-term fear memories. In other words, memory consolidation needs to occur at the same level of CaMKII activity as the one used during memory acquisition. Interestingly, similar alterations in the CaMKII expression level during the second, third, and fourth posttraining weeks were no longer capable of disrupting memory consolidation, again illustrating that memory consolidation has a critical time period during which precise reactivation of NMDA receptor-mediated cascades is required.

Additional genetic experiments suggest that the SRR process is also crucial for the consolidation of hippocampal-independent memory. For example, inducible gene knockout has demonstrated that the consolidation of conditioned taste aversion memory depends on the presence of the NMDA receptor in the cortex during the postlearning weeks. Therefore, SRR may present a common off-line mechanism for memory consolidation at both synaptic and systems levels in the brain.

The latest genetic analyses further suggest that the SRR process also plays a key role in memory storage. As mentioned before, it is generally believed that long-lasting memory is stored in the form of structural synaptic modifications triggered by original learning. Such structural changes, once laid down, were previously assumed to confer the long-term stability of stored memories. But since molecular and structural machineries at the synapse are recently shown to undergo routine metabolic turnovers, it prompted researchers to further investigate whether

the NMDA receptor-mediated SRR process serves as a quality-control mechanism for preventing the draft in synaptic efficacy.

Using forebrain-specific and inducible gene knockout technique, researchers reported that the 9 month retention of both contextual and cued fear memories can be profoundly impaired if the NR1 was temporarily knocked out by feeding mice with doxcycline (which induced disabling of NMDA receptor function) for 30 days during the seventh month after its initial memory acquisition. Interestingly, transient knockout of the NR1 (feeding doxycycline for 7 days) did not produce memory storage deficits, suggesting that synaptic draft is an accumulative process and that the prolonged absence of the NMDA receptor is detrimental to the long-term stable storage of remote memories in the brain. Therefore, the emerging view is that this NMDA receptor reactivation-mediated SRR process serves as a crucial and general cellular mechanism in linking the consolidation and storage of long-term memories from the molecular level to the systems level.

The hypothesis that the SRR process can account for both molecular and systems-level memory consolidation and storage has been further elaborated by the computational modeling. The computational analysis shows that in the presence of repeated NMDAR-mediated synaptic modification, SRR is fully capable of consolidating long-term memory traces (Figure 2). In contrast, in the absence of SRR, synaptic efficacy cannot be stably consolidated and preserved; consequently, the synaptic circuits gradually become unreliable, thereby undermining long-term stability of the neural circuits and information storage in the brain.

It is important to point out that the synaptic reactivation is different from the reactivation of memory at the cognitive level, in that the SRR-mediated off-line strengthening of synaptic connections requires only pairwise reactivation between two activated neurons. One conceivable triggering mechanism could be that conscious recall can initiate the SRR process. This is consistent with common experience, where the more frequently a particular event is recalled, the better and longer it will be remembered. Other triggering mechanisms could be the spontaneous reactivation of the brain network in either the awake or sleep state.

Network-Level Memory Patterns and Memory-Coding Units

One fundamental goal of neuroscience is to understand the organizing principles and the cellular network basis by which the brain encodes mnemonic information in real time. Although enormous information has been obtained by lesions and molecular techniques, there is little understanding of the network-level encoding mechanisms underlying information processing. Recently, researchers used a large-scale ensemble recording array to investigate the activity patterns of over 200 individual neurons in the mouse hippocampus during the encoding process of episodic memories. The large-scale ensemble recordings have revealed that various robust episodic experiences, such as free-fall inside a plunging elevator (drop), a sudden gush of air to the back of an animal's back (air blow), or an earthquake-like shake of the home cage (earthquake), have evoked diverse changes in the firing of some of the hippocampal CA1 cells (Figure 4).

Pattern-classification algorithms, namely multiple discriminant analysis (MDA) and principal component analysis (PCA) reveal that various startle-triggered ensemble responses of CA1 neurons form distinct patterns in a low-dimensional encoding subspace (Figure 5). Further application of a sliding-window technique to MDA or PCA methods enables the researchers, for the first time, to directly visualize and dynamically monitor real-time network memory encoding patterns. Moreover, the method also permits the detection and quantification of postlearning processing and reactivations of newly formed ensemble patterns (Figure 4).

Application of agglomerative hierarchical clustering and sequential MDA methods further reveals that the network-encoding power is actually derived from a set of functional coding units, termed neural cliques – a group of neurons with the similar response property and selectivity, in the CA1 cell population. For example, 'the general startle neural clique' consists of individual cells capable of responding to all types of startling stimuli including the elevator drop, earthquake, and air blow, whereas the 'subgeneral startle cliques' are neural groups that respond to a combination of two types, not all, of startling events. In addition, there are neuron groups which exhibit high specificity toward one specific type of startling events, such as elevator drop (the drop-specific neural clique), earthquake ('earthquake-specific neural clique'), or sudden air-blow events ('air-puff specific neural clique').

The key feature of neural cliques is that the individual neurons belonging to a given clique exhibit 'collective co-spiking' temporal dynamics. This collective co-spiking dynamics among neural clique members enables the memory coding units to achieve real-time network-level encoding robustness by overcoming the response variability of individual neurons.

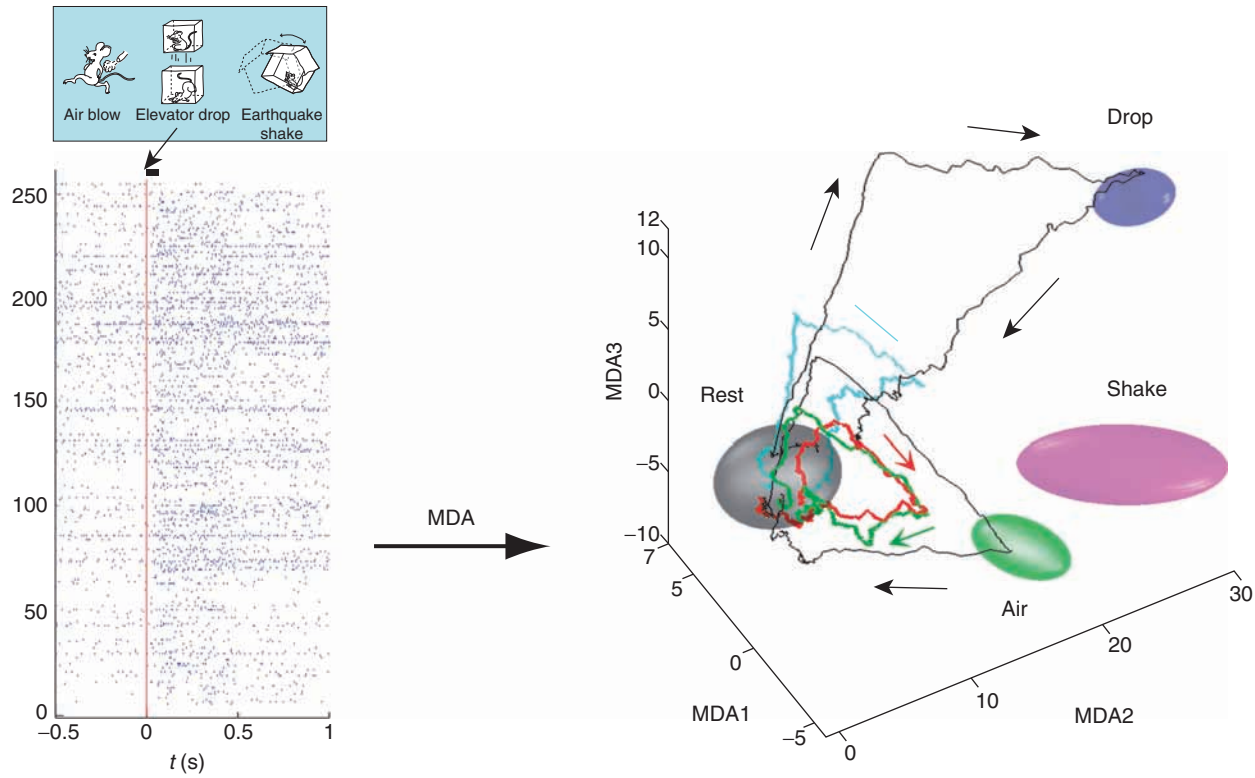

Figure 4 Visualization and dynamical monitoring of network-level memory patterns in the hippocampus. Two hundred and sixty simultaneously recorded single units from a mouse during a period of 0.5 s prior to and 1 s after the occurrence of startling episodes ($t = 0$ marked with vertical red line) forms distinct ensemble patterns during awake rest (dots, yellow ellipsoid), air-blow (circles, green ellipsoid), drop (triangles, blue ellipsoid) and earthquake (stars, magenta ellipsoid) epochs (shown in a three-dimensional sub-encoding space obtained using MDA). Dynamical monitoring of memory trace formation and reactivations are shown as trajectories of the ensemble encoding patterns. The initial response to an air-blow or drop events (black lines) is followed by two large spontaneous excursions (red and green lines for two air-blow reactivations, and purple line for drop pattern reactivation). Please note that the reactivated traces occurred as co-planar, geometrically similar lower amplitude trajectories (directionality indicated by arrows). Part of figure adapted from Lin L, Osan R, Shoham S, Jin W, Zuo W, and Tsien JZ (2005) Identification of network-level coding units for real-time representation of episodic experiences in the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* 102: 6125–6130.

Hierarchical Organization of Memory-Encoding Neural Clique Assemblies

Through examining the overall organization of neural clique assembly involved in startle memory encoding, it seems that the internal CA1 representations of any given startle episode involve a combinatorial set of neural cliques, invariantly consisting of the 'general startle clique', 'subgeneral startle clique', 'startle identity-specific clique', and context-specific startle clique. As such, each clique assembly is invariantly organized in a categorical hierarchy manner and is consisted of a 'feature-encoding pyramid' (Figure 5): starting with the neural clique representing the most general and common features (to all categories) at the bottom layer, followed by neural cliques responding to less general features (covering multiple, but not all, common categories), and then moving gradually up toward more and more specific and discriminating features (responding to a specific category), and eventually with the most discriminating feature clique (corresponding to context-specificity) on the top of the feature-encoding pyramid.

According to this hierarchical structure of network-level memory encoding (Figure 5), the 'general startle neural clique' represents the neurons engaged in the extraction of the common features among various episodes (e.g., encoding abstract and generalized knowledge that 'such events are scary and dangerous', by integrating neural inputs from the amygdala). The 'subgeneral neural cliques' are involved in identifying sub-common features across a subset of startling episodes (e.g., perhaps, the 'earthquake and drop-specific clique' for encoding the semantic memory of the fact that 'those events involve shaking and motion disturbances', by integrating inputs from the vestibular system); whereas the 'startle identity-specific cliques'

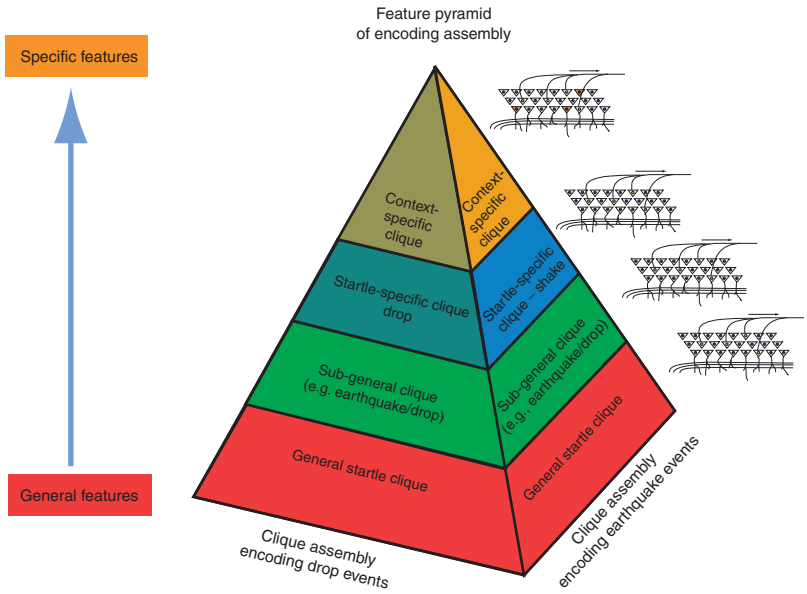

Figure 5 Hierarchical and categorical organization of memory coding units. Memory-encoding neural cliques are organized in a hierarchical and categorical fashion, and any given startling episode is encoded by a combinatorial assembly of a series of neural cliques, invariantly consisting of the general startle clique, sub-general startle clique, startle identity-specific clique, and context-specific startle clique. In this feature pyramid of the encoding clique assembly, the neural clique representing the most general, abstract features (to all categories) is at the bottom and it forms a common building block for all types of startle event encoding. The next layer of the pyramid is made by neural cliques responding to less general features (covering multiple, but not all, categories), those sub-general cliques are present in a subset of the neural clique assemblies. Moving up along this encoding feature pyramid, neural cliques become more and more specific. The neural clique at the top of the pyramid encodes the most specific and highly discriminate features, thereby defining a particular event or experience. Please note that the number of neurons for each clique does not necessarily correspond to its position in the feature pyramid. In other words, the neural clique encoding the general features does not necessarily have more neurons than the neural cliques encoding more specific features. Figure adapted from Lin L, Osan R, and Tsien JZ (2006) Organizing principles of real-time memory encoding: Neural clique assemblies and universal neural codes. *Trends in Neuroscience* 29: 48–57.

encode discriminative information about startle types (defining 'what type' of events have happened); and the 'startle context-specific cliques' provide even more specific features, such as contextual information about where a particular startling event has happened.

This invariant 'feature-encoding pyramid' of neural clique assemblies reveals four basic principles for the organization of memory encoding in the brain (Figure 5): first, the neural networks in the memory systems employ a categorical and hierarchical architecture in organizing memory coding units. Second, the internal representations of external events in the brain through such a feature-encoding pyramid are achieved not by recording exact details of the external event, but rather by recreating its own selective pictures based on the importance for survival and adaptation. Third, the 'feature-encoding pyramid' structure provides a network mechanism, through a combinatorial and self-organizing process, for creating seemingly unlimited numbers of unique internal patterns, capable of dealing with potentially infinite numbers of behavioral episodes that an animal or human may encounter during its life. Fourth, in addition to its vast memory storage capacity, this neural clique-based hierarchical extraction and parallel binding processes also enable the brain to achieve abstraction and generalization, cognitive functions essential for dealing with the complex, ever-changing situations.

The finding that the memory-encoding neural clique assembly appears to invariantly contain the coding units for processing the abstract and generalized information is particularly revealing. It provides crucial insights into how the formation of daily episodic memories is intrinsically linked to the formation of semantic memories – memories of facts, concepts, and generalized knowledge.

The observed 'feature-encoding pyramid' structure of the neural clique assembly is likely to represent a general mechanism for memory encoding across different animal species. For example, single unit recordings in human hippocampus show that some hippocampal cells fire in response to faces, or more selectively, to certain types of human facial emotions, others seem to exhibit highly selective firing to one specific person regardless (e.g., 'actress Halle Berry cell' which fires selectively to her photo portraits, Catwoman character, and even a string of her name).

Although those cells were not recorded simultaneously, the findings are nonetheless consistent with the general-to-specific feature pyramid structure. In addition, it is also reported that while some place cells in the rat hippocampus exhibit location-specific firing regardless of whether the animals engage in a random forage or goal-oriented food retrieval (or make a left-or right-turn in a T-maze), others seem to fire selectively at their place fields only in association with a particular kind of experience.

Universal Neural Codes Across Individuals and Species

With the identification of the neural clique as a basic coding unit and the feature-encoding pyramid within the clique assemblies, researchers can now mathematically convert those distinct ensemble representations observed in a low-dimensional encoding-subspace into a string of binary activation codes with 1's and 0's. This binary assignment, 1 for the active state and 0 for the inactive state of neural cliques, is based on the idea that the activity state of a neural clique can be monitored by downstream neurons using a biologically plausible binary activation function. This mathematical conversion of the activation patterns of the neural clique assembly into a binary code of 1's and 0's creates a simple and convenient way for universally comparing and categorizing network-level representations from brain to brain of the same species and even across animal species.

It is important to point out the fundamental differences between the neural clique pattern-based brain codes and the nucleotide-based genetic codes. Specifically, the neural clique-based brain codes have at least four distinct properties:

1. *Uninheritable*. Genetic codes are directly transferred through reproductions; whereas brain codes, by and large, are not inheritable and can only be acquired through experiences (perhaps with exceptions of those neural codes for controlling the primitive functions such as breathing, heartbeat, and knee-jerk reflex, etc.; those may have been genetically programmed).
2. *Self-organizing*. Genetic codes act like predetermined scaffolds, providing blueprints for the development and basic functionality of the organism, whereas brain codes are dynamical and self-organizing, arising out of internal structures and connectivity of neural networks upon behavioral experiences.
3. *Variable sizes*. The numbers of genes are exactly fixed for each individual and species, whereas the number of brain codes is highly variable in each brain, and in theory, it is only limited by the network capacity (which is determined by the convergence and divergence in connectivity), as well as the amount of behavioral experiences that an individual encounters.
4. *Modifiable*. Unless mutated, the genetic code remains static, whereas the membership of individual neurons within a given neural clique is modifiable by experience-dependent synaptic plasticity or disease states.

Thus, the above features of the brain codes are set apart in a fundamental way from the genetic codes.

The identification of neural cliques as memory coding units in the hippocampus prompts us to entertain the concept that neural cliques as basic, self-organizing

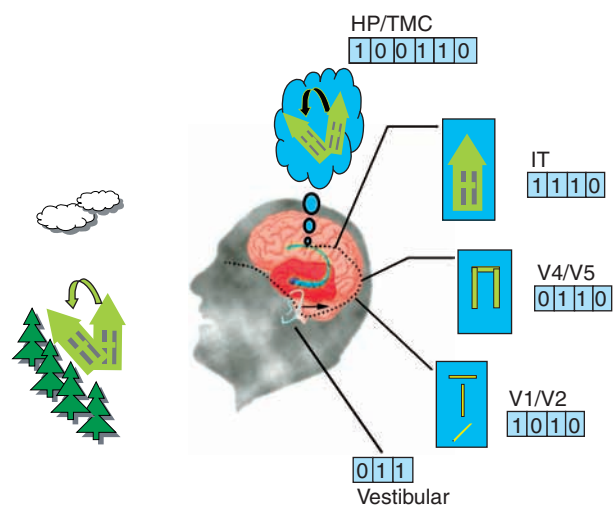

Figure 6 The hierarchical-extraction and parallel-binding process in the brain. Under this neural clique coding process, the brain achieves coherent internal encoding and processing of the external events. For example, when a person experiences a sudden earthquake, neural cliques in his primary visual cortex encode the decomposed features about edge orientation, movement, and eventually shapes of visual objects, whereas the neural cliques in the vestibular nuclei detect sudden motion disturbances. As information is processed along its pathways into deeper cortex such as the inferior temporal cortex (IT), neural cliques begin to exhibit complex encoding features such as houses. By the time it reaches high association cortices such as the hippocampus (HP) and temporal medial cortex (TMC), the neural clique assembly encodes earthquake experience and its location, with 'what and where' information. At this level, some more abstract and generalized information in experiencing the earthquake event (e.g., semantic memories of 'the earthquake is dangerous and scary') have emerged. As information is further processed into other cortical regions involving decision-making and motor-planning, a series of phased firing among various neural clique assemblies lead to adaptive behaviors, such as screaming and running away from the house or hiding under a dining table. The figure is adapted from Lin L, Osan R, and Tsien JZ (2006) Organizing principles of real-time memory encoding: Neural clique assemblies and universal neural codes. *Trends in Neuroscience* 29: 48–57.

processing units may be applicable to many, if not all, neural networks in the brain. Under this neural clique code model (Figure 6), the functionality implemented by neural cliques in a given network depends on the specializations of the corresponding regions. In primary sensory regions, neural cliques in those regions (perhaps, organized in the forms of cortical columns) encode piecemeal information by decomposing external events into various basic features (e.g., the primary visual cortex for detecting edge orientation, color or size of visual objects, whereas the vestibular nuclei for detecting motion). As information is further processed along its pathways into deeper regions, neural cliques (although no longer organized in their anatomically distinguishable maps or columns) start to encode more complex features (e.g., shapes and complex objects such as houses and faces in the inferior temporal cortex). By the time it reaches high association cortices such as the hippocampus, neural cliques have already contained both specific and generalized mnemonic information about events, places, and people with a significant amount of abstraction and generalization. Eventually, the brain areas involved in decision making, executive function, and motor planning may start coherent and phased firings among various neural cliques, thereby generating behaviors.

The emergence of a variety of new technologies, ranging from molecular genetics to high-density recording arrays, has begun to provide researchers an unprecedented opportunity to investigate the relationship between genes, brain, and mind. The latest identification of memory coding unit assemblies and their hierarchical organization in the brain has revealed as to how and why memory, semantic knowledge, abstraction, and generalization are coherently generated. It gives scientists not only a mathematical solution to translate neural activity into universal binary codes for the assembly of the codebook of the mind, but also a potential strategy for designing truly intelligent computers and robots.

See also: Episodic Memory: Assessment in Animals; Hebbian Plasticity; Hippocampus and Neural Representations; Long-Term Depression: Cerebellum; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Memory Representation; Memory Consolidation: Systems; Multiple Memory Systems; NMDA Receptors and Development; Sleep-Dependent Memory Processing; Spatial Memory: Assessment in Animals; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Cain DP, Saucier D, Hall J, Hargreaves EL, and Boon F (1996) Detailed behavioral analysis of water maze acquisition under APV or CNQX: Contribution of sensorimotor disturbances to drug-induced acquisition deficits. *Behavioural Neuroscience* 110: 86–102.

Cohen NJ and Eichenbaum H (1993) *Memory, Amnesia, and the Hippocampal System*. Cambridge, MA: MIT Press.

Corkin S (2002) What's new with the amnesic patient H.M.? *Nature Reviews Neuroscience* 3: 153–160.

Cui Z, Lindl KA, Mei B, Zhang S, and Tsien JZ (2005) Requirement of NMDA receptor reactivation for consolidation and storage of nondeclarative taste memory revealed by inducible NR1 knockout. *European Journal of Neuroscience* 22: 755–763.

Cui Z, Wang H, Tan Y, Zaia KA, Zhang S, and Tsien JZ (2004) Inducible and reversible NR1 knockout reveals crucial role of the NMDA receptor in preserving remote memories in the brain. *Neuron* 41: 781–793.

Datta S (2006) Activation of phasic pontine-wave generator: A mechanism for sleep-dependent memory processing. *Sleep and Biological Rhythms* 4: 16–26.

Debiec J, LeDoux JE, and Nader K (2002) Cellular and systems reconsolidation in the hippocampus. *Neuron* 36: 527–538.

Feng R, Rampon C, Tang YP, et al. (2001) Deficient neurogenesis in forebrain-specific presenilin-1 knockout mice is associated with reduced clearance of hippocampal memory traces. *Neuron* 32: 911–926.

Frankland PW, O'Brien C, Ohno M, Kirkwood A, and Silva AJ (2001) Alpha-CaMKII-dependent plasticity in the cortex is required for permanent memory. *Nature* 411: 309–313.

Frey U and Morris RG (1998) Synaptic tagging: Implications for late maintenance of hippocampal long-term potentiation. *Trends in Neuroscience* 21: 181–188.

Kandel ER (2001) The molecular biology of memory storage: A dialogue between genes and synapses. *Science* 294: 1030–1038.

Kim JJ and Fanselow MS (1992) Modality specific retrograde amnesia of fear. *Science* 256: 675–677.

Lin L, Osan R, Shoham S, Jin W, Zuo W, and Tsien JZ (2005) Identification of network-level coding units for real-time representation of episodic experiences in the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* 102: 6125–6130.

Lin L, Osan R, and Tsien JZ (2006) Organizing principles of real-time memory encoding: Neural clique assemblies and universal neural codes. *Trends in Neuroscience* 29: 48–57.

Lisman J, Schulman H, and Cline H (2002) The molecular basis of CaMKII function in synaptic and behavioral memory. *Nature Reviews Neuroscience* 3: 175–190.

Malenka RC and Nicoll RA (1999) Long-term potentiation – a decade of progress? *Science* 285: 1870–1874.

Malinow R, Schulman H, and Tsien RW (1989) Inhibition of postsynaptic PKC or CaMKII blocks induction but not expression of LTP. *Science* 245: 862–866.

Maren S (2001) Neurobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience* 24: 897–931.

Mayford M, Wang J, Kandel ER, and O'Dell TJ (1995) CaMKII regulates the frequency-response function of hippocampal synapses for the production of both LTD and LTP. *Cell* 81: 891–904.

McClelland JL, McNaughton BL, and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102: 419–457.

McGaugh JL (2000) Memory – a century of consolidation. *Science* 287: 248–251.

Miller RR and Springer AD (1974) Implications of recovery from experimental amnesia. *Psychological Review* 81: 470–473.

Monyer H, Sprengel R, Schoepfer R, et al. (1992) Heteromeric NMDA receptors: Molecular and functional distinction of subtypes. *Science* 256: 1217–1221.

Muller GE and Pilzecker A (1900) Experimentelle Beitrage zur Lehre vom Gedachtniss. *Zeitschrift fur Psychologie Erganzungsband* 1: 1–300.

Nakanishi S (1992) Molecular diversity of glutamate receptors and implications for brain function. *Science* 258: 597–603.

Perazzona B, Isabel G, Preat T, and Davis RL (2004) The role of cAMP response element-binding protein in *Drosophila* long-term memory. *Journal of Neuroscience* 24: 8823–8828.

Rampon C, Tang YP, Goodhouse J, Shimizu E, Kyin M, and Tsien JZ (2000) Enrichment induces structural changes and recovery from nonspatial memory deficits in CA1 NMDAR1-knockout mice. *Nature Neuroscience* 3: 238–244.

Rempel-Clower NL, Zola SM, Squire LR, and Amaral DG (1996) Three cases of enduring memory impairment after bilateral damage limited to the hippocampal formation. *Journal of Neuroscience* 16: 5233–5255.

Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurochemistry* 20: 11–21.

Shimizu E, Tang YP, Rampon C, and Tsien JZ (2000) NMDA receptor-dependent synaptic reinforcement as a crucial process for memory consolidation. *Science* 290: 1170–1174.

Silva AJ, Paylor R, Wehner JM, and Tonegawa S (1992) Impaired spatial learning in alpha-calcium-calmodulin kinase II mutant mice. *Science* 257: 206–211.

Squire LR and Alvarez P (1995) Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinion in Neurobiology* 5: 169–177.

Sutherland GR and McNaughton B (2000) Memory trace reactivation in hippocampal and neocortical neuronal ensembles. *Current Opinion in Neurobiology* 10: 180–186.

Suzuki WA (1996) The anatomy, physiology and functions of the perirhinal cortex. *Current Opinion in Neurobiology* 6: 179–186.

Sweatt JD (1999) Toward a molecular explanation for long-term potentiation. *Learning and Memory* 6: 399–416.

Tang YP, Shimizu E, Dube GR, et al. (1999) Genetic enhancement of learning and memory in mice. *Nature* 401: 63–639.

Tsien JZ (2000) Building a brainier mouse. *Scientific American* 282: 62–68.

Tsien JZ (2000) Linking Hebb's coincidence-detection to memory formation. *Current Opinion in Neurobiology* 10: 266–273.

Tsien JZ, Chen DF, Gerber D, et al. (1996) Subregion- and cell type-restricted gene knockout in mouse brain. *Cell* 87: 1317–1326.

Tsien JZ, Huerta PT, and Tonegawa S (1996) The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory. *Cell* 87: 1327–1338.

Wang H, Shimizu E, Tang YP, et al. (2003) Inducible protein knockout reveals temporal requirement of CaMKII reactivation for memory consolidation in the brain. *Proceedings of the National Academy of Sciences of the United States of America* 100: 4287–4292.

Wigstrom H and Gustafsson B (1985) On long-lasting potentiation in the hippocampus: A proposed mechanism for its dependence on coincident pre- and postsynaptic activity. *Acta Physiologica Scandinavica* 123: 519–522.

Wilson MA and McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* 265: 676–679.

Wiltgen BJ, Brown RA, Talton LE, and Silva AJ (2004) New circuits for old memories: The role of the neocortex in consolidation. *Neuron* 44: 101–108.

Wittenberg GM, Sullivan MR, and Tsien JZ (2002) Synaptic reentry reinforcement based network model for long-term memory consolidation. *Hippocampus* 12: 637–647.

Wittenberg GM and Tsien JZ (2002) An emerging molecular and cellular framework for memory processing by the hippocampus. *Trends in Neuroscience* 25: 501–505.

Metabolic Syndrome and Sleep

F W Turek, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Simply observing people – at the beach, at a shopping mall, walking down a crowded street, at a sports stadium, or in the workplace – will lead the observant individual to the inescapable conclusion that there is a major medical problem in the United States: obesity. Obesity is a disease that in just the past few decades has been increasing dramatically in developed countries, and it has reached epidemic levels in the United States. Color maps of the United States provided by the US Centers for Disease Control and Prevention, depicting the prevalence of overweight people in individual states, reveal in a dramatic fashion just how rapidly the 'obesity invasion' has occurred and the enormity of the epidemic. ('Overweight' refers to a body mass index, or BMI, between 25 and 30, and obesity is a BMI > 30; this measure of obesity is calculated as weight in kilograms divided by the square of height in meters.) The obesity epidemic is not limited to adults, as there are numerous epidemiological studies demonstrating an alarming increase in obesity in children. This increase is even leading diabetologists to rethink the term 'adult-onset diabetes,' to differentiate adult-onset (type 2) from juvenile-onset (type 1) diabetes, as many juveniles are now being diagnosed with insulin resistance and type 2 diabetes. Furthermore, this epidemic of obesity is not limited to the developed, food-rich countries of the world, as increased signs of obesity and diabetes are being found in less developed countries, including the two most populous countries, China and India.

Besides the social stigma attached to obesity, the upward trends in BMI are cause for concern because of risks of secondary complications, including signs of what has been referred to as the 'metabolic syndrome' (formerly referred to as syndrome X), including insulin resistance, hyperglycemia, dyslipidemia, hypertension, inflammation, and coagulation disorders. Obesity and the metabolic syndrome have been linked to heart attack, stroke, diabetes, cancer, and arthritis, all of which are responsible in part for some of the major health problems facing the industrialized world today.

While less easy to observe by just watching people walk down the street, in the United States there has been a parallel epidemic – an epidemic many have attributed to the fast pace of our '24/7' industrial societies – involving chronic partial sleep loss. Voluntary chronic sleep loss represents an increasingly common condition in industrialized societies and is due primarily to socioeconomic and cultural pressures. The incidence of symptoms of insufficient sleep has been increasing since about 1950. Around-the-clock operations imply chronic partial sleep loss for the large population of shift workers, and schedules compressing the workweek by the use of extended work shifts (for example, 12 h workdays) and unscheduled overtime have become routine in a number of industries.

The relationship between the rise in obesity and the loss of sleep, hypothesized to be 'interacting epidemics,' is depicted in Figure 1. However, are these epidemics interacting or are they simply correlated, with no cause-and-effect relationship to one another? The following discussions provide both epidemiological and clinical research results which indicate that long-term partial chronic sleep loss is indeed an underlying cause of obesity, diabetes, and the metabolic changes associated with the metabolic syndrome. In addition, data from animal models supporting the connection between sleep loss and metabolic dysfunction are provided, in addition to a discussion of how the two interacting epidemics could be connected on a mechanistic level. It should be noted that we are at an early stage in our understanding of how the obesity and chronic sleep loss epidemics are linked to one another, as their possible dependency on one another has only emerged in the past few years.

Sleep Loss and Metabolic Syndrome: Interacting Epidemics – Epidemiological Studies

There are now numerous epidemiologic surveys indicating a strong association between sleep disturbances, or shortened sleep durations, and cardiovascular disease, diabetes mellitus, and respiratory disorders. Indeed, as of 2006, there were at least 35 epidemiological studies linking short sleep periods with adverse health outcomes, including obesity, diabetes, and cardiovascular disease. For example, individuals who sleep less than 5 h a night have a threefold increased risk of heart attacks, and patients with coronary heart disease had more complaints of disturbed sleep than did patients without coronary heart disease. Another example is sleep apnea, which has been shown to increase the risk of cardiovascular disease and diabetes mellitus. In turn, cardiovascular and metabolic disorders, most notably hypertension, congestive heart failure, coronary heart disease, and diabetes mellitus, are associated with a higher prevalence of sleep apnea.

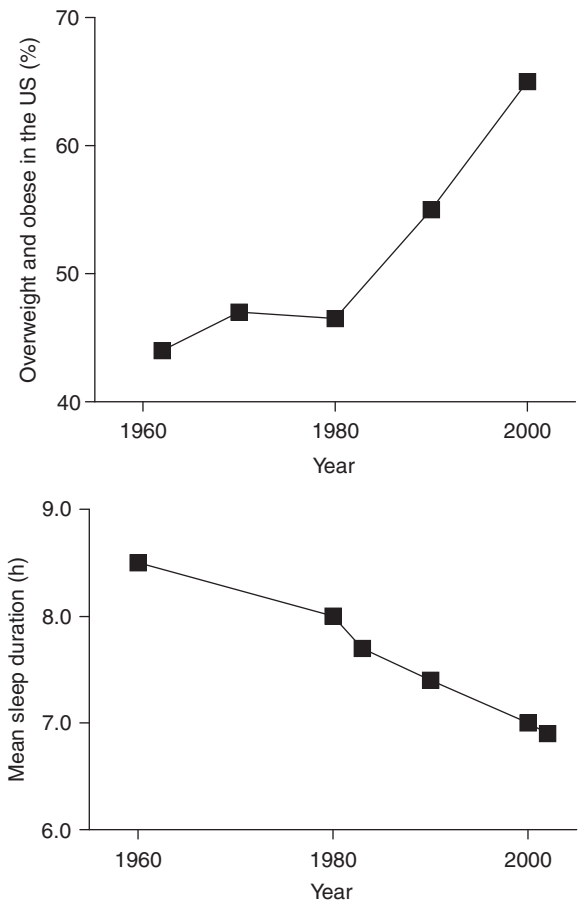

Figure 1 Time lines for the epidemics of obesity and sleep loss in the United States. The figures have been redrawn from data collected for the Centers for Disease Control and Prevention National Health and Nutrition Examination Study (obesity) and from National Sleep Foundation polls (sleep duration).

Recent studies indicate that the relationship between short sleep duration and obesity extends beyond urban settings to rural populations, for whom lifestyle factors, nutrition, physical activity levels, and work hours are often different than for city dwellers.

For individuals with type 2 diabetes mellitus, sleep disturbances may be related to obesity and sleep apnea, and recent studies have also demonstrated a strong association between obesity, impaired glucose tolerance, insulin resistance, and sleep apnea. Results from both short-term physiological and epidemiologic studies consistently suggest that decreased sleep duration or quality may increase the risk for diabetes mellitus. Indeed, in patients with type 2 diabetes mellitus, decreased subjective sleep duration and quality are predictive of poorer glycemic control. These are notable findings for health care professionals and patients because they indicate not only that voluntary chronic partial sleep deprivation may increase the risk for diabetes mellitus, but also that continuing this pattern renders it more difficult to treat the condition.

In another study that analyzed 924 patients from four primary care practices in Virginia, a reduced amount of sleep was associated with overweight and obese status, with patients in the obese group showing a near inverse linear relationship between weight and reported sleep time. Findings from the Amsterdam Growth and Health Longitudinal Study, which began collecting data on study participants at 13 years of age, on factors such as fatness, fitness, and lifestyle, revealed these factors to be important determinants of the metabolic syndrome in 10% of the studied population by the age of 36 years – an indication that the obesity epidemic in young people is not confined to the United States. An alarming statistic defines a major problem among North American youth: between 1980 and 2000, the percentage of obese young people has risen from 5% to 15%, and another 15% has moved into that classification in young adulthood. While this epidemic in such a young population has received national attention, less well known is the finding that children of all ages in the United States are sleeping 1–2 h less per night than they need, according to a recent poll taken by the National Sleep Foundation. While insufficient sleep has often been associated with the elderly population, the increasing demands and lifestyles of modern society have imposed restricted sleep regimens on youth as well. It is now critical to determine the importance of a lack of sufficient sleep during the early formative years, and how that may be involved in putting the young on a trajectory toward obesity and the metabolic syndrome – a trajectory that could be altered in some cases, if sleep loss is indeed playing a role in this epidemic.

Sleep Loss and Obesity and the Metabolic Syndrome: Interacting Epidemics – Clinical Research Studies

Van Cauter and colleagues published a pivotal study in 1995 that demonstrated in humans that sleep loss has clear negative consequences for the body, beyond the brain. In this laboratory experiment, 6 days of restricted time (4 h per night) in bed in young men was associated with alterations in glucose metabolism that, over the long term, could increase the risk of developing diabetes. A similar study demonstrated that even short-term chronic sleep loss was associated with alterations in the levels of two hormones involved in appetite regulation: leptin and ghrelin. These alterations led to increased appetite, raising the possibility that such changes could ultimately increase the risk of weight gain. Similar associations between sleep duration and leptin and ghrelin were observed in a larger sample from the Wisconsin Sleep

Cohort Study, providing further evidence for the link between sleep and appetite regulation. Such studies in chronically sleep-deprived humans has renewed interest in the role of insufficient sleep as a cause of many of the metabolic abnormalities associated with sleep apnea. However, while there is a growing awareness among some sleep, metabolic, cardiovascular, and diabetes researchers that insufficient sleep could be leading to a cascade of disorders, few in the general medical profession or in the lay public have yet made the connection.

Sleep Loss and Obesity and the Metabolic Syndrome: Interacting Epidemics – Evidence from Animal Studies

Unfortunately, most of the weight-conscious public, and even many in the health care professional community, still view weight control as a biological black box and ascribe obesity to laziness and a lack of will. However, the experience of those whose body weight is 'defended,' even though they are dieting and trying to lose weight, suggests that body weight regulation is more complex than just will power. Indeed, it is controlled much like a thermostat, with fluctuation around a programmed set point that can be altered after weight gain. The magic bullet for preventing obesity has proved elusive because, like other complex diseases (e.g., hypertension, hyperlipidemia, and cancer), multiple genetic and environmental factors contribute to body weight dysregulation and its secondary complications. An important goal in developing new therapies for obesity and the metabolic syndrome is to elucidate the elaborate pathways that regulate appetite, food intake, and energy expenditure, and to determine how interactions between behavior, metabolism, and the environment contribute to long-term energy balance. It is of particular interest to determine the cellular and neural mechanisms that underlie the effects of acute and/or chronic sleep loss on metabolism.

Genetic and clinical observations both point to interconnections between sleep and metabolic regulation. A major finding that set the stage for investigating neural interconnections linking sleep and metabolic homeostasis was the discovery of the orexins (also called hypocretins). These neuropeptides are part of signaling pathways within the energy center of the brain. A major discovery by Mignot and colleagues was the identification of the orexin receptor as the causative genetic defect in the most common form of canine narcolepsy. While these two fundamental systems appear to have distinct functions in human physiology, the discovery of overlapping molecular components suggests a co-regulation that may have coordinated the alternance between periods of fasting and sleep and periods of wakefulness, food seeking, and energy storage. A lack of sleep may perturb the homeostatic coordination of these pathways, activating metabolic pathways that trigger energy conservation (increased food intake and decreased expenditure) in excess of the energy cost of increased wakefulness. Indeed, our recent analysis of leptin-deficient animals suggests that leptin *per se* may be necessary for consolidated sleep, since animals with a genetic absence of leptin display markedly increased sleep fragmentation and experience increased periods of microsleep during the active period. In view of observations, from experimental models and human clinical studies, on the relationship between the sleep–wake cycle and fuel homeostasis, it may well be that interventions to improve sleep may ameliorate some of the metabolic deficits associated with overweight and obesity.

Recent genetic studies in mice have also provided strong evidence for there being important linkages between the sleep and body weight regulatory systems. We have recently discovered that homozygous *Clock* mutant mice have a greatly attenuated diurnal feeding rhythm, are hyperphagic and obese, and develop a metabolic syndrome of hyperleptinemia, hyperlipidemia, hepatic steatosis, and hyperglycemia, with insufficient compensatory insulin production, a hallmark of type 2 diabetes mellitus. In addition, levels in the expression of hypothalamic peptides associated with energy balance were greatly attenuated in the *Clock* mutant animals. The *Clock* gene is a major component in the transcriptional–translational feedback loops that make up the molecular circadian clock in mammals. Taken together, these results indicate that the circadian clock gene network plays an important role in mammalian energy balance involving a number of central and peripheral tissues, and that disruption of this network can lead to obesity and the metabolic syndrome in mice.

Chronic sleep loss is often associated with disrupted circadian rhythms. This is particularly true in shift workers, who not only have disturbed circadian rhythms as they continually shift their sleep and wake times to different times of the 24 h day, but who also do not obtain sufficient sleep often over the course of many years. In view of recent findings indicating that disrupted rhythms can lead to metabolic dysfunction in animal models, and that at the molecular level circadian clock genes also interact with other genes that regulate glucose and lipid metabolism, adverse effects on metabolism may in some cases be due to a combination of sleep loss and disrupted rhythms. Distinguishing between adverse metabolic outcomes as a function of sleep loss as opposed to disturbed rhythms can be problematic, since for many animal

models both sleep loss and circadian disruption occur simultaneously. For example, while we concluded that decreased longevity in cardiomyopathic hamsters exposed to a light–dark cycle that was reversed on a weekly basis was due to the chronic disruption of the normal circadian organization, it could be expected that such forced re-entrainment on a weekly basis would have adverse effects as well on the quantity and/or quality of sleep. Similarly, while the development of obesity and signs of the metabolic syndrome in the *Clock* mutant animal have been attributed to circadian dysfunction at the organismal and/or molecular levels, *Clock* mutant animals show reduced sleep of 1–2 h per day; such chronic loss of sleep could also be contributing to the adverse metabolic events. While it will be important to determine the underlying links between disturbed rhythms and/or sleep loss on metabolic function from a mechanistic point of view, from a health perspective, the finding that loss of sleep and/or circadian disruption on a chronic basis, for whatever reason, can have severe adverse metabolic consequences, has important health implications that need to be addressed as our society faces an unprecedented epidemic in obesity, diabetes, and associated diseases.

Summary

Many fingers are being pointed at the possible multitude of causes for the obesity epidemic: not enough exercise, high-fat foods, poor eating habits, watching too much television, genetics, family history, fast-food hysteria, or exposure to advertising luring us to eat unhealthy foods. Even our early prenatal and postnatal environments may be putting us on a trajectory for an unhealthy metabolic state later in life, as many studies have now demonstrated that nutritional state early in life predicts how one will respond to dietary inputs later in life. Our genes evolved in humans engaged in high levels of physical activity (much of it chasing and looking for food), with eating habits requiring them to eat as much as possible (and as much meat and fat as possible) when a meal became available (i.e., rarely); today, many humans have food available all of the time, and this modern 'food everywhere' environment is a reality that may predispose modern humans to obesity.

It is hoped that the book by Kiess and colleagues, *Obesity in Childhood and Adolescence,* has raised the clinical community's awareness of the causes, consequences, and therapies of obesity in the young. However, it is noteworthy that none of the 21 chapter titles in that book contain the word 'sleep,' indicating that the connection between sleep loss and obesity is still not a major consideration in the war on obesity. The Institute of Medicine recently published a report titled *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*, in which it was stated, "The cumulative effects of sleep loss and sleep disorders represent an under-recognized public health problem and have been associated with a wide range of health consequences." Thus the alarm has been sounded – will people and health care professionals wake up?

See also: Autonomic Dysregulation During REM Sleep; Endocrine Function During Sleep and Sleep Deprivation; Immune Function During Sleep and Sleep Deprivation; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Ayas NT, White DP, Manson JE, et al. (2003) A prospective study of sleep duration and coronary heart disease in women. *Archives of Internal Medicine* 163: 205–209.

Gangwisch JE, Heymsfield SB, Boden-Albala B, et al. (2006) Short sleep duration as a risk factor for hypertension: Analyses of the first National Health and Nutrition Examination Survey. *Hypertension* 47: 833–839.

Gluckman PD and Hanson MA (2004) The developmental origins of the metabolic syndrome. *Trends in Endocrinology Metabolism* 15: 183–187.

Laposky A, Shelton J, Bass J, Dugouic C, Perrino N, and Turek FW (2006) Altered sleep regulation in leptin-deficient mice. *American Journal of physiology regulation Integrative and Comparative phsiology* 290: R1–R10.

Liu Y and Tanaka H (2002) Overtime work, insufficient sleep, and risk of non-fatal acute myocardial infarction in Japanese men. *Occupational and Environmental Medicine* 59: 447–451.

Newman AB, Spiekerman CF, Enright P, et al. (2000) Daytime sleepiness predicts mortality and cardiovascular disease in older adults. The Cardiovascular Health Study Research Group. *Journal of the American Geriatrics Society* 48: 115–123.

Punjabi NM, Shahar E, Redline S, et al. (2004) Sleep-disordered breathing, glucose intolerance, and insulin resistance: The Sleep Heart Health Study. *American Journal of Epidemiology* 160: 521–530.

Resnick HE, Redline S, Shahar E, et al. (2003) Diabetes and sleep disturbances: Findings from the Sleep Heart Health Study. *Diabetes Care* 26: 702–709.

Rosa RR (1995) Extended workshifts and excessive fatigue. *Journal of Sleep Research* 4: 51–56.

Sanders MH, Newman AB, Haggerty CL, et al. (2003) Sleep and sleep-disordered breathing in adults with predominantly mild obstructive airway disease. *American Journal of Respiratory and Critical Care Medicine* 167: 7–14.

Schwartz SW, Cornoni-Huntley J, Cole SR, et al. (1998) Are sleep complaints an independent risk factor for myocardial infarction? *Annals of Epidemiology* 8: 384–392.

Shahar E, Whitney CW, Redline S, et al. (2001) Sleep-disordered breathing and cardiovascular disease: Cross-sectional results of the Sleep Heart Health Study. *American Journal of Respiratory and Critical Care Medicine* 163: 19–25.

Spiegel K, Leproult R, L'Hermite-Baleriaux M, et al. (2004) Leptin levels are dependent on sleep duration: Relationships with sympathovagal balance, carbohydrate regulation, cortisol, and

thyrotropin. *Journal of Clinical Endocrinology and Metabolism* 89: 5762–5771.

Spiegel K, Tasali E, Penev P, et al. (2004) Brief communication: Sleep curtailment in healthy young men is associated with decreased leptin levels, elevated ghrelin levels, and increased hunger and appetite. *Annals of Internal Medicine* 141: 846–850.

Taheri S, Lin L, Austin D, et al. (2004) Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index. *PLoS Medicine* 1: 62.

Turek FW, Joshu C, Kohsaka A, et al. (2005) Obesity and metabolic syndrome in circadian clock mutant mice. *Science* 308: 1043–1045.

Van Cauter E, Copinschi G, and Turek FW (2001) Endocrine and other biological rhythms. In: DeGroot LJ and Jameson JL (eds.) *Endocrinology,* 4th edn., vol. 1, pp. 235–256. W.B. Saunders: New York.

Vorona RD, Winn MP, Babineau TW, et al. (2005) Overweight and obese patients in a primary care population report less sleep than patients with a normal body mass index. *Archives of Internal Medicine* 165: 25–30.

Wehr TA, Moul DE, Barbato G, et al. (2005) Conservation of photoperiod-responsive mechanisms in humans. *American Journal of Physiology* 265: R846–R857.

Yaggi HK, Araujo AB, and McKinlay JB (2005) Sleep duration as a risk factor for the development of type 2 diabetes. *Diabetes Care* 29: 657–661.

Young JB and Landsberg L (2005) Stimulation of the sympathetic nervous system during sucrose feeding. *Nature* 269: 615–617.

Young JB and Landsberg L (2005) Suppression of sympathetic nervous system during fasting. *Science* 196: 1473–1475.

Relevant Websites

http://www.cdc.gov – Centers for Disease Control and Prevention (United States).

http://www.sleepfoundation.org – National Sleep Foundation (United States and Territories, and International).

Metabotropic Glutamate Receptors (mGluRs): Functions

B A Grueter and D G Winder, Vanderbilt University School of Medicine, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Metabotropic Glutamate Receptors

In the mid-1980s, glutamate was shown to stimulate inositol 1,4,5-trisphosphate (IP_3) production, indicating the potential existence of a 'metabotropic,' or non-ion-channel-forming, glutamate receptor. This led to the distinction between ionotropic and metabotropic glutamate receptors (mGluRs). There are eight subclasses of mGluRs that have been identified and cloned. mGluRs belong to family C of the G-protein-coupled receptor superfamily, which includes calcium-sensing and γ-aminobutyric acid B ($GABA_B$) receptors and act to modulate neuronal excitability in many brain regions. Eight separate gene products (mGluRs 1–8) comprise the mGluR family of receptors, several of which can exist as alternatively spliced variants. These subtypes are classified into three groups based on pharmacology, similarities in coupling mechanisms, and sequence homology (Table 1). mGluRs of the same group show about 70% sequence identity, whereas between groups this percentage decreases to about 45%. Group I receptors (mGlu1 and mGlu5) are linked to G_q, whereas, group II (mGlu2 and mGlu3) and group III (mGlu4, mGlu6, mGlu7, and mGlu8) receptors are coupled to $G_{i/o}$. Intracellularly, the large C-terminus of mGluRs can interact with a variety of signaling systems.

G-Protein-Coupled Receptors

G-protein-coupled receptors (GPCRs) are seven-transmembrane-domain (7TM) proteins that interact with G-proteins. Receptor activation leads to associated guanosine triphosphatase (GTPase) activity and ultimately to activation or inhibition of an effector protein. G-protein regulation of effectors, particularly in neurons, occurs through two broad mechanisms. One is direct G-protein binding to ion channels. More traditionally, G-proteins function to activate or inhibit second-messenger-forming enzymes, leading to direct second messenger binding to ion channels, second messenger activation of kinases or phosphatases that phosphorylate/dephosphorylate ion channels, or alterations in protein–protein interactions involving an ion channel. These mechanisms constitute the major routes for mGluR signaling, although it should be noted that these receptors have also been reported to signal through G-protein-independent ways through as yet unclear mechanisms.

Structure of mGluRs

mGluRs are composed of two main domains, a large extracellular domain (ECD), and a 7TM region, and share very little sequence homology with other GPCRs. Proposed models confirmed by X-ray crystallography suggest the unique ECD contains two globular lobes separated by a cleft which contains the orthosteric ligand-binding site. While glutamate is commonly thought to be the primary ligand at this site, it is important to note that many excitatory amino acids are present in neural tissue, many of which can also activate mGluRs. In addition, allosteric binding sites have also been identified in the 7TM region. The second intracellular loop of mGluRs infers selectivity for G-protein coupling while a highly conserved third loop plays a crucial role in G-protein activation. Interactions among the first loop, third loop, and C-terminal tail are thought to control efficacy of G-protein coupling. The intracellular C-terminal region of mGluRs plays a major role in regulation and trafficking.

Regulation of mGluR Signaling

mGluR signaling is tightly regulated by both homologous and heterologous mechanisms. Both agonist-induced (homologous) and heterologous forms of regulation of mGluRs have been observed. Phosphorylation by several different kinases can alter mGluR signaling. For instance, protein kinase C (PKC)-dependent phosphorylation of mGluRs, particularly group I mGluRs, is thought to serve as a major mechanism of desensitization. β-Arrestin-induced desensitization of mGluRs has also been described. The $\beta\gamma$ subunits of G-proteins are thought to target G-protein-coupled receptor kinases (GRKs) to mGluRs, resulting in β-arrestin-mediated sequestration and ultimately desensitization.

Group I mGluR expression at the plasma membrane and the functional coupling of group I mGluRs to intracellular signaling molecules are dependent upon scaffolding proteins. The Homer family of proteins is a major example of this. There are three genes encoding Homer proteins. Homer 1a is an immediate-early gene product while Homer 1b/c, Homer 2, and Homer 3 are constitutively expressed. Homer proteins link group I mGluRs through the C-terminal tail to signaling molecules such as the IP_3 receptor, and to scaffolding components of the postsynaptic density, such as Shank. It is suggested that Homer

Table 1 Classification of mGluRs

Family receptor	*Coupling*	*Group/subtype-selective pharmacological agents*[a]
Group I		
mGluR1	G_q-coupled	Agonists: DHPG, ACPD, quisqualate Antagonist: LY393675 Inverse agonist (allosteric antagonist): LY367385
mGluR5	G_q-coupled	Agonists: DHPG, ACPD, quisqualate, CHPG Inverse agonist (allosteric antagonist): MPEP
Group II		
mGluR2	$G_{i/o}$-coupled	Agonists: DCG-IV, LY354730, ACPD Antagonist: LY341495
mGluR3	$G_{i/o}$-coupled	Agonists: DCG-IV, LY354730, ACPD Antagonist: LY341495
Group III		
mGluR4	$G_{i/o}$-coupled	Agonists: L-AP4, L-SOP Antagonists: MSOP, MAP4
mGluR6	$G_{i/o}$-coupled	Agonists: L-AP4, L-SOP Antagonists: MSOP, MAP4
mGluR7	$G_{i/o}$-coupled	Agonists: L-AP4, L-SOP Antagonists: MSOP, MAP4
mGluR8	$G_{i/o}$-coupled	Agonists: L-AP4, L-SOP, DCPG Antagonists: MSOP, MAP4

[a]ACPD, (1*S*,3*R*)-1-amino-1,3-cyclopentanedicarboxylate; L-AP4, L-2-amino-4-phosphonobutyric acid; CHPG, (*R*,*S*)-2-chloro-5-hydroxyphenylglycine; DCG-IV, (2*S*,2′*R*,3′*R*)-2-(2′,3′-dicarboxycyclopropyl)glycine; DCPG, (*S*)-3,4-dicarboxyphenylglycine; DHPG, 3,5-dihydroxyphenylglycine; MAP4, α-methyl-2-amino-4-phosphonobutyrate; MPEP, 2-methyl-6-(phenylethynyl)pyridine; MSOP, methylserine-*O*-phosphate; L-SOP, L-serine-*O*-phosphate.

proteins may regulate the expression and function of group I mGluRs at multiple levels, including targeting, surface expression, clustering, physical linkage to other synaptic and subsynaptic complexes, and modulation of constitutive activity. Interestingly, biochemical evidence suggests the C-terminus of Homer 2a interacts with the Rho family of small GTPases, in a GTP-dependent manner. Homer proteins have also been shown to form complexes with Shank proteins, which act as scaffolding proteins and link group I mGluRs with other proteins in the postsynaptic densities (PSDs), such as Dynamin2, an important molecule implicated in endocytosis, and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor (AMPAR)–glutamate receptor-interacting protein (GRIP) complexes.

Expression of mGluRs

mGluRs have unique but overlapping distributions throughout the central nervous system (CNS). In addition, mGluRs have been reported to function in the peripheral nervous system and elsewhere. For example, mGluRs have been shown to play an important role in nociceptive transmission, as will be described in a later section.

Glutamate Sources from Which mGluRs Can Be Activated

mGluRs are often located in the areas adjacent to the synapse. Functionally, their location suggests that they become activated in situations of repeated stimulation of afferents, which results in considerable glutamate accumulation in the synaptic cleft coupled to spillover to extrasynaptic sites. In addition to the frequency and duration of synaptic activity, the extracellular concentration of glutamate also depends on the efficacy of neuronal and glial uptake, which is dictated by the number of glutamate transporters (GluTs). GluTs, along with passive diffusion, are the primary means of glutamate removal from the extracellular space. The ability of mGluRs to participate in signaling is thought to be tightly regulated by GluT activity.

Another potentially important contributor to extracellular glutamate levels is the cystine–glutamate exchanger. The nonsynaptic cystine–glutamate antiporter exchanges extracellular cystine for intracellular glutamate with a minimal contribution of synaptic glutamate release. It has been suggested that one form of cocaine-induced neuroadaptation may be due to a decrease in basal levels of extracellular glutamate in the *N*-acetylcysteine (NAc) after several weeks of withdrawal from cocaine, an effect attributed to a reduced function of the cystine–glutamate transporter.

mGluR Function

Neuromodulation

Neuromodulation is a subtle influence on synaptic efficacy or neuronal excitability. Neuromodulation

can also involve longer-lasting changes such as alterations in gene expression. The 'fine tuning' of neuronal function by neuromodulation makes receptors and pathways attractive therapeutic targets for many diseases and disorders. mGluR activation serves as a major source by which glutamatergic transmission can modulate synaptic efficacy, neuronal excitability, and gene expression.

Synaptic Plasticity

Persistent alterations in synaptic efficacy initiated by transient stimuli are referred to as forms of synaptic plasticity. Changes in synaptic efficacy can occur by presynaptic mechanisms such as altered neurotransmitter release as well postsynaptic mechanisms. Activation of mGluRs can play an important role in facilitation of multiple forms of synaptic plasticity. mGluR activation can induce long-term depression (LTD) at many glutamatergic synapses. Further activation, primarily of group I mGluRs at some synapses, can positively modulate *N*-methyl-D-aspartate receptor (NMDAR)-dependent long-term potentiation (LTP) induction.

Group I mGluRs

Group I receptors (mGluR1 and mGluR5) are linked to activation of phospholipase C (PLC), leading to phosphoinositide hydrolysis, calcium release, and PKC activation. Activation of group I mGluRs can also indirectly regulate adenylyl cyclase, leading to increased cyclic adenosine monophosphate (cAMP) formation, as well as cyclic guanosine monophosphate (cGMP) accumulation, phospholipase A, and extracellular signal-regulated kinase (ERK) activation.

Although highly homologous, studies suggest mGluR1 and mGluR5 receptors have distinct functions in regulating synaptic transmission. This implies differential regulation of these receptors and potential divergence in effector systems. Conversely, at some synapses only one type of group I mGluR is functionally important: for instance, only mGluR1 has been shown to function in the ventral tegmental area and cerebellum.

Synaptic Locus of Group I mGluRs

Mechanisms underlying group I mGluR function vary across brain regions and synapses. Much information on the function of these receptors can be gained by determining the synaptic locus of action of these receptors. Immunohistochemical and electrophysiological results reveal that group I mGluRs localize and function presynaptically, postsynaptically, and on glia. However, group I mGluRs are predominantly located at the periphery of the postsynaptic density (perisynaptically). Activation of group I mGluRs causes postsynaptic effects such as neuronal depolarization, excitation, and spike frequency adaptation in a number of brain regions. For instance, selective activation of group I mGluRs with the agonist 3,5-dihydroxyphenylglycine (DHPG) increases postsynaptic membrane excitability in hippocampus, cortex, striatum, amygdala, subthalamic nucleus, and hypothalamic nuclei.

Increases in neuronal excitability via activation of group I mGluRs are thought to occur in large part through modulation of K^+ channels. Specific activation of group I mGluRs induces net inward currents by inhibiting K^+ channel conductance in neurons in the hippocampus and hypothalamus. For instance, K^+ channels are a major target of mGluRs, and many diverse types of K^+ channels are inhibited following activation of mGluRs. Activation of mGluRs has been shown to block the I_{AHP} (I_{AHP} is an outward, Ca^{2+}-activated, K^+ current with slowly activating and inactivating ($\sim$2 s) kinetics) and to thereby reduce accommodation of spike firing in hippocampal CA1 pyramidal cells, CA3 pyramidal cells, the dentate gyrus, cultured cerebellar Purkinje cells, and basolateral amygdaloid nucleus.

Activation of group I mGluRs has also been shown to modulate voltage-gated calcium channels (VGCCs). Alterations in VGCCs can occur either presynaptically, consequently changing vesicular release probability, or postsynaptically, leading to changes in intracellular calcium levels. Although mGluR-dependent modulation of channels has been shown to be a consequence of direct G-protein-linked action – for example, inhibition of Ca^{2+} channels – many other effects occur as a result of activation of intracellular messenger pathways. Group I, in contrast to group II, mGluRs can use several distinct signal transduction pathways to inhibit Ca^{2+} channels, both Ca^{2+} intracellular-independent and -dependent mechanisms.

In addition to the postsynaptic modulation of synaptic transmission by alteration of neuronal excitability, group I mGluRs have been reported to function presynaptically. One example is in the hippocampus, at the level of the synapses between Schaeffer collaterals and CA1 pyramidal cells. Consistent with presynaptic function, an increase in paired pulse facilitation has been observed upon activation of mGluR5, suggesting a decrease in vesicular release probability.

Retrograde Signaling

Cannabinoids Of great interest in the mGluR field was the finding that endocannabinoid synthesis can be

triggered by activation of group I mGluRs. Retrograde signaling through the endocannabinoid system helped explain biophysical evidence suggesting a presynaptic change in probability of neurotransmitter release, with immunochemical data suggesting that group I mGluRs are primarily localized to postsynaptic structures.

Use of pharmacological tools as well as genetic manipulations has provided evidence for a role of arachidonic acid metabolites – more specifically, the 12-lipoxygenase metabolite of arachidonic acid, 12(*S*)-hydroxyeicosa-5(*Z*),8(*Z*),10(*E*),14(*Z*)-tetraenoic acid (HETE) – in inducing group I mGluR LTD in the hippocampus. Direct application of this metabolite mimicked and occluded mGluR5-dependent LTD at these synapses; mGluR5-dependent LTD in the hippocampus is absent in mice lacking the 'leukocyte type' 12-lipoxygenase, and group I mGluR LTD-inducing stimuli promote 12-lipoxygenase activity. A suggested potential downstream target of 12-lipoxygenase is p38 kinase. p38 kinase, like ERK, is a member of the mitogen-activated protein kinase (MAPK) family and has been implicated in group I mGluR LTD. For instance, in the dentate gyrus, group I mGluR LTD is dependent on p38 kinase signaling and in the CA1 of the hippocampus activation of group I mGluRs results in AMPA receptor endocytosis (which will be discussed in greater detail in a subsequent section) and is dependent on p38 kinase signaling as well as protein synthesis.

Kinases Involved in Group I mGluR Function

Group I mGluRs are linked to the activation of multiple kinases. Consistent with coupling of group I mGluRs with G_q and Ca^{2+} signaling, activation of these receptors can lead to activation of the Ca^{2+}/diacylglycerol (DAG)-dependent PKC family. mGluR-induced activation of PKC can lead to phosphorylation of a multitude of downstream targets, ultimately resulting in neuronal modifications such as LTP or LTD. One such target of mGluR-stimulated PKC activity is the MAPK/ERK pathway. The ERK pathway is a signaling cascade that plays a crucial role in a variety of cell regulatory events, including cell proliferation, differentiation, and survival, as well as an involvement in long-term synaptic changes and behavior. In neurons, ERK activation has been shown to be involved in processes associated with synaptic remodeling and long-term changes in synaptic efficacy. These processes include protein synthesis, changes in gene expression, dendritic spine stabilization, modulation of ion channels, and receptor insertion.

The ERK cascade is activated by a variety of extracellular agents, including growth factors, hormones, and neurotransmitters, and serves as an important point of convergence for the PKC and protein kinase A (PKA) pathways. PKC regulates ERK activity through an interaction with either Ras or Raf-1, leading to activation of mitogen-activated protein extracellular kinase (MEK) and consequently ERK. Additionally, it was discovered that DAG, another second messenger product of PLC activity, is capable of activating ERK independent of PKC activation.

Consistent with a role for PKA and PKC in some forms of group I mGluR LTD, group I mGluR LTD in the hippocampus, bed nucleus of the stria terminalis (BNST), and cerebellum is dependent on ERK activation. In addition to these kinases, receptor tyrosine phosphatases have also been shown to play an integral role in the group I mGluR-induced removal of AMPARs from the synapse, resulting in LTD of excitatory synaptic transmission.

Pharmacology of Group I mGluRs

The recent development of mGluR-selective ligands (e.g., phenylglycine derivatives acting either as agonists or as antagonists) has helped begin the elucidation of the functional roles of mGluRs in brain and behavior. The pharmacological profile of group I mGluRs has been determined in mammalian heterologous expression systems on cloned mGluRs. The rank order of potency of the most common agonists is quisqualate > DHPG = glutamate > (1*S*,3*R*)-1-amino-1,3-cyclopentanedicarboxylate (ACPD). The specific agonist for group I mGluRs used in the studies presented in the following sections is DHPG, which is devoid of activity at other mGluRs. An agonist for mGluR5 – (*R*,*S*)-2-chloro-5-hydroxyphenylglycine (CHPG) – has been reported, although it is not very potent.

The first antagonist described for group I mGluRs was α-methyl-4-carboxyphenylglycine (MCPG). However, MCPG also antagonizes group II mGluRs. There are several drugs that selectively inhibit mGluR5; the most widely used is the noncompetitive antagonist 2-methyl-6-(phenylethynyl)pyridine (MPEP), with the caveat that at high concentrations it blocks NMDA receptors.

Group I mGluRs and Plasticity

Group I mGluRs have been shown to potentiate either NMDAR- or AMPAR-mediated responses. However, a major function of group I mGluR activation is to produce a persistent weakening of glutamatergic transmission at synapses in the CNS. There appear to be at least two mechanisms through which this is accomplished. One involves the recruitment of endocannabinoid signaling and presynaptic alterations, and has been observed in the dorsal and ventral striatum.

A second major mechanism is through endocannabinoid-independent signaling mechanisms, and has been heavily studied in the hippocampus and cortical regions. Both presynaptic and postsynaptic mechanisms have been described for endocannabinoid-independent forms of group I mGluR LTD.

Group I mGluRs Can Alter iGluR Function

Group I mGluRs have been shown to modulate iGluRs in several ways. Activation of mGluRs can lead to changes in neuronal excitability and therefore influence NMDAR activity. For instance, membrane depolarization via group I mGluR stimulation can enable enhanced NMDAR-mediated synaptic plasticity. Additionally, through activation of kinases/phosphatases, group I mGluRs can alter the signaling properties of iGluRs and can influence trafficking of these receptors into and out of the synapse.

PKC phosphorylates NMDA and AMPA receptors. The activation of PKC by mGluRs might therefore modulate the phosphorylation state of these two ionotropic receptors, resulting in a potentiation of the response. Activation of group I mGluRs can lead to a removal of AMPA receptors from the membrane, resulting in a depression of synaptic transmission. More specifically, following activation of group I mGluRs there is a reduction in the density of postsynaptic AMPARs.

Group II mGluRs

Group II mGluRs (mGluR2/3) are widely distributed throughout the CNS, where they are moderately to highly expressed in brain regions that are commonly associated with anxiety disorders, including the hippocampus, prefrontal cortex (PFC), and amygdala. mGluR2 is generally localized at the periphery of the presynaptic terminal, suggesting a role for these receptors at instances of spillover of glutamate beyond the synaptic cleft in response to repetitive stimulation. By contrast, mGluR3 is more diversely localized, including both pre- and postsynaptic localization on neurons, as well as relatively heavy localization on glial cells. Group II mGluRs most commonly function presynaptically to decrease the probability of synaptic vesicle release, but at multiple synapses within the CNS postsynaptic mechanisms have been reported. Voltage-dependent Ca^{2+} channels are likely to be involved in the presynaptic inhibition mediated by mGluRs. Additionally, it has been suggested that K^+ channels mediate some of the effects induced by activation of group II or III mGluRs.

LY354740 ((1*S*,2*S*,5*R*,6*S*)-(+)-2-aminobicyclo[3.1.0] hexane-2,6-dicarboxylic acid) and several related compounds are potent and selective agonists at mGluR2 and mGluR3, with an EC_{50} of 1050 nM in the rat cortex, hippocampus, and striatum, and 10 or 30 nM in cells expressing recombinant mGluR2 or mGluR3, respectively. LY341495 ((2*S*)-2-amino-2-[(1*S*,2*S*)-2-carboxycycloprop-1-yl]-3-(xanth-9-yl) propanoic acid) is an antagonist of mGluRs, with nanomolar potency against group II mGluRs and micromolar potency against group III mGluRs.

LTD of excitatory transmission can be induced by activation of group II mGluRs in several brain regions. For example, group II mGluR LTD has been shown presynaptically in the basolateral amygdala (BLA), the nucleus accumbens, the BNST, and the striatum. In contrast, it was shown that stimulation of thalamic inputs to the lateral nucleus of the amygdala induces a group II mGluR-mediated postsynaptic LTD. Additionally, in the dentate gyrus and the medial PFC it was found that group II mGluR activation induced a postsynaptic LTD that was PKA and PKC dependent.

Behavioral studies of mice lacking mGluR2 or mGluR3 support earlier findings of a role of these receptors in anxiety behaviors. Consistent with the role of these receptors in various animal models of anxiety/stress, group II mGluR agonists have been considered promising candidates as therapeutic agents for the treatment of anxiety as well as other brain disorders.

Group III mGluRs

Like group II mGluRs, group III mGluRs are coupled to inhibition of cAMP production, and modulation of ion channels. However, in contrast to the other groups, group III mGluRs are primarily located at presynaptic active zones at the axon terminal. As with the group II mGluRs, the group III mGluRs inhibit neurotransmitter release, resulting in a reduction in glutamatergic or GABAergic synaptic transmission. Unfortunately, there are few pharmacological tools with the desired receptor subtype selectivity ideal for testing the significance of the subtypes of group III mGlu receptors. L-2-Amino-4-phosphonobutanoate (L-AP4), the most commonly used selective group III agonist, has a high affinity for mGluR4, mGluR6, and mGluR8 and a low affinity for mGluR7. One of the most promising drugs available for group III mGluRs is the positive allosteric modulator of mGluR4, (–)-*N*-phenyl-7-(hydroxyimino)cyclopropa[*b*]chromen-1a-carboxamide (PHCCC). (*S*)-3,4-Dicarboxyphenylglycine (DCPG), a more recently described group III mGluR agonist, is a relatively specific mGluR8 agonist.

While group III mGluR function has been reported in brain regions such as the BLA, central nucleus of the amygdala, BNST, and hippocampus using an agonist

pharmacology approach, behavioral analysis of specific receptor mutant mice has also proved useful. In combination with the limited pharmacology, mutants of mGluR4, mGluR7, and mGluR8 have all been shown to have behavioral phenotypes. For instance, consistent with a dysregulation of the hypothalamic–pituitary–adrenal (HPA) axis, mice with targeted deletion of mGluR7 show changes in animal behavior paradigms predictive of antidepressant and anxiolytic action. They also exhibit a deficit in fear responses and learning paradigms. Interestingly, administration of the mGluR4 allosteric potentiator, PHCCC, also has anxiolytic-like effects, implicating a role for this receptor in anxiety as well. mGluR4 knockout (KO) mice have altered glutamate and GABA release and, interestingly, mGluR4 activation protects against neurodegeneration. mGluR8 has been implicated in conditioned fear, and similar to the other group III mGluRs, mice lacking mGluR8 have an increased anxiety phenotype.

Of the mGluRs, mGluR6 has the most restricted expression. These receptors are expressed postsynaptically at ON bipolar cells in both rod and cone systems and function as the main excitatory receptor responsible for relaying synaptic signals in these retinal cells. Consistent with their important functional role in visual processing, genetic targeting of mGluR6 impairs detection of visual contrasts.

Diseases

Many psychiatric as well as neurological diseases have been linked to alterations in neuronal excitability via the glutamatergic system. As mGluRs function to regulate a large number of synapses, and are known to play a role in pathophysiological conditions, compounds that act on these receptors are prime targets for a number of therapeutic applications. For instance, drugs targeting mGluRs have been shown to prevent certain forms of pain and anxiety. There are positive potential therapeutic effects of mGluR compounds in treating Parkinson's disease as well as the conditions of drug dependence and withdrawal. Also, activating or potentiating mGluRs may help in treatment of schizophrenia and Alzheimer's disease. An interesting model suggests a specific role for group I mGluRs in fragile X mental retardation. Intense investigation reveals a role for mGluRs in nociception both centrally and peripherally. For instance, group I mGluRs have been implicated in the processes of central sensitization and persistent nociception, whereas activation of group II mGluRs is effective against neuropathic or inflammatory pain.

Pharmacology/Therapeutic Utility of mGluRs

An exciting new direction in the pharmacological targeting of mGluRs has been the development of subtype-selective allosteric modulators. An advantage of allosteric modulators over traditional compounds is that allosteric modulators rely on activation by the endogenous transmitter released in an activity-dependent manner. mGluRs are prominent targets as both positive and/or negative allosteric modulators with unique pharmacological properties. Allosteric modulators, as compared to orthosteric ligands, appear to show a unique degree of selectivity.

See also: Glutamate; Glutamate Receptor Organization: Ultrastructural Insights; Glutamatergic and Gabaergic Systems; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Depression (LTD): Endocannabinoids and Cerebellar LTD; NMDA Receptors and Development.

Further Reading

Anwyl R (1999) Metabotropic glutamate receptors: Electrophysiological properties and role in plasticity. *Brain Research Reviews* 29: 83–120.

Bordi F and Ugolini A (1999) Group I metabotropic glutamate receptors: Implications for brain diseases. *Progress in Neurobiology* 59: 55–79.

Conn PJ and Pin JP (1997) Pharmacology and functions of metabotropic glutamate receptors. *Annual Review of Pharmacology and Toxicology* 37: 205–237.

DeBlasi A, Conn PJ, Pin J, et al. (2001) Molecular determinants of metabotropic glutamate receptor signaling. *Trends in Pharmacological Sciences* 22: 114–120.

Pin JP and Duvoisin R (1995) The metabotropic glutamate receptors: Structure and functions. *Neuropharmacology* 34: 1–26.

Rouse ST, Marino MJ, Bradley SR, et al. (2000) Distribution and roles of metabotropic glutamate receptors in the basal ganglia motor circuit: Implications for treatment of Parkinson's disease and related disorders. *Pharmacology & Therapeutics* 88: 427–435.

Schoepp DD (2001) Unveiling the functions of presynaptic metabotropic glutamate receptors in the central nervous system. *Journal of Pharmacology and Experimental Therapeutics* 299: 12–20.

Sladeczek F, Pin JP, Recasens M, et al. (1985) Glutamate stimulates inositol phosphate formation in striatal neurones. *Nature* 317: 717–719.

Sugiyama H, Ito I, and Hirono C (1987) A new type of glutamate receptor linked to inositol phospholipid metabolism. *Nature* 325: 531–533.

Tanabe Y, Masu M, Ishii T, et al. (1992) A family of metabotropic glutamate receptors. *Neuron* 8: 169–179.

Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology

F Nicoletti and V Bruno, University of Rome 'La Sapienza,' Rome, Italy
G Battaglia, Istituto Neurologico Mediterraneo 'Neuromed,' Pozzilli, Italy

© 2009 Elsevier Ltd. All rights reserved.

Background

The first challenge to the general belief that all glutamate receptors were ligand-gated ion channels was the demonstration by Sladeczek, Pin, Recasens, Bockaert, and Weiss that glutamate stimulates polyphosphoinositide (PI) hydrolysis in cultured striatal neurons. Similar data were obtained in brain slices, where the PI response to glutamate is maximal in the early postnatal life and declines progressively during postnatal development. The term 'metabotropic glutamate (mGlu) receptors' was introduced in 1987 by H Sugiyama and colleagues, who examined responses to excitatory amino acids in *Xenopus* oocytes injected with rat brain mRNA. Cloning of the first mGlu receptor subtype by S Nakanishi and colleagues in 1991 was the milestone for any further development of the field.

Classification of mGlu Receptors

mGlu receptors form a family of eight subtypes subdivided into three groups based on their sequence similarities and G-protein coupling.

Group I mGlu Receptors

Group I includes mGlu1 (variants: mGlu1a, -1b, -1c, -1d, and -1e) and mGlu5 (variants: mGlu5a and -5b) receptors, which are primarily coupled to G_q/G_{11} proteins (Figure 1). Secondary coupling involves G_s and the pertussis toxin-sensitive G_o. Activation of both subtypes in recombinant cells stimulates PI hydrolysis with an ensuing formation of the two intracellular second messengers, inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol (DAG). IP_3 releases Ca^{2+} from intracellular stores, whereas DAG activates protein kinase C (PKC). Activation of group I mGlu receptors can also stimulate cyclic adenosine monophosphate (cAMP) formation, arachidonic acid release, the mitogen-activated protein kinase (MAPK) pathway, and the phosphatidylinositol 3-kinase (PI3K) pathway. The mGlu1 receptor is negatively coupled to a variety of K^+ channels and to the C1 type of transient receptor potential (TRP) channels.

Homologous desensitization of group I mGlu receptors involves phosphorylation mechanisms mediated by either PKC or G-protein-coupled receptor kinases (GRKs). Different GRK isoforms, including GRK2, GRK4, and GRK5, have been found to desensitize mGlu1 receptors in recombinant cells. GRK4 is also required for desensitization of native mGlu1 receptors expressed by cerebellar Purkinje cells. While desensitization by GRK4 is mediated by receptor phosphorylation, desensitization by GRK2 is independent of phosphorylation and may involve a direct interaction with activated $G_{\alpha q}$. The mGlu5 receptor is selectively desensitized by members of the GRK2 family (GRK2 and GRK3) through a mechanism that involves phosphorylation of the Thr840 residue in the C-terminus domain of the receptor. The Regulator of G-protein signaling, RGS4, which acts as effector antagonist by increasing the GTPase activity of $G_{\alpha q}$, inhibits mGlu1 and mGlu5 receptor signaling.

Group II mGlu Receptors

Group II includes mGlu2 and mGlu3 receptors, which are coupled to G_i/G_o proteins (Figure 1). Their activation inhibits cAMP formation, inhibits L- and N-type voltage-sensitive Ca^{2+} channels, activates K^+ channels, and stimulates the MAPK and PI3K pathways. The latter two effects are presumably mediated by the $\beta\gamma$ subunits of G_i. Activation of group II mGlu receptors can also amplify the stimulation of cAMP formation by β-adrenergic receptor agonists in cultured astrocytes, and the stimulation of PI hydrolysis by mGlu1/5 receptor agonists in brain slices.

Group III mGlu Receptors

Group III includes mGlu4, mGlu6, mGlu7 (variants: mGlu7a and -7b), and mGlu8 (variants: mGlu8a, -8b, and -8c) receptors, which are all coupled to G_i/G_o proteins (Figure 1). Activation of mGlu4, mGlu7, and mGlu8 receptors inhibits cAMP formation and N-type voltage-sensitive Ca^{2+} channels. In cells expressing mGlu4 receptors, GRK2 attenuates agonist-stimulated MAPK activation without affecting inhibition of adenylyl cyclase. Activation of mGlu6 receptors inhibits cGMP-dependent cation channels in ON bipolar cells of the retina (see later).

Structure and Mechanisms of Activation of mGlu Receptors

mGlu receptors belong to class C G-protein-coupled receptors (GPCRs), which also includes γ-aminobutyric acid B ($GABA_B$) receptors, the Ca^{2+}-sensing receptor

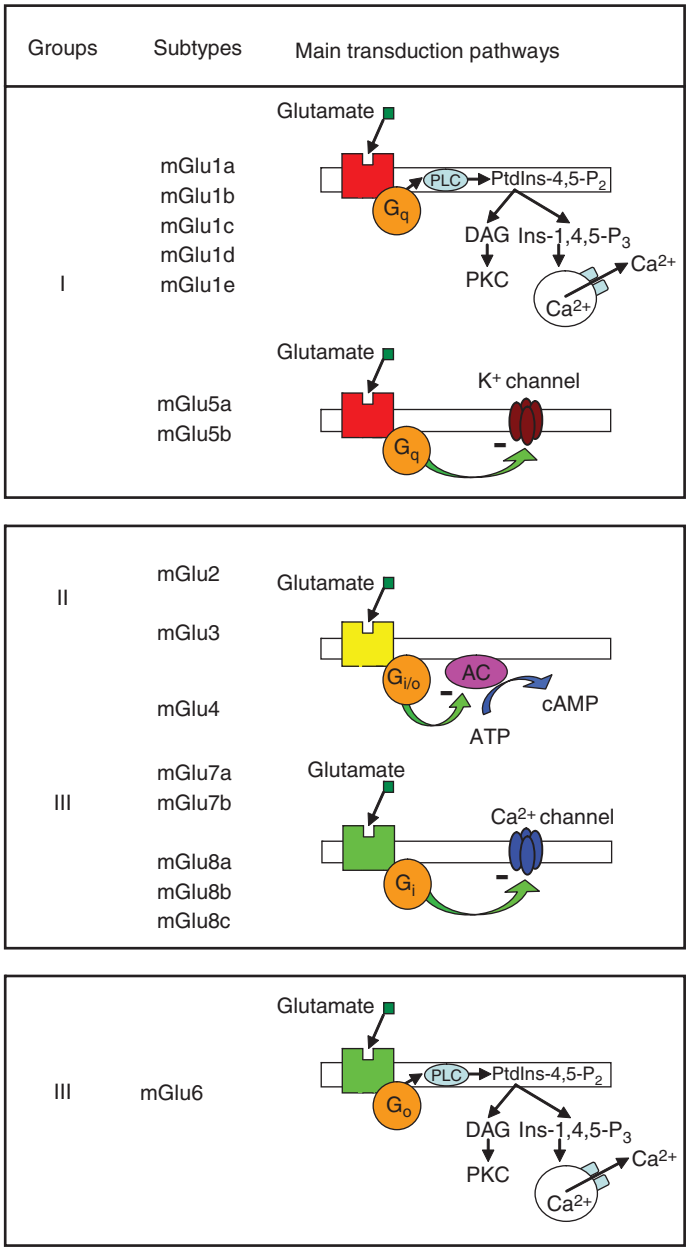

Figure 1 The mGlu receptor family. PtdIns-4,5-P_2, phosphatidylinositol 4,5-bisphosphate; Ins-1,4,5-P_3, inositol 1,4,5-trisphosphate; DAG, diacylglycerol; PKC, phosphokinase C; PLC, phospholipase C.

expressed parathyroid cells, the taste receptors that detect sweet and umami, and some pheromone receptors. mGlu receptors are formed by (1) a large N-terminus extracellular domain (about 600 amino acids) containing the glutamate binding site and a Cys-rich region, (2) a seven-transmembrane (7TM) domain, and (3) an intracellular C-terminus domain that varies in length depending on the receptor subtype/splice variant. mGlu1a, mGlu5a, and mGlu5b receptors have long C-terminus domains, which favor their interaction with adaptor and scaffolding proteins (see later). In the mGlu1b and -1d variants, the 313 C-terminus amino acids are substituted by 20 (mGlu1b) and 22/26 (mGlu1d human/rat) amino acids. The mGlu1e isoform corresponds to the soluble extracellular domain of the mGlu1 receptor. The mGlu5b variant differs from mGlu5a for the presence of 39 amino acids inserted 49 residues downstream of the seventh TM domain. The existence of truncated forms of mGlu3 and mGlu6 receptors, corresponding to the extracellular portion of the receptors, has been demonstrated in human tissue. In the mGlu7b variant, the last 16 amino acids of mGlu7a are substituted by 23 different residues.

The elegant report by Kunishima and colleagues in 2000 has afforded insights into the structure

and mechanism of activation of mGlu receptors. Glutamate binds to a 'venus fly trap' (VFT) region of the N-terminus domain constituted by two globular lobes separated by a hinge. The lobes oscillate between a close/active conformation and an open/inactive conformation. Agonist binding to the hinge facilitates the closure of the two lobes, whereas binding of competitive antagonists stabilizes the open conformation. Similarly to other class C GPCRs, mGlu receptors function as dimers. Dimerization involves a hydrophobic interaction between lobe I of the VFT in each monomer, and is stabilized by a disulfide bond. The use of receptor chimeras containing the C-terminus domains of type 1 and type 2 $GABA_B$ receptors (which form obligatory heterodimers) has allowed to establish that the mGlu receptor dimer can be activated by a single molecule of orthosteric agonist, but that binding of two molecules (one to each monomer) is required for full receptor activation. In contrast, it appears that only one 7TM domain is required for effective G-protein activation. The 7TM domain contains the binding sites for positive allosteric modulators ('potentiators' or 'enhancers'), which amplify receptor activation only in the presence of an orthosteric agonist – that is, in an activity-dependent manner. One enhancer molecule per dimer is needed for full amplification of mGlu receptors. The 7TM domain also bears the binding sites for negative allosteric modulators, which inhibit mGlu receptors independently of the concentrations of ambient glutamate.

mGlu receptors interact with a variety of adaptor and scaffolding proteins, which regulate receptor expression and coupling with other receptors or enzymes. Homer proteins, which contain both postsynaptic density/disc/zonula occludens (PDZ) and Ena/vasodilator-stimulated phosphoprotein homology-1 (EVH-1) domains, bind to a proline-rich motif in the C-terminal region of mGlu1a and mGlu5 receptors. Constitutively expressed, long isoforms of Homer proteins (Homer 1b, -1c, -2, and -3) have a C-terminal coiled-coil domain that mediates self-multimerization and allows the interaction of mGlu1a/5 receptors with other Homer-interacting proteins, such as the IP_3 receptor (Figure 2(a)). This interaction is disrupted by the short inducible isoform of Homer (Homer 1a), which lacks the coiled-coil domain. Interestingly, Homer 1a is encoded by an early inducible gene, which is expressed in response to synaptic activation. The interaction between Homer and the long isoform of the PI3K enhancer, PIKE, mediates the stimulation of PI3K by mGlu1 receptor activation. mGlu1 receptors functionally interact with ephrin-B2, a member of the ephrin/Eph receptor family of transmembrane proteins which mediate processes of cell-to-cell interaction during development and in the adult life. This interaction might have interesting implications for processes of developmental plasticity. Other proteins interacting with group I mGlu receptors are the E3 ubiquitin ligase, Seven in absentia homolog 1a, protein phosphatase 1c, tubulin, and tamalin. Protein phosphatase 2C binds selectively to, and dephosphorylates, mGlu3 receptors. mGlu7 receptors interact with calmodulin, protein interacting with PKC 1 (PICK 1), glutamate receptor-interacting protein 1 (GRIP 1), syntenin, α-tubulin, filamin A, and the catalytic γ subunits of protein phosphatase 1C. A simultaneous or exclusive binding to these proteins tightly regulates mGlu7 receptor signaling. mGlu8 receptors interact with sumo1 and other components of the sumoylation cascade, such as ube2a, Pias1, Piasγ, and Piasxβ.

Cell Biology, Pharmacology, and Implications for Human Pathology

The three groups of mGlu receptor ligands (Figure 3) have potential clinical applications, as outlined in Table 1.

mGlu1 Receptor

The mGlu1 receptor consists of 1194 amino acids in humans and 1199 amino acids in rats and mice; accession numbers are Q13255 (human), P23385 (rat), and P97772 (mouse). The gene is named *GRM1* (human) and *Grm1* (rat, mouse); chromosomal locations are 6q24 (human), 1p13 (rat), and 10 A2 (mouse).

At a subcellular level, mGlu1 proteins are mostly found in postsynaptic elements at the periphery of the postsynaptic density (PSD). This contrasts with the distribution of *N*-methyl-D-aspartate (NMDA) and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors, which are localized in the central region of the PSD. mGlu1 receptors are expressed by Purkinje cells of the cerebellar cortex, where they participate in the induction of long-term depression (LTD) at the parallel fiber–Purkinje cell synapse. LTD is a particular form of synaptic plasticity that, in the cerebellum, underlies motor learning. Mice with genetic deletion of mGlu1 receptors are ataxic and show persistent multiple climbing fiber innervation of cerebellar Purkinje cells. Interestingly, anti-mGlu1 receptor antibodies are found in a subset of patients with Hodgkin's lymphoma and paraneoplastic ataxia. The mGlu1 receptor is also expressed in the olfactory bulb, pars compacta of the substantia nigra, hippocampus, and thalamic nuclei. In the hippocampus, activation of mGlu1 receptors inhibits GABA release and mGlu1 receptor antagonists are protective against excitotoxic and hypoxic/ischemic neuronal death. Unexpectedly, that

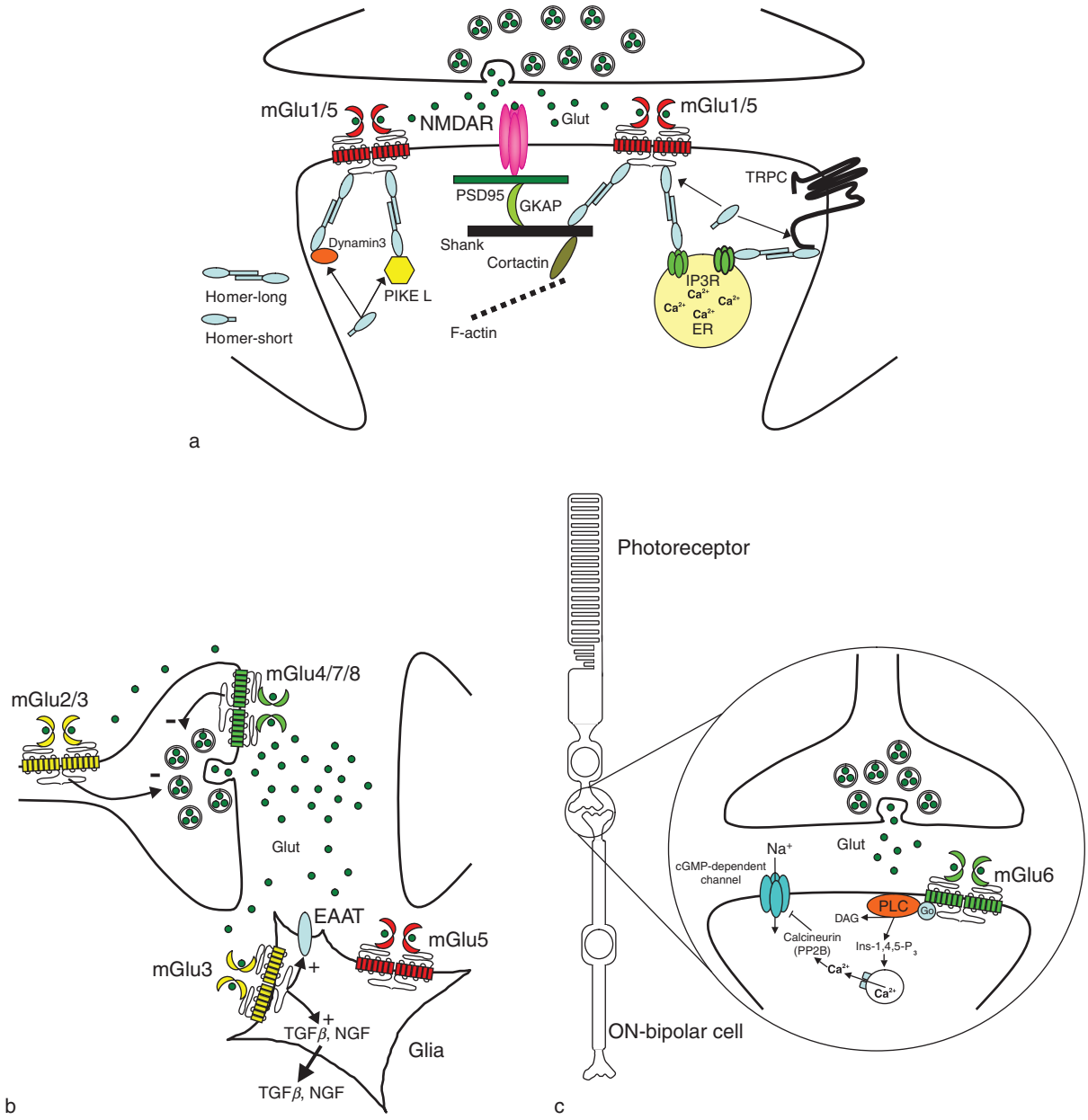

Figure 2 Synaptic localization of mGlu receptors. (a) mGlu1 and mGlu5 receptors are found at the periphery of postsynaptic elements. mGlu1a, mGlu5a, and mGlu5b receptors interact with scaffolding and adaptor proteins, including Homer. (b) mGlu2/3 and mGlu4/7/8 receptors are predominantly found in presynaptic terminals, where their activation inhibits neurotransmitter release; mGlu3 and mGlu5 receptors are also found in glial cells. (c) Activation of mGlu6 receptors by the glutamate released from photoreceptor cells inhibits the activity of ON bipolar cells. Glut, glutamate; IP3R, inositol 1,4,5-trisphosphate receptor; ER, endoplasmic reticulum; EAAT, excitatory amino acid transporter; PP2B, protein phosphatase 2B; TRPC, transient receptor potential-C. (a) Adapted from Fourgeaud L (2005) Addicted to Holmer? *Journal of Neuroscience* 25(42): 9555–9556.

ectopic expression of mGlu1 receptors in melanocytes promotes the development of multiple melanomas in mice. Because mGlu1 receptors are found in human melanomas but not in benign nevi, it is possible that receptor activation contributes to neoplastic transformation of melanocytes.

There are no selective agonists of mGlu1 receptors. 3,5-dihydroxyphenylglycine (DHPG) activates both mGlu1 and mGlu5 receptors, although it has no activity at other glutamate receptor subtypes. The following drugs are examples of competitive mGlu1 receptor antagonists (in the rank order of affinity): LY367385 = 4C3HPG > 4CPG > AIDA. Compounds Ro01–6128 and Ro67–7476 are mGlu1 receptor enhancers with nanomolar potency, whereas BAY36–7620 and CPCCOEt are negative allosteric

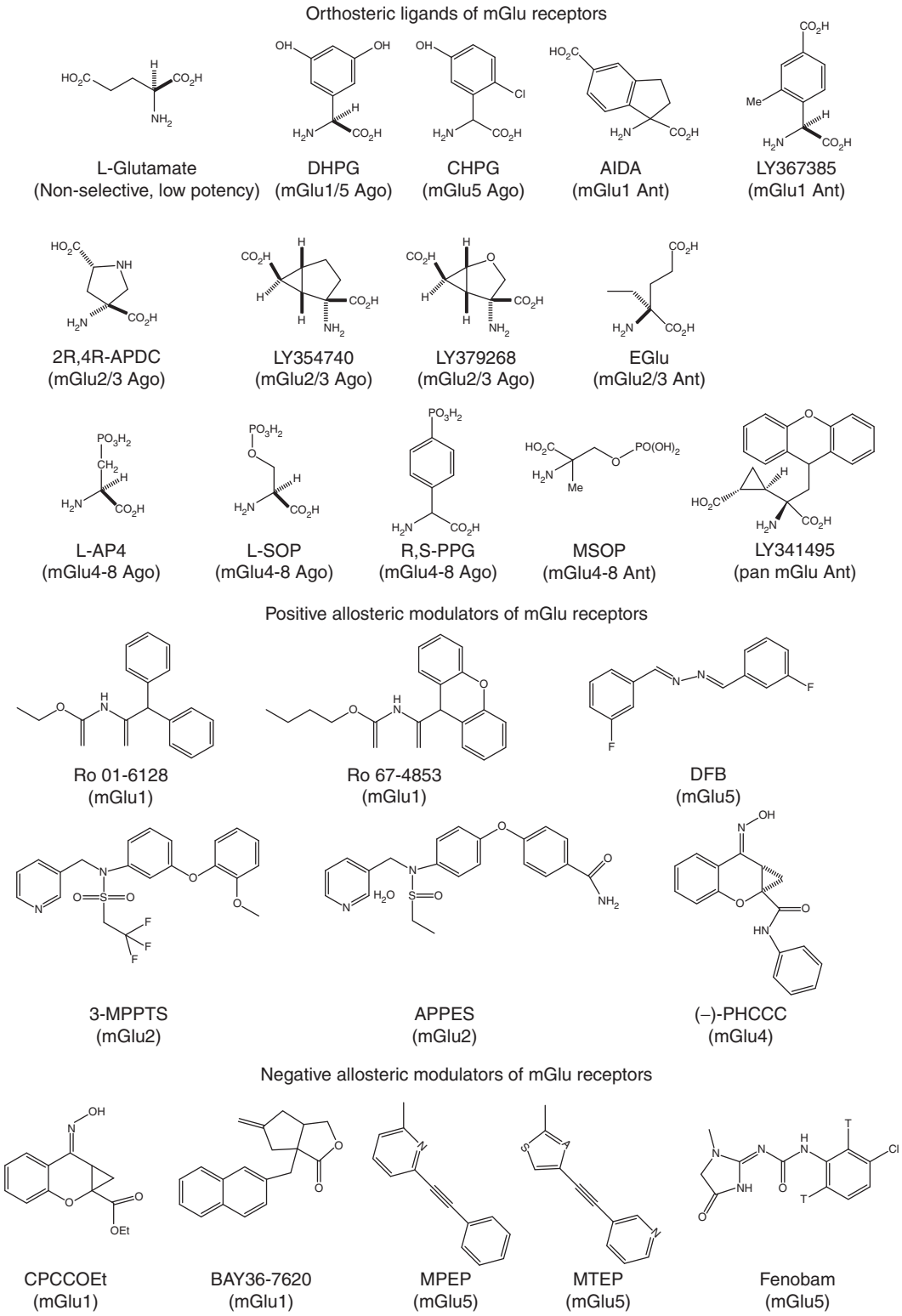

Figure 3 Selected mGlu receptor ligands. Ago, agonist; Ant, competitive antagonist.

modulators of mGlu1 receptors. (Abbreviations: LY367385, (*E*)-2-methyl-6-stryrylpyridine(–)-2-oxa-4-aminocyclo[3.1.0]exane-4,6-dicarboxylic acid; 4C3HPG, 4-carboxy-3-hydroxyphenylglycine; 4CPG, 4-carboxyphenylglycine; AIDA, (*R*,*S*)-1-aminoindan-1,5-dicarboxylic acid; Ro 01–6128, ethyl diphenylacetylcarbamate; Ro 67–7476, (*S*)-2-(4-fluorophenyl)-1-(toluene-4-sulfonyl)pyrrolidine;

Table 1 Potential clinical applications of mGlu receptor ligands

Group	*Application*
Group I	Chronic pain (mGlu1/5, Ant/NAM)
	Anxiety disorders (mGlu5, Ant/NAM)
	Drug addiction (mGlu5, Ant/NAM)
	Fragile X syndrome (mGlu5, Ant/NAM)
	Parkinsonism (mGlu5, Ant/NAM)
	Schizophrenia (mGlu5, PAM)
	Brain ischemia (mGlu1, Ant/NAM)
Group II	Generalyzed anxiety disorders and panic attack (mGlu2/3, Ago/PAM)
	Drug addiction (mGlu2/3, Ago/PAM)
	Schizophrenia (mGlu2/3, Ago/PAM)
	Chronic pain (mGlu2/3, Ago/PAM; mGlu2 inducers)
Group III	Epilepsy (mGlu4/7/8, Ago/PAM)
	Parkinsonism (mGlu4, Ago/PAM)
	Anxiety (mGlu8, Ago/PAM)

Ago, agonist; Ant, competitive antagonist; PAM, positive allosteric modulator; NAM, negative allosteric modulator.

CPCCOEt, 7-(hydroxyimino)cyclopropa[*b*]chromen-1a-carboxylate ethyl ester.)

mGlu5 Receptor

The mGlu5 receptor consists of 1212 amino acids in humans and 1203 amino acids in rats; accession numbers are P41594 (human) and P31424 (rat). The gene is named *GRM5* (human) and *Grm5* (rat, mouse); chromosomal locations are 11q14.2 (human), 1q32 (rat), and 7 D3 (mouse).

Activation of recombinantly expressed mGlu5 receptors induces oscillatory increases in intracellular Ca^{2+} release. This property, which is not shared by mGlu1 receptors, relies on the presence of a Ser residue in the C-terminus domain, which is phosphorylated by PKC, and places the mGlu5 receptor at the core of basic processes in cell biology. Similar to the mGlu1 receptor, the mGlu5 receptor is localized in postsynaptic elements, where it is physically linked to NMDA receptors via a chain of interacting proteins, which include PSD-95, Shank, guanylate kinase-associated protein (GKAP), and Homer (**Figure 2(a)**). Activation of mGlu5 receptors facilitates the opening of NMDA-gated ion channels, and activation of NMDA receptors amplifies mGlu5 receptor function by inhibiting receptor desensitization. Because of their functional interaction with NMDA receptors, mGlu5 receptors are involved in the induction of long-term potentiation (LTP), a putative electrophysiological substrate of associative learning. mGlu5 knockout mice show a defective LTP in the hippocampus and an impaired spatial learning. Activation of mGlu5 receptors also mediates a non-NMDA-dependent form of LTD in the hippocampus.

mGlu5a and mGlu5b variants show an opposite developmental pattern of expression in the central nervous system (CNS). mGlu5a receptors are abundantly expressed early after birth, and mediate the robust PI response to excitatory amino acids in the first 2 weeks of postnatal life. Expression of mGlu5b receptors increases with age and is prominent in the adult hippocampus, striatum, and cerebral cortex. mGlu5 receptors are also found in peripheral cells, including thymocytes, hepatocytes, melanocytes, cells of the male germinal line, osteoblasts, and insulinoma cell lines, where their function begins to be explored. Remarkably, mGlu5 receptors are present in both embryonic and neural stem cells, and their activation supports proliferation and self-renewal of these cells. A dysfunction of mGlu5 receptors has been associated with the fragile X syndrome, the most frequent form of inherited mental retardation. This syndrome is caused by the absence of the fragile X mental retardation protein (FMRP), an RNA-binding protein which functions as a negative regulator of protein synthesis. LTD mediated by mGlu5 receptors is enhanced in mice lacking FMRP, and pharmacological blockade of mGlu5 receptors partially corrects neurological and behavioral abnormalities in these mice.

mGlu5 receptors are selectively activated by 2-chlorohydroxyphenylglycine (CHPG), which, however, is a relatively weak agonist. No selective competitive antagonists are currently available. 2-Methyl-6-(phenylethynyl)pyridine (MPEP), 6-methyl-2-(phenylazo) pyridin-3-ol (SIB-1757), and (*E*)-2-methyl-6-stryrylpyridine (SIB-1893) are prototypes of a growing list of negative allosteric modulators of mGlu5 receptors, which includes 3-[(2-methyl-1,3-thiazol-4-yl)ethynyl] pyridine (MTEP) and the anxiolitic drug, fenobam. Compounds 3,3'-difluorobenzaldazine (DFB), *S*-(4-fluorophenyl)-{3-[3-(4-fluorophenyl)-[1,2,4] oxadiazol-5-yl]-piperidin-1-yl}methanone (AD X47273), 3-cyano-*N*-(1,3-diphenyl-1*H*-pyrazol-5-yl) benzamide (CDPPB), and *N*-{4-chloro-2-[(1,3-dioxo-1,3-dihydro-2*H*-isoindol-2-yl) methyl]phenyl}-2-hydroxybenzamide (CPPHA) are example of mGlu5 receptor enhancers which amplify receptor function in an activity-dependent manner. mGlu5 receptor ligands are of potential interests for the treatment of psychiatric and neurological disorders, such as anxiety, schizophrenia, parkinsonism, and neuropathic pain. Systemic administration of MPEP produces anxiolytic effects in a number of tests, including fear-potentiated startle, auditory and contextual fear conditioning, the elevated plus maze, conflict tests, and stress-induced hyperthermia. These effects are mediated by mGlu5 receptor blockade in the hippocampus and in the central nucleus of the amygdala, two regions that are actively involved in the acquisition and expression of fear conditioning. Pharmacological blockade or genetic

deletion of mGlu5 receptors disrupts prepulse inhibition of the acoustic startle reflex in rodents, thus mimicking the deficit in sensorimotor gating typical of schizophrenic patients. Administration of MPEP also enhances the cognitive deficit induced by the psychotomimetic drug, phencyclidine. Hence, mGlu5 receptor enhancers are currently developed as potential antipsychotic agents.

mGlu5 receptors are highly expressed in all stations of the basal ganglia motor circuit, and regulate the 'indirect pathway' of motor control by opposing the action of dopamine released from nigrostriatal terminals. The activity of this pathway is negatively regulated by the dopamine released from nigrostriatal fibers, which acts on D_2 inhibitory receptors located on medium-size spiny GABAergic neurons that project to the external portion of the globus pallidus (GP_e). In striatal projection neurons, mGlu5 receptors physically interact with A_{2A} adenosine receptors, and act synergistically with them in antagonizing the inhibitory control exerted by D_2 receptors. In addition, activation of mGlu5 receptors inhibits GP_e neurons by a mechanism of cross-desensitization with mGlu1 receptors, and stimulates neurons of the subthalamic nucleus. Pharmacological blockade or genetic deletion of mGlu5 receptors protects mice against nigrostriatal damage produced by the parkinsonian toxin, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP), or the psychostimulant, methamphetamine. This raises the attractive possibility that mGlu5 receptor antagonists/negative modulators behave as antiparkinsonian agents by relieving motor symptoms and reducing the ongoing degeneration of nigrostriatal neurons at the same time.

mGlu5 and mGlu1 receptors are implicated in the pathophysiology of chronic pain. Both receptor subtypes are localized on peripheral unmyelinated sensory afferents, and their activation increases the sensitivity to noxious heat, a phenomenon termed 'thermal hyperalgesia.' Activation of mGlu5 receptors in peripheral nociceptors stimulates PI hydrolysis and arachidonic acid release from DAG, leading to prostaglandin formation. Autocrine or paracrine stimulation of prostanoid receptors by prostaglandins in nociceptors results into an increased sensitivity of vanilloid TRPV1 receptors, thus inducing thermal hyperalgesia. Group I mGlu receptors also contribute to the regulation of nociceptive transmission and plasticity in the dorsal horns of the spinal cord, and perhaps in other stations of the pain neuraxis. Antagonists/negative modulators of mGlu5 and mGlu1 receptors produce analgesic effects in models of inflammatory and neuropathic pain.

Finally, mGlu5 receptors have been implicated in the synaptic mechanisms underlying drug addiction. These receptors positively regulate brain reward function and drive drug consumption. Blockade of mGlu5 receptors by MPEP elevates the intracranial self-stimulation threshold, decreases cocaine and nicotine self-administration, and attenuates naloxone-induced somatic signs of morphine withdrawal.

mGlu2 and mGlu3 Receptors

The mGlu2 receptor consists of 872 amino acids in humans, rats, and mice; accession numbers are Q14416 (human) and P31421 (rat). The gene isnamed *GRM2* (human) and *Grm2* (rat, mouse); chromosomal locations are 3q21.31 (human), 8q32 (rat), and 9 (mouse).

The mGlu3 receptor consists of 877 amino acids in humans and 879 amino acids in rats and mice; accession numbers are Q14832 (human), P31422 (rat), and Q9QYS2 (mouse). The gene is named *GRM3* (human) and *Grm3* (rat, mouse); chromosomal locations are 7q21.1–q21.2 (human), 4q32 (rat), and 5A1-h (mouse).

mGlu2 and mGlu3 are highly homologous and have a similar functional and pharmacological profile. Both receptors are predominantly localized in the preterminal region of the axon, and, if present in glutamatergic nerve terminals, can only be activated by amounts of glutamate sufficient to escape the clearance mechanisms around the synapses (**Figure 2(b)**). Activation of group II mGlu receptors attenuates neurotransmitter release as a result of a reduced cAMP formation and inhibition of voltage-sensitive Ca^{2+} channels. In brain slices, activation of mGlu2/3 receptors fails to stimulate PI hydrolysis *per se*, but amplifies the stimulation of PI hydrolysis mediated by group I mGlu receptors. The functional significance of this mechanism is unknown. mGlu3 receptors are also found in astrocytes, where their activation stimulates the MAPK and PI3K pathways and leads to the production of tropic factors (**Figure 2(b)**).

Both mGlu2 and mGlu3 receptors are activated by 2*R*,4*R*-4-aminopyrrolidine-2,4-dicarboxylic acid (APDC; (a selective agonist with micromolar affinity), and by the cyclopropan derivatives, 2*S*,2′*R*,3′*R*)-2-(2′,3′-dicarboxycyclopropyl)glycine (DCG-IV) and (2*S*,1′*S*,2′*S*)-2-(carboxycyclopropyl)glycine (L-CCG-I). Both drugs activate mGlu2 and mGlu3 receptors in the mid-nanomolar range; however, DCG-IV is also an NMDA receptor agonist, whereas L-CCG-I can also activate group I mGlu receptors. (1*S*,2*S*,5R,6*S*)-(+)-2-Aminobicyclo[3.1.0]hexane-2,6-dicarboxylic acid (LY354740) and (–)-2-oxa-4-aminocyclo[3.1.0]hexane-4,6-dicarboxylic acid (LY379268) are conformationally constrained glutamate analogs in which the glutamate backbone is locked into a fully extended state by incorporation into a bicycle[3.1.0]

hexane ring system. Both compounds are potent and selective mGlu2/3 receptor agonists with affinity in the low nanomolar range, and are systemically active. LY354740 and its prodrugs are currently under clinical development for the treatment of anxiety and panic disorders (see later). (1*R*,2*S*,5*S*,6*S*)-2-Amino-6-fluoro-4-oxobicyclo[3.1.0]hexane-2,6-dicarboxylic acid monohydrate (MGS0028) is another potent and selective agonist of mGlu2/3 receptors. Compounds (2*S*,1′*S*,2′*S*)-2-(9-xanthylmethyl)-2-(2′-carboxycyclopropyl)glycine (LY341495) and (1*R*,2*R*,3*R*,5*R*,6*R*)-2-amino-3-(3,4-dichlorobenzyloxy)-6-fluorobicyclo[3.1.0]hexane-2,6-dicarboxylic acid (MGS0039) are mGlu2/3 receptor antagonists with nanomolar affinity. However, LY341495 recruits other mGlu receptor subtypes at micromolar concentrations and is considered as a pan-mGlu receptor antagonist. There are only a few tools that can help discriminate mGlu2 from mGlu3 receptors. No subtype-selective mGlu2 receptor agonists are available. *N*-Acetyl-aspartyl-glutamate (NAAG) is an endogenous compound that activates mGlu3 receptors and has no activity at mGlu2 receptors. However, NAAG displays a low potency and is also active at NMDA receptors. A growing number of positive allosteric modulators, such as 2,2,2-trifluoro-*N*-[3-(2-methoxyphenoxy)phenyl]-*N*-(3-pyridinylmethyl)-ethanesulfonamide (3-MP PTS) and 2,2,2-trifluoro-*N*-[3-(cyclopentyloxy)phenyl]-*N*-(3-pyridinylmethyl)-ethanesulfonamide (cyPPTS), selectively amplify responses mediated by mGlu2 receptors. No mGlu3 receptor enhancers are currently available.

The development of mGlu2/3 receptor agonists or enhancers is aimed at the treatment of a number of psychiatric disorders, such as anxiety disorders, schizophrenia, and drug addiction. The agonist LY354740 is effective in all animal models of anxiety, and is also active in humans in preventing CO_2-induced anxiety and panic attacks and relieving symptoms of generalized anxiety disorder (GAD). Remarkably, pharmacological activation of mGlu2/3 receptors does not produce adverse effects, such as sedation, ataxia, and dependence, that are typical of benzodiazepines and other CNS depressants, although it may disrupt memory processing. mGlu2 and mGlu3 receptors are highly expressed in limbic regions, and may regulate transmission at the synapses between the basolateral nuclei and the central nucleus of the amygdala. A final effect of LY354740 appears to be an enhanced GABAergic activity within the central nucleus of the amygdala, as shown by an increased Fos protein expression in GABAergic neurons. It appears that both mGlu2 and mGlu3 receptors are necessary for the anxiolytic effect of LY354740, although selective mGlu2 enhancers can also relieve anxiety.

mGlu2/3 receptor agonists or mGlu2 receptor enhancers are active in models predictive of antipsychotic activity, including the disruption of working memory, motor activity, sensory–motor gating (only mGlu2 enhancers), and glutamate efflux induced by phencyclidine or amphetamine. In addition, there is a functional antagonism between mGlu2/3 receptors and serotonin 5-HT_{2A} receptors, which mediate the psychotomimetic action of LSD and other hallucinogenic drugs. Thus, there is a common pathway for the action of mGlu2/3 receptor agonists and atypical antipsychotic drugs, such as clozapine or olanzapine, which act as 5-HT_{2A} receptor antagonists. Although there is no association of polymorphism in the mGlu2 receptor gene with schizophrenia, an increased expression of mGlu2/3 receptors is found in the prefrontal cortex of schizophrenic patients. Thus, one expects that the brain of schizophrenic patients is highly responsive to drugs that activate mGlu2/3 receptors.

mGlu3 receptors are present in astrocytes and microglia, where their activation stimulates the synthesis of neurotropic factors, such as transforming growth factor-β (TGF-β), nerve growth factor (NGF), and brain-derived neurotropic factor (BDNF). An increased secretion of TGF-β from astrocytes mediates the neuroprotective activity of group II mGlu receptor agonists in culture. mGlu3 receptors are also found in neural stem cells and human glioma cells, and their blockade inhibits cell proliferation by dampening the stimulation of the MAPK and the PI3K pathways.

Group II mGlu receptors have been recently shown to play a role in the synaptic adaptations that occur during the development of drug dependence. Activation of mGlu2/3 receptors negatively regulates brain reward function, as agonists elevate the intracranial self-stimulation threshold and decrease relapse to drug-taking behavior during abstinence, thus regulating the negative affective state observed during withdrawal.

Finally, mGlu2/3 receptors mediate analgesic effects via a variety of mechanisms, including a reduced sensitivity of TRPV1 channels in peripheral nociceptors, and a reduced neurotransmitter release at the synapses between primary afferent fibers and second-order neurons in the dorsal horn of the spinal cord. L-Acetylcarnitine, a drug currently used for the treatment of painful neuropathies, induces the expression of mGlu2 receptors in dorsal root ganglia neurons by amplifying the activity of transcription factors of the nuclear factor-kappa B (NF-κB) family.

mGlu4, mGlu7, and mGlu8 Receptors

The mGlu4 receptor consists of 912 amino acids in humans, rats, and mice; accession numbers are Q14833 (human), P31423 (rat), and Q68EF4 (mouse). The gene is named *GRM4* (human) and

Grm4 (rat, mouse); chromosomal locations are 6p21.3 (human), 20p12 (rat), and 17 A3.3 (mouse).

The mGlu7 receptor consists of 915 amino acids in humans, rats, and mice; accession numbers are Q14831 (human), P35400 (rat), and Q68ED2 (mouse). The gene is named *GRM7* (human) and *Grm7* (rat, mouse); chromosomal locations are 3p26.1 (human), 4q42 (rat), and 6 E3 (mouse).

The mGlu8 receptor consists of 908 amino acids in humans, rats, and mice; accession numbers are O00222 (human), P70579 (rat), and P47743 (mouse). The gene is named *GRM8* (human) and *Grm8* (rat, mouse); chromosomal locations are 7q31.3–q32 (human), 4q22 (rat), and 6 A3 (mouse).

These three receptor subtypes are preferentially, although not exclusively, located in the terminal region of the axons and share the ability to inhibit neurotransmitter release at various synapses in the CNS. Their location in the vicinity of the active zones of neurotransmitter release makes these subtypes as putative glutamate autoreceptors (Figure 2(b)). L-2-Amino-4-phosphonobutanoate (L-AP4) and the endogenous compound L-serine-*O*-phosphate (L-SOP) are prototypical orthosteric agonists of all group III mGlu receptors. (+)-4-Phosphonophenylglycine (4-PPG) activates receptors with the following rank order of potency: mGlu8 > mGlu4 >> mGlu7. (*S*)-3,4-Dicarboxyphenylglycine (DCPG) is an agonist with greater affinity for mGlu8 than for mGlu4, whereas (1*S*,3*R*,4*S*)-1-aminocyclopentane-1,3,4-tricarboxylic acid (ACPT-I) activates mGlu4 and mGlu8 receptors with equal potency. Neither DCPG nor ACPT-I has any effect at mGlu7 receptors. (*R*,*S*)-α-Cyclopropyl-4-phosphonophenylglycine (CPPG) and (*S*)-α-methyl-2-amino-4-phosphonobutanoic acid (M AP-4) are competitive antagonists with higher affinity for mGlu8 receptors. CPPG has no apparent activity at mGlu7 receptors. (–)-*N*-Phenyl-7-(hydroxyimino) cyclopropa[*b*]chromen-1a-carboxamide (PHCCC) is an enhancer with high selectivity for mGlu4 receptors. At high concentrations, PHCCC may also recruit mGlu1 receptors because of its similarity with CPCCOEt (see earlier). A selective agonist of mGlu7 receptors, *N*,*N*′-dibenzhydrylethane-1,2-diamine dihydrochloride (AM N082), has recently been synthesized. This compound is atypical because it does not bind to the VFT (like all classical mGlu receptor agonists), but to a site located in the 7TM domain of mGlu7 receptors.

Activation of mGlu4 receptors is neuroprotective and mGlu4 receptor agonists/enhancers have been proposed as potential therapeutic agents in Parkinson's disease for their ability to reverse parkinsonian symptoms in animal models. This action appears to be mediated by the inhibition of GABA release in the GP_e, which lies along the 'indirect pathway' of the basal ganglia motor circuit. Activation of mGlu4 receptors is also protective against experimental parkinsonism induced by MPTP in mice. mGlu4 receptors are highly expressed by cerebellar granule cells, and are also found in medulloblastoma cells, which derive from granule cell neuroprogenitors. Interestingly, activation of mGlu4 receptors inhibits proliferation of medulloblastoma cells and limits the spontaneous growth of medulloblastomas in the cerebellum of tumor-prone mice.

The function of mGlu7 and mGlu8 receptors is only partially known. Both receptors have been proposed as drug targets in treatment of anxiety and stress-related disorders, but there are some caveats. The mGlu7 receptor agonist, AMN082, relieves anxiety, but genetic deletion of mGlu7 receptors does the same. In contrast, mGlu8 knockout mice show an increased level of anxiety, similar to that found in animals subjected to stressful conditions.

Finally, mGlu4, mGlu7, and mGlu8 receptors have been considered as potential targets for the experimental therapy of epilepsy. Drugs that activate mGlu4 or mGlu8 receptors reduce generalized seizures in epilepsy-prone mice and rats and chemically induced seizures. In addition, mice lacking mGlu7 or mGlu8 receptors show an increased susceptibility to chemically induced seizures. Expression of mGlu4 receptors is increased in the dentate gyrus of mice with a low susceptibility to seizures and of patients with temporal lobe epilepsy. In contrast, activation of mGlu4 receptors is required for the induction of absence seizures in mice.

mGlu6 Receptor

The mGlu6 receptor consists of 877 amino acids in humans and 871 amino acids in rats and mice; accession numbers are O15303 (human), P35349 (rat), and Q5NCH9 (mouse). The gene is named *GRM6* (human) and *Grm6* (rat, mouse); chromosomal locations are 5q35 (human), 10 (rat), and 11 B1.3 (mouse).

This is a peculiar receptor; it is exclusively localized in the dendrites of ON bipolar cells of the retina and responds to the glutamate released from rod and cone photoreceptor cells in the dark. Activation of mGlu6 receptors leads to hyperpolarization of ON bipolar cells by closing cGMP-dependent cation channels. This effect may involve a receptor coupling with G_o and activation of the Ca^{2+}-dependent protein phosphatase, calcineurin (Figure 2(c)). Light reduces glutamate stimulation of mGlu6 receptors, leading to opening of cation channels and depolarization of ON bipolar cells. Mice lacking mGlu6 receptors show a loss of ON response but unchanged responses to light. Interestingly, mutations in the gene encoding

the mGlu6 receptor are associated with night blindness and abnormal electroretinogram ON responses in humans. The mGlu6 receptor shows the same pharmacological profile of other group III mGlu receptors, being activated by L-AP4, L-SOP, and 4-PPG, and antagonized by CPPG, MSOP, and MAP-4.

See also: Glutamate Receptor Organization: Ultrastructural Insights; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP); Metabotropic Glutamate Receptors (mGluRs): Functions.

Further Reading

Alexander GM and Godwin DW (2006) Metabotropic glutamate receptors as strategic targets for the treatment of epilepsy. *Epilepsy Research* 71: 1–22.

Bear MF, Huber KM, and Warren ST (2004) The mGluR theory of fragile X mental retardation. *Trends in Neurosciences* 27: 370–377.

Bruno V, Battaglia G, Copani A, et al. (2001) Metabotropic glutamate receptor subtypes as targets for neuroprotective drugs. *Journal of Cerebral Blood Flow and Metabolism* 21: 1013–1033.

Conn PJ, Battaglia G, Marino MJ, et al. (2005) Metabotropic glutamate receptors in the basal ganglia motor circuit. *Nature Review Neuroscience* 6: 787–798.

DeBlasi A, Conn PJ, Pin J, et al. (2001) Molecular determinants of metabotropic glutamate receptor signaling. *Trends in Pharmacological Sciences* 22: 114–120.

Fourgeaud L (2005) Addicted to Holmer? *Journal of Neuroscience* 25(42): 9555–9556.

Kenny PJ and Markou A (2004) The ups and downs of addiction: Role of metabotropic glutamate receptors. *Trends in Pharmacological Sciences* 25: 265–272.

Kunishima N, Shimada Y, Tsugji Y, et al. (2000) Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. *Nature* 407: 971–977.

Pollock PM, Cohen-Solal K, Sood R, et al. (2003) Melanoma mouse model implicates metabotropic glutamate signaling in melanocytic neoplasia. *Nature Genetics* 34: 108–112.

Nakanishi S (1992) Molecular diversity of glutamate receptors and implication for brain function. *Science* 258: 597–603.

Nicoletti F, Wroblewski J, Iadarola MJ, et al. (1986) Excitatory amino acid recognition sites coupled with inositol phospholipid metabolism: Developmental changes and interaction with alpha 1-adrenoceptors. *Proceedings of the National Academy of Sciences of the United States of America* 83: 1931–1935.

Pin JP, Kniazeff J, Liu J, et al. (2004) Allosteric functioning of dimeric class C G-protein-coupled receptors. *FEBS Journal* 272: 2947–2955.

Schoepp DD, Jane DE, and Monn JA (1999) Pharmacological agents acting at subtypes of metabotropic glutamate receptors. *Neuropharmacology* 38: 1431–1476.

Sladeczek F, Pin JP, Recasens M, et al. (1985) Glutamate stimulates inositol phosphate formation in striatal neurones. *Nature* 317: 717–719.

Sugiyama H, Ito I, and Hirono C (1987) A new type of glutamate receptor linked to inositol phospholipid metabolism. *Nature* 325: 531–533.

Swanson CJ, Bures M, Johnson MP, et al. (2005) Metabotropic glutamate receptors as novel targets for anxiety and stress disorders. *Nature Review Drug Discovery* 4: 131–144.

Varney MA and Gereau RW 4th (2002) Metabotropic glutamate receptor involvement in models of acute and persistent pain: Prospects for the development of novel analgesics. *Current Drug Targets CNS Neurological Disorders* 1: 283–296.

Relevant Website

http://www.iuphar-db.org – IUPHAR Receptor Database.

Metal Accumulation during Aging

H M Schipper, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Central nervous system (CNS) senescence has been implicated as the most robust risk factor for a host of human neurodegenerative disorders, including sporadic (nonfamilial) Alzheimer's disease (AD), Parkinson's disease (PD), and amyotrophic lateral sclerosis (ALS). Thus, a thorough understanding of intrinsic brain aging mechanisms and their relationship to intervening neuropathological processes is essential for the development of therapeutic modalities that successfully prevent, ameliorate, interrupt, or reverse neuronal attrition and clinical decline in these conditions. Over the past two decades, oxidative stress (OS) and free radical damage, altered patterns of transition metal (especially iron) deposition, and mitochondrial insufficiency (bioenergetic failure) have been abundantly documented in the aging and degenerating mammalian CNS. Although the tendency has been to investigate these changes as disparate histopathological features of neural senescence and disease, we hypothesize that OS, abnormal iron mobilization, and mitochondrial dysfunction combine inextricably to yield a single neuropathological 'lesion' that serves as a vital link between normal brain aging physiology and neurodegeneration. In this article, support for this unifying perspective is reviewed based on evidence adduced from investigations of rodent astroglial senescence, experimental models of aging-associated neurodegenerative disorders, and human neuropathological studies. Several scientific and clinical implications of this formulation are considered.

Free Radicals, Transition Metals, and OS

Free radicals are atoms containing unpaired electrons in their outer orbitals. In the course of normal aging, and to a greater extent in disease states, free radicals and other reactive oxygen species (ROS), reactive nitrogen species (RNS), and reactive sulfur species (RSS) generated endogenously or derived from environmental sources may overwhelm tissue enzymatic and nonenzymatic antioxidant defenses (a condition broadly referred to as 'OS') and incur molecular damage to various cellular substrates, including proteins, lipids, carbohydrates, and nucleic acids. Among the transition metals (d-block elements listed between Group IIA and Group IIB of the periodic table), iron, copper, manganese, mercury, and cadmium are redox-active and are capable of mediating significant injury upon deposition within neural tissues. Ferrous iron (Fe^{2+}) and cuprous copper (Cu^{+}), the predominant transition metals in mammalian tissues, play pivotal roles in cellular redox chemistry by reducing H_2O_2 to the highly cytotoxic hydroxyl ($OH^{\bullet}$) radical (Fenton catalysis) or by behaving as nonenzymatic peroxidases that bioactivate benign compounds into toxic free radical intermediates. Within the CNS, ROS in turn may promote aberrant cell signaling and electrophysiological derangements; bioenergetic failure; proteosomal dysfunction, protein aggregation, and inclusion formation; protoxin bioactivation; and neuronal synaptolysis, apoptosis, and necrosis. Pathological transition metal deposition and attendant OS constitute an important pathway for cellular dysfunction and death (and hence a potential therapeutic target) in a host of human neurological (Table 1) and medical disorders, and may contribute to the senescence-dependant decline in cognition and other CNS activities. This article focuses primarily on mechanisms of brain iron sequestration and the relationship of the latter to the prevailing 'free radical–mitochondrial' theory of aging in the context of human neurodegenerative disorders.

The Free Radical–Mitochondrial Theory of Aging

In 1956, Denham Harman first proposed a 'free radical' theory of aging which posits that "free radical reactions are involved in the aging changes associated with the environment, disease and intrinsic aging processes." The theory gained momentum upon recognition that mitochondria are both a leading source of ROS generation and a prime target of oxidative molecular damage in aging eukaryotic cells. The free radical–mitochondrial theory states that oxidative damage to mitochondria, instigated by intrinsic metabolic processes and/or environmental insults, leads to cascading infidelity of electron transport and a self-propagating spiral of increased ROS generation within the inner mitochondrial membrane. The latter, in turn, elicits bioenergetic deficits (ATP loss and failure of ATP-dependent processes) and progressive tissue aging. Evidence supporting this theory has been garnered from a broad range of experimental and observational studies involving intact animal and *in vitro* models that span the entire phylogenetic spectrum. The theory suggests that (1) any injury or disease that exacerbates mitochondrial ROS production may accelerate senescence of the

Table 1 Redox-active transition metals implicated in the etiopathogenesis of human nervous system conditions

Iron	Copper
CNS senescence	Wilson's disease
Cerebral hemorrhage	Menkes disease
Alzheimer's disease	Acquired copper deficiency
Parkinson's disease	Alzheimer's disease?
Amyotrophic lateral sclerosis	Creutzfeldt-Jakob disease?
Huntington's disease	Manganese
Progressive supranuclear palsy	Parkinsonism (manganism)
Multiple sclerosis	Hepatic encephalopathy?
Corticobasal degeneration	Mercury
Friedreich ataxia	Inorganic mercury poisoning
X-linked sideroblastic anemia with ataxia	Minamata disease
PANK-2 deficiency (formerly Hallervorden-Spatz)[a]	Cadmium
Aceruloplasminemia	Toxic encephalopathy
Neuroferritinopathy	Peripheral neuropathy
Superficial siderosis	
Restless legs syndrome	
HIV-1 encephalitis	
Hemochromatosis?	

[a]PANK-2, Pantothenate kinase type 2.
Adapted from Schipper HM (2004) Redox neurology: Visions of an emerging subspecialty. *Annals of the New York Academy of Sciences* 1012: 342–355, with permission.

affected tissue or organ and (2) normal, aging tissues compromised by a progressively unfavorable mosaic of healthy and ROS-generating mitochondria (termed 'heteroplasmy') may become increasingly vulnerable to aging-associated conditions such as atherosclerosis, neurodegeneration, and cancer.

Iron Deposition in the Aging CNS

Iron is indispensable for normal neural development and physiology and contributes to a diverse spectrum of cellular functions, including myelination, biogenic amine metabolism, electron transport, antioxidant enzyme activity, and cytokinesis. To facilitate these vital functions, while at the same time limiting the metal's propensity to mediate toxic Fenton reactions (*vide supra*), iron homeostasis in neural and peripheral tissues is tightly controlled by an orchestra of proteins governing its absorption, valence configuration, extracellular transport, cellular flux, chemical signaling activities, and intracellular storage (Table 2). Despite this elaborate regulation, iron progressively accumulates in the mammalian CNS as a function of advancing age, and a portion of the metal in the brain parenchyma and cerebrospinal fluid (CSF) is maintained in a redox-active state. In the normal human CNS, the sequestered metal exhibits an affinity for the basal ganglia, hippocampus, certain cerebellar nuclei, and other, largely subcortical, brain regions. In these aging brain loci, the iron deposits co-register with multiple markers of oxidative substrate modifications and genetic, biochemical, and morphological indices of enhanced mitochondrial damage. Moreover, there appears to be a fairly consistent exacerbation of these co-localizing pathological features in most, if not all, of the major aging-related human neurodegenerative disorders and in some inflammatory, ischemic, and infectious CNS conditions.

Table 2 Some key proteins implicated in mammalian iron homeostasis

Protein name	*Function*
ATP-binding cassette protein, ABC7	Nonheme Fe export from mitochondria(?)
ABC-me (mitochondrial erythroid)	Mitochondrial transport related to heme synthesis
Ceruloplasmin	Ferroxidase activity/cellular Fe export
Divalent metal transporter (DMT1)/ Nramp2	Membrane transport for Fe^{2+}
Duodenal cytochrome *b*	Ferric reductase (provides Fe^{2+} for DMT1 in duodenum)
Ferritin (H and L)	Cellular Fe storage
Ferrochelatase	Fe^{2+} insertion into protoporphyrin IX (heme synthesis)
Ferroportin 1/MTP1/Ireg1	Cellular Fe export
Frascati	Mitochondrial Fe transport
Frataxin	[Fe–S] cluster synthesis
Heme oxygenase (1 and 2)	Recycling of heme Fe/cell survival
Hemojuvelin (repulsive guidance molecule C)	Unknown/cardiac and striated muscle Fe metabolism(?)
Hepcidin	Inhibitor of intestinal Fe absorption and macrophage heme Fe recycling
Hephaestin	Ferroxidase activity/enterocyte Fe export
Hereditary hemochromatosis protein	Modulates tissue Fe uptake via interaction with TfR(?)
Iron regulatory protein (1 and 2)	Cellular Fe sensors/regulators of Fe uptake and storage proteins
Lactoferrin	Fe-binding protein/antibacterial and antiviral activities
Melanotransferrin	Unknown/Fe binding protein
Mitochondrial ferritin	Mitochondrial Fe storage(?)
Sideroflexin 1	Mitochondrial transport related to Fe metabolism
Steap3	Endosomal ferrireductase
Transferrin	Plasma Fe(III) carrier
Transferrin receptor (TfR1 and 2)	Membrane receptor for diferric-Tf

Adapted from Ponka P (2004) Hereditary causes of disturbed iron homeostasis in the central nervous system. *Annals of the New York Academy of Sciences* 1012: 267–281, with permission.

Iron Deposition in Alzheimer's and Parkinson's Diseases

Aging remains the most robust risk factor thus far identified for the development of sporadic (nonfamilial) AD and idiopathic PD, two of the most common and devastating neurodegenerations in aging populations of the industrialized world. AD is a dementing condition characterized by progressive neuronal degeneration, gliosis, and the accumulation of extracellular amyloid (senile plaques) and intraneuronal inclusions (neurofibrillary tangles) in discrete regions of the basal forebrain, hippocampus, and association cortices. Idiopathic PD is an extrapyramidal (movement) disorder featuring gradual attrition of dopaminergic neurons in the substantia nigra pars compacta, formation of intraneuronal fibrillar inclusions (Lewy bodies) in this cell population, and variable depletion of noradrenaline and serotonin in other brain stem nuclei. Oxidative damage and mitochondrial insufficiency have been amply documented in the affected brain of AD and PD patients and in transgenic mouse models of these conditions. Neuropathological and magnetic resonance imaging studies have demonstrated augmented iron deposition in the cerebral cortex and basal ganglia of AD and PD patients, respectively. In both the AD and PD brain, increased expression of the intracellular iron storage protein, ferritin, parallels the distribution of the excess iron and largely implicates nonneuronal (glial, endothelial) cellular compartments. In AD, moreover, redox-active iron co-localizes to senile plaques and neurofibrillary tangles and may contribute to the biogenesis of these hallmark inclusions via free radical mechanisms. Interestingly, local concentrations of transferrin-binding sites in these neurodegenerative conditions remain unchanged or vary inversely with the elevated iron stores. Thus, although important for normal iron delivery to most systemic tissues, the transferrin pathway of iron mobilization appears to contribute little to pathological brain iron trapping in these aging-related neurodegenerations. As discussed later, experiments conducted in our laboratory have implicated the enzyme heme oxygenase-1 (HO-1) as a putative transducer of aberrant mitochondrial iron trapping under conditions of OS that likely prevail in the senescent CNS and, to a greater extent, in AD- and PD-affected neural tissues. Other iron-mobilizing proteins, such as lactoferrin, melanotransferrin (p97), ceruloplasmin, and the mutant hereditary hemochromatosis protein (encoded by the *HFE* gene), may also participate in the pathogenesis of these common neurodegenerative diseases.

Iron Deposition in Other Human CNS Disorders

Although most vigorously investigated in AD and PD, OS, mitochondrial lesions, and disturbances in iron homeostasis have also been described in other aging-related human neurodegenerative conditions, including ALS, progressive supranuclear palsy, corticobasal degeneration, Huntington's disease, and aceruloplasminemia. This pathological triad may also compromise neural tissues affected by multiple sclerosis (MS), HIV-1 encephalitis, ischemic or hemorrhagic stroke, and cerebral trauma. The extent to which transferrin-dependent and -independent pathways contribute to pathological iron mobilization in these diverse conditions remains to be determined.

Hypothesis

On the basis of their frequent co-occurrence within the aging and diseased neuraxis, we posit that OS, augmented iron deposition, and mitochondrial insufficiency represent a single cohesive neuropathological 'lesion,' with the advent of one constituent obligating the appearance of the others. If confirmed, this formulation may (1) establish a novel conceptual bridge between normal brain senescence and neurodegeneration in the context of the free radical–mitochondrial theory of aging, (2) impact our understanding of the pathogeneses and regional predilections of various senescence-dependant human neurological afflictions, and (3) possibly inform strategies for therapeutic intervention that target the molecular pathways serving to bind the components of this tripartite lesion. In the following section, a model of glial senescence is presented that lends support to our central hypothesis and provides a useful experimental paradigm for delineation of key events favoring pathological iron deposition and mitochondrial injury in the aging and degenerating CNS.

Glial Iron Sequestration in Brain Aging and Disease

Astrocytes

Astrocytes, the most abundant and diversified of the neuroglia, subserve a broad range of adaptive functions. Under certain circumstances, however, astrocytes may mediate dystrophic effects within the CNS and thereby contribute to neurological decline. Examples of the latter include the bioactivation of proneurotoxins; the release of excitotoxic amino acids following tissue hypoxia, metal exposure, and OS; formation of epileptogenic scar tissue in response to

CNS injury; and neoplastic transformation and malignant behavior. Astrocyte hypertrophy, accumulation of glial fibrillary acidic protein (GFAP)-positive intermediate filaments, and possibly hyperplasia (reactive gliosis) are characteristic pathological features of the major aging-related neurodegenerative disorders including AD, PD, Huntington's disease, and ALS. Gliosis also occurs, to a lesser extent, in the course of normal brain aging. Increases in monoamine oxidase B (MAO-B) activity accompany reactive gliosis in the aging mammalian brain, in AD, and in experimental models of PD. Under these circumstances, neuronal injury may be fostered by excessive H_2O_2 derived from the accelerated oxidative deamination of dopamine (DA) and other monoamines.

Iron Deposition in Subcortical Astroglia

In the course of normal aging, a subpopulation of astrocytes residing within the mammalian hippocampus, basal ganglia and other subcortical brain regions progressively accumulates cytoplasmic granules replete with ferrous iron and occasionally other transition metals such as copper and chromium. For reasons discussed earlier, the accumulation of these iron-rich inclusions may render the senescent nervous system increasingly prone to free radical-mediated degeneration. We have determined that the gliosomes are derived from damaged, metal-laden mitochondria engaged in a complex macroautophagic process. In turn, the gliosomes may give rise to corpora amylacea (CA), glycoproteinaceous inclusions that accumulate in astrocytes and the extracellular space in subpial and periventricular regions of the aging and degenerating human brain. In rodents, the appearance of this senescent glial phenotype can be accelerated *in vivo* and/or *in vitro* by dietary iron and copper supplementation, cranial X-irradiation, mechanical trauma, cysteamine (CSH) administration, and, in certain hypothalamic nuclei, chronic estrogenization. We have also garnered evidence that intracellular OS is a 'final common pathway' leading to the transformation of normal astrocyte mitochondria to peroxidase-positive inclusions *in vitro* and in the intact aging brain. Thus, the pathological triad of OS, iron sequestration, and mitochondrial damage is fully manifest in the development of this senescent gliopathy. The topography and intensity of endogenous glial peroxidase activity (iron-laden mitochondria) may denote CNS regions that are particularly susceptible to OS during normal aging and under pathological conditions.

In rat astroglial cultures, the pro-oxidant effects of CSH, hydrogen peroxide, menadione, DA, β-amyloid$_{40/42}$, tumor necrosis factor-α (TNF-α), and interleukin-1β (IL-1β) promote the uptake of non-transferrin-derived $^{59/55}$Fe, but not diferric-transferrin-derived $^{59/55}$Fe, by the mitochondrial compartment. These findings are commensurate with the fact that pathological iron accumulation in the degenerating human CNS appears to be a transferrin-independent process. Moreover, the targeting of excess cellular iron to mitochondria may readily explain why bioenergetic failure is a ubiquitous feature of neural tissues exhibiting iron overload. As discussed in the following section, the heme-degrading enzyme, HO-1, may be a pivotal factor promoting the establishment of this aging-associated gliopathy.

Heme Oxygenase-1: Transducer of Mitochondrial Iron Deposition in 'Stressed' Astroglia

Heme Oxygenase

Heme oxygenases reside within the endoplasmic reticulum where they catalyze, in collaboration with reduced nicotinamide adenine dinucleotide phosphate (NADPH) cytochrome P450 reductase, the oxidative cleavage of heme to biliverdin, free ferrous iron, and carbon monoxide (CO). Biliverdin is metabolized further to the bile pigment, bilirubin, via biliverdin reductase (**Figure 1**). At least two isoforms of heme oxygenase, HO-1 (also called HSP32) and HO-2, are expressed in mammalian cells. Basal HO-1 expression in the normal rodent brain is restricted to small groups of neurons and neuroglia, whereas HO-2 protein is widely distributed throughout the neuraxis. Diverse response elements in the mammalian *ho-1* promoter and enhancer regions render the gene highly inducible by numerous pro-oxidant and inflammatory stimuli, including H_2O_2, transition metals, ultraviolet light, T helper cell (Th1) cytokines, prostaglandins, lipopolysaccharide, dopamine, β-amyloid, and the natural substrate, heme. By promoting the conformational modification and accelerated degradation of vulnerable tissue hemoproteins, SW Ryter and RM Tyrrel surmised that OS may transiently increase the intracellular 'free heme pool.' In these stressed cells, the upregulation of HO-1 may confer protection by accelerating the conversion of pro-oxidant heme to the radical-scavenging bile pigments, biliverdin and bilirubin. In many tissues overexpressing HO-1, potential toxicity resulting from the intracellular liberation of heme-derived ferrous iron is mitigated by co-induction of apoferritin synthesis, a major iron sequestration pathway. However, the Piantadosi lab and others have shown that heme-derived iron and CO may at times exacerbate intracellular OS and substrate damage by facilitating ROS

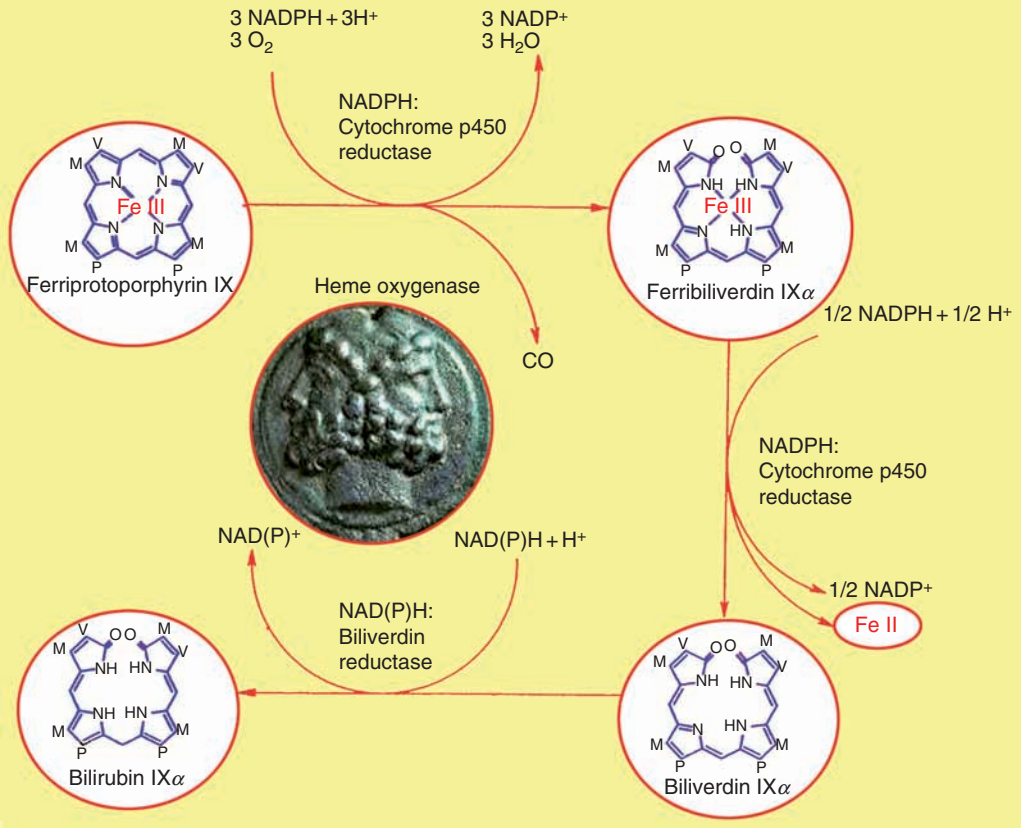

Figure 1 The heme catabolic pathway. The heme products – ferrous iron (Fe II), carbon monoxide, and biliverdin/bilirubin – may behave as either pro-oxidants or antioxidants, accounting for the disparate influences of heme oxygenase expression on cellular survival (symbolized by Janus faces). In the structures depicted, M is methyl, V is vinyl, and P is propionate. Adapted from Ryter SW and Tyrrell RM (2000) The heme synthesis and degradation pathways: Role in oxidant sensitivity. Heme oxygenase has both pro- and antioxidant properties. *Free Radical Biology & Medicine* 28: 289–309, with permission.

production within the mitochondrial and other subcellular compartments.

HO-1 and Mitochondrial Iron Sequestration

In rat astroglial cultures, we have observed that HO-1 mRNA, protein, and/or activity levels increase within hours of treatment with CSH, DA, β-amyloid$_{40/42}$, TNF-α, or IL-1β. After 3–6 days of exposure to these stimuli, uptake of non-transferrin-derived $^{59/55}$Fe (but not diferric transferrin-derived iron) is significantly augmented by the mitochondrial compartment relative to untreated control cultures. Coadministration of antioxidants (ascorbate, melatonin, or *trans*-resveratrol) suppresses the HO-1 response to CSH, DA, and the proinflammatory cytokines, indicating that OS is a common mechanism mediating glial *ho-1* gene induction under these experimental conditions. HO-1 overexpression appears necessary and sufficient for the mitochondrial iron trapping in oxidatively challenged astroglia because the latter can be (1) attenuated by tin mesoporphyrin (SnMP), a competitive inhibitor of heme oxygenase activity, or dexamethasone (DEX), a transcriptional suppressor of the *ho-1* gene and (2) recapitulated by transient transfection of the human (h) *ho-1* gene.

HO-1, Intracellular OS, and the Mitochondrial Permeability Transition

We have demonstrated that mitochondrial iron trapping by hHO-1-transfected glia and cells exposed to DA, TNF-α, or IL-1β is abrogated by co-treatment with cyclosporin A or trifluoperazine, inhibitors of the mitochondrial permeability transition pore. These findings suggest that, in 'stressed' astroglia, engagement of this membrane megachannel facilitates the import of cytosolic iron (and possibly other metal ions) to the mitochondrial matrix. We determined that (1) treatment with SnMP or antioxidants (ascorbate, melatonin or resveratrol) blocks the compensatory induction of the redox-sensitive *mnsod* gene in astrocytes challenged with DA or transiently transfected with hHO-1 cDNA and (2) hHO-1 transfection significantly augments levels of protein carbonyls (protein oxidation), 8-epiPGF2α (lipid peroxidation), and 8-OHdG

(nucleic acid oxidation) in glial mitochondrial fractions relative to sham-transfected controls and hHO-1-transfected cells incubated with SnMP. These findings indicate that HO-1 overexpression in astroglia amplifies intracellular OS that, in turn, favors mitochondrial permeability transition and iron influx into this organelle. Of note, Bruce Ames' group and U Rauen et al. have provided evidence that increases in chelatable cytosolic iron facilitate permeability transition in hepatocyte mitochondria.

HO-1 Expression in Human CNS Aging and Disease

Normal aging In 2003, W Hirose and colleagues reported that numbers of neurons and neuroglia immunoreactive for HO-1 in the normal human hippocampus and cerebral cortex increase progressively between 3 and 84 years of age. In earlier studies, we calculated that 6.8% and 12.8% of GFAP-positive astrocytes co-express immunoreactive HO-1 protein in the normal senescent human hippocampus and substantia nigra, respectively. HO-1 expression also increases in optic nerve astrocytes and in retinal ganglion cells and photoreceptors of the human eye between the second and the seventh decades of life. In the normal elderly, our group and George Perry's lab noted HO-1 staining to be particularly prominent in choroid plexus epithelial cells, ependymocytes, cerebrovascular endothelial cells, many glial and extracellular CA, and the olfactory neuroepithelium. On the basis of the previously discussed evidence and the tenets of the free radical–mitochondrial theory of aging, we propose that a self-reinforcing cascade of ambient OS, glial HO-1 expression, and mitochondrial iron deposition may contribute to the gradual depletion of high-energy substrates in aging human neural tissues.

Alzheimer's disease In the AD hippocampus, we have calculated that ~86% of GFAP-positive astrocytes exhibit HO-1 immunoreactivity, whereas only 6–7% of hippocampal astroglia in age-matched normal control tissue express this stress protein. In collaboration with David Bennett's group in Chicago, we observed significant induction of astroglial HO-1 expression in the brains of individuals with amnestic mild cognitive impairment (a harbinger of incipient AD), indicating that glial HO-1 upregulation is an early event in the pathogenesis of this dementing disorder. To the extent that our *in vitro* data can be extrapolated to the intact human brain, it is conceivable that the perturbed CNS iron homeostasis and mitochondriopathy characteristic of AD may be downstream effects of chronic HO-1 overproduction in the diseased neural tissues.

Parkinson's disease In 1998, we reported that dopaminergic neurons of the substantia nigra in both normal and PD human brain exhibit moderate HO-1 immunoreactivity. In the PD specimens, cytoplasmic Lewy bodies within affected dopaminergic neurons were prominently decorated with annular HO-1 staining, confirming an earlier observation by R Castellani and colleagues. In our study, the fraction of GFAP-positive astrocytes expressing HO-1 protein in the PD nigra was substantially increased (77.1%) relative to age-matched control preparations (12.8%). In both the PD and control specimens, relatively low levels of glial HO-1 expression were observed in other subcortical nuclei, such as the caudate, putamen, and globus pallidus. In accord with the data and our central hypothesis, sustained augmentation of HO-1 activity in nigral astroglia may promote the pathological deposition of nontransferrin iron, intracellular OS, and mitochondrial electron transport (complex I) deficits reported in the substantia nigra of PD patients.

Other disorders In MS, an autoimmune demyelinating disorder of central white matter, we have observed that astroglial HO-1 expression is significantly increased in spinal cord plaques (57.3% of GFAP-positive cells) relative to that computed in the spinal white matter of normal individuals (15.4%). As argued in the cases of AD and PD, enhanced intraglial heme degradation catalyzed by HO-1 may contribute to the aberrant iron deposits and mitochondrial lesions reported in human MS and in rodent models of this condition. Robust HO-1 induction has also been documented in the brains of patients with Pick disease, corticobasal degeneration, progressive supranuclear palsy, ischemic and hemorrhagic stroke, and *Plasmodium falciparum* malaria. The triad of oxidative substrate damage, deregulated iron homeostasis, and bioenergetic failure occurs in most, if not all, of these conditions and may be secondary to the antecedent overexpression of HO-1 in 'stressed' astroglia.

Conclusions and Implications

The mechanisms responsible for the augmented deposition of iron and other redox-active transition metals in the aging and diseased CNS remain incompletely understood. On the basis of a considerable body of neuropathological and experimental data, we propose that oxidative molecular damage, the deposition of nontransferrin iron (and possibly other metals), and mitochondrial insufficiency in these conditions are causally interdependent phenomena and inextricable components of a single cytopathological 'lesion.' This central hypothesis predicts that iron

homeostasis will invariably be perturbed in CNS tissues harboring mitochondrial lesions, that neural tissues manifesting iron overload will exhibit obligatory bioenergetic failure, and that at least some of the aberrant metal deposits will co-localize to the mitochondrial compartment. We have reviewed evidence that induction of the heme-degrading enzyme, HO-1, in astroglia stressed by aging or disease may represent a 'final common pathway' for the sequestration of nontransferrin iron by the mitochondrial compartment and transformation of these organelles into peroxidase-positive inclusions and CA, structural biomarkers of the senescent mammalian CNS. Given the exquisite sensitivity of the glial *ho-1* gene to ambient OS and other noxious stimuli, it is not surprising that pathological iron deposition and mitochondrial damage are a fairly consistent response of neural tissues to a broad range of genetic and acquired metabolic disturbances. Thus, OS accruing from an excessive amyloid burden or the elaboration of proinflammatory cytokines may be responsible for glial HO-1 induction, iron sequestration, and bioenergetic failure in the AD forebrain. Dopamine-derived ROS, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-like xenobiotics, endoneurotoxins, nitric oxide, and Th1 cytokines are plausible inducers

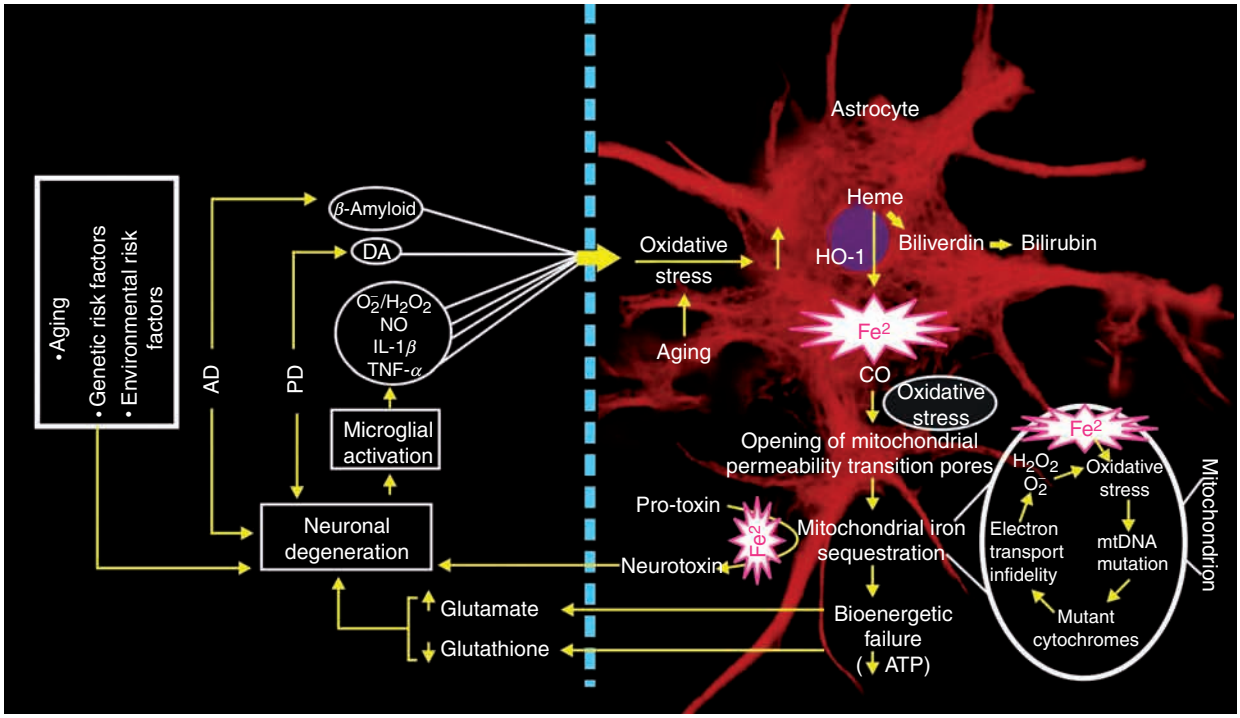

Figure 2 A model for pathological iron deposition, OS, and mitochondrial insufficiency in brain aging and disease. Extracellular pro-oxidant stressors and intrinsic aging processes upregulate heme oxygenase-1 (HO-1) in astrocytes, liberating heme-derived free ferrous iron (Fe^2) and carbon monoxide (CO). The latter generate intraglial OS that promotes opening of mitochondrial permeability transition pores and influx of (non-transferrin-derived) iron into the mitochondrial matrix. The redox-active iron augments oxidative damage to the electron transport chain machinery (mitochondrial DNA (mtDNA), cytochromes), resulting in a vicious spiral of mitochondrial ROS generation and injury in accord with the prevailing free radical–mitochondrial theory of aging. This progressive state of astroglial mitochondrial heteroplasmy endangers nearby neuronal constituents by several mechanisms: (1) The mitochondrial ferrous iron catalyzes the bioactivation of protoxins (e.g., catechols, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)) to potent neurotoxins (*o*-semiquinones, MPP^+). The latter may be extruded to the extracellular space and exert direct dystrophic effects on susceptible neuronal targets. Diminished ATP levels accruing from glial bioenergetic failure may further predispose to neuronal degeneration by (2) augmenting extracellular glutamate concentrations (excitotoxicity), (3) curtailing glutathione (GSH) delivery to the neuronal compartment, and (4) limiting the degradation of potentially toxic protein aggregates by the ubiquitin–proteosome system. Neuronal degeneration stimulates microglial activation, resulting in the release of ROS, nitric oxide (NO), and proinflammatory cytokines. The latter further induce the *ho-1* gene in indigent astroglia, completing a self-sustaining loop of pathological cellular interactions, perpetuating oxidative damage and mitochondrial insufficiency within senescent and degenerating neural tissues. Genetic and environmental risk factors may confer disease specificity by superimposing unique pathological signatures on this core lesion. In the case of AD and PD, for example, β-amyloid deposition in the hippocampus and dopamine (DA) release from dying nigrostriatal projections, respectively, feed into the pathological cascade by stimulating HO-1 activity in regional astroglia. Progressive derangement of astroglial iron mobilization and mitochondrial insufficiency would then drive the primary neurodegenerative process regardless of its etiology. H_2O_2, hydrogen peroxide; IL-1β, interleukin-1β; TNF-α, tumor necrosis factor-α. Adapted from Schipper HM (2004) Brain iron deposition and the free radical-mitochondrial theory of aging. *Aging Research Reviews* 3: 265–301.

of the astroglial heme oxygenase–mitochondrial iron axis in the PD nigra, and iron-mediated mitochondrial injury in MS may be secondary to the upregulation of glial HO-1 by IL-1β, TNF-α, and/or myelin basic protein within the affected white matter.

Importantly, neuron-like PC12 cells co-cultured with rat astrocytes replete with mitochondrial iron (induced by CSH pretreatment) or overexpressing the human *ho-1* gene were found to be far more vulnerable to OS-related killing than were PC12 cells grown atop control ('iron-poor') astroglia. Thus, the accumulation of iron-laden, metabolically deficient astroglia may seriously compromise the integrity and viability of nearby neuronal constituents and thereby promote neurological dysfunction and disease. In our model, pathological, mutually reinforcing glial–neuronal interactions may perpetuate oxidative damage, pathological iron deposition, and mitochondrial insufficiency within senescent and degenerating neural tissues long after initiating insults may have dissipated (**Figure 2**). As a putative transducer of mitochondrial iron deposition in 'stressed' astroglia, the augmented glial HO-1 activity may provide a rational target for therapeutic intervention in a host of human neurodegenerative and neuroinflammatory conditions.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: Neurodegeneration; Brain Composition: Age-Related Changes; Mitochondrial Dysfunction in Nervous System Injury; Oxidative Damage in Neurodegeneration and Injury; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Connor JR (1997) *Metals and Oxidative Damage in Neurological Disorders*. New York: Plenum Press.

Dennery PA (2000) Regulation and role of heme oxygenase in oxidative injury. *Current Topics in Cellular Regulation* 36: 181–199.

Dwyer BE, Takeda A, Zhu X, et al. (2005) Ferric cycle activity and Alzheimer disease. *Current Neurovascular Research* 2: 261–267.

Halliwell B and Gutteridge JMC (2007) *Free Radicals in Biology and Medicine*. Oxford: Oxford University Press.

Heales SJ, Lam AA, Duncan AJ, et al. (2004) Neurodegeneration or neuroprotection: The pivotal role of astrocytes. *Neurochemical Research* 29: 513–519.

Jenner P (2003) Oxidative stress in Parkinson's disease. *Annals of Neurology* 53(supplement 3): S26–S36; discussion S36–S28.

Ponka P (2004) Hereditary causes of disturbed iron homeostasis in the central nervous system. *Annals of the New York Academy of Sciences* 1012: 267–281.

Reddy PH and Beal MF (2005) Are mitochondria critical in the pathogenesis of Alzheimer's disease? *Brain Research Reviews* 49: 618–632.

Ryter SW and Tyrrell RM (2000) The heme synthesis and degradation pathways: Role in oxidant sensitivity. Heme oxygenase has both pro- and antioxidant properties. *Free Radical Biology & Medicine* 28: 289–309.

Schipper HM (2004) Heme oxygenase-1: Transducer of pathological brain iron sequestration under oxidative stress. *Annals of the New York Academy of Sciences* 1012: 84–93.

Schipper HM (2004) Brain iron deposition and the free radical-mitochondrial theory of ageing. *Ageing Research Reviews* 3: 265–301.

Schipper HM (2004) Redox neurology: Visions of an emerging subspecialty. *Annals of the New York Academy of Sciences* 1012: 342–355.

Schipper HM (2004) Heme oxygenase expression in human central nervous system disorders. *Free Radical Biology & Medicine* 37: 1995–2011.

Metaplasticity

D R Ireland, B G Mockett, and W C Abraham, University of Otago, Dunedin, New Zealand

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Metaplasticity Defined

Metaplasticity refers to an activity-dependent and persistent regulation of subsequently induced synaptic plasticity, such as long-term potentiation (LTP) or long-term depression (LTD). Essentially, metaplasticity entails a change in the physiological or biochemical state of neurons such that the ability to generate synaptic plasticity is altered. A key feature of metaplasticity is that this change, once established, outlasts the triggering ('priming') bout of activity and persists until a second bout of activity occurs to induce LTP or LTD (Figure 1). It is this persistence feature that is critical for the phenomenon, and that distinguishes it from more conventional forms of neuromodulation which occur during the induction of synaptic plasticity. There is no agreed criterion for the degree of persistence required to meet the definition of metaplasticity, but commonly studied effects have ranged from many minutes to many days. The 'meta' part of the term reflects the higher order nature of the plasticity, that is, the plasticity of synaptic plasticity.

In principle, metaplasticity can be induced by a range of priming events: cell firing, synaptic activity, paracrine signals, or endocrine signals. It is most clearly observed when the priming event by itself generates no observable LTP or LTD, because the modulation of subsequent plasticity cannot be attributed simply to a change in synaptic drive during the second bout of activity. However, the generation of plasticity during the priming episode does not preclude the simultaneous induction of metaplasticity.

When metaplasticity is induced by synaptic activity, two classes of phenomena can be distinguished: homosynaptic metaplasticity and heterosynaptic metaplasticity. Metaplasticity is homosynaptic when the priming synaptic activity regulates subsequent plasticity only at those same synapses. In contrast, heterosynaptic metaplasticity occurs when plasticity is regulated for a range of synapses terminating on a postsynaptic neuron, regardless of whether they participated in the priming activity.

BCM Model

Metaplasticity was conceptualized at a theoretical and computational level by the Bienenstock–Cooper–Munro (BCM) model of synaptic plasticity in the visual cortex, well before the term was actually coined. In this model, the degree of LTD or LTP generated by a given bout of synaptic activity is a nonlinear function of the degree of postsynaptic activation. Synaptic activity that occurs moderately above the resting level generates LTD, but if sufficiently high levels of activity occur, LTP is generated instead. The level of activity at which the plasticity function switches from LTD to LTP is known as the modification threshold (Figure 2). This model incorporates a form of metaplasticity, in that the modification threshold is itself a function of the degree of cell firing recently generated by the postsynaptic neuron. If the neuron has recently been very active, presumably as driven by synaptic inputs, then the modification threshold shifts to the right, making it harder to generate LTP and easier to induce LTD (Figure 2). Equally, if the neuron has recently been very inactive, then the modification threshold shifts to the left. This modeled form of metaplasticity is fundamentally heterosynaptic in nature, because it is immaterial which synaptic inputs drove the initial cell firing that shifted the modification threshold. The frequency–plasticity function is then set for all excitatory synapses on the postsynaptic neuron.

Metaplasticity Induction

Metaplasticity Induced by Synaptic Stimulation

The study of metaplasticity in the hippocampus has revealed that both LTP and LTD may be modified by prior synaptic activity according to the receptor systems that are activated (see the section titled 'Receptor mechanisms') and the pattern of that activation. Hence, stimulation protocols which would typically induce LTP may fail to do so if preceded by bouts of stimulation which by themselves do not induce conventional synaptic plasticity. For example, in area CA1, LTP can be completely suppressed if the LTP-inducing high-frequency stimulation (HFS) is preceded 20 min earlier by multiple weak trains of HFS. This form of metaplasticity is homosynaptic and lasts for at least 30 min. On the other hand, if multiple strong trains of HFS are delivered to an input pathway, heterosynaptic metaplasticity can be induced

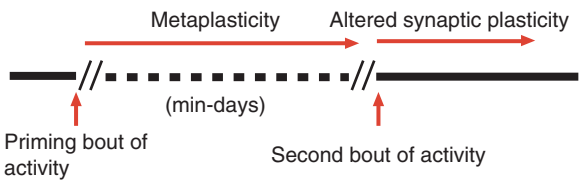

Figure 1 Time line of events during metaplasticity. Metaplasticity is defined as a lasting after-effect of an initial bout of neural activity which changes the state of the participating neurons such that the ability to induce synaptic plasticity at a later time is altered.

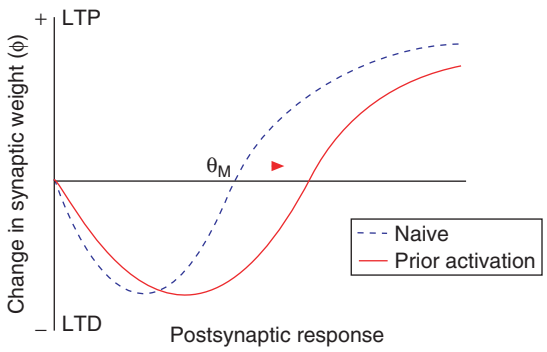

Figure 2 Bienenstock–Cooper–Munro (BCM) model. The ability of presynaptic activity to induce synaptic plasticity is a nonlinear function (Φ) of postsynaptic activity. The modification threshold (θ_M) is the crossover point between LTD and LTP. The modification threshold dynamically varies based on time-averaged postsynaptic cell firing. In this example, sufficient prior postsynaptic activity drives the modification threshold to the right, changing the Φ function from the dashed line to the solid line.

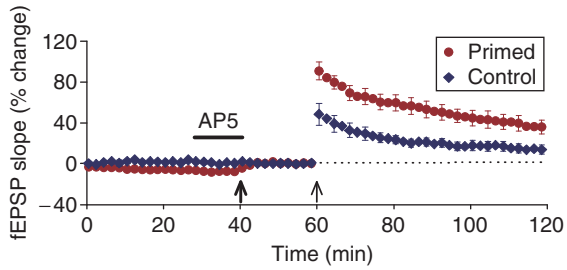

Figure 3 An example of mGluR-dependent synaptic priming of LTP in hippocampal CA1 slices. In control hippocampal slices, theta-burst stimulation (thin arrow at 60 min) delivered to the Schaffer collaterals in area CA1 induced a decaying form of LTP. Priming stimulation of group I metabotropic glutamate receptors (mGluRs) (thick arrow at 40 min) during *N*-methyl-D-aspartate receptor (NMDAR) antagonist AP5 application significantly enhanced the induction and persistence of LTP induced by theta-burst stimulation 20 min later. Data are mean ± S.E.M. of the field excitatory postsynaptic potential (fEPSP) slopes expressed as a percentage of baseline values (dotted line). Adapted with permission from Raymond CR, Thompson VL, Tate WP, et al. (2000) Metabotropic glutamate receptors trigger homosynaptic protein synthesis to prolong long-term potentiation. *Journal of Neuroscience* 20: 969–976. Copyright 2000 by Society for Neuroscience.

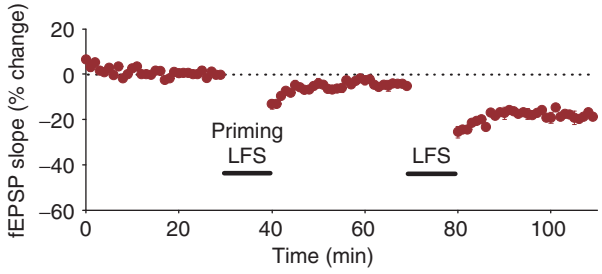

Figure 4 An example of NMDAR-dependent priming of LTD induction in hippocampal CA1 slices. An initial bout of low-frequency stimulation (LFS) to the Schaffer collaterals in area CA1 produces at best a weak short-term depression, but primes the synapses to respond with a much larger LTD when a second identical LFS is delivered 10 min later. Data are mean ± S.E.M. of the field excitatory postsynaptic potential (fEPSP) slopes expressed as a percentage of baseline values (dotted line). Adapted from Mockett B, Coussens C, and Abraham WC (2002) NMDA receptor-mediated metaplasticity during the induction of long-term depression by low-frequency stimulation. *European Journal of Neuroscience* 15: 1819–1826.

such that the induction of LTP is inhibited and the induction of LTD is facilitated for all tested pathways terminating on those postsynaptic neurons. In the dentate gyrus, priming stimulation applied to the medial perforant path raises the LTP threshold heterosynaptically in the lateral perforant pathway for up to several weeks. Although metaplasticity typically entails a change in the initial induction of synaptic plasticity, under some conditions a stimulus protocol can suppress selectively the late phase of subsequent LTP (lasting hours or more), without affecting its induction.

In contrast to the inhibitory effects of prior stimulation on LTP, priming stimulation under other conditions can facilitate LTP (Figure 3). This effect is generally homosynaptic, lasts at least 1 h, and occurs under conditions which favor the activation of metabotropic glutamate receptors (mGluRs) during the priming stimulation (see later). Prior synaptic stimulation can also regulate LTD. For example, an initial mild low-frequency stimulation (LFS) protocol primes synapses to exhibit a large LTD upon being given a second identical LFS (Figure 4). The metaplastic effect of the first LFS is input specific, effective for at least 1 h, and by itself entails little or no observable change in transmission at the primed synapses.

Behavioral Induction of Metaplasticity

While metaplasticity can be induced in its different forms by electrically stimulated synaptic activity, such protocols are not particularly naturalistic. However, metaplasticity can also be induced by environmental stimuli such as changes in sensory input, learning, and

stress. As predicted by the BCM model of synaptic plasticity, changes in visual input during development *in vivo* can cause metaplasticity of LTP and LTD as assessed *in vitro*. In slices of visual cortex from visually deprived (dark-reared) rats, the modification threshold is shifted to the left compared to light-reared animals, such that LTP and LTD are induced by a lower than normal range of stimulation frequencies (Figure 5). Similarly, olfactory discrimination learning makes LTP in rat olfactory cortex more difficult to induce in slices, while at the same time enabling LTD induction that does not normally occur in slices from untrained animals. Noxious stimuli can also affect plasticity; stress induced *in vivo* causes impairment of subsequent LTP in both CA1 and the dentate gyrus, plus an enhancement of LTD, as assessed *in vitro*. This metaplasticity can last several days after the stress occurs.

There is an exception to the general principle that increased sensory input makes LTP more difficult to attain; if a brief exposure to a novel environment is given temporally close to the induction of LTP, then its induction *in vivo* in CA1 is enhanced. This indicates that under some conditions an increase in sensory input is followed immediately by a short window of time during which synapses are in a metaplastic state, such that LTP induction is facilitated rather than inhibited.

Receptor Mechanisms

The stimulation protocols just described achieve their metaplastic effects by activation of specific postsynaptic receptors, primarily glutamatergic. These receptors then initiate a series of intracellular events through activation of second messengers or entry of calcium (see the section titled 'Signal transduction pathways'). *N*-Methyl-D-aspartate receptors (NMDARs) and mGluRs have attracted the most attention as triggers for evoking metaplasticity events in the hippocampus. Priming activation of NMDARs by either electrical stimulation of afferent pathways or by direct NMDA application inhibits the induction of LTP while promoting the development of LTD. Furthermore, NMDAR antagonists block the ability of multiple weak tetani to inhibit subsequent LTP. Similarly, NMDAR antagonists block the priming of LTD by a mild LFS. In the stress model of metaplasticity, the inhibition of LTP and facilitation of LTD are both mediated by NMDAR activation during the stressful experience, as NMDAR antagonists given during the stress block these metaplastic effects.

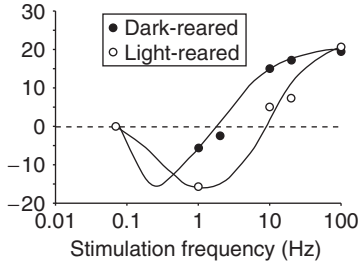

Figure 5 Visual experience can induce metaplasticity. In visual cortical slices from rats that have been visually deprived (filled circles), LTP and LTD (measured in layers 2/3 and evoked by layer 4 stimulation) can be induced by lower than normal stimulation frequencies *in vitro*. Note that stimulation at 0.5 Hz induces greater LTD in visual cortex of dark-reared rats compared with light-reared rats (data point not shown). Adapted from Philpot BD, Espinosa JS, and Bear MF (2003) Evidence for altered NMDA receptor function as a basis for metaplasticity in visual cortex. *Journal of Neuroscience* 23: 5583–5588. Copyright 2003 by Society for Neuroscience.

The consequences of prior mGluR activation on subsequent synaptic plasticity largely depend on the subtype activated and the brain region in which they are located. For example, in area CA1, group I mGluR activation facilitates LTP induced by a subsequent mild HFS, while prior activation of group II mGluRs inhibits LTD induction without affecting LTP. In contrast, the priming effect of mGluR activation on subsequent LTP/LTD in the dentate gyrus is less predictable and appears to be determined by a range of experimental variables. For example, at medial perforant path–granule cell synapses *in vitro*, activation of group I and II mGluRs by strong HFS blocks subsequent group I receptor function, and hence inhibits group I-dependent LTD/depotentiation, while activation of mGluRs in response to a weak HFS inhibits subsequent HFS-induced LTP. The latter effect requires a synergistic interaction between group I and group II mGluRs and coactivation with NMDARs. The outcome of mGluR activation also depends on the nature of the experimental preparation. Thus, while activation of group II mGluRs in freely moving rats facilitates synaptically induced LTD, *in vitro* this treatment facilitates LTP induction.

Some priming effects have been reported for other receptor types, such as the β-adrenergic and dopamine receptors. Dopamine D1/D5 receptors, for example, are involved in mediating the consolidation of tetanus-induced LTP in CA1 of freely moving rats by prior exposure to novel objects. Interestingly, *in vitro* studies show that this is a highly regional effect within the hippocampus, as similar regulation is absent in the dentate gyrus. However, in the case of β-adrenergic receptor-mediated priming, a reversal of this pattern occurs with increased persistence of tetanus-induced LTP observed in the dentate gyrus, but not in CA1. While these priming effects are of considerable physiological relevance, comparatively little work has been done to characterize the mechanisms involved in mediating these responses or their integration with other plasticity-inducing factors.

Signal Transduction Pathways

Like synaptic plasticity, metaplasticity relies on cascades of intracellular signaling molecules to turn synaptic activation of membrane-bound receptors such as mGluRs and NMDARs into lasting changes in the cell (Figure 6). Group I mGluRs couple to phospholipase C (PLC) via a $G_{\alpha q}$-type G-protein. PLC then liberates diacylglyerol (DAG) and inositol trisphosphate (IP_3) from membrane phospholipids, the former activating protein kinase C (PKC) and the latter binding to receptors on intracellular Ca^{2+} stores, causing Ca^{2+} release into the intracellular compartment. NMDARs, in contrast, act more directly by virtue of their relatively high permeability to Ca^{2+}: under conditions favorable to NMDAR activity, Ca^{2+} enters the neuron and acts as a trigger for subsequent biochemical events. A prime example of a signaling enzyme that is triggered by a rise in intracellular Ca^{2+} levels is Ca^{2+}/calmodulin-dependent kinase II (CaMKII). CaMKII is a key component not only in the induction of LTP but also in metaplasticity. This kinase is initially activated by Ca^{2+}/calmodulin complex, but once activated it autophosphorylates at the Thr-286 site and becomes active independent of Ca^{2+}. However, if CaMKII

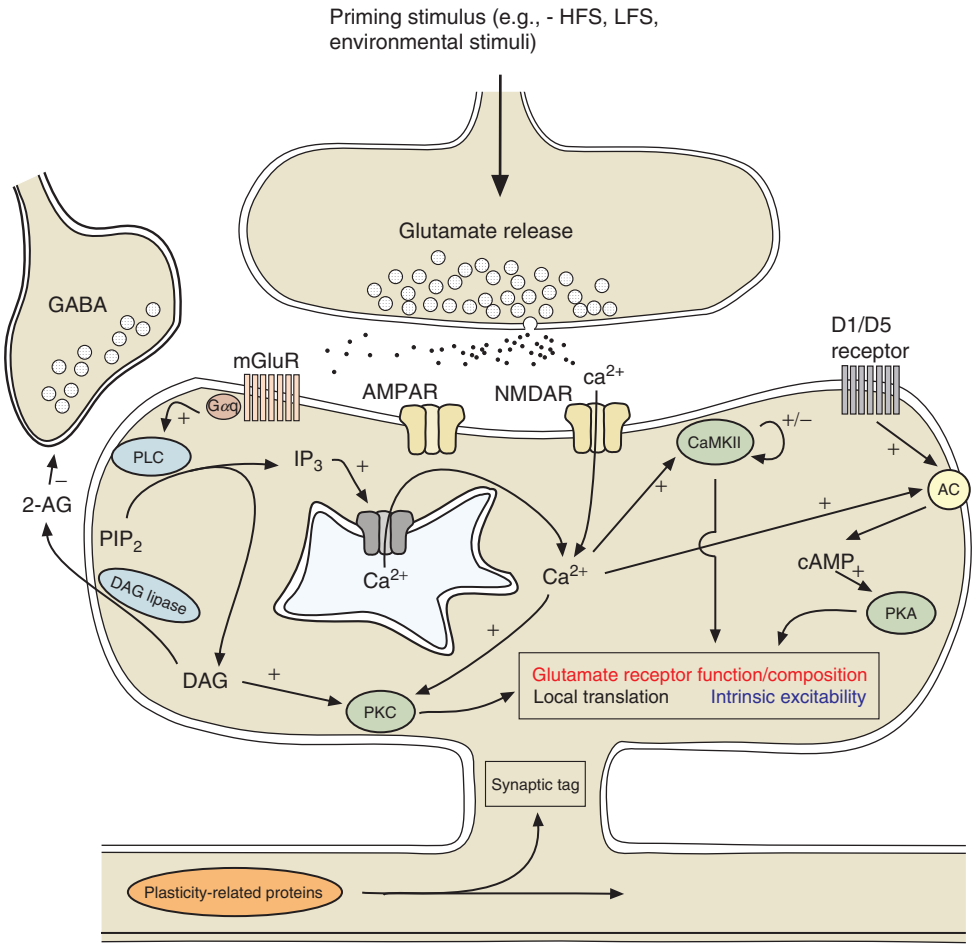

Figure 6 Schematic diagram showing some of the receptors and signal transduction pathways involved in metaplasticity. Presynaptic stimulation of an appropriate magnitude and temporal pattern causes the release of glutamate, which binds to *N*-methyl-D-aspartate receptors (NMDARs), α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs), and metabotropic glutamate receptors (mGluRs). Both NMDARs and mGluRs can initiate metaplasticity in response to different patterns of synaptic stimulation (HFS, LFS, high- and low-frequency stimulation). NMDARs act by allowing direct Ca^{2+} influx into the cell and thereby triggering a signaling cascade involving enzymes such as Ca^{2+}/calmodulin-dependent kinase II (CaMKII), whereas mGluRs induce metaplasticity by coupling to signaling pathways via a G-protein. These actions then contribute to processes such as glutamate receptor-associated changes (receptor function, expression, location, or subunit composition), changes in intrinsic excitability, and presynaptic changes which place the synapse in a metaplastic state. Additionally, metaplasticity can entail increased local translation and somatic transcription of proteins involved in plasticity induction and consolidation, and setting of a 'synaptic tag' to allow capture of newly synthesized plasticity-related proteins transported from the soma. GABA, γ-aminobutyric acid; PLC, phospholipase C; PIP_2, phosphoinositol bisphosphate; IP_3, inositol trisphosphate; DAG, diacyglycerol; PKC, protein kinase C; PKA, protein kinase A; AC, adenylate cyclase; cAMP, cyclic AMP.

autophosphorylates at the Thr-305/306 site, it inhibits its own activity. This dual regulation allows CaMKII to be involved in both mGluR- and NMDAR-dependent metaplasticity. For example, the mGluR-dependent 'molecular switch' that facilitates later LTP induction at CA3–CA1 synapses is dependent on activation of CaMKII. On the other hand, NMDAR-dependent homosynaptic inhibition of LTP by prior activity at lateral perforant path synapses in dentate gyrus relies on autoinhibition of CaMKII.

PKC is also common to several types of metaplasticity. This enzyme is activated by the second messenger DAG, but the activity of some PKC isotypes is also modulated by Ca^{2+}. PKC mediates several types of mGluR-dependent metaplasticity; at CA3–CA1 synapses both the 'molecular switch' and priming of LTP involve PKC (following $G_{\alpha q}$ and PLC activation in the latter case). Similarly, at medial perforant path synapses in the dentate gyrus, inhibition of NMDAR-dependent LTP and mGluR-dependent LTD by prior high-frequency stimulation relies on PKC. Other signaling molecules, such as p38 MAP kinase, protein phosphatase 1, and protein phosphatase 2A, have been implicated in metaplasticity.

A form of metaplasticity which deserves separate mention because of its unusual signal transduction is the mGluR- and endocannabinoid-dependent facilitation of LTP at CA1 synapses. This form of metaplasticity is dependent on activation of PLC but not a rise in intracellular calcium, instead relying on production of the endocannabinoid, 2-arachidonoyl glycerol (2-AG), from DAG. Another unusual aspect of this form of metaplasticity is the fact that it is not restricted to the postsynaptic neuron; instead, 2-AG is released to retrogradely inhibit presynaptic GABAergic synapses (Figure 6).

Metaplasticity Expression

Changes in Glutamate Receptor Function and Expression

Metaplasticity is expressed in many cases by changes in glutamate receptor function and expression. For example, some protocols which metaplastically inhibit LTP induction act by suppressing NMDAR currents and therefore Ca^{2+} flow into the spine during LTP-inducing HFS. Similarly, NMDAR function is downregulated during maturation of rat cortex through changes in NMDAR subunit composition from predominantly NR2B-containing to predominantly NR2A-containing, the latter being less Ca^{2+} permeable. Although normally a slow developmental process, this change can occur very rapidly, as exemplified by the subunit switch occurring in the rat visual cortex of dark-reared animals 2 h after light exposure. The net effect of having more NR2A-containing receptor is a rightward shift in the modification threshold such that higher stimulus frequencies are needed to induce LTP and LTD compared to non-light-exposed rats Figure 5. Similarly, olfactory training makes LTP in rat olfactory cortex more difficult to induce in slices, in conjunction with an increase in the expression of the NR2A subunit relative to the NR2B subunit of the NMDAR.

At many glutamatergic synapses, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs) provide the majority of current flow into the postsynaptic cell in response to presynaptic glutamate release under normal conditions, and thus any modification of receptor function or synaptic expression will manifest itself as an overt change in synaptic transmission (i.e., synaptic plasticity). On the other hand, regulation of AMPARs may also be involved in the expression of metaplasticity, such as a change in the threshold for addition of new AMPARs to the existing synaptic pool. Emerging evidence indicates that this can be accomplished by protein kinase A (PKA)-dependent priming of AMPAR movement into the extrasynaptic membrane. Subsequent LTP-related NMDAR activity then triggers their rapid delivery into the synaptic membrane as functional receptors.

Changes in Protein Synthesis

In some models, the expression of metaplasticity requires *de novo* protein synthesis as an intermediate step. For example, priming of LTP at CA1 synapses by activation of group I mGluRs initiates the synthesis of new proteins, and inhibitors of translation but not transcription inhibit the priming effect. This indicates that the proteins are synthesized by preexisting mRNA, probably localized to dendrites or synapses. The identities of the key proteins are not yet known. In another paradigm, a weak decaying form of LTP generated by a mild HFS in one pathway can be converted to a stable form of LTP if it can capture proteins generated in response to a strong HFS in a neighboring pathway.

How are the new proteins captured at the synapse for use? In mGluR priming of LTP, the proteins are made locally near the synapses and thus are already close to their site of action. In the case of somatodendritically localized protein synthesis, proteins, are captured through a process known as 'synaptic tagging.' Here, a mild HFS delivered to one set of synapses sets a 'tag' at those synapses to identify them as activated. Although the identity of this tag is also currently unknown, its role is to sequester newly

synthesized somatic and dendritic plasticity-related proteins into the synapse as they are transported throughout the cell, and thereby enhance the persistence of the initially weak plasticity. Synaptic tagging is an example of a heterosynaptic metaplastic process in which weakly activated synapses exhibiting decremental plasticity are primed for the capture of plasticity-related proteins necessary for consolidation of LTP/LTD (Figure 6).

Metaplasticity at Individual Synapses – Synaptic States

Although metaplasticity is usually studied in populations of synapses, individual synapses within a group could have very different histories of activity. In organotypic slice cultures, synapses between pairs of individual hippocampal neurons can exist in one of several possible 'synaptic states,' the particular state being determined by the synapse's recent history of activity. One of these states, termed 'silent,' describes a synapse that possesses functional NMDARs but no AMPAR response, probably due to a lack of (or few) AMPARs in the membrane. When triggered by appropriate synaptic activity, a silent synapse can become 'active' and exhibit an AMPAR-mediated response. From this active state, LTP and LTD can be generated. A further state, neither potentiated nor depressed, has been termed 'recently silent.' Recently silent synapses have only just acquired an AMPAR response due to recent synaptic activity but have not yet fully transitioned into the active state. Despite having a normal AMPAR response and being capable of potentiation, these synapses are temporarily immune from LTD caused by endocytosis of AMPARs that would occur at an active synapse in response to LFS. Thus two synapses with apparently identical strength but different histories of activity can differ greatly in their ability to undergo synaptic plasticity.

Plasticity of Intrinsic Excitability

Under some conditions, action potentials generated in or near the soma by summation of synaptic potentials can back-propagate into the dendrites and affect synaptic function. The relative timing between the arrival of synaptic input and the back-propagating spikes can influence the strength and direction of synaptic plasticity at a given synapse. In addition, dendrites are capable of generating spikes, which may also be important in regulating synaptic plasticity. Therefore, any persistent plasticity of intrinsic cell excitability (i.e., the ability to generate and propagate a spike) in response to activity provides a potentially powerful means of metaplasticity expression.

In several types of neurons, the ability to fire action potentials is modifiable by synaptic stimulation. For example, HFS of deep cerebellar nucleus neurons causes a persistent NMDAR-dependent increase in the number of spikes fired in response to a depolarizing current pulse. Also in cerebellar neurons, theta-burst stimulation of mossy fibers causes a persistent increase in excitability, expressed as an NMDAR-dependent increase in input resistance and decrease in spike threshold. In cortical neurons, moderate-frequency synaptic stimulation causes a highly persistent increase in spike firing in response to a depolarizing stimulus. This is due to an mGluR-dependent downregulation of the Ca^{2+}-activated K^+ channels underlying the spike afterhyperpolarization (AHP) (Figure 7). Importantly, downregulation of the AHP facilitates the induction of subsequent LTP.

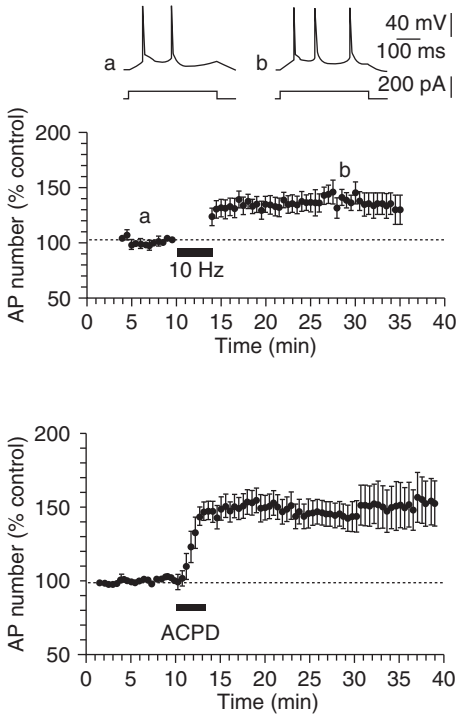

Figure 7 Synaptic or pharmacological activation of mGluRs can cause potentiation of intrinsic excitability. Potentiation of intrinsic excitability in layer V pyramidal neurons can be induced by tetanic stimulation of glutamatergic synaptic inputs at 10 Hz in the presence of ionotropic receptor antagonists. Upper panel: Number of action potentials (AP) evoked by a constant current pulse before and after tetanic stimulation. Traces a and b, respectively, recorded before and 25 min after a 10 Hz stimulation. Lower panel: A similar modification of intrinsic excitability induced pharmacologically by the mGluR agonist aminocyclopentane dicarboxylic acid (ACPD) (50 μM). Adapted from Sourdet V, Russier M, Daoudal G, et al. (2003) Long-term enhancement of neuronal excitability and temporal fidelity mediated by metabotropic glutamate receptor subtype 5. *Journal of Neuroscience* 23: 10238–10248. Copyright 2003 by Society for Neuroscience.

A key characteristic of the plasticity of intrinsic excitability is its ability to either localize to specific regions in the neuron or to be cell-wide. For example, a burst of back-propagating action potentials fired by depolarization of rat CA1 pyramidal neurons persistently depresses voltage-dependent Ca^{2+} channels in individual dendritic spines, prevents subsequent LTP. Also in CA1 neurons, LTP is accompanied by a persistent increase in dendritic excitability which manifests as an amplification of back-propagating action potentials due to a downregulation of A-type K^+ channels. Crucially, this change in intrinsic excitability is limited to the local dendritic region containing the stimulated synapses. Conversely, a persistent change in somatic spike-generating ability should influence subsequent synaptic plasticity in a more generalized way. Indeed, LTP-inducing synaptic activity at one set of synapses on a CA1 pyramidal neuron can cause a persistent increase in the AHP and a reduction in spike firing, which inhibits the induction of subsequent LTP at a separate set of synapses on the same neuron.

Persistent plasticity of intrinsic cell excitability measured *in vitro* is also associated with *in vivo* learning. For example, hippocampus-dependent learning results in persistent suppression of the AHP and an increase in spike firing measured subsequently *in vitro* in both hippocampal and olfactory cortex neurons. Similar changes in intrinsic excitability as a result of learning tasks have also been observed in invertebrates such as *Aplysia*. Importantly, learning of either a hippocampus-dependent olfactory or spatial task is enhanced by prior olfactory learning, and this enhancement occurs only within the time frame of the increase in intrinsic excitability. It is therefore likely that the persistent increase in intrinsic excitability that accompanies learning is a real-life form of metaplasticity which allows an improvement in the ability to learn a subsequent task.

Presynaptic Mechanisms

In addition to the sizable body of literature devoted to postsynaptic mechanisms of metaplasticity, there is also evidence that a persistent change in the release probability of presynaptic glutamatergic terminals due to prior activity is one way in which metaplasticity is expressed. In cultures of hippocampal neurons, a chronic reduction in synaptic activity increases the release probability and physical size of synapses, and either chronic reduction in activity or chronic blockade of AMPA receptors increases the pool of available glutamate-containing vesicles. Further, a chronic reduction in either activity or Ca^{2+} influx modifies synapses such that their number of releasable vesicles and presynaptic release probability are more likely to be increased by an LTP-inducing stimulus. Presynapthic metaplasticity also occurs *in vivo*: at mossy fiber–CA3 synapses in the rat hippocampus, intense synaptic activity *in vivo* leads to increased presynaptic release probability *in vitro* (lasting several weeks), an increase in the pool of releasable glutamate available in the presynaptic terminal, and a corresponding reduction in the ability to induce LTP *in vitro*.

Another way in which presynaptic changes can mediate metaplasticity of excitatory synapses is through plasticity of inhibitory synapses. At hippocampal CA3–CA1 synapses *in vitro*, an HFS which is subthreshold for LTP induction can nonetheless produce an LTD of GABAergic inhibitory synaptic transmission via retrograde endocannabinoid signaling (see earlier, section titled 'Signal transduction pathways,' and Figure 6). This results in a persistent enhancement of subsequently induced LTP.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Depression: Cerebellum; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; NMDA Receptors and Development; Plasticity of Intrinsic Excitability.

Further Reading

Abraham WC and Bear MF (1996) Metaplasticity: the plasticity of synaptic plasticity. *Trends in Neurosciences* 19: 126–130.

Abraham WC and Tate WP (1997) Metaplasticity: A new vista across the field of synaptic plasticity. *Progress in Neurobiology* 52: 303–323.

Bienenstock EL, Cooper LN, and Munro PW (1982) Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience* 2: 32–48.

Esteban JA, Shi SH, Wilson C, et al. (2003) PKA phosphorylation of AMPA receptor subunits controls synaptic trafficking underlying plasticity. *Nature Neuroscience* 6: 136–143.

Huang YY, Colino A, Selig DK, et al. (1992) The influence of prior synaptic activity on the induction of long-term potentiation. *Science* 255: 730–733.

Kim JJ and Yoon KS (1998) Stress: Metaplastic effects in the hippocampus. *Trends in Neurosciences* 21: 505–509.

Mockett B, Coussens C, and Abraham WC (2002) NMDA receptor-mediated metaplasticity during the induction of long-term depression by low-frequency stimulation. *European Journal of Neuroscience* 15: 1819–1826.

Montgomery JM and Madison DV (2004) Discrete synaptic states define a major mechanism of synapse plasticity. *Trends in Neurosciences* 27: 744–750.

Philpot BD, Espinosa JS, and Bear MF (2003) Evidence for altered NMDA receptor function as a basis for metaplasticity in visual cortex. *Journal of Neuroscience* 23: 5583–5588.

Raymond CR, Thompson VL, Tate WP, et al. (2000) Metabotropic glutamate receptors trigger homosynaptic protein synthesis to prolong long-term potentiation. *Journal of Neuroscience* 20: 969–976.

Sourdet V, Russier M, Daoudal G, et al. (2003) Long-term enhancement of neuronal excitability and temporal fidelity mediated by metabotropic glutamate receptor subtype 5. *Journal of Neuroscience* 23: 10238–10248.

Zhang W and Linden DJ (2003) The other side of the engram: Experience-driven changes in neuronal intrinsic excitability. *Nature Reviews Neuroscience* 4: 885–900.

Zhang L, Kirschstein T, Sommersberg B, et al. (2005) Hippocampal synaptic metaplasticity requires inhibitory autophosphorylation of Ca^{2+}/calmodulin-dependent kinase II. *Journal of Neuroscience* 25: 7697–7707.

3,4-Methylenedioxymethamphetamine (MDMA, "Ecstasy")

A R Green, University of Nottingham, Nottingham, UK
M I Colado, Universidad Complutense, Madrid, Spain

© 2009 Published by Elsevier Ltd.

Ecstasy is the name given to 3,4-methylenedioxymethamphetamine (MDMA). However, ecstasy tablets obtained illicitly often contain impurities and even other psychoactive compounds in addition to MDMA. Ecstasy is widely used by young people as a recreational drug, particularly as part of the dance party or 'rave' scene. The major pharmacological effects of MDMA, both acute and long term, primarily involve the neurotransmitter serotonin (5-hydroxytryptamine; 5-HT), but other neurotransmitters are also involved in the action of MDMA, particularly dopamine. In addition to an acute effect on serotonin release from nerve endings in the brain, high or repeated doses of MDMA produce neurotoxic damage to serotonin nerve terminals in the brains of experimental animals. However, unequivocal evidence for similar long-term damage in the brain of human users remains lacking. Nevertheless, there is increasing evidence for both psychological and physiological abnormalities occurring in persons who have used the drug extensively.

History and Human Recreational Use

MDMA was originally synthesized and patented in Germany in 1912 as an intermediate for therapeutically active drugs being developed by E Merck. It is an amphetamine derivative and structurally related to several other amphetamine derivatives (Figure 1). Its pharmacological and toxicological effects were examined in the 1950s by the US military as part of a chemical warfare program examining psychoactive drugs related to mescaline, to which it also bears some structural similarity (Figure 1). By the 1980s, MDMA was being used as an adjunct to psychotherapy, but it was classed as a Schedule 1 drug in the United States in 1985, and this was closely followed by it being listed as an illegal Class A drug in the UK. Despite (or possibly in part because of) these restrictions, MDMA rapidly gained considerable popularity as a recreational drug, particularly as part of the dance scene.

Ecstasy tablets continue to be sold in an assortment of colors, sizes, and shapes, and recent studies have suggested that most now contain primarily MDMA and that each tablet is likely to contain between 70 and 90 mg of MDMA. While many users will only take one or two tablets in an evening, high doses (upward of eight tablets) are sometimes consumed in one session, and binge dosing (taking large numbers of tablets but divided over time) has also been reported to occur, in an apparent effort by the user to reduce adverse effects. Concurrent use of other psychoactive compounds, particularly alcohol and cannabis, is common. While acute fatalities do occur, they are relatively infrequent and are estimated to be approximately 12 per year in the UK despite upward of 500 000 young persons ingesting it weekly. However, concern continues at possible long-term damage to the brain, particularly as this could be occult and only become apparent at some time in the future.

Acute Effects in Animals

Brain Neurochemistry

A single acute dose of MDMA produces a rapid release of 5-HT from its nerve terminals in the rodent brain. This has been demonstrated *in vivo* using microdialysis (Figure 2) and also using brain slice preparations *in vitro*. The brain 5-HT concentration can be depleted by as much as 80% within the first few hours following a dose of MDMA (Figure 3). This rapid 5-HT release can be prevented by the administration of a 5-HT uptake inhibitor such as fluoxetine (Figure 2) or citalopram, which indicates that MDMA is being taken up into the nerve ending via the 5-HT transporter site. The drug also inhibits tryptophan hydroxylase activity, the rate-limiting step in serotonin biosynthesis. Inhibition of this enzyme will also increase 5-HT depletion in the brain. MDMA also enhances dopamine release and again this has been demonstrated *in vitro* using brain slice preparations and *in vivo* using microdialysis, but this has little overall effect on the dopamine concentration in the brain. MDMA also exerts a stimulatory effect on cholinergic neurons resulting in the release of acetylcholine in several brain areas as well as in striatal slices. This response seems to be secondary to the enhanced 5-HT release induced by the drug. MDMA has been shown to enhance norepinephrine release *in vitro* but evidence for this effect occurring *in vivo* is lacking. In contrast to the long-term effects, the acute changes induced by MDMA on brain 5-HT and dopamine release appear to be produced by the action of the parent compound since both are evident when the drug is administered peripherally or directly into the brain.

MDMA binds to 5-HT, dopamine, and norepinephrine transporters, having the highest affinity for the 5-HT transporter. It also has some affinity for

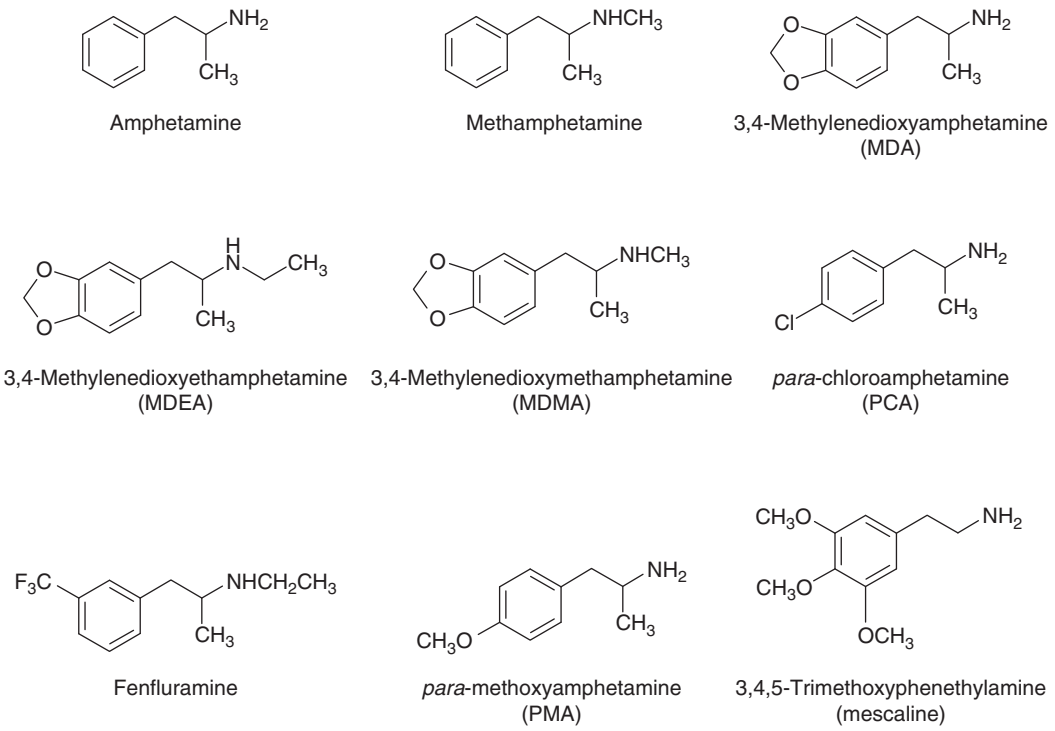

Figure 1 The structure of MDMA, MDA, some other amphetamine compounds, and mescaline.

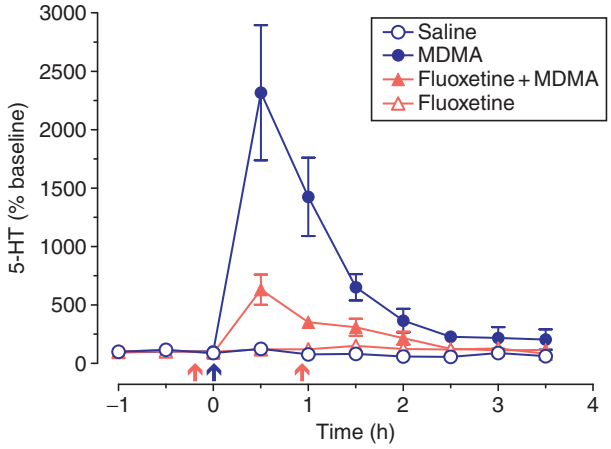

Figure 2 The acute effect of MDMA (15 mg kg^{-1} given at time 0) on 5-HT release in the hippocampus and the effect of fluoxetine (10 mg kg^{-1}) given at –5 and +55 min. Results shown as mean ± SE of mean.

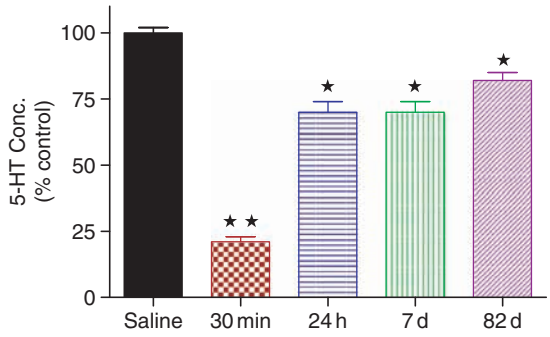

Figure 3 The concentration of 5-HT in frontal cortex of rats administered MDMA (5 mg kg^{-1} given at 3 h intervals to a total of three doses) at various times after administration of the final dose compared to saline-injected controls. Different from saline-injected rats: $^{**}P < 0.01$; $^{*}P < 0.05$.

several neurotransmitter receptors including 5-HT$_1$, 5-HT$_2$, α-adrenergic, and β-adrenergic receptors.

Administration of MDMA also results in a rapid increase in free radical production in the brain of both rats and mice together with subsequent lipid peroxidation. The free radicals are thought to result from the metabolism of MDMA and its principal metabolite 3,4-methylenedioxyamphetamine (MDA) to catechol and quinine compounds that then produce free radicals following further degradation. Dopamine has also been suggested to be involved, but this is controversial, and evidence has recently been presented which indicates that MDMA increases cerebral tyrosine which is converted to free radical-producing metabolites by nonenzymatic processes. Free radical production has been linked to the long-term neurotoxic action of MDMA (discussed later). Both oxygen-(hydroxyl) and nitrogen-based radicals have been implicated in the damage. MDMA also causes a decrease in superoxide dismutase and glutathione peroxidase activities in several regions of the brain.

An increase in the production of the proinflammatory cytokine interleukin 1-beta (IL-1β) is also detected shortly after MDMA. This effect is partially a consequence of the hyperthermic response subsequent to MDMA and it is reduced when rats are maintained at low ambient room temperature. In addition to increasing IL-1β release, MDMA also induces microglial activation, but, in contrast to the changes in IL-1β, the glial response is not significantly modified when the hyperthermia is abolished. These changes in neuroinflammatory parameters are probably involved in the long-term neurotoxic effects induced by MDMA.

Physiological and Psychological Effects

The major acute functional consequence of MDMA administration to rodents is a change in body temperature. Generally, the response is hyperthermia, which is seen if the animals are housed at normal or warm (>20 °C) ambient room temperatures. However, if rats are housed in cooler room conditions (<17 °C), a hypothermic response is observed. Increasing the room temperature results in rats displaying a greater hyperthermic response than that seen in animals housed at normal ambient temperature. The hyperthermia appears to result primarily from enhanced MDMA-induced dopamine release stimulating D1 dopamine receptors, while the hypothermia seen in rats housed in cool conditions is due to stimulation of D2 receptors. The rise in rectal temperature following MDMA involves thermogenesis. There is an increase in metabolic rate involving the hypothalamic–pituitary–thyroid axis, and β3-adrenergic stimulation in brown adipose tissue. Since the tail temperature does not increase in hyperthermic animals following MDMA, and the tail is a major heat loss organ in rodents, it can be proposed that MDMA also interferes with heat loss mechanisms. If sufficient MDMA is administered, the hyperthermic response can prove fatal. In common with other amphetamine compounds, the toxic action of MDMA can be enhanced by grouping animals together (so-called aggregation toxicity). High temperature and noise also worsen toxicity, which is a matter of concern given the use of MDMA by humans in hot, noisy, crowded dance floor conditions. A somewhat different pattern of temperature effect is seen in nonhuman primates following MDMA. While MDMA does produce a hyperthermic effect, the magnitude of the response is not altered by the ambient temperature conditions. Hyperthermia is seen in nonhuman primates housed in cool conditions and normal ambient temperature conditions and is not increased further by increasing the room temperature, which suggests that monkeys have a greater ability to lose heat even at elevated room temperature.

Rats given a high dose of MDMA display the so-called serotonin syndrome, a series of behavioral changes which include hind limb abduction, head weaving, reciprocal forepaw treading, and increased locomotor activity. This behavioral syndrome is indicative of increased 5-HT_{1A} receptor activity in the brain. In mice, the predominant behavioral consequence of MDMA administration is enhanced locomotor activity, due in part to 5-HT_{1B} receptor stimulation. While it has been suggested that the changes in body temperature observed after MDMA administration might be the result of the increased activity, this seems unlikely since studies have shown a dissociation between MDMA-induced hyperthermia and hyperlocomotion. Thus, at an elevated room temperature (30 °C), the same dose of MDMA produced a marked hyperthermia but similar locomotor response to that observed in rats given MDMA at standard room temperature.

Administration of MDMA to rats induces several neuroendocrinological changes. Serum corticosterone is increased as is prolactin, aldosterone, and renin. There have also been cardiovascular abnormalities observed, with increases in both mean arterial pressure and bradycardia.

Acute effects of MDMA in rats examined in models of anxiety remain somewhat conflicting. In rats, MDMA has been reported to have an anxiogenic-like effect when animals are tested on the elevated plus maze, while both an anxiogenic and an anxiolytic effect have been reported in the social interaction test. In mice all studies performed have indicated that MDMA has an anxiogenic-like effect. Several studies have found that the rewarding properties of the drug MDMA produces a rewarding effect in electrical and drug self-administration tests in both rats and baboons. This effect is probably associated with the increase in dopamine release from the nucleus accumbens produced by MDMA administration. Interestingly, the drug-induced dopamine release is enhanced by housing the rats in warm conditions, which suggests that the enthusiasm for taking the drug in warm dance-club conditions may not be coincidental.

It is also possible that previous exposure to MDMA may increase the reinforcing effects of other psychostimulants. Cocaine produces a higher increase in the extracellular levels of dopamine in the nucleus accumbens of animals pretreated with MDMA than that seen in control rats. These animals also show a significantly greater conditioned place preference (CPP) response to cocaine than the vehicle-treated group. Since CPP is believed to be a measure of appetitive behavior, these results provide direct evidence of an increase in the rewarding properties of cocaine in

rats preexposed to MDMA and suggest that MDMA abuse leads to increased vulnerability to cocaine addiction and dependence. Consequently, and taking into account the limitations involved in extrapolating results from animals to humans, it would seem reasonable to propose that MDMA users may be at risk of developing addiction to and dependence on other psychostimulants.

Acute Effects in Humans

While there is a considerable body of information on the acute effects of MDMA in human users, the value of many of the observations is somewhat compromised by the fact that the doses consumed and time of ingestion are often unknown, and other drugs, both known and unknown to the subject, will have been taken around the same time as the MDMA. A further complication includes the fact that MDMA metabolism appears to have saturable kinetics; that is, higher doses are metabolized relatively more slowly than lower doses. Furthermore, it is metabolized by the CYP450 enzymes, including the CYP2D6 enzyme, which means that not only will other drugs interfere with its rate of metabolism but also the rate of metabolism is influenced by whether the person is a poor or fast metabolizer.

The acute physiological effects of MDMA in humans are very similar to those reported in rodents. These include increased heart rate and increased systolic and diastolic blood pressure. These symptoms are blocked by the 5-HT uptake inhibitor citalopram. Crucially, hyperthermia is one of the major consequences of MDMA-induced toxicity, with body temperatures of over 43 °C being reported: such marked hyperthermia can prove fatal. Like rats, the hyperthermia is not blocked by a 5-HT uptake inhibitor, but, unlike rats, MDMA increases body temperature in humans regardless of the ambient temperature. Various other adverse effects have been reported in MDMA users including tachycardia, coagulopathy, rhabdomyolysis, thrombocytopenia, and pulmonary congestion, all of which are potentially fatal. Kidney failure, subarachnoid hemorrhage, and liver failure have also been seen, and many of these severe adverse effects are primarily the result of the severe hyperthermia.

Acute psychological changes following MDMA ingestion can include both desirable and undesirable effects. The presumed desirable effects of this recreational drug include elation, happiness, friendliness, calmness, and increased perception of color, sound, and touch, but confusion has also been reported. However, following the drug intake feelings of low mood, paranoia, insomnia, and irritability have been regularly reported by subjects. 'Low mood' can be experienced several days after the last ingestion of the drug. Many of the psychological effects of MDMA can be blocked with citalopram, but since MDMA is an amphetamine compound and also acutely releases dopamine, it is reasonable to propose that excitability and euphoria are associated with dopamine rather than serotonin release in the brain. Other undesirable effects reported by ecstasy users are nausea/vomiting, bruxism, muscle aches/headache, sweating, dry mouth, and dizziness.

Long-Term Neurotoxicity in Animals

There have been a substantial number of reports attesting to the fact that high or repeated doses of MDMA produce neurotoxic degeneration of 5-HT nerve terminals in the forebrain of rats. Most of these studies have used biochemical measures of 5-HT loss to indicate damage (Figure 3). It has been found that MDMA produces a loss of 5-HT in regions of rat brain lasting for many months. Similar reductions in the concentration of the serotonin metabolite 5-hydroxyindoleacetic acid (5-HIAA) and 5-HT uptake into cerebral tissue preparations have also been shown to occur. The most relevant index of terminal loss is the 5-HT uptake site on the presynaptic terminal (measured by radioligand binding of [^{3}H]-paroxetine), and MDMA administration decreases the number of these sites in a dose-dependent manner (Figure 4). The long-term loss of 5-HT occurs in most forebrain regions, but is of somewhat shorter duration in the striatum. The concentration of other monoamine neurotransmitters is unaltered. This selective 5-HT neurotoxicity is seen not only in rat but also in guinea pig and primate brain following MDMA. In contrast, MDMA administration to mice produces no long-term toxic effect

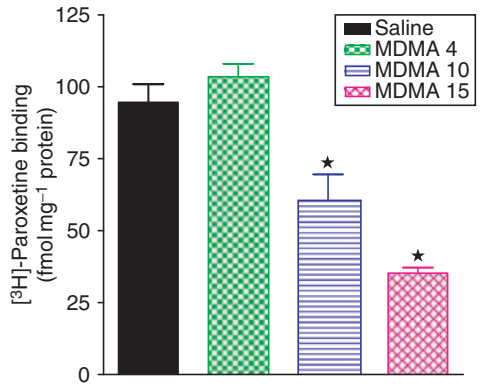

Figure 4 Long-term effect of a single administration of MDMA (4, 10, and 15 mg kg^{-1}, i.p.) on the cortical [^{3}H]-paroxetine binding 7 days after injection. Different from saline $^*P < 0.01$. Results shown as mean ± SE of mean.

on cerebral serotonin, but does produce long-term toxicity to dopamine neurons as reflected in a major loss in cerebral dopamine content and dopamine transporter.

The biochemical evidence for MDMA-induced damage to 5-HT neurons in rat brain is supported by histological observations of damage. Techniques used include staining methods of cerebral tissue and immunocytochemical analysis of brain regions indicating reductions in serotonergic axonal density and morphological evidence of terminal damage in both rat and primate brain. Recovery is region dependent but loss is still detectable in rat frontal cortex after 2–3 months (Figure 3) and in some brain regions up to a year later, while it is still apparent in primate brain after 7 years. A positron emission tomography (PET) study in baboons, using a marker for the 5-HT transporter indicated significant terminal loss for more than 9 months.

In mice, the long-term dopamine loss following MDMA leads to an enhanced expression of glial fibrillary acidic protein (GFAP), which is indicative of astrocyte hypertrophy (reactive gliosis) and thus reflective of neurodegeneration. However, in rats studies have failed to observe a GFAP increase in brain following a neurotoxic dose of MDMA, and this has led to suggestions that MDMA is not producing a true neurotoxic degeneration even though it is clearly causing 5-HT neurochemical and functional deficits.

Interestingly, MDMA itself does not cause neurotoxicity when injected directly into the brain which suggests that a neurotoxic metabolite is probably formed peripherally, and it is this metabolite that produces free radical formation and neurotoxicity in the brain.

Various approaches have been made to prevent the MDMA-induced damage in rat brain. A most effective approach has been to administer the drug when the animals are housed at low ambient temperature. This prevents both the acute hyperthermia and the long-term toxicity, and it is generally acknowledged that preventing the acute hyperthermia by either the use of a cool room or giving drugs that decrease the hyperthermic action of MDMA has the effect of severely attenuating or even preventing neurotoxic damage. There is evidence that prevention of the hyperthermia also has the effect of inhibiting the formation of free radical species in the brain following MDMA, which suggests that free radicals are a key factor in inducing the damage. This interpretation is supported by studies that have shown that administration of the free radical trapping agent α-phenyl-*N*-tert-butyl nitrone (PBN) lessens the MDMA-induced neurotoxicity, and other free radical scavengers including cysteine and ascorbic acid can also provide neuroprotection. There are compounds that provide protection against MDMA-induced neurotoxicity in rats by a mechanism not related to changes in body temperature. The selective 5-HT uptake inhibitor fluoxetine, co-administered with MDMA or even injected up to 4 days before MDMA, prevented the long-term loss of 5-HT concentration without altering MDMA-induced hyperthermia. Fluvoxamine also totally prevents MDMA neuronal damage, although protection is only observed when it is given concurrently with MDMA. Both compounds inhibit 5-HT transporter and could be blocking the entry of a toxic metabolite of MDMA into the 5-HT nerve terminal.

Not surprisingly, given the many roles of 5-HT in the brain, the damage produced by MDMA results in a variety of functional deficits in animals given a neurotoxic dose of the drug. For example, rats given a prior toxic dose of MDMA show an impaired ability to thermoregulate when exposed to a high-temperature environment; this problem is not seen in normothermic conditions. This problem of thermoregulation becomes even more apparent if lesioned rats are placed in warm room conditions and given a further small dose of MDMA, since the acute hyperthemic response is substantially prolonged. Obviously, if this observation in rats has any clinical relevance, it means that heavy recreational users might be more at risk when further doses are taken in a hot environment.

Rats with a prior neurotoxic lesion also display abnormal responses when examined in various psychological tests. For example, most studies have found that lesioned rats display greater anxiety-like behavior in the elevated plus maze and social anxiety tests.

Studies in mice are far fewer than those in rats, presumably because the long-term dopamine loss seen is thought to have little relevance to the damage seen in both primates and humans. The dopamine loss is also probably the result of increased free radical production since transgenic mice with enhanced CuZn-superoxide dismutase are resistant to MDMA-induced damage. Furthermore, toxicity is enhanced in selenium-deficient mice.

Long-Term Effects in Humans

The problems in ascertaining the doses ingested, other drugs taken, and frequency of dosing by recreational users have already been alluded to. Not only is such information not generally available, but the veracity of the subjects studied is sometimes questionable. Furthermore, direct assay of cerebral tissue is impossible except occasionally postmortem.

Proton magnetic resonance spectroscopy has been used to measure *N*-acetylaspartate a neuronal marker and myoinositol as a possible marker of gliosis. Conflicting data were obtained with one study producing data suggestive of gliosis, the other not confirming this; it is possible that differences in drug exposure in the two studies were responsible for the discrepant data. Some evidence for neuronal loss was also obtained.

There have been several imaging studies performed on recreational users, using ligands to examine the 5-HT transporter with PET and single-photon computed tomography (SPECT). Data suggested a loss of transporter site number in the brain of ecstasy users though such studies cannot determine whether this is drug related or secondary to users having a preexisting difference in serotonergic function to control subjects. Exclusive ecstasy use does not appear to decrease the striatal dopamine transporter density; that is, it is not neurotoxic to dopamine neurons. However, the combined use of amphetamine and MDMA may be associated with a reduction in striatal dopamine transporter density. These data may have substantial clinical significance since ecstasy users are often polydrug users and many of the heavy ecstasy users also take amphetamine.

Serotonin function in the brain has also been examined by use of neuroendocrine challenge tests. There was some indication that the serum prolactin response to an infusion of L-tryptophan was blunted in MDMA users, and a challenge test using fenfluramine showed a similar effect and also a blunted cortisol effect. What was striking was the fact that these abnormal responses were seen a year after the subjects had stopped using MDMA. The 5-HIAA concentration in the cerebrospinal fluid (CSF) may also be lower in MDMA users, but this is a poor marker of central 5-HT function.

There have been many studies of long-term psychological effects in recreational users of MDMA, and generally where abnormalities have been observed these have been seen long after drug use has ceased. A range of psychiatric symptoms has been reported including anxiety and panic attacks, depression, paranoid delusions, and visual hallucinations. Chronic psychosis has also been reported in a small number of subjects. The problem, however, is that these symptoms are seen in many polydrug users, and heavy users of MDMA are likely to be taking other recreational drugs, so it is unclear whether MDMA is the specific culprit for the problems seen or indeed whether some of the problems are preexisting in the personality types attracted to heavy recreational drug use.

There is much evidence that ecstasy use can produce selective cognitive impairment, and this problem can continue in drug-free persons, being more pronounced in heavy drug users. The MDMA user group display impaired immediate verbal memory and delayed visual memory as well as longer reaction time to visual and auditory stimuli. Memory deficits of MDMA users seem to be the result not only of a temporal or hippocampal dysfunction but also of a dysfunction of regions within the frontal cortex.

General Conclusions

MDMA produces a complex series of acute neurochemical events in the brain of experimental animals, primarily involving monoamine neurotransmitter function, and these changes initiate a number of physiological and behavioral changes. Similar functional changes are also seen in recreational users of the drug. The drug also produces long-term neurotoxic damage in both rodent and primate brain which is dose dependent. A major problem in interpreting the neurotoxicity data in experimental animals in terms of its relevance to human pharmacology is the doses administered. Many preclinical experimental studies on MDMA have used high doses, and it is questionable whether the effects seen are clinically relevant. While it has been argued that it is relevant to use 'interspecies scaling' to interpret results, the fact remains that similar doses (in $mg\,kg^{-1}$ quantities) can produce many comparable acute physiological and psychological effects in rodents and humans. Unfortunately, few animal studies have measured plasma drug concentrations, and without this information interpretation remains difficult. It is to be hoped that future studies will correct this oversight.

See also: Amphetamines; Dopamine – CNS Pathways and Neurophysiology; Neurotoxins; Neurotoxins and their Neurotoxicology; Serotonin-Related Psychedelic Drugs; Stimulant and Wake-Promoting Substances.

Further Reading

Baumann MH, Wang X, and Rothman RB (2007) 3,4-methylenedioxymethamphetamine (MDMA) neurotoxicity in rats: A reappraisal of past and present findings. *Psychopharmacology* 189: 407–424.

Cohen RS (1998) *The Love Drug Marching to the Beat of Ecstasy.* Binghamton: The Haworth Medical Press.

Colado MI, O'Shea E, and Green AR (2004) Acute and long-term effects of MDMA on cerebral dopamine biochemistry and function. *Psychopharmacology* 173: 249–263.

Cowan RL (2007) Neuroimaging research in human MDMA users: A review. *Psychopharmacology* 189: 539–556.

Dumont GJH and Verkes RJ (2006) A review of acute effects of 3,4-methylenedioxymethamphetamine in healthy volunteers. *Journal of Psychopharmacology* 20: 176–187.

Gouzoulis-Mayfrank E and Daumann J (2006) The confounding problem of polydrug use in recreational ecstasy/MDMA users: A brief overview. *Journal of Psychopharmacology* 20: 188–193.

Green AR, Mechan AO, Elliott JM, O'Shea E, and Colado MI (2003) The pharmacology and clinical pharmacology of 3,4-methylenedioxymethamphetamine (MDMA, 'ecstasy'). *Pharmacological Reviews* 55: 463–508.

Green AR, O'Shea E, and Colado MI (2004) A review of the mechanisms involved in the acute MDMA (ecstasy)-induced hyperthermic response. *European Journal of Pharmacology* 500: 3–13.

Parrott AC (2006) MDMA in humans: Factors which affect the neuropsychobiological profiles of recreational ecstasy users, the integrative role of bioenergetic stress. *Journal of Psychopharmacology* 20: 147–163.

Reneman L, de Win MML, van den Brink W, Booij J, and den Heeten GJ (2006) Neuroimaging findings with MDMA/ecstasy: Technical aspects, conceptual issues and future prospects. *Journal of Psychopharmacology* 20: 164–175.

Schifano F (2004) A bitter pill: Overview of ecstasy (MDMA, MDA) related fatalities. *Psychopharmacology* 173: 242–248.

Microarray use for the Analysis of the CNS

S D Ginsberg, Nathan Kline Institute, New York University School of Medicine, Orangeburg, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Conventional molecular biology techniques enable gene expression analysis in a wide variety of analytical approaches and paradigms. Experimental methods include Southern analysis (DNA detection), Northern analysis (RNA detection), polymerase chain reaction (PCR; DNA detection), reverse-transcriptase PCR (RT-PCR; RNA detection), ribonuclease (RNase) protection assay, and *in situ* hybridization, among others. The majority of these approaches typically quantitate the abundance of individual elements one at a time (or a few at a time). Recent developments in high-throughput genomic methodologies enable the assessment of dozens to hundreds to thousands of genes simultaneously in a coordinated manner. Recent advances in molecular genetics have led to sequencing of the entire human genome, as well as whole-genome sequencing of several relevant animal models. Furthermore, expression data are increasingly available for many diverse tissues throughout the body, allowing for exciting hypothesis testing of critical concepts such as development, differentiation, normative function, and, ultimately, disease pathogenesis.

Input Sources of RNA

Input sources of RNA can originate from a variety of *in vivo* and *in vitro* sources. Fresh, frozen, and fixed tissues are useful to varying degrees, depending on the paradigm and tissue quality. Many laboratories isolate genomic DNA and total RNA from paraffin-embedded fixed tissues as well as from fresh and frozen tissues. In addition, genetic material preserved by fixation is a significant resource because fresh and/or frozen tissues are not frequently available, whereas archived fixed human tissues exist in many brain banks and individual laboratories. With mRNA as a starting material, the importance of preservation of RNA integrity in tissues and cells for downstream genetic analyses cannot be emphasized enough. RNA species are particularly sensitive to RNase degradation. RNases are found in virtually every cell type and are quite stable over a broad pH range. Therefore, it is prudent to employ RNase-free precautions during all RNA extraction and analytical procedures. In addition, a laudable goal is to devise optimal methods to prepare brain tissues for downstream genetic analyses for consistency across laboratories and paradigms. A consensus protocol awaits further development and implementation. Salient variables for the accrual of postmortem human tissues include documentation of agonal state, antemortem characteristics, clinical diagnostic criteria, duration of fixation, length of storage, and postmortem interval. A useful method for assessing RNA quality in tissue sections prior to performing expression profiling studies is acridine orange (AO) histofluorescence. AO is a fluorescent dye that intercalates selectively into nucleic acids and has been used to detect RNA and DNA in brain tissues. Significantly, individual RNA species (including mRNA, rRNA, and tRNA) cannot be delineated by AO histofluorescence. Rather, this method provides a diagnostic test that is employed on adjacent tissue sections to ensure the likelihood that an individual case has abundant RNA prior to performing expensive expression profiling studies. A more thorough examination of RNA quality can be obtained via bioanalysis, which uses capillary gel electrophoresis to detect RNA quality and quantitate relative abundance.

Expression Profiling in the Central Nervous System

The central nervous system (CNS) is an exciting structure to evaluate relationships between functional genomics and expression data. Recent developments in functional genomic technologies, specifically microarray platforms, have enabled the coordinated assessment of a plethora of gene alterations that occur within the CNS. However, unlike an organ that comprises one principal cell type, the brain contains a multiplicity of both neuronal (principal neurons, interneurons, and others) and nonneuronal (glial cells, epithelial cells, inflammatory cells, vascular elements, and others) populations of cells. The intricate connectivity of synaptic-based circuits within the brain demands cognizance of rostrocaudal and mediolateral coordinates as well.

Microarray analysis has emerged as a useful and fairly cost-effective tool to assess transcript levels in a myriad of systems and paradigms. A disadvantage of these high-throughput technologies is a requirement for significant amounts of high-quality input sources of RNA for increased sensitivity and reproducibility. Whole-organism studies can generate significant input amounts of RNA species without any amplification procedures, but does not allow regional and/or

cellular specificity. Regional analysis of gene expression is a widely used paradigm due to the relatively large amounts of RNA that can be extracted from cultured cells, animal-model tissues, and postmortem human brain tissues. An advantage of regional gene expression analysis is that, in most cases, extracted RNA is sufficient to generate significant hybridization signal intensity for microarray analysis, enabling the analysis of hundreds to thousands of targets without requiring RNA amplification. However, expression profiles garnered from regional dissections cannot discriminate molecular signatures from admixed neuronal populations and nonneuronal populations within the region of interest.

Homogeneous Cell Gene Expression Analysis

An emerging goal of modern molecular neuroscience is to sample gene expression from similar cell types within a defined region without potential contamination by expression profiles of adjacent neuronal subtypes and noneuronal cells. For example, the development of a series of molecular and cellular criteria that differentiate selectively vulnerable neurons from relatively spared neurons is critical. An important factor that determines specificity of functional genomics methods is the identification and subsequent purity of captured cells for subsequent downstream genetic analyses. As such, single-cell and population-cell molecular fingerprinting requires accurate, nondestructive isolation of cells from optimally prepared tissue sections. This is a difficult task that demands a multidisciplinary approach (Figure 1). At present, an optimal methodology to assess gene expression is to evaluate single cells, either identified physiologically in living preparations or by immunocytochemical or histochemical procedures in fixed cells *in vitro* or *in vivo*. Unfortunately, the quantity of RNA harvested from a single cell, estimated to be approximately 0.1–1.0 pg, is not sufficient for standard RNA extraction. Therefore, exponential PCR-based analyses and linear RNA amplification including amplified antisense RNA (aRNA) amplification and terminal continuation (TC) RNA amplification have been used in combination with microdissection procedures to enable the use of microarray analysis. For example, specific neuronal populations can be identified by selective expression of relevant proteins, such as cholinergic basal forebrain neurons and midbrain dopaminergic neurons. Moreover, cells that lack a distinct or selective signature phenotype can be analyzed using a variety of Nissl stains (including cresyl violet and hematoxylin and eosin) for downstream genetic applications. For example, cresyl violet has been used in our laboratory to identify CA1 and CA3 pyramidal neurons in postmortem human tissue sections for subsequent single-cell microarray analysis and quantitative real-time PCR (qPCR) studies.

Microaspiration Procedures

Single-cell assessment and analysis of a homogeneous group of cells (termed population-cell analysis) requires precise nondestructive isolation from optimally prepared tissue sections. Provided that labeling procedures are performed on well-prepared tissue sections and RNase-free conditions are employed, both immunocytochemical and histochemical procedures can be used to identify specific cells of interest. Several techniques are used to aspirate individual cells or populations of cells, including single-cell microaspiration and laser capture microdissection (LCM). Single-cell microaspiration entails visualizing an individual cell (or cells) using an inverted microscope connected to a micromanipulator, a vacuum source, and an imaging workstation on an air table. Microaspiration results in the accurate dissection of the neurons of interest with minimal disruption of the surrounding neuropil and has been used on a variety of tissues and cell types, including normal samples and tissues undergoing various forms of neurodegeneration in animal models and human postmortem tissues.

Another high-throughput microaspiration device, LCM, enables rapid accession of single cells and populations for downstream molecular and cellular analyses. LCM is a popular and effective technique that was developed originally by investigators at the National Institutes of Health (NIH). Positive extraction entails pulsing an infrared laser onto thermoplastic film embedded in a specialized microfuge cap to form a thin protrusion that bridges the gap between the cap and a tissue section, effectively adhering to target cells. Lifting the thermoplastic cap separates targeted cells from surrounding undisturbed tissue. Negative extraction (or noncontact laser extraction) procedures employ a near-infrared laser source to cut around the area of interest within a tissue section, and the microdissected material is catapulted into a microfuge tube. Positive and negative extraction methods enable captured cells to be examined microscopically, confirming the identity and quality of isolated cell populations. Individual cells or dozens, hundreds, or thousands of cells can be collected rapidly by LCM. RNA, DNA, and protein extraction methods can be performed on microdissected cells. LCM is used conventionally to enable cellular isolation for RNA extraction and subsequent microarray analysis.

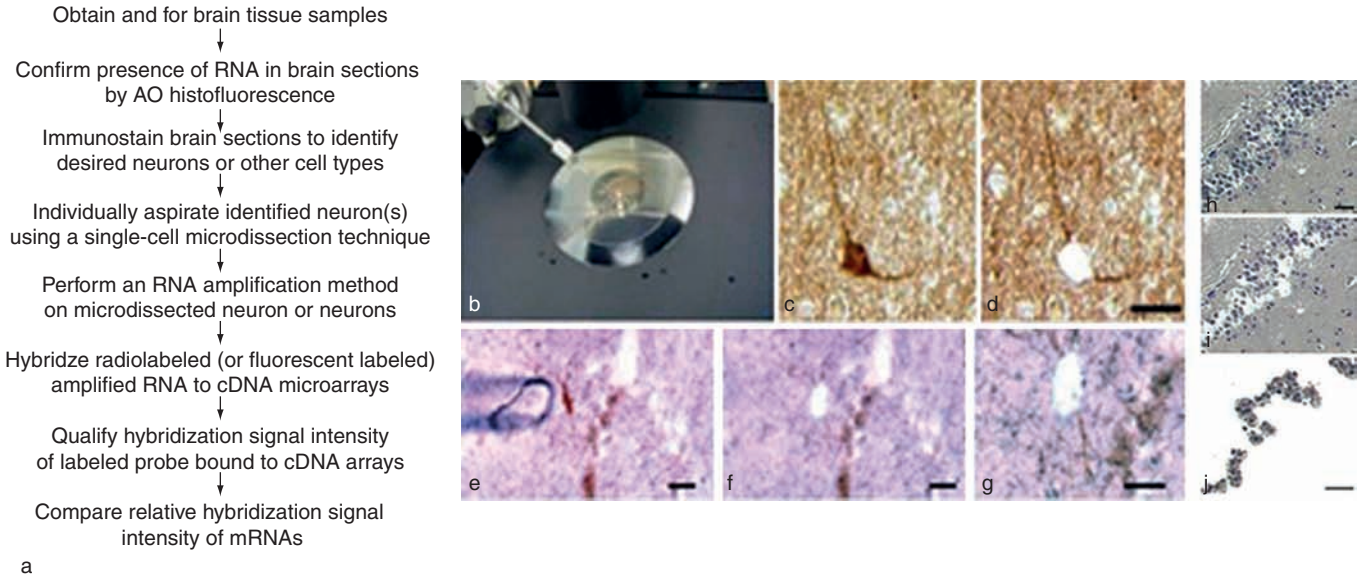

Figure 1 Microdissection combined with high-throughput gene expression analysis using array platforms: (a) outline of the general procedures; (b) photomicrograph illustrating the placement of a human postmortem tissue section of the basal forebrain onto the microaspiration apparatus; (c) section immunolabeled with an anti-neurofilament antibody depicting a layer II entorhinal cortex stellate cell obtained postmortem from a normal control human brain; (d) the same section as in (c) following microdissection of the immunostained neuron; (e) human cholinergic basal forebrain neurons from the nucleus basalis dual-labeled for the cholinergic marker $p75^{NTR}$ (brown cell bodies) and galanin (black punctate fibers; (f) micrograph of the same tissue section in (e) following microaspiration of the cholinergic neuron at low magnification; (g) same field as in (e) and (f) at higher magnification; (h) photomicrograph of cresyl violet-stained human hippocampal granule cells within a tissue section prior to LCM; (i) photomicrograph after a thermoplastic cap is removed following laser pulses over the desired cells, leaving spaces where microdissected granule cells originally resided; (j) captured cells visualized by placing the cap on a clean slide for contrast. In (e), a microaspirating pipette is visualized in the left plane of the field. AO, acridine orange; LCM, laser capture microdissection. Scale bar = 25 μm (c (scale as in (d), d), h, i(scale as in (h)), 40μm (e–g), 30 μm (j). Adapted from Ginsberg SD, Hemby SE, Mufson EJ, and Martin LJ (2006) Cell and tissue microdissection in combination with genomic and proteomic applications. In: Zaborszky L, Wouterlood FG, and Lanciego JL (eds.) *Neuroanatomical Tract Tracing, Vol. 3: Molecules, Neurons, and Systems,* pp. 109–141. New York: Springer.

Microarray Platforms

Technical advances have enabled the development of a wide variety of microarray platforms that serve as the substrate for high-throughput analysis of dozens to hundreds to thousands of genes simultaneously. Microarrays represent miniaturized, high-density dot-blots that exploit the complementary hybridization between nucleic acids. Synthesis of cDNA microarrays entails adhering cDNAs or expressed sequence-tagged cDNAs (ESTs) to solid supports such as glass slides, plastic slides, or nylon membranes. A parallel technology uses photolithography to adhere oligonucleotides to array media. The anchored cDNA/EST/oligonucleotide is commonly called the microarray target. The target length varies from short oligonucleotides to chromosomal fragments. When the input RNA is not amplified (as in the case of regional assessments), probe generation is performed via reverse transcription, and a detection moiety is incorporated directly into the cDNA strand. If input RNA is amplified through an RNA amplification methodology, then the amplified material is labeled and used to probe the microarrays. Arrays are washed to remove nonspecific background hybridization and imaged using a laser scanner for biotinylated/fluorescently labeled probes and a phosphor imager for radioactively labeled probes. A target-labeled probe complex emits a quantifiable signal that is proportional to the abundance of the labeled probe in the sample. The specific signal intensity (minus background) is expressed as a ratio of the total hybridization signal intensity, reducing variation across array platforms due to differences in the specific activity of the probe and absolute quantity of probe present. Quantification of hybridization signal intensity is performed to evaluate relative expression levels of each cDNA, EST, or oligonucleotide feature on the array platform. Gene expression is then assessed using statistical and informatics modules that coordinate statistical analyses. Computational analysis is critical due to the large number of data points that are generated from a single assay.

RNA Amplification Strategies

An RNA amplification technique is often required to perform microarray analysis and related downstream genetic analyses when using small sample inputs as starting material. PCR-based amplification methods are not optimal because exponential amplification can skew the original quantitative relationships between genes from an initial population. Linear RNA amplification is another strategy that has been used successfully to generate input RNA for molecular-profiling studies. One method of linear amplification, termed aRNA amplification, uses a T7 RNA polymerase-based amplification procedure that allows quantitation of relative gene expression levels from microdissected cells and tissues (Figure 2). aRNA maintains a proportional representation of the size and complexity of the initial mRNAs. Relative mRNA expression levels can be quantified using cDNA or oligonucleotide arrays by probing with fluorescent-and/or biotin-tagged nucleotide triphosphates incorporated directly into aRNA products. Alternatively, radiolabeled nucleotide triphosphates can be incorporated into aRNA products and used for hybridization to membrane-based cDNA arrays via reverse Northern analysis to known cDNAs and/or ESTs. aRNA products are biased toward the 3′ end of the transcript because of the initial priming at the poly(A)+ RNA tail. This 3′ bias exists for all amplified aRNA products, and relative levels of gene expression can be compared. Although aRNA is often a technically involved procedure, successful results have been generated with micro-aspirated animal model and human tissues using a wide variety of microarray platforms. Modifications of the initial aRNA procedure have been used, and several aRNA-based kits are available commercially.

A variety of strategies have been developed by independent laboratories to improve RNA amplification, including combining PCR with aRNA and isothermal RNA amplification, among others. One obstacle is the problematic second-strand cDNA synthesis. This impediment is not specific to the aRNA protocol. Rather, this issue is endemic to all current RNA amplification methods. Key factors to improving RNA amplification include increasing the efficiency and/or obviating the second-strand cDNA synthesis and allowing for flexibility in the placement of bacteriophage transcription promoter sequences.

A new RNA amplification procedure has been developed in our laboratory that uses a method of TC (Figure 3). TC RNA amplification entails synthesizing first-strand cDNA complementary to the RNA template and generating second-strand cDNA complementary to the first-strand cDNA. RNA is amplified via an *in vitro* transcription procedure using the newly formed double-stranded cDNA as template. Synthesis of the first-strand cDNA complementary to template mRNA is performed by two oligonucleotide primers, a poly(dT) primer and a TC primer. The poly(dT) primer is similar to conventional primers that anneal to the poly(A)+ sequence present on most mRNAs. The TC primer consists of an oligonucleotide sequence at the 5′ terminus and a short span of three cytidine triphosphates or guanosine triphosphates at the 3′ terminus. TC RNA transcription can

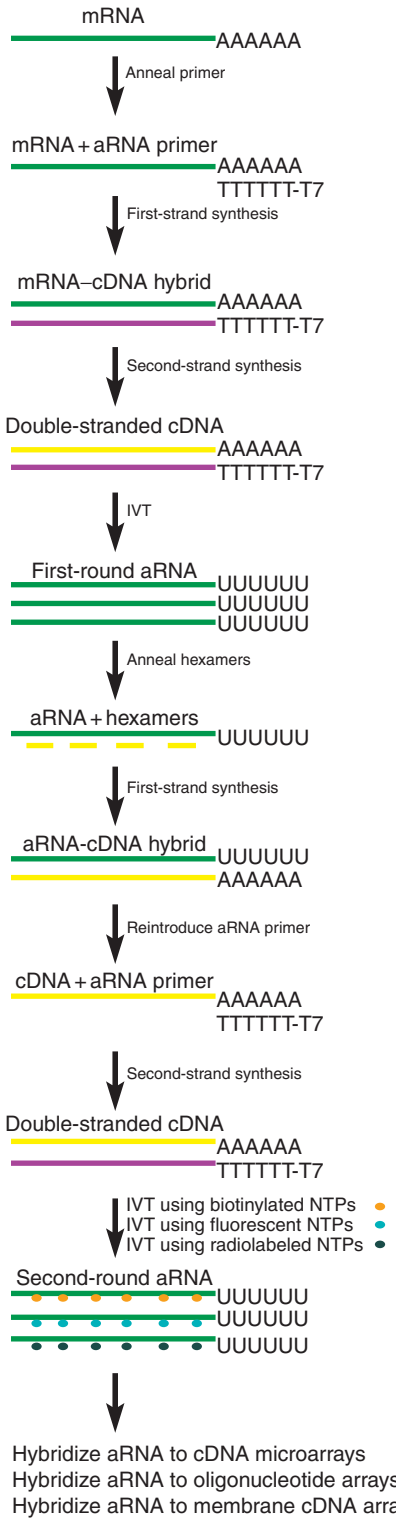

Figure 2 aRNA amplification scheme. An aRNA primer {oligo-d(T)T7} primer is hybridized directly to poly(A)+ mRNAs. A double-stranded mRNA–cDNA hybrid is formed (first-strand synthesis) by reverse transcribing the primed mRNAs with reverse transcriptase. The double-stranded mRNA–cDNA hybrid is converted into double-stranded cDNA, forming a functional T7 RNA polymerase promoter. A first round of aRNA synthesis occurs via *in vitro* transcription using T7 RNA polymerase and nucleotide triphosphates. A second round of aRNA amplification begins by annealing random hexamers to the newly formed aRNA and performing first-strand synthesis. The aRNA primer is then reintroduced, which binds to the poly(A)+ sequence on the newly synthesized cDNA strand. A double-stranded cDNA template is formed by second-strand synthesis. A second round of aRNA products is produced by *in vitro* transcription using biotinylated, fluorescent, or radiolabeled nucleotide triphosphates. Labeled second-round aRNA products are used as probes to hybridize to cDNA microarrays, oligonucleotide arrays, or membrane-based cDNA arrays. IVT, *in vitro* transcription; NTPs, nucleotide triphosphate. Adapted from Ginsberg SD (2005) RNA amplification strategies for small sample populations. *Methods* 37: 229–237; Ginsberg SD and Mirnics K (2006) Functional genomic methodologies. *Progress in Brain Research* 158: 15–40.

be driven using a promoter sequence attached to either the 3′ or 5′ oligonucleotide primers (Figure 3). Therefore, transcript orientation can be in an antisense orientation (similar to conventional aRNA methods) when the bacteriophage promoter sequence is placed on the poly(dT) primer or in a sense orientation when the promoter sequence is attached to the TC primer, depending on the design of the experimental paradigm (Figure 3). TC RNA amplification enables the quantitative assessment of a large proportion of genes, as evidenced by bioanalysis and microarray analysis in animal model and human postmortem brain tissues, including single cells and populations of homogeneous neurons. Significantly, the increased sensitivity of the TC RNA amplification method appears greatest for genes with relatively low abundance. Moreover, background hybridization is significantly attenuated when using TC RNA amplification.

Validation of Microarray Results via Quantitative Polymerase Chain Reaction

Amplification of genetic signals can be performed at both the DNA and RNA levels, and final amplified-products are either DNA or RNA. A common method for DNA amplification is PCR. However, PCR is an exponential, nonlinear amplification, and variation can occur within individual mRNA species of different molecular mass and base-pair composition. PCR-based methods tend to amplify abundant genes over rare genes and may distort quantitative relationships among gene populations. Real-time qPCR can quantitate PCR product formation during each cycle of amplification and has lessened concerns that are associated with conventional PCR methods. Other advantages of real-time qPCR include high-throughput capabilities, the ability to simultaneously multiplex reactions, enhanced sensitivity, reduced interassay variation, and lack of post-PCR manipulations. A variety

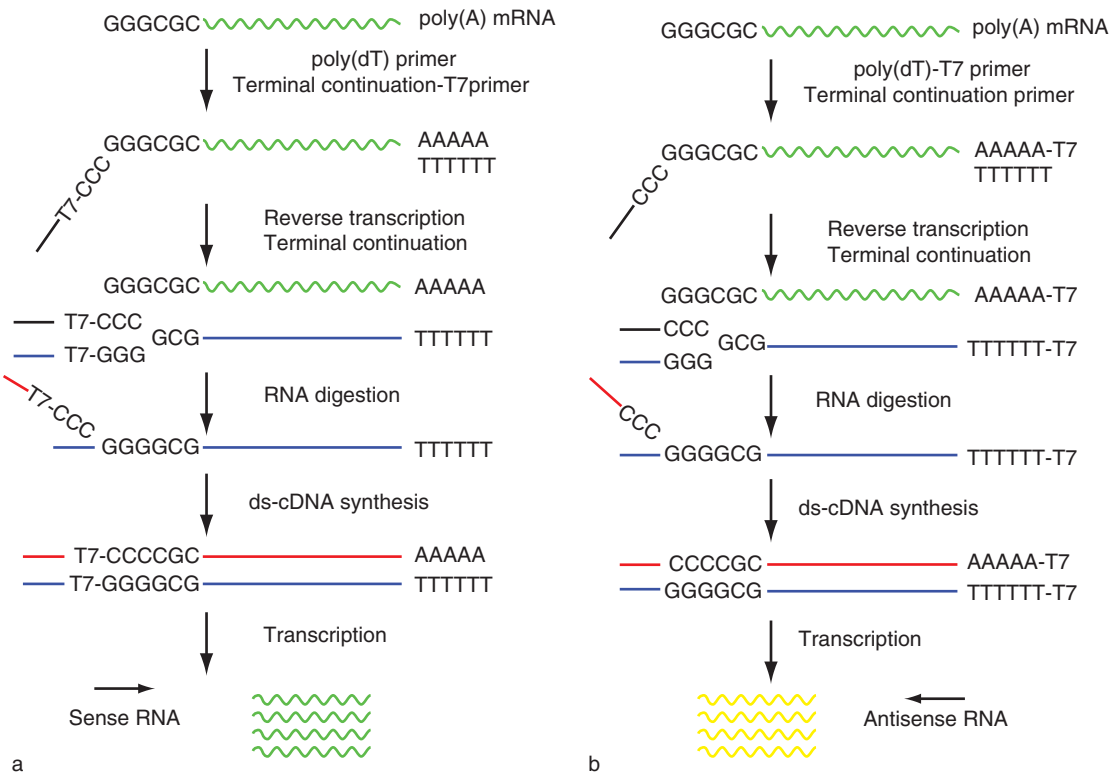

Figure 3 Overview of the TC RNA amplification method: (a) sense orientation; (b) antisense orientation. As shown in (a), a TC primer (containing a bacteriophage promoter sequence for sense orientation) and a poly(dT) primer are added to the mRNA population to be amplified (green rippled line). First-strand synthesis (blue line) occurs as an mRNA–cDNA hybrid is formed following reverse transcription and TC of the oligonucleotide primers. After an RNase H digestion step to remove the original mRNA template strand, second-strand synthesis (red line) is performed using *Taq* polymerase. The resultant double-stranded product is used as a template for *in vitro* transcription, yielding high-fidelity linear RNA amplification of sense orientation (green rippled lines). The scheme in (b) is similar to (a), illustrating the TC RNA amplification procedure amplifying RNA in the antisense orientation (yellow rippled lines). TC, terminal continuation. Adapted from Che S and Ginsberg SD (2004) Amplification of transcripts using terminal continuation. *Laboratory Investigation* 84: 131–137; Ginsberg SD (2005) RNA amplification strategies for small sample populations. *Methods* 37: 229–237.

of dye chemistries are employed in qPCR assays, including hydrolysis probes, molecular beacons, and double-stranded DNA binding dyes.

The most well-known hydrolysis probes are the TaqMan assay. In this assay, the *Taq* polymerase enzyme cleaves the TaqMan probe during the PCR extension phase. The TaqMan probe is dual labeled, with a reporter dye and a quenching dye at two separate ends. When the probe is in its free form, fluorescence emission of the reporter dye is absorbed by the quenching dye by fluorescence resonance energy transfer. An increase of reporter fluorescence emission occurs when the separation of the reporter and quencher dyes takes place during nuclease degradation in the PCR reaction. This process takes place during each PCR cycle and does not interfere with the exponential accumulation of amplified product. Molecular beacons are probes that form a stem–loop structure from a single-stranded DNA molecule. Molecular beacons are particularly useful for identifying point mutations because these probes can detect targets that differ by only a single nucleotide. DNA-binding dyes such as SYBR green incorporate selectively into double-stranded DNA. SYBR green emits undetectable fluorescence levels when it is in its free form. Once bound to double-stranded DNA, an intense fluorescent signal is emitted. qPCR has emerged as an excellent companion technology to microarray analysis in that the differential regulation of individual transcripts observed via array platforms can be tested with relative ease and quickness via qPCR-based assays.

Caveats

The application of microarray technology in the neuroscience disciplines is still in its infancy. Most high-throughput transcriptome-profiling studies are moderately expensive, labor intensive, and somewhat time consuming. To generate meaningful data sets, an appropriate experimental design is imperative prior to initiating actual studies. A concern is that

significant variation exists between and across array platforms. For example, significant differences exist in microarray production (including target deposition vs. *in situ* synthesis), microarray target length (ranging from >20 to >50 bp oligonucleotides vs. partial ESTs to full-length cDNAs and chromosomal fragments), probe-labeling procedures (e.g., radioactive labeling, direct or indirect fluorochrome incorporation, and gold-particle deposition), amplification procedures (including PCR-based amplification, T7-based RNA amplification, isothermal amplification, and microRNA amplification), and a dizzying array of hybridization procedures. Each of these procedural and technical variables may have substantial effects on the subsequent data sets that emerge from the microarrays. As a result of these potential pitfalls, negative findings on any array platform must be interpreted cautiously. Significantly, independent validation methods of transcript analysis (Northern hybridization, RNase-protection assay, and *in situ* hybridization, among others) are not made obsolete by high-throughput transcription-profiling technologies. Rather, these traditional (relatively low-) throughput methods often have important advantages over the high-throughput assays, especially when a specific gene of interest is being hypothesis tested. Also, an important caveat about microarray analysis is that coordinate changes in proteins encoded by the genes of interest are common but not absolute.

Conclusion

Simply stated, when deciding whether to use a specific functional genomic methodology, the most crucial aspect to consider is the question the researcher is interested in answering; only then can the researcher determine the method(s) that are appropriate to perform the experiments. Advancements in functional genomics have led to the development of high-throughput gene-profiling techniques to quantify the expression levels of multiple transcripts simultaneously within discrete brain regions and specific cell types. Microarray-based paradigms are enabling the analysis of coordinate changes in gene expression that were not feasible prior to the development of these exciting technologies. However, the use of high-throughput analysis does not obviate the need for traditional methods of expression-level quantitation for validation purposes in a lower-throughput fashion. Thus, a combination of multidisciplinary approaches is ideal for verification of gene-expression-level alterations, with the explicit knowledge that the sum of the evaluations may be more informative and reflect the actual biology of the system better than an individual method.

See also: Gene Expression in Normal Aging Brain; Gene Expression Regulation: Chromatin Modification in the CNS; Gene Expression Dysregulation in CNS Pathophysiology; Genomic Disorder and Gene Expression in the Developing CNS; Hippocampus: Molecular Anatomy; Single Cell PCR Coupled with Electrophysiology.

Further Reading

Blumberg DD (1987) Creating a ribonuclease-free environment. *Methods in Enzymology* 152: 20–24.

Bonner RF, Emmert-Buck M, Cole K, et al. (1997) Laser capture microdissection: Molecular analysis of tissue. *Science* 278: 1481–1483.

Bustin SA (2002) Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): Trends and problems. *Journal of Molecular Endocrinology* 29: 23–39.

Che S and Ginsberg SD (2004) Amplification of transcripts using terminal continuation. *Laboratory Investigation* 84: 131–137.

Eberwine J, Yeh H, Miyashiro K, et al. (1992) Analysis of gene expression in single live neurons. *Proceedings of the National Academy Science of the United States of America* 89: 3010–3014.

Farrell RE Jr. (1998) *RNA Methodologies*, 2nd edn. San Diego, CA: Academic Press.

Ginsberg SD (2005) RNA amplification strategies for small sample populations. *Methods* 37: 229–237.

Ginsberg SD and Che S (2004) Combined histochemical staining, RNA amplification, regional, and single cell analysis within the hippocampus. *Laboratory Investigation* 84: 952–962.

Ginsberg SD, Hemby SE, Mufson EJ, and Martin LJ (2006) Cell and tissue microdissection in combination with genomic and proteomic applications. In: Zaborszky L, Wouterlood FG, and Lanciego JL (eds.) *Neuroanatomical Tract Tracing, Vol. 3: Molecules, Neurons, and Systems*, pp. 109–141. New York: Springer.

Ginsberg SD and Mirnics K (2006) Functional genomic methodologies. *Progress in Brain Research* 158: 15–40.

Kacharmina JE, Crino PB, and Eberwine J (1999) Preparation of cDNA from single cells and subcellular regions. *Methods in Enzymology* 303: 3–18.

Mullis KB (1990) The unusual origin of the polymerase chain reaction. *Scientific American* 262: 56–65.

Sambrook J and Russell DW (2001) *Molecular Cloning: A Laboratory Manual*, 3rd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Microcephaly Vera

G H Mochida, Beth Israel Deaconess Medical Center, Massachusetts General Hospital, and Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

What Is Microcephaly?

Microcephaly refers to a condition in which the human brain fails to achieve normal growth. Clinically, a diagnosis of microcephaly is made by measuring the occiptofrontal circumference (OFC) using a non-stretchable tape measure. When the OFC is below a certain cutoff for the person's age and sex, a diagnosis of microcephaly is made. In the clinical setting, a cutoff of −2 SD below the mean is often used, but in the research setting a stricter cutoff of −3 SD is more commonly used.

Microcephaly is a highly heterogeneous condition, and its etiology can be divided into environmental and genetic causes. The environmental causes are numerous and include intrauterine infections, intrauterine exposure to toxic substances, and hypoxic–ischemic injury before the early postnatal period. Many different types of environmental insults that affect the development of the brain while it is still growing in size can lead to microcephaly. The genetic causes of microcephaly are also diverse. In many cases, microcephaly is one of the clinical signs of a syndrome that also affects other organ systems or shows characteristic dysmorphic features. Examples include Rubinstein–Taybi syndrome, Seckel syndrome, and Smith–Lemli–Opitz syndrome. However, in some cases, the brain is the only affected organ, and no dysmorphic features are seen. Microcephaly vera is a condition that belongs to this latter group.

Historical Note and Nomenclature

The term microcephaly vera (literally, 'true' microcephaly) was first coined by Giacomini in the late nineteenth century. This term was used to describe a condition in which no anatomical abnormalities of the brain (except for its smallness) were identified. In contrast to microcephaly vera, the term 'microcephaly spuria' ('false' microcephaly) was used to describe a condition in which destructive lesions were in evidence. The essence of Giacomini's definition of microcephaly vera has not changed to date. Currently, in the narrow and strict definition of the term, it refers to a condition that is characterized by the constellation of clinical manifestations including severe congenital microcephaly, normal to mildly delayed motor development, mental retardation, and a lack of motor signs such as spasticity. The term 'primary autosomal recessive microcephaly' may also be used to describe this condition. In a broader sense, the term microcephaly vera is sometimes used to refer to various forms of genetic microcephaly without other organ involvement or dysmorphic features. However, this broader definition encompasses a highly heterogeneous group of brain malformations and can lead to confusion. Thus, the usage of the term microcephaly vera in the broader definition should be avoided, and in this article, microcephaly vera in the narrow sense is discussed.

Epidemiology

Microcephaly vera has been reported from different populations and ethnic groups. However, because it shows autosomal recessive inheritance, it appears to be much more common in the countries and regions where the rate of consanguineous marriage is high. Many genetic studies have been conducted in such populations, including northern Pakistan and the Arabic Middle East. Although population-based epidemiological data of microcephaly vera are not available, the prevalence may be as high as 1/10 000 in some populations with a high rate of consanguinity. It appears that the male:female ratio is 1:1, and this is consistent with its autosomal recessive mode of inheritance.

Clinical and Radiological Manifestations

Clinical Presentation

Patients with microcephaly vera present with microcephaly at birth (congenital microcephaly). Therefore, it is important to confirm the head circumference at birth when a patient presents after the neonatal period. It is also important to measure the parents' head circumferences because they are usually normal in microcephaly vera. If one of the parents has a small head circumference, the autosomal dominant form of microcephaly may be suspected. This is a separate entity from microcephaly vera and is often associated with normal intelligence. In patients with microcephaly vera, microcephaly may be detected by prenatal ultrasound during the third trimester. Complications during the perinatal period are not common, and weight and length at birth are generally normal. Early motor milestones, such as rolling over, sitting, and independent walking, are normal or mildly delayed. Delay in cognitive development and subsequent mental retardation are always present.

The severity of mental retardation is often of moderate degree. Some patients acquire language skills, often consisting of single words. The final head circumference generally is between −4 and −12 SD. Height and weight are generally within the normal range, and when significant short stature is noted, other syndromes such as Seckel syndrome (characterized by pre- and postnatal growth retardation and unique facial features) may need to be considered. There are usually no dysmorphic features, except for sloping forehead and relative prominence of the ears, both of which are due to severe decrease in the size of the cranium. Many patients with microcephaly vera acquire basic self-care skills. Social interaction is not severely impaired, and many patients have a friendly personality. Autistic symptoms and severe behavioral disturbances appear to be rather uncommon. Epilepsy is also uncommon. Although patients with microcephaly vera and epilepsy have been reported, it is not known if the incidence of epilepsy among patients with microcephaly vera is higher than that of the general population. Organs other than the brain are apparently not affected, and usually there are no associated signs or symptoms outside the central nervous system.

Neuroimaging

Detailed radiological studies of microcephaly vera have not been reported, but available data suggest that the basic structure of the brain is generally well preserved. In terms of the brain volume decrease, the cerebrum appears to be most severely affected, although the cerebellum and brain stem are also

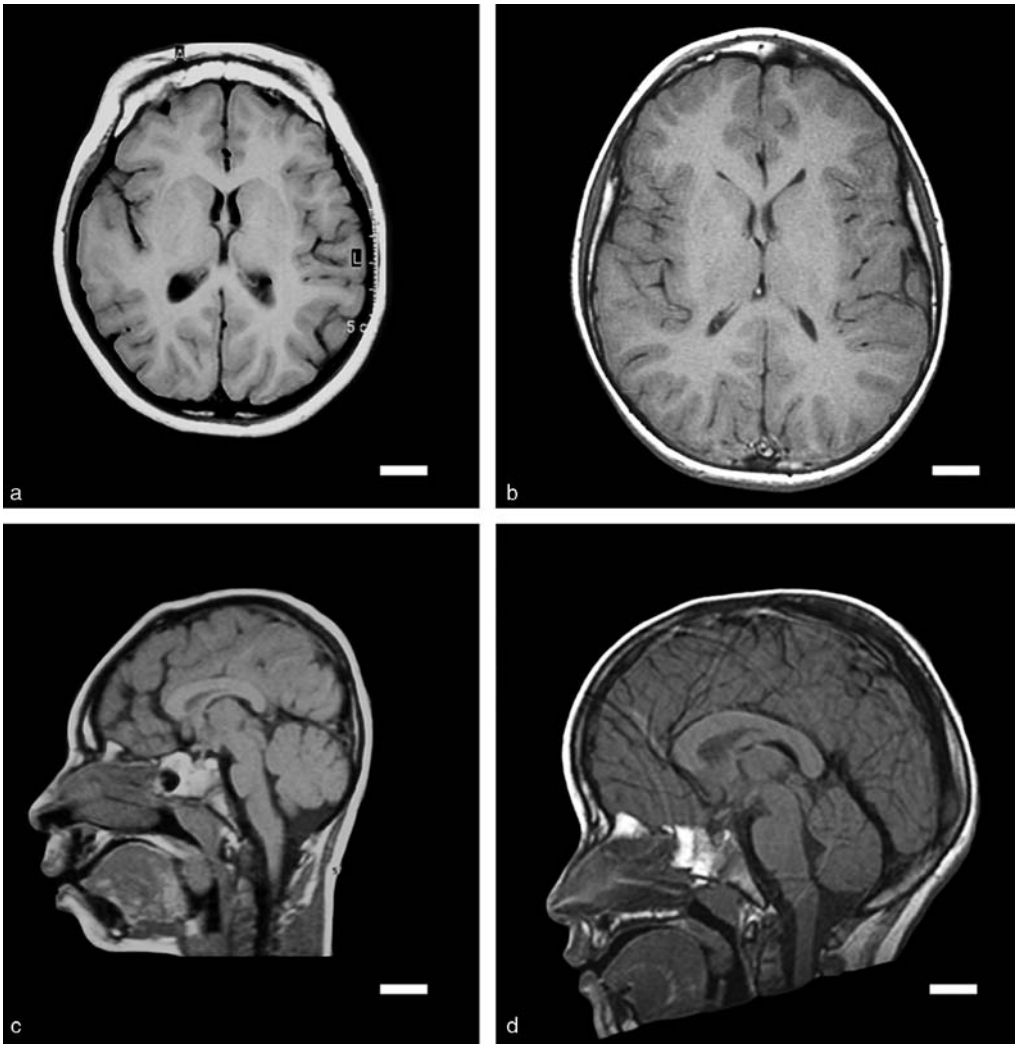

Figure 1 Brain MRI of microcephaly vera. Axial T1-weighted MRI image of a 9-year-old boy with microcephaly vera and an *ASPM* mutation (a) reveals a mildly simplified gyral pattern and significantly decreased brain volume compared to an age-matched control (b). Sagittal T1-weighted MRI image of a 5-year-old boy with microcephaly vera and an *ASPM* mutation (c). shows essentially normal architecture of the midline structure, but severe craniofacial disproportion is evident compared to an age-matched control (d). Scale bars = 2 cm.

smaller (Figure 1). Sagittal brain magnetic resonance imaging (MRI) images are often striking for severe craniofacial disproportion (i.e., a small cranium compared to the size of the facial structure). The gyral pattern of the cerebral cortex may be normal or mildly simplified. At least one patient with microcephaly vera and a mutation in the *MCPH1* gene has been shown to have periventricular heterotopia, but it is not known how common this finding is among patients with microcephaly vera.

Treatment and Prognosis

There is no treatment for the underlying biological defects of microcephaly vera, and like other disorders with mental retardation, early intervention with special education and therapy is important to the functional outcome of the individual. It is not a progressive disorder, and many patients acquire some language (often single words) and basic self-care skills. Microcephaly vera is usually not associated with epilepsy, but rare patients are known to have epilepsy. In such instances, the treatment of seizure is essentially the same as in other cases of epilepsy. No data exist on the life expectancy of patients with microcephaly vera. However, it is believed that life expectancy is not significantly shorter than that of the general population.

Genetics and Biology

Genetic Loci and Causative Genes for Microcephaly Vera

Even before the era of modern human genetic research, it was known that many microcephaly syndromes showed autosomal recessive inheritance, and microcephaly vera was no exception. Microcephaly vera that fits the previously discussed clinical description is inherited in an autosomal recessive manner. In recent years, the genetics of microcephaly vera has been extensively investigated, and it has been found that it is genetically highly heterogeneous. Six genetic loci for microcephaly vera have been identified, and these loci are named MCPH1 through MCPH6 (Table 1). There is no clear genotype–phenotype correlation, and the clinical phenotype of a patient cannot predict the underlying causative locus. It appears that MCPH5 is the most common locus for microcephaly vera, and it may account for nearly half of all cases of microcephaly vera in the limited populations studies that have been performed. Still, up to 20–25% of the pedigrees with microcephaly vera do not map to any of the known six loci, and it is almost certain that there are other loci yet to be identified. Among these six loci, the causative genes have been identified in four (Table 1). MCPH1 is caused by the gene of the same name, *MCPH1* (also known as *microcephalin* and *BRIT1* (BRCT-repeat inhibitor of TERT expression)), MCPH3 by *CDK5RAP2* (cyclin-dependent kinase 5 regulatory subunit-associated protein 2), MCPH5 by *ASPM* (abnormal spindle-like, microcephaly associated), and MCPH6 by *CENPJ* (centromeric protein J). Only a few mutations have been reported for each of *MCPH1*, *CDK5RAP2*, and *CENPJ*, but more than 20 different nonsense or frameshift mutations have been reported for *ASPM*. There appear to be no particular 'hot spots' for mutations in *ASPM*. Although the head circumference of patients with *ASPM* mutations varies widely from approximately –4 to –12 SD, the location of a mutation within the gene does not predict the severity of reduction of the brain size. Thus, it is speculated that all the mutations that have been reported so far lead to complete loss of function of the protein.

Biological Functions of the Genes for Microcephaly Vera

The essentially identical phenotype caused by the different genes for microcephaly vera suggests that these genes may be involved in the same or related biological process. This speculation has largely proven correct. All the four known genes appear to be important to regulation of cell division of neuronal progenitor cells. Among these, ASPM, CDK5RAP2, and CENPJ are

Table 1 Genetic loci and causative genes for microcephaly vera

Locus	*Gene*	*Chromosome location*	*Gene function*
MCPH1	*MCPH1*	8p23.1	DNA damage response/cell cycle checkpoint regulation
MCPH2	Not known	19q13.1–13.2	Not known
MCPH3	*CDK5RAP2*	9q34	Mitotic centrosome regulation?
MCPH4	Not known	15q14–22.2	Not known
MCPH5	*ASPM*	1q31.3	Mitotic spindle/centrosome regulation
MCPH6	*CENPJ*	13q12.2	Mitotic centrosome regulation?
Others	Not known	?	Not known

shown to localize to the spindle pole/centrosome during mitosis.

ASPM ASPM is a large protein with 3477 amino acids. It consists of four distinguishable regions: a putative N-terminal microtubule-binding domain, a calponin homology domain, a domain consisting of IQ repeats, and a C-terminal region. Calponin homology domain is found in many actin-binding proteins, and IQ repeats (the name derives from isoleucine (I) and glutamine (Q), which are key amino acids that define the repeat structure) is shown to bind to calmodulin in some proteins, although the functions of these domains in ASPM are not known. The gene is the homolog of the *asp* (abnormal spindle) gene in *Drosophila melanogaster*. The Asp protein localizes to the spindle pole/centrosome during mitosis, and mutations in *asp* result in abnormal morphology of mitotic spindle and mitotic arrest. The human ASPM protein has also been shown to localize to the spindle pole/centrosome during mitosis. *Aspm* (the mouse homolog of *ASPM*) is highly expressed in the telencephalic ventricular zone during the period of active neurogenesis and rapidly downregulated once the period is over. Interestingly, *Aspm* is also expressed in other organs where there is active cell division, and this seems to be true in humans as well. It is not known why mutations in *ASPM* do not cause obvious defects in the development of the other organs.

CDK5RAP2 and CENPJ CDK5RAP2 and CENPJ comprise 1893 and 1338 amino acids, respectively. Their roles are not precisely understood, but the protein products of both genes have been shown to localize to the centrosome during mitosis. Studies of their mouse homologs showed that both genes are highly expressed during the period of neurogenesis in the neuroepithelium of the forebrain. The brain and spinal cord showed the highest expression, but, like *Aspm*, the expression was seen widely in other organs in the developing mouse embryo.

MCPH1 MCPH1 is an 835-amino-acid protein with three BRCT (BRCA1 C-terminal) domains. BRCT domains are often found in proteins involved in regulating cell cycle checkpoint in response to DNA damage, and MCPH1 is also thought to function in this capacity. A high level of *MCPH1* gene expression is observed in the neuroepithelium along the lateral ventricles of the developing mouse brain. Interestingly, MCPH1 seems to function, at least in part, in the same DNA damage response pathway as ATR (ataxia–telangiectasia and RAD3-related), which is one of the causative genes for Seckel syndrome. Seckel syndrome, like microcephaly vera, is characterized by severe congenital microcephaly but, in addition, shows severe pre- and postnatal growth retardation. Both *MCPH1* and *ATR* mutant cells show defective G_2–M checkpoint arrest and nuclear fragmentation after DNA damage, but *MCPH1* mutant cells show premature chromosome condensation, which is not seen in ATR mutant cells, suggesting an additional ATR-independent role of MCPH1.

Genes for Microcephaly Vera and Brain Evolution

One of the most striking characteristics of the human brain is its large size in relation to the body size. The molecular and genetic basis of this dramatic volume expansion during evolution is not well understood. However, insufficient growth of the brain without major disruption in the architecture in microcephaly vera raised a possibility that the genes for microcephaly vera may have played a role in the volume expansion of the human brain during evolution. Several studies have examined the evolution of the microcephaly vera genes to explore this possibility. These studies searched for evidence of positive selection of these genes by calculating the parameter called the K_a/K_s ratio of the genes. The K_a/K_s ratio of a gene is a ratio of nonsynonymous nucleotide changes (i.e., nucleotide changes that alter resulting amino acids) and synonymous nucleotide changes (i.e., nucleotide changes that do not alter resulting amino acids) during the course of evolution. When the ratio is low, the gene is under negative selection, meaning that any change in the protein sequence is likely deleterious, and it is unlikely that the gene has played important roles in evolutionary changes seen among organisms. On the other hand, if the gene has a high K_a/K_s ratio, the gene is under positive selection (i.e., rapidly evolving), and it is more likely that the gene has played a role for the organism to have acquired new characteristics during evolution. *MCPH1* was found to have evidence of positive selection during primate evolution, particularly prior to the ancestor of great apes. In contrast, *ASPM* was shown to have evidence of positive selection in the more recent history of primate evolution. The evidence for positive selection was the strongest in the human lineage and also seen to a lesser extent in the African hominoid clade (human, chimpanzee, and gorilla) or great apes (the African hominoid clade and orangutan). There are also data to suggest evidence of positive selection for *CDK5RAP2* and *CENPJ* in primates.

Experimental data to prove the relationship between the positive selection of these genes and the evolutionary volume expansion of the human brain

are difficult to obtain. However, it is remarkable that all four known genes for microcephaly vera show evidence of positive selection, and it appears very likely that these genes were key molecules for the volume expansion of the brain in humans and other primates.

Future Directions

Although significant progress in our understanding of microcephaly vera has been made during the past decade, there are still many unanswered questions. First, it appears that there are additional microcephaly vera loci yet to be identified. Second, the functions of the known causative genes are still not well understood. For example, we know that *ASPM*, *CDK5RAP2*, and *CENPJ* are associated with the spindle pole and centrosome, but how mutations in these proteins lead to human microcephaly is far from clear. Third, the potential role of these genes in human brain evolution is strongly suggested but has been difficult to prove experimentally. By addressing these questions, we will gain knowledge not only about microcephaly vera but also about general mechanisms of human brain development and evolution, and this will likely have a major impact on much broader fields of neuroscience and biology.

See also: Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brain Scaling Laws; Holoprosencephaly; Joubert Syndrome.

Further Reading

Alderton GK, Galbiati L, Griffith E, et al. (2006) Regulation of mitotic entry by microcephalin and its overlap with ATR signalling. *Nature Cell Biology* 8: 725–733.

Bond J, Roberts E, Mochida GH, et al. (2002) ASPM is a major determinant of cerebral cortical size. *Nature Genetics* 32: 316–320.

Bond J, Roberts E, Springell K, et al. (2005) A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. *Nature Genetics* 37: 353–355.

Bond J, Scott S, Hampshire DJ, et al. (2003) Protein-truncating mutations in ASPM cause variable reduction in brain size. *American Journal of Human Genetics* 73: 1170–1177.

Gilbert SL, Dobyns WB, and Lahn BT (2005) Genetic links between brain development and brain evolution. *Nature Reviews Genetics* 6: 581–590.

Jackson AP, Eastwood H, Bell SM, et al. (2002) Identification of microcephalin, a protein implicated in determining the size of the human brain. *American Journal of Human Genetics* 71: 136–142.

Kouprina N, Pavlicek A, Collins NK, et al. (2005) The microcephaly ASPM gene is expressed in proliferating tissues and encodes for a mitotic spindle protein. *Human Molecular Genetics* 14: 2155–2165.

Kouprina N, Pavlicek A, Mochida GH, et al. (2004) Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. *PLoS Biology* 2: E126.

Mochida GH and Walsh CA (2001) Molecular genetics of human microcephaly. *Current Opinion in Neurology* 14: 151–156.

Roberts E, Hampshire DJ, Pattison L, et al. (2002) Autosomal recessive primary microcephaly: An analysis of locus heterogeneity and phenotypic variation. *Journal of Medical Genetics* 39: 718–721.

Woods CG, Bond J, and Enard W (2005) Autosomal recessive primary microcephaly (MCPH): A review of clinical, molecular, and evolutionary findings. *American Journal of Human Genetics* 76: 717–728.

Microglia Identification Methods

E Raibon and T Möller, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Brief History of Microglial Identification

Utilizing the silver staining method developed by Golgi, Ramón y Cajal first described an uncharacterized 'third element' within the central nervous system in 1913. In his definition, this term applied to all cells morphologically different from either neurons (first element) or astrocytes (second element). As we know today, this term actually applies to two distinct cell groups, oligodendrocytes and microglia. Pio del Rio Hortega formally described 'microglia' as their own unique entity in 1921 and provided the first systematic investigation of microglial cells in 1932. He described microglia as cells that "possess liberal ramified expansions and display migratory and phagocytic activity." Both Ramón y Cajal and del Rio Hortega used silver staining and relied on cell morphology for their assessments. Their method reveals resting microglia as cells with characteristically elongated, almost bipolar, cell bodies with spine-like processes having many short branches. The nucleus of microglial cells is large in comparison with the small perikaryon. While this early morphological description of microglial cells still holds true today, the staining methods of Ramón y Cajal and del Rio Hortega were neither specific nor reproducible. Consequently, for an extended period of time there was little progress in our understanding of microglial cells. This quest was further hampered by the fact that microglia show a remarkable morphological plasticity, which ranges from highly ramified cells in normal central nervous system (CNS) tissue, over rod-shaped cells in inflamed cerebral cortex, to full-blown phagocytes in tissue with necrotic or apoptotic cells. This enormous hurdle was finally overcome by the development of 'specific' stains, including lectins and monoclonal antibodies against epitopes expressed by microglia. This rekindled interest in microglial cells in the 1980s and set the stage for our current understanding of microglial cells in health and disease. However, one major caveat, inherent to all currently available microglial markers, should be pointed out early on. Virtually all available markers to identify microglia also detect other myeloid cells, such as macrophages and monocytes. The so-called microglia-specific labeling, as used here and in virtually all literature, actually only indicates that these labels do not cross-react with other CNS resident cells, such as neurons, astrocytes, or oligodendrocytes.

Lectins

Lectins are carbohydrate-binding proteins ubiquitously found in species from plants to humans. They play an important role in cell adhesion and control of protein levels in the blood. Their ability to recognize specific terminal carbohydrates on glycoproteins or glycolipids makes them valuable markers. The α-D-galactosyl-specific isolectin B4 (IB4) of *Griffonia simplicifolia*, an African medicinal plant, is a lectin widely used to label microglial cells in the rat, guinea pig, sheep, salamander, and even leech. *Ricinus communis* (castor bean) agglutinin-1 (RCA-1), a β-D-galactosyl-preferring lectin, is frequently used to stain microglial cells in normal and pathological human tissue. The lectin from *Lycopersicon esculentum* (tomato) has a high affinity to poly-*N*-acetyllactosamine and is a good marker for mouse microglia. It has been used for fixed tissue but also for labeling of living microglia in brain slices. One common caveat for all lectins is their cross-reactivity with endothelial cells. While blood vessels are rarely mistaken for microglial cells, abundant staining of endothelial cells can hamper the identification of the smaller microglia. However, because of their ease of use, lectin stainings are frequently used in routine histology.

Antibodies

The advent of monoclonal antibodies directed against specific epitopes paved the way to improved and more specific ways of microglial cells staining. Due to the myeloid lineage of microglial cells, most of the markers frequently used and presented here are also found in macrophages and monocytes. These markers are linked to common tasks of myeloid cells, such as phagocytosis or other immune functions. Even though most of the antibodies were developed against peripheral myeloid cells, protocols were successfully adapted to stain microglia in the CNS. Antibodies can be used in a variety of methods in immunocytochemistry and immunohistochemistry, flow cytometry, immunopanning, and fluorescent and magnetic cell sorting. Many antibodies can be used on fixed and/or frozen tissue and on paraffin-embedded sections (an antigen retrieval step is sometimes required for the most sensitive antibodies). Furthermore, primary antibodies from different species or different primary antibody isotypes allow for multiple labeling in the same sample.

Among the most used antibodies to detect microglia are CD11b, CD68, and F4/80 antibodies, which recognize, respectively, antigens from the integrin family, a lysosomal glycoprotein, and the F4/80 antigens. CD11b is part of the complement receptor 3

and is also known, under the name of the antibody clones, as Mac-1 (antimouse) and OX-42 (antirat). CD68, also known as macrosialin, is a lysosomal protein also found on the cell surface. Interestingly, the antibody ED1, used for a long time to stain rat microglia, was found to recognize rat CD68. F4/80 antibody is mainly used in mouse tissue and the F4/80 protein was recently found to be involved in immune functions within the eye, which constitutes a CNS-like functional entity. Antibodies against all three epitopes have been extensively used and stain both resting and activated microglia, along with circulating monocytes and macrophages. Interestingly, differentiation between monocyte-derived macrophages invading the CNS and microglial cells was made possible by using anti-CD45 antibodies. CD45, also known as the leukocyte common antigen (LCA), is a member of the receptor protein tyrosine phosphatase family. Although expressed widely by microglia, monocytes, macrophages, and other hematopoietic cells, CD45 expression levels differ between microglia and macrophages. Microglia have a low level of expression, whereas circulating monocytes and peripheral tissue macrophages express high levels of CD45. This feature has been successfully exploited for the identification of microglial cells via *ex vivo* flow cytometry.

ED2 and ED3 antibodies, raised against macrophage-specific lysosomal markers with still unknown function, stain activated microglia and macrophages in rats. However, the ED2 antibody does not label resting parenchymal microglia, and ED3 staining of resting microglia is variable. Other antibodies recognize epitopes linked to immune functions of microglial cells and are upregulated when microglia become activated. For example, major histocompatibility complex class (MHC) II antigen is virtually not present on resting microglia, but is strongly upregulated in activated cells. However, MHC II upregulation varies depending on the type of CNS injury. MHC II is highly expressed during bacterial infection whereas its level is usually low in neurodegenerative diseases. Thus antibodies against the MHC II antigen, such as OX-6 (antirat), reveal activated microglial cells. However, they can also label other MHC II-expressing cells, such dendritic cells, which have been shown to be involved in the CNS immune response.

A recent addition to the microglial marker repertoire is ionized calcium-binding protein adaptor molecule 1(Iba1). Iba1 is a calcium-elongation factor (EF) hand protein, and is specific to the monocytic lineage, including microglia and macrophages. Iba1 protein function is still not completely understood; however, the Iba1 gene is highly conserved throughout evolution, suggesting an important role for this protein. Recent molecular studies pointed out a possible role in membrane ruffling and phagocytosis. The ubiquitous intracellular localization of Iba1 provides good staining, delineating whole microglial cells when used in immunohisto- and immunocytochemistry, as shown in Figure 1.

It is worth noting that there is yet no available single marker to reliably distinguish resting from activated microglia. The distinction between the two still mainly relies on morphological criteria. Another way to distinguish resting from activated microglia is to detect the expression of activation-associated cytokines. However, cytokine secretion is not restricted to activated microglia. For example, astrocytes can secrete some of the same cytokines secreted by microglia. In order to identify the source of cytokine secretion, it is necessary to combine the intracellular cytokine

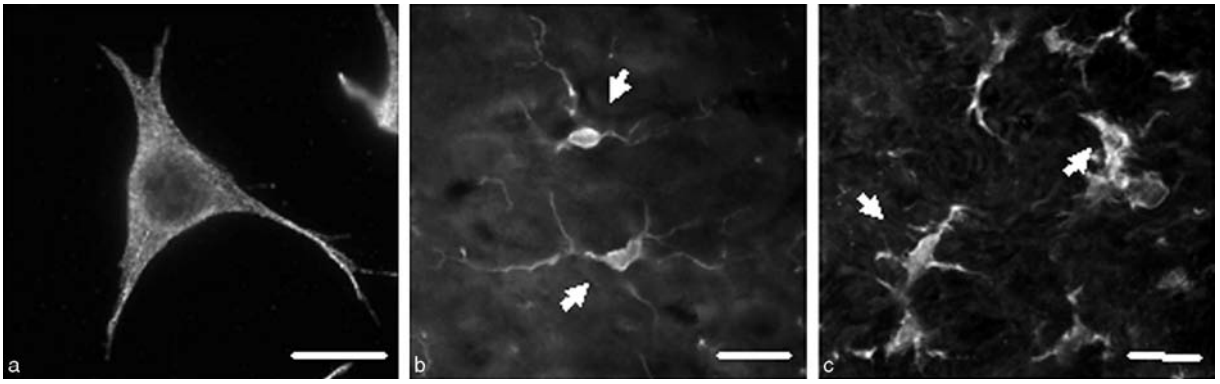

Figure 1 Fluorescent microscopy images of microglia stained with Iba1 antibody. (a) In the microglial N9 cell line, Iba1 antibody stains cytoplasm and processes, and the nucleus is clearly distinguished. (b) Iba1 staining of microglial cells (arrows) in the healthy adult mouse brain. Cells present a ramified morphology, with a small cell body and highly branched processes. (c) In the spinal cord of amyotrophic lateral sclerosis model mice with end-stage disease, Iba1 staining reveals activated microglial cells (arrows), with larger cell bodies and shortened processes. Scale bars = 5 μm (a), 10 μm (b, c).

staining with a cell-selective (e.g., microglia) marker. Double labeling is a useful approach that can help to distinguish resting and activated microglia.

Many antibodies were developed and used before their corresponding antigen was actually known. Consequently, when used as cell markers, they are often referred to by the name of the antibody clone rather than the name of the antigen. This has inevitably led to some confusion. For example, the CD11b antigen, which together with CD18 forms the CR3 receptor, is also known as Mac-1 (mouse) and OX-42 (rat), the names of the respective antibody clones. The Table 1 gives an overview of the different categories of markers currently in use, with a focus on a classification by antigen names instead of clone names.

Genetic Labels

Specific labeling of microglia (and all other cells) depends on the restricted expression of certain molecules, which are inherent only to particular cells. For example, the expression of Iba1 in microglial cells, but not in neurons, macroglia, or endothelial cells, is what makes the antibody directed against it such a valuable marker. If the genetics of a cell-type-specific molecule are known and a promoter has been identified, this information can be used to devise transgenic animals which express a readily identifiable molecule – for example, green fluorescent protein (GFP) – under the control of the promoter. Such approaches have been used to make GFP-labeled neurons, astrocytes and oligodendrocytes. However, transgene expression in microglia and closely related macrophages has proved to be more challenging than expression in neuroectodermal cells. Early experiments using CD11b or CD14 promoters were disappointing. The first report of GFP-expressing microglia was a by-product of engineering a *CX3CR1* knockin/knockout mouse. The investigators sought to interrupt the functional expression of *CX3CR1* by inserting GFP into this gene. While originally constructed to investigate the role of *CX3CR1*, the engineered mice have been used by the microglial community to investigate microglial behavior *in vivo*. However, results from these animals have to be interpreted carefully. Even though these studies used heterozygous animals ($CX3CR1^{+/-}$), the disruption of one genomic copy of *CX3CR1* might reduce *CX3CR1* expression, and thereby influence microglial motility. Other animals with GFP-expressing microglia have been reported since, using

Table 1 Commonly used microglial markers

Marker	*Name*	*Other names*	*Antigen (function)*	*Staining pattern*
Lectins	*Griffonia simplicifolia* isolectin B4	IGSA-IB4 IB4	α-D-Galactosyl group (cell adhesion and control of immune function)	Resting and activated microglia (rat) Vascular endothelium Patchy staining
	Ricinus communis agglutinin-1 lectin	RCA-1	β-D-Galactosyl group (cell adhesion and control of immune function)	Resting and activated microglia (human) Vascular endothelium Patchy staining
	Lycopersicon esculentum lectin	Tomato lectin TL	Poly-*N*-acetyllactosamine group (cell adhesion and control of immune function)	Resting and activated microglia (mouse) Vascular endothelium Patchy staining
Antibodies	CD11b	Mac-1 OX-42 (rat)	Complement receptor CR3; integrin family (adhesion)	Resting and activated microglia Macrophages
	CD14	rmC5-3; Sa2-8 (mouse)	Coreceptor for lipopolysaccharide (endotoxin immune response)	Activated microglia Monocytes
	CD45	LCA OX-1 (rat)	Receptor protein tyrosine phosphatase family (T cell receptor/B cell receptor-mediated activation)	Resting and activated microglia (low) Monocytes/macrophages (high)
	CD68	ED1 (rat)	Lysosomal glycoprotein (phagocytosis)	Resting and activated microglia Macrophages
	ED2 ED3	ED2, ED3 (rat)	Macrophage-specific lysosomal marker (unknown function)	Resting and activated microglia; not parenchymal microglia Macrophages
	F4/80	F4/80; BM8 (mouse)	Glycoprotein (role in immune regulation)	Resting and activated microglia Macrophages
	MHC II	I-A; I-E (mouse) RT1B; OX-6 (rat)	MHC class II (antigen presentation)	Activated microglia
	Iba1	Ionized calcium-binding adaptor molecule 1	Calcium-EF hand protein (membrane ruffling; phagocytosis)	Resting and activated microglia Monocytes/macrophages

Iba1, CD68, and F4/80 as microglia-specific promoters. Again, one has to keep in mind that the term 'microglia-specific' only refers to 'not expressed in other CNS resident cells.' All these animals express GFP in other myeloid cells (e.g., macrophages and monocytes).

A variation in these transgenic approaches to drive expression of a marker gene is the use of microglia-specific promoters such as F4/80 or CD11b in a cyclization recombinase/locus of 'X-over' P (Cre/Lox) system. Expression of Cre recombinase in microglial cells allows the microglia-specific knockdown of a gene flanked by *loxP* sites.

PK11195

The isoquinoline PK11195 is a ligand for the peripheral benzodiazepine receptor expressed in myeloid cells. In the CNS it selectively labels microglial cells, and PK11195 binding strongly increases in activated microglia. Based on these observations, radiolabeled PK11195 was successfully used in positron emission tomography (PET) studies to investigate microglial cells in human disease. For example, in Alzheimer's disease (AD), hot spots with PK11195-labeled activated microglia can be readily identified, providing a prognostic tool for disease progression. In patients with multiple sclerosis (MS), activated microglial cells are found in active lesions identified in gadolinium-enhanced magnetic resonance images (MRIs). Similar results have been obtained for many other neurological diseases and have kindled the interest of neuroradiologists.

Other Labels

Before specific antibodies against microglial epitopes became widely available, enzymes such as nucleoside diphosphatase or thiamine pyrophosphatase were used to identify microglial cells in tissue sections. Ecto-5′-nucleotidase (CD73), another enzymatic marker used, is found on microglia cell membranes, and its expression is strongly upregulated in activated cells; however, it is not cell-type specific. The same is true for NADPH diaphorase, which has frequently been used to detect oxidative burst in activated microglial cells. Furthermore, due to the necessity of maintaining enzyme activity, such stainings cannot be used on paraffin-embedded tissue.

A less frequently used method to identify microglial cells exploits the increased pinocytosis of myeloid cells. Microglial cells incubated with a dye solution such as lucifer yellow will take up this dye into pinocytic vesicles, and therefore the cells become labeled over time. Another indirect labeling technique takes advantage of the phagocytic capability of fully activated microglia. Fluorescent carbocyanine dye (Di-I) injected into axons retrogradely labels neuronal cell bodies. When neuronal cell death is induced, Di-I is found in cells with microglial morphology. Although interesting for studies focusing on phagocytic functions of microglial cells, this type of indirect labeling will only mark cells which have differentiated into phagocytes.

See also: Glial Cells: Microglia During Normal Brain Aging; Inflammation in Neurodegenerative Disease and Injury; Microglia Properties; Microglial Response to Injury; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines.

Further Reading

Banati R (2002) Visualising microglial activation *in vitro*. *Glia* 40: 206–217.

del Rio-Hortega P (1932) Microglia. In: Penfield W (ed.) *Cytology and Cellular Pathology of the Nervous System*, pp. 482–534. New York: Hoeber.

del Rio-Hortega P (1939) The microglia. *Lancet* 1: 103–106.

Hirasawa T, Ohsawa K, Imai Y, et al. (2005) Visualization of microglia in living tissues using Iba1-EGFP transgenic mice. *Journal of Neuroscience Research* 81(3): 357–362.

Imai Y, Ibata I, Ito D, et al. (1996) A novel gene iba1 in the major histocompatibility complex class III region encoding an EF hand protein expressed in a monocytic lineage. *Biochemistry and Biophysics Research Communications* 224(3): 855–862.

Kettenmann H and Ransom BR (eds.) (2004) *Neuroglia*, 2nd edn. Oxford: Oxford University Press.

Ramón y Cajal S (1911, 1995) *Histology of the Nervous System of Man and Vertebraters (History of Neuroscience, No. 7)*, Swanson N and Swanson L (trans.). New York: Oxford University Press.

Streit W (1990) An improved staining method for rat microglial cells using the lectin from *Griffonia simplicifolia* (GSA I-B4). *Journal of Histochemistry and Cytochemistry* 38(11): 1683–1696.

Microglia Properties

U-K Hanisch and D van Rossum, University of Göttingen, Gottingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Microglia as a Latent Central Nervous System Macrophage

Microglial cells represent the central nervous system (CNS) equivalent of a tissue macrophage. Major roles relate to the surveillance, support, protection, and, as much as possible, restoration of structural and functional tissue integrity. Microglial cells can rapidly respond to homoeostatic disturbances and potential threats by a transformation process (known as microglial activation) to mount a broad variety of reactive measures. Microglia can share or acquire many features known from macrophages of non-CNS origin, including the constitutive and inducible expression of receptors and surface structures for effector functions, cellular communication, and matrix interactions; the production of a plethora of releasable factors with signal or enzymatic activity; as well as abilities for chemotactic movement, phagocytosis, or antigen presentation. As immunocompetent cells, activated microglia can execute innate defense mechanisms and recruit the assistance of adaptive immunity. Yet resident microglia may differ from other macrophages by regulatory aspects and inducible reactive profiles because they are influenced by the specific biochemical and cellular milieu of the CNS in general and they are adapted to the structural vulnerability of neuronal networks. The anatomical restraints on tissue swelling are a feature that, besides the brain's extraordinary structural complexity and poor capacity for reconstruction, further compromises the ability of the CNS to fight defensive battles. Accordingly, there is a tight control on inflammatory and immune defense mechanisms in neural tissues to which microglia must obey.

Distribution and Morphological Features in CNS Tissues

The term microglia, introduced in the early twentieth century by del Rio-Hortega, refers to the parenchymal cell population, which is distinct from other macrophage-like cells with perivascular or meningeal localization. Estimated to account for up to 20% of glia, microglial cells are dispersed throughout the brain and spinal cord and also found in the retina. Although regional tissue densities vary, cells normally occupy and control discrete territories. The characteristic, macrophage-atypical appearance in the mature and healthy CNS is that of a small cell body with highly branched processes (Figure 1). This ramified morphology of the resting microglia is commonly revealed by specific immuno- or lectin histochemical staining methods in fixed tissue specimens. However, the dynamics of the highly motile processes have only recently been visualized by experimental *in vivo* imaging techniques. The scanning movement of the otherwise territorial microglia thereby seems to assist the environmental monitoring function. Upon activation, microglia can rapidly reorganize the processes. Cells reveal a granular appearance and turn into a more rounded, bushy or rod-like and ameboid shape. Local cell densities can increase at sites of an injury by chemotactic movement and congregation as well as proliferative expansion, building a pool of reactive cells which is assisted also by bone marrow-derived, invading monocytes. The postactivation principles that would reestablish normal microglia densities are less well understood. Apoptosis appears to be a mechanism to reduce local numbers. On the other hand, areas that donated cells to another location may be repopulated by migration or mitosis in order to guarantee sufficient microglial surveillance and protection capacity. The dramatic gradual transition in microglial morphology has been a sensitive and reliable hallmark to indicate neuropathological changes of most diverse etiology and to reveal primary or secondary engagement of these cells. Morphological appearance, however, is a poor surrogate of functional properties and may not reflect the actual functional consequences of activated microglia. The fact that isolated microglia *in vitro* present with a heterogeneous but essentially not ramified morphology (Figure 2) has been taken as sign of partial activation. Nevertheless, even though cells in culture are at a higher alert level due to the lack of certain tissue-related cues, many executive functions rapidly and tightly associated with reactive microglia, such as the synthesis and release of proteins (e.g., cytokines), are not yet effective. On the contrary, ramified appearance can be enforced without guaranteeing a fully resting, *in vivo*-like state. Moreover, due to a more complex response potential and a previously underestimated functional diversity of microglia, neither morphological signs of activation nor detection of a single marker molecule can provide reliable insight into the actual performance. As discussed later, the reactive phenotype of such a cell is defined by a molecular expression pattern and a related ensemble of cellular properties. Currently, there is also no

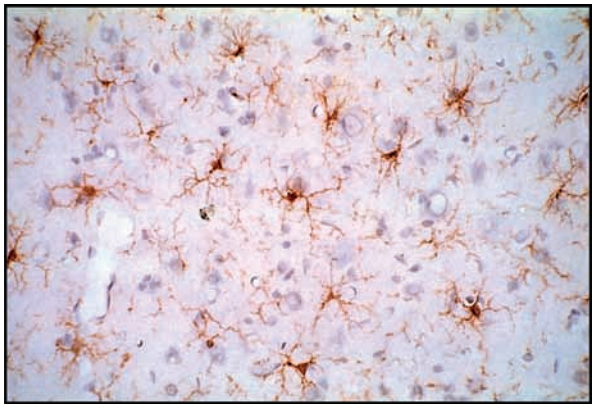

Figure 1 Microglial cells in the healthy adult CNS reveal a ramified morphology. The image represents an immunostaining for microglia-expressed Iba1 protein in adult rat cerebral cortex. Kindly provided by Dr. Shinichi Kohsaka, National Institute of Neuroscience, Tokyo, Japan.

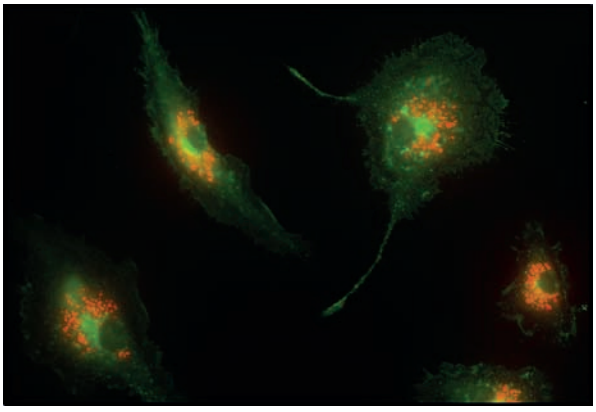

Figure 2 Microglia in culture display heterogeneous cell shapes different from the typical appearance in the CNS. *In vitro* settings are widely used to study molecular and cellular properties as well as mechanisms of activation but lack the tissue complexity. The fluorescence micrograph shows mouse microglia stained by FITC-conjugated *Griffonia simplicifolia* isolectin B4 after phagocytotic incorporation of rhodamine-labeled myelin.

established marker available that could easily distinguish microglia from other macrophages by immunostaining or *in situ* hybridization, even though expression levels of certain surface molecules allow separation by quantitative analysis. Constitutive and inducible diversity among and exchange between cells of functional microglia/macrophage orientation may even render it unlikely that a single molecule could fulfill the criteria as a simple distinguishing marker. On the other hand, gene and protein arrays deliver increasingly more molecular patterns to identify individual subtypes within regional and situation-associated cell populations.

Developmental Origin and Population Replenishment

The predominant notion of origin views microglial cells as of monocyte/macrophage lineage and thus as descendents of mesoderm. Precursors of microglia colonize the developing CNS (probably in waves) during embryonic and early postnatal periods from mesodermal tissues and via routes of invasion involving the bloodstream or the ventricles and meninges. The early ameboid cells subsequently turn into resident microglia which undergo relatively little turnover by proliferation and exchange. Although the first precursors arrive within the CNS even before vascularization, early ontogenetic settlement and some further replenishment of parenchymal microglia seem to be based on circulating monocytes, probably also through the pool of perivascular cells. Conversion of blood-borne cells into microglia proper may occur to some extent under normal conditions, whereas pathological changes are associated with enhanced monocyte infiltration. Despite the use of modern techniques such as bone marrow chimeras, the source, heterogeneity, and plasticity of brain macrophages/microglia remain subjects of controversy.

Functions in Healthy and Diseased CNSs

As an ameboid phagocyte in the developing CNS, microglial cells participate in the shaping of neuronal circuitry by removal of aberrant cells and synaptic connections. Additional contributions to a maturation of neuronal properties are likely. However, there is little information regarding microglial participation and partnership in neuronal functioning and plasticity later on, as is documented for astrocytes. On the other hand, significant effects of surrounding microglia on spinal cord neuron signaling in chronic pain suggest that these cells have the communicative assets to exert marked influence on neuronal information processing. Major known functions of the resting microglia relate to the surveillance of tissue homeostasis – an active process that involves the constant monitoring of the ionic, molecular, and cellular milieu. For this purpose, microglial cells express a large number of cognate surface molecules to sense the unusual occurrence, changed concentration, or altered structure of biochemically most different compounds, ranging from nucleotides, (co)transmitters and small peptides, and large proteins and aggregates (Table 1). Regardless of their chemical nature, microglia-alerting factors may carry either on- or off-signal character, depending on whether their appearance or disappearance is the critical trigger. The stimulation of receptors for certain ligands with external and endogenous

origin can then lead to microglial responses that aim to protect neuronal and glial cell function and integrity. Microglia can, for example, synthesize trophic factors as well as unfold a repertoire of defense actions. Molecular cues associated with pathogens, such as bacterial or viral DNA/RNA and cell wall components, are sensed through pattern recognition molecules and alert microglia to fight off a CNS-threatening infection. Breakdown of the extracellular matrix (ECM) or factors released from endangered, degenerating, or damaged neurons, such as ATP and other purines or abnormally processed membrane lipids and proteins, could indicate neuronal impairment as a consequence of acute neurotoxicity, mechanical injury, stroke, or degenerative processes. Serum components that usually do not have access to parenchymal microglia can similarly inform about a potentially dangerous situation, such as due to trauma or disruption of the blood–brain barrier, which normally guarantees controlled exchange of molecules between the central compartment and the circulation. On the other hand, deviation from a normally constitutive signaling of certain ligand–receptor systems can lift a calming influence on microglia and thus initiate an autonomous response program. Known examples of such ligand–receptor systems reveal reciprocal expression by neurons and microglia. Activated microglia can migrate along chemotactic gradients (e.g., created by chemokines or complement) to concentrate at a site of need as well as release complex cocktails of chemoattractive molecules for highly organized recruitment and guidance of further microglia as well as immune cells (granulocytes, monocytes, and lymphocytes). Microglial phagocytosis is important for the removal of apoptotic cells, clearance of debris resulting from tissue damage, as well as the elimination of opsonized or immune-complexed material. Accordingly, microglia can express a repertoire of receptors for complement, immunoglobulins, and cell-associated structures (Table 2). Expressing the required surface structures (i.e., major histocompatibility complex and accessory molecules), microglia can also act as antigen presenting cells, a feature shared with both macrophages and dendritic cells. Effective production of an enormous variety of neuromodulatory, immunoregulatory, ECM modeling, and even cytotoxic factors further serves the interaction with and support of resident neural and infiltrating immune cells, the neutralization of harmful agents, as well as attempts at repair. Together, the molecular equipment and cellular mechanism organizing the sensory and

Table 1 Examples of factors initiating and influencing microglial activation and responses

Components of infectious agents	Bacterial cell wall proteoglycans, bacterial DNA, gp41 (viral coat protein), gp120, lipopolysaccharide (LPS), prion protein (PrP), Tat (viral nuclear protein), teichoic and lipoteichoic acid (LTA)
CNS disease-related (poly) peptides	β-Amyloid aggregates, Aβ25–35, Aβ40/Aβ42
Cytokines	GM-CSF, IL-1β, IL-2, IL-4, IL-6, IL-10, IL-12, IL-15, IL-18, IFN-γ, M-CSF, TGF-β, TNF-α
Chemokines	Ligands for CCR3, CCR5, CXCR, CXCR, CXCR4, CX3CR1, IL-8R
Immunoglobulins and complement	IgG, IgM, C1q, C5a
Neurotransmitters and neurotransmission-related compounds	β-Adrenergic agonists, ATP and related purines, glutamate, kainite, NMDA, K^+
Neurotrophins	BDNF, GDNF, NGF, NT-3, NT-4
Other proteins and compounds	Albumin, ApoE, CD40L, MSH, endothelin, S100B, thrombin, VIP, cannabinoids, ceramide, gangliosides, lysophosphatidic acid, melatonine, opioids/endomorphines, PAF, PGE_2, steroids/steroid hormones, vitamin D_3

Table 2 Examples of molecules expressed and released by (activated) microglia

Receptors and cognate structures	B7.1 (CD80), B7.2 (CD86), C3bi receptor (CR3/CD11b/Mac1), CD14/TLR4 (LPS receptor/signaling complex), CD68, CD16, CD32, CD64 (FcR I, II, and III), CD54 (ICAM-1), CD45 (LCA), MHC II, several cyto- and chemokine receptors, TLR2
Ion channels	Kv1.3 outward K^+ channel
Cytokines and chemokines	CCL2 (MCP-1), CCL3, CCL4 (MIP-1α/1β), CCL5 (RANTES), CCL22 (MDC), CXCL1 (GROα/KC), CXCL2 (MIP-2), CXCL8 (IL-8), CXCL10 (IP-10), IL-1α/-1β, IL-1ra, IL-3, IL-6, IL-10, IL-12 ($p40_2$, p70), IL-15, IL-18, M-CSF, TGF-β, TNF-α
Intracellular proteins	Granule cell death-10 gene (*gcd-10*) product, Iba-1 Ca^{2+}-binding protein, microglia response factor (MRF-1)
Neurotransmitters, neuromodulators, toxins, and proteases	Glutamate, NO, reactive oxygen species (ROS), MMPs

effector functions of the resting and the activated states enable microglia to act primarily protective. In contrast, under certain conditions microglia may actively induce or facilitate neuronal death instead of trying to prevent it, especially when irreversible impairment is associated with improper function (e.g., incorrect and thus rather disturbing signaling). Furthermore, the cytotoxic weaponry and arsenal of neuro- as well as immunoregulatory messenger compounds and effector molecules can turn against the CNS when the microglial activation process is out of control. Then, microglia fuel or carry autodestructive cascades that employ a range of inflammatory mediators and can also involve recruited activities of immune cells.

Microglial Activation and Phenotypic Diversity

In order to activate the array of macrophage-like features and functions, microglial cells have to become activated (Figure 3). Traditionally, the transformation process has been considered as a stereo yped reaction. According to this view, cells would respond to different stimuli with a rather uniform profile of inducible functions. The reflex-like image was influenced by the typical transitions in cell shape as routinely found in histopathological analyses. It was also largely based on, and apparently confirmed by, experimental challenges of isolated microglia with single activating agents, such as bacterial lipopolysaccharide. In such restricted settings, complex integration of multiple signals cannot take place and the resulting reaction reflects an emergency program of defense. This view leaves little room for adjustments by the situational context or feedback regulation within a tissue. It also cannot provide a satisfying explanation of the dichotomy of protective versus destructive actions of microglia.

However, the concept is changing. Activation of microglia is probably an adaptive process with the cells committing to different phenotypes depending on the signaling context as given by more than a single factor and as subject to influences by other resident CNS or infiltrating immune cells. Much of the evidence for such diversity in the response behavior is provided by studies on extraneural macrophages. Treatment with certain cytokines, defined agonists of pattern recognition receptors or their combinations, confrontations with complex microbial agents, and exposure to tumor material or apoptotic or necrotic cells induce macrophages to acquire distinct reactive phenotypes. Reactive phenotypes are not simply characterized by a single release product or receptor expression but, rather, comprise coordinated sets of induced or selectively not induced gene products, adjustments in metabolic and signaling pathways, as well as defined functional (inter)actions. They reveal selective, partially overlapping, as well as partially reciprocal production of mediators for cellular communication and profiles of effector molecules (e.g., cytokines, metabolic or extracellular matrix modeling enzymes, and surface structures for phagocytosis or antigen presentation). These adaptations have physiological consequences, such as opposite effects on inflammation, skewing for types of immune responses, and orientation toward defense

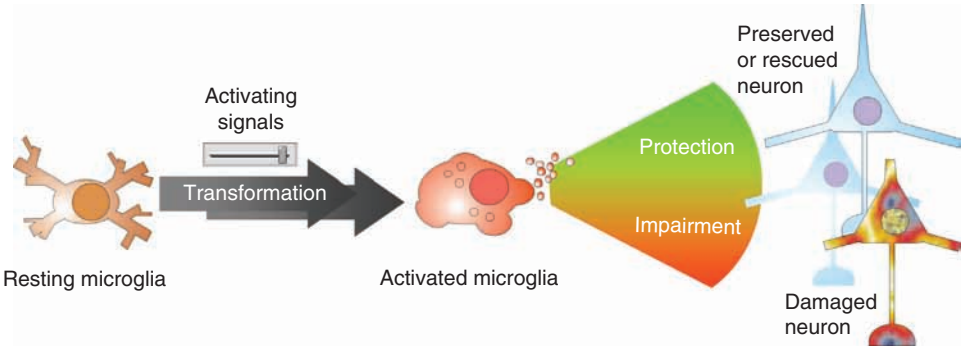

Figure 3 The process of microglial activation causes the normally resting cells in healthy CNS tissue to enter alert and reactive states which are associated with characteristic changes in morphology, increased motility, directed migration and proliferation, as well as a range of macrophage-like functions. The activation is rapidly triggered by signals derived from exogenous or endogenous sources, including structural cues of infectious agents and factors released from damaged cells and tissues. Microglia can then synthesize a huge variety of neuro- and immunoregulatory mediators, toxic compounds, and effector molecules; express surface receptors for cellular and matrix interaction; clear material from the extracellular space by phagocytosis; and present antigen to T cells. As a consequence, microglia may exert mostly protective effects, but they also contribute to cell and tissue damage in certain neuropathological situations and processes. The traditional concept, however, considered microglia activation as a rather stereotyped response, offering little explanation for the variable outcome and impact. Moreover, based on *in vivo* imaging techniques, the resting microglial cells in normal tissue appear to be very motile for actively scanning their environment. The term 'resting,' implying 'nonactive,' may thus not properly apply to describe this functional state. Consequently, the term 'microglia activation' may also require reconsideration.

or repair. Conceivably, microglia can similarly choose from variable response options to adequately cope with an exogenous or endogenous threat or to organize the recruitment of adaptive immune support. Moreover, distinct reactive phenotypes may not only occur as alternative responses to different initial activation scenarios but also interchange in a stage-specific manner during an activation process (Figure 4). Highly organized, context-determined, situation-adapted switches in anti/inflammatory messenger production, timed recruitment versus decline of immune infiltrates, phagocytotic clearance, and tissue rebuilding attempts could be the elements in versatile, (usually) self-limiting, and (mostly) efficient programs for protection and (partial) restoration. Interregional and intrapopulational heterogeneity in constitutively expressed features could also determine the response potential of microglia. In this regard, discrete expression of, for example, neurotrophic factors indicates that microglia are not uniform in the beginning, and that specialized functions could be carried by individual cells. Microglial activation could thus prove to be a much more variable process than originally assumed. Although certain functional features may be part of an initial core program in emergency situations, some are downregulated or become effective when suitable.

Neuropathological Consequences of Microglial Dysfunction

Microglial activation is observed in the neurodegenerative processes of Alzheimer's disease (AD), Parkinson's disease, and amyotrophic lateral sclerosis; in the autoimmune-driven demyelination of multiple sclerosis (MS); in infectious CNS complications associated with, for example, AIDS, meningitis, cerebral malaria, or Creutzfeldt–Jakob disease; as well as with destructive cascades following stroke and trauma. Evidence from morphological examination and staining procedures often reveal locotemporal correlation with the occurrence and development of tissue lesions. It has been interpreted as contributing to an insult or disease process. Combined with the fact that experimental challenges of microglia can cause profound neuronal cell damage

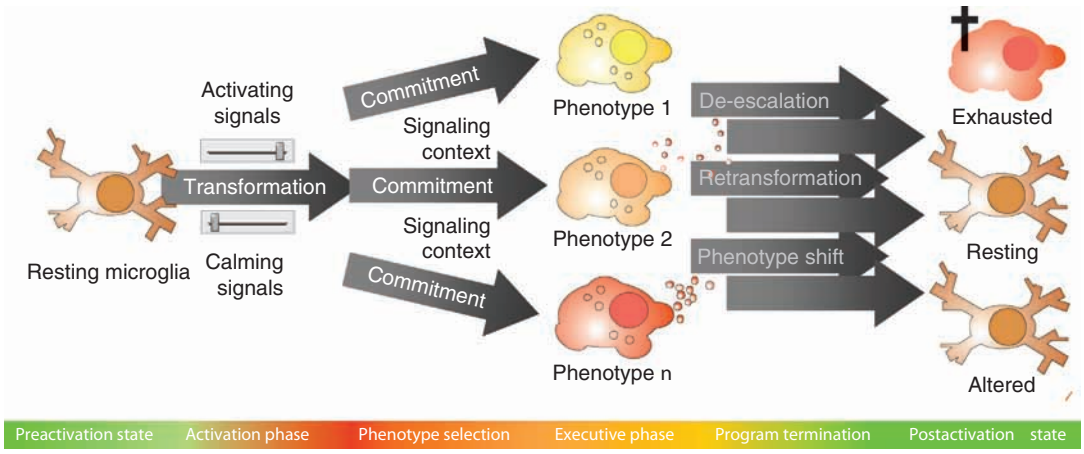

Figure 4 Based on findings and conceptual development, microglial activation is viewed as a more adaptive process. Although some factors would have to appear or to increase in a signaling-relevant form or concentration in the vicinity of microglia in order to trigger a transformation to alerted, activated, or full-blown reactive states, other factors could induce microglial activation by a kind of 'off-signaling.' They usually maintain the resting status and probably keep an attenuating influence. Interruption of their calming input would then result in activation by unchaining an autonomous program for installing increased readiness to provide support or to prepare for an attack. The functional heterogeneity among microglial populations and the diversity in the executive profiles are still poorly understood. Granting microglial cells the ability to integrate the large number of molecular signals for which they can express cognate receptors, the prejudice of a stereotyped response with little adaptive capacity may no longer hold true. Consequently, the term 'activated microglia' would not adequately describe the actual involvement of individual cells or subpopulations in the complex context of an endangered or diseased tissue. Depending on the primary activating signals, as they indicate a threat to the functional and structural integrity of the tissue, microglia may choose an immediate response program which is influenced and adjusted according to further incoming information. Surrounding neurons and glial cells as well as infiltrating immune cells could exert an important control on the initial commitment and development of reactive phenotypes. During the course of activation, affected microglia could shift phenotypes in order to exchange the defense-dominated character of the response for repair orientation. Such changes could, for example, cover the profile of released chemoattractive messengers to support organized infiltration of different immune cell populations or relate to the production of enzymes and substrates for extracellular matrix breakdown versus restoration. Finally, the activation process has to resume. Local microglial distributions and features may return more or less to preactivation conditions. Alternatively, microglia may maintain some altered status, conditioning responses to a subsequent challenge. The concept of context-dependent phenotype diversity and dynamics suggests that not only excessive stimulation (quantitative aspect of a response) but also maladapted reactions (response quality) of microglia contribute to neuropathologies.

and death, activated microglia have earned the reputation of being destructive cells. Yet most of the transient, focally restricted, and successfully resolved episodes of microglia activation may never present with clinical symptoms, and the neuropathological records may be biased for the cases in which microglial attempts to protect, support, and restore were insufficient. Due to a dysregulated engagement, microglia may fail to fight off a threat and to rescue the CNS from further impairment. Collateral (bystander) damage can result when microglial activities exceed critical limits. Whereas in acute settings, such as a bacterial invasion, massive stimulation of microglia may simply result in excessive reaction, such as exaggerated production of potentially neurotoxic molecules (reactive oxygen and nitrogen species, cytokines, and proteases), it may also be the quality and duration of the response which are inappropriate. Granting microglia diversity and adjustment of reactive phenotypes, a misleading signal or signaling context may trigger unnecessarily aggressive cellular behavior. There is reason to assume that endogenously produced factors can be read by microglia (through respective receptor systems) as a signal which otherwise would trigger an antimicrobial and therefore toxic defense action. Detrimental consequences of an active microglia may arise when certain functions are inhibited or when adjustments of phenotypic features are prevented. Microglial activation may often resolve when a particular activating stimulus vanishes. These cells might even organize for its removal, for example, by phagocytosis. Impaired phagocytotic activity or resistance of the target material could then condemn microglia to remain in a certain activation state associated with secretion of proinflammatory and neurotoxic mediators. Vicious cycles of microglia-caused damage resulting in microglia recruitment could thereby cause the progressive neuronal impairment in aggregopathies such as AD. Phagocytosis is certainly also critical for the neutralization of infectious agents. For example, bacterial structures will trigger the release of proinflammatory and toxic substances, which resumes upon their removal from the extracellular space. Phagocytotic clearance of myelin debris in MS is a prerequisite for partial remyelination and probably affects the risk of additional autoimmune challenges. Further research efforts need to address the question of whether functional properties of microglia are strictly or unavoidably coupled. For example, physical contact with molecular or cellular structures for (potentially beneficial) phagocytotic elimination may concomitantly trigger certain (potentially harmful) release activities. During CNS development, phagocytosis is carried out in the absence of inflammatory reactions, whereas the situation may be different under pathological conditions. Finally, a shift from an initial emergency program to repair-promoting phenotypes could eventually be followed by the termination of the activation process. Disturbances along such conversions could hamper de-escalation or result in essential restoration capacities being skipped. The link between the regulated, selective, and sequential execution of functional contributions is likely key to understanding the actual impact of an activated microglia in a disease setting. It is important to note that the surveillance accuracy, support performance, protective capacity, as well as regulatory influences and control mechanisms of and on microglia could fade with aging. This aspect has major implications for the CNS's ability to withstand neurodegenerative processes.

Microglia as a Target of Diagnostic Tools and Therapeutic Intervention

Microglia activation and activities are associated with virtually all challenges of CNS homeostasis. This central role and the repertoire of inducible effector functions make these cells attractive for the development of clinical methods and strategies. Imaging techniques built on radioligands in positron emission tomography aim at visualizing anatomical regions with activated microglia during and even before cell and tissue damage occurs in certain CNS diseases. Silencing of undesired or excessive activities, such as overproduction of potentially harmful substances, may be combined with sparing or enhancing beneficial functions, such as phagocytotic removal of neurotoxic or autoimmunogenic material. Whereas selective molecular manipulation could be designed to attenuate receptor stimulation, alter intracellular signal processing, or intercept with released mediators, general suppression of microglia would deprive the CNS of its supportive functions. Enhanced recruitment of monocytic cells to the resident microglia at sites of lesions may offer access to Trojan horse drug delivery systems.

Approaches to Study Microglia

Models to study microglial behavior outside CNS tissue suffer from several restrictions. In homogenic culture, the isolated cells are deprived of their natural environment. Because soluble and insoluble cues normally surrounding the microglial cells are missing, the cells may not develop the standby mode commonly considered as the 'resting state'. Regarding the principle of off signals, the calming influence normally imposed on microglia by neighboring cells is absent. After the mechanical stress of isolation and culture preparation, cells may therefore remain in a kind of

subalerted mode. Nevertheless, this setting has been very useful for identifying basic mechanisms in the range of activation processes. Despite the respective limitations, *in vitro* approaches still allow for relatively easy and direct experimental manipulations and analyses. Compared to primary cultures, microglia-derived cell lines yield larger cell numbers and require less preparational effort, but they often have deviated features. Mixed and co-cultures offer additional cellular elements and interactions. Yet the intact tissue is more than a mixture of different cell types. Therefore, microglial properties have also been studied in CNS tissue slices. Nevertheless, the surface of a brain slice represents injured tissue. Any acute study thus runs the risk of observing microglia in a response to trauma. Although expression (or lack of expression) of a receptor or ion channel protein does not change within minutes, functional properties may do so. Rapid responses of microglia are thus very likely considering their physiological role as a sensor and effector in pathological situations. Phosphorylation cascades affecting the cytoskeleton, metabolic enzymes, or features of receptor signaling occur faster than the tissue handling prior to measurements. Accommodated slice preparations may instead allow the activated microglia to readjust and to return to a more resting state. Organotypic culture preparations avoid massive tissue disintegration immediately before an investigation and maintain, at least for some time in culture, the cells within the context of their natural environment, including other cell types and their relative orientation. However, organotypic tissue cultures cannot prevent some progressive degeneration of cellular, mainly neuronal, components, which will unavoidably influence microglia. *In vivo* studies guarantee the CNS tissue context and avoid potential artifacts of disintegrative models, but they have not been suitable for directly studying characteristics of a given cell population over a period of time. Developments in *in vivo* imaging of microglia in superficial cortical layers through the thinned skull bone make it possible to visualize the living cells while avoiding massive tissue disruption and leaving the organ untouched. Further technological improvements will provide access to deeper structures as well. Still, the systemic approach has the disadvantage of being very complex. Identifying molecular mechanisms and cellular principles will thus require a combination of *in vitro* and *in vivo* approaches based on wild-type as well as genetically modified systems. Concerning microglia, transgenic mouse models already exist in which the cells are rendered fluorescent. Constitutive and conditional knockout animals have been generated as well. Cell type specificity is achieved for cells of the monocyte/macrophage lineage only, due to the current lack of microglia-restricted gene targets, but is sufficient for many applications within the CNS. Moreover, use of bone marrow chimeras and reciprocal donor–recipient experiments allow the identification of the contributions of resident microglia versus invading monocytes. Approaches aimed at insertion of a reporter or deletion of a functional protein must yet take into account the involuntary impairment of critical molecules operating in the extremely sensitive activational control of microglia.

See also: Glia and Stroke; Glial Responses to Injury; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Growth Factors; Glial Influence on Synaptic Transmission; Glial Responses to Virus Infection; Microglia identification Methods; Microglial Response to Injury.

Further Reading

Barchetti B, Kettenmann H, and Streit W (eds.) (2005) Special issue: Glia–neuron crosstalk in neuroinflammation, neurodegeneration and neuroprotection. *Brain Research Reviews* 48. 129–408.

Cuadros MA and Navascués J (1998) The origin and differentiation of microglial cells during development. *Progress in Neurobiology* 56: 173–189.

Dobrenis K (1998) Microglia in cell culture and in transplantation therapy for central nervous system disease. *Methods in Enzymology* 16: 320–344.

Hanisch UK, Kohsaka S, and Möller T (guest eds.) (2002) Special issue: Microglia. *Glia* 40: 131–269.

Nimmerjahn A, Kirchhoff F, and Helmchen F (2005) Resting microglial cells are highly dynamic surveillants of brain parenchyma *in vivo*. *Science* 308: 1314–1318.

Polazzi E and Contestabile A (2002) Reciprocal interactions between microglia and neurons: From survival to neuropathology. *Reviews in the Neurosciences* 13: 221–242.

Raivich G (2005) Like cops on the beat: The active role of resting microglia. *Trends in Neurosciences* 28: 571–573.

Raivich G and Banati R (2004) Brain microglia and blood-derived macrophages: Molecular profiles and functional roles in multiple sclerosis and animal models of autoimmune demyelinating disease. *Brain Research Reviews* 46: 261–281.

Schwartz M, Butovsky O, Brück W, and Hanisch UK (2006) Microglial phenotype: Is the commitment reversible? *Trends in Neurosciences* 29: 68–74.

Streit WJ (2005) Microglial cells. In: Kettenmann H and Ransom BR (eds.) *Neuroglia*, 2nd edn., pp. 60–71. Oxford: Oxford University Press.

Streit WJ (2006) Microglial senescence: Does the brain's immune system have an expiration date? *Trends in Neurosciences* 29: 506–509.

Wilms H (ed.) (2001) Special issue: Studies on microglia *in vitro*. *Microscopy Research and Technique* 54: 1–122.

Microglial Response to Injury

K Ohsawa and S Kohsaka, National Institute of Neuroscience, Tokyo, Japan

© 2009 Elsevier Ltd. All rights reserved.

Microglia ubiquitously exist in the neural parenchyma and exhibit ramified morphology under the normal brain, and they are called 'resting microglia.' In neuronal injuries, microglia are activated to show a number of features, including morphological changes; cell migration and proliferation at lesion sites; induction of wide-ranging myeloid markers, cytokines, free radicals, and nitric oxide; and acquisition of phagocytic phenotype in extensive neuronal cell death (Figure 1). The activated microglia are recognized during neuronal degeneration and then revert to the resting phenotype in parallel with the neuronal regeneration. Therefore, the activated microglia are thought to participate in the degeneration and regeneration of neuronal cells and to play an important role in the repair of the injured nervous system.

Microglia in Normal Brain

In the normal adult central nervous system (CNS), microglia display ramified morphology, and they have a small body with long branched processes (Figure 2). The microglia show a slow turnover from either proliferation of resident microglia or recruitment of bone marrow-derived cells through the intact blood–brain barrier (BBB). Ramified microglia are thought to be in a resting state, but they appear to participate in the immune surveillance in defense of the CNS. Fine branches of the ramified microglia are highly mobile, providing extensive and continuous surveillance of their cellular environment. The resting microglia constitutively express a calcium-binding protein Iba1 and several surface molecules, including CD11b (Mac-1 in mice and complement receptor type 3 (CR3) in rats), F4/80 (in mice), and lectin-binding sugar molecules. Expression of phagocytic marker ED1 and antigen-presenting molecules major histocompatibility complex (MHC) class I and II is undetectable or very low in the ramified microglia under the normal brain.

Animal Models of Acute Neuronal Injury

Activation of microglia following acute neuronal injury has been studied in animal models of transection of nerve axons, injection of toxins, and ischemia. The neuronal injury by a stab or stroke breaks the BBB, leading to the influx of blood-derived monocytes and macrophages to the CNS parenchyma. It is impossible for the monocytes or macrophages to be distinguished from resident microglia. Transection of nerve axons, such as rodent facial, hypoglossal, or sciatic nerve and an entorhino-dentate perforant path projection, induces retrograde changes in the central nucleus of the axotomized neuron without BBB disruption and activates local microglia around the axotomized neurons. In axotomy of the facial nerve of adult rats, which includes transection of the facial nerve in the stylomastoid foramen distant from the facial nucleus in the brain stem, most of the facial motoneurons survive after nerve axotomy, and the activated microglia accumulate around the regenerating facial nucleus (Figure 3). On the contrary, neonatal axotomy induces massive degeneration and death of facial motoneurons to activate local microglia. The rat facial axotomy has been used as the most helpful model in research of the function of activated microglia in degeneration and regeneration of injured neurons.

Morphological Transformation

In the initial stage of activation ranging from minutes to a few hours following neuronal injury, microglia cause morphological changes with partial retraction and slight hypertrophy of the microglial processes, and they become bushy or rodlike cells. Clustering of the hyperramified microglia is recognized around the injured neuronal cell bodies 1–3 days following injury. In the case of extensive neuronal degeneration, especially in the later phase of the degenerative process, microglia may become large, round phagocytic cells called 'brain macrophages.'

Accumulation around Injured Neurons

One to three days following neuronal injury, activated microglia rapidly move and accumulate around the damaged structure like injured neurons or degenerating neurite terminals. These microglia directly contact with injured neuronal cell bodies, where they displace synaptic input, a phenomenon known as synaptic stripping. In the case of laser-induced focal tissue injury, processes of nearby microglia extend toward the lesion site. Within 30 min after laser ablation, the microglial processes reached the damaged site and appeared to accumulate and fuse together around the damaged area. Microglia establish a potential barrier between healthy and injured tissues, which may play a neuroprotective role for microglia.

Microglia activated by neuronal injury increase expression of the receptors for the microglial mitogens

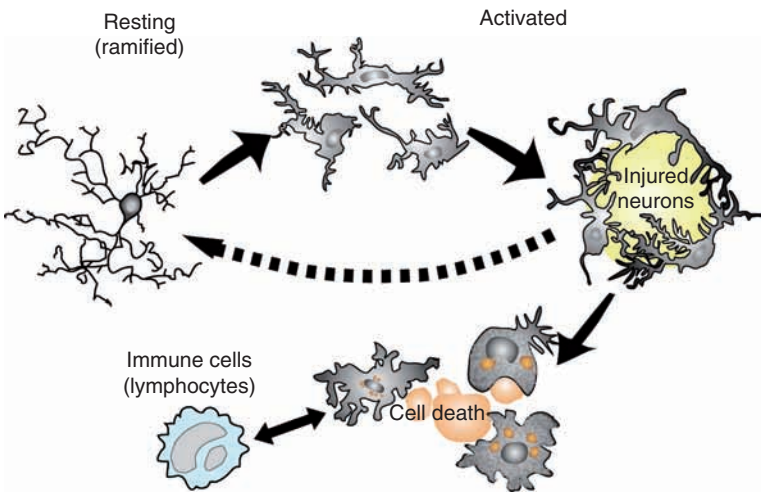

Figure 1 A schematic summary of microglial response to neuronal injury. Neuronal injury induces a rapid transformation of resting (ramified) microglia to the activated form. Microglia change their morphology, migrate toward the injured sites, and proliferate and accumulate around the injured neurons. Activated microglia produce and release various molecules in response to several stimuli from surroundings. Cell death leads to a further transformation of microglia into phagocytic cells. Activated microglia also interact with the immune cells infiltrating through the blood–brain barrier. In parallel with the regeneration of neuronal cells, microglia gradually return to a resting state.

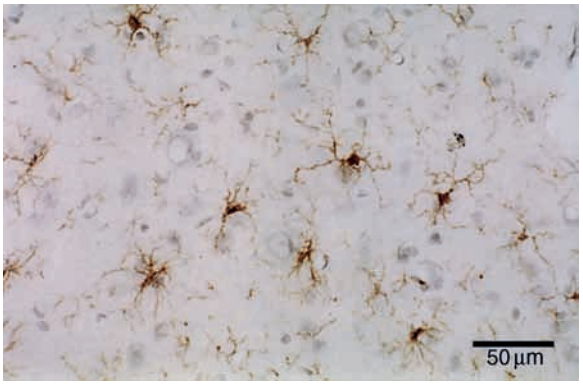

Figure 2 Ramified microglia in the adult rat cerebral cortex. A brain section was immunostained with an anti-Iba1 antibody.

macrophage colony-stimulating factor (M-CSF) and granulocyte macrophage colony-stimulating factor (GM-CSF), and they begin to proliferate. Mitotic activity in these cells reaches the maximum approximately 2–4 days after axotomy or during the anterograde degeneration.

Production of Bioactive Molecules

In parallel to their morphological transformation, microglia are also induced to upregulate constitutively expressed markers such as CD11b, F4/80, and Iba1, and they are stimulated to express MHC class I, CD45, CD4, and cell adhesion molecules such as intracellular adhesion molecule-1 (ICAM-1), very late antigen 4 (VLA-4), and the leukocyte function antigen-1 (LFA-1). The activated microglia also synthesize a number of cytokines, such as transforming growth factor (TGF-β1), interleukin-1β (IL-1β), tumor necrosis factor-α (TNF-α); growth factors, such as M-CSF, GM-CSF, and insulin-like growth factor (IGF-1); and chemokines, such as CCL2/MCP-1, CCL3/MIP-1α, CCL4/MIP-1β, CCL5/RANTES, CXCL10/IP-10, and CXCL1/fractalkine. Expression of the receptors for M-CSF, GM-CSF, and chemokines is upregulated in activated microglia, and they are involved in migration and proliferation of resident microglia. Molecules such as MHC class II and B7, which are involved in antigen presentation, are usually only expressed by subpopulations of the lesion-reactive microglia.

Phagocytosis

Neuronal cell death causes a further transformation of microglia into phagocytotic cells that remove neural debris. In neonatal rat facial nerve axotomy, microglia rapidly transform to phagocytes in parallel with the degeneration process of the axotomized motoneurons. Phagocytotic microglia express high levels of Fcγ receptors, B7, and MHC class I and II, which may serve as potent stimuli of T cell activation in the presence of specific antigen. The phagocytotic microglia can play a role in CNS antigen-presenting cells. On the other hand, microglia gradually decrease in number, lose activation markers, and revert to a resting state with ramified morphology in the absence of neuronal cell death. In adult rat facial nerve axotomy, microglia accumulate at the regenerating facial nucleus without turning into phagocytes.

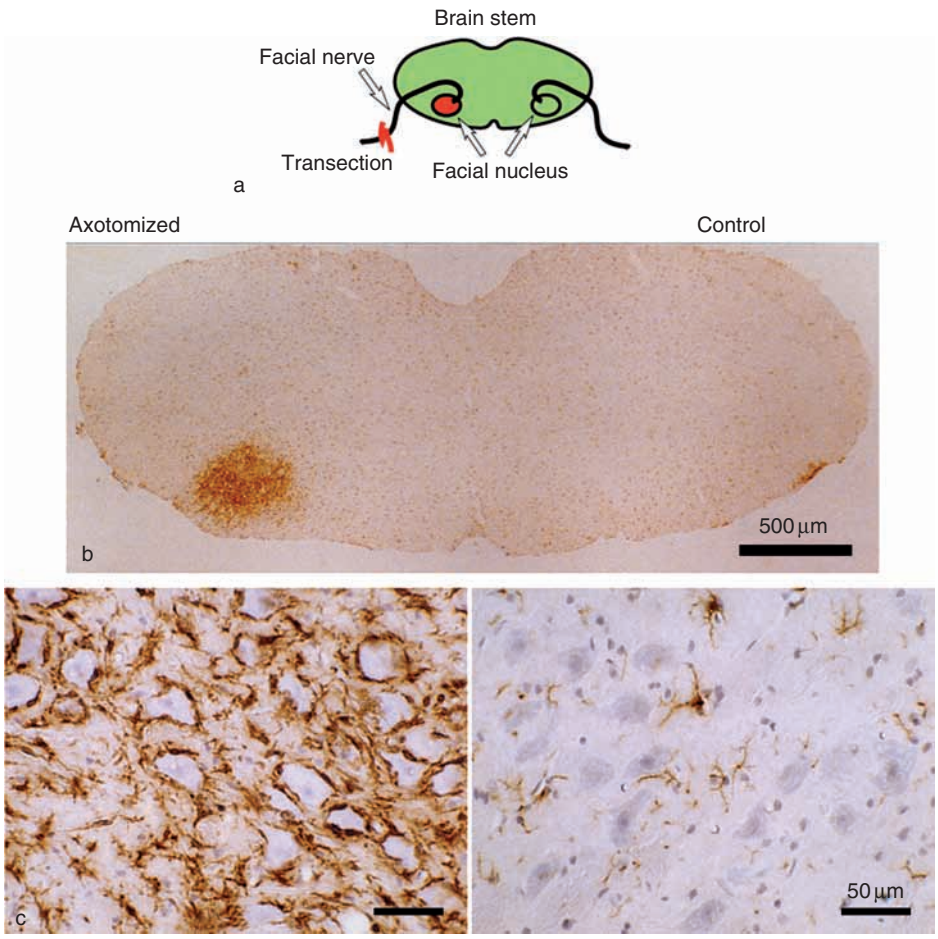

Figure 3 Facial nerve axotomy model. (a) Diagram showing the location of facial nucleus on either side in the brain stem. The facial nerve is transected at the stylomastoid foramen outside the skull. (b) Low-magnification image of a brain stem section including the facial nucleus. A section from the brain stem 1 week after facial nerve axotomy of adult rats was immunostained with an anti-Iba1 antibody. (c) High-magnification image of the axotomized (left) and nonoperated side (right) of the facial nucleus.

Interaction with Immune Cells Recruited to the CNS

The entry of lymphocytes into the neural parenchyma is recognized 1 day after injury. In a mouse model of facial nerve axotomy, the response is limited to a small number of T lymphocytes and restricted to the reacting facial nerve nucleus. Invading lymphocytes produce interferon-γ and induce expression of MHC class II and inducible NO synthase in microglia. A direct mechanical injury with the BBB disruption leads to the massive influx of a variety of immune cell types, macrophages, neutrophils, and granulocytes, as well as T lymphocytes. They show numerous and rapid recruitment to the vasculature and parenchyma in the damaged brain. A disruption of the BBB gives the microglial access to the circulating proinflammatory molecules (TNF-α, IL-1, and lipopolysaccharide), which activates microglia. Various molecules produced by activated microglia, such as cytokines, chemokines, and cytotoxic factors, are likely to contribute to leukocyte recruitment and control of CNS inflammation.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Glial Responses to Injury; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Microglia Properties.

Further Reading

Kreutzberg GW (1996) Microglia: A sensor for pathological events in the CNS. *Trends in Neuroscience* 19: 312–318.

Ladeby R, Wirenfeldt M, Garcia-Ovejero D, et al. (2005) Microglial cell population dynamics in the injured adult central nervous system. *Brain Research: Brain Research Reviews* 48: 196–206.

Moran LB and Graeber MB (2004) The facial nerve axotomy model. *Brain Research: Brain Research Reviews* 44: 154–178.

Nakajima K and Kohsaka S (2005) Response of microglia to brain injury. In: Kettenmann H and Ransom BR (eds.) *Neuroglia*, pp. 443–453. New York: Oxford University Press.

Raivich G (2005) Like cops on the beat: The active role of resting microglia. *Trends in Neuroscience* 28: 571–573.

Raivich G, Bohatschek M, Kloss CU, et al. (1999) Neuroglial activation repertoire in the injured brain: Graded response, molecular mechanisms and cues to physiological function. *Brain Research: Brain Research Reviews* 30: 77–105.

Streit WJ (2002) Microglia as neuroprotective, immunocompetent cells of the CNS. *Glia* 40: 133–139.

Microtubule Associated Proteins in Neurons

L Cassimeris, Lehigh University, Bethlehem, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

From a cell biologist's perspective, a neuron represents an extreme cell shape with unique intracellular compartments. Neurons typically have a relatively small, round cell body with long projections of axons and dendrites extending from the cell body. These projections differ in structure; axons are of uniform caliber and do not have many branches, whereas dendrites have extensive branches. The axon and dendrite compartments are unique intracellular environments without membrane-based barriers to separate them from each other or the cell body. The ability of neurons to form projections and maintain their functions requires the cytoskeleton, with the microtubule cytoskeleton playing a critical role in both growth and maintenance of the neuronal cell.

Microtubules are polymers that are required for a number of functions relevant to terminally differentiated neurons, including vesicle and organelle motility, cell polarity, and cell motility. The hollow microtubule is assembled from 13 linear chains of alpha/beta tubulin heterodimers; each linear chain is termed a protofilament. Within each protofilament, the dimers bind to each other in a head-to-tail orientation. Protofilaments bind to each other around a central open lumen, forming a polarized tube structure having two structurally different ends. These two ends are referred to as plus (beta tubulins exposed at the end) and minus (alpha tubulins exposed at the end).

The organization of microtubules differs in axons and dendrites. In each compartment, microtubules do not extend for the entire length of the projection but instead are found as numerous overlapping polymers. In axons, the microtubules all face in the same direction, with the plus ends all facing the distal end of the projection. In contrast, dendrites are filled with microtubules of mixed polarity: some have their plus ends facing the distal end of the projection, and others have their minus ends facing this direction.

Microtubule Dynamic Turnover

Microtubules continually convert between a growing state, in which subunits add to the end of the microtubule, and a shortening state, in which the protofilaments peel away from each other and fall apart, releasing individual dimers into the cytoplasm. This process is called dynamic instability to reflect the unstable character of microtubules. Dynamic instability is a nonequilibrium process and requires energy input. The energy comes from assembly of tubulin dimers bound to guanosine triphosphate (GTP). Hydrolysis of GTP to guanosine diphosphate (GDP) follows assembly, making the bulk of the microtubule composed of tubulin-GDP. The microtubule plus end is the faster growing end, and it is also more dynamic, switching to shortening and back to growth more frequently than does the minus end. In most cells, the minus end is attached to the centrosome and is not dynamic. Neurons are unique in having large numbers of microtubules with free plus and minus ends in their axons and dendrites. Under some conditions, microtubules in cells can also show net assembly at one end and net disassembly at the other, a process termed treadmilling.

Microtubule-Associated Proteins Regulate Microtubule Dynamics

For many microtubule-based functions in the neuron, such as providing tracks for vesicle transport, it is important for microtubules to remain stable and not depolymerize. In other cases, such as extension of the growth cone, microtubules must be highly dynamic and able to remodel quickly. Cells use a number of associated proteins to modify microtubule assembly, making them more or less dynamic to fit these specific cell tasks. Several microtubule-associated proteins (MAPs) are expressed only in neurons, and these can be localized specifically to either axons or dendrites. This article will summarize general features of MAPs and outline how MAPs contribute to neuron function and how genetic defects in MAPs contribute to neuronal disease.

The definition used to classify a protein as a MAP has changed over the years. Originally, MAPs were defined as proteins that cofractionate with microtubules during purification or proteins that co-localize with microtubules in cells. For this article, the definition of MAPs is proteins that modify microtubule assembly by binding either to the microtubule lattice or to soluble tubulin subunits. Besides regulating microtubule assembly, MAPs also function to define the spacing between microtubules, to regulate organelle motility, and to serve as scaffold proteins to anchor kinases and phosphatases to the microtubules.

MAPs can be broadly divided into two classes of regulatory proteins depending on whether they stabilize or destabilize microtubules. Stabilization means that the MAP either promotes polymer assembly or prevents polymer disassembly. Stabilizing MAPs

include MAP1, MAP2, doublecortin, and tau. Microtubule destabilizing proteins favor loss of tubulin dimers from microtubules. They act either by preventing new dimers from adding to the polymer, by stimulating catastrophes at microtubule ends, or by severing microtubules into pieces. Examples of destabilizing proteins include stathmin and katanin. By shifting the balance between active stabilizers and destabilizers, cells can regulate the dynamic turnover of microtubules and the amount of tubulin subunits assembled into polymers. Mutations in genes encoding either microtubule stabilizers or destabilizers lead to neuronal diseases, indicating that regulation of the microtubule cytoskeleton's dynamic turnover is critically important for cell and tissue health. The stabilizing and destabilizing MAPs are described in more detail below.

Assays to Monitor MAP Functions

MAPs were first isolated from brain tissue, and their ability to regulate microtubule assembly was measured using *in vitro* assays combining purified tubulin with purified MAP fractions. Several relatively straightforward but low-resolution assays can be used to measure tubulin assembly into microtubules and determine whether a specific MAP promotes or inhibits microtubule polymerization. A common spectrophotometric assay measures light scattering by microtubules; the amount of light scattering is proportional to the amount of microtubule formation. Often microtubule assembly is 'seeded' by adding a tiny number of preassembled microtubules to serve as nucleation sites for polymerization. MAP fractions can be assayed for their ability to speed or inhibit polymerization over 30 min or less. Microtubule assembly has also been monitored by measuring the quantity of microtubules pelleted by centrifugation. Each of these assays measures the total quantity of microtubules formed but does not give any information on the function of the MAP at the resolution of an individual microtubule.

Higher-resolution assays use microscopy to examine microtubule assembly in fixed samples taken at time points or to follow the assembly of individual microtubules in real time. Samples of microtubules and MAPs can be fixed and pelleted onto coverslips, stained with antitubulin and/or anti-MAP antibodies, and examined by fluorescence microscopy. Alternatively, fixed samples can be applied to electron microscopy grids and examined after negative staining. In these assays, it is possible to differentiate between MAPs that stimulate polymerization of long microtubules from those that act by nucleating many short microtubules or those that bundle microtubules.

To understand how a MAP regulates dynamic instability, it is best to watch microtubule dynamics in real time. Individual microtubules can be viewed by high-resolution differential interference contrast (DIC) microscopy and their length changes followed over time. Microtubules also can be visualized by fluorescence microscopy in samples containing small amounts of fluorescently tagged tubulin. By following individual microtubules as they grow and shorten in the presence of a purified MAP, one can figure out how a MAP modifies dynamic instability. While assays with purified components are critical for determining mechanisms at the molecular level, it is also important to measure MAP function in the context of a cell or organ.

Cell-based assays of MAP function use either nonneuronal cell lines or explanted neurons grown in culture. To visualize microtubule assembly, cells are engineered to express green fluorescent protein (GFP)-alpha tubulin. Images of a small region of the cell are collected at time points to record the dynamic turnover of microtubules. Cells can be modified to overexpress a MAP (typically in nonneuronal cells) or to decrease expression of a MAP. In this way, the MAP's effects on microtubule assembly can be monitored.

Expression of neuronal MAPs in nonneuronal cells can give clues to MAP function simply through examination of the morphology of the cell or the organization of microtubules in the cell. For example, long, axonlike projections filled with microtubules form in fibroblast cells expressing tau. Likewise, blocking tau expression in neuronal cells results in loss of both axon formation and growth cone motility.

MAP functions are also assayed in the context of the organism, where a protein's function during neuronal development or aging can be determined. Vertebrate MAPs have been examined in mice by gene knockout. *Drosophila* is also used as a model organism for study of MAP function in neurogenesis.

Many Stabilizing MAPs Bind to the Sides of Microtubules

The best-characterized stabilizing MAPs, including MAP2 and tau, bind to the sides of microtubules through a conserved microtubule binding domain. MAP2 and tau have similar microtubule binding domains, including the conserved sequence KXGS. Tau protein isoforms, generated by alternative splicing, vary in the number of microtubule binding domains, having either three or four microtubule binding repeats. Not surprisingly, those isoforms with four binding domains are more potent microtubule stabilizers than those having only three repeats. In addition to the microtubule binding domain, stabilizing MAPs often have a projection domain that extends out from the microtubule surface (**Figure 1**). This projection domain can regulate the spacing between microtubules

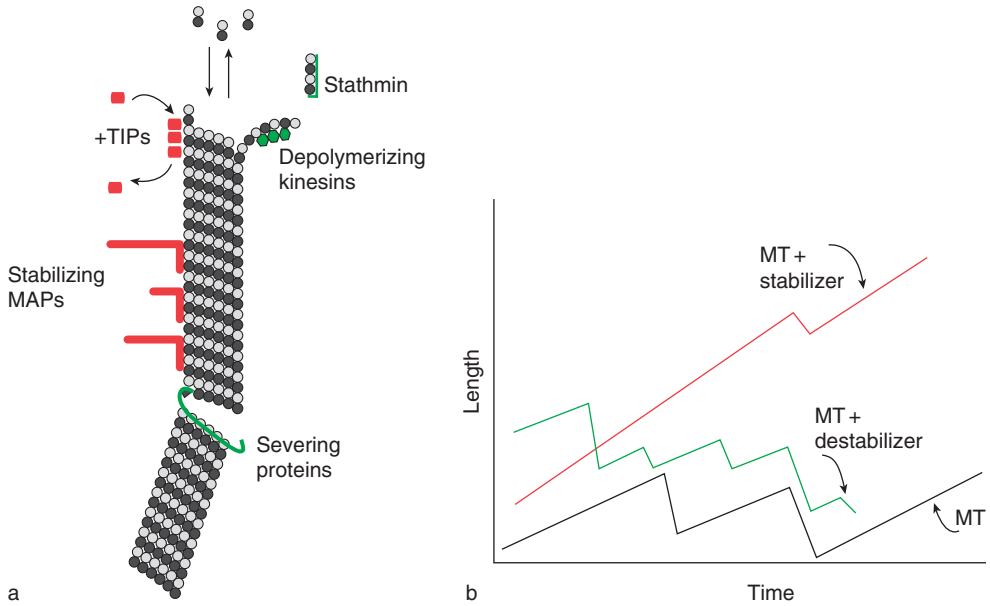

Figure 1 Microtubule-associated proteins (MAPs) stabilize or destabilize microtubules. (a) Several classes of MAPs interact with the microtubule. Stabilizing MAPs (e.g., MAP2 or tau) bind to the side of the microtubule. +TIPs associate with growing microtubule ends. Severing proteins cut the polymer into pieces. Depolymerizing kinesins disrupt the structure of the microtubule end. Stathmin binds to tubulin dimers (shown as black and gray balls) to limit polymerization. (b) Hypothetical curves of microtubule (MT) dynamic instability in the presence of stabilizing or destabilizing MAPs. The lengths of the hypothetical microtubules are shown over a timescale of several minutes.

or between microtubules and either neurofilaments or actin filaments.

MAP2 and tau compete with the motor proteins kinesin and dynein for binding to microtubules. In this way, MAP binding to microtubules can inhibit the ability of these motor proteins to transport vesicle or organelle cargos by blocking motor access to the microtubule. In experiments involving overexpression of tau, motility in the direction of microtubule plus ends (anterograde movement in the axon) is more sensitive to excess tau than retrograde movement toward microtubule minus ends.

While MAP2 and tau share a similar microtubule binding domain, they differ at the cellular level by their localizations to different compartments in the neuron. MAP2 is found in dendrites, while tau is found in axons. Consistent with these unique localizations, mice deficient in MAP2 have defects in dendrites, and those deficient in tau have defects in small-caliber axons. Mice lacking tau show an increased expression of another wall-binding MAP, MAP1A, which may explain how their nervous system is able to develop normally, perhaps using MAP1A to compensate for the loss of tau.

MAP1A and 1B are large proteins encoded on two separate genes.These polypeptides associate with smaller polypeptides, the light chains LC1, LC2, and LC3. MAP1A and 1B also bind to the sides of microtubules and stabilize them, although they are not as potent microtubule stabilizers as MAP2 and tau are.

MAP1B is the first MAP detected during neurogenesis and is required for nervous system development in mice. MAP1B levels are low in the adult central nervous system but remain high in the peripheral nervous system. This MAP is found throughout the nerve cell, with the highest concentration at the distal end of the axon.

In localization and expression levels during development, MAP1A is opposite of MAP1B. MAP1A is present in adult neurons at a concentration tenfold that of its level during development. Like MAP1B, MAP1A is found throughout the nerve cell, but MAP1A is present at higher levels in dendrites.

The final wall-binding MAP considered here is doublecortin. Doublecortin is one of several proteins linked to lissencephaly, a severe disease resulting from loss of cortical neuron migration during brain development. Doublecortin associates with microtubules through a pair of binding domains and stimulates microtubule polymerization. Mutations that result in lissencephaly are located in the microtubule binding domains of doublecortin, suggesting that the disease results from a disruption of the microtubule cytoskeleton.

Several other gene mutations give rise to lissencephaly, including mutations in genes encoding LIS1, a regulator of the microtubule-based motor dynein and

alpha-1 tubulin, a subunit of microtubules. The ability of mutations in these genes to cause lissencephaly indicates that the microtubule cytoskeleton plays a critical role in neuronal migration. It should be noted that mutations in the gene encoding reelin, an extracellular matrix protein, also result in lissencephaly.

Regulation of Wall-Binding MAPs

In general, the wall-binding MAPs are regulated by phosphorylation. Phosphorylation adds negatively charged phosphate groups to the MAP and reduces MAP binding to the microtubule, reducing the microtubule stabilizing function of MAPs (Figure 1). Phosphorylation also changes the structure of tau from a flexible protein in the dephosphorylated state to a stiffer protein when phosphorylated. Whether other MAPs undergo structural changes with phosphorylation is not known.

A large number of kinases phosphorylate MAP1, 2, and tau. For example, MAP2 is phosphorylated by protein kinase A, calcium/calmodulin-dependent protein kinase II (CaMKII), protein kinase C, extracellular signal-regulated kinases, glycogen synthase kinases, cyclin-dependent kinases, and MAP/microtubule affinity regulating kinases (MARKs). MARKs recognize the KXGS sequence found in the microtubule binding domain of MAP2 and tau and thus disrupt MAP binding to microtubules. Tau is also phosphorylated by a number of kinases, including the c-Jun N-terminal kinase (JNK) family of stress kinases. Tau phosphorylation downstream of JNK is likely important during the hyperphosphorylation of tau in Alzheimer's disease.

Several factors determine the phosphorylation state of each MAP. First, some kinases and phosphatases have specific subcellular localizations within the neuron, which limits their MAP targets to those present in that part of the cell. For example, both MAP2 and tau are phosphorylated by CaMKII *in vitro*, but only MAP2 should be phosphorylated in cells because both MAP2 and this kinase are found in dendrites. Second, kinases and phosphatases bind to the microtubule cytoskeleton by associating with MAPs, which localizes enzyme activity near their MAP substrates.

MAP phosphorylation is also linked to disease. Tau is hyperphosphorylated, which causes it to aggregate and form filaments, in Alzheimer's disease, Pick's disease, frontotemporal dementia with parkinsonism linked to chromosome 17, and several other diseases. The cause of tau hyperphosphorylation and the pathological consequences of this hyperphosphorylation are presently under study. One model has recently suggested that tau is first expressed at higher than normal levels (for reasons unknown), which inhibits the ability of motor proteins to transport vesicles on microtubules. In this model, the cell responds by hyperphosphorylating tau to release it from the microtubules, but the loss of tau from the microtubule reduces microtubule stability and leads to neurodegeneration.

In addition to phosphorylation, the wall-binding MAPs can be regulated by a nonenzymatic protein termed mapmodulin. Mapmodulin binds MAPs 1B, 2, and tau and preferentially binds to soluble MAPs, compared with those bound to microtubules. It is interesting to note that mapmodulin may sequester MAPs and prevent them from binding to microtubules. One hypothesized function for mapmodulin is to displace MAPs from microtubules to enhance motor protein-driven transport of vesicles along microtubules.

MAP Binding to Microtubule Plus Ends

A growing number of MAPs have been identified that associate with growing ends of microtubules (as comets 1–2 μm in length at the microtubule end). Collectively, these MAPs are called +TIPs to denote their association with microtubule plus ends. Some +TIPs bind directly to the assembling microtubule end, perhaps in association with incoming tubulin dimers. Other +TIPs hitch a ride by binding those proteins interacting directly with the growing microtubule end. The +TIPs bind for a short time and then dissociate from the microtubule (Figure 1). In this way, individual proteins remain stationary as the microtubule end grows; these MAPs do not move with the growing microtubule.

In general, the +TIPs stabilize the growing microtubule plus end. For example, the +TIP CLIP170 stimulates microtubule rescues, promoting the transition from a shortening state back to a growing state. CLIP170-associated protein (CLASP) has a similar function, while adenomatous polyposis coli inhibits catastrophes. These +TIPs often bind to a subset of microtubules, allowing the cell to regulate the dynamics of this subset differently from other microtubules in the same cytoplasm.

By binding to growing microtubule plus ends, the +TIPs can also localize additional proteins to this domain. The minus end-directed motor dynein and its accessory complex dynactin bind to microtubule plus ends through their interaction with CLIP170. One current model suggests that binding of dynein/dynactin to the microtubule plus end transports this motor complex to the periphery of cells, where the motor complex can pick up vesicles for delivery back to the center of the cell.

Microtubule Destabilizing Proteins

By destabilizing microtubules, this broad group of proteins likely increases microtubule turnover in cells, allowing the microtubule cytoskeleton to reorganize quickly. As mentioned above, the destabilizing proteins can act by one of three general mechanisms: sequester tubulin dimers and prevent their polymerization, disrupt the structure at the microtubule end to favor disassembly, or cut microtubules into smaller pieces.

Stathmin (also known as Oncoprotein 18) is a small cytosolic protein that binds to tubulin dimers rather than to microtubules. Each stathmin binds two tubulin dimers, preventing those dimers from polymerizing, resulting in destabilization of microtubules (Figure 1). By an unknown mechanism, stathmin can also function as a catastrophe promoter and stimulate microtubules to shorten. Stathmin may interact with tubulin dimers at the end of the microtubule and favor the splaying conformation of shortening microtubules. This latter model is consistent with the structure of the stathmin/tubulin complex but has not been demonstrated experimentally.

Stathmin is expressed ubiquitously, but several proteins related to stathmin are expressed specifically in the nervous system. These neuronal stathmin family proteins, SCG10, SCLIP, and RB3, all contain a conserved region homologous to stathmin and additional extensions that target them to membranes (golgi apparatus or secretory vesicles). Each of the neuronal proteins binds tubulin dimers and destabilizes microtubules, although the affinity of each protein for tubulin is different.

Each stathmin family member is likely to have a unique function in the brain, a conclusion based on the phenotypes of mice lacking specific proteins or on cell-based assays of protein function. Experiments in cultured neurons have demonstrated that SCG10 and SCLIP are required to control growth cone expansion and axon branching, respectively. In contrast, mice lacking stathmin are viable, but their axons lose integrity as they age. It is interesting that stathmin –/– mice have revealed a role for this protein in controlling both learned and innate fear. How the microtubule cytoskeleton contributes to fear responses is not yet apparent.

Members of the kinesin-13 subfamily also cause microtubule depolymerization. Unlike other members of the kinesin family, the kinesin-13 proteins do not possess motor activity. Instead, they use adenosine 5′-triphosphate (ATP) hydrolysis to disrupt the structure of microtubule tips, stimulating microtubules to depolymerize (Figure 1).

Kif2a, a kinesin-13 protein, functions as a microtubule depolymerizing protein in neurons. Mice lacking Kif2a have brains with numerous defects and die within a day of birth. Microtubules are less dynamic in the growth cones of neurons from the Kif2a knockout mouse. Neurons also show defects in migration during development.

Two members of the large AAA ATPase family of proteins, katanin and spastin, destabilize microtubules by cutting them perpendicular to the microtubule's long axis. Without stabilizing MAPs bound to the sides of the microtubules, these cut fragments would rapidly depolymerize. To sever a microtubule, six kananin polypeptides form a ring surrounding it and then use ATP hydrolysis to somehow break tubulin–tubulin contacts within it (Figure 1).

Spastin also functions to sever microtubules. Spastin is so named because mutations in the spastin gene are associated with hereditary spastic paraplegia. Loss of spastin expression reduces the density of microtubules in axons and at synaptic boutons, opposite to what one might expect at first glance after loss of a severing protein. The phenotype of cells lacking spastin suggests that spastin contributes to microtubule density, perhaps by generating small microtubule fragments that serve as templates for formation of longer microtubules. Stabilizing MAPs must also protect the severed fragments from depolymerization.

Concluding Remarks

The term MAP describes a number of different proteins that function to regulate microtubule assembly. These proteins regulate the assembly dynamics of microtubules to either temper the dynamics and stabilize the microtubule or enhance the dynamics to speed microtubule turnover. Expression levels of MAPs and the regulation of their activities contribute to the stability and organization of the microtubule cytoskeleton in neurons. MAP activities can be regulated within subdomains of the cell, setting up unique microtubule populations in different areas of the cell.

The microtubule cytoskeleton does not function alone in providing support for neuronal cell processes. Neurofilaments and actin filaments also play critical roles in the neuron. MAPs likely also mediate some of the cross talk between cytoskeletal filaments. In particular, MAP2c and MAP1 are able to bind both microtubules and actin filaments.

Several recent studies have highlighted the role of MAP mutations in causing human disease, particularly diseases resulting in developmental defects in brain development. Further study of how MAPs contribute to neuron development and migration are needed to understand these processes.

See also: Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Cytoskeleton in Plasticity; Microtubules: Organization and Function in Neurons.

Further Reading

Baas PW, Karabay A, and Qiang L (2005) Microtubules cut and run. *Trends in Cell Biology* 15: 518–524.

Baas PW and Qiang L (2005) Neuronal microtubules: When the MAP is the roadblock. *Trends in Cell Biology* 15: 183–187.

Cassimeris L and Spittle C (2001) Regulation of microtubule associated proteins. *International Review of Cytology* 210: 163–226.

Dehmelt L and Halpain S (2004) Actin and microtubules in neurite initiation: Are MAPs the missing link? *Journal of Neurobiology* 58: 18–33.

Dehmelt L and Halpain S (2004) The MAP2/tau family of microtubule-associated proteins. *Genome Biology* 6: 204.

Evans KJ, Gomes ER, Reisenweber SM, et al. (2005) Linking axonal degeneration to microtubule remodeling by spastin-mediated microtubule severing. *Journal of Cell Biology* 168: 599–606.

Horesh D, Sapir T, Francis F, et al. (1999) Doublecortin, a stabilizer of microtubules. *Human Molecular Genetics* 8: 1599–1610.

Kornack DR and Giger RJ (2005) Probing microtubule +TIPs: Regulation of axon branching. *Current Opinion in Neurobiology* 15: 58–66.

Kreis T and Vale R (eds.) (1999) *Guidebook to Cytoskeletal and Motor Proteins*, 2nd edn., New York: Oxford University Press.

Lee VM and Trojanowski JQ (1999) Neurodegenerative tauopathies: Human disease and transgenic mouse models. *Neuron* 24: 507–510.

Mi K and Johnson GV (2006) The role of tau phosphorylation in the pathogenesis of Alzheimer's disease. *Current Alzheimer Research* 3: 449–463.

Ozon S, Maucuer A, and Sobel A (1997) The stathmin family. *European Journal of Biochemistry* 248: 794–806.

Wordeman L (2005) Microtubule-depolymerizing kinesins. *Current Opinion in Cell Biology* 17: 82–88.

Microtubules: Organization and Function in Neurons

P W Baas, K A Myers, L Qiang, and V C Nadar, Drexel University College of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons are one of the most remarkable cells in nature. They are the cell type whose exaggerated morphology is most intimately associated with its unique functions. A typical vertebrate neuron consists of a rounded cell body (located in the brain, spinal cord, or peripheral ganglion), several dendrites, and a single elongated axon. The axon is highly specialized to transmit information outward, while the dendrites form a receptive field for the uptake of information. The axon is effectively unlimited in its growth potential and can traverse great distances within the body, extending, for example, from the spinal cord to the fingertips. Dendrites, by contrast, are generally short, stout, and highly branched. In addition to these morphological differences, axons and dendrites also differ compositionally. Axons contain only a subset of the organelles found in the cell body, while dendrites generally contain all the same types of organelles as the cell body does. Most protein synthesis and organelle biogenesis occur within the cell body of the neuron, thus demanding highly sophisticated mechanisms for actively transporting specific proteins and organelles into each type of process. During development, there is a great deal of motility involved in the wiring of the nervous system, so that the cell bodies of the neurons arrive at their proper locations and so that the tips of axons reach their appropriate target tissues. Thus, a number of challenges must be met in order for neurons to establish and maintain their high degree of morphological, functional, and compositional polarity. Meeting these challenges depends on tightly regulated architectural elements within the neuron and sophisticated mechanisms of transport and motility.

The architectural elements of the neuron consist of three cytoskeletal filaments: microtubules, actin filaments, and intermediate filaments. Together, these three filaments are responsible for the generation of neuronal morphology and function. Microtubules are particularly important for the architecture of axons and dendrites and for the specific and directional transport of proteins and organelles within these processes. Over the past few decades, enormous progress has been made in understanding the central role of microtubules in serving these functions. Microtubules are classically known as 'spindle fibers' because they are the filaments that comprise the mitotic spindle during cell division. It is interesting that microtubules are extremely abundant in neurons, even though neurons are terminally postmitotic cells that no longer form mitotic spindles. Microtubules occupy axons and dendrites as paraxially aligned arrays. These microtubule arrays provide a structural backbone for axons and dendrites that allows them to acquire and maintain their specialized morphologies. In addition, microtubules are the major railways along which proteins and organelles are actively transported in both directions within axons and dendrites. Microtubules are also critically important for early developmental stages of the neuron, such as migration of the cell body and the cue-dependent pathfinding of the growth cone at the tip of the elongating axon. A vast array of accessory proteins regulate the properties and functions of microtubules during development and throughout the life of the neuron. Mutations and flaws in the regulation of some of these microtubule-related proteins give rise to major neuropathies that plague the human population.

Organization of Neuronal Microtubules

Microtubules are hollow polymers made up of tubulin subunits. Neurons, like other eukaryotic cells, contain a pool of free tubulin subunits and a pool of the microtubule polymers. These two pools dynamically exchange with one another such that microtubules undergo bouts of assembly and disassembly. Tubulin is a heterodimer consisting of an alpha subunit and a beta subunit. Because the two subunits are different, the microtubule is an intrinsically polar filament. One end, termed the plus end, is favored for the addition and subtraction of subunits, while the other end, termed the minus end, is less favored for these dynamics. The assembly properties of a microtubule are defined by a mechanism called dynamic instability, in which rapid bouts of assembly and disassembly are ultimately attenuated by molecules that selectively stabilize certain microtubules within the array. Neurons are rich in molecules that can stabilize microtubules, as well as other molecules that can destabilize microtubules and promote their disassembly. Tight control over microtubule assembly, disassembly, stabilization, and destabilization (i.e., 'microtubule dynamics') is one key means by which neurons change the configuration of microtubules during critical stages of development. Regulation of microtubule dynamics is particularly important for rapid changes in neuronal morphology and motility,

a good example of which is the turning of the growth cone in response to environmental cues that guide it to its target.

It is important to realize that the polarity of the microtubule is not manifest only at its two ends but also exists all along the length of the polymer. This structural polarity is recognized by a class of enzymes called molecular motor proteins that 'walk' along the surface of the microtubule. Cytoplasmic dynein walks toward the minus end of the microtubule, while most members of the kinesin superfamily walk toward the plus end of the microtubule. A few unusual kinesins walk toward the minus end (Figure 1(a)). Each molecular motor protein can potentially interact with different types of 'cargo' to be moved along the polymer. Thus, the organization of microtubules relative to their polarity is key to the directionality with which different cytoplasmic constituents are transported. In the axon, microtubules have a uniform polarity orientation, with their plus ends directed away from the cell body toward the axonal tip (Figure 1(b)). By contrast, dendritic microtubules have a mixed, or nonuniform, polarity orientation, with roughly equal numbers of microtubules directed each way (Figure 1(b)). These distinct microtubule polarity patterns are one of the most essential and earliest differences to arise between axons and dendrites and are almost certainly responsible for many of the compositional and morphological differences that distinguish the two types of processes. For example, dendrites contain Golgi stacks and axons do not, because Golgi elements are transported toward minus ends of microtubules by cytoplasmic dynein. Axons may become longer and more slender than dendrites because vesicular organelles that expand the membrane have a unidirectional vector to the tip of the growing axon via their plus-end-distal microtubules, while the same is not true of dendrites. It should be noted that, despite the enormous influence of microtubule polarity patterns on neuronal morphology and organelle distribution, the distinct polarity patterns of axonal and dendritic microtubules cannot explain all aspects of neuronal polarity. For example, some molecular motors transport cargo toward plus ends of microtubules into one type of process but not the other, even though both axons and dendrites contain microtubules with a plus-end-distal orientation.

Axonal and dendritic microtubules also differ compositionally, for example with regard to the complements of fibrous microtubule-associated proteins (MAPs) that decorate each type of microtubule. Fibrous MAPs are one category of microtubule-regulatory protein; these MAPs consist of a microtubule-binding domain as well as a domain the projects away from the microtubule. MAPs bind along the surface of the microtubule and can influence properties of the microtubule, such as its stability, stiffness, and interactions with other cytoskeletal elements, such as actin filaments. Axonal microtubules are particularly rich in the MAP termed tau, and dendritic microtubules are rich in MAP2. These different MAPs are known to space the microtubules more closely in axons and farther apart from one another in dendrites because

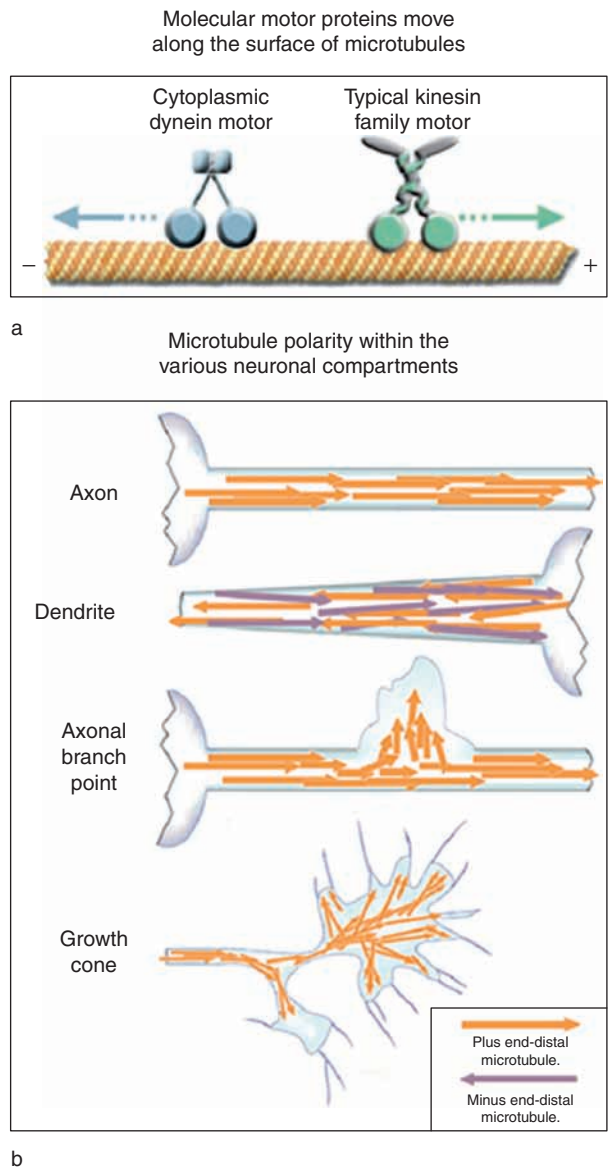

Figure 1 Neuronal microtubule organization. (a) By reading the polarity of the microtubule, molecular motor proteins are able to move along the microtubule surface toward either the plus end or the minus end. Cytoplasmic dynein moves toward the minus end, and most members of the kinesin superfamily move toward the plus end. (b) Microtubules in the axon are uniformly oriented with plus ends distal to the cell body, whereas microtubules in dendrites are nonuniformly oriented. At sites of branch formation, microtubules in the parent axon are severed into short pieces, some of which transit into the new branch with plus ends outward. In growth cones, short and long microtubules extend, plus ends outward, into filopodia.

MAP2 has a longer projection domain than tau does. In addition, the enrichment of tau in axons is relevant to a number of pathologies, termed 'tauophathies,' in which the tau molecules become hyperphosphorylated and lose their association with the microtubules. The dissociated tau often forms abnormal filaments that are detrimental to the health of the axon. For example, tau forms the 'neurofibrillary tangles' that are a hallmark feature of Alzheimer's disease.

There is great interest in understanding the mechanisms by which neurons organize microtubules into the distinct patterns that characterize axons and dendrites. In simple nonneuronal cells, microtubules are generally organized by structures such as the centrosome, which act as nucleation sites for the formation of the microtubules. The microtubules elongate, with the plus end growing away from the centrosome and the minus end associated with the centrosome. This generates an orderly array of uniformly oriented microtubules. In neurons, microtubules are also nucleated at the centrosome (located in the cell body of the neuron), but the microtubules are then severed from the centrosome so that they can be deployed into axons and dendrites. The microtubules are deployed by molecular motor proteins that move the microtubules specifically with either their plus end leading (into axons and dendrites) or with their minus end leading (into dendrites only). Thus, in neurons, the organization of microtubules relative to their polarity is determined, not by the attachment of the microtubules to their sites of nucleation, but instead by the forces that transport them into each type of neuronal process.

Another critically important feature of the neuronal microtubule arrays is that each array consists of individual microtubules that assume a variety of different lengths. The array of microtubules within an axon or dendrite extends from the cell body to the tip of the process, but individual microtubules generally do not traverse the entire length of the process. Some microtubules are extremely short, only a few micrometers in length, and other microtubules can achieve great lengths, exceeding a hundred micrometers or even hundreds of micrometers in longer axons. The longer microtubules are important for two reasons: they act as architectural struts that oppose the propensity of the axon or dendrite to retract, and they act as major long-distance railways for organelle transport. The short microtubules are important because they are highly mobile and provide the means by which tubulin is actively transported down the axon.

Axonal and Dendritic Transport

For decades, axons were considered to be incapable of sustaining local protein synthesis. Today we know that small amounts of certain proteins are synthesized in axons, at least during development. Local protein synthesis can be quite important for events such as growth cone navigation. Even so, most of the proteins destined for the axon are synthesized in the cell body and then actively transported down the length of the axon. In addition to this anterograde transport, retrograde transport moves substances from the axonal tip back toward the cell body. Retrograde transport includes endocytosed materials such as growth factors and also substances destined for degradation within the lysosomes of the cell body. Microtubules are the major railways for membrane-bound vesicular organelles (and the proteins associated with the vesicles) moving in either the anterograde or retrograde direction. Members of the kinesin superfamily provide the forces necessary for anterograde transport of these vesicular organelles, while cytoplasmic dynein is the motor for their retrograde transport. This type of transport is known as fast axonal transport and occurs at rates of about 50–400 mm per day, with larger organelles such as mitochondria moving more slowly and smaller organelles moving more quickly. Over shorter distances, similar transport can occur along actin filaments, but it is the microtubule array of the axon that serves as the major long-distance highways for fast axonal transport (**Figure 2(a)**).

The other type of axonal transport, termed slow axonal transport, has been far more controversial over the years and has proven more difficult to study mechanistically. Slow axonal transport is the transport of the cytoskeletal proteins themselves, together with their associated proteins (such as tau, for example) and a vast array of 'soluble' proteins such as cytosolic enzymes. This type of transport was originally termed 'slow' because the rate at which these proteins are conveyed down the axon was determined to be many times slower than fast transport. Transport rates were originally determined using a biochemical technique in which newly synthesized proteins are allowed to incorporate radiolabeled amino acids, after which the levels of the proteins that incorporate the label are assessed as they progress down the axon. On the basis of the slow rate, it was hypothesized that the cytoskeleton marches coherently down the axon in the form of polymers, together with their associated cytoskeletal proteins. Microtubules and neurofilaments move at roughly the same rate (0.2–1 mm per day), and actin moves a bit faster (2–8 mm per day). It was posited that the 'soluble' proteins somehow affiliate with the actin cytoskeleton and hence move at roughly the same rate as actin. This model for slow axonal transport, proposed in the early 1980s, met with numerous challenges over the ensuing decades because live-cell imaging studies failed to reveal the slow, coherent march of cytoskeletal polymers down the axon.

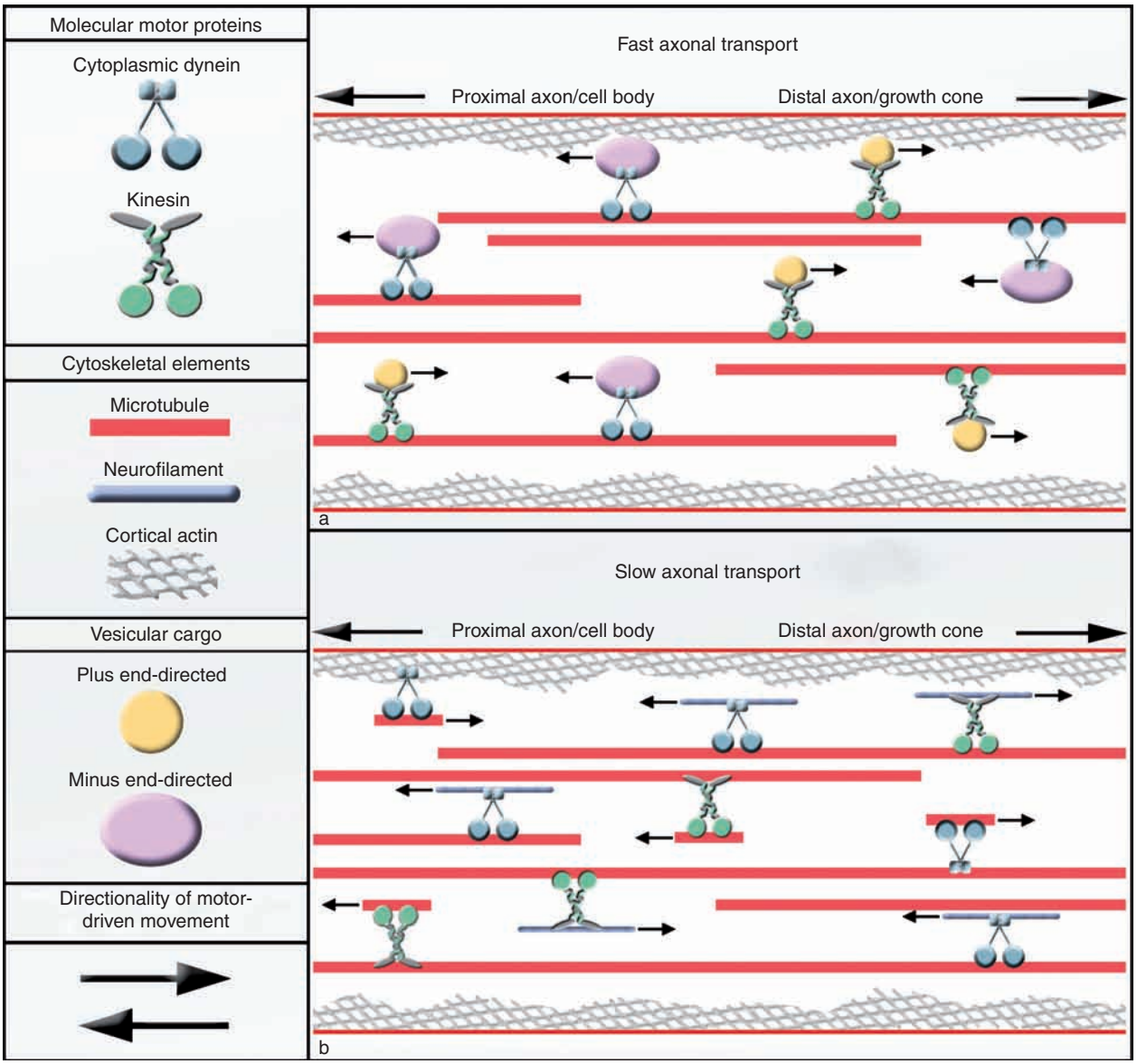

Figure 2 Axonal transport. (a) Fast axonal transport involves concerted movement of vesicular cargo by molecular motor proteins along microtubules. This type of transport occurs at rates of 50–400 mm per day, depending on the size of the particular cargo being moved. Kinesins are responsible for anterograde fast transport of organelles, which may contain proteins important for cell growth and/or growth cone navigation. Retrograde fast axonal transport is mediated by cytoplasmic dynein, which is responsible for moving structures destined for degradation, as well as packaged growth factors, to the cell body. (b) Slow axonal transport is the transport of the cytoskeletal elements and their associated proteins, as well as a variety of cytosolic proteins. The term 'slow axonal transport' was given to this type of motor-driven movement on the basis of early studies, in which microtubules and neurofilaments were documented to move at rates of 0.2–1 mm per day, much slower than rates of fast axonal transport. More-recent analyses have revealed that the movements of individual cytoskeletal elements are rapid, intermittent, and bidirectional, such that the average rate is slow although individual movements are actually quite fast. Under the current model for slow axonal transport, cytoplasmic dynein transports short microtubules (by generating forces against both long microtubules and the actin cytoskeleton) in the anterograde direction while also mediating the retrograde transport of neurofilaments against long microtubules. Kinesin motors transport neurofilaments anterogradely and short microtubules retrogradely (by generating forces against long microtubules). Little is known about the mechanisms of actin transport.

The past few years have seen a revolution in the mechanistic understanding of slow axonal transport, in large part owing to direct studies on microtubule behaviors in cultured neurons. More-sophisticated live-cell imaging approaches have revealed that microtubules and neurofilament polymers are indeed in transit down the axon. However, these studies showed that the polymers are not transported coherently or slowly but rather very asynchronously and quite rapidly. When they move, microtubules and

neurofilaments are transported at rates similar to those for fast axonal transport. However, most of the polymers are immobile for prolonged periods. Thus, slow axonal transport is an average of a small number of intermittent fast movements and a great deal of nonmovement, or 'pausing,' of the polymers. It is interesting that the microtubules that are transported are particularly short, usually on the order of 5–7 μm or shorter (Figure 2(b)). As noted earlier, microtubules in the axon assume a variety of lengths. Presumably, some of the short microtubules that are actively transported can disassemble to provide subunits for the elongation of other microtubules. Other short microtubules can elongate and thereby become immobile. In order for long microtubules to resume movement, they need to be cut into pieces that are shorter than the threshold for transport, which appears to be about 7 μm. The process of microtubule severing, therefore, is a critical determinant of where and when microtubules are most avidly transported. For example, when axonal branches form, microtubules are severed within the parent axon at the site of branch formation so that large numbers of short microtubules can transit into the new branch (Figure 1(b)). The major microtubule-severing proteins in the neuron are enzymes called katanin and spastin. Abnormalities to spastin can lead to the neuropathy known as spastic paraplegia.

Of course, there is also dendritic transport, but transport of organelles and proteins within dendrites is far less studied than their transport in axons. The abundant presence of ribosomes and Golgi elements within dendrites introduces the possibility for more local protein synthesis and processing than in the axon. Even so, only certain proteins are locally synthesized in dendrites (e.g., tubulin is not), and typical organelles are simply too large to diffuse. Dendritic transport probably utilizes machinery to similar to that of axonal transport, but the traffic would presumably be more complex because the microtubules are nonuniformly oriented in dendrites. As noted earlier, the lack of a unidirectionally polarized microtubule array may explain why dendrites remain short whereas axons appear to be effectively unlimited in their growth potential. In the dendrite, for example, certain organelles may tend to move back and forth rather than unidirectionally to the tip of the process, as is the case with the axon. Slow axonal transport is certainly also more complex in dendrites, given that microtubules would have to move into the dendrites with two different polarity orientations.

Motors Organize and Generate the Neuronal Microtubule Arrays

The profound importance of neuronal microtubules raises the question of how the microtubule arrays of axons and dendrites are established during the development of the nervous system and how they are regulated throughout the lifetime of the neuron. Studies on these mechanisms have been an active topic of research over the past two decades. As noted earlier, the machinery that establishes the microtubule polarity patterns of axons and dendrites is essentially the same as the machinery that continuously supplies microtubules to these processes throughout the life of the neuron. Microtubules originate at the centrosome and are oriented by molecular motors that convey them specifically with one or the other end of the microtubule leading. The transport of the microtubules occurs via motors that generate forces against structures with more resistance to movement in the opposite direction than the direction in which the microtubule is able to move. Microtubules are thought to move via motors pushing against either actin filaments or longer, immobile microtubules.

Ongoing studies are aimed at identifying precisely which motors are responsible for transporting microtubules in various compartments of the neuron. Cytoplasmic dynein appears to be the primary motor responsible for transporting microtubules with their plus ends leading. Because all motion is relative, a short microtubule will move with its plus end leading when cytoplasmic dynein pushes against a cargo with greater resistance to movement than the short microtubule. The microtubules within axons and the plus-end-distal microtubules in dendrites are thought to be transported in this fashion. The minus-end-distal microtubules are conveyed into dendrites by a kinesin motor called CHO1/MKLP1, which transports microtubules with their minus ends leading by pushing against microtubules of the opposite orientation. It is interesting that this motor was originally believed to be specific to the mitotic spindle, where it localizes to a region of overlap between oppositely oriented microtubules. More recently, at least two other kinesin motors (namely, Eg5 and Kif15) originally identified in the mitotic spindle have now been found to be components of the neuronal microtubule arrays. These kinesins have microtubule binding motifs outside the motor domain and are specialized for generating forces between two microtubules rather than between a microtubule and a vesicular organelle. The use of these motors in neurons suggests that the same type of motor-driven forces that organize and transport microtubules during the formation and functioning of the mitotic spindle are also important for the transport, organization, and regulation of the postmitotic microtubule arrays in axons and dendrites.

A variation on this theme occurs during the process of neuronal migration. During development, many types of neurons undergo significant movement

from their sites of birth to their ultimate destinations; this movement involves translocation of the cell body, not just process outgrowth. The migration of the neuron requires force generation within the cytoskeleton, and these forces may be a variation of the same forces that fuel slow axonal transport. Most microtubules apparently remain attached to the centrosome during neuronal migration and are far more cross-linked with each other than in a neuron with a stationary cell body and growing axon. The cell body extends a 'leading process' containing microtubules, many of which remain attached to the centrosome. Instead of individual microtubules moving independently, the same motor-driven forces that would otherwise transport the microtubules during axonal outgrowth are utilized to tow the centrosome and nucleus along, thus causing the cell body of the neuron to move along with the microtubule array. Such movements also require a cage of microtubules around the nucleus to ensure that the nucleus moves in unison with the leading process, as well as actomyosin-driven forces within the leading process.

More Proteins That Regulate Neuronal Microtubules

As noted earlier, fibrous MAPs are another major factor in the organization and regulation of the neuronal microtubule arrays. MAPs are important for microtubule transport for many reasons; they can influence how motor proteins interact with microtubules; they can stabilize microtubules so that they do not simply depolymerize once released from the centrosome; and they can regulate the access of severing proteins to the microtubule lattice. In particular, tau has been identified as an important determinant of microtubule severing in the axon because microtubules rich in tau are far less accessible to the severing proteins. This may relate to why the axonal microtubule array degenerates in tauopathies such as Alzheimer's disease, given that the microtubules would fall victim to uncontrolled severing once the tau has been dissociated from the microtubule. In addition, because tau's association with microtubules can be tightly regulated by phosphorylation and dephosphorylation, it is appealing to speculate that microtubule severing might be regulated locally, for example at axonal branch sites, by signaling cascades that determine whether tau is bound to the microtubules. MAP1b is another important MAP that is particularly important for regulating the plasticity of the microtubule array during axonal growth and has some functions that overlap with those of tau. These ideas are useful to contemplate because they demonstrate how a variety of different microtubule-related proteins can work cooperatively to regulate the neuronal microtubule arrays and because they also illustrate how flaws in any of the microtubule-related proteins can give rise to severe pathological problems.

Motors, MAPs, and severing proteins only scratch the surface of the enormous array of molecules that interact with the microtubule arrays of the neuron. For example, there are microtubule stabilizing proteins that are generally not considered to be traditional fibrous MAPs, and these include stable tubule only protein, doublecortin, and a family of cross-linker proteins called plakins. There are proteins such as stathmin, SCG10, and even some members of the kinesin superfamily that promote microtubule dynamics and disassembly. Any protein that influences the actin cytoskeleton is likely to influence microtubules indirectly because the distribution of microtubules in developing neurons and particularly within the growth cone is strongly influenced by the organization and distribution of actin filaments. Proteins such as Lis1 and members of the dynactin complex are important for regulating the functions of cytoplasmic dynein and other molecular motors. It is interesting that both Lis1 and doublecortin appear to be quite important for regulating microtubules during neuronal migration and appear to interact with one another. In addition, there is a recently discovered family of proteins, called +TIPs, that associate mainly with the plus end of the microtubule as it undergoes bouts of assembly. The +TIPs can influence the assembly properties of the microtubule but can also influence the interactions of the microtubule tip with molecular motors and with cortical structures within the leading edge of the cell. Finally, a great deal of recent evidence has shown that microtubule behaviors are influenced by a wide array of small G-proteins and kinases that can affect many of the other microtubule regulatory proteins. As already mentioned for tau and spastin, abnormalities to some of these microtubule-related proteins can give rise to neuropathies and/or developmental abnormalities. For example, abnormalities in the dynein/dynactin motor system can give rise to amyotrophic lateral sclerosis, and mutations in doublecortin can give rise to defects in the lamination of the brain during embryogenesis. Table 1 summarizes some of the microtubule-related proteins and their proposed functions but is not meant to be all-inclusive.

Concluding Remarks

Microtubules are one of the most prominent cytological structures in the neuron. They appear as dense arrays of aligned filaments in both axons and dendrites. When first observed by electron microscopy,

Table 1 Partial list of microtubule-related proteins in neurons and other cell types

Microtubule (MT)-related proteins			*Mainly in neurons*	*Putative functions*	*Related disease(s)*
MT-associated proteins (MAPs)	Fibrous MAPs	MAP1A	Yes	Promote MT assembly and stabilization; determine spacing between MTs; influence MT stiffness; regulate interactions between MTs and other structures such as actin filaments; tau and MAP2 can also protect MTs from severing and promote MT bundling	Giant axonal neuropathy
		MAP1B	Yes		Unknown
		MAP2	Yes		Unknown
		Tau	Yes		Alzheimer's disease; FTDP-17
	Novel MAPs	Doublecortin	Yes	Promotes MT assembly and stabilization; interacts with Lis1	Lissencephaly; double cortex syndrome
		STOP	No	Promotes MT assembly and stabilization, especially stability against cold	Unknown
Molecular motors	Cytoplasmic dynein		No	Transports organelles and vesicles (cargo) toward minus end of MTs; transports MTs with plus end leading; generates forces on MTs and sometimes between MTs and actin filaments relevant to several events, such as mitosis, cellular migration, axonal growth, and growth cone guidance	ALS; distal spinal and bulbar muscular atrophy
	Kinesin superfamily		No	Most family members transport cargo toward plus end of MTs; some transport cargo toward minus end; some are involved in generating forces on MTs during mitosis; one type suppresses MT assembly (see below)	Hereditary spastic paraplegia; Charcot–Marie–Tooth disease type 2; neurofibromatosis; fragile X syndrome; neurotropic viral disease
Cytoplasmic dynein-interacting proteins	Lis1		Yes	Modifies MT dynamics and dynein activity; interacts with doublecortin; plays roles in neuronal migration, neurogenesis	Lissencephaly
	Dynactin complex		No	Mediates interactions between cytoplasmic dynein and other structures; modify functions of cytoplasmic dynein; may also influence MT dynamics and interact with other proteins, such as some kinesins	ALS, Huntington's disease
MT-destabilizing proteins	SCG10		Yes	Suppress MT assembly; promote MT disassembly	Unknown
	Stathmin		No		
	Kinesin 13 superfamily	KIF2a	No		
		KIF2b	No		
		KIF2c	No		
MT-severing proteins	Katanin		No	Release MTs from the centrosome; regulate MT length; fragment MTs in discrete locales to increase MT number	Unknown
	Spastin		Yes		Hereditary spastic paraplegia
Cross-linking proteins	MT actin cross-linking factors		No	Cross-link cytoskeletal elements; anchor them to the plasma memberane	Unknown
	Plakins		No		
MT plus-end tracking proteins	End-binding proteins		No	Modify MT dynamics and regulate interaction of plus ends of MTs with other proteins and structures	Unknown
	Cytoplasmic linker proteins (CLIPs)		No		
	CLIP-associating proteins		No		

ALS, amyotrophic lateral sclerosis; FTDP, frontotemporal dementia with parkinsonism; STOP, stable tubule only protein.

microtubules were thought to be mainly structural. Today we know that microtubules are indeed structural in the neuron, but they are also highly dynamic and can interact with a vast array of regulatory proteins. They participate in a number of important functions, such as the transduction of forces necessary for neuronal migration and for the guidance of the growth cone toward its target. Microtubules are also essential for carrying out axonal and dendritic transport and for establishing many important features of neuronal polarity. Elucidating the mechanisms that regulate neuronal microtubules is essential for understanding the normal development of the nervous system, flaws in its development, and an array of debilitating neuropathies that affect the adult nervous system.

See also: Axonal Transport Disorders; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and Neurodegenerative Diseases; Microtubule Associated Proteins in Neurons; Slow Axonal Transport.

Further Reading

Akhmanova A and Hoogenraad CC (2005) Microtubule plus-end-tracking proteins: Mechanisms and functions. *Current Opinion in Cell Biology* 17(1): 47–54.

Baas PW, Karabay A, and Qiang L (2005) Microtubules cut and run. *Trends in Cell Biology* 15(10): 518–524.

Baas PW, Nadar CV, and Myers KA (2006) Axonal transport of microtubules: The long and short of it. *Traffic* 7(5): 490–498.

Baas PW and Qiang L (2005) Neuronal microtubules: When the MAP is the roadblock. *Trends in Cell Biology* 15(4): 183–187.

Brandt R, Hundelt M, and Shahani N (2005) Tau alteration and neuronal degeneration in tauopathies: Mechanisms and models. *Biochimica et Biophysica Acta* 1739(2–3): 331–354.

Galjart N (2005) CLIPs and CLASPs and cellular dynamics. *Nature Reviews Molecular Cell Biology* 6(6): 487–498.

Grenningloh G, Soehrman S, Bondallaz P, Ruchti E, and Cadas H (2004) Role of the microtubule destabilizing proteins SCG10 and stathmin in neuronal growth. *Journal of Neurobiology* 58 (1): 60–69.

Gressens P (2006) Pathogenesis of migration disorders. *Current Opinion in Neurology* 19(2): 135–140.

Hirokawa N and Takemura R (2004) Molecular motors in neuronal development, intracellular transport and diseases. *Current Opinion in Neurobiology* 14(5): 564–573.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews Neuroscience* 6(3): 201–214.

Kornack DR and Giger RJ (2005) Probing microtubule +TIPs: Regulation of axon branching. *Current Opinion in Neurobiology* 15(1): 58–66.

Levy JR and Holzbaur EL (2006) Cytoplasmic dynein/dynactin function and dysfunction in motor neurons. *International Journal of Developmental Neuroscience* 24(2–3): 103–111.

Morfini G, Pigino G, Beffert U, Busciglio J, and Brady ST (2002) Fast axonal transport misregulation and Alzheimer's disease. *Neuromolecular Medicine* 2(2): 89–99.

Myers KA, He Y, Hasaka TP, and Baas PW (2006) Microtubule transport in the axon: Re-thinking a potential role for the actin cytoskeleton. *Neuroscientist* 12(2): 107–118.

Vallee RB, Williams JC, Varma D, and Barnhart LE (2004) Dynein: An ancient motor protein involved in multiple modes of transport. *Journal of Neurobiology* 58(2): 189–200.

Midbrain Patterning

S Agarwala, The University of Texas at Austin, Austin, TX, USA
C W Ragsdale, University of Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

The midbrain, or mesencephalon, is defined by embryology. It is the middle of the three primary cerebral vesicles of the vertebrate neural tube, lying between the forebrain and hindbrain vesicles. The neural tube constriction formed at the junction of the midbrain and hindbrain vesicles is called the isthmus. This tissue harbors a major signaling center that patterns the adjoining midbrain and hindbrain. No comparable anterior signaling center has yet been identified at the junction between the midbrain and forebrain vesicles. By gene expression, however, the transition between dorsal midbrain and caudal forebrain is very abrupt, suggesting the presence of an important developmental boundary. Studies of cell lineage have provided evidence that both the isthmus and the dorsal midbrain–forebrain junction form barriers to cell mixing. Whether the midbrain vesicle constitutes a true developmental compartment, however, remains unresolved.

Pattern of the Adult Midbrain

The adult midbrain comprises a dorsal tectum (Latin for 'roof') and a ventral tegmentum (from 'tegmen,' which is Latin for 'cover'). In mammals, the cerebral peduncle, a fiber bundle of descending neocortical axons, is found beneath the tegmentum at the base of the midbrain. The tectum is dominated by a single structure, the optic tectum, or, in mammals, the superior colliculus. The optic tectum is a sensorimotor structure whose roles include a 'sentinel' function of orienting the head and eyes toward salient stimuli. It is organized into layers, with superficial layers receiving sensory input and deep layers containing premotor cells controlling eye and head movements. These sensory and motor layers contain sensory and motor maps that are in register across the layers and can be viewed as forming a coordinated representation of eye-centered external space when projected onto the tectal surface. The dominant sensory input to the optic tectum is from the retina. Retinal terminations form a coherent retinotopic map across the contralateral optic tectum, with the retinal temporal–nasal axis aligned with the tectal anterior–posterior axis and the retinal ventral–dorsal axis corresponding to the tectal medial–lateral axis. In addition to its intrinsic role in sensorimotor processing, the optic tectum serves as a relay for visual information traveling to the forebrain. The dorsal midbrain also contains an ascending relay for auditory and lateral line sensory information. This posterior midbrain structure is called the inferior colliculus in mammals, the mesencephalicus lateralis dorsalis nucleus in birds, and the torus semicicularis in other vertebrates ('torus' is Latin for 'bulge').

The ventral midbrain contains a complex constellation of nuclei, many of which are involved with motor system function. Among these are the red nucleus, which is part of the cerebellar motor system, and the nigral complex, a basal ganglia component that includes the dopamine neurons of the nigrostriatal and mesolimbic systems. The ventral midbrain also contains the most anterior motor neurons of the vertebrate nervous system. These are the somatic motor neurons of the oculomotor nucleus, which innervate extraocular muscles, and the visceral motor neurons of the Edinger–Westphal nucleus, which supply the ciliary parasympathetic ganglion. The oculomotor nucleus and the Edinger–Westphal nucleus are together called the oculomotor complex. Their motor axons travel to the eye's orbit in the third cranial nerve.

Early Midbrain Patterning: Boundaries and Signals

The regionalization of the brain into forebrain, midbrain, and hindbrain occurs early in gastrulation through the action of axial signals that organize the embryonic body plan. Once regionalized, the brain vesicles are patterned largely independently by local organizing centers. In the midbrain, the presence of two such organizing centers, the isthmic organizer (IsO) at the midbrain–hindbrain boundary and the rostral floor plate (rFP) of the ventral midbrain midline, is well established. The roof plate (RP), which runs along the dorsal midline of the midbrain, is also enriched in known developmental signaling molecules and is likely to serve as a midbrain signaling center.

Establishment of the Isthmic Organizer

Midbrain patterning depends critically on the establishment of the IsO. The IsO is initially formed anterior to the physical constriction of the midbrain–hindbrain junction. By gene expression, the IsO falls precisely at the posterior expression boundary of *OTX2*, a vertebrate homolog of the *Drosophila orthodenticle* gene, and the anterior limit of a second homeobox gene, *GBX2*, a vertebrate homolog of the *Drosophila* gene *unplugged* (Figure 1(a)). Gain- and loss-of-function

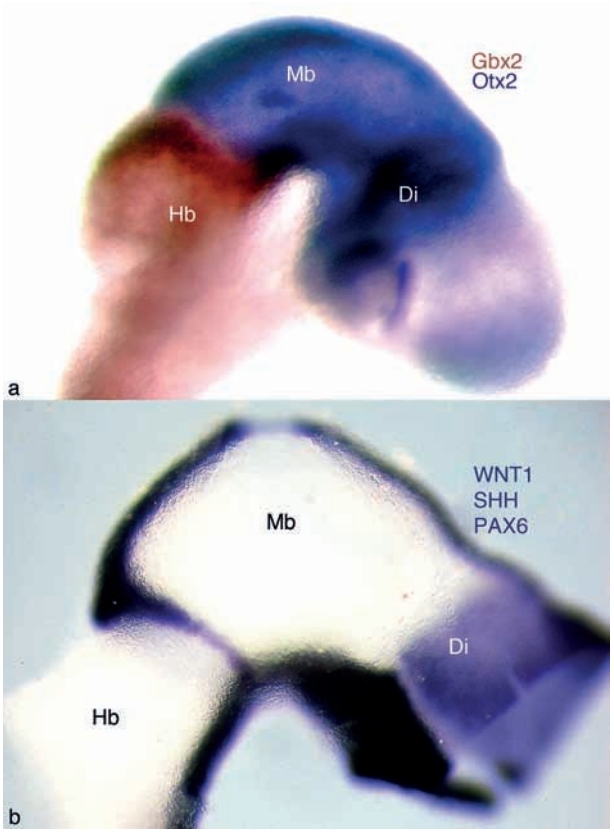

Figure 1 Identification of the major regional boundaries of the embryonic midbrain by gene expression for transcription factors and signaling molecules. (a) Plane of apposition of *Otx2* and *Gbx2* homeobox gene expression identifies the midbrain–hindbrain junction and the location of the isthmic organizer. Side view of an embryonic day 9 mouse prepared for whole mount *in situ* hybridization. Anterior is to the right. *Otx2* (blue) is expressed throughout midbrain and into forebrain, whereas *Gbx2* (brown) labeling is in hindbrain. (b) Polygon formed by gene expression markers of midbrain boundaries outlines the right midbrain of a chick embryo. Whole mount *in situ* hybridization of dissected stage 17 chick brain oriented with anterior to the right and probed for *WNT1*, *SHH*, and *PAX6* gene expression. The diencephalon–midbrain boundary is identified by *PAX6* expression in the diencephalon, the ventral midbrain midline is marked *SHH* labeling, and the isthmic organizer and the roof plate are demonstrated with *WNT1* gene expression. Di, diencephalons; Hb, hindbrain; Mb, midbrain.

experiments suggest that mutual repression between *OTX2* and *GBX2* is critical for the positioning and maintenance of the IsO.

A number of transcription factors are expressed within the midbrain and at the IsO and are required for the specification of the IsO and the midbrain. These include the paired box genes *PAX2/5/8*, the engrailed homeobox genes *EN1/2*, and the LIM homeobox gene *LMX1B*. In addition, WNT and FGF8 signaling molecules (WNT1 and FGF8/17/18) are expressed by the IsO (Figure 1(b)). The initial induction of these IsO genes is thought to depend on the mesendodermal tissues lying subjacent to the neural tube. Subsequently, within the neural tube, cross-regulation among *EN1/2*, *PAX2/5/8*, *WNT1*, and *FGF8* is required to establish and maintain IsO identity. Many of these genes are at first broadly expressed and progressively become restricted to partially overlapping but distinct domains of the isthmus, the posterior midbrain, and rhombomere 1 of the hindbrain. In mice, the loss of *Pax2/5/8*, *En1/2*, *Wnt1*, or *Lmx1b* function results in variable phenotypes depending on the genetic background. The most severe of these phenotypes is a partial to complete deletion of the IsO and the adjacent midbrain and anterior hindbrain.

Midbrain Patterning by FGF8

FGF8 at the IsO is expressed in a band around the anteriormost hindbrain immediately behind a ring of WNT1 expression. Transgenic mouse analyses, quail–chick chimeras, and FGF8 delivery by plasmid electroporation or ligand-soaked bead have demonstrated that FGF8 signaling is critical to midbrain patterning. In midbrain, FGF8 delivery shifts the isthmus forward and converts midbrain into hindbrain. In caudal diencephalon, FGF8 delivery induces an ectopic IsO (*EN1/2+*, *WNT1+*, and *FGF8+*) and the ectopic IsO converts the surrounding diencephalon into midbrain tissue (Figure 2(b)). This transformation can be quite dramatic, producing a full midbrain duplication in extreme cases. Unlike the native midbrain, the duplicated midbrain has its IsO at its anterior end, resulting in anterior–posterior patterning that is mirror image to that of the native midbrain. This induction of an ectopic midbrain cannot be produced throughout forebrain; it can only be elicited caudal to the zona limitans intrathalamica (zli), a signaling center that bisects the diencephalon. Evidence suggests that the transition from permissive (post-zli) to nonpermissive (pre-zli) tissue may be controlled by the expression domains of the iroquois homeobox gene *IRX3*, expressed posterior to the zli, and the sine oculis homeobox gene *SIX3*, expressed anterior to the zli.

The Forebrain–Midbrain Junction

Genetic fate mapping studies have shown that a lineage restriction boundary preventing the mixing of midbrain and forebrain cells is established at the diencephalon–midbrain boundary (DMB) by embryonic day (E) 9.5 in the mouse. Interestingly, this boundary may fall precisely along the midbrain–forebrain junction only dorsally. Ventrally it is shifted anteriorly into the subthalamic region. Interactions between the midbrain–hindbrain junction genes *EN/PAX2/FGF8* and the paired box homeobox gene

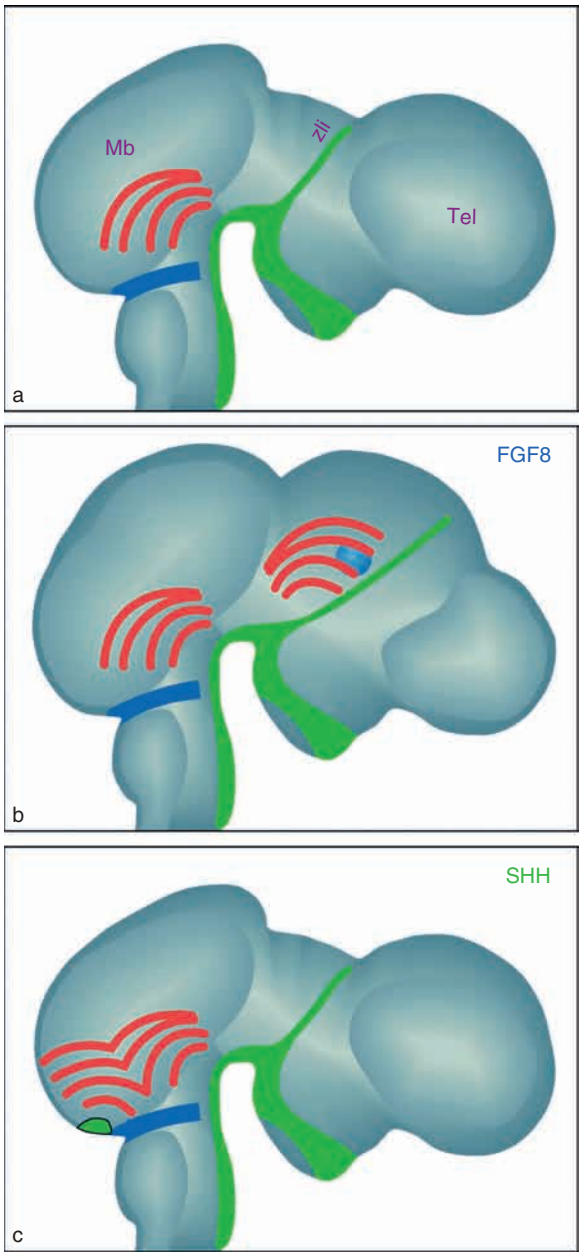

Figure 2 Regulation of midbrain pattern formation by FGF8 and Shh. (a) Chick brain at embryonic day 6, illustrating FGF8 expression at the midbrain–hindbrain boundary (blue) and Shh expression at the ventral midline and extending up the zona limitans intrathalamica (zli) in diencephalon (green). The anterior–posterior polarity of the midbrain is illustrated for ventral midbrain by a midbrain arcs schematic (red). (b) Exogenous FGF8 delivered to the diencephalon induces an ectopic isthmic organizer and a duplicate midbrain with a polarity mirror image to that of the native midbrain. (c) Ectopic Shh can ventralize dorsal midbrain and induce a duplicate set of midbrain arcs. Mb, midbrain; Tel, telencephalon.

PAX6, which is expressed in the diencephalon, have been shown to regulate DMB positioning in birds and fish. Thus, the establishment of the DMB critically depends on the prior specification of the IsO.

Induction of the Midbrain Floor Plate

The rFP and the underlying axial mesoderm (notochord and prechordal plate) express the signaling molecule Sonic hedgehog (Shh), a principal architect of ventral midbrain patterning. The origin of the floor plate and its mode of induction have long been a matter of controversy. A dual origin for spinal cord floor plate is indicated by zebra fish mutant analyses and quail–chick chimera experiments which show that the medial floor plate is derived from the node, expresses Shh, and helps induce lateral floor plate in adjoining neural tube. In the zebra fish, the specification of the medial floor plate depends on nodal signaling. Shh signaling is required for the maintenance but not the induction of the medial floor plate.

The rFP of the midbrain eventually forms a complex, wedge-shaped structure – one much broader than the floor plate of spinal cord and hindbrain. The rFP is likely to be induced by nodal and Shh signaling derived from a transient contact between the prechordal plate and the midbrain anlage. Subsequently, the notochord is briefly apposed to the ventral midbrain, but its rapid retraction strongly suggests that, unlike the arrangement in spinal cord in which notochord and floor plate collaborate on cell type specification, the rFP alone provides patterning signals for rFP expansion and the specification of ventral midbrain cell types. Fate mapping studies have shown that the rFP contains a mixed population of cells that at least in the chick incorporates a unique set of cells from a region anterior to the node called area a.

Roof Plate

The midbrain RP expresses the rich collections of WNT and BMP signaling molecules and ZIC and LMX transcription factors that are found in roof plate at other axial levels of the nervous system and that are thought to contribute to dorsal neural tube patterning. In midbrain, loss-of-function experiments for these molecules have only presented extreme phenotypes, either midbrain deletions or exencephaly, or no apparent patterning effects, possibly reflecting signaling redundancy. Consequently, the contributions of the midbrain RP to midbrain patterning remain obscure.

Interactions among Midbrain Organizers

Despite the complete orthogonality of the IsO and the rFP signaling centers, there is clear evidence for cross-regulation in their construction and maintenance. When Shh signaling is perturbed, the midbrain–hindbrain junction is disrupted, allowing mixing of midbrain and hindbrain cells. Moreover, in the mouse, loss of Shh

leads by E12.5 to a severe loss of *Wnt1* and a complete loss of *Fgf8* in the IsO. Interactions between the IsO and the RP include a direct contribution of midbrain–hindbrain boundary cells to the developing midbrain roof plate.

Ventral Midbrain Patterning

Medial Longitudinal Fasciculus

The first neurons to be born in ventral midbrain are part of a cohort of cells that straddle the midbrain–forebrain junction and contribute to the interstitial nucleus of Cajal and the nucleus of Darkschewitsch of the adult. The axons of these early born neurons travel posteriorly in a paramedian position and pioneer a major fiber tract that corresponds to the medial longitudinal fasciculus of the mature brain stem.

Midbrain Arcs

The embryonic midbrain is organized into a series of longitudinal territories arrayed bilateral to the ventral midline. These territories were first identified in the mantle layer of the chick embryonic midbrain with acetylcholinesterase histochemistry, which identifies postmitotic neurons in embryonic brain. Because these longitudinal territories constitute acetylcholinesterase-rich columns and have arcuate shapes, which are accentuated by the cephalic flexure, they are called midbrain arcs. The acetylcholinesterase-poor columns between the arcs also contain neurons and are referred to as interarcs. The arcs are numbered 1–5, from medial to lateral, and the interarcs are identified by their flanking arcs. Each arc and interarc expresses a unique signature of transcription factors and neurotransmitter-specific genes. For example, arcs 2 and 3 are enriched in the GABA synthetic enzyme *GAD2* and the zinc finger transcription factor *GATA2*, whereas the two-thirds interarcs expresses the transcription factor gene *PAX6* (Figure 3). Combined molecular and tract-tracing studies have shown a clear relationship between the first arc and nuclei of the adult midbrain. The first arc contains primordia for the oculomotor complex and the red nucleus. These primordia are separated within the first arc along the ventricular–pial axis, with the primordium of the oculomotor complex next to the ventricular layer and that of the red nucleus more pial. The possible nuclear fates of the lateral arcs and interarcs are not clear, in part because of the absence of selective markers for many adult ventral midbrain structures. Resolution of this important question will likely require genetic fate-mapping experiments.

The ventricular zone of ventral midbrain progenitor cells also contains arcuate territories. These territories are readily demonstrated as arcuate periodicities in the expression patterns of signaling molecules, including

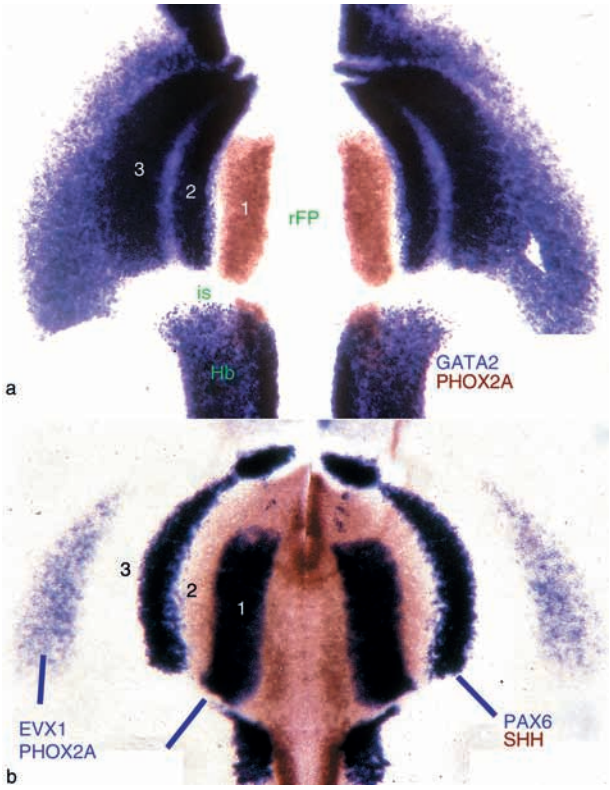

Figure 3 Arcuate organization of the embryonic midbrain tegmentum demonstrated by transcription factor gene expression. Illustrated are dissected flat mounts of embryonic day 5 chick brain stem prepared for *in situ* hybridization. The tissue documentation is arranged with the ventral midline oriented vertically, the midbrain arcs 1–3 identified on the left side only, and the isthmus (is) and hindbrain (hb) shown at the bottom of the panel. (a) *PHOX2A* labeling (brown) identifies the oculomotor complex of the most medial, or first, arc, whereas *GATA2* message enrichment (blue) distinguishes arcs 2 and 3. (b) Panel of three homeobox genes indicates regularly spaced arcuate territories (blue) corresponding to arc 1 (*PHOX2A*), interarc 2/3 (*PAX6*), and interarc 3/4 and arc 4 (*EVX1*). Gene expression for *SHH* (brown) illustrates the balloon-like expansion of the *SHH*-rich floor plate during midbrain development. rFP, rostral floor plate.

the WNT ligands WNT5A and WNT7A and ligands of the NOTCH receptor. The ventricular zone periodicities and the mantle layer arcs lie in radial registration, suggesting that midbrain arc patterning is set up in the ventricular zone and that arcs are populated by direct radial migration. Intracellular dye labeling studies demonstrating that radial clusters of coupled cells extend into the ventral midbrain mantle layer support this model.

Shh misexpression experiments have shown that the entire midbrain arc pattern, including the ventricular zone periodicities, can be generated by an ectopic source of Shh (Figure 2(c)). These ectopic arcs have precisely the same relative spatial positioning and the same molecular identity as the native arcs. The size

and shape of the ectopic arcs, however, are controlled by the size and shape of the ectopic Shh source. This type of pattern regulation is predicted by morphogen-based positional signaling and provides support for an 'action-at-a-distance' mode of Shh signaling in midbrain development. A specific requirement for Shh in the generation of the midbrain arcs and the specification of their constituent cell types has been demonstrated in loss-of-function experiments in chick and mouse.

At a morphological level, the midbrain arcs are the midbrain continuation of an embryonic columnar organization that runs from the spinal cord through brain stem to the subthalamic region of the caudal diencephalon. At a molecular level, however, the midbrain arc patterning mechanism shares only some features with those of the caudal central nervous system, and there is no coherent mapping that carries the molecular signature of the spinal cord domains V0–V1–V2–MN–V3 forward to the arc and interarc organization of the midbrain.

Oculomotor Complex

The midbrain oculomotor complex (OMC) is a derivative of the first arc. The earliest specific marker for these cells is the paired-like homeobox gene *PHOX2A*, which can be detected in still dividing ventricular zone precursor cells. In mice null for *Phox2a* function and in the zebra fish *phox2a* mutant *soulless*, the oculomotor nucleus and the nearby trochlear nucleus of anterior hindbrain are not generated. OMC neurons also express PHOX2B, which shares an identical homeodomain with PHOX2A, and the LIM homeobox gene *ISL1*, a pan-motor neuron marker, but they do not express the homeobox gene *HLXB9*, which is found apparently in all somatic motor neurons posterior to the trochlear nucleus. In this regard, the midbrain OMC is more akin to the branchial and visceral motor neurons of hindbrain than the somatic oculomotor neurons of the hindbrain abducens nucleus. Phox2b is required for the development of the hindbrain branchiovisceral motor neurons but not for the generation of OMC. Experiments in mouse in which *Phox2a* coding sequence is replaced by that of *Phox2b* demonstrate only a partial rescue, suggesting that the nonhomeodomain sequence differences between *Phox2a* and *Phox2b* are important for OMC development.

The human midbrain oculomotor system and its peripheral targeting are affected in many congenital cranial dysinnervation disorders. In one of these, CFEOM2 (congenital fibrosis of the extraocular muscle 2), patients present a bilateral ptosis (drooping eyelids) with eyes primarily in an exotropic (outward-looking) position. This presentation indicates that third cranial nerve motor function, which in humans includes innervation of the levator palpebrae superioris (upper eyelid muscle), is lost. Gene mapping studies have identified three distinct *PHOX2A* mutations in CFEOM2 patient families, at least one of which appears to be a null.

Red Nucleus

The red nucleus primordium is identified in the first arc by its specific expression of the POU homeobox gene *POU4F1/BRN3A*. Gene targeting studies have demonstrated that the absence of *Pou4f1* gene function leads to loss of the red nucleus by birth. In mice mutant for the homeobox gene *Emx2*, the *Pou4f1*-positive neurons of the red nucleus are also generated, but they are lost earlier – by midgestation.

Midbrain Dopamine Neurons

Midbrain dopamine neurons are born in ventromedial midbrain within the Shh-expressing ventricular zone of the rFP. Specific sites of origin likely include the ventral midline and the ventricular zone overlying the first arc. Some dopamine neurons remain medially to form the A10 complex of the ventral tegmental area. The A9 and A8 dopamine cell groups of the substantia nigra and retrorubral area are populated by a reelin-dependent lateral migration. Studies of transcription factors involved in midbrain dopamine development have identified a large number of genes that regulate distinct aspects of dopamine neuron production, differentiation, and survival. These factors include a suite of homeodomain transcription factors, En1/2 and Lmx1a/b, that are also expressed in the IsO or rFP, and the nuclear receptor Nr4a2/Nurr1, which is required for expression of the dopamine synthetic enzyme tyrosine hydroxylase. Particularly notable among these factors is the paired-like homeobox gene *PITX3*, which is a uniquely specific marker of apparently all midbrain dopamine neurons. In mice lacking *Pitx3* gene function, there is a severe reduction in midbrain dopamine cell generation that particularly affects the A9 dopamine cell group.

Nuclei Not of Midbrain Origin

Classical and modern embryology have shown that some nuclei found at the midbrain–hindbrain junction in the adult brain are clearly of hindbrain origin. These include the trochlear nucleus of the fourth cranial nerve and the serotonin cells of the raphe system. The cholinergic isthmic nucleus of birds, which maintains reciprocal connections with the optic tectum, is also generated in hindbrain. A hindbrain origin is therefore possible for its predicted mammalian homolog, the parabigeminal nucleus.

Dorsal Midbrain Patterning

The tectal–tegmental boundary is readily demonstrated by many developmental control genes that mark dorsal midbrain selectively, including the paired box homeobox genes *PAX3/7*, the homeobox gene *DBX1*, and the nuclear receptor *NR2E1*. The possibility that this sharp tectal–tegmental boundary may form a cell lineage boundary has not been addressed, but tissue culture experiments have shown that ventral cells progressively lose their capacity to intermix with dorsal cells. As is true for the dorsal spinal cord, dorsal midbrain identity can be switched to a ventral character by ectopic expression of Shh. Unlike the dorsal spinal cord, however, dorsal midbrain growth and specification are dependent on ventral midbrain Shh signaling. In mice lacking *Shh* or the Shh effector *Smo*, the dorsal midbrain is severely reduced in size and an early molecular marker of dorsal midbrain identity, *Dbx1* gene expression, is lost. The molecular mechanism for this Shh-dependent effect is unknown, but because of the distances involved, it seems likely to involve intermediary signals. One candidate intermediary is *Fgf15*, which is expressed in embryonic dorsal midbrain and is largely lost from this tissue in the *Shh* mutant mouse.

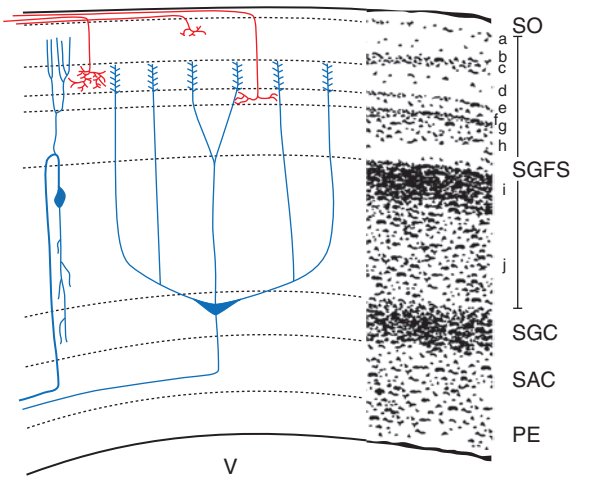

Figure 4 Cellular organization of the chick optic tectum. Cytoarchitects distinguish up to 16 tectal layers between the pial surface (top) and the midbrain ventricle (V). The layering of the optic tectum serves to organize the cell bodies, the dendritic arborizations, and the inputs of the tectal neurons. Illustrated in red are three examples of retinotectal axons that target specific lamina within the retinorecipient layers a–f. Illustrated in blue are two examples of tectal projection cells – one in stratum griseum centrale (SGC) that has 'bottlebrush' dendritic endings in layer d and issues ascending projections to the thalamus, and a second in layer i that projects to the isthmic nuclear complex. PE, the periventricular and ependymal layers; SAC, stratum album centrale; SGFS, stratum griseum and fibrosum superficialis; SO, stratum opticum.

Mesencephalic Trigeminal Nucleus

The mesencephalic trigeminal nucleus is found in all jawed vertebrates and contains primary sensory neurons for jaw muscle proprioception. Its constituent neurons are the first to differentiate in the midbrain and issue descending axons that on leaving midbrain pioneer the lateral longitudinal fasciculus of the hindbrain. Whether these primary sensory neurons are generated directly in brain or are returnees of neural crest origin has been difficult to resolve with experimental embryological methods. Molecular studies using neural crest markers and genetic fate mapping support a central origin for these cells.

Optic Tectum

Proper tectal histogenesis entails the generation of a layered structure that shares many features with cortex, including layer-specific cell types with apically oriented dendrites and topographically organized inputs that target particular layers (Figure 4). The cellular mechanisms for constructing the optic tectum, however, parallel only in part those found in cortex. The deepest layers of the optic tectum, including the output neurons of the stratum griseum centrale, are generated first. The next born neurons, however, do not follow the simple neocortical 'inside-out' pattern but, rather, are produced in a much more complex manner. The cells of the prospective stratum griseum and fibrosum superficialis layers, which are designated the tectal plate in development, are born in two major waves, with superficial layers a–g produced first and the deeper layers h–j born later. In addition, within both waves, there are further temporal gradients, with layers a–g following an inside-out pattern of birthing and layers h–j observing an 'outside-in' sequence. Retroviral lineage analysis has shown that the tectal plate is principally populated by clonally related cells traveling from the ventricular zone along radial glia. However, there are also early and late tangential migrations. Some of these tangential migrations give rise to both neurons and glia.

One role of tectal lamination is to segregate inputs to particular dendritic territories. Focusing on the retinal input layers a–f, anatomical mapping studies have identified a large number of potential cell surface cues that are expressed in specific retinorecipient laminae and could contribute to retinal axon layer targeting. These include adhesion molecules of the cadherin family and the extracellular matrix molecule versican, a member of the aggrecan family of chondroitin sulfate proteoglycans. Function-blocking experiments employing an antibody to N-cadherin or a lectin that binds to versican have demonstrated lamina-specific defects in retinal axon arborization. Inappropriate laminar targeting by incoming retinal

fibers has also been found following misexpression of the Groucho family corepressor TLE4 in the embryonic chick tectum. The TLE4 defect is likely secondary to a disruption of tectal cell-type specification which also causes lamina formation defects.

The most studied of the tectal input topographies is that of the retina. It now appears almost certain that retinotectal topography is set up by multiple independent batteries of guidance molecules acting along orthogonal axes. These guidance ligands are found in increasing gradients across the ventral-to-dorsal and anterior-to-posterior axes of the tectum, and their receptors are found in gradients across the retina. Current models of map formation are that each ligand has attractive and repulsive actions and that retinal axons distribute themselves across the tectal axes based on the attractive–repulsive set point conferred by their retinal position-dependent receptor status. The retinotectal ligand–receptor combinations identified to date are: (1) ephrin-As and EphA receptor tyrosine kinases, (2) RGM and neogenin receptors for the tectal anterior–posterior axis, (3) ephrin-Bs and EphB receptors, and (4) Wnt3 and the WNT receptors Ryk and Frizzled-5 for the tectal dorsal–ventral axis. In addition to these specific guidance molecules, direct axon competition for tectal target space is thought to regulate retinotopic map development, although the specific molecular mechanisms involved in this process remain obscure. Finally, neuronal spiking activity is needed for refinement of the retinal topographic map. Activity-dependent mechanisms are also likely to mediate sensory map registration across the tectal layers, integrating input from the isthmic nuclear complex and from nonvisual sensory structures.

The axon guidance gradients found in the tectal anterior–posterior axis are set up by the IsO. The 'naive' model is that the IsO releases FGF8, which forms an instructive, posterior high concentration gradient that is converted into a cellularized expression gradient of transcription factors (such as En1/2) and target ligands (such as eprhin-A2 and ephrin-A5). A striking modification of this model derives from the finding that soluble En2 homeodomain protein can act directly to guide retinal axons, repelling temporal axons and attracting nasal ones, which accords precisely with the posterior high gradient in EN2 expression. For the tectal dorsal–ventral gradients, it is anticipated that the RP plays a major instructive role, and there is direct evidence that rFP contributes to retinotectal dorsal–ventral patterning as well.

Midbrain Patterning and Invertebrates

Midbrain is a feature of vertebrates, and its possible status in invertebrates is uncertain. In the cephalochordate *Branchiostoma*, for example, even the presence of a tectum-like structure is disputed and the best guide for a possible midbrain is the positioning of the anteriormost motor neurons. Modern molecular developmental biology, however, has identified regional markers of anterior–posterior identity that allow informative comparisons to be made within and across phyla. For assessing the origins of midbrain, a potentially informative marker is the paired-like homeobox gene *DMBX1* (diencephalons/mesencephalon homeobox gene). *DMBX1* expression onset precedes the establishment of the IsO. It is first detected during gastrulation as a crescent in the anterior neural plate and then accumulates at the level of the prospective midbrain and, with the appearance of definitive brain vesicles, is strongly expressed in the embryonic midbrain and adjoining posterior diencephalon. Unfortunately, the functional importance of DMBX1 to midbrain development is unclear; whereas *dmbx1* zebra fish morphants show reduced tectal growth, *Dmbx1*-deficient mice present no gross abnormalities in brain development. Moreover, as development proceeds, *DMBX1* is found in other central nervous system sites including the hindbrain. Despite these handicaps, DMBX1 is the single best molecular marker of embryonic midbrain tissue.

A *DMBX1* ortholog has not been described in protostomes. However, comparative genomics has readily identified *DMBX* orthologs throughout invertebrate deuterostomes, including cephalochordates, urochordates, echinoderms, and hemichordates. The finding of one *Hydra* and six *Nematostella* DMBX sequences indicates that this gene was present in the cnidarian–bilateralian ancestor. Studies of *dmbx* gene expression in the hemichordate acorn worm *Saccoglossus* provide strong support for the conclusion that midbrain-like regional gene expression and isthmus-like organizer activity were present in at least the deuterostome ancestor. The midlevel structures of the acorn worm, the mesosome and the anterior metasome to the level of the first gill slit, express makers characteristic of midbrain and caudal forebrain, including *dmbx*, *dbx*, *otx*, *engrailed*, and *pax2*. In addition, the posterior end of this region is marked by a *fgf8*- and *wnt1*-rich signaling center. By contrast, transcription factor markers characteristic of anterior and ventral forebrain are found anteriorly, in the prosome, and *hox* genes are expressed only posterior to the first gill slit. This finding of a conserved core of midbrain patterning circuitry does not mean that the shared ancestor of vertebrates and hemichordates had a midbrain. The predicted nervous system of the deuterostome ancestor, like that of the acorn worm, is an epidermal nerve net.

See also: Forebrain Development: Prosomere Model; Forebrain Development: Holoprosencephaly (HPE); Forebrain: Early Development; Neural Patterning: Midbrain–Hindbrain Boundary; Optic Tectum: Sensorimotor Integration; Segmentation: Segmental Boundaries, Establishment and Morphogenesis (EPH); Wnt Pathway and Neural Patterning.

Further Reading

Agarwala S and Ragsdale CW (2002) A role for midbrain arcs in nucleogenesis. *Development* 129: 5779–5788.

Agarwala S, Sanders TA, and Ragsdale CW (2001) Sonic hedgehog control of size and shape in midbrain pattern formation. *Science* 291: 2147–2150.

Crossley PH, Martinez S, and Martin GR (1996) Midbrain development induced by FGF8 in the chick embryo. *Nature* 380: 66–68.

Engle EC (2006) The genetic basis of complex strabismus. *Pediatric Research* 59: 343–348.

Flanagan JG (2006) Neural map specification by gradients. *Current Opinion in Neurobiology* 16: 59–66.

Gerhart J (2006) The deuterostome ancestor. *Journal of Cellular Physiology* 209: 677–685.

Hunter E, Begbie J, Mason I, and Graham A (2001) Early development of the mesencephalic trigeminal nucleus. *Developmental Dynamics* 222: 484–493.

Inoue A and Sanes JR (1997) Lamina-specific connectivity in the brain: Regulation by N-cadherin, neurotrophins, and glycoconjugates. *Science* 276: 1428–1431.

Millet S, Bloch-Gallego E, Simeone A, and Alvarado-Mallart RM (1996) The caudal limit of Otx2 gene expression as a marker of the midbrain/hindbrain boundary: A study using *in situ* hybridisation and chick/quail homotopic grafts. *Development* 122: 3785–3797.

Nakamura H and Sugiyama S (2004) Polarity and laminar formation of the optic tectum in relation to retinal projection. *Journal of Neurobiology* 59: 48–56.

Rhinn M, Picker A, and Brand M (2006) Global and local mechanisms of forebrain and midbrain patterning. *Current Opinion in Neurobiology* 16: 5–12.

Ruthazer ES and Cline HT (2004) Insights into activity-dependent map formation from the retinotectal system: A middle-of-the-brain perspective. *Journal of Neurobiology* 59: 134–146.

Sanders TA, Lumsden A, and Ragsdale CW (2002) Arcuate plan of chick midbrain development. *Journal of Neuroscience* 22: 10742–10750.

Smidt MP and Burbach JP (2007) How to make a mesodiencephalic dopaminergic neuron. *Nature Reviews Neuroscience* 8: 21–32.

Zervas M, Millet S, Ahn S, and Joyner AL (2004) Cell behaviors and genetic lineages of the mesencephalon and rhombomere 1. *Neuron* 43: 345–357.

Mitochondrial Dysfunction in Nervous System Injury

D A Butterfield and W O Opii, University of Kentucky, Lexington, KY, USA

© 2009 Elsevier Ltd. All rights reserved.

Mitochondria Structure

Mitochondria are organelles that evolved from a symbiotic relationship between aerobic bacteria and primordial eukaryotic cells. Despite being in eukaryotic organisms, mitochondria still carry a functional relic of their original genome. They are intracellular tubular organelles delimited by two membranes. The outer mitochondrial membrane is permeable to ions and small proteins of molecular weights <10 kDa, whereas the inner mitochondrial membrane houses the multimeric enzyme complexes of the electron transport chain (ETC). Human mitochondria DNA is a circular double-stranded molecule of approximately 16.6 kb in length with the genome containing 37 genes. Of these, 13 genes encode protein subunits of respiratory chain complexes, with a significant amount of mitochondrial proteins being encoded by the nuclear genome and transported into the mitochondria. For example, complex II is solely composed of proteins encoded by nuclear genes. The varying mitochondrial density in various tissues is usually dictated by the need for oxidative phosphorylation and energy production. Cardiac and skeletal muscles and neurons have the highest density of mitochondria, hence their sensitivity to energy-dependent defects resulting from mitochondrial dysfunctions following injury.

Mitochondria Electron Transport Chain, Membrane Potential, and Energy Metabolism

Central to the role of mitochondria is energy metabolism through the production of ATP. This energy production is usually carried out at the level of the inner mitochondrial membrane through oxidative phosphorylation involving the reduction of oxygen to water by multimeric enzyme complexes. The ETC is composed of NADH–ubiquinone oxidoreductase (complex I), succinate dehydrogenase–CoQ oxidoreductase (complex II), cytochrome reductase (complex III), cytochrome oxidase (complex IV), and ATP synthase, which is sometimes referred to as complex V. These multimeric complexes are arranged in the inner mitochondrial membrane according to their reduction potentials. During oxidative phosphorylation, ETC complexes are usually involved in the reduction and oxidation reactions through reducing equivalents such as NADH or succinate, which are transported into the mitochondria from the cytosol. In the process, protons are pumped from the matrix into the intermembrane space, terminating with the reduction of O_2 to H_2O. The transfer of protons from the matrix to the inner mitochondrial membrane leads to the generation of a mitochondrial membrane potential ($\Delta\Psi_m$) of 150–180 mV, which usually determines the energetic status of the mitochondria. This reserve of potential energy through the electrochemical gradient is then coupled to the generation of ATP from ADP and inorganic phosphate P_i through the ATP synthase complex (complex V).

Ca^{2+} Regulation and Excitotoxicity

Mitochondria under normal conditions usually act as high-capacity Ca^{2+} sinks through a highly complex system for the regulation and the transportation of Ca^{2+}. Through these processes, mitochondria sense and respond to changes in cytosolic Ca^{2+} loads to maintain cellular Ca^{2+} homeostasis that is required for normal neuronal function. Disruptions in Ca^{2+} homeostasis are one of the major factors that contribute to the neuropathology following central nervous system (CNS) injury. It has been shown that following CNS injury, mitochondrial dysfunction is primarily involved in glutamate neurotoxicity. Glutamate excitotoxicity is usually triggered by the excessive influx of calcium following activation and overstimulation of the *N*-methyl-D-aspartate (NMDA) subtype of glutamate receptors. The resulting increase in intracellular calcium has been shown to lead to impaired function of the mitochondrial ETC and formation of reactive oxygen species/reactive nitrogen species (ROS/RNS), particularly superoxide anion and hydrogen peroxide. In addition, as mentioned previously, extreme loads of Ca^{2+} have been shown to lead to the opening of the mitochondrial permeability transition pore (mPTP) and induction of the apoptotic pathways, eventually leading to neuronal cell death following CNS trauma.

Mitochondria and Oxidative Stress

Oxidative stress can be defined as an imbalance in the antioxidant defense mechanisms and the amount of ROS/RNS produced. The mitochondria are a key source of ROS. Through the generation of ATP via the oxidative phosphorylation by the ETC components, there is always some leakage of electrons (1% or 2%) from the mitochondria resulting in the-partial reduction of O_2 to form superoxide anion O_2^-, a majority of which is generated from complex I

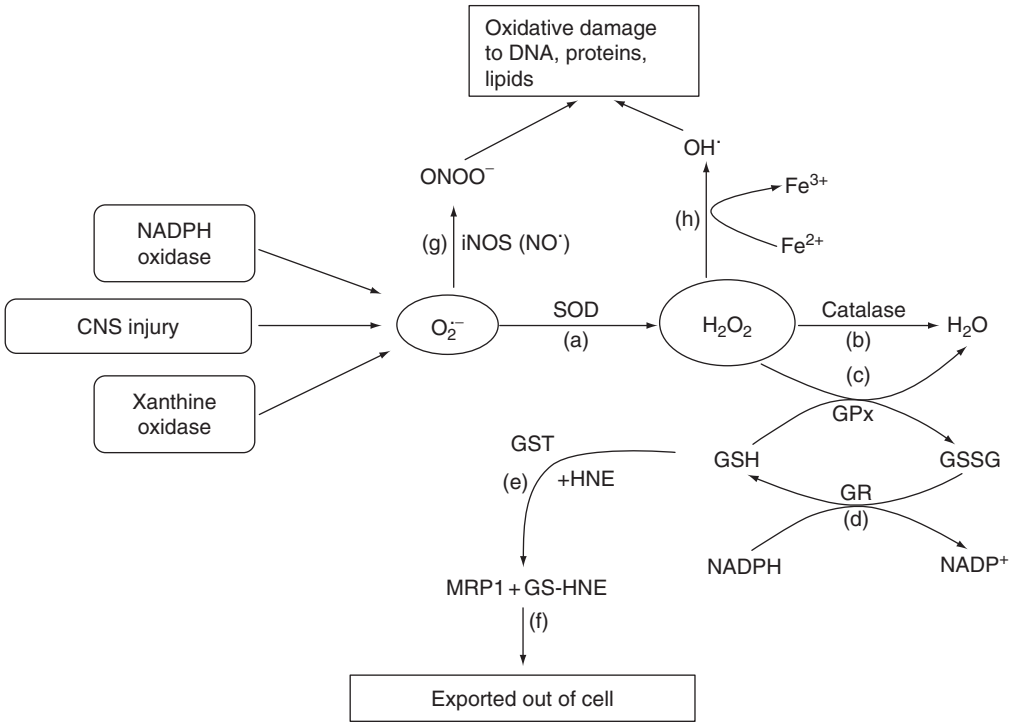

Figure 1 A summary of the production of reactive oxygen and nitrogen species (ROS/RNS) following CNS injury. In addition, the figure displays the antioxidant systems that provide protection against increased oxidative stress as seen in nervous system injury.

and complex III of the ETC. The production of superoxide anion can occur when the ETC is impeded following mitochondrial injury. Note that the production of ROS by the mitochondria is tightly linked to the mitochondrial membrane potential since hyperpolarization or high mitochondrial membrane potential promotes the production of ROS. With a high mitochondrial membrane potential, the ETC cannot transfer protons against the electrochemical proton gradient from the matrix. This, in essence, leads to a reduced state of electron carriers and an increase in the half-life of semiquinone. The intermediates thus remain in a prolonged reduced state, thereby increasing the chances of electron leakage and partial reduction of O_2 to superoxide anion O_2^-. As discussed later, it is on this basis that mitochondrial protein uncouplers are gaining importance as possible therapeutics for CNS injuries.

As noted previously, one major pathway predominantly involved in the production of ROS/RNS is the secondary activation of glutamate receptors and excitotoxicity following injury to the nervous system. Following CNS injury, the cytosolic levels of Ca^{2+} increase and are sequestered into the mitochondrial matrix through various transporters, as stated previously. However, excessive calcium loads result in the dysregulation of mitochondrial Ca^{2+} homeostasis, leading to the activation of nitric oxide synthase (iNOS) and the production of nitric oxide radical ($NO^{\cdot}$). The presence of nitric oxide NO and superoxide anion O_2^-, as seen in Figure 1(g), leads to the formation of peroxynitrite $ONOO^-$, another significant RNS, which can lead to the oxidative modification of proteins through nitration. Elevated intracellular Ca^{2+} levels also activate phospholipase A_2; this enzyme liberates the unsaturated fatty acid arachidonic acid, initiating the formation of free radicals via the cyclooxygenase-2 and lipoxygenase pathways. Arachidonic acid can also be oxidized, leading to the formation of lipid peroxidation products such as 4-hydroxynonenal (HNE), malodialdehyde (MDA), and acrolein. The activation of inflammatory response following CNS injury also leads to the production of substances such as cytokines, tumor necrosis factor-α, and interleukin-1β that can damage neurons. In addition, activation of microglia and astrocytes can also lead to the generation of more ROS. For example, NADPH and xanthine oxidase significantly contribute to O_2^- production upon activation of microglia. If the significant increase in ROS and RNS following CNS injury is not controlled, damage to DNA, proteins, or lipids would result, eventually leading to loss of function and possibly neuronal cell death.

Mitochondria Permeability Transition Pore

The mitochondrial permeability transition (mPT) is defined as an increase in the permeability of the inner mitochondrial membrane to solutes that have molecular masses lower than 1.5 kDa. The induction of this transition has since been demonstrated to involve the opening of a 2- or 3-mm inner membrane megachannel referred to as the mPTP. This pore is a voltage-dependent channel that is activated by the elevation of Ca^{2+} levels, oxidative stress, and low inner mitochondrial membrane potentials as observed in most nervous system injuries. The exact identity of this megachannel is still under investigation. However, progress has been made in trying to provide a logical model for its composition. The minimum configuration of the mPTP requires the adenine nucleotide translocator (ANT), which is located in the inner mitochondrial membrane, in association with outer mitochondrial membrane proteins such as the voltage-dependent anion channel and cyclophilin D (a peptidyl–prolyl *cis–trans* isomerase whose activity is usually inhibited by immunosuppressant cyclosporine A (CsA)). In addition, other proteins, such as Bcl-2 and Bax, have been shown to interact with ANT, providing a regulatory role for mPTP. The exact physiological function of mPTP has not been established; however, there seems to be a clear consensus on its involvement in the induction of permeability transition in the neuropathology of various CNS injuries, such as traumatic brain injury (TBI) and spinal cord injury (SCI). This is particularly evident since it has been established that CNS injury triggers the opening of mPTP, resulting in a collapse of the electrochemical gradient, the swelling of mitochondria, the inhibition of ATP synthesis, the release of apoptotic factors, and the eventual initiation of cell death pathways that are blocked by the use of the cyclophilin D binding immunosuppressant drug CsA.

Mitochondria Apoptosis and Necrosis

Mitochondria also play a central role in the regulation of both programmed cell death (also known as apoptosis) and uncontrolled cell death (i.e., necrosis). Evidence of apoptosis following CNS trauma (discussed later) has been well established. As mentioned previously, events leading toward CNS injury usually occur in a biphasic manner – that is, through a primary necrotic process that results in rapid and significant decline in the levels of ATP and in the ensuing hours, days, and weeks, a secondary injury that exacerbates the primary insult occurs through apoptosis. These two mechanisms of neurodegeneration differ in their morphological features of general cell structure. Apoptosis is marked by cell shrinkage and DNA fragmentation, whereas necrosis is marked by cellular swelling and uncontrolled cell lysis.

As noted previously, following CNS injury, there is an increase in oxidative stress and excitotoxicity leading to Ca^{2+} overload, which in turn triggers the induction of the permeability transition. This usually leads to excessive mitochondrial swelling and depolarization of the mitochondria and uncoupling of the oxidative phosphorylation, eventually leading to a rapid decline in ATP reserves which if not restrained, as seen in severe CNS injuries, would more often than not lead to necrotic cell death. However, it has been shown that transient opening of the mPTP is associated with the induction of apoptosis through the release of cytochrome c and other pro-apoptotic proteins. Released cytochrome c usually complexes with apoptosis protease-activating factor 1 (Apaf-1), dATP, and procaspase-9 to form the high-molecular-weight complex, the apoptosome. The apoptosome then activates downstream effector caspases, which then initiate the apoptotic cascade. However, apoptotic events can still occur without the opening of the permeability transition pore. Pro-apoptotic proteins, such as Bax and Bad, also have been thought to be capable of forming pores in the outer mitochondrial membrane, thereby influencing the release of other pro-apoptotic proteins such as cytochrome c, apoptosis inducing factor, endonuclease G, and Smac/Diablo, thereby activating the apoptotic cascade. However, their role in CNS injury is still not well established. These various pathways have been summarized in Figure 2.

Mitochondria and Antioxidant Defense System

Cells are equipped with efficient strategies to control and maintain the intracellular levels of ROS/RNS and antioxidants. As mentioned previously, following CNS injury there is extensive production of ROS/RNS through a wide range of mechanisms. If these levels are not controlled, these ROS/RNS could lead to oxidative damage to key mitochondrial proteins or enzymes, DNA, and lipids, leading to a loss in function of the mitochondria, decline in ATP production, and possibly eventual cell death as observed in most events following CNS injuries. As a result, mitochondria have developed a defense system to protect themselves and the cell from the possible deleterious effects of these ROS/RNS through the use of key enzymes and proteins (discussed later). The key antioxidant found in the mitochondria is manganese superoxide dismutase (MnSOD), with the glutathione and thioredoxin antioxidant systems playing key roles in the

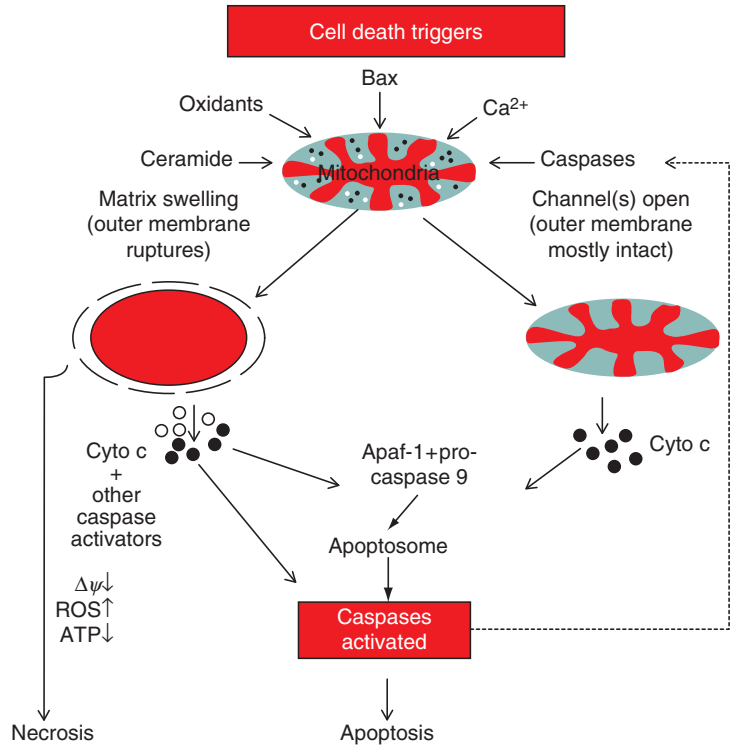

Figure 2 A summary of the molecular pathways and downstream cascades of apoptosis and necrosis following injury to the CNS. Adapted from Green DR and Reed JC (1998) Mitochondria and apoptosis. *Science* 281(5381): 1309–1312.

regulation of ROS/RNS. Figure 1 provides a summary of the key pathways of antioxidant systems which are discussed here with emphasis on CNS injury.

As a first line of defense following CNS injury and against production of ROS, the mitochondria use the MnSOD (Figure 1(a)). There are two classes of SOD; manganese SOD (Mn-SOD), which is predominately located in the mitochondrial matrix, and copper/zinc SOD (Cu/Zn SOD), which is predominately found in the cytosol, although studies have shown that Cu/Zn SOD can also be located in the intermembrane space of the mitochondria. The SOD enzymes catalyze the disproportionation of O_2^- to H_2O_2 and H_2O, thus eliminating the cytotoxic superoxide anions from the cell. However, the disproportionation of O_2^- generates H_2O_2, another ROS. H_2O_2 also can lead to oxidative damage through various mechanisms. In the presence of transition metal ions such as Fe^{2+} or Cu^+, H_2O_2 can undergo a Fenton reaction resulting in the generation of hydroxyl radicals $OH^{\cdot}$, a much more toxic ROS that can exact excessive damage at distant sites from where H_2O_2 was generated. As a result, there is need for the cell to regulate intracellular levels of H_2O_2. One immediate mechanism through which the intracellular levels of H_2O_2 are regulated is via the enzyme catalase (Figure 1(b)), which catalyzes the conversion of H_2O_2 to H_2O and dioxygen. However, catalase is not found in brain mitochondria; as a result, CNS mitochondria predominately depend on other peroxidases to regulate the levels of H_2O_2. One mechanism involves the glutathione (GSH) system in conjunction with glutathione peroxidase (GPx) and glutathione reductase (GR), as shown in Figures 1(c) and 1(d). GSH is usually synthesized in the cytosol but is then transported into the mitochondria through various mechanisms. H_2O_2 can be converted to water by GPx (Figure 1(c)) using GSH as the hydrogen donor. The oxidized glutathione (GSSG) is then recycled back to GSH by the-enzyme GR (Figure 1(d)), which uses NADPH as a cofactor. Another antioxidant system associated with glutathione is through the enzyme glutathione-*S*-transferase (GST). As mentioned previously, following CNS injury, there is a significant production of reactive aldehydes such as HNE or MDA which, if not eliminated from the cell, could lead to additional damage. The levels of these reactive aldehydes are usually regulated by the enzyme GST, which catalyzes their conjugation to GSH as shown in Figure 1(e). With the aid of multidrug-resistant protein-1 (MRP1), this GS–alkenal conjugate is exported out of the cell, removing toxic reactive aldehydes (Figure 1(f)). Peroxiredoxins are another important H_2O_2-removal antioxidant system. Peroxiredoxins belong to a family of enzymes that are homodimers and require no prosthetic groups. They use thioredoxin-2 to reduce H_2O_2

and are dependent on cysteine at the active sites. These antioxidant defense mechanisms within the mitochondria, if maintained well, are sufficient to protect against ROS/RNS. However, when they are perturbed, as in the case with most CNS injuries, the excess levels of ROS can lead to mitochondrial dysfunction, a decline in ATP levels, and eventual activation of necrotic and/or apoptotic pathways, as discussed previously. As a result, modulation and maintenance of antioxidant levels through the use of exogenous antioxidant compounds has been considered as a potential therapeutic intervention following CNS injury.

Mitochondria and Traumatic Brain Injury

TBI is a serious health care problem in the United States, with more than 400 000 individuals hospitalized each year and an estimated annual cost of $25 billion. As noted previously, for most CNS injuries the pathophysiological events in TBI occur in a biphasic manner. Following TBI, the primary insult results in a rapid and significant necrosis of cortical tissue at the site of injury, and in the ensuing period (days and weeks) a secondary injury occurs that exacerbates the primary damage. One underlying feature of this secondary injury in both experimental and clinical TBI is the loss in mitochondrial bioenergetics. This loss is usually characterized by mitochondrial dysfunction that includes exposure of neurons to excitotoxic levels of excitatory neurotransmitters with excessive uptake of Ca^{2+} and eventual overload, generation of reactive oxygen species, induction of the opening of mPTP, release of cytochrome c, inhibition of ATP production, and ultimately neuronal cell death. Following TBI, ROS/RNS can also be generated through other cellular pathways, such as calcium activation of phospholipases, iNOS xanthine oxidase, the Fenton and Haber–Weiss reactions, and by inflammatory cells, as previously discussed. It has been shown that following TBI there is a significant loss in mitochondria bioenergetics, possibly through the loss of activity of key enzymes. It has also been shown that following experimental TBI, there is increased oxidative damage in mitochondria. Using proteomics, it has since been shown that key mitochondrial-related proteins are oxidatively modified, contributing to the loss in mitochondrial-related bioenergetics. There is currently no particularly effective therapeutic intervention for TBI; thus, TBI still presents clinicians and researchers with a major challenge.

Mitochondria and Spinal Cord Injury

Spinal cord injuries are the leading causes of disability and mortality in the United States, with the estimated incidence rate of 43–55 per million annually. The primary cause of SCI is automobile collisions, followed by violence, falls, and injuries in sporting events. Like TBI, SCI also undergoes multiple secondary injuries following the initial or primary impact. SCI results in hemorrhage, edema, axonal demyelination, increased glial activity, and the eventual compromise and loss of neurological function. The pathophysiology of SCI is characterized by an immediate and irreversible primary insult that is composed of physical and mechanical trauma. The generation of sharp bone fragments, contusions, compression, and eventual concussion are often observed. This damage is more manifested as ischemia through continuous compression after injury. The secondary injury, which is also considered as the more severe, involves a highly complex cascade of molecular downstream events that range from early neuronal apoptosis at the site of injury to the degeneration of white matter tracts weeks after injury among many other pathologies. Although there are a significant number of factors that contribute to the events seen in the secondary injury during SCI, such as ischemia, edema, ionic imbalances, compromised energy metabolism, and biochemical changes, mitochondria play a significant role in the neuropathology of SCI. Following SCI, there is a sustained activation of glutamate receptors, which leads to the intracellular accumulation of high levels of Ca^{2+} and a rapid impairment of mitochondrial function. This glutamate-mediated excitotoxicity, the formation of ROS, and lipid peroxidation are prominent events thought to contribute to neuronal dysfunction and cell loss following traumatic damage and ischemic injury to the CNS. Evidence for the involvement of mitochondria in these mechanisms stems from the fact that there is impaired mitochondrial function and lipid peroxidation within 1 h following experimental SCI, with an increase in ROS or evidence of oxidative damage occurring at 4 h and 24 h and as long as 4 weeks following injury. In addition, there were changes in the levels of ROS scavenger molecules 24 h following injury, as characterized by an increase in catalase activity and GSH levels. There is no known highly effective therapeutic remedy for SCI; however, there is a time window through which therapeutic interventions would be protective, as discussed in the next section. A summary of the pathophysiology of the secondary injury following SCI is given in Figure 3.

Stroke (Cerebral Ischemia)

Stroke (cerebral ischemia) is a major cause of death and the primary cause of adult disability in most

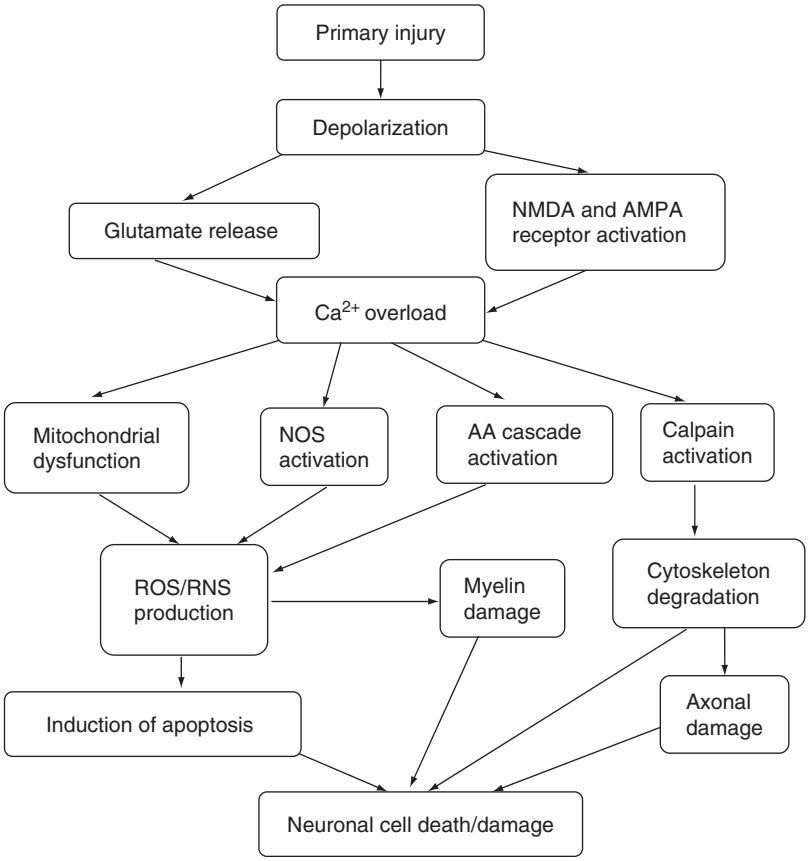

Figure 3 A summary of the pathophysiological events in SCI following the initiation of the secondary injury pathways. Following SCI, there is neuronal depolarization that leads to glutamate release and activation of NMDA and AMPA receptors. This then leads to excessive influx of calcium, resulting in an overload. Ca^{2+} as a second messenger leads to mitochondrial dysfunction and activation of iNOS, arachidonic acid (AA), and calpain. These calcium-activated events lead to production of ROS/RNS, cytoskeleton degradation, axonal and myelin damage, activation of apoptosis, and eventual neuronal cell death. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazols propionic acid. Adapted from Hall ED and Springer JE (2004) Neuroprotection and acute spinal cord injury: A reappraisal. *NeuroRx* 1(1): 80–100. Halestrap AP, McStay GP, and Clarke SJ (2002) The permeability transition pore complex: Another view *Biochimie* 84: 153–166.

industrialized countries, accounting for the largest number of hospitalizations among neurological disorders. Thrombolytic stroke occurs as a consequence of irreversible brain injury resulting from cerebral ischemia provoked by interruption of cerebral blood flow through the occlusion of cerebral arteries. As in SCI and TBI, the mitochondria also play a critical role in the pathological mechanisms leading to tissue damage and infarction following stoke. Reduction or interruption of regional blood flow in the brain produces ischemic injury when delivery of metabolic substrates (i.e., glucose and oxygen) fails to meet the energy demand of the affected tissue. As in SCI and TBI, glutamate also plays a significant role in the pathogenesis of ischemic brain injury through mechanisms already discussed. Briefly, glutamate is released at high concentrations within the core of the cerebral infarction and in the penumbral tissue, overactivating its receptors and leading to excessive influx of calcium that activates a myriad of pathways, leading to mitochondrial dysfunction and to neuronal cell death (excitotoxicity), as previously discussed.

Therapeutic Interventions for Mitochondrial Involvement in CNS Disorders

The mechanisms that follow TBI, SCI, and stroke, among other CNS injuries, are similar, consisting of a primary injury that is further compounded with a secondary injury. Since the secondary injury evolves over time, this has provided a possible window of opportunity through which therapeutic interventions can be developed. Most therapeutic strategies have successfully concentrated on facilitating the actions of antioxidant enzymes, maintaining mitochondrial function, or inhibiting the formation of ROS/RNS following CNS injury. One key therapeutic strategy following

CNS injury has been the use of CsA. As noted previously, CsA is a potent immunosuppressive drug that transiently inhibits the opening of mPTP by unbinding mitochondrial matrix cyclophilin D from the pore. In experimental TBI, CsA dramatically reduces Ca^{2+}-induced cytoskeleton changes, improves mitochondrial function, and significantly ameliorates the volume of cortical damage in rodents. However, CsA does not offer the same neuroprotective effects in SCI as seen in TBI due the intrinsic differences. These intrinsic differences have further complicated the search for therapeutic interventions for SCI. Moreover, CsA may have its own damaging effects with regard to ROS generation under certain conditions. However, there are ongoing studies that show that NIM811, a nonimmunosuppressive CsA derivative, is able to provide neuroprotective effects in both TBI and SCI models. Preliminary results indicate that NIM811 increases cortical tissue sparing, improves mitochondrial function, and reduces mitochondrial oxidative stress after TBI in adult rats. This agent could provide a promising approach in the development of a therapeutic intervention for both SCI and TBI.

Since there is an increase in oxidative stress following CNS injury, the use of antioxidants and antioxidant-related compounds has also been shown to provide neuroprotective effects following CNS injury. For example, it has been demonstrated that posttraumatic administration of the N-type calcium channel blocker SNX-111 (S) and a novel blood–brain barrier penetrating antioxidant U-101033E (U) significantly ameliorates mitochondrial dysfunction induced by TBI in rats. However, caution should be exercised since U-101033E was shown to induce protein oxidation even though it is an inhibitor of lipid peroxidation. In addition, the use of *N*-acetylcysteine (NAC) has also resulted in improvement of mitochondrial function following TBI. Moreover, NAC protected striatum in injury from the lesions modulated by the complex II inhibitor 3-nitopropionic acid. A creatine-supplemented diet significantly attenuates cortical damage after TBI in rodents and induces tissue sparing and possibly provides neuroprotection following SCI.

Mitochondrial uncoupling has also been used as an effective neuroprotective strategy following CNS injury. As noted previously, the production of ROS by the mitochondria is tightly dependent on the mitochondrial membrane potential. As a result, it has been suggested that a reduction in the membrane potential through transient uncoupling would decrease the membrane potential and reduce ROS production and Ca^{2+} uptake. Mitochondrial uncoupling *in vitro* reportedly reduces neuronal mitochondrial Ca^{2+} loading and can inhibit excitotoxic cell death. The use of the uncoupler 2,4-dinitrophenol (2,4-DNP) following stroke/cerebral ischemia has been shown to result in early stabilization of mitochondrial homeostasis and a reduction in the release of cytochrome c into the cytoplasm. Also, administration of uncouplers DNP and carbonylcyanide-*p*-trifluoromethoxyphenylhydrazone following TBI and SCI has been shown to lead to less tissue loss, improved behavioral outcomes, a reduction in mitochondrial oxidative damage, Ca^{2+} loading, and dysfunction, and significantly reduces mitochondrial dysfunction associated with injury. As a result, mitochondria uncoupling could potentially be another therapeutic intervention strategy for TBI.

Since glutamate-related excitotoxicity is one of the mechanisms involved in the secondary mechanism following SCI, it has been hypothesized that antagonists against non-NMDA receptors would provide protection following SCI. Fifteen-minute post-SCI treatment with one non-NMDA receptor antagonist, 2,3,4-tetrahydro-6-nitro-2,3-dioxo-benzol[f] quinoxaline-7-sulfonamide disodium (NBQX), resulted in improvements in mitochondrial function and a reduction in the levels of ROS and lipid peroxidation products. The therapeutic interventions are not limited to those discussed in this article. There is much ongoing promising research for the development of therapeutic interventions against nervous system injuries.

Conclusions

CNS injuries such as TBI, SCI, and stroke are among the leading causes of disabilities in the population. An urgent program for the development of therapeutic interventions is needed. There are several pathological events that occur following CNS injury/trauma that lead to neuronal cell death. These events occur in a biphasic manner, consisting of an initial or primary insult followed by a prolonged and more deleterious secondary injury. Central to these pathologies are the mitochondria involving postinjury pathological events, such as glutamate-induced excitotoxicity, Ca^{2+} dysregulation, production of ROS/RNS, increased oxidative stress, loss of mitochondrial membrane potential, induction of mPT, release of pro-apoptotic bodies, and activation of downstream caspases leading to eventual cell death. The increased Ca^{2+} levels also lead to reduced mitochondrial bioenergetics and compromised ETC capacity demonstrated by altered oxygen consumption in TBI and SCI. Promising therapeutic strategies are being developed to provide neuroprotection against nervous system injuries. The use of antioxidants and antioxidant-related compounds, dietary intervention using creatine supplementation, the use of mPTP blockers such as cyclosporin

A, mitochondria uncouplers, the use of NMDA antagonists, among many others, are potential approaches to therapeutic treatment of TBI and SCI. 'Omic' methods, such as genomics, proteomics, lipidomics, and metabolomics, may offer means of better understanding the molecular mechanisms of TBI, SCI, and stroke, as well as monitoring the efficacy of therapeutic interventions.

See also: Apoptosis in Nervous System Injury; Brain Trauma; Exocytosis: Ca^{2+}-Sensitivity; Mitochondrial Organization and Transport in Neurons; Oxidative Damage in Neurodegeneration and Injury; Programmed Cell Death; Spinal Cord Injuries; Stroke; Stroke: Injury Mechanisms.

Further Reading

Azbill RD, Mu X, Bruce-Keller AJ, Mattson MP, and Springer JE (1997) Impaired mitochondrial function, oxidative stress and altered antioxidant enzyme activities following traumatic spinal cord injury. *Brain Research* 765: 283–290.

Butterfield DA and Stadtman ER (1997) Protein oxidation processes in aging brain. *Advances in Cell Aging and Gerontology* 2: 161–191.

DiMauro S, Simonetti S, Chen X, et al. (1993) Mitochondrial dysfunction as a mechanism of CNS injury. *Research Publications – Association for Research in Nervous and Mental Disease* 71: 67–79.

Eldadah BA and Faden AI (2000) Caspase pathways, neuronal apoptosis, and CNS injury. *Journal of Neurotrauma* 17: 811–829.

Green DR and Reed JC (1998) Mitochondria and apoptosis. *Science* 281: 1309–1312.

Halestrap AP, McStay GP, and Clarke SJ (2002) The permeability transition pore complex: Another view. *Biochimie* 84: 153–166.

Hall ED and Springer JE (2004) Neuroprotection and acute spinal cord injury: A reappraisal. *NeuroRx* 1(1): 80–100.

Halliwell B (2006) Oxidative stress and neurodegeneration: Where are we now? *Journal of Neurochemistry* 97: 1634–1658.

Kahl R, Kampkotter A, Watjen W, and Chovolou Y (2004) Antioxidant enzymes and apoptosis. *Drug Metabolism Reviews* 36: 747–762.

Lewen A, Matz P, and Chan PH (2000) Free radical pathways in CNS injury. *Journal of Neurotrauma* 17: 871–890.

Liu Y, Fiskum G, and Schubert D (2002) Generation of reactive oxygen species by the mitochondrial electron transport chain. *Journal of Neurochemistry* 80: 780–787.

Maragos WF and Korde AS (2004) Mitochondrial uncoupling as a potential therapeutic target in acute central nervous system injury. *Journal of Neurochemistry* 91: 257–262.

Opii WO, Nukala VN, Sultana R, et al. (2007) Proteomic identification of oxidized mitochondria proteins following experimental traumatic brain injury. *Journal of Neurotrauma* 24: 772–789.

Springer JE (2002) Apoptotic cell death following traumatic injury to the central nervous system. *Journal of Biochemistry and Molecular Biology* 35: 94–105.

Sullivan PG (2005) Interventions with neuroprotective agents: Novel targets and opportunities. *Epilepsy and Behavior* 7(supplement 3): S12–S17.

Sullivan PG, Rabchevsky AG, Keller JN, et al. (2004) Intrinsic differences in brain and spinal cord mitochondria: Implication for therapeutic interventions. *Journal of Comparative Neurology* 474: 524–534.

Sullivan PG, Rabchevsky AG, Waldmeier PC, and Springer JE (2005) Mitochondrial permeability transition in CNS trauma: Cause or effect of neuronal cell death? *Journal of Neuroscience Research* 79: 231–239.

Tator CH and Fehlings MG (1991) Review of the secondary injury theory of acute spinal cord trauma with emphasis on vascular mechanisms. *Journal of Neurosurgery* 75: 15–26.

Mitochondrial Encephalomyopathies

S DiMauro, E Bonilla, and E A Schon, Columbia University Medical Center, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Mitochondrial myopathies were described in the early 1960s, when systematic ultrastructural and histochemical studies revealed excessive proliferation of normal- or abnormal-looking mitochondria in muscle of patients with weakness or exercise intolerance. Because, with the modified Gomori trichrome stain, the areas of mitochondrial accumulation looked purplish (Figure 1), the abnormal fibers were dubbed 'ragged-red fibers' (RRFs) and came to be considered the pathological hallmark of mitochondrial disease. However, it soon became apparent that in many patients with RRFs, the myopathy is associated with symptoms and signs of brain involvement, and the term 'mitochondrial encephalomyopathies' was introduced. It also became clear that lack of RRFs in the biopsy does not exclude a mitochondrial etiology, as exemplified by Leigh syndrome (LS), an encephalopathy of infancy or childhood invariably due to mitochondrial dysfunction but never accompanied by RRFs.

According to the widely accepted 'endosymbiotic hypothesis,' mitochondria are the remnants of protobacteria that populated anaerobic nucleated cells and endowed them with the precious gift of oxidative metabolism. Thus, mitochondria are the main source of energy for all human tissues and contain many metabolic pathways, only some of which (e.g., the pyruvate dehydrogenase complex (PDHC), the carnitine cycle, the ß-oxidation 'spirals,' and the tricarboxylic acid (TCA) cycle) are shown in Figure 2.

Respiratory Chain

Although defects in all of the aforementioned metabolic pathways are by definition mitochondrial diseases, the term 'mitochondrial encephalomyopathy' has come to indicate disorders due to defects in the respiratory chain. This is the 'business end' of oxidative metabolism, where adenosine triphosphate (ATP) is generated. Reducing equivalents produced in the TCA cycle (also known as the Krebs cycle) and in the ß-oxidation spirals are passed along a series of protein complexes embedded in the inner mitochondrial membrane (the electron transport chain). The electron transport chain consists of four multimeric complexes (I to IV) plus two small electron carriers, coenzyme Q (or ubiquinone) and cytochrome *c*. The energy generated by these reactions is used to pump protons from the mitochondrial matrix into the space between the inner and outer mitochondrial membranes. This creates an electrochemical proton gradient, which is utilized by complex V (or ATP synthase) to produce ATP in a process known as oxidation/phosphorylation coupling.

A unique feature of the respiratory chain is that it is controlled by two distinct genomes, the nuclear DNA (nDNA) and the mitochondrial DNA (mtDNA). Of the approximately 80 proteins that make up the respiratory chain, 13 are encoded by mtDNA and all others are encoded by nDNA. As indicated by the different shadings in Figure 2, complex II, coenzyme Q, and cytochrome *c* are exclusively encoded by nDNA. In contrast, complexes I, III, IV, and V contain some subunits encoded by mtDNA: seven for complex I, one for complex III, three for complex IV, and two for complex V.

Human Mitochondrial DNA

Human mtDNA (Figure 3) is a 16 569-kb circular, double-stranded molecule, which contains 37 genes: 2 rRNA genes, 22 tRNA genes, and 13 structural genes encoding the aforementioned respiratory chain subunits. In the course of evolution, mtDNA has lost much of its original autonomy and now depends heavily on the nuclear genome for the production of factors needed for mtDNA transcription, translation, and replication. Since 1988 (when mutations in mtDNA were first associated with human disease), the circle of mtDNA has become crowded with pathogenic mutations, and the principles of mitochondrial genetics should, therefore, be familiar to the practicing physician.

Heteroplasmy and Threshold Effect

Each cell contains hundreds or thousands of mtDNA copies, which, at cell division, distribute randomly among daughter cells. In normal tissues, all mtDNA molecules are identical (homoplasmy). Deleterious mutations of mtDNA usually affect some but not all mtDNAs within a cell, a tissue, and an individual (heteroplasmy), and the clinical expression of a pathogenic mtDNA mutation is largely determined by the relative proportion of normal and mutant genomes in different tissues. A minimum critical number of mutant mtDNAs is required to cause mitochondrial dysfunction in a particular organ or tissue and mitochondrial disease in an individual (threshold effect).

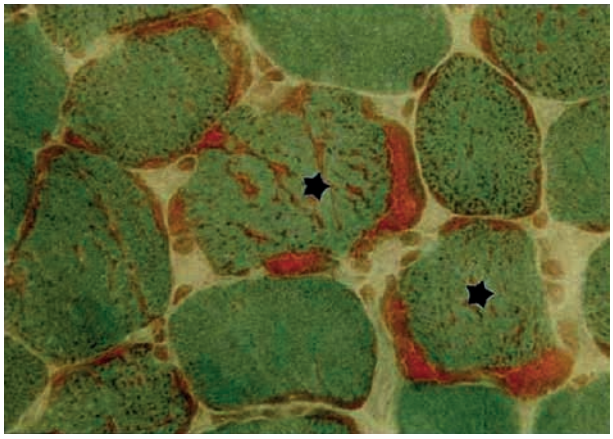

Figure 1 Illustration of ragged-red fibers (RRFs). (a) RRFs revealed by the modified Gomori trichrome stain. Abnormal accumulations of mitochondria appear as reddish blotches, mostly at the periphery of muscle fibers. (b, c) Serial cross-sections from the muscle biopsy of a patient with Kearns–Sayre syndrome. In (b), an RRF (white asterisk) is highlighted by the histochemical stain for succinate dehydrogenase (SDH), which is entirely encoded by nuclear DNA. In (c), the same RRF (black asterisk) shows no activity for cytochrome *c* oxidase, an enzyme that contains three subunits encoded by mitochondrial DNA.

Mitotic Segregation

At cell division, the proportion of mutant mtDNAs in daughter cells may shift and the phenotype may change accordingly. This phenomenon, called mitotic segregation, explains how certain patients with mtDNA-related disorders may actually shift from one clinical phenotype to another as they grow older.

Maternal Inheritance

At fertilization, all mtDNA derives from the oocyte (Figure 4). Therefore, the mode of transmission of mtDNA and of mtDNA point mutations (single deletions of mtDNA are usually sporadic events) differs from that of Mendelian inheritance. A mother carrying a mtDNA point mutation will pass it on to all of her children (males as well as females), but only her daughters will transmit it to their progeny. A disease expressed in both sexes but with no evidence of paternal transmission is strongly suggestive of an mtDNA point mutation. The best way for the clinician to chart a course toward a diagnosis in the jungle of mitochondrial encephalomyopathies is to use a classification that combines genetic and biochemical criteria (Table 1). From the genetic point of view, there are two major categories, disorders due to defects of mtDNA and disorders due to defects of nDNA.

Disorders Due to Defects of mtDNA

Disorders due to defects of mtDNA include rearrangements (single deletions or duplications) and point mutations.

mtDNA Rearrangements

Single deletions of mtDNA have been associated with three sporadic conditions:

1. Pearson syndrome, a usually fatal disorder of infancy characterized by sideroblastic anemia and exocrine pancreas dysfunction.
2. Kearns–Sayre syndrome (KSS), a multisystem disorder with onset before age 20 of impaired eye movements (progressive external ophthalmoplegia; PEO), pigmentary retinopathy, and heart block. Frequent additional signs include ataxia, dementia, and endocrine problems (diabetes mellitus, short stature, hypoparathyroidism). Lactic acidosis, elevated cerebrospinal fluid (CSF) protein (over 100 mg dl^{-1}), and RRFs in the muscle biopsy are typical laboratory abnormalities.
3. PEO with or without proximal limb weakness, often compatible with a normal life span. Deletions vary in size and location, but a 'common' deletion of 5 kb is frequently seen in patients and in aged individuals (Figure 3).

Duplications of mtDNA can occur in isolation or together with single deletions and have been seen in patients with KSS or with diabetes mellitus and deafness. Duplications and duplications/deletions (as well as the associated phenotypes) are transmitted by maternal inheritance.

Point Mutations

Over 150 pathogenic point mutations have been identified in mtDNA from patients with a variety of disorders, most of which are maternally inherited and multisystemic, but some are sporadic and tissue specific (Figure 3). Among the maternally inherited encephalomyopathies, four syndromes are more common. The first is MELAS (mitochondrial encephalomyopathy, lactic acidosis, and strokelike episodes), which usually presents in children or young adults after normal early development. Symptoms include recurrent vomiting, migrainelike headache, and strokelike episodes causing cortical blindness, hemiparesis, or hemianopia. Magnetic resonance imaging (MRI) of the brain shows 'infarcts' that do not correspond to the distribution of major vessels, raising the question of whether the strokes are vascular or metabolic in nature. The most common mtDNA mutation is A3243G in the $tRNA^{Leu(UUR)}$ (*MT-TL1*) gene, but about a dozen other mutations have been associated with MELAS, most notably a mutation (G13513A) in the *MT-ND5* gene, which encodes subunit 5 of complex I.

The second syndrome is MERRF (myoclonus epilepsy with ragged red fibers), characterized by

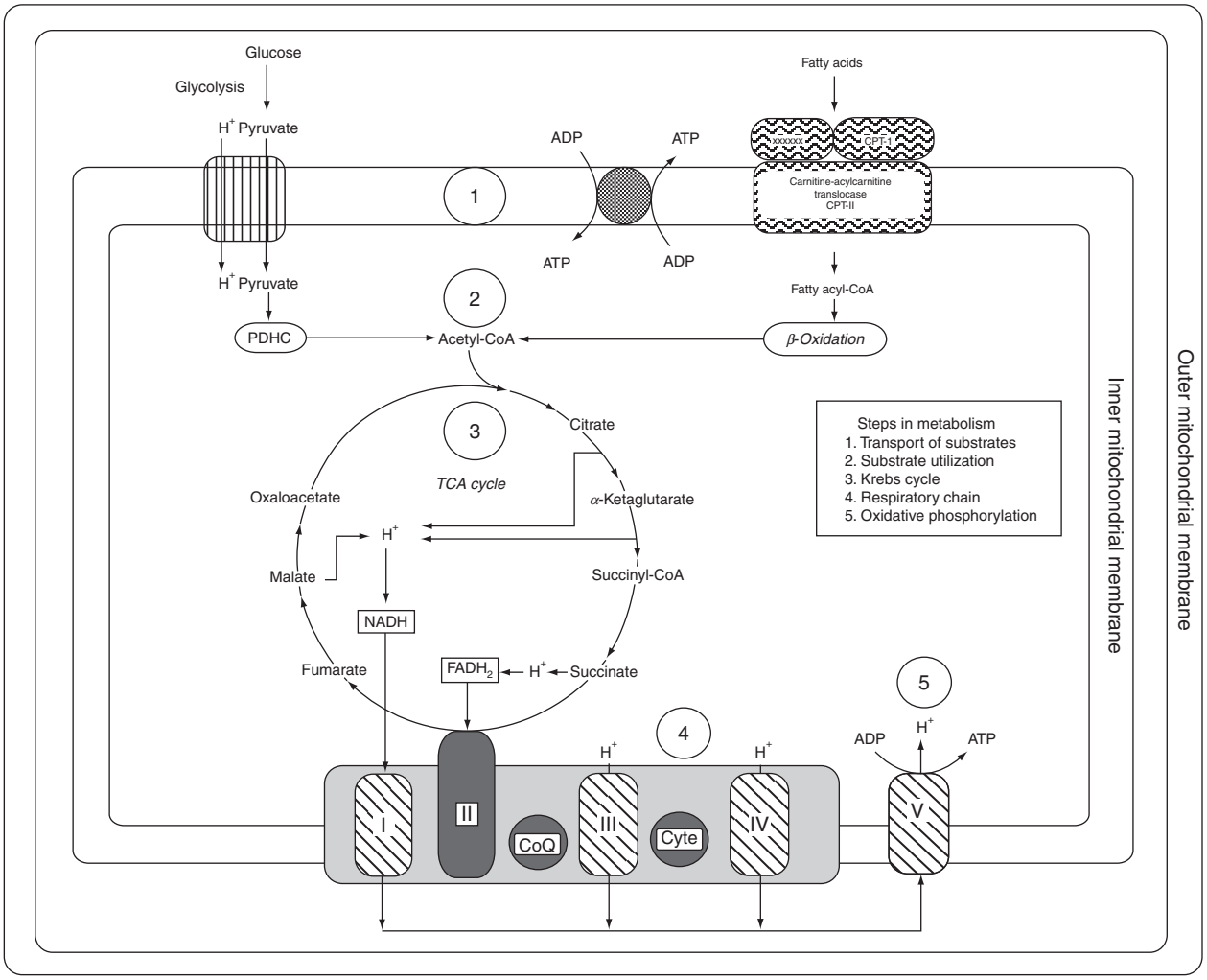

Figure 2 Schematic representation of mitochondrial metabolism. ADP, adenosine diphosphate; ATP, adenosine triphosphate; CPT, carnitine palmitoyltransferase; CoA, coenzyme A; TCA, tricarboxylic acid; FADH2, flavin adenine dinucleotide, reduced form; NADH, nicotinamide adenine dinucleotide, reduced form; I, II, III, IV, V, complexes of the respiratory chain (stippled, complexes containing both subunits encoded by nuclear DNA and subunits encoded by mitochondrial DNA; solid, complexes encoded entirely by nuclear DNA); CoQ, coenzyme Q; Cyt c, cytochrome *c*; PDHC, pyruvate dehydrogenase complex.

myoclonus, seizures, mitochondrial myopathy, and cerebellar ataxia. Less common signs include dementia, hearing loss, peripheral neuropathy, and multiple lipomas. Three mtDNA mutations (A8344G, T8356C, and G8363A) have been associated with MERRF and all are in the $tRNA^{Lys}$ (*MT-TK*) gene.

The third syndrome comes in two forms: (1) NARP (neuropathy, ataxia, retinitis pigmentosa) usually affects young adults and causes retinitis pigmentosa, dementia, seizures, ataxia, proximal weakness, and sensory neuropathy; (2) maternally inherited Leigh syndrome (MILS) is a more severe infantile encephalopathy with characteristic symmetrical lesions in the basal ganglia and the brain stem.

The fourth syndrome, LHON (Leber hereditary optic neuropathy), is characterized by acute or subacute loss of vision in young adults, more frequently males, due to bilateral optic atrophy. Three mtDNA point mutations in NADH dehydrogenase (ND) genes (encoding subunits of complex I) have been associated with LHON. These are G11778A in *MT-ND4*, G3460A in *MT-ND1*, and T14484C in *MT-ND6*.

mtDNA Mutation Effects

Not surprisingly, syndromes associated with mtDNA mutations can affect every system in the body, including the visual (optic atrophy, retinitis pigmentosa, cataracts), auditory (neurosensory deafness), and endocrine (short stature, diabetes mellitus, hypoparathyroidism) systems, and the heart (familial cardiomyopathies, conduction blocks), the gastrointestinal tract (exocrine pancreas dysfunction, intestinal pseudoobstruction, gastroesophageal reflux), and the kidney (renal tubular acidosis). Any combination of the aforementioned

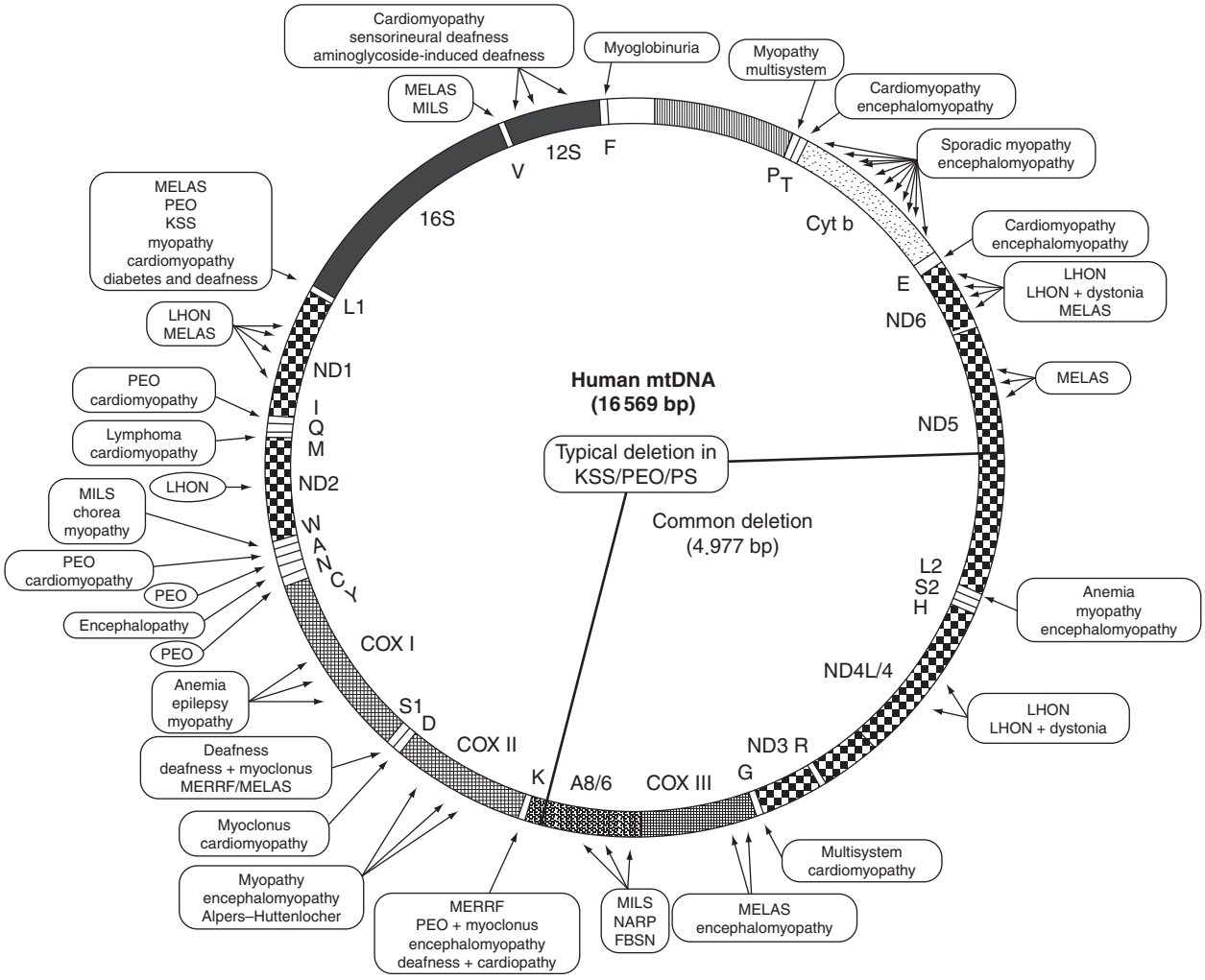

Figure 3 Morbidity map of the human mitochondrial genome as of 1 January 2003. The map of the 16 569-kb mitochondrial DNA (mtDNA) shows differently shaded areas representing the protein-coding genes for the seven subunits of complex I (NADH dehydrogenase, ND), the three subunits of cytochrome *c* oxidase (COX), cytochrome *b* (Cyt *b*), and the two subunits of ATP synthetase (A8/6), the 12S and 16S ribosomal RNAs (12S, 16S), and the 22 transfer RNAs (tRNA) identified by one-letter codes for the corresponding amino acids. FBSN, familial bilateral striatal necrosis; KSS, Kearns–Sayre syndrome; LHON, Leber hereditary optic neuropathy; MELAS, mitochondrial encephalomyopathy, lactic acidosis, and strokelike episodes; MERRF, myoclonic epilepsy with ragged-red fibers; MILS, maternally inherited Leigh syndrome; NARP, neuropathy, ataxia, retinitis pigmentosa; PEO, progressive external ophthalmoplegia.

symptoms and signs should raise the suspicion of a mitochondrial disorder, especially if there is evidence of maternal transmission. On the other hand, point mutations in mtDNA protein-coding genes often escape the rules of mitochondrial genetics in that they affect single individuals and single tissues, most commonly skeletal muscle. Thus, patients with exercise intolerance, myalgia, and, sometimes, recurrent myoglobinuria may have isolated defects of complex I, complex III, or complex IV, due to pathogenic mutations in genes encoding ND subunits, cytochrome *b*, or cytochrome *c* oxidase (COX) subunits. The lack of maternal inheritance and the involvement of muscle alone suggest that mutations arose *de novo* in myogenic stem cells after germ-layer differentiation ('somatic mutations').

Disorders Due to Defects of Nuclear DNA

Disorders due to defects of nDNA are all transmitted by Mendelian inheritance and include three major subgroups.

Mutations in Genes Encoding Subunits or Ancillary Proteins of the Respiratory Chain

As noted in the preceding discussions, mtDNA encodes only 13 subunits of the respiratory chain, while nDNA encodes all subunits of complex II, most subunits of the other four complexes, as well as coenzyme Q_{10} (CoQ_{10}) and cytochrome *c*. Nuclear DNA mutations can affect respiratory chain complexes directly or indirectly.

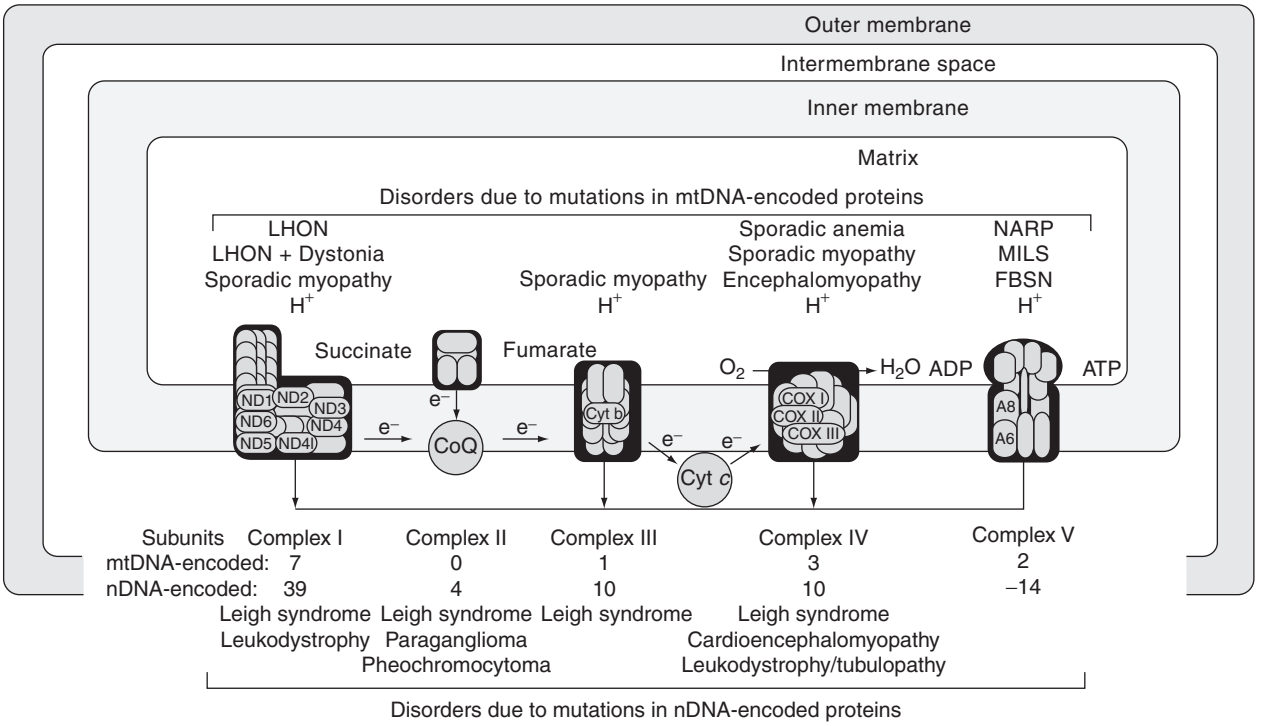

Figure 4 Schematic representation of the mitochondrial respiratory chain, showing nuclear DNA (nDNA)-encoded (unlabeled) and mitochondrial DNA (mtDNA)-encoded (labeled) subunits. Electrons (e^-) flow along the electron-transport chain, and protons (H^+) are pumped, first from the matrix into the intermembrane space through complexes I, III, and IV, then back into the matrix through complex V to generate ATP. Disorders due to mutations in nuclear genes are listed below the defective respiratory chain complexes. LHON, Leber hereditary optic neuropathy; NARP, neuropathy, ataxia, retinitis pigmentosa; MILS, maternally inherited Leigh syndrome; FBSN, familial bilateral striatal necrosis; ND, NADH dehydrogenase; CoQ, coenzyme Q; Cyt *b*/Cyt *c*, cytochromes *b*/*c*; COX, cytochrome *c* oxidase.

Direct 'hits' – that is, mutations in genes encoding respiratory chain subunits – are known only for complex I and complex II. These have been associated with autosomal recessive forms of Leigh syndrome. Primary or secondary CoQ_{10} deficiency can cause four major syndromes: (1) a predominantly myopathic disorder with recurrent myoglobinuria but also central nervous system involvement (seizures, ataxia, mental retardation), (2) a predominantly encephalopathic disorder with ataxia and cerebellar atrophy, (3) an isolated myopathy, with RRFs and lipid storage, and (4) a generalized mitochondrial encephalomyopathy, usually with onset in infancy. An example of secondary CoQ_{10} deficiency is that associated with aprataxin (*APTX*) gene mutations, which cause ataxia oculomotor apraxia type 1 (AOA1). Examples of primary CoQ_{10} deficiency include mutations in two of the eight biosynthetic genes, *COQ1* and *COQ2*. Irrespective of etiology, diagnosis is important because most patients with CoQ_{10} deficiency respond to high-dose CoQ_{10} supplementation.

Indirect 'hits' are mutations in genes encoding proteins that are not components of the respiratory chain, but are needed for the proper assembly and function of respiratory chain complexes. This 'murder by proxy' mechanism has been identified in disorders due to defects in complexes I, III, IV, and V, but is best illustrated by Mendelian defects of complex IV (cytochrome *c* oxidase). Mutations in six ancillary proteins (SURF1, SCO2, SCO1, COX10, COX15, and LRPPRC) have been associated with COX-deficient Leigh syndrome or other multisystemic fatal infantile disorders in which encephalopathy is accompanied by cardiomyopathy (SCO2, COX15), nephropathy (COX10), or hepatopathy (SCO1).

This is a burgeoning field of research with important theoretical and practical implications. From an investigative point of view, these disorders teach us a lot about the structural and functional complexity of the respiratory chain. At a more practical level, identification of mutations in the related genes renders prenatal diagnosis available and suggests approaches to therapy, such as copper administration to infants with mutations in SCO2, a copper-carrying protein.

Defects of Intergenomic Signaling

As noted previously, the mtDNA is highly dependent for its proper function and replication on numerous factors encoded by nuclear genes. Mutations in these

Table 1 Genetic classification

Defects of mitochondrial DNA	*Defects of nuclear DNA*
Mutations in protein synthesis genes	Mutations in respiratory chain subunits
tRNA, rRNA, rearrangements	Complex I, complex II, CoQ_{10}
Mutations in protein-coding genes	Mutations in ancillary proteins
Multisystemic (LHON, NARP/MILS)	Complex III, complex IV, complex V
Tissue specific	Defects of intergenomic signaling
	AR-PEO with multiple Δ-mtDNA
	mtDNA depletion
	Defective mtDNA translation
	Defects of protein transport
	Mohr–Tranebjaerg syndrome
	Defects of the lipid milieu
	Barth syndrome

tRNA/rRNA, transfer/ribosomal ribonucleic acid; LHON, Leber hereditary optic neuropathy; NARP, neuropathy, ataxia, retinitis pigmentosa; MILS, maternally inherited Leigh syndrome; CoQ_{10}, coenzyme Q_{10}; AR-PEO, autosomal recessive progressive external ophthalmoplegia; Δ-mtDNA, mutated mitochondrial DNA.

genes cause Mendelian disorders characterized by qualitative or quantitative alterations of mtDNA. Examples of qualitative alterations include autosomal dominant or autosomal recessive multiple deletions of mtDNA, usually accompanied clinically by progressive external ophthalmoplegia plus a variety of other symptoms and signs. Four of these conditions have been characterized at the molecular level. Mutations in the gene for thymidine phosphorylase (*TP*) are responsible for an autosomal recessive multisystemic syndrome called MNGIE (mitochondrial neurogastrointestinal encephalomyopathy). Mutations in the gene for one isoform of the adenine nucleotide translocator (*ANT1*) have been identified in some, but not all, patients with autosomal dominant PEO. Mutations in the *PEO1* gene, encoding Twinkle, a helicase, are associated with autosomal dominant PEO, whereas mutations in the gene encoding polymerase γ (*POLG*) may cause either autosomal dominant or autosomal recessive PEO. Autosomal recessive mutations in *POLG* also cause an mtDNA depletion syndrome (see later) involving liver and brain, known as Alpers or Alpers–Huttenlocher syndrome. Interestingly, all of these mutations affect genes involved in the homeostasis of the mitochondrial nucleotide pool and probably have similar pathogenic mechanisms.

Examples of quantitative alterations of mtDNA include severe or partial mtDNA depletion, usually characterized clinically by congenital or childhood forms of autosomal recessively inherited myopathy or hepatopathy. Mutations in five genes, four of them involved in mitochondrial nucleotide homeostasis, have been associated with mtDNA depletion syndromes, although they still do not explain all cases. Mutations in the gene encoding thymidine kinase 2 (*TK2*) are often seen in patients with myopathic mtDNA depletion syndromes, whereas mutations in the gene (*SUCLA2*), encoding the β subunit of the mitochondrial matrix enzyme succinyl-CoA synthetase, cause both myopathy and encephalopathy. Mutations in *DGUOK*, encoding deoxyguanosine kinase, predominate in patients with hepatic or hepatocerebral mtDNA depletion syndromes, but, as mentioned previously, mutations in *POLG* are the major causes of Alpers–Huttenlocher syndrome, a severe hepatocerebral syndrome with vulnerability to valproic acid. Mutations in a gene (*MPV17*) not involved in nucleotide pool homeostasis have been associated with hepatocerebral syndrome and with the Navajo neurohepatopathy (NNH) syndrome, prevalent in the Navajo population of Southwestern Unites States. An iatrogenic form of mtDNA depletion may follow treatment with nucleoside analogs such as zidovudine (AZT).

A new group of defects of intergenomic communication is due to mutations in genes encoding factors necessary for the faithful translation of mtDNA-encoded proteins, including *EFG1* (encoding elongation factor 1), *MRPS16* (encoding small subunit protein), *EFTu* (encoding elongation factor Tu), *TSFM* (controlling the expression of both EFTs and EFTu), and *PUS1* (encoding pseudouridine synthase 1). The resulting disorders usually affect infants and cause severe encephalomyopathy, cardiomyopathy, and sideroblastic anemia. Typically, both quality and quantity of mtDNA are normal in these patients, but there are – not surprisingly – multiple respiratory chain defects involving all complexes containing mtDNA-encoded subunits.

Defects of Mitochondria as Causes of Neurodegenerative Disorders

Although they do not affect the respiratory chain directly, defects of mitochondrial motility, fusion, and fission are taking center stage as causes of neurodegenerative disorders. Mitochondria travel on microtubular rails, propelled by motor proteins, usually GTPases, called kinesins or dyneins. The first defect of mitochondrial motility was identified in a family with autosomal dominant hereditary spastic paraplegia and mutations in a gene (*KIF5A*) encoding one of the kinesins. Interestingly, the mutation affects a region of the protein involved in microtubule binding.

Mutations in *OPA1* cause autosomal dominant optic atrophy, the Mendelian counterpart of LHON. Mutations in *MFN2* cause an autosomal dominant axonal variant of Charcot–Marie–Tooth (CMT) disease. Also, mutations in *GDAP1*, the gene encoding ganglioside-induced differentiation protein 1, which is located in the mitochondrial outer membrane and regulates the mitochondrial network, cause CMT type 4A, an autosomal recessive, severe, early-onset form of either demyelinating or axonal neuropathy.

Indirect Causes of Respiratory Chain Dysfunction

We include here mutations in nuclear genes that affect neither respiratory chain subunits or ancillary proteins, nor mtDNA structure or copy number. For example, the function of the respiratory chain can be impaired by alterations in the lipid composition of the inner mitochondrial membrane or by defective importation of one or more subunits. The first situation is exemplified by Barth syndrome, an X-linked recessive disorder characterized by mitochondrial myopathy, cardiopathy, and leukopenia. The gene responsible for this disorder (*TAZ*) encodes a family of proteins ('tafazzins') involved in the synthesis of phospholipids, and biochemical analysis has shown altered amounts and composition of cardiolipin, the main phospholipid component of the inner mitochondrial membrane.

An example of defective protein importation is the X-linked deafness–dystonia syndrome (Mohr–Tranebjaerg syndrome), characterized by progressive sensorineural deafness, dystonia, cortical blindness, and psychiatric symptoms. This disorder is due to mutations in *TIMM8A*, which encodes the deafness–dystonia protein (DDP1), a component of the mitochondrial protein import machinery located in the intermembrane space.

See also: Genomics of Brain Aging: Nuclear and Mitochondrial Genomes; Hereditary Spastic Paraplegia; Mitochondrial Dysfunction in Nervous System Injury; Mitochondrial Organization and Transport in Neurons; Visceral Pain.

Further Reading

DiMauro S (2006) Mitochondrial myopathies. *Current Opinion in Rheumatology* 18: 636–641.

DiMauro S and Davidzon G (2005) Mitochondrial DNA and disease. *Annals of Medicine* 37: 222–232.

DiMauro S, Hirano M, and Schon EA (2006) Approaches to the treatment of mitochondrial diseases. *Muscle and Nerve* 34: 265–283.

DiMauro S, Hirano M, and Schon EA (eds.) (2006) *Mitochondrial Medicine*. Abingdon, UK: Informa Healthcare.

Wallace DC (2005) The mitochondrial genome in human adaptive radiation and disease: on the road to therapeutics and performance enhancement. *Gene* 354: 169–180.

Mitochondrial Organization and Transport in Neurons

E E Glater and T L Schwarz, Children's Hospital, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Mitochondrial Organization in Neurons

The Need for Mitochondrial Organization and Motility in Neurons

The ability of mitochondria to move and establish a characteristic distribution within a given cell type is a general cell biological process. In dividing yeast, for example, the transport of mitochondria determines the inheritance of the proper quotient of mitochondria by the two daughter cells, and this phenomenon has been studied extensively. In neurons, however, mitochondrial transport and localization are particularly vital because of the extraordinary lengths of axons and dendrites and the complex anatomy of each cell. Neurons depend on mitochondria for adenosine 5′-triphosphate (ATP) production and for Ca^{2+} buffering. Because the demand for energy and the need for Ca^{2+} buffering differ from one neuron to another and indeed from one part of the neuron to another part, the density of mitochondria differs between neurons and within a single neuron. Though much remains to be determined about how mitochondrial number and distribution are regulated in neurons, it is apparent that mitochondria are dynamic organelles that move long and short distances and stop and start often. These movements presumably allow them to alter their distribution and thereby adapt to changes in the neuron's local energy requirements. In this regard, their movements may be more complex and more tightly regulated than those of other neuronal cargoes, such as synaptic vesicle precursors or receptors undergoing transport to dendrites. The mechanisms of mitochondrial motility, the motors employed, and the adaptor proteins and regulatory proteins that control these motors in neurons are beginning to be understood.

Complex Distribution of Mitochondria

Mitochondria accumulate in subcellular regions that have high energy demands, such as in axonal growth cones and nodes of Ranvier. The latter, for example, have more mitochondria than internodal regions of myelinated nerve, and this is likely to reflect the need for ATP to drive the extrusion of the ions that enter during action potentials at the node. Mitochondria are enriched both pre- and postsynaptically at synapses and also at areas of high metabolic demand as a consequence of the need to pump out ions that have entered through voltage- and ligand-gated channels. An example of this phenomenon occurs in the clustering of mitochondria at the presynapse of the neuromuscular junction during differentiation in frog spinal neuron–muscle cultures (Figure 1). The distribution of mitochondria in dendrites may be a limiting factor in the development and maintenance of synapses in cultures of hippocampal neurons.

The correlation of mitochondrial activity with neuronal activity has proven to be useful as a reporter of regional brain function. Dramatic examples of this correlation have been found in somatosensory and visual cortices that were stained for the activity of the mitochondrial enzyme cytochrome oxidase. When these cortices are deprived of sensory input, the corresponding regions of the cortices have decreased cytochrome oxidase staining (Figure 2). The decrease in cytochrome oxidase activity can result either from a decrease in the number of mitochondria at inactive synapses or from a decrease in the activity of mitochondrial enzymes, and in fact a combination of both these phenomena is the most likely explanation. This correlation has been examined in some detail in crustacean motor neurons that were subjected to sustained increases in activity. Mitochondrial vital dyes and counts of mitochondrial cross-sectional areas by electron microscopy have indicated that both the number and functional state of mitochondria were enhanced. Similarly, a comparison of tonic and phasic synapses in the crayfish demonstrated that tonic synapses, which are thought to be the more active class, have a higher mitochondrial density as measured by electron microscopy. Activity-dependent changes are likely to result in changes in number of mitochondria in nonsynaptic compartments of the neuron as well. For example, when the activity in the axonal growth cone of a chick sympathetic neuron is decreased by inhibiting axonal growth, the density of mitochondria in the growth cone also decreases.

Mitochondrial Functions in Neurons

Production of ATP At the synapse and throughout the neuron, mitochondria are thought to play three important roles: producing ATP, buffering Ca^{2+}, and sequestering apoptotic factors. ATP supports many essential neuronal functions. Some, such as protein synthesis and regulatory phosphorylations, are common to all cells, but others are either unique to or more significant in neurons. Thus, either directly or indirectly, ATP fuels the sodium pump for maintaining the membrane resting potential, clearing Ca^{2+} from the cytosol via membrane pumps, packaging

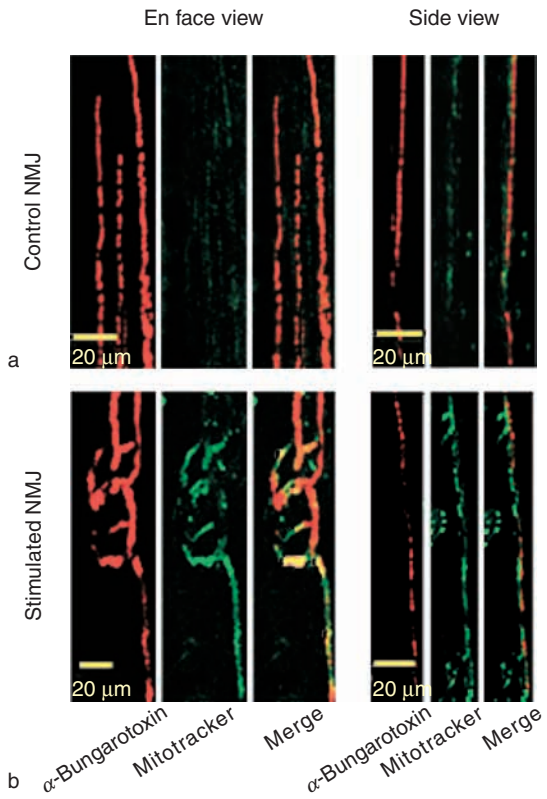

Figure 1 Mitochondria are enriched at synapses. (a) At the frog neuromuscular junction (NMJ), mitochondria labeled with a mitochondrion-selective dye (mitotracker) are found at the nerve terminal, as indicated by their presence near nicotinic acetylcholine receptors labeled with α-bungarotoxin. (b) Stimulation of the nerve terminal increases mitochondrial staining at the NMJ. Belair EL, Vallee J, and Robitaille R (2005) Long-term *in vivo* modulation of synaptic efficacy at the neuromuscular junction of Rana pipiens frogs. *Journal of Physiology* 569: 163–178.

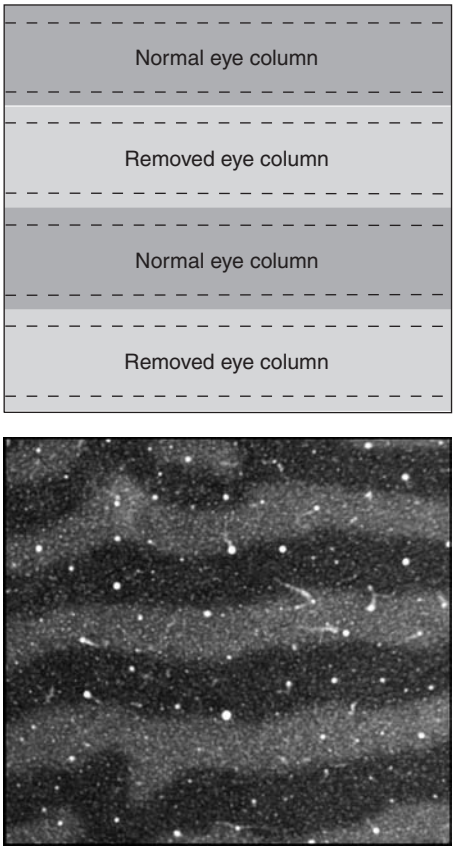

Figure 2 Neuronal activity affects distribution of active mitochondria. The mitochondrial enzyme cytochrome oxidase (CO) serves as a marker for active mitochondria in layer 4 of macaque visual cortex. Removal of one eye severely reduces the activity of neurons in the lateral geniculate nucleus that receive input from that eye. As a consequence, the regions of primary visual cortex that are innervated by those neurons also have reduced activity. This is reflected in the pattern of stripes of high and low cytochrome oxidase activity, which corresponds to regions devoted to input from the active eye and those receiving input from the silenced pathway. Horton JC and Adams DL (2005) The cortical column: A structure without a function. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 360: 837–862.

neurotransmitter into vesicles, and recovering neurotransmitter from the synaptic cleft. Indeed, maintaining ionic homeostasis through pumps is the most important energetic demand on a nerve cell.

Ca^{2+} buffering In addition to producing ATP, mitochondria play a key role in buffering local Ca^{2+} concentrations in nerve terminals. Ca^{2+} enters the mitochondrial matrix through the Ca^{2+}-permeable uniporter (presumably a channel) in the mitochondrial inner membrane. The rapid mitochondrial uptake of cytoplasmic Ca^{2+} dissipates microdomains of high Ca^{2+} and prevents diffusion of Ca^{2+} in the cell. In addition to Ca^{2+} influx into the mitochondria, the exit of Ca^{2+} from the mitochondria also affects cytoplasmic Ca^{2+}. After accumulating Ca^{2+}, mitochondria release it slowly through a sodium–Ca^{2+} exchanger. Thus, if Ca^{2+} channels in the plasma membrane open to create a large, sudden rise in cytoplasmic Ca^{2+}, Ca^{2+} will first accumulate in the mitochondrial matrix, which can decrease the amplitude and duration of the large change in cytosolic Ca^{2+}. When Ca^{2+} slowly exits the mitochondria, the effect will be to keep the cytoplasmic level of Ca^{2+} somewhat elevated, long after the original stimulation for the Ca^{2+} increase has stopped. Ultimately, this Ca^{2+} will be extruded across the plasma membrane to restore the initial resting conditions of the cell.

Apoptosis In addition to producing ATP and buffering Ca^{2+}, mitochondria are important for apoptosis, or programmed cell death. Many different factors, such as DNA damage, oncogene activation, or withdrawal of growth factors, can activate mitochondrial-dependent apoptosis. On activation of apoptosis, two proteins, Bax and Bak, oligomerize on the

mitochondrial outer membrane, and this leads to mitochondrial outer membrane permeabilization and release of cytochrome C, which in turn activates caspases, proteolysis, and cell death. It remains uncertain to what extent changing the distribution of mitochondria in a cell will alter the cell's susceptibility to localized proapoptotic signals.

Mitochondria and the Function of the Nerve Terminal

Because the availability of ATP and the regulation of cytosolic Ca^{2+} are so important to nerve terminals, and because mitochondria are abundant at synapses, it is not surprising that the particular role of mitochondria in shaping synaptic transmission has received a lot of attention. For many years this has been addressed pharmacologically, through the application of compounds that poison the electron transport chain in mitochondria. Some drugs will collapse the potential across the mitochondrial membrane (e.g., the protonophore FCCP) and thereby prevent both oxidative phosphorylation for the production of ATP and mitochondrial uptake of Ca^{2+}. Other compounds (e.g., ruthenium red) can be employed that selectively block the Ca^{2+} uptake pathway. More recently, it has been possible to examine synapses in *Drosophila* mutants that lack mitochondria because of defects in mitochondrial fission or transport. This approach can bring about the complete absence of mitochondria from the nerve terminal, but unlike the pharmacological approach, it is a chronic manipulation. In principle, genetic approaches may also be able to distinguish among types of mitochondrial function, but this has not yet been reported.

Ca^{2+} buffering by mitochondria can influence transmitter release One manner in which mitochondria can influence transmitter release is via their ability to sequester and then release Ca^{2+}. Cytosolic proteins are likely to be the major buffers of Ca^{2+} in neurons, but mitochondria can also contribute to the buffering. As discussed above, if mitochondria take up Ca^{2+} rapidly when the cytosolic concentration is high and then release it slowly as the concentration falls, a prolonged phase of somewhat elevated cytosolic Ca^{2+} arises in consequence. After a train of action potentials, this sustaining of cytoplasmic Ca^{2+} is thought to be important for synaptic plasticity. For example, it has been shown that posttetanic potentiation (PTP), the enhancement of synaptic transmission for several minutes that often occurs after high-frequency stimulation, is in part due to mitochondrial Ca^{2+} sequestration. The underlying mechanism is likely to be that cytoplasmic Ca^{2+} levels remain elevated after stimulation and thereby increase the amount of neurotransmitter release. Inhibition of mitochondrial Ca^{2+} uptake and release both blocked PTP and reduced cytoplasmic Ca^{2+} levels. Mitochondrial sequestration has also been shown to be important for terminating transmitter release after a stimulus and preventing asynchronous neurotransmitter release.

The endoplasmic reticulum (ER) may also play an important role in buffering synaptic Ca^{2+}, although this point is very controversial, and the relative importance of the ER and mitochondria for buffering Ca^{2+} may vary between synapses. At some, one compartment or the other may dominate, or both may contribute to Ca^{2+} sequestration. At other synapses, neither the ER nor mitochondria have a substantial influence on the shape of the Ca^{2+} signal. Regardless of how actively they sequester Ca^{2+}, release induced from either store can influence transmitter release. Thus multiple Ca^{2+} buffering mechanisms may allow different synapses to tailor their buffering capacity to the needs of the synapse.

Synaptic transmission in mutants lacking synaptic mitochondria The importance of mitochondria for synaptic function has been highlighted by three different *Drosophila* mutants that lack mitochondria in their axons and terminals (Figure 3). These mutants, *milton*, *miro*, and *drp1*, all have defects in synaptic transmission and indeed were identified in screens for impaired synaptic transmission at photoreceptor synapses. Transmission in the *miro* mutant has been examined closely at the neuromuscular junction, where it causes a decrease in vesicle release (decreased amplitudes of excitatory junctional potentials) during high-frequency stimulation protocols and an increase in asynchronous release after prolonged stimulation. *Drp1* mutants have a similar neuromuscular phenotype. The *drp1* phenotype was interpreted as arising from a requirement for mitochondrial ATP in the mobilization of the reserve pool of synaptic vesicles. Drp1 is the fly homolog of dynamin-related protein and plays a role in mitochondrial fission, whereas milton and miro proteins are required for axonal transport of mitochondria (see below).

These phenotypes notwithstanding, the three *Drosophila* mutants (*milton*, *miro*, and *drp1*) that lack synaptic mitochondria are not as severely compromised at their synapses as might have been anticipated. The axons develop normally and reach their targets with no gross abnormalities. Moreover, these mutants have normal synaptic transmission at the neuromuscular junction unless challenged with high-frequency stimulation. The mutations are lethal, but the homozygotes for *miro* and *drp* survive to the third instar stage. The explanation for this persistence of normal function is not yet clear but

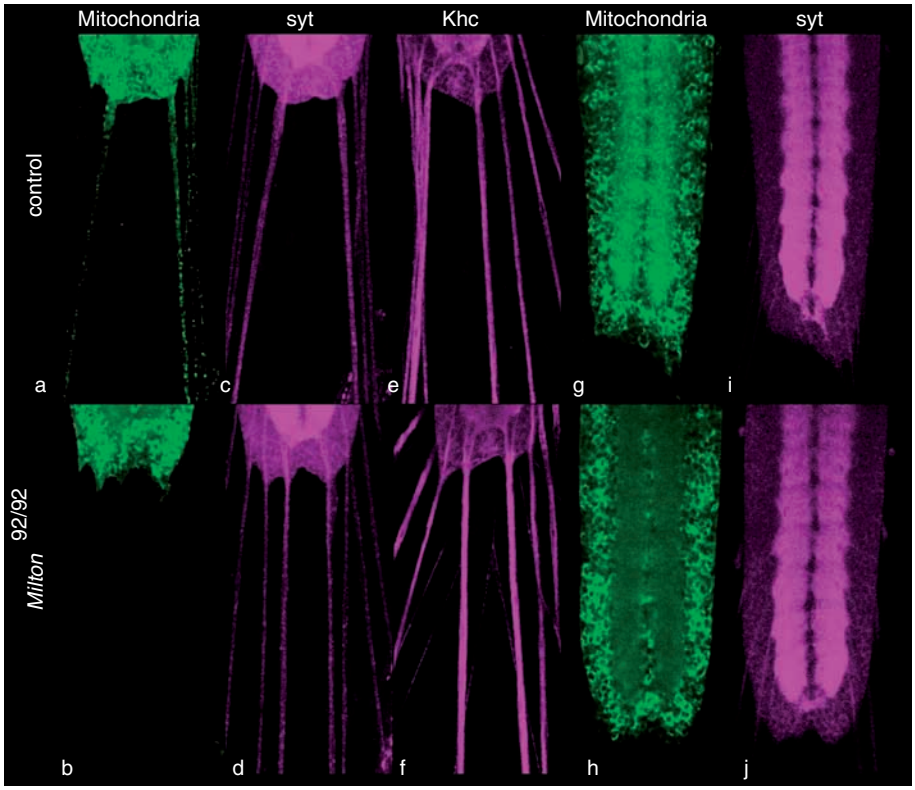

Figure 3 Mitochondria are absent from axons of flies lacking the protein milton. (a–f) Images of peripheral nerves that extend from the posterior of the nerve cord to the body wall muscles. (a) In control larvae, mitochondria labeled with green fluorescent protein (mitoGFP) are abundant in axons. (b) In mutant larvae lacking the protein milton, mitochondria are absent from axons. The defect is specific for mitochondria. Antisynaptotagmin immunolabeling of synaptic vesicles (c, d) or antikinesin heavy chain immunolabeling (e, f) is indistinguishable between *milton* mutant and control larvae. (g–j) Confocal images through the middle of the ventral nerve cord. (g) In control larvae, on either side of the midline, the synapse-rich neuropil is enriched for mitochondria (mitoGFP) relative to the surrounding cell-body region. (h) In mutant nerve cords lacking milton, the mitoGFP is primarily in cell bodies. (i, j) Synaptotagmin is concentrated in the neuropil of both control and mutant larvae. Khc, kinesin heavy chain; syt, synaptotagmin. Glater EE, Megeath LJ, Stowers RS, and Schwarz TL (2006) Axonal transport of mitochondria requires milton to recruit kinesin heavy chain and is light chain independent. *Journal of Cell Biology* 173: 545–557.

may indicate a relatively minor role for the mitochondria in Ca^{2+} buffering at these terminals and the presence of alternative sources of ATP.

Mechanisms of Mitochondrial Motility

Mitochondrial Movement

Imaging of mitochondria in neurons reveals that their neurites contain many discrete mitochondria and that a number of these mitochondria are in motion at any given moment. Thus, whereas mitochondria may exist in a continuous, interconnected reticulum in many cell types – and indeed may form such a reticulum in neuronal cell bodies and in primary dendrites – in many neuronal processes they are encountered as discrete organelles that appear to travel independently of one another. The visualization of mitochondrial motility depends on live imaging, which has been accomplished with a variety of vital dyes and, more recently, with transgenes that express green fluorescent protein to which a mitochondrial import signal has been attached.

Why do neurons contain moving mitochondria? One likely explanation is that mitochondria must move from the cell body, the site of the majority of organelle biogenesis, to the axons and their terminals. It is also possible that some replication of mitochondria occurs in the periphery of the cell. Mitochondrial movement is bidirectional; some of the mitochondria are moving from the soma to the periphery and some from the periphery back to the soma. This may represent older mitochondria that are returning to the soma for degradation or to be resupplied with newly synthesized proteins. Equally important as an explanation of mitochondrial motility may be the capacity of the cell to monitor local energy supplies and redistribute the mitochondria accordingly. If mitochondria become mobile where ATP is high and are immobilized where it is low, they will shift in such a

way as to preserve an adequate energy supply throughout the cell. In contrast, the stationary mitochondria may guarantee that a subset of mitochondria is devoted to the energetic demands of a given region.

The pattern of mitochondrial movement has been described in several systems, including hippocampal neurons and *Drosophila* nerves. Their movement is often described as saltatory, meaning they stop and start, change speed, and often reverse direction. In hippocampal axons, approximately one-third of mitochondria move at any one time, and in the same preparation, they can behave in a variety of manners: some move exclusively in one direction, others shift directions, some change their speed between fast and slow rates, and others exhibit only a single speed. Thus the complex movements of mitochondria are somewhat different from those of other axonal cargoes, such as small vesicles and endosomes, which consistently move in a single direction.

Mechanisms of Axonal Transport of Mitochondria

Mitochondria move by both microtubule and actin-dependent mechanisms. The majority of mitochondrial movements are microtubule-based and can be arrested by the pharmacological depolymerization of the microtubules. Actin depolymerization also decreases mitochondrial movement, though to a lesser extent, which has led to suggestions that only 20% of mitochondrial movement in axons is actin-based, with the remainder driven by microtubule-based motors. The actin-based movements are likely to drive only short-range movement, with long-distance transport of mitochondria transpiring via microtubules. In most neurons, the axon is a highly polarized structure, with the minus ends of the microtubules located in the cell bodies and the plus ends at axon terminals. Dendrites, in contrast, contain microtubules of both polarities. In consequence, transport can be simpler to analyze in axons than in dendrites, but in both types of neurite, the behavior of the mitochondria will be governed by the interplay of plus end-directed microtubule motors, the kinesins, with dynein, the minus end-directed motor, as well as with actin and myosin-based motors and the poorly understood machinery that is likely to tether mitochondria in place.

A kinesin-containing complex for mitochondrial transport Two kinesin motors have been implicated in axonal transport of mitochondria: conventional kinesin-1 and kinesin-3. Kinesin-1 is a tetramer made up of two kinesin heavy chains (KHCs) and two kinesin light chains (KLCs). In *Drosophila,* KHC is crucial role in mitochondrial transport. In *khc* mutants, axonal transport is severely impaired and in the segmental nerves, gives rise to large swellings of axons at irregular intervals (Figure 4). These swellings, sometimes called traffic jams, are filled with a variety of immobilized organelles, including mitochondria. In photoreceptor axons that are homozygous for *khc* null alleles, the axons are devoid of mitochondria. In live imaging of *khc* mutants, mitochondrial anterograde movement is greatly decreased. Thus the *khc* phenotype provides strong evidence for the importance of this conventional kinesin-1 in anterograde axonal transport of mitochondria.

However, the conventional kinesin may be acting in an unconventional state. The structure of a kinesin-1 motor is typically tetrameric, consisting of two heavy chains and two light chains. The heavy chains form the motor and stalk domains, and the light chains often serve as regulators and cargo-attachment areas. In *Drosophila*, however, mitochondria are transported into axons even in neurons that are homozygous for null alleles of the single KLC, and the light chain is not present in the mitochondrial motor-adaptor complex that includes the heavy chain.

The importance of the kinesin-1 motors for anterograde movement of mitochondria appears to be evolutionarily conserved. In mammals, a homolog of KHC, called KIF5B, has similarly been implicated in mitochondrial transport. In *kif5B-/*-mouse extraembryonic cells, the mitochondria were clustered in the perinuclear region of the cell and absent from the periphery, where they are normally found in wild-type cells. In subcellular fractionation studies, KIF5B was associated with the mitochondrial fraction. It is not known if this mammalian kinesin-1 heavy chain is coupled to mitochondria without a light chain, as in *Drosophila*, but two isoforms of the mammalian light chains, KLC-B and KLC-3, have been reported to associate with mitochondria in nonneuronal cells. Whether they are involved in axonal transport is unknown.

Kinesin-1 may not be the sole anterograde motor for mitochondria. A kinesin-3 family member, mammalian KIF1B, was shown to transport purified mitochondria along microtubules *in vitro*, to colocalize with mitochondria, and to be enriched in a mitochondrial fraction. It is interesting that mutants in the *Drosophila* homolog of KIF1B, *Imac*, do not have defects in axonal transport of mitochondria but do have defects in transport of synaptic vesicles. Thus there may be some variation from species to species and from cell type to cell type as to which motors are involved in the anterograde movement, although the case for the involvement of the conventional kinesin-1 is presently very strong.

How then is the kinesin attached to the mitochondrion and how is its activity regulated? Recent studies have identified a protein complex that couples the KHC

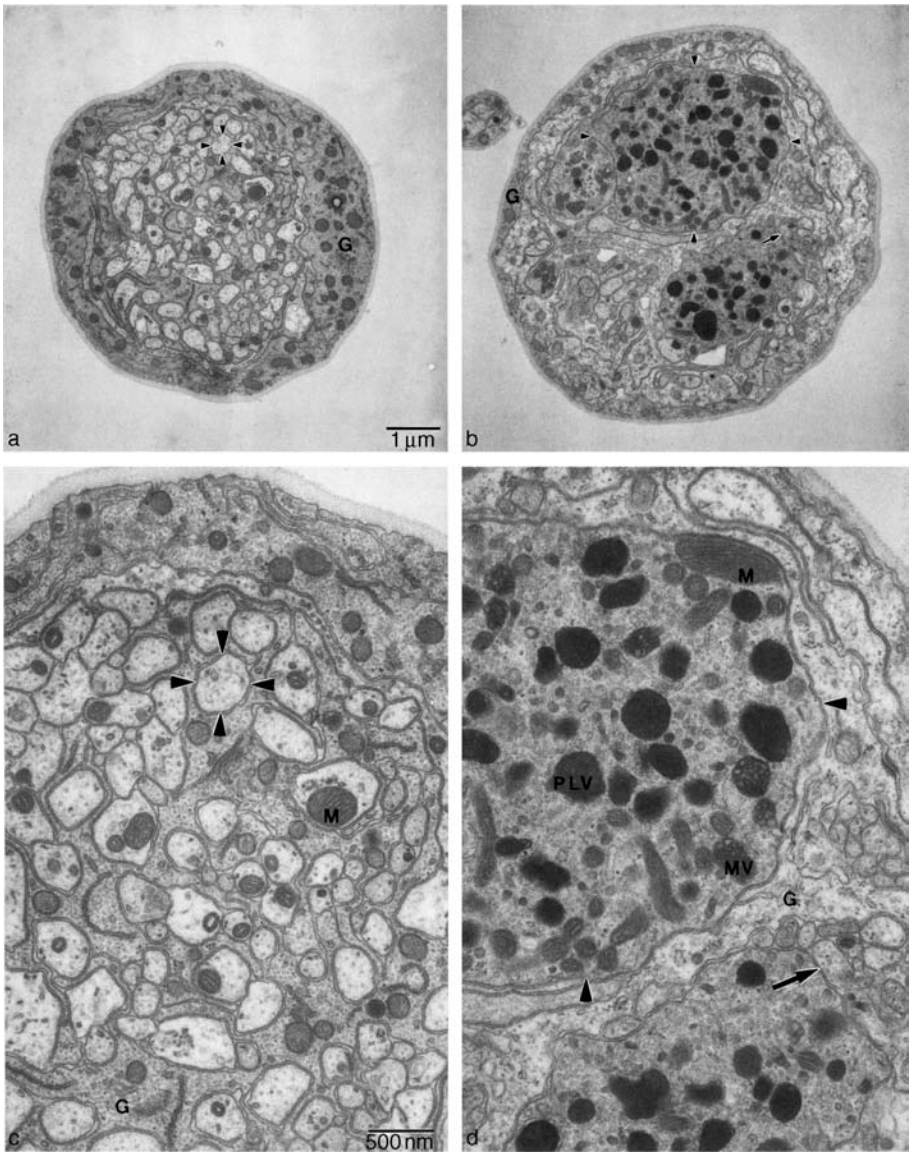

Figure 4 Axonal pathology caused by defects in kinesin. Axonal swellings include accumulations of organelles, including mitochondria. *Drosophila* segmental nerves from control larvae (left) and from larvae mutant in the heavy chain of kinesin-1 (right). Though most axons retain their normal appearance, with diameters of less than 1 μm, some axons have massive swellings within which a variety of organelles are encountered, including mitochondria (M), multivesicular bodies (MV), likely prelysosomal vacuoles (PLV), and a variety of smaller vesicles. Hurd DD and Saxton WM (1996) Kinesin mutations cause motor neuron disease phenotypes by disrupting fast axonal transport in Drosophila. *Genetics* 144: 1075–1085.

to the outer surface of the mitochondrion, and this complex is likely to be a focal point for the regulation of plus end-directed mitochondrial movement. The complex consists of at least three proteins: conventional KHC, milton, and miro. Milton is a protein of approximately 1100 amino acids that is conserved between *Drosophila* and mammals, though in mammals it has also been named *O*-linked *N*-acetylglucosamine-Interacting Protein 106, OIP98/$GABA_A$ Receptor Interacting Factor-1, and Trafficking Protein, Kinesin Binding. The importance of the protein for axonal and dendritic transport was revealed by the isolation of null mutants in *Drosophila* in a screen for mutations with defective synaptic transmission. In these mutants, neuronal mitochondria were found to be restricted to the cell bodies. Within photoreceptor cell bodies, they were broadly distributed, much as in wild-type photoreceptors, but they were never encountered beyond the axon hillock. Milton contains a coiled-coil domain through which it binds to the KHC. This association is inhibited by the presence of the KLC, which is consistent with the persistence of mitochondrial transport in light chain mutants. Milton in turn binds to the mitochondrial membrane, at least in part, via its interactions

with the third protein, miro. Miro is anchored in the mitochondrial membrane by a C-terminus transmembrane domain, but the protein is chiefly exposed on the cytoplasmic surface. It contains two guanosine triphosphatase (GTPase) domains with some similarity to the rho family and also contains a pair of EF hands. The *miro* mutants, like *milton* alleles, lack axonal mitochondria but retain functional mitochondria in their cell bodies. Thus the present model for mitochondrial transport has milton serving as an adaptor between the mitochondrial surface protein, miro, and the KHC.

An additional protein that may be important for determining the distribution of mitochondria in mammalian neurons is syntabulin, a protein without a clear homolog in the *Drosophila* genome. Syntabulin localizes to the mitochondrial outer membrane and also interacts with microtubules. Syntabulin may serve as an adapter between KHC and mitochondria or may play a role in tethering stationary mitochondria to the cytoskeleton.

Dynein and mitochondrial transport Dynein has been implicated as the major motor for microtubule-based retrograde axonal transport, and dynein has been shown to be associated with mitochondria. The most direct evidence for a role of dynein in retrograde transport of mitochondria is impairment of this process in *Drosophila* dynein heavy chain mutants. Dynein is part of a complex including dynein heavy, intermediate, and light chains as well as many associated adapter and regulatory proteins. To date, no specific dynein complex components or adapters have been shown to be specifically associated with mitochondria.

It is interesting to note that dynein seems to depend on kinesin-1 for transport and that these two motors are interconnected. Indeed, a direct interaction has been shown between dynein intermediate chain and KLCs. The interaction of these proteins may reflect a requirement for kinesin to bring dynein to the periphery so that dynein is available for returning cargoes to the soma, but it may also indicate an ongoing connection and functional interdependence. An indication of this interdependence may be the observation that *Drosophila khc* mutants are impaired in both anterograde and retrograde movements of mitochondria. A regulatory interplay between the motors would seem to be required when mitochondria switch directions in the axon.

Actin, myosin, and mitochondrial transport Actin-based transport of mitochondria is well established in nonneuronal cells and, as mentioned above, may be responsible for short-range movements of mitochondria in axons as well, though less is known about these mechanisms in neurons. In yeast, actin-based mechanisms drive mitochondrial motility. One important pathway is myosin independent and is driven instead by the polymerization of actin. In contrast, in neurons and other animal cells, the evidence presently points to unconventional myosin motors as the basis of actin-dependent mitochondrial movements. In locust photoreceptor cell bodies, mitochondrial transport has been shown to be actin-dependent, and a putative, myosin, has been implicated as the motor. This myosin was shown to localize to mitochondria in photoreceptor cells, and *in vitro* the protein shows an ATP-dependent binding to actin filaments.

Regulation of Mitochondrial Movement

Recently, several intracellular signals have been shown to play a role in specifically controlling the movement of mitochondria. Some pathways seem to affect global mitochondrial movement and some local mitochondrial movement. In general, sensitivity to local cytoplasmic Ca^{2+} concentration may be a crucial regulatory mechanism for neuronal mitochondria. High local influxes of Ca^{2+} will correlate with regions of high metabolic demand, such as nerve terminals and postsynaptic specializations. Thus a tendency for mitochondria to arrest where Ca^{2+} is elevated will cause them to cluster near synapses. In addition, a sustained elevation of cytosolic Ca^{2+} may be a symptom of insufficient local ATP to power the ATPase that extrudes the ion across the plasma membrane. In this case, it would again be advantageous for mitochondria to be attracted to areas of high Ca^{2+} or to stop in their vicinity as a means of balancing ATP production throughout the cell. One manifestation of the ability of elevated Ca^{2+} to arrest mitochondrial transport is seen in the application of glutamate to cultured hippocampal neurons and the consequent Ca^{2+} influx through *N*-methyl-D-aspartate receptors. How Ca^{2+} regulates mitochondrial motility is unknown, but the EF hands of miro are candidates for the Ca^{2+}-binding sites that mediate this effect. As part of the complex that couples kinesin to the mitochondrial surface, Ca^{2+} binding to miro potentially could regulate either the activity of kinesin or its association with the mitochondrion.

Ca^{2+} is not the only signal to regulate mitochondria. Local mitochondria respond, for example, to the application of nerve growth factor (NGF). Mitochondria, via an actin-dependent mechanism, accumulate near beads coated with NGF. In addition, zinc has been shown to inhibit mitochondrial movement. It is interesting that both the zinc and NGF effects were shown to be mediated by phosphatidylinositol 3-kinase activity (**Figure 5**).

Other sites of mechanisms of potential regulation are suggested by studies of milton and miro. Miro contains

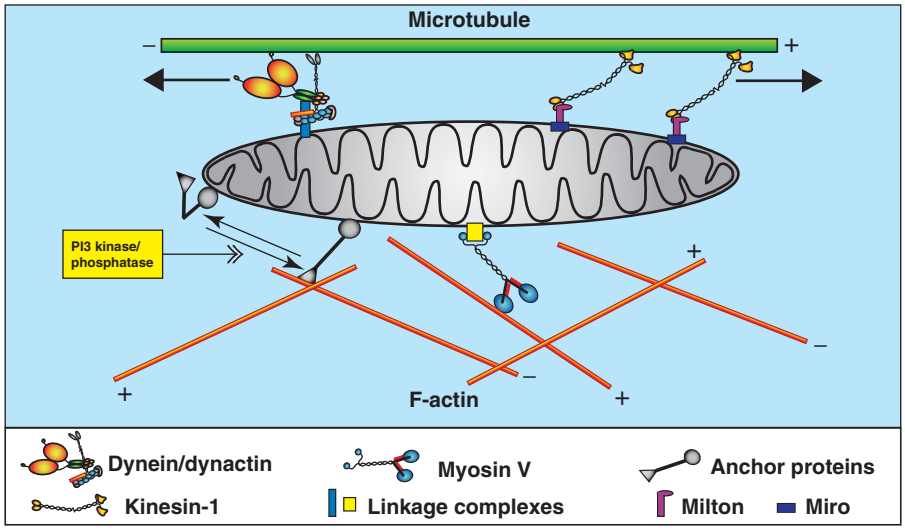

Figure 5 Model of mitochondrial transport. At least three motors are involved in the movement of neuronal mitochondria. Kinesin-1 moves mitochondria in the anterograde direction toward the plus ends of microtubules. Kinesin-1 is attached to the organelle by an adaptor complex of miro and milton. Dynein/dynactin moves mitochondria in the retrograde direction and is attached to the mitochondrion by unknown adaptors. Myosin moves mitochondria in short-range movements along F-actin. Stationary mitochondria may be anchored to F-actin by myosin or unknown anchor proteins. Phosphatidylinositol 3-kinase (PI3 kinase) has been shown to play a role in regulating mitochondrial movement. Modified from Hollenbeck PJ and Saxton WM (2005) The axonal transport of mitochondria. *Journal of Cell Science* 118: 5411–5419.

rho-like GTPase domains, and the GTP or guanosine diphosphate-bound state of these domains may therefore function as a regulator switch. Milton binds to the cytoplasmic glycosylating enzyme O-GlcNAc transferase and is glycosylated by that enzyme, though the significance of this posttranslational mechanism for mitochondrial movement is unknown. Finally, at least four isoforms of *Drosophila* milton exist as a consequence of alternative splicing of the N-terminus, and alternative splicing arises in the mammalian homologs as well. At least one of the *Drosophila* isoforms does not bind KHC when expressed in COS cells, which raises the possibility that this alternatively spliced region can regulate the association of the motor with the organelle.

Mitochondria and Neurodegenerative Disease

With the essential functions of mitochondria in ATP production, Ca^{2+} buffering, and apoptosis, it is not surprising that defects in mitochondrial function lead to disease. Moreover, recently it has been shown that defects specifically in the axonal transport of mitochondria are correlated with disease. The most well-characterized example is Huntington's disease, and several studies have shown that mitochondrial transport is disrupted in mouse and *Drosophila* models of the disorder. In transgenic mice expressing mutant huntingtin with polyglutamine repeats, mitochondria move more slowly and stop more frequently, and this occurs prior to overt neurological symptoms. The transport defect seems to be specific for mitochondria because the movement of other organelles, such as endocytic vesicles, appears to remain unaffected.

Several other neurodegenerative disorders have been shown to involve defects in axonal transport, though the extent to which mitochondrial movements are responsible for the pathology is less clear. These diseases include mutations in motor proteins implicated in hereditary spastic paraplegia, Charcot-Marie-Tooth disease Type 2A, congenital fibrosis of the extraocular muscles Type I, and motor neuron disease. In addition, axonal transport has been shown to be disrupted in Alzheimer's disease and implicated in subsequent neuronal death. Although these are complex disorders in which many cellular processes may be compromised, it seems likely that defects in the distribution of mitochondria will contribute to the manifestations of the disorders.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Adenosine Triphosphate (ATP); Apoptosis in Nervous System Injury; Apoptosis in Neurodegenerative Disease; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and Neurodegenerative Diseases; Intracellular Calcium and Neuronal Death; Mitochondrial Dysfunction in Nervous System Injury; Myosin Transport and Neuronal Function; Neuronal Motility and Structure: MARK and GSK Pathways; Programmed Cell Death; Synaptic Transmission: Models.

Further Reading

Boldogh IR, Fehrenbacher KL, Yang HC, and Pon LA (2005) Mitochondrial movement and inheritance in budding yeast. *Gene* 354: 28–36.

Chada SR and Hollenbeck PJ (2004) Nerve growth factor signaling regulates motility and docking of axonal mitochondria. *Current Biology* 14: 1272–1276.

Duncan JE and Goldstein LS (2006) The genetics of axonal transport and axonal transport disorders. *Public Library of Science Genetics* 2(9): e124.

Glater EE, Megeath LJ, Stowers RS, and Schwarz TL (2006) Axonal transport of mitochondria requires milton to recruit kinesin heavy chain and is light chain independent. *Journal of Cell Biology* 173: 545–557.

Guo X, Macleod GT, Wellington A, et al. (2005) The GTPase dMiro is required for axonal transport of mitochondria to Drosophila synapses. *Neuron* 47: 379–393.

Hollenbeck PJ and Saxton WM (2005) The axonal transport of mitochondria. *Journal of Cell Science* 118: 5411–5419.

Hurd DD and Saxton WM (1996) Kinesin mutations cause motor neuron disease phenotypes by disrupting fast axonal transport in Drosophila. *Genetics* 144: 1075–1085.

Ligon LA and Steward O (2000) Role of microtubules and actin filaments in the movement of mitochondria in the axons and dendrites of cultured hippocampal neurons. *Journal of Comparative Neurology* 427: 351–361.

Morris RL and Hollenbeck PJ (1993) The regulation of bidirectional mitochondrial transport is coordinated with axonal outgrowth. *Journal of Cell Science* 104(Pt. 3): 917–927.

Nguyen PV, Marin L, and Atwood HL (1997) Synaptic physiology and mitochondrial function in crayfish tonic and phasic motor neurons. *Journal of Neurophysiology* 78: 281–294.

Pilling AD, Horiuchi D, Lively CM, and Saxton WM (2006) Kinesin-1 and Dynein are the primary motors for fast transport of mitochondria in Drosophila motor axons. *Molecular Biology of the Cell* 17: 2057–2068.

Rintoul GL, Filiano AJ, Brocard JB, Kress GJ, and Reynolds IJ (2003) Glutamate decreases mitochondrial size and movement in primary forebrain neurons. *Journal of Neuroscience* 23: 7881–7888.

Stowers RS, Megeath LJ, Gorska-Andrzejak J, Meinertzhagen IA, and Schwarz TL (2002) Axonal transport of mitochondria to synapses depends on milton, a novel Drosophila protein. *Neuron* 36: 1063–1077.

Tanaka Y, Kanai Y, Okada Y, et al. (1998) Targeted disruption of mouse conventional kinesin heavy chain, kif5B, results in abnormal perinuclear clustering of mitochondria. *Cell* 93: 1147–1158.

Wong-Riley MT and Welt C (1980) Histochemical changes in cytochrome oxidase of cortical barrels after vibrissal removal in neonatal and adult mice. *Proceedings of the National Academy of Sciences of the United States of America* 77: 2333–2337.

Molecular Anatomy of the Mammalian Brain

M A Zapala, University of California at San Diego, La Jolla, CA, USA
C Barlow, BrainCells Inc., San Diego, CA, USA
I Hovatta, National Public Health Institute, Helsinki, Finland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Given the heterogeneity of the brain, the large numbers of genes encoded in the genomes of higher mammals, and the complexity of processes such as behavior, learning, and memory, how is it possible to identify the genes that regulate these phenotypes? How can we use genetic and genomic information to discover targets for therapeutic intervention for brain diseases? As genomic technologies such as DNA microarrays and high-throughput *in situ* hybridizations have become more commonplace, they have led to significant advances in the study of the molecular anatomy of the mouse, human, and primate brain. Researchers have utilized these data to identify genes associated with specific brain functions, behaviors, and disease-related phenotypes. Technologies, experimental design, and data analysis methods have improved so that it is now possible to reproducibly measure gene expression levels in different regions, nuclei, and even single cells of the brain, and to reliably detect very subtle expression differences under specific conditions.

An important goal in neuroscience is the development of a combined high-resolution map of the molecular, genetic, and physical anatomy of the brain, and this goal, although challenging, is now technically feasible. The purpose of obtaining such a map and of developing the tools to navigate is to provide a more complete context for neurobiological studies in much the same way that the genome sequence provides a genetic context for the study of biology. But like sequence information, such a map is more than just a reference text and a trail guide; it is also a discovery tool which will help neuroscientists to formulate and test hypotheses about the brain, speeding the identification of genes important for brain function and helping to understand the molecular bases of brain processes such as learning and memory and the sources of central nervous system (CNS) diseases.

Integrating Genomic Technologies with Neuroanatomy

The most widely used techniques for studying gene expression levels are *in situ* hybridization, DNA microarrays, quantitative real-time polymerase chain reaction (PCR), and Northern blots. In this article we discuss the use of *in situ* hybridization and microarrays, as these have been the methods of choice in the construction of molecular brain atlases. Both of these methods have their advantages and disadvantages. *In situ* hybridization provides an anatomical context for the gene expression, as it is based on hybridization of a probe complementary to the mRNA of interest on a section of the brain. The disadvantages of *in situ* hybridization are that the method is not very sensitive to small differences in expression levels when two samples are compared with each other, and that the anatomical signal is usually restricted to the nucleus of the cell. This latter problem has been solved in the case of bacterial artificial chromosome (BAC) transgenic mice (see later). Another problem with *in situ* hybridization is that usually only one gene is studied at a time and therefore analysis of large numbers of genes is time consuming, although the method can be automated to a certain extent, as in the case of the Allen Brain Atlas (ABA). The major advantage of microarrays is that expression levels of virtually the whole transcriptome can be studied at the same time. However, the anatomical resolution is dependent on the size of the tissue from which the studied RNA is extracted. Even though microarrays provide better sensitivity than *in situ* hybridization does, very small differences (<20%) in expression levels of the two samples are difficult to distinguish, which might be a problem in the brain, where even subtle changes may have dramatic effects on the phenotype.

Microarrays

Despite a great deal of initial skepticism, it is now clear that genomic approaches, such as parallel expression profiling with DNA microarrays, can be used to help elucidate the workings of the brain at the molecular level. Some of the early skepticism concerning the application of genomic techniques to neuroscience was not wholly unwarranted, however, as the study of the brain brings a variety of somewhat unique and formidable technical challenges. Studies of the nervous system are complicated by the fact that the brain is very heterogeneous, the cells of greatest interest might comprise only a small fraction of the entire tissue, high-quality material can be hard to obtain, the appropriate anatomical divisions between regions are not always clear, dissections may be inconsistent, some anatomical and functional regions are extremely small, cell bodies and processes (axons and dendrites) are often located in different parts of the brain, biologically important expression changes

may be very subtle, and the expression profiles are dynamic and subject to change in response to even minimal perturbations. Nevertheless, there have been great strides made in performing genome-wide gene expression studies in the brain, as detailed in the following sections.

In situ Hybridization

An important step in the process of developing a molecular map of the mammalian brain is also the acquisition of high-quality *in situ* hybridization patterns of genes expressed in the brain. *In situ* hybridization provides an in-depth picture of the neuroanatomical gene expression patterns for brain specific genes. These key brain-specific genes can be used to anchor large-scale microarray gene expression studies in an anatomical context. *In situ* hybridization provides information about the distribution of cells expressing a given gene in a given region (scattered vs. coherent; small population expressing at high levels vs. large population expressing at lower levels), as well as the relationship between gene expression boundaries and neuroanatomical boundaries that cannot be extracted easily from microarray data. By integrating microarray data with *in situ* hybridization data, a fuller picture of the molecular anatomy of the brain is developed. Cross-correlation of *in situ* hybridization patterns with microarray data for a relatively small number of genes makes it possible to identify combinations of microarray parameters that usefully predict whether a given gene is likely to be (1) detectable by *in situ* hybridization and (2) specifically expressed in one or more brain regions.

Brain Tissue Collection Methods

Dissection

Key to the molecular mapping effort in the brain are high-quality dissections that preserve mRNA intact. There are distinct advantages to working with mice as model organisms for the molecular mapping of the mammalian brain, as opposed to humans or other mammals. For example, the handling of the animals and the tissue dissections can be performed systematically and consistently. This is extremely important in order to minimize sample-to-sample variations. In our laboratory, all mice are singly housed for 7 days prior to sacrifice, and the dissections are carried out at specified hours of the day. Dissections are performed on the surface of petri dishes filled with wet ice. Samples are dissected as fast as possible and frozen on dry ice. For microarray studies, pooling of tissue from several animals is to be avoided whenever possible because dissection artifacts from one sample in a pool will invalidate the results of the entire pool. If dissecting small brain regions, such as the amygdala or cortical subregions, we either pool samples from two to five animals and collect them in RNAlater buffer (Ambion) or we use methods for small-sample amplification that do not require pooling.

For microarray studies, we perform many brain dissections by hand using a sophisticated dissection microscope and very fine dissection tools. It is important to dissect the exact anatomical structure of interest. Inclusion of too much additional material might 'dilute' a weak expression difference seen between two regions, and failure of including all small subnuclei might lead to missing important differences. Laser-capture microdissection (LCM) is a method for producing reasonably pure populations of targeted cells from small regions of tissue sections that cannot be obtained otherwise. LCM uses a specially designed microscope and an integrated laser to select and collect cells onto a special transfer film. It has been shown that the detailed morphology of the captured cells is maintained and that DNA and RNA of high quality can be extracted from laser-captured cells and used for a range of different analyses, including gene expression monitoring on DNA microarrays. It is important to use stringent quality control of extracted total RNA. This can be achieved by checking RNA quality on an agarose gel and by an absorption measurement or by using a bioanalyzer. It is clear that gene expression profiling of brain tissue from humans or model organisms is feasible if the special nature of the brain tissue is taken into account and rigorous quality control is performed.

Voxelation

With the advent of sophisticated imaging techniques, such as functional magnetic resonance imaging (fMRI), it is now possible to start to integrate brain imaging technologies with molecular and genetic data. To complement the aforementioned dissection strategies, a voxelation approach has been implemented which allows for microarray analysis of spatially registered voxels, or cubes, taken from the brain. Voxelation creates volumetric maps of gene expression similar to imaging techniques such as fMRI, computed tomography (CT), or positron emission tomography (PET). Voxelation has been performed in both human and mouse studies on Alzheimer's and Parkinson's disease brains. The voxelation approach can lead to volumetric resolution in the human brain from 1 cm^3 to as low as 87 μl and in the mouse brain from 7.5 μl to as low as 1 μl. As the mammalian brain is composed of greater than 700 anatomically distinct major regions and an even larger number of named subregions, hand dissection and LCM currently

remains an impractical high-throughput approach for mapping of the entire mammalian brain. The voxelation approach allows for high-throughput analysis of specific brain areas without dissection biases through the use of mathematical analysis techniques such as singular value decomposition. However, voxelation suffers information loss from what is known as 'voxel inhomogeneity,' which is the dilution or averaging of gene expression signals in any one voxel from various brain regions and cell types. Combining gene expression data from brain regions collected by voxelation, hand dissection, and LCM with *in situ* hybridization data will advance the coverage and resolution of the mammalian molecular brain maps.

Current Large-Scale Projects

Gene Expression Nervous System Atlas

The Gene Expression Nervous System Atlas (GENSAT) BAC transgenic project is a National Institute of Neurological Disorders and Stroke (NINDS)-supported two-stage approach that combines high-throughput *in situ* hybridization screening and the use of BAC transgenic reporter gene analysis. The principal investigators of the study are Nathaniel Heintz and Mary E Hatten, professors at the Rockefeller University. The project is designed to use enhanced green fluorescent protein (EGFP) reporter genes both to map gene expression in the developing and adult brain and to provide experimental access to CNS cell populations. The aim is to identify and map the expression of ~5000 of the most important CNS-expressed genes throughout development. First, a prescreen is done by *in situ* hybridization to identify the most interesting genes, and second, the chosen genes are analyzed at high resolution using BAC transgenic mice. This two-stage approach takes advantage of the sensitivity, dynamic range, and efficiency of *in situ* hybridization methodology to allow parallel analysis of large numbers of CNS-expressed genes, while exploiting BAC reporter gene technology to allow systematic analysis and high-resolution visualization of each cell type expressing a gene of interest. The data generated by the project are available in the GENSAT database, which contains a gene expression atlas of the CNS of the mouse based on BACs (see Figure 1). The technology is based on gene replacement, whereby endogenous protein-coding sequences of the studied genes have been replaced by sequences encoding the EGFP reporter gene. Thus, all the cells that would normally express the gene of interest now express the EGFP reporter gene, which can be easily visualized directly or indirectly by immunohistochemistry. This allows the visualization of gene expression patterns in the CNS (and other tissues). The GENSAT database contains histological data from given BAC transgenic mouse lines at three developmental stages: embryonic day 15.5, postnatal day 7, and adult. EGFP is visualized by staining with an anti-EGFP antibody using the diaminobenzidine (DAB) method, or by confocal microscopy of unstained tissue sections. Discoveries in several categories of developmental neurobiology can be made with the dataset, including (1) identifica-

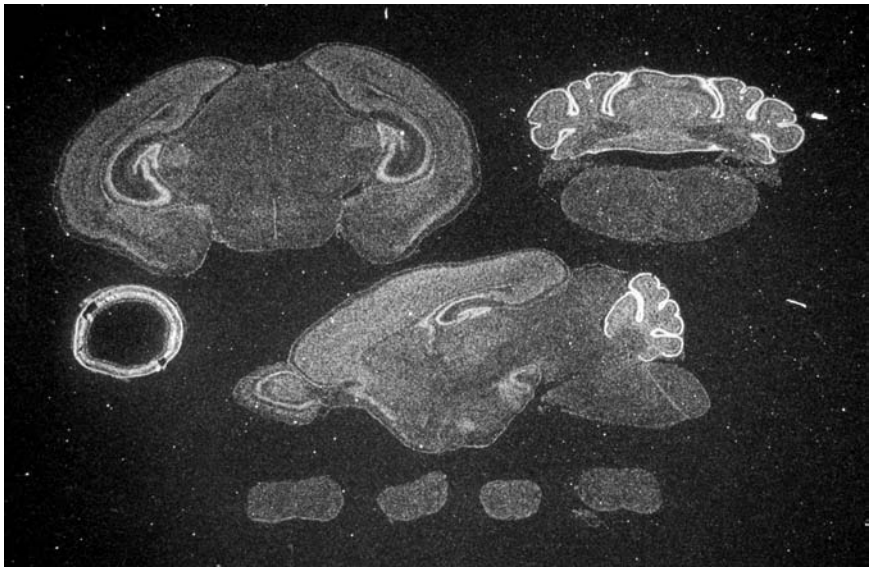

Figure 1 *In situ* hybridization darkfield images of a sagittal section close to the midline and coronal sections of a C57BL/6J postnatal day 7 mouse brain, showing mRNA distribution of the ataxia telangiectasia mutated homolog (*ATM*) gene from GENSAT.

tion of genes that mark specific regions of the brain, (2) identification of genes that mark specific cell types, and (3) identification of genes that reveal patterns of axonal connections in the emerging neural circuitry of the immature CNS.

Brain Gene Expression Map

The Brain Gene Expression Map is focused on the mouse brain and is housed at St. Jude Children's Research Hospital in Memphis, Tennessee and is associated with GENSAT. This database contains gene expression patterns from mouse nervous system tissues at four time points throughout brain development, including embryonic day 11.5, embryonic day 15.5, postnatal day 7, and adult day 42. Using high-throughput *in situ* hybridization, there are gene expression patterns from approximately 2850 genes, available in a searchable database. The database includes darkfield images from radioactive probes, reference cresyl violet-stained sections, the complete nucleotide sequence of the probes used to generate the data, and all the information required to allow users to repeat and extend the analyses. The database is linked to PubMed, LocusLink, Unigene, and the Gene Ontology Consortium housed at the National Center for Biotechnology Information (NCBI) in the National Library of Medicine in Bethesda, Maryland.

Allen Brain Atlas

The ABA is an interactive, genome-wide image database of gene expression in the mouse brain. It is based on automated high-throughput *in situ* hybridization of adult 8-week-old male C57BL/6J mouse brains. The platform used for the ABA utilizes a nonisotopic approach, with digoxigenin-labeled nucleotides incorporated into a riboprobe produced by *in vitro* transcription. Colorimetric detection of bound probe is generated by the alkaline phosphatase substrates nitroblue tetrazolium (NBT) and bromochloroindolyl phosphate (BCIP), which produce a blue/purple particulate reaction product that allows cellular localization of the signal. In addition to the *in situ* hybridization data, the project is generating reference atlases of the mouse brain that include both coronal and sagittal planes. This Nissl-stained reference facilitates annotation of the *in situ* hybridization data in the anatomical context. In order to view the *in situ* hybridization data, the project has created a web-based tool that allows searching by gene name and visualizing the corresponding *in situ* hybridization data.

Brain Gene Expression Database

The Brain Gene Expression Database contains gene expression data for various physiological and pathological processes in the mouse brain. All of the data were obtained by adaptor-tagged competitive PCR, an advanced version of quantitative PCR. Assayed genes were selected from the Brain Expressed Sequence Tag (EST) Database containing 3656 genes. RNA samples were taken under unique biological conditions using the C57BL/6J inbred strain. The datasets include postnatal cerebellar and dentate gyrus development at postnatal days 2, 4, 8, 12, and 21, and week 6. The following hippocampal regions are also available at 6 weeks: CA1, CA2, and CA3. The database also analyzes several physiological and pathological processes, such as the effects of ischemia on gene expression in the hippocampus, aging in the whole brain, and antipsychotic drugs (haloperidol and chlozapine) in the frontal lobe.

Microarray-Based Brain Gene Expression Databases

TeraGenomics This database contains microarray gene expression patterns for 24 neural tissues covering the mouse CNS measured in three inbred mouse strains (C57BL/6J, 129S6/SvEvTac, and DBA/2J) using the Affymetrix U74Av2 array. The complete collection of extensively annotated gene expression data along with data mining, visualization, and filtering tools has been made available on a publicly accessible website. A large-scale database, called TeraGenomics, was built to house and provide access to all of the quantitative, brain region-specific gene expression data, along with quality control measures and metadata in accordance with the Minimal Information About a Microarray Experiment (MIAME) standard. The body of metadata contains 75 different sample annotation fields, which include information on sample origin, dissection protocols, sample preparation, and array hybridization parameters. In addition, the anatomical hierarchy of the Neuronames taxonomy has been included as a user-friendly query tool within the database. A reference document for each dissection was created, consisting of photographs, atlas figures, and bregma coordinates of each collected brain region along with step-by-step instructions to illustrate the exact methods used in tissue collection.

SymAtlas The Genomics Institute of the Novartis Research Foundation (GNF) has a publicly available microarray database composed of expression profiles for 79 human tissues (20 brain regions), 61 mouse tissues (10 brain regions), and approximately 15 rat brain structures in as many as four different rat strains. These human, mouse, and rat gene expression atlases are available online and can be accessed through the SymAtlas web application. The web application provides searching and visualization by

Table 1 Publicly available brain atlases

Group	*Online address*
Gene Expression Nervous System Atlas (GENSAT)	http://www.ncbi.nlm.nih.gov/projects/gensat/
Allen Brain Atlas (ABA)	http://www.brain-map.org
Brain Gene Expression Map	http://www.stjudebgem.org
Brain Gene Expression Database	http://genome.mc.pref.osaka.jp/BGED/
TeraGenomics	http://www.barlow-lockhartbrainmapnimhgrant.org/
SymAtlas	http://symatlas.gnf.org/SymAtlas/

keyword, accession number, gene symbol, genome interval, sequence, and expression pattern. Queries can be partitioned by expression location and fold change above or below the median to assist in identifying genes with similar expression patterns. All of the Affymetrix datasets are also available for download. Table 1 displays online addresses of these large-scale gene-expression-based brain atlases.

Molecular Relationships within the Mammalian Brain

Using the aforementioned resources and tools, we and other neuroscientists have set out to understand how regional gene expression patterns in the brain are related to anatomical brain architecture and organization, and to identify relationships between brain regions based on both shared and restricted gene expression patterns. The existing framework to organize the mammalian nervous system is based on studies of anatomy, embryology, and evolution. It is now possible to begin to understand the mammalian brain from a molecular viewpoint.

In developmental biology, gene expression patterns for small groups of genes have been used to identify particular brain regions during embryogenesis. Dynamic changes in expression patterns have also been used to understand functional relationships between brain regions during development. Distinct neural cell precursors form at different positions along the anteroposterior (AP) and dorsoventral (DV) axes of the developing neural epithelium, which is critical for the establishment of a structured pattern of differentiated neurons. The concept of an established genetic pattern during embryogenesis has been strengthened by the expression of genes that mark morphogenetic fields during brain development, such as homeobox transcription factor genes. Spatially restricted expression of transcription factors specifies regional identity in the developing nervous system. The expression of these transcription factors underlies a crucial stage in neuronal patterning: the partitioning of the neural epithelium into domains with distinct identities that later form different groups of neural cell types.

The role of transcription factors in the developing mouse CNS has been studied using *in situ* hybridization. As mentioned previously, transcription factors are known to play a critical role in brain development by directing the formation of neurons and glia from progenitor cells. The aim of a recent study was to identify transcription factor gene expression patterns that are spatially and temporally restricted within the brain. Gene expression of 1174 transcription factors was visualized at embryonic days 10.5 and 13.5 and postnatal day 0, as well as on sections through the postnatal cerebellum at days 7, 15, and 21. The *in situ* hybridization data showed that gross anatomical diversity based on seven general areas within the CNS could be described by the expression of specific transcription factors in specific brain areas. These seven areas included the cortex, striatum, thalamus, hypothalamus, midbrain, hindbrain, and spinal cord. This study focused on embryogenesis and found that transcription factors known to play a role in developmental patterning, such as homeobox and forkhead genes, had the most restricted expression and were the most informative about anatomical diversity within the brain.

Whole-genome microarray expression data now suggest that the imprinted genetic program established during embryogenesis is still evident in the mature brain. Using gene expression patterns measured from 24 spatially registered (based on bregma coordinates) neural tissues covering the mouse CNS, it was shown that the adult brain demonstrates a transcriptional pattern consistent with where those tissues were derived embryologically. In addition to this imprinted genetic program, significant numbers of embryonic patterning and homeobox genes with region-specific expression in the adult nervous system were observed (see Figure 2). Several studies of the developing brain have demonstrated that similar sets of embryonic patterning and homeobox genes are used to establish a particular anatomical region and to maintain the cell–cell relationships of the differentiated region. The global gene expression pattern relationships between different anatomical regions and the nature of the observed region-specific genes suggest that the adult brain retains a degree of overall gene expression

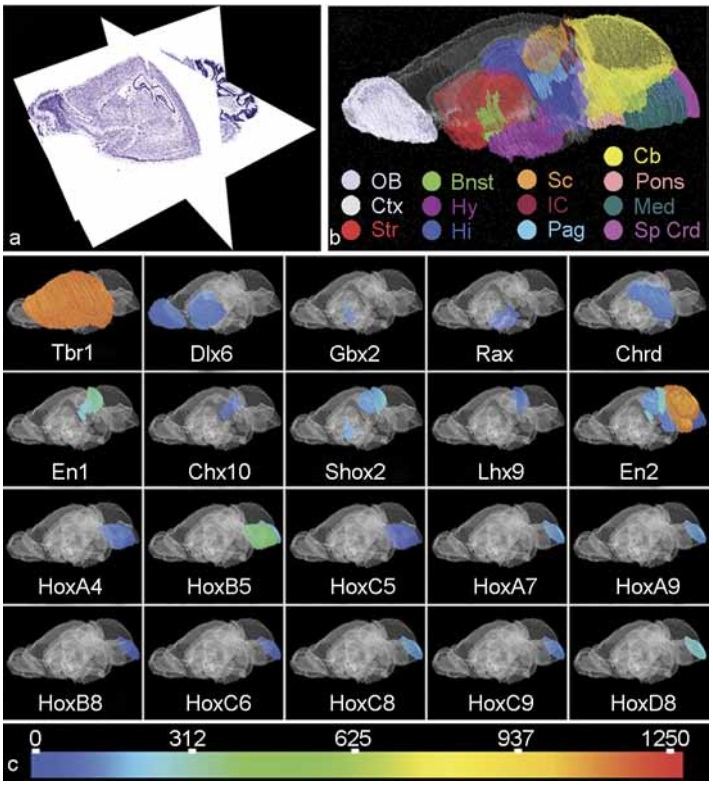

Figure 2 Genes with region-specific expression patterns function in development, pattern specification, and morphogenesis. (a) Reference brain atlas displayed in the three orthogonal planes. This atlas of Nissl-stained C57BL/6N mouse brain comprises 462 coronal sections at 30 μm thickness, digitized at a resolution of 1.3 μm per pixel. The sagittal and horizontal planes are 'virtual' sections dynamically constructed from the coronal sections. (b) Three-dimensional atlas of brain regions. Specific brain regions along the rostrocaudal neuraxis are color coded. OB, olfactory bulb; Ctx, cerebral cortex; Str, striatum; Bnst, bed nucleus of stria terminalis; Hy hypothalamus; Hi, hippocampus; SC, superior colliculus; IC, inferior colliculus; Pag, periaqueductal gray matter; Cb, cerebellum; Med, medulla; Sp Crd, spinal cord. (c) The expression levels of some of the homeobox and other embryonic patterning genes expressed in the adult mouse brain are digitally represented for each region. Adapted from Zapala MA, Hovatta I, Ellison JA, et al. (2005) Adult mouse brain gene expression patterns bear an embryologic imprint. *Proceedings of the National Academy of Sciences of the United States of America* 19: 10357–10362, with permission.

patterning, established during embryogenesis, that is important for regional specificity and for the functional relationships between brain regions in the adult.

There have also been smaller scale studies to investigate the molecular architecture of specific brain nuclei in the mouse. In a previous regional analysis of the amygdala, *in situ* hybridization had revealed that the majority of the genes specifically expressed in the amygdala displayed intra-amygdaloid expression boundaries corresponding to neuroanatomically defined subnuclei. Microarray data further confirmed these findings in that the gene expression patterns of specific amygdaloid nuclei were found to respect the ontogenetic origins of the subnuclei which derive embryologically from both pallial and subpallial structures. Similarly, the study of specific hippocampal subregions (CA1, CA2, and dentate gyrus) found that the relative enrichment and absence of genes in the hippocampal subregions support the conclusion that there is a molecular basis for the previously defined subregions, and these genes could define unique functions of these hippocampal subregions.

It has also become possible to examine anatomically distinct regional brain gene expression patterns in humans. For example, gene expression differences in the cerebellum and in the cortical areas in humans have been studied. Functional classification using gene ontology tools identified functional gene families enriched in the cerebellum and cerebral cortex. Genes with an enriched cortical gene expression pattern belong to categories such as 'calmodulin binding,' 'brain development,' 'receptor protein kinase,' and 'peptide hormones.' Several genes identified as being specifically enriched in the cortical areas have been previously implicated in psychiatric disorders, such as regulator of G-protein signaling 4 in schizophrenia; neuropeptide Y in bipolar disorder; cholecystokinin in depression; somatostatin in mania, schizophrenia,

and Alzheimer's disease; and 5-hydroxytryptamine receptor 2A in major depression and suicide.

Conclusions

There are several ongoing projects which have generated high-quality, spatially specific brain gene expression datasets. These molecular mapping efforts have provided invaluable information about specific gene expression patterns in anatomically distinct brain nuclei. They form a basis for other studies that aim to discover previously unknown functional elements in the brain, such as circuits involved in clinically important phenotypes. It is now possible to begin integrating and leveraging these unique data repositories to gain new insights into the molecular underpinnings of complex brain functions. Thus, it appears that not only are we closer to understanding the molecular anatomy of the mammalian brain, but we are also poised to begin to comprehend how it regulates complex behaviors or how disturbances in brain function manifest directly as specific neuropathologies.

See also: Genetic Influence on CNS Gene Expression: Impact on Behavior; Genomic Disorder and Gene Expression in the Developing CNS; Genomics of Brain Aging: Nuclear and Mitochondrial Genomes; Hippocampus: Molecular Anatomy; Neuroanatomy Methods in Humans and Animals; Psychiatric Genomics and Expression Profiling.

Further Reading

Bonaventure P, Guo H, Tian B, et al. (2002) Nuclei and subnuclei gene expression profiling in mammalian brain. *Brain Research* 943: 38–47.

Evans SJ, Choudary PV, Vawter MP, et al. (2003) DNA microarray analysis of functionally discrete human brain regions reveals divergent transcriptional profiles. *Neurobiology of Disease* 14: 240–250.

Gong S, Zheng C, Doughty ML, et al. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425: 907–908.

Gray PA, Fu H, Luo P, et al. (2004) Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* 306: 2255–2257.

Lein ES, Zhao X, and Gage FH (2004) Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput in situ hybridization. *Journal of Neuroscience* 24: 3879–3889.

Lein ES, Hawrylycz MJ, Ao N, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445: 168–176.

Lockhart DJ and Barlow C (2001) DNA arrays and gene expression analysis in the brain. In: Chin HR and Moldin SO (eds.) *Methods in Genomic Neuroscience*, pp. 143–169. Boca Raton: CRC Press.

Lockhart DJ and Barlow C (2001) Expressing what's on your mind: DNA arrays and the brain. *Nature Reviews Neuroscience* 2: 63–68.

Singh RP, Brown VM, Chaudhari A, et al. (2003) High-resolution voxelation mapping of human and rodent brain gene expression. *Journal of Neuroscience Methods* 125: 93–101.

Sunkin SM (2006) Towards the integration of spatially and temporally resolved murine gene expression databases. *Trends in Genetics* 22: 211–217.

Yamamori T and Rockland KS (2006) Neocortical areas, layers, connections, and gene expression. *Neuroscience Research* 55: 11–27.

Zapala MA, Hovatta I, Ellison JA, et al. (2005) Adult mouse brain gene expression patterns bear an embryologic imprint. *Proceedings of the National Academy of Sciences of the United States of America* 19: 10357–10362.

Zirlinger M, Kreiman G, and Anderson DJ (2003) Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid subnuclei. *Proceedings of the National Academy of Sciences of the United States of America* 98: 5270–5275.

Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines

J Masson and M Hamon, INSERM – Université Pierre et Marie Curie, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Monoamines are released at the synapse to exert their actions by activating specific receptors. The termination of synaptic transmission occurs through different mechanisms: degradation of the released monoamine, desensitization of the target receptors, and reuptake of monoamine by specific neuronal transporters. Monoamine transporters share several common features, particularly their dependence on Na^+ and Cl^- ions. Their encoding genes were cloned in several species (rat, human, bovine, ovine, and mouse) in the 1990s. They all belong to the same superfamily of genes (the SLC6 family), which give rise to proteins of ~600–620 residues termed DAT, NET, and SERT for dopamine, norepinephrine, and serotonin transporters, respectively. Due to the absence of a signal sequence, the N-terminus and C-terminus of these proteins are proposed to reside within the cytoplasm. To date, all experimental data have confirmed this inference, and further major progress in the knowledge of the molecular organization of Na^+/Cl^-dependent transporters of the SLC6 family has been made with the crystallization of the bacterial (from *Aquifex aeolicus*) leucine transporter. Thus, all monoamine transporters share the same topology with 12 transmembrane domains, the N-terminal and C-terminal regions located within the cytoplasm, and a large second extracellular loop endowed with N-glycosylation sites. Conservation of amino acid sequences among the three monoamine transporters seems to be highest within the putative transmembrane domains, whereas the least conserved regions are at the N-terminus and the C-terminus.

Amino acid sequence analysis of SERT, DAT, and NET proteins has revealed the presence of numerous serine (S), threonine (T), and tyrosine (Y) residues that could serve as phosphorylation sites, and many of these residues are present within consensus sequences for protein kinase C (PKC), protein kinase A (PKA), and calcium/calmodulin-dependent protein kinase (Figure 1), suggesting that second messengers play a role in posttranslational regulation of monoamine transporters. Extensive investigations aimed at identifying residues involved in the binding of monoamine to its specific pocket within the respective transporter identified residues within the first two transmembrane domains in the case of 5-HT binding into SERT. However, site-directed mutagenesis showed that substitution of amino acids at various locations within SERT (Y95, D98, I172, F335, F341, S438, and G442) affected the transporter affinity for 5-HT, indicating that the monoamine binding pocket involves several transmembrane domains.

The substrate specificity of monoamine transporters and their respective sensitivity to selective inhibitors have been examined in brain preparations and in heterologous systems with recombinant transporters. Although each monoamine is a substrate for its cognate transporter, it has been shown that both DAT and NET can transport dopamine as well as norepinephrine, whereas SERT transports only 5-HT.

This article focuses on SERT, with the aim of showing how cytokines regulating its expression and/or activity can provoke depressive symptoms. In brain, SERT is expressed only in serotonergic neurons, and immunolabeling has shown that it is targeted all over the neuronal plasmic membrane, from the somatodendritic compartment to the axonal varicosities and terminals. It is important to note that, in humans, SERT expressed in neurons, platelets, and placenta correspond to the same protein encoded by a single gene. In the following sections, we present and discuss results obtained in brain, platelets, as well as cell lines derived from placental choriocarcinoma.

SERT Is a Target for Antidepressant Drugs

Soon after radioligands became available for the quantitative labeling of SERT, it was established that uptake sites for 5-HT are generally downregulated in platelets from patients suffering from depression and in postmortem brain samples from depressed patients and suicide victims. On the other hand, the blockade of 5-HT transport by antidepressants appeared as a key pharmacological action to relieve depression, and selective serotonin reuptake inhibitors (SSRIs) are currently the most frequently used antidepressant drugs worldwide.

The cloning of the SERT encoding cDNA by Blakely et al. in 1991 allowed a more precise characterization of this protein at the molecular level and opened a new field of investigation into the identification of the binding pocket of specific inhibitors, such as fluoxetine, paroxetine, sertraline, fluvoxamine, and citalopram, within the SERT protein.

```
DAT   -----------MSKSKCSVGPMSSV----VAPAKESNAVGPREVELILVKEQNGVQLTNSTLINPPQTP----VEAQERETWSKKID
SERT  METTPLNSQKVLSECKDREDCQENGVLQKGVPTTADRAEPSQISNGYSAVPSTSAGDEASHSIPAATTTLVAEIRQGERETWGKKMD
NET   ---------MLLARMKPQVQPELGG----ADQLPEQPLRPCKTADLLVVKERNGVQCLLAS------QD----GDAQPRETWGKEID
                  *                                                                  **** *  *

              TM1                                TM2                                        TM3
DAT   FLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVP YLLFMVIAGMPLFYMELALGQFNREGAAGVW-KICPVLKGVGFTVILISFYVG
SERT  FLLSVIGYAVDLGNIWRFPYICYQNGGGAFLLP YTIMAIFGGIPLFYMELALGQYHRNGCISIWRKICPIFKGIGYAICIIAFYIA
NET   FLLSVVGFAVDLANVWRFPYLCYKNGGGAFLIP YTLFLIIAGMPLFYMELALGQFNREGAATVW-KICPFFKGVGYAVILIALYVG
      ***** * **** * ***** ** ******* **        * ***********  *  *    *  **** ** *     *  *

DAT   FFYNVIIAWALHYFFSSFTMDLPWIHCNNTWNSPNCSDAHASNSSD--GLGLNDTFGTTPAAEYFERGVLHLHQSRGIDDLGPPRWQ
SERT  SYYNTIIAWALYYLISSLTDRLPWTSCTNSWNTGNCTNYF------AQDNITWTLHSTSPAEEFYLRHVLQIHQSKGLQDLGTISWQ
NET   FYYNVIIAWSLYYLFASFTLNLPWTNCGHAWNSPNCTDPKLLNASVLGDHTKYSKYKFTPAAEFYERGVLHLHESSGIHDIGLPQWQ
        ** **** * *   * *  ***  *   **  **                  ** *   * **  * * *  * *   **

              TM4                    TM5                                                TM6
DAT   LTACLVLVIVLLYFSLWKGVKTSGKVVWITATMPYVVLTALLLRGVTLPGAMDGIRAYLSVDFYRLCEASVWIDAATQVCFSLGVGF
SERT  LTLCIVLIFTVIYFSIWKGVKTSGKVVWVTATFPYIVLSVLLVRGATLPGAWRGVVFYLKPNWQKLLETGVWVDAAAQIFFSLGPGF
NET   LLLCLMVVIVVLYFSLWKGVKTSGKVVWITATLPYFVLFVLLVHGVTLPGASNGINAYLHIDFYRLKEATVWIDAATQIFFSLGAGF
*  *        *** ************ *** ** **  **  * *****  *   **      * *  ** *** *  **** **

                                  TM7                                                TM8
DAT   GVLIAFSSYNKFTNNCYRDAIITTSINSLTSFSSGFVVFSFLGYMAQKHNVPIRDVATD-GPGLIFIIYPEAIATLPLSSAWAAVFF
SERT  GVLLAFASYNKFNNNCYQDALVTSVVNCMTSFVSGFVIFTVLGYMAEMRNEDVSEVAKDAGPSLLFITYAEAIANMPASTFFAIIFF
NET   GVLIAFASYNKFDNNCYRDALLTSTINCVTSFISGFAIFSILGYMAHEHKVKIEDVATE-GAGLVFVLYPEAISTLSGSTFWAVLFF
*** ** ***** **** **  *   *  ** * ***  *  ****          **   *  * *  * ***     *   *  **

                                                  TM9                    TM10
DAT   LMLLTLGIDSAMGGMESVITGLVDEF-QLLHRHRELFTLGIVLATFLLSLFCVTNGGIYVFTLLDHFAAGTSILFGVLIEAIGVAWF
SERT  LMLITLGLDSTFAGLEGVITAVLDEFPHIWAKRREWFVLIVVITCVLGSLLTLTSGGAYVVTLLEEYATGPAVLTVALIEAVAVSWF
NET   LMLLALGLDSSMGGMEAVITGLADDF-QVLKRHRKLFTCAVTLGTFLLAMFCITKGGIYVLTLLDTFAAGTSILFAVLMEAIGVSWF
***  ** **   * * ***   * *       *  *         *      * ** ** ***   * *   *   * **  * **

                                      TM11                                TM12
DAT   YGVQQFSDDIKQMTGQRPNLYWRLCWKLVSPCFLLYVVVVSIVTFRPPHYGAYIFPDWANALGWIIATSSMAMVPIYATYKFCSLPG
SERT  YGITQFCSDVKEMLGFSPGWFWRICWVAISPLFLLFIICSFLMSPPQLRLFQYNYPHWSIVLGYCIGMSSVICIPTYIIYRLISTPG
NET   YGVDRFSNDIQQMMGFKPGLYWRLCWKFVSPAFLLFVVVVSIINFKPLTYDDYVYPPWANWVGWGIALSSMILVPAYVIYKFFSIRG
**   *      *    *   ** **   ** ***                  *  * *    *  *  **    * *  *   *  *

DAT   SFREKLAYAITPEKDHQLVDRGEVRQFTLRHWLLL
SERT  TLKERIIKSITPETPTE-IPCGDIRMNAV------
NET   SLWERVAYGITPENEHHLVAQRDVRQFQLRHWLAI
   *      ****          *

Homology : LOW AVERAGE GOOD
```

a

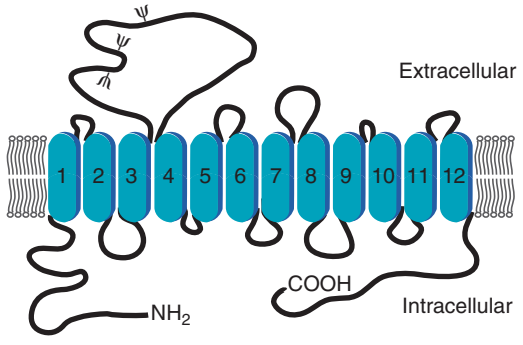

b

Figure 1 (a) Alignment of the amino acid sequences of monoamine transporters. The putative α-helical membrane spanning domains (TM1-12) are indicated by gray boxes. The degree of homology between the three transporters is indicated by a color code (blue to red) and conserved residues by stars. (b) Schematic structural organization (derived from hydropathy plot analysis) of monoamine transporters. Ψ, potentially *N*-glycosylated asparagine residue.

Blocking the SERT activity by SSRIs results in an immediate increase in extracellular 5-HT levels, and this is actually the first step, but not the only one, for the antidepressant effect of SSRIs.

Recent studies on the binding of citalopram to human SERT have revealed the presence of two binding sites, the first one involving the N-terminal transmembrane (TM) domains and the second one transmembrane domains (TM10 and TM11) close to the C-terminus portion of the transporter protein. There is evidence that the first site is directly involved in 5-HT uptake inhibition, whereas the second one exerts an allosteric influence, accounting for a reduced efficacy of escitalopram (the S enantiomer of citalopram) or paroxetine to inhibit 5-HT uptake when this site is occupied by R-citalopram. These findings suggest that in addition to SSRIs, it should be possible to develop drugs that produce fine-tuning of SERT activity through binding at this second allosteric site. Furthermore, positive allosteric influence on the first SSRI binding site can be expected to enhance the efficacy and accelerate the therapeutic action of SSRI antidepressants. Relevant research toward this goal is under way.

Consequences of the Genetic Deletion of SERT in Mice

The knockout gene strategy has been used to explore the consequences of the absence of SERT (during development and adulthood). As a model for whole life treatment with SSRIs, SERT knockout has been generated by homologous recombination involving the replacement of the second exon of the SERT gene by a neomycine cassette. This SERT–neomycine allele gives a mutant transcript coding for a mutant (truncated) protein unable to achieve its reuptake function in neurons. The loss of 5-HT reuptake function in SERT null mutants causes reduced clearance of extracellular 5-HT, leading to an increase (four- to sixfold) in basal levels of forebrain extracellular 5-HT. Furthermore, SERT knockout is associated with alterations in 5-HT neuronal firing and receptor function. In particular, the somatodendritic 5-HT_{1A} autoreceptor expressed in serotonergic dorsal raphe neurons is desensitized and downregulated, and the density of 5-HT_{1A} heteroreceptors expressed in forebrain regions also exhibits significant changes in SERT−/− mutants compared to paired wild-type mice.

General health, neurologic, and sensory functions were evaluated, and relevant data suggested that these SERT mutants are especially suitable for investigating physiopathological mechanisms underlying behavioral alterations related to 'emotion.'

It has been reported that null SERT mutants express exacerbated anxiety-like behavior in validated paradigms. This altered phenotype has notably been confirmed in relevant tests such as the elevated plus-maze, with its aversive open arms, and the dark/light box, with its 'anxiogenic' light compartment. It is important to note that this increase in anxiety-related behavior in SERT knockout mice has been observed only in SERT mutants with a C57B1/6J background. Indeed, behavioral alterations observed in knockout mice could be attributed not only to the deletion of the SERT gene but also to the background genes present in this mouse strain, indicating the complexity of gene interactions possibly associated with psychiatric diseases. Nevertheless, the key relationship between 5-HT and anxiety was further supported by the observation that SERT-overexpressing transgenic mice were much less anxious than paired wild-type animals, and that SERT blockade by the SSRI paroxetine reversed the low-anxiety phenotype of the mutants.

In addition, null SERT mutants were also investigated using validated depression-related paradigms. Results obtained with inescapable stress-related tests, such as the tail suspension test and the forced swim test, were contradictory depending on the genetic background of SERT−/− mutants. Nevertheless, as a whole, the reported data suggest that SERT−/− mutants have an exaggerated stress response, in line with the growing literature linking genetic variations in SERT function with emotional abnormalities.

On the other hand, null SERT mutants have also been described as less aggressive using the resident–intruder test. These behavioral traits are not completely similar to those observed after chronic systemic SSRI treatment or intracerebroventricular infusion of siRNA to knockdown SERT activity/expression in adult mice because the latter two treatments were regularly reported to induce clear-cut antidepressant-like effects in all validated tests. Subsequent studies that inhibited SERT selectively during the early postnatal period provided an explanation of the differences between constitutive SERT knockout and SERT inactivation in adults. Indeed, transient SERT blockade during early brain development was shown to produce paradoxical depression-related effects, which could then be subsequently reversed by SSRI antidepressants administered at the adult stage. These data highlighted the critical neurotrophic role of 5-HT for brain development, which was demonstrated more than 30 years ago by experiments showing that pharmacological alterations of 5-HT synthesis or inactivation at early stages provoke marked anatomofunctional alterations in brain. These data also illustrate that a constitutive

knockout is generally not homologous to long-term blockade of the corresponding gene product, enzyme/receptor/transporter, at the adult stage, notably because of the long-term consequences of the absence of the gene product during development.

SERT Gene Polymorphisms: Characteristics and Functional Consequences

The human SERT gene (SLC6A4; Figure 2) has been localized to chromosome 17q11.1–12, and its transcriptional activity is modulated by polymorphic repetitive elements located upstream of the transcription site, in the promoter region, and in the 5′ and 3′ untranslated regions. The SERT gene polymorphisms which have been most studied in depressed patients are (1) 5-HTTLPR in the 5′ untranslated region, described as L (long) and S (small) forms, which are associated with variable levels of SERT expression (in Caucasians, the 44-base pair insertion, which corresponds to the L variant of this polymorphism, is associated with higher expression of brain SERT compared to the S variant); and (2) STin2 variants in intron 2 (variable number of a 16- or 17-base pair tandem repeat).

In 1996, the first study of STin2 variants showed that STin2.9 allele (nine copies of the variable number tandem repeat (VNTR)) was more frequent in the unipolar depressive patient group. However, in other studies, unipolar affective disorder appeared to be associated with the STin2.12 allele (12 copies of VNTR). In fact, subsequent studies on unipolar and bipolar depressive disorder patients did not support any association with STin2.12. Interestingly, SERT gene transcription has been proposed to be determined by the combinatorial action of different promoter domains in intron 2, with the number of copies of VNTR markedly influencing the effects of specific transcription factors such as Y box binding protein 1 and CCTC-binding factor.

Much more data have been obtained on the polymorphism in the promoter region (5-HTTLPR), which indicate that this polymorphism is related to the tissue/cell levels of SERT mRNA. Evidence indicates that the distinctive structure of 5-HTTLPR gives rise to a specific DNA secondary structure that directly influences the transcriptional activity of the SERT gene promoter. When fused to a luciferase reporter gene and transfected into human SERT expressing cell lines, the short and long 5-HTTLPR variants differentially modulate transcriptional activity of the SERT gene promoter. The L allele confers a higher level of SERT mRNA expression than the S allele. The latter short allele was initially suggested to be associated with affective disorder risk, but subsequent results were contradictory. Thus, a few studies emphasized that the S allele was significantly more frequent in patients with unipolar depression. In particular, subjects with the biallelic SS promoter were found to more frequently suffer severe depression episodes in response to stressful life events than those with the biallelic LL promoter. However, no association between the L/S polymorphism and mood disorders was found in other studies. Indeed, marked heterogeneity between studies was evident, possibly reflecting differences in diagnosis, different control populations, and different ethnic origin. These factors unavoidably interfered with any reliable interpretation of the associations found in these studies.

On the other hand, the L allele of the 5-HTTLPR polymorphism was also found to be associated with a better and faster response to SSRIs in depressed patients in some but not all studies. Further investigation is needed to clarify the relationships between the SERT genotype, the susceptibility to affective disorders, and the response to serotonergic agents.

SERT Regulation

The regulation of any protein activity in cells mainly involves two types of molecular mechanisms. The first occurs after the protein has been translated from mRNA. It consists of phosphorylation/dephosphorylation of Ser, Thr, and Tyr residues by specific kinases and phosphatases. The second one concerns transcriptional regulation. Major improvements in methods for the quantification of specific mRNAs led to the characterization of gene promoter, transcriptional factors (activator and repressor), and their binding sites on genomic DNA and contributed to the current knowledge of SERT regulatory mechanisms.

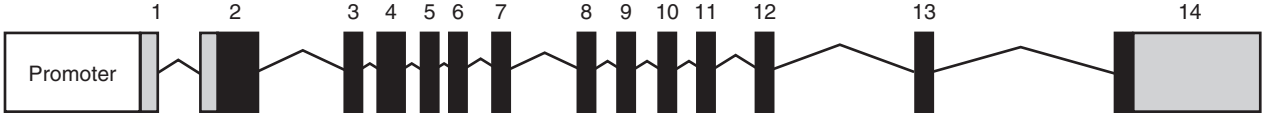

Figure 2 Organization of the human SERT gene and its 5′ flanking region on chromosome 17q11.2. The white box represents the SERT promoter containing the 5-HTTLPR polymorphism, with the short (S) or long (L) allele. Boxes depict exons (1–14). Black and gray boxes are coding and noncoding regions, respectively. Introns are represented by dashed lines. STin2 polymorphism is localized in the second intron between exons 2 and 3.

Posttranscriptional Regulation

SERT regulation by signaling pathways involving PKC, cyclic guanosine monophosphate (cGMP), and calmodulin has been the subject of numerous studies. It is well established that PKC or activators of PKC reduce 5-HT transport capacity (V_{max}), suggesting it may silence membrane-bound SERT or change its cell surface expression. Indeed, SERT is phosphorylated following PKC and PKA activation and protein phosphatase 2A inhibition (by okadaic acid). PKC-dependent phosphorylation of SERT shows a close temporal correlation with a reduction in SERT transport capacity. On the other hand, SERT substrates and antagonists can influence PKC-dependent SERT phosphorylation and surface redistribution.

Transcriptional Regulation

Transcriptional regulation of SERT was first studied in human placental choriocarcinoma, a cell line that endogenously expresses the transporter. In these cells, cholera toxin has been used to determine the role of cAMP as a molecule upregulating SERT. On the other hand, steroids are also involved in the regulation of SERT gene transcription. In particular, SERT mRNA is significantly downregulated by estrogen treatment, through the activation of nuclear estrogen receptor-beta, which acts as an inhibitory transcription factor on SERT gene promoter in serotonergic neurons. In contrast, treatment of immortalized human B-lymphoblastoid cells, which express the human SERT, with dexamethasone, a synthetic glucocorticoid receptor (GR) agonist, results in increased SERT expression. Transcriptional activation by GR depends on the repetitive GC-rich sequence located approximately 1.4 kilobases upstream of the SERT gene transcription site. In particular, the steroid effect is no longer observed in cells transfected with a deleted mutant reporter gene construct, which lacks the repetitive promoter sequence.

In addition, studies on transcriptional regulation of SERT gave rise to a number of reports concerning the effects of cytokines. The mechanisms involved and the functional/clinical relevance of the effects of cytokines on SERT are presented in the following sections.

Cytokines and Depression

Cytokines are regulatory proteins secreted by lymphocytes and other cell types. They play key roles in immune responses to inflammatory processes, but they also regulate growth, differentiation, and function of many cells. Cytokines are classified as proinflammatory (interleukin (IL)-1, IL-6, tumor necrosis factor-α (TNF-α), and interferon (IFN)) when they stimulate immune cells and as anti-inflammatory (IL-4 and IL-10) when they silence the immune system to decrease the production of further cytokines. Historically, the existence of relationships between cytokines and depression has been postulated from convergent clinical reports indicating that patients treated with some cytokines (IFN-α) can develop symptoms resembling depression.

Cytokine Levels in Depressed Patients

Numerous reports have noted that the plasma concentrations of proinflammatory cytokines (IL-1β, IL-6, IFN-γ, and TNF-α) and the density of soluble cytokine receptors (IL-2 and IL-6) in peripheral blood mononuclear cells are increased in depressed patients. However, a few studies did not confirm these results. Discrepancies among studies are possibly due to the evaluation of various subtypes of depression, differences in severity of the disease, limited sample sizes, and treatment with or without antidepressant medication. In addition, rhythmic changes in cytokine levels may also account for the variable data from one study to another. Nevertheless, immune activation has been proposed to be a marker for susceptibility to depression. This hypothesis is based on numerous clinical observations, such as marked increases in serum levels of soluble IL-2 receptor, IL-6, and IL-10 in patients suffering from chronic hepatitis C (CHC) infection who developed major depressive disorder in response to treatment with IFN-α.

Depressive Symptoms Occurring during Cytokine Treatment

In line with the clinical observations summarized in the previous section, a large body of evidence indicates that a major proportion of patients treated for infections or malignancies with proinflammatory reactions display 'sickness behavior' including neuropsychiatric symptoms, some of which are typical of depression. Depression usually occurs within the first 4 weeks of antiviral therapy, but its incidence is variable, depending on the disease being treated, the cytokine molecule and dose, and the patient's psychiatric history.

IFN is a proinflammatory cytokine with antiviral and anticancer effects. IFN treatment has been shown to increase the risk of major depressive disorder by 23–45%. Interestingly, of all the pathologies cured by IFN treatment – myelogenous leukemia, melanoma, renal carcinoma, and CHC – patients suffering from the latter are more prone to develop IFN-induced depression.

The development of clinically significant neuropsychiatric changes is also common during treatment

with IL-2 and may be difficult to manage. However, usually psychoaffective alterations occur with a rather long latency after starting the treatment and disappear once patients have been cured.

Reversal by Antidepressants of Depression Caused by Cytokine Treatment

The emergence of depressive symptoms in cytokine-treated patients may possibly lead to discontinuation of treatment. Fortunately, antidepressant drugs, notably SSRIs, are generally effective in treating this unbearable side effect.

An interesting example concerns a study that was performed on a group of CHC-suffering patients who had to be re-treated with IFN after a first therapy failure because of IFN-induced depression. These patients were psychometrically monitored during re-therapy for CHC and SSRI prophylaxis. The SSRI medication started 3 weeks before initiation of the antiviral treatment. This antidepressant therapy protected patients from the IFN-induced depression observed during the IFN therapy-naive treatment. This strategy has also been used and validated for efficacy in patients with malignant melanoma treated with high doses of IFN-α.

Immune stimulation produced during proinflammatory cytokine therapy has been shown to induce a decrease in brain 5-HT tone. This deficit in 5-HT appeared to be secondary to the stimulation of indeolamine-2,3-dioxygenase and GTP-cyclohydrolase which results in tryptophan degradation to kynurenine and neopterin production, thereby reducing the concentration of the precursor aromatic L amino acid for 5-HT synthesis (Figure 3).

These data suggest that the mechanism responsible for cytokine-induced depression involves 5-HT, and that the pharmacological blockade of SERT is sufficient to cure the depression (and even prevent the occurrence of the disease) through facilitation of 5-HT neurotransmission compensating for the cytokine-evoked deficit.

Effects of Cytokines in Animal Behavioral Models of Depression

To study the influence of cytokine treatment on depression, several groups examined the effects of cytokines on rodent behavior in validated experimental paradigms. In behavioral tests developed to evaluate depression-like behavior in rodents, such as the tail suspension test, sucrose consumption, and the force swimming test, acute as well as chronic administration of IFN-α induced depression-like responses. Thus, anhedonia, a depressive-like symptom assessed through a decreased free consumption of sucrose in rats, is observed following acute administration of IFN-α, and this behavioral deficit can be reversed by a 2-week antidepressant (desipramine or fluoxetine) treatment. In contrast, IFN-β and IFN-γ do not trigger such depression-like behaviors. However, contradictory results have also been reported suggesting that the effects of cytokines, especially IFN-α, critically depend on the experimental conditions of each laboratory.

Further studies demonstrated that IFN-α-induced depression-like symptoms could be attributed, at least in part, to a decrease in brain 5-HT neurotransmission. Indeed, as detailed in the following section, *in vitro* data indicated direct effects of cytokine on SERT expression and/or activity.

Cytokines Regulate SERT Expression and Activity

The first report on the identification of a physiologically occurring cytokine which regulates expression of the SERT gene was published in 1995. This study opened a new field of investigation on the roles of cytokines in the control of brain 5-HT neurotransmission through SERT modulations. It was found that IL-1β stimulated SERT transcription in human JAR choriocarcinoma cells and yielded up to a three- or fourfold increase in SERT mRNA levels, a 2.5-fold increase in membrane-bound SERT protein, and a 1.8-fold increase in 5-HT uptake. A minimal 4 h exposure to the cytokine was required to produce significant effects. Interestingly, IL-1β-mediated SERT upregulation did not involve cAMP, cGMP, nitric oxide, or ceramide signaling. Upregulation of SERT by IL-1β appeared to be associated with the activation of mitogen-activated protein (MAP) kinases and nuclear factor-kappa B (NF-κB p65 homodimer) in response to the stimulation of IL-1 receptor type I. Indeed, the following sequence was established: (1) IL-1β activates its specific receptor, which in turn (2) activates MAP kinase with a maximal effect after a 15 min exposure to the cytokine; then (3) NF-κB is activated (maximal effect after a 30 min treatment); and, finally, (4) SERT is upregulated (after a 4 h treatment). In these IL-1 receptor type 1-expressing JAR cells, NF-κB activation follows the classical pathway: NF-κB is present in the cytoplasm as an inactive complex with an inhibitory protein IκB. The complex dissociates after IκB ubiquitination, and free NF-κB enters the nucleus to activate transcription. In JAR cells, the inhibitory protein involved in IL-1β-induced SERT upregulation is IκBε. Interestingly, IL-1 receptor type 1 encoding mRNA is expressed over the entire midline raphe in the rat, so the sequence established in JAR cells probably occurs in brain serotonergic neurons.

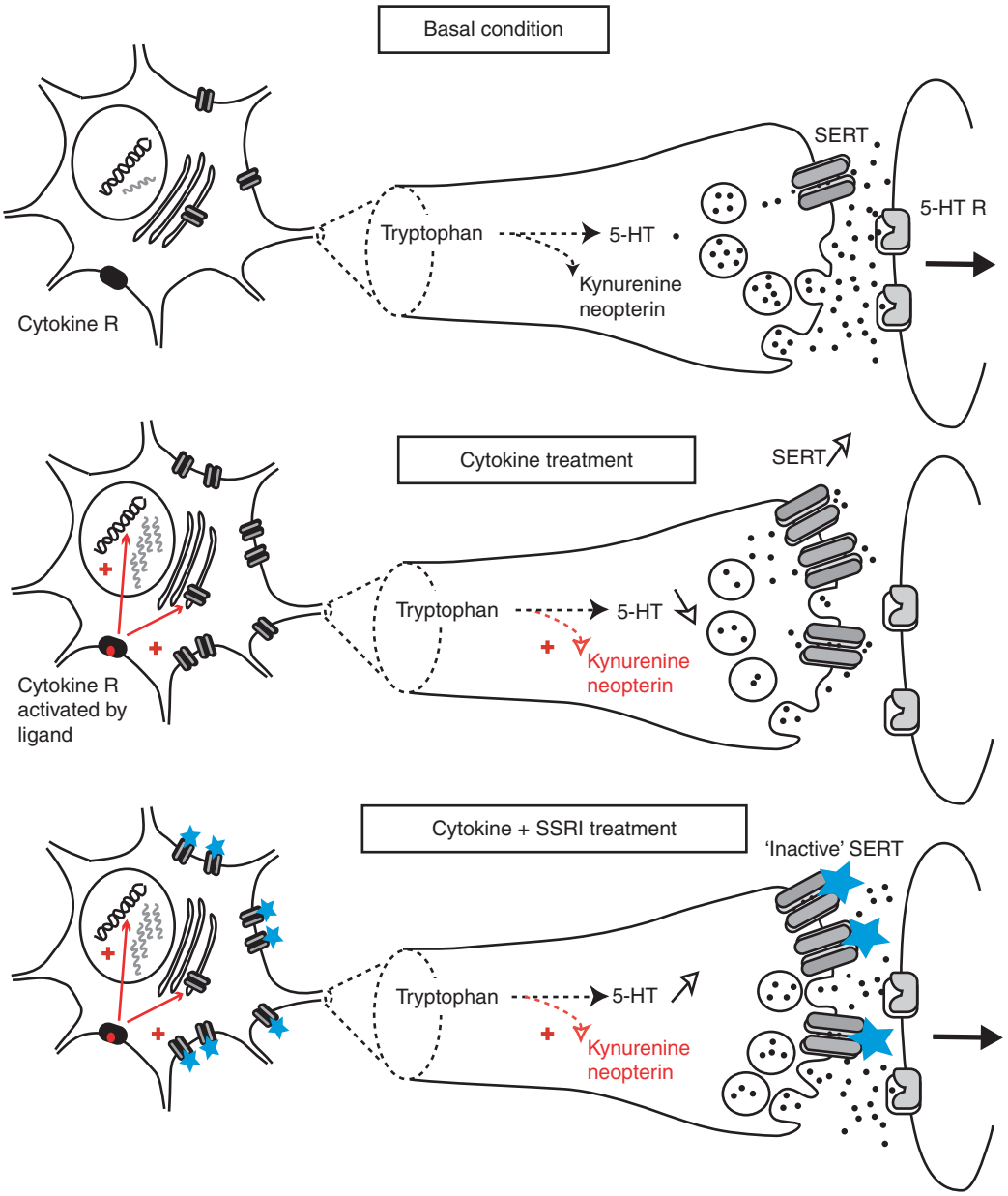

Figure 3 How cytokine binding onto 5-HT neurons influences serotonergic transmission. In basal condition, the released 5-HT activates 5-HT receptors, and receptor signaling proceeds until 5-HT is removed from the extracellular space by uptake through SERT localized on 5-HT neurons. During cytokine treatment, 5-HT synthesis is reduced because of the decrease in tryptophan, the 5-HT precursor (kynurenine and neopterin synthesis pathways are activated). Furthermore, SERT is upregulated (increase of mRNA and posttranscriptional activation of SERT protein), leading to an increase in 5-HT removal from the extracellular space and a decrease in serotonergic transmission. This hyposerotonergic activity can cause depressive symptoms in human and in animal models. Cytokine-induced 'depressive behavior' is reversed by SSRIs, which block 5-HT uptake, thereby restoring high levels of 5-HT in the extracellular space.

TNF-α was also found to enhance 5-HT uptake in JAR cells through an increase in maximal velocity (but no change of 5-HT affinity for SERT), which reached its maximum after a 48 h exposure to the cytokine. The presence of a TNF-α-responsive NF-κB element 62 base pairs upstream of the start of repetitive elements in the human SERT promoter strongly supports the idea that TNF-α-induced upregulation of SERT results from increased transcription of the SERT gene.

The interferons IFN-α and IFN-γ have also been described as SERT mRNA-regulating cytokines in another cell line, the human placental choriocarcinoma BeWo cells. Indeed, SERT expression increased after a minimal 3 h exposure to either interferon, and the maximal effect was observed after a 6 h

cytokine treatment. BeWo cells express receptors for both interferons and their signaling pathways through Janus kinase-STAT, and consensus binding sites of STAT family member were detected in the 5′-flanking region of the mouse SERT gene. It can therefore be inferred that the Janus kinase-STAT pathway is probably implicated in interferon-induced SERT upregulation in the mouse. Thorough analysis of the promoter of the murine SERT gene has resulted in the identification of several binding sites for transcription factors: cAMP-responsive element binding protein 1 (CREBP1) or its heterodimer with c-Jun (CREBP1CJUN), interferon regulatory factors 1 and 2 (IRF1 and IRF2), and transducers and activators of transcription (STAT), which mediate signal transduction triggered by interferon receptors. Accordingly, the various mechanisms underlying IFN-α and/or IFN-γ upregulation in BeWo cells probably also occur in rodents.

Direct evidence showing that cytokines increase SERT activity in serotonergic neurons has been reported by Zhu et al., who noted that IL-1β and TNF-α stimulate 5-HT uptake in the rat embryonic raphe cell line RN46A and in mouse midbrain and striatal synaptosomes. The effect of IL-1β is completely prevented by pharmacological blockade of p38 MAP kinase, indicating that IL-1β-induced p38 MAP kinase activation mediates the stimulatory effect of this cytokine on SERT. In contrast, only partial inhibition was noted with TNF-α, suggesting that additional non-p38 MAPK-mediated mechanisms participate in the stimulatory effect of this cytokine on neuronal 5-HT uptake. Unlike that found in cell lines, only short treatment (10–20 min) with IL-1β and TNF-α was needed to maximally increase 5-HT uptake in brain synaptosomes, indicating that this effect does not involve gene transcription but, rather, posttranslational regulatory steps. Indeed, this effect is short lasting, resolving within 20–60 min.

Interestingly, all cytokines with the capacity to upregulate SERT expression and/or activity are proinflammatory cytokines (IL-1β, TNF-α, and INF-γ). Conversely, the anti-inflammatory cytokine IL-4 has been found to induce a reduction of 5-HT uptake in immortalized B lymphocytes, which endogenously express SERT. Additional investigations demonstrated that IL-4-induced SERT downregulation is more pronounced in cell lines homozygous for the 5-HTTLPR long allele, corresponding to the high SERT activity. Because IL-4 is a potent inhibitor of IL-1β and TNF-α production *in vivo,* this anti-inflammatory cytokine might reduce 5-HT uptake through both its direct inhibitory effect on SERT expression and the blockade of any potential SERT upregulation by the proinflammatory cytokines.

Conclusion

Among monoamine transporters of the SLC6 family, the serotonin transporter, SERT, has been described for decades as a key protein for the physiopathology and treatment of depression. Antidepressant drugs target SERT; they act by blocking 5-HT transport activity, thereby increasing extracellular 5-HT and compensating for the deficit in serotonin release associated with the pathology. The effect of SSRI antidepressants proceeds in two key phases: First, 5-HT is increased in the extracellular space and, second, after a delay of 2 or 3 weeks, the sustained high level of extracellular 5-HT causes the desensitization of 5-HT_{1A} autoreceptors that are responsible for the negative feedback control of serotonergic neuron firing. The resulting suppression of this inhibitory control, combined with 5-HT reuptake blockade, produces a marked elevation of extracellular 5-HT concentration, which probably accounts for the antidepressant action of SSRIs. Interestingly, cytokines are able to differentially regulate SERT expression depending on their pro- or anti-inflammatory profile. Proinflammatory cytokines, which have the capacity to provoke depressive symptoms, enhance SERT activity, thereby causing a reduction in 5-HT extracellular levels and 5-HT neurotransmission (Figure 3). Cytokine-induced depression is reversed by SSRIs, confirming that the psychiatric pathology is causally related to a deficit in serotonergic transmission due to enhanced 5-HT reuptake.

Thus, the state of an immune response may be communicated to the central nervous system and lead to various adaptive changes in the plethora of brain functions that are under the influence of the 5-HT system, such as mood control.

See also: Monoamines: Human Brain Imaging; Monoamines: Release Studies; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors; Serotonin-Related Psychedelic Drugs; Stress, Cytokines and Depressive Illness; Trace Monoamines and Receptors in Mammalian CNS.

Further Reading

Asnis GM and De La Garza R II (2005) Interferon-induced depression: Strategies in treatment. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 29: 808–818.

Blakely RD, Berson HE, Fremeau RT Jr., et al. (1991) Cloning and expression of a functional serotonin transporter from rat brain. *Nature* 354: 66–70.

Caspi A, Sugden K, Moffitt TE, et al. (2003) Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* 301: 386–389.

Dunn AJ, Swiergiel AH, and de Beaurepaire R (2005) Cytokines as mediators of depression: What can we learn from animal studies? *Neuroscience & Biobehavioral Reviews* 29: 891–909.

Hamon M and Bourgoin S (2006) Pharmacological profile of antidepressants: A likely basis for their efficacy and side effects? *European Neuropsychopharmacology* 16: S625–S632.

Kekuda R, Leibach FH, Furesz TC, Smith CH, and Ganapathy V (2000) Polarized distribution of interleukin-1 receptors and their role in regulation of serotonin transporter in placenta. *Journal of Pharmacology & Experimental Therapeutics* 292: 1032–1041.

Lesch KP (2001) Variation of serotonergic gene expression: Neurodevelopment and the complexity of response to psychopharmacologic drugs. *European Neuropsychopharmacology* 11: 457–474.

Masson J, Sagné C, Hamon M, and El Mestikawy S (1999) Neurotransmitter transporters in the central nervous system. *Pharmacological Reviews* 51: 439–464.

Mossner R, Daniel S, Schmitt A, Albert D, and Lesch KP (2001) Modulation of serotonin transporter function by interleukin-4. *Life Sciences* 68: 873–880.

Mossner R, Heils A, Stober G, Okladnova O, Daniel S, and Lesch KP (1998) Enhancement of serotonin transporter function by tumor necrosis factor alpha but not by interleukin-6. *Neurochemistry International* 33: 251–254.

Neubau HA, Hansen CG, and Wiborg O (2006) Dissection of an allosteric mechanism on the serotonin transporter: A cross-species study. *Molecular Pharmacology* 69: 1242–1250.

Ramamoorthy S, Ramamoorthy JD, Prasad PD, et al. (1995) Regulation of the human serotonin transporter by interleukin-1 beta. *Biochemical & Biophysical Research Communications* 216: 560–567.

Thakker DR, Natt F, Husken D, et al. (2005) siRNA-mediated knockdown of the serotonin transporter in the adult mouse brain. *Molecular Psychiatry* 10: 782–789.

Yamashita A, Singh SK, Kawate T, Jin Y, and Gouaux E (2005) Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. *Nature* 437: 215–223.

Zhu CB, Blakely RD, and Hewlett WA (2006) The proinflammatory cytokines interleukin-1beta and tumor necrosis factor-alpha activate serotonin transporters. *Neuropsychopharmacology* 31: 2121–2131.

Monoamines

S Govoni, University of Pavia, Pavia, Italy

© 2009 Elsevier Ltd. All rights reserved.

Overview

Progressive increases in the elderly population worldwide, and the expected greater population increases in this age group in the developing countries in the next decades, create a sense of urgency to study the mechanisms of aging. Both at systemic and brain levels, it is of great importance to understand the neurochemical changes that take place, and the relationship of these changes with the clinical manifestations that limit physical and mental performance in late life. Among the age-associated changes, those involving brain monoamines, particularly dopamine, may represent strategic targets for interventions aimed at sustaining normal mood, motor, and cognitive functions.

Studies on brain neurotransmitter changes during aging trace back into the late 1960s. The not always consistent descriptions of age-associated alterations of monoamine content, synthesis, and receptors in various areas of the aging brain were actively pursued up to the late 1980s. Since then, the interest has shifted toward the investigation of age-associated neurodegenerative diseases and protein–protein interactions, rather than neurotransmitter dynamics. What is still valid at more than 15 years of distance from early research, and what deserves our continuous attention? What are the lessons learned from the first studies? Early studies on monoamines changes during aging were motivated by the seminal observations by Arvid Carlson and Olef Hornykiewcz, who focused on the role of dopaminergic transmission in motor function and its derangement in Parkinson's disease, an age-associated pathological condition. This led to the concept that it was possible to define the neurochemical fingerprints of neurological and psychiatric diseases and also of the functional motor and cognitive decline of the elderly, and to treat them by normalizing the function of the involved neurotransmitters, either by replacement therapy or by antagonizing their receptors in case of overfunction. This view, although obviously not complete, provided the rationale for the development of several drugs useful for neuropsychiatric conditions and promoted further studies on the aging brain. Both animal and human brains were subsequently investigated.

All of the data published on the age-associated changes of neurotransmitters in humans (in part summarized in Table 1) are potentially biased by the selection of and the variability of the investigated individuals and by the presence of age-associated pathologies of presymptomatic conditions. Indeed, neurotransmitter studies in humans have been performed on tissues obtained at autopsy. Investigations using postmortem material are limited by many factors, including age, sex, disease state, drug history, postmortem delay, stability of the neurotransmitter's indices, and many others. Some studies were done measuring monoamines and their metabolites in the cerebrospinal fluid (CSF) or in plasma. However CSF analyses, in addition to being invasive, present several problems, including cervical to lumbar concentration gradients of the species measured, and the fact that it is not possible to attribute the observed changes to a specific brain area. Plasma levels of monoamine and their metabolite are, in general, a poor index of events occurring at brain level and may rather reflect changes in peripheral monoaminergic nerve terminals or in platelets, in the case of serotonin. In turn, animal studies, mostly, but not exclusively, on rodents, present not only species-but also strain-dependent variability and are biased by the rather artificial conditions of laboratory animal life. All of this has to be kept in mind when taking into consideration the published data and in particular when extrapolating from laboratory conditions to human aging.

Dopaminergic Neurotransmission

In the elderly, the decline in memory and cognitive abilities is of great concern, as is motor weakening, which increases with age. The dopaminergic system mediates motor abilities as well as cognition and emotion and may be especially vulnerable to aging. Consequently, the aging of this system has both clinical and vocational aspects. The reason for the selective vulnerability of dopaminergic transmission to the aging process is only partially understood. The phenomenon in the past has been attributed to an excess of oxidative stress in dopaminergic areas linked to the metabolism of dopamine (DA) itself. Interestingly, it has been shown that dopamine-induced toxicity is age dependent. Rats of various ages receiving dopamine infusion in the substantia nigra show behavioral and histological correlates of dopamine toxicity, as indicated by motor behavioral deficits and by a loss of tyrosine hydroxylase; the effects are more severe in middle-aged rats (15 months) than in young ones. More recently the attention has been shifted to the interaction between dopamine and parkin, a key

Table 1 Synopsis of changes in some monoamine neurotransmitter systems during brain aging[a]

Parameter	*Change in humans*	*Change in laboratory animals*
Dopaminergic neurotransmission		
Presynaptic parameters		
Dopaminergic cells in substantia nigra	−	
Tyrosine hydroxylase	−	
DOPA decarboxylase	−	
Homovanillic acid	−	−
Dopamine	−	−
Dopamine transporter		−
Postsynaptic parameters		
Dopamine D1 receptors		=/−
Dopamine D2 receptors	−	−/=
Dopamine-stimulated adenylate cyclase		−
Dopamine-inhibited adenylate cyclase		=
Catabolizing enzymes		
Monoamine oxidase B	+	
Noradrenergic neurotransmission		
Presynaptic parameters		
Noradrenergic cells in locus coeruleus	−	
Noradrenaline	−/=	−[b]
Dopamine β-hydroxylase	=	
Postsynaptic parameters		
β-Adrenergic receptors	− (in cerebellum)	−[c]
Serotonergic neurotransmission		
Presynaptic parameters		
Serotonin and/or serotonin metabolites	−/=/+	−[d]
5-Hydroxyindoleacetic acid	=/+	
Imipramine binding	−/=/+	
Postsynaptic parameters		
5-HT_{1A} serotonin receptors	−	−/=[e]
5-HT_{2A} serotonin receptors	−	−[f]
Serotonin transporter	−/=	

[a]Key to changes: −, decrease; +, increase; =, no change. The table provides an overview of several age-associated changes in monoamine neurotransmission. Besides dopaminergic transmission, displaying a rather consistent view, the results on noradrenergic and serotonergic transmission are more complex, brain area specific, and in line with the concept that normal brain aging is associated with subtle morphological and functional alterations rather than large-scale neuronal loss. If not indicated, laboratory animals include various strains of rats and mice.
[b]Impaired increased release of noradrenaline following immobilization stress in aged rats.
[c]Reduced magnitude of long-term potentiation induction in hippocampus following β-adrenergic receptor stimulation.
[d]In the pineal gland, Wistar rats.
[e]Monkey brain.
[f]Dog brain.

protein for the survival of neurons in the substantia nigra and locus coeruleus. In particular, it has been shown that parkin is a direct target of dopamine quinone *in vitro*, leading to a loss of parkin activity. Since dopamine oxidation occurs in aging brains, as indicated by an age-associated accumulation of neuromelanin, it is possible that this mechanism contributes to the degeneration of substantia nigra neurons over time. Also, age-associated changes in the activity of growth factors acting on nerve cells may have a role. Experiments on transgenic mice show that the age-related changes in nigrostriatal dopaminergic function are enhanced when one allele for brain-derived neurotrophic factor (BDNF) is deleted, thus supporting the view that changes in growth factor expression may alter the susceptibility of selected neuronal populations to the aging process (somewhat similar results were obtained for glial-derived growth factor and noradrenergic neurons; see later). Gender differences may also play a role. Data on aged Fisher 344 rats indicate that there are both age and sex differences in the brain's response to chemical lesioning using 6-hydroxydopamine, and that the compensatory or neuroprotective mechanisms are more efficient in the aged female brain.

A deficit in dopaminergic tone is present in the brain areas most densely innervated by dopaminergic neurons, and classical experiments in aged rodents made by Cotzias in the late 1970s suggested that fostering dopaminergic transmission by a prolonged treatment with 3,4-dihydroxy-L-phenylalanine (L-DOPA) was able to improve physical aspect, sexual behavior,

and longevity of treated animals. Characterization of the aging dopaminergic system in laboratory rodents (in decreasing order of frequency of use in aging studies: rats, mice, guinea pigs, rabbits) is based on detailed studies considering dopamine terminal synthesis, release, and uptake of dopamine, and of the ability of postsynaptic dopamine receptors to recognize dopamine and to trigger a correct signal transduction. Aging affects the number of dopamine receptors and their ability to stimulate adenylate cyclase and to promote cyclic adenosine monophosphate (cAMP)-stimulated protein phosphorylation through cAMP-dependent protein kinase A (PKA). In particular it has been shown that the dopamine-stimulated adenylate cyclase activity, which is functionally linked to dopamine D1 receptors, is significantly reduced in dopamine-rich areas. Most of the data have been obtained working with extrapyramidal dopaminergic areas such as the neostriatum, while less systematic data are available on other brain regions where dopamine terminals are present, such as nucleus accumbens, prefrontal cortex, hippocampus, and hypothalamus. The nigrostriatal system has been more intensively studied compared to other brain areas because several diseases that affect it have strong age-related incidences, particularly Parkinson's disease, the onset of which peaks after 50 years of age. The changes at receptor level are complex, depending upon the animal species or strain used and the ligand adopted. There are two families of dopamine receptors: D1 sites, which stimulate adenylate cyclase, and D2 sites, which inhibit adenylate cyclase or are independent of a cAMP-generating system. The number of binding sites for D1 receptor ligands has been reported to be unmodified in some rodent species (Sprague–Dawley rats, C57 mice) and reduced in other (Wistar and Fisher 344 rats). Selective D2 ligands, such as substituted benzamides, and the dopamine-inhibited adenylate cyclase (a dopamine D2 receptor-associated event) are unmodified. In contrast, the labeling of dopamine receptors using butyrophenones (the most popular ligand being spiroperidol) as ligands shows a reduced receptor density, suggesting that in the experimental conditions and animal species/strain used, butyrophenones label subsets of nondopaminergic receptors that are not labeled by benzamides and that show an age-associated decline.

Not only the binding sites for dopamine receptor ligands show age-associated changes, but also their ability to adapt to a challenge. As an example, it has been shown that in old rats the supersensitivity response (seen as an increased number of receptors) to chronic receptor blockade by the dopamine antagonist haloperidol is impaired. A reduced adaptive capacity of dopamine receptors in response to a chronic reduction in receptor activation has also been reported in mice. Since similar changes also take place in humans (see later), it may be predicted that failure of the compensatory receptor response, together with reduced presynaptic transmitter availability, may contribute to the greater sensitivity of the elderly to the extrapyramidal side effects of antipsychotic drugs.

In addition to the nigrostriatal system, also the limbic and hypothalamic dopaminergic systems are sensitive to aging. In particular, tuberoinfundibular dopamine (DA) transmission shows the greatest age-associated changes. In the animal model, dopamine concentrations decrease with age in the medial basal hypothalamus and in the median eminence. The increased secretion in senescent animals of prolactin represents the consequence of impaired hypothalamic dopaminergic transmission, which is under inhibitory dopaminergic control. Indeed, age-associated changes are commonly observed in the control of prolactin in laboratory rats and similarly in other species. In male and female rats of some strains, circulating prolactin levels become elevated in middle age. This observation supports the concept that hypothalamic dopaminergic mechanisms are impaired in aging due to a diminished dopamine release at the median eminence, in turn leading to an increased prolactin secretion. Increased serum prolactin levels have been observed also in humans, although the changes appear to be small. These data open, from a theoretical point of view, another perspective – that changes concerning hypothalamic monoamines may have long-range consequences due to the role that these neurotransmitters play in regulating hypothalamic neuroendocrine functions.

It should be remembered that some age-related changes might be modulated by gender and hormonal status. For example, it has been shown that long-term estrogen therapy may enhance dopaminergic response in postmenopausal women. Finally, it should be noted that most of the studies on dopamine transmission in aged rodents were performed at a time (late 1970s, early 1980s) in which the very definition of 'aged rodents' was somewhat ill defined, and the ages of the animals labeled as aged were somewhat variable (18, 21, 24, or 28 months) among the various laboratories worldwide, accounting for at least part of the variability of the published data.

In many studies on the human senescent brain, all biochemical indices of presynaptic nigrostriatal and mesolimbic dopamine neurons have been found substantially reduced. These changes include concentration of dopamine and its main metabolite, homovanillic acid; activity of tyrosine hydroxylase (TH) and DOPA decarboxylase; and dopamine uptake

sites. In extensive studies of dopamine and other presynaptic dopaminergic markers in human striatum, dopamine levels showed a pronounced postnatal increase during adolescence, followed by a decrease during aging. The levels of tyrosine hydroxylase and of aromatic amino acid decarboxylase show a similar postnatal increase, then decline by 30 years of age. Also, the vesicular monoamine transporter and the dopamine transporter show a similar pattern. The levels of the latter slightly, but significantly, decline in the caudate during aging. Consistent with the view of a reduced monoaminergic tone during aging, the activity of the monoamine-catabolizing enzyme, monoamine oxidase – in particular, the dopamine-degrading monoamine oxidase B – seems to be increased in several regions of the senescent human brain. On the whole, the data indicate a discrepancy between the loss of dopamine and the relative preservation of proteins involved in the biosynthesis and transport, suggesting that the decreased DA function in senescence may be explained by loss of DA *per se*, rather than by, or in addition to, a loss of dopaminergic neuropil. Notably, of the two dopamine-synthesizing enzymes, tyrosine hydroxylase and DOPA decarboxylase, the age-associated changes of the latter are more widespread, making it a better neurotransmitter index to monitor during aging.

The age-related loss of spiperone binding to dopamine D2 receptors has been shown to occur also in humans, and decreased concentrations are seen in the neostriatum long before old age, being significant by age 30 years. Notably, the selective loss of dopaminergic receptor subpopulations points to the possibility that their decreased number might be the consequence of a selective loss of postsynaptic elements, an observation that may distinguish between the context of the dopaminergic innervation loss in Parkinson's disease, which is much more pronounced and concerns predominantly presynaptic dopaminergic terminals, and nonpathological aging, in which the changes are more subtle and also involve postsynaptic elements. In relation to the presynaptic component, accurate positron emission tomography (PET) imaging studies using [^{18}F]fluorodopa indicate that the storage capacity for dopamine, both in the caudate and putamen, is reduced in the healthy elderly relative to young control individuals, and much more so, up to 85%, in parkinsonian patients. The observed changes have been attributed to a loss of vesicular storage capacity. On the other hand, the loss of dopamine D2 receptors in the senescent human striatum may be due to the loss of intrinsic striatal neurons as well as to the loss of nigrostriatal dopamine terminals and corticostriatal glutamatergic terminals bearing D2 dopamine receptors, even if animal data do suggest the possibility of a reduced receptor synthesis.

At partial variance with the preceding findings describing a reduced postsynaptic receptor number, observations showing that elderly patients develop spontaneous orobuccal–lingual dyskinesias suggest development of functional denervation supersensitivity of dopamine receptors, possibly due to the loss of presynaptic dopamine terminals. This view is also supported by data indicating that the ratio of homovanillic acid to dopamine (HVA/DA) is increased in elderly individuals, suggesting an increased DA turnover, a possible compensatory mechanism to the reduction of DA terminal number. Animal studies support this hypothesis; in fact the responsiveness of aged rats to apomorphine and amphetamine is increased.

On the whole, the data (summarized in Table 1) suggest that the age-related loss of dopamine D2 receptors represents the most generally observed change in the brain, both in laboratory animals and in humans. In turn, this observation suggests that the ratio between D1 and D2 receptors also changes with age, leading to a prevalence of the former. Even if the studies on dopamine metabolism are less consistent, the observed decline in the rate of HVA delivery from the brain to the CSF in elderly individuals is consistent with other evidence that brain dopaminergic neurotransmission declines with age.

The functional significance of these changes in normal human brain has also been investigated. From a general point of view it has been observed that the reduction in brain dopamine activity occurring as a part of aging is associated with changes in motor as well as cognitive functions. Interestingly, imaging studies have shown an association between dopamine D2 measures and metabolic activity in the frontal lobe. Together, these observations corroborate an involvement of the dopamine system in the cognitive deficits of the elderly. Moreover, these results, as well as observations in parkinsonian patients, indicate the possibility to identify strategies to improve motor and cognitive functions in the elderly based on the improvement of dopamine function. This view is further supported by data showing that diminished motor memory encoding in elderly humans is enhanced by L-DOPA intake. This goal may also be attained by nonpharmacological interventions. Notably, it has been repeatedly shown that dietary restriction in rats may prevent the loss of striatal dopamine receptors; as well, exercise training may affect brain catecholamine concentrations and receptor densities. Interestingly, new research also suggests the possibility to use compounds that mimic caloric restriction (CR) effects by targeting metabolic and stress

responses that are affected by CR, without actually restricting caloric intake. Also, drugs inhibiting glycolysis or enhancing insulin activity have provided some promising results showing enhanced dopamine action and upregulated brain-derived neurotrophic factor.

Noradrenergic Neurotransmission

Several studies report reduced noradrenaline concentrations at the level of the hypothalamus and hippocampus, and a similar trend also in other brain areas. Decreased levels of the neurotransmitter might be due to loss of noradrenergic neurons in the locus coeruleus. Compared to the age-associated loss of substantia nigra cells, the locus coeruleus cell loss is less pronounced. TH levels show little changes, even in regions of significant reduction of noradrenaline levels. Noradrenergic receptors appear to be intact in several brain areas, except in the cerebellum, where they are reduced. Studies in laboratory rats suggest that β-adrenergic receptors may become less able to express supersensitivity under reduced neurotransmitter concentrations and thus develop a reduced ability to efficiently couple to postreceptor transduction mechanisms.

From a more general point of view, in relation to postreceptor transduction mechanisms, it should be noted that data published on age-associated alterations of the coupling of receptors to G-proteins and to effector systems, in addition to data in published reports on changes in activities of several protein kinases, such as protein kinase A, calcium/calmodulin-dependent protein kinase, and, prominently, protein kinase C, do suggest that the age-associated changes of the response to monoamine transmitters acting on receptors using these transduction mechanisms might be more widespread than generally thought; this raises the possibility of a synergic combination of a reduced transmitter availability and a defective postreceptor machinery. Accordingly, age-associated alterations in the functional response may be more relevant than are changes in the number of receptors. Within this context, it has been shown that β-adrenergic receptors and the related cAMP signaling pathway play an important role in learning and memory, and change in aging in the hippocampus. In experiments using slices of the hippocampus of Fisher 344 rats of different ages, and a particular stimulation protocol involving the application of the β-adrenergic agonist isoproterenol, it was observed that the magnitude of long-term potentiation (LTP), an electrophysiological correlate of memory trace formation, was attenuated in old rats. Such changes may contribute to the compromised performance on memory tasks often observed with normal aging. It has also been shown in Fisher 344 rats that impaired acquisition and retention of the water maze task in aged rats were primarily associated with decreased norepinephrine (NE) activity in the cingulate cortex. No changes in the α1-, α2-, β1-, or β2-adrenergic receptor number were observed in any region sampled.

Within the brain, not only neuronally located β-adrenergic receptors but also those present on brain capillaries show selected age-associated changes. In fact, the locus coeruleus is the site of adrenergic nerve supply to parenchymal capillaries, and stimulation of this nucleus increases capillary permeability to small molecules and water. Abnormalities in the response of microvascular β-adrenergic receptors in aged rats in response to ischemic challenge have been observed, although it is not known whether these changes depend upon altered noradrenergic innervation of the vessels or upon altered intrinsic properties of the receptors. Such alterations may be relevant to microvascular dysfunctions known to occur with aging, including changes of control of the blood–brain barrier (BBB) permeation. On the other hand, it should be noted that in rodent models, no gross changes of the BBB permeability to proteins have been observed, as shown using horseradish peroxidase.

As in the case of dopamine, changes taking place at hypothalamic level influence the neuroendocrine status. The age-related decline of noradrenergic function in the hypothalamus influences the secretion of luteinizing hormone, growth hormone, and thyroid-stimulating hormone; drugs increasing catecholamines, administered to old rats, are shown to reverse this decline. Moreover, age-associated changes of noradrenergic transmission may influence the ability of an organism to cope with stress. In particular, the effects of acute stress on indices of the hypothalamic–pituitary–adrenal axis have been examined in aged Fisher 344 rats by evaluating the extracellular fluid content of noradrenaline and its metabolites, and of 3,4-dihydroxyphenylacetic acid (DOPAC), the metabolite of dopamine, by means of the brain microdialysis technique. In young animals, the stress produced by immobilization induces increased levels of NE in the paraventricular nucleus. The extent of the stress-induced NE increase is smaller in aged rats. The decreased responsiveness of the central catecholaminergic system is paralleled by a similar defect in the peripheral catecholaminergic system and is accompanied by a progressive hypothalamic corticotropin-releasing hormone deficiency. At the pineal level, the melatonin response to isoproterenol is age dependent, and in pineals from old Wistar rats the response is approximately half of that found in young rats.

The age-associated changes of noradrenergic brain neurotransmission may in part be due to the failure of trophic factors, such as glia-derived neurotrophic factor (GDNF), for noradrenergic neurons in the locus coeruleus. Old (18 months) mice heterozygous for the GDNF gene (*Gdnf*+/−), compared to +/+ mice, present a reduced locomotor activity and a reduction in NE transporter activity in the cerebellum and brain stem, as well as reduced noradrenaline tissue content in the locus coeruleus. The data suggest that GDNF may contribute to maintain the noradrenergic system throughout life. Together with the previously mentioned data showing that BDNF helps to maintain adequate dopaminergic neuron survival, these additional data support the concept that at least some of the age-associated changes in catecholaminergic innervation may depend upon defects in the synthesis of adequate amounts of growth factors.

Serotonergic Neurotransmission

The effect of age on serotonergic transmitter parameters has been less studied than the effect on catecholamines. From a general point of view the vulnerability of the catecholaminergic systems, in particular dopamine, to aging contrasts with the relative resistance of the serotonergic system. At a glance, it appears that aging affects the serotonergic system rather heterogeneously. The published reports are frequently contradictory and reflect strong differences according to the brain area investigated. In the case of serotonergic transmission, the data on presynaptic markers of activity are inconsistent, with increase, decrease, and no changes having been reported in the literature. As a whole, serotonin and its metabolite 5-hydroxyindoleacetic acid (5-HIAA), as well as the imipramine binding site, do not seem to show age-associated alterations. Also, the deaminating enzyme monoaminoxidase A appears to be stable with age. More consistent data have been reported in the case of serotonin 5-HT_1 and 5-HT_2 receptors, showing a decreased density at the level of the frontal cerebral cortex and hippocampus. Using a selective 5-HT_{2A} receptor ligand, a decrease in the binding index of a neuroreceptor radioligand has been observed in the frontocortical area in canine brain. PET studies investigating *in vivo* in conscious monkeys the effect of aging on 5-HT_{1A} receptors, and their functional response to agonists, revealed a significant age-associated decrease of 5-HT_{1A} binding in frontal and temporal cortices, but not in cingulate gyrus, hippocampus, occipital cortex, striatum, thalamus, or raphe nucleus. Moreover, the degree of displacement of the tracer from its binding sites following systemic administration of 8-hydroxy-2-(di-*n*-propylamino) tetralin (8-OH-DPAT), a 5-HT_{1A} agonist, was smaller in aged compared to young monkeys.

The reported loss of serotonin receptors with age may be attributed to loss of neurons or cell components that receive serotonergic inputs. At the cortical level, loss of serotonin receptors may be associated with cortical cell loss, rather than to a specific impairment of serotonin receptor synthesis. Such serotonin receptor loss, as well the decreased noradrenergic tone, may contribute to the increased risk to develop depression in the elderly, even if changes in other systems, such as control of growth factors and corticosteroid activity, may be more important in this respect.

Changes in the activity of serotonin may alter neuroendocrine function. The decline in circadian activity of the suprachiasmatic nucleus serotonergic neurons may account for the blunting of circadian rhythms in the secretion of several anterior pituitary hormones in old animals. Moreover, inasmuch as catecholamine levels decrease with aging, the ratios of serotonin to noradrenaline and of serotonin to dopamine rise. At hypothalamic level, catecholamines and serotonin modulate endocrine function through action on hypothalamic neurosecretory products, thus an imbalance between these systems may underlie some of the neuroendocrine changes associated with aging (see **Figure 1** for a scheme integrating these concepts). Serotonin transmission is also affected at the level of the pineal gland. In old Wistar rats, pineal 5-HT and 5-HIAA concentrations were 41–47% of those found in young rats. These observations and those on the melatonin response to noradrenergic stimulation suggest the existence of both a noradrenergic and a serotonergic deficit at the pineal level. Due to the putative roles attributed to melatonin in immunosenescence and in light/dark adaptation, these changes may participate in determining the complex pattern of monoamine-related neuroendocrine aging.

In humans, a reduced density of serotonin 5-HT_{2A} receptors has been observed using PET tracers in various brain regions, including the anterior cingulate, posterior medial prefrontal cortex, hippocampus, and occipital cortex. Most of the decrease in receptor binding occurred up through midlife and there was less decrease in late life. In the past, postmortem studies have provided limited and conflicting data regarding aging effects on the central serotonin transporter (SERT). However, data from autopsy studies are limited by the difficulties of assessment of the acute clinical picture before death and by postmortem artifacts. On the other hand, data based on *in vivo* studies using PET techniques have also

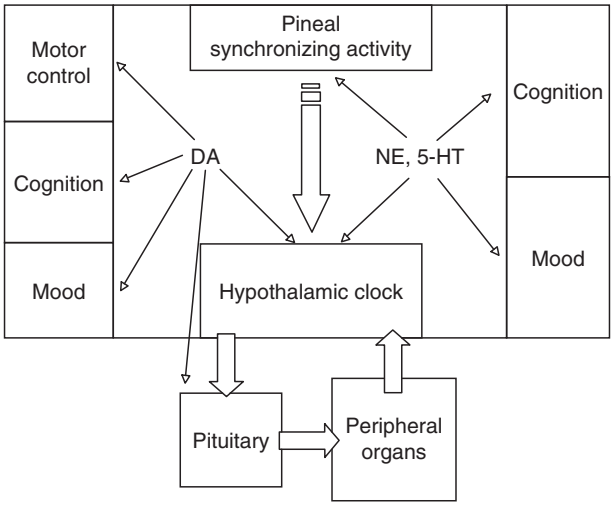

Figure 1 Illustration of the key role of monoamines in controlling several activities in the brain and in the neuroendocrine axis. Age-associated changes in synthesis, transporters, and receptors, coupling with effector systems involving any of the monoamines or in their ratios (DA, dopamine; NE, norepinephrine; 5-HT, 5-hydroxytryptamine), may profoundly affect brain function as well as peripheral organ function. Pineal and hypothalamic function should also comply with internal and external timing (consider the pulsatile secretion of several hormones); the age-associated changes may therefore occur not only in terms of absolute amount of transmitter or response, but also in terms of incorrect timing of secretion. See text for further details.

provided conflicting data. Single-photon emission tomography analyses of presynaptic SERT suggest an age-related decline in the thalamic/hypothalamic and midbrain/upper brain stem SERT density. However, it appears that when the appropriate control measures are taken, the changes of central SERT, if present, are much smaller than those of dopamine transporters. As a whole, the results of postmortem and *in vivo* imaging studies on the influence of age on serotonergic transmission are scanty and controversial, in spite of the major role that the central neurotransmitter serotonin plays in a number of neuropsychiatric human conditions.

See also: Cholinergic System; Growth Factors: Neuronal Atrophy; Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Monoamines: Human Brain Imaging; Monoamines: Release Studies; Neuroleptics; Neurotransmitters and Growth Factors: Overview; Octopamine and Other Monoamines in Invertebrates; Trace Monoamines and Receptors in Mammalian CNS.

Further Reading

Cotzias GC, Miller ST, Tang LC, et al. (1977) Levodopa, fertility, and longevity. *Science* 196: 549–551.

Govoni S and Battaini F (eds.) (1987) *Modification of Cell to Cell Signals during Normal and Pathological Aging*. Nato ASI series, vol. 9. Berlin, Heildelberg, New York: Springer.

Haycock JW, Becker L, Ang L, et al. (2003) Marker disparity between age-related changes in dopamine and other presynaptic dopaminergic markers in human striatum. *Journal of Neurochemistry* 87: 574–585.

LaVoie M, Ostaszewski BL, Weihofen A, et al. (2005) Dopamine covalently modifies and functionally inactivates parkin. *Nature Medicine* 11: 1215–1221.

Meites J (1991) Role of hypothalamic catecholamines in aging processes. *Acta Endocrinology* 125(supplement 1): 98–103.

Morgan DG, May PC, and Finch CE (1987) Dopamine and serotonin systems in human and rodent brain: Effects of age and neurodegenerative disease. *Journal of the American Geriatric Society* 35: 334–345.

Pradhan SN (1980) Central neurotransmitters and aging. *Life Sciences* 26: 1643–1646.

Roth GS, Mattison JA, Ottinger MA, et al. (2004) Aging in rhesus monkeys: Relevance to human health interventions. *Science* 305: 1423–1426.

Samorajski T (1977) Central neurotransmitter substances and aging: A review. *Journal of the American Geriatric Society* 25: 337–348.

Volkow ND, Gur RC, Wang G-J, et al. (1998) Association between decline in brain dopamine activity with age and cognitive and motor impairment in healthy individuals. *American Journal of Psychiatry* 155: 345–349.

Monoamines: Human Brain Imaging

D F Wong, The Johns Hopkins University School of Medicine, Baltimore, MD, USA
A Gjedde, Aarhus University & Aarhus University Hospitals, Aarhus, Denmark; The Johns Hopkins University School of Medicine, Baltimore, MD, USA; and McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Monoamines and Monoaminergic Neurotransmission

Monoamines are neuromodulators that arise from the decarboxylation of large neutral and aromatic amino acids in brain tissue. Depending on the specific precursor, they fall into two categories, the catecholamines that ultimately are derived from phenylalanine but in brain only come from tyrosine, and the indoleamines that are derived from tryptophan. The most important catecholamines are dopamine and noradrenaline, and the most important indoleamine is serotonin.

Neuromodulation is a special kind of neurotransmission. Neurotransmission, in turn, is the term applied to the exchange among brain cells of specific information encoded in chemicals known as transmitters. The conventional description of neurotransmission includes, at the very least, the steps of synthesis, vesicular storage, release, reuptake, and removal or breakdown. This is a simple view of the molecular interactions involving neurotransmitters, but it is the basis on which the molecular imaging of neurotransmission rests. The individual steps are implemented in extra- or intracellular proteins that act as enzymes, receptors, or transporters. Molecules that interact with the proteins are known as ligands, because the first step of the interaction is binding, regardless of the additional functions of the protein. By necessity, all neurotransmitter molecules are ligands, but not all ligands work as neurotransmitters.

Binding is the first step in the processing of the ligands, the ultimate fates of which differ, according to the proteins with which they interact. Receptors release ligands to the original pool of unbound molecules after a delay that depends on the affinity of the receptors toward the ligands and possible internalization. Transporters relocate ligands to different pools of unbound ligands. Enzymes release ligands to the same pools but in a new chemical form. Exogenous or artificial ligands may differ so much from native ligands that transporters or enzymes fail to carry out their primary function, and instead act as receptors by simply returning ligands unchanged to their original pool. Intentionally or unintentionally, this interaction with a protein may interfere with the access of any endogenous ligand. Depending on the degree of interference, exogenous or artificial ligands may thus be classified as antagonists or partial agonists of the action of the native neurotransmitters.

As neuromodulators, monoamines are not classical neurotransmitters. Classical neurotransmitters confine their action between cells to receptors located in or near synapses, or 'kissing points,' that are also the sites of release. This type of neurotransmission is 'wired'; in wired transmission, release, binding, and reuptake are linked by the very short distances, so as to prevent wide dissemination of the transmitter. In contrast, 'volume' transmission is a paracrine process, in which the transmitter molecules diffuse widely and their concentrations depend less on the moment-to-moment fluctuations of release. Instead, it is the density and location of transporters in relation to sites of release that determine the specificity of action. Thus, in volume transmission, the transporters are located far from the site of release and their location determines the nature of the transmission. Hence transmitters engaged in volume transmission carry no specific information about the sites of release. In keeping with the mechanism of neuromodulatory volume transmission, the monoamine transporters primarily are extrasynaptic, although the monoamines may serve both wired and volume transmissions, such as in the case of dopamine.

As neuromodulators, monoamines tend to reduce the excitability of target neurons, although this action does depend on the specific receptor subtypes that are expressed locally. Because the specificity of the proteins toward exogenous or artificial ligands varies, it is now possible to image presynaptic and intracellular as well as postsynaptic sites of transmitter interaction by selecting or designing exogenous or artificial ligands that match the specificities of the proteins. The main extra- and intracellular sites of binding, reuptake, and metabolism are listed in Table 1.

The expression of receptors and enzymes is highly variable and depends on the intra- and extracellular environments, making all results of positron emission tomography (PET) and single-photon emission computed tomography (SPECT) contingent upon the specific circumstances at the time of study. This is also true of the transporters, but when the first tropane-based PET probes for the dopamine transporters were used in baboons and humans, it was generally assumed that the number of transporters remained constant. However, it is now known that the number of dopamine transporters is subject to regulation by intrinsic mechanisms as well as pharmacological manipulation.

Table 1 Variables of monoaminergic neurotransmission

Location	*Variable*	*Example*
Presynaptic	Transporter availability (binding potential)	Dopamine transporters, vesicular monoamine transporter 2
	Enzyme activity	Dopa decarboxylase
	Precursor concentration	DOPA
Postsynaptic	Receptor availability, density, and affinity	Dopamine D_2-like receptors
	Receptor subtype selectivity	Dopamine D_1- or D_2-like receptors
Intrasynaptic	Endogenous neurotransmitter occupancy	Dopamine
	Endogenous neurotransmitter release	Dopamine
	Enzyme activity	Monoamine oxidase A or B

Table 2 Properties of radionuclides used in PET or SPECT imaging

Isotope	*Modality*	*Decay mode (%)*[a]	*Half-life (min)*	*Production type*[b]
^{11}C	PET	β^+ (100)	20.1	C
^{13}N	PET	β^+ (100)	9.996	C
^{15}O	PET	β^+ (100)	2.04	C
^{18}F	PET	β^+ (97), EC (3)	109.8	C/R
^{62}Cu	PET	β^+ (97), EC (3)	9.7	G[^{62}Zn]
^{68}Ga	PET	β^+ (89), EC (11)	68	G[^{68}Ge]
^{82}Rb	PET	β^+ (95), EC (5)	1.3	G[^{82}Sn]
^{99m}Tc	SPECT	IT (100)	360	G[^{99}Mo]
^{123}I	SPECT	EC (100)	780	C
^{131}I	SPECT	β^- (100)	8 (days)	C/R

[a]Decay modes: β^+, positron emission; β^-, electron emission; EC, electron capture; IT, isomeric transition.
[b]Production types: C, cyclotron; R, reactor; G, generator.

Methods of Molecular Imaging with PET and SPECT

Molecular imaging of neurotransmission *in vivo* is an extension of the principles of *in vitro* and *ex vivo* autoradiography to the living brain. Historically, much of the imaging has focused on the dopaminergic innervations of the basal ganglia by the nigrostriatal fibers, but the serotonergic, noradrenergic, and dopaminergic innervations of the limbic and cerebral cortices have since become important targets. The first reports on the use of PET and SPECT to image aspects of neurotransmission in the basal ganglia with radiolabeled drugs appeared as early as 1983. By labeling specific drug molecules of known action, the investigators confirmed the fates of these molecules in time and space in the brain and thus recorded events that could be ascribed to the known effects of these drugs on dopaminergic neurotransmission.

Although both PET and SPECT are used in these efforts, it is important to note their different characteristics, as listed in Table 2. PET tracers emit positrons, and PET detects the gamma radiation that arises from the annihilation of positrons when they collide with electrons. The most common radiolabels used with PET are the positron-emitting isotope of carbon (^{11}C) that replaces a stable carbon atom (^{12}C), and the positron-emitting isotope of fluorine (^{18}F) that either is added to some molecules or replaces stable fluorine (^{19}F) in other molecules, in most cases without significantly affecting the original kinetic properties of the molecules. PET has the advantage that the half-lives of these isotopes are very short (20 min for ^{11}C and 110 min for ^{18}F), but the disadvantage is that the nuclides must be produced in a local cyclotron.

SPECT tracers emit gamma radiation directly, arising from photon-emitting nuclides such as technetium-99m (^{99m}Tc) and iodine-123 (^{123}I). These nuclides have much longer half-lives, as listed in **Table 2.** Some SPECT tracers therefore have the advantage of being available in most nuclear medicine departments, and the relatively long half-life of ^{123}I (13 h) permits tomography to take place as late as 24 h after administration of the tracer, but it is a disadvantage that this nuclide currently is available primarily in North America only from a cyclotron facility in Vancouver, Canada (Tri-University Meson Facility, or TRIUMF).

Multiple brain images in the form of dynamic series of tomographic frames are acquired in PET studies of monoaminergic neurotransmission. Typically, 30–50 frames are obtained as two- or three-dimensional images in 60–120 min. Because of the short half-lives, this duration is the limit of most studies with ^{11}C or ^{18}F. In contrast, SPECT studies with ^{123}I may be continued for 6 or more h, as evident from Table 2. In two-dimensional imaging, the tomographic record consists of 15–64 or more sections of the brain, depending on the tomograph.

In many studies, both of animals and humans, investigators insert a catheter in a suitable artery to determine the tracer concentration in arterial blood. The blood samples are assayed for radioactivity and six to eight discrete samples of arterial plasma undergo additional high-performance liquid chromatography (HPLC), or thin-layer chromatography

(TLC), to determine the chemical nature of all labeled products in plasma, to account for metabolism in the course of circulation and to reconstruct the concentration of the parent compound as a function of time. The arterial samples and brain images establish the concentrations of the tracer in blood and brain as functions of time. The records then form the basis for the subsequent kinetic analysis that yields the specifics of the monoaminergic neurotransmission, as listed in Table 1.

The availability of the right tracers is the key to successful imaging of monoaminergic neurotransmission in animals or humans. With PET, the imaging is limited by the half-life of the tracers, which means in the case of ^{11}C-labeled tracers that all information must be extracted during the first hour of recording. The usefulness of this information in turn is dictated by the pharmacology and binding properties of the tracer in relation to the relevant proteins and not solely by the properties of the ligand.

An important challenge to PET imaging is the necessary anesthesia of animals used in research. It has been possible in very few places to train nonhuman primates to remain awake while seated comfortably under a horizontally suspended tomograph (e.g., Hamamatsu, Japan). In other places, it is necessary to use a within-subject design of drug studies to reduce the variability of the PET measures. Other factors, such as the size of brain, genotype, age, latent disease states, personality, and adaptation to stress and drugs, have not routinely been sorted out in most preclinical or clinical studies. A few studies addressed elements of psychology and personality, including introversion and social reclusiveness, and the effects of social anxiety. Nonetheless, it is clear that interpretation of these effects is difficult, as a single variable may reflect opposite changes of monoaminergic neurotransmission.

In one study of social isolation in monkeys, the availability of dopamine D_2 receptors in basal ganglia, measured by PET, was shown to rise in dominant monkeys when they moved from individual to social housing. Receptor 'availability' is the controversial term applied to the estimate of the ratio between bound and unbound ligand quantities in a tissue. This ratio is also known as the 'binding potential.' It declines when the receptors approach saturation and hence can be said to be an index of the relative number of unoccupied receptor sites. The number rises either because the total number of receptors rises, or because the concentrations of relevant ligands fall, or both. In the social isolation study, the result was interpreted, somewhat vaguely, as a 'rise' of dopaminergic neurotransmission when dominant monkeys met other monkeys that they could dominate, but the effect could just as well arise from a decline of dopamine, when dominant monkeys escape from the social isolation.

An important problem is sample size. Typical numbers in PET studies of humans range from 10 to 20 participants, and a study population of 30 individuals is rare. Adequate numbers are necessary to counteract the confounding influence of the tomograph's finite power of resolution of the recorded radioactivity, known as the 'partial volume' effect. The effect is most prominent in small structures that typically fuse below a diameter of 2–3 mm. The effect persists for diameters as large as two to three times the full-width-at-half-maximum (FWHM) property of a point source for any imaging agent. Thus, the human cerebellum usually loses about 5 or 10% of its total activity when the entire structure is imaged. The resolution of PET originally was no less than 8 mm FWHM. Current tomographs enjoy resolutions as low as 2–3 mm FWHM for the high-resolution research tomograph (HRRT), which allows visual separation of two points with distances as small as 0.5 cm. The partial volume effect is an even greater problem with SPECT, because the attenuation correction is much greater with this method.

Tracers of Monoaminergic Neurotransmission

PET and SPECT both use marker molecules (tracers) that bind to the proteins of monoamine synthesis, binding, and metabolism with sufficiently high affinity and selectivity, but ordinarily with such low concentration that pharmacodynamic effects are absent. The marker molecules are tracers because the radiation that they emit can be detected externally by the tomographs.

The typical investigation of monoaminergic neurotransmission begins with identification of an appropriate precursor of a tracer that targets a specific step in the pathway from the aromatic amino acid parent to final metabolite. Both the choice of precursor and the synthesis of the tracer from the precursor are based on evidence obtained *in vitro* or *ex vivo* in animals or humans. When a potential tracer is labeled, its usefulness must be examined carefully, first in animals and then in humans; at any time in this process, unexpected species differences may interfere with the imaging – for example, by blocking the transfer of the tracer across the blood–brain barrier, or by altering the metabolism of the tracer in the brain tissue. The necessary safety profile in the form of toxicology and radiation dosimetry must also be established. Validation in animals may involve

initial studies *in vitro* to confirm the saturability of the binding, or the competition with chemically or pharmacologically related drugs. Once the binding and metabolism characteristics are known, the tracer may finally be deemed appropriate for use *in vivo* in nonprimate mammals or nonhuman or human primates.

Table 3 lists the large number of monoaminergic tracers that have been tested with variable success since 1983. Unfortunately, of the hundreds of

Table 3 Monoamine neuroreceptor radioligands for PET and SPECT

Protein	*Dopamine*	*Serotonin*	*Noradrenaline*
Receptors	$D_{2/3}$	$5HT_{1A}$	α_2
	3-*N*-[^{11}C]Methylspiperone	[^{11}C]WAY 100635	[^{11}C]Yohimbine
	[^{11}C]Raclopride	[^{11}C]Carbonyl-WAY 100635	[^{11}C]Mirtazapine
	[^{11}C]FLB 457	*cis*-4-[^{11}C]Methoxy-WAY	
	[^{11}C-Methyl]-(−)-OSU6162	*trans*-4-[^{11}C]Methoxy-WAY	
	[^{18}F]Clebopride	[^{11}C]LY-274601	
	[^{18}F]Fallypride	[^{11}C]NAD-299	
	[^{18}F]Desmethoxyfallypride	[^{11}C]CPC-222	
	[^{18}F]Ioxipride	[^{11}C]DWAY	
	[^{18}F]Methylspiperone	[^{18}F]FBWAY	
	[^{18}F]FCP	[^{18}F]FCWAY	
	[^{18}F]Spiperone	[^{18}F]Fluoroethyl-WAY	
	[^{18}F]Fluoroethylspiperone	[^{18}F]Fluoropropyl-WAY	
	[^{18}F]Fluoropropylepidepride	[^{18}F]MeFBWAY	
	N-Allyl-5-[^{18}F]fluoropropylepidepride	*cis*-[^{18}F]FCWAY	
	[^{18}F]NCQ-15	*trans*-[^{18}F]FCWAY	
	[^{123}I]IBZM	*p*-[^{18}F]MPPF	
	[^{123}I]IBF	*p*-[^{123}I]MPPI	
	D_1	$5HT_{2A}$	
	[^{11}C]SKF 82957	*N*-[^{11}C]Methylspiperone	
	[^{11}C]SCH 23390	[^{11}C]MDL 100,907	
	[^{11}C]NNC 756	[^{18}F]Altanserin	
	[^{11}C]NNC112	[^{18}F]Setoperone	
	[^{11}C]SCH 20098	$5HT_4$	
	R-[^{11}C]SKF 82957	[^{11}C]SB207145	
	[^{18}F]Fluoropropyl SCH 38548		
Transporters	DAT	5HTT or SERT	NET
	[^{11}C]Methylphenidate	[*O*-Methyl-^{11}C]venlafaxine	[*O*-Methyl-^{11}C]venlafaxine
	[^{11}C]Cocaine	[^{11}C]McN5652	[^{11}C]-SS-MeNER
	[^{11}C]CFT (WIN 35,428)	[^{11}C]NS2456	[^{11}C]Hydroxyephedrine
	[^{11}C]SS-CIT-FE	[^{11}C]DASB	
	[^{11}C]PE2I	[^{11}C]DAPA	
	[^{11}C]Altropane	[^{11}C]AFM	
	[^{11}C]Naphthyltropane analog	[^{11}C]ADAM	
	[^{11}C]Phenyltropane analogs	[^{11}C]MADAM	
	[^{11}C]RTI 32	[^{123}I]ADAM	
	[^{18}F]Fluorococaine		
	[^{18}F]CFT		
	[^{18}F]FP-CIT		
	[^{18}F]FE-CNT		
	[β-[^{123}I]CIT		
	[^{99m}Tc]TRODAT-1		
	VMAT2	VMAT2	VMAT2
	d-[^{11}C]Tetrabenazine	d-[^{11}C]Tetrabenazine	d-[^{11}C]Tetrabenazine
Enzymes	Dopa decarboxylase	Dopa decarboxylase	Tyrosine hydroxylase
	L-[^{11}C]DOPA	L-[^{11}C]DOPA	2-[^{18}F]Tyrosine
	6-[^{18}F]Fluoro-l-dopa	5-Hydroxy-L-[β-^{11}C]Tryptophan]	4-[^{18}F]Tyrosine
	[^{18}F]Fluorodopamine	[^{18}F]FMT (fluoro-*m*-tyrosine)	[^{123}I]IMT
	O-Methyl-[^{18}F]fluoro-DOPA		[^{18}F]FMT (fluoro-*m*-tyrosine)
	[^{18}F]FMT (fluoro-*m*-tyrosine)		

thousands of compounds developed by pharmaceutical industry, university, or government institutions, only a tiny fraction can be used successfully in labeling. Radioligands are useful only when they meet specific requirements of selectivity, affinity, and kinetics of binding, transport, or metabolism. Only when these requirements are met can reliable estimates of receptor density or neurotransmitter synthesis, release, or metabolism, be obtained noninvasively.

Three tracers illustrate the historical advance of tracer development from 1983 till today, and the problems associated with the use of the tracers. In 1983, the experimental neuroleptic spiroperidol (spiperone; 8-[4-(4-fluorophenyl)-4-oxo-butyl]-1-phenyl-1,3,-8-triazaspiro[4.5]decan-4-one) was the precursor by methylation of the radioligand 3-*N*-[^{11}C]methylspiperone (NMSP), for which the dopamine D_2-like (D_2, D_3, and D_4) receptors have high affinity. Spiperone is a piperazine, originally so named because of the chemical relation to piperidine, a constituent of piperine in the black pepper plant (*Piper nigrum*), but the radioligand is known also as a butyrophenone. The radioligand yielded the first D_2-like receptor PET images in humans. However, the selectivity was not exclusive for this receptor subtype, as the serotonin $5HT_{2A}$ receptors also have significant affinity. Kinetic analysis of the binding was complicated by this high affinity; the relative influences of transport across the blood–brain barrier and binding to the receptors cannot be fully resolved when a secular equilibrium of bound and unbound ligand is not approached during the period of uptake. Therefore, ligands that bind irreversibly often have delivery-dependent, rather than affinity-dependent, uptake. In contrast, useful ligands frequently have binding characteristics that include high selectivity but only moderate affinity, commensurate with binding potentials between 1 and 6. In the case of NMSP, the binding potential exceeds 6 in humans, although partial volume effects may reduce this value in the smaller brains of experimental animals. Interestingly, NMSP has been of use in cerebral cortex, where the lower affinity of the serotonin receptors renders the binding reversible in the period available for PET, without serious interference from the lower density dopamine receptors in cortex.

The currently most common monoamine receptor ligand used with PET is [^{11}C]raclopride (3,5-dichloro-*N*-[(1-ethylpyrrolidin-2-yl)methyl]-2-hydroxy-6-metho xy-benzamide), which derives from another class of chemical, the benzamides. Benzamides are derivatives of benzoic acid, and substituted benzamides often serve as neuroleptics or antipsychotics in psychiatric practice. Two active substances from the group of benzamides are in clinical use, sulpiride and amisulpride. Raclopride binds reversibly and selectively to the dopamine D_2 and D_3 receptors, with a binding potential of 3 in human brain, and as such is an ideal tracer of these dopamine binding sites in high-density locations such as the striatum. In cerebral cortex, the lower density renders the binding potential too low to be useful. Several alternatives with higher affinity are available for this purpose. The tracer [^{18}F]fallypride ([5-(3-fluoropropyl)-2,3-dimethoxy-*N*-[(2*S*)-1-(2-propenyl)-2-pyrrolidinyl]-methyl]-benzamide) is a radioligand with binding potentials of 0.5–1.5 in cerebral cortex. The higher affinity elevates the retention in striatum, where the binding potential is so high that binding is delivery dependent. Benzamide tracers are also available for SPECT. They include [^{123}I]iodobenzamide (IBZM) and [^{123}I]- and [^{125}I]epidepride.

In the case of metabolic substrates, it is necessary that the product be retained in brain as an index of the activity of the enzyme, provided the transport across the blood–brain barrier is not rate limiting for the uptake. The first study with PET of the metabolism of the dopamine precursor [^{18}F]fluorodopa (FDOPA) was also published in 1983. The tracer is one of several substrates of the aromatic amino acid decarboxylase, which produces both dopamine and serotonin from appropriate amino acid precursors in respective terminals. The uptake appears to be irreversible but the appearance may be deceiving, as the retention may reflect only a fraction of the tracer.

For FDOPA, a host of different kinetic analyses have been attempted to solve the issue of the specificity of the signal. Increasingly, it is clear that the retention primarily reflects the transport of the labeled product fluorodopamine (FDA) into vesicles, and the retention depends on the irreversibility of this transport. Two processes interfere with the retention: First, FDA is subject to oxidation by monoamine oxidases, and second, release of FDA from the vesicles exposes the released FDA to renewed metabolism after reuptake into the terminals. Therefore, low retention can reflect both reduced vesicular incorporation due to loss of terminals or loss of dopa decarboxylase activity, or loss or inhibition of vesicular transporters, or increased release of FDA in response to increased dopaminergic neurotransmission. As indicators of neuronal function, these interpretations have completely opposite implications for the health of the patient, yet the PET signals may look exactly the same. It should be noted that FDOPA is not specific for dopaminergic terminals, as both serotonergic and noradrenergic terminals use this enzyme in the synthesis of their transmitters. However, only in dopaminergic terminals is the product FDA retained.

As markers of dopaminergic activity in presynaptic terminals, the only alternatives to FDOPA are

the tracers of the monoamine reuptake sites, which are based on the tropane structure known as nortropane (8-azabicyclo[3.2.1]octane). Tropane is a nitrogenous bicyclic organic compound, mainly famous for the large group of alkaloids that are derived from it (called tropane alkaloids). The derivatives include, among others, atropine and cocaine. These two alkaloids both contain tropinone, from which tropane is derived. Tropane alkaloids occur in plants of the family Solanaceae (mandrake, henbane, deadly nightshade, potato, and tomato) and in the coca plant.

Several potentially irreversible ligands for the cocaine receptor site on the dopamine transporter are derived from cocaine or 3β-phenyltropane. The cocaine analog CFT (2β-carboxymethoxy-3β-(4-fluorophenyl) tropane, or WIN 35,428) was first recommended as a radiolabeled probe for PET and SPECT by Madras and co-workers and was later developed as the PET tracer [^{11}C]WIN 35,428 by Kuhar and co-workers. The tracer binds irreversibly to the cocaine site on the dopamine transporter (DAT) but lesser to the serotonin transporter (5HTT, or SERT) and the norepinephrine transporter (NET). In attempts to overcome the low selectivity, the basic structure has been modified repeatedly with substitutions of fluorine or small carbon chains that may reduce affinity and raise selectivity. Versions suitable for SPECT include β-[^{123}I]CIT (β-carbomethoxy-3β-(4-iodophenyl)tropane) for the dopamine and serotonin transporters, and [^{99m}Tc] TRODAT-1 selective for the dopamine transporter (technetium(V)-oxo-3β-(4-chlorophenyl)-8-methyl-8-azabicyclo[3.2.1]oct-2-yl[*N*-(2-mercaptoethyl), *N*-(*N*′-(2-mercaptoethyl)-2-aminoethyl)]-aminomethyl).

The evidence for regulation of monoamine transporters has intensified the search for additional probes of events in monoaminergic terminals. The vesicular monoamine transporter (VMAT2) is located in the membranes of terminal vesicles and can be imaged with the radioligand D-[^{11}C]tetrabenazine (DTBZ). However, not even the imaging of the VMAT2 transporter may settle the 'transporter regulation issue,' as VMAT2 also appears to be regulated.

Pharmacokinetics of Monoaminergic Imaging

Generally, the characteristics of binding of radiotracers to the proteins of monoaminergic neurotransmission *in vivo* are consistent with insights from studies *in vitro*, particularly with respect to binding. In Figure 1, the binding of a number of compounds is expressed as the 50% inhibitory concentration (IC_{50}) of the tracer, which takes into account both the receptor's affinity toward the tracer and the concentrations of competitors.

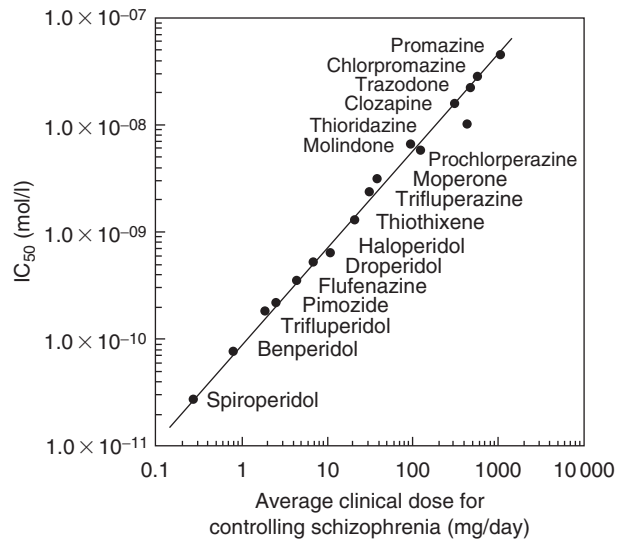

Figure 1 Relationship between 50% inhibitory concentration (IC_{50}) values at dopamine receptors and therapeutic dose of neuroleptics in treatment of schizophrenia. The IC_{50} values are the concentrations of the antipsychotic drugs that reduce the stereospecific component of [^{3}H]haloperidol binding by 50%. The abscissa indicates the average values of doses used in schizophrenic patients. Redrawn from Seeman P, Lee T, Chau-Wong M, et al. (1976) Antipsychotic drug doses and neuroleptic/dopamine receptor. *Nature* 262: 717–719.

The most commonly reported variable obtained with imaging of monoaminergic neurotransmission is the ratio between the density of receptors and the affinity constant, known as the binding potential. Both terms are part of the Michaelis and Menten steady-state solution to the binding equations,

$$B = \frac{B_{\max} C}{K_m + C}$$

where B is the bound quantity of ligand, C is the unbound ligand's concentration, and K_m is the Michaelis constant or half-saturation concentration. The binding potential is the ratio between the bound and unbound ('free') quantities of an endogenous ligand, modified by the ratio of the affinities of the receptor toward the endogenous ligand and the tracer,

$$p_B = \frac{B^*}{C^*} = \frac{B_{\max}}{C^* + K_m^*\left(1 + \frac{C_i}{K_i}\right)} = \frac{K_i B_{\max}}{K_m^*(K_i + C_i)} = \left(\frac{K_i}{K_m^*}\right)\left(\frac{B_i}{C_i}\right)$$

where p_B is the binding potential and the remaining symbols represent bound tracer (B^*), unbound tracer (C^*), maximum number of binding sites ($B_{\max}$), affinity toward the tracer (K_m^*), affinity toward an endogenous ligand (K_i), concentration in brain tissue

of endogenous ligand (C_i), and bound endogenous ligand (B_i).

Sokoloff and co-workers first presented the now standard three-compartment model of tracer kinetics *in vivo* that typically is used to obtain the estimates of binding potential or other variables of tracer binding or metabolism. When kinetic modeling was first proposed for studies of neurotransmission *in vivo* in 1983 and 1984, its use in pharmacology was novel because the Michaelis–Menten solution requires equilibrium of binding, which in principle is not achieved with PET. The Michaelis–Menten equation is the solution to the differential equations of the exchange of tracer between the circulation compartment, the pool of unchanged and unbound tracer in brain tissue, and the pool of bound or metabolized tracer in the tissue. The model assumes that the tracer or its metabolite is retained in this latter pool to such an extent that the tracer exchange with the circulation is delayed significantly.

The model shown in Figure 2 includes the coefficient of clearance (K_1) of the tracer from the circulation to the brain, and the rate constant of fractional clearance (k_2) of the tracer from the exchangeable pool of unbound tracer in the brain back to the circulation. The exchange distribution volume is simply the ratio of these two terms. The rate constants k_3 and k_4 refer to the interaction with the specific proteins, indicative of the retention associated with binding, transport, or metabolism. The specific definitions depend on the process, but generally k_3 is related to the quantity of protein and the concentration of the tracer while k_4 is simply the decay constant of the pool of retained substance. In the case of binding, the definitions are

$$k_3 = K_{\text{on}}(B_{\max} - B_i)$$

and

$$k_4 = K_{\text{off}}$$

where k_3 and k_4 define the respective association and dissociation rate constants. The ratio k_3/k_4 is a function of the amount of protein (such as density of receptors) and the affinity for the tracer, and the ratio is therefore an index of the tissue's potential for retention, a real 'retention potential.' In the case of binding, the definitions show that this ratio is also the binding potential defined earlier.

The exact solutions can be obtained from any standard text. The receptor density $B_{\max}$ is calculated from multiple estimates of the binding potential when the quantity of bound ligand is known for each estimate. The bound quantity includes the endogenous as well as exogenous ligands, such as a drug given to reduce the binding potential of the tracer by competition for binding.

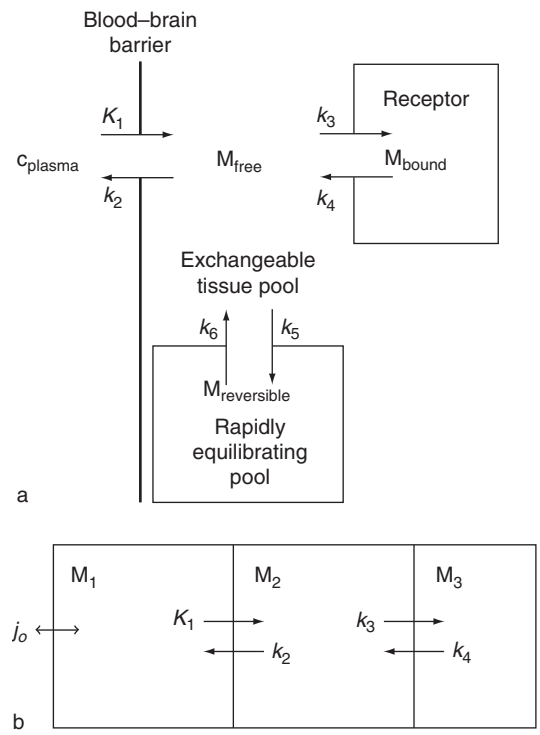

Figure 2 (a) Alternative 1: Three-compartment model (M) of receptor binding, with additional compartment for nonspecific binding of ligand. For definition of rate constants, see text. (b) Alternative 2: Three-compartment model (M) of receptor binding of radioligands. For definition of rate constants, see text. (a) From Wong DF, Gjedde A, and Wagner HN Jr. (1986) Quantification of neuroreceptors in living human brain. I. Irreversible binding of ligands. *Journal of Cerebral Blood Flow & Metabolism* 6: 137–146. (b) From Gjedde A (2003) Modeling metabolite and tracer kinetics. In: Feinendegen LE, Shreeve, WW, Eckelman WC, et al. (eds.) *Molecular Nuclear Medicine*, ch. 7, pp. 121–169. Berlin-Heidelberg: Springer.

It is preferable to obtain the records at steady state, but the exact requirements are dictated by the pharmacology of the ligand. Ideally, the uptake would approach steady state or pseudoequilibrium very closely. If the tracer is given as a bolus administration, true steady state, of course, is never reached. If the tracer is continuously infused, it is possible to get very close to true equilibrium, depending on the tracer's half-life. Equilibrium conditions are ideal, both because comparative binding *in vitro* in principle is at equilibrium, and because the binding transients decline to zero. Unfortunately, some of the most selective ligands have the highest affinity and do not approach secular equilibrium during the 90–120 min available for PET.

An important question is the number of frames necessary to optimize the kinetic analysis. The higher the

number of frames, the lower the radioactivity recorded in each frame and the higher the random variability. The image variance exceeds the Poisson distribution and thus increases with the number of frames and lower counts per frame. If multiple short frames are recorded, there is a trade-off between the number of frames and the accuracy of the kinetic analysis, expressed in terms of parameter 'identifiability.' The information content can be shown to decline as a function of the number of parameters as the variability of the radioactivity measurements increases. There is another trade-off between too many and too few parameters. Many researchers believe that the maximum number of parameters that can safely be derived from PET and SPECT is four, possibly as many as five, consistent with the existence of three compartments.

Estimates of the receptor density are not routine, because they require significant occupation of the receptors with a ligand, which in principle must then have a pharmacodynamic effect. The procedure consists of giving a pharmacologically active dose of a ligand together with the tracer (e.g., haloperidol with [^{11}C]NMSP, or raclopride with [^{11}C]raclopride). As many as five different estimates of the binding potential have been made to complete Scatchard or Eadie–Hofstee plots of the binding, as shown in Figure 3.

In the past and rather unrealistically, it was believed that the binding of radioligands to neurotransmitter receptors would not be affected by the endogenous ligand, but this appears not to be the case. Although the interference could be considered a hindrance to the proper quantification of the receptor density, it has instead permitted indirect and noninvasive investigations of the fluctuations of transmitter release, particularly with respect to the dopaminergic neurotransmission. The promise and excitement of this approach notwithstanding, the reality of the interpretations is the subject of considerable current controversy, as not only transmitter concentration but also receptor number and affinity affect the estimates of binding potentials that may change as a simple consequence of an affinity change.

Altered Monoaminergic Neurotransmission *In Vivo*

Binding potentials have been determined for a number of monoaminergic tracers in many parts of the brain in both health and disease. Figure 4 shows the coincident maps of average dopa decarboxylase activity, dopamine transporter binding potentials, and dopamine receptor binding potentials in normal human beings. The maps are mutually confirmatory, but the individual numbers are not particularly informative in the absence of related evidence of biology or pathology. The noninvasive probing of monoaminergic

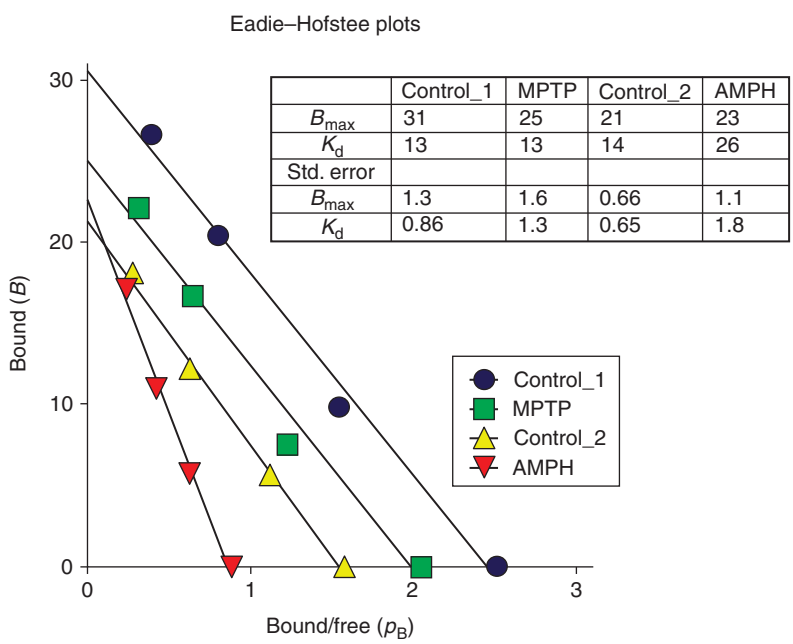

	Control_1	MPTP	Control_2	AMPH
B_{max}	31	25	21	23
K_d	13	13	14	26
Std. error				
B_{max}	1.3	1.6	0.66	1.1
K_d	0.86	1.3	0.65	1.8

Figure 3 Scatchard plots of dopamine receptor density and affinity, determined with [^{11}C]raclopride in monkeys rendered parkinsonian by means of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) and methamphetamine (AMPH) administration, in comparison with control monkeys (B, bound; F, free; P_B, binding potential; K_d is special case of K_m for receptor binding; see text for discussion). From Holden JE, Jivan S, Ruth TJ, et al. (2002) *In vivo* receptor assay with multiple ligand concentrations: An equilibrium approach. *Journal of Cerebral Blood Flow & Metabolism* 22: 1132–1141.

neurotransmission in living human brain also promises to clarify the pharmacokinetic and pharmacodynamic mechanisms of expectation, reward, and drug effects in healthy human beings. It is now possible to map the increase of dopamine occupancy that follows when humans become aware of and assign meaning to meaningful stimuli in the environment (Figure 5).

It is well known that the pathological changes occurring in any of the three monoaminergic systems precede the presentation of overt clinical symptoms, and changes of dopaminergic, serotonergic, and noradrenergic neurotransmission are therefore actively monitored in major neurological and neuropsychiatric disorders such as Parkinson's disease, schizophrenia, stroke, depression, and dementia. Not only are such studies of great diagnostic value but they also help avoid over- or underdosing of human patients with the relevant medications.

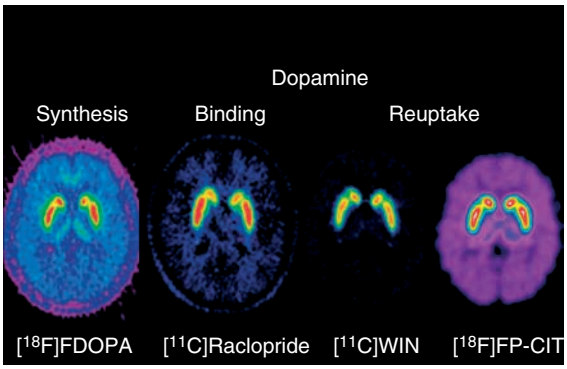

Figure 4 Images of dopaminergic neurotransmission in human striatum, with [^{18}F]FDOPA (fluoro-l-dopa), for dopamine synthesis capacity; [^{11}C]raclopride, for dopamine receptor availability; and [^{11}C]WIN 35,428 (CFT; 2β-carbomethoxy-3β-4-(fluorophenyl) tropane) and [^{18}F]FP-CIT (fluoropropyl β-carbomethoxy-3β-(4-iodophenyl)tropane), for dopamine transporter availability.

Dopaminergic Disorders

The myriad number of sites that can now be imaged are relevant to multiple disorders, as exemplified for the dopamine systems in Table 4. The imaging of dopamine's receptors has shown that they are lost as a function of normal aging. Striking clinical examples of change include abnormalities at multiple steps in the neurotransmission of dopamine in disorders as diverse as Parkinson's disease, schizophrenia, and addiction, including cocaine and methamphetamine abuse.

Parkinson's disease and schizophrenia in many ways represent opposite ends of a spectrum of perturbed dopaminergic neurotransmission in striatum and midbrain. In schizophrenia, PET and SPECT reveal increased capacity of dopamine synthesis from DOPA, increased tonic and phasic dopamine release, elevated density of dopamine D_2-like receptors, and increased extracellular dopamine. These findings explain how antagonists of dopaminergic neurotransmission can mitigate the worst effects of the disease, at least for some subtypes with positive symptoms, but complete reversal has not been achieved, probably because the primary disorder is cortical. For antipsychotics as well as a number of antidepressants it has been shown that occupancy of 60–80% of the dopamine receptors yields the best therapeutic effect. The receptor occupancy plotted against the plasma level of the drug essentially is a Michaelis–Menten graph of the therapeutic 'window of opportunity' for a neuroleptic. This appears also to be true for partial agonists as well as antagonists of the dopamine receptors.

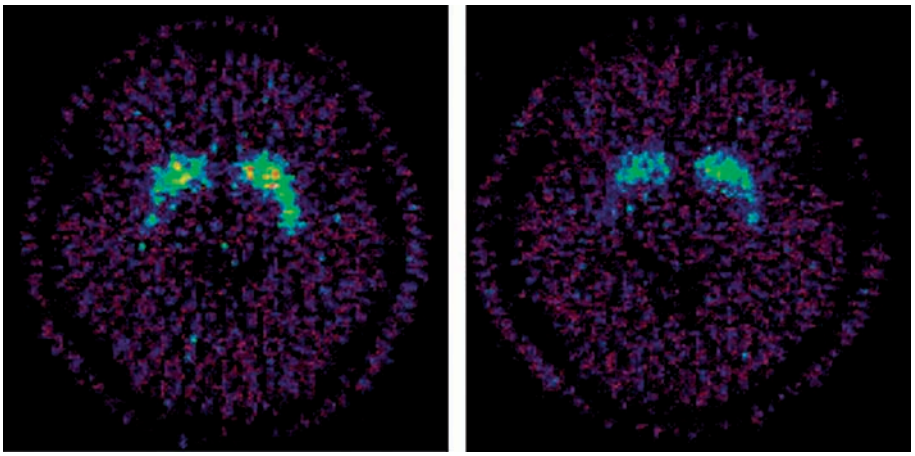

Figure 5 Reduction of dopamine receptor availability ([^{11}C]raclopride binding potential) in cognitive awareness. Left panel, control state; right panel, meditation state. Note that decline of raclopride binding is indicative of increased dopamine binding. From Kjaer TW, Bertelsen C, Piccini P, et al. (2002) Increased dopamine tone during meditation-induced change of consciousness. *Brain Research – Cognitive Brain Research* 13: 255–259.

Table 4 Changes of dopaminergic neurotransmission *in vivo*

Condition	*Variable*[a]				
	D_2-like receptor availability or density	*DDC enzyme activity*	*DA 'release' or DA occupancy*	*Extracellular or baseline DA concentration*	*DAT availability or density*
Cognitive awareness	–	–	Increase (1)	–	–
Normal aging	Decrease 7–10% per decade (2)	Decrease (3)	–	–	–
Schizophrenia	Two- to three-fold increase (some studies: 4, 5); no change (6, others)	Increase (multiple studies: e.g., 7, 8)	Increase (9, 10)	Increase (11, 12)	–
Bipolar disease/ psychosis	Two-fold increase (13)	–	–	Increase (12)	–
Gilles de la Tourette's syndrome	Increase (14)	–	Increase (15a,15b)	–	–
Attention-deficit/ hyperactivity disorder	Increase (16)	–	Increase (17)	–	Decrease (18)
Temporal lobe epilepsy w/psychosis	–	Increase (19)	–	–	–
Cocaine abuse	Decrease (20)	–	Decrease (i.e., attenuation of normal increase) (21)	–	Decrease (22)
Alcoholism	Decrease?	–	Decrease?	–	–
Parkinson's disease	Increase in untreated patients (23)	Decrease (multiple studies (24, 25)			Decrease (26, 27)

[a]Numbers in parentheses refer to the following studies: 1, Kjaer TW, Bertelsen C, Piccini P, et al. (2002) Increased dopamine tone during meditation-induced change of consciousness. *Brain Research – Cognitive Brain Research* 13: 255–259; 2, Wong DF, Wagner HN Jr., Dannals RF, et al. (1984) Effects of age on dopamine and serotonin receptors measured by positron tomography in the living human brain. *Science* 226: 1393–1396; 3, Kumakura Y, Vernaleken I, Gründer G, et al. (2005) PET studies of net blood–brain clearance of FDOPA to human brain: Age-dependent decline of [^{18}F]fluorodopamine storage capacity. *Journal of Cerebral Blood Flow & Metabolism* 25: 807–819; 4, Tune LE, Wong DF, Pearlson G, et al. (1993) Dopamine D2 receptor

density estimates in schizophrenia: A positron emission tomography study with ^{11}C-*N*-methylspiperone. *Psychiatry Research* 49(3): 219–237; 5, Wong DF, Wagner HN Jr., Tune LE, et al. (1986) Positron emission tomography reveals elevated D_2 dopamine receptors in drug naive schizophrenics. *Science* 234: 1558–1563; 6, Farde L, Hall H, Ehrin E, et al. (1986) Quantitative analysis of D_2 dopamine receptor binding in the living human brain by PET. *Science* 231: 258–261; 7, Ernst M, Zametkin AJ, Jons PH, et al. (1999) High presynaptic dopaminergic activity in children with Tourette's disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* 38: 86–94; 8, Rousset OG, Ma Y, Kamber M, et al. (1993) 3D simulations of radiotracer uptake in deep nuclei of human brain. *Computerized Medical Imaging and Graphics* 17: 373–379; 9, Breier A, Su TP, Saunders R, et al. (1997) Schizophrenia is associated with elevated amphetamine-induced synaptic dopamine concentrations: Evidence from a novel positron emission tomography method. *Proceedings of the National Academy of Sciences of the United States of America* 94: 2569–2574; 10, Laruelle M, Abi-Dargham A, van Dyck CH, et al. (1996) Single photon emission computerized tomography imaging of amphetamine-induced dopamine release in drug-free schizophrenic subjects. *Proceedings of the National Academy of Sciences of the United States of America* 93: 9235–9240; 11, Abi-Dargham A, Rodenhiser J, Printz D, et al. (2000) Increased baseline occupancy of D2 receptors by dopamine in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 97: 8104–8109; 12, Gjedde A and Wong DF (2001) Quantification of neuroreceptors in living human brain. V. Endogenous neurotransmitter inhibition of haloperidol binding in psychosis. *Journal of Cerebral Blood Flow & Metabolism* 21: 982–994; 13, Pearlson GD, Wong DF, Tune LE, et al. (1995) *In vivo* D2 dopamine receptor density in psychotic and nonpsychotic patients with bipolar disorder. *Archives of General Psychiatry* 52(6): 471–477; 14, Wong DF, Singer HS, Brandt J, et al. (1997) D2-like dopamine receptor density in Tourette syndrome measured by PET. *Journal of Nuclear Medicine* 38: 1243–1247; 15a, Singer HS, Szymanski S, Giuliano J, et al. (2002) Elevated intrasynaptic dopamine release in Tourette's syndrome measured by PET. *American Journal of Psychiatry* 159: 1329–1336; Wong DF, Brašić J, Singer H, et al. (2007) Mechanisms of dopamine serotonin in Tourette syndrome: Clues from an *in vivo* neurochemistry study with PET. *Neuropsychopharmacology* (online); 16, Lou HC, Rosa P, Pryds O, et al. (2004) ADHD: Increased dopamine receptor availability linked to attention deficit and low neonatal cerebral blood flow. *Developmental Medicine & Child Neurology* 46: 179–183; 17, Rosa-Neto P, Lou HC, Cumming P, et al. (2005) Methylphenidate-evoked changes in striatal dopamine correlate with inattention and impulsivity in adolescents with attention deficit. *NeuroImage* 25: 868–876; 18, Volkow ND, Wang GJ, Newcorn J, et al. (2007) Brain dopamine transporter levels in treatment and drug naive adults with ADHD. *NeuroImage* 34: 1182–1190; 19, Reith J, Benkelfat C, Sherwin A, et al. (1994) Elevated dopa decarboxylase activity in living brain of patients with psychosis. *Proceedings of the National Academy of Sciences of the United States of America* 91: 11651–11654; 20, Volkow ND, Fowler JS, and Wang GJ (2003) Positron emission tomography and single-photon emission computed tomography in substance abuse research. *Seminars in Nuclear Medicine* 33(2):114–128; 21, Wong DF, Kuwabara H, Schretlen DJ, et al. (2006) Increased occupancy of dopamine receptors in human striatum during cue-elicited cocaine craving. *Neuropsychopharmacology* 31: 2716–2727; 22, Gjedde A and Wong DF (2001) Quantification of neuroreceptors in living human brain. V. Endogenous neurotransmitter inhibition of haloperidol binding in psychosis. *Journal of Cerebral Blood Flow & Metabolism* 21: 982–994; 23, Rinne JO, Laihinen A, Rinne UK, et al. (1993) PET study on striatal dopamine D2 receptor changes during the progression of early Parkinson's disease. *Movement Disorders* 8: 134–138; 24, Hoshi H, Kuwabara H, Léger G, et al. (1993) 6-[^{18}F]fluoro-l-DOPA metabolism in living human brain: A comparison of six analytical methods. *Journal of Cerebral Blood Flow & Metabolism* 13: 57–69; 25, Kumakura Y, Gjedde A, Danielsen EH, et al. (2006) Dopamine storage capacity in caudate and putamen of patients with early Parkinson's disease: Correlation with asymmetry of motor symptoms. *Journal of Cerebral Blood Flow & Metabolism* 26: 358–370; 26, Kumakura Y, Gjedde A, Danielsen EH, et al. (2006) Dopamine storage capacity in caudate and putamen of patients with early Parkinson's disease: Correlation with asymmetry of motor symptoms. *Journal of Cerebral Blood Flow & Metabolism* 27: 358–370; 27, Varrone A, Marek KL, Jennings D, et al. (2001) [(^{123})I]β-CIT SPECT imaging demonstrates reduced density of striatal dopamine transporters in Parkinson's disease and multiple system atrophy. *Movement Disord*ers 16: 1023–1032.

Remarkable changes of dopamine synthetic capacity and dopamine transporter availability have been shown in Parkinson's disease. The cause of the disease is not known, but two different forms could exist in theory, one in which the initial insult is a decline of vesicular incorporation of dopamine, leading to neurotoxic levels of dopamine in the dopaminergic terminals, and one in which the initial insult is a loss of dopamine synthesis capacity. Dopamine will be low in both cases, but the processes of synthesis and metabolism of dopamine will express opposite changes.

Attention-deficit/hyperactivity disorder (ADHD) is suspected of being a dopamine deficiency disorder related to a class of disturbances known as reward deficiency syndrome. PET reveals increased dopamine receptor availability (higher than normal binding potential) in ADHD, consistent with lower than normal extracellular dopamine level, which can be rectified by blocking the dopamine reuptake sites with methylphenidate, as shown in Figure 6. The reason for the deficiency may be abnormally high activity of the dopamine transporters.

Dopaminergic neuroimaging is particularly important to the investigation of substance abuse: the challenge is to understand the interactions of receptors and enzymes with the substances abused. PET has revealed that the main effect of blockade of dopamine transporters by cocaine, amphetamine, or methamphetamine is an increase in dopamine's occupancy of its receptors because of increased extracellular dopamine, as evidenced by decline of the binding potential of a dopamine receptor ligand. Other monoamine transporters are also affected because of the comparatively low selectivity of the transporters toward the blockers.

Many studies reveal decreased extracellular dopamine concentrations and reduced dopamine D_2 receptor availability and density in abstinent abusers, which would tend to exacerbate the dopamine deficiency. Indeed, the monoamine receptors tend to be the direct or indirect, intended or accidental, targets of a wide range of illicit drugs enjoyed recreationally by humans. The drugs that are abused most frequently (opiates, ethanol, nicotine, amphetamine, and cocaine) acutely increase extracellular dopamine concentrations in striatum and nucleus accumbens. In contrast, drugs that are not abused by humans do not appear to affect the dopamine occupancy, while drugs with aversive properties that are actively avoided tend to lower dopamine occupancy, indirectly proving that activation of dopaminergic neurotransmission is a fundamental property of drugs of abuse.

Continued abuse of cocaine or other blockers of monoamine transporters attenuates the increase of occupancy elicited by amphetamine and lowers the density of the transporters in the striatum, measured shortly after the onset of abstinence. The transporter availability rises again after longer periods of abstinence, but cue-induced craving for cocaine in abstinent abusers can raise dopamine's occupancy of its receptors without the aid of cocaine, probably because of cue-induced release of dopamine. Table 5 summarizes findings of dopamine transporter availability in three groups of study participants, as measured with the dopamine transporter ligand [^{11}C]WIN 35,428.

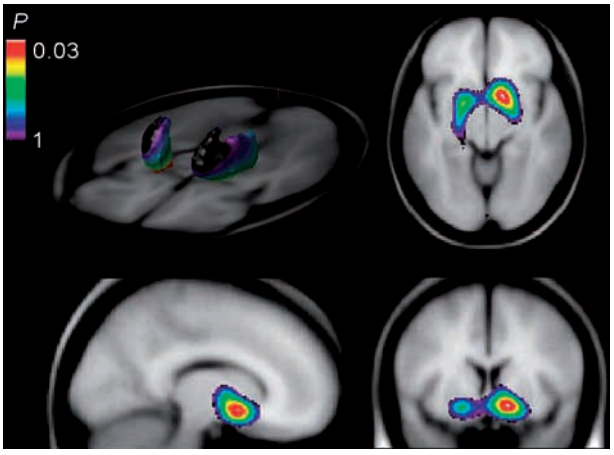

Figure 6 Reduction of dopamine receptor availability ([^{11}C] raclopride binding potential) by treatment with methylphenidate in patients with attention-deficit/hyperactivity disorder. From Rosa-Neto P, Lou HC, Cumming P, et al. (2005) Methylphenidate-evoked changes in striatal dopamine correlate with inattention and impulsivity in adolescents with attention deficit. *NeuroImage* 25: 868–876.

Serotonergic Disorders

Serotonin receptors are most abundant in cerebral cortex, particularly in the hippocampal region and the ventromedial prefrontal cortex, where PET also shows that binding potentials at $5HT_{1A}$ receptors decline with age and in depression, with return to normal after antidepressant treatment. The changes are somewhat equivocal because it is uncertain

Table 5 Dopamine transporter binding potentials in striatum

Group	*Number of participants*	*Binding potentials (k_3/k_4)*	
		Caudate nucleus	*Putamen*
Healthy volunteers	10	7.5 ± 1.7	7.3 ± 1.4
Methamphetamine abuse	6	5.8 ± 1.0	5.5 ± 0.9
Parkinson's disease	3	4.0 ± 0.9	2.3 ± 0.4

whether the changes are due to change of receptor density or change of serotonin level. Abnormalities of occupancy at these receptors have been difficult to show with the available radioligands because serotonin reuptake inhibitors have little effect on the binding potential. The decline of $5HT_{1A}$ binding potentials in aging is shown in Figure 7. Changes attributable to loss of serotonin have been observed after stroke in patients who develop pathological crying, which is relieved instantly by administration of selective serotonin reuptake inhibitors.

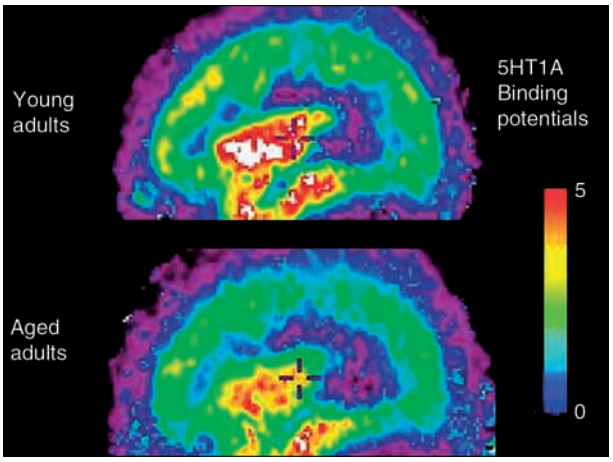

Figure 7 Reduction of serotonin ($5HT_{1A}$) receptor availability ([^{11}C]WAY binding potential) in aging. From Moller M, Jakobsen S, and Gjedde A (2007) Parametric and regional maps of free serotonin 5HT(1A) receptor sites in human brain as function of age in healthy humans. *Neuropsychopharmacology* 32(8):1707–1714.

The drug MDMA (3,4-methylenedioxy-*N*-methylamphetamine), commonly known by the street name Ecstasy, is a semisynthetic entactogen of the phenethylamine family, which affects the serotonin system by interaction with serotonergic and dopaminergic terminals. The acute effect of MDMA appears to result from the drug's interference with dopamine and serotonin reuptake. PET results obtained with the radioligands McN5652 and 3-amino-4-(2-dimethylaminomethylphenylsulfanyl)-benzonitrile (DASB) in humans and shown in Figure 8 suggest that MDMA abusers suffer a measurable reduction of serotonin transporter availability in multiple brain regions, which is relieved after abstinence, although the evidence is still equivocal.

Design of Drugs for the Monoaminergic Systems

PET and SPECT imaging techniques play an important role in the development of drugs that specifically target abnormalities that PET and SPECT have identified in disease. The applications typically involve imaging of the biodistribution of candidate drugs labeled with ^{11}C or ^{18}F for PET, or ^{123}I or ^{99m}Tc for SPECT. The areas of application include (1) therapeutic rationale, (2) justification for the study of specific biological systems, (3) mechanisms of drug action, and (4) rational drug dosing. Much of this work can be carried out in neuroreceptor imaging laboratories of universities

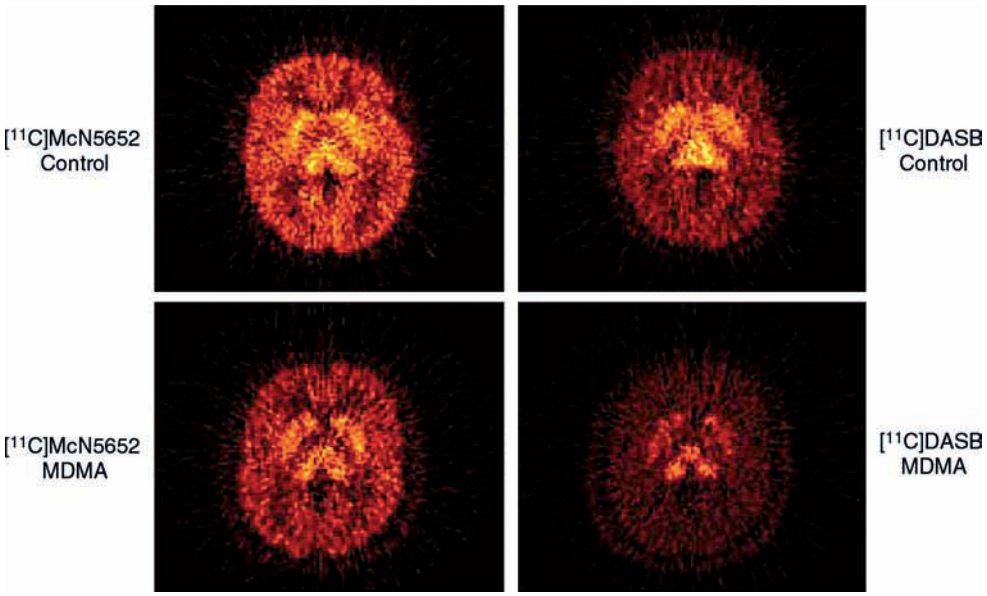

Figure 8 Effect of methylenedioxymethylamphetamine (MDMA; 'Ecstasy') on serotonin transporter availability – [^{11}C]McN5652 and [^{11}C]DASB (3-amino-4-(2-dimethylaminomethyl-phenylsulfanyl)-benzonitrile) binding potentials. From McCann UD, Szabo Z, Seckin E, et al. (2005) Quantitative PET studies of the serotonin transporter in MDMA users and controls using [^{11}C]McN5652 and [^{11}C]DASB. *Neuropsychopharmacology* 30: 1741–1750.

that negotiate master agreements with pharmaceutical companies for evaluations of ligands that will be used in the early stages of drug development.

Conclusions

In summary, imaging approaches to the study of monoaminergic neurotransmission have relevance to both neurological and psychiatric diseases. Historically, neuroreceptor imaging focused on the pathophysiology of major neuropsychiatric disorders such as schizophrenia, Parkinson's disease, and depression, but it is now also being used to study and diagnose dementia, stroke, and drug abuse, and to assist in the development of new drugs. The method of imaged neurochemistry is emerging as an effective means of clarifying the mechanisms and neuroadaptive properties of brain disease related to monoaminergic neurotransmission. Currently, a wealth of radiotracers of the monoaminergic systems can be brought to bear on the unraveling of these disorders. The efforts help provide rational drug dosing and examine mechanisms of action of drugs that are relevant to cognitive research. This will improve objectivity in research and clarify drug mechanisms and treatment pathways in the future.

See also: Dopamine – CNS Pathways and Neurophysiology; 3,4-Methylenedioxymethamphetamine (MDMA, "Ecstasy"); Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Monoamines; Monoamines: Release Studies; Norepinephrine: CNS Pathways and Neurophysiology; Octopamine and Other Monoamines in Invertebrates; Positron Emission Tomography (PET); Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors; Single Photon Emission Computed Tomography (SPECT); Trace Monoamines and Receptors in Mammalian CNS.

Further Reading

Backman L, Nyberg L, Lindenberger U, et al. (2006) The correlative triad among aging, dopamine, and cognition: Current status and future prospects. *Neuroscience & Biobehavioral Reviews* 30: 791–807.

Bantick RA, Deakin JF, and Grasby PM (2001) The 5-HT_{1A} receptor in schizophrenia: A promising target for novel atypical neuroleptics? *Journal of Psychopharmacology* 15: 37–46.

Carson RE (2000) PET physiological measurements using constant infusion. *Nuclear Medicine & Biology* 27: 657–660.

Doudet DJ (2001) PET studies in the MPTP model of Parkinson's disease. *Advances in Neurology* 86: 187–195.

Fowler JS, Logan J, Volkow ND, et al. (2005) Translational neuroimaging: Positron emission tomography studies of monoamine oxidase. *Molecular Imaging and Biology* 7: 377–387.

Frankle WG, Slifstein M, Talbot PS, et al. (2005) Neuroreceptor imaging in psychiatry: Theory and applications. *International Review of Neurobiology* 67: 385–440.

Gjedde A (2003) Modeling metabolite and tracer kinetics. In: Feinendegen LE, Shreeve WW, Eckelman WC, et al. (eds.) *Molecular Nuclear Medicine,* ch. 7, pp. 121–169. Berlin-Heidelberg: Springer.

Gjedde A, Wong DF, Rosa-Neto P, et al. (2005) Mapping neuroreceptors at work: On the definition and interpretation of binding potentials after 20 years of progress. *International Review of Neurobiology* 63: 1–20.

Holden JE and Doudet DJ (2004) Positron emission tomography receptor assay with multiple ligand concentrations: An equilibrium approach. *Methods in Enzymology* 385: 169–184.

Laakso A and Hietala J (2000) PET studies of brain monoamine transporters. *Current Pharmaceutical Design* 6: 1611–1623.

Meyer JH (2007) Imaging the serotonin transporter during major depressive disorder and antidepressant treatment. *Journal of Psychiatry and Neuroscience (Canada)* 32: 86–102.

Seibyl JP (2003) Imaging studies in movement disorders. *Seminars in Nuclear Medicine* 33: 105–113.

Volkow ND, Fowler JS, and Wang GJ (2003) Positron emission tomography and single-photon emission computed tomography in substance abuse research. *Seminars in Nuclear Medicine* 33(2): 114–128.

Volkow ND, Wang GJ, Fowler JS, et al. (2005) Imaging the effects of methylphenidate on brain dopamine: New model on its therapeutic actions for attention-deficit/hyperactivity disorder. *Biological Psychiatry* 57: 1410–1415.

Wong DF (2002) *In vivo* imaging of D2 dopamine receptors in schizophrenia: The ups and downs of neuroimaging research. *Archives of General Psychiatry* 59: 31–34.

Monoamines: Release Studies

A D Smith, A C Michael, B J Lopresti, R Narendran, and M J Zigmond, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons in the brain act on their targets via the release of chemical transmitters, a process that is generally assumed to occur across a synapse. In the case of monoamines, there is reason to believe that the field of influence can be considerably larger and often involves nonsynaptic or 'volume transmission.' Even this larger space between a given monoaminergic neuron and its target is likely to defy direct chemical monitoring, and there is a long history of attempts to overcome this hurdle through a wide variety of indirect approaches. Many of these approaches have been reviewed in the past. This article reexamines the issue from the standpoint of current technology and understanding.

Monoamine neurotransmitters, by virtue of their easily oxidized catechol or indole functional groups, can be detected by means of electrochemistry. This has spurred the development of two technologies for monoamine analysis: one comprises electrochemical detectors coupled to chromatography and electrophoresis systems; the other comprises microelectrodes implanted directly into brain tissue for *in vivo* measurements, so-called *in vivo* voltammetry. A third approach, the use of positron emission tomography (PET), is based on an entirely different principle, which is that once released, an endogenous monoamine can displace a radioligand that also binds to a receptor on which the monoamine acts. This article reviews each of these three approaches to monitoring monoamine release in the intact brain.

Voltammetry

Electrochemistry has revolutionized the capacity to detect monoamine release. Electrochemical detection and voltammetry both use electrochemistry but are otherwise fundamentally different techniques. In the former, the electrode is intended to be nonselective, giving a signal for all oxidizable components of the sample after their separation on a column. It is of great value in studies of fluid samples that can be collected and then analyzed via high-performance liquid chromatography (HPLC). It will be considered at some length in the next section. In contrast, in voltammetry, various strategies, including customized potential waveforms and tailored electrode surfaces, are used to render the electrode as selective as practicable for a target monoamine. Absolute selectivity between the monoamines is not possible by electrochemical means alone, so experimental conditions are often chosen carefully in light of this point. For example, *in vivo* voltammetry is often applied in regions of the brain where a single monoamine is predominant, as in the case of DA within the striatum, to measurements of monoamine release during selective stimulation of pathways; or to measurements of an exogenously applied monoamine. Rather than selectivity, the forte of *in vivo* voltammetric techniques is their sensitivity, speed of response, and spatial resolution. Indeed, the technology has steadily improved over the course of the last three decades to the point that voltammetry provides access to information not available by other means.

Electrode Design

Although a variety of microelectrode designs are described in the literature, the majority of recent *in vivo* measurements of monoamines have relied on carbon fiber microelectrodes (Figure 1). Carbon fibers, which are available with diameters as small as 5 μm, are usually sealed into a pulled glass capillary with epoxy or sometimes just by virtue of the glass–carbon junction. The fiber protruding from the glass tip can be cut to a desired length, or the tip can be beveled on a polishing wheel so that only the cross section of the fiber is exposed. Although these beveled electrodes are usually called microdisk electrodes, they are actually elliptical in shape. The microdisk geometry is useful when high spatial resolution is needed, such as during the recording of individual exocytotic events, whereas the longer electrodes are suitable when spatial averaging is preferred.

The tiny dimensions of carbon fiber microelectrodes offer several benefits to *in vivo* voltammetry. The electrodes, for example, do very little damage to brain tissue during implantation. Electron microscopy of striatal tissue from anesthetized rats following acute carbon fiber implantation showed that the majority of tissue damage was confined to within approximately 3 μm of the electrode track. This short distance permits faithful recording of *in vivo* chemical events with minimal diffusional distortion because small molecules need only about 10 ms to diffuse 3 μm through the brain extracellular space. Thus, the minimization of tissue trauma contributes to the rapid speed of response available with *in vivo* voltammetry.

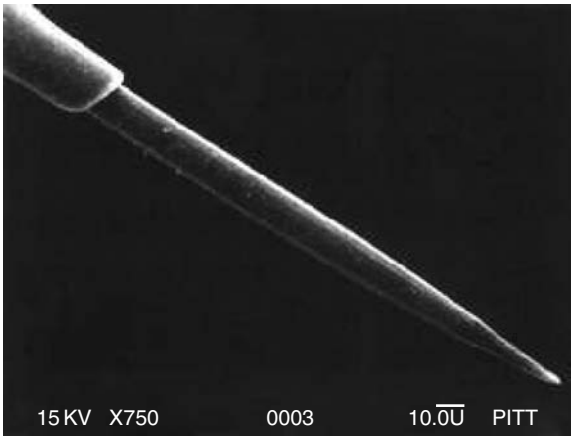

Figure 1 Scanning electron micrographs illustrate the features and dimensions of the carbon fiber microelectrodes used in the majority of *in vivo* voltammetry studies. The fibers in the image have a diameter of 7 μm. The fibers are sealed into a pulled glass pipet with epoxy. The microcylinder electrode protrudes from the pipet, whereas the microdisk is prepared by polishing the tip on a microelectrode beveller.

Fast-Scan Cyclic Voltammetry

Another benefit of the small dimensions of the microelectrodes is the small capacitance of the electrode surface. The small capacitance allows for rapid changes in the potential applied to the electrode, which has enabled the technique called fast-scan cyclic voltammetry. In this technique, the electrode is subjected to linear potential sweeps at rates of 300 V s^{-1} or more. Since monoamines can be detected with potential sweeps spanning about 3 V, individual voltammetric scans are completed in 10 ms or less. The ability to rapidly scan the potential, therefore, is another contributor to the rapid speed of response. Furthermore, as the potential sweep rate increases, voltammetry becomes selective for monoamine neurotransmitters over their acid metabolites and other oxidizable substances, including ascorbate and urate, encountered in brain tissue. This is due to a quirk of the kinetics of the oxidation and reduction reactions of the monoamines, which are substantially faster than those of the potential interfering substances.

The fast-scan cyclic technique generates a voltammogram, a plot of the electrode current versus the applied potential, which provides a fingerprint of the detected substance. Recently, as discussed further in the section titled 'Recording naturally occurring DA release on a subsecond timescale,' statistical analysis of these fingerprints by means of principal components analysis has greatly advanced the analytical capabilities of in voltammetric recording techniques.

While the rapid potential scanning approach has several advantages, they come at a price. Rapidly scanning the potential induces a flow of current called the electrode charging current. Even though the electrode capacitance is small in an absolute sense, the charging current is massive in comparison to the currents associated with the detection of monoamines. Consequently, the monoamine-associated current appears as a tiny signal component superimposed on a much, much (1000 times) larger background current. For this reason, voltammetry is well suited only to differential measurements, that is, voltammetry can be used to track changes in monoamine concentrations from a starting condition. On the other hand, voltammetry is not able to tell us 'absolute' or basal monoamine concentrations. Fortunately, much useful insight into the functional aspects of brain monoamine systems has been derived from differential voltammetric measurements.

Constant Potential Amperometry

As an alternative to fast-scan voltammetry, constant potential amperometry has proved a highly useful recording technique. Here, the electrode potential is simply poised at a value appropriate to the oxidation of monoamines, usually near 0.6 V versus Ag/AgCl. This approach provides no selectivity and thus is suitable only in those specialized experimental conditions in which there is no doubt about the identity of the substance being detected. However, several such conditions exist. One example is the use of the clearance of exogenously applied monoamine to evaluate transporter kinetics. Another example is electrical stimulation of the medial forebrain bundle to selectively release DA in the striatum and accumbens. In these experiments, the speed of the recording tends to be of the utmost concern, over and above the selectivity of the response. In the case of constant potential amperometry, the speed of the recording is determined by the data acquisition hardware rather than the features of an applied potential waveform.

Characterizing Monoamine Transporter Kinetics

One recent and extensive application of *in vivo* voltammetry has been to characterize the kinetics of the DA transporter (DAT) in several dopaminergic regions of the rat brain. DAT is responsible for clearing DA from the extracellular space to terminate the actions of DA on pre- and postsynaptic receptors. Two main strategies for evaluating DAT kinetics are described in the literature. One involves recording the clearance of extracellular DA following the electrical stimulation of DA axons in the medial forebrain. Detailed analysis of the descending phase of the stimulus response leads to values of the V_{max} and K_M of the clearance process. The ability to determine these kinetic parameters has engendered detailed study of

the factors that regulate them, include drugs and autoreceptor mechanisms. Much attention has also been focused on regional differences in DAT kinetics, a topic of long-standing interest in the field of substance abuse.

A second approach to evaluating transporter kinetics is to use *in vivo* voltammetry to monitor the clearance of exogenously applied substrate. Applications by both pressure ejection and iontophoresis have been described. A point that remains incompletely resolved is that these methods produce substantially smaller kinetic constants than those obtained by monitoring DA clearance after electrical stimulation. There has been speculation that this may be due to the voltage dependency of the transporter that might be affected by stimulation of an as yet unspecified diffusional process during the application of substrate. Nevertheless, the technique reveals modulation of clearance rates by drugs and anesthetics and reveals regional variations in transporter activity. Daws and co-workers have extended this procedure to 5-HT, reporting recently, for example, that the effect of ethanol on 5-HT clearance is enhanced in mice lacking the 5-HT transporter, a finding that sheds new light on the role of 5-HT in alcoholism. At present, however, voltammetry has not been extensively used to monitor endogenous 5-HT release.

Autoreceptor Control of DA Release

Considerable use has been made of voltammetry to examine the control of DA release by receptors present on the DA terminal itself, the so-called autoreceptors. In one such experiment by François Gonon and co-workers, amperometric recording was used to examine the time course of the control of evoked DA release by autoreceptors. Their data indicate that the D2 subtype of DA receptors can modulate DA release on a msec timescale and that the duration of the autoreceptor-mediated control of DA release is far longer than the stimulus-evoked elevation in DA itself, suggesting a role for downstream intracellular second messengers. Such studies both illustrate the power of high-speed recordings cases in which the identity of the monoamine is not in question and provide a task to which voltammetry is uniquely suited.

Recording Naturally Occurring DA Release on a Subsecond Timescale

Recently, R. Mark Wightman and his colleagues have described the voltammetric recording of naturally occurring DA release on the subsecond timescale. This represents a significant advance in the capabilities of *in vivo* voltammetry, which previously had been used mainly to monitor electrically evoked DA release. A key technical development that enabled this advance was the introduction of the principal components regression analysis of individual voltammograms recorded in the brain of conscious, freely moving rats. The regression algorithm separates the voltammetric signals arising from DA from those arising from pH fluctuations in the brain extracellular space. Fluctuations in pH trigger voltammetric signals at carbon fiber microelectrodes. However, the fingerprint voltammograms associated with pH changes are different from those associated with DA changes, which enables the regression analysis to distinguish pH and DA contributions to the signals. By separating these signal components, which statistically account for the entire signal, voltammetry achieves DA detection limits near 20 nmol l^{-1} and subsecond temporal resolution.

According to this voltammetric system, naturally occurring DA release comprises a series of subsecond transients in the nanomolar concentration range. It appears likely, although it remains to be definitively established, that the DA transients are associated with the burst firing of DA neurons. The amplitude, frequency, and duration of the transients are increased by DAT inhibitors, including cocaine and nomifensine, suggesting that the transients may play a key role in the psychostimulant properties of these drugs.

Microdialysis

Basis of Approach

Measurement of transmitter release via intracerebral dialysis is based on the combination of three separate techniques: (1) a dialysis membrane to collect sample by mass diffusion of small solutes across that membrane, (2) HPLC to separate out the several molecules that are collected in this manner, and (3) a detector to quantify the molecules of interest. In the case of monoamines, that detector is typically an electrochemical detector.

Microdialysis has a much slower timescale and provides much less anatomical resolution than voltammetry. On the other hand, microdialysis offers an unambiguous identification of a great many compounds, including the monoamines and their metabolites; provides absolute values; can be used to monitor the substances over many hours; and provides a means for delivering compounds as well as collecting them.

Diffusion is initiated when the inner surface and the outer surface of the dialysis membrane are in contact with two aqueous solutions, one containing a higher concentration of a substance of interest than the other. If the aqueous solution in contact with the inner surface of the membrane is continuously flowing and has a significantly lower concentration

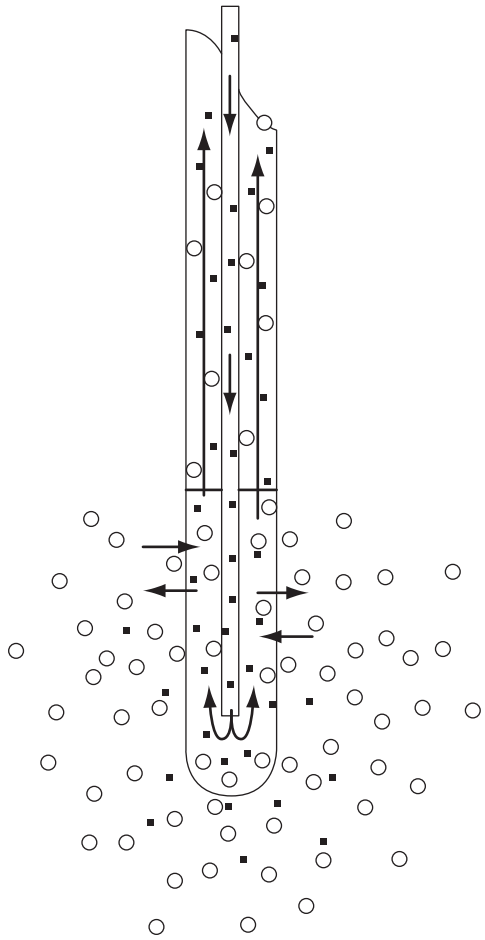

Figure 2 Schematic diagram of the dialysis process using a concentric tube microdialysis probe. A drug in the perfusate (squares) can dialyze outward into tissue, while small molecules in the tissue (circles) diffuse inward and are collected with the dialysate for subsequent assay. From Kissinger PT (1996) Electrochemical detection in bioanalysis. *Journal of Pharmaceutical and Biomedical Analysis* 14: 871–880.

of the substance of interest, a concentration gradient is created resulting in continuous diffusion of the substance from the external medium into the membrane. If the dialysis membrane is attached to an inflow and outflow line, the continuous flow of solution carries the substance to the end of the outflow, line where it can be collected for later analysis (Figure 2).

Early Development of Monitoring Techniques

Several sampling procedures preceded intracerebral dialysis as a means of measuring the chemical environment within the brain. The earlier technique of ventricular perfusion did not allow sampling of distinct brain regions, and cortical cup perfusion allowed sampling of only the cortical surface in anesthetized animals. Push–pull cannula perfusion, introduced by John Gaddum in 1961, was the first technique in which extracellular fluid from distinct subcortical

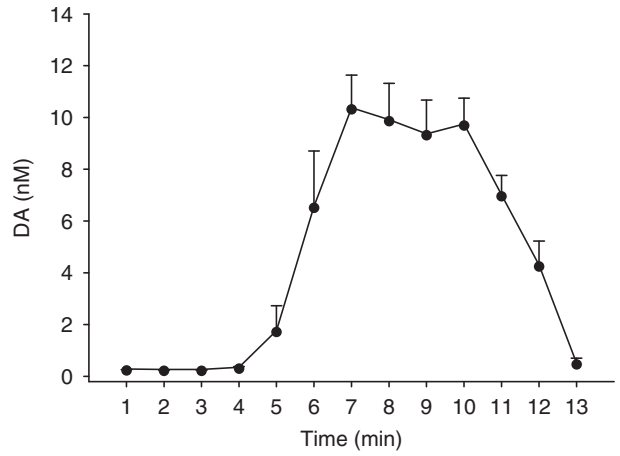

Figure 3 Temporal effect of local perfusion of KCl (100 mmol l^{-1}) on extracellular DA in the striatum.

regions could be measured. Two concentric tubes were implanted into the brain, with the tip being within the region of interest. The inner tube was used to deliver the perfusion solution and the outer tube to collect the perfusate for subsequent analysis. Because the perfusion fluid was in direct contact with brain tissue, significant tissue damage was incurred from the fluid flow.

By placing a semipermeable bag over the tip of the push–pull cannula to circumvent this problem, Jose Delgado and colleagues introduced the first intracerebral dialysis probe, called the dialytrode. With the advent of highly sensitive methods of detection, such as HPLC coupled with electrochemical detection, Urban Ungerstedt extended this approach, pioneering its use to measure endogenous DA in extracellular fluid in response to drugs such as amphetamine. Soon this was further extended to include stimuli such as KCl (Figure 3) and electrical stimulation and to monitor various transmitter releases during exogenous stimuli and behavior. *In vivo* microdialysis quickly became the most common means by which *in vivo* transmitter efflux was monitored, resulting in more than 5000 publications by 2007.

Advantages of Microdialysis

Microdialysis has several advantages over earlier *in vivo* sampling methods. The damage incurred by the surrounding extracellular environment is minimized not only by the presence of the dialysis membrane but also by a decrease in the overall size of the probe. Another advantage of the dialysis membrane is that it acts as a filter against large molecules, such as enzymes, that may act to degrade the substance of interest, and it also purifies the sample for subsequent analysis. Furthermore, because intracerebral dialysis is merely a sampling technique, in conjunction with a

suitable assay, it can be used to study a broad range of compounds. However, the most studied class of compounds has been the monoamines, DA in particular.

Although modern dialysis probes minimize the damage to the surrounding tissue, studies have shown that immediately after implantation of the probe, not all transmitter 'release' is either tetrodotoxin (TTX) sensitive or calcium dependent and thus may represent a response to damage. On the other hand, by the next day, release is both TTX sensitive and calcium sensitive, at least in the case of DA, and thus is generally considered to reflect a physiological process.

The release of DA is known to occur largely if not entirely via exocytosis, and its actions are terminated by removal from the synapse by a combination of diffusion and high-affinity uptake via DAT. These processes can be exploited to determine the functional state of the DA system in response to various treatments. For example, in a normal intact DA system, addition of a high concentration of K^+ ions to the perfusate increases the concentration of DA sampled by the probe via an increase in exocytotic release of DA. However, infusion of 6-hydroxydopamine, a catecholaminergic toxin, into the nigrostriatal DA pathway decreases potassium-evoked DA release, indicating that the system is not functioning at full capacity.

Amphetamine increases extracellular levels of DA by reversal of the DAT, a calcium-independent process. In a normal nigrostriatal DA system, amphetamine induces an approximate 50-fold increase of DA in dialysate. However, in the denervated striatum, amphetamine-induced DA release is attenuated. Monitoring extracellular monoamine levels under basal conditions or during stimulation may provide information about the functional integrity of a given projection. However, care must be taken to extrapolate from extracellular levels to release. Extracellular transmitter concentration is a net result of release and removal of transmitter, and removal in turn is a consequence of diffusion and high-affinity uptake. Likewise, care must be taken before an assumption that extracellular transmitter levels are a simple reflection of the number of nerve terminals. For example, extensive damage to a catecholaminergic system is required to reduce basal extracellular levels of either norepinephrine or DA. In addition, some trophic factors, particularly glial-derived neuronal factor, have been shown to increase the synthesis and release of DA without affecting the apparent number of DA terminals.

Problems with Microdialysis

A major disadvantage of microdialysis, because of the need for large amounts of material for detection, has been temporal resolution. For example, even in the case of extracellular DA, which is present in high concentration compared with other monoamines or even DA in other brain regions, a sampling interval of at least 10 min has typically been required. However, separation techniques such as capillary electrophoresis and microbore columns for HPLC, requiring smaller sample volumes, coupled with more sensitive detection methods ($\leq 1\,\mu l$) have allowed the temporal resolution to be decreased to 1 min. Of course, this resolution still lags far behind that of synaptic events and *in vivo* voltammetry, which are on the millisecond scale.

Another drawback of intracerebral dialysis has been an inability to measure the actual concentration of endogenous substances. At the crux of this problem is a description of the relationship between the dialysate concentration and the actual concentration of substance within the extracellular fluid, or the recovery. However, here again, techniques have been developed which allow relatively quantitative measurement of the extracellular concentration of neurotransmitters independent of the recovery. One method uses a very slow rate, at which the *in vivo* recovery approaches 100%. However, this method requires long sampling intervals, which dramatically reduces temporal resolution and also introduces problems of sample stability. A second approach to the problem is the difference method (or method of no net flux) developed by Peter Lönnroth. In this technique, concentrations of the compound of interest above and below what is expected to be present in the extracellular fluid are perfused through the probe and the resulting dialysate concentration measured. When the differences between the inflow and outflow concentrations are plotted versus the inflow concentration, the resulting regression yields the extracellular concentration (zero point) and the *in vivo* recovery. Again, a disadvantage of this technique is the time required for its execution.

A third problem with microdialysis is that the probe size, typically 250 μm, precludes fine spatial resolution and also is associated with considerable local tissue damage characterized by acute and chronic histological, physiological, and biochemical changes in adjacent tissue. An area of altered release and uptake exists within the trauma layer, which could effectively result in discrepancies between the extraction fraction and the relative recovery, rendering an extracellular concentration that is over- or underestimated.

Imaging Dopaminergic Transmission

The reduction in the availability of D2 receptors as measured with benzamide radioligands (such as ^{123}I-iodobenzamide (IBZM) and ^{11}C-raclopride) following the administration of psychostimulants known

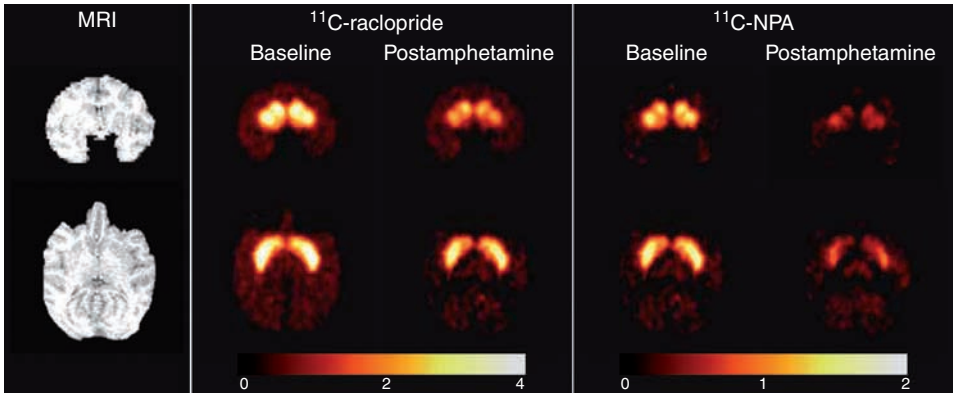

Figure 4 Effect of amphetamine on ^{11}C-raclopride and ^{11}C-(−)-*N*-propyl-norapomorphine (NPA) in a nonhuman primate (baboon). Magnetic resonance images (MRIs) and corresponding parametric binding potential (BP) maps shown in the same baboon under four conditions: ^{11}C-raclopride baseline study, ^{11}C-raclopride postamphetamine study (0.5 mg kg^{-1}), ^{11}C-NPA baseline study, and ^{11}C-NPA postamphetamine study (0.5 mg kg^{-1}). Following the same dose of amphetamine, 0.5 mg kg^{-1}, a greater decrease in BP is observed with the agonist radioligand ^{11}C-NPA (−46%) than with the antagonist ^{11}C-raclopride (−26%). From Narendran R, Hwang DR, Slifstein M, et al. (2004) *In vivo* vulnerability to competition by endogenous dopamine: Comparison of the D2 receptor agonist radiotracer (−)-*N*-[^{11}C]propyl-norapomorphine ([^{11}C]NPA) with the D2 receptor antagonist radiotracer [^{11}C]-raclopride. *Synapse* 52: 188–208.

to release DA (e.g., amphetamine or methylphenidate) is a clear example of an interaction between endogenous neurotransmitter and radioligand that has been extensively validated in both nonhuman primates and humans. In baboons, the increase in synaptic DA elicited by amphetamine is associated with a displacement of the binding of the single photon emission computed tomography (SPECT) D2 receptor radiotracer ^{123}I-IBZM and of the PET radiotracer ^{11}C-raclopride (Figure 4). By combining SPECT and microdialysis in baboons, Marc Laruelle and colleagues established that the magnitude of the ^{123}I-IBZM displacement after administration of amphetamine was related to the magnitude of DA release, and that a linear correlation existed between microdialysis measurements of DA release and radioligand displacement. Over the years, this technique of measuring the reduction in radiotracer binding following DA release via a stimulant challenge has been successfully used to study dopaminergic transmission in the living human brain in both healthy and diseased states such as schizophrenia and addiction.

Limitations of Imaging

The capacity to monitor neurotransmitter release in humans as well as experimental animals is a great advantage of *in vivo* imaging techniques employing radioligands. However, there are several limitations to these approaches. Some are similar to those of voltammetry: Both methods are applicable to a relatively small number of compounds. Indeed, only a few radioligands have been shown to have *in vivo* binding characteristics sensitive to fluctuations in endogenous neurotransmitter release. Other limitations are similar to those of microdialysis: limited spatial and temporal resolution. Although modern PET scanners can now provide images of the entire human brain with 2 mm spatial resolution, assessments of neurotransmitter release via these techniques are limited to the tissue level.

Other limitations are unique to imaging in general or to the specific radioligand employed. One example is the relatively low sensitivity of measuring endogenous transmitter by imaging its displacement of radiolabeled ligand. A 1% decrement in radiotracer binding corresponds to a 50% increase in extracellular DA as measured by microdialysis. Another limitation is the 'ceiling effect' seen when any of the available benzamide radiotracers are used: Only about half of the radiotracer's specific binding is vulnerable to changes in extracellular DA concentration.

Both the low sensitivity and the ceiling effect may be related at least in part to the fact that D2 receptors are configured in interconvertible states of high or low affinity for agonists. This is because benzamide radiotracers such as ^{123}I-IBZM and ^{11}C-raclopride are antagonists that bind with equal affinity to both states, but DA is not expected to compete efficiently with the fraction of the ^{123}I-IBZM or ^{11}C-raclopride binding to the low-affinity state. Thus, in contrast to the antagonist, a large fraction of the *in vivo* binding of an agonist radiotracer to D2 high sites will be available for endogenous competition by DA, thereby making it a more sensitive probe to study stimulant-induced DA release. Recent preclinical PET studies comparing novel agonist radioligands (in anesthetized nonhuman primates and cats) such as [^{11}C](−)-*N*-propyl-norapomorphine (NPA) and ^{11}C-PHNO with the reference antagonist radiotracer ^{11}C-raclopride tend to support this hypothesis. If the

increased sensitivity of these agonist radioligands extends to humans, they may provide researchers with a superior probe with which to report on acute changes in synaptic DA concentration. Such probes might also allow researchers to control for baseline affinity states (availability of receptors configured in D2 high) that may differ between clinical populations.

One limitation of the SPECT and PET radiotracers ^{123}I-IBZM and ^{11}C-raclopride is that the affinity of these ligands provides a quantifiable signal only in the striatum where receptors are abundant. However, a major advancement in availability of higher-resolution PET cameras has allowed the study of amphetamine-induced DA release in the functional subdivisions of the striatum with ^{11}C-raclopride, and further advances in instrumentation may provide even higher resolution.

Demonstration That Imaging Can Monitor Monoamine Release, Though with Relatively Low Sensitivity

In spite of methodological limitations and limited sample sizes, several reports indicate that imaging can detect DA release in response to cognitive challenges such as yoga, psychosocial stress tasks, and video games, as well as motor challenges such as finger tapping and foot movement.

Nonetheless, such stimuli, like psychostimulants, lead to decreases in 11C-raclopride binding that are on the order of 10%, rather than the 50–1, 500% increases in extracellular DA observed with microdialysis. Recent studies that combined PET with microdialysis highlighted the discord between microdialysis and PET as applied to measurement of DA release. For example, one such study showed that a dose of amphetamine that elicited the same degree of inhibition of 11C-raclopride binding as a given dose of methylphenidate resulted in a fourfold greater increase in DA release as measured by microdialysis. To reconcile such differences, it has been hypothesized that the DA release measured with PET radiotracers using endogenous competition techniques may be reflective of fluctuations in intrasynaptic DA, whereas the DA release measured using microdialysis techniques may be reflective of extrasynaptic DA concentrations.

Imaging Serotonergic Transmission

Defects in other monoamine systems, particularly serotonin (5-HT), have been implicated in various psychiatric disorders, including major depression, anxiety disorders, and schizophrenia. Efforts to measure endogenous release of 5-HT with PET have been modeled after similar experiments designed to measure DA release *in vivo*, whereby a pharmacologic stimulus is used to enhance or deplete extracellular 5-HT and thereby elicit a reproducible change in radioligand binding that correlates with the degree of endogenous neurotransmitter release.

It has been demonstrated, for example, that acute administration of the psychostimulant fenfluramine resulted in a two- to 15-fold increase in extracellular 5-HT concentrations in the rat hippocampus and that this endogenous release resulted in dose-dependent decreases in the binding of the 5-HT_{1A} radioligand 18F-MPPF. Other studies conducted in rodents have reported little or no effect of fenfluramine administration on the binding of another 5-HT_{1A} PET radioligand, ^{11}C-WAY-100635 or the 5-HT_{2A} radioligand 11C-MDL-100907. In nonhuman primate brain, acute administration of fenfluramine failed to elicit significant decreases in the binding of 18F-MPPF. Although one study reported modest decreases in the binding of the 5-HT_{2A} radioligand 18F-altanserin in whole human cortex (14–23%) after acute administration of the selective serotonin reuptake inhibitor clomipramine (25 mg), it is unclear whether this effect can be attributed to endogenous competition by 5-HT or to direct competition with clomipramine itself, which has been shown to have a 50–60 $\mathrm{nmol\,l^{-1}}$ affinity for the 5-HT_{2A} receptor.

Unfortunately, for reasons that are unclear, attempts to measure endogenous competition by serotonin using PET have been unsuccessful. This may have to do with the fraction of receptors configured *in vivo* in the high-affinity state for the agonist, cellular localization (intracellular versus extracellular), and their distance from the synaptic cleft.

Conclusions

Using such methods as immunohistochemistry, electrophysiology, and genomics, it is now possible to examine the characteristics of single cells in the brain, either *in vivo* or at least immediately after brain removal. Although the direct monitoring of transmitter release from a single neuron has now been achieved in culture, it is not yet possible within an intact brain. Nonetheless, the judicious application of voltammetry, capillary electrophoresis, microdialysis, and *in vivo* imaging is allowing us to draw important conclusions about the release of monoaminergic transmitters from groups of neurons, and important observations made with these approaches continue to provide enlightenment about both normal and abnormal brain function. Moreover, the sensitivity and specificity of these methods have improved significantly over the past decade, and these improvements should continue until biochemistry takes its place

alongside histology, electrophysiology, and molecular biology as a route to the monitoring of a single cell within the living brain.

See also: Cells: 5-Hydroxytryptamine Receptors; Dopamine; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; 3,4-Methylenedioxymethamphetamine (MDMA, "Ecstasy"); Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Monoamines; Norepinephrine: CNS Pathways and Neurophysiology; Octopamine and Other Monoamines in Invertebrates; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors; Single Photon Emission Computed Tomography (SPECT); Trace Monoamines and Receptors in Mammalian CNS.

Further Reading

Borland LM and Michael AC (2004) Voltammetric study of the control of striatal dopamine release by glutamate. *Journal of Neurochemistry* 91: 220–229.

Carson RE (2003) Tracer kinetic modeling in PET. In: Valk PE, Dale L, Bailey DL, et al. (eds.) *Positron Emission Tomography: Basic Science and Clinical Practice*, pp. 147–180. London: Springer.

Daws LC, Montanez S, Munn JL, et al. (2006) Ethanol inhibits clearance of brain serotonin by a serotonin transporter independent mechanism. *Journal of Neuroscience* 26: 6431–6438.

Fillenz M (2005) *In vivo* neurochemical monitoring and the study of behaviour. *Neuroscience and Biobehavioral Reviews* 29: 949–962.

Finlay JF and Zigmond MJ (1995) A critical analysis of neurochemical methods for monitoring transmitter dynamics in brain. In: Bloom FE and Kupfer DJ (eds.) *Psychopharmacology: The Fourth Generation of Progress* pp. 29–40. New York: Raven Press.

Heien MLAV, Khan AS, Ariansen JL, et al. (2005) Real-time measurement of dopamine fluctuations after cocaine in the brain of behaving rats. *Proceedings of the National Academy of Sciences of the United States of America* 102: 10023–10028.

Kissinger PT (1996) Electrochemical detection in bioanalysis. *Journal of Pharmaceutical and Biomedical Analysis* 14: 871–880.

Laruelle M (2000) Imaging synaptic neurotransmission with *in vivo* binding competition techniques: A critical review. *Journal of Cerebral Blood Flow and Metabolism* 20: 423–451.

Laruelle M, Abi-Dargham A, van Dyck CH, et al. (1995) SPECT imaging of striatal dopamine release after amphetamine challenge. *Journal of Nuclear Medicine* 36: 1182–1190.

Michael AC and Borland LM (eds.) (2007) *Electrochemical Methods for Neuroscience*. Boca Raton, FL: CRC Press.

Narendran R, Hwang DR, Slifstein M, et al. (2004) *In vivo* vulnerability to competition by endogenous dopamine: Comparison of the D2 receptor agonist radiotracer (−)-*N*-[11C]propylnorapomorphine ([11C]NPA) with the D2 receptor antagonist radiotracer [11C]-raclopride. *Synapse* 52: 188–208.

Peter JL, Miner LH, Michael AC, and Sesach SR (2004) Ultrastructure at carbon fiber microelectrode implantation sites after acute voltammetric measurement in the striatum of anesthetized rats. *Journal of Neuroscience Methods* 137: 9–23.

Robinson DL, Venton BJ, Heien ML, and Wightman RM (2003) Detecting subsecond dopamine release with fast-scan cyclic voltammetry *in vivo*. *Clinical Chemistry* 49: 1763–1773.

Rougé-Pont F, Usiello A, Benoit-Marand M, Gonon F, Piazza PV, and Borrelli E (2002) Changes in extracellular dopamine induced by morphine and cocaine: Crucial control by D2 receptors. *Journal of Neuroscience* 22: 3293–3301.

Sabeti J, Gerhardt GA, and Zahniser NR (2003) Individual differences in cocaine-induced locomotor sensitization in low and high cocaine locomotor-responding rats are associated with differential inhibition of dopamine clearance in nucleus accumbens. *Journal of Pharmacology and Experimental Therapeutics* 305: 180–190.

Smith AD, Olson RJ, and Justice JB Jr. (1992) Quantitative microdialysis of dopamine in the striatum: Effect of circadian variation. *Journal of Neuroscience Methods* 44: 33–41.

Westerink BH (2000) Analysis of biogenic amines in microdialysates of the brain. *Journal of Chromatography. B, Biomedical Sciences and Applications* 747: 21–32.

Wightman RM (2006) Probing cellular chemistry in biological systems with microelectrodes. *Science* 311: 1570–1574.

Wu Q, Reith MEA, Walker QD, Kuhn CM, Carroll FI, and Garris PA (2002) Concurrent autoreceptor-mediated control of dopamine release and uptake during neurotransmission: An *in vivo* voltammetric study. *Journal of Neuroscience* 22: 6272–6281.

Zhang MY and Beyer CE (2006) Measurement of neurotransmitters from extracellular fluid in brain by *in vivo* microdialysis and chromatography-mass spectrometry. *Journal of Pharmaceutical and Biomedical Analysis* 40: 492–499.

Monocarboxylate Transporters

K Pierre and L Pellerin, Université de Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Glucose has long been considered the only valuable energy source for brain cells. In the past few years, however, monocarboxylates such as lactate, pyruvate, and ketone bodies have attracted much attention and appear to play an important role in brain energy metabolism. Thus, it was demonstrated that monocarboxylates can be taken up by the brain from the systemic circulation in a number of circumstances. In the first few hours after birth, lactate present in the blood at high concentration provides a critical energy source for the brain during this early life period. Then, during the entire lactating period, ketone bodies become substantial brain energy substrates. Interestingly, monocarboxylates can also represent significant energy substrates for the adult central nervous system under specific conditions. It was shown, for example, that a rise in circulating lactate as observed following moderate exercise is sufficient to reduce brain glucose utilization and thus significantly contribute to satisfying brain energy needs. Similarly, in situations leading to increased production of ketone bodies (e.g., diabetes or prolonged starvation), these monocarboxylates are taken up and used by the adult brain. In addition to blood-borne monocarboxylates, some parenchymal cells also produce and release large amounts of lactate (from blood-borne glucose) within the adult brain that can be consumed in parallel by others. Because monocarboxylates are hydrophilic substances and do not cross membranes easily, specific transport systems must exist to carry them across the blood–brain barrier and within each cell type or transfer them from one cell type to the other when monocarboxylates are produced locally. Several monocarboxylate transporter subtypes have been identified and rapid progress has been made on their distribution in the brain. However, regulation of their expression has just started to be explored.

The monocarboxylate transportern (MCT) family (also termed the SLC16 gene family) is composed of 14 members based on sequence homologies, identified as MCT1–9, MCT11–14, and TAT1. MCT1–MCT4 were shown to transport several monocarboxylates that can be produced endogenously, such as lactate, pyruvate, and ketone bodies, as well as other monocarboxylates from exogenous sources, such as acetate, propionate, and butyrate. The mechanism of monocarboxylate transport is a cotransport with H^+ following a 1:1 stoichiometry. In the brain, three MCTs have been identified: MCT1, MCT2, and MCT4. Their affinity for each substrate differs substantially. For example, the K_m for L-lactate is 3.5, 0.7, and 34 mM for MCT1, MCT2, and MCT4, respectively. These kinetic differences could partly determine their distribution, with high-affinity transporters present in cells exhibiting monocarboxylate utilization and low-affinity, high-capacity transporters being preferentially expressed by cells producing large amounts of monocarboxylates.

Endothelial Cells Forming Blood Vessels Express MCT1, Whereas Astrocytes Have Both MCT1 and MCT4

MCT1 mRNA and protein are homogenously distributed throughout the whole rodent brain. In young and adult rodent brain, MCT1 protein (apparent molecular weight, ~45 kDa) is abundantly expressed in endothelial cells forming microvessels and ependymocytes lining the four brain ventricles. MCT1-positive glial end feet surrounding capillaries and glial-like processes are also visible using light and electron microscopy. Confocal microscopy using double labeling with glial markers such as GFAP or S100β confirmed association of MCT1 immunoreactivity with astrocytic processes, both *in vitro* and *in vivo* (Figure 1). Similarly, immunohistochemical investigations in the human brain showed that MCT1 is associated with blood vessels as well as expressed by astrocytes. In contrast, MCT4, a protein of ~43 kDa, is specifically expressed by glial cells in the adult rodent brain. MCT4 labeling in the cortex, the hippocampus, and the corpus callosum is restricted to astrocytes. In the cerebellum, both Bergmann glia in the molecular layer and astrocytes in the granular layer are labeled. The expression of MCT1 and MCT4 by astrocytes is consistent with their kinetic properties that confer them a prominent role in glycolytic tissues (e.g., fast-twitch muscle). Except in glia limitans, MCT2 protein expression has not been detected in glial-like cells of the mouse brain, whereas it may be visible in some endothelial cells forming brain capillaries. By contrast, MCT2 has been shown to be expressed by astrocytes in the rat brain using different antibody sources. This issue needs to be clarified.

Neurons Express Predominantly MCT2

It can be considered that MCT2 represents the major neuronal transporter in the rodent brain. MCT2 protein (molecular weight, ~40–43 kDa) is strongly

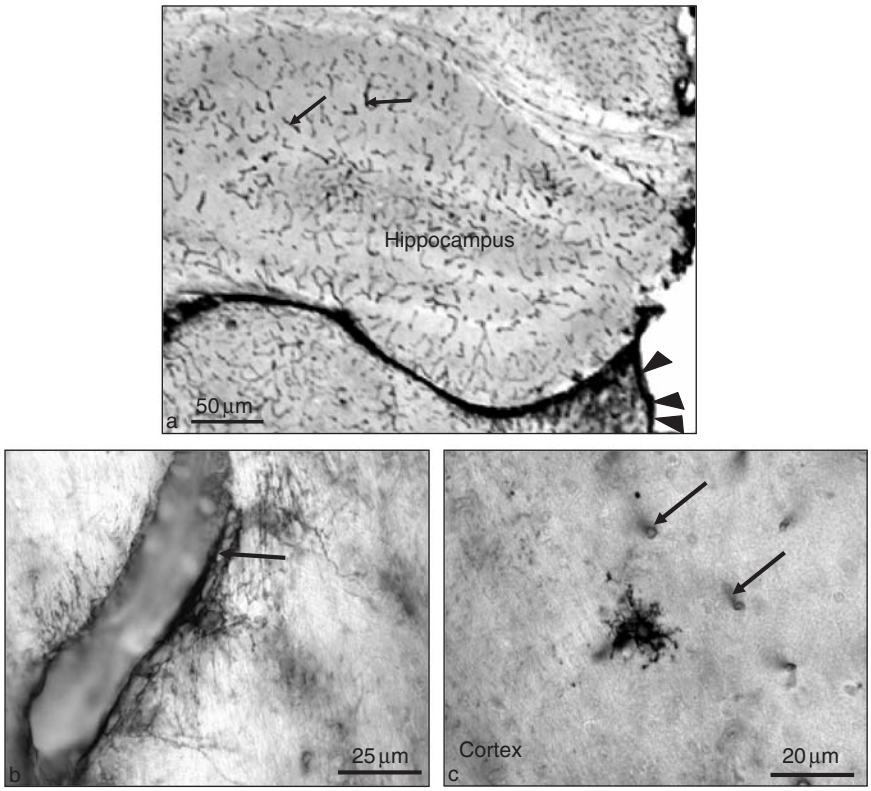

Figure 1 MCT1 immunoreactivity in the adult rat brain. MCT1 protein is strongly expressed on blood microvessels (a–c, arrows) and in astrocytes either surrounding capillaries (b) or isolated in the parenchyma (c). It is also visible in ependymocytes lining the ventricles (a, arrowheads). Immunoperoxidase stainings viewed with light microscopy. Reproduced from Pellerin L and Pierre K (2005) Monocarboxylate transporters and their dual role in efflux and influx across the blood–brain barrier and beyond. In: Taylor EM (ed.) *Efflux Transporters and the Blood–Brain Barrier*, pp. 193–217. New-York: Nova Science, with permission.

expressed in the cortex, the hippocampus, and the cerebellum of the rodent brain. Cortical neurons in culture exhibit MCT2 immunoreactivity at the surface and in the cytoplasm of their cell bodies and neurites (Figure 2). *In vivo*, MCT2 immunoreactivity is visible as a strong, punctiform labeling surrounding immunonegative neuronal somata. Double labeling experiments using the neuronal marker MAP2 showed a co-localization between MCT2 and MAP2-positive processes, both *in vitro* and *in vivo*. Moreover, intense MCT2 immunoreactivity was found in cerebellar Purkinje cell bodies and their processes as well as on mossy fibers in the cerebellum. Expression of the high-affinity transporter MCT2 by neurons is consistent with their important uptake and consumption of lactate as oxidative substrate.

Evidence suggests that part of MCT2 expression is associated with synapses. First, observation in developing neurons in cultures following double immunofluorescent labeling with the presynaptic marker synaptophysin and MCT2 revealed a parallel distribution. In the mouse brain cortex, a co-localization with PSD95, the major component of the postsynaptic density (PSD), was seen occasionally. Moreover, MCT2 immunoreactivity is present in the postsynaptic membrane of parallel fiber–Purkinje cell synapses in the rat cerebellum and co-localizes with postsynaptic δ2-glutamate receptors as revealed by electron microscopy. MCT2 has been shown to be located at the PSD of distinct glutamatergic synapses in the hippocampus and the cerebellum of the rat and mouse brain, respectively.

In addition to its main expression by astrocytes and endothelial cells, a faint MCT1 immunoreactivity was found in cultured mouse cortical neurons, although it remained undetectable using Western blot analysis on the same culture extracts. Cultured hypothalamic neurons were found to be heavily labeled for MCT1. Whether this expression is caused by cultured conditions rather than being a genuine characteristic of specific neuronal populations remains to be determined. *In vivo*, MCT1 immunoreactivity was undetectable in neurons throughout the mouse brain, whereas it was visible in a few scattered neurons in the rat brain.

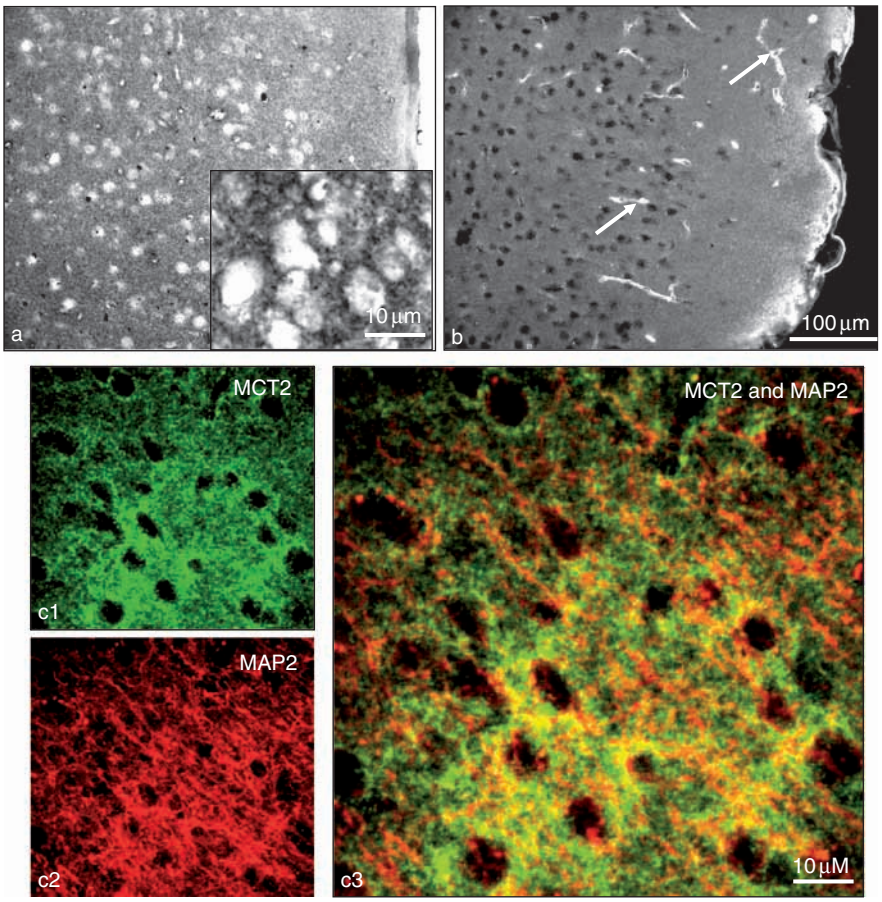

Figure 2 MCT2 immunoreactivity in the adult mouse brain. A strong punctiform immunoreactivity for MCT2 is found in the neuropil surrounding immunonegative neuronal cell bodies throughout the entire brain (a, b, and c1). Double labeling using antibodies raised against MCT2 (c1, green) and the neuronal marker MAP2 (c2, red) revealed co-localization between MCT2 and MAP2 (c3, yellow) in neuronal dendrites. In addition, MCT2 immunoreactivity is detected in microvessel-like structures (b, arrows). (a) Immunoperoxidase staining viewed with light microscopy. (b and c1–c3) Immunofluorescence stainings viewed with light (b) and confocal (c1–c3) microscopy using appropriate filters. Reproduced from Pierre K, Magistretti PJ, and Pellerin L (2002) MCT2 is a major neuronal monocarboxylate transporter in the adult mouse brain. *Journal of Cerebral Blood Flow Metabolism* 22: 586–595.

Otherwise, MCT4 has never been localized in neurons, either *in vivo* or *in vitro*.

Regulation of MCT Expression in Brain

During early postnatal development, monocarboxylate uptake increases in parallel with a similar enhancement of its utilization by the brain. Thus, it is expected that the expression of MCTs both on endothelial cells forming blood vessels and on parenchymal cells is regulated in the developing brain because the availability of these major energy substrates varies considerably during peri- and postnatal development. Indeed, evidence suggests that MCT expression changes importantly during this period. For example, it was shown that mRNAs for both MCT1 and MCT2 increased during the early postnatal period to reach a maximum at approximately postnatal day 15 and decreased rapidly after weaning (approximately postnatal day 21) to stabilize at adult levels. Moreover, data indicate that alteration of MCT expression following nutritional modifications occurs in the adult brain. In the adult rodent, a ketogenic diet was found to induce brain MCT1 as well as MCT2 and MCT4 expression. However, the precise signals and mechanisms involved in this regulation are unclear.

Other types of regulation have begun to emerge. In cultured neurons, the level of MCT2 expression has been correlated with synaptogenesis, suggesting that its expression may be regulated by factors in relation with synaptic activity. MCT2 expression in neurons appears to be under the control of specific neuroactive substances. Thus, it has been shown to be enhanced by noradrenaline in cultured cortical neurons, an effect reproduced by the activation of

the cAMP-mediated cascade. Moreover, it has been demonstrated that this type of regulation of MCT2 expression occurs at the translational level.

Roles of MCTs in Brain

The expression of MCTs on endothelial cells forming blood vessels as well as on parenchymal cells early during development fulfills an obvious role in providing an adequate supply of energy substrates in this critical developmental period, during which glucose metabolism is not yet the predominant pathway to fuel brain function. The presence in the adult brain of distinct MCT isoforms on different cell types is more intriguing. An interesting observation is the fact that not only are MCT1 and the glucose transporter GLUT1 co-expressed on endothelial cells of adult brain microvessels but also their respective densities are correlated with several metabolic parameters within specific cortical areas (e.g., cytochrome oxidase staining). These data suggest that monocarboxylates provided by the circulation could represent a significant energy source for the brain in specific circumstances and their utilization could be correlated with the degree of functional activity within each cortical area. Indeed, it has been clearly shown that lactate produced during exercise that accumulates in the circulation is consumed by the brain. Moreover, raising plasma lactate levels directly leads to reduced brain glucose utilization. Lactate has even been found to support cognitive functions during hypoglycemic episodes, further supporting a putative role of lactate as an adequate brain energy substrate even for the adult brain. In this context, MCTs would be essential to facilitate delivery of lactate to the brain as well as to each cell type that would require/prefer it as a fuel.

The expression of the high-affinity transporter MCT2 by neurons is consistent with the notion that these highly oxidative cells have a preferential use of lactate as oxidative substrate. By contrast, the expression of MCT1 and, specifically, MCT4 by astrocytes (**Figure 3**) is in agreement with the high glycolytic capacity of these cells. It appears likely that under conditions in which glucose is the main substrate delivered to the brain, part of it will be converted to lactate by astrocytes and released in the extracellular space via MCT1 and MCT4. Because the extracellular lactate concentration in rodent and human brain is ~1 or 2 mM, neurons probably consume in parallel with glucose a certain amount of lactate via MCT2. It is also suggested that net lactate production by astrocytes and lactate consumption by neurons could be enhanced following brain activation within specific areas. Again, the presence of specific MCTs

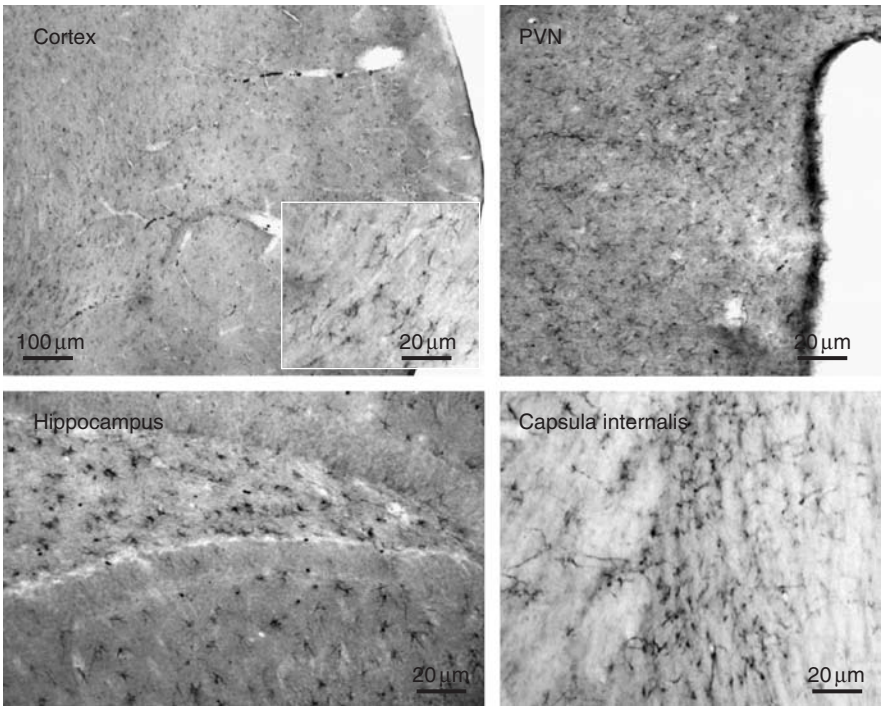

Figure 3 Expression of MCT4 by astrocytes in various rat brain regions. Immunolabelings carried out with an antibody against MCT4 show immunopositive astrocytes in the cortex (top left), hippocampus (bottom left), paraventricular nucleus (PVN; top right), and capsula internalis (bottom right). Immunoperoxidase immunoreactivity viewed by light microscopy. Reproduced from Pellerin L, Bergersen LH, Halestrap AP, and Pierre K (2005) Cellular and subcellular distribution of monocarboxylate transporters in cultured brain cells and in the adult brain. *Journal of Neuroscience Research* 79: 55–64, with permission.

on both astrocytes and neurons would be essential to permit such interchange between the two cell types. The presence of MCT2 within the postsynaptic density at glutamatergic synapses further supports a role of MCT2 and neuroenergetics in synaptic transmission, although much more research is needed in this area.

Considering the previous discussion, it is tempting to speculate that both short-term and long-term mechanisms must exist to adapt the number of MCTs at the membrane of each cellular and subcellular compartment (endothelial cell, astrocyte, neuron, and synapse) in order to handle the changing energy requirements in different situations. It will be a challenge in the future to identify such mechanisms and determine their importance in various physiological and pathophysiological conditions.

See also: Astrocyte: Neurotransmitter and Hormone Receptors; Glial Energy Metabolism: Overview; Glial Glutamate Transporters: Electrophysiology; Glial Glutamate Transporters; Glial Energy Metabolism: A NMR Spectroscopy Perspective; Neurotransmitter Release from Astrocytes.

Further Reading

Bergersen LH, Magistretti PJ, and Pellerin L (2005) Selective postsynaptic co-localization of MCT2 with AMPA receptor GluR2/3 subunits at excitatory synapses exhibiting AMPA receptor trafficking. *Cerebral Cortex* 15: 361–370.

Chiry O, Pellerin L, Monnet-Tschudi F, et al. (2006) Expression of the monocarboxylate transporter MCT1 in the adult human brain cortex. *Brain Research* 1070: 65–70.

Debernardi R, Pierre K, Lengacher S, Magistretti PJ, and Pellerin L (2003) Cell-specific expression pattern of monocarboxylate transporters in astrocytes and neurons observed in different mouse brain cortical cell cultures. *Journal of Neuroscience Research* 73: 141–155.

Gerhart DZ, Enerson BE, Zhdankina OY, Leino RL, and Drewes LR (1998) Expression of the monocarboxylate transporter MCT2 by rat brain glia. *Glia* 22: 272–281.

Halestrap AP and Meredith D (2004) The SLC16 gene family – From monocarboxylate transporters (MCTs) to aromatic amino acid transporters and beyond. *Pflugers Archives* 447: 619–628.

Leino RL, Gerhart DZ, Duelli R, Enerson BE, and Drewes LR (2001) Diet-induced ketosis increases monocarboxylate transporter (MCT1) levels in rat brain. *Neurochemistry International* 38: 519–527.

Maurer MH, Canis M, Kuschinsky W, and Duelli R (2004) Correlation between local monocarboxylate transporter 1 (MCT1) and glucose transporter 1 (GLUT1) densities in the adult rat brain. *Neuroscience Letters* 355: 105–108.

Pellerin L (2003) Lactate as a pivotal element in neuron–glia metabolic cooperation. *Neurochemistry International* 43: 331–338.

Pellerin L, Bergersen LH, Halestrap AP, and Pierre K (2005) Cellular and subcellular distribution of monocarboxylate transporters in cultured brain cells and in the adult brain. *Journal of Neuroscience Research* 79: 55–64.

Pellerin L and Pierre K (2005) Monocarboxylate transporters and their dual role in efflux and influx across the blood–brain barrier and beyond. In: Taylor EM (ed.) *Efflux Transporters and the Blood–Brain Barrier*, pp. 193–217. New-York: Nova Science.

Pierre K, Debernardi R, Magistretti PJ, and Pellerin L (2003) Noradrenaline enhances monocarboxylate transporter 2 expression in cultured mouse cortical neurons via a translational regulation. *Journal of Neurochemistry* 86: 1468–1476.

Pierre K, Magistretti PJ, and Pellerin L (2002) MCT2 is a major neuronal monocarboxylate transporter in the adult mouse brain. *Journal of Cerebral Blood Flow Metabolism* 22: 586–595.

Pierre K, Parent A, Jayet P-Y, Halestrap AP, Scherrer U, and Pellerin L (2007) Enhanced expression of three monocarboxylate transporter isoforms in the brain of obese mice. *Journal of Physiology* 583: 469–486.

Pierre K and Pellerin L (2005) Monocarboxylate transporters in the central nervous system: Distribution, regulation and function. *Journal of Neurochemistry* 94: 1–14.

Rafiki A, Boulland JL, Halestrap AP, Ottersen OP, and Bergersen L (2003) Highly differential expression of the monocarboxylate transporters MCT2 and MCT4 in the developing rat brain. *Neuroscience* 122: 677–688.

Mood Stabilizers

H K Manji, C A Zarate, G Chen, and J Du, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Bipolar disorder is a severe recurrent medical illness that afflicts approximately 1–3% of the population. Patients with bipolar disorder often present with high rates of comorbidities, relapse, recurrence, residual mood symptoms, and poor functioning and quality of life in spite of adequate treatment with current mood stabilizers. While preclinical and clinical knowledge has expanded at a tremendous rate in recent years, there have been no major breakthroughs in novel types of treatment for bipolar disorder. Developing new treatments may come from a better understanding of the mechanism of action of current mood stabilizers (e.g., lithium, valproate, and carbamazepine) and thereby designing novel drugs that mimics these mechanism(s).

Molecular and cellular targets of current mood stabilizers include lithium inhibitable enzymes (inositol monophosphatase, inositol polyphosphate 1-phosphatase, glycogen synthase kinase-3 (GSK-3), fructose 1,6-bisphosphatase, bisphosphate nucleotidase, phosphoglucomutase), valproate inhibitable enzymes (succinate semialdehyde dehydrogenase, succinate semialdehyde reductase, and histone deacetylase), and targets of carbamazepine (sodium channels, adenosine receptors, and adenylate cyclase (AC)). Tables 1–3 summarize the direct targets of lithium, valproic acid, and carbamazepine. Of these direct targets, only GSK-3 is reviewed, as brain-penetrant inhibitors for this target are likely to be soon available.

A major problem inherent in neuropharmacologic research is the difficulty in attributing therapeutic relevance to any observed biochemical finding. One powerful approach is to identify common biochemical targets, which are modified by drugs belonging to the same therapeutic class (e.g., mood-stabilizing agents) but possessing distinct chemical structures when administered in a 'therapeutically relevant' paradigm. In this context, it is noteworthy that both valproate and lithium, with different chemical structures, belong to the same therapeutic class and share common effects on downstream-signaling pathways. This approach has led to the identification of signaling pathways that are regulated by mood stabilizers of different classes (phosphoinositol/protein kinase C (PKC), cyclic AMP, arachidonic acid (AA), neurotrophic pathways); these downstream targets are the focus of this article.

Glycogen Synthase Kinase-3

GSK-3 may represent a therapeutically relevant target for lithium. GSK-3 is a serine/threonine kinase that is normally highly active in cells, and is deactivated by signals originating from numerous signaling pathways (e.g., the Wnt pathway, phosphoinositide 3-(PI3) kinase pathway, protein kinase A (PKA), PKC, among many others). Cellular targets of GSK-3 are numerous and often depend upon the signaling pathway that is acting upon it. For example, Wnt pathway inhibition of GSK-3 activates the transcription factor β-catenin, while in the insulin/PI3 kinase signaling pathway, inhibition of GSK-3 results in activation of the enzyme glycogen synthase. Targets of GSK-3 include, among others, transcription factors (β-catenin, cyclic AMP responsive element binding protein (CREB), c-Jun), proteins bound to microtubules (Tau, MAP1B), cell cycle mediators, and regulators of metabolism (glycogen synthase, pyruvate dehydrogenase).

GSK-3 is receiving foremost interest as a regulator of apoptosis and cellular resilience (Figure 1). In general, increased activity of GSK-3 is pro-apoptotic, while inhibiting GSK-3 attenuates or prevents apoptosis. At this point, it is critical to note that evidence suggests an association between mood disorders and impairments of neuroplasticity and cellular resilience, with both *in vivo* and postmortem studies suggesting neuron and/or glial cell loss or atrophy in circumscribed brain areas. Importantly, lithium has effects suggestive of neuroprotection clinically (see below), as well as in rodent and cell-based models. Lithium may exert these neuroprotective effects, at least in part, by inhibition of GSK-3.

Many of the behavioral effects of lithium may be due to inhibition of GSK-3. Pharmacological inhibition of GSK-3 attenuates d-amphetamine hyperlocomotion in rats. Further, mice overexpressing a constitutively active form of GSK-3β in the brain showed increased locomotor activity, as well as decreased habituation in an open field. Thus, underexpression of GSK-3β results in attenuation of stimulant-induced locomotion, while overexpression results in an endogenously high level of locomotion.

In addition to the effects on manic-like behaviors, recent studies have investigated the possibility that lithium may also exert antidepressant-like effects through GSK-3. Thus, O'Brien and colleagues reported that 10 days of lithium chloride administration resulted in an antidepressant-like effect (decreased immobility time) in the forced-swim test. We recently found that subacute intraperitoneal injections of ARA014418,

Table 1 Direct targets of lithium

Inositol monophosphatase (IMPase)	Rate-limiting enzyme in inositol recycling; lithium's inhibition of IMPase led to the inositol depletion hypothesis
Inositol polyphosphate 1-phosphatase (IPPase)	Enzyme involved in inositol recycling in phosphoinositol signaling; acts prior to IMPase
Bisphosphate nucleotidase (BPNase)	Removes phosphate from 3′-phosphoadenosine 5′-phosphate (PAP) to form adenosine 5′-phosphate (AMP); an increase in PAP inhibits sulfotransferases, which transfer sulfur to biological molecules
Fructose 1,6-bisphosphastase (FBPase)	Key enzyme in glyconeogenesis; catalyzes the removal of the 1-phosphate from fructose 1,6-bisphosphatase to form fructose 6-phosphate
Phosphoglucomutase (PGM)	Enzyme in glycogenolysis and glycogenesis; catalyzes the formation of glucose 1-phosphate from glucose 6-phosphate during glycogenolysis (and the reverse during glycogenesis)
Glycogen synthase kinase-3 (GSK-3)	Normally active kinase that is inhibited by the activity of many signaling pathways; inhibiting GSK-3 has been linked to neurotrophic support, neuroprotection, and possible modulation of circadian rhythms

Table 2 Direct targets of valproic acid

Succinate semialdehyde dehydrogenase (SSA-DH)	Converts succinate semialdehyde to succinate; succinate inhibits GABA-T; inhibiting GABA-T increases GABA levels
GABA transaminase (GABA-T)	Converts GABA to succinate semialdehyde; however, valproate inhibition appears to be above the therapeutic range
Succinate semialdehyde reductase (SSA-R)	Converts succinate semialdehyde to gamma-hydroxybutyrate (GHB)
Histone deacetylase (HDAC)	Enzyme that acetylates histones. Inhibition of this enzyme results in an increase in nonaceytlated histones and a resultant increase in transcription of some genes
Sodium channels	Inhibits sodium channel function at times of high-frequency firing

Table 3 Direct targets of carbamazepine

Sodium channels	Inhibits sodium channel function at times of high-frequency firing
Adenosine receptor	Antagonist at A1 subtype; increases adenosine receptor protein levels in rats. Adenosine has modulatory functions on neurotransmitter release and numerous behavioral and cognitive functions
Adenylate cyclase (AC)	AC forms cyclic AMP from ATP, and as such is a principle regulator of the AC signaling pathway; carbamazepine attenuates cyclic AMP signaling and appears to have an inhibitory effect on AC or an AC-associated protein

a GSK-inhibitor, reduced immobility time (increased activity) in the forced-swim test, compared to vehicle-treated rats.

Downstream Targets of Shared by Current Mood Stabilizers

A number of pathways that several mood stabilizers regulate and appear to be relevant to its mechanism of action include the AC, phosphoinositol, neurotrophic, and AA signaling pathways.

Cyclic AMP-Mediated Signal Transduction

In addition to the direct effects of carbamazepine on AC, significant modulation of cyclic AMP-mediated signaling has been noted with lithium (especially) and also valproate. G-proteins modulate intracellular cyclic adenosine monophosphate (cAMP) levels by mediating the effect of neurotransmitters (via extracellular receptors) on AC, an integral membrane protein of which there exist numerous subtypes. AC catalyzes the conversion of adenosine triphosphate (ATP) to cyclic AMP. Stimulation of G-proteins $G\alpha_s$ and $G\alpha_{olf}$ increases AC activity, while stimulation of $G_{\alpha i}$ results in a decrease in AC activity. The physiologic effects of cyclic AMP appear to be mediated primarily by activation of PKA, an enzyme that phosphorylates and regulates many proteins including ion channels, cytoskeletal elements, transcription factors, and other enzymes. One major central nervous system (CNS) target for the actions of PKA is the transcription factor CREB, which plays a major role in long-term neuroplasticity, and is a downstream target of monoaminergic-based antidepressants. One of the genes activated by CREB is brain-derived neurotrophic factor (BDNF), a protein implicated in neuronal survival and synaptic plasticity.

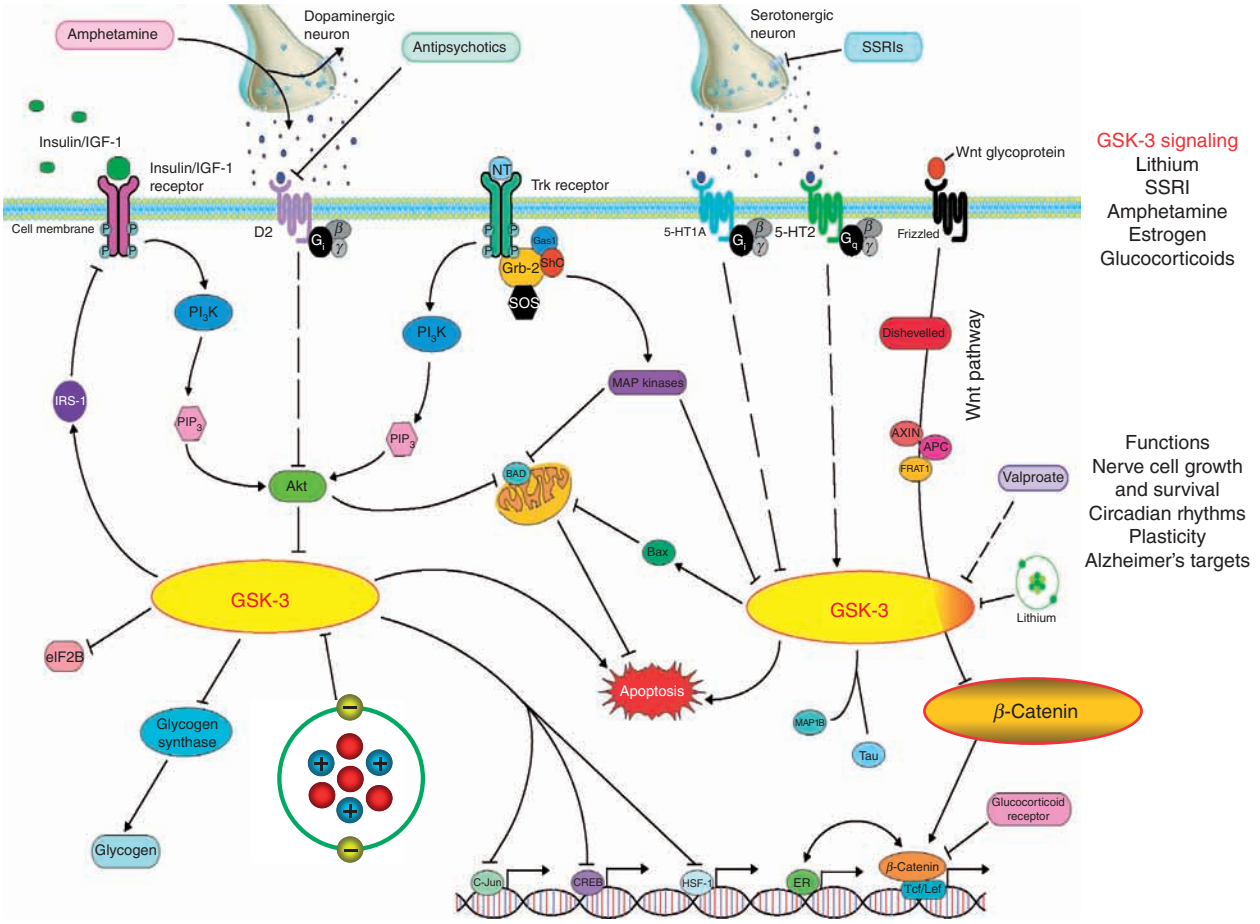

Figure 1 GSK-3 and intracellular signaling GSK-3 regulates diverse signaling pathways in the cell. These include insulin/insulin like growth factor (IGF-1) signaling, neurotrophic factor signaling, and Wnt signaling. Insulin signaling through its tyrosine receptor kinase (Trk) receptor activates PI3 kinase-mediated signaling, resulting in inhibition of GSK-3. Inhibition of GSK-3 activates glycogen synthase and eukaryotic initiation factor 2B (eIF2B) while inhibiting insulin receptor substrate-1 (IRS-1; an inhibitor of the insulin receptor). Insulin is generally thought to have minimal effects on CNS neurons; however, IGF-1 interacting with its cognate receptor appears to have similar functions. Neurotrophins (NT) act through tyrosine receptor kinase (Trk) receptors A, B, and C to activate PI3 kinase, AKT, and inhibit GSK-3. Many effectors have been implicated in GSK-3's neurotrophic effects including transcription factors (e.g., heat shock factor-1 (HSF-1), C-Jun, and cyclic AMP-responsive element binding protein (CREB)) and recently the proapoptotic Bcl-2 family member BAX. In the Wnt signaling pathway, secreted Wnt glycoproteins interact with the frizzled family of receptors and, through disheveled mediated signaling, inhibit GSK-3. Stability of this process requires the scaffolding proteins AXIN and adenomatous polyposis coli (APC). Normally active GSK-3 phosphorylates β-Catenin leading to its ubiquitin-dependent degradation. However, when GSK-3 is inhibited in the Wnt pathway, β-Catenin is not degraded, allowing for its interaction with (T cell-specific transcription factor) TCF to act as a transcription factor. β-catenin activity is modulated by the intracellular estrogen receptor (ER), which also affects transcription of an independent set of genes. As shown in the figure, and described in the text, medications useful for the treatment of mood disorders have both direct and indirect effects on GSK-3, and GSK-3 regulated cell signaling pathways. This includes direct effects of lithium and indirect effects of antipsychotics, amphetamine, and selective serotonin re-uptake inhibitors (SSRI's). These distinct pathways have convergent affects on cellular processes such as bioenergetics (energy metabolism), neuroplasticity, neurogenesis, resilience, and survival. Thus, our hypothesis is that lithium (and other medications) may act by enhancing these processes through inhibition of GSK-3. G_i refers to G_i/G_o; G_q refers to G_q/G_{11}. Modified from Gould TD, Zarate CAJ, and Manji HK (2004) Glycogen synthase kinase-3: A target for novel bipolar disorder treatments. *Journal of Clinical Psychiatry* 65: 10–21.

There is a growing body of data suggesting that agents, which directly modulate the cAMP/PKA/CREB/BDNF cascade, may have utility for the treatment of depression. In addition to antidepressant effects on cyclic AMP-mediated signaling, mood stabilizers also appear to regulate this pathway. Both lithium and valproate increase BDNF levels in the brains of rats treated chronically with these drugs and have inhibitory effects on stimulated AC-mediated signaling. Thus, it is helpful to take into account the multiple interactions between signaling pathways (e.g., CREB activity and BDNF expression are regulated by multiple signaling pathways including neurotrophic signaling pathways, as discussed later in this review), and that cyclic AMP signaling pathway does much more than simply regulate CREB activity.

In mouse cerebral cortex and cerebellar tissue, carbamazepine decreases the basal concentration of cyclic AMP. It also lowers cyclic AMP following stimulation by norepinephrine, adenosine, and the epileptogenic compounds ouabain and veratridine. In addition, carbamazepine appears to attenuate β-adrenoceptor and muscarinic cholinergic coupling to G-proteins in the rat cortex, decreases the levels of G_s and G_i, and attenuates cAMP-mediated phosphorylation of CREB in C6 glioma cells. Similarly, in pheochromocytoma (PC12) cells, carbamazepine inhibits cAMP-mediated increases in c-fos gene expression.

The literature describing the effect of lithium on G-proteins suggests that lithium both increases basal activity and inhibits stimulated AC.

There is some evidence that at therapeutically relevant concentrations in a cell line model, valproate decreased the density of βARs and attenuated both receptor- and post-receptor-stimulated cAMP production and levels of $G\alpha_s$ 45, but none of the other G-protein subunits examined were decreased. Thus, valproate may exert an effect on cAMP signaling at multiple levels.

Phosphoinositol-Mediated Signaling

Inositol phospholipids play a major role in receptor-mediated signal transduction pathways, involved in a diverse range of responses such as cell division, secretion, neuronal excitability, and responsiveness. The phosphatidylinositol (PI) pathway is initiated by the activation of G-protein-coupled receptors. Muscarinic (M1-2-3), noradrenergic (α1), and serotonergic (5HT2) receptors coupled to $G_{\alpha q/11}$ induce phospholipase C (PLC) hydrolysis of the membrane component phosphoinositide 4,5-biphosphate (PIP2). Hydrolysis of PIP2 by PLC results in the formation of the intracellular second messengers inositol-1,4,5-triphosphate (IP3) and 1,2-diacylglycerol (DAG; an endogenous activator of PKC). IP3 binds to the IP3 receptor facilitating the release of calcium from intracellular stores, in particular the endoplasmic reticulum. Among other proteins, the Ca^{2+}-receptor protein calmodulin (CaM) is activated, stimulating calmodulin-dependent protein kinases (CaM-Ks) that regulate the activity of diverse proteins, including ion channels, signaling molecules, proteins which regulate apoptosis, scaffolding proteins, and transcription factors. IPPase and IMPase (enzymes that are involved in recycling of IP3 back to PIP2) are directly inhibited by lithium (Table 1). Lithium's inhibition of these enzymes led to the inositol depletion hypothesis of lithium's action, which suggests that lithium, via inhibition of IMPase, decreases the availability of myo-inositol, and thus the amount of PIP2 available for G-protein-mediated signaling events that rely upon this pathway.

A number of studies have suggested the possibility that multiple mood stabilizers may regulate the PI signaling pathway. These include studies of the sodium/myo-inositol co-transport (SMIT), a high-affinity myo-inositol transport system that has been characterized in various cell types, including those of neural origin. The activity and expression of SMIT mRNA in cultured astrocytes is downregulated after chronic treatment with therapeutic concentrations of lithium. Downregulation of SMIT was also observed after valproate and carbamazepine treatment.

Another finding implicating phosphoinositol signaling in the actions of mood stabilizers comes from a study by Williams and colleagues. They used a novel tissue-culture assay that measures sensory neuron growth-cone stability and found that the depletion of neuronal inositol 1,4,5-trisphosphate (IP3) may be a common mechanism of action of lithium, valproate, and carbamazepine. In this study, they found that all three mood stabilizers inhibited the collapse of sensory neuron growth cones and increased growth cone area, effects which were reversed by inositol.

Both Lithium and Valproic Acid Regulate PKC Isozymes and MARCKS

PKC is a primary target of DAG, and as such has been an object of intense research with regard to the action of mood stabilizes on the PI pathway. It is an omnipresent enzyme, highly enriched in brain, where it plays a major role in regulating both pre- and postsynaptic aspects of neurotransmission. Recent studies have suggested that PKC activation may facilitate neurotransmitter release via a variety of mechanisms, including modulation of several ionic conductances regulating Ca^{2+} influx, upstream steps regulating release of Ca^{2+} from intracellular stores, recruitment of vesicles to at least two distinct vesicle pools, and the Ca^{2+} sensitivity of the release process itself. PKC has been demonstrated to be active in many other cellular processes including stimulation of transmembrane glucose transport, secretion, exocytosis, smooth muscle contraction, gene expression, modulation of ion conductance, cell proliferation, and desensitization of extracellular receptors.

PKC and PKC signaling appear to be a target of lithium and valproate. Chronic lithium treatment decreases the level of PKC isozymes α, and ε in cells and treated rodents. Further supporting the effect of lithium on PKC, lithium decreases the levels and phosphorylation of a major PKC substrate, myristoylated alanine rich C kinase substrate (MARCKS), following chronic treatment in rats.

Evidence also suggests that PKC is a target of valproate. In cultured cells, valproate reduces PKC activity in both membrane and cytoplasmic fractions. Valproate also selectively reduces the protein levels of the same PKC isozymes reduced by lithium, α, and ε, further suggesting possible importance in the treatment of bipolar disorder. The reduction in the α isozyme has also been confirmed in the brains of rats treated chronically with valproate at therapeutically relevant concentrations. Similar also to the effects of lithium, valproate decreases the levels of MARCKS. The mechanism by which valproate exerts these effects is unknown; however, they appear to be independent of myoinositol.

PKC Signaling in Animal Models of Mood Disorders

Current models of mania include kindling, behavioral/amphetamine sensitization, and glucocorticoid administration. Kindling is an animal model for epilepsy that has been proposed to have similarities with pathophysiological aspects of bipolar disorder, in which repeated administration of electrical stimulus (that are subthreshold to produce seizures *per se*) results in an epileptic focus and a permanent state of hyperexcitability to the stimulus. These studies on rats have consistently shown the upregulatory effect of hippocampal kindling on PKC activity and protein concentration, findings that were demonstrated to be valid in other brain structures such as the amygdala and neocortex as well.

Studies have also implicated alterations in PKC activity as mediators of long-term alterations in neuronal excitability in the brain following chronic stimulant use. Both acute and chronic amphetamine produce an alteration in PKC activity, its relative cytosol to membrane distribution, as well as the phosphorylation of a major PKC substrate, GAP-43 which has been implicated in long-term alterations of neurotransmitter release (Figure 2). Furthermore, PKC inhibitors have been shown to block the acute responses (as assessed by both behavioral and *in vivo* microdialysis studies) to both amphetamine and cocaine as well as cocaine-induced sensitization.

Abnormalities of circulating glucocorticoids are well known to be associated with affective symptomatology, and interestingly, elevated glucocorticoids have been associated with both depressive and manic symptomatology. Repeated administration of dexamethasone for 10 days results in a significant increase in B_{max} of [^{3}H]PDBu binding to PKC, augmented PKC activity, and raised the levels of PKCα and ε in the rat hippocampus. It is striking that behavioral sensitization and kindling (models which have been postulated to represent models of bipolar disorder and mania) as well as dexamethasone administration all produce alterations in the PKC signaling pathway in critical limbic structures, since

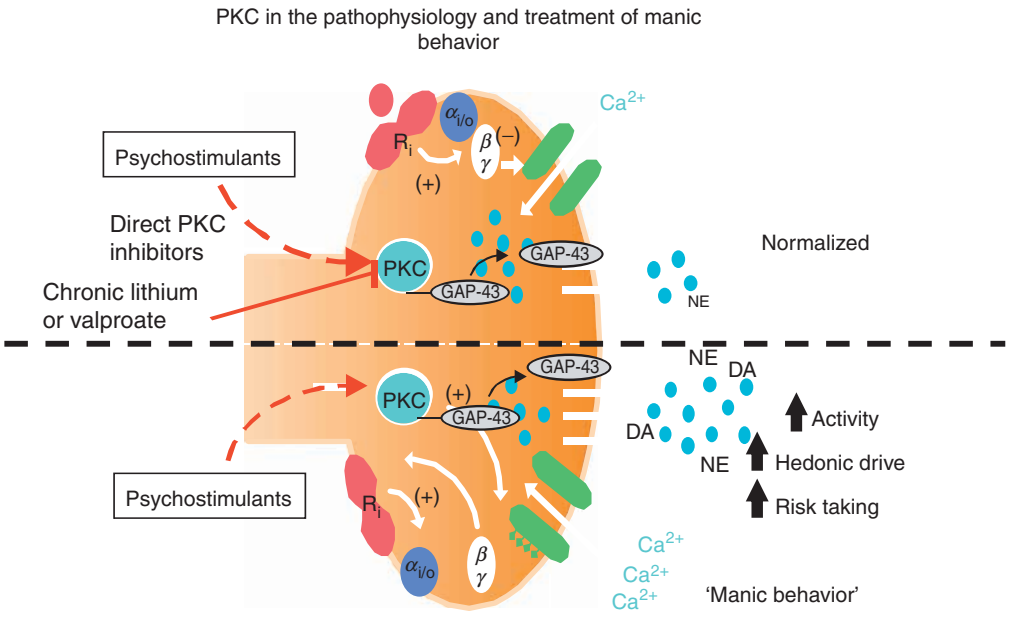

Figure 2 PKC in the treatment of mania. This figure depicts the potential mechanisms by which chronic lithium or valproate, or direct acting PKC inhibitors may have utility in the treatment of acute mania. Activation of PKC, which is known to occur by psychostimulants or stress, results in the phosphorylation of key substrates, notably GAP-43 (growth cone-associated protein) and MARCKS, thereby facilitating the release of neurotransmitters (e.g., dopamine (DA) and NE (norepinephrine)). Chronic lithium or valproate attenuates PKC signaling, effects which may be responsible for the treatment of various facets of the manic syndrome.

lithium and valproate also target the very same biochemical targets.

There is also evidence suggesting that chronic antidepressants may also modulate PKC activity in limbic and limbic-associated areas of rat brain, and PKC has been demonstrated to regulate the activity of norepinephrine, dopamine, and serotonin transporters. Whether these complex effects of antidepressants on PKC activity underlie their apparent ability to trigger manic episodes, and perhaps promote rapid cycling in susceptible individuals, remains to be determined.

PKC as potential therapeutic target In view of the pivotal role of the PKC signaling pathway in the regulation of neuronal excitability, neurotransmitter release, and long-term synaptic events, its modulation by both lithium and valproate, and supporting evidence from animal models of mania, it was postulated that the attenuation of PKC activity might have efficacy in treating mania. These findings led to a single-blind clinical trial investigating possible antimanic properties of the PKC inhibitor tamoxifen. While best known for its anti-estrogenic properties, tamoxifen is also a potent PKC inhibitor at high concentrations. The treatment resulted in a significant decrease in manic symptoms in five of seven patients enrolled in the initial trial. We recently conducted a double-blind placebo-controlled with tamoxifen in patients with bipolar mania and found it to have significant antimanic effects.

Neurotrophic Signaling Cascades

Neurotrophins (NTs) are a family of regulatory factors that mediate the differentiation and survival of neurons, as well as the modulation of synaptic transmission and synaptic plasticity. The NT family now includes, among others, nerve growth factor (NGF), BDNF, NT3, NT4/5, and NT6. BDNF and other neurotrophic factors are necessary for the survival and function of neurons, implying that a sustained reduction of these factors could affect neuronal viability. BDNF also has a number of much more acute effects on synaptic plasticity and neurotransmitter release, and facilitates the release of glutamate, γ-aminobutyric acid (GABA), dopamine, and serotonin.

Thus, BDNF is best known for its long-term neurotrophic and neuroprotective effects, which may be very important for its putative role in the pathophysiology and treatment of mood disorders. It is now clear that their survival-promoting effects are mediated in a large part by an inhibition of cell-death cascades. Increasing evidence suggests that neurotrophic factors inhibit cell-death cascades by activating the extracellular response kinase (ERK) signaling pathway, the PLC-γ cascade, and the phosphotidylinositol-3 kinase (PI-3K)/Akt pathway. Chronic stress (21 days foot-shock; an animal model of depression) induced a marked and persistent ERK1/2 hyper-phosphorylation in dendrites of the higher prefronto-cortical layers, while phospho-CREB was reduced in several cortical regions including frontal cortex. Since CREB is phosphorylated and activated by phospho-ERK1/2 directly, this reduction indicates that chronic stress could downregulate CREB phosphorylation indirectly, and subsequently downregulate the transcription of some genes such as Bcl-2 and BDNF. In this context, severe stress exacerbates stroke outcome by suppressing Bcl-2 expression. In this study, mice exposed to aggressive social stress expressed ~70% less Bcl-2 mRNA than unstressed mice following ischemia. Furthermore, stress greatly exacerbated infarct in control mice but not in transgenic mice that constitutively express increased neuronal Bcl-2. Finally, high corticosterone concentrations were significantly correlated with larger infarcts in wild-type mice but not in Bcl-2-overexpressing transgenic mice. Thus, enhanced Bcl-2 expression appears to be capable of offsetting the potentially harmful consequences of stress-induced neuronal endangerment, and suggests that pharmacologically induced upregulation of Bcl-2 may have considerable utility in the treatment of a variety of disorders associated with endogenous or acquired impairments of cellular resilience. The neurotrophic factor/ERK mitogen activated protein (MAP) kinase/Bcl-2 signaling cascade appears to play a vital role in cell survival in the CNS, and a fine balance is maintained between the levels and activities of cell survival and cell death factors. Dysregulation of the BDNF-ERK-CREB coordination may be a key mechanism by which prolonged stress induces atrophy of selective subpopulations of vulnerable neurons and/or distal dendrites.

In view of the important role of ERK MAP kinases in mediating long-term neuroplastic events, it is notable that lithium and valproate, at therapeutically relevant concentrations, have been demonstrated to activate the ERK MAP kinase cascade in human neuroblastoma SH-SY5Y cells and in critical limbic and limbic-related areas of rodent brain. Neurotrophic factors are now known to promote cell survival by activating MAP kinases to suppress intrinsic, cellular apoptotic machinery, not by inducing cell-survival pathways. Thus, a downstream target of the MAP kinase cascade, ribosomal S-6 kinase (Rsk), phosphorylates the CREB and this leads to induction of Bcl-2 gene expression. Consistent with an activation of neurotrophic signaling cascades, chronic treatment of rats with therapeutically relevant doses of lithium or valproate produces an increase in the activation of Rsk and CREB and a doubling of Bcl-2 levels in frontal

cortex, effects which are primarily due to a marked increase in the number of Bcl-2 immunoreactive cells in layers II and III of frontal cortex. Interestingly, the importance of neurons in layers II–IV of the frontal cortex in mood disorders has recently been emphasized, since primate studies indicate that these areas are important for providing connections with other cortical regions, and that they are targets for subcortical input. Further suggestive evidence that lithium and valproate activate the ERK MAP kinase pathway and/or targets of this pathway comes from the finding that both lithium and valproate increase the expression of BDNF in rodent brain following chronic treatment.

Consistent with its effects on neurotrophic signaling cascades, lithium has been demonstrated to be neuroprotective in cell culture models and animal models of ischemia and Huntington's disease. Recent evidence suggests that the neuroprotective effect of lithium in cortical neurons requires BDNF expression. Valproate also exerts neuroprotective actions in a number of cellular models including glutamate toxicity, β-amyloid toxicity, and following exposure to other toxins.

Human evidence for the neurotrophic effects of lithium Drevets and colleagues reanalyzed previous data demonstrating $\sim$40% reductions in subgenual prefrontal cortex (PFC) volumes in familial mood disorder subjects and found that the patients treated with chronic lithium or valproate exhibited subgenual PFC volumes which were significantly higher than the volumes in non-lithium- or valproate-treated patients, and not significantly different from control. In a further analysis of these data, unipolar patients exhibited reduced glial cell densities, whereas only the bipolar patients off chronic lithium or valproate exhibited similar reductions.

After four weeks of lithium treatment in depressed bipolar patients, a significant increase in *N*-aceytl-aspartate (NAA) levels (a putative marker of neuronal viability) was found with magnetic resonance spectroscopy (MRS); these effects were localized almost exclusively to gray matter. These findings indirectly support the contention that chronic lithium increases neuronal viability/function in the human brain. Furthermore, a $\sim$0.97 correlation between lithium-induced NAA increases and regional voxel gray matter content was observed, thereby providing evidence for co-localization with the regional specific Bcl-2 increases observed (e.g., gray vs. white matter) in the rodent brain cortices. These results imply that chronic lithium exerts robust neuroprotective effects (as has been demonstrated in a variety of preclinical paradigms) as well as neurotrophic effects in humans.

A follow-up volumetric MRI study demonstrated that 4 weeks of lithium treatment also significantly increased total gray matter content in the human brain, suggesting an increase in the volume of the neuropil (the moss-like layer comprised of axonal and dendritic fibers which occupies much of the cortex gray matter volume). A subregional analysis of this brain imaging data showed prominent effects in areas that have been implicated in depression, the hippocampus and caudate.

AA Metabolism

AA functions as an important mediator of second-messenger pathways within the brain. It is released from membrane phospholipids via receptor/G-protein-initiated activation of phospholipase A2. This action results in release of AA from the cellular membrane and cyclooxygenase-mediated production of eicosanoid metabolites such as prostaglandins and thromboxanes. These metabolites mediate numerous subsequent intracellular responses and transynaptic responses.

Chronic lithium and valproate treatment of rats results in selective reductions in the turnover rate in the brain phospholipids of AA. In addition, lithium downregulates the gene expression and protein levels of an AA-specific phospholipase A2 (cPLA2) and the protein levels of cycloxygenase-2 (COX-2). Valproate decreases turnover of AA and protein levels of COX-1 and COX-2. These findings suggest that effects of mood stabilizers on cell membranes, and specifically AA turnover, might be relevant to the pharmacological action of lithium and valproate. Further general support for the involvement of the AA signaling pathway in bipolar disorder comes from other preclinical and clinical studies. Administration of nonselective cycloxygenase inhibitors indomethacin and piroxicam in rats prevented amphetamine-stimulated locomotor activity and blocked cocaine sensitization (both are rodent models of mania). Furthermore, inhibition of COX-2 with NS-398 attenuates restraint stress (a model of depression)-induced oxidative changes. In humans cyclooxygenase-2 inhibitor celecoxib was recently found to have antidepressant effects in a double-blind, placebo-controlled trial.

Conclusions

There has been little progress in developing truly novel drugs for the treatment of bipolar disorder. Thus, there exists a critical need to develop novel approaches for the treatment of recurring mood disorders. We predict that they will likely derive from understanding both mechanisms of current drug actions. Of the direct targets, GSK-3 has received the most attention and GSK-3 inhibitors are rapidly being developed. In addition to the direct

targets of current mood stabilizers lithium, valproate, and carbamazepine, we also discussed signaling pathways regulated by these agents including phosphoinositol/PKC, cyclic AMP, AA signaling, and neurotrophic signaling pathways. Protein kinase C inhibition appears to be important to the antimanic effects of lithium and valproate as indicated by the efficacy of tamoxifen in reducing manic symptoms. Mechanisms to enhance neurotrophic pathways is a major focus for the treatment of neurodegenerative disorders, and it is likely that novel medications for this purpose may be soon available.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Anxiety: Drug Therapy; Cognitive Dysfunction in Psychiatric Disorders; Enteric Nervous System: Neurotrophic Factors; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Prepulse Inhibition of Startle in Humans and Laboratory Models; Psychiatric Disorders: Functional Genomics.

Further Reading

Chen G, Manji HK, Hawver DB, Wright CB, and Potter WZ (1994) Chronic sodium valproate selectively decreases protein kinase C alpha and epsilon *in vitro*. *Journal of Neurochemistry* 63: 2361–2364.

DeVries AC, Joh HD, Bernard O, et al. (2001) Social stress exacerbates stroke outcome by suppressing Bcl-2 expression. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11824–11828.

Drevets WC, Price JL, Simpson JR Jr., et al. (1997) Subgenual prefrontal cortex abnormalities in mood disorders. *Nature* 386: 824–827.

Duman RS (2002) Synaptic plasticity and mood disorders. *Molecular Psychiatry* 7(supplement 1): S29–S34.

Einat H, Belmaker RH, and Manji HK (2003) New approaches to modeling bipolar disorder. *Psychopharmacology Bulletin* 37: 47–63.

Gould TD, Zarate CAJ, and Manji HK (2004) Glycogen synthase kinase-3: A target for novel bipolar disorder treatments. *Journal of Clinical Psychiatry* 65: 10–21.

Hashimoto R, Takei N, Shimazu K, Christ L, Lu B, and Chuang DM (2002) Lithium induces brain-derived neurotrophic factor and activates TrkB in rodent cortical neurons: An essential step for neuroprotection against glutamate excitotoxicity. *Neuropharmacology* 43: 1173–1179.

Jope RS and Bijur GN (2002) Mood stabilizers, glycogen synthase kinase-3beta and cell survival. *Molecular Psychiatry* 7(supplement 1): S35–S45.

Manji HK and Duman RS (2001) Impairments of neuroplasticity and cellular resilience in severe mood disorders: Implications for the development of novel therapeutics. *Psychopharmacology Bulletin* 35: 5–49.

Manji HK and Lenox RH (1999) Ziskind–Somerfeld Research Award. Protein kinase C signaling in the brain: Molecular transduction of mood stabilization in the treatment of manic-depressive illness. *Biological Psychiatry* 46: 1328–1351.

Moore GJ, Bebchuk JM, Wilds IB, Chen G, and Manji HK (2000) Lithium-induced increase in human brain grey matter. *Lancet* 356: 1241–1242.

Nishizuka Y (1992) Intracellular signaling by hydrolysis of phospholipids and activation of protein kinase C. *Science* 258: 607–614.

Rajkowska G (2000) Histopathology of the prefrontal cortex in major depression: What does it tell us about dysfunctional monoaminergic circuits? *Progress in Brain Research* 126: 397–412.

Rapoport SI and Bosetti F (2002) Do lithium and anticonvulsants target the brain arachidonic acid cascade in bipolar disorder? *Archives of General Psychiatry* 59: 592–596.

Williams RS, Cheng L, Mudge AW, and Harwood AJ (2002) A common mechanism of action for three mood-stabilizing drugs. *Nature* 417: 292–295.

Zarate CA Jr., Singh JB, Carlson PJ, et al. (in press) Efficacy of a protein kinase C inhibitor (tamoxifen) in the treatment of acute mania: A pilot study. *Bipolar Disorders*.

Morphogens: History

L Wolpert, University College, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Morphogens refer to substances thought to be involved in the patterning of cells during embryonic development. It is a term coined by Alan Turing in 1952 for a substance whose distribution by diffusion could determine the development of cells which would respond differently to different concentrations of the morphogen. But the patterning of cells in relation to varying concentrations of a substance, particularly graded concentrations, has a long history. Of particular importance is how the position of cells in the developing embryo is specified so that they develop in an appropriate manner. It is still not clear whether diffusible morphogens provide cells with positional information and so pattern a tissue during development.

Gradients and Polarity

In 1888, Wilhelm Roux, an embryologist, reported that when one of the cells at the two-cell stage of the frog embryo was killed, the remaining cell formed a half embryo. This was interpreted to mean that different structures were already specified to develop from the two cells. But then Hans Driesch, just 3 years later, set out to repeat the experiment on sea urchin embryos. Contrary to his expectation, he found that half an embryo could give rise to a small but normal whole embryo. This was the first example in a developing system of what is known as regulation. It showed that there is a fundamental distinction between prospective and possible fates of cells – the possible fates being greater.

Further experiments by Driesch on the sea urchin embryo at a later stage led him to believe that the prospective fate of any cell in the early embryo was determined by its position in relation to the whole embryo. The 'whole' was related to the three Cartesian axes that could be drawn on the embryo – the x, y, and z axes. So in his model, each cell had its position defined. He argued that no matter which part of the early embryo is removed, it will develop normally. This is not true, but he ignored evidence to the contrary.

His model was what he called a harmonious equipotential system; the fate of any cell, or group of cells, was a function of the size of the embryo and it relative position. Driesch rejected any chemical theories to explain his complex model and proposed the concept of 'entelechy,' a vitalistic mechanism. There was no further progress along these lines, but he had emphasized the importance of position in embryonic development.

For any system based on coordinates or position, polarity plays a key role, for it can determine the direction in which a coordinate can increase or decrease. Interest in polarity initially came from studies on regeneration. Thomas Hunt Morgan in 1905 was the first to clearly state the relationship between gradients and polarity. In studying the regeneration of earthworms he put forward the idea of gradients of formative stuffs. He found, for example, that the further back a cut is made, the longer it takes to regenerate a head. "...We might speak of the cells of the worm containing some sort of stuff that is more or less abundant in different parts of the body. The head stuff would gradually diminish as we pass posteriorly...." He came to a similar conclusion studying the regeneration of *Tubularia*: "We may assume that the gradation of the material is of such a kind that the hydranth forming material decreases from the apical towards the basal end." But Morgan did not pursue these important ideas on polarity and gradients.

They were, however, taken up by Child in 1911. Studying regeneration in planaria, he concluded, "All these facts indicate that a graded difference of some sort in the dynamic processes exists along the axis.... It is this dynamic gradation along the axis, together with the complex of correlative conditions associated with it, which I regard as constituting physiological polarity. According to this idea polarity is not a condition of molecular orientation, but is essentially a dynamic gradient in one direction...."

Child went on to propose that the gradient was in metabolism, and he introduced the concept of dominance, drawing an analogy between the organism and the state – both require authority. The dominant region at one end determines this metabolic gradient by determining the rates of reaction at other levels. He gave no indication of how this was done. In support of his ideas he used evidence from experiments in which poisons such as cyanide had less effect on regions with the highest metabolism and so they would die last. All the references in his 1915 book are, with one exception, to his own work. Even as late as 1941 he was still constrained by metabolism. He argued that "...if decrease in concentration of amount of a certain substance or substance-complex occurs in one direction along a gradient, there must be increase in concentration or amount of another substance or substances unless there is a decrease in

volume." It was only when metabolism was replaced with the concept of information that new ideas emerged.

Developmental biologists were at this time becoming interested in gradients in eggs and paid no attention to the work of Morgan or Child. For example, in 1915 the Swedish embryologist Runnstrom studied the effect of different ions on the development of the sea urchin embryo and concluded that there was a chemical material expressed along the axes of the embryo and this ". . . .material is localised in a way that different parts of the embryo have different concentrations. It creates a concentration gradient. The phenomenon of polarity is an expression of the presence of this concentration gradient." Later studies by Runnstrom and Horstadius provided further evidence for gradients. One of the most important experiments in 1935 which illustrated the quantitative nature of gradients along the main axis of the sea urchin embryo – the animal vegetal axis – involved showing that in order for normal development to occur, the number of vegetal cells that were required to be added to the animal pole fragment was greater that that required for a fragment from a slightly more vegetal region.

At this time another approach to embryonic development involved induction and the concept of the embryonic field. Spemann in the 1930s discovered induction of a new axis and the organizer in the amphibian embryo. One part of the embryo, the dorsal lip of the blastopore, the site where gastrulation begins, could, when grafted to another embryo, induce almost a complete second axis. This was a most influential discovery. Yet as early as 1909 it had been shown that the head region of hydra could, when grafted to the body of another animal, induce the formation of a new axis. Spemann also introduced the field concept, referring to different fields of organization in the embryo.

Perhaps the clearest views of gradients, some 30 years after their initial conception, came from Huxley and de Beer in 1934 based largely on studies on regeneration. Their rules were that the origin of polarity is to be sought in external factors, though in some cases it may already be polarized; in regeneration the apical, or head, region is the first formed and is autonomous; and the apical region influences adjacent regions and prevents them forming an apical region.

In 1938 he lectured on gradient theory, but one of the problems he recognized was understanding how quantitative differences along the gradient could give rise to qualitative differences, namely, complex cell patterns. How could discontinuities arise from a continuous gradient? A possible solution came from Dalcq and Pasteels. Considering early amphibian development, they introduced two factors: the yolk gradient V and a cortical factor C which was graded dorsoventrally. They then defined the product CV as the morphogenetic potential as well as the concept of a threshold. How parts of the embryo developed was then determined by whether CV was below or above a threshold.

Other specific models were developed, and that of Rose in 1968 was the first gradient model to generate a pattern. He suggested that the genesis of a cellular pattern could be the consequence of a hierarchy of self-limiting reactions, together with a spread of inhibition form one differentiating region to others further down the gradient. The gradient determined the rate of differentiation, with the reactions proceeding fastest at the high point of the gradient. The reaction that predominates is then that closest to the top of the hierarchy which has not yet been inhibited. Each reaction is self-limiting by virtue of the inhibitor it produces. However, there was no evidence for the proposed inhibitors.

Insects were to provide crucial information on the possible role and nature of gradients. The gradient concept gained considerable support from Klaus Sander's 1960 experiments on the axial body pattern of the leafhopper *Euscelis* embryo, with the results on the reversal of the segment sequence providing strong evidence for gradients. At about the same time, Locke had shown that the ripple patterns in the cuticle of the insect *Rhodnius* could be explained by a gradient along the segment.

Hidegard Stumpf interpreted these results in terms of a concentration gradient in a substance produced at one margin and destroyed at the other – the first clear statement as to how a concentration gradient could be set up. Using the abdominal segments of *Galleria*, which are divided axially into three parts, she then did experiments to determine whether this pattern was based on a gradient. She concluded that this was the case and that the gradient had two functions: to orient the scales in a direction of decreasing concentration, and to provide the cells with specific concentrations to determine their fate. At the same time, and quite independently, Lawrence had proposed a gradient model for the epidermis of the insect *Oncopeltus* based on the orientation of the hairs and bristles. His gradient model was based on a gradient in sand together with active transport against diffusion.

Positional Information

Again quite independently, the author put forward the concept of positional information similar to that

proposed by Stumpf, on the basis of the French flag problem. I had been very impressed that sea urchin embryos could develop a normal pattern of tissue proportions over an eightfold size range. Again, *Hydra* could have a normal pattern over a large range of sizes. The French flag problem addressed how a line of cells, each of which could differentiate into a blue, white, or red cell, could develop the French flag pattern – one-third blue, one-third white, one-third red. Moreover, this pattern would develop irrespective of the number of cells in the line. A clear solution was that each cell knew its position in the line of cells and then could differentiate in an appropriate manner to give the French flag. One way cells could have their position specified was by a gradient in some substance whose concentration was fixed at both ends of the line, one high, the other low, and that varied linearly between the two ends – simple thresholds would give the flag reliably. A very attractive feature of positional information was that it could account for genetic mosaics in the wing and antenna of the fruit fly *Drosophila*.

Curt Stern had shown that mosaics of leg and antenna tissue developed according to where they were along the proximo-distal axis – they had the same positional values but interpreted them differently because of their position along the main body axis.

In 1952 Alan Turing, the famous mathematician, code breaker, and inventor of computers, published a totally new approach to pattern formation. It was essentially a self-organizing mechanism for setting up a pre-pattern in a diffusible substance that could cause structures to form at specific sites.

> It is suggested that a system of chemical substances, called morphogens, reacting together and diffusing through a tissue, is adequate to account for the main phenomena of morphogenesis. Such a system, although it may originally be quite homogeneous, may later develop a pattern or structure due to an instability of the homogeneous equilibrium, which is triggered off by random disturbances. Such reaction-diffusion systems are considered in some detail in the case of an isolated ring of cells, a mathematical convenient, though biologically unusual system. It is found that there are essentially six different forms it might take. In the most interesting form, stationary waves appear on the ring. It is suggested that this might account, for instance, for the tentacle pattern of Hydra and for whorled leaves.

This approach has been extended by the reaction diffusion models of Gierer and Meinhardt which can set up and regulate axial gradients.

One feature of what might be systems in which positional information is involved is that they are relatively small, none being greater than 0.5 mm maximum linear dimension, and there are typically fewer than 30 cells along an axis where position is thought to be specified. The time required to specify position appears to be on the order of hours. This led Crick to propose the existence of a diffusible morphogen for setting up positional fields. A simple model would have the morphogen produced locally at a source at one end, and then it would spread by diffusion along the axis. The resulting concentration gradient could provide positional information.

By far the best evidence for such a gradient in a morphogen patterning a developmental system is the bicoid gradient in the *Drosophila* egg. The gradient in this protein is maternal and runs from anterior to posterior. However, it acts at a time when the embryo is multinucleate, there are no cell wall membranes for it to cross, and the nuclei share a common cytoplasm. At a specific concentration it activates the gene hunchback in the nuclei in the anterior region. Another example in the early embryo relates to the dorso-ventral axis. For cellular systems things are very much more complicated.

Positional signaling by morphogens has been quite widely investigated over the past two decades but it has not been established that this mechanism provides the basis for pattern formation in any multicellular system. In principle it can account for the pattern of digits in the developing wing of the chick embryo. At the posterior margin of the limb bud is a signaling region, the polarizing region, that produces a putative morphogen, the protein sonic hedgehog. The normal pattern of digits from posterior to anterior is 4, 3, 2. The polarizing region is posterior to the future digit 4. If an additional polarizing region is grafted to the anterior margin of the bud, the resulting pattern of digits is 4, 3, 2, 2, 3, 4. The best evidence that the signal is graded comes from the observation that manufacturing a reduced signal by grafting anteriorly a small number of polarizing tissue cells results in a reduced response; if only a small piece is grafted, then the pattern is 4, 3, 2, 2. This fits nicely with a simple graded diffusible signal of sonic hedgehog, but there is no good evidence to support it. Recent studies have shown that the time that the cells are in contact with the polarizing region also plays a role. Again, the elegant model for specification of neuronal subtypes in the vertebrate ventral neural tube is based on a gradient in sonic hedgehog, but direct evidence is still lacking.

Another well-studied multicellular system is the imaginal wing disc of the *Drosophila* embryo. The wing disc is divided into anterior and posterior compartments. The diffusible molecule Hedgehog is expressed in the posterior compartment, and this causes diffusible Decapentaplegic (DPP) to be expressed at the compartment boundary. DPP is indeed

distributed in a long-range concentration gradient; DPP moves without preferential direction at a speed of more than four cells per hour through the target tissue, in spite of which the shape of the gradient remains stable. However, DPP extracellular diffusion alone does not explain its distribution as a stable gradient and that receptor-mediated endocytosis is essential for DPP long-range movement. Although there are some indications that at specific DPP thresholds the genes omb and spalt are turned on, it is notable that even in this well-studied case there is scant if any reliable quantitative evidence for a gradient-dependent positional gene activation.

In early *Xenopus* embryo there is a presumed gradient of an activin-like morphogen activity. At high concentrations of activin the gene goosecoid is activated while Xbra is turned on at lower concentrations. This has been demonstrated by placing a bead of activin in the middle of a group of animal cap cells: goosecoid is turned on nearer the bead than Xbra. These results have been interpreted in terms of the morphogens binding to receptors. However, there is no evidence for an actual diffusion-mediated gradient of activin in the embryo.

One could list many other examples, with the conclusion that no clear case of a canonical morphogen has been found. But even in principle, could diffusing morphogens by themselves specify positional information? We do not know how precise specification of position has to be, but it could be at the single cell level. There are major difficulties. The molecular concentrations involved are invariably small, hence there is an intrinsic chemical noise. Candidate morphogens and their receptors are very dilute, and they operate in a complex physicochemical environment. This potentially makes morphogen propagation extremely unreliable. The extracellular matrix most likely will present a variety of binding sites for the morphogen, trapping it. There are indeed numerous examples of interaction between proposed morphogens and other extracellular proteins. In addition, the extracellular space is tortuous. It has been estimated that this can increase the effective diffusion times as much as fivefold. Many cellular factors can be expected to perturb and/or actively modulate these various constraints. There is also almost always some endocytosis of a diffusing molecule that binds to cell receptors, and this together with tortuosity makes a precise determination of morphogen concentrations problematic. Worse still, the binding of the morphogen to the receptors can have a very significant effect on the distribution and can even prevent a gradient in receptor occupation as receptors become saturated.

It may be that cell–cell interactions play a key role in specifying positional information. Cell–cell interactions figure prominently in all mechanisms relating to tissue polarity. Favorite model systems for studying planar polarization are the *Drosophila* wing and eye. It is widely assumed that one or several (some opposed) gradients are involved. Gradient-distributed molecules such as Dachsous, Four-jointed, and Fat have been identified. Remarkably, while all these candidate players are growth factors, all are membrane proteins; they intervene in cell adhesion and communication, but likely not by being transmitted between cells.

There is a model for polarity in the abdominal epidermis of *Drosophila* based on cell interactions which also provides the cells with positional information – no morphogen playing a key role. The proteins Four-jointed, Dachsous, and Fat are involved in setting up a gradient of an unknown factor X, which determines Frizzled activity. Reading the gradient are Frizzled itself, Prickle, and VanGogh/Strabismus, enabling cells to compare their level of X with that of their neighbors and to set their value as an average of these. On the other hand, there is a model for planar cell polarity based purely on cell-to-cell interactions and local (cell-scale) graded distributions, where factor X is dispensed with, and hence no global positional information is imparted to cells. The model posits that a weak Frizzled activity gradient is read by asymmetric molecular complexes built at cell interfaces around the cadherin Flamingo. Polarization happens robustly over hundreds of cell diameters even with much noise in the Frizzled gradient. It is probable that these last two models have the potential, if suitably extended, to provide positional information.

See also: Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Segmentation: Segmental Boundaries, Establishment and Morphogenesis (EPH).

Further Reading

Child CM (1941) *Patterns and Problems of Development*. Chicago: University of Chicago Press.

Cho E and Irvine KD (2004) Action of fat, four-jointed, dachsous and dachs in distal-to-proximal wing signaling. *Development* 131: 4489–4500.

Crick F (1970) Diffusion in embryogenesis. *Nature* 225: 420–422.

Entchev EV, Schwabedissen A, and Gonzalez-Gaitan M (2000) Gradient formation of the TGF-beta homolog Dpp. *Cell* 103: 981–991.

Gierer A (1981) Some physical, mathematical, and evolutionary aspects of biological pattern formation. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* B295: 417–425.

Gurdon JB and Bourillot PY (2001) Morphogen gradient interpretation. *Nature* 413: 797–803.

Han C, Belenkaya TY, Wang B, and Lin X (2004) Drosophila glypicans control the cell-to-cell movement of Hedgehog by a dynamin-independent process. *Development* 131: 601–611.

Harfe BD, Scherz PJ, Nissim S, Tian H, McMahon AP, and Tabin CJ (2004) Evidence for an expansion-based temporal Shh gradient in specifying vertebrate digit identities. *Cell* 118: 517–528.

Huxley JS and de Beer GR (1934) *The Elements of Experimental Embryology.* Cambridge, UK: Cambridge University Press.

Kerszberg M and Wolpert L (1998) Mechanisms for positional signalling by morphogen transport: A theoretical study. *Journal of Theoretical Biology* 191: 103–114.

Kruse K, Pantazis P, Bollenbach T, Julicher F, and Gonzalez-Gaitan M (2004) Dpp gradient formation by dynamin-dependent endocytosis: Receptor trafficking and the diffusion model. *Development* 131: 4843–4856.

Lander AD, Nie Q, and Wan FY (2002) Do morphogen gradients arise by diffusion? *Developmental Cell* 2: 785–796.

Lawrence PA, Casal J, and Struhl G (2004) Cell interactions and planar polarity in the abdominal epidermis of Drosophila. *Development* 131: 4651–4664.

Le Garrec JF, Lopez P, and Kerszberg M (2006) Establishment and maintenance of planar epithelial cell polarity by asymmetric cadherin bridges: A computer model. *Developmental Dynamics* 23: 235–246.

Ma D, Yang CH, McNeill H, Simon MA, and Axelrod JD (2003) Fidelity in planar cell polarity signalling. *Nature* 421: 543–547.

Panakova D, Sprong H, Marois E, Thiele C, and Eaton S (2005) Lipoprotein particles are required for Hedgehog and wingless signalling. *Nature* 435: 58–65.

Tabata T and Takei Y (2004) Morphogens, their identification and regulation. *Development* 131: 703–712.

Teleman AA, Strigini M, and Cohen SM (2001) Shaping morphogen gradients. *Cell* 105: 559–562.

Tickle C (2003) Patterning systems—From one end of the limb to the other. *Developmental Cell* 4: 449–458.

Mother–Infant Interaction in the Variable Foraging Demand Model

J D Coplan, J G Kral, E L P Smith, and L A Rosenblum, SUNY Downstate Medical Center, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Background

Although the etiology of psychiatric disorders has been difficult to ascertain, it is clear that early-life stress is a pivotal precursor to adult mood and anxiety disorders in humans. Teasing apart the complex interplay between genes and environment that shapes the vulnerability to the emergence of psychiatric disorder has progressed significantly. Genotyping for polymorphisms that predict vulnerability to early life stress, particularly allelic variants of the serotonin transporter protein-linked promoter region (5HTTLPR), have greatly advanced the field. Nevertheless, animal models afford the advantage of random assignment to differential rearing experimental conditions and standardize the early-life stressor, thereby minimizing the contribution of genetics plays in determining the nature and perception of early-life stress. Animal models are therefore predicted to account for an ever-increasing amount of variance contributed to by genetic contributions, early-life stress, and their interaction in determining the ultimate impact of these factors on the adult phenotype. One animal model of early-life stress, developed in nonhuman primates, focuses on the grown offspring of mothers confronting, during their infants' early development, significant environmental uncertainty in regard to the ease or difficulty of maternal food procurement (variable foraging demand (VFD)). As in humans, integrity of maternal care in nonhuman primates is symbolic of survival for the infant. This VFD model is therefore argued to represent an animal analog of the impact of emotional inconsistency in human parenting. Adverse early parenting may likely represent, for the affected infant, the earliest and potentially most profound of all life stressors. We have therefore demonstrated that VFD offspring exhibit, in comparison to consistently reared controls, both sustained and pervasive alterations within a diverse range of physiological and neurobiological systems that are consistently implicated in the pathogenesis of human mood and anxiety disorders.

Perhaps of greatest significance is that VFD offspring are subject to errant social functioning and manifest concurrent behavioral disturbances reflective of affective distress, thereby mirroring certain components of human depression and anxiety (Figure 1).

VFD and Control Rearing

VFD rearing is an early-life stress paradigm in which infant bonnet macaques are reared by mothers undergoing an experimentally induced 'perception' of food uncertainty. Although adequate food is always available, randomly assigned macaque mothers face unpredictability of food availability through exposure to the VFD procedure, whose duration is only 12–16 weeks within the first year of infant life. Hence, no differences in weight change are evident between VFD and non-VFD mothers or infants. We have shown that psychopathology can be experimentally induced in the infant by this discrete period of emotional compromise in the maternal behavioral repertoire, imparting pervasive and sustained biobehavioral effects within the developmental trajectory of the offspring, into adulthood. Notable is the absence of physical intervention or frank separation in the VFD model, thereby closely approximating, despite parental presence, the far-reaching experience of emotional unavailability, so frequently reported in symptomatic human adults.

Experimental Overview of the VFD

Rosenblum and Pauley developed VFD in 1984, in order to provide a more humanlike analog of disturbed mother–infant attachment patterns than that provided by the earlier studies of Harlow, in which frank maternal separation from the infant was conducted. The effects of stress on infant in the VFD model were indirect and were intended to be communicated via disruption of mothering following her stress response to unpredictable foraging conditions. Three different experimental rearing conditions were developed, with random assignment to each condition. In general, the VFD paradigm involves two alternating conditions across either a 12- or a 16-week period: (1) Under low foraging demand (LFD) conditions, abundant food can simply be picked from pans within a cart. (2) Under high foraging demand (HFD) conditions, to obtain their daily food ration, mothers are required to search and work for food by digging through clean wood-chip bedding that hides the less abundant yet adequate food located within mobile foraging carts that allow access to internal pans from a series of ports. (3) For the VFD mothers, the foraging demand varies between LFD and HFD conditions in 2-week blocks (3–4 LFD and 3–4 HFD) across the 12- to 16-week experimental rearing period. The quantity of food placed within the cart under the HFD portion of VFD and HFD rearing is adjusted on the basis of the amount of

Long-term impact of early-life stress on primate models of anxiety and depression

- Anxious and depressive-like behavior
- CRF dysregulation
  HPA axis activity alterations
- CSF monoamine metabolite elevations
- Somatotropic axis suppression
- Pro-inflammatory cytokine expression
- Metabolic syndrome profile
- Regional spectroscopic analysis of neuroanatomical structures
- Myelin integrity compromise on diffusion tensor imaging
- Serotonin transporter protein genotypic effects

Figure 1 The long-term impact of early life stress in the form of VFD in nonhuman primates in relationship to the pathophysiology of human anxiety and mood disorders.

Figure 2 VFD stress is transmitted to the offspring via the mother, with disruption of maternal attachment and affective reciprocity. Depicted here is a mother undergoing VFD; she is noted to be distracted from her infant, who appears pensive, despite her physical presence. We have described the VFD infant as undergoing a state of 'functional' separation.

food eaten each day. Although the requirements placed on the mothers for obtaining food differs in each condition, in all instances, adequate food is always available. Thus, food deprivation is not a factor, and weekly weight measures and daily health checks are performed to ensure normal growth and health in all infants and their mothers.

Prior to and following the experimental rearing period, ad lib food is available to all animals via food bins as in the rest of the primate facility. Water, via automatic drinking spouts, is constantly available to all mothers. VFD-reared subjects were found to be most disturbed on behavioral measures, whereas HFD-reared subjects were generally indistinguishable from LFD subjects. These early studies led to the conclusion that it was the unpredictability of foraging conditions that produced long-term abnormality in offspring, rather than difficulty of foraging *per se* as seen in the HFD. In most subsequent studies, VFD-reared subjects were compared to LFD-reared subjects.

VFD stress was hypothesized to be transmitted to the offspring via the mother, via disruption of maternal attachment and normative affective reciprocity. Thus, although the mother is physically present, a form of 'functional' separation exists (Figure 2). The VFD paradigm is also accompanied by synchronized elevations of maternal and infant cerebrospinal fluid (CSF) concentrations of corticotropin releasing factor (CRF).

VFD Behavioral Observations

VFD-reared monkeys, in comparison to LFD-reared subjects, demonstrate a range of behavioral abnormalities at successive stages of maturation. During infancy, in dyadic assessments shortly after the rearing experience, VFD rearing adversely affected the

Figure 3 Infants reared under VFD conditions exhibit an increased propensity to develop a separation despair response during periods of brief maternal separation. Depicted here is a 'VFD-monkey' infant undergoing a despair response following a brief period of maternal separation.

subjects' ability to separate from mother to explore a novel physical environment in comparison to LFD subjects. VFD-reared infants also manifest timidity, social subordinance, increased separation despair, and clinging behavior in novel environments. When tested during late adolescence using a social challenge paradigm, VFD exhibited diminished capacity for affiliative social engagement relative to LFD monkeys, to whom they were socially subordinate. The VFD 'syndrome' is remarkably similar to that seen in Kagan's behaviorally inhibited children and Suomi's genetically shy rhesus monkeys (Figure 3).

In examining whether VFD rearing produced affective disturbance in unperturbed home cage behavioral

observations in comparison to LFD-reared subjects, VFD juvenile subjects exhibit reduced social engagement, increased affective withdrawal, and increased social confrontation, manifest as an increase in social timidity in comparison to controls. Because of potential co-linearity of behaviors, a multiple regression model was employed, using rearing group as the predictor variable and these latter three behaviors as dependent variables. The model accounted for 70% of the variance for group differences. Taken together, the behavioral observations demonstrated a persistently abnormal behavioral profile analogous to certain aspects of anxiety and depression in human children and adults.

VFD Behavioral Findings in Response to an Acute Fear Stimulus

One method by which further information can be obtained regarding VFD rearing is to study behavioral deficits that only become apparent under challenge conditions. One such challenge condition is an acute fear stimulus represented by a masked intruder paradigm, the mask containing large fangs and consisting of unfamiliar colors. VFD and LFD subjects were observed at three stages over 3 days: at baseline, directly following exposure to the fear stimulus, and 1 day following exposure to the fear stimulus. VFD demonstrated, at baseline, marginally reduced gregariousness scores in comparison to normally reared control subjects. However, following a brief exposure to the fear stimulus, normally reared juvenile subjects exhibited a marked reduction in gregariousness scores, whereas VFD, in contrast, failed to shift from basal levels of gregariousness. Further evidence for a muted behavioral reaction to the fear stimulus was the failure of VFD subjects to exhibit emotional responsivity to the masked intruder, whereas LFD subjects showed a fivefold increase for this same measure upon stimulus presentation. LFD responded to the fear stimulus by reducing social engagement, whereas VFD showed a slight increase in social engagement. The differential pattern of behavioral response to an acute fear stimulus supported the notion that VFD showed a reduction of emotional reactivity, and suggested long-term disturbances in affect regulation following adverse early-life experience.

Neurobiological Sequelae in VFD-Reared Subjects

Since the long-term behavioral effects of VFD rearing bore a striking resemblance to certain traitlike aspects of human psychopathology, it was deemed that the VFD model may represent an important vehicle by which to determine the long-term neurobiological correlates of these errant primate behaviors. Using the VFD model, we have subsequently documented a diverse range of persistent sequelae, of relevance to both human psychiatric and medical disease states. We first examined the noradrenergic and serotonergic systems, as parallel 'cross-talk' between these systems has been implicated in the pathophysiology and treatment mechanisms of anxiety and mood disorders. In biobehavioral studies, adult VFD subjects exhibited abnormal behavioral responsivity to the noradrenergic probe, yohimbine, and the serotonin agonist, 2-(2-methyl-4-chlorophenoxy)propionic acid (MCPP). VFD-reared subjects were hyper-responsive to the anxiogenic effects of yohimbine and hyporesponsive to the behavioral effects of MCPP. Based on norepinephrine (NE) hyper-responsivity in the VFD-reared subjects, we examined concentrations of CRF, a known stimulant of the noradrenergic system, in the CSF of differentially reared primate subjects. CRF is a neuropeptide which coordinates the body's behavioral, neuroendocrine, and neuroimmunological response to stress, and CRF dysregulation had been implicated in the pathophysiology of mood disorders and posttraumatic stress disorder. Juvenile VFD subjects exhibited persistently elevated concentrations of CSF CRF with concomitant HPA axis suppression in comparison to both LFD- and HFD-reared subjects.

We postulated that sustained alterations of CRF/HPA axis function may be a marker of early-life stress and psychiatric disorder vulnerability. CSF was also assayed for concentrations of biogenic amines and somatostatin, an inhibitory peptide of somatotropic function. VFD subjects exhibited persistent elevations in CSF 5-HIAA (5-hydroxyindoleacetic acid – the chief serotonin turnover metabolite) and HVA (homovanillic acid – the chief dopamine turnover metabolite). CRF correlated positively with all three neurochemical variables in VFD-reared but not non-VFD-reared subjects. Since the CSF CRF elevations had been noted in juvenile VFD-reared subjects, if the finding was truly persistent, it would be important to replicate the finding in the identical subjects followed into adulthood. VFD adult subjects exhibited enduring elevations of CSF CRF in comparison to normally reared controls. Moreover, there was a positive correlation between juvenile and adult VFD CSF CRF, suggesting that CRF elevations were traitlike in their expression.

In a subsequent juvenile cohort, who had been exposed to VFD rearing at a later phase of development coinciding with weaning, paradoxically low CSF CRF and high plasma cortisol concentrations were observed in comparison to LFD control subjects, suggesting that the timing of the VFD stressor could influence the directionality of CRF dysregulation. However, both the early and late forms of VFD rearing exhibited

persistent increases in brain serotonin turnover as reflected by CSF 5-HIAA. Taken together, these studies suggest that one central feature of VFD rearing entailed disturbances of CSF CRF and cortisol, although the directionality of findings evidently varied as a function of rearing age at the time of VFD exposure.

A Hypothesis of Compromised Neurotrophic State in VFD

We hypothesized that a diffuse excitatory influence, possibly imparted by the ubiquitous excitatory amino acid glutamate, could conceivably account for the synchronized neurochemical activation of CRF, somatostatin, serotonin, and dopamine, with CRF as a potential intermediate vehicle of excitation. We therefore hypothesized that persistence of this diffuse neurochemical excitation in VFD-reared subjects may impart a compromise in neurotrophic state, or counter the optimum microenvironment for neuronal activity. One peripheral measure of trophic state is reflected by secretory activity of growth hormone (GH) in response to known secretagogs. It has been well documented that anxiety and mood disorders are generally associated with a reduced GH response to the α_2 agonist clonidine. We therefore determined whether the GH signal in response to clonidine was inversely suppressed in relationship to CSF CRF concentrations. The magnitude of serial CSF CRF concentrations in VFD-reared subjects inversely predicted the GH response to clonidine challenge. The data suggest that the persistent activation of central CRF systems in VFD is associated with a peripheral antitrophic effect, and suggested that central CRF elevation may contribute to the blunted GH response to clonidine observed in humans with mood and anxiety disorders. To test whether glutamate was integral to activation of noradrenergic and central CRF systems in the bonnet macaque, as had been observed in certain cohorts of VFD-reared bonnets, we tested whether LY354740, a metabotropic presynaptic receptor agonist, would block acute yohimbine-induced cortisol increases, and if chronic administration of the glutamate blocking agent would downmodulate the HPA axis. LY354740 attenuated both acute cortisol and MHPG response when administered intravenously. Daily oral administration of the drug led to cortisol reductions several weeks into treatment. Although no rearing group differences were observed in this study, the LY354740 results, in concordance with the field, suggested that glutamate excess could conceivably underlie a neurochemical milieu of excess NE and limbic-HPA axis activation, perhaps compromising neurotrophic state.

VFD and Magnetic Resonance Spectroscopic Imaging

In order to further assess neurotrophic state and regional glutamate concentrations, we conducted neuroimaging studies in VFD adult male primates versus age-matched controls using proton magnetic resonance spectroscopic imaging (^{1}H-MRSI). We evaluated spectral signals of *N*-acetyl-aspartate (NAA), a putative marker of neuronal integrity, while using choline (Cho) and creatine (Cr) as denominators, respectively, in the anterior cingulate cortex (ACC) and hippocampus. Choline is a putative marker of membrane turnover, whereas creatine is used as a control analyte, although this view has come under question. 'Glx' is a spectral signal that represents the combination of glutamate, glutamine, and GABA, although the major contributor to the signal is glutamate. Consistent with our hypothesis that VFD subjects were subject to an antineurotrophic state, NAA/Cr and NAA/Cho resonances were significantly reduced in the ACC of VFD subjects, without overlap with normal controls. The Glx/Cr signal was significantly elevated in the VFD ACC in comparison to controls. Finally, in ACC, Glx/Cr inversely predicted NAA/Cr, suggesting that higher glutamate states may exert an antineurotrophic effect within the ACC. In addition, in hippocampus, mean concentrations of Cho/Cr were elevated in VFD-reared subjects in comparison to controls. Moreover, in left hippocampus, left NAA/Cho was reduced in VFD-reared subjects versus controls. Of note, when combining VFD and non-VFD subjects, both elevated juvenile CSF CRF concentrations and elevated adult body mass index each positively predicted an increase in the right-to-left ratio of the NAA/Cr and NAA/Cho metabolites in hippocampus, suggesting that asymmetry of metabolite profiles within the hippocampus may represent a neural substrate of relevance to certain diverse aspects of the primate phenotype.

Immune Dysregulation

Transforming growth factor-β1 (TGF-β1) is a potent anti-inflammatory cytokine that inhibits T-cell activation. We determined plasma levels of TGF-β1 in VFD-reared subjects and its relationship to affective behavior because of its cross-regulatory interactions with HPA axis activity, serotonin, and other intermediaries. In juvenile subjects, brief confinement in a novel room resulted in higher levels of plasma cortisol and serum TGF-β1 in VFD-reared macaques compared to age-matched controls. In addition, levels of TGF-β1 and cortisol after stress were directly correlated with one another and with CSF 5-HIAA.

Further studies using affymetrix gene microarray methods in monocytes indicate that VFD-reared macaques are characterized by an imbalance in inflammatory homeostasis, as evidenced by a shift in global gene expression to a pro-inflammatory profile in response to stress, and, more significantly, by a marked decrease in gene expression of the anti-inflammatory cytokine TGF-β1 (see Table 1).

The role of TGF-β1 in the expression of behavior has not been previously investigated. Serum TGF-β1 concentrations following acute stress positively predicted the behavioral propensity to affiliate with one's peers after presentation of a fearful stimulus. In this regard, the inflammatory profile of the VFD-reared macaques after stress is similar to that described for clinical depressive illness.

Primate Metabolic Syndrome

An additional line of investigation in the VFD subjects was based on the potential effects of persistent HPA axis disturbance. The metabolic syndrome, which comprises insulin resistance, visceral adiposity, dyslipidemia, and hypertension, may be induced by chronic HPA axis disturbance. We therefore hypothesized that VFD subjects would exhibit an accelerated onset of a primate form of the metabolic syndrome, which we have observed to occur spontaneously within our bonnet macaque colony. VFD-reared offspring at the juvenile phase of development had greater body mass index, abdominal circumference, and very low density lipoprotein (VLDL) lipids in comparison to controls. In addition, using a euglycemic clamp methodology, VFD subjects exhibited insulin resistance in comparison to controls. These data are relevant to the overall theme of a reduced brain neurotrophic state in VFD-reared subjects, inasmuch as reductions in brain insulin receptor activation by insulin-like growth factor are associated with reductions in neuroprotection and neurogenesis rates in the dentate gyrus of the hippocampus.

Table 1 Cytokine gene array response to restraint stress in adult VFD[a]

Proinflammatory genes which were upregulated >3-fold ($P < 0.05$) in VFD post stress
TNF-α
IL-1 β
IL-6
IL-13 receptor
Heat shock protein 90
Interleukin 7 receptor
Anti-inflammatory genes which were downregulated by >3-fold ($P < 0.05$) in VFD post stress
TGF-β1
TGF-β receptor II
Bone morphogenetic factor
IL-10

[a] Corresponding effects in non-VFD were not significant.
TNF, tumor necrosis factor; IL, interleukin; TGF, transforming growth factor.

VFD and White Matter Integrity

Diffusion tensor imaging (DTI), an application of magnetic resonance imaging (MRI), measures the extent and direction of water diffusion in neural tissue. Fractional anisotropy (FA) determines the magnitude of directionality of diffusion in brain white matter (WM), is greatest in tracts of parallel axons, and decreases in demyelinating conditions.

The rationale of performing DTI in VFD is based on the observation that humans and nonhuman primates exposed to early life stress exhibit deleteriously affected WM structures, such as the corpus callosum. One plausible mechanism whereby early life stress may affect WM integrity entails alterations in neurotrophic factor gene expression such as brain-derived neurotrophic factor (BDNF), which stimulates myelin basic protein production by differentiating oligodendrocytes. VFD exhibits long-term reductions in FA in the WM of anterior but not posterior internal capsule in comparison to control subjects. VFD also produces prominent alterations in patterns of FA concordance between anterior corpus callosum and anterior internal capsule. In control subjects, high FA correlation between anterior corpus callosum and anterior internal capsule conceivably suggests a common regulatory mechanism of myelin deposition related to neurotrophic factor gene expression. The corresponding correlations were significantly reduced in VFD subjects, indicating disruption of between-region FA homogeneity.

Integrity of frontal WM, as reflected by FA, may represent (1) a mechanism for affective disturbance via disruption of 'top-down governance' and (2) a potential indicator of neurotrophic status of brain regions implicated in human anxiety and mood disorders.

Interaction of Early-Life Stress and Serotonin Genotype on Infant Attachment Behavior

The serotonin transporter is a protein critical to the regulation of serotonin function in the brain, because it terminates the action of serotonin in the synapse via reuptake. The variant site is commonly known as 5-HTTLPR (serotonin transporter-linked promoter region). There are two common functional alleles of 5-HTTLPR, the short ('s') allele and the long ('l') allele. The s allele encodes an attenuated promoter

segment, and is associated with reduced transcription and functional capacity of the serotonin transporter relative to the l allele. Four independent investigations with child, adolescent, and young adult subjects reported associations between 5-HTTLPR variation and risk for depression following adverse life experiences. Specifically, the s allele of 5-HTTLPR was associated with the development of depression, but only in individuals with histories of child maltreatment or recent stressful life events. In the absence of these experiences, the s allele was not associated with an increased risk for depressive symptoms.

Serotonin transporter genotyping was available on infant male subjects who were studied while undergoing 16 weeks of the VFD procedure. During the final LFD/HFD cycle, repeated behavioral observations were performed that were designed to reflect the extent of infant attachment to its mother. Aggregate measures of infant proximity to mother were separately evaluated for the final LFD and final VFD period, assigning values from 1 to 5, where 1 indicated the greatest degree of infant proximity whereas 5 indicated the greatest magnitude of distance possible of infant from mother. Results indicated that, during the final LFD period, infants with the short allele of the 5-HTTLPR exhibited a significantly greater degree of proximity to mother than their long allele counterparts.

For the corresponding HFD period, no effect of genotype was noted. In conclusion, these preliminary data suggest an influence of serotonin transporter

Table 2 Summary table of biological consequences of VFD stress in comparison to experimental control subjects

Condition	*VFD effect*	*VFD implications*	*Cohort*
Oral yohimbine challenge	Behavioral hyper-responsivity	Noradrenergic dysregulation	Cohort 1 adult phase
Oral *m*-CPP challenge	Behavioral hyporesponsivity	Serotonergic dysregulation	
CSF tap under ketamine anesthesia	↑ CSF CRF (pg ml^{-1})	Central CRF activation in early VFD	Cohort 2 juvenile phase
	↓ CSF cortisol (ng ml^{-1})	HPA axis suppression in early VFD	
	↑ CSF somatostatin (pg ml^{-1})	Somatotropic axis suppression	
	↑ CSF 5-HIAA (ng ml^{-1})	Serotonin turnover ↑	
	↑ CSF HVA (ng ml^{-1})	Dopamine turnover ↑	
	↑ CSF CRF (pg ml^{-1})	Persistent central CRF activation	Cohort 2 adult phase
	↓ CSF CRF (pg ml^{-1})	Paradoxical CRF ↓ in late VFD	Cohort 3 juvenile phase
	↑ CSF 5-HIAA (ng ml^{-1})	Serotonin turnover ↑ consistent in early and late VFD	
Plasma sample under ketamine	↑ Plasma cortisol (mg dl^{-1})	Basal HPA axis activation in late VFD	
Mild novelty stress and plasma samples	↑ Plasma TGF-β1 (ng ml^{-1})	Compensatory anti-inflammatory Cytokine ↑	Cohort 3 juvenile phase
	↑ Plasma cortisol (mg dl^{-1})	HPA axis hyper-reactivity	
Acute restraint stress and monocyte gene microarray	↑ Gene expression of inflammatory, cytokines	Pro-inflammatory state under acute stress conditions	Cohort 3 adult phase
^{1}H-MRSI under ketamine anesthesia	↓ NAA/Cr in ACC	↓ Neuronal integrity in area critical for executive function	Cohort 2 and 3 adult phase
	↑ Glx/Cr in ACC	Hyperglutamatergic state (but could be GABA or glutamine)	
	↑ Cho/Cr in hippocampus	Abnormal metabolite profile in neurotrophic area	
	↓ NAA/Cho left hippocampus	Compromised neuronal integrity	
Morphometry	↑ Body mass index (kg m^{-2})	Primate metabolic syndrome	Cohort 4 juvenile phase
Ketamine and plasma samples	↑ VLDL (nmol l^{-1})	Dyslipidemia	
	↑ Glucagon-like protein 1 (pM)	Incretin compensation	
Ketamine and euglycemic clamp	↓ Glucose utilization (mg kg^{-1} min^{-1})	Insulin resistance	
DTI under ketamine	↓ FA Internal capsule anterior limb	Disruption of myelin integrity	

The table provides a comprehensive summary of the biological consequences of the VFD stress apparent in the VFD offspring. The first column provides a brief statement of the condition under which the findings were acquired; the second column summarizes the main finding observed in VFD subjects in comparison to experimental controls; the third column situates the finding within the context of its biological implications; and the fourth column lists the successive cohorts studied and the phase of development of subjects at the point of study.

VFD, variable foraging demand; *m*-CPP, *meta*-chlorophenylpiperazine; CSF, cerebrospinal fluid; CRF, corticotropin releasing factor; HPA axis, hypothalamic–pituitary–adrenal axis; 5-HIAA, 5-hydroxyindoleacetic acid; HVA, homovanillic acid; TGF-β1, tranforming growth factor β1; ^{1}H-MRSI, proton magnetic resonance spectroscopic imaging; NAA/Cr, *N*-acetyl-aspartate/creatine; 'Glx', glutamate, glutamine, and γ-aminobutyric acid (GABA); Cho, Choline; VLDL, very low density lipoprotein; DTI, diffusion tensor imaging.

genotype on infant behavior within the context of the VFD paradigm. Moreover, the findings were in the predicted direction, with VFD-reared infants possessing the s/s allele exhibiting greater degrees of 'reactive attachment', a putative measure of infant anxiety, than their l/l counterparts. Thus, the VFD-rearing experience appears to interact at a behavioral level with the allelic variant of the 5-HTT.

Conclusions

The VFD model has been extended from one providing only behavioral outcome measures, to a paradigm which demonstrates that early-life stress can significantly alter the following: (1) behavioral effects of pharmacological probes of monoamine systems, (2) stress neuropeptides and monoamine metabolite concentrations in CSF, (3) HPA axis measures, (4) neuroendocrine measures of the hypothalamic–pituitary–somatotropic, (5) neurochemical systems that are dependent on glutamatergic neurotransmission, (6) measures of cytokines and microarray studies of cytokine gene expression, (7) the occurrence of metabolic syndrome and insulin resistance, (8) regional neuroimaging data of metabolite profiles using magnetic resonance spectroscopy, and (9) measures of myelin integrity using DTI. Future studies can be interfaced with VFD genotyping to identify vulnerable versus resilient subjects. VFD-reared subjects have shown significant differences from non-VFD-reared control subjects (see Table 2) for literally all the biobehavioral measures that have been studied, suggesting that the impact of maternal stress upon the infant's developmental trajectory is persistent and widespread.

The latter observation alone serves to underscore the profound effect of a relatively nontraumatic early-life stressor on behavior and biology and serves to highlight the critical importance of emotional availabiltiy in the determining the integrity of the mother–infant experience. An additional consideration is that the abnormalities we have observed in the VFD offspring seldom occur in isolation but rather emerge within a constellation of biobehavioral abnormalities, begging the question of 'which came first'. Perhaps these findings are best conceptualized within the framework provided by McEwen, when discussing 'allostasis' as representing a chronic adapative homeostatic state in the face of chronic stress, leading to longstanding deviations in normative biology. The VFD model does have important clinical implications. It suggests that patients who have suffered early-life stress and have developed psychiatric or medical disorders are probably best conceptualized as having incurred damaging effects at multiple levels, both at the level of the central nervous system and in its interaction with the periphery.

See also: Social Interaction; Social Stress in Adult Primates; Stress Response: Sex Differences.

Further Reading

Andrews MW and Rosenblum LA (1994) The development of affiliative and agonistic social patterns in differentially reared monkeys. *Child Development* 65: 1398–1404.

Coplan JD, Altemus M, Mathew SJ, et al. (2005) Synchronized maternal–infant elevations of primate CSF CRF concentrations in response to variable foraging demand. *CNS Spectrum* 10: 530–536.

Coplan JD, Andrews MW, Rosenblum LA, et al. (1996) Persistent elevations of cerebrospinal fluid concentrations of corticotropin-releasing factor in adult nonhuman primates exposed to early-life stressors: Implications for the pathophysiology of mood and anxiety disorders. *Proceedings of the National Academy of Sciences of the United States of America* 93: 1619–1623.

Coplan JD, Rosenblum LA, and Gorman JM (1995) Primate models of anxiety. Longitudinal perspectives. *Psychiatric Clinics of North America* 18: 727–743.

Coplan JD, Smith EL, Altemus M, et al. (2001) Variable foraging demand rearing: Sustained elevations in cisternal cerebrospinal fluid corticotropin-releasing factor concentrations in adult primates. *Biological Psychiatry* 50: 200–204.

Coplan JD, Trost RC, Owens MJ, et al. (1998) Cerebrospinal fluid concentrations of somatostatin and biogenic amines in grown primates reared by mothers exposed to manipulated foraging conditions. *Archives of General Psychiatry* 55: 473–477.

Heim C, Plotsky PM, and Nemeroff CB (2004) Importance of studying the contributions of early adverse experience to neurobiological findings in depression. *Neuropsychopharmacology* 29: 641–648.

Mathew SJ, Shungu DC, Mao X, et al. (2003) A magnetic resonance spectroscopic imaging study of adult nonhuman primates exposed to early-life stressors. *Biological Psychiatry* 54: 727–735.

Meaney MJ and Szyf M (2005) Environmental programming of stress responses through DNA methylation: Life at the interface between a dynamic environment and a fixed genome. *Dialogues in Clinical Neuroscience* 7: 103–123.

Mirescu C, Peters JD, and Gould E (2004) Early life experience alters response of adult neurogenesis to stress. *Nature Neuroscience* 7: 841–846.

Nemeroff CC (2004) Early-life adversity, CRF dysregulation, and vulnerability to mood and anxiety disorders. *Psychopharmacology Bulletin* 38: 14–20.

Newport DJ, Stowe ZN, and Nemeroff CB (2002) Parental depression: Animal models of an adverse life event. *American Journal of Psychiatry* 159: 1265–1283.

Plotsky PM, Thrivikraman KV, Nemeroff CB, Caldji C, Sharma S, and Meaney MJ (2005) Long-term consequences of neonatal rearing on central corticotropin-releasing factor systems in adult male rat offspring. *Neuropsychopharmacology* 30: 2192–2204.

Reite M, Kaemingk K, and Boccia ML (1989) Maternal separation in bonnet monkey infants: Altered attachment and social support. *Child Development* 60: 473–480.

Rosenblum LA, Coplan JD, Friedman S, Bassoff T, Gorman JM, and Andrews MW (1994) Adverse early experiences affect noradrenergic and serotonergic functioning in adult primates. *Biological Psychiatry* 35: 221–227.

Motion Sickness

J R Lackner, Brandeis University, Waltham, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term 'motion sickness' was originally introduced to provide a nosological category for disorders elicited by exposure to unusual patterns of motion during passive transport, such as on boats and trains (causing seasickness and train sickness). The term came to subsume airsickness as well, and in recent years, simulator sickness, space sickness, and cybersickness (virtual environment sickness). It is now recognized that both space sickness and cybersickness are complex syndromes that include many sensorimotor and postural components in addition to motion sickness. Motion sickness is often equated with nausea and vomiting. However, the signs and symptoms of motion sickness are much broader, involving a wide range of autonomic nervous system activity, primarily of the sympathetic branch. Table 1 lists some of the more prominent features.

The pattern of motion sickness symptom development and expression depends on the nature of the exposure conditions and on an individual's inherent susceptibility. Nearly all persons with normal vestibular function are susceptible to varying degrees, although their susceptibility may differ in different background force levels (*g* levels) for the same pattern of body motion. With relatively unprovocative stimulation, such as gentle rocking, 'head signs' (e.g., headache and drowsiness) may predominate. With more provocative stimulation, such as making head movements during rotation, 'gut signs' (e.g., nausea and vomiting) are more prominent. Motion sickness can be difficult to identify under nonlaboratory conditions when it does not involve prominent pallor or nausea. For example, drowsiness is an early sign of motion sickness during exposure to relatively unprovocative motion stimulation. Individuals with total loss of vestibular function are generally immune to motion sickness and those with unilateral loss are less susceptible than normal persons. Recently, it has been reported that some individuals with total labyrinthine functional loss can be made motion sick by exposure to optokinetic stimulation, albeit their threshold is much higher than that of normal individuals.

Vestibular Function

The vestibular component of the inner ear consists of the semicircular canals and otolith organs. The receptor mechanisms of the three orthogonally oriented canals in each inner ear are influenced by angular acceleration of the head in their planes. They have high resting discharge rates that are increased by ampullopetal deflection and decreased by ampullofugal deflection. The change in discharge rate for natural voluntary head movements is proportional to head velocity. Pairs of canals in the same plane operate in a push–pull fashion and participate in the angular vestibuloocular reflex (VOR). Voluntary fixation of a target to suppress the VOR can result in motion sickness. Exposure to impulsive deceleration from constant velocity rotation leads to a pattern of afternystagmus which has a time constant longer than that of the cupula–endolymph system of the semicircular canals. This phenomenon is referred to as 'velocity storage' and reflects the fact that the oculomotor neurons driving the eyes receive a direct pathway signal from the canals, and an indirect one reflecting integration of the direct pathway signal. This integrated signal is also contributory to the sense of body rotation and displacement with respect to the external surroundings. Integration of the direct pathway signal is suppressed in weightlessness.

The otolith organs are influenced by inertial and gravitational accelerations and give rise to a linear VOR during translation of the body, and ocular counterrolling during tilt of the body relative to gravity. Vertical oscillation of the body alone can give rise to motion sickness and is an important factor in seasickness. The motion sickness response is frequency dependent and is greatest around 0.1–0.2 Hz.

The vestibular nuclei of the brain stem receive projections from the semicircular canals and otolith organs, but also visual, auditory, somatosensory, and proprioceptively related signals and a variety of other afferent signals, including from the cerebellum. They in turn have projections to the oculomotor nuclei, spinal cord, cerebellum, cerebral cortex, and other sites where they participate in a broad range of functions subserving postural and balance control, gaze stabilization, and perception of body orientation and motion. The circuitry is extremely complex, not fully identified, and operates largely automatically, although attentional factors can be contributory. Cortical projections of the labyrinth are to parietal cortex, which is involved in the representation of body orientation.

Modification of vestibular function occurs after injury – for example, half-sided loss of function, exposure to unusual background force level conditions, and during sensory 'rearrangements' that disrupt the compensatory effect of the VOR in maintaining target fixation during voluntary or passive head movements.

Table 1 Diagnostic categorization of different levels of severity of acute motion sickness

Category	*Pathognomonic 16 points*	*Major 8 points*	*Minor 4 points*	*Minimal 2 points*	*AQS*[a] *1 point*
Nausea syndrome	Vomiting or retching	Nausea[b] II, III	Nausea I	Epigastric discomfort	Epigastric awareness
Skin		Pallor III	Pallor II	Pallor I	Flushing
Cold sweating		III	II	I	
Increased salivation		III	II	I	
Drowsiness		III	II	I	
Pain					Headache
Central nervous system					Dizziness Eyes closed > II Eyes open III
Levels of severity identified by total points scored					
Frank sickness (S) >16 points	Severe malaise (M III) 8–15 points	Moderate malaise A (M IIA) 5–7 points	Moderate malaise B (M IIB) 3–4 points		Slight malaise (M I) 1–2 points

[a]AQS, Additional qualifying symptoms.
[b]III, severe or marked; II, moderate; I, slight.

Such modifications usually lead to motion sickness. Over time, however, adaptation generally occurs to restore normal or near-normal function. The extent and time course of adaptation are highly dependent on the exposure conditions, and adaptation is more complete with gradual, rather than abrupt, functional alterations. Such adaptive changes are an important component of the abatement of motion sickness with continued exposure to otherwise provocative situations.

Theories, Mechanisms, and Manifestations

At present, there is no fully adequate theory of motion sickness. Vestibular 'overstimulation' theories have largely been abandoned because, for example, low-frequency linear oscillation, even though it feels gentle, can be more provocative in eliciting symptoms than high-frequency, seemingly intense, stimulation. Variations on 'sensory conflict' theories are currently the most popular explanations of motion sickness. These theories relate the elicitation of sickness to 'conflicts' between the body motion the vestibular system is transducing and that which other senses (e.g., vision) are signaling, or between the motion the canals and otoliths are specifying about ongoing body orientation. Sensory conflict theories have great generality and applicability to circumstances eliciting motion sickness. Almost all such situations involve some form of visual–vestibular or canal–otolith or other form of sensory conflict. However, such theories lack predictive power and cannot identify why some conflicts are provocative and others not, nor why exposure to certain forms of motion, such as linear oscillation, are provocative primarily for specific frequencies, nor why some individuals are susceptible, and others are not, to particular conflicts.

Evolutionary theories of motion sickness emphasize its similarity to a poison response with two primary components, stomach emptying and a stress reaction. From this standpoint, motion sickness leads to emesis, which empties the stomach of ingested poisons and induces immobility, which would decrease the rate of blood circulation of toxins. Dogs have been found to show a profound decrease in susceptibility to many (but not all) emetic drugs after bilateral labyrinthectomy. (Dogs are commonly used in animal studies of motion sickness because of the similarity of their response to motion to that of humans.) Most of the autonomic and endocrine changes occurring during motion sickness do accompany stress responses as well. Removal of the nodulus and uvula, the vestibular projection areas of the cerebellum, renders monkeys insusceptible to motion sickness. It is yet to be determined whether such lesions also eliminate the vomiting response to emetic drugs. How motion sickness could develop evolutionarily to mimic or constitute a poison response remains unclear. The time course of emesis evoked by motion sickness is generally too slow to be of functional significance for ejecting toxins.

Emesis is achieved by the coordinated activity of both smooth and somatic muscles to generate appropriate changes in intra-abdominal and intrathoracic pressures, and opening of the esophageal sphincters. The central nervous system coordinating mechanisms are complex and only recently have become well understood. In severe motion sickness involving multiple bouts of vomiting, alkalosis may develop

because of hydrogen ion loss and may lead to increased renal excretion of potassium bicarbonate, producing a potassium deficiency that can cause muscle weakness, constipation, and cardiac arrhythmias. Sodium loss can lead to hypotension. Antidiuretic hormone release is also increased. The presence of a functioning vestibular system is not necessary for vomiting to be produced, and emesis can be evoked by activation of a variety of peripheral and central afferent mechanisms. Earlier notions of a well-defined 'vomiting center' in the brain stem have not been supported by neurophysiological and neuroanatomical studies, which point instead to a more broadly distributed neuronal representation in the medulla oblongata. The input circuitry controlling emesis is plastic, allowing changes in the sensitivities of other pathways when one pathway is interrupted.

Predicting Susceptibility to Motion Sickness

Motion sickness has been a persistent problem for the military, with air, sea, and land vehicle sickness being ubiquitous operational concerns. Consequently, there have been many attempts to develop both questionnaires and experimental tests of susceptibility to screen for motion sickness. Comprehensive questionnaires developed to screen individuals for susceptibility show that persons reporting multiple experiences of motion sickness are more likely to experience sickness when tested in experimental devices than are individuals without a history of susceptibility. Women more often than men report a past history of motion sickness, and women are also more likely to report having used anti-motion-sickness drugs. Individuals who are classified as highly susceptible on the basis of their questionnaire answers are more likely to have highly susceptible parents or siblings, which may be partially attributable to similar exposure histories. Motion sickness susceptibility may in part be genetically determined, and Asians have been found to be more susceptible than Caucasians.

Motion sickness questionnaires provide a means of identifying that small fraction, about 10% of the population, that is highly susceptible to motion sickness. However, susceptible subjects tend to avoid situations that are likely to elicit motion sickness (e.g., carnival rides or boats). Disadvantages of questionnaires are that it is not possible to determine quantitatively the individual's prior exposure to provocative stimulation, and that most people equate motion sickness with nausea and vomiting, and there are many other symptoms that may accompany motion sickness (e.g., pallor, changes in salivation, sweating, drowsiness, stomach awareness, and stomach discomfort). When a motion sickness questionnaire is being used as a selection or screening device, candidates may not answer truthfully; moreover, questionnaires can provide information about past episodes of motion sickness, but not about the person's ability to adapt to provocative motion or about his or her rate of recovery. Many questionnaires are long and cumbersome and difficult to fill out. However, Golding has developed a short questionnaire that is easy to complete and that has been validated in laboratory assessments of susceptibility. It is likely to serve as the standard in the field of motion sickness studies.

Gender Differences in Susceptibility

Women typically report a greater incidence of motion sickness on questionnaires than do men. It has been questioned whether such reports reflect actual susceptibility differences or, instead, differences in willingness to talk about illness; this is unclear, as is whether in the general population males may have had more experiences in stressful motion environments. Some experimental studies contrasting male and female susceptibility have not found any significant differences. Other studies have found that women are somewhat more susceptible than men are, and that women also vary in their susceptibility, being more susceptible during their menstrual cycle. Other studies have failed to find susceptibility variations related to the female menstrual cycle. An excellent analysis of these issues is provided by Golding.

Laboratory Tests of Motion Sickness Susceptibility

Predicting susceptibility to motion sickness in operational environments from laboratory procedures is very difficult. Susceptibility in one exposure condition may not generalize to another. Although the vestibular system is usually essential for motion sickness, susceptibility is not related to vestibular sensitivity *per se*. Motion sickness can be induced, for example, by motions that stimulate separately or in combination the semicircular canals and otolith organs of the inner ear, moving patterns of visual stimuli, or by the voluntary suppression of reflexive eye movements. Artificial alteration of the effective gain of the vestibuloocular reflex by means of lens systems readily induces symptoms until adaptive compensations are made. Head movements made during passive body rotation produce a bizarre pattern of stimulation of the canals and otolith organs; this pattern, known as Coriolis cross-coupling stimulation, is highly provocative under ground-based conditions, only mildly provocative under weightless conditions,and extraordinarily provocative under conditions of increased background force level (e.g., high-*g* conditions). In weightless conditions, velocity storage

of semicircular canal signals is attenuated, reducing the time constant of vestibular nystagmus and reducing the perception of body angular displacement. As a consequence, head movements out of the axis of passive body rotation are less provocative in weightless environments than in high-force environments.

Almost all nonlaboratory situations that are provocative involve more than one etiological or predisposing factor. Individuals often vary in their susceptibility to these different forms of stimulation. This largely accounts for why it is difficult to generalize from laboratory assessments of susceptibility involving one or a few factors to more complicated, natural exposure situations, which can involve both passive and active head and body movements and voluntary locomotion. However, persons who are extremely sensitive to one type of motion exposure (e.g., angular acceleration) tend to be very sensitive to others as well. The highly susceptible constitute about 5–10% of the population. Golding has raised the possibility that there may be a general motion sickness susceptibility factor as well as specific factors related to susceptibility for particular types of provocative motion. This perspective helps explain why there can be such variability in susceptibility across sometimes seemingly similar exposure conditions.

It has been especially difficult to predict susceptibility to space motion sickness. In spaceflight, in addition to 'lifting' of the tonic 1 *g* acceleration of earth gravity (owing to the free-fall conditions), the entire pattern of sensorimotor control and support of the body is changed. Such altered sensorimotor demands during head and body movement are provocative in themselves when mimicked under terrestrial conditions. It is not surprising, therefore, that reports of motion sickness are also commonplace with the use of virtual reality devices, which can produce alterations in visuomotor control and apparent self-motion. Such cybersickness can be as disruptive during exposure to virtual environments, as space sickness is in space flight. The 'sopite syndrome' is an important aspect of space motion sickness and of cybersickness (from the Greek *sopor*, for 'drowsiness'). It refers to the tiredness, mood swings, lack of initiative, and apathy that can develop and persist during prolonged exposure to relatively unprovocative situations. Under some conditions, and for some individuals, the sopite syndrome can be the primary or sole expression of motion sickness. Its presence can be difficult to identify in nonlaboratory settings.

Preventing and Treating Motion Sickness

Prevention of motion sickness and recovery from motion sickness can be complex. A small percentage of normal individuals are highly susceptible for nearly all exposure situations, a small percentage are highly insusceptible, and most are in between. The best prevention for the highly susceptible is avoidance. For other individuals, spaced exposure of short duration can lead to a buildup of adaptation to the provocative situation. This is especially effective if only minor symptoms of motion sickness are allowed to develop before terminating each exposure period. Alternatively, incremental exposure to gradually increasing levels of provocative stimulation (e.g., making head movements during exposure to passive body rotation at higher and higher rotation velocities) can allow adaptation to be achieved without motion sickness being elicited, even at stressor levels that, if achieved in a single step, would be intolerably provocative.

Behavior modification techniques, such as autogenic feedback, have been promoted as a way of preventing motion sickness. In autogenic feedback training, the individual is trained to detect changes in heart rate, respiration, skin temperature, and other parameters and learns how to prevent alterations in these parameters. The notion is that preventing the physiological changes associated with motion sickness will prevent the development of motion sickness. One comprehensive study, with proper controls, evaluated the efficacy of autogenic feedback training. Test participants were initially required to make head movements during passive body rotation to determine baseline susceptibility. They were then divided into three groups, matched for susceptibility. One group received autogenic training with accurate feedback about their autonomic nervous system responses, one control group received false feedback, and another control group received no treatment. Autogenic feedback training did not enable test group members to withstand exposure to provocative stimulation any better than was seen in the two control groups. Acupressure devices, worn on the wrist, are available in various forms commercially, and, like autogenic feedback, are proposed as a way of preventing motion sickness. There is no scientific evidence to support such claims. Such devices may, however, have a placebo effect in individuals who are convinced that they work. Placebo effects are commonplace in experimental studies of anti-motion-sickness drug efficacy, in which placebo controls are essential.

Some studies have reported a close correlation between tachygastria – electrogastrogram activity elevated above the normal 3 cycles per minute (cpm) rate, to 6–12 cpm – and motion sickness, raising the possibility that preventing tachygastria could suppress motion sickness. Controlled studies have shown, however, that tachygastria can occur before, during, or after symptoms have developed or even

abated. Tachygastria, in fact, appears to be most correlated with anxiety and apprehension. Importantly, it has been found that individuals remain abnormally sensitive to provocative motion for 3–4 h after all overt symptoms seem to have abated following provocative motion testing.

A variety of drugs are available to reduce susceptibility to motion sickness. Over-the-counter drugs with some efficacy include dimenhydrinate (Dramamine), meclizine (Bonine), and cyclizine (Marezine). The prescription transderm patch can provide protection for 24 h or more. The most effective anti-motion-sickness drugs, including the relatively weak ones mentioned here, have adverse side effects, generally including drowsiness and dry mouth. More potent drugs and drug combinations include scopolamine (often used in association with amphetamine, 'Scopdex,' to counteract sleepiness) and promethazine (often with ephedrine). In cases of fully developed, severe motion sickness, intramuscular injections of scopolamine or promethazine may be necessary to provide relief. Effective antiemetic drugs have anticholinergic, antihistaminic, or sympathomimetic properties; however, not all of the central sites of drug action are understood, and these drugs tend to be relatively unselective in their effects. New families of receptor subtypes are being discovered that open the possibility of developing anti-motion-sickness drugs that block only certain types of neurotransmitter binding sites, thereby leading to fewer undesirable side effects.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Nervous System; Balance and Posture Control; Canal–Otolith Interactions; Gravitational Effects on Brain and Behavior; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vestibular System; Vestibulo-Autonomic Responses; Vestibulo-Ocular Reflex; Visual–Vestibular Interactions.

Further Reading

Anson BG, Harper DB, and Winch TG (1968) *Third Symposium on the Role of the Vestibular Organs in Space Exploration*. NASA SP-152., Houston, TX. Washington, DC: US Government Printing Office.

Cheung B and Vaitkus P (1998) Perspectives of electrogastrography and motion sickness. *Brain Research Bulletin* 47(5): 421–431.

Cohen B, Matsuo V, and Raphan T (1977) Quantitative analysis of the velocity characteristics of optokinetic nystagmus and optokinetic afternystagmus. *Journal of Physiology* 270: 321–344.

Cowings PS and Toscano WB (1982) The relationship of motion sickness susceptibility to learned autonomic control for symptom suppression. *Aviation Space and Environmental Medicine* 53(6): 570–575.

DiZio P and Lackner JR (2002) Proprioceptive adaptation and after-effects. In: Stanney K (ed.) *Handbook of Virtual Environments*, pp. 751–771. New York: Lawrence Erlbaum.

Golding JF (2006) Predicting individual differences in motion sickness susceptibility by questionnaire. *Personality and Individual Differences* 41: 237–248.

Golding JF, Mueller AG, and Gresty MA (2001) A motion sickness maximum around the 0.2 Hz frequency range of horizontal translational oscillation. *Aviation Space and Environmental Medicine* 72(3): 188–192.

Graybiel A and Knepton J (1976) Sopite syndrome: A sometimes sole manifestation of motion sickness. *Aviation Space and Environmental Medicine* 47: 873–882.

Johnson WH, Sunahara FA, and Landolt JP (1999) Importance of the vestibular system in visually induced nausea and self-vection. *Journal of Vestibular Research* 9(2): 83–87.

Jozsvai EE and Pigeau RA (1996) The effect of autogenic training and biofeedback on motion sickness tolerance. *Aviation Space and Environmental Medicine* 67(10): 963–968.

Lackner JR and DiZio P (2006) Space motion sickness. *Experimental Brain Research* 175: 377–399.

Lackner JR, Graybiel A, Igarashi M, et al. (1985) Head movements elicit motion sickness during exposure to microgravity and macrogravity acceleration levels. In: *Proceedings of the VII International Symposium: Vestibular and Visual Control on Posture and Locomotor Equilibrium*, pp. 170–176. Basel: Karger.

Money KE, Lackner JR, and Cheung RSK (1996) The autonomic nervous system and motion sickness. In: Crampton GH (ed.) *Motion and Space Motion Sickness*, pp. 147–173. Boca Raton, FL: CRC Press.

Reason JT and Brand JJ (1975) *Motion Sickness*. New York: Academic Press.

Stern RM, Koch KL, Stewart WR, et al. (1987) Spectral analysis of tachygastria recorded during motion sickness. *Gastroenterology* 92: 92–97.

Yates BJ and Bronstein AM (2005) The effects of vestibular system lesions on autonomic regulation: Observations, mechanisms, and clinical implications. *Journal of Vestibular Research* 15(3): 119–129.

Yates BJ, Miller AD, and Lucot JB (1998) Physiological basis and pharmacology of motion sickness: An update. *Brain Research Bulletin* 47(5): 395–406.

Motor Autonomic Transmission

V Ralevic, Queen's Medical Centre, Nottingham, UK

© 2009 Elsevier Ltd. All rights reserved.

Distribution and Neurotransmitter Content of Sensorimotor Nerves

Sensorimotor nerves are a subpopulation of nociceptive primary afferent nerves, which have a dual sensory (afferent) and motor (efferent) function (Figure 1). They comprise predominantly unmyelinated, slowly conducting C-fibers with small-diameter cell bodies, together with a smaller population of myelinated, fast-conducting Aδ-fibers with large-diameter cell bodies. Their cell bodies lie in dorsal root, trigeminal, and nodose ganglia. Each sensorimotor nerve has a single axon which divides at the cell body into a centrally projecting and peripherally projecting branch. The central branch forms synapses in the central nervous system (CNS) with second-order sensory neurons. The peripheral branch, which may have a number of branches or collaterals, terminates in its target organs and tissues in the periphery. In the periphery, there are no specialized synapses between the nerve and the target tissue; instead, neurotransmitters are released from specialized swellings of the axon known as varicosities.

In the CNS, glutamate, aspartate, and adenosine 5′-triphosphate (ATP) are the principal neurotransmitters of sensorimotor nerves, being released along with other neurotransmitters, including the peptides calcitonin gene-related peptide (CGRP), substance P (SP), and neurokinin A (NKA). In the periphery, the archetypal sensorimotor neurotransmitters are CGRP, SP, and NKA. The peptides are synthesized in cell bodies in sensory ganglia and transported to the periphery. There is considerable overlap in the expression of CGRP and SP, consistent with their being cotransmitters in sensorimotor nerves, although both neurotransmitters are not always coexpressed and/or coreleased. There is evidence that a number of other peptides and nonpeptides (e.g., vasoactive intestinal polypeptide, ATP, and bombesin) are cotransmitters in the peripheral terminals of sensorimotor nerves.

Classically, the distribution of sensorimotor nerves in tissues has been studied by mapping immunoreactivity for sensory neuropeptides (mainly CGRP and SP) and, importantly, by showing that the immunoreactivity is not present in experimental animals treated with capsaicin to destroy selectively C- and Aδ-fibers. Functional studies examining tissue responses to stimulation of sensorimotor nerves and measurements of neurotransmitter release (mainly CGRP and SP) have confirmed the widespread distribution of sensorimotor nerves in peripheral tissues. Collectively, these provide evidence for sensorimotor innervation of a variety of organs and tissues, including the heart, blood vessels (including those of skin), smooth muscle of the bladder and airways, skeletal muscle, and gastrointestinal tract.

Sensory neuropeptides may be differentially expressed in the peripheral endings of C- and Aδ-fibers. For example, in trigeminal sensorimotor nerves innervating cranial blood vessels, SP is contained predominantly in C-fibers, whereas CGRP is contained in both C- and Aδ-fibers. In general, CGRP mediates vasodilatation and SP mediates plasma extravasation. Accordingly, low-intensity, brief electrical stimulation, which recruits only trigeminal Aδ sensorimotor nerves, causes CGRP release and neurogenic vasodilatation, whereas only the additional activation of C-fibers, by high-intensity sustained electrical stimulation, leads to a release of both SP and CGRP from activated C-fibers, which in turn causes plasma extravasation as well as vasodilatation (neurogenic inflammation). Similarly, there is evidence in skin that Aδ-fibers are involved in vasodilatation but not plasma extravasation, indicating that these Aδ-fibers also contain predominantly CGRP and not SP.

The use of selective antagonists has shown a differential role of various sensorimotor neurotransmitters in different tissues. For example, the CGRP receptor antagonist CGRP[8–37] blocks sensorimotor neurogenic vasodilatation in mesenteric arteries, whereas the NK1 selective antagonist CP 96 345 blocks sensorimotor neurogenic contraction in the iris (indicating involvement of SP/NKA).

Activation of Sensorimotor Nerves

Sensorimotor nerves are a subpopulation of nociceptive primary afferent nerves (not all nociceptive primary afferents have an efferent/motor function). The majority are polymodal nociceptors, being activated by a variety of stimuli including noxious levels of pressure, heat and pH, and a variety of agents (classically by the vanilloid capsaicin), acting through specific cell surface receptors. When the peripheral terminals of sensorimotor nerves are activated by one of these noxious stimuli, there is a release of neurotransmitter at the central terminals, which elicits pain and triggers avoidance/protective reflexes. Concomitantly, neurotransmitter is released from the peripheral terminals of sensorimotor nerves in an efferent or motor function, which leads, in general, to the promotion of tissue recovery and repair.

The peripheral release of sensory neurotransmitter can take place through different mechanisms

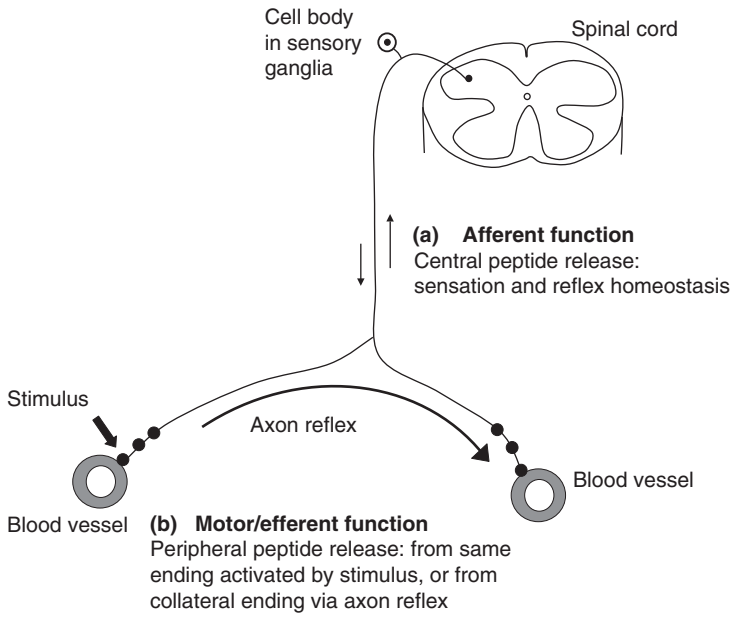

Figure 1 Dual role of sensorimotor nerves. As their name suggests, sensorimotor nerves have a dual function, being able to convey information about a stimulus to the central nervous system and being able to release neurotransmitter (typically neuropeptides) from their peripheral endings to elicit a motor or efferent function. The motor function can be elicited by three mechanisms: direct release of neurotransmitter from the same terminal that is activated by the stimulus, release of neurotransmitter from a collateral ending following antidromic conveyance of the impulse in the axon reflex, and electrical stimulation of the nerve trunk to cause antidromic activation (not shown). Adapted from Holzer P (1992) Peptidergic sensory neurons in the control of vascular functions: Mechanisms and significance in the cutaneous and splanchnic vascular beds. *Reviews of Physiology, Biochemistry, and Pharmacology* 121: 49–146.

(Figure 1). There may be a direct and local release of sensory neuropeptides from the same peripheral terminals that are activated by the stimulus. Because this does not involve axonal conveyance of impulses, it is insensitive to tetrodotoxin blockade of voltage-gated sodium channels. Another mechanism involves a release of neuropeptides at varying distances from the stimulus site that is due to antidromic conveyance of the impulse to collateral terminals (axon reflex) and accordingly is tetrodotoxin sensitive. Its physiological manifestation can be seen in the flare response in skin, a spread of increased blood flow some distance outward from the stimulus, which can be blocked by capsaicin defunctionalization of sensorimotor nerves. The efferent function of sensorimotor nerves can also be elicited by antidromic activation of the nerve trunk by electrical stimulation. This effect is not mimicked by periaxonal administration of capsaicin (which locally defunctionalizes the nerves) and is sensitive to blockade by tetrodotoxin.

Sensorimotor nerves express a number of different cell surface receptors through which both their afferent and efferent functions are activated. The receptor which has undoubtedly received the most interest, because of its key role in nociceptive signaling, is the vanilloid type 1 receptor (or TRPV1), which is part of the transient receptor potential (TRP) superfamily of ion channels. The other TRPV family members are thermosensitive nonselective cation channels (TRPV1, TRPV2, TRPV3, and TRPV4) and calcium selective channels involved in calcium absorption and reabsorption across epithelia (TRPV5 and TRPV6). Only TRPV1 is activated by capsaicin, a vanilloid and naturally occurring pungent compound found in hot peppers, and by the structurally related plant-derived compound resiniferatoxin. Because TRPV1 receptors are not expressed by most other autonomic neurons, motorneurons, and central neurons, this confers on capsaicin a remarkable (but not exclusive) selectivity for primary afferent C- and Aδ-fibers and has led to the widespread use of this agent in the study of the physiology and pharmacology of sensorimotor nerves. TRPA1 is a novel channel expressed on sensorimotor nerves that is activated by tetrahydrocannabinol and sulfur compounds derived from mustard oil and garlic, and it mediates pain and vasodilatation.

Capsaicin and other stimuli of TRPV1 act at TRPV1 receptors to cause depolarization, calcium influx, and consequent neurotransmitter release from sensorimotor nerves. Following activation, capsaicin can block sensorimotor nerve conduction and neurotransmitter release through a combination of effects including TRPV1 desensitization and blockade of voltage-gated Ca^{2+} channels. Used chronically, capsaicin is neurotoxic. These effects have been widely utilized in the identification and characterization of sensorimotor

nerves. Blockers of TRPV1, including ruthenium red, 5′-iodoresiniferatoxin, 6-iodonordihydrocapsaicin, and capsazepine, have also been useful in characterizing TRPV1 responses and the properties of sensorimotor nerves.

TRPV1 receptors act as polymodal nociceptors and confer on sensorimotor nerves their sensitivity (their afferent and efferent responsiveness) to specific noxious stimuli. In addition to capsaicin, TRPV1 receptors, and hence sensorimotor nerves, are activated by noxious heat (threshold of approximately 43 °C), acidosis (pH < 6) (which may occur upon ischemia, inflammation, muscle exercise, or damage to the stomach mucosal barrier), resiniferatoxin, phenylacetylrivanil, olivanil, allicin, camphor, some eicosanoids (e.g., 12-HPETE, 15-HPETE, 5-HETE, and leukotriene B_4), and certain endocannabinoids (anandamide, *N*-arachidonoyl-dopamine, and *N*-oleoyl-dopamine). TRPV1 receptors can be sensitized by a variety of mechanisms, including phospholipase C-mediated hydrolysis of PIP_2; PKA-mediated phosphorylation, which prevents rapid desensitization of TRPV1; PKB-mediated phosphorylation, which enhances the probability of TRPV1 channel gating; and mild acidosis (pH 6–7), which does not itself activate the receptor but lowers the threshold for TRPV1 activation by heat.

Sensorimotor nerves express a number of excitatory receptors in addition to TRPV1. 5-Hydroxy tryptamine (5-HT), which is released from activated platelets and mast cells, acts via 5-HT_3 receptors to activate sensorimotor nerves, and 5-HT_3 receptor antagonists are anti-inflammatory, analgesic, and block 5-HT-evoked reflexes. Histamine, released from mast cells, acts principally via H_1 receptors to activate sensorimotor nerves. Bradykinin is a locally produced blood plasma protein which has a number of proinflammatory effects including stimulation of sensorimotor nerves, leading to pain and reflex responses (largely through B_2 receptors), and CGRP and tachykinin release, leading to motor effects including plasma protein extravasation in skin and airways and to positive chrono- and inotropic effects in isolated atria. Activation of TRPV1 can cause sensorimotor nerves to become unresponsive to the previously mentioned stimuli.

Arachidonic acid metabolites are key players in inflammation because they are produced by other inflammatory mediators and can activate sensory nerves both via TRPV1 (12-HPETE and 15-HPETE) and independently of TRPV1 (e.g., prostacyclin and lipoxin A_4). They can also potentiate, by causing sensitization, both the sensory and the efferent functions of sensorimotor nerves. Platelet-activating factor is a lipid mediator of inflammation that activates sensorimotor nerves. Sensorimotor nerves express nicotinic receptors that mediate a release of tachykinins and/or CGRP, which accounts for the motor effects of acetylcholine in a variety of tissues, including heart, bronchus, and lung. Other substances shown to activate sensorimotor nerves include *N*-formyl-methionyl-leucyl-phenylalanine (FMLP), interleukin-1, neurotensin, nitric oxide, γ-aminobutyric acid (GABA; via $GABA_A$), adenosine (via A_2 receptors), ouabain, and Ca^{2+}.

ATP can be released from many different types of cells during normal processes and tissue injury, and there is some evidence that it can be released as a neurotransmitter from both the central and the peripheral terminals of some sensorimotor nerves. ATP and its analog cause depolarizing currents in dorsal root ganglion (DRG) and trigeminal neurons, an action mediated via $P2X_3$ homomeric receptors and/or $P2X_{2/3}$ heteromeric receptors. In DRG cells, the depolarization can lead to glutamate and peptide release. ATP also excites the peripheral terminals of C- and Aδ-nociceptive afferents and elicits pain responses *in vivo*. However, ATP has not been shown to be effective at evoking an efferent function at the peripheral terminals of sensorimotor nerves. Indeed, a lack of involvement of capsaicin-sensitive sensorimotor nerves has been shown in functional responses to ATP and α,β-methylene ATP (a P2X-selective ligand) in mesenteric arteries. Similarly, in atria ATP does not activate the efferent function of sensorimotor nerves but, rather, inhibits, via adenosine P1 receptors, the positive inotropic efferent function caused by transmural stimulation of sensorimotor nerves.

Inhibition of Sensorimotor Nerves

A number of agents inhibit the release of sensorimotor neurotransmitter at the peripheral and central terminals. Inhibitory neuromodulators include opioids, noradrenaline (via α_2-adrenoceptors), neuropeptide Y (NPY) (via Y_2 receptors), adrenomedullin, galanin, corticotrophin-releasing factor, and somatostatin. Adenosine acts via A_1 and A_{2A} receptors found centrally, and A_1 receptors found peripherally, to inhibit pain and sensorimotor neurotransmitter release. An adenosine receptor-independent inhibitory modulation through direct binding of adenosine and its analogues to TRPV1 has been described. Thus, adenosine, which can be formed by ectoenzymatic metabolism of ATP, provides a potential brake for the P2X receptor-mediated excitatory actions of ATP on sensorimotor neurotransmission. An important source of ATP, noradrenaline, and NPY is sympathetic nerves, from which they are released as cotransmitters, suggesting

that interactions between sympathetic and sensorimotor nerves may occur.

Some substances that are able to inhibit prejunctionally sensorimotor neurotransmission also have stimulatory effects (described previously), which is explained by actions at their different receptor subtypes. Prejunctional inhibitory effects of histamine (via H_3 receptors), 5-HT (via 5-HT_1-like receptors), and GABA (acting via $GABA_B$ receptors) on sensorimotor neurotransmission have been shown. Whether a substance excites or inhibits sensorimotor neurotransmission may depend on the relative affinities of the different receptor subtypes involved.

Certain cannabinoids can elicit both excitatory and inhibitory effects on sensorimotor nerves. Anandamide and methanandamide inhibit capsaicin-evoked plasma extravasation in rat skin via CB_1 receptors. These cannabinoids are also ligands at TRPV1, and virtually all TRPV1-expressing DRG neurons express CB_1. Thus, cannabinoids are able to mediate opposite actions at sensorimotor nerves, namely excitation via CB_1 and inhibition via TRPV1. In rat skin and DRG neurons, anandamide is more potent at CB_1 than at TRPV1 receptors, so low doses of anandamide inhibit neurotransmission, whereas higher doses are excitatory in a TRPV1-dependent manner, the latter possibly involved in sensorimotor nerve excitation in pathophysiological conditions. Similar potencies of anandamide at recombinant CB_1 and TRPV1 receptors suggest that other more subtle forms of differentiation might exist, perhaps involving relative rates of activation and/or inactivation of the receptors or differential modulation by factors such as pH and second messengers. Inhibitory effects of cannabinoids on sensorimotor neurotransmission in mesenteric arteries appear to be mediated via CB_1 and non-CB_1, non-CB_2 receptors.

Prolonged activation of TRPV1, classically by capsaicin, inhibits prejunctionally sensorimotor neurotransmission through a combination of effects, including TRPV1 desensitization and blockade of voltage-gated Ca^{2+} channels.

Inhibition of sensorimotor neurotransmission can be achieved through the use of selective antagonists to block the postjunctional receptor and, therefore, the actions of the released sensory neurotransmitter on its target tissue. This has provided important information about the roles of sensorimotor nerves and of the identity of the sensorimotor neurotransmitter(s) involved in the motor response.

Roles of Sensorimotor Nerves

The primary role of sensorimotor nerves appears to be as a system of defense acting against challenges to homeostasis. They are ideally designed for this purpose because their sensory role allows them to convey information to the central nervous system about the stimulus and to elicit reflexes to take evasive action if necessary, whereas the motor component promotes tissue recovery and repair. Many of the stimuli activating sensorimotor nerves are mediators of inflammation, which accounts for the involvement of sensorimotor nerves in the inflammatory response ('neurogenic inflammation'). This was first described in skin, and in this process sensorimotor nerves both activate and are activated by cells of the immune system through the actions of their neurotransmitters/mediators. However, there is evidence for wider roles of sensorimotor nerves under different physiological as well as pathophysiological conditions (**Figure 2**). A dissociation of the pain and cardiovascular effects of sensorimotor nerves has been reported, in that cardiovascular effects can be elicited with stimuli that are subthreshold for pain stimulation. This suggests that sensorimotor nerves may operate at different levels according to the duration or intensity of stimulation, with an efferent-only function operating under physiological conditions and a dual sensorimotor function under pathophysiological conditions. In addition, sensorimotor nerves in some mesenteric arteries have synaptic contacts with enteric neurons which allow them to respond to physiological stimuli (cholycystokinin, bile, and distension) of the intestine. Finally, there may be an ongoing release of sensorimotor neurotransmitter that is involved in trophic effects.

In neurogenic inflammation, sensorimotor nerves convey information about the inflammatory response to the central nervous system and promote recovery and tissue repair by causing arteriolar vasodilatation, an increase in vascular permeability, degranulation of mast cells, and effects on leukocytes and fibroblasts. The arteriolar vasodilatation is caused mainly by sensory nerve CGRP, a potent vasodilator, and the increase in blood flow promotes the delivery of phagocytic cells to the affected area. Ultrastructural evidence suggests that a close relationship exists between sensorimotor nerves and mast cells, and that a bidirectional functional interaction exists between them. SP released from sensorimotor nerves causes a release of histamine, 5-HT, prostaglandins, and nitric oxide from mast cells. The increase in vascular permeability that occurs is caused by SP and other tachykinins released from sensorimotor nerves, as well as by mast cell histamine. This allows antimicrobial chemicals and clotting proteins to seep into the interstitial space of the damaged tissue ('plasma protein extravasation'). Histamine, acting via H_3 receptors on sensorimotor nerves, can act as

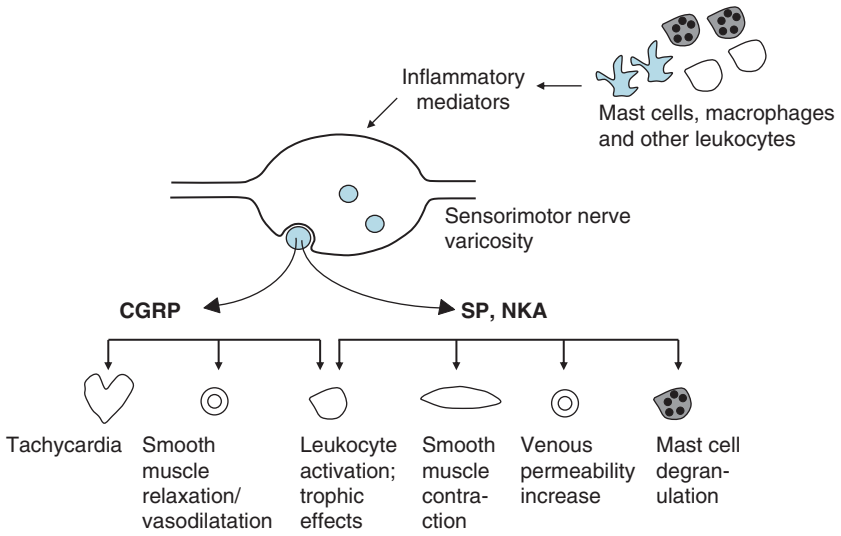

Figure 2 Motor/efferent function of sensory nerves. Calcitonin gene-related peptide (CGRP), substance P (SP), and neurokinin A (NKA) are coreleased from stimulated sensorimotor nerves to elicit a variety of different effects on the target tissue innervated. There are a number of different stimuli of sensorimotor nerves, including inflammatory mediators released from mast cells, macrophages, and other leukocytes. Following stimulation of sensorimotor nerves, the released sensory neurotransmitters evoke motor responses which lead to the promotion of tissue recovery and repair and also restoration of homeostasis.

part of an inhibitory feedback loop on sensorimotor neurotransmission. However, the functional interactions between mast cell mediators and sensory nerves are complex because prostaglandins, for example, can sensitize sensory nerves. Interactions can also take place between sensorimotor neurotransmitters. For example, although CGRP does not induce an increase in vascular permeability, it facilitates the SP-induced increase in permeability. Both CGRP and SP have effects on monocytes/macrophages, including a release of cytokines, prostaglandins, leukotrienes, and thromboxanes, which in turn affect vessel diameter, permeability, and sensorimotor nerve excitability (typically causing sensitization).

The study of sensorimotor nerves in a range of blood vessels outside of skin has allowed a fuller understanding of their roles in the cardiovascular system. This has confirmed that sensorimotor neurogenic vasodilatation is mediated principally by CGRP. Accordingly, electrical stimulation of nonadrenergic, noncholinergic nerves in certain blood vessels produces a long-lasting vasorelaxation that is abolished by capsaicin, blocked by CGRP receptor antagonism, and mimicked by exogenous CGRP (and capsaicin). Moreover, CGRP release can be detected during vasodilatation due to electrical stimulation of capsaicin-sensitive nonadrenergic, noncholinergic perivascular nerves. SP and other tachykinins do not appear to have a role in sensorimotor vasodilatation. In many tissues and organs, parasympathetic nerves provide a functional counterpart to the actions of sympathetic nerves. Most blood vessels, however, are not innervated by parasympathetic nerves. It seems as though in this case vasodilator sensorimotor nerves provide a neurogenic counterpart for vasocontractile sympathetic nerves. A decrease in vascular sensorimotor innervation and motor function (nonadrenergic, noncholinergic vasodilatation) has been shown in animal models of aging and hypertension.

In the heart, capsaicin is a potent cardiostimulant mediating positive inotropic and chronotropic effects due to activation of sensory nerves. CGRP, SP, and NKA are coreleased from cardiac sensory nerves, and CGRP mimics the actions of capsaicin. Cardiac sensorimotor nerves are activated by several mediators of inflammation and ischemia.

Activation of reflexes by sensorimotor nerves has been shown *in vivo*. In experimental models, administration of capsaicin into the hindlimb induces a fall in blood pressure and an increase in ventilation that is blocked by capsaicin pretreatment and the vanilloid antagonist capsazepine, indicating an action at TRPV1 on sensorimotor nerves. The cardiovascular and respiratory responses to capsaicin were blocked by sectioning of the femoral and sciatic nerves, indicating the involvement of a reflex action. Additional local effects of capsaicin at TRPV1 in the femoral artery, involving the efferent function of sensory nerves, are likely. The type of response is likely to be dependent on the dose of capsaicin because low doses, which may be more physiologically relevant, may activate the efferent function of sensorimotor nerves and cause local femoral vasodilatation without eliciting cardiovascular reflexes.

Sensorimotor nerves are present in the genitourinary and gastrointestinal tracts and can be activated by distension as well as chemical stimuli. Activation of sensorimotor nerves causes contraction and relaxation of smooth muscle in ureter. In the iris sphincter, electrical stimulation evokes large, slow contractions which are capsaicin sensitive and partly blocked by an NK1 antagonist. Activation of sensorimotor nerves in the airways mediates bronchoconstriction and protective reflexes such as coughing and wheezing. Tachykinins can cause bronchus constriction, airway edema, and mucus secretion. Capsaicin pretreatment prevents the induction of airway hyperresponsiveness in an experimental model of asthma and attenuates antigen-induced bronchospasm.

There may be an ongoing basal release of neurotransmitters from the peripheral terminals of sensorimotor nerves and these may be involved in trophic effects. In wound healing, SP, NKA, and CGRP stimulate proliferation of endothelial cells, smooth muscle cells, and skin fibroblasts, whereas destruction of capsaicin-sensitive afferents reduces survival of skin flaps and leads to the appearance of skin wounds. These trophic effects of sensorimotor nerves may occur in the normal physiological control of body function.

See also: Neuropeptides: Sensory Systems; Nociceptor Responses; Sensorimotor Integration: Models; Sensorimotor Control of Manipulation.

Further Reading

Barja F, Mathison R, and Huggel H (1983) Substance P-containing nerve fibres in large peripheral blood vessels of the rat. *Cell and Tissue Research* 229: 411–422.

Burnstock G (2000) P2X receptors in sensory neurons. *British Journal of Anaesthesiology* 84: 476–488.

Caterina MJ and Julius D (2001) The vanilloid receptor: A molecular gateway to the pain pathway. *Annual Review of Neuroscience* 24: 487–517.

Holzer P (1988) Local effector functions of capsaicin-sensitive sensory nerve endings: Involvement of tachykinins, calcitonin gene-related peptide and other neuropeptides. *Neuroscience* 24: 739–768.

Holzer P (1991) Capsaicin: Cellular targets, mechanisms of action, and selectivity for thin neurons. *Pharmacological Reviews* 43: 143–201.

Holzer P (1992) Peptidergic sensory neurons in the control of vascular functions: Mechanisms and significance in the cutaneous and splanchnic vascular beds. *Reviews of Physiology, Biochemistry, and Pharmacology* 121: 49–146.

Holzer P (1998) Neurogenic vasodilatation and plasma leakage in the skin. *General Pharmacology* 30: 5–11.

Holzer P (2004) TRPV1 and the gut: From a tasty receptor for a painful vanilloid to a key player in hyperalgesia. *European Journal of Pharmacology* 500: 231–241.

Holzer P and Maggi CA (1998) Dissociation of dorsal root ganglion neurons into afferent and efferent-like neurons. *Neuroscience* 86: 389–398.

Lundberg JM (1996) Pharmacology of cotransmission in the autonomic nervous system: Integrative aspects on amines, neuropeptides, adenosine triphosphate, amino acids and nitric oxide. *Pharmacological Reviews* 48: 113–178.

Maggi CA (1991) The pharmacology of the efferent function of sensory nerves. *Journal of Autonomic Pharmacology* 11: 173–208.

Maggi CA and Meli A (1988) The sensory–efferent function of capsaicin-sensitive sensory neurons. *General Pharmacology* 19: 1–43.

Wharton J, Gulbenkian S, Mulderry PK, et al. (1986) Capsaicin induces a depletion of calcitonin gene-related peptide (CGRP)-immunoreactive nerves in the cardiovascular system of the guinea pig and rat. *Journal of the Autonomic Nervous System* 16: 289–309.

Wood JN (1997) Chemical activators of sensory neurons. *Annual Review of Physiology* 59: 457–482.

Motor Control of Feeding and Drinking

J B Travers, The Ohio State University, Columbus, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Organization of Feeding

Feeding behavior is commonly divided into appetitive and consummatory phases. The appetitive phase of feeding, seeking out and finding food, is highly variable. We can, for example, make elaborate dinner plans and travel great distances to find our favorite restaurant, or simply stumble into the kitchen and open a can of soup. Equally diverse are the antecedents or coincidents of appetitive behavior, which involve a vast array of neural and hormonal signals reflecting metabolic need or surfeit, and the subsequent activation of hypothalamic pathways that lead to the search for food. Habits, time of day, and social pressure further induce one to eat. However, once food is in the mouth, the behaviors exhibited are far less variable, as one need only reduce the food to a size commensurate with swallowing, perhaps pausing long enough to enjoy the flavor, before propelling it toward the stomach. In comparison to the vast array of sensory and regulatory signals associated with body weight regulation or appetitive behavior, the highly stereotyped motor actions of mastication, licking, suckling, and swallowing are guided by a considerably more restricted set of gustatory and orotactile receptors.

Consummatory behavior can be subdivided based on behavioral, cinefluorographic, and electromyographic analyses. Although highly species specific, there are behavioral commonalities. Thus, the consummatory phase of eating begins with ingestion (putting food in the mouth), followed by transport of the bolus to an interdental position for mastication, and followed by a further transport phase to the back of the tongue for swallowing. Mastication can be further subdivided into a fast-closing phase that ceases when food makes contact with the bolus, at which point there is a slow-closing phase. Subsequent jaw opening begins with a slow jaw-opening phase followed by a fast jaw-opening phase. The act of swallowing is subdivided into oral, pharyngeal, and esophageal phases.

The seemingly simple motor problem of transporting food from the front of the mouth to the stomach belies an underlying complexity of coordination among a large number of muscles. Although mastication/licking and swallowing are inextricably linked during eating, sharing muscles of the tongue and jaws, historically they have been studied largely independently. This view is beginning to change as it is recognized that the generation of these behaviors does not derive from entirely dedicated circuits, but rather from multifunctional networks coordinating other oral motor behaviors, including respiration and, perhaps, even vocalization. Nevertheless, herein, the discussion will proceed from the perspective of acute neurophysiological studies in which mastication/licking and swallowing appear to have relatively distinct, spatially separate brain stem circuitry.

Mastication and Licking

The rhythmic consummatory actions of mastication, licking, and swallowing are organized by central pattern generators. By definition, this means that the critical timing information necessary to affect motor pattern switching (e.g., the transition from jaw opening to jaw closing) is not of peripheral origin. Rather, rhythmic jaw opening and closing can be centrally elicited by electrical brain stimulation (EBS) or pharmacological stimulation in select brain areas, and these rhythmic patterns can be induced in the absence of peripheral feedback following selective deafferentation or paralysis. Because mastication/licking and swallowing can be elicited in decerebrate preparations, the fundamental motor patterns are organized in the lower brain stem. However, there are intimate connections between brain stem preoromotor sites and forebrain sites involved with metabolic regulatory function, including the hypothalamus, circuits involved with motivation and reward (such as the nucleus accumbens and striatum), and cortical circuits mediating voluntary movements. Sensory feedback from intraoral gustatory and somatosensory fibers further modulates the motor pattern. Also, the motor neurons cannot be excluded from active participation in pattern generation. Although central pattern generators for mastication/licking and swallowing are not organized within the motor nuclei, oromotor neurons are not simply passive recipients of excitatory and inhibitory inputs. Rather, intrinsic membrane properties of the motor neurons facilitate rhythmic bursting behavior.

Muscles

Muscles of mastication The muscles of mastication are divided into jaw-openers (mandible depressors) and jaw-closers (mandible elevators); however, natural masticatory movements are rarely so simple. In humans, the mandible rotates forward during

opening and the temporomandibular joint allows lateral movements as well. Muscles that depress the mandible and thus open the jaw include the anterior digastric, mylohyoid, and inferior head of the lateral pterygoid. Jaw-closer muscles consist of the masseter, temporalis, medial pterygoid, and superior head of the lateral pterygoid. Because the jaw-closing muscles do most of the work of chewing during the closing powerstroke, these muscles are larger with more motor units.

The muscles of mastication show specializations indicative of their specific roles during chewing. For example, only the jaw-closing muscles are endowed with muscle spindles. Efferent feedforward to the intrafusal fibers of these spindles means that when the jaw-closing muscles meet resistance as they contact a food bolus, intrafusal fibers will become momentarily stretched in relation to their extrafusal counterparts. This stretch initiates a jaw-closing reflex via mesencephalic trigeminal nucleus afferents; this reflex monosynaptically further excites jaw-closing motor neurons, and thus facilitates the powerstroke. A (minimally) disynaptic jaw-opening reflex serves a different function. Unlike jaw-closing motor neurons that are centrally inhibited during the jaw-opening phase of mastication, jaw-openers are not subject to inhibition during jaw closing. This lack of inhibition on jaw-opener motor neurons allows a protective jaw-opening reflex to be initiated by high-threshold nociceptors, should, for example, the (sharp) teeth slide off a food bolus and make contact with the soft tissues of the mouth. In further evidence of specialization, the muscles of mastication have a greater number of hybrid isoforms of contractile proteins that endow them with intermediate contraction speeds and fatigue resistance, compared with muscles of the leg or trunk. In particular, the jaw-closing muscles, which do most of the work associated with mastication, have a larger number of slow and hybrid muscle isoforms, compared with jaw-opening muscles. Jaw opening is a relatively ballistic movement that is not against a load, and consequently relies on myosin heavy-chain isoforms (e.g., MHC-IIA) with fast conduction and fatigue-resistant qualities.

Lingual muscles The lingual muscles are divided into those intrinsic to the body of the tongue (affecting tongue shape), and the extrinsic tongue muscles with bony insertions (affecting tongue position). Intrinsic muscles include the superior and inferior longitudinal muscles in which fibers are parallel to the long axis of the tongue, vertical muscles that are perpendicular to the long axis, and transverse muscles that course across the breadth of the tongue. Of the extrinsic muscles, the genioglossus is the major lingual protrudor, and the hyoglossus and styloglossus muscles are the major lingual retractors. The geniohyoid is not, strictly speaking, a lingual muscle, but its contraction against the hyoid bone also functions to protrude the tongue. The tongue is remarkably flexible and serves prehensile actions in some species. This flexibility derives, in part, from the synergy between intrinsic and extrinsic muscles. The hydrostatic theory of lingual action makes clear this synergistic action. Stated simply, the tongue is considered a closed bag. Thus, contraction of the intrinsic longitudinal muscles will shorten the body of the tongue and facilitate retraction if coactivated with the styloglossus and hyoglossus muscles. Likewise, contraction of the vertical and horizontal fibers will lengthen the tongue and facilitate protrusion, together with the genioglossus and geniohyoid.

As with the muscles of mastication, there is specialization of muscle fiber phenotypes. The anterior tongue, which is composed exclusively of intrinsic fibers, has a large proportion of MHC-IIA fast fibers. The posterior tongue, in contrast, has many slow-contraction MHC-I fibers as well as hybrid isoforms. There are a small number of muscle spindles within the lingual muscles, but whether they provide monosynapatic input to hypoglossal motor neurons is unclear, and a role in rhythmic protrusion and retraction has not been demonstrated. Lingual reflexes in general play important protective roles, such as protecting the tongue from the teeth during chewing. Other muscle groups that contribute to ingestive oromotor behaviors include the posterior belly of the digastric (innervated by the accessory facial nucleus) and the obicularis oris and buccinator and zygomatic muscles (innervated by the facial nucleus) that position food within the mouth and form a seal to facilitate suckling or drinking.

Motor Nuclei

Trigeminal The motor neurons innervating the muscles of mastication are myotopically organized within the motor trigeminal nucleus (**Figure 1(a)**). Jaw-closer motor neurons that constitute the majority of the motor neurons are dorsomedial to those innervating the jaw-openers. During mastication, jaw-closer and jaw-opener motor neurons receive alternating waves of excitatory glutamatergic input. In addition, during the jaw-opening phase of mastication, jaw-closer motor neurons are under phasic glycinergic inhibition. Such inhibition may dampen jaw-closer reflexes initiated by passive stretching of muscle spindles in jaw-closer muscle fibers, thus making for a smoother and more rapid widening of the mandible. The membrane properties of trigeminal motor neurons have been studied in detail by SH Chandler's group. In the presence of serotonin, trigeminal motor neurons

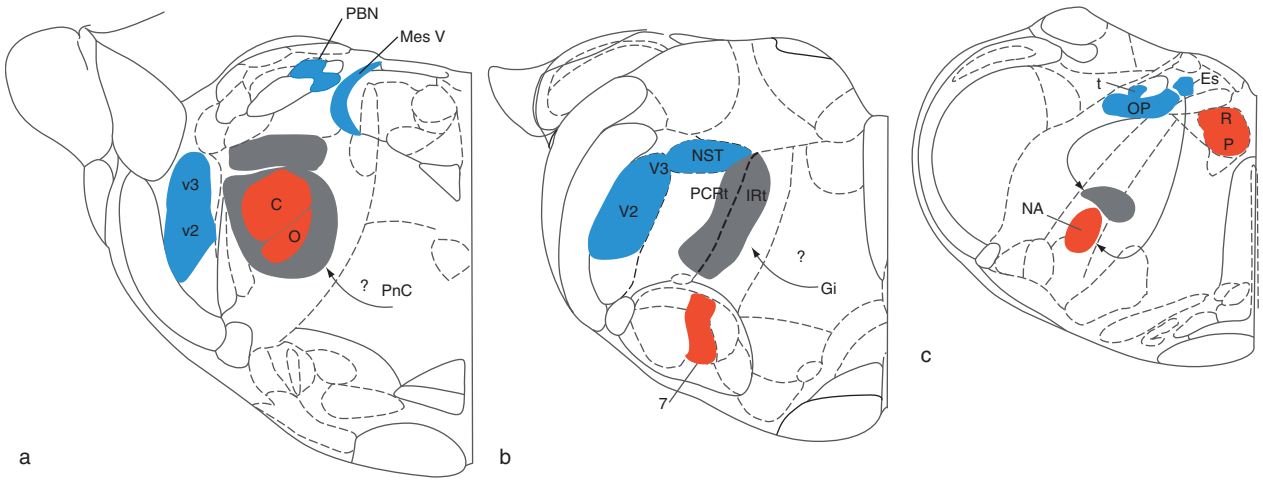

Figure 1 Major brain stem structures mediating consummatory responses of licking/mastication and swallowing. Sensory areas in blue, reticular formation structures in gray, and motor nuclei in red. C, jaw closure motor neurons in dorsal subdivisions of motor trigeminal nucleus; Es, esophageal generator neurons in the central subdivision of the nucleus of the solitary tract; Gi, nucleus gigantocellularis; IRt, intermediate zone of the medullary reticular formation; Mes V, mesencephalic trigeminal nucleus; NST, nucleus of the solitary tract; O, jaw opener motor neurons in the ventromedial subdivision of the motor trigeminal nucleus; OP, oropharyngeal generator neurons in the lateral and ventrolateral nucleus of the solitary tract; P, lingual protrudor motor neurons in dorsal subdivision of hypoglossal nucleus; PBN, waist area of parabrachial nucleus; PCRt, parvocellular zone of the medullary reticular formation; PnC, nucleus pontis caudalis; R, lingual retractor motor neurons in ventral subdivision of hypoglossal nucleus; t, solitary tract; V2, maxillary subdivision of principal trigeminal nucleus (a) or oral subdivision of spinal trigeminal complex (b); V3, mandibular subdivision of principal trigeminal nucleus (a) or oral subdivision of spinal trigeminal complex (b); 7, facial nucleus. Coronal sections adapted from Paxinos G and Watson C (2005) *The Rat Brain in Stereotaxic Coordinates*, 5th edn. Amsterdam: Elsevier.

display plateau potentials and rhythmic bursting behavior, produced by conductance changes in L-type calcium channels. The membrane properties of these motor neurons are further dependent on Ca^{2+}-dependent K^+ channels and Na^+/K^+-ATPase pump currents, which regulate burst termination during the afterhyperpolarization. Serotonergic input to the motor trigeminal nucleus comes from brain stem raphe nuclei and could thus facilitate rhythmical bursting via these intrinsic motor neuron membrane properties to augment excitatory phasic input coming from a central pattern generator.

Hypoglossal The hypoglossal nucleus controls tongue movement and is myotopically organized with lingual retractor motor neurons dorsal and protrudor motor neurons ventral (Figure 1(c)). Motor neurons controlling intrinsic lingual muscles are located along the midline of the nucleus. The nucleus is further spatially organized such that intrinsic muscle motor neurons controlling the tip of the tongue are located caudally toward the spinal cord and more posteriorly located extrinsic lingual muscles innervated by motor neurons rostral within the nucleus (i.e., the 'motor map' is reverse-oriented, with the tip toward the spinal cord). Hypoglossal motor neurons do not show phased inhibitory currents during electrical brain stimulation-induced rhythm movements. The implication is that they receive primarily excitatory inputs during behaviors such as licking.

Reticular Formation

The largest concentration of preoromotor neurons is in the brain stem reticular formation (RF), and it is all but certain that these neurons are important elements in a central pattern generator for mastication/licking. These preoromotor neurons are concentrated in the intermediate zone of the medullary reticular formation (IRt), but significant numbers of preoromotor interneurons are also found further lateral in the parvocellular RF (PCRt), as well as in regions of the RF immediately adjacent to the oromotor nuclei (Figure 1). The significance of the IRt/PCRt substrate is evident from behavioral pharmacology studies. Consummatory licking is reduced or eliminated following focal RF infusions of the γ-aminobutyric acid ($GABA_A$) agonist muscimol or excitatory amino acid antagonists, but appetitive responses such as approach and posturing to drink from the water bottle are not affected.

Neurophysiological evidence also indicates that this RF substrate provides excitatory and inhibitory interneurons that are essential to the normal functions of the oromotor nuclei. A majority of these premotor neurons are glutamatergic, with a relatively

smaller number of GABA-containing neurons. Neurons in the IRt/PCRt are rhythmically active during both natural consummatory behavior or following EBS, and three distinct classes of premototrigeminal interneurons have been identified: excitatory interneurons with projections to (1) jaw-opener or (2) jaw-closer motor neurons, and (3) inhibitory interneurons with projections only to jaw-closer motor neurons. Similar populations of excitatory interneurons to the hypoglossal nucleus have been proposed. Neurons in the IRt/PCRt thus have the necessary connectivity to be considered important pattern-generating neurons for mastication/licking. A small population of these neurons has projections to more than one oromotor nucleus, and they provide a basis for synergistic action between lingual, masticatory, and facial movements.

While there is general agreement over the location of final common interneurons that provide the drive to the appropriate oromotor neurons, there are alternative theories with regard to the origin of the fundamental rhythm for mastication/licking. Based on studies using EBS of the orbital gyrus to evoke masticatory-like (fictive) responses, the medial core of the medullary RF comprising multiple subdivisions of the nucleus gigantocellularis (Gi) was initially thought to provide the basis for the fundamental masticatory rhythm (Figure 1(b)). This rhythm-generating substrate was thought to project to pattern-generating neurons located further lateral in the IRt/PCRt region. However, recent studies have shown (1) that infusions of muscimol into Gi did not disrupt the rhythm of natural licking and (2) that interneurons adjacent to the motor trigeminal nucleus were capable of producing a masticatory-like rhythm (Figure 1(a)). Thus, unlike the respiratory central pattern generator, which has a clear locus for rhythm generation in the pre-Bötzinger complex, the origin of a masticatory/licking substrate remains obscure.

Sensory Inputs

Oromotor consummatory circuits are modulated by oro- and viscerosensory inputs at the brain stem level. As demonstrated by decerebrate preparations, the gustatory system has access to oromotor circuits. Intraoral infusions of preferred sweet stimuli lead to ingestive oromotor responses, while infusions of bitter nonpreferred stimuli, such as quinine, initiate an oral rejection response. These gustatory modulatory inputs can reach the IRt/PCRt via either the rostral (gustatory) nucleus of the solitary tract (NST) (Figure 1(b)) or via projections from the waist (gustatory) region of the parabrachial nucleus (Figure 1(a)). There is also evidence that visceral afferents from the gut can influence oromotor circuitry via brain stem pathways. Loading the stomach diminishes ingestive responses in decerebrate preparations, suggesting that second-order vagal afferents from the caudal NST influence the oromotor pattern generator, in much the same way as their rostral (gustatory) counterparts do.

Trigeminal afferents play key roles in oromotor ingestive function. In addition to trigeminal mechanoreceptors that guide the intraoral transport of food, Ruffini endings in the periodontal ligament reflexively trigger salivation during chewing, providing an important lubricatory and predigestive substrate. Afferent fibers of mesencephalic trigeminal neurons may contribute to the masticatory rhythm. In addition to innervating the periodontal ligament, mesencephalic trigeminal neurons innervate muscle spindles in jaw-closer muscles and are thus active during chewing. By virtue of a persistent inward sodium current, mesencephalic neurons burst when depolarized and may 'resonate' with the masticatory rhythm. Because the central connection of these afferent neurons is on trigeminal motor neurons, this resonant property could facilitate the masticatory rhythm. Trigeminal afferents are also subject to modulation by the central patterning substrate, and orosensory afferents are presynaptically inhibited during masticatory movements in order to dampen unnecessary reflex function.

Descending Control

Widespread regions of the midbrain and forebrain control the motor substrates for ingestion organized in the lower brain stem. Forebrain projections descend in parallel (Figure 2). For example, EBS of a cortical

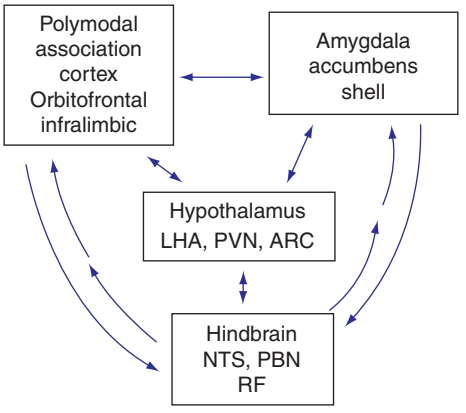

Figure 2 Parallel projections from the forebrain to hindbrain regions organizing consummatory ingestive responses. ARC, arcuate nucleus; LHA, lateral hypothalamus; NTS, nucleus of the solitary tract; PBN, parabrachial nucleus; PVN, paraventricular nucleus; RF, reticular formation. Adapted from Berthoud HR (2002) Multiple neural systems controlling food intake and body weight. *Neuroscience & Biobehavioral Reviews* 26(4): 393–428.

masticatory area (CMA) located in the frontal, orbital gyrus elicits rhythmic mouth movements, as well as movements of the tongue, salivation, and swallowing. EBS of different cortical areas can elicit mouth movements that resemble different components of the ingestive sequence (e.g., gnawing, mastication, intraoral transport, and even licking). Direct projections from these cortical sites reach the medullary RF, including the IRt/PCRt, but they also access the RF immediately adjacent to the motor trigeminal nucleus, and perhaps even the motor neurons. However, the prefrontal cortex is also intimately connected with the nucleus accumbens, the amygdala, and other ventral forebrain structures involved with motivation and reward, structures that also make connections with the RF. In fact, the consummatory licking response from pharmacological manipulation of the nucleus accumbens is prevented by inhibiting the lateral hypothalamus, which also has direct projections to the medullary RF. Thus, the ventral forebrain can potentially influence oromotor consummatory circuits via both cortical and lateral hypothalamic pathways. In addition to cortical, lateral hypothalamic, and amygdala projections to a lower brain stem central pattern generator, the substantia nigra and superior colliculus located in the midbrain offer additional indirect projections from the forebrain to lower brain stem circuits.

The cerebellum further modulates consummatory ingestive behavior. Ensembles of Purkinje cells show time-locked activity to rhythmic licking, and they project to the deep cerebellar nuclei. Outputs from the deep cerebellar nuclei reach the IRt/PCRt. Although not necessary to the generation of rhythmic ingestive responses, stimulation of the deep cerebellar nuclei, (nucleus lateralis and nucleus intermedius) alters the masticatory rhythm, while ablation of the cerebellum slows the chewing rate.

Swallowing

Swallowing is divided into a voluntary, preparatory 'oral' phase, followed by involuntary pharyngeal and esophageal phases. Because the pharyngeal phase includes activation of several oral muscles, this phase is more accurately described as the oropharyngeal phase. Thus, the involuntary reflex components of swallowing are generally described as consisting of oropharyngeal and esophageal phases. The oropharyngeal phase consists of contracting the posterior tongue and soft palate together, in order to force a food bolus into the upper pharynx. The lowest threshold site for initiating the involuntary components of the oropharyngeal phase consists of receptors lining the mucosa of the larynx, the epiglottis, and the portal to the pharynx (vallecullae, fauces), which are innervated by the superior laryngeal branch of the vagus (SLN). Glossopharyngeal-innervated receptors on the posterior tongue are somewhat less effective trigger sites. Sequential activation of pharyngeal constrictor muscles propels the bolus into the esophagus. Thus, the apposition of the posterior tongue against the soft palate forms an initial upper seal that demarcates the initiation of the oropharyngeal phase, and the contraction of the esophageal sphincter (cricopharyngeal muscle), once the bolus is in the esophagus, demarcates the initiation of the esophageal phase.

As many as 25 muscles are recruited into the swallowing sequence. These muscles include those innervated by the motor trigeminal, facial, hypoglossal, and ambiguus nuclei as well as motor neurons located in the upper cervical spinal cord. Muscle groups can be divided into those essential for swallowing (i.e., obligate muscles) and those that may or may not be recruited (i.e., facultative muscles). Facultative muscles include those of the tongue and face that participate in the oral phase of sequestering and organizing the bolus within the mouth. Obligate muscles of the leading complex of the oropharyngeal phase include trigeminal (mylohyoid, anterior digastric), lingual (styloglossus, genioglossus), hyoid (geniohyoid, stylohyoid), palatal (palatoglossus, palatopharyngeus), and pharyngeal (superior constrictor) muscles. Sequential activation of obligate middle and inferior constrictor muscles propels the bolus past the relaxed esophageal sphincter. Contraction of the cricopharyngeal muscle closes the esophageal sphincter, and esophageal peristalsis propels the bolus into the stomach.

Electromyographic recordings indicate that central inhibition plays an important role in the swallowing sequence, with muscle excitation preceded by activity levels below background levels. Thus, the middle and inferior pharyngeal constrictor muscles as well as the cricopharyngeal muscles are relaxed during activation of the leading complex. Such inhibition facilitates movement of the bolus. A similar pattern of sequential inhibition followed by excitation is also evident during esophageal peristalsis.

Central Control of Swallowing

Swallowing is organized by two major constellations of neurons in the caudal medulla (Figure 1(c)). A dorsal group includes neurons in the caudal NST and subjacent RF. These neurons receive peripheral afferent input from sites that reflexively trigger swallowing, in particular the SLN. The dorsal group is also the target for descending cortical input that initiates voluntary swallowing. Electrical stimulation, lesion, and recording studies all suggest that the dorsal swallowing group

functions as a pattern generator to sequentially activate the premotor (and esophageal motor) neurons. Thus, the caudal NST is the lowest threshold central site for eliciting swallowing via central stimulation, and NST lesions prevent both cortical-induced and peripherally induced swallows. Recording studies confirm that single neurons in the caudal NST fire during different phases of the swallowing reflex.

These dorsal group neurons, however, do not project directly to the motor neurons innervating the swallowing musculature. Rather, they project to premotor neurons located in the ventrolateral RF; these constitute a ventral group of swallowing-related neurons (**Figure 1(c)**). These neurons fire later than do their dorsal group counterparts, and tracing studies confirm that they connect to the oropharyngeal motor nuclei. In fact, some neurons in this region project to multiple oromotor nuclei, further establishing a role for this region in orchestrating the simultaneous activation of the diverse musculature that characterizes the oropharyngeal phase. Tracing studies using the transynaptic properties of viruses further confirm that esophageal 'generator' neurons in the dorsal group contact motor neurons in the nucleus ambiguus in the absence of additional (reticular formation) interneurons. Sections along the midline indicate the existence of a central pattern generator for swallowing on each side of the medulla. Coordination between the two central pattern generators to produce bilateral muscle activation is mediated by connections across the midline (e.g., between the two dorsal or ventral groups) and by bilateral projections from each ventral group to oromotor neurons.

The dorsal group of swallowing neurons is further differentiated into neurons generating the oropharyngeal and the esophageal phases. Neurons active during the oropharyngeal phase are located in the interstitial, intermediate, lateral, and ventrolateral subdivisions of the caudal NST (**Figure 1(c)**). Esophageal neurons are located in the intermediate and central subdivisions (**Figure 1(c)**). Lesion studies indicate that these populations of generator neurons can function independently – that is, a lesion in the central subnucleus will leave the oropharyngeal phase intact. Despite this spatial separation, the neuron groups are closely coupled through excitatory cholinergic and inhibitory GABAergic pathways (**Figure 3**). Infusions of cholinergic antagonists into the caudal NST suppress the esophageal phase while leaving the oropharyngeal phase intact. Similarly, infusions of the $GABA_A$ antagonist bicuculline 'releases' esophageal peristalsis, presumably from oropharyngeal generator neuron-mediated inhibition. Activation of the oropharyngeal phase requires excitatory amino acid receptors, particularly *N*-methyl-D-aspartate (NMDA) receptors.

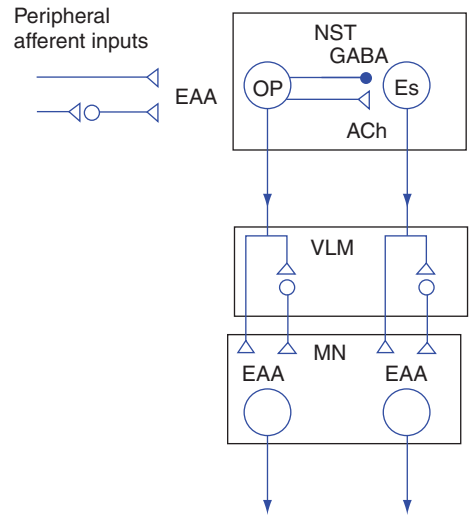

Figure 3 Neurochemistry of central pattern generator for swallowing. ACh, acetylcholine; EAA, excitatory amino acids; GABA, γ-aminobutyric acid; OP, oropharyngeal generator neurons; Es, esophageal generator neurons; MN, oropharyngeal motor neurons; NST, nucleus of the solitary tract; VLM, ventrolateral medulla. Adapted from Jean A (2001) Brain stem control of swallowing: Neuronal network and cellular mechanisms. *Physiological Reviews* 81(2): 929–969.

The pattern generators for swallowing organized in the dorsal group are also modulated by monoamines, and there is increasing evidence for roles by other neuromodulators such as leptin. Leptin receptors are located in the dorsal vagal complex, and local infusions suppress SLN-elicited swallowing, reinforcing the view put forth by H Grill that the brain stem contains neural substrates with the capacity for ingestive function regulation.

A theory of pattern generation for swallowing has been put forth by A Jean. The essential element of this theory is that generator neurons in the dorsal group that are active early in the swallow sequence (e.g., oropharyngeal-related neurons) excite a second population of generator neurons, which in turn excite a third population, and so forth. In addition to a chained excitatory network, neurons inhibit all of the generator neurons down the chain. Thus, neurons last in the chain receive the most inhibition. Such a schema could explain the sequential activations of specific muscles that are preceded by relaxation (inhibition) below basal rate. The demonstration of transient K^+ currents (A currents) in potential generator neurons in the caudal NST adds weight to the theory. Neurons with A currents display delayed excitation as a function of the strength of hyperpolarization. Thus, generator neurons destined to drive esophageal muscles would receive more inhibition than would generator neurons innervating preinferior constrictor neurons, and, perhaps, generator neurons for the

oropharyngeal phase would have the least inhibition. The strength of this inhibition is further dependent on sensory feedback and can be increased by peripheral stimuli that mimic bolus size. As such, a particularly large bolus might suppress excitation, delaying transit. Thus, although swallowing is centrally programmed, it is sensitive to sensory feedback.

Descending Control of Swallowing

Cortical projections to the NST likely account for initiation of voluntary swallows. Although imaging studies in humans suggest numerous cortical sites involved in swallowing (sensorimotor, prefrontal, anterior cingulate, and insular cortex), only relatively restricted sites in the anterolateral cortex evoke 'pure' swallowing when stimulated in experimental animals. Because stimulation of cortical swallowing sites affects primarily brain stem neurons active during the 'early' (oropharyngeal) phase, the cortex likely functions to trigger swallowing.

Coordination between Consummatory Responses

To what extent do mastication/licking and swallowing share central substrates? Clearly they share oromotor neuron pools innervating the tongue, jaws, and facial muscles. However, swallowing further recruits pharyngeal and esophageal muscles as well as laryngeal muscles to protect the airway. At the interneuron level, although the oropharyngeal phase of swallowing can be evoked from the caudal NST, a structure that appears to have little to do with initiating mastication/licking – the ventral group of swallowing interneurons – appears to overlap with substrates identified as necessary to mastication/licking. Thus, the ventral group of swallowing-related interneurons extends into the IRt/PCRt, though perhaps not extending dorsal to the ventral border of the rostral NST, an area rich in prelingual/masticatory interneurons. Chronic unit-recording studies confirm that some neurons in the IRt/PCRt that are active during licking are also active during swallowing. Other RF neurons are active during various combinations of swallowing, vocalization, respiration, licking, and oral rejection. Though unproved, it is thought that such interneurons form a multifunctional substrate to coordinate oral behaviors. And oral behaviors are clearly coordinated: the lick cycle is interrupted by swallowing, and swallowing is closely coupled to the respiration rhythm. Swallowing generally takes place between late inspiration and expiration and produces a brief period of apnea. During the oropharyngeal phase, the airway is protected by reflex laryngeal elevation as well as by closure of the larynx by laryngeal muscles. Although the initiation of the normal respiratory rhythm is usually attributed to the ventral lateral RF (e.g., pre-Bötzinger complex), gasping responses, which recruit both lingual and masticatory muscles to a greater extent than in normal respiration, appear also to be organized in the IRt/PCRt substrate, further emphasizing the critical role this area plays in orchestrating oromotor activity.

See also: Anterior-Posterior Spinal Cord Patterning of the Motor Pool; Basal Ganglia: Motor Functions; Central Gustatory System and Ingestive Behavior; Computational Approaches to Motor Control; Motor Neuron Specification in Vertebrates; Orofacial Motor Control; Reticular Formation and the Brain Stem.

Further Reading

Berthoud HR (2002) Multiple neural systems controlling food intake and body weight. *Neuroscience & Biobehavioral Reviews* 26(4): 393–428.

Blessing WW (1997) *The Lower Brainstem and Bodily Homeostasis*. New York: Oxford University Press.

Dellow PG and Lund JP (1971) Evidence for central timing of rhythmical mastication. *Journal of Physiology* 215(1): 1–13.

Ertekin C and Aydogdu I (2003) Neurophysiology of swallowing. *Clinical Neurophysiology* 114(12): 2226–2244.

Grill HJ and Kaplan JM (2002) The neuroanatomical axis for control of energy balance. *Frontiers in Neuroendocrinology* 23(1): 2–40.

Hiiemae KM and Crompton AW (1985) Mastication, food transport, and swallowing. In: Hildebrand M, Bramble DM, Liem KF, et al. (eds.) *Functional Vertebrate Morphology*, pp. 262–290. Cambridge, MA: Belknap Press.

Jean A (2001) Brain stem control of swallowing: Neuronal network and cellular mechanisms. *Physiological Reviews* 81(2): 929–969.

Nakamura Y and Katakura N (1995) Generation of masticatory rhythm in the brainstem. *Neuroscience Research* 23(1): 1–19.

Orchardson R and Cadden SW (1998) Mastication. In: Linden RWA (ed.) *The Scientific Basis of Eating*, vol. 9, pp. 76–121. Basel: Karger.

Paxinos G and Watson C (2005) *The Rat Brain in Stereotaxic Coordinates*, 5th edn. Amsterdam: Elsevier.

Travers JB (2004) Oral motor nuclei. In: Paxinos G (ed.) *The Rat Nervous System*, pp. 295–319. San Diego: Academic Press.

Travers JB, DiNardo LA, and Karimnamazi H (1997) Motor and premotor mechanisms of licking. *Neuroscience & Biobehavioral Reviews* 21(5): 631–647.

Travers JB and Travers SP (2005) Oral physiology. In: Cummings CW, Fredrickson JM, Harker LA, et al. (eds.) *Otolaryngology – Head and Neck Surgery*, vol. 2, pp. 1409–1436. St. Louis: Elsevier Mosby.

Motor Gait and Falls

W Mathiyakom and J L McNitt-Gray, University of Southern California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Falls are a major cause of morbidity and mortality in older adults. More than one-third of adults older than 65 years of age fall each year (Figure 1). The incidence of falls increases with age (Figure 1) and is also higher in institutionalized older adults (Figure 2). The consequences of falls experienced by older adults often extend far beyond minor injury to significant loss of functional independence and death. By 2020, the cost of fall-related injuries is expected to be $43.8 billion. Epidemic studies suggest that falls are associated with multiple factors, such as medical, biomechanical, physiological, psychological, and environmental factors. The interaction between these factors is complex and is further complicated by the heterogeneity of the aging process. Current clinical research related to fall prevention focuses on population-specific strategies for maintaining balance and balance recovery.

A fall can be defined as a sudden drop in body position. The position and velocity of the body can be controlled by generating reaction forces when the body is in contact with the environment. Maintaining balance requires that the body (often represented as a single point, the center of mass (CM)) is positioned over the base of support. During standing, the trajectory of the CM is kept within the base of support by regulating the reaction forces generated at the foot–surface interface. A reaction force (RF) is a push or pull that occurs during contact with the environment. The point of force application or center of pressure (CP) describes the centroid of the pressure distribution at the foot–surface interface. The CP location represents where a resultant reaction force would be applied to the foot to reflect the effect of all forces applied to the foot. The magnitude and direction of the RF and the CP are regulated by coordinated activation of multiple muscles. If the trajectory of the CM moves beyond the base of support, compensatory responses (e.g., step in the direction of movement) may be required to avoid a fall. If the net response is ineffective in controlling the CM relative to the base of support, a sudden drop in body position will result (a fall).

Maintenance of balance while performing activities of daily living is a critical factor contributing to aging in place. Effective interaction between the control (neural) and dynamic (musculoskeletal) systems is required to maintain and recover balance during both expected and unexpected situations. As an individual develops, he or she learns how to take advantage of the ever-changing set of neural and musculoskeletal capabilities to reliably perform functional movements. As one ages, alterations in the nervous and/or musculoskeletal system capabilities introduce age-related constraints that influence how an individual plans and executes daily tasks.

Age-related alterations in the neural system affect balance. The aging process alters sensory input (visual, vestibular, and somatosensory systems), the central processing (the planning and executing of movement), and cognitive domain in both expected and unexpected perturbations during interaction with the environment. Visual input provides the central nervous system with information regarding vertical and horizontal orientation of the body in relation to a reference system. During standing, vision detects slight changes in body orientation relative to the environment. The vestibular system provides information to the central nervous system (CNS) regarding head position and movement and helps resolve conflicting information between the visual input and the real movement. Somatosensory (e.g., proprioceptive and vibratory) input also provides important information to the CNS regarding segment and joint motion with respect to the support surface.

Previous research demonstrates that the abundance of sensory input allows the nervous system to effectively maintain balance. The absence of sensory input from multiple sensory systems amplifies body sway and increases the incidence of falls. For example, in younger adults, alteration in vision (no visual and inaccurate visual input) slightly increases body sway during stance. However, body sway significantly increases (approximately 3 times greater) when alterations in visual input occur in combination with inaccurate somatosensory (unstable surface) input. Likewise, in older adults, alteration in visual input in combination with inaccurate somatosensory input results in significant increases (approximately 2 times greater) in body sway compared to that of younger adults. In patients with vestibular loss, altered visual input or inaccurate somatosensory input slightly increases body sway, whereas vestibular loss in combination with altered visual input and inaccurate somatosensory input results in a high incidence of falls. These observed behaviors under compromised sensory conditions illustrate the role of multiple sensory systems in maintaining balance in a standing posture.

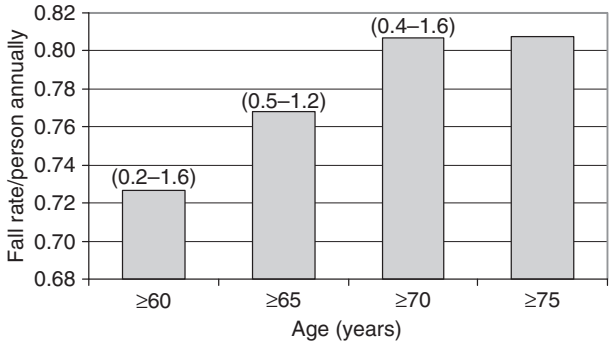

Figure 1 Mean (range) of annual fall incidence reported by different age groups of community-indwelling older adults. The mean of fall incidence increased as the age increased. Adapted from Rubenstein LZ and Josephson KR (2002) The epidemiology of falls and syncope. *Clinics in Geriatric Medicine* 18: 141–158.

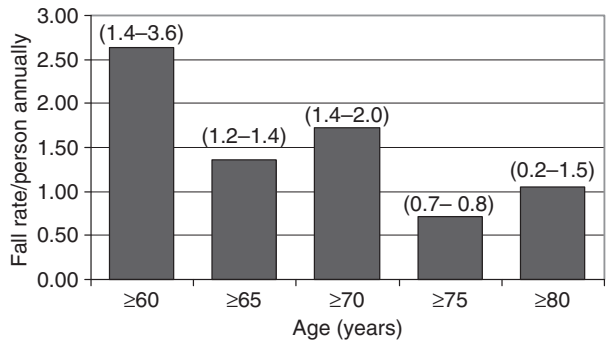

Figure 2 Mean (range) of annual fall incidence reported by different age groups of institutionalized (hospital and longer-term care) older adults. Note that the fall incidence was much higher than that reported by community-dwelling older adults within the same age group. Large variation in the number of falls reflects the interaction and the differences between the age-related pathology, activity levels, and fall prevention strategies and programs among facilities. Adapted from Rubenstein LZ and Josephson KR (2002) The epidemiology of falls and syncope. *Clinics in Geriatric Medicine* 18: 141–158.

Motor Strategies Used to Maintain Postural Stability

Classic research by Horak and Nashner provides a framework for classifying motor strategies used to maintain balance during upright postures under a variety of biomechanical and environmental constraints. This line of research has explored how the nervous system uses the sensory input, processes it in the context of previous learning experience, and executes a modified preprogrammed motor program appropriate for the perturbation, biomechanical, and environmental constraints. Control strategies used to maintain balance when unexpectedly perturbed are described as ankle, hip, and stepping strategies.

Ankle Strategy

An ankle strategy is commonly observed when the perturbation applied to a body is relatively small and the base of support is relatively wide. In this strategy, the body moves about the ankle joint as if it is a rigid body. To counter a forward sway of the body in response to a backward movement of the platform relative to the body, the muscles posterior to the ankle, knee, and hip are activated in a distal-to-proximal order (e.g., gastrocnemius, hamstrings, and paraspinal). In contrast, to counter a backward sway of the body, muscles anterior to the ankle, knee, and hip are activated (tibialis anterior and quadriceps). The latency of muscle activation is approximately 100–120 ms post perturbation. The set of observed muscle activation patterns appears to serve as a mechanism for redirecting the RF relative to the CM. If the RF acts anterior to the CM, a backwards moment about the CM is generated (to counter a forward sway). If the RF acts posterior to the CM, a forward moment about the CM is generated (to counter a backward sway). The magnitude of the moment about the CM is determined by the vector cross-product ($\mathbf{r} \times \mathrm{RF}$) of the position vector ($\mathbf{r}$), from the axis of rotation (CM) to the point of force application (CP), and the reaction force vector (RF). If the line of action of the RF vector is aligned with the CM, the RF does not cause rotation about the CM.

Hip Strategy

A hip strategy is typically observed when the perturbation is relatively large and the base of support is relatively small. In this strategy, the primary movements are observed about the hip joint. During a forward sway, the anterior muscles of the trunk (abdominal), hip, and knee (quadriceps) are activated, resulting in the use of hip flexion as a means of returning the CM over the base of support. In contrast, during a backward sway, posterior muscles of the trunk (paraspinal), hip, and knee (hamstrings) are activated. As a result, a hip extension movement is used to return the CM over the base of support. Motion of the trunk serves as a mechanism for relocating approximately 50% of the body mass (mass above the hip).

Muscle activation patterns used by older adults during balance recovery vary among individuals and populations. In some older adults (~40%), the order of muscle activation is reversed (distal to proximal). Older adults without a history of falls also exhibit a longer latency in distal muscle activation compared to young adults. Older adults who have fallen exhibit a longer latency of the distal muscle activation (tibialis anterior) and less ankle plantar flexor and dorsiflexor

strength than older adults without a history of falls. A reduction in muscular strength in combination with a longer latency in muscle activation may contribute to reductions in RF generation and/or the ability to reorient the RF relative to the CM to restore balance. Variations in balance control strategies among populations may represent the differences in scaling of the sensory inputs, previous experience, cognitive status, the state of the musculoskeletal system, and the ability to effectively interact with the environment.

Stepping Strategy

When a forceful perturbation is applied to the body (e.g., tripping), the ankle and hip strategies may not be sufficient for regaining balance and a stepping strategy is introduced. The stepping strategy requires coordination of both the supporting (stance) and recovery (swinging) legs. The role of the recovery leg is to establish a larger base of support and to provide a mechanism for redirecting the RF relative to the CM to counter the forward translation and rotation of the CM resulting from the trip. The stepping strategy is initiated by stabilizing the supporting leg and moving the recovery leg in the direction of the CM trajectory. While the recovery leg is in contact with the ground, the forward placement of the foot relative to the hip and CM allows the lower extremity to generate a backward and upward directed RF that passes anterior to the CM. The orientation of the RF relative to the CM counters the movement (translation and rotation) of the body associated with the trip, thereby providing a mechanism for regaining balance.

Age-related delays associated with the use of a recovery leg to regain balance influence the ability to regain balance during a forward fall. As the forward lean of the body prior to the fall increases, older adults are less likely to recover their balance than are younger adults. At small lean angles, the reaction and movement times of the recovery leg are similar between young and older adults. However, at maximum lean angle prior to the fall, the velocities of lower extremity joint flexion and extension during swing phase and the reaction and movement times are significantly different between young and old adults. These relationships were observed in both male and female groups. Variation in muscle activation patterns used to control the recovery leg also varied between older and younger adults. During the swing phase of the recovery leg, the onset of vastus lateralis activation occurs earlier (~40 ms) and the onset of rectus femoris activation occurs later (~90 ms) in older adults compared to younger adults. No significant differences in activation of the ankle plantar flexors, dorsiflexors, and hamstrings are observed between younger and older adults. These results indicate that the ability to regain balance from a forward lean position is limited by the ability to quickly orient the recovery leg.

Load distribution among the lower extremity joints of the recovery leg during contact also differ between young and older adults. Early in foot contact (~30 ms), ankle plantar flexor, knee flexor, and hip extensor net joint moments (NJMs) contribute to the control RF. Later in foot contact (~50 ms), ankle, knee, and hip extensor NJMs are observed. Older adults use smaller knee extensor but larger hip and ankle extensor NJMs compared to younger adults. These differences in load distribution strategy suggest that, compared to younger adults, older adults may impose a greater mechanical demand on the muscles controlling the ankle and hip when regaining balance using a stepping strategy.

Early RF redirection by the support leg also serves a mechanism to counteract the forward rotation of the body. Younger adults recover from a forward trip by using the support leg to redirect the RF anterior to the CM at ~150 ms post trip initiation. This RF creates a backward moment about the CM, thereby counteracting the forward rotation created by the trip. The upward component of the RF generated by the stance leg also translates the CM upward as indicated by hip elevation, thereby allowing more time for the recovery leg to be placed anterior to the CM. Redirection of the RF by the stance leg is achieved by a set of ankle plantar flexor, knee flexor, and hip extensor NJMs that are significantly different from those observed during the propulsive phase of gait. Generation of these moments by the stance leg during balance recovery requires significantly greater activation of the biarticular hamstrings, gastrocnemius, and uniarticular soleus muscles than during gait. These results indicate that effective balance recovery relies on the ability to quickly alter muscle activation patterns so that moment generated by the RF about the CM arrests the fall.

In older adult nonfallers, the patterns of the RF redirection relative to the CM, elevation of CM, lower extremity NJMs, and activation of muscles controlling the support leg during forward fall recovery are similar to those of younger adults. However, the magnitude and rate of development of muscle activation, rate of development of knee NJMs, and hip elevation are significantly slower for the older adult nonfallers than for younger adults. Older adults with a history of falls do not use their stance leg to redirect the RF anterior to the CM during forward fall recovery. After trip initiation, the RF generated by the stance leg is directed forward but passes posterior to the CM, contributing to both forward translation and rotation of the CM about the support foot.

The inability to redirect the RF anterior to the CM observed in older adult fallers is accompanied by relatively lower ankle plantar flexor NJMs and rate of development of ankle, knee, and hip NJMs compared to those of younger adults. These results indicate that alterations in lower extremity dynamics influence the RF redirection and forward fall recovery in older adult fallers. The reduction in magnitude and rate of force generation by lower extremity muscles may place older adults at risk for falls.

Postural Stability in Patients with Parkinson's Disease

Parkinson's disease (PD) is one of the most common neurological disorders associated with falls and recurrent falls. It has been estimated that 70% of PD patients fall at least once each year. The prevalence of severe injury resulting from falls (e.g., fractured hip) is 2.5–3.4 times greater in patients with PD compared to those without PD. Extrapyramidal signs (e.g., bradykinesia, axial rigidity, and tremor), as a result of the affected dopaminergic pathway in basal ganglia, alter several important mechanisms used to restore balance. These altered mechanisms include postural instability during unexpected perturbations, inability to regulate and redirect the RF relative to the CM during transition phases of goal-directed tasks, and coordinated activation of muscles responsible for coupling and decoupling body segment motion.

Research has demonstrated direction-specific instability in PD patients. When PD patients are perturbed by a moving platform, a high incidence of falls is observed in the backward and lateral directions. The incidence of falls is amplified when PD patients assume a more narrow base of support during stance. The decline in balance control observed in PD patients is primarily attributed to the inability to regulate the displacement of the CP at the foot–surface interface, thereby reducing the functional base of support and the ability to redirect the RF relative to the CM. When standing in a narrow stance, the peak CP displacement in all directions is significantly smaller for patients with PD than for healthy older adults. However, in wide stance condition, the reduction in peak CP displacement in PD patients is mainly observed in the backward direction. In healthy older adults, a decrease in base of support results in an increase in the rate of development of peak CP. In PD patients, the rate of development in peak CP during both narrow and wide base of support stance is slower than in healthy older adults. These results indicate that balance control in patients with PD is limited by the ability to regulate CP displacement.

Regulation of the CM displacement during multidirectional perturbation also contributes to postural instability observed in PD patients. During multidirectional perturbations, PD patients displace the CM significantly more than older healthy subjects. In narrow stance conditions, the magnitude of CM displacement is significantly larger in PD patients than in healthy adults in all directions except the forward direction. In wide stance conditions, the magnitude of CM displacement is comparable between healthy adults and PD patients. However, PD patients are slower in reversing the CM displacement compared to healthy adults, particularly when using a narrow base support. Significantly larger magnitudes of CM displacement and slower rates of CM reversal are attributed to significantly smaller amplitude of frontal hip movements and sagittal plane knee movements in order to bring the CM back over the base of support observed in PD patients compared to healthy older adults. Slower rates of CM regulation, reversal of CM orientation, and reduction in knee and hip motion indicate that postural bradykinesia commonly observed in PD patients may be related to higher background muscle activation and coactivation of muscles crossing both sides of the joints.

Restrictions in CP displacement in combination with increases in CM displacement contribute to smaller stability margins in PD patients. Stability margins during narrow stance are significantly smaller in all directions of perturbation for PD patients compared to healthy older adults. During wide stance, the stability margins are significantly smaller during backward and backward/lateral directions for PD patients compared to healthy older adults. Inability of PD patients to regulate safety margins in backward and lateral directions appears to be linked to limitations in regulating CP displacement in the backward direction and controlling the RF relative to the CM.

Regulation of the horizontal component of the RF serves as a mechanism for redirecting the RF relative to the CM during balance recovery. PD patients, however, demonstrate differences in horizontal RF generation in response to multidirectional perturbations compared to healthy older adults. Healthy older adults appear to use an asymmetrical strategy in controlling the horizontal component of the RF applied to the loaded and unloaded legs. In contrast, PD patients use a more symmetrical strategy. In healthy older adults, the loaded leg (ipsilateral side as the body sway) directs the horizontal RF in the lateral direction opposite to the direction of perturbation. The RF generated by the loaded leg returns the CM over the base of support. Simultaneously, the unloaded leg directs the horizontal RF in the anterior/posterior direction. This action by the unloaded leg serves to redirect the RF relative to the CM, thereby counteracting the rotation of the body. This differentiated role

of the loaded and unloaded legs allows healthy older adults to simultaneously control both translation and rotation of the body. In contrast, PD patients use both legs to generate diagonal horizontal RFs to counteract the translation and/or rotation imposed by the perturbation. Although this symmetrical RF redirection strategy may allow PD patients to maintain balance, the stability margin (CP–CM) is relatively small, particularly in the backward and backward/lateral direction. The symmetrical strategy used by PD patients may be related to an inability to initiate differentiated motor patterns when the conditions quickly change.

Coordinated and direction-specific activation of lower extremity and trunk muscles affects control of the RF relative to the CM. For example, during the right-backward body sway, older healthy adults exhibit direction-specific activation patterns of the lower extremity and trunk muscles. In the loaded (right) leg, the soleus, tibialis anterior, tensor fasciae latae, and erector spinae are activated, whereas in the unloaded (left) leg, the hip adductor and left abdominal muscles are activated. Evidence of direction-specific muscle activation patterns is also observed in PD patients. However, greater magnitude and shorter latency of the antagonistic muscles, acting in opposition to these muscles, are observed in PD patients compared to older healthy adults. The presence of muscle coactivation may result in more symmetrical net horizontal RFs observed in PD patients compared to older healthy adults. As the stance width decreases, the activation levels of the agonistic muscles significantly increase in healthy older adults. However, PD patients do not exhibit changes in muscle activation levels in response to the change in the base of support. These results suggest that coactivation of the antagonist muscles may impede the ability of PD patients to scale the magnitude of the agonistic muscles relative to the demand of the tasks and to regulate horizontal RFs in response to changes in the base of support.

Deficits in the nervous and musculoskeletal systems associated with the aging process and diseases prevalent in older adults introduce age-related constraints on mechanisms used for maintaining balance and balance recovery. These limitations in control appear to impede the ability of older adults and patients with neurological disorders such as PD to (1) regulate the segment and joint movements as indicated by reducing hip and knee displacement during perturbation, (2) reorient the total body CM relative to the base of support as a result of reduction of segment and joint motion, (3) reestablish a functional base of support as indicated by restricted excursion of the CP during stance and limitations imposed on stepping strategies, and (4) redirect the horizontal RF relative to the CM.

See also: Animal Models of Parkinson's Disease; Balance and Posture Control; Balance and Posture Control: Human; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Bloem BR, Steijns JA, and Smits-Engelsman BC (2003) An update on falls. *Current Opinion in Neurology* 16: 15–26.

Dimitrova D, Horak FB, and Nutt JG (2004) Postural muscle responses to multidirectional translations in patients with Parkinson's disease. *Journal of Neurophysiology* 91: 489–501.

Dimitrova D, Nutt J, and Horak FB (2004) Abnormal force patterns for multidirectional postural responses in patients with Parkinson's disease. *Experimental Brain Research* 156: 183–195.

Horak FB, Dimitrova D, and Nutt JG (2005) Direction-specific postural instability in subjects with Parkinson's disease. *Experimental Neurology* 193: 504–521.

Horak FB and Nashner LM (1986) Central programming of postural movements: Adaptation to altered support-surface configurations. *Journal of Neurophysiology* 55: 1369–1381.

Horak FB, Nashner LM, and Diener HC (1990) Postural strategies associated with somatosensory and vestibular loss. *Experimental Brain Research* 82: 167–177.

Madigan ML and Lloyd EM (2005) Age-related differences in peak joint torques during the support phase of single-step recovery from a forward fall. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 60: 910–914.

Madigan ML and Lloyd EM (2005) Age and stepping limb performance differences during a single-step recovery from a forward fall. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 60: 481–485.

Pijnappels M, Bobbert MF, and van Dieen JH (2004) Contribution of the support limb in control of angular momentum after tripping. *Journal of Biomechanics* 37: 1811–1818.

Pijnappels M, Bobbert MF, and van Dieen JH (2005) Control of support limb muscles in recovery after tripping in young and older subjects. *Experimental Brain Research* 160: 326–333.

Pijnappels M, Bobbert MF, and van Dieen JH (2005) Push-off reactions in recovery after tripping discriminate young subjects, older non-fallers and older fallers. *Gait and Posture* 21: 388–394.

Rodgers MM and Cavanagh PR (1984) Glossary of biomechanical terms, concepts, and units. *Physical Therapy* 63: 1886–1902.

Rubenstein LZ and Josephson KR (2002) The epidemiology of falls and syncope. *Clinics in Geriatric Medicine* 18: 141–158.

Thelen DG, Muriuki M, and James J (2000) Muscle activities used by young and old adults when stepping to regain balance during a forward fall. *Journal of Electromyography and Kinesiology* 10: 93–101.

Thelen DG, Wojcik LA, Schultz AB, et al. (1997) Age differences in using a rapid step to regain balance during a forward fall. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 52: M8–M13.

Woollacott MH, Shumway-Cook A, and Nashner LM (1986) Aging and posture control: Changes in sensory organization and muscular coordination. *International Journal of Aging and Human Development* 23: 97–114.

Motor Neuron Specification in Vertebrates

M-R Song and S L Pfaff, The Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Motor Neuron Progenitors: Extrinsic Signals

During early stages of neural tube development, multiple signaling activities intersect with one another and are responsible for specifying the identity of neuronal subtypes within the spinal cord. In the ventral neural tube where motor neurons and multiple classes of interneurons differentiate, cell fate specification occurs in a highly stereotypical manner based on the position of cells along the rostrocaudal (RC) and dorsoventral (DV) axes. Embryonic cell position and fate are linked because the location of undifferentiated progenitors is correlated with the types and concentrations of extrinsic factors they encounter – and it is the extrinsic factors that activate signaling pathways involved in establishing cell identity.

Motor neurons are located within the caudal regions of the nervous system, including the hindbrain and spinal cord. In the absence of extrinsic factors, progenitor cells default to an anterior identity, and as a consequence their progeny have characteristics of forebrain neuronal classes. A variety of extrinsic factors produced by Hensen's node and the mesoderm flanking the neural tube have been implicated in 'caudalizing' progenitor cells, including wnts, retinoic acid (RA), fibroblast growth factors (FGFs), and growth/differentiation factor 11 (gdf11). Although the interplay between these factors is still poorly understood, their graded activity appears to shift cells toward a progressively more posterior fate. Caudalizing signals play a critical role in establishing precursor cells that can give rise to motor neurons, whereas progenitors located at anterior levels do not encounter these transforming signals and therefore are not competent to generate motor neuron progeny. Based on studies of RC signaling in embryos, *in vitro* differentiation protocols using mouse and human embryonic stem cells to generate motor neurons typically incorporate RA and FGF to shift the rostrocaudal identity of the neural progenitors to a caudal identity appropriate for motor neuron specification.

RC patterning is not only involved in establishing progenitor cells competent to produce motor neurons but also contributes to the further diversification of motor neurons into individual subclasses. In large part, motor neurons develop in RC register with their peripheral targets. For example, limb-innervating motor neurons are located at limb levels (lower cervical and lumbar), whereas craniofacial-innervating motor neurons are located within the hindbrain. Grafting studies in chick embryos have shown that the early RC position of cells within the neural tube has a major influence on the eventual formation of motor neuron subclasses. Studies that alter RA, FGF, and/or Gdf11 signaling have shown that the identity of developing motor neuron subclasses is shifted along the RC axis of the neural tube. The influence of these signals also contributes to the precise assignment of motor neuron 'pools,' the unit representing a group of motor neurons that target specific muscles. Thus, caudalizing signals not only play a role in establishing precursor cells that can give rise to motor neurons but also operate at an extremely high resolution in defining what subtypes of motor neurons will develop along the RC axis of the embryo. These RC signals likely begin to influence progenitor cell identity prior to, and independently of, the DV signals that trigger motor neuron development.

Progenitor cells within the ventral portion of the caudal neural tube generate many cell types, including astrocytes, oligodendrocytes, floor plate cells, a diverse array of interneuronal classes (e.g., V0–V3), and motor neurons. Classical cell grafting studies have shown that graded signals along the DV axis of the neural tube influence cell fate specification. Sonic hedgehog (Shh) is produced as a secreted glycoprotein from the notochord and floor plate, establishing a high ventral-to-low dorsal gradient. With remarkable accuracy, neuroepithelial cells are able to readout different concentrations of Shh and respond by initiating specific differentiation programs. *In vitro* studies suggest that Shh activates signaling pathways for altering gene expression in the 0.5–10 nM range. Concentration differences as small as twofold lead to the activation of entirely different differentiation programs. At spinal cord levels, a narrow range of Shh triggers motor neuron development, whereas higher or lower levels of Shh trigger the differentiation of other cell types. Although Shh appears to only trigger the differentiation of a single generic population of ventral-exiting motor neurons (v-MNs) at spinal cord levels, the situation is more complex at hindbrain levels. Here, graded Shh appears to contribute to the specification of two major motor neuron classes. In the hindbrain, high concentrations of Shh have been linked to the development of dorsal-projecting visceral motor neurons (d-MNs), whereas progenitor cells located more dorsally in the neural

tube encounter lower levels of Shh and respond by differentiating as ventral-projecting somatic motor neurons (v-MNs). Thus, the intersection of RC and DV signals is integrated to specify motor neurons of appropriate subtypes at specific locations within the developing nervous system.

Motor Neuron Progenitors: Intrinsic Factors

RC and DV extrinsic signals are converted into intrinsic regulatory programs by activating the expression of multiple families of transcription factors, many of which contain a homeodomain for DNA binding. Along the DV axis, graded levels of Shh lead to the expression of unique combinations of transcription factors that define five separate progenitor cell domains within the ventral neural tube. An interesting cell-intrinsic mechanism is used to ensure that cells which fall close to the threshold level of Shh for two different progenitor cell domains are decisively resolved into a genetic program for a single cell type. This is achieved by having transcription factors for one progenitor cell type negatively regulate the expression of the factors expressed in adjacent cell domains. If cells initially start to express transcription factors for two different cell types, the reciprocal inhibition of the factors ensures co-expression is not tolerated, which rapidly leads to one set winning the cross-inhibitory 'battle.' Thus, a gradient of Shh is converted into all-or-nothing responses at the level of gene regulation controlled by progenitor domain-specific transcription factors.

The progenitor cells for motor neurons within the spinal cord reside within a 'domain' called the pMN region of the ventricular zone. pMN cells express a unique constellation of homeodomain transcription factors that distinguishes them from the progenitor cells for other neuronal types in the spinal cord. Furthermore, progenitor cells for d-MNs (restricted to hindbrain and upper cervical levels), such as trigeminal and spinal accessory neurons, arise from a unique pMN domain marked by Nkx2.2, Nkx2.9, Nkx6.1, Nkx6.2, and phox2b, whereas pMN cells that generate v-MNs express Nkx6.1, Nkx6.2, Pax6, and Lhx3. v-MN progenitors also express basic helix–loop–helix (bHLH) transcription factors that control the process of neurogenesis, a genetic subroutine involved in converting neural stem cells into postmitotic neurons. Two bHLH factors have been identified in pMN cells that have antagonistic functions toward one another: Olig2, which seems to hold pMN cells in a dividing progenitor cell state, and Ngn2, which favors neuronal differentiation. The counterrelationship between Olig2 and Ngn2 might serve as a mechanism for first expanding the pMN precursor population before allowing progeny cells to differentiate since precocious motor neuron differentiation would likely deplete the number of progenitor cells for motor neurons. The coordination of pMN growth and differentiation appears to be based on the timing of Olig2 and Ngn2 expression in pMN cells: Olig2 is expressed initially, but then it activates the expression of Ngn2, which begins to accumulate in the cells as development proceeds. The high Olig2 level in early pMN cells favors their growth, whereas the accumulation of Ngn2 beyond a threshold level over time appears to drive neuronal differentiation (neurogenesis) at later stages.

The high caudal-to-low rostral gradient of FGF in the embryo influences the expression of nested sets of Hox class transcription factors along the RC axis of the neural tube. Motor neurons that innervate the same muscle cluster their cell bodies into pools within the spinal cord, which typically span two to four spinal cord segments, vary in cell number in relationship to the size of the muscle they innervate, and extensively overlap with other motor pools spread along the RC axis. Gain- and loss-of-function studies have shown that Hox transcription factors are involved in establishing motor pool subtype identity. Like the progenitor cell factors involved in DV patterning, Hox genes involved in the development of different motor pools have cross-inhibitory regulatory interactions that help to convert graded extrinsic signals of RC position into unambiguous patterns of gene regulation for cell identity. By tuning the relative strength of the cross-inhibitory interactions for Hox factors, this could represent a mechanism for allowing some motor pools to expand in size at the expense of others so that motor neuron numbers within individual motor pools are roughly matched to the size (i.e., muscle fiber number) of the peripheral target.

Stepwise Progression in Motor Neuron Diversification

As progenitor cells generate postmitotic motor neurons, many changes in gene expression occur with the final cell division. The postmitotic birth of d-MN cells is tightly linked to the initiation of expression of the LIM-homeodomain (LIM-HD) factor Isl1 and the homeodomain factor Phox2a. On the other hand, v-MN cells express Isl1 and homeodomain factor Hb9 when they are born and maintain their expression of the LIM-HD 'progenitor' factor Lhx3 (Figure 1). Nevertheless, many of the other progenitor transcription factors found in pMN cells such as Olig2 and Pax6 are downregulated in newly formed postmitotic motor neurons. Little is known about the molecular basis for the further diversification of d-MN cell

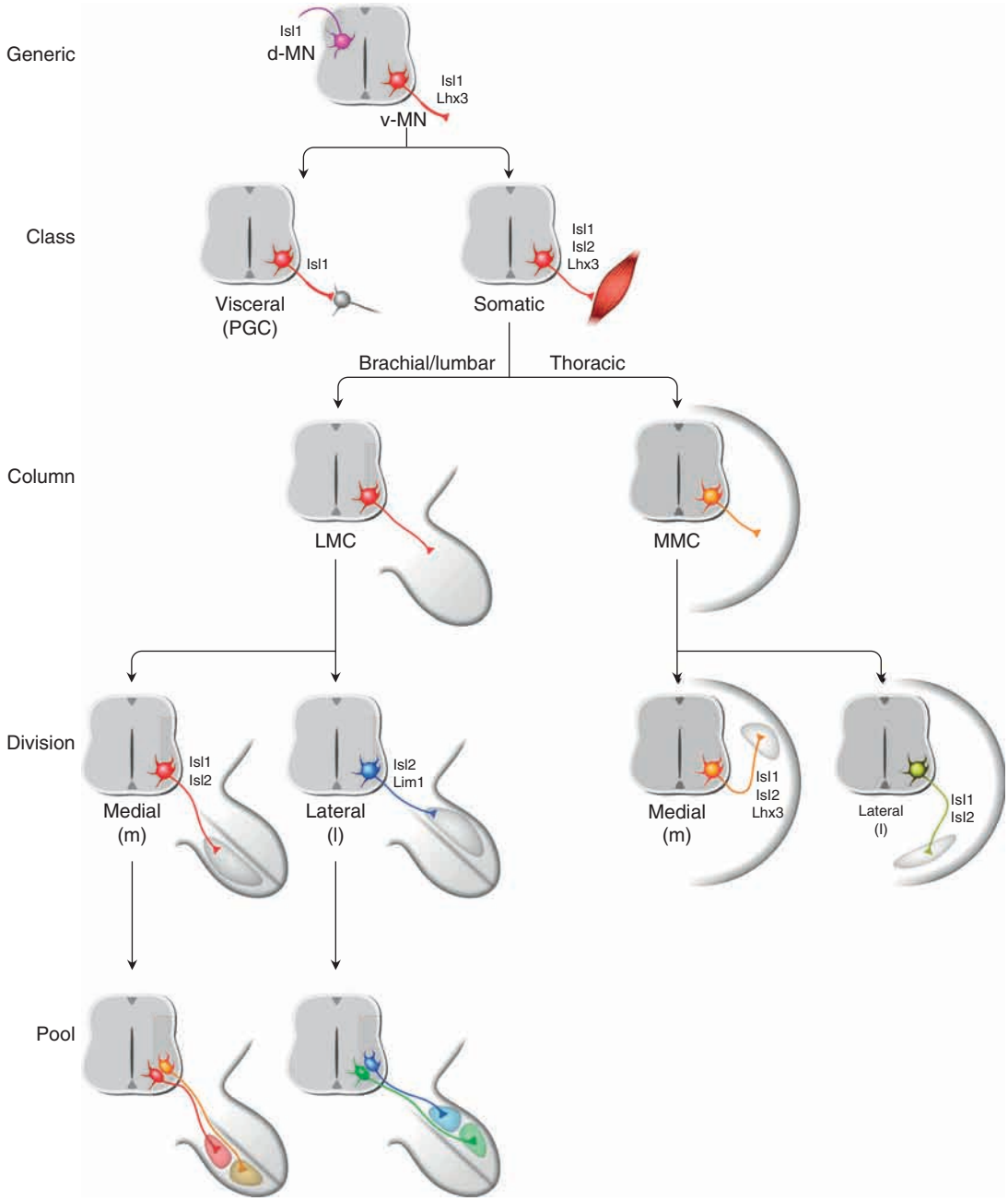

Figure 1 Motor neuron subtypes and their stepwise specification by LIM-HD factors. Motor neurons are classified based on their cell body position, axonal trajectory, and targets. Generic: Dorsal-projecting motor neurons (d-MNs) located at the hindbrain/upper cervical levels occupy dorsolateral positions and extend axons dorsally from the neural tube. There are many subclasses of d-MNs (not shown), and each has a slightly different cell body position and axonal trajectory. This diagram is a schematic representation of d-MNs that is not intended to indicate their precise position and axonal trajectory. Ventral-projecting motor neurons (v-MNs) extend axons ventrally from the neural tube forming ventral roots. Class: v-MNs at spinal levels are divided into two subclasses – somatic motor neurons innervate skeletal muscles and visceral motor neurons (forming a preganglionic motor column (PGC) in the intermediolateral spinal cord) innervate sympathetic neurons. Column: Somatic motor neurons are composed of multiple motor columns named based on their mediolateral cell body position – median motor column (MMC) neurons and lateral motor column (LMC) neurons. Division: Individual motor columns are divided into medial and lateral subdivisions – medial MMC cells extend axons dorsally and innervate axial muscles, lateral MMC cells extend axons ventrally and innervate body wall (intercostal) musculature, medial LMC (LMCm) cells extend axons into the ventral limb, and lateral LMC (LMCl) cells extend axons into the dorsal limb. Pool: Subsets of motor neurons clustered together within a single motor column such as the LMCl and LMCm that innervate the same muscle. The expression profile of LIM-HD factors is listed. Lhx3 has a paralogous gene Lhx4 (Gsh4) that has the same expression as Lhx3 and therefore is not listed for simplicity.

types, but numerous additional steps in the subspecialization of v-MN cells have been defined (Figure 1). Newly formed v-MN cells all extend axons out through the ventral root of the neural tube, but soon after this step additional changes occur in their profile of transcription factor expression reflecting the emergence of different v-MN subtypes. These dynamic and highly selective patterns of gene expression correlate with the migration of subclasses of v-MN cell bodies to specific locations within the neural tube and the growth of their axons along specific pathways in the periphery (Figure 1).

At thoracic spinal cord levels, v-MN cells can be divided into two major subclasses: somatic motor neurons innervating skeletal muscles and visceral motor neurons innervating sympathetic targets. Somatic motor neurons upregulate the LIM-HD factor Isl2, whereas visceral motor neurons downregulate Lhx3 and Hb9 but continue to express Isl1 as their cell bodies migrate dorsomedially to form a preganglionic motor column (PGC). Somatic motor neurons contribute to the formation of many different motor columns that are classified as either median (MMC) or lateral (LMC) based on their mediolateral cell body position in the ventral horn. The median and lateral motor columns are further subdivided into medial and lateral divisions to create four motor columns containing cells that innervate distinct peripheral targets: MMCm cells innervate axial muscles, MMCl cells innervate body wall musculature, LMCm cells innervate ventral limb targets, and LMCl cells innervate dorsal limb targets (Figure 1). Each of these motor columns is marked by a specific combinatorial pattern of LIM-HD genes. MMCm cells express Isl1, Isl2, Lhx3 (also called Lim3), and Lhx4 (also called Gsh4); MMCl cells found only at thoracic levels and LMCm cells found only at limb levels express Isl1 and Isl2; and LMCl cells express Isl2 and Lim1 (also called Lhx1). A variety of genetic studies have shown that the LIM-HD factors function in a combinatorial manner to control the expression of genes within motor columns that mediate axon pathfinding.

Motor columns can be further subdivided into motor pools which correspond to cells that are clustered together within a column and are related to one another by virtue of innervating the same muscle, expressing the same pattern of cadherins, and being gap junctionally interconnected. Individual motor pools express specific patterns of Hox and Ets class transcription factors. Ets factors such as Pea3 and Er81 are upregulated in motor pools by retrograde neurotrophin signaling activated by neurotrophin 3 (NT-3) and glial-derived neurotrophic factor (GDNF) produced by cells encountered by motor axons in the periphery. The Ets factors mediate a number of later functions in motor pools, whereas gain- and loss-of-function studies with *Hox* genes such as *Hoxc8*, *Hox5*, and *Hoxc6* have linked these factors to the actual specification of motor pool identity within the brachial (lower cervical) spinal cord. Taken together, the stepwise process whereby the subclass identity of 'generic' v-MN spinal motor neurons becomes progressively more refined is reflected in the sequential changes in the transcription factor 'coding' that take place during postmitotic stages of motor neuron development (Figure 1).

The LIM Code and Motor Neuron Diversification

LIM-HD genes evolved in early metazoans, and many studies have linked their function to the regulation of cell fate during the development of numerous tissues. The *Caenorhabditis elegans* LIM-HD factors Mec-3 and Lin11 are involved in specification of mechanosensory neurons and vulval cells, respectively. In *Drosophila*, the LIM-HD factor Apterous specifies dorsal cell identity during wing development. Likewise, LIM-HD factors in vertebrates are found in diverse organs, such as the developing pituitary, retina, limbs, pancreas, spinal cord, and brain, and are expressed with highly restricted expression patterns in specific cell types. As noted previously, motor neurons express unique combinations of LIM-HD genes during several phases of their development (Figure 1). v-MN cells express Isl1, Lhx3, and the Lhx3 paralog Lhx4 when they extend axons out through the ventral root, whereas d-MN cells express only Isl1. The first demonstration of the combinatorial function of LIM-HD genes was based on the finding that ectopic expression of Lhx3 in d-MN cells converted them to v-MNs, and conversely double knockouts of Lhx3 and Lhx4 converted v-MNs into d-MNs. Thus, the presence or absence of Lhx3/4 activity in newly formed motor neurons is thought to control the v-MN versus d-MN fate decision.

Interestingly, LIM-HD factors also contribute to the further diversification of v-MN subclasses at later stages of motor neuron development. Following the generation of v-MN cells, the emergence of motor columns is associated with the appearance of unique combinatorial arrays of five different LIM-HD factors (Figure 1). The refinement of motor neuron subtype identity is associated with both up- and downregulation of specific LIM-HD genes at postmitotic stages of neuronal development. Genetic studies in mice have shown that at later stages Lhx3 is involved in MMCm motor column development, Lim1 (Lhx1) is involved in LMCl development, and Isl2 is involved in PGC development. Taken together, functional studies on

the LIM-HD transcription factors indicate that they contribute to the development and subclass refinement of motor neurons by acting in a combinatorial manner at several stages in the stepwise development of these cells. Thus, they contribute to motor neuron subtype diversification at two levels: by functioning at multiple time points and by functioning in a combinatorial manner. These features of the LIM-HD factors enhance the ability of this relatively small gene family to contribute to the generation of an extraordinary level of cellular diversity within the spinal cord. Since all 12 members of the LIM-HD gene family are expressed by developing neurons in the peripheral and central nervous system, it is not surprising that studies have defined functions for these transcription factors in cortical, retinal, pituitary, and sensory neuron development as well.

The Molecular Underpinnings of the LIM Factors

The LIM motif is a cysteine–histidine-rich sequence which creates a tandem zinc finger structure that functions as a protein–protein interaction domain. Although numerous studies support the view that the LIM domain mediates interactions for higher-order ternary complexes, only a limited set of binding partners are well defined. Structural analyses and biochemical studies have demonstrated that the LIM domains derived from the nuclear LIM proteins encoded by 4 LIM-only genes and 12 LIM-HD genes can interact with the nuclear LIM interactor (NLI) protein (also called Clim and Ldb) (Figure 2). NLI is a widely expressed (probably ubiquitous) nuclear factor containing a self-dimerization domain and a LIM interaction domain. The protein-interaction characteristics of NLI suggest that it serves as a scaffold for assembling complexes with multiple LIM-HD proteins (Figure 2). A simple example of these LIM-HD complexes is found in V2a interneuron development – a population of Lhx3-expressing cells that form dorsal to motor neurons in the spinal cord. Biochemical and structure–function studies indicate that stoichiometric complexes of 2-Lhx3:2-NLI molecules regulate gene expression and specify V2a interneuron cell identity (Figure 2).

v-MN motor neurons expressing Isl1, Lhx3, and NLI represent a more complicated situation that illustrates a general issue with combinatorial codes based on reusing elements in a context-dependent manner. The co-expression of multiple LIM-HD factors, such as Isl1 and Lhx3 in the case of v-MN cells, raises the possibility that a mix of LIM complexes could form (i.e., 2-Lhx3:2-NLI, 2-Isl1:2-NLI, and Lhx3:2-NLI:Isl1). However, if this occurred, it is expected that the 2-Lhx3:2-NLI complex would inappropriately activate V2a interneuron genes, leading to a hybrid phenotype in v-MN cells. In this case, biochemical studies found that the LIM domain of Lhx3 is also able to bind with high affinity to the C-terminal portion of Isl1, leading to the formation of a hexameric complex

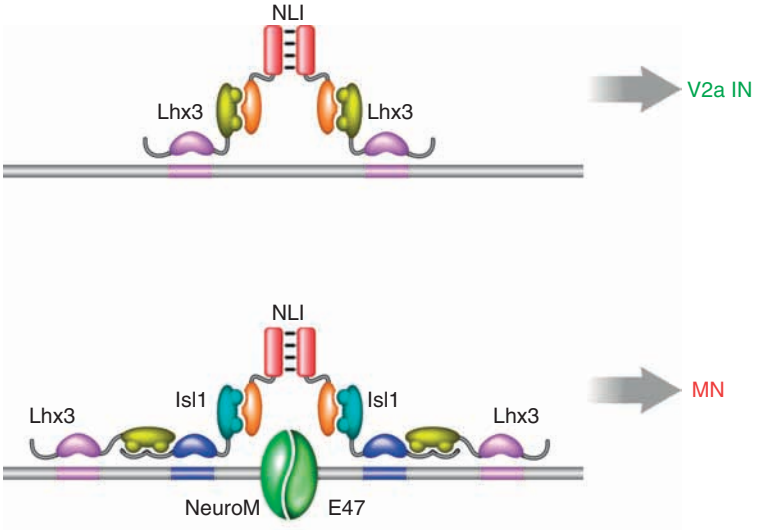

Figure 2 Ternary LIM-HD complexes regulate gene expression. The LIM domain of Lhx3 can bind to both NLI and Isl1. In V2a interneurons in which Lhx3 is expressed without Isl1, Lhx3 binds to NLI, which dimerizes to form a tetrameric 2-Lhx3:2-NLI complex for DNA binding. In v-MN motor neurons, Isl1 displaces Lhx3 from NLI, thereby preventing the V2a interneuron tetrameric complex from forming. Instead, Lhx3 binds to Isl1's C-terminus leading to the formation of a hexameric complex composed of 2-Lhx3:2-Isl1:2-NLI. This complex synergizes with bHLH factors such as Ngn2, NeuroM, and NeuroD (green) to activate high transcription levels. This synergy couples the activity of LIM-HD factors to proneural bHLH factors ensuring that neuronal differentiation and subtype specification are linked.

composed of 2-Lhx3:2-Isl1:2-NLI (Figure 2). The binding of Lhx3 to Isl1 results in a competitive situation whereby Lhx3–NLI interactions needed for V2a interneuron gene regulation are depleted at the expense of forming hexameric complexes for v-MN specification. Thus, neuron-specific protein–protein interactions represent one way in which LIM-HD factors can be used to regulate gene expression in different cellular contexts.

Although Isl1 and Lhx3 specify v-MN cell identity, until recently it was unclear how the activity of LIM-HD factors was tied to bHLH class transcription factors in developing motor neurons. Ngn2, NeuroM, and NeuroD are expressed by motor neuron progenitor cells and differentiating motor neurons and have proneural activity, meaning that they trigger cell cycle exit and activate generic neuronal genes such as β-tubulin. Transcription studies have found that the hexameric LIM complex containing Isl1 and Lhx3 (Figure 2) has a low inherent transactivating capacity. However, the binding sites of Isl1 and Lhx3 within the motor neuron-specific gene Hb9 promoter were found to flank E-box elements for bHLH factors. In the presence of Ngn2, NeuroM, or NeuroD, the activity of the LIM-HD and bHLH factors synergizes to greatly enhance v-MN gene activation. This synergistic link requires the dimerization of NLI (Figure 2), suggesting a conformational change occurs that mediates efficient transcription in the presence of LIM-HD and bHLH transcription factors. This cooperative link serves as a coincident detector that provides a means for integrating and coordinating the activity of factors for neuronal differentiation with those that control subtype properties.

Motor Neuron Soma Migration

Once generated, newborn motor neurons undergo stereotypical patterns of cell migration. Individual subclasses of motor neurons often settle in different positions from one another. For example, somatic motor neurons within the spinal cord typically settle in the ventral horn in specific medial and lateral columnar and pool positions (Figure 3). In contrast, thoracic visceral motor neurons migrate dorsomedially away from the ventral horn and settle at intermediate positions in response to reelin. Multiple transcription factors have been implicated in the control of motor neuron cell body positioning. For instance, misexpression of Lhx3 (normally in medial MMC cells) among non-MMCm neurons in *Lhx3* knockin mice shifted the settling position of motor neurons to a medial (MMCm-like) position. Likewise, *Isl2* null mice display defects in the mediolateral distribution of MMC neurons, and LMCl neurons require Lim1 induction for their normal migration. Ets class transcription factors such as Pea3 have been implicated in the positioning of motor pools within motor columns. The clustering

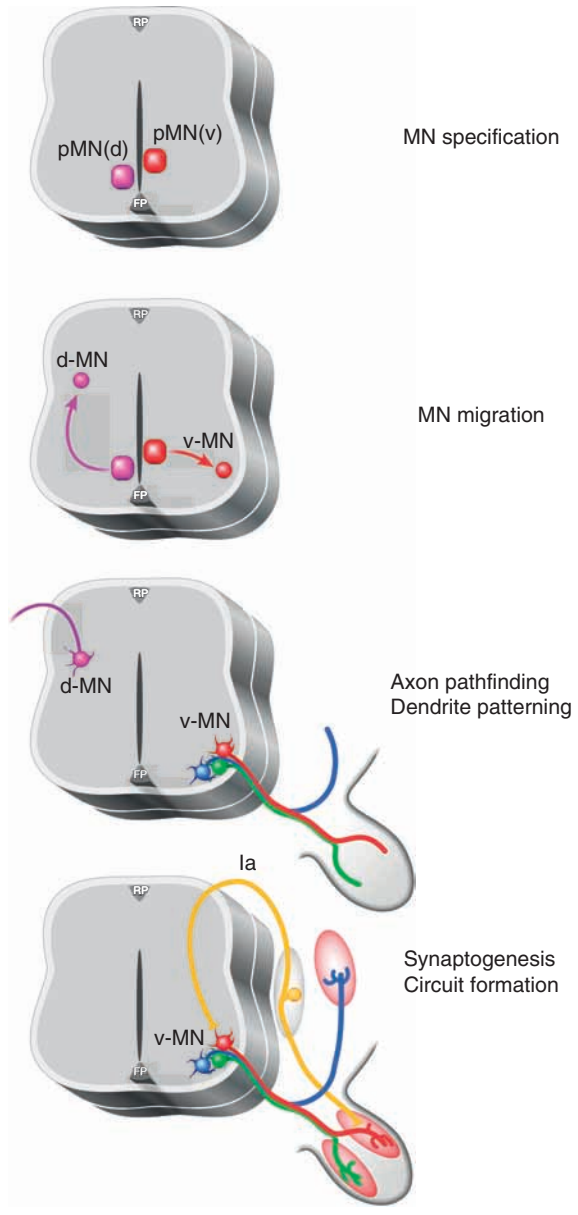

Figure 3 Stages of motor neuron development and maturation. v-MNs and d-MNs arise from different progenitor cell populations (pMN(d), purple; pMN(v), red) generated in response to different levels of Shh. The cell bodies of d-MNs move dorsolaterally and their axons exit the neural tube dorsally. The cell bodies of somatic v-MNs migrate laterally and form longitudinally aligned columns: medial MMC (MMCm, blue), medial LMC cells (LMCm, green), and lateral LMC cells (LMCl, red). All v-MN axons exit the spinal cord ventrally but select specific pathways in the periphery. LMC neurons grow to the base of the limb. LMCm cells innervate the ventral limb, whereas LMCl cells innervate the dorsal limb. MMCm neurons turn dorsally to innervate axial muscles. As motor neuron subtypes mature, they form peripheral branches and develop characteristic dendritic patterns which may contribute to the formation of specific connections for controlling their activity.

of like-motor neurons into motor pools has been linked to the cadherin gene family, encoding adhesion molecules that are selectively expressed by different motor neuron subtypes.

At hindbrain levels, the cell bodies of d-MNs (also called branchiomotor and visceromotor cranial motor neurons) migrate extensively. These classes of motor neurons often migrate dorsally along dorsolateral pathways, radially to the pial surface, and, in the case of facial cranial motor neurons, caudally along 'tangential' pathways. Although individual classes of d-MN cells follow different migratory pathways, some of the regulatory factors that control diverse migratory behaviors are shared among d-MN cells, including Tbx20 and Nkx6.1. In contrast, factors such as Ebf1, the planar cell polarity signaling pathway, and VEGF164 signaling appear to control a single process – the tangential migration of facial motor neurons.

Axon Pathfinding

Motor neurons share the unique property of extending axons into the periphery. All v-MN cells initially follow a common ventral pathway out of the spinal cord, but soon after exiting they are confronted with a choice point that branches toward three targets: MMCm cells select a dorsal pathway toward axial muscles; LMC and MMCl motor neurons select a ventral pathway toward the limbs and body wall muscles, respectively; and PGC axons grow ventrally toward sympathetic ganglia (Figure 1). Once LMC neurons reach the base of the limb, LMCl cells choose to grow into the dorsal region, whereas LMCm neurons enter the ventral limb compartment (Figure 1). Although arrays of LIM-HD transcription factors such as Isl1 and Lhx3 initially play a pivotal role in assigning motor neuronal fates, the expression of LIM-HD factors becomes progressively restricted to the selective motor neuron subtypes in postmitotic motor neurons (Figure 1). The specific expression of these factors precedes the selective axonal pathway and target innervation of neuronal subtypes, indicating that the combinatorial expression of LIM-HD factors (the LIM code) may guide the subtype-specific axon projection toward the target muscles. For instance, whereas most motor neurons rapidly extinguish Lhx3 and Lhx4 expression, MMCm neurons continue to express Lhx3 and Lhx4. Mouse genetic studies demonstrated that in the absence of Lhx3 and Lhx4, MMCm axons abnormally choose their exit points dorsally from the neural tube. Furthermore, when Lhx3 is stably expressed in all motor neurons as in *Lhx3* knockin mice, non-MMCm neurons extend axons toward axial muscles. This phenotype has been linked to Lhx3's regulation of FgfR1, which mediates chemoattraction toward the Fgf4/8-expressing axial muscles. LMCm neurons which innervate ventral limb muscles express Isl1, whereas LMCl neurons which innervate dorsal limb muscles express Lim1 (Lhx1). LMCl axons grow normally to the base of the limb in *Lim1* mutant mice, but they randomly select both the dorsal and the ventral limb compartment. Lim1 has been linked to regulation of EphA4, which mediates repulsion from the ephrin-A5-expressing region of the ventral limb. Individual motor columns contain motor pools which innervate the same muscle. Misexpression and knockdown studies with Hox genes such as *Hox5* and *Hoxc8* have shown that these factors control motor pool subtype identity and their altered expression influences axonal targeting.

Locomotor Circuitry: Dendrite Patterning and Synaptogenesis

Motor neuron cell bodies occupy specific positions and develop elaborate dendritic arbors for receiving input from presynaptic connections (Figure 3). Many classes of interneurons, corticospinal neurons, and sensory neurons synapse with motor neurons to control their activity; however, the mechanisms that control the precise pattern of input are just beginning to emerge. Individual motor pools display unique dendritic morphologies with regard to their branching pattern, size, and orientation. Genetic studies of *Pea3* mutant mice have found that the pattern of dendrite organization within specific motor pools requires the Ets transcription factor Pea3. After motor axons reach their target muscles, axons branch across the surface of the muscle and form neuromuscular synapses. Axons of specific motor pools display muscle-specific branching patterns, suggesting the presence of intrinsic branching programs. This is supported by the finding that *Pea3* mutants exhibit defects in the terminal arborization of motor neurons innervating the cutaneous maximus and latissimus dorsi muscles.

Motor neurons receive proprioceptive feedback input from Ia sensory neurons (Figure 3). This connectivity develops with incredible precision so that homologous motor and sensory neurons form monosynaptic connections. Interestingly, Ets factors Pea3 and Er81, which are selectively expressed in specific motor pools, are also present in subsets of sensory afferent neurons. Labeling studies have shown that there are preferential connections between sensory neurons expressing a particular Ets factor and the motor cells expressing the same factor. This finding suggests that Ets factors control multiple aspects of motor neuron maturation, including cell body position, dendritic pattern, and possibly Ia sensory input.

Conclusion

During the past decade, rapid progress has been made in defining the genetic pathways that control motor neuron subtype diversification. Characterization of the inductive signals and downstream transcription factors in motor neurons has provided great insight into the process of neuronal diversification. It seems likely that other neuronal populations will use similar DV and RC inductive cues and cross-repressive interactions to become diversified. Although our understanding is still fragmentary, the characterization of the factors involved in motor neuron subtype regulation is beginning to provide insight into the processes that control the development of locomotor circuitry.

See also: Anterior-Posterior Spinal Cord Patterning of the Motor Pool; Differentiation: the Cell Cycle Instead; Drosophila Apterous Neurons: from Stem Cell to Unique Neuron; Helix–Loop–Helix (bHLH) Proteins: Proneural; Helix–Loop–Helix (bHLH) Proteins: Hes Family; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Terminal Differentiation: REST.

Further Reading

Dasen JS, Tice BC, Brenner-Morton S, and Jessell TM (2005) A Hox regulatory network establishes motor neuron pool identity and target-muscle connectivity. *Cell* 123: 477–491.

Haase G, Dessaud E, Garces A, et al. (2002) GDNF acts through PEA3 to regulate cell body positioning and muscle innervation of specific motor neuron pools. *Neuron* 35: 893–905.

Jessell TM (2000) Neuronal specification in the spinal cord: Inductive signals and transcriptional codes. *Nature Reviews Genetics* 1: 20–29.

Kania A and Jessell TM (2003) Topographic motor projections in the limb imposed by LIM homeodomain protein regulation of ephrin-A:EphA interactions. *Neuron* 38: 581–596.

Kania A, Johnson RL, and Jessell TM (2000) Coordinate roles for LIM homeobox genes in directing the dorsoventral trajectory of motor axons in the vertebrate limb. *Cell* 102: 161–173.

Landmesser LT (2001) The acquisition of motoneuron subtype identity and motor circuit formation. *International Journal of Developmental Neuroscience* 19: 175–182.

Lee SK and Pfaff SL (2003) Synchronization of neurogenesis and motor neuron specification by direct coupling of bHLH and homeodomain transcription factors. *Neuron* 38: 731–745.

Livet J, Sigrist M, Stroebel S, et al. (2002) ETS gene Pea3 controls the central position and terminal arborization of specific motor neuron pools. *Neuron* 35: 877–892.

Sharma K, Leonard AE, Lettieri K, and Pfaff SL (2000) Genetic and epigenetic mechanisms contribute to motor neuron pathfinding. *Nature* 406: 515–519.

Sharma K, Sheng HZ, Lettieri K, et al. (1998) LIM homeodomain factors Lhx3 and Lhx4 assign subtype identities for motor neurons. *Cell* 95: 817–828.

Shirasaki R, Lewcock JW, Lettieri K, and Pfaff SL (2006) FGF as a target-derived chemoattractant for developing motor axons genetically programmed by the LIM code. *Neuron* 50: 841–853.

Shirasaki R and Pfaff SL (2002) Transcriptional codes and the control of neuronal identity. *Annual Review of Neuroscience* 25: 251–281.

Thaler JP, Lee SK, Jurata LW, Gill GN, and Pfaff S (2002) L. LIM factor Lhx3 contributes to the specification of motor neuron and interneuron identity through cell-type-specific protein–protein interactions. *Cell* 110: 237–249.

Vrieseling E and Arber S (2006) Target-induced transcriptional control of dendritic patterning and connectivity in motor neurons by the ETS gene Pea3. *Cell* 127: 1439–1452.

Motor Primitives

S F Giszter, Drexel University College of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Data from several lines of research suggest that the lower levels of the motor system may possess a set of motor primitives. These are compositional or constructive 'building blocks.' In a fashion similar to primitives used in computer graphics or robotics, they may simplify the problem of early movement construction and form an initial scaffold for motor learning. The motor primitives for limb movement are likely to reside in the vertebrate spinal cord. This article reviews the rationale for primitives, the development of an experimental framework, the current theoretical approaches, and recent data relating to spinal cord motor primitives and then concludes with a discussion of important open questions for future work.

The Degrees of Freedom Problem and the Rationale for Suggesting Primitives

The degrees of freedom problem in motor control was first strongly stated by Bernstein. The excess degrees of freedom available at the joint level for positioning the limbs, and the redundancy of muscles available to generate torques at these joints, make the issue of choosing the best limb postures, trajectories, and muscle activation patterns potentially complex (the so-called 'curse of dimensionality'). Wolpert, Ghahramani, and Flanagan have pointed out that human 'muscle space' comprises about 600 muscles and considering only contracted versus relaxed muscles leads to 2^{600} patterns, a number larger than the sum of atoms in the known universe! Of course, simultaneously, the available degrees of freedom enable immense flexibility. They can allow new motor strategies and the rapid use of novel tools. These new motor strategies may include types of movement never employed routinely or perhaps never employed previously in evolutionary history, (e.g., the Fosbury-flop strategy in the high jump). However, the great flexibility available with these degrees of freedom comes at a cost.

To make the problems and potential costs of the flexibility of the motor apparatus concrete, consider the wildebeest calf born in the Serengeti or a newly hatching turtle on a Caribbean beach. Both are subject to predation by a range of carnivores and are likely to soon be eaten if they fail to rapidly organize directed locomotion. Natural selection will thus favor their rapid development of effective movement and their efficient management of motor complexity through development. Modularity is an obvious way to solve several problems in motor control. It can help (1) manage complexity, (2) provide a basis to build more-complex motions, and (3) provide a 'seed' or bootstrap to motor learning, based on expectations about the world. Finer adjustments of the motor nervous system to the specifics of the individual motor apparatus, and novel motor adaptations or inventions, can then be added on timescales of development and individual life span. Notions of modularity in this spirit have arisen in ethology and neuroethology (e.g., fixed action patterns), in developmental biology, and in neurophysiology (the Sherringtonian reflex and interneuron classes in the work of Lundberg and Jankowska).

Research on motor primitives examines several issues: (1) What modularity can be identified in motor behavior? (2) Does modularity simplify the organization and control of biomechanical properties and movement? (3) Is modularity adequate to support both flexibility and the construction and learning of new movements? and (4) is modularity (a) hardwired, (b) developed in critical periods, or (c) simply an optimal short-term solution to an animal's specific and immediate task? The relationships of motor primitives to limb and body biomechanics and to the control and computation needed for movement are thus central issues. To this extent, a review such as this must draw on several approaches and avenues of research.

The remainder of this article will first introduce research on frogs (the section titled 'Motor primitives in frog spinal cord from microstimulation') which has shown that the spinal cord could potentially organize a collection of simple constructive or compositional elements quite directly linked to the limb biomechanics. Prior to these experiments, work on spinal cord already supported modularity from several different perspectives, and these data and approaches will be reviewed, putting the idea of motor primitives in context (the section titled 'Other historical, converging, and intersecting approaches to spinal modularity'). A body of theory on motor primitives has been elaborated from several different perspectives. A brief overview of these theoretical approaches to

primitives will be presented (the section titled 'Computational approaches to motor primitives'). The more recent research on spinal motor primitives in tetrapod vertebrates including mammals and man will follow, with a discussion of the relationship between reflex systems, pattern generators, and primitives indicated by some of these data (the section titled 'Current approaches to spinal motor primitives'). Finally, several important open questions are discussed that must be addressed in future work (the section titled 'Future issues in spinal motor primitives').

Motor Primitives in Frog Spinal Cord from Microstimulation

What the Spinal Cord Can Do

The spinal cord of vertebrates can organize complex and adapted movements in isolation. The spinal cord of the lamprey or dogfish can organize swimming; the spinal cord in rat or cat can organize the motor patterns needed to walk, trot, and gallop; and the spinal cord of rat, cat, turtle, and frog can organize precisely directed scratches. These involve coalescing and controlling many degrees of freedom of the joints and muscles.

Microstimulation: Hypothetical Possible Outcomes, Experimental Approach

The spinal basis for the abilities of frogs to make fine directed movements was investigated by Bizzi, Mussa-Ivaldi, and Giszter. By using microstimulation, they activated, with currents less than 10 μA, small regions of spinal cord in the intermediate gray matter in the spinalized frog. Both muscle activity and forces were recorded with the limb held isometrically at different limb locations. *A priori*, several kinds of outcomes appear possible with this experiment: First, random assemblies of muscles might be activated. Second, muscles might be activated in proportion to the underlying motor pool motor neuronal densities. Third, the circuitry of cord might be organized so as to define a continuum of kinematic actions or goals, such as final postures or specific directions or velocities of motion. Fourth, the circuitry might define a modular set of groups of muscles and feedback able to organize specific forces in the limb. The last two outcomes are not strictly exclusive of one another, but they do differ in various details, since Newton's laws specify it is not possible to control position and force simultaneously. Results of stimulation were expressed as force fields (Figure 1).

Results of Stimulation: Stable Force Fields

By moving the limb about the limb workspace, a set of force field patterns resulting from microstimulation of different regions of cord can be constructed. Passive (prestimulus) and active forces are distinguished (Figure 2). In stimulation directly of motor pools, patterns similar to those predicted based on motor pool composition are found. However, the outcome of stimulation in intermediate gray were found to be a collection of force fields which were conservative (i.e., emulated behavior of a passive mechanical system). These were all convergent in Cartesian (or external) space. All elicited fields were represented in joint space as torque/angle fields, and it was clear that all elicited fields were elastic. However, some fields did not exhibit torque reversals and therefore did not strictly define equilibria. Nonetheless, their elastic forces approached zero near joint limits. The stimulation defined neither a single unique force direction across the limb workspace nor a particular and unique movement direction. Instead, the elastic force fields (actually viscoelastic force

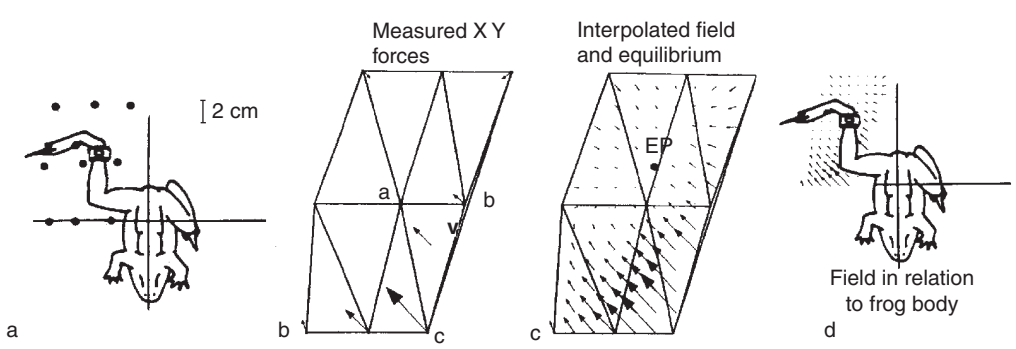

Figure 1 Construction of a force field following microstimulation or skin stimulation. (a) A collection of forces is recorded at several different spatial locations (circles). (b) A minimum perimeter (Delaunay) tessellation of the nine points is constructed. Within each triangle, vectors are estimated using an exact linear interpolation based on the three corner vectors. Thus, vector **v** is calculated using vectors at vertices a, b, and c. (c) The interpolated field is used to find any equilibria (EP). (d) The interpolated force field is shown in relation to the frog in the apparatus. From Giszter SF, Mussa-Ivaldi FA, and Bizzi E (1993) Convergent force fields organized in the frog spinal cord. *Journal of Neuroscience* 13: 467–491.

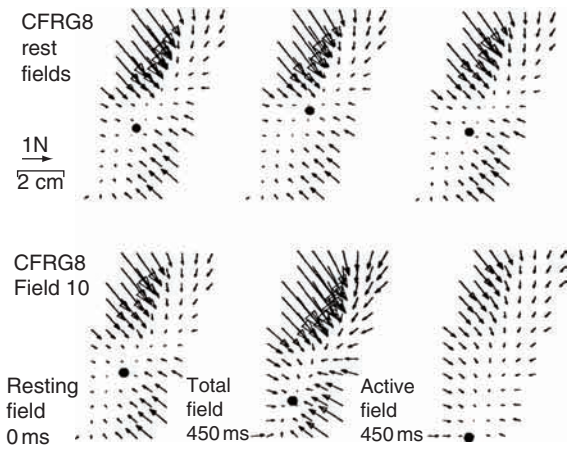

Figure 2 Estimation of the force field activated by stimulation. 'Resting fields and separation of active and resting fields'. Top, three different resting fields that were collected from a single frog during 6 h. These have an almost constant structure. Bottom, an example of the removal of the resting field (measured at 200 ms prior to stimulation) from the total peak field measured at 450 ms (250 ms following stimulation). The vector-by-vector subtraction used to obtain the active data is equivalent to the removal of baseline forces. The removal generates an 'active field.' It can be seen that in this case, the active field was convergent. Parameters: 9 μA at 60 Hz for 300 ms at 400 pm depth in cord. From Giszter SF, Mussa-Ivaldi FA, and Bizzi E (1993) Convergent force fields organized in the frog spinal cord. *Journal of Neuroscience* 13: 467–491.

fields measured isometrically) could move the limb to postures in the workspace or to workspace limits (Figure 3, panel 2). These fields showed several interesting properties of scaling and superposition.

Scaling of Force Fields over Time

Following stimulation at a site in the spinal cord, force fields that were elicited first developed smoothly and then relaxed (Figure 3). Throughout this process they remained a scaled version of the peak or maximum field. This property was not *a priori* expected. Spinal feedback systems activated by initial contractions could (in principle) alter muscle and torque balances later in the activity, but this was not observed. Either feedback played no part in the patterns seen or feedback conspired to support the force structures rather than dissolve them. In fact, subsequent experiments have shown that feedback regulation does contribute but that these contributions are not strictly necessary to the basic structure. Force fields also scaled with stimulation. If stimulation was increased, the peak amplitude was scaled, but structure was preserved.

A Small Number of Behaviorally Relevant Force Fields

If the electrode is moved throughout the intermediate gray of lumbar spinal cord in many frogs, a thorough sampling of the force patterns elicited throughout the cord can be made. Surprisingly, only about six different force patterns are elicited. Clustering all patterns using inner product measures shows that six fields account for the results. Some of the fields elicited by microstimulation are very similar to those found by skin stimulation to elicit reflex wiping or withdrawal. The small number of field structures are thus behaviorally relevant elements.

Combination of Force Fields

Dual stimulation of spinal cord to elicit different fields showed that these combined vectorially in 85% of tests and that one or the other field was suppressed in the remaining 15% (Figure 4). This observation is an extremely significant one and is the basis for considering the fields compositional or constructive elements of the spinal cord. It is interesting that to replicate blending, switching, and transition zones of scratch or wipe, a mixture of blending combinations and occasionally exclusive switching mechanisms is necessary.

Capacity to Drive Movement

The identification of a combination mechanism suggested that arbitrary limb postures and motions might be constructed through vector superposition of the few fields. Tests by Lemay showed that stimulation in the frog allowed adjusted control of limb position. Taken together, the fixed force field structures, the small number of structures, the combination mechanisms, and the capacity to organize and control movements led Bizzi, Mussa-Ivaldi, and Giszter to suggest the microstimulation results represented force field motor primitives from which a repertoire might be constructed. A theoretical framework, reviewed in the section titled 'Computational approaches to motor primitives,' was further elaborated.

Spinal Interneuronal Origins: Chronic Deafferentation

A clear concern, given the close relationship of stimulation results to reflex behaviors, has been whether afferents are the primary stimulation targets and whether the apparent organization observed is dynamic and transient. If so, the primitives are less likely to be available or present in other contexts. A number of experiments speak to this issue, but the most salient regarding afferent stimulation is to perform a chronic deafferentation and transection before microstimulation. After this manipulation, while interneuronal targets of afferents remain, the peripheral sensory fibers and descending fiber projections and synapses in the spinal cord are all stripped

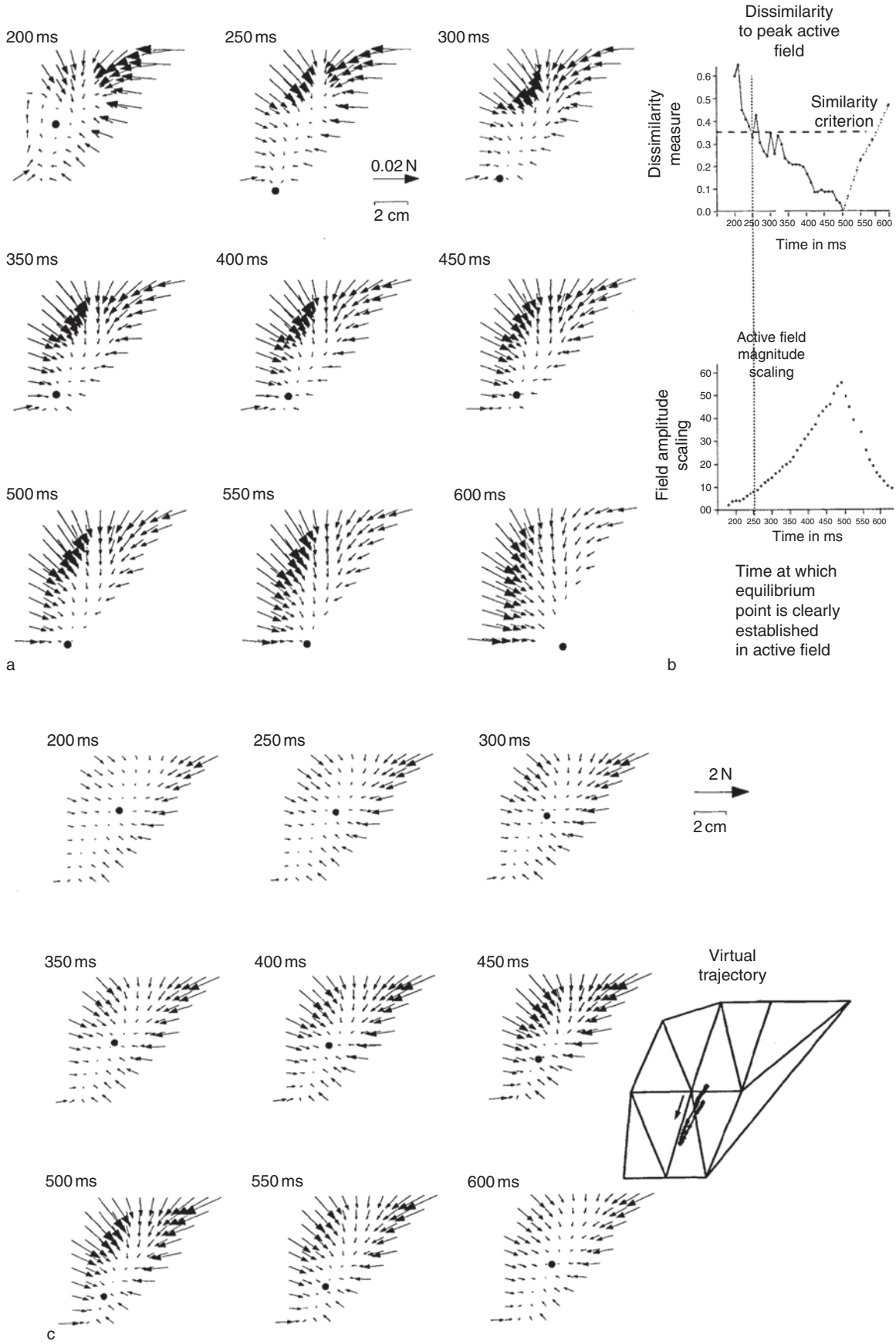

Figure 3 Continued

away and cannot contribute to, or organize, the patterns that result from stimulation. This experiment has been conducted in both frogs and rats, and in both preparations the basic observation of a collection of a few force fields that can be recruited by microstimulation survives this procedure.

Reflex and Sensorimotor Topography in Frogs

One idea that has arisen from microstimulation is that a topography of the motor primitives may exist. This observation of a topography of 'hot spots' for recruitment of primitives provides little information about the circuitry for primitives. The circuitry itself could be localized or distributed. Stimulation hot spots could be very important for prosthetics. However, the idea is not necessary for thinking about motor primitives. Topography has been much harder to pin down across species than the motor modularity and force field primitives. However, there are known topographies associated with afferents and sensorimotor reflex mappings in frog, rat, and cat. These topographies and their targets in the frog might be responsible for some of the spatial organization observed in microstimulation. The motor pools contributing to motor primitives are broadly distributed over multiple segments, and thus a broadly projecting collection of distribution interneurons must be involved in organizing primitives. It remains possible that the cell bodies of these interneurons may be highly distributed rather than localized. The topography observed in the frog, while experimentally convenient, is unlikely to be an important feature of all spinal motor primitives.

The observation of a small number of primitives forming a modular capacity to organize limb mechanical properties and movements drawn from the frog spinal cord is actually in keeping with a great many other avenues of research. The next section discusses these relationships.

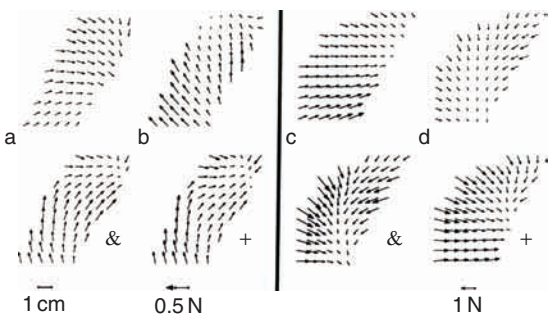

Figure 4 Vector superposition of force fields from spinal motor primitives. Results of costimulation of the lumbar gray matter in a frog. Two results are found: (1) vector summation (~85% of cases) and (2) winner-take-all (~15% of cases). Examples of vector summation. Upper left, fields (a) and (b) were obtained by separate stimulation of two different sites, A and B. Below, the observed costimulation fields (&) and summation field (+). This was by far (87.8% of cases) the most common outcome. The quality of fit of summation to observed is determined as the inner product derived Cosine. cos(a/b) close to 1.0 indicates near identity of two tested fields, a and b. In this panel cos(&/+) = 0.967. (Right) An example in which the costimulation (&) did not correspond to the sum (+) of the two fields (c) and (d) from two different stimulation sites C and D, but instead to a 'winner-take-all' case (D is the winner). Cos(D/&) = 0.949; but Cos(+/&) = 0.770. From Mussa-Ivaldi FA, Giszter SF, and Bizzi E (1994) Linear combination of primitives in vertebrate motor control. *Proceedings and National Academy of Sciences of the United States of America* 91: 7534–7538.

Other Historical, Converging, and Intersecting Approaches to Spinal Modularity

Reflex Circuitry and Central Pattern Generation

Spinal cord modularity arises naturally in the Sherringtonian conception of reflexes and its subsequent developments. The interneuron targets of reflex pathways from group II afferents, and the last-order interneurons described by Lundberg, Jankowska, Schomburg, and colleagues, may be closely related to the primitives

Figure 3 Behavior and characteristics of force fields from spinal motor primitives. (a) Normalized active fields are almost constant. To obtain this figure, active fields were obtained by removing the resting field from the fields shown in Panel 2. Stimulation parameters are 1.5 mA and 0.6 ms-duration pulses at 40 Hz for 300 ms. These active fields were then scaled to a constant average value to produce a normalized active field. (b) The magnitude of the scalar by which each of the normalized fields would have to be multiplied in order to obtain the original active field magnitude is shown in the graph to the lower right. Notice that after about 250 ms, the normalized field had an almost constant structure until 600 ms (when it begins to relax). A similarity measure for each active field relative to the peak active field was calculated and is in the upper right. After 250 ms, the similarity measure value dropped below the criterion value. This means the fields can be viewed as the same by this measure. The action of the microstimulation over time can thus be taken to be a scaling of the amplitude of a normalized active field of invariant structure. (c) Temporal evolution of the total force field. The total force field specifies the mechanical equilibrium of the limb. Over time, the field stiffens and alters as a result of microstimulation. The equilibrium point of each total field is shown. The motion of this equilibrium forms the virtual trajectory, which is shown to the right. The equilibrium alters as a result of microstimulation. The equilibrium point of each total field is shown. The time shown in the upper left of each field was measured relative to the stimulus train onset. Compare the pattern of motion of the equilibrium in the figure with the active field behavior shown in (a) for the same field. From Giszter SF, Mussa-Ivaldi FA, and Bizzi E (1993) Convergent force fields organized in the frog spinal cord. *Journal of Neuroscience* 13: 467–491.

identified as biomechanical and muscle synergies. Similarly, the C3–C5 neurons of Alstermark and L2–L3 propriospinal interneurons of Edgely and Jankowska may be very closely related.

Modular oscillators or unit pattern generators have been proposed by Grillner and Wallen, and unit burst generators have been suggested by Jordan. In turtle, Stein has identified modular burst generators or shared core structures. A central issue in relating pattern generator elements to primitives is how basic rhythmogenesis, multilimb and multijoint phase synthesis for gait, and pattern shaping are organized and interrelated and whether they are separable (see the section titled 'Supporting data in mammals').

Kinematic Decompositions of Movement

Kinematic modularity and kinematic trajectory strokes, submovements, or primitives have been identified and evaluated in a number of ways in human movements and in monkeys. This effort began with studies of Terzuolo and Viviani, and continued in work by Soechting and Flanders; Flash, Hogan, and Krebs; Sanger, Milner, and their various colleagues (see Figure 5). Miller, Houk, and colleagues have

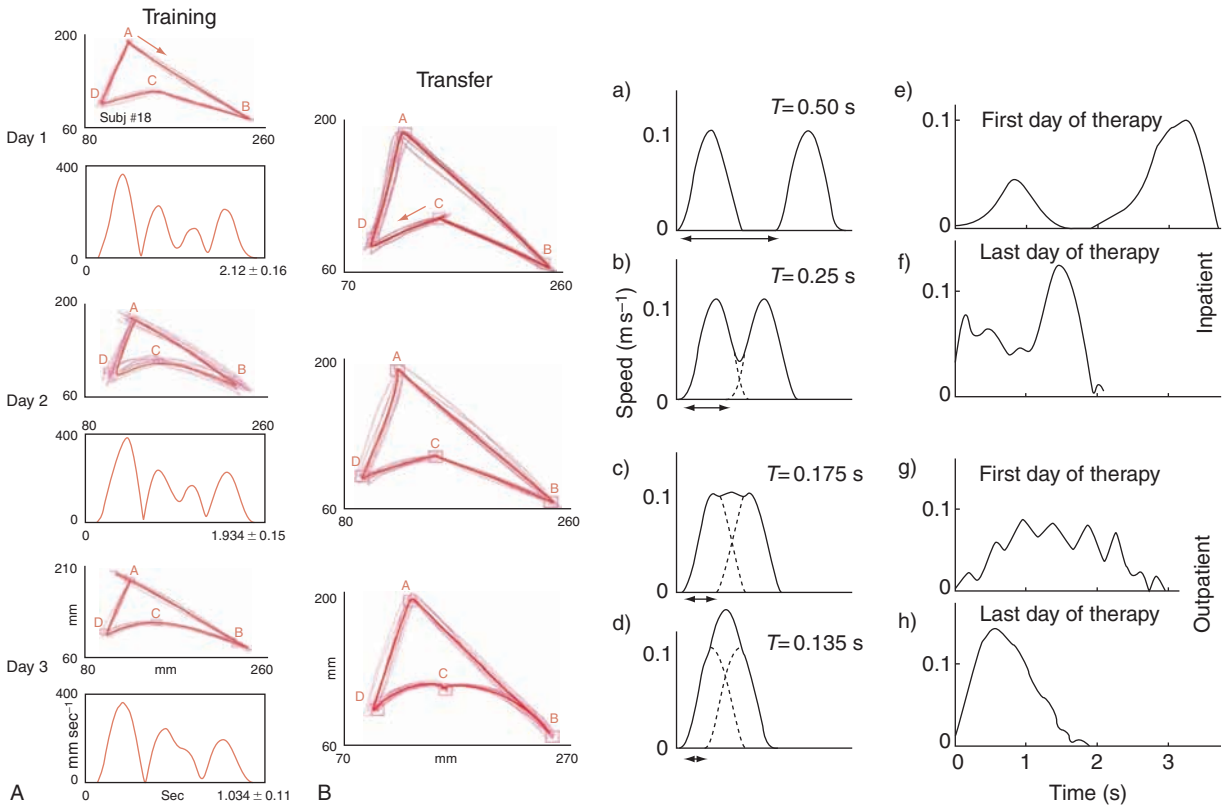

Figure 5 Kinematic submovements as primitives. Kinematic primitives may exist at the movement planning level. These are likely to differ from execution primitives. It is not known whether the kinematic primitives for voluntary movements are represented in spinal cord. Panel 1 (left): submovements in motor adaptation. Normal individuals learn to 'coarticulate' primitive submovements into new compound kinematic 'primitives' through adaptation. The learned strokes transfer from clockwise to counterclockwise motions about the points. The evolution of the geometrical 'primitive' during adaptation. (A) Training trajectories and transfer trajectories, generated in the same training day, for one representative participant. For each training day, the upper left plot denotes 20 training paths (pink) and average training path (red); lower left plot denotes average velocity profile. The rightmost plots denote 20 transfer paths (pink) and average transfer path (red). (B) Transfer paths of three participants who practiced for 7 days only the transfer condition (single block each day). Data from Sosnik R, Hauptmann B, Karni A, and Flash T (2004) When practice leads to co-articulation: The evolution of geometrically defined movement primitives. *Experimental Brain Research* 156: 422–438. Panel 2 (right): submovements in recovery from stroke; simulated vs. actual speed profiles. (a)–(d) Progressive blending of two minimum-jerk curves at various states of blending. See 'Materials and methods' of paper for a detailed description of the simulation. (e)–(h) Actual patient speed profiles; (e) and (f) are taken from the first and last therapy sessions of an inpatient; (g) and (h) are taken from the first and last therapy sessions of an outpatient. Simulated speed profiles qualitatively resemble the actual patient data. (a) Two distinct speed peaks, just as the patient speed profile (e). Continuing down the columns, (b)and (f) are qualitatively similar, (c) somewhat resembles (g), and (d) is similar to (h). Progression from the first to the last therapy sessions qualitatively suggests an increase in submovement blending. In addition, the movements of the participant that are longer poststroke (outpatient) have characteristics of more highly blended submovements compared with those of the inpatient. From Rohre B, Fasoli S, Krebs HI, et al. (2002) Smoothness during stroke recovery *Journal of Neuroscience* 22(18): 8297–8304.

tested similar decompositions in monkey corrections. In 'lower' animals, similar ideas of kinematic primitives have been proposed and evaluated by Full and Koditschek in locomotion, and in an ethological framework by Golani and Fentress. Related execution-level kinetic analyses of human movement in terms of force pulses have been developed by Gottlieb, Argawal, and Corcos.

The Russian physiologist Bernstein in the 1930s put forth the notion of a more fluid and transient development of modularity and coalescing of degrees of freedom, following his initial discussion of the degrees-of-freedom problem. This perspective is carried forward by experimental and theoretical work of Berkinblitt, Feldman, and Latash, with subsequent additions of Schoner and Scholz and others. Many of the issues raised by Bernstein in this perspective are now being addressed formally in several ways by Harris and Wolpert, Kawato, and Jordan and Todorov.

Computational Approaches to Motor Primitives

In Marr's decomposition of problems in computational neuroscience, analysis of a problem can be made at the task, algorithm, or implementation level. Motor primitives could be considered useful descriptors at each level, depending of the formulation. Several types of computational motor primitives have been suggested, and these differ in how they would fit in a Marr scheme. A primary division of theoretical work is into analysis of (1) kinematic motor primitives offering decomposition at a more abstract planning level and (2) force field primitives, more directly associated with execution. Kinematic primitives are probably best classified in task or algorithm level in Marr's scheme, and force field primitives are best classified as algorithmic or implementation. Although the emphasis here is on spinal motor primitives and these are thought to be force field execution elements, kinematic representations and body schemes may exist in spinal cord, and briefly considering both types of elements and their elaboration and implications is useful here.

Kinematic primitives have been developed by Schaal, Ijspeert, and colleagues. They address primarily planning issues: how to build compact dynamical representations of multijoint task kinematics that guarantee task achievement and stability. Their representations are of two types. The first type is discrete point-to-point kinematic motions, with static attractors. The second type is limit cycle oscillators at joints. If these two classes of elements are coupled, a scheme is developed that allows optimization of dynamic patterns in a relatively simple and a highly principled fashion. The scheme supports imitation learning in biomorphic robots and gives a plausible account of kinematic imitation learning in man. Execution details are managed downstream from these kinematic representations (by a separated inverse-dynamics capacity). These primitives are thus in the tradition both of kinematic decompositions (Figure 5) and of unit oscillators reviewed in the section titled 'Reflex circuitry and pattern generation.'

Force field primitives related to the work in spinal cord have been developed by Mussa-Ivaldi and colleagues and represent execution elements. Following their descriptions, activity of muscles and feedback pathways as a group can be generically represented as a multidimensional time-varying elastic field in joint space:

$$F(q, \dot{q}, t) = C(q, \dot{q}, u(t)) \qquad [1]$$

where q, $\dot{q}$ are joint angles and angular velocities, $u(t)$ represents the vector of control variables, t is time, F is the force field in joint torque, and general joint coordinates, C is a (noninvertible) function transforming muscle activations to F.

As F drives motion, limb or body dynamics can be written as follows:

$$M(q)\ddot{q} + G(q, \dot{q}) + E(q, \dot{q}, t) = F(q, \dot{q}, t) \qquad [2]$$

where M represents intertial terms; G, interaction terms; E, environmental forces; and F, the torque generation from eqn [1].

The problem of movement generation is to deliver an appropriate set of forces F to satisfy the task constraints, which may be purely kinematic (e.g., pointing, gesture, or dance), purely counteracting an environmental force/load (e.g., stationary support of a book while reading), or a combination of target motions and interaction forces with the environment (e.g., locomotor stance phases or manipulation).

Spinal motor primitives provide a modular basis for constructing an arbitrary field. Mussa-Ivaldi showed an arbitrary smooth static field may be approximated with a combination of n static circulating and conservative fields:

$$F(q) = \sum_i \Phi_i(q) + \sum_i \Psi_i(q) \qquad [3]$$

where the Ψ are circulating fields and Φ are conservative fields

The fields found in practice in experimental work show no significant circulation and thus emulate passive conservative systems, and Ψ can be eliminated in the approximation in eqn [3]. This is consistent with constraints on guaranteeing passive stability

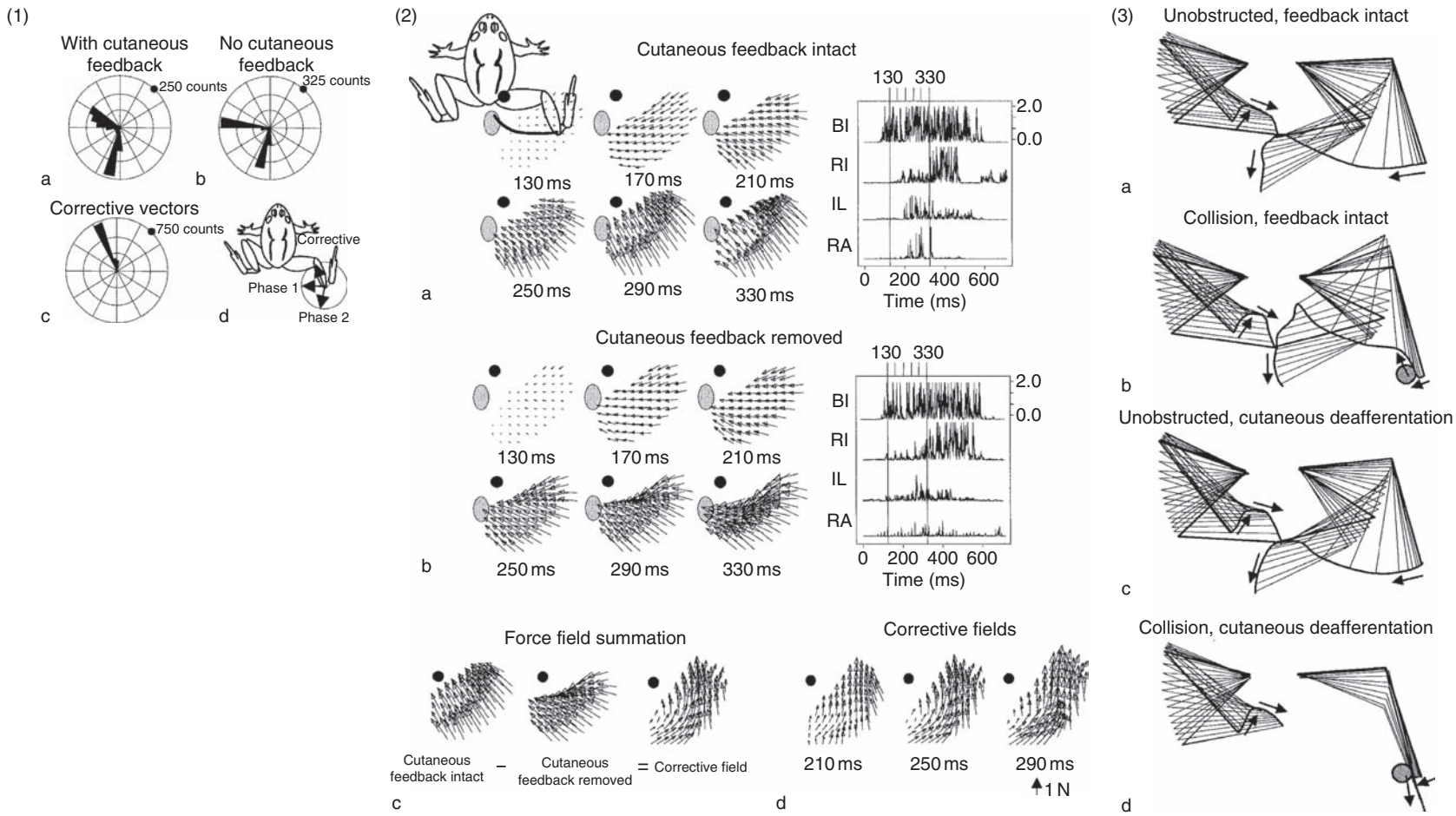

Figure 6 Continued

in interaction that were formalized and identified by Hogan and Colgate. Thus, for the biological primitives, the approximation needed can be written as follows:

$$F(q, \dot{q}, t) = \sum_i a_i(t)\Phi_i(q, \dot{q}) \qquad [4]$$

where a viscous term is added and $a(t)$ represents the time course of a viscoelastic force field.

Mussa-Ivaldi has used this framework to create fields from pulse- and step-activated primitives that generate two-joint planar limb kinematics.

In the spinal and semi-intact frog, experimental data support a fixed time course pulsatile force field element potentially initiated at arbitrary phases, and the trajectory formation can be expressed as follows:

$$F(q, \dot{q}, t) = \sum_i A_i a(t + \tau_i)\Phi_i(q, \dot{q}) \qquad [5]$$

where A is the amplitude of the pulse; τ, the timing of the pulse; and $a(t)$ similar for all primitives.

The problem of movement construction is then simply the selection of the number, timing, and amplitude of a sequence or repeating rhythm of such pulses. For example, a rhythmic central pattern generator (CPG) comprising a collection of coupled oscillators forming interlimb gait and timing synthesis could organize such elements in a pattern shaping layer for gait.

However, this degree of simplicity, namely, a fixed duration pulsed force field as primitive, may be peculiar to 'lower' tetrapods or the preparations

Figure 6 Superposition of motor primitives' force fields during online correction behaviors executed by the spinal cord. A single kinematic stroke, motion to a target with a unimodal velocity profile during wiping/scratch movements is executed by sequencing of motor primitives. Rapid online correction of this single stroke trajectory employs superposition of a new primitive, added into the pattern when triggered by cutaneous input from collision with an obstacle. Other aspects of the motor pattern are unaltered by the summation of the new element. Panel 1: corrective force components are directionally stable. (a)–(c) Circular histograms of the directions of force vectors generated over wiping in five trials (600 ms duration trials, 250 Hz sampling rate; thus 750 samples total) are shown for each condition: with feedback and correction response (a) and without cutaneous feedback or a correction response (b). The derived corrective vectors are shown in (c). Each bin represents 12°. See (d) for orientation of the circular histograms to the frog's body and wiping limb. (a) With feedback intact, vector directions exhibited two clusters: an initial cluster directed toward the midline and a second cluster directed away from the frog. The first cluster was broadly spread and spanned four bins. (b) Without cutaneous feedback, the initial cluster of vector directions was more focally distributed to the midline, that is, mainly in one bin (78 –90°), whereas the second cluster of forces was similar to that with feedback intact (252–264°). (c) Corrective vectors were obtained by subtracting time-matched forces in (b) from (a). The directions of the corrective vectors clustered tightly in a single bin from 108° to 120° and were directed to the hip joint. Thus, the corrective vectors summed with the initial set of wiping vectors directed to the midline and resulted in the rostral spread of these midline-directed force clusters when feedback was intact. Panel 2: cutaneous reflex pathways signaling obstacle contact recruited a corrective force field that summed with an underlying force field sequence produced during wiping. Time series of force fields are shown after initiation of wiping both with feedback intact and after cutaneous deafferentation. A typical electromyograph (EMG) pattern of representative muscles associated with field generation is shown to the right of each time series. For the fields, a single vector represents the force generated at that position at the specified latency after wiping initiation. The black dot marks the location of the hip joint. The lightly shaded ellipse indicates the region of contact with the target limb in free limb trials. (a) With feedback intact, force fields were initially directed to the midline, where target limb contact normally occurred (bold line overlaid on the 130 ms field shows the unrestrained ankle path up to limb contact. At 210 ms up to 330 ms, force fields rotated rostrally toward the hip and away from the target region. At 330 ms, force fields began to rotate back to the target region. The hip extensor burst began around this time (see EMG record to the right; rectified EMGs from BI knee flexor, RI hip extensor, and IL/RA-hip flexors are shown for a single trial taken from the middle of the workspace. EMGs are in millivolts). (b) After removal of cutaneous feedback, initial force fields were directed to the midline. Subsequent fields (210–330 ms) converged to the target region and did not rotate to the hip joint. The strong IL/RA activation was absent without cutaneous feedback (see EMGs to the right). At 330 ms, the field began to rotate caudally because of hip extensor activation. (c) Procedure used to produce corrective force fields. Time-matched fields in (b) were subtracted from corresponding fields in (a). The procedure was performed at each time point after wiping initiation. (d) Corrective fields are shown at three times, at 40 ms intervals. Corrective fields were structurally similar (similarity measures, 0.95) over a 100 ms period starting shortly after IL EMG onset. These rostrally directed fields summed with the underlying wiping force field patterns. Panel 3: path obstacles evoked force fields in panel 2 causing trajectory correction responses during kinematics of hindlimb wiping movements. (a)–(d) Digitized stick figures of the wiping and target limbs (thigh and calf) and the ankle path of both limbs (traced by a bold line; arrows mark direction of movement along path) are shown after initiation of the wiping reflex in four experimental manipulations. Each stick figure is separated in time from the next by 16.67 ms (1/60 s). Bold stick figures mark the initial limb configurations and their configuration at target limb contact. The frog's body would be above each stick figure. (a) The unobstructed kinematics of the wiping limb with feedback intact consisted of a knee flexion and hip extension that resulted in the ankle moving along a trajectory directed to a midline region of stimulus contact. A final knee extension ('whisk') occurred on stimulus contact. (b) When the wiping limb collided with a path obstacle en route to stimulus contact, a hip flexion was evoked that enabled the limb to clear the obstacle. Once the limb was clear, there was an enhanced knee flexion. Target limb contact occurred at a similar midline site and at a time similar to unobstructed trials. (c) After local removal of cutaneous feedback from the wiping limb, unobstructed free limb kinematics showed no apparent differences from the kinematics seen with feedback intact (compare (a) and (c)). (d) When obstacles were placed in the path of the locally deafferented limb, corrective responses were no longer evoked, and the limb was driven further into the obstacle (compare (b) and (d)). BI, biceps femoris (Ecker) or iliofibularis; RI, rectus internus; IL, iliopsoas, RA, rectus anticus. From Kargo WJ and Giszter SF (2000) Rapid corrections of aimed movements by combination of force-field primitives. *Journal of Neuroscience* 20: 409–426.

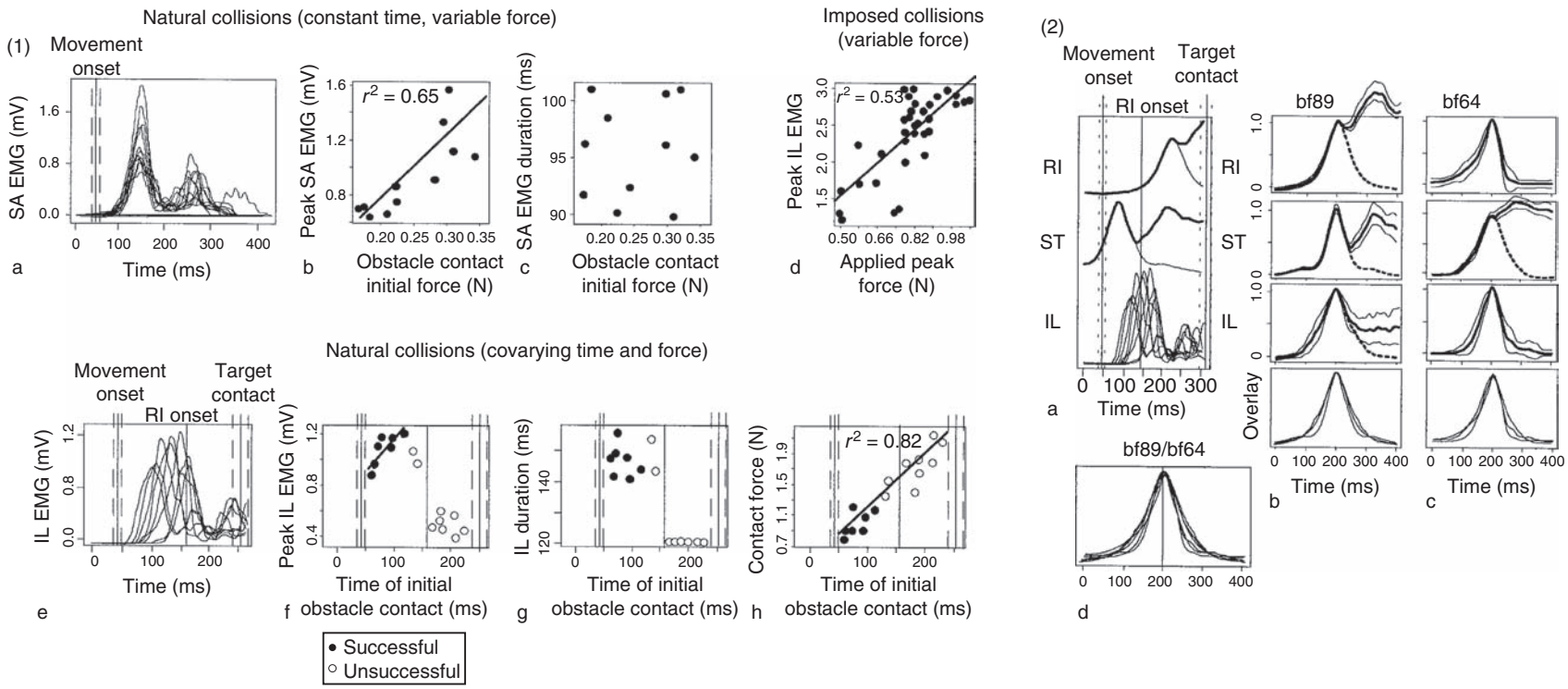

Figure 7 Continued

tested. Adding a dilation term ω to each pulse is supported by motor pattern decompositions in humans (see the section title 'Support in humans.'):

$$F(q, \dot{q}, t) = \sum_i A_i a(\omega_i t + \tau_i)\Phi_i(q, \dot{q}) \qquad [6]$$

Construction of Primitives: How Do Primitives Arise?

Several methods of resolving motor control redundancy issues *de novo*, via optimization approaches, have been developed by Sanger and Todorov, Gharamani, and Jordan, and specific examples have been elaborated by Ting, by Berniker, and by Haruno and Wolpert. Probably the major issue in relation to thinking about spinal primitives, besides whether they account for the spinal organization observed, is the extent to which such optimizations may operate transiently on the short timescales of movement construction in adults and the extent to which they are deeply embedded in the neural organization by development and by evolution.

Current Approaches to Spinal Motor Primitives

Support in Lower Vertebrates

Experiments geared to examine motor primitives have been performed, most specifically in the frog. However, earlier and current work in the turtle, which is couched in modular pattern generator terms, corresponds closely. Both frog and turtle show occasional modular deletions of synergy bursts. Often these deletions do not cause any alteration of other activity or change the ongoing timings or rhythmic structure. These spontaneous deletions are consistent with superposition of modular elements or primitives downstream from rhythmogenesis and phase synthesis.

In the frog, a trajectory correction primitive has been identified (Figure 6). It represents one of the set

Figure 7 Pulsatile organization of motor primitives. Motor patterns in frogs are organized from almost fixed duration pulses. This is demonstrable in wiping corrections. Panel 1: correction response amplitude was regulated both by motor phase and by feedback arising from obstacle contact. (a) Rectified and filtered SA electromyographs (EMGs) are overlaid for multiple trials ($n = 11$) in which obstacle collision occurred at a fixed latency. For each trial, obstacle collision occurred within 16.67 ms after movement onset (the mean time of movement onset and the standard deviations (SDs) from the mean are shown as bold and dotted vertical lines, respectively). Collision occurred slightly after the second dotted line. Evoked responses represent scaled versions of one another. (b) The peak magnitude of the evoked SA EMG is plotted versus the highest contact force before SA onset as the limb collided with the obstacle. The peak SA EMG was significantly correlated with the highest preceding contact force ($p < 0.05$). This relationship could indicate regulation of corrective strength by contact force or covariation of corrective strength with wiping motor pattern strength. (c) SA EMG duration at half peak showed no correlation with the magnitude of obstacle collision. (d) In a second frog held under isometric conditions, impacts of varying strength were generated with a stepping motor to accelerate a mass into the normal site of obstacle contact. Rectified and filtered peak IL EMG ($n = 34$ trials) significantly covaried with the applied force ($p < 0.01$) independent of motor program strength, strength of correction, and time of collision. (e) The rectified and filtered EMGs of evoked IL response are shown for multiple trials ($n = 12$) in which obstacle collision occurred at different distances (and therefore times) along the movement path of the wiping limb. The IL EMG was evoked at progressively longer latencies from movement onset because contact occurred at later times. (f) The peak magnitude of the evoked IL EMG is plotted vs. the time of obstacle contact. Successful responses (f) were positively correlated with the time of obstacle contact ($r^2 = 0.73$). After RI onset (second solid vertical line), IL responses were no longer evoked. Close to RI onset, some evoked correction responses failed to successfully clear the obstacle (e). Evoked IL responses had reduced amplitudes in these trials, departing from the linear relationship with impact force. (g) The duration of evoked IL responses showed no correlation with the time of obstacle contact and varied by <10 ms. (h) The contact force showed a significant linear correlation with the time of obstacle contact ($p < 0.001$). Panel 2: the activation waveform of EMGs comprising the correction response had a shape and duration that were similar to the activation waveforms of other muscles' EMGs in different phases of the wiping motor pattern. (a) In the top two rows, the averaged EMG waveforms of ST and RI from movement onset up to target contact are shown. In the bottom row, the individually evoked IL responses are shown (from panel 1). IL responses were evoked by obstacle contact that occurred at different times in the interval between movement onset and RI onset. This was a 100 ms time window. The averaged time of movement onset is marked by the first set of vertical lines; dotted lines represent 1 SD about this. (b) The averaged EMG waveforms of ST, RI, and the corrective IL burst during trajectory correction trials ($n = 11$) for the data in (a) are shown in the top three rows. Averages were constructed by aligning individual EMGs from each trial ($n = 11$) at their initial peak value (ST and RI exhibit multiple bursts in the wiping motor pattern) and normalizing this peak to a value of 1 before averaging the resultant waveforms. SDs are shown around the averaged EMGs. The dotted lines indicate the inferred relaxation of the activation waveform in the absence of the later, overlapping phases (note, however, that in frog BF64 in (c), the relaxation of RI can be directly observed). These were obtained by assuming waveform symmetry and reflecting the rising phase of the EMG. In the bottom row, the three averaged EMG waveforms for RI, ST, and IL are overlaid. These waveforms have a similar shape and duration despite occurring at different phases of the motor program, and all lie within 1 SD of the IL waveform. (c) The same averaging procedure was performed in a second frog. Individual EMG waveform averages were similar and showed low variance. Overlaid EMG waveforms have a similar shape and duration. (d) The ST, RI, SA, and IL EMG waveforms for the two frogs shown in (b) and (c) are overlaid. The EMG waveforms in these spinal frogs all exhibit similar shape and duration. The data suggest that different motor primitives may all have common activation dynamics even when recruited in different phases and combinations during reflex behaviors. SA, sartorius; RI, rectus internus; IL, iliopsoas; ST, semitendinosus outer head. Data from Kargo WJ and Giszter SF (2000) Rapid corrections of aimed movements by combination of force-field primitives. *Journal of Neuroscience* 20: 409–426.

of six primitives which is added to the motor pattern following a collision. This addition acts to modify motion so as to circumvent an obstacle. Superposition of primitives is thus used in a contingency fashion by the spinal cord. Remarkably, as a result of the primitive addition, the limb reaches the target region within 50 ms of the unperturbed arrival time. In addition, the spinal cord recruits the primitive only following a collision during the period of motion when superposition can achieve this. The primitive addition initiates a hip flexion. Collisions can no longer recruit the primitive as soon as the more powerful hip extensor muscles begin being recruited into the pattern (Figure 7, panel 1). The spinal cord thus appears to assemble and use these elements based on their potential utility in achieving the task goal and in some way represents these contingent relationships. The degree to which this result should be thought of as representing the existence of a predictive internal model of the task is currently unclear, but it is an interesting issue.

Corrections in the frog are pulsatile and of fixed duration (Figure 7), and this observation leads directly to the idea that the frog uses fixed-duration pulses of activity more broadly. Indeed, the spinal motor primitives and the associated force fields in frog appear to have a fixed time course. Adjustments in phase and amplitude of the premotor drive pulses are thus the primary method of movement construction and adjustment in frogs.

This perspective can be confirmed by detailed decomposition of the electromyogram motor patterns of frog into premotor drives and their driving pulses (similar to eqn [5] above). This was accomplished using independent components analysis and wavepacket analysis. The premotor drive pulses can be observed directly in the electromyograms of the correction behavior but are extracted statistically from the whole repertoire of behavior. The drive structures found across different behaviors imply neural and sensory projection patterns and properties. The sensory control of the pulse amplitudes and timing of pulses, as well as the neural underpinnings, are presently being investigated.

The competence of a few modules to capture the electromyogram patterns of 'intact frogs' is now also established. In the intact frog, the method of nonnegative matrix factorization has been used. Time-varying synergies were extracted rather than pulsed muscle synergies. However, it is likely that these represent 'chunking' of simpler pulses in sequences since no additional and sometimes less variance was accounted for by time-varying synergies, compared with the other methods.

Supporting Data in Mammals

Many of the same experiments conducted in frog and turtle have also been conducted in mammals. Tresch and Bizzi showed a set of a few force fields or spinal motor primitives can be elicited from the isolated unanesthetized rat spinal cord (Figure 8). These had properties similar to those of frog: fixed force structures scaling in amplitude with stimulation. The fields had major actions in the parasagittal plane, consistent with roles in locomotion. Vector superposition was demonstrated. In addition, the fields were also shown to be recruited from sets of chronically spinalized and deafferented frogs and so did not simply represent afferent or descending projection patterns.

Lemay and Grill used the same paradigm as in frog and rat in the decerebrate cat or chloralose anesthetized cat. They also showed a modularity of a small number of fields (Figure 9). However, winner-take-all interactions, which were seen in only 20% of frog costimulations, occurred in all sites examined in their data.

Corrections, omissions, and deletions in the cat have been demonstrated by McCrea and colleagues in fictive preparations (Figure 10). Like the phenomena in frog and turtle, these data comprise additions or deletions of bursts, often without perturbation of the remaining pattern. This behavior is difficult to achieve readily with a nonhierarchical pattern

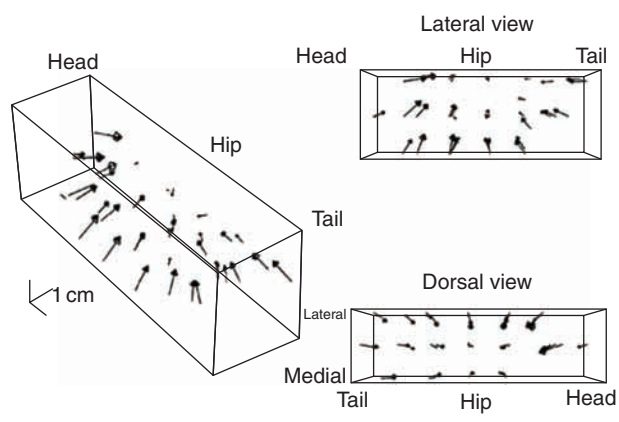

Figure 8 Spinal motor primitives in mammals: the rat. An example of a force field measured in the three-dimensional workspace of the rat hindlimb. Forces evoked from microstimulation in the spinal cord were measured at ankle positions in three sagittal planes, one directly below the hip, one 1 cm lateral to the hip, and one 1 cm medial to the hip. Three views of the force field are shown here. In addition to the perspective view shown on the left, a lateral view and a dorsal view are shown. This response was obtained from microstimulation of a site at a depth of 750 μm with a current of 8 μA in L5. From Tresch MC and Bizzi E (1999) Responses to spinal microstimulation in the chronically spinalized rat and their relationship to spinal systems activated by low threshold cutaneous stimulation. *Experimental Brain Research* 129(3): 401–416.

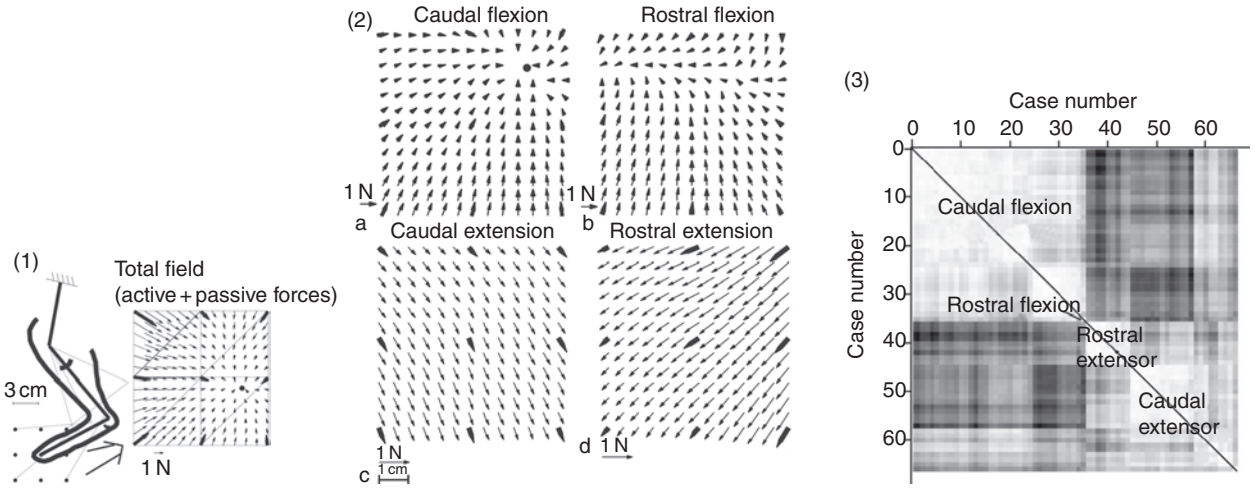

Figure 9 Spinal motor primitives in mammals: the cat. Panel 1: field measurements in the cat used in Panels 2 and 3. Construction of a force field from measurement of isometric force vectors at the endpoint of the limb. Left: With the femur held fixed via pins, the endpoint was moved on a 3 x 3 cm grid where the dots indicate the spatial locations where forces were recorded. Thin lines indicate orientation of the limb at different endpoint positions. Right: measured force vectors (thick dark arrows), triangles dividing the workspace, and the interpolated force vectors (thin light arrows). Forces represented are total forces evoked by intraspinal microstimulation (active and passive) minus forces due to gravity. The field exhibited convergence, that is, net endpoint forces went to 0 at a location in the workspace indicated by the black circle. Panel 2: four types of active force field evoked at the endpoint of the hindlimb by intraspinal microstimulation. (a) Caudal flexion (depth, 3400 μm); (b) rostral flexion (depth, 3400 μm); (c) rostral extension (depth, 4600 μm); and (d) caudal extension (depth, 5000 μm). Panel 3: dissimilarity matrix of grouping of endpoint force patterns. Matrix of the squared euclidean distance was arranged so that fields in the same cluster are adjacent to one another in rows and columns. Lighter squares indicate smaller distances, and darker squares indicate larger distances. Lighter squares along the diagonal represent clusters. Caudal flexion responses extend from cases 0 to 34, rostral flexion is represented by a single square, rostral extensors extend from cases 36 to 45, and caudal extensors extend from cases 46 to 68. Note how the caudal extensor square along the diagonal is composed of two squares corresponding to two subgroups, 2–48 and 24–43, and similarly the caudal flexion square is formed by two squares corresponding to two subgroups. From Lemay MA and Grill WM (2004) Modularity of motor output evoked by intraspinal microstimulation in cats. *Journal of Neurophysiology* 91(1): 502–514.

generator. Using modeling and data, McCrea and colleagues suggested a rhythmogenic part of the CPG that synthesizes gait patterns, and a pattern-shaping layer that can add or subtract burst modules as demanded by task constraints and events, without necessarily altering the normal pattern. This latter is similar to a collection of motor primitives, and their model organization is reminiscent of recent ideas for organizing kinematic primitives in biomorphic robots advanced by Schaal and colleagues.

Ting, MacPherson, and colleagues have shown that a small number of muscle synergies account for phasic postural adjustments in the cat and that the structure and recruitment patterns of these may also closely resemble the patterns predicted by optimization schemes.

Support in Humans

Predating most work on spinal motor primitives, in an analysis of human gait, Vaughan showed that factor analysis of human electromyograms allowed their representation by a reduced set of four to six muscle groups. Zajac and colleagues also attacked the issue in cycling. More recently, Ivanenko, Poppele, Lacquaniti, and colleagues have very elegantly used principal components/factor analysis to clearly demonstrate modularity of human locomotion (Figures 11 and 12). The modular elements they identified can be related to both pattern generation and motor primitives and are likely to have a strong spinal component. Their data show conservation of most aspects of the modular elements across many voluntary tasks incorporated in walking or running, such as kicking a soccer ball. Their elements, while all similar in duration within a step cycle, are also modulated in duration along with the step-cycle duration and in phase depending on the activity (Figure 12). Thus the bursts can be dilated or compressed in time as a function of cycle period, as the locomotor cadence changes. Leg muscle synergies or balance and composition within bursts showed small adjustments but also remarkable consistency.

Future Issues in Spinal Motor Primitives

Development

The development of motor primitives and their subsequent plasticity are important issues. Are these

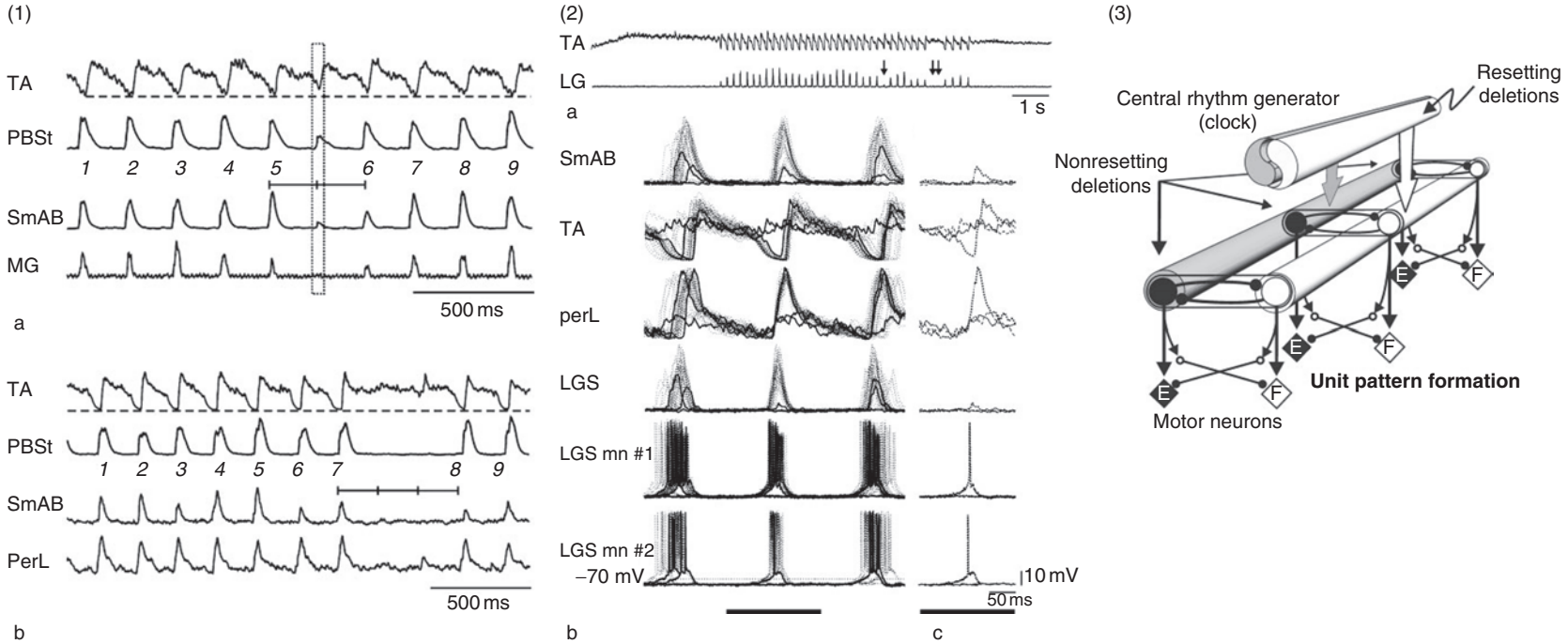

Figure 10 Deletions of parts of motor patterns as units in the decerebrate cat. Panel 1: deletions of extensor activity during fictive scratch in the decerebrate cat. (a) Dotted box indicates where, after the five rhythmic bursts illustrated, the medial gastrocnemius nerve became inactive (a complete deletion). During this period the posterior biceps and semitendinosis (PBSt) and semimembranosis and anterior biceps (SmAB) bursts were reduced in amplitude (partial deletions). Ankle flexor activity (tibialis anterior; TA) was modulated but did not return to baseline. Intervals on the horizontal bar were calculated from the immediately preceding five cycle periods (197 ms, SD 2). Interval spanning the deletion was not significantly different from control two-cycle intervals ($p > 0.40$). (b) Another example of a deletion of activity during fictive scratch in the hip extensor (SmAB) as well as the bifunctional PBSt and the ankle flexor (peroneous longus; PerL) nerves. This deletion is accompanied by sustained activity in another ankle flexor (TA). Scratch cycle period indicated on the horizontal bar was 204 ms (SD 14). Rhythmic bursts after the deletion reappear at an interval corresponding to three control step-cycle periods (three vertical tick marks). Interval spanning this deletion was not significantly different from three-cycle intervals taken from the control period ($p > 65$). Panel 2: maintenance of cycle period timing during deletions of fictive scratch activity. (a) A bout of fictive scratch in which there were three deletions (arrows) of extensor motor neuron (lateral gastrocnemius; LG) activity. (b) Overlay of all the cycles of rhythmic activity in an episode of fictive scratch. Top four traces are rectified and integrated electroneurographic records, and the bottom two traces are simultaneous recordings from a pair of LG and soleus (LGS) motor neurons. Records of normal activity (dotted lines) are overlaid with those obtained during extensor deletions (solid lines). Deletion interval was not significantly different from two-cycle intervals in the control period ($p = 0.95$). (c) Portions of the same traces during partial deletions (dotted lines) overlaid with the corresponding trace during the complete deletion (solid line). Panel 3: schematic representation of the proposed architecture for the central pattern generator, which consists of separate networks for rhythm generation (clock) and for pattern formation (i.e., the distribution of excitation and inhibition to motor neuron pools). Pattern formation modules include circuitry for reciprocal inhibition of antagonist motor pools and control the activity of subsets of motor neurons. Deletions that include a change in the locomotor rhythm are indicated by the arrow affecting the rhythm generator network. Deletions in which cycle timing is maintained are hypothesized to occur by disturbances at the level of the pattern formation networks and may affect individual pattern formation networks differentially. E, extensor; F, flexor. Data from Lafreniere-Roula M and McCrea DA (2005) Deletions of rhythmic motoneuron activity during fictive locomotion and scratch provide clues to the organization of the mammalian central pattern generator. *Journal of Neurophysiology* 94(2): 1120–1132.

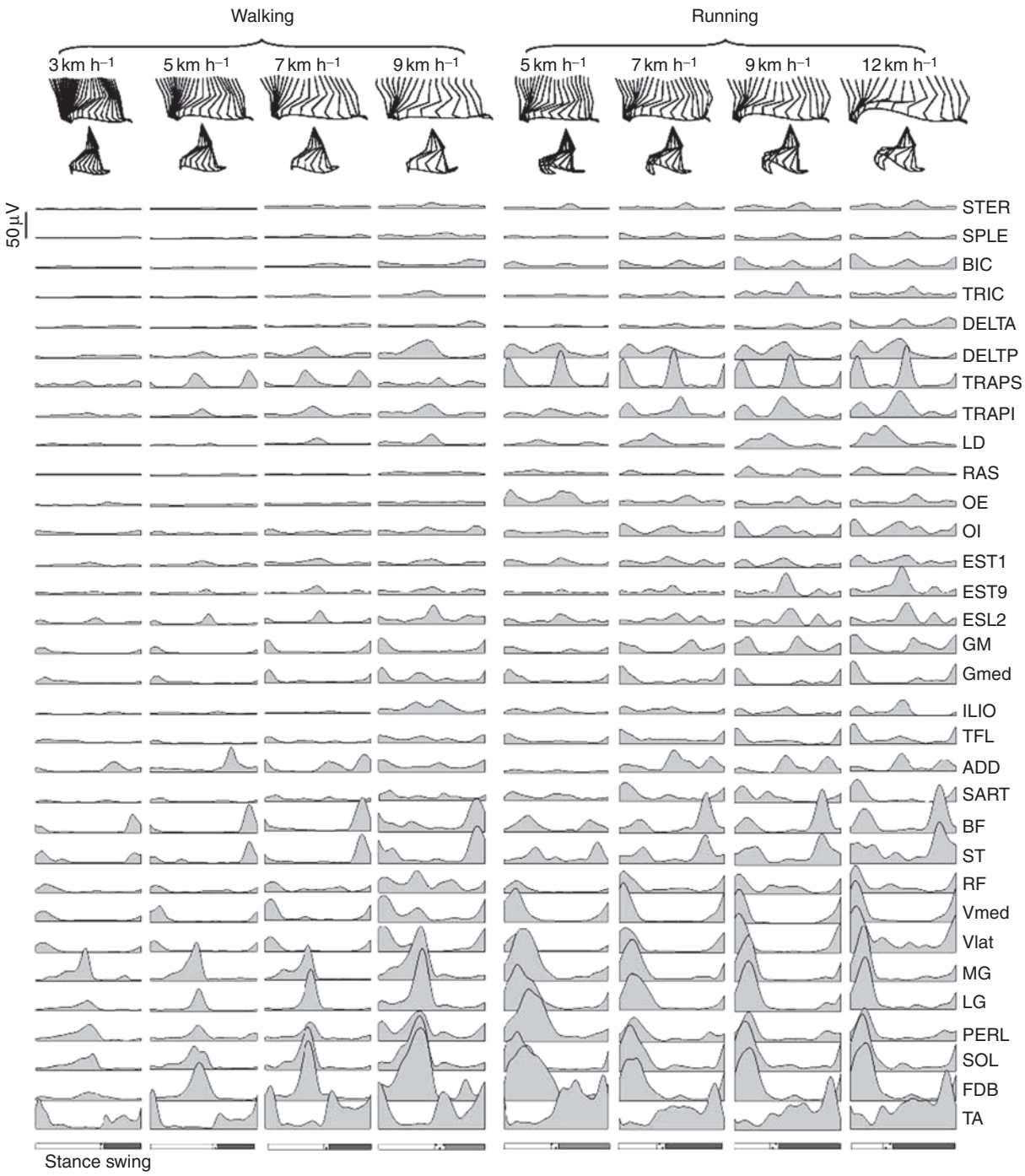

Figure 11 Motor patterns in human locomotion. Ensemble averaged activity patterns recorded from eight participants for 32 muscles. Electromyographic records were filtered with a low-pass cutoff of 10 Hz. Top: stick diagrams of locomotion stepping. Bottom: EMG Data were plotted vs. a normalized cycle. As the relative duration of stance varied across participants, a hatched region on the bar below indicates an amount of variability in the stance phase duration across participants. These complex pattern data form the basis for the decomposition and analysis in **Figure 12**. Data from Cappellini G, Ivanenko YP, Poppele RE, and Lacquaniti F (2006) Motor patterns in human walking and running. *Journal of Neurophysiology* 95(6): 3426–3437.

elements hardwired or self-organizing, and are they long-term stable entities or transient neural organizations? Specific work on spinal motor primitives is currently lacking. However, *in ovo* and *in utero* and subsequent neonatal motor development of reflexes and pattern generation are the subject of intense research by several laboratories, such as Kiehn, Clarac, Bradley, Stehouwer, and Schouenborg. Initial motions

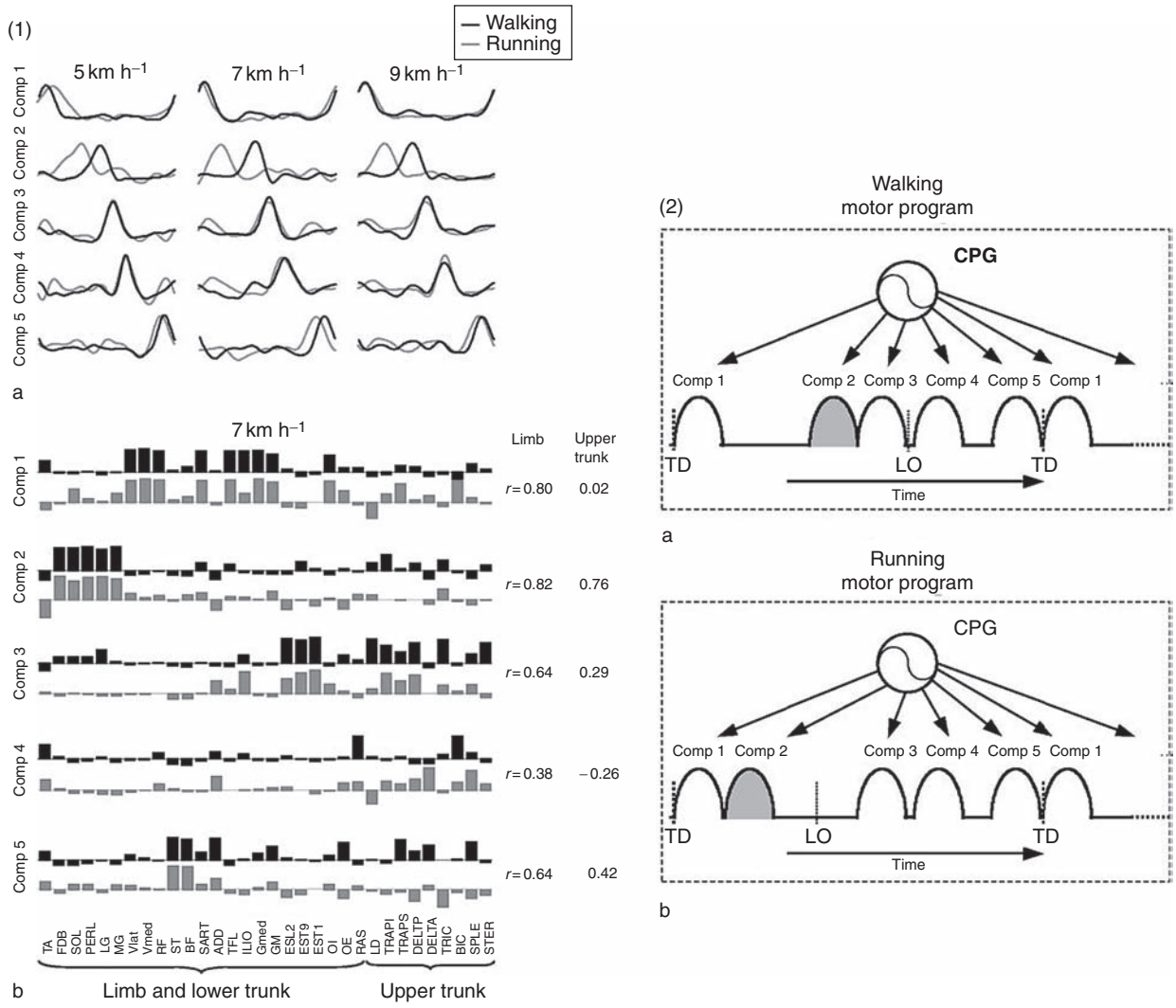

Figure 12 Motor pattern modularity in human locomotion. All walking and running motor patterns in the limb can be simply expressed as a collection of five appropriately timed pulses, which represent coordinated recruitment of muscles as a group. Pulse durations scale to the locomotor step-cycle duration. Many muscle groups represented in the pulses are also stable across the conditions, although there are some variations (e.g., component 4 in panel 2). This scheme somewhat resembles the scheme in **Figure 10** and the formulation in eqn [6] in the text. From Cappellini G, Ivanenko YP, Poppele RE, and Lacquaniti F (2006) Motor patterns in human walking and running. *Journal of Neurophysiology* 95(6): 3426–3437. Panel 1: activation temporal components derived from averaged muscle activity patterns (see **Figure 11**). (a) Superimposed activation components in walking and running. Note a prominent phase shift of the timing of the main peak of the second component in running. (b) Weighting coefficients (across 32 muscles) of the five basic temporal components (comps) during walking and running at the same speed (7 km h^{-1}). The correlation of weighting patterns between walking and running is shown at the right of each record for both the leg/lower trunk (tibialis anterior (TA) – rectus abdominis, superior portion (RAS)) and upper trunk/shoulder (latissimus dorsi (LD) – sternocleido mastoideus (STER)) muscles separately. Panel 2: the authors' interpretations are based on the data of the original study. Hypothetical motor programs for walking (a) and running (b) in terms of the characteristic timing of muscle activations and their pulsed activity. In both programs, four activation components (comp 1, 3, 4, and 5) occur relative to the touchdown (TD) phase of the cycle, and comp 2 occurs relative to the liftoff (LO) phase, at about 20% of the cycle preceding LO. As one gait mode is switched to the other, the timing of LO relative to TD changes along with the relative timing of comp 2 with respect to the other four comps. CPG, central pattern generator. Data from Cappellini G, Ivanenko YP, Poppele RE, and Lacquaniti F (2006) Motor patterns in human walking and running. *Journal of Neurophysiology* 95(6): 3426–3437.

both *in ovo* and *in utero* are largely unconstrained other than by viscosity and limb inertia. Subsequently the embryo becomes more confined and the limbs become loaded by the egg or uterine wall. Thus some opportunity for both kinematic and kinetic experiences of self-generated motion exist in a protected and simplified environment. However, much of the basic patterning seen in neonates can be developed without descending influences and without sensory feedback. The hardwiring of motor pools and primary afferent patterns is consistent with the possibility of hardwired spinal motor primitives. However, proceeding more

centrally, such hardwired structures are likely to be replaced with far more flexible topologies and far more adjustable synaptic weights. Schouenborg and colleagues have shown how distal reflexes can become exquisitely refined. Where primitives reside on this continuum is not yet known.

Plasticity and Recovery after Injury

The role of motor primitives in recovery after injury, rehabilitation, and neuroprosthetics is likely to depend on the relative stability of primitives. Infant motor patterns such as grasp, routing, and Babinski sign reveal themselves intact after adult injury, suggesting long-term stability. However, it has also been established by Wolpaw and colleagues that the spinal circuitry is subject to continuous adjustment by corticospinal and other mechanisms. Hogan, Krebs, and colleagues, examining stroke recovery, have shown that a set of coarse kinematic submovements in point-to-point movement reveal themselves early following stroke and then become blended or coarticulated through recovery. They are also examining the electromyographic basis for this process.

Relationship to Other Ideas about Spinal Organization

The detailed relationships of pattern generators, unit burst generators, pattern shaping systems, and the spinal motor primitives are still an area of investigation. In particular the neural underpinnings, circuitry, and processes that give rise to the modular separability at the level of function are not well understood despite the great wealth of excellent research on spinal circuits.

Relationship to Motor Learning and Evolution

From a Bayesian and evolutionary perspective, low-level execution primitives are useful only insofar as they adequately capture *a priori* expectations of the useful movement types and the adaptations needed (e.g., as in the wildebeest example). At the same time, they should be robust to and adapted to any structural variations in the musculoskeletal system throughout life. The best compromise between these potentially competing requirements of adaptive flexibility and long-timescale hardwiring of motor structures is not currently well understood. If primitives represent near optimal decompositions of tasks, as some suppose, are they organized transiently by spinal cord and brain, are they organized in restricted critical periods, or are they hardwired into development by evolution?

How spinal motor primitives are incorporated into learned behaviors and the extent to which they constrain early learning or skilled behaviors are not understood. The relations of learning basis functions, which have also been termed 'primitives' by Thoroughman and Shadmehr, to spinal motor primitives is unclear. Bracha and Bloedel have shown a strong influence of cerebellar adaptation processes on spinal reflex elements, supporting their use as a partial basis for organizing learning. Work of Bastian on human locomotor adaptation further supports this notion.

It is worth noting with regard to the developmental, learning, and theoretical issues here that the biological motor apparatus is a structure with a long evolutionary history. Coadaptation of mechanical plant and control are almost certain to have occurred. These might be extremely well tuned. However, both the mechanical structure and its controls may also be rife with historical 'appendices' and 'spandrels.' There may be a historically important structure with better algorithms overlaid on the cruder and older competences. An ideal *de novo* engineered structure might bypass this potential mass of 'legacy' structure, but the biology may be stuck with it, at least on the timescale of modern science!

Conclusions

Motor primitives laid down by evolution simplify early motor control in developing organisms, seed motor learning, and allow rapid adapted reflex trajectories and stepping adjustments. Evidence for such modularity now exists across limbed vertebrates. These spinal motor primitives may be equivalent to the keyboards on which voluntary and pattern generator systems play out motor patterns and with which they initially synthesize novel motion.

See also: Bayesian Models of Motor Control; Central Pattern Generators: Sensory Feedback; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Kinematics and Dynamics; Pattern Generation; Reflex Circuits; Segmentation: Spinal Cord Segmentation and A–P Somite Patterning; Sensorimotor Integration: Models.

Further Reading

Bizzi E, Tresch MC, Saltiel P, and d'Avella A (2000) New perspectives on spinal motor systems. *Nature Reviews of Neuroscience* 1(2): 101–108.

Cappellini G, Ivanenko YP, Poppele RE, and Lacquaniti F (2006) Motor patterns in human walking and running. *Journal of Neurophysiology* 95(6): 3426–3437.

Flash T and Hochner B (2005) Motor primitives in vertebrates and invertebrates. *Current Opinion in Neurobiology* 15(6): 660–666.

Giszter SF, Mussa-Ivaldi FA, and Bizzi E (1993) Convergent force fields organized in the frog spinal cord. *Journal of Neuroscience* 13: 467–491.

Kargo WJ and Giszter SF (2000) Rapid corrections of aimed movements by combination of force-field primitives. *Journal of Neuroscience* 20: 409–426.

Lafreniere-Roula M and McCrea DA (2005) Deletions of rhythmic motoneuron activity during fictive locomotion and scratch provide clues to the organization of the mammalian central pattern generator. *Journal of Neurophysiology* 94(2): 1120–1132.

Lemay MA and Grill WM (2004) Modularity of motor output evoked by intraspinal microstimulation in cats. *Journal of Neurophysiology* 91(1): 502–514.

Mussa-Ivaldi FA and Bizzi E (2000) Motor learning through the combination of primitives. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 355(1404): 1755–1769.

Mussa-Ivaldi FA, Giszter SF, and Bizzi E (1994) Linear combination of primitives in vertebrate motor control. *Proceedings and National Academy of Sciences of the United States of America* 91: 7534–7538.

Poppele R and Bosco G (2003) Sophisticated spinal contributions to motor control. *Trends in Neurosciences* 26(5): 269–276.

Rohre B, Fasoli S, Krebs HI, et al. (2002) Smoothness during stroke recovery. *Journal of Neuroscience* 22(18): 8297–8304.

Schaal S and Schweighofer N (2005) Computational motor control in humans and robots. *Current Opinion in Neurobiology* 15(6): 675–682.

Schouenborg J (2004) Learning in sensorimotor circuits. *Current Opinion in Neurobiology* 14(6): 693–697.

Stein PS (2005) Neuronal control of turtle hindlimb motor rhythms. *Journal of Comparative Physiology A Neuroethology, Sensory and Neural Behavior Physiology* 191(3): 213–229.

Tresch MC and Bizzi E (1999) Responses to spinal microstimulation in the chronically spinalized rat and their relationship to spinal systems activated by low threshold cutaneous stimulation. *Experimental Brain Research* 129(3): 401–416.

Wolpert DM, Ghahramani Z, and Flanagan JR (2001) Perspectives and problems in motor learning. *Trends in Cognitive Science* 5(11): 487–494.

Motor Psychophysics

C A Buneo, Arizona State University, Tempe, AZ, USA
J F Soechting, University of Minnesota, Minneapolis, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

The motor plan for an arm movement to a target is ultimately defined by the distance and the direction from the hand to the target. Most commonly, such movements are effected to a seen target and, accordingly, the target's location is defined by its projection onto the retina. That is, the target's location is initially specified in a retinotopic frame of reference. Conversely, vision of the arm is not needed to make a reasonably accurate movement and the hand's location can be derived from proprioception. Thus, the problem of making goal-directed arm movements is one of combining information across sensory domains. This process of sensorimotor integration has been viewed as a mapping of information into a common frame of reference. Hand location derived proprioceptively is defined in a frame of reference fixed to the body, and because the eyes can move in the head and the head can move relative to the body, the retinotopic and proprioceptive frames of reference will generally not be aligned.

Initial behavioral experiments that addressed this problem led to a model in which visual information was mapped into the proprioceptive (body-fixed) frame of reference and a movement vector was defined in this coordinate system. However, recent behavioral experiments and electrophysiological recordings have refined this model to incorporate additional steps in the formulation of a movement plan. In this model, hand location is initially remapped into a retinocentric (eye-fixed) frame of reference and the initial movement vector is defined in this frame of reference. At subsequent stages, this movement vector is ultimately remapped into an arm-fixed frame of reference. This model appears counterintuitive because it introduces additional coordinate transformations and, accordingly, potentially more sources of error. However, as argued here, such a process would facilitate visually mediated, online error corrections – corrections that would be especially important when there is motion of the body or the target.

How Is the Motor Plan Defined?

The arm movement vector is defined by the location of the target relative to the hand. There are several lines of evidence to support this contention. The first is that the arm posture at the end of a reaching movement depends on the initial arm posture. The arm (neglecting the wrist) has a redundant degree of freedom since there are three degrees of freedom of motion at the shoulder joint and one at the elbow joint, but only three degrees of freedom are required to define the location of the hand in space. The extra degree of freedom can be characterized simply by the plane defined by the upper arm and the forearm. It has been found that the arm plane at the target depends in a predictable and consistent manner on the initial arm posture.

Second, electromyographic activity in arm muscles is tuned to the direction of movement. Generally, the pattern of muscle activity for whole arm movement has roughly the same characteristics as the pattern for simpler movements restricted to a single joint: an initial agonist burst for acceleration followed by an antagonist burst to decelerate the motion. The amplitude of these bursts shows directional tuning in that it decreases by an amount proportional to the cosine of the angle between the direction of motion and the 'best' direction of that particular muscle. Furthermore, neuronal activity in primary motor cortex (M1) is also tuned with respect to the direction and the amplitude of the arm movement.

The distance from the hand to the target is independent of the reference frame in which this vector is described. However, this is not the case for the movement's direction. Two frames of reference come to mind readily: an extrinsic one (perhaps fixed to the body) and another that is fixed to the arm. An extrinsic frame of reference is motivated by perception: we commonly define movements with respect to the cardinal axes, such as 'straight ahead' or '45 degrees to the right.' However, the mechanical actions of muscles depend on the posture of the arm and, accordingly, the tuning of arm muscle activity will change (for the same direction of movement defined in trunk-fixed coordinates) when the posture of the arm changes. In fact, early experiments suggested that the tuning of neurons in M1 was constant in an arm-fixed frame of reference.

Recent experimental results, however, point also to a frame of reference that is intermediate to these two – that is, a frame of reference that rotates by half of the amount of arm rotation. The mechanical actions of muscles at the shoulder and at the wrist have been found to follow such a 'half-angle' rule, as have the extraocular muscles of the eye. Furthermore, a considerable number of neurons in the wrist area of M1 show tuning that is 'muscle-like' in that their

preferred direction changed with wrist orientation (pronation–supination). However, half of the neurons in M1 show tuning in a trunk-fixed frame of reference.

How Is Hand Location Defined?

Although vision can be used to define the initial location of the hand, proprioceptive information is generally used for this purpose. The strongest evidence that proprioception is needed to encode the location of the hand in space comes from the fact that patients with proprioceptive deficits resulting from large-fiber sensory neuropathy exhibit large directional errors in hand movements. However, vision also contributes to the encoding of hand location since these directional errors were reduced when these patients could see their arm prior to the onset of the movement. Movement errors can also be induced in normal subjects if the apparent visual location of the hand relative to the target is altered. Behaviorally, subjects typically foveate the target prior to movement onset, rather than the hand, and it appears that vision contributes primarily indirectly by calibrating proprioceptive information about hand location. Visually mediated feedback of hand location does not appear to be prominent during arm movements. In fact, when executing movements to a sequence of targets, subjects typically saccade to the subsequent target before the hand arrives at the previous one.

How proprioceptive information from the muscles and skin is processed to derive the location of the hand in space is not well understood, either behaviorally or electrophysiologically. Behavioral experiments suggest that arm posture is perceived with respect to cardinal axes (vertical and parasagittal) as an angular elevation and yaw of each of the limb segments.

Coordinate Transformations Deduced from Psychophysical Studies

The question of how the location of the hand and that of the target are represented has been addressed behaviorally by analyzing the pattern of errors in pointing movements executed under open-loop conditions. In such experiments, subjects are presented with a target, asked to remember its location, and then, after the target is removed, asked to point to the remembered location. Constant errors (the average error for a particular target location) and variable errors (the variance about this mean) have been analyzed.

Errors (constant and variable) were not homogeneous in three-dimensional space. Furthermore, they depended on the location of the target relative to the subject, but neither on the starting location of the hand nor on the distance from the hand to the target. The hand's location at the end of the pointing movement tended to match the target's direction to a much better degree than its distance from the subject; directional errors were much smaller than distance errors. The initial experiments were performed in the dark. In this condition, distance errors in pointing depended only on the target's distance when distance and direction were measured from the shoulder. They were skewed when an eye-centered coordinate system was used. Thus, the pattern of errors suggested that target location was encoded in a spherical coordinate system that was shoulder centered. This interpretation implied a transformation of the representation of target location from an eye-fixed to a body-fixed frame of reference (with an intermediate head-fixed representation because the eyes can move relative to the head). It was suggested that using a shoulder-centered spherical coordinate system simplified the mapping of target location into intended arm posture by means of simple, linear transformations that were only approximately correct and led to the observed errors.

Other experiments were consistent with this interpretation. Subjects did not make errors in distance when they put the tip of a handheld pointer on the remembered target location, implying that the pointing errors did not arise from a misperception of the target's location. Furthermore, they were able to point to a location halfway between the target and the shoulder, a task that is simple if target location is encoded in shoulder-centered coordinates but becomes much more difficult if the location is encoded in eye-centered coordinates. Another analysis that made no *a priori* assumptions about the origin of the coordinate system for target location indicated that it was centered at a point approximately halfway between the shoulder and the eyes in the frontal plane. This estimate was obtained by taking the error vectors on each trial (the difference between target location and hand location) and extrapolating them to determine their point of intersection. Since not all of the vectors intersected at a common point, there was some uncertainty in the estimated location; however, the 95% confidence limits did not include the shoulder or the center of the head.

A similar analysis of results obtained when subjects could view their hand, but not the target, suggested that in this instance the target was represented in a retinocentric frame of reference. Other experimenters, analyzing the pattern of variable errors, arrived at the same conclusion. The major axis of the ellipsoid defining the variance of the errors pointed toward

the head, irrespective of the location of the target or the starting location of the hand. For movements performed in the dark but with the eyes open, the orientation of the major axis of the error ellipse depended on the interval between the presentation of the target and the onset of the movement. For short memory delays, these vectors pointed toward the head, shifting toward the shoulder at longer time delays. Other experimenters introduced errors in the visually perceived location of the finger at one target location and examined how such a visuomotor remapping generalized to other target locations. The remapping was best described in a spherical eye-centered coordinate system.

Typically, subjects foveate the target to which they point. For eccentric targets in a retinocentric frame of reference, subjects overestimate the eccentricity when making a pointing movement in the absence of visual feedback of the target location. Results of experiments that took advantage of this phenomenon also indicated that target location was represented in a retinocentric frame of reference and that this location could be remapped if there was an intervening eye movement.

These experimenters presented subjects with a target straight ahead and asked them to foveate it and to make a saccade to an eccentric location after the target was removed, before pointing to the remembered target location. Directional errors in pointing reflected the eccentricity of the gaze and were indistinguishable from errors in pointing to eccentric target locations in the absence of eye movements.

All of these results suggest that, in the initial stages of processing, the hand movement vector is defined in a retinocentric frame of reference – that is, in the frame of reference in which the target location is initially defined visually. This conclusion implies that the location of the hand is also mapped into this frame of reference, even when vision of the hand is unavailable. In the dark, this representation gradually shifts to a body-fixed, shoulder-centered frame of reference. It also implies that the movement vector is subsequently remapped from an eye-centered frame of reference into one that is more closely aligned with arm posture (i.e., from retinal coordinates to muscle-based coordinates). Electrophysiological evidence from behaving primates is in accord with this conclusion.

Neural Correlates of the Coordinate Transformations for Reaching

The neural correlates of coordinate transformations have been investigated most extensively in the posterior parietal cortex (PPC). Although the initial focus of these studies was on eye movements, reaching movements have recently received the most attention. In these studies, animals made reaching movements under conditions in which both the hand location at the start of the movement and the target location were held fixed in one frame of reference but varied in others. Under such conditions, the criterion for identifying the frame of reference that best captures the neural responses is reasonably straightforward. For example, if neurons encode the target and initial hand location in a retinocentric frame of reference, then the firing rates of these neurons should not vary if hand and target position are held fixed with respect to the direction of gaze, even if gaze direction varies with respect to the body or in space. Similarly, if neurons encode hand and target position in a body-centered reference frame, then their activity should not vary as long as these positions are held fixed with respect to the body, even if body position changes in space.

In the parietal reach region (PRR), neurons encode reach-related information in a manner that is consistent with a retinocentric reference frame. Neural responses in this area are relatively invariant as long as the target location for a reach falls on the same part of the retina. A subset of PRR neurons are also modulated when the initial hand position is changed. Importantly, varying the initial hand position does not alter the tuning of a neuron for target position; that is, the cell continues to respond best for the same retinal location of the target. Instead, varying the hand position modulates the gain of target-related responses. Such 'gain fields' for body position signals are ubiquitous in the brain and have been postulated to play a role in coordinate transformations, among other functions. A surprising feature of these hand position gain fields in PRR is that they do not appear to encode the location of the hand with respect to the body, as one might expect of a signal derived in part from proprioceptors. Rather, PRR neurons appear to encode hand position in the same reference frame that is used to encode the target (i.e., a retinocentric one). Since neurons in PRR encode information about the locations of both the target and the hand with respect to gaze, PRR can be said to provide a retinocentric representation of the hand movement vector.

The issue of reference frames has been taken up in another part of the PPC as well, namely area 5. Neuronal activity in this area seems to represent the remapping of the movement vector into a frame of reference more closely aligned with arm posture. Some neurons in area 5 do not encode the movement vector in any single frame of reference; rather, the responses of these neurons appear to reflect the use

of a retinocentric frame in conjunction with one that is more closely linked to the arm. Other area-5 neurons do not appear to use retinocentric coordinates at all; these neurons explicitly encode the difference between the current target and hand locations (motor error) in a manner that is independent of gaze direction. It is unclear whether this latter activity is best described as encoding a motor error in extrinsic coordinates or a motor error that is expressed in the intrinsic coordinates of the joints or muscles (an arm-centered one). In any event, within the PPC there is a continuum of responses, from neurons whose tuning is purely eye centered (PRR) to neurons that are purely hand/arm centered or that show a mixture of these two reference frames (area 5). The presence of such a continuum, as well as the paucity of neurons encoding reach-related information in head-centered reference frames, has led to the proposition that, at least in contexts in which the hand and target positions can be directly compared via vision, coordinate transformations for reaching involve a direct transformation from a retinocentric reference frame to a hand/arm-centered one (Figure 1).

Understanding subsequent transformational stages depends critically on the precise nature of the hand/arm-centered representation in area 5. An arm-centered representation is ultimately required by the motor system, that is, simply knowing the direction and distance the hand must move to reach the target is insufficient to specify a control signal for the reach, as this control signal must account for the fact that the arm's mechanical properties vary as a function of arm configuration. Thus, the continuum of responses described previously either must reflect the direct remapping of a retinocentric movement vector onto arm coordinates or transformations occurring 'downstream' of area 5 must incorporate a mechanism for transforming a hand-centered (extrinsic) representation into the appropriate reference frame for activating the muscles.

An early multiplexing of visual signals with signals related to eye, head, and arm orientation would provide support for the idea that the PPC directly maps a retinocentric representation of the movement vector onto arm muscle-based coordinates. In support of this idea, neurons in PRR, which is thought to occupy a processing stage functionally antecedent to area 5, appear to encode eye orientation as well as the retinocentric location of the target and hand. It is unknown whether head position signals are encoded by PRR neurons in the same way as eye position because the relevant experiments have yet to be performed, but access to such signals is implied by the responses of neurons in other parts of the PPC, namely 7a and the lateral intraparietal area. Regarding arm

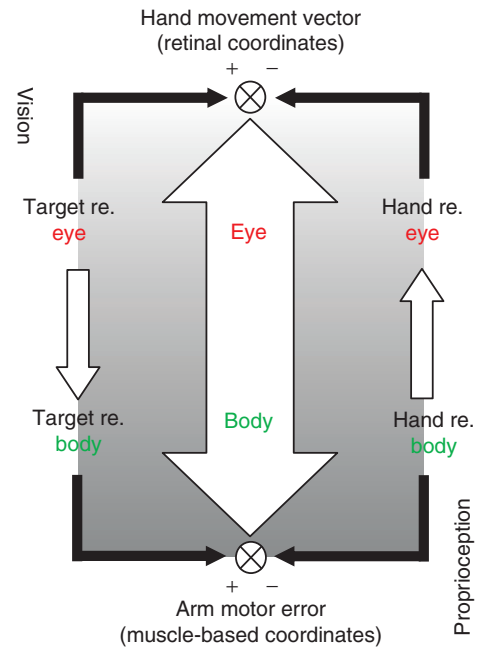

Figure 1 Schematic of the reference frame transformations for reaching, based on psychophysical and neurophysiological studies. In general, a hand movement vector in retinal coordinates is mapped directly to an arm motor error expressed in muscle coordinates (center arrow). The precise mechanism underlying this transformation may be context dependent. In the absence of continuous and reliable visual feedback (e.g., for movements generated in the dark or following long delays), targets are remapped from eye to body-centered coordinates and then compared to the body-centered hand location to derive the arm motor error (left arrow). In contrast, when the hand and target are easily viewed, hand position is remapped from body to eye coordinates to determine the hand movement vector in eye coordinates (right arrow), which is subsequently remapped into muscle coordinates. Note that although explicit head-centered representations are not required for either mechanism, head (and eye) position signals play an important role in both mechanisms.

orientation, it has been shown that the preferred directions of neurons in area 5 rotate with changes in arm posture in much the same way as they do in M1. This phenomenon is at least consistent with a role for the PPC in transforming retinocentric information into a reference frame that is in line with the one used to specify muscle activations.

The existence of an intrinsic, arm-centered frame of reference in area 5 would not necessarily mean that the PPC is involved in specifying the joint torques or muscle activations required to move the arm. As stated previously, neurons in M1 and the dorsal premotor cortex also show evidence of 'muscle-like' tuning and thus any one or all three of these areas could support the transformation from kinematic to kinetic signals. Moreover, it should be noted that area 5 activity does not appear to reflect the time course of the forces that generate arm movements as well as activity in M1. Thus, area 5 could simply represent an

intermediate transformational stage that is intrinsic in nature but 'dynamics free.' Alternatively, the representation in area 5 may be part and parcel of a role in estimating limb position on the basis of visual and somatosensory information and previous motor commands. Integration of the latter two signals would seem to require an arm-centered frame and would also predict some degree of sensitivity to changing arm dynamics, which has been reported in area 5. This last scenario would also fit well with the results of transcranial magnetic stimulation (TMS) studies in humans, which have reported impairments in the learning of novel arm dynamics with stimulation of the PPC.

Conclusion

Remapping hand position into an eye-centered frame of reference requires at least as many transformational stages as a scheme requiring remapping of target position into body-centered coordinates. What, then, is the benefit of a retinocentric representation of the hand? Recoding hand position in retinal coordinates would facilitate online updating of reaching movements because the target and hand could then be rapidly and directly compared. This may be especially important in conditions in which there is relative motion of the target with respect to the body, a situation that has been studied rarely experimentally but one that occurs quite commonly in everyday tasks. Interestingly, TMS studies of normal human subjects as well as studies of patients strongly implicate the PPC in rapid online corrections of arm movements. Visual feedback of the hand is generally considered to be too slow to effectively mediate such corrections; thus, a retinocentric representation of hand position would be most beneficial for online control if it were based on other signals besides vision. The existence of this type of representation is supported by recordings of neurons in the PPC: in at least some of these neurons, hand position appears to be encoded in a retinocentric frame of reference even in the absence of concurrent visual input.

See also: Corticomotoneuronal System; Finger Movements: Control; Kinematics and Dynamics; Motor Sequences; Motor Timing; Posterior Parietal Cortex and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape; Reaching and Grasping.

Further Reading

Batista AP, Buneo CA, Snyder LH, and Andersen RA (1999) Reach plans in eye-centered coordinates. *Science* 285: 257–260.

Buneo CA, Jarvis MR, Batista AP, and Andersen RA (2002) Direct visuomotor transformations for reaching. *Nature* 416: 632–636.

Buneo CA, Soechting JF, and Flanders M (1997) Postural dependence of muscle actions: Implications for neural control. *Journal of Neuroscience* 17: 2128–2142.

Caminiti R, Johnson PB, Galli C, Ferraina S, and Burnod Y (1991) Making arm movements within different parts of space: The premotor and motor cortical representation of a coordinate system for reaching to visual targets. *Journal of Neuroscience* 11: 1182–1197.

Flanders M, Helms Tillery SI, and Soechting JF (1992) Early stages in a sensorimotor transformation. *Behavioral and Brain Sciences* 15: 309–320.

Georgopoulos AP (1995) Current issues in directional motor control. *Trends in Neurosciences* 18: 506–510.

Gordon J, Ghilardi MF, and Ghez C (1995) Impairments of reaching movements in patients without proprioception: I. Spatial errors. *Journal of Neurophysiology* 73: 347–360.

Henriques DY, Klier EM, Smith MA, Lowy D, and Crawford JD (1998) Gaze-centered remapping of remembered visual space in an open-loop pointing task. *Journal of Neuroscience* 18: 1583–1594.

Johansson RS, Westling G, Backstrom A, and Flanagan JR (2001) Eye–hand coordination in object manipulation. *Journal of Neuroscience* 21: 6917–6932.

Kakei S, Hoffman DS, and Strick PL (1999) Muscle and movement representations in the primary motor cortex. *Science* 285: 2136–2139.

Lacquaniti F, Guigon E, Bianchi L, Ferraina S, and Caminiti R (1995) Representing spatial information for limb movement: Role of area 5 in the monkey. *Cerebral Cortex* 5: 391–409.

McIntyre J, Stratta F, and Lacquaniti F (1997) Viewer-centered frame of reference for pointing to memorized targets in three-dimensional space. *Journal of Neurophysiology* 78: 1601–1618.

Vetter P, Goodbody SJ, and Wolpert DM (1999) Evidence for an eye-centered spherical representation of the visuomotor map. *Journal of Neurophysiology* 81: 935–939.

Vindras P, Desmurget M, Prablanc C, and Viviani P (1998) Pointing errors reflect biases in the perception of the initial hand position. *Journal of Neurophysiology* 79: 3290–3294.

Motor Sequences

R Sosnik, Bar Ilan University, Ramat Gan, Israel
F Polyakov and T Flash, Weizmann Institute of Science, Rehovot, Israel

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Motor learning refers to two different processes: adaptation and skill acquisition. Adaptation relates to our ability to compensate for environmental changes, including visual, kinematic, or dynamic changes in the characteristics and attributes of external stimuli and/or objects. On the other hand, the ability to learn a movement sequence and the incremental improvement in its performance refer to the sequence learning aspects of still acquisition which are the main focus of our present discussion. The generation of sequential actions is central to most of human intelligent behaviors ranging from everyday skills such as tying one's shoe laces to using a computer or playing the piano. The acquisition, representation, and retention of motor sequences are believed to build on and benefit from the availability of basic, simple motion elements ('primitives'), which are then ordered and concatenated into particular sequences, thus giving rise to the diverse, complex, goal-directed motor repertoire of humans and other animals. Long-lasting changes in motor performance related to sequence learning are known to evolve following several distinct phases and usually require many practice sessions. Subsequent sections discuss several topics related to sequence learning, starting with the presentation of two different mechanisms of sequence learning (i.e., explicit vs. implicit). The article then describes up-to-date theories concerning the specificity of these different learning mechanisms and their time course and discusses the neural substrates and brain areas subserving sequence learning. Finally, it reviews some of the current literature, pertinent to sequence learning, concerning shifts in representation, consolidation, and the effect of sleep on the retention and off-line improvement in the performance of learned sequences.

Implicit versus Explicit Sequence Learning

When considering the learning of sequential tasks, one needs to distinguish between explicit and implicit learning. Explicit learning is frequently assumed to be similar to the processes operating during conscious problem solving and includes conscious attempts to construct a representation of the task, a directed search of memory for similar or analogous task-relevant information, and conscious attempts to derive and test hypotheses related to the structure of the task. This type of learning has been distinguished from alternative models of learning, termed 'implicit' learning. The term 'implicit' learning denotes learning phenomena in which more- or less-complex structures are reflected in the overt behavior although the learners are unable to verbally describe these structures. Numerous studies have examined implicit learning of serial-order information using the serial reaction time (SRT) task established by Nissen and Bullemer. In this task, participants see a target stimulus, typically an asterisk, appearing at one of four locations on a computer display and are required to indicate its location by making a key press. Participants are initially trained on repeating a given sequence. However, learning is assessed by presenting participants with a block of experimental trials (e.g., a block of random trials) and then examining differences in reaction time (RT) existing between sequence and random blocks of trials. It was shown that the RTs in trials in which an overpracticed sequence was presented were significantly shorter than for random sequences, suggesting that participants have unconsciously learned the sequence and anticipated the next element (Figure 1).

There is a vast literature debating what is really learned in the implicit sequence paradigm in general and in the SRT task in particular. From a theoretical point of view, the description of a given sequence structure is not trivial because a given structure typically has several different structural components, such as frequency-based statistical structure (i.e., the redundancy of a specific element appearance in the pattern), relational structures (the occurrence of a specific element after or before a different element), and spatial structure (the position of the different elements in space, as in the case of fluent finger spelling). A literature review shows that all these components have an influence on the rate in which a sequence is learned.

Stages in Motor Learning: Transferability and Effector Specificity

A study conducted in monkeys, in which a sequence of 10 button presses is learned by trial and error, has shown that the time course of improvement of two performance measures, key press errors and RT, was different (Figure 2). The key press errors reached an asymptote within a shorter period of training compared with the RTs, which continued to decrease throughout a longer time. This finding suggested that the acquisition of sequence knowledge (as measured

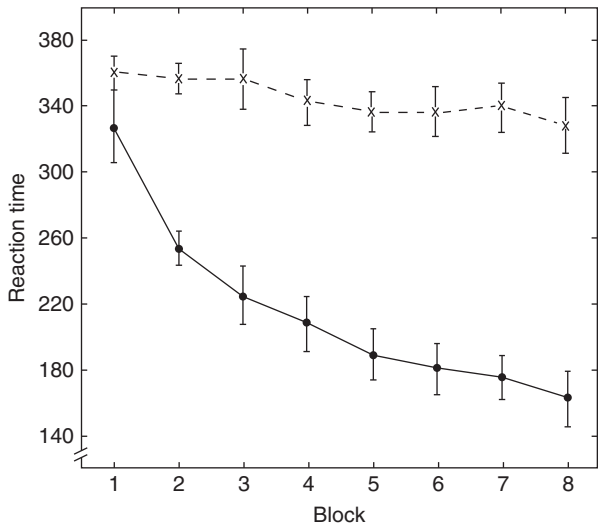

Figure 1 Implicit acquisition of a practiced sequence. Participants see an asterisk appearing at one of four locations on a computer display and are required to indicate its location by making a key press. After a training period on repeating a given sequence, the participants are presented with a block of random trials. The differences in reaction time (RT) existing between the trained sequences and the random blocks of trails are shown by depicting the mean of median reaction times in milliseconds in each block. Filled circles, repeating sequence; x's, random sequence. Bars represent standard errors. From Nissen MJ and Bullemer PT (1987) Attentional requirements for learning: Evidence from performance measures. *Cognitive Psychology* 19:1–32.

by key press errors) may take place quickly, but long-term motor sequence learning (as measured by RT) may take longer to become established. Thus, different aspects of the task are learned in different timescales. It was shown that the learned procedure is effector-dependent: when monkeys were asked to use the hand that has not been used for long-term practice of the given sequence, their performance became very poor. However, this was not true for a newly learned sequence. These results suggest that the memory for a visuomotor sequence, which is accessible for both the trained hand and the untrained hand in the early stage of learning, becomes largely inaccessible to the untrained hand in the late stage of training, that is, the performance gain is not transferred to the untrained hand. It was further demonstrated that the whole sequence, rather than the individual elements, is learned: When the order of the component elements of a well-learned sequence was reversed, the performance became very poor, as if the monkeys were performing completely new sequences. The results suggested that the monkeys acquired the memories for individual elements separately. This was not true in the early learning stages. Further studies in monkeys and in humans demonstrated that although effector-dependent and -independent learning occur simultaneously,

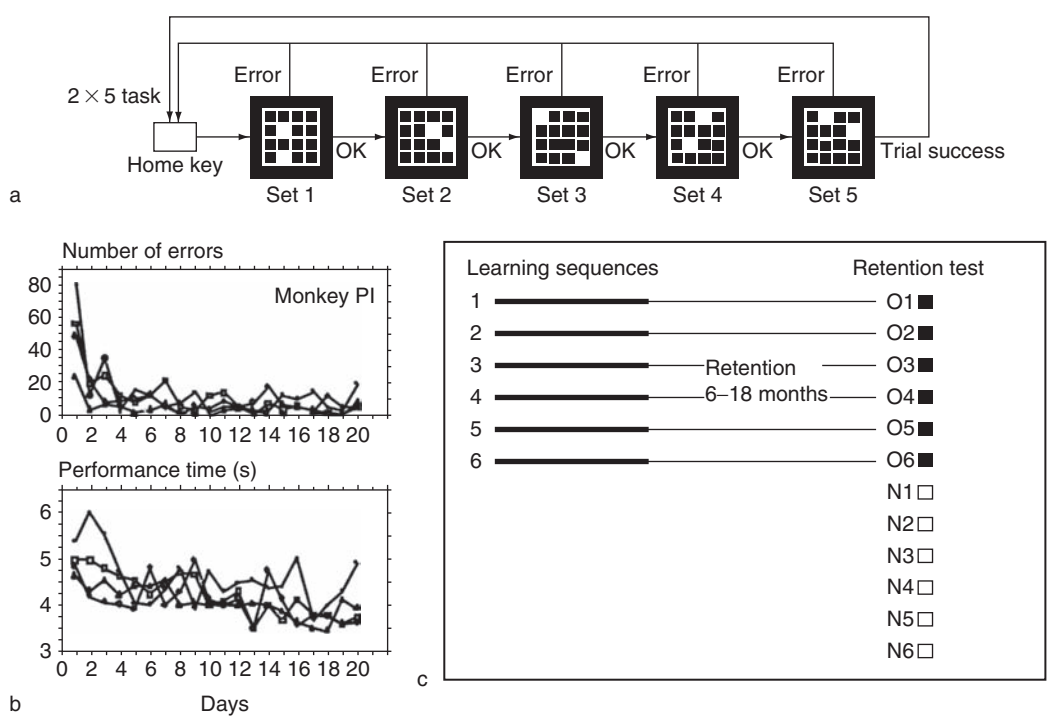

Figure 2 Long-term retention test of motor skill in monkeys. (a) A procedure of the 2 × 5 sequence task. A monkey has to press the two illuminated buttons in the correct order to proceed to the next set until completing five sets (called a 'hyperset'). Each hyperset is presented repeatedly in a block until the monkey performs the whole hyperset for a total of 20 trials. A different hyperset is then used for the next block. (b) The learning curves of monkey PI for four hypersets during the initial practice period of 20 days are shown as the number of errors before completing 20 trials successfully (top) and the mean performance time for the successful trials (bottom).

effector-dependent representation might take longer to establish than effector-independent representation does and that they are mediated by different neural substrates.

It is reasonable to assume that a gain in performance reflects a change in brain processing triggered by practice. Many skills, while acquired, are retained over long time intervals, suggesting that training can induce long-lasting neural changes. Previous results from neuroimaging studies in which performance is modified over time have shown that many different stages can occur. As an effect of repetition or practice, several studies report that activations in particular brain areas decreased. In other studies, specific brain areas showed increase in the magnitude or extent of activation. In some reports, a shift in activated areas was observed between unskilled and more skilled performance of a task. Thus, decreases, increases, and shifts in activations are reported as an effect of practice. The differences in activations might be related, in part, to the amount of practice or repetition a participant has had with a task. A study conducted by Karni et al. demonstrated that motor skill learning (sequential finger opposition task) requires time and has two distinct phases, analogous to those subserving perceptual skill learning. An initial, fast improvement phase ('fast learning'), when the extent of activation in primary motor cortex (M1) decreases (habituation-like response), is followed by slowly evolving, posttraining incremental performance gains ('slow learning'), when the activation in M1 increases compared with control conditions. Figure 3 depicts the differential evoked responses in M1 in the trained versus the untrained (control) sequence. Overall, this study showed that limited training experience can be sufficient to trigger performance gains that require time to become evident. It was proposed that skilled motor performance is acquired in several stages: fast learning, an initial, within-session improvement phase, followed by a period of consolidation of several hours duration, and then slow learning, consisting of delayed, incremental gains in performance emerging after continued practice. This time course may reflect basic mechanisms of neuronal plasticity in the adult brain that subserve the acquisition and retention of many different skills. Later, using a simple single finger opposition task, Tracy et al. showed that even during a brief practice, there was a vast difference in activation in the corpus striatum, thalamus, and cingulate gyrus so that the neural circuit implementing this simple task varied over a brief window of practice.

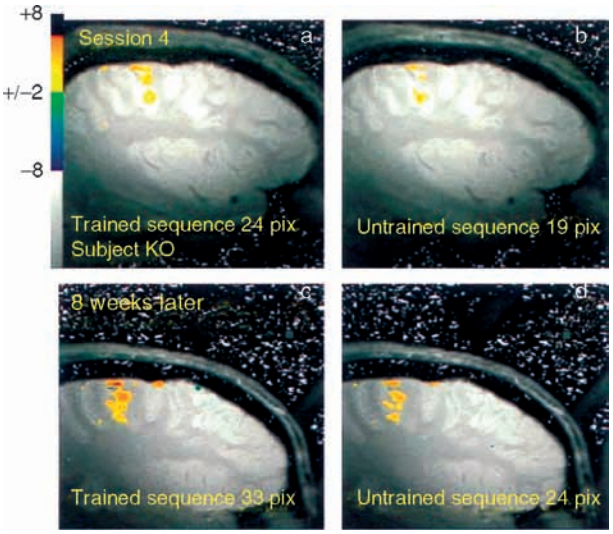

Figure 3 Differential evoked responses in human M1 to the trained vs. the untrained (control) sequence. (a, b) Training and performance during scanning done with the left (nondominant) hand. (c, d) Emergence of differential activation after 3 weeks of daily practice on the designated training sequence. Maintained differential activation 8 weeks later with no additional training in the interval. *Z*-score values are indicated by the psuedo-color scale. Reproduced from Karni A, Meyer G, Hipolito CR, et al. (1998) The acquisition of skilled motor performance: Fast and slow experience-driven changes in primary motor cortex. *Proceedings of the National Academy of Sciences of the United States of America* 95: 861–868. Copyright (1998) National Academy of Sciences, USA.

A Schematic View of the Neural Representations of Motor Sequence Learning

Out of numerous views and results, sometimes controversial, that can be found in the literature on neural representation of learning motor sequences, this article discusses two major approaches, one claiming that there exist two parallel systems for neural coding of sequences and the other differentiating between the explicit and implicit learning processes. This approach bridges different observations and notions concerning sequence learning.

Two Parallel Systems of Coding

When people learn to perform a novel motor task, a corresponding memory trace is generated in the

(c) General schedule of the long-term retention test. After retention periods of 6–18 months, the monkey was instructed to perform the previously learned hypersets (O1–O6) and new hypersets (N1–N6) in a 1-day session. Reproduced from Hikosaka O, Rand MK, Nakamura K, et al. (2002) Long term retention of motor skill in macaque monkeys and humans. *Experimental Brain Research* 147: 494–504, with kind Permission of Springer Science and Business Media.

brain. This type of memory encodes procedural aspects of the task rather than facts and is termed 'procedural memory.' In the early stages of learning, a sequential motor procedure is composed of a series of sensorimotor processes, executed in a discrete manner (Figure 4(a)). However, the amount of computation needed to generate a novel sequential task is highly demanding. Such computations include the need to perform a series of coordinate transformations (e.g., from allocentric to egocentric or from eye-centered to hand-centered coordinate frames) and to deal with ordinal and temporal aspects of sequence generation, such as concatenating among sequential elements within the sequence. In order to address the question of how the system simplifies such complex computations and in order to relate to the finding that different learning stages are characterized by ability/inability to transfer the performance gain to the untrained hand, a hypothetical scheme for sequence learning has been suggested. This scheme, which attempted to suggest a plausible organization of motor sequence learning in the brain, proposed that a motor procedure is acquired and encoded independently by two cortical systems: one system encoding the sequence in spatial, extrinsic coordinates and the other one encoding the sequence in motor, intrinsic coordinates (Figures 4(b) and 4(c)). In the suggested scheme, each of these two parallel processes operates using a single type of coordinate system (either spatial or motor coordinates), hence providing a more parsimonious mode of representation. Based on a number of empirical studies conducted in monkeys and humans, the model proposes that (1) the procedural sequence is acquired in spatial, visual coordinates (green connections in Figure 4) before it is acquired in motor, intrinsic coordinates (blue connections in Figure 4); (2) the motor coordinate representation persists permanently, whereas the spatial representation is only temporary; and (3) the motor coordinate representation is effector specific and the spatial coordinate representation is not effector specific. The scheme may also be valid for the paradigms in which procedural memory is acquired implicitly, such as the SRT task.

As discussed earlier, the acquisition of a motor sequence in the brain is subserved by two distinct types of processes: implicit and explicit. An important question is whether the neural substrates of the two learning processes overlap and whether the acquisition of one learning type interacts or interferes with the other. Some studies reported that the neural substrates of implicit and explicit learning overlap at the levels of the prefrontal cortex and basal ganglia and that explicit learning of sequences, which exerts

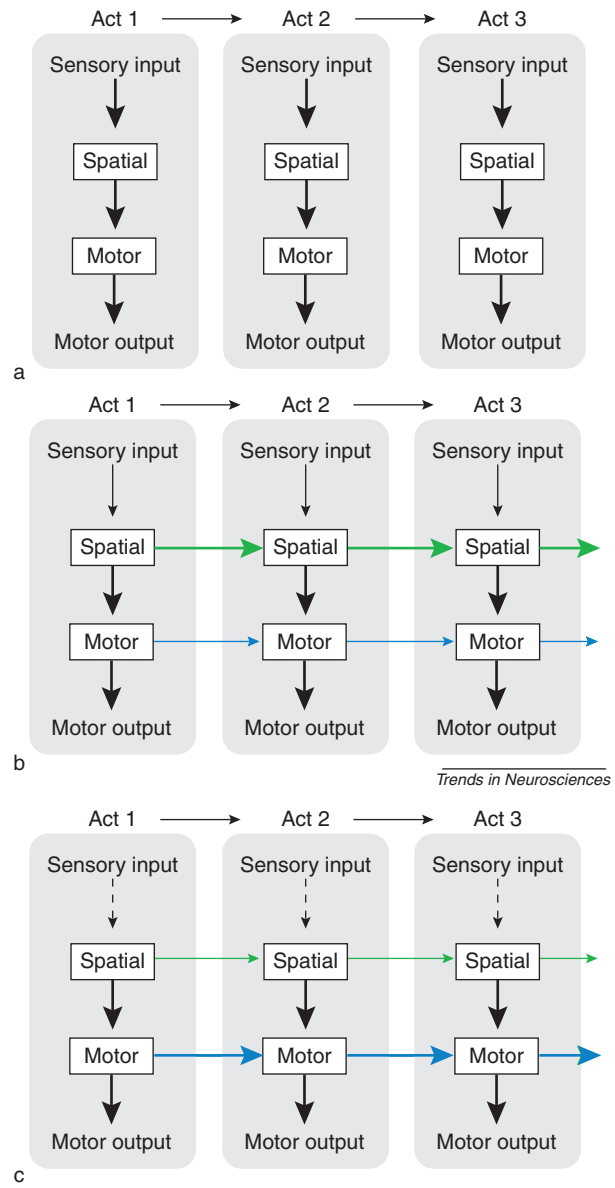

Figure 4 A hypothetical scheme for information processing during sequence learning. (a) The pre-learning stage, in which the participant performs three actions (Acts 1–3) step by step by relying on sensorimotor transformations for each action (vertical connection). (b), (c) By performing the actions in the same order, sequential processes are formed across the actions (horizontal connections). This occurs at each of the two stages of the sensorimotor transformation so that a given sequence is learned independently but from different perspectives: as a sequence in visual coordinates (spatial sequence, shown in green) and as a sequence in motor coordinates (motor sequence, shown in blue). The spatial sequence mechanism acquires the sequence quickly and flexibly by relying on attention and working memory (b), whereas the motor sequence mechanism acquires the sequence slowly and steadily through long-term practice (c). The thicker arrows indicate the dominant process occurring in the early (b) and late (c) stages and also the corresponding coordinates, that is, spatial in (b) and motor in (c). From Hikosaka O, Nakahara H, Rand M, et al. (1999) Parallel neural networks for learning sequential procedures. *Trends in Neurosciences* 22: 464–471.

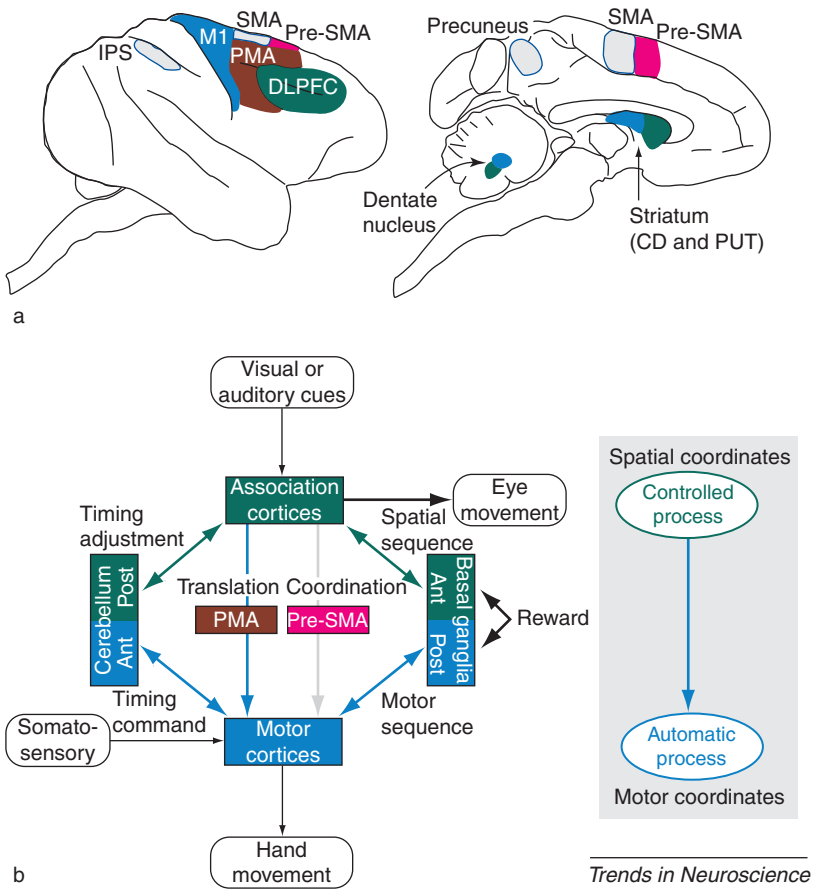

Figure 5 The association cortices (particularly prefrontal cortex) and the anterior part of the basal ganglia in monkeys constitute a loop circuit that operates as a spatial sequence mechanism (shown in green); the motor cortices (including primary motor cortex (M1) and supplementary motor area (SMA)) and the middle part of the basal ganglia constitute a loop circuit that operates in terms of motor coordinates (shown in blue). In addition, the anterior and posterior parts of the cerebellum contribute to the motor and spatial aspects of sequence generation by forming loop circuits with the motor and association cortices; their functions in the control of timing (indicated in the figure) are still speculative. The spatial and motor loops are capable of operating independently, acquiring and reproducing a given sequence in different coordinates. In addition, they have their own input and output: visual and auditory inputs and oculomotor outputs for the association cortices; somatosensory inputs and skeletomotor outputs for the motor cortices. The interaction between the two learning mechanisms is made possible in two ways: (1) by translation from spatial into motor coordinates by the premotor area and (2) by coordination or switching by the pre-SMA. From Hikosaka O, Nakahara H, Rand M, et al. (1999) Parallel neural networks for learning sequential procedures. *Trends in Neurosciences* 22: 464–471.

functional changes in the basal ganglia, directly inhibits implicit learning.

Figure 5 presents an anatomically supported model that describes the coordination between different brain areas mediating motor sequence learning, either in spatial or in motor coordinates, whereas Figure 6 depicts a scheme highlighting the hierarchy of the cortical structures based on their involvement in explicit/implicit representation. The organization of the cortical areas within the hierarchies depicted in both schemes is similar, and basically the two schemes supplement each other. Taken together, they suggest that the cortical structures that are strongly involved with explicit or implicit processes are mostly engaged with representing a new acquired sequence in spatial or motor coordinates, respectively. Hence, the interference between implicit and explicit sequence learning may be related to the functional and anatomical overlap among the brain structures mediating the spatial and motor learning mechanisms.

The Role of Different Brain Areas in Sequence Acquisition, Representation, and Implementation

The prefrontal cortex is located at the top of both schemes describing learning of motor sequences (Figures 5 and 6). Compelling evidence exists in favor of the prefrontal cortex involvement in the process of explicit sequence learning and the acquisition and representation of the spatial aspects of a newly acquired sequence. Furthermore, it is considered to be related to several functions relevant to sequence leaning, such as working memory, rehearsal,

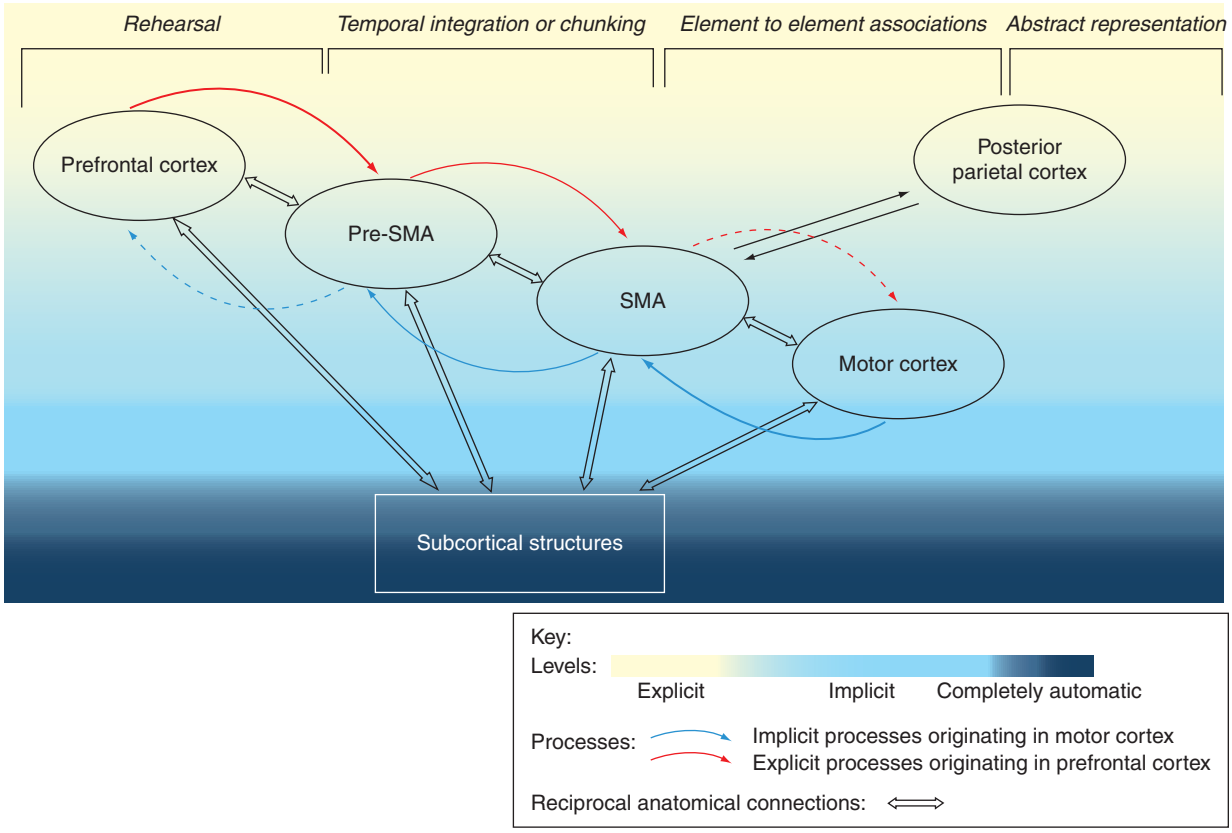

Figure 6 A set of cortical structures subserve the control of motor sequences. Depending on the instructional set, a participant's intentions, and the development of awareness, implicit or explicit processes predominate or interact during sequence learning, triggering specific mechanisms and leading to the formation of different types of representations of the learned sequence. For instance, when the intention to learn the sequence is explicit, either as a consequence of awareness or as an instructional set, explicit processes will originate in prefrontal cortex and will propagate to premotor areas. In this case learning will be based on mechanisms such as rehearsal and chunking. When learning is implicit, as a result of simply performing sequential movements without intention to learn or awareness of the sequence, implicit processes originate in motor cortex and propagate to the premotor areas. In such instances learning might be based on element-to-element associations or temporal coding. Of course, in most cases, these two types of processes will interact with each other and will change in importance depending on the stage of the learning process. From Ashe J, Lungu O, Basford A, and Lu X (2006) Cortical control of motor sequences. *Current Opinion in Neurobiology* 16: 213–221.

and executive control. Lately, it has been shown that the prefrontal cortex is strongly activated during explicit learning in the SRT task, a standard visuomotor RT paradigm, and that single dorsolateral prefrontal neurons encode the temporal order of a sequence of actions.

Based on the hierarchy depicted in Figure 6, the medial premotor cortex is situated a step below the prefrontal cortex. A large body of work, including lesion studies, suggests that the medial premotor cortex has an important role in the temporal organization of sequences, including the chunking together over time of multiple, separately performed movements.

While the functional properties of the supplementary motor area (SMA) and pre-SMA overlap, it is still generally accepted that the repetition of earlier acquired sequences is more strongly manifested in the activity of SMA neurons and that there is a greater tendency of the activity of pre-SMA neurons to reflect representation of new sequences. Pre-SMA and premotor areas (including SMA) may be involved in coordinating between the spatial and motor sequence mechanisms (Figure 5). A specific feature of the pre-SMA with respect to learning is that its neurons tend to become active during acquisition of novel explicit sequences, and their inactivation causes deficits in learning of novel sequences.

M1 is strongly related to the execution of movements, so establishing its role in the acquisition of a novel movement sequence is difficult. However, experiments designed to differentiate between the two contexts (instantaneous encoding of movements and encoding of motor sequences), such as generating different sequences composed of the same elements but in different temporal order, implied that the neural activity in M1 and dorsal premotor cortex also encodes sequential attributes of motor tasks and that its role in sequence production is independent of its

role in basic control of movements. Moreover, M1 is engaged at the early stage of motor consolidation. Thus, in contrast to the commonly held view of the motor cortex solely as being important for motor execution, it may also be fully equipped for the planning and production of sequential behavior.

The function of the cerebellum in sequence learning is controversial. The cerebellum is suggested to be involved in the coordination of the movements of different body parts and in the acquisition of timing or rhythm. Coherent changes of activity in the cerebellum and M1 were found during learning of timed motor sequences. In particular, during the initial stages of learning (when performance is poor), the activity in the cerebellum was greater than the cerebellar activity at the advanced learning stage. In M1, conversely, greater activity corresponded to better performance at the advanced learning stage.

Neuropsychological research suggests that implicit sequence learning in the SRT task is spared in patients with organic amnesia, so implicit SRT learning does not appear to depend on the medial temporal and diencephalic brain regions that are critical for explicit memory. Conversely, patients with Huntington's or Parkinson's disease have consistently shown SRT impairments, so the basal ganglia seem to be critically involved in SRT learning. Recent studies indicate that the anterior striatum affects learning of new sequences whereas the posterior striatum is engaged in recalling a well-learned sequence. It is important to note that the hypothetical parallel learning mechanisms (motor vs. spatial) provide advantages in parsimony, robustness, and adaptability of the system. The basal ganglia may have a key role in motivating procedural learning. Reinforcement (via dopaminergic neurons) might take place separately for the spatial (via the anterior striatum) and motor (via the middle striatum) sequence mechanism (Figure 5).

Movement Planning and Anticipation: Submovements and Coarticulation

The stereotyped kinematic patterns of planar reaching movements are not the expression of a prewired or an inborn motor pattern but the result of learning during ontogenesis. When infants start to reach, their reaching is characterized by multiple accelerations and decelerations of the hand, whereas experienced infants reach with much straighter hand paths and with a single bell-shaped hand velocity profile. It has been shown that it is possible to decompose a large proportion of infant reaches into an underlying sequence of submovements that resemble simple adult reaching movements. Based on several studies with adult participants, it has been suggested that the CNS uses elementary simple movements, primitives, which are smoothly combined in order to compose more complicated trajectories.

Other evidence for the existence of primitive submovements has come from work on hemiplegic stroke patients, which shows that their first movements are clearly segmented and exhibit a remarkably invariant speed profile. The idea that curved and obstacle avoidance movements result from the superposition of more elementary movements was modeled by Burdet, who suggested a mathematical model for the learning of accurate human arm movements. Its main features are that the overall movement results from the superposition of smooth smaller submovements (primitives). A similar notion was also elaborated by Henis and Flash, who formulated the 'superposition' model for target switching experiments. It was found that arm trajectory modification in a double target displacement paradigm might involve the vectorial summation of two independent plans, each coding for a smooth point-to-point trajectory. The first plan is the initial unmodified plan for moving between the initial hand position and the first target location. The second plan is a time-shifted trajectory plan that starts and ends at rest and has the same amplitude and kinematic form as a simple point-to-point movement between the first and second target locations.

After extensively practicing a particular sequence, most of the participants undergo implicit or explicit anticipation, which results in a coarticulation – the emergence of both spatial and temporal overlap among adjacent articulatory units. It is well known that as humans learn to speak, the speech becomes smoother and more fluent. Coarticulation in speech production is a phenomenon in which the articulatory movements for a given speech sound vary systematically with the surrounding sounds and their associated movements. Coarticulation was also studied in hand motor sequence generation. It was shown that pianists could anticipate a couple of notes before they are being played, which resulted in hand and finger kinematic divergence; that is, the spatial configuration of the fingers was different according to the context in which a specific note was played. Such a divergence implies an anticipatory modification of sequential movements of the hand, akin to the phenomenon of coarticulation in speech. Moreover, a study of fluent finger spelling has shown that rather than simply an interaction whereby a preceding movement affects the consecutive one; the anticipated movement in a sequence can systematically affect the one preceding it (Figure 7).

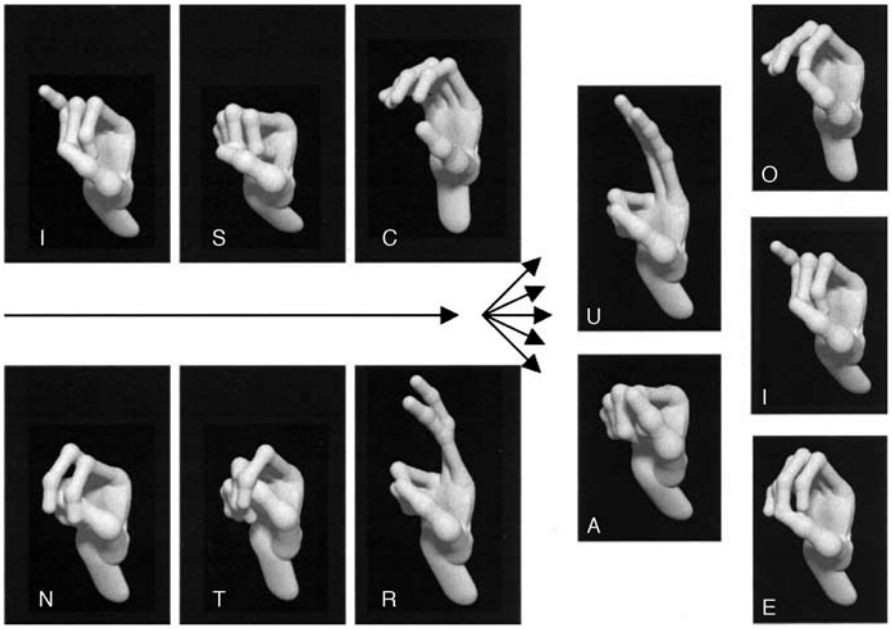

Figure 7 Cartoon images of hand shapes representing letters of interest in the American Sign Language manuscript alphabet, rendered with POV-Ray software. The layout illustrates the experimental design for a study of coarticulation: two fixed-letter strings (I-S-C and N-T-R) were followed by one of the five vowels (A, E, I, O, or U). Reproduced from Jerde TE, Soechting JF, and Flanders M (2003) Coarticulation in fluent finger spelling. *Journal of Neuroscience* 23(6): 2383–2393, with permission from Society for Neuroscience.

A recent study conducted by Sosnik et al. has shown that extensive training on writing like sequence movements, whereby participants have to connect four target points "as rapidly and as accurately as possible," results in the coarticulation of consecutive segments of a given sequence of planar movements generated between a series of target points and that the outcome of prolonged training can be a novel, curved trajectory, which although corresponding to a longer path, affords smooth and rapid performance with no loss in accuracy (Figure 8). With accumulated experience, a qualitative change occurs in motor planning: from the generation of a sequence of individually planned component movements to the generation of globally planned, maximally smooth movement units, providing a highly effective solution to the challenge imposed by the double requirement to move "as rapidly and as accurately as possible." These findings suggested that a profound hierarchical change can occur in motor planning: from syntax-dependent performance to unitary (modular), globally planned movement (coarticulation).

Consolidation and Sleep in Sequence Learning

So far in this discussion, only a distinction between an early versus a late stage of sequence learning has been made. However, as was recently summarized in a review by Doyon et al., sequence learning actually occurs over several stages. The first stage is an early and fast one in which considerable improvements in performance occur very rapidly within a single training session. This stage is followed by a slow stage, during which further gains in performance are achieved over several practice sessions. Third, there is a consolidation stage, in which off-line, spontaneous increases in performance are achieved provided that another task is administered only beyond a critical period of 4–6 h; otherwise interference caused by the administration of the new sequence learning task can occur. The fourth automatic stage is one during which the skilled behavior requires minimal cognitive resources and is resistant to interference and to the effects of the passage of time. Finally, there is a fifth retention period in which the motor skill can be performed after long delays without further practice of the learned task.

An important mechanism for the retention of learned sequences in memory is consolidation. The term consolidation, which was coined more than a century ago, describes the reduction of the fragility of declarative memory – memory for facts and events after it has been acquired and encoded. Through consolidation, an initially fragile memory is transformed into a more robust and stable one. Consolidation has been exhaustively investigated with respect to declarative memory, but more recently it has been

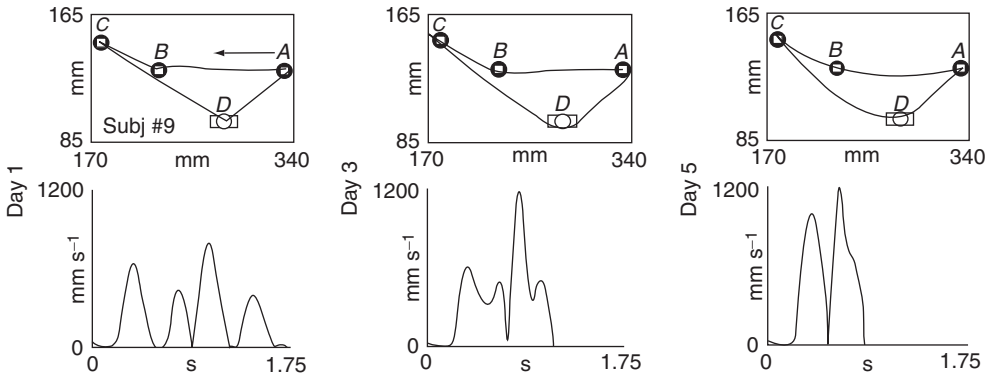

Figure 8 The evolution of coarticulated trajectories throughout training. Participants have to connect four target points "as rapidly and as accurately as possible," resulting in the coarticulation of consecutive segments of a given sequence of planar movements generated between the series of points. There was a relatively high spatial alignment (obtuse angle) between the lines connencting target pairs *AB* and *BC* or *CD* and *DA*. For each training day, upper and lower plots denote trajectories and velocity profiles, respectively. Note the novel, curved trajectory that results. Reproduced from Sosnik R, Hanptmann B, Karni A, and Flash T (2004) When practice leads to co-articulation: The evolution of geometrically defined movement primitives. *Experimental Brain Research* 156(4): 422–438, with kind permission of Springer Science and Business Media.

extended to other memory systems, including procedural memory. Beyond the reduction in the fragility of the memory trace that occurs following the acquisition of a new skill, consolidation refers also to another phenomenon, namely, the off-line improvement in performance that can occur either between practice sessions or after the practice period has ended. There is evidence for both types of consolidation processes. This evidence will be discussed especially in the context of sequence learning.

Off-line improvements in the performance of a skill occur over time without practice and often depend on sleep. Overnight improvements are seen in the performance of short finger sequences. Only a negligible amount of improvement in performance is seen when the interval between sessions is spent awake. Learning another sequence before sleep can block this overnight improvement, showing that this learning requires an active and sequence-dependent mechanism. Moreover, the improvement depends on the time spent in a particular component of sleep or in the combination of several sleep components. Depriving the participant of the critical sleep component or of any sleep at all greatly reduces off-line improvements. Fatigue or circadian factors cannot account for this phenomenon. Instead, off-line improvements in sequence performance seem to reflect marked changes in the brain areas subserving the mechanisms involved in memory storage, retrieval, and sequence generation. There is evidence to suggest that overnight skill improvements are supported by replay of the previously learned task and involve both enhancement of neuroplastic mechanisms and reorganization of the learned information. Overnight reorganization of information and of the memory trace might lead to changes in the representations of motor sequences that previously have been attributed only to practice. Moreover, the dependence of off-line improvement on sleep holds for the learning of sequential tasks (e.g., sequential finger movements) only when the skill is acquired intently and involves explicit learning. In the case of implicit learning, off-line learning is not sleep dependent but develops equally well during the day and overnight. It takes then 4–6 h for such off-line improvements to appear.

Historically, consolidation was described as the reduction in the fragility of memory for a new task. The capability to perform a new acquired skill is interfered with when the participant practices another new skill immediately after the first one. The amount of interference depends on the time that passes between the learning of the two skills. In particular, in the case of dynamic adaptation, learning a second task disrupts the memory trace for the first task, a phenomenon called retroactive interference. As the time between the two learning sessions increases, the memory trace of the first skill becomes more stable, and the interference of learning a second skill becomes weaker. There is some evidence to suggest that retention in memory of learned sequences might also be susceptible to disruption. Doyon et al. have suggested an integrated view of the plasticity that motor memory traces can undergo across learning stages. In the first learning stage, both motor adaptation and sequence learning recruit the corticostriatal and corticocerebellar systems, depending on the cognitive processes required for the task.

When the skill has been learned well, the neural representation of this motor behavior is thought to be distributed in a network of structures involving only one of these circuits. The corticostriatal and corticocerebellar systems are crucial for consolidation and for maintaining in long-term memory a new motor sequence or a new motor adaptation skill, respectively. Thus in the case of sequence learning, Doyon and Ungerleider have shown that the cerebellum is no longer essential, and long-lasting retention of the skill involves representational changes only in the striatum and in the associated cortical areas.

See also: Finger Movements: Control; Kinematics and Dynamics; Motor Skill Learning; Motor Timing; Posterior Parietal Cortex and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape; Premotor Cortex in Primates: Dorsal and Ventral; Premotor Areas: Medial.

Further Reading

Ashe J, Lungu O, Basford A, and Lu X (2006) Cortical control of motor sequences. *Current Opinion in Neurobiology* 16: 213–221.

Burdet E and Milner TE (1998) Quantization of human motions and learning of accurate movements. *Biological Cybernetics* 76: 307–318.

Carpenter AF, Georgopoulos AP, and Pellizzer G (1999) Motor cortical encoding of serial order in a context-recall task. *Science* 283: 1752–1757.

Doyon J and Benali H (2005) Reorganization and plasticity in the adult brain during learning of motor skills. *Current Opinion in Neurobiology* 15: 161–167.

Henis EA and Flash T (1995) Mechanisms underlying the generation of averaged modified trajectories. *Biological Cybernetics* 72(5): 407–419.

Hikosaka O, Nakahara H, Rand M, et al. (1999) Parallel neural networks for learning sequential procedures. *Trends in Neurosciences* 22: 464–471.

Hikosaka O, Rand MK, and Nakamura K (2002) Long term retention of motor skill in macaque monkeys and humans. *Experimental Brain Research* 147: 494–504.

Jerde TE, Soechting JF, and Flanders M (2003) Coarticulation in fluent finger spelling. *Journal of Neuroscience* 23(6): 2383–2393.

Karni A, Meyer G, Rey-Hipolito C, et al. (1998) The acquisition of skilled motor performance: Fast and slow experience-driven changes in primary motor cortex. *Proceedings of the National Academy of Sciences of the United States of America* 95: 861–868.

Korman M, Raz N, Flash T, and Karni A (2003) Multiple shifts in the representation of a motor sequence during the acquisition of skilled performance. *Proceedings of the National Academy of Sciences of the United States of America* 100: 12492–12497.

Lu X and Ashe J (2005) Anticipatory activity in primary motor cortex codes memorized movement sequences. *Neuron* 45: 967–973.

Matsuzaka Y, Picard N, and Strick PL (2007) Skill Representation in the primary motor cortex after long-term practice. *Journal of Neurophysiology* 97: 1819–1832.

Morasso P and Mussa-Ivaldi FA (1982) Trajectory formation and handwriting: A computational model. *Biological Cybernetics* 45(2): 131–142.

Nissen MJ and Bullemer PT (1987) Attentional requirements for learning: Evidence from performance measures. *Cognitive Psychology* 19: 1–32.

Penhunea VB and Doyon J (2005) Cerebellum and M1 interaction during early learning of timed motor sequences. *NeuroImage* 26: 801–812.

Roberston EM, Pascual-Leone A, and Miall RC (2004) Current concepts in procedural consolidation. *Nature Reviews Neuroscience* 5(7): 576–582.

Seidler RD, Purushotham A, Kim SG, et al. (2002) Cerebellum activation associated with performance change but not motor learning. *Science* 296: 2043–2046.

Sosnik R, Hanptmann B, Karni A, and Flash T (2004) When practice leads to co-articulation; The evolution of geometrically defined movement primitives. *Experimental Brain Research* 156(4): 422–438.

Tanji J (2001) Sequential organization of multiple movements: Involvement of cortical motor areas. *Annual Review of Neuroscience* 24: 631–651.

Tracy JI, Faro SS, Mohammed F, Pinus A, Christensen H, and Burkland DA (2001) Comparison of "early" and "late" stage brain activation during brief practice of a simple motor task. *Brain Research. Cognitive Brain Research* 10: 303–316.

Walker MP (2005) A refined model of sleep and the time course of memory formation. *Behavioral and Brain Sciences* 28(1): 51–86.

Motor Skill Learning

S P Wise, National Institute of Mental Health, Bethesda, MD, USA
D T Willingham, University of Virginia, Charlottesville, VA, USA

Published by Elsevier Ltd.

Introduction

The vertebrate brain evolved to control movements and thereby achieve biological goals, including procreation, harm avoidance, ingestion of essentials, exploration, and energy economics. To cope with changing conditions, early vertebrates developed – and their descendants inherited – the ability to expand their innate repertoire of motor capabilities. This facility, called motor skill learning, increased their ability to attain biological goals through interaction with the environment. What motor skill learning did for our ancestors 500 Ma, it does for us today.

Some motor skills involve a single movement; others require movement sequences. These two types of motor skills are described in the following sections.

Skilled Single Movements

Single skilled movements can involve any skeletal muscles, but most research centers on reaching, pointing, or throwing with the arm and hand. The target of movement might come from memory, but often it arises from sensory receptors. In that case, the motor system must transform sensory information about the target's location into the motor commands necessary to bring to a target either the hand, tools held in the hand, or objects thrown from the hand. The motor system learns these computations, called perceptual–motor or sensorimotor mappings, by adjusting them from movement to movement based on errors and uses these mappings and feedback to adjust ongoing movements.

Types of Perceptual–Motor Mappings

Perceptual–motor learning includes both standard and nonstandard mappings. Because vision dominates the sensory life of primates, including humans, most research deals with visuomotor mappings. In standard mapping, visible objects serve as the targets of movement. In nonstandard mapping, visual inputs provide information about targets, but movements are aimed elsewhere. One type of nonstandard mapping, called transformational mapping, requires an algorithm to transform a cue's location into that of the target. For example, a ball should be aimed at a moving target's projected location, not its current one. In another type of nonstandard mapping, called arbitrary mapping, a symbolic cue specifies the goal. Learning to stop at a red light is an example of arbitrary mapping. Figure 1 (pink shading) illustrates these three kinds of perceptual–motor mappings, along with some of the brain regions thought to play key roles in computing them.

Motor Plans

Reaching, pointing, and throwing require changes in the hand's location relative to the rest of the body, mostly because of movements around the shoulder, elbow, and wrist joints. Figure 1(b) illustrates someone ready to reach for a coffee cup. He might be able to see his hand, as illustrated, but vision of the hand's starting location is not necessary to compute its initial position. Typically, people fixate on the target (dashed lines), not on the hand, so proprioceptors provide the best signals for estimating the hand's initial position (Figure 1(a), green shading). This estimate depends on motor learning: the experience gained by relating viewed hand position to proprioceptive inputs over many movements. Prisms can perturb these relationships, which leads to misalignments of viewed and proprioceptively sensed hand position. These misalignments, in turn, lead to reaching errors. The estimated initial hand position can then be brought back into alignment by the practice gained in several reaches, which eliminates the errors. Once the prisms are removed, the misalignment returns and an error of opposite sign, called an aftereffect, appears. This error, too, diminishes with practice.

Once the motor system has computed the target's location and the hand's starting position, it can develop a plan for moving the limb. The difference between initial hand position and target location defines a vector, called the difference vector, which has distance and direction as its dimensions. A net movement of the hand by that distance in that direction will bring it to the target. Neurophysiological research in monkeys indicates that parts of the posterior parietal cortex (PPC) play the largest role in computing target location, initial hand position, and the difference vector.

The difference vector is an important aspect of motor planning, but not the only one. Much of the literature on perceptual–motor learning explores the principles that guide the generation of motor

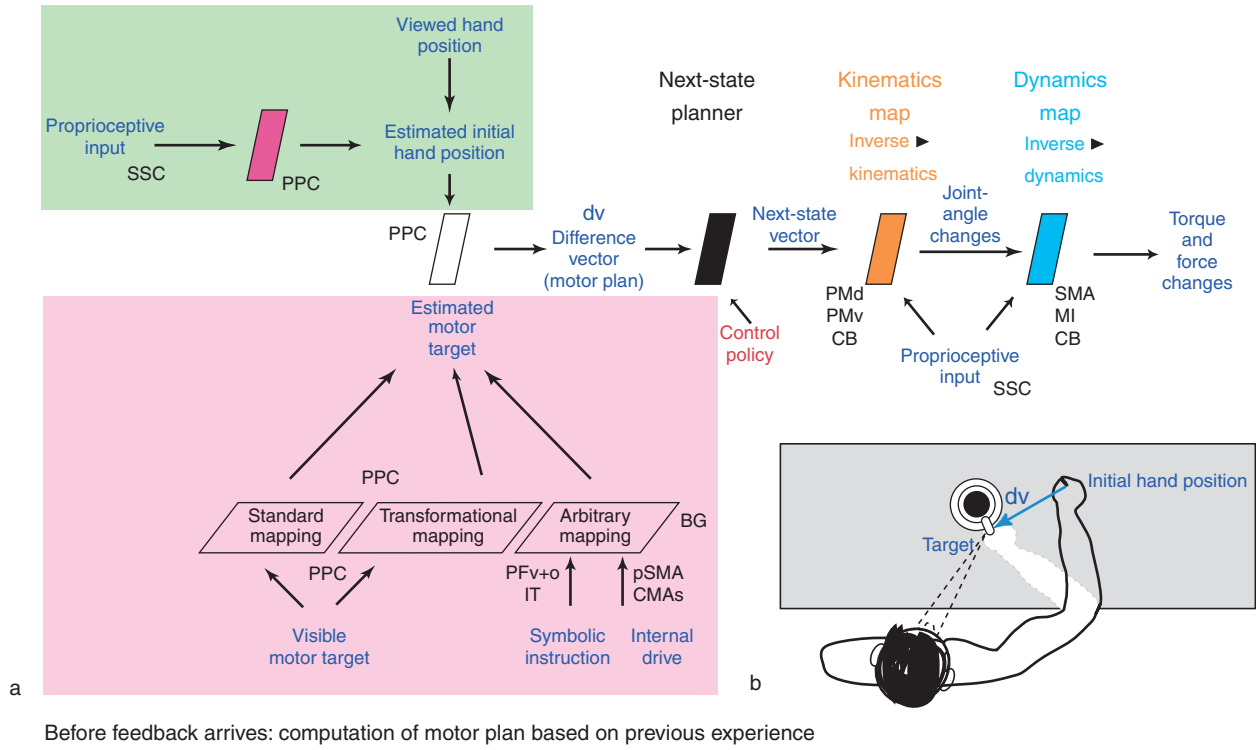

Before feedback arrives: computation of motor plan based on previous experience

Figure 1 (a) Model of skilled reaching. Each parallelogram represents a neural network, with some of their brain regions indicated nearby. One network (pink parallelogram) aligns proprioceptive input with the viewed position of the hand. A second network (white parallelogram) computes the difference between the current hand position and that of the target, called the difference vector. The next-state planner (black parallelogram) produces the next incremental change in hand position needed to reach the target in accord with a control policy, such as minimizing endpoint error. A kinematics map (orange parallelogram) is an internal model that learns to transform this next-state vector into the desired joint-angle changes. Finally, the dynamics map transforms the desired joint-angle changes into the required torques and forces. Abbreviations: BG, basal ganglia; CB, cerebellum; CMAs, cingulate motor areas; dv, difference vector; IT, inferior temporal cortex; PFv+o, ventral and orbital prefrontal cortex; PMd, dorsal premotor cortex; PMv, ventral premotor cortex; PPC, posterior parietal cortex; pSMA, pre-supplementary motor area; SMA, supplementary motor area; SSC, somatosensory cortex. (b) A person planning to grasp a coffee cup. The arm's configuration after the reaching movement appears in white. Dashed line: orientation of gaze. Adapted from Shadmehr R and Wise SP (2005) *The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge: MIT Press, with permission from MIT Press.

commands for a given difference vector. These ideas focus on constraining redundant means of reaching a target or minimizing the jerkiness of movements, coactivation of antagonist muscles, endpoint or intermediate-point errors, or curvature in visual coordinates. The dominant factors appear to be the last two. Indeed, among the others, some properties said to be 'invariant' often vary dramatically depending on a person's intentions. Furthermore, according to Fitts' law, the motor system must trade-off speed and accuracy. So the coffee drinker in Figure 1 cannot move his hand as fast as possible to get the coffee. Variations in motor commands for a given difference vector, called control policies, compose crucial aspects of motor skills.

Motor Commands

The difference vector represents the desired displacement of the hand. Neural networks must convert this information into command signals that cause the joint-angle changes and forces necessary to reach the target. The former computation is called inverse kinematics, the latter, inverse dynamics (Figure 1). The neural networks that compute these transforms implement the kinematics and dynamics maps that underlie, respectively, inverse kinematics and inverse dynamics. Inverse models of this kind are one kind of internal model: a neural representation of how the body interacts with the world and itself. Another kind, called a forward model, is taken up below. An inverse model uses desired movements to compute motor commands; a forward model uses motor commands to compute predicted movements.

The mappings computed by internal models must be learned. For example, prisms perturb kinematics maps, as noted above for proprioceptively sensed hand position. They cause reaching errors by altering the visual consequences of movements. Likewise, dynamics maps can be perturbed by imposing forces on the limb during movement. These networks adapt

their mappings on a movement-to-movement basis to reduce errors. Indeed, every movement teaches the motor system something about the motor commands needed for future movements.

Neuronal Mechanisms

Neurophysiological and neuroimaging research indicates that in addition to the cerebellum, the PPC and certain nonprimary motor areas, such as the dorsal and ventral premotor cortex, compute kinematics maps, while the primary motor area (M1) and the supplementary motor area (SMA) compute dynamics maps.

Neurons in all of these structures have tuning curves, which peak at a preferred direction (PD) of movement. A population of neurons with different PDs can serve as the motor primitives for representing motor parameters. These computational components are also known as basis functions. The neuronal mechanisms of motor skill learning involve changes in both the PDs of the tuning curves, which encode a parametric value, and their width, which reflects encoding sensitivity. The encoding sensitivity affects how errors generalize from one kind of movement to another.

A common perturbation of the kinematics map, called the 'rotation' task, involves the use of a mouse or joystick to control a cursor on a computer screen. Initially, when the mouse moves left, the cursor also moves left, when the mouse moves away from the body, the cursor moves up, and likewise for all directions. After a perturbation of 90°, for example, when the mouse moves left, the cursor moves up. Therefore, to use the mouse–cursor tool to reach an upward target, subjects must learn to move the mouse to the left rather than away from the body. Training on the 'rotation' task can be for one direction or for several. After training on one direction of movement, humans and monkeys show limited generalization of learning. They generalize to other directions as a function of angular distance from the training direction, with a fall-off similar to the width of M1 tuning curves (and those elsewhere). As noted above, computational models suggest that the width of these tuning curves – their encoding sensitivity – could account for this pattern of generalization. Furthermore, changes in tuning-curve width might contribute to skill learning. After training, changes in M1 activity, which consist of narrowing the width of its tuning curves, occur only in neurons with PDs close to the training direction. By virtue of this increase in encoding sensitivity, M1 neurons contain more information about movement direction after learning, and this greater information might enhance motor skill.

The activity of M1 neurons also changes during the learning of new dynamics maps (**Figure 2**). About one-fifth of M1 neurons change their PDs during adaptation to a force field (shaded area), but return to their baseline properties after removal of those forces, called the null field. For another fifth of the cells, their PD changes during the force-adaptation period, but they maintain those changes after monkeys readapt to the null field (**Figure 2(c)**). Finally, another fifth of the cells change their PD after a return to the null field, not during force adaptation (**Figure 2(d)**). Because, as a population, the changes depicted in **Figures 2(c) and 2(d)** balance, the population shows no net change in PD. Nevertheless, over 40% of M1 neurons have changes in PD that outlast the adaptation period. Thus, the sum of M1 output remains unchanged, which supports the computation of movement direction in the null field, yet the network as a whole retains an influence of monkey's adaptation to the force field.

In addition to M1, the cerebellum plays a central role in learning dynamics maps. Studies of force-field adaptation suggest a reliance on motor primitives that resemble the tuning curves of cerebellar Purkinje cells. These neurons have two peaks in their tuning curves, one for hand velocity in the cell's PD, and a smaller peak for the opposite direction. This tuning property could account for the transfer of learning from a practiced movement direction to the opposite direction. Furthermore, patients with cerebellar disease adapt poorly, if at all, to forces imposed on the limb during a movement. Unlike healthy subjects (**Figures 3(a) and 3(c)**, blue), who decrease their reaching errors as they gain experience in counteracting imposed forces, cerebellar patients continue to make errors without noticeable improvement (**Figures 3(a) and 3(b)**, green). Cerebellar diseases also affect the learning of other motor skills, including mirror writing and learning to copy shapes rapidly.

It seems likely that sometimes the motor system uses multiple controllers that can be selected and switched according to task context. Adapting to two force fields based on contextual cues, such as color, may be impossible under certain circumstances, but when random and frequent switching of context occurs, people can learn multiple, parallel skills, based on mappings cued by context. Learning a prism skill is also possible. Given a contextual cue, people can immediately call up the kinematic mappings needed for a given prism-induced displacement of the visual field.

Adjusting Ongoing Movements

In addition to adjusting motor commands from movement to movement, the motor system can adjust

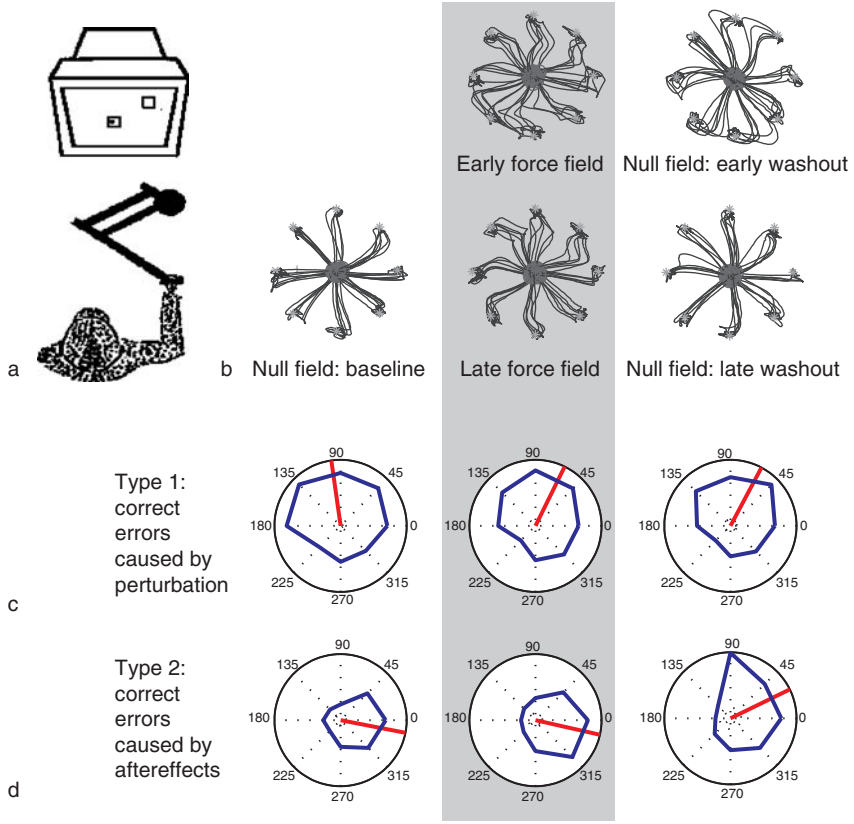

Figure 2 Perturbation of the dynamics map: (a) The monkey grasps a handle, and motors in that handle exert variable forces during a reaching movement. The monkey reaches one of eight directions from center to a visible target. In a baseline period, the motors do not apply forces, a condition called the 'null field.' (b) After practice in the force field, the monkey makes straighter movements, demonstrating the learning of a new dynamics map (see **Figure 1**). When the force disappears, aftereffects occur, which also decrement with practice (early washout vs. late washout). (c, d) The activity of two neurons in the primary motor cortex (M1). Activity for each of the eight movement directions is depicted in polar coordinates (blue), with the cell's PD in red. (c) As the monkey adapts to the clockwise force, the PD of one M1 cell shifts by approximately 30° clockwise. As a population, PDs in M1 cells shift by 16°. After a return to the null field (washout), the monkey readapts as the kinematics and dynamics return to baseline conditions. If M1 neurons processed only movement dynamics, all PDs should return to baseline in the washout condition, and many do. But in this and other M1 neurons, the changes in PD outlast the perturbation. (d) In a different M1 cell, a shift in PD occurs during readaptation to the null field. Overall, 41% of M1 cells exhibit the properties shown in (c, d), in equal proportions. (a, c, d) Adapted from Li CS, Padoa-Schioppa C, and Bizzi E (2001) Neuronal correlates of motor performance and motor learning in the primary motor cortex of monkeys adapting to an external force field. *Neuron* 30: 593–607, with permission from Elsevier. (b) Adapted from Padoa-Schioppa C, Li CS, and Bizzi E (2004) Neuronal activity in the supplementary motor area of monkeys adapting to a new dynamic environment. *Journal of Neurophysiology* 91: 449–473, with permission from American Physiology Society.

ongoing movements. Adjustments become necessary if movements begin inaccurately, if the target moves during a reach, or if some force perturbs the limb, among other reasons. If the pencil in **Figure 4** rolls away during the movement, an updated computation of the difference vector will restore a trajectory toward the target.

Much like some of the sequence learning discussed below, this kind of online correction can occur without conscious awareness. For instance, when people reach to a target that jumps to a new location during a movement, they often deny that it jumped at all. When told to suppress a movement to a target if it jumps, they cannot do so. It appears that a mechanism like that shown in **Figure 4** functions something like the automatic pilot of an airplane, adjusting to perturbations of the limb or unexpected movements of the target with fast, corrective actions. Patients with PPC lesions lack these autopilot adjustments, as do healthy people when transcranial magnetic stimulation disrupts PPC activity.

According to the model illustrated in **Figure 4**, the ability to adjust ongoing movements depends on a forward model, which predicts proprioceptively sensed hand movements based on ongoing motor commands, called efference copy. In essence, they predict where the hand should be at a given time after a motor command, taking feedback delays into account. Why should the motor system predict the sensory consequences of motor commands? First, a

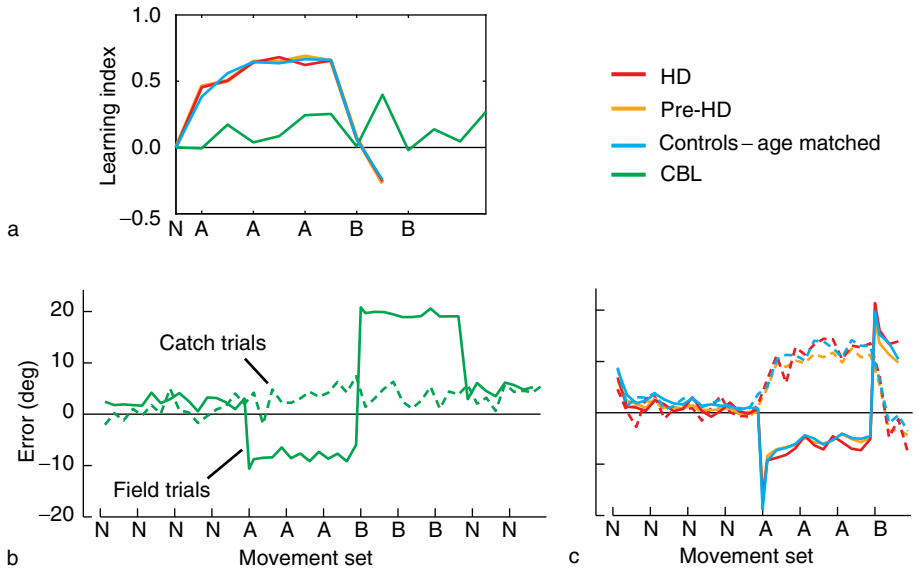

Figure 3 Effects of cerebellar disease on motor learning: (a) Learning index for movement-to-movement learning of force fields. Prior to imposing a complex pattern of forces, participants make straight movements to visible targets in a null field (N). The participants then learn to make straight movements in one force field, of pattern A, which is the same as used in **Figure 2**. Subjects practice in field A for three blocks of trials. HD (red) and control (blue) participants learn to counteract the imposed forces and return to making relatively straight movements, indicated by the relatively high learning index. This learning does not transfer to the opposite pattern of forces, pattern B. Patients with cerebellar damage (green) show very little movement-to-movement learning. (b) Data leading to B for cerebellar patients. After four blocks of movements in the null field, the patients practiced for three blocks in the A field. Once every seven movements, a catch trial (dashed line) occurred in which the null field was presented. The lack of change in error and the lack of aftereffects in catch trials indicates a severe learning deficit. (c) HD (red) patients and asymptomatic carriers of gene for HD (orange) learn as effectively as control subjects, as indicated by the decrease in error for the second and third blocks in the A field and the aftereffects on the corresponding catch trials. Reprinted from Smith MA and Shadmehr R (2005) Intact ability to learn internal models of arm dynamics in Huntington's disease but not cerebellar degeneration. *Journal of Neurophysiology* 93: 2809–2821, with permission from the American Physiological Society.

forward model allows the motor system to monitor motor output and predict the sensory state of the limb ahead of the actual sensory feedback. If neural noise, for example, causes the motor system to generate erroneous motor commands, they can be corrected before sensory feedback arrives. Second, once feedback arrives, the motor system has two sources of information to guide action: the predicted state of the limb and the sensed state. By combining them, it acquires a more accurate (less noisy) estimate of limb state than it can from either source alone.

Research on Huntington's disease (HD) suggests that parts of the basal ganglia contribute crucially to these computations. Although HD involves more than the basal ganglia, it is most involved in the disease, especially in its early phases. Even earlier, asymptomatic carriers of the HD gene have a subtle reaching abnormality that worsens once symptoms appear. Unlike cerebellar patients (**Figures 3(a) and 3(b)**, green), HD gene carriers (**Figures 3(a) and 3(c)**, orange) and symptomatic HD patients (**Figures 3(a) and 3(c)**, red) learn to correct reaching errors in a force field as well as controls (**Figures 3(a) and 3(c)**, blue). Instead, they have problems adjusting ongoing movements based on error feedback, and make jerky movements as they approach the target.

In summary, neural networks involving the cerebellum, motor cortex, and PPC adjust perceptual–motor mappings for the next movement based on the errors encountered on previous ones, whereas networks involving the basal ganglia and cortex adjust ongoing movements.

Sequential Movements

Although it takes considerable skill to make a single movement with spatial and temporal accuracy, the epitome of skilled behavior – exemplified by writing, speaking, typing, and dancing – entails learning a limited set of movements and then combining them into new sequences. It might seem that generating a sequence of movements need not differ fundamentally from generating a single movement – the motor system could simply select a series of movement targets, in turn. But the motor system does not plan sequences in this manner. If it did, then it would complete the first movement in a sequence before planning and completing the next. Skilled movements occur much too fast for this sort of planning. A skilled typist initiates the second keystroke of a sequence long before the motor system receives any proprioceptive feedback about the first. The fact that

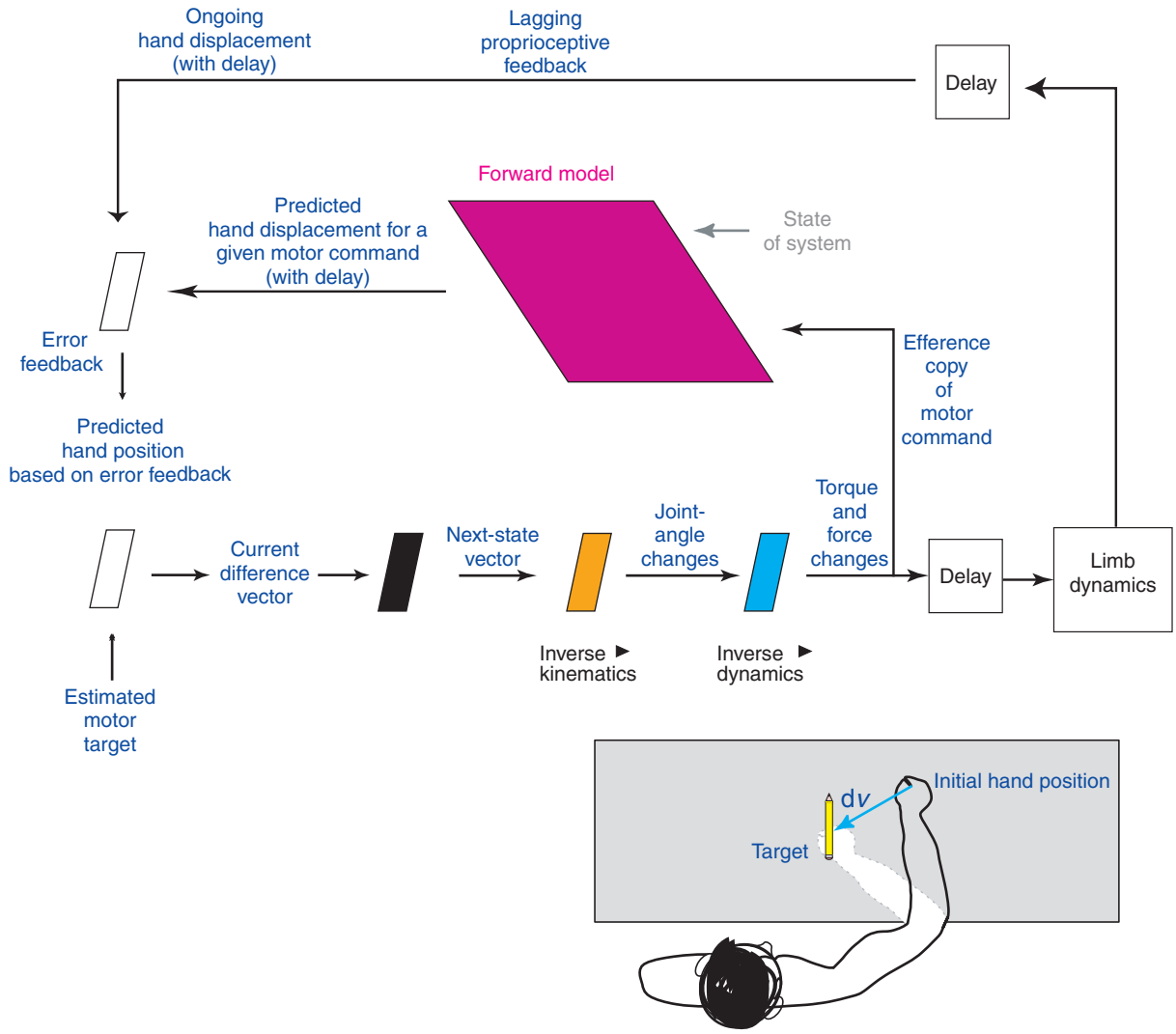

Figure 4 Forward models predict feedback and learn to correct ongoing movements. Format as in **Figure 1**, but incorporating feedback and the prediction of feedback. Forward models compute the movement that should occur given a motor command. (Inset) Adapted from Shadmehr R and Wise SP (2005) *The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge: MIT Press, with permission from MIT Press.

the time to make the first movement in a sequence increases with sequence length further demonstrates that the motor system plans sequences coherently, not piecemeal.

Perhaps the most fundamental problem to be solved in sequencing lies at its heart: the ability to combine a limited set of elements into a limitless set of motor acts. But how then can the motor system distinguish sequences? For example, how can it plan to type 'pig' and 'gap' and do both efficiently? In one case, 'g' follows 'p,' and in the other, 'p' follows 'g.' Simple associations chained together will not suffice: the motor system needs some mechanism for sequence planning, and this mechanism makes sequencing powerful.

Control of Sequencing

There is a considerable overlap in the neural machinery that supports moment-to-moment control of motor sequencing and that which supports sequence learning. **Figure 5** shows a heuristic representation of a linear organization of cortical areas that contribute to motor sequencing. Sequencing is initiated by the dorsolateral prefrontal cortex (DLPFC), and perhaps other prefrontal areas, which may be considered to employ more cognitive representations, meaning that they are flexible in terms of how sequencing behavior may be expressed. The hierarchy of cortical areas culminates in M1, which employs the least flexible representation, meaning that sequencing is closely tied to motor expression in terms of force production.

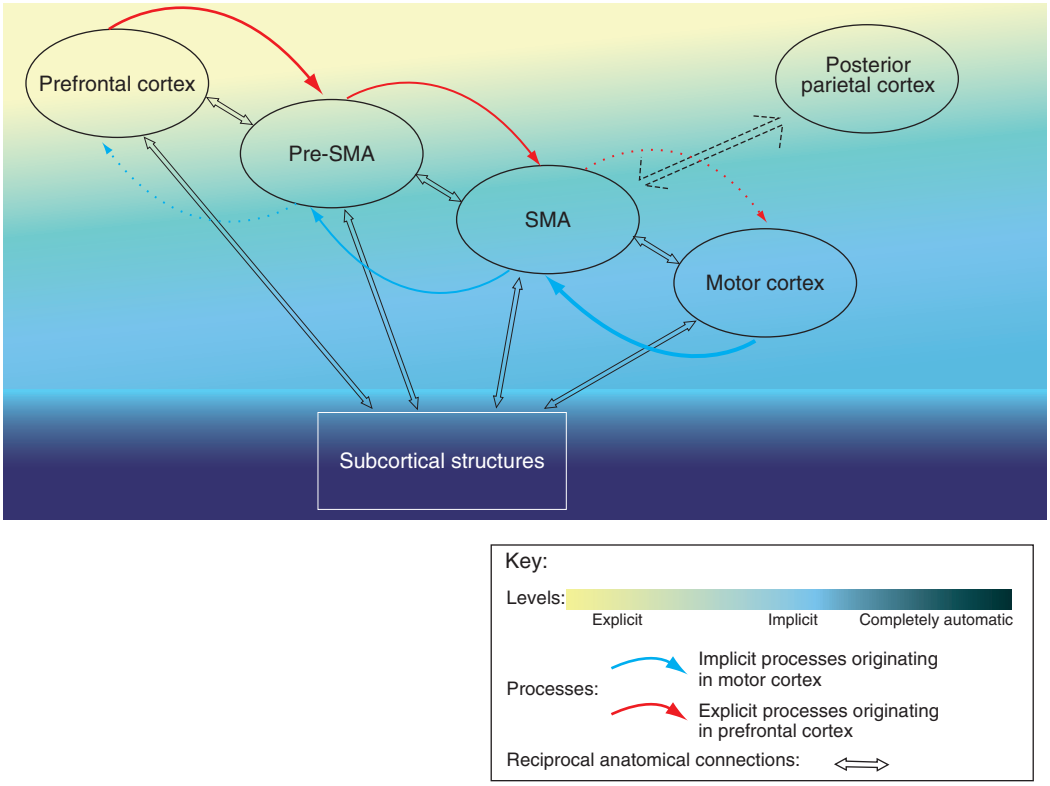

Figure 5 A schematic of cortical structures that underlie sequence learning. The darker regions are not available to awareness, but lighter regions are. Explicit learning originates in the DLPFC, and is expressed via the supplementary motor area (SMA), the pre-SMA, and the motor cortex. The red arrows show that, with practice, motor areas show learning consistent with the explicit sequence – that is, automaticity is developed. Implicit motor sequence learning progresses in the opposite direction, beginning in motor areas. According to this model, with sufficient practice this information propagates to prefrontal cortex and so becomes accessible to awareness. Reprinted from Ashe J, Lungu OV, Basford AT, and Lu X (2006) Cortical control of motor sequences. *Current Opinion in Neurobiology* 16: 213–221, with permission from Elsevier.

Among the areas illustrated in Figure 5, DLPFC has long been implicated in temporal order, dating to neurological observations of sequencing deficits in patients with frontal lesions. Neurophysiological studies show that these symptoms cannot be accounted for solely on the basis of dysfunction of an executive planner or deficits in working memory: some DLPFC neurons code temporal order, others code object identity, and yet others integrate object and order information. The pre-supplementary motor area (pre-SMA) receives prominent projections from DLPFC and sends projections to the SMA. There is overlap in the functionality of the pre-SMA and SMA; both contain neurons that code ordinal information (e.g., the second movement in the sequence, irrespective of the specific movement), neurons that code sequence initiation or the transition from one movement to another, and neurons that code a specific movement, but only for a particular sequence. For example, some cells become active when a monkey will push a lever, but only when it follows a pulling movement (Figure 6). Pre-SMA shows a higher proportion of neurons for order, initiation and transition, and the SMA for specific movement pairs, such as pull–push. Striatal neurons show similar properties, as might be expected due to the substantial projections from SMA and pre-SMA to the striatum, and projections that loop back to these cortical areas from striatum, via other basal ganglia structures and the thalamus. The SMA also projects to M1, which has recently been shown to be sensitive to sequencing. Activity patterns for simple movements change in M1 depending on the sequence context, and reversible blockade affects sequence production while having little impact on single movements.

Sequence Learning Tasks

Most of the work on motor sequencing has examined sequencing of spatial responses, and much of it has used a model task: the serial reaction time task (SRTT). The subject sits before a computer monitor, with the middle and index finger of each hand resting lightly on four response buttons. On each trial, a stimulus appears in one of four possible locations, and the subject presses the button directly below the stimulus. The stimulus disappears, and after a brief

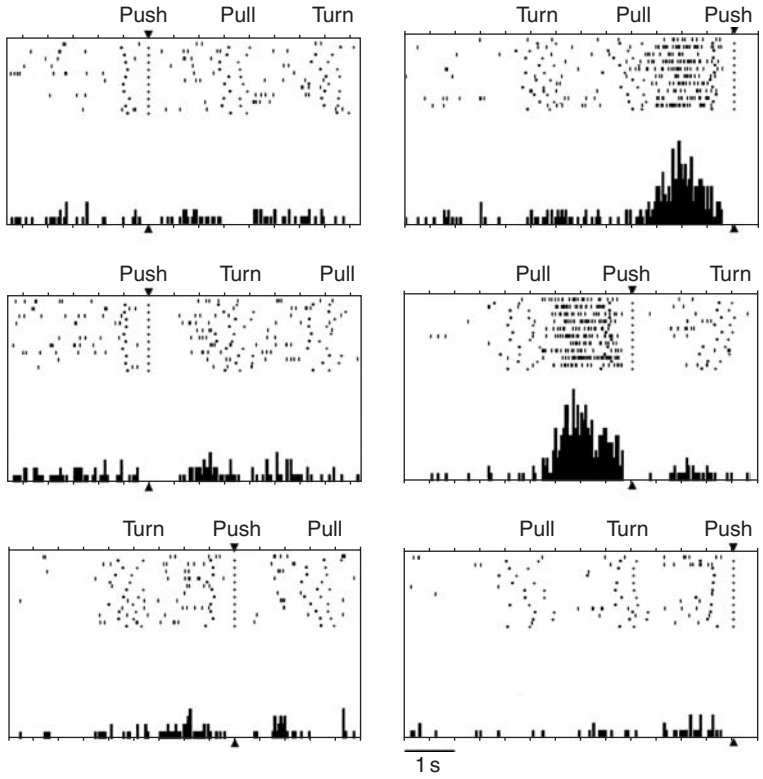

Figure 6 A SMA neuron that is active before a push movement, but only if the monkey has just completed a pull movement. In the raster display, each row shows the neuron's activity for one trial, the plus represents movement initiation, and each dot represents one discharge of the neuron. The histogram indicates the activity average. Reprinted from Shima K and Tanji J (2000) Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology* 84: 2148–2160, with permission from the American Physiological Society.

delay a new stimulus appears. This procedure continues for several hundred trials. For some trials the stimuli appear randomly, and for others they appear in a repeating sequence, typically 12 units long. This task has been used in hundreds of sequencing experiments because of its simplicity, and because it can, with little modification, be used to test two types of sequence learning: implicit or explicit. In implicit sequence learning, the subject remains unaware that he or she is learning anything. If subjects are not informed about the presence of the sequence, they often remain unaware of it. Nevertheless, their response times to individual stimuli decrease with practice, and increase if the sequence changes. Despite this implicit learning, subjects perform at chance on a recall or recognition test of the sequence. Hence, subjects show behavioral evidence of learning, despite their unawareness of the sequence.

The same task may be used to measure explicit sequence learning, in which the subject is aware of learning the sequence. In this version, the experimenter simply informs the subject that the stimuli appear in a sequence, and encourages the subject to learn it, as a way of responding more quickly. With these instructions, subjects will exhibit sequence learning through faster response times to the sequenced stimuli than to the random stimuli, and will also successfully produce the sequence during a later recall or recognition task.

Tracking tasks have also proven useful in studies of sequencing. In a standard pursuit tracking task, the subject uses a manipulandum to control a cursor, which is to be kept as close as possible to a moving target on a screen. The target appears to move randomly, but a repeating sequence is embedded in the otherwise random movement, as shown in Figure 7. Learning is assessed by comparing tracking error in the sequenced and random segments. Such learning is virtually always implicit and, in this respect, resembles the force adaptation learning illustrated in Figure 2(b).

Implicit Sequence Learning

Behavioral studies indicate that when subjects learn a sequence implicitly, they learn a sequence of targets for the movements. That is, they do not learn the sequence of stimulus locations, nor a sequence of specific motor responses in terms of joint-angle changes or forces. Instead, they learn the sequence of targets to which movements will be made. After

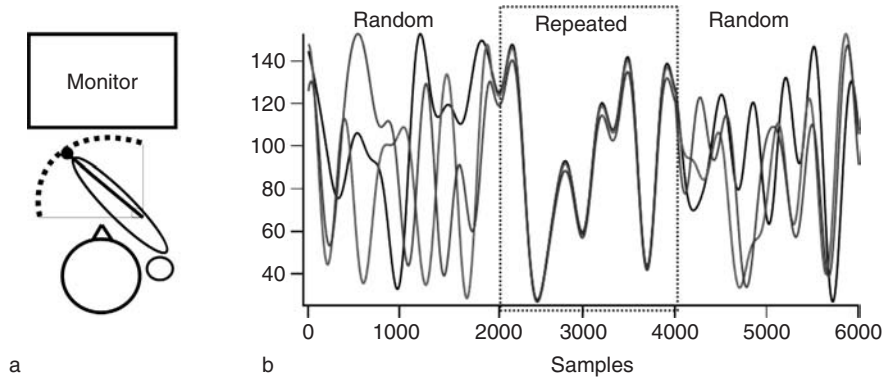

Figure 7 A continuous tracking task: (a) The subject sits before a computer screen with the forearm resting on a lever, capable of movement via shoulder rotation through a 90° arc. Lever movement manipulates a screen cursor, which the subject uses to track a moving target. The target moves in one dimension, from the subject's left to right. (b) Example of target movement on three different trials. Note that the segments for the first and last (third) trials differ (and are random) but the intervening segment repeats for each trial. Reprinted from Boyd LA and Winstein CJ (2004) Cerebellar stroke impairs temporal but not spatial accuracy during implicit motor learning. *Neurorehabilitation and Neural Repair* 18: 134–143, with permission from Sage Publications.

training, the stimulus locations can be changed, or the effectors used to respond can change, without any cost to learning, as long as the sequence of target locations remains the same.

These behavioral data aid in interpreting neuroimaging and neurophysiological studies of implicit sequence learning. Neuroimaging studies of the SRTT and of the tracking task (Figure 7) show consistent activation in four brain regions: the SMA and pre-SMA, striatum, PPC, and cingulate cortex. The PPC represents the sequence of targets in an abstract spatial code (Figure 1, pink shading), and the striatum and SMA represent it in a motor code, consistent with neurophysiological data implicating these structures in motor sequencing. Cingulate cortex may monitor the success of motor prediction, as indicated by the findings that it is especially active during sequence violations and that its activity decreases as the skill develops and errors decrease. Consistent with this interpretation, lesions to SMA or PPC devastate implicit sequence learning. Striatal dysfunction caused by HD or Parkinson's disease also leads to deficits in sequence learning, although only partially, likely because patients in a testable stage of the disease have only incomplete striatal damage.

Cerebellar activation is observed in some neuroimaging studies, and cerebellar patients show drastically impaired implicit sequence learning. But behavioral paradigms that separate learning and the behavioral expression of learning in the SRTT show cerebellar activation associated with the latter, not the former. This interpretation is consistent with the putative role of the cerebellum in timing; it may calculate when the next position in the sequence will appear, which facilitates motor preparation, but not the identity of the next sequence element for which to prepare. Also relevant are data from the tracking task. Cerebellar patients are impaired on this task, but only on the timing component; the spatial response profile shows normal learning of the sequence, but their responses lag behind the target and the timing of responses is variable. Thus, the cerebellum appears to support the behavioral expression of implicit sequence learning, but does not appear to be the site of learning the specific motor commands.

Explicit Sequence Learning

Explicit sequence learning has been examined in the SRTT, and in conceptually similar studies in which subjects repeatedly perform a simple sequence of finger-thumb opposition movements. Behavioral studies show that learning in these explicit sequencing tasks is based on knowledge of the stimulus locations, in contrast to implicit sequences, which are learned as sequences of target locations.

Neuroimaging studies show consistent activation in PPC and in DLFPC. These two regions are associated with spatial working memory, attention, and goal selection, consistent with the interpretation that explicit sequence learning is not tied to the motor system, at least early in training. Explicit learning is open to flexible expression, for example through verbal responses or through motor responses different than those used in training. Structures linked to explicit learning, such as the hippocampus and other parts of the medial temporal lobe, are also active. Furthermore, amnesic patients, who often have damage somewhere in the white or gray matter of the medial temporal lobe, are profoundly impaired in explicit sequence learning, although they learn implicit sequences normally.

As expected from the discussion above, the SMA and pre-SMA are also consistently active in neuroimaging studies of explicit sequence learning. Although

no one knows whether monkeys learn these tasks explicitly or implicitly, neurophysiological studies in monkeys show that the pre-SMA is especially active during the learning of new sequences, and that once the sequence is well learned, some pre-SMA neurons discharge only at the initiation of the sequence. Furthermore, functional blockade of pre-SMA impairs the learning of new sequences by trial and error. Thus, pre-SMA activity in monkeys has some of the properties to be expected for explicit sequence learning.

Practice-Related Changes

A salient behavioral consequence of practicing an explicit skill is the decreasing role of attention, and eventual automaticity of behavior. Characterizing the neural changes associated with long-term practice has been difficult because of variation in experimental designs and because behavior changes with training – skills are performed more quickly and with fewer errors – and those changes must be accounted for when interpreting physiological changes. Figure 5, based on neuroimaging studies of long-term sequence learning in humans, depicts a rough summary of the pattern that has emerged. The red arrows represent the progression of explicit learning, which DLPFC and other prefrontal areas are thought to underlie. With sufficient training, motor areas become tuned to the specifics of the explicitly learned sequence, and then the contribution of prefrontal cortex diminishes. The blue arrows represent changes in implicit learning, which begins in motor areas, and with sufficient practice may propagate to prefrontal cortex, which presumably contributes to awareness of the sequence.

As skilled performance becomes more automatic, attentional resources can be directed to other matters, such as perceptual decisions and the choice of future actions. Although the implicit nature of highly practiced skills has led some to consider them as a 'lower' form of learning, the early-twentieth-century mathematician, Alfred North Whitehead (1911), put the matter in its proper perspective when he noted that it was mistaken to "cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them."

See also: Motor Control of Feeding and Drinking; Procedural Learning in Animals; Procedural Learning in Humans; Procedural Learning: Cerebellum Models.

Further Reading

Ashe J, Lungu OV, Basford AT, and Lu X (2006) Cortical control of motor sequences. *Current Opinion in Neurobiology* 16: 213–221.

Bischoff-Grethe A, Goedert KM, Willingham DT, and Grafton ST (2004) Neural substrates of response-based sequence learning using fMRI. *Journal of Cognitive Neuroscience* 16: 127–138.

Boyd LA and Winstein CJ (2004) Cerebellar stroke impairs temporal but not spatial accuracy during implicit motor learning. *Neurorehabilitation and Neural Repair* 18: 134–143.

Hikosaka O, Nakahara H, Rand MK, et al. (1999) Parallel neural networks for learning sequential procedures. *Trends in Neurosciences* 22: 464–471.

Krakauer JW, Ghilardi MF, and Ghez C (1999) Independent learning of internal models for kinematic and dynamic control of reaching. *Nature Neuroscience* 2: 1026–1031.

Li CS, Padoa-Schioppa C, and Bizzi E (2001) Neuronal correlates of motor performance and motor learning in the primary motor cortex of monkeys adapting to an external force field. *Neuron* 30: 593–607.

Osu R, Hirai S, Yoshioka T, and Kawato M (2004) Random presentation enables subjects to adapt to two opposing forces on the hand. *Nature Neuroscience* 7: 111–112.

Padoa-Schioppa C, Li CS, and Bizzi E (2004) Neuronal activity in the supplementary motor area of monkeys adapting to a new dynamic environment. *Journal of Neurophysiology* 91: 449–473.

Paz R and Vaadia E (2004) Learning-induced improvement in encoding and decoding of specific movement directions by neurons in the primary motor cortex. *PLoS Biology* 2: 264–274.

Paz R, Wise SP, and Vaadia E (2004) Viewing and doing: Similar cortical mechanisms for perceptual and motor learning. *Trends in Neurosciences* 27: 496–503.

Seidler RD, Purushotham A, Kim SG, Ugurbil K, Willingham D, and Ashe J (2005) Neural correlates of encoding and expression in implicit sequence learning. *Experimental Brain Research* 165: 114–124.

Shadmehr R and Wise SP (2005) *The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge: MIT Press.

Shima K and Tanji J (2000) Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology* 84: 2148–2160.

Smith MA, Brandt J, and Shadmehr R (2000) Motor disorder in Huntington's disease begins as a dysfunction in error feedback control. *Nature* 403: 544–549.

Smith MA and Shadmehr R (2005) Intact ability to learn internal models of arm dynamics in Huntington's disease but not cerebellar degeneration. *Journal of Neurophysiology* 93: 2809–2821.

Whitehead AN (1911) *Introduction to Mathematics*, p. 61. New York: Henry Holt.

Willingham DB (1998) A neuropsychological theory of motor skill learning. *Psychological Review* 105: 558–584.

Wise SP, Moody SL, Blomstrom KJ, and Mitz AR (1998) Changes in motor cortical activity during visuomotor adaptation. *Experimental Brain Research* 121: 285–299.

Wolpert DM and Kawato M (1998) Multiple paired forward and inverse models for motor control. *Neural Networks* 11: 1317–1329.

Motor Timing

A Wing, University of Birmingham, Edgbaston, Birmingham, UK
R M Bracewell, University of Wales, Bangor, UK

© 2009 Published by Elsevier Ltd.

Issues

Accurate timing of movements is a key element of most motor skills and includes the ability to coordinate separate elements of an action and to synchronize the action with external cues. For instance, in tennis, good relative timing of stepping and swinging the racquet with the arm results in an accurately directed shot, but this must be synchronized with the approach of the ball. In a music duet, each player produces series of notes timed according to the score but must also maintain relative phase with the other player. In these examples the timed intervals fall in the tens and hundreds of milliseconds range. In the study of motor timing, the issues are defined by seeking answers to important questions: What brain mechanisms support motor timing? Is there evidence of hierarchical timing with more complex temporal patterns represented at higher, cortical levels? Are there different timing systems for paced and unpaced responding? Is there generality of timing across different movement systems or is timing modality specific? How are movements of separate limbs coordinated? How is timing integrated with other aspects of movement such as force?

Experimental Tasks

A common paradigm in the study of motor timing, providing control over interresponse interval, comprises an initial period of paced responding, in which movement events (e.g., tapping the finger on a response button that records the time of each response) are synchronized with a metronome, followed by unpaced tapping at the same rate (see Figure 1(a)). In the paced phase, synchronization error between each tap and corresponding metronome pulse is the main measure of interest. Synchronization error is typically found to exhibit a small lead (negative onset asynchrony). People are very sensitive to changes in synchronization error. For instance, an advance in metronome phase, produced by one metronome interval being longer than the rest, causes a reduction in negative onset asynchrony. This results in compensatory timing adjustment to produce an advance in response phase, even when the metronome phase shift is so small that it is not consciously detectable (Figure 1).

In the unpaced phase of tapping, measures of interest include the mean and variability of the intervals produced. A commonly observed characteristic is that the variability increases with the mean, with the ratio of the standard deviation to the mean (Weber fraction) approximately constant except for an increase as the interval goes to zero, partly due to the positive intercept at zero (see Figure 1(b)). The form of the variability in unpaced responding is also of interest, with successive interresponse intervals commonly exhibiting a short–long pattern (negative lag-one autocorrelation), while, over the longer term, there is often a tendency for drift, producing positive autocorrelation at long lags. In some studies the pacing sequence comprises a rhythmic pattern, rather than all equal intervals (isochronous sequence). In such cases it is found that the longer intervals in the rhythm are more variable than the shorter ones and the overall level of variability is higher than when the same intervals are produced in an isochronous context.

Tapping the finger on a button introduces a clear event, or response, within what might otherwise be a smooth cycle of movement. A variant of the paced–unpaced tapping paradigm is one in which the movements are made with continuously smooth direction change, as in circle drawing or flexing and extending the finger without contact. In the unpaced phase, timing of continuously smooth circling or flex–extend movements are less variable than for discontinuous event-based movements. However, with continuous movements, the synchronization error in the paced phase is greater. Such differences suggest contrasts in the underlying nature of timing, with, for example, corrective adjustment in paced but not in unpaced tapping, but also emphasize the need to document the form of movement when reporting the accuracy of motor timing.

A contrast with timing involving just a single effector is provided by a paradigm focusing on bilateral interlimb coordination, with two effectors moving with specified relative phase. For instance, left and right index fingers might flex concurrently giving 0°, in-phase movements. Or flexion of one digit might be combined with extension of the other, resulting in 180° out-of-phase movements. A characteristic finding is that the stability of relative phase for in-phase movements is greater than for movements at other phases. Another important finding is that the stability of relative phase for out-of-phase movements decreases as frequency increases (the movements are typically paced by a metronome, the rate of which is progressively increased), even to the extent that relative phase undergoes a spontaneous transition to the more stable in-phase movement (see Figure 2).

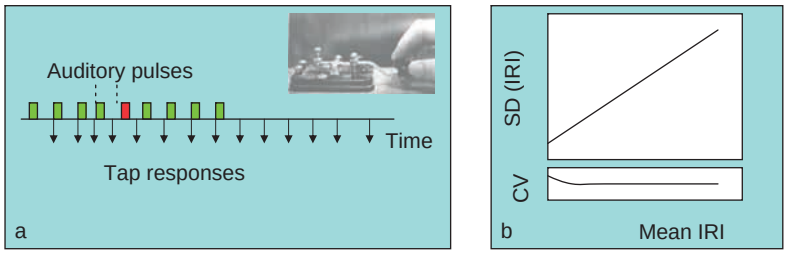

Figure 1 (a) Response key with time line showing paced (with phase shift) and unpaced (interval) tapping phases. A metronome phase delay (at red pulse marker) results in rapid compensation in tapping. (b) Increase in interresponse interval (IRI) variability with the mean interval.

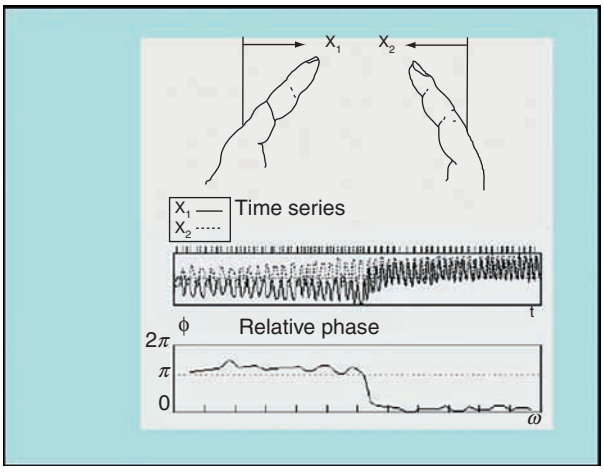

Figure 2 Isofrequency phase transition.

The form of movement in interlimb coordination studies is usually smoothly continuous (e.g., making circles in the air). It is possible that synchronization between two effectors is difficult and control of phase is unstable, because there are no discrete events in the two effectors that can be used to monitor produced phase. If event structure is introduced (for instance, by requiring tapping on a surface), relative phase becomes more stable. This effect, which is referred to as anchoring, presumably relates to improved synchronization seen with discontinuous, event-based tapping, compared to circling. A variant of the interlimb coordination task is one in which the two limbs move at different frequencies. In this case certain frequency combinations are found to be more stable than others (e.g., 2 against 3 is more stable than 4 against 5). Increasing the rate of responding tends to result in transitions from less stable to more stable frequency combinations.

Models

A widely cited theoretical account of the variability of unpaced interresponse intervals in single-effector timing assumes dissociable sources of variance due to central timing and the implementation of successive responses (see Figure 3). This two-level account separates out the variance contributions of timer intervals and delays in motor implementation, and predicts the negative lag-one autocorrelation generally seen in unpaced interresponse interval data, although this may be partly masked by drift in central timing.

A second focus of motor timing models has been the nature of the coupling, which maintains relative phase between motor events and the metronome in the paced tapping task, or between movement trajectories produced by each of the two limbs in the interlimb coordination task. In the former, linear first-order correction gives a good account of the data while nonlinear coupled oscillator models provide important insights into the latter (see Figures 3(b) and 3(c)). In the coupled oscillator models a contrast is made between instantaneous and delayed linkage between effector states. Instantaneous linkage is suggestive of central origin of the coupling – for instance, due to motor command interactions. Coupling that includes a delay between effector states, which might be identified with feedback delays, is consistent with placing the origin of coupling effects at the level of motor consequences rather than at the motor command level.

The form of the central timer or oscillator has been a matter of debate. In one perspective, a neural pulse source is assumed and motor timing depends on comparing an accumulated count with a remembered standard (see Figure 4(a)). If the neural event source is modeled by a Poisson process (random times between events), the variance of the interval being timed increases linearly with the mean. However, in motor timing it is usually found that standard deviation, rather than variance, increases with the mean (recall Figure 1(b)), and this suggests an additional source of variance that is rate dependent. For instance, the counter mechanism might be error-prone in proportion to the total count, perhaps due to inattention. Consistent with a role for attention in the central timer is the finding that performance of a secondary task

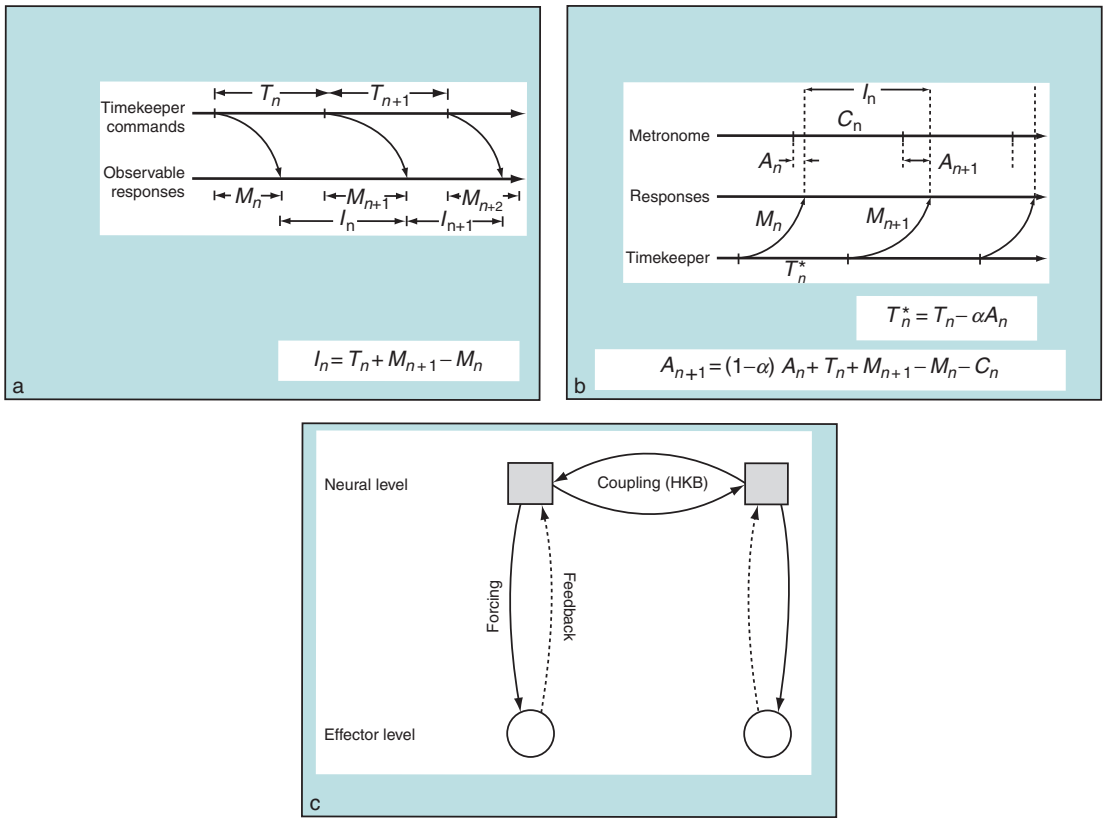

Figure 3 (a) Wing and Kristofferson model of timing: interresponse intervals in unpaced responding. (b) Linear phase correction model. (c) Coupled (two-level) oscillator model (including implementation delay to account for negative lag-one autocorrelation). The Haken–Kelso–Bunz (HKB) equation expresses the dynamics of interlimb rhythmic coordination. (a) From Wing AM and Kristofferson AB (1973) Response delays and the timing of discrete motor responses. *Perception and Psychophysics* 4: 5–12. (b) From Vorberg D and Wing AM (1996) Modelling variability and dependence in timing. In: Heuer H and Keele S (eds.) *Handbook of Perception and Action*, vol. 2, pp. 181–262. New York: Academic Press.

(e.g., solving anagrams) reduces timer accuracy, but not motor implementation variability, during tapping. Other alternatives to accumulative counter-based timing have been proposed, including models involving sets of oscillators in which a combination of oscillators approximates the target interval or timing based on the spatial spread of activation, with longer intervals represented by greater neural distance.

Lesion Studies

Given various models of motor timing, a key question concerns how their component functions map onto brain structures. In this section and the next, we consider neural bases for timing from two perspectives, lesion and brain activation studies. In both cases we examine the extent to which motor timing functions can be localized to individual brain regions or networks of regions. Unpaced tapping is more variable in various neurological disorders, including Parkinson's disease (PD), cerebellar disorders, and hemiparetic stroke. Comparison of these three groups has indicated that the central timer is more variable in cerebellar disease, whereas motor implementation is more variable in PD and stroke (see Figure 5(a)). Another study revealed a double dissociation in timing deficits exhibited by cerebellar patients with medial versus lateral lesions, the former having elevated implementation variability and the latter having increased timer variability (Figure 5(b)). However, the reliability of cerebellar timing deficits is a matter of current debate, and not all classes of timing are affected by cerebellar disease. Patients with cerebellar lesions are selectively impaired on motor timing tasks with discontinuous movements such as tapping, but are not affected in circling movements that lack definite event structure. Other studies have demonstrated elevated timer variability, and no evidence of increased variability of motor implementation, in impaired basal ganglia function associated with PD and in right frontal cortical stroke. Thus, the suggestion that there is a clear mapping of clock timer variability and motor variability onto different brain structures remains contentious.

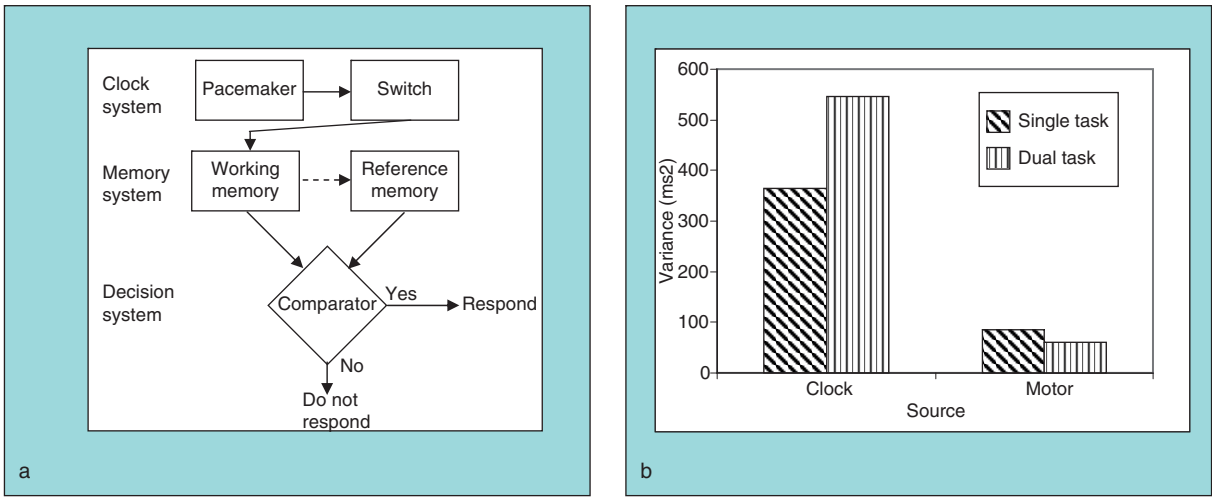

Figure 4 (a) Counter model of the timer. (b) Compared to baseline, concurrent secondary task increases timer, not motor implementation variance. (a) From Treisman and Gibbon. (b) From Sergent et al. (1993).

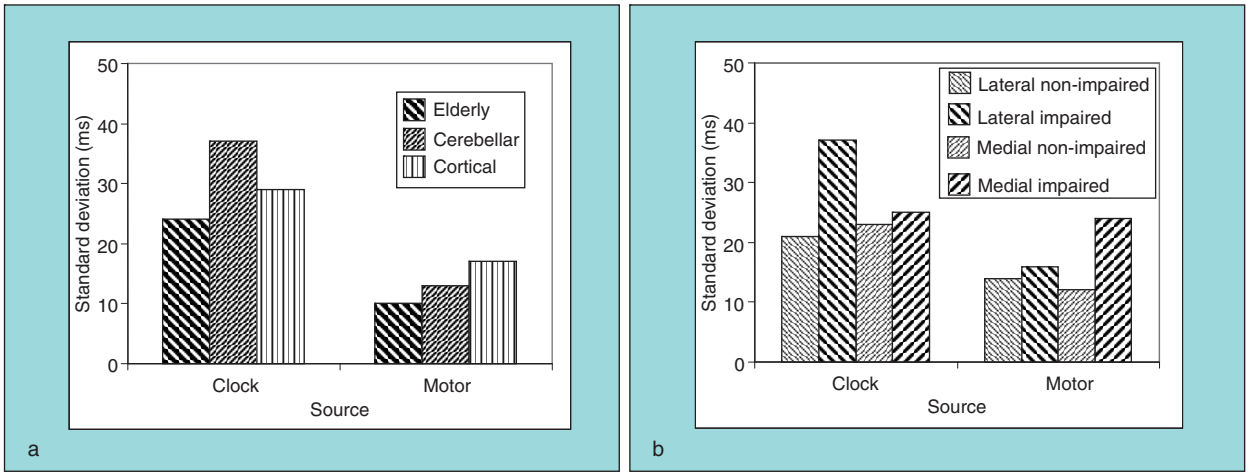

Figure 5 (a) Increased timer variability in study participants with cerebellar damage, compared to those with Parkinson's disease (PD) and stroke. (b) Medial vs. lateral cerebellar differences. (c) Phase correction in Parkinson's disease vs. cerebellar impairment. (a, b) From Ivry and Keele (1989). (c) After Diedrichsen et al.

So far we have focused on deficits in motor timing in unpaced responding. Maintenance of phase in paced tapping has also been investigated in cases of neurological disorders. Cerebellar patients exhibit larger clock variability, whereas PD patients show weaker error correction together with larger negative onset asynchrony (see Figure 5(c)). In an interesting single case study, a musician with impaired expressive and receptive music abilities following right temporoparietal stroke showed more variable timing but normal motor implementation. There was evidence of stronger correction of phase errors between responses and metronome pulses. This might have been due to strategic adaptation to impaired memory for temporal interval, or it may have reflected compensation for loss of tempo, evidenced as an increased tendency to drift in unpaced timing.

Transcranial magnetic stimulation (TMS) affords a means of introducing temporary disruption in cortical processing. In studies of motor timing this has been used, both in single-pulse TMS mode, delivered at various phases with respect to the tapping movements, and with repetitive stimulation (rTMS), using tests before and after to determine temporary disruptive effects on motor timing. rTMS over primary motor cortex reduces the phase lead (negative-onset asynchrony) in paced responding but, surprisingly, has no effect on the response to phase shift (see Figure 6). This suggests that these two aspects of paced responding reflect different underlying mechanisms. One possible interpretation of the changed asynchrony is that the interval being produced by the central timer is lengthened. Single-pulse TMS over motor cortex during unpaced tapping causes selective effects

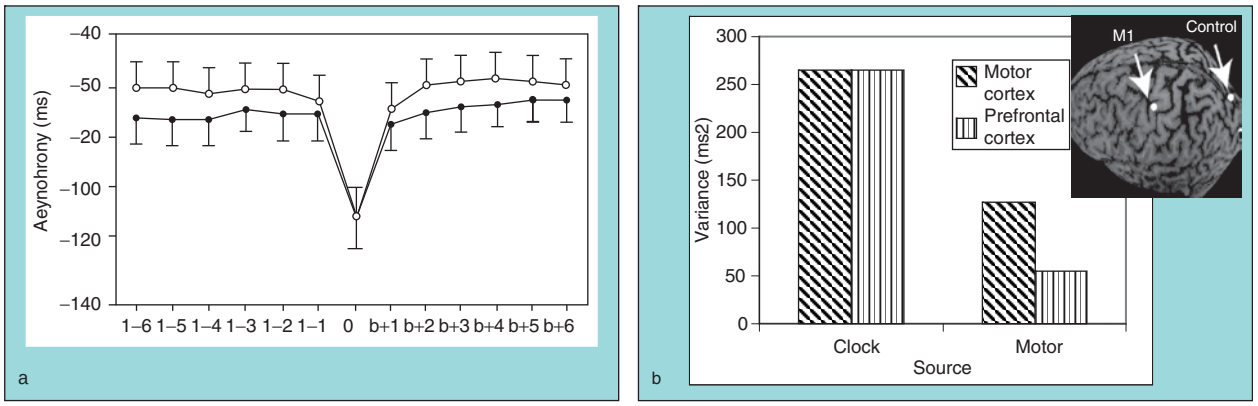

Figure 6 (a) Baseline, but not correction, affected by repetitive transcranial magnetic stimulation. (b) Motor implementation affected by single-pulse transcranial magnetic stimulation (TMS); timer variance not affected. (a) From Doumas. (b) From Verstynen.

on motor implementation but not timer variability (see Figure 6(b)).

Motor impairments in neurological disorders are often lateralized to a marked degree. For example, a patient with a left cortical stroke will show impairment of movement of the right (contralesional) limbs, whereas the deficit following a cerebellar stroke is on the same (ipsilesional) side. The motor symptoms of PD are typically quite lateralized, particularly in the early to moderate stages of disease that are most commonly studied. This has a methodological benefit for studies of motor timing in that each participant can serve as her own control. Thus, the affected side can be compared with the unaffected (or less affected) side. More importantly, such lateralization of motor timing function leads to the conclusion that there are two separate timing systems, one for either side of the body. It is then interesting to note that, in patients with lateralized lesions, bilaterally simultaneous movements result in timing that is intermediate between each side alone. This indicates that multilimb integrative processes operate in motor timing. Multisegment integrative effects have also been reported in normal participants, with two limbs showing less variable timing compared to either limb alone, an effect attributed to statistical benefits arising from cumulating the output of separate timers.

Patients with missing or incomplete corpus callosal connections between the two cerebral hemispheres also show a reduction in variance with bilateral movements, suggesting that subcortical, rather than cortically interconnected, structures support integration of the timers. Interestingly, this bilateral timing advantage depends on the movements having an event structure and disappears if the movements are continuous, again suggesting two distinct classes of motor timing behavior. From this one may infer that discrete movements are coupled at a subcortical level, whereas corticocortical connections underpin the coupling of continuous movements. In this context it is of interest that patients with left parietal lesions show marked difficulty in synchronizing their hands in making antiphase circling movements, which may rely on corticocortical connections.

As noted earlier, patients with PD may show impairment in timing actions with a single limb. The administration of drugs that enhance or depress dopaminergic neural transmission affects timing functions in healthy volunteers. For example, the dopamine antagonist haloperidol impairs perception of timing intervals of 500 ms or less. The most prominent deficit in PD is in dopaminergic innervation of the striatum by the substantia nigra. The striatum is that part of the basal ganglia that receives the largest input from the cerebral cortex. Therefore it is of interest to review studies of bilateral coordination in PD. There is general agreement that patients have great difficulty in maintaining antiphase movements of the arms, whereas there may be little or no impairment of in-phase movements, thus PD patients may be considered to show pathologically strong coupling between the sources of control for movements of each side.

It is interesting to compare the performance of PD patients with that of patients with Huntington's disease (HD). In the latter, the major pathology falls in the striatum rather than in the substantia nigra, and the dopaminergic innervation of the striatum from the substantia nigra is intact. Patients with HD produce bimanual cyclical movements with lower accuracy and greater variability than do controls. They are unable to produce antiphase movements, generating in-phase movements instead. Interestingly, the same study reveals a similar pattern in PD patients; however, the performance of PD, but not HD, patients during in-phase movements improves with an external auditory

pacing signal. One might speculate that the dopaminergic nigrostriatal system plays a specific role in conveying external timing signals to the striatum.

A lesion affecting one hemisphere of the cerebellum causes ipsilateral limb ataxia. Phase relationships between pairs of limbs during cyclical in-phase and antiphase movements have been investigated in a cerebellar patient with left arm and leg ataxia (see Figure 7). The phase relationships between the unaffected right arm and leg were normal, as expected. When one unaffected and one affected limb were moved together (the 'diagonal' conditions), the phase relationships were close to normal. However, the mean relative phase and its variability were very high when moving the left arm and left leg together. One possible explanation for these findings is that because the patient was unable to generate accurate timing signals for the left side, then maintenance of an accurate and stable phase relation between the left limbs was impossible, whereas in the diagonal conditions the right-sided timing signal was able to support both sides.

In summary, the human lesion literature suggests that a widespread network of cortical and, critically, subcortical areas is involved in motor timing. One controversial claim is that the cerebellum has an important role as a central timer, for both paced and unpaced actions. This may only be true for discontinuous movements, in which there are discrete motor events. In contrast, impaired performance following cortical and basal ganglia damage has been attributed to variability in motor implementation. There is evidence for largely independent timers for the two sides of the body; integration between the two sides may rely more on subcortical interconnections for discrete movements, and callosal connections for continuous movements. One interesting idea is that damage to the basal ganglia may result in pathologically strong coupling between the two sides. Comparison of PD and HD raises the possibility that the dopaminergic nigrostriatal pathway may be important in paced movements. With the marked exception of cerebellar studies, there has been little patient work contrasting discontinuous and continuous movements, which may well rely on different timing mechanisms.

Brain Activation Studies

One of the earliest brain activation (functional magnetic resonance imaging) study of motor timing involved paced followed by unpaced responding. The results indicated that the two modes of tapping draw on partially separate networks in the brain, with the network for continuation including that for synchronization as a subset. Increased activation when tapping at 300 or 600 ms with the right hand, compared to rest, involved right cerebellum, left sensorimotor cortex, and right superior temporal gyrus in both paced and unpaced phases of the task (see Figure 8). In addition, during unpaced, but not during paced, responding, activation was observed in the right inferior frontal gyrus, supplementary motor area (SMA), putamen, and ventrolateral (motor) thalamus. These results lead to the suggestion of a contrast between 'implicit' timing, in keeping pace with the metronome, and 'explicit' timing, in continuation. The latter was assumed to involve processes to sustain an auditory working memory representation of the target interval, based on right inferior frontal gyrus

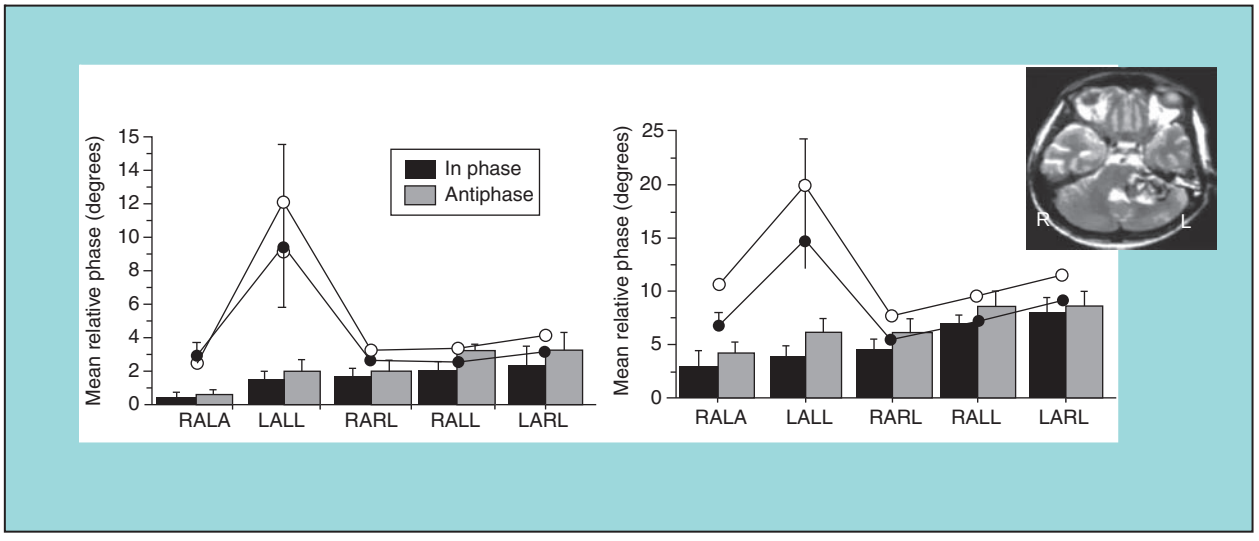

Figure 7 Upper–lower limb coordination. Impaired control of relative phase in cerebellar disease. RALA, right arm, left arm; LALL, left arm, left leg; RARL, right arm, right leg; RALL, right arm, left leg; LARL, left arm, right leg. From Bracewell et al. (2005).

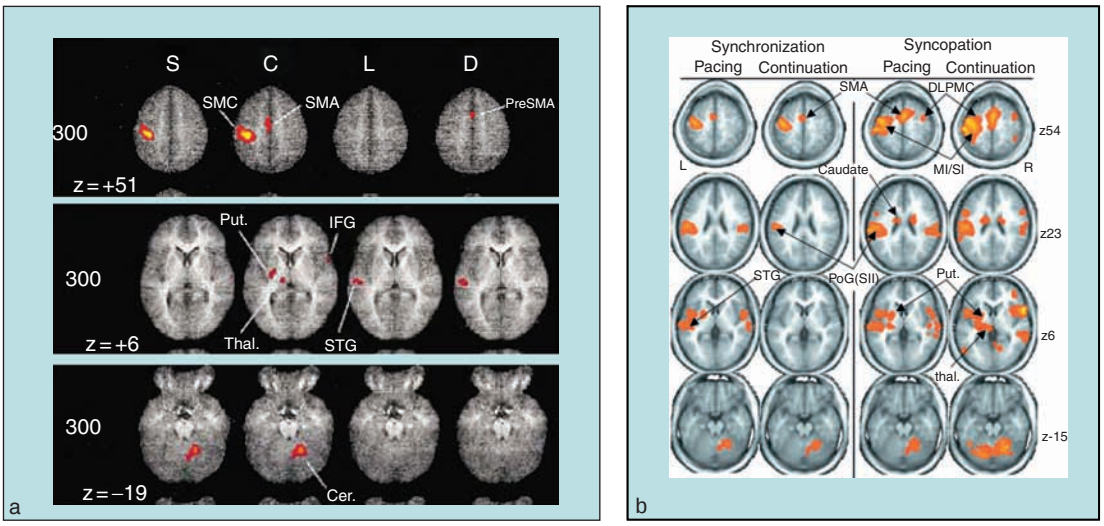

Figure 8 Activation studies: the functional magnetic resonance imaging of timing. Paced and unpaced tapping is compared with rest. SMC, supplementary motor complex; SMA, supplementary motor area; put, putamen; thal, thalamus; IFG, inferior frontal gyrus; STG, superior temporal gyrus; Cer, cerebellum. From Rao. (b) From Jantzen et al.

linked to right superior temporal gyrus, plus explicit timing control based on SMA, putamen, and ventrolateral thalamus.

A recent imaging study of motor timing emphasized similarities rather than differences between paced and unpaced tapping. Participants were asked to tap in phase (synchronize) or in antiphase (syncopate) with a metronome and then to continue with unpaced tapping after the metronome stopped. The results showed that activation was more widespread for syncopation than for synchronization, including increased activity in SMA, left premotor area (PMA), right thalamus, bilateral inferior frontal gyri, and cerebellum. Interestingly, there were no differences between the paced and unpaced phases, with the exception that auditory temporal cortex was not active in the absence of the metronome. That is, the difference between synchronized and syncopate tasks persisted into the unpaced phase, even though in this phase the movements were indistinguishable. It is intriguing that the brain state in unpaced responding should so closely reflect those in the paced phase, and it suggests that mental structures that are set up during pacing are preserved in unpaced responding. This might correspond to a subjective feel of swing that, set up by the offbeat syncopate task, is preserved even when the metronome stops.

Several imaging studies have sought to determine the neural underpinnings of sequential information in combination with motor timing. One important dimension of sequential complexity is rhythm, in which grouping of motor events derives from patterning of intervals. Given that rhythmic sequences are more complex than are equal-interval sequences, it is interesting to note that the cerebellum, already implicated in equal-interval timing, shows decreasing contribution with advancing skill level in rhythm production. Thus, in a study which tracked activation over several days of learning a temporal pattern, cerebellar activation progressively declined. In another study the temporal patterns were learned prior to the brain imaging in order to identify parametric dependence of activity on complexity of the rhythm. Participants tapped six different temporal patterns that varied in complexity but were matched in overall frequency of responding. Activations in several regions, including bilateral SMA and PMA, and right dorsolateral prefrontal and motor cortices, were observed to be correlated with rhythm complexity in the paced phase (and in a preparatory phase before that), but no areas of correlated activation were observed in unpaced tapping. This difference may have reflected preparatory processes in retrieval and setting up of the correct time program in the initial phases of paced responding, which then ran with lower attention demands during unpaced tapping.

A rhythm conveys a sense of grouping among responses in an otherwise undifferentiated event series. Such grouping can be defined by patterning of the force used in responding (e.g., low and high) as well as by patterning of the time intervals (e.g., short and long). In a study evaluating the effect of alternation of force and time, participants produced squeeze responses at equal (600 ms) or alternating (400 and 800 ms) intervals with either equal (12 N) or alternating (8 and 16 N) force (see Figure 9).

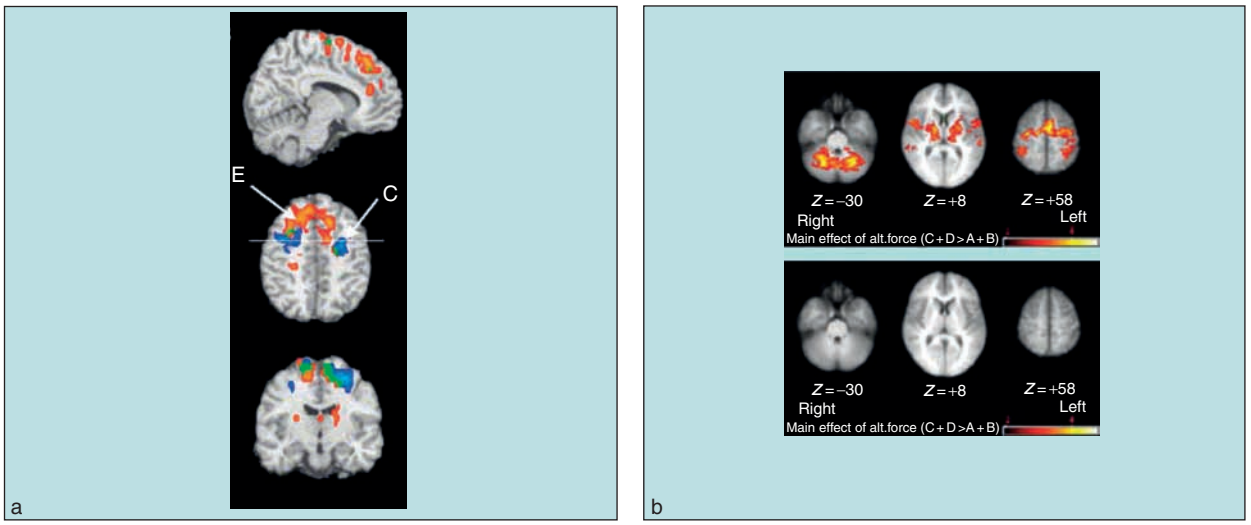

Figure 9 Functional magnetic resonance imaging study of alternation of force but not time, showing basal ganglia activation. From Pope.

Analysis of the unpaced phase compared to rest showed activation of M1, SMA, basal ganglia, thalamus, and cerebellum during the production of sequences marked by equal interval and force. Although there was no reliable main effect of alternating interval, these regions exhibited greater activity with the extra demands of responding with alternating force pulses. These data are important in identifying a significant role of the basal ganglia in the control of force and also point to the need to control force in activation studies when the focus is on basal ganglia function in timing (since patterning of force alone is sufficient to activate these structures). Interestingly, one region, right posterior parietal cortex, showed significant activation related to the effect of alternation of force and time, compared to alternation of either force or time.

As noted earlier, the motor timing literature is replete with studies of contrasting in-phase and antiphase bilateral movements, which generally reveal, at higher frequencies, greater variability with antiphase movements and a tendency to switch spontaneously from antiphase to in-phase movements. People can be trained to produce polyrhythms, in which there is a noninteger relationship between the two limbs (e.g., 2:3). In a study that compared in-phase, antiphase, and polyrhythm production during functional magnetic resonance imaging (fMRI), it was shown that the latter two coordination modes activate a much larger network than do in-phase movements. This network included SMA, pre-SMA, inferior parietal gyri, PMA, and the superior temporal gyri, whereas in-phase movement activated the cingulate motor area (CMA) and anterior cerebellum. In addition, polyrhythms activated the posterior cerebellum and part of the CMA. These imaging data are largely in agreement with those derived from lesion studies, in suggesting an extensive frontoparietotemporal network, with strong cerebellar involvement. The imaging studies tend not to emphasize basal ganglia activation, although many of the active cortical areas are strongly interconnected with the basal ganglia.

Magnetoencephalography (MEG) offers greater temporal resolution than does fMRI in assessing changes in cortical activity accompanying polyrhythm production. A recent study reported MEG signal changes suggestive of cross-hemispheric inhibition of coupling during polyrhythm production. A picture is emerging of bilateral coupling of independent timers, for the two sides of the body, which occurs mainly at a subcortical level. The tendency for the two sides to produce similar, in-phase actions most readily must be overcome by intercortical inhibition (mediated by the corpus callosum) in order to produce independent continuous movements of the two arms (for example, fast antiphase movements, or polyrhythms). Participants find making antiphase movements of intermediate difficulty, between the level of difficulty of making in-phase movements and polyrhythms. Patients with PD reliably show difficulty in antiphase movements. These results suggest that the basal ganglia might form part of the subcortical network that supports bimanual coupling. Many of the cortical areas that show activity in antiphase movements project to the basal ganglia. 'Descending' commands from these cortical areas might modulate basal ganglia activity to 'allow' decoupling of the limbs during antiphase and polyrhythm production.

Conclusion

Several themes have emerged in this review of motor timing studies. We have stressed the importance of models to help formalize our approach to motor timing. For example, a general framework in which to approach unpaced timing is the two-level model that allows the experimenter to distinguish variability due to central timer functions (which encompasses a number of processes, including motor programming) from that due to motor implementation variability. Working in this framework, lesions of the (lateral) cerebellum have been implicated in central timing difficulties, whereas motor implementation errors have been associated not only with medial cerebellar damage, but also with cortical and basal ganglia dysfunction. However, the allocation of specific functions to individual brain regions on the basis of human lesion studies remains contentious, and functional imaging studies strongly suggest that networks of cortical and subcortical structures are crucial in timing functions.

It is important to consider in greater detail the type of movement made. One can study paced movements, typically involving synchronization with an external cue (such as a metronome), or unpaced movements in which the participant attempts to continue the movement in the absence of an ongoing cue. An extension of this approach is to compare movements made in synchrony with an external pacer with movements made to syncopate with the beat. Another distinction is that between continuous and discontinuous movements. In each case wide but generally overlapping underlying networks have been demonstrated.

We have also seen evidence of separate timing systems for the two sides of the body. Synchronization of the two may depend on subcortical connections, particularly for discrete movements. Patients with diseases of the basal ganglia (PD being the most extensively studied) show deficits in interlimb coordination when required to make antiphase movements. This may reflect inflexibility in the coupling of the two sides. In contrast, patients with corpus callosal lesions show greater ease than do persons with healthy brains in making independent movements of the two limbs. The corpus callosum may be crucial for inhibiting these interactions to allow the generation of complex patterns such as fast antiphase movements and polyrhythms.

Progress in motor timing will come from a multifaceted approach. Behavioral studies in healthy volunteers and patients with well-defined neurological deficits should be informed by models, and should in turn inform these theoretical accounts. A range of different types of single and multilimb movements should be studied. These same tasks should be used during imaging (with both fMRI and MEG) and with carefully timed application of TMS to disrupt cortical function. In general, imaging studies show extended networks involved in a given task; it is, however, often difficult to determine the precise role or roles of an area by the traditional subtraction techniques (what is active in task B but not A?). Lesion studies can be complementary, but the same set of tasks needs to be used in a range of patient groups in order to make sense of these data.

See also: Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia: Motor Functions; Basal Ganglia: Physiological Circuits; Cerebellar Microcircuitry; Deep Brain Stimulation and Movement Disorder Treatment; Motor Sequences; Prefrontal Cortex; Premotor Areas: Medial; Procedural Learning: Cerebellum Models.

Further Reading

Haken H, Kelso JAS, and Bunz H (1985) A theoretical model of phase transitions in human hand movements. *Biological Cybernetics* 51: 347–356.

Haken HH, Peper CE, Beek PJ, et al. (1996) A model for phase transitions in human hand movements during multifrequency tapping. *Physica D* 90: 179–196.

Harrington DL, Lee RR, Boyd LA, et al. (2004) Does the representation of time depend on the cerebellum? Effect of cerebellar stroke. *Brain* 127: 561–574.

Ivry RB, Diedrichsen J, Spencer RCM, et al. (2004) A cognitive neuroscience perspective on bimanual coordination. In: Swinnen S and Duysens J (eds.) *Neuro-behavioural Determinants of Interlimb Coordination*, pp. 259–295. Boston, MA: Kluwer.

Ivry RB and Spencer RMC (2004) The neural representation of time. *Current Opinion in Neurobiology* 14: 225–232.

Lewis PA and Miall RC (2003) Distinct systems for automatic and cognitively controlled time measurement: Evidence from neuroimaging. *Current Opinion in Neurobiology* 13: 250–255.

Mauk MD and Buonomano DV (2004) The neural basis of temporal processing. *Annual Review of Neuroscience* 27: 307–340.

Pressing J (1999) The referential dynamics of cognition and action. *Psychological Review* 106: 714–747.

Repp BH (2005) Sensorimotor synchronisation: A review of the tapping literature. *Psychonomic Bulletin & Review* 12: 969–992.

Schirmer A (2004) Timing speech: A review of lesion and neruoimaging findings. *Cognitive Brain Research* 21: 269–287.

Swinnen SP and Wenderoth N (2004) Two hands, one brain: The neuroscience of bimanual skill. *Trends in Cognitive Sciences* 8: 18–25.

Vorberg D and Wing AM (1996) Modelling variability and dependence in timing. In: Heuer H and Keele S (eds.) *Handbook of Perception and Action*, vol. 2, pp. 181–262. New York: Academic Press.

Wing AM (2002) Voluntary timing and brain function; an information processing approach. *Brain and Cognition* 48: 7–30.

Wing AM and Beek PJ (2002) Movement timing – a tutorial. In: Prinz W and Hommel B (eds.) *Attention and Performance 19*, pp. 202–226. Oxford: Oxford University Press.

Wing AM and Kristofferson AB (1973) Response delays and the timing of discrete motor responses. *Perception and Psychophysics* 4: 5–12.

MPTP Parkinsonism Model

T Wichmann and M R DeLong, Emory University School of Medicine, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Recent progress in neuroscience research has led to major insights into the pathophysiologic basis of Parkinson's disease (PD). The availability of suitable animal models, in particular the primate treated with 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP), has been crucial in this progress. This article reviews some of the essential pathologic changes found in PD, gives an overview of the mechanism by which MPTP induces parkinsonism, and summarizes, from a systems perspective, the pathophysiologic concepts that have arisen from the studies carried out in MPTP-treated primates.

Pathologic Substrate in PD

Idiopathic PD is characterized by the cardinal signs of akinesia (impaired movement initiation and poverty of movement), bradykinesia (slowness of movement), muscular rigidity, and tremor at rest. The etiology of PD is most likely multifactorial, with both genetic and environmental factors playing a role. Multiple genetic forms of PD have now been identified, and epidemiologic studies have identified a number of environmental risk factors, such as rural living, consumption of well water, and head injury. The motor features of PD result from degeneration of dopaminergic neurons in the substantia nigra pars compacta (SNc), which project to the striatum and, to a lesser extent, to other basal ganglia nuclei such as the external and internal segments of the globus pallidus (GPe and GPi, respectively), the subthalamic nucleus (STN), and the substantia nigra pars reticulata (SNr). At the time of onset of clinical signs, dopamine loss affects particularly the sensorimotor portion of the striatum, the putamen, resulting in the appearance of motor disturbances early in the disease. In later stages, more-widespread dopamine loss, affecting other regions of the basal ganglia (such as the caudate nucleus and the pallidum) and extrabasal ganglia areas, such as cortex, hypothalamus, and thalamus, may contribute to the clinical manifestations of the disease.

Although dopamine loss is the defining pathologic feature of PD, it is now clear that other brain systems are also affected, even early in the disease. Indeed, recent studies indicate that the earliest pathologic changes occur in the mesenteric plexus of the gut and in the olfactory bulb. The disease appears then to involve brain stem neurons in an ascending fashion, involving the dorsal motor nucleus of the vagus, the nucleus gigantocellularis, the locus coeruleus, and the raphe nuclei before reaching the SNc. Subsequent steps in the progression of the disease involve the amygdala, thalamus, and cerebral cortex. Damage to these extranigral areas may significantly contribute to the nondopaminergic manifestations of the disease, in particular the accompanying anxiety and depression, sleep disturbances, dementia, and autonomic disturbances resulting in orthostatic hypotension, constipation, and urinary problems.

Animal Models of PD

None of the available animal models for parkinsonism faithfully reproduce the progressive pathology and clinical phenotype of human PD. Instead, animal models for PD can be grouped into genetic models and toxin-based models. Genetic models, such as α-synuclein overexpression mice, model the process of neurodegeneration, often without resulting in a convincing clinical phenotype. By contrast, toxin or drug models mimic the consequences of dopamine depletion but generally employ toxins unlikely to be etiologically relevant for human so-called sporadic PD. The MPTP model falls into the latter category of models and is described in detail below. It is one of a family of compounds that can be employed to disturb dopaminergic functions in the brain. Other members of this group are selective dopamine receptor blockers and nonselective catecholamine-depleting agents such as reserpine or tetrabenazine. These agents are used primarily for acute *in vitro* experiments or for studies requiring a reversible dopaminergic deficit. Models produced by toxins that permanently destroy dopaminergic cells more faithfully reproduce the pathology of the human disorder and are often used for *in vivo* studies. In rodents, injections of 6-hydroxydopamine into the SNc or treatment with mitochondrial complex I inhibitors such as rotenone or MPTP can produce oxidative damage to dopaminergic neurons.

Historical Aspects and Clinical Findings

MPTP was first synthesized and described in 1947 by Hoffman LaRoche chemist Albert Ziering as a by-product of the synthesis of the opioid analgesic meperidine

analog 1-methyl 4-phenyl 4-propionoxypiperidine. Starting in the 1950s, MPTP was evaluated in pharmaceutical trials in animals and humans for possible anticholinergic, antihypertensive, and even antiparkinsonian effects. The realization that exposure to MPTP can result in almost instantaneous parkinsonism came after drug addicts became parkinsonian in the early 1980s after having injected themselves with a meperidine analog which was contaminated by MPTP. Acutely after MPTP injections, such individuals experienced euphoria, sometimes associated with disorientation, transient visual illusions, or hallucinations followed by intermittent appendicular jerking movements. Over the next few weeks, this reaction was followed by the gradual development of more-typical parkinsonian features, including slowness of movement, dysarthria, dysphagia, and tremor (which may be asymmetric). In the ensuing chronic parkinsonian phase, patients showed all the cardinal motor features of idiopathic PD, as well as freezing, paradoxical kinesia, dystonic posturing of trunk or limbs, and flexed posture. MPTP-induced parkinsonism in humans is a chronic but nonprogressive condition. The symptoms of the disease respond to commonly used antiparkinsonian medications, such as levodopa and dopamine receptor agonists. It is interesting that patients exposed to MPTP also develop cognitive abnormalities comparable to those seen in moderately advanced idiopathic PD, but they do not develop some of the other nonmotor abnormalities seen in PD (such as olfactory dysfunction).

In 1983, it was first reported that MPTP treatment in primates induces movement abnormalities similar to those seen in PD. The MPTP-treated monkey has since been used in numerous investigations into the pathophysiology and pharmacology of PD. In addition, MPTP was shown to produce dopamine depletion in mice. The phenotype of the parkinsonian syndrome resulting from dopamine loss in these animals is not as convincing as the MPTP-induced phenotype in primates. However, the use of the toxin in mice has greatly helped clarify its mechanism of action, and MPTP is often used for pharmacologic and molecular or genetic experiments.

MPTP: Mechanism of Action

After systemic administration, MPTP rapidly crosses the blood–brain barrier and is taken up by glia and serotonergic cells. The compound is metabolized by monoamine oxidase B into an unstable intermediate product (MPDP), which undergoes spontaneous conversion to a more stable metabolite, 1-methyl-4-phenylpyridium, MPP^+ (see Figure 1). MPP^+ is released into the extracellular space and taken up via the plasma membrane dopamine transporter into dopaminergic cells. Following the transmembrane mitochondrial potential gradient, MPP^+ then accumulates within mitochondria. It has long been known that the toxin is an inhibitor of the mitochondrial complex I and inhibits adenosine triphosphate production. Failure of energy production is the direct cause of the dying back of nigrostriatal axon terminals and the eventual death of some dopaminergic neurons. In addition, mitochondrial inhibition also results in overproduction of reactive oxygen species, which in turn triggers apoptotic signaling and cell death.

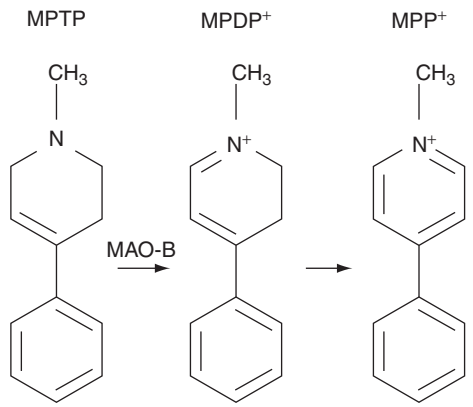

Figure 1 Metabolism of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP). MPTP is a protoxin which is converted to 1-methyl-4-phenylpyridium (MPP^+), the actual toxin in glia cells. MPP^+ is a substrate for the dopamine transporter, which is found in dopaminergic neurons. Once taken up by the dopaminergic cell, MPP^+ acts to inhibit the mitochondrial complex I and may also enhance apoptotic cell death via an increase in oxidative stress. MAO-B, monoamine oxidase B; MPDP, 1-methyl-4-phenyl-2,3-dihydropyridinium.

Pathologic Changes Seen in MPTP-Induced Parkinsonism

MPTP administration in primates results in pathologic and biochemical changes in the basal ganglia that are similar to those seen in human idiopathic PD. Immunohistochemical and other studies have revealed a marked cell loss in the SNc, particularly affecting the ventrolateral portion of the SNc, similar to pathology found in humans. Additional, although less pronounced, damage has been found in the ventral tegmental area, locus coeruleus, raphe nuclei, hypothalamus, retina, and cerebral cortex. The loss of dopaminergic cells in the SNc has been demonstrated to be associated with iron accumulation and an increased astrocytic reaction. The most prominent biochemical changes are loss of dopamine in putamen and caudate nucleus (often reaching more than 90% in symptomatic animals), associated with an increase in the turnover rate of dopamine. The pattern of

striatal dopamine loss is similar to that seen in PD. However, while dopamine loss in the putamen is greater than that in the caudate nucleus in MPTP-treated primates, the difference is not as striking between these regions as it is in human PD. Other neurotransmitter systems, such as the norepinephrine system, and serotonin systems may also be affected.

Two primary areas have been identified in which the pathology of MPTP-induced parkinsonism in monkeys differs from that of idiopathic PD in humans. One is the fact that MPTP treatment results in an acute parkinsonian syndrome rather than a slowly progressive disease, although attempts have been made to generate a chronic model through the use of small amounts of MPTP administered daily over months. The second difference is that MPTP-treated monkeys lack clearly defined intracytoplasmic eosinophilic inclusion bodies, the so-called Lewy bodies. Lewy bodies are one of the pathologic hallmarks of idiopathic PD in humans and contain aggregates of α-synuclein, as well as other proteins. Lewy bodies are found in many brain regions, most notably the SNc and cortex. Levels of α-synuclein in the SNc are known to be upregulated in MPTP-treated monkeys, and α-synuclein aggregates may occur. Eosinophilic inclusion bodies have also been described, particularly in aged monkeys treated with MPTP. However, the inclusion bodies do not resemble the ultrastructure and composition of Lewy bodies as they are found in parkinsonian patients. In contrast to idiopathic PD, MPTP does not cause the formation of cortical Lewy bodies.

In nonprimate species, MPTP has somewhat different effects. While causing little pathologic or biochemical damage in rats, it induces prominent dopamine depletion in mice, albeit at much higher doses than those required in monkeys. Acutely, the toxin has been shown to lead to vacuolation and nuclear condensation of dopaminergic cells in the SNc, associated with perivascular edema and astrocytic and microglial activation. As in monkeys, MPTP administration eventually results in degeneration of DA neurons in the SNc, VTA, and retrorubral field, as well as lesser degeneration in regions outside the basal ganglia, such as the midline and intralaminar thalamic nuclei.

The relationship between the MPTP-induced changes and the expression of α-synuclein has been closely scrutinized in these animals. MPTP treatment results in an increase in the number of neurons expressing α-synuclein in the SNc. In some studies, animals overexpressing α-synuclein have been more susceptible to MPTP, while α-synuclein knockout animals are partly protected from the effects of MPTP. Some authors have reported cytoplasmic inclusion bodies in surviving dopaminergic neurons in MPTP-treated mice, particularly with chronic MPTP exposure, although they lack the typical Lewy body morphology.

Basal Ganglia Circuits Affected by Treatment with MPTP

The use of MPTP in animals has led to important discoveries regarding the pathophysiology and potential environmental mechanisms involved in the degeneration of dopaminergic cells in PD. All the research that has been done with this toxin is beyond the scope of this article; it focuses instead on a description of the use of the MPTP model of PD in one specific area of research, that is, research into the pathophysiology of parkinsonism from a 'systems' perspective. This field has been revolutionized by the availability of the MPTP model, and the knowledge gained from experiments in MPTP-treated primates has been translated into new therapies of this disorder.

The loss of dopamine impacts basal ganglia function. The basal ganglia are a group of functionally related subcortical nuclei that include the neostriatum (comprised of the caudate nucleus and the putamen), ventral striatum, GPe, STN, GPi, SNr, and SNc. They are anatomically related to large portions of the cerebral cortex, thalamus, and brain stem. The cortex–basal ganglia–thalamocortex circuits appear to be organized in an orderly arrangement of segregated reentrant circuits. Of these circuits, the 'motor circuit' has been particularly studied because of its likely involvement in the pathophysiology of movement disorders. In primates, it involves projections from the somatosensory, motor, and premotor cortices, which terminate in the postcommissural putamen. Cortical inputs also terminate in the STN. A second major group of inputs to striatum and STN arises from the intralaminar thalamic nuclei, the centromedian, and parafascicular nucleus. In primates, the centromedian projects to the motor portions of putamen and STN. Brain stem projections to the STN, GPi, and SNr also arise from the pedunculopontine nucleus (PPN), which receives major output from the GPi/SNr.

Cortical information is conveyed from the striatum to the output nuclei of the basal ganglia (GPi and SNr) via the the so-called direct and indirect pathways. The direct pathway arises from striatal neurons that project monosynaptically to neurons in GPi and SNr, whereas the indirect pathway arises from a separate set of neurons that projects to GPe. GPe conveys the information it receives either directly to GPi/SNr, or via the STN. Anatomically, some striatofugal neurons appear to collateralize more extensively,

reaching GPe, GPi, and SNr. Dopamine modulates the activity of the basal ganglia output neurons in GPi and SNr by facilitation of transmission over the direct pathway and inhibition of transmission over the indirect pathway. The net effect of striatal dopamine release appears to be to reduce basal ganglia output, whereas a reduction of dopamine release, as occurs in PD, results in a net increase in basal ganglia output.

Basal ganglia output arises from both GPi and SNr. The motor territory of GPi projects almost exclusively to the posterior part of the ventrolateral nucleus, which in turn sends projections toward the supplementary motor area, the primary motor cortex, and premotor cortical areas. GPi and SNr also project to the noncholinergic neurons of the PPN and to the intralaminar thalamic nuclei (centromedian and parafascicular nuclei).

Movement-Related Basal Ganglia Activity in Normal Animals

Since the majority of neurons in GPi exhibit a spontaneous high-frequency (inhibitory) discharge and increase their firing rate with movement, it has been speculated that a major role of the basal ganglia is to inhibit the activity of the thalamocortical network. Voluntary movements are initiated at the cortical level of the motor circuit with output to brain stem and spinal cord and to multiple subcortical targets, including the thalamus, putamen, and STN. Cortical activation of putaminous direct pathway neurons leads to a reduction of the inhibitory basal ganglia output and subsequent disinhibition of related thalamocortical neurons and facilitation of the movement. In contrast, activation of striatal neurons that give rise to the indirect pathway will lead to increased basal ganglia output and a suppression of movement. The basal ganglia may play a role in the control of specific kinematic parameters, such as amplitude, velocity, or direction, allowing intended movements to proceed and suppressing unintended movement. They may also be involved in self-initiated (internally generated) movements, in motor (procedural) learning, and in movement sequencing.

The functions of the nonmotor portions of the basal ganglia may be analogous to the motor functions, that is, participating in the respective roles of the anatomically related cortical areas, but they are less explored.

Parkinsonism

Early studies using the MPTP model of parkinsonism suggested that metabolic activity (as measured with the 2-deoxyglucose technique) is increased in both pallidal segments but decreased in the STN. This was interpreted as evidence for increased activity of the striatum–GPe connection and the STN–GPi pathway or, alternatively, as evidence for increased activity via the projections from the STN to both pallidal segments. It was then shown with microelectrode recordings of neuronal activity, that MPTP-induced parkinsonism in primates is associated with reduced tonic neuronal discharge in GPe and increased discharge in STN, GPi, and SNr, compared with normal controls. In parkinsonian patients undergoing pallidotomy, it has also been shown that the average discharge rate in GPe is significantly lower than that in GPi.

The changes in discharge rates in the basal ganglia nuclei have been interpreted as indicating that striatal dopamine depletion leads to increased activity of striatal neurons of the indirect pathway, resulting in inhibition of GPe and subsequent disinhibition of STN and GPi/SNr. In addition, loss of dopamine in the striatum may lead to reduced activity via the direct pathway. Increased basal ganglia output to the thalamus and increased inhibition of thalamocortical neurons have been corroborated by 2-deoxyglucose studies in which increased (synaptic) activity in the ventral anterior and ventrolateral nucleus of thalamus was demonstrated. Positron emission tomography studies have shown that the activation of motor and premotor areas in parkinsonian patients is reduced although no changes have been seen in the thalamus. Alterations of cortical activity in motor cortex and supplementary motor areas have also been demonstrated with single-cell recording in hemiparkinsonian primates.

Increased inhibition of thalamocortical neurons by excessive output from GPi/SNr may reduce the responsiveness of cortical mechanisms involved in motor control or in the initiation of movement or may interfere with 'set' functions that have been shown to be highly dependent on the integrity of basal ganglia pathways. Because parkinsonism is a circuit disorder, surgical or pharmacologic interventions at a variety of targets within the network could be successful. This can indeed be appreciated when considering the results of lesion studies in parkinsonian primates. One of the most important and dramatic in this regard was the demonstration that lesions of the STN in MPTP-treated primates reverse all the cardinal signs of parkinsonism, presumably by reducing GPi activity. Similarly, GPi and SNr inactivation have been shown to be effective against at least some parkinsonian signs in MPTP-treated primates. These results have rekindled interest in functional neurosurgical approaches to the treatment of medically intractable PD. This was first employed in the form of GPi lesions (pallidotomy) and, more recently, with STN lesions. In addition, high-frequency deep brain

stimulation (DBS) of both the STN and the GPi have been shown to reverse parkinsonian signs, probably by multiple modes of action, for instance by inhibition of STN neurons through 'depolarization block' or activation of inhibitory afferents or by true activation of STN efferents to the pallidum.

Although the simple rate-based circuit models were consistent with many aspects of PD, several findings in patients are not compatible with the rate-based model. For instance, the model would predict that thalamic lesions would worsen parkinsonism, but such lesions are, in fact, beneficial in the treatment of tremor and rigidity. Similarly, the fact that lesions of GPi are effective against drug-induced dyskinesias contradicts the rate-based model. Dyskinesias were initially thought to arise from pathologic reduction in basal ganglia output and thus should not respond positively to further reduction of pallidal outflow. These and other findings suggest that factors other than changes in the discharge rates are important in the development of PD. It has now been extensively documented that not only rates but also discharge patterns and the level of synchronization between neighboring neurons undergo substantial changes in MPTP-treated monkeys and patients. For instance, the proportion of cells in STN, GPi, and SNr with discharge in oscillatory or nonoscillatory bursts is greatly increased in the parkinsonian state. There is also a marked increase in the synchronization of discharge between neurons in the basal ganglia, and the intensity and prevalence of neuronal responses to passive limb manipulations in STN, GPi, and thalamus increase after treatment with MPTP.

At least some of these abnormal discharge patterns may be caused by intrinsic membrane properties of neurons in the dopamine-depleted basal ganglia. Local circuit interactions may also contribute to bursts and oscillations in the basal ganglia. Oscillations and other firing abnormalities throughout the entire basal ganglia–thalamocortical network may be tightly related to each other so that in most cases, no single generator of such abnormalities may be identifiable.

Conclusions

The discovery that a relatively selective dopaminergic neurotoxin, MPTP, produces a typical parkinsonian syndrome in humans and monkeys is one of the major breakthroughs in neuroscientific research. The use of this toxin in monkeys has resulted in the formulation of a highly valuable model of the disease. More importantly, the work in MPTP-treated primates and other species has contributed to the development of new insights into the pathophysiologic basis of PD and to the application of surgical interventions that can be used to help patients with PD and has provided researchers with a tool for examining the properties of new antiparkinsonian drug treatments.

See also: Animal Models of Parkinson's Disease; Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia: Motor Functions; Basal Ganglia: Acetylcholine Interactions and Behavior; Basal Ganglia: Internal Organization; Cognition: Basal Ganglia Role; Dopamine in Perspective; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopamine Neurons: Reward and Uncertainty; Neurogenesis in the Intact Adult Brain; Parkinsonian Syndromes; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Bove J, Prou D, Perier C, and Przedborski S (2005) Toxin-induced models of Parkinson's disease. *NeuroRx* 2: 484–494.

Langston JW and Ballard P (1983) Chronic parkinsonism in humans due to a product of meperidine-analog synthesis. *Science* 219: 979–980.

Langston JW and Palfreman J (1996) *The Case of the Frozen Addicts: How the Solution of an Extraordinary Medical Mystery Spawned a Revolution in the Understanding and Treatment of Parkinson's Disease*. New York: Pantheon.

Vila M and Przedborski S (2003) Targeting programmed cell death in neurodegenerative diseases. *Nature Reviews Neuroscience* 4: 365–375.

Wichmann T and DeLong MR (2003) Functional neuroanatomy of the basal ganglia in Parkinson's disease. *Advances in Neurology* 91: 9–18.

Müller Cells

A Bringmann and A Reichenbach, University of Leipzig, Leipzig, Germany

© 2009 Elsevier Ltd. All rights reserved.

The retina of all vertebrates contains specialized radial glial cells even in the adult stage. They were discovered in 1851 by Heinrich Müller and were later termed radial fibers of Müller or Müller cells. It is interesting to note that Müller observed the cells, together with some adhering photoreceptor cells, in dissociated retina preparations (Figure 1(d)). In many lower vertebrates, Müller cells are the only retinal glia; in mammals, they constitute the vast majority of glial cells together with the minorities of astrocytes and microglial cells, which latter are both confined to the innermost retinal layers (nerve fiber, ganglion cell, and inner plexiform layers).

Basic Properties of Müller Cells

Morphology

Müller cells display a characteristic bipolar morphology. Their soma is located in the inner nuclear layer; it gives rise to two main stem processes which pass through the tissue in opposite directions, together spanning the entire thickness of the retinal tissue (Figures 1–3). The inner stem process terminates at the inner (i.e., vitread) retinal surface with an endfoot which contributes, together with a basal lamina, to the inner limiting membrane. The inner stem process contains bundles of intermediate filaments, consisting mainly of vimentin in the healthy retina but also of increasing amounts of glial fibrillary acidic protein (GFAP) in virtually all forms of retinal injury. The endfoot is densely packed with smooth endoplasmic reticulum (Figure 2); its vitread surface membrane contains many orthogonal arrays of membrane particles, most likely representing ion and water channels, and other transport proteins. Similar *en-passant* endfeet are formed by Müller cell processes where they contact retinal blood vessels in species with vascularized retinae (i.e., in most mammals including humans). The outer stem process, arising from the opposite pole of the soma, contains microtubules and ends up at the outer limiting membrane, formed by zonulae adherentes between these Müller cell processes and the inner segments of the photoreceptors. From there, this process extends many microvilli into the subretinal space. Within both synaptic (plexiform) layers, many side branches originate from the two stem processes which form complex, filigrane sheaths around the synapses. Furthermore, thin velate sheaths arise from the Müller cell in all nuclear layers, enveloping the neuronal somata. This is particularly evident in the outer nuclear layer where the outer Müller cell process forms honeycomb-like structures (Figures 1–3).

Whereas this general description applies to Müller cells in all vertebrate retinae, there is considerable adaptive variation in cell morphology (Figure 3). For instance, in the Müller cells of reptiles and birds the inner stem process is usually split into sheaves of thin processes, each with its own small vitread endfoot. Moreover, in the thick central retina (where the retinal cells are densely packed) the Müller cells are longer and thinner than in the retinal periphery. In the retinae of most lower vertebrates and of mammals with avascular retinae (such as rabbits and guinea pigs), the Müller cells have only few mitochondria, located at the sclerad end of the outer stem process (i.e., close to the choroid as the only source of oxygen; cf. Figure 2), whereas in vascularized retinae mitochondria are distributed along the entire length of both Müller cell processes.

As a rule, Müller cells occupy approximately 10% of the total retinal volume (up to >70% in the nerve fiber layer of the retinal periphery where the adjacent Müller cell endfeet form a continuous cobblestone pattern and only a few nerve fiber bundles pass through). The surface-to-volume ratio of the Müller cell processes is high, particularly within the two plexiform layers (14–18 μm^{-1}), in the outer nuclear layer (approximately 20 μm^{-1}), and at the subretinal microvilli (up to 50 μm^{-1}). Thus, virtually all neuronal elements are contacted by Müller cell processes separated only by narrow extracellular clefts. As a whole, the Müller cell population forms a regular pattern (Figure 1(c), middle and bottom rows) with strikingly similar cell densities among the various vertebrate retinae (some 5000 Müller cells mm^{-2} in the retinal periphery, up to 15 000 mm^{-2} in the retinal center; only in the perifoveal retina of primates are higher densities of more than 25 000 mm^{-2} found). Each of these cells can be considered the core of a columnar micro-unit of retinal neurons, in mammals constituted by approximately one cone photoreceptor, three to six interneurons of the inner nuclear layer, less than one retinal ganglion cell, and a varying number of rod photoreceptor cells (approximately two in diurnal squirrels; approximately ten in average mammals such as human, rabbit, horse, and others; and slightly more than 20 in nocturnal rodents and carnivores).

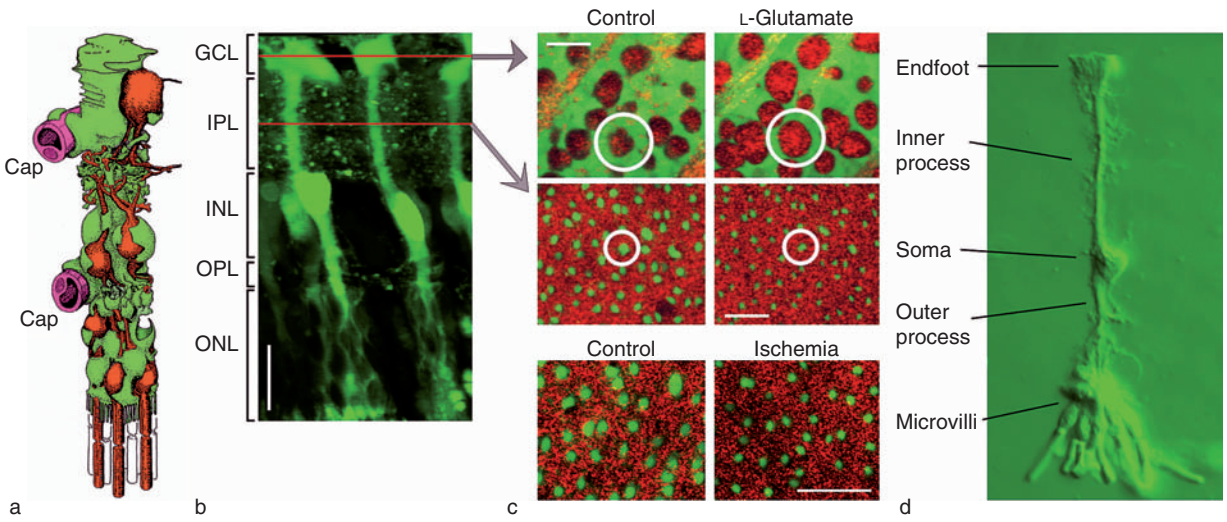

Figure 1 Müller cell morphology and its activity-dependent changes: (a) semi-schematic drawing of a Müller cell (green) together with the retinal neurons ensheathed by it (red) and retinal capillaries (violet) contacted by en-passant-endfeet; (b), Müller cells (in vital retinal slice) selectively stained by the vital dye, Mitotracker Orange (green); (c) confocal views on to the GCL (top) and the IPL (middle and bottom) of retinal wholemounts, revealing cross sections through the Müller cell endfeet and inner processes (green) and neuronal structures (red; reflexion mode of the microscope); (d) enzymatically dissociated vital Müller cell from a guinea pig retina, together with some adhering photoreceptor cells (bottom); Nomarski optics. In (c), the acute application of the excitatory neurotransmitter, L-glutamate (1 mM for 10 min) causes a swelling of the ganglion cell bodies in the GCL and of the synapses in the IPL. The swelling of neuronal elements is partly compensated by a decrease in the thickness of the Müller cell processes (which is also observed after retinal ischemia; bottom). These dynamic changes in glial cell shape prevent dramatic changes in the volume of the extracellular spaces. In (d), adhering photoreceptor cells can be used for electrophysiological (patch-clamp) and fluorimetric studies. GCL, ganglion cell layer; INL, inner nuclear layer; IPL, inner plexiform layer; ONL, outer nuclear layer; OPL, outer plexiform layer. Scale bar = 20 μm. (a) Adapted from Newman EA and Reichenbach A (1996) The Müller cell: A functional element of the retina. *Trends in Neurosciences* 19: 307–312. (b–c) Adapted from Uckermann O, Vargova L, Ulbricht E, et al. (2004) Glutamate-evoked alterations of glial and neuronal cell morphology in the guinea-pig retina. *Journal of Neuroscience* 24: 10149–10158.

Functions

Müller cells constitute an anatomical and functional link between the retinal neurons (of their column) and the compartments with which these need to exchange molecules (i.e., the retinal blood vessels, the vitreous body, and the subretinal space). Müller cells thus play a wealth of roles in supporting neuronal information processing that is carried out by astrocytes, oligodendrocytes, and ependymal cells in other regions of the central nervous system (Figure 4). Müller cells are essential in the ontogenetic creation and the adult maintenance of the neuroretinal architecture; they also (1) support neuronal survival, (2) provide retinal neurons with nutrients such as lactate and pyruvate for their oxidative metabolism, and (3) remove the metabolic waste of the neurons. In this context, Müller cells maintain the homeostasis of the retinal extracellular milieu, including ion, water, and pH homeostasis. Moreover, Müller cells regulate the retinal blood flow, contribute to the formation and maintenance of the blood–retina barrier, and act as modulators of immune and inflammatory responses and as phagocytes of fragments of degenerating retinal cells and foreign substances. They contribute to neuronal signaling processes, particularly by fast uptake and recycling of neurotransmitters, by providing precursors of neurotransmitters to neurons, and by the release of factors (e.g., D-serine and glutamate) which control the excitability of neurons. Moreover, Müller cells are probably involved in the recycling of photopigments. Any impairment of the supportive functions of Müller cells must therefore cause or aggravate retinal dysfunction and loss of retinal neurons by increasing the susceptibility of neurons to stressful stimuli.

Retinal Development

In the vertebrate retina, there are no glia- or neuron-specific progenitor cells; even the final division of a late progenitor cell typically generates one rod photoreceptor or bipolar cell and one Müller cell. Before this last division, each of the late progenitor cells undergoes a species-specific varying number of cell divisions, eventually generating a columnar aligned group of neurons together with one Müller cell in its core. Thus, most of the cells in a given columnar micro-unit which is constituted by one Müller cell and surrounding neurons (Figure 1(a)) are clonally as well as functionally related.

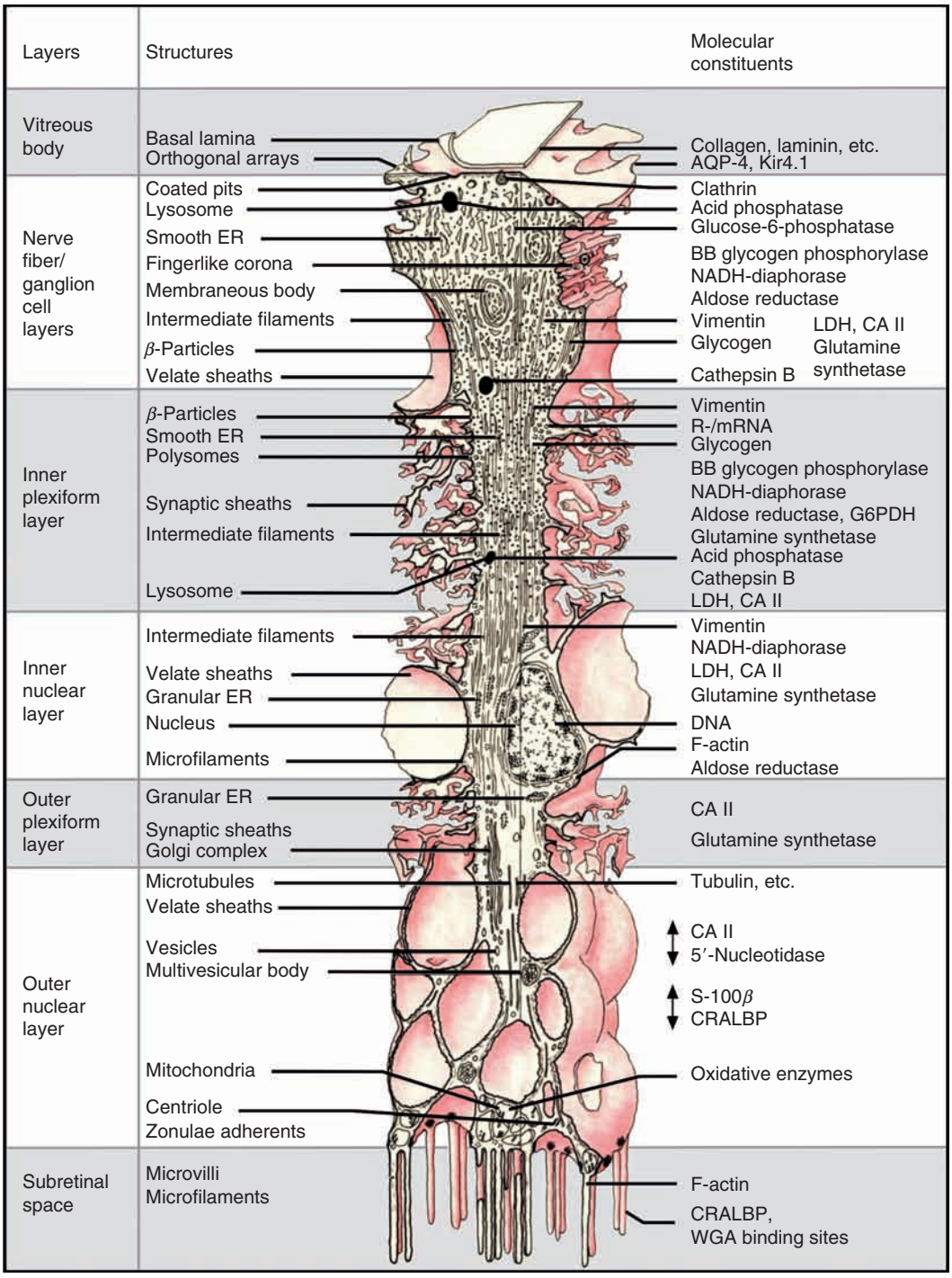

Figure 2 Cytotopographic specializations of a rabbit Müller cell (semi-schematic drawing). AQP-4, aquaporin-4 water channels; BB, brain-specific isozyme; CA II, carbonic anhydrase II; CRALBP, cellular retinoic acid binding protein; ER, endoplasmic reticulum; G6PDH, glucose-6-phosphate dehydrogenase; Kir4.1, inwardly rectifying potassium channels; LDH, lactate dehydrogenase; NADH, nicotinamide adenine dinucleotide (reduced form); S-100β, calcium-binding protein; WGA, wheat germ agglutinin. Adapted from Reichenbach A (1989) Attempt to classify glial cells by means of their process specialization using the rabbit retinal Müller cell as an example of cytotopographic specialization of glial cells. *Glia* 2: 250–259.

To take on the multitude of supportive functions for retinal neurons, the young Müller cell must undergo a radical differentiation. In rodents, this differentiation occurs during the first 2–3 postnatal weeks (Figure 5). This period includes the cessation of cell proliferation (during the first postnatal days) and the establishment of the mature synaptic wiring of retinal neurons (end of the first week) as well as the

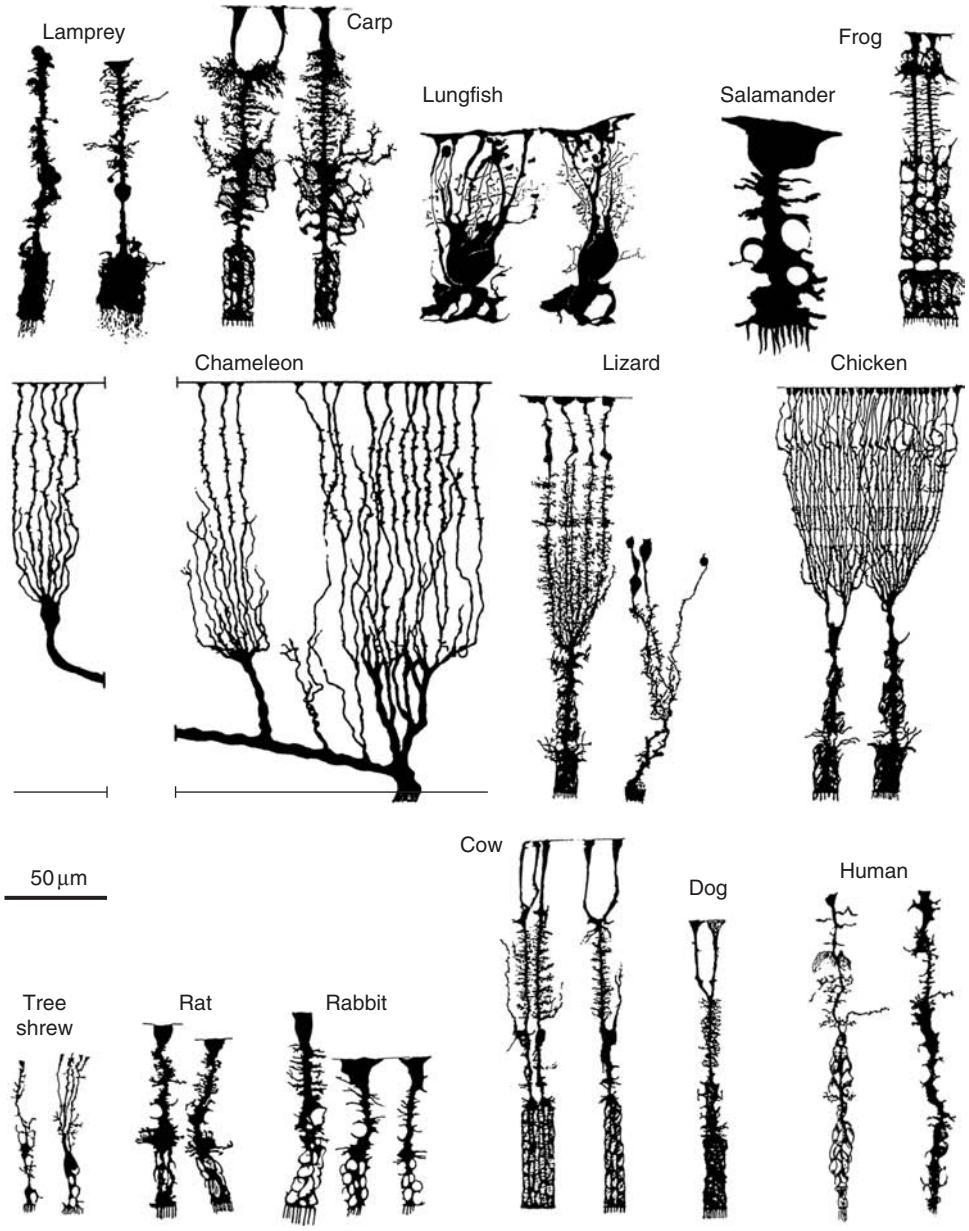

Figure 3 Adaptations of Müller cell morphology in various vertebrates. Most cells were drawn from Golgi-stained preparations, some are camera-lucida drawings of dye-filled cells. As far as possible, all cells are shown at the same magnification. Adapted from Reichenbach A and Wolburg H (2004) Astrocytes and ependymal glia. In: Kettenmann H and Ransom BR (eds.) *Neuroglial*, 2nd edn., pp.19–35. Oxford: Oxford University Press.

full development of photoreceptor outer segments (the second week). Accordingly, the retinal glia differentiates from a progenitor-like state, via a stage of immature radial glia, to mature Müller cells. This differentiation involves morphological, biochemical, and (electro-)physiological changes; the cells extend side branches between the neuronal somata and synaptic processes, accumulate glia-specific enzymes such as glutamine synthetase, and dramatically change the expression pattern of membrane proteins such as ion channels. In differentiating Müller cells, a dramatic increase of the transmembranous K^+ conductance occurs (Figure 5), accompanied by a hyperpolarization of the plasma membrane toward the adult levels of approximately −80 mV. This is apparently due to the insertion of K^+ channels into the Müller cell membrane during the second week of life and provides the essential precondition for most of the neuron-supportive functions of mature Müller cells. For example, in neurotransmitter recycling,

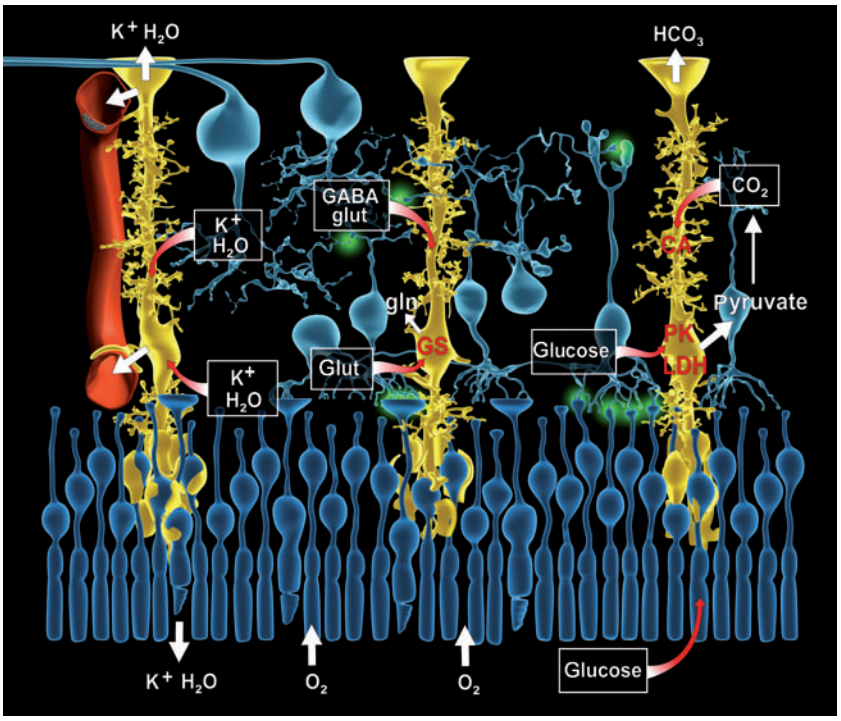

Figure 4 Summary of Müller cell functions, exemplified in a semi-schematic drawing of the mammalian retina. Spatial buffering of K^+ ions and water from extracellular clefts around active neurons into the vitreous body, the blood vessels, and the subretinal space (left). Transmitter recycling; the amino acid transmitters, GABA and glutamate, are taken up by the Müller cells and are converted into the nonactive metabolite glutamine, which is then released for neuronal resynthesis of the transmitters (middle). Metabolic symbiosis; the glycolytic metabolism of Müller cells generates lactate/pyruvate, which is delivered to the neurons as a substrate of their Krebs cycle (right). The neurons then release CO_2, which is either used by the Müller cells for their lipid synthesis or disposed of via the glial enzyme carbonic anhydrase. CA, carbonic anhydrase; GABA, γ-aminobutyric acid; gln, glutamine; glut, glutamate; GS, glutamine synthetase; LDH, lactate dehydrogenase; PK, pyruvate kinase. Adapted from Bringmann A, Pannicke T, Grosche J, et al. (2006) Müller cells in the healthy and diseased retina. *Progress in Retinal and Eye Research* 25: 397–424.

after the neural synapses are established at the early second week of life, the Müller cells (1) send side branches toward the synapses and ensheath them, (2) express glutamate uptake transporter molecules in the membrane of these perisynaptic sheaths, (3) hyperpolarize their resting membrane potential to provide the necessary driving force for the transporters, and (4) express the enzyme glutamine synthetase to convert the internalized glutamate molecules into glutamine. In parallel, the young Müller cells are involved in the control of retinal vascularization including the induction of the blood–retina barrier.

Postmitotic Müller cells retain a remarkable plasticity even in adult retinas; they are capable of dedifferentiation and proliferation. As important steps of this dedifferentiation, the cells reduce the K^+ conductance of their membrane (Figure 5) and increase their intracellular Ca^{2+} signaling (Figure 5(a)). The reduction of the K^+ conductance must cause a severe loss of functions involved in normal neuron–glia interaction, most of which require a hyperpolarized membrane potential. Moreover, a subpopulation of Müller cells expresses markers for neural progenitor (stem) cells even in the adult tissue (Müller stem cells). Transient Müller cell proliferation induced by *N*-methyl-D-aspartate (NMDA) excitotoxicity in postnatal chicken retinas causes a dedifferentiation of Müller cells into retinal progenitor-like cells. These cells may remain undifferentiated for longer time periods, may generate new Müller cells, or may transdifferentiate into retinal neurons. Similarly, Müller cells may be stimulated to transdifferentiate into neurons and photoreceptor cells in the adult rat retina after toxic injury produced with NMDA. These observations suggest that Müller cells are a potential source of neural regeneration in the postnatal retina.

Metabolic Symbiosis

As in many types of glial cells, the energy metabolism of Müller cells relies mainly on anaerobic glycolysis, even in the presence of sufficient oxygen supply. Therefore, Müller cells display a low rate of oxygen consumption and may withstand even long-lasting

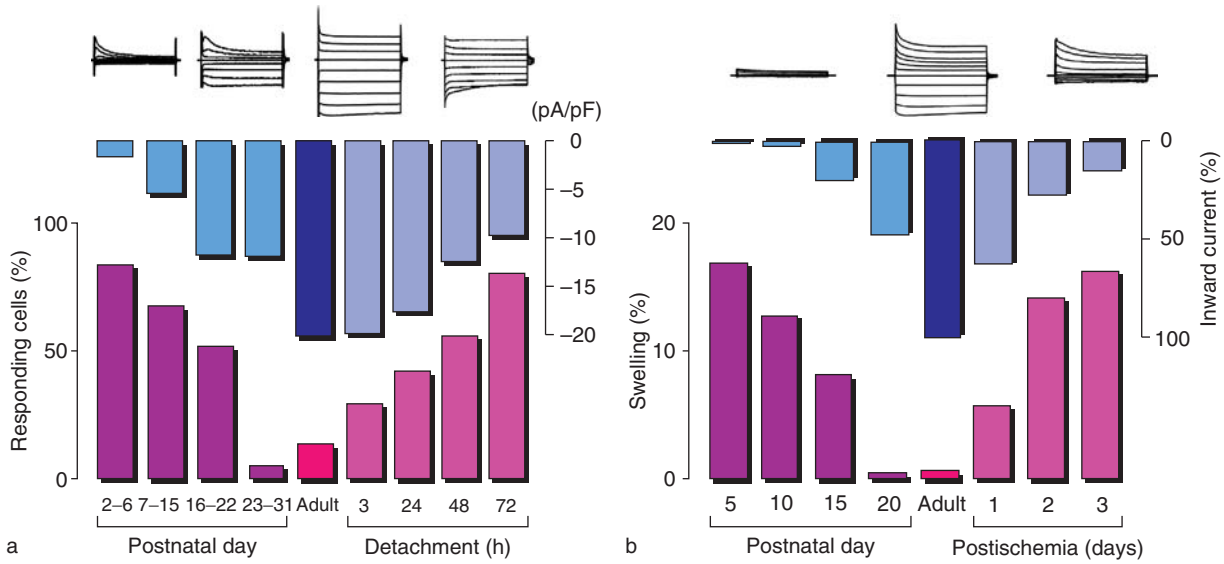

Figure 5 Differentiation and dedifferentiation of Müller cells: (a) in rabbit retina; (b) in rat retina. During the early postnatal differentiation of Müller cells from immature radial glial/progenitor cells, there occurs an increased expression of inwardly rectifying K^+ channels (mainly, Kir4.1), which results in increased inward current densities (top row, original traces; and blue bars) as well as in more negative membrane potentials (not shown). In parallel, the percentage of Müller cells which respond to adenosine 5′-triphosphate (ATP) application with an intracellular Ca^{2+} rise (dark violet/red bars in (a)) decreases, and the Müller cell swelling in response to hypotonic stress gradually disappears (dark violet/red bars in (b)). In many cases of retinal injury, such as retinal detachment (a) and ischemia/reperfusion (b), a de-differentiation of the cells occurs. The density of K^+ inward currents decreases dramatically (top row, original traces; and blue bars), accompanied by a depolarization of the mean membrane potentials (not shown), and the majority of Müller cells again display intracellular Ca^{2+} rises upon ATP-stimulation (bright-violet bars in (a)) and swell on hypotonic challenge (bright-violet columns in (b)).

anoxia. As long as oxygen is available, Müller cells are also resistant to the absence of glucose because other substrates such as lactate, pyruvate, glutamate, or glutamine may be used as energy substrates for their (although normally nondominant) Krebs cycle. Short periods of glucose deficiency and even ischemia (associated with a complete lack of glucose and oxygen) may be compensated by the glycogen deposits in Müller cells. Müller cells are abundantly endowed with the brain-specific glycogenolytic isoenzyme, brain-specific isozyme (BB) glycogen phosphorylase. The elevation of the extracellular K^+ concentration causes rapid glycogenolysis in Müller cells, indicating that the amount of glial glycogenolysis depends on the momentary neuronal activity in the retina.

The glycolytic metabolism of glial cells results in the generation of lactate, which is converted into pyruvate. Pyruvate is then released from the cells, probably via their monocarboxylate transporter MCT2, and is taken up by the neurons which use it as a substrate for their Krebs cycle (Figure 4, right). This interaction has beneficial consequences for both neurons (which are being fed by the substrate of their metabolism, without the necessity of their own unefficient but costly glycolytic pathway) and glial cells (which are getting rid of the unwarranted acidic end product of their metabolism). A mirror image of this mutual benefit arises at the end of the aerobic pathway when CO_2 and water are generated by the neurons. By means of the glia-specific enzyme carboanhydrase, the Müller cells transfer CO_2 to HCO_3^- and either transport it via a H^+/ HCO_3^- exchanger into the vitreous body and the blood vessels or use it for their lipid synthesis. Again, the neurons (which cannot get rid of CO_2 on their own) are liberated from a potentially dangerous end product, and the Müller cells (which do not generate enough CO_2 in their own glycolytic metabolism) receive a substrate for their metabolism (Figure 4, right).

Homeostasis of the Extracellular Space Milieu

Neurotransmitter Recycling

For the fast clearance of the neurotransmitters from synaptic spaces, the Müller cell membranes contain uptake systems for various amino-acid transmitters such as glutamate, γ-aminobutyric acid (GABA), and glycine. The clearance of synaptic glutamate by Müller cells is required for the normal function of excitatory synapses and for the prevention of neurotoxicity. After experimental inhibition of glutamate uptake by Müller cells, even low concentrations of

extracellular glutamate become neurotoxic. The main glutamate transporter expressed by Müller cells is the glutamate/aspartate transporter (GLAST) (excitatory amino acid transporter 1 (EAAT1)). GLAST is the predominant transporter for the removal of glutamate within the retina, also mediating the rapid termination of the postsynaptic action of glutamate. The Na^+-dependent glutamate uptake of Müller cells is electrogenic and, therefore, energy dependent; a very negative membrane potential is necessary for efficient uptake. Cell depolarization (e.g., by activation of glial ionotropic receptors or by elevated extracellular K^+) decreases the uptake rate. After its uptake, glutamate is intracellularly converted into glutamine by the enzyme glutamine synthetase, which is expressed only by Müller cells. Likewise, after the uptake of GABA into Müller cells, it is converted to glutamate, which then is fed into the glutamine synthetase pathway. Glutamine is transported back to neurons as a precursor for their synthesis of glutamate and GABA (glutamate–glutamine cycle of the retina). Due to the efficiency of the glutamine synthetase, demonstrable levels of glutamate within Müller cells are present only when the enzyme is experimentally inhibited or downregulated under pathological conditions. When the glutamine synthetase activity is experimentally blocked in Müller cells, retinal neurons lose their glutamate and the animals become rapidly (within 2 min) functionally blind.

The uptake of glutamate by Müller cells is also important for the retinal defense against free radicals and toxic factors. Müller cells synthesize the tripeptide glutathione from glutamate, cysteine, and glycine. Reduced glutathione is provided to neurons and acts as a scavenger of free radicals and reactive oxygen compounds. As the glutamine synthetase of Müller cells is the only enzyme available for ammonia detoxification in the retina, this is an important sideways function of neurotransmitter recycling.

K^+ Homeostasis

Müller cells are crucially involved in retinal K^+ homeostasis by mediating transcellular spatial buffering K^+ currents, which counteract the changes in extracellular K^+ concentration associated with neuronal activity. A dysregulation of the K^+ homeostasis causes K^+-evoked neuronal hyperexcitation and glutamate toxicity. Active neurons release K^+ ions predominantly into the extracellular clefts of the plexiform (synaptic) layers. This causes a local driving force for K^+ inward currents into Müller cells; thus, Müller cells take up the neuron-derived excess K^+ from the plexiform layers and release a similar amount of K^+ into fluid-filled spaces outside of the neural retina (blood, vitreous, and subretinal space) (Figures 4 and 6). Passive K^+ currents through inwardly rectifying K^+ (Kir) channels play a major role in counteracting fast extracellular K^+ increases.

The plasma membranes of Müller cells are highly permeable to K^+ as a consequence of the high density of specialized K^+ channels. The high K^+ permeability causes a very negative membrane potential of the cells close to the equilibrium potential of K^+ ions (approximately -80 mV), which is an essential precondition for virtually all neuron-supportive functions of the cells. A block or downregulation of the K^+ channels depolarizes the membranes. In addition to Kir channels, which are active at the resting membrane potential, Müller cells express various other types of K^+ channels that are activated on membrane depolarization, including voltage-gated, outwardly rectifying K^+ (Kv) channels, tandem-pore (TASK) channels, and Ca^{2+}-activated, big conductance K^+ (BK_{Ca}) channels. Among the various subtypes of K^+ channels expressed by Müller cells, two subtypes of Kir channels have been mainly implicated in mediating the K^+ buffering currents: Kir4.1 and Kir2.1. The proteins of both channels are expressed in a polarized fashion in the plasma membrane of Müller cells. The Kir4.1 protein shows a strong enrichment in such membrane domains across which the cells release excess K^+ during neuronal activity (i.e., in perivascular membrane sheets and at the inner and outer limiting membranes) (Figures 6(b) and 6(c)). Kir4.1 channels are weakly rectifying; that is, they mediate outward and inward K^+ currents with similar amplitudes (Figure 6(a)). The Kir2.1 channel is strongly rectifying and, thus, predominantly mediates inward K^+ currents. The Kir2.1 channel protein is expressed by Müller cells in such membrane domains which abut neurons of the inner retina and within both plexiform layers (Figures 6(b) and 6(c)). The differential expression of Kir4.1 and Kir2.1 channels (together with the local transmembranous K^+ gradients) determines the direction of the transglial K^+ currents; excess K^+ is absorbed by Müller cells from extracellular spaces around neurons and is distributed into spaces outside the neuroretina.

Water Homeostasis

Water accumulates in the retinal tissue due to (1) the metabolic production of water associated with the oxidative synthesis of adenosine 5′-triphosphate (ATP), (2) the influx of water from the blood coupled to the uptake of metabolic substrates such as glucose, and (3) the forcing of water into the retina by the intraocular pressure. The accumulated water is continuously redistributed out of the neuroretina by the pigment epithelium, which dehydrates the subretinal space, and by Müller cells, which dehydrate the inner retina. Both the pigment epithelium and Müller

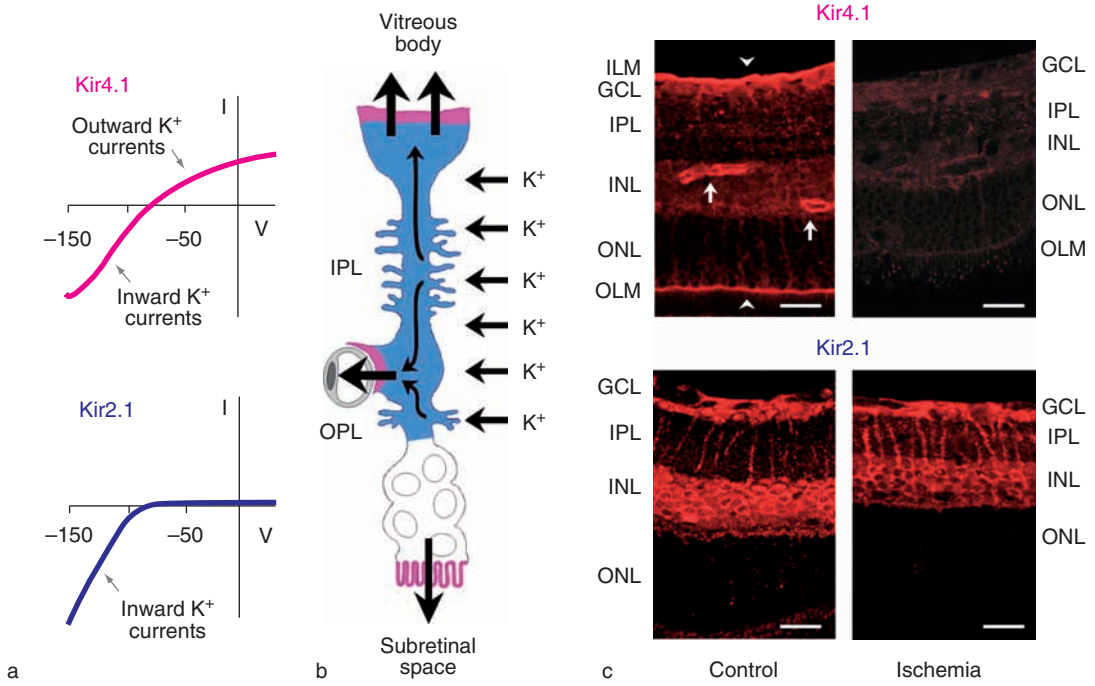

Figure 6 Müller cell physiology and K^+ (Kir) channels: (a) weakly rectifying Kir4.1 channel and strongly inward-rectifying Kir2.1 channel; (b) the differential expression of Kir channels causing a directional flow of K^+ ions through the Müller cells during neuronal activation, that is, Müller cells take up the neuron-derived excess K^+ ions and release them into extraretinal spaces; (c) after transient ischemia of the rat retina, the expression of the Kir4.1 channels strongly decreases while the Kir2.1 channels remain expressed in Müller cells, impairing K^+ homeostasis in the injured retina. Müller cells express different subtypes of inwardly rectifying Kir channels. The weakly rectifying Kir4.1 channel (cf. the current–voltage relation of the currents in (a)) is expressed at the contacts of Müller cells to extraretinal spaces, that is, at both limiting membranes (arrowheads in (c)) and around the blood vessels (arrows in (c)). The strongly inward-rectifying Kir2.1 channel is expressed in Müller cell membranes which contact neurons. This differential expression of Kir channels causes a directional flow of K^+ ions through the Müller cells during neuronal activation. This clearance mechanism is called spatial buffering or potassium siphoning. GCL, ganglion cell layer; ILM, inner limiting membrane; INL, inner nuclear layer; IPL, inner plexiform layer; ONL, outer nuclear layer; OLM, outer limiting membrane; OPL, outer plexiform layer. Scale bar = 20 μm. Adapted from Bringmann A, Pannicke T, Grosche J, et al. (2006) Müller cells in the healthy and diseased retina. *Progress in Retinal and Eye Research* 25: 397–424.

cells mediate transcellular water fluxes that are coupled to the fluxes of osmolytes (mostly ions) and which are facilitated by the polarized expression of water-transporting proteins, the aquaporins. The co-expression of Kir4.1 and aquaporin-4 water channels in Müller cell membranes at interfaces to extraretinal spaces (Figure 7) has led to the suggestion that the water transport through Müller cells is tightly coupled to the spatial buffering K^+ currents mediated by the cells. In addition, aquaporin-4 water channels are also expressed in the plexiform layers; thus, aquaporin-4 is expressed in all membrane domains through which Müller cells take up and release excess K^+. The coupling of water and K^+ transport through Müller cells is also suggested by the observation that in aquaporin-4 knockout mice the b-wave potential of the electroretinogram is reduced, indicating a reduced or disturbed excitability of second-order neurons after indirect inhibition of the transglial K^+ clearance. The aquaporin-4 water channels expressed by Müller cells may facilitate the constitutive redistribution of metabolic water from the retinal parenchyma into the blood as well as the activity-dependent rapid water fluxes necessary for regular neuronal activation and synaptic transmission.

pH Homeostasis

Neuronal activity leads to substantial variations in the concentration of H^+ in the extracellular space; these may affect synaptic transmission. Müller cells play an essential role in regulating the extracellular H^+ concentration by the expression of carbonic anhydrase (which converts CO_2 into HCO_3^-) and of the electrogenic Na^+/HCO_3^- cotransporter and by other acid/base transport systems or pumps. Light stimulation causes an extracellular alkalinization generated by synaptic activity, as well as a reduced metabolic activity of photoreceptor cells. Activation of the Na^+/HCO_3^- cotransporter system of Müller cells by increasing the extracellular K^+ concentration results in acid efflux and an acidification of the extracellular

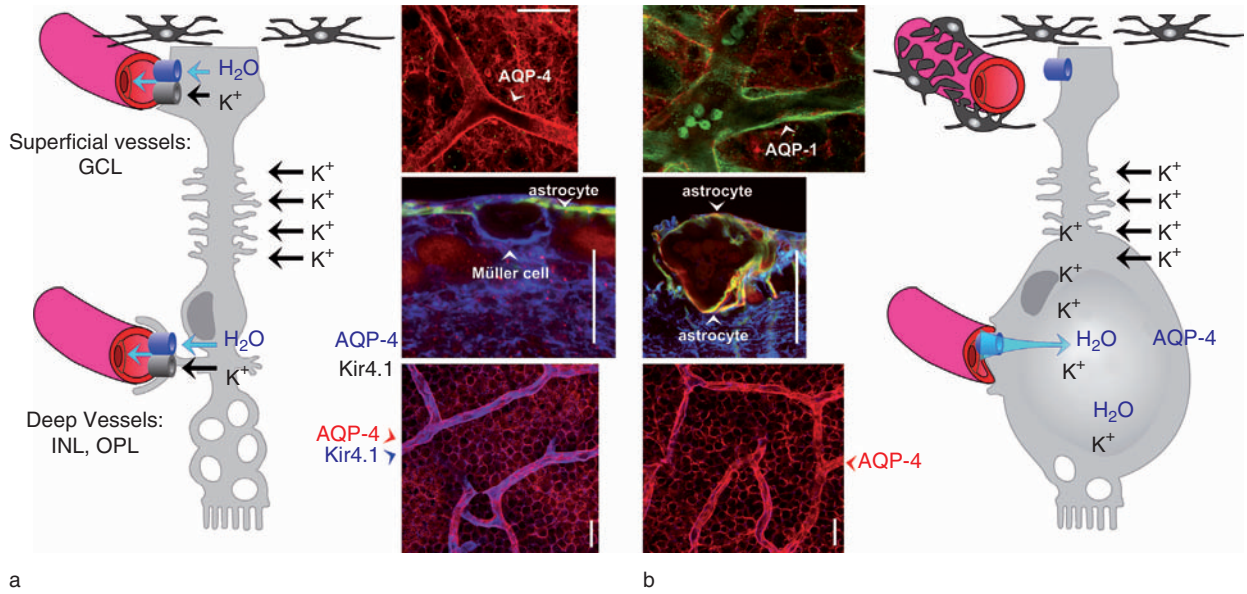

Figure 7 Water channels and Müller cell swelling in the normal and in the diabetic rat retina. In (a), aquaporin-4-expressing Müller cell membranes abut the superficial vessels in the ganglion cell layer (red immunostaining in the upper photograph, retinal wholemount; blue immunostaining in the middle photograph, retinal slice) and the deep vessels in the inner nuclear layer and in the outer plexiform layer (red staining in the lower photograph). The co-expression of aquaporin-4 and Kir4.1 (shown exemplary for the deep vessels in the lower photograph) makes sure that, osmotically coupled to the transglial K^+ currents (see **Figure 6**), water is absorbed from the retinal tissue and released into the blood (schematic drawing). Astrocyted (green immunoreactivity in the middle photograph) in the vicinity of superficial vessels do not contact the vessels. In (b), the superficial vessels are surrounded by astrocytes which express aquaporin-1 (green immunoreactivity in the upper photograph), whereas the immunoreactivity for aquaporin-4 is less pronounced. The middle photograph shows an astrocyte (green immunoreactivity for GFAP) surrounding a vessel in the ganglion cell layer. Müller cells decrease their expression of Kir4.1 around the vessels (lacking blue immunostaining in the lower photograph) and at the limiting membranes. Because the Müller cells may take up neuron-derived K^+ through other subtypes of K^+ channels (e.g., Kir2.1; cf. **Figure 6**), K^+ may accumulate within the cells, which causes an increase of the intracellular osmotic pressure. The alteration in the osmotic gradient at the glio-vascular interface results in inflow of water from the vessels into the Müller cells, which subsequently swell (schematic drawing). Therefore, Müller cell swelling may contribute to cytotoxic edema in the diabetic (and ischemic) retina. AQP, aquaporin; GCL, ganglion cell layer; INL, inner nuclear layer; Kir4.1, inwardly rectifying potassium channels; OPL, outer plexiform layer. All scale bars = 20 μm. Adapted from Bringmann A, Pannicke T, Grosche J, et al. (2006) Müller cells in the healthy and diseased retina. *Progress in Retinal and Eye Research* 25: 397–424.

space. The acidification may counterbalance the alkalinization generated by neuronal activity.

Homeostasis of the Müller Cell Volume

Neuronal activity in the retina is associated with volume changes of the neuronal cells and, particularly, their synapses. This occurs because the activation of ionotropic receptors causes an influx of Na^+ and Cl^- ions into the synapses, which is associated with a water influx and, therefore, with swelling of neuronal structures (**Figure 1(b)**). For the maintenance of the extracellular space volume, which is necessary to avoid excessive increases in the concentration of ions such as K^+ (which would disturb regular neuronal activity), Müller cells display compensatory changes of their cell shape, with a decrease in the thickness (and an elongation) of their stem processes (**Figure 1(b)**).

In addition to the synaptic structures which undergo a volume increase on activation of ionotropic receptors, the blood vessels show volume alterations during neuronal activity. Müller cells are crucially implicated in the activity-dependent regulation of the local blood flow. Light-evoked dilation and constriction of arterioles is mediated by a purinergic neuron-to-glia signaling in the retina, probably via ATP released from activated neurons, suggesting that Müller cells contribute to neurovascular coupling. It has been suggested that the release of K^+ ions from perivascular endfeet plays a role in the local dilation of arterioles. To avoid changes in the extracellular space volume around the vessels when they dilate during neuronal activation, Müller cells should decrease the volume of their perivascular processes, despite an increase of the intracellular osmotic pressure due to the uptake of neuron-derived osmolytes such as K^+ and Na^+-glutamate. It has been suggested that glutamate, released by retinal neurons on activation, causes the activation of a purinergic volume regulatory mechanism in Müller cells via the activation of metabotropic glutamate receptors and

subsequent stimulation of autocrine ATP- and adenosine-mediated signaling. The activation of adenosine A1 receptors results in an opening of additional K^+ and Cl^- channels in the Müller cell membrane; these allow for an ion extrusion, which is coupled with a water outflow and thus inhibits swelling of the cells. This mechanism may mediate the homeostasis of the extracellular space volume during neuronal activation. It remains to be determined whether long-range Ca^{2+} waves, which propagate through the network of retinal astrocytes and Müller cells during strong neuronal activation and which proceed by autocrine ATP signaling, transmit volume regulatory signals over a wide area within the retina and therefore mediate ascending vasodilation.

Modulation of Synaptic Transmission

It has recently become obvious that in the retina there is a two-way communication between neurons and glial cells, with a significant contribution of the glial cells to neuronal information processing. Light-evoked activity in retinal neurons results in an increase in the frequency of Ca^{2+} transients in Müller cells, probably evoked by release of ATP from neurons and the subsequent activation of purinergic receptors on glial cells. Conversely, the light-evoked activity of retinal ganglion cells is modulated by retinal astrocytes and Müller cells. Ca^{2+} waves in retinal glial cells are associated with the enhancement or depression of light-evoked spiking and excitatory postsynaptic currents of retinal ganglion cells. Enhancement of ganglion cell activity is probably mediated by Ca^{2+}-dependent release of glutamate from Müller cells, whereas ganglion cell hyperpolarization is mediated by the release of ATP from Müller cells, extracellular conversion of ATP into adenosine, and activation of neuronal adenosine A1 receptors, which finally inhibits the neurons by increasing K^+ conductance. Müller cells may also stimulate neuronal activity by the release of the neuromodulator D-serine. In the retina, Müller cells appear to be the primary source of D-serine. D-Serine plays a role as an endogenous co-agonist of the NMDA receptor and is therefore involved in setting the sensitivity of ganglion and amacrine cells.

Support of Neuronal Survival

Neurotrophic factors such as basic fibroblast growth factor (bFGF), brain-derived neurotrophic factor (BDNF), ciliary neurotrophic factor (CNTF), and glial cell line-derived neurotrophic factor (GDNF) promote the survival of retinal neurons and photoreceptor cells. Müller cells play an active role in the survival of retinal neurons and photoreceptor cells, in particular by releasing such neurotrophic factors and/or by mediating the effects of these factors. The protective effects of neurotrophic factors on photoreceptor cells may be mediated directly (in the case of bFGF) or indirectly (in the cases of BDNF and CNTF, whose receptors are not present on photoreceptor cells) via the activation of the Müller cells. Factors secreted by activated microglial cells, or extracellular glutamate, boost the production of trophic factors by Müller cells, thus increasing the survival of photoreceptor cells. Neurotrophic factors such as CNTF may cause a shift of Müller cells toward a more neuroprotective phenotype, which is characterized by, for example, a more efficient buffering of high concentrations of glutamate. Additional (previous) retinal injuries, such as mechanical stress, or preconditioning with bright light may protect the photoreceptors from degeneration because these stimuli cause an upregulation of bFGF and CNTF expression by Müller cells.

Blood–Retina Barrier

Müller cells are communicators between vessels and neurons, responsible for the uptake of glucose from the circulation, the glycolytic metabolism of glucose, and the transfer of substrates such as lactate and pyruvate to neurons. Retinal capillaries consist of endothelial cells and pericytes which are covered by a basement membrane and are ensheathed by membranes of astrocytes and Müller cells. Müller cells participate in the establishment of the blood–retina barrier, which is constituted by the tight junctions between the vascular endothelial cells. *In vitro*, Müller cells enhance the endothelial cell barrier function under normoxic conditions and impair the barrier function under hypoxic conditions. Müller cells produce factors that increase (GDNF, neurturin, pigment epithelium-derived growth factor (PEDF)) and others that decrease (tumor necrosis factor, vascular endothelial growth factor, etc.) the tightness of the endothelial barrier. Under normal conditions, Müller cells provide a permanent antiproliferative condition for vascular endothelial cells by the release of anti-angiogenic factors such as PEDF and thrombospondin-1.

See also: Glutamate; Retinal Development: An Overview; Retinal Development: Cell Type Specification; Retinal Ganglion Cells: Receptive Fields; Retinal Glia.

Further Reading

Bringmann A, Pannicke T, Grosche J, et al. (2006) Müller cells in the healthy and diseased retina. *Progress in Retinal and Eye Research* 25: 397–424.

Bringmann A and Reichenbach A (2001) Role of Müller cells in retinal degenerations. *Frontiers in Bioscience* 6: E77–E92.

Chalupa LM, and Finlay BL (eds.) (1998) *Development and Organization of the Retina from Molecules to Function*. New York: Plenum.

Hertz L (ed.) (2004) *Non-neuronal Cells in The Nervous System: Function and Dysfunction*. Amsterdam: Elsevier.

Kettenman H, and Ransom BR (eds.) (2005) *Neuroglia*, 2nd edn. Oxford: Oxford University Press.

Kofuji P and Newman EA (2004) Potassium buffering in the central nervous system. *Neuroscience* 129: 1045–1056.

Newman EA (2003) New roles for astrocytes: Regulation of synaptic transmission. *Trends in Neurosciences* 26: 536–542.

Newman EA (2004) Glial modulation of synaptic transmission in the retina. *Glia* 47: 268–274.

Newman EA and Reichenbach A (1996) The Müller cell: A functional element of the retina. *Trends in Neurosciences* 19: 307–312.

Ransom BR, and Kettenmann H (eds.) (1995) *Neuroglia*. Oxford: Oxford University Press.

Reichenbach A (1989) Attempt to classify glial cells by means of their process specialization using the rabbit retinal Müller cell as an example of cytotopographic specialization of glial cells. *Glia* 2: 250–259.

Reichenbach A and Wolburg H (2004) Astrocytes and ependymal glia. In: Kettenman H, and Ransom BR (eds.) *Neuroglia*, 2nd edn., pp. 19–35. Oxford: Oxford University Press.

Reichenbach A, Henke A, Eberhardt W, Reichelt W, and Dettmer D (1992) K^+ ion regulation in retina. *Canadian Journal of Physiology and Pharmacology* 70: S239–S247.

Reichenbach A and Robinson SR (1995) Phylogenetic constraints on retinal organization and development: An Haeckelian perspective. *Progress in Retinal and Eye Research* 15: 39–171.

Reichenbach A, Schneider H, Leibnitz L, Reichelt W, Schaaf P, and Schümann R (1989) The structure of rabbit retinal Müller (glial) cells is adapted to the surrounding retinal layers. *Anatomy and Embryology* 180: 71–79.

Reichenbach A, Stolzenburg JU, Eberhardt W, Chao TI, Dettmer D, and Hertz L (1993) What do retinal Müller (glial) cells do for their neuronal "small siblings"? *Journal of Chemical Neuroanatomy* 6: 201–213.

Sarthy V and Ripps H (2001) *The Retinal Müller Cell. Structure and Function*. New York: Kluwer Academic/Plenum Publishers.

Uckermann O, Vargova L, Ulbricht E, et al. (eds.) (2004) Glutamate-evoked alterations of glial and neuronal cell morphology in the guinea pig retina. *Journal of Neuroscience* 24: 10149–10158.

Uga S and Smelser GK (1973) Comparative study of the fine structure of retinal Müller cells in various vertebrates. *Investigative Ophthalmology* 12: 434–448.

Multimodal Signaling in Animals

P Otovic, University of South Florida, Tampa, FL, USA
S Partan, Hampshire College, Amherst, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Animals communicate using their entire bodies and faculties. The scientific study of animal communication has progressed from early ethologists' observations of display behavior, which included detailed descriptions of elaborate combinations of visual, vocal, and chemical signals, to the current focus on experiments conducted to discover the meaning of the signals in a more controlled format. The gains in controlling variables through experimentation came at the cost of decreased complexity, however; in order to isolate a signal of interest, experimenters typically had to simplify the situation by examining one component of the display at a time. Advances in experimental techniques now allow for simultaneous manipulation of multiple facets of a display, allowing for a more complete understanding of an animal's display behavior. This renewed interest in multimodal communication (communication using signals in multiple sensory channels) has spurred new research on the topic, which is summarized here. This article discusses terminology, methods for studying multimodal communication, and the classification of multimodal and multicomponent signals and hypotheses. It also discusses the incidence of multimodal signals and channels used, and theory concerning the evolution and function of multimodal signaling. In addition, it briefly discusses the neural integration of multimodal stimuli.

Terminology of Multimodal Signals

A communication signal is defined here as any aspect of an organism that modifies the behavior of the signal recipient. This definition is broad and includes static or enduring components such as body color and size, along with dynamic or transient components such as vocal or chemical emissions. Although much attention has been focused on communication signals that contain only one component, signals that contain more than one component (composite signals) have attracted increased interest in recent years. There are two types of composite signals: multicomponent and multimodal. Multicomponent signals contain components that are transmitted in the same sensory channel. For example, male jungle fowl (*Gallus gallus*) court females with brightly colored combs, long tails, and display movements all in the visual modality (Figure 1(a)). Multimodal (or multisensory) signals contain components that are transmitted via multiple sensory channels. Many animals communicate by simultaneously producing signals from different sensory channels or modalities during courtship, aggressive, or antipredator displays. For example, arctiid moths (*Rhodogastria leucoptera*) emit an odor from their thoracic glands while moving their wings to reveal red coloration on their abdomen in an attempt to ward off potential predators (Figure 1(b)).

Methods for Studying Multimodal Signals

There are many ways to study the meaning of multimodal signals. Observational methods provide invaluable baseline data concerning the incidence of multimodal signal usage in the field. It is imperative that fieldwork be conducted to understand how these signals are used under natural conditions. Experimental methods allow the control of extraneous variables while signals of interest are manipulated. With experimental techniques, we can classify and assess the function of multimodal signals.

Video and audio playback is an experimental method that has gained popularity in recent years. This method has the potential to provide adequate control conditions and to manipulate the signal content in each condition, which allows one to posit a plausible cause-and-effect relationship between signal and response. For example, playback methods have been used to assess the function of each component in the multimodal courtship display of wolf spiders, which consists of visual foreleg waving and vibration of the substrate. These methods allow one to present each unimodal component to an organism in isolation as well as simultaneously in order to assess and compare the elicited responses. Playback methods can also be used for presenting conflicting cues to an organism. A visual component from a conspecific signal can be presented along with an auditory component from, for instance, a heterospecific signal. This is especially useful for determining the relative importance of each signal channel.

However, video playback has several limitations that may be circumvented by an alternative method – mechanical modeling of a live conspecific. First, video stimuli may be perceived differently by humans than by nonhuman animals, with problems arising from differences in critical flicker fusion frequency or spectral sensitivity. Second, a mechanical model has three dimensions, which makes it more lifelike by adding cues for depth perception and allowing active perception by the recipient (in which the recipient can get multiple views by moving its head).

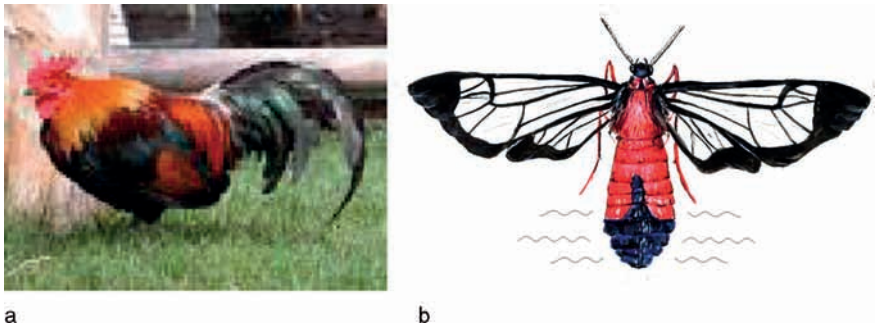

Figure 1 (a) Multicomponent visual display: Red jungle fowl use their long tail feathers and brightly colored combs in courtship. Reproduced from Hogle Zoo. (b) Multimodal visual/chemical display: An arctiid moth exposes its red abdomen while emitting a thoracic odor to deter predators. Adapted from a drawing by Jane Carlson, Louisiana State University, with permission.

Third, mechanical modeling allows the live animals to interact with the model in real time and space. Robotic models have successfully elicited species-specific behavior in animals, including frogs, bowerbirds, honeybees, and lizards. This method is ideal for examining multimodal signals since models can perform visual movements, mechanical vibrations, or emit sounds via speakers.

Classification of Multimodal Signals

In order to understand the plethora, variety, and complexity of animal signals, researchers have devised categorization systems for classifying them within or across taxonomic boundaries. Communication signals have traditionally been classified within species according to behavioral context (aggression or courtship, etc.) or channel (auditory or visual, etc.). Researchers have devised new classification schemes specifically for composite, multicomponent signals. These new approaches to categorizing composite signals are discussed in the following sections along with a classification scheme created specifically for multimodal (as opposed to multicomponent) signals.

Categorization by Receiver Fitness

Multicomponent communication signals have been classified by hypotheses proposed to explain how they affect the survival, growth, and reproductive success of a signal receiver; these fitness effects can be positive, negative, or neutral (Table 1). For example, signals that are assumed to enhance receiver fitness comprise one group, such as feather color in greenfinches (*Carduelis chloris*). In these birds, the brightness of color patches on the tail reflects an individual's ability to resist parasites and the brightness of color patches on the wing conveys age information, so these multiple visual signals facilitate accurate mate assessment. Signals that have a negative impact on receiver fitness constitute a second group. Signals evolving via intersexual conflict, such as deceptive signals used by males to overcome mating resistance on the part of females, are classified in this group. Signals that are assumed to have negligible or only indirect effects on receiver fitness encompass a third group. For instance, some signal components may be remnants of past selection for sexually selected traits and may in fact be useful in the present only as attractants to ensure that a male meets the threshold level of attraction for a female.

Categorization by Other Selective Pressures

Signals have also been grouped according to other selection pressures affecting their evolution (Table 2). One category includes composite signals that are thought to have evolved due to changes in the environment. For instance, yellow and green colors are most salient in green-colored forest shade, whereas blue colors are most salient in blue-colored open woodland shade, so one would expect composite visual displays of animals in each of these habitats to emphasize these respective colors. Signals that have been shaped primarily by a receiver's sensory apparati, such as Fisherian runaway traits, or by individual variability in the perceptual abilities of receivers form another group. Individuals of the same species vary in their capacity to perceive signals from different modalities. For example, human children who are musically inclined process verbal stimuli more effectively than those inclined to the visual arts, whereas the artistically inclined recognize visual stimuli faster than the musically inclined children.

Categorization by Meaning

If signal meaning can be determined, for example, by correlating a feature such as redness of neck feathers with nutrient quality, then composite signals can be classified by their intrinsic meaning or information content (Table 3). Hence, signal components that

Table 1 Categorization of signals by how they affect receiver fitness

Adaptiveness of preferences	*Form of cues*	*Proposed hypotheses*	*Purpose of different cues*	*Proximate benefit gained by the receiver*
Adaptive, increase fitness through less errors or lower costs	Informative cues	Multiple messages	Indicator traits that reflect different qualities	
			Sensitive to different environmental or intrinsic factors	Evaluated together, traits reflect overall quality
			Develop at different times	Individuals can emphasize cue most relevant for them
		Backup cues	Indicator traits that reflect the same quality	Increased accuracy of assessment
		Species recognition	Different cues for species recognition and mate assessment	Avoidance of costly matings with other species
		Fisherian cues	Indicate heritable attractiveness	Offspring inherit the attractive trait
	Uninformative cues	Unreliable cues	Exploitation of sensory biases in the receiver	Facilitates detection of the signaler
		Receiver psychology	Additional cues enhance detectability, discriminability, or memorability of the signal	Facilitates mate assessment
	Informative or uninformative cues	Multiple sensory environments	Indicator or unreliable cues that are used	Facilitates mate assessment or increases the amount of information gained
			At different distances	
			At different stages of the courtship ritual	
			In different signal transmission habitats	
Nonadaptive, no effect on fitness	Uninformative cues	Unreliable cues, threshold cues	Exploitation of sensory biases in the receiver	
			Remnants from past selection	
Maladaptive, reduce fitness	Uninformative, manipulating cues	Intersexual conflict	Manipulation of mating resistance by exploitation of preexisting sensory biases	

Adapted from Candolin U (2003) The use of multiple cues in mate choice. *Biological Reviews* 78: 575–595.

Table 2 Categorization of signals by other selection pressures

Multiple sensory environments	Signal transmission and reception are influenced by environmental conditions. Efficacy backup: each signal acts as a backup to the other in the presence of environmental variability. Efficacy trade-off: the transmission characteristics of each signal vary; each one is specialized to overcome a particular transmission impediment.
Sensory constraints	The receiver's sensory system influences the detection and processing of signals. Parallel processing: multiple bits of information can be processed simultaneously, potentially reducing the latency to receiver response. Sensory bias for multisensory integration: given the existence of multisensory neurons, a multimodal signal enhances the neuronal response, thereby increasing the salience of the event. Sensory overload: a signaler can 'jam' a receiver's reception and/or processing system by providing more information than can be handled by the receiver's sensory system. Perceptual variability: each signal acts as a backup to the other in the presence of perceptual variability among receivers.

Adapted from Hebets EA and Papaj DR (2005) Complex signal function: Developing a framework of testable hypotheses. *Behavioral Ecology and Sociobiology* 57: 197–214, with permission.

Table 3 Categorization of signals by content

Multiple messages	Different signals provide different information: either different types of information or information about different aspects of signaler quality. Quality: both signals convey different aspects of signaler quality; referred to in the literature as 'multiple messages.' Static vs. dynamic: one signal conveys more static properties (i.e., genetic indicators) while a second conveys more dynamic properties (i.e., current condition). Species recognition: one signal conveys species identity while another conveys a different type of information, such as signaler quality. Quality plus: one signal conveys signaler quality while another provides a different type of information such as signaler location. Receiver variability: different signals are directed at different receivers due to variable receiver preferences.
Redundant signal	Different signals provide the same information, allowing for increased accuracy of receiver response.

Adapted from Hebets EA and Papaj DR (2005) Complex signal function: Developing a framework of testable hypotheses. *Behavioral Ecology and Sociobiology* 57: 197–214, with permission.

transmit the same information may be classified in one group, such as the red-orange bill color and redundant red feathers under the wing of female northern cardinals (*Cardinalis cardinalis*), which are both correlated with body condition. Composite signal components that each convey different information may be classified in another group, such as multiple visual signals of size and color pattern in male pygmy swordtails (*Xiphophorus pygmaeus*). In this case, size indicates mate quality, whereas black bars on the body are likely used for species identification.

It is important to note that practitioners must be careful in their conclusions regarding signal meaning. It may be possible for two components to carry the same general message, such as that the individual is in good condition, yet the specific meanings or literal translations can differ. For example, in zebra finches (*Taeniopygia guttata*), higher song rate and redder beak color are both preferred by females, suggesting that they may be redundant signals of male quality. Nevertheless, song rate ostensibly indicates a mate's present condition, whereas beak color indicates past condition, because after controlled nutrition and exercise the finches' song rate changed much faster than their beak color. Therefore, two signal components may both function in mate choice but have slightly different meanings or act on different time courses.

Categorization by Signal Interaction

Composite signals may be demarcated by how the components influence one another (Table 4). Composite signals in which one component facilitates learning or memory for the other component could be grouped together. For instance, when learning to avoid water with an aversive taste, rats (*Rattus norvegicus*) learn

Table 4 Categorization of signals by their interactions

Multitasking	A signaler's ability to generate one signal is constrained by, and thus negatively correlated with, his or her ability to generate another signal.
Increased detection and discrimination	The presence of one signal either increases the probability and/or speed of detection of a second signal or increases the accuracy with which a receiver responds to a second signal. Amplifier: one signal increases the probability of detection of the second signal and/or decreases the detection threshold, making it more conspicuous. Alerting and attention altering: one signal either alerts a receiver to a second signal, thereby decreasing the reaction time, or influences the information-filtering mechanism of the receiver such that the receiver's attention is focused on another signal. Context: the presence of one signal provides a context in which a receiver can interpret and respond to a second signal. Emergence: a receiver is better able to discriminate among signalers due to the emergent properties of combining multiple signals in such a way that a new, unique signal is formed.
Increased learning and memory	The presence of one signal increases a receiver's ability to learn and/or remember another signal or the entire display.
Increased deception	The presence of one signal decreases the accuracy with which a receiver responds to another signal.

Adapted from Hebets EA and Papaj DR (2005) Complex signal function: Developing a framework of testable hypotheses. *Behavioral Ecology and Sociobiology* 57: 197–214, with permission.

much faster when an aversive odor is also perceived. Signals in which the production of one component limits the capacity to produce another component could form another group. Male brown-headed cowbirds (*Molothrus ater*) use visual and vocal signals in their communication displays; however, they often remain still while vocalizing in order to prevent any detrimental effects to the sound quality of their vocalizations. Composite signals in which one component affects the ability to accurately perceive the other component could comprise another group, such as when a male stickleback (*Culaea inconstans*) releases a chemical signal in order to attract a female's attention toward his visual signal components of courtship, such as his red ventral patch and zigzag movements.

Categorization by Receiver Response

One classification system has been set up to categorize specifically multimodal signals (Figure 2). It is based on receiver response to both the multimodal signal and the individual signal components. Signals are then categorized based on the relationship between their components. Multimodal signals can be divided into those that are redundant and those that are nonredundant. In order to determine which applies, the receiver's response to each unimodal component as well as to the composite signal must be analyzed. This classification system is discussed in more detail because it is meant to be used specifically for multimodal signals.

Response to redundant multimodal signals

A redundant multimodal signal occurs when a receiver's response to a composite signal is qualitatively identical to its response to each of the unimodal components. Therefore, a redundant composite signal contains components with the same general meaning, sometimes referred to as backup signals. For example, when begging for food from their parents, young magpies (*Pica pica*) use a multimodal display that consists of a gaping beak (visual component) and begging vocalizations (auditory component), both of which elicit parental feeding behavior.

Redundant signal components can be parsed into two subcategories according to the quantitative nature of the receiver's response: equivalence and enhancement (Figure 2). Equivalent signals refer to those in which the receiver's response to the composite signal and each signal component is qualitatively and quantitatively homogeneous. For example, male moths (*Cyncia tenera*) emit acoustic clicks and chemical pheromones during courtship displays. Females are equally willing to copulate with the males no matter whether acoustic, chemical, or both channels of information are available, suggesting that this composite signal would be classified in the equivalence subcategory.

Enhancement, a heightened response to the composite signal relative to the response to either single component, is more common than equivalence. For example, female fruit flies (*Drosophila melanogaster*) are much more likely to mate if simultaneously presented with auditory (wing vibration song) and chemical (hydrocarbons from male cuticle) components of the male's courtship display than if they are presented with either unimodal component alone. Female pigeons (*Columba livia*) respond with more courtship behavior after being exposed to the male's multimodal bow and coo display than after being exposed to only the bow or the coo.

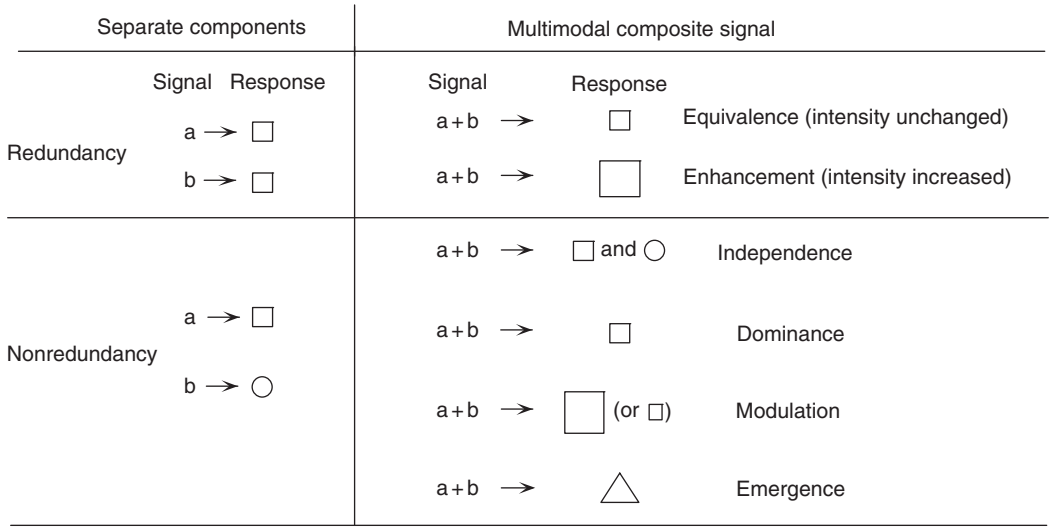

Figure 2 Classification of communication signals by receiver response. First, the responses to each signal component alone must be compared in order to classify the signal as redundant or nonredundant. Then, to classify into subcategories, the responses to the signal components must be compared to the response to the whole composite signal. Different shapes indicate qualitatively different responses, whereas different sizes represent quantitatively different responses. Reproduced with permission from Partan SR and Marler P (1999) Communication goes multimodal. *Science* 283: 1272–1273. Copyright (1999) American Association for the Advancement of Science.

Technically, enhancement may be parsed into three further subcategories: minor enhancement, summation, and multiplicative enhancement (Figure 3). Minor enhancement occurs when the response to the composite, multimodal signal is greater than the response to each isolated component but less than the sum of the responses to each component. Summation occurs when the response to the multimodal composite signal is equal to the sum of the responses to each unimodal component. Seitz's law of heterogeneous summation is an example of this subcategory; he found that the aggressive response of cichlids (*Haplochromis* spp.) to models that consisted of various visual components was equal to the sum of the aggressive responses to each of the components presented in isolation. Finally, multiplicative enhancement occurs when the response to the composite, multimodal signal is greater than the sum of the responses to each unimodal component. The three subcategories of enhancement can be difficult to accurately distinguish when strictly studying behavior; this type of distinction is easier to measure when observing neuronal response to multiple stimuli.

Response to nonredundant multimodal signals Nonredundant multimodal signals refer to composite signals in which the response to each unimodal component is qualitatively different, or when the response to the composite signal is qualitatively different than the response to at least one of the unimodal components. For example, rhesus macaques (*Macaca mulatta*) are likely to respond to a silent threat (e.g., an open-mouthed gesture (**submissive**)) with fearful and submissive responses. However, when this visual threat is combined with an aggressive vocalization (e.g., a bark), the monkeys are as likely to respond with aggressive responses as they are with submissive responses. Nonredundant signals may be divided into four subcategories: independence, dominance, modulation, and emergence.

Independence indicates that the response to the composite, multimodal signal consists of both responses to each of the unimodal components. Hence, the response to the multimodal signal is merely a combination of the (qualitatively different) responses to each unimodal component. For example, in the tropical wandering spider (*Cupiennius salei*), males display undirected courtship behavior in the presence of female pheromones alone. After being exposed to tactile cues on the web due to movement of the abdomen and legs, the males move in the female's direction. When both chemical and tactile cues are present, the males direct courtship behavior in the female's direction; the response to the composite signal is simply the combination of the responses to each unimodal component.

Dominance refers to a situation in which the response to the composite, multimodal signal is equivalent to the response to only one of the unimodal components. When soliciting play, domestic dogs (*Canis familiaris*) may combine a paw lift gesture with a growl. Growls generally indicate aggression in dogs, and paw lifts generally indicate playfulness; when the two signals are combined, play behavior is

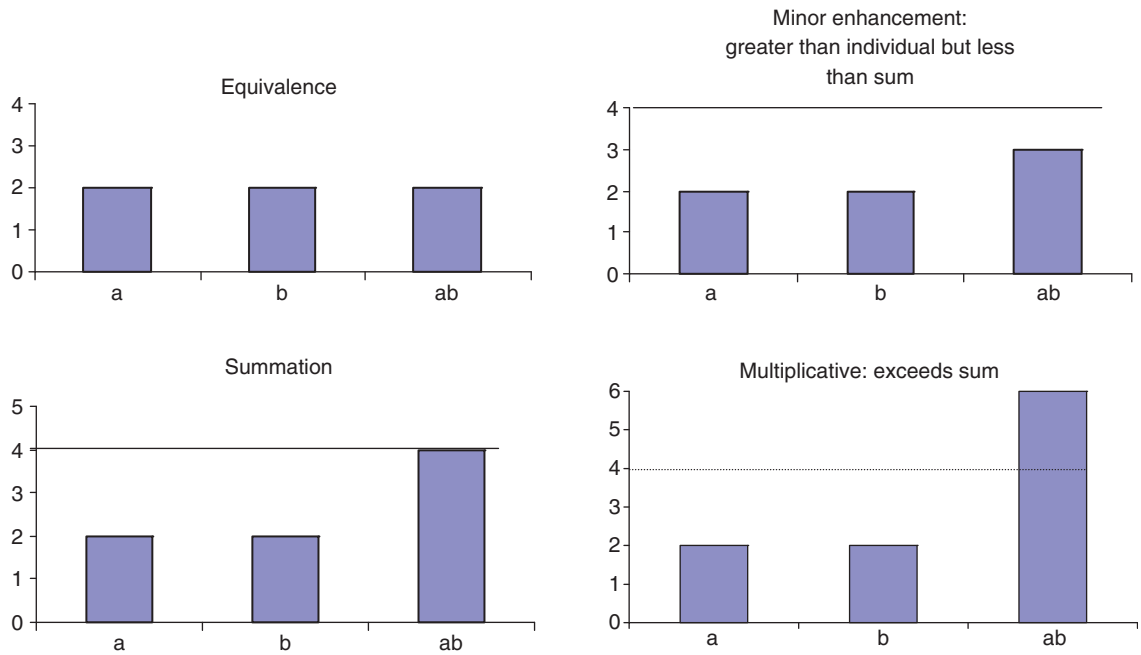

Figure 3 Redundant signals: equivalence and enhancement. Vertical axis indicates a measure of response to the signals. Equivalence occurs when the response to each component is qualitatively similar to the response to the composite signal. Minor enhancement, summation, and multiplicative enhancement occur when the response to the composite signal is less than, equal to, or greater than the sum of the responses to each signal component, respectively. Reproduced from Partan SR (2004) Multisensory animal communication. In: Calvert GA, Spence C, and Stein BE (eds.) *The Handbook of Multisensory Processes*, pp. 225–240. Cambridge, MA: MIT Press, with permission.

elicited. Therefore, the play component is thought to dominate the aggressive component when the two are combined to create a multimodal signal.

Modulation occurs when the response to the composite signal is qualitatively similar to the response to one of the components but quantitatively dissimilar. Dart-poison frogs (*Epipedobates femoralis*) use a multimodal display during territory defense that consists of vocal sac pulsations and vocalizations. The frogs do not display any noticeable responses to the pulsating vocal sac alone, but some frogs move in the direction of the sound source after hearing vocalizations without a visual component. When the multimodal signal is presented, nearly all males approach the displaying frog. Hence, the response to the multimodal signal is an amplified version of the response to the auditory signal component. In fact, the term 'amplifier' has been used to describe this situation when it occurs solely in the visual modality.

Modulatory multimodal signals can also reduce the response that is elicited from one of the unimodal components. Male snapping shrimp (*Alpheus heterochaelis*) respond aggressively after viewing a female conspecific's open chela, or claw, but do not respond aggressively when only female chemical cues are detected. When the males have access to the chemical and the visual components, their aggressive response diminishes considerably relative to their response to the visual-only component.

Finally, emergence refers to a situation in which the composite signal elicits a response that is new, in that it is qualitatively different from the responses elicited to either unimodal component in isolation. For instance, young chickens (*Gallus domesticus*) display a robust aversive response to stimuli with a combination of pyrazine odor and red or yellow coloration (aposematic warning signals). However, they do not avoid stimuli that contain only the visual or the chemical component. Therefore, a novel response is elicited from the multimodal signal.

Incidence of Multimodal Signaling and Channels Commonly Combined

It is likely that all animals (including humans) communicate via multiple sensory channels simultaneously. Surveys of the literature show that multimodal signals have been observed in both invertebrates and vertebrates. Depending on their sensory acuity and on habitat features, different taxa rely on different combinations of channels. Insects commonly combine chemical with visual or vibrational signal components, for example, whereas birds usually combine auditory and visual signals. In a survey of 30 species whose multimodal signaling behavior has been examined in

detail, invertebrates employed only those multimodal signals that included chemical or vibrational components; vertebrates strongly preferred auditory/visual composites (Table 5). Almost all (26 of 30) examples include at least one short-distance communication channel (visual, tactile, or vibratory). Therefore, it appears that one component of a multimodal signal often serves to relay information to nearby receivers.

Out of a subset of 18 examples from the same survey that were definitively classified into one of the categories shown in Figure 2, most fell into the redundant enhancement or the nonredundant modulation groups (Table 6). Interestingly, auditory components were more common in those signals classified as redundant (six of the seven redundant examples contained an auditory component compared to 3 of 11 nonredundant ones), whereas visual and chemical components were more common in the nonredundant group (9 of 11 nonredundant examples contained visual components compared to three of seven redundant ones; 7 of 11 nonredundant examples contained chemical components compared to three of seven redundant ones). This suggests that auditory signals may more often be merely repetitive, whereas visual and chemical signals carry unique information.

Function and Evolution of Multimodal Communication

Function of Redundant and Nonredundant Signals

Redundant signal components provide insurance that the signal will be received. This is especially useful in environments with biotic or abiotic impediments to signal transmission. For instance, if there is an abundance of auditory noise in a particular environment, incorporating a visual component to the signal will increase the chances of it being detected. Examples of abiotic impediments include dense habitats that impede visual displays or broken substrate unfavorable for seismic signaling. In these cases, the incorporation of an additional signal channel will improve transmission. Given the assumption that each signal is perceived by a receiver with a certain

Table 5 Survey of sensory channels used by 30 species whose multimodal signaling has been examined in detail[a]

Sensory channels used	*Invertebrates (n = 11)*	*Vertebrates (n = 19)*
Auditory/visual ($n = 11$)	0	11
Chemical/visual ($n = 9$)	5	4
Auditory/chemical ($n = 4$)	2	2
Chemical/tactile ($n = 2$)	0	2
Chemical/vibrational ($n = 2$)	2	0
Visual/vibrational ($n = 1$)	1	0
Auditory/vibrational ($n = 1$)	1	0

[a]Species names and signals used in these examples are listed in Partan SR and Marler P (2005) Issues in the classification of multimodal communication signals. *American Naturalist* 166(2): 231–245, tables 2 and A1. This is neither an exhaustive nor necessarily representative list of species that use multimodal signals; rather, it is a survey of species that have been studied by biologists who explicitly examined the use of each signal component in multimodal communication.

Table 6 Sensory channels used, by signal category, for all 18 signals from the survey that were able to be categorized according to Figure 2[a]

Redundant			*Nonredundant*			
Sensory channels	*Redundant Equivalence*	*Enhancement*	*Nonredundant Independence*	*Dominance*	*Modulation*	*Emergence*
Auditory/chemical	1	1				
Auditory/vibrational	1					
Auditory/visual		3		1	2	
Chemical/vibrational		1	1			
Chemical/visual					4	1
Chemical/tactile					1	
Visual/vibrational					1	

[a]Species names and signals used in these examples are listed in Partan SR and Marler P (2005) Issues in the classification of multimodal communication signals. *American Naturalist* 166(2): 231–245.

amount of error, redundant signals may also increase the probability that a signal is detected accurately. In addition, redundant signals may increase the salience of a display. Hence, redundant signals have the potential to increase stimulus intensity, such as when sex pheromones increase the visual attractiveness of a human of the opposite sex. In other cases, they also may have the potential to jam a receiver's sensory apparati in order to diminish or prevent a response (Table 2). Finally, redundant signals may facilitate stimulus recognition; humans utilize redundant visual and vocal components when speaking, and their combination enhances the recognition of the spoken word.

Nonredundant signal components have the potential to increase the amount of information sent per unit time. It is likely that the 'coo' component of the male Barbary dove (*Streptopelia risoria*) courtship display indicates individual identity, whereas the 'bow' component indicates sex. Male jumping spiders (Salticidae) likely signal different aspects of mate quality in their multimodal display that includes visual and seismic signals. Although there are fewer examples of nonredundant than of redundant multimodal signals, many animals perform multicomponent, rather than multimodal, displays in which each component represents a different aspect of mate quality. For example, the black cap of the American goldfinch (*Carduelis tristis*) is likely influenced by dominance rank, whereas the color of the plumage and bill indicate intestinal health.

There are a number of other potential functions that apply to both redundant and nonredundant multimodal signals. Multimodal stimuli are generally detected faster and elicit a shorter reaction time than their unimodal components. In addition, the detection threshold is lower and discriminability is facilitated for multimodal versus unimodal stimuli. For example, cats (*Felis cattus*) are more accurate in perceiving the intensity of light when it is presented with an auditory stimulus than when the auditory component is missing. Multimodal signals can also facilitate learning and memory relative to their unimodal counterparts. For instance, male nightingales (*Luscinia megarhynchos*) learn more songs and repeat them more accurately if they are accompanied by a pulsating strobe light during acquisition. Finally, a component from one modality can often provide the context for a second signal. Pheromones from the snapping shrimp (*Alpheus heterochaelis*) ostensibly indicate sex, as implied previously, and these chemical cues modify the response to the open chela. If female pheromones are detected, a male conspecific's response depends on chela size. In contrast, responses are independent of chela size when male pheromones are detected. Therefore, the chemical cue appears to provide a context in which to interpret the visual chela cue.

Game Theory and Sexual Selection

Game theoretical models have been proposed to explain the evolution of multicomponent and multimodal communication signals, particularly as used in mate choice. Some models have predicted that signals with multiple components will only be stable over evolutionary time if each signal component plays an equal role in indicating the quality of a mate. The backup signal hypothesis states that each component within a composite signal is indicative of the same quality. Hence, backup signals serve to reduce the probability of a receiver assessing a quality inaccurately. However, individuals may vary in the traits they find important in mate choice. In addition, a composite display may be directed at multiple receivers. Consequently, other hypotheses, such as the multiple messages hypothesis, have been posited to explain the evolution of signal components that reveal different types of information about an organism.

Particular attention has been paid to signals that have evolved via sexual selection. Two common theories to explain this process are exploitation of a mate's preexisting sensory biases and Fisherian runaway selection. Exploitation of sensory biases occurs when animals produce signals that are sent via specific channels or ranges that take advantage of a conspecific's preexisting sensory receptors for stimuli within those channels or ranges. For example, courting male water mites (*Neumannia papillator*) produce vibrations that mimic prey items in order to attract females. It is thought that this courtship display evolved subsequent to female sensory systems becoming attuned to detecting the vibratory stimuli produced from nearby copepods. Fisherian runaway selection, in which male traits evolve due to arbitrary female preferences, may explain the evolution of traits such as the costly, elaborate tail feathers of male peacocks that do not facilitate their owners' survival. Males most likely produce these feathers for the sole purpose of increasing their attractiveness to females.

Another theory proposed to explain the evolution of composite signals involves amplifiers or modifiers. These are signal components that may have no intrinsic meaning but instead can function to alter the perception of another component. For example, the feathers of many birds are outlined in dark colors. Although the outline coloration may not convey any information alone, it may attract more attention

toward the feathers or make it easier to accurately perceive feather condition; the outline is a modifier to feather condition in this multicomponent visual signal. Similarly, male blue-black grassquits (*Volatinia jacarina*) combine a visual leaping display with a vocalization in order to facilitate localization of the acoustic signal by the receiver. The amplifier theory can also apply to signals that are not involved in

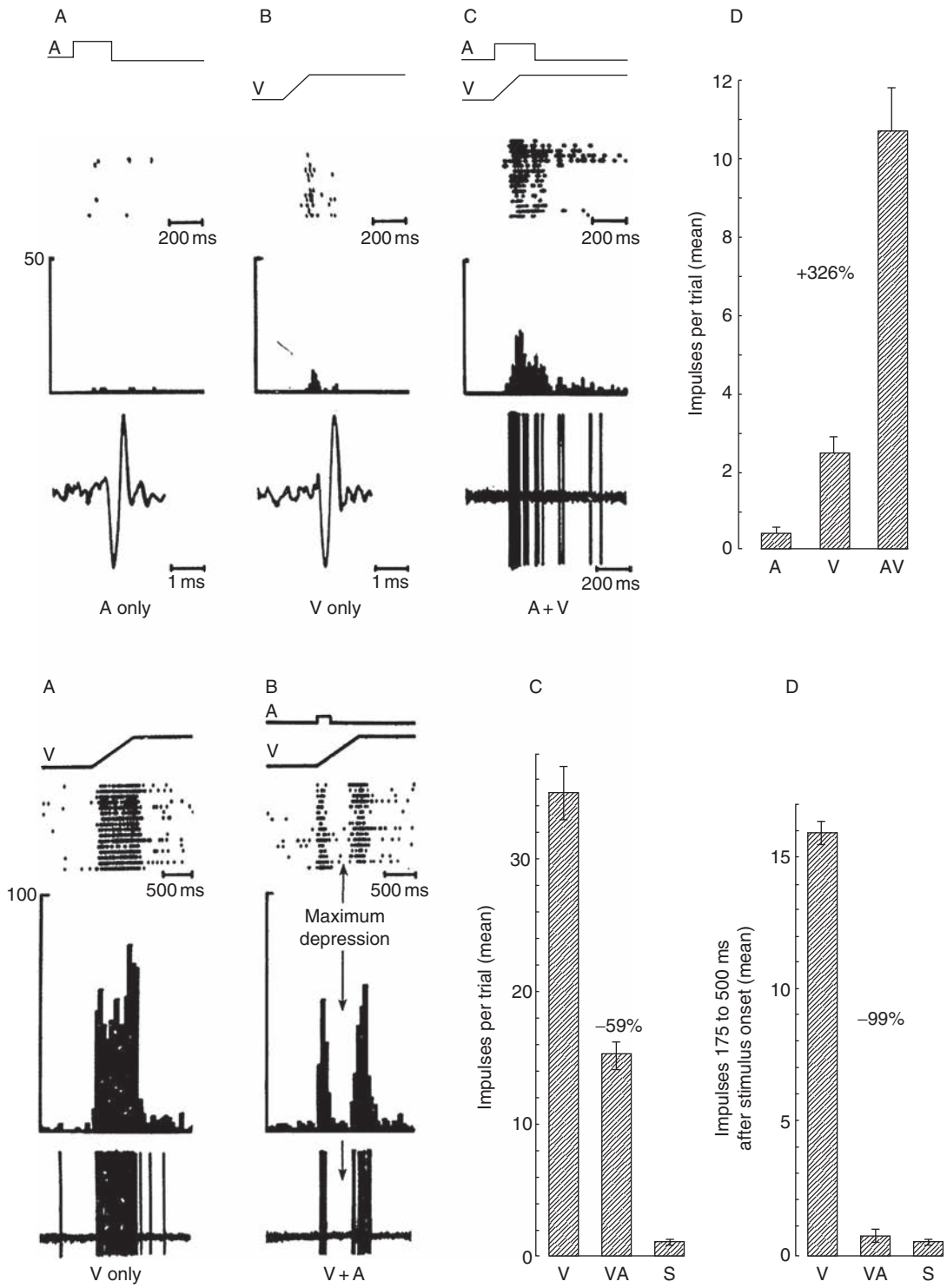

Figure 4 (Top) Response enhancement in the cat brain: the neural response to the multimodal stimulus (visual plus auditory) is greater than the response to either the audio or the visual stimuli presented in isolation. (Bottom) Response depression: the neural response to the multimodal stimulus is less than the response to the unimodal stimulus (visual only). Reproduced with permission from Meredith MA and Stein BE (1983) Interactions among converging sensory inputs in the superior colliculus. *Science* 221: 389–391. Copyright (1983) American Association for the Advancement of Science.

sexual selection. For instance, one component can help to clarify an ambiguous signal from another channel, such as when female red-winged blackbirds add a vocalization to their visual displays in order to indicate either aggression or courtship.

Neural Factors (Multisensory Integration)

The neural integration of multisensory information has traditionally been tested using lights and tones as stimuli. Current research, however, uses recordings of actual communication signals as stimuli to test biologically relevant multisensory input. For example, monkey neural responses to auditory/visual input are tested using video of monkey faces combined with audio recordings of monkey vocalizations.

Multisensory integration can be the result of higher-level neural integration of information from primary sensory areas or of single neurons that preferentially respond to multimodal stimuli. Techniques such as functional magnetic resonance imaging can reveal whether a particular brain area in general responds to multimodal stimuli, but single cell recordings are necessary to assess the presence of multimodal neurons. Multimodal neurons can show any of the same responses that were described previously for behavior: response enhancement, depression, modulation, and even emergence (Figures 2 and 4). This provides a mechanism for the behavioral reactions.

Evidence of a neural substrate for multisensory integration does not necessarily imply that this mechanism is used for purposes of communication. For example, some organisms that have neural connections between different sensory areas, such as rattlesnake (*Crotalus viridus*) visual and infrared modalities, are not known to use multimodal signals that combine these two sensory channels for communication purposes. It is plausible that this sort of preexisting connection between different sensory modalities initially subserved other behaviors, such as predation, and may later facilitate the evolution of multimodal communication.

See also: Communication in Terrestrial Animals; Electrical Perception and Communication; Game Theory and the Economics of Animal Communication; Kinship Signals in Animals; Multisensory Convergence and Integration; Seismic and Vibrational Signals in Animals; Sexual Selection and the Evolution of Animal Signals; Signal Transmission in Natural Environments; Signal Design Rules in Animal Communication; Visual Signaling in Animals.

Further Reading

Candolin U (2003) The use of multiple cues in mate choice. *Biological Review* 78: 575–595.

Endler JA (1992) Signals, signal conditions, and the direction of evolution. *American Naturalist* 139: S125–S153.

Ghazanfar AA and Santos LR (2004) Primate brains in the wild: The sensory bases for social interactions. *Nature Reviews Neuroscience* 5(8): 603–616.

Grether GF, Kolluru GR, and Nersissian K (2004) Individual colour patches as multicomponent signals. *Biological Review* 79: 583–610.

Hasson O (1989) Amplifiers and the handicap principle in sexual selection: A different emphasis. *Proceedings of the Royal Society of London* 235: 383–406.

Hebets EA and Papaj DR (2005) Complex signal function: Developing a framework of testable hypotheses. *Behavioral Ecology and Sociobiology* 57: 197–214.

Johnstone RA (1996) Multiple displays in animal communication: 'Backup signals' and 'multiple messages'. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 351: 329–338.

King AJ and Calvert GA (2001) Multisensory integration: Perceptual grouping by eye and ear. *Current Biology* 11(8): 322–325.

Leger D (1993) Contextual sources of information and responses to animal communication signals. *Psychological Bulletin* 113(2): 295–304.

Meredith MA and Stein BE (1983) Interactions among converging sensory inputs in the superior colliculus. *Science* 221: 389–391.

O'Conner N and Hermelin B (1983) The role of general ability and specific talents in information processing. *British Journal of Developmental Psychology* 1: 389–403.

Partan SR (2004) Multisensory animal communication. In: Calvert G, Spence C, and Stein BE (eds.) *The Handbook of Multisensory Processes*, pp. 225–240. Cambridge, MA: MIT Press.

Partan SR and Marler P (1999) Communication goes multimodal. *Science* 283: 1272–1273.

Partan SR and Marler P (2005) Issues in the classification of multimodal communication signals. *American Naturalist* 166(2): 231–245.

Rowe C (1999) Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour* 58: 921–931.

Stein BE and Meredith MA (1993) *The Merging of the Senses*. Cambridge, MA: MIT Press.

Uetz GW and Roberts JA (2002) Multisensory cues and multimodal communication in spiders: Insights from video/audio playback studies. *Brain, Behavior, and Evolution* 59: 222–230.

Multiple Memory Systems

N M White, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

The term multiple memory systems (MMSs) refers to the idea that different types of information are processed and stored in different parts of the brain. It is an application of the concept of localization of function, based on the premise that processing and storing particular types of information are unique functions. Each of several series of connected brain structures forms parallel pathways or systems. Each system contains microstructure specialized to process a different type of information, represented by patterns of neural activity. As these activity patterns move through the systems, they are altered (processed) in various ways. The outputs of the systems influence thought and behavior. These ideas are illustrated and expanded in Figure 1.

In its weakest form the MMS concept is a convenient way to organize information about the localization of memory functions in the brain. In its strongest version MMS theory postulates that information is processed and memories of that information are stored in several parallel, independently functioning systems. In some part of each system, neuroplastic properties of the microstructure may allow information being processed to alter synaptic relationships. These changes constitute memories. Because of them, processing similar information on subsequent occasions results in altered output of the system, generating new (learned) thoughts and behaviors.

Although direct interactions among the systems are possible, the idea that they function independently implies that their outputs converge to influence behavior and thought. The most common form of interaction among the outputs is probably cooperation, when the information processed by each system makes a unique contribution to a complex memory or performance of a complicated task. The systems can also process information leading to different outcomes resulting in competition for control of behavior. These situations, which can be created in the laboratory, provide valuable insights into the operation and interaction of the systems.

Origin of the MMS Idea

The MMS idea was originally suggested by the study of individuals in whom damage to a particular part of the brain impairs one type of memory but not others. The most common case is damage to the medial temporal lobes, variously including parts of the hippocampus and the overlying parahippocampal cortex. These people are unable to remember most personal experiences that occurred prior to their brain damage or current events a few minutes after they occur. They are unable to find their way around even in familiar environments. However, the same people can learn new skills and procedures, and acquire and retain new emotional responses. An example is the patient HM, described by Milner, Penfield, and Scoville. Most of HM's medial temporal lobes had been surgically removed to relieve otherwise intractable epilepsy. After repeated practice HM became proficient at solving a short maze by moving a stylus on a tablet through a series of left and right turns. However, when asked to demonstrate this ability he invariably replied that he was willing to try but doubted he would be very good at it because he had never seen the task before. He then proceeded to solve the maze easily. This behavior pattern demonstrated HM's inability to recall his previous experience with the maze (memory normally dependent on the medial temporal lobe) and his simultaneous ability to remember the sequence of moves required to solve the maze. The memory for this sequence must have depended on some part of the brain other than the medial temporal lobes.

Types of Information/Memory

Declarative and Procedural Memory

The kind of information HM was unable to recall is usually called declarative memory because it can be described verbally. His intact memory for solving the maze is a type of skill, often called procedural memory.

Sherry and Schacter have suggested that procedural and declarative memories consist of functionally incompatible information. Procedural memories represent constant relationships among the elements of a situation. They are acquired by repeated exposure to such constancies, usually in situations where a specific stimulus comes to evoke a specific behavior pattern or response. These characteristics led Mishkin to suggest the term habit memory as a synonym for procedural memory. Procedural memory is prescriptive (e.g., the sight of a particular door invariably elicits the response of opening it). In contrast, most declarative memories are acquired as the result of a single experience and are descriptive in nature (the sight of the door elicits a memory of what to expect when it is opened). Descriptive information about the door can be combined with other information in a flexible way that could result in a decision to open or not to open the door. No such flexibility is possible

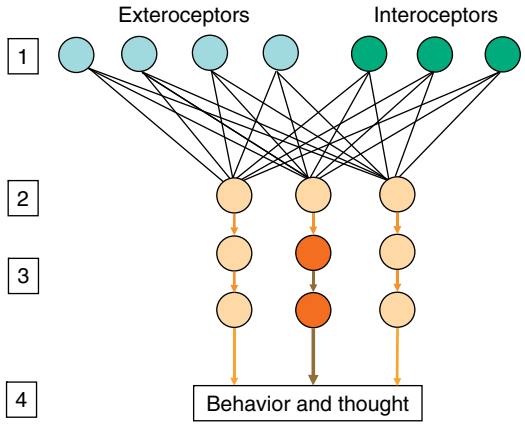

Figure 1 Diagram of MMSs concept: (1) Information enters the brain from various external and internal sources and is represented by patterns of neural activity. As the individual's situation changes, the activity patterns change. (2) The changing patterns converge on various brain areas and move, or flow, through several parallel series of interconnected structures (circles in the figure), known as systems. (3) Due to differences in internal microstructure each system is specialized to represent activity patterns consisting of different kinds of information. In this way, each system can represent different information about the same situation. When an activity pattern corresponds to the microstructure of a system the system responds in a coherent manner – indicated by the dark brown circles in the figure. As the activity patterns change the system responding most coherently may change. (4) The outputs of the system converge on brain structures that result in behavior and thought. The outputs may promote similar behaviors or parts of complex behaviors and thoughts (cooperation), or they may promote different behaviors (competition). The output of the system responding most coherently at any time exerts the strongest influence on thought and behavior.

in the procedural system where a specific stimulus always evokes the same specific response.

Because it is purely descriptive, declarative memory can be described as a series of associations among stimuli (S–S associations). Procedural memory has been described as consisting of a series of unique stimulus–response (S–R) associations.

Emotional Memory

Emotions and internal motivational states result from patterns of activity in the autonomic nervous and endocrine systems, and in certain brain structures. These activity patterns are innate, unconditioned responses (URs) to specific kinds of stimuli – an approaching speeding car or the smell of certain foods (even if they have not previously been tasted) – called unconditioned stimuli (US) or reinforcers. A neutral stimulus present at the same time as a US becomes a conditioned stimulus (CS) which can, in the absence of the US, elicit a conditioned response (CR) that is similar to the UR. This is Pavlovian conditioning, also described as stimulus–reinforcer (S–Rf) learning.

Although emotional states can be described and act to influence thought and behavior, no declarative memory of the relationship between a CS and the conditioned emotional state it elicits is required for this form of memory. Emotional memories are limited to USs and the URs they elicit innately, and do not require the performance of any other response. In contrast, procedural memories can be formed from associations between any stimulus that can be perceived and any response that can be performed.

Anatomical Basis of Memory Types

The preponderance of evidence is that damage to the medial temporal lobes (including either the hippocampus or the fimbria-fornix, a major input–output pathway of the hippocampus) or the medial thalamus impairs declarative memories but leaves procedural and emotional memory functions intact. Procedural memories are impaired by damage to the basal ganglia (including the caudate nucleus), which does not impair declarative or emotional memories. Damage to the amygdala impairs emotional memories with no effect on procedural or declarative memory.

These dissociations are strong evidence for localized processing of different information types. In this experimental design (illustrated in Figure 2) the effects of damage to different brain structures (usually two) on an equal number of memory tasks are compared. A dissociation is demonstrated when performance on each task is impaired by damage to a different structure and damage to each structure affects only one of the tasks.

Each of the structures implicated in one of the three types of memory has extensive efferent and afferent connections that allow information to flow through it. Each structure, together with its efferents and afferents constitutes a system. Evidence dissociating the information types processed in each system form the basis of the MMS idea.

Dissociations in Human Memory Functions

Declarative versus Procedural Memory

Especially in their early stages, certain pathologies selectively affect specific brain structures. Alzheimer's disease primarily impairs hippocampal function; Korsakoff's syndrome affects the medial thalamus. Patients with either of these pathologies suffer severe declarative memory loss and are referred to as amnesic. Parkinson's and Huntington's diseases primarily affect the function of the basal ganglia, including the caudate nucleus.

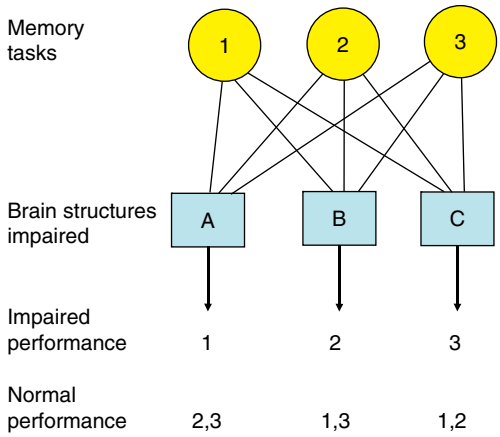

Figure 2 Dissociation paradigm. In this experimental paradigm, subjects are given several different behavioral tests (1–3 in the diagram) for their ability to learn and recall specific types of information. Different subjects or groups of subjects have impairments of specific brain areas (A–C in the diagram) due to accidental trauma, disease, or surgery in humans or experimental manipulations in animals. Each of the brain lesions impairs performance on one of the tasks but leaves the others intact. Demonstrations of this type are evidence that the brain area required to process the information necessary for performing each of the tasks is not required for the other tasks. This shows that the information processing functions for each of the tasks occur in different parts of the brain. The conclusion is that A, B, and C are parts of systems that function independently to process different kinds of information. Most dissociations involve two tasks and two brain areas (double dissociation), but McDonald and White have described a triple dissociation of this kind among memory types and brain systems.

Mirror reading Cohen and Squire found that with practice, amnesic patients could learn to read fluently complex words seen in a mirror (a skill or procedural memory) but were subsequently unable to recall any of the words they had read (declarative memory). This pattern of deficits was replicated in Korsakoff's patients by Heindel, Butters and colleagues. They also showed that Huntington's patients were unable to improve their mirror reading skill but could remember the words they had read. This double dissociation localizes declarative (S–S) memory to the hippocampus and medial thalamus, and procedural (S–R) memory to the basal ganglia. The dissociation also shows that both systems can process and store these types of information independently of the other system (although it remains possible that both structures could contribute to both behaviors in normal individuals).

Weather prediction task In the weather prediction task (Knowlton, Mangels, and Squire) subjects are shown displays of four cards at a time selected from a larger set, each printed with a unique pattern. The cards predict either of two outcomes, rain or shine, with a probability of less than 100%. Participants indicate their responses by pressing either of two buttons and receive feedback as to whether or not their response was correct. The probabilistic association between each stimulus and its correct response provides sufficient repetition to produce procedural (S–R) learning while retarding declarative (S–S) learning about the correct responses. Parkinson's patients are severely impaired on the weather prediction task while amnesiacs perform almost as well as normal controls. Notwithstanding their near normal weather prediction, amnesic patients are unable to answer questions about the testing experience (declarative memory); the Parkinson's patients' memory of the testing experience is normal. These findings are a double dissociation of declarative memory (in the amnesiacs with impaired function of medial temporal lobe and/or medial thalamus) and procedural memory (in the Parkinson's patients with impaired function of basal ganglia).

Brain imaging with functional magnetic resonance imaging (fMRI) while doing the weather prediction task (Poldrack and colleagues) revealed increased activity in the basal ganglia (caudate nucleus) and decreased activity in the medial temporal lobe. When the same information was learned by presenting the cards together with the outcome they predicted (a form of paired associate declarative learning known to depend on the medial temporal lobe), activity in the hippocampus increased while activity in the caudate nucleus decreased. This double dissociation is consistent with the evidence from brain-damaged patients.

Radial maze In another imaging study (Iaria, Bohbot, and colleagues) normal participants searched for virtual rewards at the ends of the arms of a computer-generated radial maze (a circular center area with eight arms radiating from it in a sunburst pattern). The maze was surrounded by a landscape with mountains, trees, and the sun. Since rewards obtained were not replaced, participants had to remember which maze arms they had visited to avoid reentries and obtain the rewards with the fewest possible errors. After performing the task, about half of the participants said they had used the landmarks to remember and avoid previously visited arms (a spatial strategy). Hippocampal activity was increased in these participants. The other half said they had arbitrarily labeled or counted the arms (a nonspatial strategy). These participants showed increased activity in the caudate nucleus. A subgroup that started out using a spatial strategy reported that they switched to a nonspatial strategy during training; their patterns of brain activation shifted from hippocampal to caudate activation in parallel with their reported strategy switch.

Declarative versus Emotional Memory

Bechara and Damasio compared a patient with Urbach–Wierthe disease, which produces brain damage confined to the amygdala complex, to a patient with hippocampal damage due to anoxia. The patients were shown slides of different colors, one of which (the CS) was followed by the sound of a very loud air horn (US). In addition to a startle response, the horn produced increased skin conductance (UR), a standard measure of aversive emotional state. When the CS slide was shown without the horn it elicited increased skin conductance (a CR) in normal participants, an instance of Pavlovian conditioning. The patient with amygdala damage did not exhibit this CR but was subsequently able to recall most details of the experimental experience (declarative memory). In contrast, the hippocampal patient showed normal conditioning, but was unable to recall most experimental details.

Johnsrude and colleagues compared surgical patients with unilateral amygdala resection to a group with unilateral frontal cortex resection. The patients were exposed to a series of patterns on a computer screen while performing a highly distracting counting task. The patterns were accompanied by a reward (a raisin or small candy) on either 90% or 10% of their appearances. After training, the participants were asked for the result of the counting task (declarative memory) and were then shown pairs of the patterns and asked which one they preferred (Pavlovian conditioning).

Normal participants performed the counting task accurately and preferred the patterns that had been paired with the reward 90% of the time over those that had been paired 10% of the time, a conditioned preference. Patients with amygdala damage performed the counting task accurately but did not exhibit the conditioned preference. In contrast, patients with frontal lobe damage were unable to perform the counting task but exhibited normal conditioned preferences.

These studies dissociate the processing of information required for amygdala-based aversive and appetitive Pavlovian conditioning (S–Rf memory) from the processing of declarative (S–S) information.

Summary of Dissociations in Humans

Figure 3 illustrates the anatomical dissociations of the different types of information processing described. Some of the justification for this way of organizing the data derives from experiments with rats, reviewed next.

Dissociations of Memory Systems in Rats

Although animals cannot verbalize recalled information, it is possible to show that they can learn and recall the same three types of information as is suggested by the human studies, and that these types are processed in neural systems consisting of the same or homologous brain structures.

Declarative (S–S) versus Procedural (S–R) Memory

Cross maze An experimental paradigm for dissociating declarative and procedural memory in rats, introduced by Tolman in 1946, makes use of a

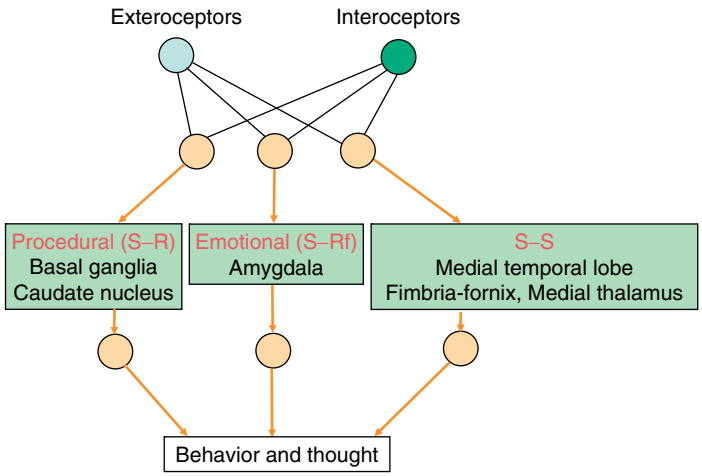

Figure 3 Dissociations of information types and brain structures. The figure is based on evidence that performance on tasks requiring memory for three different kinds of information (S–S, S–R, and S–Rf) is impaired by lesions of the medial temporal lobes, basal ganglia, and amygdala, respectively, and that none of these lesions affects processing of the other information types. Imaging data also dissociate activity levels in hippocampus and caudate nucleus with processing declarative and procedural information, respectively. The figure presents this information in the context of the MMS concept: the idea that the three kinds of information are processed in different, parallel neural systems. The figure shows the systems interacting only when their outputs converge to influence behavior and thought.

cross-shaped maze, described in Figure 4. Using this paradigm, Packard and McGaugh found that when tested after 8 days of training, most rats ran to the correct spatial location of the food, even though this required them to make a turn opposite to the one which had been reinforced during training. This is an example of the flexible use of information by the declarative system. After 8 more days of training, most of the same rats made the same turn at the choice point that had been reinforced during training, even though this response led away from the food. This is an example of an inflexible S–R relationship of the kind mediated by the procedural system. This shift from the control of behavior by declarative information to control by procedural information is similar to the shift reported by humans on Iaria et al.'s virtual radial maze task, many of whom switched from using spatial information to nonspatial methods during the course of training.

In other groups of rats, Packard and McGaugh temporarily inactivated either the hippocampus or the caudate nucleus with lidocaine during the test trials. The results (explained in detail in Figure 5) revealed that the tendency to turn toward the food on the early test trial was due to hippocampus-based processing of spatial information, and the tendency to turn away from the food on the later trial was due to caudate-based processing of S–R information. These results illustrate several features of hippocampus-based declarative and caudate-based procedural memory and their interaction. (1) Declarative memory can be used flexibly, depending on the current context, to generate behavior leading to an outcome; procedural memory is inflexible: a stimulus (the choice point in the maze) always elicits the same response regardless of the outcome. (2) Declarative memory is acquired and stored relatively quickly; procedural memory requires more repeated trials for an S–R association to build up sufficient strength to influence behavior. (3) The outputs of the declarative and procedural memory systems compete with each other for control of behavior. Such evidence of competition between behaviors produced by different kinds of information mediated in different brain systems is evidence that the systems acquire and process information independently (see Figure 5).

Water maze Morris and colleagues described a task for studying spatial learning using a large circular pool of opaque water situated in a room with a variety of cues. Rats swim around until they encounter a small platform, invisible below the surface, which allows them to escape from the water. Over repeated trials, the rats quickly learn to swim directly to the platform even when they are placed into the pool at different points on the perimeter. Since no local cues indicate the platform location it must be learned in relation to the spatial cues in the pool environment. This instance of spatial learning can be contrasted with learning to swim to a platform that is visible above the surface of the water. The presence of a local cue makes this a form of S–R learning.

McDonald and White trained rats to find a platform that remained in a constant location on training trials when it was visible and when it was submerged. On a final test trial the visible platform was moved to a new location (Figure 6). Rats with fimbria-fornix lesions tended to swim directly to the visible platform;

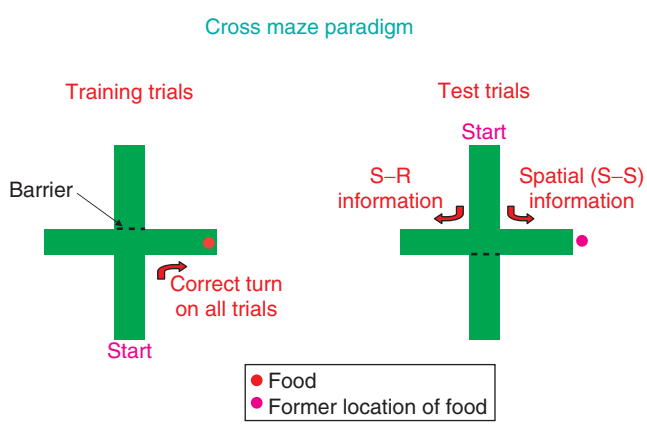

Figure 4 The cross maze task. On training trials (left) rats are placed at the end of one of the arms. When they reach the choice point, the barrier forces them to turn either left or right. Food is always placed at the end of the arm on the right. After repeated training trials, during which rats learn to turn right, a test trial (right) is given: the rats are placed at the opposite end of the start arm which is again blocked, forcing them to turn left or right at the choice point. The direction of the turn reveals the nature of the processed and stored information that produced it. A right turn, the same response as the one made during training – but now leading away from the food – is due to procedural or S–R learning. A left turn – opposite to the one made during training but leading toward the food – is due to spatial learning, a subtype of declarative or S–S memory.

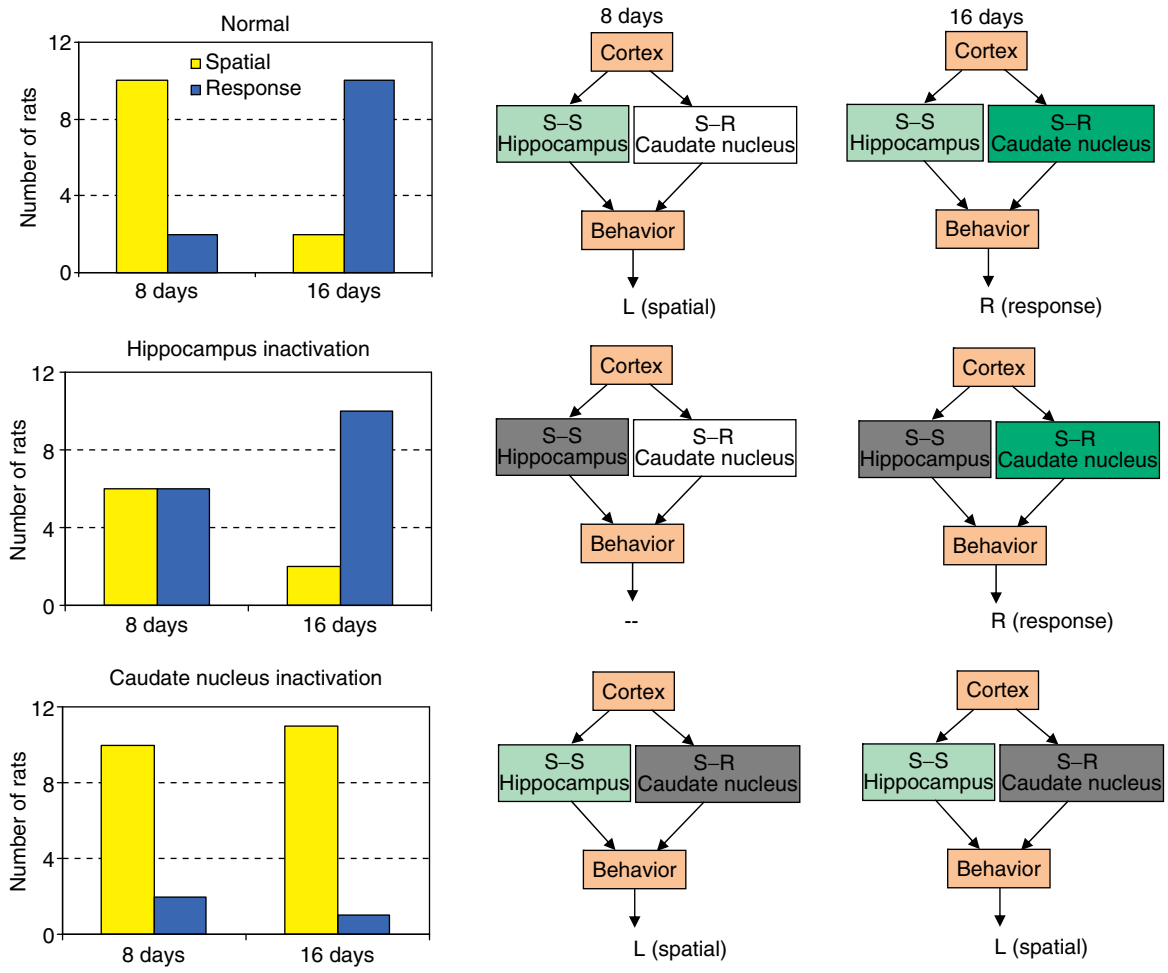

Figure 5 Results and explanation of cross maze experiment: Top row: normal rats. The graph shows the number of rats that turned in the direction of the food indicating use of spatial information, or away from the food indicating use of S–R information on the two test trials. After 8 days of training most rats turned toward the food due to hippocampus-based processing of spatial information (light green). At this time insufficient caudate-based S–R learning had occurred to influence the response (white). After 16 training days, S–R learning had become strong enough for the output of the caudate system (dark green) to compete with the hippocampus-based response and take control over the rats' behavior, causing them to make the response acquired during training, leading away from the food. Middle row: inactivation of the hippocampus during the test trials. After 8 days hippocampal inactivation (gray) produced random responding, as predicted by the hypothesis that turning toward the food was due to hippocampus-based spatial information processing. After 16 days hippocampal inactivation (gray) did not affect the tendency to turn away from the food, consistent with the hypothesis that the hippocampus was not involved in this S–R type behavior. Bottom row: inactivation of the caudate nucleus during the test trials. After 8 days caudate inactivation (gray) had no effect, as predicted by the idea that this structure was not involved in processing the spatial information that produced the turn toward the food. After 16 days caudate inactivation (gray) resulted in turning toward the food, reversing the behavior of normal rats, which turned away from the food at this point (top row). This is predicted by the hypothesis that the caudate nucleus mediated the S–R information that produced the turn away from the food in the normal rats. It also shows that the hippocampus-based spatial information was still present and able to influence behavior, but that it was prevented from doing so in normal rats by competition from the output of caudate-based S–R information. Eliminating the caudate output allowed the spatial information to resume its influence, allowing the rats to turn toward the food. Data from Packard MG and McGaugh JL (1996) Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory* 65: 66.

rats with caudate lesions tended to swim to the former location of the invisible platform first; upon failing to find it there they then swam to the visible platform.

These results (explained in detail in Figure 6) show that two kinds of information competed for control of the rats' behavior on the test trial. Eliminating spatial information processing with a fimbria-fornix lesion resulted in expression of S–R memory (swimming directly to the visible platform without regard to its new location), and eliminating S–R processing with caudate lesions resulted in swimming to the previously learned platform location notwithstanding that the platform was visible in a different location.

Analysis of the behavior of the control rats (explained in Figure 7) showed that at the start of training half of them relied on spatial learning to find

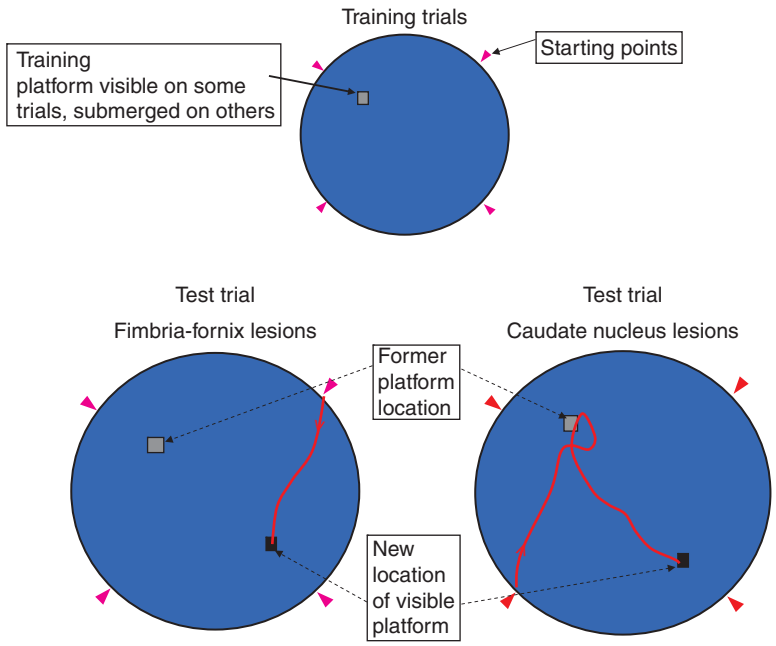

Figure 6 Competition between visible and invisible platform in the water maze. The blue circles are top views of a circular swimming pool about 1.5 m in diameter. Rats were placed into the pool once at each of the four starting points on each training day. The platform was always in the same location; it was visible on some training days and submerged (invisible) on others. On the test trial the rats were placed into the pool with the visible platform in a new location. In this situation, spatial learning about the location of the platform competed with the S–R tendency to swim to the visible stimulus. As shown by the typical swim path (red line) of a rat with fimbria-fornix lesions on the bottom left, the inability to process spatial information resulted in a tendency to swim directly to the visible platform due to the expression of caudate-based processing of S–R information. The bottom right shows a typical swim path for a rat with caudate nucleus lesions that could not process S–R information. This rat swam directly to the old spatial location of the platform due to the expression of hippocampus-based spatial information. After failing to find the platform in the old location the rat swam to the visible platform. Adapted from McDonald RJ and White NM (1994) Parallel information processing in the water maze. *Behavioral and Neural Biology* 61: 260.

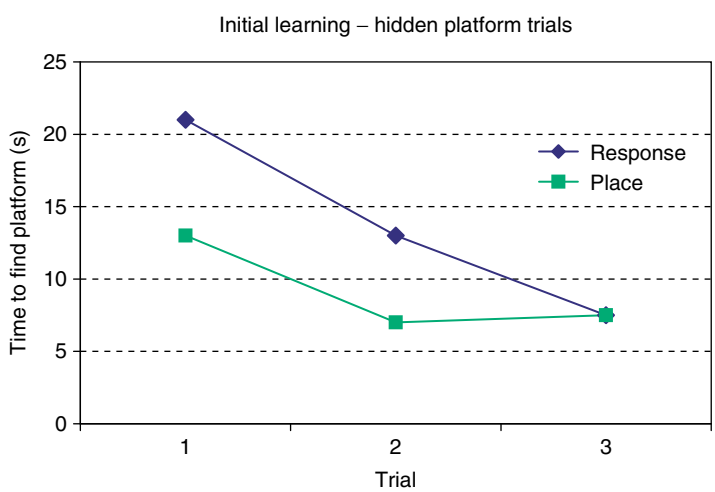

Figure 7 Performance of normal rats grouped according to behavior on the competition test in the water maze. Of eight normal control rats, four swam directly to the visible platform (like the rats with fimbria-fornix lesions and intact caudate nucleus) and four swam to the old platform location first (like the rats with caudate lesions and intact hippocampus system) on the competition test (**Figure 6**). This figure shows the mean time required to find the hidden platform after each of three successive 3-day blocks of training with a visible platform in the same location. Learning the location of a hidden platform requires spatial information processing. The group that used spatial information on the competition test learned the spatial location of the hidden platform during the first three visible platform trials, but the group that used S–R information on the competition test required considerably more training to learn its spatial location. Adapted from McDonald RJ and White NM (1994) Parallel information processing in the water maze. *Behavioral and Neural Biology* 61: 260.

the maze while the other half acquired an S–R association. This parallels the finding of Iaria et al. that about half of humans solving a virtual radial maze initially used each of the two kinds of information and suggests that there may be constitutional differences in the relative influence of the two kinds of information on individual behavior.

Radial maze The radial maze task was introduced by Olton and colleagues in 1975 as a measure of working memory for spatial information. The maze is illustrated in Figure 8. A food pellet is placed at the end of each arm. A hungry rat is placed on the center platform and allowed to forage freely for the available food. Since consumed pellets are not replaced, the rat should enter each arm only once (leading to the description of the behavior as win-shift).

This task has a number of features in common with declarative memory. Entering each arm only once requires a memory of which arms have been entered

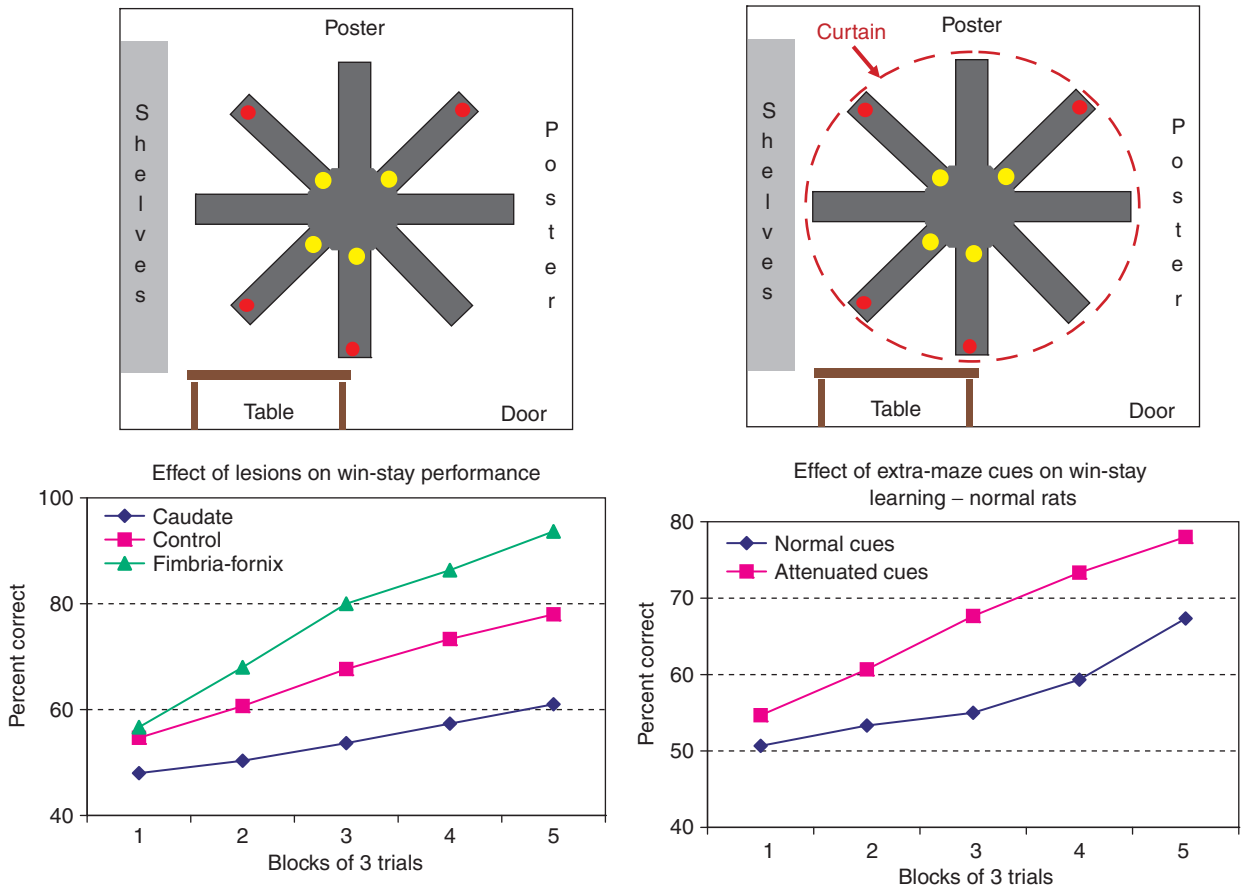

Figure 8 Effects of lesions and attenuation of extra-maze cues on win-stay performance. The top part of the figure shows an eight-arm radial maze in a room with extra-maze cues. On each daily trial, four of the eight maze arms have lights at their entrances (yellow dots) and only those four arms contain food (red dots). These four arms change on each daily trial. When a rat obtains the food pellet from one of the arms it is replaced. When the rat obtains the second pellet from an arm the light at the entrance is extinguished and no further food is placed in that arm. The bottom left shows the number of lit arms entered as a percent of total arms entered while retrieving the eight pellets. The graph shows that caudate nucleus lesions impaired win-stay performance relative to the normal controls, but that rats with fimbria-fornix lesions learned the task faster and attained a higher level of performance than normal rats. The superior performance of the rats with fimbria-fornix lesions can be explained by assuming that the caudate and fimbria-fornix are parts of independent information processing systems that acquire different kinds of information while the rat performs the task. Since win-stay performance is impaired by lesions of the caudate nucleus, that structure is assumed to be part of a system that processes the S–R information required to perform the task correctly. At the same time, the medial temporal lobe system, of which the fimbria-fornix is a part, processes information about the spatial locations of the food in the maze environment. However, since the spatial locations of the food change on each trial, information about food location acquired on any given trial and recalled on a subsequent trial is likely to be wrong and result in errors. Eliminating this source of errors by impairing the function of the hippocampal system with fimbria-fornix lesions therefore improves performance. The top right of the figure shows the maze surrounded by a curtain which attenuates extra-maze cues, retarding spatial learning with respect to those cues. The graph on the bottom right shows that rats trained with attenuated extra-maze cues learn the win-stay task faster and perform better than rats tested with normal exposure to the cues. This demonstration is consistent with the explanation for the facilitation of win-stay learning by fimbria-fornix lesions. Both the curtain and the lesions impair spatial learning, eliminating its interference with correct responding based on S–R learning. Adapted from Packard (1987), unpublished Master's thesis, McGill University.

as the trial progresses. The rat must therefore be able to identify (discriminate among) the arms. With appropriate controls this can only be done by learning the locations of the visited arms in relation to the cues visible in the maze environment. The rat must use information about arms entered to label each visited location so it can be avoided for the remainder of the trial. This information about each entered arm is acquired in a single experience without repetition and is different on every trial. Note that consumption of the food at the ends of the arms does not function as a reinforcer, which would make reentries more, not less, likely. Instead, information about which arms no longer contain food may contribute to the tendency to avoid those arms as the trial progresses.

Packard, Hirsh, and White used this apparatus to compare declarative (win-shift) learning and S–R learning using the win-stay task (see Figure 8). Since the presence of food at the ends of the arms in the win-stay task is always indicated by lights at the arm entrances spatial learning is not required to discriminate among the arms or to remember which arms have been entered. Instead, the rats repeatedly enter lit arms and each entry into any such arm is followed by food functioning as a reinforcer by increasing the probability of entering lit arms.

Rats with lesions of the fimbria-fornix or hippocampus are impaired on the win-shift but not on the win-stay task. In contrast, rats with lesions of the caudate nucleus are impaired on win-stay, but not on win-shift. This double dissociation of the neural substrates of information processing required to perform these declarative and procedural tasks is consistent with the other human and animal evidence described.

Figure 8 shows the impaired win-stay performance of the rats with caudate lesions and also shows that the rats with fimbria-fornix lesions acquired this task more quickly and reached a higher level of performance than normal rats. This observation is another demonstration of competition between caudate-based S–R learning and hippocampus-based spatial learning, explained in detail in Figure 8.

These radial maze findings are further evidence for the idea that different parts of the brain process and may also store declarative and S–R information independently, and that outputs of these systems compete for control of behavior.

Declarative (S–S) versus Emotional (S–Rf) Memory

Radial maze White and McDonald used the radial maze to study Pavlovian conditioning of emotional memory with the conditioned cue preference (CCP) task, explained in Figure 9. Hungry rats are confined in one of the maze arms with a supply of food. Eating the food (US) elicits internal responses perceived as reward (UR). The neutral cues visible from the arm acquire CS properties by becoming associated with the UR. On the test trial the rats are given a choice between the food-paired and another arm, with no food in either arm. Normal rats spend more time in the food-paired arm than in the unpaired arm. This is due to conditioned reward elicited by the conditioned cues visible from the food-paired arm.

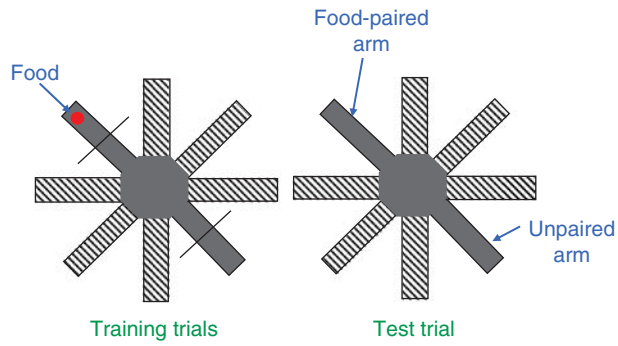

Figure 9 The CCP task. On the training trials (left) rats are confined on one arm of the maze with a supply of food, and on a nonadjacent arm with no food for 30 min, on alternate days. One or more of these 2-day training trials are given. On a subsequent test trial (right) the rats are allowed to forage for food on the maze with access to both arms, neither of which contains food. Normal rats spend more time in the food-paired arm than in the unpaired arm. With appropriate controls this preference can be attributed to the formation of an association between the extra-maze cues visible from the food-paired arm and the effects of consuming the food, an S–Rf association. Since the cues visible from the food-paired arm cannot be seen from the no-food arm during training this arm discrimination does not require spatial information. Since the rats consume the food while confined at the end of an arm no specific response is associated with the food-paired stimuli. It must therefore be attributed to Pavlovian conditioning: a rewarding internal response (UR) elicited by the food (US) is associated with the cues visible from the food-paired arm (CS). Elicitation of the CR on the test trial causes the rats to enter and remain in the empty food-paired arm.

The CCP is impaired by lesions of the amygdala, but unaffected by lesions of the caudate nucleus, fimbria-fornix, or hippocampus. In contrast, amygdala lesions do not affect win-stay or win-shift learning (a triple dissociation – see Figure 2). Accordingly the S–Rf information required for the CCP is processed in a different neural system from those that mediate the S–R and S–S information processing required for the win-stay and win-shift tasks.

Emotional (S–Rf) versus Procedural (S–R) Memory

Radial maze McDonald and Hong trained rats on the win-stay task and then gave them a preference test on the same maze with no food available. All eight

arms of the maze were open and four of them were lit. The rats spent more time in lit than in dark arms, showing that they acquired a CCP during win-stay training. When tested with the same procedure, rats with amygdala lesions were normal on win-shift learning but did not express CCPs; rats with dorsal striatum lesions were impaired on win-shift learning but expressed normal CCPs. This double dissociation of caudate-based S–R learning and amygdala-based S–Rf learning shows that the information leading to the two types of behavior is acquired and expressed independently by different neural substrates.

Summary

The data from both humans and rats show that, consistent with the concept of functional localization, processing each of three kinds of information is impaired by damage to a different neural structure. In humans, imaging reveals complimentary information relating the kinds of information being processed and activated brain structures. These relationships are summarized in Figure 10. The figure organizes the data by suggesting the existence of three independent neural systems, each of which is specialized to process a different kind of information. The outputs of these processes influence thought and behavior.

The dissociations that provide the evidence for these ideas are based on specific learning tasks that depend on information processing in single systems. Most normally occurring behaviors undoubtedly depend on information processing in more than one of the systems and on cooperation among them at the output level. Competition among systems may also occur normally, possibly accounting for phenomena of indecision, hesitation and 'mind changing.'

Although considerable evidence for the idea that different kinds of information are processed in different brain areas has been discussed, no direct evidence has been presented that the memories for these information types are actually stored in these systems. While it is possible that this is the case, it is also possible that memory storage for all types of information is stored in a single location (e.g., the cortex) and that the systems selectively utilize cortically stored information. Given evidence that short-term memories may be stored in subcortical systems but that memories retained for months or years may be stored in the cortex, both of these hypotheses remain possible.

Most of the evidence for the MMS idea comes from the study of just three central brain structures, the hippocampus (or fimbria-fornix), the amygdala, and the basal ganglia (caudate nucleus). In each case, these central structures together with their afferents and efferents must constitute systems. Further research will define the anatomical divergence and convergence of the neural pathways (known as memory systems) that process and store different types of information.

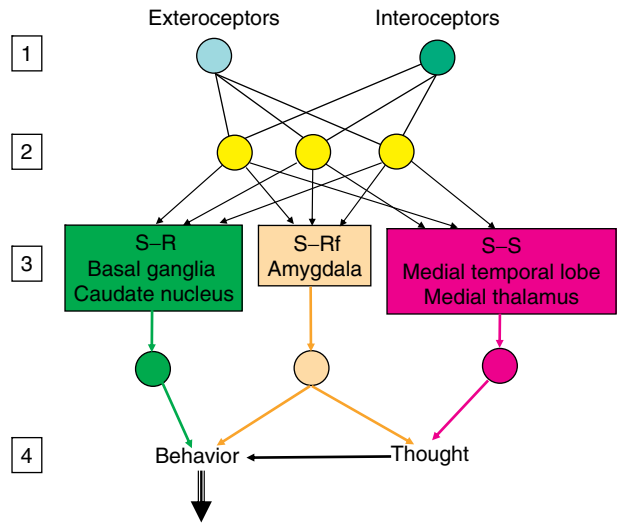

Figure 10 How multiple memory systems influence thought and behavior. (1) As shown in Figure 2, patterns of neural activity originate in exteroceptors and interoceptors. (2) These patterns converge on sensory thalamic and cortical areas where they are processed and combined into activity patters that represent the current external and internal environments. (3) These activity patterns are transmitted to subcortical areas, each of which is activated by a specific set of relationships among the elements represented: basal ganglia are activated by S–R relationships, the amygdala is activated by S–Rf relationships, and the medial-temporal lobe system is activated by stimulus–stimulus (S–S) relationships. Most situations activate each of the systems to a different degree. (4) The outputs of the systems converge in various ways on brain areas that mediate behavior and thought. Basal ganglia-mediated S–R output influences behavior in an inflexible manner. Amygdala-mediated S–Rf output can influence behavior directly by producing conditioned approach or withdrawal responses, and can produce conditioned emotional states whose perception can influence behavior by instrumental learning. Medial-temporal lobe mediated S–S learning results in knowledge which can be used flexibly in combination with current affective and sensory information to produce behavior.

See also: Amygdala: Structure and Circuitry in Primates; Cognition: Basal Ganglia Role; Declarative Memory System: Anatomy; Dementia; Emotional Hormones and Memory Modulation; Hippocampus and Neural Representations; Memory Disorders; Procedural Learning: Classical Conditioning; Reward Systems: Human; Spatial Cognition and Executive Function.

Further Reading

Hartley T and Burgess N (2005) Complementary memory systems: Competition, cooperation and compensation. *Trends in Neurosciences* 28: 169–170.

McDonald RJ, Devan BD, and Hong NS (2004) Multiple memory systems: The power of interactions. *Neurobiology of Learning and Memory* 82: 333–346.

McDonald RJ, Hong NS, and Devan BD (2004) The challenges of understanding mammalian cognition and memory-based behaviours: An interactive learning and memory systems approach. *Neuroscience and Biobehavioural Reviews* 28: 719–745.

McDonald RJ and White NM (1993) A triple dissociation of memory systems: Hippocampus, amygdala and dorsal striatum. *Behavioral Neuroscience* 107: 3–22.

McDonald RJ and White NM (1994) Parallel information processing in the water maze. *Behavioral and Neural Biology* 61: 260.

Milner B, Corkin S, and Teuber H-L (1968) Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of H.M. *Neuropsychologia* 6: 215–234.

Packard MG and McGaugh JL (1996) Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory* 65: 66.

Poldrack RA and Packard MG (2003) Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia* 41: 245–251.

Poldrack RA and Rodriguez P (2004) How do memory systems interact? Evidence from human classification learning. *Neurobiology of Learning and Memory* 82: 324–332.

Sherry DF and Schacter DL (1987) The evolution of multiple memory systems. *Psychological Review* 94: 439–454.

White NM (2004) The role of stimulus ambiguity and movement in spatial navigation: A multiple memory systems analysis of location discrimination. *Neurobiology of Learning and Memory* 82: 216–229.

White NM and McDonald RJ (2002) Multiple parallel memory systems in the brain of the rat. *Neurobiology of Learning and Memory* 77: 125–184.

Multisensory Convergence and Integration

B E Stein, B Rowland, P Laurienti, and T R Stanford, Wake Forest University School of Medicine, Winston-Salem, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Interest in the neurobiology of multisensory integration extends well beyond the biomedical sciences. Engineers have long been interested in producing an effective sensor-fusion device based on biological principles for such purposes as air traffic control, detecting the presence of obstacles under the sea, searching the ocean floor for unexploded mines, and hundreds of other purposes that depend on the remote sensing of objects and/or events. This engineering interest was sparked by inherent problems in using single-sensor devices, which have been effectively minimized by biological organisms through the process of multisensory integration. Consider that the effectiveness of commonly used single-sensor devices such as radar or sonar is greatest when the event to be detected provides a clear signal whose signature is readily recognizable. If the signal is weak, ambiguous, and/or embedded in noise, it can be very difficult to detect and/or disambiguate. Simply lowering the threshold of the sensor in an attempt to detect weak signals is not a viable solution because it makes the systems increasingly vulnerable to false positives. However, this problem can be mitigated by coupling different sensors (e.g., for light and sound), each tuned to detect events with specific features of interest, to a central processor capable of signaling their synchronous activation. The probability that anything other than the event of interest would mimic its signature in multiple-sensor domains is comparatively low.

Fusing or, more accurately, integrating signals that are derived from different sensory systems is a ubiquitous phenomenon in extant biological organisms. Its utility is evident regardless of how elaborate and specialized individual senses become; pooling information from them enhances not only the likelihood of detecting a given environmental event but also of identifying it. Because each sense is tuned to a different type of environmental energy, events that cannot readily be unambiguously identified along any single sensory dimension (e.g., how they look, feel, or sound) may be identified through a combination of sensory dimensions.

Converging Channels of Sensory Information

There are many areas of the brain that receive inputs from multiple-sensory systems. Best known among these is a midbrain structure known as the superior colliculus (SC), as well as a host of higher-order areas of cortex, sometimes referred to as association cortices. These areas contain mixtures of unisensory neurons, representing each of the structure's input modalities and each of the possible modality convergence patterns. Most of these multisensory neurons are capable of synthesizing their convergent cross-modal information and of producing a signal that is quite different from those evoked by either of the individual inputs. The product signal can be enhanced or degraded. Signal enhancement is generally the result of the synthesis of input signals that are initiated by the same events because these inputs are closely linked in space and time and thus meet the neural specifications for enhancement (discussed later). From an operational perspective, these cues are coherent or meaningful and their enhancement makes them more salient. When the signals are not linked in space and/or time, as is often the case when they are derived from different events, they are either not integrated or are depressed, and thereby rendered less salient. The benefits of such integrated products are obvious in determining which events have the highest priority in gaining an animal's immediate attention and influencing behavior. The process also adds depth to our sensory experiences and provides nuanced perceptions that could not be achieved if the sensory channels always operated independently of one another.

Illusions Illustrate the Potency of the Multisensory Product

The potency of multisensory integration is particularly evident in circumstances in which small discrepancies disrupt the normal spatial and/or temporal relationships among cross-modal stimuli. A good example is the Ventriloquism Effect, in which we think we hear the entertainer's voice coming from the dummy's mouth. This effect is really a general phenomenon that we experience regularly when watching a movie: the sounds all come from the same place (the acoustical speakers) but seem to be linked to the location of specific events or characters on the screen. The illusion depends on the brain preferring visual over auditory information for the purpose of localization. There are many such illusions, all of which represent the brain's effort to properly match cross-modal cues with one another in order to best evaluate an event.

Multisensory Integration as a Statistically Optimal Inference

As already noted, sensory information is frequently ambiguous, and thus environmental events cannot

be detected and identified with complete certainty. Bayes's theory provides a model for how a statistically optimal inference about an environmental event (e.g., its location) can be made based on ambiguous sensory information and how multiple sensory streams can be combined to make the best possible use of all of the available information. Human performance on tasks requiring visual-auditory and visual-tactile (or haptic) integration is consistent with that predicted by a Bayesian model, giving credence to the idea that organisms integrate sensory information in a statistically optimal fashion. Furthermore, Bayesian models correctly predict the neurophysiological response properties of multisensory neurons in areas of the brain such as the SC. Thus, Bayesian models may offer more than a yardstick of statistically optimal integration. The computations internal to Bayesian models may provide insights into how information is actually integrated in the brain and point to likely underlying neural mechanisms.

The Superior Colliculus Neuron: A Model of Multisensory Integration

The SC neuron has been used most extensively as a model to understand the neural processes underlying multisensory integration. This structure is involved in orientation toward targets in contralateral space and uses incoming visual, auditory, and somatosensory information to do so. These inputs adhere to a similar maplike pattern to ensure that the different sensory inputs to a given SC neuron provide information from common areas of sensory space. As a result, the different receptive fields of an individual multisensory neuron overlap one another, as shown in Figure 1(d). This pattern of spatial register is found in multisensory neurons throughout the nervous system.

Although the SC's best-known role involves initiating and controlling gaze shifts (i.e., eye and head movement) to objects or events in contralateral sensory space, it is actually involved in coordinated orientation of the various sensory organs (e.g., eyes, head, ears, limbs, whiskers, and mouth) that best position the animal to both assess and react to the stimuli of interest. It plays a role in these orientation responses via its output projections to those areas in the brain stem and spinal cord that are in more direct contact with the muscles. Anatomically, the SC is situated in the midbrain (Figure 1(a)), where it receives visual, auditory, and somatosensory inputs from a variety of sources. Many of these inputs converge on the same neurons, thereby rendering them multisensory. Despite the variety of afferent sources and their ultimate convergence patterns, there is a characteristic constancy to these projections – they are topographically ordered. In each case, the different inputs to a common SC target neuron provide information from common areas of sensory space. This pattern of topographic fidelity is not unique to the SC but, rather, is characteristic of multisensory neurons throughout the nervous system.

A Multisensory Map of Sensory Space

Organizing the receptive fields throughout the SC into overlapping visual, auditory, and somatosensory maps creates a multisensory map of contralateral sensory space, with the central-peripheral axis (or head to rump) laid out from front to back in the SC and the superior–inferior axis laid out from medial to lateral in the structure. This multisensory map is in register with the motor representation used to orient the eyes, ears, and head. However, unlike the organization found in the primary sensory projection pathways, whose maps are referenced to the relevant peripheral sensory organ (eyes, ears, head, etc.) and which are more involved in processing sensory-specific information, the SC maps have been transformed from their individual modality-specific frame of reference to a common motor frame of reference. In primates and carnivores, for example, the coordinate frame in the SC is linked to the direction of gaze. By making such a transformation, the brain has made sure that, regardless of the positions of the peripheral sensory organs with respect to one another, the activity initiated by any sensory cue in the SC represents its position relative to the current line of sight. This ensures that the site of activity in the SC is always that which initiates a gaze shift to look at the target.

Multisensory Integration: The Physiology of Response Enhancement and Response Depression

The principal job of multisensory neurons, regardless of where in the brain they are located, is to pool and integrate information carried by the different senses. This is evident when two different sensory inputs, such as sound and touch, produce a response that is significantly greater or less than the response produced by either of these individual stimuli. Indeed, this is the definition of multisensory integration and is an effect often noted in individual SC neurons. As previously noted, one of the critical factors in this process is the positions of the cross-modal stimuli

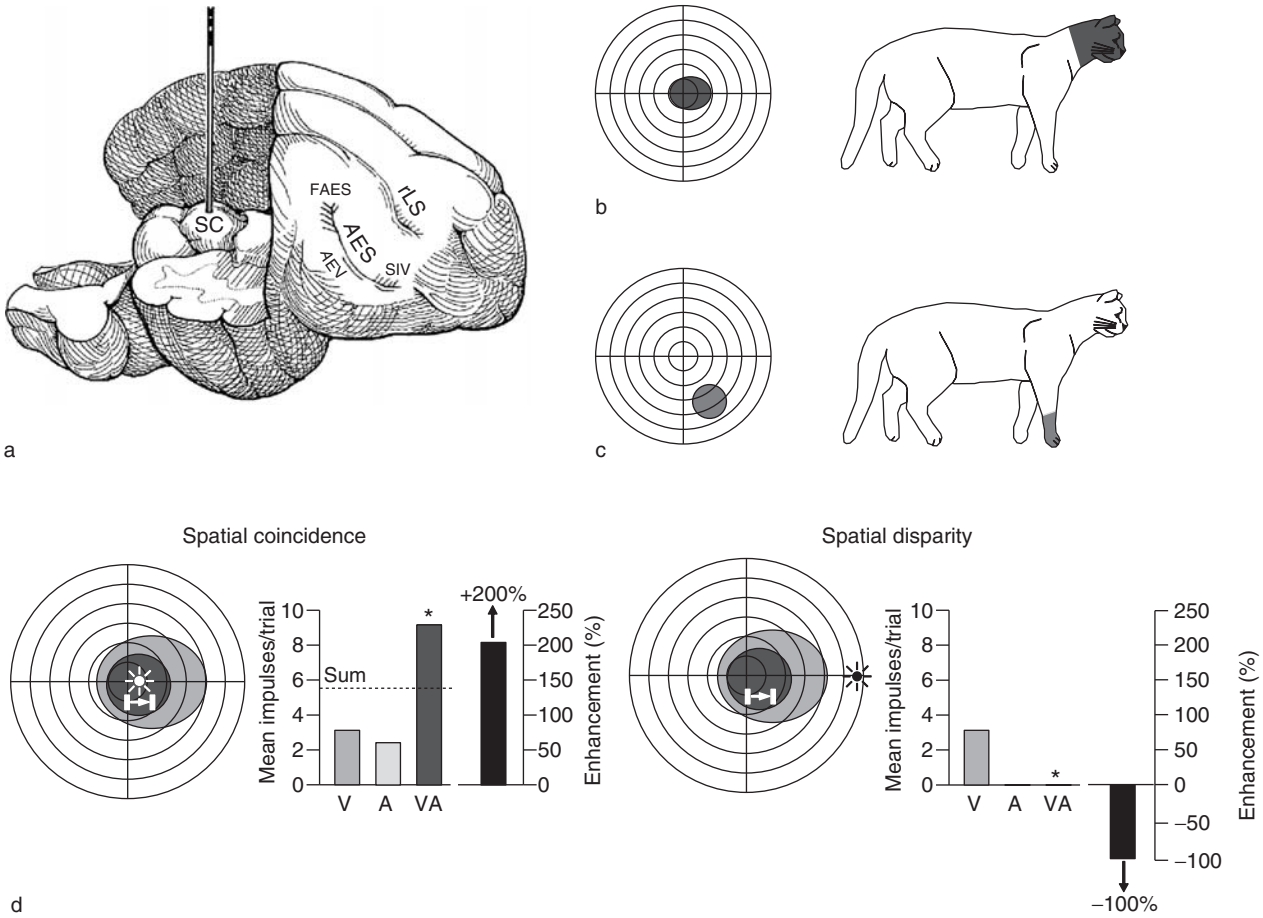

Figure 1 Multisensory enhancement and depression depend on the locations of cross-modal stimuli relative to their overlapping receptive fields: (a) posterior cortex removed from the cat brain schematic to reveal the superior colliculus (SC) in the midbrain; (b) receptive fields (shading) of a visual-somatosensory neuron; (c) receptive fields (shading) of an auditory-somatosensory neuron; (d) opposing physiological responses of a visual-auditory neuron when the cross-modal stimuli are spatially coincident (left) and spatially disparate (right). In (a), the rostrolateral suprasylvian sulcus (rLS) and the anterior ectosylvian sulcus (AES, composed of subregions FAES, AEV, and SIV) are also labeled. In (b–c), the schematics of visual and auditory spaces are polar coordinate maps in which each concentric circle represents 10° of space. In (d), when the visual and auditory stimuli are within their respective receptive fields and, thus, at a similar location in sensory space, a significant response enhancement is evoked by their pairing (left). But, when the auditory stimulus is in an inhibitory area surrounding its receptive field (inhibitory surround not shown), it counters the effect of the excitatory visual stimulus. The percentage change from the best component unisensory response is calculated using the formula $[(\mathrm{VA} - S_{\max})/S_{\max})] \times 100$, where VA is the response to the combined modality presentation and $S_{\max}$ is the response to the most effective of the two modality-specific stimuli. The predicted sum of the combined modality-specific responses is depicted by the dashed line. *indicates $p < 0.05$. A, auditory stimuli; AEV, visual region of the anterior ectosylvian sulcus; FAES, auditory region of the anterior ectosylvian sulcus; SIV, somatosensory region of the anterior ectosylvian sulcus; V, visual stimuli; VA, combined visual-auditory stimuli.

with respect to the receptive fields of a given multisensory neuron. This spatial principle of multisensory integration, first noted in SC neurons, is intuitively obvious; when multiple stimuli from the different senses are provided by the same event, they generally originate from the same location, but when they are derived from different events, they are likely to be derived from different locations. In the former case, the stimuli falling into the respective receptive fields of the same neuron provide excitatory inputs that interact synergistically to amplify neural activity and the physiological salience of the event. The magnitude of this enhancement is greatest when the impact of the individual component stimuli is weakest; a principle known as inverse effectiveness. If an interaction is induced in the latter case, when the stimuli are spatially disparate, it is one of depression because the excitation induced by one stimulus falling within its receptive field is counterbalanced by the other stimulus falling into an inhibitory region bordering its excitatory receptive field (Figure 1(d)). We can see why it is important to maintain the register among

the different sensory maps and the receptive fields of individual multisensory SC neurons even if the peripheral sensory organs move relative to one another when scanning the visual field or moving the limbs and not the eyes or head.

Mechanisms Underlying Multisensory Integration

As previously stated, the operational definition of multisensory integration is that it is a statistically significant difference between the neural responses to two concurrently presented stimuli in different modalities (e.g., a cross-modal stimulus that is visual-auditory) and the response to the most effective of the individually presented modality-specific component stimuli (e.g., visual or auditory). This definition, which speaks to the relative efficacy of cross-modal versus modality-specific stimulation, addresses the potential benefit of pooling the influences of individual cues, but does not speak to the mechanisms that give rise to the multisensory response. Clearly, the simplest mechanism for producing multisensory enhancement from spatially congruent cross-modal stimuli is one that sums the influences of the individual modality-specific inputs. Indeed, it has been demonstrated that, for many cross-modal stimulus combinations, neurons within the SC combine their modality-specific influences linearly such that the response to the cross-modal stimulus is statistically indistinguishable from the sum of the unisensory responses to the modality-specific component stimuli. Thus, in many cases, a relatively simple mechanism of summation underlies the integration of excitatory influences to give rise to multisensory enhancement.

Although simple summation is an adequate description of multisensory integration in many cases, there are notable exceptions. In particular, superadditive interactions, wherein the multisensory response exceeds the sum of the unisensory responses have often been reported. Similarly, subadditive interactions have also been observed. Recent studies examining the conditions that give rise to different integrative modes in SC neurons have shown that their relative probability is dictated by the efficacy of the modality-specific component stimuli. Combinations of very weak stimuli, especially those near the threshold of the neuron, are those most likely to produce superadditivity, whereas combinations of highly effective stimuli are those most likely to produce subadditivity. Thus, the operations that multisensory neurons apply to their modality-specific inputs appear to be a manifestation of the inverse effectiveness principle, which describes the inverse relationship between stimulus efficacy and response enhancement.

Converging Unisensory Inputs Render Superior Colliculus Neurons Multisensory

SC multisensory neurons receive their converging sensory inputs from unisensory neurons in many structures. Some ascend from the periphery, and others descend from cortex. It is interesting to note that these descending corticocollicular inputs come from association cortices that contain both unisensory and multisensory neurons but are derived from the former. Of specific interest in the model species used here (i.e., the cat) is the anterior ectosylvian sulcus (AES), shown in Figure 2(a). The AES is functionally segregated into three largely unisensory subregions from which the corticocollicular projections are derived: a somatosensory region (SIV), an auditory region (FAES), and a visual region (AEV). Between these areas are transitional zones in which non-corticocollicular multisensory neurons are concentrated.

Multisensory Integration in Anterior Ectosylvian Sulcus Cortex Is Similar to Multisensory Integration in Superior Colliculus

Despite the fact that these transitional-zone AES multisensory neurons are independent of (i.e., do not project to) those in the SC and are involved in a separate circuit, they share many fundamental features. As is true in the SC, the different receptive fields of multisensory AES neurons are in the spatial register. They also exhibit the same spatial principle of multisensory integration and the same principle of inverse effectiveness. This suggests that parallel multisensory processes going on at different levels in the neuraxis are governed by many of the same integrative principles. In this way, the different networks of multisensory neurons involved in different aspects of a coordinated behavior are rendered coherent.

Superior Colliculus Multisensory Integration Depends on Descending Inputs from Two Cortical Areas

Although SC neurons receive parallel ascending and descending converging inputs from the different senses, they do not have equivalent effects on this process. It is the descending projection system that is essential for multisensory integration but not for rendering an SC neuron multisensory. Thus, temporarily deactivating these inputs by lowering the cortical temperature until constituent neurons became inactive revealed that the modality convergence pattern of SC neurons did not change but that they could no longer integrate their cross-modal inputs (Figure 2(b)). Two cortical areas proved to be involved in this process: the AES and an adjacent corticocollicular area in the rostral aspect of the lateral suprasylvian

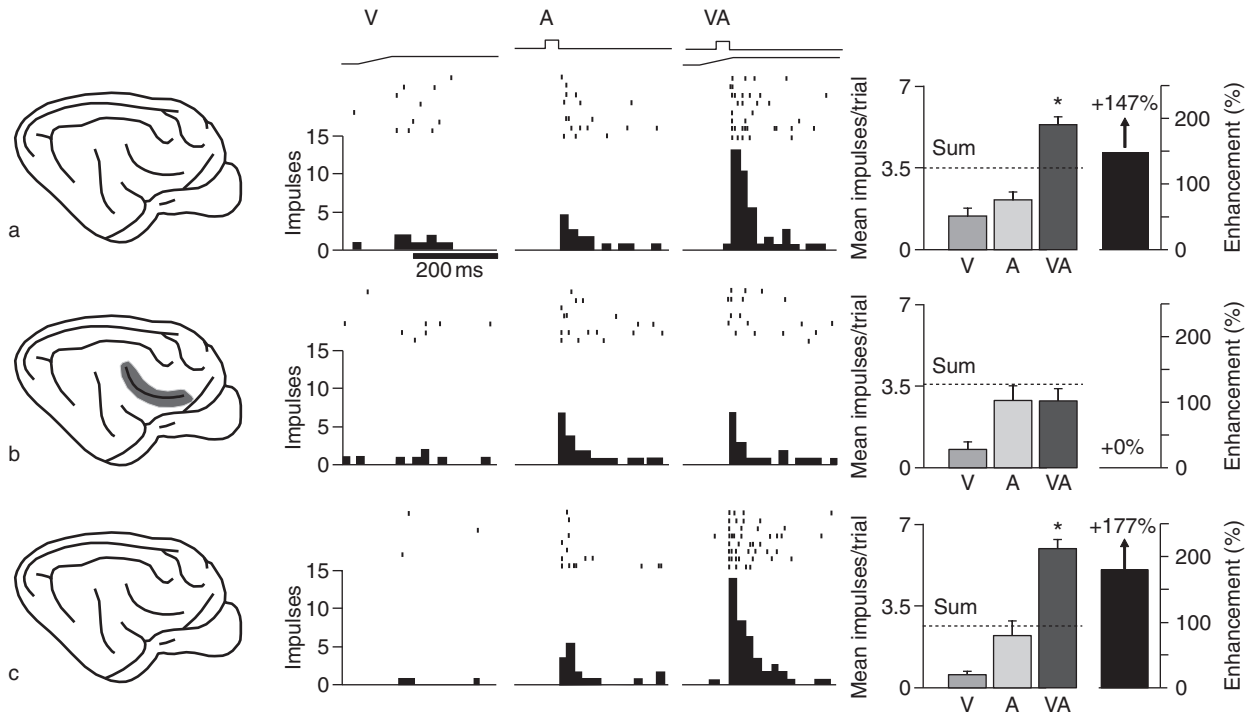

Figure 2 Deactivation of the AES and multisensory integration in a population of the SC neurons: (a) control; (b) deactivated AES; (c) reactivated AES. The deactivation of the AES abolishes multisensory integration in a population of SC neurons. Traces at the top of the figure show the time courses of visual (V) and auditory (A) stimuli. Impulse rasters (each dot represents 1 impulse and each row of dots represents the response to one trial) and histograms show the neural responses. In (a), note the enhanced multisensory response in the control condition of this visual-auditory neuron, which, in this case, exceeds the sum of the unisensory responses. In (b), multisensory enhancement is lost during cryogenic deactivation of AES (other SC neurons depend on the AES and rLS or on the rLS alone for multisensory integration). Note that the unisensory visual and auditory responses are not significantly affected. In (c), multisensory integration is reinstated when the AES is reactivated by rewarming. *indicates $p < 0.05$. AES, anterior ectosylvian sulcus; rLS, rostrolateral suprasylvian sulcus; SC, superior colliculus; VA, combined visual-auditory stimuli. Reproduced from Jiang W, Wallace MT, Jiang H, Vaughan JW, and Stein BE (2001) Two cortical areas mediate multisensory integration in superior colliculus neurons. *Journal of Neurophysiology* 85: 506–522.

sulcus (rLS). Parallel effects on behavior were also demonstrated – when the AES or rLS was deactivated, there was little change in an animal's ability to orient to a visual cue, but the multisensory enhancement due to the presence of a spatially coincident auditory cue was lost and the multisensory depression induced by a disparate auditory cue was significantly reduced. This effect was due to the loss of multisensory integration capabilities in the SC and was also seen when the SC was damaged by ibotenic acid lesions, leaving cortex intact.

These results highlight the close interplay between areas of association cortex and the SC. Such circuits not only are a likely substrate for the modulation of subcortical processes by cognitive factors (e.g., attention and motivation), but may also serve as the conduit through which experience can shape sensory processing in the SC, both during development and in adulthood. Whether the evidence ultimately supports this contention and whether other species use similar circuits for this purpose remain to be determined. However, it is likely that these intriguing possibilities will be topics of discussion and exploration in the near future.

Evaluating Multisensory Integration in Humans Using Functional Brain Imaging

As our understanding of multisensory integration expanded based on the studies performed in the SC and cortex of animal models, a growing interest emerged in the neural mechanisms of multisensory integration in humans. Many of these studies have used recordings from scalp electrodes, from which event-related potentials (ERPs) can be recorded. These ERP records represent the sum of activity from many cortical neurons simultaneously and have revealed integrative phenomena in human subjects like those already discussed in animal experiments. The rapidly growing number of multisensory

studies in humans has recently been fueled by the advances in technology that allow the imaging of human brain activity. Of these technologies, functional magnetic resonance imaging (fMRI) has provided the greatest advances in understanding multisensory processing in humans. To date, most studies that have evaluated multisensory integration have been designed based on the physiological studies performed in the SC of animals. This translational science approach, in which data from animal models are applied to human studies, is quite logical, and studies have identified superadditive responses in several brain regions of humans during multisensory integration. In particular, the superior temporal sulcus (STS) has been implicated repeatedly as a multisensory region in human brain imaging studies. Other important regions have also been identified, including (but not limited to) regions in the parietal and frontal lobes. Of growing interest is the possibility that multisensory interactions occur at the level of the sensory-specific cortices. It has been demonstrated that stimulation of one sensory modality in relative isolation can suppress activity in other nonmatching sensory cortices. More recently, a vigorous debate has emerged concerning the actual integration of multisensory stimuli in the sensory-specific cortices. Although this debate calls into question the existence of purely sensory-specific cortices, much more work must be done to resolve this important issue. With continued advances in brain imaging technology, our current understanding of sensory interactions will serve as a foundation of what is certain to be an exciting future for studies of multisensory integration.

See also: Cross-Modal Interactions Between Vision and Touch; Eye and Head Movements; Multimodal Signaling in Animals; Neural Coding of Spatial Representations; Optic Tectum: Sensorimotor Integration; Perception and Eye Movements; Saccadic Eye Movements; Sensorimotor Integration: Models; Spatial Cognitive Maps; Superior Colliculus; Vision for Action and Perception.

Further Reading

Beauchamp MS (2005) See me, hear me, touch me: Multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology* 15: 145–153.

Calvert GA, Spence C, and Stein BE (2004) *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.

Jiang W, Wallace MT, Jiang H, Vaughan JW, and Stein BE (2001) Two cortical areas mediate multisensory integration in superior colliculus neurons. *Journal of Neurophysiology* 85: 506–522.

Laurient PJ, Perrault TJ, Stanford TR, Wallace MT, and Stein BE (2005) On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research* 166: 289–297.

Meredith MA and Stein BE (1983) Interactions among converging sensory inputs in the superior colliculus. *Science* 221: 289–291.

Schroeder CE and Foxe J (2005) Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology* 15: 454–458.

Stanford TR, Quessy S, and Stein BE (2005) Evaluating the operations underlying multisensory integration in the cat superior colliculus. *Journal of Neuroscience* 25: 6499–6508.

Stein BE, Huneycutt WS, and Meredith MA (1988) Neurons and behavior: The same rules of multisensory integration apply. *Brain Research* 448: 355–358.

Stein BE, Magalhaes-Castro B, and Kruger L (1975) Superior colliculus: Visuotopic–somatotopic overlap. *Science* 189: 224–226.

Stein BE and Meredith MA (1993) *The Merging of the Senses*. Cambridge, MA: MIT Press.

Stein BE, Meredith MA, Huneycutt WS, and McDade L (1989) Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli. *Journal of Cognitive Neuroscience* 1: 12–24.

Walk RD and Pick HL (1981) *Intersensory Perception and Sensory Integration*. New York: Plenum.

Relevant Websites

http://www.multisense.info – Multisense.info.

http://www1.wfubmc.edu – Wake Forest University Baptist Medical Center.

Munc13 and Associated Molecules

R Weimer and J Richmond, University of Illinois, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurotransmitter release from presynaptic nerve terminals is mediated by the fusion of neurotransmitter-filled synaptic vesicles with the plasma membrane. Synaptic vesicle fusion is tightly coupled to voltage-induced Ca^{2+} influx and has latency on the microsecond timescale. This rapidity suggests that within a nerve terminal a population of synaptic vesicles is competent, or primed, to undergo membrane fusion immediately upon Ca^{2+} entry.

Synaptic vesicle priming and fusion requires members of the conserved soluble *N*-ethylmaleimide-sensitive factor (NSF) attachment protein receptor (SNARE) protein families, specifically, the plasma membrane-associated SNAREs (t-SNAREs, also called Q-SNAREs) syntaxin and synaptosomal-associated protein of 25 kDa (SNAP-25) and the vesicle membrane-associated SNARE (v-SNARE, also called R-SNARE) synaptobrevin. These SNARE proteins interact through their SNARE domain to form a parallel four α-helical bundle termed the SNARE complex. SNARE complex assembly *in trans* is predicted to bring the vesicle membrane into close proximity to the plasma membrane, a prerequisite for membrane fusion.

Because SNARE interactions can be detected prior to neurotransmitter release, interactions between SNARE proteins are thought to represent the molecular underpinnings of vesicle priming. By extension, proteins that interact with the SNAREs are candidates for regulators of vesicle priming. UNC-13/Munc13 is a presynaptic protein that interacts with syntaxin. Here we review the evidence implicating UNC-13/Munc13 in vesicle priming and discuss the current models for how UNC-13/Munc13 functions in vesicle priming and how its function is regulated.

Characterization of UNC-13/Munc13

UNC-13 was first implicated as a mediator of synaptic transmission from pioneering forward-genetic screens in the nematode *Caenorhabditis elegans* conducted by the now Nobel laureate Sydney Brenner. Brenner reasoned that mutations which disrupt genes required for synaptic transmission at neuromuscular junctions should result in animals that exhibit defects in locomotion, or are uncoordinated. Mutations in the *unc-13* locus result in severe locomotory defects; null alleles are essentially paralyzed. Paralysis in these mutants is not due to a gross defect in neurodevelopment or postsynaptic functions, suggesting that UNC-13 is required presynaptically for normal neurotransmission.

In *C. elegans*, *unc-13* encodes for at least three UNC-13 proteins derived from alternative splicing. All isoforms characterized to date contain a C1 domain, two C-terminal C2 domains, and two Munc homology domains (MHDs) (**Figure 1(a)**). Most contain an N-terminal extension containing a third C2 domain, a Rim-interacting domain (RID), and a calmodulin-binding domain (CBD). In general, these isoforms are grouped into two categories – long (L) and short (S) – based on the presence of this extended N-terminus.

UNC-13 is a conserved protein; homologs have been identified in nearly every organism examined that contains a nervous system. In *Drosophila*, the UNC-13 homolog (Dunc13) is encoded by a single locus which gives rise to three isoforms. In mice, UNC-13 homologs are encoded by four loci (Munc13-1, -2, -3, and -4), which may also exhibit alternative splicing. UNC-13, Dunc13, and Munc13-1, -2, and -3 are expressed primarily in neurons; UNC-13 and Dunc13 are expressed almost exclusively in neurons of the worm and fly, whereas Munc13-1, -2, and -3 exhibit tissue and temporal expression specificity.

In addition to similarities in sequence and tissue expression, mutational perturbation of UNC-13 and its homologs results in virtually identical synaptic transmission defects – a nearly complete block of neurotransmitter release from presynaptic nerve terminals. Although synaptic vesicles are abundant at presynaptic sites, virtually none are competent to undergo fusion with the plasma membrane. These phenotypes suggest that UNC-13/Munc13 is required for synaptic vesicle priming.

The SNARE complex crystal structure predicts that the assembly of *trans*-SNARE complexes during priming will bring the synaptic vesicle membrane into close apposition with the plasma membrane. However, previous analyses of synapses in which SNARE protein function was perturbed did not reveal changes in the number of plasma membrane-contacting vesicles. In contrast, a recent study using rapid high-pressure freeze fixation, rather than classic chemical fixation, to preserve the ultrastructure of *C. elegans* neuromuscular junctions found vesicle contact with the plasma membrane to be largely UNC-13 dependent. These data provide the first evidence for a morphological correlate of priming and further implicate UNC-13 in this process.

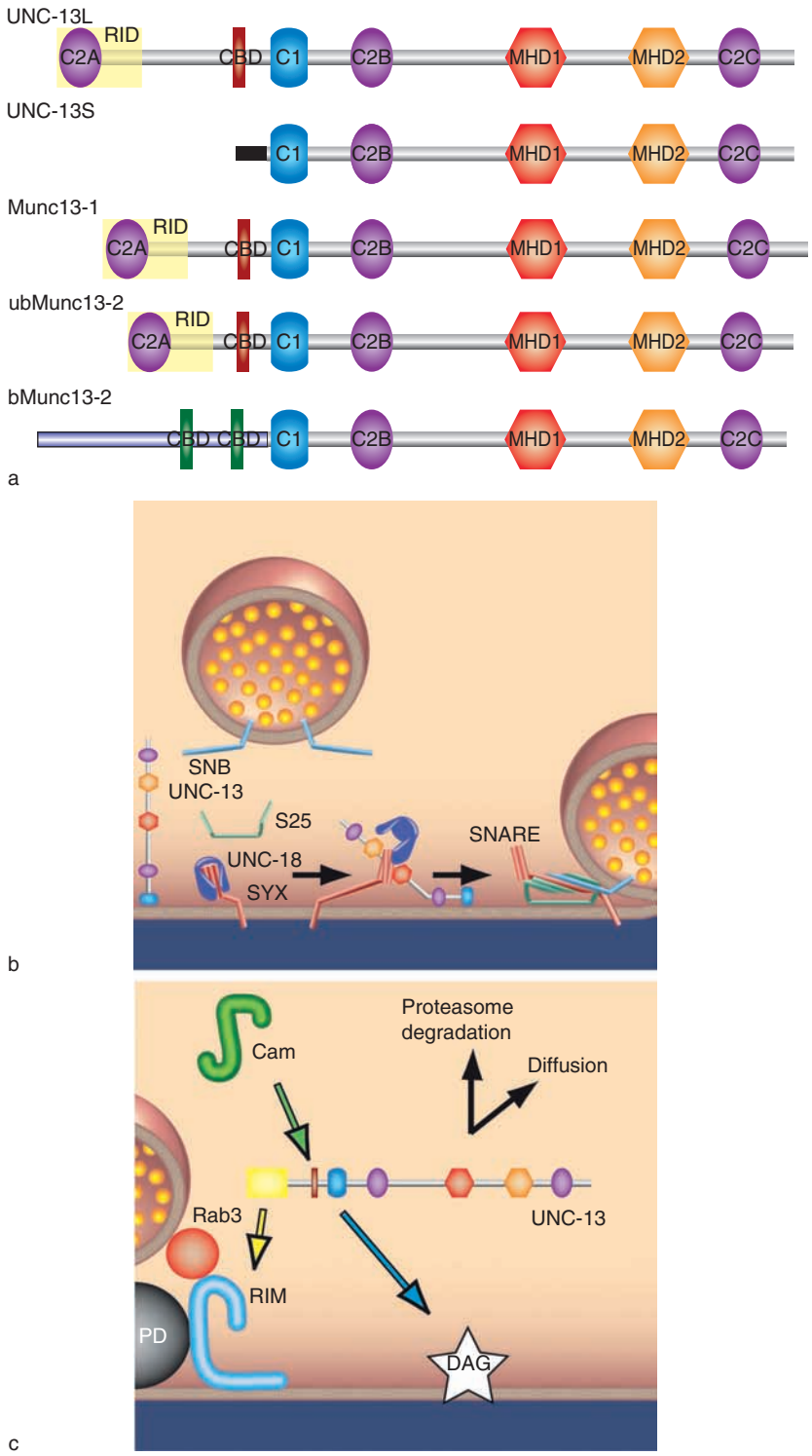

Figure 1 UNC-13/Munc13 and its molecular pathway: (a) UNC-13 and its homologs illustrated are their functional domains; (b) the probable priming by UNC-13/Munc13 of synaptic vesicles through interactions with UNC-18 and the SNARE protein syntaxin (SYX) which interacts with the SNARE proteins SNAP-25 (S25) and synaptobrevin (SNB); (c) several pathways that probably regulate UNC-13/Munc13 function at the synapse. C1, C1 domain; C2, C2 domain; Cam; calmodulin; CBD, calmodulin-binding domain; DAG, diacylglycerol; MHD, Munc homology domain; PD, presynaptic density; RID, Rim-interacting domain; Rim, Rab3-interacting molecule; SNAP-25, synaptosomal-associated protein of 25 kDa; SNARE, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor; ub, ubiquitous.

Mechanism of Action of UNC-13/Munc13

To define the region of UNC-13/Munc13 required for vesicle priming, several groups conducted experiments to identify a minimal UNC-13/Munc13 rescuing fragment. Rescue of synaptic transmission in cultured hippocampal neurons derived from Munc13-1/2 double knockout mice requires a large fragment spanning virtually the entire region between C_2B and C_2C (amino acids 859–1531) containing both MHD1 and -2. This minimally rescuing construct produced a partial rescue of evoked synaptic transmission at excitatory synapses resulting from ~50% rescue of the fusion-competent vesicle pool.

In chromaffin cells, overexpression of full-length Munc13-1 can produce a doubling in the size of the primed pool of dense-core vesicles – dense-core and synaptic vesicles exocytosis utilize much of the same molecular machinery. To attain this elevated level of dense-core vesicle priming, a slightly larger fragment of Munc13-1, amino acids 1100–1735, which extends beyond the MHDs to include the third C2 domain, is required. This suggests that a fully functional Munc13-1 rescuing fragment contains both MHDs and C2C. Consistent with this premise, rescue of *C. elegans unc-13* mutants requires a similar construct that includes both MHDs and C_2C (amino acids 676–1420).

An earlier study using this conserved C-terminal region of Munc13-1 (amino acids 1181–1736) as bait in a yeast two-hybrid screen identified syntaxin-1B as a binding partner. The clones isolated in this screen encoded an N-terminal region of syntaxin covering the first two α-helices (HA and HB), suggesting that UNC-13/Munc13's role in vesicle priming is through its interactions with the syntaxin N-terminus. Consistent with this hypothesis, point mutations in the Munc13-1 minimally rescuing fragment that disrupted syntaxin binding abolished its dense-core vesicle priming ability. Similarly, alanine-scanning mutagenesis studies of the *C. elegans* UNC-13 MHD2 also identified residues that eliminated syntaxin binding and, as a consequence, reduced evoked neurotransmitter release. Together, these experiments strongly support the notion that UNC-13/Munc13 promotes synaptic vesicle exocytosis, and probably vesicle priming, through its interaction with syntaxin.

In solution, syntaxin adopts a default-closed confirmation in which its N-terminus, containing the H_{ABC} helices, folds over and occludes its C-terminal SNARE domain (also referred to as H3 domain). This observation led to speculation that Munc13 binding to the N-terminus of syntaxin may promote the opening of syntaxin to expose its SNARE domain, thereby facilitating SNARE complex assembly and vesicle priming. Through alanine scanning, two residues in the linker region between the H_{ABC} and SNARE domains of vertebrate Syntaxin 1A (LE165/166AA) were found to destabilize the closed conformation, rendering the protein constitutively open. Expression of the corresponding mutant form of syntaxin in *C. elegans* (syntaxin LE166/167AA) partially bypasses the requirement of UNC-13 in synaptic vesicles priming, suggesting that UNC-13/Munc13 promotes vesicle priming by promoting the opening of syntaxin (**Figure 1(b)**). However, evoked synaptic transmission and behavior are not fully rescued in these animals, implicating UNC-13 in additional steps in exocytosis.

Regulation of UNC-13/Munc13

Whereas the basic function of Munc13/UNC-13 in priming has been mapped to its C-terminus, the N-terminus contains several domains that are considered important in regulating this priming function.

Rab3/RIM Interaction

UNC-13/Munc13 interacts directly with the active-zone protein Rim through its N-terminal RID (**Figures 1(a)** and **1(c)**). Rim, like UNC-13/Munc13, is required for normal levels of priming, and its function requires interaction with UNC-13/Munc13. In addition, Rim interacts with the synaptic vesicle-associated guanosine triphosphatase (GTPase) Rab3 (hence its name, Rab3-interacting molecule). Rab3 has been implicated in regulating the release probability of primed synaptic vesicles.

Rab3, Rim, and Munc13 interact to form a tripartite complex. Rim, therefore, might regulate priming by recruiting Munc13 and synaptic vesicles to active zones through this tripartite complex. Indeed the active-zone recruitment of Munc13-1 and the ubiquitous (ub)Munc13-2 splice variant is regulated by Rim. Although UNC-13 is not grossly mislocalized, fewer synaptic vesicles are located near active-zone dense projections in *C. elegans* Rim/*unc-10* and *rab-3* mutants. If Ca^{2+} entry occurs preferentially at this site, regulating the distribution of primed synaptic vesicles could account for the postpriming functions of Munc13.

C1-Diacylglycerol Regulation

Phorbol esters are known to increase presynaptic efficacy by mimicking the actions of the endogenous second messenger, diacylglycerol. In addition to protein kinase C (PKC), Munc13/UNC-13 is a well-established phorbol ester receptor in which the N-terminal C1

domain is the binding target. Although the C1 domain is not essential for the basic priming activity of Munc-13/UNC-13, it is required for survival in mice, indicating that the C1 domain performs an important function. Genetic studies in *C. elegans* have also established an important regulatory pathway in which diacylglycerol regulates motor neuron output through an UNC-13-dependent mechanism.

Phorbol ester treatment has been shown to promote the plasma membrane association of Munc13 in human embryonic kidney (HEK) cells and chromaffin cells, suggesting the membrane recruitment of Munc13 may be a potential mechanism by which phorbol esters enhance release (**Figure 1(c)**). Consistent with this hypothesis, paralysis and impaired cholinergic signaling due to loss of diacylglycerol production in *C. elegans* phospholipase-C mutants (*egl-8*) can be partially bypassed by expressing a constitutively membrane-associated form of UNC-13 produced by N-terminal myristoylation. Furthermore, introducing a point mutation at a conserved histamine within the C1 domain of *C. elegans* UNC-13(H173K), which is predicted to disrupt phorbol ester binding, eliminates phorbol ester-induced enhancement of cholinergic signaling and reduces agonist-stimulated recruitment of UNC-13 to synapses. It is worth noting that the UNC-13 isoform in these studies lacks the N-terminal C2A domain implicated in active-zone recruitment and priming via interactions with Rim/Rab3a; therefore, the membrane translocation observed may be specific to this isoform. Indeed, analysis of the corresponding point mutation in the C1 domain of Munc13-1(H567K) was not associated with a detectable defect in membrane targeting.

What, then, is the nature of the phorbol ester-dependent increase in synaptic efficacy? Surprisingly, analysis of phorbol ester effects on wild-type synapses suggests that the level of vesicle priming is unaltered but that the release probability is increased. Furthermore, the release probability of Munc13(H567K) C1 domain mutants is the same as phorbol ester-treated wild-type synapses. This observation, coupled with the fact that phorbol ester application has no further stimulatory effect on the release properties of Munc13-1(H567K) mutant synapses, suggests this is a gain-of-function mutation, mimicking the wild-type phorbol ester-activated C1 state.

Based on these observations, it has been proposed that the C1 domain regulates the functional state of Munc13 downstream of priming, under basal conditions. Munc13 may be held in a folded inactive state through a C1 intramolecular interaction with the catalytic C-terminus, this folded conformation possibly favoring interactions between Munc13 C2A and Rim/Rab3a involved in basal priming. It has been proposed that the binding of diacylglycerol/phorbol ester to the C1 domain induces a conformational change in Munc13 by disrupting C1 intramolecular binding and driving C1–plasma membrane interactions, thereby unleashing an undefined catalytic post-priming action of Munc13 that increases vesicle release probability.

In a recent analysis of phorbol ester-dependent potentiation of hippocampal synapses, both phosphorylation of Munc18 by PKC and activation of Munc13 were shown to be required. Munc18 is a syntaxin-binding partner that plays an essential role in synaptic vesicle exocytosis. Interestingly, PKC-dependent phosphorylation of Munc18 reduces the binding affinity of Munc18 for syntaxin. Therefore, it appears that the convergent involvement of both Munc18 and Munc13 in synaptic potentiation may relate to their common interaction with syntaxin. Precisely how these diacylglycerol/phorbol ester-induced events lead to increased release probability has yet to be elucidated.

Double C2-Containing Protein Family

Small elevations of presynaptic Ca^{2+} (0.1–1 μM) are thought to augment release by enhancing priming in neurons and chromaffin cells, indicating that priming is a Ca^{2+}-regulated process. Although UNC-13/Munc13 proteins contain three C2 domains, their ability to bind Ca^{2+} has yet to be tested. However, members of the double C2-containing protein family (DOC2) are positive regulators of neurotransmitter release that exhibit Ca^{2+}-dependent translocation to the plasma membrane in the Ca^{2+} range implicated in promoting priming. Because DOC2 proteins are known to interact with Munc13, the Ca^{2+}-dependent translocation of DOC2 to the plasma membrane may be a mechanism to recruit and/or activate Munc13 in the micromolar Ca^{2+} range. This hypothesis may be supported by interference experiments in which introduction of the DOC2-interacting domain of Munc13 (Did) reduces both vesicle priming in superior cervical ganglion neurons and phorbol ester-induced potentiation at the calyx of Held. Furthermore, overexpression of DOC2 in PC12 cells enhances the primed vesicle pool, implicating DOC2 in this process. Although these experiments support a positive role for DOC2 in neurosecretion, the molecular mechanism by which the DOC2–Munc13 interaction may enhance priming remains to be determined.

Calmodulin-Binding Domain

Calmodulin, which is known to co-localize with Munc13 in the presynaptic compartment, is also activated in the 500 nM Ca^{2+} range implicated in

enhanced priming. In a screen for calmodulin-binding proteins in *Drosophila*, an N-terminal Cam-binding motif in Dunc13 was identified. Following this observation, Ca^{2+}-dependent interactions between the N-terminals of Munc13-1 and ubMunc13-2 and mammalian calmodulins (Calm1–3) were also found. Sequence homology between the amphipathic helix constituting the calmodulin-binding motifs of Munc13, Dunc13, and UNC-13 indicate that this domain, located adjacent to the C1 domain, is evolutionarily conserved (Figures 1(a) and 1(c)). Arginine substitution of a conserved tryptophan residue in the calmodulin-binding motif of both Munc13-1(W464R) and ubMunc13-2(W387R) completely abolish calmodulin binding. Munc13 isoforms carrying this tryptophan to arginine mutation could therefore be expressed in autaptic hippocampal neurons from Munc13-1/2 double knockout to test the function of the calmodulin–Munc13 complex.

Neurons expressing calmodulin binding-deficient Munc13 variants had normal evoked synaptic transmission and vesicle priming, indicating that basal priming and release does not require calmodulin binding to Munc13. However, striking differences were observed in Munc13-dependent synaptic plasticity in the absence of calmodulin binding. Specifically high-frequency stimulation-induced synaptic depression associated with wild-type Munc13-1 became more pronounced in the presence of Munc13-1(W464R) and the moderate augmentation normally observed following the stimulation train was reduced, results that suggest a deficiency refilling the primed pool of vesicles. Expression of calmodulin-insensitive ubMunc13-2 (W387R) also converted the synaptic facilitation normally observed with ubMunc13-2 to synaptic depression and greatly reduced posttrain augmentation. Consequently, in the absence of calmodulin binding, the normal differences in synaptic plasticity between Munc13-1 and ubMunc13-2 were eliminated. It therefore appears that calmodulin–Munc13 complexes are essential for Ca^{2+}-dependent short-term facilitation. Precisely how Cam-binding to Munc13 proteins promotes this activity-dependent facilitation has yet to be elucidated.

Abundance

UNC-13/Munc13 activity is also probably regulated by its abundance at the synapse; two mechanisms have been implicated in this mode of regulation: proteasome-mediated degradation and diffusion/translocation.

Acute inhibition of proteasome-mediated protein degradation at *Drosophila* neuromuscular junctions causes a rapid increase in synaptic transmission. Quantitative immunohistochemical analysis of these neuromuscular junctions indicates that Dunc13 levels increase with proteasome inhibition, suggesting that Dunc13 levels at neuromuscular junctions may be under the control of proteasome-mediated degradation and, at least in part, accounts for the increase in synaptic transmission. Consistent with this hypothesis, the N-terminal region of UNC-13 was shown to interact with the F-box protein ERI-1. F-box proteins provide substrate specificity to the ubiquitination pathway. However, neither Dunc-13, UNC-13, nor Munc13 has been shown to be ubiquitinated, and proteasome inhibitors failed to increase Munc13-1 levels in neuronal cell cultures, suggesting that this potential mechanism of regulation is cell-type specific.

Although synapses may appear isolated along a neurite, they do not act in complete isolation; synapses share their constituents. Fluorescence recovery after photobleaching (FRAP) and photoactivation experiments have shown that proteins can be exchanged between synapses, including Munc13-1. FRAP experiments using a yellow fluorescent protein-tagged version of Munc13-1 suggest that Munc13-1 within a synapse is exchanged with a half-life on the order of minutes to tens of minutes. This exchange may be a mechanism for the rapid redistribution of synaptic resources during plasticity. Consistent with this hypothesis, chronic suppression of spontaneous activity decreases the exchange of Munc13-1 within synapses.

Future Directions

In summary, functions have now been assigned to each of the UNC-13/Munc13 domains. Current evidence indicates that the basic priming activity of UNC-13/Munc13 is achieved primarily through an interaction of the conserved C-terminal domain either directly or indirectly with the N-terminus of syntaxin. Domains within the N-terminus of UNC-13/Munc13 appear to regulate the location and degree of priming and may potentiate release via additional mechanisms in response to various cues. Specifically, diacylglycerol/phorbol ester-dependent potentiation acts via the C1 domain, Ca^{2+}-dependent potentiation acts via the calmodulin-binding domain, and short- and long-term potentiation act via interactions of the C2A domain with Rim and Rab3.

What remains to be elucidated are the molecular mechanisms that underlie these important Munc-13/UNC-13-dependent functions. It is anticipated that high-resolution structural information gleaned by solving the crystal structures of Munc-13/UNC-13

domains interacting with proposed binding partners will provide important insights into the molecular rearrangements occurring during basal priming and synaptic plasticity. For example, the molecular mechanism coupling UNC-13/Munc13 to SNARE complex assembly has yet to be fully understood. Biochemical approaches will help to elucidate the timing and affinities of interactions between UNC-13/Munc13 and syntaxin as well as with the SNARE complex. Similarly, the proposed conformational change of Munc13 in response to C1 activation by diacylglycerol/phorbol ester, akin to that of PKC, can be explored using these approaches.

The synaptic membrane localization of Munc13 is potentially regulated by several molecular events, including Rim/Rab3a interactions, activation of C1 with diacylglycerol/phorbol ester, as well as interactions with DOC2. Recent advances in fluorescent imaging using tagged knockin mutagenesis will be helpful in addressing the role of specific domains in UNC-13/Munc13 dynamics and localization. Combining knockin technology with Förster resonance energy transfer (FRET) and FRAP should elucidate temporal and spatial interactions between Munc13 and its binding partners under various conditions. The use of high-pressure freeze fixation and total internal reflection velocimetry (TIRV) has also begun to provide insights into the changes in vesicle distribution accompanying the activity of UNC-13/Munc13 and its binding partners. It is likely that the continued application of these techniques in combination with domain specific mutagenesis will provide a more complete picture of the crucial role this fascinating protein plays in synaptic function.

See also: Chromaffin Cells: Model Cells for Neuronal Cell Biology; Development of Drosophila Neuromuscular Junctions; Large Dense Core Vesicles (LDCVs); Munc18; NSF and SNAPs; Rab3; Rab3A Interacting Molecules (RIMs); SNAREs; Synaptic Vesicles; Vesicular Neurotransmitter Transporters.

Further Reading

Aravamudan B, Fergestad T, Davis WS, Rodesch CK, and Broadie K (1999) Drosophila UNC-13 is essential for synaptic transmission. *Nature Neuroscience* 2: 965–971.

Augustin I, Rosenmund C, Sudhof TC, and Brose N (1999) Munc13-1 is essential for fusion competence of glutamatergic synaptic vesicles. *Nature* 400: 457–461.

Basu J, Betz A, Brose N, and Rosenmund C (2007) Munc13-1 C1 domain activation lowers the energy barrier for synaptic vesicle fusion. *Journal of Neuroscience* 27: 1200–1210.

Basu J, Shen N, Dulubova I, et al. (2005) A minimal domain responsible for Munc13 activity. *Nature Structural and Molecular Biology* 12: 1017–1018.

Becherer U and Rettig J (2006) Vesicle pools, docking, priming, and release. *Cell and Tissue Research* 326: 393–407.

Betz A, Okamoto M, Benseler F, and Brose N (1997) Direct interaction of the rat unc-13 homologue Munc13-1 with the N terminus of syntaxin. *Journal of Biological Chemistry* 272: 2520–2526.

Junge HJ, Rhee JS, Jahn O, et al. (2004) Calmodulin and Munc13 form a Ca^{2+} sensor/effector complex that controls short-term synaptic plasticity. *Cell* 118: 389–401.

Kalla S, Stern M, Basu J, et al. (2006) Molecular dynamics of a presynaptic active zone protein studied in Munc13-1-enhanced yellow fluorescent protein knock-in mutant mice. *Journal of Neuroscience* 26: 13054–13066.

Lonart G (2002) RIM1: An edge for presynaptic plasticity. *Trends in Neuroscience* 25: 329–332.

Madison JM, Nurrish S, and Kaplan JM (2005) UNC-13 interaction with syntaxin is required for synaptic transmission. *Current Biology* 15: 2236–2242.

Richmond JE, Davis WS, and Jorgensen EM (1999) UNC-13 is required for synaptic vesicle fusion in *C. elegans*. *Nature Neuroscience* 2: 959–964.

Richmond JE, Weimer RM, and Jorgensen EM (2001) An open form of syntaxin bypasses the requirement for UNC-13 in vesicle priming. *Nature* 412: 338–341.

Stevens DR, Wu ZX, Matti U, et al. (2005) Identification of the minimal protein domain required for priming activity of Munc13-1. *Current Biology* 15: 2243–2248.

Varoqueaux F, Sigler A, Rhee JS, et al. (2002) Total arrest of spontaneous and evoked synaptic transmission but normal synaptogenesis in the absence of Munc13-mediated vesicle priming. *Proceedings of the National Academy of Sciences of the United States of America* 99: 9037–9042.

Wierda KD, Toonen RF, de Wit H, Brussaard AB, and Verhage M (2007) Interdependence of PKC-dependent and PKC-independent pathways for presynaptic plasticity. *Neuron* 54: 275–290.

Munc18

I Dulubova, Reata Pharmaceuticals, Irving, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Munc18 is one of the most enigmatic proteins in the field of synaptic transmission. Loss-of-function mutations of Munc18-1 or of its orthologs in several eukaryotic species invariably lead to one of the most dramatic synaptic phenotypes – either a complete loss or a severe impairment of synaptic transmission. However, in spite of the availability of genetic knockouts, a wealth of biochemical and cell biological data, and the existence of three-dimensional structures at atomic resolution, the physiological function(s) of Munc18 remain unclear, although recent advances have brought us several important insights into the biology of Munc18.

Discovery, Molecular Structure, and Biochemical Properties

Munc18-1 (also referred to as Munc18a), was originally isolated in 1993 from a solubilized rat-brain homogenate by virtue of its tight binding to syntaxin 1, a member of a class of presynaptic soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) proteins. Munc18-1 turned out to be a mammalian homolog of the product of *Caenorhabditis elegans* gene *UNC18* (thus, Munc18 stands for the mammalian unc18 homolog). In *C. elegans*, *UNC18* belongs to a group of genes identified on the basis of severe locomotion defect (uncoordinated phenotype). These genes are defined by mutations causing paralysis, resistance to the acethylcholine esterase inhibitors, and accumulation of acetylcholine suggesting impairments in neurotransmitter release. Munc18-1 is also homologous to Sec1p, the product of a long-known yeast secretory mutant gene. Almost simultaneously with the discovery of Munc18-1 based on its syntaxin binding properties, cDNAs for the same protein were independently identified in two separate screens of rodent brain cDNA libraries based on its sequence homology with Sec1. The cloned protein was named n-Sec1 or rbSec1 (for neural Sec1 homolog or rat brain Sec1, respectively). To avoid confusion, we refer to the neuronal protein as Munc18-1 and to the entire gene family as SM proteins (for Sec1/Munc18-like proteins).

In mammals, Munc18s are represented by three highly homologous isoforms, referred to as Munc18-1, -2, and -3 (or as Munc18a, 18b, and 18c). Munc18-1 is a predominantly neuronal isoform essential for synaptic vesicle (SV) exocytosis. Munc18-1 is highly expressed throughout mammalian brain; its expression profile at the transcriptional level parallels the pattern of synapse formation in the developing rodent brain. In addition to brain, Munc18-1 mRNA is abundant in the spleen, lung, kidney, and testis and is detected at lower levels in several other nonneuronal tissues. At the protein level, Munc18-1 is also highly expressed in adrenal chromaffin cells and in pituitary glands, where it is involved in large dense-core vesicle (LDCV) release. Munc18-2 and 18-3 are widely expressed outside of the neuronal system and are implicated in multiple exocytotic and trafficking processes, such as mast cells granule exocytosis, Ca^{2+}-triggered exocytosis of dense-core vesicles in platelets, insulin-regulated glucose transporter (GLUT4) translocation to the plasma membrane in adipocytes and skeletal muscle, and glucose-stimulated insulin granule exocytosis from pancreatic islet β-cells. Genetic knockout of Munc18-3, a ubiquitous Munc18 isoform, suggested that it is important for normal development of embryos, including proper formation of the central nervous system (CNS).

Munc18s, proteins of apparent molecular mass 67–70 kDa, are composed of 592–594 amino-acid residues and share 52–63% amino acid identity evenly distributed throughout their sequence. Munc18s are highly conserved in evolution (rat Munc18-1 is 100% identical with the corresponding human protein, and there is 63%, 59%, and 25% identity between mammalian Munc18-1 and its orthologs *Drosophila* Rop, *C. elegans* Unc18, and *Saccharomyces cerevisiae* -Sec1p, respectively). In the human genome, Munc18 genes are localized to chromosomes 9 (Munc18-1), 19 (Munc18-2), and 1 (Munc18-3). Although Munc18s are hydrophilic proteins with no apparent transmembrane regions, in living cells they are found peripherally associated with plasma membranes, partly, although not exclusively, through their interactions with syntaxins.

Binding to Syntaxins and Functional Coupling to SNAREs

Mutations in genes encoding for Munc18 homologs in lower eukaryotes suggested a potential involvement of this protein family in regulated secretion. *Sec1* mutants in yeast accumulate post-Golgi secretory vesicles, reflecting a defect in a late step in exocytosis. Mutations in the *Unc18* gene in nematodes lead to the accumulation of acetylcholine, indicating an impairment of one of the final steps in the vesicular

traffic of acetylcholine in the neurons. Loss-of-function mutations in the *Drosophila* Rop gene (the fly version of Munc18 that is highly expressed in the fly nervous system and in other tissues actively engaged in secretion) results in a loss of neurotransmitter release in adult photoreceptor cells as well as in the mislocalization of a number of normally secreted products. Subsequent genetic knockout of Munc18-1 in mice produced one of the most severe synaptic phenotypes known so far. Null mutant embryos were alive prenatally, but died immediately after birth. Synapses from Munc18-1-deficient mice were totally silent, demonstrating a general, complete, and permanent loss of synaptic transmission. Interestingly, the brains of the mutant mice initially developed normally, but underwent massive apoptosis and degeneration after neuronal assembly was completed, indicating that initial synaptic connectivity does not depend on synaptic activity.

The tight binding of Munc18-1 to syntaxin 1 suggested a direct link between Munc18-1 and the synaptic fusion machinery. Genetic evidence also indicated a functional connection between Sec1p and the yeast syntaxin 1 homologs Sso1p and Sso2p. Overexpression of *Sso1* or *Sso2* in yeast partially complemented the *Sec1* phenotype. These observations have prompted a thorough biochemical characterization of direct interactions between Munc18 and syntaxins.

Munc18-1 binds to neuronal syntaxin 1 in a strikingly tight interaction, allowing the one-step purification of native Munc18-1 from a total mammalian brain homogenate using a syntaxin 1 affinity column. Attempts to directly measure the affinity of Munc18-1/syntaxin 1 binding *in vitro* resulted in K_d values ranging from 1.4 to 60 nM, but it is possible that even the lower number represents an overestimation due to the intrinsic limitations imposed by the methods. Further biochemical experiments demonstrated that all Munc18s directly interact with exocytotic plasma membrane syntaxins (syntaxins 1–4), although with different specificity. The specificity of Munc18/syntaxin interactions was extensively characterized using multiple *in vitro* binding assays. Thus, Munc18-1 and -2 directly interact with syntaxins 1, 2, and 3 but not with syntaxin 4, whereas Munc18-3 specifically binds to syntaxin 4 but not to syntaxins 1–3. However, the precise affinities for most of these interactions have not been measured, and it is likely that each Munc18 isoform has a different spectrum of binding affinities for the various syntaxins.

Tight binding of Munc18 to syntaxin is an evolutionarily conserved feature that is also observed in lower eukaryotes. Thus, Unc18 protein directly interacts with *C. elegans* syntaxin 1, and *Drosophila* Rop firmly binds to both mammalian and fly syntaxin 1 *in vitro*. However, several studies indicated that in yeast Sec1p instead binds to the assembled SNARE complexes.

The physiological relevance of Munc18/syntaxin 1 interaction was further strengthened by the observations that the levels of syntaxin 1 were dramatically decreased in Munc18-1-knockout mice (by 70%), and in *C. elegans* Unc18 mutants (by ~50%). Conversely, Munc18 levels are significantly reduced (by ~25%) in syntaxin 1a-knockout mice. However, detailed characterization of the Munc18–syntaxin interaction revealed a basic and still unresolved conundrum; in order to bind Munc18, syntaxin 1 has to be in a so-called closed conformation. This is the default conformation for syntaxin 1 in isolation, which involves an intramolecular interaction between three-helical N-terminal domain of syntaxin 1 (Habc domain) and its SNARE motif. *In vitro,* this conformation is incompatible with the participation of syntaxin in the SNARE-complex formation with two other neuronal SNAREs, synaptosomal-associated protein of 25 kDa (SNAP-25) and synaptobrevin/vesicle-associated membrane protein (VAMP). Thus, the essential positive role in neuronal exocytosis suggested for Munc18-1 by the gene knockout experiments came in a conceptual clash with the apparently inhibitory mode of Munc18-1 interaction with the closed conformation of syntaxin 1.

Structural Studies of Munc18

A detailed picture of the syntaxin 1a–Munc18-1 interaction at atomic level was obtained when the crystal structure of the complex was determined at 2.6 Å resolution (referred to as nSec1-syntaxin 1a complex in the crystallography study). The structure revealed an arch-like shape for Munc18-1 with a central cavity of ~15 Å formed by three domains (**Figure 1**). Domain 1 spans the 134 N-terminal residues of munc18-1 and consists of a five-stranded parallel β-sheet flanked by five α-helices. The second domain, which is composed of residues 135–245 and 480–592, also adopts an α-β-α fold. The sequence forming the third Munc18-1 domain (residues 246–479) is inserted in the middle of the sequences making up domain 2. The third domain is composed of two subdomains, 3a and 3b, that are bent at a 60° angle with respect to one another. Subdomain 3a also contains a mixture of α- and β-structural elements, whereas the subdomain 3b represents a cluster of α-helices. Syntaxin 1A binds in its closed conformation in the central cavity of Munc18-1. The interacting surfaces of the two proteins are largely polar and complementary in charge. Both the Habc domain and the SNARE motif of the closed

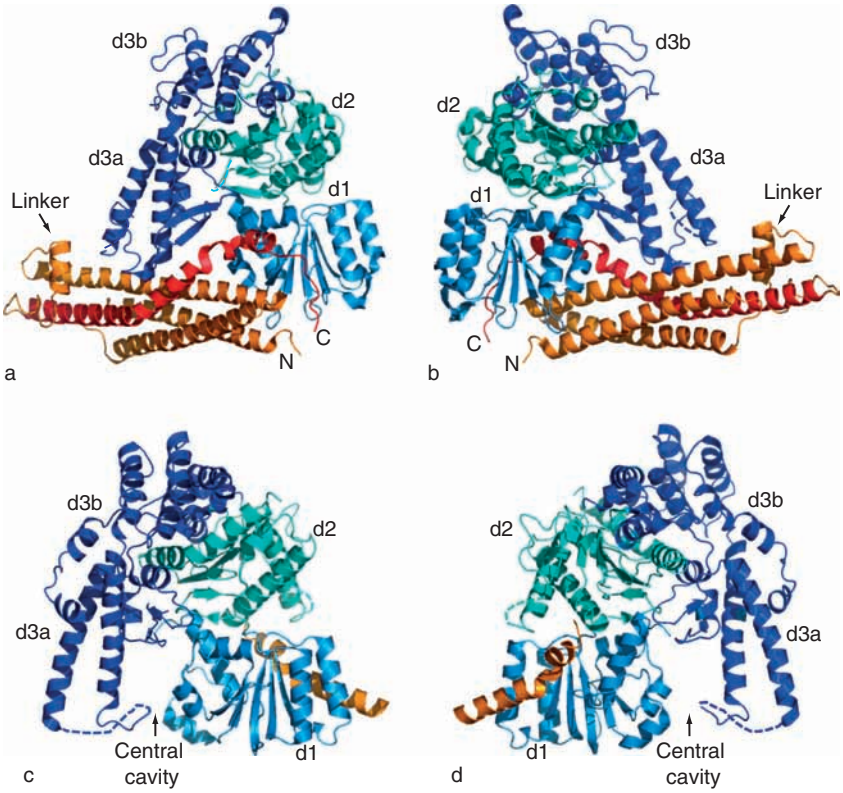

Figure 1 X-ray structures of SM proteins: (a) ribbon diagrams of the Munc18-1–syntaxin 1a complex (PDB accession code: 1DN1); (b) 180° rotation of the view depicted on the (a); (c) ribbon diagrams of the yeast SM protein Sly1p in a complex with the N-terminal peptide of the yeast syntaxin Sed5p (PDB accession code: 1MQS); (d) 180° rotation of the view depicted on the (c). In (a) and (b), the three structural domains of Munc18-1 are colored in different shades of blue; the SNARE motif of syntaxin 1A is colored in red, and the rest of the syntaxin molecule (the linker region, the three-helical Habc domain, and the N-terminal sequence) is shown in orange. The linker region and the N- and C-termini of syntaxin are indicated. In (c–d), the color code used for the structural domains of Sly1p is the same as (a) and (b); the N-terminal syntaxin peptide is shown in orange. SNARE, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor. The figure was made using PyMOL (open source software, http://www.pymol.org). Note that recently revised structure of the Munc 18-1–syntaxin 1a complex (PDB accession code : 3C98) revealed an additional binding site for the N–terminal sequence of syntaxin 1a on the outer surface of Munc 18-1 (see text).

syntaxin make extensive contacts, predominantly with domains 1 and 3a of Munc18-1. Because subdomain 3a of Munc18-1 is involved in relatively few intramolecular interactions, it was suggested that this part of the molecule might be able to move with respect to the rest of the protein due to the presence of a single nonhelical glycine residue placed between the two subdomains. Such outward movement could affect the syntaxin conformation, creating more favorable conditions for the ordered association of the syntaxin SNARE motif with the SNARE motifs of SNAP-25 and synaptobrevin.

The crystal structure of isolated Munc18 from the squid (referred to as s-Sec1 in that study) is overall very similar to the structure observed for the mammalian protein in the complex with syntaxin 1. It did confirm the structural plasticity of subdomain 3a, but it also revealed a substantial flexibility in the orientation of domain 1 of Munc18. Because both subdomain 3a and domain 1 of Munc18 are primarily responsible for the binding to the closed syntaxin, these observations offer initial clues toward understanding the mechanisms regulating this interaction. Both structures have also revealed the presence of potential ligand-binding pockets on the surface of Munc18 opposite the syntaxin binding interface (e.g., a large negatively charged cavity with approximate dimensions $20 \times 15 \times 15$ Å localized between domains 1 and 2). It has been speculated that the binding of an effector molecule to this surface of Munc18-1 could cause a rotation in domain 1 with respect to domain 2 and therefore lead to a release of syntaxin.

A more recent crystal structure of mouse Munc 18-3 (Munc 18c) in a complex with syntaxin 4 peptide (see below) revealed the same modular structure and overall fold as for the neuronal Munc 18-1 and for the squid Munc18 (1.5 Å rmsd with 45% sequence identity between Munc 18-3 and Munc 18-1, and 1.3 Å rmsd

with 51% sequence identity between Munc 18-3 and the squid protein).

The crystal structure of the syntaxin 1a–Munc18-1 complex fully confirmed the binding mode suggested by the biochemical and nuclear magnetic resonance (NMR) spectroscopy studies and uncovered the atomic basis for the incompatibility between the syntaxin–Munc18-1 complex and the SNARE complex. Altogether, these studies indicated that syntaxin 1 must experience a large conformational change to switch between the two complexes. This premise has led to the proposal that an unidentified factor(s) helps to dissociate the syntaxin 1–Munc18-1 complex or causes conformational changes in Munc18-1, after which Munc18-1 is able to perform its essential positive function in neuronal exocytosis.

Phosphorylation of Munc18

In a search for potential factors or processes that could dissociate Munc18 from syntaxin, effects of posttranslational modifications of Munc18 were examined.

Phosphorylation, in particular by conventional protein kinase C (PKC), for many years has been implicated in regulating multiple stages of exocytosis in neurons and neuroendocrine cells. Munc18 is a likely candidate to mediate some of these effects because phosphorylation affects its biochemical properties and all three mammalian isoforms of Munc18 are phosphorylated in intact cells in response to different secretory stimuli or modulators of protein kinase activities.

In vitro, Munc18-1 can be phosphorylated by PKC and cyclin-dependent kinase 5 (Cdk5), but not by protein kinase A (PKA) or calmodulin-dependent protein kinase II (CaMKII). PKC phosphorylates Munc18-1 at two serine residues, Ser306 and Ser313, and this phosphorylation inhibits Munc18/syntaxin interaction *in vitro*. At present, only phosphorylation of Ser313 of Munc18-1 has been confirmed to occur *in vivo*. In bovine adrenal chromaffin cells and in rat brain synaptosomes, phorbol ester treatments or the application of secretagogs caused a marked increase in phosphorylation of Ser313, but not of Ser306, as revealed by phosphospecific antisera generated to these sites. The crystal structure of the Munc18-1–syntaxin 1a complex indicated that phosphoserine in position 313 could be sterically accommodated without disrupting the syntaxin-binding surface; however, the introduction of a negative charge in this area would have an effect on the electrostatically positive character of domain 3a and perturb its charge complementarity with the syntaxin 1a linker region. Ser306 is localized further away from syntaxin 1a in the complex, and it is not immediately apparent how the phosphorylation of this site could directly affect interaction between the two proteins.

A double-point mutant of both Ser306 and Ser313 to glutamate that demonstrated reduced binding to syntaxin 1a *in vitro* was employed as a phosphomimetic mutant to examine the role of Munc18-1 phosphorylation by PKC in the regulation of release. These studies, performed in transfected chromaffin cells, used carbon-fiber amperometry to analyze the extent and kinetics of single-vesicle release events from permeabilized cells. Although the mutant did not affect the number of exocytotic events, it did produce changes in single-vesicle release kinetics, implicating PKC-mediated phosphorylation of Munc18 in this process. However, a single-site phosphomimetic mutant, S313E, failed to reproduce this effect. In addition, in Munc18-1-knockout mice, which exhibit a severe reduction (but not a complete abolishment) of LDCV secretion in chromaffin cells, no changes in the kinetic properties of single LDCV fusion events were observed.

Munc18-1 phosphorylation by PKC was also implicated in a process of efficient replenishment of the releasable pools of vesicles in bovine chromaffin cells. Rescue experiments in *munc18-1* null mutant autaptic neurons with PKC phosphorylation deficient mutant of Munc18-1 (referred to as $M18_{PKCi}$) also demonstrated that Munc18-1 is an essential substrate in the PKC-dependent pathway and implicated Munc18-1 phosphorylation by PKC in short-term synaptic plasticity.

In summary, there is little doubt that Munc18 is physiologically phosphorylated by PKC, but the molecular mechanism by which PKC phosphorylation of Munc18-1 alters the parameters of release and synaptic plasticity is still unclear. One plausible model is that PKC-dependent phosphorylation of Munc18-1 dynamically regulates the strength of Munc18-1 interaction with syntaxin 1 by decreasing their binding affinity. Alternatively, PKC phosphorylation could promote the translocation of Munc18-1 to a different protein.

Cdk5 has been also shown to phosphorylate Munc18-1 *in vitro*. Phosphorylation occurs at Thr574 and results in a strong reduction of syntaxin binding, leading to a model whereby the phosphorylation of Munc18-1 by Cdk5 mediates the disassembly of the syntaxin–Munc18 complex. Cdk5, a serine/threonine protein kinase that is highly and ubiquitiously expressed in the adult brain, was also implicated in the regulation of neuronal activity and synaptic plasticity. It attracted special attention because Cdk5 was shown to modulate the dynamics of fusion pore during exocytosis. It was suggested

that this effect could be at least partly mediated by the phosphorylation of Munc18-1. However, the dynamics of M18-1 phosphorylation/dephosphorylation in rat brain nerve terminals indicated that the PKC, rather than Cdk5, is a primary kinase for the presynaptic Munc18. In addition, in the crystal structure of Munc18-1 Thr-574 is buried and structurally inaccessible to the kinase. Thus, the precise understanding of how phosphorylation of Munc18-1 by Cdk5 could modulate exocytosis needs further clarification.

Outside of the neuronal system, tyrosine phosphorylation of Munc18-3 (Munc18c) in response to glucose stimulation in adipocytes and insulin-secreting cell line was reported. The phosphorylation, which negatively affected Munc18-3–syntaxin 4 binding, was implicated in regulation of insulin granule exocytosis and in platelet-derived growth factor (PDGF)-stimulated glucose transporter 4 translocation in adipocytes. Crystal structure of Munc18-3 in a complex with syntaxin 4 peptide suggested a molecular basis for this regulation. Phosphorylation of Tyr219 on Munc18-3 could potentially affect Munc18 loop that forms a hydrogen bond with syntaxin 4 N-terminal peptide (see below).

Altogether these data emphasize the notion that protein phosphorylation could well be among conserved mechanisms regulating Munc18 function *in vivo*.

Munc18 Interacting Proteins and Proteins Implicated in Regulating Munc18–Syntaxin Interaction

Multiple proteins were reported to either directly bind to Munc18 or regulate Munc18–syntaxin interaction. Among them are: Mints, DOC2, granuphilin/Slp4-a, phospholipase D, Munc13, and Tomosyn.

Mint1 and Mint2 (Munc18-1-interacting proteins) are multidomain neuronal adaptor proteins that were initially suggested as mediating the function of Munc18-1 in neurotransmitter release. However, genetic knockout of Mint1 in mice demonstrated selective changes in presynaptic plasticity in inhibitory synapses, but did not reveal global impairments of synaptic functions. More recent studies on mice with genetic disruption of all Mint isoforms uncovered a decrease in Munc18-1 levels and a global change in neurotransmitter release, consistent with a modulatory role for the Mint–Munc18 interaction in release.

Double C2 domain (DOC2) proteins are vesicular adapter proteins that bind to Munc18 in multiple *in vitro* assays. Their capability to shift the equilibrium between syntaxin–Munc18 and DOC2–Munc18 complexes suggested that they may also regulate the release of syntaxin from the Munc18–syntaxin complex *in vivo*. Disruption of endogenous Munc18c–DOC2β complexes or manipulations with expression levels of DOC2β affected glucose-stimulated insulin secretion in transfected cells. However, deletion of DOC2α, the predominant neuronal isoform, in mice demonstrated that DOC2 is not essential for Ca^{2+}-dependent neurotransmitter release *per se*. Thus, at least in the neuronal system, it seems unlikely that DOC2s serve as direct regulators of Munc18–syntaxin interaction *in vivo*.

Several studies have been published on Munc18 interaction with granuphilin (also known as Slp4-a, a member of a family Rab3/Rab27 effectors specific for pancreatic beta-cells and pituitary glands). Granuphilin directly interacts with Munc18-1, syntaxin 1a, and with Munc18-1-syntaxin 1a complex in multiple binding assays leading to a proposal that granuphilin molecularly docks insulin granules to the fusion machinery. RNAi experiments in insulin secreting cells suggested that granuphilin functionally collaborates with Munc18 during insulin granules exocytosis; however, precise molecular mechanisms of this interaction require further elucidation.

It was also suggested that Munc18-1 directly associates with and negatively regulates the basal activity of mammalian phospholipase D (PLD); however, the physiological meaning of this interaction for regulated exocytosis is unclear.

Several proteins were also evaluated for their potential role in regulating Munc18-1–syntaxin 1 interaction; however, only Munc13 still firmly holds its position as a potential candidate for such a function.

Munc13-1, a mammalian homolog of the *C. elegans* Unc13, is a large neuronal multidomain protein implicated in priming of SVs during exocytosis. Disruption of *(M)UNC13* gene function in mice, *C. elegans,* and *Drosophila* results in the loss of evoked synaptic transmission. Although SVs are morphologically docked at the plasma membrane, they do not fuse in response to stimuli that normally release the entire readily releasable primed vesicle pool, suggesting a role for Munc13 in priming. The potential ability of Munc13 to regulate the Munc18-1–syntaxin interaction was suggested by the initial observation that Munc13 directly binds to the N-terminal sequence of syntaxin 1 *in vitro* and by the follow-up report showing that the *C. elegans* Unc13 protein can displace Unc18 from syntaxin. However, the strongest support for this idea comes from the fact that the overexpression of an open form of syntaxin 1 (called LE mutation, which switches the equilibrium toward the open syntaxin conformation in isolation and lowers the affinity of Munc18–syntaxin binding) rescued evoked SV release in *UNC13*-null mutants. These results

strongly suggested that Unc13 functions in priming by regulating the conformational state of syntaxin 1 to control SNARE complex formation. Interestingly, open syntaxin also restored evoked release in *UNC10* mutants. Unc10 represents the *C. elegans* homolog of RIM (Rab3-interacting molecule), which also directly interacts with Munc13. Together, these data suggest that RIM can facilitate Munc13 function. Thus, Munc13 in conjunction with RIM may represent an essential part of a mechanism releasing syntaxin from Munc18-1; however, the molecular details of this mechanism are still unclear.

Tomosyn is a large cytosolic syntaxin-binding protein implicated in the modulation of both constitutive and regulated exocytosis. In *C. elegans* its ortholog TOM-1 negatively regulates synaptic transmission presumably by affecting the priming step of SV exocytosis. *In vitro*, Tomosyn can displace syntaxin from the Munc18-1–syntaxin 1 complex; however, the physiological function of Tomosyn is most likely to be attributed to its ability to act as a competitive inhibitor of the SNARE complex. Tomosyn contains a C-terminal SNARE motif that can take place of the Synaptobrevin/VAMP SNARE motif in the genuine SNARE complex, as convincingly demonstrated by the X-ray crystallography studies.

Extended Family of SM Proteins

Whereas Munc18s and their yeast ortholog Sec1 are selectively involved in regulated secretion in mammals and yeast, respectively, their homologs function at different transport steps in eukaryotic cells. Altogether, these proteins form a large family known as the SM proteins. There are four SM proteins in yeast. and seven SM proteins in mammals (Table 1). In pairwise comparisons, SM proteins exhibit approximately a 17–27% overall sequence identity that is evenly distributed over their entire length. In yeast, the SM protein family is represented by Sec1p, Sly1p, Vps33p (also known as Slp1p or Vam5p), and Vps45p/Stt10p. Sec1p operates at the plasma membrane, Sly1p regulates traffic between the endoplasmic reticulum (ER) and Golgi, Vps45p functions in the *trans*-Golgi network (TGN)/endosome system, and Vps33p controls vacuole traffic. Their mammalian counterparts, Munc18-1, -2, and -3, mSly1, mVps45, and two mVps33 (Vps33a and -b) operate in analogous cellular compartments in a mammalian cell.

A wealth of genetic and biochemical data demonstrated a critical involvement of SM proteins in SNARE-mediated intracellular membrane fusion reactions. Most, although not all, of the SM proteins directly interact with their cognate syntaxins with high specificity and affinity. However, initial studies of the mechanisms of their binary interactions suggested that the binding of neuronal Munc18-1 to the closed conformation of syntaxin 1 is rather unique. The most prevalent mode of a direct interaction between SM proteins and syntaxin is irrespective of the open or closed syntaxin conformation and is largely mediated by a short N-terminal syntaxin sequence preceding the Habc domain (Figure 2). This interaction mode, which was originally found between SM proteins and syntaxins involved in vesicular transport between the ER and Golgi and through the TGN/early endosomes (Table 1), is evolutionarily conserved and, significantly, is compatible with SNARE complex formation. The N-terminal-binding mode has been demonstrated for several syntaxins and their cognate SM proteins: yeast Tlg2p with Vps45p, Sed5p and Ufe1p with Sly1p, and their mammalian counterparts (syntaxin 16 with mammalian Vps45, and syntaxin 5 and syntaxin 18 with mammalian Sly1). The crystal structure of yeast Sly1p in a complex with the N-terminal peptide of Sed5p and NMR studies of mammalian Sly1/syntaxin 5 interaction uncovered the structural basis for this binding mode. Whereas Sly1 shares a common arch-shaped fold with the neuronal SM proteins, the N-terminal syntaxin 5/Sed5p peptide binds in α-helical conformation to the groove on the outside surface of domain 1 of Sly1 (Figures 1(c) and 1(d)), a completely different interface compare to the binding of the closed syntaxin 1 to Munc18-1. Thus, the big central cavity of the SM protein occupied by the closed syntaxin in the Munc18-1–syntaxin 1 complex is free to mechanistically serve as the site of action for the positive function of the SM proteins in fusion (e.g., as a platform for the SNARE complex assembly).

Although syntaxin N-terminal-binding paradigm was initially uncovered for the SM proteins involved in trafficking between intracellular compartments, it

Table 1 SM proteins in yeast and mammals

Yeast SM protein	*Cellular compartment*	*Mammalian SM protein*	*Cellular compartment*
Sec1p	Plasma membrane, concentrated at the sites of exocytosis	Munc18-1, -2, -3	Plasma membrane
Vps45p	TGN, early and late endosomes	mVps45	Golgi/endosomes
Sly1p	ER/Golgi	mSly1	ER/Golgi
Vps33p	Vacuoles, enriched in the vertices of vacuole–vacuole contacts	mVps33A and B	n.d.

ER, endoplasmic reticulum; n.d., not determined; SM, Sec1/Munc18-like; TGN, *trans*-Golgi network.

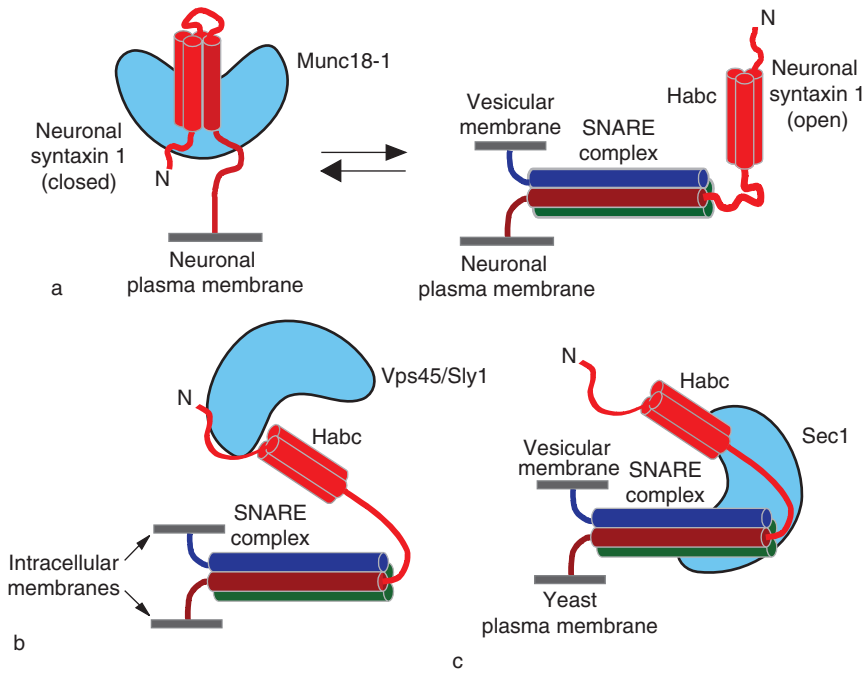

Figure 2 Mechanisms of SM protein–SNARE coupling in membrane traffic: (a) Munc18-1 interacts with the closed conformation of neuronal syntaxin 1a (left) while SNARE complex formation requires an open conformation of syntaxin (right); (b) N-terminal peptide-based interaction among Vps45p/mVps45, Sly1p/mSly1, and their cognate syntaxins (Tlg2p/syntaxin 16 for Vps45p/mVps45; Sed5p/syntaxin 5 and Ufe1p/syntaxin 18 for Sly1p/mSly1); (c) Sec1p interaction with the assembled SNARE complex. Syntaxin is colored in different shades of red (the SNARE motif is in dark red; the linker region, the three-helical Habc domain, and the N-terminal sequence are shown in bright red). The other SNARE motifs involved in the SNARE complex are in green and dark blue. The SM proteins in all three panels are in light blue. N, N-terminus of syntaxins; SNARE, soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor. Note that rapidly emerging data (see text) suggest a significant convergence between these models. Binding to the assembled SNARE complexes containing cognate syntaxins has been recently described as a second interaction mode for Vps45p, Sly1p, Munc18-3 and neuronal Munc18-1, as well as high affinity binding of Vps45 to the likely 'closed' syntaxin 16.

was recently extended toward Munc18 proteins involved in regulated exocytosis. Crystal structure of Munc18-3 in a complex with syntaxin 4 N-terminal peptide revealed a striking evolutional conservation of syntaxin N-terminal-binding mode from yeast to mammals and further emphasized potential physiological significance of this interaction.

Another coupling mode is found for the SM proteins implicated in the vesicular trafficking through the 'digestive tract' of the eukaryotic cell, specifically through the vacuolar and prevacuolar/lysosomal compartments. Here, the yeast vacuolar SM protein Vps33p is engaged in a large multisubunit protein complex, including SNAREs, but not via direct interaction with the cognate syntaxin, Vam3p. Vps45p, which has been genetically proven to operate at the same trafficking step as yeast syntaxin Pep12p, also does not directly interact with this syntaxin but rather operates as an integral part of a large protein complex.

To add to the complexity of Munc18-SNARE coupling, the interaction between Munc18-1 and syntaxin-SNAP-25 heterodimer has also been reported.

The most striking distinctions in the mechanisms of Munc18–SNARE coupling are found between SM proteins involved in exocytosis in yeast and in the SV exocytosis at neuronal plasma membrane (Figure 2). Yeast Sec1p preferentially binds to the assembled SNARE complex, in sharp contrast to the high-affinity binding of neuronal Munc18-1 to syntaxin 1. Altogether, the apparent diversity of the Munc18–SNARE coupling mechanisms has considerably complicated the process of uncovering the biological function(s) of SM proteins. However, several most recent discoveries suggest that these apparently diverse mechanisms have more in common than previously thought.

Potential Role(s) of SM Proteins in Fusion and Munc18-1 Function(s) in Neuronal Exocytosis

The high degree of sequence conservation and the overwhelming evidence that SM proteins are critically involved in all SNARE-dependent intracellular

fusion reactions suggest that SM proteins must have a common biological function. A function for SM proteins in vesicular docking was suggested by observations that the deletion of Munc18-1 in mice results in a significant reduction of docked LDCVs in adrenal chromaffin cells and that disruption of *UNC-18* gene in *C. elegans* leads to a decrease of morphologically docked vesicles in the mutant synapses. However, docking of SVs in the prenatal neurons of Munc18-1-knockout mice is apparently normal, speaking against the universal character of the docking phenotype. Overall, the formulation of a unifying hypothesis to account for the molecular mechanisms of SM protein function was largely precluded by the lack of a common SM protein–SNARE coupling mechanism. Such a mechanism has recently started to surface.

A number of studies demonstrated that, in addition to Sec1p, several other SM proteins have a capacity to interact with assembled SNARE complexes. Both yeast Sly1p and Vps45p, in addition to their well-characterized binding to the N-terminal sequences of the corresponding syntaxins (**Figure 2**), exhibit a second mode of interaction with SNARE complexes containing their cognate syntaxins. This observation was later expanded to the mammalian proteins. An N-terminal syntaxin-binding mode was reported for the Munc18-3 (Munc18c)–syntaxin 4 interaction. Munc18-3 is also shown to directly bind to syntaxin 4-containing SNARE complexes *in vitro*. Moreover, the preformation of Munc18-3–syntaxin 4 complex is reported to increase the rate of SNARE complex assembly. This is in line with previous observations in yeast suggesting that Sly1p binding to Sed5p confers fidelity to the SNARE complexes *in vitro* and that the loss of Vps45p function negatively affects the ability of Tlg2p to form functional SNARE complexes *in vivo*. Altogether, these results suggest that it is the association of the SM proteins with the assembled SNARE complexes that might represent a missing unifying mechanism underlying a common physiological function(s) of the SM proteins in fusion. One of such common function(s) could be to provide a platform for a concerted process of SNARE-complex assembly.

How can this notion be reconciled with the inhibitory character of the neuronal Munc18-1–syntaxin 1a complex? Recent data showed that Munc18-1 also directly binds to and accelerates SNARE complex formation *in vitro*, and that removal of the N-terminal sequence of syntaxin 1 abolishes both the interaction and the activation effect on the complex formation. Moreover, the N-terminal syntaxin 1 sequence is critical for its biological function, as demonstrated by multiple cell-based assays and by the electrophysiological experiments in the calyx of Held synapses. NMR spectroscopy experiments revealed that this sequence is actually engaged in the interaction with Munc18-1 in the Munc18-1–syntaxin 1a complex. In the originally published crystal structure of the complex, the N-terminal residues of syntaxin 1a were apparently disordered and thus were absent from the final structure. However, subsequent reexamination of the electron density guided by the available structures of two other complexes (Munc18-3 with syntaxin 4 N-terminal peptide, and Sly1p with Sed5p N-terminal peptide) revealed that the N-terminal sequence of syntaxin 1a also binds to the predicted site on the outer surface of Munc18-1 in Munc18-1 syntaxin 1a complex. Despite overall similar binding mode, syntaxin 1a N-terminal sequence binds weaker and forms only 4 hydrogen bonds with Munc18-1, while syntaxin 4 N-terminal sequence forms 13 hydrogen bonds with Munc18-3. These variations in binding strength are primarily due to isoform-specific sequence differences. Interestingly, molecular modeling revealed potential presence of the syntaxin N-terminal binding mode in most but not all SM protein–syntaxin pairs suggesting that this property might have coevolved to satisfy a particular regulatory function. Altogether, these results once again point toward a unified syntaxin/SM-protein-binding mechanism and also support the notion that the universal biological function of the SM proteins in fusion could be performed via their interaction with SNARE complexes. The emerging view is that *in vivo* SM proteins are switching between distinct binding modes with their cognate SNARE proteins or protein complexes during different stages of the SNARE assembly/disassembly cycle. These interactions, although overall similar for most if not all SM proteins, could be of greatly different affinities dictated by the specific needs of a particular transport step. Some of these complexes could be rather transient by nature and difficult to detect. High-affinity binding of the neuronal Munc18-1 to the closed conformation of syntaxin 1 could be evolved to meet very special spatial and temporal requirements of the neuronal exocytosis. Physiological relevance of the closed syntaxin was confirmed by recent genetic studies in mice. The analysis of transgenic mice expressing open syntaxin 1 revealed massively accelerated rate of SV fusion confirming a critical role for the closed conformation of syntaxin 1 in regulating neurotransmitter release. The data also suggested that the closed conformation operates upstream of fusion reaction and controls the entry of SV into the fusion reaction that is then mediated by Munc18-1 binding to assembled SNARE complexes.

See also: Chromaffin Cells: Model Cells for Neuronal Cell Biology; Large Dense Core Vesicles (LDCVs); Munc13 and Associated Molecules; Rab3A Interacting Molecules (RIMs); SNAREs; Synaptic Vesicles; Vesicular Neurotransmitter Transporters.

Further Reading

Aalto MK, Keranen S, and Ronne H (1992) A family of proteins involved in intracellular transport. *Cell* 68: 181–182.

Barclay JW, Craig TJ, Fisher RJ, et al. (2003) Phosphorylation of Munc18 by protein kinase C regulates the kinetics of exocytosis. *Journal of Biological Chemistry* 278: 10538–10545.

Bracher A and Weissenhorn W (2002) Structural basis for the Golgi membrane recruitment of Sly1p by Sed5p. *EMBO Journal* 21: 6114–6124.

Carr CM, Grote E, Munson M, Hughson FM, and Novick PJ (1999) Sec1p binds to SNARE complexes and concentrates at sites of secretion. *Journal of Cell Biology* 146: 333–344.

Craig TJ, Evans GJ, and Morgan A (2003) Physiological regulation of Munc18/nSec1 phosphorylation on serine-313. *Journal of Neurochemistry* 86: 1450–1457.

de Vries KJ, Geijtenbeek A, Brian EC, de Graan PN, Ghijsen WE, and Verhage M (2000) Dynamics of munc18-1 phosphorylation/dephosphorylation in rat brain nerve terminals. *European Journal of Neuroscience* 12: 385–390.

Dulubova I, Khvotchev M, Lui S, Huryeva I, Südhof TC, and Rizo J (2007) Munc18-1 binds directly to the neuronal SNARE complex. *Proceedings of National Academy of Sciences of the United States of America* 104: 2697–2702.

Dulubova I, Sugita S, Hill S, et al. (1999) A conformational switch in syntaxin during exocytosis: Role of munc18. *EMBO J* 18: 4372–4382.

Dulubova I, Yamaguchi T, Arac D, et al. (2003) Convergence and divergence in the mechanism of SNARE binding by Sec1/Munc18-like proteins. *Proceedings of the National Academy of Sciences of the United States of America* 100: 32–37.

Garcia EP, Gatti E, Butler M, Burton J, and De Camilli P (1994) A rat brain Sec1 homologue related to Rop and UNC18 interacts with syntaxin. *Proceedings of the National Academy of Sciences of the United States of America* 91: 2003–2007.

Hata Y, Slaughter CA, and Südhof TC (1993) Synaptic vesicle fusion complex contains unc-18 homologue bound to syntaxin. *Nature* 366: 347–351.

Hu S-H, Latham CF, Gee CL, James DE, and Martin JL (2007) Structure of the Munc18c/Syntaxin4 N-peptide complex defines universal features of the N-peptide binding mode of Sec1/Munc18 proteins. *Proceedings of National Academy of Sciences of the United States of America* 104: 8773–8778.

Jahn R, Lang T, and Südhof TC (2003) Membrane fusion. *Cell* 112: 519–533.

James DE (2005) MUNC-ing around with insulin action. *Journal of Clinical Investigations* 115: 219–221.

Latham CF, Lopez JA, Hu S-H, et al. (2006) Molecular dissection of the munc18c/syntaxin4 interaction: implications for regulation of membrane trafficking. *Traffic* 7: 1408–1419.

Misura KM, Scheller RH, and Weis WI (2000) Three-dimensional structure of the neuronal-Sec1–syntaxin 1a complex. *Nature* 404: 355–362.

Oh E and Thurmond DC (2006) The stimulus-induced tyrosine phosphorylation of Munc18c facilitates vesicle exocytosis. *Journal of Biological Chemistry* 281: 17624–17634.

Pevsner J, Hsu SC, and Scheller RH (1994) n-Sec1: A neural-specific syntaxin-binding protein. *Proceedings of the National Academy of Sciences of the United States of America* 91: 1445–1449.

Richmond JE, Weimer RM, and Jorgensen EM (2001) An open form of syntaxin bypasses the requirement for UNC-13 in vesicle priming. *Nature* 412: 338–341.

Rizo J and Südhof TC (2002) Snares and Munc18 in synaptic vesicle fusion. *Nature Reviews Neuroscience* 3: 641–653.

Shen J, Tareste DC, Paumet F, Rothman JE, and Melia TJ (2007) Selective activation of cognate SNAREpins by Sec1/Munc18 proteins. *Cell* 128: 183–192.

Toonen RF and Verhage M (2003) Vesicle trafficking: pleasure and pain from SM genes. *Trends in Cell Biology* 13: 177–186.

Verhage M, Maia AS, Plomp JJ, et al. (2000) Synaptic assembly of the brain in the absence of neurotransmitter secretion. *Science* 287: 864–869.

Weimer RM and Richmond JE (2005) Synaptic vesicle docking: A putative role for the Munc18/Sec1 protein family. *Current Topics in Developmental Biology* 65: 83–113.

Muscarinic Receptors: Autonomic Neurons

R S Aronstam and P Patil, University of Missouri – Rolla, Rolla, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Muscarinic receptors recognize the neurotransmitter acetylcholine, translating this recognition into electrical transients and altered cell behavior by activating and suppressing an assortment of signaling pathways. Muscarinic receptors comprise one of the two classes of receptors for the neurotransmitter acetylcholine, with nicotinic receptors comprising the other class. Muscarinic receptors are selectively activated by the alkaloid muscarine from the mushroom *Amanita muscaria* and are blocked by belladonna alkaloids, such as atropine and scopolamine (Figure 1). Muscarinic receptors are involved in the transduction of cholinergic signals in the central nervous system, autonomic ganglia, smooth muscles, and other parasympathetic end organs.

The history of muscarinic systems is intimately associated with the development of receptor theory, pharmacology, and the discovery of neurotransmitter transmission. Muscarinic receptors are members of the superfamily of G-protein-coupled receptors, specifically class A (rhodopsin-like) receptors. Muscarinic receptors are related to the ionotropic nicotinic acetylcholine receptors only insofar as their physiological activator is acetylcholine; muscarinic and nicotinic receptors share little similarity in their structure, physiological functions, or pharmacology (except for a few close analogues of acetylcholine).

Muscarinic Receptor Subtypes

The genes for five subtypes of muscarinic receptors, M1–M5, were identified, cloned, and sequenced between 1986 and 1990. These receptor subtypes differ in their primary structure, distribution, pharmacology, and signal transduction activity (Table 1). This heterogeneity presents the possibility of selectively affecting specific muscarinic functions in the brain and other organs; accordingly, the pharmacology of muscarinic receptor subtypes has been the subject of intensive investigation. Unfortunately, the selectivity of antagonists for receptor subtypes rarely exceeds tenfold. In the absence of pharmacological uniqueness, knockout mice are proving exceptionally useful in delineating functions associated with specific receptor subtypes.

Our understanding of muscarinic receptor biology is largely dependent on (1) physiological and behavioral analysis of the actions of muscarinic agonists and antagonists (notably, smooth muscle contraction, cardiac function, vascular tone, glandular secretion, arousal, attention, and memory); (2) ligand binding studies employing radiolabeled probes (notably, the antagonists [^{3}H]*N*-methylscopolamine and [^{3}H]quinuclidinyl benzilate); (3) biochemical characterization of receptor proteins (including primary, secondary, and tertiary protein structure, and patterns of phosphorylation, glycosylation, phosphorylation, and lipid modification, dimerization, nitrosylation, and internalization); (4) measurement of receptor influence on intracellular second messengers (notably, inositol triphosphate (IP_3), diacylglycerol, Ca^{2+}, nitric oxide, cAMP, cGMP, and arachidonic acid); (5) analysis of receptor gene sequence and control; (6) genetic manipulation of receptor expression; (7) genetic manipulation of receptor protein structure; and (8) elimination of specific receptor subtypes using gene knockout technology.

The properties of the different subtypes are summarized in Table 1. A schematic diagram of the human M2 receptor is presented in Figure 2, a possible arrangement of the seven transmembrane domains is schematically represented in Figure 3, the amino acid sequence alignment of the five subtypes is presented in Figure 4, and an alignment tree of receptor amino acid sequences suggesting evolutionary relations is shown in Figure 5. The receptors were named on the basis of the order of their discovery, which reflects their relative abundance. Muscarinic receptors have been traditionally divided into two groups based on both their pharmacology and signaling properties: M2 and M4 comprise one group, and M1, M3, and M5 comprise the other (Figure 6).

Muscarinic Receptor Structure

Muscarinic receptor proteins are single polypeptides of 460–590 amino acids with an extracellular N-terminus and an intracellular C-terminus (Figures 2 and 3). The N-termini do not possess consensus sites for protein cleavage, and epitope tags applied to the N-terminus provide convenient markers to monitor receptor expression and transport.

The coding regions of muscarinic receptor genes are contained within a single exon. Hydropathic analyses of the amino acid sequences reveal seven regions of 20–24 amino acids that are likely to form membrane spanning α-helical structures. The amino acid composition of the membrane-spanning regions is

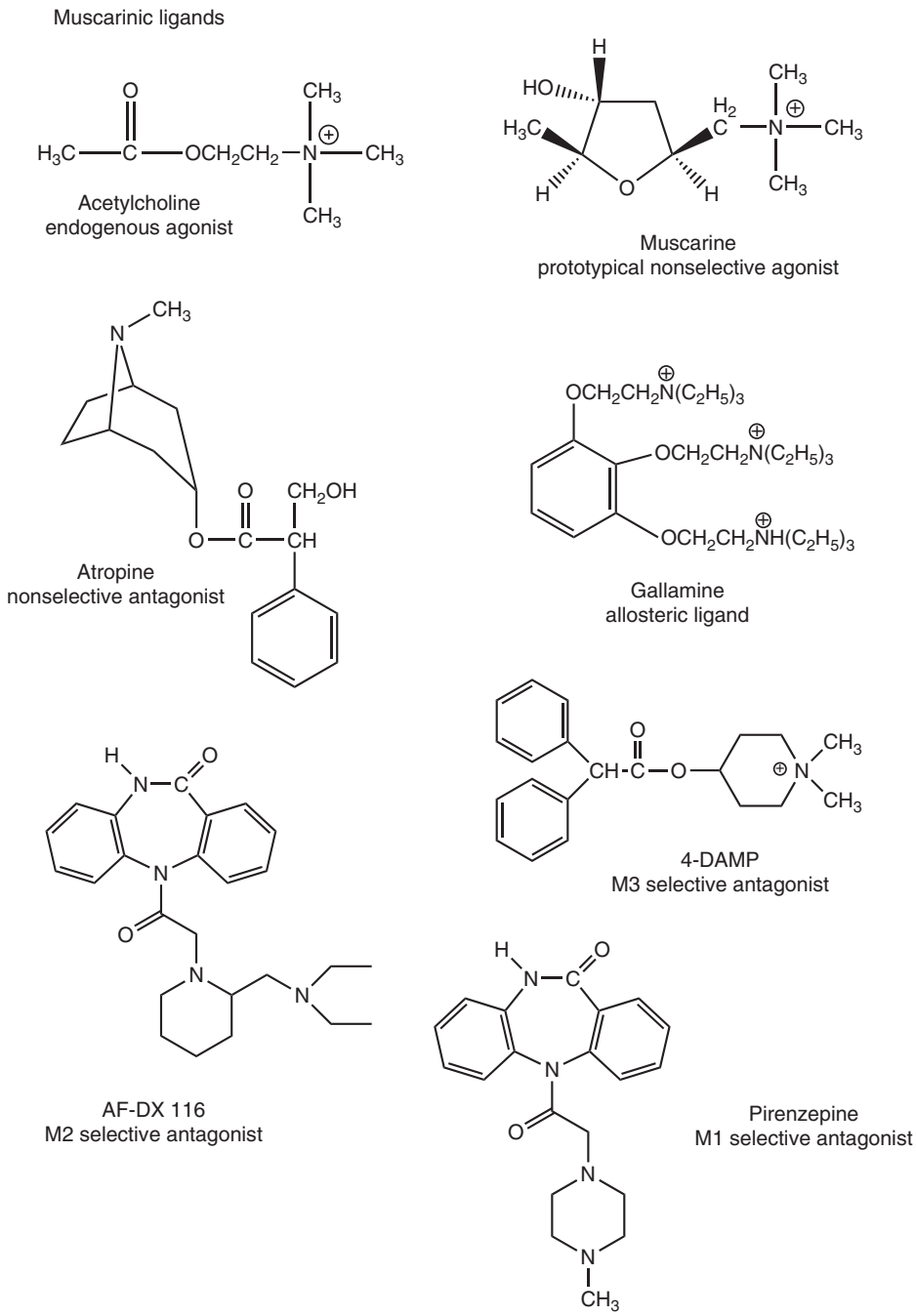

Figure 1 Structures of muscarinic ligands. Acetylcholine is the physiological agonist. Muscarine and atropine are the prototypical agonist and antagonist which define the receptor class. Gallamine is an allosteric receptor antagonist. Pirenzepine, AF-DX 116, and 4-DAMP are antagonists with a degree of selectivity for the M1, M2, and M3 receptor subtypes, respectively.

highly conserved (90% sequence similarity) among the five subtypes, as it is among the larger family of G-protein-coupled receptors.

Between the fifth and sixth membrane-spanning units is a large intracellular loop that is highly variable in composition and size. Several consensus sites for phosphorylation are located on the third intracellular loop, as well as on the C-terminal chain. A disulfide bridge is formed between a conserved cysteine adjacent to the third transmembrane segment and a cysteine in the middle of the second extracellular loop in all five subtypes. Chemical modification of these cysteine groups decreases agonist binding affinity as well as the ability of the receptor to couple to transducer

Table 1 Properties of muscarinic receptor subtypes

Property	*Subtype*				
	M1	*M2*	*M3*	*M4*	*M5*
Molecular weight[a]	51 240	51 715	66 127	53 058	60 186
Amino acids	460	466	590	479	532
Genomic location	11q13	7q31–q35	1q43	11p12–p11.2	15q26
G-protein coupling	G_q, G_{11}	G_i, G_o	G_q, G_{11}	G_i, G_o	G_q, G_{11}
Second messengers[b]	IP_3/DAG, NO	cAMP (↓)	IP_3/DAG, NO	cAMP (↓)	IP_3/DAG, NO
Tissues[c]	Brain, autonomic ganglia, secretory glands, vas deferens, sympathetic ganglia	Brain, heart, sympathetic ganglia, lung, ileum, uterus and other smooth muscles	Brain, secretory glands, smooth muscles	Brain, lung	
Brain regions	Cerebral cortex, hippocampus and dentate gyrus, striatum, olfactory bulb and tubercle, amygdala	Cerebellum, medulla, pons, basal forebrain, olfactory bulb, diencephalon	Cerebral cortex, thalamus, piriform cortex, olfactory bulb, brain stem nuclei	Occipital cortex, caudate and putamen, visual nuclei, olfactory tubercle, hippocampus	Midbrain
Subtype-selective antagonists[d]	Pirenzepine, telenzepine, mamba venom toxins	AF-DX-116, methoctramine, gallamine, himbacine	4-DAMP, hexahydro-sila-difenidol	Tropicamide, himbacine	
Functions revealed by analysis of knockout mice	Agonist-induced seizures, learning, control of locomotor activity; indirect inhibition of N- and L-type calcium channels; M current inhibition in sympathetic ganglion neurons; MAPK pathway activation	Agonist-induced akinesia and central tremor; corticosterone release; analgesia; hypothermia; smooth muscle contraction; bradycardia; inhibition of N- and P/Q-type calcium channels	Papillary and urinary bladder constriction; salivary secretion; smooth muscle contraction; control of dopamine release; maintenance of body mass and food intake	Dopamine release; acetylcholine release (hippocampus); depression of motor control; analgesia	Dilation of cerebral arteries and arterioles; dopamine release

[a]Molecular weights are calculated for the polypeptide contribution only.
[b]IP_3/DAG, stimulation of phospholipase C to release the second messenger inositol triphosphate (IP_3) and diacylglycerol (DAG) from phosphatidylinositol; NO, increase in nitric oxide production. cAMP(↓): inhibition of adenylate cyclase, thereby lowering cellular content of the second messenger cAMP.
[c]Only low levels of M5 receptor expression have been detected in tissues.
[d]Antagonist selectivity of the subtypes is relative; antagonists generally differ in their interactions with the different subtypes by a factor of less than 20.

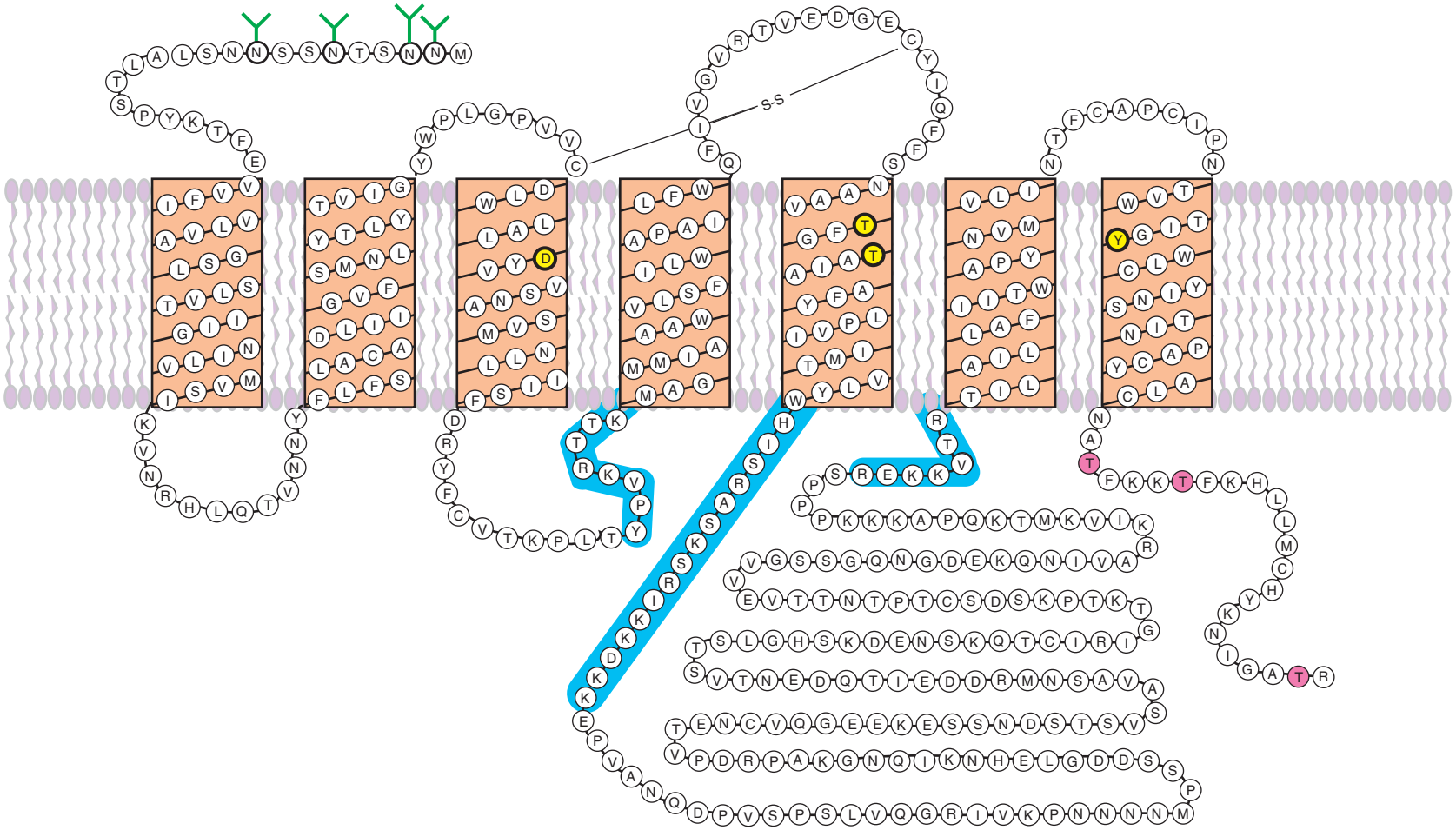

Figure 2 Schematic diagram of a human M2 muscarinic receptor showing the primary amino acid sequence. Four consensus sites of glycosylation at asparagine residues are indicated on the extracellular N-terminal domain. The aspartate and tyrosine and threonine residues in the membrane-spanning regions that are thought to be involved in agonist binding are highlighted in yellow. Regions of the third intracellular loop involved in G-protein recognition and coupling are indicated by blue shading. Possible sites of threonine phosphorylation in the C-terminal domain are also indicated by highlighting.

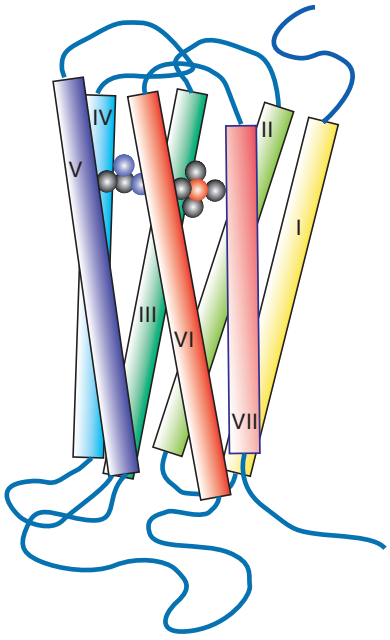

Figure 3 Possible arrangement of muscarinic receptor transmembrane domains, as suggested by structural analysis of rhodopsin. Amino acid moieties involved in the binding of acetylcholine are located approximately one-fourth of the way into the plane of the membrane.

G-proteins. Moreover, nitrosylation of receptor cysteines after reduction of the disulfide bond severely diminishes ligand binding to the receptor. Elimination of either of these cysteines, however, precludes expression of the receptor. A conserved cysteine in the N-terminus may be a site for palmitoylation (thereby creating another intracellular loop) and an N-terminal polybasic region in four of the subtypes (all except the M2 receptor) may be involved in recruitment of adapter proteins in certain signaling pathways.

M3 receptors (at least) have been found to form disulfide-linked dimers on the cell surface. Two to four consensus sites for N-glycosylation are present in the extracellular N-terminal domain of each receptor, and 25% of the receptor mass may be contributed by these oligosaccharides. Although this carbohydrate component may affect the processing and orientation of the receptor, it plays no discernible role in ligand binding or signal transduction: enzymatic removal of the carbohydrate has little effect on the binding and signal transduction potential of mature receptors.

Chimeric receptors and mutational analysis have revealed sites on the receptor proteins that are specifically involved in ligand binding and coupling to transducer G-proteins. Acetylcholine binds to a site within a pocket formed by the roughly circularly arranged transmembrane domains (Figure 3). An aspartate moiety in the third transmembrane region participates in an ionic interaction with the quaternary nitrogen of acetylcholine, whereas a series of tyrosine and threonine residues located in the membrane-spanning segments approximately one-fourth of the distance through the membrane likely form hydrogen bonds with muscarinic ligands. In agreement with classical pharmacological analysis, the binding site for competitive antagonists is thought to overlap the acetylcholine recognition site but additionally involves contiguous hydrophobic areas of the receptor protein and membrane. Muscarinic receptors also possess a site(s) or region that participates in allosteric regulation by a variety of compounds, including gallamine (Figure 1). These ligands compete with classical muscarinic ligands but slow the dissociation of previously bound probes, indicating the existence of a ternary complex (receptor plus allosteric and orthosteric ligands). Gallamine binding is sensitive to modification of the sixth transmembrane domain as well as the third extracellular loop.

A number of regions have been identified that participate in the interactions of muscarinic receptors with transducer G-proteins. In particular, the second intracellular loop and the N- and C-terminal portions of the third intracellular loop are involved in coupling to G-proteins. Desensitization of muscarinic receptors may entail phosphorylation of conserved threonine residues on the C-terminal portion of the receptor, as well as other sites on the third intracellular loop. Prolonged stimulation of muscarinic receptors induces a rapid phosphorylation that is associated with a decrease in agonist affinity and an uncoupling from transducer G-proteins. Agonist-induced phosphorylation involves a G-protein receptor kinase (GIRK) and not kinases activated by Ca^{2+}, cAMP, cGMP, or phorbol esters. Phosphorylation-dependent internalization of muscarinic receptors appears to be mediated by arrestins.

The evolutionary relationship of muscarinic receptor subtypes is suggested by amino acid sequence (Figure 5). M1, M3, and M5 comprise one cluster, with M1 and M3 being somewhat more closely related. M2 and M4 comprise a second cluster. Analysis of nucleotide sequence of the coding regions reveals an essentially similar pattern. Sequence analyses reveal more genomic divergence within the two clusters than between the two groups. The five muscarinic receptor genes are scattered throughout the genome; the only two on the same chromosome (M1 and M4 on chromosome 11) are only distantly related.

Muscarinic Receptor Gene Structure

Whereas the coding sequences of the gene for each subtype are uninterrupted or intronless, the genes are

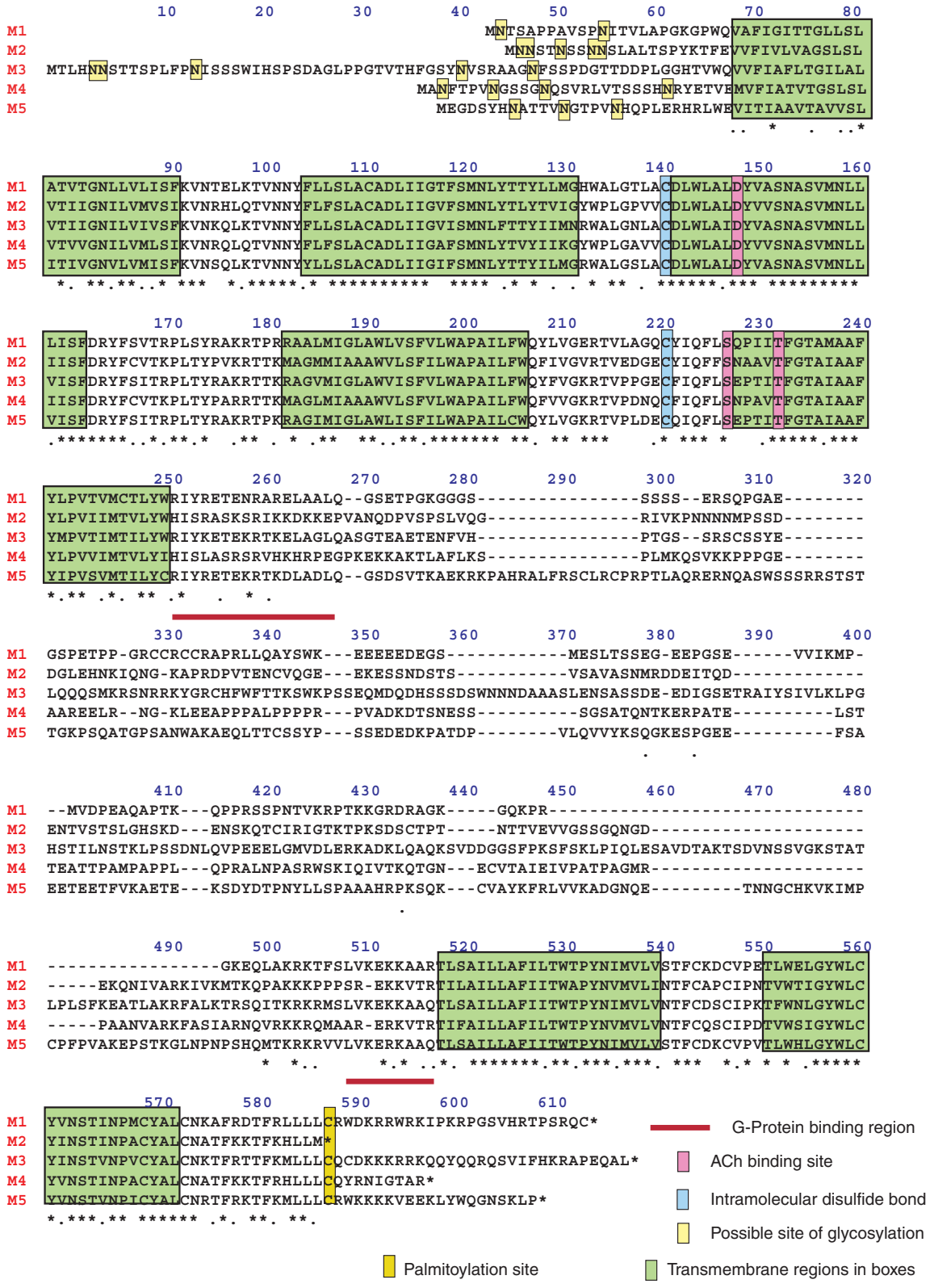

Figure 4 Amino acid sequence alignment of the five human muscarinic acetylcholine receptor subtypes. Conserved and semiconserved residues are indicated by the asterisks and dots below the sequences; possible N-terminal glycosylation sites are indicated by yellow shading; the conserved cysteines involved in disulfide bond formation are indicated by blue shading; the N-terminal site of palmitoylation is highlighted; sequences believed to be involved in receptor interactions with G-proteins are indicated by red bars; transmembrane regions are identified by the green shading. Alignment was performed using the ClustalW procedure provided in MacVector.

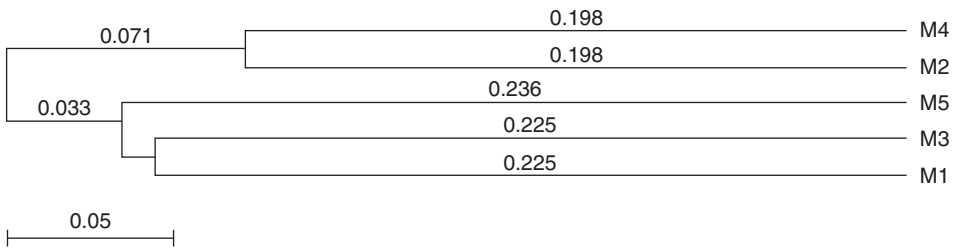

Figure 5 UPGMA distance tree illustrating the evolutionary relationship of the muscarinic receptor subtypes revealed by amino acid sequence. M1, M3, and M5 comprise a cluster; M2 and M4 comprise a second cluster. Analysis of nucleotide sequence of the coding regions reveals an essentially similar pattern. The 0.05 scale marker is approximately equivalent to a 5% difference in sequence.

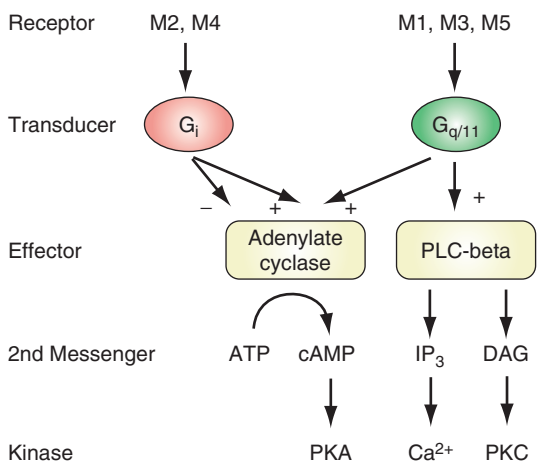

Figure 6 Muscarinic receptor signaling pathways. The initial stages of signal transduction include receptor, transducer (G-protein), effector (typically an enzyme or ion channel), second messenger production (e.g., Ca^{2+}, cAMP, cGMP, DAG, and IP_3), and protein kinase activation. It has long been appreciated that M2 and M4 receptors inhibit adenylate cyclase through activation of G_i, whereas M1, M3, and M5 receptors stimulate phospholipase C-beta through activation of $G_{q/11}$. In fact, muscarinic signaling involves an impressive variety of other pathways. For instance, under the right conditions all receptors can stimulate adenylate cyclase through the activation of G_s. Other pathways are activated directly or secondarily. PKA, protein kinase A; PKC, protein kinase C; IP_3, inositol triphosphate; DAG, diacylglycerol.

complex, spanning large distances, and often contain multiple exons that result in multiple splice variants that are under the control of tissue-specific promoters. TATA and CAAT boxes are not involved in the transcriptional control of muscarinic receptor gene expression. The promoter structures of several muscarinic acetylcholine receptor subtypes have been analyzed. Consensus regulatory sequences, including binding sites for SP1, AP-1, and AP-2, as well as tissue-specific silencer elements, have been identified. All muscarinic receptor genes contain a large (10–26 kb) intron close to the 5′ end of the translational start site. The relatively simple M1 gene contains an upstream 657 bp exon with a flanking 900 bp promoter region that contains consensus sites for SP1, NZF-1, AP-1, AP-2, E-box, NF-κB, and Oct-1. The M2 gene spans over 126 kb on chromosome 7, and it contains four exons that result in at least eight different splice variants. The M2 gene has at least three distinct promoters that result in distinct tissue-specific expression, and these promoters are highly conserved between mammalian species. The M3 gene spans at least 285 kb encompassing eight exons (the coding sequence is in exon 8). Multiple AP-2 regulatory sequences are located in the 1.2-bp 5′-flanking promoter region, and the transcription factor AP-2α increases M3 expression.

Muscarinic gene sequences tend to be highly conserved; relatively few nonsynonymous polymorphisms have been reported.

Physiological Functions

Muscarinic receptors subserve a number of physiological functions. Muscarinic receptors are present in autonomic ganglia and on the target organs of postganglionic parasympathetic fibers. Thus, muscarinic receptors mediate such parasympathetic effects as smooth muscle contraction, blood vessel dilation, decreases in heart rate and cardiac contractility, and glandular secretion.

Muscarinic agonists have relatively little use in therapeutics, being limited to stimulating muscles of the gastrointestinal tract and bladder and to lowering intraocular pressure. A great deal of work has focused on developing M1 subtype-specific agonists that might be useful in alleviating the cholinergic losses that characterize Alzheimer's disease. It has been sporadically proposed that subtype-selective cholinomimetics would be useful in the treatment of psychiatric diseases. Arecoline, an alkaloid found in the betel nuts that are chewed by a significant portion of the world's population, is undoubtedly the most widely consumed muscarinic agonist. Arecoline causes parasympathetic stimulation (notably increased salivation) and produces euphoria and enhanced cognitive function.

Muscarinic antagonists are used clinically to inhibit parasympathetic activity in a variety of situations. For example, antimuscarinic drugs have been used in anesthesia (to reduce secretions), parkinsonism, overactive bladder, bradycardia, and irritable bowel syndrome.

In the central nervous system, cholinergic fibers arise in nuclei in the midbrain and pons to diffusely innervate muscarinic receptors on neurons in the thalamus and other diencephalic structures, as well as the reticular formation and other brain stem and cranial nerve nuclei. A second major projection system arises from the basal forebrain cholinergic system (which includes the medial sepal nucleus, diagonal band nuclei, and nucleus basalis) and innervates muscarinic receptors on neurons throughout the telencephalon. Cholinergic interneurons innervating muscarinic receptors are located in several brain structures, including the cerebral cortex, nucleus accumbens, and striatum. Muscarinic receptor density is particularly high in the striatum, cerebral cortex, hippocampus, and other specific nuclei. Muscarinic receptor subtypes display striking regional and laminar heterogeneity (Table 1). Central muscarinic systems are prominently involved in arousal, attention, learning, and memory. Other functions involving central muscarinic actions include movement, analgesia, and temperature regulation.

Subsets of M2 and M4 receptors have a presynaptic localization and play a role in the control of neurotransmitter release; the other receptors are predominantly postsynaptic.

Analysis of knockout mice lacking expression of specific receptor subtypes has shed considerable light on the physiological functions of receptor subtypes (Table 1). Muscarinic knockout mice are both healthy and fertile; recognition of deficits frequently requires specific focused analysis. These studies reveal that M1 receptors are involved in M current (potassium channel) regulation, agonist-induced seizures mitogen-activated protein kinase (MAPK) activation, certain learning paradigms, and inhibition of a slow, voltage-independent calcium current. M2 receptors are involved in the production of bradycardia; smooth muscle contraction (stomach, urinary bladder, and trachea); and central tremor, analgesia, and hypothermia. M3 receptors are seen to be involved in salivary secretion, pupillary constriction, and bladder detrusor contraction. M4 receptors are seen to be involved in limiting locomotor activity (including modulation of motor responses to dopamine). There is an interesting relationship between M5 receptors and dopamine neurotransmission: M5 receptors facilitate dopamine release, thereby affecting brain reward systems. M5 receptors also appear to play a specific role in the dilation of cerebral arteries and arterioles.

Muscarinic Signal Transduction

Muscarinic receptors alter the activity of receptive cells by a number of different signal transduction pathways (Figure 6). The biochemical pathways activated depend on the nature and quantity of the receptor subtype, transducer proteins, effector molecules, and protein kinase substrates expressed in the tissue, as well as the potential for cross-talk between the various transduction pathways. Conventional wisdom holds that the odd-numbered receptor subtypes – M1, M3, and M5 – efficiently couple through pertussis toxin-insensitive G-proteins of the $G_{q/11}$ family to a stimulation of phospholipase C (β subtype). Phospholipase C releases the second messengers diacylglycerol and IP_3 from the membrane phospholipid, phosphatidylinositol-4, 5-bisphosphate. Diacylglycerol activates protein kinase C, whereas IP_3 releases Ca^{2+} from intracellular stores by acting on specific receptors on the endoplasmic reticulum. M1, M3, and M5 also activate the MAP kinase pathway as well as phospholipase D (releasing arachidonic acid). Cross-talk between signaling pathways makes the distinction between proximal and indirect signaling events problematic.

M2 and M4, inhibit adenylate cyclase through actions involving G_i (any of the three G_i proteins may be involved). A few, but by no means all, isoforms of mammalian adenylate cyclase are inhibited by $G_{i\alpha}$ subunits (e.g., AC5 and AC6). An even larger number of adenylate cyclase subunits are inhibited by $G_{\beta\gamma}$ dimers. Thus, the molecular basis for inhibition of the enzyme is not always obvious. Inactivation of G_i by ADP-ribosylation catalyzed by pertussis toxin uncovers activation of adenylate cyclase by all receptor subtypes. In the case of M3 at least, this activation seems to be mediated by a muscarinic receptor–G_s interaction.

Thus, our initial binary classification of muscarinic signaling paradigms (i.e., receptor:transducer:effector = M1/M3/M5:G_s:phospholipase C_β or M2/M4:G_i:adenylate cyclase) has been extended in response to the identification of pathways that involve additional transducer proteins as well as an appreciation of secondary effects and cross-talk between the pathways. Other well-documented muscarinic actions include stimulation of adenylate cyclase, inhibition of phosphodiesterase, stimulation of phospholipase A2 (M1, M3, and M5), and stimulation of phospholipase D (M1 and M3).

Muscarinic receptor stimulation activates a number of depolarizing and hyperpolarizing currents through both direct and indirect mechanisms. Prominent effects include (1) stimulation of inwardly rectifying potassium conductances by M2 and M4 receptors; (2) activation of calcium-dependent potassium, chloride,

and cation conductances by M1, M3, and M5 receptors; (3) inhibition of the M current, a voltage- and time-dependent potassium conductance, by M1 and M3 receptors; and (4) inhibition of a calcium conductance by M2 and M4 receptors, directly via G_o and indirectly via reductions in cAMP.

Other novel biochemical effects of muscarinic receptor activation have been described. This signaling heterogeneity may reflect (1) multiple isoforms of signal transduction effectors (e.g., at least 11 isoforms of adenylate cyclase), (2) cross-talk between transduction pathways due to a lack of specificity or the presence of indirect and secondary effects, or (3) cell type-specific differences due to expression of different assortments of signal transduction molecules. Muscarinic receptor-mediated activation of both phosphodiesterase and nitric oxide synthetase has been reported.

Muscarinic receptors have substantial mitogenic potency. The molecular details of muscarinic receptors of the pathways that mediate these effects are incompletely understood. The activities of multiple kinases and phosphatases are affected by muscarinic receptors under specific conditions, including Raf, tyrosine kinases, MAPK, and phosphatidyl inositol-3-kinase. The activities of multiple small GTPases, including Ras and Rho, are also affected by muscarinic receptor activation.

Muscarinic signaling is subject to regulation by agonist-induced phosphorylation and downregulation (internalization). Mice deficient in G-protein-coupled receptor kinase-5 display muscarinic supersensitivity. Although domains located in the third intracellular loop are prominently involved in receptor phosphorylation and downregulation, a number of subtype-dependent variations in arrestin association and internalization pathways have been described.

Muscarinic receptors regulate the expression of a variety of genes, notably a variety of transcription factors and signaling factors. An appreciation of this regulation promises to increase our understanding of muscarinic physiology.

See also: Acetylcholine Neurotransmission in CNS; Basal Ganglia: Acetylcholine Interactions and Behavior; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Nicotinic Receptors: Autonomic Neurons; Parasympathetic Nervous System.

Further Reading

Ashkenazi A and Peralta EG (1994) Muscarinic acetylcholine receptors. In: Peroutka SJ (ed.) *Handbook of Receptors and Channels. G Protein-Coupled Receptors*, pp. 1–27. Boca Raton, FL: CRC Press.

Beltran B, Orsi A, Clementi E, and Moncada S (2000) Oxidative stress and S-nitrosylation of proteins in cells. *British Journal of Pharmacology* 29: 953–960.

Birdsall NJ and Lazareno S (2005) Allosterism at muscarinic receptors: Ligands and mechanisms. *Mini-Reviews in Medicinal Chemistry* 5: 523–543.

Brann MR, Klimkowski VJ, and Ellis J (1993) Structure/function relationships of muscarinic acetylcholine receptors. *Life Sciences* 52: 405–412.

Buckley NJ, Bachfischer U, Canut M, et al. (1999) Repression and activation of muscarinic receptor genes. *Life Sciences* 64: 495–499.

Eglen RM (2005) Muscarinic receptor subtype pharmacology and physiology. *Progress in Medicinal Chemistry* 43: 105–136.

Felder CC (1995) Muscarinic acetylcholine receptors: Signal transduction through multiple effectors. *FASEB Journal* 9: 619–625.

Hulme EC, Birdsall NJM, and Buckley NJ (1990) Muscarinic receptor subtypes. *Annual Review of Pharmacology and Toxicology* 30: 633–673.

Ma W, Li BS, Zhang L, and Pant HC (2004) Signaling cascades implicated in muscarinic regulation of proliferation of neural stem and progenitor cells. *Drug News and Perspectives* 17: 258–266.

Nadler LS, Rosoff ML, Hamilton SE, et al. (1999) Molecular analysis of the regulation of muscarinic receptor expression and function. *Life Sciences* 64: 375–379.

Nathanson NM (2000) A multiplicity of muscarinic mechanisms: Enough signaling pathways to take your breath away. *Proceedings of the National Academy of Sciences of the United States of America* 97: 6245–6247.

Saffen D, Mieda M, Okamura M, and Haga T (1999) Control elements of muscarinic receptor gene expression. *Life Sciences* 64: 479–486.

Tobin AB and Budd DC (2003) The anti-apoptotic response of the $G_{q/11}$-coupled muscarinic receptor family. *Biochemical Society Transactions* 31: 1182–1185.

von der Kammer H, Demiralay C, Andresen B, et al. (2001) Regulation of gene expression by muscarinic acetylcholine receptors. *Biochemical Society Transactions* 67: 131–140.

Watson S and Arkinstall S (1994) *The G-Protein Linked Receptor Facts Book*. London: Academic Press.

Wess J (1993) Molecular basis of muscarinic acetylcholine receptor function. *Trends in Pharmacological Sciences* 14: 308–313.

Wess J (2004) Muscarinic acetylcholine receptor knockout mice: Novel phenotypes and clinical implications. *Annual Review of Pharmacology and Toxicology* 44: 423–450.

Relevant Websites

http://www.alomone.com – Alomone Labs.
http://www.calbiochem.com – Calbiochem.
http://www.gpcr.org – GPCRDB: Information system for G-protein-coupled receptors.
http://www.scbt.com – Santa Cruz Biotechnology.
http://www.sigma-aldrich.com – Sigma–Aldrich.
http://stke.sciencemag.org – Signal transduction knowledge environment.
http://www.cdna.org – University of Missouri – Rolla cDNA Resource Center.

Music

P J Donnelly and C J Limb, The Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Music is a phenomenon that appears to have been universally present throughout human history and across all human populations. The perception and performance of music are often delicately intertwined with a variety of perceptual (e.g., listening), sensorimotor (e.g., playing an instrument), visual (e.g., reading a musical score), mnemonic (e.g., playing a piece from memory), and affective (e.g., emotional interpretation) processes. For these reasons, music offers potentially powerful insights into complex brain organization and function. Additionally, musical aptitude and the unique effects of musical training are of specific interest to the study of the nature and effects of brain plasticity and reorganization.

Music is often compared to language, and indeed music does share many abstract similarities with language. In both language and music, perceptually distinct elements are combined into hierarchically organized sequences according to rules of syntax. While music and language are highly analogous in many respects, however, they are also quite different. Therefore the cognitive investigations of music perception provide the opportunity to reveal similarities and differences in the neural structures and functions involved in music and language perception. Furthermore, the search for the cognitive substrates of music perception complements the musicological investigations of the origins of music, its perceptual distinction from speech and other auditory input, and its evolutionary relationship to the development of language.

The past few decades have seen a major increase in the systematic study of the cognitive neuroscience of music. Recent advances in neuroscientific techniques, such as functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and positron emission tomography (PET), provide new data to build upon findings from neurophysiologic and lesion-based studies. Because of the breadth of processes and brain regions involved in musical perception, the cognitive investigations of music perception traditionally dissect music into its distinct fundamental elements.

Pitch and Melody

Pitch is the fundamental building block of music. For this reason, the majority of research on the neurological correlates of music perception has focused on pitch and, to a lesser extent, on melodic phrase. The perception of pitch relies on the acoustic properties of the spectra (the ratios of the fundamental frequency to the harmonics) and the envelope (the rise and decay time) of each harmonic constituent. Melody is a musical phrase created when a series of pitches are organized into sequential temporal patterns of varying musical contour and interval.

Numerous studies spanning nearly 50 years strongly indicate a 'pitch center' in the posterior superior temporal lobe, composed of the auditory cortices and secondary auditory association areas. Early studies of patients with temporal lobe excisions note the prominence of the right hemisphere in pitch and melodic processing. Excisions that included the right lateral Heschl's gyrus significantly impaired the perception of changes in pitch, pitch direction, interval size, melodic contour, and complex tones. However, left-lateralized damage can also cause selective deficits in pitch and melodic processing. The results of a neuroimaging study using PET indicate a left lateral specialization of the core auditory cortex for temporal processing of sound, while the spectral processing of pitch correlated to activity in the right hemisphere belt cortical areas (Figure 1).

While considerable debate surrounds possible hemispheric specialization in pitch perception, recent studies have begun to investigate specialization of areas within the auditory cortices. Studies of primates have demonstrated the importance of the secondary auditory cortex in the tonotopic organization of pitch (Figure 2). Recent neuroimaging studies of humans have also implicated the importance of the secondary auditory association centers in complex pitch processing. When individuals were examined while listening to sequentially varied pitches – that is, a melody – increased activity was found beyond Heschl's gyrus in the superior temporal gyrus and planum polare. Research in music perception implicates possible dissociation between the perception of pitch height (the octave of a pitch) and pitch chroma (the note within the octave). A recent fMRI study offers a potential neural correlate, indicating that regions anterior to the primary auditory cortex are important for pitch

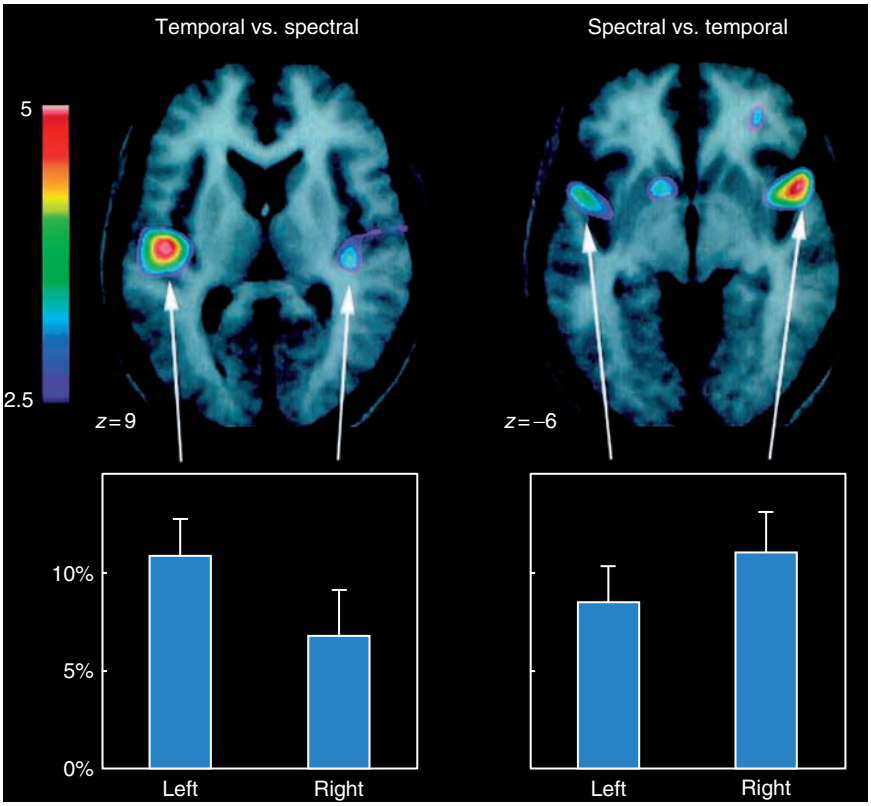

Figure 1 Merged positron emission tomography and magnetic resonance image volumes corresponding to the direct comparison of temporal and spectral conditions to one another. The image on the left shows a horizontal section taken through the region of Heschl's gyri, which showed significantly greater activity in the combined temporal conditions than in the combined spectral conditions. The section on the right is taken through the maxima corresponding to the anterior superior temporal region, which showed a greater response to the spectral conditions than to the temporal conditions. The bar graphs illustrate the percent difference between condition in regions of interest taken from corresponding locations. Reprinted from Zatorre RJ and Belin P (2001) Spectral and temporal processing in human auditory cortex. *Cerebral Cortex* 11: 949, with permission from Oxford University Press.

chroma while the posterior regions are important in the perception of pitch height.

Absolute Pitch

Absolute pitch perception is of specific relevance to the discussion comparing the effects of musical training, musical exposure, and innate musical predisposition. Absolute pitch is the ability to identify a musical note by pitch without reference to any other sound. Genetic factors are implicated, since absolute pitch is exceedingly rare – even among musicians, it has a sibling recurrence rate of 8–15%, and seems to be distributed differently across human populations. However, environmental factors are clearly important, since absolute pitch requires musical training before the age of 12 years in order to manifest itself.

Through neuroimaging studies, the functional and structural correlates of absolute pitch are beginning to be identified. Investigations of the structural correlates of absolute pitch show that musicians with absolute pitch show a greater leftward asymmetry of the planum temporale than do nonmusicians or musicians without absolute pitch. A follow-up study indicated that this asymmetry results not from an enhancement of the left superior temporal cortex, but rather from a structural reduction of the right planum temporale in those with absolute pitch. While these results suggest that a developmental 'pruning effect' may be needed for acquisition of absolute pitch, the possible relationship between innate brain structure and predisposition for absolute pitch needs further exploration.

A functional neuroimaging study using PET demonstrated that musicians without absolute pitch show greater activity in the right inferior frontal cortex, a region important for working memory for pitch, when making judgments of musical intervals compared to those with absolute pitch. Additionally, absolute pitch possessors show greater activation in the left posterior dorsolateral frontal cortex, a region associated with conditional associations in memory.

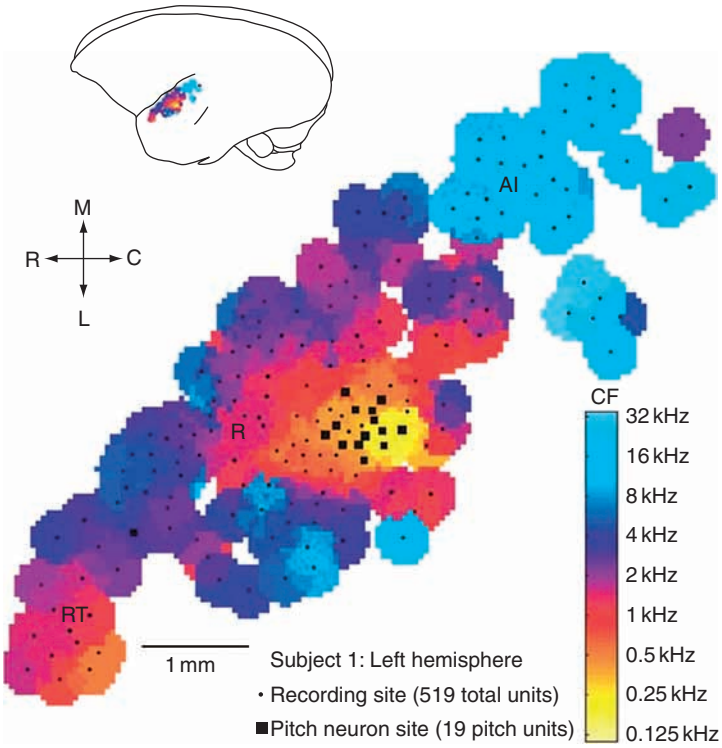

Figure 2 Characteristic frequency (CF) topographical map from the left hemisphere of one marmoset. Pitch-selective neurons (black squares) were found clustered near the anterolateral border of primary auditory cortex (AI). Frequency reversals indicate the borders between the AI/R and R/RT (rostral temporal) fields (R, rostral; C, caudal; L, lateral; M, medial). Reprinted from Bendor D and Wang X (2005) The neuronal representation of pitch in primate auditory cortex. *Nature* 436: 1163, with permission from Macmillan Publishers Ltd. Copyright (2005).

This region was active in all individuals in tasks requiring the identification of musical intervals, but was uniquely active in those with absolute pitch when listening to single tones, possibly correlating mnemonic association with the ability to identify pitch without relative reference.

Timbre

Another fundamental aspect of music is timbre, or tone color, the psychoacoustic property of sound that permits us to differentiate between two different musical instruments playing at the same pitch. The perception of timbre relies on the acoustic properties of the spectra (harmonics) and timing (envelope) and therefore the cognition of timbre relies upon the neural regions important in the processing of pitch.

Imaging studies focusing on the alteration of spectrally dissimilar sounds demonstrate activity in the superior temporal lobes, especially Heschl's gyrus and the superior temporal sulcus, with a degree of right lateralization. These findings are reinforced by data from patients who have undergone temporal lobectomy. Patients with right temporal lobe excisions show deficits in the discrimination of changes in timbre compared to left temporal lobe excision and normal control groups. Since the perception of timbre relies upon the fine spectral and temporal aspects of sound, deficits in timbral perception are usually described in conjunction with deficits in pitch perception. However, in one particularly unique case, a patient, following a right temporal stroke, lost the ability to discriminate between the timbre of previously familiar musical instruments while retaining normal perception of melody, rhythm, and other musical aspects.

Rhythm

The term 'rhythm' is often used to describe broad macroscopic temporal patterns in music that encompass the structural musical elements of pulse, meter, and rhythmic grouping. Pulse or beat is the periodic repetition of acoustic stimuli in time, such as the unaccented 'ticks' of the metronome. The speed of the underlying pulse of music defines the tempo of a piece of music, often measured in 'beats per minute.'

Meter is defined by regularly recurring patterns of stressed ('strong') and unstressed ('weak') beats, and in some notated music is regulated by the presence of a time signature. A more precise definition of rhythm frames it as the subjective perception of groupings of accented and unaccented beats, which vary in length and relationship over time.

Because rhythm develops over time, the perception of rhythmic patterns involves brain regions important in temporal processing. In music, meter and rhythm are often intimately linked with movement, such as in dance or the urge to tap your foot to a beat. The production of a timed motor response from an external stimulus involves a network that includes the sensorimotor cortices, superior temporal gyrus, cerebellum, and basal ganglia. Less well understood are the correlates of temporal perception without motoric reproduction, such as in tasks of passive rhythmic perception.

Early and contradicting lesion studies implicated both the left and right temporal cortices in the process of rhythmic sequences. More recent studies implicate a possible divergence in the perception of metric and rhythmic groupings. Damage to the temporal lobe in the right hemisphere impairs perception of meter, whereas left-lateralized damage impairs the discrimination of rhythmic patterns; however, such differences may only reflect the subtle differences in task and stimuli.

The specialization and complex interactions of the left and right hemispheres in the perception of temporal structures are not well understood. One fMRI study indicates separate processes in the cognition of metric (musical) rhythms and nonmetric (nonmusical) rhythms. Musical rhythms causes activation in the left premotor and left parietal areas and the right anterior cerebellar lobe, while nonmusical rhythms leads to activity in the right prefrontal, right premotor, and right parietal areas as well as in the bilateral cerebellar posterior lobe.

Further evidence of a leftward dominance for temporal processing implicates the recruitment of the 'language areas' in the perception of rhythmic patterns. Imaging studies of passive rhythmic perception show increased activation in the left inferior frontal gyrus, including Broca's area (BA 44/5) and BA 47. A case study of a subject with Wernicke's aphasia demonstrated a severe deficit in the perception of rhythm compared to pitch and melodic perception, indicating the importance of Wernicke's area (BA 22) in rhythmic perception. Additionally, musicians show increased activity in the left hemisphere in the passive perception of rhythm compared to nonmusician, suggesting that hemispheric specialization for rhythm perception can be enhanced by musical training.

Musical Structure

Music, much like language, is governed by rules and syntax (grammar). In Western tonal music, these rules are governed by a metric structure, a hierarchy of musical keys, the order of chord progressions (tonality), and rules of harmony (counterpoint). Only very recently have researchers begun to investigate the neurological correlates of the perception of musical structure and syntax. The majority of studies have focused on identifying the effects of presenting syntactically irregular chords within a structured harmonic sequence.

Converging data from neuroimaging and electroencephalographic studies indicate that violations of tonality activate the inferior frontal regions, including Broca's area (BA 44/5) and its homolog in the right hemisphere and BA 47, as well as the posterior temporal regions, including Wernicke's area. Since both Broca's and Wernicke's areas are crucial in the processing of language, this evidence indicates a significant overlap of these 'language areas' in the syntactic processing of both language and music. However, musical syntax involves a complex set of rules and harmonic relationships that extend beyond simple expectation of correct or incorrect chords. Isolating brain regions involved in this syntactic processing is rather difficult, especially since the correlates of the basic elements of music perception, such as pitch and rhythm, are not fully understood. For instance, a specific musical key (e.g., C major or G minor) determines both the pitch content (musical scale) and chords (harmony) permitted in a tonal musical phrase. In Western tonality, the various 24 different major and minor keys are cyclically interrelated based on a hierarchy prescribed in the musical 'circle of fifths,' and this hierarchy regulates the perceptual distance of musical keys and chords. An fMRI study designed to investigate the perception of melodic contour as a function of tonality (modulated through all 24 musical keys) implicates the rostromedial prefrontal cortex as a unique region that maintains a relative topographic map of tonality. These findings suggest a possible neural substrate for the perceptive ability to maintain the integrity of a melody despite musical transposition.

Musical Semantics

Although the issue of whether nonvocal music carries any inherent semantics is highly controversial, music certainly has the power to convey culturally relevant interpretations and associations. One study measured the N400 component of the event-related brain potential during the presentation of 'semantically' related words and musical excerpts, such as the word

'staircase' and an ascending musical scale. The findings argue that words with descriptive, qualitative, structural, or abstract similarities to a particular musical excerpt activate a response pattern similar to semantically related word and sentence pairs, indicating that superficial semantic interpretations of certain musical aspects might be shared. These findings, while provocative, underscore the need for further investigation of the effects of environment factors such as culture, language, and training on music perception and interpretation.

Musical Memory and Imagery

Memory is crucial to the perception of music, since the macroscopic structures of melody, rhythm, and harmony unfold over periods of time. As indicated by imaging and lesion studies, working memory for pitch requires the interaction of the right temporal and right frontal cortices. Measurements of cerebral blood flow using positron emission tomography during tasks of recognition and identification of music as 'familiar' show a broad network of activation, including the bilateral superior temporal cortices; bilateral orbital, medial, and inferior frontal regions; precuneus; angular gyrus; and parahippocampal gyrus.

Musical imagery is the subjective task of imaging music in the absence of auditory input. When individuals were asked to imagine the continuation of a truncated musical stimulus, activation was observed in the inferior frontal and superior temporal regions as well as the supplementary motor areas. In addition, measurements of the topography of event-related potentials (ERPs) found that activity in the imagination of the continuation of a melody is similar to the pattern activity when auditory stimuli are not withheld, further indicating that musical imagination employs auditory regions even in the absence of acoustic stimuli.

Emotion

Since music is ultimately abstract, the interpretation and understanding of music are highly subjective, depending on culture, degree of musical training, and personal preference. Nevertheless, the power of music to elicit particularly strong emotive responses is well recognized. At first glance this association seems quite unusual; however, neurophysiological investigations indicate that these emotive responses to music are associated with primitive neural systems associated with basic survival mechanisms, perhaps explaining why music has persisted through history despite the absence of any obvious evolutionary benefits. These associations between limbic structures and musical perception provide a potentially contradictory viewpoint to theories that postulate music to be the evolutionary side effect of an auditory system that has become specialized for language.

A seminal study used PET imaging to monitor neural responses to individual-selected music that elicited 'chills' down the spine. Increases in the subjective intensity of the emotive experience correlated to changes of cerebral blood flow in the ventral striatum, dorsal midbrain, amygdala, orbitofrontal cortex, and ventral medial prefrontal cortex – regions associated with processing of reward and pleasure, such as hunger and sexual drive. Similar imaging studies have also shown correlation between pleasant responses to music and activity in the subcallosal cingulate gyrus, prefrontal anterior cingulate, retrosplenial cortex, and hippocampus; the nucleus accumbens, ventral tegmental area, hypothalamus, and insula; and the inferior frontal gyrus, Heschl's gyrus, and the Rolandic operculum. Increasingly dissonant music correlated with increased activity in the right parahippocampal gyrus and right precuneus (Figure 3).

Lesion-based studies indicate both functional and structural dissociations between music perception and emotional responses to music. The recognition of emotional tone can be spared even when basic perception and recognition of music are impaired. Conversely, deficits in response to emotion in music are described even when basic music perception is unimpaired. A single case study found that following an infarction involving the left amygdala and insula, the patient lost the 'shiver down the spine' phenomenon he had previously experienced when listening to Rachmaninoff, but retained both pitch and speech perception.

Furthermore, the neurological correlates of the emotional responses to music seem to vary based on the subjective perception of emotional tone. Patients with either left or right medial temporal lobe resections, including the amygdala but sparing the superior temporal gyrus, were found to be impaired in the recognition of 'scary music', but unimpaired in the recognition of 'happy' or 'sad' music. Another patient with complete bilateral damage relatively restricted to the amygdala was selectively impaired in the recognition of 'scary music', but was able to identify 'happy' music as well as the control group could. In a similar study, patients with significant loss of the parahippocampal cortex were unable to identify 'unpleasant' music but were able to identify consonant music as pleasing. Another patient with significant impairment in tasks of music perception caused by bilateral lesions to the auditory cortex was found to be indifferent to musical dissonance but was able to identify music as either 'happy' or 'sad' as well as the

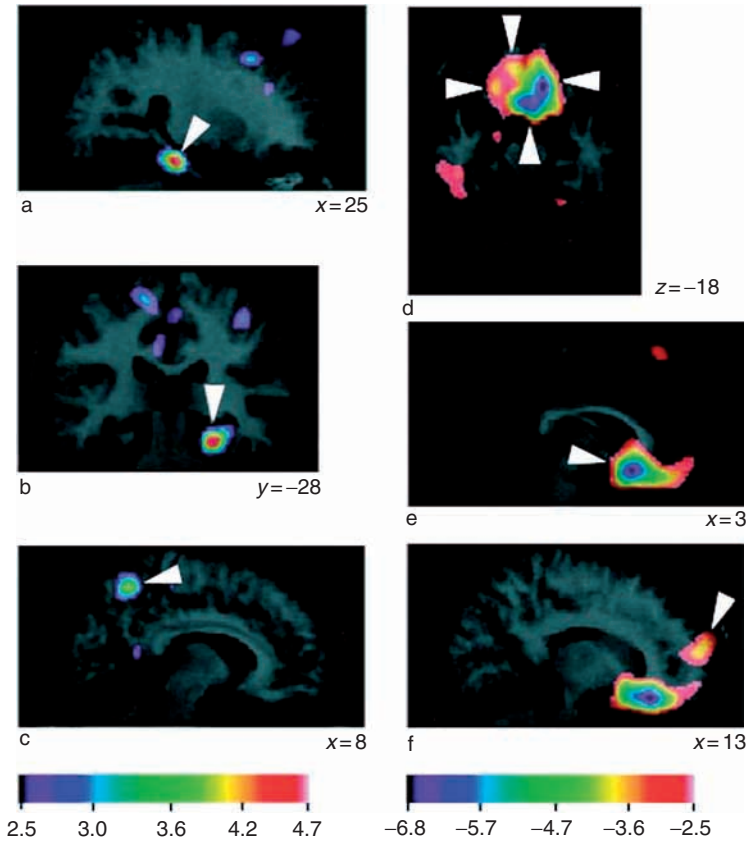

Figure 3 Cortical regions demonstrating significant regional cerebral blood flow correlations with dissonance level. Correlations are shown as *t*-statistic images superimposed on corresponding averaged magnetic resonance image scans. The *t*-statistic ranges for each set of images are coded by color scales below each column, corresponding to images a–c and d–f. (a–c) Positive correlations with increasing dissonance demonstrated relative cerebral blood flow activity in right parahippocampal gyrus and right precuneus. (d–f) Negative correlations with increasing dissonance (equivalent to positive correlations with increasing consonance) demonstrated relative cerebral blood flow activity in bilateral orbitofrontal cortex, medial subcallosal cingulate, and right frontal polar regions. Reprinted from Blood AJ, Zatorre RJ, Bermudez P, and Evans AG (1999) Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nature Neuroscience* 2: 382–387 with permission from Macmillan Publishers Ltd. Copyright (1999).

normal controls could. When taken together, these results imply that the emotive interpretation of consonance and dissonance relies initially on the perceptional organization of sound (in the superior temporal gyri and other neocortical regions) that is subsequently relayed to limbic and paralimbic regions responsible for affective responses.

Musical Training

Professional musicians achieve expertise through years of rigorous training, usually starting at an early age. Studies comparing skilled musicians and nonmusicians demonstrate both morphological and functional differences in the musical brain. Such differences provide the opportunity to investigate the effects of musical training on brain plasticity and reorganization. Imaging studies comparing musicians and nonmusicians have shown structural differences in the motor cortex, the somatosensory and superior-parietal areas, the auditory cortex, the corpus callosum, and cerebellum.

In addition to the anatomical differences seen in brains of musicians and nonmusicians, functional differences in brain organization have also been observed. Neuroimaging studies comparing musicians and nonmusicians have demonstrated differences of activity in the auditory cortex, motor cortices, and cerebellum. These studies reveal both enhancements and reductions of activity, suggesting the complex effects of musical training on brain reorganization. One such MEG study demonstrated an enhanced cortical representation of the four fingers of the left hand used in string playing, compared to the nonmusicians and to the right hand of the string players themselves. Furthermore, increases in the cortical representation of the somatosensory and auditory areas in skilled musicians have been shown to correlate with age at which musical training was begun.

Some of the plastic effects of musical training can relate to the musician's specific area of training. Orchestra conductors show a superior spatial processing in the auditory periphery, reflecting the skill of localizing the sounds of various musical instruments in space. Instrumental performers show a preference to their own musical instrument in both functional and morphological enhancements of the auditory cortex, further indicating long-term training-induced plasticity. But even short periods of intense musical training can cause functional brain reorganization. While differences certainly exist in the musical brain, the precise nature and complex interaction of these structural and function brain changes caused by musical training are little understood. Furthermore, the relationship between brain structure and function and innate musical predisposition prior to training has not yet been fully explored.

Music and Culture

While music is universally present across cultures, the subjective interpretation of music depends on a host of factors, including culture, musical exposure, and musical training. Culturally specific definitions of music vary to such a degree that what is deemed highly musical in one culture may be perceived as completely unmusical in another. A majority of studies of the neural correlates of music perception have focused on individuals and stimuli from Western musical traditions. Yet, the functional and structural differences found between musicians and non-musicians invite investigation into possible effects of cultural exposure on music perception. Persons culturally exposed to Western music react to violations of musical syntax at a very young age, when musically untrained, when distracted, and even when explicitly primed for the violation. Furthermore, acquisition of implicit musical expectation seems to occur early in development, possibly within the first year of development.

Only a few studies have explicitly investigated the correlates between music perception and cultural exposure. Individuals show a behavioral preference in the ability to synchronize with culturally specific music, and electroencephalography (EEG) data indicate cultural-specific differences in listening strategies of musical phrase, musical instrument, and musical scale. However, while the processing of a familiar language may use distinct brain regions, one functional imaging study found no differences in the processing of culturally familiar and unfamiliar music – only differences correlating to the degree of musical training that were culturally independent.

Conclusion

The various ways music is perceived, performed, learned, identified, and interpreted indicate a broad system of complex neurological interactions involving a wide array of brain regions and structures. Historically, investigations of the cognitive neuroscience of music strive to isolate regions and processes associated with the individual aspects of music. Emerging neuroimaging literature continues to isolate substrates responsible for the perception of individual elements of music. While many argue this dissection of musical stimuli threatens the 'ecological validity' of music, this increased understanding of basic musical elements undoubtedly provides a foundation for future macroscopic comparisons and investigations. Ideally, an eventual convergence of musicologic, psychologic, and neuroimaging data will extend our knowledge to advanced concepts of musical learning, performance, and composition, as well as the complex questions surrounding the innate musical ability and musical creativity.

See also: Auditory Cortex: Models; Auditory Evoked Potentials; Auditory Scene Analysis; Auditory System: Central Pathways; Cognition: An Overview of Neuroimaging Techniques; Connectionist Models of Language Processing; Language Development; Language Evolution; Language: Auditory Processes; Language: Cortical Processes; Musical Illusions; Neuroimaging.

Further Reading

Avanzini G, Faienza C, Minciacchi D, et al. (eds.) (2006) *The Neurosciences and Music II: From Perception to Performance*. New York: New York Academy of Sciences.

Avanzini G, Lopez L, Koelsch S, et al. (eds.) (2003) *The Neurosciences and Music,* vol. 999. New York: New York Academy of Sciences.

Bendor D and Wang X (2005) The neuronal representation of pitch in primate auditory cortex. *Nature* 436: 1163.

Blood AJ and Zatorre RJ (2001) Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11818–11823.

Blood AJ, Zatorre RJ, Bermudez P, and Evans AC (1999) Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nature Neuroscience* 2: 382–387.

Deutsch D (ed.) (1999) *The Psychology of Music,* 2nd edn. San Diego, CA: Academic Press.

Elbert T, Pantev C, Wienbruch C, et al. (1995) Increased cortical representation of the fingers of the left hand in string players. *Science* 270: 305–307.

Juslin PN and Sloboda JA (eds.) (2001) *Music and Emotion: Theory and Research*. Oxford: Oxford University Press.

Krumhansl CL (1990) *Cognitive Foundations of Music Pitch*. New York: Oxford University Press.

Mithen S (2006) *The Singing Neanderthals: The Origins of Music, Language, Mind and Body.* Cambridge, MA: Harvard University Press.

Peretz I and Zatorre RJ (eds.) (2003) *The Cognitive Neuroscience of Music.* Oxford: Oxford University Press.

Peretz I and Zatorre RJ (2005) Brain organization for music processing. *Annual Review of Psychology* 56: 89–114.

Plack CJ, Oxenham AJ, Fay RR, et al. (eds.) (2005) *Pitch: Neural Coding and Perception.* New York: Springer.

Wallin NL, Merker B, and Brown S (eds.) (2000) *The Origins of Music.* Cambridge, MA: MIT Press.

Zatorre RJ, Belin P, and Penhune VB (2002) Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Neuroscience* 6(1): 37–46.

Zatorre RJ, Evans AC, and Meyer E (1994) Neural mechanisms underlying melodic perception and memory for pitch. *Journal of Neuroscience* 14: 1908–1919.

Zatorre RJ and Belin P (2001) Spectral and temporal processing in human auditory cortex. *Cerebral Cortex* 11: 949.

Zatorre RJ and Peretz I (eds.) (2001) *The Biological Foundations of Music,* vol. 930. New York: New York Academy of Sciences.

Musical Illusions

D Deutsch, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Over the past decade, there have been significant advances in understanding the brain mechanisms underlying music perception. Functional imaging studies of normal individuals have provided considerable information concerning the anatomical structures involved. Studies of patients with musical deficits resulting from brain lesions have supplied much complementary information, as have studies of individuals with congenital amusia. Investigations of musical illusions – misperceptions of music by normal individuals – provide further insights into the brain circuitry underlying music perception. This entry reviews those musical illusions which have clear implications for the neuroscience of music. In exploring these illusions, we focus in particular on three issues:

1. *Innate differences in music perception.* Can we assume that music perception is the function of innate brain circuitry that is common to most individuals, or are there widespread and substantial differences in music perception that reflect innate differences in brain organization? Some of the illusions to be described point to widespread individual differences in the perception of music that correlate with handedness; these indicate that there are substantial variations in the innate brain circuitry underlying music perception.

2. *Neural plasticity in the development of music perception.* Recent imaging studies have indicated that musical practice produces alterations in brain organization, and that this happens particularly at an early age, when the effects of experience on brain function are particularly strong. In addition, a musical illusion to be described here has been shown to be heavily influenced by the speech patterns to which the individual has been exposed, particularly in childhood. The characteristics of this illusion indicate that the brain circuitry underlying music perception can be influenced by extramusical input, particularly at an early age. More specifically, it points to a strong influence of exposure to speech patterns in childhood on how music is later perceived.

3. *Modularity in musical processing.* Findings from neurophysiological and neuropsychological studies have indicated that there are distinct 'what' and 'where' pathways in the auditory system. In accordance with these findings, certain musical illusions indicate that there are separate decision mechanisms underlying what pitches we hear and underlying where the tones appear to be coming from, and that the combined outputs of these decision mechanisms can lead to illusory conjunctions. More specifically, we can think of a musical tone as a bundle of attribute values – having a pitch, a location, a loudness, a timbre, and so on, with the different attributes being subserved by distinct neural pathways. When more than one tone is presented at a time, the different outputs of these pathways must be correctly conjoined in order to produce veridical percepts. Some of the illusions described here can be explained as due to incorrect conjunctions of attribute values.

Stereo Illusions

Powerful musical illusions can be produced when two streams of tones arise in parallel from different regions of space. There are striking differences among listeners in the way these illusions are perceived. These differences have been found to correlate with the handedness of the listener, and so can be taken to reflect innate differences in brain organization.

The Octave Illusion

The octave illusion is produced by the pattern shown in the upper portion of Figure 1. Two tones which are spaced an octave apart (at 400 and 800 Hz) are repeatedly presented in alternation. The tones are 250 ms sine waves at equal amplitude, and they follow each other without pause. The identical sequence is presented via headphones to both ears simultaneously; however, when the right ear receives the high tone the left ear receives the low tone; and vice versa.

Surprisingly, this simple pattern is almost never perceived correctly, and instead gives rise to a variety of illusions. The type of illusion varies substantially across listeners, and the one most commonly obtained is shown in the lower portion of Figure 1. A single tone appears to be switching from ear to ear while at the same time its pitch appears to be switching between high and low. In other words, the sequence as heard at the right ear consists of 'high tone – silence – high tone – silence' while the sequence as heard at the left ear consists of 'silence – low tone – silence – low tone.'

The illusion becomes even more surprising when the listener's earphones are placed in reverse position. Now most people experience exactly the same thing – the tone that had appeared in the right ear still appears in the right ear, and the tone that had

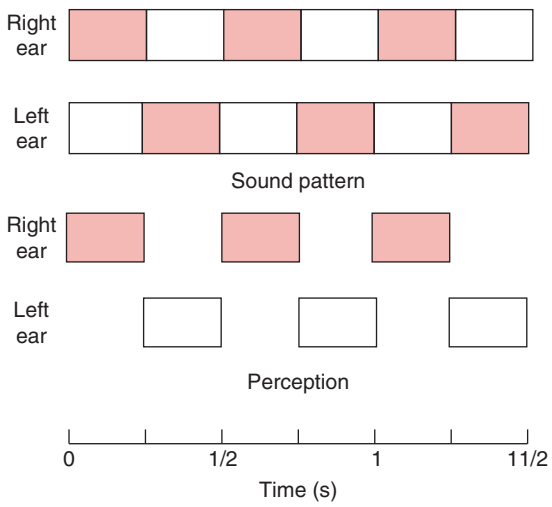

Figure 1 The sound pattern that gives rise to the octave illusion, and the percept most commonly obtained by right-handers. Filled boxes indicate tones of 800 Hz and unfilled boxes indicate tones of 400 Hz. When this pattern is played continuously through headphones, most right-handed listeners perceive an intermittent high tone in one ear that alternates with an intermittent low tone in the other ear. Adapted from Deutsch D (1974) An auditory illusion. *Nature* 251: 307–309.

appeared in the left ear still appears in the left ear. This gives rise to the peculiar impression that the earphone that had been producing the high tone is now producing the low tone, and that the earphone that had been producing the low tone is now producing the high tone.

There is clearly no simple explanation for this illusory percept. However, if we assume that there are two separate decision mechanisms, one determining what pitch we hear, and the other determining where the sound is coming from, we are in a position to advance an explanation. The model is shown in Figure 2. To provide the perceived sequence of pitches, the frequencies arriving at the dominant ear (i.e., the right ear in most right-handers) are followed, while those arriving at the nondominant ear are suppressed from conscious perception. However, to provide the perceived locations, each tone is localized at the ear that receives the higher frequency signal, regardless of whether a pitch corresponding to the higher or the lower frequency is perceived.

More specifically, we can consider a listener who perceives the pitches corresponding to the frequencies delivered to the right ear. When the high tone is delivered to the right ear and the low tone is delivered to the left, this listener hears a high tone, since it is delivered to the right ear, and localizes the tone in the right ear, since this ear is receiving the higher frequency signal. However, when a low tone is delivered to the right ear and a high tone is delivered to the left, this listener now hears a low tone, since it is delivered to the right ear, but hears the tone in the left ear instead, since this ear is receiving the higher frequency signal. So the full sequence is heard as a high tone to the right, alternating with a low tone to the left. It can be seen from Figure 2 that reversing the position of the headphones would not alter this basic percept. However, given a listener who perceives the pitches corresponding to the frequencies arriving at the left ear instead, holding the localization rule constant, the same sequence is heard as a high tone to the left, alternating with a low tone to the right.

Evidence for this model has been obtained in a number of studies. The octave illusion therefore provides a strong indication, at the perceptual level, that the neural pathways underlying what pitch we hear, and those underlying where the sound is located, are at some stage distinct and separate, and operate in accordance with different rules. The combined outputs of these two mechanisms can then result in striking illusory conjunctions.

As another aspect of the octave illusion, it has been found, on a statistical basis, that right-handers and left-handers show different patterns of localization for the high and low tones at the two ears. Right-handers tend strongly to hear the high tone on the right and the low tone on the left; however, left-handers as a group do not show the same tendency. This handedness difference is in accordance with the literature showing that while most right-handers have clear left-hemisphere dominance, patterns of cerebral dominance among left-handers vary considerably. In one study individuals were categorized as strongly right-handed, mixed-handed, or left-handed, and these groups were subdivided into those who had only right-handed parents and siblings, and those who had a left-handed parent or sibling. As shown in Table 1, this categorization produced an orderly pattern of results: subjects who were strongly right-handed and with no left-handers in their family were most likely to hear the high tone on the right and the low tone on the left, while left-handers with left-handers in their family were least likely to do so. These findings are as expected from neuropsychological studies relating hemispheric dominance of function to patterns of handedness and familial handedness background, since those individuals with left-handers in their family are less likely to be strongly left-hemisphere dominant than are those with only right-handers in their family. As a further handedness correlate, left-handers as a group tend more than right-handers to obtain complex percepts on listening to the octave illusion. This is in accordance with studies showing that left-handers, on a statistical basis, tend to have weaker patterns of cerebral dominance than do right-handers.

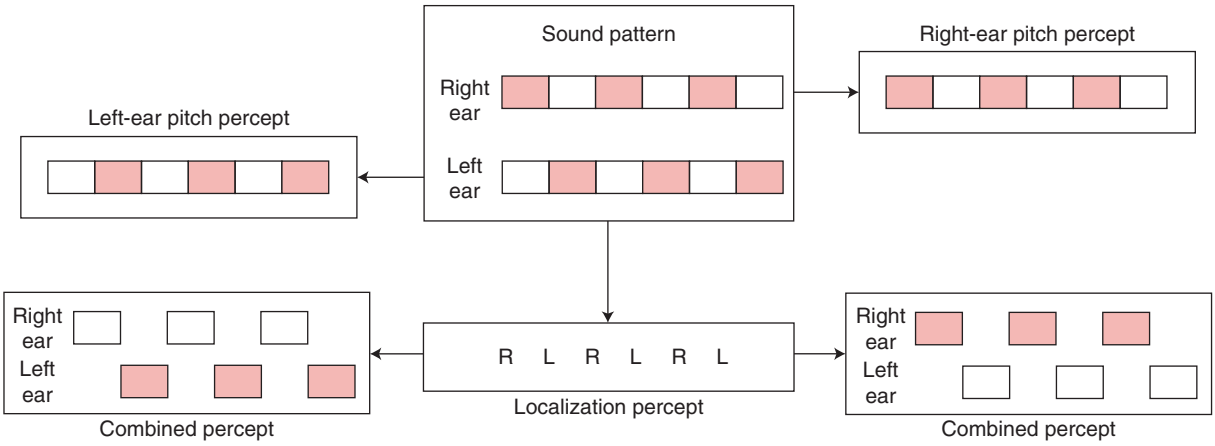

Figure 2 Model showing how the outputs of two decision mechanisms, one determining perceived pitch and the other determining perceived location, combine to produce the octave illusion. Filled boxes indicate tones of 800 Hz and unfilled boxes indicate tones of 400 Hz (R, right; L, left). Adapted from Deutsch D (1983) Auditory illusions, handedness, and the spatial environment. *Journal of the Audio Engineering Society* 31: 607–618.

Table 1 Percentages of individuals reporting the high tone on the right and the low tone on the left on listening to the octave illusion, tabulated by handedness and familial handedness background

	Dextral	*Left- or mixed-handed parent*	*Left- or mixed-handed sibling*
Right-handed	87.2% ($N=86$)	64.0% ($N=25$)	79.3% ($N=29$)
Mixed-handed	67.7% ($N=31$)	44.4% ($N=9$)	58.3% ($N=12$)
Left-handed	55.6% ($N=9$)	50.0% ($N=16$)	33.3% ($N=6$)

From Deutsch D (1983) The octave illusion in relation to handedness and familial handedness background. *Neuropsychologia* 21: 289–293.

The Scale Illusion

We next inquire what happens when a pattern is presented that consists of more than two tones. The scale illusion provides an example. As shown in the upper portion of Figure 3, the sequence that produces this illusion consists of a major scale, with successive tones alternating from ear to ear. The scale is presented simultaneously in both ascending and descending form, such that when a tone from the ascending scale is in the right ear, a tone from the descending scale is in the left ear; and vice versa. In consequence, the right ear receives one disjunct pattern of pitches, while the left ear receives a different, and overlapping, disjunct pattern of pitches. The tones are 250 ms sine waves at equal amplitude, and the sequence is repeatedly presented without pause.

This scale pattern also produces a number of illusions, which vary strikingly across listeners. The illusion most commonly obtained by right-handers is shown in the lower portion of Figure 3. A melody that corresponds to the higher tones appears to be coming from the right earphone, and melody corresponding to the lower tones appears to be coming from the left earphone. When the earphone positions are reversed, the higher tones continue to be heard as in the right ear

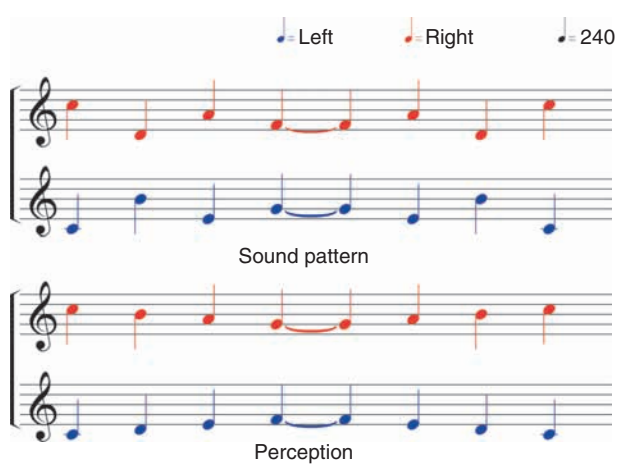

Figure 3 The sound pattern that gives rise to the scale illusion, and the percept most commonly obtained by right-handers. A major scale is presented simultaneously in ascending and descending form. The components alternate between earphones such that when a component of the ascending scale is delivered to one ear a component of the descending scale is delivered to the other ear. The pattern is played continuously without pause. Most right-handers perceive a melody that is composed of the higher tones in the right ear and a melody composed of the lower tones in the left ear. Adapted from Deutsch D (1999) Grouping mechanisms in music. In: Deutsch D (ed.) *The Psychology of Music*, 2nd edn., pp. 299–348. San Diego: Academic Press.

and the lower tones as in the left ear. This produces the strange impression that the procedure of reversing earphone positions has caused the higher tones to migrate from one earphone to the other, and the lower tones to migrate in the opposite direction. As with the octave illusion, right-handers and left-handers differ statistically in the way they perceive the scale illusion: right-handers tend to hear the higher tones on the right and the lower tones on the left; however, left-handers as a group do not show the same tendency. This indicates that there is a perceptual tendency to perceive higher tones as coming from the dominant side of space and lower tones as from the nondominant side.

Variants of the scale illusion are easily produced. Figure 4 depicts a version involving a two-octave chromatic scale. In reality, the sounds coming from each earphone leap around in pitch; however, listeners instead hear two smooth melodies that correspond to the higher and lower tones in the combined pattern, and they perceive the higher tones as coming from one earphone (in right-handers, this is usually the right earphone) and the lower tones as from the other.

The scale illusion can also occur with the sounds presented to the listener via two loudspeakers in a normal room environment. In one experiment, piano tones were used as signals, and differences in timbre were introduced for the tones presented through the two loudspeakers. The differences in timbre gave rise to yet another perceptual dissociation: a new sound quality was heard, but it appeared to be coming simultaneously from both loudspeakers. This perceptual dissociation of timbre from pitch and location may be related to findings indicating that brain lesions can give rise to deficits in timbre perception with spared pitch perception.

Why, then, should the illusory conjunctions that give rise to perception of the scale illusion occur? One approach involves the notion of unconscious inference. When faced with a complex sound configuration the brain engages in a process of inference concerning the sources that produced the sounds. In everyday life, similar sounds are likely to be coming from the same source and different sounds from different sources. So the best interpretation of this implausible musical pattern, in terms of the real world, is that the higher tones are all coming from one source and the lower tones from a different source. We therefore mislocalize the tones in accordance with this interpretation.

A related approach derives from the writings of the Gestalt psychologists, who argued that we form perceptual configurations based on certain organizational principles such as proximity. In the case of vision, for example, we group together dots that are positioned in spatial proximity, and we perceive motion between neighboring lights that are turned on and off in succession. The principle of proximity is a useful heuristic, since in the real world elements that are proximal in an array are likely to be coming from the same source. In the scale illusion we group

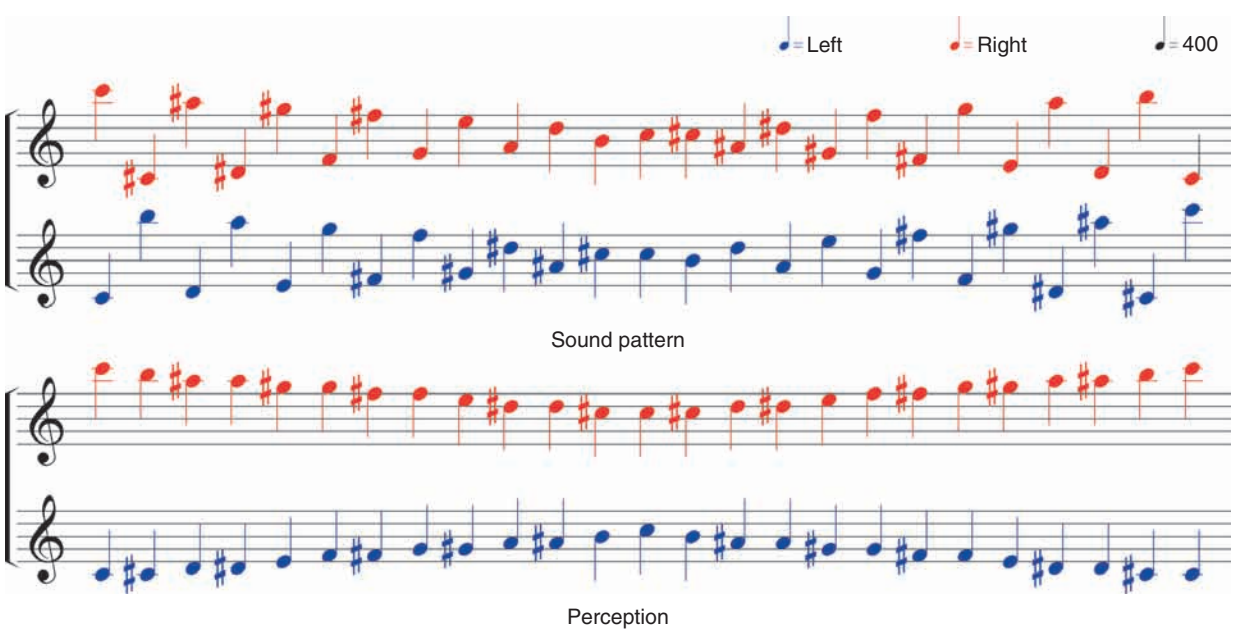

Figure 4 The sound pattern that gives rise to the chromatic illusion, and a way it is often perceived by right-handers. A two-octave chromatic scale is played simultaneously in ascending and descending form. The components alternate between earphones such that when a component of the ascending scale is presented to one ear a component of the descending scale is presented to the other ear. Most right-handers perceive a melody that is composed of the higher tones in the right ear and a melody composed of the lower tones in the left ear.

together successive tones that are related by pitch proximity, and separate out those that are further apart, and we mislocalize the tones on this basis, so overriding low-level localization cues.

The Glissando Illusion

The glissando illusion is best heard with the listener seated in front of two stereophonically separated loudspeakers, with one to the left and the other to the right. The pattern that gives rise to the illusion consists of two components; the first is a synthesized oboe tone of constant pitch, and the second is a sine wave that glides up and down in pitch. As illustrated in Figure 5, these two components are presented simultaneously via the two loudspeakers, and constantly switch between the speakers such that when the oboe tone is emanating from the speaker on the right a portion of the glissando is emanating from the speaker on the left; and vice versa. A fragment of the pattern producing this illusion is shown in Figure 5. When each channel is presented separately, the listener correctly hears the oboe tone in alternation with a portion of the glissando. However, when both channels are presented together, the percept changes dramatically. The oboe tone is heard correctly as switching back and forth between the loudspeakers. However, the portions of the glissando appear to be joined together quite seamlessly, so that a single, continuous tone is heard that appears to move around in space in accordance with its pitch motion. When the switching rate speeds up and slows down, the speed at which the glissando appears to move around in space does not change, but instead remains tied to its pitch motion.

Handedness correlates again emerge in terms of the apparent spatial positions of the higher and lower portions of the glissando: right-handers tend strongly to hear the glissando move from left to right as its pitch moves from low to high, and then back from right to left as its pitch moves from high to low. However, non-right-handers as a group do not show the same tendency. In this sense, the glissando illusion is similar to the octave and scale illusions, since it reflects a tendency to hear higher tones as on the dominant side of space and lower tones as on the nondominant side. However, in the octave and scale illusions the apparent spatial positions of the higher and lower tones generally remain fixed for the duration of the pattern, whereas in the glissando illusion the gliding tone is perceived as moving slowly in space, in accordance with its pitch motion.

In considering why the glissando illusion occurs we can refer to the same principles as were discussed with respect to the scale illusion. In this case we link together portions of the glissando that are proximal in pitch, and that follow smooth trajectories. In the real world this perceptual strategy is likely to lead to the best conclusion in interpreting our auditory environment, since it is very unlikely that a sound which is changing smoothly in pitch is switching abruptly between two widely different locations.

Illusions Based on Octave Ambiguity

It has long been recognized that tones that are related by octaves have a certain perceptual equivalence. This is acknowledged in the system of notation for the traditional musical scale. The core of this scale consists of 12 tones, which correspond to the division of the

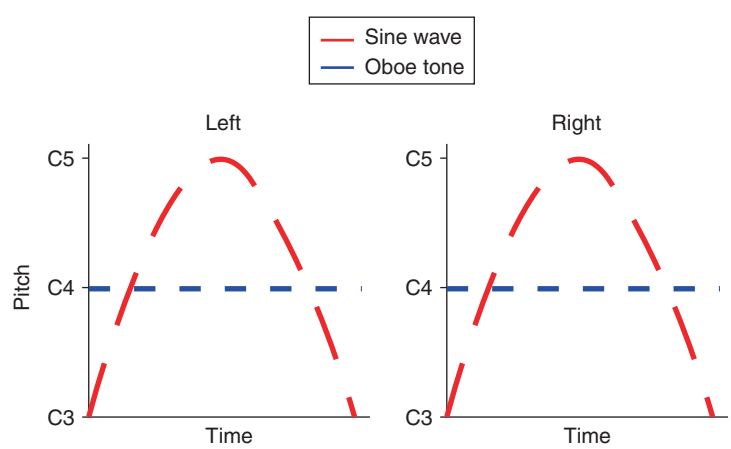

Figure 5 The sound pattern that produces the glissando illusion. This consists of an oboe tone of constant pitch and a sine wave that glides up and down in pitch, with the two repeatedly switching between loudspeakers. Listeners perceive the oboe tone correctly as alternating between loudspeakers; however, the glissando appears to be joined together quite seamlessly, and to be moving around in space in accordance with its pitch motion.

octave into semitones, and each tone is given a name (C, C#, D, D#, E, F, F#, G, G#, A, A#, and B). The entire scale, as it ascends in height, consists of repeating the same sequence of note names across successive octaves. Because all Cs sound in a sense equivalent, as do all C#s, all Ds, and so on, pitch can be characterized as varying both along a monotonic dimension of height and also along a circular dimension of pitch class (or note name). The circular dimension is illustrated in Figure 6. The two dimensions can be largely separated by the use of computer-produced tones with pitch classes that are clearly defined, but that are ambiguous in terms of height. For example, one can employ tones consisting only of components that are related by octaves. With the use of such ambiguous tones, striking illusions are produced.

Illusions of Pitch Circularity

When two octave-ambiguous tones are played in succession, the listener hears either an ascending pattern or a descending one, depending on which is the shorter distance between the tones along the pitch class circle. So, for example, the tone pair consisting of C# followed by D is always heard as ascending, since the shorter distance between these tones is clockwise. Similarly the tone pair consisting of F# followed by F is always heard as descending, since the shorter distance between these tones is counterclockwise. So when a series of octave-ambiguous tones is presented that repeatedly traverses the pitch class circle in clockwise direction (C, C#, D, and so on) listeners perceive a sequence that appears to ascend endlessly in pitch. When, instead, the tones repeatedly traverse the pitch class circle in counterclockwise direction (C, B, A#, and so on) listeners hear a sequence that appears to descend endlessly. So just as with the scale illusion, this circularity illusion provides an example of the perceptual tendency to link together elements of an array of tones on the basis of pitch proximity.

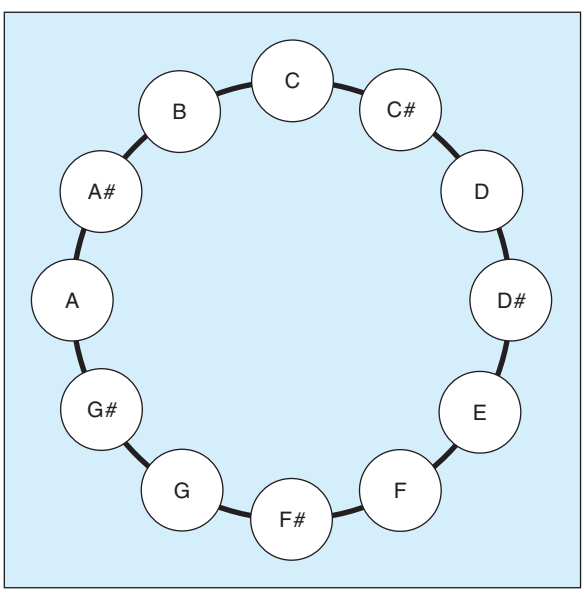

Figure 6 The pitch class circle. The notes of the Western musical scale are formed by dividing the octave into 12 semitone steps, and each note is given a name: C, C#, D, and so on. The entire scale is formed by repeating this sequence of note names across octaves.

The Tritone Paradox

We can then ask what happens when listeners are presented with pairs of octave-ambiguous tones that are related in such a way that proximity cannot be used as a cue in judging their relative heights. For example, what happens when listeners are presented with two tones in succession that stand in opposite positions along the pitch class circle, so that the same distance between them along the circle is traversed in either direction? Such tones form an interval of six semitones, known as a tritone. When such tone pairs are presented, striking perceptual differences among listeners emerge. For example, when D is presented followed by G#, some listeners clearly hear an ascending pattern, while other listeners clearly hear a descending one. Yet when, instead, A is presented followed by D#, the first group of listeners now hears a descending pattern, whereas the second group now hears an ascending one. Furthermore, for any given listener the pitch classes generally arrange themselves perceptually with respect to height in a systematic way: tones in one region of the pitch class circle are heard as higher, and tones in the opposite region as lower. This is illustrated in Figure 7, which reproduces judgments of the tritone paradox made by four different listeners. It can be seen that the judgments of each listener showed an orderly relationship to the positions of the tones along the pitch class circle; however, the direction of this relationship varied substantially across listeners. As a result, extended patterns formed of such tone pairs are heard by listeners in entirely different ways, as depicted in Figure 8.

It has been shown that the way the tritone paradox is perceived correlates with the pitch range of the listener's speaking voice. Perception of this illusion also varies with the language or dialect to which listeners have been exposed. For example, striking differences have been found between listeners who had grown up in California and those who had

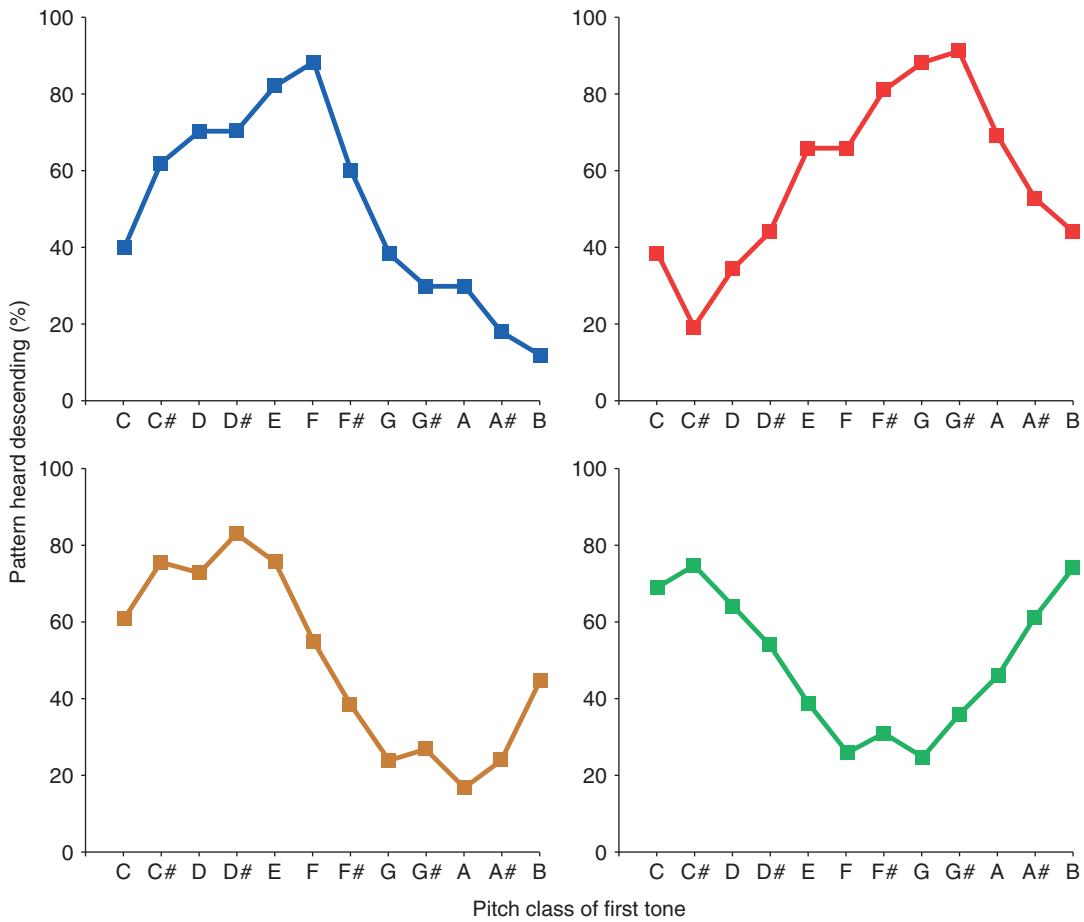

Figure 7 Judgments of the tritone paradox made by four different listeners. Each graph shows the percentages of judgments that a tone pair formed a descending pattern, plotted as a function of the pitch class of the first tone of the pair. The judgments of all listeners showed orderly relationships to the positions of the tones along the pitch class circle; however, the direction of this relationship varied substantially across listeners. Data supplied by the author.

grown up in the south of England: in general, when the California group tended to hear a pattern as ascending, the English group tended to hear it as descending, and vice versa. It appears, therefore, that perception of the tritone paradox is based on an acquired pitch template that is used primarily for speech communication.

Further work has provided evidence that the pitch representation that influences perception of the tritone paradox is formed early in life. On study explored perception of the tritone paradox in mothers and their children. The children were all Californians; however, the mothers had grown up in many different geographical regions, including England, the European continent, and various parts of the United States. As expected, the group of mothers perceived this pattern in ways that differed considerably; however, although the children were all native Californians, their percepts were very similar to those of their mothers. In a further study, perception of this illusion was explored in individuals who were born in Vietnam, as well as in a group of English-speaking Californians. The Vietnamese group was divided into two subgroups: those who had arrived in the United States as adults and spoke fluent Vietnamese but little English, and those who had arrived as infants or young children and now spoke perfect English, though most were not fluent in Vietnamese. Perception of the tritone paradox by the two Vietnamese groups was very similar, and clearly differed from the paradox's perception by English-speaking Californians. The study therefore indicates that exposure to speech sounds during an early critical period can have a substantial influence on perception of music in adulthood.

Conclusion

The perceptual phenomena described in this article demonstrate that the auditory system is very prone

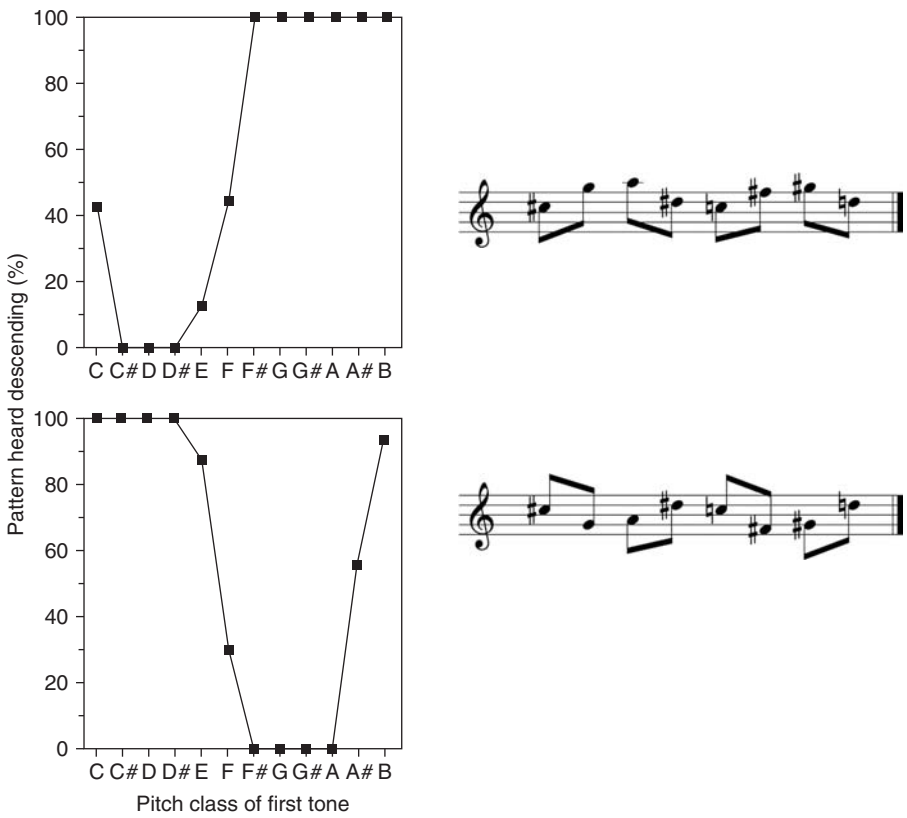

Figure 8 Perception of the tritone paradox by two listeners who experienced the illusion in a pronounced way. Each graph shows the percentages of judgments that a tone pair formed a descending pattern, plotted as a function of the pitch class of the first tone of the pair. The notations show how the identical series of tone pairs was perceived by these two listeners in entirely different ways. From Deutsch D (1999) Processing of pitch combinations. In: Deutsch D (ed.) *The Psychology of Music,* 2nd edn., pp. 349–412. San Diego: Academic Press.

to illusion. This is not surprising, considering the paucity of information that is supplied by the peripheral auditory system. There are roughly 16 000 peripheral receptors in each ear, compared with roughly 125 million receptors in each eye. In addition, the amount of brain tissue devoted to hearing is far less than is devoted to vision. As a further consideration, sound signals are mixed together and are subject to numerous distortions as they travel from their sources to our ears, so that the brain is faced with the very difficult task of reconstructing the original signals from the mixture of sounds that arrive at the ears. The musical illusions described here provide valuable information concerning the strategies adopted by the nervous system in interpreting our auditory environment, and they reflect the operation of both innate mechanisms and neural plasticity in the process.

See also: Language: Auditory Processes; Music; Speech Perception: Adult; Speech Perception: Cortical Processing; Speech Perception: Development; Speech Perception: Neural Encoding.

Further Reading

Deutsch D (1974) An auditory illusion. *Nature* 251: 307–309.

Deutsch D (1975) Musical illusions. *Scientific American* 233: 92–104.

Deutsch D (1975) Two-channel listening to musical scales. *Journal of the Acoustical Society of America* 57: 1156–1160.

Deutsch D (1983) Auditory illusions, handedness, and the spatial environment. *Journal of the Audio Engineering Society* 31: 607–618.

Deutsch D (1983) The octave illusion in relation to handedness and familial handedness background. *Neuropsychologia* 21: 289–293.

Deutsch D (1991) The tritone paradox: An influence of language on music perception. *Music Perception* 8: 335–347.

Deutsch D (1992) Paradoxes of musical pitch. *Scientific American* 267: 88–95.

Deutsch D (1992) Some new sound paradoxes and their implications. Auditory processing of complex sounds. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 336: 391–397.

Deutsch D (1999) Grouping mechanisms in music. In: Deutsch D (ed.) *The Psychology of Music,* 2nd edn., pp. 299–348. San Diego: Academic Press.

Deutsch D (1999) Processing of pitch combinations. In: Deutsch D (ed.) *The Psychology of Music,* 2nd edn., pp. 349–412. San Diego: Academic Press.

Deutsch D, Henthorn T, and Dolson M (2004) Speech patterns heard early in life influence later perception of the tritone paradox. *Music Perception* 21: 357–372.

Peretz I and Zatorre R (2003) *The Cognitive Neuroscience of Music*. Oxford: Oxford University Press.

Rauschecker JP and Tian B (2000) Mechanisms and streams or processing 'what' and 'where' in the auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America* 97: 1180–1186.

Shepard RN (1964) Circularity in judgments of relative pitch. *Journal of the Acoustical Society of America* 36: 2346–2353.

Relevant Website

http://psy.ucsd.edu–ucsd psychology, Deutsch's Musical Illusions.

Myelin: Molecular Architecture of CNS and PNS Myelin Sheath

S S Scherer and E J Arroyo, University of Pennsylvania Medical School, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

The Function of Myelin Sheaths

A myelin sheath is a multilamellar spiral of specialized cell membrane that surrounds a segment of an axon that is typically larger than 1 μm in diameter (Figure 1). Myelinated axons are completely covered by myelin sheaths (internodes) except at nodes of Ranvier, the small gaps (less than 1 μm in length) directly exposed to the extracellular milieu. By reducing the capacitance and clustering ion channels at nodes, myelin sheaths reduce current flow across the internodal axonal membrane, thereby facilitating saltatory conduction, one of the fundamental adaptations of vertebrates.

The Circumferential Organization of Peripheral Nervous System Myelin Sheaths

Ramon y Cajal (1928) understood that a myelinated peripheral nervous system (PNS) axon "consists of an external solid membrane – sheath of Schwann – of a tubular cell – Schwann cell – of a fatty sheath, and of an axon." He understood that the sheath of Schwann was not part of the Schwann cell, and that it continued across nodes of Ranvier. With the advent of electron microscopy (EM), the sheath of Schwann was shown to be composed of a basal lamina and associated collagen fibers. This basal lamina contains laminin 2 (composed of a trimer of α_2, β_1, and γ_1 laminin chains; Figure 2), type IV collagens, entactin/nidogen, fibronectin, agrin, and glypican. Of these, α_2 and γ_1 laminin are essential because eliminating either one disrupts the formation of the basal lamina and results in developmental defects in myelination.

By analogy to epithelial cells, the basal/abaxonal cell membrane apposes the basal lamina (Figures 1–3). The abaxonal membrane contains at least two receptors for laminin 2 – the integrin $\alpha_6\beta_4$ and dystroglycan (Figures 2 and 3). In epithelia such as the skin, $\alpha_6\beta_4$ heterodimers are found in hemidesmosomes, which join epithelial cells to their basal lamina via intermediate filaments. Schwann cells, however, do not have hemidesmosomes; $\alpha_6\beta_4$ integrin may be linked to the actin cytoskeleton and not to intermediate filaments. Mice lacking either the α_6 or the β_4 integrin subunit die soon after birth because of a profound skin defect, but myelination appears normal in co-cultures established from mice lacking β_4 integrin. Dystroglycan is composed of α and β subunits, which are products of the same gene. As depicted in Figure 2, dystroglycan binds to laminin 2 and to agrin. The intracellular domain of β dystroglycan forms a complex with periaxin, dystroglycan-related protein 2 (DRP2), a short dystrophin isoform (Dp116), utrophin, and β, δ, and ε sarcoglycan. In contrast to the continuous localization of $\alpha_6\beta_4$ integrin and even dystroglycan, DRP2 and periaxin are specifically localized to regions lacking cytoplasm (Figure 4), indicating that DRP2 and/or periaxin may bind to a component of compact myelin.

The importance of the dystroglycan complex in myelinating Schwann cells remains to be fully elucidated, but dystrophin as well as α, β, δ, and γ sarcoglycan are all essential in skeletal muscle cells since the absence of any one of these components results in myopathy. In mice, eliminating dystroglycan or especially the β_1 integrin gene in Schwann cells causes developmental defects in myelination. Recessive mutations in the mouse *Large* gene, one of the genes involved in glycosylation of β dystroglycan, also results in abnormal myelination. In humans, Dp116 may be required for proper myelination since a splice site mutation in the *Dystrophin* gene that abolishes dystrophin expression in peripheral nerve has been reported to cause a demyelinating neuropathy. The best evidence for an essential role of the dystroglycan complex is that recessive *Periaxin* mutations in humans cause an inherited demyelinating neuropathy known as Charcot–Marie–Tooth disease 4F (CMT4F). In addition to a demyelinating neuropathy, *periaxin* null mice also show a loss of the cytoplasmic channels/Cajal bands, the cytoplasmic channels on the outside of the myelin sheath (Figure 4). Cajal bands contain microtubules that provide the highways for trafficking cargo to and from the perinuclear region, where the rough endoplasmic reticulum and Golgi apparatus are located. The cell membranes of Cajal bands are enriched in caveole (and caveolin-1), as well as the K^+ channels Kir2.1, Kir2.3, and a Ca^{2+}-activated K^+ channel. In *periaxin* null mice, the cytoplasm uniformly surrounds the compact myelin, resulting in abnormally short myelin internodes.

Compact Myelin

Figure 5(a) shows a schematic PNS myelinated fiber. It depicts two internodes, one of which has been unrolled to reveal its trapezoidal shape. The myelin sheath has two domains – compact and noncompact myelin – each of which contains a nonoverlapping set

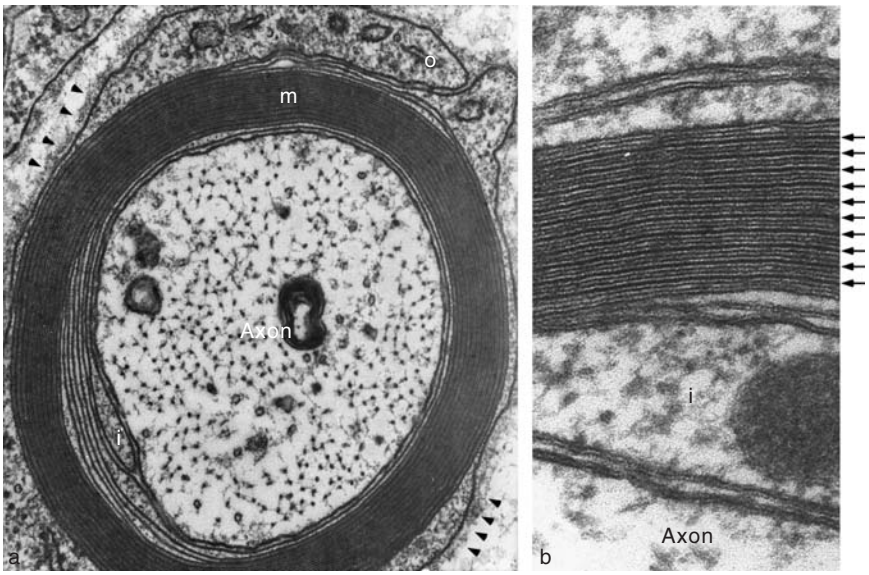

Figure 1 Ultrastructure of a myelinated fiber in the PNS. (a) This electron micrograph shows an axon, its myelin sheath (m), basal lamina (arrowheads), as well as the inner (i) and outer (o) mesaxons (from a 4-day-old mouse). (b) This electron micrograph shows that the compact myelin sheath is composed of alternating major dense (arrows) and intraperiod lines. Note the 'double nature' of the intraperiod line. From SS Scherer (1997) Molecular genetics of demyelination: New wrinkles on an old membrane. *Neuron* 18: 13–16, with permission of Elsevier.

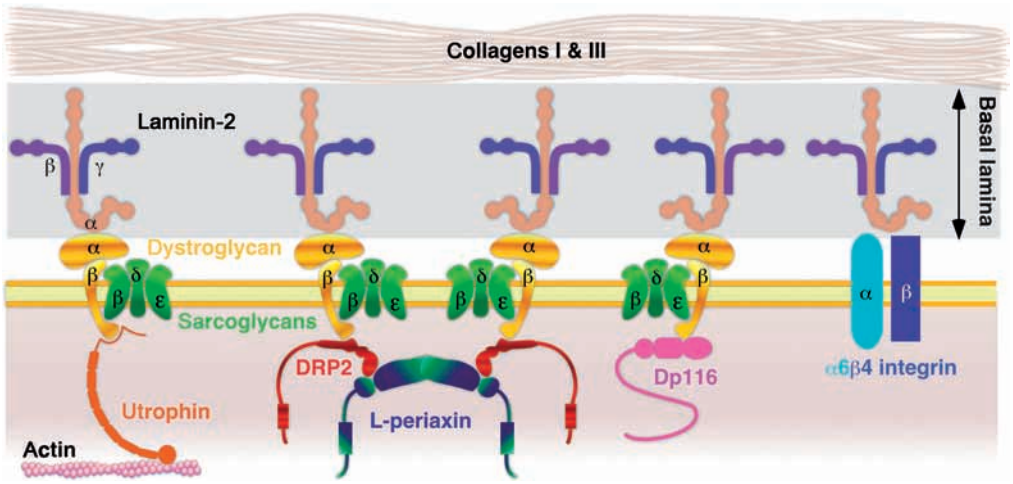

Figure 2 Extracellular matrix receptors of myelinating Schwann cells. The dystroglycan receptor complex of myelinating Schwann cells is linked to three members of the dystrophin family – utrophin, DRP2, and Dp116. Of these, only utrophin has an actin binding domain, whereas DPR2 binds periaxin, which is shown binding to itself via its PDZ domains. Three sarcoglycans – β, δ, and γ – associate with dystroglycan. $\alpha_6\beta_4$ integrin also binds to laminin-2, which is a trimer of α, β, and γ subunits. The figure was modified from the one kindly provided by Drs. Diane Sherman and Peter Brophy.

of proteins (Figure 5(b). Compact myelin forms the bulk of the myelin sheath and is a spiral of closely packed membranes, with a periodicity of approximately 14 nm in conventional transmission EM (Figures 1 and 6) and slightly greater (18 nm) in fresh nerves by X-ray diffraction. The apposed membranes form the intraperiod line; the thin space that separates the membranes is contiguous with the extracellular space. The major dense line is contiguous with the cytoplasm.

Compact myelin is largely composed of lipids, mainly cholesterol and sphingolipids, including galactocerebroside and sulfatide. The targeted disruption of the genes required for the synthesis of sulfatide (*cst* null mice) or galactocerebroside and sulfatide (*cgt*

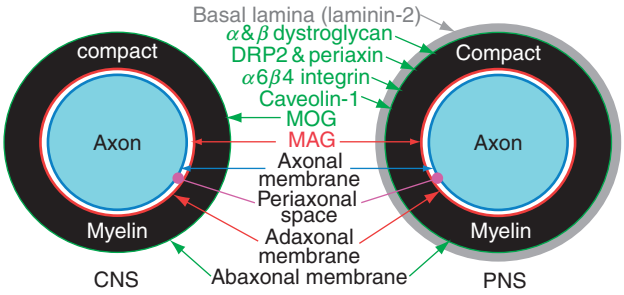

Figure 3 Circumferential organization of myelinating glial cells. This schematic drawing depicts the circumferential organization of a myelinated axon from the CNS and the PNS, including representative molecules. Adapted from Scherer SS and Arroyo EJ (2002) Recent progress on the molecular organization of myelinated axons. *Journal of the Peripheral Nervous System* 7: 1–12, with permission of The Peripheral Nerve Society.

null mice) provided the unprecedented opportunity to perturb the lipid components of myelin. Surprisingly, except for the splitting of some central nervous system (CNS) myelin sheaths, compact myelin has a normal ultrastructure. Despite this normal appearance, axonal conduction velocity is dramatically slowed, which is likely related to abnormal paranodes.

The main myelin proteins of compact myelin are myelin protein zero (P_0), peripheral myelin protein 22 kDa (PMP22), and myelin basic protein (MBP). P_0 is by far the most abundant and is the main adhesive molecule in compact myelin. It is an immunoglobulin (Ig)-like cell adhesion molecule (CAM); its extracellular domain forms tetramers in the plane of the membrane. As shown in Figures 5 and 6, these tetramers interact with each other both in *cis* (in the plane of the membrane) and in *trans* (from apposed membranes). The Schwann cells in *Mpz* null mice still form a multilamellar spiral of membrane around axons, but the myelin is not compact. Mice that are heterozygous for a null *Mpz* allele also develop a late-onset demyelinating neuropathy, indicating that a reduction in P_0 causes instability of compact myelin. Thus, one of the ways that mutations in the human *MPZ* gene could cause demyelination (the disease is called CMT1B) is by haplotype insufficiency resulting in a dose-related decrease in the amount of P_0 protein.

PMP22 is a hydrophobic, intrinsic membrane protein of unknown function, and it is much less abundant than P_0. Nevertheless, the amount of PMP22 in compact myelin is critical. Humans with three (instead of two) copies of the *PMP22* gene have CMT1A, the most common inherited demyelinating neuropathy, probably because there is slightly more PMP22 in compact myelin. Conversely, humans with one copy

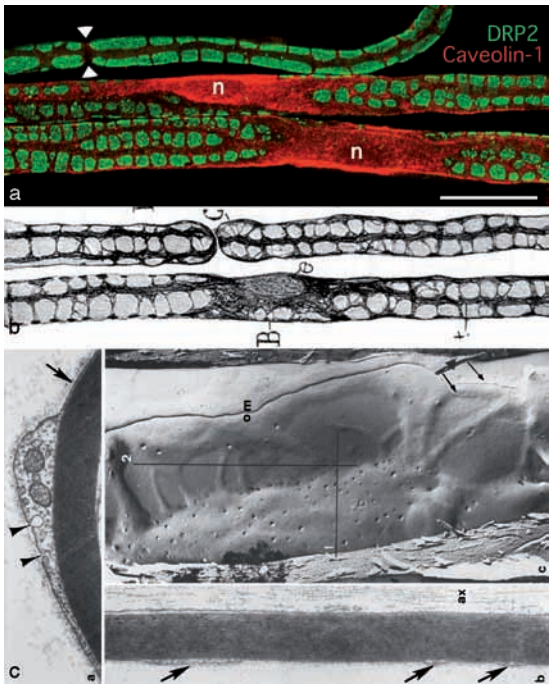

Figure 4 Cytoplasmic domains in myelinating Schwann cells. (a) This is a confocal image of three teased fibers from a rat sciatic nerve, double labeled with a rabbit antiserum against DRP2 (green) and a mouse monoclonal antibody against caveolin-1 (red). The positions of a node (apposed arrowheads) and two nuclei (n) are indicated. Reproduced from Scherer SS and Arroyo EJ (2002) Recent progress on the molecular organization of myelinated axons. *Journal of the Peripheral Nervous System* 7: 1–12, with permission of The Peripheral Nerve Society. Scale bar = 10 μm. (b) This is a drawing made by Ramon y Cajal (1928) of teased fibers from an adult cat, stained with reduced silver. He subdivided the Schwann cell cytoplasm into a 'perinuclear mass,' 'longitudinal stripes,' and 'transverse trabeculae'; as shown in a, the Schwann cell membrane in these regions contains caveolin-1 but not DRP2. Used with permission of Oxford University Press. (c) These are transmission (a and b) and scanning (c) EM images of the outer surface of a myelin sheath. a, Cross-cut myelinated fiber in thin section. Arrowheads indicate two caveolae. b, Longitudinally cut myelinated fiber in thin section. The plane of section passes through three thin cytoplasmic regions (arrows) devoid of large organelles. c, Outer surface of freeze-fracture sheath. The P-face of the Schwann cell membrane has been exposed. Conspicuous convexities of the fracture face studded with caveolar stomata correspond to organelle-rich cytoplasmic regions such as is represented in a. The fingerlike elevations correspond to organelle-poor cytoplasmic regions shown in b. The flat portions in between correspond to semicompact regions. Lines 1 and 2 demarcate planes corresponding approximately to those shown in thin section in a and b, respectively. om, outer mesaxonal furrow. The triangle indicates the cross-fractured mesaxonal cytoplasm, and the arrow indicates the mesaxonal tight junctional strands. Reproduced from Schnapp BJ and Mugnaini E (1978) Membrane architecture of myelinated fibers as seen by freeze-fracture. In: Waxman SG (ed.) *Physiology and Pathobiology of Axons*, pp. 83–123. New York: Raven Press, with permission of Blackwell Publishing.

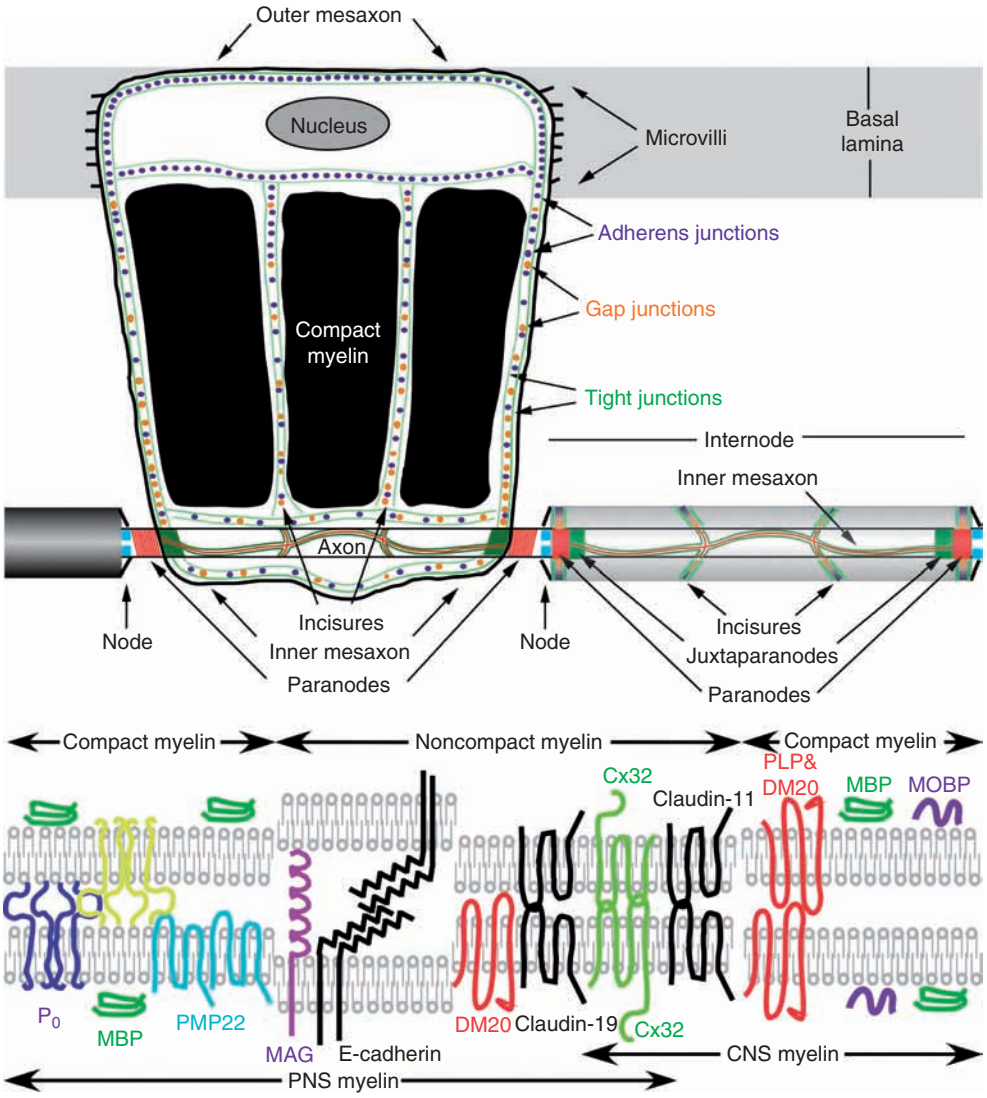

Figure 5 Schematic PNS myelinated axon; the proteins of CNS and PNS myelin sheaths. (a) One myelinating Schwann cell has been unrolled, revealing the regions that form noncompact myelin – the incisures and paranodes. Adherens junctions are depicted as two continuous (purple) lines; these form a circumferential belt and are also found in incisures. Gap junctions are depicted as (orange) ovals; these are found between the rows of adherens junctions. The nodal, paranodal, and juxtaparanodal regions of the axonal membrane are colored blue, red, and green, respectively. (b) CNS and PNS myelin sheaths contain distinct sets of proteins. In the PNS, compact myelin contains P_0, PMP22, and MBP; in the CNS, it contains PLP, MOBP, and MBP. In the PNS, the noncompact myelin contains E-cadherin, MAG, and Cx32; in the CNS, it contains Cx32 and claudin-11. Note that P_0 and MAG have extracellular Ig-like domains (semicircles), and that claudins, PLP, PMP22, and Cx32 all have four transmembrane domains. Adapted from Scherer SS and Arroyo EJ (2002) Recent progress on the molecular organization of myelinated axons. *Journal of the Peripheral Nervous System* 7: 1–12, with permission of The Peripheral Nerve Society.

of the *PMP22* gene have a different kind of inherited demyelinating neuropathy – hereditary neuropathy with liability to pressure palsies – probably owing to a modest decrease of PMP22 in compact myelin. The complete absence of PMP22, which has been created in *Pmp22* null mice (but has not been described in humans), leads to a severe demyelinating neuropathy, further demonstrating an essential role of PMP22 in compact myelin.

MBP is a cytoplasmic protein, and it comprises the bulk of the major dense line of compact myelin (Figures 1 and 6), as shown by its absence in *shiverer* mice, which completely lack MBP because of a recessive mutation in the *Mbp* gene. PNS myelin is not affected in *shiverer* mice at least in part due to the basic cytoplasmic domain of P_0, which may effectively substitute for MBP, as demonstrated by the phenotype of mice that lack both MBP and P_0.

These data demonstrate that both P_0 and PMP22 play essential roles in myelinating Schwann cells, and that a modest reduction in P_0 or PMP22 or a modest increase in PMP22 results in demyelination. Because

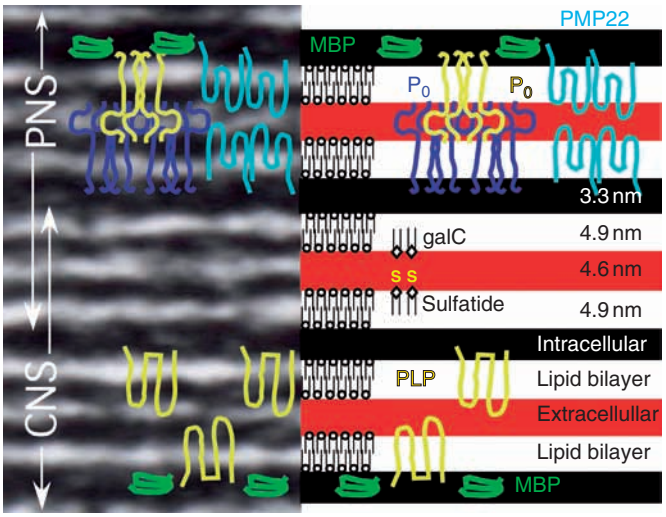

Figure 6 The localization of PNS and CNS myelin proteins in compact myelin. (Left) Electron micrograph of compact (PNS) myelin, which consists of alternating layers known as the intraperiod line (actually a double line) and the major dense line. (Right) Schematic depiction of how apposed cell membranes create the intraperiod and major dense lines. The disposition of P_0 (tetramers), PMP22 (dimers), PLP, and MBP monomers as well as the glycolipids galactocerebroside and sulfatide are shown. The approximate thickness of the lipid bilayers (based on X-ray diffraction), as well as the intracellular and extracellular spaces, is indicated for PNS myelin. Adapted from Scherer SS and Arroyo EJ (2002) Recent progress on the molecular organization of myelinated axons. *Journal of the Peripheral Nervous System* 7: 1–12, with permission of The Peripheral Nerve Society.

perturbations in the stoichiometry of any one component appear to alter the integrity of entire structure, compact myelin has been likened to a liquid crystal. How lipids, PMP22, and the other minor intrinsic membrane proteins 'fit' into the lattice of P_0 tetramers remains to be determined.

Noncompact Myelin

As depicted in Figures 5 and 7, noncompact myelin is found in the paranodes, the lateral borders of the trapezoid, and in Schmidt–Lanterman incisures (the funnel-shaped interruptions in the compact myelin). Noncompact myelin contains adherens junctions, tight junctions, and gap junctions; these join adjacent layers of the myelin sheath. Some of these junctions are also found in the inner and/or outer mesaxons, where the innermost or outermost layers of the myelin membrane contact the adjacent layer.

Adherens junctions (or 'desmosome-like' junctions in the older literature) typically span many layers and are most prominent in the outer layers of the myelin sheath (Figures 5 and 7). The adherens junctions of myelinating Schwann cells are composed of E-cadherin, perhaps the most prevalent member of the cadherin family. E-cadherin is a Ca^{2+}-dependent CAM that forms 'strand dimers' that bind homophilically with those on the apposing membrane (in *trans*). The cytoplasmic domain of E-cadherin binds p120 catenin and β-catenin; the latter links E-cadherin to the actin cytoskeleton. The role of adherens junctions

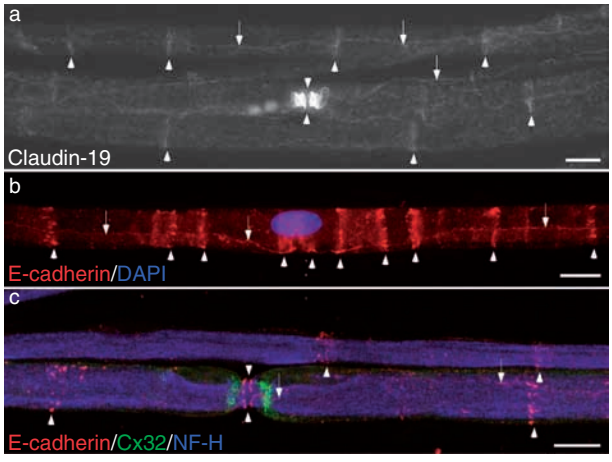

Figure 7 Tight junctions, adherens junctions, and gap junctions in PNS myelin sheaths. These are teased fibers that have been immunostained for claudin-19, E-cadherin, connexin32 (Cx32), and neurofilament heavy (NF-H), as indicated; DAPI is a nuclear counterstain. Claudin-19 and E-cadherin labeling are found in incisures (arrowheads) and mesaxons (arrows). Claudin-19, E-cadherin, and Cx32 labeling are found at paranodes (nodes are marked by apposing arrowheads). Scale bars = 10 μm.

is uncertain because deleting the E-cadherin gene in myelinating Schwann cells does not have an obvious effect on myelin sheaths. On the other hand, expressing a dominant negative p120 in myelinating Schwann cells has been reported to result in a loss of incisures.

Tight junctions limit the diffusion of ions and small molecules between apposed cell membranes, the paracellular pathway (Figures 5 and 7). A family

of molecules, the claudins, are the essential functional component of tight junctions. In myelinating Schwann cells, claudin-19 forms the tight junctions between the adjacent paranodal loops as well as at the inner and outer mesaxons; it is unsettled whether claudin-5 forms the tight junctions in some incisures. Claudins have a PDZ binding domain, and claudin-19 (and claudin-5) may bind to PATJ, MUPP1, ZO-1, ZO-2, and Par-3, all of which are PDZ proteins localized to paranodes and/or incisures. Except for the lack of tight junctions, the PNS myelin sheaths of claudin-19 null mice appear normal. The mice have subtle motor findings, but a subset of myelinated axons have slowed conduction, indicating that the tight junctions between the layers of the myelin sheath contribute to the normal propagation of action potentials. Curiously, no neurologic phenotype has been noted in humans with homozygous recessive mutations in claudin-19.

Small gap junction plaques had been noted between the rows of tight junctions in the PNS myelin sheath by freeze-fracture EM for decades. Their obscurity ended when mutations in *GJB1* were discovered to cause CMT1X, an inherited demyelinating neuropathy. *GJB1* encodes connexin32 (Cx32), which belongs to a family of approximately 20 connexin genes that encode gap junction proteins. Gap junctions formed by Cx32 are found in the paranodes and incisures of myelinating Schwann cells (Figures 5 and 7) and appear to provide a direct pathway for the diffusion of ions and small molecules across the layers of the myelin sheath. Mutations in *GJB1* may disrupt the diffusion of small molecules across the myelin sheath. Cx29 (the human ortholog is Cx31.3) is also localized to paranodes and incisures, but its functional role is unclear because it does not appear to form functional gap junctions on its own and does not appear to interact with Cx32.

The CNS Myelin Sheath

Unlike Schwann cells, which generate only one myelin sheath, each oligodendrocyte typically makes multiple myelin sheaths; the number varies from tract to tract and appears to relate to the caliber of the axons. Oligodendrocytes make fewer sheaths in tracts containing large myelinated fibers, which is the result of axon–oligodendrocyte interactions rather than an intrinsic trait of the oligodendrocytes themselves, as shown by transplantation studies. Oligodendrocytes also differ from Schwann cells in that their myelin sheaths lack a basal lamina, and their incisures lack distinguishing molecular markers.

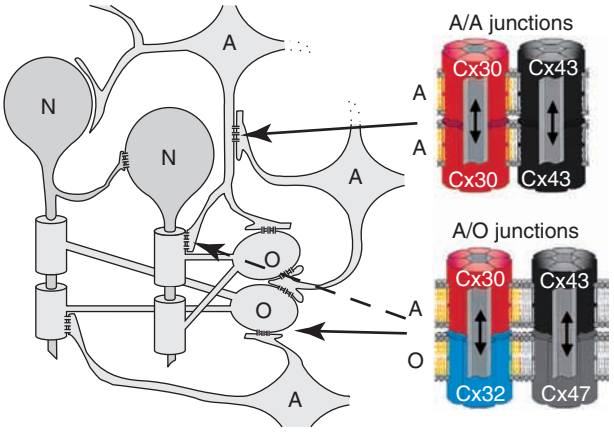

Figure 8 Astrocyte/oligodendrocyte coupling. This figure shows that astrocytes (A) are coupled to astrocytes and oligodendrocytes (O) but not to neurons (N). Note that oligodendrocytes are not coupled to themselves and that A/A junctions are homotypic (Cx30/Cx30 or Cx43/Cx43), whereas A/O junctions are heterotypic (Cx30/Cx32 or Cx43/Cx47). Adapted from Rash JE, Yasumura T, Dudek FE, and Nagy JI (2001) Cell-specific expression of connexins and evidence of restricted gap junctional coupling between glial cells and between neurons. *Journal of Neuroscience* 21: 1983–2000, with permission of The Society for Neuroscience.

Another key difference is that oligodendrocytes are coupled to astrocytes (but not to themselves) by gap junctions. As depicted in Figure 8, astrocyte/oligodendrocyte (A/O) gap junctions are probably formed by two discrete pairs of connexins – Cx43/Cx47 and Cx30/Cx32. Astrocytes express Cx43 and Cx30; oligodendrocytes express Cx47 and Cx32. Cx47 is mainly found on oligodendrocyte somata, whereas Cx32 is also found on the outer aspects of myelin sheaths. There is compelling genetic evidence that A/O coupling is essential because recessive mutations in *GJA12/Cx47* cause a devastating CNS dysmyelinating disease in humans known as Pelizaeus–Merzbacher-like disease, and mice that lack both Cx32 and Cx47 develop severe CNS demyelination. The functional role of these A/O junctions is incompletely understood, but buffering the K^+ that is released during neural activity has been implicated. In addition, Cx32 is localized to the paranodal region of CNS myelin sheaths, although not as prominently as in PNS myelin sheaths.

CNS myelin sheaths are polarized. In addition to Cx32 and Cx47, their outer cell membrane contains myelin–oligodendrocyte glycoprotein (MOG), and their inner cell membrane contains myelin-associated glycoprotein (MAG) and Nogo (Figure 3).

The molecular components of compact myelin in the CNS myelin sheaths partially overlap with those of the PNS (Figures 5 and 6). Both contain high amounts of lipids, especially cholesterol and sphingolipids, including galactocerebroside and sulfatide.

P_0 and PMP22 are not present in CNS compact myelin; instead, proteolipid protein (PLP) and its alternatively spliced isoform, DM20, are the main proteins. Conversely, PLP is not present in PNS myelin, although DM20 is present in noncompact PNS myelin. Both PLP and DM20 are highly hydrophobic, intrinsic membrane proteins. Furthermore, either deletion or duplication of the human PLP gene results in CNS dysmyelination (Pelizaeus–Merzbacher disease), demonstrating again the importance of gene dosage for myelin proteins. Remarkably, CNS myelin appears nearly normal in *Plp* null mice – a finding that underscores our lack of knowledge about the function of PLP and DM20. As shown in Figure 5, the major dense line contains MBP and myelin–oligodendrocyte basic protein (MOBP). As demonstrated by the severe dysmyelinating in *shiverer* mice, MBP plays an essential role in CNS myelin, whereas CNS myelin appears normal in *Mobp* null mice.

CNS myelin has a distinctive structural feature that is not seen in PNS myelin – the so-called 'radial component.' This is a series of radially arranged, intralamellar tight junctions that span the thickness of the compact myelin, usually in a single sector (Figure 9). These tight junctions are composed of claudin-11, which also forms tight junctions between the paranodal loops. Like claudin-19 null mice, the neurological findings in claudin-11 null mice include mild conduction slowing. In contrast to mice lacking either PLP or claudin-11 alone, compact myelin in mice lacking both PLP and claudin-11 is decompacted, indicating that both of these proteins stabilize its structure.

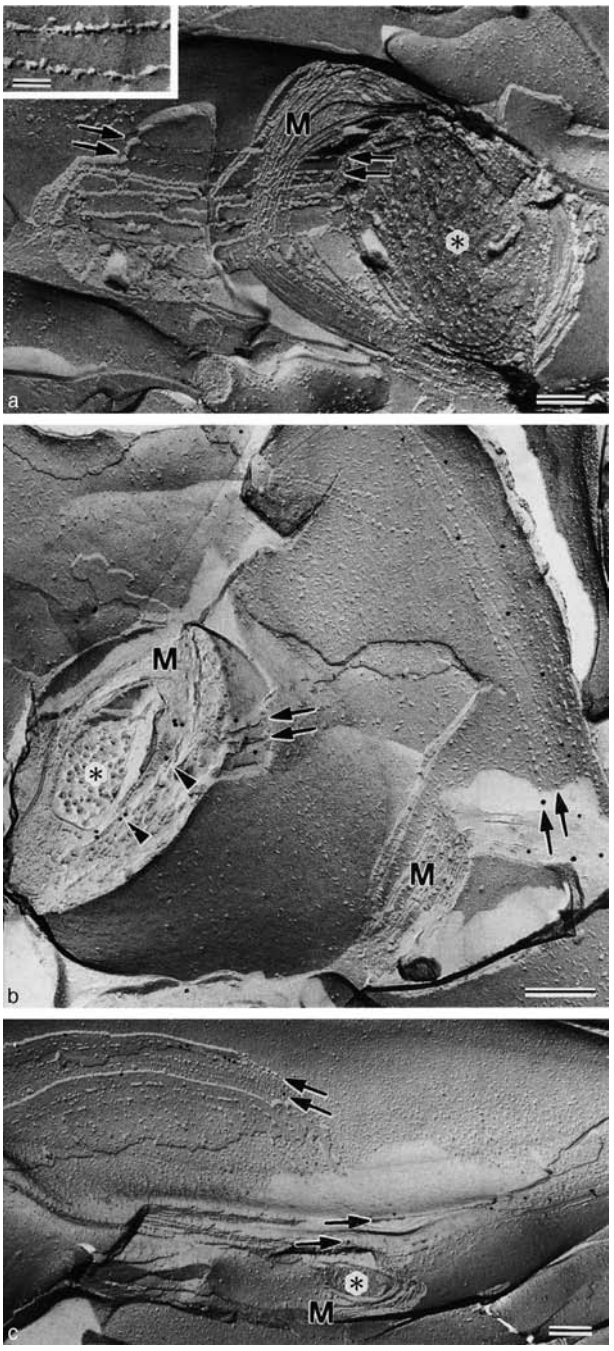

Figure 9 Claudin-11 is localized to the radial component of compact myelin. (a) Freeze-fracture EM of a mouse brain. A series of interlamellar strands (arrows; enlarged in the inset) are found between each layer of the compact myelin sheath. (b and c) Immunoelectron microscopy localizes claudin-11 (arrows) to the interlamellar strands of freeze-fractured mouse brain (b) and optic nerve (c). The asterisks indicate axons. Reproduced from Morita K, Sasaki H, Fujimoto K, Furuse M, and Tsukita S (1999) Claudin-11/OSP-based tight junctions of myelin sheaths in brain and Sertoli cells in testis. *Journal of Cell Biology* 145: 579–588, with permission of The Rockefeller University Press.

Axonal Specializations at Nodes of Ranvier

Despite the differences between PNS and CNS myelin sheaths, the organization of the nodal region is quite similar. Transmission EM of the nodal axolemma reveals an electron-dense undercoating and scanning EM reveals a high density of large intramembranous particles (1200 μm^2). These large particles are mainly voltage-dependent Na^+ (Na_V) channels, which comprise a family of ten genes; however, $Na_V1.6$ appears to be the main one because it is expressed at all mature nodes. The gene encoding $Na_V1.6$ is mutated in mice with a recessively inherited motor endplate disease (*med*), which die of respiratory failure approximately 20 days after birth. $Na_V1.6$ is probably not the only nodal Na_V channel because nerve conduction velocity of sensory axons is only minimally slowed in *med* mice; $Na_V1.2$ and $Na_V1.8$ are the best substantiated.

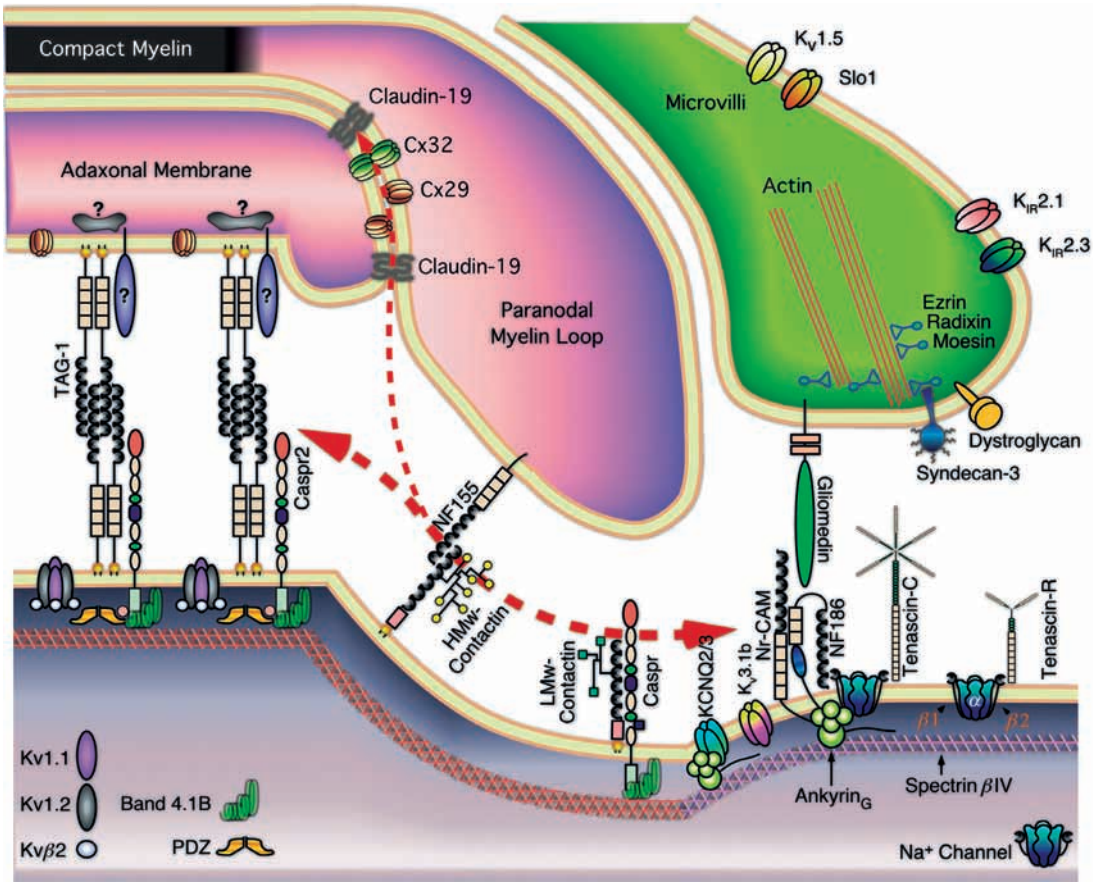

Figure 10 The molecular organization of the nodal region. Schematic drawing of possible *cis* and *trans* interactions between the molecular components of nodes, paranodes, and juxtaparanodes. The pathways for diffusion of molecules through the paranodal tight junctions and the septate-like junctions are shown by the small and large red dotted lines, respectively. Adapted from Scherer SS and Arroyo EJ (2002) Recent progress on the molecular organization of myelinated axons. *Journal of the Peripheral Nervous System* 7, 1–12, with permission of The Peripheral Nerve Society.

The actual Na_V channel is composed of a single α subunit – a large, glycosylated, polytopic protein composed of four homologous domains. Mature channels are composed of one α and two β subunits (Figure 10); the latter are members of the Ig CAM superfamily. Four genes encode different β subunits; β_1 and β_3 are more closely related to each other than to β_2 or β_4. Both β_1 and β_2 have been localized to nodes; whether β_3 or β_4 are localized to nodes remains to be determined. In addition, it is unknown whether different combinations of β subunits have different interactions with, and thus different effects on, $Na_v1.2$, $Na_v1.6$, and $Na_v1.8$.

A variety of functions have been proposed for β subunits. In transfected cells, β subunits increase the delivery of α subunits to the cell membrane and can alter their electrophysiological characteristics. As CAMs, β subunits have multiple potential interactions with various nodal components: β_1 and β_2 can also interact homophilically and heterophilically with the extracellular matrix molecules tenascin-C and tenascin-R; β_1 and β_3 (but not β_2) can interact in *cis* with neurofascin and Nr-CAM; and β_1 (but not β_2) can interact with contactin and phosphacan, a secreted isoform of receptor protein tyrosine phosphatase β (RPTPβ). Furthermore, the cytoplasmic domains of β_1 and β_2 can recruit ankyrin-G. It remains to be determined which of these potential interactions play important roles at nodes of Ranvier, but the findings that β_2 is required for normal density of nodal Na_v channels *in vivo* and that β_4 interacts with $Na_V1.6$ to produce the 'resurgent' current that enables action potentials to be generated at high rates are provocative.

Two members of the L1 family of CAMs, neurofascin186 and Nr-CAM, are found in the nodal axolemma. Both bind to the ankyrin-G isoforms that are found at nodes (ankyrin-G 480 and 270 kDa). The binding of ankyrin-G 480/270 kDa to neurofascin can be modulated since phosphorylation of a highly conserved tyrosine in neurofascin abolishes binding. As depicted in Figure 10, these ankyrin-G isoforms also

interact with the α subunits of Na_V channels and with the K^+ channels KCNQ2 and KCNQ3. Like Na_V channels, these K^+ channels are localized to nodes and axon initial segments; KCNQ2 appears to be localized to all nodes and initial segments, whereas KCNQ3 is localized to a subset. KCNQ2 and/or KCNQ3 likely produce the slow K^+ current (K_s) that has been recorded from patches of nodal membrane. The importance of K_s is shown by a dominant mutation in the human *KCNQ2* gene; the mutant KCNQ2 reduces the activity of both KCNQ2 and KCNQ3, and it produces repetitive firing of motor axons (known as neuromyotonia or neuromyokymia). Both KCNQ2 and KCNQ3 have an amino acid sequence in their C-terminus that is strikingly similar to the sequence (found in the intracellular loop between the second and third domains) of Na_V channels that binds to ankyrin-G; this sequence is necessary and sufficient for binding to ankyrin-G 480/270.

Ankyrins are a family of three genes, each with extensive alternative splicing. The ankyrin-G 480/270 isoforms found at nodes and initial segments are distinguished by ANK repeats, a spectrin-binding domain, and a serine/threonine-rich domain. Ankyrins interact with the spectrin cytoskeleton, a lattice composed of α and β heterodimers that are localized just beneath the cell membrane. There are two α and four β spectrin genes, leading to many potential combinations. One particular splice variant of spectrin βIV (βIVΣ1) is specifically localized to initial segments and nodes, and it presumably binds ankyrin-G 480/270. Evidence for a role for spectrin βIV in this regard comes from the analysis of *quivering* mice, in which recessive mutations in the *spectrin 4* gene cause altered ion channel distributions in myelinated axons.

These data indicate that ankyrin-G 480/270 is a multivalent protein and links CAMs (neurofascin186, Nr-CAM, and β subunits of Na_V channels) and ion channels (Na_V, KCNQ2, and KCNQ3) to the spectrin cytoskeleton at nodes. In keeping with this idea, inactivation of the ankyrin-G gene in the cerebellum results in the mislocalization of neurofascin186, Nr-CAM, $Na_V1.6$, KCNQ2, KCNQ3, and spectrin βIVΣ1 in the initial segments of Purkinje cells. Since initial segments and nodes share many molecular characteristics, the nodal membranes of Purkinje cells should be similarly affected.

Glial Specializations at Nodes of Ranvier

The lateral borders of the Schwann cell cytoplasm are tipped with microvilli (Figures 5 and 10). These contain F-actin and the ERM proteins erzin, radixin, and moesin, which probably link intrinsic membrane proteins such as syndecan-3 and -4 to the actin cytoskeleton. Electron microscopists have long known that the tips of these microvilli contact the nodal axolemma, indicating that there are molecular interactions between the two cell types. These interactions likely begin in developing nerves, as clusters of Na_V channels are found at the lateral edges of developing myelin sheaths; the clusters (heminodes) of two adjacent internodes fuse to form a node. Furthermore, clusters of neurofascin and Nr-CAM appear before those of ankyrin and Na_V channels. An intrinsic membrane protein called gliomedin has been found based on its ability to bind to both neurofascin and Nr-CAM. Gliomedin is expressed by Schwann cells, localized to their microvilli, appears before ankyrin and Na_V channels, and is both necessary and sufficient to cluster neurofascin186, Nr-CAM, ankyrin-G, and Na_V channels. Thus, Schwann cells appear to induce the formation of nodes via their focal expression of gliomedin. Dystroglycan also appears to play a role at nodes because deleting dystroglycan in Schwann cells results in disorganized microvilli, reduced density of Na_V channels, and slowed conduction.

In the CNS, the processes of 'perinodal astrocytes' (astrocytes and glial progenitors) have been postulated to have a similar function as Schwann cell microvilli. These processes express tenascin-R, which can bind to Na_v channels and alter their electrophysiological properties; this may be the reason why myelinated optic axons in mice lacking tenascin-R have slowed conduction. Although CNS nodes contain RPTPβ, which can interact with tenascin-R and both the α and the β subunits of Na_v channels, the distribution of nodal Na_v channels as well as conduction velocity of CNS myelinated axons are normal in RPTPβ-deficient mice. Both tenascin-R and RPTPβ also interact with the high-molecular-weight form of contactin, which (in contrast to the PNS) is found at CNS nodes. Contactin regulates the surface expression of Na_v channels, and mice lacking contactin have reduced nodal Na_v channels clusters.

Specializations at Paranodes

At the paranode, the lateral edge of the myelin sheath contacts the axon; a loop of cell membrane slightly indenting the axon as it spirals around it, forming the axoglial junctions (Figure 11). For most myelin sheaths, the entire lateral edge remains in contact with the paranodal axolemma; some of the paranodal loops of the largest PNS myelin sheaths, however, do not touch the axonal membrane. A unique specialization known as septate-like junctions (or transverse bands) is found at axoglial junctions; freeze-fracture EM reveals that these are composed of rows of large

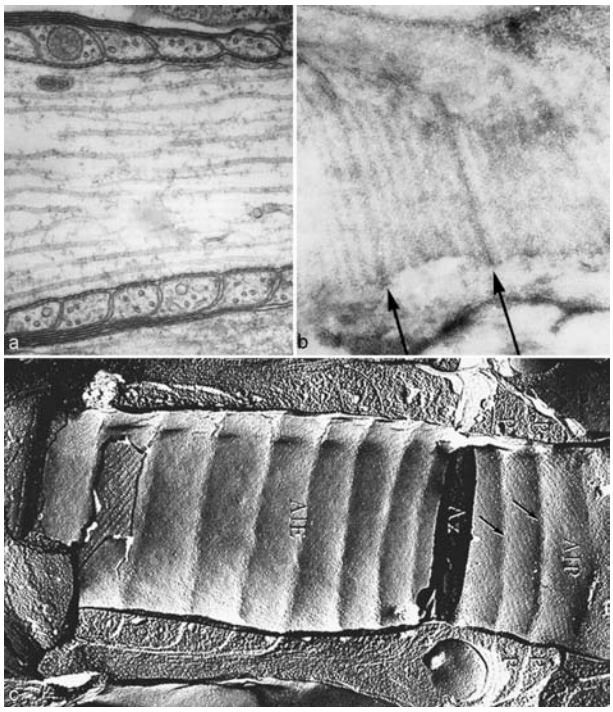

Figure 11 The ultrastructure of the paranode. (a) A transmission electron micrograph showing a CNS paranode in longitudinal section. The terminal loops of the myelin sheath indent the axonal membrane, to which they are connected to septate-like junctions. (b) A transmission electron micrograph showing lanthanum outlining the space between the septate-like junctions of CNS axons. The arrows indicate the locations of the spaces between adjacent glial loops; these spaces are larger than the spaces between adjacent rows of septate-like junctions. (c) A longitudinal freeze-fracture EM through a CNS paranode. The P face of the paranodal axolemma (AIP) shows scalloping (arrows) produced by the pockets of paranodal cytoplasm that are visible at the sides (P). The fracture plane then passes across the axon (Ax) so that the B face (AIE) of the axolemma on the other side comes into view. (a) Reproduced from Hirano A and Dembitzer HM (1967) A structural analysis of the myelin sheath in the central nervous system. *Journal of Cell Biology* 34, 555, with permission of Rockefeller University Press. (b) Reproduced from Hirano A and Dembitzer HM (1982) Further studies on the transverse bands. *Journal of Neurocytology* 11, 861–866, with permission of Elsevier. (c) Reproduced from Peters A, Palay SL, and Webster H de F (1991) *The Fine Structure of the Nervous System*, 3rd edn. New York: Oxford University Press, with permission from Oxford University Press.

particles on the glial membrane that are in register with a double row of smaller particles on the axolemma. These junctions not only resemble invertebrate septate junctions but also contain similar molecules that are expressed by the axon or the myelinating Schwann cell: Contactin (a GPI-linked protein) and contactin-associated protein (Caspr, an intrinsic membrane glycoprotein) are localized to the paranodal axolemma, and neurofascin-155 (an intrinsic membrane protein) is localized to the paranodal loops. There is some uncertainty regarding whether neurofascin-155 binds to contactin or to contactin–Caspr heterodimers, but eliminating any of these three proteins in mice demonstrates that all of them are essential for the formation of septate-like junctions. Gene knockout studies in mice also show that galactocerebroside and sulfatide, as well as the tetraspanin protein CD9, are also required for septate-like junctions; all are expressed by myelinating Schwann cells and may be required to deliver and/or stabilize neurofascin-155 to septate-like junctions. In the absence of septate-like junctions, the paranodal loops of myelinating Schwann cells are slightly separated from the axolemma, whereas those of oligodendrocytes are often detached.

The cytoplasmic domain of Caspr contains a binding site for protein 4.1, which in turn can bind to the spectrin cytoskeleton. Of the family of 4.1 proteins, only 4.1B has been found in axons, and it is found throughout the internode but not at nodes (**Figure 10**). The cytoplasmic domain of Caspr also contains a SH3 binding site, which could allow Caspr to participate in signaling pathways within the axon.

Septate junctions function similarly to vertebrate tight junctions, forming intercellular junctions that prevent the diffusion of small molecules and ions. Because septate-like axoglial junctions do not prevent the diffusion of lanthanum (**Figure 11**) or microperoxidase into the periaxonal space, they are not 'tight' in the conventional sense. In addition to limiting the diffusion of large molecules, paranodal septate-like junctions have also been shown to sequester Kv1.1 and Kv1.2 K^+ channels. In the absence of septate-like junctions, erstwhile juxtaparanodal Kv1.1/Kv1.2 channels occupy the paranodal region, thereby apposing nodal Na_V channels, leading to current shunting and the slowing of action potential propagation.

Specializations at Juxtaparanodes

By freeze-fracture EM, the axolemma in the region extending 10–15 μm from the paranode contains clusters of five or six particles. The distribution of these juxtaparanodal particles corresponds to the distribution of delayed rectifying K^+ channels, likely composed of *Shaker*-type Kv1.1 and Kv1.2, which mix in varying proportions to form tetramers, the functional channels. The size of the particles (10 nm in diameter) compares well to the expected size of K^+ channels. Each Kv1.1 and 1.2 subunit is associated with a Kvβ2 subunit. Although they appear to be concealed under the myelin sheath, juxtaparanodal Kv1.1/Kv1.2 channels are thought to have an important physiological function, dampening the

excitability of myelinated fibers. The finding that *Kcnac1/Kv1.1* null mice have abnormal impulse generators near the neuromuscular junctions supports this idea. Similarly, mutations in the human *KCNAC1* gene cause a form of familial episodic ataxia that is associated with ectopic impulse generators in the distal aspects of motor axons, causing neuromyotonia/neuromyokymia.

Caspr2 and TAG-1 (transiently expressed glycoprotein-1, an Ig CAM) are also found in the juxtaparanodal axolemma in both the CNS and the PNS. Caspr2 has a similar domain structure to Caspr, especially in the extracellular region; its intracellular C-terminus contains a band 4.1 binding domain, a PDZ binding domain, but not an SH3 domain (Figure 10). Caspr2 co-localizes and even co-immunoprecipitates with Kv1.1 and Kv1.2, perhaps because they all bind to a common PDZ protein. Although PSD95 is localized to juxtaparanodes, Caspr2, Kv1.1, Kv1.2, and Kvβ2 are not mislocalized in mice that lack PSD95. Rather, *cis* interactions of Caspr2 with homodimers of TAG-1 may localize the complex of Caspr2/Kv1.1/Kv1.2/Kvβ2 to juxtaparanodes because TAG-1 can form homodimers with TAG-1 on the adaxonal membrane of the myelin sheath (Figure 10). In keeping with this model, Kv1.1 and Kv1.2 are mislocalized in mice that lack either TAG-1 or Caspr2 but not in mice that lack Kv1.1, Kv1.2, or Kvβ2.

The adaxonal Schwann cell membrane in the juxtaparanodal region contains hexamers (12 nm in diameter) by freeze-fracture EM. These particles have the appearance of hemichannels and are composed of Cx29. The importance of these putative hemichannels is unknown because the myelinated axons in *cx29* null mice appear normal, and how such hemichannels might function is a matter of speculation. Because TAG-1 and Cx29 are unlikely to interact directly, there must be other proteins that mediate this interaction (Figure 10).

The Internodal Region

The apical/adaxonal surface of the myelin sheath apposes the axon and is highly enriched in MAG in both the PNS and the CNS (Figure 3). MAG can bind to certain complex gangliosides, including GD1a and GT1b, but it remains to be shown that these are preferentially distributed on the internodal axonal membrane. In the CNS, the adaxonal membrane is also enriched in the Nogo, which may bind to the Nogo receptor on the internodal axonal membrane.

The internodal axonal membrane lacks the conspicuous structural specializations of the nodal region. Nevertheless, at least one important molecule, Na,K-ATPase, is distributed rather uniformly throughout the internode of myelinated axons, including the juxtaparanodes. Na,K-ATPase is composed of an α and a β subunit; the α subunit exchanges intracellular Na^+ for extracellular K^+. In the PNS, myelinated axons express α_1, α_3, or a combination of the two; in the CNS, myelinated axons express α_3. How α subunits are tethered to the internodal axonal cytoskeleton is unknown; in other cell types, Na,K-ATPase interacts with ankyrins.

In the PNS, a tripartite strand of glial proteins – consisting of a central strand of NF155 flanked on either side by strands of Cx29 – is localized on the adaxonal Schwann cell membrane at the inner mesaxon and at the innermost aspect of incisures (Figure 12). A tripartite strand of axonal proteins – consisting of a central strand of Caspr staining, flanked on either side by strands of Kv1.1/1.2/β2 and Caspr2 immunoreactivity – apposes the inner mesaxon ('juxta-mesaxonal') and the innermost aspect of incisures ('juxta-incisural'). This precise apposition suggests that the same *trans*-molecular interactions between myelinating Schwann cells and axons that are found at paranodes and juxtaparanodes extend into the internodal region. In contrast, there is no mesaxon-like staining of Cx29 or NF155 in CNS myelin sheaths,

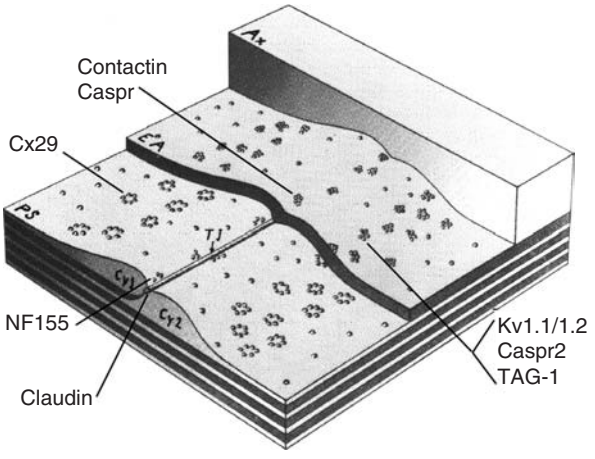

Figure 12 Internodal specialization of myelinating Schwann cells. This is a schematic summary of the putative identities and distributions of intramembranous particles on the axonal and adaxonal Schwann cell membranes in the internodal region, as visualized by freeze-fracture electron microscopy. The location of the tight junction (TJ) at the inner mesaxon is indicated. Adapted from Stolinski C, Breathnach AS, Thomas PK, Gabriel G, and King RH (1985) Distribution of particle aggregates in the internodal axolemma and adaxonal Schwann cell membrane of rodent peripheral nerve. *Journal of Neurological Sciences* 67: 213–222, with permission of Elsevier.

and myelinated CNS axons only have internodal staining of Caspr (and not Kv1.1/1.2/β2 and Caspr2) at the inner mesaxon.

See also: Demyelinating Diseases; Demyelination and Demyelinating Antibodies; Neuropathy: Peripheral; Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells; Oligodendrocyte and Schwann Cell Identification Methods; Oligodendrocyte Morphology; Schwann Cell Development; Schwann Cell Morphology; Schwann Cells and Axon Relationship; Schwann Cells and Plasticity of the Neuromuscular Junction; Transplantation of Myelin Forming Cells.

Further Reading

Bennett V and Baines AJ (2001) Spectrin and ankyrin-based pathways: Metazoan inventions for integrating cells into tissues. *Physiological Reviews* 81: 1353–1392.

Dyck PJ and Thomas PK (2005) *Peripheral Neuropathy*, 4th edn. Philadelphia: Saunders.

Hille B (2001) *Ion Channels of Excitable Membranes*. Sunderland, MA: Sinauer.

Kamasawa N, Sik A, Morita M, et al. (2005) Connexin-47 and connexin-32 in gap junctions of oligodendrocyte somata, myelin sheaths, paranodal loops and Schmidt–Lanterman incisures: Implications for ionic homeostasis and potassium siphoning. *Neuroscience* 136: 65–86.

Lazzarini RA (2004) *Myelin Biology and Disorders*. San Diego: Elsevier.

Peters A, Palay SL, and Webster H de F (1991) *The Fine Structure of the Nervous System*, 3rd edn. New York: Oxford University Press.

Poliak S and Peles E (2003) The local differentiation of myelinated axons at nodes of Ranvier. *Nature Reviews Neuroscience* 4: 968–980.

Previtali SC, Feltri ML, Archelos JJ, Quattrini A, Wrabetz L, and Hartung HP (2001) Role of integrins in the peripheral nervous system. *Progress in Neurobiology* 64: 35–49.

Ramón y and Cajal S (1928) May RM (ed.). *Degeneration and Regeneration of the Nervous System*. Oxford: Oxford University Press.

Salzer JL (2003) Polarized domains of myelinated axons. *Neuron* 40: 297–318.

Scherer SS and Arroyo EJ (2002) Recent progress on the molecular organization of myelinated axons. *Journal of the Peripheral Nervous System* 7: 1–12.

Schnapp BJ and Mugnaini E (1978) Membrane architecture of myelinated fibers as seen by freeze-fracture. In: Waxman SG (ed.) *Physiology and Pathobiology of Axons*, pp. 83–123. New York: Raven Press.

Waxman SG, Kocsis JD, and Stys PK (1995) *The Axon*. Oxford: Oxford University Press.

Willison HJ and Yuki N (2002) Peripheral neuropathies and anti-glycolipid antibodies. *Brain* 135: 2591–2625.

Relevant Websites

http://www.ncbi.nlm.nih.gov – National Center for Biotechnology Information.

http://www.neuro.wustl.edu – Neuromuscular Disease Center of Washington University in St. Louis.

http://www.molgen.ua.ac.be – University of Antwerp, Molecular Genetics Department.

Myopia

J Wallman, City University of New York, New York, NY, USA

© 2009 Published by Elsevier Ltd.

Introduction

Myopia is an important topic not only because it afflicts a disproportionately large fraction of educated people but also because myopia research has shown that the visual system has a function beyond light reception and perception, that of the homeostatic control of eye growth. Under normal circumstances, this neural control adjusts the growth of the eye so that the length of the eye matches the focal length of its optics, causing distant objects to be imaged on the photoreceptor layer of the retina, a condition called emmetropia. In myopia, the eye is too long relative to its focal length, causing distant objects to be imaged in front of the photoreceptors. Only nearby objects are in focus on the photoreceptors.

Much evidence points to the eye growing under homeostatic control. Thus if a neonate is born hyperopic (ocular length shorter than focal length), the eye elongates more rapidly than usual, thus overtaking the increasing focal length and correcting the hyperopia. Conversely, if an eye starts out myopic (ocular length greater than focal length), the rate of elongation is decreased, allowing the normal increase in focal length to correct the myopia. It is clear that vision provides much of the feedback guidance of growth. This has been shown by having young animals wear spectacle lenses that displace the plane of focus of the image, creating either hyperopia (image focused behind the photoreceptors) by negative lenses or myopia (image focused in front of the photoreceptors) by positive lenses. Within days or weeks, depending on the species, the eye increases or decreases its rate of growth, bringing the eye toward emmetropia while the animal is wearing the lenses. This control occurs in chicks, fish, guinea pigs, tree shrews, marmosets, and rhesus monkeys. On this basis, it is difficult to argue that it does not also occur in human children. Furthermore, another type of visually induced myopia occurs when animals are deprived of vision, but not of light, as by wearing diffusers over their eyes (form deprivation myopia). This also occurs in children with visual deprivations caused by cataracts, eyelid problems, and so forth.

Before this animal work, there were two principal views of human myopia. The predominant view, especially among ophthalmologists, was that myopia was a largely genetic condition, either in the sense of genetic determination of eye shape or in the sense of genetic defects of scleral structure, leading to excessive expansion as a consequence of the intraocular pressure and thus to excessive ocular elongation. The alterative view, frequently considered radical, was that excessive reading or other visual tasks involving viewing nearby objects perhaps drove the emmetropization mechanism to make the eye 'emmetropic' for the distance of the page. The radical aspect of these theories was that they led to the notion that myopia was partly iatrogenic, in that the fitting of progressively stronger corrective lenses by the physician or optometrist prevented the myopia from reaching an asymptote at which the page was naturally in focus without requiring accommodation. The consequence of the experimental work on animals in the past three decades has been that blur is now seen as an important cause of myopia.

This article summarizes first the experimental work on animals and then the studies on humans and then speculates on how the former can inform the latter.

Experimental Studies on Animals

All Blur Is Not Alike

As implied by the spectacle-lens compensation experiments mentioned earlier, the old story that myopia results simply from unsharp images is almost certainly untrue. Although eyes of many species (monkeys, guinea pigs, tree shrews, chicks, fish) become myopic by wearing negative lenses, eyes wearing positive lenses become hyperopic even when the environment is arranged to make all images blurred (at least in chick and tree shrew). These results imply that the eye can grow to compensate for opposite signs of blur (whether the image is focused in front or behind the retina), and it suggests that the sign of blur is more important than its magnitude. Furthermore, even though wearing diffusers or negative lenses causes myopia, there is evidence from experiments in chicks that form deprivation myopia is different from lens compensation myopia in that various manipulations of lighting and also optic nerve section have differential effects on the two types of myopia.

Spatial Factors: Myopia Is Local

When chick eyes that are blind because of optic nerve section wear diffusers or spectacle lenses, they

become myopic or hyperopic, much as intact eyes would, although some differences exist. This finding means that however the visual system determines that defocus is present and whether the sign of defocus is myopic or hyperopic, it must do it with the retina alone, without the processing power of the brain. Furthermore, if chick eyes wear diffusers or lenses that affect only part of their visual field, only the specific parts of the eye exposed to the defocus or deprivation become myopic or hyperopic. This finding means that local regions of retina must modulate the growth of the immediately underlying sclera, whether or not the brain is involved. Furthermore, the fact that the eyes can compensate for either positive or negative lenses worn over part of the visual field suggests that the visual system decodes the sign of the blur locally, rather than relying on differences in blur across the retina, although this suggestion has not been tested explicitly.

Surprisingly, the highly foveate retina of primates also shows evidence of control by the peripheral retina. Monkeys with foveal ablations emmetropize normally, and eyes with visual deprivation of the peripheral retina become myopic, as assessed at the fovea. Thus, whereas the peripheral retina is important in both chicks and monkeys, the effects of deprivation are less localized in monkeys, perhaps because their sclera is less rigid as it lacks the layer of cartilage present in nonmammalian vertebrates. The implications of peripheral defocus for the development of human myopia are discussed below.

Temporal Factors: Blur Does Not Add Linearly

Once it is accepted that the growth of the eye is not controlled simply by the fovea, the fact must be dealt with that even if ocular accommodation (focusing) kept the image at the fovea perfectly sharp (which it does not do), all other points on the retina would experience a rapid fluctuation in the degree and sign of blur as the eye looked from a near object (retinal regions seeing the more distant parts of the scene would experience myopic defocus) to a far one (retinal regions seeing the closer parts of the scene would experience hyperopic defocus).

How do these episodes of myopic and hyperopic defocus add together? It would be convenient if doubling the duration of hyperopic defocus doubled the compensatory myopia. However, experiments in chicks show this not to be the case. When defocus is presented in episodes of different length (darkness in between), one finds (1) that frequent brief episodes have more effect than do less frequent longer ones if the total daily amount of defocus is the same, but (2) that episodes lasting less than 2 min have no effect, no matter how often they are repeated. If myopic and hyperopic defocus (e.g., positive and negative lenses) are alternated, the eyes become hyperopic in compensation for the myopic defocus, even if the hyperopic defocus is five times as long as the myopic defocus. However, if the alternation is extremely rapid (more than once per second), the episodes of hyperopic and myopic defocus cancel. Although the experiments just described have been done only in chicks, a related manipulation was done in several mammalian species: If diffusers or negative lenses are removed for an hour each day, their effect is reduced in half (during the clear vision, the eyes would have experienced myopic defocus, as in the lens-switching experiments). One can view these results as arising from internal signals related to defocus that build up over minutes and then persist in the absence of defocus for much longer times (hours). One implication of these results is that perhaps in children, the development of myopia as a result of long periods of reading has as much to do with the pattern of interruptions from reading as it does with the total amount of reading.

Molecular Signals Correlated with Myopia

An implication of eye growth being under local control is that the retina must directly or indirectly influence the tissues subjacent to it. One would like to know what signals the retina sends to increase or decrease eye growth and what the recipient tissue is. (It seems more probable that there is a signal cascade than that the retina directly controls the sclera.) A number of molecular signals have been found that either change in level when animals develop form deprivation myopia or, when agonists or antagonists are injected, affect the degree of myopia. These include dopamine, acetylcholine, vasoactive intestinal peptide, fibroblast growth factor, transforming growth factor, melatonin, and sonic hedgehog. One must view these findings with caution because depriving an eye of vision affects many processes, most having little to do with control of eye growth. Of greater relevance, administration of some agonists or antagonists of these signals can also reduce the severity of form deprivation myopia without having much effect on normal eyes. However, it is unclear what the relation of form deprivation myopia is to normal control of eye growth.

The most intriguing and puzzling molecular signal is acetylcholine, acting at muscarinic receptors. For many years, it has been known that giving animals or children atropine or pirenzepine, which blocks these receptors, greatly reduces myopia. Originally, atropine was thought to act by blocking accommodation; once this possibility was eliminated, atropine was

then thought to act via the retina. This too has been largely ruled out, in part because even eyes in which the cholinergic neurons have been destroyed pharmacologically become myopic to form deprivation, and have the myopia blocked by atropine, and in part because the pattern of responses to various cholinergic drugs is difficult to reconcile with a single mode of action. Because most of the tissues of the eye, including sclera, have most or all the six muscarinic receptors, the relevant tissue has not been identified. This issue has some clinical importance because one could argue that one should not give healthy children drugs with unknown actions.

Dopamine has also been a persistent candidate for a signal controlling eye growth, both because its levels change with form deprivation or negative lenses and because dopamine agonists ameliorate myopia, but how and where it acts are obscure. The fact that dopamine varies strongly with patterns of illumination and is a signal that synchronizes circadian rhythms makes it risky to infer a role in eye growth from effects of deprivation. On the other hand, it hints at the possibility that circadian rhythms may be involved in eye growth.

Only two retinal signals have been identified which change in opposite directions when negative lenses (which stimulate ocular elongation and myopia) and positive lenses (which do the opposite) are worn: (1) Retinoic acid is synthesized by retina and choroid in both birds and mammals, and the amount of synthesis changes in opposite directions for negative and positive lenses; retinoic acid also directly influences scleral growth. (2) Glucagon is implicated by the fact that in chicks, the expression of the immediate early gene ZENK (also called zif 268, egr-1, NGF1A, and Krox24) in glucagonergic amacrine cells is turned up by positive lens wear and turned down by negative lens wear, and the general pattern of effects of glucagon agonists and antagonists is consistent with a role in eye growth. In mammals, however, glucagonergic amacrine cells seem not to exist, casting doubt on the generality of this eye growth control pathway.

Choroidal and Scleral Components

Because myopia generally results from an excessively elongated eye, and because the sclera defines the size of the eye, it is perhaps not surprising that some theories of myopia have held that the fault lay in a defective sclera. Indeed, when the mammalian eye elongates, the sclera thins, as though it were a balloon under pressure. A more modern view is that scleral remodeling underlies ocular elongation, a view upheld by the changes in synthesis of proteins, proteoglycans, and DNA, as well as of metalloproteinases, inhibitors of metalloproteinases, and of growth factors that accompany experimental myopia or hyperopia. Complicating our comparison of experimental results from birds and mammals is the fact, mentioned above, that unlike other vertebrates, most mammals have lost the layer of cartilage in the sclera. Thus in myopic eyes, proteoglycan synthesis increases in birds but decreases in mammals. However, if the two layers of the avian sclera are separated and analyzed separately, the fibrous layer decreases its synthesis as does that of mammals, whereas the cartilaginous layer increases its synthesis.

Coincident with the changes in synthesis of the sclera are changes in the mechanical properties of the sclera, which can be appraised by measuring the size of the eye when the intraocular pressure is increased or by stretching excised pieces of sclera. Birds and mammals differ in this assay as well, perhaps because the mammalian sclera contains contractile myofibroblasts. A theory integrating these mechanical and biochemical changes would go a long way toward linking the humoral changes resulting from visual conditions to changes in ocular length.

Interposed between the retina and the sclera lies the rather mysterious choroid, which contains a vascular layer nourishing the outer retina and which synthesizes various molecules that influence scleral growth. Beyond these roles, the choroid changes its thickness in response to defocus, becoming thicker if the image plane lies in front of the photoreceptors and becoming thinner if the image plane lies behind the photoreceptors, in both cases pushing or pulling the retina toward the image plane. Thus, this thickening and thinning is a second component of emmetropization, in addition to changes in ocular length. Furthermore, there is some evidence that the thickening by itself might inhibit ocular elongation, even if the thickening is transient and therefore easily missed by investigators. If this role is confirmed, it would mean that the small magnitude of choroidal changes in mammals (20–50 μm) compared with that in birds (100–1000 μm) might not mean that its biological role is any less important.

One can also ask which of the layers of the eye (retina, retinal pigment epithelium, choroid, sclera) is the one that determines the size of the eye and is thus the principal locus of control of eye size. Although it seems reasonable that the sclera, being both stiffest and most external, is the locus of control, there is no reason that the sclera might not adapt itself to the size of any of the other layers. Just as the expansion of the retina is thought to be controlled by the tensile strength of the retinal blood vessels, so the expansion of the eye might be controlled by the tensile strength of the choroidal blood vessels.

Human Epidemiology

Across many cultures, the prevalence of myopia is correlated to the level of education, reaching levels of 90% among medical students in Singapore (Figure 1). The stronger prevalence of myopia among the educated in Asia than elsewhere is unlikely to be principally a matter of genetics because young Chinese children in Australia have one-tenth the prevalence of myopia found in Singapore. Although there are substantial differences in lifestyle between the two countries (more preschool coaching in Singapore, more outdoor activities in Australia), statistical analysis suggests that known factors cannot account for the magnitude of the difference in the prevalence of myopia.

In several studies, children with two myopic parents are more likely to be myopic than those with no myopic parents. Because children resemble their parents in many ways beyond their genes, studies of monozygotic and dizygotic twins have been used to dissociate the genetic from the environmental. At least some of these studies found an interaction between environmental and genetic factors. In a study in Singapore, the effect on the prevalence of myopia of reading more than two books a week was greater in those with two myopic parents than in those with none (Figure 2). Recently, the prevalence of myopia has been shown to be less in children with large amounts of outdoor activity.

With respect to genetic explanations of myopia, there is good evidence that extreme myopia has a tight genetic linkage. However, two developments have made it difficult to hold that common myopia of moderate degree was entirely genetic in origin: First, the animal experiments mentioned earlier have shown that every animal can acquire myopia by visual manipulations. (Some authors argue that most of the animal experiments are done on young animals, when emmetropization is in process, and therefore doubt the relevance to the human myopia that occurs later in development.) Second, societies in which people spend most of their time outside have low prevalence of myopia, but when compulsory education (and other accompaniments of modernity) is introduced, the prevalence of myopia rises tenfold within a generation or so. Less dramatically, when rural regions become industrialized, the incidence of myopia goes up quickly. This, of course, could not be a genetic change. Therefore, one must view the 'geneticness' of myopia as being more like that of lung cancer or heart disease, in which environmental factors interact with genetic predispositions, than like that of blue eyes, which are simply genetic in the Mendelian sense.

It is unclear why intensive reading should predispose the eye to myopia or why outdoor activities should have a protective effect. The simplest explanation is that the eye is simply emmetropizing normally, but to a nearby world, which requires the eye to become myopic to have clear vision without constant accommodation. This explanation is likely to be too simple, in part because it fails to take into account the spatial and temporal aspects of the control of eye growth discussed earlier with respect to the animal experiments. Related to this, much work has been based on the notion that any blurred vision promotes myopia, and thus the blur caused by imperfect accommodation (the accommodation system is a negative feedback system which does not correct all the defocus presented to it) could cause myopia. Even the brief myopic blur that results from accommodation's not relaxing immediately after a long period of reading has been considered a cause of myopic progression.

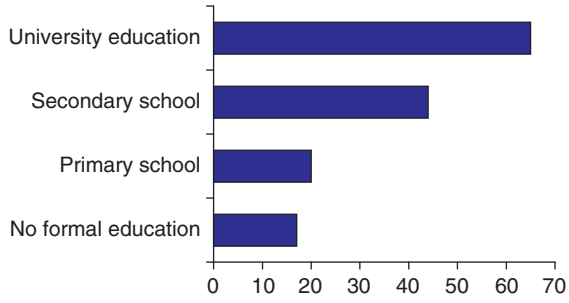

Figure 1 The prevalence of myopia among a large cohort of military recruits was strongly correlated with the level of education they had attained. Adapted from Morgan I and Rose K (2005) How genetic is school myopia? *Progress in Retinal and Eye Research* 24(1): 1–38; data from Eong KG, Tay TH, and Lim MK (1993) Education and myopia in 110 236 young Singaporean males. *Singapore Medical Journal* 34: 489–492.

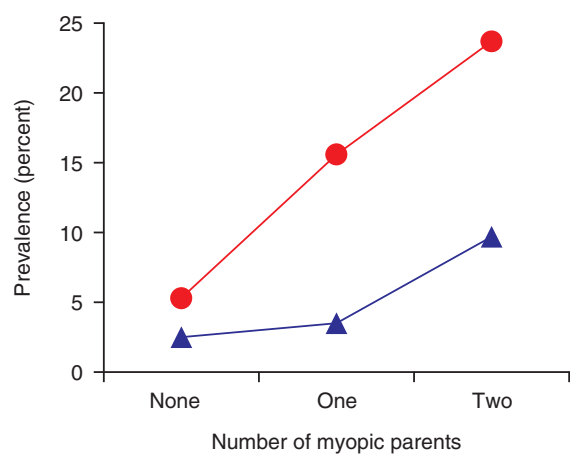

Figure 2 Evidence of the interaction of genetics and environment in determining the prevalence of myopia (< –3 D) in children reading more (red circles) or less (blue triangles) than 2 books per week. Data from Saw SM, Hong CY, Chia KS, Stone RA, and Tan D (2001) Nearwork and myopia in young children. *Lancet* 357(9253): 390.

Spatial Factors: Influence of Peripheral Refractive Errors

Much evidence has accumulated during the past 35 years that the pattern of refractive error across the eye – resulting from differences in eye shape – is correlated with whether the eye is myopic or hyperopic at the fovea (the only location typically measured by clinicians). Myopic eyes tend to be relatively hyperopic in the periphery, whereas hyperopic eyes tend to be relatively myopic in the periphery. Furthermore, a study of emmetropic students at a school for pilots in Holland showed that those with the peripheral pattern of refractions typical of myopia tended to become myopic much more often than those who had the pattern typical of hyperopia (Figure 3).

One can envision three explanations for this association between eye shape and myopia. First, one could imagine that children who read a lot develop elongated myopic eyes and that this elongation causes the sides of the eye to be drawn in closer to the optic axis, making the periphery hyperopic. The tension exerted by the zonule fibers that support the lens could exacerbate this tendency to pull the sides of the eye in. Once the periphery is hyperopic, the homeostatic growth control will be faced with opposite demands from the periphery (grow more) and from the center (grow less). The balance between these conflicting signals would probably be tipped in favor of further elongation of the eye because there is much more periphery than there is fovea (60 times the area from 30° to 31° eccentricity than between 0° and 1° eccentricity). Correcting the central myopia with negative spectacle lenses would increase the hyperopia of the periphery. Second, one can imagine that in a child who reads a lot, the periphery is subjected to myopic defocus by the world more distant than the book. This could cause the periphery to slow its growth and become hyperopic in compensation for the myopic defocus. Children who spend a lot of time outdoors would experience a flatter world in the sense that the defocus would be uniform across the visual field. Third, it could be that some children at birth have an eye the shape of which predisposes them to myopia or hyperopia. Longitudinal studies of peripheral refractive error in young children may clarify the influence of each of these factors.

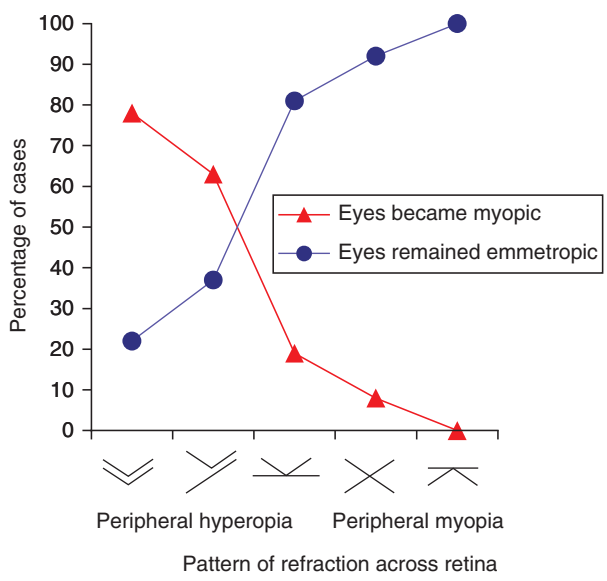

Figure 3 Association between the refractive state across the retina and subsequent development of myopia in students in a school for pilots. The diagrams under the x-axis represent the pattern of hyperopia (upward lines) and myopia (downward lines) across the retina for two orthogonal meridians. Data from Hoogerheide J, Rempt F, and Hoogenboom WP (1971) Acquired myopia in young pilots. *Ophthalmologica* 163(4): 209–215; plotted by Earl Smith, University of Houston.

Human Interventions

The logic of many visual explanations of myopia has been that all blur (at least all blur during reading) is bad, leading to myopia. Thus, many studies have attempted to reduce blur by having children wear bifocals or progressive addition lenses (like those worn by adults of a certain age who have the age-related diminution of accommodation called presbyopia). Although some positive results have been reported, the magnitude of the amelioration of myopic progression has been quite modest. This might be because children have so great an accommodative range that they have little incentive to use the lower segment of the bifocals or progressive lenses. Alternatively, it may be that under normal circumstances, the level of accommodation is determined more by the needs of the binocular fusion (vergence–accommodation) than by blur, and so reducing the accommodative demand alone may not reduce blur as much as expected. In a large multicenter trial of progressive addition lenses, the children who benefited the most were those who read more and had esophoria, less myopia, and two myopic parents. Overall, there is little support for the prophylactic value of blur reduction. More directly, a recent small study had children wear a fully correcting lens over their dominant eye and a weaker lens over the other eye, so that this eye would experience not only reduced hyperopic defocus, but myopic defocus, when accommodation focused the dominant eye; this myopic defocus slowed the myopic progression, suggesting that the sign of defocus is more important than its magnitude.

The most consistent slowing of the progression of myopia has come from treating children with atropine. Many studies over the past decades have shown that

daily treatment greatly reduces myopic progression, but the basis of this effect is not understood. Because atropine also blocks accommodation and papillary constriction, recent studies have looked to pirenzepine as a more benign prophylactic agent.

If spatial and temporal factors have the importance that is emerging from human and animal studies, it would seem that searching for a magic bullet to stop myopia might be a futile endeavor. Rather, one would need to understand how different spatial and temporal distributions of blur are integrated to determine the refractive distribution across the retina, measurements not easily made with existing technology. With respect to the spatial factors, if correcting foveal myopia exacerbates peripheral hyperopia, which in turn increases foveal myopia, then a treatment different from existing spectacles, contact lenses, or refractive surgery would be required – perhaps lenses which have opposite corrections for central and peripheral retina. With respect to the temporal factors, if children are like chicks in having long periods of hyperopic defocus canceled by briefer periods of myopic defocus, it may be that finding the correct timing of distant vision to offset the effects of long periods of reading might do more to stop eyes from becoming myopic than could any pharmacological intervention.

Implications of Animal Studies for Human Myopia

The finding that fish, birds, guinea pigs, tree shrews, marmosets, and rhesus monkeys all show compensation for imposed defocus from spectacle lenses makes it difficult to argue that human eyes do not also respond in a compensatory way to imposed defocus. Once one accepts that eye growth is under visually guided homeostatic control and that entire the retina controls eye growth, it becomes difficult to ignore that large areas of retina experiencing hyperopic defocus might participate in driving the eye toward myopia, even if the small region around the fovea is emmetropic. Similarly, once one accepts that brief episodes of myopic defocus imposed by positive lenses can compensate for long periods of hyperopic defocus, it suggests novel prophylactic measures to prevent human myopia, either by having children interrupt reading at intervals or by having myopic defocus imposed by more direct means. Thus the biology of eye growth control may well lead to an understanding of the etiology of human myopia.'

See also: Autonomic Nervous System: Ophthalmic Control; Non-Photoreceptor Photoreception; Optic Nerve, Optic Chiasm and Optic Tracts; Optic Tectum: Development and Plasticity; Retina: An Overview; Vertebrate Eyes: Evolution; Visual Development.

Further Reading

Hoogerheide J, Rempt F, and Hoogenboom WP (1971) Acquired myopia in young pilots. *Ophthalmologica* 163(4): 209–215.

McBrien NA and Gentle A (2003) Role of the sclera in the development and pathological complications of myopia. *Progress in Retinal and Eye Research* 22(3): 307–338.

Millodot M (1981) Effect of ametropia on peripheral refraction. *American Journal of Optometry and Physiological Optics* 58(9): 691–695.

Morgan I and Rose K (2005) How genetic is school myopia? *Progress in Retinal and Eye Research* 24(1): 1–38.

Mutti DO, Mitchell GL, Moeschberger ML, Jones LA, and Zadnik K (2002) Parental myopia, near work, school achievement, and children's refractive error. *Investigative Ophthalmology and Visual Science* 43(12): 3633–3640.

Phillips JR, Khalaj M, and McBrien NA (2000) Induced myopia associated with increased scleral creep in chick and tree shrew eyes. *Investigative Ophthalmology and Visual Science* 41(8): 2028–2034.

Rada JA, Shelton S, and Norton TT (2006) The sclera and myopia. *Experimental Eye Research* 82(2): 185–200.

Saw SM, Hong CY, Chia KS, Stone RA, and Tan D (2001) Nearwork and myopia in young children. *Lancet* 357(9253): 390.

Stone RA and Flitcroft DI (2004) Ocular shape and myopia. *Annals of the Academy of Medicine, Singapore* 33(1): 7–15.

Wallman J and Winawer J (2004) Homeostasis of eye growth and the question of myopia. *Neuron* 43(4): 447–468.

Wildsoet C (2003) Neural pathways subserving negative lens-induced emmetropization in chicks: Insights from selective lesions of the optic nerve and ciliary nerve. *Current Eye Research* 27(6): 371–385.

Myosin Transport and Neuronal Function

P C Bridgman and M Ahmed, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Transport Requirements of Neurons

Neurons are highly polarized cells with a cell body containing the nucleus and the majority of the synthetic machinery interposed between two distinct functional compartments, axons and dendrites (Figure 1). The vast majority of the cell cytoplasm is located in these long processes, indicating that the requirement for transport from the main site of synthesis and assembly to functional compartments in neuronal processes is greater than in any other cell type in the body. Even localized synthesis of materials that occurs far from the cell body requires the initial transport to the appropriate compartment or site.

Microtubules act as the main transport tracts in both axons and dendrites. However, the orientation of microtubules differs. Axons have microtubules oriented with their plus or fast-growing ends distal, whereas dendrites have microtubules with a roughly equal mixed polarity distribution. This means that the requirements for long-range transport are different for the two types of processes. Delivery of essential components to distal sites in axons appears to be straightforward because of the uniform microtubule orientation and the known directionality of microtubule-based motors kinesins and dyneins. However, because of the approximately equal mixed polarity in dendrites, it is not clear how net transport to distal sites occurs in these processes or how microtubule motors differentiate dendrites from axons. Multiple mechanisms resembling those used to maintain polarity in epithelial cells probably contribute to the maintenance of these two distinct main compartments.

In addition to axons and dendrites, most neurons have two much smaller functional compartments required for chemical neurotransmission: presynaptic terminals and dendritic spines (Figure 1). Microtubules usually do not fully penetrate into these regions, indicating that they cannot normally serve as efficient transport tracts. Dendritic spines are rich in actin filaments. Presynaptic terminals also have actin filaments, although the amount is less than in spines. Actin filaments have been shown to serve as transport tracks in other cell types as well as neurons and therefore may provide the main transport tracks for these actin-rich, microtubule-poor regions. Furthermore, it is possible that actin filaments and associated motor proteins may act as 'traps' in dendrites to facilitate delivery of components to distal sites similar to that seen in melanocytes, a cell type responsible for producing and delivering pigment containing organelles (melanosomes) to neighboring cells in skin via dendritic processes. The motor proteins responsible for actin-dependent transport are members of the large family of myosins that has at least 17 members, the majority of which were discovered within the past two decades.

Neuronal Myosins and Transport Function

The recent period of new myosin discovery primarily expanded the category of 'unconventional' myosins and was eye opening because it provided new possibilities for transport motors. Neurons have multiple isoforms of conventional nonmuscle myosin II as well as a number of unconventional myosins. Not surprising is the finding that these unconventional myosins appear to be involved in transport processes in neurons. It is also possible that myosin II may contribute to transport. Somewhat unexpected is the specific type of transport function that appears to involve unconventional myosins. Two types predominate: transport of plasma membrane components including transport required for receptor recycling and local delivery of proteins or translational machinery including mRNA. Both potentially play important roles during development and in synaptic plasticity. Therefore, the focus is on myosin-dependent transport in neurons that is potentially linked to pathfinding during development and synaptic plasticity in the mature nervous system. The general proposal is that myosins act primarily to mediate local transport, whereas microtubules mediate long-range transport. This general hypothesis is supported by studies and provides an appropriate basis for understanding myosin-based transport functions in neurons. It also implies that there is targeted switching between the two motor systems at specific locations in both axons and dendrites. In part, this switching will depend on the relative amounts and close proximity of the two polymers that act as the transport tracks in the different locations. However, the regulation of this switching may also involve signaling mechanisms that are cell type specific.

Properties of Neuronal Myosins Involved in Transport Functions

Although there is evidence for conventional myosin II being involved in axonal transport, most transport

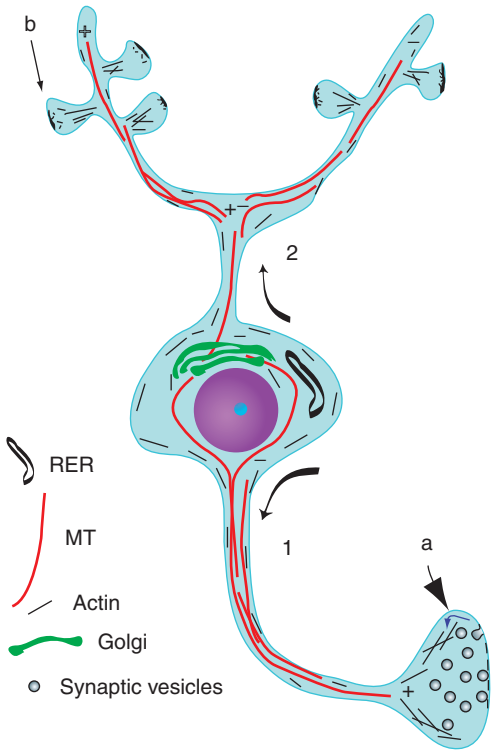

Figure 1 Neuronal compartments include the axon (1), dendrites (2), and smaller, more discreet regions such as the presynaptic terminal (a) and spines (b). Microtubule-based transport is responsible for the net movement of newly synthesized materials from the main site of synthesis in the cell body (arrows) into axons and dendrites. Actin-based transport predominates within actin-rich compartments such as presynaptic terminals and spines. The compartments are not shown to scale.

functions in neurons are associated with the myosins that are classified as unconventional. Three types of unconventional myosins predominate: myosin I, myosin V, and myosin VI. However, another unconventional myosin, myosin X, has been added to the list. The unconventional myosin I and V categories contain multiple members.

There are 30 isoforms of the single-headed myosin 1, at least eight of which are present in vertebrates. The primary neuronal forms appear to be Myo1b,-1c, and-1d. The N-terminal motor domain of myosin 1 is the most conserved region and is similar to other myosins (Figure 2). The neck region binds calmodulin as a light chain and is also suspected to be a lipid binding site in high calcium. The C-terminal tail domain has low sequence similarity between myosin 1 isoforms, allowing for a diversity of binding partners. In Myo1c, this region also has a high-affinity binding site for lipid headgroups. Similar to single-headed kinesins, myosin 1 transport function requires the binding of multiple molecules to cargo to allow movement along actin filament tracks.

Myosin V exists in three isoforms – Va, Vb, and Vc – but only the first two are highly expressed in nervous tissue. All isoforms are dimers with a long neck region that binds multiple light chains (mainly calmodulin) and a globular tail portion that contributes to cargo binding (Figure 2). The properties of myosin Va have been extensively studied. It is a processive motor, meaning that it can make a large number of steps (up to several hundred) along an actin filament before being released. This ability is a consequence of its biochemical cycle that allows at least one of its two heads to be bound to actin at all times. It moves hand over hand along the actin filament. This property makes it a good candidate for a general cargo carrying transport motor, a function that has been most directly demonstrated in melanocytes. Myosin Va binds to melanosomes through Rab27a and an adaptor protein, melanophilin/Slac2. Binding of melanophilin activates myosin Va ATPase activity, indicating that cargo binding may trigger association and movement along actin filaments. The major portion of the melanophilin-binding domain is coded by the melanocyte-specific exon F of the myosin Va gene. Exon F is not expressed in neurons, resulting in very weak binding of the neuronal form of myosin Va to melanophilin. Adaptor proteins that facilitate vesicular binding and activation of the neuronal form of myosin Va have not been identified, although interactions with several synaptic proteins have been described. Myosin Vb and Vc have a similar association with Rab proteins (11a and 8, respectively) and have functional interactions with endosomes.

Myosin VI was discovered in *Drosophila* and was immediately implicated in transport functions. It is widely expressed in nervous and sensory tissue, with particularly high expression in hair cells. Mutations in myosin VI are linked to deafness. The Snell's waltzer (sv) mouse is a myosin VI null and exhibits deafness, as well as circling, head-tossing, and hyperactivity.

Similar to myosin Va, myosin VI is processive and walks along actin in large step sizes of 30–60 nm. This is usual given that the neck region is short (Figure 2). Although it shares many structural elements with other myosins, it has been observed to walk to the minus end of actin filaments. It is one of only two myosins that show this unusual characteristic and is the only two-headed minus end-directed motor protein (although the predominant native form may be single headed and only dimerize in some circumstances). This directionality is consistent with a role in endocytosis because in general actin filaments are polarized so that their fast-growing or barbed end is directed toward or associated with the plasma membrane. Similar to myosin V, its interaction with

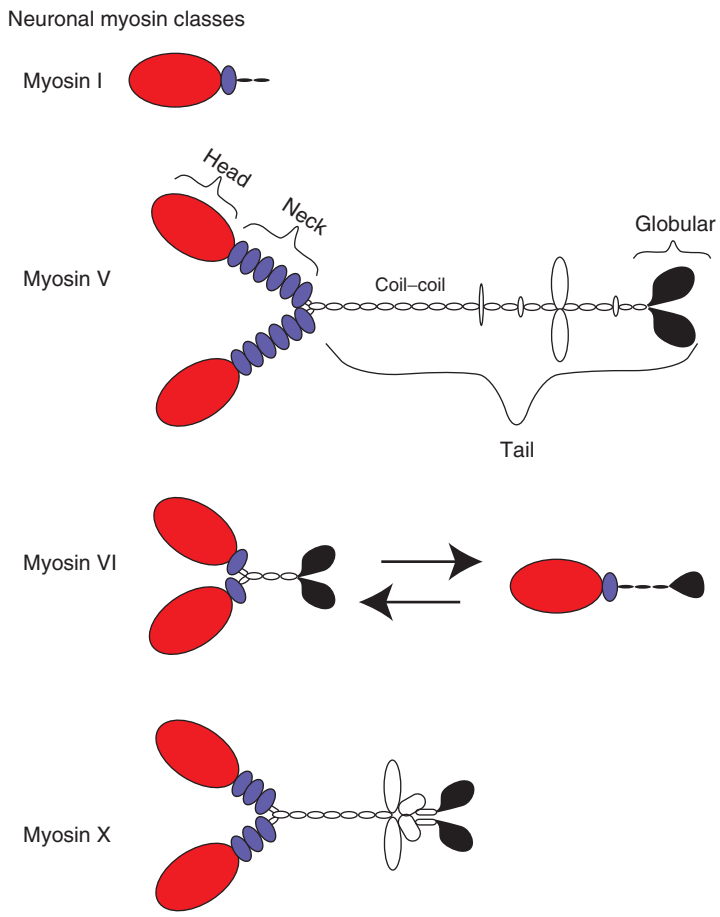

Figure 2 Diagrams comparing the different classes of myosin found in neurons. The general protein domains found in most forms of myosin are shown in the diagram for myosin V and are color coded for comparison in the other classes. The neck regions (blue) show the bound light chains. Myosin 1 exists only as a monomer. Myosin V and X are known to form dimers. Myosin VI may exist as both a monomer and, in some circumstances, a dimer. Coil–coil regions are responsible for dimer formation.

vesicular components or other cargo is likely to involve one or more adaptor proteins.

Myosin X is differentially expressed in adult nervous tissue, having detectable levels only in the cerebellum. However, during development it is more widely expressed, with a peak in the cerebrum between postnatal days 5 and 15. A headless form of myosin X is also expressed in brain with a similar expression profile in cerebrum to the full-length protein. It may act as negative regulator of myosin X motor function. The headless form is not detectable in adult cerebellum. Myosin X is also upregulated following peripheral nerve injury, indicating that it may play a role in nerve regeneration. Myosin X has head and neck domains similar to other unconventional myosins and probably forms dimers (Figure 2). The tail domain is unusual because it contains myosin homology 4 (MyTH4) and band 4.1, ezrin, radixin, moesin (FERM) domains. The MyTH4 domain has been shown to bind to microtubules, whereas the FERM domain facilitates interactions with proteins such as integrins.

Myosin Transport during Nerve Outgrowth

Growth cones are actin-rich structures that contain a variety of myosins (Figure 3), including multiple isoforms of myosin II and myosins 1, Va, and VI. Particles have been observed to move along actin filaments in growth cones, but the significance of this motility has not been fully determined. Evidence has accumulated for actin-dependent trafficking of local translational machinery for β-actin in motile cells. Local translation of β-actin contributes to growth cone turning, suggesting that this actin-based transport may be a key step in growth cone-dependent steering in response to certain guidance cues. The specific myosin(s) involved has yet to be identified, but the direction of movement (toward the periphery) is most consistent with a (+) end-directed myosin such as myosin 1, V, or II.

A functional organelle transport role for myosin 1, V, or VI in growth cones has not been firmly established. Data on the roles of myosin V and VI in the

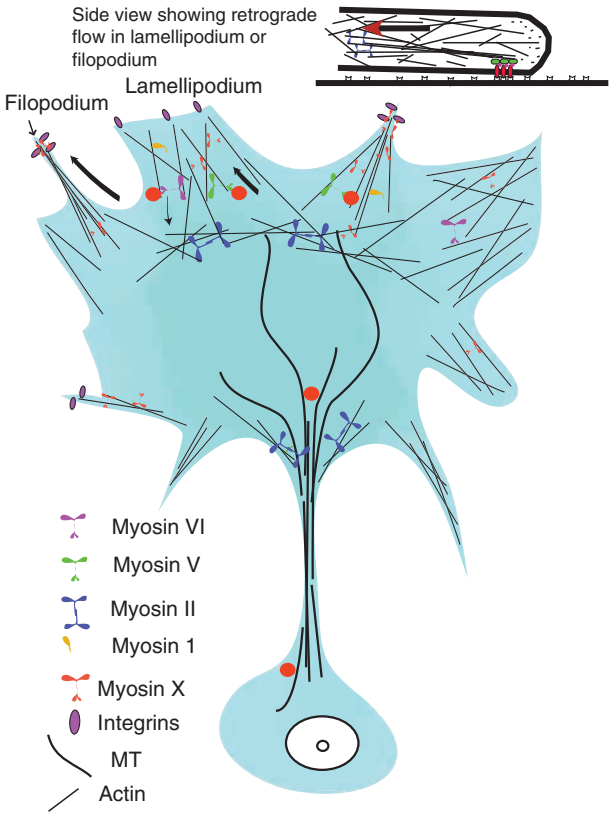

Figure 3 Actin-based transport in nerve growth cones. Growth cones contain multiple forms of myosin. Integrins are transported to the periphery along actin filaments by myosin X. Vesicles can be transported bidirectionally by myosin Va and myosin VI. Retrograde flow of actin filaments is driven by myosin II (with possible contributions by Myo1c). The side view at the top indicates the interaction between myosin II and actin that results in retrograde flow (arrow) or traction force when the flow mechanism is coupled to the substrate via adhesion complexes (multicolored structures).

regulation of neurotransmitter release may shed light on myosin VI's function in growth cones. Myosin V has been shown to be present in chromaffin cells and PC-12 cells, where it regulates release of secretory granules. In presynapatic terminals, myosin VI regulates brain-derived neurotropic factor (BDNF)-dependent neurotransmitter release characteristics. It seems plausible that the roles of these myosins in growth cones may be similar. Growth cones respond to neurotrophins such as BNDF and nerve growth factor, release neurotransmitters, and recycle synaptic vesicle membrane locally similar to presynaptic terminals.

Both Myo1c and myosin II have been shown to contribute to the retrograde flow of actin filaments (bulk rearward movement of actin networks or bundles from the leading edge without a shape change) in growth cones. Although this movement is not typically associated with transport functions, it is clearly a locally important phenomenon for motility that is at least partially driven by myosins. The mechanism of retrograde flow has been linked to the generation of traction forces necessary for the advance of growth cones.

Myosin X is a relatively recently discovered unconventional myosin that is uniquely located and appears to have a specific transport function. Myosin X localizes primarily to filopodia (long actin-rich peripheral protrusions) of migrating cells, including filopodia of growth cones. It interacts with integrins and functions in transporting integrins to the tips of filopodia. Knockdown of myosin X results in loss of filopodia, whereas overexpression promotes filopodia formation. It may also act as a molecular clutch linking integrins that have bound their ligand to the underlying actin cytoskeleton. Because of the location in filopodia and interaction with integrins, myosin X may have an important role in growth cone steering in response to guidance cues.

Myosin Transport in Mature Neurons

Myosin transport in mature neurons appears to be related primarily to synaptic plasticity. Myosins are present in both presynaptic terminals and dendritic spines. Receptor recycling in dendritic spines is the most completely described function. Myosin V and VI are both involved. In addition to recycling, transport of smooth endoplasmic reticulum (ER) into spines is myosin Va dependent. Myosin Va and VI have been found to be located presynaptically, although their functions in this location are less well documented.

Transport in Dendritic Spines

Spine actin is organized with the barbed or fast-growing end oriented primarily toward the spine head membrane. This means that myosins that walk toward the barbed end will move toward the spine head and those that walk toward the minus end will move away from the spine head and toward the stalk. This is significant for understanding transport in spines because spines contain both myosin V and VI, as well as myosin II (Figure 4).

Myosin VI has been shown to bind to a complex of proteins from brains that contain the AMPA receptor subunits: GluR1 and GluR2, as well as AP-2 and SAP97. SAP97 acts as a postsynaptic scaffold protein and AP-2 is involved in receptor-mediated internalization. Cells derived from mutant mice that lack myosin VI expression show defective internalization of AMPA receptors but do not show general defects in clathrin-mediated endocytosis. This indicates that myosin VI is critical for the specific uptake of AMPA receptors and thus may play an important role in regulating some forms of synaptic plasticity. The orientation of actin filaments in spines is consistent with

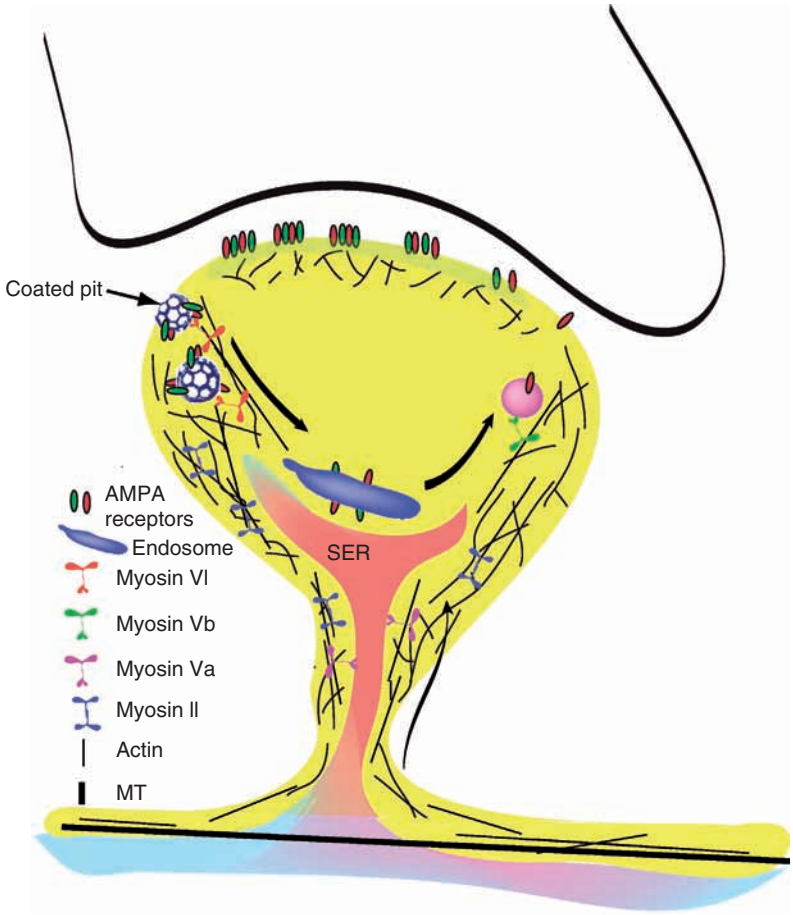

Figure 4 Actin-based transport in dendritic spines. Myosin VI is responsible for the movement of clathrin-coated endocytotic vesicles containing AMPA-type glutamate receptors away from the spine head membrane to endosomes. Myosin Vb contributes to the movement of recycled AMPA receptors back to the membrane of the spine head. Myosin Va contributes to the transport of smooth endoplasmic reticulum (SER) into spines. Myosin II regulates spine shape.

internalization and myosin VI-dependent transport away from the postsynaptic membrane to endosomes that may be within the spine or dendrite. Early, sorting and recycling endosomes are present in dendrites or spines and thus are likely to process the protein complex for degradation or recycling. For recycling, the return pathway to the postsynaptic membrane remains uncertain, but it has been shown that myosin Vb interacts with a complex of proteins that includes the GluR1 subunit of the AMPA receptor and therefore may provide the transport motor for the return of this subunit. This interaction involves Rab11, which has also been shown to mediate interactions between myosin Vb and muscarinic acetylcholine receptors. Since myosin Vb is a barbed end-directed motor and is found in dendrites and spines of hippocampal neurons (after DIV 28), it may provide this receptor subunit delivery transport function. Rab8 has been shown to be critical for AMPA receptor recycling. Myosin Vc has been shown to interact with Rab8, indicating the possibility that this isoform of myosin V may also contribute to AMPA receptor recycling. However, no data are available on myosin Vc expression in hippocampal neurons.

Myosin Va has an important role in transporting smooth ER into dendritic spines of Purkinje cells. Transporting smooth ER appears to be a general property of myosin V because it has been shown to have this function in a variety of species, including in yeast, squid axons, and *Xenopus* oocytes. Surprisingly, hippocampal neurons from dilute-lethal mice (myosin Va null) often still have spine-associated smooth ER, indicating that some spines do not rely on myosin Va for transport. Hippocampal neurons also express myosin Vb, which may possibly take over this function in the absence of Va.

Transport in Presynaptic Terminals

Both myosin Va and myosin VI have been localized to axons and presynaptic terminals. Some initial information about possible functional roles in presynaptic terminals is now available.

Myosin Va has been shown to bind to SV2-positive vesicles in neurons and the absence of myosin Va results in abnormal accumulation of these vesicles in presynaptic terminals, as well as a general increase in terminal volume. Myosin Va binds to a synaptobrevin–synaptophysin complex in a calcium-dependent manner, indicating that it may modulate some aspect of presynaptic function. The neck region of myosin Va, which contains multiple calmodulin binding sites, has been shown to interact with syntaxin-1A in chromaffin cells. This interaction appears to regulate the late sustained phase of chromaffin granule exocytosis in response to increases in intracellular calcium above 0.3 μM. Presumably, myosin Va acts to transport vesicles from the reserve to the readily releasable pool in chromaffin cells. Whether a similar function occurs in presynaptic terminals is not clear because electrophysiological experiments have not found significant changes in release properties in the absence of myosin Va.

Myosin VI has a role in regulating BDNF-mediated neurotransmission in hippocampal neurons. Myosin VI interacts with a complex of proteins that includes GIPC1 and TrkA or TrkB. GIPC1 appears to have an adaptor function and binds directly to myosin VI. The absence of myosin VI or GIPC1 impairs basal synaptic transmission from hippocampal slices and BDNF-induced long-term potentiation. The exact mechanism through which myosin VI influences synaptic transmission remains to be determined, but considering its general role in endocytosis, it seems likely that it acts mainly in the retrieval of synaptic vesicles following exocytosis.

Myosin Transport of mRNA, Translational Machinery, and Organelles in Axons and Dendrites

In mature neurons, myosin Va has been shown to associate with ribonucleoprotein complexes (that can include mStaufen, Pur α, or fragile X mental retardation protein) and mRNA that participate in local translation of proteins. This association has been observed in both axons and dendrites. The axonal form of this complex is known as periaxoplasmic ribosomal plaques. Although it is not clear if myosin Va has a direct transport role for these complexes, the association and known function of myosin V in other organisms such as yeast is consistent with a role in the transport of synthetic machinery for local synthesis. It is probable that this role will be expanded as more is learned about myosin V transport functions in actin-rich regions such as spines.

Myosin Va has also been shown to affect the amount, transport, and distribution of neurofilaments in axons. However, neurofilament transport does not require actin filaments, making it unlikely that neurofilaments are directly transported by myosin Va. Thus, what role does myosin Va play in axonal neurofilament formation and maintenance? Upregulation and local translation of neurofilament subunits has been described following nerve injury, suggesting that local production may contribute to axonal neurofilament formation. If myosin Va localizes or transports components necessary for local synthesis of proteins, including neurofilaments, then it is possible that disruption of the machinery required for this function could alter neurofilament transport and formation. Alternatively, local translation of microtubule motors is myosin Va dependent, and disruption in the production of microtubule motors interferes with neurofilament transport.

Organelle transport in axons and dendrites does not require myosin motors, but rate and directional changes may involve myosin 1 or Va activity. Transient binding to actin could cause pauses in transport and allow organelles to interact with microtubule-based motors with the opposite directional polarity. Alternatively, myosins may ride as passive passengers on organelles transported by microtubule motors and then become activated when microtubules terminate within actin-rich regions. The activation of myosins could immobilize or locally transport organelles into actin meshworks at specific sites, preventing them from being transported along microtubules in the opposite direction. This could lead to accumulation of components at distal functional sites in dendrites.

Myosin Transport and Neuronal Function

Data on myosin transport functions in neurons are consistent with a general role in the polarized delivery or retrieval of components (including synthetic machinery) from the actin-rich cytoplasmic domains that include dendritic spines and presynaptic terminals. Myosins also appear to directly participate in receptor and synaptic vesicle turnover or recycling in these compartments through transport functions closely associated with exocytosis and endocytosis. Whereas transport between dendrites or axons and spines or presynaptic terminals requires switching between microtubule and actin filament tracks, as well as motor proteins, recycling within these compartments is independent of microtubules and therefore may be solely dependent on myosin transport.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal

and Dendritic Identity and Structure: Control of; Axonal mRNA Transport and Functions; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; Growth Cones; Microtubule Associated Proteins in Neurons; Microtubules: Organization and Function in Neurons.

Further Reading

Bridgman PC (2004) Myosin-dependent transport in neurons. *Journal of Neurobiology* 58: 164–174.

Buss F, Spudich G, and Kendrick-Jones J (2004) Myosin VI: Cellular functions and motor properties. *Annual Review of Developmental Biology* 20: 649–676.

Dillon C and Goda Y (2005) The actin cytoskeleton: Integrating form and function at the synapse. *Annual Review of Neuroscience* 28: 25–55.

Hasson T (2003) Myosin VI: Two distinct roles in endocytosis. *Journal of Cell Science* 116: 3453–3461.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews* 6: 201–214.

Kennedy MJ and Ehlers MD (2006) Organelle and trafficking machinery for postsynaptic plasticity. *Annual Review of Neuroscience* 29: 325–362.

Langford GM (2002) Myosin-V, a versatile motor for short-range vesicle transport. *Traffic* 3: 859–865.

Ming GL (2006) Turning by asymmetric actin. *Nature Neuroscience* 9: 1201–1202.

Ostap ME and Pollard TD (1996) Overlapping function of myosin-1 isoforms? *Journal of Cell Biology* 133: 221–224.

Seabra MC and Coudrier E (2004) Rab GTPases and myosin motors in organelle motility. *Traffic* 5: 393–399.

Sotelo-Silveira JR, Calliari A, Kun A, Koenig E, and Sotelo JR (2006) RNA trafficking in axons. *Traffic* 7: 508–515.

Sousa AD and Cheney RE (2005) Myosin-X: A molecular motor at the cell's fingertips. *Trends in Cell Biology* 15: 533–539.

Wagner W and Hammer JA III (2003) Myosin V and the endoplasmic reticulum: The connection grows. *Journal of Cell Biology* 163: 1193–1196.

Relevant Website

http://www.mrc-lmb.cam.ac.uk – The Medical Research Council, Laboratory of Molecular Biology.

Napping

S C Medrick and S P A Drummond, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

To answer the question of what the function of napping may be, we must first define 'function.' Though this article concentrates mostly on the function of napping for behavior, it would be possible also to write about the function of napping across cultures and history. For example, there is a long history of napping in southern regions of the globe, in which it is practiced in the form of a siesta. By the first century BC, the Romans had coined a word for the afternoon break, *meridiari*, derived from the Latin word for 'midday.' Later, the church divided the day into periods designated for specific activities, such as meals, prayer, and rest. Midday became known as *sexta*, as in the sixth hour (noon, by their way of counting), a time when everyone would take rest and pray. The word has survived as the familiar term, 'siesta.' This postprandial sleep custom disappeared in northern Europe with the industrial revolution, but remained a strong tradition in southern Europe and many Latin American cultures, where people would shutter their businesses for a few hours to return home, eat a meal, sleep, and then return to work from 4 p.m. until 9 p.m. While the 'siesta' culture continues today in these regions, enthusiasm for the siesta has cooled substantially.

Presently, napping is most popular in highly industrious cultures such as Japan and Germany, whereas countries such as Spain or Italy have begun phasing out the practice. Current beliefs about midday rest are being reassessed from an empirical perspective, as evidenced by the fact that this topic is enjoying the spotlight of an entire entry in an encyclopedia of neuroscience for the first time. One reason for this resurgence of interest in the function of naps is due to the changing needs of modern culture. An increase in 24 h work cycles, the regularity of international travel and communication with global markets, and long work days and longer commutes have brought on a host of secondary problems that require attention – reports show continual decreases in nocturnal sleep, increases in sleep disorders, and increases in sleepiness-related accidents. The nap is thus being increasingly investigated as an inexpensive, noninvasive, short period of sleep strategically implemented for sleepiness and fatigue management, sustained productivity and alertness, and optimal cognitive processing.

This emerging body of research is summarized in the following sections, with emphasis on the possible biological basis of effects of napping, implementation during night shift and sleep deprivation studies, and promising new directions investigating the dose- and sleep stage-dependent benefits possible through targeted napping schedules designed to improve performance on a wide variety of behavioral measures.

Definitions and Demographics of Napping

What Is a Nap?

Naps, at least as discussed herein, are defined as intended periods of sleep that can last anywhere from 3 min to 3 h and can be taken anytime during the day or night. Meta-analysis of surveys and diary studies of American populations show that naps usually range between 0.5 and 1.6 h. Naps should be distinguished from microsleep, which is a brief, involuntary period of sleep lasting from seconds to minutes. In addition, naps should be distinguished from 'major' sleep periods that typically occur overnight and typically last at least 5 h or more (of course, in night shift workers, these major sleep periods occur during the day).

Changes in Napping Behaviors across the Life Span

Napping patterns shift throughout life. In infancy, two basic types of sleep emerge: quiet sleep and active sleep. These are the infant equivalent, respectively, of nonrapid eye movement (NREM) and rapid eye movement (REM) sleep seen in adults. Over 50% of neonate sleep is active (though premature babies can achieve levels as high as 80%), but that number will drop to 30% by the end of the first year. Infants have shorter sleep cycles than adults do, with the typical cycle lasting 50–60 min as opposed to 80–100 min in adults. Interestingly, infants also commonly show sleep-onset active sleep especially after feeding, a sleep pattern considered pathological in adults. By the age of 2 years of age, REM occupies 20% of total sleep, a figure that remains relatively common throughout the rest of life. Napping emerges a little before the first birthday, when sleep coalesces into a nocturnal sleep period, a shorter nap (30 min to 1 h) in the morning around 10 a.m., and a longer nap (1–3 h) in the afternoon. Eventually napping consolidates into one long afternoon nap and then the nap disappears in most children around age

four, reappearing in teenage years and into college years, when up to 60% of students report regular napping habits.

Recent evidence suggests that men take more naps than women do, even though nocturnal sleep hours average about the same for both sexes. On average, men's nocturnal sleep appears to be less efficient, so men may simply need a nap more than women do. Specifically, studies of young adults show that males not only spend more time in bed awake compared to females, but also enjoy less slow-wave sleep (SWS) and REM compared to females. Troubled sleep occurs in women as well, of course, but typically begins later in life, resulting in part from hormone fluctuations that occur across a woman's lifetime.

Napping behavior also changes as a function of aging. Older adults nap more frequently and later in the day compared with younger adults. There may be multiple underlying causes for these differences, including nocturnal sleep deterioration, weaker circadian rhythm and circadian phase advance, or simply that the elderly have more time to practice elective napping. Some studies have shown that napping in older adults is related to decreases in slow-wave activity and reduced sleep efficiency, whereas napping is not related to nocturnal sleep parameters in normal adults or insomniacs.

A number of studies have investigated how napping interferes with nocturnal sleep parameters; effects are likely due to the relieving of sleep pressure, the biological drive to go to sleep that increases as a function of time awake. In general, naps are reported to not interfere with nocturnal sleep. Exceptions to this are that (1) naps can decrease slow-wave activity in subsequent sleep episodes when naps are taken within a 2–3 h window from nocturnal sleep (bedtime) and that (2) when naps are taken during a night shift, the subsequent daytime sleep can show decreased slow-wave activity.

Is Napping a Natural Part of Our Circadian Rhythm?

Timing of Naps

When napping is examined in the laboratory, the consistent finding is that daytime sleepiness is a regularly occurring phenomenon. The afternoon 'nap zone,' first proposed by Broughton, is a period between 14:00 and 16:00, when daytime sleep propensity is highest. Such a propensity for diurnal sleep has been demonstrated in a variety of different experimental milieus, including with removal of all temporal constraints or 'free-running' conditions during ultrashort routines, in which sleep–wake schedules occur over a 90 min period, and as evidenced in the classic 'M-shaped' time-of-day function in studies of sleep propensity using the Multiple Sleep Latency Test (MSLT), a test in which the time it takes to fall asleep, sleep onset latency, is measured at regular intervals across the day. Even in studies when individuals are specifically asked not to nap, resistance to daytime sleep has been most weak during these afternoon hours. Due to the increased sleep propensity because of the time of day, as well as the historical develop of napping behaviors, it has been (perhaps misleadingly) termed 'the postprandial dip.' Studies have shown, however, that the energy slump occurs even in the absence of lunch and/or without knowledge of the time of day.

Physiological Evidence for 'Nap Zone'

A fluctuation in core body temperature (CBT), a fundamental measure of circadian rhythms, represents the best physiological marker to correlate with increased afternoon sleep propensity. Generally speaking, there is a rise in temperature across the daytime and a decrease during the night. The falling temperature traditionally has been considered one important trigger for onset and method of sustainment for nocturnal sleep. CBT starts to fall prior to habitual bedtime and reaches its lowest point approximately two hours prior to habitual wake time (typically between 03:00 and 05:00 in most adults). Though circadian fluctuation of CBT can be fit with a simple sinusoid function spread across the 24 h period, further investigations have found that CBT is better described by adding a 12 h bicircadian component to the model, which corresponds to a robust finding of a dip in temperature in the afternoon (Figure 1).

The afternoon dip in temperature corresponds to the time when individuals show a greater sleep propensity. Although decreases in CBT are temporally correlated with increases in sleepiness, a direct mechanistic link has yet to be discovered. Sedative-hypnotics such as melatonin and benzodiazpines decrease CBT and increase peripheral heat loss, which has been directly related to sleep onset latency. In contrast, agents such as caffeine, amphetamines, nicotine, and cocaine decrease sleep propensity and increase CBT. Further, studies have shown that the best predictor of sleep onset (better than melatonin) was the distal–proximal skin gradient, an index of peripheral heat loss. Thus, it is likely that a decrease in temperature (CBT and peripheral heat) is a trigger for sleep in general, and possibly also for afternoon naps. The duration of a sleep episode may also be related to the direction of change in CBT – that is, long nocturnal sleep occurs during an extended period of decreased temperature while short sleep occurs

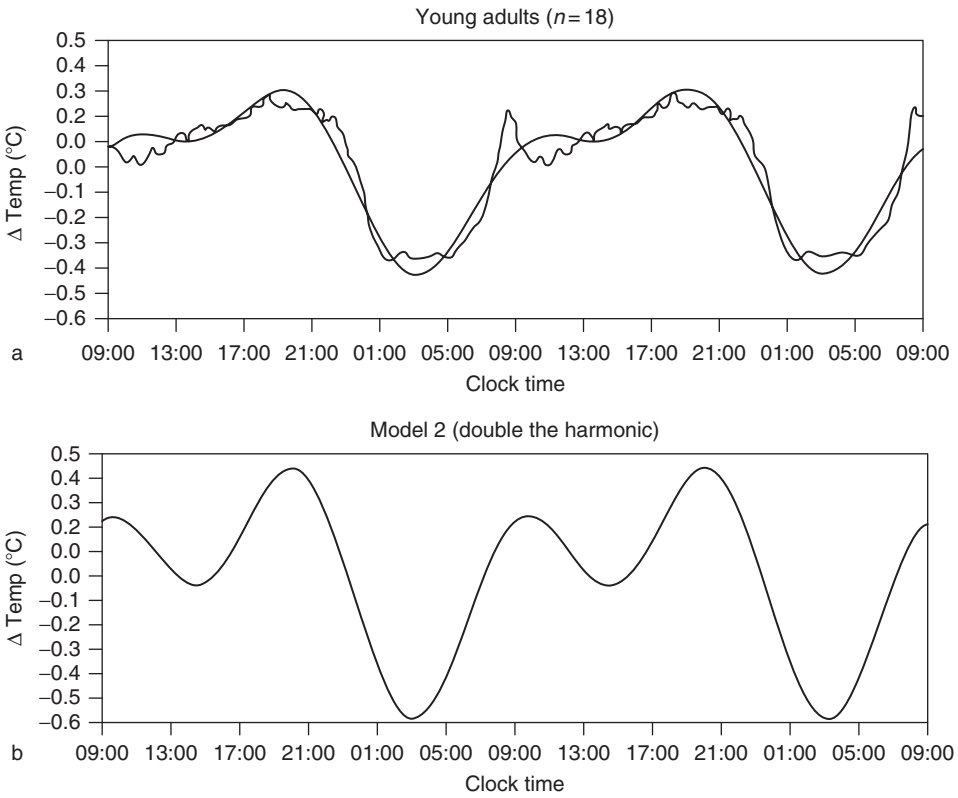

Figure 1 (a) Double-plotted 24 h rectal temperature rhythms from 18 young adults on a nycthemeral routine, together with best fitting composite 24 and 12 h sinusoids (see text). (b) The same fitted model, but with the 12 h amplitude raised from 0.152 to 0.304 °C. Copyright (1996) from 'Circadian determinants of the postlunch dip in performance' by Monk TH, Buysse DJ, Reynolds CF 3rd, and Kupfer DJ. Reproduced by permission of Taylor & Francis Group, LLC.

when CBT increases. Studies have found that increases in CBT are related to more frequent awakenings. Further research is needed to disentangle the causal relationship between changes in CBT and daytime sleep onset.

Behavioral Markers for the 'Nap' Zone

Anecdotal evidence for behavioral measures of the 'nap' zone can be easily found in any typical workplace, with decreased productivity and increased caffeine consumption during the afternoon, as well as an increase in traffic accidents during this time of day (even after taking into account the increased number of cars on the road). On the other hand, it has been more difficult to find consistent laboratory evidence for behavioral markers. Two factors are most likely the cause. First, there may be specific cognitive processes that are vulnerable to circadian peaks and troughs, while others are not. For example, tasks that rely on continuous performance show dips in the afternoon, such as card sorting, serial search, a variety of signal detection tasks, and physical exertion such as sprinting. Memory tasks and perceptual tasks, on the other hand, have not shown a strong circadian component. Second, individual differences are likely to produce a large source of variation. Intriguing findings of individual differences in vulnerability to a midafternoon performance deficit on a monotonous visual vigilance task demonstrate that only half of the individuals show the 14:00 performance dip. Monk and colleagues compared the CBT of the 'dippers' to the 'nondippers' and showed that performance decreases coincided with a flattening of the CBT in the dippers, whereas CBT in the nondippers continued to increase in a linear manner during this period. Monk proposed that an individual's propensity for midday decreases in performance may be predicted by a combination of the individual's endogenous circadian pacemaker and the length of time the individual has been awake (i.e., magnitude of sleep pressure). Specifically, Monk stated that "the size (or timing) of the 12 h temperature rhythm component might be predictive of the size (or presence) of a post-lunch dip in performance." Further research investigating other biological and genetic determinants of midday dip vulnerability, such as morningness and eveningness measures proposed by Horne, will be extremely interesting pieces of information for answering these questions.

What Is the Function of Naps?

Now that we have provided evidence in favor of naps being a natural part of sleep–wake cycles, at least in some individuals, we turn our attention to the possible function of naps. As already stated, the concept of 'function' can take many forms. For example, one can ask whether the function of a nap is the same as the function of sleep in general, or if it serves a purpose separate from that of the major sleep period. While we do not yet fully understand why we sleep, we do know that a variety of molecular, genetic, and physiological processes occur exclusively, or primarily, during sleep. For the most part, research has not examined whether these same changes occur during naps. What has been examined, though, are the behavioral and cognitive benefits of a nap under a variety of conditions. In other words, research has examined how naps help individuals function better. Thus, it is this more operational definition of function that we discuss here.

Sleepiness versus Fatigue Countermeasures

In the broadest sense, naps are used operationally as either sleepiness countermeasures or fatigue countermeasures. The distinction between sleepiness and fatigue is one that is, unfortunately, often blurred or confused, but is an important distinction to keep in mind. Sleepiness is the physiological propensity to fall asleep, either intentionally or unintentionally. As discussed earlier, the circadian rhythm of the CBT is one traditional biomarker of sleepiness. When naps are used as a sleepiness countermeasure, the intended function is to increase arousal and alertness. This increased alertness level may, in turn, produce better performance. In contrast, fatigue refers to a decrease in physical or cognitive efficiency related to time-on-task and workload, independent of whether someone has a propensity to fall asleep. When naps are used as a fatigue countermeasure, the intended function is to directly boost performance. Most often, the term 'fatigue countermeasure' is incorrectly used to cover both functions (of not only naps, but also interventions such as caffeine) when used in operational settings.

Napping in Sleep-Deprived Conditions

Until recently, napping research has focused primarily on treating sleepiness during extended work periods, such as long-haul truck driving, transatlantic airplane routes, and NASA space flight, as well as during nontraditional work schedules such as night shifts. In both of these circumstances sleepiness due to sleep deprivation is a common danger and has been implicated as contributing to accidents during work and transit from work to home, as well as to increased health problems in these workers.

Night shift work is particularly vulnerable to extreme performance decrements due to unintentional sleep, increased sleepiness, and decreased performance for most skills, including vigilance, reaction time, serial addition/subtraction, spatial orientation, and flight simulator operation. Critical hours for increased errors and slowness are between 03:00 and 05:00 (coinciding with CBT nadir). The two main sources of reduced alertness and performance during night work are (1) the circadian rhythm of sleepiness and alertness (as discussed earlier) and (2) increased homeostatic sleep pressure. Contributing to this difficulty with night shift work is the poor adaptability of circadian regulated processes such as endocrine, sleep, CBT, adrenaline, alertness, and other physiological rhythms to even long-term (2–3 months) reversal of sleep–wake cycles. Even workers on permanent night shift schedules do not show a change in the timing of the circadian system and continue to show reduced performance and increased mistakes after years on the night shift. Thus, the need to find sleepiness and fatigue management solutions for this population is an imperative.

Work-related napping strategies have categorized three types of naps: prophylactic napping (taken in anticipation of sleep deprivation), compensatory napping (taken after sleep deprivation has begun), and operational napping (napping during working hours). Overall, prophylactic napping taken just before the work night seems to best enhance performance overnight, although a combination of prophylactic napping and caffeine may work even better. With respect to operational napping, it has been found that both pilots and truck drivers unofficially nap during night shifts and long-haul transportation trips. Research has shown that a 20 min operational nap between 01:00 and 03:00 significantly improves speed of response on a vigilance task measured at the end of the shift, compared with a control condition. The potential problem with napping during the night is that there is increased risk for waking with sleep inertia. Sleep inertia is the feeling of slowness, irritability, and poor decision-making ability that can occur during the first 20–30 min after waking from deep sleep, although this period may be shorter with naps than with longer sleep periods. Methods for combating sleep inertia, however, have been proved successful, including exposure to bright light, washing the face, exercise, and, of course, a dose of caffeine. Apart from performance on specific measures, napping has been shown generally to improve alertness, productivity, and mood, and this may be especially so under sleep-deprived conditions,

during night shift work, and during prolonged periods of driving.

Cognitive Benefits of Napping Linked to Specific Stages of Sleep

It is perhaps unsurprising that napping has been shown to be an effective sleepiness countermeasure, and that improved alertness can, in turn, have a positive impact on performance. What may be even more interesting to explore is whether napping can be a fatigue countermeasure independent of alerting effects, and whether naps can even enhance normal performance when fatigue and/or sleepiness are not an issue. Recent studies investigating the impact of napping on a variety of memory consolidation measurements have in fact provided evidence for both of these functions. Studies on the benefit of napping on performance have demonstrated that short daytime sleep episodes not only can decrease sleepiness but also show selective enhancement of various forms of memory, as well as increase alertness and physical and mental stamina. Furthermore, these studies have related these napping-related improvements to specific sleep stages.

An important methodological aspect of napping that has been exploited by these studies is that sleep during naps can be titrated to have specific stages of sleep without disturbing the napper's sleep. This is accomplished primarily by manipulating the duration and timing of the nap. Such manipulations have been attempted during nocturnal sleep studies by depriving individuals of sleep during the first or the second half of the night, in order to isolate SWS or REM sleep. Although selective cognitive benefits have been shown, this method does not actually isolate stage two, SWS, or REM. Naps, on the other hand, have shown a high degree of specificity in performance between stage one alone versus stage one and two combined, and naps with SWS alone versus SWS and REM combined.

Hayashi et al. demonstrated the recuperative power of napping for sleepiness and alertness measurements even with a 5-min nap limited to stage one sleep. They report that performance improvement on a visual detection task and digit symbol task, as well as decreases in slow eye movements during testing, required a nap with stage two. Walker demonstrated that a nap can improve performance on a motor learning task to the same degree as can a full night of sleep, with stage two sleep playing an important role.

Naps rich in SWS have been shown to improve declarative memory for pictures or word pairs after a nap, as well as prevent deterioration in performance that develops across the day. Mednick and colleagues have reported a series of studies establishing the efficacy of naps in combating performance deterioration. These studies utilized a visual perceptual task in which individuals reliably show significant decreases in performance with repeated testing across the day, even when the test is only given twice. Importantly, they found that a 60 min midday nap rich in SWS can reverse perceptual deterioration and restore performance to baseline, with long-lasting benefits to performance (Figure 2).

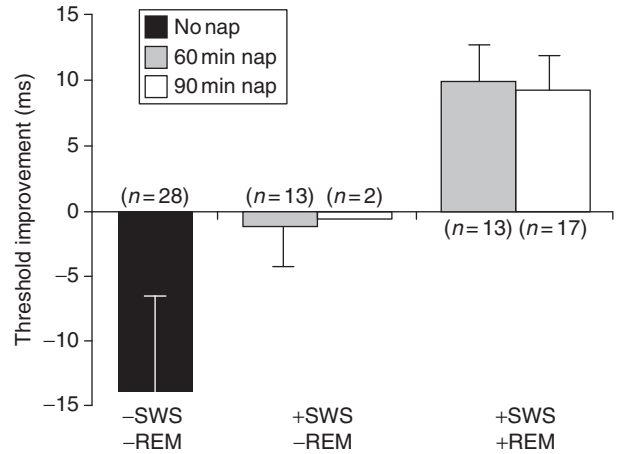

Figure 2 Same-day improvement in no nap, 60 min nap, and 90 min nap groups, with and without rapid eye movement (REM) and slow-wave sleep (SWS). (Left) No-nap group shows deterioration at 19:00 from baseline test at 09:00. (Center) Performance after naps with SWS but without REM shows neither deterioration nor improvement. (Right) Naps with SWS and REM led to significant improvement. Only two individuals in the 90 min nap group showed no REM. Reprinted by permission from Macmillan Publishers Ltd: *Nature Neuroscience* (Mednick S, Nakayama K, and Stickgold R (2003) Sleep-dependent learning: A nap is as good as a night. *Nature Neuroscience* 6(7): 697 –698.), copyright (2003).

Naps including both SWS and REM actually led to an improvement in perceptual performance equivalent to that following a full night of sleep. Furthermore, when individuals are tested after a nap and a full night of sleep, they demonstrate as much benefit as with two nights of sleep, indicating that sleep-dependent learning is similarly effective whether it is from daytime naps or nocturnal sleep. Also, the benefits from napping and nocturnal sleep are additive (Figure 3). It should be emphasized that these sets of studies examined performance after a normal night of sleep, rather than following a period of sleep deprivation. Thus they showed that naps can enhance performance beyond even 'normal' levels.

Summary

In examining the function of naps it appears that modern culture has redefined the functionality to suit the needs of an increasingly 24 h society. The

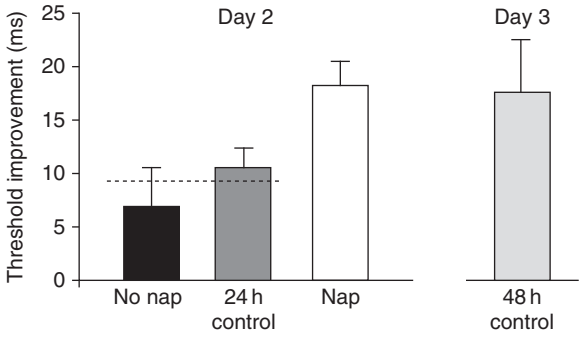

Figure 3 Improvement for nap and no-nap groups. (Left) Improvements 24 h after training for the no-nap group's second retest, the 24 h control group's first retest, and the 90 min nap group's second retest, all at 09:00 on day 2. Dashed line shows nap group's improvement on day 1. (Right) Improvement 48 h posttraining with no nap. Reprinted by permission from Macmillan Publishers Ltd: *Nature Neuroscience* (Mednick SC, Nakayama K, and Stickgold R (2003) Sleep-dependent learning: A nap is as good as a night. *Nature Neuroscience* 6(7): 697–698.), Copyright (2003).

amount of nocturnal sleep continues to decrease as labor demands increase in duration, as well as with the increase in around-the-clock work schedules. Thus the nap serves as both a fatigue and sleepiness countermeasure. Some studies demonstrate that napping in the afternoon may in fact be an inherent part of our natural sleep–wake cycle, as evidenced by physiological and performance decreases during the afternoon that temporally coincide with increased propensity to sleep. Most research on napping has examined either alertness or specific cognitive benefits. An important consideration that has emerged from this body of research is that specific stages of sleep can confer specific benefits to performance. Overall, naps seem to confer a number of possible benefits, but this area of investigation is still young. Present directions of research are investigating the cognitive benefits of napping and the possibility of fitting a nap to an individual's needs by adjusting the duration and time of day of the nap. Important areas for future research concerning the function of naps will include the medical, physiological, and psychological benefits of napping that have been reported to occur with nocturnal sleep.

See also: Autonomic Dysregulation During REM Sleep; Circadian Rhythms: Influence of Light in Humans; Circadian Rhythms in Sleepiness, Alertness, and Performance; Immune Function During Sleep and Sleep Deprivation; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Research and Sleep Medicine in Historical Perspective; Sleep in Adolescents; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function; Sleep and Circadian Rhythm Disorders in Human Aging and Dementia; Sleep in Aging; Sleep: Development and Circadian Control; Sleep-Dependent Memory Processing; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Carskadon MA and Dement WC (1986) Effects of a daytime nap on sleepiness during sleep restriction. *Sleep Research* 15: 69.

Dinges DF and Broughton RJ (eds.) (1989) *Sleep and Alertness: Chronobiological, Behavioral and Medical Aspects of Napping*, pp. 171–204. New York: Raven Press.

Hayashi M, Motoyoshi N, and Hori T (2005) Recuperative power of a short daytime nap with or without stage 2 sleep. *Sleep* 28(7): 829–836.

Mednick SC, Nakayama K, Cantero JL, et al. (2002) The restorative benefit of naps on perceptual deterioration. *Nature Neuroscience* 5(7): 677–681.

Mednick SC, Nakayama K, and Stickgold R (2003) Sleep-dependent learning: A nap is as good as a night. *Nature Neuroscience* 6(7): 697–698.

Monk TH, Buysse DJ, Carrier J, et al. (2001) Effects of afternoon "siesta" naps on sleep, alertness, performance, and circadian rhythms in the elderly. *Sleep* 24(6): 680–687.

Monk TH, Buysse DJ, Reynolds CF 3rd, et al. (1996) Circadian determinants of the postlunch dip in performance. *Chronobiology International* 13(2): 123–133.

Rosekind MR, Gander PH, and Dinges DF (1991) Alertness management in flight operations: Strategic napping. SAE Technical Paper Series 912138.

Stampi C (ed.) (1992) *Why We Nap: Evolution, Chronobiology, and Functions of Polyphasic and Ultrashort Sleep*. New York: Springer.

Walker MP and Stickgold R (2005) It's practice, with sleep, that makes perfect: Implications of sleep-dependent learning and plasticity for skill performance. *Clinics in Sports Medicine* 24(2): 301–317, ix.

Relevant Website

http://www.nationalsleepfoundation.org – List of Organizations Involved in Sleep Research.

Narcolepsy

E Mignot and L Lin, Stanford University, Palo Alto, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

What Is Narcolepsy?

The sleep disorder narcolepsy was first recognized, named, and described in the late nineteenth century when an unusual association of excessive daytime sleepiness and episodic muscle weakness that was believed to be more than simply 'epileptoid' phenomenon was noted. The unusual muscle weakness episodes triggered by emotions that occur in this condition were called 'cataplexy' soon thereafter. Until relatively recently, whether or not narcolepsy was a symptom or a disease entity was much debated. The discovery of a human leukocyte antigen (HLA) association in 1980 added credence to the hypothesis of a unique disease entity with a discrete etiology. This was subsequently confirmed by the discovery of hypocretin (orexin) deficiency as the major cause of narcolepsy–cataplexy.

In its typical form, symptoms include sleepiness, cataplexy, sleep paralysis, hypnagogic hallucinations, and disturbed nocturnal sleep. Sleepiness, abnormal sleep onset into rapid eye movement (REM) sleep, and cataplexy are the most important symptoms. Whereas much is known about the pathophysiology of narcolepsy with cataplexy, little is known regarding the cause of narcolepsy without cataplexy. For this reason, international classifications and research studies typically separate narcolepsy into cases with and without cataplexy. As discussed later, narcolepsy without cataplexy is more likely to be etiologically heterogeneous.

Approximately 1 person in 2000 in the United States and Western European countries has narcolepsy with cataplexy. The exact prevalence of narcolepsy without cataplexy (defined as isolated unexplained daytime sleepiness with abnormal REM sleep onset periods) is unknown but likely to be higher. Narcolepsy is typically diagnosed many years after disease onset. Early detection and treatment is critical to ensure normal schooling and to prevent associated loss of social status. The disorder is socially isolating, and many patients have problems holding a job and maintaining an active social and affective life.

Narcolepsy Symptoms

Excessive daytime sleepiness is the most troublesome symptom and most often causes patients to consult. Typical complaints are frequent and overwhelming sleep attacks and the need to nap frequently during the day. A wide range of severity may be seen, from simply needing a nap everyday or nodding when at rest to an almost constant feeling of irresistible drowsiness. Sleepiness in narcolepsy may be confused with fatigue and depression.

Cataplexy, a brief and sudden loss of muscle strength triggered by emotions, is very important clinically for the diagnosis of narcolepsy. It is the only symptom specific of narcolepsy, but unfortunately it is not present in all patients. Typically, patients feel weak in the knees and have to sit down when they are emotionally excited – for example, in the context of a funny joke, a good laugh, or when becoming angry. Other attacks of muscle weakness may affect the head (head dropping), the jaw (jaw dropping), the face, or the arms. Cataplexy is frequently mild. It may occur only a few times per month or several times a day. Strong attacks may escalate to a complete body paralysis episode lasting up to a few minutes. Cataplexy may be confused with epilepsy or other neurological/psychiatric problems. Importantly, patients are awake and conscious during cataplexy. Cataplexy should also not be confused with catalepsy, a symptom characterized by increased muscle tone and body rigidity, most often in the context of schizophrenia.

The other symptoms (sleep paralysis, hypnagogic hallucinations, and insomnia) are only weakly indicative of narcolepsy. They are also observed in normal individuals or in association with other sleep disorders. Sleep paralysis is characterized by an inability to move when waking up or when falling asleep. The first episode of sleep paralysis is often frightening, but the paralysis always ends after a few seconds to a few minutes. Hypnagogic hallucinations ('hypnagogic' means 'when falling asleep') are dreamlike visions or auditory perceptions that occur when patients are tired or actually falling asleep. Patients with vivid dreamlike hallucinations have occasionally been misdiagnosed as schizophrenic.

Contrary to popular belief, patients with narcolepsy do not typically sleep more than normal people. Rather, they are unable to stay awake or asleep for long periods of time. Insomnia is thus common in patients with narcolepsy. Patients with narcolepsy typically fall asleep easily at any time during the day and night but have great difficulties staying asleep at night. The total amount of sleep during the 24-h day is thus often normal.

Dissociated REM Sleep in Narcolepsy

Many of the symptoms in narcolepsy are due to abnormal transitions from wakefulness to REM

sleep (called sleep-onset REM periods), a state associated with vivid dreaming, rapid eye movements, and a complete muscle paralysis. In normal subjects, REM sleep first occurs 90–120 min after falling asleep and then reoccurs periodically throughout the night. In narcolepsy, abrupt transitions from wakefulness to REM sleep often occur before the patient is completely asleep. This problem produces 'dissociated' states in which the patient is half awake and half in REM sleep, for example, being paralyzed (sleep paralysis) or dreaming (hypnagogic hallucinations) but awake. Cataplexy is also considered similar to REM sleep paralysis but occurs when the brain is excited by an emotion such as laughing or anger.

Genetic Predisposition to Narcolepsy–Cataplexy

The occurrence of narcolepsy–cataplexy involves both genetic predisposition and environmental triggers. Approximately 75% of reported monozygotic twins are discordant for narcolepsy, suggesting the importance of environmental factors. Multiplex families are rare, but a 10- to 40-fold increase in relative risk is reported in first-degree relatives.

One of the major genetic susceptibility factors for narcolepsy–cataplexy is HLA-DQB1 (Figure 1(a)). Almost all patients with typical cataplexy carry DQB1*0602, an HLA subtype found in 12% of Japanese, 25% of Caucasian, and 38% of African American controls. HLA-DQB1*0602 is always associated with DQA1*0102, another HLA allele encoded by the nearby DQA1 gene, located less than 20 kb telomeric to DQB1 (Figures 1(a) and 1(b)). The DQB1 association is especially tight in subjects with hypocretin deficiency; only four HLA DQA1*0102–DQB1*0602 negative narcolepsy subjects with low or undetectable cerebrospinal fluid (CSF) hypocretin-1 have been identified to date (out of several hundred subjects with CSF testing).

The DQB1*0602 association in narcolepsy is primary and not due to linkage disequilibrium with other loci. It is mostly due to DQA1*0102–DQB1*0602, but it is not simply a dominant effect. Indeed, DQB1*0602 homozygotes have a two- to fourfold higher risk than DQB1*0602/X heterozygotes, and specific DQB1*0602 heterozygotes are either at increased (e.g., DQB1*0602/DQB1*0301) or decreased (e.g., DQB1*0602/DQB1*0601 or DQB1*0602/DQB1*0501) risk (Figure 1(c)). Additional smaller effects have also been suggested for specific DRB1 alleles, another HLA locus located in the same region. The complexity of the HLA allele association in narcolepsy mirrors that reported in various autoimmune diseases, such as type I diabetes.

Genetic factors other than HLA are likely to be involved in the genetic susceptibility to narcolepsy–cataplexy but are not as well established or replicated. They are also likely to contribute far less to overall genetic risk than HLA-DQ. Studies in multiplex families suggest linkage to 4p13–q31 and 21q11.2 and possible associations with tumor necrosis factor (TNF)-α, catechol-O-methyl-transferase, and TNF receptor-2 polymorphisms. Notably absent in this list is the preprohypocretin gene and its two receptors, hypocretin receptor-1 and -2. Indeed, only a single mutation in any of these three loci has been identified. This case, an HLA-DQB1*0602-negative boy with hypocretin deficiency, had an unusually early onset of narcolepsy–cataplexy at 6 months of age. This boy had a potential dominant mutation in the signal peptide of the preprohypocretin gene that produces impaired hypocretin trafficking and, presumably, cell death.

Narcolepsy–Cataplexy Is Tightly Associated with Hypocretin Deficiency

The role of hypocretin in narcolepsy was first demonstrated in familial canine narcolepsy, a disease caused by mutations in the hypocretin receptor-2 gene, and through the study of a mouse preprohypocretin gene knockout model. This was followed by the demonstration of low CSF hypocretin-1 in 90% or more of cases with narcolepsy and cataplexy. Postmortem studies have shown a selective loss of 50 000–100 000 posterior hypothalamic neurons that produce the neuropeptide hypocretin (orexin) (Figures 2(a) and 2(b)). Loss, rather than decreased expression, is suggested by the concomitant loss of associated co-localized markers dynorphin and neuronal pentraxin-2. The HLA association and associated hypocretin deficiency is particularly strong (>90%) in patients with definite cataplexy (Figure 3). This association lends support to the hypothesis that narcolepsy is an autoimmune disorder.

Neurobiology of the Hypocretin Systems and Relevance to Narcolepsy

The hypocretin peptides (also called orexins) are encoded by the preprohypocretin gene. This gene encodes two related amidated polypeptides, hypocretin-1 and hypocretin-2, with high sequence homology. Hypocretin-1 (orexin-A), a 33-amino-acid peptide, has two disulfide bonds and is remarkably stable in biological fluids. It has an approximately equal affinity for the two known hypocretin receptors, hypocretin receptor-1 (HCRTR1) and hypocretin receptor-2 (HCRTR2). Hypocretin-2 (orexin-B), a

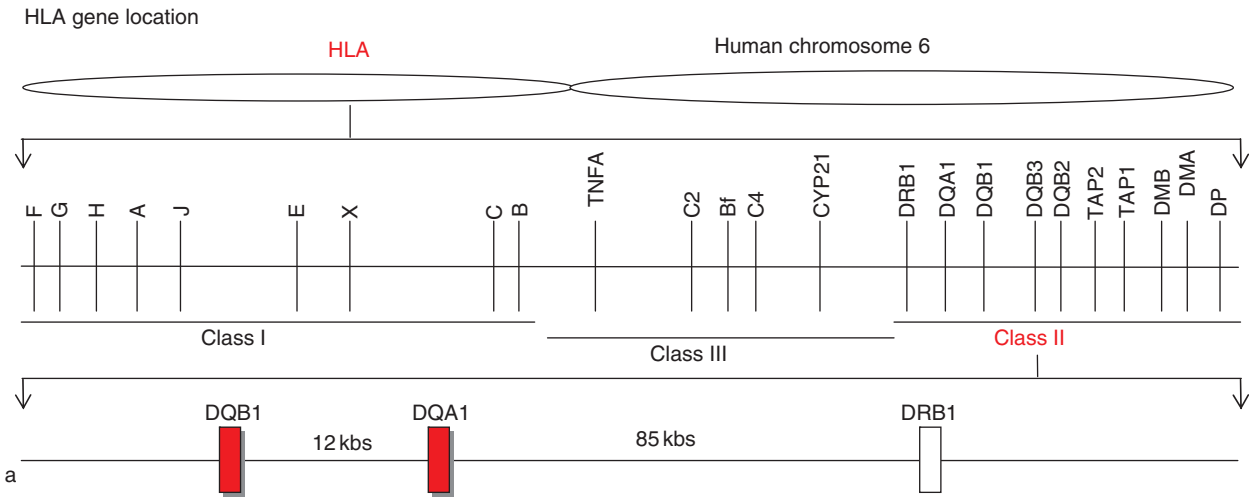

DQB1*0602 binding grove

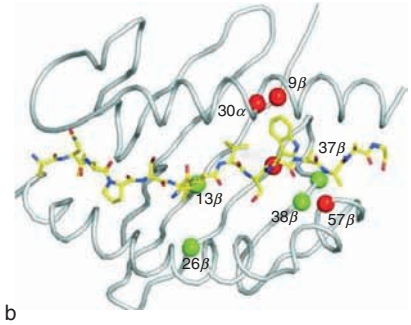

b

Susceptibility effects of various DQB1*0602 positive genotype combinations on narcolepsy in Caucasian, African American, and Japanese

DQA1-DQB1 allele combinations	Caucasian Narcolepsy $n=238$(%)	Caucasian Control $n=146$(%)	OR	African American Narcolepsy $n=77$(%)	African American Control $n=243$(%)	OR	Japanese Narcolepsy $n=105$(%)	Japanese Control $n=698$(%)	OR
0102–0602/0102–0602	43 (18.1)	3 (2.1)	10.5	18 (23.4)	9 (3.7)	7.9	16 (15.2)	5 (0.7)	24.9
0102–0602/X-0301	49 (20.6)	4 (2.7)	9.2	12 (15.6)	8 (3.3)	5.4	20 (19.0)	6 (0.9)	27.1
0102–0602/0101–0501	12 (5.0)	2 (1.4)	3.8	5(6.5)	3 (1.2)	5.6	2 (1.9)	5 (0.7)	2.7
0102–0602/0103–0601	0	0	0.0	0	0	0.0	8 (7.6)	15 (2.1)	3.8
0102–0602/other DQB1	109 (45.8)	25 (17.1)	4.1	37 (48.1)	56 (23.0)	3.1	59 (56.0)	54 (7.7)	15.3
non 0602/non 0602	25 (10.5)	112 (76.7)	0.0	5 (6.5)	167 (68.7)	0.0	0	613 (87.8)	0.0

c

Figure 1 Genetic susceptibility to narcolepsy: human leukocyte antigen effects. (a) Human leukocyte antigen (HLA) genes are numerous and located within the major histocompatibility complex on human chromosome 6. The HLA genes with the tightest genetic association with narcolepsy–cataplexy are the HLA class II region genes DRB1, DQA1 (encoding the DQα chain of the DQ molecule), and DQB1 (encoding the β chain of the DQ molecule). (b) Crystal structure of a DQ molecule associated with a peptide. The HLA-DQ molecule is a heterodimer formed by the polymorphic DQα and DQβ chains (encoded by DQA1 and DQB1 genes, respectively). The two chains form a peptide binding groove in which antigenic peptide (here, the hypocretin-1 peptide) can bind, to be presented to other cells of the immune system for the development of the immune response. (c) Specific DQA1–BQB1 allele combinations and genetic risk for narcolepsy–cataplexy in various ethnic groups. Allele groupings are ranked from the most predisposing to the most protective combinations. Almost all combinations with DQA1*0102–DQB1*0602 are at increased risk; one copy of DQA1*0102–DQB1*0602 is required in almost all cases for developing narcolepsy with cataplexy and hypocretin deficiency. Note that the highest odds ratio (OR) is observed for DQA1*0102–DQB1*0602 homozygotes. Other allele combinations confer various degrees of susceptibility, as often observed in other HLA-associated diseases. (b) Reproduced from Siebold C, Hansen BE, Wyer JR, et al. (2004) Crystal structure of HLA-DQ0602 that protects against type 1 diabetes and confers strong susceptibility to narcolepsy. *Proceedings of the National Academy of Sciences of the United States of America* 101: 1999–2004. (c) Reproduced from Mignot et al. (2001).

smaller peptide of 29 amino acids, has a tenfold higher affinity for HCRTR2.

Hypocretin neurons have tight functional interactions with cholinergic and monoaminergic systems regulating sleep (Figure 2(c)). These cell groups are enriched in hypocretin receptors, with a complementary distribution of the two receptors at the anatomical level. HCRTR1 is particularly abundant in the adrenergic locus coeruleus, whereas HCRTR2 is highly expressed on histaminergic cells of the tuberomammillary nucleus. Other regions of interest include hippocampus, paraventricular thalamus, and ventromedial hypothalamic nucleus for HCRTR1, and cortex, septal nuclei, paraventricular thalamus, paraventricular and arcuate hypothalamic nuclei, and pontine gray area for HCRTR2. In almost all cases, these two receptors have excitatory effects, although there is suggestion of promiscuous coupling of these receptors to multiple G-proteins with occasional inhibitory effects.

It has been hypothesized that the loss of excitatory input to monoaminergic cell groups mediates the sleepiness and short REM sleep in narcolepsy. Indeed, the activity of most monoaminergic cell groups such as locus coeruleus, raphe, and tuberomammillary nuclei, but not the dopaminergic systems in the

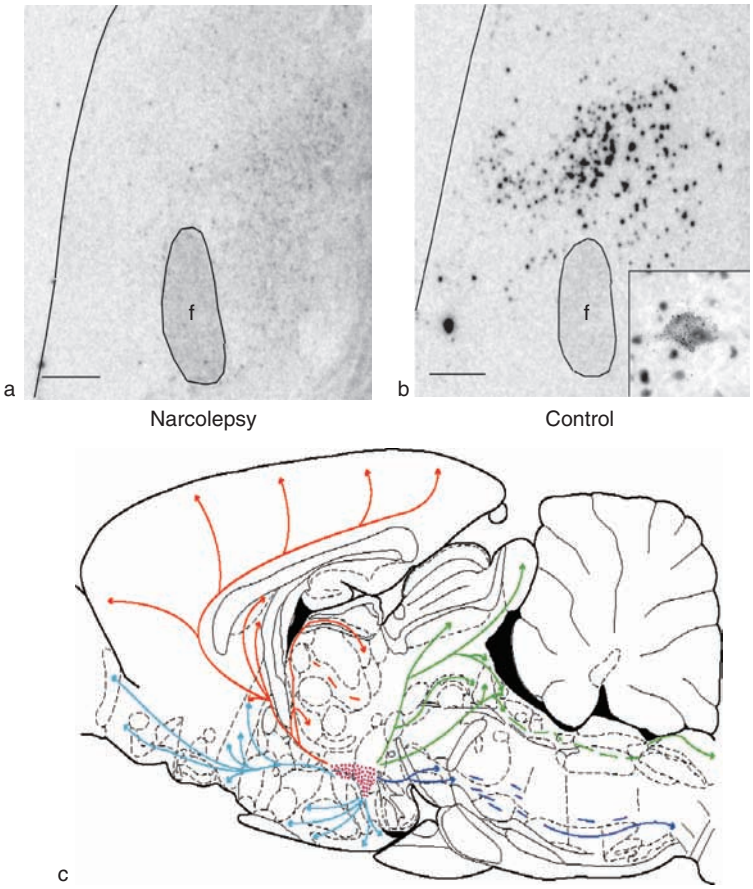

Figure 2 Preprohypocretin *in situ* hybridization in the posterior hypothalamic region in a narcoleptic (a) and control (b) human subject. Note loss of signal in narcolepsy, suggesting neuronal loss of this population. (c) Location and cell bodies of hypocretin-producing cells in a rat. Note the hypothalamic location of cell body and widespread projection through the entire neuroaxis. Scale bar = 1 cm (a, b). Reproduced from Peyron C, Faraco J, Rogers W, et al. (2000) A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nature Medicine* 6: 991–997. (c) Reproduced from Peyron C, Tighe DK, van den Pol AN, et al. (1998) Neurons containing hypocretin (orexin) project to multiple neuronal systems. *Journal of Neuroscience* 18: 9996–10015.

ventral tegmental area and substantia nigra, is depressed during sleep and almost silent during REM sleep. The fact that narcolepsy is typically treated with drugs that indirectly stimulate monoaminergic transmission (e.g., amphetamine-like compounds and antidepressants) is also in line with this hypothesis.

Interestingly, hypocretin effects on wakefulness after intracerebroventricular injections are largely blocked or attenuated by monoaminergic blockers, most notably antihistaminergic compounds, suggesting a primary role for downstream histaminergic systems. An important role for histamine is also suggested by findings indicating significantly decreased CSF histamine levels in narcolepsy–cataplexy, a result contrasting with more variable changes in other monoamine and metabolites.

Without hypocretin, narcoleptic patients have sleepiness, inappropriate REM paralysis during wakefulness (cataplexy and sleep paralysis), REM dreaming before falling asleep (hypnagogic hallucinations), and disorganized nighttime sleep. Rapid transitions into REM sleep can be observed during nocturnal sleep and while napping. Finally, narcolepsy is often associated with an increased body mass index and obesity, a possible consequence of sleepiness and decreased energy expenditure.

Little is known regarding the normal function of hypocretins. Hypocretin neurons integrate metabolic and sleep/wake-related inputs. The relative contribution of these various factors may vary across species. In rodents, these cells have receptors for leptin and ghrelin, and they are sensitive to glucose changes, likely facilitating wakefulness in reaction to changes in food availability. In rodents, electrophysiological recording of these neurons indicate maximal activity during wakefulness, especially during purposeful activity. Activity is low during non-REM and REM

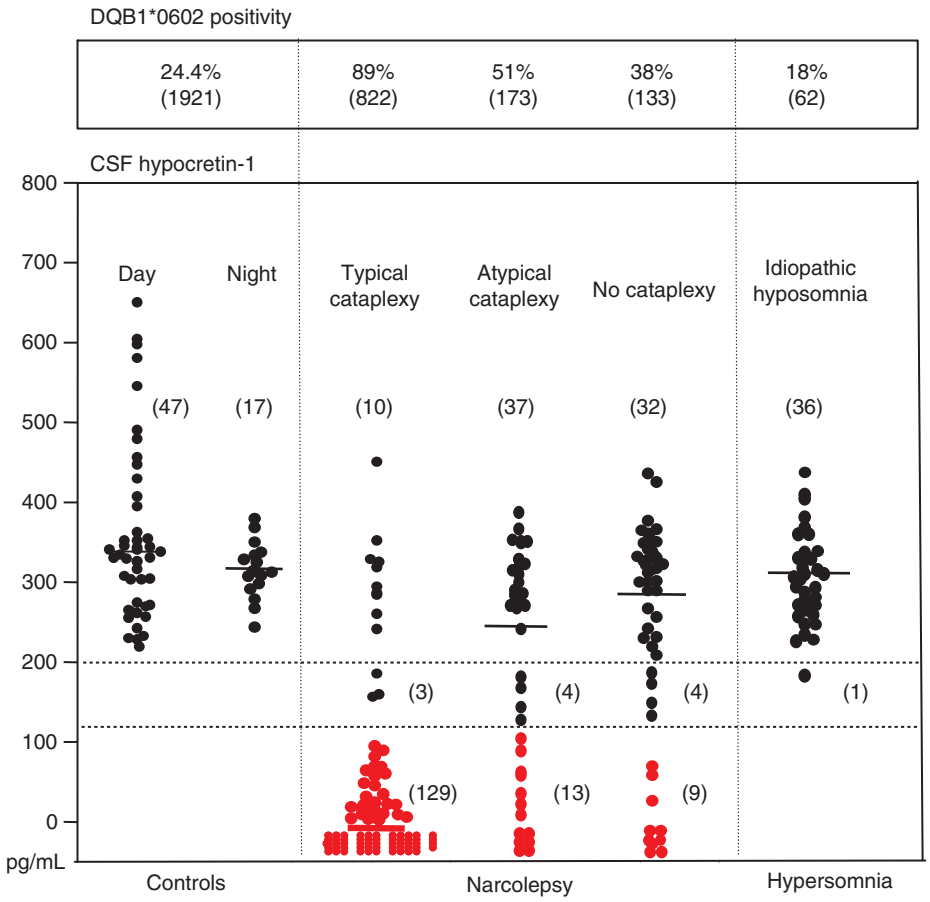

Figure 3 HLA-DQB1*0602 positivity (top) and hypocretin deficiency (bottom), as measured in the cerebrospinal fluid (CSF). (Top) Almost all patients with typical cataplexy are DQB1*0602, suggesting etiological homogeneity. DQB1*0602 positivity decreases in patients with atypical and no cataplexy, but it is still higher than control values. Subjects with idiopathic hypersomnia (isolated sleepiness and no REM transitions) have a normal DQB1*0602 frequency, suggesting no pathophysiological overlap. (Bottom) CSF hypocretin-1 in controls, narcolepsy, and idiopathic hypersomnia. The large majority of subjects with typical cataplexy have low CSF hypocretin-1. Quantitative receiver curve analysis has indicated that values below 110 pg ml^{-1} are most predictive of a narcolepsy diagnosis, although control values are typically above 200 pg ml^{-1}. Only 20–40% of cases without cataplexy have low CSF hypocretin-1. None of the controls or cases with idiopathic hypersomnia have low CSF hypocretin-1. (Top) Reproduced from Lin L and Mignot E (2007) HLA and narcolepsy: Present status and relationship with familial history and hypocretin deficiency. In: Bassetti C, Billiard M, and Mignot E (eds.) *Narcolepsy and Hypersomnia*. New York: Dekker/Taylor & Francis. (Bottom) Reproduced from Mignot E, Lammers GJ, Ripley B, et al. (2002) The role of cerebrospinal fluid hypocretin measurement in the diagnosis of narcolepsy and other hypersomnias. *Archives of Neurology* 59: 1553–1562.

sleep. Hypocretin release is also tightly coupled with locomotor activity in rodents, an effect less evident in wake-consolidated species such as the squirrel monkey. Hypocretin release is also driven by the circadian clock and peaks at the end of the active (wake) period. Finally, hypocretin release is increased by even a brief sleep deprivation, suggesting a role for the hypocretin system in consolidating wakefulness in the face of a mounting sleep debt. This last role of hypocretin may explain the inability of narcolepsy patients to stay awake for periods typically longer than a few hours, as sleep debt is increasing. As expected for a wake-promoting signal, these cells are more active during active wake, decrease their activity during non-REM sleep, and are least active during REM sleep. The sleep stage-specific activity is consistent with a role in promoting wakefulness and inhibiting REM sleep.

Is Narcolepsy an Autoimmune Disorder?

The tight HLA association and the involvement of both genetic and nongenetic factors in the pathophysiology of narcolepsy suggest an autoimmune basis for the disorder. This hypothesis is also consistent with the peripubertal onset of the disease but not the even sex ratio found in the disorder. The discovery of hypocretin-producing cells as a potential target has given further credence to the autoimmune hypothesis. HLA-DQ is expressed on the surface of B cells,

macrophages (including microglia), and activated T cells. It is a heterodimer composed of an α and a β chain encoded by the polymorphic DQA1 and DQB1 genes, two genes in tight linkage disequilibrium. HLA-DQ binds peptide fragments and can present the resulting complex to other cells of the immune system, as recognized by the T cell receptor complex (Figure 1).

How HLA molecules in general and HLA-DQ in particular predispose to autoimmune disorders is not well understood. It probably involves the ability of some HLA alleles to bind and consequently present specific autoantigens to the rest of the immune system. In other HLA-associated diseases, however, antibodies directed toward the potential target (e.g., islet cell antibodies in type I diabetes) or T cell clones with primary responsiveness to target antigens (e.g., against myelin basic protein in multiple sclerosis) can be detected. In narcolepsy, however, all attempts to demonstrate T cell reactivity against hypocretin peptides or to detect autoantibodies directed at hypocretin-producing cells have failed.

These generally negative results do not necessarily exclude the possibility of an autoimmune mechanism, but they do raise the possibility of other, more complex neuroimmune mechanisms. Another possibility is an infectious agent with particular tropism toward hypocretin-producing cells, although HLA association in infectious diseases is usually not very tight or it is primarily to a single allele. Alternatively, it has been suggested that these neurons could be unusually sensitive to excitotoxicity. Interest in the autoimmune arena has been rekindled by case reports that intravenous immunoglobulin may reduce the development of narcolepsy severity if administered within a year of onset.

Narcolepsy without Cataplexy

The importance of cataplexy in the diagnosis of narcolepsy has long been suggested by clinical studies. The finding that almost all cases with definite cataplexy (defined as recurrent episodes of muscle paralysis triggered at least sometimes by typical emotions such as laughing or joking) are HLA-DQB1*0602 positive strongly suggests etiological homogeneity in these cases. This was further confirmed by the finding that almost all HLA-positive patients with cataplexy have low CSF hypocretin-1 (Figure 3). In contrast, HLA-DQB1*0602 is typically found in 40% of patients without cataplexy or with doubtful or atypical cataplexy (rare events or only events triggered by unusual emotions). This is matched by a much lower percentage of patients without cataplexy who have low CSF hypocretin-1 (5–30% of patients, depending on the specific case series) (Figure 3). Some of these cases (approximately one-third) are children with recent onset ($\leq$4 years) who are likely to develop cataplexy within a few years.

These results are consistent with two possible, nonexclusive pathophysiological models for narcolepsy without cataplexy. In the first model, narcolepsy both with and without cataplexy is part of the same disease continuum, with similar pathophysiological effects on hypocretin transmission. Narcolepsy with cataplexy is a generally more severe form of the disease, most commonly associated with an almost complete destruction of the hypocretin system. HLA-DQB1*0602 may be a severity factor more likely to predispose to complete hypocretin destruction. Only in rare cases with cataplexy would hypocretin cell loss be partial but sufficient to produce normal levels of hypocretin and a narcolepsy phenotype. In these cases, compensation by the remaining hypocretin cells likely maintains CSF hypocretin-1 at normal levels. Projections with more effects on CSF hypocretin-1 but of less functional importance for the phenotype (e.g., hypocretin afferents to the spinal cord) also may be more spared in cases with cataplexy and normal CSF levels.

In narcolepsy without cataplexy, cell loss would be less pronounced. The severity of the disease would be generally lower and the cell loss insufficient to result in low CSF hypocretin-1. The HLA-DQB1*0602 association with narcolepsy without cataplexy is weaker. Only in rare cases might neuroanatomical destruction be pronounced, resulting in low CSF hypocretin-1 but still sparing a select hypocretin cell subpopulation projecting to cataplexy-triggering pathways. In support of this model is a study by Thannickal et al. in which 14% of hypocretin cells remained in a case without cataplexy versus 4.4–9.4% remaining cells in five other subjects with cataplexy. Lesion studies in rats also indicate a 50% decrease in CSF hypocretin-1 with a 77% destruction of hypocretin cells, suggesting some degree of compensation.

The fact that HLA-DQB1*0602 is slightly increased in frequency in cases without cataplexy but with normal CSF hypocretin-1 is also consistent with this hypothesis. This finding, however, remains to be confirmed in a population-based sample because HLA positivity may have been increased in clinical samples due to a bias in inclusion (some clinicians use HLA typing to confirm the diagnosis of narcolepsy without cataplexy). The model also predicts that a large number of narcolepsy without cataplexy cases may exist in the general population but have a milder phenotype, consistent with our observation of slightly decreased REM latency in normal population subjects with HLA-DQB1*0602. Similar milder forms of the disease or long latent periods have been suggested for

other autoimmune diseases, such as DQ2-associated celiac disease, B27-associated spondylarthropathies, and DR3/DR4-associated type I diabetes. In type I diabetes, for example, a larger number of individuals, especially relatives of affected probands, may be positive for islet cell antibodies without ever developing the disease.

In the second model, the cause of cases without low CSF hypocretin does not involve partial hypocretin cell loss. Other systems downstream of hypocretin, such as hypocretin receptors, histamine, or neuroanatomical systems not associated with hypocretin neurobiology, may be involved. Of note, the two models may not be entirely exclusive. A partial hypocretin cell loss may, for example, not be sufficient in itself to produce symptoms but could lead to narcolepsy when associated with additional defects downstream. A similar model, albeit speculative, has been proposed in some obese-type diabetes children in which both types I and II may coexist. In these cases, subjects with partially reduced islet cell numbers may be asymptomatic when lean but develop insulin resistance if obesity develops and the remaining islet population is unable to produce enough insulin to keep up with increased tissue demand.

To distinguish between these models, neuropathological studies of narcolepsy without cataplexy cases are urgently needed. It is also imperative to study narcolepsy without cataplexy not only in clinical samples but also in the general population because clinical cases may represent a more severe and selected subpopulation. We believe that the most likely explanation will be a combination of these models, including both disease heterogeneity with respect to hypocretin neuropathology and severity gradients for hypocretin cell loss as discussed previously. The extent of this overlap remains to be defined.

The previous discussion is important because the real population-based prevalence of narcolepsy without cataplexy is unknown. Data by our group and others suggest that several percent of the population have unexplained sleepiness and REM transition. Whether a small or large portion of these subjects have a mild degree of hypocretin dysfunction (possibly related pathophysiologically to narcolepsy–cataplexy) is unknown.

Therapies for Narcolepsy

Good sleep hygiene, education, and treatment compliance are important to the management of narcolepsy. Referral to support groups such as the Narcolepsy Network is helpful. Fixed wake-up times, sleep diaries, and regular sleep schedule are recommended. Obesity may develop especially in young children when disease onset is abrupt. It is useful to restrict diet, encourage exercise, and treat sleepiness aggressively at this stage. The risk of driving while sleepy, especially prior to adequate therapy, must be discussed and, if appropriate, regulatory agencies notified. Patients with narcolepsy must avoid jobs that put others in danger and should consider activities that are less sedentary. Jobs that involve repetitive tasks or sitting and looking at a computer all day can be difficult. Employers or teachers can be asked to accommodate 15- to 30-min scheduled naps.

Modafinil is now first-line, standard-of-care treatment for sleepiness associated with narcolepsy. Amphetamine-like stimulants, such as methylphenidate, dextroamphetamine, amphetamine racemic mixture, and methylamphetamine, can also be used but have a number of disadvantages. They are potentially addictive and should preferably be reserved for patients with a well-established diagnosis (e.g., those with cataplexy). Additionally, side effects such as palpitations, hypertension, and nervousness may occur because of their action on the autonomic nervous system. The mode of action of amphetamine-like stimulants and modafinil on wakefulness is thought to involve increased dopaminergic transmission through the inhibition of dopamine reuptake (modafinil and methyphenidate), by increasing dopamine release (methyphenidate), or through more complex effects on release and storage (amphetamine).

Antidepressants are commonly used to treat cataplexy. The rationale for their efficacy is the profound REM suppressant effects most antidepressant share. They are also effective for sleep paralysis and hypnagogic hallucinations, a sometime disabling symptom. Tricyclics are very effective but have anticholinergic side effects. Serotonin reuptake specific inhibitors are efficacious, but relatively high doses are generally needed. Newer medications targeting norepinepherine reuptake, such as venlafaxine and atomoxetine, are most effective. The dual serotoninergic/noradrenergic reuptake inhibitor venlafaxine is a typical first-line treatment. Atomoxetine, a specific adrenergic reuptake inhibitor typically prescribed for attention deficit hyperactivity disorder, can also be used and as it is mildly stimulant. Patients should be warned that rebound cataplexy usually occurs when medications are discontinued, changed, or skipped.

Sodium oxybate (γ-hydroxybutyrate (GHB)) is unique because it is efficacious on all symptoms. This naturally occurring compound is a potent but short-acting hypnotic that consolidates slow-wave and REM sleep, most likely via $GABA_B$ agonistic effect. It has been available since the 1970s but was only recently approved for the treatment of cataplexy. It is thought that the consolidation of nocturnal sleep and REM sleep leads to decreased daytime symptoms. The drug

is carefully controlled and can be addictive. It is typically reserved for narcoleptic with both disturbed nocturnal sleep and cataplexy.

Narcolepsy is frequently associated with other sleep disorders. Sleep apnea, a condition in which patients snore and gasp during sleep, is common and must be treated especially if GHB is to be used. Periodic leg movements (repetitive leg twitching occurring during sleep) are also common but very rarely need treatment.

Perspectives for New Treatments

The realization that narcolepsy and related pathologies are not rare is fueling pharmaceutical interest in this area. Novel symptomatic agents are being developed. Agents that act on the histaminergic system, such as H3 antagonists (compounds that increase histamine release through the blockade of the H3 autoreceptor), are being studied. These compounds are wake promoting and effective in animal models of narcolepsy.

New treatments aimed at replacing the missing hypocretin hormone are being developed and will probably be much more effective. These may involve developing hypocretin receptor agonists, preferentially of the HCRTR2 type, using gene therapy or using cell-based replacement therapy. Research aimed at discovering the process that destroys the hypocretin-containing cells in the brain is ongoing. Understanding this process is necessary before truly preventive and curative strategies can be developed.

See also: Cataplexy; Hypocretin/Orexin and MCH and Receptors; Sleep Apnea; Sleep in Adolescents.

Further Reading

Bassetti C, Billiard M, and Mignot E (2007) *Narcolepsy and Hypersomnias of Central Origin*. New York: Dekker/Taylor & Francis.

Blouin AM, Thannickal TC, Worley PF, et al. (2005) Narp immunostaining of human hypocretin (orexin) neurons: Loss in narcolepsy. *Neurology* 65: 1189–1192.

Crocker A, Espana RA, Papadopoulou M, et al. (2005) Concomitant loss of dynorphin, NARP, and orexin in narcolepsy. *Neurology* 65: 1184–1188.

Dauvilliers Y, Carlander B, Rivier F, et al. (2004) Successful management of cataplexy with intravenous immunoglobulins at narcolepsy onset. *Annals of Neurology* 56: 905–908.

Honda Y, Asaka C, Tanimura M, and Furusho T (1983) A genetic study of narcolepsy and excessive daytime sleepiness in 308 families with a narcolepsy or hypersomnia proband. In: Guilleminault C and Lugaresi E (eds.) *Sleep/Wake Disorders: Natural History, Epidemiology and Long-Term Evolution*, pp. 187–199. New York: Raven Press.

Lin L, Faraco J, Li H, et al. (1999) The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* 98: 365–376.

Lin L and Mignot E (2007) HLA and narcolepsy: Present status and relationship with familial history and hypocretin deficiency. In: Bassetti C, Billiard M, and Mignot E (eds.) *Narcolepsy and Hypersomnia*, ch. 37, pp. 411–426. New York: Dekker/Taylor & Francis.

Mignot E (1998) Genetic and familial aspects of narcolepsy. *Neurology* 50(2 supplement 1): S16–S22.

Mignot E, Lammers GJ, Ripley B, et al. (2002) The role of cerebrospinal fluid hypocretin measurement in the diagnosis of narcolepsy and other hypersomnias. *Archives of Neurology* 59: 1553–1562.

Mignot E, Lin L, Finn L, et al. (2006) Correlates of sleep onset REM periods during the Multiple Sleep Latency Test in community adults. *Brain* 129: 1609–1623.

Mileykovskiy BY, Kiyashchenko LI, and Siegel JM (2005) Behavioral correlates of activity in identified hypocretin/orexin neurons. *Neuron* 46: 787–798.

Nishino S and Mignot E (1997) Pharmacological aspects of human and canine narcolepsy. *Progress in Neurobiology* 52: 27–78.

Peyron C, Faraco J, Rogers W, et al. (2000) A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nature Medicine* 6: 991–997.

Peyron C, Tighe DK, van den Pol AN, et al. (1998) Neurons containing hypocretin (orexin) project to multiple neuronal systems. *Journal of Neuroscience* 18: 9996–10015.

Scammell T (2006) The frustrating and mostly fruitless search for an autoimmune cause of narcolepsy. *Sleep* 29: 601–602.

Siebold C, Hansen BE, Wyer JR, et al. (2004) Crystal structure of HLA-DQ0602 that protects against type 1 diabetes and confers strong susceptibility to narcolepsy. *Proceedings of the National Academy of Sciences of the United States of America* 101: 1999–2004.

Thannickal TC, Moore RY, Nienhuis R, et al. (2000) Reduced number of hypocretin neurons in human narcolepsy. *Neuron* 27: 469–474.

Wisor JP, Nishino S, Sora I, et al. (2001) Dopaminergic role in stimulant-induced wakefulness. *Journal of Neuroscience* 21: 1787–1794.

Yoshida Y, Fujiki N, Nakajima T, et al. (2001) Fluctuation of extracellular hypocretin-1 (orexin A) levels in the rat in relation to the light–dark cycle and sleep–wake activities. *European Journal of Neuroscience* 14: 1075–1081.

Natriuretic Peptides

M L Kuroski de Bold and A J de Bold, University of Ottawa Heart Institute, Ottawa, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The natriuretic peptide (NP) system comprises three polypeptides hormones: atrial natriuretic factor (ANF or ANP), brain natriuretic peptide (BNP), and C-type natriuretic peptide (CNP). These peptides are encoded by three different genes which encode a similar 17-amino-acid disulphide loop present in all three peptides (Figure 1). ANF and CNP amino acid sequences are highly conserved. In contrast, the amino acid sequence of BNP varies considerably among the species.

Peripherally, ANF and BNP are synthesized and stored in atrial cardiocytes. Increased secretion of these peptides over baseline is observed in response to stretch of the atrial wall and various neurohumoral stimuli. ANF and BNP are also present in the ventricles of the heart in concentrations that are many times lower than that of the atria. These NPs are expressed in tissues other than cardiac, such as in kidney and gonads. ANF and BNP gene expression in neural tissue is widespread.

ANF and BNP share potent vasodilatory, diuretic, natriuretic, and antigrowth properties. They also inhibit renin and aldosterone production and sympathetic tone (Figure 2). The diuretic and natriuretic actions exerted by these hormones occur without the expected rebound response by the renin–angiotensin–aldosterone and sympathetic systems.

CNP does not have natriuretic activity but is a potent vasodilator and inhibits smooth muscle proliferation. CNP is present in significant amounts in the central nervous system (CNS) and in the vascular endothelium. In cerebrospinal fluid CNP is the most abundant NP.

There are three subtypes of receptors for NPs: NP receptor A (NPR-A), NP receptor B (NPR-B), and NP receptor C (NPR-C) (Figure 3). ANF and BNP bind to NPR-A, which has guanylyl cyclase activity. The ligands thus induce an increase in intracellular cyclic guanosine monophosphate (cGMP) levels as a first step in a signaling cascade that may involve enzymes or ion channels. ANF has 50 times more affinity for NPR-A than BNP. NPR-A is expressed in a variety of tissues including kidney, adrenals, lung, thymus, and the gastrointestinal tract. In the brain, NPR-A is found in the cortex, hippocampus, thalamus, amygdala, and cerebellum.

CNP binds exclusively to NPR-B, which is also a guanylyl cyclase. It is the most abundant NP receptor in the rat hypothalamus.

NPR-C is a clearance receptor. The affinity of this receptor for NP is ANF > CNP ≥ BNP. There is some evidence showing that NPR-C might be coupled to G-protein signaling. Peripherally, NPR-C is expressed by many tissues. In the CNS, NPR-C is localized in the striatum, neocortex, hippocampus, amygdala, and pituitary gland, among others.

Natriuretic Peptides in the CNS

ANF

ANF is present in areas that are involved in neuroendocrine regulation such as the hypothalamus, thalamus, septum, olfactory bulb, and cortex. In the CNS, most ANF-containing cell bodies are found in the anterolateral region of the third ventricle. High levels of ANF are found in the anteroventral third ventricular area (AV3V) and septal regions. The distribution of ANF and its receptor in the brain, much of it shared with the angiotensin 2 receptor, suggest that they might be involved in the modulation of blood pressure, fluid regulation, and sympathetic tone. In the spontaneously hypertensive rat, a reduction in ANF-binding sites in all areas of the brain has been reported as compared with normotensive Wistar rats. An upregulation of ANF receptors in the subfornical organ has been shown after acute dehydration.

ANF and inflammation In the brain, microglial cells, the macrophage-like cells that are activated by neuroinflammation induced by lipopolysaccharide, brain injury, or infection, have been shown to express the message for the ANF receptors NPR-A, NPR-B, and NPR-C. Exposure of microglial cells *in vitro* to bacterial endotoxin induces a significant intracellular increase in nitric oxide and interleukin-1β. These effects are inhibited by ANF.

The hypothalamic–pituitary–adrenocortical (HPA) system and ANF Several studies indicate that ANF plays an important role in the modulation of the neurohumoral response to stress hormones (e.g., vasopressin, cortisol, and catecholamine). An increase in the release of corticotrophin-releasing hormone stimulates the secretion of ANF, which in turn decreases cortisol and catecholamine secretion and decreases sympathetic activity. Thus, the decrease in the stress hormone response due to ANF supports the notion that this peptide might exert a direct interaction with

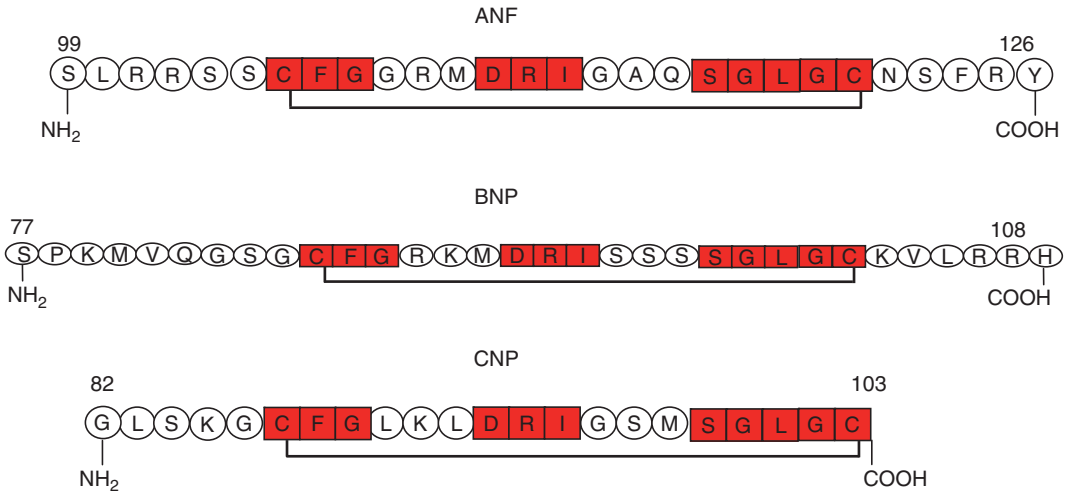

Figure 1 Human natriuretic peptides amino acid sequence. Atrial natriuretic factor (ANF), brain natriuretic peptide (BNP), and C-type natriuretic peptide (CNP) share a 17-amino-acid ring created by a disulphide bond between two cysteine residues. The conserved amino acid residues are shown.

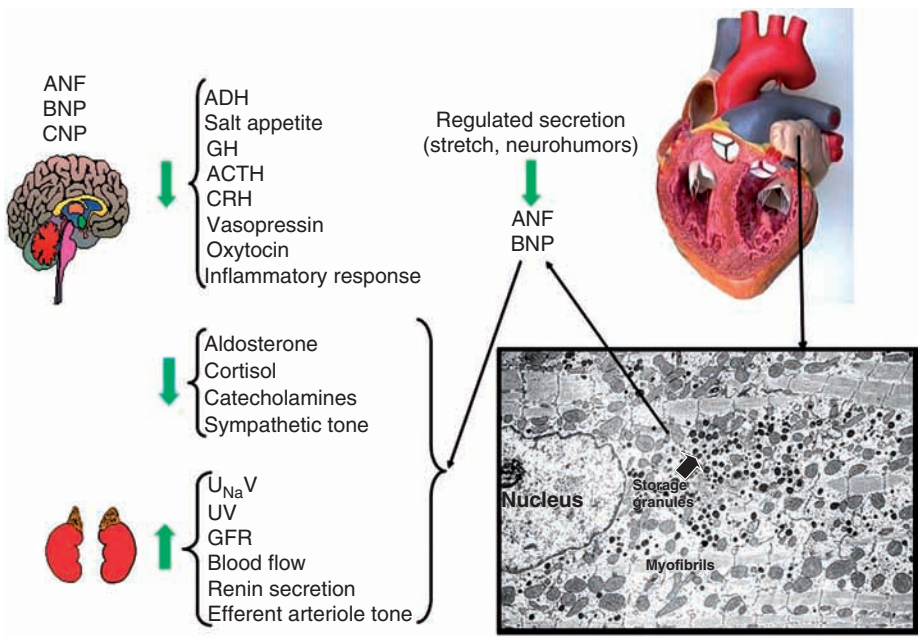

Figure 2 Main sites of synthesis for natriuretic peptides (ANF, atrial natriuretic factor; BNP, brain natriuretic peptide; CNP, C-type natriuretic peptide) and their biological actions are shown. ADH, antidiuretic hormone; ACTH, adrenocorticotropin; CRH, corticotrophin-releasing hormone; $U_{Na}V$, urinary sodium excretion; UV, urine volume; GFR, glomerular filtration rate.

the hypothalamic–pituitary–adrenocortical (HPA) system and the autonomic nervous system in conditions such as panic and anxiety disorders. Most of the studies to date have been short term; the long-term effects of ANF upon the CNS need to be investigated.

CNP appears to have the opposite effects of ANF, but there are few studies reported on this peptide's effects on the brain.

ANF and alcohol withdrawal High ANF circulating levels and brain-secreted ANF appeared to be involved in modifying the clinical symptoms of alcohol withdrawal (AW) reactions. In AW there are significant changes in volume and electrolyte homeostasis due to the changes in the secretion of hormones such as ANF, aldosterone, and renin. In humans, it has been shown that in patients with severe AW, the plasma ANF,

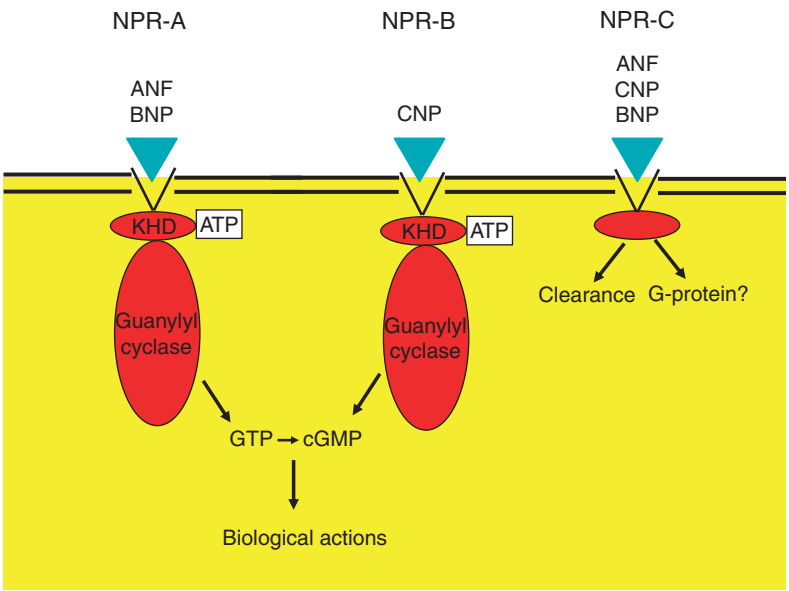

Figure 3 Natriuretic peptide receptors. Atrial natriuretic factor (ANF) and brain natriuretic peptide (BNP) binding activates natriuretic peptide receptor A (NPR-A), a process that requires adenosine triphosphate (ATP), which results in the increased production of cyclic guanosine monophosphate (cGMP). NPR-B, natriuretic peptide receptor B; NPR-C, natriuretic peptide receptor C; CNP, C-type natriuretic peptide; KHD, kinase homology domain; GTP, guanosine triphosphate.

renin activity, and aldosterone levels were increased as compared to age-matched controls. During partial recovery, plasma renin activity and aldosterone returned to normal levels but ANF remained high for the duration of the observation period. It was also noted that patients who developed delirium tremens showed plasma ANF levels significantly higher than in those who were undergoing withdrawal but who did not exhibit delirium tremens. This study confirmed the fact that in alcoholic patients there are significant alterations of the hormones that regulate fluid and electrolyte homeostasis and that measurement of ANF might be a useful diagnostic tool to predict the onset of delirium tremens.

The relationship between ANF and alcohol has been investigated in alcohol-tolerant/dependent mice. Upon alcohol withdrawal these animals exhibit hyperreactivity to noise, handling, and convulsion similar to humans. The hyperreactivity to handling was measured following removal of alcohol before and after intracerebral administration of ANF. It was found that medium to high doses of ANF significantly reduced handling-induced convulsions. These studies suggest that ANF plays a significant role in the control of the CNS symptoms of alcohol withdrawal.

BNP

BNP was originally isolated from the porcine brain, but its main site of synthesis, storage, and release is, like ANF, in the atria of the heart. BNP-binding sites are found in the subfornical organ, supraoptic nucleus, and paraventricular nucleus. BNP was shown to compete with the majority of the high-affinity binding sites for ANF in the rat's brain. Immunohistochemically, BNP is localized in the hypothalamus of the rat and the cynomolgus monkey. These findings suggest that BNP in concert with ANF is involved in CNS mechanisms for modulating cardiovascular homeostasis.

CNP

CNP is abundantly expressed in the brain and vascular endothelial cells. CNP and its cognate receptor NPR-B have been detected in various regions of the brain (e.g., the subfornical organ, median preoptic nucleus, and interpeduncular nucleus). CNP and NPR-B are also found in the olfactory epithelium, suggesting that they might play a role in the differentiation of olfactory receptor neurons. CNP and NPR-B gene expression is present in the embryonic mouse nervous system as early as embryonic day 10.5. The structures included developing ganglia, cranial trigeminal ganglion, and segmental dorsal ganglia. These studies show that CNP through its receptor NPR-B exerts multiple effects such as stimulation and inhibition of DNA synthesis in Schwann cells and embryonic hindbrain precursors. CNP has also been shown to inhibit proliferation of olfactory precursors; thus, it is apparent that through NPR-B, CNP could play area- and phase-specific tasks throughout the development of the peripheral and central nervous

systems. Nevertheless, the exact function of CNP in the CNS has not been elucidated as yet.

See also: Atrial Natriuretic Peptide: Fluid/Mineral Balance; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Inflammation in Neurodegenerative Disease and Injury; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Discovery.

Further Reading

Antoni FA, Hunter EF, Lowry PJ, Noble JM, and Seckl JR (1992) Atriopeptin: An endogenous corticotropin-release inhibiting hormone. *Endocrinology* 130: 1753–1755.

Brown J and Czarnecki A (1990) Autoradiographic localization of atrial and brain natriuretic peptide receptors in rat brain. *American Journal of Physiology* 258: R57–R63.

DiCicco-Bloom E, Lelievre V, Zhou X, Rodriguez W, Tam J, and Waschek JA (2004) Embryonic expression and multi-functional actions of the natriuretic peptides and receptors in the developing nervous system. *Developmental Biology* 271: 161–175.

Imura H, Nakeo K, and Itoh H (1992) The natriuretic peptide system in the brain: Implications in the central control of cardiovascular and neuroendocrine functions. *Frontiers in Neuroendocrinology* 13: 217–249.

Kovác GL (2000) The role of atrial natriuretic peptide in alcohol withdrawal: A peripheral indicator and central modulator? *European Journal of Pharmacology* 405: 103–112.

Kurihara M, Saavedra JM, and Shigematsu K (1987) Localization and characterization of atrial natriuretic peptide binding sites in discrete areas of the rat brain and pituitary gland by quantitative autoradiography. *Brain Research* 408: 31–39.

Ma KK and de Bold AJ (2004) Atrial natriuretic factor and family of natriuretic peptides. *Encyclopedia of Endocrine Diseases* 1: 295–300.

McGrath MF, de Bold ML, and de Bold AJ (2005) The Endocrine function of the heart. *Trends in Endocrinology and Metabolism* 16(10): 469–477.

Moriyama N, Taniguchi M, Miyano K, Miyoshi M, and Watanabe T (2006) ANP inhibits LPS-induced stimulation of rat microglial cells by suppressing NF-κB and AO-1 activations. *Biochemical and Biophysical Research Communications* 350: 322–328.

Strohle A and Holsboer F (2003) Stress responsive neurohormones in depression and anxiety. *Pharmacopsychology* 36(supplement 3): S207–S214.

Vollmar AM (2005) The role of atrial natriuretic peptide in the immune system. *Peptides* 26: 1086–1094.

Wiedemann K, Jahn H, and Kellner M (2000) Effects of natriuretic peptides upon hypothalamo-pituitary-adrenocortical system activity and anxiety behaviour. *Experimental and Clinical Endocrinology and Diabetes* 108: 5–13.

Natural Images: Coding Efficiency

D J Graham and D J Field, Cornell University, Ithaca, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Images and movies of the natural world are known to share a variety of statistical regularities. Such stimuli show consistent spatial statistics, spatiotemporal statistics, contrast and intensity distributions, and chromatic structure. Consider the images shown in Figure 1. The top row shows white noise and the bottom row shows natural scenes. A process that generates random white noise images will generate all possible images with equal probability. However, because of the statistical regularities of natural scenes, the possibility of seeing a natural scene generated from a white noise process is extremely low. For even a small 8×8 pixel patch, the entropy of the white noise patches with 8 bits of gray level is $8 \times 8 \times 8$, or 512, bits. This results in a total of 2^{512}, or about 10^{154} equally probable images. Natural scenes are considerably redundant (i.e., they have much lower entropy), with estimates that they contain approximately 40% of the entropy of white noise for small (8×8) patches, and still lower relative entropy for larger patches. This difference would in turn imply that a single white noise pattern out of 10^{90} would have the basic statistics of natural scenes (assuming a flat distribution). Although the precise ratio requires an estimate of the true distribution, the estimate does demonstrate just how redundant natural scenes are. Because of this redundancy, it is possible to build a visual system that is efficient if it is specifically dedicated to representing such an environment. A visual system can reduce the size of its problem space and focus its efforts on what it is likely to encounter in the world. Because of this redundancy, visual systems can be shaped by evolution and development to take advantage of the environment.

Using the supposition that the visual system employs codes that approach optimal efficiency with respect to predictable (i.e., redundant) structure, one can go some way in explaining why visual neurons show the coding properties they do. A number of studies have shown success at predicting many of the basic linear properties of visual neurons. In the following sections we will review some of this work. But we note that this work and any extension of it depend on the appropriate definition of efficiency. In the classical definition of efficiency used in engineering, a system is most efficient if all of its work (within limits imposed by thermodynamics) is done in service of its task. In this view, efficiency presumes that the system has some well-defined tasks, and that the system's design is a direct reflection of the need to do as little work as possible to achieve those tasks. But how do we define these tasks for biological systems?

Efficient for What Task?

Many sensory systems, the task or 'goal' of which is often readily definable and measurable, have sensitivity that is near their physical limits. With respect to the brain as a whole, however, the only clearly defined task is the most general one in biology: differential reproductive success. But this goal tells us little about why a visual system would have the structure it does. Clearly, the visual system is involved in vision, but to treat it as an independent subsystem with some basic input/output relationship is too much of a simplification. Certainly, the output of the visual system cannot be reduced to a behavior. And although it may seem computationally reasonable, there is little reason to presume that the output of the visual system is simply some 'object-detector.' There are three reasons for this: (1) object recognition is not a well-localized processing task in cortex, (2) no known congenital disorder obliterates this ability exclusively, and (3) patients with visual agnosia caused by lesions typically show severe deficits in other visual faculties. Even though object recognition is distributed, one might still propose that object recognition could be the collective goal of the visual system's many parts. But object recognition surely competes for neural real estate with areas whose goals include accurate spatial mapping, motion predictions, and a host of other tasks.

This difficulty in defining the visual system's general goal is likely to prove a significant hurdle for studies attempting to apply efficient coding theories to later stages of processing. In addition, our current ability to predict a neuron's behavior is significantly limited at these later stages. Much of the success of efficient coding techniques has been in the application to neurons relatively early in the visual pathway (retina, lateral geniculate nucleus (LGN), and V1). For neurons in these areas, the family of response properties is relatively well defined and it is possible to talk about how the information in the visual image is represented with the array of neurons in each of these areas. This in turn allows us to address the question of whether the array of neurons is processing the information efficiently. However, the approach still requires a precise definition of efficiency.

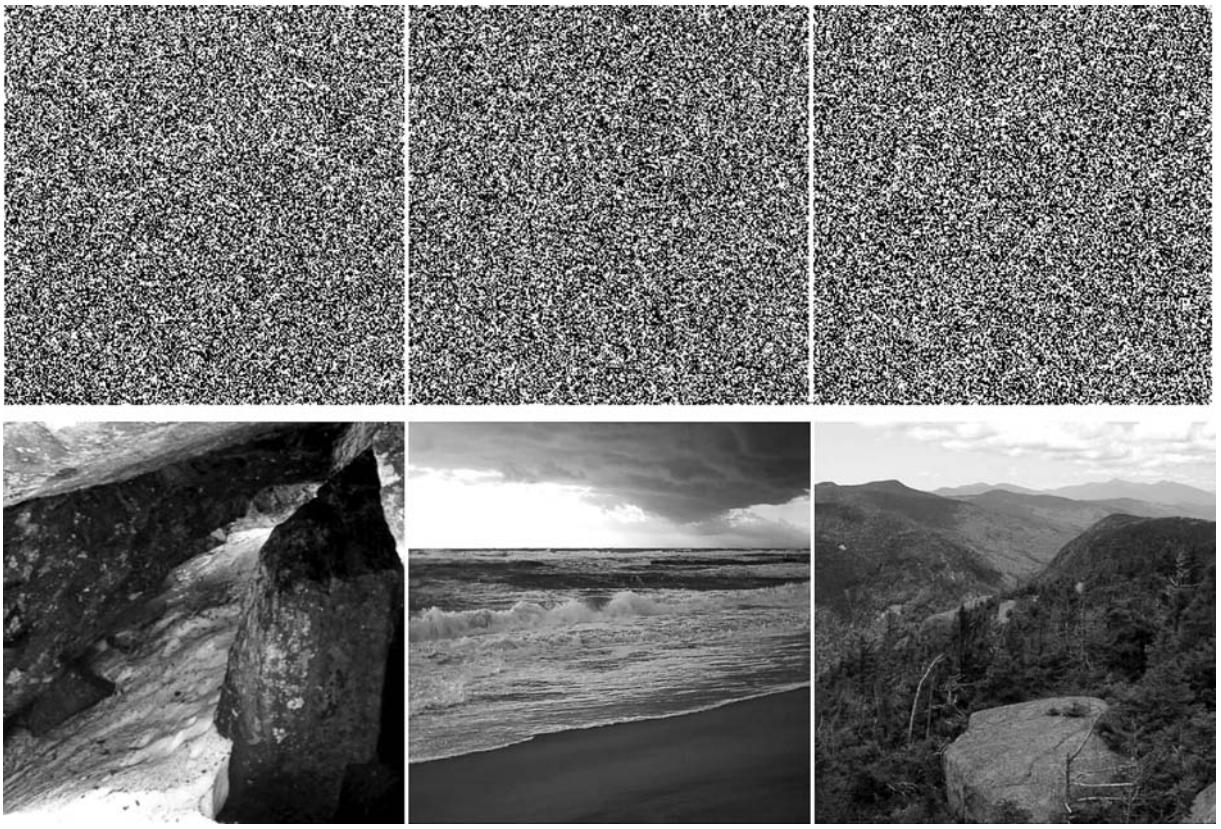

Figure 1 A process that chooses pixel intensity randomly will produce white noise images (top). Images resembling natural scenes (bottom), would almost never occur in such a process.

Defining Efficiency

For those who have proposed theories of optimality in early visual coding, much of the discussion has centered on the metric of efficiency. The majority of the work in the field focuses on what we will call representational efficiency. Such studies employ the tools of information theory, and they have explored the properties of neurons that are involved in representing the image. These studies have focused on issues relating to correlations and statistical independence in the firing rates of neurons. Much of this work has involved neural networks and computational models of visual areas. A second line of research has focused on what we call metabolic efficiency. Several influential reports have described investigations of the metabolic costs of generating spikes, and others have argued that constraints that minimize wiring are important for explaining known neuronal properties. In this article, we propose a third form of efficiency that we call learning efficiency. We argue that an important consideration for any sensory system is the challenge of learning about the relative probability of events in the world based on a handful of samples. To the extent that the visual system is optimally efficient at carrying out its multitude of tasks, these efficiency rules are likely to each contribute significantly to a description of why the visual system is designed as it is. Visual systems may approach optimality (in the engineering sense of the word) across each of these three dimensions, but as we will argue, one of these dimensions – learning efficiency – is one that engineers rarely consider. These dimensions are not necessarily orthogonal to one another, nor do they currently all have well-defined units of measurement. These dimensions are mere sketches of the terrain over which the human visual system appears to have been optimized through evolution. Furthermore, we believe these dimensions can find application in other modalities as well, though here we focus on the visual system. Together, these efficiency dimensions may also suggest ways to design efficient artificial visual systems.

Representational Efficiency

Marcelja, in 1980, was the first to propose that neurons found in the primary visual cortex (V1) show a number of similarities to the mathematical functions

described by Gabor's 1946 theory of communication. Later studies (by Field and Tolhurst, in 1986; and Jones and Palmer, in 1987) confirmed that Marcelja's Gaussian-modulated sinusoid model (i.e., the Gabor function model) provided a good first-order approximation to the receptive field properties of these neurons. Some of the first computational models of the early visual system (e.g., by Watson, in 1983; and Daugman, in 1985) demonstrated how an array of neurons with these properties can represent a visual image. However, this work left unresolved the important question of 'why' such a solution might evolve.

One approach suggests that an understanding of the 'why' question requires consideration of the environment in which the system functions. This approach suggests that theories should be guided by an understanding of the statistics of natural scenes. Early television researchers, and later, in 1987, Field and Burton and Moorhead, found that the Fourier spatial frequency spectra of natural scenes typically fall off as $1/f^k$, where f is spatial frequency and k is approximately 1.2. This regularity expresses the same redundancy as the autocorrelation function of scenes, which measures how similar neighboring points are in terms of luminance. The $1/f$ structure follows from two properties: (1) neighboring points are correlated and (2) the images are roughly scale invariant. A number of studies have shown that there are a wide variety of forms of redundancy found in natural scenes that go beyond the pairwise correlations and spatial frequency spectra (see later). However, pairwise correlations have been the source of a number of theories regarding efficient coding and many researchers imply that these are the most relevant statistics, so we begin with a discussion of such correlations.

Correlation and Decorrelation

When the notion of stimulus redundancy is considered, most analyses typically refer to the pairwise correlations in the data. Many of the earliest theories of efficient sensory coding developed from the notion that if the neural responses to two stimuli are correlated, then an efficient system should strive to represent the data with reduced correlation. This is certainly one important form of redundancy. In recent work, Chandler and Field argue that the pairwise correlations (as represented by the power spectra) account for approximately 40% of to the total redundancy in natural scenes (for 8×8 natural scene patches).

In theory, any representation with significant correlations implies that most of the signal lies in a subspace within the larger space of possible representations. By choosing a representation that codes for only that subspace, it is possible to represent the data with significantly reduced dimensionality (e.g., a reduced number of neurons, or a smaller dynamic range of responses). Srinivasan, Laughlin, and Dubs showed that the center-surround structure of fly large monopolar cells is well matched to the correlations in a collection of scenes they considered relevant to the fly. Given a certain level of noise, they found that a weighted, linear sum over space transmits the greatest amount of information about their collection of natural inputs. Atick and Redlich continued this line of inquiry, arguing that the amplitude spectra of natural scenes are effectively flattened once they reach retinal outputs. The argument is that the roughly linear rise in sensitivity with increasing spatial frequency is inversely related to the $1/f$-distributed fall-off in natural scene spectra, thus allowing neighboring neurons to be uncorrelated (i.e., have a flat spectrum). Atick and Redlich argued that this decorrelation is reflected in a flattening of the spatial frequency response of retinal ganglion cells. However, this argument is dependent on appropriate spacing of retinal neurons. Moreover, the current evidence suggests that neighboring neurons are significantly correlated in the presence of natural scenes.

We have recently argued that an independent goal of retinal coding is to achieve compression while maintaining equal response magnitude across the array of neurons of different sizes. This approach achieves a form of response spectrum flattening but it is quite independent of whether the neurons are tuned and spaced ideally to achieve decorrelation. This work suggests that, in order to explain the details of neural response properties, we must acknowledge that decorrelation in the retina and LGN is far from complete and that efficient coding schemes must consider additional constraints.

Optimal Information Transfer

Another strategy attempts to utilize each level of response magnitude with equal frequency. To this end, Laughlin compared the distribution of local contrasts in the blowfly environment with contrast responses in large monopolar cells in the creature's eye. The intensity-response function measured for these cells is well matched to a distribution of natural contrasts, such that the range of possible responses approximates the information-theoretic ideal. In other words, contrast responses in the fly are distributed across the range of environmental contrasts in such a way as to minimize the number of 'code words' necessary for transmitting information about the entire range of contrasts at a given fidelity. Later studies have shown that adaptation of visual neurons with respect to local luminance,

orientation, and contrast can also be addressed using this maximum information transfer (infomax) approach. A more elaborate extension of infomax models combines feedback of stored predictors from higher cortical areas with typical natural inputs. These schemes optimize over families of representations that minimize error between the input and a top-down representation. Feedback representation optimization models of this sort effectively produce a system whereby primary visual areas reduce their response as higher areas provide better descriptions of input 'content,' in line with recent imaging findings. It should be noted that in many infomax models, optimal redundancy reduction (decorrelation) and maximum information transfer strategies are equivalent.

Beyond Correlations: Sparseness and Independence

The pairwise correlations found in natural scenes represent only one form of redundancy. Figure 2 shows two images – a natural scene and noise – with similar $1/f$ structure and therefore similar pairwise correlations. There are a number of ways to describe the differences between these two images. For example, they differ in their phase spectra. But the two images can also be described in terms of differences in their sparse structure. For the $1/f$ noise image, all linear representations produce response distributions that are Gaussian. However, for the natural scene, some projections of the data are non-Gaussian. That is, when the appropriate array of linear filters is used to represent a natural scene, the histogram of activity will be a non-Gaussian histogram.

As noted by Field, a non-Gaussian histogram implies low entropy in the first-order responses and relatively high entropy in the higher order relationships between the filters for a linear transform (i.e., more independent). In other words, a system that produces maximally non-Gaussian histograms produces a representation wherein the neurons are maximally independent. This is the basic idea behind sparse coding algorithms and independent components analysis (ICA).

In a code with maximal independence, the firing of each neuron provides maximal unique information (i.e., the sharing of information with other neurons has been minimized). If the data consist of an array of relatively rare, sparse events, then matching the neurons to those events will produce activity that is sparse. The definitions of sparseness have varied in the literature. In general, 'sparse' implies a relatively high probability of no activity across the population, and some proportion of relatively active neurons. In the computational literature, in which neurons are often modeled as linear operators, the kurtosis (the fourth statistical moment) of the response histogram can be used to describe relative sparseness. Other metrics, such as the sparseness index, have proved more useful for spike trains.

Field demonstrated that arrays of linear neurons with properties like those found in primary visual cortex appear to maximize the sparse response to natural scenes. Olshausen and Field further demonstrated that a neural network that attempts to represent natural scenes and maximize sparseness will produce an array of neurons with spatial properties like those found in cortical simple cells. That is, a system that is forced to produce a faithful representation of the input using only a handful of neurons (each firing near its maximum response when it is active) gives simple-cell-like receptive fields. This suggests that at least at the level of primary visual cortex, visual system representations show evidence of being efficiently matched to the sparseness of natural scenes. Similar results have been found for spatiochromatic stimuli and spatiotemporal patterns.

It must be emphasized that sparse outputs of these networks result from the sparse structure of the data. It would be relatively simple to produce a

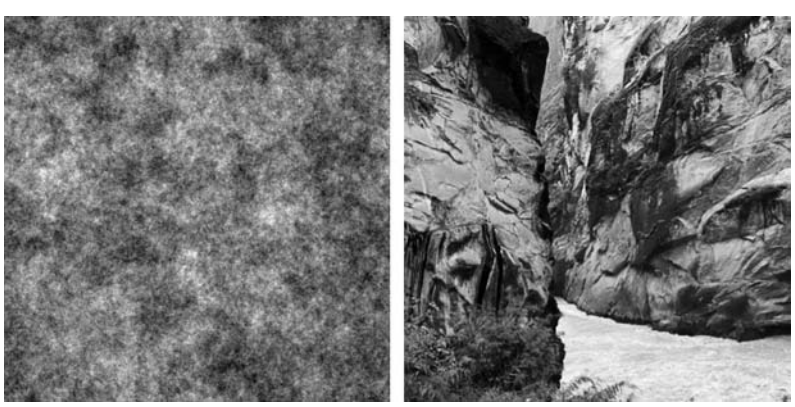

Figure 2 Noise with a spatial frequency amplitude spectrum like that of natural scenes (left) has the same pairwise correlations as does the natural scene (right), but lacks other statistical regularities of scenes.

nonlinearity that forces a sparse output independent of the input, one that would not be an efficient coding strategy. The networks described earlier expressly search for the sparse structure that exists in the data. Moreover, recordings of primate visual neurons in response to natural scenes have also been shown to produce a sparse output. Visual neurons show sparse responses both in early stages of cortical processing, such as V1, and in higher levels, such as inferotemporal cortex. Highly sparse firing is a widespread phenomenon in a variety of brain areas and species, including monkey association areas, rabbit motor areas, rat somatosensory areas, rat auditory areas, rat hippocampus, and centers believed to be involved in bird song generation.

However, the sparse response of these neurons is not direct evidence that these neurons are efficiently representing the environment. All of these neurons are nonlinear, and higher level neurons are very nonlinear. A proof that these codes are efficient would require a clear understanding of how the array of neurons represents the input image. We are not at this level of understanding – and we may never be. Most visual neurons beyond the retina can be modulated by higher levels of processing. Although a number of studies have implied that a decent theory of primary visual cortex is close at hand, Olshausen and Field have argued that we are still a long way from such a theory. Part of the argument comes from the work of the Gallant lab. This work has been an effort to measure the responses of primate visual neurons to a set of natural stimuli and to predict from these responses how each neuron would respond to an arbitrary natural scene. Gallant and colleagues have found that even in V1, typically less than 50% of the response variance to new stimuli can be predicted. In higher levels of the visual system, that prediction accuracy is further reduced. In addition, as noted by Olshausen and Field in 2005, recording studies select only a small portion of the neurons in any given area, and researchers typically record only from large pyramidal neurons that produce large spikes. Other neuron types (e.g., granular neurons that produce smaller spikes) are often not surveyed in these studies, which has led to the interesting suggestion by Shoham and colleagues that there exists a neural 'dark matter' problem.

Optimality with Nonlinear Systems

Despite the difficulties in providing a clear metric for defining efficiency in nonlinear systems, a number of efforts have been made to generalize efficiency arguments to account for the known nonlinear properties of visual neurons. One particularly fruitful approach involves a technique called slow feature analysis. According to this model, a stimulus moving across an image results in quick changes to the neurons responding to the input (for a linear array of neurons). If we consider the histogram of neural activity integrated over some time period, we would find that the activity is less sparse than in a single time frame. However, the changes in images are not random and are often caused by objects or backgrounds showing consistent movement. Slow feature analysis attempts to take into account this redundancy in the movements of features. The technique attempts to find nonlinear solutions that are capable of describing the moving images with a relatively consistent set of neurons, despite the changes that are occurring. Slow feature analysis has been demonstrated to produce nonlinear behavior like that shown with complex cells. It has been argued that current implementations of slow feature analysis are mathematically equivalent to a form of spatiotemporal ICA. Given that object identities are relatively invariant over time, there is hope that these lines of work will eventually be capable of producing object-level representations like those found in inferotemporal cortex. However, we have not yet reached that point.

There exist a variety of techniques that come under the general heading of nonlinear ICA. Although this area of research is too vast to be reviewed here, it should be noted that many of these studies have generated some of the nonlinear properties of visual neurons. However, since we do not yet have a complete model of the nonlinearities in these neurons, we cannot yet argue that such techniques are capable of accounting for the full array of neural properties in any region of visual cortex.

Without a clear account of how information is represented in any given area, we are left without any kind of proof that the information is represented efficiently. Linear models that optimize sparseness and independence do produce simulated neurons with many properties like those found in V1. This certainly supports the notion that V1 is directed toward an efficient code for natural scenes. But as we will argue in the next sections, there are both metabolic and learning costs for high sparseness and therefore a more complete understanding of efficient coding requires examining factors beyond those considered in proposals of representational efficiency. Both the design of artificial visual systems and the development of algorithms for representing natural scenes will likely benefit from our understanding of a larger range of evolutionary constraints, like those described in the following sections. These additional biological constraints require a broadening of current notions of efficient design beyond representational

efficiency. We may someday find that the best artificial systems look remarkably like systems we find in biological systems, and it is likely that a full account of biological systems will require an understanding of these biological requirements.

Metabolic Efficiency

Consideration of representational efficiency as discussed in the preceding sections leads to important insights into the design of visual systems. It provides a number of metrics of efficiency that are generally independent of the limitations of energy or of neural hardware. In this section, we consider attempts to explain the properties of neurons from a consideration of metabolic constraints. We discuss two broad hypotheses of efficient neural design: spike efficiency and neural wiring optimization.

Spike Efficiency

Any information-processing strategy in the nervous system will incur metabolic costs. If we accept that information is primarily transmitted through the use of spikes, then a relevant question is whether one of the constraints on information processing is the number of spikes. Two studies have made detailed efforts to estimate the metabolic cost of spikes, and both have come to the conclusion that the high cost of spikes indeed results in an important constraint on neural processing. Both Attwell and Laughlin, in 2001, and Lennie, in 2003, found that the total available metabolic resources and the cost of a spike limit the firing rate to less than 1 Hz – and probably less than 0.2 Hz. As subsequently noted by Olshausen and Field, most studies have found firing rates significantly higher, suggesting that the cells that are typically recorded have unusually high firing rates. Levy and Baxter earlier argued that when the cost of spiking is considered, maximum information transfer is attained when only 2–16% of the neurons are firing. Lennie estimated that the limited resources imply that at any given moment, only 1/50th of the population of cortical neurons will show high firing rates. Thus, in order to save energy individually and across populations, visual neurons must adopt highly sparse patterns of firing. Therefore, we can argue that from both a representational point of view and a metabolic point of view, sparse firing is efficient.

A related debate centers on the optimal histogram of activity for a real neuron that can only produce a positive firing rate. If the goal of a neuron is to maximize the total information rate while minimizing the mean activity, the most efficient distribution is the exponential distribution. There is certainly a similarity between the histograms of activity of visual neurons (in response to natural signals) and the exponential distribution. However, on close inspection, the exponential model does not appear to fit. For example, Treves et al. found that for neurons in inferotemporal cortex, the exponential model could be rejected in 84% of cases. The lack of any particular successful model led these authors to conclude that there was "no special optimization principle or purpose to the firing distributions found."

Moreover, we question whether the goal should be to optimize the information rate. Consider a case in which there are n possible causes or features in an image and there exist n neurons available to represent those features. One could argue from a representational viewpoint that one should match the neurons to the features. If each feature had a particular response probability, then matching that feature would produce a histogram of activity that matched the histogram of the feature, not one that necessarily matched the histogram corresponding to the optimal compression algorithm. Certainly we do not expect that there is a match between the number of features and the number of neurons. But this example is meant to show that if we expect neurons to provide an explicit representation of their environment, part of the impetus for that explicit representation may be to match the probability distributions of the environment – not simply to maximize information rates.

Minimum Wiring

In addition to metabolic constraints, any neural architecture will be dependent on the anatomical 'wiring' available. This so-called wiring optimization principle dates back to Ramón y Cajal. This type of metabolic efficiency is typically measured by comparing the wiring volume and distribution of real neurons with calculations of the minimum volume needed to connect model cells and cortical areas to one another. Mitchison has argued that stripe patterns in cortex (e.g., ocular dominance columns) are efficient at minimizing the volume of dendritic wiring needed by areas receiving input from two separate sources, compared to alternative arrangements. Durbin and Mitchison found that a highly reduced model of cortical wiring arrangement matched real wiring in certain ways and minimized a model of wiring cost. Along the same lines, Koulakov and Chklovskii found that pinwheel patterns of orientation preference in cortex allow efficient wiring. In a related study by Chklovskii et al., neural wiring density and local signal delay have similarly reached an optimal compromise. Together, these studies suggest that intracortical wiring optimization can help explain why

cortical maps have the topography they do. Cherniak et al. and others have shown that minimal wiring may also help explain why brain areas are placed and interconnected as they are. In an early study using a reduced anatomical model of the *Caenorhabditis elegans* ganglia, Cherniak employed a connection-optimizing algorithm to search for connectivities that minimized total wiring. The actual wiring of the ganglia matched the optimal arrangement from a family of possible orderings.

Some have argued that brain evolution and development impose invariant connectivity ratios, which may tend to enforce optimal wiring. For example, Changizi and Shimojo suggest that across phyla the average number of synapses per neuron scales with the number of neurons per physiologically defined area, and that the number of connections among areas (per area) scales with the number of areas. However, this approach has several limitations. It should be noted that the number of ways that an efficient wiring system could strive to optimize the cost of adding cell volume, increasing cell metabolism, delaying and attenuating the signal, and making projections during development is essentially unlimited. But when considered in concert with representational constraints, minimum wiring arguments can provide insights about efficient receptive field design.

Learning Efficiency

In addition to the fact that neural systems, unlike human-engineered ones, are difficult to assess in terms of their goals and costs, they also differ from traditionally engineered systems because they are produced through development and learning rather than through carefully planned assembly. Indeed, part of the reason why the level of efficiency of a brain is hard to measure is that it becomes specialized to a variety of tasks – often concurrently – during development, the combination of which helps the adult creature reproduce. Therefore, in order to explain why the visual system of the adult has the structure it does, given a proposal of efficient processing of natural scenes, we must consider how visual system representations may depend on efficient learning and development. This efficiency dimension in particular is an example of somewhere our intuition about the brain seeking desperately to conserve energy can lead us astray. Indeed, learning efficiency could be at odds with metabolic cost optimizations, since the proportion of metabolic energy consumption in infant human brains (which are most subject to this constraint) is roughly 3 times that for adults.

In discussions of efficiency in sensory systems, many studies have used various types of neural networks and other learning algorithms to generate early sensory processes. There are certain basic problems that any learning algorithm must overcome. We examine one of these issues – invariance. Then we describe evidence of contributions to learning efficiency from innate, purely learned, and hybrid strategies, which span the well-known nature/nurture debate.

Sparseness and Invariance

Certain constraints are imposed on a system that must learn about its environment from a relatively small number of examples. In essence, efficient learning is dependent on finding a balance between selectivity of neurons for specific features and invariance across examples having features that vary in irrelevant ways (e.g., lighting). This task is further complicated because each example is likely to be seen just once in a lifetime. A second instance is unlikely to be presented with the same lighting, the same position, the same size, and the same orientation as the first. To allow any calculation of the probability of what is likely, it is critical that the system be capable of generalizing to multiple instances of that object.

Sparse coding, in the strict version, does not help with this problem. The typical argument against hierarchical sparse codes is that a system that develops a neuron for every object would require too many object detectors in order to function. This 'grandmother cell' hypothesis has been critiqued elsewhere, but it is worth noting that plausible models of object recognition suggest the need for grandmother cell-like coding for certain tasks (e.g., distinguishing between different faces) and sparse population codes for other tasks (e.g., categorization). That is, object recognition may require a variety of strategies that vary in their degree of sparseness.

It has also been argued that optimizing for sparseness and independence can assist a system in identifying novel inputs and in detecting new relationships. Consider the case in which objects are uniquely represented by just one neuron versus the case in which the object is identified by the relative activity of 100 neurons. In the single-neuron case, to learn that two objects often co-occur requires a relatively simple algorithm (e.g., Hebbian learning). But when identification requires a particular activity profile among many neurons, the learning algorithm would require a far more complex association among the 100 units involved in the representation.

Overcompleteness

An important property of visual representations beyond the optic nerve is that they are highly overcomplete. Visual cortex in humans contains on the order of 1000 times more neurons than do the two optic

nerves. In macaque V1 alone, there are approximately 50 times more output fibers than input fibers. These overcomplete codes must involve significant redundancy. Overcompleteness has been suggested by Barlow as an optimally efficient way "to model the redundancy in images, not necessarily to reduce it." And as Riesenhuber and Poggio pointed out, overcompleteness allows any particular signal to be represented with a higher degree of sparseness than is possible with complete codes, a property which is useful for generalization during learning, as described earlier. In order to achieve efficient learning, one must consider solutions that may not be optimally efficient from a strict representational efficiency point of view. That is, solutions that are representationally efficient (e.g., maximum information transfer arguments) may not help explain how the system achieves efficient learning.

'Hard-Coded' Efficiency

Though every part of the visual system undergoes developmental change, some efficient properties of the adult cells are thought to be innate. Maloney found that the distribution of natural chromaticities could be efficiently and accurately coded using just three wavelength vectors, and that the three cone spectral sensitivities in particular may optimally achieve color constancy based on typical surface spectral reflectances in the natural world. Others have proposed that the distribution of spectral sensitivities in primates is optimally weighted so as to separate nutritious fruit from surrounding foliage, a genetic adaptation that could have aided propagation of trichromatic primates.

Moreover, at the level of the retina, there is a deep conservation of 'virtually all functional and structural features of importance' and of the developmental program in the vertebrate retina according to Finlay et al. This may indicate that visual system development is optimized to allow successful elaboration, robustness, and specialization (what is referred to as 'evolvability') for a range of species that live in a variety of habitats using a single developmental and organizational scheme. It can therefore be seen as an efficiency which is innate, but which is connected to the types of variations present in a changing visual environment.

The development of classical receptive field organization in cortex does not appear to require learning, and relatively little spatial refinement is needed to achieve adult level acuity. One area of debate centers on the question of the necessity of visual experience in order to achieve adult receptive field structures, which presumably contribute to efficient representations in the brain. Recent studies have suggested the prevailing view – which proposes that animals dark-reared during the critical period of development are left with unrefined receptive fields because visual experience had no chance to 'prune' dendritic arbors – could be in need of reevaluation.

Efficient Learning from the Environment

In higher levels of cortex, learning leads to a marked decrease in mean response in prefrontal and inferotemporal cortex, which has been linked to the detection of novel stimuli. In particular, learning based on natural stimulus matching in monkeys led to systematically lower mean responses to learned natural stimuli in prefrontal cortex, and these responses did not vary with the addition of noise. At lower levels of cortex, learning has a minor effect on classical response properties such as orientation tuning and receptive field size in V1 cells). Learning also has a relatively small effect on responses in V4 to noise-degraded natural stimuli. Therefore, efficient learning strategies may be an important principle, especially in higher levels of visual cortex.

Hybrid Strategies: Efficient Innate Learning

Finally, there is evidence that spontaneous activity in the visual system, in a time period before a creature's eyes open, may help refine the organization of the visual system. There is good evidence that suppressing spontaneous activity in the retina (so-called retinal waves) can affect the refinement of retinal projections to the LGN. As some have noted, spontaneous, patterned activity is observed in cortex, hippocampus, thalamus, retina, and spinal cord in a variety of creatures. Simple programs of the sort that could theoretically produce spontaneous, patterned activity could well be coded into genes, and the 'running' of these programs could produce the necessary statistical properties that a developing visual system needs to operate at a basic level once the eyes open. We refer to this as notion as 'innate learning,' since it employs elements of both learning and innateness. These innate learning strategies are efficient in the sense that in the absence of external stimuli, they require far fewer genetic instructions in order to develop proper response properties, compared to full genetic specificity.

However, innate learning in the form of spontaneous retinal activity is not required for the initial formation of ocular dominance columns or of orientation selectivity. There is strong evidence that retinotopic maps and ocular dominance patterns fail to refine properly in animals reared in altered visual environments (e.g., strobe environments, binocular deprivation, environments with altered spatial statistics). But it should be noted, especially considering the brief time during which spontaneous activity has been so far studied,

as stated by Carrasco and colleagues, that "the extent to which spontaneous and visually driven activity contribute to the development and maintenance of stimulus specificity is unclear."

In short, learning efficiency may require solutions that would not be predicted by representational efficiency arguments or metabolic efficiency arguments alone. The difficulty of learning invariant properties of objects based on a handful of presentations is one reason why this is so. The degree to which the visual system is efficiently engineered to learn is dependent on innate properties, on learned associations, and on patterns of spontaneous activity, the combination of which may approach 'optimal' efficiency in concert with metabolic and representational constraints.

Conclusion

Visual systems have been argued to be operating near theoretical limits of optimality – not simply parsimoniously – for certain tasks. But given the profusion of ways that brain structure, cell development and function, and neural representation are believed to be optimal in some way, we argue that visual system design represents a compromise among these many demands for efficiency. We have attempted to elucidate a variety of ways in which visual systems can strive for efficiency. No single quantity (e.g., energy consumption) applies strict limits to brain design (except in the absolute sense), nor is the system optimized to perform a single, 'optimizable' task. There is much to be learned from investigating the contribution from a variety of constraints on efficiency.

See also: Activity in Visual Development; Animal Models of Inherited Retinal Degenerations; Fovea: Primate; Information Coding; Retina: An Overview; Retinal Pharmacology: Inner Retinal Layers; Retinal Color Mechanisms; Retinal Development: An Overview; Retinal Development: Cell Type Specification; Retinal Models; Visual Cortical Models of Orientation Tuning.

Further Reading

Atick JJ and Redlich AN (1992) What does the retina know about natural scenes? *Neural Computation* 4: 196–210.

Attwell D and Laughlin SB (2001) An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism* 21: 1133–1145.

Barlow HB (2001) Redundancy reduction revisited. *Network: Computer Neural Systems* 12: 241–253.

Carrasco MM, Razak KA, and Pallas SL (2005) Visual experience is necessary for maintenance but not development of refined retinotopic maps in superior colliculus. *Journal of Neurophysiology* 94: 1962–1970.

Chandler DM and Field DJ (2007) Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America* 24: 922–941.

Chklovskii DB and Koulakov AA (2004) Maps in the brain: What can we learn from them? *Annual Review of Neuroscience* 27: 369–392.

Daugman JG (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America* 2: 1160–1169.

David SV, Vinje WE, and Gallant JL (2004) Natural stimulus statistics alter the receptive field structure of V1 neurons. *Journal of Neuroscience* 24: 6991–7006.

Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A* 4: 2379–2394.

Field DJ (1993) Scale-invariance and self-similar 'wavelet' transforms: An analysis of natural scenes and mammalian visual systems. In: Farge M, Hunt JCR, and Vassilicos JC (eds.) *Wavelets, Fractals and Fourier Transforms: New Developments and New Applications*, pp. 151–193. Oxford: Oxford University Press.

Field DJ (1994) What is the goal of sensory coding? *Neural Computation* 6: 559–601.

Finlay BL, Silveira LCL, and Reichenbach A (2005) Comparative aspects of visual system development. In: Kremers J (ed.) *The Structure, Function and Evolution of the Primate Visual System*, pp. 37–72. New York: Wiley.

Graham DJ, Chandler DM, and Field DJ (2006) Can the theory of 'whitening' explain the center-surround properties of retinal ganglion cell receptive fields? *Vision Research* 46: 2901–2913.

Jones JP and Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology* 58: 1233–1258.

Laughlin SB, de Ruyter van Steveninck RR, and Anderson JC (1998) The metabolic cost of neural information. *Nature Neuroscience* 1: 36–41.

Lennie P (2003) The cost of cortical computation. *Currents in Biology* 13: 493–497.

Maloney LT and Wandell BA (1986) Color constancy: A method for recovering surface spectral reflectance. *Journal of the Optical Society of America A* 3: 29–33.

Mitchison G (1991) Neuronal branching patterns and the economy of cortical wiring. *Proceedings of the Royal Society of London – Series B* 245: 151–158.

Olshausen BA and Field DJ (1996) Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature* 381: 607–609.

Olshausen BA and Field DJ (2005) How close are we to understanding V1? *Neural computation* 17: 1665–1699.

Osorio D and Vorobyev M (1996) Colour vision as an adaptation to frugivory in primates. *Proceedings of the Royal Society of London – Series B* 263: 593–599.

Shoham S, O'Connor DH, and Segev R (2006) How silent is the brain: Is there a 'dark matter' problem in neuroscience? *Journal of Comparative Physiology A* 192: 777–784.

Srinivasan MV, Laughlin SB, and Dubs A (1982) Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London – Series B: Biological Sciences* 216: 427–459.

Treves A, Panzeri S, Rolls ET, et al. (1999) Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. *Neural Computation* 11: 601–631.

Nausea and Vomiting

P L R Andrews, University of New Mexico, Albuquerque, NM, USA

© 2009 Published by Elsevier Ltd.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

Biological Significance

Although nausea (unpleasant sensation associated with the urge to vomit) and vomiting (forceful oral expulsion of gastric contents) may be most frequently encountered in a clinical context as a symptom of disease (e.g., raised intracranial pressure, migraine, hypertrophic pyloric stenosis, labyrinthitis, renal failure, Addison's disease), as a symptom of viral (e.g., Norwalk virus) or bacterial infection (e.g., *Staphylococcal enterotoxin* B, *Salmonella typhimurium,* meningococcal meningitis), as a side effect of drug therapy (e.g., antitumor chemotherapy, digoxin, morphine, oral L-DOPA, theophylline, oral contraceptives, some antibiotics (e.g., erythromycin), as accidental ingestion of plant (e.g., Ackee toxin Jamaican vomiting sickness), nicotine, veratridine, tremetone (milk sickness) or animal (e.g., batrachotoxin, ciguatoxins) toxins, or as sequelae to treatment (e.g., radiotherapy, postoperative nausea and vomiting (PONV)), it is important to remember that both are components of a biologically important, hierarchically organized system presumed to have evolved to protect the body against the accidental ingestion of toxins. Usually, every effort is made to avoid ingestion of hazardous agents by taking evasive action based on taste, smell, and visual cues, but not all toxins taste bitter or have a warning coloration and so may be inadvertently swallowed. In addition, toxins may be elaborated postingestion. Therefore, a mechanism is clearly required by which the toxin can be detected and ejected once it is in the body; hence, detection systems are located at both pre- (e.g., gut mucosa) and post- (e.g., circulation) absorptive sites.

A curious feature of vomiting is why it is evoked by such a wide range of stimuli. In contrast to vomiting evoked by ingested toxins, which is clearly appropriate in that it leads to expulsion, that induced by most clinical and some natural stimuli, such as snake bite, motion, and pregnancy (especially in the first trimester, when it is often called 'morning sickness'), appears 'inappropriate' as the stimulus remains. However, for the reflex to be evoked, all that is required is for the stimulus to activate one of the triggering systems either directly or indirectly. It has been argued that motion sickness is also part of the defensive system against ingested toxins. Morning sickness is argued to protect both the mother and embryo against potentially hazardous foods (e.g., teratogens) at a time when both are vulnerable, the mother because of immunosuppression and the fetus because it is at a key period of organ development.

Vomiting occurs in representative fish (e.g., shark, skate, trout, tuna), amphibia (e.g., frog, salamander), birds (e.g., pigeon, petrel), reptiles (e.g., snake, crocodile), and mammals (e.g., musk shrew, cat, dog, ferret, pig, cynomologous monkey, marmoset, sperm whale). Reflexes for the oral ejection of upper gut contents have been described in some invertebrates (e.g., arthropods such as grasshoppers, sea anemone, *Pleurobranch* mollusk). While the reflex may be triggered by ingested toxins, many species appear to use vomiting or regurgitation for other functions such as voiding indigestible food residues such as squid beaks (e.g., sharks, sperm whales), bones (owls), or trichobezoars (e.g., crocodile, cat); for feeding the young or mates (e.g., albatross, African wild dog); and for defense (e.g., petrel). (In mammals this can be distinguished from vomiting by the patterns of neuromuscular activity, particularly the diaphragm and abdominal muscles, but as they are not present in all species, such a distinction may be semantic.)

Despite its potential biological significance, rodents (e.g., rat, mouse) and lagomorphs (e.g., rabbits) do not have this reflex. While the reason is not known, it is thought that sensitivity of the taste system and the ability to generate conditioned taste (food) aversions (CTAs) in these species enable them to avoid toxin ingestion, but they may lack some of the brain stem pathways required for the vomiting reflex (perhaps in the nucleus tractus solitarius (NTS); see subsequent discussion), although both vomiting and nonvomiting species have an area postrema (AP). In addition, these animals may have anatomical features (e.g., a long narrow abdominal esophagus) or physiological differences (e.g., inability to shorten the esophagus sufficiently) that would make ejection of gastric contents difficult, or they have behavioral responses (e.g., pica, the ingestion of nonnutritive substances such as clay) or biochemical mechanisms to allow them to deal with ingested toxins. Rodents without a vomiting reflex do have a gag reflex elicited from the pharynx. The presence of the reflex does not depend upon the size of the mammal, as it is present in both a shrew (*Sorex unguiculatus*) weighing approximately 10 g and a sperm whale. Research studies of the emetic reflex have used nonhuman primates (e.g., cynomologous monkey,

marmoset) and cats and dogs, but recently the ferret (a carnivore), piglet (an artiodactyl), and *Suncus murinus* (the house musk shrew, an insectivore) are being used increasingly.

Motor Components

Irrespective of how the vomiting reflex is triggered in mammals, the motor components are the same, and in this respect the reflex may be viewed as an example of a stereotypical motor program.

Preejection or Prodromal Phase

Prior to the actual expulsion of upper gut contents, a considerable number of motor changes mediated by the autonomic nervous system have taken place, particularly in the gastrointestinal tract (Figure 1). These include relaxation of the proximal stomach to confine the food together with any toxin to the stomach from where it can be ejected. This is mediated predominantly by vagal efferent activation of myenteric nonadrenergic, noncholinergic inhibitory nerves, probably using nitric oxide (NO) and vasoactive intestinal peptide (VIP) as transmitters. Another motor change is a retropulsive giant contraction sweeping from the mid-small intestine to the stomach, which is proposed to return any previously emptied and contaminated gastric contents to the stomach for ejection. It has also been proposed that this contraction may propel alkaline pancreatic and intestinal fluid into the stomach to buffer gastric contents before ejection. The contraction is vagally mediated, mainly involving activation of atropine-sensitive cholinergic nerves. When expulsion occurs, these gastrointestinal changes always precede it, but they may also occur in the absence of vomiting. Tachycardia, cutaneous vasoconstriction, and sweating also occur in the prodromal phase and are probably mediated by the sympathetic nerves. Increases in the plasma levels of adrenaline, adrenocorticotropic hormone (ACTH), growth hormone, prolactin, and vasopressin (ADH) have been reported during the prodromal phase. Some changes are likely to be a reflection of general stress (e.g., adrenaline), whereas others (e.g., vasopressin) may be specific to nausea.

Ejection Phase: Retching and Vomiting

The expulsion of gastric contents (vomiting) is brought about by the coordinated rhythmic contraction of the diaphragm, external intercostals (inspiratory), and abdominal muscles (rectus abdominus, external oblique) compressing the stomach and ejecting the contents through the mouth via the esophagus (see Figure 1). The propulsive pressure exceeds 100 mmHg. Some evidence also suggests that during expulsion longitudinal shortening of the esophagus occurs, and there may be an esophagopharyngeal retrograde contraction that could contribute to the oral expulsion of gastric contents. Retching, which usually precedes vomiting and in contrast to vomiting occurs as multiple events, differs in that gastric contents are not ejected orally. This is thought to be because during retching the entire diaphragm contracts rhythmically under the influence of the

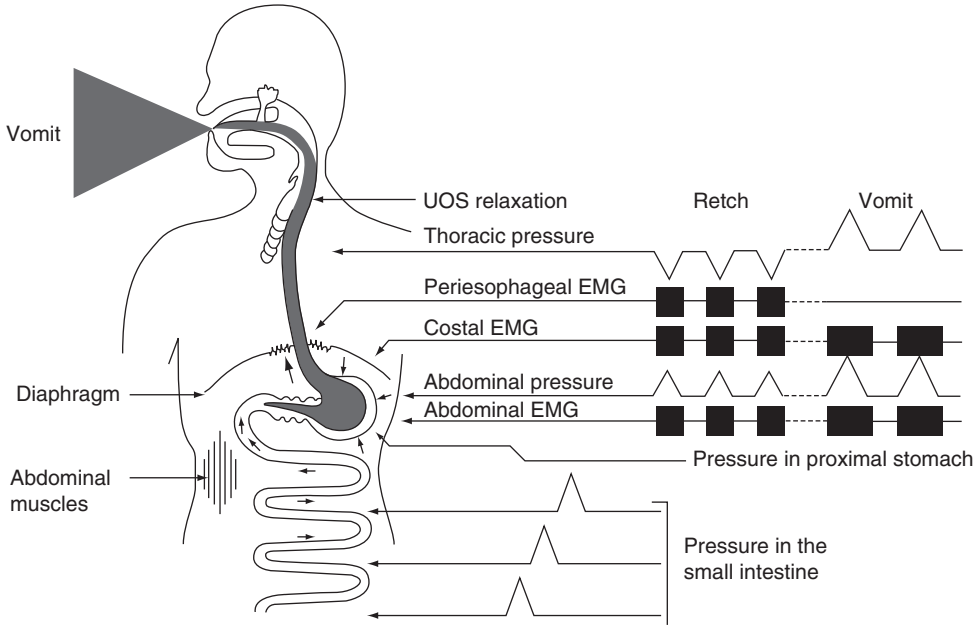

Figure 1 A diagram summarizing the major visceral and somatic motor changes comprising the vomiting reflex. EMG, electromyogram; UOS, upper esophageal sphincter.

phrenic nerve, whereas during vomiting the bulk of the diaphragm contracts but the periesophageal region (crural fibers) is inactive to permit ejection of gastric contents. The function of retching is not known, although it is suggested that it helps to overcome the antireflux barrier present in the region of the gastro-esophageal junction. The factors that regulate the number of retches in a burst of retching or determine whether a vomit follows retching are not known, although studies in the ferret showed an inverse relationship between gastric volume and the number of retches prior to a vomit. Blood gas levels have also been implicated in this transition.

In nonmammals other neuromuscular mechanisms may be used for ejection of gastric contents. These include retrograde contraction of the stomach (e.g., dogfish) and oral eversion of the stomach (toad, skate); the mechanism by which the latter occurs is unclear but may involve relaxation of the stomach and an increased intra-abdominal pressure.

Coordination of the Motor Components

The motor systems that comprise the vomiting reflex are almost all employed in nonemetic reflexes and hence the magnitude, combination, and sequence of activation of these outputs make the reflex unique. The individual brain stem nuclei involved in the control of the autonomic and somatic motor components of emesis have largely been identified. For example, the dorsal motor vagal nucleus and nucleus ambiguus provide the vagal output to the gut and heart; the presympathetic neurons in the ventrolateral medulla regulate the drive to the preganglionic sympathetic neurons in the spinal cord that supply the cutaneous vasculature; and the ventral respiratory group of neurons (e.g., nucleus retrombigualis, Bötzinger complex) provides the drive to the diaphragm and intercostal muscles involved in the somatic motor components of emesis. However, it is still not known exactly how the various components are coordinated. In particular, the visceral and somatic motor nuclei must interact to ensure that the gastrointestinal and other preejection changes precede contraction of the diaphragm and abdominal muscles. Whether there is a single nucleus that corresponds to the old concept of a 'vomiting center' is still hotly disputed, but the parvicellular reticular formation, the Bötzinger complex (part of the ventral respiratory group of neurons), and the NTS (the major integrative nucleus for visceral information and location of the dorsal respiratory group of neurons) have all been proposed to fulfill this function to varying degrees.

Triggering the Emetic Reflex

How the multitude of stimuli that can evoke vomiting activate the central pathways outlined previously is not known, but three major pathways have been identified, and they provide a framework within which some emetic stimuli may be discussed (Figure 2).

Gastrointestinal Afferents

Vomiting can be evoked by overdistention of the upper gut (e.g., overeating or motility disorder) and by the presence of chemical irritants (e.g., hypertonic solutions or copper sulphate) or toxins (e.g., plant alkaloids, such as emetine, a constituent of the clinical emetic ipecacuanha, or bacterial toxins) in food. These stimuli activate mechanoreceptors or mucosal chemoreceptors with afferents principally in the vagus nerve, which projects to the medulla (NTS and AP), to trigger the emetic motor pathways. Biologically, this is probably the most important route for activation of emesis.

Studies in several species (dog, monkey, ferret, and *Suncus*) have implicated mucosal vagal afferents as the predominant pathway in the acute phase of vomiting induced by systemically administered antitumor agents (e.g., cisplatin) and whole-body radiation. It is proposed that these stimuli induce the formation of free radicals in the enterochromaffin cells of the upper gastrointestinal tract. This triggers a calcium-dependent release of 5-hydroxytryptamine (and probably other neuroactive substances, e.g., substance P), which activates 5-HT_3 receptors (see subsequent discussion) located on the vagal afferents terminating in close proximity.

Vagal afferents from the heart can also evoke nausea and vomiting and are responsible for these symptoms during infarction.

Area Postrema

The AP is a circumventricular organ located at the caudal extremity of the fourth ventricle outside the blood–brain and cerebrospinal fluid (CSF)–brain barriers and is therefore ideally suited to its proposed function as a chemoreceptor. Ablation studies have implicated the AP in the mediation of emesis by a wide variety of systemic agents (e.g., cytotoxic drugs and the dopamine D_2 agonist apomorphine) and led to it being called the chemoreceptor trigger zone (CTZ) for emesis. Although there is no dispute that suitable activation can trigger emesis and perhaps influence the sensitivity of other emetic inputs, it is probably more accurate to regard the AP as having a wider chemoreceptive role with an involvement in conditioned taste aversion, salt balance, and food intake. Because of the close proximity of the AP to

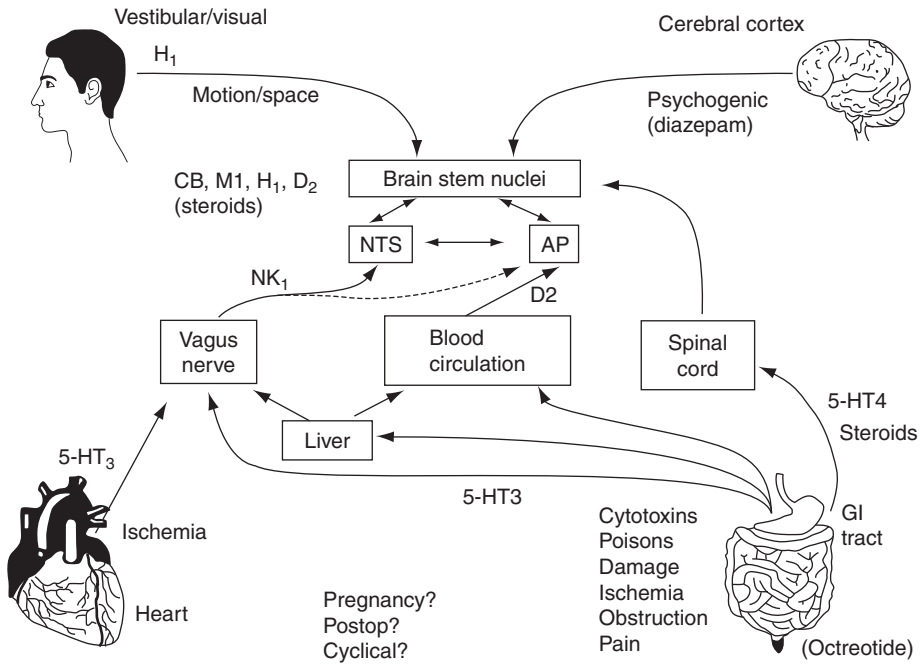

Figure 2 The different operative pathways and pharmacologies used by major emetic stimuli. NTS, nucleus tractus solitarius; AP, area postrema; CB, cannabinoid; NK_1, neurokinin$_1$ receptor; D_2, dopamine D_2 receptor; M_1, muscarinic M_1 receptor; H_1, H_1 receptor antagonist. From Sanger GJ and Andrews PLR (2001) Emesis. In: Farthing MJG and Ballinger AB (eds.) *Drug Therapy for Gastrointestinal and Liver Diseases*, pp. 45–61. London: Martin Dunitz Ltd.

the central emetic circuitry and sites of termination of abdominal vagal afferents (particularly the NTS), collateral damage may be difficult to avoid and thus could confuse the interpretation of lesion studies. Teleologically, it may be argued that a toxin detector in the central nervous system (CNS) would be of little use because after the toxin has been absorbed it cannot be expelled by vomiting. Hence, the emetic role of the AP is probably a reflection of its general chemoreceptive role and connectivity, a role supported by its presence in nonvomiting species.

Vestibular system

In 1881, Irwin observed that the vestibular system (semicircular canals and otolith organs) was critical for motion-induced emesis. This and other studies led to the idea that overstimulation was the major trigger, but this did not account for all the experimental observations. Current hypotheses are based on the idea of sensory conflict or rearrangement. Some view this as a disparity between different sets of incoming information (e.g., from vision and the vestibular system), a sensory conflict, whereas others tend more toward the view that the conflict lies not between modalities but between what the body expects to happen in a given situation and what actually occurs: a sensory motor or efference–copy conflict. Psychophysical studies support both hypotheses. Conclusive neurophysiological studies are awaited, although there are several regions where visual and vestibular information converge, for example, second-order vestibular-nucleus neurons and the flocculus, nodulus, and uvular regions of the cerebellum. Ablation of the latter renders the dog refractory to motion-induced emesis. The exact pathways by which the medullary emetic circuitry is activated are not known, but both the lateral medullary reticular formation and the NTS receive vestibular inputs, and individual neurons in the NTS have been shown to receive vestibular and gastrointestinal inputs.

Sensitivity to motion sickness is a predictor of the emetic response to other emetic stimuli such as pregnancy sickness, postoperative nausea and vomiting, and anticancer chemotherapy-induced emesis. This suggests that the vestibular system may have a 'modulatory' influence on the brain stem emetic pathways, setting their sensitivity. Based on studies of different racial sensitivity to motion stimuli (Asians are more sensitive than Westerners), some evidence suggests genetic determinants of sensitivity.

Nausea

The term nausea is derived from the Greek word for ship and well illustrates man's early experience of sea travel. Although we know that activation of the

pathways described previously, usually at a lower intensity than that required to induce vomiting, can induce nausea, the precise CNS pathways by which this sensation is generated are not known. Magnetic source imaging has implicated the inferior frontal gyrus. The main reason for the relative paucity of knowledge of the mechanisms underlying nausea is that while in animals retching and vomiting are obviously the same as in humans, in animals we can only study what we conclude are the 'behavioral equivalents' of nausea (e.g., conditioned taste aversion, pica) and not the actual sensation experience (cf. pain studies). In the preejection phase in humans, increases in vasopressin (ADH) may be many times maximal antidiuretic levels, and this has led to the suggestion that it may be a useful index of nausea, or even involved in its genesis. As nausea can be induced in patients with diabetes insipidus, it is clear that ADH is not obligatory for its genesis. The involvement of disturbed gastrointestinal motility (e.g., tachyantria) in the genesis of nausea from all causes is not resolved, but in cases in which there is gastric stasis resulting in distension, it is possible that nausea (and vomiting) is caused by visceral afferent activation. Understanding nausea is still a major challenge but, fortunately, most drugs alleviating vomiting also mitigate nausea, although usually at a reduced efficacy.

Biologically, the function of nausea appears to be the induction of a learned aversion to the stimulus responsible (e.g., contaminated food, sea travel), although this does not appear to apply to nausea induced by alcohol or pregnancy.

Anti-emetic Agents

The selection of anti-emetic agents relies on knowledge of the pharmacology of the pathways involved in triggering emesis in the various clinical settings. Some of the better-understood pharmacological therapies are outlined in this section.

Motion sickness is well controlled using centrally acting muscarinic (e.g., hyoscine) or histamine (H_1, e.g., diphenhydramine) receptor antagonists. Neurophysiological studies have shown that these agents modulate activity in the vestibular nuclei. Recent studies using more selective muscarinic receptor antagonists (e.g., zamifenacin) have implicated M_3 receptors in motion-induced emesis in humans, but the site of action is not known.

The severe acute (first 24 h) vomiting that may be induced by cytotoxic drugs (e.g., cisplatin) or radiation therapy for carcinoma is completely controlled in approximately 60% of patients (particularly in combination with steroids, see below) and reduced in most of the remainder using antagonists selective for 5-hydroxytryptamine3 (5-HT_3) receptors (e.g., granisetron, ondansetron, tropisetron, dolasteron) located on the peripheral terminals of abdominal vagal afferents (see previous discussion) and neurons in the nucleus tractus solitarius and AP. These agents are not effective against the delayed phase (2–7 days) of emesis associated with cytotoxic therapy. The anti-emetic efficacy of 5-HT3 receptor antagonists in chemotherapy-induced nausea and vomiting can be enhanced by the coadministration of the steroids dexamethasone and methyprednisolone, but the mechanism of action is not known, although allosteric modulation of the receptors has been suggested in addition to the well-known anti-inflammatory actions of glucocorticoids. Cannabinoids (e.g., nabilone) have been shown to be slightly superior to conventional anti-emetics in chemotherapy-induced emesis. 5-HT_3 receptor antagonists also have demonstrable efficacy in patients with postoperative nausea and vomiting.

When nausea and vomiting are induced by gut motility disorders (e.g., delayed gastric emptying), they can be alleviated by using agents that stimulate motility (e.g., the D_2 receptor antagonist/5-HT_4 receptor agonist metoclopramide) and hence have an indirect anti-emetic effect. However, if the vomiting is due to a physical obstruction (e.g., hypertrophic pyloric stenosis in neonates) then surgery is the appropriate treatment.

Pregnancy sickness, affecting more than 80% of women in the first trimester, is arguably the most common form of emesis. Its etiology is unknown, but it is proposed that the steroids secreted in early pregnancy sensitize the central emetic pathways. Empirical therapy ranges from pyridoxine to diphenhydramine; because of concern about potential adverse effects of drugs on the fetus, treatment of this type of emesis is difficult and arguably is unnecessary because the condition is self-limiting. Rarely (in approximately 1 in 1000 pregnancies) severe emesis continues during pregnancy and is termed hyperemesis gravidarum, but as with the sickness in early pregnancy, the etiology is uncertain. Treatment is essentially supportive, correcting any nutritional, fluid, or ionic imbalances.

While research has focused on pharmacological therapies, complementary therapies, such as ginger, acupressure (P6 point), and acupuncture have been claimed to successfully treat nausea and vomiting induced by motion, pregnancy, chemotherapy, and postoperatively. In anticancer therapy-induced emesis, behavioral interventions such as cognitive behavioral therapy are being investigated together with the role played by the patient's expectations of outcome in influencing the magnitude of the emetic response.

Research into novel pharmacological therapies continues with attempts to identify anti-emetics with

a broad spectrum of action against a variety of stimuli. Preclinical studies demonstrate that 5-HT_{1A} receptor agonists (e.g., 8-OH-DPAT), μ opioid receptor agonists (e.g., fentanyl), and the vanilloid receptor agonist resiniferatoxin have some broad-spectrum anti-emetic activity. Cannabinoid (CB1) receptor agonists also have anti-emetic activity in preclinical models, as do NMDA (LY 233053), non-NMDA receptor antagonists (e.g., NBQX), mixed dopamine D_2/D_3 receptor antagonists, and cyclo-oxygenase inhibitors (meloxicam, indomethacin). Preclinical studies with a number of highly selective and potent $neurokinin_1$ (substance P receptor) receptor antagonists (e.g., vofopitant GR-205171, ezlopitant CJ-11974, aprepitant MK-869, L-758, 298 (water-soluble prodrug of MK-869, CP-99, 994) have shown that they are capable of blocking vomiting induced by cytostatic drugs, radiation, oral copper sulphate, intravenous apomorphine, and motion. This broad spectrum of action suggests a brain stem site of action most likely in the NTS, but other sites cannot be excluded. Clinical studies have been undertaken with representative agents. Efficacy has been demonstrated in limited studies of PONV (GR-205171) and more extensive studies in chemotherapy-induced nausea and vomiting (CJ-11, 974, CP-122, 721, MK-869/L-758, 298), with the agents usually given in combination with dexamethasone plus or minus a 5-HT_3 receptor antagonist, demonstrated improved control of both acute and delayed emesis. The effect on delayed emesis is of particular interest, and the first drug in this class (aprepitant, Emend) was approved by the FDA in April 2003 for the treatment of acute and delayed chemotherapy-induced nausea and vomiting. Further studies are awaited to fully assess the clinical potential of this class of agent.

See also: Autonomic Nervous System: Gastrointestinal Control; Autonomic Nervous System: Central Control of the Gastrointestinal Tract; Motion Sickness; Vertigo; Vestibular System.

Further Reading

Andrews PLR (1999) Postoperative nausea and vomiting. In: Herbert MK, Holzer P, and Rower N (eds.) *Problems of the Gastrointestinal Tract in Anesthesia*, pp. 267–288. Berlin: Springer.

Crampton GH (1990) *Motion and Space Sickness*. Boca Raton, FL: CRC Press.

Diemunsch P and Grelot L (2000) Potential of substance P antagonists as antiemetics. *Drugs* 60: 533–546.

Ernst E and Pittler MH (2000) Efficacy of ginger for nausea and vomiting: A systematic review of randomised clinical trials. *British Journal of Anaesthesiology* 84: 367–371.

Fink-Jensen A, Judge ME, Hansen JB, et al. (1992) Inhibition of cisplatin-induced emesis in ferrets by the non-NMDA receptor antagonists NBQX and CNQX. *Neuroscience Letters* 137: 173–177.

Flaxman SM and Sherman PW (2000) Morning sickness: A mechanism for protecting mother and embryo. *Quarterly Review of Biology* 75: 113–148.

Girod V, Dapzol J, Bouvier M, and Grelot L (2002) The COX inhibitors indomethacin and meloxicam exhibit anti-emetic activity against cisplatin-induced emesis in piglets. *Neuropharmacology* 42: 428–436.

Hesketh PJ, Gralla RJ, Webb RT, et al. (1999) Randomised phase II study of the neurokinin-1 receptor antagonist CJ-11,974 in the control of cisplatin-induced emesis. *Journal of Clinical Oncology* 17: 338–343.

Koch KL (1995) Approach to the patient with nausea and vomiting. In: Yamada T (ed.) *Textbook of Gastroenterology*, 2nd edn., vol. 1, pp. 731–749. Philadelphia: Lippincott.

Kris MG, Radford J, Pizzo BA, et al. (1997) Control of emesis following cisplatin by CP-122,721, a selective NK1 receptor antagonist. *Journal of the National Cancer Institute* 89: 817–818.

Lehmann A and Karrberg L (1996) Effects of *N*-Methyl-D-aspartate receptor antagonists on cisplatin-induced emesis in the ferret. *Neuropharmacology* 35: 475–481.

Lien H-C, Sun WM, Chen Y-H, Kim H, et al. (2003) Effects of ginger on motion sickness and gastric slow-wave dysrhythmias induced by circular vection. *American Journal of Physiology and Gastrointestinal Liver Physiology* 284: G481–489.

Lucot JB (1998) Effects of *N*-methyl-D-aspartate antagonists on different measures of motion sickness in cats. *Brain Research Bulletin* 47: 407–411.

Mayer DJ (2000) Acupuncture: An evidence-based review of the clinical literature. *Annual Review of Medicine* 51: 49–63.

Miller AD and Grelot L (1996) The neural basis of nausea and vomiting. In: Yates BJ and Miller AD (eds) *Vestibular Autonomic Regulation*, pp. 85–94. Boca Raton, FL: CRC Press.

Morrow GR, Roscoe JA, Hickock JT, Andrews PLR, et al. (2002) Nausea and emesis: Evidence for a biobehavioural perspective. *Support Care Cancer* 10: 96–105.

Navari RM, Reinhardt RR, Gralla RJ, et al. (1999) Reduction of cisplatin-induced emesis by a selective neurokinin receptor antagonist. *New England Journal of Medicine* 340: 190–195.

Quigley EMM, Hasler WL, and Parkman HP (2001) AGA Technical review on nausea and vomiting. *Gastroenterology* 120: 263–286.

Reynolds DJM, Andrews PLR, and Davis CJ (eds.) (1995) *Serotonin and the Scientific Basis of Anti-emetic Therapy.* Oxford: Oxford Clinical Communications.

Roscoe JA, Hickok JT, and Morrow GR (2000) Patient expectations as predictor of chemotherapy-induced nausea. *Annals of Behavioral Medicine* 22: 121–126.

Rudd JA, Cheng CHK, and Naylor RJ (1999) Modulation of emesis by fentanyl and opioid receptor antagonists in Suncus murinus (house musk shrew). *European Journal of Pharmacology* 374: 77–84.

Sanger GJ and Andrews PLR (2001) Emesis. In: Farthing MJG and Ballinger AB (eds.) *Drug Therapy for Gastrointestinal and Liver Diseases*, pp. 45–61. London: Martin Dunitz Ltd.

Simoneau II, Hamza MS, Mata HP, et al. (2001) The cannabinoid agonist WIN55, 212-2 suppresses opioid-induced emesis in ferrets. *Anesthesthesiology* 94: 882–887.

Sleisenger MH (ed.) (1993) *The Handbook of Nausea and Vomiting.* New York: CRC Press-Parthenon Publishing.

Tramer MR, Moore RA, Reynolds DJM, and McQuay HJ (1997) A quantitative systematic review of ondansetron in treatment of established postoperative nausea and vomiting. *British Medical Journal* 314: 1088–1092.

Tramer MR, Carroll D, Campbell FA, et al. (2001) Cannabinoids for control of chemotherapy induced nausea and vomiting: quantitative systemic review. *British Medical Journal* 323: 1–8.

Treisman M (1977) Motion sickness: An evolutionary hypothesis. *Science* 197: 493–495.

Van Belle S, Lichinitser MR, Navari RM, et al. (2002) Prevention of cisplatin-induced acute and delayed emesis by the selective neurokinin-1 antagonists, L-758,298 and MK-869. *Cancer* 94: 3032–3041.

Van Sickle MD, Oland LD, Ho W, et al. (2001) Cannabinoids inhibit emesis through CB1 receptors in the brainstem of the ferret. *Gastroenterology* 121: 767–774.

Yates BJ, Miller AD, and Lucot JB (1998) Physiological basis and pharmacology of motion sickness: An update. *Brain Research Bulletin* 47: 395–406.

Neglect Syndrome and the Spatial Attention Network

M M Mesulam, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Damage to the right side of the brain can trigger a debilitating impairment known as the left hemineglect (selective inattention) syndrome. Patients with this syndrome minimize the salience of events in the contralesional side of the environment, display a reluctance to shift attention in a leftward direction, and cannot effectively search for targets embedded within the left hemispace. The left hemineglect syndrome is assumed to reflect damage to a large-scale frontoparietal network that normally sustains the adaptive distribution of spatial attention.

In this context, the term 'spatial attention' designates a set of interrelated sensory, motor, and cognitive processes that collectively enable the selective allocation of neural resources to motivationally relevant parts of the environment. The spatial representation of events according to their emotional salience, their targeting for oculomotor fixation or manual grasp, the anticipatory biasing of regions where interesting events are expected to occur, and the search for objects embedded among distractors are components of spatial attention.

Experiments in monkeys traced the two major nodes of an analogous spatial attention network to posterior parietal cortex (e.g., 7a in the inferior parietal lobule (IPL) and lateral intraparietal (LIP) in the intraparietal sulcus) and a premotor/prefrontal area that includes the frontal eye fields (FEF). These two areas are reciprocally interconnected and contain neurons that encode spatial coordinates in multiple reference frames, represent extrapersonal space in terms of salience rather than object characteristics, and direct orientating and search behaviors toward salient events. They collectively provide a coarse mapping of salient extrapersonal events and encode the intent to generate saccadic eye movements or arm reaching movements toward behaviorally relevant events. Both of these areas receive peristriate information and project to the superior colliculus in a way that would allow them to exert a top-down modulation of oculomotor and visual areas involved in the motor and sensory aspects of spatial attention.

In general, area 7a has more sensory properties and may play the major role in encoding a salience-based representation of the extrapersonal events, whereas areas LIP and FEF tend to have a closer relationship to the programming of saccades in the direction of behaviorally relevant targets. Although Chafee and Goldman-Rakic have shown that neuronal firing is greater in the FEF just before a memory-guided saccade and in the LIP just after the onset of a salient spatial cue, these two areas are functionally interdependent and display relatively few differences in response patterns. They collectively represent an advanced state of sensorimotor synthesis in which perception can be represented as salience rather than sensory features, and action as intent rather than movement.

Axonally transported tracer experiments in the monkey brain showed that area 7a/LIP and the FEF are reciprocally interconnected not only to each other but also to the cingulate gyrus. Cingulate neurons in the awake animal respond to behaviorally relevant extrapersonal events and may participate in the remapping of salience following saccadic eye movements. Through its connections with the amygdala and parahippocampal complex, the cingulate gyrus provides a limbic component of the spatial attention network.

The complexity of this network and the heterogeneity of lesion sites associated with the neglect syndrome account for the large variety of clinical manifestations and explanatory theories. Research in this field has proceeded in two major directions: investigation of patients with the hemispatial neglect syndrome and functional imaging of neurologically intact subjects performing tasks of spatial attention that are impaired in these patients. This research has led to exciting new insights but has also generated two major questions. The first is related to hemispheric asymmetry. Although severe contralesional neglect is far more common after right than after left hemisphere lesions, functional imaging has not yet revealed a consistent asymmetry of neural activity during tasks of spatial attention in normal subjects. The second question concerns the location of neglect-causing lesions. Although functional imaging highlights the engagement of dorsal frontoparietal cortex in spatial attention tasks, clinical evidence links the neglect syndrome to more ventral lesions of the inferior parietal lobule and temporoparietal junction. This review summarizes clinicoanatomical correlates of the neglect syndrome and functional imaging experiments on spatial attention. The two questions discussed previously will not be resolved, but the facts that generate them will be reviewed.

The Neuropsychology of Neglect

Contralesional neglect is more lasting and severe after lesions in the right side of the brain. This asymmetry

is consistent with a neural model postulating that the left hemisphere directs attention primarily to the contralateral hemispace, whereas the right hemisphere directs attention to both hemispaces or, alternatively, serves a more global orienting function. According to this model, left hemisphere lesions do not cause prominent neglect because the right hemisphere continues to direct attention to either side of space. When neglect is severe, the patient with a right hemisphere lesion may behave almost as if one-half of the universe had abruptly ceased to exist in any meaningful form. One patient may shave, groom, and dress only the right side of the body; another may fail to eat food placed on the left side of the tray; another may omit to read the left half of each sentence or even the left side of every word printed anywhere on the page; still another may fail to copy detail on the left side of a drawing and may show a curious tendency to leave an uncommonly wide margin on the left side of the paper when asked to write.

Neglect behavior has perceptual, motor, and motivational components. Perceptual aspects of neglect can be probed with tests of extinction, line bisection, and covert attentional shifts. Extinction is said to exist when patients who respond accurately to unilateral stimulation from either the left or the right side consistently ignore the stimulation on the left under conditions of bilateral simultaneous stimulation. If extinction occurs in only one modality, it can conceivably reflect a subtle disruption of relevant sensory pathways or even a callosal disconnection syndrome. Multimodal extinction, however, almost always reflects the representational aspect of neglect. In the traditional version of the line bisection task, patients are asked to mark the midpoint of a horizontal line drawn on a sheet of paper. Patients with left hemineglect tend to place their mark substantially rightward of center, suggesting that the representational impact of the left side becomes diminished and that it takes a longer segment on the left to balance the salience of a shorter segment on the right.

A rightward skewing of mental representations was demonstrated by Bisiach and colleagues in experiments in which patients were asked to mentally retrieve prominent features along the Piazza del Duomo in Milan as they imagined themselves looking toward the cathedral. Patients with left neglect were more accurate in listing details situated on the right side of the square as viewed from that vantage point. Upon being asked to imagine looking at the square while facing away from the cathedral, the same patients showed better recall of the items that had been omitted when assuming the former vantage point. Thus, the impaired evocation of left-sided details in the first part of the experiment was not due to an obliteration of the information but to an inability to activate the part of the representation which fell to the left of the imaginary perspective. In addition to this difficulty in activating the left side of previously encoded mental representations, patients with left neglect also display a relative deficit in encoding the left side of sensory experience during the compilation of new representations.

A task developed by Posner and colleagues has played an influential role in this field of investigation. No eye movements are allowed, central fixation is required at all times, and the attentional orientation is shifted covertly within a mental representation of the ambient visual scene. These shifts can be triggered exogenously through sensory priming or endogenously through centrally presented arrows which induce directional expectancy. Patients with left hemineglect but no hemianopia show much longer reaction times when responding to targets in the left. They also display an excessive difficulty disengaging attention from a cue in the right hemispace when the task requires a subsequent leftward shift of covert orientation to targets on the left. This resistance to disengagement is partially analogous to the phenomenon of extinction.

Patients with neglect can also display a reluctance to scan and explore the left hemispace even in the absence of obvious gaze or limb paresis. The resultant impairments of search behaviors are readily elicited by tasks which require the patient to circle or check targets on a sheet of paper. Patients omit many more targets on the left, need more time to find left-sided targets, use a disorganized scanning strategy, make fewer and lower amplitude eye movements to the left, and have longer visual fixation times on right-sided targets. The use of a 90° mirror, which reverses the direction of a manual movement needed to mark a target from the direction of the shift in visual attention that leads to its detection, shows that target cancellation impairments reflect a failure to look left in some patients and a failure to reach to the left in others. The frequent observation that targets oriented randomly on the page are far less efficiently detected than targets organized into rows and columns implies that an inability to endogenously impose an orderly scanning strategy contributes to the severity of the neglect.

Patients with left unilateral neglect also display difficulties in manual search. Blindfolded search for small objects by manual palpation is intact on the right side of the table not only with the right hand but also with the left hand, whereas it becomes ineffective on the left side of the table even when the intact right hand is being used. These observations show that left neglect is associated with impaired tactile exploration within the left hemispace regardless of the limb

that is being used. Thus, motor programs involved in exploration appear to be organized not according to the muscle groups that are being activated but, rather, according to the hemispace within which the movement is to be discharged.

A major role of any attentional system is to shift the attentional spotlight toward extrapersonal events of emotional and motivational significance. Patients with unilateral neglect devalue the left side of the world and behave not only as if nothing is actually happening in the left but also as if nothing of any importance could be expected to emanate from that side. The influence of this factor can be probed by varying motivational valence. For example, a patient showed marked improvement in detecting targets on the left when he was promised one penny for each accurate detection. Another patient with severe left hemineglect failed to reach for food on the left side of a tray and would bitterly complain that his tea had been left out. On a day when the nurse was instructed to withhold breakfast, the patient became unusually hungry by noontime but did not show any change in the severe neglect of left-sided targets in a letter-cancellation task. When his lunch tray was brought, however, he had no reluctance to reach for his tea on the left side of the tray. These anecdotal observations suggest that a devaluation of sensory events on the left may contribute to the emergence of neglect and that the relationship between neglect and motivation may be material specific so that hunger decreases the spatial distribution of neglect for edible items but not for letters on a test sheet.

Many aspects of neglect are based on top-down processes that impose a spatial bias on the encoding of a percept or the reactivation of its representation. In patients with extinction, for example, reaction times to unilateral right-sided stimuli are longer than those to bilateral stimuli. Since both conditions are reported as unilateral right-sided stimulation by the patient, the extinguished stimulus appears to gain access to the central nervous system and exert a covert influence on behavior. According to a case report by Marshall and Halligan, a patient with left hemineglect who was shown line drawings of two houses, one of which had flames coming out of the left side, judged the two drawings to be visually identical but chose the one without flames when asked to select the house she would prefer to live in.

In the preceding account, the word 'left' has deliberately been used without further qualification, as if it constituted a fixed attribute such as color or texture. This is clearly not the case since there is no left or right without a specific frame of reference. At least four frames of reference define the leftness of an extrapersonal event: egocentric (defined with respect to the observer), allocentric (defined with respect to another extrapersonal event), world-centered (defined with respect to a fixed landmark in the environment), and object-centered (defined with respect to a principal axis in the canonical representation on an object). Since the eyes, head, and trunk can rotate and tilt with respect to each other, the egocentric frame of reference contains retinocentric, cephalocentric, somatocentric, and gravitational coordinates so that an event which is in the left according to one egocentric coordinate could be on the right according to another. Furthermore, perceptual and conceptual factors influence the segmentation of the stimulus field into individual clusters with their own left and right sides. Left hemineglect is defined as a behavioral state in which a location in one of the egocentric lefts, in the environmental left, in the canonical or object-centered left, in the allocentric left, and in the segmentational left is endowed with less perceptual salience and becomes less likely to attract attentional capture through search and orienting behaviors.

Spatial Attention Network in the Human Brain

Contralesional neglect can arise in patients with parietal, frontal, cingulate, or subcortical lesions. In the group of patients with parietal lesions, some investigators point to the IPL (or temporoparietal junction) as the critical site, whereas others emphasize the importance of the superior temporal gyrus, an opinion that is difficult to reconcile with the auditory functions of the superior temporal gyrus. A study by Hillis and colleagues based on the remote cortical dysfunction caused by subcortical lesions suggested that interference with IPL function may cause the egocentric (observer-centered) aspects of neglect, whereas interference with superior temporal gyrus function may cause its allocentric and object-centered aspects. Reaching deficits, extinction, and extinction-like phenomena (e.g., a slowing of attentional disengagement in tasks of covert attentional shifts) may be associated with superior parietal lobule (SPL) lesions, whereas the distortions of spatial representation may be associated with IPL lesions. In the frontal lobes, some studies show that the critical lesion site is in the FEF, whereas others implicate the inferior frontal gyrus. It is also not clear if subcortical lesions cause neglect because of the tissue they destroy or because of the remote cortical dysfunction they induce. A relatively attractive hypothesis involves the possibility that 'parietal neglect' might be predominantly perceptual, whereas 'frontal neglect' might be predominantly motor. In support of this possibility, several studies have shown that 'perceptual' tasks such as extinction and line bisection are more likely

to be associated with parietal lesions, whereas 'motor' tasks such as target cancellation are more likely to be associated with frontal lesions. Ingenious experiments with pulleys and mirrors have shown that errors in line bisection and target-detection tasks can be attributed to representational biases in patients with parietal lesions and to directional hypokinesia in patients with frontal lesions. Other studies, however, have not been able to confirm the presence of a relationship between directional hypokinesia and frontal lesions.

The strong interconnectivity between the frontal and parietal components of this network raises the possibility that damage to one may induce distal hypometabolism in the other through the process of diaschisis. It therefore seems unrealistic to expect a strict behavioral dissociation between the manifestations of neglect that result from parietal versus frontal lesions since these two components of the attentional network collectively subserve a level of sensorimotor integration in which the boundaries between action and perception become blurred.

Functional Imaging of Spatial Attention

In neurologically intact volunteers, numerous tasks of attentional shifts in the visual or auditory modalities, with or without eye movements, guided by exogenous or endogenous cues, and entailing variable components of manual or oculomotor search, have elicited conjoint activations in the FEF and in the banks of the intraparietal sulcus (IPS), together with additional but more variable clusters of activity in other posterior parietal, temporoparietal, cingulate, and prefrontal regions. These experiments have supported the contention that the FEF and posterior parietal cortex exert a top-down modulation of the superior colliculus and visual areas of the cerebral cortex so that the representation of significant events and the actions necessary for their attentional capture can be promoted. It is interesting to note that spatial attention tasks most commonly activate the IPS and FEF, whereas the clinical observations suggest that the lesions responsible for neglect tend to be more ventral in location.

According to functional imaging experiments, frontal components of the spatial attention network (FEF, supplementary eye fields, and middle frontal gyrus) are more closely related to the compilation of a prospective motor code for overt search, whereas the parietal components (IPS, SPL, and IPL) are more closely related to the maintenance of the underlying spatial representations that guide the search. The IPL/SPL region may also mediate the cue-induced voluntary reorientation of spatial attention and the remapping of the attentional landscape based on the spatial distribution of salience.

The FEF has an internal functional segregation characterized by a partially overlapping mosaic of areas mediating saccadic eye movements, visual search, working memory, and covert shifts of attentional orientation. Functional heterogeneity is also emerging within the cingulate component of the network where an anterior sector seems to mediate processes related to cognitive control, target detection, vigilance, conflict resolution, and response selection, whereas a posterior sector seems to mediate the anticipatory allocation of attention toward locations where relevant events are expected to occur. As shown by Small and colleagues, posterior cingulate activations linked to anticipatory spatial biasing are further enhanced by monetary incentive, supporting the contention that posterior cingulate neurons mediate the motivational modulation of spatial attention.

Some functional imaging experiments have yielded results consistent with right hemisphere specialization, whereas others have not. The critical role of the right hemisphere for spatial attention, so convincingly illustrated by clinical observations on neglect patients, is therefore not obvious in functional imaging experiments.

Conclusions

It is becoming increasingly clear that neglect cannot be attributed to a unitary deficit of orientation, representation, or intention. Instead, it represents an outcome of interacting perturbations in each of these processes. As in the case of aphasia and amnesia, neglect is a 'network syndrome.' It represents damage to one or more interactive components of the distributed network shown in Figure 1, in which each component has a different pattern of physiological and anatomical specialization.

The spatial attention network revolves around cortical epicenters in the posterior parietal cortex, FEF, and the cingulate gyrus. Each of these macroscopic components serves a dual purpose: it provides a local network for regional neural computations and also a nodal point for the linkage of distributed information. Although each component displays a relative specialization for specific behaviors, spatial attention as a whole is the emergent property of the entire network. Compiling a salience map of the extrapersonal space and transforming spatial coordinates into targets for reorienting attention can be included among the relative specializations of the parietal component; translating representations into systematic search behaviors reflects one of the relative specializations of the FEF, and calibrating spatial expectancy according to motivational valence represents one of the specializations of the cingulate component. The coherent

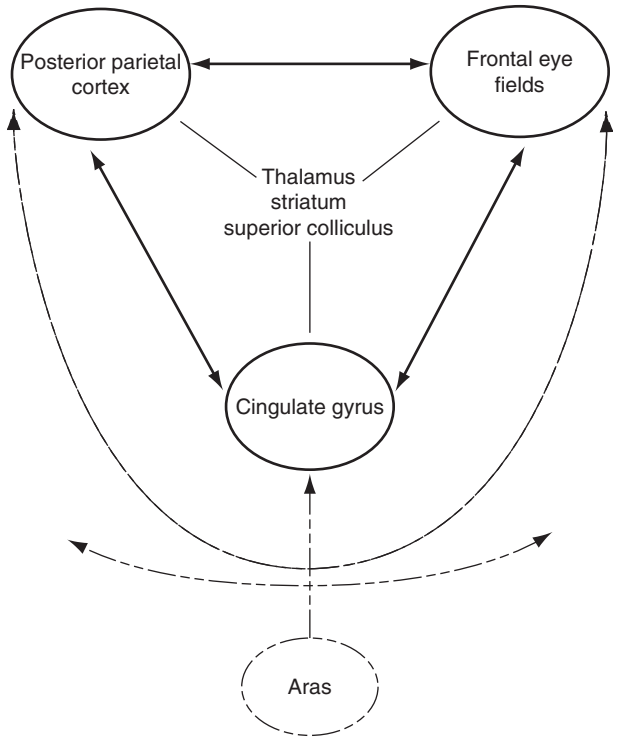

Figure 1 A large-scale network for spatial attention.

engagement of this network as a whole enables motivationally relevant extrapersonal events to become the selective targets of attentional behaviors.

Although much has been learned about the spatial attention network shown in Figure 1, many crucial questions remain. In particular, the computational features of the network, its interactions with other partially overlapping networks, the internal functional heterogeneity of its major components, and the physiological basis of right hemisphere specialization are among the major questions that will be attracting new and exciting insights.

See also: Attention and Eye Movements; Attentional Networks; Attentional Networks in the Parietal Cortex; Frontal Eye Fields; Neural Coding of Spatial Representations; Parietal Cortex and Spatial Attention; Spatial Cognition and Executive Function; Spatial Cognition; Visual System: Multiple Visual Areas in Monkeys; Visual Attention.

Further Reading

Andersen RA, Snyder LH, Bradley DC, and Xing J (1997) Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* 20: 303–330.

Bisiach G, Geminiani G, Berti A, and Rusconi M (1990) Perceptual and premotor factors of unilateral neglect. *Neurology* 40: 1278–1281.

Chafee MV and Goldman-Rakic PS (2000) Inactivation of parietal and prefrontal cortex reveals interdependence of neural activity during memory-guided saccades. *Journal of Neurophysiology* 83: 1550–1566.

Corbetta M, Kincade JM, Ollinger JM, McAvoy MP, and Shulman GL (2000) Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience* 3: 292–297.

Curtis CE, Rao VY, and D'Esposito M (2004) Maintenance of spatial and motor codes during oculomotor delayed response tasks. *Journal of Neuroscience* 24: 3944–3952.

Daffner KR, Ahern GL, Weintraub S, and Mesulam MM (1990) Dissociated neglect behavior following sequential strokes in the right hemisphere. *Annals of Neurology* 28: 97–101.

Gitelman DR, Parrish TB, Friston KJ, and Mesulam M-M (2002) Functional anatomy of visual search: Regional segregations within the frontal eye fields and effective connectivity of the superior colliculus. *NeuroImage* 15: 970–982.

Gottlieb JP, Kusunoki M, and Goldberg ME (1998) The representation of visual salience in monkey parietal cortex. *Nature* 391: 481–484.

Heilman KM, Watson RT, and Valenstein E (1985) Neglect and related disorders. In: Heilman KM and Valenstein E (eds.) *Clinical Neuropsychology*, pp. 279–336. New York: Oxford University Press.

Hillis AE, Newhart M, Heidler J, Barker PB, Herskovits EH, and Degaonkar M (2005) Anatomy of spatial attention: Insights from perfusion imaging and hemispatial neglect in acute stroke. *Journal of Neuroscience* 25: 3161–3167.

Hopfinger JB, Buonocore MH, and Mangun GR (2000) The neural mechanisms of top-down attentional control. *Nature Neuroscience* 3: 284–291.

Karnath H-O, Ferber S, and Himmelbach M (2001) Spatial awareness is a function of the temporal not the posterior parietal lobe. *Nature* 950–953.

Marshall JC and Halligan PW (1988) Blindsight and insight in visuo-spatial neglect. *Nature* 336: 766–767.

Mesulam M-M (1999) Spatial attention and neglect: Parietal, frontal, and cingulate contributions to the mental representation and attentional targeting of salient extrapersonal events. *Philosophical Transactions of the Royal Society B* 354: 1325–1346.

Nobre AC, Coull JT, Maquet P, Frith CD, Vandenberghe R, and Mesulam M-M (2004) Orienting attention to locations in perceptual versus mental representations. *Journal of Cognitive Neuroscience* 16: 363–373.

Posner MI (1980) Orienting of attention. *Quarterly Journal of Experimental Psychology* 32: 3025.

Small DM, Gitelman D, Simmons K, Bloise SM, Parrish T, and Mesulam M-M (2005) Monetary incentives enhance processing in brain regions mediating top-down control of attention. *Cerebral Cortex* 15: 1855–1865.

Summerfield JJ, Lepsien J, Gitelman DR, Mesulam M-M, and Nobre AC (2006) Orienting attention based on long-term memory experience. *Neuron* 49: 905–916.

Vallar G (1993) The anatomical basis of spatial hemineglect in humans. In: Robertson IH and Marshall JC (eds.) *Unilateral Neglect: Clinical and Experimental Studies*, pp. 27–59. Hillsdale, NJ: Erlbaum.

Weintraub S and Mesulam M-M (1988) Visual hemispatial inattention: Stimulus parameters and exploratory strategies. *Journal of Neurology, Neurosurgery & Psychology* 51: 1481–1488.

Yantis S, Schwartzbach J, Serences JT, et al. (2002) Transient neural activity in human parietal cortex during attentional shifts. *Nature Neuroscience* 5: 995–1002.

Neocortex: Origins

F Aboitiz, Pontificia Universidad Católica de Chile, Santiago, Chile

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The origin of the mammalian neocortex is probably one of the most controversial themes in modern comparative neurobiology. There have been long-standing arguments on this issue (dating from early last century), and still there is no general consensus about it. This is partly due to the enormous anatomical diversification that is observed in the cerebral hemispheres of distinct vertebrate groups. Starting from a relatively conserved embryonic plan that divides the hemispheres into a dorsal pallium and a ventral subpallium, different brain components have expanded in each vertebrate class, producing profound distortions of the ancestral adult condition. For example, cartilaginous fishes are characterized by the development of a central pallial nucleus located at the dorsal midline, while bony fish develop a brain in which the ventricular region everts to the outside, separating the medial hemispheres in the dorsal aspect. Amphibians possess quite a simple brain with little neuronal migration outside the internal ventricular zone, and almost no expansion of cell masses. Their pallium has been subdivided into medial, dorsal, and lateral components. On the other hand, the pallium of reptiles and birds (sauropsida) displays a cortical, laminated architecture containing a medial/dorsomedial (hippocampal) cortex, a lateral (olfactory) cortex, and a small dorsal cortex interposed between these two (in birds, the region corresponding to the dorsal cortex has been termed the Wulst, or more recently, the hyperpallium). These components correspond to the different pallial subdivisions in amphibians. In addition, the brains of sauropsids display a conspicuous nuclear structure that bulges into the ventricle, the dorsal ventricular ridge (DVR), which receives many thalamic sensory projections (Figure 1).

Mammals are characterized by the presence of a neocortex (Figure 1). This structure is bound medially by the medial cortex or hippocampus, and laterally by the lateral or olfactory cortex. Deep to some cortical areas like the insular region there are structures such as the claustral complex. These and other dorsal structures make up the pallium, while the ventral subpallium includes the basal ganglia and related nuclei. The neocortex of mammals is indeed a unique structure, consisting of six laminae that remain somewhat conserved across species and cortical regions. Furthermore, it is subdivided into several distinct areas in the tangential domain, some of them being sensory, others motor, and still others receiving multimodal input. In several mammalian groups there is a clear trend of tangential neocortical expansion that results in the increase of both surface area (leading to the formation of fissures and convolutions) and the number of cortical areas. The neocortex receives most of its extrinsic input from dorsal thalamic nuclei, many of them conveying sensory information from somatosensory, auditory, or visual modalities, and projects back to the thalamus, to the amygdala, to the entorhinal cortex (which in turn projects to the hippocampus), and to several subcortical structures, including the spinal cord.

Historical Background

Early in the last century, concepts on the evolution of the amniote brain (i.e., reptiles, birds, and mammals) were dominated by Elliot Smith's interpretation that sauropsids were characterized by the expansion of subpallial structures like the basal ganglia, leading to the development of the DVR. On the other hand, mammals would have expanded the pallial region, producing the neocortex. Interestingly, other authors, such as Nils Holmgren and Bengt Källén, proposed a pallial origin for the reptilian DVR, which they described as arising from a deep proliferative region that overlaps with the lateral cortex. Holmgren termed this region as the hypopallium, and considered it to be related to deep-brain mammalian structures like the claustrum or parts of the amygdalar complex. However, these views were eclipsed by Elliot Smith's claim for a subpallial origin of the DVR.

Subsequent analyses performed in the 1970s, taking advantage of newly developed histochemical procedures, revealed that the reptilian DVR does not contain acetylcholinesterase activity as the basal ganglia do, and is therefore a pallial structure. Furthermore, tract-tracing experiments performed by Harvey Karten and collaborators, showed that the avian DVR (more specifically, its anterior part, ADVR, or nidopallium) received thalamic sensory projections similar to those reaching some parts of the mammalian neocortex. More precisely, collothalamic nuclei (i.e., thalamic nuclei receiving input from mesencephalic sensory centers, conveying visual and auditory information) project to the ADVR/nidopallium of birds, and to the lateral neocortex of mammals (visual extrastriate and

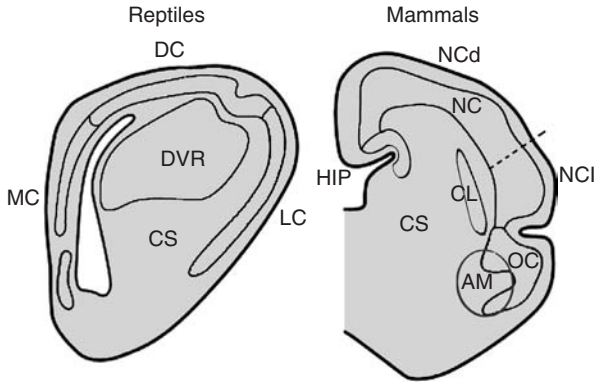

Figure 1 The cerebral hemispheres of reptiles and mammals. In the dorsal aspect of the hemisphere, reptiles display a cortex, with medial/dorsomedial (MC), dorsal (DC), and lateral (LC) components, and a dorsal ventricular ridge (DVR) consisting of an anterior part and a posterior part (not shown). The anterior DVR receives most sensory projections from mesencephalic-receiving thalamic nuclei. The medial and dorsomedial cortices of reptiles (MC) have been compared to the hippocampal formation (HIP) of mammals, and the lateral cortex (LC) is considered homologous to the mammalian olfactory cortex (OC). The dorsal cortex (DC) of reptiles has characteristics of both the entorhinal cortex (adjacent to the hippocampal formation; not shown) and the neocortex (NC) of mammals. The latter appears subdivided into a dorsal aspect, NCd, and a lateral aspect, NCl. There is no consensus regarding the homology of the anterior DVR. According to some authors, this structure corresponds to the mammalian lateral neocortex (NCl), while others consider that it relates to the claustroamygdaloid complex (AM, CL). Reproduced from Northcutt RG and Kaas JH (1995) The emergence and evolution of mammalian neocortex. *Trends in Neuroscience* 18: 373–379, with permission from Elsevier.

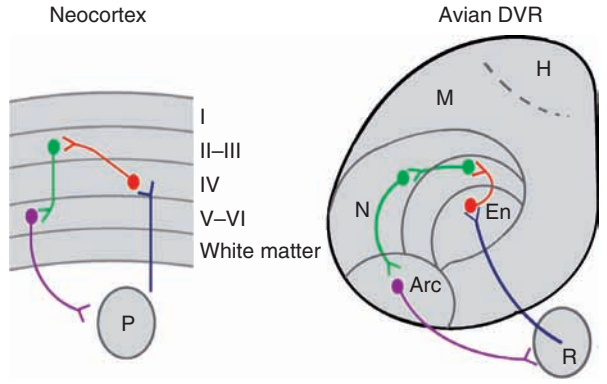

Figure 2 Harvey Karten's original cell-equivalent hypothesis, in which visual processing circuits of the avian nidopallium (N) and arcopallium (Arc) are homologous to those of the different layers in the mammalian extrastriate visual cortex. En, entopallium (a component of the nidopallium); H, hyperpallium; M, mesopallium; P, pulvinar nucleus; R, rotundus nucleus.

auditory cortices). On the other hand, lemnothalamic nuclei (receiving direct sensory input from ascending lemniscal systems, carrying visual or somatosensory input) project to the avian hyperpallium, and to the dorsal neocortex of mammals (primary visual and somatosensory cortices). Thus, Karten proposed that the avian hyperpallium (dorsal cortex of reptiles) was homologous to the dorsal neocortex of mammals (both receiving lemnothalamic input), while the ADVR/nidopallium of birds and reptiles was homologous to the mammalian lateral neocortex (both receiving collothalamic input). Furthermore, relying on similarities in intrinsic connectivity, it was proposed that different components of the avian ADVR corresponded to specific neocortical layers in mammals (Figure 2).

This hypothesis dominated the views of early mammalian brain evolution and neocortical origins for several years. Nevertheless, there were always some dissenters, backed by the early embryological studies and by the different topographical position of the mammalian neocortex and the reptilian DVR. In these alternative views, the most likely homolog of the whole mammalian neocortex was the reptilian dorsal cortex, which is located in a similar relative position as the neocortex: interposed between the lateral (olfactory) and the medial (hippocampal) cortices. Thus, in the origin of mammals, collothalamic projections originally destined to the ADVR or its equivalent region might have deviated or sent collaterals toward an expanding dorsal pallium or dorsal cortex. This change was considered to be more likely to occur than a transformation of the topographic positions of the embryonic proliferative zones and the adult structures. Supporting this view, the mammalian neocortex bears other patterns of connectivity that are not found in sauropsids, like the interhemispheric connections and the corticospinal tract. This is to be expected from a structure that increases in size and consequently increases its connectivity domains. Additional considerations on the topography, development, and connectivity of the sauropsidian ADVR led to the revival of Holmgren's hypothesis of homology between this structure and either the dorsal claustral complex or the basolateral amygdala of mammals.

Developmental and Genetic Evidence

An important breakthrough in this controversy was provided by the analysis of regulatory gene expression patterns in the developing telencephalon, which identified three main compartments during the specification of the cerebral vesicles: the subpallium, being positive for regulatory genes such as *Dlx1* and *Dlx2*; the medial, dorsal, and lateral pallia (or medial, dorsal, and lateral cortices), expressing genes such as *Emx1*, *Tbr1*, and *Pax6*; and the intermediate territory or ventral pallium located between these two, expressing *Tbr1* and *Pax6* but not *Emx1* (Figure 3).

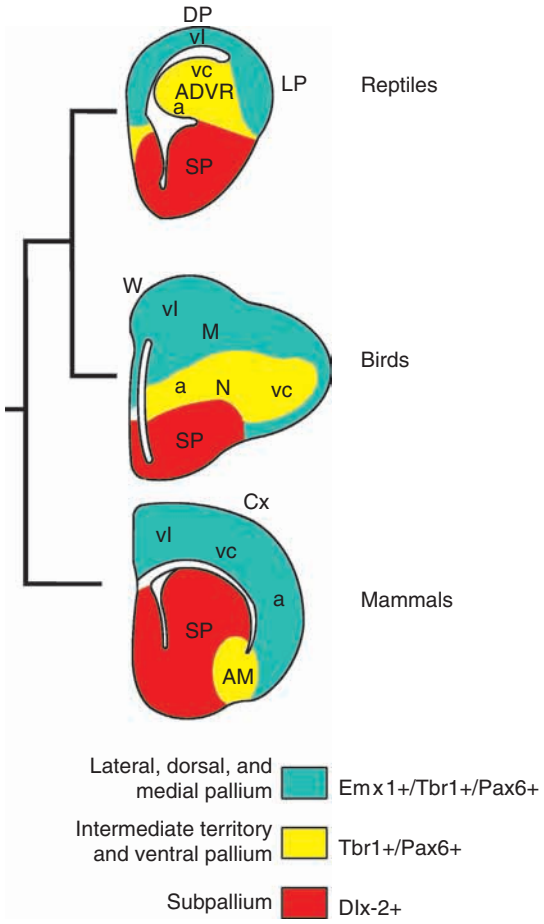

Figure 3 Patterns of regulatory gene expression in the developing telencephalon of sauropsids and mammals. All pallial components (medial, dorsal, and lateral; blue) express the markers Emx1, Tbr1, and Pax6, with the exception of the ventral pallium (yellow), which is positive for Tbr1 and Pax6 but not for Emx1. The subpallium (SP, red) expresses *Dlx* genes. According to this evidence, the ventral pallium gives rise to the reptilian anterior dorsal ventricular ridge (ADVR), to the avian nidopallium (N), and to the claustroamygdalar complex (AM) of mammals. a, vc, termination of auditory and visual collothalamic pathways, respectively; vl, termination of visual lemniscal pathways; DP, dorsal pallium; LP, lateral pallium; Cx, cortex; M, mesopallium; W, Wulst or hyperpallium. Adapted from Aboitiz F, Morales D, and Montiel J (2003) The evolutionary origin of the mammalian isocortex: Towards an integrated developmental and functional approach. *Behavioral and Brain Sciences* 26: 535–586, with permission from Cambridge University press.

Interestingly, the ventral pallium maps into parts of the claustrum and the amygdala of mammals and into the DVR of sauropsids, indicating a common developmental origin for these structures, and supporting embryonic homology between the territories producing the reptilian ADVR and parts of the claustroamygdalar complex of mammals. Nevertheless, it is not clear that in the adult state there is any precise homology between the different ventral pallial derivatives in sauropsids and mammals, as both structures are likely to have undergone distinct evolutionary histories. In fact, fossil evidence indicates that the lineage of mammal-like reptiles diverged quite early from the reptilian line, and that their brains were extremely simple in their gross morphology, resembling more the brains of modern amphibians than those of modern reptiles.

In an attempt to accommodate these data to Karten's original hypothesis, it has been proposed that the lateral neocortex of mammals arose by virtue of a transformation of a ventral pallial structure into a dorsal pallial phenotype. This may have resulted either from the acquisition of dorsal characteristics of a yet unspecified territory, or from duplication of the ventral pallium and subsequent transformation of its dorsal moiety. Nevertheless, these hypotheses might be difficult to refute, as any dorsal pallial feature in the lateral neocortex could be considered to result from the transformation from a ventral to a dorsal pallial structure. Instead, these proposals require positive evidence, such as a direct sign of ventral pallial origin in the lateral neocortex.

Reptilian Dorsal Cortex: Neocortical and Entorhinal/Subicular Features

Furthermore, the presumed homology – between visual and somatosensory areas in the reptilian dorsal cortex, and primary visual and somatosensory areas of the mammalian neocortex – has been questioned by some authors. For example, somatosensory projections to the dorsal cortex have not been observed in either lizards or turtles, and thus may not represent an ancestral character of sauropsids. In addition, visual projections to the dorsal pallium reach a region that in turtles includes the dorsal cortex and the pallial thickening, and in lizards, are restricted to the pallial thickening. The pallial thickening is located in the lateral aspect of the pallium of some reptiles and its homologies and embryonic origins are unclear. Moreover, the multimodal dorsomedial pallium of amphibians and reptiles has several features reminiscent of the mammalian entorhinal cortex. Like the latter, the reptilian dorsal cortex has important connections with the medial/dorsomedial or hippocampal cortices. In addition, the reptilian dorsal cortex does not participate in vision but rather in spatial learning and memory. Thus, the dorsal cortex of reptiles shares features with the entorhinal/subicular cortex of mammals, but also contains a direct visual (and sometimes somatosensory) sensory input that, if perhaps not being strictly equivalent to the mammalian visual (or somatosensory) cortex, may serve to feed sensory information to the hippocampus for establishing sensory associations during learning.

The Dorsalization Hypothesis

Considering the evidence discussed herein, Aboitiz and collaborators proposed that the mammalian neocortex arose as an expansion of an ancestral pallial region with dorsal characteristics, which acquired a massive lemnothalamic and collothalamic input. This may be conceived as a dorsalization effect, in which the activity of genes (*FGFs*, *Wnts*, *Emxs*, *Lhx2*, *Pax6*, and others) or signaling pathways involved in dorsal pallial specification was enhanced, thus producing an increase in proliferative activity leading to the expansion of a primitive dorsal cortex into a neocortex. More precisely, the embryonic dorsal pallium represents a transitional zone that receives influence from several signaling sources, including the anterior forebrain expressing FGF8; the medially located cortical hem expressing Wnts and BMPs, and regulating other genes like *Emxs* (these signals specify both the medial pallium and the neocortex); and the *Pax6*-dependent antihem that specifies the lateral and ventral pallium. The expansion of this transitional tissue may have somehow resulted from the interaction between these three morphogenetic fields.

This process was not only the result of a developmental transformation; probably it was also favored for functional reasons. As nocturnal animals, early mammals probably had refined senses of smell and hearing, which was reflected in the development of a large olfactory bulb, a middle ear, and external outer ears. Strong associative networks involving olfaction and other senses developed in the dorsal pallium and the hippocampus, which participated in the establishment of learned maps of space. These networks benefited from the expansion of the dorsal pallium, which in turn facilitated the projection of collothalamic pathways that were originally directed to the ventral pallium or to the subpallium into this structure.

Lamination of the Neocortex: The Preplate

The neocortex is not only the result of the tangential expansion of the dorsal pallium. There is also a dramatic increase in radial thickness, as the reptilian cortex is only a few cells deep while the mammalian neocortex is organized in vertical columns about 100 cells in depth. As mentioned, the adult neocortex consists of six laminae, the most superficial (layer I) being a cell-poor layer in which tangential axons from Martinotti cells and other axons run along various distances. The development of the neocortex is quite a conspicuous one, in which the first cells to arrive make up a transient embryonic structure termed the preplate, consisting of Cajal–Retzius cells and pioneer neurons. The former derive in large part from the medially located cortical hem and migrate laterally into the neocortex, while pioneer neurons derive mainly from the subpallium. Reelin-secreting Cajal–Retzius cells are involved in radial glia maintenance and in controlling the latest steps of neuronal migration to the cortex, and pioneer neurons display an axon directed subcortically that acts as a guide for growing thalamocortical axons. Subsequently, radially migrating neurons that make up the true neocortical plate arrive from the cortical ventricular zone and split the preplate into two layers: a superficial marginal zone, containing the Cajal–Retzius cells and other cells, and a deep subplate below the emerging cortical plate, containing most pioneer neurons and other cells such as excitatory neurons (**Figure 4**). The latter receive transient contacts from arriving thalamic axons before they enter into the mature cortex. Most of the cells in the marginal zone and the subplate die during development or dilute with cortical expansion, while cells of the cortical plate remain in adulthood. Interestingly, the development of the cortical plate occurs in an inside-out sequence in which the earliest-produced cells end up in the deepest cortical layers, and the latest-produced cells

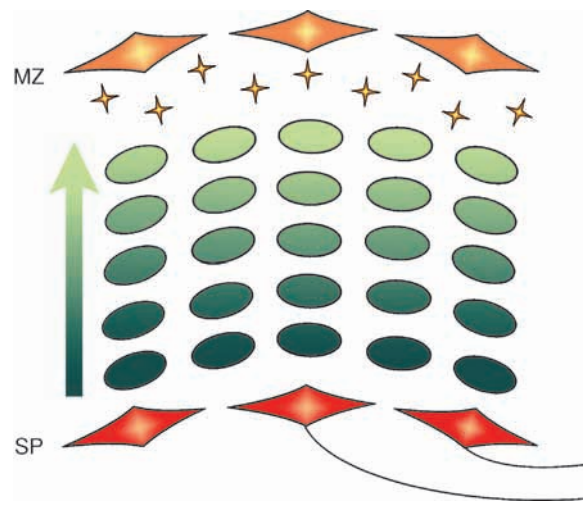

Figure 4 Lamination of the mammalian neocortex. The earliest cells to arrive at the embryonic cortex make up a preplate (not shown). The preplate becomes divided into two layers: a superficial marginal zone (MZ, future layer I), containing Cajal–Retzius cells (diamond-shaped cells above, brown) that secrete reelin (yellow stars), and a deep subplate (SP), containing pioneer neurons (diamond-shaped cells below, red) whose axon serves as a guide for growing thalamocortical axons. Cells that locate between these layers make up the cortical plate (green ovals) and arrange in five laminae (VI–II in the mature neocortex). These laminae are generated in an inside-out sequence in which early-produced neurons locate above late-produced ones (arrow represents progressively later neurogenetic times).

end up in the more superficial layers. This implies that late-produced neurons have to cross the layers of early-produced neurons and locate above them. Additionally, inhibitory neurons originating in the subpallium populate the developing cortical plate, also following an inside-out gradient in which early-produced cells arrive at deeper layers and late-produced cells arrive at superficial layers. This whole process contrasts with the situation of the much thinner reptilian cortex, where cells migrate in an outside-in sequence, early layers locating more superficially than late-produced layers.

Is there any correspondence between the reptilian cortical cells and the different layers making up the mammalian neocortex? Miguel Marín-Padilla originally proposed that neurons of the mammalian preplate (more specifically, those making up the subplate) were homologous to the reptilian cortex, the cortical plate being a developmental addition to the mammalian brain. However, other authors proposed that the mammalian preplate and the subplate emerged as devices to support the development of corticocortical connectivity in mammals. Instead of corresponding to the mammalian preplate, the reptilian cortex was claimed to be homologous to the deep neocortical layers V and (especially) VI of the mammalian cortical plate.

There are reports of a subplate-like structure in reptiles consisting of tangentially migrating interneurons, but no excitatory neurons have yet been described in it. On the other hand, in a marsupial, no subplate was detected and corticothalamic axons were seen directly penetrating the cortical plate, instead of establishing transient contacts below it as they do in placental mammals. If this observation is correct and true for all marsupials, it would imply that at least part of the subplate (its excitatory component, which receives arriving thalamic axons) is an acquisition of placental mammals. Moreover, in some mammals, including some marsupials, there is an adult cortical layer inferior to layer VI, termed layer VIb or VII, which contains persisting subplate cells. These cells have been considered to be closely related to the deepest neurons in the cortical plate. Furthermore, there is molecular evidence for an intimate relation between the inferior layers of the cortical plate and the subplate. The gene *Tbr1* marks both subplate neurons and the earliest cortical plate neurons, and mutations in this gene cause defects in both structures. Thus, in the origin of mammals (or of placental mammals), a subplate-like structure containing tangentially migrating neurons might have incorporated excitatory cells that originally belonged to the cortical plate, that aided in the guidance of thalamocortical axons.

Origin of the Superficial Neocortical Layers and the Inside-Out Neurogenetic Gradient

Additional molecular evidence supports the concept of the superficial neocortical layers IV–II as developmentally distinct from the inferior layers VI and V. Among other features, mutants for the gene *Pax6* exhibit cortical migration defects, consisting of the inability to migrate and differentiate supragranular neocortical neurons belonging to layers IV to II. Furthermore, early-produced cortical plate neurons destined to neocortical layers VI and V are produced in the embryonic ventricular zone and express the marker gene *Otx1*. On the other hand, late-produced neurons, which depend on *Pax6* for their development and are destined to superficial layers IV–II, are produced in the subventricular zone surrounding the ventricular zone, and express the marker gene *Svet1*. Thus, *Pax6* may be involved in the origin of the superficial neocortical layers by recruiting the subventricular zone in the neurogenetic process. Further findings indicate that the proneural genes *Ngn1* and *Ngn2* specify the laminar characteristics of early-born cortical neurons, but upper-layer cortical neurons are specified in an *Ngn*-independent manner, requiring instead the synergistic activities of *Pax6* and *Tlx*. There is also comparative evidence indicating a progressive development of the subventricular zone in large-brained mammals, its near absence in the reptilian embryonic dorsal cortex, and its independent development in the avian nidopallium. It will be of interest to determine what, if any, is the contribution of the subventricular zone and of the genes *Pax6*, *Svet1*, and *Tlx* to dorsal cortical development in reptiles.

In this context, the effect of mutations in some genes involved in the regulation of cortical cell migration has provided further clues to the development of the laminar organization of the neocortex. For example, the reeler mutant mouse, deficient in the extracellular protein reelin, shows a striking inversion of the developmental gradient in the cortical plate, in which early-produced neurons become located superficially and late-produced neurons arrange in the inferior layers. Furthermore, in this mutant the preplate is not split and is compressed above the cortical plate. In the adult, no cell-poor layer I is detectable, as the cortical plate cells are observed immediately below the subpial layer. Electron microscopy analyses indicate abnormal attachments between the neurons and the radial glia, suggesting that these neurons did not detach correctly from the latter. There have been several hypotheses for the participation of reelin in the control of neuronal migration, including a role as

a stop signal, as a glial detachment signal, or as an attractor for migrating neurons. Overall, the evidence indicates that reelin has a role in the final stages of neuronal migration and, more important to our purposes, in detaching migrating neurons from the radial glia, and in the generation of a cell-free superficial marginal zone. Reelin is present in several other brain regions besides the cerebral cortex, and its patterns of expression reflect an evolutionarily stable framework. However, reelin expression is strongly amplified in the mammalian cortical marginal zone in relation to the reptilian marginal zone, indicating that this molecule has been one key element in the origin of the neocortex. Furthermore, the level of reelin expression correlates with the degree of columnarity present in the dorsal cortex of different reptiles, suggesting that it may have contributed to the radial organization of the mammalian neocortex.

Another molecule of interest is the cyclin-dependant kinase cdk5; cdk5 and its activator, p35, are fundamental for the adequate lamination of the neocortex. Recently, two modes of radial migration have been characterized: nuclear translocation, in which the leading process of the neuron is attached to the subpial layer and the cell body is dragged toward the periphery as a consequence of the shortening of this process; and cellular locomotion, in which cells migrate along the radial glia in their pathway to the cortex. Interestingly, early-produced, deep-layer cortical neurons migrate mainly by translocation, while late-produced, superficial neurons use glial-guided locomotion to reach the cortical plate. Mutants for the enzyme cdk5 show strong deficits in locomotion, and therefore the superficial layers of the cortex fail to form. In these animals, neurons are not able to reach the cortical plate by birth and remain in the subcortical intermediate zone. However, early-produced layer VI cells which migrate by translocation do so correctly and are able to split the preplate normally. Moreover, mutants for the cdk5 activator p35 display milder defects in locomotion, and neurons from all layers are able to reach the cortical plate by birth. However, within the cortical plate, cells arrange in an inverted outside-in gradient reminiscent of the reeler mouse, but with the difference that the preplate has been partially split and there is a normal marginal zone. *N*-Cadherin-mediated cell aggregation in migrating neurons is inhibited by cdk5/p35, which provides a possible mechanism by which normally migrating cells may use this cascade to bypass a migration-suppressing signal (*N*-cadherin or a related molecule) present in postmigratory neurons, to move above them. Considering that early-produced cortical plate cells are more likely to correspond to reptilian cortical cells, it is conceivable that the cdk5-independent mechanism of translocation represents an ancestral condition, while the activation of the cdk5 pathway represents an evolutionary innovation of the mammalian brain, associated with new modes of neuronal migration and, by virtue of the activator p35, contributing to the generation of the inside-out neurogenetic gradient that is characteristic of the neocortex.

A Scenario for the Origin of Neocortical Lamination

Based on this evidence, a possible scenario for the evolution of cortical lamination is that initially there was a radial expansion of the primitive mammalian cortex, producing late-born, phylogenetically new neurons which participated in local processing circuits, as opposed to the older neurons that function as an output layer. These new elements originated in an expanded subventricular zone, by virtue of the action of the gene *Pax6*, among others. The acquisition of an inside-out neurogenetic gradient was possibly related to the ancestral organization of afferents to the primitive cortex. The neocortex differs from other structures, including the hippocampus, the olfactory cortex, and the reptilian cortex, in the strong radial organization of its inputs, in which afferents penetrate the cortical plate from the underlying white matter (aided by subplate cells and pioneer neurons); in these other structures the afferents arrange tangentially in the superficial marginal zone (**Figure 5**). Radial expansion of the primitive cortex in the context of a primitive outside-in developmental pattern would have left the new, local processing neurons located below the early-produced output neuronal layer, separating them from the superficially located afferents. In such context, there would be a benefit to those individuals in which late-produced neurons were able to cross the layer of early-produced cells, generating an intermediate layer between the input axons and the output cells. In other words, the inside-out neurogenetic gradient was produced in order to maximize synaptic contacts between the new, late-born neurons, and the superficially located cortical afferents. In this process, the cdk5 signaling pathway facilitated new mechanisms of radial migration, and especially its activator p35 was important in permitting late-born neurons to cross the barrier formed by previously migrated cells. Finally, an important increase in the expression of reelin by Cajal–Retzius cells permitted the preservation of a neuron-free marginal zone. Reelin also contributed to stop radial migration by detaching migrating neurons from the radial glia, thus facilitating the migration of young neurons past old ones.

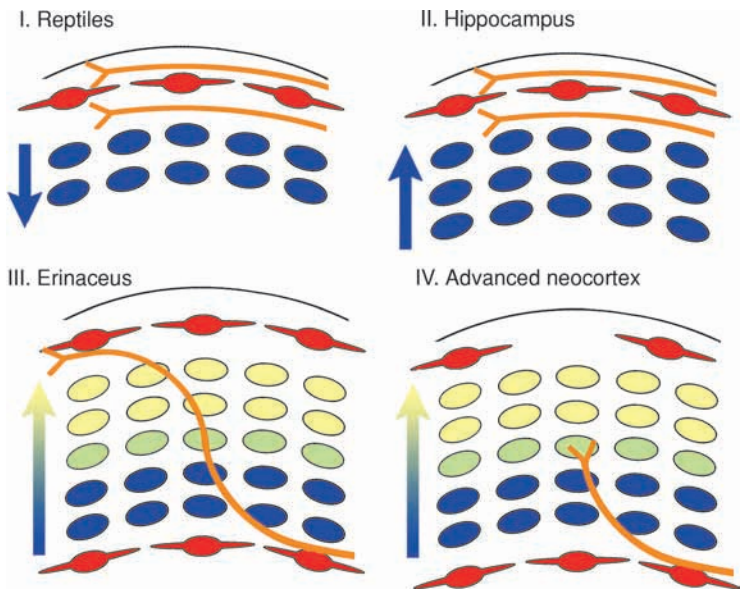

Figure 5 Postulated sequence in the evolution of neocortical lamination. In a primitive state (I) like that of reptilian cortex, with an outside-in neurogenetic gradient (arrow downwards), pioneer neurons (pointed ovals, red) locate in the superficial marginal zone and guide thalamocortical axons (Y-shaped lines) tangentially above the cortex (ovals, blue, green, and yellow). A second stage (II) may be exemplified by the hippocampus, which already displays an inside-out gradient (arrow upwards) while its afferents are located superficially. The neocortex of the hedgehog (Insectivora) represents a third stage (III), where axons enter the cortical plate from below but many run obliquely and end tangentially in the marginal zone. Finally (IV), the primary sensory cortex of advanced mammals displays axons radially entering the cortex and synapsing with neurons in layer IV, and a large proportion of pioneer neurons locate in the subplate below the cortex.

As the neocortex expanded tangentially, however, the distance traveled by tangential afferent axons became increasingly longer, with a consequent impairment in transmission time. A solution to this problem was provided by the rerouting of these axons via the subcortical white matter, making a shortcut to their final targets. This was possible by virtue of the displacement of the pioneer neurons attracting the thalamic afferents, from a superficial position in the marginal zone of reptiles, to a deeper position in the subplate of mammals. Thus, there were several steps in the acquisition of a neocortex, including the increase in cortical thickness, the origin of the inside-out gradient, and the displacement of axonal trajectories from a superficial to a subcortical position. Interestingly, intermediate stages may be found between the reptilian cortex, displaying superficial, tangential inputs, and the advanced mammalian neocortex, with radially organized inputs. For example, like the neocortex, the hippocampus has an inside-out developmental gradient, but also a superficial localization of afferents, indicating a step above the reptilian cortex. A subsequent stage may be represented by the neocortex of the primitive insectivore *Erinaceus*, in which axons penetrate the cortex radially from the underlying white matter, cross obliquely to the marginal zone, and run tangentially there for some distance. Finally, in the sensory neocortex of advanced mammals, afferents enter the cortex radially and synapse in spiny stellate neurons of layer IV (Figure 5).

Final Comments

Summarizing the views presented here, the mammalian neocortex displays several unique characteristics and therefore the term 'neocortex' is indeed appropriate. One main dicussion in relation to its origins concerns the ancestral structure that gave rise to it. There are two main proposals in this regard, one that claims that the neocortex partly arose from the DVR of reptiles, and the other that prescribes that it originated as an expansion of the dorsal pallium. Current developmental evidence tends to favor the latter, as the reptilian DVR has embryonic origins similar to those of the mammalian claustroamygdalar complex in the ventral pallium. The neocortex is also unique in its laminar structure. Recent genetic and developmental analyses have provided important insights as to the mechanisms and possible sequence of events that gave rise to the neocortex, producing its characteristic inside-out neurogenetic gradient and the rerouting of thalamic afferents from a superficial to a deep, subcortical position. All this evidence strengthens the view that development may not only give clues to evolutionary processes, but also that variations in the developmental processes are the main events in the generation of new structures in evolution.

See also: Cerebral Cortex; Evolution of Vertebrate Brains; Homology and Homoplasy.

Further Reading

Aboitiz F, Montiel J, and García RR (2005) Ancestry of the mammalian preplate and its derivatives: Relicts or embryonic adaptations. *Reviews in the Neurosciences* 16: 359–376.

Aboitiz F, Morales D, and Montiel J (2003) The evolutionary origin of the mammalian isocortex: Towards an integrated developmental and functional approach. *Behavioral and Brain Sciences* 26: 535–586.

Bar I, Lambert de Rouvroit C, and Goffinet A (2000) The evolution of cortical development. An hypothesis based on the reelin signaling pathway. *Trends in Neuroscience* 23: 633–638.

Bruce LL and Neary TJ (1995) The limbic system of tetrapods: A comparative analysis of cortical and amygdalar populations. *Brain, Behavior and Evolution* 46: 224–234.

Butler AB and Hodos W (1996) *Comparative Vertebrate Neuroanatomy. Evolution and Adaptation.* New York: Wiley-Liss.

Caviness VS, Crandall JE, and Edwards MA (1988) The reeler malformation. Implications for neocortical histogenesis. In: Peters A and Jones EG (eds.) *Cerebral Cortex, Vol. 7. Development and Maturation of the Cerebral Cortex*, pp. 59–89. New York: Plenum.

Doron NN and LeDoux JE (1999) Organization of projections to the lateral amygdala from auditory and visual areas of the thalamus in the rat. *Journal of Comparative Neurology* 412: 383–409.

Goffinet AM, Daumerie C, Langerwerf B, and Pieau C (1986) Neurogenesis in reptilian cortical structures: ^{3}H-thymidine autoradiographic analysis. *Journal of Comparative Neurology* 243: 106–116.

Karten HJ (1997) Evolutionary developmental biology meets the brain: The origins of mammalian neocortex. *Proceedings of the National Academy of Sciences of the United States of America* 94: 2800–2804.

Kriegstein AR and Noctor SC (2004) Patterns of neuronal migration in the embryonic cortex. *Trends in Neuroscience* 27: 392–399.

Marín-Padilla M (1978) Dual origin of the mammalian neocortex and evolution of the cortical plate. *Anatomy and Embryology* 152: 109–126.

Molnár Z, Métin C, Stoykova A, et al. (2006) Comparative aspects of cerebral cortical development. *European Journal of Neuroscience* 23: 921–934.

Nadarajah B and Parnavelas JG (2002) Modes of neuronal migration in the developing cerebral cortex. *Nature Reviews Neuroscience* 3: 423–432.

Northcutt RG and Kaas JH (1995) The emergence and evolution of mammalian neocortex. *Trends in Neuroscience* 18: 373–379.

Powers AS (1990) Brain mechanisms of learning in reptiles. In: Kesner RP and Olton DS (eds.) *Neurobiology of Comparative Cognition*, pp. 75–87. New York: Erlbaum Press.

Puelles L, Kuwana E, Puelles E, et al. (2000) Pallial and subpallial derivatives in the embryonic chick and mouse telencephalon, traced by the expression of the genes Dlx-2, Emx-1, Nkx-2.1, Pax-6 and Tbr-1. *Journal of Comparative Neurology* 424: 409–438.

Schuurmans C, Armant O, Nieto M, et al. (2004) Sequential phases of cortical specification involve neurogenin-dependent and -independent pathways. *EMBO Journal* 23: 2892–2902.

SmithFernández A, Pieau C, Repérant J, et al. (1998) Expression of the Emx-1 and Dlx-1 homeobox genes defines three molecularly distinct domains in the telencephalon of mouse, chick, turtle and frog embryos: Implications for the evolution of telencephalic subdivisions in amniotes. *Development* 125: 2099–2111.

Striedter GF (2005) *Principles of Brain Evolution.* Sunderland, MA: Sinauer Associates.

Supèr H and Uylings HB (2001) The early differentiation of the neocortex: A hypothesis on neocortical evolution. *Cerebral Cortex* 11: 1101–1109.

Sur M and Rubenstein JL (2005) Patterning and plasticity of the cerebral cortex. *Science* 310: 805–810.

Voogd J, Nieuwenhuys R, and VanDongen PAM (1998) Mammals. In: Nieuwenhuys R, Ten Donkelaar HJ, and Nicholson C (eds.) *The Central Nervous System of Vertebrates,* vol. 3, pp. 1637–2098. Berlin: Springer.

Neocortical Organization in Monotremes

L Krubitzer and K Campi, University of California at Davis, Davis, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Monotremes are a unique order of mammals that includes only three extant species: the duck-billed platypus (*Ornithorynchus anitinus*), the short-billed echidna (*Tachyglossus aculeatus*), and the western long-billed echidna (*Zaglossus bruijni*). Like other mammals, monotremes possess a segmented jaw, have three bones that comprise the middle ear, are warm blooded, and lactate. However, they have retained at least some characteristics of therapsid reptiles, such as laying eggs and the presence of a cloaca (in fact, the term 'monotreme' literally means 'single hole'). Although these animals are often referred to as primitive or ancestral, they are not the ancestors of all mammals. Instead, monotremes formed a very early radiation of mammals that originally evolved from early synapsids, which are considered to have been reptile-like mammals (Figure 1). Monotremes are of interest from a neurobiological perspective because it is thought that they may have retained features of brain organization that were present in the first mammals.

Of course any extant mammal is a compilation of both retained characteristics and newly derived features, and monotremes possess both general and specialized features of the body and the brain. For example, as determined by the fossil records, the first mammals appeared about 200 Ma (Figure 1), during the late Triassic or early Jurassic period, and are thought to have laid eggs; however, they were shrew-like animals that did not possess a bill, unlike the three extant monotreme species today. Thus, the distinct bill of the platypus and the unique bill and tongue of the echidna are newly derived features that are associated with navigation, mate selection, and prey capture, and these features are related to highly specialized lifestyles of these three species. This idea also holds true for features of brain organization. Studies of monotreme neocortex suggest that they have retained a number of general mammalian characteristics, as well as incorporated a number of newly derived features associated with their unique morphological specializations and behavior. This article describes the organization of somatosensory, visual, auditory, and motor cortex, as determined from architectonic analysis, electrophysiological recordings, and anatomical connection studies in two of the three extant monotreme species, the duck-billed platypus and the short-billed echidna (to our knowledge, the neocortex of western long-billed echidna has not been examined). We will compare the neocortical organization in monotremes with that of other species to determine which features are common to all mammals. Such a comparison allows us to infer the organizaton of neocortex of early mammals.

Gross Brain Morphology

The brains of the platypus and the echidna are distinctly different from each other; the platypus has a smooth, relatively small brain (Figures 2(a) and 2(b)), and the echidna has a convoluted, large brain with commensurate expansion of the cortical sheet (Figures 2(c) and 2(d)). In the echidna, the major sulci of the brain are termed the α sulcus and the β sulcus (Figure 2(c)), and much of the cortical surface is buried within the fissures (Figure 2(d)). In the platypus, sensory cortical areas occupy the caudal two-thirds of the entire cortical sheet, and motor cortex occupies much of the remaining rostral cortex. Similarly, in the echidna, sensory cortex is located in the caudal portion of the neocortex, caudal to the α sulcus, and motor cortex is located just rostral to sensory cortex, between the α and β sulci. While there is very little cortex rostral to motor cortex in the platypus, in the echidna there is an enormous expansion of the cortical sheet rostral to motor cortex, the function of which is unknown.

The size of pyriform cortex also differs between the platypus and the echidna. The pyriform cortex is separated from the neocortex by the rhinal sulcus and is located lateral (ventral) to the neocortex. In the echidna, the relative size of pyriform cortex with respect to the neocortex is very large compared to the pyriform cortex of the platypus. As in many other mammals, the olfactory bulb of the platypus is smooth and relatively small; however, the olfactory bulb in the echidna is extremely large and contains a number of fissures. The increased size of pyriform cortex and the olfactory bulbs in echidnas may be related to the expansion of the frontal lobe (discussed in the section titled 'Multimodal and frontal regions of cortex').

Somatosensory and Electrosensory Cortex

Much of what is known about the organization of the neocortex in monotremes comes from studies of the somatosensory system. This is likely due

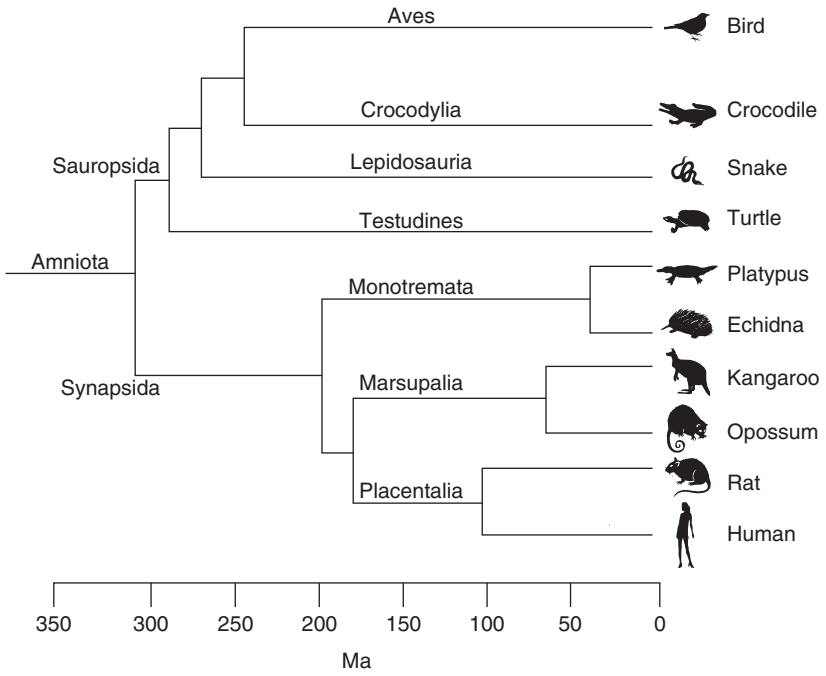

Figure 1 A cladogram depicting the phylogenetic relationships of reptiles, birds, and the three major groups of mammals, including monotremes. A large vertebrate group, the synapsids, arose more than 300 million years ago (Ma) and includes all mammals. Early synapsids are considered to be mammal-like reptiles that ultimately gave rise to the first mammals more than 200 Ma. Monotremes arose about 200 Ma, before the radiation of marsupials and placental mammals, approximately 180 Ma. Fossil records indicate that extant monotremes, which include the platypus and two species of echidnas, are highly derived, particularly in the structure of their bill, and in this respect do not resemble the first mammals.

to the unique morphological specialization of the bill, which is critical for the survival of these species and dominates the organization of the nervous system along the entire neuroaxis. Previous functional mapping studies have demonstrated that the somatosensory neocortex of both the platypus and the echidna is composed of three separate fields, a primary somatosensory area (S1 or area 3b), a rostral field (R), and a field just caudal to S1 that is in the position of the parietal ventral area (PV) in other animals (Figure 2).

The primary somatosensory area is a large field in both the platypus and the echidna and is topographically organized, with the tail and feet represented most medially, followed by the representation of the trunk, forelimbs, and face more laterally. In both the platypus and the echidna, the most striking feature of S1 is the bill representation, and in the platypus, neocortical specializations associated with their electrosensory abilities are remarkable. Specifically, the platypus has evolved electrosensory receptors that are located in a striped pattern down the bill and are interdigitated with the mechanosensory receptors, which are also arranged in stripes (Figure 3). Unlike electroreception in some electric fish, electroreception in the platypus is passive. Electrosensory receptors enable the platypus to locate electrical signals emitted from the muscle of small prey such as insects and crayfish, which generate electrical fields that can be as small as $20\,\mu V\,cm^{-1}$. Unlike other mammals, the platypus uses electrosensation combined with somatosensation (instead of vision, audition, and olfaction) for most daily activities, including navigation, mating, and feeding. Correspondingly, the bill representation comprises about 90% of S1. In cortex that has been sectioned tangentially and stained for cytochrome oxidase (CO), the bill representation of S1 stains heterogeneously, with CO light regions interdigitated with CO dark regions (Figure 3). These dark and light regions contain neurons responsive to mechanosensory and electrosensory stimulation, respectively.

Like the platypus, the echidna also has an extremely large representation of the bill in S1. However, there are substantially fewer electroreceptors in the bill of the echidna (only a few hundred to 2000) compared with the 40 000 receptors found on the platypus bill. Thus, it is likely that electroreception is not used to the same extent in the echidna as it is in the platypus. Nevertheless, the echidna does use its bill and elongated tongue when foraging for termites and other insects, and this specialization is reflected in the organization of the somatosensory cortex. Most notably, the rostral 1 cm of the bill assumes the greatest amount

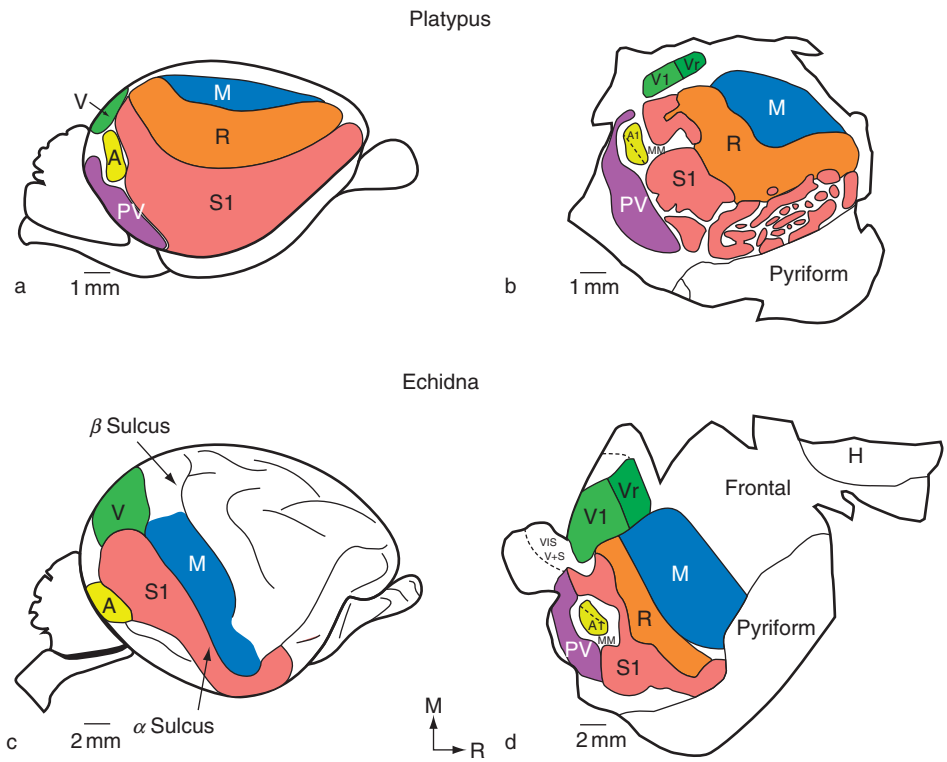

Figure 2 Gross brain morphology in the echidna and the platypus, depicting the location of sensory and motor areas (a,c), and flattened cortex depicting all the known sensory and motor areas of the cortex in each species (b,d). The echidna has a large, highly fissured, gyrencephalic brain, while the platypus has a smooth, lissencephalic brain. In the platypus, much of the entire neocortex is composed of the primary somatosensory area (S1), the rostral somatosensory field (R), the parietal ventral area (PV), and motor cortex (M). In the echidna, sensory cortex, including S1, is located caudally in the brain, and M is located between two major sulci, termed α and β. The rostral field (R) in echidnas is buried in the β sulcus. Primary sensory areas (S1, A, and V1) have been identified in both species. A small amount of cortex surrounding A has been identified as multimodal (MM). The relative amount of S1 cortex in the platypus is quite large, and most of this is occupied by the representation of the bill (see **Figure 3**). The echidna cortex is of particular interest because frontal cortex has undergone an extreme expansion, and such an expansion is not observed in the platypus. H, hippocampus; VIS, visual cortex; Vr, rostral visual field; V+S, visual and somatosensory cortex. Arrow directions M and R refer to medial and rostral, respectively.

of cortical space within the entire representation of the bill, and the receptive fields for neurons in this rostral representation are extremely small. As in the platypus, S1 in the echidna stains darkly for myelin, and architectonic analysis by Hassiotis and colleagues has demonstrated that S1 can be readily distinguished with acetylcholinesterase stains (AChE) and a monoclonal antibody reaction to neurofilament protein (SMI-32).

A second field, just rostral to S1, termed the rostral field (R), has been identified electrophysiologically and architectonically in both the platypus and the echidna. As in S1, R contains a complete representation of the contralateral body with a similar medial to lateral organization; however, neurons in R respond to stimulation of deep receptors whereas neurons in S1 respond to cutaneous stimulation. Like S1, R also contains an enlarged representation of the bill in both the platypus and the echidna. R can be distinguished from S1 architectonically because it stains lightly for myelin and CO compared with S1. Furthermore, architectonic studies show that the laminar distribution of AChE in R is distinct from that of S1.

Finally, a cortical field just caudal to S1, termed PV, has been described in both the platypus and the echidna. This field contains a complete representation of the cutaneous receptors of the contralateral body that is distinct from the representation in S1. This field is in the relative position of PV described in other mammals and has a similar topographic organization. Within PV, the tail is represented medially, the hind limb is represented caudolateral to the tail, the forelimb is represented rostrolateral to the tail, and the bill is represented lateral to these representations. Like S1 and R, there is an enormous magnification of the representation of the bill. Together, the bill representation in all three fields comprises approximately 64% of the entire cortical surface. Such magnification of a single body part is rarely seen in other mammals. PV, in both platypus and echidna

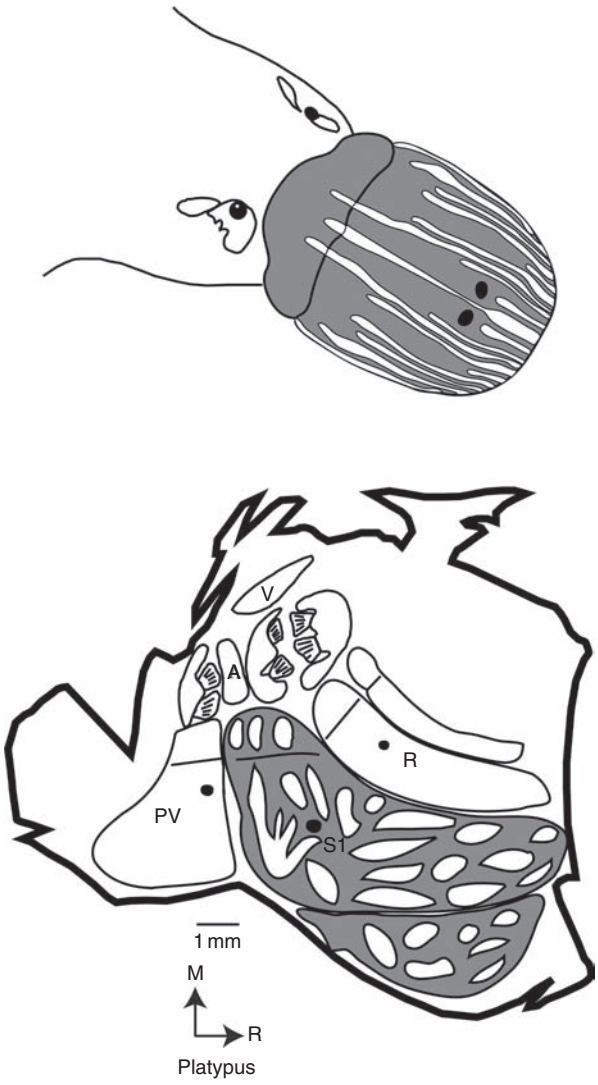

Figure 3 An illustration of the platypus (above), depicting the location of electrosensory receptors (white stripes) on its bill. The enormous size of the bill representation within the three separate somatosensory fields in the platypus neocortex (below) reflects the importance of the bill for survival behaviors. Within the primary somatosensory area (S1), mechanosensory receptors (shaded regions), interdigitate with the representation of electrosensory receptors. See **Figure 2** for abbreviations.

brain sections, stains moderately for myelin, and architectonic analysis in the echidna indicates that PV can be readily distinguished from S1 with a number of different stains, including AChE, parvalbumin, and SMI-32.

Anatomical studies have been done only in the echidna. Specifically, degeneration studies by Welker and Lende and neuroanatomical tracer studies by Ulinski have demonstrated that the major projection from the thalamus to S1 is from the ventral posterior nucleus. Cortical area R also receives input from the ventral posterior nucleus. Cortical connection experiments have shown that S1 has dense connections with R and PV, moderate connections with motor cortex, and sparse connections with visual cortex.

Taken together, the architectonic, electrophysiological, and anatomical data indicate that three somatosensory fields are present in the platypus and the echidna: S1, R, and PV. All these fields have a greatly magnified representation of the bill, and in the platypus, S1 modules associated with electrosensory reception have been identified.

Visual Cortex

Relatively little is known about the organization of visual cortex in monotremes. In the platypus, studies of visual cortex organization are rare, but the limited data that exist from electrophysiological recording studies demonstrate that two fields are present, a caudal (Vc) and rostral (Vr) visual field (**Figure 4(a)**). The caudal visual field, Vc, contains neurons that are responsive only to visual stimulation. Although the visuotopic organization of this field is not as precise as in other mammals, Vc contains a representation of both the upper and lower central visual hemifield. Because neurons in Vc respond vigorously to visual stimulation and have relatively smaller receptive fields than neurons in Vr, this field may be homologous to V1, as described in other mammals. The rostral visual field, Vr, contains a representation of the visual hemifield; however, neurons in this field have larger receptive fields and respond less vigorously to visual stimulation. Architectonic analyses of these fields indicate that both are coextensive with a darkly myelinated zone, although Vc is slightly less myelinated than Vr.

In the echidna, early evoked potential studies by Lende demonstrated that visual cortex is located just medial to somatosensory cortex, and this has been confirmed in more-recent microelectrode recording studies by Krubitzer. Cortex in which neurons are responsive to visual stimulation has been subdivided into two and sometimes three distinct fields on the basis of architectonic analysis and anatomical connections. For example, architectonic analysis by Hassiotis and colleagues is consistent with electrophysiological recording studies by Lende and by Krubitzer and demonstrates that there is a visual field representation in the position of V1 which can be readily distinguished from the surrounding fields using myelin stains and AChE reactivity. Based on architecture, these investigators also distinguish a second field rostromedial to V1 that may correspond to V2 in other mammals. Furthermore, injections placed in the caudolateral portion of this field (which may be V1 plus the rostral area as defined by Hassiotis and colleagues; and Vc plus Vr as defined in

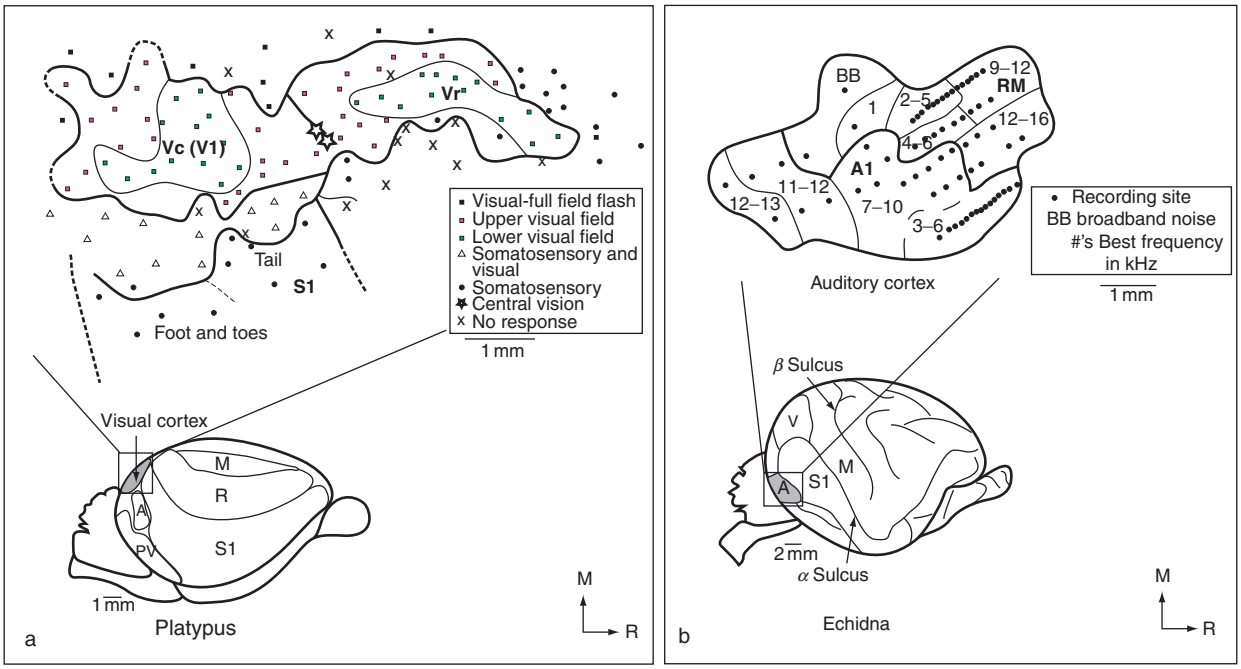

Figure 4 The functional organization of visual cortex in the platypus (a) and auditory cortex in the echidna (b). For both, the illustration at the top is an enlarged drawing of the area in the inset in the whole brain below. (a) Although the data are limited to a few platypus, there appear to be two representations of the contralateral hemifield. There is a caudal field, which appears to be homologous to V1 described in other mammals, and a rostral field, which may correspond to V2 in other mammals. On the basis of their location, these are termed Vc and Vr, respectively. (b) There appear to be two auditory fields in the echidna. In the caudolateral field, frequencies are represented from high caudally to low rostrally in isofrequency bands, much like A1 in other mammals. Because of the similarities between this field and the primary auditory area described in other mammals, we have termed this field A1. A rostromedial field (RM) is also present, in which frequencies are represented from low to high in a caudal to rostral progression. See **Figure 2** for abbreviations.

platypus) resulted in labeled cells and axon terminals immediately around the injection site, in cortex far rostral to the injection site, and in cortex lateral to the injection site. Finally, limited architectonic analysis and anatomical connection studies show that the cortex located caudal to the darkly myelinated V1, which has been previously demonstrated to contain neurons responsive to visual stimulation, has reciprocal connections with V1.

On the basis of cortical connections, architectonic analysis, and electrophysiological recordings in the platypus and the echidnas, it appears that at least two and possibly three fields exist in visual cortex. One of these fields, based on architecture, appears to be homologous to V1 described in other mammals. However, thalamocortical connections of visual cortex suggest that major differences exist between visual cortex in monotremes and other mammals.

Thalamocortical connections of visual cortex have been examined only in the echidna by Welker and Lende. Using degeneration methods, these investigators found that lesions in visual cortex resulted in degeneration of a nucleus that they termed the occipital pole region (OP) of the thalamus. This is different from other mammals in which the major source of input to V1 is the lateral geniculate nucleus (LGN). In a previous review we proposed several alternative interpretations of their result. The first is that the lesioned region may not have been V1, as other data described above indicate; at least two and possibly three fields exist in this region. A second possibility is that the LGN was misidentified, although this seems unlikely since most other investigators who examined the architecture of the thalamus in monotremes place the LGN in the same location as did Welker and Lende. Third, it is possible that OP corresponds to the pulvinar in other mammals and that connections have been altered in marsupial and placental mammals. Finally, it is possible that none of the visual areas defined in the platypus and the echidna correspond to V1 and that V1 and its unique pattern of connectivity evolved later in marsupial and placental mammals. Support for the latter two hypotheses (that OP corresponds to the pulvinar and that V1 is a more recently evolved cortical field in marsupials and placental mammals) comes from studies of retinal projections in the echidna. These studies demonstrate that the superior colliculus, rather than the poorly differentiated OP, receives the majority of input from the retina.

Taken together, the data indicate that at least two and possibly three visual fields are present in monotremes, but the connections of these fields in monotremes appear to be somewhat different from those in other groups of mammals. It is uncertain whether this difference is a reflection of a primitive form of brain organization present in the common ancestor or whether these differences represent derivations of animals that have reduced visual pathways and rely very little on vision.

Auditory Cortex

The location of auditory cortex in both the platypus and the echidna was described in early evoked potential studies by Lende and by Bohringer and Rowe, respectively, and in more-recent microelectrode recording studies by Krubitzer. In the platypus, early studies indicated that auditory cortex completely overlaps visual cortex; however, recent studies have provided evidence for a small region of cortex, just caudal to the medial portion of S1 and lateral to V1, in which neurons respond solely to auditory stimulation (**Figure 4(b)**). Architectonic analysis of this cortex demonstrates that this field, which appears to be homologous to A1, stains darkly for myelin. A small belt of cortex immediately surrounding auditory cortex in the platypus contains neurons responsive to more than one modality.

In the echidna, auditory cortex has been examined using electrophysiological recording procedures as well as architectonic analysis. Early functional mapping studies demonstrated that, as in the platypus, auditory cortex in the echidna is unimodal and is located just lateral to V1 and caudal to the medial portion of S1. In studies in which the neural response to pure tone frequencies was determined, the data show that two tonotopically organized fields are present: a rostromedial field, in which frequencies are represented from low frequencies caudally to high frequencies rostrally (**Figure 4(b)**), and a caudolateral field, in which low frequencies are represented rostrally and high frequencies are represented caudally. The internal organization of this caudal field suggests that it is homologous to A1, as described in other animals. Architectonic analyses of the auditory cortex in echidnas demonstrate that these auditory fields are darkly myelinated and that they are particularly distinct in AChE- and parvalbumin-stained tissue.

There are only a few studies of the anatomical connections of auditory cortex in monotremes, and these are limited to studies in the echidna. For example, Dann and Buhl demonstrated that auditory cortex has patchy connections with adjacent fields, which may be other auditory fields. Degeneration studies of auditory cortex by Welker and Lende reported that lesions to auditory cortex result in degeneration of posterior regions of the thalamus, in the location of the medial geniculate nucleus (MG), although they did not specifically identify the MG nucleus architectonically.

Together, the data indicate that monotremes have a separate region of cortex that is devoted to processing unimodal auditory inputs. This cortex is architectonically distinct from surrounding cortex and receives inputs from posterior portions of the thalamus, likely MG. Limited electrophysiological and connection data in monotremes suggest that within this unimodal auditory region, two fields are present, one of which may be homologous to A1 described in other mammals.

Motor Cortex

Studies of motor cortex in monotremes began more than a century ago. In these early experiments, electrical stimulation of both the platypus and the echidna neocortex produced movement maps that were roughly topographically organized from foot to forelimb to face in a mediolateral progression. Subsequent studies by Lende and by Bohringer and Rowe confirmed these early findings and provided more details about the topography of motor cortex and its relationship to sensory cortical areas. In the platypus, it appears that almost all motor cortex completely overlaps somatosensory cortex, including what would be S1, R, and PV. In contrast, in the echidna, motor cortex appears to form a separate representation that may only partially overlap R. Architectonic studies in the echidna by Ulinski have illustrated that just rostral to cortical area R is a reduced layer IV and a very thick pyramidal cell layer (layer V), similar to area 4 or motor cortex described in other mammals. A recent architectonic analysis by Hassiotis and colleagues has confirmed this and has also identified a region of transitional cortex between M and R.

There are only a few studies on the anatomical connections of motor cortex in monotremes, and these are limited to studies of thalamocortical connections in echidnas. Degeneration studies by Goldby and by Welker and Lende demonstrate that the region of cortex defined as M1 (located between the α and β sulci) receives projections from the ventral lateral nucleus (VL) of the thalamus. However, it should be noted that the region of degeneration in these lesion studies was large and included posterior portions of the dorsal thalamus other than VL. These data are consistent with the idea that at least echidnas have a separate motor

cortex that receives input from the motor nucleus of the thalamus, as in other mammals.

Taken together, the data indicate that there may be two distinct types of motor cortex organization in monotreme neocortex. The first is a sensory–motor amalgam in which somatosensory and motor cortex are completely overlapping. This type of sensory–motor amalgam appears to be present in platypus and has been previously described by Lende in some marsupials. The second type of organization consists of a separate motor representation that is distinct from somatosensory cortex. This type of organization occurs in echidnas and is commonly observed in placental mammals. As in placental mammals, the primary source of input to motor cortex in echidnas is from VL of the thalamus.

Multimodal and Frontal Regions of Cortex

An interesting feature of cortical organization in both the platypus and the echidna is that the amount of nonspecific sensory cortex that resides between known unimodal sensory areas is extremely small. In both species, the small belt of cortex lateral to V1 and caudal to S1 and surrounding A1 contains neurons that respond to auditory and somatosensory stimulation, visual and auditory stimulation, or visual and somatosensory stimulation. In most other mammals, this region is somewhat larger, even in small-brained mammals, and is occupied by multimodal cortex (MM; cortex in which neurons respond to more than one modality). In primates, the geographic location of this region (termed posterior parietal cortex) has shifted away from auditory cortex with the expansion of the cortical sheet and is dramatically expanded and contains multiple cortical fields. The observation that sensory cortex in monotremes is dominated by primary sensory areas (S1, V1, and A1) and only one or two other fields devoted to processing sensory information for each sensory system, combined with the limited extent of multimodal or association cortex, suggests that unimodal sensory cortical areas emerged early in evolution. These data in monotremes, when compared with the neocortex of marsupial and placental mammals, suggest that association cortex, including regions of posterior parietal cortex, became increasingly differentiated, and in some groups of placental mammals such as primates, it came to occupy either as much cortical territory as unimodal sensory fields or more.

One of the most remarkable observations in the neocortex of the echidna is the enormous expansion of cortex rostral to motor cortex. This cortex contains neurons that are unresponsive to any mode of sensory stimulation, at least in anesthetized animals. Based on the position of this area on the cortical sheet as well as patterns of projections from the mediodorsal nucleus of the thalamus, entorhinal cortex, amygdala, basal forebrain, pyriform cortex, substantia nigra, and the ventral tegmental areas, previous studies have proposed that this cortex is similar to prefrontal cortex in some mammals. However, the lack of the expanded frontal cortex in the platypus, as well as other mammals, including a variety of marsupial and placental mammals, which have only a very small region of cortex rostral to motor or sensory motor cortex, indicates that this expanded neocortex in the echidna arose independently and is a derivation of the echidna brain. Previously we proposed that the expansion of cortex rostral to motor cortex in echidnas might be associated with the elaboration of the olfactory system. As noted above, both pyriform cortex and the olfactory bulb in the echidna (**Figure 2**) are relatively large compared with its other brain structures and compared with these same structures in the platypus. Although olfaction has not been well studied in the echidna, it is possible that the expanded frontal cortex in the echidna is comparable to orbitofrontal cortex in other mammals and that it is related to the enhanced olfactory abilities of echidnas.

Comparisons with Other Mammals

While the organization of monotreme neocortex is interesting in and of itself, the importance of understanding cortical organization in monotremes is underscored when comparisons are made across all three mammalian lineages (monotremes, marsupials, and placentals). For example, as in other mammals, monotremes have undergone several modifications, which include a reconfiguration of sensory receptor morphology; expansions of cortex related to behaviorally relevant body parts, such as the bill representation in S1 and the expansion of frontal cortex in echidnas; and alterations in cortical connectivity. Similar modifications to the neocortex have been observed in marsupial and placental mammals, suggesting that the types of alterations made to evolving brains is highly constrained such that brain organization, even in distantly related species, often takes a similar form.

In addition to these modifications, a comparative analysis allows us to appreciate that there are a number of features of cortical organization that are similar in all three groups of mammals. First, there is a constellation of cortical fields that includes the primary sensory and motor areas (S1, V1, A1, and M1), as well as additional unimodal fields such as PV, R, Vr, and a second auditory field. Another interesting

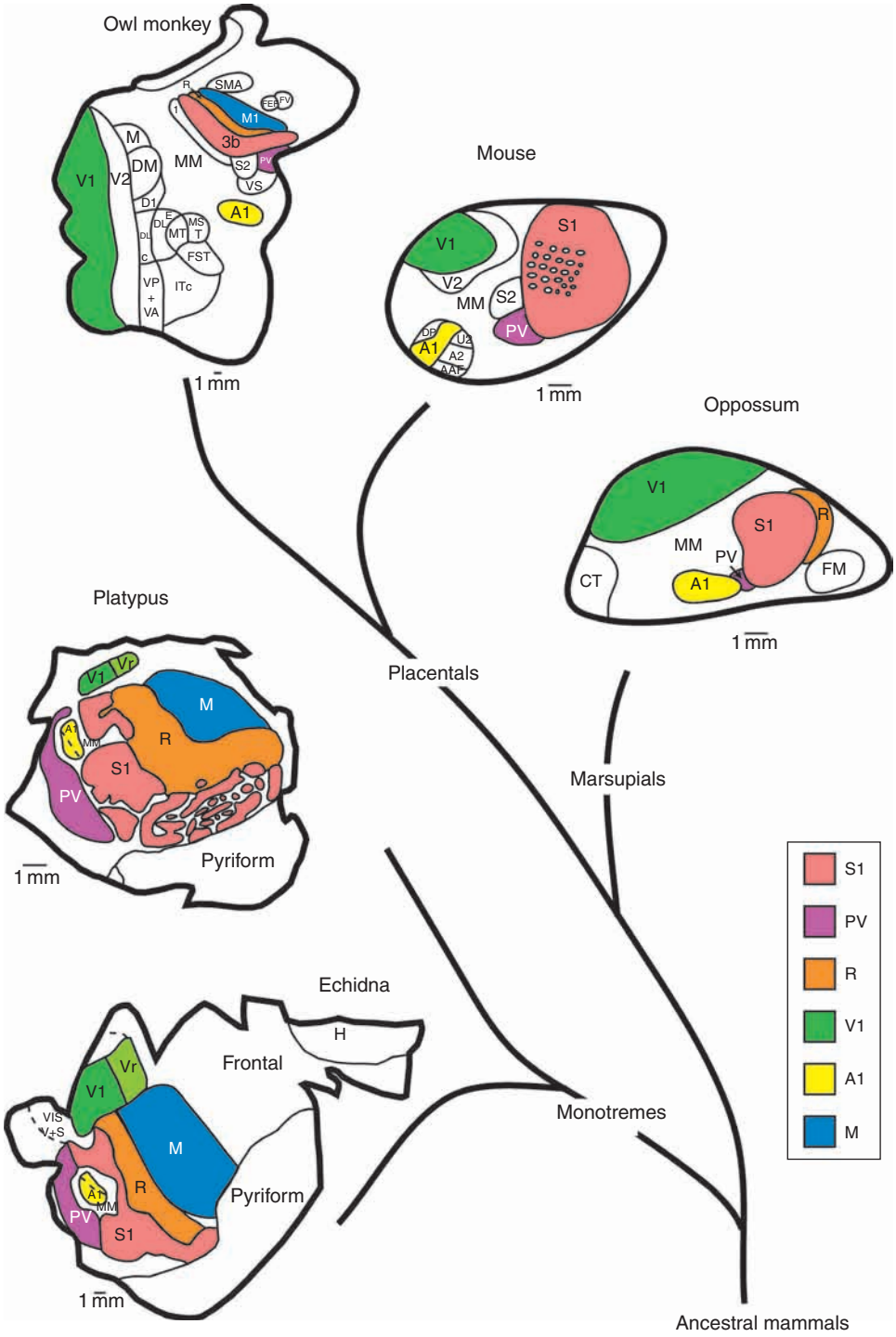

Figure 5 General phylogenetic relationship of the three major groups of mammals, and the cortical organization in representative mammals from each group. A comparative analysis indicates that all mammals have a constellation of cortical fields, including S1, R, PV, V1, A1, and two or three additional unimodal sensory areas. All mammals appear to possess either a separate motor area or a sensory–motor amalgam. Finally, all mammals have multimodal cortex. This type of analysis allows us to deduce the cortical organization of the common ancestor, which likely had these features of neocortical organization as well. Changes to the plan of cortical organization observed in different lineages include an expansion of the cortical representation of behaviorally relevant body surfaces, the expansion of the cortical sheet, the addition of new cortical fields, and alterations in connectivity. 1 and 3b are somatonsensory areas; A and A1, auditory cortex; A2, second auditory area; AAF, anterior auditory field; C, caudal; CT caudal temporal field; DI, dorsal intermediate visual area; DL, dorsal lateral visual area; DM, dorsal medial visual area; FEF, frontal eye field; FM, frontal myelinated field; FST, fundal superior temporal area; FV, Frontal ventral visual eye field; ITc, caudal inferotemporal area; M, motor cortex; M1, primary motor area; MM, multimodal cortex; PV, parietal ventral areas; R, rostal somatosensory field; S1, primary somatosensory area; SMA, supplementary motor area; S2, second somatosensory area; U2, ultrasonic area; VA and VP, divisions of the ventral visual area; V1, visual cortex; V2, second visual area; Vr, rostral visual field; VS, ventral somatosensory area.

observation that can be made by comparing the cortical organization in monotreme, marsupial, and placental mammals is that there is more than one sensory field present for any given modality in all lineages. Until recently, it was believed that only primary sensory areas existed in monotremes. This notion led to the assumption that primary cortical fields are the evolutionarily oldest fields and that second and third sensory areas were added hierarchically in marsupial and placental mammals. The presence of at least three somatosensory fields, two visual fields, and two auditory fields in all mammals investigated suggests that the first mammals had two to three sensory areas for each modality and possibly a separate motor field as well. Thus, there is no clear evidence that primary cortical areas are phylogenetically older than other unimodal sensory areas.

Finally, a comparative analysis reveals that MM is present in all mammals, although its relative size and location is altered in different species. For instance, in monotremes, MM assumes a very small proportion of the neocortex and surrounds auditory cortex (**Figure 5**). In marsupials, MM has been identified just caudal to somatosensory cortex, lateral to visual cortex, and medial to auditory cortex. In placental mammals such as primates, this cortex has undergone a dramatic expansion and is composed of a number of cortical fields, although its location between visual, somatosensory, and auditory cortex is similar to that of MM in smaller-brained mammals such as mice. The presence of MM in all mammals investigated suggests that the first mammals had a small MM which was the forerunner of the more expanded posterior parietal cortex in marsupial and, particularly, placental mammals. This suggests that the potential for cross-modal integration of sensory inputs, which ultimately leads to the planning and execution of complex overt and covert behaviors, was present in a very primitive form in the common ancestor of all mammals.

See also: Auditory Cortex Structure and Circuitry; Auditory Cortex: Models; Marsupial Neocortex; Molecular Anatomy of the Mammalian Brain; Neocortex: Origins; Visual Cortical Models of Orientation Tuning.

Further Reading

Bohringer R and Rowe M (1977) The organization of the sensory and motor areas of cerebral cortex in the platypus (*Ornithorhynchus anatinus*). *Journal of Comparative Neurology* 174: 1–14.

Dann J and Buhl E (1995) Patterns of connectivity in the neocortex of the echidna (*Tachyglossus aculeatus*). *Cerebral Cortex* 5: 363–373.

Goldby F (1939) An experimental investigation of the motor cortex and pyramidal tract of *Echidna aculeata*. *Journal of Anatomy* 73: 509–524.

Hassiotis M, Paxinos G, and Ashwell K (2004) Cyto- and chemoarchitecture of the cerebral cortex of the Australian echidna (*Tachyglossus aculeatus*). I. Areal organization. *Journal of Comparative Neurology* 475: 493–517.

Johnson J (1990) Comparative development of somatic sensory cortex. In: Jones EG and Peters A (eds.) *Cerebral Cortex*, pp. 335–449. New York: Plenum.

Krubitzer L, Manger P, Pettigrew J, and Calford M (1995) Organization of somatosensory cortex in monotremes: In search of the prototypical plan. *Journal of Comparative Neurology* 351: 261–306.

Lende R (1964) Representation in the cerebral cortex of a primitive mammal: Sensorimotor, visual, and auditory fields in the echidna (*Tachyglossus aculeatus*). *Journal of Neurophysiology* 27: 37–48.

Royal Society Publishing (1998) *Special Issue: Platypus Biology: Recent Advances and Reviews. Philosophical Transactions of the Royal Society B: Biological Sciences* 353(1372).

Scheich H, Langner G, Tidemann C, Coles R, and Guppy A (1986) Electroreception and electrolocation in platypus. *Nature* 319: 401–402.

Ulinski P (1984) Thalamic projections to the somatosensory cortex of the echidna, *Tachyglossus aculeatus*. *Journal of Comparative Neurology* 229: 153–170.

Welker W and Lende R (1980) Thalamocortical relationships in echidna (*Tachyglossus aculeatus*). In: Ebbesson SOB (ed.) *Comparative Neurology of the Telencephalon*, pp. 449–481. New York: Plenum.

Neonatal Circuits

A Lev-Tov, The Hebrew University School of Medicine, Jerusalem, Israel
M J O'Donovan, The National Institutes of Health, Bethesda, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Rhythmic patterns of coordinated movement are produced by networks of spinal neurons known as 'central pattern generators' (CPGs). These circuits have been studied using isolated spinal cord preparations that can generate a pattern of motor discharge that resembles locomotion (locomotor-like activity). The isolated spinal cord is more experimentally accessible than the equivalent *in vivo* preparation and allows network function to be studied under controlled ionic and pharmacological conditions. For technical reasons, studies of central pattern generation in isolated preparations are restricted to the neonatal period and cannot be performed on adult mammals. In this article, we first consider the general properties of the locomotor-like rhythm in the neonatal rat and mouse cords and then discuss our knowledge of the specific circuitry responsible for this behavior.

Segmental and Spatial Distribution of the Pattern-Generating Circuitry in Neonatal Spinal Cord

The hind limb locomotor CPGs of the neonatal rat are distributed along the lumbar cord with a rostrocaudal decline in the rhythmogenic capacity of each segment. In the rat and mouse, rhythmogenic activity can also be observed in isolated sacrocaudal segments (Figure 1), indicating that rhythmogenic circuitry is not unique to the limb moving segments. The sacrocaudal network controls the axial and tail musculature during locomotor and other movements. Experimental activation of this network produces alternating left/right rhythmic bursts in sacral motor neurons. The thoracolumbar and sacrocaudal networks are coupled in the rostrocaudal direction via propriospinal/descending neurons, the axons of which travel in the lateral, ventral, and ventrolateral funiculi. These connections coordinate activation of the limb, axial, and tail musculature during movement.

It has been proposed that the locomotor rhythm is generated by antagonistic 'half-centers' activated by common tonic excitation and linked by reciprocal inhibitory pathways. Consistent with this idea, intracellular recordings from lumbar motor neurons in the neonatal rat show that the locomotor drive potentials produced in lumbar motor neurons during rhythmic activity induced by bath application of *N*-methyl-D-aspartate (NMDA) and 5-hydroxytryptamine (5-HT; serotonin) are generated by an alternating sequence of phasic excitation and inhibition. However, several findings argue against the half-center model as a mechanism for spinal rhythmogenesis. For example, the locomotor-like rhythm can be generated in isolated hemicords that necessarily lack contralateral inhibitory connections. Furthermore, studies of sacrocaudal networks in the neonatal rat spinal cord have shown that reciprocal inhibitory pathways between flexor and extensor motor neurons are not activated during the sacrocaudal rhythm. As a result, flexor, extensor, and abductor tail muscles are coactive on one side of the tail during rhythmic activation of the sacrocaudal networks (Figure 1(c)). Because the sacrocaudal rhythm persists after separating the cord midsagittally, it is clear that this rhythm does not require activation of either crossed or reciprocal synaptic inhibition (Figure 1(d)).

It is now believed that the function of the contralateral inhibitory and excitatory connections is to regulate the timing of motor neuron activation. Because contralateral motor neurons alternate during locomotion it is assumed that the contralateral inhibitory connections are substantially more powerful than are the excitatory connections during normal locomotion. However, an increase in the excitatory crossed connections – such as occurs in certain mutant mice or as in the presence of inhibitory amino acid receptor blockers – can change the contralateral motor neuron activation from alternation to synchrony (see later).

Sensory Pathway Interneurons and Pattern Generation in Neonatal Spinal Cord

Studies of Lundberg and colleagues have revealed that segmental pathways activated by high-threshold flexor reflex afferents (FRAs) produced a locomotor-like alternating flexor–extensor discharge in cats subjected to spinal cord transection ('spinal cats') treated with L-3,4-dihydroxyphenylalanine (L-DOPA) and monoamine oxidase blockers. It has been suggested that these FRA pathways are actually part of the locomotor pattern generators, because stimulation of FRA components can reset the L-DOPA-induced rhythm. In addition, application of opioids to spinal

cats can block the spontaneous locomotor-like rhythm induced by L-DOPA and the alternating flexor–extensor pattern produced by stimulation of various FRA components.

Alternatively, FRA interneurons may be interposed between the locomotor CPGs and their target motor neurons. Studies of the sacrocaudal network in the neonatal rat and mouse have revealed that low- and high-intensity stimulation of sacrocaudal afferents (SCAs) can produce a coordinated activation of the locomotor and sacrocaudal CPGs (Figure 2(a)). The high accessibility of the sacrocaudal network and the extraordinary capacity of SCA pathways to produce rhythmic patterns in the lumbar and sacral cord have been recently exploited to study the role of the pathway interneurons in pattern generation.

This work has revealed that the SCA-induced locomotor rhythm involves synaptic activation of relay neurons in the sacrocaudal cord. It was found that the SCA-induced locomotor rhythm could be

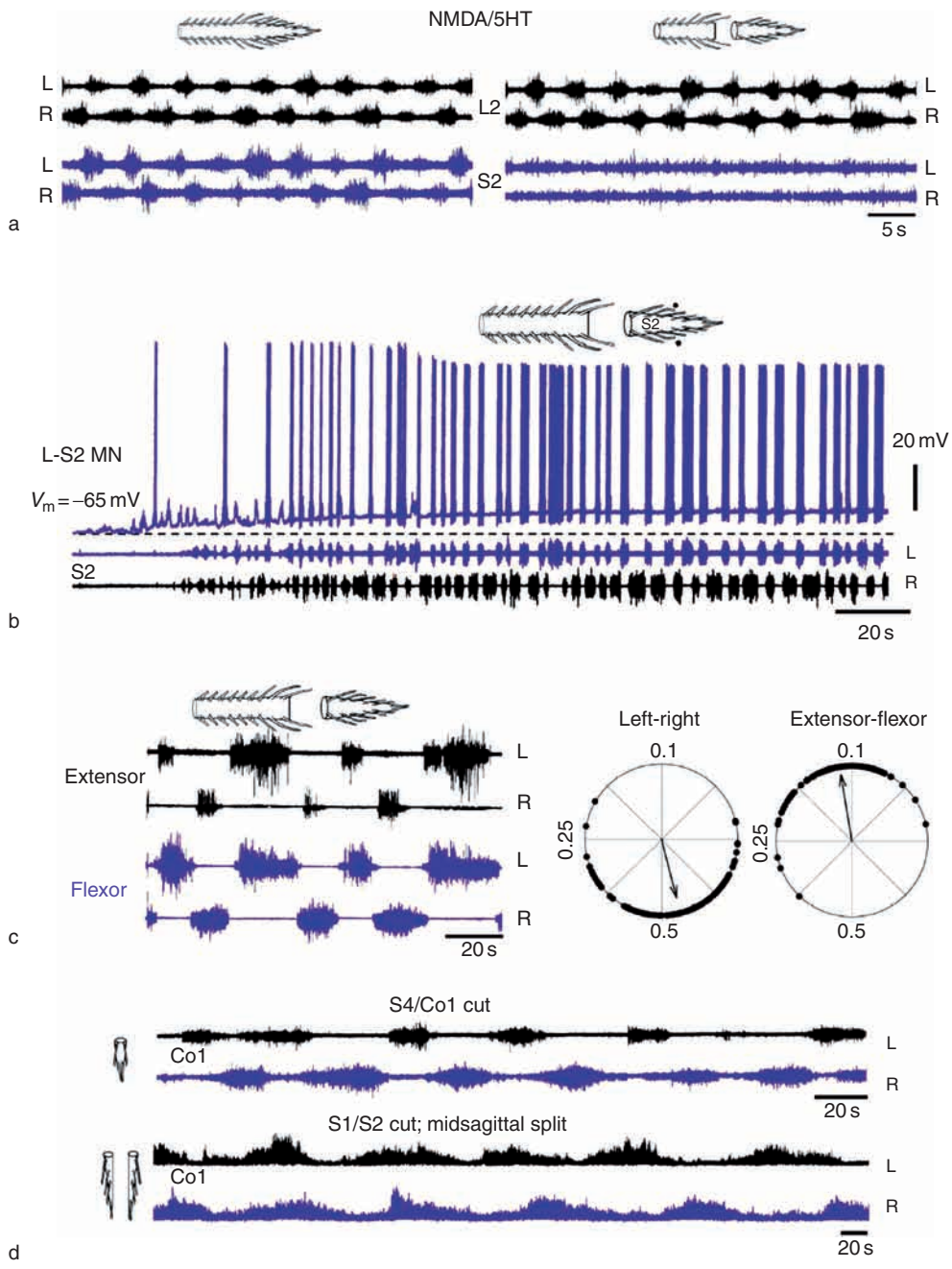

Figure 1 Continued

abolished after blocking synaptic transmission in the sacrocaudal segments by low-calcium, high-magnesium solution (**Figure 2(b)**). Moreover, specific application of μ-opioid receptor agonists to the sacrocaudal segments blocks SCA-induced rhythmic activity in the cord, although the rhythmogenic capacity of these networks is not impaired. This is because they can still be activated by bath application of NMDA/5-HT or by stimulation of the ventromedial medulla in the presence of μ-opioid receptor agonists. Physiological studies have shown that the axons of the majority of the sacrocaudal relay neurons activated by SCAs cross the spinal cord at the sacrocaudal level and ascend rostrally to the lumbar CPG mainly through the ventral funiculi (**Figures 3(a) and 3(b)**). More recent anatomical and physiological studies have indeed identified groups of sacrocaudal interneurons in the ventral horn, the intermediate zone of the gray matter, and the ventral parts of the dorsal horn (**Figure 3(c)**). These interneurons are innervated by SCAs and project their axons rostrally through the contralateral ventral funiculus. SCA stimulation activates these interneurons, an effect that is blocked by μ-opioid receptor agonists and which recovers in the presence of naloxone. It is suggested that these sensory pathway interneurons do not participate directly in the genesis of locomotion, and further studies are required to determine whether they contribute to sacrocaudal rhythmogenesis.

Specific Interneuronal Circuits

To understand the operation of locomotor networks requires identification of the constituent interneurons, their synaptic connections, and their intrinsic properties. In adult animals, spinal interneurons are generally defined by their synaptic connections. Such characterization is more difficult in the neonate because the connections may change during development and because the inputs and outputs of developing interneurons are frequently inaccessible. For these reasons, researchers have sought to identify subsets of developing neurons by their expression of various genetic markers. Among the most useful of these are combinations of transcription factors – proteins that regulate the expression of several genes. The interneuronal classes that have been identified using this approach contain the precursors of several different adult interneuron types.

We first consider the development of two types of interneurons that have been characterized in detail in the adult cat: the Renshaw cell and the 1a inhibitory interneuron. Both of these neurons have been shown to be rhythmically active during adult locomotion and are presumed to play an important role in locomotor activity. Renshaw cells, first described in cat in the 1940s, are the only known synaptic targets of motor neurons within the spinal cord. The 1a inhibitory interneurons receive direct inputs from primary muscle spindle afferents and project monosynaptically to antagonist motor neurons. Despite these differences, Renshaw cells and 1a inhibitory interneurons share many aspects of their connectivity. For example, they are each reciprocally interconnected among their own class and between each other (Renshaw–1a interneuron) and they both project monosynaptically to motor neurons.

In the neonatal mouse, shortly after birth, it has been possible to record activity from interneurons that receive direct monosynaptic connections from motor

Figure 1 Pattern generation in the isolated spinal cord of the neonatal rat. (a) Bath application of *N*-methyl-D-aspartate (NMDA) and 5-hydroxytryptamine (5HT) produces coordinated rhythmic activity in the lumbar (L2) and sacrocaudal (S2) segments of the spinal cord. The sacrocaudal rhythmicity is virtually blocked while the lumbar rhythm persists following transection of the cord at the lumbosacral junction. Recordings were obtained from the left (L) and right (R) ventral roots of L2 and S2 in preparations isolated from postnatal day 3 pups. (b) The α1 adrenoceptor agonist methoxamine (100 μM) produces rhythmic activity in the isolated sacrocaudal (S2) segments of the neonatal rat spinal cord. Intracellular recordings from a left S2 motor neuron (L-S2 MN; upper record, blue) show the development of tonic depolarization and superimposed sub- and then suprathreshold membrane potential oscillations in the presence of methoxamine. Ventral root population recordings from the left (L; blue) and right (R; black) S2 ventral roots are shown below the intracellular trace. (c) The sacrocaudal rhythmicity does not depend on activation of reciprocal inhibitory pathways between flexor and extensor centers. Electromyelogram recordings from left (L) and right (R) flexor and extensor muscles of the tail are shown following transection of the cord at the lumbosacral junction. The rhythmic pattern is characterized by alternating left/right activation of the tail muscles and by coactivation of flexor and extensor muscles within a given side of the tail. This flexor–extensor synchronicity indicates that reciprocal inhibitory pathways are not activated during the rhythm. The rhythm was induced by 10 μM noradrenaline and 3 μM NMDA. The circular diagrams (right) show the phase relation between the activities of the left and right muscles and between those of the ipsilateral extensor and flexor muscles during the rhythm, in five different experiments. (d)The sacrocaudal rhythm does not depend on activation of crossed inhibitory pathways between the left and right hemicords. Recordings from left and right ventral roots of Co1 in the isolated coccygeal cord (S4/Co1 cut) and in a midsagittally split S2/Co3 preparation (S1/S2 cut, midsagittal split; different preparation; data are rectified and integrated). Rhythmic activity was induced by 5 μM noradrenaline and 4 μM NMDA. (a) From Strauss I and Lev-Tov A, unpublished data. (b) Adapted from Gabbay H and Lev-Tov A (2004) Alpha-1 adrenoceptor agonists generate a 'fast' NMDA-receptor independent motor rhythm in the neonatal rat spinal cord. *Journal of Neurophysiology* 92: 997–1010. (c, d) Adapted from Gabbay H, Delvolvé I, and Lev-Tov A (2002) Pattern generation in caudal-lumbar and sacrococcygeal segments of the neonatal rat spinal cord. *Journal of Neurophysiology* 88: 732–739.

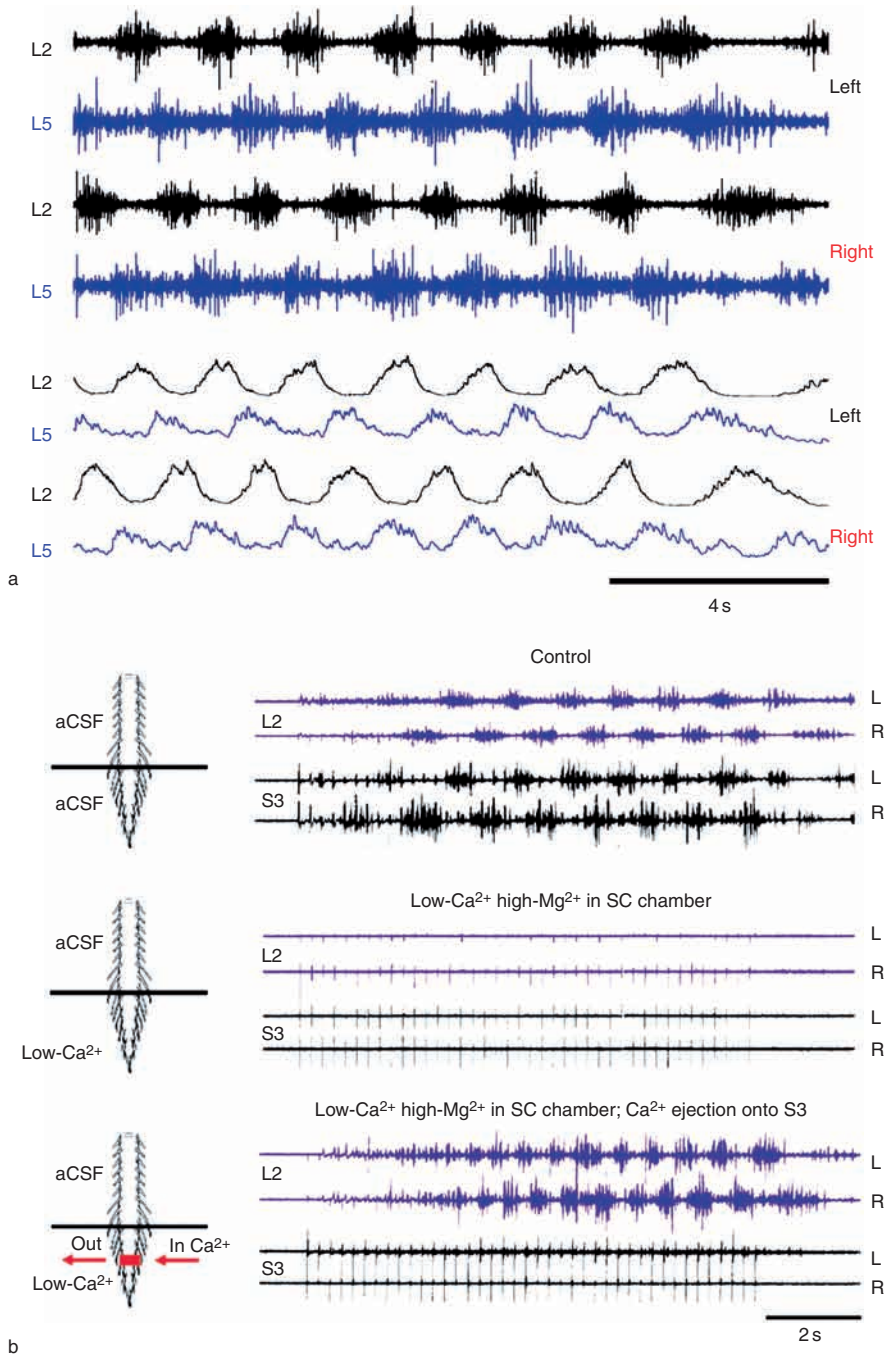

Figure 2 Sensory activation of the locomotor rhythm. (a) Recordings from the left and right ventral roots of the flexor dominating L2 (black) spinal segments and the extensor dominating L5 (blue) segments are shown during a stimulus train (50 pulse, 3 Hz) applied to the left coccygeal Co1 dorsal root at 1.8 T threshold. Upper set of recordings, 50 Hz–5 KHz; lower set of recordings, the same data following rectification and low-pass filtering at 10 Hz. Note the typical left/right alternating pattern as well as the ipsilateral alternation of flexor and extensor bursts. (b) The sacrocaudal afferent-induced locomotor rhythm is mediated by synaptic activation of sacrocaudal interneurons. Recordings of the activity produced by stimulus trains (40 pulse, 4 Hz, 2 T) applied to S4 dorsal roots, before (control; top), after bathing the sacrocaudal segments in low-calcium/high-magnesium Kreb's saline (in a sacrocaudal (SC) chamber; middle), and after repeating the stimulus train following ejection of calcium from a micropipette onto the S3 segment (bottom). The region exposed to Ca^{2+} ejection (red) was controlled using a suction pipette positioned across the ejection site. The experimental bath was divided into sacrocaudal and thoracolumbar chambers using an L6/S1 Vaseline wall. (a) From Etlin A and Lev-Tov A, unpublished data. (b) Adapted from Strauss I and Lev-Tov A (2003) Neural pathways between sacrocaudal afferents and lumbar pattern generators in neonatal rats. *Journal of Neurophysiology* 89: 773–784.

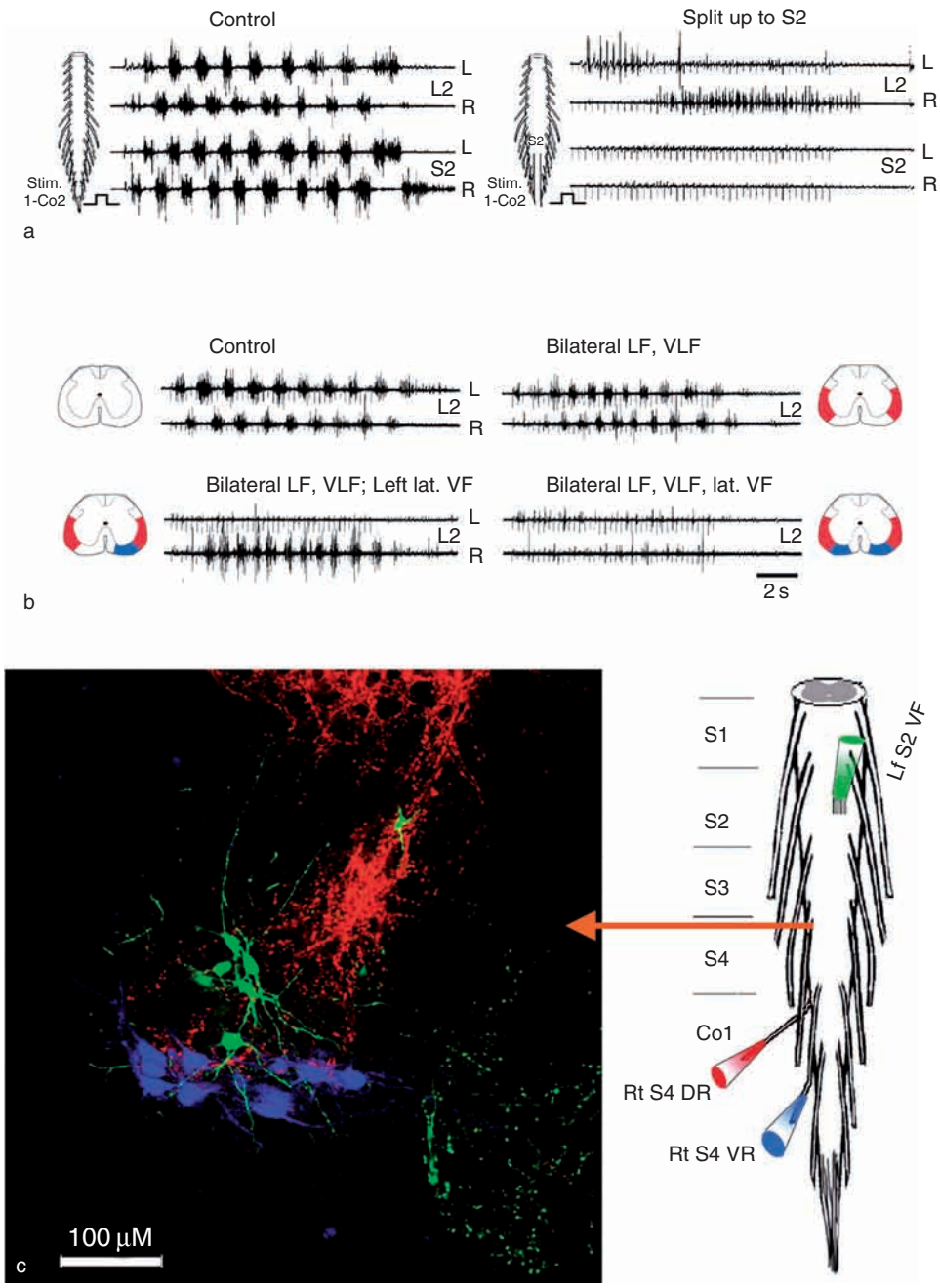

Figure 3 Sensory activation of locomotor CPGs by SCAs: the mediating pathways. (a, b) Sacrocaudal afferent stimulation activates sacrocaudal interneurons, the axons of which cross the sacrocaudal spinal cord and project rostrally through the ventral funiculi. The activity produced by stimulation of the left coccygeal Co2 dorsal root (40 pulse, 4 Hz, 1.5 T) and recorded from the left and right lumbar L2 and sacrocaudal S2 ventral roots (a; control) is severely impaired following a midsagittal split extending from the caudalmost end of the cord to caudal S2 (a; split to S2). The rhythmic activity induced in the lumbar cord by sacrocaudal afferent stimulation (40-pulse trains at 4 Hz and 1.5 T) (b; control) is virtually unaffected by bilateral lesion of the lateral (LF) and ventrolateral (VLF) funiculi at the L6–S1 junction (b; bilateral LF, VLF). Extension of the lesion to the lateral part of the left ventral funiculus (VF) blocks the activity at the ipsilateral side (b; bilateral LF, VLF; left lat. VF). Extension of the lesion to the right lateral VF (b; bilateral LF, VLF, lat. VF) virtually eliminates the rhythmic activity in the lumbar cord. (c) Sacrocaudal relay neurons and their afferent innervation. Projected confocal microscope image of a cross section at the rostral S4 segment in the neonatal mouse spinal cord. The right S4 dorsal roots (DR) were filled with Texas Red Dextran (red), the right S4 ventral roots (VR) were filled with Cascade Blue Dextran (blue), and a strip of the left ventral funiculus was filled with Fluorescein Dextran (green) at the rostral half of S2. The micrograph shows a dense ipsilateral neuropil of labeled ventral funiculus fibers with retrogradely labeled interneurons in the contralateral side of the segment. Putative contacts between the labeled interneurons and S4 afferents were found following 3-D reconstruction of the confocal images (not shown). (a, b) Adapted from Strauss I and Lev-Tov A (2003) Neural pathways between sacrocaudal afferents and lumbar pattern generators in neonatal rats. *Journal of Neurophysiology* 89: 773–784. (c) From Blivis D, Mentis GZ, O'Donovan MJ, and Lev-Tov A, unpublished data.

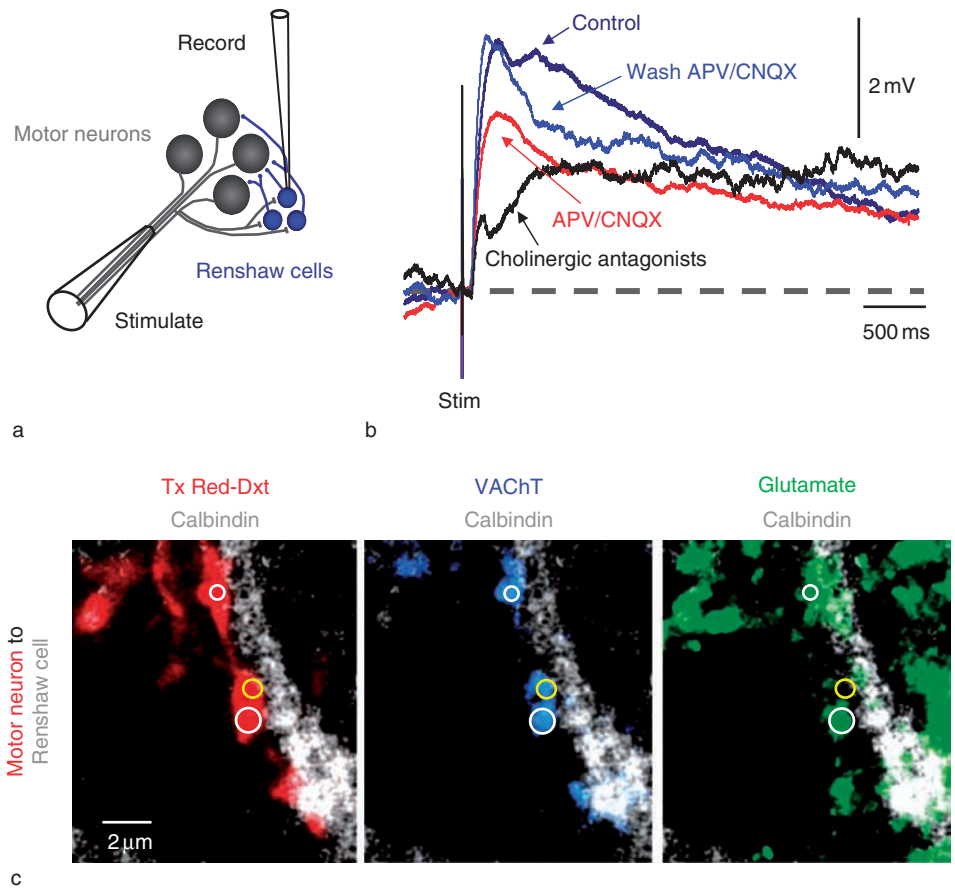

Figure 4 Synaptic potentials in a neonatal mouse Renshaw cell evoked by stimulation of motor axons are mediated by cholinergic NMDA and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid/kainate receptors. (a) Recording and stimulating arrangement. (b) Superimposed synaptic potentials recorded in a Renshaw cell in response to a single stimulus applied to motor axons of the sciatic nerve (control). The membrane potential of the cell was held at −70 mV throughout the recordings. Application of the glutamate receptor antagonists 2-amino-5-phosphonovaleric acid (APV) and 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) for 25 min depressed the amplitude of the response (APV/CNQX). The response fully recovered after 22 min of washout (wash APV/CNQX). Subsequent exposure to cholinergic antagonists depressed the amplitude of the potential. Each response is an average of five stimuli delivered at 0.1 Hz. (c) Glutamate is enriched in some motor neuron synaptic boutons contacting Renshaw cells. Confocal images of a calbindin-immunoreactive dendrite (gray) from a presumed Renshaw cell and several retrogradely labeled motor neuron varicosities labeled with Texas Red Dextran (TxRed-Dxt) and co-immunolabeled for the vesicular acetylcholine transporter (VAChT; blue) and glutamate (green). The white circles identify boutons immunoreactive for all three markers, and the yellow circles show a cholinergic motor neuron terminal without glutamate immunoreactivity. Adapted from Mentis GZ, Alvarez FJ, Bonnot A, et al. (2005) Noncholinergic excitatory actions of motoneurons in the neonatal mammalian spinal cord. *Proceedings of the National Academy of Sciences of the United States of America* 102: 7344–7349.

neurons and also from primary muscle afferents. Such cells exhibit many of the characteristics of adult Renshaw cells. These include their ventromedial location, their rapid firing following a single antidromic stimulus, and their expression of calbindin, a calcium-binding protein (see Figure 4). The number of terminals expressing a vesicular glutamate transporter specific for primary muscle afferents (VGLUT1) increases during the first three neonatal weeks and then decreases as the animal matures into adulthood. However, even in the adult mouse, some primary afferent terminals remain on putative Renshaw cells, although it has not yet been established if such terminals are physiologically active. In the neonatal mouse, the 1a inhibitory interneurons are believed to be located dorsally to the Renshaw cells but they have not yet been studied electrophysiologically. The great similarity between the synaptic connections of Renshaw cells and 1a inhibitory interneurons is perhaps not surprising given that they both derive from a common precursor population – the V1 class – of interneurons that express the *engrailed-1* transcription factor.

Collectively, findings in the adult cat and the neonatal mouse raise two possibilities: first, that the

properties of the Renshaw cell and the 1a inhibitory interneuron differ between the mouse and cat, and second, that the two interneuronal types differentiate from a common population that shares the connections of both cell classes. According to the second idea, Renshaw cells would differentiate from this precursor population by shedding their primary afferent connections whereas the future 1a inhibitory interneurons would selectively lose their motor neuronal inputs. It is also possible that these two identified interneuronal populations are not truly distinct but rather are drawn from a population continuum. At the extremes would be the classical Renshaw cell and 1a inhibitory interneuron as defined from studies of the adult cat. Between these extremes, might be a population that shares the connectivity patterns of both.

In the adult cat, the motor neuronal synaptic inputs to Renshaw cells were believed to be mediated entirely by the neurotransmitter acetylcholine. In the neonatal mouse, however, it has been found that the synaptic potentials produced in putative Renshaw cells by stimulation of motor neuron axons are mediated by cholinergic and glutamatergic/aspartergic neurotransmission (Figure 4). Such dual transmission has been observed in several different cell types in the developing nervous system, and its discovery has challenged the long-held view that motor neuron actions are exclusively cholinergic. At present, the function of this dual transmission is unknown. Clues to its possible role come from the observation that stimulation of the ventral roots in the neonatal mouse can trigger locomotor-like activity that persists in the presence of cholinergic antagonists but which is blocked by glutamatergic antagonists (Figure 5). This observation raises the possibility that neonatal motor neurons may make glutamatergic/aspartergic synaptic contacts with interneurons that have access to the locomotor central pattern generator.

Interneurons Involved in the Patterning of Motor Neuron Activity during Locomotion

In the past 10 years, several classes of developing spinal neurons have been identified based on their expression of transcription factors. As discussed in the previous section, it is not known how these developing neurons mature into the interneurons of the adult cord. Nonetheless, the participation of these primitive interneuronal types in the locomotor-like activity expressed by the neonatal cord has been studied by several groups. In considering this work, it is important to distinguish between two distinct aspects of locomotor behavior. The first is the rhythmicity exhibited by both motor neurons and interneurons, about which very little is known. The second is the timing of activity in the various motor neuron pools that innervate the limbs, the axial muscles, and the tail. For example, during locomotor-like activity both antagonistic and contralateral motor neurons alternate their discharge during each cycle whereas synergistic motor neurons fire bursts at the same time. Most studies of developing interneurons have focused on their contribution to the timing of motor neuron discharge.

One of the first interneuronal classes to be identified genetically was a group of ventrally located neurons expressing the ephrin-A4 (EphA4) receptor. This receptor and its ligand ephrin-B3 are thought to prevent axons from crossing the midline. Animals with a mutation in the EphA4 receptor exhibit an abnormal gait in which both hind limbs are activated in synchrony (hopping) rather than in alternation. The spinal neurons expressing the EphA4 are thought to be excitatory glutamatergic neurons, the contralateral projections of which are abnormally profuse in the mutant animals. More recently, it has been shown that cells expressing the EphA4 receptor form a heterogeneous population, although some of them fire rhythmically during locomotion and have direct glutamatergic projections to ipsilateral motor neurons. A second class of excitatory interneuron has been identified; it located in the lower thoracic and upper lumbar segments and expresses the homeobox gene HB9 transcription factor. This factor was known to be involved in consolidating motor neuron fate, so it was surprising to find it expressed in interneurons. Nonetheless, the $HB9^{+}$ interneurons have been shown to fire in phase with ipsilateral flexor motor neurons during locomotor-like activity. Dual recordings from these cells have shown that they are connected by electrical synapses that contribute to their synchronization during a flexor burst.

It is important to emphasize that experiments of this type do not reveal the functional role of the interneurons during locomotion. One way to approach this problem is to study locomotor deficits when particular interneurons have been prevented from developing using genetic means. This has been accomplished in mice by knocking out the homeobox DBX1 transcription factor, which results in the loss of the V0 class of interneuronal precursors. These are inhibitory interneurons that send their axons across the midline of the spinal cord. Animals lacking V0 interneurons have defects of left/right alternation suggesting that some of the V0 interneurons are involved in regulating the timing of contralateral motor neuron activity during locomotion. The same approach has been used to study a second class of inhibitory interneurons with projections that are

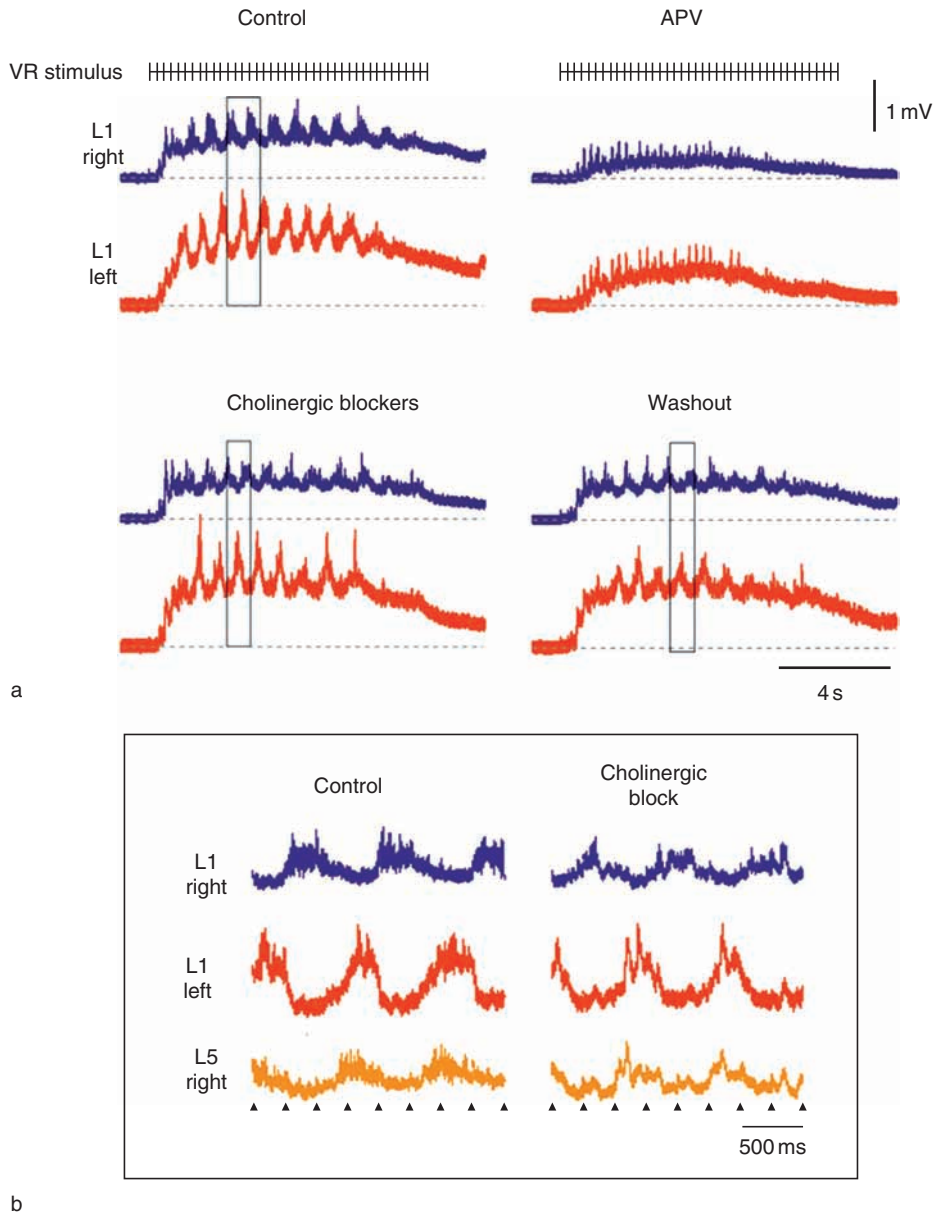

Figure 5 Locomotor-like activity in the neonatal mouse (P3) can be triggered by motor axon stimulation. (a) Ventral root recordings (L1 right and L1 left) during an episode locomotor-like activity induced by a train of stimuli applied to the motor axons of the sciatic nerve (dorsal roots cut). The recordings were made before (control), after 1 h of cholinergic block (mecamylamine 50μm, DHβE 50μm, atropine 5μm) followed by the NMDA receptor antagonist APV (100 μm) and washout of all drugs. The stimuli are shown at the top of the traces. Note that the alternating locomotor-like rhythm persists in the presence of cholinergic antagonists (see expanded traces in (b)) but is abolished by APV. (b) Expanded timescale display of the traces shown in (a), and of the simultaneously recorded L5 activity (not shown in (a)). Adapted from Mentis GZ, Alvarez FJ, Bonnot A, et al. (2005) Noncholinergic excitatory actions of motoneurons in the neonatal mammalian spinal cord. *Proceedings of the National Academy of Sciences of the United States of America* 102: 7344–7349.

believed to be exclusively ipsilateral. Knockout of this class (V1) of interneuron results in a slowing of the locomotor rhythm induced by drugs.

Intracellular recordings from contralaterally projecting commissural interneurons located in the ventromedial part of the lumbar cord have revealed a heterogeneous set of firing patterns and connectivity. Included in this population are cells projecting monosynaptic excitation or inhibition to contralateral motor neurons and other cells with connectivity that changes from rest to locomotor activity. It is likely that these recordings included some of the cells derived from the V0 category that have been implicated in the control of left/right alternation.

See also: Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Motor Primitives; Pattern Generation; Primate Interneurons; Sensorimotor Plasticity and Control of Movement Following Spinal Cord Injury +D18; Spinal Motor Neurons: Properties.

Further Reading

Alvarez FJ, Jonas PC, Sapir T, et al. (2005) Postnatal phenotype and localization of spinal cord V1 derived interneurons. *Journal of Comparative Neurology* 493: 177–192.

Bonnot A, Whelan PJ, Mentis GZ, et al. (2002) Spatiotemporal pattern of motoneuron activation in the rostral lumbar and the sacral segments during locomotor-like activity in the neonatal mouse spinal cord. *Journal of Neuroscience* 22(RC203): 1–6.

Butt SJ and Kiehn O (2003) Functional identification of interneurons responsible for left–right coordination of hindlimbs in mammals. *Neuron* 38: 953–963.

Butt SJ, Lundfald L, and Kiehn O (2005) EphA4 defines a class of excitatory locomotor-related interneurons. *Proceedings of the National Academy of Sciences of the United States of America* 102: 14098–14103.

Cazalets JR, SqalliHoussaini Y, and Clarac F (1992) Activation of the central pattern generators for locomotion by serotonin and excitatory amino acids in neonatal rat. *Journal of Physiology (London)* 455: 187–204.

Cowley KC and Schmidt BJ (1997) Regional distribution of the locomotor pattern-generating network in the neonatal rat spinal cord. *Journal of Neurophysiology* 77: 247–259.

Gabbay H, Delvolvé I, and Lev-Tov A (2002) Pattern generation in caudal-lumbar and sacrococcygeal segments of the neonatal rat spinal cord. *Journal of Neurophysiology* 88: 732–739.

Gabbay H and Lev-Tov A (2004) Alpha-1 adrenoceptor agonists generate a 'fast' NMDA-receptor independent motor rhythm in the neonatal rat spinal cord. *Journal of Neurophysiology* 92: 997–1010.

Gosgnach S, Lanuza GM, Butt SJ, et al. (2006) V1 spinal neurons regulate the speed of vertebrate locomotor outputs. *Nature* 440: 215–219.

Grillner S and Zangger P (1979) On the central generation of locomotion in the low spinal cat. *Experimental Brain Research* 34: 241–261.

Hinckley CA, Hartley R, Wu L, et al. (2005) Locomotor-like rhythms in a genetically distinct cluster of interneurons in the mammalian spinal cord. *Journal of Neurophysiology* 93: 1439–1449.

Hultborn H, Conway BA, Gossard JP, et al. (1998) How do we approach the locomotor network in the mammalian spinal cord? *Annals of the New York Academy of Science* 860: 70–82.

Jankowska E, Jukes MG, Lund S, et al. (1967) The effect of DOPA on the spinal cord. 6. Half-centre organization of interneurones transmitting effects from the flexor reflex afferents. *Acta Physiologica Scandinavica* 70: 389–402.

Kjaerulff O and Kiehn O (1996) Distribution of networks generating and coordinating locomotor activity in the neonatal rat spinal cord *in vitro*: A lesion study. *Journal of Neuroscience* 16: 5777–5794.

Kremer E and Lev-Tov A (1997) Localization of the spinal network associated with generation of hindlimb locomotion in the neonatal rat and organization of its transverse coupling system. *Journal of Neurophysiology* 77: 1155–1170.

Kudo N and Yamada T (1987) *N*-Methyl-D,L-aspartate-induced locomotor activity in a spinal cord-hindlimb muscles preparation of the newborn rat studied *in vitro*. *Neuroscience Letters* 75: 43–48.

Kullander K, Butt SJ, Lebret JM, et al. (2003) Role of EphA4 and EphrinB3 in local neuronal circuits that control walking. *Science* 299: 1889–1892.

Lanuza GM, Gosgnach S, Pierani A, et al. (2004) Genetic identification of spinal interneurons that coordinate left–right locomotor activity necessary for walking movements. *Neuron* 42: 375–386.

Lev-Tov A, Delvolvé I, and Kremer E (2000) Sacrocaudal afferents induce rhythmic efferent bursting in isolated spinal cords of neonatal rats. *Journal of Neurophysiology* 83: 888–894.

Mentis GZ, Alvarez FJ, Bonnot A, et al. (2005) Noncholinergic excitatory actions of motoneurons in the neonatal mammalian spinal cord. *Proceedings of the National Academy of Sciences of the United States of America* 102: 7344–7349.

Smith JC, Feldman JL, and Schmidt BJ (1988) Neural mechanisms generating locomotion studied in mammalian brain stem–spinal cord *in vitro*. *FASEB Journal* 2: 2283–2288.

Strauss I and Lev-Tov A (2003) Neural pathways between sacrocaudal afferents and lumbar pattern generators in neonatal rats. *Journal of Neurophysiology* 89: 773–784.

Tanabe Y and Jessell TM (1996) Diversity and pattern in the developing spinal cord. *Science* 274: 1115–1123; (1997) erratum. *Science* 276: 21.

Wenner P, O'Donovan MJ, and Matise MP (2000) Topographical and physiological characterization of interneurons that express engrailed-1 in the embryonic chick spinal cord. *Journal of Neurophysiology* 84: 2651–2657.

Nerve Growth Factor

J C Petruska and L M Mendell, State University of New York at Stony Brook, Stony Brook, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Early History

The discovery of nerve growth factor has its origin in studies of the relationship between the number of sensory and motor neurons and the size of the peripheral tissue to be innervated. In the course of these investigations, studies carried out in culture indicated that a diffusible factor from mouse sarcomas could elicit substantial outgrowth of fibers from sensory and sympathetic ganglia. Further experiments demonstrated that nerve fiber growth-promoting factors were located in snake venom and also in mouse salivary glands, the mammalian homolog of the reptilian glands secreting the active venom. The factor responsible for this effect was called nerve growth factor (NGF). It was eventually purified and found to exist in two forms called NGF 7S (S is sedimentation constant) and NGF 2.5S. These two different forms are the consequence of being composed of different peptide chains: NGF 7S (MW 130 kDa) consists of two copies of each of three chains (α, β, and γ). NGF 2.5S (MW 26 kDa) contains only the β chains and is the active form of NGF isolated from mouse salivary glands.

Neurotrophin Family

It is now known that NGF is a member of a family of growth factors that collectively are referred to as neurotrophins. Other members of the neurotrophin family are brain-derived neurotrophic factor (BDNF), neurotrophin-3 (NT-3), and neurotrophin-4/5 (NT-4/5 was so named because in the early days it was known as NT-4 or NT-5). These molecules exhibit a homology of approximately 55% with NGF.

A precursor form of neurotrophins (proNGF, pro-BDNF, and proNT-3) has been identified in neural tissues *in vivo*. ProNGF has been shown to activate the p75 receptor and thus elicit neural apoptosis.

Neurotrophin Receptors

NGF exerts its effects by activating two transmembrane receptors. One of these is the trkA receptor, a member of the tropomyosin receptor kinase (trk) family. The other is the p75 receptor. These are often called the high-affinity and low-affinity receptor, respectively; however, the differences relate more to kinetics of binding than to affinity, which is in the nanomolar range for both receptors. There is evidence that trkA and p75 function cooperatively to mediate the cellular effects of NGF, with p75 acting as a chaperone to improve the binding to trkA. Structural studies suggest that the interaction is more likely to take place within the cell via convergence of signaling pathways from trkA and p75. This issue remains to be definitively resolved. However, the different intracellular signaling elicited by activated p75 and trkA suggests that NGF could elicit some effects via independent pathways.

Members of the neurotrophin family signal largely through different trk receptors: NGF, trkA; BDNF, trkB; NT-3, trkC; and NT-4/5, trkB. However, they all signal via the p75 receptor. Therefore, it is the presence of specific trk receptor types that determines which neurotrophin can be active in any particular tissue. Trk receptors can also be present in a truncated form with the cytoplasmic end absent. Neurotrophins bind to such receptors but elicit no cellular action, effectively diminishing their functional effect.

Neurotrophin Signaling

NGF exerts its trkA-mediated effects by becoming internalized and activating diverse signaling pathways. In many situations, NGF is internalized in the axon terminals and is transported to the cell body where the cell signaling takes place. This requires specialized cellular machinery. It has been suggested that clathrin plays a crucial role in internalizing the NGF/trkA complex into endosomes that are transported to the cell body. Another suggestion is that a novel membrane protein named Pincher is responsible for internalizing NGF/trkA complex via a process known as macroendocytosis which has properties of both macropinocytosis and endocytosis.

Reports have also indicated that the neurotrophin NT-3 can signal via trkA with a molecular mechanism different from that mediating NGF/trkA signaling. The NT-3/trkA complex is not internalized, unlike NGF/trkA. It activates local membrane signaling, which is believed to play a role in determining local axonal growth, rather than the cell body signaling elicited by the transported NGF/trkA complex which promotes survival and differentiation.

TrkA Signaling

The trkA receptor is a 140 kDa glycoprotein which binds NGF. The NGF-activated trkA receptor undergoes dimerization and autophosphorylation at several tyrosine residues that selectively trigger activity

in several intracellular signaling pathways via binding of specific effector proteins to phosphorylated docking sites. Phosphorylation at Tyr751 activates PI3-kinase and the Akt protein kinase, which has been associated with NGF-induced cell survival. Similarly, phosphorylation at Tyr490 is required for association of the Shc adapter proteins that triggers the Ras–MAP kinase signaling cascade and nuclear transcription factors such as CREB and c-fos. The resulting changes in gene expression elicit differentiation and neurite growth. Phosphorylation at Tyr785 activates phospholipase C-γ, causing release of intracellular Ca^{2+} and activation of diacyl glycerol via IP_3. This leads to increased levels of protein kinase C (PKC).

p75 Signaling

Activation of p75 leads to the activation of numerous intracellular pathways, notably ones promoting apoptosis and cell survival. This suggests that the balance of activity between these two pathways is of crucial importance. In addition, activation of the p75 receptor also activates the enzyme sphingomyelinase promoting synthesis of ceramide. This pathway may be responsible for an independent p75-mediated action on cell physiology.

NGF and Development

Axon Growth

NGF promotes fiber outgrowth of sensory neurons cultured from embryonic mouse. *In vitro* studies suggest that NGF elicits its effects on axonal growth by activating PI3K in growth cones. This signaling pathway enhances actin polymerization. It also inactivates glycogen synthase kinase 3β, leading to upregulation of adenomatous polyposis coli (APC), a microtubule-binding protein. Blockade of NGF signaling removes APC from growth cones and results in their collapse.

Cell Survival

The neurotrophic hypothesis was advanced to describe the effect of neurotrophins on neuronal survival. It states that neurotrophins are required for survival of neurons innervating peripheral tissues but that they are present there in limiting amounts. Sensory, motor, and autonomic cells grow into peripheral tissue and compete for the appropriate trophic molecule. Sensory and sympathetic neurons which express trkA compete for the available NGF. The 'winners' survive; the 'losers' die via apoptosis. This is believed to be an important mechanism for matching the number of central neurons to the size of the periphery to be innervated.

Early studies in the dorsal root ganglion (DRG) demonstrated that prenatal depletion of NGF using an antibody to NGF results in reduction of cell number in the ganglion. The loss of cells was shown to be limited to small-diameter cells identified as nociceptors because it was largely restricted to cells expressing peptides known to be involved in nociception (substance P (SP) and calcitonin gene-related peptide (CGRP)). Other work has shown that cells dying as a consequence of NGF depletion project to lamina II, the known termination site of nociceptive afferent fibers. *In vivo* studies suggest that the effect of NGF on DRG cell number in rats lasts until approximately postnatal day 2.

Postnatal Development

NGF also has developmental effects on nociceptors beyond postnatal day 2 since treatment of neonatal rats with an antibody against NGF depletes Aδ high-threshold mechanonocieptors (HTMs) in the sensory nervous system. This depletion is most profound when NGF is antagonized between postnatal days 4 and 11, and it is accompanied by an increase in the proportion of Aδ low-threshold mechanoreceptors known as D-hairs. The proportion of another population of Aδ receptors innervating deep tissue is unaffected, suggesting that the effect of NGF is not to simply kill nociceptors since all remaining populations of cells would be expected to increase in number. Thus, Aδ HTMs require NGF available in the skin in order to undergo final differentiation to HTMs. If epidermal NGF is not available, these terminals migrate to the dermis to innervate D-hair follicles. This change has been referred to as a phenotypic switch.

The cell loss of sensory neurons related to disrupting the NGF/trkA system during development was recognized to be much greater than would have been expected based on the numbers of neurons expressing trkA and responsive to NGF in adult animals. Subsequent work demonstrated that although trkA is expressed by virtually all small- (and many medium-) diameter sensory neurons up to early postnatal development, by 4 weeks of age this expression pattern changed dramatically such that one-third to one-half of small-diameter neurons replaced trkA expression with components of the glial-derived neurotrophic factor (GDNF) receptor complex (GFR-α and c-Ret). Neurons that convert from trkA to ret expression lack p75 from early postnatal times (**Figure 1**) and become the nonpeptidergic small-diameter sensory neurons (in that they lack CGRP and SP, although some express somatostatin). These GDNF-sensitive nociceptive neurons have spinal terminations in the inner portion of lamina II, distinct from the superficial portion of lamina II where trkA-positive nociceptors

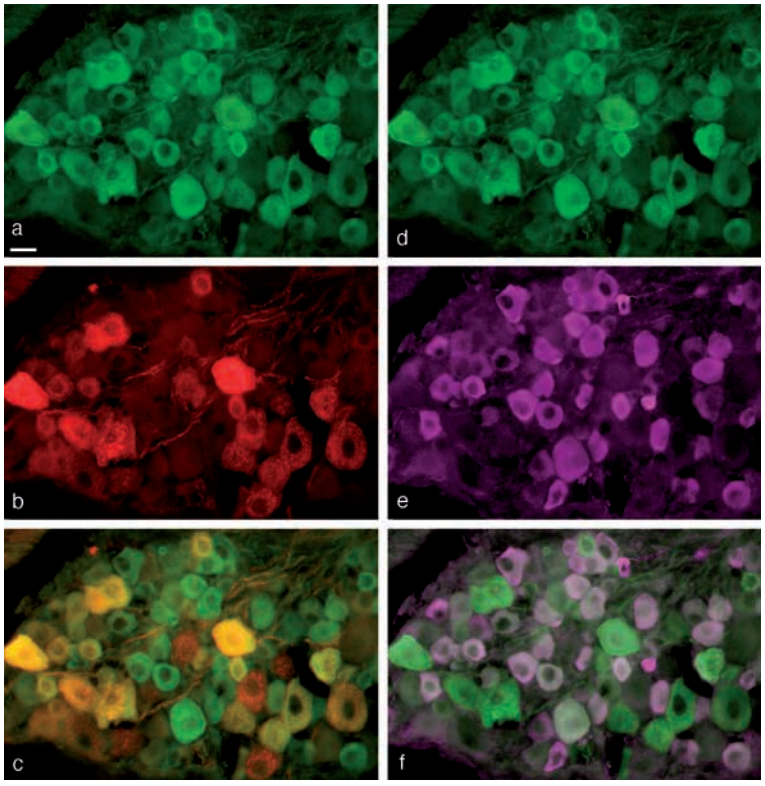

Figure 1 Section of P10 rat lumbar DRG demonstrating the distribution of different receptors using immunohistochemistry; trkA (green; a, d), p75 (red; b), and TRPV1 (purple; e). Bottom panels show the combined labels; trkA/p75 (c) and trkA/TRPV1 (f). Cells that were trkA+ but p75 (green; c) will lose their trkA expression and develop c-ret/GFR-α expression as described in the text. TRPV1-expressing cells also express trkA (whitish cells; f) throughout development. Based on unpublished data from A Davenport, JC Petruska, and LM Mendell.

terminate. They also differ in that most trkA+ sensory neurons do not bind the lectin IB4.

Developmental Effects of Other Neurotrophins

The neurotrophin NT-3 has been demonstrated to have effects on the development of large-diameter muscle afferent fibers known to express its high-affinity receptor, trkC. Depriving the developing rodent of a functional NT-3/trkC signaling system results in animals with an absence of sensory fiber projections into the ventral horn. This occurs primarily because the spindle afferent fibers responsible for these projections fail to survive. NT-3 also has been shown to be an important factor in promoting the final stages of muscle spindle axon elongation into the ventral horn to synapse with motor neurons. Furthermore, NT-3 is required for postnatal development of certain cutaneous receptors, specifically D-hairs and the Merkel cells associated with type I slowly adapting mechanoreceptors. Interestingly, BDNF appears to regulate the sensitivity of slowly adapting mechanoreceptors in the absence of any effect on Merkel cells; that is, the effect may be on the properties of the mechanosensitive channels in the axon terminals. However, NT4/5, like BDNF, a ligand for the trkB receptor, is required for the survival of D-hair receptors (i.e., the two trkB ligands can elicit quite different actions).

Biological Effects of NGF *In Vivo*: Nociception

It has become clear in the past few years that administration of NGF in the adult results in painful sequelae. It is now appreciated that this results from the binding of NGF to trkA receptors, although a possible independent role for the p75 receptor has also been demonstrated. The administration of NGF does not elicit spontaneous pain; rather, it sensitizes subsequent responses to nociceptive inputs (Figure 2). The duration of the sensitization in experimental models depends on the mode of administration of the NGF. Systemic administration of NGF results in sensitization lasting several days, whereas the effects

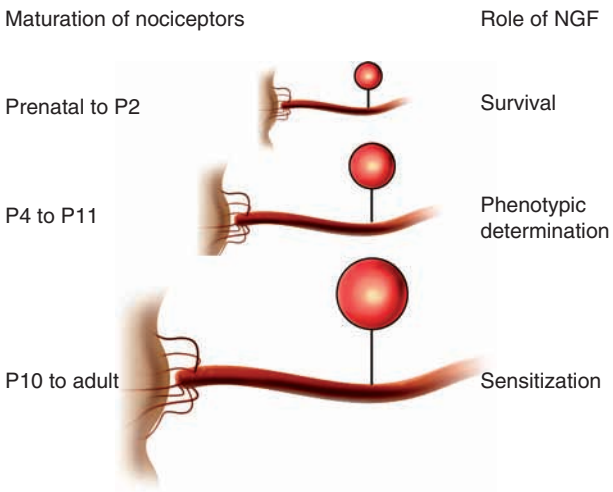

Figure 2 Differing roles of NGF on nociceptors during development. Data obtained from rat.

of local administration last on the order of hours to 1 day. The onset of hyperalgesia occurs within minutes of administration, with thermal hyperalgesia occurring within 5 min. Mechanical hyperalgesia has been reported to begin several hours after administration, but in some behavioral studies it has been found to begin within 10 min (i.e., only slightly longer than the latency for thermal hyperalgesia). This is an important issue since the thermal hyperalgesia is generally ascribed to a direct peripheral action of NGF on sensory neurons. The locus of sensitization of mechanical stimulation is not well documented; a delayed response might be attributed to an indirect central action of NGF, whereas a rapid response might indicate a direct effect in the peripheral terminals.

The nociceptive effects of NGF administration have been observed in animal models in which they are inferred from changes in the response to aversive stimuli; for example, the Hargreaves test, which measures putative nociceptive effects as a decrease in the latency of limb withdrawal from a standard noxious heat stimulus. However, more definitive evidence of the nociceptive effects of NGF administration has been obtained in humans as an unanticipated outcome of a phase I clinical trial for the use of NGF to counter the degenerative changes considered to contribute to the symptoms of Alzheimer's disease. This trial had to be terminated early because of the pain experienced by the otherwise healthy subjects when NGF was administered systemically. In a subsequent study, NGF was administered locally to human volunteers and was also found to be associated with pain. These human studies provide important direct proof of the nociceptive effects of NGF.

Mechanisms of Acute NGF-Induced Sensitization of Nociceptors

Via trkA Although the rapid sensitizing effects elicited by local injections of NGF indicate a peripheral site of action, these studies cannot definitively prove that such a mechanism exists, nor do they indicate which cell types in the periphery are involved. Numerous cell types in the periphery are endowed with trkA receptors (e.g., mast cells and autonomic efferents). The effects of NGF can be reduced or even abolished by inactivation or elimination of sympathetic efferents and mast cells. Additionally, neutrophils have been shown to contribute to NGF-induced hyperalgesia. The relative contribution of these different cell types to NGF-elicited hyperalgesia is not known and their action may be coordinated. For example, mast cells contain NGF as well as expressing trkA receptors on their surface membrane, raising the possibility that NGF activation of mast cells causes an explosive release of NGF due to recurrent action of NGF (i.e., positive feedback). TrkA receptors are also well-known to be expressed on nociceptors themselves. These are also likely to contribute to the sensitization by NGF since numerous investigators have demonstrated that exposure of dissociated small-diameter sensory neurons to NGF acutely enhances their response to capsaicin or noxious heat (measured using Ca^{2+} imaging or patch clamp recording) (Figure 3(a)). The enhanced response of TRPV1 is likely due to its phosphorylation by various intracellular messengers activated by NGF.

There is evidence that each of the intermediates mentioned previously in the general discussion of signaling pathways activated by NGF participates in sensitizing the response of TRPV1 to noxious heat or capsaicin. However, these studies, carried out under different conditions with different methods (patch clamping, Ca^{2+} imaging, and behavior), suggest that the signaling depends very much on the conditions of the experiment. Thus, the effectors activated by these signaling pathways appear to switch from axon growth and survival to sensitization after the animal is born, a proposition confirmed directly by the finding that NGF does not sensitize the response of TRPV1 to noxious heat in trkA-expressing cells on postnatal day 2, despite the fact that the two receptors are co-expressed.

Via p75 Evidence suggests that NGF can acutely influence the response of nociceptors by its action on the p75 receptor. As noted previously, NGF acting through the p75 receptor elicits enhanced levels of ceramide, which has been demonstrated to increase

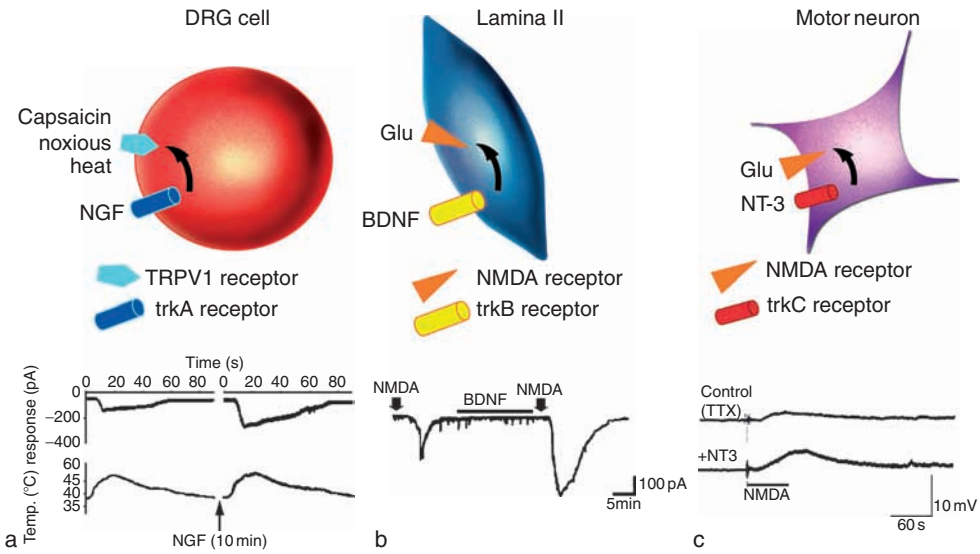

Figure 3 Examples of acute sensitization by neurotrophins: cell type, trk receptors, and receptors sensitized by neurotrophin action are displayed schematically (top). (a) The response (inward current, top traces) of a trkA-expressing cell (recorded in voltage clamp using dissociated DRG cell patch clamp) to noxious heat (lower traces) is acutely sensitized by NGF. (b) The response (inward current) of a lamina II cell (recorded in voltage clamp using spinal slice patch clamp) to exogenously applied NMDA is acutely sensitized by BDNF. (c) The response (voltage change in current clamp) of a spinal motor neuron (sharp electrode in neonatal hemisected spinal cord) to application of NMDA is sensitized by NT-3. (a) Adapted from Galoyan SM, Petruska JC, and Mendell LM (2003) Mechanisms of sensitization of the response of single dorsal root ganglion cells from adult rat to noxious heat. *European Journal of Neuroscience* 18: 535–541. (b) Adapted from Garraway SM, Petruska JC, and Mendell LM (2003) BDNF sensitizes the response of lamina II neurons to high threshold primary afferent inputs. *European Journal of Neuroscience* 18: 2467–2476. (c) Adapted from Arvanov VL, Seebach BS, and Mendell LM (2000) NT-3 evokes an LTP-like facilitation of AMPA/kainate receptor-mediated synaptic transmission in the neonatal rat spinal cord. *Journal of Neurophysiology* 84: 752–758.

the firing rate of dissociated nociceptors in response to a depolarizing ramp stimulus. This change in firing rate is associated with an increase in Na^+ conductance and a decrease in K^+ conductance elicited by membrane depolarization. This action is in parallel with and independent of NGFs sensitization of the TRPV1 receptor via its effect on the trkA receptor.

Inflammation-Induced Release of NGF into Peripheral Tissues

The role of NGF in hyperalgesia has been demonstrated by showing that the pain induced in various experimental models can be reduced or abolished by delivering an antibody to NGF in conjunction with the sensitizing agent. Another approach is to administer the immunoadhesin trkA-IgG, which binds with endogenous NGF. Both of these reduce or abolish inflammatory hyperalgesia. The implication of these experiments is that NGF is upregulated in inflamed skin, and this has been confirmed by direct measurement. Enhanced NGF levels have also been detected in the inflamed bladder and in the synovial fluid of inflamed joints, suggesting that NGF contributes generally to inflammatory hyperalgesia in peripheral tissues. The release of NGF is mediated by prior release of the cytokines tumor necrosis factor-α and interleukin-1β; keratinocytes as well as mast cells are believed to be important sources of NGF (Figure 4).

Long-Term Effects of NGF on Nociceptive Signaling

As indicated previously, regulation of gene activity is an important consequence of NGF binding to the trkA receptor. The expression of a number of genes has been shown to be upregulated by NGF administration in culture and *in vivo*, and these are considered to be important in mediating the long-term effects of NGF (see below). They include genes for ligand-gated ion channels involved in transducing nociceptive stimuli such as TRPV1 (gated by capsaicin/noxious heat), $P2X_3$ (ATP), and ASIC3 (protons); genes for the neuropeptide transmitters SP and CGRP, important in nociception; and Na^+ channel genes, both TTX sensitive (Nav1.3) and TTX insensitive (Nav1.8 and Nav1.9). Expression of the gene for another neurotrophin, BDNF, is also upregulated, and this is considered to be particularly important for the central effects induced by NGF.

These and other similar changes are instrumental in mediating longer-lasting effects of NGF. The synthesis of additional receptors responsible for transducing nociceptive stimuli (TRPV1, $P2X_3$, and ASIC3), enhances the sensitivity of nociceptors to noxious

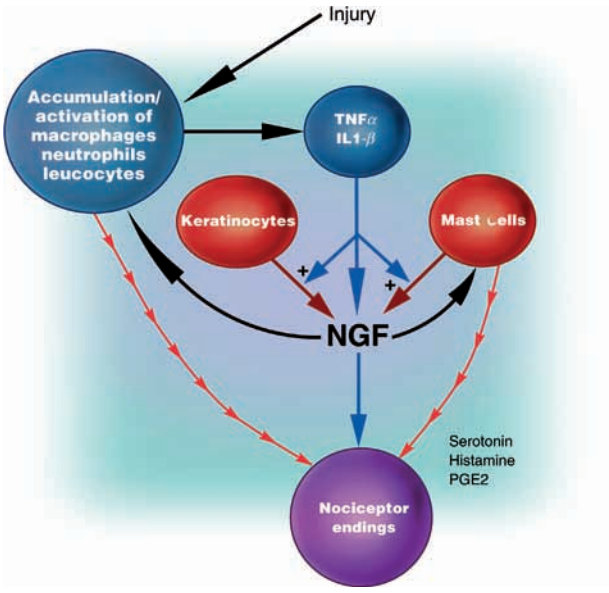

Figure 4 Some cellular mechanisms mediating NGF-induced sensitization of nociceptive afferents. NGF levels are upregulated in peripheral tissues after injury as a consequence of activation of inflammatory cells. NGF directly sensitizes nociceptor endings acutely and also contributes indirectly to sensitizing of these endings via the action of other agents released from immunocompetent cells. Direct administration of NGF activates many of these pathways and has virtually immediate sensitizing effects. After injury, the time course of sensitization will depend on the kinetics of NGF release, but it is generally several hours. NGF will also exert slower effects due to upregulation of other channels, such as TRPV1, bradykinin, and Nav1.8.

heat and other stimuli such as ATP and lowered pH. NGF also enhances expression of BK1 receptors, augmenting the effect of the sensitizing agent bradykinin. The elevated level of BDNF in the somata of trkA-expressing nociceptors after long-term exposure to NGF takes several hours to develop *in vivo*. This phenomenon is significant because BDNF has been shown to be released into the dorsal horn and to sensitize the glutamatergic synaptic response of cells in lamina II. This phenomenon is known as central sensitization (Figure 3(b)), to be distinguished from peripheral sensitization of nociceptors themselves (Figure 3(a)). Another long-term effect of NGF is the induction of terminal sprouting of nociceptors in the skin, which could also result in heightened sensitivity of these fibers.

The Effect of Other Neurotrophins and Glial-Derived Neurotrophic Factors on Peripheral Sensitization

BDNF and NT-4, but not NT-3, have also been shown to sensitize the response to capsaicin in isolated DRG cells and noxious heat in the *in vitro* skin–nerve preparation. Interestingly, NT-3 reduces the response of small DRG cells to NGF despite the fact that there are no trkC receptors on small-diameter DRG cells.

Members of the family of GDNFs, consisting of artemin, neurturin, and GDNF, have also been shown to sensitize the response of TRPV1 receptors in nociceptors. The sensitization is elicited at molar concentrations 10–100 times lower than that required for NGF-elicited sensitization of TRPV1-mediated responses of nociceptors. Although all these substances elicit sensitization, artemin appears to be the most significant since it is the only one whose level in the skin is increased by the inflammatory mediator complete Freund's adjuvant, as is its coreceptor GFR-α3. NGF and members of the GDNF family administered individually sensitize the behavioral response to noxious heat for several hours, but the combination of NGF and artemin sensitizes the response for up to 6 days.

Neurotrophins, Central Sensitization, and Neuropathic Pain

Sensitization of AMPA/kainate synapses on lamina II cells by BDNF as described previously is distinct from the sensitization induced by neurotrophins at peripheral nerve endings. There are many examples of central and peripheral synapses whose efficacy is enhanced by neurotrophins. In the case of nociceptor synapses on cells of lamina II, this sensitization does not occur directly but, rather, as a consequence of sensitization of NMDA receptors located on the same neurons (Figure 3(b)), an effect also found in the NT-3/trkC system of spinal motor neurons (Figure 3(c)). PKC has been implicated in the coupling between sensitization of NMDA and AMPA/kainate receptors. As in the case of NGF-induced sensitization of the response of nociceptive neurons to capsaicin or noxious heat, sensitization of synaptic transmission persists for at least several minutes to an hour after removal of the neurotrophin, suggesting that the neurotrophin triggers long-lasting changes in the nociceptive pathway.

BDNF has also been implicated in other mechanisms of central sensitization, specifically in eliciting neuropathic pain via microglia that are brought to an activated state by nerve injury. Nociceptor terminals co-release ATP and glutamate. ATP acts on activated spinal microglia via $P2X_4$ and P2Y receptors to induce BDNF release. BDNF has been shown to inhibit the KCC2 Cl^- transporter in lamina I cells, shifting the Cl^- reversal potential to become positive with respect to the resting potential. This leads to a GABA-induced depolarization (rather than hyperpolarization) of lamina I cells, thereby enhancing

activity in nociceptive pathways. The microglial and synaptic (trkA-expressing sensory neurons sensitized by NGF) sources of BDNF may act in synergy to produce postsynaptic sensitization throughout the superficial dorsal horn.

NGF is an important intermediate in peripheral sensitization in inflammatory pain models (e.g., after complete Freund's adjuvant or carrageenan). In neuropathic pain models (e.g., chronic nerve constriction injury), the evidence for direct NGF involvement is much less convincing. Nonetheless, neurotrophins may participate since there is evidence of increased BDNF expression in neighboring uninjured sensory neurons. Most of these models are essentially partial nerve injuries in that the system being examined involves plexus nerves. This is important because the spared neurons/axons project through a peripheral nerve containing fibers whose processes have been injured resulting in Wallerian degeneration. This degeneration is similar to peripheral tissue inflammation in many ways, not least of which is that denervated Schwann cells release neurotrophins, including NGF. Furthermore, since peripheral axonal injury induces influx of immune cells to the DRG, noninjured neurons resident in the same DRG as injured neurons are exposed to conditions known to induce sensitization (release of NGF by satellite cells, increased levels of immune mediators such as MCP-1, etc.).

Effects of Neurotrophins on Axonal Growth in Adults

The ability of neurotrophins to promote neurite elongation during development as well as in culture has prompted efforts to determine whether they might enhance neurite growth in the adult animal *in vivo*. BDNF enhances regeneration of motor axons, and both NT-3 and BDNF improve elongation of spinal axons in conjunction with different permissive substrates. NGF has not been found to be as effective as BDNF and NT-3 in promoting regeneration of spinal axons, presumably since they express trkB and trkC but not trkA. On the other hand, regeneration of sensory neurons can be enhanced by NGF as well as by NT-3 and GDNF, but not BDNF. Regenerating dorsal root axons normally do not penetrate the spinal cord, but when exposed to neurotrophins, some growth into the spinal cord has been reported. Interestingly, the identity of the regenerating fibers is appropriate for the neurotrophin that was applied (i.e., nociceptors with NGF or GDNF, and spindle afferent fibers with NT-3).

In addition to elongation, neurotrophins play a role in both adaptive and pathologic axonal branching plasticity in the adult. NGF is known to be a key player in the peripheral collateral sprouting of axons from nociceptive DRG neurons and sympathetic neurons, a process by which noninjured neurons extend new axonal branches to innervate denervated skin. A pathologic version of this process may also occur after spinal cord injury where autonomic dysreflexia is associated with new central connections established by trkA-expressing DRG neurons into the sympathetic outflow circuitry. Remodeling of hippocampal neurons as part of memory processes involves BDNF, and BDNF, NGF, and NT-3 are likely involved to some degree in pathologic (e.g., epileptiform) processes. Evidence for significant NT-3-mediated adaptive plasticity is limited to conditions of nerve or spinal cord injury with exogenous NT-3.

Clinical Implications of Neurotrophin Biology

A number of disorders of the peripheral nervous system have been attributed to alterations in NGF metabolism. The importance of neurotrophins as survival factors predicts that their absence during development should deplete specific populations of neurons according to their neurotrophin dependence. This has been advanced as the explanation of the rarely observed hereditary sensory and autonomic neuropathy type IV, in which the insensitivity to pain and reduced sweating are correlated with an absence of the NGF high-affinity receptor trkA. The infectious disease leprosy is characterized by hypoalgesia and reduced sweating from affected skin areas; this is accompanied by diminished NGF levels in skin biopsies from this areas. Diabetic neuropathy results in reduced nociceptive function; the magnitude of the capsaicin-evoked skin axon reflex (a measure of nociceptive function) is correlated with NGF levels in the skin.

In adults in whom NGF acts to sensitize the response to certain nociceptive inputs, treatments that reduce NGF activity have been explored as possible treatment modalities in various pain states. In animal models, this approach has proven useful in models of experimental arthritis, visceral pain (pancreatitis and interstitial cystitis), or metastatic bone pain where more conventional treatments have not proven effective or have significant side effects. A humanized antibody to NGF (RN624) has been developed and is in clinical trials.

See also: Growth Factors: Neuronal Atrophy; Neurotransmitters and Growth Factors: Overview; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Neurotrophins: Physiology and Pharmacology.

Further Reading

Anand P (2004) Neurotrophic factors and their receptors in human sensory neuropathies. *Progress in Brain Research* 146: 477–492.

Arvanov VL, Seebach BS, and Mendell LM (2000) NT-3 evokes an LTP-like facilitation of AMPA/kainate receptor-mediated synaptic transmission in the neonatal rat spinal cord. *Journal of Neurophysiology* 84: 752–758.

Chao MV and Hempstead BL (1995) p75 and Trk: A two-receptor system. *Trends in Neurosciences* 18: 321–326.

Cowan WM (2001) Viktor Hamburger and Rita Levi-Montalcini: The path to the discovery of nerve growth factor. *Annual Review of Neuroscience* 24: 551–600.

Galoyan SM, Petruska JC, and Mendell LM (2003) Mechanisms of sensitization of the response of single dorsal root ganglion cells from adult rat to noxious heat. *European Journal of Neuroscience* 18: 535–541.

Garraway SM, Petruska JC, and Mendell LM (2003) BDNF sensitizes the response of lamina II neurons to high threshold primary afferent inputs. *European Journal of Neuroscience* 18: 2467–2476.

Hefti FF, Rosenthal A, Walicke PA, et al. (2006) Novel class of pain drugs based on antagonism of NGF. *Trends in Pharmacological Sciences* 27: 85–91.

Huang EJ and Reichardt LF (2001) Neurotrophins: Roles in neuronal development and function. *Annual Review of Neuroscience* 24: 677–736.

Levine JM and Mendell LM (2006) Trophic factors and their influence on regeneration. In: Selzer M, Clarke S, Cohen LG, Duncan PW, and Gage FH (eds.) *Textbook of Neural Repair and Rehabilitation,* vol. 1, pp. 405–420. Cambridge, UK: Cambridge University Press.

Lewin GR and Barde YA (1996) Physiology of the neurotrophins. *Annual Review of Neuroscience* 19: 289–317.

Lewin GR and Mendell LM (1993) Nerve growth factor and nociception. *Trends in Neurosciences* 16: 353–359.

McMahon SB, Mendell LM, Phillips HS, and Wall PD (1996) Neurotrophins and sensory neurons: Role in development, maintenance and injury. *Philosophical Transactions of the Royal Society of London* 351: 361–467.

Mendell LM (1995) Neurotrophic factors and the specification of neural function. *The Neuroscientist* 1: 26–34.

Mendell LM, Albers KM, and Davis BM (1999) Neurotrophins, nociceptors and pain. *Microscopy Research and Technique* 45: 252–261.

Mendell LM and Arvanian VL (2002) Diversity of neurotrophin action in the postnatal spinal cord. *Brain Research Reviews* 40: 230–239.

Petruska JC and Mendell LM (2004) The many functions of nerve growth factor: Multiple actions on nociceptors. *Neuroscience Letters* 361: 168–171.

Pezet S and McMahon SB (2006) Neurotrophins: Mediators and modulators of pain. *Annual Review of Neuroscience* 29: 507–538.

Shu X and Mendell LM (1999) Neurotrophins and hyperalgesia. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7693–7696.

Thoenen H and Sendtner M (2002) Neurotrophins: From enthusiastic expectations through sobering experiences to rational therapeutic approaches. *Nature Neuroscience* 5: 1046–1050.

Zweifel LS, Kuruvilla R, and Ginty DD (2005) Functions and mechanisms of retrograde neurotrophin signalling. *Nature Reviews Neuroscience* 6: 615–625.

Network Control

D R McCrimmon, Northwestern University, Chicago, IL, USA
G S Mitchell, University of Wisconsin, Madison, WI, USA
J L Feldman, University of California at Los Angeles, Los Angeles, CA, USA
G F Alheid, Northwestern University, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

The Neuroanatomy of Breathing

Breathing is generated in neural circuits ranging from the telencephalon to the spinal cord. The rhombencephalon (medulla and pons) contains the neurons that generate respiratory rhythm and then sculpt that neural activity to produce respiratory motor output in cranial and spinal motor neurons. Distinct respiratory output patterns (Figure 1) are expressed in cranial motor neurons in the rhombencephalon to regulate airway caliber and in cervical and thoracic spinal motor neurons that generate the rhythmic pumping action of the lungs. Brain stem and spinal motor neurons contribute to other behaviors as well, including chewing, swallowing, vocalization, vomiting, posture, locomotion, defecation, and reproduction.

Sensory inputs that convey information concerning arterial blood gases and status of the lungs and airways modulate respiratory pattern. Forebrain regions modify breathing by providing for voluntary and emotional changes in breathing patterns and induce feed-forward changes anticipating changes in metabolic demand. Important (and sometimes overlooked) elements of respiratory circuits are the ascending respiratory-related sensory pathways that drive adaptive short- and long-term behavioral responses to assure that the individual is maintained in a favorable environment. Perceived difficulties in breathing, termed dyspnea, are potent stimuli for motivating behavior.

Respiratory Neurons

Patterns of activity in respiratory motor and premotor neurons reflect respiratory modulation, characterized by bursts of action potentials during active periods, alternating with periods of little or no activity. These brain stem and spinal neurons mediate the coordinated motions of the airways and chest wall during breathing and are regulated to mediate respiratory reflexes, the preservation of homeostasis, and voluntary changes in respiratory activity. Respiratory neurons are classified as inspiratory (I) or expiratory (E) by the phase of the respiratory cycle in which they are active. These categories are subdivided into incrementing (augmenting; I-Aug or E-Aug), decrementing (I-Dec or E-Dec), or constant (I-Con or E-Con) types. Neurons with activity spanning the inspiratory and expiratory phases (I–E or E–I) are also observed. Preinspiratory (pre-I) neurons typically fire one brief burst of action potentials immediately before the onset of inspiration and a second burst immediately following the end of inspiration. This cell type has been observed mainly in neonatal rodent *in vitro* preparations, but it is of major significance to current models of respiratory rhythm generation. Commonly recognized respiratory neuron categories are summarized in Figure 2.

Respiratory-Related Neurons

In addition to phasic respiratory neurons, there are significant groups of cells that might be classed as respiratory-related in that they are nonphasic cells that, nonetheless, modulate respiration via tonic afferents to brain stem respiratory circuits. This, for example, is probably the case for second-order neurons in the nucleus of the solitary tract (NTS) relaying information from peripheral chemoreceptors in the carotid body.

Brain Stem Respiratory Neuron Groups

Respiratory neurons are concentrated (but not found exclusively) within three regions of the rhombencephalon (Figure 3). The dorsal respiratory group (DRG; Figures 3(a) and 3(b)) is in the dorsomedial medulla in caudal to mid-rostral portions of the NTS. The ventral respiratory column (VRC; Figures 3(a) and 4) forms a continuous field in the ventrolateral medulla extending from around the facial nucleus to the medullary spinal junction (Figure 4). This large aggregate of respiratory neurons encompasses five or more physiologically distinct subregions (Figure 3(b)). The pontine respiratory group (PRG; Figure 3(b)) is in the rostral dorsolateral pons and encompasses portions of the parabrachial complex and the ventrally adjacent Kölliker–Fuse nucleus.

Nucleus of the Solitary Tract and Dorsal Respiratory Group

The NTS is a complex structure that relays visceral afferents ascending in the vagus, glossopharyngeal, and facial nerves. Afferents from the airways, lungs, and peripheral chemoreceptors directly target various NTS subnuclei, largely localized at mid rostrocaudal (corresponding to the length of area postrema) to caudal (posterior to area postrema) NTS levels (including

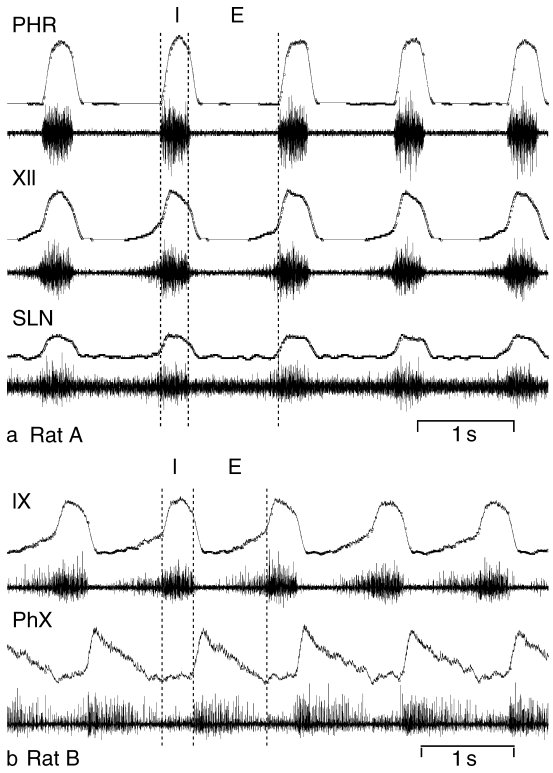

Figure 1 Patterns of respiratory motor activity on selected spinal (phrenic) and cranial nerves in two rats (a and b). Vertical dotted lines indicate phase transitions between inspiration (I) and expiration (E). Note that the onset of inspiratory activity on the cranial nerves precedes that on the phrenic nerve. The top trace for each nerve is a low-pass filtered version of the raw activity shown in the trace immediately below, showing the nerve's overall discharge pattern. IX, glossopharyngeal nerve; Phr, phrenic nerve; PhX, pharyngeal branch of the vagus nerve; SLN, superior laryngeal branch of the vagus nerve; XII, hypoglossal nerve. Reproduced from Hayashi F and McCrimmon DR (1996) Respiratory responses to cranial nerve afferent stimulation in rats. *American Journal of Physiology* 271: R1054–R1062.

its commissural subdivision). The DRG designates NTS neurons with centrally generated respiratory activity (most commonly inspiratory), thus excluding neurons whose respiratory activity is largely dependent on peripheral afferent inputs. The DRG is a source of input to spinal cord respiratory motor neurons in some species, but is less so in others.

Ventral Respiratory Column

Distinct anatomical compartments within the VRC are inferred from clustering of similar types of respiratory neurons (e.g., inspiratory or expiratory discharge patterns), presence or absence of spinal cord projections, and respiratory responses to chemical stimulation. Specific respiratory neuron types are not restricted to individual compartments but exhibit overlapping distributions with the peak concentrations of particular neuron types centered within given compartments.

The caudal half of the VRC, termed the ventral respiratory group (VRG), contains bulbospinal respiratory premotor neurons (Figure 5) that receive converging inputs from VRC rhythm-generating neurons and other inputs from outside the VRC, sculpting the pattern of motor output distributed to various respiratory motor neurons. The VRG is subdivided into rostral (rVRG) and caudal (cVRG) based on the peak concentrations of inspiratory (rVRG) versus expiratory (cVRG) bulbospinal neurons. Inspiratory bulbospinal neurons provide mainly excitatory input to spinal inspiratory motor pools, including monosynaptic input to phrenic motor neurons. Expiratory bulbospinal neurons in the cVRG are mainly excitatory and project to interneurons innervating thoracic and upper lumbar motor neurons that drive expiratory pump muscles (e.g., internal intercostal muscles and abdominal muscles). Inhibitory expiratory neurons in the vicinity of the Bötzinger complex (BötC) and cVRG contribute to the inhibition of spinal inspiratory motor pools.

Respiratory rhythm generation occurs mainly in the rostral half of the VRC, which contains at least three functionally distinct compartments (Figure 3(b)). The BötC occupies a portion of the VRC adjacent to the caudal surface of the facial nucleus and contains mainly inhibitory glycinergic expiratory neurons. The pre-Bötzinger complex (pre-BötC), an area between the BötC and anterior rVRG, is a key region for rhythm generation and is hypothesized to be essential for inspiratory rhythm. At least in the rat, the majority of respiratory neurons identified within the pre-BötC are excitatory medullary interneurons that project extensively within the ipsilateral and contralateral VRC. Some pre-BötC propriobulbar neurons express neurokinin-1 (NK1) receptors, a feature distinguishing between the pre-BötC and the BötC. NK1 receptor-positive neurons also partially differentiate the pre-BötC from the rVRG, although some large NK1 receptor-positive bulbospinal neurons are found in anterior rVRG.

The retrotrapezoid nucleus (RTN) is located at the rostral end of the VRC, ventral to the rostral BötC and continues rostrally beneath the facial nucleus near the surface of the brain to just caudal to the trapezoid body (Figures 3(b), 3(c), and 4). RTN neurons project to the pons and caudal VRC. The RTN is (part of) a region containing potential central chemosensory cells, an important source of respiratory drive and chemoreflexes. RTN neuronal activity monotonically increases with extracellular acidity (associated with increased CO_2).

The parafacial respiratory group is located ventral to the facial nucleus, at least in neonatal rats (pFRG; Figures 3(b) and 3(c)). Many pFRG neurons have a

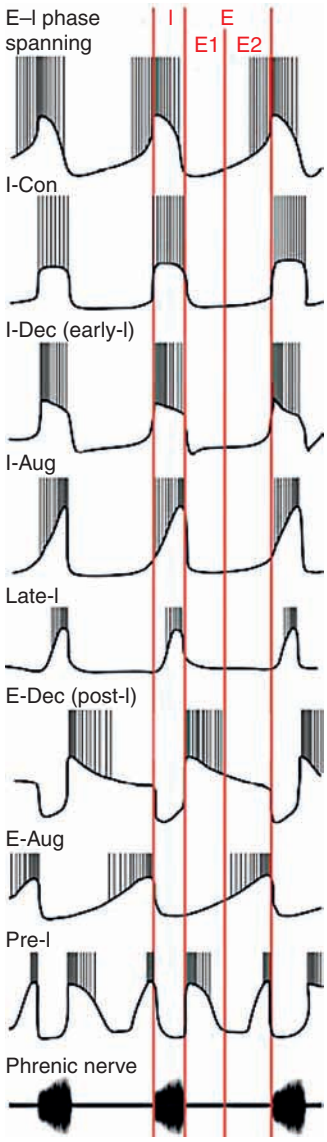

Figure 2 Membrane potential and representative discharge patterns of major classes of respiratory neurons found in the medulla and pons. Alternative names remain in common usage for these neurons; some of these are shown in parentheses. Note that a rapid decline in phrenic nerve activity is associated with the inspiratory-to-expiratory transition. Inspiratory and expiratory periods are indicated by the vertical red lines. Expiration is often divided into two phases: E1 and E2. E1 is defined by activity on expiratory-decrementing (E-Dec) (also called postinspiratory (post-I)) neurons and can include residual but declining phrenic nerve activity. Expiratory-augmenting (E-Aug) neuronal activity occurs mainly in E2. Expiratory–inspiratory phase spanning (E–I) neurons begin to depolarize and discharge prior to the onset of inspriation, with their peak depolarization near the transition from expiration to inspiration. Such neurons are concentrated in the pre-BötC, and at least a subset appears to be critical for inspiratory rhythm generation. Inspiratory-augmenting (I-Aug, or I-ramp) neurons start to depolarize at a time approximating the onset of phrenic nerve discharge and then progressively depolarize throughout inspiration, resulting in a graded increase in discharge rate. Late inspiratory (late-I) neurons begin to depolarize and discharge only late in inspiration. Inspiratory-constant (I-Con) neurons discharge at a relatively constant level throughout inspiration. A final group of inspiratory neurons, inspiratory-decrementing (I-Dec) cells, depolarize abruptly at the onset of inspiration and fire at a high rate. As inspiration progresses, these neurons slowly repolarize and decrease their discharge rate. Expiratory neurons can be divided into three predominant groups. Analogous to I-Dec neurons, E-Dec neurons begin to discharge at a high frequency at the onset of expiration. They then progressively repolarize and decrease their discharge rate during expiration. Some investigators distinguish between E-Dec neurons, which discharge at a declining rate throughout expiration, and post-I neurons, which tend to discharge only during the early portion of expiration (the E1, or post-I, interval). E-Aug neurons typically begin to discharge about midway through expiration (i.e., during the late or E2 phase) and, like I-Aug neurons, exhibit a progressive depolarization and a consequent increase in firing rate. A final group of expiratory neurons, termed pre-I neurons, have been identified in neonatal rats and appear to be important in generating expiratory rhythm. These neurons typically depolarize and fire a brief burst of action potentials immediately prior to inspiration onset; they are then hyperpolarized during inspiration and cease firing. Immediately following inspiration termination, they again briefly depolarize and fire a brief burst of action potentials. pre-BötC, pre-Bötzinger complex. Adapted from Feldman JL and McCrimmon DR (1999) In: Zigmond MJ, Bloom FE, Landis SC, Roberts JL, and Squire LR (eds.) *Fundamental Neuroscience*, pp. 1063-1090. London: Academic Press.

pre-I activity pattern. The pFRG is located beneath lateral portions of the facial nucleus in the neonate and may extend caudally beneath the BötC complex. Studies in neonatal and preadolescent juvenile rats have implicated the pFRG in the generation of respiratory rhythm and appear to be critical for active expiratory activity. Whether RTN neurons in the adult rat brain are the same population identified as pre-I neurons in the pFRG of the neonate is not yet known.

Pontine Respiratory Group

The PRG is a group of respiratory neurons in the dorsolateral pons that includes neurons in ventral parts of the medial and lateral parabrachical complex and Kölliker–Fuse nucleus (**Figures 3(b), 3(c) and 4**). The PRG and the VRC are at either end of an essentially continuous column of respiratory and respiratory-related neurons in the lateral pons and medulla (**Figures 3(c), and 4**). It extends rostrally from the medulla, essentially surrounds the facial nucleus, traverses the ventrolateral pons and regions adjacent to both the motor trigeminal and principal sensory trigeminal nuclei before merging ventrally with the PRG. These neurons merge medially with VRC-projecting neurons in the subcoeruleus area.

The lateral pontine column of neurons intervening between the PRG and VRC is inhomogeneous, but it contains neurons with axons that project to the VRC, PRG, and/or DRG, as well as cranial motor neurons (e.g., those innervating airway and orofacial regions,

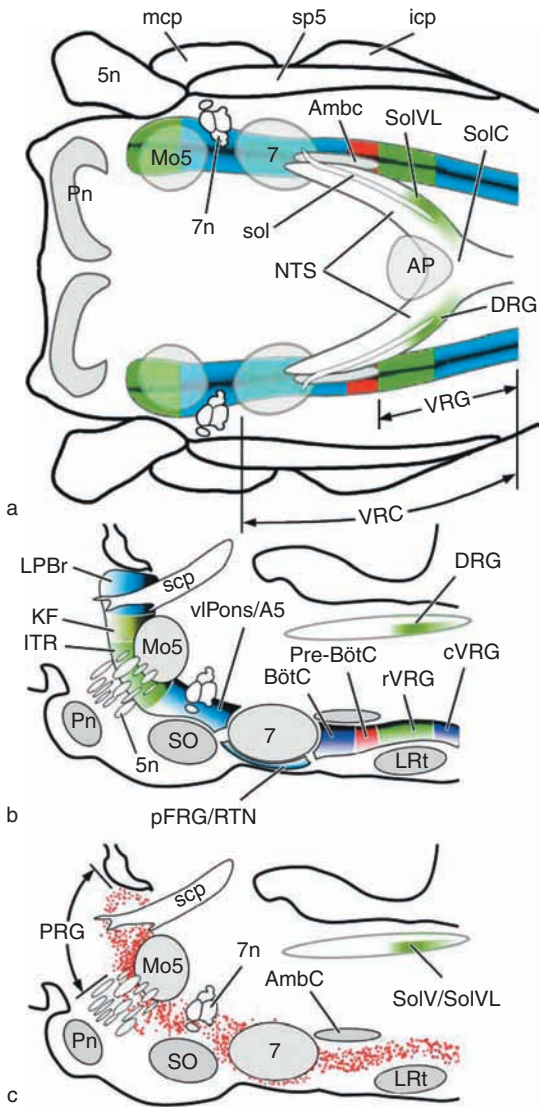

Figure 3 Respiratory neurons in the rhombencephalon of the rat brain: (a) horizontal view; (b, c) sagittal views. In (a) and (b), functionally distinct brain stem compartments related to breathing are indicated within the shaded columns by different colors. Nearby brain stem structural landmarks are indicated in gray for prominent neuronal nuclei or outlined for major fiber bundles. In (c), neurons are depicted in lateral portions of the pons and medulla that are retrogradely labeled (red dots) following an injection of FluoroGold into the general vicinity of the pre-BötC. Note that the nearly continuous column of propriobulbar neurons occupies regions that have been shown in independent physiological experiments to consist of respiratory neurons or respiratory-related neurons influencing the control of breathing. The column of neurons is interrupted or displaced by the facial nucleus that migrates caudally from the pons in ontogeny to reach its final ventral location at the rostral end of the medulla. Respiratory interneurons are found at the borders of the facial nucleus, most prominently ventrally in the retrotrapezoid nucleus (RTN) in (b). Portions of the RTN, especially caudally, have been associated with central chemoreceptors (sensing CO_2 and pH), providing drive (excitation) to other respiratory brain stem neurons; other neurons in the RTN have been associated with (expiratory) rhythm generation and encompassed by the term parafacial respiratory group (pFRG). The degree to which chemoreception and expiratory rhythm generation are served by independent groups of neurons remains to be determined. It is postulated that pFRG neurons are coupled (reciprocally) to the pre-BötC, a region in the rostral medulla that appears to be essential for the generation of the inspiratory rhythm. Respiratory neurons are most commonly found in the ventral respiratory column of the medulla (VRC) in (a), in the dorsal respiratory group (DRG, in (b)); located within ventrolateral portions of the nucleus of the solitary tract (SolVL), and in the pontine respiratory group (PRG, in (c)) in the dorsolateral pons. Note that, although only the Kölliker–Fuse nucleus and the lateral part of the parabrachial complex are depicted, the term PRG also encompasses respiratory neurons in medial portions of the parabrachial complex. An area termed the ventral respiratory group (VRG; in (a)) is mainly associated with medullary bulbospinal premotor neurons and is subdivided into a rostral aggregate of mainly inspiratory bulbospinal neurons (rVRG) and a more caudal aggregate of mainly expiratory bulbospinal neurons (cVRG). 5n, trigeminal nerve; 7, facial nucleus; 7n, facial nerve; A5, A5 noradrenergic neuronal group; AmbC, compact part of nucleus ambiguus; AP, area postrema; BötC, Bötzinger complex; cVRG, caudal division of ventral respiratory group; DRG, dorsal respiratory group; icp, inferior cerebellar peduncle; ITR, intertrigeminal KF, Kölliker–Fuse nucleus; LPBr, lateral parabrachial region; LRt, lateral reticular nucleus; mcp, medial cerebellar peduncle; Mo5, motor nucleus of the trigeminal nerve; NTS, nucleus of the solitary tract; pFRG, parafacial respiratory group; Pn, basilar pontine nuclei; pre-BötC, pre-Bötzinger complex; RTN, retrotrapezoid nucleus; rVRG, rostral division of ventral respiratory group; scp, superior cerebellar peduncle; SO, superior olive; sol, solitary tract; SolC, commissural subdivision of the nucleus of the solitary tract; SolV, ventral subdivision of the nucleus of the solitary tract; SolVL, ventrolateral subdivision of the nucleus of the solitary tract; sp5, spinal trigeminal tract; vlPons, ventrolateral pontine region; VRC, ventral respiratory column of the medulla; VRG, ventral respiratory group. (b, c) From Alheid GF, Milsom WK, and McCrimmon DR (2004) Pontine influences on breathing: An overview. *Respiratory Physiology & Neurobiology* 143: 105–114.

including the tongue). Although their specific roles in breathing remain incompletely characterized, neurons within this region can be subdivided into distinct compartments (Figures 3(b) and 4). For example, the area between the motor and sensory trigeminal nuclei (the intertrigeminal area), along with portions of the overlying Kölliker–Fuse nucleus, has been implicated in the generation of apneas as part of airway and orofacial protective reflexes.

Mesencephalic Influences on Breathing

Periaqueductal gray Periaqueductal gray (PAG) neurons that project to the parabrachial complex, ventrolateral pons, and VRC modulate breathing. PAG neurons are arranged in rostrocaudally organized columns with individual columns related to emotional coping, pain responses (including pain suppression), autonomic activation and suppression, vocalizations, and reproduction. Neurons in the caudal-lateral and ventrolateral PAG are related to vocalizations

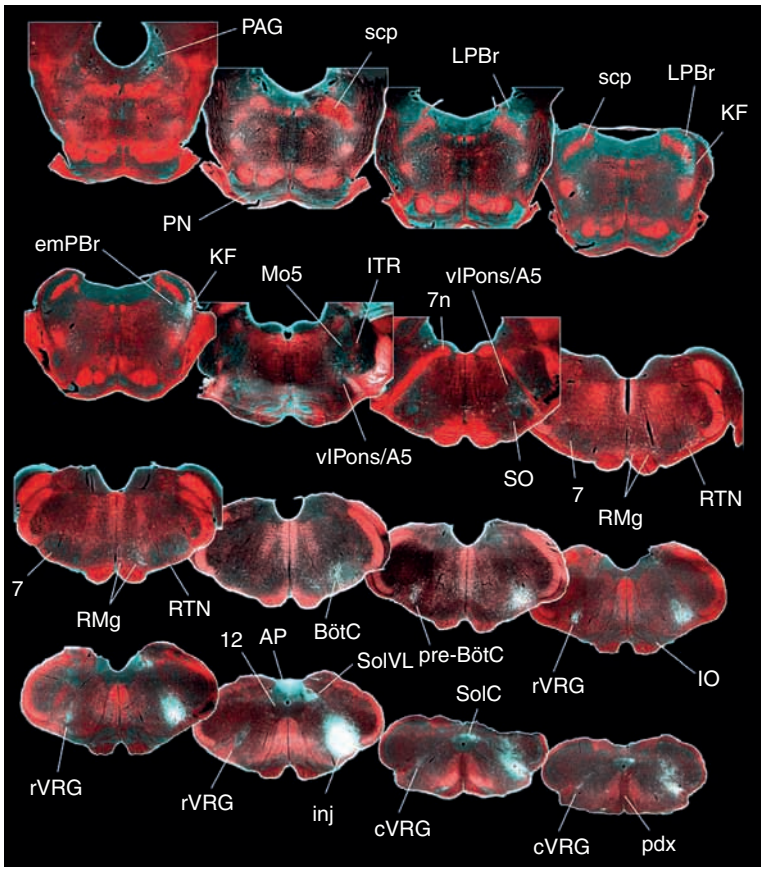

Figure 4 Coronal sections through the rhombencephalon depicting the location of neurons projecting to the ventral respiratory column of the rat. The neurons were retrogradely labeled by a tracer injection (inj) in the rVRG. The retrogradely labeled neurons are shown in white projected against pseudo-colored dark-field images (red) of the brain stem sections. Note that compact aggregates of neurons in the medulla contralateral to the injection site identify the pre-BötC, rVRG, and cVRG. In contrast, areas such as respiratory-related regions of the NTS, the Kölliker–Fuse nucleus, RTN, and BötC, project mainly to the ipsilateral medulla and consequently are only prominently labeled on the same side as the injection. It should be noted that, although the area postrema (AP) appears brightly labeled (last row), this is an artifact resulting from the relatively high amounts of endogenous peroxidase in the area postrema that is demonstrated by the immunoperoxidase procedure used to detect the retrogradely labeled neurons. It is also notable that the AP is a midline structure in small mammals but a bilateral structure in larger mammals such as cats and dogs. Despite this structural difference, respiratory-related NTS neurons are found in similar locations alongside the AP and caudal to it. 7, facial nucleus; 7n, facial nerve; 12, hypoglossal nucleus; A5, A5 noradrenergic neuronal group; AP, area postrema; BötC, Bötzinger complex; cVRG, caudal division of ventral respiratory group; emPBr, external medial parabrachial nucleus; IO, inferior olive; KF, Kölliker–Fuse nucleus; LPBr, lateral parabrachial region; Mo5, motor nucleus of the trigeminal nerve; NTS, nucleus of the solitary tract; PAG, periaqueductal gray; pdx, pyramidal decussation; Pn, basilar pontine nuclei; pre-BötC, pre-Bötzinger complex; RMg, raphe magnus; RTN, retrotrapezoid nucleus; rVRG, rostral division of ventral respiratory group; scp, superior cerebellar peduncle; SO, superior olive; sol, solitary tract; SolC, commissural subdivision of the nucleus of the solitary tract; SolVL, ventrolateral subdivision of the nucleus of the solitary tract; sp5, spinal trigeminal tract; vlPons, ventrolateral pontine region; VRG, ventral respiratory group.

(particularly in response to emotional or sensory inputs) and project to cVRG neurons premotor to laryngeal and abdominal motor neurons. PAG vocalization areas also synapse on vocalization-related neurons in the ventrolateral pons. This latter area integrates afferents from the PAG, cochlea (auditory), and forebrain that modify vocalization/breathing by projections to the VRC (e.g., cVRG) as well as to laryngeal motor neurons. Projections from cortical laryngeal motor areas terminate in the ventrolateral pons, but do not target laryngeal motor neurons directly; an exception is the human brain, in which stimulation of motor cortex directly activates laryngeal motor neurons. Dorsolateral PAG projections to the parabrachial complex appear necessary for orienting responses, such as changes in heart rate, blood pressure, and breathing elicited by novel stimuli.

Forebrain influences on breathing In addition to its role in vocalization, the forebrain affects breathing related to emotions, metabolic and hormonal homeostasis, and sleep–wake cycles. Few of these influences are mediated by direct projections to the VRC. Direct projections arise from posterior hypothalamus, paraventricular hypothalamic nucleus, and perifornical and lateral hypothalamus. These projections are

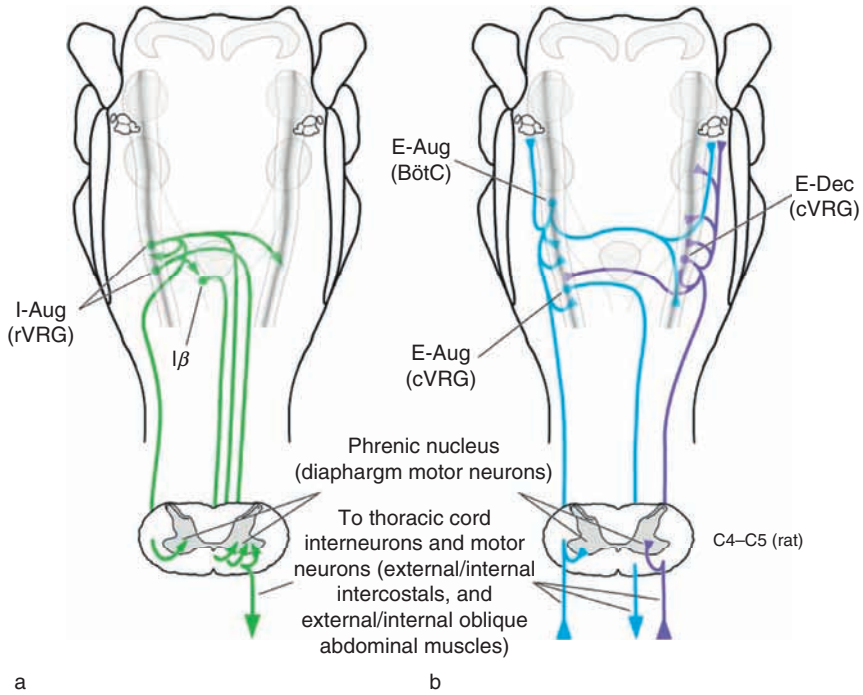

Figure 5 Bulbospinal neurons in the ventral respiratory column of the rat: (a) inspiratory bulbospinal neurons and their axons; (b) expiratory bulbospinal neurons and their axons. These are depicted on horizontal sections of the rhombencephalon and in transverse sections of the cervical spinal cord (C4–C5) from the level of the phrenic nucleus. As shown in (a), in the rat I-Aug neurons in the rVRG are the main source of excitatory (glutamatergic) inspiratory afferents innervating the phrenic nucleus and driving the diaphragm. Most I-Aug neurons use enkephalin as a co-transmitter. I-Aug neurons project via axons crossing in the medulla that descend both ipsilaterally and contralaterally to the parent cell to terminate in the spinal cord. I-Aug neurons send collaterals to other premotor neurons in the ipsilateral and contralateral ventral respiratory column and to a subset of respiratory-related cranial motor neurons. At least some of the I-Aug axons provide inspiratory input to thoracic inspiratory motor neurons (e.g., external intercostal motor neurons). In the cat, (excitatory) Iβ neurons in the ventrolateral subdivision of the NTS receive an inspiratory respiratory drive as well as excitatory SAR afferent input and project to the phrenic nucleus in the contralateral spinal cord; these cells do not provide local collaterals to respiratory areas of the medulla. Iγ neurons have also been identified in the rat. They receive a central respiratory drive, DAR afferent input, and project to the spinal cord. These neurons are included within the population of RAR/DAR cells depicted in **Figure 6**. As shown in (b), E-Aug neurons in the cVRG provide the excitatory drive to expiratory motor neurons in the thorax and abdomen. These neurons send axons that cross in the medulla and descend in the contralateral spinal cord to innervate thoracic and upper lumbar spinal interneurons and motor neurons, ultimately controlling expiratory pump muscles, such as the internal intercostal muscles of the rib cage and the abdominal muscles. No local medullary collateral axons arise from the cVRG E-Aug neurons, but their axons bifurcate at the level of their target neurons to innervate expiratory motor pools both ipsilateral and contralateral to the parent cell. E-Aug neurons in the BötC send inhibitory (glycinergic) axon collaterals to the ipsilateral VRC and a substantially lower number of branches to the contralateral VRC (see also **Figure 4**). Ascending axonal branches are also provided to the pons, including the PRG. E-Dec neurons are found throughout the VRC, particularly within the BötC and cVRG. Many E-Dec neurons within the VRC are inhibitory (glycinergic) and their axons arborize in the ipsilateral, and to a lesser extent, the contralateral medulla. Some E-Dec neurons, particularly those located caudally in the VRC provide axon collaterals to the ipsilateral spinal cord. Both BötC E-Aug neurons and E-Dec neurons are potential sources of afferents to phrenic motor neurons, which are actively inhibited during expiration. BötC, Bötzinger complex; cVRG, caudal division of ventral respiratory group; DAR, deflation-activated receptor; E-Aug, expiratory-augmenting; E-Dec, expiratory-decrementing; I-Aug, inspiratory-augmenting; I-Dec, inspiratory-decrementing; NTS, nucleus of the solitary tract; RAR, rapidly adapting stretch receptor; rVRG, rostral division of ventral respiratory group; scp, superior cerebellar peduncle; SAR, slowly adapting stretch receptor; VRC, ventral respiratory column of the medulla.

supplemented by telencephalic contributions originating in insular and frontal cortex and in portions of the extended amygdala, such as the central nucleus of the amygdala and the bed nucleus of the stria terminalis. Despite modest direct projections to the VRC, the forebrain substantially engages pontine relay neurons in areas such as the PAG, parabrachial complex, and ventrolateral pons that integrate voluntary, emotional, hormonal, and reflexive motor demands on the respiratory system.

Monoamine systems and breathing In addition to the effects of monoamines on breathing that correspond to its roles in arousal, control of motor tone and behavioral state (waking/sleeping), specific respiratory functions may also be served by select portions of the monoamine systems.

Serotonin. While many serotonin neurons in the caudal brain stem target both the dorsal and ventral spinal cord, serotonergic innervation of the VRC appears to arise from raphe magnus in the rostral

medulla-caudal pons. In victims of sudden infant death syndrome, postmortem serotonin receptor density is decreased, and the area with the most obvious receptor changes is the arcuate nucleus, a medullary nucleus composed of serotonergic and nonserotonergic neurons (a likely homolog of the raphe magnus and nearby parapyramidal neurons described in nonprimate mammals; both regions have mixed serotonergic and nonserotonergic neurons). The arcuate region in the human probably corresponds to one of several CO_2-chemosensitive regions located in the medulla and caudal pons. Medullary serotonin neurons target broad regions of the spinal cord, including respiratory motor neurons. Serotonergic innervation of motor pools appears to play an important role in regulating their excitability and in initiating/orchestrating plasticity in respiratory motor neurons. As with other monoamines, there is a circadian cycle in serotonergic neuronal activity, and this could directly impact the regulation of respiration during sleep–wake cycles.

Catecholamines. Norepinephrine, epinephrine, and dopamine have diverse effects on breathing, including effects on breathing frequency that depend on the adrenergic receptor subtype. The organization of catecholamine input to respiratory-related areas is somewhat confusing because brain stem noradrenergic cell groups vary as a function of developmental age and between species (or even between rat strains).

Noradrenergic neurons (A5 and A6 groups) modulate respiratory motor output at multiple sites, including both brain stem and spinal cord. Activity in catecholaminergic neurons also varies with the circadian cycle, and they probably mediate some changes in respiratory activity characteristic of the sleep–wake cycle. In addition, locus coeruleus neurons may participate in central chemoreception. Noradrenergic neurons, including neurons in the subcoeruleus region, probably project to the VRC. An inhibition of respiration mediated by α_2-adrenergic receptors is observed after central injections of catecholamines, and in *in vitro* preparations pontine noradrenergic afferents suppress breathing. Noradrenergic and adrenergic neurons located in the medial and commissural nuclei of the NTS are also a likely source of afferents to the ventrolateral medulla, but whether these neurons reach respiratory circuits is unclear.

Dopamine neurons in the posterior hypothalamic A11 cell group are a likely source of dopaminergic projections to the VRC and DRG. However, the central effects of dopamine on breathing are not well understood. When the pons is included, dopamine facilitates fictive breathing in brain stem–spinal cord preparations *in vitro*, but depresses respiration when the pons is removed; the depression appears to result in part from the effects of dopamine D4 receptors on a subset of pre-I neurons in the pFRG that depolarize and begin to fire tonically when stimulated with dopamine. Important dopaminergic influences are also found in the NTS, where dopamine (as well as histamine) is released from chemosensory afferent fibers arising in the carotid body.

Respiratory Reflexes

Sensory receptors related to breathing have essential regulatory and airway-protective roles. Coughing, sneezing, mucus secretion, and airway constriction protect the airways from irritants and facilitate the removal of inhaled substances that are potentially harmful to the lungs and airways. Sensory feedback from arterial and central chemoreceptors as well as lung mechanoreceptors modulates the breathing pattern (e.g., tidal volume and breathing frequency) and is essential for matching ventilation to changing metabolic requirements. Feedback is also essential for adapting respiratory motor output to other behaviors such as changes in posture or locomotion.

Receptors in the Nasal Passages and Pharynx

Respiratory reflexes can be evoked from the nasal cavities via receptors in the mucous membrane. Sensory information is transmitted largely via branches of the trigeminal nerve, including the anterior ethmoidal nerve and branches of the maxillary nerve. Nasal receptors respond to inhaled irritants such as ammonia, cigarette smoke, and volatile anesthetics, as well as to cold and decreased pressure associated with respiratory airflow. Irritants elicit various cardiorespiratory reactions, including sneezing, mucus secretion, bradycardia, hypertension, and inhibition of breathing.

Submersion of the face in water elicits a stereotypic response consisting of apnea, bradycardia, and increased peripheral vascular tone (the diving reflex). The diving reflex is elicited by receptors with afferent fibers in the trigeminal nerve that terminate primarily in ventral aspects of the spinal trigeminal nucleus with smaller terminal fields in the NTS, paratrigeminal nucleus, ventrolateral medulla, and PRG. The diving reflex can be anticipatory. This aspect involves higher central nervous system (CNS) processing, probably interacting with the dorsolateral pons. The parabrachial complex relays somatic and visceral sensations to both forebrain and brain stem; conversely, the parabrachial complex and Kölliker–Fuse nucleus relay descending cortical pathways to cardiorespiratory regions in ventolateral pons and medulla.

Laryngeal receptors elicit powerful respiratory reflexes via vagal afferent fibers that course in the

recurrent and superior laryngeal nerves. Receptors are divided into several different subgroups based principally on their activating stimulus. Cold/flow receptors respond to laryngeal cooling, which occurs during inspiration as room temperature air moves through the air passages. Pressure receptors sense laryngeal pressure changes, especially the decreased pressure that exerts a collapsing force on the airway during inspiration. Respiratory drive receptors respond to the magnitude of contraction of airway muscles, especially the sole airway abductor (opener), the posterior cricoarytenoid. Irritant receptors are activated by inhaled irritant chemicals (e.g., ammonia or cigarette smoke) as well as the presence of water (solutions lacking chloride) in the larynx. Finally, C-fiber receptors respond to chemical stimuli; significant percentages (~40% in mouse) express transient receptor potential vanilloid (TRPV1) receptors that are activated by protons, heat, and vanilloid ligands such as capsaicin. The relative inaccessibility of laryngeal receptors to experimental manipulation has limited our understanding of their specific roles in the production of respiratory reflexes. However, given their location at the entry to the trachea, they have the capacity to produce a variety of respiratory protective reflexes, including cough and apnea, as well as reflexes that promote airway patency by activation of upper airway muscles (e.g., genioglossus and posterior cricoarytenoid). Beyond the termination of the primary afferent fibers in the NTS, little is known about the central pathways by which these afferent fibers exert their influence on breathing.

Receptors in the Lower Airways

Generally, three categories of receptors are identified within the airways below the larynx: slowly adapting stretch receptors (SARs), rapidly adapting stretch receptors (RARs), and bronchopulmonary C-fibers. Receptors termed deflation-activated receptors (DARs) are prevalent in small mammals and share some homologies with the RARs in larger mammals – both elicit augmented inspiratory efforts and both are inhibited by γ-aminobutyric acid (GABA)ergic neurons receiving input from SARs. SARs, RARs, and DARs constitute airway mechanoreceptors innervated by rapidly conducting, myelinated, vagal afferent fibers that terminate with distinct topographies in the NTS (Figure 6).

C-fibers are slowly conducting unmyelinated axons constituting the majority (~75%) of pulmonary vagal afferent fibers. They terminate in several caudal NTS subdivisions (Figure 6). They are polymodal and nociceptive, responding to a variety of inflammatory mediators and inhaled irritants. Reflex changes in breathing in response to C-fiber activation involve rapid shallow breathing interspersed with apneas; additional reflex effects include bronchoconstriction, mucus secretion, hypotension, bradycardia, and airway mucosal vasodilation. Beyond the identification

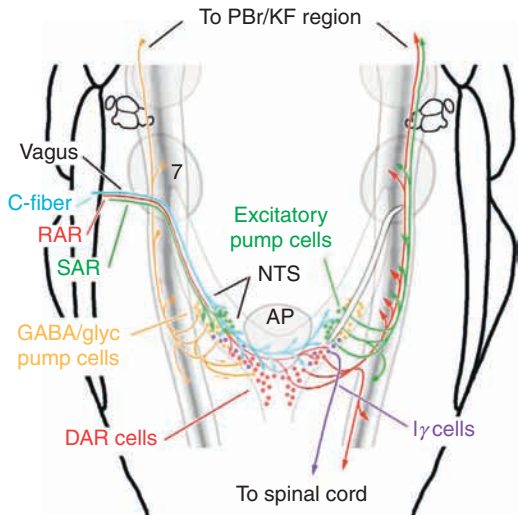

Figure 6 Respiratory-related afferents to the NTS. Vagal afferents to the respiratory-related areas of the rat NTS are shown. SARs, RARs, and C-fibers travel in the vagus nerve to enter the ventrolateral quadrant of the medulla at mid-rostral levels. These axons then travel caudally in the solitary tract to terminate in the NTS. SARs (green axons) terminate ipsilaterally in the ventral, ventrolateral, intermediate, and interstitial NTS subnuclei at the rostrocaudal level of the area postrema (AP). RARs (red vagal axons) terminate in a region beginning near the caudal end of area postrema (also called the calamus scriptorius) and extending across the midline at the commissural subdivision of the NTS. The terminations of vagal C-fibers (blue) are more widely dispersed and terminate bilaterally in the NTS. Inhibitory (GABAergic) SAR second-order neurons (pump cells, orange) in SolV and SolVL project extensively to the ipsilateral VRC and also send ascending axon collaterals to the lateral pons. Excitatory pump cells (green) are also likely to exist and their postulated axonal arborization is depicted as similar to that of GABAergic pump cells. E-Dec neurons are probable targets of excitatory pump cell axons. RAR neurons are less commonly observed in the rat than in the cat. DAR neurons (red), however, are more frequent and are topographically located in NTS regions similar to those inhabited by RAR neurons in the cat. DAR neurons are excitatory and give off extensive collaterals to the ipsilateral VRC and to respiratory related targets in the ipsilateral pons. Some also project to the spinal cord. Iγ and DAR neurons both receive input from deflation-sensitive receptors. Iγ neurons (purple), however, also receive a central respiratory drive and usually project to the spinal cord. AP, area postrema; BötC, Bötzinger complex; DAR, deflation-activated receptor; E-Dec, expiratory-decrementing; GABA, γ-aminobutyric acid; KF, Kölliker–Fuse nucleus; NTS, nucleus of the solitary tract; PBr, parabrachial nucleus; RAR, rapidly adapting stretch receptor; SAR, slowly adapting stretch receptor; SolV, ventral subdivision of the nucleus of the solitary tract; SolVL, ventrolateral subdivision of the nucleus of the solitary tract; VRC, ventral respiratory column of the medulla.

of the regions of C-fiber termination within the NTS, little is known concerning central pathways mediating these responses.

SARs are located in airway smooth muscle, particularly in large airways, and modulate breathing pattern, airway smooth muscle tone, systemic vascular resistance, and heart rate. SAR-reflex effects on breathing pattern are known as the Breuer–Hering reflexes and include alterations of both tidal volume and breathing frequency. Expansion of the lung during inspiration increases SAR activity and leads to inspiratory termination, thereby limiting the tidal volume (i.e., the Breuer–Hering inspiration-inhibiting reflex). If inflation is maintained or lung volume is increased, the expiratory duration is lengthened (i.e., the Breuer–Hering expiratory facilitatory reflex). Breuer–Hering reflexes are active under relaxed breathing conditions in all mammals, although their contribution in adult humans during resting conditions is relatively minor. SAR influences in humans increase during conditions such as physical exercise, when tidal volume is two to three times its resting value.

RARs are arranged circumferentially in airway epithelium and submucosa. They respond to lung inflation and irritant chemicals, but tend to have little activity under normal conditions. RARs fire irregularly with only a few action potentials during lung inflation and give rise to A∂-fibers in the vagus nerve. They are implicated in a number of potent airway protective reflexes, including augmented breaths (sighs), airway constriction, mucus secretion, and laryngeal closure. Coughs appear to be triggered by a select subset of polymodal A∂-fibers (cough receptors).

Second- and Higher-Order Neurons in Respiratory Reflexes

Central Pathways of Slowly Adapting Stretch Receptors

SAR primary afferent fibers have a restricted arborization within the mid to caudal portions of the NTS (Figure 6). Despite diverse autonomic and respiratory motor reflexes elicited by SAR activation, only two functional classes of NTS neurons receive monosynaptic input from SARs (Figure 6). One type, termed Iβ neurons, exhibit inspiratory discharge patterns independent of, but augmented by, SAR inputs. Iβ neurons are bulbospinal premotor neurons that monosynaptically excite phrenic motor neurons, although the extent of this projection may vary among species. This disynaptic activation of phrenic motor neurons represents a positive feedback excitation of phrenic nerve discharge by SAR afferents during lung inflation.

A second group of NTS neurons, termed pump neurons, presumably mediate Breuer–Hering reflex changes in respiratory rhythm. Pump neurons receive monosynaptic SAR input but can be readily distinguished from Iβ neurons by their general lack of inspiratory discharge in the absence of SAR input. Consistent with the broad effects they elicit on respiratory pattern, they have extensive axonal arborizations within the DRG, VRC, and lateral pontine respiratory areas.

Many pump neurons are GABAergic, with a small percentage also containing glycine. Their axons terminate throughout the VRC, where both rhythm-generating neurons (pre-BötC) and bulbospinal premotor neurons (rVRG and cVRG) are potential targets.

In addition to inhibitory pump cells, there is probably a population of excitatory pump neurons. The excitatory and inhibitory pump neurons may both contribute to Breuer–Hering inspiratory-inhibiting and expiratory-facilitatory reflexes. VRC I-Dec and E-Dec (or post-I) neurons (Figure 2) are inhibitory and influence inspiratory and expiratory duration. It appears that I-Dec neurons inhibit expiratory neurons and E-Dec neurons (glycinergic) inhibit inspiratory neurons. Their declining discharge pattern represents a declining inhibition that, when low enough, permits the transition from expiration to inspiration or vice versa. Consequently, the inhibition of I-Dec neurons by pump neurons contributes to a more rapid decline in the discharge of I-Dec neurons and a shortening of inspiratory duration, thereby decreasing tidal volume. In contrast, SAR activation provides strong excitation of E-Dec neurons (probably from excitatory pump neurons), and subsequent E-Dec inhibition of inspiratory neuron activity lengthens expiratory duration.

Central Pathways of Rapidly Adapting Stretch Receptors

RAR afferent fibers monosynaptically activate NTS excitatory neurons termed RAR interneurons (Figure 6). Like pump neurons, these interneurons provide extensive axonal arborizations to VRC and PRG neurons. Despite the extensive arborization pattern of RAR interneurons, the identity of targeted respiratory neuron subtypes is not yet known. These targets presumably include neurons that facilitate inspiration because RAR receptors evoke reflexes that are preceded by large inspirations, such as coughs and augmented breaths (sighs). Sighs are triggered in response to decreased lung compliance caused by alveolar collapse; the ensuing deep inspiration reopens collapsed alveoli.

Deflation-Activated Receptor-Mediated Reflexes

Lung deflation causes expiratory termination and increased inspiratory effort. This reflex is particularly important in young mammals, including human infants, or other small mammals that have highly a compliant chest wall. In the absence of compensating muscular activity, increased chest wall compliance permits greater inward recoil of the lung during passive expiration, leading to low end-expiratory lung volumes. Low lung volumes have potentially negative implications, including the greater likelihood of alveolar collapse and impaired gas exchange. This tendency is offset by slowing expiratory airflow via glottal adduction and/or braking activity (postinspiratory activation) of the diaphragm and earlier onset of the subsequent inspiratory effort. Lung deflation alters the activity of respiratory neurons controlling all these factors, producing appropriate changes in airway caliber as well as decreased expiratory duration and increased peak phrenic nerve activity.

DAR afferent fibers terminate in the NTS, where they activate two general classes of excitatory (glutamatergic) neurons. One type, DAR relay neurons, is analogous to the pump cells in having a discharge pattern determined almost completely by their primary afferent input. They are localized caudally within the commissural subnucleus of the NTS, immediately medial to the solitary tract, and are inhibited by input from GABAergic/glycinergic pump cells. The second class of neurons receiving DAR afferent input is a subset of DRG inspiratory neurons, termed inspiratory-γ (Iγ) neurons. Iγ neurons project to the spinal cord and, although their spinal targets have not yet been identified, their inspiratory facilitatory effect suggests that phrenic motor neurons are the likely targets. Like pump cells and RAR interneurons, DAR relay neurons give rise to axons that arborize extensively in the VRC and PRG, although the target neuron identities are not yet known.

Central Pathways of Peripheral Chemoreceptors

Chemoreceptors in the carotid bodies and aortic arch are sensitive to changes in arterial carbon dioxide, oxygen, and pH. The carotid bodies are generally more important in mediating this response and provide the principal mechanism by which mammals sense lowered levels of oxygen. Afferents from the carotid body travel in the carotid sinus nerve, a branch of the glossopharyngeal cranial nerve. The latter enters the brain stem at the posterior levels of the facial nucleus and travels caudally to join the solitary tract. The afferents terminate in caudal portions of the NTS, mainly within the commissural and medial subdivisions. A variety of transmitters and modulators are found in carotid chemoreceptor afferent fibers, including dopamine, histamine, adenosine, serotonin, and glutamate. Subsequent projections from the NTS to the VRC and particularly to the RTN appear to be glutamatergic.

Summary

From a physiological perspective, the idiom "as easy as breathing" is an oxymoron because breathing is an extraordinarily complex event that requires multiple layers of organization and integration. At odds with this complexity, our individual personal experiences, when healthy, mislead us into thinking it is a simple process because it is so automatic and reliable, leading to a smooth integration of respiratory activity into nearly every facet of our daily lives. To the contrary, the entire process depends on multiple structural systems that have evolved to serve gaseous homeostasis and on the parallel evolution of behavioral systems that avail themselves of voluntary and reflex components of breathing and airway regulation in order to eat, drink, speak, swim, walk, run, copulate, defecate, vomit, and the myriad other activities that require adjustments in respiratory rhythm and pattern so that they may also proceed in a smoothly integrated and coordinated fashion. It is usually only under challenges, such as during sustained physical activity, disease, or emotional stress that we begin to appreciate that breathing is not easy. The consequences of failure at this essential task are devastating.

See also: Autonomic Nervous System: Respiratory Control; Autonomic Nervous System: Central Respiratory Control; Brainstem Respiratory Circuits; Evolution of Vertebrate Respiratory Control; Respiration.

Further Reading

Coleridge JC and Coleridge HM (1984) Afferent vagal C fibre innervation of the lungs and airways and its functional significance. *Reviews of Physiology Biochemistry and Pharmacology* 99: 1–110.

Feldman JL and McCrimmon DR (2003) Neural control of breathing. In: Squire LR, Bloom FE, Roberts JL, Zigmond MJ, McConnell SK, and Spitzer NC (eds.) *Fundamental Neuroscience*, 2nd edn., pp. 967–990. New York: Academic Press.

Hage SR and Jürgens U (2006) On the role of the pontine brainstem in vocal pattern generation: A telemetric single-unit recording study in the squirrel monkey. *Journal of Neuroscience* 26: 7105–7115.

Huey KA, Szewczak JM, and Powell FL (2003) Dopaminergic mechanisms of neural plasticity in respiratory control: Transgenic approaches. *Respiratory Physiology and Neurobiology* 135: 133–144.

Kubin L, Alheid GF, Zuperku EJ, and McCrimmon DR (2006) Central pathways of pulmonary and lower airway vagal afferents. *Journal of Applied Physiology* 101: 618–627.

Lazarov N, Rozloznik M, Reindl S, Rey-Ares V, Dutschmann M, and Gratzl M (2006) Expression of histamine receptors and effect of histamine in the rat carotid body chemoafferent pathway. *European Journal of Neuroscience* 24: 3431–3444.

Lee LY, Lin YS, Gu QH, Chung E, and Ho CY (2003) Functional morphology and physiological properties of bronchopulmonary C-fiber afferents. *Anatomical Record* 270A: 17–24.

Mathew OP and Ghosh TK (1995) Role of airway afferents on upper airway muscle activity. In: Dempsey JA, and Pack AI (eds.) *Regulation of Breathing,* 2nd edn., pp. 511–541. New York: Dekker.

McCrimmon DR, Alheid GF, and Zuperku EJ (2006) Reflexes from the lungs and chest wall. In: Laurent GJ and Shapiro SD (eds.) *Encyclopedia of Respiratory Medicine*, pp. 618–626. Oxford: Academic Press.

Otake K, Nakamura Y, Tanaka I, and Ezure K (2001) Morphology of pulmonary rapidly adapting receptor relay neurons in the rat. *Journal of Comparative Neurology* 430: 458–470.

Sant'Ambrogio G, Tsubone H, and Sant'Ambrogio FB (1995) Sensory information from the upper airway: Role in the control of breathing. *Respiration Physiology* 102: 1–16.

Thach BT (2001) Maturation and transformation of reflexes that protect the laryngeal airway from liquid aspiration from fetal to adult life. *American Journal of Medicine* 111(supplement 8A): 69–77.

Widdicombe J (2001) Airway receptors. *Respiration Physiology* 125: 3–15.

Widdicombe JG (1986) Reflexes from the upper respiratory tract. In: Fishman AP, Cherniack NS, Widdicombe JG, and Geiger SR (eds.) *Handbook of Physiology, The Respiratory System, Vol. 2: Control of Breathing*, pp. 363–393. Bethesda, MD: American Physiological Society.

Yu J (2005) Airway mechanosensors. *Respiration Physiology & Neurobiology* 148: 217–243.

Neural Cell Adhesion Molecules and Synapse Regulation

M Schachner, I Leshchyns'ka, and V Sytnyk, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

© 2009 Elsevier Ltd. All rights reserved.

The Neural Cell Adhesion Molecule

The neural cell adhesion molecule (N-CAM) was discovered as the first member of the immunoglobulin superfamily in neural cells. It occurs in three major isoforms and minor, often very cell-type-characteristic, alternatively spliced variations. All three major isoforms of N-CAMs share the same extracellular domain encompassing five immunoglobulin-like (Ig-like) domains and two fibronectin type III modules. Two transmembrane isoforms with apparent molecular masses of 180 and 140 kDa (designated N-CAM180 and N-CAM140, respectively) contain a single transmembrane region and intracellular domains of different lengths. The shortest N-CAM isoform, with the apparent molecular mass of 120 kDa (designated N-CAM120), is attached to the plasma membrane via a glycosylphosphatidylinositol anchor, predisposing the localization of N-CAM120 to lipid-enriched microdomains, which are also called lipid rafts and which are distinct from the lipid microenvironment of N-CAM140 and N-CAM180 (Figure 1).

N-CAM is unique in that it carries an unusual carbohydrate, the α-2,8-linked polysialic acid, which has not been found on any other recognition molecule. Polysialylation and isoform expression undergo changes during development. Polysialylation decreases during embryonic progression to the adult stage, but remains prevalent at sites where neurogenesis is active, such as on neural stem cells in the dentate gyrus of the hippocampus and the rostral migratory stream emanating from the subventricular zone of the lateral ventricles and terminating in the olfactory bulb.

N-CAM was discovered as a homophilically interacting (i.e., self-binding) molecule. Reports from different laboratories have shown that homophilic binding between N-CAM molecules is accomplished via antiparallel association of all five immunoglobulin-like domains, the first and the second immunoglobulin-like domains, or only via interaction between the third immunoglobulin-like domains of the molecule. Polysialic acid regulates homophilic interactions of N-CAMs in that it reduces these interactions, supposedly by its highly negatively charged sialic acid residues, in addition to a large hydration volume surrounding the glycan chains. Besides *trans*-interactions (i.e., interactions between molecules located on opposed cell surfaces), *cis*-interactions between N-CAMs located in the plasma membrane of the same cell can also occur. These may modulate *trans* homophilic interactions between N-CAMs.

N-CAM also interacts with non-self-binding partners in so-called heterophilic *cis*- or *trans*-interactions. Among the heterophilic binding partners are the receptors for fibroblast growth factor and the glial cell line-derived neurotropic factor receptor, both of which are involved in N-CAM-dependent signal transduction. N-CAM interacts with other members of the immunoglobulin superfamily, such as L1, with other recognition molecules, such as prion protein, and with extracellular matrix molecules, such as phosphacan.

N-CAM and Synaptic Plasticity

The polysialylated form of N-CAM and the protein isoforms of N-CAM are differentially distributed in cell surface membranes. While N-CAM180 is mainly, if not exclusively, found on postsynaptic membranes in the central nervous system, N-CAM140, the smaller transmembrane isoform, is not only found postsynaptically, but also presynaptically. The glycosylphosphatidylinositol (GPI)-anchored isoform of N-CAM is mainly, if not exclusively, found on glial cells which abut onto the pre- and postsynaptic membranes as potentially homophilic interaction partners, thus providing a molecular basis for a glial influence on synaptic activity. Interestingly, the levels of expression of both N-CAM140 and, in particular, N-CAM180 depend on synaptic activity.

The first observation suggesting that expression of N-CAM180 in postsynaptic densities of some, but not all, synapses between axons of granule cells and Purkinje cell dendrites in the molecular layer of the adult mouse cerebellum pointed to a particular functional distinction between these synapses. Together with the observation that N-CAM180 is less mobile in the surface membrane than N-CAM140 is, this finding gave rise to the idea that N-CAM180 is well poised to immobilize cell surface components at crucially interacting apposing cell surface membranes at the synapse. An increase in the proportion of hippocampal spine synapses expressing N-CAM180 following long-term potentiation in rats verified this conclusion. In chickens, it was also observed that training increased the proportion of synaptic active zones containing N-CAM in the lobus paraolfactorius of the striatum and altered the distribution of N-CAM in synaptic membranes in the hyperstriatum ventrale of chicken forebrain.

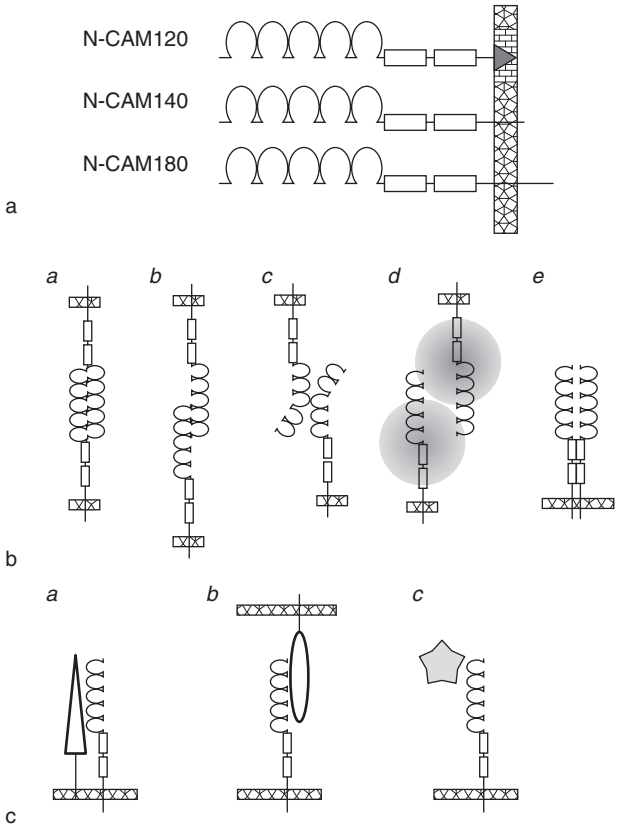

Heterophilic binding partners of N-CAM:

Receptors:	FGFR (*a*), GFRα (*a*)
Other recognition molecules:	L1 (*a,b*), RPTPζ/β (*b*), prion protein (*a,b*)
Extracellular matrix molecules:	phosphacan (*c*), heparin (*c*), heparan sulfate proteoglycans (*c*)

Figure 1 Structure and homo- and heterophilic interactions of N-CAMs. (a) The three major N-CAM isoforms are shown. All isoforms share the same extracellular part, composed of five immunoglobulin-like and two fibronectin type III domains. N-CAM140 and N-CAM180 are transmembrane isoforms, with N-CAM180 containing the largest intracellular domain. N-CAM120 is anchored to the plasma membrane via glycosylphosphatidylinositol and is located in lipid-enriched microdomains which are distinct from the lipid microenvironment of N-CAM140 and N-CAM180. (b) A scheme of the homophilic interactions of N-CAM is shown. Homophilic *trans*-interactions of N-CAM (*a*–*d*) can be mediated by all five immunoglobulin-like domains (*a*), the first and the second immunoglobulin-like domains (*b*), or only the third immunoglobulin-like domain (*c*). Polysialic acid (shown in gray) attenuates homophilic interactions of N-CAM (*d*). N-CAM molecules can also interact with each other in a *cis* configuration, being located in the surface plasma membrane of one cell (*e*). (c) A scheme of the heterophilic interactions of N-CAM is shown. N-CAM can associate with heterophilic binding partners located in the same plasma membrane (*a*), on adjacent membranes (*b*), or in the extracellular matrix (*c*). Bottom: Examples of the heterophilic binding partners of N-CAM, with the type of their heterophilic interactions with N-CAM (as described in (c)) indicated in parentheses.

Not only the number of synapses expressing N-CAMs, but also the distribution of N-CAMs, and in particular N-CAM180, within postsynaptic densities is changed in an activity-dependent manner, correlating with changes in localization of the NR2A subunit of the *N*-methyl-D-aspartate (NMDA) receptor subtype of glutamate receptors. This is not the case for the GluR2/3 subunit of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor subtype and the mGluR1 subunit of the metabotropic glutamate receptor: in synaptic membranes of nonstimulated hippocampal spine synapses, N-CAM180 and NR2A accumulated in the center of the postsynaptic density, whereas GluR2/3 and mGluR1 were distributed evenly. Twenty-four hours following induction of long-term potentiation *in vivo*, N-CAM180 and NR2A accumulated at the edges of postsynaptic densities, whereas GluR2/3 localized more centrally.

Functional validation of the importance of N-CAM in synaptic plasticity derived from observations that antibodies against N-CAM interfered with hippocampal long-term potentiation in slice preparations *in vitro*. Recovery of emotional behavior in N-CAM-deficient mice through transgenic expression of N-CAM180 driven by the neurofilament promoter

showed that N-CAM180 is indeed involved in certain types of synaptic interactions. Further functional links between synaptic activity and N-CAM expression were provided by observations on N-CAM-deficient mice which showed increased anxiety and serotonin 1A receptor response. Furthermore, mice constitutively deficient in N-CAM and mice in which N-CAM was conditionally ablated in the second and third postnatal week in excitatory neurons, under the control of calcium/calmodulin kinase IIα (CaMKII) promoter, showed reduced long-term potentiation and long-term depression *in vitro* and abnormalities in behavior. Mice deficient in the two polysialyltransferases responsible for N-CAM polysialylation also showed altered synaptic plasticity *in vitro* and abnormalities in learning and memory.

The combined observations emphasize the importance of N-CAM and its isoform N-CAM180 in synaptic plasticity. The exact mechanisms by which N-CAM participates in the regulation of synaptic efficacy remain poorly understood, but recent findings indicate that N-CAM may play an important role in this scenario by activating signal transduction pathways, remodeling of the cytoskeleton in postsynaptic densities, and/or gene transcription. These findings are discussed in greater detail in the following sections.

N-CAM Is a Signal Transducer

Early findings indicated that N-CAM, like other transmembrane cell surface receptors, is indeed a signal transducer. Activating N-CAM at the cell surface by using N-CAM antibodies or triggering homophilic N-CAM *trans*-interactions produced signaling responses in the N-CAM-expressing cells, including enhanced calcium influx via G-protein-dependent channel activation, pH change, and phosphoinositide turnover, all of which had been suggested to contribute to neurite outgrowth and synapse function. Additional observations that N-CAM180, the largest N-CAM transmembrane isoform, is accumulated at sites of focused signal transduction – namely in the postsynaptic density, where cytoskeletal elements act together with pivotal signal-transducing enzymes – fostered further activities aiming to elucidate signal transduction pathways that were first discovered to influence neuronal migration and neurite outgrowth, but were later found to be pivotal in signal transduction in synapses (Figure 2).

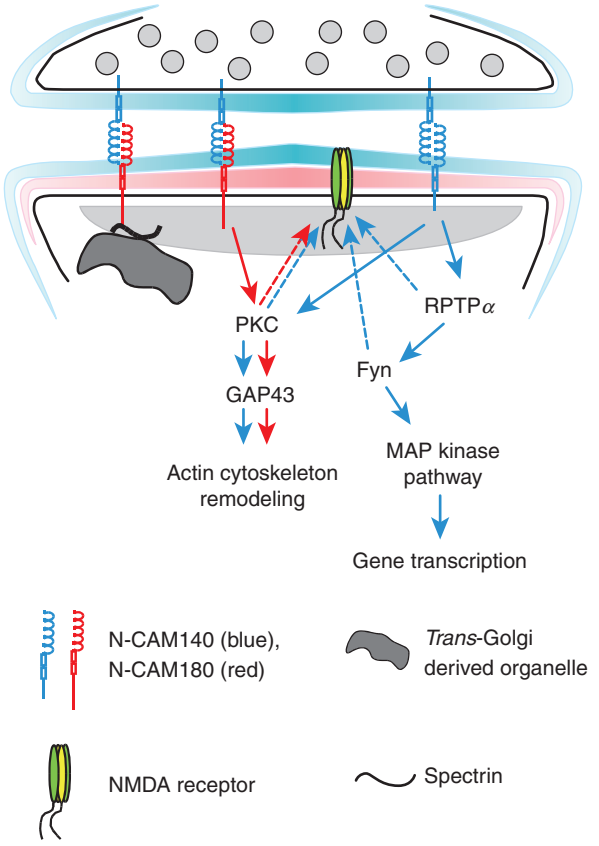

Figure 2 Schematic diagram showing the distribution and multiple functions of N-CAM in synapses and postsynaptic densities. The topography of the distribution of N-CAM140 and N-CAM180 in synapses is indicated in blue and red, respectively. N-CAM140 is present both pre- and postsynaptically, while N-CAM180 is present only postsynaptically. Homophilic and probably heterophilic (not shown) interactions of N-CAM in the synaptic cleft contribute to synapse stabilization and modulation of function. Via the intracellular domain, N-CAM180 associates with spectrin-enriched *trans*-Golgi-derived organelles, and anchors this complex in the postsynaptic density. Homophilic and heterophilic interactions of N-CAM140 and N-CAM180 activate signaling cascades in the postsynaptic density. Two of the most well-studied pathways, including MAP kinase pathway activation via RPTPα/p59fyn (blue arrows) and PKC/GAP43 through N-CAM140 (blue arrows) and N-CAM180 (red arrows), are shown. These pathways result in activation of gene transcription in the nucleus and cytoskeleton remodeling in the postsynaptic density, respectively. PKC, RPTPα, and p59fyn regulate the phosphorylation of other postsynaptic proteins, including NMDA receptors, thus providing additional mechanisms for N-CAM-mediated control of synapse efficacy. Signaling pathways that have been characterized are shown with solid arrows; hypothetical signaling pathways are shown with dashed arrows.

N-CAM Associates with and Reorganizes the Cytoskeleton

A prerequisite for the precise topographic distribution of N-CAM isoforms in the postsynaptic membrane derives from the capacity of N-CAM180 and N-CAM140 to interact with the cytoskeleton. N-CAM binds to the membrane cytoskeleton linker protein spectrin, which is one of the most abundant proteins in the postsynaptic density and an important platform for docking of receptors and ion channels. In initial experiments, N-CAM180, because of its direct interaction with spectrin, was shown to be less mobile than N-CAM140 within the surface membrane. However,

more refined techniques have recently shown that all three isoforms of N-CAM accumulate, co-localize, with and co-immunoprecipitate spectrin. In another recent study, α- and β-tubulin as well as α-actinin were observed to associate with N-CAM180 and N-CAM140. Moreover, N-CAM180, but not N-CAM140, bound to microtubule-associated protein 1a, β-actin, and tropomyosin. These proteins are also associated with the cytoskeleton, particularly in postsynaptic compartments.

Spectrin is a pivotal platform for activation of neuromodulatory receptors at the cell surface. The NMDA glutamate receptor subtype regulating synaptic plasticity in long-term potentiation and depression is cross-linked to N-CAM via spectrin. Furthermore, CaMKIIα, highly enriched in postsynaptic densities and involved in synaptic plasticity, is also bound to the spectrin platform. Clustering of N-CAM induces formation of detergent-insoluble complexes enriched in postsynaptic proteins and resembling postsynaptic densities. Disruption of the N-CAM/spectrin complex in N-CAM-deficient mice or by an interfering spectrin fragment in wild-type cells decreases the size of postsynaptic densities and reduces synaptic targeting of N-CAM/spectrin-associated postsynaptic proteins, not only including spectrin, but also the NMDA receptor and CaMKIIα. Furthermore, N-CAM-deficient neurons are unable to recruit CaMKIIα to synapses in response to NMDA receptor activation. These observations indicate that N-CAM regulates formation of spectrin-based postsynaptic specializations with one of its quantitatively most important ingredients – CaMKIIα – to promote activity-dependent postsynaptic rearrangements leading to long-lasting changes in synaptic strength.

Clustering of N-CAM Induces Signaling through Lipid Rafts

Spectrin plays an important role in the interaction of N-CAM with kinases and phosphatases in lipid rafts, which are lipid-enriched microdomains serving as lipid-based signaling platforms in postsynaptic densities. N-CAM120 redistributes spectrin predominantly to lipid rafts, while N-CAM140 and N-CAM180/βI spectrin complexes do not depend on raft integrity and are located both in rafts and raft-free membrane domains, thus suggesting the possibility of integration of different signal transduction mechanisms. Spectrin acts as a bridge between the transmembrane N-CAM isoforms and protein kinase C β2 (PKCβ2) and the receptor-like protein tyrosine phosphatase α (RPTPα). Clustering of N-CAM enhances the formation of the N-CAM/PKCβ_2 and N-CAM/RPTPα complexes via spectrin. Interestingly, redistribution of PKCβ2 and RPTPα to N-CAM/spectrin complexes is blocked by a specific fibroblast growth factor receptor inhibitor. Activation of N-CAM is accompanied by a redistribution of N-CAM and N-CAM-associated proteins to lipid rafts, where N-CAM interacts with the growth-associated protein 43 (GAP43), which is highly expressed not only in the developing nervous system, but also in synapses in the adult. GAP43 promotes actin reorganization, which is regulated by its phosphorylation through protein kinase C. RPTPα, a known activator of lipid-raft-localized protein tyrosine kinase p59fyn, links N-CAM to this kinase. Association between N-CAM and p59fyn is lost in RPTPα-deficient brains, and is disrupted by dominant-negative RPTPα mutants, showing that RPTPα functions as a link between N-CAM and p59fyn. N-CAM-mediated p59fyn activation is abolished in RPTPα-deficient neurons, implicating RPTPα as a major phosphatase involved in N-CAM-mediated signaling.

p59fyn is an organizer of signal transduction pathways activated by clustering of N-CAM molecules at the cell surface. Phosphorylation of p59fyn activates the focal adhesion kinase (FAK). After activation of p59fyn and FAK by N-CAM, further downstream signaling is induced, resulting in activation of mitogen-activated protein kinases (MAPKs). Activation of the Erk/MAPKs commences at the cell surface also by activation of the small guanosine triphosphate (GTP)-binding protein Ras. From Ras, the signal progresses through sequential activation of the serine/threonine kinases Raf and MAP extracellular kinase (MEK), leading to phosphorylation of Erk/MAPKs. Activation of MEK, the kinase immediately upstream of extracellular signal-regulated kinases ERK-1 and ERK-2, rescues the inhibition of p59fyn. A docking protein, ShcA, has been shown to contribute to p59fyn/FAK-mediated MAPK activation in integrin signaling by recruiting the adaptor protein Grb2, which associates with the GTP exchange factor SOS to induce Ras activation. In addition, MAPK activation by N-CAM may result from an increase in cytoplasmic calcium, which induces activation of MAPK, and/or by PKC, which can activate Raf.

More indirect ways, also probably involving the spectrin cytoskeleton, have become evident by the ability of N-CAM to regulate cell surface delivery of G-protein-activated inwardly rectifying potassium channels via lipid rafts. Mice deficient in N-CAM exhibited a strikingly increased sensitivity to serotonin 1A receptor agonists. Studies of the relationship between N-CAM and serotonin receptor signaling pathways mediating G-protein-activated inwardly rectifying potassium (Kir3) channels showed that delivery of these channels to the cell surface is increased in N-CAM-deficient mice as opposed to wild-type mice.

N-CAM140 and N-CAM180 specifically reduced inwardly rectifying currents generated by neuronal, but not cardiac, channels through impairment of cell surface transport of these channels. Interestingly, mutational and pharmacological disruption of the lipid raft association of N-CAM140, including elimination of palmitoylation-targeted delivery of N-CAM to lipid rafts, normalizes surface delivery of channels. This observation alludes to a novel mechanism controlling long-term excitability of neurons.

These observations indicate a complex interplay between N-CAM function, neurotransmitter and cytokine receptors, ion channels, kinases, and receptor tyrosine phosphatases that themselves act as recognition molecules and impinge on the cytoskeleton and common signaling cascades of the MAP kinase pathway and transcriptional elements.

N-CAM Activates Gene Transcription

Cyclic AMP response element-binding protein (CREB) activates genes that contain regulatory cre-elements, such as the immediate-early gene c-*fos*, and both CREB and the activated genes are important for N-CAM-mediated neurite outgrowth. CREB is activated by the MAPK pathway, leading to its phosphorylation and activation of transcription. Signaling from N-CAM to CREB also proceeds through protein kinase A (PKA), since inhibition of PKA inhibits N-CAM-dependent neurite outgrowth, and neither activation of PKA or MAPK rescues inhibition of N-CAM-mediated neurite outgrowth by expression of a dominant-negative CREB protein. These findings imply that PKA and MAPK signaling converges on CREB activation. It has been suggested that calcium influx is an important ingredient in stimulating cAMP production to activate PKA, which in turn is required for nuclear translocation of MAPK. PKA also induces MAPK activation through the Ras-related small GTP-binding protein Rap1. Implicating yet another N-CAM-related molecule, it has been found that CaMKII, together with PKA, can directly phosphorylate CREB to regulate its activity. CaMKII also regulates the transcriptional activator nuclear factor-kappa B (NF-κB), which has been shown to be activated by N-CAM functions. A link from the MAPK pathway to transcription involves phosphorylation of the transcription factor ELK-1. The genes activated by ELK-1 and NF-κB, as well as those activated by CREB in response to N-CAM triggering in synapses, remain to be identified.

N-CAM Regulates Synapse Formation

N-CAM is not only important in the functioning of mature synapses, but also during synapse development, as observed in cell culture of hippocampal neurons. Time-lapse video recordings in cell cultures of hippocampal neurons showed that N-CAM accumulates at initial contacts between an axon and dendrite within several minutes after contact establishment. These initial contacts then often give rise to synapses, but can be disassembled via retraction of an axon or dendrite, with both neurites separating from each other. Interestingly, absence of N-CAM in cultured neurons from N-CAM-deficient mice leads to an enhanced disassembly of initial contacts within several minutes after contact establishment, when compared to wild-type neurons, indicating that N-CAM is required to stabilize interacting neurites.

Again, the mechanisms of the transformation of a contact between an axon and a dendrite into a fully functional synapse depend on N-CAM. *Trans*-Golgi network organelles that are linked via spectrin to the initially formed clusters of N-CAM in the plasma membrane are vitally important for the transformation of contacts to nascent synapses. The complexes of N-CAM and *trans*-Golgi network organelles are translocated along neurites of cultured hippocampal neurons and are trapped at sites of initial neurite-to-neurite contacts. The accumulation of *trans*-Golgi network organelles at contacts with neurons from N-CAM-deficient mice is reduced when compared to neurons from wild-type mice, confirming that N-CAM mediates the anchoring of intracellular organelles at nascent synapses.

To evaluate the contribution of the pre- versus postsynaptic expression of N-CAM in regulation of synaptic efficacy, dissociated hippocampal cells from N-CAM-deficient and wild-type mice were cultured in homo- and heterogenotypic combinations. Electrophysiological recordings from synaptically coupled neurons showed that synaptic strength of excitatory, but not inhibitory, synapses depends on expression of N-CAM postsynaptically, but not presynaptically, correlating with higher levels of potentiation and synaptic coverage of N-CAM-expressing neurons compared to N-CAM-deficient neurons in these heterogenotypic co-cultures. Synaptic density was the same in homogenotypic cultures of N-CAM-deficient and wild-type neurons as well as in heterogenotypic co-cultures in which several types of glutamate receptors were pharmacologically inhibited. These observations indicate that the relative levels of postsynaptic N-CAM expression controlled synaptic strength in an activity-dependent manner by regulating the number of synapses.

While postsynaptic N-CAM promotes synapse formation, presynaptic N-CAM is involved in organization of the presynaptic machinery as revealed by analysis of neuromuscular junctions (NMJs) in N-CAM-deficient mice. N-CAM-deficient NMJs

show multiple alterations in the presynaptic organization including reduced presynaptic accumulation of Ca^{2+} channels, Rab 3-interacting molecule (RIM) and syntaxin. N-CAM-deficient mice exhibit transmission failures in NMJs, subjected to repetitive stimulation, and locomotor behavioral defects. Presynaptically, N-CAM140 plays the major role as revealed by analysis of primarily N-CAM180-deficient mice. However, deficiency in N-CAM180 also affects molecular organization of the presynaptic terminals.

The role of polysialic acid in the synaptogenic activity of N-CAM was also studied in cell culture. Conveniently, a very specific enzyme removing only polysialic acid, but not other glycan chains, from the N-CAM protein backbone abolished preferential formation of synapses on N-CAM-expressing neurons in heterogenotypic co-cultures of wild-type and N-CAM-deficient hippocampal neurons. Expression of any of the three major isoforms of N-CAM in N-CAM-deficient neurons stimulated preferential synapse formation on N-CAM-expressing neurons. Additionally, removal of heparan sulfates from the cultured neurons and mutation in the heparin-binding domain of N-CAM diminished synaptogenic activity of neuronally expressed N-CAM carrying polysialic acid, suggesting that interaction of N-CAM with heparan sulfate proteoglycans mediates synapse formation. N-CAM-driven synaptogenesis was also inhibited by antagonists to the fibroblast growth factor receptor and the NMDA receptor, but not by blockers of non-NMDA glutamate receptors and by blockers of voltage-dependent sodium channels. Enzymatic removal of polysialic acid and heparan sulfates also inhibited the increase in the number of perforated synapses associated with NMDA receptor-dependent long-term potentiation in the CA1 region of organotypic hippocampal cultures. Thus, neuronally expressed N-CAM, with its associated polysialic acid in complex with heparan sulfate proteoglycans, promotes synaptogenesis and activity-dependent remodeling of synapses.

Dysfunction of N-CAM Leads to Psychiatric Disorders

Disturbances of N-CAM function in forming synapses and at mature synapses, studied both in cell culture and in intact animals with the help of N-CAM-specific mutations, show wider implications, not only in laboratory animals, but also in humans: genetic variations in the N-CAM gene have been shown to confer risks associated with bipolar affective disorders. Furthermore, the hippocampus and prefrontal cortex of patients with schizophrenia and bipolar disorders accumulate the secreted extracellular cleavage products of N-CAMs in tissue and cerebral spinal fluid, most likely interfering with homophilic interactions of N-CAMs. Accordingly, transgenic mice ectopically expressing secreted N-CAM display disturbed synaptic connectivity and abnormal behavior, with abnormalities that are particularly relevant to schizophrenia. Thus, disruption of the many functions of N-CAMs by mutations may affect the structural organization of postsynaptic densities and reduce the efficacy of synapse formation and maintenance, possibly being directly relevant to the etiology of psychiatric disorders. The combined observations establish N-CAM as an important player in the assembly and function of a unique invention: the synapse.

See also: Bipolar Disorder; Calcium-Calmodulin Kinase II (CaMKII) in Learning and Memory; Cell Adhesion Molecules at Synapses; Cytoskeleton in Plasticity; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Dopamine; Long-Term Depression: Cerebellum; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation.

Further Reading

Bodrikov V, Leshchyns'ka I, Sytnyk V, et al. (2005) RPTPalpha is essential for NCAM-mediated p59fyn activation and neurite elongation. *Journal of Cell Biology* 168: 127–139.

Bukalo O, Fentrop N, Lee AY, et al. (2004) Conditional ablation of the neural cell adhesion molecule reduces precision of spatial learning, long-term potentiation, and depression in the CA1 subfield of mouse hippocampus. *Journal of Neuroscience* 24: 1565–1577.

Delling M, Wischmeyer E, Dityatev A, et al. (2002) The neural cell adhesion molecule regulates cell-surface delivery of G-protein-activated inwardly rectifying potassium channels via lipid rafts. *Journal of Neuroscience* 22: 7154–7164.

Dityatev A, Dityateva G, Sytnyk V, et al. (2004) Polysialylated neural cell adhesion molecule promotes remodeling and formation of hippocampal synapses. *Journal of Neuroscience* 24: 9372–9382.

Kleene R and Schachner M (2004) Glycans and neural cell interactions. *Nature Reviews Neuroscience* 5: 195–208.

Kiselyov VV, Soroka V, Berezin V, et al. (2005) Structural biology of NCAM homophilic binding and activation of FGFR. *Journal of Neurochemistry* 94: 1169–1179.

Leshchyns'ka I, Sytnyk V, Morrow JS, et al. (2003) Neural cell adhesion molecule (NCAM) association with PKCbeta2 via betaI spectrin is implicated in NCAM-mediated neurite outgrowth. *Journal of Cell Biology* 161: 625–639.

Panicker AK, Buhusi M, Thelen K, et al. (2003) Cellular signalling mechanisms of neural cell adhesion molecules. *Frontiers in Bioscience* 8: d900–d911.

Sytnyk V, Leshchyns'ka I, Dityatev A, et al. (2004) Trans-Golgi network delivery of synaptic proteins in synaptogenesis. *Journal of Cell Science* 117: 381–388.

Vawter MP (2000) Dysregulation of the neural cell adhesion molecule and neuropsychiatric disorders. *European Journal of Pharmacology* 405: 385–395.

Neural Circuitry in the Somatosensory System

K D Alloway, The Pennsylvania State University, Hershey, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Somatosensory Circuits

Cutaneous stimulation evokes neuronal responses in the central nervous system that are tightly coordinated both spatially and temporally. As in other neural systems, the coordination of stimulus-induced neuronal activity in the somatosensory system is dictated primarily by the detailed connections of its neural circuits. Within this context, the main factors that influence neuronal coordination in the somatosensory system include its topographic organization, the morphology and density of its synaptic contacts, and the biophysical properties of the neurons embedded within its neural circuits. Although the coordination of neural activity is constrained by the connectivity of the neuronal substrate, somatosensory circuit operations are dynamic. Consequently, the functional strength of the connections between brain regions or between specific pairs of neurons in the somatosensory system can fluctuate with changes in stimulation, neuronal bursting, corticofugal processing, and other factors.

Ascending Somatosensory Pathways

The somatosensory system consists of the ascending neuronal pathways that convey somesthetic information to the cerebral cortex as well as the descending corticofugal projections from the somatosensory cortex. Hence, awareness of the body and its interactions with the environment requires the transmission of kinesthetic and tactile information through the peripheral nerves and then through a series of brain regions located progressively higher in the neuraxis. Innocuous stimulation of the skin, for example, activates peripheral nerve fibers that, depending on the body part that is stimulated, terminate in one of the following brain stem regions: the nucleus gracilis (leg and lower trunk), the nucleus cuneatus (arm and upper trunk), or the trigeminal nuclei (face and head). Each of these brain stem nuclei gives rise to ascending axons that transmit cutaneous information to neurons in the ventrobasal part of the contralateral thalamus, which includes the ventroposteromedial (VPM) nucleus for the head representation and the ventroposterolateral (VPL) nucleus for the trunk and limbs. In turn, the ventrobasal thalamus transmits cutaneous information to the primary somatosensory cortex and surrounding regions in the parietal lobe.

The somatosensory representations in the brain stem, thalamus, and cortex have a precise somatotopic organization that has important consequences for processing spatial information. Each somatosensory region contains a spatially organized map of the entire body, and within each map, adjacent neuronal populations represent neighboring sites on the body. The ascending connections through successive levels of the neuraxis act as linkages between neuronal populations that represent the same part of the body. For example, the thalamic neurons that represent the thumb have many synaptic contacts on the cortical neurons that represent the thumb, but have few synaptic contacts on cortical neurons that represent the palm or the arm.

Time-Locked Neuronal Responses

Tactile stimulation evokes a time-locked sequence of neuronal activity that reflects the sequential order of the ascending somatosensory pathways. Following punctate stimulation of the hand, for example, microelectrode recordings demonstrate that neurons in the nucleus cuneatus have response latencies (i.e., time delays) that are shorter than those recorded in the ventrobasal thalamus. Likewise, thalamic response latencies are shorter than those recorded in the somatosensory cortex. Therefore, response latencies in the somatosensory system increase both with the length of the axonal route and with the number of intervening synapses that are traversed as stimulus-induced neural activity advances from the periphery to the somatosensory cortex.

Local Inhibitory Circuits

The somatosensory nuclei in the brain stem and thalamus contain elementary circuits in which inhibitory interneurons modulate the transmission properties of the relay neurons. By comparison, the somatosensory cortex contains even more complicated inhibitory circuits that modulate the response properties of the pyramidal neurons. In fact, the majority of synaptic contacts on the pyramidal soma originate from local inhibitory neurons.

In all levels of the somatosensory system, inhibitory connections provide a mechanism for sharpening the temporal and spatial tuning properties of the excitatory neurons. Lateral inhibitory connections, for example, regulate the spatial extent of neuronal responsiveness. This view is based on experiments showing that local blockade of γ-aminobutyric acid (GABAergic) transmission substantially increases the receptive field size of individual neurons in each level of the somatosensory system.

Recurrent inhibition also plays a major role in processing somatosensory information. In the thalamus, for example, projections from the reticular nucleus form recurrent inhibitory connections with the thalamocortical relay neurons in VPM or VPL. Composed exclusively of a slender band of GABAergic neurons along the lateral edge of the thalamus, the output of the reticular nucleus modulates the spatial and temporal activity of the thalamocortical relay neurons. In addition to receiving collateral excitatory inputs from thalamocortical relay neurons, the reticular nucleus also receives collateral inputs from corticothalamic feedback projections. By activating the reticular nucleus, corticothalamic projections indirectly inhibit selective parts of the ventrobasal thalamus while directly activating other parts of the thalamus. Hence, coordinated interactions between the corticothalamic projections and the recurrent inhibitory circuits of the reticular nucleus provide a mechanism for selectively filtering or enhancing transmission of tactile information to the cerebral cortex.

Somatosensory Neuronal Interactions

To characterize neuronal coordination within a neural circuit, neurophysiologists record the action potentials of several neurons simultaneously and then use cross-correlation analysis to identify those neurons that have correlated discharges. When electrodes are placed in brain regions that are anatomically connected, cross-correlation analysis permits inferences about whether specific pairs of neurons are interconnected. Furthermore, if two neurons are synaptically connected or are located in serially connected brain regions, cross-correlation analysis provides a convenient method for measuring the functional strength of these connections during a variety of experimental manipulations.

In essence, a cross-correlogram (CCG) illustrates time-varying fluctuations in the incidence or probability of a target neuron discharge given that a reference neuron discharged at time zero. As shown in Figure 1, two neurons located in different neuronal circuits have coordination patterns that reflect their relative circuit locations and synaptic relationships. For serially connected neurons, the CCG is used to calculate the probability that a reference neuron discharge is associated with a target neuron discharge at a specific time delay. This probability, or efficacy, is calculated as follows: efficacy $= (CE/NR) \times 100$, where CE indicates the number of correlated events in the CCG peak and NR indicates the total number of reference discharges. Hence, the efficacy ratio indicates the proportion of presynaptic reference neuron discharges that are correlated with subsequent time-locked discharges in the postsynaptic target neuron. For neurons that are not serially connected, the correlation coefficient is used to indicate the amount of correlated activity. As shown in Figure 1, the formula for the correlation coefficient is similar to the efficacy ratio.

Feed-Forward Transmission

Simultaneous neuronal recordings in successive levels of the somatosensory system reveal substantial changes in neuronal responsiveness as somesthetic information is transmitted to higher levels of the neuraxis. In general, peripheral afferent nerves transmit information with a high degree of fidelity to their postsynaptic targets in the brain stem, but subsequent transmission to the thalamus and cortex is characterized by a decline in postsynaptic responsiveness.

When afferent nerve fibers and their postsynaptic responses in the nucleus cuneatus are recorded simultaneously, nearly 90% of the axonal discharges in the afferent fiber are associated with a time-locked discharge in the postsynaptic cuneate neuron. In fact, a single impulse in the afferent nerve fiber often produces a burst of two or three discharges in the cuneate neuron. This high degree of synaptic security is related to the morphologic relationship between the central terminals of the afferent nerve fiber and the postsynaptic dendrites of the cuneate neuron. Ultrastructural examination indicates that afferent fibers have massive terminals that are invaginated by the cuneate neuron dendrites, and these synaptic contacts are completely ensheathed by glia. Presumably, this glomerular-like structure enables high-fidelity transmission of somesthetic information.

Comparisons of neuronal responses recorded simultaneously in the nucleus cuneatus and ventrobasal thalamus indicate that the rate of stimulus-induced activity is much lower in the ventrobasal thalamus than in the nucleus cuneatus. Consistent with this decline in responsiveness, cross-correlation analysis indicates that efficacy rates for cuneothalamic interactions reach a maximum of 30%. Hence, cuneothalamic efficacy is much lower than the synaptic efficacy of the primary afferent nerve fibers (see earlier).

Simultaneous neuronal recordings in the ventrobasal thalamus and somatosensory cortex reveal that cortical responsiveness is lower than in the thalamus. Cross-correlation experiments indicate that the maximum thalamocortical efficacy rate is about 20%, but most pairs of thalamic and cortical neurons exhibit efficacy rates well below 10%.

Differences in the efficacy rates for cuneothalamic and thalamocortical interactions are also related to the synaptic morphology of these connections.

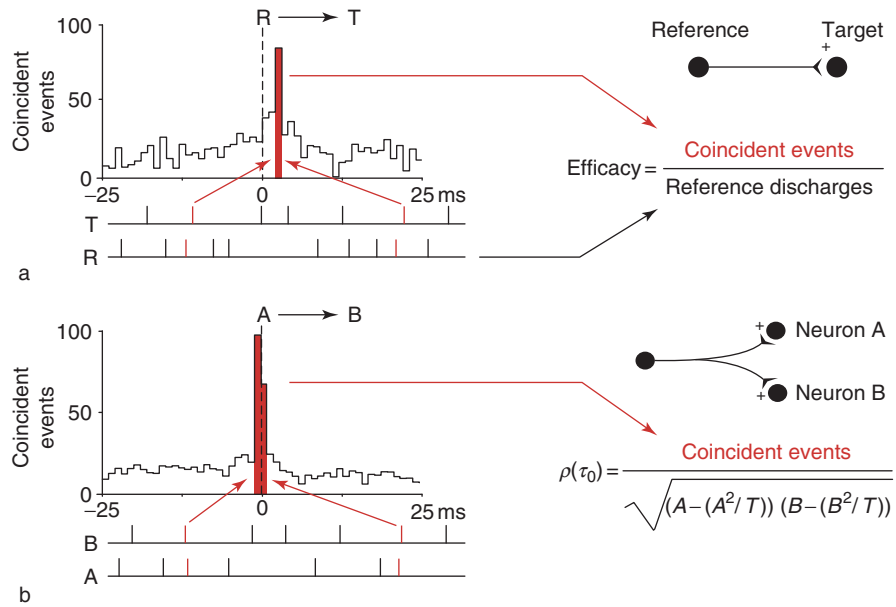

Figure 1 Cross-correlation analysis and the measurement of functional connection strength. (a) For neurons that are serially connected, the CCG shows time-varying fluctuations in the probability of a discharge in the target neuron (T) given that a reference neuron (R) discharged at time zero. The height of the CCG peak represents the number of instances in which a discharge in the reference neuron was time-locked to a subsequent discharge in the target neuron. The delay of the peak with respect to time zero reflects the time needed for action potentials to travel along the axon of the reference neuron. The efficacy ratio, which is obtained by dividing the number of coincident events in the CCG peak by the number of reference spikes, represents the probability that a reference spike is associated with a subsequent target neuron discharge at a specific time delay. (b) For neurons that are not serially connected with each other, the amount of correlated activity is determined by calculating a correlation coefficient, $\rho(\tau_0)$, in which A and B represent the number of discharges recorded from neurons *A* and *B* over a time period *T*.

Although not as large as the terminals formed by afferent nerve fibers, cuneothalamic terminals form glomerular-like synapses that are enveloped by a glial sheath. By contrast, thalamocortical synaptic contacts are much smaller, are not invaginated by the postsynaptic dendrites, and are not ensheathed by glial processes. Hence, thalamocortical connections are substantially weaker than the synaptic connections in the earlier stages of the somatosensory system.

Thalamocortical Feed-Forward Interactions

Feed-forward neuronal interactions between the ventrobasal thalamus and somatosensory cortex vary as a function of the laminar structure of the cerebral cortex. Cross-correlation analysis indicates that the incidence and strength of stimulus-induced thalamocortical interactions are highest when thalamic neurons are paired with cortical neurons in layer IV or in the deep part of layer IIIb. In fact, thalamocortical efficacy values above 10% are obtained only when the cortical neuron is in granular layer IV. This is consistent with the fact that neurons in the middle cortical layers receive the highest density of synaptic contacts from thalamocortical relay neurons. Furthermore, stimulus-induced thalamic discharges precede neuronal discharges in the middle cortical layers by 1–2 ms, which corresponds to the time needed for conducting action potentials from the thalamus to the cortex. This tight temporal coordination undoubtedly reflects interactions between neurons that have strong, monosynaptic connections.

Thalamocortical efficacy rarely exceeds 5% for cortical neurons recorded outside the middle cortical layers. Nonetheless, a comparison of thalamic interactions with cortical neurons in the supra- or infragranular layers reveals functional distinctions between these layers. For example, the probability of detecting correlated thalamocortical interactions is twice as great when the target neuron is a supragranular neuron instead of an infragranular neuron. Furthermore, the proportion of time-locked thalamocortical interactions that transpire within a 3 ms interval is much higher for target neurons in the supragranular than in the infragranular layers. These findings suggest that information transmitted from the ventrobasal thalamus must be processed by many synaptic connections before it can influence neurons in the deep cortical layers.

Laminar differences in the strength and relative timing of thalamocortical interactions have prompted

the view that tactile information follows a multisynaptic serial route through somatosensory cortex that commences in granular layer IV and proceeds to the supragranular layers before arriving in the infragranular layers. While this view overlooks many details of cortical connectivity and other aspects of cortical processing, it is consistent with the predominant circuit connections within a column of somatosensory cortex. Thus, most axons arising from granular neurons in layer IV terminate in the supragranular layers, which project to infragranular layer V. In turn, layer V pyramidal neurons send collateral projections to the neurons in infragranular layer VI. Finally, cortical output projections to the thalamus, neostriatum, pons, and spinal cord originate almost exclusively from infragranular layers V and VI.

Somatotopic Influences

The incidence and strength of stimulus-induced neuronal interactions in the somatosensory system are related to the somatotopic properties and other characteristics of the recorded neurons. Cross-correlation analysis indicates that neighboring cortical neurons with similar receptive fields and other response properties show high rates of correlated activity. As the amount of receptive field overlap declines, however, the amount of correlated activity that is produced by a punctate stimulus also declines.

Similar findings have been observed for pairs of neurons that reside in successive levels of the somatosensory system. Thus, in response to a simple punctate stimulus, cuneothalamic or thalamocortical interactions are strongest for those neurons that have identical or highly overlapping receptive fields. Neurons that represent nonoverlapping receptive fields do not exhibit correlated responses to a punctate stimulus. These results are consistent with the precise topographic organization of the ascending somatosensory projections.

Dynamic Interactions

The coordination of stimulus-induced neuronal activity varies substantially with stimulus intensity, location, and other factors that affect the spatiotemporal distribution of neural responsiveness. For example, if two neurons with similar receptive field properties are recorded simultaneously in the thalamus and cortex, their correlated activities can be altered merely by changing the peripheral stimulation site. Even though the synaptic contacts made by the thalamic neuron onto the cortical neuron are constant within a recording session, small shifts in stimulus position effectively activate slightly different populations of thalamic and cortical neurons. Because the two neurons are embedded in a larger network of neuronal connections, any shift in the network distribution of activity can produce a concomitant shift in the correlated activities of the two neurons. Hence, the dynamic operations of a circuit, including the effectiveness of communication between two connected neurons, may fluctuate instantaneously as the distribution of neural activity is altered by changes in the parameters of sensory stimulation.

Functional connection strength in the somatosensory system also varies with changes in the temporal structure of neuronal activity. Simultaneous neuronal recordings in VPM and somatosensory barrel cortex have shown that thalamocortical efficacy depends on the time interval between successive thalamic impulses. Thalamic discharges that occur after a quiet period of several hundred milliseconds are much more effective in evoking a cortical response than are thalamic discharges that occur in the midst of a train of thalamic discharges. Although the exact biophysical mechanism for this effect remains uncertain, activity-dependent depression of the thalamocortical synapse is the likely explanation for this dynamic shift in connection strength.

Neuronal Synchronization

Administration of a punctate stimulus to the skin activates intermingled populations of different types of peripheral receptors and their afferent projections to the central nervous system. Even if a single afferent fiber could be activated by a sensory stimulus, afferent fibers have terminal arborizations that form multiple synaptic contacts on neighboring postsynaptic neurons. This pattern of divergence is repeated at successive levels of the somatosensory system. Hence, the rich innervation of the skin and the divergence of successive ascending projections ensure that populations of neighboring neurons are activated simultaneously within each level of the somatosensory system.

Cross-correlation analysis of neurons located in close proximity consistently reveals the presence of synchronized activity during somatosensory stimulation. Thus, when neighboring neurons are recorded by two separate electrodes during cutaneous stimulation, the CCG exhibits a peak of coincident discharges at time zero (see Figure 1(b)). Such experiments have demonstrated stimulus-induced synchronization in all levels of the somatosensory system, including the brain stem, thalamus, and in primary (SI) and secondary (SII) somatosensory cortex.

The widespread incidence of neuronal synchronization in different brain regions during sensory, motor, and cognitive processing has prompted the

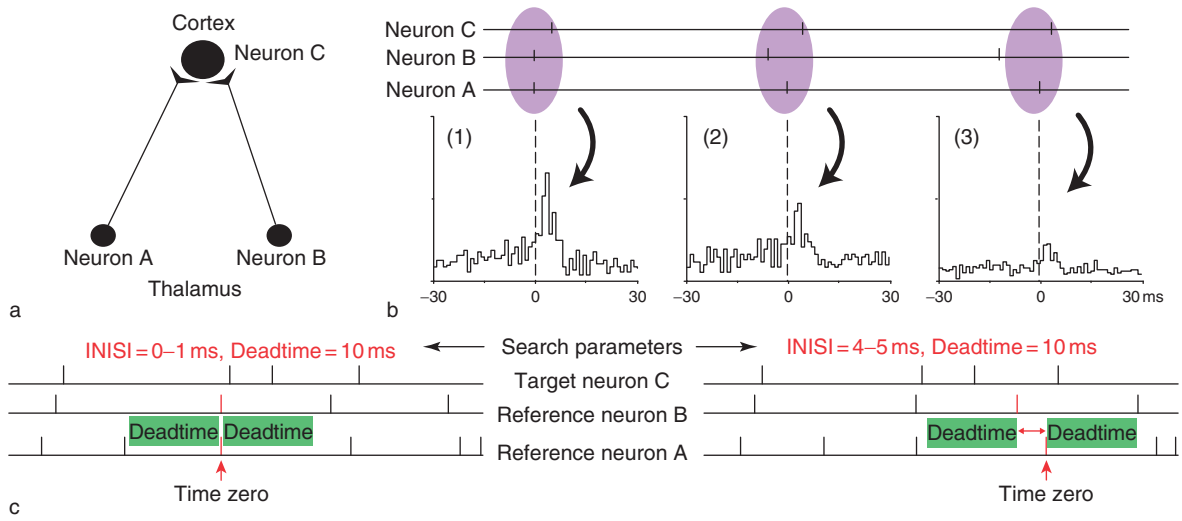

Figure 2 Thalamocortical cooperativity in the somatosensory system. (a) Simple neuron circuit in which a pair of thalamocortical relay neurons converge on a target neuron in somatosensory cortex. (b) The impact of thalamic synchronization on cortical responsiveness is determined by comparing the efficacy values from a series of CCGs that are constructed on specific types of reference events occurring in the spike trains of the two thalamic neurons. For example, individual CCGs are constructed (1) for synchronous thalamic discharges, (2) for short interneuronal interspike intervals (INISIs), or (3) for longer INISIs. (c) Specific reference events are identified by inspecting the thalamic spike trains and identifying INISIs in which a pair of thalamic discharges in the two neurons are isolated from other discharges by a deadtime period.

view that neuronal synchronization facilitates neurotransmission in many neural systems. According to this view, the probability of a postsynaptic neuronal discharge depends on the relative timing of presynaptic inputs and is highest when multiple presynaptic neurons discharge synchronously.

Thalamocortical Cooperativity

The impact of thalamic synchronization on thalamocortical transmission has been analyzed by using conditional cross-correlation analysis to measure cortical responsiveness as a function of the relative timing of thalamic discharges. In this procedure, the probability of a cortical neuron discharge following synchronous thalamic discharges is compared to the probability of a cortical response when each thalamic neuron discharges asynchronously (i.e., alone). If thalamic neurons cooperate with each other to activate a cortical neuron, the efficacy rate of their synchronous thalamic discharges should exceed the efficacy rate that is calculated from the joint probability of their asynchronous thalamic events. For example, if each thalamic neuron has an efficacy rate of only 10% when it discharges by itself, then each thalamic neuron also has a failure rate of 90%. Consequently, when both thalamic neurons discharge simultaneously, the predicted joint failure rate is 81% (the product of $90\% \times 90\%$). Hence, the predicted efficacy rate when both thalamic neurons discharge simultaneously is 19% ($100\% - 81\%$). If the observed efficacy exceeds 19%, then synchronization of these thalamic neurons must represent 'cooperative' rather than independent events.

Conditional cross-correlation analysis of thalamocortical interactions requires that two thalamic neurons with similar somatotopic properties are recorded simultaneously with a cortical neuron that also has similar response properties. As shown by Figure 2, a computer program searches the two thalamic spike trains and sorts those instances in which discharges of the two thalamic neurons are separated by a specific time interval. A set of conditional CCGs is then constructed for each type of reference event. Thus, the first conditional CCG is constructed based only on instances in which both thalamic neurons discharged at exactly the same time. A second CCG is then constructed based on reference events consisting of a discharge in one thalamic neuron followed by a discharge 1 ms later in the other thalamic neuron. A third CCG is constructed based on reference events consisting of a discharge in one thalamic neuron followed by a discharge in the other thalamic neuron 2 ms later. This process is repeated until a set of CCGs has been constructed for the range of discharge intervals across the two thalamic neurons. In each conditional CCG, the reference event at time zero represents the time of the second discharge of the two thalamic neurons.

The computer program also requires the presence of a deadtime period before and after each pair of thalamic neuron discharges. Hence, each reference event for a conditional CCG must represent an instance in which the discharges in the two thalamic neurons occur in isolation of other thalamic discharges. As shown in Figure 2, the deadtime period is at least as long as the interval between the two neuronal discharges. Without a deadtime period, the cortical neuron might discharge because it received a burst of homosynaptic inputs from the same thalamic neuron. Hence, a deadtime period removes cortical responses due to homosynaptic integration of the recorded neurons and ensures that the conditional CCGs are based on cortical responses to their heterosynaptic inputs.

Conditional cross-correlation analysis indicates that synchronized pairs of thalamic neurons have an efficacy rate that is more than twice the efficacy rate that is measured when the thalamic neurons discharge asynchronously (i.e., separated by 10 ms or more). Cortical responsiveness declines rapidly when thalamic discharges are separated by more than 3 ms and reaches an asymptotic level at intervals larger than 10 ms. This suggests that synchronous thalamic discharges represent cooperative events that facilitate the transmission of sensory information to the cerebral cortex.

Cortical Coordination

The presence of stimulus-induced neuronal synchronization in the cerebral cortex has prompted substantial debate and speculation, but its function remains unknown. In contrast to the impact of thalamic synchronization on cortical responsiveness, no studies have shown that cortical synchronization increases the responsiveness of target neurons in other cortical or subcortical regions.

Local populations of neurons in somatosensory cortex exhibit synchronized responses that vary as a function of cortical layer, neuronal proximity, and changes in the pattern of tactile stimulation. In rodent barrel cortex, for example, stimulus-induced synchronization is much more prevalent among infragranular neurons than in the other cortical layers. This finding is consistent with the infragranular location of intrinsically bursting neurons, which have extensive horizontal connections that may synchronize cortical outputs. Nonetheless, many results indicate that neurons in all layers of somatosensory cortex show increases in synchronization during certain experimental paradigms.

Tactile stimulation evokes neuronal synchronization among neighboring neurons in both somatosensory cortical regions SI and SII, but some distinctions between these regions have been reported. During cutaneous stimulation of a focal site, neuronal synchronization is more extensive in SII than in SI, possibly because SII neurons receive more thalamocortical convergence and have larger receptive fields. Neuronal synchronization in SII is also modulated by shifts in selective attention. In monkeys that alternately perform tactile and visual tasks, SII neurons are weakly synchronized during the visual task, but are strongly synchronized during the execution of a tactile discrimination.

Corticocortical Synchronization

Simultaneous neuronal recordings in separate cortical regions demonstrate that tactile stimulation can synchronize neurons that are widely separated from each other. In cats, for example, SI and SII neurons with overlapping receptive fields become synchronized during focal stimulation. This is consistent with the fact that the ventrobasal thalamus sends parallel, somatotopically organized projections to both cortical regions. Presumably, the dense corticocortical interconnections between SI and SII facilitate and reinforce the temporal precision of the synchronization that is observed across these cortical regions.

Direct corticocortical connections are not necessary, however, to synchronize separate cortical areas during tactile stimulation. In monkeys, for example, neuronal populations in Brodmann areas 3b (SI) and 4 (primary motor cortex) become synchronized during manipulative behaviors that require skill and attention, but not during simple repetitive movements. Although these cortical regions are not directly connected, it is conceivable that they become synchronized by subcortical inputs or by multisynaptic corticocortical pathways. While the function of neuronal synchronization across widespread parts of sensorimotor cortex is not clear, the acquisition of skilled movements that rely on somesthetic guidance may require coordination of multiple cortical regions that represent specific tactile sensations related to the control of specific motor actions.

Corticocortical and Corticofugal Interactions

Corticocortical and corticofugal projections play a major role in processing somatosensory information. Somatosensory cortex provides a rich source of projections to motor cortex as well as to other parts of the motor system, including the neostriatum, the pons, and the superior colliculus. The importance of these projections is related to the close interactions that transpire between the somatosensory and motor

systems during motor learning, active touch, and other somesthesis-guided behaviors. In addition, somatosensory cortex provides feedback projections to each level of the somatosensory system, including the spinal cord, brain stem, and thalamus. The ability to make fine tactile discriminations or to ignore somesthetic information in certain situations is facilitated by cortical feedback projections that modulate the feed-forward transmission of tactile information at each level of the somatosensory system.

The influence of the somatosensory cortex on the thalamus has been demonstrated by many physiological studies. Reversible inactivation of somatosensory cortex, for example, produces both suppressive and augmenting responses in the thalamus. Such opposite effects are undoubtedly due to the fact that corticothalamic projections can produce direct excitation or indirect inhibition via the reticular nucleus.

Simultaneous recordings of neuronal activity in SI cortex and the ventrobasal thalamus during different behavioral states suggest that corticothalamic projections have a modulatory influence on the feed-forward transmission of somesthetic information. For example, recordings in SI and VPM of awake rats reveal neuronal oscillations in the range of 7–12 Hz that are initiated in SI and then appear in the thalamus several milliseconds later. This top-down progression of neuronal coordination almost always precedes rhythmic whisker movements at frequencies of 7–12 Hz. Furthermore, SI and VPM responses to repeated electrical stimulation of the infraorbital nerve, which innervates the whiskers, are not constant but vary in magnitude according to whether the animal is actively whisking or is in a state of quiet rest. Although these results are subject to interpretation and may rely on several corticofugal mechanisms, such findings indicate that neuronal responses in the somatosensory system represent the outcome of dynamic circuit operations that entail more than the simple feed-forward transmission of sensory information along the ascending somatosensory pathways.

See also: Barrel Cortex Circuits; Corticothalamic Connections: Ultrastructure; Corticothalamic Connections: Structure and Function; Sensorimotor Integration: Barrels, Vibrissae and Topographic Representations; Somatosensory Cortex: Functional Architecture; Somatosensory Receptive Fields; Somatosensory Pathways (Ascending): Functional Architecture; Somatosensory Cortex.

Further Reading

Aertsen AMHJ, Gerstein GL, Habib MK, et al. (1989) Dynamics of neuronal firing correlation: Modulation of 'effective connectivity.' *Journal of Neurophysiology* 61: 900–917.

Alloway KD, Wallace MB, and Johnson MJ (1994) Cross-correlation analysis of cuneothalamic interactions in the rat somatosensory system: Influence of receptive field topography and comparisons with thalamocortical interactions. *Journal of Neurophysiology* 72: 1949–1972.

Bruno RM and Simons DJ (2002) Feedforward mechanisms of excitatory and inhibitory cortical receptive fields. *Journal of Neuroscience* 22: 10966–10975.

Johnson MJ and Alloway KD (1996) Cross-correlation analysis reveals laminar differences in thalamocortical interactions in the somatosensory system. *Journal of Neurophysiology* 75: 1444–1457.

Metherate R and Dykes RW (1985) Simultaneous recordings from pairs of cat somatosensory cortical neurons with overlapping peripheral receptive fields. *Brain Research* 341: 119–121.

Murthy VN and Fetz EE (1996) Synchronization of neurons during local field potential oscillations in sensorimotor cortex of awake monkeys. *Journal of Neurophysiology* 76: 3968–3982.

Nicolelis MAL, Fanselow E, and Henriquez C (2002) A critique of the pure feedforward model of touch. In: Nelson RJ (ed.) *The Somatosensory System*, pp. 299–334. New York: CRC Press.

Roy SA and Alloway KD (2001) Coincidence detection or temporal integration? What the neurons in somatosensory cortex are doing. *Journal of Neuroscience* 21: 2462–2473.

Steinmetz PN, Roy A, Fitzgerald PJ, et al. (2000) Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature* 404: 187–190.

Swadlow HA and Gusev AG (2001) The impact of 'bursting' thalamic impulses at a neocortical synapse. *Nature Neuroscience* 4: 402–408.

Vickery RM, Gynther BD, and Rowe MJ (1994) Synaptic transmission between single slowly adapting type I fibres and their cuneate target neurones in cat. *Journal of Physiology* 474: 379–392.

Zachariah MK, Coleman GT, Mahns DA, et al. (2001) Transmission security for single, hair follicle-related tactile afferent fibers and their target cuneate neurons in cat. *Journal of Neurophysiology* 86: 900–911.

Neural Coding in Primary Motor Cortex

S H Scott, Queen's University, Kingston, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

What neural code is conveyed in the discharge of neurons in motor cortex? While it is unlikely that neurons code explicitly some movement parameter, it remains valuable (if not essential) to correlate neural activity to some aspect of sensorimotor performance. The neural code expressed in a given brain region is an important issue since it is closely linked to its function.

This article focuses on neural coding related to primary motor cortex (M1), a region intimately involved in volitional motor function and a region that has received the greatest attention for addressing neural coding in the motor system. The neural code used by M1 is important not only for understanding neural processing in this brain region but also for understanding processing in other regions of the central nervous system. For example, if M1 activity reflects the goals of a motor task, it suggests that the details of motor execution must be performed anatomically downstream from this structure, notably in the spinal cord. Alternatively, if M1 activity codes much of the details about muscle activity patterns, then far less neural processing is required at the spinal cord.

This article begins with two background sections. First, it describes the concept of sensorimotor transformations, a theoretical framework that is implicitly or explicitly assumed by many studies in this field. This is followed by a description of the basic physics of limb movement in order to clarify the types of information that may be coded by the brain.

There are many behaviors and animal models that have been used to explore neural processing in M1. This article focuses predominantly on neural processing in nonhuman primates during reaching movements since these are the most common task and species used to address this issue.

Sensorimotor Transformations and Coordinate Frames

The most common framework for addressing the neural basis of reaching has been based on the problem of sensorimotor transformations. For a task such as reaching to a target, it is supposed that the brain converts visual information on target location into patterns of muscle activity through a series of intermediary coordinate frames. One putative series of transformations is shown in Figure 1, in which spatial target location is sequentially converted into hand motion, joint motion, joint muscle torques, and finally into muscle activation patterns.

These types of variables can be divided into several categories. For example, above the dotted line are kinematic variables that describe motion of the hand and joints, or the spatial goal. In contrast, below the line are kinetic variables that describe the physical forces required to generate movement and muscle activity. The dashed line denotes another common separation of variables into extrinsic variables, related to the global goal of the task (spatial goal or hand motion), and intrinsic variables, related to the details of the motor periphery (joint motion, muscular torques, or muscle activity).

One compelling argument to support the hypothesis that the brain computes intermediate representations to control motor action is based on the observation that humans (and nonhuman primates) make relatively straight hand trajectories with bell-shaped velocity profiles. Novel mechanical loads applied to the limb result in deviations in hand trajectory, but with practice, individuals once again generate relatively straight hand trajectories. This consistency of hand trajectories across different spatial directions and mechanical conditions certainly appears to support the idea that hand motion is computed and controlled by the brain. One could imagine that the discharge patterns of neurons somewhere in the motor system should reflect hand motion. Other representations, such as joint motion, could facilitate the integration of proprioceptive information from muscle afferents into motor commands.

Kinesiology of Reaching Movements

In order to identify what information may be represented in the discharge pattern of neurons in M1, it is important first to understand the physics of limb motion. Figure 2 illustrates how variables change during a reaching movement performed in the horizontal plane where movement is restricted to flexion and extension motions at the shoulder and elbow joints. Parameters related to the spatial goal, hand motion, joint motion, joint muscular forces, and muscle electromyographic (EMG) activity create a rich mix of time-varying signals. There are two basic patterns: tonic signals that shift monotonically during movement and phasic signals that rapidly change with

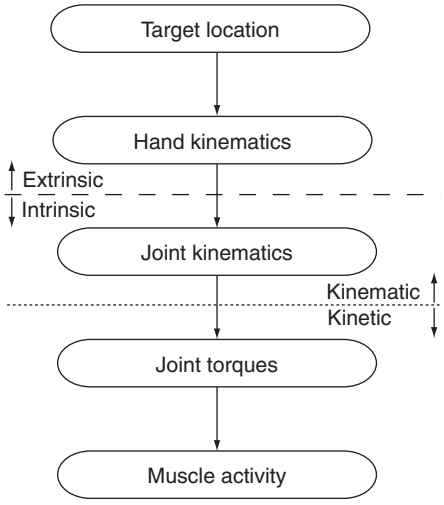

Figure 1 Putative coordinate transformations performed by the brain to convert information on the spatial location of a target into patterns of muscle activity. The various coordinate frames can be divided in several different ways. The dashed line denotes a separation between extrinsic representations related to the global goal versus intrinsic representations related to the details of the motor periphery. The dotted line denotes a separation between kinematic and kinetic representations.

one or more inflections during movement. In some cases, these patterns, such as hand and joint velocities, look quite similar. Also note that joint acceleration and muscular torque are not identical, due to intersegmental dynamics, because muscular torque generated at a joint can cause motion at other joints.

Across movement directions, the pattern of joint motion, muscular torque, and EMG activity can change dramatically (Figure 3). Movements away and toward the body require the largest motion at the shoulder and elbow joints, whereas movements to the right and left require little motion. This is simply a property of limb geometry, or how the motion of each joint contributes to hand motion for a given movement direction. Variations in muscular torque do not parallel joint motion. Shoulder muscular torque is largest for movements to the left and away (flexor), and to the right and toward the monkey (extensor). Elbow muscular torque is smaller in magnitude compared with that at the shoulder and is maximal for movements in the opposite quadrants. The directions of maximal activity of muscles do not match muscular torque at the spanned joint since muscle activity is influenced by a number of other factors, including muscle length, velocity, and the ways other muscles contribute to a given behavior.

The description of limb movements in Figures 2 and 3 are for planar movements involving only two mechanical degrees-of-freedom (DOF), one at the shoulder and one at the elbow. Free whole-arm movements involve at least two more DOF at the shoulder, abduction–adduction and internal–external rotation. Therefore, the patterns of joint motions, muscular torques, and EMG activities are even richer in this context.

Correlations Based on Single-Cell Activity

One common approach for identifying what information is coded in M1 is to correlate neural activity to variables of movement. Figure 4 illustrates the patterns of activity of four representative neurons observed during reaching to the same target displayed in Figure 2. Cells show tonic and phasic patterns and combinations of these two basic patterns. A given neuron could code any of a number of different parameters or a combination of parameters. How does one choose among all the possible parameters?

Perhaps the most common variable correlated to neural discharge in M1 is related to hand motion. This approach was adopted in the early 1980s in the first studies of whole-arm reaching performed by Apostolos Georgopoulos and colleagues. Only hand motion was recorded because there wasn't any technology readily available to record joint motion or estimate joint torques. Figure 5 illustrates how neural activity is broadly tuned to the direction of hand motion. Cell activity is maximal for movements to the lower right, its preferred direction of movement (PD), and the magnitude of activity is reduced gradually for movements at greater angles from this PD. This broad tuning of activity is a very common characteristic of neurons in M1 and elsewhere in the motor system. Therefore, implicitly this and other studies using this approach illustrate that neural activity in M1 correlates with the direction of hand motion and with other hand-based variables, such as movement speed and distance.

Correlations Based on Population Codes

Another approach for examining neural coding in M1 (and elsewhere) is to consider what parameters are specified at the population level. This seems reasonable since motor function certainly reflects the combined action of large populations of neurons. The basic idea is that a neural code is highly distributed, so a neural code may be observed across the cell population.

While Humphrey in the 1970s first demonstrated that the combined activity of neurons could predict the pattern of muscle EMG activity during a motor task, the most influential study examining population coding in M1 was again provided by the reaching studies by Georgopoulos and colleagues. This approach, developed initially by Ken Johnson, is

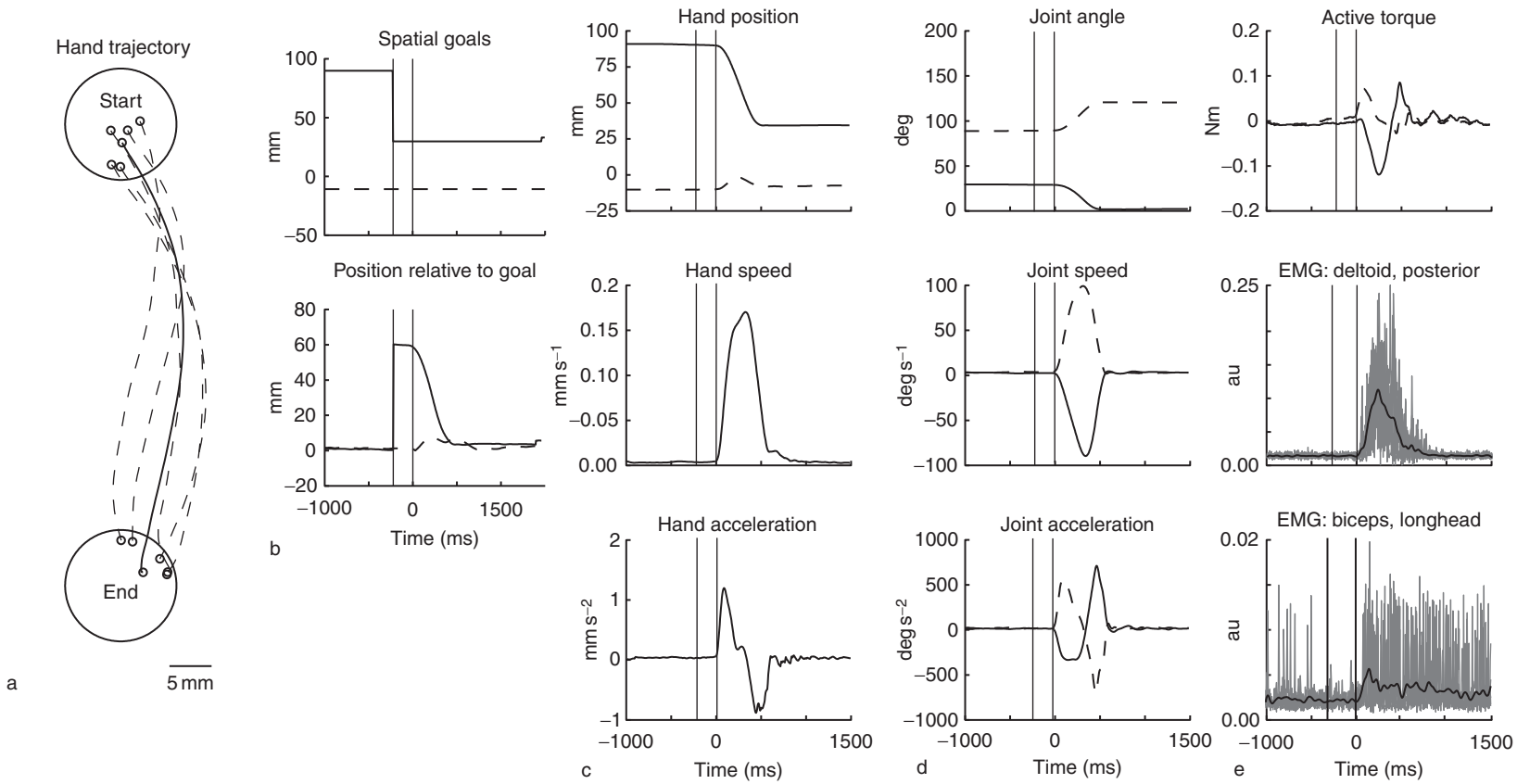

Figure 2 Kinesiology of a reaching movement. (a) Hand trajectory for five repeat reaches (dashed lines) and the mean trajectory (solid). Large circles denote the size of the central and peripheral targets. Small circles denote the identified start and finish of hand motion based on 5% of peak hand velocity during each trail. (b) Variables related to the spatial goal including its absolute position (top panel), which shifts from the central to the peripheral target when the latter is illuminated. The position of the hand relative to the goal (bottom panel) abruptly jumps when the peripheral light is illuminated but decreases during hand motion. Dashed and solid line denote *x*- and *y*-parameters, respectively. (c, d) Variation in the position, speed, and acceleration of the hand (c) and joints (d) during reaching. Solid and dashed lines in (d) denote the shoulder and elbow, respectively. (e) Active muscular torque and example muscle activity patterns. EMG, electromyographic activity. EMG scale is in arbitrary units (au).

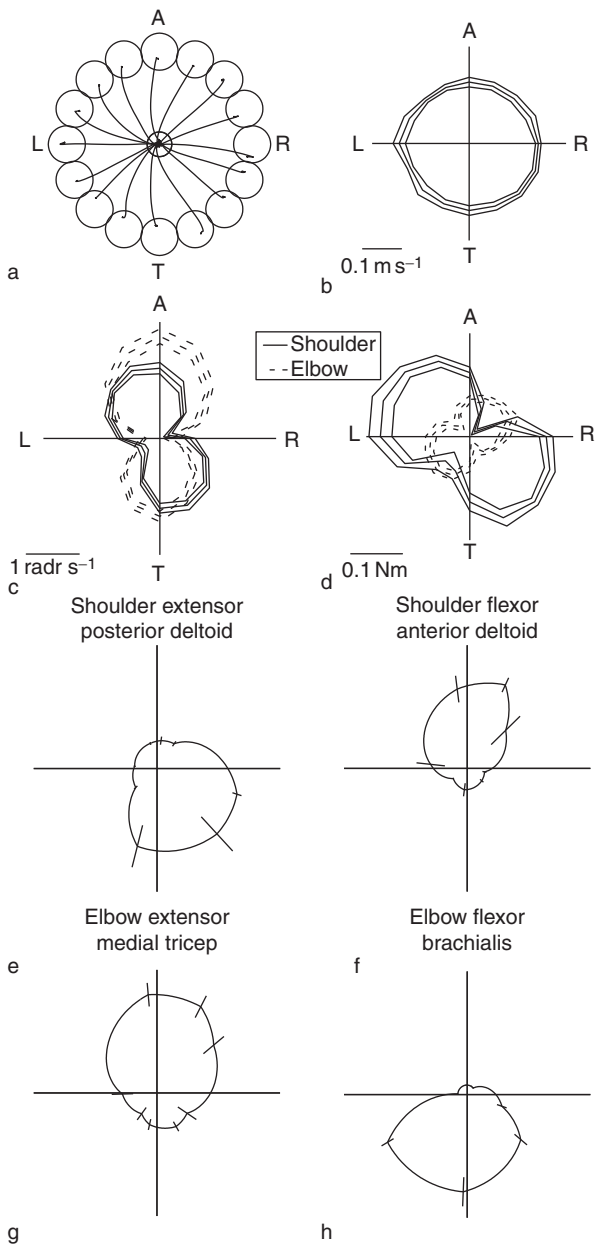

Figure 3 Kinesiology of reaching in different spatial directions. (a) Mean hand trajectories for reaching movements from a central target to 16 peripheral targets. A, R, T, and L denote directions of movement away, right, toward, and left relative to the monkey's body. (b) Polar plot of peak hand speed. Angle defines the direction of hand motion in Cartesian space, and distance from origin defines the magnitude of the peak hand speed. Thick and thin lines denote mean and standard deviation across repeat trials. (c, d) Polar plots of peak joint speed and joint muscular torque. (e)–(h) Electromyographic activity of four representative muscles. Thick lines denote standard deviation for each movement direction. (a–d) Adapted from Graham KM, Moore KD, Cabel DW et al. (2003) Kinematics and kinetics of multijoint reaching in non-human primates. *Journal of Neurophysiology* 89: 2667–2677, used with permission from The American Physiological Society.

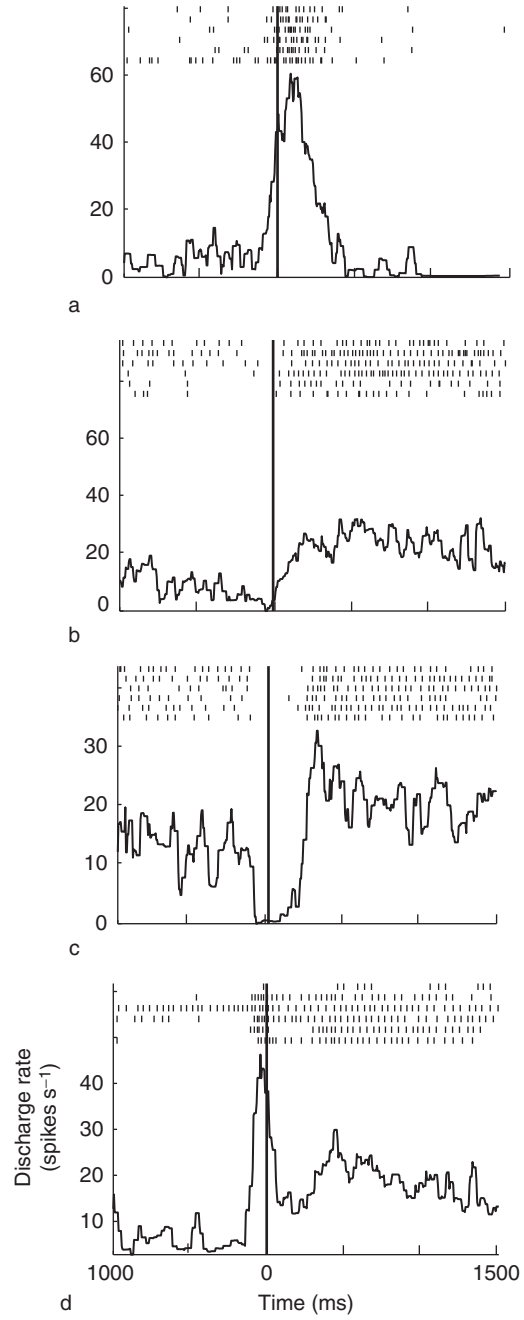

Figure 4 Different patterns of cell activity in primary motor cortex during reaching. (a)–(d) Each panel shows a raster of the time of action potentials for six repeat trials, and the solid line denotes the instantaneous firing rate of the neuron. The vertical line denotes the onset of movement.

called the population vector hypothesis, or method (PVM). Each neuron provides a 'vote' along a vector aligned with its PD. For a given movement, the magnitude of a cell's vote is scaled proportional to its firing rate for that movement direction. The population vector is computed by summing vectorially all the individual cell vectors. The key observation is that the population vector tends to point in the direction of hand motion (Figure 6). Several studies have used the PVM or more complex mathematical algorithms to predict movement of the hand during spiral movements, circles, or tracking movements throughout the workspace. Based on the number of

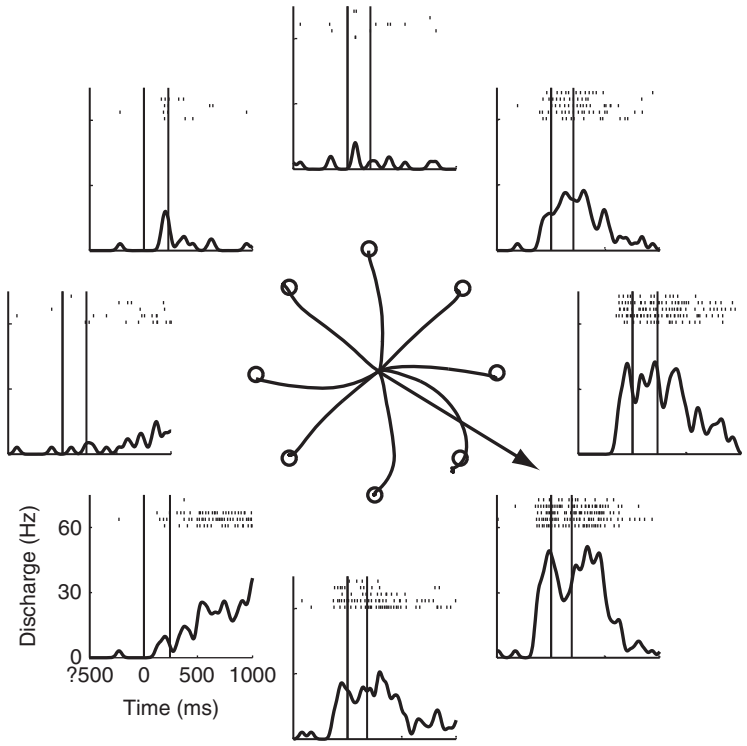

Figure 5 Cell activity in primary motor cortex during reaching. Central panel illustrates the mean hand trajectory for five repeat movements to each spatial target. Corresponding cell activity is adjacent to each peripheral target. Each raster plot shows the discharge rate for five repeat trials; each small vertical tick denotes the time of an action potential. Data are aligned to movement onset. Below each raster is the instantaneous firing rate of the neuron. Vertical lines denote movement onset (time = 0 ms) and mean peak hand velocity. The overlying arrow on the central panel denotes the neurons' preferred direction of hand motion.

studies published in the literature, this hand-based approach is certainly the most popular one for interpreting motor cortical activity. The use of correlations between hand motion and M1 activity is also becoming popular for applied research on neuroprosthetics, with the goal of helping individuals who cannot control their muscles directly.

What is the significance that neural activity in M1 is broadly tuned to the direction of hand movement and that population vectors or other mathematical algorithms can predict the direction of hand movement? It could mean that M1 activity is coding high-level extrinsic information related to the global goal of the task, moving the hand to a spatial target. However, several articles have raised the difficulty of interpreting neural coding based on these correlates of hand motion. In particular, when one examines the physics of limb motion, one finds that many variables are broadly tuned to the direction of hand motion, including joint motion, joint muscular torques, and EMG activity patterns (Figure 3). Further, mathematical algorithms can be used to predict virtually any parameter of movement from M1 activity. The basic problem is that understanding what is coded in M1 (and elsewhere in the motor system) cannot be addressed easily using only correlational techniques at the single-cell or population level. Simply recording only one type of information, such as hand motion, and ignoring all other variables do not resolve this problem.

While population vectors tend to point in the direction of movement when monkeys perform free arm movements, some systematic errors between the population vector and the movement direction have been observed for planar limb movements. For these planar movements, there is a substantial bias in the distribution of preferred directions such that most neurons have PDs that are oriented either away from and to the left, or toward and to the right (Figure 7). The result is that population vectors tend to be skewed in these same spatial directions and rarely predict the direction of hand motion. This nonuniform distribution of PDs is also paralleled by variations in joint power and in the distribution of PDs of the shoulder and elbow muscle activities. One could make the argument that this bias in the distribution of M1 neurons means that neurons code limb mechanics or muscle activities but can be countered by the argument stated above regarding the difficulty of interpreting hand-based codes. At best, these results suggest that the mechanics of the limb has some influence on the patterns of activity in M1 at the population level. They do not resolve what information is expressed at the single-cell level.

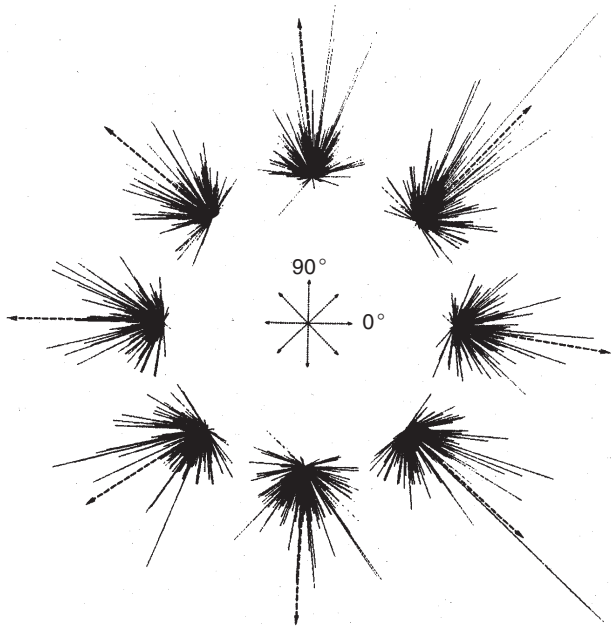

Figure 6 Population vectors (dashed line with arrow) for movements in different spatial directions. Each cluster includes population vector and the individual neuron vectors for a given direction of movement (0° is to the right, 90° away, etc.). Each black solid line in a cluster denotes the contribution of an individual neuron; the orientation of the line reflects the neuron's preferred direction, and the length of the line reflects the discharge rate of the neuron for that particular direction of movement. The population vector is the vector summation of all the individual neuron vectors. Reproduced from Georgopoulos AP, Kalaska JF, Caminiti R, and Massey JT (1983) Spatial coding of movement: A hypothesis concerning the coding of movement direction by motor cortical populations. *Experimental Brain Research Supplement* 7: 327–336, with kind permission of Springer Science and Business Media.

Dissociation of Parameters

Given the fact that movement parameters are highly correlated, the best approach to identify which parameters are coded in M1 activity is based on experiments designed to dissociate different types of variables. The key aspect of such studies is that they hold one or more variables constant while systematically manipulating another variable. If neural activity changes between the two conditions, then neural activity reflects, to some degree, the latter variable (or another variable with similar attributes).

This important point was recognized by Ed Evarts in the first study to record single-neuron activity in M1 in awake, behaving monkeys. Monkeys were trained to make wrist flexion and extension movements with loads that either assisted or resisted wrist motion. The key observation is that about half of M1 neurons modified their activity when mechanical loads were varied. However, half of M1 neurons were insensitive to loads, demonstrating that they reflected only kinematic aspects of the task.

A more recent study separated three features of movement (joint motion, target direction, and EMG activity) by training monkeys to make wrist movements involving flexion–extension and/or radial-ulnar deviation, with the forearm in three different orientations: supinated, neutral, and pronated. These different forearm orientations shifted the orientation of the hand so that wrist flexion required vertical motion in the supinated position but downward motion in the pronated position. Movement of the wrist was displayed on a video monitor. If neural activity specified the direction of hand motion, then neural activity would rotate 180° from supination to pronation. If neural activity specified the spatial direction of the target, then neural activity would not rotate between the different forearm orientations. Ultimately, the response of muscles was approximately halfway between these two extremes, rotating ~90°. The investigators found that some neurons rotated their activity similarly to muscles, whereas other neurons showed no rotation (Figure 8). No neurons rotated 180°. This suggests that there are at least two populations of neurons, one related to the spatial goal and/or visual feedback of hand motion and another population related to the motor action of the peripheral muscles.

Many other studies have performed dissociations between other classes of variables: intrinsic versus extrinsic variables, spatial goals versus limb motion, and kinematics versus kinetics. In each study, the same basic observation is usually found: some neurons changed their activity across conditions, and other neurons showed no change. Across all these studies, it appears that neural activity can reflect virtually every class of variable. One must conclude that there is no single, unified, coordinate frame to explain the activity of all neurons in M1. Rather, neural coding in M1 reflects many different classes of movement variables.

Context-Dependent Change in Neural Coding

Recent work suggests that there is an additional level of complexity in neural coding in M1. It appears that neural coding can change abruptly across different types of motor behaviors such as switching from maintaining a constant arm posture to reaching to another spatial location. Specifically, the response of neurons to constant loads applied to a monkey's shoulder, elbow, or both joints was observed prior to, during, and after reaching movements (Figure 9). Planar regression models were used to relate cell discharge at a given period (during posture or during movement) to shoulder and elbow torques. About half of recorded neurons in M1 modulated their discharge for these loads as the monkey maintained its hand at a central target

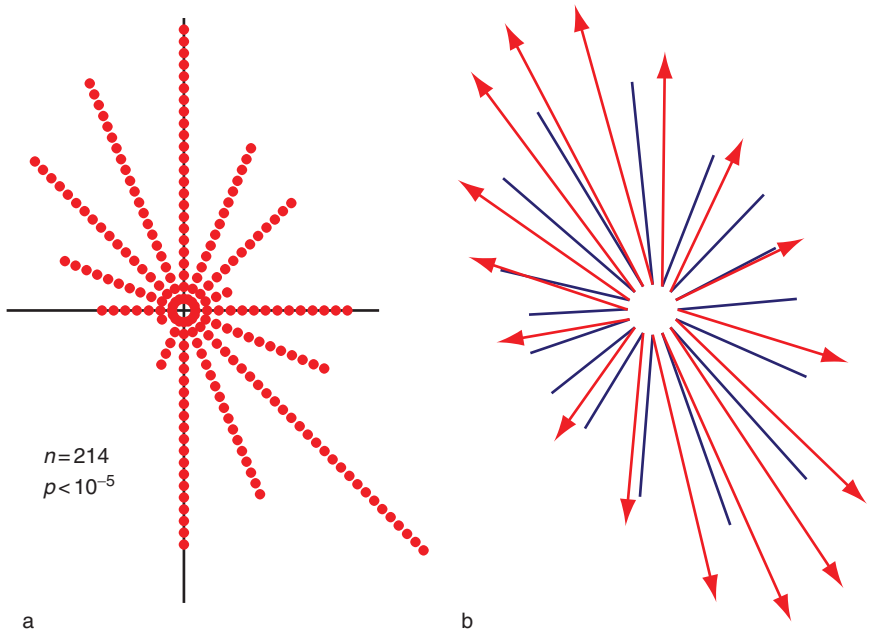

Figure 7 Neural responses during reaching with the limb in the horizontal plane. (a) Distribution of preferred directions of neurons recorded in M1. Each red dot denotes the preferred direction of movement for an individual neuron, and cells are grouped into 16 equal-size bins (bin size = 22.5°). The *p* value in the panel reflects the statistical level at which the population was found to be not uniform compared with a bimodal distribution. (b) Comparison between population vectors (red) and the direction of hand motion (blue). Population vectors did not point in 13 of the 16 directions of movement. Adapted from Scott SH, Gribble PL, Graham KM, and Cabel DW (2001) Dissociation between hand motion and population vectors from neural activity in motor cortex. *Nature* 413: 161–165.

(postural control). However, just before the monkey made a movement toward another spatial target, half of the neurons modulated by loads during posture lost their load-related activity ('posture only' cell, Figure 9), and half the neurons not influenced by load during posture instantly became modulated by load ('reach only' cell, Figure 9). In addition, neurons that responded to loads across posture and movement epochs often changed the magnitude of their load-related discharge. This change in load-related activity was not evident at the muscular level, where most shoulder and elbow muscles responded to loads during both posture and movement. This dramatic change in central processing of load-related activity suggests that the information represented or coded in M1 is not fixed but can change from one behavioral context to another.

Neural Coding in Premotor Cortex

Neural coding in dorsal and ventral premotor cortex appears to be far less complex with regard to the range of variables expressed in the discharge of individual neurons. Spatial goals and the direction of movement are well represented at the single cell level, with less influence by parameters related to the motor periphery, such as joint motion and forces. Also, far more preparatory activity can be observed when the animal is instructed where it will move to but must wait for a go signal before executing the movement. This suggests that neural processing in these brain regions is related to planning and higher-level aspects of the task rather than the details of motor execution.

Neural processing in these premotor regions also shows several complexities. Neural activity can be observed for movements of the contralateral limb and the ispilateral limb and even for movements generated by other individuals. Neural activity can also be influenced by factors such as where the individual is looking. It is interesting that the magnitude of this influence depends on whether gaze angle is controlled experimentally and maintained at a specific spatial location or allowed to move freely during the motor task: Gaze angle influences neural activity in the former context but minimally in the latter.

Beyond Neural Coding

It is worthwhile to reflect on what has been learned about neural coding in M1. Virtually, every variable displayed in Figure 2 correlates with neural activity in this brain region. Even experiments that specifically dissociate different classes of variables illustrate considerable heterogeneity in the type of information conveyed in the activity of M1 neurons.

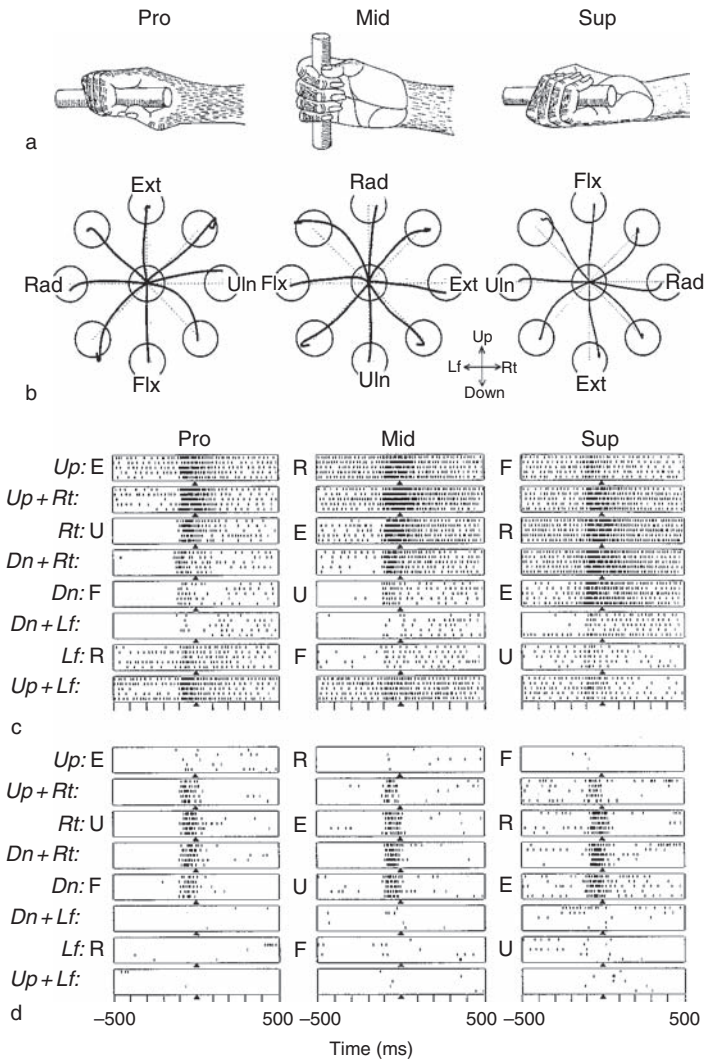

Figure 8 (a) Wrist movements made in different spatial directions using different forearm orientations: pronated (Pro), middle (Mid), and supinated (Sup). (b) Wrist movements were performed in eight different spatial directions. The shift in forearm orientation shifted the relationship between wrist motion and spatial direction. Upward movements required wrist extension when the forearm was pronated but required wrist flexion when the forearm was supinated. (c) Discharge pattern of a neuron in M1 for the three different arm orientations. Each row of rasters shows the cell's discharge for the same spatial direction, and data are aligned to movement onset. This neuron showed a shift in its preferred direction across the three forearm orientations similar to that observed for wrist muscles. (d) A neuron that showed minimal changes in its pattern of activity when movements were performed with different forearm orientations. Rad, radial deviation; Ext, extension; Uln, ulnar deviation; Flx, flexion; R, radial deviation; E, extension; U, ulnar deviation; F, flexion; Rt, right; Dn, down; Lf, left. Reprinted with permission from Kakei S, Hoffman DS, and Strick PL (1999) Muscle and movement representations in the primary motor cortex. *Science* 285: 2136–2139. Copyright (1999) AAAS.

This diversity in M1 activity is difficult to fit with the idea that the brain performs a series of sensorimotor transformations to convert spatial goals into muscle activities. The bias in neurophysiological recording techniques toward sampling larger neurons means that most recorded neurons are likely corticospinal neurons from layer V that project down to the spinal cord. Therefore, signals reflecting extrinsic features of movement are in parallel with signals related to the intrinsic features of movement. A further challenge to the notion of sensorimotor transformations is that the variable expressed by an individual neuron can abruptly switch as soon as one switches behavioral context.

It is not clear that continued focus on identifying which parameter is expressed in the discharge pattern of M1 neurons will provide substantive progress in understanding M1 function. What is needed perhaps is to refocus attention on the basic motor problem, sensorimotor control.

The 1970s saw considerable interest in interpreting M1 activity on the basis of feedback control. M1 is

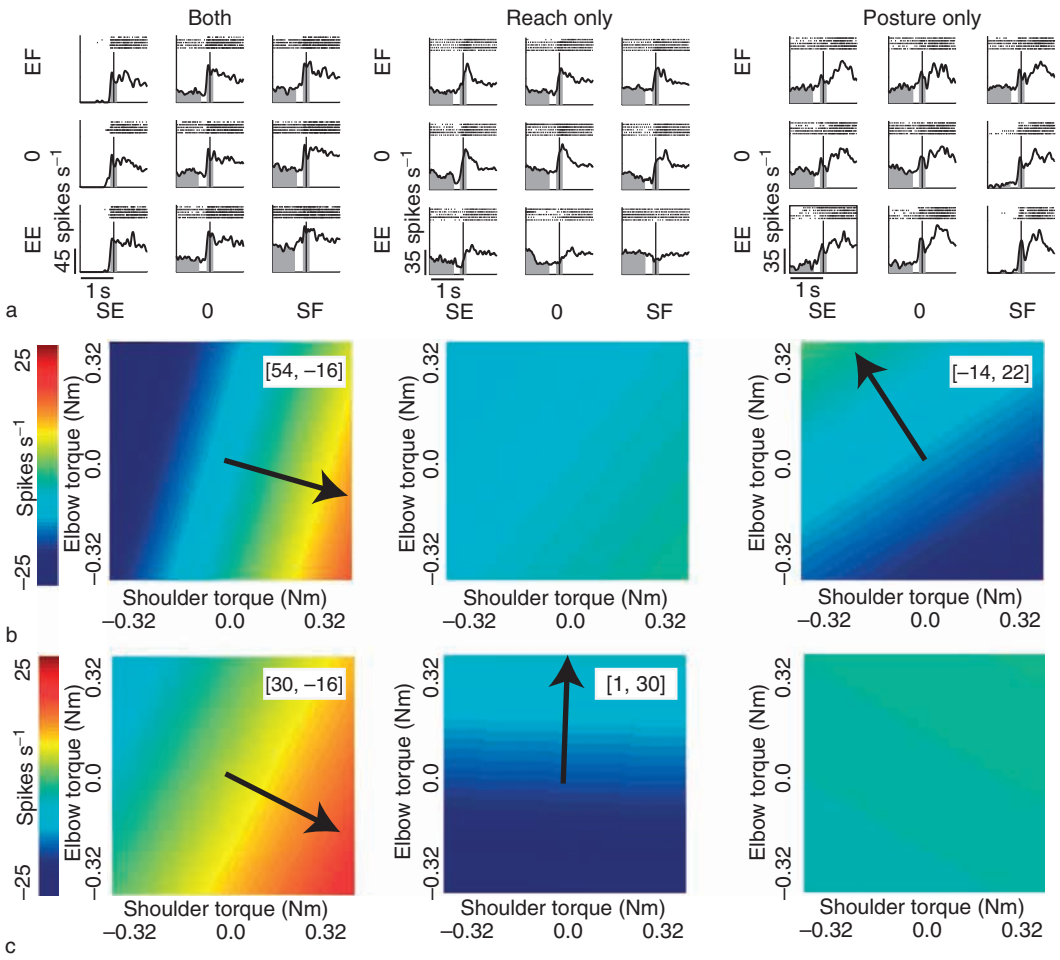

Figure 9 (a) Discharge pattern of three neurons when a reaching movement is made with and without constant-magnitude loads applied to the shoulder or elbow joints. (a)–(c) Vertical columns show sample neurons expressing load-related activity during posture and movement (both), only during reaching (reach only), or only during posture (posture only). (a) Spike rasters and mean instantaneous firing rate under the nine load conditions, organized in a joint–torque coordinate frame, shoulder flexor (SF), shoulder extensor (SE), elbow flexor (EF), elbow extensor (EE), and no joint–torque (0). Shaded regions denote posture and movement epochs used for plane fits displayed in subsequent panels. (b, c) Plane fits of change in neural activity versus change in joint torque during posture (b) and movement (c). Arrows and coefficients denote significant plane fit and regression values, respectively. Reproduced from Kurtzer I, Herter TM, and Scott SH (2005) Random change in cortical load representation suggests distinct control of posture and movement. *Nature Neuroscience* 8: 498–504.

known to receive rich sensory information related to limb motion, and several studies illustrated how M1 neurons responded rapidly to transient perturbations, providing a long-latency reflex response that could be modified on the basis of the goal of the task. However, several factors appeared to diminish the utility of this line of research. First, experiments examined single-joint movements, which lack much of the mechanical complexity of whole-limb movements. Feedback control was viewed as simple servocontrol with a fixed mapping between sensory input and motor output (single input–single output). The principles of control for a single joint do not simply scale for more-natural whole-limb tasks. Second, substantial time delays in sensory feedback from the periphery meant that simple servocontrol strategies were untenable.

There was a dramatic shift in experimental paradigms in the 1980s toward experiments that examined open-loop control processes. The introduction of the whole-arm reaching task shifted attention to correlates of hand motion and higher-level features of movement, and studies focused on neural activity prior to the initiation of movement, before feedback arrived in the brain. Conceptually, sensorimotor transformations and putative coordinate frames became and have remained the most popular approach for interpreting M1 activity.

However, several advances in our knowledge again make feedback control a viable approach for

interpreting the role of M1 in motor function. The fatal problem that sensory information from the periphery was transmitted too slowly to support feedback control has been curtailed by the recognition that the brain likely uses forward models to convert motor commands into estimates of limb motion. These estimates of limb motion can be computed with minimal delays and can be adjusted as sensory feedback from the limb and vision arrives in the central nervous system. The rapid conversion of motor commands into estimates of limb motion provides a simple explanation for efferent copy and why there are several internal feedback loops within the motor system.

More-advanced feedback controllers are also being considered for interpreting the neural basis of sensorimotor control, such as optimal feedback control (Figure 10). Instead of a simple one-to-one fixed mapping between sensory feedback and motor output, these controllers can connect all sensory feedback signals to all motor outputs (multiple input, multiple output). How this sensory feedback is connected to motor output depends on the goal of the task. These controllers effectively create smart reflexes so that errors that influence the goal of the task are corrected and errors that do not influence the goal are ignored.

Smooth bell-shaped velocity profiles and relatively straight hand paths have often been used as evidence that the brain directly controls and computes hand motion. An interesting feature of optimal feedback controllers is that such patterns of hand motion can be generated even though hand motion is not computed explicitly. Such controllers can use any set of state variables: hand motion is an emergent property, reflecting the interplay between noise/errors and feedback gains. These gains are set on the basis of the behavioral goal which in many motor tasks is related to the hand.

If one views the brain as behaving like an optimal feedback controller, the problem of neural coding in a brain region like M1 changes dramatically. Vision, proprioceptive signals from muscle, cutaneous receptors in the skin, and even auditory signals can all provide valuable feedback on motor performance. It should not be surprising to see neural activity in a region like M1 related to each of these types of signals. Spatial signals are not higher-level representations of movement as compared to signals related to joint motion, as assumed in the sensorimotor transformation framework. Rather, these signals are parallel when viewed from optimal feedback control since visual and proprioceptive signals both provide feedback to guide motor action.

Behavioral context also plays a crucial role in modifying neural processing in M1 by dictating how sensory information is used to guide action. Changes in the behavioral goal necessarily lead to changes in the use of sensory feedback. During posture, the appropriate motor response to a perturbation is to rapidly return to the same spatial location. This is an inappropriate response during a reaching movement, when one should correct and continue to move toward the new spatial location. Thus, it should not be surprising to see substantial changes in neural processing of variables, such as applied loads, between contexts, as shown in Figure 9. Variations in the experimental paradigm, such as controlling gaze angle during reaching or viewing the arm directly versus using a computer monitor which dissociates proprioceptive and visual alignment of the limb, should also modify neural processing.

In conclusion, the concept of neural coding has been influential for exploring the type of information conveyed in the discharge patterns of M1 neurons in awake, behaving animals. These studies have not found a common code to describe neural processing

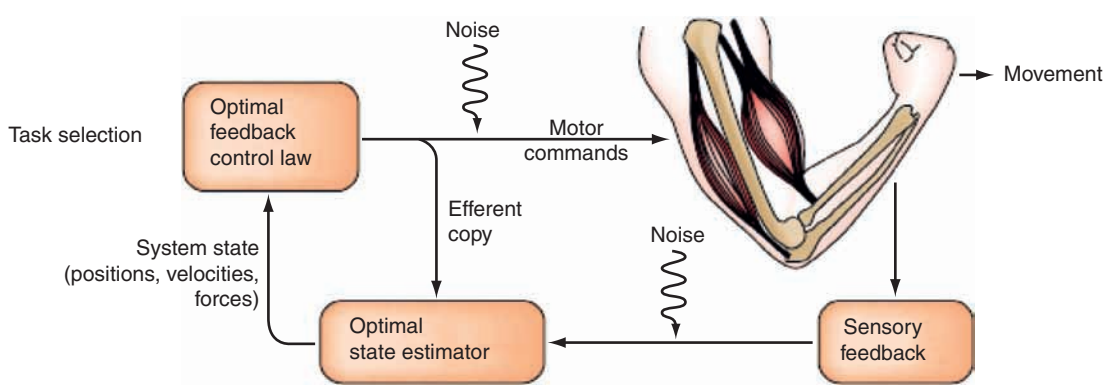

Figure 10 Basic principle of optimal feedback control. Feedback gains (control law) are optimized based on the global goal of the task. Feedback of motor performance is based on efferent copy and sensory feedback from the limb and vision. Such systems are particularly well suited for managing noise/errors inherent in the sensorimotor system. Reproduced from Scott SH (2004) Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience* 5: 532–546.

in this brain region. Rather, these studies have uncovered a highly heterogeneous mix of signals or codes. This seems perplexing within the framework of sensorimotor transformations but provides a foundation for developing highly flexible control signals as viewed from sophisticated control processes, such as optimal feedback control.

See also: Computational Approaches to Motor Control; Corticomotoneuronal System; Finger Movements: Control; Motor Timing; Motor Psychophysics; Premotor Cortex in Primates: Dorsal and Ventral; Primate Interneurons; Reaching and Grasping; Sensorimotor Control of Manipulation.

Further Reading

Cheney PD and Fetz EE (1980) Functional classes of primate corticomotoneuronal cells and their relation to active force. *Journal of Neurophysiology* 44: 773–791.

Cisek P, Crammond DJ, and Kalaska JF (2003) Neural activity in primary motor and dorsal premotor cortex in reaching tasks with the contralateral versus ipsilateral arm. *Journal of Neurophysiology* 89: 922–942.

Cisek P and Kalaska JF (2002) Modest gaze-related discharge modulation in monkey dorsal premotor cortex during a reaching task performed with free fixation. *Journal of Neurophysiology* 88: 1064–1072.

Desmedt JE (1978) *Cerebral Motor Control in Man: Long Loop Mechanisms*. Basel: Karger.

Evarts EV (1968) Relation of pyramidal tract activity to force exerted during voluntary movement. *Journal of Neurophysiology* 31: 14–27.

Fetz EE (1992) Are movement parameters recognizably coded in the activity of single neurons? *Behavioral and Brain Sciences* 15: 679–690.

Georgopoulos AP, Kalaska JF, Caminiti R, and Massey JT (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience* 2: 1527–1537.

Georgopoulos AP, Kalaska JF, Caminiti R, and Massey JT (1983) Spatial coding of movement: A hypothesis concerning the coding of movement direction by motor cortical populations. *Experimental Brain Research Supplement* 7: 327–336.

Graham KM, Moore KD, Cabel DW, et al. (2003) Kinematics and kinetics of multijoint reaching in non-human primates. *Journal of Neurophysiology* 89: 2667–2677.

Kakei S, Hoffman DS, and Strick PL (1999) Muscle and movement representations in the primary motor cortex. *Science* 285: 2136–2139.

Kakei S, Hoffman DS, and Strick PL (2001) Direction of action is represented in the ventral premotor cortex. *Nature Neuroscience* 4: 1020–1025.

Kalaska JF and Crammond DJ (1992) Cerebral cortical mechanisms of reaching movements. *Science* 255: 1517–1523.

Kurtzer I, Herter TM, and Scott SH (2005) Random change in cortical load representation suggests distinct control of posture and movement. *Nature Neuroscience* 8: 498–504.

Mussa-Ivaldi FA (1988) Do neurons in the motor cortex encode movement direction? An alternative hypothesis. *Neuroscience Letters* 91: 106–111.

Sanger TD (2003) Neural population codes. *Current Opinion in Neurobiology* 13: 238–249.

Scott SH (2004) Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience* 5: 532–546.

Scott SH, Gribble PL, Graham KM, and Cabel DW (2001) Dissociation between hand motion and population vectors from neural activity in motor cortex. *Nature* 413: 161–165.

Scott SH and Kalaska JF (1997) Reaching movements with similar hand paths but different arm orientations. I: Activity of individual cells in motor cortex. *Journal of Neurophysiology* 77: 826–852.

Scott SH, Sergio LE, and Kalaska JF (1997) Reaching movements with similar hand paths but different arm orientations. II: Activity of individual cells in dorsal premotor cortex and parietal area 5. *Journal of Neurophysiology* 78: 2413–2426.

Shadmehr R and Wise SP (2005) *The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge, MA: MIT Press.

Shen L and Alexander GE (1997) Preferential representation of instructed target location versus limb trajectory in dorsal premotor cortex. *Journal of Neurophysiology* 77: 1195–1212.

Todorov E (2000) Direct cortical control of muscle activation in voluntary arm movements: A model. *Nature Neuroscience* 3: 391–398.

Todorov E and Jordan MI (2002) Optimal feedback control as a theory of motor coordination. *Nature Neuroscience* 5: 1226–1235.

Tanji J and Evarts EV (1976) Anticipatory activity of motor cortex neurons in relation to direction of an intended movement. *Journal of Neurophysiology* 39: 1062–1068.

Thach WT (1978) Correlation of neural discharge with pattern and force of muscular activity, joint position, and direction of intended next movement in motor cortex and cerebellum. *Journal of Neurophysiology* 41: 654–676.

Neural Coding of Spatial Representations

N M Bentley and E Salinas, Wake Forest University School of Medicine, Winston-Salem, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Neuronal Signals That Encode Spatial Location

In mammals, spatial information is derived primarily from three sensory modalities: vision, audition, and somatosensation. In many regions of the neocortex, and across sensory modalities, individual neurons fire action potentials at high rates only when objects to which they are sensitive appear in a particular region of space – the neuron's receptive field. For example, a neuron in the primary visual cortex (V1) responds vigorously only when a stimulus to which it is sensitive stimulates a particular region of the retina of approximately 1° of visual angle. Similarly, in primary somatosensory cortex, neurons respond only when a particular part of the body is touched, and tactile receptive fields on the fingertips are typically only a few millimeters wide. In the auditory system, some neurons respond only when a sound originates at some particular angle in space relative to the head. This case is special, however, because, unlike visual and somatic primary receptors on the eyes and skin, primary receptor neurons in the cochlea are not space-specific; the neurons that become selective for sounds from specific regions in space are constructed in a series of computational steps. In all of these cases, however, spatial information is conveyed basically through a labeled-line code, which means that the stimulus location can be inferred from the identity of the activated neuron or neurons. So, to determine where is the stimulus, one needs to know which neurons are responding.

As these examples illustrate, in general, spatial information is carried by firing rate distributions, and it is widely held that brains rely on such distributions to track object location. Cortical microstimulation experiments have largely corroborated this hypothesis. For instance, eye movements toward specific points in space can be elicited with high reliability by microinjecting electrical current at specific points in the cortex, a manipulation that increases the firing rate of a restricted population of neurons around the electrode. It remains possible, however, that brains also rely on other aspects of neural activity to encode object location, for example, correlations between the activity of different neurons or precise spike-train sequences. That firing rates carry a large amount of spatial information is an important premise on which many of the conclusions discussed here depend.

Receptive Field Properties

Receptive field sizes vary widely across regions of the cortex. Consider, for example, the visual system. From the thalamus, visual information first reaches the cortex in the back of the head, in the occipital lobe, where the V1 is located. Receptive fields in V1 are small, typically covering on the order of 1° of visual angle, and they are topographically arranged, meaning that neighboring neurons have receptive fields that overlap or are close to one another in visual space. From V1, visual information is transmitted to a number of different areas in the parietal, occipital, and temporal lobes. In general, the closer a visual area is to the front of the brain, the larger the receptive fields become and the less ordered their topography. The largest visual receptive fields can be almost as large as the whole visual field, as may be found, for example, in the inferotemporal cortex (IT). In general, receptive fields not only increase in size toward more central structures but also become more complex; for instance, V1 neurons are selective for small oriented bars or gratings (stripes), whereas IT neurons respond most strongly to highly complex images such as faces or whole objects.

Areas That Specialize in Spatial Processing

Cortical visual areas have traditionally been divided into two groups, which lie along the ventral and dorsal pathways. The distinction rests both on anatomical grounds and on functional evidence based on brain-damaged patients and lesioned primates, both of whom show consistent kinds of deficits depending on which pathway is affected. Damage to the ventral pathway, which includes areas V2, V4, and IT, generally leads to problems in object recognition (agnosias), whereas damage to the dorsal pathway, which includes V3, the middle temporal area (MT), and several parietal regions, such as areas 7a and the lateral intraparietal area (LIP), leads to problems in navigation and the perception of space (ataxias and visual neglect). The response properties of neurons in these areas also reflect their differing functional roles. Neurons in the ventral pathway are often highly selective for specific objects, but their responses to those objects are relatively invariant to their color, size, position, or perspective, among other properties. On the other hand, neurons in the dorsal pathway are typically not as selective for object identity, but their

activity often depends on the spatial relation between the viewed object and the organism's body. The general conclusion here is that parts of the brain are specialized for processing spatial information. Why this separation of function has evolved remains unclear. It should also be emphasized that the dorsal pathway is not exclusively involved in vision. Rather, it processes auditory and somatosensory signals in addition to visual signals and appears to be important for spatial processing and movement planning in general.

The Precision of Spatial Representations in the Brain

How precisely can we locate an object in space, and what determines this precision? First, spatial accuracy depends on the modality and the species in question. For example, auditory localization acuity varies tremendously across mammals, from hundreds of degrees in subterranean species to approximately 1° in humans and elephants. Second, acuity can depend on the form of the stimulus itself. For example, a person's ability to discriminate between fine-grained textures or Braille letters through touch depends on whether and how those stimuli move across the skin. Third, accuracy also typically depends on where an object is located. For example, spatial acuity in humans is much lower in the periphery than in the center of the visual field (i.e., the fovea), and somatosensory spatial acuity is much higher on the fingertips than on the palm of the hand. Physiologically, in both cases the differences are due to differences in the densities of primary receptors and cortical neurons dedicated to cover each region of space. Touch mechanoreceptor density is approximately four times higher at the fingertips, where acuity is high, than at the palm, where acuity is lower. Differences in photoreceptor numbers across the retina impose similar limits on acuity across the visual field. Discrimination performance in difficult perceptual tasks is often limited by such fundamental constraints as the density of primary receptors.

Tuning Curve Width and Optimal Sensory Coding

Naively, we might also expect receptive field sizes in a population of cortical neurons to limit the precision with which the population can represent an object's position. After all, if only one neuron is firing, the larger its receptive field, the larger is the uncertainty about the location of the stimulus that triggered the response. This intuition, however, is not always correct. First, information about stimulus location is typically spread out across a population of neurons with overlapping receptive fields, in which case spatial accuracy depends not only on the receptive field width but also on the density of neurons and their variability. An array of narrow tuning curves may be extremely ineffective if they leave gaps between them (a tuning curve here is the mean response of a neuron plotted as a function of stimulus location). On the other hand, an array of neurons with wide and overlapping receptive fields may be highly accurate if their variability is low. Some of these possibilities are illustrated in Figure 1.

A second and more profound factor that affects spatial accuracy was identified by Zhang and Sejnowski. They used a measure of accuracy known as the Fisher information to characterize the efficiency of populations of model cortical neurons with tuning curves of different widths. They found that the relationship between tuning width and Fisher information depends on the dimensionality of the stimulus to be represented. That is, whether narrow or wide curves are better depends crucially on the number of parameters that the neurons are representing. Recent work along similar lines by Brown and Bäcker showed that, when an infinite number of neurons cover one dimension (say, distance to a target) or two dimensions (say, target location on a surface), encoding accuracy always increases with decreasing tuning width – narrower curves are always better. However, for three dimensions or higher, the advantage of narrow curves goes away; there is an optimal tuning width in terms of Fisher information, even though the number of neurons remains infinite.

Another theoretical study, by Pouget and colleagues, demonstrated that whether wide or narrow tuning curves are better also depends on the mechanisms that generate those tuning curves. In many models, tuning curve width is determined by the strength of lateral interactions between neurons in a population, such that stronger interactions are needed to generate narrow tuning than wide tuning. Strong lateral interactions, however, also tend to produce strong noise covariations across neurons; that is, the random response fluctuations of pairs of neurons become more similar to one another. In practice, this may limit information transmission quite severely, and therefore, a population with wide tuning curves may be much more informative than a population with narrow curves if the narrowing involves traditional mechanisms that increase noise covariations. This result suggests that generating arrays of narrow tuning curves from arrays of wide ones is probably a rather inefficient strategy, even if the encoded stimulus is one- or two-dimensional, because narrowing mechanisms effectively increase the noise.

In summary, then, in terms of the accuracy for encoding an object's position, the optimal tuning curve width depends on the number of neurons in

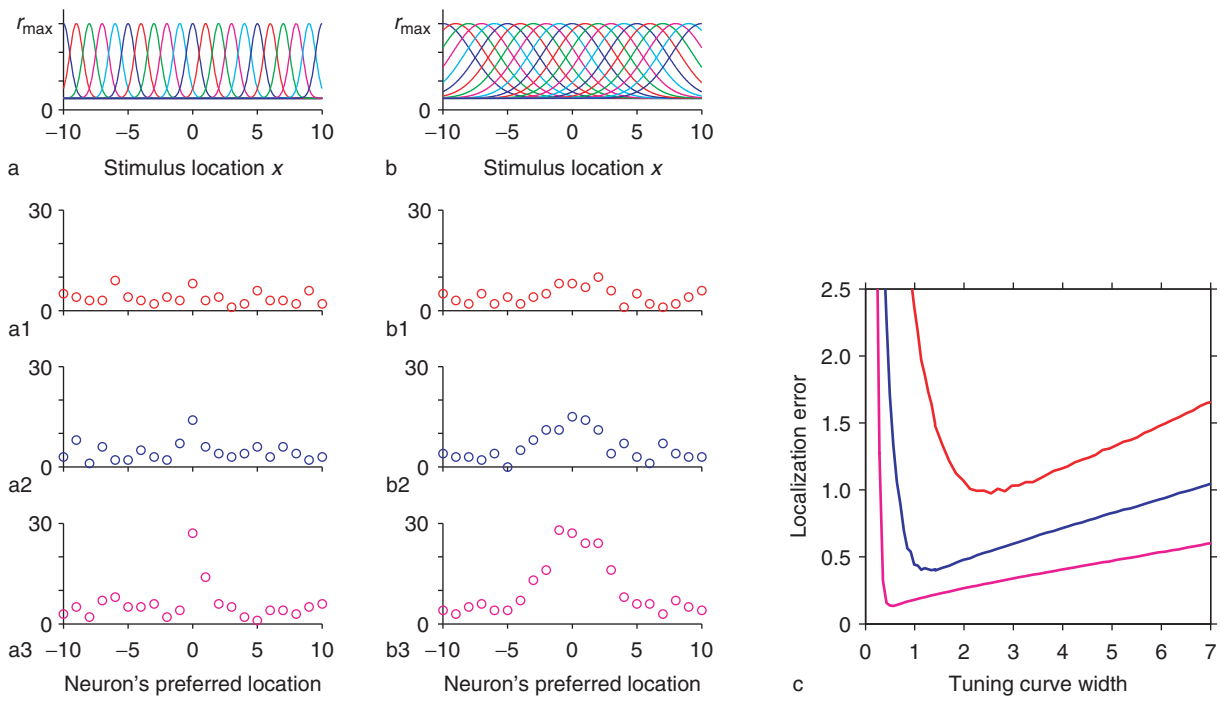

Figure 1 Encoding of spatial location by populations of neurons with Gaussian tuning curves: (a) a population of narrowly tuned neurons, tuning curve width = 0.5; (b) a population of widely tuned neurons, tuning curve width = 2.3; (c) localization error as a function of tuning curve width. In (a), each curve represents the mean response of a neuron as a function of the location of a stimulus x. There are 21 neurons, and x may be between −10 and 10 (arbitrary units). The tuning curves are narrow (width $\sigma = 0.5$), so only one or two neurons fire significantly above the baseline in any trial (a trial is one stimulus presentation). The graphs (a1)–(a3) show the responses of the 21 neurons in three single trials with low (red points), medium (blue points), and high (magenta points) signal-to-noise ratios. Variability across trials was modeled as a Poisson process, so the variance of each response was equal to the mean. The signal-to-noise ratio was controlled by varying the maximum response of the neurons: $r_{max} = 10$, 15, and 30 spikes s^{-1} for the red, blue, and magenta points, respectively. Background firing was constant at 4 spikes s^{-1}. (b) As in (a), but for wide tuning curves (width $\sigma = 2.3$). In this case, approximately eight neurons typically fire above baseline in each trial. (c) Localization error as a function of tuning curve width for the three signal-to-noise ratios in (a) and (b). The error is the average deviation between the true stimulus location, x, and the location that is estimated from the responses of the 21 neurons in each trial, x_{est} (more specifically, $E^2 = <(x - x_{est})^2>$, where E is the localization error and the brackets indicate an average over trials and over all values of x). The width that is optimal for representing stimulus location depends on the signal-to-noise ratio. If it is high (magenta curve), narrow tuning curves are better because they generate smaller localization errors; in contrast, if the signal-to-noise ratio is low (red curve), broader tuning curves are more accurate.

the population, their noise properties, and the number of dimensions of the encoded position vector.

Tuning Curve Width and Motor Constraints

An interesting and perhaps counterintuitive observation about the neural activity that encodes spatial location is that it should not be more accurate than required by the animal's behavior. Just as commuting between home and school in a Formula 1 race car would be costly and inefficient, localizing objects with greater accuracy than necessary would waste resources. A recent theoretical study by Salinas highlighted the crucial role that motor constraints play in determining optimal sensory representations, and an impressive data set collected by Heffner and colleagues illustrated the point. Based on the analysis of more than 30 species of mammals, Heffner and collaborators found that sound localization capacity varies tremendously in this group – discrimination thresholds range from approximately 1°in humans and elephants to approximately 25°in gerbils and horses. Discrimination threshold in this case corresponds to an angular separation between two sound sources, such that a subject can distinguish with 75% accuracy which source generated a short noise burst. According to Heffner's analysis, the measured differences across species are not accounted for by variations in sound environment, interaural distance, or animal lifestyle. Rather, sound localization acuity in mammals seems to depend on the precision required for the visual orienting response to sound. The logic is as follows. The primary function of sound localization is to produce an orienting response, that is, to bring the sound source into the most sensitive part of the field of view (the fovea) for visual inspection. Therefore, species that have small areas of best vision

(e.g., humans and elephants) need to make highly precise movements, whereas species that have large areas or streaks of optimal vision (e.g., gerbils and horses) do not. This is an important observation because it means that when behavior does not require high accuracy, the sensory representation should be correspondingly coarse, even if, in principle, it could be made more precise.

Reference Frames in the Cortex

References for Sensory Responses

Because there are no absolute spatial coordinates, an object's position in space can only be specified relative to a landmark that serves as reference. A crucial question therefore is: relative to which reference or references does the neocortex represent object locations? Experiments suggest that different parts of the cortex use different references, and some neurons seem to combine multiple ones. Those regions of cortex that receive sensory information most directly often use the topography of peripheral sensory organs themselves as references. For example, in the early visual cortex the reference is the retina – V1 neurons are said to encode visual information in retino-centered or eye-centered coordinates. In the early somatosensory cortex, the reference is the skin. Audition is different because firing rate distributions among the inner ear's sensory neurons do not depend in a straightforward way on where in space a sound originates. Thus, early auditory neurons do not represent spatial information in an ear-centered coordinate frame. Further downstream, however, through a series of computations involving the combination of interaural level differences, monaural spectral cues, and interaural time differences, cortical neurons create a representation of sound-source location in a head-centered or eye-centered reference frame.

References for Motor Responses

Just as sensory neurons require a reference for encoding the position of objects that are seen, heard, or touched, the motor system requires one or more references in order to move a limb to a desired position. In more central areas, the reference frames of cortical neurons that contribute to generating behavior are more complex than those of traditional sensory neurons. In a series of landmark papers, Gross, Graziano, and their colleagues showed that in the premotor cortex of monkeys many neurons fire strongly only when an object is seen at a particular place relative to a specific body part; that is, they have body-part-centered reference frames. For example, many premotor neurons are bimodal, in that they respond to both visual and somatosensory stimuli, and their visual receptive field is typically a three-dimensional region of space near the somatosensory receptive field. The key finding is that both of these receptive fields move where the body part moves. Neurons with receptive fields anchored to the arms, upper body, and head are especially prevalent. This multiplicity of reference frames suggests that neurons in the motor system rely on reference points specific to and convenient for the types of behavior they help to generate. Visual receptive fields attached to the foot may be useful for playing soccer, visual receptive fields attached to the mouth may be useful for biting an apple, and so on. Thus, the responses of such neurons may allow a sensory signal to act directly as a motor signal. Another implication is that sensory processing is not limited to sensory cortex, traditionally conceived. In fact, sensory processing seems to occur all over the cortex and to be inseparably intermingled with motor planning.

Changing Reference Frames

That sensory and motor cortical areas exhibit different sets of reference frames indicates that information about object location must be transformed from one set of reference frames to another as it propagates between brain regions. This operation is known as a coordinate transformation. According to the work of Richard Andersen and collaborators, the posterior parietal cortex (part of the dorsal pathway, not surprisingly) is largely responsible for coordinate transformations in the cortex. Many posterior parietal neurons are multimodal, and many represent locations in mixed reference frames, which means that receptive fields are partially anchored to more than one body part. For example, parietal neurons often have receptive fields that are sensitive to the position of the eyes in the orbit as well as to the position of the head relative to the body. Several theoretical studies have demonstrated that such mixed representations can underlie coordinate transformations. Lesion studies also suggest that the parietal cortex is crucial for coordinate transformations. Parietal lesions famously lead to neglect, a condition wherein a person loses awareness of a region of space. The condition consists not merely of blindness or of a failure to locate objects within a region but of a total lack of awareness of the region's existence. Neglect does not have one characteristic reference frame; rather, it depends on the nature of the task at hand (and of the lesion). The neglected region of space may be defined relative to the eye, the arm, the body, or even an object in the environment. These observations support the idea that the parietal cortex represents the location of objects in many different reference frames.

As an example, consider the problem of reaching for an object, say, a teacup. To grasp the teacup, the brain must compute the difference between the position of the hand and the position of the cup (with positions specified by vectors). To compute this motor error, the two positions can be defined with respect to a number of reference frames. It is thought, however, that at some point in the visuomotor cascade the two locations must be encoded in the same reference frame and that the corresponding neural signals should be combined to produce the motor error. Indeed, neurons in a cortical area called the parietal reach region respond to both target and current hand positions in an eye-centered reference frame. Let us call these E neurons. For E cells, there is a special point in the visual field (the preferred target location) such that their activity is strongest when the target is located there; but there is also another point (the preferred hand location) such that their activity is strong when the hand is seen there. When the subject moves his eyes, each E neuron changes its firing rate by some amount, but as a population the E responses still encode the hand and target positions (if these have not changed). In contrast, neurons in another parietal area, area 5, seem to encode the displacement vector explicitly, independently of eye position. We call these M neurons. They have a preferred displacement vector, so their activity can be described as signaling target position in hand coordinates. The key difference it that, when the subject moves his or her eyes, the activities of all M neurons remain unchanged. This resembles much more closely a motor command. There is yet another type of parietal neuron that seems to use a mixed or intermediate representation. The activity of those intermediate cells signals target location and is sensitive to both hand and eye positions, but the dependence is more complicated than for E and M neurons. Their properties are similar to those observed in intermediate layers of artificial network models that are trained to transform object location from one reference frame to another, so they are thought to be an important link for constructing M neurons out of E neurons.

How exactly such coordinate transformations occur remains unclear, but experimental and modeling studies indicate that an important component is a mechanism called gain modulation. Gain modulation is observed when the sensitivity of a neuron to a stimulus changes as a function of another variable or parameter. For example, when the response of a parietal neuron is plotted as a function of the position of a spot of light, the resulting curve typically has a bell shape. For a neuron that is gain modulated by, say, eye position, the amplitude of the curve varies as a function of eye position. Andersen and colleagues documented the modulation of visual responses depending on a variety of parameters, such as eye and head position, arm position, and eye velocity, in many parts of the parietal cortex. This particular way to integrate information from multiple sources is powerful because populations of gain-modulated neurons do not represent spatial information in a fixed reference frame; rather, by combining those responses appropriately, different groups of downstream neurons can extract spatial information in the reference frame that best suits their function.

Keeping Objects Separate

As previously mentioned, evidence suggests that separate parts of the cortex are specialized for processing information about object identity and location. Psychophysics corroborates this idea because brain damage can destroy a person's ability to locate an object in space while leaving relatively intact his ability to identify it, or vice versa. Therefore, it has been suggested that there must be some mechanism by which the brain can keep track of which visual features appear in which parts of visual space; otherwise, it would be impossible for an organism to avoid confusing the attributes of multiple objects in different locations. This is often referred to as the binding problem. Evidence that such a mechanism exists comes from studies of patients with Balint's syndrome, which can occur following bilateral damage to the parietal lobes. A Balint's patient is capable of seeing only one object at a time. When there is more than one object in the visual field, the patient often claims that the object he or she sees has some of the attributes of other objects that he or she does not see, such as color, even though it does not. It is unknown whether such patients actually see objects or just jumbles of features, but it is undeniable that they somehow lose track of the locations of different features in space, suggesting that they have lost some mechanism by which normal brains avoid such problems.

One of the earliest and most influential attempts to explain our ability to keep track of object attributes in space is called feature integration theory, developed by Anne Treisman. The theory posits that, when the brain directs attention to an object, the neural representations of its features and its location are selectively enhanced (the firing rates of responding neurons are increased); this allows these representations to be singled out from those of other features and locations, so that there is no longer an ambiguity regarding which belongs with which. Although this is a powerful theory and it is consistent with known effects of attention on neural activity, a number of objections have been raised over the years. For

example, numerous studies suggest that a person can attend to multiple objects simultaneously, without confusing their attributes. Standard feature integration theory does not specify how the brain resolves location ambiguities in such cases.

Another theory, which was popularized by von der Malsburg, Singer, and Gray, among others, is based on binding by neural synchrony. The idea is that the neurons representing an object's features and location fire spike trains in synchrony with one another and out of synchrony with neurons representing the features and locations of other objects. The synchrony between neurons tags the encoded features as belonging together. Unlike in feature integration theory, in this mechanism features and locations of multiple objects can be simultaneously grouped. Although there is substantial evidence that is consistent with this hypothesis, in recent years important violations of its predictions have also appeared, so its validity remains an open question.

It should also be pointed out that the binding problem may not be quite as difficult as it seems at first. Although it is true that we rely on different areas of the cortex to track object identity and location, those areas are not totally devoid of information regarding their conjunction. For example, inferotemporal neurons, thought previously to be largely insensitive to object location, can be very sensitive to location when the visual field contains multiple objects. However, because the brain does not rely directly on this spatial information to locate an object in space (we know it relies on mechanisms in the dorsal pathway, and the parietal regions in particular), object representations in the ventral pathway may still have to be matched to spatial representations in the dorsal pathway. The conjunction information in IT could provide a basis for this match.

See also: Attentional Networks in the Parietal Cortex; Balint Syndrome; Information Coding; Motor Psychophysics; Natural Images: Coding Efficiency; Neglect Syndrome and the Spatial Attention Network; Neural Synchrony and Feature Binding; Parietal Cortex and Spatial Attention; Spatial Transformations for Eye–Hand Coordination; Spatial Cognition.

Further Reading

Avillac M, Deneve S, Olivier E, Pouget A, and Duhamel JR (2005) Reference frames for representing visual and tactile locations in parietal cortex. *Nature Neuroscience* 8: 941–949.

Brown WM and Bäcker A (2006) Optimal neuronal tuning for finite stimulus spaces. *Neural Computation* 18: 1511–1526.

Buneo CA and Andersen RA (2006) The posterior parietal cortex: Sensorimotor interface for the planning and online control of visually guided movements. *Neuropsychologia* 44: 2594–2606.

Cohen YE and Andersen RA (2002) A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience* 3: 553–562.

Deneve S, Latham PE, and Pouget A (2001) Efficient computation and cue integration with noisy population codes. *Nature Neuroscience* 4: 826–831.

Graziano MSA, Hu TX, and Gross CG (1997) Visuospatial properties of ventral pre-motor cortex. *Journal of Neurophysiology* 77: 2268–2292.

Heffner RS (1997) Comparative study of sound localization and its anatomical correlates in mammals. *Acta Otolaryngological (Stockholm)* 532(supplement): 46–53.

Johansson RS and Vallbo AB (1979) Tactile sensibility in the human hand: Relative and absolute densities of four types of mechanoreceptive units in glabrous skin. *Journal of Physiology* 286: 283–300.

Knudsen EI (2002) Instructed learning in the auditory localization pathway of the barn owl. *Nature* 417: 322–328.

Palanca BJA and DeAngelis GC (2005) Does neuronal synchrony underlie visual feature grouping? *Neuron* 46: 333–346.

Pouget A and Sejnowski TJ (2001) Simulating a lesion in a basis function model of spatial representations: Comparison with hemineglect. *Psychological Review* 108: 653–673.

Salinas E (2006) How behavioral constraints may determine optimal sensory representations. *Public Library of Science, Biology* 4(12): e387.

Series P, Latham PE, and Pouget A (2004) Tuning curve sharpening for orientation selectivity: Coding efficiency and the impact of correlations. *Nature Neuroscience* 7: 1129–1135.

Singer W and Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* 18: 555–586.

Xing J and Andersen RA (2000) Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames. *Journal of Cognitive Neuroscience* 12: 601–614.

Zhang KC and Sejnowski TJ (1999) Neuronal tuning: To sharpen or broaden? *Neural Computation* 11: 75–84.

Neural Crest

H M Young and R B Anderson, University of Melbourne, VIC, Australia
D Newgreen, Murdoch Childrens Research Institute, Parkville, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The development of the nervous system begins overtly with the formation of the neural plate in the surface ectoderm. Soon after the neural plate forms, it folds in on itself to form neural folds. The dorsal lips of the neural folds subsequently fuse to form the neural tube, a process called primary neurulation. Just before, during, and slightly after the fusion of the neural folds, a population of neuroepithelial cells in the dorsal lips of the neural fold and the dorsal region of the neural tube undergo an epithelial–mesenchyme transition (EMT) and emigrate away from neural tube; this cell population is the neural crest (Figures 1(a) and 1(b)). The formation of the lumbosacral neural tube in birds and mammals, and the entire neural tube in fish, occurs by a different process, known as secondary neurulation. Here, a compact mass of cells forms and then cavitates to form a neural tube. Neural crest cells also form at the dorsal regions of the neural tube, adjacent to the epidermis, during secondary neurulation. The neural crest is a transient structure that only exists around the time of formation of the neural tube. Neural crest cells migrate along particular pathways before settling in many different locations. They give rise to most of the facial skeleton, the peripheral nervous system, melanocytes, and many other derivatives.

The neural crest was first described and recognized as a migratory population in the 1860s. From the 1920s on, many methods have been used to experimentally probe its development, including ablation (removal) of neural crest in living embryos and later observations of deficits, and application of various stains to neural crest cells to follow them. However, the method that has been pivotal in studying the migration pathways and fates of neural crest cells is the quail–chick chimera technique. This method was devised in the late 1960s by Nicole Le Douarin, who saw that the nuclei of quail cells can be distinguished from those of chick cells using histological or immunohistochemical techniques (Figure 2(c)). Segments of neural primordium containing premigratory neural crest cells from specific regions of the neural axis in quails are transplanted into chick embryos (Figures 2(a) and 2(b)), and the location of the quail cells within the chick host embryos can be determined later, thus revealing the migratory pathways and fate of neural crest cells from different axial levels.

The neural crest can be subdivided into three main axial populations – cranial, vagal, and trunk – although the vagal population includes the most caudal cranial and the most rostral trunk neural crest cells (Figure 3). The different axial populations migrate along different pathways and each gives rise to a range of cell types. Some of these cell types are common to several axial levels, whereas some are unique. For example, all neural crest levels give rise to sensory neurons. In contrast, cranial neural crest cells give rise to most of the bone and cartilage of the head and neck, whereas trunk neural crest cells have minimal ability to form skeletal elements. This article discusses the induction of the neural crest and the migratory pathways and derivatives of cranial, vagal, and trunk neural crest cells.

Induction and EMT

The first portents that the neural crest is different from the more medial ectoderm of the neural plate (which produces the central nervous system) and the more lateral epidermis (which produces the skin) are seen in gastrulation, even before neurulation. At this stage, a restricted region of epiblast (surface cell layer) fated to contribute to the neural crest expresses the paired-box transcription factor gene *Pax-7*. This region of epiblast is able to produce cells with neural crest markers when isolated from all inductive influences, and suppression of Pax-7 prevents this.

Later, important signaling processes occur across the border between the neural and epidermal ectoderm to mark this zone as different from the adjacent but contiguous epithelia. Important in this process are secreted growth factors of the Wnt family and of the transforming growth factor family, especially bone morphogenetic proteins such as BMP-4. The lineage-specifying action of BMP-4 in particular has a quantitative nature; high levels specify epidermal lineage and moderate levels promote neural crest characters, whereas low levels permit neural plate character. BMP-4 gene expression is homogenetically induced in the dorsal neural ectoderm, giving a positive feedback capacity. A gradient of BMP-4 activity is sharpened by a countergradient of its secreted inhibitor Noggin. These signals from the epidermis act within the ectodermal layer on the edge (later dorsal) cells of the neural ectoderm. In addition,

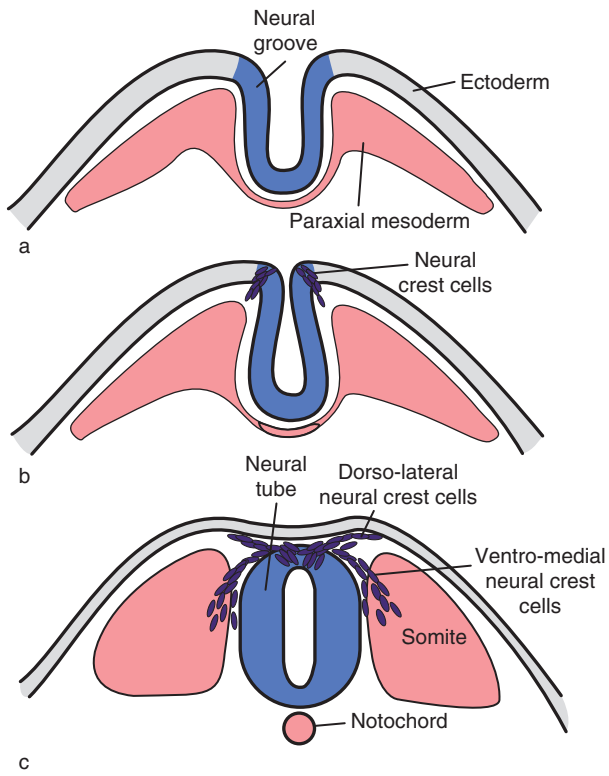

Figure 1 (a, b) Primary neurulation, which involves the formation of the neural plate and folds, and subsequent closure of the folds to form the neural tube. Neural crest cells delaminate from the dorsal regions of the neural folds and neural tube. (c) In the trunk region, neural crest cells migrate along two main pathways; first, they migrate along a ventral pathway through and between the somites to form sympathetic ganglia, Schwann cells, and dorsal root ganglia. Later, emigrating cells migrate along a dorsolateral pathway to form melanocytes.

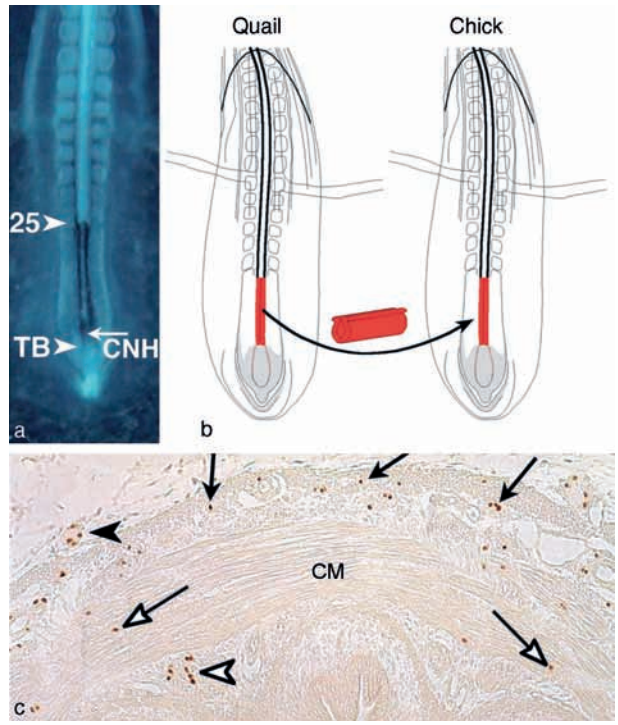

Figure 2 Quail–chick grafting technique devised by Nicole Le Douarin and colleagues. A defined region of the neural tube containing premigratory neural crest cells, in this case sacral neural tube, is removed from chick embryos (a) and replaced with transplanted neural tube from quail embryos (b). Following subsequent development, the location of the quail cells within the chick host embryos can be determined using immunohistochemical methods. CNH, chordoneural hinge; TB, tail bud. (c) Transverse section through the hindgut following quail–chick sacral neural crest grafting. Quail cells, which are stained brown, are found within nerve bundles outside the gut (arrowheads), external (arrows) and internal (open arrowhead) to the circular muscle layer (CM) of the gut, as well as within the circular muscle layer (open arrows). These experiments show that sacral neural crest cells give rise to derivatives both within and outside the hindgut. Reproduced from Burns AJ and Douarin NM (1998) The sacral neural crest contributes neurons and glia to the post-umbilical gut: Spatiotemporal analysis of the development of the enteric nervous system. *Development* 125: 4335–4347, from The Company of Biologists.

fibroblast growth factor signaling from another layer, the underlying mesoderm, is also important for neural crest identity.

The outcome of this signaling is the expression of specific genes in the lateral neural plate border region which includes the dorsal neural tube as well as cells fated for neural crest; these 'border genes' include *Zic3* and *Pax-3*. Expression of other genes such as *Sox-2* in the medial region act to limit the neural crest zone to the neural plate border. On the lateral epidermal side, the high levels of BMP-4 signaling induce genes such as *Dlx* and *Msx* family members, whose products repress a variety of proneural genes such as *Zic3* and *Sox-2*, while promoting expression of epidermis-defining keratins.

Experimental manipulation of these border genes indicates that they control expression of genes that are closely related to neural crest fate and to neural crest behavior such as EMT and migration. Included in this gene repertoire are the zinc finger transcription factors *Snail-1* or *Snail-2* (formerly called *Slug*). These genes have an interesting evolutionary history: the neural crest of fish, amphibians, and mammals expresses *Snail-1*, whereas that of reptiles and birds employs *Snail-2*. In a very short period of time, expression of the forkhead gene *FoxD3*, and the Sry HMG-box genes of the *Sox-E* subfamily (*Sox-8*, *-9*, and *-10*), is induced in the neural crest region. When overexpressed (or suppressed), these genes can induce (or inhibit) neural crestlike behavior such as migration in nonneural crest neural ectoderm, as shown first by Angela Nieto for *Snail2*. These genes are sharply temporally regulated, and with the exception of *Sox-10*, their mRNA levels decline around or just after the time when neural crest migration commences. *Sox-10*, however, continues to be expressed

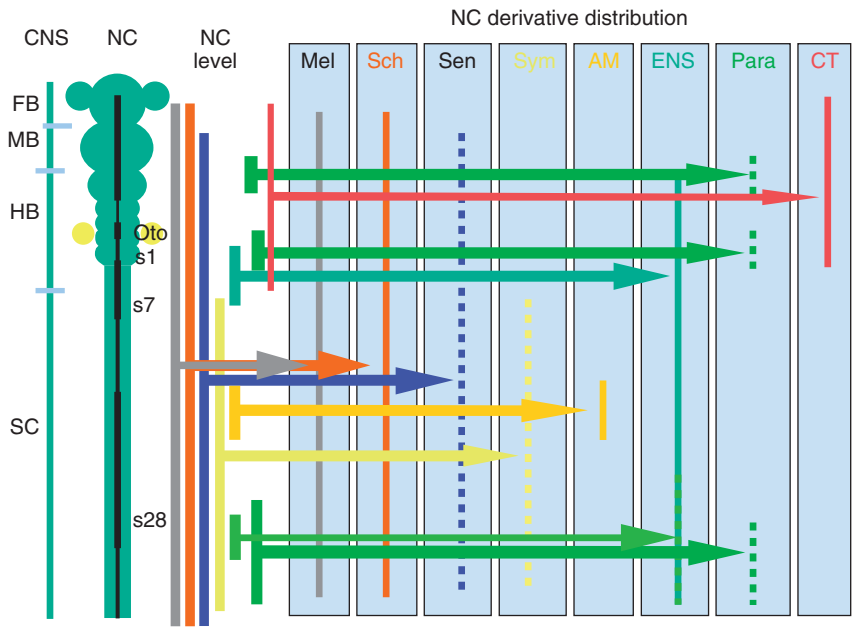

Figure 3 Axial level of origin of neural crest derivatives in chick embryos. Melanocytes arise from all levels of the neural axis, whereas only cranial neural crest contributes to skeletal elements (the cranial skeleton). Cardiac neural crest, which arises adjacent to somites 1–3 and contributes to the outflow tracts of the heart, is not included in the diagram. AM, adrenomedullary; CNS, central nervous system; CT, connective tissue; ENS, enteric nervous system; FB, forebrain; HB, hindbrain; MB, midbrain; Mel, melanocyte; NC, neural crest; oto, otic vesicle; Para, parasympathetic ganglion; s, somite; Sch, Schwann cell; Sen, sensory neuron; SC, spinal cord; Sym, sympathetic ganglia.

especially in vagal and trunk neural crest and has other functions; including maintenance of an undifferentiated state. *Dlx* and *Sox-9* expression reappears later in postmigratory cranial neural crest-derived cells prior to their formation of skeletal tissues. Work in many species by many authors, particularly Marianne Bronner-Fraser and colleagues, has opened up this field of neural crest induction. Yet the regulatory circuits between these genes are complex and not fully elucidated (Figure 4). Moreover, although these cells are specified as neural crest (since they express many appropriate genes), they are not necessarily committed to this fate. The morphogenetic event of EMT might be thought of as marking a watershed in neural crest cell lineage commitment, yet experimental transplantation back into the neural tube reveals that at least some can still contribute to the neural tube.

The neural crest at any one level produces several different types of differentiated cells, and different axial levels produce some unique types. At any one level, during the premigratory phase, some cells are broadly multipotent within the neural repertoire and differentiation is guided by exposure to instructive cues they encounter later. Others become more restricted; the last cells to commence emigration are more likely to be restricted to a melanocyte lineage, and this is related to prolonged exposure to the neural tube. The intrinsic differences between axial levels of the neural crest are tied to the general positional identity of axial structures. Broadly stated, the cranial neural crest has a less restricted range of potentiality, whereas more caudal neural crest lacks some potentials. This is most notably demonstrated by the inability of trunk neural crest to produce connective tissue derivatives such as cartilage. A similar but less complete restriction is also seen in the limited ability of trunk neural crest to generate an enteric nervous system. This seems to involve the nested expression of *Hox* genes, and there is genetic cross-modulation between these axial position genes and some to the genes previously mentioned that specify the neural crest.

The outcome of this is the EMT and onset of migration of neural crest cells, sometimes termed delamination. This is the archetypal example of a process repeatedly seen in development and which seems to be replicated pathologically in carcinoma invasion. This involves relocation and reduction of the molecules that hold the neural epithelial cells together, most prominently the classic adherens junction molecule, N-cadherin. Experiments indicate that N-cadherin inactivation can trigger migration in neural epithelial cells. Other nonclassic cadherins are induced, and these may be important in maintaining the transient cell–cell adhesions seen between migrating

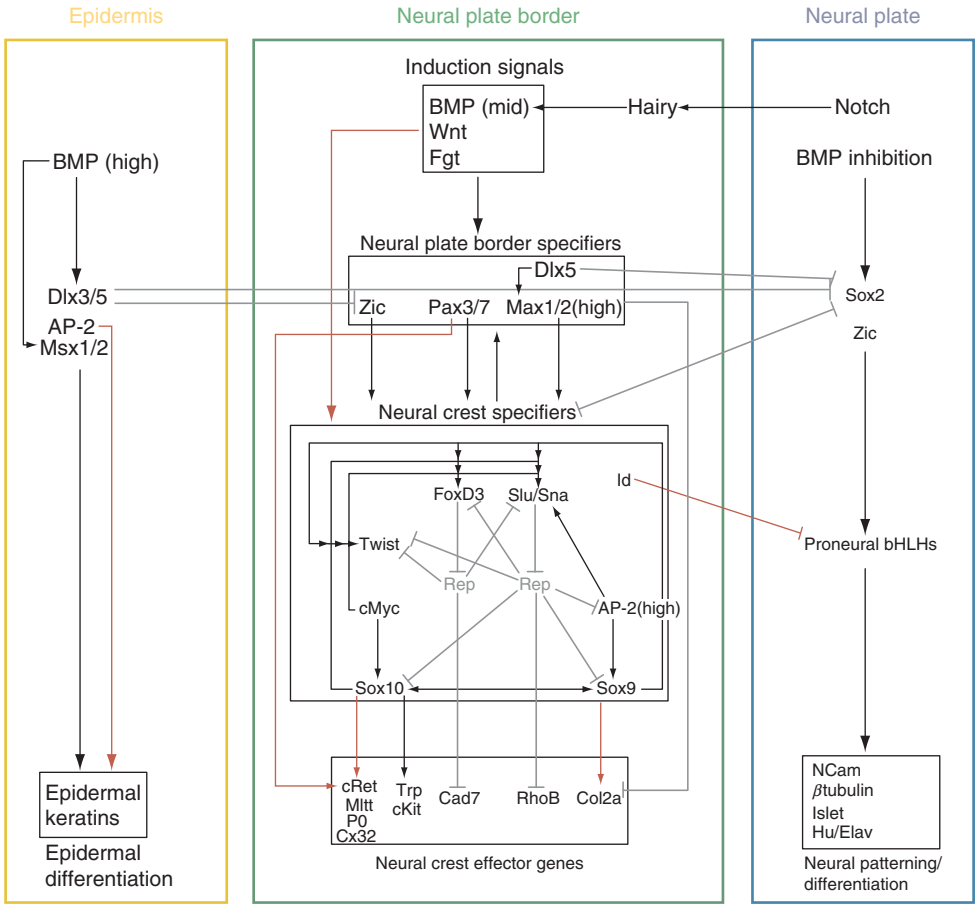

Figure 4 Putative gene regulatory and signaling interactions in action at the neural plate border of vertebrates. Red arrows, proven direct regulatory interactions; black arrows, genetic interactions suggested by gain- and loss-of-function analyses largely in *Xenopus*; gray lines, repression. Reproduced from Meulemans D and Bronner-Fraser M (2004) Gene-regulatory interactions in neural crest evolution and development. *Developmental Cell* 7: 291–299, with permission from Cell Press, Elsevier.

neural crest cells. In addition, there are changes in the suite of integrins expressed on the cell surface. The integrins allow new adhesive contacts to develop with extracellular matrix (ECM) molecules, which form the substrate for the first steps of neural crest cell migration. Like other migrating cells, neural crest cells modify their surrounding ECM via expression of several matrix metalloproteases (MMPs). Inhibition of MMPs can prevent neural crest EMT and subsequent migration, and in particular may be required to penetrate the neural tube basal lamina, as seen at the cranial level in rodents. In addition, the actin-based cytoskeleton of these cells is reorganized from the apico-basal polarity typical of neural epithelial cells to the more flexible network arrangement which allows cell motility. The transient expression of the *RhoB* gene immediately prior to EMT may be involved in this.

The timing of neural crest EMT is stereotyped but not precisely related to other morphogenetic events in the neural epithelium. Sometimes, the newly mesenchymal neural crest cells individually crawl out of the neural epithelium before neural tube closure (cranial level of rodents), but this event occurs after neural tube closure in birds and in the trunk of rodents. In amphibians, the neural crest separates as a coherent population from the neural tube and only commences migration later, when permissive ECM changes occur.

Cranial Neural Crest

Cranial neural crest cells contribute to much of the bone, cartilage, and connective tissue in the head, including most of the head skeleton and parts of the teeth. Many congenital craniofacial deformations are due to defects in the development of cranial neural crest cells. Cranial neural crest cells also give rise to some smooth muscle in the head, melanocytes, all cranial parasympathetic neurons and their support

cells, and all glia and some, but not all, neurons in cranial sensory ganglia (Figure 3). Sensory neurons in cranial ganglia that do not arise from the neural crest arise from placodes, which are specialized regions of the ectoderm. Some cells within the thyroid, parathyroid, and thymus glands also arise from the cranial neural crest.

In the hindbrain, the segmental rhombomeres not only play a vital role in the patterning of branchiomotor nerves and cranial ganglia but also influence neural crest cell migratory pathways in this region. Hindbrain neural crest cells migrate in three distinct streams adjacent to rhombomeres 2, 4, and 6 and populate the first, second, and third branchial arches, respectively (Figures 5(a) and 5(b)). The regions adjacent to rhombomeres 3 and 5 are therefore neural crest cell-free zones. Some studies attributed the absence of neural crest cells adjacent to rhombomeres 3 and 5 to apoptosis of premigratory neural crest cells at these axial levels induced by signals derived from the even-numbered rhombomeres. However, subsequent studies by Peter Farlie, Paul Kulesa, Scott Fraser, Paul Trainor, and Rob Krumlauf have shown that neural crest cells are generated at odd-numbered segments but migrate anteriorly and posteriorly to join the streams of neural crest cells migrating from the even-numbered rhombomeres. The authors of these studies have proposed that the environment adjacent to the neural tube determines the behavior of cranial neural crest cells – inhibitory signals present adjacent to rhombomere 3 induce neural crest cells at this level to migrate anteriorly and posteriorly and limit their generation, whereas the otic vesicle provides a physical barrier to neural crest cells generated at rhombomere 5.

Cranial neural crest cells, through their interactions with facial epithelium and neuroectoderm, contribute to the development of most craniofacial structures in vertebrates (Figure 6). Different species have different facial morphologies. To examine whether species-specific facial patterning is intrinsic to neural crest cells or determined by cues from the facial ectoderm, neural crest cells destined to form the upper beak from duck were transplanted into quail embryos and vice versa. The resultant chimeric 'qucks' (duck embryos containing quail frontonasal neural crest cells) had short beaks typical of quail, and 'duails'

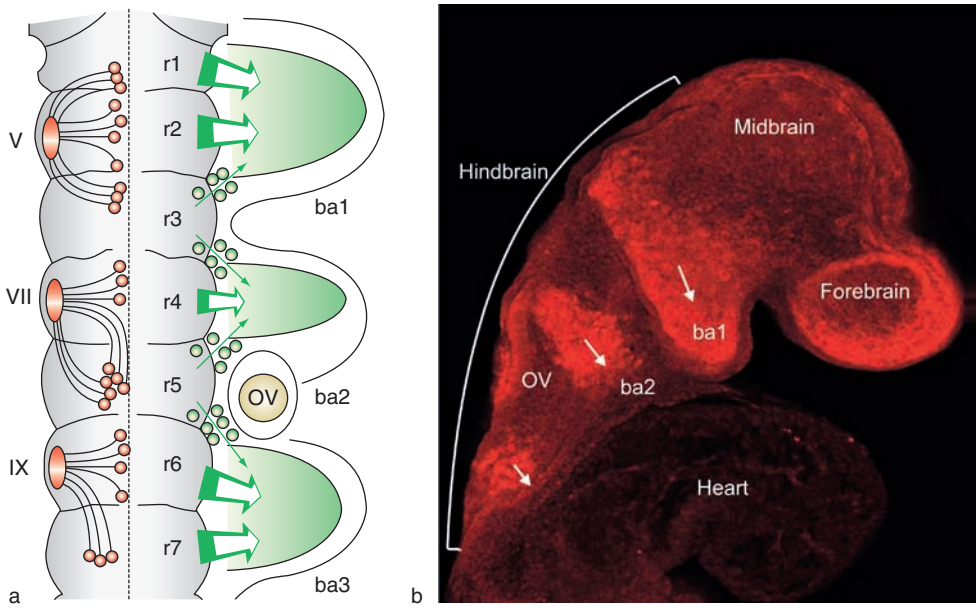

Figure 5 (a) Diagram showing migration of hindbrain neural crest cells in three streams. Neural crest cells from rhombomeres 1 and 2 (r1, r2) colonize branchial arch 1 (ba1). Only a small number of neural crest cells are generated in rhombomere 3 (r3), and those that are generated migrate anteriorly or posteriorly to join the streams lateral to r2 and rhombomere (r4), respectively. Neural crest cells from r4 colonize branchial arch 2 (ba2). The otic vesicle (OV) forms a physical barrier to neural crest cells from rhombomere 5 (r5) and they consequently migrate anteriorly or posteriorly to join the streams lateral to r4 and rhombomere 6 (r6), respectively. Neural crest cells from rhombomeres 6 and 7 colonize branchial 3 (ba3). Cranial motor nerves are shown on the left-hand side: V, trigeminal; VII, facial; IX, glossopharyngeal. (b) Head of an 8.5-day-old mouse embryo that had been processed for immunohistochemistry using an antibody to the neurotrophin receptor p75, which labels neural crest cells; p75-immunoreactive cells are stained red. There are three streams of neural crest cells (arrows) migrating from the hindbrain into the branchial arches (ba). At this developmental stage, crest cells from r1 and r2 are already colonizing ba1, but migration of neural crest cells caudal to the OV from r6 has just commenced. (a) Reproduced from Trainor PA, Sobieszczuk D, Wilkinson D, and Krumlauf R (2002) Signalling between the hindbrain and paraxial tissues dictates neural crest migration pathways. *Development* 129: 433–442, from The Company of Biologists. (b) Adapted from Anderson RB, Stewart AL, and Young HM (2006) The phenotypes of neural crest-derived cells in vagal and sacral pathways. *Cell Tissue Research* 323: 11–25, from Springer-Verlag.

(quail embryos containing duck neural crest cells) possessed flattened ducklike beaks (Figure 7). These transplantation experiments showed that neural crest cells intrinsically contain species-specific information, which is a major determinant of facial morphology. The genes involved in facial patterning are well conserved between species, and different facial morphologies are probably in part generated by differences in the temporal patterns of expression of these genes. Quail cranial neural crest cells transplanted into duck not only retained their own temporal pattern of gene expression but also altered the pattern of gene expression in nonneural crest-derived host tissue.

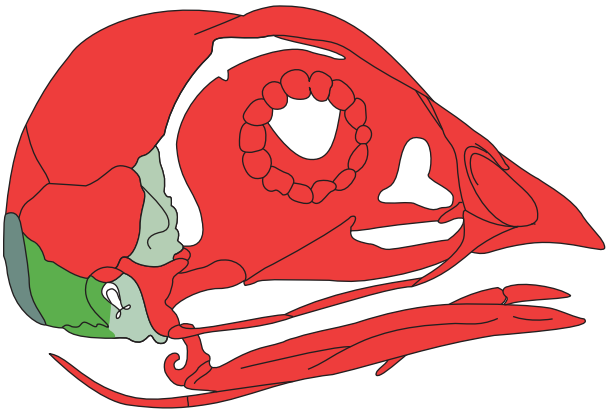

Figure 6 Lateral view of an avian skull showing the contribution of cranial neural crest (red), cephalic mesenchyme (blue), and somitic mesoderm (green). Adapted from Le Douarin NM and Kalcheim C (1999) *The Neural Crest*, 2nd edn. Cambridge, UK: Cambridge University Press.

Studies have also examined whether cranial neural crest cells at different rostrocaudal levels are prespecified to a particular fate prior to emigration from the neural tube. For example, neural crest cells destined to give rise to first branchial arch skeletal derivatives have been transplanted to a more posterior location in the hindbrain. In the 1970s, Drew Noden found that such embryos possessed duplicate first arch derivatives, such as two upper jaws, suggesting that cranial neural crest cells are prespecified prior to emigration. Studies by Paul Trainor and colleagues have shown that if the transplanted tissue excludes the mid-hindbrain isthmus, the transplanted neural crest cells show considerable plasticity. Because the isthmus was included in Noden's experiments, and because it expresses high levels of fibroblast growth factor 8 (fgf-8), it appears to act as a patterning center. Our current understanding is that cranial neural crest cells are not preprogrammed to a particular fate prior to emigration, nor is their fate determined solely by the environment into which they migrate but, rather, cranial neural crest cell fate depends on

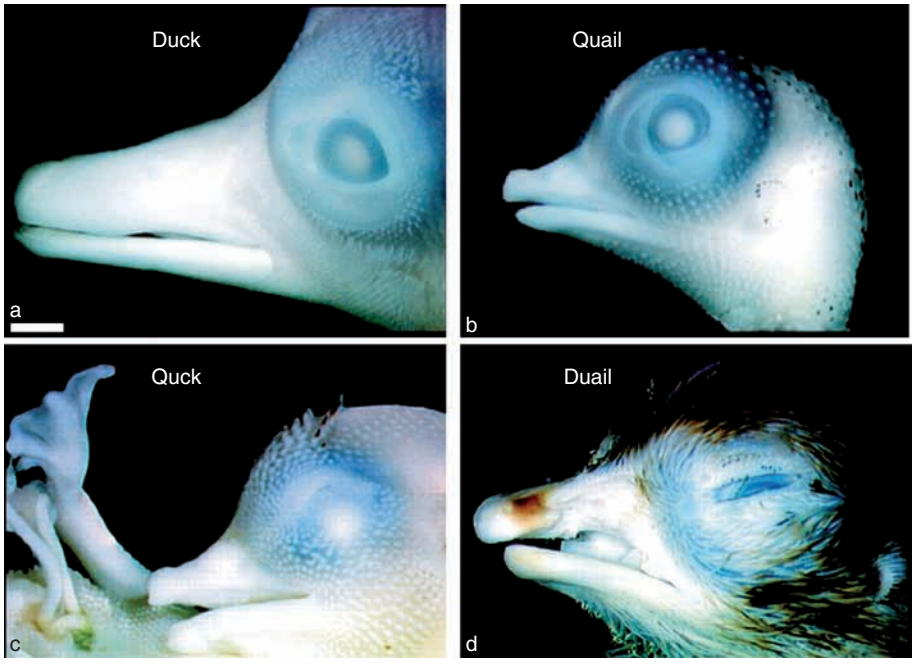

Figure 7 Differences in morphology of duck (a) and quail (b) beaks; the duck beak is flat and broad, whereas the quail beak is short and narrow. When neural crest cells of the presumptive beak region from duck embryos are removed and replaced with equivalent tissue from quail embryos, the 'quck' embryo has a beak that resembles a quail beak (c). Conversely, when neural crest cells of the presumptive beak region from ducks are transplanted into quail embryos ('duail'), the embryo has a beak that resembles a duck beak (d). Adapted from Schneider RA and Helms JA (2003) The cellular and molecular origins of beak morphology. *Science* 299: 565–568, from the American Association for the Advancement of Science.

a combination of signals from the hindbrain and from the branchial arch environment, as well as interactions between neural crest cells.

Vagal Neural Crest

Vagal neural crest cells are defined as neural crest cells that originate adjacent to somites 1–7 (Figure 3). The head–neck boundary corresponds approximately to somite 5, and thus vagal neural crest cells include the most caudal cranial neural crest cells and the most rostral trunk crest cells. Vagal neural crest cells give rise to two well-characterized derivatives – the cardiac outflow tracts and the enteric nervous system.

Cardiac Neural Crest

In the 1980s, using the quail–chick transplantation method described previously, Margaret Kirby showed that neural crest cells that emigrate adjacent to somites 1–3 contribute to the aorticopulmonary septation complex within the cardiac outflow tract of the heart and to smooth muscle (tunica media) of the aortic arch and its major branches; thus, these neural crest cells are known as the cardiac neural crest. Studies using a number of different strains of transgenic mice have largely confirmed the cardiac neural crest derivatives first identified in birds (Figure 8). However, some extra cardiovascular derivatives have also been reported in transgenic mice, including a neural crest contribution to the epicardium, but it is unclear whether these represent ectopic transgene expression or actual neural crest contribution to the cardiovascular system in mice.

A large number of genes and signaling pathways are now known to be involved in cardiac neural crest development in mice. The molecules include growth factors including fgf-8, transforming growth factor-β2, bone morphogenetic proteins, neurotrophin-3, Wnt1, and vascular endothelial growth factor; adhesion molecules including NCam, N-cadherin, and connexin-43; transcription factors including Pax3, Nkx2.6, FoxD3, Tbx1, Sox9 and Snail-1/Snail-2; and guidance cues including semaphorin 3C and notch signaling. Many of the aforementioned genes were identified from mouse models of congenital heart defects, but only a small number (e.g., Notch ligands and Tbx1) have been implicated in congenital heart anomalies in humans.

Enteric Nervous System

Within the wall of the gastrointestinal tract there is a vast network of neurons and glial cells called the enteric nervous system. Pioneering ablation experiments performed by Chester Yntema and Warner Hammond in the 1950s, and subsequent chick–quail transplantation studies performed by Nicole Le Douarin and colleagues in the 1970s, showed that all enteric neurons and glial cells arise from the neural crest, and the vast majority arise from vagal neural crest cells. Vagal neural crest cells migrate toward the gut along a pathway that is later followed by the vagus nerve (Figure 9). After

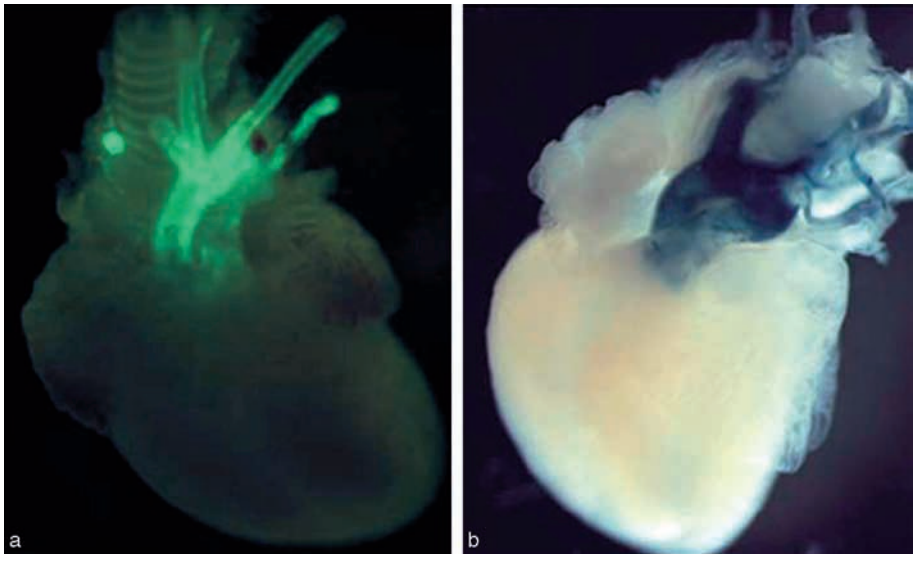

Figure 8 The heart and nearby blood vessels of transgenic mice in which cardiac neural crest cells express green fluorescent protein (GFP) (a) or the reporter gene, *lacZ*, and thus appear blue after histochemical processing (b). Cranial neural crest cells contribute to the aortic arch and great vessels. Reproduced from Stoller JZ and Epstein JA (2005) Cardiac neural crest. *Seminars in Cell and Developmental Biology* 16: 704–715, with permission from Elsevier.

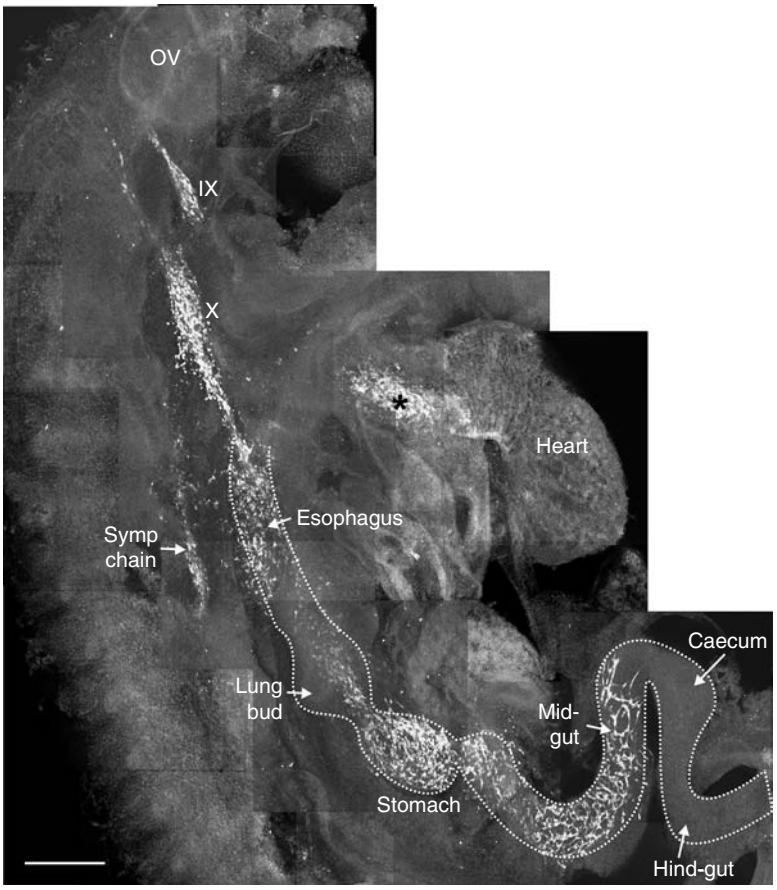

Figure 9 Lateral view of a 10.5-day-old transgenic mouse embryo in which cells expressing the receptor tyrosine kinase, Ret, also express green fluorescent protein (GFP). The epithelium had been removed to reveal the developing organs. Ret (and thus GFP) is expressed by vagal (including cardiac) and trunk neural crest cells. There is a stream of GFP+ cells that migrate from the hindbrain caudal to the otic vesicle (OV), adjacent to somites 1–7, into the gut along the pathway that will later be followed by the vagus nerve (X). At this developmental stage, GFP+ vagal neural crest cells have colonized the stomach and part of the midgut but not the caecum or hindgut; the neural crest cells within the gut will form the enteric nervous system. Vagal neural crest cells also give rise to ganglia of cranial nerves IX and X, and those arising from somites 1–3 contribute to the cardiac outflow tracts (asterisk). Trunk neural crest cells forming sympathetic chain ganglia (symp chain) also express GFP+. Reproduced from Anderson RB, Stewart AL, and Young HM (2006) The phenotypes of neural crest-derived cells in vagal and sacral pathways. *Cell Tissue Research* 323: 11–25, from Springer-Verlag.

entering the foregut, they migrate caudally through the gut mesenchyme to colonize the entire gastrointestinal tract.

A congenital disease called Hirschsprung's disease, or congenital megacolon, occurs in humans when neural crest cells fail to colonize the most distal regions of the gastrointestinal tract. Genetic studies in humans with Hirschsprung's disease and a variety of transgenic or mutant mice with defects in the enteric nervous system have been pivotal in identifying the genes required for the development of the enteric nervous system. Mutations in the genes encoding members of two signaling pathways account for approximately 50% of the cases of Hirschsprung's disease – these are glial cell line-derived neurotrophic factor acting through the receptor tyrosine kinase, Ret, and endothelin-3 acting through endothelin receptor B (EDNRB). Other signaling pathways involving neurotrophin-3, BMPs, hedgehog proteins, netrins/DCC (deleted in colorectal cancer), Slit/Robo, the cell adhesion molecule L1, and a number of transcription factors, including Sox8 and 10, Phox2b, Mash1, and Sip1, are also involved in the development of the enteric nervous system.

Trunk Neural Crest

Trunk neural crest cells give rise to dorsal root (sensory or spinal) ganglia, sympathetic postganglionic neurons, Schwann cells, adrenal chromaffin cells, and melanocytes. After emigrating from the neural tube, trunk neural crest cells follow two main pathways: neural crest cells destined to give rise to dorsal root ganglia, sympathetic ganglia, Schwann cells, and adrenal chromaffin cells follow a ventromedial pathway between the somites and the neural tube, within

the rostal half of each somite and in the intersomitic spaces, whereas melanocyte precursors follow a dorsolateral pathway, immediately under the ectoderm (Figure 1(c)). Neural crest cells that follow a ventral pathway emigrate from the neural tube prior to the cells that follow a dorsolateral pathway. Migration pathways are defined as a complex spatial array of guidance molecules, both attractive and repulsive. The substrates for migration include both extracellular matrix and cell surfaces.

Ventral Pathway

Early migrating neural crest cells follow a ventral pathway and appear to be prevented from entering a dorsolateral pathway by repulsive cues (Figure 10(a)). Ephrins and Slits are expressed in the dorsal regions of the developing somites (the dermamyotome), and their receptors – EphBs and Robos, respectively – are expressed by early migrating trunk neural crest cells; these guidance cues appear to repel early emigrating neural crest cells from the dorsolateral pathway. As they move ventrally, the crest cells diverge laterally due to repulsive signals from the medially placed notochord. These signals are mediated in part by the lectican (large chondroitin sulfate proteoglycan) aggrecan.

Neural crest cells migrate through the rostral halves of the somites and between somites but avoid the caudal half of each somite (Figure 10a). The development of segmental dorsal root ganglia is linked to the segmentation of the mesoderm. In particular, the selective migration of neural crest cells through the rostral half of each somite is essential for establishing the segmental organization of dorsal root ganglia; if the caudal halves of somites are replaced with rostral halves prior to the emigration of neural crest cells from the neural tube, a single continuous 'polyganglion' develops instead of segmental ganglia (Figure 11). A number of molecules are differentially expressed by the rostral and caudal halves of somites. Some molecules that were initially identified as repulsive axon guidance cues, including semaphorin 3A (Sema3A) and ephrin-B1, are expressed by the caudal, but not rostral, half of each somite, and the receptors for Sema3A and ephrin-B1 are expressed by migrating neural crest cells. *In vivo* and *in vitro* experiments have shown that Sema3A and ephrin-B1 are repulsive to migratory trunk neural crest cells, and thus trunk neural crest cells are channeled into the rostral half of each somite by the presence of repulsive cues in the caudal regions of the somites (Figure 10). Some ECM molecules such as F-spondin, tenascin-C, and lecticans that are selectively expressed in the caudal half of each somite are inhibitory to migrating crest cells and most likely contribute to the exclusion of neural crest cells from the caudal half of each somite.

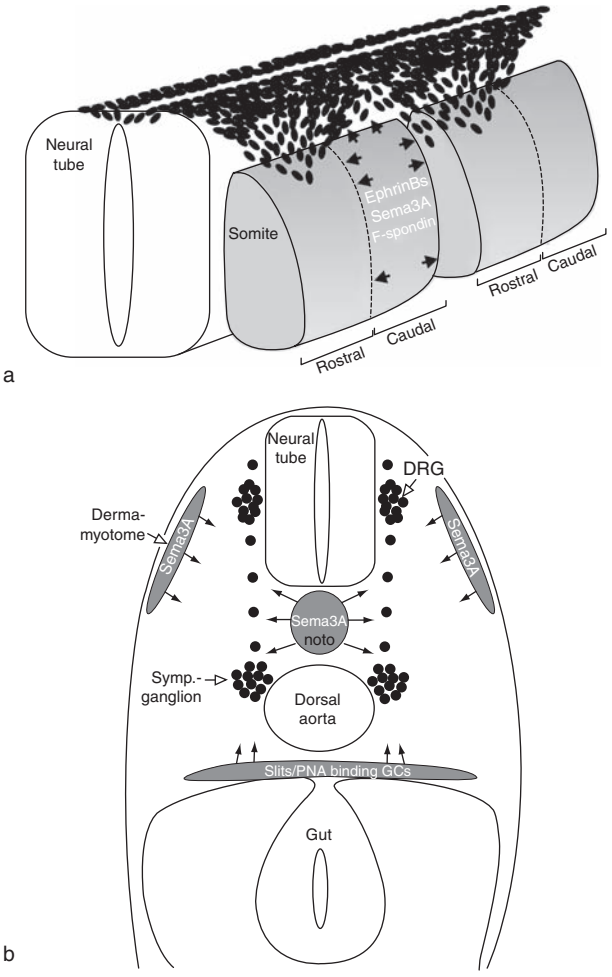

Figure 10 Pathway of trunk neural crest cells that migrate along a ventral pathway and the molecules that influence their pathways. Neural crest cells in the dorsolateral pathway are not shown. (a) Neural crest cells delaminate all along the neural tube and are then channeled into streams of cells through the rostral half of each somite by repulsive cues expressed exclusively in the caudal half of each somite; the repulsive cues include Ephrin-Bs, semaphorin 3A (Sema3A), and F-spondin. (b) The migration of cells to form sympathetic ganglia is influenced by repulsive cues in the dermamyotome and notochord and in tissue dorsal to the gut which prevents trunk neural crest cells from entering the gut. DRG, dorsal root ganglia. Reproduced from Young HM, Anderson RB, and Anderson CR (2004) Guidance cues involved in the development of the peripheral autonomic nervous system. *Autonomic Neuroscience* 112: 1–14, with permission from Elsevier.

Some trunk neural crest cells coalesce to form dorsal root ganglia adjacent to the neural tube, whereas others migrate further ventrally and coalesce into sympathetic ganglia adjacent to the dorsal aorta (Figure 10(b)). In fact, the first trunk neural crest cells that emigrate from the neural tube migrate directly to the dorsal aorta and form sympathetic ganglia, whereas dorsal root ganglia arises from crest cells

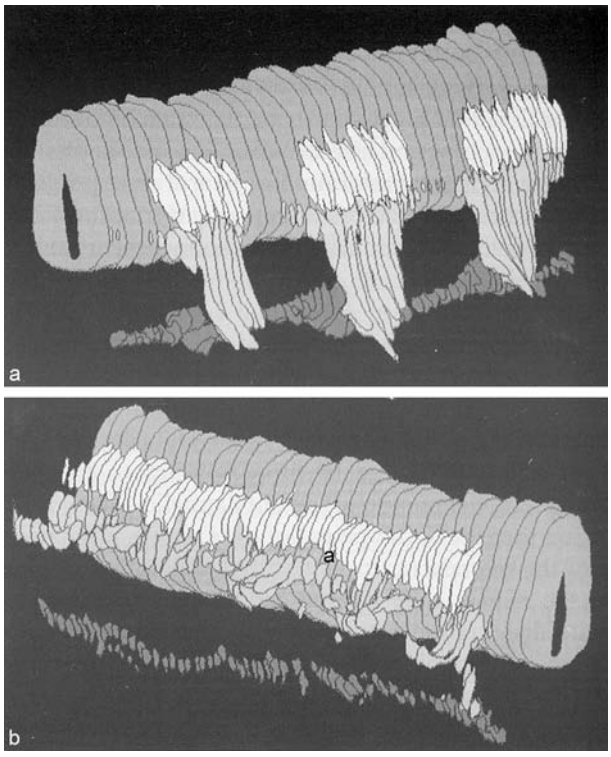

Figure 11 Three-dimensional reconstructions of the neural tube and dorsal root ganglia (DRG) from an embryo in which the caudal halves of each somite were replaced with rostral halves on one side of the embryo only. (a) On the control side of the embryo, DRG are segmental and the location of each ganglion corresponds to the rostral half of each somite. (b) On the experimental side, which contains only rostral halves of somites, there is a 'polyganglion' instead of individual segmental ganglia. Reproduced from Le Douarin NM and Kalcheim C (1999) *The Neural Crest*, 2nd edn., Cambridge, UK: Cambridge University Press.

that migrate slightly later. The general rule seems to be that for any locus of the neural crest, the first cells to migrate tend to disperse the farthest. Neuregulin, a growth factor whose effects are mediated by members of the ErbB family of receptor tyrosine kinases, is required for the migration of cells ventral to the neural tube to form sympathetic ganglia but not for the formation of dorsal root ganglia. Guidance cues expressed by the dermamyotome, notochord, and tissue dorsal to the gut are also thought to influence the migration of sympathetic ganglion precursors.

Dorsal root ganglion neurons and sympathetic postganglionic neurons have different phenotypes and use different neurotransmitters. Sympathetic postganglionic neurons are induced to express the appropriate phenotype by bone morphogenetic proteins produced by the dorsal aorta.

Dorsolateral Pathway

The last neural crest cells that emigrate from the neural tube follow a dorsolateral pathway, just beneath the ectoderm, and give rise to melanocytes (pigment cells) (Figure 1(c)). Emigration of neural crest cells along a dorsolateral pathway commences as migration along a ventromedial pathway is ceasing – late migrating neural crest cells seem to express different molecules from the early migrating cells that enable them to overcome the repulsive cues in the dermamyotomes. Melanocytes probably arise from all axial levels of the neural crest, but their development has been most extensively studied from trunk neural crest cells.

Two signaling pathways are known to be essential for the development of melanocytes. Endothelin-3 acting at EDNRB receptors on neural crest cells promotes the proliferation and migration, and delays the differentiation, of melanocyte precursors. Stem cell factor is produced by the dermamyotome and acts at Kit tyrosine kinase receptors on neural crest cells. This signaling pathway is required for the survival of melanocyte precursors until the late stages of differentiation in the skin, and it may direct their migration dorsolaterally. The transcription factor Mitf (microphthalmia-associated transcription factor) is also required for melanocyte development. Mitf has been shown to regulate genes required for pigmentation. However, since melanocyte precursors undergo apoptosis or adopt an alternative cell fate in the absence of Mitf protein, Mitf must also regulate genes required for the survival and proliferation of melanocyte precursors and the maintenance of melanocyte identity.

Sacral Neural Crest Cells

The location of the sacral region of the neural axis varies between species. Sacral neural crest cells are those that emigrate caudal to somite 28 in chick embryos and caudal to somite 24 in embryonic mice. Sacral neural crest cells give rise to neurons and glial cells in pelvic parasympathetic ganglia and to some enteric neurons, mostly in the hindgut (Figures 2 and 12). They are also thought to give rise to melanocytes and Schwann cells.

Plasticity of Trunk Neural Crest Cells

Like cranial neural crest cells, a large number of studies have examined whether trunk neural crest cells are predetermined or biased to a particular fate prior to emigration from the neural tube. Studies by Le Douarin and colleagues, in which premigratory neural crest cells from one axial level were transplanted to different axial levels, showed that at least some neural crest cells show considerable plasticity, and their fate was largely dependent on the environment. Experiments by Marianne Bronner-Fraser, Scott Fraser, and colleagues, in which individual premigratory neural crest cells were labeled, showed

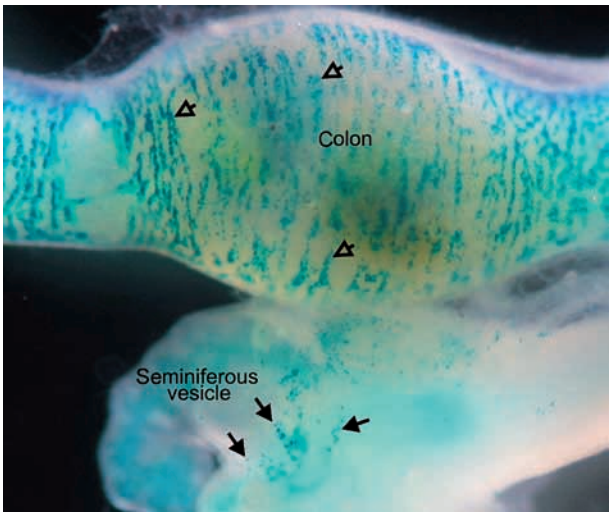

Figure 12 Colon and nearby reproductive organs from a 4-day-old transgenic mouse in which many vagal, trunk, and sacral neural crest cell derivatives express the reporter gene, *lacZ*, and so appear blue after histochemical processing. A dense network of stained neurons is present in the gut (open arrows) – most of these will have arisen from vagal neural crest cells, but some sacral crest cells also colonize the hindgut (**Figure 2(c)**). There are also stained cells (closed arrows) associated with the seminiferous vesicles – these are pelvic ganglion neuron precursors.

that individual premigratory neural crest cells could give rise to a variety of derivatives. For example, one labeled crest cell gave rise to melanocytes and Schwann cells as well as cells in dorsal root ganglia and the adrenal medulla. However, other labeled cells gave rise to only a single derivative. There is also evidence, particularly for melanocyte precursors, that some neural crest cells are prespecified, or at least biased to a particular fate, prior to emigrating from the neural tube. For example, melanocyte precursors express cell type-specific markers while still within the dorsal neural tube. Evidence suggests that some premigratory trunk neural crest cells are multipotent and their fate is probably determined by the environment through which they migrate or by tissue they colonize, whereas other premigratory trunk neural crest cells are biased or perhaps even prespecified prior to emigration. That the early migrating cohort includes multipotent cells and the late emigrating cohort comprises determined cells may indicate that restrictive lineage decisions are time regulated without regard to the morphogenetic decision of migration onset.

Conclusion

The neural crest is a transitory, migratory population of cells that gives rise to a remarkable diversity of cell types. It only occurs in vertebrates, and because of its role in craniofacial development, it is thought to have played a pivotal role in the evolution of vertebrates.

See also: Autonomic Neuroplasticity: Development; Autonomic Nervous System Development; Enteric Nervous System Development; Neural Crest Cell Diversification and Specification: Melanocytes; Neural Crest Cell Diversification and Specification: ErbB Role; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation.

Further Reading

Anderson DJ (2000) Genes, lineages and the neural crest: A speculative review. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 355: 953–964.

Anderson RB, Stewart AL, and Young HM (2006) The phenotypes of neural crest-derived cells in vagal and sacral pathways. *Cell Tissue Research* 323: 11–25.

Burns AJ and Douarin NM (1998) The sacral neural crest contributes neurons and glia to the post-umbilical gut: Spatiotemporal analysis of the development of the enteric nervous system. *Development* 125: 4335–4347.

Krull CE (2001) Segmental organization of neural crest migration. *Mechanisms of Development* 105: 37–45.

Kulesa PM and Fraser SE (1998) Neural crest cell dynamics revealed by time-lapse video microscopy of whole embryo chick explant cultures. *Developmental Biology* 204: 327–344.

Le Douarin NM and Kalcheim C (1999) *The Neural Crest*, 2nd edn. Cambridge, UK: Cambridge University Press.

Meulemans D and Bronner-Fraser M (2004) Gene-regulatory interactions in neural crest evolution and development. *Developmental Cell* 7: 291–299.

Newgreen D and Young HM (2002) Enteric nervous system: Development and developmental disturbances – Part 1. *Pediatric and Developmental Pathology* 5: 224–247.

Newgreen D and Young HM (2002) Enteric nervous system: Development and developmental disturbances – Part 2. *Pediatric and Developmental Pathology* 5: 329–349.

Schneider RA and Helms JA (2003) The cellular and molecular origins of beak morphology. *Science* 299: 565–568.

Stoller JZ and Epstein JA (2005) Cardiac neural crest. *Seminars in Cell and Developmental Biology* 16: 704–715.

Trainor PA, Ariza-McNaughton L, and Krumlauf R (2002) Role of the isthmus and FGFs in resolving the paradox of neural crest plasticity and prepatterning. *Science* 295: 1288–1291.

Trainor PA, Sobieszczuk D, Wilkinson D, and Krumlauf R (2002) Signalling between the hindbrain and paraxial tissues dictates neural crest migration pathways. *Development* 129: 433–442.

Young HM, Anderson RB, and Anderson CR (2004) Guidance cues involved in the development of the peripheral autonomic nervous system. *Autonomic Neuroscience* 112: 1–14.

Neural Crest Cell Diversification and Specification: ErbB Role

K-A Nave and M H Schwab, Max Planck Institute of Experimental Medicine, Göttingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

The peripheral nervous system (PNS) mainly derives from the neural crest and develops in a process of highly organized migration of pluripotent precursor cells that become progressively committed to fully differentiated cell types, including neurons and glia cells. The main glial cell type in the PNS are two types of Schwann cells (SCs) that serve distinct functions. Remak (or nonmyelinating) SCs engulf several small caliber C-fiber axons without enwrapping them. This ensheathment is nevertheless required for long-term axonal integrity. Myelinating SCs form compact myelin sheaths around single axons of large-caliber motor and sensory neurons, an essential prerequisite for rapid and precise impulse conduction. Myelination critically depends on reciprocal molecular interactions between neurons and glia cells. In the PNS, signaling of a neuronal growth factor, Neuregulin-1 (NRG1), to glial receptor tyrosine kinases of the ErbB family has emerged as a central regulator of SC development and myelination. This is in addition to other presumed roles of NRG1–ErbB signaling in the mammalian nervous system, such as neuronal migration, synaptogenesis, and synaptic plasticity.

Components of the NRG–ErbB Signaling System

Neuregulins (NRGs) are a family of membrane-associated growth factors with an extracellular epidermal growth factor (EGF)-like signaling domain. The term neuregulin derives from neu differentiation factor (NDF) and heregulin, the first cloned ligands for the oncogene ErbB2 (a.k.a., neu and HER2). Four genes (*Nrg1–Nrg4*) are present in mammals, but only *Nrg1* has been studied in detail. Due to alternative promoter usage and mRNA splicing, multiple (>16) NRG1 isoforms are produced that are subgrouped by their distinct N-terminal domains (Figure 1(a)). NRG1 type I (also known as heregulin, NDF, or acetylcholine receptor-inducing activity) and NRG1 type II (also known as glial growth factor (GGF)) have N-terminal immunoglobulin-like domains. Transmembrane forms of NRG1 undergo proteolytic cleavage by metalloproteinases, including tumor necrosis factor-α converting enzyme and BACE1 (β-site amyloid precursor protein–cleaving enzyme 1), a type I transmembrane aspartyl protease. As a consequence, NRG1 types I and II are shed from the neuronal cell surface and act as paracrine signaling molecules. NRG1 type III is defined by its cysteine-rich domain, which functions as a second transmembrane domain. Consequently, type III is thought to remain tethered to the cell surface after cleavage and to act as a juxtacrine signal (Figure 1(a)). In addition, exons encoding shorter amino termini of NRG1 have been identified (referred to as types IV–VI), but these isoforms have not been further characterized.

The EGF-like domain (contained in all NRG1 isoforms) is necessary and sufficient for binding and activation of transmembrane receptor tyrosine kinases of the ErbB family. The term ErbB originally derived from its homology to the erythroblastoma viral gene product, v-erbB. The basic structure of all four members of the ErbB receptor family (ErbB1–4) includes two extracellular cysteine-rich ligand binding domains, a transmembrane region, the tyrosine kinase domain (inactive in ErbB3), and a C-terminal domain (Figure 1(b)). NRG1 directly binds to ErbB3 and ErbB4. ErbB2, which is devoid of an activating ligand binding site, is the preferred dimerization partner. ErbB receptors dimerize not by virtue of a bridging effect of NRG1 but, rather, following a ligand-activated conformational change in the ectodomain of ErbB3 or ErbB4. Crystallographic data indicate that ErbB2 constitutively exposes a dimerization loop and can form heterodimers with ligand-activated ErbB3 or ErbB4 receptors. NRG1 binding induces homodimer (ErbB4/ErbB4) or heterodimer formation (ErbB2/ErbB3, ErbB2/ErbB4, and ErbB3/ErbB4), which leads to receptor cross-phosphorylation, recruitment of SH3-containing adapter molecules, and activation of downstream signaling pathways. Complexity of NRG1–ErbB signaling emerges from isoform-specific receptor affinity and dimer-dependent stimulation of signaling pathways. Moreover, several observations suggest that NRG1–ErbB signaling is bidirectional. Upon binding to an ErbB receptor and/or following neuronal depolarization, the C-terminal intracellular domain (ICD) of NRG1 is proteolytically cleaved off in a γ-secretase-dependent process and subsequently translocated to the cell nucleus. Here, depending on cell type and interactions with other transcription factors, the ICD might act as an enhancer or repressor of gene expression, resulting, for instance, in increased neuronal survival or enhanced synaptic plasticity. ErbB receptors have also been identified to be translocated to the nucleus, either the entire protein or its cytoplasmic domain. In particular, γ-secretase-dependent proteolysis (after NRG1 binding) and nuclear translocation have been observed for the cytoplasmic domain of

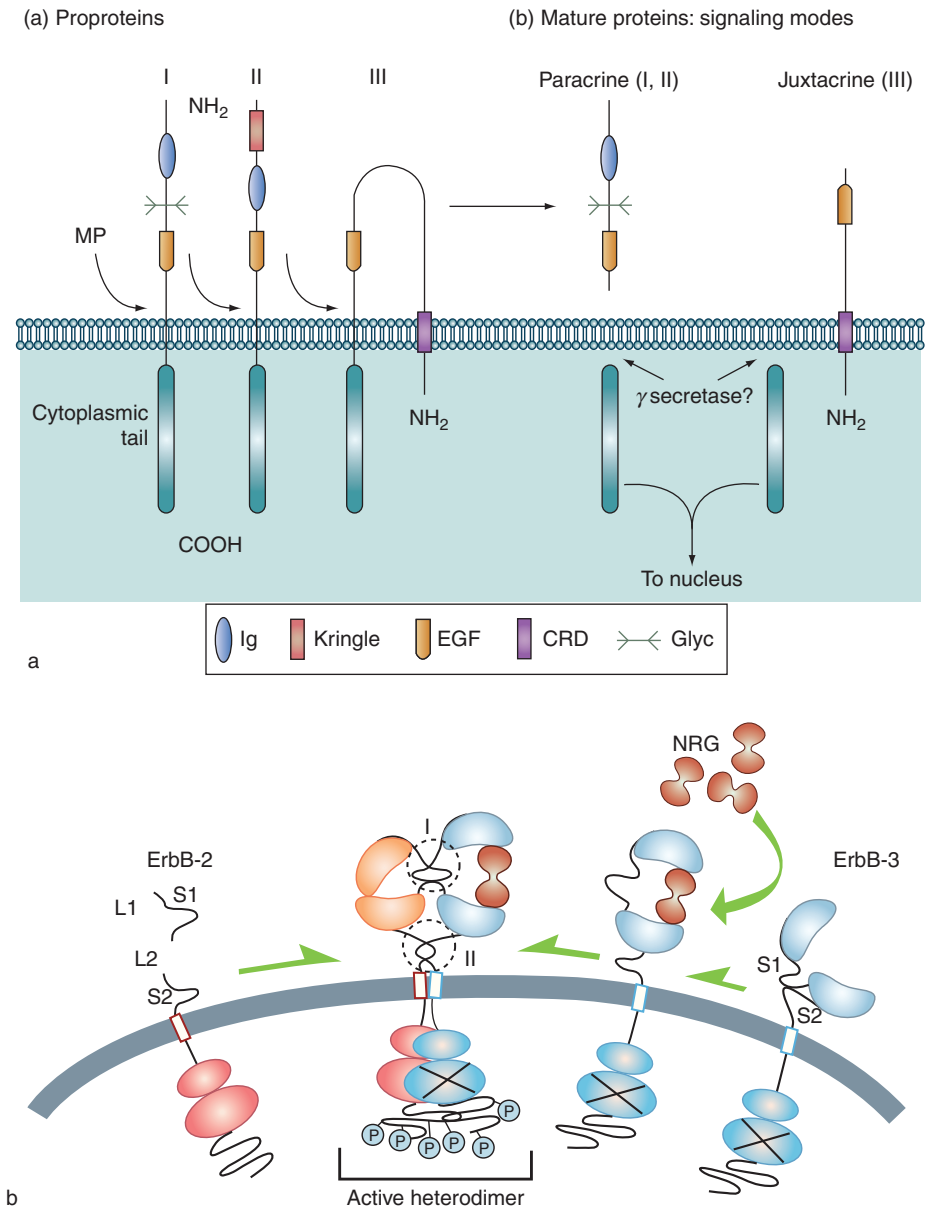

Figure 1 Structure of NRG1/ErbB signaling components. (a) NRG1 types I and II are synthesized as single pass transmembrane proteins; type III has two transmembrane domains. With metalloproteinase (MP) cleavage, types I and II are shed as paracrine signals, type III remains tethered through its cysteine-rich domain (CRD) and is a juxtacrine signal. The cytoplasmic domain might undergo further cleavage stimulated by binding of ErbB receptors to NRG1, followed by translocation to the nucleus. (b) Proposed mechanism of NRG1-induced ErbB2/ErbB3 heterodimer activation. Extracellular ErbB receptor domains are represented by two CRDs (S1 and S2) and two cysteine-free, ligand-binding domains (L1 and L2). ErbB3 (blue) lacks a functional intracellular tyrosine kinase domain and exists on the cell surface in an autoinhibited conformation (resulting from the interactions between the S1 and S2 domains). Bivalent binding of NRG1 (NRG, red dumbbell) to the L1 and L2 domains of ErbB3 rearranges the conformation of the extracellular domain, leading to heterodimer formation with ErbB2, which is devoid of a functional ligand binding site. Dimerization between a ligand-bound ErbB3 and an ErbB2 molecule results in receptor auto-phosphorylation and activation of downstream signaling pathways. (a) Reproduced from Nave KA and Salzer JL (2006) Axonal regulation of myelination by neuregulin1. *Current Opinion in Neurobiology* 16: 492–500. (b) Reproduced from Citri A and Yarden Y (2006) EGF–ERBB signalling: Towards the systems level. *Nature Reviews Molecular Cell Biology* 7: 506–516.

ErbB4, suggesting that both ErbB receptors and their ligands can function as regulators of gene expression.

NRG1–ErbB Signaling Regulates SC Lineage Development

In rodents, development of SCs from pluripotent neural crest cells progresses through two intermediate cell types, SC precursors (SCPs) and immature SCs, which proliferate and migrate along preexisting axons. Early migratory neural crest cells and mesenchymal cells express NRG1, ErbB2, and ErbB3, but not ErbB4. Before embryonic day (E) 10, only type I isoforms of NRG1 are expressed. Later, NRG1 type III becomes the predominant isoform in spinal cord motor neurons and

sensory neurons of dorsal root ganglia (DRG). In contrast, ErbB2 and ErbB3 are strongly expressed in SCPs and immature SCs but are mostly absent from sensory and motor neurons, at least during embryogenesis.

Multiple studies in which a recombinant EGF domain (derived from NRG1) was added to cultured SCs or injected into rodent sciatic nerves have demonstrated that ErbB signaling operates during all stages of SC development and has potent effects on SC survival, proliferation, migration, differentiation, and myelination.

The essential requirement of NRG1–ErbB signaling for the development of the SC lineage *in vivo* was demonstrated through gene targeting in mice. Mouse mutants lacking NRG1, ErbB2, or ErbB4 die at midgestation (E10.5) due to heart malformation. Most ErbB3 null mutants die between E11.5 and E13.5 (also due to defective heart development), but some embryos develop to term. In NRG1, ErbB2, and ErbB3 (but not ErbB4) null mutants, the number of SCPs that line peripheral axons is severely diminished already at E10.5, as evidenced by the reduced expression of several SC markers (e.g., ErbB3, p75, Sox10, and S100). The severity of SCP loss, however, depends on the axial level. Likewise, glial cells in the enteric nervous system are absent in the mutants. Analysis of ErbB3-deficient mice at later embryonic stages revealed not only loss of SCPs along peripheral nerves but also a reduction of up to 80% of motor and sensory neurons, again depending on the axial level. Neuronal death is likely to be a secondary effect and may be attributed to the absence of SC-derived neurotrophic factors (e.g., nerve growth factor, ciliary neurotrophic factor, or glial cell line-derived neurotrophic factor) that support neuronal survival.

The importance of ErbB2 signaling at later stages of prenatal SC development became apparent through transgenic rescue of the heart defect. Rescued ErbB2 mutants die at birth and completely lack SCPs and immature SCs in peripheral nerves at this time. In contrast, SCPs are present, albeit at reduced numbers, in DRG (at E12). This observation was attributed to the compromised ability of mutant SCPs to migrate out of the DRG and into peripheral nerves. However, the true lineage relationship between SCs in dorsal and ventral nerve roots, DRG, and peripheral nerves is not fully understood. Thus, the presence of some SCs in DRG of ErbB2 (and ErbB3) mutants could also reflect a differential dependency on ErbB2 signaling of certain SCP subpopulations, such as satellite SCs in DRG.

Do different NRG1 isoforms serve distinct functions during PNS development? Several lines of evidence strongly suggest that signaling of a membrane-tethered type III isoform of NRG1 is most critical for SC maturation (and myelination). Mouse mutants selectively lacking NRG1 isoforms with an Ig-like domain (affecting types I and II) die at E10.5 and display abnormalities in heart and cranial ganglia, which are very similar to those observed in pan NRG1 mutants. In contrast, SCPs that line spinal nerves are unaffected. The generation of mice that specifically lack NRG1 type III variants allowed researchers to directly address the role of this isoform in SC development. NRG1 type III mutants die at birth due to an inability to breathe. Interestingly, the consequences of type III-restricted gene targeting on SC development appear milder than those in ErbB2 or ErbB3 mutants. SCPs (defined by ErbB3 expression) are present along peripheral nerves already at E11.5, albeit at reduced numbers. At E14.5, immature SCs (Sox10/ErbB3/S100+), although dramatically reduced, are still present close to dorsal and ventral roots and along peripheral nerves. Nevertheless, S100+ cells are completely absent from nerve terminals. DRG explants from type III mutants prepared at E14.5 display a marked SC deficiency at first but are partially repopulated by endogenous SCs during the course of several weeks. However, many of these SCs fail to attach and ensheathe neurites properly. Thus, NRG1 type III isoforms are essential for the peripheral migration and/or survival of SCs, but signaling by other NRG1 isoforms is sufficient to also allow the initial generation and early differentiation of SCs.

Taken together, neuronal NRG1 type III signaling to SCs is mediated through an ErbB2/3 heterodimer and critically important for multiple aspects of SCP and premature SC development, including migration, differentiation, and survival.

NRG1–ErbB Signaling as a Master Regulator of Myelination in the Peripheral Nervous System

Immature SCs, which are generated from SCPs (between E13 and E15 in late embryonic nerves), first communally ensheathe large groups of axons. Subsequently, by radial sorting, individual SCs associate either with large-diameter axons in a 1:1 ratio or with several small-diameter (<1 μm) axons. During this process, the emergence of SC polarity, reflected by the asymmetric distribution of molecules (e.g., the polarity protein Par-3), is a prerequisite for focal deposition of myelin components. Early postnatally, SCs engulfing a large-diameter axon are selectively activated to differentiate into myelinating SCs that wrap multiple layers of myelin membrane around this axon. In contrast, SCs associated with small-diameter axons mature into 'nonmyelinating' SCs that ensheathe 5–25 (up to 50) small-caliber axons (in a so-called Remak bundle) without forming compact myelin sheaths. Since SC proliferation and survival both depend on axonal signals, limited axonal contact provides a

means by which the number of SCs and axons are matched.

In contrast to the control of myelination, the role of NRG1–ErbB signaling in radial sorting, SC polarity, timing of myelination onset, and the survival of SCs in postnatal nerves has not been determined *in vivo*. Several studies, however, support involvement of NRG1 in some of these processes. For instance, rescue of axotomy-induced SC death by exogenous NRG1 suggests that it is a component of an axon-derived survival signal at perinatal stages. The role of NRG1 in timing the onset of myelination through a phosphatidylinositol 3-kinase (PI3K)-dependent pathway is supported by the observation that treatment of rat SC cultures with recombinant NRG1 (similar to axonal contact) activates PI3K and mitogen-activated protein kinase (MAPK). In early SC–DRG neuron cocultures, elongation of SCs and subsequent myelination can be blocked by inhibition of PI3K (but not MAPK), whereas PI3K stimulation promotes myelination. In contrast, the maintenance of myelin sheaths in mature cocultures does not critically depend on PI3K activity. In agreement with this observation, pharmacologic inhibitors of ErbB receptors demonstrate a requirement of NRG1 signaling for the initiation of myelination also in zebra fish.

The first demonstration that NRG1–ErbB signaling plays a direct role in myelination came from the analysis of mouse mutants with a conditional deletion of the *ErbB2* gene in myelinating SCs. ErbB2 deficiency causes abnormally thin myelin sheaths, containing fewer myelin wraps. Similarly, mice that overexpress a dominant negative ErbB receptor in SCs (under control of the CNPase promoter) are hypomyelinated.

The level of NRG1 type III on the axonal surface, rather than axon diameter *per se*, has been identified to act as an instructive signal to induce myelination *in vitro*. Expression of NRG1 type III on the axon correlates with the ensheathment fate of axons *in vivo*: Unmyelinated, autonomic neurons express low levels of NRG1 type III on the axon surface, whereas BDNF/NT3-dependent DRG neurons, whose axons are heavily myelinated, express high levels. DRG axons from NRG1 type III null mice are not myelinated by SCs in cocultures and do not induce myelin-specific structural proteins or transcription factors, thus demonstrating that axonal NRG1 type III is essential for myelination (**Figure 2(A)**). Furthermore, mice haploinsufficient for NRG1 type III have a significantly higher proportion of axons that are persistently unmyelinated. In contrast, forced expression of NRG1 type III in the postganglionic fibers of sympathetic neurons converts these normally unmyelinated fibers to myelination *in vitro* (**Figure 2(B)**). Importantly, and despite addition of a soluble NRG1 type III ectodomain, cocultures of DRG neurons from type III null mutants and wild-type SCs fail to myelinate. Also, juxtacrine signaling (from heterologous cells stably expressing NRG1 type III) is not sufficient to induce myelin protein expression in cocultured SCs, demonstrating that NRG1 type III requires additional signals present on the axon to induce SC myelination. Together, these results suggest that threshold levels of axonal NRG1 type III provide the long-sought instructive signal that triggers SC myelination, at least *in vitro* (**Figure 3**).

Finally, NRG1–ErbB signaling has also been recruited for the regulation of the final stage of SC differentiation: the quantitative control of myelin sheath thickness (**Figures 2(C) and 2(D)**). Rapid impulse propagation is a function of axon caliber and myelination, and the optimal myelin thickness is reached when the 'g-ratio' is close to 0.68 (i.e., the numeric ratio between the diameter of the axon cylinder and that of the myelinated axon), a ratio remarkably well maintained for peripheral axons. Thus, axon size is monitored by myelinating SCs in order to assemble the correct number of myelin wraps. Mouse mutants with altered expression levels of NRG1 demonstrated that this information is encoded, at least in part, by the amount of membrane-associated NRG1 type III on the axon surface. Heterozygous NRG1 type III null mutant mice, which display approximately 50% NRG1, are hypomyelinated and exhibit reduced nerve conduction velocity. In contrast, transgenic mice that overexpress NRG1 type III in DRG sensory neurons and spinal cord motor neurons (under control of the neuronal Thy1 promoter) become hypermyelinated (**Figures 2(C) and 2(D)**). This effect appears to be specific to NRG1 type III because transgenic mice that overexpress the secreted NRG1 type I are not hypermyelinated. NRG1 has emerged as the rate-limiting factor of myelin growth control but not SC survival since SC numbers are unaltered in NRG1 +/− sciatic nerves. ErbB2 and ErbB3 are expressed by SCs at saturating levels. Only a dramatic reduction in number disrupts myelination (i.e., in a conditional mouse mutant with an SC-specific null mutation of the ErbB2 gene).

Whereas the requirement of the type III isoform for myelination is well established, the role of posttranslational modifications of NRG1 during myelination is only beginning to be addressed. First insight into the importance of proteolytic processing came from mouse mutants expressing only 'noncleavable' NRG1 variants. These mice display a phenotype analogous to NRG1 null mice (impaired heart development),

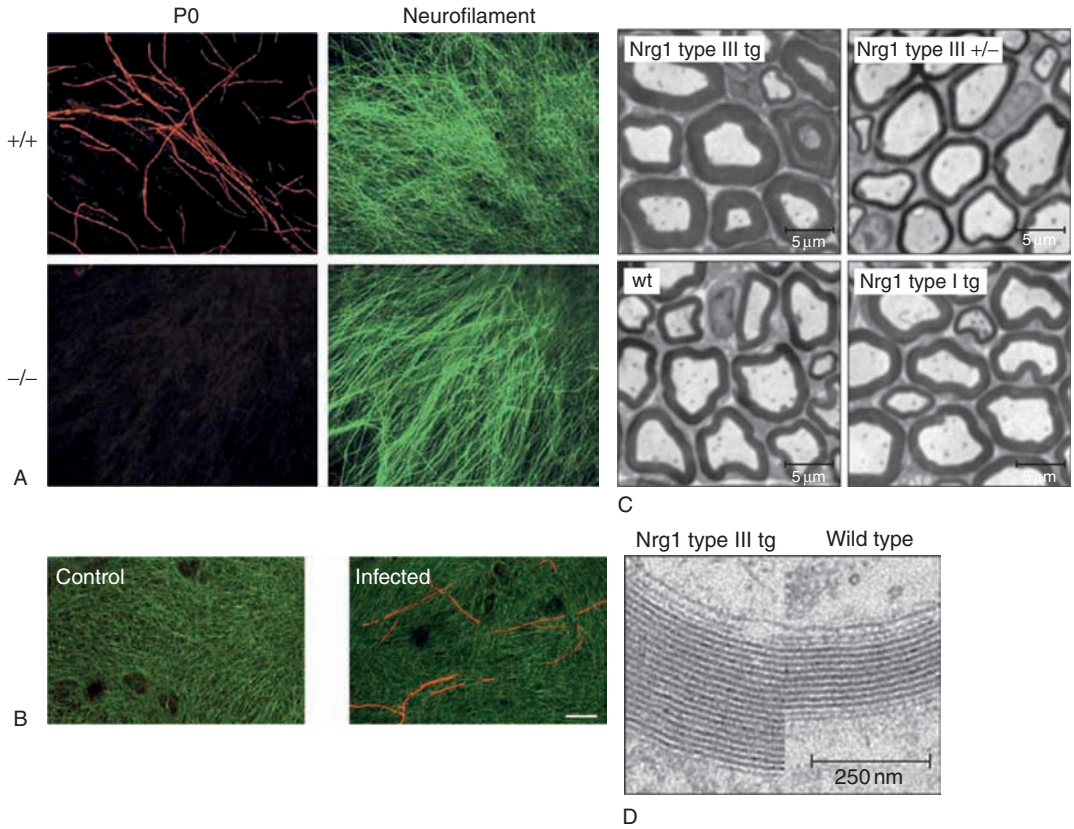

Figure 2 NRG1 type III is a master regulator of peripheral myelination. (A) Myelination by Schwann cells is completely blocked in the absence of NRG1 type III. Co-cultures of wild-type rat Schwann cells and sensory neurons from wild-type mice (+/+) or NRG1 type III null mutants (–/–) were maintained *in vitro* for 3 weeks, fixed, and stained for myelin (protein P0, red) and axons (neurofilament, green). Numerous myelin segments are evident in wild-type co-cultures, in contrast, no myelin is formed in co-cultures containing sensory neurons derived from NRG1 type III mutants. (B) Ectopic expression of NRG1 type III induces myelination of normally unmyelinated sympathetic neurons. Co-cultures of Schwann cells and control (left) or superior servical ganglion (SCG) neurons infected with NRG1 type III lentivirus (right) were stained for MBP (red) and neurofilament (green). (C) The level of axonal NRG1 type III determines myelin sheath thickness. Semithin cross sections (0.5 μm) of the sciatic nerve comparing transgenic NRG1 overexpressing mice with a wild-type control at age 5 months. Type III transgenics are hypermyelinated (Nrg1 type III tg), but not the type I transgenics (Nrg1 type I tg). The specific function of NRG1 type III is confirmed by a corresponding hypomyelination of mice with a type III isoform-specific null allele (Nrg1 type III +/–). (D) The ultrastructure and periodicity of compact myelin in hypermyelinated fibers, prepared from NRG1 type III transgenic mice (left), is indistinguishable from myelin prepared from wild-type mice (right). This demonstrates that hypermyelination is a consequence of additional membrane wraps. The two axons shown are of about equal diameter. Scale bar = 100 μm (B). (A, B) Reproduced from Tavegia C, Zanazze G, petrylak A, et al. (2005) Neuregulin-1 type III determines the ensheathment fate of axons. *Neuron* 47: 681–694. (C, D) Reproduced from Michailov GV, Sereda MW, et al. (2004) Axonal neuregulin-1 regulates myelin sheath thickness. *Science* 304: 700–703.

indicating that cleavage of NRG1 is critical for its function. Studies have identified BACE1 as essential for normal peripheral (and central) myelination. BACE1 is coexpressed with NRG1 in motor and sensory neurons at postnatal stages. BACE1 deficiency in mouse mutants results in the accumulation of unprocessed NRG1, hypomyelination of peripheral nerves, and aberrant segregation of small-diameter axons by SCs. The exact number and sequence of proteolytic cuts, the participation of other proteases, and the mechanisms that regulate proteolysis remain to be determined.

Taken together, the published studies support a model in which a threshold level of NRG1 type III is required to trigger myelination. At higher levels, the amount of myelin formed is a function of the amount of NRG1 type III presented by the axon to the SCs. This model also suggests a mechanism by which axons, via differing levels of NRG1 type III, coordinate SC numbers as they adopt alternative phenotypes. For example, axons expressing higher levels of NRG1 type III generate the additional SCs required to establish the one-to-one relationship characteristic of myelinated fibers. Open questions include the regulation of

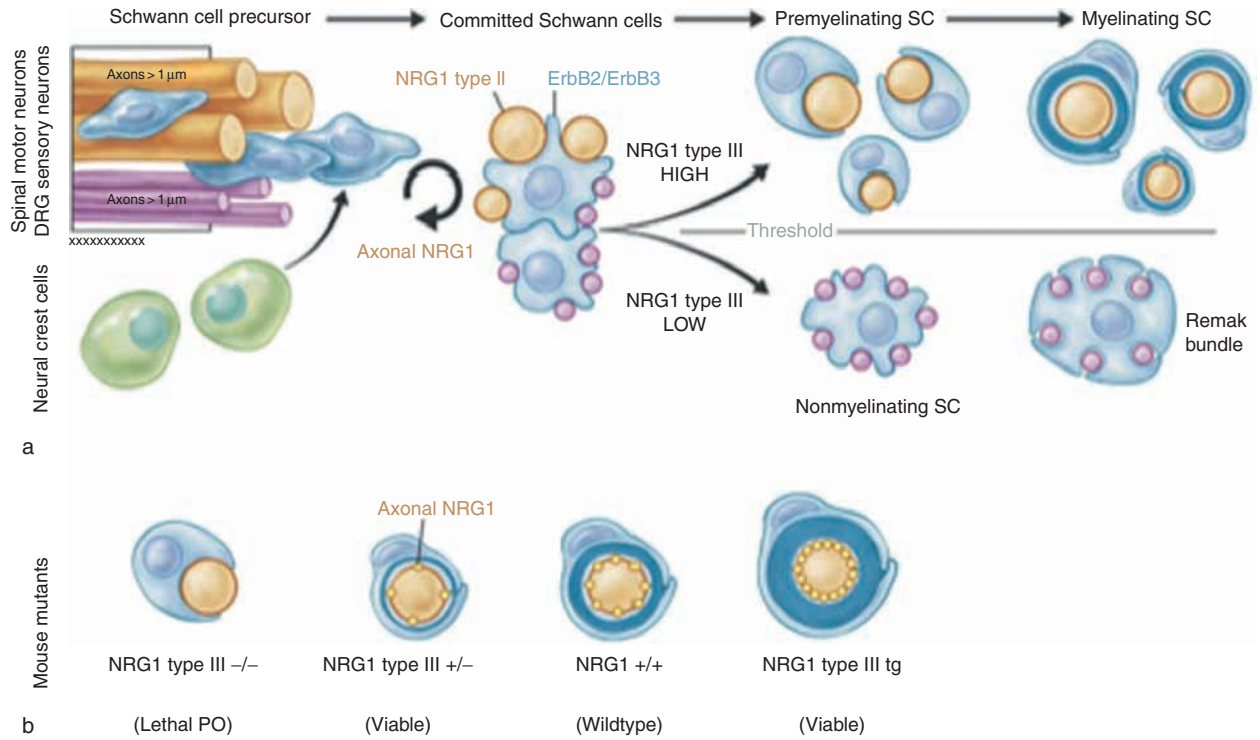

Figure 3 Axonal NRG1 regulates successive steps of Schwann cell differentiation. (a) Schwann cells (in blue) arise from neural crest precursor cells (in green) and interact with both large and small caliber axons of spinal motor and sensory neurons. During embryogenesis, NRG1 on the axon regulates Schwann cell development by activating ErbB signaling cascades, thereby promoting Schwann cell differentiation and expansion. The amount of NRG1 type III on the axon detected by committed Schwann cells, which is a function of axon size and NRG1 levels, then drives them either into segregating single axons and myelination (top), or into a nonmyelinating phenotype and formation of a Remak bundle (bottom). Above threshold levels, NRG1 type III signals axon size to Schwann cells to optimize myelin sheath thickness. (b) In mouse mutants lacking NRG1 type III (–/–), heterozygous NRG1 type III mutants (+/–) and transgenic NRG1 type III overexpressing mice, the amount of myelin made by Schwann cells varies directly as a function of axonal NRG1 type III levels (indicated by yellow dots) rather than as a function of axon diameter. Reproduced from Nave KA and Salzer JL (2006) Axonal regulation of Mylination by neuregulin 1. *Current Opinion in Neurobiology* 16: 492–500.

NRG1 expression in individual neurons to match NRG1 steady-state levels to axon caliber and length.

ErbB Signaling in the Mature Peripheral Nervous System and in Pathological Conditions

Adult sensory (DRG) and spinal motor neurons continue to express NRG1 type III. Likewise, expression of ErbB2 and ErbB3 is maintained in mature SCs, albeit at reduced levels. Moreover, differentiated SCs, at least in culture, express several NRG1 isoforms, and autocrine NRG1 signaling in SCs may explain why SCs become progressively less dependent on neuronal NRG1 signaling during development. Indeed, expression of ErbB receptors has also been reported in adult motor and sensory neurons. Thus, NRG1-expressing SCs might signal to neurons that they myelinate. Moreover, ErbB2 expression is induced in SCs upon sciatic nerve transection in adult rats. SC proliferation in Wallerian degeneration follows a fixed time course, and the expression of NRG1 and ErbB2 is coordinately induced, implicating this pathway in SC mitogenesis and remyelination after axotomy.

Surprisingly, conditional inactivation of ErbB2 in adult myelinating SCs (induced at 2.5 months of age) does not impair normal myelin sheath maintenance or the proliferation of SCs after nerve axotomy. Whether remyelination is unaffected in these mutants awaits further analysis. Thus, the molecular control of SC function differs significantly between development and regeneration.

In contrast to myelinated motor axons, sustained ErbB signaling is required for nonmyelinating SCs and small axons (C-fibers) in the mature nerve. Disruption of ErbB signaling in adult nonmyelinating SCs by expression of a dominant negative ErbB receptor causes cell death and proliferation of nonmyelinating SCs and, indirectly, loss of unmyelinated axons in transgenic mice.

Aberrant ErbB receptor activity contributes to the development of many human cancers. Receptor

overexpression, kinase domain mutations, and autocrine ligand production can cause ErbB activation in tumors. Since SCs express various forms of NRG1 and autocrine NRG1–ErbB signaling has been implicated in their proliferation, it is plausible that NRG1/ErbB2 may be involved in the pathogenesis of human nerve sheath tumors and Schwannomas. This is also supported by the development of nerve sheath tumors in transgenic mice that overexpress a soluble NRG1 type II variant (GGFβ3) in myelinating SCs. Overexpression of ErbB2 and ErbB3 by SCs has been noted in nerves from patients with Charcot–Marie–Tooth disease type 1A, the most common hereditary peripheral neuropathy. The upregulation of these receptors may correlate with the generation of supernumerary SCs ('onion bulbs') or play a role in the inhibition of myelination or demyelination.

Aberrant signaling through the ErbB2 receptor is also a cause of demyelination in leprosy. *Mycobacterium leprae*, the causative agent of leprosy, directly binds and activates ErbB2 on the surface of myelinating SCs without the usual dimerization with ErbB3. The subsequent activation of MAPK1/2 by phosphorylation results in demyelination. This process can be blocked at the receptor activation stage by Herceptin, a monoclonal antibody directed against the ectodomain of ErbB2 (Figure 4).

Taken together, these studies have demonstrated that ligand-independent ErbB activation does not necessarily recapitulate NRG1 signaling but, rather, orchestrates a distinct spectrum of potentially pathological cellular responses.

Future Research

Although principal mechanisms of myelination control by NRG1–ErbB signaling have been defined, a number of gaps remain in our understanding of the molecular details. For example, the relevance of intracellular processing, transport, and localization for isoform-specific functions (e.g., myelination versus neuromuscular junction formation) remains poorly defined, and the exact composition and localization along the axon of the core NRG1 type III–ErbB2/3

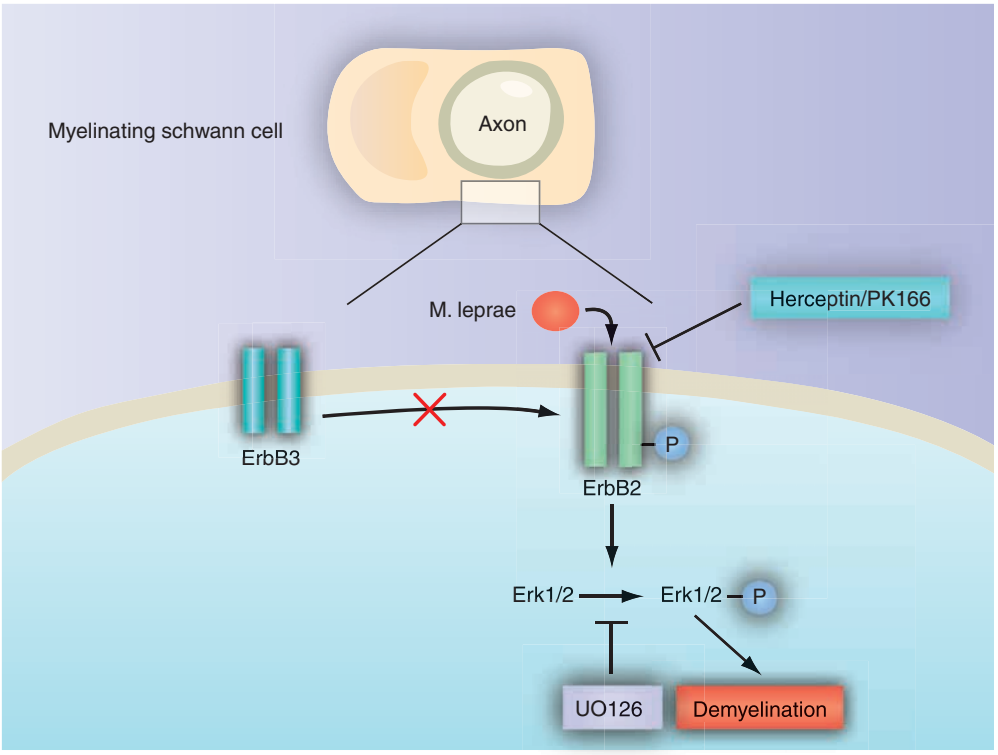

Figure 4 Leprosy induced peripheral demyelination is mediated by ErbB2, *Mycobacterium leprae,* the causative agent of leprosy, activates ErbB2 receptors on the surface of myelinating Schwann cells. This leads to their activation by phosphorylation (P) without the usual dimerization with ErbB3. The subsequent activation of the extracellular signal-regulated kinases (Erk) 1/2 by phosphorylation, a mitogen-activated protein kinase kinase (MAPKK)-dependent event, results in demyelination. The process can be blocked at the receptor activation stage by Herceptin (a monoclonal antibody directed against ErbB2) or at the Erk1/2 phosphorylation stage using a MAPKK inhibitor (UO126). Reproduced from Franklin RJ and Zhad C (2006) Tyrosine kinases: Maining in leprosy. *Nature Medicine* 12: 889–890.

signaling complex are unknown. It will be of particular interest to determine whether the C-terminal fragment of NRG1, after proteolytic cleavage by BACE1 (or other proteases), remains associated with the N-terminal EGFL domain-containing fragment or whether, upon γ-secretase cleavage, it mediates back-signaling by nuclear translocation. The N-terminal fragment might be further proteolytically processed such that the EGFL domain is released into the adaxonal space. Because the activity of the core NRG1 type III–ErbB2/3 signaling complex is most likely modulated by multiple interactions with other molecules on the axonal and glial surface and the extracellular matrix, the identification of interacting partner molecules is another important task. Finally, the downstream signaling pathways that transmit information provided by the activated ErbB2/3 heterodimer from the growing myelin membrane to the SC nucleus and the set of target genes which is transcriptionally regulated by those pathways have yet to be identified.

See also: Drosophila Apterous Neurons: from Stem Cell to Unique Neuron; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Neural Crest; Neural Crest Cell Diversification and Specification: Melanocytes; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation; Schwann Cells and Axon Relationship; Schwann Cell Morphology.

Further Reading

Citri A and Yarden Y (2006) EGF–ERBB signalling: Towards the systems level. *Nature Reviews Molecular Cell Biology* 7: 505–516.

Esper RM, Pankonin MS, and Loeb JA (2006) Neuregulins: Versatile growth and differentiation factors in nervous system development and human disease. *Brain Research Reviews* 51: 161–175.

Falls DL (2003) Neuregulins: Functions, forms, and signaling strategies. *Experimental Cell Research* 284: 14–30.

Franklin RJ and Zhad C (2006) Tyrosine kinases: Maining in leprosy. *Nature Medicine* 12: 889–890.

Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

Michailov GV, Sereda MW, et al. (2004) Axonal neuregulin-1 regulates myelin sheath thickness. *Science* 304: 700–703.

Nave KA and Salzer JL (2006) Axonal regulation of myelination by neuregulin1. *Current Opinion in Neurobiology* 16: 492–500.

Saint-Jeannet J-P (2006) *Advances in Experimental Medicine and Biology,* vol. 589. Berlin: Springer.

Tavegia C, Zanazze G, petrylak A, et al. (2005) Neuregulin-1 type III determines the ensheabment Jabe of axons. *Neuron* 47: 681–694.

Neural Crest Cell Diversification and Specification: Melanocytes

K Bismuth and H Arnheiter, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

Melanocytes are pigment-producing cells of vertebrates. They are found in the integument, the eye, the inner ear, and, depending on the species, also in a variety of other inner organs, such as the lung, the heart, and around the aorta. Their pigment, termed melanin, is a biopolymer that is synthesized through a series of catalytic steps starting with the conversion of the amino acid tyrosine to dihydroxyphenylalanine (DOPA) by a copper-binding enzyme called tyrosinase. There are two types of melanin : eumelanin, which is black, and pheomelanin, which is yellow. Both types are made in membrane-bound intracellular organelles, called melanosomes, in which they can be found in pure form or mixed. Not every cell that contains melanosomes is a melanocyte, however, because melanocytes can transfer their melanosomes to other cell types that are not capable of producing melanin. In human skin, for instance, melanosomes are transferred from melanocytes to keratinocytes and into hair shafts, and in birds they are found in feathers. Clearly, a major role of melanocytes is to provide a body's visible pigmentation and so to aid in camouflage, display, and sun protection.

However, melanocytes are also critical for other organs. In the eye, for instance, they are found in the iris, which serves as an aperture that controls the depth of focus and regulates the amount of light that enters the eye. They are also found in the choroid behind the retina, where they serve as a light screen, blocking the passage of light to deeper tissues. In the inner ear of mammals, they are found in a portion of the wall of the cochlear duct, called the stria vascularis, where they regulate the ionic composition of the endolymph and ensure that auditory hair cells function properly. In fact, diminished numbers of strial melanocytes can lead to hearing impairment, as is likely the case in human Waardenburg syndromes, which are characterized by pigmentary abnormalities in skin, hair, and eyes and by varying degrees of deafness. Importantly, the function of melanocytes in hearing is independent of their pigmentation because mammals that lack melanin but retain melanocytes usually have normal hearing. Hence, melanocytes have important roles beyond fashioning the variety and beauty of an organism's coloration.

Development of Melanocytes

Melanocytes are generated from the neural crest, which is an embryonic population of cells that forms at the border between the neural plate and the adjacent surface ectoderm. Neural crest cells are induced by extracellular signals such as bone morphogenic proteins (BMPs), the Wingless/Int family of proteins (WNTs), and fibroblast growth factors (FGFs). These ligands induce transcription factors that hierarchically serve to specify the neural plate border, the neural crest, and, finally, a variety of cellular derivatives. In addition to melanocytes, these include the glial cells and neurons of the peripheral nervous system, smooth muscle cells, and bone and cartilage of the head.

The precursors to melanocytes, called melanoblasts, emerge at all axial levels between the region of the diencephalon and the tail end of the neural tube. This is in contrast to other crest derivatives, which are usually made at more restricted sites. Soon after their generation, neural crest cells accumulate in an area, called the migration staging area, close to the dorsal lip of the neural tube. From there, they migrate to their ultimate destinations, first leaving from the head region and then progressively from regions further toward the tail. In many vertebrates, melanoblasts migrate only after other crest derivatives have left. In regions where somites are present, melanoblasts travel between the somites and the surface ectoderm, with a major direction from back to belly but also longitudinally from head to tail and tail to head. This pathway is called the dorsolateral pathway and is different from the pathway taken by nonmelanoblastic crest derivatives, which migrate between the somites and the neural tube. In regions where no somites are present, particularly in the area of the otic vesicle, melanoblasts travel close to the neural tube as well as laterally toward the surface ectoderm. The majority of melanoblasts enter the dermis and then cross the basal membrane to take up residence in the epidermis, feather buds, or hair follicles. Others migrate toward the developing inner ear, the choroid, the ciliary body, the iris, and other inner organs. While traveling, melanoblasts begin to express a series of proteins, including the melanin-synthetic enzymes dopachrome tautomerase (DCT), tyrosinase-related protein-1, and tyrosinase, and the melanosomal protein pMEL17. In mice, expression of these proteins follows a characteristic sequence over several days. This sequence culminates with the induction of tyrosinase, which is the rate-limiting enzyme in melanin synthesis. Importantly, both cell proliferation and migration can continue even after

melanin has accumulated in the cells. The principal stages of melanoblast development in mouse embryos are shown in Figure 1.

This brief description of melanocyte development leaves us with a number of specific questions. What molecular signals specify melanoblasts and segregate them from other crest derivatives? Is specification irreversible, and if so, when does irreversibility occur – before migration or only later when the cells are closer to their final destination? How many cells are specified as melanoblasts, and how much do they have to proliferate to populate an entire body? What are the mechanisms that control melanoblast proliferation, guide them to their destinations, and control the progression toward their differentiation? What are the mechanisms that replenish melanocytes that are lost in adults, for instance, during hair and feather molting, when there is no recourse to the embryonic neural crest? We briefly discuss our current knowledge on these questions.

Melanoblast Specification: The Interplay between Signaling Pathways and Transcription Regulation

The analysis of both *in vivo* genetic models and *in vitro* culture systems made it possible to identify a number of factors involved in melanoblast specification. Three common notions have emerged from these studies. First, many of the factors involved in

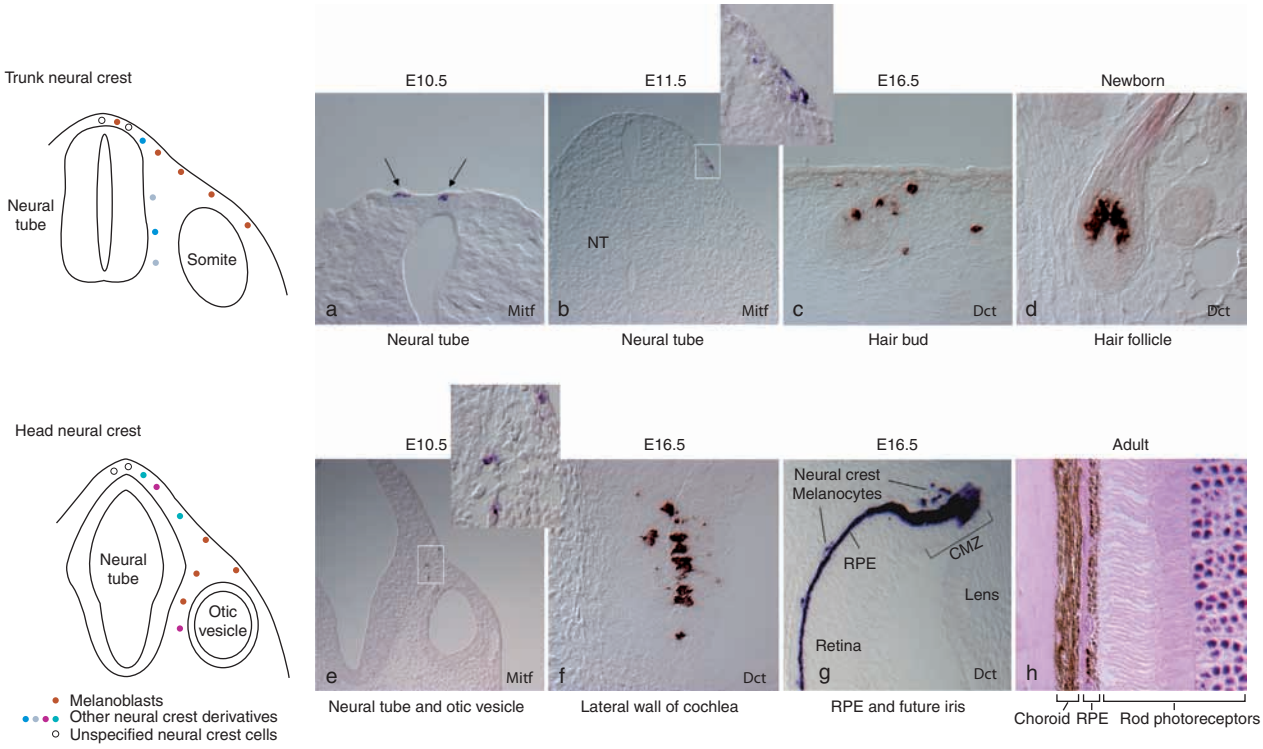

Figure 1 Melanoblast development and migration from the neural crest in mouse embryos. The left side shows schematic drawings of sections through the trunk and head areas as well as the respective migration pathways of melanoblasts and other neural crest derivatives. (a–g) Represent cryostat sections prepared from albino embryos or newborns that were labeled by nonradioactive *in situ* hybridization using riboprobes. Hence, the label marks gene expression, not pigmentation. (a, b, and e) Expression of the transcription factor gene *Mitf* is, together with KIT expression (not shown), one of the first signs of melanoblast specification in the neural crest. In the trunk area, *Mitf*-expressing cells are first found in the region of the roof plate of the neural tube (arrows in a). Then they migrate on a dorsolateral pathway underneath the surface ectoderm (b, and magnification of boxed area in b). (c and d) At later stages, the *Mitf* marker becomes progressively weaker and melanoblasts can best be labeled by expression of the melanogenic gene *Dct*. (c) Shows *Dct*-positive cells that have entered the epidermis and developing hair buds, and (d) shows *Dct*-positive cells in hair follicles of newborns. (e) In the head region, neural crest cells express *Mitf* only after they have migrated away from the crest. They are found underneath the surface ectoderm and on a pathway toward the otic vesicle. (f) *Dct*-positive melanoblasts have entered the lateral wall of the cochlear duct at E16.5. (g) Other melanoblasts have migrated close to the optic cup and start to populate the region behind the retinal pigment epithelium (RPE). This region will form the pigmented choroid. Neural crest-derived melanoblasts have also migrated close to the ciliary margin zone (CMZ), that will form the ciliary body and the iris of the eye. Note that the RPE is also positive for *Dct* but is derived directly from the optic neuroepithelium, not the neural crest. (h) H&E-stained section through the choroid and RPE of an adult eye of a pigmented mouse. Only the choroidal pigment cells are neural crest derived.

early specification operate not at a single stage but during several phases of melanocyte development. Second, many of the factors can either promote or inhibit melanocyte formation, depending on the conditions. Third, many of these factors also operate in other organ systems, and their mutations can lead to disorders unrelated to melanocyte function, including embryonic or adult lethality.

A list of factors currently implicated in melanocyte development is shown in Table 1. The factors are arranged by functional groups and alphabetically within groups, starting with a group of factors operating from outside the cells and ending with a group functioning inside. Not listed are the typical differentiation (pigmentation) genes, and we refer the reader to other reviews for more detailed information on the factors not specifically covered here.

Numerous genetic and expression studies as well as biochemical and cell-biological analyses have demonstrated that among the early effectors of melanoblast development, the cell surface receptor KIT and the transcription factor MITF mark melanoblasts selectively in the neural crest and are the ones most specifically involved in their development. KIT is a receptor tyrosine kinase that binds a single ligand, KIT ligand (KITL), and activates the mitogen-activated protein (MAP) kinase pathway and the phosphoinositide-3 (PI3) kinase pathway. Based on natural and genetically engineered mouse mutants, the MAP kinase pathway is critical for melanocyte development, but the PI3 kinase pathway is dispensable. MITF is a basic–helix–loop–helix–leucine zipper transcription factor that was originally identified in mice with transgenic insertional mutations at the microphthalmia locus (hence its name, microphthalmia-associated transcription factor). Such mice are entirely white and deaf because their melanoblasts do not develop properly, and they have small eyes (microphthalmia) because the development of the pigmented epithelium of the retina, a neuroepithelial derivative, also depends on MITF. MITF has several isoforms, one of which, the M-MITF isoform, is under the control of its own specific promoter, the M-promoter, and is the major isoform of melanoblasts.

There is *in vitro* evidence that KIT and MITF engage in cross-talk. The KIT-mediated activation of the MAP kinase pathway leads to serine phosphorylation of MITF, which regulates MITF's transcriptional activity and stability. Conversely, MITF is needed for KIT to accumulate in melanoblasts. KIT and MITF are thus part of a feedback loop. This feedback loop intersects with a second signaling pathway that also acts early in melanoblast development and involves the ligand endothelin-3 (EDN3) and the G-coupled endothelin receptor-B (EDNRB). Consistent with the observation that EDNRB signaling stimulates melanoblast proliferation *in vitro*, EDNRB-deficient mouse embryos have few melanoblasts, and after birth melanocytes are largely missing except in spots around the neck and the base of the tail. Although EDNRB is expressed in melanoblasts where it acts cell autonomously, it is also expressed in other cell types, notably in cells of the neural tube. Interestingly, co-cultured, EDNRB wild-type neural tube or nonmelanoblastic neural crest cells can rescue the proliferation of EDNRB-deficient melanoblasts and let them develop at least to the tyrosinase-positive stage, but similar cells genetically deficient in KITL cannot. Because KITL alone can rescue EDNRB-deficient melanoblasts *in vitro*, it is conceivable that EDNRB signaling also helps melanoblast development in a cell nonautonomous manner by stimulating KITL production or secretion in other cells. Hence, it appears that all three factors – EDNRB, KIT, and MITF – are functionally linked during the early stages of melanoblast development.

What, then, induces these key factors? Much work has gone into the identification of transcription factors that directly stimulate the M-MITF promoter. They include SOX10, a member of the SRY-box-containing high-mobility group of transcription factors; PAX3, a paired domain transcription factor; LEF1/TCF, another high-mobility group protein; and CREB, a basic–leucine zipper protein. None of these are specific to melanoblasts, however. Thus, much as with many other promoters, it is likely the combinatorial network of these and other factors, each with its own specific transcriptional activities, that regulates the M-MITF promoter. Similar reasoning applies to KIT and EDNRB even though much less is known about their transcriptional regulation in melanoblasts.

On the next level up, extracellular ligands have been identified that regulate the regulators of MITF. Among them are WNT proteins, which are cysteine-rich secreted proteins that bind members of the G-coupled receptor family called Frizzled receptors (FZD) and receptors related to low-density lipoprotein receptor (LPR; particularly LPR5 and LPR6). The coactivation of these two types of receptors induces the so-called canonical WNT signaling pathway, whose hallmark is the accumulation of β-catenin and its transfer to the cell nucleus, where it interacts with LEF/TCF to stimulate the transcription of specific target genes. Together with BMPs and FGFs, WNTs are critical in specifying the neural crest by operating upstream of a series of transcription factors, including PAX3, SOX9, SOX10, SLUG/SNAIL, AP-2, and FOXD3 (Table 1). Many of these factors are also engaged in extensive yet species-specific cross-regulatory networks. Interestingly, after its early role in neural

Table 1 Factors implicated in melanocyte specification and development

Functional group	*Symbol*	*Name*	*Function*	*Human neural crest-related disorders (OMIM No.)*[a]
Extracellular matrix and cell adhesion	ADAMTS20	Adamts20 metalloprotease	Involved in ECM remodeling	
	ECM	Extracellular matrix (collagen, fibronectin, laminin)	Migration of melanoblasts	
	EFNB	Ephrin B	Early migration of melanoblasts	
	EPHB	Ephrin B receptor	Early migration of melanoblasts	
	INT	Integrins	Migration of melanoblasts	
Growth factors and growth factor receptors	BMP	Bone morphogenic protein	Promotes or inhibits melanoblast specification depending on conditions	
	CDH	Cadherins	Migration of melanoblasts regulated by various combinations of cadherin family members	
	EDN3	Endothelin-3 (a 20-residue peptide)	Melanoblast proliferation and survival	Waardenburg syndrome IV (277580), Hirschsprung disease and related disorders (142623, 209880)
	EDNRB	G-coupled endothelin receptor B	Melanoblast proliferation and survival	Waardenburg syndrome IV (277580), Hirschsprung disease and related disorders (142623, 600501)
	FZD	Frizzled family of G-coupled receptors for WNT	Activation of canoncial WNT signaling, induction of neural crest and promotion of melanoblast development	
	KITL	KIT ligand	Melanoblast survival, proliferation, migration	
	KIT	KIT receptor tyrosine kinase	Activates Ras/Raf/MAP kinase pathway, melanoblast survival, proliferation, migration	Piebaldism (172800)
	LRP	Family of receptors related to low-density lipoprotein receptors	Co-receptor for FZD, activation of canoncial WNT signaling, induction of neural crest and promotion of melanoblast development	
	NOTCH	Family receptors related to *Drosophila* Notch	Melanoblast survial and melanocyte stem cell maintenance	
	WNT	Wingless-related/MMTV integration site	Induction of neural crest, promotion of melanobalst development	
Transcription factors and cofactors	AP-2	AP-2 transcription factor	Melanoblast differentiation	
	CATNβ	β-Catenin	One of the effectors of WNT signaling	
	FOXD3	Forkhead box D3	Neural crest induction, can inhibit melanoblast specification	
	LEF1/TCF	Lymphoid enhancer factor/T cell-specific transcription factor	One of the effectors of WNT signaling	
	MITF	Microphthalmia-associated transcription factor, basic helix–loop–helix–leucine zipper family	Melanoblast specification, survival, proliferation, migration	Waardenburg syndrome IIa (193510)
	PAX3	Paired domain protein-3	Melanoblast specification, survival, and proliferation; regulation of adult melanocyte stem cells	Waardenburg syndrome I, II, III (193500, 193510, 148820)
	SNAI2/SLUG	Snail homolog protein-2, zinc finger protein	Melanoblast specification downstream of MITF?	
	SOX9	SRY-box containing protein-9	Induction and differentiation of the neural crest	
	SOX10	SRY-box containing protein-10	Melanoblast specification, survival, proliferation, and differentiation	Waardenburg syndrome IV (277580)

[a]*OMIM, Online Mendelian Inheritance in Man.*

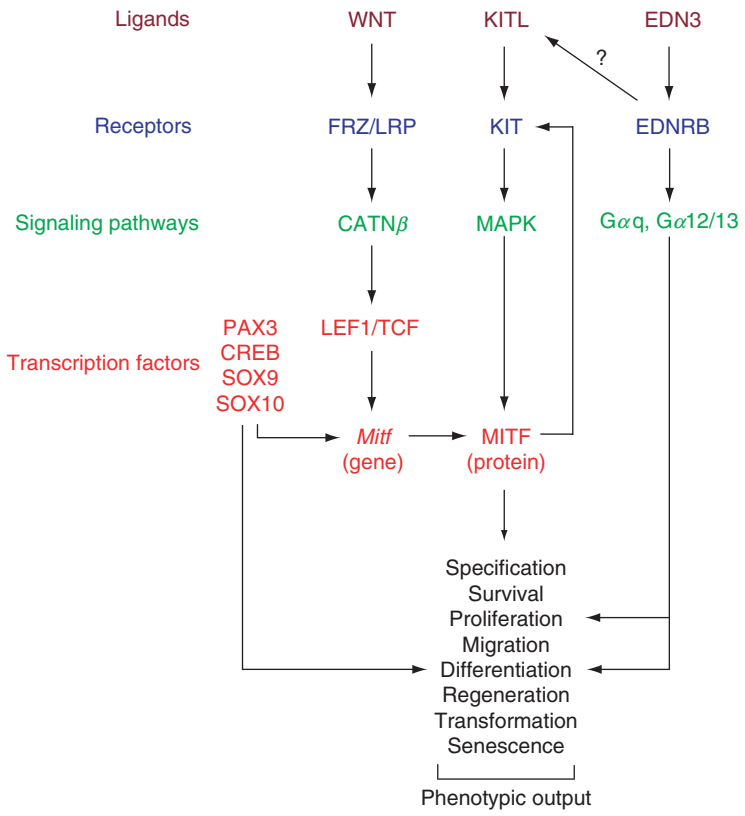

Figure 2 Simplified schematic representation of the interplay between signaling pathways and transcription factors involved in melanocyte development. For abbreviations, see Table 1. Biochemical evidence shows that the MAPK pathway regulates MITF activity and stability through phosphorylation. Genetic evidence suggests that EDNRB signals through Gnaq and Gna11, two members of the Gαq family of G-proteins, as well as through Gα12/13.

crest specification, canonical WNT signaling is also critical in promoting the melanoblast fate. For instance, in mouse embryos, elimination of β-catenin in neural crest stem cells leads to a lack of melanoblasts and sensory neuronal lineages but not other crest derivatives. In zebra fish, overexpression of β-catenin in premigratory neural crest cells leads to excess pigment cell formation through the stimulation of MITF. In transgenic mice, however, stimulation of MITF by stably active, nuclear β-catenin can also lead to inhibition of melanoblast proliferation, illustrating that potent factors such as β-catenin need to be finely tuned with respect to their levels and specific time points of expression. This is all the more important because β-catenin also acts downstream of cadherins, which together with integrins regulate cell adhesion and migration. A simplified scheme of the interactions of the extracellular ligands and transcription factors involved in melanoblast development is shown in Figure 2.

Melanoblast Specification: Cell Number and Location

In previous decades, the existence of melanoblasts could be assumed only indirectly from the existence of their descendents, the melanocytes. Using pigmentation as the readout, two types of studies have suggested that in mice, melanocytes must arise from a very small number of precursors. Mintz produced chimeras of embryos of pigmented and unpigmented, tyrosinase-deficient (albino) mice. From the width and arrangement of the pigmentation stripes that appear in adults, she concluded that there are precisely 34 melanocyte precursors, 17 equally spaced on either side of the embryo – three for the head, six for the body, and eight for the tail. Support for a small number of melanocyte precursors, although not exactly 34, came from the infection of albino embryos with retroviral vectors expressing wild-type tyrosinase. The resulting mice had small pigmented stripes whose widths were reminiscent of those observed in Mintz's chimeras. In later studies, the use of *in situ* hybridization or antibody probes specific for melanoblasts indicated that the number of initially specified melanoblasts is indeed small, particularly in the trunk region. These studies showed, however, that melanoblasts are unevenly distributed between head and tail. The use of a line of transgenic mice in which the bacterial *lacZ* gene is expressed under control of the promoter of the early melanogenic gene *Dct*

also showed an uneven head-to-tail distribution of melanoblasts. However, the *Dct–lacZ* marked melanoblasts were found to mix extensively along the head-to-tail axis and to be considerably more numerous than suggested from the earlier studies. Each study has to be considered with caution, however. The first assumes that one stripe represents a clone derived from a single cell, which is only true if the cells are thoroughly mixed early in the chimeric embryos but their descendents do not mix later in development. The second may not reflect the marking of the earliest melanoblasts because cells could be infected at different stages during lineage development. The third is limited by the sensitivity of the *in situ* probes, and the fourth may overestimate the number of melanoblasts because it cannot be excluded that *lacZ*-marked cells may give rise to other cells besides melanocytes. Thus, the truth may well lie somewhere in between: more than 34 melanoblasts in mice but less than 1000.

The question of where and when melanoblasts are committed to the melanocyte lineage has also not been answered definitively. On the one hand, the coexpression of KIT and MITF in premigratory mouse neural crest cells strongly argues for an early commitment of melanoblasts. On the other hand, there is *in vitro* evidence that *Dct*-positive neural crest cells, usually considered to be melanoblasts, are capable of generating both melanocytes and smooth muscle cells. In the quail, even epidermal melanocytes have been observed to switch their fate and become Schwann cells, the myelin-forming cells of the peripheral nervous system (and vice versa, Schwann cells have been seen to become melanocytes). In fact, the question of commitment cannot be answered universally for all vertebrate species, and probably not even universally for all axial levels in the same species. In the head region of mouse embryos, for instance, *Kit*-positive cells, identified by expression of a knock-in Kit^{lacZ} allele, are first observed at the dorsal midline of the neural tube, but MITF expression, detected by antibody labeling or *in situ* hybridization, is usually seen only in postmigratory melanoblasts. In contrast, at the trunk level, both KIT and MITF are first expressed together in cells at the top of the neural tube (**Figures 1(a) and 1(e)**).

Although there are thus arguments for both premigratory and postmigratory commitment, the proof of a true restriction in a melanoblast's developmental potential requires systematic challenges with conditions permissive for alternative fates: Fate maps are not indicative of fate restrictions. For instance, it is not known whether MITF/KIT double-positive melanoblasts of mice, which normally develop only into melanocytes, are entirely incapable *in vivo* of developing into other cell types. This point is not of mere academic interest because a prolonged retention of a broader developmental potential in one cell lineage may provide an embryo with the means to correct developmental aberrations in another, and thus still yield a healthy adult.

Melanoblast Survival, Proliferation, and Migration

The developing population of melanoblasts is regulated by the parameters of cell survival, cell proliferation, and cell migration. The molecular dissection of these parameters has indicated a series of important players. Recent evidence suggests, for instance, that Notch signaling, which is involved in many developmental processes, is also critical for melanoblast survival. Notch receptors represent a family of conserved transmembrane proteins that are activated by specific ligands, Delta and Jagged, through direct cell–cell interactions. In fact, Notch signaling, acting through the downstream target gene *Hes1*, plays a critical role in preventing apoptosis of melanoblasts. Moreover, the previously mentioned extracellular factors WNT, EDN3, and KITL, are critical for cell survival and for promoting cell proliferation. As alluded to previously, the actions of these factors converge on MITF, the all-purpose melanoblast transcription factor, which has been shown to regulate cell survival and proliferation. *In vitro* experiments in melanoma cells suggest, for instance, that MITF stimulates the expression of the antiapoptotic genes *Bcl2* and hypoxia-inducible factor-1α (*Hif1α*), but a C-terminal fragment of MITF, liberated by the proapoptotic, activated caspase-3, enhances apoptosis. MITF also seems to have both positive and negative effects on cell proliferation. It promotes cell proliferation by stimulating the transcription of the gene encoding cyclin-dependent kinase-2 (CDK2), which promotes the G_1-to-S transition during the cell cycle. MITF also stimulates the transcription factor TBX2, which represses the expression of p21, a protein inducing growth arrest. Furthermore, MITF appears to be essential, and in fact is increasingly expressed, in highly aggressive, metastasizing melanoma cells, earning it the name 'lineage addiction oncogen.' On the other hand, MITF is also clearly antiproliferative: it activates the gene encoding INK4a, which promotes cell cycle exit, and it directly stimulates the gene encoding p21, which induces G_1 arrest. What controls the net result of these opposing actions is not entirely clear but may rest in part on the balance of expression of distinct MITF isoforms. For instance, when expressed *in vitro*, isoforms that lack certain subexons are less antiproliferative, or may even promote proliferation, compared to isoforms

that contain these subexons. Hypothetically, then, distinct MITF isoforms might be expressed at different developmental time points or during different phases of the cell cycle and may regulate different subsets of cell cycle regulators. This might occur either at the transcriptional level or by direct interactions at the protein level. MITF and cell cycle regulators may hence be integrated at multiple levels.

While melanoblasts proliferate, they also migrate, but proliferation and migration are not synchronized because there are spurts of growth that alternate with spurts of migration. Migration is controlled in part by cell adhesion and the extracellular matrix. Cadherins form a family of glycoproteins involved in calcium-dependent cell adhesion, regulation of the cytoskeleton, and cell signaling. Melanoblasts express distinct cadherins very dynamically during development. For instance, when neural crest cells delaminate from the epithelium, they express cadherin 6 but only low levels of N-cadherin. In migrating melanoblasts, E- and P-cadherins are weakly upregulated, but upon arrival in the dermis, neither is expressed. E-cadherin increases 200-fold when the cells enter the epidermis and interact with E-cadherin-expressing keratinocytes. Melanoblasts that enter hair follicles, however, downregulate E-cadherin again and exclusively express P-cadherin. The few melanocytes remaining in the epidermis continue to express E-cadherin, and those in the dermis express N-cadherin. These dynamic changes in cadherin expression seem to adapt the cells to their local environment, a hypothesis that has been tested by both gain-of-function and loss-of-function mutations in select cadherins. Because cadherins have wide-ranging actions in other tissues, however, specific effects on melanoblasts are not always easy to discern.

The parameters of cell migration have also been explored in heterologous cell grafting experiments using unrelated cell types, such as tumor cells or embryonic stem cells. These experiments point out the importance of β_1 integrin, a member of a family of transmembrane proteins that assemble in heterodimers and serve as major binding partners for extracellular matrix components, but their relevance to melanoblast migration is unclear. Support for the importance of the extracellular matrix in melanoblast migration comes from genetic observations on the role of proteases known to remodel the extracellular matrix. The mouse coat color mutation *belted*, for instance, is due to a mutation in the metalloprotease gene *Adamts20*. This gene is homologous to *Gon-1* of *Caenorhabditis elegans*, which encodes a metalloprotease required for cell migration. In summary, the regulation of cell migration involves a variety of partners that all collaborate to ensure that the cells end up on the right tracks and in the right locations.

Melanoblast Differentiation

Melanoblast differentiation, characterized by melanization and the acquisition of a dendritic phenotype, relies heavily on MITF, whose targets include many of the genes involved in melanin biosynthesis (such as the previously mentioned *Dct*, tyrosinase-related protein-1, and tyrosinase). A common *cis* element in the promoters of these genes is the so-called M-box, which in its core contains an Ephrussi (E) box of the sequence 5'CATGTG3'. E boxes are widespread in the genome and bound by a wide array of basic–leucine zipper, basic–helix–loop–helix, and basic–helix–loop–helix–leucine zipper transcription factors. Although flanking sequences influence the efficiency with which specific factors bind a given E box, they are insufficient to explain the restriction of the melanogenic genes to the melanocyte lineage. Again, it is likely the combination of several distinct transcription factors that is critical for this regulation. As alluded to previously, the activity of MITF is also regulated by post-translational modifications such as phosphorylation. Moreover, it has been described that MITF is dynamically modified by small ubiquitin-like modifier (SUMO) side chains, which are proteins of approximately 100 residues that are added to two lysines of MITF. Sumoylation, as the modification is called, reduces MITF's activity on promoters with multiple E boxes, but not on those with single E boxes, and thus may influence target gene specificity. Furthermore, both *in vitro* and *in vivo* tests show that some melanogenic genes are stimulated not only by MITF but also directly by SOX10, a regulator of MITF, suggesting the existence of feed-forward loops between SOX10 and MITF target genes. By regulating genes involved in the dynamics of the actin cytoskeleton, MITF also modulates cell shape. Nevertheless, although MITF is expressed at high levels in melanoblasts as they differentiate into melanocytes, it soon fades away in most differentiated cells; thus, it is conceivable that maintenance transcription of melanogenic genes in mature melanocytes, and possibly the maintenance of cell shape, is controlled by other E-box binding factors expressed in mature melanocytes. Without functional MITF, however, melanocytes can never mature and become pigmented, but when they remain unpigmented because they have mutations in the melanin biosynthetic genes, they stay alive and thrive.

Adult Melanocyte Stem Cells

Tissue injury, be it the result of physical or metabolic processes, is a common feature of life, and hence the mechanisms of tissue repair are of great importance for the maintenance of health. Much excitement has

been created by the recognition that adult mammals contain stem cells capable of regenerating lesioned tissues. Stem cells are usually defined as undifferentiated cells capable of self-renewal and of generating a variety of differentiated cells. Many stem cells reside in a particular compartment, called the niche, that provides the necessary environment for their maintenance. An example of such a niche is the bulge region of hair follicles which lies between the hair bulb and the sebaceous gland. As the hair goes through its cycles of loss and regrowth (technically, the phases are know as catagen, telogen, and anagen), multipotent stem cells from the bulge replenish all types of epithelial cells of the hair structure. A particular slow-cycling cell, marked in mouse transgenic studies by the previously mentioned *lacZ* reporter under the control of the *Dct* promoter, is capable of self-renewal and returning to a quiescent stage; of populating neighboring, vacant bulges; and of giving rise to melanocytes. These cells share many features with developmental melanoblasts but have been dubbed melanocyte stem cells. However, the usual stem cell feature, multipotentiality, has not been shown to be among their traits, and their relationship to multipotent stem cells in the niche is unclear.

Major questions in stem cell biology concern the mechanisms of stem cell maintenance and the distinction between self-renewal and differentiation. Interestingly, several of the signaling pathways and transcription factors mentioned previously in the context of melanoblast development are also involved in melanocyte stem cell regulation. For instance, Notch signaling is required for the maintenance of melanocyte stem cells in the hair follicle, and PAX3, SOX10, and MITF, together with WNT signaling, are involved in regulating differentiation. Whereas SOX10 and MITF collaborate to stimulate *Dct* transcription, PAX3 represses *Dct* transcription. It does so by recruiting the Groucho corepressor GRG4 and by interacting with LEF/TCF, whose binding to the *Dct* promoter competes with MITF binding, resulting in low *Dct* expression. What can flip the switch is extracellular signaling, particularly the activation of the canonical WNT signaling pathway and its concomitant β-catenin activation. Activated β-catenin interferes with the binding of PAX3 to LEF/TCF and GRG4 and dislodges PAX3 from the *Dct* promoter. This relieves repression of the *Dct* promoter and frees PAX3 to stimulate MITF expression to higher levels, which in turn activates the *Dct* promoter and sends the cells on the way to differentiation. Of course, this process is not allowed to act indiscriminately in all melanocyte stem cells, lest there would be no stem cell left for the next hair cycle. How some cells of the niche escape the effects of WNT signaling is not yet known, however. Intriguingly, premature differentiation and concomitant loss of stem cells seem to be a feature of a mutant allele of *Mitf*, $Mitf^{vit}$, which leads to premature graying. This and other mutants in MITF and in other factors underscore the importance of the mentioned pathways not only for development but also for tissue regeneration.

In summary, the biology of melanocytes and their regulation from birth to death has provided us with many mechanistic insights that are not readily attainable in other systems. Melanocytes, in fact, are highly suitable to elucidate basic principles of cell lineage determination in the neural crest.

See also: Macroglial Lineages; Neural Crest; Neural Stem Cells: Adult Neurogenesis; Neural Crest Cell Diversification and Specification: ErbB Role; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation.

Further Reading

Arnheiter H, Hou L, Nguyen MTT, et al. (2006) Mitf – A matter of life and death for the developing melanocyte. In: Hearing V and Leong SPL (eds.) *From Melanocytes to Malignant Melanoma*, pp. 27–49. Totowa, NJ: Humana.

Dupin E, Glavieux C, Vaigot P, and LeDouarin NM (2000) Endothelin 3 induces the reversion of melanocytes to glia through a neural crest-derived glial–melanocytic progenitor. *Proceedings of the National Academy of Sciences of the United States of America* 97: 7882–7887.

Hari L, Brault V, Kleber M, et al. (2002) Lineage-specific requirements of beta-catenin in neural crest development. *Journal of Cell Biology* 159: 867–880.

Lang D, Lu MM, Huang L, et al. (2005) Pax3 functions at a nodal point in melanocyte stem cell differentiation. *Nature* 433: 884–887.

Le Douarin NM, Creuzet S, Couly G, and Dupin E (2004) Neural crest cell plasticity and its limits. *Development* 131: 4637–4650.

Moriyama M, Osawa M, Mak S-S, et al. (2006) Notch signaling via Hes 1 transcription factor maintains survival of melanoblasts and melanocyte stem cells. *Journal of Cell Biology* 173: 333–339.

Meulemans D and Bronner-Fraser M (2004) Gene-regulatory interactions in neural crest evolution and development. *Developmental Cell* 7: 291–299.

Nishimura EK, Granter SR, and Fisher DE (2005) Mechanisms of hair graying: Incomplete melanocyte stem cell maintenance in the niche. *Science* 307: 720–724.

Read A (2000) Waardenburg syndrome. In: Scriver C, Beaudet A, Valle D, and Sly W (eds.) *The Metabolic and Molecular Basis of Inherited Disease,* 8th edn., pp. 6097–6116. Toronto: McGraw-Hill.

Shin MK, Levorse JM, Ingram RS, and Tilghman SM (1999) The temporal requirement for endothelin receptor-B signalling during neural crest development. *Nature* 402: 496–501.

Van Raamsdonk CD, Fitch KR, Fuchs H, De Angelis MH, and Barsh GS (2004) Effects of G-protein mutations on skin color. *Nature Genetics* 36: 961–968.

Wilkie AL, Jordan SA, and Jackson IJ (2002) Neural crest progenitors of the melanocyte lineage: Coat colour patterns revisited. *Development* 129: 3349–3357.

Wilson YM, Richards KL, Ford-Perriss ML, Panthier JJ, and Murphy M (2004) Neural crest cell lineage segregation in the mouse neural tube. *Development* 131: 6153–6162.

Relevant Website

http://www.ncbi.nlm.nih.gov – National Center for Biotechnology Information, *Online Mendelian Inheritance in Man*.

Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation

M Wegner, Universität Erlangen-Nürnberg, Erlangen, Germany

© 2009 Elsevier Ltd. All rights reserved.

The Role of Transcription Factors in the Schwann Cell Lineage

Schwann cells and satellite glia are the two main glial cell types of the peripheral nervous system (PNS). Whereas satellite glia are found within ganglia in close association with neuronal somata, Schwann cells are found in close contact with axons in the peripheral nerves. Both Schwann cells and satellite glia arise from neural crest stem cells. Following their specification, Schwann cells go through several consecutive developmental stages, including the Schwann cell precursor and the immature Schwann cell, before they finally undergo terminal differentiation and become myelinating or nonmyelinating Schwann cells of the adult PNS. Each developmental stage is characterized by specific patterns of gene expression, and progression from one stage into the next requires corresponding changes. Taking the role of transcription factors as regulators of gene expression into account, each stage of Schwann cell development and each transition from one stage into the next must be brought about by changes in transcription factor occurrence or activities.

Transcriptional Control of Schwann Cell Specification

Transcription factors with importance for Schwann cell development are already present in pluripotent neural crest stem cells. This includes the α isoform of transcription factor AP2, the forkhead protein FoxD3, the paired domain protein Pax3, and the high-mobility-group domain containing protein Sox10. Because all these transcription factors are expressed during the specification event, their presence must at least be compatible with the transformation of a neural crest stem cell into a Schwann cell precursor. Except for Sox10, whether these transcription factors are also required for specification has not been rigorously tested.

In neural crest stem cells, Sox10 is important for survival and for the maintenance of pluripotency; trunk neural crest stem cells show a reduced survival rate in mice deficient for Sox10. However, enough neural crest stem cells survive to aggregate to dorsal root ganglia and populate roots and proximal portions of peripheral nerves. Whereas neuronal specification in the dorsal root ganglia is initially undisturbed in the absence of Sox10, neural crest cells fail to undergo specification to glia. These loss-of-function studies show that Sox10 is indeed required for the specification of Schwann cell precursors as well as satellite glia precursors, both of which can be distinguished from neural crest stem cells by the expression of B-FABP. Schwann cell precursors also start to express low levels of the myelin protein zero (MPZ) as a marker. This basal level of MPZ expression is probably due to a the direct activation of the MPZ promoter by Sox10 and is orders of magnitude below the level attained later during active myelination.

Because the specification of both Schwann cells and satellite glia is under control of Sox10, Sox10 alone cannot account for the difference between both glial cell types. Rather, different transcription factors must be present in presumptive glia along nerves and in dorsal root ganglia, so that cooperation of these transcription factors with Sox10 leads to different glia, depending on their localization. This model fits the general perception of Sox proteins as transcription factors that function mainly in combination with partner proteins. Although this limits the intrinsic impact of Sox proteins on developmental processes, it also endows them with functional versatility. Sox10, for example, exerts many additional functions at later stages of Schwann cell development with changing partner proteins.

Hooked on Neuregulins: The Schwann Cell Precursor

Newly formed Schwann cell precursors depend for survival and proliferation on neuregulin 1 signaling. In particular, the membrane-bound isoform III of neuregulin 1 is strongly expressed on axons at this time, thus making survival of Schwann cell precursors dependent on axonal contact. Schwann cells recognize the neuregulin 1 signal with a heterodimeric receptor consisting of ErbB2 and ErbB3. In the absence of Sox10, the expression of ErbB3 is strongly reduced in neural crest-derived cells along the nerve, proving that Sox10 is genetically upstream of ErbB3 expression. Ectopic expression of Sox10 in a neural cell line, furthermore, leads to a rapid induction of ErbB3 expression, arguing that ErbB3 transcription may be

directly dependent on Sox10 in neural cells. Sox10 may, thus, also ensure the survival of Schwann cell precursors by making them responsive to neuregulin 1.

Neuregulin 1 continues to be important for Schwann cells and promotes both the transition of Schwann cell precursors to immature Schwann cells and myelin formation. Neuregulin 1 signals seem predominantly transduced through the mitogen-activated protein kinase (MAPK) pathway in Schwann cells and through Ets proteins as their nuclear effectors. Accordingly, several Ets proteins, in particular Net and GA binding protein (GABP-)α, are prominently expressed in Schwann cell precursors, immature Schwann cells, and throughout myelination.

Neuregulin 1 also stimulates the expression of the transcription factor cyclic adenosine monophosphate (cAMP) response element binding protein (CREB) and induces sustained phosphorylation of CREB on its serine residue 133 in Schwann cells. Signaling through CREB appears to contribute to Schwann cell proliferation during the early stages of lineage development.

Molting into an Immature Schwann Cell

Several days after specification, Schwann cell precursors develop further into immature Schwann cells. In the mouse, this transition occurs around 13 days postcoitum, requires continued neuregulin signaling, and is characterized by changes in gene expression. Most important, from this stage onward Schwann cells establish an autocrine signaling mechanism that makes them less dependent on axonal signals for survival and proliferation. It is not yet known which transcription factor or which combination of transcription factors is needed to drive Schwann cell precursors into the immature Schwann cell stage. However, evidence exists to implicate AP2-α as a negative regulator of this process. As already mentioned, AP2-α expression has continued in Schwann cell precursors from the neural crest cell stage. Immature Schwann cells normally arise once AP2-α expression is extinguished. Reciprocally, prolonged expression of AP2-α interferes with the timely progression from the Schwann cell precursor to the immature Schwann cell stage. Taking into account that endothelins exert a similar effect on Schwann cell development through the endothelin B receptor, a link may exist between AP2-α and endothelin signaling in Schwann cell precursors.

Maintaining the Immature State

In contrast to AP2-α, Sox10 continues to be expressed in immature Schwann cells, possibly ensuring continued ErbB3 expression during this stage of Schwann cell development. The same continued expression from neural crest stem cells into the immature Schwann cell stage is also observed for FoxD3 and Pax3.

In addition, immature Schwann cells start to express transcription factors that are undetectable at earlier stages of their development. This includes the two related POU homeodomain transcription factors – Oct6 (also known as SCIP and Tst-1) and Brn2 – as well as the zinc finger protein Krox20 (also known as Egr2). Initial expression of these transcription factors in immature Schwann cells is at a low basal level, orders of magnitudes below the maximal levels achieved at later times of Schwann cell development. Expression of Krox20 is, furthermore, inversely correlated to the expression of its close relative Krox24 (also known as Egr1), which is expressed at highest levels in Krox20-negative Schwann cell precursors and is completely absent later in myelinating Schwann cells which exhibit maximal Krox20 levels.

More prominently expressed than Oct6 and Krox20 in immature Schwann cells is the high-mobility-group domain protein Sox2. In contrast to its distant relative Sox10, it is specifically turned on in the Schwann cell lineage and not expressed in neural crest stem cells. In fact, forced expression of Sox2 in the presumptive neural crest at the time of neurulation leads to a fate switch and causes these cells to become part of the neuroepithelium. Sox2 expression is important to maintaining Schwann cells in the immature state. Similarly, this appears to be the case for Pax3 and FoxD3, as well as for the bZip transcription factor cJun. Similar to phosphorylated CREB, these transcription factors all stimulate Schwann cell proliferation. They also appear to repress terminal differentiation.

Schwann cell proliferation is also enhanced by glucocorticoids, indicating that immature Schwann cells express the glucocorticoid receptor. Glucocorticoid receptor is not the only nuclear hormone receptor in Schwann cells. Thyroid hormone receptor and progesterone receptor have also been detected, as well as several members of the p160 coactivator family of steroid receptor coactivators (SRC1–SRC3) as crucial mediators of nuclear hormone receptor effects. Interestingly, Schwann cells also have the capacity to synthesize progesterone, indicating that progesterone signaling may be part of the autocrine loop that exists in Schwann cells from the immature state onward.

Choosing between Two Alternative Fates during Terminal Differentiation

Progesterone and thyroid hormone have also been reported to promote myelin gene expression in Schwann cells and may thus be involved in triggering

terminal differentiation at the end of embryogenesis. During terminal differentiation, Schwann cells choose between two mutually exclusive pathways. Schwann cells in contact with small-caliber axons ensheath these axons as nonmyelinating Schwann cells. These cells no longer actively divide, but are not permanently withdrawn from the cell cycle. They, furthermore, continue to express many of the same transcription factors that were already expressed in the immature Schwann cells. Myelinating Schwann cells, on the other hand, exit the cell cycle, establish a one-to-one relationship with large-caliber axons and form myelin sheaths around an axonal segment. Myelin formation, in addition, requires dramatic changes in the expression pattern of myelinating Schwann cells as lipid biosynthesis and the production of myelin proteins (such as MPZ, MBP, PMP-22, and Connexin-32) and the synthesis of specific glycolipids are steeply increased.

Triggering Myelination by Axonal Signals

Of all the steps in Schwann cell development, transition from the immature Schwann cell to the myelinating Schwann cell is best understood. It is critically dependent on at least three groups of transcription factors, the first being nuclear factor (NF)-κB proteins, in particular p65/relA. NF-κB is already present in Schwann cells before myelination, and it is mainly localized during this stage in a complex with I-κB proteins in the cytoplasm and, thus, in an inactive state. At the onset of terminal differentiation, an axonal signal activates the low-affinity nerve growth factor (NGF) receptor p75 and thereby leads to nuclear translocation of NF-κB and induction of its transcriptional activity. In the mouse, blockade of NF-κB activity or the genetic deletion of p65 attenuates peripheral myelination significantly.

NF-κB also stimulates Oct6 expression in Schwann cells and is thus part of the strong axon-dependent upregulation of Oct6 gene expression immediately prior to myelination. Axonal signals influence Oct6 gene expression also through increased intracellular cAMP levels. This inducibility by cAMP has, in fact, led to the designation of Oct6 as suppressed cAMP-inducible POU (SCIP) protein.

Escorting Schwann Cells into the Myelinating Stage

Maximal Oct6 levels coincide with the phase during which Schwann cells in contact with large-diameter axons, establish a one-to-one relationship, and prepare for myelination. Oct6 is, therefore, strongly associated with this promyelinating stage. Schwann cells in Oct6-deficient mice become arrested in their development at the promyelinating stage, supporting the notion that Oct6 mediates the progression from the promyelinating into the myelinating stage. However, the arrest is only transient, so that Schwann cells eventually overcome this block in Oct6-deficient mice and form myelin sheaths with a delay of approximately 2 weeks.

The transient nature of this developmental defect can probably be attributed to the closely related POU protein Brn2, which is co-expressed with Oct6 in Schwann cells. During normal development, Brn2 is expressed at lower levels than Oct6 in Schwann cells. As a result, the loss of Brn2 does not lead to a noticeable change in pattern or timing of Schwann cell differentiation. Once Oct6 is experimentally deleted, a compensatory upregulation of Brn2 occurs, and these increased Brn2 levels may now be able to overcome the arrest of Schwann cell development in Oct6-deficient mice.

In agreement with functional redundancy between Oct6 and Brn2, the overexpression of a Brn2 transgene rescues the developmental delay of Oct6-deficient Schwann cells. Simultaneous loss of Oct6 and Brn2, on the other hand, causes affected Schwann cells to persist in the promyelinating stage much longer than if they were only Oct6-deficient. Accordingly, compound mutant mice show a myelination delay of more than 8 weeks, compared to the 2 weeks in Oct6-deficient mice. Nevertheless, some residual capacity to enter the myelination stage is retained. At present, it is unclear why that is, but Oct-1, a ubiquitously expressed POU homeodomain protein distantly related to Oct6 and Brn2, could be responsible. In contrast, Brn-5, another POU homeodomain protein that is expressed in the Schwann cell lineage, was excluded as a compensating factor because it is predominantly expressed in myelinating Schwann cells and was unable to rescue the Oct6 phenotype when overexpressed as a transgene in Oct6-deficient Schwann cells.

It has long been controversial whether Oct6 (and by analogy Brn2) functions as a transcriptional activator or repressor. Differential display studies on the sciatic nerves of wild-type and Oct6-deficient mice have preferentially identified genes that are downregulated in the absence of Oct6, thus pointing to a primarily activating role of Oct6 in Schwann cells, although it has not been established that the identified genes (including the genes for the fatty acid transporter P2, the amino acid transporter tramdorin1, and the LIM domain protein cysteine-rich protein (CRP)2) are directly activated by Oct6.

There is very good evidence that the Krox20 transcription factor is under the direct positive control of Oct6. Studies in transgenic mice have identified an

enhancer 35 kb upstream of the transcriptional start site of the mouse Krox20 gene that is responsible for Krox20 expression in Schwann cells and that at the same time depends for its activation on Oct6 binding. Interestingly, Oct6 is not the only transcription factor that binds to the enhancer. There also are several recognition elements for Sox10, which is still strongly expressed in Schwann cells during the promyelinating stage. The enhancer is cooperatively activated by Oct6 and Sox10, and it is likely that *in vivo* cooperative activation is responsible for the dramatic upregulation of Krox20 expression in promyelinating Schwann cells.

It is also assumed that Oct6 is not only needed for the progression from the promyelinating to the myelinating stage but has additional, possibly repressive, functions during the early phases of myelination. This assumption is supported by the continued Oct6 expression in young myelinating Schwann cells. However, proof of these additional functions in myelinating Schwann cells is difficult to obtain in the currently available mouse models. In older myelinating Schwann cells, Oct6 is eventually downregulated. Interestingly, Oct6 itself is involved in its downregulation, pointing to the existence of a negative autoregulatory loop.

Finalizing Myelination

The dramatic upregulation of Krox20 in promyelinating Schwann cells follows that of Oct6, in accord with the epistatic relationship between both transcription factors. One of the functions of Krox20 is to suppress the immature state. It inactivates the proliferative response of Schwann cells toward their major axonal mitogen neuregulin 1. At the same time, Krox20 makes the cells less susceptible to cell death. To achieve this, Krox20 prevents the c-Jun N-terminal kinase (JNK)-dependent phosphorylation of cJun which is required for both the proliferative and the apopotic response of immature Schwann cells. As a consequence, Krox20-deficient Schwann cells are arrested in the promyelinating stage, but in contrast to Oct6-deficient Schwann cells, the block is permanent. There is additional evidence that Krox20 is involved in the downregulation of its activator Oct6 because Oct6 expression continues in Krox20-deficient Schwann cells.

Krox20 is also actively involved in establishing the myelinating stage. As is evident from gene expression profiling, most myelin genes and lipid biosynthetic enzymes are under direct or indirect control of Krox20. Confirming the link between Krox20 and myelin gene expression, several human patients have been identified with heterozygous mutations in Krox20. These heterozygous mutations manifest themselves as congenital hypomyelinating neuropathies or peripheral neuropathies of the Charcot–Marie–Tooth type 1 and Dejerine-Sottas syndrome, which are more frequently observed for genetic disturbances of myelin genes, including MPZ, PMP22, and Connexin-32. The patients' phenotype is thus compatible with a function of Krox20 during terminal differentiation of myelinating Schwann cells.

Regulating Myelin Gene Expression

In myelinating Schwann cells, Krox20 directly binds to the Nab1 and Nab2 cofactors. This cofactor interaction is crucially important for Krox20 function because mice deficient for Nab1 and Nab2 show defects in peripheral myelination similar to Krox20-deficient mice. Although Nab1 and Nab2 have been shown to function as repressors of Krox20 *in vitro*, the observed phenotype in Nab1/Nab2-deficient mice suggests that *in vivo* they are coactivators as well as corepressors.

It is unlikely that the Krox20–Nab1/Nab2 complex alone is responsible for the activation of all myelin genes and key enzymes in lipid metabolism. Thus, Schwann cells express transcription factors that have been associated with the transcriptional regulation of lipid metabolism in other tissues, including sterol regulatory element binding protein (SREBP)1 and SREBP2, the δ isoform of C/EBP, and the δ isoform of peroxisome proliferator-activated receptor (PPAR). Their increasing expression levels during myelination is compatible with a similar role in Schwann cells. Several SREBP target genes have been shown to be synergistically activated by SREBPs and Krox20 in tissue culture.

Another transcription factor with potential relevance for the myelinating stage of Schwann cells is the still expressed Sox10. Although a role of Sox10 during peripheral myelination has not been formally proven *in vivo*, there are several lines of evidence to implicate Sox10 in the process. The first comes again from the analysis of myelin gene expression in tissue culture. Thus, Schwann-cell specific regulatory regions of the MPZ, MBP, and Connexin-32 genes are all activated by Sox10. In case of the MPZ gene, an intronic enhancer appears to mediate this activation and is synergistically activated by Sox10 and Krox20. The Connexin-32 promoter is likewise synergistically activated by Sox10 and Krox20. Sox10 and Krox20 may thus cooperate in the activation of many genes during myelination, just as Sox10 cooperated during earlier phases of Schwann cell development with other transcription factors such as Oct6 during Krox20 activation in promyelinating Schwann cells.

The second piece of evidence for a role of Sox10 in peripheral myelination comes from patients with peripheral myelinopaties. Some carry heterozygous mutations in Sox10. Others have been found to carry mutations in a region of the Connexin-32 promoter that binds Sox10 and mediates Sox10-dependent activation. A role in myelination has already been proven for Sox10 in the oligodendrocytes of the central nervous system. Similar to Schwann cells, oligodendrocytes express Sox10 from earliest time, but in contrast to Schwann cells, they do not depend on Sox10 until terminal differentiation because of the additional presence of the closely related and functionally redundant Sox9. When Sox9 is downregulated in oligodendrocytes immediately before terminal differentiation, Sox10 function becomes evident and myelination is disrupted in the central nervous system of Sox10-deficient mice.

Open Questions

Although many transcription factors have been identified over the last years that contribute to Schwann cell development, there are still obvious gaps in our knowledge. Transcription factors that drive the transition from the Schwann cell precursor to the immature Schwann cell are, for instance, still unknown. In addition, several transcription factor families with important roles as developmental regulators in other neural cell types are missing from the list of Schwann cell regulators. The most conspicuous example concerns cell-specific (class B) basic helix-loop-helix (bHLH) proteins which, as proneural proteins, influence fate decisions in central and peripheral neurons as well as central glia. Because both Id proteins as the negative regulators of bHLH proteins and class A proteins as the heterodimerization partners for class B bHLH proteins have already been detected in Schwann cells, it is only a matter of time before a Schwann-cell-specific bHLH protein is identified. Many transcription factors thus remain to be discovered to complete the picture.

More and more findings point to functional interactions between already identified transcription factors, as exemplified by the cooperative activation of myelin genes by Sox10 and Krox20. Such interactions and cross-regulatory activities are probably essential for Schwann cell development. It is highly unlikely that the regulation of Schwann cell development can be explained by the isolated actions of a handful of cell- and stage-specific transcription factors; rather, their combination and cross-talk with more widely expressed transcription factors drive Schwann cell development. A second challenge thus is to find out how Schwann cell transcription factors influence one another's activity and to establish an integrated picture of the transcriptional network in Schwann cells.

Given the fact that lineage progression in other cell types is strongly associated with changes in chromatin structure, it can be safely assumed that such epigenetic changes are also relevant to Schwann cell development. Up until now, the question of how transcription factors influence chromatin structure has been barely addressed in Schwann cells and will certainly be an issue in the future.

Last, transcriptional regulation of Schwann cell development is thought to reflect many of the principles that are active during remyelination following nerve injury. Several transcription factors have indeed been shown to contribute to both processes. However, whether transcriptional control of remyelination is indeed a recapitulation of the regulatory events during Schwann cell development still requires rigorous testing.

See also: Glial Responses to Injury; Neural Crest; Neural Crest Cell Diversification and Specification: Melanocytes; Neural Crest Cell Diversification and Specification: ErbB Role; Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells; Oligodendrocyte and Schwann Cell Identification Methods; Peripheral Nerve Regeneration: An Overview; Schwann Cells and Axon Relationship; Schwann Cells and Plasticity of the Neuromuscular Junction; Schwann Cell Morphology.

Further Reading

Bermingham JR, Scherer SS, O'Connell S, et al. (1996) Tst-1/Oct-6/SCIP regulates a unique step in peripheral myelination and is required for normal respiration. *Genes & Development* 10: 1751–1762.

Britsch S, Goerich DE, Riethmacher D, et al. (2001) The transcription factor Sox10 is a key regulator of peripheral glial development. *Genes & Development* 15: 66–78.

Ghislain J and Charnay P (2006) Control of myelination in Schwann cells: A Krox20 cis-regulatory element integrates Oct6, Brn2 and Sox10 activities. *EMBO Report* 7: 52–58.

Jaegle M, Ghazvini M, Mandemakers W, et al. (2003) The POU proteins Brn-2 and Oct-6 share important functions in Schwann cell development. *Genes & Development* 17: 1380–1391.

Jaegle M, Mandemakers W, Broos L, et al. (1996) The POU factor Oct-6 and Schwann cell differentiation. *Science* 273: 507–510.

Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

Le N, Nagarajan R, Wang JY, Araki T, Schmidt RE, and Milbrandt J (2005) Analysis of congenital hypomyelinating Egr2Lo/Lo nerves identifies Sox2 as an inhibitor of Schwann cell differentiation and myelination. *Proceedings of the National Academy of Sciences of the United States of America* 102: 2596–2601.

Monuki ES, Weinmaster G, Kuhn R, and Lemke G (1989) SCIP: A glial POU domain gene regulated by cyclic AMP. *Neuron* 3: 783–793.

Nagarajan R, Svaren J, Le N, Araki T, Watson M, and Milbrandt J (2001) EGR2 mutations in inherited neuropathies dominant-negatively inhibit myelin gene expression. *Neuron* 30: 355–368.

Nickols JC, Valentine W, Kanwal S, and Carter BD (2003) Activation of the transcription factor NF-kappaB in Schwann cells is required for peripheral myelin formation. *Nature Neuroscience* 6: 161–167.

Topilko P and Meijer D (2001) Transcription factors that control Schwann cell development and myelination. In: Jessen KR and Richardson WD (eds.) *Glial Cell Development: Basic Principles and Clinical Relevance,* 2nd edn., pp. 223–244. Oxford: Oxford University Press.

Topilko P, Schneider MS, Levi G, et al. (1994) Krox-20 controls myelination in the peripheral nervous system. *Nature* 371: 796–799.

Warner LE, Mancias P, Butler IJ, et al. (1998) Mutations in the early growth response 2 (EGR2) gene are associated with hereditary myelinopathies. *Nature Genetics* 18: 382–384.

Wegner M (2000) Transcriptional control in myelinating glia: Flavors and spices. *Glia* 31: 1–14.

Wegner M (2000) Transcriptional control in myelinating glia: The basic recipe. *Glia* 29: 118–123.

Neural Induction in Chicks

C Linker and C D Stern, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction – Neural Induction as a Complex Process

Embryonic development involves the progressive restriction of the fates of initially pluripotent cells to different cell types, arising at the correct locations in the embryo and at the right time. Surprisingly, only a few signaling pathways (as few as seven or eight different classes) appear to be involved in all of these cell fate choices. As in music, where an infinite number of tunes can be played by arranging only a few different notes in precise combination, order, and duration, the timing of the sequence of signaling events is critical in directing cells to their ultimate fates.

Over the past two decades, we have learned considerably about the roles of these different signaling pathways, in large part due to research on embryos of the South African clawed toad frog, *Xenopus laevis*. This species lays large eggs which do not change in volume during early development, permitting genes to be misexpressed or their function inhibited in the entire embryo or a part thereof. This is done by injection of RNAs encoding a gene product of interest, or inhibitory, mutant versions of it (including antisense morpholino oligonucleotides), into one of the cells (blastomeres) arising during the first few cell divisions, and studying the consequences at a later stage of development. This approach has been particularly powerful in identifying the critical signaling pathways that influence cell fate choices in the early embryo. However, because all signaling pathways are used repeatedly during development, the usefulness of the misexpression approach is limited to the study of the very earliest developmental events. It was this approach that led to the identification of fibroblast growth factors (FGFs) and Nodal (a member of the transforming growth factor-β family) as critical players in the induction of the mesoderm, when misexpression and loss of function approaches were combined with the 'animal cap assay' designed by Pieter Nieuwkoop to study this particular inductive event in the late 1960s. The animal cap assay involves isolation of the animal pole of the blastula-stage embryo (which does not normally contribute to mesoderm), which is then cultured in a neutral medium – the cap can be cut from embryos injected with control RNAs or with experimental constructs, or can be treated with protein factors added to the medium.

Although the animal cap assay was designed specifically to study mesoderm induction (precisely because it does not normally contribute to this tissue), it has also been extended to study neural induction and to other, somewhat later developmental events. This approach, along with other experiments (summarized in Figure 1), led to the influential 'default model,' which proposes that inhibition of bone morphogenetic protein (BMP; another member of the transforming growth factor-β family) signaling is sufficient to cause animal pole cells to adopt a neural fate. However, BMP signaling is also critical for other patterning and inductive events that take place either earlier than or at the same time as neural induction, making it rather difficult to separate these processes from a direct role purely in neural induction. In particular, inhibition of BMP is first required to specify the 'dorsal' side of the whole embryo (where gastrulation will begin) and where Spemann's organizer will arise. Then, different levels of BMP specify different types of mesoderm: the higher the level, the more ventral the mesoderm. The mesoderm arises from the equatorial region of the embryo. Therefore, treatments that alter the levels of BMP signals from the earliest stages of development will affect all of these processes and the assay will reveal the cumulative effects of the treatment on all of the intervening events as well as any indirect consequences.

An additional problem with studying neural induction using this approach is that the animal cap does contain cells whose normal fate is to contribute to part of the nervous system (especially neural crest, but also some central nervous system, depending on the size of the cap excised). It is therefore impossible to exclude the possibility that by the time it is isolated, the cap has already received some of the signals required by cells to be specified as neural. Thus, the animal cap assay makes it impossible to distinguish a truly 'instructive' induction from a merely 'permissive' event that is part of a more complex cascade.

It is therefore not surprising that, although there is considerable evidence consistent with the proposals of the default model (summarized in Figure 1), some experiments in both the chick and in the frog (especially those that do not rely exclusively on early misexpression and on the animal cap assay) are not so consistent and reveal more complexity (summarized in Figure 2). In *Xenopus*, inhibition of BMP signaling in one early blastomere (A4) whose progeny does not include any part of the nervous system is not sufficient to induce expression of neural markers. In addition, when the FGF pathway is inhibited with a dominant-negative FGF receptor or with a chemical

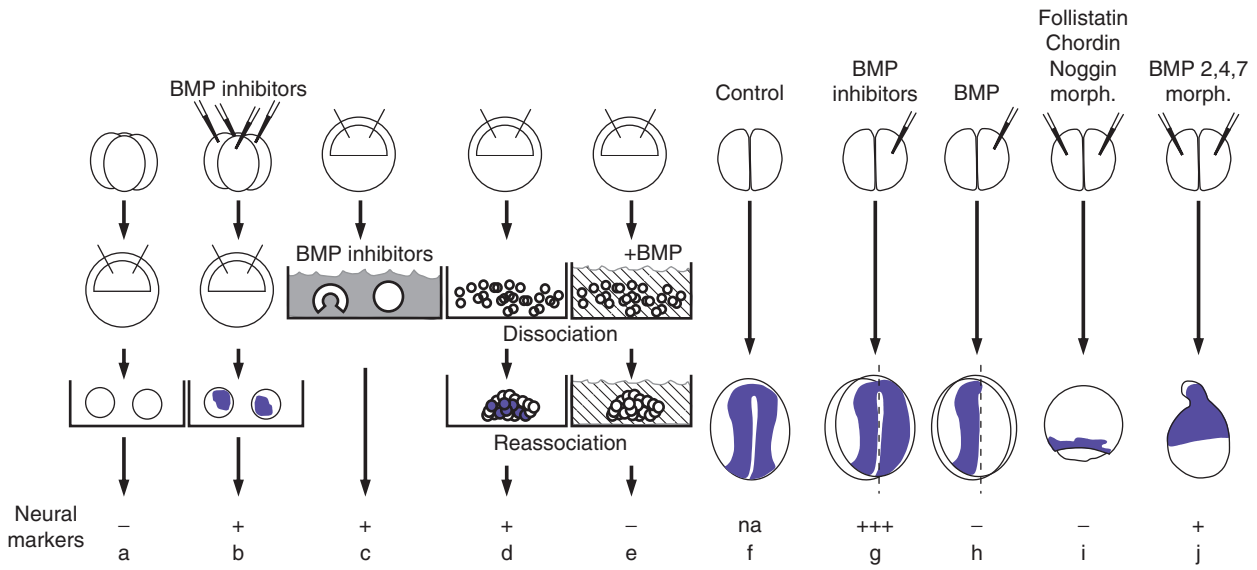

Figure 1 Experimental evidence consistent with the default model. Control animal caps (a) will develop as epidermis; neural markers (+, expressed; –, not expressed; na, not applicable) are expressed when caps are excised from embryos injected with constructs encoding bone morphogenetic protein (BMP) inhibitors (b), or when a BMP antagonist is added to the medium in which caps are cultured (c). Dissociation followed by reaggregation of animal cap cells causes neural differentiation (d), while addition of BMP proteins to the culture medium reverts this effect (e). In whole embryos, the neural plate (blue) develops from the dorsal side (f); when BMP inhibitor constructs are injected into one of the blastomeres at the two-cell stage, the neural plate is expanded in the injected side (g), while injection of BMP at the same stage leads to a drastic reduction in the size of the neural plate (h). Early depletion of BMP antagonists, by injection of antisense morpholinos against Noggin, Chordin, and Follistatin in the whole embryo, leads to expression of BMP molecules in the neural territory and causes severe ventralization of the whole embryo and almost complete abrogation of the neural plate (i). The opposite experiment (j), depletion of three BMPs (BMP2, BMP4, and BMP7), which effectively eliminates BMP signaling in the whole embryo, produces dorsalized radial embryos expressing neural markers throughout the ectoderm.

inhibitor (SU5402), animal caps from embryos injected with a BMP antagonist no longer form neural tissue, suggesting that FGF signals are required in addition to BMP inhibition to specify neural fates. Interestingly, in the ascidian *Ciona intestinalis* (a basal chordate), it is FGF rather than BMP inhibition that is responsible for inducing neural fates, suggesting that a requirement for FGF in neural induction is an ancestral feature of this process.

In the chick embryo it is possible to introduce DNA constructs or morpholinos at virtually any chosen stage of development and in precisely selected groups of cells. In addition, the chick embryo contains a very large region which never contributes to any part of the nervous system in normal embryos, although it is competent to do so in response to grafts of the organizer (Hensen's node, at the tip of the primitive streak, which is analogous to the amphibian dorsal lip of the blastopore). Experiments (Figure 2) combining grafts of the organizer with a time-course analysis of gene expression at intervals following the graft first revealed that neural induction is a relatively long process: if an organizer is grafted and then removed after different periods of time, it can be shown that a period of contact as long as 13 h is required before the cells that had been exposed to organizer signals express definitive central nervous system (CNS) markers (such as the transcription factor Sox2). Within a shorter time, the organizer does induce transient expression of some markers, such as Sox3 (which appears after about 3 h) and the 'early response to neural induction' marker (ERNI; which is induced after just 1 h). Despite the induction of these markers, however, cells expressing them revert to a nonneural fate and do not express Sox2 or form a neural plate if the organizer is removed. Inhibition of BMP after 5 h of contact with the organizer can, however, stabilize the expression of Sox3, but is still not sufficient for induction of either Sox2 or of a neural plate.

The evidence suggesting that BMP inhibition is not sufficient for neural induction in the chick embryo also includes the fact that the timing of expression of the BMP antagonists implicated in this process in the frog (such as Chordin, Noggin, and Follistatin) does not entirely fit with the period during which the chick organizer is able to induce neural markers: Chordin is expressed in the organizer at a stage when this tissue can no longer induce, while Noggin and Follistatin are not present at all at the inducing stages and appear only in nonorganizer tissue at later stages. Likewise BMP4 and BMP7 (the BMPs implicated as the inhibitors of neural fate in the frog) are not ubiquitously present

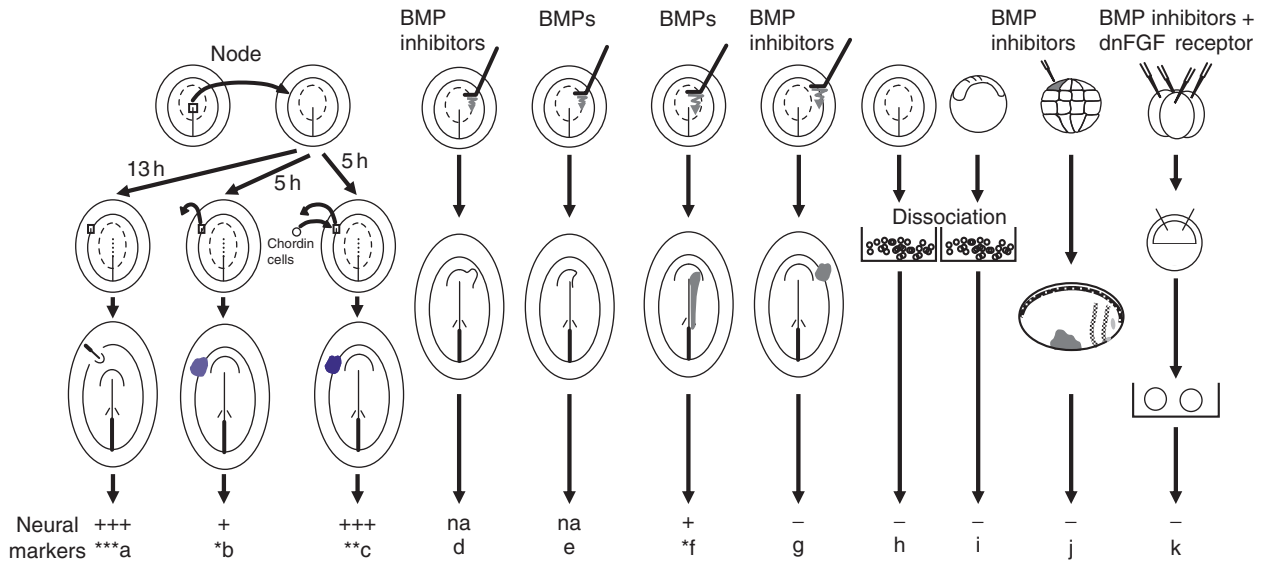

Figure 2 Experimental evidence inconsistent with the default model. Analysis of the time course of neural induction, after grafts of the organizer (Hensen's node), shows that while 13 h of contact between the organizer and the host ectoderm is required for full neural induction (a), 5 h of contact induces transient expression of early neural markers (e.g., Sox3 and ERNI) (b). Contact for 5 h is also required to sensitize the ectoderm to bone morphogenetic protein (BMP) inhibitors such as Chordin (c). Manipulation of BMP levels at the border of the neural plate affects the size of the neural plate: when Chordin (BMP antagonist) is expressed at this border, the neural plate is expanded (d); misexpression of BMP4 at this site causes narrowing of the neural plate (e). Overexpression of BMP in the neural plate territory does not affect the expression of early neural markers (f) (although it does abolish Sox2 expression at a later stage), while inhibition of BMP in the nonneural ectoderm does not induce neural markers (g). Dissociation of either chick epiblast (h) or of zebra fish animal caps (i) does not lead to neural differentiation. In *Xenopus*, injection of BMP antagonists at the 32-cell stage into a blastomere (A4) that is fated to give rise predominantly to ventral epidermis is not sufficient to trigger neural marker expression in the progeny of the injected cell (j). Animal caps excised from embryos injected with both a dominant-negative fibroblast growth factor (dnFGF) receptor and BMP antagonists (Chordin or Noggin) do not express neural markers. *, Expression of only early neural markers; **, expression of early and definitive neural markers; ***, induction of a neural plate; na, not applicable.

throughout the early embryo before neural induction as they are in *Xenopus*. Finally, inhibition of BMP (by Chordin, Noggin, or intracellular antagonists) is not sufficient to induce any neural marker in the chick, even when combined with or preceded by application of FGFs to the same cells. However, inhibition of BMP is undoubtedly required for neural induction, since when BMP4 is misexpressed in the prospective neural plate it prevents expression of Sox2 (but not Sox3). Moreover, when BMP levels are manipulated at the border of the neural plate (prospective neural crest/placode territory, somewhat equivalent to the animal cap of *Xenopus*) the size of the neural plate is affected: higher levels of BMP narrow the neural plate, while lower levels expand it.

Experiments using explants of prospective neural plate or prospective epidermis from pregastrula-stage embryos, either cultured alone or in the presence of various peptide growth factors, suggested that in addition to FGF and BMP inhibition, inhibition of Wnt signaling may also be important. However, in whole embryos, it is still impossible to impart neural identity to competent cells even with a combination of all three of these signals.

These results suggest that neural induction is likely to be the result of a more complex and relatively long cascade of sequential signaling events, involving both FGF signals and inhibition of BMP, but also other signals which are still unknown. They also suggest that cells at the anterior and lateral borders of the neural plate are particularly sensitive to the level of BMP signaling.

The Timing of Neural Induction

The classical experiments of Spemann and Mangold (published in 1924) and subsequent studies in various species suggested that neural induction probably occurs at the gastrula stage, after the mesoderm has been specified and patterned to some extent. However, experiments in the chick using explanted tissues as well as studies of the expression patterns of very early 'preneural' markers suggested that cells are already started along the cascade that will lead to the acquisition of neural fates at much earlier stages, even preceding gastrulation and before the organizer forms. ERNI and Sox3 are two such markers. Both are induced by, and require, FGF, but neither this

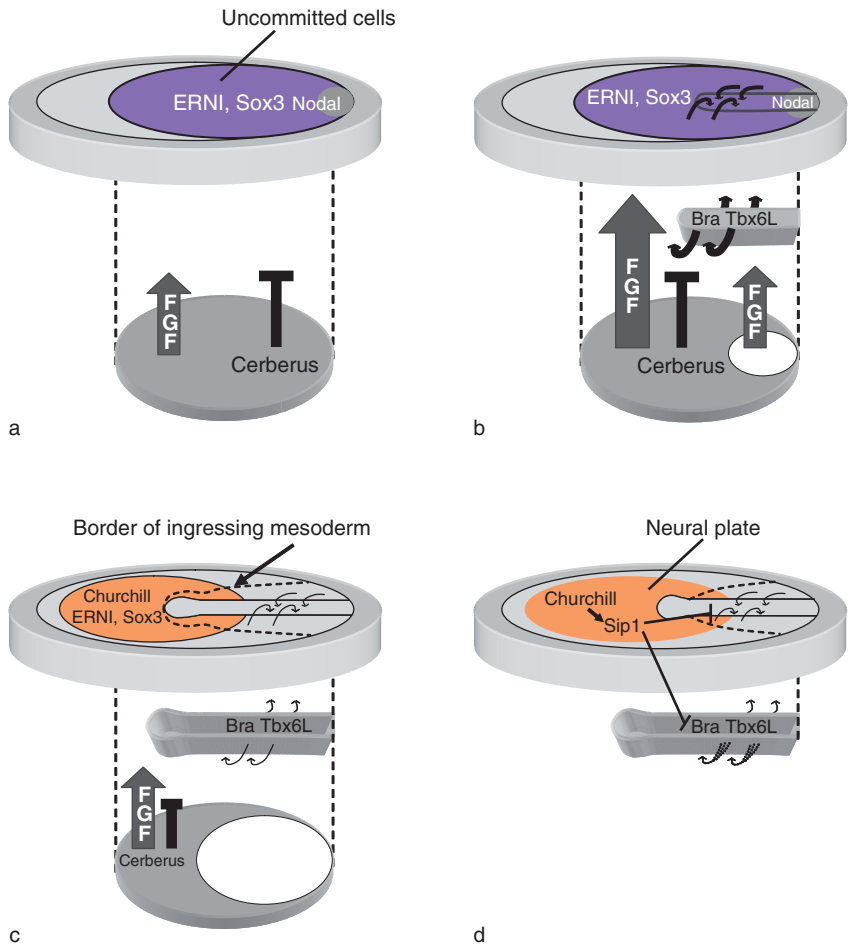

Figure 3 A model for the sequence of events that initiate neural induction in the chick embryo. (a) Before gastrulation, the epiblast (upper disk) is induced to express the early, 'preneural' markers ERNI and Sox3 by fibroblast growth factor (FGF) secreted by the extraembryonic endoderm (hypoblast; lower gray disk). At this time the posterior (right) margin of the epiblast expresses Nodal, but its activity is blocked by its inhibitor Cerberus, which is also produced by the hypoblast. (b) Gastrulation begins when the hypoblast (and Cerberus expression) moves away from the Nodal-expressing region: this releases Nodal signaling, which cooperates with FGF to induce the formation of the primitive streak. Cells ingress from the surface of the epiblast into the primitive streak to form mesoderm and endoderm. (c) Continued exposure of the epiblast that still remains on the surface to FGF from the hypoblast now induces expression of the zinc finger transcription factor Churchill. (d) A target gene activated by Churchill is *Sip1*, which encodes another zinc finger transcription factor that blocks two genes (*Brachyury* and *Tbx6L*) required for ingression of cells into the primitive streak. This effectively ends the process of gastrulation next to the anterior parts of the primitive streak and therefore causes *Churchill*$^{+}$/*Sip1*$^{+}$ cells to remain on the surface and become committed to a neural plate fate.

signal nor the expression of these early markers is sufficient for neural plate formation (Figure 3). At this early stage, FGF8 (the most likely candidate for an initial inducing signal) is expressed in an extraembryonic tissue (called the hypoblast in chick, rabbit, and human embryos, the equivalent of which is the anterior visceral endoderm (AVE) in the mouse); this tissue can induce the expression of these early preneural markers when grafted to an ectopic site, even though neither the chick hypoblast nor the mouse AVE can induce a neural plate in the same assay. It is therefore likely that neural induction is initiated by signals from the hypoblast/AVE but that acquisition of a neural plate fate occurs only after the initial induction has been reinforced and completed by signals from other tissues, most likely to be the organizer itself and its derivatives (such as the notochord, prechordal mesendoderm, and perhaps paraxial mesoderm). Importantly however, the early organizer contains both the early and the later signals (or the cells that are a source of the latter), explaining why grafts of the organizer are sufficient to induce a complete nervous system even if it does not normally do all this in the embryo.

FGF8 can induce expression of ERNI, Sox3, and Churchill (a zinc finger transcription factor, which in normal embryos is expressed in the forming neural plate from the end of gastrulation) in a competent

region of the chick embryo that is not fated to contribute to the neural plate. However, these three markers are induced at different times following exposure to FGF: ERNI is induced after 1–2 h and Sox3 after 3 h, and Churchill induction requires 4–5 h. Loss-of-function experiments suggest that both Sox3 and Churchill are functionally required for neural induction to produce a neural plate expressing Sox2. These findings suggest that even a single initial signal can trigger events that follow each other in time. Although we do not yet understand the mechanisms that regulate the timing of onset of expression of these different markers, this is likely to involve complex interactions and feedback mechanisms within the responding cells. It may well turn out that, as in music, the duration of each signal is just as important as the sequence in which they are experienced by cells.

Not Just a Decision between Neural Plate and Epidermis

In the experiments of Spemann and Mangold and those that followed them, the organizer was usually grafted ventrally, adjacent to a region of the embryo whose fate is normally to give rise to epidermis. This led to the generally assumed view that neural induction is a decision between these two fates: skin as opposed to nervous system. However, a study of the expression and functions of early 'preneural' markers suggested that, in addition, neural induction also requires cells to decide between mesodermal and neural fates, close to the midline of the embryo. Functional studies of transcription factor Churchill suggested that it encodes a transcriptional activator, the target of which is another zinc finger transcription factor called Sip1 (Smad-interacting protein-1), which in turn inhibits expression of the primitive streak gene *Brachyury*, required for ingression of cells to form mesoderm. Therefore the onset of Churchill expression marks cells that will no longer ingress to form mesoderm and which will therefore remain on the surface of the embryo, where they can form a neural plate if they receive the remaining signals required to complete the process. Interestingly, Sip1 was originally isolated as a direct partner of Smad-1, a critical effector of BMP signaling, suggesting perhaps that Churchill/Sip1 are part of a mechanism required to sensitize cells to levels of BMP signals, perhaps explaining why cells that have not been exposed either to an organizer or to FGF signals for at least 5 h do not appear to be responsive to BMPs or their antagonists. Thus, neural induction comprises not only a decision between neural plate and epidermis at its lateral boundary, but also the decision to stop ingressing (gastrulation) so that cells can remain on the surface to give rise to neural plate, at the midline. These results are summarized as a model in Figure 3.

Importantly, exposure to FGF and absence (or low levels) of BMP are both required along with Nodal signaling to initiate gastrulation and for specifying the organizer. Although the hypoblast/AVE do produce FGF8 and BMP antagonists, both of these tissues paradoxically inhibits primitive streak formation in both chick and mouse embryos. This is due to their expression of Cerberus and Lefty-1, which inhibit Nodal. Thus, FGF signals are required in parallel to specify mesendoderm and for the initiation of neural induction, but different cells within the epiblast are destined to form mesendoderm and neural plate. It is the duration and timing of these signals as well as the position of the cells when they receive them that determine which of these fates (mesendoderm, neural plate, or epidermis) they will adopt.

Taken together, these findings reveal several distinct stages in the neural induction process: first, an FGF signal induces, sequentially, expression of ERNI, Sox3, and Churchill. Then, Churchill (via Sip1) inhibits ingression of cells, preventing them from forming mesendoderm and causing them to stay on the surface. These surface cells become sensitive to both BMP inhibition and to other signals required for formation of a Sox2-expressing, mature neural plate. BMP inhibition stabilizes the expression of early markers (Sox3) but is not sufficient to induce Sox2.

Conclusions – Neural Induction Is Not Yet a Solved Problem

The experimental findings discussed here reveal that neural induction is likely to be the result of a sequence of signaling events rather than a single signal. Although BMP inhibition is required for cells to acquire a neural plate fate, this is insufficient, and other signals, including FGF (but not only this factor), are also required. Moreover, formation of a neural plate also requires mechanisms that cause cells to remain superficial in the embryo and prevent them from gastrulating. While some of these signals are starting to be uncovered, it is becoming clear that there are several other pathways that must be involved. These other pathways still remain to be found, as are the precise interactions between these different pathways, which are likely to be very complex.

The timing and spatial pattern of expression of the earliest marker, ERNI – which is initially expressed in a large territory, including the entire future neural plate, and is downregulated just before the appearance of Sox2 expression – raise the possibility that ERNI might act as an antagonist of commitment to neural plate identity. However, BMP inhibitors and Wnt

antagonists (either alone or when combined with each other and with FGF) are not sufficient to downregulate ERNI expression. Among the many questions that still remain open about neural induction, it will be particularly interesting to explore whether the signals that cause the downregulation of ERNI expression are the missing signals required alongside FGF and BMP inhibition to commit cells to a neural plate fate, and which trigger initial expression of Sox2.

See also: Bone Morphogenetic Protein (BMP) Signaling in the Neuroectoderm; Neural Patterning: Eye Fields.

Further Reading

Bainter JJ, Boos A, and Kroll KL (2001) Neural induction takes a transcriptional twist. *Developmental Dynamics* 222: 315–327.

Linker C and Stern CD (2004) Neural induction requires BMP inhibition only as a late step and involves signals other than FGF and Wnt antagonists. *Development* 131: 5871–5881.

Sheng G, Dos Reis M, and Stern CD (2003) Churchill, a zinc finger transcriptional activator, regulates the transition between gastrulation and neurulation. *Cell* 115: 603–613.

Stern CD (2001) Initial patterning of the central nervous system: How many organizers? *Nature Reviews Neuroscience* 2: 92–98.

Stern CD (2005) Neural induction: Old problem, new results, yet more questions. *Development* 132: 2007–2021.

Streit A, Berliner AJ, Papanayotou C, et al. (2000) Initiation of neural induction by FGF signaling before gastrulation. *Nature* 406: 74–78.

Wilson SI and Edlund T (2001) Neural induction: Toward a unifying mechanism. *Nature Neuroscience* 4: 1161–1168.

Wilson SI, Rydstrom A, Trimborn T, et al. (2001) The status of Wnt signaling regulates neural and epidermal fates in the chick embryo. *Nature* 411: 325–330.

Neural Integrator Models

M S Goldman, University of California at Davis, Davis, CA, USA
A Compte, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain
X-J Wang, Yale University School of Medicine, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Since the 1906 publication of Charles Sherrington's *The Integrative Action of the Nervous System*, the concept of neurons as integrators of synaptic inputs has become a central tenet of neuroscience. Neural circuits integrate information in various ways: summating spatially, receiving converging inputs from different sensory modalities, bridging sensory and motor pathways, combining reward and cognitive signals, and so forth. Integration of information in the temporal domain, the topic of this article, has been increasingly recognized in recent years to be important for a variety of brain functions. Time integration is defined as the mathematical accumulation of an input signal so that the resulting output reflects the running total of the input. Networks in the nervous system can exhibit a great diversity in the 'integration time' during which they compute and store this running total. Typically, neurons are leaky devices with an integration time of tens of milliseconds. However, some neurons or neural networks can have a very short integration time constant (a few milliseconds or less); such 'coincidence detectors' can be activated only by synaptic events that are precisely synchronized. At the other extreme, neurons of certain brain areas behave as if their characteristic time constant was effectively infinite. Not only do they perform time integration of inputs without noticeable leak during stimulation but also they maintain the memory of the integrated signal after the stimulus is withdrawn, in the form of self-sustained persistent activity.

The first neural integrators were discovered in the oculomotor system in the 1960s when it was realized that motor command neurons signaled the velocity of eye movement, whereas downstream motor neurons encoded the eye position (Figure 1(a)). Therefore, an integration operation is required to convert velocity signals into position signals. Functionally, such neural integrators are well suited to implement oculomotor control, just like integrators commonly used in feedback control systems by engineers. Similarly, in the head direction system, which is part of the spatial navigation system and signals the animal's directional heading, the inputs from the semicircular canals encode the head velocity, which is integrated by head direction cells into the positional signal that conveys the orientation of an animal's head in space (Figure 1(b)). When the head is still, head direction cells show elevated and directionally tuned persistent activity in the absence of external inputs.

These examples illustrate two encoding schemes of neural integrators – 'rate code' and 'location code.' In a rate code, exemplified by the oculomotor system (Figure 1(a)), the integrator accumulates the signal in the firing rate of the neurons and the persistent firing rate of each neuron varies monotonically with the encoded feature. By contrast, in a location code as used in the head direction system (Figure 1(b)), each neuron exhibits a bell-shaped tuning curve of an encoded feature with peak response at a particular 'preferred' feature. When a particular value of the encoded feature (e.g., a particular head direction) is being stored, neurons with preferred features similar to this value will be highly excited and neurons with more dissimilar preferred features will be less excited. This creates a localized activity profile across the population known as a 'bump' of neural activity, and the encoded feature (e.g., the current direction of the head) can be read out from the peak location of this bump. If velocity signals smoothly move this bump of activity around the neural population, then the location of the bump can represent the integral of the velocity input. In other words, the stimulus feature is specified by 'which neurons are active' in a location code and by 'how high the neural activities are' in a rate code. Mathematically, a perfect integrator is described by $dX/dt = I(t)$, where $I(t)$ is the input, and X is either the peak locus of network activity (in a location code) or the neural firing activity level (in a rate code).

Based on the mathematical properties of integrals in the sense of calculus, neural integrators are expected to be capable of certain computations of special importance to higher cognitive functions. For example, if a signal suddenly vanishes, its integral remains at exactly the same level as that before disappearing. This is suggestive of a memory system. Indeed, in a delayed discrimination task in which a monkey is required to actively remember the frequency of a vibrotactile stimulus during a memory period of a few seconds, neurons in the prefrontal cortex (a cortical area known to play a critical role in working memory) have been found to exhibit graded persistent activity that changes monotonically with the intensity of the transient somatosensory stimulus (Figure 1c). As another example, the integral

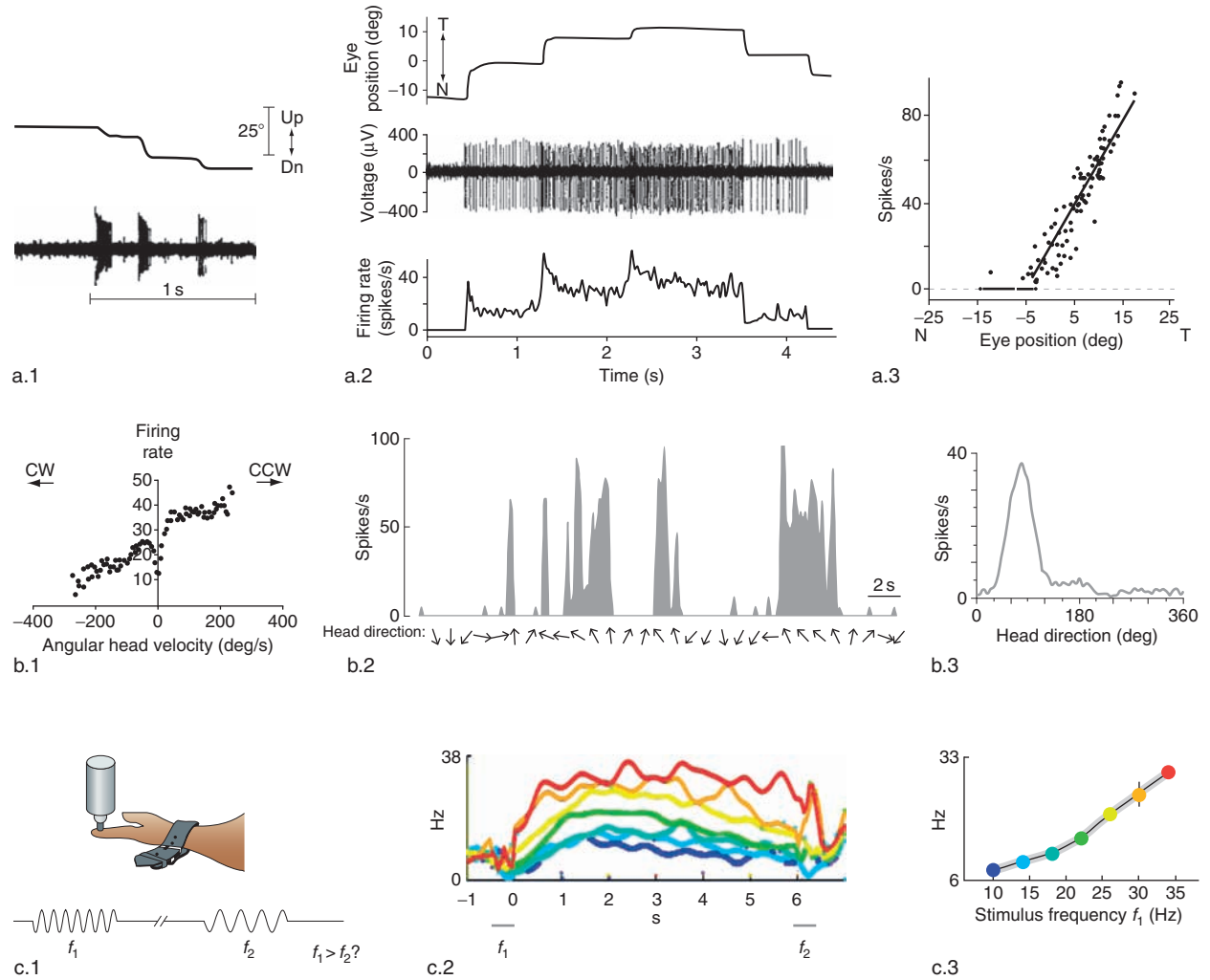

Figure 1 Experimental evidence for neural integrators in various systems reveals two principal coding schemes: 'rate code' and 'location code.' (a) In the oculomotor system, eye velocity-encoding motor commands from neurons in brain stem nuclei occur in the form of bursts at the time of each saccade but not during gaze fixation. However, neurons downstream fire tonically during fixations to maintain the eye muscles' tension and, thereby, the eye position (a.2). The neural integrator of the oculomotor system was hypothesized to explain how transient inputs as in a.1 (bottom) result in sustained responses as in a.2. Neurons in the oculomotor neural integrator show a rate code: The eye position is encoded in their average rate of sustained firing (a.3). (b) Input to head direction (HD) neurons comes from neurons in the brain stem that fire at the time of head rotations, with the firing rate encoding the angular velocity of head rotations. (b.1). HD neurons encode the integral of this signal and fire maximally only when the head points in a specific direction (b.2). The ensuing code, when plotting the average rate of the cell as a function of the HD, is a location code where neurons firing maximally encode the current HD (b.3). (c) In a task in which the vibration frequency f_1 of a tactile stimulus needs to be remembered over a delay period in order to compare it with another frequency f_2 (c.1), neurons in the prefrontal cortex present sustained firing during the delay period, which depends on f_1 (c.2; colors correspond to different values of f_1 as indicated in c.3). Using a rate code, these cells encode in their average delay firing rate the frequency f_1 (c.3). (a.1: top, eye position; bottom, extracellular voltage of motor command neuron) Reproduced from Kaneko CR, Evinger C, and Fuchs AF (1981) Role of cat pontine burst neurons in generation of saccadic eye movements. *Journal of Neurophysiology* 46: 387–408. (a.2, a.3) Reproduced from Aksay E, Baker R, Seung HS, and Tank DW (2000) Anatomy and discharge properties of pre-motor neurons in the goldfish medulla that have eye-position signals during fixations. *Journal of Neurophysiology* 84: 1035–1049. (b.1) Reproduced from Taube JS and Bassett JP (2003) Persistent neural activity in head direction cells. *Cerebral Cortex* 13: 1162–1172. (b.2, b.3) Adapted from Taube JS (1995) Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *Journal of Neuroscience* 15: 70–86. (c.1–c.3) Reproduced from Romo R, Brody CD, Hernández A, and Lemus L (1999) Neural correlates of parametric working memory in the prefrontal cortex. *Nature* 399: 470–473.

of a constant signal is a linearly increasing function. A neural integrator will thus be able to accumulate evidence in favor of alternative propositions in a decision process. When alternatives with a barely discernible difference in their sensory inputs are presented over an extended period of time, an integrator can accumulate the small differences over time until it can eventually discriminate the alternatives through

its ever growing output. This accumulation dynamics is at the heart of the 'drift diffusion' model for reaction time decisions in the psychological literature. Neurophysiological studies using behaving monkeys have found single neuron activity in the posterior parietal cortex that is correlated with accumulation of sensory information during perceptual decisions. Also, drawing on the response of a neural integrator to a constant input, a clock can be built. If the neural integrator is reset each time it reaches a prespecified value, the resets will occur at uniform intervals as the ticks of a clock. Some authors have hypothesized that this scenario represents a candidate neural mechanism by which time intervals can be estimated.

This body of experimental evidence reveals that integration is a fundamental neural computation in the brain, and quantitative neural models have begun to yield clues as to what likely mechanisms underlie this operation in neural circuits and what theoretical concepts need to be explored experimentally. In this article, we present an overview of the computational models of neural integrators. We first define and discuss key concepts that underlie the construction of these models, and then we present some representative models that operate through several specific mechanisms. Finally, we discuss challenges and perspectives for future research on neural integrators.

From Concepts to Cellular Mechanisms

Time integration involves two basic operations. First, during stimulation, a neural integrator needs to additively increase or decrease its output in response to excitatory or inhibitory input signals, respectively. Second, a neural integrator should have a memory capability to store the running total of inputs. Neither integration over a long timescale nor memory of the running total is obvious from a neurophysiological standpoint because leakage of neuronal currents out of the membrane tends to make a neuron forgetful. As a result, the duration over which a neuron can temporally sum incoming synaptic potentials is limited by its membrane time constant. To achieve long integration times, most models therefore incorporate tuned positive feedback that serves to offset the intrinsic leakage of membrane currents. This positive feedback is most often in the form of synaptic feedback in a recurrently connected network, but it has also been implemented as feedback loops in the intracellular signal transduction pathways within a single neuron.

To illustrate how positive feedback can lead to integration, consider a simple model of a neuronal pool with a mean firing rate r and a synaptic connection to itself of weight w (Figure 2a). A pulse of external input to the neuronal pool leads to an immediate change of firing rate. When the input terminates, the induced change in firing rate is sustained by reverberating input arriving through the recurrent synaptic connection. In this manner, the network both summates and stores external signals.

Quantitatively, integration by this simple network can be described by the firing rate equation:

$$\tau_{\text{neuron}} \frac{\mathrm{d}r}{\mathrm{d}t} = -r + wr + I(t) \qquad [1]$$

where r represents the mean firing rate of the neuronal pool, w is the strength of the synaptic connection, $I(t)$ is the external input to be integrated, and τ_{neuron} is the intrinsic decay time constant of neurons in the pool. The first term on the right-hand side represents the intrinsic neuronal 'leak' processes which, in the absence of feedback ($w = 0$), lead to rate decay with time constant τ_{neuron}. The second term, wr, gives the total positive feedback, which is proportional to the strength of the feedback connection and increases with the neural activity level. The effective time constant of the network is given by

$$\tau_{\text{integrator}} = \frac{\tau_{\text{neuron}}}{|1 - w|} \qquad [2]$$

For instance, when a constant external input is presented, activity in this model exponentially approaches a steady-state level (given by $I/(1-w)$) with time constant $\tau_{\text{integrator}}$. This means that the system cannot integrate an input I for a time much longer than $\tau_{\text{integrator}}$, beyond which the neural activity level becomes insensitive to the duration of the stimulus presentation. Moreover, after the input is turned off, the system decays back to zero, and hence the memory about the input is lost, with the same time constant. The system's integration time is fairly short in the absence of positive feedback (when $\tau_{\text{integrator}} = \tau_{\text{neuron}}$), but the integration time can be increased with stronger positive feedback (a larger w).

Perfect integration is achieved when the positive feedback balances the intrinsic neuronal leak processes (Figure 2a, $w = 1$) so that

$$\tau_{\text{neuron}} \frac{\mathrm{d}r}{\mathrm{d}t} = I(t) \Rightarrow r = \frac{1}{\tau_{\text{neuron}}} \int I(t)\mathrm{d}t \qquad [3]$$

This balancing of feedback and leak implies that such models require fine-tuning in order to achieve perfect integration. The tuning criterion can be viewed graphically by plotting the magnitude of the leak term (r) and the feedback term (wr) on the same set of axes (Figure 2, top row). For firing rates at which these terms balance, the network performs perfect integration. When feedback is less than leak, the stored total

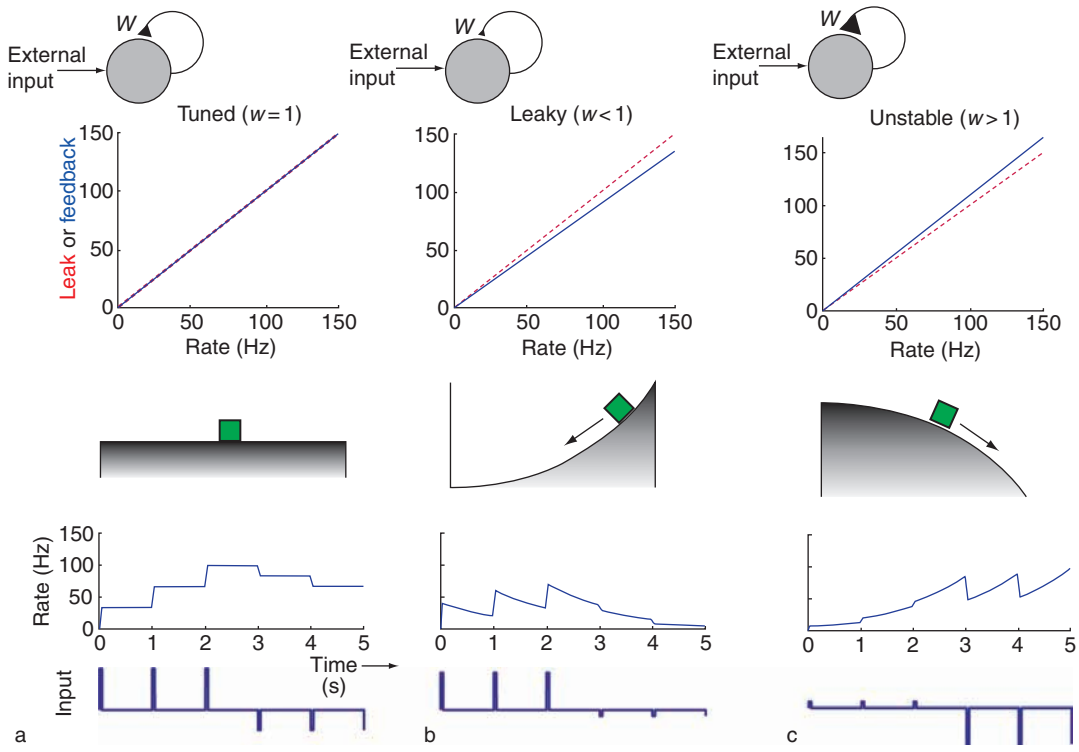

Figure 2 Positive feedback as a mechanism for neural integration. (a) Simple model of a neuronal population with positive feedback onto itself carried by a synaptic connection of weight *w*. (Top) When properly tuned, the feedback at any firing rate offsets the intrinsic leak processes that pull the firing rate back toward zero. Dashed red line, intrinsic leak; solid blue line, synaptic feedback. (Middle) The integrated total of external inputs can be conceptually represented by the location of a block along a surface. External inputs that are accumulated by the integrator move the block along the surface. When well tuned, the integrator can stably maintain any accumulated value, corresponding to the block not slipping along the flat surface. (Bottom) Accumulation of pulsed excitatory or inhibitory inputs into sustained firing rates. (b, Top) Model in which the synaptic weight is decreased so that the positive feedback is less than the intrinsic leak at all firing rates. Firing rate drift is proportional to the difference between the feedback and leak traces. (Middle) The model firing rate behavior can be represented by the location of a block on a curved surface, where the curvature is proportional to the firing rate drift in the absence of input. (Bottom) Decrease in the synaptic weight leads to exponential decay of firing rates in the absence of input. (c) A model in which feedback is too large leads to exponential growth of firing rates.

in the integrator leaks away at a rate proportional to the difference between the leak and feedback. In the absence of input, this leads to exponential decay on a timescale $\tau_{\text{integrator}}$ (leaky integrator; Figure 2(b)). Similarly, when feedback is greater than leak, activity exponentially increases with a time constant given by the same formula (unstable integrator; Figure 2(c)).

In a useful conceptual description, the network's output can be represented by the location of a block being pushed along a surface, with the movement of this block representing the dynamics of the network. External inputs slide the block with a velocity proportional to the size of the input. For a flat and level surface (Figure 2(a)), corresponding to a well-tuned network, the position of the block thus represents the integral of the external input. In the absence of input, the block stays in place, corresponding to successful maintenance of the accumulated input. For a mistuned network, the slope of the surface is given by the degree of mistuning – that is, by the difference between the feedback and leak. In this case, even in the absence of input, the position of the block drifts with a velocity proportional to the mistuning. For the leaky network, the mistuning brings the activity back to the origin (Figure 2(b), bottom), whereas for the unstable network, the mistuning leads to unbounded increase in activity (Figure 2(c), bottom).

Although the previously discussed model is highly simplified, it captures the essence of the positive feedback mechanism used by more biologically detailed network models. Because of the strong coupling between neurons that mediates the positive feedback in such networks, only certain activity patterns can be stably maintained internally in the absence of external inputs. These activity patterns are typically assumed to form a continuous, one-dimensional line in the high-dimensional space of all possible firing

patterns; the distance along this line (analogous to the distance of the block along the surface in Figure 2) represents the accumulated total of inputs being stored by the integrator. Activity patterns not along the line decay rapidly to a nearby point along the line so that the line is said to be an 'attractor,' or 'line attractor,' of the networks dynamics.

Specific biological networks differ in their implementation of positive feedback, which may be through excitatory pathways or disinhibitory (hence, effectively excitatory) pathways. They also differ in the topographic organization of the pathways conveying this feedback and in the means by which tuning of this feedback is accomplished: in networks utilizing a rate code, the positive feedback provided to each neuron typically changes monotonically as input is accumulated, and the magnitude of feedback connections is tuned to ensure that activity does not leak or exhibit unstable growth. In networks that utilize a location code, the pathways that mediate the positive feedback are spatially structured to allow a localized bump of activity to be maintained, and tuning of the network is typically accomplished by making symmetric connections between neurons so that the bump of activity can be maintained equally well at any location. Finally, the biophysical substrate of the positive feedback can differ between models, from synaptic feedback in network models to feedback between intracellular variables such as intracellular calcium concentration, membrane potential, and conductance levels in single neuron integrator models. Next, we discuss each of these specific mechanisms and their applications to neural integrators.

Diversity of Model Scenarios

Network Integrator Models

Rate code It is often observed in neural integrator networks that two opposing populations of cells are responsible for accumulating the input. Inputs to these populations are typically arranged in a 'push–pull' manner such that when one population is excited (pushed) by the external input, the other population is inhibited (pulled). Both the oculomotor integrator, which is located in the brain stem, and neocortical integrator networks display such a push–pull organization. In the oculomotor neural integrator, external commands that encode the velocity of an intended eye movement are sent through excitatory connections to one cell population and through inhibitory connections to the opposing population. The network integrates these inputs into sustained differences in firing rate between the two populations, with the excited population exhibiting above-background firing rates and the inhibited population exhibiting below-background firing rates.

Such networks contain two potential synaptic pathways to mediate positive feedback. Excitation between neurons within a population could provide a source of excitatory positive feedback. Inhibition between opposing populations could provide a source of mutual-inhibitory positive feedback whereby the excited population inhibits the opposing population and thereby disinhibits itself.

In the simplest models of this circuit, two neurons representing the opposing populations are assumed to be connected by mutual inhibition that provides a disinhibitory positive feedback loop (Figure 3a, middle). Such a circuit requires tonic excitation to drive neuronal activity. Thus, external inputs (of vestibular origin) consist of two components: a tonic component responsible for the background firing rate of neurons and a modulated component that is excitatory to one population and inhibitory to the opposing population. Common, same-direction changes in the tonic background input to each population lead to a corresponding change in the background firing rates of the two populations but are not integrated over time (Figure 3(a), first step in inputs). Because the eye position in this system is determined by the difference in the outputs of the two populations, such inputs do not lead to a change in eye position. By contrast, oppositely directed changes in inputs to the two populations about this tonic level do lead to an integrated response. The population that receives increased external input exhibits an upward ramp in activity, whereas the opposing population exhibits a downward ramp (Figure 3(a), second step in inputs). The difference between these two outputs, which is a ramp of twice the slope of either individual population, represents the integral of the difference between the commands received by either population. An advantage of this network is that it properly integrates such differences while remaining robust against same-direction perturbations in input rates that might be caused by global changes in background rates. In linear network theory, the integration of input differences and nonintegration of common inputs are termed 'differential mode' integration and 'common mode' rejection, respectively.

Alternatively, integration in the oculomotor neural integrator could be instantiated by excitatory feedback within each population, as illustrated in Figure 2. This idea has been quantitatively examined using realistic conductance-based spiking neurons and synaptic connections. Figures 3(b) and 3(c) show a model of a single population of the goldfish oculomotor neural integrator network, based on finely tuned recurrent excitatory positive feedback. Neurons were

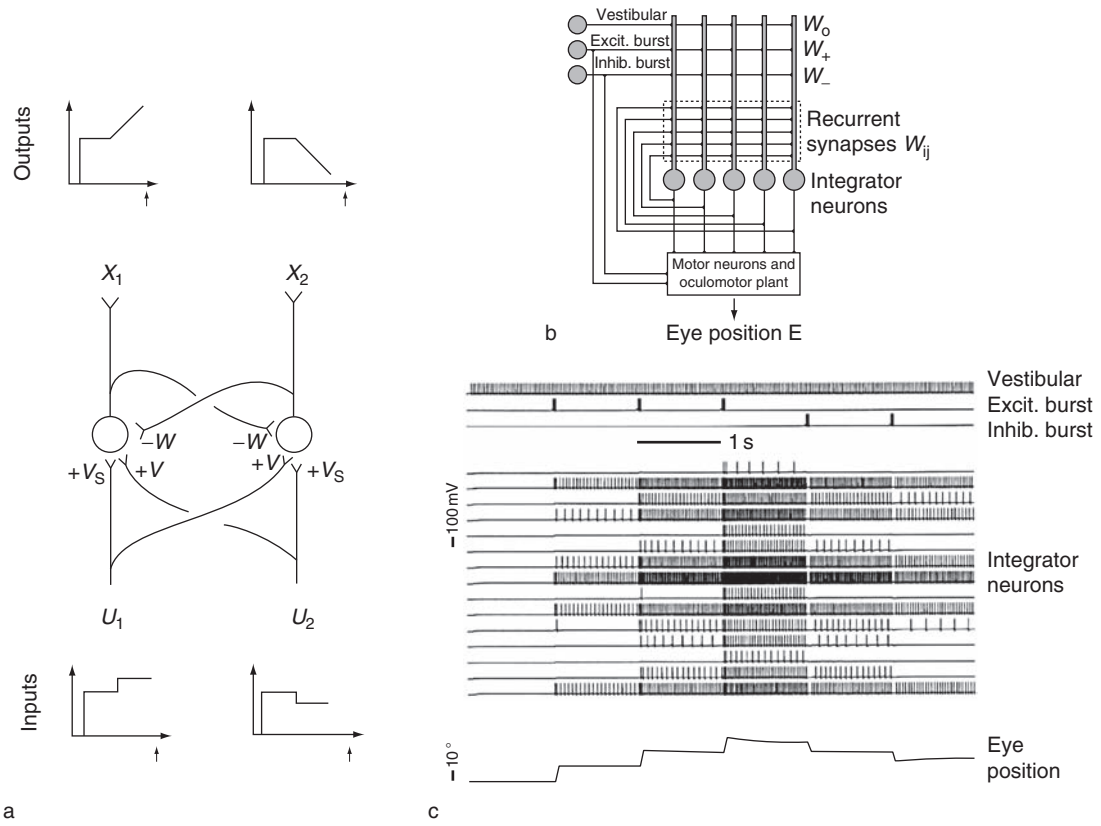

Figure 3 Rate code models of the oculomotor neural integrator. (a) Network model based on recurrent inhibition. Two neurons connected by mutual inhibition receive external inputs from ipsilateral and contralateral sources. Positive feedback is provided by each neuron inhibiting its neighbor and therefore disinhibiting itself. External inputs to the network consist of two components – a 'common' background input that is identical for each of the two neurons and a 'differential' command input that is excitatory to one neuron and inhibitory to the other. Changes in the common component (bottom, first step in inputs) lead to a change in background output levels but are not integrated (top traces). Differences in inputs (bottom, second step in inputs) are integrated by the network, in this case from a step into a ramp. In the oculomotor system on which this model was based, eye position is controlled by a difference between the outputs of two populations represented by these cells. Thus, the network output is robust to common changes in background inputs but is sensitive to differential inputs. (b, c) Integration by a network based on recurrent excitation. (b) Model network consisting of a population of neurons with recurrent excitatory synaptic connections. External commands from excitatory and inhibitory bursting neurons provide the velocity-encoding input for saccades, and vestibular signals provide tonic background input. Eye position is decoded from the network via a first-order model of the motor neurons and oculomotor plant. (c) Integration of bursts into steps of firing rate and eye position by the tuned network. Steps of firing rate in the absence of external command inputs are maintained by recurrent excitation. Different neurons begin firing at different eye positions due to differences in their background input levels. (a) Adapted from Cannon SC, Robinson DA, and Shamma S (1983) A proposed neural network for the integrator of the oculomotor system. *Biological Cybernetics* 49: 127–136. (b, c) Adapted from Seung HS, Lee DD, Reis BY, and Tank DW (2000) Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron* 26: 259–271.

connected in an all-to-all manner, with each neuron receiving excitatory and inhibitory command inputs to be integrated as well as tonic vestibular background inputs that differed from neuron to neuron (Figures 3(b) and 3(c)). As in Figure 2, active neurons exchanged positive feedback through recurrent synaptic excitation, allowing them to persistently maintain changes in firing rate in response to transient command inputs. However, unlike the simple model of Figure 2, the size of the active population also changed in response to inputs because neurons were sequentially recruited into the network with increasing network activity (with recruitment order determined by the amount of background input each neuron received). To offset the extra source of positive feedback that newly recruited neurons provided to their neighbors, the model assumed that synaptic transmission saturates as presynaptic firing rates increase. This idea of synaptic saturation balancing recruitment of neurons represented a significant step in the realistic modeling of rate-code integrators because it demonstrated how linear accumulation of inputs could occur even in systems with realistic nonlinearities. Experimentally, studies of the goldfish

oculomotor integrator have suggested that integration in this system is generated within each population of the integrator and not through recurrent inhibition between populations. This result is consistent with the idea of integration through excitatory positive feedback, but it is unknown whether single neurons in the system are endowed with intrinsic properties that could also contribute to the integration.

Location code Figure 4 shows schematically a typical neural integrator model in which the encoded feature is specified by the peak location of the network activity profile (location code). Neural networks that show this type of encoding has been used to model many neurobiological computations, from encoding of spatial location by place cells in the hippocampus to spatial working memory in dorsolateral prefrontal cortex. The realization of a neural integrator that utilizes a location code requires two ingredients. First, a continuous family of localized persistent activity patterns (bumps; Figure 4(b)) must be generated, each representing the graded value of the accumulated signal. Such localized activity profiles can be sustained by excitatory reverberation among neighboring neurons as the positive feedback mechanism (Figure 4(a)) or, alternatively, using tonic excitation combined with structured synaptic inhibition that sculpts the network activity profile. Second, a shifting mechanism must enable transient inputs to smoothly move the bump in such a way that the moving speed is linearly proportional to the input amplitude so that the final location of the bump accurately represents the time integral of the input (Figure 4(c)). The head direction (HD) system captures especially well these two aspects of neural integrators using a location code. When an animal turns its head, the angular velocity signal carried by vestibular inputs (Figure 1(b.1)) is integrated over time by HD cells into a positional signal, and the latter is sustained internally when the animal keeps the HD fixed (Figure 1(b.2)). HD cells are selective for angular HD according to a Gaussian (bell-shaped) tuning curve (Figure 1(b.3)).

Interestingly, evidence indicates that HD cells are generated in a neural circuit characterized by a paucity of local excitatory connections. Consistent with this observation, Figure 5 shows a model of HD neurons based on a mechanism without recurrent excitation in which the most active neurons push down their neighbors' activities through cross-directional inhibition rather than lifting up their own activities through isodirectional excitation. Baseline neuronal activities in this model are provided by constant-rate excitatory background inputs, and the direction-selective 'bump of persistent activity' reflects inhibition-driven decreases in activity below these externally driven levels. Moreover, the model surmises that this bump of activity is shifted around the network by 'rotation cells' consisting of two inhibitory cell populations (Figure 5(a)). When the HD is fixed, the inputs from the two inhibitory neural populations are balanced with each other, and the bump of activity in the excitatory neural population is maintained fixed (Figure 5(b), time epochs between input pulses). As the head turns, the angular velocity signal increases the firing of one inhibitory population while decreasing the firing of the other. The resulting asymmetric inhibitory inputs induce the activity pattern in the excitatory neural population to move at a constant speed proportional to the input amplitude (angular velocity) (Figure 5(b), time epochs during input pulses).

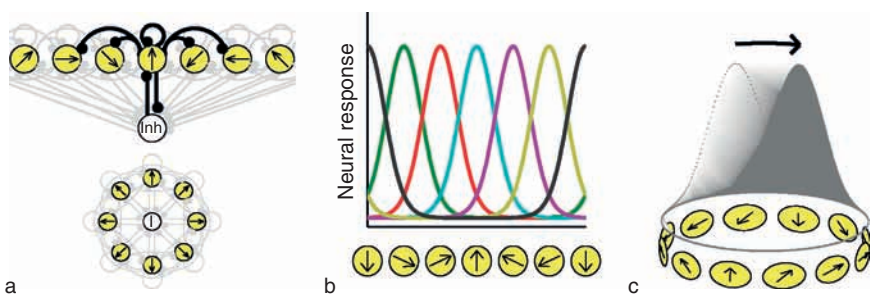

Figure 4 A bump attractor ring model as the substrate for a location code integrator. (a) Example of connectivity between neurons in a ring model attractor network. Recurrent excitation among near neighbors underlies positive feedback in the network, whereas growth of firing rates is controlled by nonselective feedback inhibition. When neurons at either end of the network (top) are made to coincide, a ring model network is obtained (bottom). (b) Given appropriate connectivity strengths, the network shows a whole family of stable neural activity states in the absence of external input ('bump' states, in different colors), each of which peaks at a different location on the network. The *x*-axis represents different neurons in the population, labeled by their preferred feature. For each curve, the *y*-axis values represent a possible distribution of firing rates across the population. (c) Bump states remain stably at their locations but can be shifted along the network smoothly by a 'push.' The final location of the shifting bump will be a read-out of the integral of this push force over the time during which it acts on the bump.

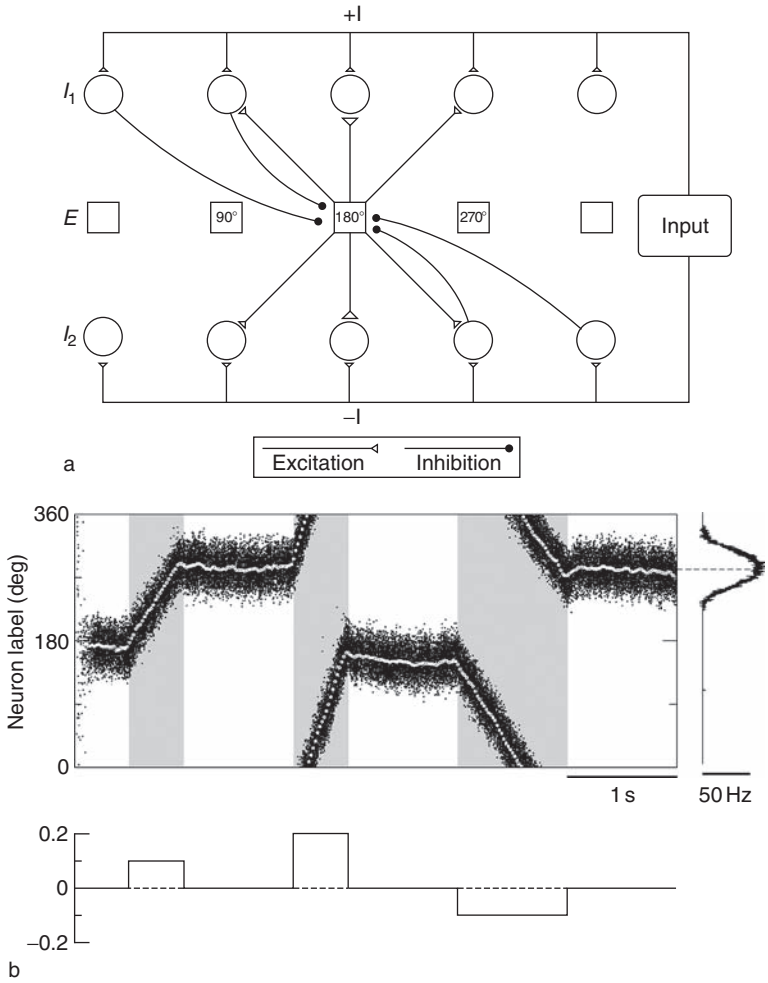

Figure 5 Time integration by a bump attractor network. (a) Model scheme. An excitatory neural network, encoding a directional angle (0–360°), receives inputs from two inhibitory neural populations that are balanced with each other when there is no input. A velocity signal increases input to one of the inhibitory populations (+I) and decreases input to the other (−I), leading to a bias for the excitatory network. (b) Network firing pattern (top) in response to a series of input signal steps (bottom). Neurons in the network are aligned along the *y*-axis, labeled by their preferred directional angles. The *x*-axis represents time. Each dot is a spike. The network activity pattern has the form of a bell-shaped profile (top right), with peak location encoding the directional angle (white line). In the absence of an external input, the directional information is maintained by a persistent firing pattern. An input induces the bump of activity to move at a speed proportional to the stimulus amplitude; hence, the network performs a time integral of the input. Simulation by P Song using the model published in Song P and Wang X-J (2005) Angular path integration by moving 'hill of activity': A spiking neuron model without recurrent excitation of the head-direction system. *Journal of Neuroscience* 25: 1002–1014.

Computation performed by the model is quite similar to an integral operation in calculus. When the input intensity is doubled (second vs. first input pulse), the hill of activity moves twice as fast. Moreover, the network can integrate both positive (first and second pulses) and negative (third pulse) inputs. In fact, with the third input of an amplitude half that of the second input, but lasting twice as long, the bump of activity moves back to the position before the second pulse, as expected by perfect integration.

Single Neuron Integrator Models

Neural integrators can also reside in the internal biochemical machinery of a single neuron. Experimental evidence that single cells can exhibit a range of graded persistent activity has been based on recordings in an *in vitro* preparation of rat entorhinal cortical slices (Figure 6(a)). A number of computational models have been built to explore the mechanistic basis of this activity (Figures 6(b)–6(d)). All of these alternative models implicate two principal actors: the muscarinically activated calcium-dependent nonselective cation current (I_{CaN}) and the concentration of calcium in the cytoplasm. The interaction between these two substrates can subserve integration through a positive feedback loop. During neural activity intracellular calcium increases because of the influx of calcium ions through high-threshold calcium channels, increased

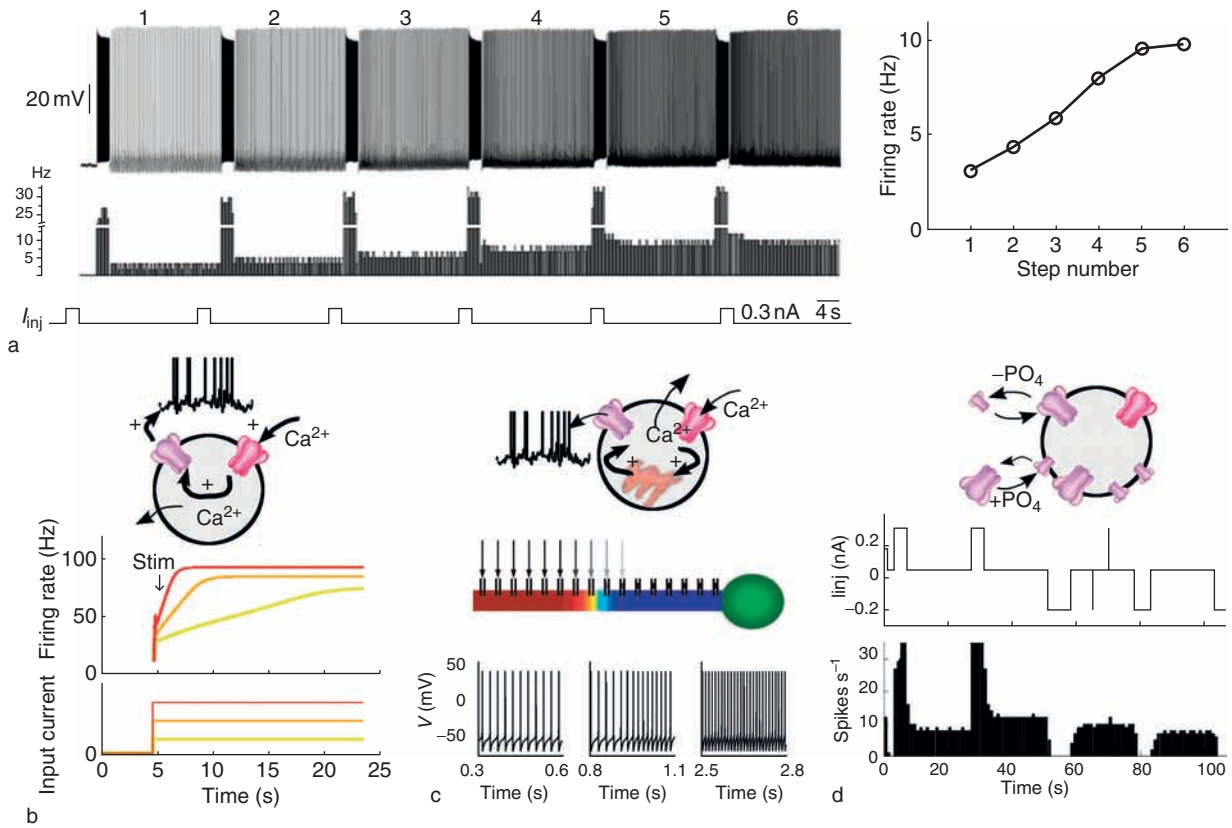

Figure 6 Mechanisms of integration in single neurons. (a) When activated cholinergically, synaptically isolated neurons of the entorhinal cortex integrate successive, brief pulses of current injection (I_{inj}; bottom left) into stable graded firing rate levels (top left). This also corresponds to a rate code, where the average rate of the neuron is linearly related to the number of stimuli applied (right). (b) Positive feedback (thick arrows indicated with a + sign) between intracellular calcium entering through high-threshold calcium channels (pink) and calcium-dependent nonspecific cation channels (purple). (Lower panels) Linearly increasing neural response (up to integrator saturation limit) for three distinct step pulse currents. (c) Positive feedback (thick arrows indicated with a + sign) between intracellular calcium and calcium release from internal stores (orange). (Middle) With uniform distribution of these feedback units along a dendrite, stable localization of the calcium front (color code in dendrite) determines the total amount of current I_{CaN} (channels open on the dendrite) summated at the soma. (Bottom) Stable neural firing rate after three consecutive pulses of current. (d) During external stimulation, phosphorylation/dephosphorylation of I_{CaN} channels modifies the balance of intrinsic currents in the neuron and its spontaneous firing rate. No positive feedback is involved in this mechanism. (Lower panels) A computational model based on this mechanism reproduces faithfully the experimental results in (a) (top graph, injected current; bottom graph, neural output). (a) Adapted from Egorov AV, Hamam BN, Fransen E, Hasselmo ME, and Alonso AA (2002) Graded persistent activity in entorhinal cortex neurons. *Nature* 420: 173–178. (b) Adapted from Durstewitz D (2003) Self-organizing neural integrator predicts interval times through climbing activity. *Journal of Neuroscience* 23: 5342–5353. (c) From Loewenstein Y and Sompolinsky H (2003) Temporal integration by calcium dynamics in a model neuron. *Nature Neuroscience* 6: 961–967. (d) Adapted from Fransén E, Tahvildari B, Egorov AV, Hasselmo ME, and Alonso AA (2006) Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron* 49: 735–746.

intracellular calcium activates the I_{CaN} current, which in turn admits cations into the neuron, depolarizing it and generating further neural activity (Figure 6(b)). Hence, this is a positive feedback mechanism similar to that of Figure 2, but it is realized at the single cell level rather than by a synaptically interconnected circuit. It has been proposed that a learning algorithm can be used to dynamically adjust the parameters of such a mechanism to generate linearly increasing responses to step current pulses of varying intensity (Figure 6(b)). Alternatively, the fundamental positive feedback loop could be established between the intracellular calcium and intracellular stores of calcium (e.g., the endoplasmic reticulum). The positive feedback occurs because intracellular stores of calcium have been found to release calcium to the cytoplasm upon increases in intracellular calcium (calcium-induced calcium release). In this scenario, I_{CaN} is not part of the positive feedback loop but, rather, serves to translate the calcium level into neuronal activity by providing calcium-dependent depolarizing currents (Figure 6(c)). Models of neural integrators built around this

scenario require a spatial distribution of the positive feedback units in order to generate graded levels of activity. One possibility would be to distribute them continuously along the neural dendrites so that a front between high- and low-calcium regions stabilizes at arbitrary locations on the dendrite (which again requires fine-tuning). Strong bursts of action potentials would progressively move this front toward the soma, enhancing the total amount of I_{CaN} and the stable firing rate of the cell (Figure 6(c)).

The two mechanisms described previously (Figures 6(b) and 6(c)) rely on a biochemical feedback loop that is highly sensitive to small modifications of the neuronal properties (fine-tuning). However, experiments on isolated, cholinergically activated entorhinal cortex neurons (Figure 6(a)) have challenged this modeling framework. In these neurons, both the careful matching of membrane currents and the release of calcium from internal stores do not seem to be necessary for neural integration. This has prompted another mechanistic explanation that does not rely on a positive feedback loop. Instead, a computational model was built based on the hypothesis that phosphorylation or dephosphorylation of I_{CaN} channels is triggered by excessively high or low concentrations of intracellular calcium, respectively. During normal, autonomous neural function it was assumed that intracellular calcium levels do not reach the extreme values required to trigger phosphorylation or dephosphorylation in the model. However, during the long pulses of external current injection tested experimentally, the phosphorylation states of the I_{CaN} channels could be changed. Because the two forms of the I_{CaN} channel have significantly different maximal conductances, these changes permanently modify the balance of currents present in the neuron and thus the basal firing rate (Figure 6d). Furthermore, because the associated basal calcium concentrations are far from the thresholds for phosphorylation or dephosphorylation of the I_{CaN} channel, the normal positive feedback loop between voltage activity, calcium concentration, and I_{CaN} conductance state is broken and the firing rates remain perfectly robust against small fluctuations. However, it is unclear whether it is biologically plausible to assume that at high calcium levels phosphorylation is not counteracted by dephosphorylation so that effectively there is no 'biochemical leak.' Moreover, the robustness does occur at a cost relative to positive feedback models because it also implies that the network is incapable of integrating small inputs. This trade-off between robustness against small fluctuations and insensitivity to small inputs is a feature of several models of neural integrators that attempt to solve the fine-tuning problem.

Hybrid Models Incorporating Synaptic Feedback and Intrinsic Bistability

Integrator models based solely on network mechanisms require fine-tuning of synaptic connection weights in order to perform accurate integration over long timescales. Some models have demonstrated that this need for fine-tuning can be reduced if neurons contain one or more digital elements that only turn on when their inputs are sufficiently large and only turn off again when their inputs are sufficiently small. In the simplest realization of this scheme, an integrator network could be made from N 'bistable' neurons that can exist in either of two stable states for a given level of input – an 'on' (active) state and an 'off' (inactive) state. If external inputs sequentially turn neurons on or off at a rate proportional to the size of the input, then the number of active units at any given time reflect an integral of the external input.

Unfortunately for this simple scheme, experimentally observed integrator neurons have not been found to exhibit such discrete on or off neuronal firing. However, this observation does not preclude the possibility that some fraction of a neuron's inputs are transmitted through bistable elements such as dendritic plateau potentials. For example, Figure 7 shows a network in which a conventional recurrent excitatory positive feedback model was modified to include multicompartmental neurons that each contain multiple bistable dendrites (Figure 7(a)). Dendritic compartments were assumed to generate plateau potentials that turn on when synaptic input exceeds a value s_{on} but do not turn off until the input falls below a lower value s_{off} (Figure 7(b)). The firing rate of each neuron reflects the number of its dendrites that are active, in addition to background and external command inputs. Because the dendrites operate in a digital manner, stored activity in the integrator network can be maintained stably even if synaptic inputs are altered by moderate amounts. As shown in Figure 7(d), the network integrates pulses of inputs into steps of firing rate even when synaptic weights (and hence recurrent synaptic inputs) are altered by an amount that would cause gross decay in firing rates in fine-tuned networks (compare to Figure 2(b)). Conceptually, the effect of the bistability in this model is to add ridges to the bottom of the energy surfaces described in Figure 2 (Figure 7(c)). External inputs to the network, which can be represented by (large) tiltings of the surface, slide the block to a location representing the new accumulated value. Imprecision in tuning of the network corresponds to (smaller) warping of the surface. Unlike the fine-tuned network, this warping does not cause the network's

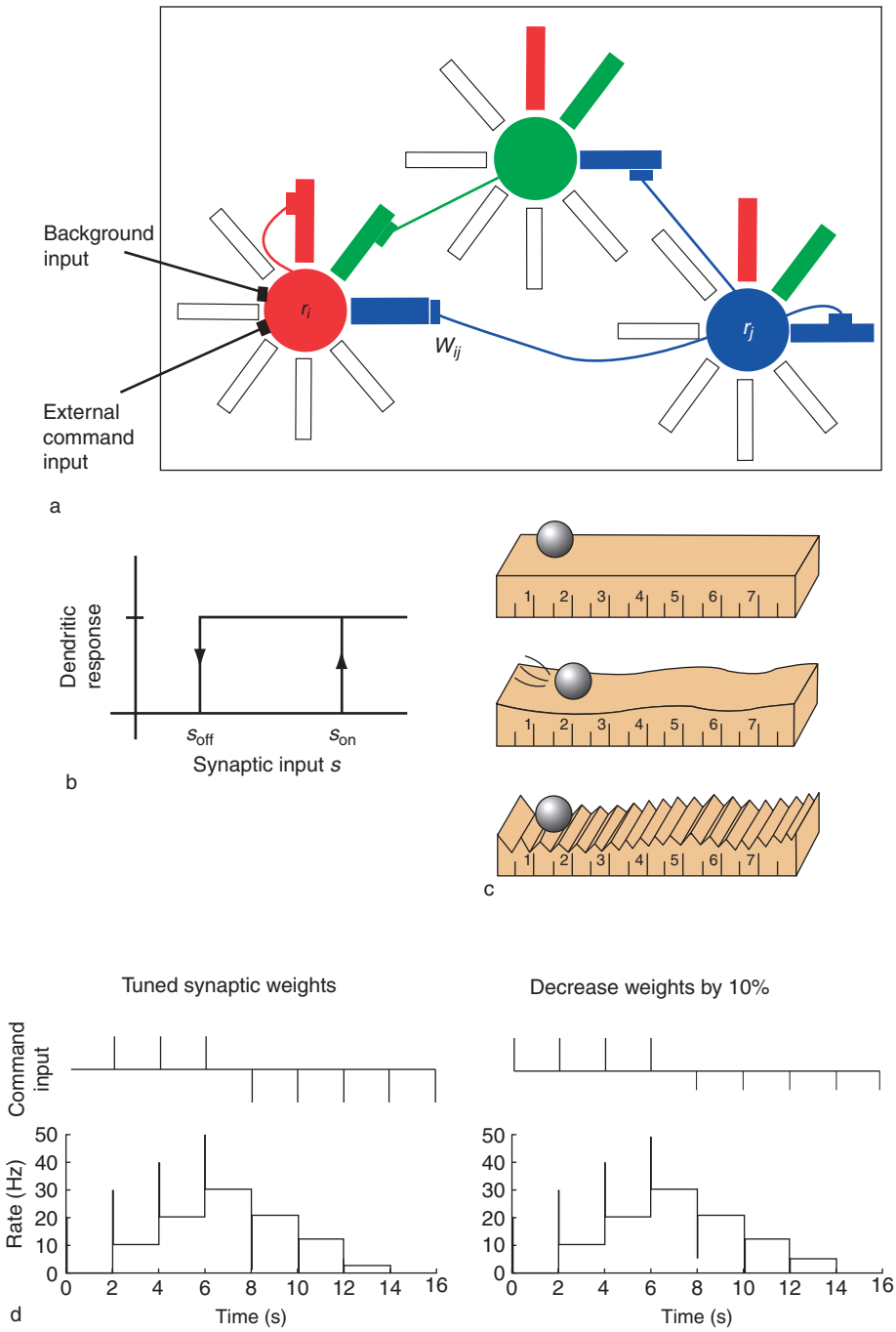

Figure 7 A hybrid intrinsic network mechanism of integration. (a) Network of multicompartment model neurons. Each neuron projects recurrent excitation of weight W to a single bistable dendrite of each of its neighboring neurons. Neurons also receive external background and command inputs at the soma. (b) Dendrites have a bistable input–output relationship characteristic of a plateau potential. Dendrites are activated when synaptic input exceeds a value s_{on} and deactivated when input falls back below a value s_{off}. (c) The bistable nature of the dendrites can be conceptualized as creating ridges in the surface along which the ball representing the state of the system moves (see **Figure 2**). These ridges make the network output resistant to drift that otherwise would be caused by small perturbations in synaptic connection strength that correspond to warping and tilting of the surface. (d) Robust maintenance of activity in the network. Pulses of input are accumulated into different levels of stable output (rate code). Firing rates can be maintained in the absence of command inputs even when weights are changed by an amount that would cause severe drift in a network without bistable elements (right; compare to **Figure 2b**). (a, b, d) Adapted from Goldman MS, Levine JH, Major G, Tank DW, and Seung HS (2003) Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. *Cerebral Cortex* 13: 1185–1195. (c) Adapted from Pouget A and Latham P (2002) Digitized neural networks: Long-term stability from forgetful neurons. *Nature Neuroscience* 5: 709–710 (comment on Koulakov AA, Raghavachari S, Kepecs A, and Lisman JE (2002) Model for a robust neural integrator. *Nature Neuroscience* 5: 775–782).

representation of its integrated input to slip as long as the warping is insufficiently large to knock the block out of the divots. However, this robustness against imprecise tuning also implies that the network is unresponsive to small inputs that do not tilt the surface enough to cause the block to slide. A challenge for future models incorporating bistability is to overcome this inherent trade-off between robustness to imperfect tuning and inability to integrate small external inputs.

Accumulation of Information versus Graded Memory

So far, our discussion has focused on perfect time integration in the sense of calculus, which is exemplified by oculomotor neural integrators and HD cells. However, it is important to note that different components of the nervous system are likely to exhibit different types of temporal integration, some less perfect than others, to satisfy a variety of computational demands. In particular, it is useful to distinguish two types of time integration: approximately linear summation of inputs during stimulation (e.g., slow ramping activity in response to a prolonged constant input) and parametric working memory (graded conversion of pulse-like input to plateau-like persistent activity). Figure 8(a) shows a line attractor network model of parametric working memory, which was developed to simulate the monkey delayed somatosensory comparison experiment (Figure 1(c)). Like the prefrontal neuron shown in Figure 1(c), this model produces graded persistent activity, which is approximately proportional to the strength of a transient input pulse and is self-sustained after the stimulus offset. The mechanism underlying graded persistent activity is precisely the tuned positive feedback described in Figure 2 but implemented with a realistic circuit model of spiking neurons. Positive feedback in the model was mediated by NMDA receptors, which have a relatively slow decay time constant of 50–100 ms, reducing the requirement for fine-tuning of synaptic weights. Such a line attractor network can sum inputs in the sense of calculus. However, working together with other neural circuits, or using additional gating input signals, such a network has been shown to be capable of encoding the intensity rather than the duration of a stimulus; that is, the graded persistent activity is proportional to the input amplitude, but if the stimulus is prolonged for increasingly longer durations, the level of neural firing would not grow linearly with time, contrary to a neural integrator. In the somatosensory delayed discrimination task the subject has to encode and remember across a delay the frequency of a somatosensory vibration, not the number of vibrations applied on a fingertip. Thus, the behavioral task requires parametric working memory but not strict time integration of input during stimulation.

Conversely, a neural circuit capable of nearly perfect integration of inputs during stimulation does not necessarily exhibit graded self-sustained persistent activity when the stimulus is withdrawn. This is illustrated in a neural circuit model of decision making (Figure 8(b)). In simulations of a two-alternative forced choice task, firing activity of neurons increases quasi-linearly over time (up to a point) in response to a constant and long-lasting stimulus. Such ramping activity gradually accumulates information about the evidence in favor of one of the two alternative choices, and the rate of ramping activity increases with the input amplitude (corresponding to the quality of sensory evidence) (Figure 8(b)). Later in time, however, competition between neural populations representing the two alternatives ultimately leads to a categorical choice that is manifested by the divergence in the neural activity of the two neural populations. One of them increases while the other decays. Thus, if a decision-based motor response must be delayed, persistent activity during working memory maintenance would be binary. The mnemonic neural firing pattern would not depend on the strength of input in a graded manner but reflect only the categorical choice made during stimulation. This was exactly what was found in lateral intraparietal area neurons in the two-alternative forced-choice version of a visual motion discrimination task and in a discrete attractor network model of decision making and working memory. Graded time integration during stimulation and binary persistent activity during a memory period are not incompatible with each other because the former is a property of neural responses to external inputs, whereas the latter reflects stable states in the absence of inputs. Thus, time integration in decision making does not necessarily require line attractors. Of course, some decision tasks require graded integration of information across temporal gaps, which engage both time integration and parametric working memory. Novel task designs are required to identify the neural activity underlying such decision processes.

Challenges and Future Directions

How neural integrators can be realized robustly by plausible biophysical mechanisms remains a topic of active current research. Current models are based on three ideas. First, positive feedback can create long time constants from much faster decaying biophysical processes. This scenario requires a need for

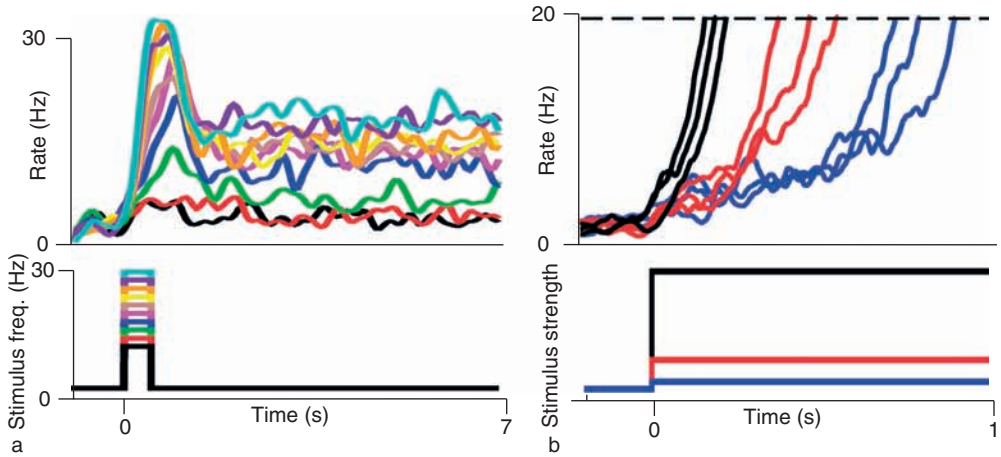

Figure 8 Graded memory and accumulation of information. (a) Graded memory in a network that converts a pulse-like input into persistent activity. The transient input is presented with a range of amplitudes, and the induced persistent activity after the stimulus offset increases approximately linearly with the input strength. The simulations are from a line attractor model of spiking neurons for the somatosensory delayed discrimination experiment (**Figure 1(c)**). (b) Linear ramping activity during stimulus presentation. Neural activity increases quasi-linearly in time, and it varies stochastically from trial to trial (three trials are shown for each input amplitude). The ramping rate increases with the input amplitude. Such 'ramping-to-threshold' probabilistic neural activity can serve to accumulate information for decision making, or can be used to estimate a time interval. The simulation results are from a discrete attractor network model of perceptual decision making and working memory. (a) Adapted from Miller P, Brody C, Romo R, and Wang X-J (2003) A recurrent network model of somatosensory parametric working memory in the prefrontal cortex. *Cerebral Cortex* 13: 1208–1218. (b) Adapted from Lo CC and Wang X-J (2006) Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience* 9: 956–963.

fine-tuning that seems unreasonably strict. For neural integrators using a rate code, fine-tuning amounts to adjustment of synaptic weights in the networks, whereas for neural integrators using a location code, fine-tuning implies symmetry of connections and homogeneity of neuronal properties in networks. Moreover, such systems typically display random drifts of neural activity in the absence of external inputs, which would be functionally detrimental. Second, very slow biophysical and biochemical processes naturally provide long time constants. However, systems that are built on processes with very slow kinetics, such as the single neuron integrator cells of entorhinal cortex, suffer from a sluggish response to inputs that does not match the rapid accumulation of inputs exhibited in other systems. Third, nonlinearities such as plateau potentials or thresholds have been used to provide robust behavior in some models but also place constraints on the types of functions that can be integrated. Further progress depends on new experiments that place more constraints on, and establish a firmer mechanistic basis for, computational modeling in this area. What types of inputs a given system does or does not integrate may provide a guide to future models because different mechanisms may be tailored to the system's specific needs. Nevertheless, certain overarching themes, such as positive feedback and analog memory storage of accumulated input levels, are likely to provide a common framework in which to view many systems.

Although most of the current research on modeling of integrators has focused on neural dynamics and robustness to perturbations on fast timescales, an open question concerns what learning rules can assemble and maintain the tuning of such circuits. Experimental evidence for plasticity rules in neural integrators is lacking, and modeling offers the opportunity to suggest appropriate experiments to address this question. Work in the oculomotor system has shown that with altered visual feedback, the oculomotor integrator can be tuned to become leaky or unstable. This suggests that reinforcement learning signals, such as retinal slip for the oculomotor system, can be utilized to tune integrator dynamics. In addition, theoretical work on this system has suggested how spike timing-dependent plasticity could be used in an unsupervised manner to maintain stable firing in the absence of external inputs. In networks utilizing a location code, where proper performance depends on homogeneity across the network so that a bump of activity can be stably maintained at any location, theoretical work has shown how a homeostatic synaptic plasticity rule (synaptic scaling) can effectively create this homogeneity in a network of neurons with diverse intrinsic parameters. Moreover, a modeling study has proposed that line attractors using a rate code can be realized using a symmetry principle for the construction of the network connectivity. If so, homeostatic regulation mechanisms may also offer a possible solution to the fine-tuning

problem for neural integrators using a rate code. Other theoretical work has noted that the higher order statistics of neuronal voltage fluctuations can exhibit special properties when networks are tuned to a line attractor, and has suggested how a learning mechanism that monitors such fluctuations could be used to tune the integrator dynamics. Close collaboration between theory and experiments will be needed to test and further develop such hypotheses at both the single neuron and the systems levels.

See also: Attractor Network Models; Multisensory Convergence and Integration; Neural Coding of Spatial Representations; NMDA Receptors and Development; NMDA Receptor Function and Physiological Modulation; Oculomotor Control: Anatomical Pathways; Oculomotor System: Models; Olfactory Coding; Sensorimotor Integration: Models; Short Term and Working Memory; Working Memory: Capacity Limitations.

Further Reading

Aksay E, Baker R, Seung HS, and Tank DW (2000) Anatomy and discharge properties of pre-motor neurons in the goldfish medulla that have eye-position signals during fixations. *Journal of Neurophysiology* 84: 1035–1049.

Brody CD, Romo R, and Kepecs A (2003) Basic mechanisms for graded persistent activity: Discrete attractors, continuous attractors, and dynamic representations. *Current Opinion in Neurobiology* 13: 204–211.

Cannon SC, Robinson DA, and Shamma S (1983) A proposed neural network for the integrator of the oculomotor system. *Biological Cybernetics* 49: 127–136.

Durstewitz D (2003) Self-organizing neural integrator predicts interval times through climbing activity. *Journal of Neuroscience* 23: 5342–5353.

Egorov AV, Hamam BN, Fransen E, Hasselmo ME, and Alonso AA (2002) Graded persistent activity in entorhinal cortex neurons. *Nature* 420: 173–178.

Fransen E, Tahvildari B, Egorov AV, Hasselmo ME, and Alonso AA (2006) Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron* 49: 735–746.

Goldman MS, Levine JH, Major G, Tank DW, and Seung HS (2003) Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. *Cerebral Cortex* 13: 1185–1195.

Kaneko CR, Evinger C, and Fuchs AF (1981) Role of cat pontine burst neurons in generation of saccadic eye movements. *Journal of Neurophysiology* 46: 387–408.

Koulakov AA, Raghavachari S, Kepecs A, and Lisman JE (2002) Model for a robust neural integrator. *Nature Neuroscience* 5: 775–782.

Lo CC and Wang X-J (2006) Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience* 9: 956–963.

Loewenstein Y and Sompolinsky H (2003) Temporal integration by calcium dynamics in a model neuron. *Nature Neuroscience* 6: 961–967.

Machens CK and Brody C (2008) Design of continuous attractor networks with monotonic tuning using a symmetry principle. *Neural Computation* 20: 452–485.

Machens CK, Romo R, and Brody C (2005) Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science* 307: 1121–1124.

Major G and Tank D (2004) Persistent neural activity: Prevalence and mechanisms. *Current Opinion in Neurobiology* 14: 675–684.

Miller P, Brody C, Romo R, and Wang X-J (2003) A recurrent network model of somatosensory parametric working memory in the prefrontal cortex. *Cerebral Cortex* 13: 1208–1218.

Miller P and Wang X-J (2006) Inhibitory control by an integral feedback signal in prefrontal cortex: A model of discrimination between sequential stimuli. *Proceedings of the National Academy of Sciences of the United States of America* 103: 201–206.

Pouget A and Latham P (2002) Digitized neural networks: Long-term stability from forgetful neurons. *Nature Neuroscience* 5: 709–710.

Renart A, Song P, and Wang X-J (2003) Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron* 38: 473–485.

Robinson DA (1989) Integrating with neurons. *Annual Review of Neuroscience* 12: 33–45.

Roitman JD and Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience* 22: 9475–9489.

Romo R, Brody CD, Hernández A, and Lemus L (1999) Neural correlates of parametric working memory in the prefrontal cortex. *Nature* 399: 470–473.

Seung HS (1996) How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences of the United States of America* 93: 13339–13344.

Seung HS (2003) Amplification, attenuation, and integration. In: Arbib MA (ed.) *The Handbook of Brain Theory and Neural Networks,* 2nd edn., pp. 94–97. Cambridge, MA: MIT Press.

Seung HS, Lee DD, Reis BY, and Tank DW (2000) Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron* 26: 259–271.

Song P and Wang X-J (2005) Angular path integration by moving 'hill of activity': A spiking neuron model without recurrent excitation of the head-direction system. *Journal of Neuroscience* 25: 1002–1014.

Taube JS (1995) Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *Journal of Neuroscience* 15: 70–86.

Taube JS and Bassett JP (2003) Persistent neural activity in head direction cells. *Cerebral Cortex* 13: 1162–1172.

Usher M and McCelland J (2001) On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review* 108: 550–592.

Wong K-F, Huk AC, Shadlen MN, and Wang X-J (2007) Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision-making. *Neuroscience Frontier of Cognitive* 1: 6, doi:10.3389/neuro.10/006.2007.

Wang X-J (2002) Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* 36: 955–968.

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience* 16: 2112–2126.

Neural Oscillators and Dynamical Systems Models

G B Ermentrout, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Suppose that a synaptically isolated neuron is injected with a constant current. As the amount of current increases, the neuron might begin to fire action potentials. Or perhaps it will fire a burst of action potentials and then stop firing. Or, if it was already firing, maybe the spikes will stop. On the other hand, consider what happens as a walking horse speeds up: he switches first to a trot and then to a gallop. In both of these cases, some qualitative change in behavior has occurred as some parameter (current in the first case, frequency in the second) changes. How can we put these wildly different phenomena at wildly different spatial and temporal scales on a common footing? Nonlinear dynamics or dynamical systems theory is an area of mathematics which attempts to classify the qualitative differences between systems as their initial states or their parameters change.

Most modeling of neurons, circuits, and systems depends on the simulation of large systems of coupled differential equations: rates of change for voltages, synaptic strengths, etc., are given in terms of the variables themselves. Thus, in this article, we focus on how nonlinear dynamics can help us understand the behavior of the kinds of differential equations that occur in computational neuroscience.

To gain an intuitive feel for the concepts that are introduced in this article, it helps to think of some simple physical systems. A familiar example is a damped pendulum, that is, a solid rod with a weight at the end and with friction. If you move the pendulum, it will rock back and forth several times, eventually returning to the downward position. This position is called an attractor, since perturbations of the pendulum end up back in the downward state. Let us consider a more complicated example. Place a flexible plastic ruler narrow edge down, and place your finger along the top edge. As long as you do not apply too much pressure, the ruler will remain completely straight. If you transiently bend the ruler from the side, keeping the same pressure on it, it will return to a the straight position. Now, apply more pressure on the top of the ruler. Once the pressure becomes too great, the ruler will buckle (bend) to one side or the other. The physical system has gone to a new attractor as the parameter (the applied pressure) changes. The plastic ruler and the pendulum are governed by sets of nonlinear equations, and the goal of dynamical systems theory is to understand their behavior both qualitatively and quantitatively.

Equilibria

We start with a very simple example, a model for a passive membrane:

$$\frac{dV}{dt} = (-g_L(V - E_L) + I)/C \qquad [1]$$

Here, $1/g_L$ is the input resistance, E_L is the resting potential, I is the applied current, and C is the capacitance. We divided by the capacitance because it is traditional to put just the derivatives of variables on the left-hand side of the equation. This is a linear differential equation, and we can find a closed form solution, or, even easier, we could pop this onto a computer and solve it. However, to do this, we need one more piece of information: the voltage at a particular point in time, for example, $t = 0$. Thus, we specify $V(0) = V_0$, which is called the initial condition. Rather than solve the equation exactly, we will, instead, introduce the concept of state space and use this to deduce what solutions look like.

In Figure 1 is a graph showing dV/dt on the y-axis and V on the x-axis. The x-axis represents all possible voltages and is called the state space for this system. The y-axis tells us how fast and, more importantly, which direction the voltage goes – to the left means it is decreasing and to the right, increasing. For example, if the initial voltage is to the left of the solid black circle, then $dV/dt > 0$ and the voltage increases. Note that as it approaches the filled circle, its rate, dV/dt, gets smaller so that it decelerates as it gets closer to the filled circle. Initial conditions to the right will decrease until the filled circle is reached. An initial condition starting exactly on the filled circle will stay there forever. For this reason, we call this particular voltage an equilibrium point for eqn [1]. Other terms which are commonly used are steady state and fixed point. Note that all initial conditions move to this unique equilibrium. Thus, we say that it is globally asymptotically stable or that it is a global attractor. No matter what the current injected, there will always be one and only one equilibrium point and it is always stable. In fact, we can write it down explicitly:

$$V_{eq} = \frac{I + g_L E_L}{g_L}$$

Thus, for this simple linear model, if we call $V(t) = \phi(t; V_0)$ the solution to eqn [1] given the initial condition, $V(0) = V_0$, we see that $V(t) \to V_{eq}$ as $t \to \infty$

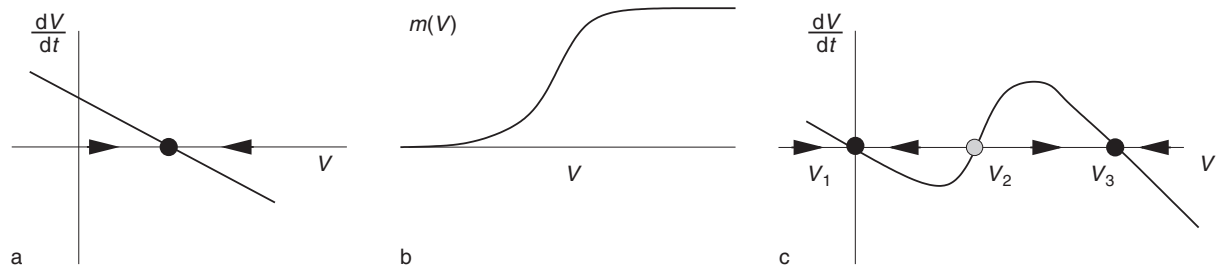

Figure 1 One-dimensional dynamical system showing stable (black) and unstable (gray) equilibria.

for any initial condition. (Note that the function ϕ is introduced to make explicit that the solution depends on the initial value.)

This graphical method may at first seem to be of little use since we can write an explicit solution to the differential equation. However, suppose that in addition to a leak conductance, our membrane has an instantaneous inward current:

$$\frac{dV}{dt} = (-g_L(V - E_L) - g_{Na}m(V)(V - E_{Na}) + I)/C \equiv f(V, I) \qquad [2]$$

where

$$m(V) = 1/(1 + \exp[-(V - V_1)/V_2])$$

(see **Figure 1(b)**). If the inward conductance is large enough, then dV/dt has the form shown in **Figure 1(c)**. It is no longer possible to write down a solution to the differential equation! However, we can use the graphical method to completely understand it. Unlike the left-hand figure, there are now three equilibria. Looking at the arrows on the x-axis, it is clear that initial conditions starting near V_1, V_3 will be driven to respectively, V_1, V_3 while initial conditions near V_2 will be pushed away from V_2. Thus, there are two asymptotically stable equilibria and there is one unstable equilibrium point. Physically, we would never be able to reach the unstable equilibrium, but we would be able to sense it in an experiment. For any initial condition $V(0) > V_2$ will tend to V_3 as time increases, and any initial condition $V(0) < V_2$ will go to V_1. Thus, eqn [2] is bistable; it has two asymptotically stable states and V_2 serves as a separatrix or threshold between the two stable equilibria. Suppose the neuron is resting at V_1 and a brief current stimulus is applied. If the stimulus is small, $V(t)$ will not cross V_2 and will relax back to V_1. If the stimulus is large enough to cause $V(t)$ to cross V_2, then the neuron will tend to the depolarized state, $V = V_3$. All we need now is some sort of delayed outward current and we will have an action potential. Nonlinear dynamics provides a geometric description for the threshold for an active neuron.

More generally, consider a set of n differential equations in n variables:

$$\frac{dx_j}{dt} = f_j(x_1, \ldots, x_n) \quad j = 1, \ldots, n \qquad [3]$$

with initial conditions $x_j(0) = x_{j,0}$. Under fairly general conditions on the functions f_j there is a solution to this which we write in vector form as $X(t) = \Phi(t; X_0)$. In this case the state space is R^n, the set of all n-dimensional vectors, and eqn [3] defines an n-dimensional dynamical system which we can write concisely as

$$\frac{dX}{dt} = F(X) \qquad [4]$$

We call X_{eq} and equilibrium solution to eqn [4] if $F(X_{eq}) = 0$. In our simple example, above, $n = 1$. The familiar space-clamped Hodgkin–Huxley equation consists of four variables, V, m, h, n, and forms a four-dimensional dynamical system. In absence of an applied current, the nerve membrane sits at rest, which is an equilibrium point. We say that an equilibrium point, X_{eq}, is asymptotically stable if all the solutions to eqn [4] with initial conditions that are close to the equilibrium point decay back to the equilibrium point as t increases. The nice geometric picture in **Figure 1(c)** does not help us for higher dimensional systems.

Suppose that A is an $n \times n$-dimensional matrix and that $F(X) = AX$. Then we call this a linear dynamical system. The study of linear dynamical systems is insightful, for, except in special cases (see below), the behavior of any nonlinear system in eqn [4] near an equilibrium point is the same as that of an associated linear system. Consider the following linear dynamical system:

$$\frac{dX}{dt} = AX \qquad [5]$$

Clearly $X = 0$ is an equilibrium point. The general solution to a linear differential equation is a sum of terms of the form:

$$Y(t) = Ve^{at}t^m(\cos bt - c) \qquad [6]$$

where m is a nonnegative integer, a, b, c are real numbers, and V is a constant vector. If $a < 0$, then

$Y(t)$ decays to zero as $t \to \infty$; if $a > 0$, then $Y(t)$ grows exponentially; and if $a = 0$, then the $Y(t)$ grows when $m > 0$ and oscillates when $m = 0$ and $b > 0$. The numbers $\lambda = a \pm ib$ are the eigenvalues of the matrix, A, that is, there is a constant vector, V, such that

$$AV = \lambda V$$

For an $n \times n$ matrix, there are n eigenvalues, roots of the nth order characteristic polynomial

$$pA(\lambda) \equiv \det(\lambda I - A)$$

Combining this fact with eqn [6], we see that all solutions to eqn [5] will decay to the equilibrium point at 0 if $a < 0$ for every eigenvalue of A and that if 'any' $a > 0$, then at least some solutions to eqn [5] will exponentially grow. In the former case, when all the a's are negative, we say that system in eqn [5] is linearly asymptotically stable, and in the latter case when at least one $a > 0$, the system is unstable. The special case $a = 0$ will play an important role later in this article.

Let X_{eq} be an equilibrium solution of eqn [3] or equivalently eqn [4]. Form the following $n \times n$ matrix:

$$B = \begin{pmatrix} \frac{\partial f_1(x_1,\ldots,x_n)}{\partial x_1} & \cdots & \frac{\partial f_1(x_1,\ldots,x_n)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(x_1,\ldots,x_n)}{\partial x_1} & \cdots & \frac{\partial f_n(x_1,\ldots,x_n)}{\partial x_n} \end{pmatrix} \qquad [7]$$

and evaluate it at the equilibrium point, X_{eq}. This matrix is called the Jacobi matrix for eqn [4] about the equilibrium point (also often called the Jacobian), and the system $Y' = BY$ is called the 'linearization.' One of the most important results in dynamical systems states that if the linear system defined by the matrix $B(X_{eq})$ is linearly asymptotically stable (all eigenvalues have negative real parts), then the equilibrium point X_{eq} is asymptotically stable as a solution to eqn [4]. This theorem provides a simple way to test the stability of any equilibrium point: form the matrix B and look at its eigenvalues. If all have negative real parts, then the equilibrium is stable; if any have positive real parts, then it is unstable.

Returning to Figure 1, the Jacobian is just the slope of dV/dt at the equilibrium point. Thus, if dV/dt has a positive slope through the equilibrium, it is unstable, and if it has a negative slope, the equilibrium is stable. Let us continue to analyze eqn [2] and Figure 1(c), but now we will change the applied current, I. Changing the current just lifts the curve in Figure 1(c) up or down. This is shown in Figure 2 on the left. For large positive or negative currents, there is only one equilibrium point. However, if we start at $I = 0$ where there are three equilibria and hyperpolarize the membrane (decrease I), then the right-most pair of equilibria will move closer to each other (compare blue with black line). As we continue to decrease the current, there is a current value, $I = I_*$, in which both equilibria come together (cyan curve, green circle), and for $I < I_*$, they no longer exist. Unlike the situation when $I = 0$, when $I < I_*$, there is only one equilibrium point (magenta curve, magenta circle). This is qualitatively different from three equilibria. A similar phenomenon occurs when we raise the current; the left-most equilibria merge when $I = I^*$ (red curve, gold circle). In dynamical systems, any time that a qualitative change in behavior occurs as a parameter varies, we say that a bifurcation has occurred. Specifically, when two equilibria merge and disappear, we say that there is a saddle-node bifurcation. This is often called a limit point or turning point bifurcation. Clearly the left-hand picture is useful for a one-dimensional equation, but obviously, if the function $f(V,I)$ is more complicated, we will run out of space and have to draw many different curves for many different values of I. A concise summary of the behavior of this nonlinear system as a parameter varies is shown in the right side of Figure 2. A plot is made of the equilibrium values of one of the variables on the y-axis, and the parameter is shown on the x-axis. This curve is called a bifurcation diagram. Traditionally, stable equilibria and unstable equilibria are distinguished by changing the color or line type of the curve. In our figure, stable equilibria are solid and unstable are dashed. The diagram should be read as follows. Pick a value of I and mentally draw a vertical

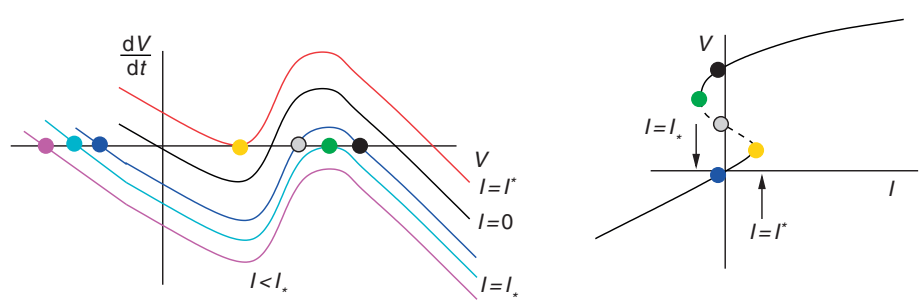

Figure 2 Behavior of nonlinear membrane as the current changes. (Left) Phase space picture. (Right) Bifurcation diagram. Special points are colored green and gold.

line through this value of I. Intersections with the bifurcation diagram correspond to equilibria. Saddle nodes are points where the diagram turns around (hence the name turning point); here, they are at I_* and I^*. Figure 2, right, shows that there will be bistability for all currents between I_* and I^*. Suppose that we do the following experiment. Starting at $I = 0$, we gradually increase the current. As soon as it exceeds I^*, the voltage will jump up to the top part of the diagram. Now, we reduce the current. The voltage stays at the upper part of the diagram until the current falls below I_*. Thus, it jumps down at a different current than it jumped up. This phenomenon is called hysteresis, and our neuron has now been endowed with 'memory.' Once in the 'up' state, it will remain there even when the current is brought to zero. Bistability between states (here, equilibria) plays a central role in almost every aspect of the dynamics of single cells and networks of coupled cells.

Before turning to more complex dynamical systems, we close this discussion with an important point. Looking again to Figure 2, left, observe that when $I = 0$, the slope through the gray circle is positive while that through the black circle is negative. The linearized system for our model is

$$y' = f'(\bar{V})y$$

where $\bar{V}$ is an equilibrium. If the slope (f') is negative, the equilibrium is stable, and if it is positive it is unstable. At $I = I_*$, the slope at the equilibrium is zero and the linearized system is

$$y' = 0$$

which tells us nothing about stability of the nonlinear system. However, whenever a linearized system has a zero eigenvalue, this is the signature for a possible bifurcation for the nonlinear system. This is true no matter what the dimension of the system.

The methods for finding equilibria and their stability, as well as looking for bifurcations, are not confined to one-dimensional models. Nor are they restricted to single neuron models. As an example of a bifurcation with symmetry, let us consider a mutually inhibitory neural network (Figure 3, left):

$$\frac{du_1}{dt} = -u_1 + F(I - gu_2)$$
$$\frac{du_2}{dt} = -u_2 + F(I - gu_1)$$

u_1, u_2 represent the firing rates of two different populations of neurons which are receiving equal inputs, I. The strength of the inhibition is g, and the function $F(u)$ increases with increasing values of u. The reader can imagine this as representing two objects competing for the attention of the network. Since the two populations are identical, there is a symmetric equilibrium point in which $u_1 = u_2 = \bar{u}$. The stability is determined by looking at the linearized system:

$$\frac{dy_1}{dt} = -y_1 - \alpha g y_2$$
$$\frac{dy_2}{dt} = -y_2 - \alpha g y_1$$

where $\alpha = F'(I - g\bar{u})$. Notice that $\alpha > 0$. The Jacobian matrix for this linear system is

$$A = \begin{pmatrix} -1 & -\alpha g \\ -\alpha g & -1 \end{pmatrix}$$

Two-variable systems (also called planar systems) are easy to analyze for stability. Let T be the sum of the diagonal entries of A (called the trace of a matrix) and let Δ be the determinant of the matrix. Then the linear system (and therefore the nonlinear system) will be stable provided that $T < 0$ and $\Delta > 0$. For this example, $T = -2 < 0$ and $\Delta = 1 - g^2\alpha^2$. If the mutual inhibition g is small, or the inputs are small (so that α is

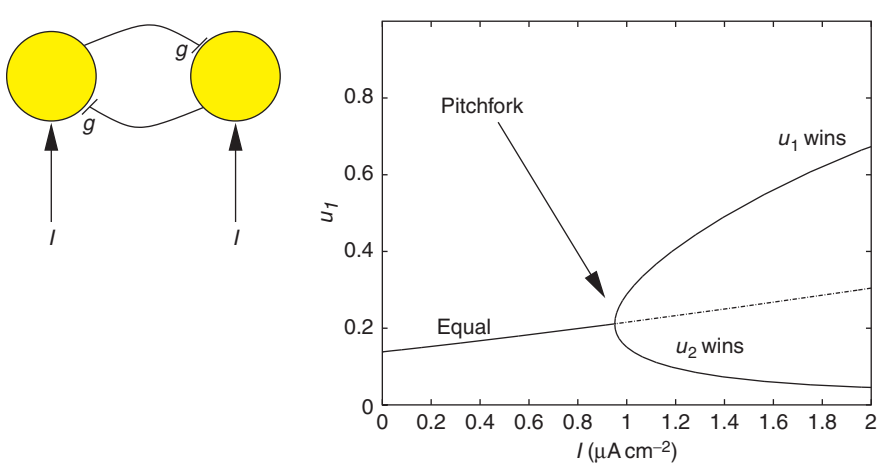

Figure 3 Mutually inhibitory network. (Right) Bifurcation diagram as the input, I, increases.

small), then $\Delta > 0$. If the inputs increase, then Δ could become negative. The equilibrium point is no longer stable. In fact, when the determinant is negative for a planar system, it means that there is one positive eigenvalue and one negative eigenvalue and the equilibrium is called a saddle point. Because of the symmetry of our example, the eigenvalues and corresponding eigenvectors of A are easy to write down:

$$\{\lambda_1 = -1 - g\alpha, \mathbf{v}_1 = [1, 1]^T\}$$

$$\{\lambda_2 = -1 + g\alpha, \mathbf{v}_2 = [1, -1]^T\}$$

λ_1 is always negative, but λ_2 becomes positive when αg exceeds 1. This means that there will be growth along the vector $\mathbf{v}_2$ and decay along the vector $\mathbf{v}_1$. For this reason, $\mathbf{v}_1$ is called the stable set or the stable manifold for the saddle point and $\mathbf{v}_2$ is called the unstable set or unstable manifold. Amazingly, if the linearized system has a saddle point and stable and unstable sets, then the nonlinear system does as well. However, they will not be straight lines.

When the input reaches a critical value so that $\alpha g = 1$, then the determinant is zero and we see that $\lambda_2 = 0$. From the previous discussion, this means that we can expect a bifurcation. Figure 3 shows the bifurcation diagram for this system. This is very different from the diagram shown in Figure 2. Rather than losing solutions as the parameter changes, a solution becomes unstable and two new 'stable' solutions appear. Because of its shape, this is known as a pitchfork bifurcation. It occurs only in systems which have some kind of intrinsic symmetry, but, as this is very common in many physical and biological problems, the pitchfork is a familiar bifurcation. The central line corresponds to the equilibrium point $u_1 = u_2$. As the input increases, when it reaches approximately 1, the symmetric equilibrium point becomes unstable – symmetry is broken. The new equilibria have the form $u_1 = a$, $u_2 = b$ and $u_1 = b$, $u_2 = a$ where $a > b$. A network like this is called winner-take-all since one of the two objects dominates the other. Notice that when $I = 2$, the value a is about ten times larger than the value b.

Planar systems have another advantage over high-dimensional models. Like the one-dimensional models in Figure 1, it is possible to sketch the complete behavior of the differential equation. Instead of plotting u_1 or u_2 versus time, plot one against the other. The resulting diagram is called a phase-plane. Figure 4 shows the phase-plane for the network. Representative solutions (called trajectories; black) appear as curves moving toward the stable equilibrium points, which are filled circles. The advantage of the phase-plane is that it is possible to infer qualitative behavior without actually solving the equations. At each point in the (u_1, u_2) plane, the equations tell us the derivatives, du_1/dt and du_2/dt. These derivatives are tangent to the solutions, so that they show the direction of trajectories. Magenta arrows are drawn at various points in the plane to show the direction. These arrows are called direction fields. Often only a crude approximation of the direction fields is enough: just the eight points on a compass. For example, northeast corresponds to both u_1 and u_2 increasing in time. To determine where u_1 is increasing or decreasing, we only need to determine where u_1 is not changing at all. The graph $du_1/dt = 0$ is called the u_1-nullcline. For our network, $du_1/dt = 0$ when $u_1 = F(I - gu_2)$. This curve is plotted as in red in the phase-plane. The u_2-nullcline, $du_2/dt = 0$ is plotted in green. The nullclines break the phase-plane into regions where the direction fields are similar and their intersections are the equilibria. Thus, much can be gained without ever solving the differential equations. In the right phase-plane, there are two stable equilibrium points separated by an unstable

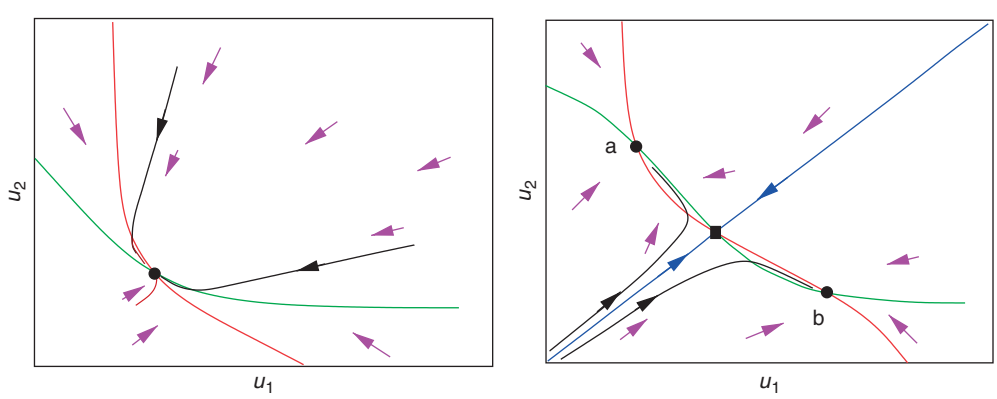

Figure 4 Phase-plane for mutual inhibitory network. Direction fields (magenta arrows) and trajectories (black) are shown. The u_1 nullcline is in red and the u_2 nullcline is in green. For stronger inputs, there are three equilibria. Two (black circles) are stable. Stable (blue) sets for the saddle point (black square) are shown.

saddle point. The stable set of the saddle point (drawn in blue) divides the plane in two. All initial values of (u_1, u_2) above the stable set will end up at the equilibrium point where u_2 wins and u_1 loses, labeled a. The opposite occurs when below the line. The upper triangular region of the plane is called the 'basin of attraction' for the equilibrium point a.

Limit Cycles

When a neuron is injected with current, it begins to fire repetitively. If the interspike interval is constant, then this pattern of firing is periodic and, in dynamical systems language, is called a limit cycle. A limit cycle is an isolated periodic solution to eqn [4]. All autonomous rhythmic phenomena seen in biology are limit cycles ranging in scales from the 24 h circadian oscillation in the suprachiasmatic nucleus to the 200 Hz ripple in hippocampus. There are several ways in which limit cycles can appear in nonlinear systems. To illustrate these, we will use the simple membrane model of Morris and Lecar:

$$C\frac{\mathrm{d}V}{\mathrm{d}t} = I - g_L(V - E_L) - g_k n(V - E_k) - gcam_\infty(V)(V - E_{\mathrm{Ca}})$$

$$\frac{\mathrm{d}n}{\mathrm{d}t} = a_n(V)(1 - n) - b_n(V)n$$

Depending on the parameters of the model, as the current increases, rhythmic firing appears in three different ways: (1) Hopf, (2) SNIC, and (3) homoclinic. These three possibilities have been found in almost all published simulations, and there is strong evidence that they are also observed in real neurons.

The best-known mechanism by which a neuron begins to fire periodically is via the Hopf or Andronov bifurcation. Recall that an equilibrium point is stable as long as the real parts of the eigenvalues of the linearized system are all negative. There are two ways in which the real part can switch to positive: (1) a real negative eigenvalue becomes a positive eigenvalue through a zero eigenvalue and (2) the real part of a complex eigenvalue changes sign, leading to an imaginary eigenvalue, $\pm i w_0$. In the former case, saddle node and pitchfork bifurcations occur – the number of equilibria changes. In the latter, there is no change in the number of equilibria, but instead, limit cycles appear. Figure 5 shows the bifurcation diagram for the Morris–Lecar model as the applied current increases. There is a single equilibrium point, and it changes from stable to unstable as the current increases past about 95. The magenta and green curves plotted in this figure correspond to limit cycles. They depict the maximum and minimum amplitude of the voltage. Like equilibria, the notion of stability can be defined for limit cycles. The magenta curves correspond to unstable and the green to stable limit cycles. The point HB denotes the point of the Hopf bifurcation. Models which undergo Hopf bifurcations have a limited frequency band and a nonzero minimum frequency (often close to w_0). For this model, the frequency ranges from 8 to 14 Hz. The small figures labeled a, b, and c are the phase-planes corresponding to points along the bifurcation diagram. At point (a) there is only a stable equilibrium and the neuron always returns to rest. In point (b), taken between the currents I_1 and I_2, there are two stable states: a stable

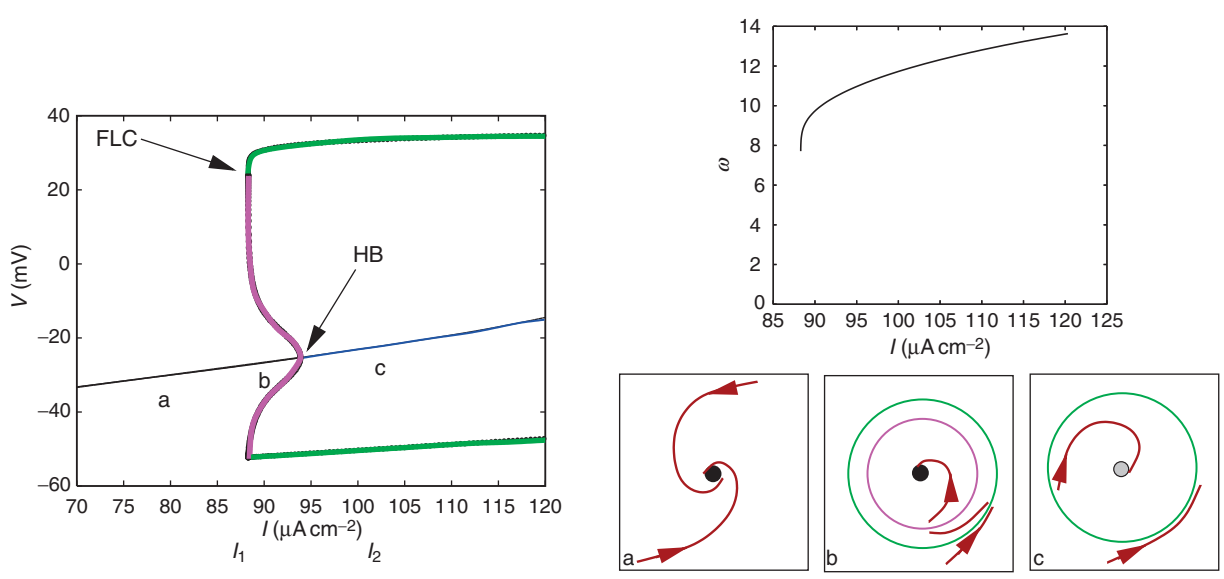

Figure 5 Hopf bifurcation formation of limit cycle. HB, Hopf bifurcation; FLC, fold of limit cycles.

equilibrium and a stable limit cycle. The unstable limit cycle (magenta) separates the plane into two regions. Everything outside of the unstable cycle ends up on the stable cycle and everything inside ends up at the equilibrium. The well-known Hodgkin–Huxley model behaves exactly like this. As the current approaches the value I_1 from the right, the magenta limit cycle and the green limit cycle collide and annihilate. This is called a fold of limit cycles (FLC) and is very much like the saddle-node bifurcation for equilibria: instead of losing two equilibria, two limit cycles are lost. In the transition from (b) to (c), the unstable magenta limit cycle shrinks to the equilibrium and renders it unstable, leaving only the stable limit cycle. There is compelling evidence that fast-spike interneurons in the cortex have this type of bifurcation.

A less well-known, but very important bifurcation to limit cycles is called the saddle-node invariant cycle (SNIC) bifurcation. Most computational models for cortical neurons make the transition from rest to repetitive firing via this bifurcation. An example is shown in Figure 6. Unlike the Hopf case, there is never bistability. Rather, there is a smooth transition from rest to rhythmicity. At the point of bifurcation, the frequency of the oscillation goes to zero. Note the huge dynamic range: 0–25 Hz for the simple model. For subthreshold currents, there are three equilibria; two are unstable and one, the rest state, is stable. Phase-plane (a) shows the big picture. The turquoise equilibrium serves as a threshold. If the potential crosses this, then a spike will be emitted before the cell returns to rest. As the current increases, the threshold and the rest state merge (phase-plane (b)) and at larger currents, these two equilibria disappear, leaving just the stable limit cycle and the unstable equilibrium. Comparing phase-plane (c) in the Hopf and SNIC shows that for large enough currents, they cannot be distinguished. However, most neurons (at least in cortex) operate near rest so that the bifurcation matters. We note that near the critical current, I^*, the frequency grows like:

$$\omega = K\sqrt{I - I^*}$$

A key difference between the SNIC and the Hopf is that the Hopf has a much more restricted frequency range and thus is very sensitive to the frequency of the inputs, while the SNIC has a very broad frequency range and no preferred input frequencies. These differences could have important computational consequences: SNIC neurons respond to the total number of inputs (integrators), while Hopf neurons are more sensitive to timing.

The third way in which a limit cycle can appear also involves an arbitrarily low frequency but with a different scaling from the SNIC. The bifurcation diagram is shown in Figure 7, and it is more complex than the other two. The frequency also has a large dynamic range, between 0 and 50 Hz. At point (a), there are three equilibria, one stable and two unstable. In phase-plane (a) the unstable set (gold) from the saddle point (turquoise) lies on the outside of the stable set (magenta). As the current approaches the value denoted (b) in the bifurcation diagram, the stable and unstable sets merge to form the red loop in phase-plane (b). This is called a 'homoclinic' solution. Just past the bifurcation point, the homoclinic becomes a stable limit cycle, at point (c) and depicted by the green cycle in phase-plane (c). Note that there is both a stable

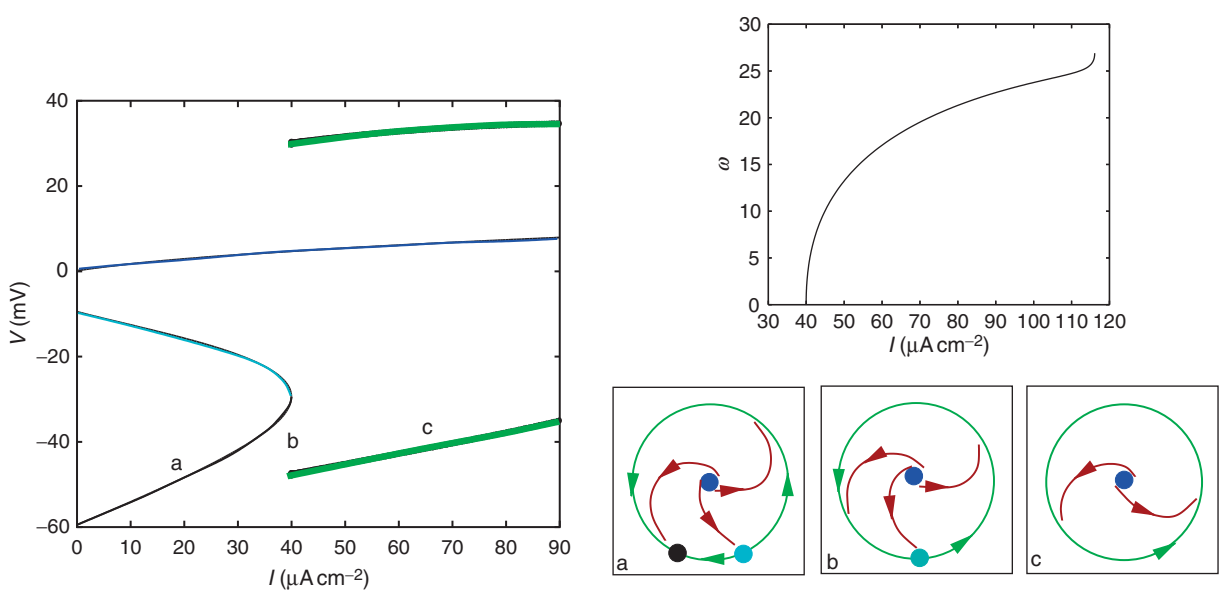

Figure 6 Saddle-node formation of limit cycle.

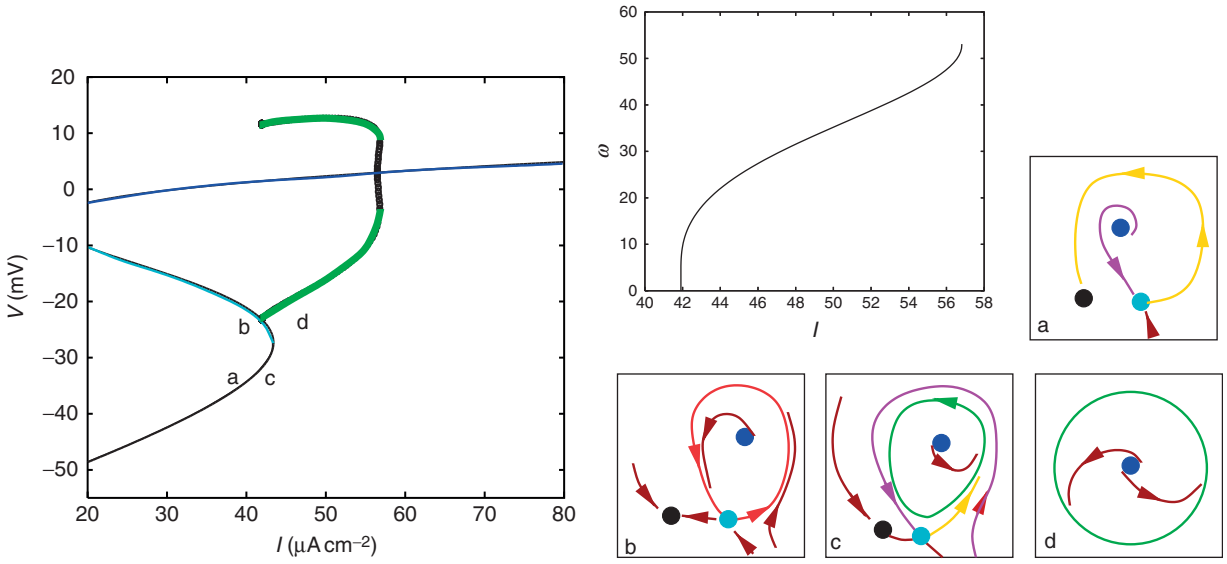

Figure 7 Homoclinic formation of limit cycle.

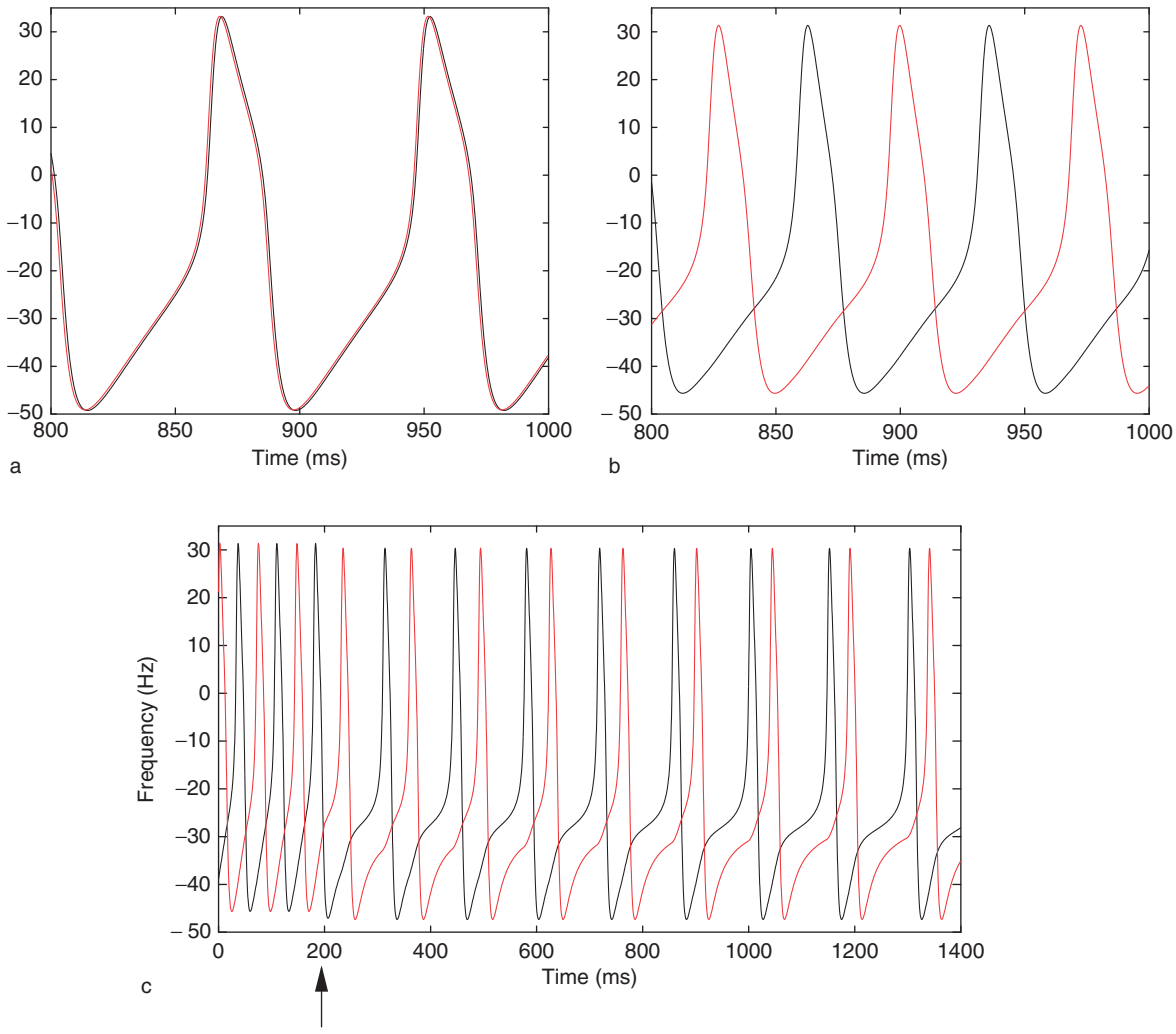

Figure 8 Bifurcation type determines synchronization properties. (a) Two Hopf neurons synchronize with excitatory coupling; (b) SNIC neurons fire in antiphase. (c) Frequency matters as well – if the frequency is halved, a pair of SNIC neurons switch from antiphase to a nonsymmetric rhythm.

limit cycle and a stable equilibrium (black circle). The stable set of the saddle point (magenta) separates the phase-plane into regions which tend to either the limit cycle or the equilibrium. Finally, as the current increases, the stable equilibrium and the saddle point merge (a saddle-node bifurcation), leaving only the limit cycle and the unstable equilibrium.

The reader may well wonder if these differences matter. Indeed, there seems to be little difference in the shapes of the spikes between the SNIC and the Hopf cases when the neuron is driven to fire at, say, 12 Hz. However, the mechanism by which a neuron makes the transition from rest to firing makes a large difference in how it responds to inputs. This fact is illustrated when two neurons are synaptically coupled into a network. Networks of coupled neural oscillators play a central role in many motor pattern circuits and may be important in sensory perception as well. Figure 8 demonstrates that the type of bifurcation matters. In (a), the neurons use parameters like those in Figure 5 and are driven to 12 Hz. Two identical cells are coupled via a synaptic current and, no matter what the initial state, will eventually synchronize. In (b), the parameters are as in Figure 6, driven also to 12 Hz. However, no matter what is the initial state, the two neurons will fire alternately in 'antiphase,' a half-cycle apart. Not only does the type of bifurcation matter, but also the frequency Panel (c) shows what happens to a pair as the frequency is halved from 12 to 6 Hz. Instead of firing in a perfect antiphase pattern, there is a switch so that the neurons fire with a short time between respective spikes followed by a long silent period. In the context of the questions posed at the beginning of this article, one can view the frequency-dependent switch shown in Figure 8(c) as analogous to a gait switch as an animal speeds up or slows down.

Summary

Nonlinear dynamics provides a means of characterizing the temporal behavior of neurons and networks of neurons. Transitions between different qualitative behaviors as parameters change occur in only a few different ways. Despite differences in timescales and dimensions of the problem, these bifurcations share common features which provide insight into how neurons react to stimuli and the number and variety of states that networks can achieve.

See also: Attractor Network Models; Hodgkin–Huxley Models; Spiking Neuron Models.

Further Reading

Guckenheimer J and Holmes P (1986) *Applied Mathematical Sciences, Vol. 42: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* New York: Springer.

Izhikevich E (2006) *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting (Computational Neuroscience).* Cambridge, MA: MIT Press.

Koch C and Segev I (1998) *Methods in Neuronal Modeling, 2nd edn.: From Ions to Networks (Computational Neuroscience).* Cambridge, MA: MIT Press.

Strogatz SH (1994) *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering.* Cambridge, MA: Perseus.

Neural Patterning: Arealization of the Cortex

D D M O'Leary and T T Kroll, The Salk Institute, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The cerebral cortex is the largest and most complex component of the brain, and more so than any other brain component has been affected by evolutionary processes resulting in a tremendous increase in size and complexity, exhibited by primates, including humans, by cetaceans (e.g., dolphins, whales), and by elephants, which have the largest brains among all species. The neocortex is the largest region of the cerebral cortex; it exhibits the most substantial increase in size and complexity in more advanced species, disproportionately so compared to the midbrain and hindbrain, and also exhibits arguably the most significant phylogenetically distinct specializations.

Early in embryonic development of mammals, the neural plate differentiates into the neural tube, and during this process the rostral end of the neural tube develops three vesicles, the forebrain, midbrain, and hindbrain, that ultimately give rise to the brain. Soon thereafter, the dorsolateral aspects of the forebrain evaginate to generate the telencephalic vesicles, and the remainder of the forebrain becomes the diencephalon. The expression of individual transcription factors (TFs) or combinations of TFs correlates with morphologic boundaries within the telencepahlon and has important roles in telencepahlon patterning into two major subdivisions – the ventral telencephalon, which gives rise to the striatum and basal ganglia, and the dorsal telencephalon, which gives rise to the cerebral cortex. The diencephalon differentiates into several components, the major ones including the dorsally positioned epithalamus (habenula) and dorsal thalamus, which reciprocally connect with the neocortex, and the ventrally positioned hypothalamus.

Putting the Neocortex in Its Place

Although the focus here is on mechanisms that control area patterning of the neocortex, it is important to place the neocortex into some context. Within the cerebral cortex, the neocortex is sandwiched between what have classically been considered the more primitive archicortex (including entorhinal cortex, retrosplenial cortex, subiculum, and hippocampus) and paleocortex (olfactory piriform cortex). The neocortex is the most highly differentiated and intricately organized region of the cerebral cortex, and its laminar patterning is distinct from that of the other cortical regions, having the largest number of layers: six major, radially stacked, stratified layers, each containing a heterogeneous population of neurons that are morphologically, connectionally, and functionally distinct from those of other layers.

In its tangential dimension, the neocortex is organized into subdivisions referred to as cortical 'fields' by some authors, but more commonly as cortical 'areas.' Areas are distinguished from one another by major differences in their cytoarchitecture and chemoarchitecture, input and output connections, and patterns of gene expression. In the adult, the transition from one neocortical area to another is typically, but not always, abrupt, with borders that can be sharply defined by area differences in cytoarchitecture and chemoarchitecture, and in some instances by the distributions of projection neurons, input projections, or gene expression patterns. These attributes form a specific combination of properties that is unique for each area, and together with unique combinations of gene expression, determine the functional specializations that characterize and distinguish areas in the adult.

The neocortex has four 'primary' areas. In addition, scores of higher-order areas are positioned between the primary areas, serving as higher-order processing centers focused on a particular modality – for example, particularly features of vision, movement, or attention related to the visual field; many higher-order areas are multimodal. Three of the primary areas are sensory: the primary visual (V1), somatosensory (S1), and auditory (A1), which process primary information received from the eye/retina (vision), body (somatosensation), and inner ear/cochlea (audition), respectively. The only major sense not processed by the neocortex is smell. Distinct orders are sensed by olfactory and vomeronasal receptors in the nose and related peripheral organs and are processed in the olfactory piriform cortex, positioned within the cerebral cortex, usually ventrolateral to more rostral parts of neocortex. The fourth primary area of the neocortex is the motor (M1) area, which controls voluntary movement of body parts.

These primary sensory and motor areas are conserved among all mammals, as is the general spatial relationship between them: V1 is positioned caudally; M1 is positioned rostrally, and S1 is located between them; A1 is located caudolaterally to S1. However, some animals with unusual or large and atypical peripheral appendages/sense organs (e.g., the platypus' bill or the echo-location system in bats) have modifications on this general geometrical scheme of area patterning.

The relationships between a primary cortical area and nuclei in dorsal thalamus are critical both for

adult function and for development and differentiation of areas. Dorsal thalamus has four principal thalamic nuclei that functionally and connectionally parallel the four primary cortical areas. Each primary cortical area sends outputs from layer 6 neurons to a principal thalamic nucleus, and receives, from the same nucleus, thalamocortical afferent (TCA) inputs that terminate primarily in layer 4, thereby generating the reciprocal area- and nuclei-specific connections between cortex and thalamus. Like cortical areas, individual nuclei are responsible for processing modality-specific sensory information. The four principal thalamic nuclei receive modality-specific sensory information from peripheral sense organs or receptors and form reciprocal projections with their related primary cortical areas. A simplified wiring circuit has the ventroposterior (VP) nucleus of dorsal thalamus reciprocally connecting with S1, the dorsal lateral geniculate nucleus connecting with V1, the ventral part of the medial geniculate nucleus connecting with A1, and the ventrolateral nucleus connecting with M1. Thus, the modality of a primary cortical area is determined by the modality of the principal thalamic nucleus with which it is reciprocally connected. Higher-order areas are often multimodal and their modalities are determined largely by the modality of the inputs that they receive from other areas – primary as well as higher-order areas, through intracortical axonal projections.

Corticogenesis

Areas of the neocortex differentiate within an earlier, more or less uniform structure composed of postmitotic neurons, termed the cortical plate (CP) (**Figure 1**). Most neocortical neurons, including all projection neurons and glutamatergic neurons, are generated in the ventricular zone (VZ) of the dorsal aspect of the lateral ventricle, or, at later stages, in a second germinal zone, the subventricular zone (SVZ). The VZ generates deeper layer neurons, including the subplate (SP) and the major classes of layer 6 and layer 5 projection neurons, whereas the SVZ is the primary source of superficial layer neurons. In primates, the SVZ is substantially larger than in other mammals (e.g., mice). In addition, the primate SVZ shows substantial differences in mitotic activity in caudal cortex (and possibly elsewhere); these differences may underlie the major increase in the numbers of neurons in the upper layers in V1 compared to adjacent higher-order visual areas (e.g., V2), which in turn likely contributes to the substantial differences in cytoarchitecture between V1 and V2.

The first postmitotic neurons accumulate on the top of the VZ, forming the preplate (PP), positioned just beneath the pial surface. PP neurons are derived not only from the VZ, but also from the cortical hem or other sources external to the neocortex. Neurons subsequently generated in the VZ migrate along the processes of radial glia, which are actually the VZ progenitor cells, then aggregate within the PP, thereby forming the CP, which splits the PP into a superficial marginal zone (MZ) (future layer 1) and a deep SP. The CP gradually differentiates in a deep to superficial pattern, and layers 6 through 2 emerge. The MZ contains Cajal–Retzius neurons that express reelin, a large, secreted protein required for the proper radial migration of CP neurons and their formation of layers. Cajal–Retzius neurons are generated in multiple sites external to the cortical VZ, primarily within the cortical hem. SP neurons have been proposed to serve a number of critical roles in cortical development, including the pioneering of the internal capsule, the path of the major input and output projections between the cortex and the rest of the central nervous system (CNS).

A majority of cortical interneurons, which are γ-aminobutyric acid (GABA)ergic and also typically express a distinct neuropeptide (e.g., somatostatin, substance P, or neuropeptide Y), are generated primarily within the medial and caudal ganglionic eminences. These interneurons migrate along multiple pathways to reach the cortex, and migrate within the cortex along tangentially aligned pathways, being directed by many molecules, including members of the slit, semaphorin, and neuregulin families. Once interneurons reach an appropriate position, they migrate perpendicular to their original migratory plane and migrate radially into the CP.

Areas Differentiate within a 'Uniform' Cortical Plate Characterized by Exuberant Distribution of Projection Neurons and a Lack of Other Area-Specific Features Except Area-Specific Thalamocortical Afferent Input

It has been assumed that the specification and differentiation processes of neocortical areas are controlled by interplay between intrinsic mechanisms (i.e., genetic mechanisms that operate within the cortex) and extrinsic mechanisms (TCA input or information relayed by it, for example). However, only in the present millennium has compelling evidence emerged for the genetic regulation of arealization, including

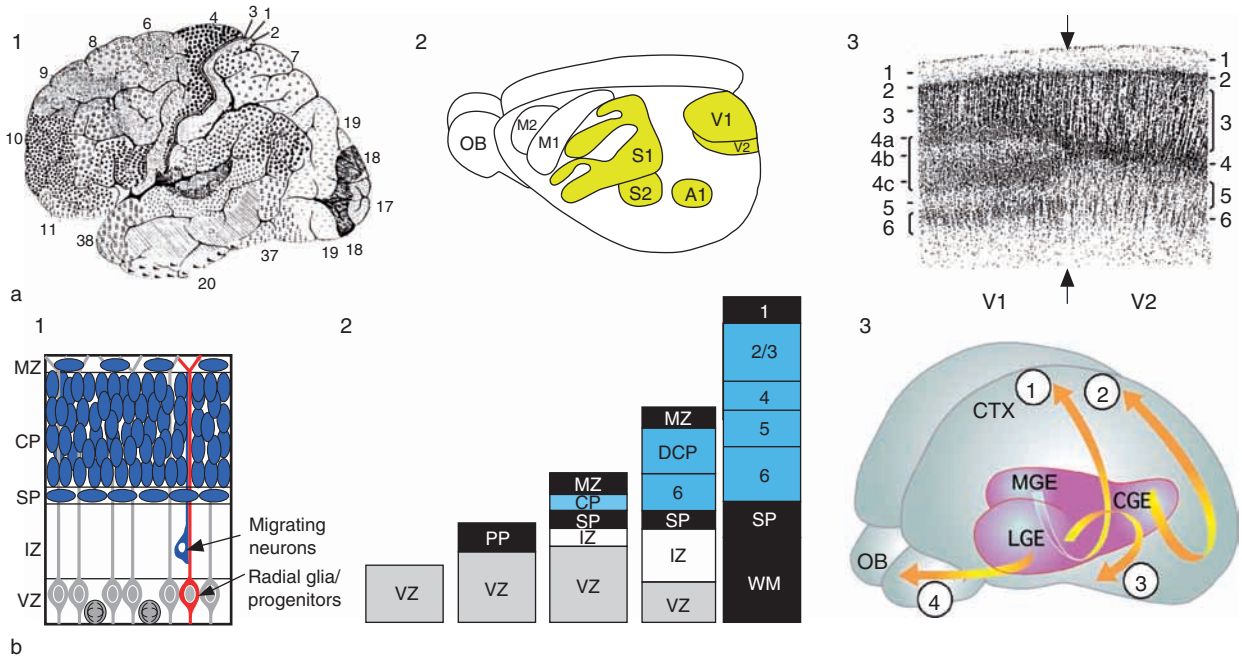

Figure 1 Basics of corticogenesis. (a) Organization of the neocortex into areas. (1) Areas of the human cerebral cortex as defined by Brodmann. View of the lateral surface of the human cerebral hemisphere shows a number of cytoarchitectonically distinct areas, such as the primary visual area (area 17, also termed V1) at the caudal pole, the primary somatosensory area (area 3, also termed S1) in the middle, and, just rostral to it, the primary motor area (area 3, also termed M1). (2) Selected areas of the rat neocortex from Kaas and colleagues showing sensory areas (V1, primary visual; V2, secondary visual; S1, primary somatosensory; S2, secondary somatosensory; A1, primary auditory) as well as the primary and secondary motor areas (M1, M2) and the olfactory bulb (OB). (3) An example of abrupt borders between areas (from an 8-month human fetus), showing the V1–V2 cytoarchitectonic border, revealed using a Nissl stain. Each area has six primary layers, but their architecture and internal sublayering differ. (b) Generation, migration, and lamination of neocortical neurons. (1) Most neocortical neurons, including all glutamatergic and projection neurons, are generated in the neocortical ventricular zone (VZ) and later in the subventricular zone (not shown). Most cortical plate (CP) neurons (SP, subplate) migrate radially on radial glial fibers from the VZ, through the intermediate zone (IZ), and aggregate in the CP. The radial glia are also the progenitors in the VZ that give rise to cortical neurons. (2) Layer development of the rodent neocortex. The first neurons generated in the VZ, as well as neurons that migrate tangentially into the marginal zone (MZ) from external germinal zones, such as the cortical hem, aggregate on top of the VZ and form the preplate (PP), which is later split into MZ and SP by the later generated CP neurons (WM, white matter). (3) Most neocortical interneurons (i.e., GABAergic interneurons of various peptide phenotypes) are generated in the medial ganglionic eminence (MGE, pathway 1) and caudal ganglionic eminence (CGE, pathway 2) and migrate tangentially through the intermediate zone and marginal zone to distribute across the neocortex, and then turn perpendicular to their tangential path and migrate radially into the cortical plate (CTX, cortex). The lateral ganglionic eminence (LGE) is the source of GABAergic interneurons that migrate into the piriform cortex (pathway 3) and to the olfactory bulb (pathway 4). Cortical plate neurons are generated in an inside-out fashion, and layers differentiate from the cortical plate in the same pattern: earlier born neurons form the deeper layers, while later born neurons migrate past them and form the more superficial layers. Interestingly, interneurons generated external to the cortex also become distributed in an inside-out gradient dependent upon their birth date. (b3) Adapted from Corbin JG, Nery S, and Fishell G (2001) Telencephallic cells take a tangent: Non-radial migration in the mammalian forebrain. *Nature Neuroscience* 4(supplement): 1177–1182.

the first demonstration of regulatory genes that control area specification and evidence for patterning centers and signaling molecules that establish the initial patterning of these genes.

Cytoarchitecture and Exuberant Projection Neurons

The CP lacks the many features that distinguish areas in the adult cortex, even after all CP neurons have been generated and layers begin to differentiate within it. The sharp cytoarchitectonic borders evident between areas in the adult cortex are lacking; indeed, other than differences in thickness, CP cytoarchitecture is uniform across its tangential extent. Also absent are the restricted, area-specific distributions of distinct types of projection neurons characteristic of the functional specializations of different cortical areas in adults. Instead, cortical projection neurons have widespread distributions early on that include parts of areas, and even entire areas, in which they are not found in the adult cortex, and their restricted adult cortical distributions come about by the elimination of functionally inappropriate axon segments and branches. This mechanism is used to generate the characteristic patterning of callosal, intracortical, and subcortical projections of the mammalian neocortex. Although the area-specific

adult distributions of projection neurons are sculpted primarily through selective axon elimination, areal differences in the initial distribution of projection neurons have been reported. For example, the adult M1 has a higher density of layer 5 corticospinal neurons than does S1; prior to the phase of axon elimination this difference is present but not to the same degree. Therefore, even in this situation, selective axon elimination contributes to the generation of the distribution and density of projection neurons found in adult brains.

Area-Specific Thalamocortical Afferent Input and Potential Roles in Area Patterning

To sum, all classes of projection neurons initially exhibit an 'exuberant' distribution far broader than that found in the adult brain. In contrast to this lack of area specificity in the early distribution of projection neurons, the area-specific projection of TCAs originating in the principal sensory nuclei of dorsal thalamus is evident even at the early stages in their development. Progress has been made in defining mechanisms of TCA pathfinding, particularly subcortically from dorsal thalamus to the neocortex, but the molecular control of TCA targeting of specific areas remains poorly defined relative to mapping in other systems, such as retinal projections. As in the visual system, area-specific TCA targeting is likely primarily controlled intracortically by guidance molecules, and influenced by neural activity. SP neurons and their axons have also been implicated in the development of area-specific TCA targeting, but their role and the molecular basis are vague.

TCAs form the major input to the neocortex, and relay visual, auditory, and somatic sensations to the primary sensory areas of the neocortex in an area-specific manner. Since TCAs are the sole source of modality-specific sensory information to the neocortex, the functional specializations of the primary sensory areas are defined by, and dependent upon, TCA input. The differentiation of anatomical features that distinguish cortical areas, including architecture and distributions of output projection neurons, depends to a large extent upon TCA input. Consistent with this role, the TCA projection exhibits area specificity throughout its development, and the gradual differentiation of areas within the CP parallels the elaboration of the TCA projection within it. In addition, a variety of manipulations, including peripheral manipulations and transplantation experiments, have demonstrated that the CP exhibits considerable plasticity in the development of area-specific features, and that diverse parts of the CP initially have similar potentials to develop features unique to a specific area. Again, TCA input has been implicated as a major influence controlling this plasticity in the differentiation of area-specific features.

The role of TCAs in shaping cortical architecture is not limited to these later events in the differentiating CP. *In vitro* experiments suggest that TCAs release a diffusible mitogenic activity that promotes the production of both glia and neurons by explants of the cortical VZ. If a similar mechanism operates *in vivo*, such an early influence of TCAs on corticogenesis could contribute to reported areal differences in neuronal production in the SVZ in occipital areas (V1 vs. V2), and therefore, as previously described, to the cytoarchitectural differences between areas that become evident later in development.

Indirect Evidence for Intrinsic Genetic Regulation of Area Identity

Evidence for the genetic control of arealization is recent and limited, but compelling. The initial evidence was indirect and based upon the differential expression patterns of numerous genes, ranging from TFs, cell adhesion molecules, and axon guidance receptors and ligands, within cortical progenitors in the VZ or their progeny in the CP. These patterns, which were typically graded across the anteroposterior (AP) and dorsoventral (DV) cortical axes, were shown to develop independently of TCA input by analyses of mutant mice with targeted deletion of TFs (Gbx2 or Mash1, neither of which is expressed in the cortex but which are required for TCAs to reach the cortex). However, the most dramatic areal changes in gene expression, and changes that match expression patterns to cortical areas or their borders, occur postnatally, coincident with the development of patterned TCA input to the CP. Thus, although the early differential expression of many genes may be independent of TCA input, TCAs may later influence the differentiation of more complex patterns.

Roles for Morphogens and Transcription Factors in Control of Area Identity

Two different major forms of gene products influence the intrinsic generation of positional or area identity in the developing neocortex. These two forms of proteins are TFs, which are typically localized to the nucleus of the producing cell, and secreted morphogens, which are secreted by the producing cells. TFs are expressed by neocortical progenitors and directly regulate the positional or area identity that is inherited by their progeny, whether these progeny are additional progenitors or are neurons that form the CP and SP. Secreted morphogens are produced in discrete signaling centers and are believed to be secreted and to diffuse from their source, forming gradients across the dorsal telencephalon (neocortical) VZ that are highly

concentrated at the site of production and progressively diminish in concentration distal to their source. Here we first consider TFs that impart area identities to cortical progenitors, then in the subsequent sections consider roles for morphogens in area patterning.

Intrinsic Control of Area Identity by Differential Expression of Transcription Factors in Cortical Progenitors

The key criteria for a gene that specifies area identity in cortical progenitors is straightforward: it must be expressed by progenitors in the VZ and/or SVZ and must function in a differential manner across the cortical axes, either through differential expression or expression of cofactors or other mechanisms that differentially influence its function, and must impart areal properties. Functionally, genes that regulate arealization, in principle, could have a range of effects – for example, from conferring the complete set of properties that comprise the area identity of a cortical neuron, to a subset of these properties, to regulating the expression of axon guidance molecules that control the area-specific targeting of TCAs. Scores of genes meet the differential expression criteria to be the source of TFs that regulate area patterning, but as of this writing, only four TFs have been shown to function in regulating arealization: Emx2, Pax6, COUP-TFI, and Sp8. Emx2 is a homeodomain TF related to the *Drosophila* gene *empty spiracles* (*ems*), Pax6 is a paired box domain TF, COUP-TFI is an orphan nuclear receptor, and Sp8 is a zinc-finger TF related to the *Drosophila* gene *buttonhead* (*btd*).

Within the cortex, all TFs but COUP-TFI have their expression almost exclusively limited to progenitors in the embryonic VZ. COUP-TFI is expressed in a high caudolateral to low rostromedial gradient both in progenitors and in their CP progeny. Emx2 is expressed in a low rostrolateral to high caudomedial gradient, and Pax6 is expressed in an opposing pattern, a high rostrolateral to low caudomedial gradient across the VZ of embryonic cortex. All three of these TFs are expressed in progenitors throughout embryonic cortical neurogenesis. In contrast, Sp8 is expressed only during the early phase of neurogenesis, and is expressed in a high rostromedial to low caudolateral gradient. Loss-of-function studies have been done in mice for each of the four genes, and gain-of-function studies have been reported for all but the gene encoding COUP-TFI.

Emx2 To date, of the four TFs that function in regulating arealization, Emx2 has been most extensively studied. Genetic manipulations that change the levels of Emx2 expression in cortical progenitors result in disproportionate changes in the sizes of the primary sensory and motor cortical areas. For example, increasing Emx2 expression in neocortical progenitors has no effect on overall cortical size, but results in a significant decrease in the sizes of rostral cortical areas, including S1 and motor areas, and a significant increase in caudal areas, including V1. Decreasing levels of Emx2 expression has the opposing effect on area sizes. Changes in the levels of Emx2 in cortical progenitors appear to result in a complete change in the area identity of their neuronal progeny, to match the contracted or expanded areas.

COUP-TFI Cortical deletion of COUP-TFI results in a massive expansion of anterior areas, including motor, to occupy most of parietal and occipital (sensory) cortex, paralleled by substantial reduction in the sizes of the three primary sensory areas that become compressed and aligned mediolaterally along the caudal pole of the cortical hemisphere. Thus, COUP-TFI is required to balance the patterning of neocortex into motor and sensory areas, and appears to operate through a novel genetic pathway to repress the identities of rostral cortical areas, such as motor, allowing for appropriate specification of the sensory cortical areas. COUP-TFI might also have a role in specifying the identities of sensory areas.

Pax6 Roles for Pax6 in area patterning are presently vague. The initial studies that implicated Pax6 in area patterning depended upon marker analyses of the small eye mutant mouse; these mice are deficient for functional Pax6 protein and die at birth, before cortical areas differentiate, and in addition lack TCA input. Nonetheless, these marker analyses implicated Pax6 in specifying anterior area identities associated with motor areas, consistent with its highest expression in progenitors that give rise to anterior areas. However, a recent gain-of-function study of Pax6 that used a yak transgenic approach to overexpress Pax6 severalfold in cortical progenitors reports minimal or no changes in area patterning. This discrepancy could be explained in several ways. One appealing way is that another gene, for example, that for COUP-TFI, normally represses, in the cortical fields, Pax6 function that would give rise to sensory areas, and therefore represses the effect of Pax6 overexpression. In any case, additional work must be done to properly define the role for Pax6 in the regulation of area patterning.

Sp8 Sp8 has been recently shown to have multiple distinct roles in area patterning, one by its influence on the expression of the morphogen fibroblast growth factor 8 (FGF8) within the commissural plate (CoP), and the other through its expression directly

in cortical progenitors. Sp8 is expressed in a high to low, anteromedial to posterolateral, gradient across the VZ, and is also transiently expressed in the CoP coincident with the expression domain of FGF8, a morphogen implicated in area patterning of the neocortex. Sp8 and FGF8 exhibit reciprocal induction in the embryonic telencephalon, but ectopic expression of Sp8 and FGF8 in anterior dorsal telencepahlon around the normal FGF8 expression domain in the CoP has opposing effects on area patterning, with ectopic expression of FGF8 inducing a posterior shift of cortical areas and ectopic expression of Sp8 inducing an anterior shift – suggesting that in parallel to regulating FGF8 expression, Sp8 also activates a distinct signaling pathway for cortical area patterning. Further, deletion of Sp8 from cortical progenitors results in a reduction of anterior cortical areas, such as motor areas.

Interactions between regulatory genes in controlling area patterning Other TFs are likely involved as primary regulators of area patterning and cooperate with Emx2, Pax6, COUP-TFI, and Sp8. In addition, Emx2, Pax6, COUP-TFI, and Sp8 have inductive or repressive effects upon one another that affect their influence, their level of expression, and even their domain of expression. For example, Sp8 is a direct transcriptional activator of FGF8, and Sp8 induction of FGF8 is repressed by Emx2, which binds Sp8. This provides a mechanism to limit FGF8 expression to the CoP. In addition, many of these TFs influence the expression of one another. For example, Emx2 and COUP-TFI appear to repress Pax6, and Pax6 appears to repress Emx2. Further, as described in the following sections, FGF8 influences the expression of many of these TFs. In summary, gradients of TFs expressed within cortical progenitors in the VZ/SVZ impart area identities, and cooperate with one another, and with as yet unidentified TFs, to activate signaling pathways that control area patterning of the neocortex.

Morphogens/Signaling Molecules Establish Graded Patterns of Transcription Factor Expression in Cortical Progenitors

Current evidence indicates that morphogens secreted by signaling centers located at the perimeter of the developing dorsal telencephalon (dTel) establish gradients of TFs across the dTel VZ, which in turn determine the area fate of cortical progenitors and their progeny. These morphogens include FGF8 expressed by the anterior neural ridge (ANR), which later becomes the CoP, located at the anterior midline of the dTel, and also members of the wingless-int (Wnt) and bone morphogenetic protein (BMP) families, expressed by the cortical hem and choroid plexus epithelium (CPe), located at the dorsal midline of more posterior dTel (**Figure 2**).

Recent studies have begun to better define these patterning centers and, early in development, the action of their signaling molecules to establish and maintain across the neocortical VZ the graded expression of TFs that directly impart area identity to cortical progenitors. The initial domain of FGF8 expression in the ANR/CoP is located at the rostral edge of the prospective neocortex, and later this expression domain extends caudally along the dorsal midline toward diencephalon. Bone morphogenetic proteins (BMP2, BMP4, BMP5, BMP6, and BMP7) and Wnts (Wnt2b, Wnt3a, and Wnt5a) are expressed in the cortical hem, a midline structure adjacent to the hippocampal anlage.

FGF8 FGF8 is the intensively studied member of the FGF family that is expressed in the ANR/CoP. At the neural tube stage, approximately embryonic day 8.5 (E8.5) to E9.5 in mouse, the FGF8-expressing domain is positioned in a discrete dorsomedial domain. After invagination of the neural tube (~E10.5), the FGF8 expression domain is positioned in a small domain in the medial wall of the two cortical hemispheres called the ANR, with the highest expression of FGF8 localized in the anterior region of the cortex. FGF8 expression continues robustly through E12.5 in mouse and then tapers off significantly. Experimental evidence shows that the high rostromedial expression of FGF8 establishes, at least in part, the graded expression of Emx2 and COUP-TFI through repression. Further, changes in levels of FGF8 in embryonic mice result in area shifts that mimic those expected by decreasing or increasing the graded expression of Emx2 across the neocortical VZ, and are likely due to such an effect on Emx2 levels (**Figure 3**).

Wnts and BMPs A second signaling center is the cortical hem that expresses members of the Wnt and BMP families of ligands. These factors begin to be expressed slightly later than FGF8 (~E9.5) and are expressed immediately lateral to the FGF8 domain during the neural tube stage. After invagination of the neural tube, Wnts and BMPs are localized within a domain called the cortical hem, which is positioned just immediately dorsal to the FGF8 expression domain. Along the anteroposterior axis of the dTel, the cortical hem shares a similar anterior boundary with FGF8 but extends significantly further caudally within the medial cortical wall. Expression of signaling molecules from the cortical hem appears to be active until slightly later in development, as expression of Wnts and BMPs remains

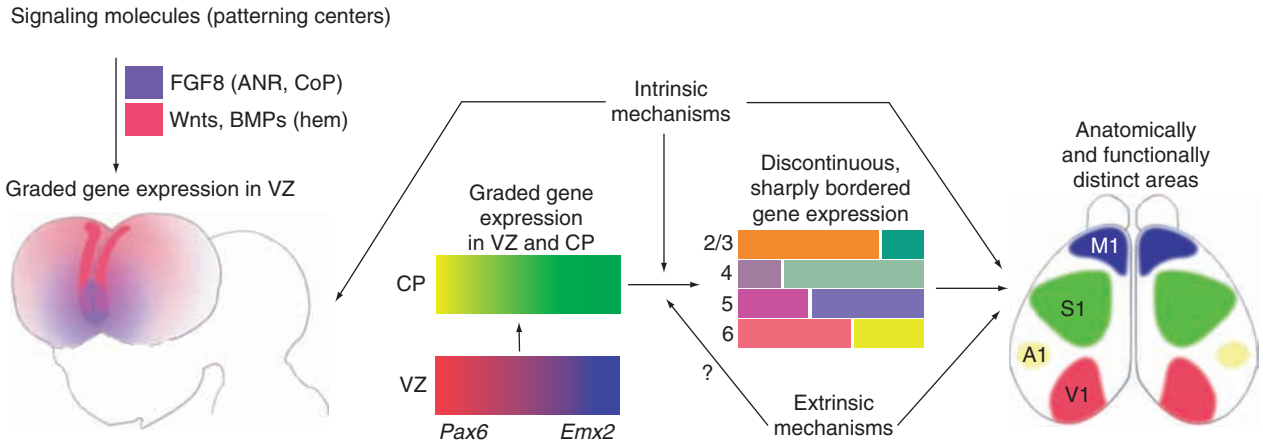

Figure 2 Mechanisms of specification and differentiation of neocortical areas. The initial, tangential gradients of transcription factors in the ventricular zone (VZ) are likely established by signaling molecules/morphogens secreted from patterning centers, such as fibroblast growth factor 8 (FGF8) from anterior neural ridge (ANR), which later becomes the commissural plate (CoP), and Wnts and bone morphogenetic proteins (BMPs) from the cortical hem. The graded expression of certain transcription factors, such as Pax6, Emx2, COUP-TFI, and Sp8, imparts positional or area identities to cortical progenitors, which is imparted to their neuronal progeny that form the cortical plate (CP). The CP also initially exhibits gradients of gene expression that are gradually converted to distinct patterns with sharp borders. Coincident with this process, distinct cortical layers (2–6), and the anatomically and functionally distinct areas seen in the adult (M1, motor; S1, somatosensory; A1, auditory; V1, visual), differentiate from the CP. Genes that are differentially expressed across the cortex are often expressed in different patterns in different layers, suggesting that area-specific regulation of such genes is modulated by layer-specific properties, raising questions about the definition of area identity. Although the initial establishment of the graded gene expression in the embryonic CP is controlled by mechanisms intrinsic to the telencephalon, the much more complex differentiation patterns established postnatally, often both spatially and temporally in parallel with the development of thalamocortical afferent input, suggest that these later events might well be regulated influences that arise extrinsic to the cortex, such as thalamocortical afferents. Future studies are required to determine and clarify these mechanisms. Adapted from O'Leary DDM and Nakagawa Y (2002) Patterning centers, regulatory genes and extrinsic mechanisms controlling arealization of the neocortex. *Current Opinion in Neurobiology* 12: 14–25.

strong until approximately E13.5, after which the cortical hem shrinks in size, develops into the fimbria (an axonal tract of the hippocampal formation), and slowly downregulates its expression of these signaling molecules. Although Wnts have been shown to enhance proliferation of cortical progenitors, little evidence implicates Wnts and BMPs in regulating area patterning.

Anteroposterior Patterning of the Cerebral Cortex: Are Arealization and Regionalization Related?

For TFs that exhibit low to high AP graded expression patterns in the neocortex, their graded expression patterns often continue beyond the posterior extent of the neocortex and into other regions of the cerebral hemisphere and through the hippocampal fields. Examples of TFs that function in neocortical arealization and exhibit this form of continuing graded expression include Emx2 and COUP-TFI. These continuous, graded expression patterns suggest a relationship between arealization of the neocortex and arealization of other regions of the cerebral cortex, as well as a relationship between arealization and regionalization of the cerebral cortex. This observation warrants further investigation into these potential relationships.

Translation of Graded Expression of TFs by Cortical Progenitors and Their CP Progeny into Sharp Borders Exhibited by Cortical Areas

The graded expression of TFs expressed by cortical progenitors in the VZ/SVZ, and early on by their neuronal progeny in the CP, eventually transforms into sharply bordered patterns of gene expression that often relate to borders of cortical areas. Studies in *Drosophila* embryos have suggested mechanisms that might operate in the mammalian cortex to accomplish this change. For example, the graded distribution of dorsal, a regulatory protein, generates (through concentration-dependent differences in binding efficacy to promoter and repressor elements) expression patterns of downstream genes with sharp borders that align with the boundaries of different embryonic tissues and related patterns of gene expression. A different example is the development of the sharply patterned expression of the *even-skipped* gene through the combined

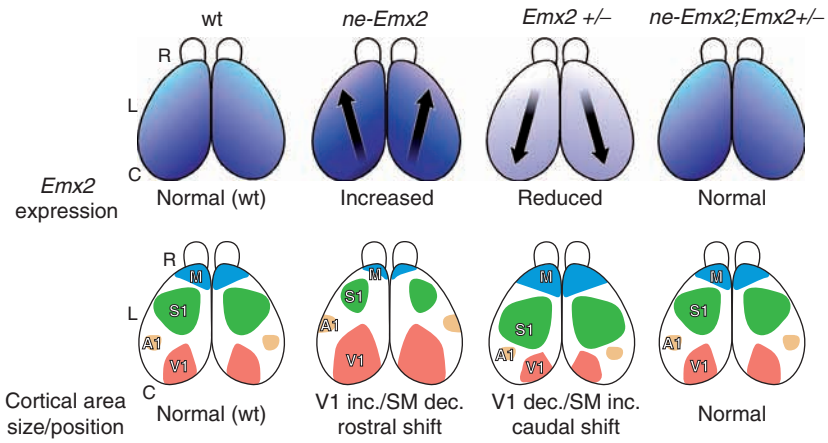

Figure 3 The *Emx2* gene imparts area identities to cortical progenitors and their progeny: levels of *Emx2* expression control sizes of cortical areas. Top panel shows relative levels of graded *Emx2* expression in cortical progenitors in four genotypes: wild type (wt), with normal levels of *Emx2*; *ne-Emx2* transgenic mice, in which an *Emx2* transgene is driven by a nestin promoter in cortical progenitors and results in about a 50% increase in overall *Emx2* levels; *Emx2* heterozygous mice ($Emx2^{+/-}$) with only one wt *Emx2* allele, resulting in about a 50% decrease in overall *Emx2* levels; and *ne*-Emx2; $Emx2^{+/-}$ mice with overall levels of *Emx2* being essentially normal due to the combination of endogenous *Emx2* and the *Emx2* transgene. The arrows show the predicted and observed shifts in area patterning. The bottom panel shows area patterning using a schematic of the four primary areas, motor (M), somatosensory (S1), auditory (A1), and visual (V1). The *ne-Emx2* transgenic mice exhibit decreased sizes of S1 and M, an increase in V1, and a rostral shift in all areas. The $Emx2^{+/-}$ mice exhibit increased sizes of S1 and M and decreased size of V1. The sizes of the primary cortical areas in the *ne-Emx2*; $Emx2^{+/-}$ mice revert toward their normal sizes observed in wt. Therefore, the sizes of the areas depend on the overall level of graded *Emx2* expression, and changes in area sizes due to increases or decreases in *Emx2* can be rescued by restoring *Emx2* to wild-type levels in cortical progenitors. Adapted from Hamasaki T, Leingartner A, Ringstedt T, et al. (2004) EMX2 regulates sizes and positioning of the primary sensory and motor areas in neocortex by direct specification of cortical progenitors. *Neuron* 43: 359–372; and from Leingartner A, Thuret S, Kroll TT, et al. (2007) Sizes of somatosensory and motor cortical areas determine proficiency at tactile and motor behaviors. *Proceedings of the National Academy of Sciences of the United States of America* 104: 4153–4158, Copyright (2007) National Academy of Sciences, USA.

action of multiple activators and repressors of its transcription; *even-skipped* is expressed in stripes perpendicular to the AP axis of the *Drosophila* axis where expression of repressors is subthreshold and that of activators is suprathreshold.

Similar mechanisms appear to operate in the developing spinal cord, where sonic hedgehog (shh), a gene product secreted by the notocord and floorplate, represses or induces the expression of different classes of TFs in the VZ of ventral spinal cord. Initially, these TFs have graded, overlapping expression patterns which progressively sharpen through mutual repression, to generate sharply bordered patterns of expression. This mechanism generates genetically distinct domains of progenitors defined by their expression of unique subsets of TFs, which in turn generate unique classes of spinal interneurons and motor neurons.

Similar mechanisms likely operate in the neocortex to generate areas, but important differences are evident. For example, at no time during neocortical neurogenesis are sharply bordered patterns of regulatory genes observed in the neocortical VZ; all have graded expression patterns. Initially, even expression in the CP is graded, but at later stages of development many genes acquire expression patterns with an abrupt border that, in some instances, matches the border between areas.

What Is 'Area Identity'?

The findings and issues discussed in the preceding sections lead directly to a major issue that remains unsettled: the extent to which areas are genetically distinct. Area-specific genes *per se* either do not exist or are very rare at the developmental stages when area identities are being genetically specified, and perhaps even later, when the anatomical features that distinguish areas begin to differentiate within the CP. Thus, in terms of gene expression, a neocortical area is not defined by the expression of a specific set of genes restricted to that area. Instead, a neocortical area is defined by the expression of a unique subset of genes, each of which is also expressed in other areas. However, the actual scenario is even more complex, since each layer has a unique profile of gene expression. Each gene differentially expressed in the neocortex, and expressed in more than one layer, has different expression patterns in each layer. An excellent example is *MDGA1*: this gene encodes a protein that is an immunoglobulin cell adhesion molecule. It is expressed in layers 2/3 through the neocortex,

>and is also expressed in layers 4 and 6 within S1. Thus, it has both layer- and area-specific patterns of expression, and can be used to define an area, S1, but its expression is not limited to that area. Thus, the term 'area identity' in the strictest sense might not truly exist for neurons across layers, and uniquely marks neurons with the specific identity of a given area.

See also: Brain Development: The Generation of Large Brains; Cerebral Cortex: Inhibitory Cells; Cerebral Cortex: Symmetric vs. Asymmetric Cell Division; Cerebral Cortex; Evolution of Vertebrate Brains; Frontal Cortex Evolution in Primates; Memory Consolidation: Cerebral Cortex; Motor Neuron Specification in Vertebrates; Neocortex: Origins; Prefrontal Cortex: Structure and Anatomy; Spontaneous Patterned Activity in Developing Neural Circuits.

Further Reading

Bishop KM, Goudreau G, and O'Leary DDM (2000) Regulation of area identity in the mammalian neocortex by Emx2 and Pax6. *Science* 288: 344–349.

Corbin JG, Nery S, and Fishell G (2001) Telencephalic cells take a tangent: Non-radial migration in the mammalian forebrain. *Nature Neuroscience* 4: 1177–1182.

Fukuchi-Shimogori T and Grove EA (2001) Neocortex patterning by the secreted signaling molecule FGF8. *Science* 294: 1071–1074.

Garel S, Huffman KJ, and Rubenstein JLR (2003) Molecular regionalization of the neocortex is disrupted in Fgf8 hypomorphic mutants. *Development* 130: 1903–1914.

Hamasaki T, Leingartner A, Ringstedt T, et al. (2004) EMX2 regulates sizes and positioning of the primary sensory and motor areas in neocortex by direct specification of cortical progenitors. *Neuron* 43: 359–372.

Kriegstein AR and Noctor SC (2004) Patterns of neuronal migration in the embryonic cortex. *Trends in Neuroscience* 27: 392–399.

Krubitzer L and Kaas J (2005) The evolution of the neocortex in mammals: How is phenotypic diversity generated? *Current Opinion in Neurobiology* 15: 444–453.

Leingartner A, Thuret S, Kroll TT, et al. (2007) Sizes of somatosensory and motor cortical areas determine proficiency at tactile and motor behaviors. *Proceedings of the National Academy of Sciences of the United States of America* 104: 4153–4158.

Lukaszeqicz A, Savatier P, Cortay V, et al. (2005) G1 phase regulation, area-specific cell cycle control, and cytoarchitectonics in the primate cortex. *Neuron* 47(3): 353–364.

Mallamaci A, Muzio L, Chan CH, et al. (2000) Area identity shifts in the early cerebral cortex of Emx2$^{-/-}$ mutant mice. *Nature Neuroscience* 3: 679–686.

Manuel M, Georgala PA, Carr CB, et al. (2007) Controlled overexpression of Pax6 *in vivo* negatively autoregulates the Pax6 locus, causing cell-autonomous defects of late cortical progenitor proliferation with little effect on cortical arealization. *Development* 134: 545–555.

Marin O and Rubenstein JLR (2003) Cell migration in the forebrain. *Annual Review of Neuroscience* 26: 441–483.

Miyashita-Lin EM, Hevner R, Wassarman KM, et al. (1999) Early neocortical regionalization in the absence of thalamic innervation. *Science* 285: 906–909.

Nakagawa Y, Johnson JE, and O'Leary DDM (1999) Graded and areal expression patterns of regulatory genes and cadherins in embryonic neocortex independent of thalamocortical input. *Journal of Neuroscience* 19: 10877–10885.

O'Leary DDM (1989) Do cortical areas emerge from a protocortex? *Trends in Neuroscience* 12: 400–406.

O'Leary DDM and Koester SE (1993) Development of projection neuron types, axonal pathways and patterned connections of the mammalian cortex. *Neuron* 10: 991–1006.

O'Leary DDM and Stanfield BB (1989) Selective elimination of axons extended by developing cortical neurons is dependent on regional locale. Experiments utilizing fetal cortical transplants. *Journal of Neuroscience* 9: 2230–2246.

O'Leary DDM and Nakagawa Y (2002) Patterning centers, regulatory genes and extrinsic mechanisms controlling arealization of the neocortex. *Current Opinion in Neurobiology* 12: 14–25.

Rakic P (1988) Specification of cerebral cortical areas. *Science* 241: 170–176.

Rash BG and Grove EA (2006) Area and layer patterning in the developing cerebral cortex. *Current Opinion in Neurobiology* 16: 25–34.

Sahara S, Kawakami Y, Izpisua Belmonte JC, et al. (2007) *Sp8* exhibits reciprocal induction with *Fgf8* but has an opposing effect on anterior-posterior cortical area patterning. *Neural Development* 2: 10 (online; doi:10.1186/1749-8104-2-10).

Shimogori T, Banuchi V, Ng HY, et al. (2004) Embryonic signaling centers expressing BMP, WNT and FGF proteins interact to pattern the cerebral cortex. *Development* 131: 5639–5647.

Sur M and Rubenstein JLR (2005) Patterning and plasticity of the cerebral cortex. *Science* 310: 805–810.

Takeuchi A, Hamasaki T, Litwack ED, et al. (2007) Novel IgCAM, MDGA1, expressed in unique cortical area- and layer-specific patterns and transiently by distinct forebrain populations of Cajal-Retzius neurons. *Cerebral Cortex* 17: 1531–1541.

Zembrzycki A, Griesel G, Stoykova A, et al. (2007) Genetic interplay between the transcription factors Sp8 and Emx2 in the patterning of the forebrain. *Neural Development* 2:8(online; doi:10.1186/1749-8104-2-8).

Zhou C, Tsai SY, and Tsai MJ (2001) COUP-TFI: An intrinsic factor for early regionalization of the neocortex. *Genes & Development* 15: 2054–2059.

Neural Patterning: Eye Fields

M Noda, H Takahashi, and H Sakuta, National Institute for Basic Biology, Okazaki, Japan

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Patterning of the central nervous system (CNS) has a crucial role in shaping functional organization. Owing to its external position in the embryo, the eye is often used as a model to study CNS patterning. Establishment of the anteroposterior (A-P) and dorsoventral (D-V) axes constitutes a fundamental step in retinal development. Several morphogens and transcription factors expressed asymmetrically in the early retina play key roles in the establishment of the A-P and D-V axes of the retina. The neural retina is a highly organized sensory organ that receives, integrates, and transmits visual information. Light is captured by the photoreceptor cells, and its information is converted into chemical signals to be sent through a series of interneurons to the projection neurons of the retina, the retinal ganglion cells (RGCs).

During embryonic development, the axons of RGCs grow from the eye to the contralateral optic tectum (OT), the primary visual center in lower vertebrates, or the superior colliculus (SC) in mammals and form a topographic map of visual space. RGC axons emanating from the anterior (nasal) retina project to the posterior OT/SC, while those from the posterior (temporal) retina project to the anterior OT/SC. Likewise, axons from the dorsal retina project to the lateral (L, or ventral) OT/SC, while those from the ventral retina project to the medial (M, or dorsal) OT/SC. For the formation of topographic maps, molecular gradients in origins and targets are essential, as predicted by Sperry in the classical chemoaffinity hypothesis. The family of Eph receptor tyrosine kinases and their ligands, the ephrins, have been shown to fulfill these criteria and control topographic map development.

Members of the EphA family expressed with a temporal high-nasal low gradient in the retina and of the ephrinA family expressed with a posterior high-anterior low gradient in the OT/SC determine the retinotectal projection along the A-P axis. The receptors of the EphB family expressed with a ventral high-dorsal low gradient in the retina, and the ephrinB ligands expressed with a medial high-lateral low gradient in the OT/SC control the mapping of retinal axons along the M-L axis. Recent advances in the study of the molecular mechanisms for retinal patterning are discussed below, and the molecular cascade controlling the events in the chick are highlighted. It is not yet clear whether the patterning mechanism revealed in the chick retina is wholly valid for mammals.

Retinal Patterning along the A-P Axis

When is the A-P axis determined in the developing retina? Fate mappings of the avian anterior forebrain have revealed that there is a distinct area for the presumptive retina in the neural plate, and the A-P axis of the presumptive territory in the neural plate corresponds to the A-P axis of the retina. In the A-P subdivision during the brain's development, boundaries are established at the zona limitans intrathalamica and between the forebrain and midbrain and the midbrain and hindbrain by mutual repression between *Six3* and *Irx3*, *Pax6* and *Pax2*/*En1*, and *Otx2* and *Gbx2*, respectively. Fibroblast growth factor-8 (*FGF-8*) is expressed at the anterior neural ridge and the midbrain–hindbrain boundary and helps to further subdivide the brain. The homeobox genes modulate the response to FGF-8 in their territories – within the *Six3* territory in the anterior forebrain, FGF-8 can induce the expression of *FoxG1*, a winged-helix/forkhead (Fox) transcription factor, whereas in the *Irx3* domain, FGF-8 in the midbrain–hindbrain boundary activates *En1*.

The forebrain is further subdivided by the mutual repression between *FoxG1* (also referred to as Brain Factor 1, *BF1*) and the related gene *FoxD1* (Brain Factor 2, *BF2*). These two *Fox* transcription factors are expressed in adjacent domains in the prospective eye region at the time the optic vesicle evaginates. It is thus considered that the establishment of the retinal polarity along the A-P axis is triggered by FGF signaling. In the chick embryo, misexpression of *FGF-8* upregulates the expression of *FoxG1* and downregulates that of *FoxD1* in the optic vesicle. After the formation of the optic cup, *FoxG1* is expressed in the nasal retina and *FoxD1* is expressed in the temporal retina in a complementary pattern by counteracting each other. Because surgical manipulations of the chick optic vesicle begin to induce severe disturbances to the visual projection map from Hamburger–Hamilton (HH) stage 11/12 (on E1.5), retinotopic specification along the A-P axis of the retina is considered to be determined at or prior to the optic vesicle stage. Ectopic misexpression of *FoxG1* or *FoxD1* in the chick retina reversed the topographic map in the retinotectal projection along the A-P axis. These factors are thus supposed to play a key role in

the initiation of retinal regional specification along the A-P axis.

Subsequently, two homeobox transcription factors, *GH6* and *SOHo1*, are known to be expressed specifically in the nasal region of the developing chick retina. Both *GH6* and *SOHo1* are downstream targets of the Fox genes. Misexpression of *FoxG1* induces the expression of *GH6* and *SOHo1* in the temporal retina. In contrast, *FoxD1* inhibits the expression of these two homeobox genes in the retina. Misexpression of *GH6* or *SOHo1* in the retina results in errors in the projection of retinal axons to the OT along the A-P axis, due to the repression of *EphA3* expression in the temporal retina.

In contrast to *EphA3*, four EphA receptors (*EphA4*, *EphA5*, *EphA6*, and *EphA7*) are uniformly expressed in the chick retina. On the other hand, *ephrinA2* (see below) and *ephrinA5* are also expressed with a nasal high-temporal low gradient in the retina. Overexpression of ephrinAs in temporal axons leads to errors in the topographic targeting of temporal axons, suggesting a role for retinal ephrinAs in the formation of topographic projections. Thus, EphAs uniformly expressed in the retina are also thought to be involved in the topographic projection along the A-P axis, with the help of ephrinAs expressed with a gradient in the retina.

It has been revealed that *FoxG1* controls the expression of all the molecules distributed asymmetrically along the A-P axis through multiple mechanisms in the chick. *FoxG1* is known to act as a transcriptional repressor. The expression of downstream targets of *FoxG1* in the retina is affected by the misexpression of a chimeric protein that consists of a *Drosophila* even-skipped repression domain and a winged-helix (WH) DNA-binding domain of FoxG1. The repressing construct of FoxG1 regulates the expression of *SOHo1*, *GH6*, *EphA3*, *FoxD1*, and *ephrinA5*, similar to the wild type, but not the expression of *ephrinA2*. On the other hand, mutant FoxG1 deficient in DNA-binding ability exerts similar effects on the expression of *SOHo1*, *GH6*, *EphA3*, *FoxD1*, and *ephrinA2* as the wild-type FoxG1, but not on the expression of *ephrinA5*. These results suggest that *FoxG1* controls *ephrinA5* by a DNA-binding-dependent mechanism, *ephrinA2* by a DNA-binding-independent mechanism, and *FoxD1*, *SOHo1*, *GH6*, and EphA3 by dual mechanisms. It is important to note that FoxG1 interferes in bone morphogenetic protein 2 (BMP2) signaling and thereby induces oblique-gradient expression of ephrinA2. The BMP2 signal resultantly plays a pivotal role in the topographic projection along both axes (see below).

What would be the point of using two distinct mechanisms, DNA binding-dependent and DNA binding-independent, to regulate the expression of topographic molecules? One answer is for security to maintain regulation of the EphA/ephrinA system. For instance, if the DNA binding-dependent mechanism is lost because of a mutation in the WH domain, FoxG1 can still regulate the asymmetrical distribution of *EphA3* and *ephrinA2* through the remaining DNA binding-independent mechanism. On the other hand, if the DNA binding-independent mechanism is lost, FoxG1 can still regulate the asymmetrical distribution of *EphA3* and *ephrinA5* through the DNA binding-dependent mechanism. Gene knockout of *ephrinA2* and *ephrinA5* in mice suggests the importance of this redundancy: in the single knockout mice, a substantial proportion of the retinal axons projected normally onto the SC. Since the total EphA/ephrinA system plays an essential role in the formation of the retinotectal map, this dual regulatory system might have evolved during the evolution of the visual system.

In *FoxD1*-deficient mice, ephrinA(s) is highly expressed in both halves of the retina. Misexpression of *FoxD1* inhibits the expression of *ephrinA5* in the chick nasal retina. These results suggest that *FoxD1* is essential for the establishment of the temporal regional specificity in the retina. To reveal the function of *FoxD1* in detail, further experiments are needed. In addition, there is the possibility that yet unknown factors, which are expressed specifically in the temporal retina as the counterpart of *SOHo1* and *GH6*, regulate the regional specification of the temporal retina. *FoxG1* and *FoxD1* are thus thought to be located at the top of the gene cascade for the regional specification along the A-P axis and determine the nasal or temporal specificity in the retina through multiple mechanisms. The cascade of control genes for the formation of the A-P axis of the chick retina is shown in **Figure 1**.

Mouse *FoxD1* plays a role in the specification of the ventrotemporal (VT) retina. In the chick, ipsilateral projection of retinal axons is not observed. In contrast, in mammals, ipsilaterally projecting axons arise from the VT region. The decision of RGCs from the VT retina to project ipsilaterally at the optic chiasma is controlled by the zinc-finger transcription factor *Zic2*. The VT axons are positive for EphB1 and repelled by ephrinB2 expressed in the midline glial cells at the optic chiasma. Conversely, an LIM homeodomain transcription factor, *Islet2*, is expressed in a pattern complementary to *Zic2*, repressing *Zic2* expression in the retina. In the *FoxD1*-deficient mice, the VT axons aberrantly project contralaterally because the expression of *Zic2* and *EphB1* is missing. It is not clear whether *FoxD1* also regulates the ipsilateral projection in other species such as *Xenopus*, in which ipsilateral projections develop after metamorphosis.

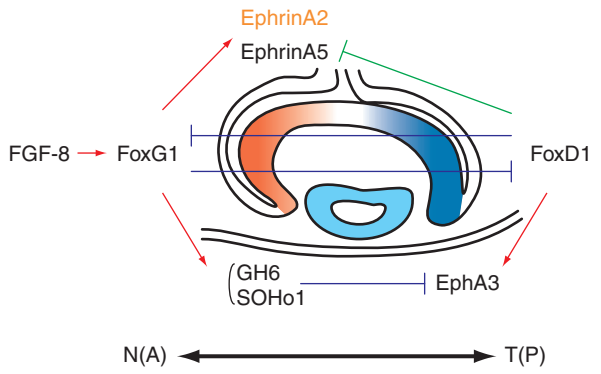

Figure 1 Molecular control of anteroposterior (A-P) patterning. FGF-8 derived from the anterior neural ridge induces *FoxG1* expression in the retina. The counteraction between *FoxG1* on the nasal side and *FoxD1* on the temporal side governs the retinal patterning along the A-P axis early on in development. Subsequently, the transcription factors GH6 and SOHo1 in the nasal retina begin to be expressed under the control of *FoxG1*. *GH6* and *SOHo1* inhibit the expression of *EphA3*. *FoxG1* and *FoxD1* also counteract each other to determine the expression pattern of *ephrinA5*. *FoxG1* induces the expression of *ephrinA2*. Arrows and T-bars indicate positive and negative effects, respectively. These effects are not necessarily induced by direct action. Nasal (N; anterior, A) is left, and temporal (T; posterior, P) is right.

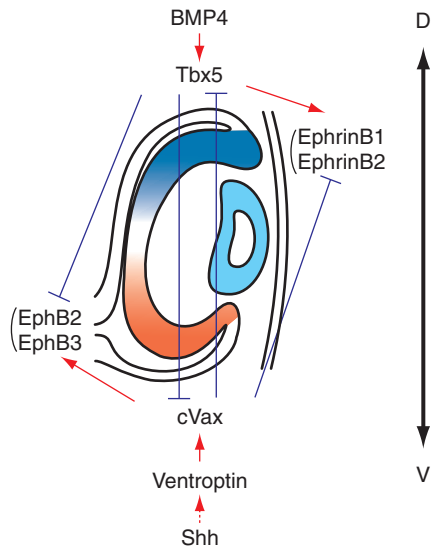

Figure 2 Molecular control of dorsoventral (D-V) patterning early in development. The counteraction between *BMP4* on the dorsal side and *Ventroptin* on the ventral side governs the retinal patterning along the D-V axis early in the developmental process. Subsequently, transcription factors Tbx5 in the dorsal retina and cVax in the ventral retina begin to be expressed under the control of the BMP4 signal. *Tbx5* and *cVax* counteract each other to determine the expression pattern of EphB receptors and their ligands, ephrinBs. Sonic hedgehog (Shh) derived from the ventral midline presumably induces expression of *Ventroptin* in the retina. Arrows and T-bars indicate positive and negative effects, respectively. These effects are not necessarily induced by direct action. Dorsal (D) is upwards, and ventral (V) is downwards.

Retinal Patterning along the D-V Axis at the Early Stages

Soon after A-P polarity is determined, initial D-V patterning also develops through the actions of morphogens and transcription factors. Figure 2 shows the schema for the establishment of the D-V axis of the chick retina early on. BMP signaling is required for the D-V patterning of the CNS in vertebrates. The retinal patterning along the D-V axis also appears to be controlled by a gradient of BMP4 signaling, emanating from the dorsal retina. Ventroptin, a BMP antagonist, in the ventral retina counteracts BMP4. Overexpression of *BMP4* in the chick retina upregulates a dorsally expressed T-box transcription factor, *Tbx5*, and downregulates a ventrally expressed *Emx* homeobox gene, *cVax* and *Ventroptin*. In the *BMP4* knockout mouse, the expression of *Tbx5* is lost. Overexpression of *Ventroptin* in the chick retina leads to a downregulation of *BMP4* and *Tbx5* expression and to upregulation of *cVax* expression. The counteraction between *BMP4* and *Ventroptin* thus governs the regional specification along the D-V axis in the retina by controlling the expression of the downstream transcription factors, *Tbx5* and *cVax*.

Sonic hedgehog (*Shh*) is a ventralizing morphogen in many areas of the CNS. Shh expression in the ventral neural tube and the underlying axial mesendoderm extends to the neural and mesodermal tissue located between the two eyes. Therefore, it has been suggested that Shh acts as a long-range morphogen and controls the D-V patterning in the retina, although it is not expressed in the retina. Consistent with this view, misexpression of *Shh* in the chick and *Xenopus* retina suppresses the expression of *BMP4* and *Tbx5* and enhances the expression of *cVax/Vax2*. Because *Ventroptin* also controls the expression of these molecules, Shh supposedly ventralizes the retina through induction of *Ventroptin* expression.

Tbx5 begins to be expressed in the dorsal retina at the early optic cup stages. Misexpression of *Tbx5* in the chick retina enhances the dorsal expression of *ephrinB1* and *ephrinB2* and represses the ventral expression of *EphB2*, *EphB3*, and *cVax*. Additionally, ectopic expression of *Tbx5* causes ventral retinal axons to exhibit defects in topographic mapping. As the optic cup invaginates, *cVax* expression commences in the ventral regions of the retina. Overexpression of *cVax* downregulates the expression of the dorsalizing factors, *BMP4* and *Tbx5*, and upregulates that of the ventral markers, *EphB2* and *EphB3*. When *cVax* is misexpressed in the chick retina, the dorsal axons do not project to their correct target zones, whereas ventral axonal projections appear to be normal.

Thus, *cVax* is sufficient for retinal ventralization. *Tbx5* and *cVax* counteract each other similar to the relation between *BMP4* and *Ventroptin*. Although they are located downstream of *BMP4* and *Ventroptin*, their expression domains have a gap of significant width, in which neither of them is expressed. Two other members of the T-box gene family, *Tbx2* and *Tbx3*, are expressed with a dorsal high-ventral low gradient the same as *Tbx5*, but they cover the gap region. They are also involved in the D-V patterning of the retina. Therefore, a combination of these transcription factors may control the retinal patterning and retinotectal projection along the D-V axis.

Retinoic acid (RA) is also asymmetrically distributed along the D-V axis in the developing retina. The presence of dorsal and ventral zones of RA activity in the retina separated by an RA-free middle zone is conserved among vertebrate species such as zebra fish, chick, and mouse. The two zones of RA activity are sculptured by the expression of two different RA-synthesizing enzymes, RALDH1 in the dorsal retina and RALDH3 in the ventral retina. The presence of a mid-zone free of RA is due to the expression of multiple members of the Cyp26 family of RA-degrading enzymes. RA has been proposed to control the D-V patterning in the retina. In support of this hypothesis, mice deficient in some retinoid receptors show marked defects in development of the ventral eye. Addition of RA to zebra fish embryos during the eye's development results in ventral retinal duplication, while addition of an inhibitor of the ventral RA-synthetic enzyme causes deletion of the ventral retina. These results indicate that RA is essential for proper development of the ventral eye. When analyzed by gene expression, RA treatment of *Xenopus* embryos causes a reduction of *Tbx5* expression and expansion of *Vax2* expression, suggesting that RA has ventralizing activity. However, inhibition of RA signaling in chick or mouse embryos by overexpression of dominant-negative RA receptors or a double knockout of *RALDH1* and *RALDH3* does not significantly alter the expression of *Tbx5* and *cVax/Vax2*, indicating that RA is not involved in the early D-V patterning of the retina in these species. The role of RA in retinal D-V patterning should be clarified by further studies.

Retinal Patterning along the D-V Axis at the Later Stages

The counteraction between *BMP4* on the dorsal side and *Ventroptin* on the ventral side governs the regional specification in the chick retina along the D-V axis early on in the developmental process (HH stage 11 to E5), as described above. At later stages (from E5 onward), in proportion to the disappearance of *BMP4* expression from the dorsal retina, *Ventroptin* begins to be expressed with both a nasal high-temporal low and a ventral high-dorsal low gradient (oblique gradient) in the chick retina, or V/N-high pattern. At the same time, *BMP2* begins to be expressed with an oblique-gradient pattern complementary to that of *Ventroptin* along the two axes, or D/T-high pattern. Thus, the counteraction between *BMP4* and *Ventroptin* is relieved by that between *BMP2* and *Ventroptin*. *FoxG1* represses *BMP2* expression through inhibition of the BMP signaling, and thereby the oblique-gradient expression patterns of *BMP2* and *Ventroptin* are induced. Expectedly, misexpression of *Ventroptin* in the developing chick retina alters the retinotectal projection not only along the M-L axis but also along the A-P axis, owing to the induction of *ephrinA2* expression. It has been long believed that the A-P and D-V axes in the developing retina are determined independently, and also that the retinotectal projection along the two axes is controlled independently. However, these findings indicate that the retinal patterning and topographic mapping along both the A-P and D-V axes are determined in a coordinated manner.

BMP2 knockdown and *BMP2* misexpression in the chick retina at later stages alter the expression of the topographic molecules so far reported to have a gradient only along the D-V axis and that of *ephrinA2* so far only along the A-P axis. Furthermore, these molecules are expressed with an oblique gradient at the later stages, similarly to *BMP2* and *Ventroptin*. Although the axis along the fissure is defined as the D-V axis early on in development, these findings indicate that the D-V axis is tilted to the posterior side at around E5 through the switch from *BMP4* to *BMP2*, along with the corresponding change in the expression pattern of the downstream topographic molecules: the revised oblique D-V axis (D/T-V/N axis) is no longer perpendicular to the A-P axis from E6 onward, when retinal axons begin to project to the tectum in chick embryos. The polarity along the D-V axis was assumed to be determined between HH stage 8 and 14 in the chick retina based on transplant experiments. However, our findings indicate that the plasticity of the regional specificity along the first (orthogonal) and second (tilted) D-V axes is long maintained by BMP signaling and that *BMP2* is responsible for the maintenance of the second D-V axis at the later stages in the retina.

The presence of oblique-gradient molecules in the retina was not recognized until the discovery of Ventroptin. EphrinA2 has long been believed to be a topographic molecule only along the A-P axis. Presumably, this was because previous studies of retinal patterning and retinotectal projection focused on

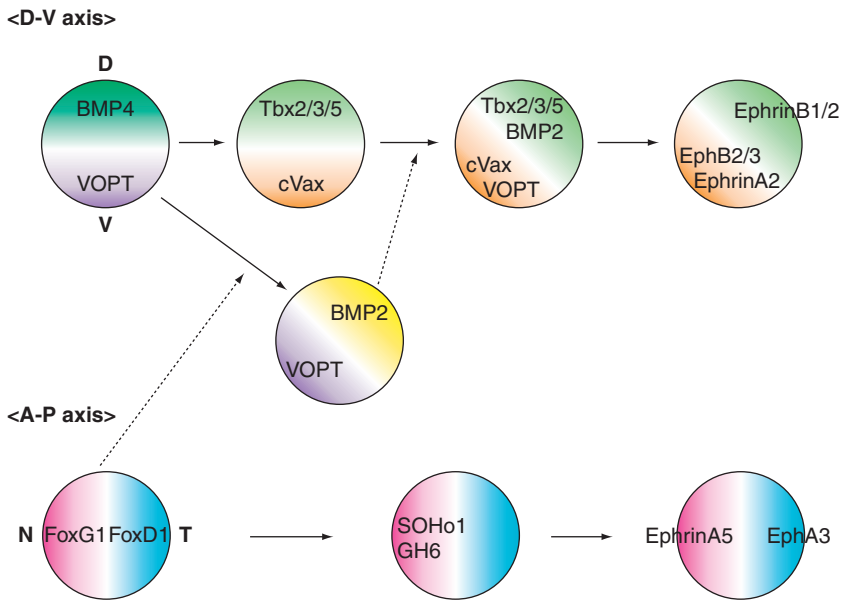

Figure 3 Developmental expression of topographic molecules in the chick retina. The counteraction between *FoxG1* and *FoxD1* determines the regional specificity along the anteroposterior (A-P) axis in the retina. *FoxG1* controls *ephrinA5* by a DNA binding-dependent mechanism, *ephrinA2* by a DNA binding-independent mechanism, and *FoxD1*, *SOHo1*, *GH6*, and *EphA3* by dual mechanisms. The counteraction between bone morphogenetic protein 4 (*BMP4*) and *Ventroptin* (VOPT) governs the regional specification in the retina along the dorsoventral (D-V) axis early in development (Hamburger–Hamilton stage 11 to E5). At around E5, BMP4 expression in the dorsal retina rapidly begins to disappear, and Ventroptin comes to be expressed in an oblique-gradient manner from E6 onward (in a ventral/nasal-high) gradient). Concomitantly, *BMP2* begins to be expressed in a pattern complementary to that of *Ventroptin* (in a dorsal/temporal-high gradient). The inhibitory action of FoxG1 on BMP signaling is attributable to the change in the expression of *Ventroptin* and *BMP2* to the oblique-gradient pattern. The counteraction between *Ventroptin* and *BMP2* determines the expression of the formerly dorsoventrally asymmetric molecules and *ephrinA2* in an oblique-gradient fashion and resultantly governs the retinal patterning and retinotectal projection along both the A-P and D-V axes. *EphrinA2* begins to be expressed much later than does *ephrinA5* from E6 in an oblique-gradient manner from the beginning under the control of the BMP2 signal. Development proceeds from left to right. Reproduced from Sakuta H, Takahashi H, Shintani T, et al. (2006) Role of BMP2 in retinal patterning and retinotectal projection. *Journal of Neuroscience* 26: 10868–10878.

only the A-P or the D-V axis individually. However, the oblique-gradient expression appears not to be specific to the chick, because *Vax2* is also reported to be expressed with an oblique-gradient (V/N-high) pattern in the mouse retina. Phenotypic defects in topographic mapping along the M-L axis are often accompanied by those along the A-P axis, and vice versa. That the revised D-V axis is not perpendicular to the A-P axis should explain why misexpression or knockdown of the topographic molecules along the revised D-V axis induces changes in the topographic mapping along the two axes in the OT/SC. In support of this view, *ephrinA2* and *Ventroptin* misexpression and *BMP2* knockdown in the developing chick retina all induce phenotypic defects in the topographic mapping along both axes. Moreover, overexpression of *BMP2* in the mouse eye apparently leads to topographic errors among ventral axons along the A-P axis in addition to the M-L axis in the SC. Thus, all the topographic molecules with the oblique-gradient expression probably affect the topographic mapping along both axes. The gene cascades of topographic molecules for the retinal patterning and retinotectal projection in the chick are shown in Figure 3.

Other Aspects of Retinal Patterning

The establishment of retinal polarity is the basis for both the topographic expression of the axon guidance molecules and the formation of the retinotectal map as described above. In addition, many retinal cells are asymmetrically organized within the retina. For instance, chick RGCs and rod photoreceptor cells show an asymmetric distribution along the D-V and A-P axes. Moreover, melanopsin-containing RGCs involved in the circadian photoentrainment show a D/T-concentrated distribution in the rat retina. The distribution of these cells in the retina is also thought to be determined by the topographic transcription factors. This view is supported by the observation that the V/N-concentrated distribution of rod photoreceptor

cells is severely disturbed by the ectopic expression of *cVax* and a dominant-negative form of *Tbx5*.

The distribution of photoreceptor cells is organized differently in the retina among the vertebrate species. Cone cells are asymmetrically distributed along both the D-V and A-P axes in the mouse retina, but only along the D-V axis in the rabbit retina. This appears to be the result of adaptations to the life style of each species during evolution and must be precisely controlled by a set of genes during development. The degree of inclination of the D-V axis may vary between species, leading to a species-specific distribution of particular retinal neuronal types.

See also: Bone Morphogenetic Protein (BMP) Signaling in the Neuroectoderm; Floor Plate Patterning of Ventral Cell Types: Ventral Patterning; Morphogens: History; Retinal Ganglion Cells: Anatomy; Retinal Ganglion Cells: Receptive Fields; Sonic Hedgehog and Neural Patterning.

Further Reading

Hattar S, Liao HW, Takao M, et al. (2002) Melanopsin-containing retinal ganglion cells: Architecture, projections, and intrinsic photosensitivity. *Science* 295: 1065–1070.

Herrera E, Marcus R, Li S, et al. (2004) Foxd1 required for proper formation of the optic chiasm. *Development* 131: 5727–5739.

Koshiba-Takeuchi K, Takeuchi JK, Matsumoto K, et al. (2000) Tbx5 and the retinotectum projection. *Science* 287: 134–137.

McLaughlin T, Hindges R, and O'Leary DD (2003) Regulation of axial patterning of the retina and its topographic mapping in the brain. *Current Opinion in Neurobiology* 13: 57–69.

Mey J and Thanos S (2000) Development of the visual system of the chick. In: Cell differentiation and histogenesis. *Brain Research. Brain Research Reviews* 32: 343–379.

Ross SA, McCaffery PJ, Drager UC, et al. (2000) Retinoids in embryonal development. *Physiological Reviews* 80: 1021–1054.

Sakuta H, Suzuki R, Takahashi H, et al. (2001) Ventroptin: A BMP-4 antagonist expressed in a double-gradient pattern in the retina. *Science* 293: 111–115.

Sakuta H, Takahashi H, Shintani T, et al. (2006) Role of BMP2 in retinal patterning and retinotectal projection. *Journal of Neuroscience* 26: 10868–10878.

Schulte D and Cepko CL (2000) Two homeobox genes define the domain of EphA3 expression in the developing chick retina. *Development* 127: 5033–5045.

Schulte D, Furukawa T, Peters MA, et al. (1999) Misexpression of the Emx-related homeobox genes cVax and mVax2 ventralizes the retina and perturbs the retinotectal map. *Neuron* 24: 541–553.

Schulte D, Peters MA, Sen J, et al. (2005) The rod photoreceptor pattern is set at the optic vesicle stage and requires spatially restricted cVax expression. *Journal of Neuroscience* 25: 2823–2831.

Shintani T, Kato A, Yuasa-Kawada J, et al. (2004) Large-scale identification and characterization of genes with asymmetric expression patterns in the developing chick retina. *Journal of Neurobiology* 59: 34–47.

Takahashi H, Shintani T, Sakuta H, et al. (2003) CBF1 controls the retinotectal topographical map along the anteroposterior axis through multiple mechanisms. *Development* 130: 5203–5215.

Yuasa J, Hirano S, Yamagata M, et al. (1996) Visual projection map specified by topographic expression of transcription factors in the retina. *Nature* 382: 632–635.

Zhang XM and Yang XJ (2001) Temporal and spatial effects of Sonic hedgehog signaling in chick eye morphogenesis. *Developmental Biology* 233: 271–290.

Neural Patterning: Midbrain–Hindbrain Boundary

S Scholpp, King's College London, London, UK
M Brand, University of Technology, Dresden, Germany

© 2009 Elsevier Ltd. All rights reserved.

Neural Induction and Posteriorization

The nervous system arises from the neural plate, which is induced during gastrulation (neural induction). Initially, the whole neural plate is of anterior character, and later it acquires posterior values leading to the formation of forebrain, midbrain, hindbrain, and spinal cord (posteriorization). Neural identity is induced by the interplay of antagonists of the bone morphogenetic proteins (BMPs) and the direct inducing activity of fibroblast growth factors (FGFs). In addition to FGFs, other signaling molecules, such as Wnts, retinoic acid (RA), and Nodals, are among the signaling molecules that have been proposed to elicit caudalizing activity. In frog and fish, high levels of BMP activity suppress anterior neural development and conversely, abrogation of BMP activity can promote neural specification. However, evidence from chick suggests that suppression of BMP activity is not sufficient to induce neural identity and that earlier signals, most likely FGFs, promote a 'prospective' state which is followed by maintenance of neural identity by BMP antagonists. Data suggest a model in which inhibition of BMP signaling promotes induction of anterior neural ectoderm, whereas in addition to its function in posteriorization, FGF signaling is necessary for specification of posterior neural ectoderm independent of BMP activity. This model links the two concepts of 'neural induction' and 'posteriorization' of neural tissue closely together by the combinatorial role of FGF signaling.

Positioning of the Brain Primordia

Identification of posteriorizing factors has been difficult because the various signaling pathways that modulate early from anterior to posterior (AP) pattern are often involved in many other events by influencing each other. Wnt molecules have been studied extensively and have been found to be good candidates for involvement in AP regionalization. Neural ectoderm is first induced with anterior character, and the most rostral cells retain this. The cells of the anterior border of the neural plate (ANB) in fish or anterior neural ridge (ANR) in chicken and mouse, located at the rostralmost tip of the neural ectoderm, have been identified as a signaling source promoting anterior forebrain gene expression. In fish, one of the secreted proteins responsible for the activity of the ANB is Tlc, a member of the secreted Frizzled related protein (sFRP) family generally considered as Wnt antagonists. Implanted Tlc-releasing cells are able to restore anterior forebrain identity in embryos lacking endogenous ANB cells. In mouse, Fgf8 has been shown to be required in the ANR for induction of Foxg1 expression and influences cell survival, and in zebra fish Fgf8 is required for patterning anterior commissural territories derived from the anterior neural plate. In general, anterior neural identity is reflected by the ubiquitous expression of *otx2*, a later marker of the forebrain and midbrain primordia. Following the induction of anterior neural tissue, Wnt signaling has been shown to directly repress the anterior expression domain of *otx2* and induces the expression of the posterior marker *gbx1* (the functional homologue of *Gbx2* in fish) without Fgf or Nodal signaling from the adjacent germ layers – respectively the endoderm and the mesendoderm. The signal is Wnt8 emanating from the posterior end of the neural plate, the margin of the embryo. Regionalized expression of the posterior neural markers *gbx1* and *gbx2* positions the boundary between midbrain and hindbrain by mutual repression of *otx2*. Following the establishment of the Otx2–Gbx interface, an MHB-specific cassette of markers is induced around the MHB: Pax2 and Wnt1 anterior to the MHB in the Otx2-positive midbrain domain, Fgf8 posterior to the MHB in the Gbx-positive hindbrain domain, and Engrailed 1 and 2 spanning the whole midbrain–hindbrain territory (Figure 1).

Similar to the interface between *Otx2* and *Gbx* at the MHB, a mechanism for the formation of the prospective ZLI has been proposed during mid-somitogenesis. The interface between the anterior Six3 expression domain and a posterior Irx3 domain should determine the future position of the ZLI. Chick explant culture experiments have suggested that the position of this interface is also controlled by Wnt signaling. Wnt3a induces expression of Irx3 and represses Six3. In addition, in *Six3-/-* mice, the prosencephalon is severely truncated, and ectopic expression of Six3 in chick and fish embryos shows that Six3 is a direct negative regulator of Wnt1 expression. Nevertheless, the role of Six3 in positioning the ZLI is still debatable due to the subsequent rapid retreat of its expression domain to the most anterior part of the forebrain prior to *shh* expression in the ZLI. In addition, there are no *in vivo* data supporting this mechanism, and genetic loss-of-function phenotypes have yet to prove this

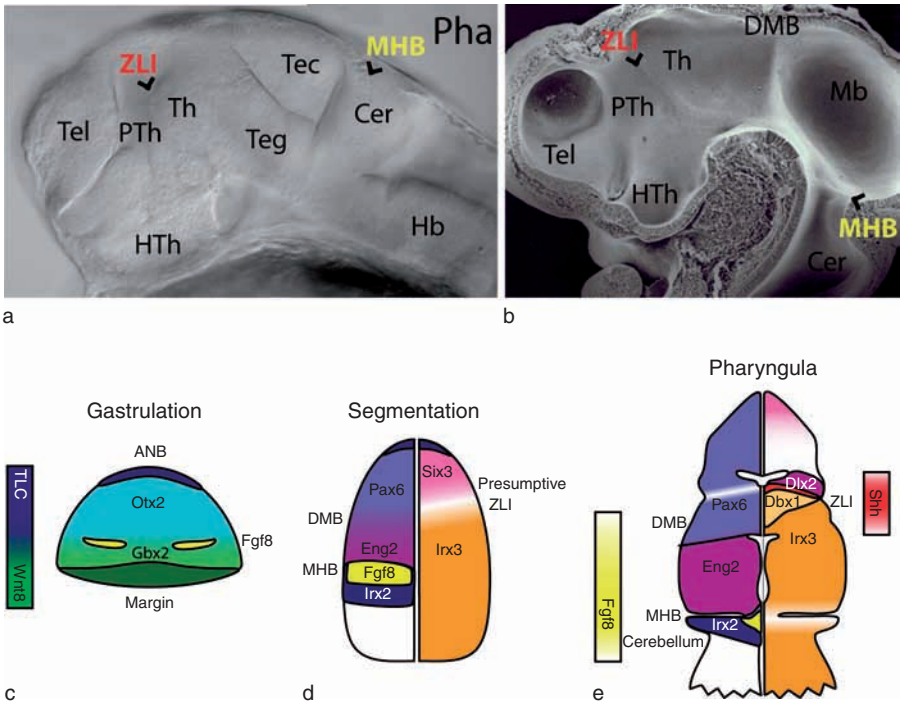

Figure 1 Positioning of organizing centers during development of the neural tube. (a and b) Comparative brain anatomy of a fish embryo (32hpf) and a chicken embryo (HH26) at the pharyngula stage. The two organizing centers, the ZLI (red) and the MHB (yellow), are located at morphological outstanding places (arrows). At gastrulation stages, mutual repression of Otx and Gbx positions the principal signal of the MHB organizer, Fgf8 (yellow wings in c). During the segmentation period (d), Anteriorly to the MHB, Fgf8 pattern the midbrain and posteriorly Fgf8 induces cerebellar fate via activation of Irx2 (shown in the left half of the embryo). At the same time, the forebrain organizer is positioned anteriorly to Irx3 in chick or Irx1 in mouse and fish. At the pharyngula stage (e), Fgf8 maintains the characteristics of the midbrain and cerebellum and sets the anterior border of the midbrain via repression of pax6. At the same time (shown in the right half of the embryo), Hh signaling from the ZLI induces the proneural genes in the forebrain (Dlx2 and Dbx1a). The range of the principal signals is shown in the columns located to the left and right of the schematic neural tube. Cer, cerebellum; Hb, hindbrain; HTh, hypothalamus; MB, midbrain; MHB midbrain–hindbrain boundary; PTh, prethalamus; Tec, tectum; Teg, Tectum; Tel, telencephalon; Th, thalamus; ZLI, zona limitans intrathalamica.

concept. Evidence from work done in mouse and fish suggests a rather different mechanism: The transcription factors Fez/Fezl in the prethalamic anlage form an interface with Otx1/2, which is expressed in the anlagen of the ZLI and thalamus. This interface coincides exactly with the anterior boundary of the presumptive ZLI. Subsequently, the expression of the thalamic Irx genes, such as Irx1b in fish, represses the ZLI territory and therefore sets the posterior boundary and gives the ZLI its characteristic tapering shape.

In contrast to the induction of Fgf8 at the MHB, sonic hedgehog (Shh) expression in the ZLI is gradually recruited from ventral to dorsal. Cell movement provides a limited contribution in the dorsal extension of *shh* expression as the ZLI matures. Our understanding is limited with regard to the mechanisms leading to this dorsoventral induction of Shh expression in the ZLI. Induction via ventral Hh signaling or via AP signaling takes center stage in these arguments. In a chick explant approach, it was suggested that Shh from the basal plate is required for the formation of the ZLI. However, induction of Hh expression is independent of Hh signaling, or indeed any ventral signaling, as shown by *in vivo* studies in fish. By taking advantage of zebra fish mutants that lack the entire basal plate of the neural tube and therefore lack ventral *Shh* expression in the midline, namely the one-eyed pinhead and cyclops mutants, establishment of the ZLI was observed as an independent process. Furthermore, all Hh signaling is dispensable for the formation of the ZLI, as shown by the slow muscle omitted mutant embryo (*smu*), which carries a mutation in the coreceptor Smoothened, lacking any Hh signaling. Although the ZLI appears narrower in *smu* mutant embryos compared with wild-type siblings, grafting of wild-type cells in an *smu* mutant background shows that Hh signaling from the ZLI still has the ability to regionalize the territory appropriately. One possible explanation for the differences between the results observed in fish and chick is that in the latter, experimentally induced reduction of ventral Hh signaling causes the ZLI to mature more

slowly. Alternatively, the positive feedback autoregulatory mechanism for *hh* expression, a plausible mechanism in chick, is less evident in fish. The persistence of *shh* expression in the ZLI in *smu* mutant embryos indicates that a positive feedback autoregulation is indeed of little importance in zebra fish.

Function of Local Organizers

As mentioned previously, posteriorization of the neural tube sets the rough position of the brain primordia. To build such a complex structure as the brain, further regionalization processes are required. Indeed, following the coarse AP organization set up during gastrulation, local organizers refine the neural plate. The concept of 'local organizers' describes specific groups of cells that can direct development of the surrounding tissue by signaling molecules. To define a cell population as an 'organizer,' the tissue has to fulfill defined characteristics:

1. The activity resides in a defined population of cells.
2. Removal of these leads to the lack of specific structures.
3. Translocation of an organizer population to an ectopic competent location such as by grafting causes the induction of an ectopic structure.
4. Organizers refine simple spatial organizations into a more elaborate pattern of subregions – a complex spatial diversification of fields. Solely ectopic provision of its organizing or so-called principal signal is able to mimic the function of an organizer population.
5. The efficiency range of organizers depends on the timing of their appearance, how their effective signaling molecules spread, and how signals are translated into specific values at the target sites.

The isthmic organizer forms at the boundary between midbrain and hindbrain and is the best understood local signaling center in the developing central nervous system. Its inductive properties were first described in chicken by grafting isthmic tissue into ectopic locations in the neural tube, resulting in the respecification of the neural tissue to ectopic midbrain and/or cerebellum. Implanted beads soaked with Fgf8 protein can mimic these effects and show that it is therefore the principal signaling molecule secreted by the MHB organizer. A requirement for Fgf8 in MHB development was demonstrated in zebra fish and mouse. The global knockout of Fgf8 in mouse fails to gastrulate and dies very early. In contrast, the zebra fish mutant *acerebellar* (*ace*) overcomes the gastrulation defect and allows the study of specific late functions of Fgf8, supported by antisense studies in fish and conditional knockout in mice. The lack of functional Fgf8 results in loss of the isthmic organizer and subsequently in a loss of the cerebellum. Fgf8 is required to maintain marker gene expression in the midbrain and isthmus but not to induce midbrain, which has already been set up during gastrulation stages. Moreover, the analysis of the midbrain in *ace* mutant fish shows that the MHB is required for AP polarization of the tectum, including the graded expression of ephrin ligands in the midbrain neuroepithelium, and for proper retinotectal map formation. In fish and mouse, Fgf8, together with Engrailed 2, is also necessary to maintain the position of the boundary between the diencephalon and the mesencephalon (DMB) by repressing the expression of *pax6*, a key regulator of forebrain development. Fgf8 secreted from the MHB organizer is also involved in patterning the anterior hindbrain. The most anterior hindbrain segment, namely rhombomere 1, which lies closest to the MHB, does not express any *Hox* genes. However, after transplantation to an ectopic location, within the hindbrain, rhombomere 1 tissue starts to express *Hox* genes. Consistent with this observation, it has been shown that both isthmic tissue and Fgf8 can inhibit expression of *Hox* genes in chick. It has also been shown that alternative splicing of Fgf8 accounts for the higher receptor-binding affinity of Fgf8b relative to Fgf8a, and for the unique ability of Fgf8b to transform midbrain into cerebellum. Taken together, these observations suggest that Fgf signaling from the MHB is required for repression of diencephalic fate, correct midbrain patterning, cerebellum induction, and anterior hindbrain patterning.

Initially, another secreted factor, Wnt1, is broadly expressed throughout the midbrain until its expression becomes restricted to the dorsal midline and a narrow band anterior to the MHB, abutting the expression domain of Fgf8 anteriorly. Wnt1 mutant mice show severe midbrain defects. In fish, Wnt1, Wnt10b, and Wnt3a are partially redundant in their capacity to regulate gene expression at the MHB, and they are required to maintain *pax2* and *fgf8*. In addition, Wnt1 is dependent on proper expression of Fgf8 and Lmx1b as well as Pax2 and Engrailed function during mid-somitogenesis and serves therefore as an example of cross-regulation of different signaling pathways in fish and chicken. However, ectopic application of Wnt1 protein does not show inductive effects comparable to those of Fgfs, suggesting a permissive role for Wnt signaling in pattern formation at the MHB and an important role in cell survival and proliferation.

The function of a further local organizer, the ZLI, has been studied in various species. Similar to the

MHB organizer, the ZLI is located at a boundary between the prethalamus and the thalamus, and members of various signaling families such as Fgfs, Wnts, and Hh are expressed at these structures. This organizing center has been characterized and it was found that ZLI exerts its function by releasing signaling molecules from the hedgehog family, sonic hedgehog-a (Shh-a), and additionally in fish Shh-b (formerly known as tiggy-winkle hedgehog). Hh signaling from the ZLI directly regulates the acquisition of the cellular identity in the diencephalon, namely the acquisition of prethalamic neuronal fate anterior to the ZLI and thalamic fate posterior to the ZLI. Lack of Hh signaling results in a loss of the expression domains and therefore in the loss of cellular identity.

As described previously, a definitive characteristic of an organizer is its ability to induce ectopic cell fate in host tissue following heterotopic transplantation. This capability has been shown for the MHB organizer, which can induce ectopic tectal and/or cerebellar structures according to the place where the organizer is grafted provided it is competent. The ZLI has yet to be confirmed as a true organizer in this way.

Boundaries and Lineage Restriction

The position of organizers often correlates with the position of constrictions in the neural plate. These constrictions are mainly manifestations of gene expression boundaries. To regulate organizer activity, a mechanism must exist that not only ensures the position of the organizer but also defines the actual number and position of cells contributing to the signaling center. One control mechanism is to restrict cell movement across boundaries, which subsequently results in the polyclonal formation of lineage restriction compartments.

A study that followed the fate of hundreds of cells in the developing MHB territory in fish showed a cell lineage restriction boundary between midbrain and anterior hindbrain that corresponds to the Otx–Gbx boundary. A study using an inducible transgenic marker in mice also indicates restricted mixing behavior between the midbrain, isthmic area, and the forming cerebellum. Therefore, the isthmus might show the characteristics of a compartment rather than a single boundary, at least on the dorsal aspects of the neural tube. Eventually, the dorsal part of the midbrain–isthmus boundary seems to allow a small number of cells to cross, although the mouse study could not exclude initiation of the lineage marker prior to the establishment of lineage restriction. Late morphological cell movement at the MHB has also been addressed in chicken. Cells from the dorsal isthmus populate the dorsal midline of the midbrain as well as the cerebellum, and application of Fgf8 protein triggers cell movement and expression of roof plate marker ectopically in, for example, the diencephalon. In summary, there is evidence for restricted cell movement at the MHB in the ventricular zone, and the isthmus may therefore form a separate compartment, but lineage restriction is not absolute and allows cell movement under certain conditions and at specific stages, such as matured neurons in the mantel zone.

Compartmentalization in the forebrain territory has been analyzed for the past 15 years mostly in chick. Various models have been proposed, correlating morphological data and gene expression domains. Analysis of boundaries in the forebrain revealed that there are only three true lineage restriction boundaries: the boundary between pallium and subpallium within the telencephalon, the boundary between the forebrain and midbrain, and the interface between the presumptive prethalamus and the thalamus – the ZLI. Interestingly, the boundary between the prethalamus and the ZLI as well as the boundary between the ZLI and the thalamus are lineage restrictive and, thus, the ZLI is a true compartment. In chick, the presumptive ZLI territory is flanked by the expression domain of lunatic-fringe (Lfng), a glycosyl transferase that modulates the activity of the Notch receptor. The wedge-shaped Lfng-free territory later narrows to a band marking the definitive ZLI territory. Whether horizontal intercalation of cells or differential growth of the adjacent territories causes this phenomenon is unclear.

First Step in Competence: Integration of the Signal by the Surrounding Tissue

The ability of a tissue to respond to signals from an organizer is called competence. With regard to signaling molecules, competence is first defined by availability of their receptors and connected downstream pathway. For the MHB organizer, Fgf receptors (FgfRs) are expressed in the presumptive territory of the midbrain and hindbrain, and ligand activation triggers the intracellular signal transduction cascade of the MAP kinase (MAPK). High levels of expression of FgfR1 can be detected directly around the boundary. Consequently, knockdown of the FgfR1 and downstream players of the MAPK pathway such as the MAPK and Erk2 causes defects in gene expression at the boundary and a loss of the MHB structure. In parallel, Fgf signaling activates a number of repressors that act in a negative feedback loop, such as the coreceptor Sef and the intracellular inhibitors Spry and MKP3. Functional studies in fish have shown that membrane-bound Sef acts as an inhibitor *in vivo*. This repressive activity is elicited by direct binding

to FgfR1 *in vitro,* resulting in prevention of autophosphorylation of the receptor, which is necessary for signal transduction. Sprouty was originally described in *Drosophila* as an inhibitor of growth factor signaling, and experiments in vertebrates confirmed this function with regard to Fgf signaling. Misexpression of Spry4 inhibits the MAPK pathway and consequently leads to the loss of the MHB territory – a phenotype comparable to the Fgf8 mutant *acerebellar.* At the level of the nucleus, competence to respond to Fgf8 signaling is specifically regulated in the neuroectoderm by the zebra fish *spiel-ohne-grenzen* gene encoding the zebra fish oct4 homologue.

For the Hh family of signaling molecules, expressed at the ZLI, the situation is more complex. The ligand-binding component of the receptor of the morphogen is Patched (Ptc), which, being induced by Hh, is a bona fide target gene of Hh signaling. After binding of Hh to Ptc, the signal-transducing component Smoothened mediates the Hh response, presumably via activation of the Gli genes. Loss of Hh signaling can be achieved by interfering with different players of the signaling pathway: by misexpression of a dominant negative form of Patched construct, lacking an extracellular binding site for Hh; analysis of the *smu* mutant embryos; or analysis of embryos treated by the Hh pathway inhibitor cyclopamine. In fish and chick, these approaches revealed a direct dependency of prethalamic as well as thalamic development on functional Hh signaling from the ZLI. Dlx2-positive neuronal precursors in the prethalamic territory as well as Dbx1a- and Gbx2-positive neuronal progenitors in the thalamus do not appear in the absence of Hh signaling, suggesting a definitive requirement for the signal.

A Further Aspect of Competence: Predetermination of Cellular Fate

The definitive fate of the competent tissue is defined less by the principal signal than by the set of transcription factors exerting this function upon activation by the signal. The homeobox transcription factor Pou2/Oct4 is expressed in the early MHB anlage and is disrupted in the zebra fish mutant *spiel-ohne-grenzen*. Besides an early role for regulation of MHB genes, Pou2/Oct4 has a permissive role in mediating hindbrain competence to respond to FGF signaling and therefore restricts influence of the organizer posterior to the MHB.

For Hox gene-mediated positional information, it has been shown that Pbx proteins are necessary and sufficient, and a further facet of its role as competence factor has been revealed: Engrailed, together with Pbx, is required for proper establishment of the midbrain territory. Similar to the Hox genes, Pbx binds to Engrailed and its absence leads to comparable phenotype with the Engrailed knockdown: down-regulation of Fgf8 and caudal shift of the posterior forebrain into the presumptive midbrain territory (Figure 2).

Interestingly, a family of molecules that are able to modulate the response to signaling molecules and are asymmetrically expressed at boundaries has been identified: the Iroquois genes, which are involved in establishing neuronal prepattern in *Drosophila*. Irx2 is expressed posterior to the MHB in the presumptive hindbrain before the onset of Fgf8 expression and has been shown to mediate the competence of this region to form the cerebellum in response to Fgf signaling. Fgf8 via MAPK signaling is required to convert Irx2 from a transcriptional repressor to an activator: Thus, an activated form of Irx2 can convert presumptive tectum into cerebellum in the absence of FGF signaling. Irx2 fused to the engrailed repressive domain has the opposite effect when electroporated into the hindbrain: The entire cerebellum is transformed into

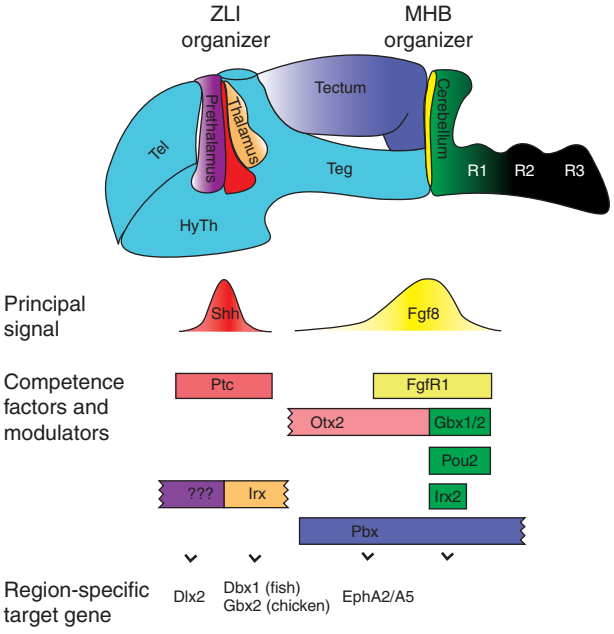

Figure 2 Function of the organizing centers. The lateral view of a schematic fish brain shows the influencing range and function of the two organizers: the ZLI organizer releases Shh (red), which binds to its receptor Ptc (light red; competence). In addition, the tissue around the organizer is already prepatterned (prethalamus, purple; thalamus, beige) and awaits the activation by Hh signaling to switch on the prethalamic program, induction of Dex2 expression, or thalamic program, induction of Dbx1a or Gbx2. Similarly, the tissue around the Fgf8-releasing center (yellow) is prepatterned and competent by the expression of the Fgf receptor 1. Anteriorly, in the tectum Fgf8 induces EphA2/A5, whereas posteriorly Irx2 and Pou2 (green) are needed to induce cerebellar fate.

midbrain tectum. This suggests that the midbrain/rhombomere 1 region has a tectal default state. The Irx2-related gene Irx3 is expressed exclusively posterior to the ZLI in the presumptive thalamus, and its ectopic misexpression anteriorly endows the prethalamus with thalamus-specific expression. This indicates that the Iroquois genes mediate positional competence on either side of signaling centers such as the MHB organizer and the ZLI organizer. Interestingly, the expression pattern of the Iroquois genes is independently induced of the principal signal and is more likely directly established by the early Wnt gradient in the neural plate. Further confirmation of this model derives from the finding that in fish, when wild-type cells are transplanted into a *smu* mutant embryo, they start to activate autonomously the proper gene expression with regard to the position of the ZLI, suggesting a correct prepattern independent of Hh signaling. Thus, principal signals from the organizers trigger the timing of the initiation of the downstream cascade rather than determination of the definitive fate of the surrounding tissue.

Functional Range of Organizers Is Determined by the Receiving Field

Apart from accurately controlling expression of the signaling molecule, a further way to control the functional range of signaling molecules has been identified. The receiving tissue actively regulates the range and the steepness of gradients from molecules released from source tissue. These mechanisms are the facilitation of signal spreading by modulation of extracellular matrix components, uptake of the signal molecule and direct degradation by 'restrictive clearance,' or uptake and re-release of signaling molecules by 'planar transcytosis.'

For signaling molecules in general, it has been shown that components of the extracellular matrix, such as the heparan sulfate proteoglycans (HSPGs), influence signaling range. For Fgf signaling, HSPGs function as coreceptors in the receptor complex and facilitate spreading of Fgf molecules through the extracellular space (ECS) in *Drosophila*. Consistently, enzymes involved in the biogenesis of heparan sulfate chains were found to be important for FGF signaling in flies and in mice. In mouse, a conditional knockout of Ext1, a glycosyltransferase involved in chain elongation of HSPGs, causes defects in the inferior colliculi and the cerebellum, a phenotype reminiscent of a loss of Fgf-dependent MHB function. A mutation in the heparan sulfate copolymerase *tout velu* and a mutation in the cell surface HSPG *dally* cause a decrease of movement of Hh molecules from their site of production. Furthermore, digestion of HSPGs by heparinase treatment strongly reduces the level of Shh signaling range in mice.

Advances have been made in understanding how signaling factors spread dynamically in embryos on a subcellular level. Identification of cellular components involved in signaling will yield a full understanding of the inductive mechanisms involved. The direct observation of tagged versions of signaling molecules has been correlated with the indirect evidence of the induction of gene expression. For Fgf8, an increase in internalization through activating Rab5-dependent endocytosis leads to a shortened functional range of the signaling factor. Conversely, reduced Fgf8 uptake by downregulation of Rab5 activity causes a wider signaling range of Fgf8 and subsequently a broader induction of target genes, such as Spry4 or the Ets transcription factors Erm and Pea3, a mechanism referred to as 'restrictive clearance' (Figure 3).

In the *Drosophila* wing, similar observations were made for wingless, for which different internalization rates operate posterior and anterior to the signaling center, which leads to different signaling ranges and creates an asymmetric response. By analogy, endocytosis might regulate the difference between

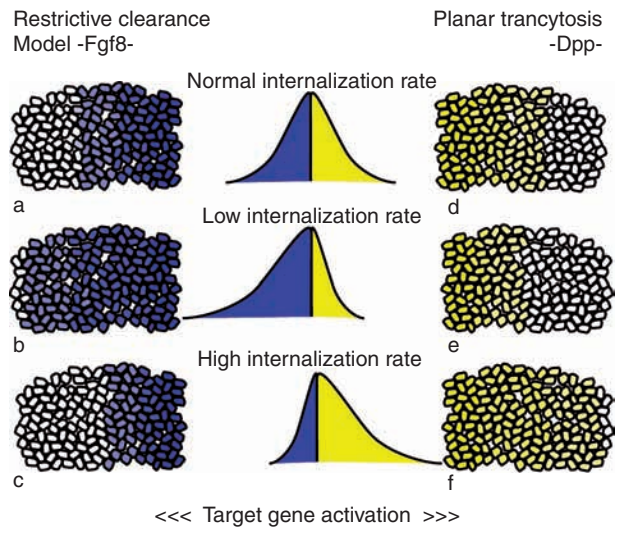

Figure 3 Influence of the competent tissue on the range of a signaling center. During gastrulation, in zebra fish, the range of Fgf8 is directly influenced by the capacity of internalization: Low internalization leads to a broader activation of the target gene (blue), whereas enhanced uptake leads to a restriction of target gene activation (c). Internalization is linked to degradation and restrictive clearance of the principal signal. In *Drosophila*, in wing imaginal discs, the signaling factor decapentapledgic (Dpp) behaves exactly the opposite: Low uptake restricts very strongly the induction of the target gene shown in yellow (d and e), whereas a higher level of internalization leads to a broad activation of target genes (f). Internalization is therefore linked to the propagation of the signaling factor and called planar transcytosis.

long-range activity of Fgf8 anterior to the MHB and short-range function in cerebellar induction and Hox gene repression. A further concept for signal spreading is provided by studies of the signaling molecule Dpp in the *Drosophila* wing imaginal disk. A higher rate of internalization led to a further distribution due to the fact that Dpp is only partially internalized for degradation, but a significant amount of the factor is released again. Thus, a planar transcytosis model allows the possibility that tight control of signaling range by the balance between degradation and release is determined by the receiving tissue as well as by the level of signal. It will be interesting to determine whether planar transcytosis exists in vertebrate embryos and whether this mechanism is employed in organizer signaling.

In this article, we described the formation and the function of two organizing centers in the early neural plate, the MHB organizer and the ZLI organizer. Apparently independent of organizer activity, the neural tube shows an initial expression pattern of modulating and determining factors, anticipating the actual fate of the regions. The organizers define the exact time points when these fates are going to be realized. Subsequently, the surrounding tissue reacts to the signal and initiates the downstream program. Furthermore, the surrounding tissue can influence the signaling range directly. The elaborate interplay between modulating factors and organizer activity sets the stage for the next level of complexity during brain development. It is obvious that this is only a next step and that many more have to follow to build as complex a structure as the human brain. Nevertheless, the prime importance of the exact timing of organizer activity is manifest, and alteration in these early events leads to gross malformation of the brain.

See also: Forebrain Development: Prosomere Model; Forebrain Development: Holoprosencephaly (HPE); Forebrain: Early Development; Midbrain Patterning; Sonic Hedgehog and Neural Patterning.

Further Reading

Entchev EV and Gonzalez-Gaitan M (2002) Morphogen gradient formation and vesicular trafficking. *Traffic* 3: 98–109.

Kiecker C and Lumsden A (2005) Compartments and their boundaries in vertebrate brain development. *Nature Reviews Neuroscience* 6: 553–564.

Larsen CW, Zeltser LM, and Lumsden A (2001) Boundary formation and compartition in the avian diencephalon. *Journal of Neuroscience* 21: 4699–4711.

Lumsden A and Krumlauf R (1996) Patterning the vertebrate neuraxis. *Science* 274: 1109–1115.

Puelles L and Rubenstein JL (2003) Forebrain gene expression domains and the evolving prosomeric model. *Trends in Neuroscience* 26: 469–476.

Raible F and Brand M (2004) Divide et Impera – The midbrain–hindbrain boundary and its organizer. *Trends in Neuroscience* 27: 727–734.

Rhinn M, Picker A, and Brand M (2006) Global and local mechanisms of forebrain and midbrain patterning. *Current Opinion in Neurobiology* 16: 5–12.

Scholpp S and Brand M (2004) Endocytosis controls spreading and effective signaling range of Fgf8 protein. *Current Biology* 14: 1834–1841.

Scholpp S, Wolf O, Brand M, and Lumsden A (2006) Hedgehog signalling from the zona limitans intrathalamica orchestrates patterning of the zebrafish diencephalon. *Development* 133: 855–864.

Wilson SW and Houart C (2004) Early steps in the development of the forebrain. *Developmental Cell* 6: 167–181.

Wurst W and Bally-Cuif L (2001) Neural plate patterning: Upstream and downstream of the isthmic organizer. *Nature Reviews Neuroscience* 2: 99–108.

Zeltser LM, Larsen CW, and Lumsden A (2001) A new developmental compartment in the forebrain regulated by Lunatic fringe. *Nature Neuroscience* 4: 683–684.

Neural Prostheses for Reaching

H Scherberger, University and ETH Zürich, Zürich, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During the past 40 years, much progress has been made on prosthetics for the functional stimulation of muscles and of sensory pathways. In particular, cochlear prostheses for the hearing impaired have been tremendously successful, and promise has been shown for the restoration of vision with the stimulation of cortex, lateral geniculate nucleus, and the retina. These stimulation devices can help visually or auditory impaired people by transmitting meaningful sensory information into the brain. Paralyzed patients, in contrast, require a neural prosthesis of the opposite kind. They need a brain–computer interface (BCI) that can read motor plans originating from the brain and send them to external effectors such as computer screens or robotic arms. Ideally, of course, such BCIs can read the patient's brain noninvasively, that is, without the need to implant electrodes inside the skull.

This has been pursued using electroencephalographic (EEG) signals. For example, the group of Birbaumer used EEG slow cortical potentials to control an electronic spelling device that allows quadriplegic patients to write messages on a computer screen, and Wolpaw and colleagues achieved two-dimensional control of a cursor on a computer screen using EEG mu (8–12 Hz) and beta rhythms (18–26 Hz). Although this technique is noninvasive, major disadvantages are that it requires extensive training of the subjects and that the amount of information transmitted through the device is limited to approximately two characters or a few movements per minute. This is probably attributable to the fact that EEG decoding is more closely related to unspecific biofeedback than to reading out the movement intentions of the brain for specific actions. For these reasons, several researchers have been trying to access motor and premotor signals more directly, namely by placing bundles of microelectrodes into specific premotor, motor, and parietal brain areas of nonhuman primates and human patients.

Progress in Neural Prostheses

Using chronically implanted electrodes in motor cortex and thalamus of rats, Chapin and colleagues reliably recorded multiple action potentials of many neurons, or multiunit activity, simultaneously over days and weeks. The rat was trained to position a one-dimensional robot arm to obtain water by pressing a lever. The real-time signals of the recorded neurons were analyzed by an artificial neuronal network that was configured to estimate the robot arm trajectory. This population readout was used to control the position of the robot arm, and in some cases the animals could position the lever even without moving their own forelimbs. This result was a first proof-of-concept that cortical multiunit recordings can be used to read out movement intentions for controlling a robotic device, and a follow-up study of the same group could link the outcome of a learning behavior in a motor task with the predictability of the animal's response from the neural ensemble activity in motor cortex.

Following this success, similar attempts to read out movement plans have been made from motor cortex of nonhuman primates. Nicolelis and colleagues predicted arm trajectories in three-dimensional space from ensembles of cells in motor cortex using a decoding algorithm that was based on a neural network approach. At the same time, Schwartz and colleagues implanted arrays of microwire electrodes permanently in monkey primary motor cortex and recorded multiple single-unit activity over weeks and months. During arm-reaching trials, Taylor and colleagues found the time-varying speed and direction of the animal's arm to be encoded in motor cortex and the population vector approach was successfully applied to decode three-dimensional arm movement trajectories from the neuronal signals. Using principal component analysis, the temporal modulation of the neuronal ensemble was utilized to improve the movement prediction accuracy and the three-dimensional endpoint position of a robot arm could be controlled online. Similar results were also reported by Serruya and colleagues from the lab of Donoghue, who employed silicone-based electrode arrays. Finally, more cognitive brain areas involved in motor control have also been investigated for their suitability to control a motor prosthesis, which is highlighted in the following section.

Decoding Cognitive Motor Signals

Cortical areas such as the parietal reach region (PRR) in the posterior parietal cortex and the dorsal premotor area (PMd) in premotor cortex play an important role in the sensorimotor transformation of visual and other sensory information for the generation of arm-reach signals. Such high-level planning areas might represent reach intentions in a particularly simple

manner, for example, if they encode the movement goal and not the details of the movement trajectory or the movement dynamics. The group of Andersen at Caltech is developing a neural prosthesis for reach movements that is based on signals from PRR. In their decoding experiments, Musallam and colleagues trained macaque monkeys first in a delayed reach task and then in a brain control task after microelectrode arrays had been implanted permanently in PRR and PMd (Figure 1(a)). In the reach task, animals first touched and looked at a central fixation spot before a visual cue was presented for a few hundred milliseconds on a touch-sensitive screen in one of eight possible locations. In the following memory period, the animal could plan the upcoming reach movement but had to withhold its execution until a 'go' signal was given that instructed the animal to reach to the peripheral target, which was rewarded with a small amount of water or juice.

During brain control trials, single- and multiunit activity was recorded simultaneously from many electrodes during the memory period and used to decode the intended reach location. As a feedback to the animal, a visual target was placed on the screen reflecting the decoded target location. Importantly, the animal was not obliged to reach with its own arm and was rewarded if the 'decoded' location matched the required location target (cue). Interestingly, animals quickly adopted a strategy in which they planned, but did not execute, their intended reach movement.

Figure 1(b) shows an example neuron during the brain control task with a preferred direction upward and to the right. Activity increased from baseline after cue onset and remained elevated throughout the memory period (interval delineated by vertical dotted lines). However, trials in which the animal knew in advance it would receive a preferred reward (orange juice) produced a significantly stronger response (black curves and tick marks) than in trials in which the animal received a less preferred reward (water; red curves and tick marks). Similar findings were made when the expected amount of reward or the reward probability was varied instead of the reward type. This suggests that individual cells in PRR represent not only the target location but also the expected value of the intended action.

The decode performance in the brain control task steadily improved over the weeks and months of the experiment. Figure 1(c) shows the overall success rate for predicting the correct target position (one out of four possible positions) in approximately 70 daily sessions in two animals (different colors). Although

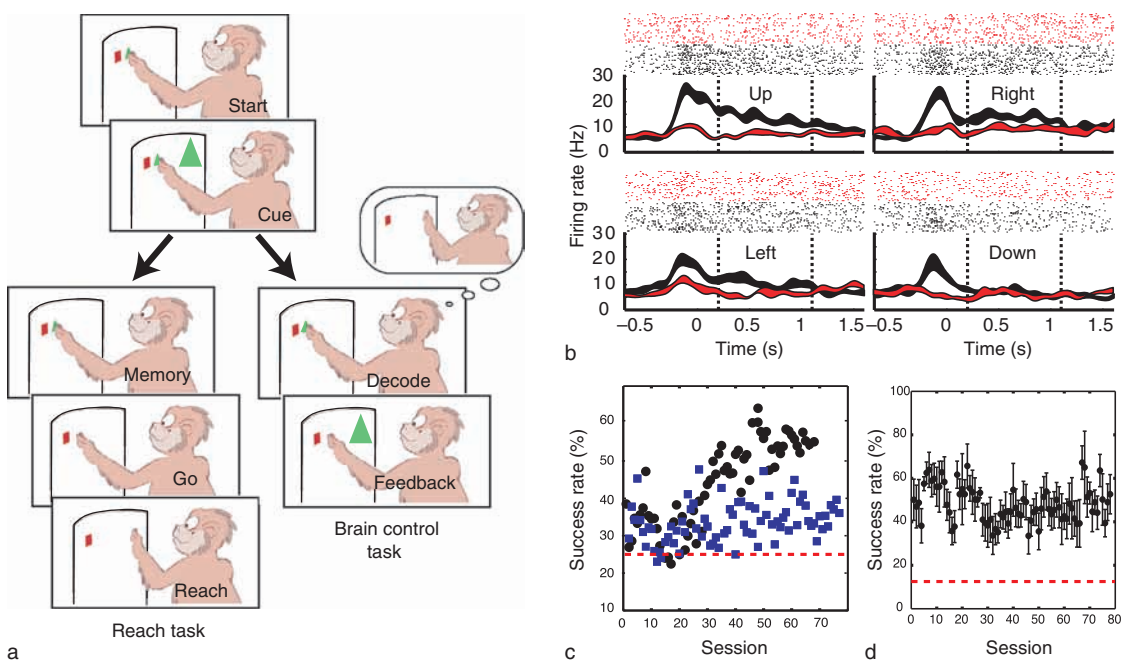

Figure 1 Cognitive control signals for neural prosthesis. (a) Delayed reach task and brain control task. (b) Example neuron during the brain control task with instructions to the up, right, left, and down direction that were rewarded either with 'orange juice' (black) or 'water' (red). Clearly, this neuron was more strongly active for trials with the preferred reward 'orange juice.' (c) Overall success rate for decoding the intended movement from four possible targets in subsequent recording sessions in two animals (black and blue dots). Dotted line, chance level (25%). (d) Success rate in off-line simultaneous decoding of the reach direction (one of four) and the expected value (preferred or nonpreferred reward) in subsequent sessions of one animal. Adapted from Musallam S, Corneil BD, Greger B, et al. (2004) Cognitive control signals for neural prosthetics. *Science* 305: 258–262, with permission from AAAS.

performance was only slightly above chance in the beginning (25%, dashed line), it steadily increased to levels of approximately 55% in the first and 35% in the second animal. Similarly, Figure 1(d) depicts the success rate of the simultaneous (off-line) prediction of the movement direction (one out of four) and the expected value (high or low) for one animal (dashed line: chance level, 12.5%). These results indicate that the target goal, as well as the expected value of the action can be decoded simultaneously from the neural activity in PRR and suggest that cognitive signals from PRR could be a viable source for controlling a motor prosthetic.

Undoubtedly, it is important to identify the brain areas most suitable for the decoding of a specific action, particularly if the goal is to restore a lost motor function without interfering with other, still intact motor modalities. Second to this goal is the speed by which information can be transmitted in a BCI, and in this case it holds: the faster the better.

Fast Decoding

As stated previously, human-operated noninvasive decoding methods (e.g., EEG-based) can only transmit a few letters per minute (in the order of 1 bit s^{-1}), whereas for invasive methods the transmission rate can be substantially higher. At Stanford University, Santhanam and colleagues from the group of Shenoy have developed a high-performance BCI using a 96-channel silicon electrode array that was implanted in the premotor arm area PMd. Macaque monkeys were first trained in the previously mentioned delayed reach movement task, from which a statistical description of the tuning properties of the recorded neurons was obtained. For speed decoding, three or more prosthetic cursor trials followed in quick succession (Figure 2(a)). In all prosthetic trials, the animal continuously touched and looked at the central fixation spot, while a visual cue appeared in the periphery to indicate the target (yellow circle). The following delay period was then subdivided into a first epoch of 150 ms that was ignored and a second epoch of variable duration (10–300 ms), during which the neural activity of the electrode array was acquired and subsequently decoded. After that, the cue disappeared and the decoded target location was presented on the screen (dashed circle), resulting in a total trial duration of approximately 200–400 ms. Immediately after, either another prosthetic trial was started or the trial sequence was completed with a delayed reach trial and the animal was rewarded (not shown).

The selection of the optimal trial duration was found to be critical for maximizing the BCI performance. Decoding from longer trials led to more

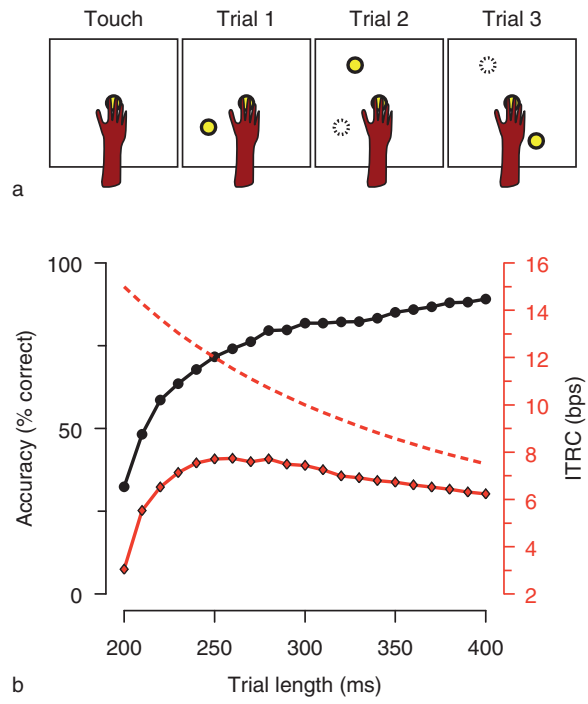

Figure 2 Single trial accuracy and information transfer rate capacity during fast prosthetic cursor trials. (a) Sequence of three prosthetic cursor trials. After the animal has placed its hand on the board, a target (solid circle) appears, the intended reach movement is decoded, and a cursor is placed on the screen at the predicted location (dashed circle in the following trial). If the predicted location was correct, the next cursor trial started immediately. After three or more cursor trials, a standard real reach trial followed and the animal was rewarded (not shown). (b) Decode accuracy (black curve) and information transfer rate capacity (ITRC) (red solid curve) for different trial lengths, as assessed from an eight-target control experiment. Whereas decoding accuracy steadily increased with trial length, ITRC was maximal for a trial length of approximately 260 ms and then declined. Red dotted line, theoretical ITRC assuming 100% accuracy independent of trial length. Reprinted by permission from Macmillan Publishers Ltd: *Nature* (Santhanam G, Ryu SI, Yu BM, et al. (2006) A high-performance brain–computer interface. *Nature* 442: 195–198), copyright (2006).

accurate predictions because of the reduced noise in the integrated signal. On the other hand, longer trial durations also reduced the number of possible trials per second, thus leading to a decreased BCI performance. Figure 2(b) illustrates this fundamental speed–accuracy trade-off using data from the standard delayed reach task. Decode accuracy (measured in percentage of correctly decoded trials) increased with trial duration but saturated for trials longer than 250 ms (black curve). As a second performance measure, the information transfer rate capacity (ITRC) gives the rate at which information is transmitted. It is defined as the information per trial (which depends on the decode accuracy and the number of possible targets) divided by the trial duration. ITRC

did not increase monotonically with trial duration (red solid curve) but showed a peak of 7.7 bit s^{-1} at 260 ms. This was expectedly lower than the theoretical maximal ITRC that assumes 100% decode accuracy regardless of the trial duration (red dotted curve).

Using sequences of prosthetic trials with variable number of possible targets (2, 4, 8, or 16), the authors found in two animals a maximal ITRC of 6.5 and 5.3 bit/s, respectively (using the 8-target task), which by far exceeded current noninvasive (e.g., EEG-based) methods. These results highlight the superior performance of systems with implanted electrodes and suggest that invasive BCIs could be very beneficial for paralyzed patients.

Clinical Applications

The group of Kennedy was the first to report the restoration of neural output in a patient using a single, chronically implanted neurotrophic electrode. The electrode was inserted in primary motor cortex of a 'locked-in' patient who was completely paralyzed. Several weeks after implantation, activity from a single neuron that had grown into the cone electrode could be recorded and the patient was trained to control the firing rate of that neuron using visual and aural feedback.

Building on this initial success and their own experience with multielectrode recordings in monkeys, Patil and colleagues from the group of Nicolelis recorded intraoperatively from neurons in the subthalamic nucleus and thalamic motor centers using a 32-microwire array in patients undergoing deep brain stimulator surgery. Neural activity was recorded while patients performed intraoperatively a hand-gripping task with visual feedback, and subsequent off-line analysis revealed the viability of such neural signals to drive a human BCI. These findings were encouraging; however, there was concern that the recorded signals were largely somatosensory and would not correlate well with the movement intentions of paralyzed patients who are unable to move their arms.

These shortcomings, as well as the inherent risk of subcortical hemorrhages were avoided in the most advanced demonstration of a functional neural prosthesis using signals from human primary motor cortex. Hochberg and colleagues from the group of Donoghue recorded ensembles of motor neurons in a tetraplegic patient who was chronically implanted with a silicone-based electrode array (**Figure 3(a)**) 3 years after suffering a massive spinal cord injury. Intended hand motion of the patient modulated the recorded motor cortical activity, and using suitable decoding algorithms, the patient was able to control a cursor on a computer screen, open and close a prosthetic hand, and perform rudimentary actions with a robotic arm. Notably, these actions were possible for the patient with relatively little effort and

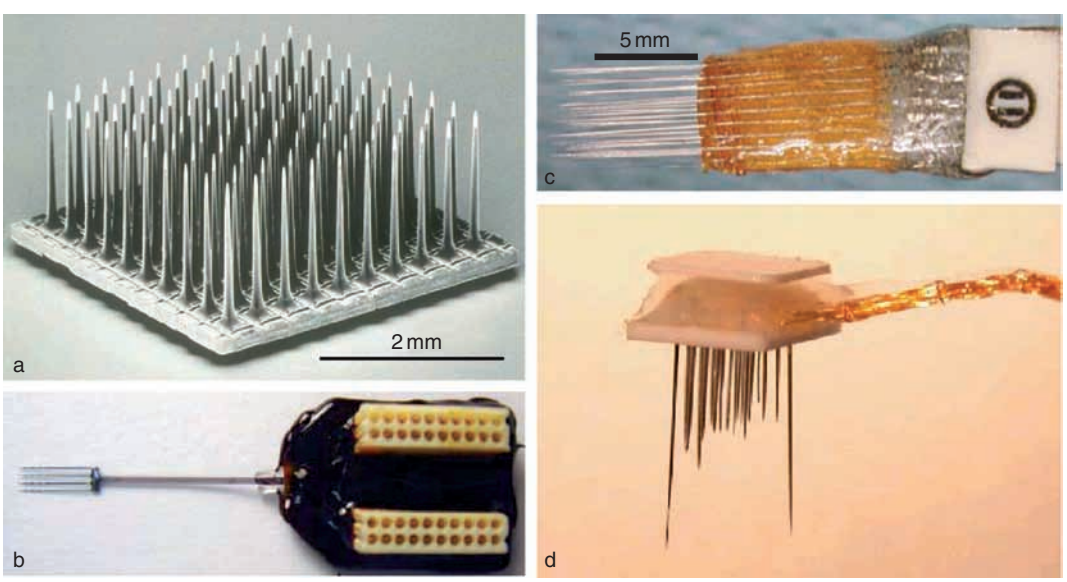

Figure 3 Types of recording electrodes. (a, b) Micromachined electrodes. (a) Silicone-based 10 × 10 electrode array (electrode length, 1.5 mm). (b) Silicon microelectrode array ('Michigan probe') with multiple recording sites along each of the four shanks. (c, d) Metal electrodes. (c) Multielectrode recording array with four rows of 8 electrodes rigidly attached to a connector block. (d) Floating metal electrode array for chronic implantation with 16 recording electrodes (1.25–1.75 mm long) and 2 stabilizing electrodes (2.5 mm long) used for the ground and reference. (a) Adapted with permission from Cyberkinetics Neurotechnology Systems (Salt Lake City, UT) (originally developed at the University of Utah). (b) Adapted with permission from NeuroNexus Technologies (Ann Arbor, MI). (d) Adapted with permission from Micro Probe Incorporated.

even while conversing. This ease of operation is most likely related to the inherent cortical plasticity that constantly keeps numerous sensory and motor systems in register and that might be involved in the fine-tuning of the recorded signals from visual feedback. Although more such patient studies are needed before final conclusions can be drawn, the existing evidence suggests a clear advantage for using highly localized brain activity for decoding, instead of the spatially more blurred signals to which noninvasive methods are currently confined.

The main difficulty of using BCIs with chronically implanted electrodes is the progressively deteriorating signal quality of the recorded spiking activity, which restricts the lifetime of the implanted electrodes to often less than 1 year. A better understanding of the physiological and biochemical nature of this signal degradation, as well as effective countermeasures, is therefore necessary before human patients can benefit in the long term from implantable BCIs that rely on spiking activity.

Electrode Design and Biocompatibility

Recording electrodes for chronic implantation in the brain have been manufactured from various materials and a number of protocols exist for their insertion into cortical and subcortical areas. Electrode arrays have been made from metal wires out of tungsten, stainless steel, or platinum/iridium with electrode shafts electrically isolated using varnish, parylene, or other coatings and with tips that are either left blunt or sharpened. Individual electrodes are arranged to bundles or arrays in a multitude of lengths and spacings. Figure 3(c) shows an example of such a multielectrode array, in which electrodes are rigidly attached to a connector. This ensemble is fixed to the skull after being slowly inserted into the brain using speeds of approximately 0.1 mm min^{-1}. Such designs have been extensively used and produced a wealth of neural data. However, since electrodes fixed to the skull cannot follow the intracranial movements of the brain due to respiration, circulation, and shock, these designs, although successful, have always been considered imperfect.

In a different approach, electrodes have been micromachined from a solid block of silicon (Figure 3(a)). This electrode array, originally developed by the group of Norman in Utah, consists of a 10×10 array of sharp electrodes that is cut and etched with electrode lengths of 1–1.5 mm and is implanted into the cortical surface with a pneumatic inserter at speeds of approximately 8 m s^{-1}. It is electrically connected with a flexible wire bundle to a connector that is fixed on the skull. This floating design allows the electrode array to follow the movements of the brain. It has been successfully used for targeting brain areas on the cortical surface, particularly motor and premotor cortex in monkeys and humans. However, implantation in cortical sulci and subcortical structures is essentially not possible due to the design limitation of approximately 1.5 mm for the electrode length.

To overcome this problem, silicone-based probes with long shanks have been developed that allow recording of neural signals from multiple sites along a single shank (Figure 3(b)). Such probes have a long history of success in acute experiments, but few data are available on the feasibility of this design for chronic recordings in monkeys or humans. Finally, a floating electrode array has been developed (Figure 3(d)) that combines the recording properties of metal electrodes with the advantages of a floating device. The suitability of this design has been demonstrated in a feasibility study in rabbits, but long-term results in monkeys or humans are not available yet.

Unfortunately, the signal quality of all these recording devices severely degrades within a few months after implantation, and the recorded spiking activity becomes unsuitable for operating a BCI. Apparently, the brain seems to develop an acute and, later, a chronic tissue reaction against all sorts of implanted electrodes. Although the biochemical reactions are not completely understood, histological examinations have shown that both reactive astrocytes and activated microglia are involved in this tissue response.

The acute tissue reaction seems to be triggered by the mechanical trauma during electrode insertion and is characterized by the presence of activated microglia in the first weeks after insertion. Functionally, the acute reaction alone might not be a significant problem. However, as soon as it diminishes, a chronic foreign body reaction starts and forms a sheath of activated microglia that insulates the electrodes from the nearby neurons. This reaction leads to a local neurotoxic environment that kills or significantly reduces the number of neurons immediately adjacent to the electrodes. As a result of these effects, the neural signal quality progressively degrades in the weeks and months after insertion until all spiking activity is eventually lost.

Although there may be some differences in the acute and chronic tissue reactions for different devices and implantation procedures, no systematic study exists for primates and the number of reported cases is relatively low. It has been argued that floating devices should be less traumatic in the long term and therefore should produce fewer tissue reactions than electrodes that are rigidly attached to the skull. However, tissue reactions seem to be quite similar

for the different electrode materials and insertion procedures. It remains to be seen whether advancements in the biocompatibility of the implanted materials or the use of drugs, such as neurotrophic factors or anti-inflammatory agents, can improve the long-term stability of the recorded neural signals in the future, such that these devices can be used in paralyzed patients over years and possibly decades. In a different approach, several groups have started to investigate other brain signals that are more robust, such as the local field potential, for the decoding of movement intentions.

Decoding from Local Field Potential Signals

The local field potential (LFP) is a summation signal of excitatory and inhibitory dendritic potentials from a large number of neurons in the neighborhood of the recording site. Spiking activity, in contrast, originates from a single neuron located close to the electrode tip. Because of its larger 'listening sphere,' LFP activity is much less affected by the tissue reactions following electrode implantation. Therefore, the question is interesting whether LFP signals could in principle be used to control a neural prosthetic device.

LFP activity has been investigated in cortical and subcortical areas. Several studies that related LFP activity in motor, parietal, and premotor cortex with the behavior of the animal showed that LFP signals are modulated during the task, as well as by the different behavioral conditions. For example, Scherberger, Jarvis, and Andersen investigated the LFP spectral power in the PRR during the planning and execution of a center-out reach or saccade task. The panels in Figure 4(a) show the LFP power spectrogram (in decibels relative to the baseline period) of a population of recording sites for reaches and saccades and for movements to the preferred and nonpreferred direction. Evidently, the low-frequency components (0–15 Hz) encoded the movement type (reach vs. saccade) during planning and execution independently from the movement direction. At higher frequencies (15–50 Hz), however, the LFP was also modulated by the movement direction, particularly for reaching.

Furthermore, decoding simulations based on these sequentially recorded data showed that LFP activity from PRR could predict the behavioral state of the animal, as well as the movement direction. Figure 4(b) illustrates the error rates for predicting five behavioral states (baseline fixation, reach and saccade planning, and reach and saccade execution) with a Bayesian classifier using various numbers of recording sites. For LFP activity, prediction errors were less than 10% when 20 or more sites were used for decoding. In contrast, errors were much larger when the predictions were made with spiking activity that was recorded simultaneously at the same sites.

The capabilities of the LFP to predict the movement direction are illustrated in Figure 4(c), which shows a confusion matrix for a decoding simulation of the reach and saccade directions using the LFP activity of 89 tuned sites during the planning period. Correct predictions lined up on the unity line, whereas directional errors occurred predominantly in the saccade task and to the neighboring directions. Notably, reaches and saccades were never confused, and the overall decoding accuracy was 81% for reach and 40% for saccade directions (chance, 6.25%). In comparison, 97% of the reach and 72% of the saccade predictions were correct when the decoding was made from the simultaneously recorded spiking activity at the same sites (not shown). Taken together, LFP activity in PRR can predict the behavioral state and the planned reach direction, and the somewhat diminished accuracy for using LFP instead of spiking activity might be compensated with an increased number of LFP recording sites that are relatively easy to acquire. Similar results have also been obtained in other studies that used LFP activity from motor cortex to predict reach movements and from the lateral parietal area to predict saccades.

It therefore appears that the LFP is a suitable and robust signal for decoding movement intentions and might help overcome some of the major drawbacks of current recording technology with respect to the long-term quality and stability of chronic recording.

More research is needed to determine whether spiking activity recorded from electrodes with improved biocompatibility characteristics or LFP or ECoG signals, or a combination of these, are best suited to overcome the current limitations of long-term recordings in invasive BCIs.

Finally, other recording techniques, such as the electrocorticogram (ECoG), might also play an important role in future neuroprosthetic devices. The ECoG is recorded with electrodes that are placed on the surface of the brain or between the skull and dura. It is a cortical summation signal that is more broadly localized than individual LFPs but still much finer than the EEG. Since ECoG electrodes do not penetrate the brain or dura, they are clinically safer than intracortical recordings, whereas the signal characteristics are still more closely related to the LFP than to the extracranial EEG. Tuning properties of the ECoG are not well explored. However, studies have demonstrated that the high-gamma frequency components of the ECoG can easily be modified through biofeedback and used for the online control of a curser. ECoG signals might therefore be advantageous for future neuroprosthetic devices with improved robustness and long-term signal stability.

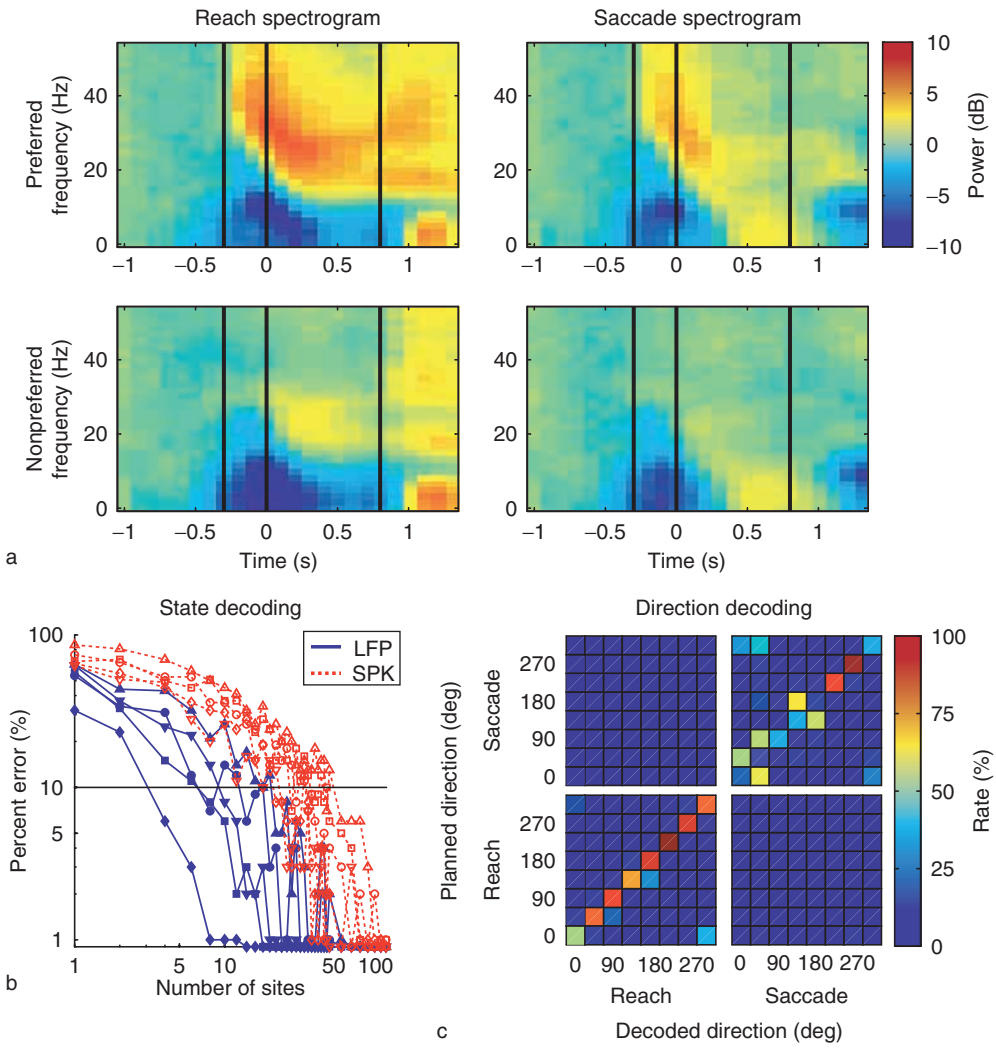

Figure 4 Decoding of reach movements from LFP activity in the parietal reach region. (a) Population spectrogram of LFP activity during the reach and saccade task is presented for movements to the preferred and nonpreferred directions. Color code indicates LFP power relative to baseline activity. Vertical lines mark the start of the cue, planning, and movement period. (b) Decoding simulation of five behavioral states using LFP (blue) or spiking activity (red) as a function of recording sites. Behavioral states are indicated by markers: baseline (circles), reach planning (squares) and execution (diamonds), and saccade planning (upward-pointing triangles) and execution (downward-pointing triangles). (c) Decoding simulation of reach and saccade directions from LFP activity during movement planning. Color codes the distribution of the decoded directions for each direction planned by the animal (chance level, 6.25%). Reproduced from Scherberger H, Jarvis MR, and Andersen RA (2005) Cortical local field potential encodes movement intentions. *Neuron* 46(2): 347–354, with permission from Elsevier.

Conclusions

Much progress has been made to bring functional neural prostheses for reaching into clinical practice, but more research is needed to better understand the fundamental coding properties of brain areas that contribute to the planning and execution of arm movements. The question is still open regarding which brain areas and what kind of decoding algorithms are best suited to operate a neural prosthesis and how well the different neural signals, spiking activity, LFP, and EcoG can be employed to operate a BCI. Furthermore, the longevity of these signals during the long-term recordings needs to be further explored. However, the most important issue to make neural prosthetics a reality for paralyzed patients is to improve the biocompatibility of the implanted electrodes in order to ensure the long-term stability of the recorded signals following the months, years, and, it is hoped, decades after electrode implantation.

See also: Computational Approaches to Motor Control; Finger Movements: Control; Games in Monkeys: Neurophysiology and Motor Decision-Making; Parietal Cortex and Spatial Attention; Posterior Parietal Cortex

and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape; Sensorimotor Control of Manipulation; Spatial Transformations for Eye–Hand Coordination; Vision for Action and Perception.

Further Reading

Birbaumer N, Ghanayim N, Hinterberger T, et al. (1999) A spelling device for the paralysed. *Nature* 398: 297–298.

Carmena JM, Lebedev MA, Crist RE, et al. (2003) Learning to control a brain–machine interface for reaching and grasping by primates. *PLoS Biology* 1: E42.

Chapin JK, Moxon KA, Markowitz RS, and Nicolelis ML (1999) Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience* 2: 664–670.

Hochberg LR, Serruya MD, Friehs GM, et al. (2006) Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* 442: 164–171.

Kennedy PR and Bakay RA (1998) Restoration of neural output from a paralyzed patient by a direct brain connection. *Neuro Report* 9: 1707–1711.

Leuthardt EC, Schalk G, Wolpaw JR, Ojemann JG, and Moran DW (2004) A brain–computer interface using electrocorticographic signals in humans. *Journal of Neural Engineering* 1: 63–71.

Mehring C, Rickert J, Vaadia E, Cardosa de Oliveira S, Aertsen A, and Rotter S (2003) Inference of hand movements from local field potentials in monkey motor cortex. *Nature Neuroscience* 6: 1253–1254.

Musallam S, Corneil BD, Greger B, Scherberger H, and Andersen RA (2004) Cognitive control signals for neural prosthetics. *Science* 305: 258–262.

Nicolelis MAL (1999) *Methods for Neural Ensemble Recordings*. Boca Raton, FL: CRC Press.

Patil PG, Carmena JM, Nicolelis MA, and Turner DA (2004) Ensemble recordings of human subcortical neurons as a source of motor control signals for a brain–machine interface. *Neurosurgery* 55: 27–38.

Polikov VS, Tresco PA, and Reichert WM (2005) Response of brain tissue to chronically implanted neural electrodes. *Journal of Neuroscience Methods* 148: 1–18.

Santhanam G, Ryu SI, Yu BM, Afshar A, and Shenoy KV (2006) A high-performance brain–computer interface. *Nature* 442: 195–198.

Scherberger H and Andersen RA (2003) Senosorimotor transformations. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 1324–1336. Cambridge, MA: MIT Press.

Scherberger H, Jarvis MR, and Andersen RA (2005) Cortical local field potential encodes movement intentions in the posterior parietal cortex. *Neuron* 46: 347–354.

Schwartz AB, Cui XT, Weber DJ, and Moran DW (2006) Brain-controlled interfaces: Movement restoration with neural prosthetics. *Neuron* 52: 205–220.

Serruya MD, Hatsopoulos NG, Paninski L, Fellows MR, and Donoghue JP (2002) Brain–machine interface: Instant neural control of a movement signal. *Nature* 416: 141–142.

Taylor DM, Tillery SI, and Schwartz AB (2002) Direct cortical control of 3D neuroprosthetic devices. *Science* 296: 1829–1832.

Wessberg J, Stambaugh CR, Kralik JD, et al. (2000) Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature* 408: 361–365.

Wolpaw JR and McFarland DJ (2004) Control of a two-dimensional movement signal by a noninvasive brain–computer interface in humans. *Proceedings of the National Academy of Sciences of the United States of America* 101: 17849–17854.

Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines

A Hartmann, INSERM UMR 679, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Inflammation is a specialized immune response of the organism to an invading pathogen, a traumatic event, or, in general terms, an injurious agent. The agent may be foreign or self, as in a necrotic cell, and inflammation may be acute or chronic. Inflammation in the classical sense involves innate and adaptive immune responses. Innate immune responses involve typically macrophages, natural killer cells, the complement system, and numerous cytokines, chemokines, acute phase proteins, and arachidonate metabolites. The adaptive immune response uses the same soluble mediators as well as T and B lymphocytes and specific antibodies. The inflammatory reaction that characterizes most neurodegenerative diseases is often called 'neuroinflammation' and consists mainly of elements of the innate immune response. Activated microglia – the resident macrophages of the brain – and astrocytes are the main neural populations that participate in this response.

Whereas an inflammatory response is necessary and crucial if the agent is a pathogen or dying cell, aberrant inflammatory responses are well-known causes of tissue damage and disease. Persistent inflammatory responses can lead to permanent scarring and tissue damage. Many wrongly equate the term 'inflammation' with these detrimental outcomes only. However, several observations have challenged this definition: Whereas the healthy central nervous system (CNS) is hostile to immune activity and is destroyed by it, immune activity may be an essential requirement for the protection and maintenance of the damaged CNS. The possibility that inflammation is a purposeful response in the CNS has long been ignored due to the fact that the beneficial effect of inflammation does not come free of charge (in terms of additional neuronal loss) and that there is a need for an optimal balance in terms of the time of onset, duration, phenotype, and intensity of inflammation.

In the healthy CNS, microglia and astroglia are quiescent, but in the event of injury to axons or cell bodies they exercise their neural function by buffering harmful self-compounds and clearing debris from the damaged site and their immune function by providing immune-related requirements for recovery. Proper regulation of the inflammatory response to injury will arrest degeneration and promote regrowth, whereas inappropriate regulation will lead to ongoing degeneration. Regulation is achieved by the operation of a T cell-mediated response directed to abundant self-antigens residing in the damaged site.

Cytokines

Cytokines are low-molecular-weight glycoproteins that may either function as membrane-bound complexes or are secreted into the surrounding tissues, sometimes packaged into vesicles. Cytokines are now known to be synthesized and secreted by many different types of cells, including nonlymphoid cells such as endothelial cells, epithelial cells, fibroblasts, and microglia. They comprise a large group of proteins and have a wide range of activities. Cytokines are generally classified into five families: the interleukins (ILs), interferons (IFNs), colony-stimulating factors, chemokines, and peptide growth factors. Like hormones, their names reflect the activities first recognized, but additional functions are frequently being discovered.

Cytokines act in a paracrine or autocrine fashion by interacting with specific cell surface receptors, and most cytokines can function as ligands for more than one receptor. This has led to the growing recognition that it is the property of the specific receptor that is activated which defines the functional properties of any one specific cytokine. Because these receptors are themselves regulated by cytokines, it becomes evident that the potential physiological and pathological effects of these mediators may be dramatically altered by the presence of other cytokines in the surrounding tissues.

Microglia

Microglia are commonly described as the CNS equivalent of tissue macrophages. However, resting microglia in the normal brain should rather be viewed as cells with macrophage-like potential because the resting state is characterized by ramified morphology and weak expression of molecules associated with macrophage function. Access of resting microglia to the macrophage-like repertoire of response options requires their transformation, which is triggered by appropriate stimulation. However, resting microglia are far from inactive. They are constantly on the lookout for changes in their environment and always stand ready to support endangered neurons or to fight off any threat to tissue integrity. Microglia can sense homeostatic disturbances in the form of alterations in

the biochemical composition or structural organization of their extracellular milieu. Even relatively minor deviations from normal metabolic neuronal activity seem to be sufficient to alert microglia. More generally, signals vary in origin and chemical nature, but they can be classified according to whether the phenotypes they trigger in microglia can be tolerated by the CNS and are thus beneficial or cannot be tolerated by the CNS and are thus destructive. Specifically, injured neurons may generate rescue signals that trigger microglia to produce trophic or other factors that promote neuronal recovery, and this microglial response is amenable to well-controlled boosting by T cells, whose specificity to CNS self-antigens directs them to the afflicted sites.

Upon activation, microglia release soluble effectors that can directly or indirectly cause damage to neural cells. One of the first effectors to be produced is IL-1β. In addition to producing IL-1β, microglia release IL-3, IL-6, tumor necrosis factor-α (TNF-α), vascular endothelial growth factor (VEGF), lymphotoxin (LT), macrophage inhibitory protein-1α, matrix metalloproteinases, nitric oxide (NO), and other reactive oxygen species. The cytokines IL-1, IL-6, TNF-α, and LT alter vascular adhesion molecule expression, which recruits lymphocytes and macrophages to sites of injury. In addition, LT, TNF-α, NO, and reactive oxygen species can directly kill cells. Although astrocytes also produce these potentially toxic factors, microglia produce much higher levels of these cell-damaging intermediates. Therefore, the initial response to CNS injury is mediated by microglia, which then recruit other immune cells to clear pathogens and infected or damaged cells. Feedback mechanisms are then initiated to prevent excessive tissue destruction, and there is good evidence that the astrocyte reaction is indeed an important compensatory response to bring about the resolution of the injury response and to restore homeostasis.

Astroglia

Astrocytes are ubiquitous glia whose finely branching processes envelope all cellular components throughout the CNS. In the uninjured CNS, astrocytes provide many supportive activities essential for neuronal function, including homeostatic maintenance of extracellular ionic environment and pH, clearance and release of extracellular debris, provision of metabolic substrates for neurons, coupling of cerebral blood flow to neuronal activity, and possibly the sculpting and maintenance of synapses. In addition, astrocytes become 'reactive' in response to all CNS insults. Also, astrocytes express numerous receptors that enable them to respond to virtually all known neuroactive compounds, including neurotransmitters, neuropeptides, growth factors, cytokines, small molecules, and toxins. These receptors enable astrocytes to not only participate in signal processing but also function as sentinels, as do microglia. Astrocytes also establish and maintain CNS boundaries, including the blood–brain barrier (BBB) and the glial limitans, through interactions with endothelial and leptomeningeal cells.

The characteristics that are most frequently associated with the astrocytic response to insult involve cellular hypertrophy, astrocyte proliferation, process extension and interdigitation, and increased production of the intermediate filaments glial fibrillary acidic protein (GFAP), vimentin, and nestin. This constellation of responses, called anisomorphic gliosis, is the consequence of gross tissue damaging injuries and results in the formation of a tightly compacted limiting glial margin termed the astrogliotic scar. These reactive astrocytes may also exacerbate tissue damage because they can release proinflammatory cytokines such as TNF-α, which can inhibit neurite outgrowth and kill oligodendrocytes, and they can produce and release arachidonic acid metabolites, NO, and reactive oxygen species that can adversely affect cell survival. However, as stated previously, microglia produce much higher levels of these cell-damaging intermediates than astrocytes; therefore, astrocytes may not be culpable for the bystander damage that occurs subsequent to reactive gliosis. There is another type of astrocyte response to insult that is less dramatic and frequently transient which is classified as isomorphic gliosis. This astroglial response is associated with improved recovery from tissue-damaging insults. At sites distant from traumatic injury or in regions where the tissue has been merely disturbed (e.g., following axotomy, nerve crush, or after a mild stroke), the astrocytes become larger in size and undergo nuclear hypertrophy; transform into a more pronounced stellate shape; and increase their production of many cytosolic enzymes, antioxidants, structural proteins, and organelles. Additionally, these 'activated astrocytes' produce soluble trophic and growth factors that enhance the survival of adjacent neurons and glia, as well as coordinate tissue remodeling. Thus, these isomorphic changes should be considered adaptive and beneficial to restoring homeostasis.

The Role of Adaptive Immunity in Regulating the Glial Inflammatory Response

Whether the CNS inflammatory response is beneficial or detrimental is determined by how the response is controlled. Studies indicate that inflammation is

mediated not only by the innate arm of the immune system (represented by microglia and macrophages) but also by its adaptive arm in the form of peripheral immune cells. This view recognizes that inflammation can be beneficial for traumatized CNS tissue by promoting clearance of cell debris and secretion of neurotrophic factors and cytokines. Their activation by T cells allows microglia to express dual activity: as immune cells with the function of antigen presentation and as neural cells that supplement the buffering of destructive self-compounds. Thus, an insult to the CNS signals the immune system for assistance in protecting the tissue against the local threat. Accordingly, inflammation should rather be described as a series of local immune responses that are recruited to damaged tissue, and its ultimate outcome depends on its regulation, which in turn depends on both genetic and environmental factors (timing and intensity).

Thus, in addition to resident microglia and infiltrating macrophages, peripheral lymphocytes (T and B cells) also participate in the local response to an acute or chronic CNS insult. The T cell-mediated assistance is aimed at providing the activated microglia with optimal conditions, leading to a beneficial outcome while minimizing the risk. Thus, microglia that encounter adaptive immunity can acquire a phenotype capable of presenting antigens and engaging in dialog with T cells. Depending on the nature and amount of T cell-derived cytokine that they encounter, such microglia can become activated without producing the potentially cytotoxic cytokine TNF-α themselves, and they can even downregulate its production in other cells. For example, in response to relatively small amounts of IFN-γ, T helper (Th)-1 cells, which are classically viewed as proinflammatory, can cause microglia to buffer glutamate (commonly involved in neurodegenerative diseases) and to protect neural tissue, provided that the amount of IFN-γ is tightly controlled. Likewise, Th2 cells, which are commonly viewed as anti-inflammatory, via IL-4 stimulate microglia to produce insulin-like growth factor (IGF)-1, which is associated with cell renewal. Microglia activated by IFN-γ or IL-4 protect neurons and can support both neurogenesis and oligodendrogenesis.

Excitoxicity and Adaptive Immunity

Glutamate, a key neurotransmitter, is pivotal to CNS function. Alterations in its concentration can be dangerous, as seen in acute injuries of the CNS and also in chronic neurodegenerative disorders. Its homeostasis is attributed to the efficient removal of glutamate from the extracellular milieu by reuptake via local transport mechanisms. Thus, the levels of glutamate transporters and glutamine synthetase increase when astrocytes become activated. Astrocytes possess two glutamate transporters that sequester excess glutamate, GLT-1 and GLAST, which may thus have neuroprotective functions. Moreover, studies suggest that glutamate, either directly or indirectly, elicits a purposeful systemic T cell-mediated immune response directed against immunodominant self-antigens that reside at the site of glutamate-induced damage. This harnessed autoimmunity may help the resident microglia in their dual function as antigen-presenting cells (serving the immune system) and as cells that clear the damaged site of potentially harmful material (serving the nervous system). The interplay between glutamate and an adaptive immune response illustrates the bidirectional dialog between the immune and nervous systems.

Regulation of Cytokine Release: Intensity and Timing

Because the ability of the CNS to tolerate any deviation from homeostasis is poor, even defensive activity on the part of activated microglia can exacerbate a chaotic situation rather than resolve it. This might explain why, for example, in mouse organotypic hippocampal slice cultures exposed to toxic concentrations of the glutamate-like compound AMPA, small doses of TNF-α (which is considered to be proinflammatory) can render microglia protective of neural cells, whereas large doses render them cytotoxic. Similarly, if microglia are exposed to small amounts of IFN-γ, they buffer any toxic excess of glutamate and support neuronal survival and neurogenesis from adult neural progenitor (stem) cells (NPCs). However, if they are exposed to IFN-γ in large amounts, they become cytotoxic to both neurons and oligodendrocytes generated from NPCs. Interestingly, these adverse effects can be counteracted not by total suppression of IFN-γ but by neutralization of TNF-α, a cytokine that is upregulated when microglia encounter large concentrations of IFN-γ. These results thus suggest that IFN-γ in small amounts endows microglia with a phenotype that favors neural tissue survival. This phenotype is not altered by an increase in the concentration of IFN-γ to which the microglia are exposed, but it is accompanied by undesirable activities such as TNF-α production.

Regarding the timing of the inflammatory response, examination of the processes of neuronal regrowth and neuronal protection suggests that within the same process the requirements for recovery might vary at different postinjury stages. For example, the anti-inflammatory cytokine IL-10, which has a strong positive effect on recovery when applied soon after spinal

cord injury, is not beneficial when first applied at later posttraumatic stages. Similarly, anti-inflammatory treatment was found to be neuroprotective only if applied within the first few hours after ischemia in experimental stroke models.

Astroglial Scar Formation

Reactive astrocytes form scar tissue after CNS insults, but the purpose of scar formation has been unclear. Ablation of scar-forming reactive astrocytes in transgenic mice shows that these cells are essential for the spatial and temporal regulation of inflammation after CNS injury. In the presence of reactive astrocytes, scar tissue composed of astrocytes and fibroblast lineage cells demarcates and surrounds damaged tissue. Within the walled-off area, a robust inflammatory reaction occurs with the release of potent cytotoxic agents targeted at potential invading microorganisms but that also sacrifices local neural cells. Outside of and immediately adjacent to the astrocyte scar, inflammation is minimal and cytoprotective mechanisms are active. Although the astrocyte scar may serve primarily as a migration barrier that keeps inflammatory leukocytes from invading adjacent healthy tissue, the redundancy of migratory guidance cues among neurons and leukocytes may account for the inhibition of axon regeneration by this barrier. Evolutionary pressure seems likely to have favored mechanisms able to wall off small injuries and keep them uninfected rather than mechanisms able to deal with large and disabling injuries. Thus, in transgenic mice in which reactive astrocytes are ablated after brain or spinal cord injuries, fibroblast lineage cells do not organize, no scar forms that demarcates damaged tissue from viable tissue, and inflammation spreads over a large area and continues unabated for prolonged times after injury.

Central Nervous System Pathologies Associated with a Beneficial Inflammatory Response

Inflammatory mechanisms beneficial for neuronal plasticity can take place at different sites: the BBB, the neuronal cell body, the axon, and the synapse. Also, evidence suggests that both microglia and astroglia can participate in neurogenic responses through the release of cytokines. As discussed in the following sections, a large amount of knowledge regarding the beneficial effects of inflammatory mediators has been gained through the use of mouse models with deletions of both major cytokines (IL-1β, IL-6, and TNF-α) and astroglia. A knockout model for microglia has been published (CD11b-HSVTK mice) but not characterized with regard to the potentially deleterious effects of microglial deletion on CNS plasticity.

Blood–Brain Barrier Integrity

Astrocytes not only maintain the BBB but also participate in reforming the BBB following CNS injury. When a stab wound is inflicted to the forebrain of adult mice selectively ablated for reactive astrocytes adjacent to the injury site, there is a failure of BBB repair, increased leukocyte infiltration, and increased neuronal degeneration. Also, ablation of reactive astrocytes after experimental brain injury leads to massive local tissue edema compared with control mice, suggesting that reactive astrocytes help to regulate tissue fluid levels and that toxic edema may occur when astrocytes have exhausted their capacity to do so. Interestingly, subsequent grafting of nontransgenic astrocytes restores BBB function.

Further evidence implicating cytokine-stimulated astrocytes in BBB restoration comes from experiments on IL-1β null mice: Animals lacking IL-1β show less astrocyte reactivity 2 or 3 days following cortical lesion, and increased permeability of the BBB at 7 days postlesion, compared with wild-type controls. These results link the early activation of astrocytes and their subsequent role in BBB repair following traumatic injury to the inflammatory response and, specifically, to the production of IL-1β.

In addition, after an injury with disruption of the BBB, angiogenesis and neovascularization must occur to provide nutrients and oxygen. Angiogenesis can be induced by numerous factors, including VEGF, which is a potent mitogen for endothelial cells and is rapidly produced in the brain in response to both hypoxia and cytokines. Several lines of evidence implicate activated astrocytes and subsequent cytokine release in VEGF-mediated angiogenesis following CNS injury.

Neuronal Survival

Astrocytes produce a large array of neurotrophic factors, including nerve growth factor (NGF), brain-derived neurotrophic factor, activity-dependent neurotrophic factor (ADNF), hepatocyte growth factor, ciliary neurotrophic factor (CNTF), and fibroblast growth factor-2 (FGF-2), in response to injury, disease, and activation of cytokine receptors.

Experiments demonstrating that IL-1β potently increases astrocytic expression of NGF provided the first evidence that glial activation may promote neuronal survival and perhaps regeneration. When cultured spinal cord astrocytes are activated with CNTF *in vitro*, they support the survival of a significantly greater number of ventral spinal motor neurons compared to unstimulated astrocytes. CNTF likely achieves these effects by increasing astroglial

neurotrophic factor production, particularly by inducing the expression of FGF-2.

IL-6 is a cytokine that has often been termed as anti-inflammatory. It shares the signaling receptor gp130 with several growth factors, including CNTF, leukemia inhibitory factor, oncostatin M, and IL-1β in the CNS. IL-6 has been shown to promote neuronal survival and neurite growth. Unlike IL-1β and TNF-α, which are responsible for the induction of multiple proinflammatory genes, IL-6 often fails to induce these genes and, furthermore, can downregulate the expression of TNF-α. Moreover, IL-6 is able to induce the nerve growth factor and counteract *N*-methyl-D-aspartate-mediated toxicity. Together, these features are suggestive of IL-6's role in repair and regulation of inflammation in the CNS. However, IL-6 is not completely harmless because the proinflammatory potential of IL-6 has been clearly shown in transgenic mice that display CNS inflammation and neurodegeneration. Therefore, IL-6 may affect inflammation and neuronal regeneration through a number of mechanisms, whose involvement would be expected to vary as a function of physiological state.

In experimental traumatic brain injury studies on TNF knockout mice, it was demonstrated that there may be an early neurotoxic function (1 or 2 days) and a neuroprotective TNF function later in the posttraumatic phase (2–4 weeks). Furthermore, mice lacking either or both TNF and IL-6 showed a higher mortality rate than wild-type mice after closed head injury. However, no significant changes in the neurologic outcome among the survivors in the IL-6 knockout group were observed. Also, a slower rate of recovery and a higher permeability of cerebral vessels were found in IL-6 knockout mice after trauma, whereas transgenic GFAP–IL-6 mice with a higher astrocytic IL-6 expression showed accelerated tissue revascularization and repair. Independent of brain injury, previous studies examining two different types of IL-6 transgenic expression reported a higher toxicity in GFAP–IL-6 overexpression (astrocytic) compared to neuron-specific enolase–IL-6 overexpression (neuronal) that did not reveal any neurologic abnormalities, although both exhibited astrogliosis and microglial cell activation. This finding enhances the idea of the diverse effect of cytokines, depending on their cellular origin.

In stroke models, one study demonstrated that a mild hypoxic insult increases the basal levels of glycogen within the brain. As such, mild insults 'precondition' the brain and decrease the extent of damage subsequent to more severe traumatic events. Levels of glycogen storage in astrocytes have been shown to be directly regulated by IGF-1 that is downstream of IL-1β signaling, which is induced after injury, specifically after a stroke. These data support the hypothesis that astrocytes, by responding to cytokines, comprise an important substrate of preconditioning insults. Furthermore, in line with the contention that microglia can benefit tissue repair, intracerebroventricular microinjection of exogenous microglia during occlusion of the middle cerebral artery in rats significantly inhibits the behavioral dysfunction induced by focal ischemia. The injected microglia possibly protect the neural tissue against ischemia-induced neurodegeneration, but the mechanism remains unclear.

Finally, in Alzheimer's disease transgenic mice models, inflammation has been shown to induce beneficial responses that include the activation of microglia to phagocytose dying cells or antibody assemblies. TGF-β1 and acute lipopolysaccharide (LPS) treatment seem to promote this phagocytic state. In contrast, chronic LPS treatment induces a neurotoxic state, highlighting the importance of duration of the microglial response. However, it must be stressed that in Alzheimer's disease, the inflammatory response is not directly neuroprotective but, rather, indirectly neuroprotective by targeting amyloid plaques; this is the rationale behind vaccine studies that have failed, probably because the inflammatory reaction could not be limited to plaques alone but affected surrounding neural tissue.

Remyelination

Emerging evidence implicates cytokine-activated astrocytes in the regenerative phases of demyelinating diseases. In particular, the cytokines IL-1β and CNTF play important roles. Animal experiments revealed that IL-1β is present early during the course of demyelination, whereas CNTF is present later, during the remyelination phase. These studies indicate that an abrogated astrocyte reaction due to a lack of IL-1β signaling significantly impairs remyelination. Astrocytes also become activated after injury by CNTF, and CNTF has been implicated in the production of FGF-2, which can enhance oligodendrocyte precursor proliferation.

Transgenically targeted ablation of reactive astrocytes after CNS injury markedly exacerbates tissue degeneration in both gray and white matter, increasing the death rate of both neurons and oligodendrocytes and increasing demyelination. NMDA receptor antagonists attenuate neuronal death, suggesting that astrocyte loss leads to accumulation of extracellular glutamate with subsequent excitotoxicity toward both neurons and oligodendrocytes.

TNF-α expression is induced in the CNS of cuprizone-treated mice, a model of demyelination and remyelination in which both microglia and astrocytes show evidence of cytokine production. The TNFR2 (p75

receptor), but not the TNFR1 (p55 receptor), is upregulated in the CNS of these mice. Using mice with targeted deletions of TNF-α and/or TNFR1 or TNFR2, studies showed that lack of TNF-α leads to a significant delay in remyelination. Further studies showed that this reparative function of TNF-α was dependent on expression of the TNFR2 because mice lacking this receptor fail to show regeneration of oligodendrocytes. Oligodendrocytes, as well as astrocytes and microglia, have been found to upregulate both TNFR1 and TNFR2 in response to inflammation; thus, changes in the relative levels of these receptors could alter the balance between cytotoxicity (p55 receptor) and growth and differentiation (p75 receptor). Furthermore, in the cuprizone model it was shown that IL-1β-deficient mice fail to remyelinate properly and that this failure correlates with delayed differentiation of oligodendrocyte precursors and with a lack of astrocytic and microglial production of IGF-1.

Axogenesis and Synaptogenesis

Reactive scar-forming astrocytes have long been regarded as a major obstacle to axon regeneration after CNS injury. Thus, it is important to acknowledge other data demonstrating that astrocytes support neurite outgrowth during development and that mature astrocytes may be essential for local nerve fiber sprouting and plasticity in adults. Moreover, reactive astrocytes express fibronectin and can support the regrowth of certain types of CNS axons along their cell surfaces. Also, CNTF-stimulated astrocytes promote neurite outgrowth better than untreated astrocytes. It will be important to understand the different contexts in which astrocytes might release inhibitory guidance cues or present permissive cues and to determine the signaling mechanisms that might regulate these different activities.

Few data are available regarding the role of inflammation at the synaptic level. However, it has been reported that ADNF stimulates synapse formation, and vasoactive intestinal peptide (VIP) elicits the release of ADNF from astrocytes. Therefore, the VIP-activated, astrocyte-derived factor ADNF acts directly on neurons to enhance synaptic connectivity and alter neuronal morphology.

Neurogenesis

Studies have demonstrated that astrocytes stimulate neurogenesis from adult neural stem cells (NSCs). Coculturing adult NSCs with primary hippocampal astrocytes increases the number of newly formed neurons tenfold. These results suggest that astrocyte-derived factors regulate neurogenesis during development. Also, neurogenesis in highly expanded NSCs can be rescued by coculture with microglial cells or microglia-conditioned medium, indicating that microglia provide a secreted factor(s) essential for neurogenesis but not for NSC maintenance, self-renewal, or propagation. These findings suggest an instructive role for microglial cells in contributing to postnatal neurogenesis the mammalian brain.

With regard to specific cytokines (from astrocytic origin) displaying neurogenic potential in the adult brain (e.g., in the adult dentate gyrus), TGF-β was shown to promote neurogenesis in NSC primary cultures. Furthermore, IL-1β and IL-6, and a combination of factors that included these two interleukins, were also able to promote NSC neuronal differentiation. The finding that cytokines and chemokines can promote NSC neuronal differentiation may help us to understand how injuries induce neurogenesis in adult brains.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Injury: Neuronal Responses; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Inflammation in Neurodegenerative Disease and Injury; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Neuronal Plasticity after Cortical Damage; Sleep and Sleep States: Cytokines and Neuromodulation.

Further Reading

Arnett HA, Mason J, Marino M, Suzuki K, Matsushima GK, and Ting JP (2001) TNF alpha promotes proliferation of oligodendrocyte progenitors and remyelination. *Nature Neuroscience* 4: 1116–1122.

Bush TG, Puvanachandra N, Horner CH, et al. (1999) Leukocyte infiltration, neuronal degeneration, and neurite outgrowth after ablation of scar-forming, reactive astrocytes in adult transgenic mice. *Neuron* 23: 297–308.

Davalos D, Grutzendler J, Yang G, et al. (2005) ATP mediates rapid microglial response to local brain injury *in vivo*. *Nature Neuroscience* 8: 752–758.

Faulkner JR, Herrmann JE, Woo MJ, Tansey KE, Doan NB, and Sofroniew MV (2004) Reactive astrocytes protect tissue and preserve function after spinal cord injury. *Journal of Neuroscience* 24: 2143–2155.

Gao X, Gillig TA, Ye P, D'Ercole AJ, Matsushima GK, and Popko B (2000) Interferon-gamma protects against cuprizone-induced demyelination. *Molecular and Cellular Neuroscience* 16: 338–349.

Herx LM and Yong VW (2001) Interleukin-1 beta is required for the early evolution of reactive astrogliosis following CNS lesion. *Journal of Neuropathology & Experimental Neurology* 60: 961–971.

Liberto CM, Albrecht PJ, Herx LM, Yong VW, and Levison SW (2004) Pro-regenerative properties of cytokine-activated astrocytes. *Journal of Neurochemistry* 89: 1092–1100.

Mason JL, Suzuki K, Chaplin DD, and Matsushima GK (2001) Interleukin-1beta promotes repair of the CNS. *Journal of Neuroscience* 21: 7046–7052.

Nimmerjahn A, Kirchhoff F, and Helmchen F (2005) Resting microglial cells are highly dynamic surveillants of brain parenchyma *in vivo*. *Science* 308: 1314–1318.

Schroeter M and Jander S (2005) T-cell cytokines in injury-induced neural damage and repair. *NeuroMolecular Medicine* 7: 183–195.

Schwartz M, Butovsky O, Bruck W, and Hanisch UK (2006) Microglial phenotype: Is the commitment reversible? *Trends in Neurosciences* 29: 68–74.

Shaked I, Porat Z, Gersner R, Kipnis J, and Schwartz M (2004) Early activation of microglia as antigen-presenting cells correlates with T cell-mediated protection and repair of the injured central-nervous system. *Journal of Neuroimmunology* 146: 84–93.

Song H, Stevens CF, and Gage FH (2002) Astroglia induce neurogenesis from adult neural stem cells. *Nature* 417: 39–44.

Neural Stem Cells and CNS Diseases

J-P Lee and E Y Snyder, The Burnham Institute for Medical Research; UCSD School of Medicine, La Jolla, CA, USA
S McKercher and J F Loring, The Burnham Institute for Medical Research, La Jolla, CA, USA
F-J Mueller, The Burnham Institute for Medical Research, La Jolla, CA, USA; and UKSH School of Medicine, Kiel, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The pathologic lesions of many neurologic disorders in the central nervous system (CNS) are often globally dispersed throughout the brain. But other disorders are manifest by a loss of more discrete neural populations affecting a more limited area of the CNS (Table 1). Transplantation of neural stem cells (NSCs) offers promising regenerative therapies for many CNS diseases. Results of transplantation experiments suggest that NSCs appear to be able to respond to neurogenic signals, reminiscent developmental cues evoked by degenerative processes. NSCs possess other inherent biologic properties that make them attractive for treatment of metabolic, degenerative, traumatic, or other lesions in the brain, particularly those that are widely distributed. They are easy to administer (often directly into the cerebral ventricles) and are readily engraftable. Additionally, preconditioning regimes such as total body irradiation are not required prior to administration (as required for bone marrow transplantation). In the following sections, we discuss these characteristics and the therapeutic potential of NSCs in CNS diseases.

Neural Stem Cell Biology

Definition of NSCs

NScs are the most primordial and least committed cells of the nervous system. Because immunocytochemically detectable markers that are sufficiently specific and sensitive have not yet been defined, the NSC is still best defined operationally. The NSC must have the following functional properties: (1) 'self-renewal' (produces unaltered daughter cells in response to mitogens) and (2) 'multipotency' (generates all three fundamental neuroectodermal lineages – neurons, astrocytes, and oligodendrocytes – in a regional and developmental stage-appropriate manner). There is no consensus about how many intermediary steps exist between the stem cell stage and the emergence of a mature functional cell. If a cell type appears to have a more limited self-renewal capacity and a narrower range of fates, it is by convention called a 'progenitor' or 'precursor' rather than a stem cell.

Stem Cell Niche and Function of NSCs in the Developing and Adult CNS

The population of NSCs in the adult brain is harbored in special locations of the CNS called 'niches.' To sustain homeostasis of NSC populations, NSCs proliferate and differentiate following an innate genetic program and in response to the signals from this specialized microenvironment. It is not yet known to what extent NSC's commitment is preprogrammed within the cell (i.e., cell autonomous), compared to the contribution of external signals (i.e., cell nonautonomous). It is possible that the number of cell divisions – and hence the degree to which multipotency and environmental responsiveness can be maintained – changes with the age of the stem cell or the age of the organism within which the stem cell resides.

In the early stages of mammalian cerebrogenesis, NSCs are located in a primary germinal zone known as the ventricular zone (VZ). In the telencephalon of an embryonic day-10 (E10) mouse, it is estimated that stem cells account for 5–20% of the cells in the VZ. The earliest NSCs divide symmetrically to amplify identical daughter cells. A stem cell retains continuous mitotic potential while remaining multipotent. Somewhat later in cerebrogenesis, in response to specific developmental signals, the NSCs then divide asymmetrically, to generate partially differentiated neural progenitor cells. Neural precursor cells begin to express differentiation markers while remaining mitotically active. They differentiate once they exit from the cell cycle, with the final fate of a given progenitor cell depending on extrinsic morphogenic signals, growth factors, and specific patterns of transcriptional activation, generally resulting in the generation of neurons prior to the generation of glia during development.

Secondary germinal zones, believed to develop from the VZ, includes the subventricular zone (SVZ) of the forebrain, the subgranule layer (SGL) of the hippocampus, and the external germinal layer (EGL) of the cerebellum. The EGL disappears once cerebellar development is completed, by 3 weeks postnatally in rodents and by 2 years of age in humans. NSCs are believed to populate these areas as well, perhaps allocated there during early development to maintain homeostasis throughout life. In the SVZ, NSCs are tightly packed and communicate with one another via

gap junctions. The SVZ is known to serve as an endogenous source of multipotent neural precursors that give rise to neurons and glia to support cell turnover and/or compensate for cell death. In the rodent SVZ, newly generated neuronal precursors reach their final destination in the olfactory bulb after long-distance migration through a well-defined path called the rostral migratory stream (RMS).

Although a rather limited region of the adult CNS has been documented to harbor NSCs – these secondary germinal zones and perhaps a dispersed group of NSCs throughout the gray matter and white matter – the very presence of such endogenous cells has raised the possibility that they may be recruited for neural reconstruction. It remains unclear whether adult-derived NSCs can make projection neurons and necessary connections as precisely as they are naturally generated in the embryo.

Table 1 Neuronal disorders in the CNS with potential stem cell action

Disease	*Potential stem cell action*[a]
Global diseases	
Lysosomal storage diseases	1, 2, 3, 4
Leukodystrophies	1, 2, 3, 4
Hypoxic–ischemic encephalopathy	3, 4, 5
Multiple sclerosis	1, 3, 4
Focal/discrete diseases	
Amyotrophic lateral sclerosis	2, 3, 4, 5
Parkinson's disease	1, 2, 3, 4
Huntington's disease	1, 2, 3, 4

[a]Potential action: 1, enzyme supply; 2, cell replacement; 3, neurotrophic factors; 4, anti-inflammatory molecules; 5, antioxidants.
The pathologic lesions of neurologic central nervous system (CNS) disorders are either globally dispersed or are in discrete areas of the CNS. Potential actions of transplanted stem cells are indicated for each disease.

Isolation, and Propagation of Murine and Human NSCs

There has been rapid progress in stem cell biology since it was first demonstrated that a single neural cell with stemlike qualities could be isolated from a mouse brain and reimplanted into another mouse brain, wherein progeny of the reimplanted cell then differentiated into integral members of the CNS (Figure 1). Since that time, neural cells with stem cell properties have been derived from the embryonic, neonatal, and adult rodent and human CNS using several different *in vitro* expansion methods.

Both mouse and human NSCs can be isolated from the ventricular zone of the fetal telencephalon and

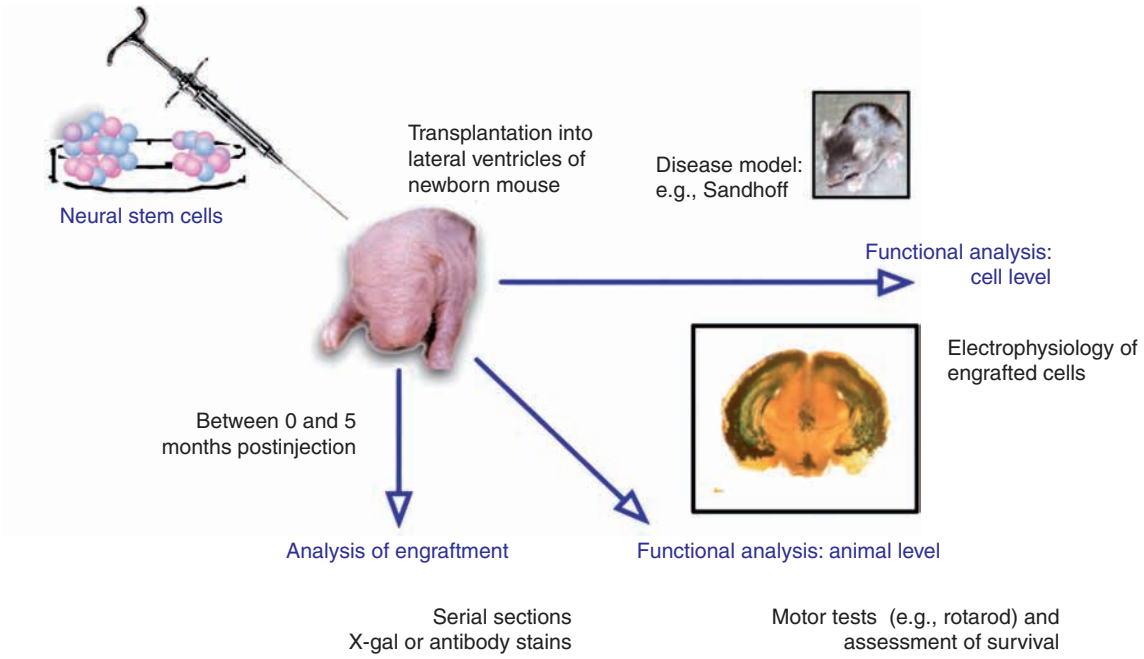

Figure 1 Implantation of neural stem cells into neonate brain of the mouse models of central nervous system diseases. *LacZ* β-galactosidase-labeled neural stem cells are traced by an X-gal reaction (blue) and/or anti-β-galactosidase immunocytochemistry in serial coronal sections of recipient brains. Assessment times of neural stem cell impact are shown in implanted mice in terms of survival and motor function and vary according to disease model. Electrophysiology on acute slices from engrafted brains can be performed to assess the identity and functional integration of donor neural stem cell-derived cells.

the SVZ, from the SGL, and even from the cortical parenchyma of the adult brain. NSCs are then selected and propagated using a serial growth factor selection procedure to select small numbers of cells that respond to both basic fibroblast growth factor (bFGF) and epidermal growth factor (EGF). The principal differences so far observed between mouse and human NSCs seem to be the length of the cell cycle (to 4 days in humans), the predilection of human NSCs to senesce (after ~100 cell divisions), and the relative NSC responsiveness to various types of mitogens. The ability to recognize marker proteins consistent with an immature phenotype has contributed to the characterization of such cells. Commonly used NSC markers include: nestin and vimentin, intermediate filament proteins; musashi 1, an RNA-binding protein; Sox1, a transcription factor; and CD133, a surface marker. These markers are selective rather than specific for NSCs because they are also expressed (at lower levels) in neural progenitor cells (i.e., cells with a more limited potential and self-renewal capacity) and even in nonneural cells (e.g., muscle and endothelium). Therefore, performing functional assays is better than looking for the presence of markers.

NSCs can also be derived secondarily *in vitro* from embryonic stem cells (ESCs), which are derived from the inner cell mass (ICM) of a blastocyst. ESCs can give rise to differentiated progeny from all three embryonic germ layers, including neural derivatives. Also, ESCs, by definition, never senesce, and they require a great deal of instruction to direct their differentiation fates unequivocally toward a given lineage. Using ESCs as a source of NSCs would obviate the problems of abstracting a relatively small population of 'native' NSCs from a given organ, maximize their proliferation, and avoid senescence. It is still under debate whether ESCs progress through a neural stem cell stage before they give rise to a particular neural cell type. Furthermore, being able to 'tame' an ESC is still quite challenging – undifferentiated ESCs can be tumorigenic following transplantation. For example, dopaminergic neurons were derived from human ESCs (hESCs) by co-culturing with a cell line composed of midbrain astrocytes. Transplantation of these cells into parkinsonian rats improved motor function. However, engrafted regions exhibited undifferentiated mitotic neoplastic-appearing neuroepithelial cells. There is growing evidence that tumorigenic potential may be reduced by significantly extending the neuronal differentiation time of the hESCs *in vitro* prior to transplantation.

Another approach is being pursued to create graft material for the CNS. It is somatic cell nuclear transfer (NT), whereby donor nuclei from a mature somatic cell of the intended recipient (e.g., a fibroblast) are inserted into an unfertilized oocyte, which, when stimulated, will divide to the blastocyst stage, from which ESCs can be derived. NSCs can theoretically be derived from these ESCs, but there are caveats to this approach. For example, nuclei from mature somatic cells may have developmentally imposed epigenetic changes, such as specific DNA methylation patterns that might be difficult to safely overcome. That the oocyte cytoplasm harbors molecules that can mediate a process of dedifferentiation is a fascinating concept, and this requires greater exploration and may be exploited in contexts even broader than NT (perhaps making 'adult cells' act stemlike).

The success in isolating stemlike cells from the CNS led to attempts to derive neural stemlike cells from nonneural organ systems, including bone marrow mesenchyme, muscle, skin, retina, fat, blood, and liver. Whether such nonneural stem cells can yield neural stem cells remains unresolved and exceptionally controversial.

NSCs and CNS Repair

Sources of NSCs for CNS Repair

Despite the presence of endogenous NSCs in the adult mammalian brain, it is recognized that following CNS injury, intrinsic self-repair activity of the adult CNS (at least for the most devastating of injuries) is inadequate or ineffective. This poor regenerative ability may be due to the restricted location and limited number of NSCs and/or to limitations imposed by the surrounding microenvironment, which may not be supportive or instructive for neuronal differentiation. Therefore, NSCs expanded *ex vivo* in culture and then implanted intracerebrally to regions in need of repair may overcome those limitations that are related to inadequate numbers of NSCs in proximity to the affected region.

Maintaining NSCs in a proliferative state *in vitro* by epigenetic or genetic manipulation does not appear to subvert their ability to respond to appropriate environmental cues *in vivo* following transplantation into the developing brain. Such cells have the ability to migrate throughout the brain and to differentiate appropriately and nontumorigenically into neurons, astrocytes, and oligodendrocytes. It remains to be determined, however, whether prolonged passaging of cells *in vitro* produces fundamental changes in these cells that begin to change them from their *in vivo* predecessors and counterparts.

It is important to learn what will be the best means to derive a source of NSCs: (1) isolating

NSCs directly from the neuroectoderm of fetuses or adults, (2) deriving NSCs from ESCs that have been directed toward a neuroectodermal and NSC lineage, or (3) isolating NSCs from nonneural but more accessible structures. Finding better genetic and epigenetic markers of NSC identity may be useful in mapping the precise location of adult NSCs as potential targets for new pharmaceuticals aimed at specific and selective activation, mobilization, and recruitment. Genomic methods (such as gene expression microarrays) and proteomic approaches may provide valuable tools for identifying these genes and their products. The neural differentiation genes may be useful targets for agents aimed at producing specific cell types from these NSCs. Certain genes, including those that help mediate mitogen action, can also be introduced into stem cells to extend the cells' ability to self-renew and preserve their multipotency. Such genes are *MYC*, telomerase reverse transcriptase, and SV40 large T antigen gene.

Identification and Fate of Donor NSCs in the CNS

NSCs can be labeled with the *lacZ* β-galactosidase protein, with bromodeoxyuridine (BrdU), or with fluorescent proteins such as enhanced green fluorescence protein (eGFP) for identification and tracing purposes following transplantation in mouse models of CNS diseases. *LacZ* encodes β-galactosidase (β-gal), thus allowing easy histochemical tracing of donor cells by the 5-bromo-4-chloro-3-iodolyl-β-D-galactosidase (X-gal) reaction and/or anti-β-gal immunocytochemistry in recipient brains. BrdU is incorporated into DNA during S phase and can be detected after fixation of the tissue by specific anti-BrdU antibodies. Fluorescently labeled cells can be viewed directly in live or fixed tissue sections. The presence of engrafted-human NSC-derived cells in transplanted mice can be confirmed by their immunoreactivity to antibodies against human-specific proteins.

The fate of engrafted NSCs in the recipient brain can be analyzed by using cell type-specific immunomarkers. These may include oligodendrocyte markers (2′,3′-cyclic nucleotide 3′-phosphodiesterase (CNPase), monoclonal antibodies O1 and O4, myelin basic protein (MBP)), astrocyte markers (glial fibrillary acidic protein (GFAP)), neuron markers (β-III-tubulin, neurofilament, NeuN, microtubule-associated protein-2 (MAP2), Hu-D), and immature progenitor markers (nestin, vimentin, musashi 1, NG2, A2B5). NSC fate can also be analyzed by looking at the ultrastructure for cell type identification, formation of synaptic contacts, presence of normal myelin, and condition of subcellular organelles. Retrograde tracing methods can be performed to see whether neuronal projections extend to the proper targets. Electrophysiology (often on live slices isolated from engrafted brains) can be performed to assess functional integration of the donor NSC-derived cells (Figure 1).

Properties of NSCs for CNS Therapy

Ideally, when producing engraftable cell populations, NSCs should be in an undifferentiated state and still capable of mitosis. Immediately postgrafting, NSCs replicate a limited number of times, then exit the cell cycle relatively quickly, migrate, and differentiate in response to local microenvironmental cues.

NSCs and their progeny should intermingle nondisruptively with endogenous neural progenitor/stem cells, responding to the same spatial and temporal cues in a similar manner. Because of their ability to develop into integral cytoarchitectural components of many regions throughout the host brain, as neurons, astrocytes, oligodendrocytes, and even incompletely differentiated but quiescent progenitors, NSCs may be able to replace a range of missing or dysfunctional neural cell types within the same region. NSCs, particularly in certain differentiation states, might express proteins that support regeneration (e.g., many neurotrophic factors, lysosomal enzymes, angiogenic factors, anti-inflammatory molecules, anti-oxidants).

NSCs can migrate and integrate widely throughout the brain, particularly when implanted into germinal zones, and may be attracted, even across long distances, to regions of neurodegeneration in brains of all ages.

Representative Therapeutic Uses

NSC therapy for globally affected CNS diseases Many CNS diseases are characterized by global degeneration or dysfunction. Treatment for these disorders requires widespread gene and/or cell replacement. Globally affected CNS diseases include the inherited neurodegenerative diseases of the pediatric age group (inborn errors of metabolism, including lysosomal storage diseases (LSDs) such as leukodystrophies and hypoxic–ischemic encephalopathy) and some adult CNS diseases (e.g., multiple sclerosis). The ability of NSCs to treat these diseases was first observed in a mouse model of genetically based neurodegenerative LSD known as mucopolysaccharidosis type VII (MPS VII; β-glucuronidase deficiency). Implanting wild-type, diffusely engrafting, enzyme-producing NSCs into the telencephalic ventricles of these mutant mice as newborns, to create 'chimeric forebrains,' essentially reversed the cerebral manifestations of the disease through cross-correction. This process is essentially like performing a bone marrow transplant on the brain.

NSC-mediated treatment of many LSDs may not require genetic manipulation but may rely on the

NSC's intrinsic production of the required enzyme, particularly when that enzyme is secreted and can be readily taken up by deficient cells and is required in only small quantities to restore normal metabolism. This strategy has proved promising for other enzyme deficiencies, including the gangliosidoses, in which β-hexosaminidase (Hex) requires repletion to thwart the pathological accumulation of GM_2 ganglioside in the brain.

A specific example of NSC use in this way is in the mouse model of Sandhoff disease (SD), a gangliosidosis characterized by a mutation in the Hex β subunit. SD is a uniformly fatal, rapidly progressive, autosomal recessive LSD caused by an absence of both HexA and HexB, leading to accumulation of GM_2 and GA_2 gangliosides in the CNS (**Figure 2(a)**). Transplantation of NSCs delayed disease onset and prolonged the survival of SD mice. Mouse and human NSCs (from both neuroectoderm and ESCs) migrate extensively throughout the brain following neonatal transplantation, thereby reducing widespread ganglioside storage. The small percentage of implanted NSCs generated electrophysiologically active neurons in SD mouse brain. Of note is that there is a synergistic effect from adding substrate-reducing pharmacological therapy to standard cellular transplantation treatment. In other words, the combined administration of NSCs plus drugs prolonged life span beyond what is seen from a single treatment alone (**Figure 2(b)**).

The preceding study provided additional insight into the role NSCs may play in LSDs, which is the provision of anti-inflammatory molecules. CNS inflammation, characterized by microglial activation and macrophage infiltration, is a hallmark of Sandhoff disease, found in humans and mice. Stem cells appear to home preferentially to regions characterized by an inflammatory signature (**Figure 3(a)**). Mouse and human NSCs constitutively express molecules that help mediate their homing to the pathological source of inflammatory chemokines. Transplanted NSCs reduced CNS inflammation in the SD brain, as measured using microglia/macrophage activation markers (**Figure 3(b)**).

In addition to widespread molecular therapy, one can imagine widespread neural cell replacement therapy as well. For example, the oligodendroglia of the dysmyelinated *shiverer* (*shi*) mouse are dysfunctional because they lack myelin basic protein, which is essential for effective myelination. Therapy, therefore, requires widespread replacement with normal MBP-expressing oligodendrocytes. When undifferentiated murine and human NSCs were transplanted at birth into the ventricles of newborn *shi* mice, NSCs became integrated throughout the *shi* brain and remyelinated some of the axons. In some recipient animals, the symptomatic tremor decreased. The ability of human NSCs to remyelinate is of particular importance because dysmyelination/demyelination plays an important role in many genetic and acquired neurodegenerative processes. For example, patients suffering from Krabbe leukodystrophy, also called globoid cell leukodystrophy (GCL), are devoid of galactosylceramidase activity (also called galactocerebrosidase, or GalC). The primary substrate of GalC is a constituent of the myelin sheath. GCL is characterized by CNS demyelination, degeneration of oligodendrocytes, and infiltration of macrophages. There is a long-standing question regarding the pathophysiology of GCL: Is the observed pathology a cell autonomous defect (i.e., a sick cell that dies) or is the disease caused by a toxic environment that would kill not only host cells but also new cells entering that environment?

The absence of GalC activity in *twitcher* (*twi*) mice (as in human GCL patients) may permit the buildup of toxic glycolipid psychosine. *In vitro* experiments suggest that the absence of GalC activity in *twi* mice predisposes neural cells (especially oligodendrocytes) to be more susceptible to the toxic metabolite. Wild-type murine NSCs (i.e., expressing a normal amount of GalC) tend to be resistant to those toxic effects. When transplanted into newborn *twi* mice, NSCs can survive and differentiate into myelinating oligodendrocytes throughout the *twi* cerebral environment. This suggests that the environment in the *twi* brain is not inherently nonpermissive for NSC survival and differentiation. Murine NSCs engineered to overexpress GalC appear to be more resistant to higher levels of psychosine than are unengineered, immature NSCs, and transplant of these overexpressing cells produces modest increases in the *twi* mouse life span. These results suggest that replacing both cells and molecules (e.g., NSCs that can become oligodendrocytes and secrete GalC to antagonize psychosine toxicity) may be a promising strategy to treat the disease. Thus, NSC-mediated cell replacement is feasible not only in diseases in which the disorder is cell intrinsic, but also in disorders in which the toxic environment can be rendered more benign via genetic manipulation.

In the preceding studies, multipotent NSCs appeared to shift their fate to yield more oligodendroglia, the particular type of cell that is needed. Consequently, NSCs might be capable of sensing the deficiency of a particular cell type, or perhaps a particular niche is created by the absence of a particular cell type and can alter NSC fate to compensate. This phenomenon was observed for neurons in two other circumstances. The *meander tail* (*mea*) mouse

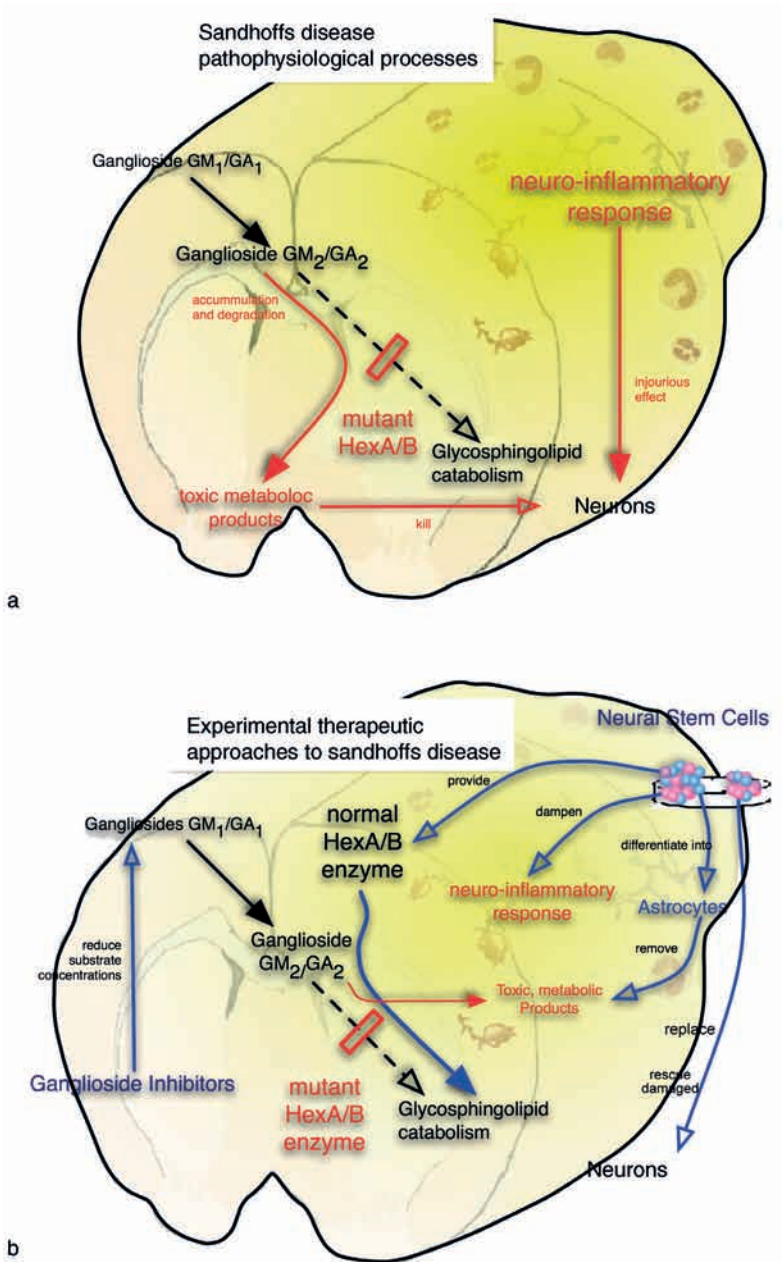

Figure 2 (a) Sandhoff disease pathophysiology. Lysosomal storage diseases, Tay–Sachs disease, and Sandhoff disease result in abnormal accumulation of GM_2 and GA_2 gangliosides due to genetic deletion of the enzyme β-hexosaminidase A (HexA) in Tay–Sachs disease, and of both HexA and HexB in Sandhoff disease, the latter disease thus being more severe. Besides ganglioside accumulation and toxic by-product production with concomitant neural dysfunction, these diseases are also marked by a prominent injurious inflammatory signature. (b) Multimodal therapeutic potential of neural stem cells in lysosomal storage diseases. In this model, neural stem cell transplantation (trophic effects and/or cell replacement), substrate reduction therapy (inhibition of ganglioside production and supply of Hex enzymes), and reduction of inflammation are complementary therapies.

mutant is characterized by the failure of a sufficient number of granule cell (GC) neurons to develop in certain regions of the cerebellum. NSCs implanted at birth produced GCs at the expense of other cell types, suggesting that environmental signals may encourage NSC differentiation into this inadequately developed cell type. This phenomenon is not restricted to young animals. When NSCs were implanted into an adult rodent neocortex in which a narrow region was depleted of projections, graft cells differentiated into pyramidal neurons only within the abnormal region. Outside that region, they became nonneurons in the cortex.

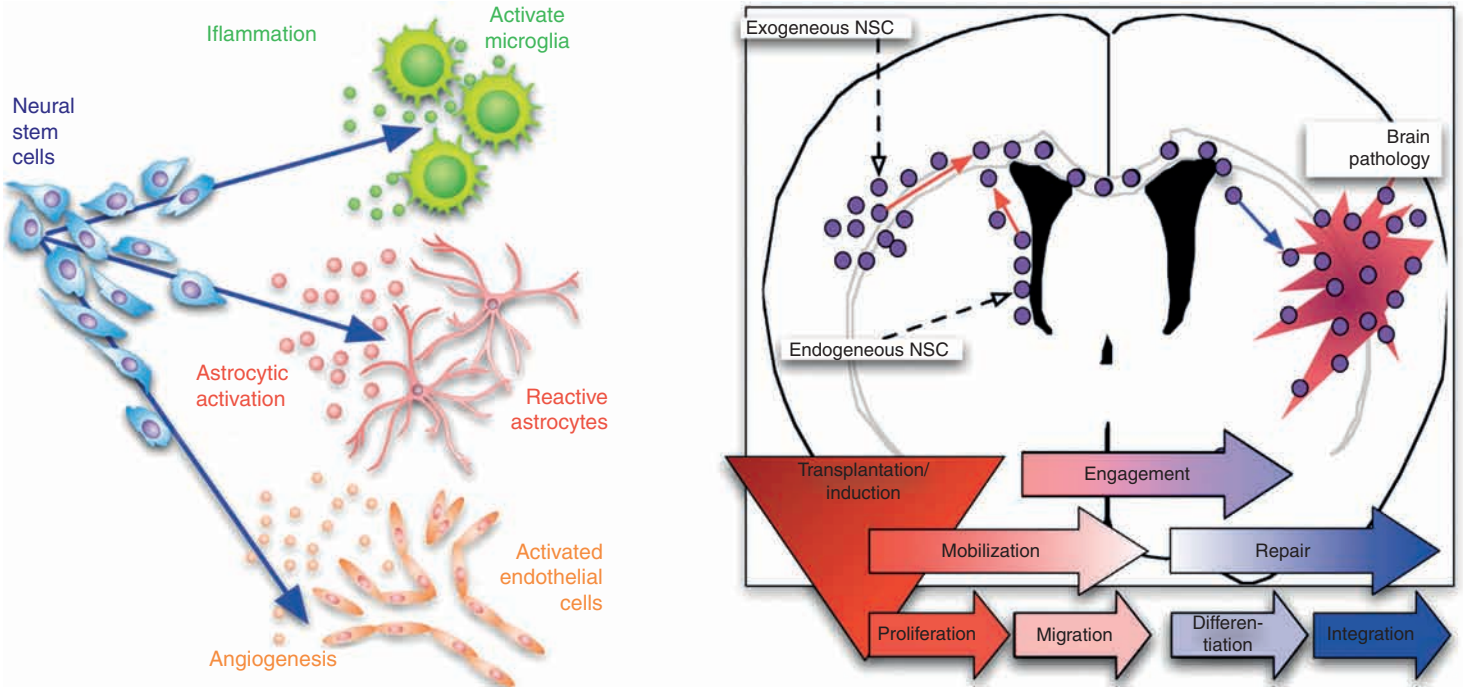

Figure 3 (a) Physiological processes common to many brain pathologies attract neural stem cells. Inflammation: activated microglia mediate pathology-induced central nervous system inflammation by secreting cytokines and chemokines, resulting in activation of astrocytes and endothelial cells. Reactive astrocytosis: direct pathology effects and microglial signals activate astrocytes, which secrete their own set of chemotactic factors, increasing the inflammatory reaction. Lesion-induced angiogenesis and inflammation-activated endothelial cells: secretion of more chemoattractant factors enhances neural stem cell homing to the area of brain pathology, where the stem cells are presented with an atypical, perivascular niche for their support. (b) Transplanted neural stem cells (NSCs) show tropism for areas of brain pathology. Note the phenomenon of exogenous NSCs migrating to the area of brain pathology from a distant implantation site. This homing ability may have therapeutic applications for targeting widespread metastatic central nervous system pathology.

The examples discussed here highlight the potential for an abnormal environment to direct the behavior of transplanted NSCs. However, we are beginning to learn that the response between an NSC and the injured host is one of ongoing reciprocal interactions. In response to implanted NSCs in the context of an injured CNS, the host nervous system can also contribute to its own repair. For example, hypoxia–ischemia (HI), a common cause of neurologic disability in adults and children, causes much of its damage from an extensive loss of cerebral parenchyma and loss of the cells and connections that reside there. When NSCs were implanted into animal models of this disorder, the exogenous implant and the injured host brain both participated in the reconstitution of parenchyma and anatomic connections, as well as in the reduction of parenchymal loss, secondary cell loss, inflammation, and scarring. Similar results were observed in the hemisectioned adult rodent spinal cord, in which there was evidence of an upregulated host neuronal regenerative response.

NSC therapy in locally affected CNS disorders In most neurodegenerative disorders (e.g., Parkinson's disease), neurons do not die abruptly but instead undergo a gradual, inexorable deterioration leading to death. Often the neurons that die comprise a discrete subtype. To explore the effect of NSCs on this type of pathological process, a mouse model of Parkinson's disease was generated by sequential systemic administration of high doses of the selective dopaminergic (DA) toxin 1-methyl-4-phenyl-1,2,3, 6-tetrahydropyridine (MPTP) to aged mice. The MPTP permanently impaired (but did not kill outright) the DA neurons of the mesostriatal nuclei. When such mice received implants of murine NSCs, the DA cell population was reconstituted, as demonstrated by renewed expression of such DA-specific molecules as the DA transporter (DAT) and, importantly, by behavior. Although restoring this DA function was graft dependent, it was not predominantly due to differentiation of the donor NSCs into DA neurons. The majority of the functionally restored DA neurons were host neurons. It appears that in this case, the transplanted NSCs were serving a supportive function, perhaps by secreting glial cell line-derived neurotropic factor (GDNF), a neuroprotective factor of DA neurons. A similar observation is beginning to emerge from the implantation of human NSCs into the MPTP-lesioned subhuman primate model of Parkinson's disease, as well as the implantation of murine and human NSCs into the spinal cords of the superoxide dismutase 1 (SOD1) transgenic mouse model of amyotrophic lateral sclerosis (ALS). These observations suggest that exogenous NSCs may not only replenish inadequate pools of endogenous NSCs for the purposes of compensating for missing neural cells, but may also serve to enhance endogenous regenerative capacities.

Future Needs and Directions

NSCs hold great promise for regenerative medicine through cell or enzyme replacement or as a means to enhance intrinsic regenerative processes. However, successful translation of the NSCs therapeutic potential in animal models to that in human diseases will rely on further characterization of the following aspects: (1) detailed characterization of human NSCs, including identification of their markers and mapping of the neural cell lineages involved in brain development; (2) better understanding of methods for controlling the propagation of and phenotype specification of NSCs *in vitro* and *in vivo*; (3) identification of the genetic and epigenetic processes that confer NSC identity to an immature cell and ultimately to mature neural cell-type specification; (4) examination of functional integration of engrafted NSCs, such as a neuron's ability to form synaptic connections that are electrophysiogically active and appropriate; and (5) observations of behavioral recovery following NSC engraftment.

We also require a better understanding of the pathophysiology of the diseases to be targeted – that is, knowing what aspects actually require repair and which cell types really require replacement or rescue. Finally, it is clear that complex diseases such as those affecting the nervous system will require complex and multifaceted solutions, including pharmacologic, genetic and molecular, cell replacement, tissue engineering, angiogenic, anti-inflammatory, antiapoptotic, proregenerative, and proneurite-outgrowth-promoting therapies. The NSC, as a key player in a set of fundamental developmental mechanisms, may serve as the glue that holds many of these strategies together, by participating dynamically in multiple aspects of the therapeutic intervention. However, determining how these strategies may be intelligently, effectively, and safely orchestrated in clinical practice will require a great deal more careful investigation.

See also: Gene Therapy: Genetically Modified Cells; Gene Therapy: Direct Viral Delivery; Gene Therapy and Protection from Stress-Induced Brain Damage; Inflammation in Neurodegenerative Disease and Injury; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neural Stem Cells: Adult Neurogenesis; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Transplantation of Myelin Forming Cells.

Further Reading

Imitola J, Raddassi K, Park KI, et al. (2004) Directed migration of neural stem cells to sites of CNS injury by the stromal cell-derived factor 1alpha/CXC chemokine receptor 4 pathway. *Proceedings of the National Academy of Sciences of the United States of America* 101: 18117–18122.

Lee JP, Jeyakumar M, Gonzalez R, et al. (2007) Stem cells act through multiple mechanisms to benefit mice with neurodegenerative metabolic disease. *Nature Medicine* 13(4): 439–447.

Myerowitz R, Lawson D, Mizukami H, et al. (2002) Molecular pathophysiology in Tay–Sachs and Sandhoff diseases as revealed by gene expression profiling. *Human Molecular Genetics* 11: 1343–1350.

Ourednik J, Ourednik V, Lynch WP, et al. (2002) Neural stem cells display an inherent mechanism for rescuing dysfunctional neurons. *Nature Biotechnology* 20: 1103–1110.

Reubinoff BE, Itsykson P, Turetsky T, et al. (2001) Neural progenitors from human embryonic stem cells. *Nature Biotechnology* 19: 1134–1140.

Roy NS, Cleren C, Singh SK, et al. (2006) Functional engraftment of human ES cell-derived dopaminergic neurons enriched by coculture with telomerase-immortalized midbrain astrocytes. *Nature Medicine* 12(11): 1259–1268.

Snyder EY, Taylor RM, and Wolfe JH (1995) Neural progenitor cell engraftment corrects lysosomal storage throughout the MPS VII mouse brain. *Nature* 374: 367–370.

Taylor RM, Lee JP, Palacino JJ, et al. (2006) Intrinsic resistance of neural stem cells to toxic metabolites may make them well-suited for cell non-autonomous disorders: Evidence from a mouse model of Krabbe leukodystrophy. *Journal of Neurochemistry* 97(6): 1585–1599.

Temple S (2001) The development of neural stem cells. *Nature* 414: 112–117.

Zhang SC, Wernig M, Duncan ID, et al. (2001) *In vitro* differentiation of transplantable neural precursors from human embryonic stem cells. *Nature Biotechnology* 19: 1129–1133.

Neural Stem Cells: Adult Neurogenesis

C Zhao and F H Gage, Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Adult neural stem cells (NSCs) are cells in the adult nervous system that can self-renew and differentiate into all types of neural cells, namely neurons, astrocytes, and oligodendrocytes. Free-floating neurosphere and adherent monolayer cultures are the methods that are commonly used to expand and characterize NSCs. Both of these types of cultures contain heterogeneous populations of cells, including the putative NSCs and their progeny. The self-renewal of adult NSCs is determined by serial passage and clonal analysis. The multipotency of adult NSCs is characterized by their ability to differentiate into all types of neural cells from single neurospheres or single cell clones. The existence and characteristics of adult NSCs have been mostly analyzed *in vitro* and have yet to be corroborated *in vivo*. Therefore, the term progenitor has often been used to loosely describe all proliferating precursor cells. Progenitor is a generic term for any dividing cell with the capacity to differentiate. This includes putative stem cells in which the self-renewal and/or multipotency have not yet been demonstrated. Precursors are cells from which more mature cells can be formed.

Location of Adult NSCs in the Mammalian Brain

The phenomenon that new neurons are generated in the postnatal brain was first described by Altman and colleagues in the 1960s. It is now known that neurogenesis continues in the adult central nervous system at two locations: the subventricular zone (SVZ; also known as the subependymal layer) of the lateral ventricles and the subgranular zone (SGZ) of the dentate gyrus of the hippocampus. In rodents, new neurons generated in the SVZ migrate through the rostral migratory stream (RMS) and incorporate into the olfactory bulb, whereas new neurons generated in the SGZ migrate a short distance into the granular cell layer and become granule neurons (Figures 1(a)–1(c)).

The identification and characterization of adult NSCs began in the 1990s, when Reynolds and Weiss found that proliferating cells could be isolated from adult mouse brain and expanded in culture when supplemented with epidermal growth factor (EGF). These cells form neurospheres in culture and secondary neurospheres when dissociated. This characteristic fulfilled one criterion for a stem cell: self-renewal. Cells in the neurospheres can differentiate into astrocytes and neurons, as defined by their expression of glial fibrillary acidic protein (GFAP) and neural-specific enolase, respectively. This phenomenon fulfilled the other criterion for a stem cell: multipotency. A collaborative effort between the groups of Weiss and van der Kooy later found that the neurosphere-forming cells reside in the periventricular region of the lateral ventricles.

The periventricular region of the lateral ventricles consists of the ependymal cells lining the ventricles and the SVZ. An observation by Frisen and colleagues suggested that the ependymal cells were the stem cells that gave rise to the proliferating cells in the SVZ. However, independent studies by the groups of Alvarez-Buylla, van der Kooy, and Temple showed that ependymal cells did not form neurospheres in culture and suggested that adult NSCs resided in the SVZ. A lentiviral labeling study by Vescovi, Naldini, and colleagues provided more definitive evidence that cells in the SVZ, but not the ependymal layer, are responsible for the long-term, continuous neurogenesis in the olfactory bulb. Lentiviral vectors can integrate into the chromosomal DNA of the host cells; therefore, the progeny of labeled cells can be traced. Even 3 and 6 months after lentiviral injection into SVZ, labeled cells were found in the RMS. The long-term presence of labeled cells in the RMS was not seen when the virus was injected into the ventricles to label ependymal cells. Therefore, SVZ has been considered to be the site of adult NSCs in the periventricular zone.

Gage and colleagues were the first to characterize NSCs from the hippocampus, the other brain region in which adult neurogenesis occurs. Adult hippocampal progenitor (AHP) is a term that has often been used to describe all proliferating precursor cells in the adult hippocampus. The progenitors/NSCs isolated from hippocampus could be passaged over an extended period of time in monolayer cultures when exposed to mitogens such as basic fibroblast growth factor (bFGF; also known as FGF2). Clonally derived adult hippocampal NSCs in monolayer culture could differentiate into all three neural lineages. In addition, cultured adult rat hippocampal progenitors were capable of neuronal differentiation when grafted into the hippocampus and the RMS of the adult rat. Contrary to these findings, van der Kooy and colleagues suggested that the hippocampus does not have multipotential NSCs but only progenitors because they were not able to isolate cells

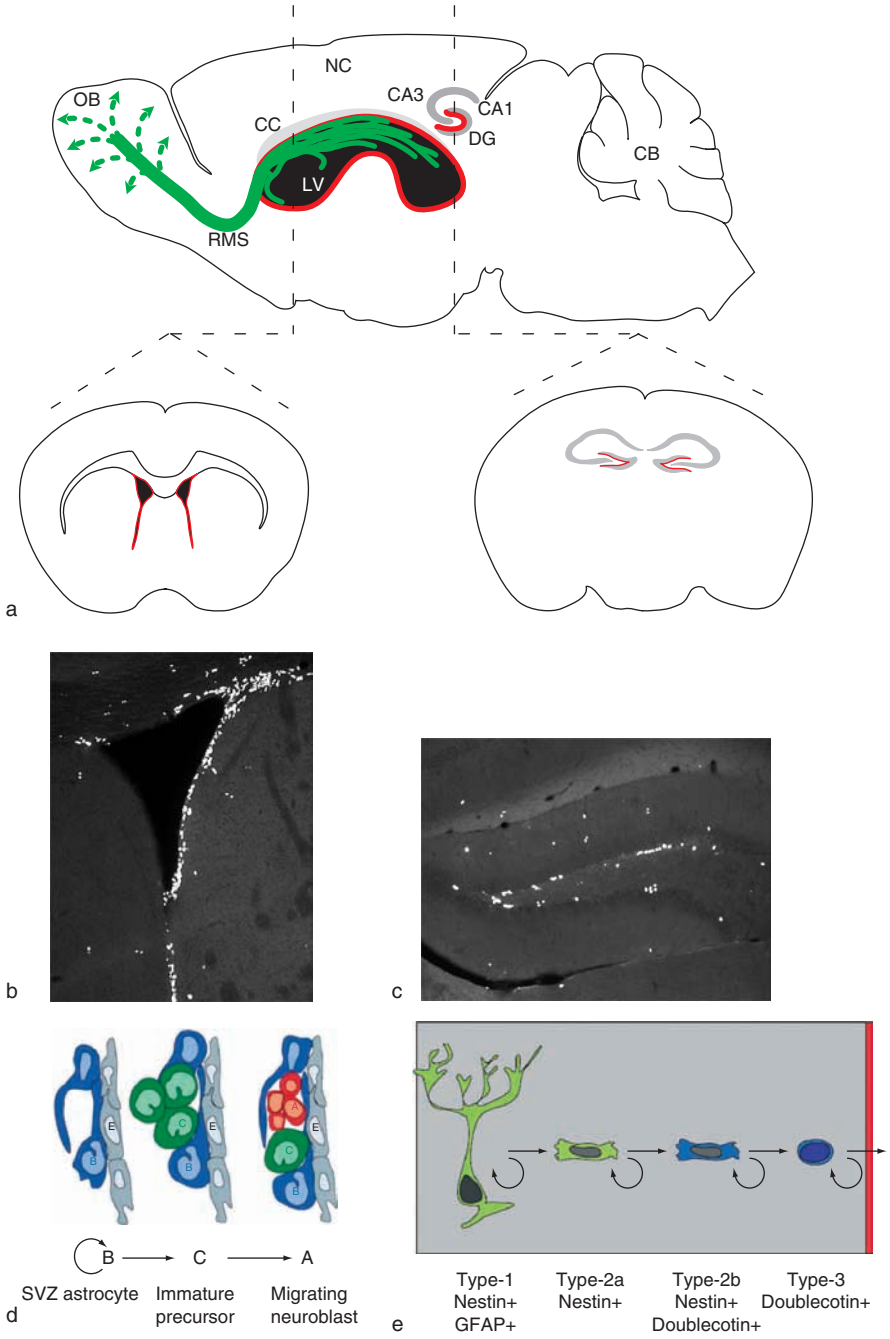

Figure 1 Neurogenesis occurs in the subventricular zone (SVZ) and the subgranular zone (SGZ). (a) Sagittal and coronal views of the mouse brain with the two main germinal zones shown in red. New neurons are generated in the SVZ of the lateral ventricles (LV) and the SGZ of the dentate gyrus (DG) of the hippocampus. Neurons generated in the SVZ migrate through the rostral migratory stream (RMS) and incorporate into the olfactory bulb (OB). Neurons generated in the SGZ migrate a short distance and become granule neurons in the dentate granule cell layer. CB, cerebellum; cc, corpus callosum; NC, neocortex. (b) An example of cell proliferation in the SVZ. Mice were given three shots of BrdU over 24 h and the sample was taken 2 h after the last BrdU administration. Note that more proliferating cells are found on the lateral side of the SVZ. (c) An example of cell proliferation in the SGZ. Images in (b) and (c) were kindly provided by Wei Deng. (d) It has been proposed that three types of cells exist in the SVZ. Type B cells are GFAP-expressing NSCs that give rise to C (immature precursor) cells and B (migrating neuroblast) cells. Reprinted from Doetsch F, Caille I, Lim DA, Garcia-Verdugo JM, and Alvarez-Buylla A (1999) Subventricular zone astrocytes are neural stem cells in the adult mammalian brain. *Cell* 97: 703–716, Copyright 1999, with permission from Elsevier. (e) Three types of proliferating cells have been proposed to exist in the subgranular layer of the DG. The identification of these cells is mainly based on their specific expression patterns of molecular markers. Adapted from Kempermann G, Jessberger S, Steiner B, and Kronenberg G (2004) Milestones of neuronal development in the adult hippocampus. *Trends in Neuroscience* 27: 447–452, Copyright 2004, with permission from Elsevier.

from the hippocampus that form neurospheres; furthermore, they only found neuronal differentiation. However, neurogenesis continues in the dentate gyrus throughout the life of mammals, although at a reduced rate upon aging. Therefore, it is unlikely that these new neurons are derived from cells with limited self-renewal capacity.

Identity of Adult NSCs

Adult NSCs in the SVZ are thought to be slow-dividing cells. Following a high dose of ^{3}H-thymidine to kill proliferating cells in the SVZ, these cells can be repopulated within a few days, suggesting that the proliferating cells are derived from some relatively quiescent cells that are not susceptible to the antimitotic treatment. A similar phenomenon was seen when another antimitotic drug, cytosine arabinoside (Ara-C), was used. The proliferating cells in the SGZ of the dentate gyrus could also recover within 2–3 weeks after the animals were treated with the DNA methylation agent methylazoxymethanol acetate (MAM); however, whether this recovery is caused by the expansion of the remaining proliferating cells or whether slow-dividing NSCs also exist in the SGZ is not known.

Three types of cells have been proposed to exist in the SVZ: type B NSCs, type C transit-amplifying cells, and type A migrating neuroblasts. Type B cells appear to express GFAP and they are the only cells in the SVZ that survived the antimitotic treatment. Type C and A cells recover sequentially soon after cessation of the treatment, leading to a lineage hypothesis from B to C to A cells (**Figure 1(d)**). The identification of B, C, and A types of cells mainly relies on the morphological analyses at the electron microscopic level. In parallel to the SVZ, three types of precursor cells appear to exist in the SGZ: type 1 stem/progenitors and type 2 and type 3 precursor cells. The identification of these different types of cells in the SGZ is mainly determined by the specific expression patterns of molecular markers (**Figure 1(e)**).

Identification of NSCs by Cell Division Markers

One of the features of NSCs is their ability to divide; therefore, the thymidine analogues ^{3}H-thymidine and 5-bromo-2-deoxyuridine (BrdU) are often used to characterize adult NSCs. Both chemicals can be administered systemically, and the labeled cells can be detected through autoradiography and immunohistochemistry. Because NSCs in the SVZ are considered to be slow dividing, and because new neurons generated in the SVZ migrate to the olfactory bulb through the RMS, cells that remain labeled in the SVZ long (1 month) after BrdU or ^{3}H-thymidine labeling are presumably mainly NSCs. In addition to nucleic acid analogues, molecular markers such as the proliferating cell nuclear antigen and Ki67 have been used to recognize mitotic cells by immunohistochemistry. Another way of identifying NSCs is to use retroviruses because retroviral vectors can integrate into the chromosomal DNA and their integration is dependent on mitosis. The ubiquitous distribution of the molecular markers allows detailed morphological analyses of labeled cells. Furthermore, when retrovirus is used to mediate the expression of fluorescent proteins, live cells can be followed in both cell culture and brain slice preparations. All of these methods label both NSCs and their progeny and are biased toward cells with an active cell cycle. Since SVZ NSCs are proposed to be relatively quiescent, the labeling efficiency of NSCs is hypothetically lower than that of rapidly dividing, transit-amplifying cells.

Identification of NSCs by Molecular Markers

Using a transgenic mouse line in which avian leucosis virus-mediated reporter gene transfer only takes place in cells with an active GFAP promoter, Doetsch, Alvarez-Buylla, and colleagues showed that C and A cells in the SVZ were indeed derived from GFAP-expressing B cells. Studies using transgenic mice expressing cre recombinase from the murine GFAP promoter provided consistent results. In addition, in transgenic animals that expressed the herpes simplex virus thymidine kinase (TK) from the murine GFAP promoter, adult neurogenesis in the olfactory bulb and the dentate gyrus was diminished when the animals were given elaidic acid ganciclovir, a drug that can be converted to a toxin by TK. These studies suggest that a subpopulation of GFAP-expressing cells is the precursor of newborn neurons. GFAP-expressing cells have two different morphologies: bipolar or unipolar GFAP cells are considered to be stem cells, and multipolar GFAP cells are astrocytes. Because GFAP-expressing astrocytes are much more abundant than GFAP-expressing NSCs, GFAP has not been considered to be an optimal marker for adult NSCs.

The intermediate filament protein nestin is specifically expressed in the neuroectoderm during development and is highly expressed in the germinal zones in the adult brain. Bartlett and colleagues used fluorescence-activated cell sorting (FACS) to separate cells based on their size (>12 μm), low binding to peanut agglutinin, and low level of the heat stable antigen (mCD24a). Approximately 80% of purified cells can form neurospheres, and all of them express nestin.

Sox2 is a member of the Sry-related HMG box transcription factors. It is expressed in totipotent

inner cell mass stem cells and in the neuroepithelium of the developing central nervous system. In the adult brain, Sox2 is expressed in both SVZ and SGZ, although its expression is not restricted to these two neurogenic areas. Most Sox2-expressing cells appear to be proliferating neuroblasts. A subpopulation of Sox2 cells also expresses GFAP with one major process spanning the granular cell layer, and these cells have been proposed to be NSCs. Although Sox2 deficiency was shown to cause impaired neurogenesis, no fate mapping studies have been documented using Sox2 as a marker.

A number of cell surface markers have been reported to be able to enrich NSCs through FACS. LeX/ssea-1 (also known as CD15), fucose-containing trisaccharide, was found on the surface of mouse embryonic stem cells and primordial germ cells. Approximately 25% of FACS-sorted LeX+ cells from the adult mouse SVZ generated neurospheres *in vitro* and approximately 18% of LeX+ cells are GFAP+. A monoclonal antibody 473HD that recognizes the chondroitin sulfate epitope was also able to enrich the population of neurosphere-forming cells.

Several other markers have also been reported based on studies in other systems or during early development, including CD133 and Musashi homologue 1. However, these markers have not been extensively characterized in adult NSCs. All of the molecular markers discussed here label heterogeneous populations of cells *in vivo*. Cells that express both GFAP and nestin (or Sox2) have been proposed to be adult NSCs. However, it is not clear whether these cells represent only a subpopulation of adult NSCs.

Identification of NSCs by Genetic Methods

The second intron of the nestin gene contains an enhancer element that is specific for its expression in the central nervous system. This regulatory element has been used either in conjunction with the nestin promoter or with hsp68 minimum promoter to label NSCs. When the nestin promoter and the enhancer from the second intron were used to drive the expression of green fluorescent protein (GFP), GFP expression was more or less restricted to neurogenic regions in the adult brain, including both SVZ and SGZ. GFP was also found in the RMS. FACS-sorted nestin:GFP cells can self-renew and are multipotent. Nestin:GFP cells in the dentate gyrus appear to have two distinct populations, GFAP-positive radial glia-like cells and polysialylated neural cell adhesion molecule (PSA-NCAM)-positive cells. The GFAP+ nestin–GFP cells have been proposed to be stem/early progenitors (type I cells) that give rise to the PSA-NCAM+ nestin–GFP cells (type II cells). Transgenic mice have also been generated using the Sox2 promoter fragment to drive the expression of the reporter β-geo or of GFP.

Regulation of Adult NSCs

Adult NSCs can be regulated at different levels, including the self-renewal and fate determination of NSCs, proliferation and differentiation of their immediate progeny, and cell survival at different stages. All of these forms of regulation may eventually be reflected in a change in adult neurogenesis. Indeed, many conditions have been reported to alter adult neurogenesis. Aging and stress appear to have a negative effect on cell proliferation in both SVZ and SGZ. Hippocampal neurogenesis is enhanced by enriched environment, voluntary exercise, antidepressant treatment, seizure, and many other conditions. It is not clear which component of the process of neurogenesis is affected by these different conditions. Determining the microenvironment in which the NSCs reside (also known as the neurogenic niche) and the intrinsic mechanisms that instruct NSCs will help us better understand the physiological and pathological conditions that are intricately associated with adult neurogenesis.

Extracellular Influences

The behavior of adult NSCs and their progeny is directly influenced by their surroundings – that is, the microenvironment (neurogenic niche) in which they reside. NSCs have been isolated from almost all areas of the adult brain and from the spinal cord. These cells display multipotency in culture; however, neuronal differentiation only occurs in the SVZ and SGZ.

A clear example of the microenvironment playing a critical role in adult neurogenesis came from a transplantation study in which cultured spinal cord NSCs were able to give rise to neurons when grafted into the hippocampus but they showed no neuronal differentiation when grafted back into the spinal cord. Gage and colleagues later found that hippocampus-derived astroglia were able to promote both the neuronal differentiation of AHPs and the maturation of differentiated neurons in culture. Astrocytes from the spinal cord did not have such effects, suggesting that the presence of hippocampal astrocytes might be one of the determinants that permit adult neurogenesis in the dentate gyrus. Wnt-3 might be one of the factors that are responsible for astroglia-induced neurogenesis. Overexpression of Wnt-3 was sufficient to promote neurogenesis, whereas expression of a dominant negative form of Wnt-3 blocked neurogenesis *in vivo*. Neurogenesin-1, a bone morphogenesis protein (BMP)-interacting molecule, has

also been proposed to be a component secreted by hippocampal astrocytes that promote neuronal differentiation.

Dividing cells in the dentate gyrus were found to be closely associated with the vasculature, and possibly coupled with angiogenesis. Endothelial co-culture can significantly increase the number of new neurons in adult SVZ NSC adherent clones. Studies using embryonic cortical NSCs have proposed that factors derived from endothelial cells promote symmetric cell division and, therefore, self-renewal of NSCs. Interestingly, adult mouse NSCs can even transdifferentiate into endothelial lineage when co-cultured with human endothelial cells. Supporting a role for the vasculature niche in adult neurogenesis, the level of vascular endothelial growth factor (VEGF) is closely linked to the level of neurogenesis in the adult hippocampus. VEGF is required for enriched environment- and running-induced neurogenesis in SGZ. VEGFR2 (also known as fetal liver kinase-1 (Flk-1) and kinase insert domain receptor (KDR)) may be expressed by Doublecortin+ cells in the neurogenic zones. Expression of a dominant negative VEGFR2 is able to antagonize VEGF-induced neurogenesis. One study has suggested that, in addition to VEGF, another factor secreted by endothelial and ependymal cells, pigment epithelium-derived factor, can promote self-renewal of adult SVZ NSCs.

Ependymal cells were also reported to express Noggin, an antagonist of BMPs. Noggin might promote SVZ neurogenesis in the adult brain. Ectopic expression of Noggin in striatum was able to promote neuronal differentiation of transplanted NSCs. BMPs promote astrocytic fate in embryonic NSCs. Similarly in the adult, BMP and the leukemia inhibitory factor have been used to induce astrocytic differentiation of cultured AHPs.

The existing neurons may influence adult NSCs through neurotransmitters. Indeed, neurotransmitters such as dopamine, γ-aminobutyric acid (GABA), serotonin, and glutamate have been shown to regulate adult NSCs and/or their progeny. It is controversial whether dopamine promotes or inhibits the proliferation of NSCs in the SVZ. The dopaminergic neurons in the substantia nigra may project to the SVZ and affect SVZ neurogenesis. In the SGZ, GABA could mediate the input to the putative transit-amplifying progenitors and promote the differentiation of these cells and the maturation of the newly generated granule neurons. Antidepressant drugs that act to inhibit serotonin reuptake can consistently upregulate neurogenesis in the SGZ. In addition, glutamate agonist and antagonist could reduce and increase proliferation in the dentate gyrus, respectively. Contrary to these observations, glutamate agonist and antagonist were shown to increase and decrease the proliferation of AHPs *in vitro*.

In addition to the effects of different types of cells within the neurogenic niche, NSCs are subjected to regulation by soluble factors including growth factors, growth hormones, and cytokines. EGF and FGF2 are potent growth factors maintaining SVZ NSCs in culture. In fact, EGF has been shown to convert transit-amplifying cells into multipotent stem cells *in vitro*. EGF also has been claimed to convert adult NSCs to radial glial cells. The morphogen sonic hedgehog (Shh) could induce the proliferation of AHPs in culture. A transgenic study showed that NSCs were responsive to Shh *in vivo*, and that Shh-responsive cells might represent the slow-dividing cells described by van der Kooy and colleagues. Although many factors have been implicated in regulating adult NSCs, it is not clear which cells in the microenvironment secrete these factors.

Intracellular Signaling

It is perhaps not surprising that genes involved in cell cycle regulation, DNA repair, and chromosomal integrity are found to regulate NSCs. SVZ NSCs from p53 knockout mice form more and larger neurospheres compared to wild-type controls. Mice deficient in the cell cycle inhibitor p21 have more NSCs in the SVZ initially, but the number of NSCs decreases after they are 8 months old. It has therefore been proposed that NSCs have a limited number of cell divisions. An intact telomere appears to be important to maintain adult neurogenesis. Reduced proliferation was seen both *in vivo* and *in vitro* in the absence of the telomerase RNA component (Terc) after the mice were bred for four or five generations and the condition worsened when the mice were also deficient in ataxia-telangiectasia mutated (Atm). More interesting, one study suggested that the retrotransposon long interspersed nuclear element-1 (LINE-1 or L1) is actively expressed in adult hippocampal NSCs, and an engineered human L1 was shown to be mobile in the mouse brain.

Adult NSCs are also regulated at the epigenetic level and by transcription factors and their cofactors. Valproic acid, an inhibitor of histone deacetylase, could induce neuronal differentiation in cultured AHPs. Adult NSCs lacking the methyl-CpG binding protein 1 exhibited increased genomic instability and reduced neuronal differentiation. TLX, an orphan nuclear receptor, is required for the maintenance of adult forebrain NSCs. As mentioned previously, sox2 deficiency causes defects in neurogenesis. NeuroD, a neuronal differentiation basic helix–loop–helix transcription factor, is required for the formation of the

dentate gyrus. In addition, the polycomb family transcriptional repressor Bmi-1 was found to be required for the maintenance of both embryonic and adult NSCs. During development, the maintenance and fate specification of embryonic NSCs are defined by a series of bHLH transcription factors. The precise mechanisms of how self-renewal and differentiation are modulated in the adult NSCs still need to be determined.

The Function of Adult NSCs

Adult NSCs have been isolated from many species, including humans. Studies have clearly demonstrated that newly generated neurons in the dentate gyrus and the olfactory bulb are functionally integrated into the preexisting circuitry. By using a retrovirus-mediated labeling approach, Gage and colleagues showed that new neurons born in the adult hippocampus became morphologically indistinguishable from mature granule neurons within 4–8 weeks. More important, the newborn neurons fired action potentials in response to perforant path stimulation. Detailed electrophysiological characterization of new neurons born in the adult SGZ showed that these neurons adopt a developmental process that is similar to that in neurons born during embryonic and postnatal stages. In the olfactory bulb, newly generated periglomerular and granule neurons are also similar to their preexisting counterparts based on their morphological and electrophysiological properties. Interestingly, new neurons in the dentate gyrus have been shown to have a lower threshold for the induction of long-term potentiation at their immature stage, suggesting that the function of adult NSCs might depend on the unique plasticity of newborn neurons before they reach maturation.

Because the hippocampus is required for certain forms of learning and memory, the putative function of adult neurogenesis in the SGZ has been studied extensively in recent years. The level of adult neurogenesis has been well correlated with certain forms of learning and memory. For example, voluntary exercise and enriched environment were both shown to significantly enhance adult neurogenesis, and rodents exposed to these conditions could perform better in a hippocampus-dependent spatial learning task. Similarly, in the olfactory bulb, the survival of new neurons is enhanced by enriched odor exposure.

In addition to the correlative observations, several studies have reported a potential causal relationship between neurogenesis and different forms of learning and memory. Systemic antimitotic treatment with the DNA methylation agent MAM could reduce the level of adult neurogenesis, and the treated animals displayed defects in a trace conditioned eye-blinking task. Low-dose irradiation has also been used to inhibit adult neurogenesis, and treated rats showed impairment in long-term spatial memory and in a nonspatial nonmatching-to-sample task. Furthermore, adult neurogenesis in the SGZ appears to be required for the behavior effects of antidepressant treatment. Localized X-ray irradiation in the hippocampal area significantly reduced cell proliferation in the SGZ, and the irradiated mice failed to respond to the fluoxetine treatment in an anxiety-related, novelty-suppressed feeding test. These studies await corroboration from more spatial- and temporal-specific ablation of adult neurogenesis. In addition, conflicting results have been reported on certain learning paradigms such as the context fear conditioning test. Nevertheless, these are the first results indicating that neurogenesis in the adult brain might be required for hippocampus-dependent learning and memory.

Computational modeling has also been used to predict the function of adult neurogenesis in learning and memory. Among other putative functions in memory formation, an interesting hypothesis is that the continuous neurogenesis in the dentate gyrus might contribute to the formation of time-associated memories.

Conclusion

The discovery of continuous neurogenesis in adult rodent brains was embraced with skepticism for many years. It has now become clear that this phenomenon is conserved in all mammals examined, including humans. The facts that neurons generated in the adult brain functionally integrate into the existing circuitry and that adult neurogenesis is regulated by physiological and pathological conditions suggest that adult neurogenesis has survived evolution to serve some specific functions. More interesting, although NSCs have been purified from both SVZ and SGZ of the human brain, neuronal differentiation has been found in the hippocampus only but not the SVZ in humans. This appears to correlate with the functional differences between the two structures. Compared to rodents, humans rely much less on olfaction for daily activities and probably rely more on the hippocampus for emotions, learning, and memory. With the current advances in the field of adult neurogenesis, it is hoped that in the near future we will understand how new neurons born in the adult brain contribute to the normal brain functions and why the specific functions of adult neurogenesis cannot be fulfilled with the preexisting neuronal population.

The discovery of adult NSCs has triggered new therapeutic studies to treat diseases related to the

central nervous system. Such approaches include grafting NSCs to the site of injury or degeneration and targeting of endogenous stem cells. Indeed, it has been reported that neurogenesis could be induced *in vivo* in some of the nonneurogenic regions. The identification of small molecules that can induce neurogenesis from endogenous NSCs in damaged areas of the adult nervous system will change how we currently view central nervous system therapies.

See also: Neural Stem Cells and CNS Diseases; Neural Stem Cells: Ocular; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Aimone JB, Wiles J, and Gage FH (2006) Potential role for adult neurogenesis in the encoding of time in new memories. *Nature Neuroscience* 9: 723–727.

Altman J and Das GD (1965) Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *Journal of Comparative Neurology* 124: 319–335.

Consiglio A, Gritti A, Dolcetta D, et al. (2004) Robust *in vivo* gene transfer into adult mammalian neural stem cells by lentiviral vectors. *Proceedings of the National Academy of Sciences of the United States of America* 101: 14835–14840.

Doetsch F, Caille I, Lim DA, Garcia-Verdugo JM, and Alvarez-Buylla A (1999) Subventricular zone astrocytes are neural stem cells in the adult mammalian brain. *Cell* 97: 703–716.

Eriksson PS, Perfilieva E, Bjork-Eriksson T, et al. (1998) Neurogenesis in the adult human hippocampus. *Nature Medicine* 4: 1313–1317.

Gage FH (2000) Mammalian neural stem cells. *Science* 287: 1433–1438.

Kempermann G (2006) *Adult Neurogenesis: Stem Cells and Neuronal Development in the Adult Brain*. New York: Oxford University Press.

Kempermann G, Jessberger S, Steiner B, and Kronenberg G (2004) Milestones of neuronal development in the adult hippocampus. *Trends in Neuroscience* 27: 447–452.

Leuner B, Gould E, and Shors TJ (2006) Is there a link between adult neurogenesis and learning? *Hippocampus* 16: 216–224.

Lie DC, Song H, Colamarino SA, Ming GL, and Gage FH (2004) Neurogenesis in the adult brain: New strategies for central nervous system diseases. *Annual Review of Pharmacology and Toxicology* 44: 399–421.

Ming GL and Song H (2005) Adult neurogenesis in the mammalian central nervous system. *Annual Review of Neuroscience* 28: 223–250.

Morshead CM, Reynolds BA, Craig CG, et al. (1994) Neural stem cells in the adult mammalian forebrain: A relatively quiescent subpopulation of subependymal cells. *Neuron* 13: 1071–1082.

Palmer TD, Takahashi J, and Gage FH (1997) The adult rat hippocampus contains primordial neural stem cells. *Molecular and Cellular Neuroscience* 8: 389–404.

Reynolds BA and Weiss S (1992) Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science* 255: 1707–1710.

van Praag H, Schinder AF, Christie BR, et al. (2002) Functional neurogenesis in the adult hippocampus. *Nature* 415: 1030–1034.

Neural Stem Cells: Ocular

T A Reh, University of Washington School of Medicine, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Most of the neurons in the mature central nervous system (CNS) are generated during the embryonic or neonatal periods of development. In a few 'neurogenic' regions of the mature CNS, new neurons continue to be generated throughout life. The subventricular zone and hippocampal progenitor zone are the two neurogenic regions of the mammalian brain that have been best characterized, whereas in nonmammalian vertebrates, other neurogenic regions include the HVC of the songbird and, most important for this article, the ciliary margin of the retina. In all these regions, it is thought that a region-specific 'neural stem cell' maintains the progenitors that produce the new neurons and glia. Several lines of evidence indicate that a 'retinal stem cell' exists throughout life in nonmammalian vertebrates. Evidence from both rodents and primates, including humans, indicates that multipotential, mitotically competent cells exist at the retinal margin and can be activated *in vitro* to generate neuronal cells. This article discusses the literature on retinal development and the developmental origins of the retinal stem cells in nonmammalian vertebrates, describes their function in retinal growth and regeneration, and details efforts to identify and manipulate the remnants of this zone in mammals. Finally, common features among neural stem cell zones throughout the mature nervous system are discussed.

Retinal Progenitors in Normal Development

Each retina is initially derived from one of two evaginations of the ventral diencephalon of the neural tube, known as optic vesicles. Soon after they form, the optic vesicles undergo a further morphological transformation to form the optic cups. The cells of the optic cup resemble neural progenitors from other regions of the CNS in that they have a simple bipolar morphology and span the width of the neuroepithelium. Clonal analysis of the progeny of the mitotically active progenitor cells shows that they can give rise to all the different types of retinal neurons, and that the clones have mixed neuronal and glial lineages (Figure 1(a)). Birthdating studies have demonstrated that the different types of retinal neurons are generated by the progenitor cells in a sequence that is conserved among vertebrates. Ganglion cells, cone photoreceptors, amacrine cells, and horizontal cells are generated during early stages of development, and most rod photoreceptors, bipolar cells, and Müller glia are generated in the latter half of the period of retinogenesis.

Progenitor cells in the retina have gene expression profiles that are similar to those of progenitors from other regions of the CNS. They express many of the same homeodomain transcription factors. Several homeodomain transcription factors, such as Pax6, Chx10, Prox1, Sox2, and Six3, are expressed in retinal progenitors and are sometimes collectively called the eye field transcription factors (EFTFs) since their early expression defines the presumptive eye-forming region of the neural tube. However, these genes are also expressed in progenitors in other anterior regions of the CNS. The retinal progenitors also express some unique transcription homeodomain transcription factors, such as Rx and Crx. A similar pattern is observed for proneural transcription factors, such as Mash1, Neurogenin2, and Math3; these genes are expressed in progenitors from retina as well as other regions of the CNS. In addition, retinal progenitors express at least one unique member of this family of transcription factors, Math5. Presumably, it is through unique combinations of the common CNS factors and the retinal-specific factors that the unique retinal cell types, such as photoreceptors and retinal ganglion cells, are specified.

In addition to the complex intrinsic factors regulating retinal progenitors, these cells are also influenced by extrinsic signaling factors. Most of these factors have similar effects on the retinal progenitors as they would elsewhere in the CNS. For example, epidermal growth factor (EGF) was first described as a mitogen for retinal progenitor cells prior to its identification as a mitogen for brain-derived neural stem cells. Fibroblast growth factors (FGFs) and sonic hedgehog are additional mitogens for retinal progenitor cells that have also been demonstrated to have mitogenic effects in other regions of the CNS. Negative regulators of proliferation are also present in the developing retina to control the extent and duration of neurogenesis. Transforming growth factor-β2 is the most well-described inhibitor of proliferation in the retina, as it is in other regions of the CNS, and it likely functions through the regulation of the cell cyclin-dependent kinase inhibitor p27kip.

Another key regulator of progenitor cells in the retina is the Notch pathway. At least two Notch

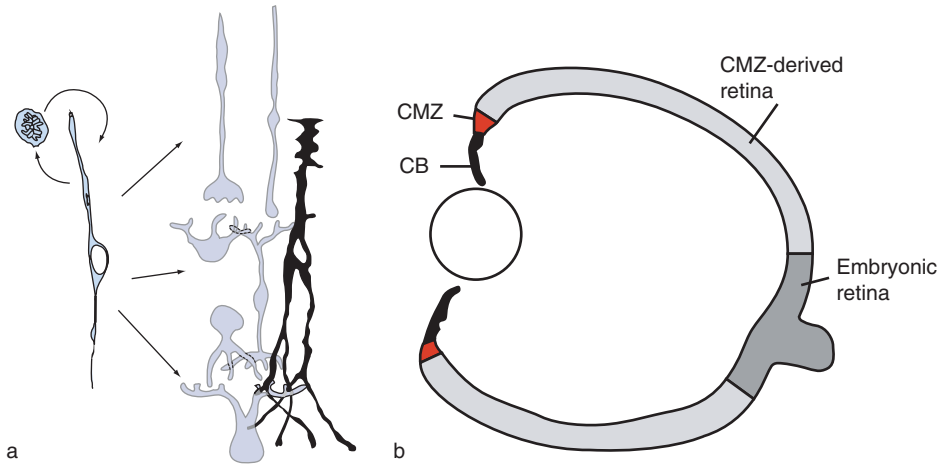

Figure 1 (a) Diagram of the multipotent progenitor that generates neurons (pink) and Muller glia (black) during embryonic development. (b) These cells persist at the retinal margin, in the 'ciliary marginal zone' (CMZ (red)) and are then called retinal stem cells, since they can persist the lifetime of the animal. The CMZ is located between the ciliary epithelium of the ciliary body (CB). The relative contributions of the retina derived from the embryonic progenitors (dark gray) and the CMZ (light gray) in a frog is also shown in this figure.

receptors and several Notch ligands are expressed by the retinal progenitor cells. Loss-of-function and gain-of-function studies have shown that Notch activity regulates progenitor differentiation. If the Notch receptor is activated in progenitors, they remain as progenitors, whereas inactivation of the Notch receptor occurs just prior to differentiation into one of the retinal neuronal types. The downstream effectors of the Notch pathway are the Hes genes. Hes1 and Hes5 are both expressed by subsets of the mitotically active cells in the developing retina, and both respond to Notch activation. Knockout and overexpression studies of the Hes genes indicate that they are necessary for both maintenance of the undifferentiated state of the progenitors and the development of the Müller glia.

In summary, the mitotically active cells in the developing retina have many characteristics of neural/glial progenitors in many other regions of the CNS. They share many of the same transcription factors, signaling molecules, and cell biology. In the retina, these cells have been called multipotent progenitors because at early stages of retinal development, these cells are competent to generate the entire complement of retinal neurons and glia. Thus, these cells have many characteristics of neural stem cells that have been isolated from other regions of the CNS. In addition, several groups have shown that the early progenitors can be cultured as 'neurospheres' (i.e., nonadherent aggregate cultures of neural stem and progenitor cells that can be passaged in EGF and/or FGF-containing medium). As described later, the adult retina of some vertebrates continues to add new neurons and glia at the peripheral margin, and thus true retinal stem cells exist. Presumably, these cells were derived from a population of similar cells in the developing retina, but there is currently no definitive way to distinguish the stem cells from the progenitors during retinogenesis.

Retinal Stem Cells in the Ciliary Marginal Zone

In the mammalian retina, the neurons and glia are generated in a continuous period of time from fetal to neonatal stages. In the mouse, for example, retinal neurons and glia become postmitotic starting at embryonic day 12.5 and ending approximately 1 week after birth. In all vertebrates, there is a central-to-peripheral pattern to histogenesis and differentiation in the retina. The first neurons are produced in the central retina and the last are produced at the periphery.

In contrast to the mammalian retina, in the amphibian, fish, or avian retina, histogenesis continues after the embryonic or neonatal period. In these animals, the retina continues to add new neurons into adulthood and, in some, throughout life. Teleost fish, for example, have a dramatic growth of the eye during their lifetime of up to 100-fold. The growth of the retina is accomplished by the continued production of new retinal neurons from a zone of cells at the peripheral margin, adjacent to the non-neuronal ciliary body, called the ciliary marginal zone (CMZ; Figure 1(b)). This zone is similar to neural stem cell zones in other regions of the CNS, such as the hippocampal progenitor zone or the subventricular zone. However, in contrast to these other zones, which typically generate only one type of neuron, the CMZ cells generate all types of retinal neurons

and Müller glia. In fact, most of cells in the retina of the mature frog or fish are generated by the cells of the CMZ. Lineage tracing studies of CMZ cells have shown that these cells can give rise to clones that contain all types of retinal neurons. The CMZ is thought to be organized with the most primitive cells (true stem cells (?)) in the most peripheral part (possibly mixed with the cells of the ciliary epithelium) and progressively more mature progenitor cells (transit amplifying cells (?)) closer to the differentiated retina.

The CMZ is highly productive in fish and some amphibians, but in birds it is greatly reduced. In birds, most of the retina is generated during embryonic development and only a small number of retinal neurons are generated by the CMZ. It is not known whether this zone persists throughout the lifetime of the bird, but new retinal neurons are generated at the peripheral edge of the retina in chickens up to 1 month of age and in the quail eye for up to 1 year. Like those of fish and amphibians, the CMZ cells of birds resemble the early progenitor cells of the retina, in terms of both their gene expression and their response to mitogenic factors.

The CMZ is greatly reduced or absent in the eyes of mammals that have been examined to date, including rodents and nonhuman primates. However, there is evidence that a CMZ-like zone can form in rodents under certain conditions. For example, mice with a single functional allele of the patched gene, a negative regulator of Shh signaling, retain a small number of proliferating, nestin-expressing cells at the retinal margin into adulthood. Moreover, when these mice are bred onto a background in which photoreceptors degenerate, the proliferation increases, reminiscent of the response to retinal damage observed in the CMZ cells of lower vertebrates. Studies have found that the normal period of proliferation of retinal progenitors can be extended by the injection of specific growth factors, but no study has successfully 'resuscitated' cells at the retinal margin of an adult mammal *in vivo*.

The CMZ in fish, amphibians, and birds can provide a source of regeneration following damage to the retina. In fish and amphibians, destruction of retinal neurons with neurotoxins or by surgery causes an increase in the proliferation of the CMZ cells. The cells produced by the CMZ then differentiate into neurons and in some cases migrate considerable distances to replace the cells lost by the experimental injury. In birds, neurotoxin damage does not by itself stimulate the proliferation of the CMZ cells, but destruction of retinal ganglion cells, followed by insulin-like growth factor and FGF injections, causes the CMZ cells to produce new ganglion cells, replacing the lost ganglion cells in the peripheral retina.

Additional Sources of Retinal Stem and Progenitor Cells

The Pigmented Epithelium

One of the most striking examples of regeneration in vertebrates is that of the amphibian retina. Following complete removal of the neural retina in larval frogs and adult urodeles (newts and salamanders), the adjacent pigmented epithelial (RPE) cells reenter the cell cycle, lose their pigmentation, and begin to express markers of retinal progenitors. The de-differentiated RPE cells go on to generate new retinal neurons in a recapitulation of normal retinal histogenesis. During a period of just a few weeks, histogenesis is complete and new retinal ganglion cells reconnect with visual centers in the brain. This process of RPE transdifferentiation into retinal progenitors does not occur in fish, adult birds, or adult mammals. However, a similar process occurs in birds and mammals at very early stages of eye development in embryos. The molecular mechanisms underlying the transition of the pigmented epithelial cells to retinal progenitors are only beginning to be understood. The de-differentiating pigment cells begin to resemble retinal progenitors soon after the retinal injury and express many of the EFTFs. FGF is a key stimulus for retinal regeneration from the RPE cells in both amphibians and chick embryos. When FGF is added to cultures of RPE cells or implanted into the vitreous *in ovo*, the RPE cells adopt a retinal progenitor identity and new, laminated retina is generated. Additional factors, such as activin and Shh, antagonize the transdifferentiation of RPE into regenerated retina, and blocking these factors enhances the retinal regeneration from the RPE that is induced by FGF.

The Ciliary Epithelium

The ciliary body is also considered a possible source of retinal stem cells or progenitors. The ciliary body is composed of a neural tube-derived part, called the ciliary epithelium, as well as a neural crest-derived part. The ciliary epithelium has two layers (**Figure 1**) – a pigmented layer that is continuous with the RPE and a nonpigmented layer that is continuous with the neural retina. The ciliary epithelium is anatomically and developmentally analogous to the choroid plexus in the rest of the CNS. The nonpigmented layer of the ciliary epithelium can generate neurons under certain conditions *in vivo* in birds. Intraocular injection of growth factors (insulin, FGF2, and EGF) stimulates the proliferation and neuronal differentiation of a subset of ciliary epithelial cells. Like the CMZ, the cells of the nonpigmented ciliary epithelium also express at least some of the EFTFs. The neurons

generated in this region following intraocular growth factor injections resemble amacrine cells and ganglion cells but not bipolar cells or photoreceptors. Thus, these cells may not actually be true multipotent progenitors but, rather, retain only some of their characteristics.

The mammalian ciliary epithelium has also received attention as a potential source of new neurons in adult animals. Several groups have found that cells can be isolated from the rodent ciliary epithelium that display characteristics of neural stem cells isolated from the mature brain. Although the cells appear initially pigmented, suggesting they are resident in the pigmented layer of the ciliary epithelium, they give rise to spheres of both pigmented and nonpigmented cells. Some of the spheres derived from the ciliary epithelium can be passaged to form new spheres, and although they appear to have a more limited life *in vitro* than neurospheres derived from the brain, they have been labeled retinal stem cells by the authors of these studies. The spheres are also capable of differentiation into cells that express markers of retinal neurons, including the rod photoreceptor-specific protein, rhodopsin. Human eyes also contain these cells, and they can be grown *in vitro* for extended periods of time and transplanted. However, the cells derived from the spheres typically express markers of specific neuronal types without taking on the morphological characteristics of the neurons. This is particularly clear in the case of photoreceptors, which have a distinct morphology when derived from retinal progenitors during normal development but appear flat and glial-like when derived from ciliary epithelial-derived cells. Thus, it is unclear whether the situation for these cells is more like what was previously described for the bird ciliary epithelium – that is, the cells take on some, but not all, of the characteristics of retinal progenitors. The nonpigmented epithelial cells from mammalian eyes can also be maintained as dissociated cell cultures grown on adherent substrates. Under these conditions, the cells are also capable of expressing neuronal markers, but again only some of the normal markers of retinal progenitors are expressed.

Intrinsic Stem Cells, Rod Precursors, and Müller Glia

In fish, there are several sources of new retinal neurons within the differentiated retina. In teleosts, new rod photoreceptors are added throughout the retina as it grows. The addition of new rods to the circuitry allows the fish to maintain a constant rod density as the retina stretches with ocular growth. Thus, sensitivity to light is maintained with increasing size. To accomplish this feat, the fish has a specific progenitor, called the rod precursor, which is normally restricted to generating new rods. These cells are found in association with the Müller glia and divide in the outer nuclear layer. In addition to the rod precursors, fish also have a relatively quiescent stem cell in the inner nuclear layer. Like the cells of the ciliary epithelium and those of the CMZ, this cell has also been labeled the retinal stem cell. The role of this slow-cycling retinal stem cell in normal retina is not clear, but when a small patch of central retina is surgically removed, both the rod precursor and the intrinsic retinal stem cell proliferate and form a blastema. The blastemal cells also resemble retinal progenitors, and over a few weeks they fill in the excised patch of retina.

The final source of retinal progenitors that have potential to replace lost neurons from within the retina are the Müller glia. Müller glia are the only glial cells in the retina produced by the multipotent progenitor. Müller glial cells are among the last cell types produced by the retinal progenitors during the normal period of histogenesis, and gene expression profiling studies have shown a high degree of overlap in the genes expressed by Müller glial cells and retinal progenitors. Despite their similarity, Müller glial cells in mammals and birds do not normally express many of the critical transcription factors that define progenitors in the retina and elsewhere in the CNS, such as Neurogenin2, Pax6, or Mash1. However, in posthatch chickens, damage to the retina by neurotoxins causes some of the Müller glia to reenter the cell cycle and reexpress both proneural genes, such as Mash1, and EFTFs, such as Pax6 and Chx10. In addition, after neurotoxic damage to the retina, some of the proliferating Müller glia go on to generate cells that express markers and morphology of neurons. The regenerative process initiated by retinal damage in young birds is largely abortive, however, and the majority of the Müller glial progeny remain as undifferentiated cells.

A similar process of regeneration from Müller glial cells may occur in teleost fish. As noted previously, regeneration of the teleost retina is particularly robust, and both rod precursors and intrinsic retinal stem cells are thought to contribute to the process. However, lesions to the retina of mature zebra fish also lead to Müller glial proliferation, and these cells may also contribute to the process, as in the chick. Attempts to stimulate regeneration from Müller cells in mammalian retina are under way and initial findings are promising, although there are still major hurdles to overcome.

See also: Neural Stem Cells and CNS Diseases; Neural Stem Cells: Adult Neurogenesis; Retinal Pharmacology: Inner Retinal Layers; Retinal Development: An Overview;

Retinal Glia; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Coles BL, Angenieux B, Inoue T, et al. (2004) Facile isolation and the characterization of human retinal stem cells. *Proceedings of the National Academy of Sciences of the United States of America* 101: 15772–15777.

Fischer AJ and Reh TA (2000) Identification of a proliferating marginal zone of retinal progenitors in postnatal chickens. *Developmental Biology* 220: 197–210.

Fischer AJ and Reh TA (2003) Potential of Müller glia to become neurogenic retinal progenitor cells. *Glia* 43(1): 70–76.

Haruta M, Kosaka M, Kanegae Y, et al. (2001) Induction of photoreceptor-specific phenotypes in adult mammalian iris tissue. *Nature Neuroscience* 4: 1163–1164.

Moshiri A, Close J, and Reh TA (2004) Retinal stem cells and regeneration. *International Journal of Developmental Biology* 48(8–9): 1003–1014.

Ooto S, Akagi T, Kageyama R, et al. (2004) Potential for neural regeneration after neurotoxic injury in the adult mammalian retina. *Proceedings of the National Academy of Sciences of the United States of America* 101: 13654–13659.

Otteson DC and Hitchcock PF (2003) Stem cells in the teleost retina: Persistent neurogenesis and injury-induced regeneration. *Vision Research* 43(8): 927–936.

Perron M, Kanekar S, Vetter ML, and Harris WA (1998) The genetic sequence of retinal development in the ciliary margin of the *Xenopus* eye. *Developmental Biology* 199: 185–200.

Reh TA and Levine EM (1998) Multipotential stem cells and progenitors in the vertebrate retina. *Journal of Neurobiology* 36(2): 206–220.

Spence JR, Madhavan M, Ewing JD, Jones DK, Lehman BM, and Del Rio-Tsonis K (2004) The hedgehog pathway is a modulator of retina regeneration. *Development* 131(18): 4607–4621.

Tropepe V, Coles BL, Chiasson BJ, et al. (2000) Retinal stem cells in the adult mammalian eye. *Science* 287: 2032–2036.

Wang JC and Harris WA (2005) The role of combinational coding by homeodomain and bHLH transcription factors in retinal cell fate specification. *Developmental Biology* 285(1): 101–115.

Wehman AM, Staub W, Meyers JR, Raymond PA, and Baier H (2005) Genetic dissection of the zebrafish retinal stem-cell compartment. *Developmental Biology* 281(1): 53–65.

Neural Synchrony and Feature Binding

W Singer, Max Planck Institute for Brain Research, Frankfurt/Main, Germany

© 2009 Elsevier Ltd. All rights reserved.

Linguistic and mathematical descriptions consist of symbols for objects, qualities, and relations. Through rule-based recombination of these symbols, a virtually infinite number of different descriptions can be generated with a relatively small set of symbols. The efficiency of this versatile strategy is further enhanced by chunking. In order to reduce the length of descriptions, frequently occurring constellations are represented by symbols of higher order that summarize sets of component features and their respective relations in a single term. This chunking strategy requires a larger number of different symbols, but it saves time because it reduces the length of descriptions. There is thus a trade-off between required storage capacity and processing speed. It appears as if the brain uses a similar strategy to represent perceptual objects and motor programs and that it possesses learning mechanisms to optimize the trade-off between storage requirement and processing speed. Perceptual objects can be represented by scripts that consist of symbols for their components and their respective qualities and symbols describing the relations between these items. The same is true for movements which can be decomposed into elementary motion components and a specific set of relations that define the respective combination and sequence of components.

Analyses of the response properties of neurons encountered at the various levels of the processing hierarchy of sensory systems suggest the following sequence of processing steps. Neurons in the sensory organs encode in their responses very elementary, local properties and qualities of the components of perceptual objects and barely any relations. An exception is the retina of the eye. Unlike most other sensory organs, it possesses a complex, multilayered network that permits, first, recombination of signals in convergent feed-forward architectures and, second, extensive lateral interactions that modulate responses in a context-dependent way. Thus, the rate modulated output of ganglion cells signals not only the presence of a local property of an object, in this case the brightness and spectral composition of a small part of its surface, but also the neighborhood relations of these particular features. However, chunking at larger scales occurs only once signals are processed at the cortical level. As one proceeds along the hierarchically arranged cortical processing areas, one encounters neurons that respond selectively to increasingly complex constellations of elementary features. It is commonly held that this chunking is achieved by iterative recombination of feed-forward connections from lower to higher order neurons. If the thresholds of the respective higher order neurons are adjusted such that they respond only if the full set of the feeding neurons is simultaneously active, they assume the function of conjunction detectors. Their responses signal not only the presence of certain sets of component features but also the way in which these features are related to each other. The latter information is implicitly encoded in the architecture of feed-forward connections that link selected sets of feeder neurons to higher order conjunction detectors.

Several theoretical arguments suggest that these chunking operations are complemented at each level of processing by additional mechanisms that permit flexible definition of relations among the responses of distributed neurons. Natural scenes, for example, usually contain a large number of different objects, the contours of which may be overlapping or partially occluded. Hence, a single border may be shared by several objects. This introduces ambiguities that need to be resolved in order to provide appropriately sorted signals to the feed-forward chunking circuits. Without such prior sorting it would be difficult to avoid accidental formation of false conjunctions. Thus, responses to contours belonging to the same object need to be grouped together for further joint processing and chunking, and they need to be segregated from responses to other objects and the embedding background. This selection of 'chunkable' responses has to occur in a context-dependent way and hence needs to be based on an evaluation of neighborhood relations. Responses evoked by coherent contours need then be tagged as related in a way that ensures their selective binding by subsequent chunking. These grouping operations must occur at early levels of the processing hierarchy because successful scene segmentation is a prerequisite for the later identification of individual objects. However, context-sensitive dynamic definition of relations is also required at the highest levels of processing where neurons are encountered that respond to very complex constellations of features, such as the various components of a face, the mouth, the eyes, or the nose. The main argument for the need to flexibly define relations also at these high levels of processing derives from the evidence that perceptual objects are represented not only by individual highly complex chunking neurons but also by distributed assemblies of cells. First, chunking neurons whose response properties are sufficiently complex and selective to encode

only a single perceptual object are rare and seem to exist only for highly overlearned objects or objects of particular behavioral relevance. Second, it is inconceivable that novel objects can be represented by preestablished chunking neurons because the required feed-forward architectures would have to be specified *a priori* to support formation of the appropriate chunks. Third, objects that are simultaneously encoded in different sensory modalities elicit responses in several different sensory systems, and these need to be related to each other in order to arrive at a comprehensive polymodal description of this object. These considerations suggest that objects not representable by individual chunking neurons are encoded by assemblies of distributed neurons, each of which represents only a particular component of the object. In assembly coding, however, a relation-defining mechanism is again required that tags responses as related that are evoked by the components of the same object. The reason is that assemblies, just as the previously mentioned scripts, consist of symbols representing particular features – here neurons tuned to particular components and qualities of perceptual objects – and relation-defining codes that indicate which symbols have actually been recruited into the description of a particular perceptual object.

Synchrony as Tag of Relatedness

Psychophysical and electrophysiological evidence indicates that attentional mechanisms play an important role in grouping operations both at the level where scene segmentation is accomplished as well as at the higher levels where object identification is thought to occur. However, the mechanisms underlying these relation-defining grouping operations are still poorly understood. One proposal is that attentional mechanisms modulate the discharge rate of neurons and enhance the amplitude of responses to attended features. In this scenario, the signature of relatedness is the concomitant enhancement of discharge frequency. Responses selected for chunking or the formation of an assembly would be distinguished from all others by their higher discharge frequency. This interpretation has been challenged by the argument that response amplitude may be an ambiguous signature of relatedness because it depends on too many other stimulus-related variables, requires long read-out times, and makes it difficult to segregate assemblies from one another that are simultaneously configurated within the same neuronal network. Therefore, it has been proposed that neurons exploit two independent coding strategies in order to convey two messages in parallel. First, they should signal that the feature or the chunk of features for which they code is present, and second, they should indicate with which of the other simultaneously active neurons their responses are related. This latter code should ensure that responses tagged as related are processed jointly at subsequent stages – that is, are routed together into the appropriate chunking channels and/or are recognizable without ambiguity as originating from cells of the same assembly. It is commonly held that the first message is encoded in the discharge rate of the neurons because at lower levels of processing discharge rate reflects reliably different physical aspects of elementary stimuli and at higher levels the presence of complex chunks. Following the discovery that neurons in the primary visual cortex can synchronize their spike discharges with a precision in the millisecond range, it has been proposed that the synchronization of responses could serve as the required tag of relatedness. One way to select subsets of responses for further joint processing, and thus for binding them together, is to selectively raise their saliency. Most models on the binding function of selective attention are based on such a mechanism, but they usually assume that saliency is enhanced by rate increases. However, precise temporal synchronization of spike discharges is an equally efficient means to raise selectively the saliency of neuronal responses. The reason is that synchronized input to target neurons has a much stronger impact than temporally uncoordinated input. Simultaneously arriving excitatory postsynaptic potentials (EPSPs) summate much more effectively than temporally dispersed EPSPs, and this coincidence sensitivity of neurons is further augmented in cortical neurons by a number of specific mechanisms: (1) active dendritic conductances that amplify fast-rising depolarizations of large amplitude; (2) the frequency adaptation of synaptic release and postsynaptic receptors which attenuates temporal summation of EPSPs; and (3) a dependence of firing threshold on the rising slope of depolarizations, favoring responses to fast-rising depolarizations. These mechanisms increase selectively the impact of synchronous inputs, and they do so with a temporal resolution in the millisecond range. Thus, relations can be defined within narrow temporal windows (<10 ms), and hence different relations can be encoded with less ambiguity and in much more rapid alternation than if relations were expressed by joint rate increases.

The proposal that precise temporal synchrony is used as a tag of relatedness in neuronal processing agrees well with the temporal sensitivity of mechanisms supporting synaptic plasticity and Hebbian learning. Known mechanisms of synaptic plasticity exploit temporal correlations among the discharges of input connections and/or the discharges of inputs

and those of the postsynaptic target cells. The temporal resolution of the mechanism that classifies discharges as synchronous (asynchronous) – that is, related (unrelated) – and causes synapses to strengthen (weaken) also operates with a precision in the millisecond range. Thus, there is a perfect match between the signatures of relatedness used in signal processing and Hebbian learning. This cannot be otherwise because both processes have to rely on the same relation-defining code to avoid learning of false conjunctions.

The Role of Oscillations and Spike Synchronization

Investigations of response synchronization in the visual system have revealed that precise synchronization of discharges is often associated with an oscillatory patterning of the neuronal responses. Because individual cells tend to skip cycles of these oscillations, they are rarely detectable in the spike trains of single cells but they are readily seen in data representing the responses of large populations of neurons (i.e., in multiunit recordings or recordings of local field potentials). *In vitro* experiments in cortical slices and simulation studies have in the meantime established causal relations between the two phenomena. The oscillatory patterning of the responses is mainly due to oscillations generated within the various pools of inhibitory interneurons that are coupled through both chemical and electrical synapses and capable of sustaining oscillatory activity patterns. These oscillatory inhibitory inputs to pyramidal cells veto their discharges during the inhibitory troughs and favor discharges at the depolarizing peaks, thus causing synchrony in firing. These locally synchronized oscillatory responses can become synchronized over large distances due to reciprocal coupling of the oscillatory networks via excitatory cortico-cortical connections. It follows from this mechanism that the precision with which spikes can be synchronized increases with oscillation frequency. A relation also exists between oscillation frequency and the distance over which synchronization is maintained. Synchronization among remote groups of neurons or among large assemblies of neurons tends to occur at lower oscillation frequencies than synchronization of local clusters of cells.

The Duration of Synchronized Events

Early studies were based mostly on conventional cross-correlation analysis of cell discharges and/or local field potentials. This method reliably detects synchronous firing if it is sustained over prolonged periods of time but it fails if synchronous events occur only a few times in a response. Therefore, more sensitive measures have been developed that allow assessment of brief events of coincident firing. One of these methods, the unitary event analysis, uses statistical methods to identify single, nonaccidental incidences of coincident firing, whereas the other evaluates consistent phase relations between the discharges of individual neurons and local field potential (LFP) oscillations (spike-field coherence). Application of these methods to data obtained from awake behaving animals has revealed that episodes of synchronized firing are often restricted to short epochs of particular behavioral sequences and may be as short as a few tens of milliseconds. This agrees with measurements of the minimal time required to segment scenes and identify objects. It was estimated that the grouping operations required for scene segmentation and object identification should not take more than 10 to 20 ms per processing stage. This implies that a substantial amount of information about the accomplished grouping must be encoded in the precise timing relations between individual discharges of distributed neurons. The reason is that not much information can be encoded in variations of discharge rates of individual cells because they can generate only few spikes within such short time windows.

Synchrony and Feature Binding

Evidence from studies in the visual system suggests that response synchronization may be used throughout all processing stages, from the retina to the highest cortical areas, in order to establish relations among distributed responses – that is, to bias grouping of responses for subsequent chunking and to tag responses of assembly members as related. In all cases, synchronization probability reflects some of the Gestalt criteria that are used for scene segmentation and perceptual grouping. In the retina, ganglion cell responses synchronize with millisecond precision if evoked by continuous contours or coherent objects, and there is evidence from studies on the escape response of frogs that synchronicity of ganglion cell firing is actually carrying behaviorally relevant information. Retinal synchronization is associated with high-frequency oscillations (up to 90 Hz) and based on horizontal interactions within the network of coupled amacrin cells.

In the primary visual cortex, synchrony is often associated, especially when it is observed over larger distances, with an oscillatory patterning of spike discharges in the gamma frequency range (30–60 Hz). At this processing stage, synchronization probability correlates well with elementary Gestalt rules. It is enhanced between responses evoked by continuous

contours, by contours moving with the same speed in the same direction, by colinearly aligned contour segments, and by contours belonging to the same surface. It is maximal among responses evoked by coherent patterns such as regular gratings, and it is minimal or absent among responses to incoherent stimuli such as random dot patterns. In the cortex, response synchronization among spatially distributed neurons is mediated by the network of tangential horizontal connections, and if it occurs across the midline of the visual field it is mediated by callosal connections. One of the reasons why synchrony is stronger among responses to continuous or colinearly aligned contours or contours moving in the same direction is the anatomical anisotropy of these tangential connections. Those spanning larger distances connect preferentially columns with similar feature preference. Thus, elementary grouping criteria are implemented in the anisotropies of the network of tangential connections and translated into synchronization probability. In the cat, such stimulus-specific synchronization phenomena have been observed both within and across different visual areas, both within and across hemispheres, and between the visual cortex and the superior colliculus. In primates, especially in the awake behaviorally trained animal, multisite recordings have been applied much less frequently and therefore less data are available on response synchronization. However, the results obtained from primary visual cortex closely resemble those obtained from cats, but oscillation frequencies tend to be higher and the distances over which synchrony is observed tend to be shorter. In the motion-sensitive area MT of the dorsal processing stream response, synchronization was found to reflect the Gestalt rule of common fate, which is one of the strongest binding cues for perceptual grouping. Presentation of two spatially overlapping bars moving in different directions led to the formation of two distinct assemblies of neurons whereby those responding to the same contour synchronized their discharges, whereas those responding to different contours did not. In the inferior temporal cortex of the ventral processing stream, synchronization probability reflected the binding of chunking neurons into assemblies representing individual objects. Neurons responding to the components of faces (eyes, nose, mouth, etc.) synchronized their responses when the arrangement of these components was such that the animals signaled having recognized a face while they did not synchronize when the components were scrambled or presented in a way that was judged by the animal as incompatible with the appearance of a normal face. In the case of neither MT nor IT was it possible to distinguish between the various arrangements of the presented stimuli if only the discharge rate of the neurons was evaluated. This is compatible with the interpretation that discharge rate signals the presence of particular features, whereas the correlations among the discharges of neurons indicate how these features are related to each other.

The Role of Attention

Grouping operations based on elementary Gestalt rules and the binding of the stereotyped feature constellations of highly familiar objects can occur preattentively. This automatic, attention-independent grouping is thought to be based on chunking in fixed feed-forward architectures. However, several arguments suggest that synchronization may also serve as a mechanism for automatic grouping. The synchronization of retinal responses cannot be influenced by attentional mechanisms because there are no efferent projections capable of conveying the required information. The fact that feature-specific response synchronization is readily observed in anesthetized preparations also suggests that binding through synchrony can occur preattentively. Interestingly, this may turn out to be a feature distinguishing automatic grouping by chunking or by synchronization. Automatic grouping by synchrony is highly context dependent, whereas chunking is not. Despite anesthesia, grouping by synchronization remains sensitive to the global configuration of stimuli. In cat areas 17 and 18, synchronization probability changes when variations in stimulus context require a change in grouping, and this reorganization of synchrony patterns occurs even if stimulus configurations are changed in a way that leaves the stimuli appearing within the aperture of the classical receptive fields of the recorded neurons unchanged.

In addition to this evidence for attention-independent grouping by synchrony, recent results clearly indicate that synchronization is also highly susceptible to top-down, attention-dependent influences and that it plays an important role in attention-dependent response selection and binding. Various measures have been used to assess the influence of selective attention on neuronal synchrony: correlations among spike discharges, spike-field coherence, correlations in phase locking between oscillatory field potentials, and, finally, the amplitude and the phase locking of oscillatory responses in magnetoencephalography (MEG) and electroencephalogram (EEG) recordings. Because the amplitude of these latter signals depends to a crucial extent on the synchronicity of large populations of neurons, not only variations in phase locking but also variations in the power of oscillations can be taken as a measure of synchrony. These data indicate that focusing

attention on a particular stimulus or to a particular modality increases the synchrony of responses in the neuronal networks that process the attended stimulus. Again, this enhanced synchronization is associated with and most likely caused by an oscillatory patterning of neuronal activity in the beta and especially in the gamma frequency range. At the same time, one observes a reduction of oscillatory activity in lower frequency bands (alpha and delta). Evidence also indicates that anticipation of a particular stimulus or a motor act is associated with the generation of oscillatory activity in the beta and gamma frequency band in cortical areas required for the processing of the stimulus or the execution of the task. For tasks involving sensory discrimination and motor responses, this anticipatory synchronization can extend across widely distributed networks of cortical areas. This anticipatory modulation of oscillatory activity is usually not associated with major changes in the discharge activity of neurons, suggesting that it consists mainly of a synchronous pacing of excitability through oscillatory activity generated in the network of inhibitory interneurons. It has been proposed that this subthreshold modulation of excitability facilitates rapid synchronization of responses once stimuli are available, thereby enhancing transmission across multiple cortical stages. Recent results from MEG studies in human subjects take this proposal one step further and suggest that the anticipatory induction of coherent oscillations across distributed cortical areas and executive structures facilitates selective routing of activity and rapid handshaking among the involved processing stages. However, it is unknown which centers coordinate this attention-dependent modulation.

In principle, synchronization could be used as an alternative mechanism to rate modulations in order to raise the saliency of responses. Experiments on binocular rivalry support this conjecture. In cat primary visual cortex, the responses to the respective perceived stimulus differed from those to the suppressed stimulus because they were more synchronized and not because they were more vigorous. A similar conclusion is suggested by experiments on perceived brightness. If a small grating is superimposed on a large grating, the perceived contrast of the former increases with increasing orientation or phase offset between the two gratings. This effect is closely related to changes of neuronal responses in primary visual cortex. Neurons responding to the small grating increase their discharge rate but not their synchrony with increasing orientation offset, whereas they increase the synchrony of their discharges but not the rate with increasing phase offset. This indicates that the saliency of responses can be enhanced by increasing either the rate or the synchronicity of discharges. The fact that the effects are perceptually indistinguishable illustrates the complementarity of rate codes and temporal codes.

Oscillations and Read-Out

Self-generated oscillatory activity and the associated synchronization of spike discharges are likely to also play a role in the read-out of information stored in the architecture of neural networks. Data from multisite recordings and optical imaging have revealed that spontaneous activity is not simply noise but exhibits a high degree of spatial and temporal organization. In the visual cortex, the spontaneous activity fluctuations are coherent among columns sharing similar orientation preferences, probably because these columns are interconnected more strongly through cortico-cortical projections than are columns with dissimilar preferences. These spontaneous fluctuations have a strong impact on the latency and amplitude of light-evoked responses. Multisite recordings of spiking activity and field potentials from primary visual cortex revealed that columns preferring contours with similar orientation and in particular colinearly aligned contours engage in highly synchronized oscillations in the gamma frequency range when the cortex is in an activated state (i.e., when the EEG exhibits high power in the beta and gamma frequency range). The effect of these self-generated coherence patterns is that columns oscillating in synchrony respond with precisely synchronized latencies when activated by light stimuli, whereas response latencies fluctuate unsystematically and over a wide range for columns that have not been oscillating in synchrony prior to light stimulation. Therefore, the output of columns coding for features that tend to be grouped perceptually (same orientation, colinearity) is more synchronized than the output of columns coding for features that are less likely to be grouped. Thus, self-generated gamma oscillations translate the anisotropies in the network of horizontal connections into spatially selective patterns of coherence which in turn bias grouping by rapid synchronization of the very first components of responses to contours. Further support for this notion comes from multisite recordings in V1 of monkeys trained to freely inspect complex visual scenes. Soon after the onset of fixation (40–100 ms) one observes a brief burst of highly synchronized high-frequency oscillations in the local field potential that are precisely phase-locked across recording sites. These fixation-related oscillations are in turn associated with excess synchronization of spike discharges in the responses to the contours of the scene. Because these oscillations also occur when the animal scans a blank screen, they are most likely due to corollary activity that is generated in anticipation of having to

process new constellations of features once a new segment of the scene is fixated. In analogy to the effect of the spontaneous oscillations, this self-generated coherent activity could serve the read-out of grouping criteria residing in the network of tangential connection and to translate these criteria into specific synchronization patterns. The saccade-related oscillations and the associated spike synchronization precede by several tens of milliseconds the peak of the neurons' rate responses, which reach a maximum only approximately 100 ms after the eyes have come to rest. Grouping cues encoded in latency adjustments and spike synchronization are thus available long before the changes in the neurons' discharge rate can be fully evaluated. Such rapid processing at early stages of the visual system appears desirable given that the animals changed gaze direction on average four or five times in a second. This implies that scene segmentation, the eventual resolution of ambiguities, the selection of signals for chunking, and the subsequent dynamic grouping of chunks into object representations must have been accomplished within approximately 200 ms.

Conclusions

The data reviewed in this article suggest that sensory systems exploit two complementary ways to evaluate and represent relations between features of perceptual objects. One strategy consists of the generation of specialized neurons in feed-forward architectures that respond selectively to particular constellations of features. The discharge rate of these chunking neurons encodes both the presence of particular features and the way in which they are related to each other. This coding strategy is fast but can encode only relations defined *a priori* by the convergence patterns of the feed-forward connections. As multisite recordings suggest, there is a second strategy that exploits the precise temporal relations between the discharges of distributed neurons to encode relations. This mechanism permits flexible and context-dependent definition of relations. It exploits the coincidence sensitivity of neurons and uses precise temporal synchronization of discharges as tag of relatedness. Interestingly, the same tag appears to be exploited by the mechanisms mediating use-dependent synaptic plasticity and associative learning. As synchronization enhances the impact of the synchronized responses, it appears to be used not only to define relations among distributed responses but also in a more general way to select responses for further processing to raise their perceptual saliency and to support selective rooting of activity under the control of attentional mechanisms. Because temporal codes can only be assessed with multisite recordings and because these have a relatively short history, we are still just beginning to understand coding strategies based on the dynamic interactions among large numbers of neurons. It may turn out that precise synchronization is only one, albeit a very important, signature of the many potentially significant dynamical states. Precisely timed phase offsets and sequences of patterns defined by specific temporal relations are likely to play an equally important role. To analyze these more complex patterns and to examine whether they contain information that can be related to behavior is one of the great challenges in future systems neurobiology.

See also: Cerebral Cortex; Contextual Interactions in Visual Perception; Contextual Interactions in Visual Processing; Memory Consolidation: Cerebral Cortex; Neural Oscillators and Dymanmical Systems Models; Perception of Surfaces and Forms; Perceptual Learning: Neural Mechanisms.

Further Reading

Azouz R and Gray CM (2003) Adaptive coincidence detection and dynamic gain control in visual cortical neurons *in vivo*. *Neuron* 37: 513–523.

Castelo-Branco M, Goebel R, Neuenschwander S, et al. (2000) Neural synchrony correlates with surface segregation rules. *Nature* 405: 685–689.

Cook EP and Maunsell JHR (2002) Attentional modulation of behavioral performance and neuronal responses in middle temporal and ventral intraparietal areas of macaque monkey. *Journal of Neuroscience* 22(5): 1994–2004.

Engel AK, Fries P, and Singer W (2001) Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience* 2: 704–716.

Engel AK, König P, Kreiter AK, et al. (1991) Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science* 252: 1177–1179.

Engel AK, König P, and Singer W (1991) Direct physiological evidence for scene segmentation by temporal coding. *Proceedings of the National Academy of Sciences of the United States of America* 88: 9136–9140.

Fries P, Neuenschwander S, Engel AK, et al. (2001) Rapid feature selective neuronal synchronization through correlated latency shifting. *Nature Neuroscience* 4(2): 194–200.

Fries P, Reynolds JH, Rorie AE, et al. (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291: 1560–1563.

Fries P, Schröder J-H, Roelfsema PR, et al. (2002) Oscillatory neuronal synchronization in primary visual cortex as a correlate of stimulus selection. *Journal of Neuroscience* 22(9): 3739–3754.

Gray CM, König P, Engel AK, et al. (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338: 334–337.

Gray CM and Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 86: 1698–1702.

Kreiter AK and Singer W (1996) Stimulus-dependent synchronization of neuronal responses in the visual cortex of the

awake macaque monkey. *Journal of Neuroscience* 16(7): 2381–2396.

Löwel S and Singer W (1992) Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science* 255: 209–212.

Markram H, Lübke J, Frotscher M, et al. (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275: 213–215.

Murthy VN and Fetz EE (1996) Oscillatory activity in sensorimotor cortex of awake monkeys: Synchronization of local field potentials and relation to behavior. *Journal of Neurophysiology* 76(6): 3949–3967.

Neuenschwander S and Singer W (1996) Long-range synchronization of oscillatory light responses in the cat retina and lateral geniculate nucleus. *Nature* 379: 728–733.

Phillips WA and Singer W (1997) In search of common foundations for cortical computation. *Behavioral and Brain Sciences* 20: 657–722.

Roelfsema PR, Engel AK, König P, et al. (1997) Visuomotor integration is associated with zero time-lag synchronization among cortical areas. *Nature* 385: 157–161.

Salinas E and Sejnowski TJ (2001) Correlated neuronal activity and the flow of neural information. *Nature Reviews Neuroscience* 2(8): 539–550.

Schoffelen J-M, Oostenveld R, and Fries P (2005) Neuronal coherence as a mechanism of effective corticospinal interaction. *Science* 308: 111–113.

Singer W (1999) Neuronal synchrony: A versatile code for the definition of relations? *Neuron* 24: 49–65.

Thorpe S, Fize D, and Marlot C (1996) Speed of processing in the human visual system. *Nature* 381: 520–522.

Tsunoda K, Yamane Y, Nishizaki M, et al. (2001) Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience* 4(8): 832–838.

Volgushev M, Chistiakova M, and Singer W (1998) Modification of discharge patterns of neocortical neurons by induced oscillations of the membrane potential. *Neuroscience* 83(1): 15–25.

Wespatat V, Tennigkeit F, and Singer W (2004) Phase sensitivity of synaptic modifications in oscillating cells of rat visual cortex. *Journal of Neuroscience* 24(41): 9067–9075.

Whittington MA, Doheny HC, Traub RD, et al. (2001) Differential expression of synaptic and nonsynaptic mechanisms underlying stimulus-induced gamma oscillations *in vitro*. *Journal of Neuroscience* 21(5): 1727–1738.

Neurexins

M Ahmad, Georg-August University, Göttingen, Germany
M Missler, Georg-August University, Göttingen, and Westphalian Wilhelms University Munster, Munster, Germany

© 2009 Elsevier Ltd. All rights reserved.

Neurexins constitute a family of cell-surface proteins that are expressed in neurons and (neuro)endocrine cells. Their hallmark is a high degree of polymorphism with potentially thousands of isoforms. Neurexins are essential for synaptic transmission and are potent inducers of synapse formation *in vitro*.

Gene Structure

Neurexins are encoded by three genes in vertebrates (NRXN1, -2, and -3). Each gene contains two promoters: an upstream promoter that directs the transcription of the larger α-neurexins and a promoter in the middle of the gene that initiates the transcription of shorter β-neurexins. NRXN1 (Figure 1) and NRXN3 genes are among the largest of mammalian genes, with sizes of ~1.1 and 1.6 Mb, respectively. NRXN2 is smaller, with a size of 110 kb. The three neurexin genes have a similar exon–intron structure with an almost identical number of exons. The neurexin genes are conserved in vertebrates, being present in mammals, birds, and frogs. However, in contrast to vertebrates, *Caenorhabditis elegans* and *Drosophila melanogaster* have only one NRXN gene, which codes for an α-neurexin-like molecule. Thus, β-neurexins are not present within the invertebrates.

Protein Structure

α-Neurexins and β-neurexins are type I membrane proteins with a large extracellular region and a small intracellular C-terminal region (Figure 2). α-Neurexins contain much longer extracellular sequences than β-neurexins, but both share an identical transmembrane region and intracellular part. The extracellular region of α-neurexins has a multidomain structure, consisting of three overall repeats followed by a serine-threonine-rich sequence before the transmembrane domain, which is O-glycosylated. Each repeat contains a central epidermal growth factor (EGF)-like domain and two flanking laminin-neurexin-sex-hormone binding protein (LNS) domains. The extracellular region of β-neurexins has a single LNS domain preceded by a short β-specific sequence. The LNS domain was identified as a repeat in Laminin A and is named for Laminin, Neurexins, and Sex hormone-binding globulin, the three proteins in which it was initially discovered. Each LNS domain consists of 160–200 residues and forms a ligand-binding domain in many proteins. The crystal structure of the LNS domain of neurexin 1β (which is equivalent to the sixth LNS domain of neurexin 1α) revealed two 7-stranded β-sheets forming a jelly-roll fold with structural similarity to lectins. The two β-sheets are roughly parallel and curved, giving rise to a β-sandwich structure. The ligand-binding site is believed to be at one edge of the β-sandwich, designated as the hypervariable surface. The crystal structure of the second LNS domain of neurexin 1α revealed a similar overall fold, along with a Ca^{2+}-binding site at the hypervariable surface. The short intracellular region of neurexins (55 aa) is highly conserved among isoforms and contains a terminal class II postsynaptic density-ZO protein (PDZ) domain-binding motif. The invertebrate neurexin has a domain structure identical to vertebrate α-neurexins and modest sequence homology to the human α-neurexins, with *C. elegans* possessing 23–24%, and *D. melanogaster* neurexin possessing 29–30%, identity.

A characteristic feature of neurexins is the extensive alternative splicing of their transcripts, giving rise to >1000 potential isoforms. There are five sites for alternative splicing in α-neurexins, designated 1–5, with the last two being shared by β-neurexins. Alternative splicing involves the exclusion or insertion of sequences, which can range from a few amino acids at some sites to >30-amino-acids at others. Three of the sites in α-neurexins and one in β-neurexins are present within the LNS domain and splice insertions in some of these have been shown to interfere with ligand binding. Some insertions in splice site 5 of neurexin 3α/β contain stop codons that may truncate the protein before the transmembrane region, generating putatively secreted variants. The sites for alternative splicing as well as the splice insert sequences are evolutionarily conserved among mammals. Homologous splice sites (1–5) and splice-insert sequences were found for neurexin 1α in chick sympathetic ganglia. Splice variants for three α-neurexins in *Xenopus* embryos were detected at splice sites 1–2 but not at 3; however, splice sites 4–5 have not yet been analyzed. Significantly, alternative spliced sequences have not been detected in invertebrates and no evidence is as yet available to indicate alternative splicing in the invertebrate neurexin.

Sequence analysis of a putative additional *Drosophila* neurexin (*Drosophila* nrxnIV) identified it as the ortholog of vertebrate CASPR. This has now been placed, together with other proteins with similarities

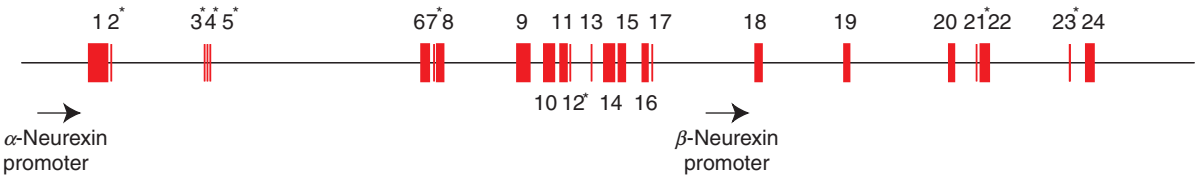

Figure 1 Structure of the human/mouse neurexin 1 gene. Shown here are 24 coding exons of the gene as red boxes or lines along with their numbers. The exons which are alternatively spliced are denoted by asterisk. The promoter for neurexin 1α lies upstream of coding exon 1, whereas the promoter for β-neurexin is located between coding exons 17 and 18. Exon 18 represents the β-specific exon.

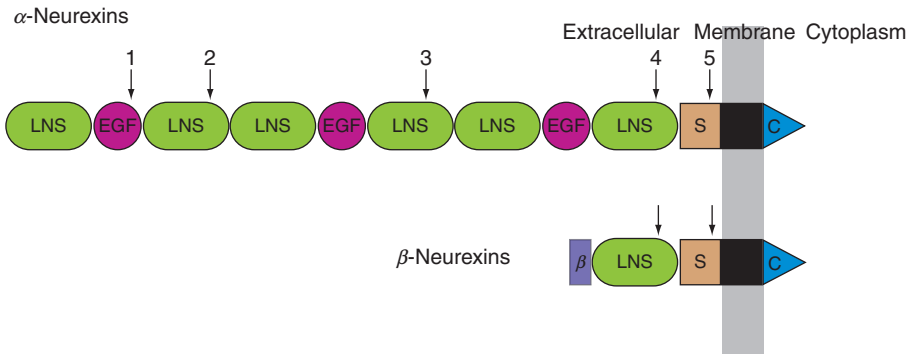

Figure 2 Domain structure of α- and β-neurexin proteins. The extracellular region of α-neurexins consists of six LNS domains and three EGF-like domains. The three LNS-EGF-LNS repeats in α-neurexins are followed by a serine-threonine-rich sequence (abbreviated S) that is O-glycosylated. The extracellular region of β-neurexins has a single LNS domain preceded by a short β-specific sequence (abbreviated β). The intracellular region of α- and β-neurexins is identical (abbreviated C). Arrows indicate the sites for alternative splicing. EGF, epidermal growth factor; LNS, laminin-neurexin-sex-hormone binding protein domain.

to neurexins, in the neurexin superfamily, which also includes CASPR/Paranodin (vertebrate) and BAM-2 (*C. elegans*). These proteins contain LNS and EGF-like domains, but the organization of their domains is distinctly different from that of the neurexins, as are their proposed roles.

Interaction Partners

Extracellular Binding Partners

The particular domain structure of the extracellular part of neurexins points toward its role as a ligand receptor. Extracellularly, α- and β-neurexins bind to the cell surface proteins neuroligins and dystroglycan, as well as to the exogenous neurotoxin α-latrotoxin. In addition, α-neurexins (but not β-neurexins) interact with neurexophilins, which constitute a family of neuropeptide-like glycoproteins (Nphl 1–4). Neurexophilin 1 and 3 bind to the second LNS domain of neurexins, and this binding is not affected by alternative splicing. Whereas neurexophilin 1 is predominantly expressed by inhibitory interneurons, neurexophilin 3 occurs in the excitatory neurons of distinct areas of the cortex and cerebellum, and has been found to play a role in sensorimotor gating and motor coordination in knockout mice, possibly in a tight complex with α-neurexins. Neuroligins are a family of postsynaptic cell-surface proteins that were discovered as endogenous ligands for β-neurexins. Neuroligins are encoded by five genes (NLGN1, -2, -3, -4, and -4Y) in humans and at least three genes in rodents. Neuroligins are alternatively spliced in their extracellular sequences at two sites, designated A and B. Neuroligins bind to the single LNS domain of β-neurexins and to the sixth LNS domain of α-neurexins in a Ca^{2+}-dependent manner that is regulated by alternative splicing on both proteins. β-Neurexins that lack an insert in splice site 4 bind to all neuroligins; however, β-neurexins which have an insert in splice site 4 and all α-neurexins may bind to neuroligins which lack an insert in their splice site B.

Intracellular Binding Partners

The intracellular region of α- and β-neurexins is identical and has been found to bind to multiple interaction partners (Table 1). A membrane-associated guanylate kinase (MAGUK) protein, calcium/calmodulin-dependent serine protein kinase (CASK) has been found to bind to the C-terminus of neurexins via its PDZ domain. The MAGUK family consists of adaptor proteins that have a catalytically inactive guanylate kinase domain. An important feature of MAGUK proteins is their multidomain structure, which can potentially help in the formation of multipartite

Table 1 Binding partners of neurexins[a]

Binding partners	*Subcellular localization*	*Binding site on neurexin*	*Binding site on partner*	*Phenotype of knockout mice*
Extracellular				
Neuroligins	Postsynaptic membrane	LNS6 of α-neurexins; single LNS of β-neurexins	Cholinesterase-like domain	Neuroligin 1 KO: no obvious phenotype Neuroligin 2–3 KOs: not known
Neurexophilins	Not known; ?Extracellular matrix bound to α-neurexins	LNS2 of α-neurexins	Not known	Nph1 KO: no obvious phenotype Nphl3 KO: impaired sensorimotor gating and motor coordination
Dystroglycan	Postsynaptic membrane	LNS2 and LNS6 of α-neurexins; single LNS of β-neurexins	Not known	KO: early embryonic lethality, defect basal lamina Conditional KO: no changes in synaptic differentiation
α-Latrotoxin	n.a.	LNS6 of α-neurexins; single LNS of β-neurexins	N-terminal cysteine-containing domain + (?) central ankyrin-repeats domain	n.a.
Intracellular				
CASK	Pre- and postsynaptic, cytosolic	PDZ domain-binding motif at the C-terminus	PDZ domain	Males show cleft palate and die on the first day of birth. Brain-specific conditional KO: not known
Mints	Pre- and postsynaptic, cytosolic	PDZ domain-binding motif	PDZ domain	Mint1 KO: No change in excitatory synaptic transmission, increased paired-pulse depression at inhibitory synapses
Syntenin	Pre- and postsynaptic, cytosolic	PDZ domain-binding motif	PDZ domain	Not known
Synaptotagmin 1	Presynaptic, vesicle membrane	PDZ domain-binding motif	Not known	Abolished fast component of Ca^{2+}-triggered exocytosis

[a]Additional intracellular binding partner such as CIPP and AF6 have been proposed, but evidence for these candidates is currently very limited. CASK, calcium/calmodulin-dependant serine protein kinase; KO, knockout; LNS, laminin-neurexin-sex-hormone binding protein domain; n.a., not applicable; PDZ, postsynaptic density-ZO protein domain.

complexes. Consequently, multiple interaction partners for CASK have been found, including Mints, Velis, α1B Ca^{2+}-channel subunit, protein 4.1, calmodulin, rabphilin 3a, and syndecan. It has been proposed that the tripartite complex of CASK, Mint, and Veli couple neurexin-mediated cell adhesion to synaptic vesicle exocytosis machinery as well as to Ca^{2+} channels at the active zone. In addition, the binding of CASK to protein 4.1 has been shown to nucleate the local assembly of actin/spectrin filaments on the intracellular tail of neurexin, linking cell adhesion to the actin cytoskeleton. In addition to CASK, the C-terminus of neurexins has been shown to interact with Mint1–2, syntenin, channel-interacting PDZ domain protein (CIPP), AF6, and synaptotagmin1. However, the number and promiscous nature of all these biochemical interactions raise doubts about their physiological significance, and further work is required to establish their specific role for neurexin function.

Expression

Neurexins are expressed throughout all brain regions and neuron types, although the different isoforms have partly overlapping and partly differential distributions. For example, CA1 pyramidal cells of hippocampus express neurexin 1α, 2α, 2β, and 3β, but lack neurexin 3α and 1β. In contrast, the interneurons in CA1 region express all three α-neurexins and neurexin 3β, whereas CA3 pyramidal neurons express all six principal neurexin isoforms. In addition to the brain, neurexins are also expressed in the spinal cord motor neurons, sympathetic neurons, and (neuro)

endocrine cells (from the pituitary, pancreas, and adrenal gland). Moreover, the alternatively spliced sequences have also been shown to differ in regional distribution. For example, neurexin 1 and 2mRNAs without an insert at splice site 4 are present at higher levels in the hippocampus and cerebral cortex than in the striatum, substantia nigra, and cerebellar nuclei. In contrast, neurexin 1 and 2 with an insert at splice site 4 show the opposite pattern. In the brain, neurexins have been shown to be enriched at synapses by biochemical fractionation methods. They are believed to be localized at presynaptic boutons on the basis of the action of their ligand α-latrotoxin, which causes massive neurotransmitter release from presynaptic terminals. However, a precise ultrastructural localization of neurexins has not yet been published, and the mechanism that targets neurexins to the synapse is still unclear.

The transcription of neurexin genes in the murine embryonic nervous system begins within 1 day of neurons becoming postmitotic. Immunoblot analysis using pan-neurexin antibodies revealed that levels of α- and β-neurexins increase dramatically after birth and reach peak expression during the second postnatal peak, which is a period of intense synaptogenesis. Expression then decreases, reaching a steady level in older animals. Another study, which used *in situ* hybridization, revealed an essentially similar picture for the neurexin 2α transcript, but found no major change in the expression of neurexin 1, 2β, and 3 transcripts during postnatal development. The expression of alternatively spliced variants of neurexins have also been shown to change during development. The expression of neurexin 1 transcript with no insert at splice site 4 increased dramatically in chick sympathetic ganglia from embryonic day 7.5 to embryonic day 18, whereas the expression of neurexin 1 with an insert at splice site 4 did not change strongly.

Function of Neurexins

Role of Neurexins in Synaptic Transmission

Neurexins were discovered as receptors for α-latrotoxin, a component of black widow spider venom, which causes massive neurotransmitter release. The action of α-latrotoxin immediately indicated a presynaptic localization of neurexins, as well as a role for neurexins in neurotransmitter release. The generation and analysis of knockout mice for all three α-neurexin genes recently confirmed the importance of α-neurexins for synaptic transmission and for Ca^{2+}-triggered release in general.

Mice lacking all three α-neurexins died on the first day of birth. Analysis of excitatory synapses in acute brain stem slices and inhibitory synapses in neocortical slice cultures revealed that evoked synaptic transmission was severely impaired in α-neurexin-knockout mice. This was accompanied by a reduction in the frequency of miniature synaptic transmission without affecting its amplitude. This indicated that α-neurexins are essential for neurotransmitter release at both the excitatory and inhibitory synapses. Further analysis revealed that the effect of deleting α-neurexins involved an impairment of Ca^{2+}-channel function. The synapses from knockout mice showed a markedly reduced sensitivity to blockers of N-type and P/Q-type Ca^{2+} channels, implying that α-neurexins might be responsible for coupling specific Ca^{2+} channels to the release machinery. In addition, it was found that whole-cell Ca^{2+} currents in brain stem neurons were reduced in knockout mice. Rescue experiments showed that transgenic neurexin 1α (but not neurexin 1β) alleviated the functional deficits of knockout mice, confirming that α-neurexins are directly involved in regulating the efficacy of neurotransmission and suggesting that the longer extracellular sequences of α-neurexins are required for this process. Other functional defects in α-neurexin knockout mice included a reduced *N*-methyl-D-aspartate (NMDA)/α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor ratio in neocortical neurons (presumably due to a compensatory reaction to the strongly reduced inhibitory transmission) and a profound inability of neuromuscular junctions to upregulate presynaptic release in response to postsynaptic receptor blockage. Significantly, a recent study revealed that deletion of α-neurexins from the pituitary gland melanotrophs also leads to an impairment of the secretory granule release, validating and extending the importance of α-neurexins as general organizer molecules of Ca^{2+}-triggered exocytosis in various systems (**Figure 3**).

Because mice lacking all β-neurexins are not yet available, a similar investigation for the function of β-neurexins in synaptic transmission has not been performed. However, the continued presence of β-neurexins in α-neurexin-knockout mice was not able to compensate for the profound deficits in neurotransmission. Therefore, it seems likely that α-neurexins have an essential function in synaptic transmission that is not shared by β-neurexins.

Role of Neurexins in Synaptogenesis

Considerable evidence has accumulated from cell culture studies to show that β-neurexins and their postsynaptic binding partners neuroligins are sufficient to induce *de novo* formation and differentiation of functional synaptic contacts. These studies concentrated on β-neurexins and not α-neurexins because the binding of

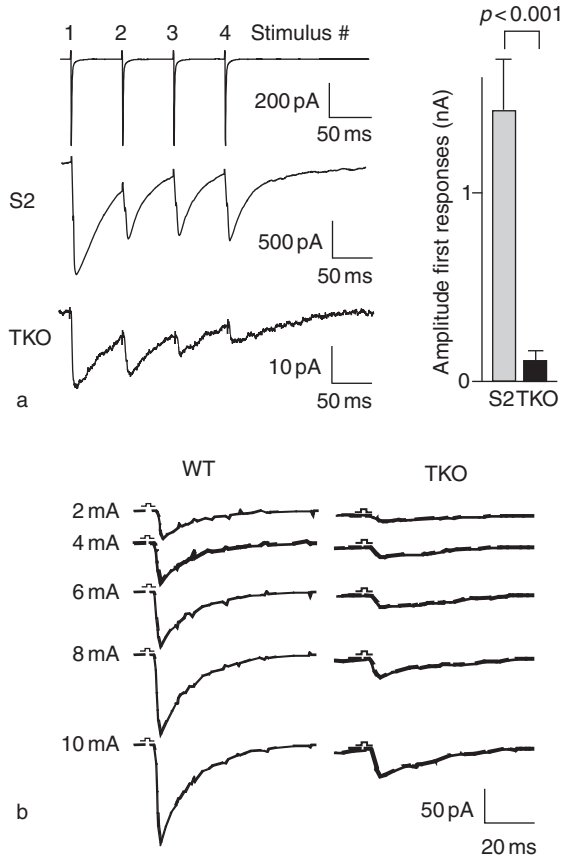

Figure 3 Impaired synaptic transmission at inhibitory and excitatory synapses in α-neurexin triple knockout (TKO) mice: (a) paired recordings of evoked $GABA_A$-receptor-mediated postsynaptic currents in pyramidal neurons of cultured neocortical slices from control (S2) and TKO mice; (b) glutamatergic-evoked responses in hypogoglossal motor neurons after field stimulation of axons from the pre-Bötzinger complex. In (a), the evoked release is reduced by 95% in α-neurexin TKO mice. In (b), the TKO mice show impaired evoked responses at different stimulation intensities. GABA, γ-aminobutyric acid; WT, wild type. From Missler M, Zhang W, Rohlmann A, et al. (2003) α-Neurexins couple $Ca2^+$ channels to synaptic vesicle exocytosis. *Nature* 423: 939–948.

α-neurexins to particular neuroligin isoforms was unknown until recently. In addition, the data from α-neurexin knockout mice showed a stronger effect on synaptic transmission than on synapse formation. Multiple groups have now detailed the role of β-neurexins and neuroligins in various aspects of synaptogenesis using cell culture assays; however, *in vivo* evidence from knockout mice is pending.

Presynaptic assembly The initial evidence for the involvement of β-neurexins in presynaptic assembly came from a chimeric culture model. The presentation of neuroligin 1 on nonneuronal HEK293 cells triggered presynaptic differentiation in contacting axons of pontine explants. There was accumulation of synaptic markers and the adaptor protein CASK at the site of induced synaptic contacts, where accumulated vesicles could also undergo depolarization-dependent turnover. Significantly, the synaptogenic effect on these *de novo*-formed contacts was lost when soluble neurexin 1β was added to the culture, implying that the interaction of neuroligins with β-neurexins is responsible for inducing presynaptic assembly. A subsequent study showed that clustering β-neurexins on the presnaptic surface is sufficient to induce presynaptic differentiation and that the synaptogenic activity of β-neurexin depends on the protein interactions of its intracellular part, although the downstream binding reactions are still unclear. One possibility is the binding of the neurexin C-terminus to modular adaptor proteins such as CASK and Mint, which can generate a scaffold of presynaptic machinery. Another possibility is the polymerization of actin on the cytoplasmic tail of neurexins through CASK and actin binding-protein 4.1, followed by F-actin stabilizing the clusters of synaptic vesicles at the induced presynaptic terminal.

Postsynaptic assembly As expected from a reciprocal interaction between neurexins and neuroligins, β-neurexins have also been shown to induce the assembly of the postsynaptic apparatus via their interaction with neuroligins. The presentation of neurexin 1β on nonneuronal COS cells triggered postsynaptic differentiation in contacting dendrites of primary hippocampal neurons, resulting in the clustering of scaffolding proteins postsynaptic density (PSD-)95 and gephyrin, as well as the neurotransmitter receptor (NR)1 subunit of NMDA receptors and γ-aminobutyric acid A ($GABA_A$) receptor subunits α2 and γ2. Interestingly, no accumulation of AMPA receptor subunits glutamate receptor (GluR)1 and GluR2 was observed, implying that the β-neurexin–neuroligin interaction might first assemble silent synapses. Later, recruitment of GluR1 subunit into the β-neurexin-induced scaffolds occurred following brief activation of NMDA receptors. This conversion of silent synapses into active synapses on NMDA receptor activation mimics the developmental maturation of synapses.

Development of synapse specificity When neurexins were discovered, the presence of multiple isoforms gave rise to the hypothesis that neurexins might act as a component of a recognition code for generating different types of synapses. This view was strengthened by the finding of a differential distribution of neurexins in different brain regions as well as in pyramidal cells and interneurons. However, the distribution was found to be overlapping, with no evidence for the selective expression of specific neurexin isoforms

in glutamatergic or GABAergic cells. However, the situation turned out to be different for neuroligins. It was shown that neuroligin 1, 3, and 4 are localized to glutamatergic (but not GABAergic) synapses, whereas neuroligin 2 is localized mainly to GABAergic synapses. The presentation of neurexin 1β on COS cells induced clustering of neuroligin 1 and 2 in contacting dendrites which subsequently recruited both gephyrin and PSD-95-containing postsynaptic scaffolds. However, more neuroligin 2 puncta were associated with gephyrin clusters than with PSD-95 clusters. Recent evidence points toward a stronger role for neurexin isoforms in determining synapse specificity. In contrast to insert-negative neurexin 1β (at splice site 4) used in the studies previously discussed, insert-positive neurexin 1β (at splice site 4) and neurexin 1α exclusively triggered the assembly of gephyrin clusters through selective recruitment of neuroligin 2. Interestingly, α-neurexin-knockout mice showed a reduction in the area density of inhibitory synapses, although the overall number of excitatory synapses also appears to be reduced due to a neurophil defect evident in older knockout animals that lack only two out of three neurexin genes. These findings suggest a role for some neurexin isoforms in determining glutamatergic versus GABAergic synapse induction (Figure 4).

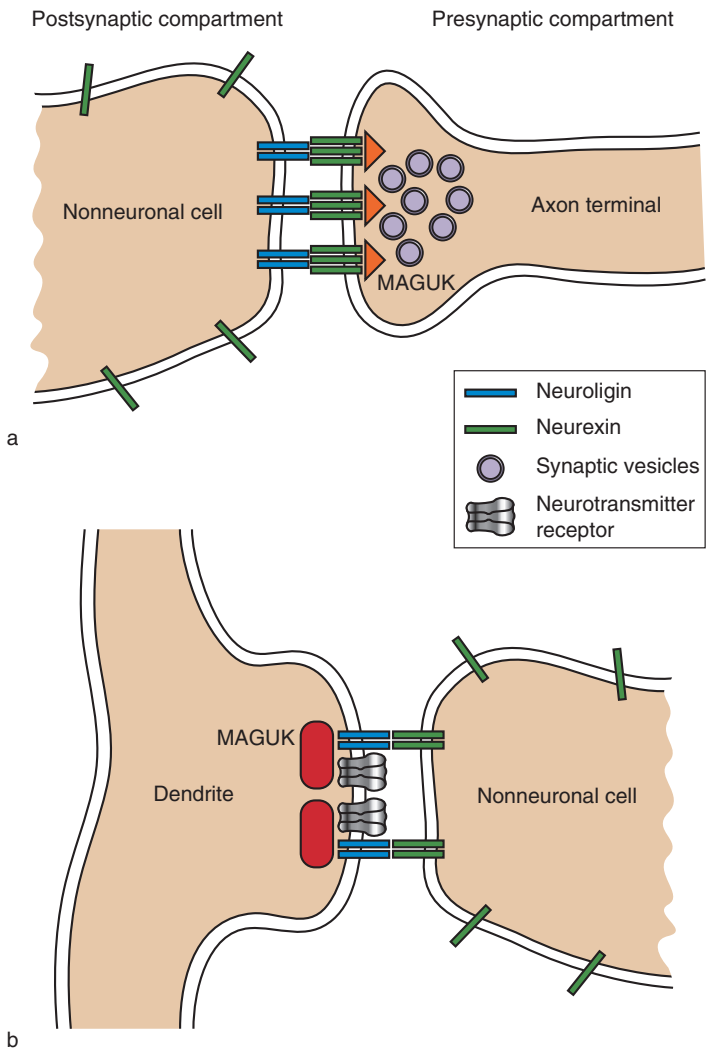

Figure 4 Diagram summarizing the synaptogenic potential of the neurexin–neuroligin interaction in chimeric culture assays: (a) nonneuronal cells transfected with neuroligin cluster neurexin molecules at the presynaptic terminus of contacting axons, leading to the recruitment of synaptic vesicles and scaffolding proteins (e.g., MAGUKs such as CASK); (b) presentation of neurexin on nonneuronal cells induces the recruitment of postsynaptic scaffolding proteins (e.g., MAGUKs such as PSD-95 and gephyrin) and neurotransmitter receptors (e.g., NMDA and $GABA_A$ receptor subunits) via neuroligins. CASK, calcium/calmodulin-dependent serine protein kinase; GABA, γ-aminobutyric acid; MAGUKs, membrane-associated guanylate kinases; NMDA, *N*-methyl-D-aspartate; PSD-95, postsynaptic density 95.

Synaptic cell adhesion Some studies also point toward the role of the neurexin–neuroligin interaction in mediating mechanical cell adhesion at synapses, which could help in aligning the pre- and postsynaptic specializations. Normally nonadhesive insect cell lines expressing neurexin 1β or neuroligin 1 formed cell aggregates only when these two cell lines were mixed together in the presence of Ca^{2+} ions, indicating that neurexin–neuroligin binding is capable of mediating actual cell adhesion. Moreover, neurexins have been found to be enriched in the growth cones of isolated axons emerging from pontine explants in culture. Their concentration in growth cones localizes them appropriately to catch neuroligin-containing postsynaptic targets. However, α-neurexin-knockout mice did not reveal an increase in the width of the synaptic cleft or a misalignment of pre- or postsynaptic densities.

Clinical Aspects of the Neurexin–Neuroligin Complex

Mutations in neurexins have not yet been directly implicated in any major disorder. Nonetheless, a recent study using pooled association genome scanning has revealed a putative link of a single-nucleotide polymorphism (SNP) residing within the neurexin 3 gene with addiction vulnerability to polysubstance abuse. The involvement of the neurexin–neuroligin interaction in autistic spectrum disorders is currently under intensive investigation because some studies have reported a linkage of mutations in neuroligin 3 and 4 to cases of autism, without, however, any implication of mutations in neurexin genes.

Along this line, the expression of neurexins appears to be under dynamic control and has been found to vary under pathophysiological conditions. One study reported increased expression of neurexin 2 mRNA following kainate- and pentylenetetrazole-induced seizures in the dentate gyrus of rats. Transient global ischemia has also been shown to induce the elevation of neurexin 2α and neurexin 3α mRNA in the hippocampal CA1 and CA3 regions, respectively. In addition, there was an appearance of an insert-positive isoform of neurexin 3α (at splice site 4) in the CA3 region following ischemia. Gene expression analysis of spontaneously hypertensive rat cerebral cortex revealed altered levels of neurexin 1β following transient focal ischemia. It is unclear, however, whether these changes in neurexin expression are part of a compensatory homeostatic response, or whether they participate in the progression of the pathological insult.

See also: Calcium Channels; Calcium Channels and SNARE Proteins; Cell Adhesion Molecules at Synapses; Exocytosis: Ca^{2+}-Sensitivity; Neural Cell Adhesion Molecules and Synapse Regulation; Neuroligins; Neuromodulation of Calcium Channels; Neurosecretion (Regulated Exocytosis in Neuroendocrine Cells); Voltage-Gated Calcium Channels.

Further Reading

Beglopoulos V, Montag-Sallaz M, Rohlmann A, et al. (2005) Neurexophilin 3 is highly localized in cortical and cerebellar regions and is functionally important for sensorimotor gating and motor coordination. *Molecular and Cell Biology* 25: 7278–7288.

Boucard AA, Chubykin AA, Comoletti D, Taylor P, and Sudhof TC (2005) A splice code for trans-synaptic cell adhesion mediated by binding of neuroligin 1 to α- and β-neurexins. *Neuron* 48: 229–236.

Butz S, Okamoto M, and Sudhof TC (1998) A tripartite protein complex with the potential to couple synaptic vesicle exocytosis to cell adhesion in brain. *Cell* 94: 773–782.

Dean C, Scholl FG, Choih J, et al. (2003) Neurexin mediates the assembly of presynaptic terminals. *Nature Neuroscience* 6: 708–716.

Dudanova I, Sedej S, Ahmad M, et al. (2006) Important contribution of α-neurexins to Ca^{2+}-triggered exocytosis of secretory granules. *Journal of Neuroscience* 26: 10599–10613.

Fairless R, Reissner C, and Missler M (2006) Role of neuroligin binding to neurexins in synaptic organization. In: Dityatev A, and El-Husseini A (eds.) *Molecular Mechanisms of Synaptogenesis*, pp. 111–124. New York: Springer.

Graf ER, Zhang X, Jin SX, Linhoff MW, and Craig AM (2004) Neurexins induce differentiation of GABA and glutamate postsynaptic specializations via neuroligins. *Cell* 119: 1013–1026.

Ichtchenko K, Hata Y, Nguyen T, et al. (1995) Neuroligin 1: A splice site-specific ligand for β-neurexins. *Cell* 81: 435–443.

Missler M, Hammer RE, and Sudhof TC (1998) Neurexophilin binding to α-neurexins: A single LNS domain functions as an independently folding ligand-binding unit. *Journal of Biological Chemistry* 273: 34716–34723.

Missler M and Sudhof TC (1998) Neurexins: Three genes and 1001 products. *Trends in Genetics* 14: 20–26.

Missler M, Zhang W, Rohlmann A, et al. (2003) α-Neurexins couple Ca^{2+} channels to synaptic vesicle exocytosis. *Nature* 423: 939–948.

Rudenko G, Nguyen T, Chelliah Y, Sudhof TC, and Deisenhofer J (1999) The structure of the ligand-binding domain of neurexin Iβ: Regulation of LNS domain function by alternative splicing. *Cell* 99: 93–101.

Ullrich B, Ushkaryov YA, and Sudhof TC (1995) Cartography of neurexins: More than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron* 14: 497–507.

Ushkaryov YA, Petrenko AG, Geppert M, and Sudhof TC (1992) Neurexins: Synaptic cell surface proteins related to the α-latrotoxin receptor and laminin. *Science* 257: 50–56.

Ushkarov YA, Petrenko AG, Geppert M, and Sudhof TC (1992) Neurexins: Synaptic cell surface proteins related to the α-latrotoxin receptor and laminin. *Science* 257: 50–56.

Neuroanatomy Methods in Humans and Animals

P Lavenex, University of Fribourg, Fribourg, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

The study of the structural organization of the nervous system dates back many centuries, but our understanding of the detailed organization of the brain and other nervous tissue has been largely dependent on the development of technologies within the past 150 years. In particular, two pieces of equipment developed in the middle of the nineteenth century have been fundamental to the study of the nervous system: the microtome and the compound microscope. Although the first compound microscope was developed at the end of the sixteenth century by Hans and Zacharias Janssen (by placing lenses at each end of a tube), major technological improvements were made in the mid-nineteenth century, leading to the predecessors of modern-day light microscopes. The microtome, used to cut thin slices of biological preparations adequate for microscopic evaluation, similarly came into common use during the mid-nineteenth century. Eventually, detailed examination of the fine structure of axons, dendrites, and synapses required the development of additional technologies. The limited magnification and resolution of the light microscope led to the development of the electron microscope in the early 1930s. Since the advent of these technologies, numerous biological staining procedures have been developed enabling the study of the fine structure of the nervous system, thus ushering in this modern era of neuroscience. Interestingly, however, despite the great advances in modern *in vivo* imaging techniques and informatics, some neuroanatomical procedures developed more than 100 years ago are still central to current investigations of the structural organization of the nervous system of humans and other animals.

Classical Histochemical Techniques

Our understanding of the detailed structural organization of the nervous system is still largely dependent on the microscopic examination of postmortem tissue. To produce these specimens, nervous tissue needs to be prepared with a fixative solution that preserves the integrity of the structures to be examined. This is generally accomplished by perfusion of aldehyde solutions through the vasculature of deeply anesthetized experimental animals or by immersion in a fixative solution for the case of postmortem human tissue. The block of tissue is then sectioned either frozen or after being embedded in media such as paraffin or celloidin, and the micrometers-thick sections are stained by different methods. The most widely used staining procedure is undoubtedly the technique published in 1894 by Franz Nissl and now commonly known as the Nissl stain (Figure 1). Basic dyes, such as thionin or cresyl violet, stain the acidic and basophilic components of the Nissl bodies (stacks of granular endoplasmic reticulum in the cytoplasm of cells), as well as the chromatin in the nuclei and nucleoli of cells. This technique has the advantage, and the disadvantage, of indiscriminately staining neuronal and glial cells. Because of its simplicity, the Nissl stain can easily be implemented in any neuroscience laboratory. It works well on tissue fixed with almost any kind of fixative and even on unfixed tissue. Nevertheless, great care is necessary to produce consistent and reliable Nissl-stained preparations in order to perform quantitative analyses of the nervous system. Although modern neuroanatomical definitions of cortical and subcortical brain areas are based on a convergence of evidence linking cytoarchitectonic, neurochemical, molecular, connectional, and developmental criteria, they always refer eventually to the cytoarchitectonic description of the distinctive types and arrangement of cells in a given area as described in Nissl-stained tissue.

Another important technique that remains of current interest is the Golgi technique, allowing the visualization of the dendritic and axonal processes of neurons (Figure 2). This technique was first published in 1873 by Camillo Golgi and employed extensively by Ramón y Cajal for his seminal work *Histologie du Système Nerveux de l'Homme et des Vertébrés*. Although adult brain material is generally considered more difficult to stain than tissue from younger organisms, variants of the technique can be adapted to produce good quality preparations from a broad variety of ages and species, and with either fresh or fixed tissue. Small pieces of tissue are immersed in solutions containing potassium dichromate for several days, followed by immersion in silver nitrate solutions kept in the dark for several more days. Serial sections of 150–250 μm are then cut, revealing neurons filled with a black precipitate of silver chromate. The strength of the approach is that, for unknown reasons, it labels only a small portion of all neurons in a nonspecific manner, enabling the reconstruction of the dendritic and umyelinated axonal plexus of individual cells. One shortcoming of the Golgi technique is that it is difficult to accurately reconstruct entire dendritic and axonal arborizations

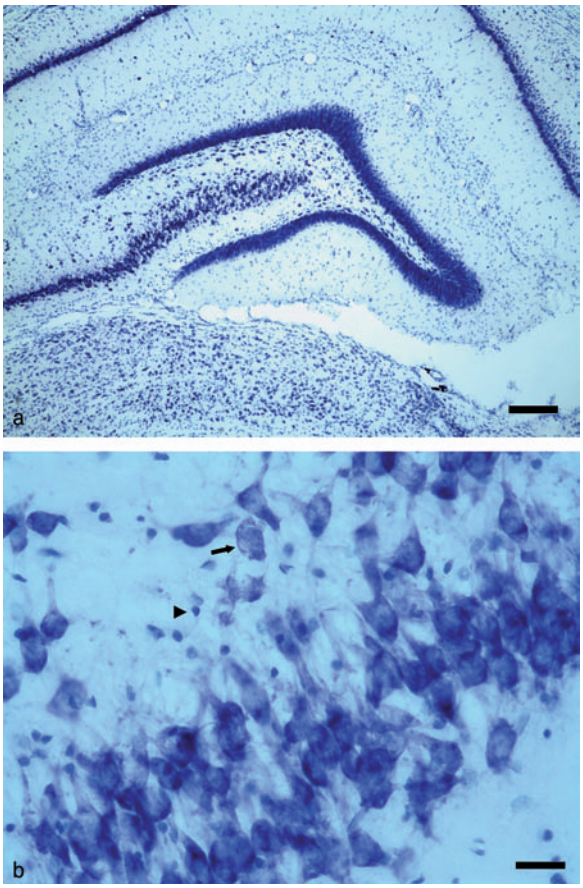

Figure 1 (a) Photomicrograph of a Nissl-stained coronal section through the dentate gyrus of an 180-day-old Sprague–Dawley rat. (b) High-magnification photomicrograph of the pyramidal cell layer of the CA3 field of the hippocampus. Pyramidal neuron indicated by an arrow; nucleus of a glial cell indicated by an arrowhead. Scale bar = 250 μm (a) and 25 μm (b).

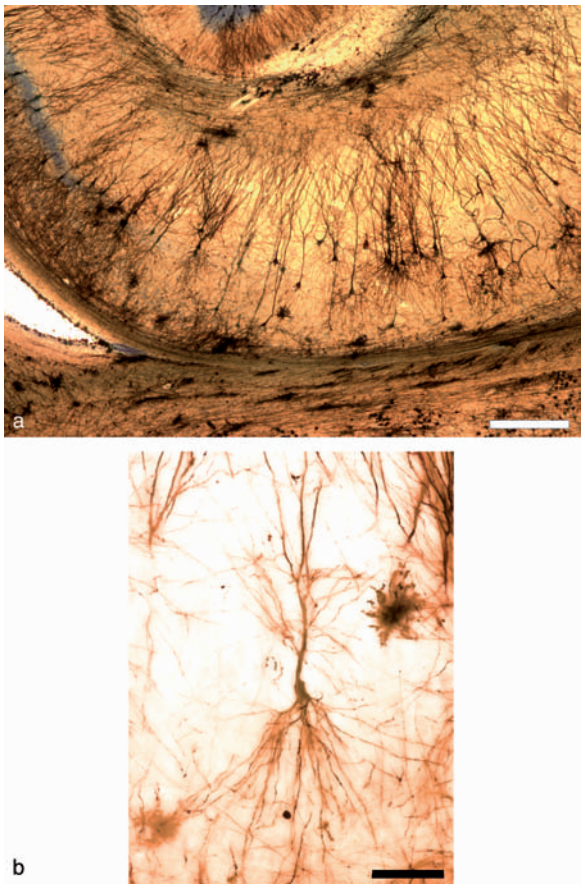

Figure 2 (a) Photomicrograph of a Golgi-stained section illustrating the dendritic arborization of CA1 pyramidal neurons in the macaque monkey. (b) High-magnification photomicrograph of a Golgi-stained CA1 pyramidal neuron in the macaque monkey. Scale bar = 1mm (a) and 100 μm (b). Modified from Altemus KL, Lavenex P, Ishizuka N, and Amaral DG (2005) Morphological characteristics and electrophysiological properties of CA1 pyramidal neurons in macaque monkeys. *Neuroscience* 136(3): 741–756, with permission.

due to the necessity of using thin sections in order to visualize the stained processes. Quantitative measurements of dendritic processes derived from Golgi-filled neurons therefore underestimate such parameters, compared to intracellularly labeled neurons (which can be visualized in thicker sections). Nevertheless, it remains a powerful technique to study the morphological characteristics of neurons, including the shape and size of their soma, the length of dendrites, and the number of branches and spines, in both human and experimental animal tissue.

A number of other histochemical staining procedures have been developed during the past 150 years in order to visualize various molecules, such as myelin, or study specific cell types or neuron populations. Histochemical myelin stains remain common in fundamental research because variations in myelination patterns provide additional information by which to define structurally different brain regions. Neuropathology laboratories also rely on myelin stains to evaluate the demyelination processes that occur in diseases such as multiple sclerosis. The first histochemical myelin staining method was published in 1884 by Weigert, but it has largely been replaced by simpler and more reliable methods using silver nitrate or gold chloride staining procedures. Immunohistochemical procedures using antibodies against myelin basic protein are also becoming more common.

Other stains developed to study specific cell populations are based on the detection of enzymatic activity linked to the synthesis or degradation of particular neurotransmitters. One example is acetylcholinesterase, the enzyme that breaks down acetylcholine and is present at all cholinergic synapses and therefore in many neurons that receive cholinergic innervation

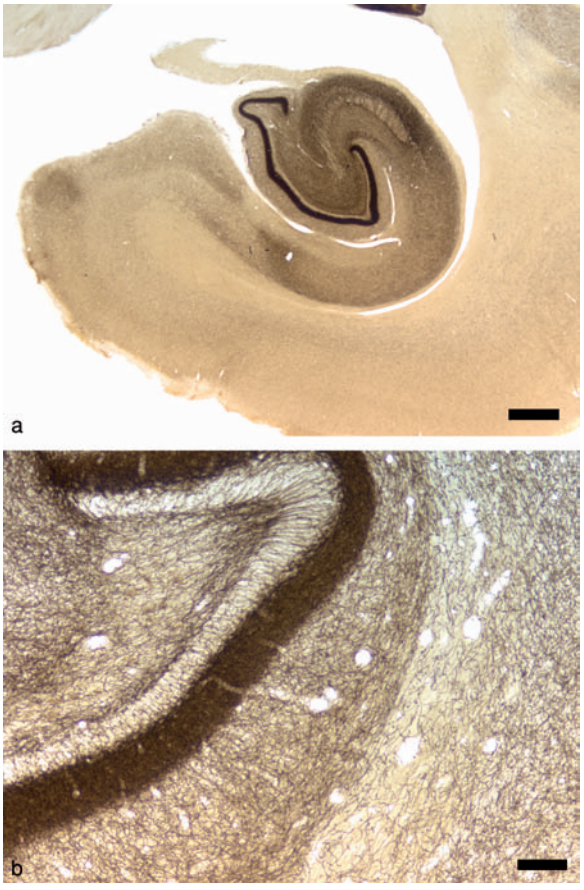

Figure 3 (a) Low-magnification photomicrograph of a coronal section through the monkey hippocampal formation stained for acetylcholinesterase enzymatic activity. (b) High-magnification photomicrograph of the monkey dentate gyrus stained immunohistochemically with an antibody raised against acetylcholinesterase. Note the high level of labeling in the inner portion of the molecular layer of the dentate gyrus observed with the two different staining methods. Scale bar = 1 mm (a) and 100 μm (b).

but that are not necessarily cholinergic. Interestingly, the activity of this enzyme is very robust and does not seem to be impacted by delays between death and fixation, making it particularly suitable for the study of human postmortem tissue, as long as this material is not overfixed (Figure 3(a)). In contrast, choline acetyltransferase, the enzyme necessary for the synthesis of acetylcholine and which is present exclusively in cholinergic neurons, is detected immunohistochemically (acetylcholinesterase can also be detected immunohistochemically; Figure 3(b)). Thus, acetylcholinesterase enzymatic activity and choline acetyltransferase immunohistochemistry can be used as differential markers of the cholinergic system.

Modern Molecular Techniques

A major advance in the study of the detailed organization of the nervous system has been the development of molecular detection techniques such as immunohistochemistry and *in situ* hybridization. The great advantage of these techniques is that they enable researchers to study specific subpopulations of neuronal or glial cells. Immunohistochemical staining procedures entail incubating tissue sections in a solution containing antibodies raised against particular molecules or peptides that are constitutive elements of neurotransmitters, receptors, enzymes, or structural proteins. These primary antibodies can be tagged by a marker that can be visualized under the microscope, but most often they are visualized via the use of a secondary antibody raised against the IgG of the host animal producing the primary antibody (indirect detection system). The secondary antibody can be tagged with one of various fluorescent markers (fluorophores), which can be visualized using fluorescent light microscopy, or with biotin, which can be visualized using avidin–horseradish peroxydase (avidin–biotin complex method) as the reagent and diaminobenzidine tetrachloride as the chromogen to produce a permanent labeling for bright field or dark field light microscopy. The indirect detection of antigens, which relies on two antibodies and a separate detection system, enables the amplification of the signal and thus increased sensitivity. Literally thousands of molecules present in neuronal and glial cell populations can now be detected with such techniques (Figures 3(b) and (4)). However, the spurious staining of unrelated molecules that happen to cross-react with the antiserum is an important concern. Primary antibodies are produced by injection of a known molecule into an animal such as a mouse or a rabbit, followed by the purification of the IgG immunoglobulins directed against this molecule from the animal's serum. Primary antibodies can be either polyclonal (i.e., produced by several B cell lines and typically recognizing several epitopes) or monoclonal (i.e., derived from a single B cell line and recognizing one single epitope). Thus, the specificity of the antibody for a given molecule should always be verified, particularly when working with novel molecules or specimens in which the normal distribution of the target molecule is not known. Finally, immunohistochemical reactions are also very sensitive to the quality and the nature of the fixation. An antibody might recognize a given epitope in a protein's native state but not when the protein has been exposed to a fixative which modifies its three-dimensional configuration. The most commonly used fixative is paraformaldehyde, which ensures the preservation and detection of the greatest number of proteins, thus allowing for the use of multiple antibodies directed at multiple target molecules. However, certain molecules, particularly smaller ones such as γ-aminobutyric acid (GABA), require

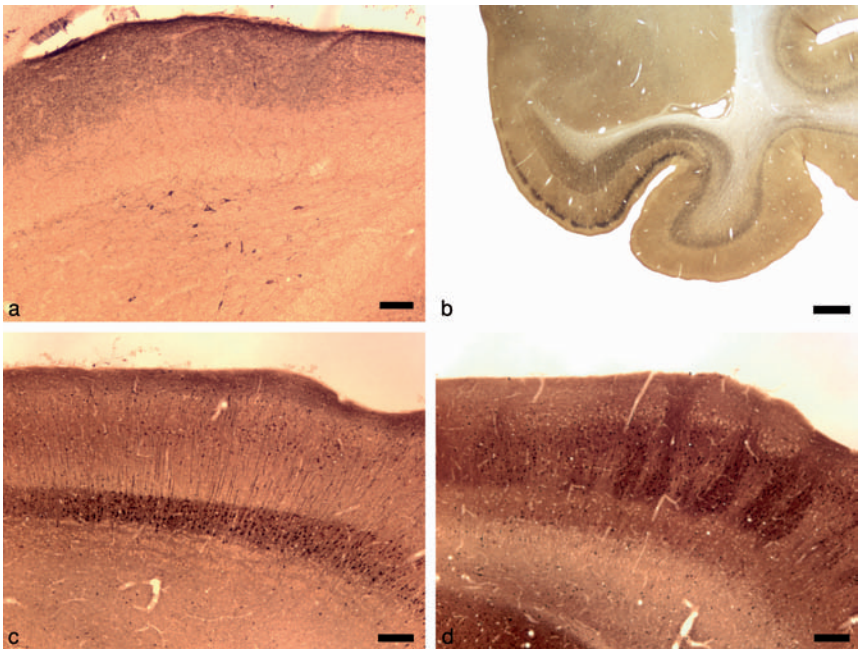

Figure 4 (a) Coronal section through the monkey dentate gyrus stained immunohistochemically with a primary antiserum which recognizes somatostatin-281-12 (S320). (b) Coronal section through the monkey medial temporal lobe stained immunohistochemically with a primary antibody against nonphosphorylated high-molecular-weight neurofilaments (SMI-32). (c) Coronal section through the monkey entorhinal cortex stained immunohistochemically with a primary antibody against the calcium-binding protein, calretinin. (d) Coronal section through the monkey entorhinal cortex stained immunohistochemically with a primary antibody against the calcium-binding protein, calbindin. Note the separate neuronal populations revealed by the two antibodies raised against different calcium-binding proteins. Scale bar = 100 μm (a), 1 mm (b), and 250 μm (c, d).

the addition of glutaraldehyde to the fixative solution in order to produce reliable results. Glutaraldehyde is also often required to produce quality material for electron microscopy examination.

In situ hybridization techniques, in contrast, target the mRNA that codes for the proteins. Chemically or radioactively labeled RNA or DNA antisense probes that are complimentary to the mRNA sequence of interest are used to localize the expression of that mRNA in the tissue. The major advantage of *in situ* hybridization is its specificity because probe sequences are generated to match one single protein-coding mRNA. Sense strand probes, which contain the same nucleic acid sequence in reverse order, are used to control for nonspecific binding of a given probe. However, *in situ* hybridization techniques detect specific mRNAs rather than the proteins that ultimately contribute to the function of the nervous system. Posttranslational modifications, such as phosphorylation, may have such a fundamental impact on protein function that *in situ* hybridization is not necessarily sufficient to obtain a comprehensive description of the phenomenon under study.

In addition to immunohistochemistry and *in situ* hybridization techniques, high-throughput molecular biology techniques, such as cDNA microarrays, serial analysis of gene expression, and real-time quantitative reverse-transcriptase polymerase chain reaction, enable researchers to study the expression of a large number of genes in defined brain regions or even in individual cell types. Nervous tissue is generally frozen or treated with an RNA preservation solution. Careful microdissection or sophisticated laser capture microdissection techniques are used to obtain tissue specimens. Following extraction, mRNA is typically amplified before the sample is analyzed. Commercially available DNA microarrays, for example, can be used to evaluate gene expression over the entire genome in humans as well as in other animals. Such techniques, however, require follow-up experiments using immunohistochemistry and *in situ* hybridization techniques to provide precise information about the regional and cellular localization of target genes.

Confocal Microscopy

In order to realize their full potential, the development of the described sophisticated molecular biology techniques capable of detecting small quantities of particular molecules has required the development of equally sophisticated imaging techniques. One such development has been that of the confocal

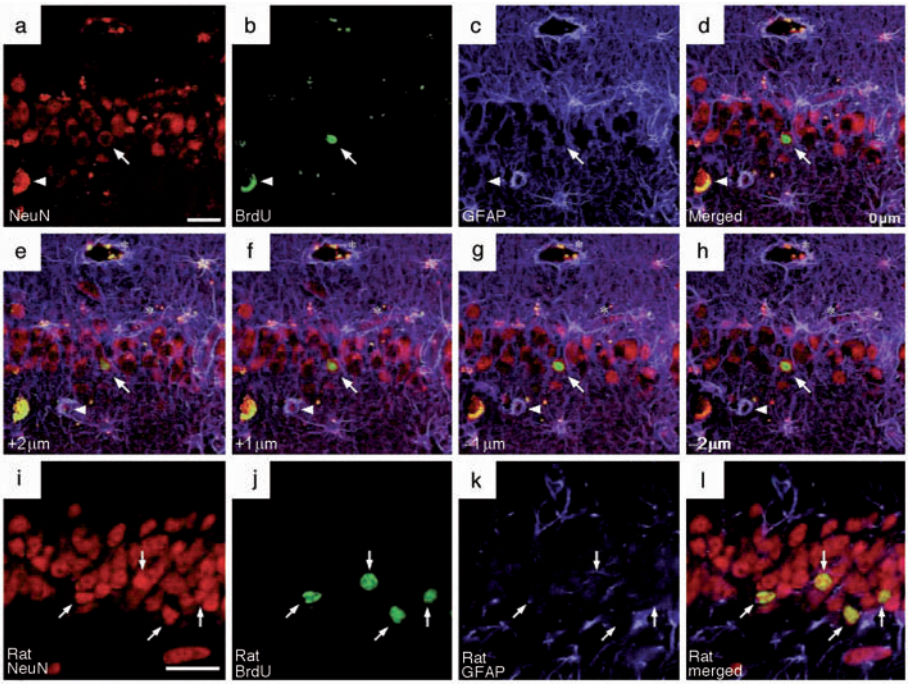

Figure 5 Confocal microscope images demonstrating the co-localization of cell division and neuron-specific markers in the adult human dentate gyrus. (a) Immunodetection of the neuronal marker NeuN (arrowhead). (b) Immunodetection of the cell-division marker BrdU (arrowhead). (c) Immunodetection of the glial marker GFAP (arrowhead). (d) Merged images demonstrating the co-localization of NeuN and BrdU (arrowhead). The arrows in (a–d) indicate nonspecific fluorescence. The specificity of BrdU–NeuN co-expression in three dimensions is demonstrated by a series of focal planes above (e, f) and below (g, h) the focal plane shown in (d). The appearance of adult BrdU-positive neurons in the human dentate gyrus is equivalent to BrdU-labeled neurons in the adult rat dentate gyrus. Scale bar = 25 μm (a–l). Reprinted by permission from Macmillan Publishers Ltd: *Nature Medicine* 4(11): 1313–1317, Eriksson PS, Perfilieva E, Bjork-Eriksson T et al., Neurogenesis in the adult human hippocampus, copyright (1998).

microscope. Confocal microscopes use pinholes to filter the light arriving from a defined focal plane in the specimen, thus improving the *z*-resolution of the light microscope to the submicrometer range. This increased resolution enables the localization of labeled substances within defined cellular compartments of individual cells. Confocal microscopy is now routinely used in studies employing multiple biological markers or labels. First, biological specimens are incubated with several primary antibodies, raised in different species, directed against different target molecules. Secondary antibodies, marked by different fluorophores which can be activated by different lasers to emit light at specific wavelengths, are then used to detect the primary antibodies and therefore determine the precise location of the various molecules of interest. Because of the increased resolution in the vertical axis, the use of confocal microscopy can verify whether different molecules that are co-localized in the *x* and *y* axes when observed in conventional light microscopy are also co-localized in the *z* axis.

Confocal microscopy has been particularly useful for the study of neurogenesis in the developing and adult vertebrate brain. Early studies using the autoradiographic detection of the cell division marker ^{3}H-thymidine did not possess the necessary resolution to unequivocally demonstrate the neuronal or glial nature of new cells that had incorporated the marker. The advent of immunohistochemical markers of cell division such as 5-deoxy-bromo-uridine (BrdU) and cell-specific markers such as NeuN for neurons and GFAP or S100beta for glia, combined with confocal microscopy, has enabled researchers to disprove the dogma that no new neurons were generated in the adult mammalian nervous system (Figure 5).

Quantitative Studies of the Nervous System

Modern neuroanatomical studies increasingly include the quantification of various aspects of the central nervous system. Quantitative estimates of the basic morphological characteristics of the nervous system are an essential prerequisite to generating realistic models of human brain function and pathology. It is only with the development of design-based stereological techniques that researchers have finally been able to provide reliable estimates of volume and neuron number in the nervous system. Volume estimates are typically calculated with the Cavalieri

method, which is based on a mathematical principle described in 1635. The volume of a structure can be estimated from the profile areas of the structure on cut sections, given that the first section to be measured is chosen randomly and all subsequent sections are at fixed intervals (e.g., systematic random sampling). The area of the structure on each section is estimated by overlying a grid and counting the number of points present in the structure (Figure 6(A)). The volume of the structure is obtained by multiplying the total number of points contained within the structure across all sections (P_i) by the area associated with each point ($a(i)$) and the distance between sampled sections (t):

$$\text{volume} = \sum P_i \times a(i) \times t$$

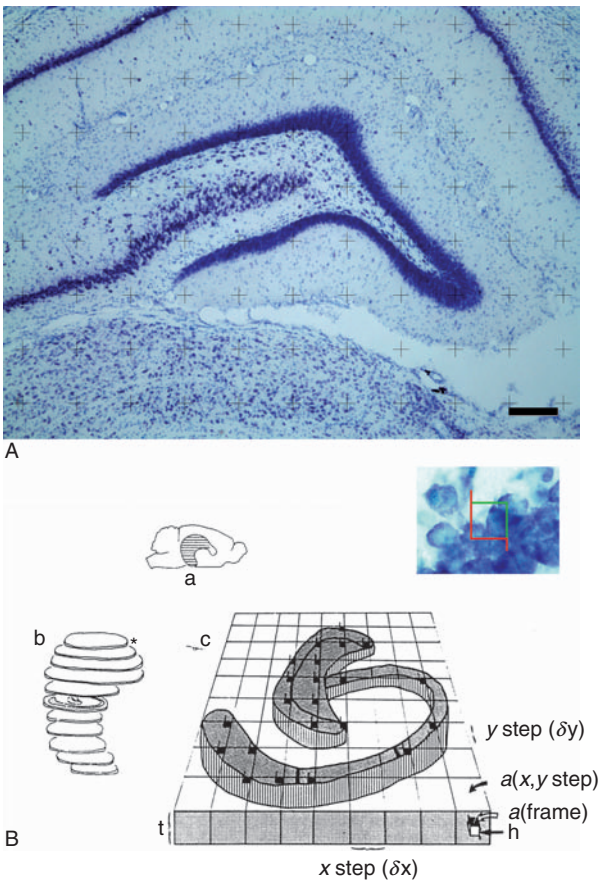

Figure 6 (A) Point grid overlay used to measure the volume of the molecular layer of the rat dentate gyrus using the Cavalieri method. The number of points (cross' intersections) are counted for each coronal section in the series. (B) Diagramatic representation of the optical fractionator sampling scheme. The area sampling fraction is the area of the counting frame (inset) relative to the area associated with each stepping frame: asf = a(frame)/a(x,y step). The thickness sampling fraction is the height of the dissector relative to section thickness: h/t. Modified from West MJ (1993) New stereological methods for counting neurons. *Neurobiology of Aging* 14(4): 275–285, with permission.

Note that the Cavalieri principle does not require any sophisticated equipment and can be applied to any kind of preparation, ranging from photomicrographs of electron microscopy material to printouts of magnetic resonance images.

Another important stereological probe is the optical fractionator used to estimate the total number of neuronal or glial cells in a defined structure (Figure 6(B)). This method also depends on the systematic random sampling of the structure of interest. That is, all the cells in the target space must have an equal probability of being sampled and therefore the integrity of the entire structure of interest is necessitated. The optical fractionator provides an estimate of the number of objects in a given space, and it has the major advantage of being free from the confounds of differential shrinkage and does not require any knowledge of the volume of the structure. The total number of neurons is counted in a known fraction of the structure of interest:

$$N = \sum Q \times (1/\text{ssf}) \times (1/\text{asf}) \times (1/(h/t))$$

where ssf is the fraction of sections being sampled, asf is the fraction of the surface area of every section being sampled, and h/t is the fraction of the thickness of the section being sampled. As with the Cavalieri method, the optical fractionator can be performed with all sorts of preparations, but in contrast it requires the assistance of a computerized microscope analysis system. Other stereological probes based on systematic random sampling have been developed to measure fiber length, cell size, and other biological parameters of interest.

Intracellular Techniques and Three-Dimensional Reconstruction of Individual Cells

Modern intracellular labeling techniques, combined with thick slice preparations, have provided a unique contribution to the study of the morphology and the intracellular electrophysiology of neurons and glia. Tissue is typically blocked to produce 400-μm-thick slices that are maintained in oxygenated artificial cerebrospinal fluid. Glass recording electrodes are filled with a tracer substance, such as biocytin, horseradish peroxidase (HRP), or Lucifer Yellow, and inserted into an individual cell. Following intracellular recordings, the tracer is infused by passing hyperpolarizing current into the cell. Once the section is fixed with paraformaldehyde, the intracellular tracer can be visualized by common histological procedures. Labeled neurons can be photographed and reconstructed at high magnification with a computer-assisted analysis system (Figure 7). The strength of this

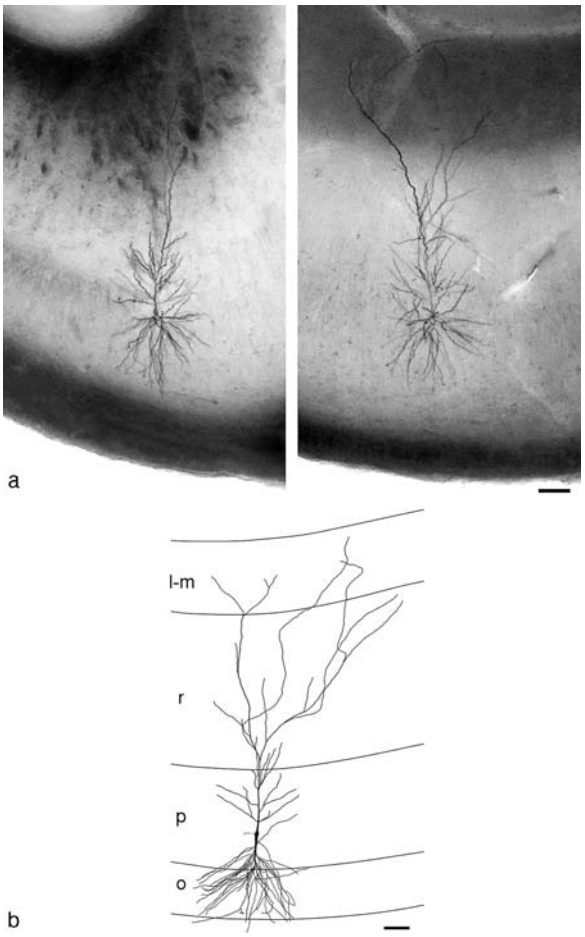

Figure 7 (a) Photomicrographs of monkey CA1 pyramidal neurons labeled intracellularly with biocytin. (b) Three-dimensional, computer-aided reconstruction of a typical monkey CA1 pyramidal neuron. l-m, stratum lacunosum moleculare; o, stratum oriens; p, stratum pyramidale; r, stratum radiatum. Scale bar = 100 μm (a, b). Reproduced from Altemus KL, Lavenex P, Ishizuka N, and Amaral DG (2005) Morphological characteristics and electrophysiological properties of CA1 pyramidal neurons in macaque monkeys. *Neuroscience* 136(3): 741–756, with permission.

approach is that it enables researchers to establish relationships between the structure and function of individual cells. Electrophysiological data about basic intrinsic electrical properties, firing patterns, and the pharmacological characterization of neurotransmitter receptors and ion channels can be gathered for individual cell types that can be fully characterized morphologically. Although intracellular injections of tracer substance can also be performed in experimental animals *in vivo*, intracellular recordings in whole brain preparations tend to be technically more challenging and typically do not permit a detailed study of the electrophysiological properties of these cells. Thick-slice techniques can be applied to nervous system preparations from experimental animals, as well as to human pathological material removed during neurosurgeries. Potential differences in electrophysiological or morphological characteristics between animals and humans are often difficult to interpret, however, because they could reflect true species differences or the pathology of the human material (obvious ethical concerns preclude the use of brain tissue from healthy human individuals). Nevertheless, parallel studies using classical histochemical techniques, such as the Golgi technique, can be used in healthy human postmortem preparations to determine the normal morphological characteristics of the cells.

Neuroanatomical Tract-Tracing Techniques

Although knowledge of the anatomical pathways of the brain cannot in itself explain how the brain works, the mapping of the central nervous system is absolutely fundamental to such understanding. Early studies of the connectivity of the nervous system were based on the physicochemical properties of circuits and utilized the Golgi technique, as well as other silver-impregnation methods, for detecting degenerating neurons following restricted lesions (Wallerian degeneration). The Golgi silver impregnation technique enables the description of the fundamental features of local connectivity (Figure 8). Major limitations of this technique, however, are that neuron impregnation is nonspecific and that it does not enable the visualization of myelinated axons and is therefore inadequate to study the long-distance connections of the brain. The development of a silver-staining method for selective labeling of degenerating axoplasm by Nauta and Gygax in the early 1950s constituted a major advance for the mapping of the brain. Although these physicochemical techniques have largely been replaced by modern tracing methods that are dependent on the active, anterograde and retrograde, transport processes of living neurons, they are still of interest for postmortem studies of the connectivity of the human brain in patients who suffered focal injury.

A number of substances dependent on active intraneuronal transport mechanisms have been used to define the connectivity of the central nervous system. These substances are injected into the brain through glass micropipettes, either by simple injection (with a syringe or a pressure injection system) or by iontophoresis (passing a low electric current through the pipette containing electrically charged tracer molecules). Following a survival period, the experimental animal is killed, its brain cut in thin sections with a microtome, and the tracer visualized in tissue sections. Most tracers are transported preferentially in one direction, either anterogradely from the cell body along the axon to the axon terminals or retrogradely

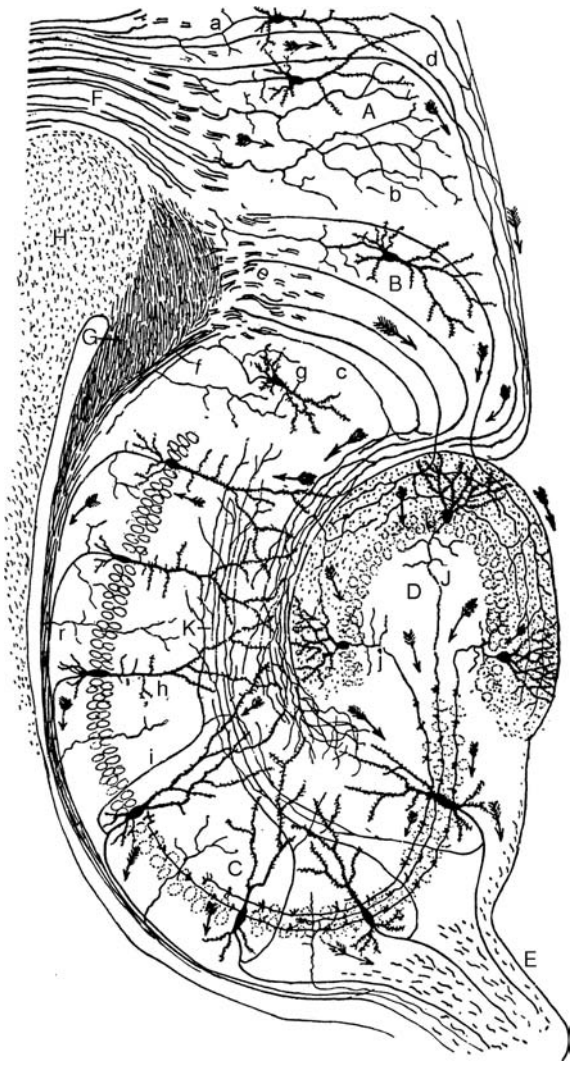

Figure 8 Diagram illustrating the structure and connections of the hippocampus drawn by Ramón y Cajal. A, retrosplenial area; B, subiculum; C, Ammon's horn; D, dentate gyrus; E, fimbria; F, cingulum; G, angular bundle or dorsal hippocampal commissure (crossed temporoammonic path); H, corpus callosum; K, recurrent collaterals from pyramidal cells to the stratum lacunosum of Ammon's horn (Schaffer collaterals); a, axon entering the cingulum; b, cingulum fibers ending in the retrosplenial area; c, fibers of the perforant or direct temporoammonic path; d, perforant fibers of the cingulum; e, plane of dorsal perforant path fibers; g, subicular cell; h, pyramidal cells in field CA1 (regio superior) of Ammon's horn; i, ascending (Schaffer) collaterals of large pyramidal cells; r, collaterals of alvear fibers. Reproduced from Ramón y Cajal S (1995) *Histology of the Nervous System of Man and Vertebrates*. New York: Oxford University Press, with permission.

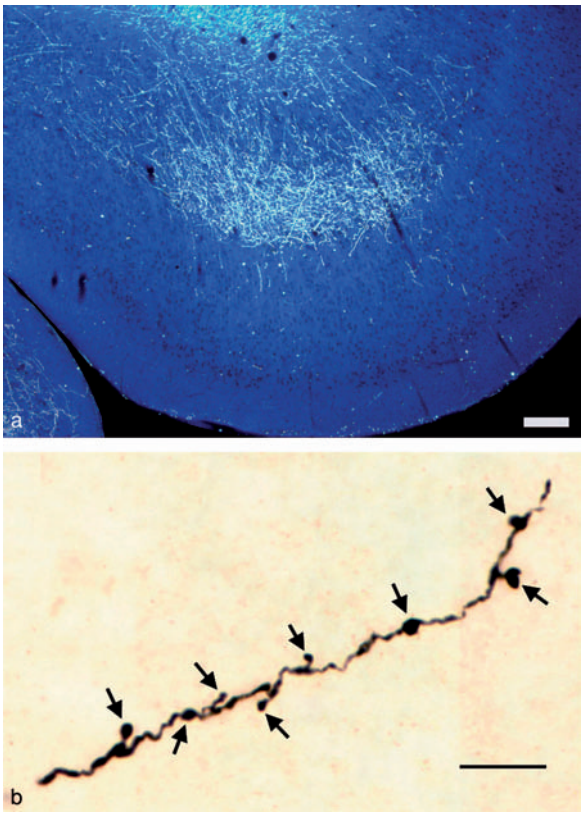

Figure 9 (a) Low-magnification, dark field photomicrograph of terminal labeling and fibers in the deep layers of the monkey entorhinal cortex following a PHA-L anterograde tracer injection in the CA1 field of the hippocampus. (b) High-magnification photomicrograph of a PHA-L-labeled fiber located in the superficial layers of area TE (following tracer injection in the amygdala). The arrows point to varicosities along the fiber. Scale bar = 250 μm (a) and 25 μm (b). Reproduced from Freese JL and Amaral DG (2006) Synaptic organization of projections from the amygdala to visual cortical areas TE and V1 in the macaque monkey. *Journal of Comparative Neurology* 496(5): 655–667, with permission.

from the terminal arborizations to the neuronal soma. The first anterograde tracer developed that was selectively taken up by neuronal cell bodies, and not by axons traveling through the injection site, was based on the uptake and transport of radioactive amino acids, typically a mix of ^{3}H-leucine and ^{3}H-proline, which are incorporated into proteins and transported toward the axon terminals. The radioactively labeled proteins are then detected by autoradiographic techniques. Many other tracers that can be detected by immunohistochemistry have been developed, enabling a very sensitive and precise visualization of the axonal arbors and synapses of the neurons that have incorporated the tracer. Among those, PHA-L, the plant lectin *Phaseolus vulgaris-leucoagglutinin*, is perhaps one of the most effective and selective anterograde tracing substance (**Figure 9**). Biotinylated dextran amine as well as other dextran amines marked with various fluorophores are also commonly used as anterograde tracers. They are taken up by cell bodies and not by fibers of passage, and they are transported toward the axon terminals; however, some retrograde transport is observed in certain conditions. Other substances are used specifically for retrograde labeling. HRP was the first retrograde label to be developed in the 1970s, and many modifications have been made throughout the years to

improve its uptake and transport by using HRP conjugated to various substances. Fluorescently labeled dyes and cholera toxin subunit B marked by different fluorophores are now commonly used as more specific retrograde tracers, especially in double-labeling studies of the divergence of fiber tracts originating from a defined brain area.

Although autoradiography and HRP tracing methods were originally developed to trace monosynaptic connections, these compounds can also cross the synaptic cleft and thus be transferred from neuron to neuron. These tracers, however, have not proven efficient enough to reliably trace multineuronal pathways because they are diluted at each synapse they cross. Alternate methods, such as attenuated live viruses, have therefore been developed for this purpose. Neurotropic viruses are particularly effective because the tracer multiplies itself in every neuron to which it is transported and can therefore label a series of neurons connected to each other, forming a defined neuronal circuitry.

Most of our understanding of the detailed connectivity of the human brain has been derived from experimental studies conducted in other animals using the methods described previously. However, lipophilic substances, such as the carbocyanins DiI or DiO, can be used to verify this information in postmortem human preparations. Dye crystals are deposited in a defined area of the brain. They are not transported actively but diffuse a few millimeters within the lipid membrane of neurons and can therefore be used to study short-distance fiber tracts in the human brain.

The full potential of neuroanatomical tracers is still being established by researchers who combine these techniques with modern molecular biology, electrophysiology, and microscopy techniques to decipher the complexity of brain connectivity.

Electron Microscopy

The study of the finest details of the structure of the nervous system, such as synapses, requires the use of an electron microscope, whose resolution of approximately 1 nm goes well beyond that of light microscopes (200 nm). This technique requires the tissue to be extremely well fixed, typically by adding glutaraldehyde to the fixative solution. The tissue is also treated with osmium tetroxyde to make it electron dense and then embedded in plastic for cutting. Semithick sections of a few micrometers are first cut and stained for examination at the light microscope level. These preparations are then cut into thin sections of approximately 10 nm on ultramicrotomes and further stained with lead and uranyl salts for examination in the transmission electron microscope. Immunohistochemical techniques using antibodies marked with electron-dense markers can be used to study specific components or circuits of the nervous system (Figure 10).

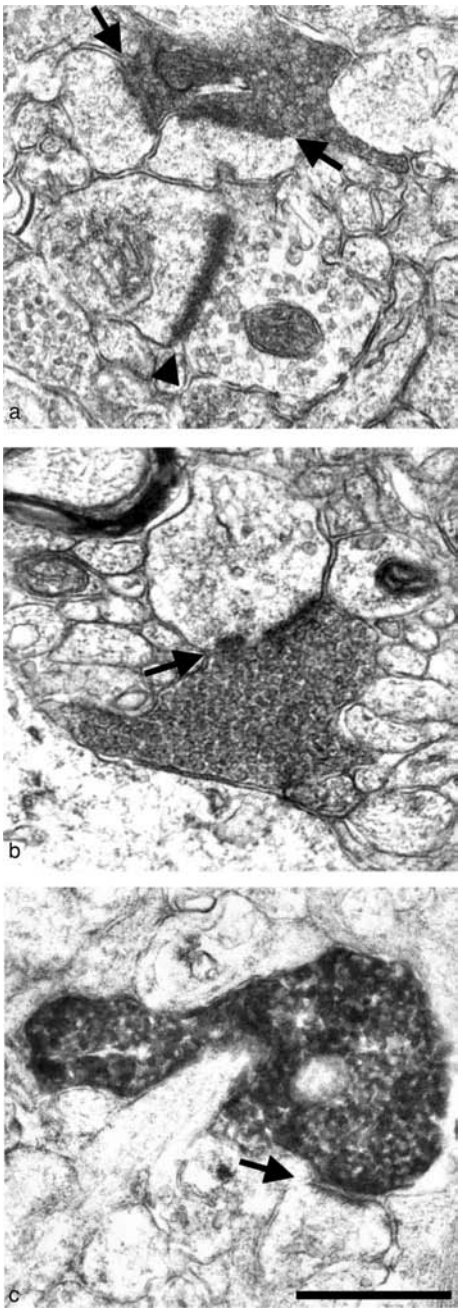

Figure 10 Electron micrograph of PHA-L-labeled varicosities forming asymmetric axospinous synapses in (a) the superficial layers of area TE, (b) the deep layers of area TE, and (c) the superficial layers of area V1 of the monkey. Arrows point to the postsynaptic density. For comparison, the arrowhead in a points to an asymmetric synapse formed by an unlabeled presynaptic bouton. Scale bar = 0.5 μm (a–c). Reproduced from Freese JL and Amaral DG (2006) Synaptic organization of projections from the amygdala to visual cortical areas TE and V1 in the macaque monkey. *Journal of Comparative Neurology* 496(5): 655–667, with permission.

Live Imaging Techniques

On the other end of the spectrum, new neuroanatomy techniques have been developed to study the organization of the nervous system in the living organism. Magnetic resonance imaging (MRI) techniques have been developed to study the connectivity of the brain. In experimental animals, MRI-detectable paramagnetic substances such as manganese (Mn^{2+}) are injected into the brain and the distribution of the tracer can then be evaluated after different delays. Transsynaptic transport of manganese in the central nervous system can be used to delineate multisynaptic networks in the living animal. Similarly, diffusion tensor imaging has also provided an invaluable tool to studying neuronal connectivity in the living human brain. This technique relies on the detection of anisotropic water diffusion across tissue (i.e., the alignment of water molecules along fiber tracts). Sophisticated algorithms can reconstruct fiber bundles originating from a particular point within the brain. Nevertheless, despite the advances of live imaging techniques, postmortem evaluation of neuroanatomical tracers injected into experimental animals remains the most precise and reliable mechanism for studying the detailed structural organization of the connectivity of the mammalian nervous system.

See also: Autonomic Nervous System: Neuroanatomy; Cognitive Neuroscience: An Overview; Diffusion Tensor Imaging (DTI); Fluorescence Microscopy in the Neurosciences; History of Neuroscience: Early Neuroscience; Microarray use for the Analysis of the CNS; Perfusion MRI.

Further Reading

Altemus KL, Lavenex P, Ishizuka N, and Amaral DG (2005) Morphological characteristics and electrophysiological properties of CA1 pyramidal neurons in macaque monkeys. *Neuroscience* 136(3): 741–756.

Cowan WM, Gottlieb DI, Henderson AE, Price JL, and Woolsey TA (1972) The autoradiographic demonstration of axonal connections in the central nervous system. *Brain Research* 37: 21–55.

Eriksson PS, Perfilieva E, Bjork-Eriksson T, et al. (1998) Neurogenesis in the adult human hippocampus. *Nature Medicine* 4(11): 1313–1317.

Freese JL and Amaral DG (2006) Synaptic organization of projections from the amygdala to visual cortical areas TE and V1 in the macaque monkey. *Journal of Comparative Neurology* 496(5): 655–667.

Gerfen CR and Sawchenko PE (1984) An anterograde neuroanatomical tracing method that shows the detailed morphology of neurons, their axons and terminals: Immunohistochemical localization of an axonally transported plant lectin. *Phaseolus vulgaris leucoagglutinin (PHA-L). Brain Research* 290: 219–238.

Nissl F (1894) Ueber eine neue Untersuchungsmethode des Centralorgans zur Feststellung der Localisation der Nervenzellen. *Neurologisches Centralblatt* 13: 507–508.

Peters A, Palay SL, and Webster Hd (1990) *The Fine Structure of the Nervous System.* New York: Oxford University Press.

Ramón y Cajal S (1909, 1911) *Histologie du Système Nerveux de l'Homme et des Vertébrés.* Paris: Maloine.

Ramón y Cajal S (1995) *Histology of the Nervous System of Man and Vertebrates.* New York: Oxford University Press.

Saleem KS, Pauls JM, Augath M, et al. (2002) Magnetic resonance imaging of neuronal connections in the macaque monkey. *Neuron* 34(5): 685–700.

Saper CB and Sawchenko PE (2003) Magic peptides, magic antibodies: Guidelines for appropriate controls for immunohistochemistry. *Journal of Comparative Neurology* 465(2): 161–163.

West MJ (1993) New stereological methods for counting neurons. *Neurobiology of Aging* 14(4): 275–285.

Zaborszky L, Wouterlood FG, and Lanciego JL (2006) *Neuroanatomical Tract-Tracing.* New York: Springer.

Neurodegeneration in Psychiatric Illness

R J Schloesser, M Lan, and H K Manji, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Definitions and Research Limitations

In this article, the term 'neurodegeneration' is defined as chronic and progressive loss of structure or function of neurons. Classic neurodegenerative disorders such as Alzheimer's, Huntington's, and Parkinson's diseases display gross tissue or cell loss, usually including glial reactions such as scarring. For the psychiatric disorders discussed in this article (schizophrenia, bipolar disorder, and unipolar depression), there is no evidence pointing toward similar large-scale cell loss, gliosis, or severe morphologic alterations, however. Other common histopathological hallmarks of neurodegenerative disorders, such as inclusion bodies, neurofibrillary tangles, or amyloid plaques, are also not present in these disorders. Neurodegeneration in psychiatric disorders is therefore likely to be of comparatively limited scope and severity. The fact that degenerative processes are less apparent, however, does not imply that they are of low importance. Further research determines their importance in the development of the symptoms of the disorders.

Schizophrenia, bipolar disorder, and major depression are defined clinically by distinct behavioral and psychosocial symptoms, although other abnormalities, such as endocrine, cognitive, and motor dysfunction, usually accompany these symptoms. So far, no biological markers for the disorders have been identified, and the diagnoses depend on symptomatology only. It is therefore likely that several disorders with a distinct pathophysiology are subsumed under one clinical category, leading to unfortunate problems in research data interpretation.

Evidence for Neurodegeneration in Psychiatric Illness

Numerous observations from different research perspectives support the idea of neurodegenerative processes in the pathophysiology of these major psychiatric illnesses. The disorders have a characteristic progressive course. Imaging and postmortem data show structural and functional abnormalities in brains of psychiatric patients. Preclinical experiments on animal and cell culture models display neuronal death in response to stimuli that model risk factors for psychiatric disorders. In addition, in some cases drugs with proven efficacy in the treatment of these disorders exert neuroprotective and neurotrophic properties both *in vivo* and *in vitro*.

Clinical Course

In classic neurodegenerative disorders, a deteriorating illness course with increasing severity and frequency of symptoms is common. To various degrees, the clinical course of schizophrenia, bipolar disorder, and major depression also displays a progression. The following section illustrates general clinical characteristics of psychiatric disorders and highlights evidence for illness progression over a patient's lifetime.

Schizophrenia Schizophrenic disorders present with episodes of fundamental distortions of thinking and perception as well as blunted or inappropriate affect. The symptoms consist of positive ones, comprising hallucinations and delusions, and negative ones, such as flat affect, apathy, and poverty of speech. Typically, the onset of the disorder occurs in late adolescence or young adulthood, manifesting with an episode of predominantly positive symptoms. Prodromal symptoms, such as a marked social withdrawal, often occur before and lead up to a 'psychotic break.' Two-thirds of patients experience more than one episode, and in at least one-third of cases, the disorder becomes chronic, lasting throughout the patient's life.

Emil Kraepelin first noted a progressive nature to the symptoms of schizophrenia, labeling the disorder 'dementia praecox,' implying a progressive deterioration leading to severe cognitive dysfunction. Many longitudinal studies have corroborated his early observations and show an increased duration and number of periods of active positive symptoms with time. Negative symptoms also progress, often becoming more clinically prevalent than positive symptoms. The disorder also leads to severe cognitive dysfunction, thought disorganization, and motor symptoms as it progresses. Early initiation and maintenance of antipsychotic treatment, in particular with atypical antipsychotics, helps to enhance clinical outcomes and improves cognitive ability, suggesting a protective effect of drug treatment rather than mere symptom suppression.

The buildup of symptoms to a psychotic break and the worsening of symptoms with time are consistent with a neurodegenerative course to the disorder.

Bipolar disorder and major depression Episodes of debilitating depressed mood, decreased interest and pleasure, neurovegetative symptoms including changes

in diet and sleep, and thoughts of death or suicidal ideation sustained over a prolonged period of time characterize a major depressive episode. In major depressive disorder, these depressive states are often episodic and frequently recur throughout the patients' lives, with periods of relative remission between episodes. A nonrecurring major depressive episode can also occur, however.

Patients with bipolar disorder usually experience episodes of clinical depression also, but the presence of episodes of mania define the disorder. Manic symptoms are in many ways the opposite of what is experienced during depressive episodes. Patients have racing thoughts, increased self-esteem or grandiosity, distractability, and increased goal-directed and risk-taking behavior. Patients frequently experience an alternation between depression and mania over time, with each episode lasting weeks or even months. Between episodes, a resolution of many symptoms typically occurs to some extent. Similar to schizophrenia, the onset of symptoms typically occurs in late adolescence or early adulthood, and the disorder often remains chronic and lifelong.

In the past, both major depression and bipolar disorder were thought to have a *restitio ad integrum,* or complete remission of symptoms between episodes. They were also considered to have a good chance for patient recovery. Recent evidence, however, suggests that for a large percentage of patients, the outcome is poor, with a high rate of chronicity. Also, residual depressive or manic symptoms are common between episodes, and the likelihood of developing other chronic illnesses, such as cardiovascular disease or diabetes, is higher. Over the course of the disorder, episode frequency increases, with the risk of recurrence positively correlated with the number of prior episodes. Early episodes often occur when the patient experiences severe psychosocial stress, whereas later episodes arise more autonomously. With time, patients often have a progression of residual symptoms, cognitive impairment, psychosocial disability, and diminished well-being as well, illustrating progressive clinical decline. Some evidence shows that mood stabilizers, medications for bipolar disorder, can prevent or decrease the occurrence of further episodes, suggesting a neuroprotective effect of these drugs. Taken together, these characteristics demonstrate a continuous, progressive etiology.

The clinical course for bipolar and unipolar disorders is consistent with a neurodegenerative etiology. In addition, the major affective disorders increase the risk of developing several classic neurodegenerative diseases, including Parkinson's and Alzheimer's disease. A possible explanation for this is a higher susceptibility of neurons in affective disorders, likely caused by divergent or common degenerative pathophysiology. Notably, however, schizophrenia in itself does not increase the risk for neurodegenerative disorders.

Neuroimaging

Clinical studies using magnetic resonance imaging (MRI) and computer-assisted tomography (CT) have identified structural tissue alterations in people with schizophrenia, bipolar disorder, and unipolar depression when compared with matched controls. In schizophrenia, volumetric increases in the third and lateral ventricles have been robust and consistent. These increases suggest possible periventricular tissue deficits in schizophrenia. Other regions showing changes are the medial temporal lobe area, including the hippocampus and subregions of the prefrontal cortex. Less consistent changes were reported in subcortical regions. These deficits are consistent with a neurodegenerative process. Several longitudinal studies demonstrate an increase in severity of these deficits with time, also consistent with a neurodegenerative course. Structural deficits found in first-onset patients suggest the changes may occur before the onset of symptoms. The latter finding would be consistent with the initiation of neurodegeneration prior to first onset.

Affective disorders are also associated with volumetric changes. Patients with bipolar disorder have increases in third and lateral ventricle volumes, although this finding has been less consistent than in people with schizophrenia. Findings of decreased volume and blood flow in the subgenual prefrontal cortex of patients with bipolar and unipolar depression have complemented postmortem findings showing deficits in that region. The subgenual prefrontal cortex is of functional interest because of its role in executive function and modulation of emotions, important processes disrupted in acute affective episodes. The cerebellar vermis and the hippocampus and amygdala also show decreased volumes in bipolar and unipolar depression. These structural findings are consistent with a neurodegenerative process.

By means of magnetic resonance spectroscopy (MRS), certain metabolites can be quantified in particular regions of the brain tissue *in vivo*. *N*-acetyl aspartate (NAA), a major metabolite quantifiable with this technique, serves as a marker for neuronal viability. It increases in the brain with neurogenesis and decreases with neurodegeneration or damage. Neural damage causes a decrease in the metabolite even before structural loss can be noted. In people with either schizophrenia or bipolar disorder, NAA

decreases in various brain regions, including the orbital and prefrontal cortex, compared with controls, which is consistent with a neurodegenerative process in the disorders. Shigher frequency in the brains of patients with bipolar, unipolar, or schizophrenic disorders compared with matched control participants, consistent with a neurodegenerative process. These small lesions can occur after neural damage such as demyelination, astrogliosis, or microvascular infarcts. They also increase in frequency in the normal aging process.

These clinical imaging findings could all conceivably be caused by subtle neurodegeneration. None of the findings constitute definitive proof for degeneration, however, and are consistent with other hypotheses as to the etiologies of the disorders. The subtle morphological changes, for example, occur in both neurodegenerative and nonneurodegenerative neurological disorders. Structural brain deficits also occur with chronic stress and high glucocorticoid levels in animals (see the section titled "Unipolar depression and bipolar disorder"), calling into question the role stress plays in their development. It is unknown whether the structural and MRS changes are state dependent (occurring with symptoms only) or trait specific (occurring at all times with the disorders), a question imperative to identify possible irreversible neurodegeneration. Except for ventricular enlargement in schizophrenia, most imaging findings warrant further confirmation, and there is a need for further longitudinal studies. More sensitive and accurate imaging modalities may provide better data concerning these early findings of neural pathology.

Postmortem Neuropathological Findings

Postmortem brain tissue of psychiatric patients shows several distinct morphological alterations. Such changes in cell numbers, densities, and cytomorphological characteristics appear mostly in areas of the brain important for emotion perception and control: the limbic system, anterior cingulate gyrus, and the prefrontal and orbitofrontal cortex.

Schizophrenia Early postmortem studies of schizophrenic patients suggested decreases in brain weight and smaller cortical volumes. Although some studies examining cortical volumes did not reveal significant differences, several more-recent studies, using advanced stereological analysis, report decreased gray matter volume. Neuronal abnormalities have been reported in various brain regions with schizophrenia. Some reports show fewer neurons in the thalamic nuclei, notably in the mediodorsal nucleus and pulvinar. Other reasonably well-replicated findings include smaller cell bodies of pyramidal neurons in the hippocampus and neocortex, with fewer dendritic spines and reductions of dendritic arborizations. These findings could be due to less extensive or less active axodendritic trees and correlate with reported decreases in molecular markers for neuronal plasticity such as microtubule-associated protein 2 and spinophilin. Decreased densities of neurons, especially nonpyramidal cells or interneurons in the anterior cingulate cortex and hippocampus, have also been reported. Decreased expression of genes associated with synaptic plasticity, such as synaptophysin, SNAP-25, and complexin-II, support these data. Molecular changes that indicate a general cellular pathology have also been reported in postmortem tissue. Several studies have identified changes in mitochondrial molecules related to oxidative stress and energy metabolism altered in the postmortem brain tissue of patients with schizophrenia. Abnormalities in myelination have also been reported, consistent with oligodendrocyte dysfunction. Although postmortem studies cannot completely control for differences in the agonal effects of death or the postmortem interval, these findings are consistent with a degenerative neuropathology of the disorder.

Unipolar depression and bipolar disorders Gross morphological examination of the postmortem brain from major depressive or bipolar patients does not expose dramatic changes such as gross tissue loss or the presence of focal lesions. However, specific brain regions, including the orbitofrontal cortex, dorsolateral prefrontal cortex, anterior cingulate cortex, and hippocampus, show changes in cell number and density and in cell body size. Pyramidal neurons were found to have reductions in cell body size in specific layers of the cortex in areas associated with emotion perception and control. This reduction indicates a general atrophic process likely involving the dendritic tree. It is possible that these neurons are less resilient toward degenerative stimuli and thus show more atrophy than the other cells in the tissue. Decreased pyramidal neuron densities have been reported in older patients with a lifelong history of depression, corroborating the idea of increased susceptibility in these cells that ends in neuronal death. It is interesting that the density of γ-aminobutyricacidergic interneurons, which are particularly vulnerable to apoptotic stimuli, is lower in cortical and hippocampal areas of patients with major affective disorders. Postmortem brain tissue from bipolar patients also displays myelination changes consistent with oligodendrocyte dysfunction and mitochondrial changes consistent with apoptosis and cell death. These molecular findings are also consistent with a neurodegenerative process.

Classic neurodegeneration and depression As discussed above, unipolar depression is associated with a higher risk for other recognized neurodegenerative disorders. Histopathological features characteristic of these disorders, such as gliosis, neurofibrillary tangles, plaques, or inclusion bodies, are absent in psychiatric disorders. However, there is evidence for higher levels of both plaques and tangle formation in the hippocampus of patients with Alzheimer's disease and a lifelong history of depression compared with patients with Alzheimer's disease but without a history of depression.

Preclinical Data

Preclinical studies into the mechanism of psychiatric disorders and their treatments are hampered by a lack of accurate models. These limitations are due to the still modest insights into genetic and biological underpinnings of the disorders, as well as a lack of the complexity of associated psychosocial function in other species that is characteristic of the disorders in humans. Certain aspects of etiological importance in these disorders, such as chronic stress or hyperdopaminergic states, can be modeled and studied in both animal models and cell cultures, however. In addition, *in vivo* and *in vitro* experiments can provide data on cellular and intracellular effects of drugs with known efficacy in the treatment of psychiatric disorders. Such experiments provide data showing neuroprotective effects of several psychiatric medications as well as their ability to increase neurotrophic signaling and neurotrophic function on the molecular and cellular level.

Animal Models for Psychiatric Disorders

Psychosis and mania Several psychotropic drugs elicit symptoms similar to those seen in psychiatric disorders. Psychostimulants such as amphetamine and its derivates can induce mania and schizophrenia-like symptoms in humans. These compounds are indirect agonists of monoamine neurotransmitters; they cause the increased release of the transmitters and block their reuptake from the synapse. In animals, amphetamine also induces behavior resembling mania and psychosis, such as hyperactivity, stereotypic behavior, and deficits in prepulse inhibition. It is interesting that these drugs have been found to be neurotoxic and induce degeneration of monoamine nerve terminals, especially when given chronically. In addition, they are capable of eliciting apoptosis and mitochondrial dysfunction in neurons. Thus, pharmacologic manipulations that model the symptoms of both bipolar disorder and schizophrenia also have a neurotoxic effect, consistent with a neurodegenerative component to the disorders' etiologies.

Depression In humans, the likelihood of developing a depressive disorder increases with the experience of negative life events. Stress, and the accompanying activation of the hypothalamic–pituitary–adrenal axis, is thus considered to be a factor contributing to the course and severity of the disorder. This element of the disorder can be effectively modeled in animals such as rodents and primates. In fact, acute and chronic stressors can induce depressive-like behavior such as learned helplessness and anhedonia in animal models. For example, chronically restraining rodents by limiting their movement can both induce the physiologic stress response and elicit depressive-like behaviors. What is more, antidepressant treatment prevents these behavioral changes. In these stress-induced depression models, animals have subtle brain atrophy in certain areas in a manner similar to that found in mood-disordered and schizophrenic patients. This finding implies that the patients' cell loss and atrophy, indicative of a neurodegenerative process, could potentially be involved in the pathophysiology of the disorders. High levels of glucocorticoids that accompany stress can also lead to neurodegeneration and can weaken the capacity of neurons to survive coincident neurotoxic insults. This effect has been found especially in the hippocampus. In this way, stress has been shown to cause a neurotoxicity, implying that stress may either cause or exacerbate a neurodegenerative process in the disorders.

High glucocorticoid levels can also lead to a decline in proliferation of neural progenitor cells in the dentate gyrus of the adult hippocampus, whereas antidepressants and the mood stabilizer lithium, as well as atypical antipsychotics, increase proliferation of adult hippocampal progenitor cells. Antidepressants block the decrease in adult neurogenesis induced by stress and high glucocorticoid levels. A decrease in neuroproliferation with stress and an enhancement with psychiatric medications implies that neurotrophic processes that can cause these changes may be involved in the pathophysiology of the disorder and the mechanism of its treatment. This finding is consistent with a neurodegenerative etiology of the disorder.

Neuroprotective Effects

In vivo and *in vitro* experiments show neuroprotective properties of the mood stabilizers, lithium, and valproate. In cell culture models, they show protective capacity against several damaging environments, such as glutamate toxicity and low potassium. Furthermore, mood stabilizers are neuroprotective against many damaging insults in animal models. Both medications limit the damage from ischemia in the brain, for example. These two medications have distinct molecular structures, and this common function thus might underlie their shared mood-stabilizing ability.

Several different types of antipsychotics are efficient in the treatment of psychotic symptoms of schizophrenia and mood disorders. One class of compounds, the phenothiazine derivatives, including clozapine, olanzapine, and quetiapine, has shown neuroprotective properties against several insults both *in vitro* (including oxidative stress) and *in vivo* (including A beta toxicity). Risperidone, a member of the benzisoxazole derivative family, also has neuroprotective properties. These two classes of compounds are 'atypical' antipsychotics with complicated pharmacological profiles. 'Typical' antipsychotics such as haloperidol are thought to act primarily by inhibiting dopamine receptors and do not show neuroprotective properties, however.

Neurotrophic Effects

The mood stabilizers lithium and valproate, and atypical antipsychotics and antidepressants, induce biochemical changes consistent with neurotrophic effects. Lithium and valproate both increase the mRNA and protein levels of bcl-2, for example, a protective compound that stabilizes the cell's mitochondrial membrane and protects against apoptosis. Both lithium and valproate also increase the levels of bcl-2's chaperone protein BAG-1. Lithium and valproate have also been shown to stimulate the extracellular signal-regulated kinase (ERK) and mitogen-activated protein kinase (MAPK) pathways, cell survival, and growth signaling, consistent with a neurotrophic effect. Similarly, atypical antipsychotics induce the ERK and MAPK pathways as well as the phosphoinositide-3 kinase (PI3K)/protein kinase B (Akt) pathway, also consistent with a neurotrophic effect. A number of antidepressant therapies, including the selective serotonin reuptake inhibitors, the norepinephrine selective reuptake inhibitors, the monoamine oxidase inhibitors, atypical antidepressants, and electroconvulsive seizures, also increase neurotrophic cascades, indicated by increased levels of brain-derived neurotrophic factor (BDNF) in the animal brain after chronic treatment. BDNF in turn increases intracellular neurotrophic cascades and thereby modulates cell survival and neuronal outgrowth in the brain.

The protective capability and the neurotrophic effects of these compounds indicate that processes associated with cellular survival are stimulated with the medications. This effect is consistent with a neurodegenerative etiology to the disorders that is reversed or corrected with the drugs. The 'typical' antipsychotics are the exception to this paradigm, although they could potentially exert their effects through an independent mechanism.

Summary

Evidence from the clinical course of the disorders, molecular and cellular postmortem studies, *in vivo* imaging of patients, and preclinical studies modeling the disorders and their treatments suggests that a neurodegenerative component exists in the pathophysiology of the psychiatric disorders. These data are neither exhaustive nor definitive, but this perspective may be useful in understanding the etiology of these disorders and in the quest for novel treatments.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Antipsychotic Drugs; Apoptosis in Neurodegenerative Disease; Bipolar Disorder; Cognition: Neuropharmacology; Cognitive Deficits in Schizophrenia; Cognitive Dysfunction in Psychiatric Disorders; Depression and the Brain; Mood Stabilizers; Psychotherapeutic Approaches to Psychiatric Disorders; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Schizophrenia: Genetics.

Further Reading

Charney DS and Nestler EJ (2004) *Neurobiology of Mental Illness*. Oxford: Oxford University Press.

Drevets WC (2000) Functional anatomical abnormalities in limbic and prefrontal cortical structures in major depression. *Progress in Brain Research* 126: 413–431.

Drevets WC (2000) Neuroimaging studies of mood disorders. *Biological Psychiatry* 48(8): 813–829.

Goodwin FK and Jamison KR (2007) *Manic-Depressive Illness: Bipolar and Recurrent Depression*. New York, NY: Oxford University Press.

Harrison PJ and Weinberger DR (2005) Schizophrenia genes, gene expression, and neuropathology: On the matter of their convergence. *Molecular Psychiatry* 10(1): 40–68.

Lieberman JA (1999) Is schizophrenia a neurodegenerative disorder? A clinical and neurobiological perspective. *Biological Psychiatry* 46(6): 729–739.

Phillips ML, Drevets WC, Rauch SL, and Lane R (2003) Neurobiology of emotion perception I: The neural basis of normal emotion perception. *Biological Psychiatry* 54(5): 504–514.

Phillips ML, Drevets WC, Rauch SL, and Lane R (2003) Neurobiology of emotion perception II: Implications for major psychiatric disorders. *Biological Psychiatry* 54(5): 515–528.

Rajkowska G (2000) Postmortem studies in mood disorders indicate altered numbers of neurons and glial cells. *Biological Psychiatry* 48(8): 766–777.

Rajkowska G (2003) Depression: What we can learn from postmortem studies. *Neuroscientist* 9(4): 273–284.

Schloesser RJ, Chen G, and Manji HK (2007) Neurogenesis and neuroenhancement in the pathophysiology and treatment of bipolar disorder. *International Review of Neurobiology* 77: 143–178.

Zarate CA Jr., Singh J, and Manji HK (2006) Cellular plasticity cascades: Targets for the development of novel therapeutics for bipolar disorder. *Biological Psychiatry* 59(11): 1006–1020.

Neuroeconomics: History

P W Glimcher, New York University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

Within the neurobiological community this rapprochement had its origins in the early 1990s. At that time, significant progress was being made toward understanding the neurobiological basis of simple sensory–motor decision making in nonhuman primates. The catalyst for that progress was Newsome and colleagues' work on the relationship between activity in the middle temporal area of visual cortex (area MT) and perceptual decisions. The goal of this work was to relate the activity of single neurons in area MT to the perceptual decisions of monkeys. They accomplished this using a formal mathematical model that defined an optimal perceptual decision-making process, a normative psychophysical theory, which was rooted in the theory of signal detection that had dominated behavioral studies of perception for several decades. As other research groups began to extend this work, studying more complicated and less perceptual forms of decision making, it became clear that the normative models of psychophysics that define optimal performance in perception could not be extended to situations in which choices relied on complex functions of nonsensory variables: a point made decades earlier, but often overlooked, by the mathematical psychologists Duncan Luce and Patrick Suppes.

At the same time, cognitive psychologists employing functional magnetic resonance imaging (fMRI) also became interested in understanding the neural basis of decision making. In the late 1980s and early 1990s, a number of early studies were conducted that also demonstrated the need for a unified theory of choice in the neurobiological setting.

This gap between theory and experiment led a number of physiologists and psychologists to begin to explore alternative models for understanding what constitutes efficient decision making. What quickly became clear was that the academic discipline of economics, although far behind neuroscience experimentally, possessed an extremely well-developed corpus of elegant normative models that could serve the same role for neuroscientific studies of decision making that signal detection theory had served for neuroscientific studies of perception.

At the same time that this gradual realignment was taking place in theoretical neuroscientific circles, economics was also experiencing the beginnings of an important sea change. Until the early 1970s, economics had developed as a theoretical and analytic discipline that lacked an experimental tradition. The goal of economics, since the neoclassical revolution of the first half of the twentieth century, had been to relate external-world variables to the choices made by individuals, or groups of individuals, using purely theoretical approaches. Within neoclassical economics, only preexisting data about choices in real-world situations served as acceptable dependent variables for tests of theories, and unsurprisingly many dominant theories were only weakly tested by these existing data sets. In the 1960s, a number of empirical demonstrations raised serious challenges to this approach, suggesting that the existing corpus of normative theory was not as predictive of choice as had been hoped. Partly in response to these challenges and in response to others, in the 1970s a number of economic scholars began to develop a homegrown experimental methodology to empirically explore how theories could explain behavior in laboratory, or artificial, markets and games, and how the trading behavior of human subjects was influenced by changes in the rules governing the operations of those markets.

Together, these two groups of scholars laid the foundations for what was essentially a new economic approach, one in which hypotheses could be posed and addressed experimentally. Like those that had come before them, however, they also believed strongly that the goal of economics was to predict and explain choice. As a result, they argued that while studies of hidden psychological or neurobiological processes might be of interest to scholars studying the algorithms by which choice was accomplished, these biological studies would add little in the way of predictive power to models of choice itself, the object of economic study.

The rise of fMRI in the 1990s, however, changed that assumption for some economists. During that period, brain scanners became relatively cheap to use and widely distributed in the academic community. The popular press also showed a tremendous interest in brain scanning, and a constant flood of articles describing 'how the brain worked' appeared in the general press. The result was an interest, particularly among younger economists, in the possibility that measurements of neural state could be used to test or refine economic models.

The two communities, neuroscience and economics, thus came toward a union from two very different directions. Neurobiologists largely turned toward economics as a source of normative theory while economists turned toward neurobiology for new measurements that could be used to test or constrain existing theory. Both of these approaches were, and remain, controversial in many circles. Neurobiologists opposed to the use of economic theory argued that the complex normative models of economics would be of little value for understanding the behavior of real humans and real animals. In a similar vein, many economists argued that algorithmic-level studies of decision making were unlikely to improve the predictive power of normative models in any meaningful way.

Critical First Steps

Despite the vocal criticism of mainstream elements in both communities, these social and technological trends led to a convergence of interest among a group of economists, neuroscientists, and cognitive psychologists in the late 1990s. This community of scholars, who began to call themselves neuroeconomists around 2002 or 2003, aggregated around a small set of meetings and papers during this period. Although the first explicitly neuroeconomic papers preceded the first interdisciplinary workshops by several years, these meetings were probably the critical step that allowed these social and natural scientists to develop a common language.

Probably the first of these interdisciplinary interactions was held in 1997 at Carnegie-Mellon University, organized by the economists Colin Camerer and George Loewenstein. After a hiatus of several years this was followed by two meetings in 2001, one held by the Gruter Foundation for Law at their annual meeting in Squaw Valley. At that meeting the Gruter Foundation chose to focus its workshop on the intersection of neuroscience and economics and invited several speakers active at the interface of these converging disciplines. The second meeting was focused more directly on what would later become neuroeconomics and was held at Princeton University. The meeting was organized by the neuroscientist Jonathan Cohen and the economist Christina Paxson. This meeting is often seen as the inception point for the present-day Society for Neuroeconomics. At that meeting, economists and neuroscientists met to explicitly discuss the growing convergence of these fields and to debate the value of such a convergence. There was, however, no consensus among the scholars present at that meeting that such a convergence would be desirable. In fact, when asked specifically at the end of the meeting what he thought of such a convergence, the internationally acclaimed economist Ariel Rubenstien (then at Princeton) concluded, essentially, that there was nothing in neuroscience that would change how he thought about economics.

Nonetheless, the Princeton meeting generated significant momentum and 2 years later (2003) a small invitation-only meeting that included nearly all of the active researchers in the emerging area was held on Martha's Vineyard and organized by Greg Berns of Emory University. This 3-day meeting marked a clear turning point in which a group of economists, psychologists, and neurobiologists began to identify themselves as neuroeconomists and began to explicitly shape the convergence between the fields. This led, the subsequent year, to an open registration meeting at Kiawah Island, South Carolina that was attended by 83 people in which the Nobel Laureate in economics Vernon Smith delivered the keynote address. At that meeting a decision was made to incorporate the existing group as an academic society that would sponsor an annual meeting that would serve as a focal point for neuroeconomics internationally. At that meeting Paul Glimcher was elected President of the Society for Neuroeconomics which held its first formal meeting the following year (2005) at Kiawah Island with a group of 114 people.

At the same time a group of European scholars undertook the establishment of a similar meeting in 2002. The first meeting of the European conference on neuroeconomics (ConNEcs) was held at the Westfalian Wilhelms University of Muenster in northern Germany. This was followed by a second meeting in 2004, establishing neuroeconomics as an international discipline.

Landmark Papers

Against this backdrop of meetings, a series of critical papers and books were emerging that shaped the interactions between these scholars and also served to communicate the goals of the emerging neuroeconomic community to the larger neurobiological and economic communities. Probably the first neurobiological paper to explicitly rest on a normative economic theory was Peter Shizgal and Kent Conover's 1996 review in *Current Directions in Psychological Science* entitled: 'On the neural computation of utility.' This was followed the next year by a related paper published in *Current Opinion in Neurobiology* entitled 'Neural basis of utility estimation' authored by Shizgal. The reason that these papers can be viewed as the first neuroeconomic papers is that they attempt to describe the neurobiological substrate for a behavioral choice using a normative theory rooted in neoclassical

economics. In these papers, Shizgal analyzed the results of studies of intracranial self-stimulation in rats using a type of utility theory related loosely to the standard expected utility theories of Savage and VonNeuman and Morgenstern. The papers argue that the choices an animal makes about whether or not to work for direct electrical activation of the medial forebrain bundle can be construed as an effort to maximize the animal's instant-to-instant utility. In their analysis, changes in the desirability of brain stimulation reward, for example, as a function of stimulation frequency, can be formally interpreted as changes in the utility of the electrical stimulation. Unlike standard theories of utility, however, Shizgal proposed that the expected utility of an action is perceived by the animal as the expected utility of that action divided by the sum of the expected utilities of all available actions. This particular formulation had its origin in the work of the psychologist Richard Herrnstein who proposed that many choices reflect this 'normalization' with regard to the value of other alternatives, a phenomenon he referred to as the matching law (1997).

This equation was introduced to self-stimulation studies about 5 years earlier by Shizgal's mentor C Randy Gallistel. In the early 1990s, Gallistel used Herrnstein's work to inspire quantitative choice-based experiments intended to cardinally scale the desirability of any given set of self-stimulation parameters. Shizgal's extension of this work is critical for the history of neuroeconomics because he retained Gallistel's goal of quantitatively measuring, or more precisely cardinally scaling, the desirability of a neurobiological manipulation while focusing on the idea that such a scale implies an underlying utility theory. What Shizgal's work did not do, however, was to fully incorporate the standard economic model, but rather a more normative version of Herrnstein's approach.

In 1999, this work was followed by a paper published by Michael Platt and Paul Glimcher (another student of Gallistel's) in *Nature* that argued quite explicitly for a normative utility-based analysis of choice behavior in monkeys. As they put it in that paper: "Neurobiologists have begun to focus increasingly on the study of sensory–motor processing, but many of the models used to describe these processes remain rooted in the classic reflex. ... Here we describe a formal economic–mathematical approach for the physiological study of the sensory–motor process, or decision-making." The paper goes on to argue for an approach to the study of decision making rooted in standard expected utility theory. At an experimental level, the paper demonstrates that the activity of single neurons in the posterior parietal cortex is a lawful function of both the probability and magnitude of expected rewards. This was significant because standard expected utility theory predicates choice on lawful functions of these same two variables. The paper, however, makes a misstep in its final examination of choice behavior. The authors examine a matching-law type behavior, which they interpret in terms of normative expected utility theory. This is problematic because there is no normative standard for the analysis of the classical matching-law. The result is a mixing of normative and nonnormative approaches that characterized many of the early neuroeconomic papers from neurobiologists.

At the same time that this paper appeared in print, the economists Colin Camerer, George Lowenstein, and Drazen Prelec began circulating in economic circles a manuscript by the name of 'Grey matters.' In this manuscript the authors also argued for a neuroeconomic approach, but this time from an economic perspective. Over the course of the preceding decade behavioral economists, a group of scholars interested in deviations of actual choice behavior from normative theory, had argued that the choices of humans often did not reflect optimal calculations that perfectly maximized utility. Instead, these behavioral economists had noted that, particularly under conditions in which choosers experience strong emotions, decisions typically deviate systematically from the predictions of standard normative models. What these three scholars argued was that these systematic deviations may reflect neurobiological constraints on decision making and that future neurobiological studies might serve to reveal and define these constraints that cause deviations of behavior from normative theory. What was striking about this argument, in economic circles, was that it proposed an algorithmic analysis of the physical mechanism of choice, a possibility that has been explicitly taboo in economic circles. Prior to the 1990s it was a completely ubiquitous view in economic circles that models of behavior, like utility theory, were 'as if' models: the model was to be interpreted as if utility was represented internally by the chooser. But critically it was irrelevant whether this was actually the case because the models sought to link options to choices not to make assertions about the mechanisms by which that process was accomplished. Camerer and colleagues argued against this view suggesting that deviations from normative theory should be embraced as clues to the underlying neurobiological basis of choice. In a real sense then, the economists turned to neurobiology for exactly the opposite reason that the neurobiologists had turned to economics. They embraced neuroscience as a principled complement to the axiomatic reasoning that characterized the core features of standard economic practice.

In sum then, some members of the neurobiological community had turned toward economics in an effort to embrace a normative theory that they were just beginning to understand, but a normative theory that had only limited predictive power. Members of the economic community had turned toward neuroscience in an effort to find an alternative to normative theory that could account for the choices of real individuals. Of course, these conflicting goals were well aligned; both groups were becoming poised to meet in the intersection of their disciplines.

At this point there was a rush by several research groups to perform an explicitly economic experiment that would mate these two disciplines in human choosers. Two groups succeeded in this quest in 2001. The first of these papers appeared in the journal *Neuron* and reflected a collaboration between the fMRI pioneer Hans Breiter, Shizgal, and the Princeton psychologist/economist Daniel Kahneman (who would win the Nobel prize for his contribution to behavioral economics the following year). That paper was loosely based on prospect theory, a nonnormative form of expected utility theory that guided much research in judgment and decision-making laboratories throughout the world. In that paper Breiter and colleagues manipulated the perceived desirability of a particular lottery outcome (in this case, winning zero dollars in a 'wheel of fortune' game) by changing the values of the two other possible lottery outcomes it was competed with. When winning zero dollars is the worst of three possible outcomes, Kahneman and Tversky's prospect theory predicts, as do several related theories, that subjects should view it negatively, but when it is the best of the three outcomes then subjects should view it more positively. The scanning experiment revealed that brain activation in the ventral striatum matched these predicted subjective valuations.

The other paper published that year reflected a collaboration between the economist Kevin McCabe, his colleague Vernon Smith (who would share the Nobel prize with Kahneman the following year for his contributions to experimental economics), and a team that included economists, a psychologist, and a biomedical engineer. Their paper, which appeared in the pages of the *Proceedings of the National Academy of Sciences* in the United States, examined behavior and neural activation while subjects engaged in a strategic game. This also represented the first use of game theory, an economic tool for the study of social decision making, in a neurobiological experiment. In that paper, subjects played a trust game either against an anonymous human opponent or against a computer. In their game, two players interact regarding an investment. The game begins when the first player is given 45 monetary units and given a choice between keeping that money or investing it with an unseen partner, the trustee. Any money given to the trustee is then multiplied by nine, and the partner must decide whether or not to share 180 units of this windfall to the original investor. What is interesting about the game is that standard normative theory (which assumes a self-interested chooser) assumes that the trustee wishes to maximize her gain. This predicts that in a consequence-free laboratory environment the trustee should refuse to share the windfall. Her partner, the investor, is presumed to know this about her opponent and thus is expected, in the first round, to keep all of the money for herself. What is striking about the game, however, is that humans do not behave in accord with these normative predictions when playing against another, even anonymous, human. Some players, however, do approximate this approach when facing a computer opponent. What McCabe and colleagues therefore did was to scan subjects while they played against both humans and computers and to take the difference between activation in the two states. These neurobiological data revealed that in some subjects the medial prefrontal cortex is differentially active under these two conditions, becoming more active when subjects play this cooperative strategy that deviates from the normative prediction. From these data the authors hypothesized that this nonnormative pattern of cooperation has its origin in circuits of the prefrontal cortex.

The following year, many of these emerging trends were reviewed in an important special Society for Neuroscience conference issue of the journal *Neuron*, edited by Jonathan Cohen and Kenneth Blum, entitled 'Reward and decision.' As these editors wrote in the introduction to that issue: "Within neuroscience, for example, we are awash with data that in many cases lack a coherent theoretical understanding (a quick trip to the poster floor of the Society for Neurosciences meeting can be convincing on this point). Conversely, in economics, it has become abundantly evident that the pristine assumptions of the 'standard economic model' – that individuals operate as optimal decision-makers in maximizing utility – are in direct violation of even the most basic facts about human behavior." In that issue, although all of the articles are written by neurobiologists, particular attention is drawn to normative theories of decision. Of particular interest are articles by Montague and Berns, Schultz, Dayan and Balleine, Gold and Shadlen, and Glimcher, which all point toward the interaction of normative models and neurobiology. In particular, the issue draws attention to the ongoing debate about the role of the neurotransmitter dopamine in reward processing and draws upon previous work that had identified normative or near-normative models of

learning that posit a role for dopamine. What followed was a literal flood of decision-making studies in the neuroscientific literature, many of which relied on normative economic theory.

At the end of this initial period a set of summary reviews began to emerge that served as manifestos for the emerging neuroeconomic discipline. In 2003, Glimcher published a book, directed primarily at neuroscientists, that reviewed the history of neuroscience and argued that this history was striking in its lack of normative models for higher cognitive function. Glimcher proposed explicitly in that book that economics could serve as the source for this much needed normative theory for cognitive neuroscience. Shortly thereafter, the Camerer, Loewenstein, and Prelec paper was published (2005), which also served as a manifesto for neuroeconomics, but this time from the economic side. More recently, an invited review detailing many of these advances appeared in the journal *Science*, marking the advancing visibility of neuroeconomics in the scientific community at large.

The next major step in neuroeconomics was probably the incorporation of game theoretic normative models into a broader neurobiological spectrum. In 2004, two key papers appeared that employed normative game theory for the study of single neurons in awake-behaving primates. A paper by Daeyeol Lee and colleagues studied animals trained to play a very simple classic game called matching pennies. In that game the monkeys had to choose between one of two responses, A and B. The monkeys' computer opponent made a similar choice. If they both chose the same response, the monkey earned a drop of juice, and if they made opposite choices, the monkey earned nothing. Lee varied across long blocks of trials the game theoretic sophistication of the monkey's computer opponent while recording from single neurons in the frontal cortex. What he and his colleagues found was that the behavior of the monkeys was responsive to the sophistication of the opponent and that neurons of the frontal cortex encoded properties related to both the monkeys' strategies and the recent history of the game.

In a similar vein, Dorris and Glimcher trained monkeys to play a slightly more complex game called work or shirk. Game theory, in this case Nash equilibrium theory (REF), predicts both a particular outcome for any set of payoffs in that game and that at behavioral equilibrium the monkeys should mix their responses, sometimes working and sometimes shirking. Regardless of the ratio of shirking to working, the normative theory (somewhat counterintuitively) predicts that the subjective value of working and shirking should be nearly identical at this equilibrium. They found that the behavior of the monkeys was fairly well predicted by the normative theory, and neurons in the posterior parietal cortex seemed to fire at a rate that was also predicted for elements encoding expected utility. Together, these two papers probably represented the first use of normative game theory for single-unit studies.

Within the economic community a role similar to that of the *Neuron* special issue was played by a special issue on Neuroeconomics presented by the journal *Games and Economic Behavior* and edited by the economist Aldo Rustichini, which appeared shortly after this in 2005. Within the economic community this issue was hugely influential and served to define neuroeconomics to a large degree. That issue included articles by several economists and neuroscientists, including scholars ranging from Gallistel to Smith.

Another major advance was presented in 2005 by Michael Kosfeld and his colleagues in Ernst Fehr's research group at the University of Zurich. This paper was important because it was the first demonstration of a neuropharmacological manipulation that alters behavior in a manner that can be interpreted with regard to normative theory. Subjects were asked to play a trust game much like the one examined by McCabe and colleagues. Fehr's critical manipulation was to increase brain levels of the neurotransmitter oxytocin (by an intranasal application of the compound) before some of the players made their decision. What Kosfeld and colleagues found was that trustees in the game (those players who had received money from the investors) returned more to the investors (in contravention of normative theory) if they had received oxytocin than if they had been treated with a control substance. What was most interesting about this study from a neuroeconomic point of view was the demonstration that administration of this endogenously produced hormone altered the choice behavior of subjects in a way that went against an existing and well-described normative economic theory.

Summary

The future of neuroeconomics, however, remains unclear. There is still significant opposition to neurobiological approaches within much of mainstream economics, and, in fairness, neuroeconomics has yet to revolutionize either economic theory or the culture of neuroscience. The next decade should prove a critical period for this discipline. During this second decade neuroeconomics, if it is to succeed, will have to make major contributions to both of its parent disciplines.

In 1998 the evolutionary biologist EO Wilson wrote that "The full understanding of utility will come from biology and psychology by reduction to the elements of human behavior followed by a

bottom-up synthesis, not from the social sciences by top-down inference and guesswork based on intuitive knowledge. It is in biology and psychology that economists and social scientists will find the premises needed to fashion more predictive models, just as it was in physics and chemistry that researchers found the premises that upgraded biology." It remains to be seen whether neuroeconomic studies will provide major insights in either neuroscience, economics, or both over the next few years. Only time will determine whether this emerging discipline actually can revolutionize its parent disciplines.

See also: Decision-Making in Financial Markets; Decision-Making and Neuroeconomics; Delayed Reinforcement: Economics; Game Theory and the Economics of Animal Communication; Games in Monkeys: Neurophysiology and Motor Decision-Making; Prediction Errors in Neural Processing: Imaging in Humans; Reward Decision-Making; Social Cognition.

Further Reading

Barraclough DJ, Conroy ML, and Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience* 7: 404–410.

Bechara A, Damasio H, Tranel D, and Damasio AR (1997) Deciding advantageously before knowing the advantageous strategy. *Science* 275: 1293–1295.

Breiter HC, Aharon I, Kahneman D, Dale A, and Shizgal P (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* 30: 619–639.

Camerer C, Loewenstein G, and Prelec D (2005) Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature* 43: 9–64.

Dayan P and Balleine BW (2002) Reward, motivation, and reinforcement learning. *Neuron* 36: 285–298.

Dorris MC and Glimcher PW (2004) Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* 44: 365–378.

Ellsberg D (1961) Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Gallistel CR (1994) Foraging for brain stimulation: Toward a neurobiology of computation. *Cognition* 50: 151–170.

Gallistel CR (2005) Deconstructing the law of effect. *Games and Economic Behavior* 52: 410–423.

Glimcher P (2002) Decisions, decisions, decisions: Choosing a biological science of choice. *Neuron* 36: 323–332.

Glimcher P (2003) *Decisions Uncertainty and the Brain: The Science of Neuroeconomics*. Cambridge, MA: MIT Press.

Glimcher PW and Rustichini A (2004) Neuroeconomics: The consilience of brain and decision. *Science* 306: 447–452.

Gold JI and Shadlen MN (2002) Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* 36: 299–308.

Herrnstein RJ, Rachlin H, and Laibson DI (1997) *The Matching Law: Papers in Psychology and Economics*. New York/Cambridge, MA: Russell Sage Foundation/Harvard University Press.

Houser D, Bechara A, Keane M, McCabe K, and Smith V (2005) Identifying individual differences: An algorithm with application to Phineas Gage. *Games and Economic Behavior* 52: 373–385.

Kahneman D and Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.

Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, and Fehr E (2005) Oxytocin increases trust in humans. *Nature* 435: 673–676.

Luce RD and Suppes P (1965) Preference, utility, and subjective probability. In: Luce RD, Bush RR, and Galanter E (eds.) *Handbook of Mathematical Psychology*. New York: Wiley.

McCabe K, Houser D, Ryan L, Smith V, and Trouard T (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11832–11835.

Montague PR and Berns GS (2002) Neural economics and the biological substrates of valuation. *Neuron* 36: 265–284.

Newsome WT, Britten KH, and Movshon JA (1989) Neuronal correlates of a perceptual decision. *Nature* 341: 52–54.

Platt ML and Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400: 233–238.

Plott CR and Smith VL (1978) An experimental examination of two exchange institutions. *Review of Economic Studies* 45: 133–153.

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, and Cohen JD (2004) Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport* 15: 2539–2543.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, and Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. *Science* 300: 1755–1758.

Savage LJ (1954) *The Foundations of Statistics*. New York: Wiley.

Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36: 241–263.

Shizgal P (1997) Neural basis of utility estimation. *Current Opinion in Neurobiology* 7: 198–208.

Shizgal P and Conover K (1996) On the neural computation of utility. *Current Directions in Psychological Science* 5: 37–43.

Smith VL (1991) *Papers in Experimental Economics*. Cambridge: Cambridge University Press.

VonNeumann J and Morgenstern O (1944) *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wilson EO (1998) *Consilience: The Unity of Knowledge*. New York: Knopf (distributed by Random House).

Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women

N E Rance, University of Arizona College of Medicine, Tucson, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The loss of female reproductive function in midlife constitutes the most dramatic and consistent change during aging in humans. Because menopause occurs early, at a mean age of 51 years, the average woman will spend approximately one-third of her life span in the postmenopausal phase. The hallmark of menopause is ovarian failure, which leads to a reduction in circulating blood levels of estrogen and progesterone to castrate levels. In turn, the loss of ovarian hormones leads to hypersecretion of pituitary gonadotropins and the clinical symptoms experienced by postmenopausal women, including hot flushes, vaginal dryness, and mood changes. Postmenopausal women also have an increased risk for osteoporosis and cardiovascular disease, which may reflect the effects of estrogen withdrawal on age-related diseases. Thus, the repercussions of menopause can impact a women's health for a significant portion of her life span.

Overview of the Hypothalamic–Pituitary–Ovarian Axis in Young Women

Reproduction is controlled via reciprocal interactions among the hypothalamus, anterior pituitary gland, and ovaries. Neurons in the medial basal hypothalamus that synthesize gonadotropin-releasing hormone (GnRH) project to fenestrated capillaries in the median eminence. In the median eminence, GnRH is secreted from axon terminals into hypophyseal portal vessels, a capillary system that carries blood to the pituitary gland. At the level of the anterior pituitary gland, GnRH stimulates the production and release of the gonadotropins, luteinizing hormone (LH) and follicle-stimulating hormone (FSH), into the systemic circulation. LH and FSH target the ovaries via the systemic circulation to regulate folliculogenesis, ovulation, and the secretion of steroid and glycoprotein hormones. In turn, the ovarian steroids, estrogen and progesterone, exert complicated feedback actions on both the hypothalamus and the pituitary. The ovary also secretes glycoprotein hormones (such as inhibin and activin) that modulate FSH secretion at the level of the anterior pituitary gland. The timing of the menstrual cycle is dependent upon the pattern of ovarian hormone secretion. In addition, multiple neural inputs to hypothalamic GnRH neurons convey information on the status of physiologic systems to optimize the timing for reproduction.

GnRH is secreted in a pulsatile manner, leading to short-duration peaks of LH in peripheral plasma. Studies in the rhesus monkeys have delineated the medial basal hypothalamus as the control center for pulsatile LH secretion. The medial basal hypothalamus is the site of multiunit electrical activity (the GnRH pulse generator) that is synchronized with pulses of LH in the peripheral circulation. Neurons with high levels of GnRH gene expression are scattered within the primate medial basal hypothalamus, further evidence of its importance in the control of the reproductive axis. The frequency and amplitude of GnRH pulses vary over the menstrual cycle, and the pattern of GnRH pulsatility plays an important role in pituitary responsiveness. For example, continuous (rather than pulsatile) administration of GnRH (or one of its analogs) actually suppresses LH secretion. Moreover, the pattern of LH and FSH pulses in the systemic circulation influences ovarian follicle development.

Ovarian Aging and the Menopausal Transition

Ovarian aging is a critical determinant of menopause. At the time of birth, there are approximately 1 million primordial follicles. The absolute number of ovarian follicles declines exponentially until approximately 37 years of age, at which time follicular loss accelerates (Figure 1). At the time of the perimenopausal period, the follicle population has been reduced to less than 1000, and after menopause the ovary is virtually depleted of follicles. Fertility begins to decline in the third and fourth decades of life and this process accelerates after 35 years of age.

A staging system for the phases of reproductive life in women is shown in Figure 2. The staging system uses the final menstrual period as a reference point. In the reproductive phase, menstrual cycles are regular, except for a variable time immediately following puberty. The reproductive phase is followed by the menopausal transition, heralded by the onset of cycles of variable lengths. It is important to note, however, that alterations in hormone secretion occur quite early, even before the onset of irregular cycles.

The first endocrine marker of reproductive aging consists of an elevation of serum FSH early in the follicular phase. The rise in serum FSH has been termed monotropic because it is unaccompanied by

an elevation in serum LH. The monotropic increase in FSH secretion in the late reproductive stage is inversely correlated with declining serum levels of inhibin B (Figure 3). Inhibin B is secreted by granulosa cells in developing antral follicles. Lower levels of inhibin B may reflect a diminished number of primordial follicles recruited into the developing pool or decreased numbers of granulosa cells for each follicle. Loss of the restraining effect of inhibin B on the anterior pituitary gland results in the reciprocal increase in serum FSH. Early development of the dominant follicle, secondary to increased FSH, may lead to shortening of the follicular phase and marked increases in estrogen secretion.

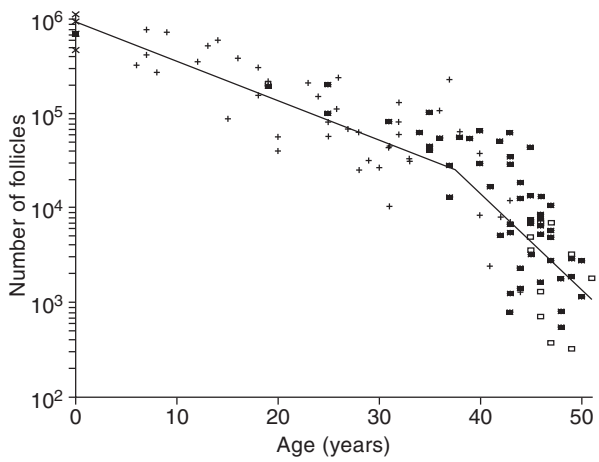

Figure 1 Biexponential model of declining follicle numbers in pairs of human ovaries, from neonatal age to 51 years old. Each symbol represents one person. The different symbols represent data obtained from several studies. Reprinted from Faddy MJ, Gosden RG, Gougeon A, et al. (1992) Accelerated disappearance of ovarian follicles in mid-life: Implications for forecasting menopause. *Human Reproduction* 7: 1342–1346, with permission.

The menopause transition is a prolonged period of time characterized by irregular cycle lengths and periods of amenorrhea (Figure 2). During the early menopausal transition, there may be elevated levels of FSH throughout the cycle and lower levels of inhibin A, inhibin B, and progesterone. Serum estrogen levels of the early menopausal transition are either preserved or increased. In the late menopausal transition, frequent anovulatory cycles occur, with substantial variability in hormone levels. The multiinstitutional Study of Women's Health across the Nation (SWAN) has classified the anovulatory cycles into three types based on daily urine hormone levels: (1) cycles in which estrogen fails to induce an LH surge, (2) cycles with an LH surge but disrupted or dysfunctional corpus luteum formation, and (3) cycles with tonically elevated LH and FSH but with no evidence of a rise in estrogen. The variable cycles in the late menopausal transition lead to markedly fluctuating levels of ovarian hormones. The fluctuating levels of estrogen in particular may account for many of the symptoms experienced at this stage, including breast tenderness, menorrhagia, hot flushes, sleep disturbances, and mood changes.

The monotropic rise in FSH secretion has been postulated to produce the accelerated phase of follicular

Final menstrual period (FMP) 0

Stages:	−5	−4	ϖ3	−2	−1	+1	+2
Terminology:	Reproductive			Menopausal transition		Postmenopause	
	Early	Peak	Late	Early	Late*	Early*	Late
				Perimenopause			
Duration of stage:	Variable			Variable		(a) 1 yr	(b) 4 yrs / Until demise
Menstrual cycles:	Variable to regular	Regular		Variable cycle length (>7 days different from normal)	≥2 skipped cycles and an interval of amenorrhea (≥60 days)	Amen × 12 mos	None
Endocrine:	Normal FSH		↑ FSH	↑ FSH		↑ FSH	

*Stages most likely to be characterized by vasomotor symptoms ↑ = Elevated

Figure 2 The Stages of Reproductive Aging Workshop staging system, showing endocrine levels of follicle-stimulating hormone (FSH) at different stages. Reprinted from Soules MR, Sherman S, Parrott E, et al. (2001) Executive summary: Stages of Reproductive Aging Workshop (STRAW). *Fertility and Sterility* 76: 874–878, with permission.

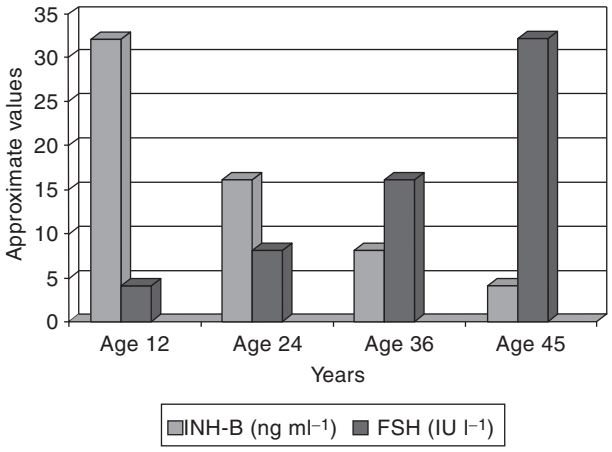

Figure 3 As ovarian reserve declines throughout reproductive life, inhibin B (INH-B), a product of small, growing follicles, decreases. Concomitant with this relative loss of INH-B, follicular-phase follicle-stimulating hormone (FSH) increases in a reciprocal fashion. Reprinted from Santoro N (2005) The menopausal transition. *American Journal of Medicine* 118(supplement 12B): 8–13, with permission.

decline that occurs after the age of 37 years. Higher levels of FSH (secondary to decreased inhibin B) could recruit increased numbers of follicles into the growing pool. Once this happens, follicles must either ovulate or become atretic. An abnormal FSH signal could also be detrimental to ovarian function independent from its effects on follicular recruitment. The steady loss of ovarian follicles in late reproductive life could reach a critical threshold and initiate a cascade of events leading ultimately to complete ovarian failure.

A hypothalamic contribution to the acceleration of follicular loss has also been proposed. For example, in experimental animals, altered patterns of LH pulsatility (secondary to GnRH pulses) are detrimental to ovarian function. Moreover, reducing the frequency of GnRH pulses can produce isolated FSH secretion in experimental animals, providing a potential mechanism whereby an altered hypothalamic signal could contribute to the monotropic rise in FSH and the subsequent acceleration of follicular decline. The slowing of LH pulses in older, regularly cycling women, however, has only been demonstrated in one study, and this finding has not yet been replicated. Therefore, the proposal that hypothalamic factors trigger the ovarian failure of menopause is controversial.

Effects of Ovarian Failure on the Hypothalamic–Pituitary Axis

The hormonal milieu of the postmenopausal phase of life is characterized by profound ovarian hormone deficiency. There is a marked increase in serum gonadotropins in postmenopausal women, similar to the hormone profile seen in young, oophorectomized women. The postmenopausal increase in serum gonadotropins is secondary to both loss of steroid negative feedback and removal of the restraining action of inhibin on FSH secretion from the pituitary gland. In addition, the loss of ovarian hormones shifts the composition of the gonodotropins to more acidic isoforms. The shift of gonadotropin isoforms contributes to elevated levels of LH and FSH in the peripheral circulation by retarding their clearance. Administration of gonadal steroids still suppresses the high levels of serum gonadotropins in elderly women, indicating that the postmenopausal hypothalamus remains sensitive to negative feedback effects. Overall, many aspects of hypothalamic and pituitary function remain intact in postmenopausal women, in contrast to the complete failure of the ovaries.

Based on the central role of GnRH in the reproductive axis, research has focused on the status of GnRH neurons in postmenopausal women. Because direct measurements of GnRH secretion require invasive procedures, information on the function of these neurons has been obtained through indirect methods. For example, assessment of GnRH pulses can be inferred by examination of pulses of the free α subunit (FAS) of the pituitary glycoprotein hormones in peripheral plasma. This method was developed because measurements of LH pulses are complicated by the prolonged half-life of LH and the rapid frequency of pulses in postmenopausal women. The quantity of GnRH secretion can also be estimated by pharmacological methods such as competitive receptor blockage. Another approach to studying GnRH neurons is to quantify relative levels of gene expression in postmortem tissues using *in situ* hybridization histochemistry. This method is well suited for the investigation of human autopsy material due to the relative stability of mRNA species with prolonged postmortem intervals. A further level of anatomical resolution is provided by the localization of mRNA species at the cellular level within tissue sections.

The function of GnRH neurons in postmenopausal women appears to be dominated by the removal of steroid negative feedback. Studies have shown that negative feedback effects of estrogen and progesterone on LH secretion are maintained in postmenopausal women. The hypothalamic content of the GnRH peptide decreases in postmenopausal women, similar to response in young women who have had their ovaries removed. Therefore, it has been hypothesized that the decreased GnRH content in postmenopausal women is secondary to increased

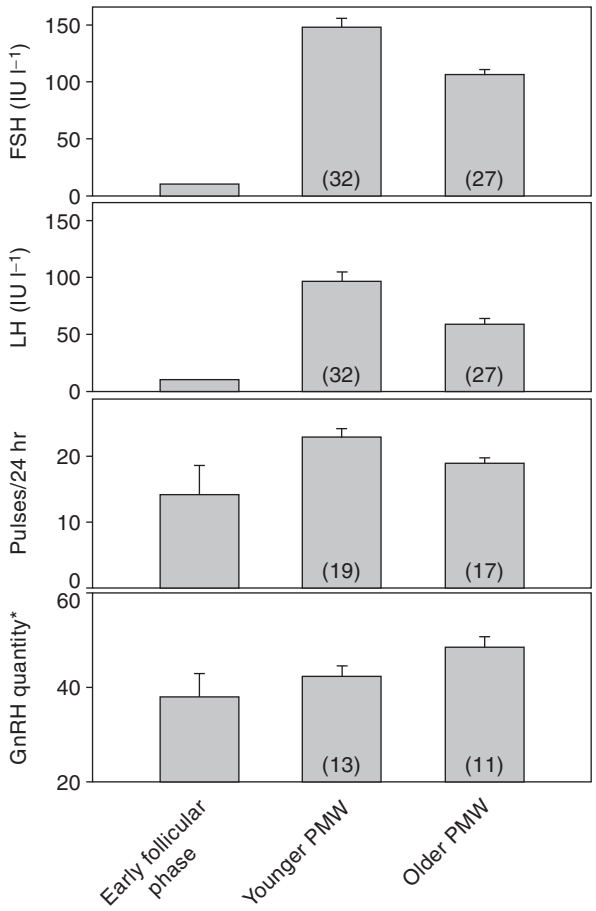

Figure 4 Studies in younger (age 45–55 years) and older (age 70–80 years) postmenopausal women (PMW) indicate that follicle-stimulating hormone (FSH) and luteinizing hormone (LH) levels increase after menopause, but decrease with aging. Gonadotropin-releasing hormone (GnRH) pulse frequency, assessed by pulsatile secretion of the free α subunit, is also increased in relation to the early follicular phase, but is not different from the frequency in the late follicular phase of midcycle surge in healthy women. After menopause, pulse frequency decreases with aging. The quantity of GnRH (100% inhibition of LH in response to submaximal GnRH receptor blockade) increases with aging. The number of individuals contributing to each bar is indicated in parentheses. Reprinted from Hall JE (2004) Neuroendocrine physiology of the early and late menopause. *Endocrinology and Metabolism Clinics of North America* 33: 637–659, with permission.

secretion into the portal vessels. As would be expected with the removal of steroid negative feedback, GnRH pulse frequency (assessed by peripheral FAS pulses) is increased in young postmenopausal women compared with premenopausal women in the early follicular phase (Figure 4). Moreover, the quantity of GnRH secreted (estimated by competitive receptor blockage) is increased in young postmenopausal women and continues to increase with age. These findings agree with the demonstration of increased GnRH gene expression in the medial basal hypothalamus of postmenopausal women. Taken together, these studies provide evidence that GnRH neurons in postmenopausal women, instead of age-related signs of degeneration, exhibit compensatory responses due to removal of ovarian hormones.

Effects of Menopause on Neuronal Morphology and Neuropeptide Gene Expression in the Human Hypothalamus

More than four decades ago, Sheehan and Kovacs described differences in hypothalamic neuronal morphology in premenopausal and postmenopausal women. In postmenopausal women, pronounced enlargement of neurons was observed in the infundibular nucleus of the medial basal hypothalamus. The cellular hypertrophy was characterized by increased Nissl substance (indicative of increased protein synthesis) and enlarged nuclei and nucleoli (Figure 5). Based on similar findings in other hypogonadal conditions, the authors proposed that neuronal hypertrophy in the infundibular nucleus was secondary to loss of ovarian function.

Subsequent studies using *in situ* hybridization have characterized the phenotype of the hypertrophied neurons and have shown numerous alterations in hypothalamic neuropeptide gene expression in comparing tissues of premenopausal and postmenopausal women. The most dramatic change in gene expression occurs in the subpopulation of infundibular neurons that undergoes neuronal hypertrophy. These neurons co-express estrogen receptor-α and two tachykinin neuropeptides, neurokinin B (NKB) and substance P (SP). In addition to the marked increase in cell size, the number of neurons expressing NKB and SP gene transcripts in the infundibular nucleus increases by more than sixfold, indicating a marked increase in tachykinin gene expression. GnRH gene expression is also elevated in postmenopausal women, but in a separate subpopulation of neurons within the medial basal hypothalamus. The elevation in GnRH mRNA provides evidence of increased hypothalamic GnRH secretion contributing to the rise in plasma gonadotropins in postmenopausal women.

In contrast to elevated NKB and GnRH gene expression, the number of neurons expressing proopiomelanocortin (POMC) gene transcripts in the infundibular nucleus is significantly reduced in older women. Moreover, the number of neurons expressing POMC mRNA is inversely correlated with age and the expression of neuropeptide Y mRNA. Numerous studies in experimental animals have revealed that POMC and neuropeptide Y neurons in the arcuate nucleus (the rodent homolog of the infundibular nucleus) are essential components of hypothalamic

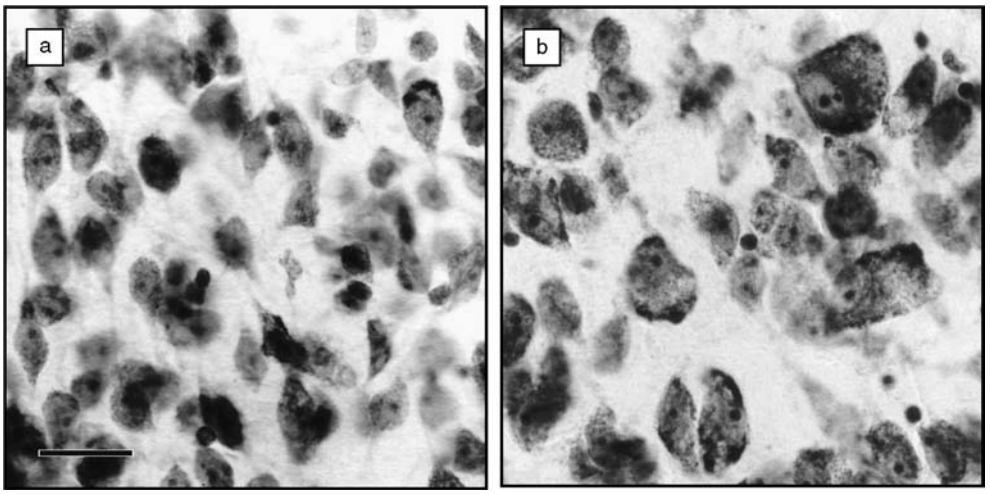

Figure 5 Representative photomicrographs of cresyl violet-stained sections of the infundibular nucleus of young, premenopausal (a) woman and an older, postmenopausal (b) woman. In the section from the older woman, note the considerably enlarged neurons, with both increased nuclear size and increased Nissl substance. Scale bar = 25 μm (a, b). Reprinted from Abel TW and Rance NE (2000) Stereologic study of the hypothalamic infundibular nucleus in young and older women. *Journal of Comparative Neurology* 424: 679–688, with permission.

regulatory circuits controlling body weight and reproduction. Arcuate POMC and neuropeptide Y neurons express leptin receptors and exert opposing actions on feeding behavior and energy balance. Experimental paradigms in laboratory animals in which nutritional status or serum leptin is altered produce reciprocal changes in POMC and neuropeptide Y gene transcripts. Thus, the finding of a reciprocal relationship between POMC and neuropeptide Y gene expression in the human arcuate (infundibular) nucleus suggests that this relationship is conserved among species.

A major limitation in interpreting changes in gene expression in premenopausal and postmenopausal women is the confounding influence of two variables, age and ovarian status. To control for the effects of aging, young cynomolgus monkeys have been used as a model to determine the effects of ovariectomy and steroid hormone replacement on hypothalamic neuropeptide gene expression (Table 1). Importantly, ovariectomy of young monkeys induces both hypertrophy and increased NKB gene expression in the infundibular nucleus. Similarly, GnRH gene expression in the medial basal hypothalamus is increased by ovariectomy of young cynomolgus monkeys. Conversely, estrogen replacement of young ovariectomized monkeys decreases NKB and GnRH gene transcripts. These data provide strong support for the hypothesis that increased NKB and GnRH gene expression observed in the hypothalamus of postmenopausal women is secondary to ovarian failure and not due to age *per se*. In contrast to the marked effect of ovarian hormones on NKB gene expression, ovariectomy or steroid replacement had no effect on POMC or neuropeptide Y gene expression in the primate model. These data suggest that factors other than ovarian failure may be more important in influencing this peptide system in postmenopausal women.

The association of increased NKB gene expression with estrogen withdrawal and gonadotropin hypersecretion suggests that these neurons participate in the hypothalamic circuitry regulating steroid negative feedback. This hypothesis is supported by the demonstration of estrogen receptor-α within the hypertrophied neurons and the location of these neurons within the primate control center for reproduction. Studies in experimental animals have also provided strong support for this hypothesis. Expression of NKB gene transcripts is increased by ovariectomy and reduced by estrogen replacement in a variety of species. NKB gene expression in the arcuate nucleus varies with the estrous cycle, and these neurons are sexually dimorphic. Moreover, anatomic studies have also revealed a close relationship between NKB and GnRH neurons. In the rat median eminence, there is close apposition of NKB and GnRH fibers at a site where GnRH fibers express the receptor for NKB, the neurokinin-3 receptor. Finally, LH secretion is modulated by central injections of a neurokinin-3 receptor agonist. Thus, multiple lines of evidence support the hypothesis that the hypertrophied neurons in the infundibular nucleus of postmenopausal women play a role in the regulation of steroid negative feedback on gonadotropin secretion.

Table 1 Effects of ovariectomy and estrogen replacement on neuropeptide gene expression in the medial basal hypothalamus of young cynomolgus monkeys

Neuron phenotype	*Changes in postmenopausal women*	*Effects of ovariectomy of cynomolgus monkeys*	*Effects of estrogen replacement in ovariectomized monkeys*
NKB	Hypertrophy, ↑ mRNA	Hypertrophy, ↑ mRNA	↓ mRNA below levels of detection
GnRH	↑ mRNA	↑ mRNA	↓ mRNA
POMC	↓ mRNA	No change	No change
NPY	↑ mRNA	No change	Not tested

NKB, neurokinin B; GnRH, gonadotropin-releasing hormone; POMC, proopiomelanocortin; NPY, neuropeptide Y.

Aging of the Reproductive Neuroendocrine Axis, Independent of the Effects of Gonadal Status

As chronological age advances, hypothalamic aging and secondary effects from the aging of other neuroendocrine systems become increasingly important. Age-associated changes occur in the reproductive neuroendocrine axis independently from those caused by ovarian failure. For example, plasma LH is lower in postmenopausal women (between 51 and 70 years old), compared to women (between 25 and 40 years old) with premature ovarian failure. A consistent decline in serum LH has also been observed in association with aging after the menopause. Moreover, some studies have reported an age-associated decline in the frequency of LH pulses, although these studies are confounded by the prolonged half-life of LH after the menopause.

Recent studies by Hall and colleagues have provided a detailed picture of the aging of the hypothalamic–pituitary axis by comparing hormone levels in postmenopausal women who are 45–55 years of age and 70–80 years of age. Using a sampling interval of 5 min and measurements of the gonadotropin free α subunit, a reduction in GnRH pulse frequency was observed in the older postmenopausal women. Because the GnRH pulse generator resides within the hypothalamus, these studies provide evidence of hypothalamic neuroendocrine aging that is independent of the changes in gonadal function.

Despite slower GnRH pulses and reduced serum LH, older postmenopausal women retain the capacity to decrease GnRH secretion in response to steroid feedback. Interestingly, indirect pharmacological measurements provide evidence that GnRH secretion increases with age in postmenopausal women. The age-related rise in GnRH secretion is consistent with the demonstration of increased GnRH secretion documented by push–pull perfusion in older rhesus monkeys. Moreover, the reciprocal changes in POMC and neuropeptide Y (NPY) gene expression that occur in the medial basal hypothalamus of older women also appear to be independent of gonadal status. It is hypothesized that loss of an inhibitory effect of POMC, combined with stimulation by NPY, could explain why GnRH secretion could continue to rise, even after the menopausal transition.

Summary

The hallmark of reproductive aging in women is degeneration of ovarian follicles, which accelerates around the age of 37 years. An early sign of reproductive aging is a rise in serum FSH, which is inversely correlated with decreased inhibin B secretion from developing antral follicles. After a transition phase characterized by irregular cycles and fluctuating hormone levels, the final menstrual period marks the cessation of cycles. In the postmenopausal period, the reproductive neuroendocrine axis is dominated by the withdrawal of ovarian hormones. Age-associated changes in the hypothalamic pituitary axis also occur and are independent of gonadal status.

See also: Emotional Hormones and Memory Modulation; Gene Expression Regulation: Steroid Hormone Effects; Gonadotropin, Neural and Hormonal Control; Gonadotropin-Releasing Hormone: GnRH-1 System; Hormones and Memory; Hormones and Behavior; Hypothalamo–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH; Neuroendocrine Aging: Pituitary Metabolism; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Pituitary-Adrenal Axis; Photoperiodic Regulation of Reproductive Cycles; Pituitary Gland (Cell Types, Mediators, Development); Sexual Behavior: Neuroendocrine Control.

Further Reading

Abel TW and Rance NE (2000) Stereologic study of the hypothalamic infundibular nucleus in young and older women. *Journal of Comparative Neurology* 424: 679–688.

Escobar CM, Krajewski SJ, Sandoval-Guzmán T, et al. (2004) Neuropeptide Y gene expression is increased in the infundibular nucleus of postmenopausal women. *Journal of Clinical Endocrinology & Metabolism* 89: 2338–2343.

Faddy MJ, Gosden RG, Gougeon A, et al. (1992) Accelerated disappearance of ovarian follicles in mid-life: Implications for forecasting menopause. *Human Reproduction* 7: 1342–1346.

Gill S, Lavoie HB, Bo-Abbas Y, et al. (2002) Negative feedback effects of gonadal steroids are preserved with aging in postmenopausal women. *Journal of Clinical Endocrinology & Metabolism* 87: 2297–2302.

Gill S, Sharpless JL, Rado K, et al. (2002) Evidence that GnRH decreases with gonadal steroid feedback but increases with age in postmenopausal women. *Journal of Clinical Endocrinology & Metabolism* 87: 2290–2296.

Gore AC, Windsor-Engnell BM, and Terasawa E (2004) Menopausal increases in pulsatile gonadotropin-releasing hormone release in a nonhuman primate (*Macaca mulatta*). *Endocrinology* 145: 4653–4659.

Hale GE and Burger HG (2005) Perimenopausal reproductive endocrinology. *Endocrinology and Metabolism Clinics of North America* 34: 907–922, ix.

Hall JE (2004) Neuroendocrine physiology of the early and late menopause. *Endocrinology and Metabolism Clinics of North America* 33: 637–659.

Hall JE, Lavoie HB, Marsh EE, et al. (2000) Decrease in gonadotropin-releasing hormone (GnRH) pulse frequency with aging in postmenopausal women. *Journal of Clinical Endocrinology & Metabolism* 85: 1794–1800.

Knobil E (1980) The neuroendocrine control of the menstrual cycle. *Recent Progress in Hormone Research* 36: 53–88.

Rance NE and Young WS, III (1991) Hypertrophy and increased gene expression of neurons containing neurokinin-B and substance-P messenger ribonucleic acids in the hypothalami of postmenopausal women. *Endocrinology* 128: 2239–2247.

Sandoval-Guzmán T, Stalcup ST, Krajewski SJ, et al. (2004) Effects of ovariectomy on the neuroendocrine axes regulating reproduction and energy balance in young cynomolgus macaques. *Journal of Neuroendocrinology* 16: 146–153.

Santoro N (2005) The menopausal transition. *American Journal of Medicine* 118(supplement 12B): 8–13.

Soules MR, Sherman S, Parrott E, et al. (2001) Executive summary: Stages of Reproductive Aging Workshop (STRAW). *Fertility and Sterility* 76: 874–878.

Weiss G, Skurnick JH, Goldsmith LT, et al. (2004) Menopause and hypothalamic–pituitary sensitivity to estrogen. *JAMA: Journal of the American Medical Association* 292: 2991–2996.

Neuroendocrine Aging: Pituitary Metabolism

W E Sonntag, University of Oklahoma Health Science Center, Oklahoma City, OK, USA
C B Herenu and R G Goya, National University of La Plata, La Plata, Argentina

© 2009 Elsevier Ltd. All rights reserved.

Overview

A substantial volume of empirical and scientific evidence has accumulated that demonstrates that biological aging is associated with functional deficits at the cellular, tissue, organ, and system levels. Several theories have been proposed to explain these changes as well as the increased risk of disease with age. However, to date, no single explanation satisfactorily accounts for the physiological and pathological events related to biological aging, and it is apparent that several factors interact, resulting in the decline in tissue function that creates diversity in the aging phenotype. Current concepts of aging encompass genetic, free radical, somatic mutation, DNA repair, and cellular theories, all of which are not mutually exclusive. Although there is evidence to support a relationship between these theories of aging and end-of-life pathologies and/or life span, there is no overwhelming evidence that the progressive development of the aging phenotype can be adequately explained by these classical theories.

The neuroendocrine theory of aging evolved from observations that animals exhibit predictable and progressive impairments in a number of physiological processes over time, including, but not limited to, decreases in reproductive and immune function, decreases in muscle mass and function, increased adipose tissue mass, and decreases in glucose utilization and cognitive function. The role of the neuroendocrine system in these processes is related to the observation that many of the hormones regulated by the neuroendocrine system have an important trophic and integrative role in maintaining tissue function, withdrawal of hormonal support mimics some of the phenotypes observed in aging animals and humans, and hormone replacement therapy in many cases improves function in older individuals and, in some unique cases, life span.

The neuroendocrine system includes the hypothalamus and associated brain structures as well as the pituitary gland. This system includes neurotransmitters and neuropeptides within the brain that regulate hypothalamic-releasing and -inhibiting hormones secreted into hypophyseal portal blood that reaches the pituitary gland. These hormones influence the secretion of anterior pituitary hormones into the bloodstream and subsequently regulate tissue function. The posterior pituitary also is an important part of the neuroendocrine system but, in contrast to the anterior pituitary gland, is composed of long axons from specific hypothalamic nuclei. The hypothalamus and pituitary gland have the capacity to detect neural activity and/or humoral secretions from target tissues and adjust activity to maintain an optimal internal environment or 'milieu' for tissue function. It is well established that the neuroendocrine system has a critical role in regulating tissue growth and metabolism through the release of growth hormone (GH) and thyroid stimulating hormone (TSH); reproductive function through the release of luteinizing hormone, follicle stimulating hormone, prolactin, and oxytocin; and plasma electrolytes and responses to stress through secretion of vasopressin and adrenocorticotropin, respectively. In addition, the hypothalamus has an important role in the integration of parasympathetic and sympathetic nervous system activity and can thereby influence a wide variety of functions, including heart rate, blood pressure, vascular responses, and glucose metabolism, among others. Specific nuclei within the hypothalamus also regulate biological rhythms. More recently, the regulation of food intake, fat metabolism, and body composition has been shown to be regulated through the hypothalamus by its response to leptin, cholecystokinin, and ghrelin via the synthesis of neuropeptide Y and other orexigenic peptides. Unfortunately, the categorization of hormones and their primary functions noted above is an overly simplistic and only partial view of the neuroendocrine system, and it is now known that critical interactions occur between hormones that contribute to the regulation of cellular function. Because many of the early events of aging include alteration in systems regulated by the neuroendocrine axis, it was proposed (and subsequent studies supported the conclusion) that age-dependent alterations in the neuroendocrine system result in a progressive series of events that are manifest as biological aging.

Although the etiology of the age-related changes in the neuroendocrine system is unknown, it has been proposed that cellular and molecular events in specific subpopulations of neurons within the hypothalamus and brain and/or supporting structures are a contributing factor in the dysregulation of this system. The cause of the specific perturbations may be related to genetic errors or increased free radicals that lead to progressive aberrations in tissue function. As a result, the contribution of neuroendocrine

theories to aging is unique compared with other theories of aging in that alterations in this system are not considered the primary causative factor in biological aging but rather are important mediators of aging initiated by cellular changes in specific subpopulations of neurons or systems that closely interact with hypothalamic neurons. In this article, the major alterations within the neuroendocrine axis are discussed, with special emphasis on the regulation of GH, insulin-like growth factor 1 (IGF-1), TSH, and prolactin since these hormones mediate or have the potential to mediate some of more important physiological changes that occur with age. Insulin has been included in this review because of its importance for biological aging, homology to IGF-1, and potential interactions with several neuroendocrine systems.

GH and IGF-1

Invertebrate Studies

Results from studies conducted over the past several years have revealed that life span in invertebrates can be substantially increased by disrupting the insulin/IGF-1 pathway. In *Caenorhabditis elegans*, insulin and IGF-1 share a common signaling pathway, and the protein receptor encoded by the daf-2 gene binds to a circulating insulin-like substrate initiating a cascade of events, including activation of age-1, a homolog of mammalian phosphoinositide-3-OH kinase. This kinase, in turn, activates the Akt-encoded protein kinase B (PKB). PKB is a serine/threonine kinase that blocks the function of a forkhead/winged helix transcription factor (daf-16), at least in part, by preventing its translocation to the nucleus.

Disruption of the daf-2 pathway (including mutations in daf-2, age-1, or Akt genes) in *C. elegans* extends life span from 30% to 100%. This appears to be the result of the release of the Akt/PKB 'brake' on daf-16, which is the primary target of the daf-2 pathway. Although the specific mechanisms of the increased life span are not well understood, it is known that daf-16 regulates gene expression that is essential to environmental stress resistance, development, and dauer formation.

In addition to the studies in *C. elegans*, results indicate that disruption of the insulin/IGF-1 signaling pathway in flies and yeast also increases life span. These effects on life span in multiple species raise the possibility that these signaling molecules regulate a conserved pathway that influences longevity in mammals. Although there are promising data in this regard, these issues are more complex in mammals because IGF-1 and insulin regulate two independent but interacting signaling pathways and appropriate models to address these questions are not readily available.

Mammalian Studies: Background

Our understanding of GH and IGF-1 regulation and actions in mammals began in the early part of this century. It was known that a substance present in blood promotes body growth, but it was only after the isolation of bovine GH from the pituitary gland in 1945 that the biological effects of the hormone became evident. It was subsequently shown that this protein hormone stimulated cellular amino acid uptake, DNA, RNA, and protein synthesis and has a role in cellular division and hypertrophy.

Plasma levels of GH reveal a pattern of discrete pulses that increase after the onset of sleep. Although the precise function of this ultradian pattern remains unknown, the pulsatile release of GH has been confirmed in every species examined to date and is closely related to the biological actions of the hormone. GH in adolescents and young adults is characterized by relatively low-amplitude pulses throughout the day and a large pulse after the onset of sleep. In male rats, GH is characterized by an ultradian rhythm with high amplitude secretory pulses every 3–3.5 h; between pulses, GH decreases to almost undetectable levels. This secretory pattern is regulated by two different hormones released from the hypothalamus: GH-releasing hormone (GHRH) increases GH release, and somatostatin inhibits GH release. Both hormones are secreted in a phasic manner, with GHRH contributing to high-amplitude GH pulses and somatostatin to trough levels. The dynamic interrelationship between these hypothalamic hormones is necessary for pulsatile secretion. A number of other factors either within the hypothalamus or in plasma contribute to the regulation of GH release, either by acting directly on the pituitary gland or by regulating hypothalamic somatostatin and GHRH secretion. Both GH and IGF-1 inhibit GH secretion in a typical feedback relationship: They increase somatostatin and decrease GH RH release from the hypothalamus, resulting in a suppression of GH release.

More recently, there has been increased interest in small peptides that stimulate the release of GH. One of these peptides, ghrelin, is produced in the stomach and gastrointestinal tract. Although it does not interact with the GHRH receptor, it has been shown to act synergistically with GHRH to regulate both GH release and food intake. This peptide is particularly important from a clinical standpoint since oral administration of peptide and nonpeptide analogs of ghrelin increase GH secretion.

Although GH is one of the most potent anabolic factors in the body, most of the anabolic actions of the hormone are regulated through the secretion of IGF-1. IGF-1 binds with high affinity to the type 1

IGF receptor found in tissues throughout the body. The IGF-1 receptor shares 50% amino sequence homology with the insulin receptor, and competitive binding and affinity cross-linking studies have demonstrated that IGF-1 binds to the insulin receptor with 100 times lower affinity than to the IGF-1 receptor. IGF-1 is synthesized mainly in the liver under the regulation of GH but is also synthesized in smaller quantities in almost all tissues. Although regulation of local or 'paracrine' IGF-1 activity is poorly understood, it represents an important source of the hormone that also appears to be regulated by GH.

IGF-1 secreted from the liver circulates in the blood, either free or bound to specific binding proteins that prolong the half-life of the peptide. At present, at least six binding proteins have been identified and constitute an intricate transport system for the IGFs that regulate their availability to specific tissues. It is now clear that IGF binding proteins are critical regulators of IGF-1 activity yet may have actions independent of IGF-1.

The actions of GH are not entirely mediated through IGF-1, and in muscle and adipose tissue, GH can have direct effects. Specifically, GH decreases insulin sensitivity (see the section titled 'Insulin and glucose regulation') by direct effects on skeletal muscle and increases release of free fatty acids (and decreases fat mass) by increasing the activity of hormone-sensitive lipase in adipose tissue. These findings have led to the concept that the major anabolic actions of GH are mediated through IGF-1 (through either endocrine or paracrine regulation), whereas the effects of GH on free fatty acid levels are mediated by direct actions on muscle and adipocytes (Figure 1).

Effects of Age on the GH/IGF-1 Axis

Alterations in the regulation of GH have an important role in the physiological and biochemical changes normally associated with aging. Early studies in humans indicated that the amplitude of GH pulses and the rise in the concentrations of GH after insulin-induced hypoglycemia were blunted in older individuals. Subsequent studies demonstrated a prominent decrease in the amplitude of GH pulses in older animals although there were no changes in either the basal concentrations or the ultradian rhythm of GH. The age-related decline in GH results in a progressive reduction in plasma concentrations of IGF-1 in both animals and humans. Subsequent studies in various strains of rats and mice, nonhuman primates, and humans consistently confirmed the decline in GH and IGF-1 concentrations with age and suggested that changes in these hormones are a robust marker of biological aging in mammalian species.

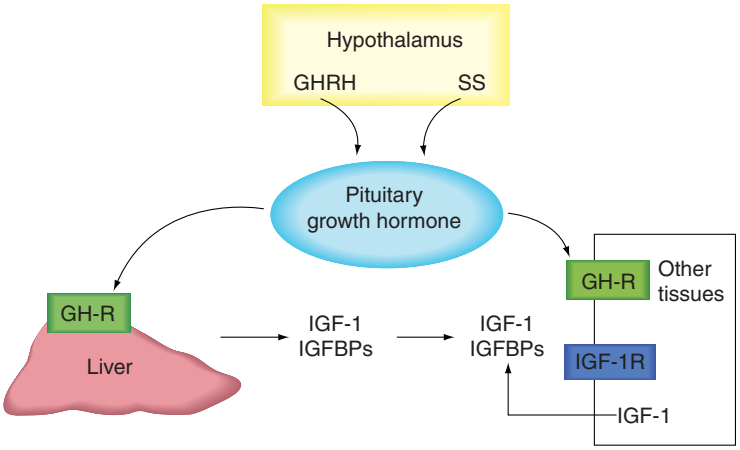

Figure 1 Growth hormone (GH), produced in the anterior pituitary, is modulated by two hypothalamic hormones: growth hormone-releasing hormone (GHRH), which stimulates both the synthesis and secretion of GH, and somatostatin (SS), which inhibits GH release in response to GHRH. GH also feeds back to inhibit GHRH secretion and probably has a direct inhibitory effect on secretion from the somatotroph (GH-producing cells). In mammals, GH is secreted in pulsatile bursts from the anterior pituitary gland, a pattern that is necessary to achieve full biological activity. GH binds with high affinity to its receptor (GH-R), found in tissues throughout the body, and activation of this receptor stimulates the synthesis and secretion of insulin-like growth factor 1 (IGF-1). Although 90% of circulating IGF-1 is synthesized and secreted by the liver, many types of cells, including some found in the brain and vasculature, are capable of IGF-1 production. Binding of the hormone to the IGF-1 receptor (IGF-1R) causes potent mitogenic effects, including increases in DNA, RNA, and protein synthesis. Although heterogeneity exists in the processing of IGF-1 messenger RNA, these transcripts appear to produce a single peptide that is homologous to the structure of proinsulin. Blood and tissue levels as well as activity of the peptide are regulated by IGF-1 binding proteins (IGFBPs). Although it was initially proposed that all the actions of GH were mediated through IGF-1, data from several studies support direct roles for GH in the regulation of lipolysis and insulin sensitivity that are independent of IGF-1. Reproduced from Carter CS, Ramsey MM, and Sonntag WE (2002) A critical analysis of the role of growth hormone and IGF-1 in aging and lifespan. *Trends in Genetics* 18: 295–301, with permission from Elsevier.

Restoration of GH and IGF-1 levels with purified GH preparations reverses the age-related decline in cellular protein synthetic capacity in both skeletal and cardiac muscle. Although these studies did not address the question of whether tissue response to GH diminished with age, they clearly indicated that (1) cellular protein synthesis decreases in an environment of reduced GH and IGF-1 levels and (2) GH and/or IGF-1 replacement has the capacity to attenuate a classical symptom of aging, that is, the decrease in cellular protein synthesis in older people. Furthermore, these studies were supported by clinical observations that GH replacement to humans for 6 months reversed the decline in bone mass, the negative nitrogen balance (an indicator of decreased whole-body protein synthesis), and the increase in fat mass characteristic of aging in humans. Subsequent reports demonstrated that GH or IGF-1 administration could partially reverse the decline in immune function; increase the expression of aortic elastin, lean body mass, skin thickness, and vertebral bone density; and improve cognitive function in older animals and humans. Although the beneficial effects of GH administration have been confirmed repeatedly, deleterious side effects of GH replacement severely limit its usefulness as a treatment to modulate physiological changes with age. For example, GH increases cartilage growth, contributes to carpal tunnel syndrome, impairs insulin sensitivity, increases blood glucose, and in several cases has been found to increase neoplastic disease. In fact, elevated IGF-1 levels are a risk factor for breast, gastrointestinal, prostate, and lung cancer. These studies provide compelling data suggesting that the development of the aging phenotype results, at least in part, from dysfunction within the neuroendocrine axis and, more specifically, deficiencies in the secretion of GH and IGF-1. Therefore, it is evident that decreases in GH and IGF-1 have clinical significance and are a contributing factor in the genesis of tissue dysfunction in older animals and humans.

Actions of GH and IGF-1 in Older Animals

Although GH and IGF-1 are well known for their ability to increase cellular protein synthesis, muscle and bone mass, and immune function and to decrease fat mass, recent data suggest that the age-related decrease in these hormones has an important role in the genesis of brain aging. Studies indicate that IGF-1 reverses the age-related decline in cerebral microvascular density and blood flow and that IGF-1 crosses the blood–brain barrier and enhances learning and memory, long-term potentiation, glucose metabolism, synaptic density, neurogenesis, and neuronal survival and reduces oxidative stress in older animals. These results and recently published reports have led to the hypothesis that age-related deficiency in circulating GH and IGF-1 leads to vascular insufficiency and deficits in synaptic function and cell replacement in the central nervous system and are a contributing factor in the genesis of brain aging (Figure 2).

Animal Models of GH/IGF-1 Excess and Deficiency

Our understanding of the effects of GH/IGF-1 deficiency and replacement in aging humans and animal models is complicated by results of studies with transgenic and mutant animals. Transgenic mice with varying levels of GH overexpression demonstrate an inverse relationship between GH expression and life span. The reduction in life span observed with

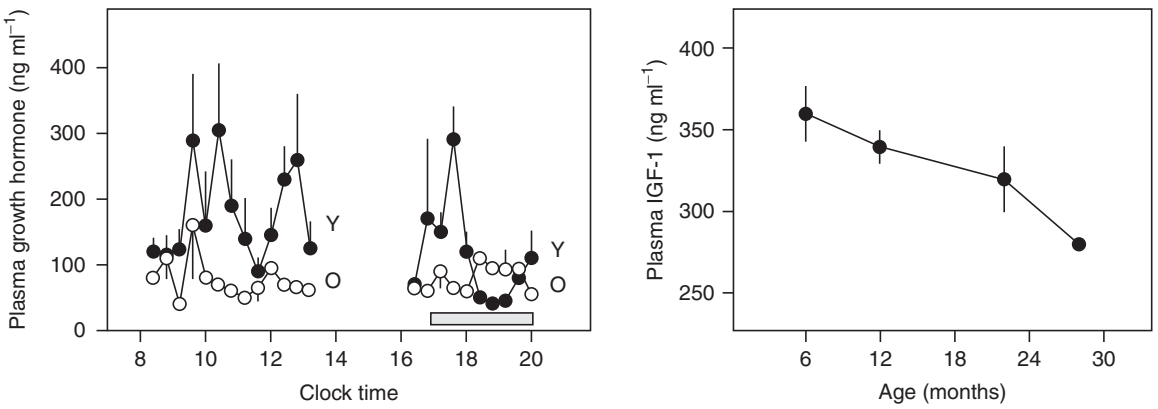

Figure 2 Age-related changes in the growth hormone/IGF-1 axis. Left, growth hormone secretory pulses in young (6-month-old; solid circles; Y) and old (20-month-old; open circles; O) Sprague–Dawley male rats. Serial samples of blood were removed at 20-min intervals from conscious, freely moving animals and analyzed by radioimmunoassay. Hatched area represents the dark phase of the light/dark cycle. Right, age-related changes in plasma IGF-1. Data represent mean ± SE. Reproduced from Carter CS, Ramsey MM, and Sonntag WE (2002) A critical analysis of the role of growth hormone and IGF-1 in aging and lifespan. *Trends in Genetics* 18: 295–301, with permission from Elsevier.

overexpression of GH is generally attributed to increased pathologies, primarily renal lesions and a very high frequency of hepatocellular tumors, including adenoma and carcinoma. While the reduced life span observed in models of GH overexpression appears to support the hypothesis that GH reduces life span, it is important to note that GH levels in these overexpression models are 1000–10 000-fold greater than those found under physiological conditions. Therefore, the distinction between the potential of pharmacological doses of GH to increase pathology and reduce life span has to be considered separately from the physiological changes in growth that occur during the normal life span of most mammalian models. Therefore, it remains unclear whether transgenic models of GH overexpression provide relevant information on the role of GH and IGF-1 in normal life span regulation (Figure 3).

Perhaps the most commonly used animal models to investigate the effects of GH and IGF-1 are those that show deficiency in GH expression or have mutations in the GH/IGF-1 axis. Some of these models include those with deficiencies in transcription factors necessary for the generation of GH, TSH, and prolactin-producing cells in the pituitary gland (e.g., Snell and Ames dwarf mice), and others have selective impairments in the GH/IGF-1 axis (e.g., $GHRHR^{-}/GHRHR^{-}$, GHRKO, and $IGFR^{+/-}$ mice). Despite the limitations of the models with multiple hormonal deficiencies, even models with specific deficiencies in GH and IGF-1 during early development result in impaired development of related endocrine systems, and alterations in corticosterone, insulin, glucose, and thyroid hormones are routinely observed. As a result, data from these models must be interpreted cautiously since they demonstrate the potential of alterations in GH/IGF-1 early during development to affect developmental processes that regulate life span but cannot address whether these effects are mediated directly by the GH/IGF-1 actions on target tissues or are mediated indirectly via the developmental other endocrine systems dependent on GH and IGF-1. The distinction is important to resolve for the direction of future scientific studies.

The marked contrasts noted in the studies described above related to the beneficial effects of GH replacement on function in humans and animal models and the contrasting studies demonstrating that deficiencies in GH and IGF-1 increase life span represent the crux of the current issue in the GH/IGF-1 field. Recent studies have attempted to circumvent the developmental effects of GH/IGF-1 deficiency by developing a model of adult-onset GH/IGF-1 deficiency. In this model, a mutant animal that exhibits reduced levels of GH around puberty is administered GH replacement to increase circulating IGF-1 and ensure the development of other endocrine systems until adulthood. The GH injections are then terminated and the animal allowed to age normally, creating the model of adult-onset GH deficiency (AO-GHD), and these animals are compared to dwarf (GH-deficient) or wild-type (normal) animals. In these studies, no differences in life span were evident between normal and GH-deficient animals, but it is interesting that AO-GHD animals lived 15% longer than either group. End-of-life pathology indicated that both AO-GHD and GH-deficient animals had less kidney pathology and neoplastic disease compared with wild-type animals but died primarily from cardiac thrombus and intracerebral hemorrhage. The increased life span in the AO-GHD animals resulted from a corresponding delay in the appearance of cardiac thrombus and intracerebral hemorrhage. This study provided compelling data that GH and IGF-1 levels around puberty have an important role in determining life span, that life span and the factors that determine life span may be a separate entity from age-related pathology, and finally that functional changes with age are separate from life span.

The effects of GH and IGF-1 on aging and life span highlight several unresolved issues in the field

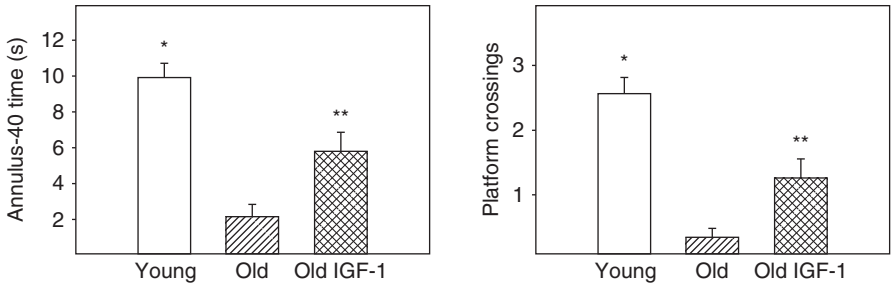

Figure 3 Probe trial in the Morris water maze assessed spatial reference memory. Performance in both Annulus-40 time (left) and platform crossings (right) that are measures of memory declined with age ($p < 0.001$ comparing young and old control (old) animals). IGF-1 treatment for 28 days reversed this decline in both measures (*$p < 0.05$ comparing young to old and old IGF–1; **$p < 0.01$ in both measures comparing old control (old) and old IGF-1 treated animals). Data are shown as mean ± SE (unpublished data).

of gerontology. Deletion or mutation of a single gene, as occurs in transgenic or mutant animals, can result in developmental anomalies, but consequences for data interpretation remain uncertain. Most investigators recognize the importance of the age-related decline in tissue function, pathology, and life span; however, few studies of life span include either functional or pathological analyses. Until gerontologists agree on standards for the conduct of aging studies designed to assess the effects of these interventions on aging, such issues will remain unresolved (Table 1).

Insulin and Glucose Regulation

Since the initial studies demonstrating that suppression of signaling in the insulin/IGF-1 pathway in *C. elegans* increases life span, there has been much interest in insulin signaling in mammalian models and its potential contribution to life span. In contrast to the common insulin/IGF-1 pathway in invertebrates, separate insulin and IGF-1 signaling pathways have evolved in mammals.

In humans, the regulation of blood glucose and insulin secretion is perhaps one of the most important areas of endocrinology because pertubations in this pathways are closely associated with cardiovascular disease and many other pathological conditions that increase with age. The postprandial rise in insulin secretion in mammals results from an increase in blood glucose that affects the pancreatic β cell. Also, glucagon-like peptide-1 release from the intestine sensitizes the β cells, resulting in optimal insulin output into the circulation. Although insulin action is a more complicated process than described here, its primary action is to stimulate glycogen synthesis in liver and, after binding to its receptor on muscle and adipose tissue, to activate glucose transporter-4 proteins in the membrane of the cell through a complicated signal transduction mechanism. The result is glucose uptake in these tissues and a reduction in blood glucose levels. In nonobese older study participants, there are minor, albeit significant, changes in both insulin secretion and glucose tolerance tests with age. These changes alone are not sufficient to result in pathology. However, in obese study participants regardless of age, basal glucose levels rise, resulting in increased pathology. The interaction of body composition (adipose tissue mass) and glucose levels has been widely recognized and results in insulin receptor insensitivity, potentially from a rise in cytokines (e.g., tumor necrosis factor-α and interleukin-6) and other hormones (e.g., leptin) secreted by adipose tissue, that interferes with insulin signaling mechanisms designed to lower blood glucose. In addition, recent studies indicate that adipose-like tissue accumulates within skeletal muscle and is an important factor in the increased incidence of disability with age. Whether these changes in adipose tissue are important corollary events or actually cause the increase in disability remains to be determined.

Unfortunately, it is unclear whether alterations in insulin levels or insulin signaling directly contribute to the development of the aging phenotype. Certainly, one might expect that obesity and the pathologies associated with obesity synergize with the pathologies of aging to shorten life span. But the interactions between these pathways are complex and remain underinvestigated. In this regard, there are data to

Table 1 End-of-life pathology in a model of adult-onset growth hormone deficiency[a]

Pathology	*Heterozygous*	*GHD*	*AO-GHD*
Tumor bearing (%)	88	73 ($p = 0.34$)	63 ($p = 0.04$)
Tumor burden (mean)	1.3 ± 0.2	0.92 ± 0.2 ($p =$ ns)	1.0 ± 0.2 ($p =$ ns)
Fatal tumor (%)	57	36 ($p = 0.21$)	31 ($p = 0.11$)
Fatal pituitary adenoma (%)	39	21 ($p = 0.30$)	13 ($p = 0.08$)
Grade of pituitary adenoma	1.57 ± 0.3	1.69 ± 0.4 ($p =$ ns)	1.19 ± 0.3 ($p =$ ns)
Fatal chronic nephropathy (%)	74	0 ($p < 0.0001$)	0 ($p < 0.0001$)
Severity of nephropathy	4.0 ± 0.3	0.8 ± 0.2 ($p < 0.001$)	1.0 ± 0.2 ($p < 0.001$)
Intracranial hemorrhage/thrombus (%)	17	50 ($p = 0.06$)	44 ($p = 0.14$)
Age of occurrence of hemorrhage/thrombus (months)	31.9	30.7	34.1
Disease burden (total)	6.6 ± 0.5	5.0 ± 0.5 ($p = 0.009$)	5.5 ± 0.4 ($p = 0.067$)

[a]End-of-life pathology was determined in the following experimental groups: (1) heterozygous (normal) male Lewis rats injected with saline from 4 to 14 weeks of age (normal size, normal GH/IGF-1 levels); (2) GH-deficient (GHD) dwarf Lewis males administered saline from 4 to 14 weeks of age (dwarf, GHD); and (3) GHD dwarf Lewis males administered porcine growth hormone from 4 to 14 weeks of age (Adult-onset GHD (AO-GHD)). AO-GHD animals exhibited a 12–15% increase in life span compared with either heterozygous or GHD animals ($p < 0.05$). ns, not significant.

Reproduced from Sonntag WE, Carter CS, Ikeno Y, et al. (2005) Adult-onset growth hormone and insulin-like growth factor I deficiency reduces neoplastic disease, modifies age-related pathology, and increases life span. *Endocrinology* 146: 2920–2932, copyright 2005, The Endocrine Society.

suggest that a knockout of the insulin receptor specifically in adipose tissue increases life span in mice, but no age-related pathological analyses or performance measures were done, and therefore this remains an important area for future research.

Endocrine Effects of Moderate Caloric Restriction: Interactions with Insulin and IGF-1

Despite the absence of changes in insulin and glucose levels in nonobese humans with age, there is important information to suggest that the insulin and IGF-1 pathways may have important roles in determination of life span in invertebrate and vertebrate subjects. Consistently, moderate caloric restriction (60% of *ad libitum* feeding) has been shown to increase life span in rodent models. Although there are data to suggest that the decrease in nutrients alters several cellular metabolic pathways (e.g., SirT1, PGC1α) that may contribute to regulation of life span, alterations in caloric intake have important effects on several hormonal systems that may affect pathology and life span. For example, reduction in calories increases GH, cortisol, and catecholamine secretion and reduces circulating insulin and IGF-1 levels. These changes are a normal physiological response designed to access nutrients that have been stored during periods of nutrient excess. The result of these changes is an increase in gluconeogenesis, glycogenolysis, ketogenesis, and fatty acid metabolism and a decrease in insulin sensitivity and cellular protein synthesis, all of which are designed to maintain blood glucose and other nutrients in the bloodstream. The endocrine changes that result from a reduction in caloric intake appear to have a direct effect on life span and age-related pathology in small animals. For example, chemical-induced bladder cancer is markedly reduced in caloric restricted compared with *ad libitum* animals, and replacement of IGF-1 to calorie restricted animals increases the incidence of this pathology. These data suggest that the decrease in IGF-1 that occurs during caloric restriction is an important part of the protective mechanism against chemical-induced cancer and provides indirect evidence that chronically lower levels of IGF-1 may protect animals from other types of neoplastic diseases that increase with age. This conclusion is supported by data from many transgenic and mutant animal strains and studies using the model of AO-GHD that show protection against dimethylbenz[a]anthracene-induced mammary cancer. In the studies using the AO-GHD model, GH replacement resulted in a dose-related rise in mammary cancer. Since many mouse strains are highly susceptible to neoplastic disease, any intervention that delays or prevents this pathology will naturally increase life span. Although it is unlikely that the increased life span from moderate caloric restriction results from reduced levels of IGF-1 and/or insulin alone, the relationship between IGF-1 and neoplastic disease is compelling, and further research to understand the specific hormones that may mediate the actions of moderate caloric restriction is required (Figure 4).

Prolactin

Basic Regulation and Actions

Prolactin secretion from the pituitary lactotroph in humans and animal models is regulated by hypothalamic dopamine secretion into the hypothalamic-pituitary portal vessels. The dopaminergic neurons of the rat hypothalamus are grouped into two main areas, A_{12} and A_{14}. The dopaminergic perikarya of the A_{12} area are located in the arcuate nucleus, with their axons projecting into the external zone of the median eminence. There, dopamine is released from the axon terminals into the portal vessels of the median eminence and travels to the anterior pituitary lobe. The A_{14} dopamine neurons are primarily located within the paraventricular and periventricular

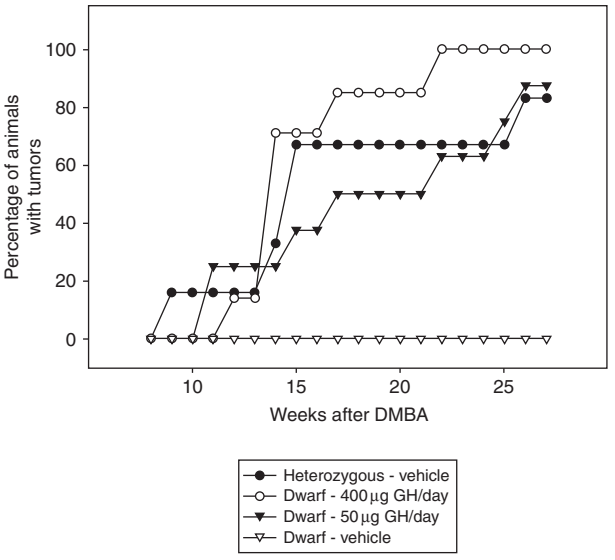

Figure 4 Cumulative tumor incidence after dimethylbenz[a] anthracene (DMBA) injection in growth hormone-deficient dwarf animals and heterozygous (normal) control female rats. At 27 weeks after DMBA injection, 80% of heterozygous animals exhibited mammary tumors, while dwarf animals administered saline, 50 µg, or 400 µg porcine growth hormone per day demonstrated 0%, 83%, and 100% tumor incidence, respectively. Reproduced from Ramsey MM, Ingram RL, Cashion AB, et al. (2002) Growth harmone-deficient dwarf animals are resistant to dimethylbenzanthracine (DMBA)-induced mammary carcinogenesis. *Endocrinology* 143: 4139–4142, copyright 2002, The Endocrine Society.

hypothalamic nuclei and project their axons through the internal portion of the median eminence to terminate on both the neural and intermediate pituitary lobes. The A_{12} area and its corresponding axon terminals constitute the tuberoinfundibular dopaminergic (TIDA) system, whereas the A_{14} area and its fibers are known as the periventricular dopaminergic system. Both systems regulate lactotrophs exerting a tonic inhibitory control on prolactin secretion and lactotroph proliferation.

Plasma concentrations of prolactin are lowest during waking hours in humans and highest during sleep. The variable patterns of secretion are dependent on gender, environmental conditions including stress, reproductive status (including lactation), estrous and menstrual cycles, pregnancy, and other factors. Prolactin participates in multiple roles throughout the body although its primary actions are related to reproduction and maternal behavior. In some mammals, prolactin is important for the secretory activity of the corpus luteum. Within the mammary gland, prolactin induces growth and synthesis of milk (e.g., uptake of some amino acids; synthesis of the milk proteins casein and α-lactalbumin; uptake of glucose, lactose, and fats) and maintains milk secretion. Also, prolactin has a role in the immune system, stimulating mitogenesis in normal lymphocytes. Prolactin also appears to have a role in the regulation of solute and water transport across mammalian cell membranes. In keeping with its lactogenic properties, it affects water transport across amniotic membranes and acts as an osmoregulatory hormone.

Prolactin Secretion and Aging

The dopaminergic neurons of the central nervous system are among the cells most susceptible to damage produced by the aging process. For example, Parkinson's disease is a human condition characterized by a degeneration of nigrostriatal dopamine neurons. This disease affects 0.1–0.3% of the population and is the most conspicuous reflection of the vulnerability of dopamine neurons to age. In female rodents, degeneration and loss of TIDA neurons during normal aging is associated with progressive hyperprolactinemia and the development of pituitary prolactinomas. It is interesting that patients with Parkinson's disease usually reveal functional alterations within the hypothalamoprolactin axis.

Female rats ranging in age from 6 to 32 months have been studied to ascertain the impact of aging on the number and morphology of dopamine neurons of the A_{12} and A_{14} areas as well as on serum prolactin levels. In these studies, a progressive loss of both TIDA and periventricular dopaminergic neurons was evident, and these changes became more conspicuous at extreme ages. It is interesting that in old female rats, restoration of hypothalamic IGF-1 using recombinant adenoviral vectors that express rat IGF-1 was highly effective in restoring hypothalamic dopamine function, suggesting important interactions between the age-related decrease in IGF-1, the deterioration of dopamine activity, and the age-related rise in prolactin levels.

In addition to IGF-1, hypothalamic lactotrophs are controlled by a number of factors. For example, estrogen has been demonstrated to have an important effect on lactotrophs and to stimulate the development of tumorigenesis within the pituitary. Although there are data demonstrating that estrogen acts on dopaminergic cells within the hypothalamus, several investigators have proposed a lactotroph, folliculostellate cell interaction as a potential mechanism for the role of estrogen in the genesis of pituitary prolactinomas. Under basal conditions, prolactin cells are under a tonic inhibitory autofeedback of transforming growth factor-β (TGFβ)-1 acting on the TGFβII receptor. When estrogen levels rise, lactotroph cells are stimulated to release high levels of TGFβ3, which then acts on TGFβII receptors on folliculostellate cells, inducing the production and release of basic fibroblast growth factor, which in turn acts on the lactotroph to stimulate both cell proliferation and prolactin secretion. These novel interactions of estrogen on the pituitary gland may interact synergistically with the reduced levels of dopamine to produce the pronounced rise in pituitary microadenomas and frank tumors that occurs with age.

TSH: Tri-iodothyronine, Thyroxine, and Thyroid Function

Cellular metabolism is regulated by the secretion of TSH release from the anterior pituitary gland. This hormone enters the circulation and stimulates the thyroid gland to produce both tri-iodothyronine (T_3) and thyroxine (T_4). These hormones circulate bound to thyroid binding globulin and after transport are taken up by the cell. T_3 is the more potent thyroid hormone, but T_4 can be converted to T_3 with the enzyme deiodinase. Levels of these hormones regulate basal metabolic activity throughout the body. T_3 and T_4 are controlled by negative feedback regulation, primarily at the level of the anterior pituitary gland, by suppression of TSH release although the hypothalamus also produces thyroid-releasing hormone (TRH) that can stimulate TSH release. TRH has only a minor role in the regulation of TSH release under normal conditions, but in response to cold stress or other challenges, TRH exerts an important influence. In addition to prominent effects on cellular metabolism and temperature regulation, thyroid hormones have a permissive role in the regulation of

many organs and tissues, including reproductive function, bone metabolism, and glucose metabolism. Therefore, changes in the secretion or response to thyroid hormones would be expected to have broad-ranging effects on the health of the organism.

Apart from the appearance of specific thyroid diseases that increase with age, only subtle changes in the levels of thyroid hormones seem to occur in otherwise healthy older animals or humans. Investigators have reported a reduction in deiodinase activity that leads to reduced levels of T_3 within target organs, but the significance of these changes for the development of physiological changes in tissue function have not been well defined. Similarly, it is of interest that numerous animal models that exhibit increased life span also exhibit a decrease in thyroid hormone levels, but the significance of these changes remains to be determined.

Conclusions

There are now convincing data that alterations within the neuroendocrine system contribute to both the physiological and pathological manifestations of aging. Decreases in GH and IGF-1, for example, are important factors in the accumulation of body fat mass, the reduction in cellular protein synthesis, and impairments in some aspects of immune function with age. The effects of moderate caloric restriction are mediated, at least in part, through physiological changes affected by the neuroendocrine system. These studies, together with results from transgenic and mutant animals, provide compelling data of an important link between the neuroendocrine axis and the development of the aging phenotype. Nevertheless, controversies remain that can be resolved only by the development of transgenic animal models permitting close regulation of both GH and IGF-1 and assessing function, pathology, and life span.

Although the specific etiology of the age-related neuroendocrine dysfunctions is unknown, increases in oxidative damage in specific hypothalamic nuclei, impairments in neurotransmitter levels and turnover, and diminished plasticity within hypothalamic nuclei are potential mechanisms. An alternative explanation for the alterations within the neuroendocrine system is that an active regulation of hypothalamic function exists throughout the life span. Decreases in GH and IGF-1, for example, are initiated immediately after puberty and progress with age. In fact, age-related changes in the regulation of this system may be consistent with a model of antagonistic pleiotropy, a theory on the evolution of aging that suggests that the expression of particular genes is beneficial early in life but becomes detrimental as the organism ages. Certainly, the beneficial actions of GH and IGF-1 on muscle mass and organ development early in life would be expected to increase general fitness as well as reproductive fitness, whereas continued high levels may increase pathology and limit life span. Thus, the age-related decrease in GH and IGF-1 as well as other hormones regulated by the neuroendocrine axis may be a physiologically regulated, adaptive process rather than the result of damage to specific hypothalamic nuclei. The potential significance of such changes for this and other systems regulated by the neuroendocrine axis awaits further investigation.

See also: Growth Hormone ; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Insulin-Like Growth Factor Signaling and Actions in Brain; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Neuroendocrine Aging: Pituitary-Adrenal Axis.

Further Reading

Bartke A (2005) Minireview: Role of the growth hormone/insulin-like growth factor system in mammalian aging. *Endocrinology* 146: 3718–3723.

Bartke A, Masternak MM, Al-Regaiey KA, and Bonkowski MS (2007) Effects of dietary restriction on the expression of insulin-signaling-related genes in long-lived mutant mice. *Interdisciplinary Topics in Gerontology* 35: 69–82.

Carter CS, Ramsey MM, and Sonntag WE (2002) A critical analysis of the role of growth hormone and IGF-1 in aging and lifespan. *Trends in Genetics* 18: 295–301.

Goya RG, Sarkar DK, Brown OA, and Herenu CB (2004) Potential of gene therapy for the treatment of pituitary tumors. *Current Gene Therapy* 4: 79–87.

Harper JM, Durkee SJ, Dysko RC, Austad SN, and Miller RA (2006) Genetic modulation of hormone levels and life span in hybrids between laboratory and wild-derived mice. *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 61: 1019–1029.

Herenu CB, Brown OA, Sosa YE, et al. (2006) The neuroendocrine system as a model to evaluate experimental gene therapy. *Current Gene Therapy* 6: 125–129.

Herenu CB, Cristina C, Rimoldi OJ, et al. (2007) Restorative effect of insulin-like growth factor-I gene therapy in the hypothalamus of senile rats with dopaminergic dysfunction. *Gene Therapy* 14: 237–245.

Kenyon C (2005) The plasticity of aging: Insights from long-lived mutants. *Cell* 120: 449–460.

Ramsey MM, Ingram RL, Cashion AB, et al. (2002) Growth harmone-deficient dwarf animals are resistant to dimethylbenzanthricine (DMBA)-induced mammary carcinogenesis. *Endocrinology* 143: 4139–42.

Sarkar DK (2006) Genesis of prolactinomas: Studies using estrogen-treated animals. *Frontiers in Hormone Research* 35: 32–49.

Sonntag WE, Carter CS, Ikeno Y, et al. (2005) Adult-onset growth hormone and insulin-like growth factor I deficiency reduces neoplastic disease, modifies age-related pathology, and increases life span. *Endocrinology* 146: 2920–2932.

Sonntag WE, Ramsey M, and Carter CS (2005) Growth hormone and insulin-like growth factor-1 (IGF-1) and their influence on cognitive aging. *Ageing Research Reviews* 4: 195–212.

Neuroendocrine Aging: Pituitary-Adrenal Axis

E R de Kloet, LACDR, Leiden University Medical Center, Leiden, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Aging is characterized by decreased ability to maintain homeostasis and consequent less efficient adaptation to change. The impact of this age-related breakdown of adaptation is far more disruptive when integrative homeostatic communication systems are also affected. These systems include the sympathetic and parasympathetic branches of the autonomic nervous system and the neuroendocrine hypothalamic–pituitary–adrenal (HPA) axis (Figure 1). The amine, neuropeptide, and steroid components of these systems signal specific information to ensure coordination of brain and body functions during adaptation. Such coordination involves the control of emotions and cognitive performance as well as energy metabolism, cardiovascular, inflammatory, and immune regulation. The question addressed here is whether aberrant HPA activity is the cause or the consequence of signs of aging and age-related pathology. The focus is on the glucocorticoids, which operate in a bimodal fashion crucial for defense and recovery of homeostasis.

HPA Axis and Glucocorticoids

Glucocorticoid Cascade Hypothesis

The HPA axis promotes an organisms' ability to cope with environmental stressors and facilitates behavioral adaptation. At senescence, coping and adaptation usually become less efficient and as a consequence HPA reactions may become inadequate, excessive, or prolonged. In particular, the excessive and prolonged secretion of glucocorticoids has received attention. Elevated glucocorticoid levels imply that the cost to defend homeostasis has become too high. The outcome is enhanced wear and tear, a process termed 'allostatic load' by Bruce McEwen. This notion materialized in Robert Sapolsky's glucocorticoid cascade hypothesis, which focused on the damaging outcome of chronic glucocorticoid overexposure during the aging process.

While acute rises in glucocorticoid levels are essential for recovery, adaptation, and survival, prolonged elevated glucocorticoid levels may indeed become damaging to brain and body. It is thought that such elevated hormone levels are primarily caused by slowing down of cognitive operations and less efficient homeostatic control mechanisms. In the brain, the damage particularly affects the hippocampus, which is an important glucocorticoid target and is crucial for learning and memory processes. Because of ensuing feedback resistance, glucocorticoid levels further rise. Hence a feed-forward cascade develops, leading ultimately to degeneration and death. This appealing hypothesis clearly has supporting evidence: for instance, rainbow trout swimming upriver for reproductive purposes eventually die of cushingoid symptoms.

HPA Axis in Aging Rodents

In support of the glucocorticoid cascade hypothesis, excessive HPA axis activation in aging has been reported. Sprague–Dawley, Long–Evans, Fisher, and Wistar rats generally show prolonged stress-induced corticosterone levels during aging. Several scenarios have linked rises in corticosterone level to aging. First, stress and hypercorticism have been proposed to promote cognitive aging. Second, aging may cause initial hypercorticism, which then, in a feed-forward cascade of increasing feedback resistance, would further promote the aging process. Third, stress, together with hypercorticism, would amplify the individual variation in cognitive aging. However, hypercorticism during aging appears not at all to be a common phenomenon.

In aged Brown Norway rats, exposure to a novel environment or a conditioned emotional stressor produces excessive adrenocorticotropic hormone (ACTH) peak values, while corticosterone responses are not different from those of young controls, indicating an efficient feedback operation. The relatively low corticosterone concentration is partly compensated by reduced corticosteroid-binding globulin (CBG) capacity, yielding an increased free fraction of nonbound corticosterone. Hence, Brown Norway rats show adrenocortical deficiency during aging, suggesting that healthy aging and low corticosterone levels coincide. This demonstrates, however, that hypercorticism is not an inevitable consequence of aging.

The profound differences in 50% survival age between different rat strains raise the question of the appropriate rodent model for aging research. The criteria for the 'ideal' model include an almost rectangular survival curve and thus a high estimated 50% survival age. Included also is the occurrence of age-related, multiple pathologies rather than inbred-strain specific diseases (e.g., the high prevalence of testis- or pituitary-tumors in F344 or Wistar rats). The animals used in many aging studies in fact either fail to meet these criteria or should be considered middle-aged rather than old. The Brown Norway strain does

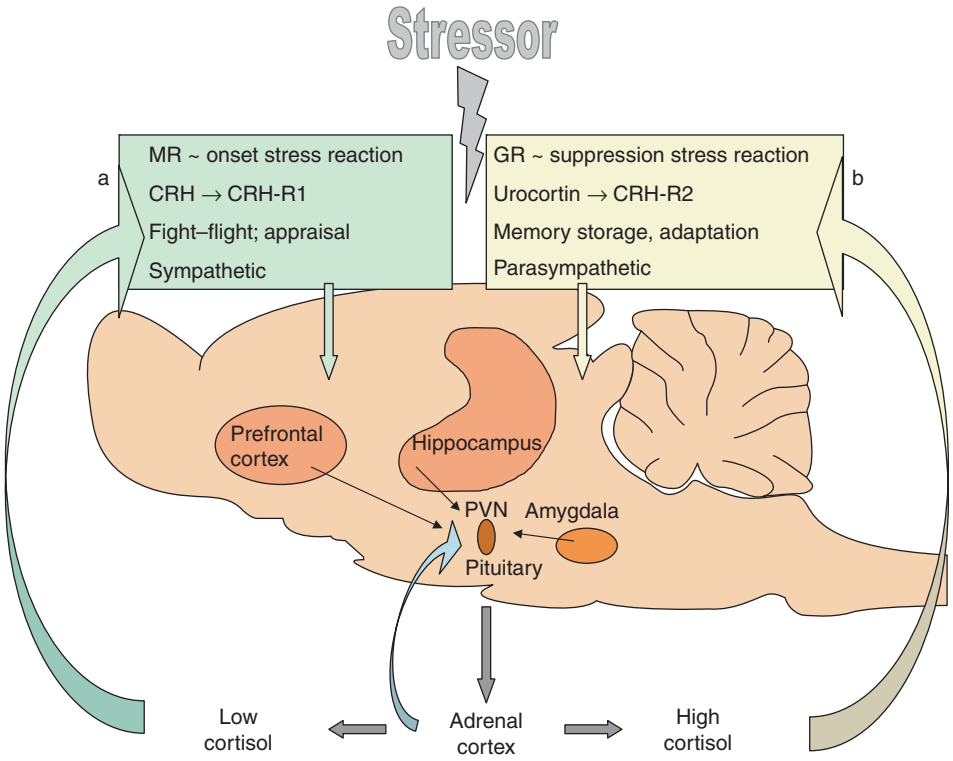

Figure 1 Bimodal organization of neuroendocrine stress response. Stressful stimuli, through afferent pathways, stimulate corticotropin-releasing hormone (CRH) and vasopressin neurons in the paraventricular nucleus (PVN), thus orchestrating behavioral, autonomous, and neuroendocrine response patterns. Physical stressors (e.g., pain, blood loss, heat, and cold) directly stimulate the PVN via ascending aminergic pathways. Psychological stressors require processing in limbic–cortical brain structures (e.g., hippocampus, amygdala, and prefrontal cortex) that communicate transsynaptically with the PVN. The neuroendocrine response pattern enhances the secretory bursts of cortisol (primates) or corticosterone (rodents) from the adrenal cortex. The corticosteroids feed back on the brain circuits that have produced their own secretion and prevent them from overexcitation. These actions exerted by the steroids are mediated by (a) mineralocorticoid receptors (MR) and (b) glucocorticoid receptors (GR). Under psychosocial stress conditions, MRs regulate the onset of the stress response, which – through CRH receptor 1 (CRH-R1) – determines the extent of CRH action on hypothalamic–adrenal–pituitary activation, sympathetic nervous system responses, and flight–flight behavior. After stress progressively activates GRs, the rise in corticosterone levels facilitates adaptation, promotes memory storage, and hence prepares for coping with future events. CRH-R2 activation by urocortins and parasympathetic nervous activity is thought to synergize with GRs to promote recovery. MRs and GRs operate in complementary fashion through a (non)genomic mechanism to facilitate defense and recovery of homeostasis; imbalance is thought to promote signs of aging and age-related pathology. From de Kloet ER, de Rijk RH, and Meijer OC (2007) Is there an imbalanced response of mineralocorticoid and glucocorticoid receptors in depression? *Nature Clinical Practice Endocrinology & Metabolism* 3: 168–179.

meet these criteria and is, as such, generally accepted as a good rodent model to study aging.

Another important consideration in these types of studies is the individual variation within a genetically homogeneous cohort of animals – variation that becomes much more pronounced at old age. With respect to many behavioral and neurobiological parameters, such as memory performance, individual animals can often be classified in impaired, intermediate, and nonimpaired subgroups. In aged Long–Evans or Brown Norway rats, approximately 30% will show excellent memory performance, whereas the rest will be partially or severely impaired; in Long–Evans rats, the extent of impairment appears related to the persistence in corticosterone secretion after stress.

HPA Axis Pulsatility

Glucocorticoids are secreted in a circadian pattern, with peak levels before the active period, which, for the nocturnal rat, is in early evening. However, hormones are secreted in distinct pulses, as is well documented for the gonadotropin-releasing hormone and growth hormone systems. Also, the HPA axis displays an ultradian rhythm and produces a complex pattern of approximately 1-h-long glucocorticoid secretory pulses. The variation in amplitude of the glucocorticoid pulses produces the circadian profile.

A stressor initiates an HPA pulse. The pulse frequency and amplitude increase, but it makes a difference at what phase of the ultradian rhythm the

stressor is experienced. As shown in Stafford Lightman's laboratory, a stressor at the ascending phase of the pulse results in a profound glucocorticoid response, while at the descending phase a refractory period exists, leading to strongly attenuated glucocorticoid secretion. This raises the possibility that modulation of glucocorticoid pulsatility serves as means of communication. Frequency encoding by glucocorticoid hormones implies that the pattern of hormone secretion matters, rather than absolute hormone concentration. It implies also that glucocorticoid pulses throughout the circadian cycle likely are required for synchronization and coordination of daily and sleep-related events.

Glucocorticoid pulsatility has an immediate consequences for HPA aging studies. This is one may wonder about the relationship between cortisol changes observed after a stressor if disparate pulsatility patterns among individuals are grouped. In some rat strains, aging is characterized by a prolonged response to stress. Taking the ultradian rhythm into account, this apparent persistence may be due to desynchronization of the pulsatile corticosterone pattern among individuals: the pulse may be delayed in some and advanced in others. Such disordered pulsatile patterns may appear seemingly prolonged after grouping the data of each individual. Hence the pulsatile pattern of corticosterone needs to be examined during the aging process.

Glucocorticoid Feedback

The recognition of pulsatile patterns of hormone secretion casts an entirely new light on the classical view of regulation and function of the HPA axis. The pulse generator is not known, but probably resides proximal to or within the hypothalamic corticotropin-releasing hormone (CRH) neurons. Furthermore, hormone pulses can be amplified and dampened at each level of the HPA axis. Central modulation occurs via afferent inputs that are involved in processing of stressful information. In this respect, excitatory afferents from the hippocampus and prefrontal cortex that process contextual information innervate a γ-aminobutyric acid (GABA)ergic inhibitory neuronal network controlling the CRH neurons. Emotional arousal through nervous input from the amygdala excites the paraventricular nucleus (PVN), as do the monosynaptic ascending aminergic pathways from the brain stem. How these inputs are integrated and translated in a CRH pulse is not known; current evidence suggests that a stressor advances the onset of the pulse and can enhance pulse frequency and amplitude.

Glucocorticoids feed back precisely in those cells and circuits that initially activated the HPA pulse generator, but in what modes does the feedback action operate? The answer to this question has been guided by the discovery of the receptors for the glucocorticoids. Two types of receptors have been identified: mineralocorticoid receptors (MRs) and the glucocorticoid receptors (GRs), which together control a spectrum of actions over time domains ranging from seconds to hours. The GRs prevail in the PVN and pituitary primary feedback sites, where they exert negative control over the stress-induced HPA pulse. The limbic–midbrain afferents express both MRs and GRs that modulate the input secondarily in either feed-forward or feedback mode, depending on time and context, and thus are capable of modulating the HPA pulse either way. The MR is promiscuous and binds aldosterone as well as corticosterone and cortisol, but is exposed to much larger amounts of the latter steroid. This is because the limbic brain lacks the enzyme 11β-hydroxysteroid dehydrogenase type 2 (11-HSD-2), which confers aldosterone selectivity to MRs in kidney cells. Instead, limbic neurons express 11-HSD-1, a reductase that regenerates bioactive glucocorticoids.

Glucocorticoid action mediated by limbic MRs and GRs modulates higher brain functions involved in behavioral and physiological adaptations to the stressor. GRs become substantially occupied at glucocorticoid peaks during ultradian pulses and after stress, while MRs are substantially occupied already at the troughs. The MR- and GR-mediated glucocorticoid actions act in concert with the other stress signaling systems. Figure 1 depicts the bimodal nature of the stress system, with CRH/CRH receptor 1 (CRH-R1) and corticosterone/cortisol through MRs controlling the onset of the pulse, and possibly urocortin/CRH-R2, with GRs facilitating the termination of the pulse. This dualism in the stress system presents a new working model to examine the HPA axis during the aging process.

Brain Corticosteroid Receptors

Corticosteroid Receptor Balance and Aging

Crucial for stress and aging are two fundamental observations. First, MRs and GRs act as transcription factors and they are co-localized in limbic neurons, but have distinctly different properties. MRs have a tenfold higher affinity than GRs do for cortisol/corticosterone, and both act through transactivation at overlapping but distinct hormone response elements (HREs). Additionally, GRs are capable of interacting with common transcription factors such as activator protein-1 (AP-1) and nuclear factor-kappa B (NF-κB), with rapid effects on gene transcription by transrepression. Thus,

in response to a brief corticosterone pulse, waves of primary and secondary gene responses occur and initially act through predominant transrepression, probably dampening the initial stress reactions. Then transactivation prevails to exert recovery of homeostasis and adaptation.

Second, it was recently discovered by Henk Karst and Marian Joëls that the very same nuclear MRs also mediate nongenomic actions in the same neurons. Glucocorticoid pulses were found to rapidly enhance excitatory glutamate transmission in the hippocampus through nongenomic MRs; CRH release from the PVN was found to be suppressed through membrane GR-like receptors. Thus, it is very likely that these nongenomic effects have an important control function in activation and termination of the HPA pulse.

Previously, the slower genomic MR actions in the hippocampus appeared important for maintaining stability and integrity of limbic neuronal circuits. Situations with mostly nuclear MRs, but few GRs, activated are associated with small calcium currents, reduced spike frequency accommodation, stable responses to repeated stimulation of glutamatergic pathways, and relatively small responses to biogenic amines. Nuclear MR activation thus seems to guarantee a stable background of neuronal firing. Activation of co-localized nuclear GRs (and MRs) with higher glucocorticoid concentrations was found to result in enhanced calcium influx, stronger spike frequency accommodation, changes in ion channel conductances, and marked responses to biogenic amines such as serotonin. Thus, nuclear MRs also control, on another time scale, the onset of the stress reactions, and GRs control recovery and adaptation.

How this complementary function of membrane and nuclear MRs and GRs accommodates glucocorticoid pulsatility needs to be investigated. One testable prediction is that membrane MRs control the ascending, and membrane GRs control the descending, arm of the pulse. The genomic MRs promote long-term stability and defense of homeostasis, while genomic GRs facilitate recovery after a stressful challenge and help to prepare for the future. These functions are indicated for MRs as being proactive, determining the future state versus the reactive mode mediated by GRs, referring to the consequence of a stress-induced glucocorticoid pulse.

The MR/GR balance hypothesis predicts that, upon imbalance, threats to homeostasis are less well communicated. At a certain threshold, such hampered communication may lead to a condition of neuroendocrine dysregulation and impaired behavioral adaptation, which potentially can aggravate age-related deterioration and promote susceptibility to age-related disease to which an individual is predisposed. Aging, therefore, is characterized by a changing MR/GR balance and enhanced HPA lability.

Modulation of Brain Corticosteroid Receptors and HPA Axis: Implications for Aging

Interestingly, individual differences in cognitive impairment that appear at old age are subject to manipulation by early life events, such as handling of newborns and depriving newborns of maternal contact. In rats, maternal deprivation or postnatal handling during the first 2 weeks of life can alter stress responsiveness permanently. Aged Long–Evans animals handled postnatally showed an attenuated stress-induced rise in corticosterone levels and are cognitively less impaired than are animals not handled as infants. Handled Long–Evans animals show a higher expression of hippocampal GRs, an effect that is established during the handling procedure by an epigenetic mechanism acting in GR promoter regions.

The outcome of maternal deprivation depends on the age at which separation occurred and duration, and strain and gender are also important variables. Deprivation of maternal contact in Brown Norway rats for 24 h at postnatal day 3 was recently shown to cause a profound midlife surge in stress-induced corticosterone output, but senescent animals and their control, mother-reared littermates displayed comparable attenuated corticosterone responses. It was expected that the early adverse experience would cause cognitive decline in all of the aged rats. This was, however, not observed: rather, maternal deprivation amplified individual differences in old age, and gave rise to a complete shift in the percentages of animals classified as impaired, partially impaired, or nonimpaired, on the basis of their performance in the Morris water maze. Most senescent rats that were mother-reared as pups showed average partially impaired performance, but senescent rats subjected to maternal deprivation were either fully impaired or nonimpaired, at the expense of the average (Figure 2).

Whether such a previous 'midlife' surge in HPA activity is of relevance for the subsequent aging trajectory has been tested in various ways. In the classical Phil Landfield experiments, Fisher rats adrenalectomized at midlife and provided with low amounts of corticosterone, sufficient to occupy MRs, showed at senescence less age-related decline in various behavioral and neuroanatomical markers than did their untreated controls. Likewise, in Jonathan Seckl's laboratory, mutant mice in which the 11-HSD-1 gene was deleted, preventing local reactivation of corticosterone, age-related learning impairments were ameliorated. Alternatively, rats receiving high doses of corticosterone during midlife deteriorated cognitively toward old age. However, Carmen Sandi showed that such

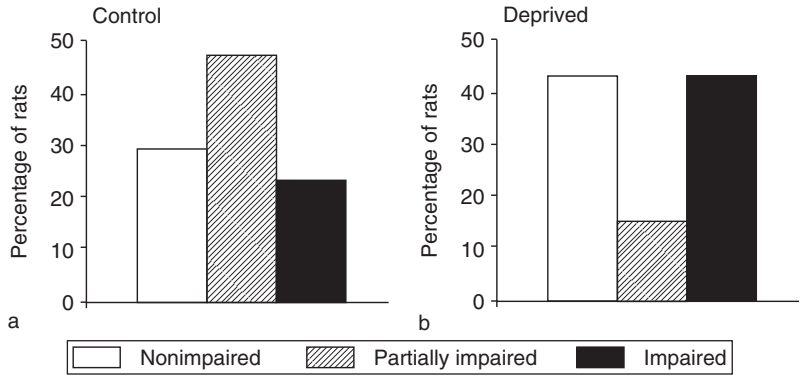

Figure 2 Adverse early life experience amplifies individual differences in cognitive performance. At senescence, the distribution of the quality of spatial performance differs between (a) control ($n = 34$) and (b) maternally deprived ($n = 47$) animals, yielding a bell-shaped vs. a U-shaped distribution, respectively. Spatial learning of maternally deprived rats is impaired, at the expense of average performance (i.e., partially impaired). From de Kloet ER and Oitzl MS (2003) Who cares for a stressed brain? The mother, the kid or both. *Neurobiology of Aging* 24: S61–S65.

corticosterone-induced decline depends on the context in which corticosterone operates. In these experiments, Wistar rats were distinguished at midlife by high or low reaction to novelty exposure and were subsequently exposed for 28 days to stressors. Interestingly, both high and low responders showed enhanced corticosterone reactivity at senescence, but only the high novelty responders deteriorated in cognitive performance at senescence. Which factor determines this context-dependent action of corticosterone is not known.

A consistent finding has been the downregulation of hippocampal MRs in aged animals, irrespective of whether these were rodents, dogs, or primates. Strategies to enhance MR expression indeed were able to enhance cognitive performance. One example is conferred by nature: the successfully aging Brown Norway rat strain has a gain-of-function mutation in the MR. Another example is provided by tricyclic antidepressants, which in the senescent animal had no effect on cognitive performance. If, however, treatment started at midlife and was maintained until senescence, the antidepressant ameliorated cognitive impairment in the aged rats, while the drug induced MRs and downregulated corticosterone secretion. Collectively, these data suggest that the outcome of midlife corticosterone for the aging process depends on genetic and environmental contexts, in a fashion involving the corticosteroid receptors.

Glucocorticoids and Aging: Mechanisms

Age-Related Changes during Hypercorticism: The Relevance of Context

The relevance of context in corticosteroid action is provided by studies using a certified strategy for longevity: caloric restriction (CR). It has been known for a long time that reduction of caloric intake by at least 20% slows down the aging process and extends the life span of rats and other organisms. The aged CR rats have improved cognitive performance in a maze test, long-term potentiation (LTP) is facilitated, and neurogenesis is enhanced. However, this condition of energy deprivation also results in high circulating glucocorticoid levels. How is it possible that in some conditions chronically elevated glucocorticoids cause damage, but in others, such as in CR, elevated glucocorticoid levels exert protection? According to Caleb Finch, this paradox may be explained by the following processes: CR promotes (1) maintenance of microvasculature integrity, (2) attenuation of oxidative damage, (3) enhanced synthesis of growth factors, and (4) attenuation of glial cell proliferation. It is thought that these effects exerted by CR override the negative effects of elevated glucocorticoids. It might be that CR imposes a different context for corticosterone, resulting in suppression of oxidative damage and glial cell activation; the factors determining this context are not known.

Possible Mechanisms Underlying Glucocorticoid-Related Damage

Excessive glucocorticoid exposure is associated with age-related impairment in memory performance, but what is the underlying mechanism? One factor is the overall increase in calcium currents, which has a number of consequences, including suppressed synaptic efficacy, as seen from impaired LTP. Glial cell proliferation and function are affected and dendritic branching in the CA3 pyramidal cell field is suppressed. Other factors related to energy metabolism, cell birth, viability, recovery, and growth also could be implicated in corticosteroid-dependent degenerative changes accompanying the aging process.

Cortisol can suppress hippocampal glucose metabolism, which is known to reflect learning and memory capacity. Positron emission tomography (PET) measurements of cerebral glucose metabolism in elderly persons indeed show in cortical and hippocampal areas a generalized hypometabolism that precedes subsequent hippocampal volume reductions and reduced performance in long-term memory tasks. The beneficial effects of glucose, and also of insulin administration, on aspects of rat and human cognitive performance are of interest, although variable results have been reported.

Brain-derived neurotrophic factor (BDNF) and nerve growth factor (NGF), and their receptors, may be implicated, since age-associated spatial learning impairments in rats can be reversed by administration of some but not all growth factors. Most attention is on the age-related atrophy of cholinergic neurons innervating hippocampus and neocortex. Recently, in rhesus monkeys, learning impairment as well as the associated atrophy of subcortical cholinergic neurons could even be reversed using neurotrophin gene therapy. This suggests that growth factors hold some potential for treating age-related changes.

Another possibility is hippocampal dentate gyrus neurogenesis, which persists from early life but slows down at old age. It is thought that neurogenesis allows the dentate gyrus to continuously rejuvenate its neuronal population in order to adapt to change. Newborn cells become functional after 2–3 weeks as judged from increases in synaptic plasticity and LTP. Stress and elevated corticosterone concentrations suppress the rate of neuronal birth in young, adult, and aged rats. Alternatively, high corticosterone levels – as a result of calorie restriction or increased physical activity through voluntary running, and also estrogen treatment, enriched environmental housing, or training in a hippocampal learning task – increase neurogenesis into old age. Further studies are needed to explore the significance of neurogenesis for aging in these various contexts.

HPA Axis, Aging, and Age-Related Pathology in Humans

Similar to that in rodents, increased HPA activity is not a universal feature of human aging. A flattened circadian rhythm in circulating cortisol is the most consistent finding. In support of enhanced HPA activity, increased numbers of CRH/vasopressin-expressing neurons were found in the PVN of aging males. Stress responsiveness is affected by aging, as a (stressful) public speech task revealed that the subgroup with an early (anticipatory) increase in cortisol in response to the stimulus was impaired in declarative memory, whereas elderly subjects without this early rise in cortisol showed no memory deficits. These data suggest that some of the results from rat studies can be extrapolated to the human situation. However, the design of the experiments and the nature of the stressors are for obvious reasons quite different.

New information also suggests that corticosteroid receptor polymorphisms are associated with aging. In one study from Steven Lamberts' group, the ER22/23EK polymorphism, conferring a relative resistance to the GRs, was found associated with a healthy metabolic profile. The GR variant (frequency 8.9% in adults) had a significantly higher frequency in the older population. The MR Il80V polymorphism discovered by Roel de Rijk shows loss of function, and, as a result, carriers of this gene variant (frequency, 14%) display strongly enhanced neuroendocrine and autonomic responses in the Trier social stress test. Aged individuals carrying this gene variant showed a higher prevalence of depressive symptoms accumulating in the aging population.

In a prospective study, cortisol appeared related to cognitive function. Sonia Lupien observed that only aged individuals with rising cortisol levels over the 4-year period were correlated with reductions in explicit memory and attention tasks. In parallel, a reduction of 14% in hippocampal volume occurred. These volume reductions are already apparent in elderly individuals who display only mild cognitive impairment preceded by glucose hypometabolism. Similar changes in cognitive performance and hippocampal volume have been observed after exogenous glucocorticoids and in Cushing's disease. Significant correlations have, furthermore, been found between hippocampal size and delayed, but not primary or immediate, memory performance, independent of gender, generalized cerebral atrophy, or age. The extent of these early atrophic changes of the hippocampus was found to be a risk factor for accelerated memory dysfunction in normal aging.

Enhanced HPA activity, cognitive decline, and hippocampal volume reduction strongly depend on the genotype of a patient. These measures are reduced in carriers of the apolipoprotein E4 allele, which is a risk factor for age-related cognitive deficits and Alzheimer's disease. In general, basal and stress-induced HPA activation are further enhanced by Alzheimer's disease pathology. This raises the question whether Alzheimer's disease hippocampal damage is the cause or the consequence of excessive glucocorticoids. Since Alzheimer's disease pathology is absent in patients treated with corticosteroids or in severely depressed patients suffering from hypercortisim, it has been argued by Dick Swaab that these hormones are obviously not the trigger. Rather, aberrant HPA activity is considered of relevance in the variable onset and progression of Alzheimer's disease.

Future Directions

A fundamental question in neuroendocrine aging is to what extent the HPA axis contributes to physical and behavioral signs of aging and age-related pathology. This question has received renewed interest because HPA hormones, particularly the glucocorticoids, have been found to be released in pulsatile fashion, to coordinate cell and organ function. Moreover, at the tissue level, the glucocorticoids operate through complementary MR- and GR-mediated actions in the limbic brain. While MRs are involved in stability and the onset of the stress response, recovery and adaptation after stress are mediated by GRs. In balance, MRs and GRs are crucial for homeostasis and health; imbalance and consequent HPA lability are characteristic features of aging.

A major issue is how glucocorticoids should be positioned in the current theories of aging. The thrifty genotype (James Neel) and phenotype (David Barker) hypotheses of aging hinge on nutrition and metabolism. The underlying genes and physiology do not act by themselves, but need to be regulated. Glucocorticoids are essential for this purpose, but how glucocorticoid action can change from protective to damaging is not known. Also, Caleb Finch's 'gero-inflammatory manifold,' favoring systemic and local inflammations as common denominators in age-related degenerations, has an obvious link to glucocorticoids, but glucocorticoid's place in the sequence of events is virtually unknown. Whether susceptibility genes and adverse life experiences aggravate age-related processes unidirectionally in every individual, or act on a common unknown substrate, amplifying individual differences *per se*, is an unresolved issue as well.

Thus, the issue on aberrant HPA axis and glucocorticoids, as either the cause or the consequence of aging processes, is unresolved. Yet, the significance of HPA pulsatility and MR/GR responsiveness provides ample opportunity for aging research. The current evidence suggests that MR/GR imbalance compromises the ability of organisms to defend and to restore homeostasis. Such imbalance may result in HPA dysregulation enhancing vulnerability through the very same genes that are otherwise essential for maintaining homeostasis, resilence, and health. The identity of these genes is not known.

See also: Aging of the Brain; Chronic (Repeated) Stress: Consequences, Adaptations; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Neuroendocrine Aging: Pituitary Metabolism; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Pituitary Gland (Cell Types, Mediators, Development); Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Buckley TM and Schatzberg AF (2005) Aging and the role of the HPA axis and rhythm in sleep and memory consolidation. *American Journal of Geriatric Psychiatry* 13: 344–352.

de Kloet ER, de Rijik RH, and Meijer OC (2007) Is there an imbalanced response of mineralocorticoid and glucocorticoid receptors in depression? *Nature Clinical Practice Endocrinology & Metabolism* 3: 168–179.

de Kloet ER, Joëls M, and Holsboer F (2005) Stress and the brain: from adaptation to disease. *Nature Reviews Neuroscience* 6: 463–475.

de Kloet ER and Oitzlr MS (2003) Who cares for a stressed brain? The mother, the kid or both. *Neurobiology of Aging* 24: S61–S65.

de Kloet ER, Vreugdenhil E, Oitzl MS, et al. (1998) Brain corticosteroid receptor balance in health and disease. *Endocrine Reviews* 19: 269–301.

de Nicola AF, Saravia FE, Beauquis J, et al. (2006) Estrogens and neuroendocrine hypothalamic–pituitary–adrenal axis function. *Frontiers of Hormone Research* 35: 157–168.

Finch CE and Crimmins EM (2004) Inflammatory exposure and historical changes in human life-spans. *Science* 305: 1730–1739.

Finch CE and Stanford CB (2004) Meat adaptive genes and the evolution of slower aging in humans. *Quarterly Review of Biology* 79: 3–50.

Karst H, Berger S, Tronche F, et al. (2005) Mineralocorticoid receptors are indispensable for nongenomic modulation of hippocampal glutamate transmission by corticosterone. *Proceedings of the National Academy of Sciences of the United States of America* 102: 19204–19207.

Landfield PW (1978) Hippocampal aging and adrenocorticoids: Quantitative correlations. *Science* 202: 1098–1102.

Lucassen PJ and De Kloet ER (2001) Glucocorticoids and the aging brain. In: Hof PR and Mobbs CV (eds.) *Functional Neurobiology of Aging*, pp. 883–903. San Diego: Academic Press.

McEwen BS (2003) Interacting mediators of allostasis and allostatic load: Towards an understanding of resilience in aging. *Metabolism* 52: 10–16.

Meaney MJ and Szyf M (2005) Maternal care as a model for experience-dependent chromatin plasticity? *Trends in Neurosciences* 28: 456–463.

Oitzl MS, Workel JO, Fluttert M, et al. (2000) Maternal deprivation affects behaviour from youth to senescence: Amplification of individual differences in spatial learning and memory in senescent Brown Norway rats. *European Journal of Neuroscience* 12: 3771–3780.

Patel NV and Finch CE (2002) The glucocorticoid paradox of caloric restriction in slowing brain aging. *Neurobiology of Aging* 23: 707–717.

Sapolsky RM (1992) *Stress, the Aging Brain and the Mechanism of Neuron Death*. Cambridge, MA: MIT Press.

Young EA, Abelson J, and Lightman SL (2004) Cortisol pulsatility and its role in stress regulation and health. *Frontiers in Neuroendocrinology* 25: 69–76.

Relevant Website

http://www.lifespannetwork.nl – LifeSpan Network (a consortium of European laboratories studying mechanisms of aging and development).

Neuroendocrine Aging: Pituitary–Gonadal Axis in Males

P Y Liu, P Takahashi, and A Nehra, Mayo Clinic, Rochester, MN, USA
S M Pincus, Guilford, CT, USA
D M Keenan, University of Virginia, Charlottesville, VA, USA
J D Veldhuis, Mayo Clinic, Rochester, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

The human hypothalamic–pituitary–testicular axis maintains testosterone production and spermatogenesis through an intricate feedforward and feedback system involving the intermittent exchange of neurohormonal signals. The chief components of this ensemble are hypothalamic gonadotropin-releasing hormone (GnRH), pituitary luteinizing hormone (LH), and testicular sex steroids (testosterone and estradiol). Epidemiological data and clinical experiments jointly indicate that healthy older men have lower bioavailable and free testosterone concentrations, higher sex hormone-binding globulin (SHBG) concentrations, and decreased daily testosterone secretion rates, compared with young individuals. Mechanistic analyses predict each of impaired GnRH outflow, blunted LH-stimulated testosterone secretion, and reduced negative feedback by testosterone in the older male. Aging also disrupts quantifiable synchrony among sleep stage, the secretion of LH, testosterone, prolactin, and follicle-stimulating hormone (FSH), and the oscillations in nocturnal penile tumescence, denoting erosion of coordinate central neurohormonal outflow. Prolactin, estradiol, and inhibin B concentrations fall and FSH concentrations rise with age. How age affects spermatogenesis in healthy populations is not known. In summary, aging in men is marked by multisite regulatory failure within the gonadal axis. Further studies are required to define the tempo, variability, reversibility, and import of such interlinked deficits, and to delineate safe and effective interventional strategies to obviate aging-related frailty.

The human hypothalamic–pituitary–testicular axis maintains a eugonadal state via reciprocal feedforward and feedback connections. Multiple changes in this network occur with age, resulting in a gradual (1–2% yr^{-1}) decline in testosterone availability. Whether aging also decreases spermatogenesis is not established due to the inherent difficulty in obtaining semen from representative populations. Less testosterone production may contribute to aging-associated increases in visceral fat, insulin resistance, falls, and fractures, and decreases in bone mineral density, muscle mass and strength, physical performance, perceived quality of life, libido, and potentia. Mechanistic inquiries dissecting the bases for testosterone depletion with age should facilitate the development and optimization of preventive and therapeutic strategies. Such strategies must carefully balance potential risks of androgen-induced prostatic disease, cardiovascular disease, erythrocytosis, and/or obstructive sleep apnea against psychological and physical benefits.

Rate of Evolution and Extent of Relative Hypoandrogenemia in Older Men

Testosterone depletion in aging is an increasingly pertinent clinical issue, given epidemiological data linking hypoandrogenemia with sarcopenia, osteopenia, visceral adiposity, insulin resistance, reduced physical stamina, sexual dysfunction, impaired quality of life, and (possibly) depressive mood and waning cognitive function. Impoverished testosterone availability in older men has been affirmed by (1) longitudinal investigations in healthy cohorts, (2) meta-analysis of cross-sectional data, and (3) direct sampling of the spermatic vein. In general, bioavailable (non-SHBG-bound) and free testosterone concentrations fall more markedly than do total values, as illustrated for cross-sectional data in Figure 1. Total testosterone concentrations are partially maintained by a concomitant increase in SHBG concentrations in aging men. A 15-year prospective analysis in the United States (New Mexico) reported that total testosterone concentrations fall by approximately 0.39 nmol l^{-1} yr^{-1} in men after the age of 60 years; the Massachusetts Male Aging Cohort studies estimated a 0.8–1.3% annual decrement in bioavailable testosterone concentrations and the Baltimore Longitudinal Study of Aging predicted an annual decline of 4.9 pmol testosterone per nanomole SHBG. In the last study, the age-related prevalence of hypogonadal testosterone:SHBG ratios, judged by young-adult normative criteria, exceeded 20%, 30%, and 50% at ages 60, 70, and 80 years, respectively. Thus, reduced systemic testosterone availability is especially significant in societies with high longevity.

Comorbidity in aged individuals can markedly exacerbate androgen deficiency. Comorbid risk factors include chronic systemic disease, acute illness, stress, weight loss, hospitalization, institutionalization, and the use of certain drugs. The degree to which illness-induced androgen decrements are reversible in various settings is not yet well articulated. Whether certain comorbidities represent indications for testosterone replacement also requires further investigation.

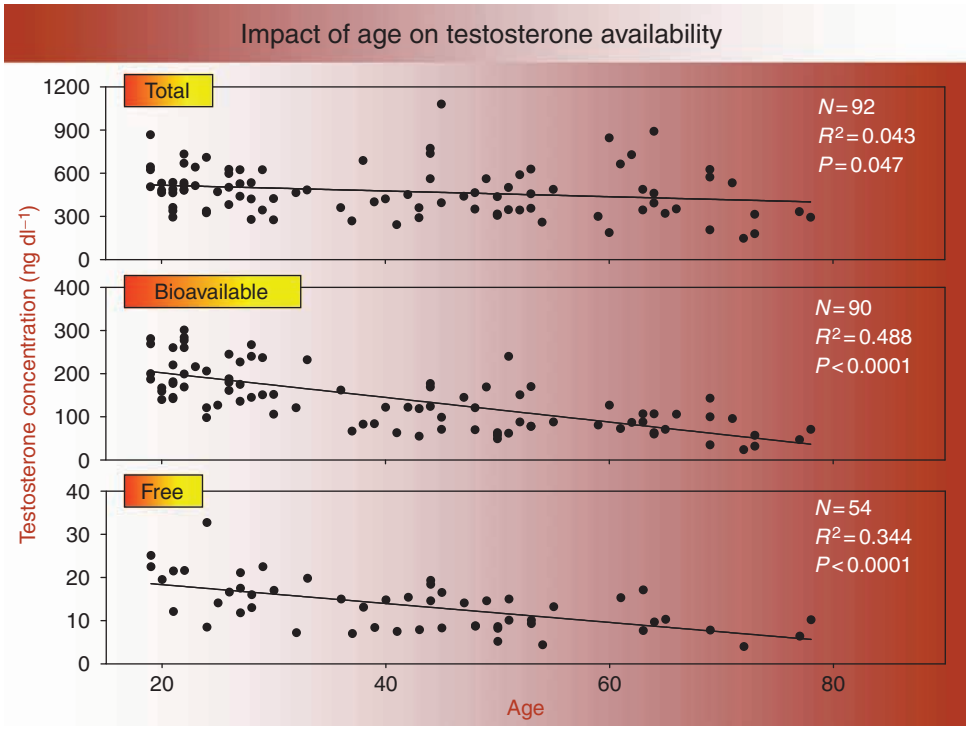

Figure 1 Plots of total, bioavailable, and free (unbound) testosterone concentrations regressed on age in healthy volunteers from Olmsted County, Minnesota, USA. The R^2 values denote the fraction of variability accounted for by age.

Mechanistic Bases of Testosterone Depletion

The fundamental mechanisms that cause hypoandrogenemia in the aging male remain unknown. However, recent clinical investigations have disclosed multiple alterations in the gonadal axis in older men: (1) lower amplitude LH pulses, suggesting reduced drive by GnRH; (2) more frequent LH pulses and less orderly patterns of LH release, consistent with diminished negative feedback; (3) preserved LH and heightened FSH responses to exogenous GnRH pulses, indicating lack of primary pituitary failure; (4) reduced pulsatile and thereby total daily testosterone secretion; and (5) impaired testosterone responses to elevated endogenous LH concentrations (stimulated by flutamide, tamoxifen, GnRH, or anastrozole) and to infused pulses of recombinant human LH. The collective findings indicate that relative androgen deficiency in the older male reflects multisite failure of the GnRH–LH–testosterone axis.

An investigative challenge arises in verifying specific mechanisms that mediate gradual aging-related adaptations in the human gonadal axis. A recent strategy is to exploit an ensemble regulatory concept of reciprocal interactions among GnRH, LH, and testosterone (Figure 2). An integrative biomathematical approach allows one to formalize and quantify underlying feedback and feedforward interactions objectively. The analytical rationale is to reconstruct aging-related adaptations among all three primary signals – GnRH, LH and testosterone – rather than evaluate any one signal in isolation. According to ensemble analyses, a parsimonious explication of collective clinical data is that aging (1) attenuates hypothalamic GnRH outflow, (2) impairs testicular responsiveness to LH pulses, and (3) decreases androgenic negative feedback. Longitudinal studies will be required to establish the relative importance of individual regulatory deficits in healthy aging (Figure 3).

Putative Pathophysiological Mechanisms in the Aging Male

Hypothalamic GnRH Deficiency in Older Men

Indirect clinical observations support the hypothesis that hypothalamic GnRH outflow is reduced in older men (Table 1). In particular, aging individuals maintain lower LH pulse amplitudes (due to a reduced mass of LH secreted per burst) and more irregular LH secretory patterns (increased approximate entropy), yet respond normally to exogenous GnRH pulses. Although such observations are congruent with decreased endogenous GnRH drive, hypothalamic GnRH secretion cannot be monitored directly

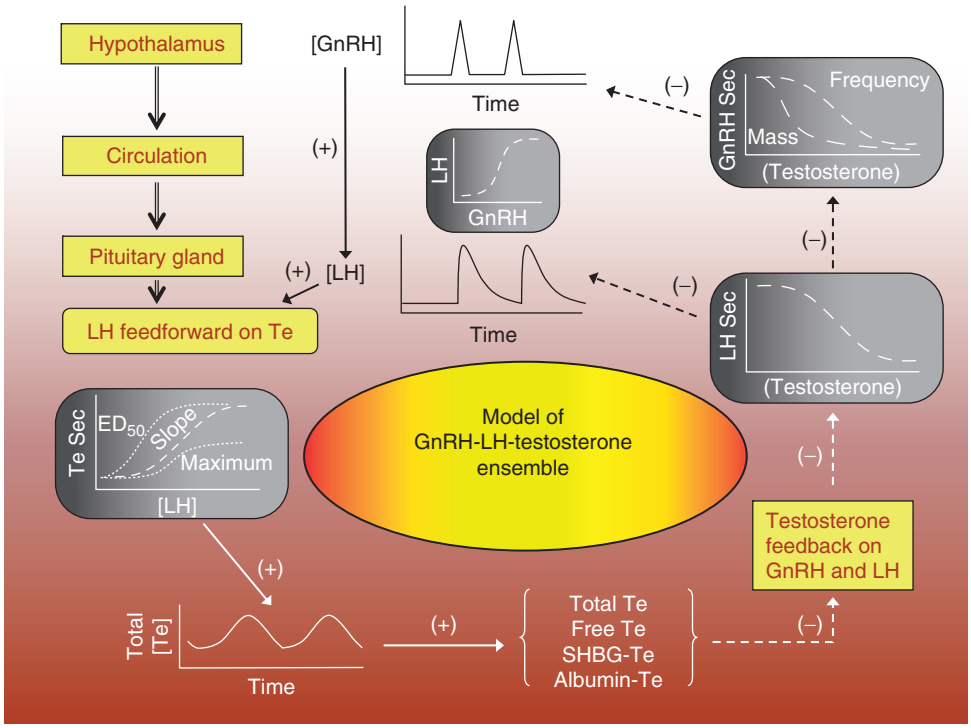

Figure 2 Simplified structure of the male gonadotropin-releasing hormone (GnRH), luteinizing hormone (LH), and testosterone (Te) (i.e., hypothalamic–pituitary–Leydig cell) axis with key feedback (–, inhibitory) and feedforward (+, stimulatory) interactions. Smooth curves denote *in vivo* dose–response interfaces mediating GnRH-induced LH secretion (top middle), LH-stimulated testosterone secretion (Te Sec; bottom left; SHBG, sex hormone-binding globulin), and testosterone-enforced negative feedback on the pituitary gland (right middle) and hypothalamus (rightmost top). Possible adaptations in LH → Te feedforward sensitivity (slope), potency (one-half-maximal stimulus concentration, or ED_{50}), and efficacy (maximal responsiveness) are illustrated (left middle).

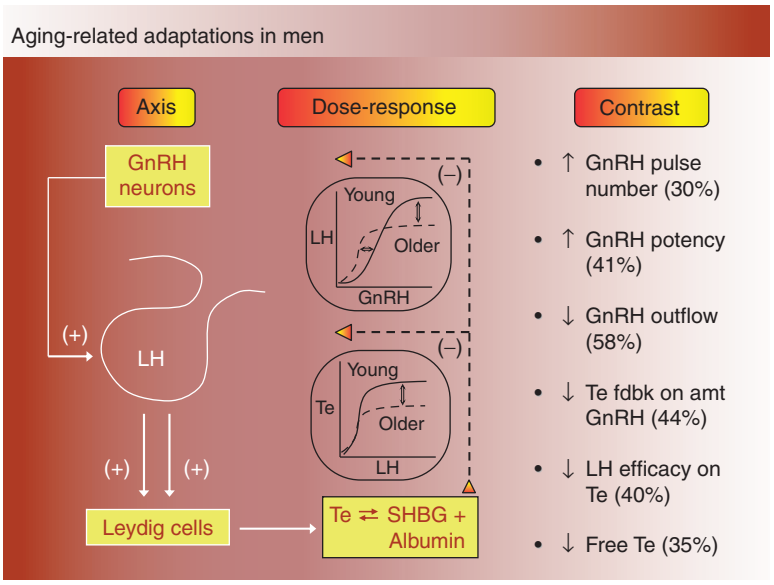

Figure 3 Schema of inferred anatomic sites (axis; left) and adaptive mechanisms (contrast; right) that subserve relative testosterone (Te) deficiency in healthy older men (age >60 years), compared with young (age <35 years) men (GnRH, gonadotropin-releasing hormone; LH, luteinizing hormone; SBHG, sex hormone-binding globulin). Arrows define stimulation (+) and inhibition (–). The two encircled insets (dose–response; center) depict age-related contrasts in GnRH → LH drive (top center) and LH → Te feedforward (bottom center).

Table 1 Indirect evidence for impaired gonadotropin-releasing hormone drive in aging men

Low-amplitude spontaneous luteinizing hormone pulses
Increased or normal potency of exogenous gonadotropin-releasing hormone pulses
Blunted unleashing of pulsatile luteinizing hormone secretion after blockade of sex hormone feedback or opiate receptors
Normal luteinizing hormone bioactivity
Quantifiably more disorderly patterns of luteinizing hormone release
Asynchrony of central neurohormone outflow: luteinizing hormone, follicle-stimulating hormone, prolactin, sleep stage, and nocturnal penile tumescence
Mathematical predictions based on an ensemble concept

in the human. One means to estimate central GnRH outflow noninvasively entails graded competitive inhibition of both endogenously driven and exogenous GnRH-induced LH secretion. Inhibition can be enforced by escalating doses of a potent and selective GnRH-receptor antagonist (such as ganirelix). In this paradigm, age does not affect LH responses to an exogenous GnRH stimulus, whether or not ganirelix is administered. In contrast, increasing age potentiates suppression of endogenously driven LH pulses by any given blood concentration of the receptor antagonist. Given that GnRH and ganirelix compete for the same pituitary receptor, greater inhibition of LH pulses by the receptor antagonist in older men points to less hypothalamic GnRH outflow to drive gonadotrope cells. This important inference is confirmed analytically using a model in which one assumes that (1) unobserved hypothalamic GnRH concentrations stimulate LH secretion via a positive logistic function and (2) testosterone inhibits GnRH outflow (release and action) via a negative exponential function.

Laboratory experiments have also delineated significant (but not exclusive) central hypogonadotropism associated with hypoandrogenemia in the aged male rodent. Relevant evidence includes (1) reduced *in vivo* LH pulse amplitude, (2) diminished *in vitro* GnRH secretion by hypothalamic tissue, (3) attenuated castration-induced LH release, (4) normal LH responses to injected GnRH, (5) altered GnRH neuronal synaptology, and (6) restoration of sexual activity in the aged rat following hypothalamic neuronal transplantation. Nonetheless, potential species differences limit facile extrapolation of these data to humans.

Enhanced Gonadotrope Secretory Responsiveness to Small Amounts of GnRH

In principle, smaller LH pulses in older men could reflect diminished hypothalamic GnRH secretion and/or decreased pituitary sensitivity to available GnRH. In the latter context, initial studies using a single supramaximal dose of GnRH were contradictory. Recent investigations incorporating physiological doses of GnRH disclosed that aging potentiates submaximal stimulation (Figure 3). One dose–response study that injected single intravenous pulses of GnRH in random order on a single morning estimated 1.6-fold higher GnRH-stimulated LH, FSH, and free α-subunit secretion in aging men. Another dose–response analysis administered individual intravenous doses of GnRH spanning a 1000-fold range on separate, randomly ordered mornings. This stratagem disclosed that older men maintain nearly twofold greater gonadotrope sensitivity to GnRH (steeper slope of the dose–response) and twofold greater GnRH potency (50% lower dose of GnRH required to stimulate LH secretion one half-maximally) than do young counterparts, but have comparable maximal LH secretion (denoting age-equivalent GnRH efficacy). Heightened acute gonadotrope responsiveness to submaximal amounts of GnRH (2.5, 10, and 25 ng kg^{-1}) in older men could reflect upregulation of GnRH receptors, augmentation of pituitary LH stores, and/or feedback withdrawal due to lower concentrations of total, bioavailable, or free testosterone. The last consideration has been affirmed analytically. A third investigation infused pulses of GnRH (100 ng kg^{-1}) or saline intravenously every 90 min for 14 days by portable infusion pump in young and older men. The GnRH clamp increased pulsatile LH secretion equivalently in the two age groups, indicating comparable gonadotrope responsivity. However, bioavailable and free testosterone concentrations remained 50% lower in elderly men, suggesting reduced bioactivity of secreted LH or impaired testicular steroidogenesis.

Impaired Leydig Cell Steroidogenesis in Older Men

Administration of the LH surrogate, human chorionic gonadotropin (hCG), fails to stimulate maximal young-adult testosterone concentrations in many, but not all, aging men. Impaired testis responsivity in this setting is difficult to interpret definitively, because the hCG stimulus is not physiological. Three concerns exist: (1) the half-life of hCG is 20–30 h (compared with 1 h for LH), thus enforcing nonpulsatile drive, (2) hCG rapidly downregulates Leydig cell steroidogenesis *in vivo* and *in vitro*, and (3) conventional hCG doses exceed the lutropic activity of an LH pulse by >300-fold, thus testing only maximal testis responses. The last issue is relevant, because low testosterone production occurs in older men when LH concentrations are normal or only minimally elevated.

A near-physiological protocol to examine Leydig cell responsiveness involves intravenous infusion of pulses of recombinant human LH during acute suppression of

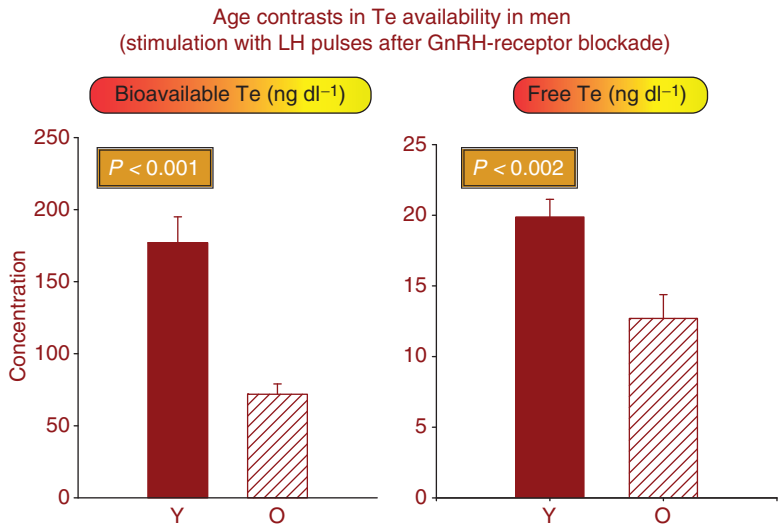

Figure 4 Pulses of recombinant human luteinizing hormone (LH) after gonadotropin-releasing hormone (GnRH) receptor blockade fail to normalize bioavailable (left) and free (right) testosterone (Te) concentrations in older men. Adapted from Liu PY, Takahashi PY, Roebuck PD, et al. (2005) Aging in healthy men impairs recombinant human luteinizing hormone (LH)-stimulated testosterone secretion monitored under a two-day intravenous pulsatile LH clamp. *Journal of Clinical Endocrinology & Metabolism* 90(10): 5544–5550. Copyright 2005, The Endocrine Society.

LH secretion with a GnRH-receptor antagonist. Application of this paradigm for 48 h unmasked 50% lower exogenous LH-stimulated bioavailable and free testosterone concentrations in older, as compared to young, men (Figure 4). Whether more extended delivery of normal-amplitude LH pulses would restore young-adultlike testosterone secretion in older individuals is not known. Although experiments in the aged Brown Norway rat suggest a fixed steroidogenic defect in Leydig cells, the exact relevance of the animal model to the human is not certain.

Aging Restricts Feedback by Endogenous Testosterone on GnRH and LH Secretion

Clinical studies have reported both greater and less reduction of LH concentrations following supplementation with testosterone or 5α-dihydrotestosterone in older men, as compared to young men. Discrepancies could relate to nonphysiological routes, amounts, preparations and patterns of androgen administration, confounding effects of endogenous testosterone, dissimilar study populations, and/or single measurements of LH concentrations. On the other hand, aging appears to mute negative feedback by endogenous testosterone consistently. The latter inference applies to comparisons made by cross-correlation, approximate entropy, and ensemble-model analyses of frequently sampled LH and testosterone concentrations. Feedback attenuation is not due solely to systemic androgen depletion in older individuals, because feedback strength decreases per unit testosterone concentration (Figure 5). The molecular mechanisms that mediate impaired feedback are not known. However, histochemical analyses have quantitated reduced androgen-receptor expression in the brain and pituitary gland of the aged rat and in genital fibroblasts of the older human.

Reduced negative feedback by androgens could contribute to the higher frequency, lower amplitude, and more irregular LH secretion patterns that typify aging individuals. In fact, blocking testosterone or estradiol synthesis with steroidogenic enzyme inhibitors elicits rapid and irregular LH pulses even in healthy young men. The sex steroid-depleting regimens fail to increase LH secretory burst size equally in older and young men. Therefore, smaller LH secretory bursts in aging individuals are not attributable to hypoandrogenemia or hypoestrogenemia *per se*. On the other hand, more frequent LH secretory bursts and less orderly LH secretion patterns may reflect both reduced feedback efficacy and reduced testosterone availability for feedback. Verification of relative age- and androgen-dependent altered LH secretion would require experimentally clamping systemic testosterone availability in different age groups.

FSH and Inhibin B

Hypothalamic–pituitary–gonadal control of FSH secretion is less well understood because (1) feedforward relationships among FSH, testosterone, estradiol, and the 90-day spermatogenic cycle in the human are complex; (2) sperm production and

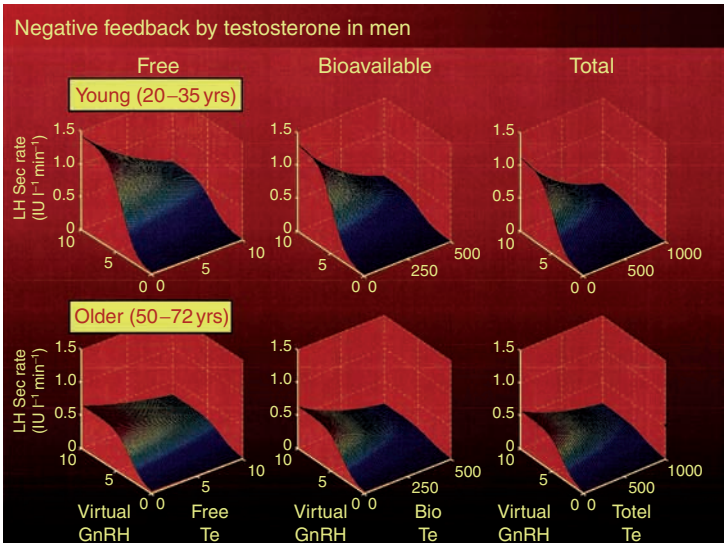

Figure 5 Analytically reconstructed surfaces depicting virtual (calculated) gonadotropin-releasing hormone (GnRH) feedforward and free testosterone (Te) feedback onto luteinizing hormone (LH) secretion in men in the indicated age cohorts. Each surface interlinks pulsatile LH secretion (vertical axis) with estimated GnRH drive (oblique axis) and Te feedback (horizontal axis). Te feedback is calculated for each of free (left), bioavailable (middle), and total (right) Te concentrations. Aging is marked in three ways: by decreased maximal LH secretion, defined at maximal virtual GnRH outflow at zero (extrapolated) Te concentrations (vertical axis intercept of the surface); by less Te feedback on LH secretion at any given GnRH signal strength (right-to-left downslope at top of surface due to increasing free Te concentrations); and by greater potency of virtual GnRH (leftward shift in GnRH → LH dose–response function at any given Te concentration). Adapted from Keenan DM, Takahashi PY, Liu PY, et al. (2006) An ensemble model of the male gonadal axis: Illustrative application in aging men. *Endocrinology* 147(6): 2817–2828, with permission. Copyright 2006, The Endocrine Society.

quality are difficult to monitor serially; (3) FSH secretion is restrained by all three of testosterone, estradiol, and inhibin B; (4) both GnRH and activin appear to maintain FSH secretion in animals; (5) recombinant human activin and inhibin B are not available for clinical studies; and (6) the prolonged half-life of FSH makes reliable pulsatility analysis more difficult (Figure 6). Available clinical studies show that greater age is associated with higher serum FSH concentrations but lower inhibin B (especially before age 35 years) and bioavailable sex-steroid concentrations, and loss of coordinate LH–FSH secretion. Although estradiol, testosterone, and nonaromatizable androgens inhibit FSH secretion in older men, how age influences sex-steroid feedback potency and efficacy remains unknown.

Prolactin

Disorders associated with hyperprolactinemia are associated with suppressed LH pulsatility and reduced sex-steroid production. How hyperprolactinemia represses GnRH secretion is not established in humans. Data derived from animal models are species dependent. For example, prolactin administration suppresses LH

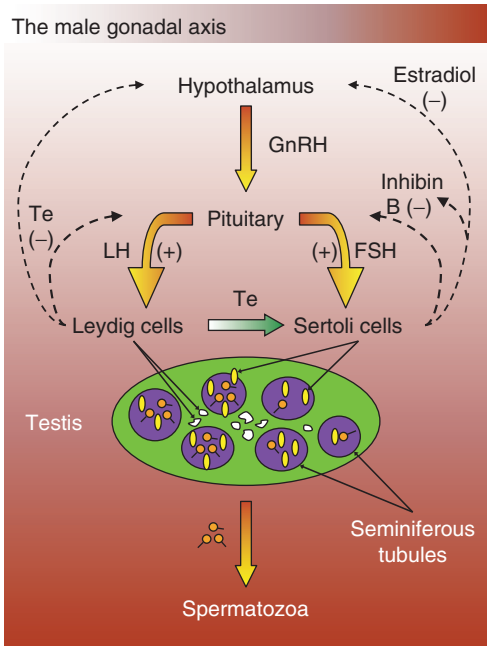

Figure 6 Overall gonadal axis, including the gonadotropin-releasing hormone (GnRH)–luteinizing hormone (LH)–testosterone (Te) component (left) and the seminiferous tubular compartment with GnRH, follicle-stimulating hormone (FSH), estradiol, and inhibin B (right).

pulse frequency and amplitude in sheep and castrated rats. On the other hand, a prolactin receptor-knockout mouse strain had normal concentrations of LH, FSH, and testosterone, with preserved fertility and testis morphology. Lactotrope cell number does not change, but pulsatile prolactin secretion declines with age in humans. Prolactin is not known to regulate testis function directly in men, unlike in the rodent.

Neurohormonal Network Regulation in Aging Men

Available analyses indicate that aging in men impairs hypothalamic GnRH outflow, LH-stimulated Leydig cell steroidogenesis, and testosterone-dependent negative feedback (Figure 3). How this tripartite conclusion is reached requires explanation. Strong interdependencies among GnRH, LH, and testosterone secretion signify the operation of networklike (ensemble) control of the gonadal axis. In ensemble systems, no one signal acts alone or can be viewed alone. Thus, to ensure valid mechanistic interpretation, all three signals must be analyzed together without perturbing their feedback or feedforward connections. The approximate entropy (ApEn) measure provides one means to do this. ApEn is a model-free statistic that quantifies the overall impact of regulatory inputs within an interlinked system. The endpoint is the relative regularity, orderliness, or consistency of recurring subpatterns in serial measurements. Thus, ApEn does not replace, but instead complements, the analysis of pulses or circadian rhythms. ApEn analyses unveil more irregular (less orderly) patterns of both LH and testosterone secretion with increasing age. In contrast, FSH release is quite irregular at all ages.

As an extension of the ApEn metric, the cross-ApEn statistic quantifies the joint pattern synchrony of two coupled processes, specifically paired signals. Unlike the linear cross-correlation coefficient, cross-ApEn provides a probabilistic measure of pattern similarities (reproducibility) independently of time lag. Cross-ApEn analyses unmask prominent age-related deterioration of LH's coupling with each of testosterone, FSH, and prolactin secretion, sleep-stage transitions, and nocturnal penile tumescence (NPT) oscillations. Directional cross-ApEn comparisons indicate that older individuals have symmetric erosion of LH → testosterone feedforward and testosterone → LH feedback linkages (Figure 7). Thus, aging alters ensemble control among LH, FSH, prolactin and testosterone secretion as well as sleep and autonomic nervous system outflow, which govern NPT.

Investigation of the male gonadal axis is enhanced by biomathematical formalism, which encapsulates nonlinear, dose-responsive, and time-delayed feedforward and feedback coupling among GnRH, LH, and testosterone. One analytical construct allows one to estimate endogenous dose–response properties noninvasively

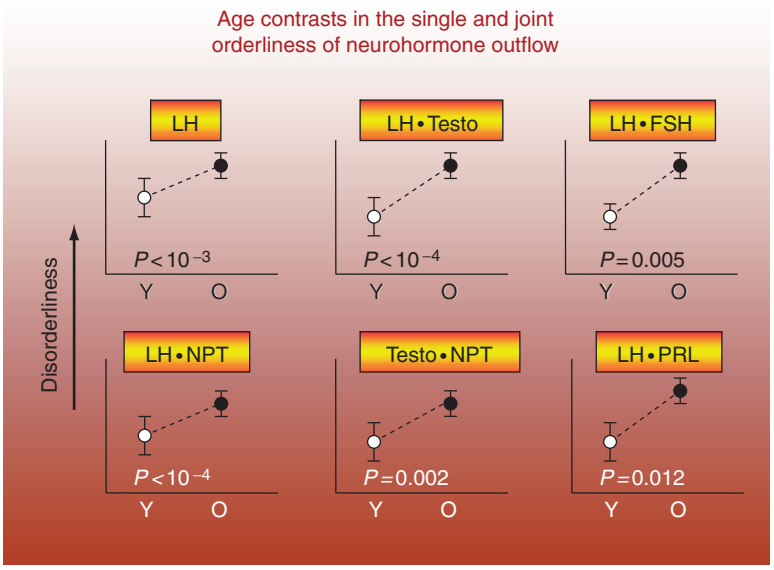

Figure 7 Disruption of central neurohormone outflow in the aging male quantified by the cross-approximate entropy (ApEn) statistic, which quantifies loss of orderliness. Cross-ApEn values for each pair of neurohormone outflow measures are elevated in older men, compared with those in young men, signifying deterioration of joint synchrony (LH, luteinizing hormone; Testo, testosterone; NPT, nocturnal penile tumescence; FSH, follicle-stimulating hormone; PRL, prolactin). Greater asynchrony of paired signals in turn denotes disruption of the physiological pathways connecting the signals. Adapted from Veldhuis JD, Iranmanesh A, Godschalk M, et al. (2000) Older men manifest multifold synchrony disruption of reproductive neurohormone outflow. *Journal of Clinical Endocrinology & Metabolism* 85(4): 1477–1486, with permission. Copyright 2000, The Endocrine Society.

without the injection of GnRH, LH, or testosterone or disruption of feedback linkages. Experimental validation required repetitive direct sampling of cavernous sinus and internal jugular venous concentrations of GnRH, LH, and testosterone in the awake stallion and ram, and peripheral LH and spermatic vein testosterone concentrations in the human. Figure 8 illustrates the results of noninvasive analytical reconstruction of feedforward dose–response functions linking LH concentrations and testosterone secretion rates, using only paired systemic measurements of the two hormones in 15 older and 15 young men. The methodology is illustrated schematically in Figure 9. This approach predicts that aging selectively impairs

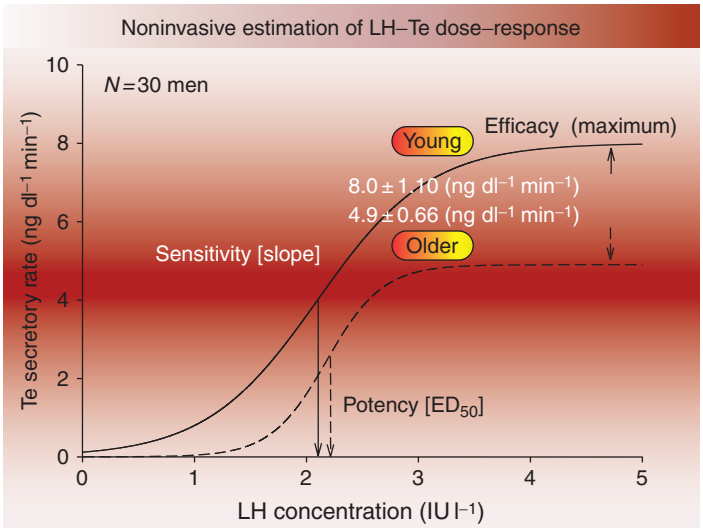

Figure 8 Analytically estimated luteinizing hormone (LH) → testosterone (Te) feedforward dose–response functions in young ($N = 15$) and older ($N = 15$) men. LH efficacy is defined as asymptotically maximal Te secretion; potency is defined as [ED_{50}], the half-maximally effective stimulatory concentration (dose) of LH, and sensitivity is defined as the maximal positive slope. Only LH efficacy differed by age in these analyses. Adapted from Keenan DM and Veldhuis JD (2004) Divergent gonadotropin–gonadal dose–responsive coupling in healthy young and aging men. *American Journal of Physiology* 286(2): R381–R389.

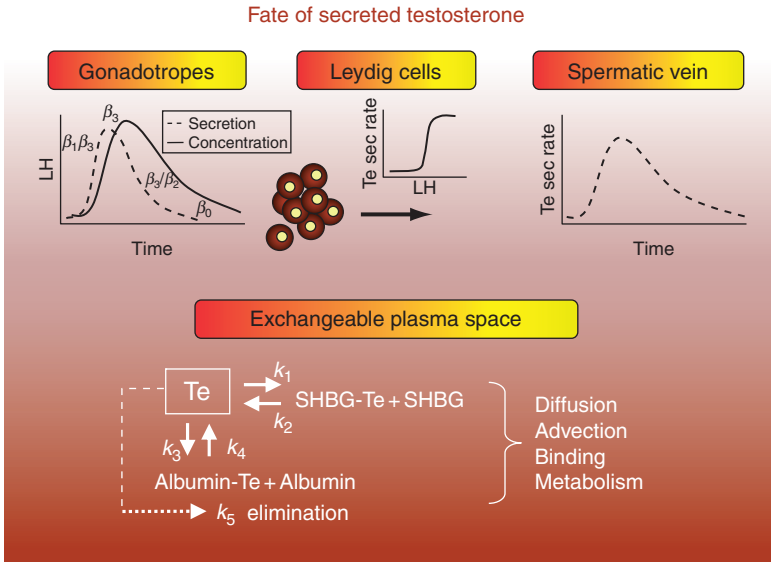

Figure 9 Model of feedforward mechanism by brain gonadotropin-releasing hormone onto pituitary gonadotrope cells to induce a burst of luteinizing hormone (LH) secretion (top left). LH concentrations drive gonadal testosterone secretion (Te Sec) via a monotonic dose–response function (top middle). Burstlike Te secretion over time (top right) is the output of the simultaneously calculated LH → Te feedforward function. Testosterone entering the blood has the five possible fates, denoted as k_1–k_5, representing random diffusion, linear advection in blood, binding to albumin or sex hormone-binding globulin (SHBG), and irreversible removal or chemical metabolism.

Table 2 Model-simulated implications of selected regulatory deficits in (aging) male gonadal axis

Deficit examined[a]	*Predicted responses*			*Orderliness or regularity*
	Available testosterone	*Luteinizing hormone pulses*		
		Amplitude	*Frequency*	
(A) ↓ GnRH outflow	↓	↓		↑
(B) ↓ LH → Te drive	↓	↑	↑	↓
(C) ↓ Te feedback action	↑	↑	↑	↓
Combined A, B, and C[a]	↓	↓	↑	↓

[a]Mimics observed features in older men.
GnRH, gonadotropin-releasing hormone; LH, luteinizing hormone; Te, testosterone.

endogenous LH efficacy; that is, aging reduces maximally stimulated testosterone secretion without altering LH potency or testis sensitivity to LH. Randomly ordered dose-varying intravenous infusions of recombinant LH during GnRH-receptor blockade will be needed to establish such inferences empirically.

Computer-assisted simulations confer predictive insights. Recent model-based simulations of gonadal axis adaptations in the older male illustrate that (1) isolated attenuation of central GnRH drive forecasts lower LH pulse amplitude and attendant hypoandrogenemia, but not higher LH pulse frequency or greater irregularity; (2) exclusive impairment of Leydig cell steroidogenesis presages lower testosterone concentrations, higher LH pulse frequency, and greater disorderliness of LH secretion patterns, but not smaller LH pulses; (3) pure restriction of negative feedback projects normal or elevated testosterone concentrations with elevated LH pulse frequency, amplitude, and irregularity; and (4) combined reductions in GnRH outflow, LH-stimulated testosterone secretion, and negative feedback by testosterone predict all four of hypoandrogenemia, attenuated LH pulse amplitude, accelerated LH pulse frequency, and disorderly LH release. This tetrad is characteristic of older men. Table 2 summarizes these concepts.Accordingly, from an objective modeling perspective, no single regulatory defect can explicate the neuroendocrine phenotype observed in healthy older men.

Summary

Androgen deprivation adversely impacts libido and sexual potency, psychological well being, mood and (possibly) cognition, exercise tolerance, muscle mass, bone mineral density, intraabdominal adiposity, and lipid and carbohydrate metabolism. Clinical investigations in older men support triple impairment of hypothalamic GnRH secretion, pulsatile LH-stimulated testosterone synthesis, and testosterone-mediated negative feedback. Understanding the primary mechanistic bases of testosterone depletion in aging men should foster novel therapeutic strategies to maintain anabolism, enhance quality of life, and minimize physical frailty.

See also: Gene Expression in Normal Aging Brain; Gene Expression Regulation: Steroid Hormone Effects; Gonadotropin, Neural and Hormonal Control; Gonadotropin-Releasing Hormone: GnRH-1 System; Hypothalamo–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH; Neuroendocrine Aging: Pituitary Metabolism; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Neuroendocrine Aging: Pituitary-Adrenal Axis; Pituitary Gland (Cell Types, Mediators, Development); Sleep and Circadian Rhythm Disorders in Human Aging and Dementia.

Further Reading

Feldman HA, Longcope C, Derby CA, et al. (2002) Age trends in the level of serum testosterone and other hormones in middle-aged men: Longitudinal results from the Massachusetts male aging study. *Journal of Clinical Endocrinology & Metabolism* 87: 589–598.

Foresta C, Bordon P, Rossato M, et al. (1997) Specific linkages among luteinizing hormone, follicle stimulating hormone, and testosterone release in the peripheral blood and human spermatic vein: Evidence for both positive (feed-forward) and negative (feedback) within-axis regulation. *Journal of Clinical Endocrinology & Metabolism* 82: 3040–3046.

Harman SM, Metter EJ, Tobin JD, et al. (2005) Longitudinal effects of aging on serum total and free testosterone levels in healthy men. Baltimore Longitudinal Study of Aging. *Journal of Clinical Endocrinology & Metabolism* 86: 724–731.

Keenan DM, Alexander S, Irvine CH, et al. (2004) Reconstruction of *in vivo* time-evolving neuroendocrine dose-response properties unveils admixed deterministic and stochastic elements. *Proceedings of the National Academy of Sciences of the United States of America* 101: 6740–6745.

Keenan DM, Takahashi PY, Liu PY, et al. (2006) An ensemble model of the male gonadal axis: Illustrative application in aging men. *Endocrinology* 147: 2817–2828.

Keenan DM and Veldhuis JD (2001) Hypothesis testing of the aging male gonadal axis via a biomathematical construct. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 280: R1755–R1771.

Liu PY, Swerdloff RS, and Veldhuis JD (2004) The rationale, efficacy and safety of androgen therapy in older men: Future research and current practice recommendations. *Journal of Clinical Endocrinology & Metabolism* 89: 4789–4796.

Liu PY, Takahashi PY, Roebuck PD, et al. (2005) Aging in healthy men impairs recombinant human luteinizing hormone (LH)-stimulated testosterone secretion monitored under a two-day intravenous pulsatile LH clamp. *Journal of Clinical Endocrinology & Metabolism* 90: 5544–5550.

Longcope C (1973) The effect of human chorionic gonadotropin on plasma steroid levels in young and old men. *Steroids* 21: 583–592.

Mahmoud AM, Goemaere S, DeBacquer D, et al. (2000) Serum inhibin B levels in community-dwelling elderly men. *Clinical Endocrinology (Oxford)* 53: 139–140.

Morley JE, Kaiser FE, Perry HM, 3rd, et al. (1997) Longitudinal changes in testosterone, luteinizing hormone, and follicle-stimulating hormone in healthy older men. *Metabolism: Clinical and Experimental.* 46: 410–413.

Mulligan T, Iranmanesh A, Kerzner R, et al. (1999) Two-week pulsatile gonadotropin releasing hormone infusion unmasks dual (hypothalamic and Leydig cell) defects in the healthy aging male gonadotropic axis. *European Journal of Endocrinology* 141: 257–266.

Pincus SM, Mulligan T, Iranmanesh A, et al. (1996) Older males secrete luteinizing hormone and testosterone more irregularly, and jointly more asynchronously, than younger males. *Proceedings of the National Academy of Sciences of the United States of America* 93: 14100–14105.

van den Beld AW, De Jong FH, Grobbee DE, et al. (2000) Measures of bioavailable serum testosterone and estradiol and their relationships with muscle strength, bone density, and body composition in elderly men. *Journal of Clinical Endocrinology & Metabolism* 85: 3276–3282.

Veldhuis JD, Iranmanesh A, and Keenan DM (2004) An ensemble perspective of aging-related hypoandrogenemia in men. In: Winters SJ (ed.) *Male Hypogonadism: Basic, Clinical, and Theoretical Principles*, pp. 261–264. Totowa, NJ: Humana Press.

Zmuda JM, Cauley JA, Kriska A, et al. (1997) Longitudinal relation between endogenous testosterone and cardiovascular disease risk factors in middle-aged men. A 13-year follow-up of former Multiple Risk Factor Intervention Trial participants. *American Journal of Epidemiology* 146: 609–617.

Neuroendocrine Control of Energy Balance (Central Circuits/Mechanisms)

R Coppari and J K Elmquist, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Energy balance is maintained when energy intake equals energy expenditure over an extended period of time. This simple statement suggests that regulating body weight is straightforward. Unfortunately, the incidence of obesity is rising at an alarming rate. Fortunately, however, in approximately the past decade, there has been a dramatic increase in the understanding of the mechanisms that the brain uses to coordinate food intake and energy expenditure. Energy intake takes the form of feeding. Energy expenditure takes the form of physical activity, basal metabolism, and adaptive thermogenesis (i.e., energy dissipated as heat in response to changes in environmental temperature and diet). The brain, especially the hypothalamus and regions of the brain stem, serves as an integrator of sensory information and integrates these signals into coordinated responses to regulate energy balance.

Early studies (based on anatomically restricted brain lesions) performed in the 1940s and 1950s demonstrated that the hypothalamus is required for coordinated regulation of energy homeostasis. These studies revealed that bilateral destruction of the hypothalamus in a region centered on the medial and basal hypothalamus (including the ventromedial hypothalamic and arcuate nuclei) results in obesity, whereas bilateral destruction of the lateral hypothalamic area (LHA) causes loss of appetite and weight in rats. This early result has been largely validated by modern neuroanatomic, molecular, and genetic data, and today we have a rough map of the key brain regions required for coordinated control of energy balance. As outlined later, a key component of coordinated control of food intake and body weight is the ability of the hypothalamus to integrate a growing list of metabolic cues, including hormones, nutrients, and feedback from sensory neurons. The hypothalamus integrates all of this information and, in turn, provides neuronal outputs to key effector sites, including the anterior pituitary gland, the posterior pituitary gland, the cerebral cortex, premotor and motor neurons in the brain stem and spinal cord, and autonomic (parasympathetic and sympathetic) preganglionic neurons. The hypothalamic outputs to these effector sites ultimately result in coordinated endocrine, behavioral, and autonomic responses that maintain energy homeostasis.

Within the hypothalamus, several hypothalamic sites are thought to be key in regulating body weight. The first group is composed of neurons located in the medial hypothalamus, including the arcuate nucleus, ventral medial nucleus, dorsal medial nucleus, and paraventricular nucleus. In addition, the lateral hypothalamus (LHA and perifornical hypothalamus) is key in regulating food intake and body weight. The arcuate nucleus is a key hypothalamic nucleus involved in the control of energy balance. For example, the arcuate nucleus contributes to the actions of metabolic signals such as leptin, insulin, and ghrelin. Specifically, proopiomelanocortin (POMC) neurons within the arcuate nucleus are required for regulating body weight, food intake, and glucose homeostasis. POMC is a multifunctional pro-peptide that is differentially processed into key peptides in different tissues. In the brain, α-melanocyte-stimulating hormone (α-MSH) is an important product that regulates food intake and body weight. α-MSH is the agonist for the melanocortin-4 receptor (MC4-R), which is a key regulator of food intake and body weight. An endogenous MC4-R antagonist exists, agouti-related protein (AgRP), that is coexpressed with the orexigenic neuropeptide Y (NPY) in neurons in the arcuate nucleus of the hypothalamus.

Hypothalamic neurons also orchestrate the major neuroendocrine responses occurring upon starvation. The thyroid gland-secreted hormone thyroxine positively acts on basal metabolism. Reduced energy content leads to decreased secretion of thyrotrophin-releasing hormone (TRH) by neurons located in the paraventricular hypothalamic nucleus (PVH). These neurons release TRH into the pituitary portal circulation to stimulate thyroid-stimulating hormone (TSH) secretion by the pituitary gland. Since TSH is essential for normal production of thyroid hormones, its decrease causes reduced blood thyroxine levels and therefore low basal metabolism. The hypothalamic–pituitary–adrenal (HPA) axis is also positively influenced by starvation. The HPA axis consists of PVH neurons secreting corticotrophin-releasing hormone (CRH) directly into the pituitary portal circulation. CRH stimulates the secretion of ACTH by the pituitary.

Key Hormonal Signals

Leptin

Leptin (Greek root *leptos*, 'thin') is one of the most important peripheral signals for energy balance.

Leptin is secreted from adipocytes proportionally to the body fat content. Leptin was discovered in 1994 by Zhang and colleagues, who used a positional cloning approach to identify the mutation causing massive obesity and neuroendocrine abnormalities in *ob/ob* mice. Its receptor was cloned a year later by Tartaglia and colleagues, and five different splice variants have been identified. The mRNA for the signaling form of the leptin receptor (*Lepr-b*) is expressed by selected nuclear groups in the brain. Within the hypothalamus, abundant expression is observed in the arcuate (ARH), dorsomedial (DMH), ventromedial (VMH), and ventral premamillary nuclei. Also, moderate expression is found in PVH and LHA. Outside the hypothalamus, *Lepr-b* mRNA is found in thalamus, midbrain (e.g., in dopamine neurons), caudal brain stem, and cortex. A key concept regarding leptin physiology is that starvation leads to a rapid decrease in leptin levels, which has profound effects on several neuroendocrine systems including activation of the HPA axis, inhibition of the growth hormone and thyroid axes, and inhibition of reproductive function. Therefore, many physiological responses that occur when there is a lack of leptin are also found in a state of starvation. Notably, many of these starvation-induced endocrine and autonomic changes are blocked or blunted by pretreatment with systemic leptin. These observations support the model that leptin is a signal to the brain that energy stores are sufficient.

Soon after the discovery of leptin, it became clear that most of the varied effects of leptin are mediated by the brain. During the past few years, studies have begun to unravel some of the complex circuitry involved in leptin signaling. For example, electrophysiological and neuroanatomical studies have established the identity of some of the neurons that are directly activated (or inhibited) by leptin (i.e., first-order, leptin-sensitive neurons). In the ARH, two classes of neurons respond oppositely to leptin: The orexigenic (i.e., feeding-stimulating) neurons expressing *Agrp* (agouti-related peptide) and *Npy* (neuropeptide Y) are inhibited, whereas the anorexigenic (i.e., feeding-suppressing) neurons expressing *Pomc* (proopiomelanocortin) and *Cart* (cocaine and amphetamine-related transcript) are activated by leptin. In the VMH, neurons expressing steroidogenic factor-1 (SF1) are also activated by leptin. NPY/AGRP and CAR/POMC neurons belong to the melanocortin systems since both secrete neuropeptides that bind to the melanocortin receptors 3 and 4 (MC3Rs and MC4Rs), which are required for normal energy balance. AGRP is the antagonist, whereas the POMC-derived α-MSH is the agonist at these melanocortin receptors.

Studies have selectively deleted leptin receptors in specific neurons. This work has directly tested the physiological relevance of leptin signaling in selected first-order, leptin-sensitive neurons. These studies have demonstrated that mice lacking leptin signaling only in POMC in the arcuate nucleus and SF1 neurons in the VMH develop increased body fat mass. These results have established that ARH and VMH neurons are critical mediators of leptin's body weight-reducing effects. Nevertheless, the degree of obesity in these mice was much less than in mice completely lacking leptin signaling. Therefore, other neuronal populations are also involved in mediating leptin's effects on energy balance. These neurons are likely located in other hypothalamic sites and in neurons in the midbrain and the caudal brain stem. For example, midbrain dopamine neurons within the ventral tegmental area (VTA) are inhibited by leptin, and by perturbing leptin signaling in these VTA neurons, feeding is altered in rats. In addition, delivery of leptin to the dorsal vagal complex (located in the caudal brain stem) reduces food intake and body weight in rats (Table 1). In summary, it is likely that the body weight-reducing effects of leptin are distributed

Table 1 Peripheral signals and central neurotransmitters governing energy balance

Producing tissues	*Hormones*	*Receptors*	*Brain sites*
Hormonal systems with partially identified sites of action in the brain			
Fat	Leptin	LEPR-b	ARH, VMH, VTA, DVC
Pancreas	Insulin	IR	ARH
Stomach and intestine	Ghrelin	GHSR	ARH, VTA
Intestine	PYY	NPY-Y2R	ARH
Hormonal systems with unidentified sites of action in the brain			
Intestine	GLP-1	GLP-1R	?
Intestine	Oxyntomodulin	GLP-1R	?
Stomach and intestine	Obestatin	GPR39	?
Neurotransmitter systems with partially identified sites of action in the brain			
Brain	Serotonin	5-HT2CR and 5-HT1BR	ARH, LPBN
Brain	Dopamine	D1-R and D2-R	CPu

ARH, arcuate nucleus; CPu, caudate putamen; DVC, dorsal vagal complex; LPBN, lateral parabrachial nucleus; VMH, ventromedial nucleus; VTA, ventral tegmental area.

across the brain and involve neurons within the mediobasal hypothalamus, midbrain, and caudal brain stem.

Insulin

Insulin (Latin root *insula*, 'island') was discovered in 1922 by Banting and Best when they first demonstrated its glucose-lowering effects. Insulin is secreted from β cells located in the pancreatic islets of Langerhans, and since its discovery, its role in glucose homeostasis has been established. In the 1970s, insulin's effects on energy balance by acting in the brain were first explored. In 1979, Woods and Porte suggested a role for insulin in the central regulation of energy balance. First, they found that insulin levels circulate in proportion to adipose tissue mass in most mammals. Second, they showed that intracereventricular (ICV) administration of insulin results in a dose-dependent reduction in food intake and body weight in baboons. Other studies have provided evidence for the presence of insulin receptors in the hypothalamus, including POMC neurons. In 2000, Brüning and colleagues assessed the physiological role of insulin signaling in central neurons by generating mice lacking insulin receptors only in the central nervous system (CNS). Their findings established that central insulin receptors are required for normal energy balance since mice with CNS-specific deletion of insulin receptors developed increased adiposity. In 2002, Rossetti's group obtained data supporting the idea that the mediobasal hypothalamus is a key brain area mediating insulin's antifeeding effects. By selective reduction of insulin receptors in the ARH, they caused hyperphagia and increased fat mass in rats. The intracellular molecular players that relay insulin signal into coordinated energy balance and the neuronal populations involved are still to be clarified. To identify these pathways, several experimental approaches designed to perturb insulin signaling in a neuron-specific manner in mice are under way.

Ghrelin

Ghrelin (Proto-Indo-European root of the word 'grow') is secreted by gastrointestinal endocrine cells, which are mainly located in the fundus of the stomach. Ghrelin was discovered in 1999 while searching for ligands of the growth hormone secretagogue receptors (GHSRs). This gut hormone is a ligand at ghrelin receptors and a potent stimulus for pituitary somatotrophs to release growth hormone. In addition to pituitary, ghrelin receptors are expressed in several brain areas, including ARH NPY/AGRP neurons, midbrain neurons within the VTA and substantia nigra nuclei, and in all three components of the dorsal vagal complex (area postrema, nucleus of the solitary tract, and the dorsal motor nucleus of the vagus). Ghrelin levels spike before meals and decrease within 1 h of food intake. This pattern led scientists to investigate the role of ghrelin in feeding behavior and to the finding that exogenous ghrelin administrations (either ICV or peripheral) stimulate feeding. The importance of ghrelin signaling in energy balance has been tested by means of gene deletion. Both the *ghrelin* gene and its receptor (*Ghsr)* gene have been studied in mice null for both genes. These studies support that ghrelin-related pathways are required for normal energy balance, especially when food is high in fat content. In accordance with ghrelin's feeding-stimulating effects, mice lacking either one of these genes eat less and are slimmer than wild-type controls. The sites in the brain that mediate the effects of ghrelin are still to be determined. Studies using gene knockout mouse models have provided evidence that ghrelin's feeding-stimulating effects are mediated by ARH NPY/AGRP neurons in the arcuate nucleus of the hypothalamus. Studies have also proposed a role for VTA neurons as important mediators of ghrelin's feeding effects. It is worth noting that, as for leptin, neurons within the mediobasal hypothalamus and the midbrain VTA are likely important mediators of ghrelin's actions on feeding.

Peptide YY

Peptide YY (PYY) is an anorectic hormone peptide secreted by endocrine L cells in the gut. Its release depends on caloric intake because PYY serum levels increase postprandially. When exogenously delivered, PYY(3-36) reduces food intake in mammals (from rodents to humans). The importance of PYY(3-36) as a negative regulator of energy balance has been established by the finding that *Pyy* null mice have increased food intake and body weight. Pharmacological and gene knockout studies have demonstrated that PYY(3-36) effects on food intake are mediated by NPY Y2 receptors (NPY2Rs). These receptors are expressed in several brain areas, including the presynaptic terminals of NPY/AGRP neurons in the arcuate nucleus. Electrophysiological studies have shown that PYY(3-36) inhibits the electrical activity of NPY/AGRP neurons. It is worth noting that some of these neurons have GABAergic terminals projecting to nearby POMC neurons. Therefore, PYY(3-36)'s inhibitory effects on NPY/AGRP neurons lead indirectly to the activation of adjacent POMC neurons and supposedly to reduced feeding.

Glucagon-like Peptide 1

Glucagon-like peptide 1 (GLP-1) is also secreted by endocrine L cells in the endocrine gut in response to nutrients (mainly free fatty acids (FFAs) and carbohydrates). GLP-1 is one of the peptide products of the prohormone proglucagon peptide that is

cleaved via tissue-specific posttranslational processing. The biologically active form of GLP-1 is the amino-terminally truncated GLP-1(7-36) amide. GLP-1(7-36) amide exerts an array of different physiological effects: It stimulates the release of insulin and inhibits the secretion of glucagon from the endocrine pancreas, and it inhibits gastric emptying and glucose disposal. The effects of GLP-1(7-36) amide on energy intake have been identified by studies from Bloom et al.'s laboratory. They showed that ICV administrations of GLP-1(7-36) amide markedly inhibit food intake in rats. Also in humans, peripheral administration of this peptide induces satiety. Nevertheless, the physiological importance of GLP-1(7-36) amide's effects on feeding have been questioned due to data indicating that GLP-1(7-36) amide causes conditional taste aversion in rats. It is interesting to note that GLP-1Rs are found in key hypothalamic nuclei controlling energy balance (e.g., the VMH, ARH, LHA, and DMH). Moreover, GLP-1Rs are also found in caudal brain stem centers (area postrema and nucleus tractus solitarius) important for the autonomic control of energy balance. Whether GLP-1(7-36) amide activates these centers to suppress feeding has yet to be established. Oxyntomodulin (OXM) is also a satiety peptide originated by the posttranslational cleavage of proglucagon in L cells of the gut. OXM has potent anorectic effects in both rodents and humans. These effects are thought to be mediated by activation of GLP-1R on ARH neurons. However, studies are needed to clarify the physiological effects of these two hormones and also to understand whether their action on energy balance is solely mediated by central neurons.

Given the rising rates of obesity and diabetes, there is intense interest in the development of new therapies that safely and effectively lower body weight and blood glucose in diabetic subjects. A new class of drugs called incretin mimetics have been approved for treatment of type II diabetes. One of these drugs mimics the effects of GLP-1, enhancing glucose-dependent insulin secretion, slowing gastric emptying, inhibiting gastric acid secretion, and reducing food intake. Possibly, this drug acts by modulating GLP-1-sensitive neurons in the CNS involved in the control of energy homeostasis.

Metabolic Signals

During the past few years, it has become established that in addition to key hormones, the hypothalamus also directly responds to the levels of several circulating metabolites (including FFAs and glucose) to serve as a reflection of the status of the body energy supplies. For example, the food intake-reducing effects of long-chain fatty acids are mediated by neurons in the mediobasal hypothalamus. In the cytoplasm of neurons (and perhaps glial cells), FFAs are rapidly esterified to fatty acyl-coenzyme A (fatty acyl-CoA) and then may undergo β-oxidation in the mitochondria. The entering of long-chain fatty acyl-CoAs into the mitochondria requires two mitochondrial proteins – carnitine-dependent long-chain acyltransferases 1 and 2 (CPT1 and CPT2). The observation that inhibition of hypothalamic CPT1 activity causes increased accumulation of long-chain fatty acyl-CoAs in the cytoplasm and marked inhibition of food intake has provided new insights into the molecular mechanisms of brain lipid sensing. The key neurons mediating this effect remain to be determined.

Blood glucose levels are also sensed by neurons within the CNS. More than 40 years ago, it was shown for the first time that changes in blood glucose levels affect the frequency of action potentials recorded from single hypothalamic neurons in living cats and dogs. It was found that the majority of neurons within the VMH are activated by an increase in extracellular glucose levels, whereas LHA neurons are mainly inhibited by it. The identification of single populations of glucose-activated and glucose-inhibited neurons has became more expedite only recently by means of electrophysiological studies on hypothalamic slice preparations obtained from transgenic mice expressing green fluorescent protein in a neuron-specific manner. Such studies revealed that POMC neurons in the arcuate nucleus are glucose activated. In contrast, LHA neurons expressing the feeding-stimulating peptide orexin are glucose inhibited neurons. The glucose-induced modulation of the activity of these neurons is in agreement with their physiological role in energy balance. It is also important to note that brain stem neurons are also critical in sensing changing levels of glucose, especially hypoglycemia.

Neurotransmitters

Serotonin

Serotonin (5-hydroxytryptamine (5-HT)) is a biogenic amine synthesized from the essential amino acid tryptophan and produced in cells present in peripheral tissues as well as in neurons in the brain. Central neurons producing this neurotransmitter (serotoninergic neurons) are located in the raphe nuclei of the brain stem and project to virtually all parts of the CNS (from cortex to spinal cord). Serotonin binds to several receptors, all of which are G-protein-coupled receptors except for the subtypes 3, which are ligand-gated ion channels receptors. The serotonin systems are essential for several brain functions, such

as attention, cognition, mood, perception, and sleep. Its involvement in appetite was first discovered in 1976 by Breisch and colleagues. They used a pharmacological approach to deplete the brain of serotonin and observed increased feeding and obesity in rats.

In the 1990s, gene targeting approaches aimed at disrupting the gene encoding for either the serotonin 2C receptors (5-HT_{2C}Rs) or serotonin 1B receptors (5-HT_{1B}Rs) revealed that these receptors are required for normal energy balance because mice lacking either one of these genes were found to have increased body weight. Thus, the serotonin systems, with their feeding-reducing effects, are crucial components of the central circuitries that coordinate feeding behavior. Which neurons are involved in mediating serotonin's hypophagic effects? There is evidence that neurons located in the brain stem and in the hypothalamus are important. In the hypothalamus, among many neuronal groups, POMC and NPY/AgRP neurons express serotonin receptors. Specifically, the anorexigenic POMC neurons express 5-HT_{2C}Rs, and NPY/AgRP neurons express the G_i-coupled 5-HT_{1B}Rs. Thus, serotonin can directly activate POMC neurons and directly inhibit NPY/AgRP neurons. In addition to direct action, serotonin exerts indirect stimulating effects on CART/POMC neurons. This is achieved by inhibition of NPY/AGRP neurons that have GABAergic terminals projecting to nearby POMC neurons. As noted previously, these two hypothalamic neuronal populations are known to have opposite effects on melanocortin receptor-expressing neurons. Thus, serotoninergic neurons control feeding behavior in part by engaging central circuitries with upstream hypothalamic melanocortin (NPY/AGRP and CART/POMC) neurons and downstream MC4Rs-expressing neurons. Since the serotonin systems are thought to be altered in some psychiatric disorders, such as depression and anorexia nervosa, it is plausible that the serotonin–melanocortin pathways may be the link between these pathological conditions and some of their associated complications (e.g., alteration of feeding and weight loss).

Dopamine

Dopamine (DA) is a neurotransmitter synthesized from the amino acid tyrosine. DA-secreting (dopaminergic) neurons reside in many brain areas, but the majority of them lie in midbrain nuclei such as the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc). DA binds to five different receptors (D1–D5R), some of which, when activated, elicit an increase in intracellular cyclic AMP (D1R and D5R). The dopamine system is essential for several mammalian behaviors, including coordinated movement, reward responses, and learning. Although pharmacological experiments provided support for a role of dopamine in feeding behavior, the critical studies were performed using the generation of mice deficient in dopamine. These mice developed locomotor deficits (bradykinesia) and severe hypophagia. The fact that mice lacking dopamine starve if not otherwise treated with exogenous L-DOPA supports the hypothesis that dopaminergic neurons are critical components of the central circuitries controlling energy balance. In agreement with dopamine's effects on energy balance, mice lacking D1R or D2R have reduced body weights. VTA and SNc dopaminergic neurons send inputs to many brain sites, including the caudate nucleus and the putamen in the telencephalon. Although the physiological relevance of these dopaminergic networks has been established for normal and coordinated ambulatory movements, studies performed by Palmiter's group indicate an additional role for the VTA + SNc innervations in the caudate putamen (CPu). They showed that restoration of dopamine signaling to the CPu (mainly from SNc neurons) is sufficient to restore normal feeding (as well as motor behavior) in mice otherwise deficient in DA. The identity of the key neurons within the dopamine systems involved in energy balance is unknown. However, as previously described, studies have brought to light a potentially important role for VTA DA neurons in responding to leptin and ghrelin to coordinate feeding behavior.

In summary, dopamine circuitries appear to be critical for controlling feeding behavior, and dysregulation of these circuitries may contribute to obesity and may also underlie the wasting syndrome characteristic of patients with Parkinson's disease.

See also: Energy Homeostasis: Paraventricular Nucleus (PVN) System; Food and Water Intake: Regulation; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Hypothalamo–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH; Neuroendocrinology; Neuropeptides: Food Intake.

Further Reading

Badman MK and Flier JS (2005) The gut and energy balance: Visceral allies in the obesity wars. *Science* 307: 1909–1914.

Bruning JC, Gautam D, Burks DJ, et al. (2000) Role of brain insulin receptor in control of body weight and reproduction. *Science* 289: 2122–2125.

Elmquist JK, Elias CF, and Saper CB (1999) From lesions to leptin: Hypothalamic control of food intake and body weight. *Neuron* 22: 221–232.

Flier JS (2006) Regulating energy balance: The substrate strikes back. *Science* 312: 861–864.

Heisler LK, Jobst EE, Sutton GM, et al. (2006) Serotonin reciprocally regulates melanocortin neurons to modulate food intake. *Neuron* 51: 239–249.

Morton GJ, Cummings DE, Baskin DG, and Schwartz MW (2006) Central nervous system control of food intake and body weight. *Nature* 443: 289–295.

Spiegelman BM and Flier JS (2001) Obesity and the regulation of energy balance. *Cell* 104: 531–543.

van den Pol AN (2003) Weighing the role of hypothalamic feeding neurotransmitters. *Neuron* 40: 1059–1061.

Zhou QY and Palmiter RD (1995) Dopamine-deficient mice are severely hypoactive, adipsic, and aphagic. *Cell* 83: 1197–1209.

Neuroendocrine Control: Maternal Behavior

R S Bridges and B C Nephew, Tufts University, North Grafton, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Maternal behavior emerges and wanes over the course of development in female mammals. In recent years, our understanding of the biochemical regulation of maternal care and the neural substrates involved in regulating the expression of maternal behavior has significantly increased. This article focuses primarily, but not exclusively, on research conducted in rodents, species that have provided a foundation for our understanding a complexity of processes associated with raising and nurturing young. The initial objective is to identify or define maternal behavior. The underlying physiological events that transpire during pregnancy and lactation are then presented in light of their roles in affecting maternal care. Next, the actions of the hormones of pregnancy on the neural substrates that regulate maternal responsiveness are discussed as well as the role of genetic and epigenetic factors and their interactions over the course of development. Finally, the effects of experience on the modification of parental care are explored. This article is not fully comprehensive; rather, it is designed to present a broad overview of the neuroendocrine and endocrine–neural regulation of maternal behavior.

Measurement of Maternal Behavior

The neurobiological capacity to respond in a maternal or parental-like manner is present throughout most of prepubertal and adult development in female mammals. The capacity to display enhanced maternal care, however, is most pronounced coincident with the birth of young and during lactation. Specifically, the onset of maternal behavior is associated with increased interest in the sensory modalities of the amniotic fluid and pup-related olfactory cues and an enhanced motivation to respond to the young.

What behaviors emerge at approximately the time of parturition that help promote maternal care and reproductive success? In the rat and other rodents, the new mother ingests amniotic fluid as well as engages in placentophagia. These behaviors are thought to both reduce the likelihood of predators finding the nest site by removing olfactory cues and provide possible hormonal sources from the amniotic fluid and placentas that appear to help promote lactogenesis and provide energy and fluid sources for the new mother (Table 1). During the postpartum period, mothers display a range of young-oriented as well as non-young-oriented responses that are unique to the lactational phase of care. Among these behaviors are retrieval to place the young into the nest site, pup licking to stimulate urination and to enhance sensory stimuli from the mother, and crouching over the young to permit suckling and modulate thermal loss and needs.

In addition to these young-oriented responses, the dam exhibits increased levels of nest building to provide a centralized location for the young; a heightened level of aggression toward intruders, a response that first emerges during late pregnancy and is maintained at high levels throughout much of lactation; increased feeding to generate food resources for the young (i.e., milk); and reductions in anxiety.

Developmental Aspects of Maternal Behavior

Maternal behavior is separated into an onset phase and a maintenance phase. The onset phase focuses on the initial expression of maternal care at parturition. In contrast, the maintenance phase consists of maternal care during lactation and care beyond the end of lactation. As noted previously, the neurobiological ability to display parental care is present independent of the reproductive states of pregnancy, parturition, and lactation. In female rats, maternal care is typically measured by the animal's latencies to display the more active components of pup-oriented care toward young (i.e., pup retrieval, grouping of the young, and crouching over the young). Interestingly, when exposed to 3- to 8-day-old foster pups, prepubertal females display a very rapid onset of maternal behavior as early as 20–24 days of age. As shown in Figure 1, this response latency increases from approximately 1 or 2 days at 24 days of age to approximately 5 or 6 days by 30 days of age. The latencies to respond maternally remain at this level unless a female becomes pregnant. The gravid female's response latency toward foster young decreases prior to parturition, indicating that the expectant mother is ready to exhibit spontaneous maternal care. Once maternal care is initiated, postpartum maternal behavior persists until weaning.

The ability of the parous female to display a rapid onset of maternal behavior, once established, persists well into adulthood. It has been shown that as little as 1 or 2 h of postpartum maternal experience is sufficient to establish a 'maternal memory.' That is, parturient mothers will remember or retain how to be a mother for more than 1 month following a brief

maternal postpartum experience. The maternal response latencies remain at approximately 1 day compared with a latency score of 5 or 6 days for a female that is cesarean delivered 1 day prior to expected parturition but not given any immediate pup contact and maternal experience following surgery.

In addition to pup-directed behaviors, most lactating mammals exhibit heightened levels of maternal aggression. Maternal aggression is a specialized form of aggression exhibited by females just prior to parturition and during lactation. This type of aggression is remarkable in that females are often able to display high levels of both maternal aggression and parental care in the same period. In mice and voles, maternal aggression is actually inversely correlated with attacks on their own young. Maternal aggression may increase the fitness of the female by increasing the probability of offspring survival. The vast majority of studies investigating various aspects of maternal aggression have used rodent models, in which aggression is most robust during the first postpartum week, followed by a decline in intensity as lactation progresses. The behavior consists of frontal attacks, lateral attacks, jump attacks, rearing, biting, arching of the back, piloerection, and upright boxing. Olfaction is an important component in the expression of maternal behavior in rats, mice, and hamsters. Whereas intruder vocalizations and sensory stimulation of the nipples are important stimuli in mice, maternal aggression in rats is more dependent on olfaction and hormonal priming by estrogen and progesterone. Nevertheless, although slight differences exist in the expression of maternal behavior across rodent species, the general behavioral ethogram is similar for most species.

Table 1 Maternal Behaviors

Parturitional responses
Consumption of amniotic fluid
Placentophagia
Young-directed responses
Retrieval
Licking and grooming
Crouching/nursing
Nest building
Young-related responses
Protection of the young – maternal aggression
Increased food consumption
Reduced anxiety

The responses displayed by mothers include a set of parturitional behaviors together with responses directed toward the young. An additional set of related responses are present during lactation that further support the development and protection of the offspring.

Hormonal Regulation of Maternal Behavior

The endocrine events accompanying pregnancy, in addition to maintaining a healthy environment for fetal growth and development and preparing the mother to meet the nutritional needs of her offspring, also play an important role in altering the affective state of the female during pregnancy and postpartum. In studies of rats, critical roles for both steroid and protein hormones have been demonstrated, especially during the initial pregnancy. These hormones appear to act synergistically to prepare the newly parturient dam to be fully maternal at the time of birth.

Steroid Hormones

Estradiol Estradiol (E_2) which is synthesized by the ovaries and conceptus during pregnancy has an

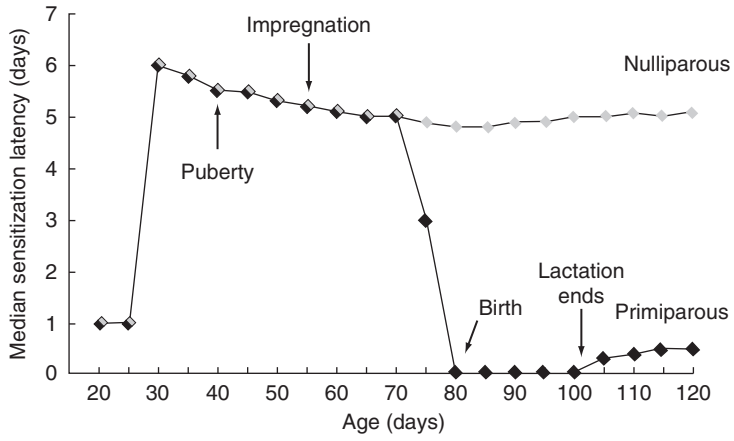

Figure 1 A developmental profile of maternal responsiveness in female rats. Points represent the median latencies in days for a female to respond maternally to pups – that is, retrieve, group, and crouch over the young during a 1 h daily test session. Reproduced from Bridges RS (1990) Endocrine regulation of parental behavior in rodents. In: Krasnegor NA and Bridges RS (eds.) *Mammalian Parenting: Biochemical, Neurobiological, and Behavioral Determinants*, pp. 93–117. New York: Oxford University Press, with permission.

important role in stimulating the onset of maternal behavior in rodents as well as sheep. The pattern of increasing blood levels of E_2 during pregnancy is shown in Figure 2. In addition to its own action as a behavioral inducer, E_2 acts permissively to allow other hormones and neurotransmitters to exert stimulatory actions on maternal care. That is, progesterone (P), prolactin (PRL), as well as oxytocin (OT) all require the presence of E_2 for their actions. Studies using a rat model have further established that E_2 acts centrally at the level of the medial preoptic area (MPOA) to stimulate a rapid onset of maternal behavior. Other studies have found that E_2's actions are mediated through the alpha subtype of the estrogen receptor (ER), since mice with deletions of this receptor display deficits in maternal care.

Progesterone – 'Priming and timing' P, the progestational hormone, is synthesized in large amounts during the course of pregnancy in most mammals (Figure 2). In the pregnant rat, the corpora lutea in the ovaries are the primary source of P, whereas in primates the placenta is the major site of P production. P has two distinct roles in the induction of maternal behavior. First, the prolonged exposure to P over the course of pregnancy primes the maternal brain to respond intensively to a host of stimulatory signals at approximately the time of parturition. In addition to this 'priming action,' P exerts an inhibitory action on the expression of the rapid onset of maternal behavior. The prepartum decline in P prior to parturition serves as a 'timing signal' that coordinates the stimulation of a range of physiological and neurobiological processes. These processes include stimulation of uterine contractions, enhanced sensitivity to the actions of OT on parturition, stimulation of lactogenesis, and stimulation of maternal care. The behavioral and physiological actions of P apparently

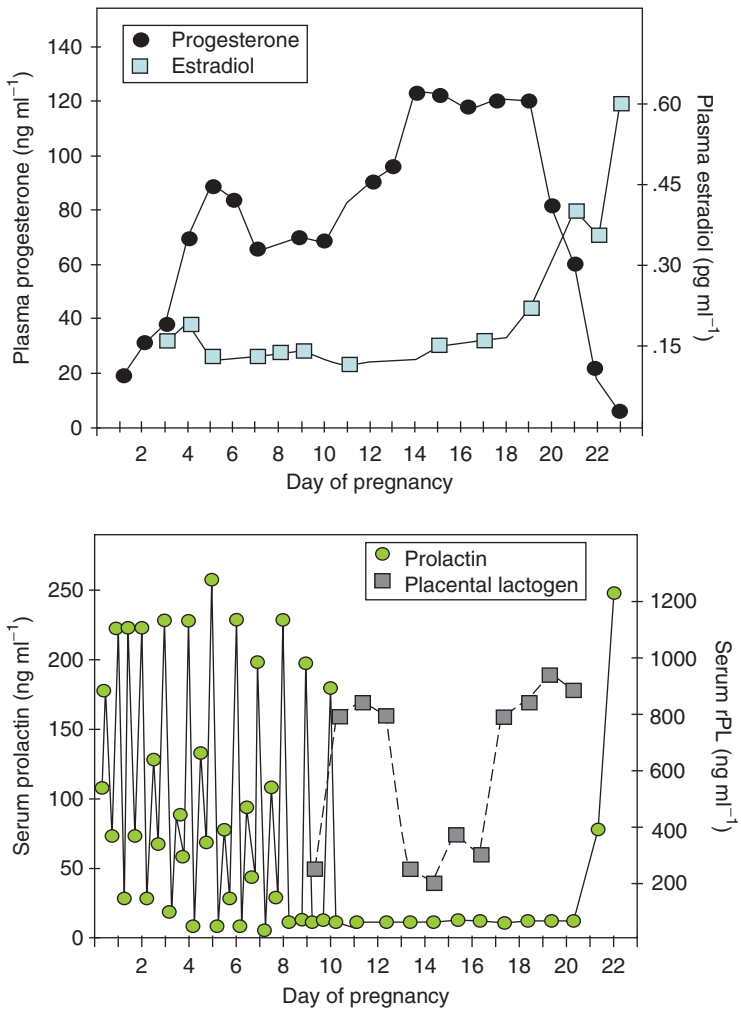

Figure 2 Hormone profiles of estradiol, progesterone, prolactin, and rat placental lactogens during pregnancy in the rat. Adapted from Bridges RS (1990) Endocrine regulation of parental behavior in rodents. In: Krasnegor NA and Bridges RS (eds.) *Mammalian Parenting: Biochemical, Neurobiological, and Behavioral Determinants*, pp. 93–117. New York: Oxford University Press, with permission.

are mediated through the progesterone receptor (PR) since treatment of rats with the PR antagonist RU-486 stimulates a more rapid onset of maternal behavior after pregnancy termination. It has been reported that aggression toward pups is reduced and caregiving behaviors are increased in male mice in which the PR is knocked out. Thus, deactivation or inactivation of the P system helps to facilitate parental care. It is of interest that in humans P levels do not decline until after delivery of the placenta. However, expression of the PR in the uterus declines prepartum, suggesting that in women a functional P withdrawal occurs in the uterus and perhaps in the brain at this time.

Protein Hormones

Prolactin/placental lactogens The lactogenic hormones, PRL and placental lactogens (PLs), are secreted in copious amounts during the course of pregnancy and lactation (Figure 2). Both PRL and PLs bind to PRL receptors (PRL-Rs) located within the choroid plexus and brain. Treatment of inexperienced, virgin rats with a combination of PRL, E_2, and P can stimulate a rapid onset of maternal behavior, reducing latencies from 5 or 6 days to 1 or 2 days. Blocking secretion of PRL in hormonally primed virgin rats prevents the rapid onset of maternal care, an effect that is reversed by concurrent treatment with PRL. The actions of PRL as well as PLs in priming the brain to respond maternally at parturition appear to occur during the course of pregnancy and not just at parturition. Within a physiological setting, therefore, prolonged exposure to the hormonal milieu of gestation alters the neural processes that account for the subsequent rapid onset of maternal care. It is of interest that in both pregnant rats and women, PLs gain access to the cerebrospinal fluid and bathe the gestating brain in hormonal signals that communicate that the female is pregnant. Evidence in support of a role for the PRL-R in mediating the behavioral actions of PRL is provided by studies of steroid-primed virgin rats in which infusions of the PRL-R antagonist, S179-D PRL, into the brain delay the rapid onset of maternal behavior. Likewise, mice with null mutations of the PRL-R respond slower to foster pups than do wild-type subjects.

Neural Basis of Maternal Behavior

Neural Network of Maternal Behavior

The research of neuroscientists Michael Numan and Alison Fleming has provided a solid foundation on which our understanding of the neural regulation of maternal behavior is based. Numan's early finding that the integrity of the MPOA was crucial for the expression of ongoing maternal care in rats provided an important foundation on which to establish a maternal neural network. Overall, the MPOA appears to be a convergence point that processes information from sensory and motivational systems and translates this information into downstream motor output. Fleming's work, which first established a key role for olfactory input, has gone on to identify neural sites, such as the nucleus accumbens (NAcc), which can affect motivation, as well as cortical inputs from the prefrontal cortex that influence maternal affect. A schematic representation of a neural network involved in the regulation of young-oriented maternal behavior is depicted in Figure 3.

Although a complete characterization of the neural connections and actions is beyond the scope of this article, it is important to recognize that the olfactory system plays a valued role in regulating maternal care in

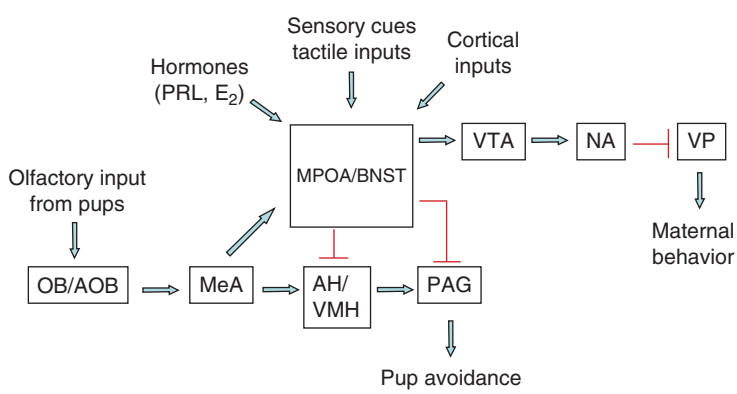

Figure 3 Schematic of key neural regions and connections that constitute the maternal neural network in the rat. Blue arrows represent stimulatory inputs, whereas red connections represent inhibitory inputs AH, anterior hypothalamus; AOB, accessory olfactory bulb; BNST, bed nucleus of the stria terminalis; MeA, medial amygdala; MPOA, medial preoptic area; NA, nucleus accumbens; OB, olfactory bulb; PAG, periaqueductal gray; VMH, ventromedial hypothalamus; VP, ventral pallidum; VTA, ventral tegmental area.. Adapted from Numan M (2006) Hypothalamic neural circuits regulating maternal responsiveness toward infants. *Behavioral and Cognitive Neuroscience Reviews* 5: 163–190, with permission.

numerous species. During adult development in the inexperienced rat, this system exerts a primary inhibitory role on the expression of maternal care. Lesions of the olfactory bulb and tract as well as its principal projection site (i.e., amygdala) actually stimulate maternal care. The projections from the amygdala terminate in various hypothalamic nuclei, including the MPOA and anterior hypothalamus, with the latter region also being inhibitory to maternal behavior. It has been proposed that naive females find the smell of young aversive and must overcome this aversion in order to exhibit maternal care. In support of this concept are the findings that rats find the smell of young more attractive during late pregnancy and ewes are attracted to the odors of amniotic fluid soon before birth.

Whereas the amygdala may influence MPOA function, the output from the MPOA to lower brain centers involves downstream projections to the ventral tegmental area (VTA). The VTA in turn projects to the NAcc, a structure that reflects the motivation state of the female to engage in maternal care during the postpartum period and lactation. Lesions of the shell region of the NAcc prior to parturition disrupt the establishment of maternal care and maternal memory. Other research in prairie voles suggests that this region mediates OT-related pair bonding. Finally, the NAcc inhibits activity of the ventral pallidum, resulting in enhanced behavioral reactivity to pups.

Neural Underpinnings of Maternal Aggression

Compared to the wealth of knowledge on the neuroanatomy of parental behavior, the neuroanatomy of maternal aggression is less well understood. Studies indicate that there are numerous areas of overlap in the neural regulation of parental care and aggression. Brain regions that increase in activity in response to maternal aggression include the bed nucleus of the stria terminalis (BNST), MPOA, paraventricular nucleus (PVN), amygdala, septum, peripedunacular nucleus (PPN), and periaqueductal gray (PAG).

Ventral sensory inputs associated with suckling appear critical for the display of maternal aggression. One key neural region involved in the transmission and processing of this sensory information is the PPN, which transmits sensory information to the midbrain, hypothalamus, and several limbic structures. Lesions of the PPN decrease maternal aggression.

The PVN is another key neural component in the control of maternal aggression. PVN lesions may increase or decrease aggression in lactating dams, depending on the stage of lactation in which lesioning is performed. PVN lesions early postpartum increase maternal aggression, whereas late postpartum lesions decrease aggression. The effects of early lesions may be due to a decrease in OT neurons since the disruption of OT signaling in the PVN can increase some components of maternal aggression. In the PAG, both pre- and postpartum lesions reduce attack latency and increase the number of attacks, indicating a substantial inhibition of maternal aggression by this nucleus in intact animals.

Finally, increases in cell activity are seen in the MPOA in response to maternal aggression, and inhibition of nitric oxide (NO) synthase suppresses aggression and reduces MPOA c-Fos expression in this region, suggesting that the MPOA is part of the neural circuitry controlling both pup-related maternal behavior and maternal aggression.

Hormone–Brain Interactions

Studies have localized neural sites of action for estrogen, PRL, and OT in the regulation of maternal behavior. The central site(s) of P action in regulating the priming and timing actions on maternal care has not been demonstrated, although hypothalamic areas are likely involved. One common key central site of E_2 and PRL action is the MPOA. Using a pregnancy termination model, it was first demonstrated that local application of estrogen to the MPOA stimulated a more rapid onset of maternal care in first-time pregnant rats than in control animals. A more detailed study of PRL's action showed that central infusions of PRL into either the lateral ventricle or bilaterally into the MPOA stimulate a rapid onset of maternal care in steroid-primed, virgin rats (Figure 4). PRL

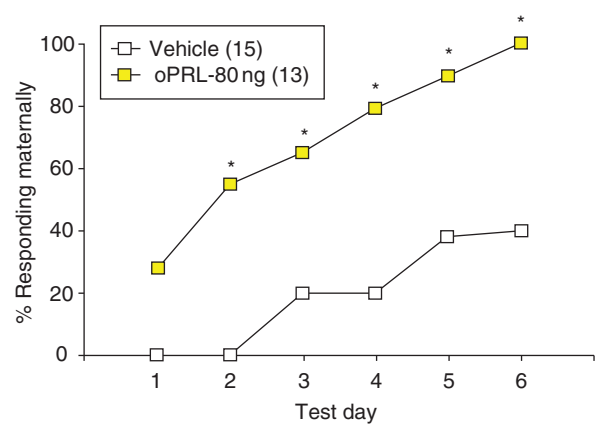

Figure 4 Stimulation of maternal behavior after bilateral infusions of ovine (o) PRL (40 ng/side) into the MPOA region of steroid-primed nulliparous rats. $^*p < 0.01$ versus vehicle controls. Reproduced from Bridges RS, Numan M, Ronsheim PM, Mann PE, and Lupini CE (1990) Central prolactin infusions stimulate maternal behavior in steroid-treated, nulliparous female rats. *Proceedings of the National Academy of Sciences of the United States of America* 87(20): 8003–8007, with permission.

infusions were ineffective in the absence of steroid priming. Infusions of PRL into the ventromedial hypothalamus (VMH) were also suggestive of a similar stimulatory role for PRL in this region. It is noteworthy that MPOA has numerous PRL and E_2 receptors which can bind their respective ligands. A related study in mice examined the role of STAT5B in maternal behavior. STAT5B is a component of a key transduction pathway that mediates PRL's actions after PRL binds to its receptor. In STAT5B knockout animals, maternal behavior was unaffected. Thus, the STAT5B pathway's integrity appears not to be essential for normal maternal care. Further studies are needed to investigate the pathways through which PRL and E_2 stimulate maternal care.

In addition to PRL's ability to stimulate maternal behavior in inexperienced rats, related research has demonstrated that rat PL-I and rat PL-II gain access to the cerebrospinal fluid during pregnancy and thus may play an important physiological role in priming the maternal brain to respond maternally. Central infusions of these PLs into the MPOA, like PRL, stimulate the onset of maternal care, presumably through their action on PRL receptors.

One intriguing possibility that merits study is whether the increase in neurogenesis during gestation in rodents that is reported under PRL and perhaps PL regulation may alter the mother's responsiveness to her young. New neurons produced within the subventricular zone migrate along the rostral migratory stream and become integrated into the olfactory bulb. Certainly, given the key role for olfaction in maternal care in numerous mammals, an involvement of these new neurons in olfactory recognition of young is possible.

In addition to pituitary PRL's role in stimulating maternal behavior, a role for centrally derived PRL in maternal care is also possible. The expression of neural PRL in the hypothalamus of lactating rats increases in response to both stress and suckling in the hypothalamus of lactating rats (Figure 5), in which neural PRL may act locally to reduce anxiety and affect other behaviors. The possible role of brain-derived PRL in maternal care merits further study. It is conceivable that once maternal care is induced via an endocrine mechanism at parturition, the central sensitivity to neural-derived PRL is sufficient to maintain maternal care.

An important role for the neuropeptide OT in maternal behavior in rodents and sheep and in affiliative behavior in voles has been well documented. Specifically, when OT is infused into either the ventricular system or the MPOA of E_2-primed, inexperienced rats, a rapid onset of maternal behavior occurs. In sheep, during the periparturitional period vaginal distension and stimulation induces a central release of OT which contributes to the ewe's acceptance of her newborn lamb. OT can also act within the NAcc to promote maternal-like behavior in juvenile rats and to stimulate pair bond formation in prairie voles. Thus, OT can be conceived of as a more general promoter of affiliative responses. The neurobehavioral relationships between PRL and OT have not been firmly established. It is of interest, however, that PRL receptors are located on OT neurons in the PVN. Therefore, it seems tenable that these two crucial lactational hormones may also act in concert to stimulate given aspects of maternal behavior.

Neurotransmitter Systems

Pup-Directed Maternal Behavior

Roles for the more classic brain dopaminergic, noradrenergic, and opioidergic systems have been

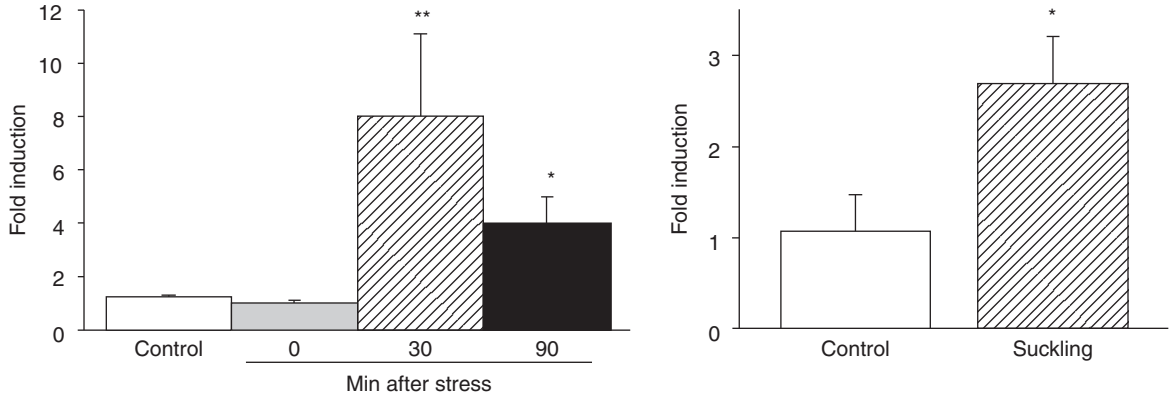

Figure 5 Effects of a 30 min immobilization stress and suckling on prolactin mRNA in the rat hypothalamus. Determination of fold induction of PRL mRNA is expressed as a ratio with mean ± SEM. $^*p < 0.05$ and $^{**}p < 0.01$ versus the control group. Reproduced from Torner L, Maloumby R, Nava G, Aranda J, Clapp C, and Nuemann ID (2004) *In vivo* release and gene upregulation of brain prolactin response to physiological stimuli. *European Journal of Neuroscience* 19: 1601–1608, with permission.

identified in the expression of maternal care in mammals. The dopaminergic system's involvement includes both motoric and motivational elements. Activation of the nigrostriatal dopaminergic system produces increases in pup retrieval and activity levels. The striatal dopaminergic system, through the actions of dopamine on the D2 receptors, also modulates maternal motivation and the establishment of maternal memory. The involvement of the noradrenergic system is somewhat less defined and equivocal. Perhaps the most convincing support for a role for norepinephrine (NE) is its role in olfactory-mediated maternal care in ewes. Stimulation of the birth canal during parturition results in the release of NE within the olfactory bulbs, and pharmacological blockade of NE at this time interferes with the establishment of maternal behavior in ewes. Whereas surgically induced depletions of hypothalamic NE only modestly affect the onset of maternal care at birth in rats, some studies have indicated a role for the ascending noradrenergic systems in the onset of maternal behavior. It is possible that increased activity of the noradrenergic system at the time of parturition facilitates a more general increase in attention and arousal which enables the new mother to respond to the sensory stimuli provided by the offspring.

Research also indicates that the endogenous opioid system may modulate maternal responsiveness. In rats, infusions of opiates that bind to the mu opiate receptor block ongoing maternal behavior when infused into the MPOA but not the VMH. Since suckling is a strong stimulus for inducing opiate release within the brain, it has been suggested that endogenous opiates, such as beta endorphin, may function physiologically to modulate nursing bouts in lactating rats by quieting the female once suckling has begun. Termination of the suckling bout may be produced by the increase in body temperature that accompanies endogenous opiate release and activity. Opioid peptides may also affect maternal behavior through their ability to alter brain dopamine activity. For example, beta endorphin can suppress dopamine function and perhaps in that manner reduce either maternal motivation or motor function.

Another neurotransmitter, NO, likewise has been implicated in regulating ongoing maternal care in rodents. Intracerebroventricular infusions of an NO synthase inhibitor (L-NAME) into rats eliminated pup retrieval during early lactation as well as decreased milk ejections. Possible sites of NO action include the MPOA and amygdala. Further studies are warranted to investigate the relationships between the endocrine events of pregnancy and lactation and NO activity in light of these behavioral findings.

Maternal Aggression

Neurotransmitters implicated in the expression of maternal aggression include serotonin, GABA, dopamine, OT, arginine vasopressin (AVP), corticotrophin-releasing hormone (CRH), NO, and opioids. In a correlational study of humans, central serotonin was inversely correlated with male aggression but not maternal aggression. A similar result is seen in biparental prairie voles, in which selective serotonin reuptake inhibitors decrease male aggression but have no effect on maternal aggression. However, work in pigs has recorded decreased serotonin activity in the medial amygdala and lateral septum in aggressive female pigs.

GABA receptor antagonists injected into the VMH or medial amygdala decrease the number of female attacks, and benzodiazepine agonists increase maternal aggression at low doses. In mice, benzodiazepine agonists increase aggression toward males but decrease maternal aggression toward females in aggressive dams. Evidence of the role of dopamine in maternal aggression is inconclusive. Whereas decreasing catecholaminergic activity by global or directed lesioning increases rat maternal aggression, treatment with haloperidol, a D1/D2 receptor antagonist, actually decreases mouse maternal aggression. It is possible that the actions of haloperidol in mice may be due to nonspecific effects on locomotion.

Maternal aggression increases in rats after infusion of OT into the central nucleus of the amygdala. Lesion studies, on the other hand, have not found effects of neural lesions of OT nuclei on this behavior. It has been proposed that OT may mediate a dam's ability to identify intruders through its established role in social recognition. OT cells in the PVN and supraoptic nucleus are activated during maternal aggression, and OT concentrations increase in the central amygdala of 'high-anxiety' rats during resident intruder challenges. When OT activity is blocked in these high-anxiety dams with antagonist injections into the PVN or central amygdala, maternal aggression decreases. Conversely, exogenous OT infusions into the PVN increase maternal aggression.

Studies suggest that AVP may also modulate maternal aggression. Cellular activity in the PVN and SON of lactating rats increases in association with maternal aggression, as reflected by increased double-labeling for c-Fos and AVP. Likewise, AVP receptor expressing cells in the central amygdala and BNST have increased c-Fos activity. Moreover, in lactating rats bred for high levels of anxiety, both OT and AVP release increases in the central amygdala during resident–intruder challenge, and exogenous AVP injected into the central amygdala increases aggression.

NO also appears to modulate the display of maternal aggression. Treatment with the NO synthase antagonist L-NAME suppresses maternal aggression. Neuroanatomical studies have found that female mice and prairie voles increase citralline levels, a by-product of NO synthesis, in several brain nuclei in response to the display of aggressive behavior. Additional work in prairie voles indicates that the blockade of NO synthesis depresses maternal aggression. It is possible that the effects of NO on maternal aggression are related to its ability to modulate CRH release. Studies using central CRH injections show a decrease in maternal aggression in response to the peptide, and since NO and CRH may be negatively correlated, this mechanism is feasible. When NO increases, there may be a corresponding decrease in CRH that potentiates increased expression of aggression.

Genes and Motherhood

Null Mutations

With the development of techniques in genetics and molecular biology, a number of studies have identified roles for specific genes in maternal responsiveness. One of the earlier studies reported that knocking out the immediate early oncogene, Fos-B, resulted in severe parenting deficits in mice. Similar deficits in maternal care in mice were reported after null mutations of the long form of the PRL-R but not after deletion of the PRL gene. A careful reexamination of these findings is warranted. Likewise, mice in which the OT gene was deleted, while failing to lactate, did not display deficits in maternal care. In contrast, in a related study, deletion of the CD38 gene, which is a transmembrane glycoprotein that regulates OT release within the brain, resulted in notable deficits in both maternal and social behaviors. A similar deficit in pup retrieval was present in mice when the gene dopamine β-hydroxylase, the rate-limiting enzyme controlling NE and epinephrine synthesis, was deleted.

Overall, it is worth noting that interpretation of the behavioral consequences of gene deletions should be made with caution since prolonged developmental absence of a ligand or receptor may result in physiological compensation or emphasis on redundant systems. Moreover, the behavioral specificity of such deletions needs to be established to ensure that the deletions do not result in more global biochemical modifications that can indirectly affect behavior.

Parent of Origin Effects

The genetic materials inherited from a mother and father can be differentially expressed in the offspring. In some instances, select genes inherited from the father, but not the mother, are expressed. This can result in differences in maternal responsiveness. One example of this is provided by studies of Peg 1, Peg 3, and Mest knockout mice that display deficits in maternal care. The behavioral responses of the female offspring can differ based solely on the paternal gene that they receive since genes inherited from the father are expressed by the daughter offspring, whereas the maternal alleles are methylated and thus inactive. This phenomenon, referred to as imprinting, can account for variation in parental care based on the paternal allele. Other studies have shown that maternal behavior can be affected by the maternal genotype. When mice are reciprocally crossbred, the maternal care of the offspring is often similar to the genetic background of the maternal phenotype, supporting the idea that the maternal genotype can itself impact the maternal care of its offspring.

Epigenetic Transmission of Maternal Behavior

In addition to the genetic transmission of behavior, the behavioral experience of the subject can be transferred partially through modifications of genetic material passed on to her offspring. Provocative studies by Meaney and colleagues have demonstrated that higher levels of crouching and pup licking–grooming (LG) behaviors in lactating rats are associated with increased expression of ER-α in the MPOA. Furthermore, these behavioral traits are transmitted across generations (Figure 6). The differences in LG behaviors of the females appear to be associated with differences in the methylation of the ER-α1b promoter within the MPOA, such that in low LG mothers the available ER-α in the MPOA is reduced (due to increased DNA methylation of the promoter) and hence lower levels of maternal responsiveness result. What perhaps is most intriguing is that environmentally induced changes in methylation patterns appear to be passed on from one generation to the next in almost a Lamarckian fashion.

Experimental Modifications of Maternal Behavior

Maternal Memory

At the time of parturition, the female rat forms a behavioral memory that is reflected by shortened latencies to respond to foster young weeks following delivery and a brief period of parenting. This memory formation is thought to be associated with increased gene expression and new protein synthesis since treatment with protein synthesis inhibitors such as actinomycin D can interfere with the formation of this

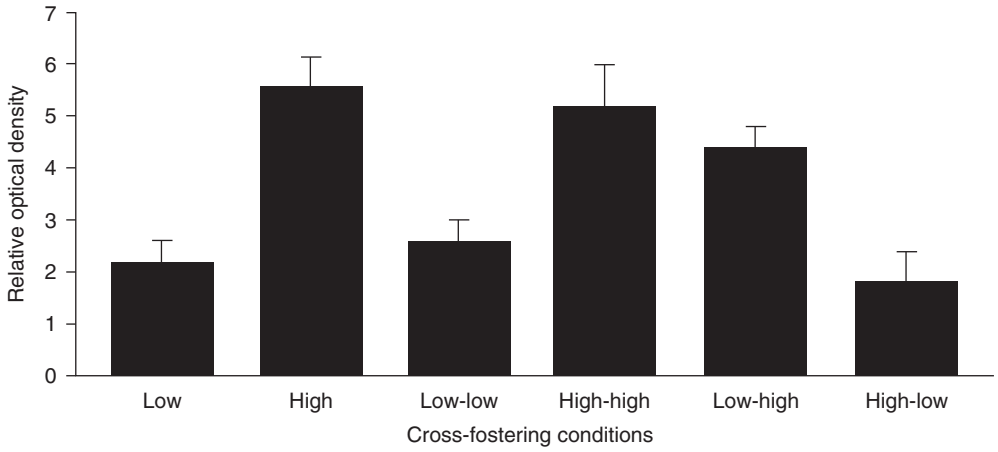

Figure 6 Effects of maternal care of young on their subsequent expression of ERα mRNA in the MPOA as adults. Females cross fostered from a high or low licking–grooming (LG) mother to a low LG mother had equivalent levels of ERα mRNA and had less ER expression in the MPOA than females cross-fostered from a high or low LG to a high LG mother. The original mother is designated on the left and the cross-fostered mother on the right of hyphenated terms (i.e., low-low). Adapted from Champagne FA, Weaver ICG, Diorio J, Dymov S, Szyf M, and Meaney MJ (2006) Maternal care associated with methylation of the estrogen receptor-α1b promoter and estrogen-receptor-α expression in the medial preoptic area of female offspring. *Endocrinology* 147: 2909–2915, with permission.

memory. Maternal memory is also promoted by treatment of a pregnancy hormone regimen consisting of E_2 and P, suggesting that the endocrine changes associated with pregnancy prime the maternal brain to enhance maternal memory at parturition. One key neural region that has been implicated in regulating the formation of maternal memory is the shell region of the NAcc. Lesions of this region prior to parturition prevent the normal formation of maternal memory. Moreover, increased dopaminergic activity within this region appears to participate in this process. Dopamine acts through D2 receptors to mediate this process since treatment with the D2-like antagonist clebopride, but not a D1 antagonist, delayed the reestablishment of maternal behavior 1 week postpartum. The possible involvement of other neurotransmitters, neuropeptides, and hormones in this process remains to be determined. However, it appears that central sensitivities may shift on a long-term basis as a function of behavioral and reproductive experiences, which likely impacts the female's subsequent maternal care and memory.

Reproductive Experience and Maternal Behavior

Studies have established that giving birth and raising young can produce profound long-term changes in neuroendocrine function, hormone secretion, and behavior. Specifically, prior parity in rats as well as women results in reductions in PRL secretion and an increase in central dopaminergic tone. It has further been shown that parity results in a long-term increase in expression of the long form of the PRL receptor within both the MPOA and the arcuate nucleus and

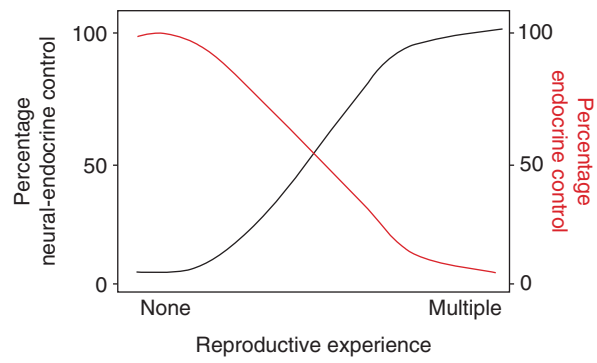

Figure 7 Schematic representation of the effects of reproductive experience on the relative control of maternal behavior by the neural–endocrine and endocrine systems. With increasing reproductive experience, the reliance of the female on endocrine signals declines, whereas the involvement of the neural–endocrine system (i.e., neural prolactin) increases.

an increase in the ability of PRL to activate the PRL-R transduction pathway as measured by increased SOCS-1 and SOCS-3 expression in the MPOA. It thus appears that the neural–endocrine axis within the MPOA undergoes an extended, if not permanent, shift as a function of reproductive experience, shifting from an endocrine to a neural–endocrine basis of control (Figure 7). What does this mean in terms of the neuroendocrine regulation of maternal behavior? It may mean that key hormone-sensitive behavioral centers, such as the MPOA, become more responsive to PRL stimulation of maternal behavior in experienced mothers. If this is true, then the rapid reactivation of maternal behavior in animals displaying

maternal memory may reflect this altered and enhanced central sensitivity to biochemical signals that trigger maternal care.

Finally, the fact that experienced animals display a shift from an endocrine trigger to a neural endocrine trigger, perhaps of neural origin, may shed light on the regulation of maternal care in species that do not rely as heavily on endocrine triggers to stimulate maternal care. For example, women may in fact rely on similar neural–endocrine signals which may mediate aspects of their affective responses, and perhaps when they go amiss, postpartum mood disorders emerge.

See also: Estrus and Menstrual Cycles: Neuroendocrine Control; Gonadal Steroid Actions on Brain; Hormones and Behavior; Neuroendocrinology; Neuroendocrinology of Affective Disorders; Neuroendocrinology of Puberty; Neurosteroids.

Further Reading

Bridges RS (1990) Endocrine regulation of parental behavior in rodents. In: Krasnegor NA and Bridges RS (eds.) *Mammalian Parenting: Biochemical, Neurobiological, and Behavioral Determinants*, pp. 93–117. New York: Oxford University Press.

Bridge RS (ed.) (2008) *Neurobiology of the Parental Brain.* New York: Academic Press

Bridges RS, Numan M, Ronsheim PM, Mann PE, and Lupini CE (1990) Central prolactin infusions stimulate maternal behavior in steroid-treated, nulliparous female rats. *Proceedings of the National Academy of Sciences of the United States of America* 87(20): 8003–8007

Bridges RS and Byrnes EM (1999) Neuroendocrine regulation of maternal behavior. In: Conn PM and Freeman ME (eds.) *Neuroendocrinology in Physiology and Medicine*, pp. 301–315. Totawa, NJ: Humana Press.

Champagne FA, Weaver ICG, Diorio J, Dymov S, Szyf M, and Meaney MJ (2006) Maternal care associated with methylation of the estrogen receptor-α1b promoter and estrogen-receptor-α expression in the medial preoptic area of female offspring. *Endocrinology* 147: 2909–2915

Fleming AS and Li M (2002) Psychobiology of maternal behavior and its early determinants in nonhuman mammals. In: Bornstein MH (ed.) *Biology and Ecology of Parenting,* vol. 2, pp. 141–182. Mahwah, NJ: Erlbaum.

Kinsley CH and Lambert KG (2006) The maternal brain. *Scientific American* 294: 72–79.

Lonstein JS and Gammie SC (2002) Sensory, hormonal, and neural control of maternal aggression in laboratory rodents. *Neuroscience and Biobehavioral Reviews* 26: 869–888.

Lonstein JS and Morrell JI (2007) Neuroendocrinology and neurochemistry of maternal motivation and behavior. In: Blaustein J (ed.) *Behavioral Neurochemistry, Neuroendocrinology, and Molecular Neurobiology: Handbook of Neurochemistry and Molecular Neurobiology,* 3rd edn., pp. 2–51. Berlin: Springer.

Numan M (2006) Hypothalamic neural circuits regulating maternal responsiveness toward infants. *Behavioral and Cognitive Neuroscience Reviews* 5: 163–190.

Numan M and Insel TR (2003) *The Neurobiology of Parental Behavior.* New York: Springer.

Shingo T, Gregg C, Enwere E, et al. (2003) Pregnancy-stimulated neurogenesis in the adult female forebrain mediated by prolactin. *Science* 299: 117–120.

Torner L, Maloumby R, Nava G, Aranda J, Clapp C, and Nuemann ID (2004) *In vivo* release and gene upregulation of brain prolactin response to physiological stimuli. *European Journal of Neuroscience* 19: 1601–1608

Neuroendocrine Peptide Processing

N G Seidah, Clinical Research Institute of Montreal, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The human and mouse genome are predicted to encode approximately 30 000 gene products. The estimated number of primary protein products, however, is 10–30 times higher, mainly due to alternative splicing and posttranslational modifications. The latter can lead to both reversible (e.g., phosphorylation and sulfation) and irreversible (e.g., N-glycosylation, proteolytic cleavage, C-terminal amidation, and N-terminal acetylation) modifications of proteins (Figure 1). Some of the major irreversible posttranslational modifications are afforded by proteases, which either cleave within proteins and peptides, resulting in two or more products, or trim them from the N-terminus (aminopeptidases) or C-terminus (carboxypeptidases), leading to the generation or loss of functional activity. Proteases comprise approximately 1.7% of the predicted proteins of the human genome. Thus, it is estimated that the approximately 500–600 mammalian proteases expressed as active enzymes belong to one of the five classes of eukaryotic proteases, which are classified based on the required nucleophile and on the identity of the catalytic subunit between each family members (Figure 2). Proteases are either of a digestive nature, resulting in protein degradation via multiple cleavages of the protein substrate, or they participate in the limited proteolysis of precursor proteins (proproteins) to generate diversity within the proteome by increasing the number of polypeptide products that can be generated from a single proprotein. The cleavage process can take place either intracellularly at the cell surface or in the extracellular milieu. Within the cell, processing can occur in various subcellular locations and organelles, including the cytosol, nucleus, mitochodria, endoplasmic reticulum (ER), various lamellae of the Golgi apparatus, and within secretory granules, endosomes, and lysosomes. This article focuses on the cleavage of proteins within the so-called secretory pathway. This includes the processing of proteins that enter the ER (via a signal peptide or stop transfer hydrophobic sequences) and that either remain there or are transported to the Golgi apparatus and beyond, including the cell surface, dense core secretory granules, recycling endosomes, and lysosomes. Within the secretory pathway, there exists at least two routes for cell surface secretion – a rapid (minutes) default constitutive secretion from small vesicles fusing with the plasma membrane and a slower (minutes to hours) regulated one allowing the stimulated secretion of dense core secretory granule contents. This article focuses on the role of specific proteolytic enzymes within both pathways.

The Proprotein Convertase Gene Family

The mammalian proprotein convertases (PCs) represent a family of secretory serine proteinases distinct from the trypsin/chymotrypsin subfamily and much more closely related to bacterial subtilisin. Across the phylae, six different subtilase clans can be distinguished, but the nine members of the secretory mammalian PC family belong to only three of these (Figure 2). The seven mammalian convertases related to yeast kexin (Figures 2 and 3) cleave after basic amino acids (basic aa-specific PCs) and are known as PC1/3, PC2, Furin, PC4, PC5/6, PACE4, and PC7. These enzymes cleave after single or paired basic residues predominantly within the motif (Arg/Lys)-X_n-Arg↓, where $n = 0$, 2, 4, or 6, which are variable amino acids separating the two canonical basic residues required for cleavage recognition (Figure 2). The other two convertases (SKI-1/S1P and PCSK9) (Figures 2 and 3) cleave at nonbasic residues within the motifs R-X-(hydrophobic)-Z↓ and (V/I/L) FAQ↓, respectively (Figure 2). All these convertases are initially synthesized as inactive zymogens (proPCs) in the ER. Upon folding, most of these (except for PC2) undergo an initial autocatalytic processing in the ER at an N-terminal site in their prosegment, resulting in a tight binding complex of the inhibitory N-terminal prosegment noncovalently associated with the main protease (Figure 4). Usually, a second autocatalytic zymogen activation event is needed to liberate the protease from its prosegment, thus specifying the location within the cell where the protease can then cut substrates in *trans*. This zymogen activation step can occur in at least five different subcellular organelles: in the *cis*/medial-Golgi (SKI-1); in the early *trans*-Golgi network (TGN), as is the case for PC5/6B; in the last lamella of the TGN that is insensitive to brefeldin A (BFA), as is the case for Furin and PC7; at the cell surface, as seems to be the case for PC5/6A and PACE4; or in dense core secretory granules (PC1/3 and PC2) (Figures 4 and 5).

Analysis of the subcellular distribution of the convertases allowed their classification into at least five distinct groups (Figure 5): First, the membrane-bound convertase SKI-1 is mostly localized in the *cis*/medial-Golgi and is autocatalytically shed, resulting in the secretion of a soluble form of the proteinase.

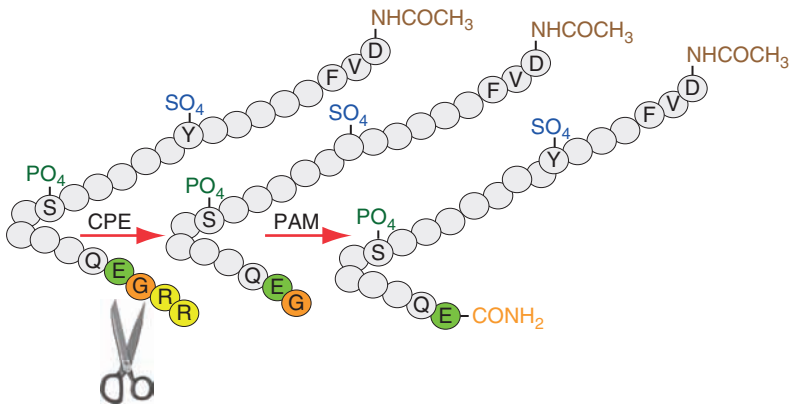

Figure 1 Posttranslational modifications of secretory proteins. On the left is shown part of a hypothetical secretory protein posttranslationally modified by phosphorylation at either a Ser (S) or a Thr, sulfation at Tyr (Y), and N-acetylated at an Asp acid (D) residue. Cleavage of this protein at the Arg–Arg (RR) pair of basic amino acids by one of the basic aa convertases (represented schematically by a pair of scissors) exposes a C-terminal Arg residue, which is then trimmed by a basic aa-specific carboxypeptidase such as CPE (others include CPD and CPX). This exposes a C-terminal Gly (G) residue that is a substrate for the amidation enzyme PAM, which then amidates the penultimate Glu (E) residue.

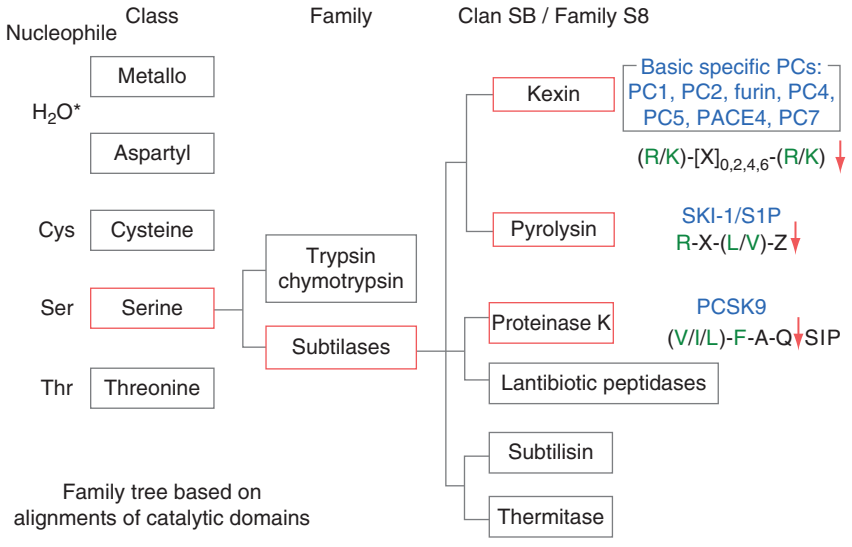

Figure 2 Classification of proteases with emphasis on the secretory PCs. The five classes of proteases are represented on the left, with the nucleophile used with each enzyme class. The emphasis is on serine proteases of the subtilisin type, of which six subfamilies within the clan SB can be identified. The basic amino acid PCs are related to yeast kexin, whereas the convertases SKI-1/S1P and PCSK9 are related to pyrolysin and proteinase K, respectively. The three different motifs generally recognized by the seven basic aa-specific convertases, SKI-1/S1P and PCSK9, are also shown, with the downward arrow indicating the cleavage site. X, any amino acid; Z, any amino acid except Glu. Asp, Val, Pro, and Cys.

The second group includes the membrane-bound convertases Furin, the isoform PC5/6-B, and PC7, which cycle from the cell surface to the TGN through the endosomal pathway. Their subcellular trafficking is regulated by signals in their cytosolic tail. These membrane-bound convertases process most precursors in the Golgi (including the TGN), cell surface, or endosomes. The membrane-bound convertases Furin and PC5/6B (but not PC7) are shed into soluble secreted enzymes. Shedding/cleavage of Furin and PC5/6B is performed by membrane-bound metalloproteases, including ADAMs. Third, the secretory convertases PC5/6A and PACE4 contain a specific Cys-rich domain (Figure 3) that binds to the C-terminal segment of tissue inhibitors of metalloproteases (TIMPs). This complex is retained at the cell surface by heparin sulfate proteoglycans (HSPGs), possibly a site where processing of specific HSPG-bound precursors takes place. Fourth, the regulated convertases PC1/3 and PC2, and sometimes PC5/6A, are sorted to dense core secretory granules and process most polypeptide hormones, which are then secreted following specific

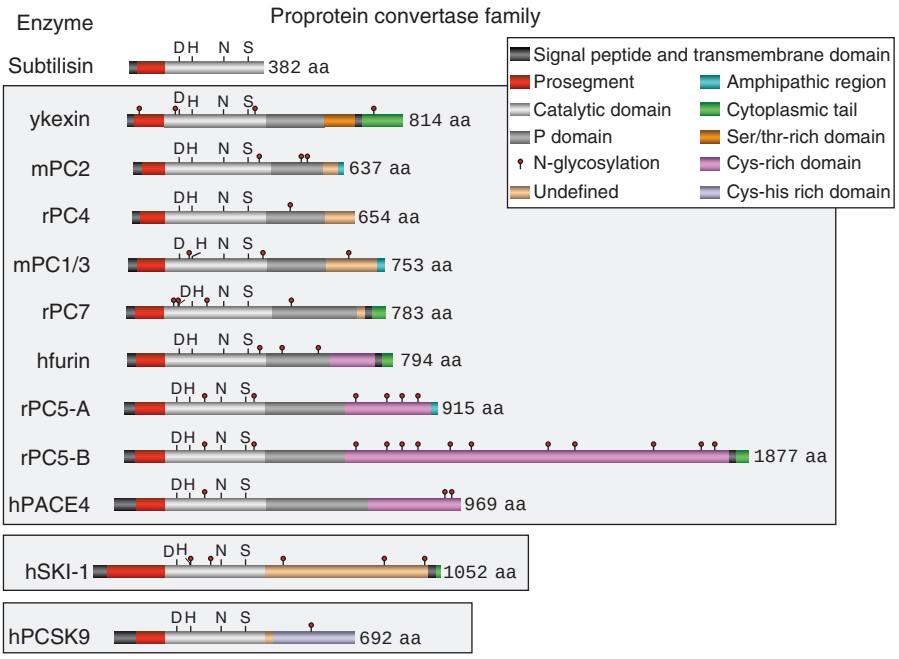

Figure 3 Schematic representation of the primary structure and domains of the PCs. The primary structures and domains of the nine mammalian PCs (h, human; r, rat; and m, mouse) are compared to the original subtilisin and yeast (y) kexin. The basic aa-specific PCs together with yeast kexin, SKI-1, and PCSK9 are individually boxed to emphasize their distinct subclasses. Only PC5 exhibits two validated alternative spliced forms, namely PC5A and PC5B. The membrane-bound PCs include Furin, PC5B, PC7, and SKI-1. The presence of a signal peptide, a prosegment, and catalytic domain is common to all convertases that exhibit the typical catalytic triad residues Asp, His, and Ser and the oxyanion hole Asn. Except for SKI-1 and PCSK9, all the basic aa-specific convertases exhibit a β barrel P domain that apparently stabilizes the catalytic pocket. The C-terminal domain of each convertase contains unique sequences regulating their cellular localization and trafficking. PC5 and PACE4 contain a specific Cys-rich domain, which in combination with TIMPs binds to HSPG at the cell surface. In contrast, PCSK9 exhibits a Cys–His-rich domain that is required for cell surface binding in an LDLR-dependent manner.

stimuli. Finally, the last member of the family, PCSK9, is secreted and then reinternalized into clathrin-coated endosomes, where it exerts its action on LDL-R as an inactive enzyme.

The processing of protein precursors by Furin, PC5/6, PACE4, PC7, and SKI-1 can be inhibited by overexpression of their prosegments, by specific chloromethylketone (CMK) containing membrane-permeable decanoylated (dec-) peptides such as dec-RVKR-CMK or dec-RRLL-CMK, or by variants of the serpin α_1-PDX (Figure 5). These inhibitors have been shown to block metastasis and tumor growth, suggesting that the basic aa-specific Furin-like PCs, and possibly other PCs, should be considered as major targets for the control of tumor progression and cancer metastasis.

Neural and Endocrine Functions of the Proprotein Convertases

Because a number of excellent reviews have been written on the general functions of the proprotein convertases and their targets, we consider here only the implications of the PCs in neuronal functions.

Neuroendocrine Convertases PC1/3 and PC2

PC1/3 and PC2 are the principal activators of prohormones and proneuropeptides within the regulated secretory pathway of neural and endocrine cells. PC1 and PC2 process many precursors into their functional hormones: proopiomelanocortin (POMC; precursor of adrenocorticotropic hormone (ACTH) and β-endorphin), proinsulin (Figure 6), pro-melanin concentrating hormone (MCH), secretogranin II, proenkephalin, prodynorphin, prosomatostatin, chromogranins A and B and PACAP, proenkephalin, and the neurotrophic factor pro-BDNF. The functional activation of the convertase PC2 is regulated by an approximately 30-kDa neuroendocrine PC2-binding protein 7B2 that is coregulated with it. Both proteases are calcium-dependent enzymes that function within the confines of dense core secretory granules (Figure 5), where they exert their maximal activity.

The physiological importance of PC1/3 and PC2 was deduced from studies of the phenotypes of their gene knockout in mice and the discovery of some human patients with defects in PC1/3. In all cases, mice were viable, suggesting that individually the genes

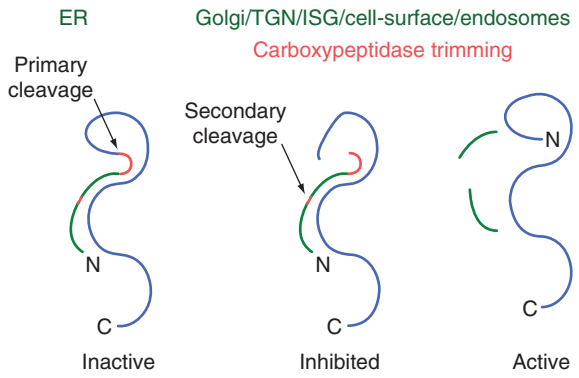

Figure 4 Zymogen activation of the PCs. A PC usually undergoes a first autocatalytic zymogen cleavage in the endoplasmic reticulum (ER) generating an N-terminal inhibitory prosegment that remains noncovalently tightly bound to the enzyme. This bimolecular complex then exits the ER and is directed toward the Golgi apparatus. For SKI-1, a second autocatalytic processing event in the prosegment activates the enzyme in the *cis*/medial-Golgi. Zymogen activation of Furin, PC7, and PC5B occurs in the *trans*-Golgi network (TGN), where the acidic (pH 6.2) and high calcium concentrations (~5–10 mM) favor a second autocatalytic cleavage of the prosegment, thereby liberating the enzyme from its inhibitory prosegment. These enzymes can then cleave *in trans* other substrates. In the case of PC5 and possibly PACE4, the second autocatalytic cleavage may occur either in the TGN or at the cell surface, where these proteinases may then cleave and sometimes inactivate specific proteins. The zymogen activation steps of the regulated endocrine and neural convertases PC1/3 and PC2 occur in the highly acidic (pH ~5.2) immature secretory granules, which explains why they are maximally active only in dense core secretory granules. Finally, the secreted PCSK9 is complexed with its prosegment even after prosegment in recycling endosomes upon reinternalization of the secreted bimolecular complex.

of these convertases are not essential for life. Nevertheless, although PC2 null mice appear normal at birth, they exhibit retarded growth. Analysis of these mice reveals chronic fasting, hypoglycemia, and a deficiency in circulating glucagon. PC2 is known to process various neuroendocrine precursors, and many of these were not fully processed in PC2 null mice, including prosomatostatin, neuronal proCCK, neurotensin, neuromedin N, prodynorphin, proorphanin FQ/nociceptin, and POMC-derived peptides.

Contrary to PC2, PC1/3 gene disruption results in developmental abnormalities. PC1/3 null mice exhibit growth retardation. Adult mutant mice are approximately 60% of the normal size and phenotypically resemble mice that have mutant growth hormone-releasing hormone (GHRH) receptor. Interestingly, insulin growth factor-1 (IGF-1) and GHRH levels are significantly reduced along with pituitary GH mRNA levels, suggesting that this reduction contributes to the growth retardation observed in these mice. Similarly, analysis of several protein precursors known to be processed by PC1/3 revealed that these mice, like PC2 mutant mice, exhibit multiple defects in various hormone precursor processing events. These include the hypothalamic GHRH, pituitary POMC, pancreatic proinsulin, and intestinal proglucagon. In contrast to PC2 null mice, PC1/3 null mice process normally pituitary POMC to adrenocorticotropic hormone (ACTH) (Figure 6) and have normal levels of blood corticosterone. Like PC2 null mice, they also develop hyperproinsulinemia.

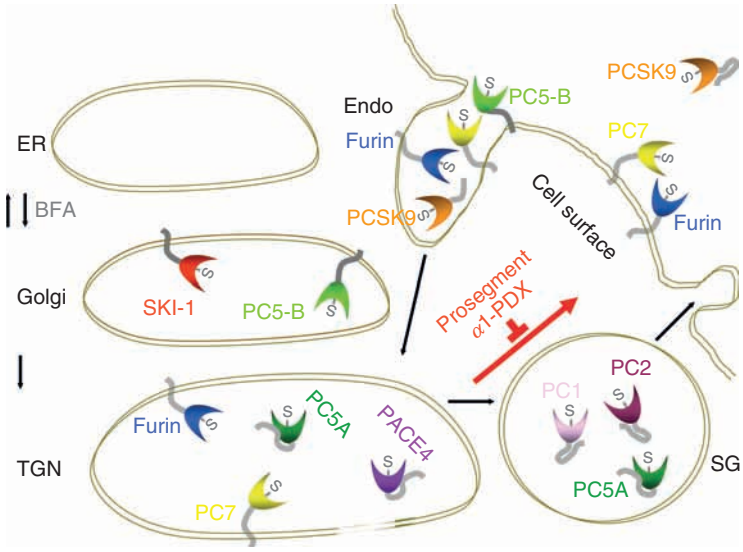

Figure 5 Subcellular localization and sorting of the PCs. The subcellular localization of the color-coded convertases in the Golgi, TGN, endosomes (Endo), cell surface, and dense core secretory granules (SG) is shown. The scheme also shows that the drug brefeldin A (BFA) can cause the reversible retrieval of both SKI-1 and the isoform PC5B into the endoplasmic reticulum (ER). Some of the known inhibitors (⊥) of the basic aa PCs are also shown. Finally, each enzyme exhibits the active site Ser (S) residue within the catalytic subunit (crescent shape).

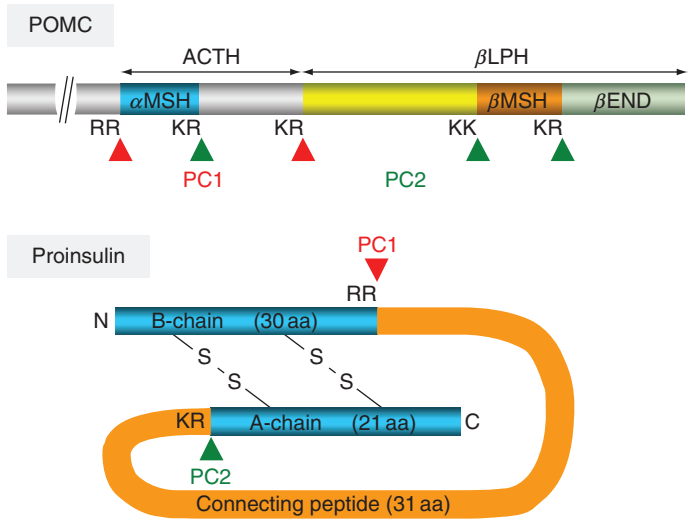

Figure 6 Stepwise processing of POMC and proinsulin by PC1 and PC2. PC1 and PC2 cleave POMC and proinsulin in immature secretory granules at different dibasic sites, represented in this figure. This results in the major production of β endorphin and β-MSH by PC2, ACTH and β-LPH by PC1, and α-MSH and insulin by PC1 + PC2.

Since PC2 is the major convertase that cleaves POMC and proenkephalin to generate the morphinomimetic peptides β-endorphin (Figure 6) and Met- and Leu-enkephalins, respectively, it was important to investigate the role of PC2 in pain perception. Unexpectedly, after a short forced swim in warm water, PC2 null mice were significantly less (rather than more) responsive to the stimuli than wild-type mice, an indication of increased opioid-mediated stress-induced analgesia. The enhanced analgesia in PC2 null mice may be caused by an accumulation of opioid precursor processing intermediates with potent analgesic effects or by loss of antiopioid peptides. Thus, the presence of abnormal cocktails of pain neuropeptides in the brain of PC2 knockout mice is likely to disturb pain perception mechanisms in ways that remain to be fully elucidated.

PC1/3 deficiency in a female patient compound heterozygote for both splicing and nonsynonymous mutations resulted in very low expression of the protein. This subject exhibited neonatal obesity and abnormal glucose homeostasis as well as the presence of other endocrine defects, including very high circulating levels of proinsulin and multiple forms of partially processed POMC (intermediate ACTH precursors), low-serum estradiol, follicle-stimulating hormone, and luteinizing hormone. Another PC1/3 deficiency female subject presented with severe diarrhea, which started on the third postnatal day. Clinical investigations revealed a defect in the absorption of monosaccharides and fat, revealing the role of PC1/3 in the absorptive function of the small intestine. Although the phenotypes of the PC1/3 null mice differ from those observed in these patients (PC1/3 null mice are not obese), the findings confirmed the importance of PC1/3 as a key neuroendocrine convertase. Interestingly, obesity, hyperphagia, and increased metabolic efficiency have been identified in PC1/3 mutant mice exhibiting a homozygote mutation N222D/N222D that results in an approximately 60% decrease in PC1/3 activity, suggesting that it is the dose of PC1/3 and possibly reduced hypothalamic α-MSH (Figure 6) that may define the obesity phenotype.

Ubiquitous Convertases Furin, PC7, and SKI-1/S1P

Furin, a ubiquitous membrane protein, is initially produced as an approximately 104-kDa precursor and is rapidly converted into an active, approximately 98-kDa form. This autocatalytic cleavage, occurring in the ER, is a prerequisite for the exit of mature Furin molecules from the ER to the TGN and cell surface. Furin and PC5 seem to exhibit partial redundancy of their *in vitro* cleavage selectivity of a number of substrates and sensitivity to certain modified serpin inhibitors, such as α_1-PDX or their prosegments (Figure 6). The varied types of constitutively secreted proteins processed by the Furin-like convertases are exemplified in Table 1. In the central nervous system, Furin is responsible for the processing of a number of growth factors, including the neurotrophins pro-nerve growth factor (NGF) and neural cell adhesion and cueing proteins such as L1-CAM and semaphorins.

Inactivation of the *fur* gene (*Pcsk3*) causes embryonic death on approximately embryonic day 11 (E11)

Table 1 Major Furin and Furin-like substrates[a]

Growth factors: TGF-β, IGF-1 and -2, PDGF-A and -B, VEGF-C
Neurotrophins: NGF, BDNF, NT3 and-4
Prohormones: PTH, endothelin, somatostatin, NPY, 7B2, MCH, POMC
Blood coagulation factors: Von Willebrand factor, factor VII, factor IX, factor X
Integrins α chains: α_3, α_4, α_5, α_6, α_8, α_v
Neural molecules: L1, semaphorins, CLAC, BRI3, SorLA, BACE1, N-CAM
Receptors: Insulin-R, IFG1-R, R-PTPμ, Notch, LRP1
Metalloendopeptidases: ADAMs, MMPs, MT-MMPs
Viral envelope glycoproteins: HIV-gp160, RSV-Pr95^{env}, NDV-F, influenza hemagglutinin-HA, CMV-gB, SFV-p62
Bacterial surface proteins: Anthrax and diphtheria toxins

[a]Furin and Furin-like PCs (PC5, PACE4, and PC7) can process a wide variety of secretory precursor proteins as exemplified here. This also includes cell surface proteins originating from opportunistic viral and bacterial pathogens.
BDNF, brain-derived neurotropic factor; IGF, insulin-like growth factor; MCH, melanin concentrating hormone; NGF, nerve growth factor; NPY, neuropeptide Y; NT, neurotrophin; PDGF, platelet-derived growth factor; POMC, proopiomelanocortin; PTH, parathyroid hormone; TGF-β, transforming growth factor β_1; VEGF-C, vascular endothelial growth factor-C.

due to hemodynamic insufficiency and cardiac ventral closure defects. Mutant embryos failed to develop large vessels despite the presence of endothelial cell precursors. Transforming growth factor β_1 has been shown to be efficiently processed by Furin, and inactivation of its gene produces a phenotype similar to that of Furin null embryos. A conditional knockout in liver, with deletion of exon 2 dependent on the Cre expression from the Mx1-cre transgene, resulted in viable Pcsk3flox/flox Tg(Mx1-cre) mice with almost no phenotype. This showed some redundancy with other PCs since some typical Furin substrates were cleaved to a lesser extent.

Different from the other PC null mice, multiple studies on PC7 null mice embryos revealed no apparent abnormal phenotype under resting conditions. This may be explained by the fact that PC7 expression extensively overlaps with that of Furin. Many reports show that PC7 and Furin process the same substrates, such as PDGF-AA, PDGF-BB, vascular endothelial growth factor C, and bone morphogenic protein. Alternatively, the most conserved convertase PC7 may be involved in the processing of a panel of nonessential substrates. Double knockouts of Furin and PC7 may help resolve the issue of the possible critical functions of PC7 during embryogenesis in a Furin null background.

SKI-1 is a key enzyme in the regulation of lipid metabolism and cholesterol homeostasis that cleaves the transcription factors sterol regulatory element binding proteins (SREBP-1 and SREBP-2). The latter are synthesized as precursors harboring two transmembrane domains separated by a short ER luminal loop and comprising N- and C-terminal cytosolic domains. These precursors are cleaved in an SREBP cleavage-activating protein (SCAP) and insulin-induced gene (Insig)-dependent manner. When cellular cholesterol levels are high, the insulin-regulated proteins Insig-1 and/or Insig-2 bind and retain the SCAP–SREBP complex in the ER. When cells are deprived of sterols, Insigs separate, allowing the transport of the SREBP–SCAP complex to the Golgi. Therein, a two-step proteolytic process starting with SKI-1 at the sequence R-X-V-L↓ (Figure 2) and then site 2 protease (S2P) releases the cytosolic N-terminal segments of SREBPs from cell membranes, allowing their translocation to the nucleus (nSREBP), where they activate transcription of more than 35 mRNAs coding for proteins/enzymes required for the biosynthesis and uptake of cholesterol and unsaturated fatty acids, as well as the low-density lipoprotein receptor (LDL-R).

Similar to SREBPs, the ER-anchored type II membrane-bound transcription factor ATF6 plays a major role in the unfolded protein response. Under normal conditions, it is held in the ER by the chaperone BIP, with its N-terminal DNA-binding domain facing the cytosol and its COOH terminus in the ER lumen. Accumulation of improperly folded proteins in the ER, which can be induced by calcium depletion (thapsigargin) or inhibition of N-glycosylation (tunicamycin), leads to an ER stress response resulting in BIP dissociation from proATF6. The latter is then translocated in a SCAP-independent manner to the Golgi, where it is first cleaved by SKI-1 and then by S2P. This releases the cytosolic N-terminal domain, which reaches the nucleus (nATF6) to activate ER stress target genes.

Other type II membrane-bound substrates include at least six CREB-like basic leucine zipper transcription factors. BDNF is a soluble substrate and the study of its processing led to the initial cloning of SKI-1. We also showed that the soluble prosomatostatin is cleaved by SKI-1 to release the N-terminal peptide Antrin.

The essential roles played by SKI-1/S1P are evident from the fact that deletion of its gene results in

embryonic death at the earliest stages of cell division (i.e., from one- to two-cell stage), thus preventing blastocyst formation. The importance of this enzyme in the central nervous system is yet to be unraveled using tissue-specific knockout mice. When this was done in liver, the level of circulating total cholesterol was decreased by 50%, emphasizing its critical control of cholesterol synthesis and uptake. This gene has not been found to be polymorphic, and no single nucleotide polymorphic variation has been identified that could be genetically associated with a known pathology.

Convertases PC5/6 and PACE4

The convertases PC5/6 and PACE4 seem to form a class of their own based on their primary structures and their ability to bind the cell surface via their C-terminal Cys-rich domains, which bind TIMPs and HSPGs and, in many cases, inactivate HSPG-bound proteins such as endothelial and lipoprotein lipases and possibly adhesion molecules. Because in cellular experiments and *in vitro* many of the substrates processed by either enzyme can also be cleaved by Furin and/or PC7, the specific physiological substrates of PC5/6 and PACE4 have not been identified. In the central nervous system, it has been shown that PC5/6 can process the neural adhesion molecule L1, assisting in neuronal repair and migration. Ontogeny and tissue distribution analysis of these convertases showed that PC5/6 expression is detected early during embryonic development, appearing first in extraembryonic tissues. By E9, it is also specifically expressed in cells of the maternal–embryonic junction, where no other convertase is expressed. Data strongly suggest unique tissue-specific functions of PC5 and PACE4. Thus, PC5/6 mRNA was detected only in neuronal cells, whereas PACE4 mRNA was expressed in both neuronal and glial cells. In areas that are rich in neuropeptides, such as cortex, hippocampus, and hypothalamus, mRNA levels of PC5 were high but PACE4 levels were low or undetectable. In regions such as the amygdaloid body and thalamus, distinct but complementary distributions of PC5 and PACE4 mRNAs were observed. The medial habenular and cerebellar Purkinje cells expressed very high levels of PACE4 mRNA.

The knockout of PACE4 and PC5/6 genes in mice resulted in different phenotypes. Thus, whereas the PACE4 knockout results in a 75% viable phenotype with bone morphogenesis defects, that of PC5/6 causes death at birth, due in part to the lack of processing of the TGF-β-like growth factor GDF11 and to the lack of various homeotic transformations. What are the other precursors that need such specific processing events and what is the role of TIMPs and HSPGs in this process are open questions that may be resolved by tissue-specific knockouts and careful analysis of the cell surface proteome of various tissues in the presence or absence of these convertases.

Convertase PCSK9 (Originally Named NARC-1)

PCSK9 was first characterized by our group in 2003. It is highly expressed in adult liver, gut, and kidney. During development, PCSK9 mRNA levels are transiently high in telencephalon (E12–E15), in the rostral extension of the olfactory peduncle, and in cerebellar neurons. Under resting conditions, the expression of PCSK9 in adult brain is restricted to the olfactory peduncle and cerebellum. Overexpression of PCSK9 in telencephalic progenitor cells enhances proliferation and neuronal differentiation. Whether specific signals can regulate the levels of PCSK9 in the central nervous system is yet to be defined. PCSK9 expression in liver hepatocytes is upregulated by SREBP-2 and downregulated by cholesterol. We established the first association between single point mutations in PCSK9 and autosomal dominant hypercholesterolemia in two French families. Later, it was shown that nonsense PCSK9 mutations, likely resulting in a loss of function, are associated with hypocholesterolemia in approximately 2% of black subjects. Accordingly, mutations associated with hypercholesterolemia result in a gain of function of PCSK9 that triggers the degradation of LDL-R in acidic compartments, likely endosomes. By an unknown mechanism, high levels of PCSK9 lead to a faster rate of degradation of cell surface LDL-R, resulting in increased circulating LDL-cholesterol as LDL uptake in hepatocytes by LDL-R is diminished. In agreement, *Pcsk9* null mice exhibit increased LDL-R protein, but not mRNA, and a twofold decrease in circulating cholesterol, whereas mice overexpressing PCSK9 following recombinant adenoviral infections or in liver transgenic mice exhibit high levels of circulating cholesterol. A 32-year-old educated female aerobic instructor with two compound heterozygote mutations leading to complete loss of PCSK9 expression was recently identified. Although healthy, fertile, and normotensive, with normal liver and renal functions, her LDL-cholesterol is remarkably low ($14\,\mathrm{mg\,dl^{-1}}$). This and the observation that loss-of-function nonsense mutations could lead to 88% reduction in the risk of developing cardiovascular artery disease (CAD) during a 15-year period, indicate that inhibition of PCSK9 may represent a safe and effective strategy for the primary prevention of CAD. The role of PCSK9 in the brain is less clear, and it may only become apparent under stressful or pathogenic conditions such as cerebral stroke.

Summary

The nine-member family of PCs comprises seven basic amino acid-specific subtilisin-like serine proteinases related to yeast kexin, known as PC1/3, PC2, Furin, PC4, PC5/6, PACE4, and PC7, and two other subtilases that cleave at nonbasic residues called SKI-1/S1P and PCSK9. Except for testicular PC4, all other convertases are expressed in brain and play a critical role in various neuronal functions, including the production of diverse neuropeptides as well as neural growth factors and receptors, the regulation of cellular adhesion/migration, cholesterol and fatty acid homeostasis and repair mechanisms, and growth/differentiation of progenitor cells. Some of these convertases process proteins implicated in neuropathologies such as Alzheimer's disease β-secretase (BACE1), neural growth such as NGF, as well as adhesion molecules such as L1. N-CAM and N-cadherin, which are important for cancer malignancies and neuronal regeneration and repair following injury. The development of specific inhibitors/modulators of convertases may find applications in the control of some of these pathologies.

See also: Neuroendocrinology; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Endocrine Cells.

Further Reading

Attie AD and Seidah NG (2005) Dual regulation of the LDL receptor – some clarity and new questions. *Cell Metabolism* 1: 290–292.

Benjannet S, Rhainds D, Essalmani R, et al. (2004) NARC-1/PCSK9 and its natural mutants: Zymogen cleavage and effects on the low density lipoprotein (LDL) receptor and LDL cholesterol. *Journal of Biological Chemistry* 279: 48865–48875.

Benjannet S, Rhainds D, Hamelin J, Nassoury N, and Seidah NG (2006) The proprotein convertase (PC) PCSK9 is inactivated by Furin and/or PC5/6A: Functional consequences of natural mutations and post-translational modifications. *Journal of Biological Chemistry* 281: 30561–30572.

Cohen J, Pertsemlidis A, Kotowski IK, et al. (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nature Genetics* 37: 161–165.

Essalmani R, Zaid A, Marcinkiewicz J, et al. (2008) *In vivo* functions of the proprotein convertase PC5/6 during mouse development: Gdf11 is a major substrate. *Proceeding of the National Academy of Sciences of the United States of America* (in press).

Kotowski IK, Pertsemlidis A, Luke A, et al. (2006) A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *American Journal of Human Genetics* 78: 410–422.

Mayer G, Hamelin J, Asselin M-C, et al. (2008) The regulated cell-surface zymogen activation of the proprotein convertase PC5A directs the processing of its secretory substrates. *Journal of Biological Chemistry* 283: 2373–2384.

Maxwell KN, Fisher EA, and Breslow JL (2005) Overexpression of PCSK9 accelerates the degradation of the LDLR in a post-endoplasmic reticulum compartment. *Proceedings of the National Academy of Sciences of the United States of America* 102: 2069–2074.

Pullikotil P, Vincent M, Nichol ST, et al. (2004) Development of protein-based inhibitors of the proprotein of convertase SKI-1/S1P: Processing of SREBP-2, ATF6, and a viral glycoprotein. *Journal of Biological Chemistry* 279: 17338–17347.

Rashid S, Curtis DE, Garuti R, et al. (2005) Decreased plasma cholesterol and hypersensitivity to statins in mice lacking Pcsk9. *Proceedings of the National Academy of Sciences of the United States of America* 102: 5374–5379.

Seidah NG (2006) Unexpected similarity between the cytosolic West Nile virus NS3 and the secretory Furin-like serine proteinases. *Biochemistry Journal* 393: e1–e3.

Seidah NG, Benjannet S, Hamelin J, et al. (1999) The subtilisin/kexin family of precursor convertases. Emphasis on PC1, PC2/7B2, POMC and the novel enzyme SKI-1. *Annals of the New York Academy of Sciences* 885: 57–74.

Seidah NG and Chretien M (1999) Proprotein and prohormone convertases: A family of subtilases generating diverse bioactive polypeptides. *Brain Research* 848: 45–62.

Seidah NG, Khatib AM, and Prat A (2006) The proprotein convertases and their implication in sterol and/or lipid metabolism. *Biological Chemistry* 387: 871–877.

Seidah NG, Mayer G, Zaid A, et al. (2008) The activation and physiological functions of the proprotein convertases. *International Journal of Biochemistry and Cell Biology* (in press).

Seidah NG and Prat A (2002) Precursor convertases in the secretory pathway, cytosol and extracellular milieu. *Essays in Biochemistry* 38: 79–94.

Steiner DF (1998) The proprotein convertases. *Current Opinion in Chemical Biology* 2: 31–39.

Taylor NA, Van de Ven WJ, and Creemers JW (2003) Curbing activation: Proprotein convertases in homeostasis and pathology. *FASEB Journal* 17: 1215–1227.

Thomas G (2002) Furin at the cutting edge: From protein traffic to embryogenesis and disease. *Nature Reviews Molecular Cell Biology* 3: 753–766.

Zaid A, Roubtsova A, Essalmani R, et al. (2008) PCSK9: Hepatocyte-specific LDL receptor degradation and critical role in liver regeneration. *Hepatology* (in press).

Zhao Z, Tuakli-Wosornu Y, Lagace TA, et al. (2006) Molecular characterization of loss-of-function mutations in PCSK9 and identification of a compound heterozygote. *American Journal of Human Genetics* 79: 514–523.

Neuroendocrinology

L A Frohman, University of Illinois at Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Neuroendocrinology encompasses the brain's regulation of and interaction with the endocrine system. The neuroendocrine system can be subdivided by anatomic, physiologic, and biochemical characteristics. The major anatomic components are the hypothalamus and extrahypothalamic regions, the hypothalamic–pituitary stalk and portal vessels, and the pituitary. Once designated as the master gland, the pituitary is to a large extent subservient to the neuroendocrine apparatus of the medial basal hypothalamus. The pituitary is divided anatomically into three regions that are embryologically distinct and separately regulated.

The anterior pituitary (adenohypophysis) secretes six well-characterized hormones important for distant endocrine regulation: adrenocorticotropin (ACTH) stimulates the adrenal cortex; thyroid-stimulating hormone (TSH), or thyrotropin activates the thyroid gland; luteinizing hormone (LH) and follicle-stimulating hormone (FSH), both gonadotropins, act on the ovaries and testes; prolactin stimulates milk production in the breast; and growth hormone (GH) or somatotropin exerts metabolic effects essential for attainment of full somatic growth and maintenance of normal metabolic homeostasis. These hormones are produced by individual cell types, except for LH and FSH, which are produced in the same cell. The major cell types of the anterior pituitary are called corticotropes, thyrotropes, gonadotropes, lactotropes, and somatotropes, and can be distinguished from one another by immunocytochemistry and electron microscopy. An additional cell type, the somatomammotrope, produces both GH and prolactin, and another, the folliculostellate cell, produces several substances, including vascular endothelial growth factor, interleukins, and lipocortin. The adenohypophysis originates embryologically from the joining of Rathke's pouch, an outpouching of the primitive gut, with portions of the ventral placodes of the neural ridges, which also give rise to the hypothalamus.

The posterior pituitary (or pars nervosa) is derived from primitive hypothalamic tissue. As it evaginates from the base of the brain, it forms the infundibulum of the third ventricle, the median eminence of the hypothalamus, and the pituitary stalk. The neural lobe contains the axons of hypothalamic neurons located in the paired supraoptic and paraventricular nuclei and forms the site of terminal storage in cells, termed pituicytes, for the neurohypophysial hormones vasopressin and oxytocin. Synthesized in the neurons of the supraoptic and paraventricular nuclei, vasopressin and oxytocin are transported axonally through the pituitary stalk to the posterior pituitary and are released by axonal depolarization. Such nerve cells – capable of synthesis, transport, storage, and depolarization-induced hormone release – are called neurosecretory neurons.

The intermediate lobe of the pituitary, while prominent in lower mammals, is rudimentary in primates. It is the source of β-endorphin and MSH (melanin-stimulating hormone) in some species, but has little known function in humans. It is innervated in rodents primarily by dopaminergic nerve fibers arising in the hypothalamus.

The pituitary is connected to the hypothalamus by a stalk that permits two types of communication: vascular and neuronal. The vascular connections are essential for the regulation of anterior pituitary function by the hypothalamus. The anatomical features of the vascular supply are unique. Shortly after exiting from the cavernous sinus, the internal carotid arteries provide bilateral arterial branches, the superior hypophyseal arteries which penetrate the median eminence of the hypothalamus, dividing into a capillary network that extends into the outer layer of the basal hypothalamus. The capillaries are juxtaposed on the axonal terminals of the hypothalamic hormone-secreting neurons. The fenestrated endothelial cells, unlike other capillaries of the brain, permit diffusion of the neurohormones from the nerve terminals into the portal capillaries. In most mammals, the capillaries of the median eminence coalesce to form six to ten portal veins that traverse the pituitary stalk from the brain downward, and then divide to form the vascular sinusoids that bathe the secretory cells of the anterior pituitary.

Neurons located in the anterior and medial basal hypothalamus synthesize hormones and transport them axonally to the vicinity of the capillaries of the median eminence, from which they freely diffuse into the vessels and are carried to the anterior pituitary. The hypothalamic hormones, all but one of which are peptides, are frequently called releasing or inhibiting hormones (or factors) and are relatively specific for each of the anterior pituitary hormones (Figure 1). Their isolation and characterization were initially accomplished by quantitative purification of hundreds of thousands of sheep or pig hypothalamic extracts. The work of Schally and Guillemin and their collaborators elucidated the structure of thyrotropin-releasing hormone (TRH), gonadotropin-releasing hormone (GnRH), and somatostatin (somatotropin release-inhibiting

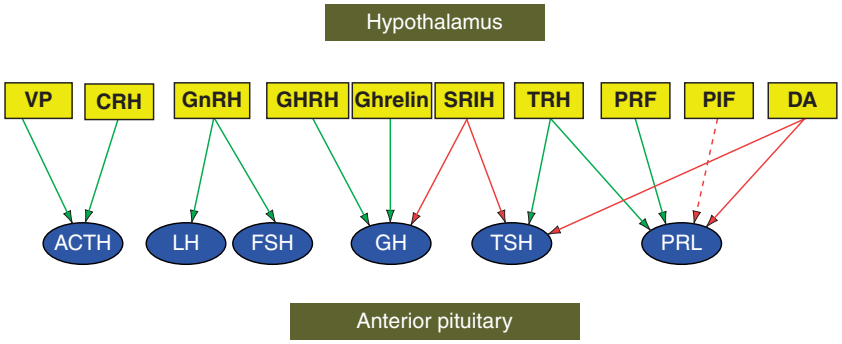

Figure 1 Hypothalamic hormone–pituitary hormone relationships. Symbols in interrupted lines have not yet been fully characterized. Green arrows indicate stimulatory effects; red arrows indicate inhibitory effects. CRH, corticotropin-releasing hormone; VP, vasopressin; GnRH, gonadotropin-releasing hormone; GHRH, growth hormone-releasing hormone; SRIH, somatostatin; TRH, thyrotropin-releasing hormone; DA, dopamine; PIF, prolactin-inhibiting factor; PRF, prolactin-releasing factor; ACTH, adrenocorticotropin; LH, luteinizing hormone; FSH, follicle-stimulating hormone; GH, growth hormone; TSH, thyrotropin; PRL, prolactin.

hormone, SRIH) for which they were awarded the Nobel Prize in 1971. Corticotropin-releasing hormone (CRH) was isolated from rat hypothalamus and growth hormone-releasing hormone (GHRH) from pancreatic tumors and subsequently human hypothalamus one decade later.

Neuroanatomical studies, using immunocytochemical and *in situ* hybridization techniques, have demonstrated distinct hypothalamic localization within neurons for each of the peptides. TRH is found primarily in the anterior periventricular region, with some cells interposed with the cells of the periventricular nucleus. GnRH is found in arcuate nucleus of primates and in the medial preoptic area of rodents. Somatostatin is located in anterior periventricular cells, and to a lesser extent in the arcuate nucleus. CRH is located in periventricular cells and, as with TRH, in cells within the magnocellular paraventricular nuclei. GHRH is found almost exclusively in the arcuate nucleus. Prolactin secretion is regulated principally by an inhibitory hypothalamic influence. The primary prolactin inhibitory factor (PIF) is dopamine, found also in arcuate neurons. There is now considerable physiologic evidence for the presence of one or more prolactin-releasing factors (PRFs). Vasoactive intestinal peptide is believed to represent one such PRF, and there is evidence for another PRF in the neurointermediate lobe. The hypothalamic hormones are not entirely cell specific, because TRH stimulates both TSH and prolactin release and somatostatin inhibits both GH and TSH release.

The actions of the releasing hormones on pituitary cells are mediated by cell receptors. Receptors for TRH, GnRH, somatostatin, CRH, GHRH, and dopamine have been cloned. GHRH and ACTH activate the adenylyl cyclase-protein kinase A system and increase intracellular Ca^{2+} within somatotropes and corticotropes, respectively; TRH and GnRH act primarily through the phosphoinositol-protein kinase C system in thyrotropes and gonadotropes, respectively; and somatostatin inhibits adenylyl cyclase in somatotropes. The actions of the releasing hormones are rapid; secretion of hormones can be detected within minutes after administration both *in vitro* and *in vivo*. In addition, at least one hypothalamic hormone (GHRH) exhibits independent effects on pituitary hormone (GH) synthesis and mitogenesis in somatotropes.

The discipline of neuroendocrinology, broadly defined, encompasses the mechanisms, processes, and physiologic responses that coordinate and integrate the internal hormonal milieu. These responses are necessarily complex and diverse in the intact animal. All pituitary hormones are secreted dynamically in episodic bursts or pulses. Each hormone has its own characteristic pattern of secretion, during short periods (ultradian), on a daily basis (circadian), and during more prolonged periods (seasonal). These complex patterns of secretion are generated in the brain and transduced into appropriate hormonal signals by the neurosecretory cells of the hypothalamus. Hormone secretion is determined by numerous behavioral events (e.g., stress, sleep, and emotions) and metabolic events. For example, vasopressin release is regulated by osmotic changes in the circulating blood, volume changes induced by hemorrhage, and stress. GH secretion is stimulated by nutrient ingestion (protein), stress, slow-wave sleep, and pharmacological agents that activate β-adrenergic, dopaminergic, or serotonergic neurotransmitter systems.

The hypothalamus is situated in anatomical and functional proximity to the limbic system. Neural structures important in regulating neuroendocrine responses include the amygdala, hippocampus, septum, and limbic midbrain. In the latter, the biogenic amine systems, which originate in the brain

stem and include the noradrenergic systems arising in the locus coeruleus and other nuclei and the serotonergic system located in the raphe nuclei, are of particular importance.

Recent developments have considerably extended concepts of the boundaries of neuroendocrinology. In addition to the releasing hormones and the influence of biogenic amines, many other neurotransmitters and neuropeptides can act to modify neuroendocrine responses. Histamine, γ-aminobutyric acid (GABA), glutamate, and acetylcholine each influence releasing hormone secretion. The opioid peptides are involved in the secretion of prolactin, GH, and the gonadotropins, and nitric oxide has been implicated in the regulation of several releasing hormones, particularly GHRH. Disorders of gonadotropin release caused by intensive exercise (e.g., ballet, prolonged athletic activity) or by anorexia nervosa and other forms of caloric deprivation or malnutrition may be attributable, in part, to overactivity of central opioid systems. A number of other neuropeptides including substance P, neurotensin, and bombesin are found in abundance in the hypothalamus and are localized in nerve terminals in the median eminence. Their physiologic role in neuroendocrine regulation remains to be fully characterized.

Considerable interest has recently developed in the role of peripheral peptides with actions on hypothalamic feeding regulation, such as leptin, and hypothalamic peptides with similar actions, such as the orexins and neuropeptide Y. In addition to their effects on metabolic homeostasis, these peptides influence the secretion of several pituitary hormones, such as GH and the gonadotropins, through hypothalamic-mediated mechanisms. These findings have provided evidence for the mediation of the effects of nutrition on reproductive function. For example, kisspeptin, a recently identified neuropeptide with profound stimulatory effects on GnRH secretion through its GPR-54 receptor, appears to act as an intermediary between leptin and GnRH neurons. Another peptide, ghrelin, is found primarily in the stomach, although it is also present in the hypothalamus. Its receptors are present in both hypothalamus and pituitary, and the peptide has potent effects on GH secretion through a protein kinase C pathway, thereby explaining the profound effects of alterations in nutritional status and GH secretion. It is of interest that these effects may be mediated through changes not only in peptide concentrations but also in their receptors.

Finally, considerable interest has emerged in the interaction of the immune and neuroendocrine systems (neuroimmunoendocrinology). Interleukins 1 (IL-1) and 6 (IL-6), as well as other cytokines, have been implicated in the release of SRIH, GH, and ACTH and may serve as central nervous system peptide modulators of the stress response. In addition, they may function as regulators of the pathophysiologic responses to stimuli that provoke immunologic responses.

See also: Gonadotropin, Neural and Hormonal Control; Gonadotropin-Releasing Hormone: GnRH-1 System; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Hypothalamo–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH; Immune System–Neuroendocrine Interactions; Neuroendocrine Aging: Pituitary-Adrenal Axis; Neuroendocrine Control of Energy Balance (Central Circuits/Mechanisms); Neuroendocrine Control: Maternal Behavior.

Further Reading

Chan JL and Mantzoros CS (2005) Role of leptin in energy-deprivation states: Normal human physiology and clinical implications for hypothalamic amenorrhoea and anorexia nervosa. *Lancet* 366: 74–85.

Chrousos GP (1995) The hypothalamic–pituitary–adrenal axis and immune-mediated inflammation. *New England Journal of Medicine* 332: 1351–1362.

de la Iglesia HO and Schwartz WJ (2006) Minireview: Timely ovulation: Circadian regulation of the female hypothalamo-pituitary-gonadal axis. *Endocrinology* 147: 1148–1153.

Dungan HM, Clifton DK, and Steiner RA (2006) Minireview: Kisspeptin neurons as central processors in the regulation of gonadotropin-releasing hormone secretion. *Endocrinology* 147: 1154–1158.

Fernald RD and White RB (1999) Gonadotropin-releasing hormone genes: Phylogeny, structure, and functions. *Frontiers in Neuroendocrinology* 20: 224–240.

Frohman LA and Kineman RD (1998) Growth hormone-releasing hormone: Discovery, regulation, and actions. In: Kostyo J (ed.) *Handbook of Physiology: Hormonal Control of Growth*, pp. 189–221. New York: Oxford University Press.

Gatford KL, Egan AR, Clarke IL, and Owens PC (1998) Sexual dimorphism of the somatotrophic axis. *Journal of Endocrinology* 157: 373–389.

Gaylinn BD (1999) Molecular and cell biology of the growth hormone-releasing hormone receptor. *Growth Hormone and IGF Research* 9: 37–44.

Hinuma S, Habata Y, Fujii R, et al. (1998) A prolactin-releasing peptide in the brain. *Nature* 393: 272–276.

Howard AD, Feighner SD, Cully DF, et al. (1996) A receptor in pituitary and hypothalamus that functions in growth hormone release. *Science* 273: 974–977.

Hoyle CH (1999) Neuropeptide families and their receptors: Evolutionary perspectives. *Brain Research* 848: 1–25.

Johnson AK and Thunhorst RL (1997) The neuroendocrinology of thirst and salt appetite: Visceral sensory signals and mechanisms of central integration. *Frontiers in Neuroendocrinology* 18: 292–353.

Kakar SS (1997) Molecular structure of the human gonadotropin-releasing hormone receptor gene. *European Journal of Endocrinology* 137: 183–192.

Kojima M, Hosoda H, Matsuo H, and Kangawa K (2001) Ghrelin: Discovery of the natural endogenous ligand for the growth hormone secretagogue receptor. *Trends in Endocrinology and Metabolism* 12: 118–122.

Lechan RM and Fekete C (2006) The TRH neuron: A hypothalamic integrator of energy metabolism. *Progress in Brain Research* 153C: 209–235.

López FJ, Merchenthaler IJ, Moretto M, and Negro-Vilar A (1998) Modulating mechanisms of neuroendocrine cell activity: The LHRH pulse generator. *Cellular and Molecular Neurobiology* 18: 125–146.

Patel YC (1999) Somatostatin and its receptor family. *Frontiers in Neuroendocrinology* 20: 157–198.

Rexford S, Ahima RS, Saper CB, et al. (2000) Leptin regulation of neuroendocrine systems. *Frontiers in Neuroendocrinology* 21: 263–307.

Rosenfeld MG, Bach I, Erkman L, et al. (1996) Transcriptional control of cell phenotypes in the neuroendocrine system. *Recent Progress in Hormone Research* 51: 217–238.

Rostène WH and Alexander MJ (1997) Neurotensin and neuroendocrine regulation. *Frontiers in Neuroendocrinology* 18: 115–173.

Sakurai T, Amemiya A, Ishii M, et al. (1998) Orexins and orexin receptors: A family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* 92: 573–585.

Schwartz MW and Seeley RJ (1997) Neuroendocrine responses to starvation and weight loss. *New England Journal of Medicine* 336: 1802–1811.

Sternberg EM (2001) Neuroendocrine regulation of autoimmune/inflammatory disease. *Journal of Endocrinology* 169: 429–435.

Turek FW (1998) Circadian rhythms. *Hormone Research* 49: 109–113.

Wade GN and Jones JE (2004) Neuroendocrinology of nutritional infertility. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology* 287: R1277–R1296.

Neuroendocrinology of Affective Disorders

D A Gutman and C B Nemeroff, Emory University School of Medicine, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Our understanding of the neurobiology of depression and other major psychiatric disorders has been advanced to a considerable extent by the systematic analysis of various circulating hormones in patients and normal comparison subjects. The occurrence of prominent psychiatric symptoms in patients with well-documented endocrine disorders, including Cushing's disease and primary hypothyroidism, provided the original rationale for exploring the connection between endocrine function and both affective and cognitive function. Work during the past 50 years has highlighted the complex interplay between endocrine gland secretions and the central nervous system (CNS).

The neuroendocrine control systems are generally organized in a hierarchical manner and are composed of three major components (Figure 1) that are briefly outlined here. At the zenith of each is a collection of neurons in the hypothalamus that integrate signals from higher order centers in the brain. These neurons synthesize and release factors that either promote or inhibit the release of anterior pituitary hormones, the hypothalamic hypophysiotropic hormones also known as release or release-inhibiting factors. These peptide hormones, as summarized in Table 1, are synthesized in the neuronal perikarya by transcription of the DNA sequence for the peptide prohormone. After translation in the endoplasmic reticulum, these prohormones are processed during axonal transport and packaged into vesicles destined for the nerve terminals. These now biologically active peptides are then released, following appropriate physiological stimuli or as part of a circadian cycle, from the most ventral portion of the hypothalamus termed the median eminence. These peptides, secreted directly into the hypothalamo-hypophyseal portal vessels, are transported humorally in this collection of blood vessels that connects the hypothalamus (which is contained inside the CNS) and the anterior pituitary gland (Figure 2). These peptides arrive in high concentration in the sinusoids of the anterior pituitary (adenohypophysis), where they bind to specific membrane receptors on their targets, the pituitary trophic hormone-producing cells. Activation of their receptors promotes (or inhibits) the release of the anterior pituitary hormones into the systemic circulation. The increase or decrease in the plasma concentrations of these pituitary hormones produces a corresponding increase or decrease in the plasma concentrations of the end organ hormones. The hormones of the end organ axes, such as gonadal and adrenal steroids or thyroid hormones, feedback on both pituitary and hypothalamic cells to reduce further release of trophic hormones or releasing factors, respectively, often referred to as 'long-loop' negative feedback. Short-loop negative feedback circuits have also been identified in which pituitary hormones directly feedback upon hypothalamic neurons to prevent further release of hypothalamic releasing factors.

Disturbances in the feedback regulation of the hypothalamic–pituitary–end organ axes are of considerable interest in psychiatry. As noted previously, the common occurrence of psychiatric symptoms in many primary endocrine disorders, such as hypothyroidism and Cushing's syndrome, served as an impetus for investigation into the regulation of neuroendocrine systems in common and severe psychiatric disease states such as depression, schizophrenia, and bipolar disorder. Thus, a large part of psychoneuroendocrinology has focused on identifying changes in basal levels of hypothalamic, pituitary, and end organ hormones in patients with psychiatric disorders.

Corticotropin-Releasing Factor in Depression and Anxiety Disorders

Basic Biology

The hypothalamic–pituitary–adrenal (HPA) axis is among the most intensely studied systems in psychiatry. At the pinnacle of this system is the neuropeptide corticotropin-releasing factor (CRF), also known as corticotropin-releasing hormone, a 41-amino-acid peptide. From a classical neuroendocrine standpoint, CRF is the master hormone controlling the activity of the HPA axis (see Figure 3).

CRF is primarily synthesized in parvocellular neurons located in the paraventricular nucleus (PVN) of the hypothalamus and projects to the median eminence. CRF neurons in the PVN (Figure 1) receive input from a variety of brain regions, including the amygdala, bed nucleus of the stria terminalis, and other brain stem nuclei. In response to stress, the CRF PVN neurons are activated, releasing CRF from the median eminence into the hypothalamo-hypophyseal portal system, where it is transported to the anterior pituitary activating CRF receptors on corticotrophs. In addition, CRF promotes the synthesis of proopiomelanocortin

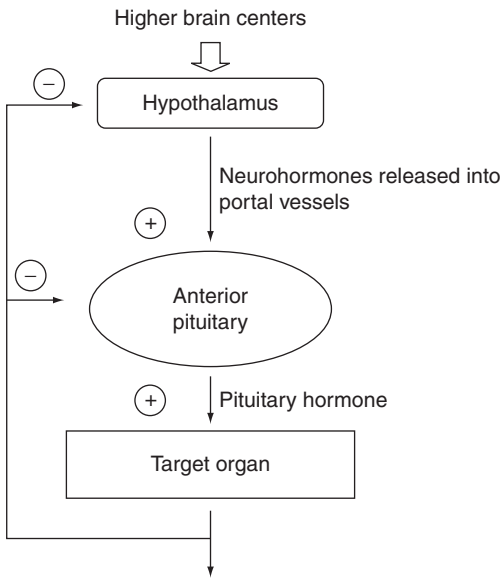

Figure 1 Overview of the common organizational motif of the neuroendocrine axis. The neurosecretion of hypothalamic factors into hypophyseal portal vessels is regulated by a set point of activity from higher brain centers. Neurohormones released from the hypothalamus into hypophyseal portal vessels in turn stimulate cells in the pituitary. These adenohypophyseal hormones then regulate the hormone output from the end organ. The end organ then exerts negative feedback effects at the pituitary and hypothalamus to prevent further neurohormone and pituitary hormone release via 'long-loop' negative feedback. Short-loop negative feedback may also occur where pituitary hormones feed back directly on hypothalamic neurons to prevent further neurohormone release. Adapted from Gutman DA and Nemeroff CB (2002) Neuroendocrinology. In: D'Haenen H, Den Boer JA, and Willner P (eds.) *Biological Psychiatry*, p. 98. London: Wiley.

Table 1 Major releasing factors and their hormonal targets

Neurohormone/releasing factor	*Hormone stimulated*
Corticotropin-releasing factor (CRF)	Adrenocorticotropic hormone (ACTH)
Thyrotropin-releasing hormone (TRH)	Thyroid-stimulating hormone
Gonadotropin-releasing hormone (GnRH)	Follicle-stimulating hormone (FSH)
Somatostatin (SRIF)	(Inhibits) growth hormone (GH)
Growth hormone-releasing hormone (GHRH)	GH
Arginine vasopressin (AVP)	ACTH, prolactin
Oxytocin	Prolactin

Adapted from Gutman DA and Nemeroff CB (2002) Neuroendocrinology. In: D'Haenen H, Den Boer JA, and Willner P (eds.) *Biological Psychiatry*, p. 98. London: Wiley.

and the release of its posttranslation products, adrenocorticotropic hormone (ACTH), β-endorphin, and others. ACTH released from the anterior pituitary into the general circulation then stimulates the production and release of cortisol from the adrenal cortex. (For a general schematic of HPA axis regulation, see Figure 2.)

HPA Axis Abnormalities in Depression

Evidence linking HPA axis abnormalities to psychiatric symptoms dates back more than 100 years. The relatively frequent occurrence of depression and other psychiatric symptoms in both Cushing's and Addison's diseases, associated with excess and deficient levels of circulating cortisol, respectively, served as an impetus for researchers to scrutinize HPA axis function in patients with primary psychiatric disorders. Some of the earliest studies demonstrated abnormalities in glucocorticoid function in depressed patients, including elevated plasma and cerebrospinal fluid (CSF) cortisol concentrations, increased 24 h urinary free cortisol concentrations, and increased levels of cortisol metabolites in urine (Table 2). The next major advance occurred with the discovery of the neuropeptide CRF in 1981, the primary physiological secretagogue of ACTH, and ultimately cortisol.

A number of tests have been developed to assess HPA axis activity, and one of the most common is the dexamethasone suppression test (DST). In this standardized test, 1 mg of dexamethasone is administered orally at 11 p.m. and blood samples are obtained at various times on the following day (e.g., 8 a.m., 4 p.m., and 11 p.m.) and cortisol concentrations measured. Dexamethasone is a synthetic glucocorticoid and, similar to cortisol, it suppresses ACTH secretion, and consequently cortisol release, in healthy volunteers. Nonsuppression of plasma glucocorticoid concentrations postdexamethasone is relatively common in depression, at much higher rates than in normal volunteers. The rate of cortisol nonsuppression after dexamethasone administration generally correlates with the severity of depression; in fact, nearly all patients with major depression with psychotic features exhibit DST nonsuppression. Although this test is used primarily in research labs and not routine clinical practice, the overwhelming conclusion is that a sizeable percentage of depressed patients exhibit HPA axis hyperactivity. Unfortunately, the specificity of the test for diagnostic purposes is unacceptably low, with nondepressed patients with eating disorders, anxiety disorders, and dementia, among others, exhibiting false positives. Moreover, we have suggested that HPA axis hyperactivity may be more related to early life trauma than depression.

A CRF stimulation test has also been developed to assess HPA axis activity. In this paradigm, CRF is administered intravenously (usually in a dose of

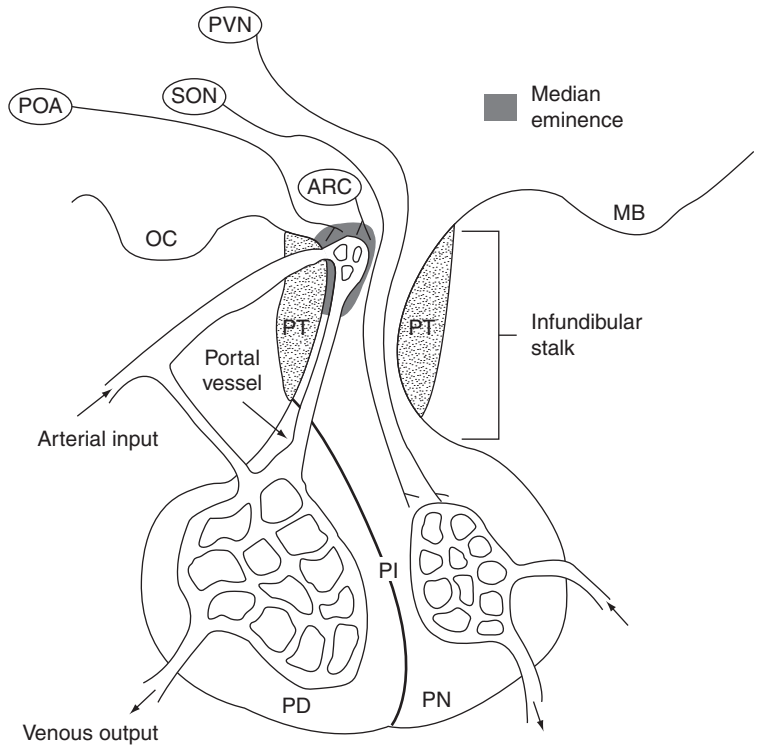

Figure 2 Diagram of the neurovascular anatomy of the hypothalamic–pituitary axis. ARC, arcuate nucleus; MB, mamillary body; OC, optic chiasm; PD, pars distalis; PI, pars intermedia; PN, pars nervosa; POA, preoptic area; PT, pars tuberalis; PVN, paraventricular nucleus; SON, supraoptic nucleus. Adapted from Gutman DA and Nemeroff CB (2002) Neuroendocrinology. In: D'Haenen H, Den Boer JA, and Willner P (eds.) *Biological Psychiatry*, p. 98. London: Wiley.

1 μg kg^{-1} or a fixed dose of 100 μg) and the ensuing ACTH and cortisol response is measured at 30 min intervals during a 2 or 3 h period. Numerous studies have demonstrated a blunted ACTH and β-endorphin response to exogenously administered ovine CRF or human CRF in depressed patients compared to non-depressed subjects. The blunted ACTH response to CRF occurs in depressed DST nonsuppressors but not in DST suppressors. The attenuated ACTH response to CRF is likely due, in part, to chronic hypersecretion of CRF from nerve terminals in the median eminence, which results in downregulation of CRF receptors in the anterior pituitary. This receptor downregulation results in a reduced responsivity of the anterior pituitary to CRF, as has been demonstrated in laboratory animals. Chronic hypercortisolemia likely also contributes to the blunted ACTH response to CRF by cortisol action on the corticotroph. The abnormalities in the CRF stimulation test and DST also appear to be state dependent because they generally normalize following recovery from depression.

A combined dexamethasone/CRF test has also been developed, and it is arguably the most sensitive measure of HPA axis function. In this test, 1.5 mg of dexamethasone is administered orally at night (11 p.m.), and subjects receive an i.v. bolus of 100 μg of human CRF at 3 p.m. the following day. Patients with HPA axis dysfunction, which is reportedly seen in 40–60% of depressed patients, display a paradoxically increased release of ACTH and cortisol following CRF challenge relative to controls. These abnormalities disappear following remission of depression, and normalization of HPA axis function seems to precede full clinical remission. Conversely, the absence of normalization of the HPA axis in clinically remitted depressed patients is associated with later relapse. The combined DEX/CRF test appears to exhibit much higher sensitivity for detecting subtle alterations in HPA axis function; approximately 80% of patients with major depression exhibit an abnormal response to the DEX/CRF test. In contrast, only approximately 44% of patients with major depression demonstrate an abnormal response when DST is administered alone. Furthermore, otherwise healthy individuals with first-degree relatives with an affective illness, which greatly increases their own risk for a mood disorder, had cortisol and ACTH responses in the DEX/CRF test which were greater than those of a control group but considerably reduced in magnitude compared to those of patients currently suffering

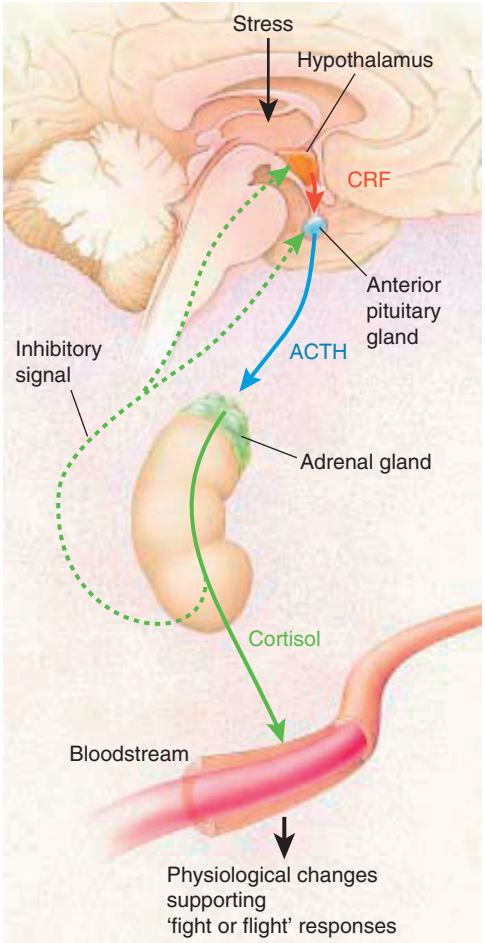

Figure 3 Overview of the feedback mechanism of the HPA axis. Following relevant stimuli, including stress, CRF is released from the hypothalamus into the hypophyseal portal vessels, where it is transported in high concentrations to the pituitary gland. CRF then promotes the release of ACTH, which in turn promotes the release of cortisol from the adrenal glands. Cortisol acts as an inhibitory signal at both the hypothalamus and the pituitary, preventing further CRF and ACTH release, respectively. Reproduced with permission of Tomo Narashima from Nemeroff CB (1998) The neurobiology of depression. *Scientific American* 278:42–49.

from major depression. This suggests that a defect in corticosteroid receptor function may render these individuals more susceptible to developing affective disorders.

Extrahypothalamic CRF and Depression

A rich body of literature primarily on rodents and nonhuman primates has highlighted the extensive network of CRF-containing neurons in extrahypothalamic sites. In rodents, primates, and humans, CRF and its receptors have been heterogeneously localized in a variety of brain regions, including the amygdala, thalamus, hippocampus, and prefrontal cortex. These brain regions are important in regulating many aspects of the mammalian stress response and in regulating affect.

Table 2 HPA axis changes demonstrated in depression

↑ Corticotropin-releasing factor (CRF) in cerebrospinal fluid[a]
↓ Adrenocorticotropic hormone response (ACTH) to CRF stimulation
↓ Density of CRF receptors in frontal cortex of suicide victims
Enlarged pituitary gland in depressed patients[a]
Adrenal gland enlargement in suicide victims and depressed patients
↑ Plasma cortisol during depression[a]
↑ Urinary free cortisol concentrations[a]
Nonsuppression of plasma cortisol and ACTH after dexamethasone administration[a]

[a]State dependent.
Adapted from Gutman DA and Nemeroff CB (2002) Neuroendocrinology. In: D'Haenen H, Den Boer JA, and Willner P (eds.) *Biological Psychiatry*, p. 100. London: Wiley.

Involvement of extrahypothalamic CRF systems in the pathophysiology of depression is supported by numerous studies showing elevated CRF concentrations in the CSF of patients with major depression, although discrepant results have been reported. Elevated CSF CRF concentrations have also been observed in depressed suicide victims. A reduction in CSF CRF concentrations has been reported in healthy volunteers treated with the tricyclic antidepressant desipramine and in depressed patients following treatment with fluoxetine, amitriptyline, or ECT, providing evidence of interactions between antidepressants, noradrenergic and serotonergic neurons, and CRF systems. Elevated CSF CRF concentrations appear to represent a state, rather than a trait, marker of depression (i.e., a marker of the state of depression rather than a marker of vulnerability to depression). Furthermore, high and/or increasing CSF CRF concentrations despite symptomatic improvement of major depression during antidepressant treatment may be the harbinger of early relapse, as we have previously reported for DST nonsuppression.

Although the exact mechanism contributing to CRF hyperactivity remains obscure, studies from our group and others have documented long-term persistent increases in HPA axis activity and extrahypothalamic CRF neuronal activity after exposure to early untoward life events, such as neglect and child abuse, in both laboratory animals (rat and nonhuman primates) and humans. A large database indicates that early life stress permanently sensitizes the HPA axis and increases the activity of CRF neurons leading to a greater risk of developing depression later in life. In one schema, the persistent increase in the activity of CRF neuronal systems results in heightened responses

to stress later in life. To measure HPA responsivity to stress in humans, the Trier Social Stress Test (TSST) was developed. This laboratory paradigm involves a simulated 10 min public speech coupled with a mental arithmetic task. The TSST has been validated as a potent activator of the HPA axis in humans. Our group has reported increased plasma ACTH and cortisol concentrations, presumably due to hypersecretion of CRF, after exposure to the TSST in depressed women who were exposed to severe physical and emotional trauma as children. Studies by our group, using the combined DEX/CRF test described previously, in adult men with early life trauma and comorbid major depression revealed markedly increased HPA axis activity, consistent with the TSST test results in women and with the CRF hypersecretion hypothesis. These data provide evidence for hyperactivity of CRF systems associated with early adverse life events.

CRF and Anxiety Disorders

Involvement of CRF in anxiety disorders has been less well studied in both animal models and human studies than has depression, but the results are nevertheless quite promising. Patients with posttraumatic stress disorder (PTSD; i.e., Vietnam combat veterans) have been reported to exhibit elevated CSF CRF concentrations, as well as a blunted ACTH response to CRF challenge. A study using an indwelling cannula for the repeated collection of lumbar CSF for several hours after the initial lumbar puncture further confirmed elevated CSF CRF concentrations in combat veterans with PTSD. In contrast, low serum cortisol and urinary free cortisol concentrations have repeatedly, yet unexpectedly, been observed in PTSD, especially after dexamethasone administration. One potential mechanism for this finding is heightened negative feedback within the HPA axis in chronic PTSD resulting in reduced basal cortisol concentrations (Figure 3).

Although CSF CRF concentrations are not increased in panic disorder patients, a blunted ACTH response to CRF administration has been observed. Increased or normal concentrations of CSF CRF have been reported in patients with obsessive–compulsive disorder, although significant decreases in CSF CRF concentrations occur with a therapeutic response to clomipramine. Patients with generalized anxiety disorder, however, exhibit similar CSF CRF concentrations as normal controls. In summary, although HPA axis hyperactivity exists in patients with certain anxiety disorders, such perturbations do not exist in the patterns suggestive of chronic CRF hypersecretion as documented in patients with major depression.

Glucocorticoid Receptor Antagonists and Depression

As discussed previously, nonsuppression of cortisol in the DST is very common in patients with major depression with psychotic features. Therefore, one proposed treatment for psychotic depression is the use of mifepristone (RU-486), which at high concentrations is a glucocorticoid receptor antagonist. An initial open-label study showed some efficacy in psychotic major depression following 1 week of treatment with mifepristone. A subsequent larger (~110 patients per treatment group), placebo-controlled trial demonstrated that 7 days of mifepristone followed by 21 days of treatment as usual resulted in significant improvement in psychotic but not depressive symptoms. Overall, this large-scale, double-blind, placebo-controlled trial did not demonstrate efficacy of mifepristone in psychotic depression, although this may be due to the unusually high placebo response rate observed. A second double-blind, placebo-controlled study also failed to demonstrate any separation of mifepristone from placebo, although this study also showed abnormally high placebo response rates. However, similar to the initial study, an abnormally high response rate was noted in both groups, with 95% of patients in both arms of the study demonstrating a response by day 56. There is also evidence suggesting that antiglucocorticoid therapy may improve depressive symptoms in patients in the depressed phase of bipolar disorder. It is reasonable to propose that 1 week of antiglucocorticoid therapy may serve to reset the HPA axis, resulting in subsequent improvement in clinical symptoms.

The HPA Axis in Depression and Anxiety: Conclusions

Numerous studies have documented that when CRF is directly injected into the CNS of laboratory animals, it produces many effects reminiscent of the cardinal symptoms of major depression and certain anxiety disorders, including decreased libido, reduced appetite and weight loss, sleep disturbances, and neophobia. These findings, taken together with the numerous clinical studies described previously, suggest that CRF_1 receptor antagonists represent a novel putative class of antidepressants and/or anxiolytics. Such compounds show activity in nearly every preclinical screen for antidepressants and anxiolytics currently employed. To date, only one open-label study examining the effectiveness of a CRF_1 receptor antagonist, R121919, in patients with major depression has been published. Severity measures of both anxiety and depression were significantly improved. Several studies with novel CRF_1 receptor antagonists are ongoing.

Thyrotropin-Releasing Factor, the Hypothalamic–Pituitary–Thyroid Axis, and Depression

Biology

Dysfunction of the hypothalamic–pituitary–thyroid (HPT) axis has for many years been known to be associated with numerous psychiatric disturbances ranging from mild depression and anxiety to overt psychosis. The HPT axis controls the release of the thyroid hormones tri-iodothyronine (T_3) and thyroxine (T_4). The tripeptide thyrotropin-releasing hormone (TRH) causes the release of adenohypophysial thyroid-stimulating hormone (TSH) into the general circulation, and TSH in turn acts directly on the thyroid to induce thyroid hormone secretion. TRH was the first of the hypothalamic-releasing hormones to be isolated and characterized. It is synthesized predominantly in the hypothalamic paraventricular nucleus, stored in nerve terminals in the median eminence, and released into the vessels of the hypothalamic–hypophyseal portal system. Like the other releasing hormones, TRH is then transported to the sinusoids in the anterior pituitary, where it binds to TRH receptors on the thyrotropes, promoting the release of TSH into the systemic circulation. TSH acts on receptors in the thyroid gland to release the major thyroid hormones, T_3 and T_4. Besides their well-known metabolic effects on a variety of tissues, T_3 and T_4 feed back to the hypothalamus and adenohypophysis, respectively, to regulate the synthesis and release of TRH and TSH.

Within the CNS, the actions of thyroid hormone are mediated by thyroid hormone receptors, which are members of the large superfamily of nuclear hormone receptors. Upon binding to its receptor, the thyroid hormone/receptor complex binds to thyroid hormone response elements located in the promoter region of specific target genes. Within the CNS, rodent studies have demonstrated a relatively high density of thyroid hormone receptors in cortex, hippocampus, hypothalamus, and olfactory bulbs. The importance of thyroid hormone in the CNS is further highlighted by the devastating effects of hypothyroidism on the developing CNS. Fetal and neonatal hypothyroidism can lead to cretinism, which is associated with a number of fetal abnormalities, including severe mental retardation, short stature, thickened skin, and neurologic impairments.

Psychiatric Manifestations of Thyroid Dysfunction

Primary thyroid axis disorders can be associated with a wide range of psychiatric manifestations ranging from depression and anxiety to overt psychosis. Regardless of the precise etiology, hypothyroidism leads to a number of clinical manifestations, including slowed mentation, forgetfulness, decreased hearing, cold intolerance, and ataxia. Decreased energy, weight gain, depression, cognitive impairment, or overt psychosis ('myxedema madness') may also result. Due to the overlapping symptoms with major depression, thyroid hormone deficiency must be ruled out when evaluating patients with depression.

Studies conducted in the late 1960s and early 1970s scrutinized psychiatric morbidity in patients with hypothyroidism, and a substantial number of these patients demonstrated signs of depression, anxiety, and/or cognitive disturbances. Treatment of the underlying thyroid condition alone generally resulted in significant improvement or resolution of these symptoms. Cognitive disturbance is also common in hypothyroidism. Depression and anxiety symptoms are also common in hypothyroidism, occurring in up to 50% and 30% of unselected patients, respectively. Manic and hypomanic states are rarely reported in hypothyroid patients. Although psychosis is the most common reported symptom in the case literature on hypothyroidism, it accounts for only approximately 5% of the psychiatric morbidity in an unselected sample, presumably because the more severe symptoms are more likely to be documented in the clinical literature. It is important to note that occult primary hypothyroidism is among the leading medical causes of refractory depression.

TRH and TSH in Depression

Considerable evidence indicates an increased incidence of HPT axis dysfunction in patients with major depression. More than 30 years ago, research groups led by Prange and by Kastin demonstrated that approximately 25% of patients with major depression exhibit a blunted TSH response to TRH. This has been postulated to be due to hypersecretion of TRH from the median eminence, which leads to TRH receptor down-regulation and consequent reduced sensitivity of the pituitary to exogenous TRH. This hypothesis is supported by reports of elevated TRH concentrations in the CSF of drug-free depressed patients. Depressed patients have also been shown to have an increased rate of symptomless autoimmune thyroiditis, defined by the abnormal presence of antithyroglobulin and/or antimicrosomal thyroid antibodies consistent with grade 4 hypothyroidism. Studies of the elderly have suggested that subclinical hypothyroidism substantially increases the risk for depression, perhaps even more than overt hypothyroidism. However, a separate large-scale study of the elderly found no clear association between TSH, T_4, and depression, anxiety, or cognition, although the authors discussed a number

of confounding factors that may have impacted their results. A large epidemiological study evaluated a population of more than 30 000 individuals and measured plasma TSH concentrations in all women and 50% of the men in the study population. Individuals with abnormally high TSH levels were further screened for the presence of thyroid peroxidase autoantibodies. In this large study of nonselected patients, the authors were unable to detect a significant association between antithyroid antibodies and the prevalence of depression or anxiety.

T_3 augmentation has been reported to be effective in the treatment of refractory depression when combined with tricyclic antidepressants (TCAs). A metareview of the literature on combination TCAs–thyroid hormone supported their efficacy in refractory depression. Since then, selective serotonin reuptake inhibitors (SSRIs) have replaced TCAs as first-line treatment for depression. A small study of 12 patients evaluated the efficacy of thyroid hormone augmentation in patients with treatment-resistant depression. The patients in this study had failed to demonstrate a satisfactory antidepressant response after at least 6 weeks of antidepressant treatment with an SSRI (17-item Hamilton Rating Scale for Depression score >18 at study entry). These patients were also carefully screened and had both normal basal TSH concentrations and normal results of a TRH stimulation test. Although this study was not blinded, thyroid hormone (T_3) augmentation (25 $\mu g\,day^{-1}$ increased to 50 $\mu g\,day^{-1}$ after the first week) was shown to be highly effective in producing a clinical response in 42% of patients during the 3-week study follow-up. The efficacy of thyroid hormone augmentation was also highlighted in the STAR*D study. In this large multicenter, placebo-controlled trial of antidepressant efficacy, 142 patients were randomized to augmentation with either lithium (up to 900 $mg\,day^{-1}$) or T_3 (50 $\mu g\,day^{-1}$) after failing to adequately respond to an initial trial with citalopram and then a second medication switch or augmentation trial. During the 14-week follow-up, in this relatively treatment-resistant population, 24.7% of patients receiving T_3 attained clinical remission. However, a blinded placebo-controlled trial which evaluated T_3 augmentation (25 or 50 $\mu g\,day^{-1}$) in depressed patients receiving paroxetine showed no advantage of thyroid hormone compared to placebo.

There is also some question as to the ideal dosing regimen when using thyroid hormone as an augmenting agent. In one study, hypothyroid patients being treated with T_4 were monitored for two 5-week periods and underwent physiological, biochemical, and neuropsychiatric testing. During one of the periods, their normal T_4 replacement dose was decreased by 50 μg and replaced with 12.5 μg of T_3. The T_3/T_4 combination improved results on several mood and neuropsychologic measures. Although these results are by no means conclusive, they do raise the possibility that combined T_3/T_4 augmentation strategies may have superior efficacy than T_4 alone. Because thyroid hormone is both inexpensive and safe, it remains a viable augmentation strategy compared to other newer strategies that are not necessarily more efficacious.

Although there is evidence suggesting hypersecretion of TRH in a subset of depressed patients, it is unclear how common or important this finding is. Interestingly, lumbar intrathecal infusion of 500 μg of TRH in medication-free inpatients with refractory depression produced a clinically robust, but short-lived, improvement in mood and suicidality, which led the authors to propose that elevated CSF TRH levels might be a compensatory response in depression. Similar antidepressant responses were seen after intravenous nocturnal TRH administration in patients with bipolar disorder, type I or II, currently in a depressed phase. Although this work is preliminary, it raises the interesting possibility that a systemically administered TRH receptor agonist may possess antidepressant properties.

Bipolar Disorder and HPT Abnormalities

HPT axis abnormalities have also been reported in patients with bipolar disorder. Both elevated basal plasma concentrations of TSH and an exaggerated TSH response to TRH have been observed. There is also evidence that bipolar patients with the rapid cycling subtype have a higher prevalence of hypothyroidism (grades I–III) than other bipolar subtypes. A blunted or absent evening surge of plasma TSH, a blunted TSH response to TRH, and the presence of anti-thyroid microsomal and/or anti-thyroglobulin antibodies at higher rates than those of normal comparison subjects have also been demonstrated in bipolar patients. High doses of T_4 have also been reported to be effective in the treatment of rapidly cycling bipolar disorder patients.

Thyrotropin-Releasing Hormone Alterations in Anxiety Disorders

The HPT axis has been relatively poorly studied in anxiety disorders. A blunted TSH response to TRH has been observed in panic disorder patients, but in one of these studies, this effect was only demonstrated in patients with concurrent major depression. Another study found a normal TSH response to TRH in panic disorder patients. No alterations in CSF TRH concentrations have been reported in patients with panic disorder, generalized anxiety disorder, or obsessive–compulsive disorder.

Growth Hormone and Somatostatin

Biology

Growth hormone (GH) is synthesized and secreted from somatotrophs in the anterior pituitary. Its release is unique in that it is controlled by two hypothalamic peptide hormones – growth hormone-releasing factor (GHRF) and somatostatin – and secondarily by classical neurotransmitters such as dopamine (DA), norepinephrine (NE), and 5-hydroxy tryptamine (5-HT), which innervate the releasing factor-containing neurons (Table 3). Somatostatin, also known as growth hormone-release inhibiting hormone or somatotrophin-release inhibiting factor (SRIF), was first isolated from ovine hypothalamus in 1974. It is a tetradecapeptide containing a disulfide bridge linking the two cysteine residues. Somatostatin is synthesized predominantly in neurons in the periventricular and paraventricular nucleus of the hypothalamus, is released from nerve terminals in the median eminence, and inhibits GH release. Somatostatin also inhibits the release of both CRF and ACTH and has been termed 'panhibin.' Somatostatin has a wide extrahypothalamic brain distribution, including the limbic system, cerebral cortex, hippocampus, and amygdala.

GHRF is a 44-amino-acid peptide that has the most limited CNS distribution of all the hypothalamic-releasing hormones thus far identified. GHRF-containing neurons are concentrated primarily in the infundibular and arcuate nuclei of the hypothalamus and stimulate the synthesis and release of GH. Dopamine-, NE-, and 5HT-containing neurons innervate GHRF neurons and thereby modulate GH release. Both GHRF and SRIF are released from nerve terminals in the median eminence into the hypothalamo-hypophyseal portal system, where they act on somatotrophs in the anterior pituitary to regulate GH release. Negative feedback is provided by GH, which stimulates SRIF release, thus preventing further GH release. The hypothalamic–growth hormone axis is unique in that it does not have a single target endocrine gland but instead GH acts directly on a myriad of tissue targets, including bone, muscle, and liver. However, GH does release somatomedin from the liver.

Table 3 Releasing and inhibiting factors for growth hormone

Growth hormone-releasing factors
Growth hormone-releasing hormone
Dopamine
L-DOPA
Apomorphine (dopamine receptor agonist)
Norepinephrine
Clonidine (nonspecific α-adrenergic receptor agonist)
Serotonin
L-Tryptophan
5-Hydroxytrytophan
Factors which inhibit growth hormone release
Somatostatin
Phentolamine (nonspecific α-adrenergic receptor antagonist)
Methysergide (serotonin receptor antagonist)
Cyproheptadine (serotonin receptor antagonist)

Adapted from Gutman DA and Nemeroff CB (2002) Neuroendocrinology. In: D'Haenen H, Den Boer JA, and Willner P (eds.) *Biological Psychiatry*, p. 106. London: Wiley.

GH and Affective Disorders

Several findings suggest dysregulation of GH secretion in depression. A blunted nocturnal GH surge in depression has been demonstrated, whereas daylight GH secretion has been reported to be exaggerated in both unipolar and bipolar depressed patients. A number of studies have also demonstrated a blunted GH response to drugs that increase noradrenergic activity (clonidine and desipramine) and, to a lesser extent, dopaminergic agonists (apomorphine) in depressed patients. Two studies have suggested that the blunted GH response to apomorphine in depressed patients is predictive of suicidal behaviors. The blunted GH response to clonidine has been demonstrated to be unrelated to age or sex and may be secondary to decreased α_2-adrenergic receptor sensitivity in depression. No changes in basal GH secretion were observed in patients with panic disorder who were monitored by blood sampling every 15 min during a 24 h period. Another group failed to find changes in GH secretion following apomorphine in either unipolar depressed or bipolar patients.

A GHRF stimulation test, analogous to the CRF and TRH stimulation tests, has also been developed. Two studies found a blunted GH response to GHRF in depressed patients, whereas our group failed to confirm the results in a study of 19 depressed and 19 matched controls.

A comprehensive review of GHRF stimulation tests in depression, anorexia nervosa, bulimia, panic disorder, schizophrenia, and Alzheimer's disease was conducted, and it was found that the results of this test are apparently not always consistent and in some cases are contradictory. Some of the confounding factors, including the variability of GHRF-stimulated GH release among controls and age- and gender-related effects, may account for some of this variability. Additional studies to standardize the GHRF stimulation test for application in psychiatry are clearly needed.

Somatostatin and Psychiatric Disorders

Decreased CSF SRIF concentrations in patients with depression, dementia, schizophrenia, and Alzheimer's disease have been demonstrated. Reduced CSF SRIF concentrations have been reported in DST nonsuppressors, whether diagnosed with schizophrenia or depression, and are negatively correlated with the maximum postdexamethasone cortisol plasma concentration in patients with major depression. Although there is evidence of normalization of CSF SRIF concentrations after recovery from depression, other studies have noted no such changes despite clinical improvement with antidepressants or ECT treatment. Somatostatin concentrations are markedly reduced in cerebrocortical and hippocampal postmortem tissue in Alzheimer's disease. Somatostatin also inhibits the release of both CRF and ACTH, indicating a possible direct interaction between the growth hormone and HPA axes. No differences in CSF SRIF concentrations between patients with generalized anxiety disorder and controls were observed. No published studies measuring GHRH concentrations or GHRH mRNA expression have been conducted of postmortem tissue of depressed patients and matched controls that, in light of the evidence presented here, is of interest. Similarly, no CSF studies of GHRH in psychiatric patients have been reported.

Alterations of the hypothalamic–pituitary growth hormone axis and SRIF concentrations have been reported in major depression. Whether these changes represent primary pathophysiological alterations or epiphenomena remain to be determined. Arguably, the blunted GH response to clonidine seen in depression may be one of the most reproducible and specific findings in the biology of affective disorders. Diminished availability of SRIF might plausibly contribute to CRF hypersecretion and increased HPA axis activity. Further elucidation of SRIF receptor function and the effects/utility of SRIF receptor agonists and antagonists may provide important information regarding the pathophysiology of major depression.

Hypothalamic–Prolactin Axis

Biology

Unlike other anterior pituitary hormones, prolactin (PRL) release is largely regulated via tonic inhibition by a prolactin-inhibitory factor, determined not to be a neuropeptide but, rather, the monoaminergic neurotransmitter DA. Neurons in the tuberoinfundibular system of the hypothalamus release DA from nerve terminals in the median eminence that directly inhibit PRL release from lactotrophs. Prolactin can also inhibit its own release by a short-loop negative feedback to the hypothalamus. Prolactin is believed to play a role in the behavioral aspects of reproduction and infant care. Serum PRL concentrations are normally low throughout life in males. Basal PRL levels increase in females following parturition, and suckling stimulates PRL release. Prolactin stimulates breast growth and lactation. TRH, oxytocin, serotonin, estrogen, and other neuroregulators have prolactin-releasing factor (PRF) activity.

Prolactin-Releasing Factor

A putative peptidic PRF has also been identified. PRF-containing cell bodies have been detected in a number of brain regions, including the nucleus tractus solitarius, the ventral and lateral reticular nuclei in the medulla, and the dorsomedial nuclei of the hypothalamus. These neurons project nerve fibers to a wide range of brain areas. After its initial characterization and the demonstration both *in vivo* and *in vitro* that PRF can indeed induce PRL secretion, considerable debate persists as to whether PRF is a major physiological secretagogue of PRL. It is important to note that PRL may also be involved in the regulation of energy and food balance, sleep regulation, and may also interact with CRF and the HPA axis.

PRL and Psychiatric Disorders

Excess circulating PRL can lead to a number of clinical symptoms. The most common causes of hyperprolactinemia are tumors, usually microadenomas of pituitary lactotrophs. Treatment with typical antipsychotic medications and some atypical antipsychotics (e.g., risperidone at high doses), because of their potent blockade of D_2 DA receptors, can also cause hyperprolactenemia. Hyperprolactinemia often leads to reduced testosterone secretion in men, and decreased libido, lactation, and gynecomastia can be seen in both men and women. Women may also display menstrual cycle irregularities including amenorrhea. Patients may also complain of depression, stress intolerance, anxiety, and increased irritability that usually resolve following treatments that reduce serum PRL levels. Consistent alterations in the hypothalamic–prolactin axis have not been clearly demonstrated in psychiatric disorders. Because PRL release is inhibited by DA, the PRL response to infusions of DA agonists has also been used to estimate CNS dopaminergic activity, although it is likely only an accurate reflection of the tuberoinfundibular hypothalamic DA system.

Many reports have used provocative tests that increase serotonergic neurotransmission and PRL secretion in patients with psychiatric disorders.

Briefly, these tests use agents that increase serotonergic transmission such as L-tryptophan, 5-hydroxytryptophan, and fenfluramine. In general, the PRL response to agents that increase serotonergic activity is blunted in depression, as well as in patients with cluster B personality disorders. In contrast, a study of 20 patients with major depression demonstrated an exaggerated PRL response to L-tryptophan. These authors suggested that concomitant psychotropic medications or hypercortisolemia may have confounded earlier studies. However, the available data suggest that the altered PRL response seen in depression is, in part, mediated by an alteration in 5-HT_{1A} receptor responsiveness.

Arginine–Vasopressin

Biology

Arginine–vasopressin (AVP), also known as antidiuretic hormone, is a nonapeptide synthesized in the lateral magnocellular neurons of the paraventricular and supraoptic nuclei of the hypothalamus, and it is released directly into the bloodstream from axon terminals in the posterior pituitary. These AVP-containing neurons terminate in the neurohypophysis of the pituitary and secrete AVP into the systemic circulation, although they also project nerve terminals to the hypothalamo-hypophyseal portal system. Another group of AVP-containing neurons project from the medial parvocellular subdivision of the PVN to the median eminence. Within the median eminence, the parvocellular-derived AVP is released from axon terminals, secreted into the hypothalamo-hypophyseal portal circulation, and transported to the anterior lobe of the pituitary gland. Extrahypothalamic AVP-containing neurons lie within limbic structures such as the septum and amygdala, as well as in the brain stem and spinal cord. AVP and the other major posterior pituitary hormone, the nonapeptide oxytocin, are believed to play a role in modulating neural activity in hypothalamic, limbic, and autonomic circuits.

AVP has prominent roles in controlling fluid balance via its effects on the kidney, in regulating blood pressure by its vasoconstrictive effects on blood vessels, and it can directly promote thirst. AVP can also act synergistically with CRF to promote the release of ACTH in humans and laboratory animals following stressful stimuli. Chronic stress or adrenalectomy increases the activity of the parvocellular AVP system. CRF and AVP are co-localized in the parvocellular cells of the human hypothalamus and may be secreted together into the human hypothalamic-hypophyseal portal circulation. The amount of AVP and CRF released into the hypothalamic-hypophyseal portal circulation varies in different species and in response to different stressors.

AVP in Anxiety and Depression

CSF AVP concentrations in patients with major depression, particularly those with atypical depression, were reported to be reduced in comparison to those of control subjects. CSF AVP is likely largely extrahypothalamic and not reflective of PVN AVP circuits. Basal plasma concentrations of AVP (secreted from the magnocellular neurons of the PVN after osmotic and/or baroreceptor stimulation) in depressed patients are also reportedly decreased in comparison to those of age-matched controls, although others have found no such difference. In one study, reduced plasma AVP concentrations, as well as a blunted GH and AVP response to both clonidine and apomorphine, were reported in depressed patients; no differences in basal or stimulated oxytocin concentrations were observed.

Combined infusion of lysine–vasopressin and human CRF produced nonsuppression of cortisol following the DST in healthy controls; in these same patients, neither agent infused separately produced DST nonsuppression. This suggests that both CRF and AVP contribute to the HPA axis hyperactivity characteristic of depression. AVP serum concentrations following the DST have also been reported to be elevated in both patients with chronic major depressive disorder and patients with symptomatic and remitted bipolar disorder.

A mildly blunted ACTH response to exogenous AVP administration has been reported in depression, but this finding was not replicated in two other studies. In a postmortem tissue study, an increase in the number of PVN AVP neurons, as well as those that co-localize AVP and CRF neurons, was reported in depressed patients compared to controls. Increased AVP mRNA expression has been reported in the PVN and supraoptic nucleus (SON) of nine depressed patients compared to eight control subjects. This is of interest in view of the ability of AVP to potentiate the actions of CRF at corticotrophs.

Interestingly, elevations in CSF and plasma AVP concentrations have been reported in manic patients in comparison to patients with unipolar depression and controls. Clearly, hypothalamic and extrahypothalamic AVP circuits are regulated independently. Additional work is warranted, especially in view of the emerging evidence of antidepressant activity of AVP receptor antagonists in a variety of animal models. Whether the perturbations of AVP secretion in patients with neuropsychiatric disorders are state or trait dependent has yet to be elucidated.

Oxytocin

Biology

Oxytocin is a nonapeptide synthesized primarily in magnocellular neurons located in the SON and PVN of the hypothalamus. It shares significant homology with AVP, differing in only two of nine amino acids. Similar to AVP, it is primarily released into the general circulation from the posterior lobe of the pituitary gland (i.e., the neurohypophysis). Oxytocin-producing neurons in the PVN also project to a number of extrahypothalamic sites, including the hippocampus, amygdala, striatum, hypothalamus, nucleus accumbens, locus coeruleus, and the nucleus of the tractus solitarius.

Oxytocin serves a number of physiologic functions, but much of the research on this peptide has focused on its role in reproduction. Oxytocin promotes contraction of the uterus during labor, as well as milk letdown during lactation. A role for both oxytocin and vasopressin in affiliative behavior and social bonding in animals has been demonstrated. Compelling preclinical studies have also suggested a very significant role for oxytocin in the entire panoply of maternal behaviors, including nest building, licking and grooming pups, and nursing.

Oxytocin and Psychiatric Disorders

In an investigation of 25 medication-free depressed patients and 25 age- and gender-matched controls, no significant differences in serum oxytocin concentrations were detected either at baseline or following both apomorphine and clonidine stimulation tests. In this same experiment, basal vasopressin concentrations were significantly lower in depressed patients and reduced following apomorphine and clonidine challenge. GH responses were similarly blunted in depressed patients after either clonidine or apomorphine treatment. Intranasal oxytocin has also been shown to influence willingness to engage in social risks. Due to its role in affiliative behavior and social interactions, there is increasing interest in understanding what role, if any, oxytocin may have in autism spectrum disorder. A putative role for oxytocin in postpartum depression is another area worth exploring.

Clinical Implications and Conclusions

Careful clinical observations of patients with psychiatric disorders associated with primary endocrine disorders such as Cushing's syndrome and hypothyroidism have led to a broader understanding of the role of neuroendocrine disturbances in a variety of psychiatric disorders, most notably in mood disorders. These studies have led to major advances in biological psychiatry by helping to elucidate the brain circuits involved in the pathophysiology of mood and anxiety disorders. Foremost among these is the CRF theory of depression, which is supported by studies using a variety of experimental approaches and has led to the development of a putative novel therapeutic class of antidepressants – CRF receptor antagonists. Furthermore, this body of work has provided a neurobiological mechanism to explain the increase in depression seen in patients exposed to trauma early in life, first suggested by Freud in the early twentieth century. If CRF truly is the 'black bile' of depression, CRF antagonists may represent a novel class of antidepressants with a unique mechanism of action distinct from currently available antidepressants.

In addition to the HPA axis and CRF alterations observed in depression, HPT axis abnormalities are also very common; the majority of depressed patients, in fact, exhibit alterations in one of these two axes. Furthermore, the blunted GH response to clonidine or apomorphine and the blunted PRL response to serotonergic agonists in depressed patients have been widely replicated.

It is obvious that the vast majority of studies have focused on patients with mood disorders, particularly unipolar depression. Other disorders, including eating disorders, anxiety disorders, autism, schizophrenia, and axis II diagnoses, should also be evaluated for neuroendocrine dysfunction. The availability of selective receptor ligands that can be utilized with positron emission tomography will mark the next major leap in our understanding of the neuroendocrine axes in psychiatric disorders. The ability to determine peptide receptor alterations in the brain and pituitary of patients with psychiatric disorders will contribute immensely to our understanding of the neurobiological underpinnings of such disorders.

Finally, an increasing number of studies have demonstrated that depression is a systemic disease that increases vulnerability to other disorders. Depressed patients demonstrate an increased incidence of coronary artery disease and stroke, osteoporosis, and perhaps cancer. These observations may at least, in part, be attributed to the endocrine alterations observed in depression. Neuroendocrine alterations in depression may well contribute to this phenomenon.

See also: Anxiety Disorders; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Corticotropin-Releasing Hormone and Urocortins:

Binding Proteins and Receptors; Depression and the Brain; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Hypothalamo–Pituitary–Thyroid Axis: Organization, Neural/Endocrine Control of TRH; Neuroendocrinology; Neuroendocrinology of Social/Affiliative Behavior; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress, the HPA Axis and Depressive Illness.

Further Reading

Burbach JP, Young LJ, and Russell J (2006) Oxytocin: Synthesis, secretion and reproductive functions. In: Neill JD (ed.) *Knobil and Neill's Physiology of Reproduction*, pp. 3055–3128. New York: Elsevier.

Checkley SA, Slade AP, and Shur E (1981) Growth hormone and other responses to clonidine in patients with endogenous depression. *British Journal of Psychiatry* 138: 51–55.

Fink G (2000) Neuroendocrine regulation of pituitary function: General principles. In: Conn P and Freeman ME (eds.) *Neuroendocrinology in Physiology and Medicine*, pp. 112–120. Totowa, NJ: Humana Press.

Gutman DA and Nemeroff CB (2002) Neuroendocrinology. In: D'Haenen H, Den Boer JA, and Willner P (eds.) *Biological Psychiatry*. London: Wiley

Hauger R and Dautzenberg FM (2000) Regulation of the stress response by corticotropin releasing factor. In: Conn P and Freeman ME (eds.) *Neuroendocrinology in Physiology and Medicine*, pp. 267–293. Totowa, NJ: Humana Press.

Insel TR (1997) A neurobiological basis of social attachment. *American Journal of Psychiatry* 154(6): 726–735.

Levine J (2000) The hypothalamus as a major integrating center. In: Conn P and Freeman ME (eds.) *Neuroendocrinology in Physiology and Medicine*, pp. 75–95. Totowa, NJ: Humana Press.

Lim MM and Young LJ (2006) Neuropeptidergic regulation of affiliative behavior and social bonding in animals. *Hormones and Behavior* 50(4): 506–517.

Nicholas L, Dawkins K, and Golden RN (1998) Psychoneuroendocrinology of depression: Prolactin. *Psychiatric Clinics of North America* 21(2): 341–358.

Nierenberg AA, Fava M, Trivedi MH, et al. (2006) A comparison of lithium and T(3) augmentation following two failed medication treatments for depression: A STAR*D report. *American Journal of Psychiatry* 163(9): 1519–1530.

Prange AJ Jr., Lara PP, Wilson IC, Alltop LB, and Breese GR (1972) Effects of thyrotropin-releasing hormone in depression. *Lancet* 2(7785): 999–1002.

Raadsheer FC, Hoogendijk WJG, Stam FC, Tilders FJH, and Swaab DF (1994) Increased numbers of corticotropin-releasing hormone expressing neurons in the hypothalamic paraventricular nucleus of depressed patients. *Neuroendocrinology* 60: 436–444.

Rubinow D, Dais C, and Post R (1995) Somatostatin in the central nervous system. In: Bloom F and Kupfer D (eds.) *Psychopharmacology: The Fourth Generation of Progress*, pp. 553–562. New York: Raven Press.

Steckler T and Holsboer F (1999) Corticotropin-releasing hormone receptor subtypes and emotion. *Biological Psychiatry* 46(11): 1480–1508.

Yehuda R, Teicher MH, Trestman RL, Levengood RA, and Siever LJ (1996) Cortisol regulation in posttraumatic stress disorder and major depression: A chronobiological analysis. *Biological Psychiatry* 40(2): 79–88.

Neuroendocrinology of Puberty

A Lomniczi, V Matagne, and S R Ojeda, Oregon Health and Science University, Beaverton, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

General Concepts

Puberty is the phase of life when individuals become capable of reproducing, and females are able to sustain a pregnancy to term. In all mammalian species, secondary sexual characteristics become established at puberty as a result of increasing sex steroid levels. In males, an increased androgen production induces enlargement of the testes and penis, as well as growth of the accessory sex organs (seminal vesicles and prostate). In females, estrogen induces development of the mammary gland, as well as development of the internal (uterus and oviduct) and external (vagina and vulva) genitalia. In human males, puberty results in increased body hair and muscle mass, along with voice deepening. In human females, there is a widening of the pelvis and accumulation of fat in the hips, thighs, buttocks, and breasts. Humans also undergo a growth spurt at puberty, but this is not observed in other species (except great apes).

The pubertal phase of genital development, initiated and sustained by gonadal steroids, is known in humans as gonadarche. Gonadarche results from reactivation of the hypothalamic–pituitary–gonadal axis, which is active during early postnatal life and operates at a reduced pace during childhood. As a consequence of this reawakening, gonadal steroid secretion increases, leading to virilization in boys and breast development (thelarche) followed by the first menstruation (menarche) in girls. Gonadarche is preceded by a phase of increased adrenal androgen secretion known as adrenarche. A peculiarity of humans and great apes, adrenarche in humans is responsible for the appearance of pubic and axillary hair. No such changes in adrenal activity are observed in other species.

The Hypothalamic–Pituitary–Gonadal Axis

The medial basal hypothalamus in primates, and the preoptic area (POA) in rodents, contains 800–1000 neurosecretory neurons that govern pituitary gonadotropin secretion via release of gonadotropin-releasing hormone (GnRH). GnRH reaches the pituitary gland by a vascular route provided by the portal vessels that connect the median eminence of the hypothalamus to the pituitary gland. Pituitary gonadotropes respond to GnRH with release of the gonadotropins luteinizing hormone (LH) and follicle-stimulating hormone (FSH), which act on the gonads to promote the secretion of steroids and peptides, in addition to maturation of germinal cells. In turn, gonadal hormones control the release of GnRH and gonadotropins via negative and positive feedback mechanisms (Figure 1).

Central Role of the GnRH Neuron

For many years, it has been accepted that GnRH neurons arise from the olfactory epithelium and migrate to the forebrain along the olfactory nerve pathway. This view has been challenged by evidence suggesting that subpopulations of GnRH neurons originate from different embryonic neural sites, and that one of these populations associates transiently with the olfactory placode en route to the basal forebrain. Notwithstanding these results, the subject remains controversial because other authors have presented findings confirming the validity of the initial interpretation.

In rodents, most GnRH neurons take residency in the POA; in primates, they continue caudally to colonize the medial basal hypothalamus. This migration requires at least three genes – *KAL1* (which encodes an extracellular matrix glycoprotein), *FGFR1* (a receptor for fibroblast growth factor), and *NELF* (a cell surface migratory factor). Loss-of-function mutations of each of these genes result in isolated hypothalamic hypogonadism (IHH), a syndrome characterized by the relative absence of gonadotropin secretion and failure of the gonads to develop.

GnRH neurons transduce neuronal and glial signals into periodic discharges of GnRH. This cell–cell communication complex, collectively known as the 'pulse generator,' is responsible for maintaining a pulsatile mode of GnRH release. Intriguingly, immortalized GnRH-secreting neurons also release GnRH in a pulsatile manner, suggesting that periodic GnRH secretion is an intrinsic property of these neurons. Studies have shown the involvement of cyclic AMP and Na^+ channels associated with oscillatory increases in intracellular calcium and gap junction communications to be key components of this self-sustained system.

Pulsatile GnRH release is already evident during GnRH neuronal migration into the brain. In humans and nonhuman primates, gonadotropin secretion (likely driven by changes in GnRH output) reaches peak levels by midgestation and decreases toward term due to the establishment of steroid negative feedback. In rodents, gonadotropins secretion is initiated near the end of gestation. In all these species,

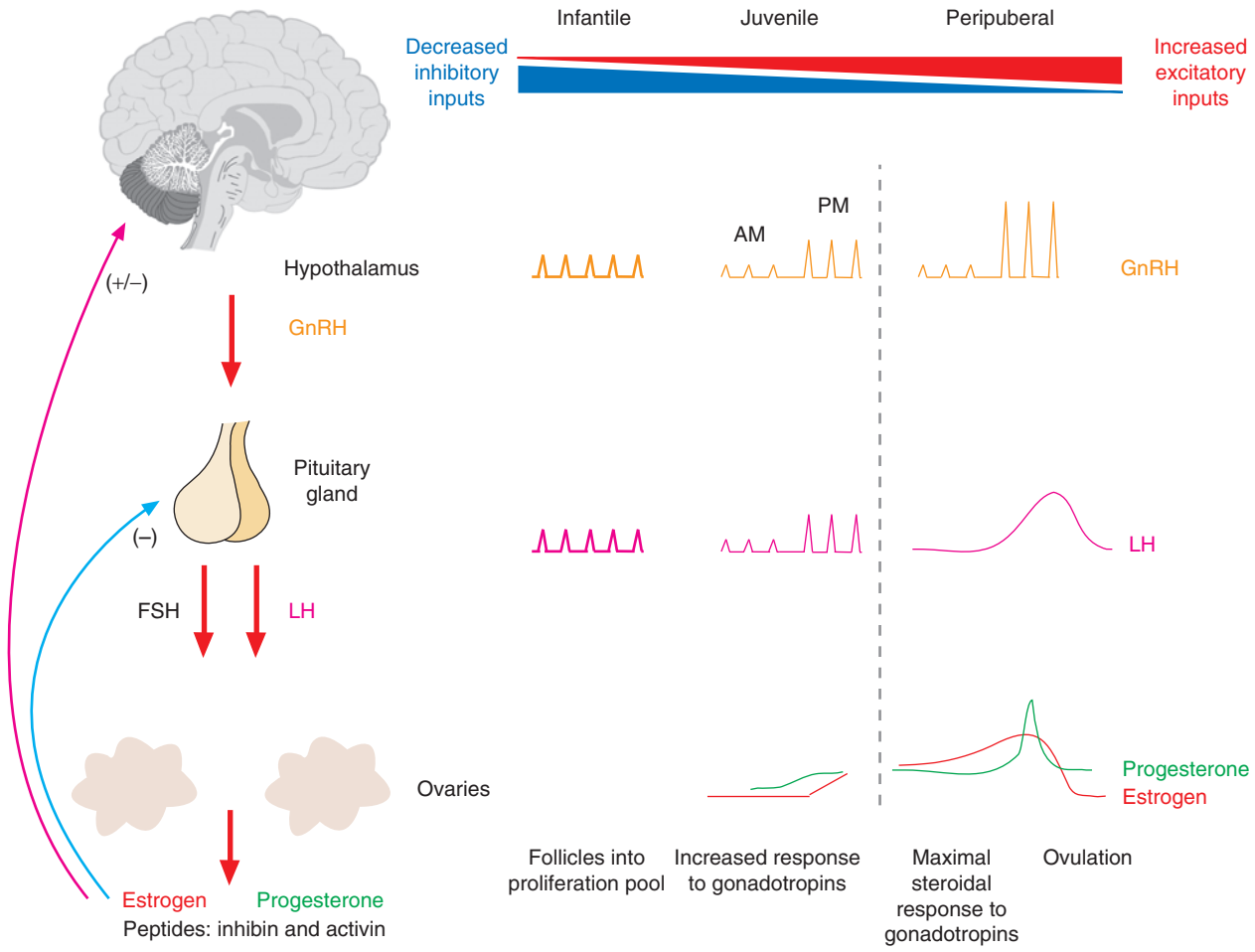

Figure 1 Current hypothesis concerning the central mechanism responsible for setting puberty in motion. Hypothalamic GnRH promotes the secretion and synthesis of pituitary gonadotropins (LH and FSH), which in turn stimulate the secretion of gonadal steroids (estrogen and progesterone) and gonadal peptides, such as inhibin and activin. During the juvenile period, GnRH secretion is low because of a reduced excitatory tone, in the presence of relatively unopposed inhibitory transsynaptic inputs. Puberty is initiated in a gonad-independent manner after a coordinated increase in excitatory inputs (transsynaptic and glial), coupled with a reduction in inhibitory transsynaptic inputs to the GnRH network. The initiation of puberty is accompanied by increased gonadal responsiveness to gonadotropins.

circulating gonadotropin levels are elevated during the infantile period and decrease thereafter (Figure 1). In primates, this decrease is gonad independent. The central mechanisms causing it have been traditionally considered to result from enhancement of a restraining influence; nevertheless, a paucity of excitatory inputs to GnRH neurons is a likely alternative.

The most striking difference between rodents and primates as animal models for the study of the pubertal process is the apparent absence in rodents of the juvenile hiatus in gonadotropin secretion. This difference might not be absolute, however, because an ovary-independent juvenile decrease in FSH release has been reported to occur in rats. Importantly, pulsatile gonadotropin secretion increases in a diurnal manner at the end of juvenile development in both primates and rodents, signaling the initiation of the pubertal process (Figure 1).

What are the mechanisms underlying the pubertal increase in GnRH release? Work from several laboratories has led to the conclusion that both the reduced gonadotropin output during juvenile development and the reemergence of an increased GnRH pulsatility result not from an intrinsic alteration in GnRH neuronal activity or a modification in gonadal feedback control but, rather, as a consequence of coordinated changes in transsynaptic and glial inputs to the GnRH neuronal network. According to this concept, such events are initiated in a gonad-independent manner. The changes in transsynaptic communication that initiate puberty involve an increase in excitatory inputs accompanied by a decrease in inhibitory

neurotransmission. The changes in glial inputs involve an enhanced growth factor-mediated glial-to-neuron signaling. Figure 1 schematizes these events.

Transsynaptic Control of the GnRH Neuronal Network

The transsynaptic control of GnRH secretion is mostly exerted by neuronal networks that utilize amino acids, catecholamines, and neuropeptides for either excitatory or inhibitory neurotransmission.

Classical Inhibitory Neurotransmission

GABA γ-Aminobutyric acid (GABA) is the dominant inhibitory neurotransmitter in the hypothalamus. GABA is synthesized from glutamate via a decarboxylation reaction catalyzed by the enzyme glutamic acid decarboxylase (GAD). Although the initiation of puberty is not accompanied by major changes in expression of the genes encoding GAD-65 and GAD-67, the two GAD enzymes, GABA release decreases in both rats and monkeys during the peripubertal stage, suggesting that in both species the initiation of puberty is preceded by a loss in GABA-mediated inhibitory neurotransmission.

GABA appears to mainly use $GABA_A$ receptors to reduce GnRH secretion because puberty is strikingly advanced in female monkeys when these receptors are blocked. For many years, it has been assumed that an inhibitory, $GABA_A$ receptor-mediated effect of GABA on GnRH secretion is directly exerted on GnRH neurons. However, studies indicate that these neurons respond to $GABA_A$ receptor stimulation with excitation instead of inhibition, and that a direct GABAergic inhibition may be mediated by $GABA_B$ receptors. Some researchers, however, argue that activation of $GABA_A$ receptors in GnRH neurons is inhibitory instead of excitatory.

Opioid peptides The function of different opioid peptides on pubertal GnRH release has not been completely elucidated. An inhibitory effect is indicated by the ability of naloxone (a μ-opioid receptor antagonist) to elicit precocious puberty when administered to female rats soon after birth. More direct evidence is the decreased effectiveness of opioids to suppress LH secretion detected during the onset of rat puberty. Although opioid peptides as a group do not appear to restrain primate puberty, the role of specific opioid systems (enkephalins, endorphins, etc.) has not been examined in detail.

Classical Excitatory Neurotransmission

The most important excitatory neuronal system controlling puberty utilizes glutamate as neurotransmitter. Another well-established stimulatory system, perhaps subservient to glutamatergic neurons, utilizes norepinephrine for neurotransmission.

Glutamate Glutamate is the major excitatory neurotransmitter in the hypothalamus. GnRH neurons are directly innervated by glutamatergic fibers. The facilitatory control they exert on GnRH neurons increases gradually during the infantile–juvenile period as shown by the increasing ability of glutamate receptor stimulation to enhance GnRH release during this phase of development. In the rat, hypothalamic glutamate concentrations increase with age, reaching a maximum after the onset of puberty. In the female rhesus monkey, prepubertal glutamate release is very low at the level of the median eminence and increases markedly during the early pubertal period, remaining elevated at midpuberty.

Endogenous glutamate exists as four different pools: transmitter, metabolic, glial, and GABA precursor. There is little information concerning the dynamics of each at the time of puberty, but as indicated previously, changes in the neurotransmitter pool appear to be critical for the pubertal activation of glutamatergic synaptic transmission.

Glutamate is the endogenous ligand of two major receptor families: (1) ionotropic receptors, consisting of NMDA, AMPA, and kainate receptor subtypes, and (2) metabotropic receptors, which are G-protein-coupled receptors. All these receptors are abundant in the hypothalamus. Activation of NMDA, AMPA, and kainate receptors stimulates the release of GnRH and LH in several species, including rats and monkeys. The importance of NMDA receptors in the control of sexual development is evidenced by the advancement of puberty elicited by NMDA receptor stimulation in both female rats and male monkeys and by the delay in puberty caused by pharmacological blockade of these receptors. Further evidence for this involvement is provided by the gonad-independent increase in ligand-dependent activation of hypothalamic NMDA receptors seen at puberty and the greater number of GnRH neurons expressing NMDA receptors in pubertal than in immature female rats. AMPA receptors may play a complementary role, facilitating the progression of puberty, because they are necessary for the preovulatory surge of gonadotropins. Also, kainate receptors appear to be involved because the number of GnRH neurons expressing kainate receptor subunits essential for receptor function increases during the days encompassing the initiation of rat puberty. These observations indicate that an increased direct excitatory tone to the GnRH neuronal network is an intrinsic component of the pubertal

process. A pubertal augmentation in dendritic arborization and spine formation observed in GnRH neurons reinforces this concept because spine density reflects the degree of excitatory inputs to neurons. Electrophysiological studies employing green fluorescence protein-tagged GnRH neurons have provided irrefutable evidence that GnRH neurons are indeed subjected to a strong glutamatergic control, exerted via NMDA and AMPA/kainate receptors.

Catecholamines A role for catecholamines in the onset of puberty was first shown by the delayed vaginal opening that follows inhibition of brain catecholamines synthesis. In both rats and monkeys, pharmacological approaches have shown that the principal catecholamine involved is norepinephrine, which acts via α_1 and α_2 receptors to elicit GnRH secretion. Norepinephrine turnover, a measure of catecholaminergic activity, increases in the hypothalamus at the time of rat puberty, with maximal values being attained on the day of first proestrus. Noradrenergic neurons located in the brain stem are thought to be responsible for the increased noradrenergic stimulation of GnRH release at the time of the first preovulatory surge of gonadotropins. However, it does not appear that this change is the primary change responsible for the initiation of puberty because it is estrogen dependent.

Neuropeptide Y A facilitatory effect of neuropeptide Y (NPY) on the initiation of rat puberty was suggested by the delay in vaginal opening that results from the intraventricular infusion of an NPY antiserum and by the advancement of the vaginal opening caused by the intraventricular injection of NPY. NPY neutralizing antibodies delivered into the median eminence of monkeys also resulted in suppression of GnRH release in pubertal, but not prepubertal, monkeys. In contrast to these results, an inhibitory role of NPY has been reported in rats and monkeys. The reasons for these discrepancies remain to be determined.

The Novel Kisspeptin–GPR54 Signaling System

The kisspeptin–GPR54 receptor signaling pathway is a novel system involved in the hypothalamic control of GnRH secretion. In 2003, two groups reported that mutations in the GPR54 gene resulted in IHH. One of these groups also showed that mice lacking *Gpr54* exhibit a similar phenotype. Proteolytic cleavage of the primary product of the *KiSS1* gene, a 53-amino-acid peptide known as kisspeptin or metastin, generates the decapeptide kisspeptin-10, which is a potent stimulator of LH release. Because kisspeptin is equally effective in stimulating LH release when administered either intracerebrally or systemically, it may have a dual hypothalamic–pituitary site of action or a main effect at the median eminence of the hypothalamus. Kisspeptin-containing neurons are mainly located in the arcuate nucleus (ARC) of the hypothalamus and the anteroventral periventricular nucleus (AVPV) of the POA, with some neurons also present in the periventricular nucleus. The kisspeptin receptor, GPR54, is more diffusely distributed, suggesting that kisspeptin affects the activity of more than one neuronal subset.

GnRH neurons contain GPR54 receptors, suggesting that they are directly targeted by kisspeptin neurons. Evidence for this direct site of action was provided by the demonstration of an excitatory effect of kisspeptin on the bursting activity of GnRH neurons. This study also showed that although mouse GnRH neurons already display GPR54 receptors during early postnatal development, these receptors acquire the ability to excite GnRH neurons in response to kisspeptin only during the transition from the juvenile stage into adulthood. Interestingly, kisspeptin expression is differentially regulated by gonadal steroids in the hypothalamus and POA. Whereas the number of KiSS1-expressing neurons increases in the ARC in response to ovariectomy and decreases after estrogen or androgen treatment, KiSS1 expression is increased in the AVPV neurons in response to sex steroids. These opposite changes suggest that kisspeptin may be involved in mediating the positive feedback effects of estrogen in the POA and estrogen negative feedback in the ARC.

Glial Control of the GnRH Neuronal Network

A fascinating aspect of the GnRH neuronal network is the intimate association that exists between GnRH neurons and glial cells. Studies of both nonhuman primates and rodents have shown that this association is dynamic and remarkably plastic. It involves both astrocytes (**Figure 2(a)**) and a type of modified ependymoglial cells lining the wall of the third ventricle, known as tanycytes (**Figure 2(b)**). Astroglial processes abundantly appose the GnRH cell membrane, and this apposition is particularly evident along the secretory axons projecting to the median eminence. Although species differences exist (mostly manifested as a much less abundant astroglial ensheathment of GnRH neuronal perikarya in rats than in monkeys), the degree of glial apposition changes in relation to the endocrine status of the animal in both species. This plastic relationship is most evident in the median eminence, in which tanycytes control the access of GnRH nerve terminals to the portal vasculature via end feet specializations that alternatively advance or

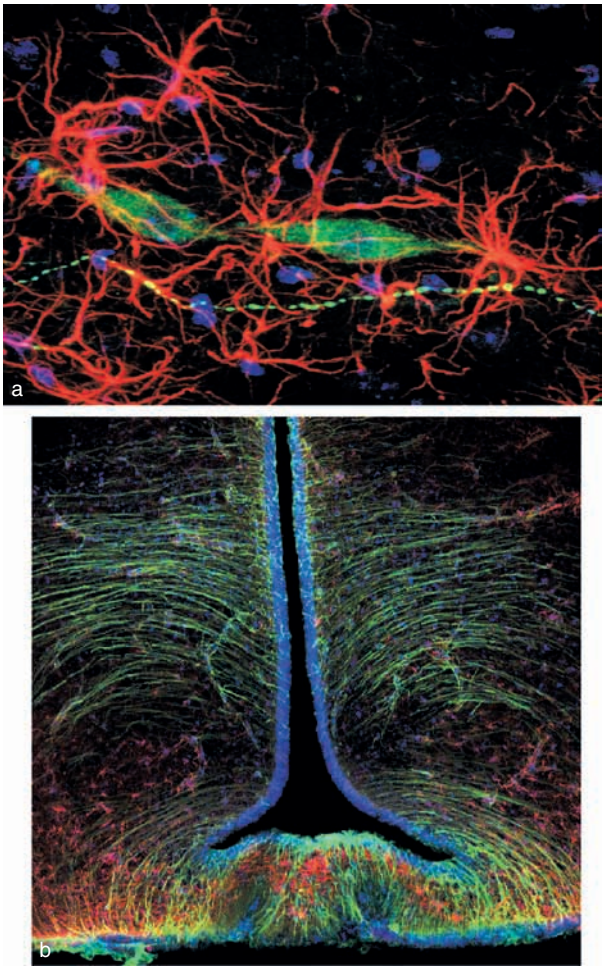

Figure 2 Immunohistofluorescence confocal microscopy images showing astroglia–GnRH neuron and astroglia–tanycyte associations in the prepubertal female rat hypothalamus. (a) Astrocytes identified by their glial fibrillary acid protein (GFAP) staining (red) wrap GnRH cell bodies and neuronal processes (green) in the preoptic area. (b) Tanycytes (identified by the presence of vimentin (green)) line the floor of the third ventricle and send long processes to the external layer of the median eminence to contact the endothelial wall of the portal system. Astrocytes (identified by the presence of GFAP (red)) are located immediately below the layer of tanycytic cell bodies, with which they establish intimate contact and also send processes to the external layer of the median eminence.

retract according to the stage of the estrous cycle. When estradiol levels are high, preceding the preovulatory surge of gonadotropins, the end feet retract, allowing direct access of the GnRH nerve ending to the portal system and thus resulting in unhindered release of the decapeptide to the portal system. In addition to plastic rearrangements, glial cells contribute to the control of GnRH release by releasing growth factors and other chemical messengers, and by amplifying the facilitatory effects of glutamatergic neurotransmission on GnRH neurons.

Glial Growth Factors

Astroglial cells synthesize and release several growth factors, including transforming growth factor β_1 (TGF-β_1), basic fibroblast growth factor (bFGF), insulin-like growth factor (IGF-1), and members of the epidermal growth factor (EGF) family. In this article, we consider only the latter. It is well-established that the EGF relatives, transforming growth factor-α (TGF-α) and the neuregulins (NRGs), are key components of this glia-to-neuron communication pathway. By binding to glial erbB receptors (erbB1 and erbB4, respectively), TGF-α and NRGs recruit an auxiliary receptor termed erbB2 and elicit the release of chemical messengers, such as prostaglandin E_2 (PGE_2), that act directly on GnRH neurons to stimulate GnRH secretion.

Various approaches have been used to demonstrate the involvement of this erbB receptor-mediated astrocytic signaling system in the control of female sexual development. Pharmacological blockade of erbB1 receptors or a natural mutation in the erbB1 gene result in delayed puberty. Conversely, transgenic or anatomically targeted (to the median eminence) TGF-α overexpression advances female puberty. The TGF-α–erbB1 signaling complex is also a key determinant of the plastic changes that tanycytes exhibit during the preovulatory surge of GnRH. Activation of erbB1 receptors in tanycytes results in release of PGE_2, which enhances TGF-β_1 release. In turn, TGF-β_1 induces tanycytic retraction by facilitating dissolution of the extracellular matrix.

The importance of erbB2 for appropriate hypothalamic erbB-mediated signaling has been shown by the delay in puberty resulting from the *in vivo* disruption of hypothalamic erbB2 receptor synthesis via infusion of an antisense oligodeoxynucleotide.

Finally, involvement of a glial NRG–erbB4 signaling complex in the control of female puberty was initially suggested by the presence of both the ligands and the receptor in astrocytes of the immature female rat hypothalamus and then by the increase in hypothalamic content of erbB2 and erbB4 mRNA that occurs at the end of the juvenile period. This increase occurs first in a gonad-independent manner and is subsequently driven by ovarian steroids, suggesting that the initial activation is a component of the centrally mediated, steroid-independent change in hypothalamic activity (the 'central drive') that initiates puberty. Transgenic animals expressing a dominant mutated form of the erbB4 receptor targeted to astrocytes exhibit a reduced astrocytic (PGE_2) response to NRG stimulation, a blunted GnRH response to astrocyte-conditioned medium, a decrease in plasma gonadotropin levels, and delayed puberty. The use

of double mutant animals carrying both a dominant negative form of erbB4 and a point mutation of the erbB1 receptor demonstrated that both receptors act coordinately to control GnRH secretion and, hence, female sexual development.

Glutamate

Astrocytes control extracellular glutamate levels in a variety of ways, including uptake reversal of glutamate transporters, the Cl^-/HCO_3^- anion exchange system, calcium-dependent exocytosis, and diffusion through ionotropic purinergic receptors. Astrocytes can also modify intracellular glutamate levels directly via astrocyte–astrocyte gap junction communication. The identification of these mechanisms has led to the recognition of astrocytes as a third component of synaptic communication, in addition to the pre- and postsynaptic neuronal specializations.

Astrocytes also play a major role in glutamate metabolism. During the onset of female puberty, there is an increased ability of the hypothalamus to release glutamate. There are also opposite changes in the hypothalamic content of glutamate dehydrogenase (GDH), an enzyme responsible for the synthesis of glutamate, and glutamine synthetase (GS), the enzyme that catalyzes the metabolism of glutamate to glutamine. GDH content increases and GS decreases at the time of puberty, whereas the abundance of the neuronal-specific glutamate synthesizing enzyme phosphate-activated glutaminase remains unchanged. Because both GDH and GS are predominantly expressed in astrocytes, these results suggest that a glial glutamatergic input to GnRH neurons is an intrinsic component of the hypothalamic regulatory system controlling the onset of female puberty.

Small Bioactive Molecules

As indicated previously, PGE_2 mediates the stimulatory actions that TGF-α and NRGs exert on GnRH release. Glutamate also directly stimulates astrocytic release of PGE_2, which in turn promotes a calcium-dependent glial release of glutamate. In addition, glutamate can increase its own availability by enhancing the release of astroglial arachidonic acid, which inhibits glutamate uptake into glial cells. Another small molecule released by astrocytes in a calcium-dependent manner is ATP. GnRH neurons respond to extracellular ATP, through $P2X_2$ and $P2X_4$ receptors, with an immediate increase in intracellular calcium and release of GnRH. These interactions are likely to be important for the regulation of GnRH secretion during sexual development.

Neuron-to-Glia Communication

During the past few years, it has become clear that the activity of glial cells is influenced by neuronal inputs. In the neuroendocrine brain, stimulation of astrocytic glutamatergic receptors induces the activation of the erbB signaling system. Hypothalamic astrocytes express both metabotropic and ionotropic (AMPA) glutamatergic receptors. Combined metabotropic/AMPA stimulation results in phosphorylation of erbB receptors. This activation requires the activity of a metalloproteinase that promotes the cleavage of both TGF-α and NRG precursors, making them available for interaction with their respective receptors. In the case of TGF-α, the metalloproteinase involved is termed tumor necrosis factor-α converting enzyme (TACE). TACE activity increases in the rat median eminence at the time of the first preovulatory surge of gonadotropins, suggesting a role for TACE in the control of GnRH release. Consistent with this view, blockade of TACE activity in the median eminence decreases GnRH secretion and delays puberty in the female rat.

It is therefore evident that glutamate controls GnRH neuronal activity not only via a transsynaptic mode of communication but also by utilizing the astrocytic erbB signaling system coupled to PGE_2 release to facilitate GnRH secretion.

Transcriptional Control of the Onset of Puberty

The complexity of the aforementioned cell–cell communication pathways suggests a level of control provided by gene networks that, operating within the neuroendocrine brain, coordinate neuronal–glial interactions in a developmentally regulated manner. Regulatory control systems such as this have the potential for thousands of interactions. Nevertheless, the 'heart' of such complex systems always consists of a handful of genes encoding transcription factors. Little is known about the hierarchy of such 'upstream' transcriptional regulators controlling the onset of mammalian puberty, but three potential candidates have been identified. One is Oct-2, a transcriptional regulator of the POU domain family of homeobox-containing genes. Because Oct-2 proteins are more abundant in cultured astrocytes than in neurons, they may be important for the transregulation of astroglial genes. TGF-α is one of these target genes because the transcriptional activity of the TGF-α promoter is increased by Oct-2. The physiological implications of this relationship have been made evident by several observations, including the gonad-independent increase in hypothalamic Oct-2 mRNA levels that occurs during juvenile development, the concomitant activation of both Oct-2 and TGF-α expression in astrocytes near the site of puberty-advancing lesions of the anterior hypothalamus, and the reduction in astrocytic TGF-α synthesis associated with a delay in the age at first ovulation resulting from

blockade of hypothalamic Oct-2 synthesis with antisense oligodeoxynucleotides.

The second candidate is another homeobox gene known as TTF-1 (NKx2.1). TTF-1 is required for diencephalic morphogenesis; like Oct-2, it remains expressed after birth in discrete neuronal and glial populations of the hypothalamus, where it acts to regulate cell-specific functions. For instance, TTF1 enhances GnRH, KiSS1, and erbB2 promoter activity, but it inhibits preproenkephalin gene transcription. Neuron-specific deletion of the TTF-1 gene results in delayed puberty and decreased reproductive capacity – deficiencies associated with increased preproenkephalin mRNA and reduced GnRH and KiSS1 mRNA levels in the hypothalamus.

A third candidate was discovered using cDNA arrays examining the primate hypothalamus at the time of puberty. This gene, previously known as chromosome 14 open reading frame 4 (C14ORF4), has been termed EAP-1 (enhanced at puberty). EAP-1 encodes a nuclear protein that, like TTF-1, is expressed in neurons involved in the stimulatory and inhibitory control of GnRH secretion, including GnRH neurons, as well as glutamatergic, GABAergic, and proenkephalinergic neurons. EAP-1 also has dual transregulatory activity; it activates the GnRH promoter but represses the preproenkephalin gene, which is inhibitory to the pubertal process. Reducing hypothalamic EAP-1 expression using siRNAs delayed puberty and disrupted estrous cyclicity, indicating that EAP-1 is necessary for the timely activation of GnRH secretion at puberty.

These observations suggest that upstream genes, such as TTF-1, Oct-2, and EAP-1, control the pubertal process by coordinating the activity (and therefore the interactions) of neuronal and glial subsets involved in the developmental control of the GnRH neuronal network (Figure 3).

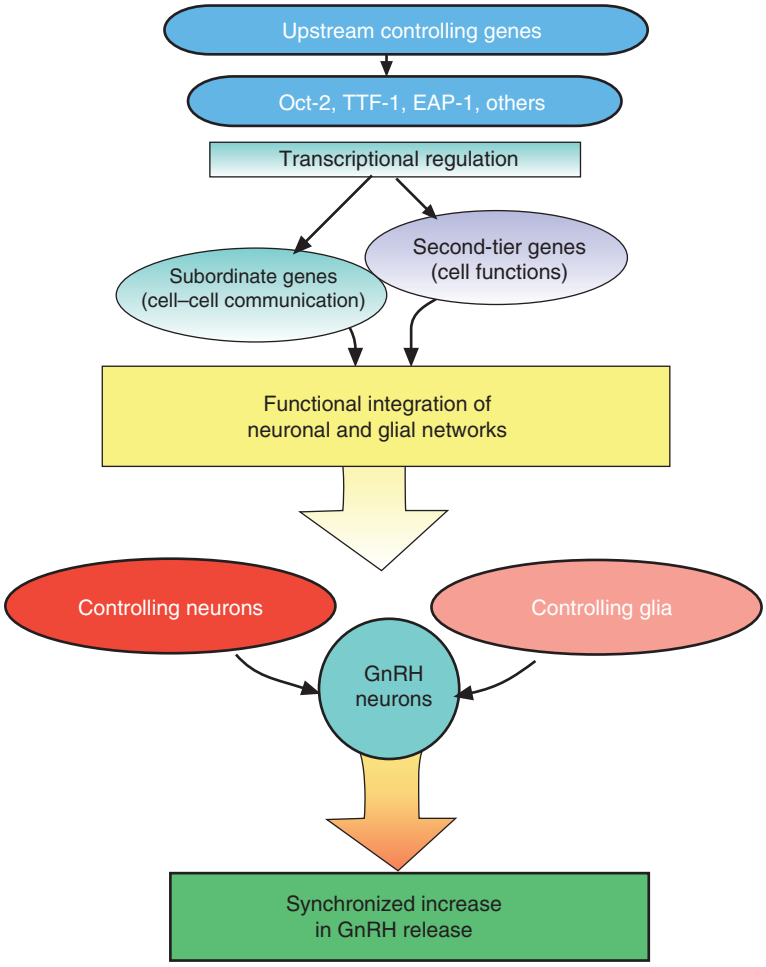

Figure 3 The transcriptional control of the GnRH neuronal network at puberty by upstream controlling genes. Changes in the secretory activity of GnRH neurons are specified by transsynaptic and glial inputs. These inputs are determined by subordinate genes, which in turn are controlled at the transcriptional level by upper-echelon genes such as Oct-2, TTF-1, and EAP-1. It is believed that this hierarchical arrangement is required to initiate and maintain an enhanced level of pulsatile GnRH secretion at puberty. Modified from Ojeda SR, Lomniczi A, Mastronardi C, et al. (2006) Minireview: The neuroendocrine regulation of puberty: Is the time ripe for a systems biology approach? *Endocrinology* 147: 1166–1174, copyright 2006, The Endocrine Society.

An Integrated View

The principal cellular interactions and cell–cell communication molecules believed to control the initiation of mammalian puberty are illustrated in Figure 4. As indicated previously, the onset of puberty appears to result from the coordinated activation of excitatory transsynaptic inputs to GnRH neurons (e.g., glutamate and kisspeptin) coupled with a reduction in inhibitory neurotransmission (e.g., GABA and opioid peptides) and an increase in growth factor-mediated glia-to-neuron signaling (e.g., TGF-α and NRGs). Absent from this model is the clock mechanism

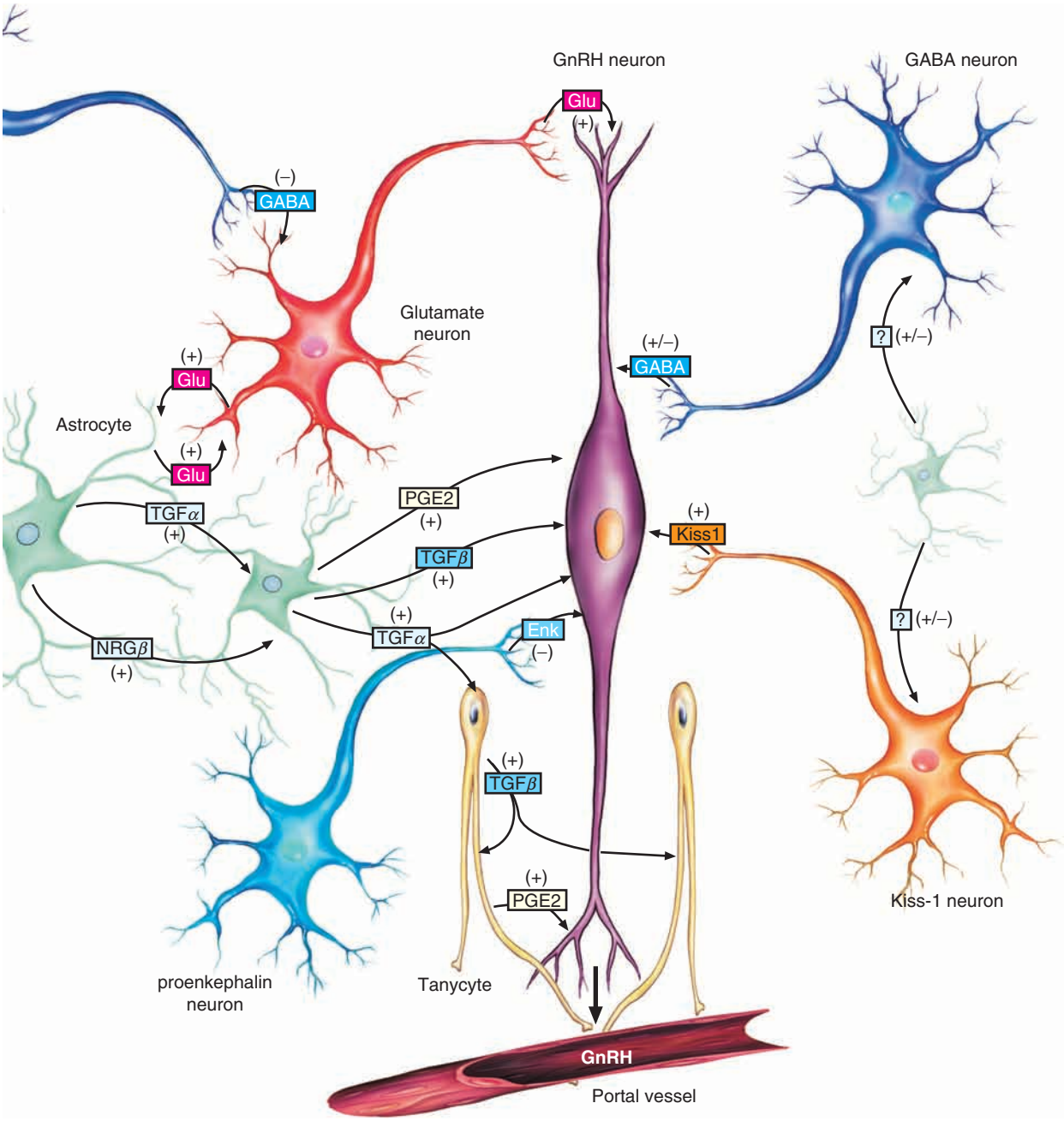

Figure 4 Neuronal and glial factors involved in the control of GnRH secretion. The transsynaptic control of GnRH secretion is exerted by excitatory neurons that utilize glutamate or kisspeptin for neurotransmission/neuromodulation and inhibitory neurons that employ GABA or opioid peptides (represented by preproenkephalinergic neurons) for inhibition. Astrocytes and ependymoglial cells stimulate GnRH secretion via release of growth factors that can act directly on GnRH neurons (e.g., TGF-β, bFGF, and IGF-1) or growth factors (e.g., TGF-α and NGRs) that need the intermediacy of small bioactive molecules (e.g., PGE_2). At the time of puberty, the excitatory component of this regulatory network predominates over the inhibitory component, resulting in enhanced GnRH secretion. +, stimulation; −, inhibition; ?+/−, not known.

that coordinates the changes in diurnal activity of this regulatory system. Such an entrainment is characteristic of the pubertal process in both primates and rodents because in these species the GnRH-dependent activation of gonadotropin secretion has a diurnal rhythm (Figure 1). Emerging evidence suggests that integrations of neuron-to-neuron and reciprocal glia–neuron interactions leading to the initiation of the pubertal process require a higher level of control provided by a transcriptional machinery composed of genes able to modify gene expression in a cell- and promoter-dependent manner (Figure 3).

See also: Gonadal Steroid Actions on Brain; Gonadotropin, Neural and Hormonal Control; Gonadotropin-Releasing Hormone: GnRH-1 System; Kisspeptins and their Receptors; Neuroendocrinology; Neurosteroids; Sexual Behavior: Neuroendocrine Control.

Further Reading

Bourguignon J-P, Lebrethon MC, Gérard A, et al. (2000) Amino acid neurotransmission and early ontogeny of pulsatile GnRH secretion from the rat hypothalamus. In: Bourguignon J-P and Plant TM (eds.) *The Onset of Puberty in Perspective*, pp. 119–129. Amsterdam: Elsevier.

Clarkson J and Herbison AE (2006) Development of GABA and glutamate signaling at the GnRH neuron in relation to puberty. *Molecular and Cellular Endocrinology* 254/255: 32–38.

de la Iglesia HO and Schwartz WJ (2006) Minireview: Timely ovulation: Circadian regulation of the female hypothalamo-pituitary-gonadal axis. *Endocrinology* 147: 1148–1153.

Dungan HM, Clifton DK, and Steiner RA (2006) Minireview: Kisspeptin neurons as central processors in the regulation of gonadotropin-releasing hormone secretion. *Endocrinology* 147: 1154–1158.

Grumbach MM (2002) The neuroendocrinology of human puberty revisited. *Hormone Research* 57(supplement 2): 2–14.

Ojeda SR, Lomniczi A, Mastronardi C, et al. (2006) Minireview: The neuroendocrine regulation of puberty: Is the time ripe for a systems biology approach? *Endocrinology* 147: 1166–1174.

Ojeda SR, Lomniczi A, Mungenast A, et al. (2005) Towards understanding the neurobiology of mammalian puberty: Genetic, genomic and proteomic approaches. In: Kordon C, Gaillard R, and Christen Y (eds.) *Hormones and the Brain*, pp. 47–60. Berlin: Springer.

Ojeda SR and Skinner MK (2006) Puberty in the rat. In: Neill JD (ed.) *The Physiology of Reproduction,* 3rd edn., pp. 2061–2126. San Diego: Academic Press.

Palmert MR and Hirschhorn JN (2003) Genetic approaches to stature, pubertal timing, and other complex traits. *Molecular Genetics and Metabolism* 80: 1–10.

Parent AS, Teilmann G, Juul A, Skakkebaek NE, Toppari J, and Bourguignon JP (2003) The timing of normal puberty and the age limits of sexual precocity: Variations around the world, secular trends, and changes after migration. *Endocrine Reviews* 24: 668–693.

Plant TM and Witchel SF (2006) Puberty in nonhuman primates and humans. In: Neill JD (ed.) *The Physiology of Reproduction,* 3rd edn., pp. 2177–2230. San Diego: Academic Press.

Sisk CL and Foster DL (2004) The neural basis of puberty and adolescence. *Nature Neuroscience* 7: 1040–1047.

Terasawa E and Fernandez DL (2001) Neurobiological mechanisms of the onset of puberty in primates. *Endocrine Reviews* 22: 111–151.

Terasawa E, Keen KL, Grendell RL, and Golos TG (2005) Possible role of 5′-adenosine triphosphate inx synchronization of Ca^{2+} oscillations in primate luteinizing hormone-releasing hormone neurons. *Molecular Endocrinology* 19: 2736–2747.

Whitlock KE (2005) Origin and development of GnRH neurons. *Trends in Endocrinology and Metabolism* 16: 145–151.

Neuroendocrinology of Social/Affiliative Behavior

C D Cole and L J Young, Emory University School of Medicine, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Affiliative behavior can be defined as any prosocial behavior that brings two or more individuals into close proximity. The formation of social units, such as mating pairs, family units, colonies, or even societies is critical for the survival of most species, including humans. Affiliative behavior depends on several neurobiological processes, including the ability to recognize familiar individuals based on social stimuli, the motivation to engage in social interactions, and the formation of selective social bonds. In recent years, great strides have been made in understanding the underlying neurobiological processes that drive affiliative behaviors, including social bond formation. Much attention has focused on the role of the neuropeptides oxytocin (OT) and arginine vasopressin (AVP) in the regulation of specific types or components of affiliative behavior, including social recognition, parental care, and social bonding. This article discusses the roles of these peptides as well as other neuroendocrine factors in the regulation of affiliative behaviors in animal models, and it concludes with a brief discussion of the implications for this research on human behavior.

Oxytocin and Vasopressin: Neuropeptides of Affiliation

OT and AVP are closely related nonapeptides that are synthesized predominantly in magnocellular neurons of the paraventricular nucleus (PVN) and supraoptic nucleus (SON) of the hypothalamus, which project heavily to the posterior pituitary, where these peptides are secreted into the bloodstream. Peripherally released OT facilitates uterus contractions during parturition and milk ejection during lactation. Peripherally released AVP, also known as antidiuretic hormone, controls water reabsorption in the kidney and vascular tone. OT and AVP are also synthesized in parvocellular neurons in the PVN as well as extrahypothalamic sites, such as the medial amygdala and the bed nucleus of the stria terminalis, which project centrally and are thought to mediate the central effects of the peptides. Because of the blood–brain barrier and independent release, the peripheral and central effects of OT and AVP are considered to be independent. Interestingly, the production of AVP in certain brain regions is sexually dimorphic and highly dependent on testosterone such that males have significantly more AVP than females within these specific circuits. The actions of OT and AVP are mediated by four G-protein-coupled receptors: oxytocin receptor (OTR) and V1a, V1b, and V2 subtypes of the vasopressin receptor. The V1a and V1b subtypes are localized in the brain, and the majority of the behavioral effects of AVP appear to be regulated by the V1a receptor (V1aR).

Centrally released OT and AVP have been implicated in the regulation of a variety of social behaviors, including parental care, social recognition, sexual behavior, and aggression in a wide range of vertebrate taxa. Not surprisingly, due to its sexually dimorphic expression in the brain, AVP has predominately been implicated in the modulation of male-typical social behaviors. This article focuses on the roles of OT and AVP and other neuroendocrine factors in animal models of social recognition and parental care in rodents, mother–infant bonding in sheep, pair bond formation in prairie voles, and the potential role of OT and VP in human social behavior.

Social Recognition in Rodents

Social recognition is best defined as the ability to recognize a familiar conspecific based on salient social cues. This process is the foundation for the formation of all social relationships and is therefore necessary for the survival of a species. Visual and auditory cues are primarily used in human and nonhuman primate social recognition, whereas olfactory cues are predominantly used to differentiate familiar and nonfamiliar individuals in rodents.

Since rodents rely on olfactory cues to distinguish individuals, social recognition in rodents can be assessed by determining the amount of time an animal spends investigating a familiar individual upon subsequent exposures or the relative amount of time an individual spends investigating a novel compared to a familiar individual. This social recognition test is based on the fact that rodents spend less time investigating familiar individuals compared to novel individuals. Upon repeated exposures to the same individuals, rodents typically habituate to the familiar individual, resulting in decreased duration of olfactory investigation.

The role of AVP in social recognition was first investigated using Brattelboro rats, which are natural mutants for AVP. Male Brattelboro rats display severe deficits in social recognition, but normal social recognition is

restored by infusion of AVP into the lateral septum. Later studies reported that V1aR selective AVP antagonist infused into the lateral septum of normal rats blocks social recognition, whereas infusion of AVP into the lateral septum increases the duration of the memory from less than 2 h to more than 24 h. AVP appears to be involved in the consolidation or recall of the social memory since AVP is necessary only after the initial social encounter.

Several molecular and transgenic studies have confirmed the role of AVP in modulating social recognition. Infusion of a V1aR antisense oligonucleotide into the lateral septum of rats decreases V1aR density and inhibits social recognition. Artificially increasing V1aR expression in the lateral septum of rats using a viral vector enhances social recognition. Finally, V1aR knockout mice fail to display social recognition since they do not habituate to familiar individuals. These mice perform normally on nonsocial learning and memory tasks, such as the Morris water maze, as well as tests for sensorimotor gating, such as prepulse inhibition, indicating that nonsocial cognitive processes are intact. This deficit is also not due to a general disruption in olfaction since V1aR knockout mice habituate to nonsocial odors. Social recognition in these animals can be restored by expressing the V1aR in the lateral septum using viral vector-mediated gene transfer.

OT is also required for social recognition, although it appears to have a different function than AVP. OT antagonists inhibit social recognition, and mice genetically deficient in the OT gene fail to habituate to familiar individuals upon repeated exposures. Like V1aR knockout mice, OT knockout males display normal olfactory memory for nonsocial scents and perform normally in nonsocial cognitive tasks, confirming that this deficit is selective for social learning and memory. This social recognition deficit was reversible with a single central injection of OT before, but not after, the initial encounter with a stimulus female. Thus, unlike AVP, OT appears to be critical for the neural processing of the social cues required for the formation of the memory rather than consolidation or recall. Furthermore, OT appears to be acting in the medial amygdala, rather than in the lateral septum, since infusion of OT in this region restores social recognition (Figure 1).

Studies of female OT knockout mice suggest that OT is also involved in assessing mate quality. Female mice use olfaction to distinguish between suitable, healthy mates and those infected with parasites. Interestingly, OT knockout females fail to display a preference for uninfected males, further suggesting that OT is involved in the processing of social cues and potentially in mate selection.

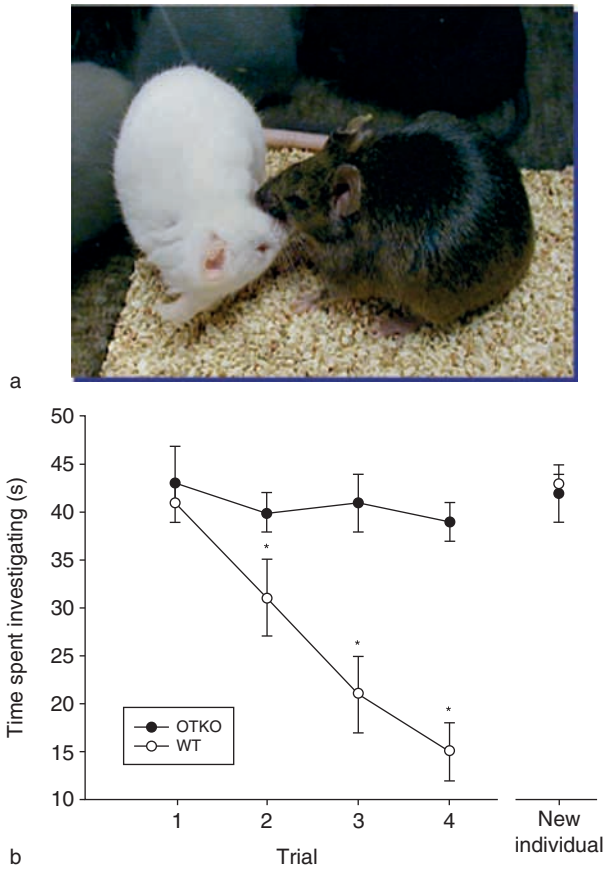

Figure 1 Social recognition in wild-type (WT) and oxytocin knockout (OTKO) male mice. (a) Olfactory investigation in mice. During a social recognition test, the duration of time in which the experimental mouse investigates the stimulus mouse is quantified after multiple exposures. Typically, WT mice will show a decline in investigation time after subsequent exposures. (b) Olfactory investigation of ovariectomized female mice by WT and OTKO male mice. Olfactory investigation is measured during four successive 1 min trials separated by 10 min. A fifth dishabituation trial using a novel female demonstrates that the decline in investigation in WTs is due to recognition of the initial female. A significant decrease (asterisks) occurred between each trail for WT males, but no decline was detected in the OTKO males. Data depict mean ± 1 SEM. Reprinted by permission from Macmillan Publishers Ltd: *Nature Genetics* (Ferguson JN, Young LJ, Hearn EF, et al. (2000) Social amnesia in mice lacking the oxytocin gene. *Nature Genetics* 25(3), 284–288), copyright 2000.

Neuroendocrine Control of Parental Behavior in Rodents

Maternal Care in Rats and Mice

The proper regulation of parental care is critical for reproduction in all mammals. Virgin female rats, like most mammals, show no interest in or even avoid pups, but at parturition they rapidly switch to displaying high levels of maternal behavior and being attracted to pups. This change in behavior represents a change in motivational state rather than

the development of selective social attachments since female rats will readily adopt any pup that they encounter.

The neuroendocrine changes associated with pregnancy and parturition play a critical role in the induction of maternal responsiveness. Circulating estrogen, progesterone, and prolactin concentrations increase during pregnancy, followed by a decrease in progesterone secretion at parturition. These hormonal fluctuations are critical not only for priming the uterus and mammary tissue in preparation for birth but also for the induction of maternal care through their actions on receptors in specific brain regions. Priming with estrogen and progesterone, followed by progesterone withdrawal, as well as infusion of prolactin, stimulates the initiation of maternal care in ovariectomized, virgin rats. Interestingly, these hormones are not essential for the maintenance of maternal care because once maternal behavior has been established, it is not diminished by removal of these hormones.

The medial preoptic area (MPOA) plays a significant role in the neuroendocrine regulation of maternal behavior. Lesions of the MPOA or its projections disrupt the onset of maternal behavior. The MPOA has a high concentration of estradiol, and progesterone receptors and prolactin receptors increase in the MPOA as parturition approaches. Implants of estradiol into the MPOA facilitate maternal behavior. Studies on knockout mice confirm the role of estrogen and prolactin in regulating the onset of maternal behavior. Estrogen receptor-α knockout mice have poor maternal behavior, whereas estrogen receptor-β knockout mice successfully raise their young, suggesting that estrogen receptor-α is the substrate with which estrogen interacts in the MPOA. Prolactin receptor knockout mice also have deficits in pup retrieval and nesting deficits.

In addition to estrogen, progesterone, and prolactin, centrally released OT is involved in the initiation of maternal responsiveness in rodents. OT is released from the posterior pituitary into the bloodstream at parturition and during lactation via the synchronous firing of magnocellular PVN and SON neurons. More than 25 years ago, two studies demonstrated that central infusion of OT facilitated the onset of rapid maternal behavior in virgin female rats. Later studies confirmed that OT antagonists and lesions of the PVN delay the onset of maternal behavior in primiparous female rats. However, PVN lesions do not disrupt maternal behavior in animals that are already maternal, again demonstrating that neuroendocrine factors mediate the onset but not the maintenance of maternal behavior. Finally, OT receptor antagonists injected directly into the MPOA inhibit maternal behavior.

OT also appears to be acting in the olfactory bulb to modulate the olfactory response to pup odors. Infusion of OT into the olfactory bulbs of parturient rats facilitates the onset of maternal behavior, whereas the application of an OT antagonist inhibits the onset of maternal behavior. Initial reports on OT knockout mouse suggested that maternal behavior was not disrupted. However, more careful analysis of their maternal behavior demonstrated that these mutant mice do in fact have impaired maternal responsiveness. In addition, OT receptor knockout mice display profound impairments in maternal behavior.

Spontaneous Maternal Care in Prairie Voles

Prairie voles are monogamous rodents that display high levels of affiliative behavior and have thus become important models for investigating the neurobiology of affiliation. Interestingly, juvenile female prairie voles, and approximately half of adult female prairie voles, display 'spontaneous' maternal care as virgins. OT receptor density in the nucleus accumbens is correlated with the quality of spontaneous maternal care both in juveniles and in adults. That is, individual prairie voles with higher densities of OT receptor in the nucleus accumbens display higher levels of maternal care. Furthermore, infusion of OT receptor antagonist into the nucleus accumbens inhibits spontaneous maternal care in female prairie voles. Interestingly, mice and meadow voles do not show spontaneous maternal care as juveniles, and they also have little or no OT receptors in the nucleus accumbens. In contrast, juvenile rats, which do display maternal behavior prior to puberty, have moderate levels of OT receptors in the accumbens, but these levels decrease as the developmental pattern of maternal behavior declines.

Paternal Care

Monogamous mammals, representing only 3–5% of mammalian species, display paternal care in addition to maternal care. Several studies have implicated a role for AVP in the regulation of paternal behavior, particularly in male prairie voles. Male prairie voles display an increase in paternal behavior after cohabitating with a female. AVP immunoreactivity in the lateral septum decreases and AVP mRNA in the bed nucleus of the stria terminalis increases concomitantly with these changes in behavior. This is consistent with a release of AVP from terminals in the septum and an increase in AVP synthesis in the neurons giving rise to those terminals. A more direct indication of a role for AVP in inducing paternal behavior came from studies in which infusions of AVP into the lateral septum facilitated, and a V1aR

antagonist blocked, paternal care in male prairie voles. In a separate study, male prairie voles with elevated levels of V1aR expression in the lateral septum displayed higher levels of paternal care than males with lower levels of receptor in the same region. The role of AVP in regulating paternal behavior has not been explored in other species; however, one study found a dramatic increase in V1aR protein in the cortex of monogamous male marmosets associated with paternal experience.

Social Bonding

Social bonding is a special form of affiliative behavior in which selective social attachments strengthen social relationships. Social bonding requires both social recognition processes and motivational systems to be activated. In many species, neither mating pairs nor parent–infant relationships develop selective bonds. However, two models of social bonding – mother–infant bonding in sheep and pair bonding in monogamous prairie voles – have provided significant insights into the neuroendocrine regulation of social bonding.

Mother–Infant Bonding in Sheep

As in rats, hormonal fluctuations during pregnancy and parturition facilitate the onset of maternal behavior in sheep. However, unlike rats, the ewe forms a selective bond with her lamb such that she provides maternal nurturing only to her own lamb and rejects other lambs. This bonding process is necessary for sheep since many ewes give birth to precocial (i.e., mobile) lambs simultaneously, and it ensures that all maternal resources are directed to the ewe's offspring. The selective bond between the ewe and the lamb is formed within a few hours after birth, after which the mother rejects all other lambs.

OT has been shown to be critical for the formation of mother–infant bonds in sheep. Estrogen and progesterone stimulation increases maternal responsiveness in sheep, but this behavior is disorganized and slow to develop. However, vaginocervical stimulation in these same animals results in the rapid acceptance of a foreign lamb, often within 5 min. Vaginocervical stimulation, either artificially or during parturition, results in the central release of OT in the brain and the olfactory bulb. Finally, central infusion of OT in a hormonally primed nonpregnant ewe will stimulate her to form a bond with an unfamiliar lamb within a few minutes. After this bonding process, the ewe will behave maternally to the bonding lamb and reject all other lambs.

The olfactory bulb appears to be a critical brain area responsible for developing a social memory for a lamb. OT release in the olfactory bulb during parturition is thought to synaptically recognize neurons in the mitral cell layer to selectively respond to the familiar lamb. In addition to OT, noradrenaline, acetylcholine, glutamate, and GABA levels increase following parturition or vaginocervical stimulation. It appears that OT modulates the release of other neurotransmitters, suggesting that the release of OT coordinates the changes in the olfactory bulb after parturition or vaginocervical stimulation.

Pair Bonding in Monogamous Prairie Voles

Like sheep, the monogamous prairie vole (*Microtus ochrogaster*) forms selective, enduring bonds. These bonds, however, are with their mates and can last the lifetime of the pair. Prairie voles are highly social and typically live in communal burrows occupied by a breeding pair and multiple generations of offspring. Laboratory studies suggest that mating facilitates pair bond formation, although mating is not strictly required. Because of its role in maternal bonding, it was hypothesized that OT may play a role in pair bonding in voles. In the laboratory, pair bonding is assessed using a partner preference paradigm. After a period of cohabitation, one of the pair is tethered to one chamber of a three-chambered testing arena. A novel animal of the same sex and stimulus value is tethered to the opposite chamber. The test animal is placed in the central chamber and is allowed to move freely through all three chambers. Typically, a prairie vole that has formed a pair bond will spend more than twice as much time in contact with the partner than with the stranger.

Infusions of OT in female prairie voles result in the development of a partner preference in the absence of mating. Likewise, mating-induced partner preferences are blocked by central infusion of an OTR antagonist in females. Similarly, central injections of AVP into the brain facilitate pair bond formation in the absence of mating and V1aR antagonist blocks pair bond formation in male prairie voles.

Unlike prairie voles, montane voles (*M. montanus*) are non-monogamous, relatively asocial, and do not typically form social bonds. Both species have similar levels and distributions of OT and AVP in the brain. However, infusions of either peptide do not stimulate affiliative behavior in montane voles. In contrast to the peptides, the distribution of OT and AVP receptors in the brain differs dramatically between these species, suggesting that receptor distribution may be responsible for the differences in behavior (**Figure 2**).

Prairie voles have high densities of OT receptor and V1aR in brain regions implicated in reward and reinforcement, such as the nucleus accumbens and ventral pallidum, respectively. In contrast, non-monogamous

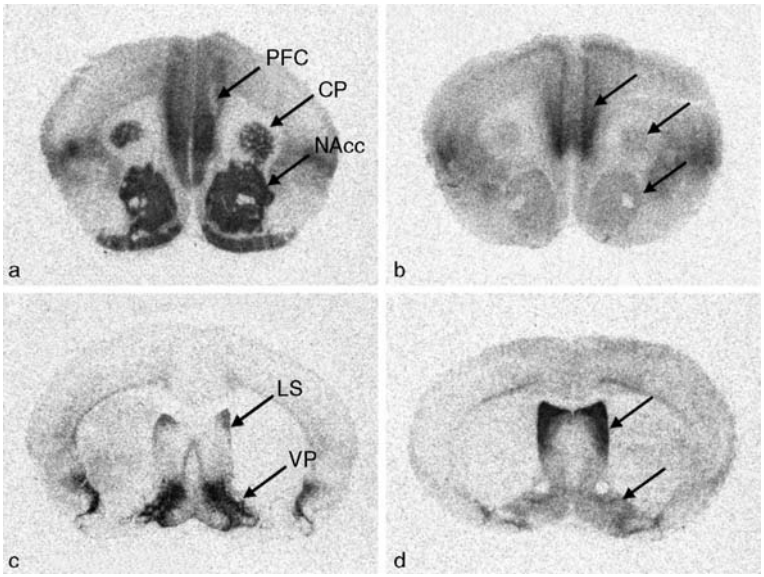

Figure 2 Oxytocin receptor (a, b) and vasopressin 1a receptor (c, d) binding in the brains of monogamous prairie voles (a, c) and non-monogamous montane voles (b, d). Prairie voles have higher densities of oxytocin receptors in the nucleus accumbens (NAcc) than montane voles. Additionally, prairie voles have higher densities of vasopressin receptors in the ventral pallidum (VP) than do montane voles. CP, caudate putamen; LS, lateral septum; PFC, prefrontal cortex. Adapted from Young LJ and Wang Z (2004) The neurobiology of pair bonding. *Nature Neuroscience* 7: 1048–1054.

montane and meadow voles have low levels of receptors in these regions. Microinjection of OT receptor antagonist in the nucleus accumbens blocks pair bond formation in female prairie voles, whereas microinjection of V1aR antagonist into the ventral pallidum blocks pair bonding in male prairie voles. Artificially elevating V1aR levels in the ventral pallidum of the non-monogamous meadow vole (*M. pennsylvanius*) was found to enhance mating-induced partner preference formation in this species. Thus, it appears that differences in mating strategies in these species may be due in part to the species differences in OT receptor and V1aR expression patterns in the brain.

The involvement of the nucleus accumbens in pair bond formation suggests a role for the mesolimbic dopamine reward pathway in pair bond formation. Extracellular dopamine levels rise in the nucleus accumbens during mating. Infusion of dopamine D2-like receptor agonist and antagonist into the nucleus accumbens of prairie voles facilitates and blocks partner preference, respectively. However, simultaneous blockade of OT receptors in females prevents dopamine agonist-induced partner preference formation in females. This suggests that the OT and dopamine systems interact, resulting in pair bond formation.

Since OT and AVP are critically involved in social recognition processes and dopamine is important in reward conditioning, it has been proposed that pair bond formation results from the convergent activation of social recognition circuits (necessary for long-term memory of the mate) and the reward pathway. That is, simultaneous activation of both circuits during mating results in a conditioned association between the olfactory signature of the partner and the rewarding nature of mating, resulting in a conditioned partner preference (Figure 3).

Genetic studies suggest that individual differences in V1aR gene structure and expression in the brain contribute to individual differences in affiliative behavior. In the wild, not all male prairie voles form bonded pairs, and there are significant individual differences in brain V1aR patterns. Comparison of the V1aR gene structure between prairie and montane voles revealed a highly repetitive microsatellite element in the promoter of the gene which differed significantly between the species. Microsatellite elements are highly unstable and prone to mutation. There is significant individual variation in the length of the microsatellite among individual prairie voles. Interestingly, males with long alleles of the microsatellite have higher levels of V1aR expression in the olfactory bulb and lateral septum than do males with short alleles. In addition, males with longer microsatellites also display higher levels of paternal behavior and are more likely to form partner preferences than males with shorter microsatellites. These data suggest that variation in V1aR expression may underlie both individual and species differences in affiliative behavior in voles.

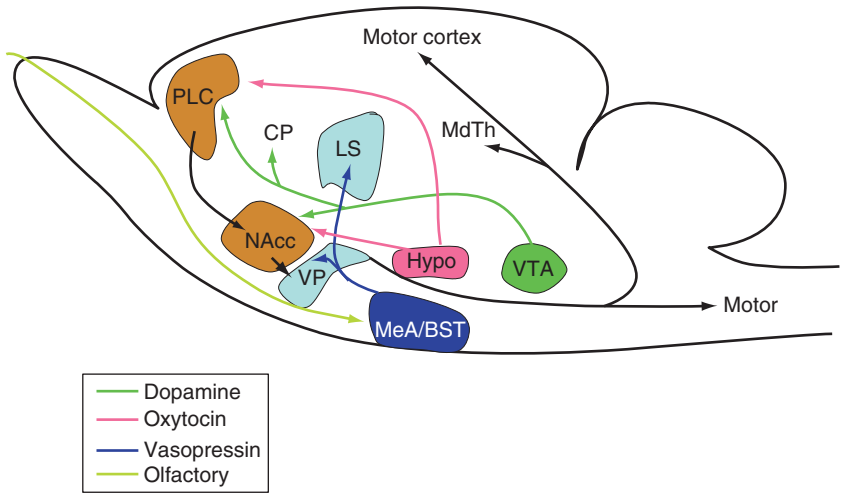

Figure 3 Proposed neural circuitry underlying pair bonding in prairie voles. Mating activates the ventral tegmental area (VTA), resulting in increased dopamine activity in the prefrontal cortex (PLC) and nucleus accumbens (NAcc). Concurrently, olfactory signals from the mate are transmitted via olfactory bulb to the medial nucleus of the amygdala (MeA). Mating also stimulates increased extracellular concentrations of oxytocin in the PFC and NAcc of females and vasopressin in the ventral pallidum (VP) of males. Vasopressin fibers in the lateral septum (LS) and VP originate from cell bodies in the MeA and bed nucleus of the stria terminalis (BST). Oxytocin projections to the NAcc and PFC most likely arise from cells in the hypothalamus (Hypo). The concurrent activation of the dopaminergic system and the oxytocin (in females) or arginine vasopressin (in males) system in the NAcc or VP potentially results in the development of a conditioned partner preference. The VP is a major output relay of the NAcc and modulates motor output in response to reinforcing stimuli via projects to the mediodorsal thalamus (MdTh) and cortical and mesencephalic motor nuclei. Adapted from Young LJ and Wang Z (2004) The neurobiology of pair bonding. *Nature Neuroscience* 7: 1048–1054.

Implications for Human Social Behavior

It is impossible to directly translate findings in animal models to human behavior. However, the animal studies discussed in this article suggest that it is possible that OT and AVP may also modulate human social behavior and that disregulation in these systems may contribute to the behavioral phenotypes observed in psychiatric disorders associated with disruptions in social behavior, such as autism. Findings from a study that used intranasal infusion of OT suggest that OT increases trust in humans. A separate brain imaging study demonstrated that intranasal OT infusion alters amygdala activity in response to viewing socially salient scenes. OT infusion also acts synergistically with the presence of a partner to decrease the cortisol stress response resulting from a public speaking task.

One study has reported that autistic children have significantly lower levels of plasma OT compared to age-matched normal subjects. Interestingly, nursery-reared rhesus monkeys (raised by humans instead of their mothers), which have severely disrupted social behaviors, have lower levels of OT in the cerebrospinal fluid (CSF) than monkeys reared by their mothers. In another primate model, the relatively asocial pigtail macaques also had significantly lower levels of OT in CSF than the more social bonnet monkeys. Notably, studies examining the therapeutic value of OT in autism have reported that intravenous OT infusions reduce repetitive behaviors and enhance social cognition in adults with autism, although these studies need to be replicated.

Several studies suggest that central AVP may modulate human social behavior as well. Intranasally administered AVP altered human facial responses (as measured by electromyograms) to normally neutral social stimuli in a sex-specific manner, suggesting that AVP might modulate the processing of social stimuli in humans similar to that observed in rodent models. Polymorphisms in microsatellites in the human V1aR gene have also been associated with variation in human social behavior as well as psychopathology. Specifically, three independent studies have reported associations between these polymorphisms and autism. Thus, as in prairie voles, genetic variability in the V1aR gene may contribute to individual differences in affiliative behavior in humans.

Conclusions

Affiliation results from a complex set of behaviors that involve many cognitive processes, each of which is regulated by many neuroendocrine factors, neurotransmitters, and brain circuits. We have made

significant progress in understanding the neural mechanisms underlying specific examples of affiliative behavior in animals and are just beginning to understand how these same mechanisms may modulate human behavior. The studies of social recognition, parental care, and social bonding discussed in this article reveal specific roles for neuropeptides (OT, AVP, and prolactin), steroids (estrogen and progesterone), and dopamine. However, it should be recognized that a multitude of other hormones and neurotransmitters are acting in concert with these systems to produce the complex set of events that ultimately lead to the development of social relationships which are critical for our own well-being.

See also: Agonistic and Affiliative Signals: Resolutions of Conflict; Neuroendocrine Control: Maternal Behavior; Neuroendocrine Peptide Processing; Neuroendocrinology; Reward Neurophysiology and Primate Cerebral Cortex; Social Interaction Effects on Reward and Cognitive Abilities in Monkeys; Vasopressin/Oxytocin and Receptors.

Further Reading

Bartz JA and Hollander E (2006) The neuroscience of affiliation: Forging links between basic and clinical research on neuropeptides and social behavior. *Hormones and Behavior* 50: 518–528.

Bielsky IF, Hu SB, Szedga KL, Westphal H, and Young LJ (2004) Profound impairment in social recognition and reduction in anxiety-like behavior in vasopressin V1a receptor knockout mice. *Neuropsychopharmacology* 29: 483–493.

Bielsky IF, Hu SB, Szedga KL, Westphal H, and Young LJ (2005) The V1a vasopressin receptor is necessary and sufficient for normal social recognition: A gene replacement study. *Neuron* 47: 503–513.

Carter CS, DeVries AC, and Getz LL (1995) Physiological substrates of mammalian monogamy: The prairie vole model. *Neuroscience and Biobehavioral Review* 19: 303–314.

Dantzer R, Bluthe RM, Koob GF, and LeMoal M (1987) Modulation of social memory in male rats by neurohypophyseal peptides. *Psychopharmacology* 91: 363–368.

Ferguson JN, Aldag JM, Insel TR, and Young LJ (2001) Oxytocin in the medial amygdala is essential for social recognition in the mouse. *Journal of Neuroscience* 21: 8278–8285.

Ferguson JN, Young LJ, Hearn EF, et al. (2000) Social amnesia in mice lacking the oxytocin gene. *Nature Genetics* 25(3): 284–288

Ferguson JN, Young LJ, and Insel TR (2002) The neuroendocrine basis of social recognition. *Frontiers in Neuroendocrinology* 23: 200–224.

Hammock EAD and Young LJ (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* 308: 1630–1634.

Kendrick KM, Da Costa AP, Broad KD, et al. (1997) Neural control of maternal behavior and olfactory recognition of offspring. *Brain Research Bulletin* 44: 383–395.

Lim MM, Wang Z, Olazabal DE, et al. (2004) Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene. *Nature* 429: 754–757.

Lim MM and Young LJ (2006) Neuropeptidergic regulation of affiliative behavior and social bonding in animals. *Hormones and Behavior* 50: 506–517.

Numan M and Insel TR (2003) *The Neurobiology of Parental Behavior.* New York: Springer.

Rosenblatt JS, Mayer AD, and Giordano AL (1998) Hormonal basis during pregnancy for the onset of maternal behavior in the rat. *Psychoneuroendocrinology* 13: 29–46.

Winslow JT, Hastings N, Carter CS, Harbaugh CR, and Insel TR (1993) A role for central vasopressin in pair bonding in monogamous prairie voles. *Nature* 365: 545–548.

Young LJ and Wang Z (2004) The neurobiology of pair bonding. *Nature Neuroscience* 7: 1048–1054.

Neuroethics

S J Bird, Wrentham, MA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Neuroethics is a newly identified field at the intersection of neuroscience and ethics, the area of philosophy that deals with the study of standards of conduct and moral judgment. (Neuroscience intersects with other areas of philosophy as well, such as aesthetics, epistemology, metaphysics, and mind–body theory, but these are not directly relevant to neuroethics.) Neuroethics includes both the ethics of neuroscience and the neuroscience of ethics; that is, it is concerned with ethical issues uniquely or primarily associated with the practice or application of neuroscience and also the neurobiology of moral and ethical thinking and decision making (Figure 1).

History

The term neuroethics was initially introduced in 1993 in the context of treatment of young neurological patients and reintroduced almost a decade later in discussions of the application of technological advances in the area of neuroscience and the biological elements underlying ethical thoughts and behavior. However, many of the issues with which neuroethics is concerned have been the focus of philosophical thought for centuries. Attention to the ethical concerns associated with the application of neuroscience findings, and with the conduct of neuroscience research, has steadily increased during the past several decades. For example, debates about free will and determinism which were initially grounded in the role of the gods in controlling the lives of mere mortals evolved into a recognition and discussion of the strong influence of cultural factors and experience in behavioral decisions. These discussions were the forerunners of current debates about responsibility, self-control, and the neurobiological underpinnings of moral decision making.

In the twentieth century, discussions of ethics and neuroscience primarily centered around the application of neuroscience findings or techniques to address individual or societal concerns. For example, the actual and proposed use of psychosurgery to treat mental conditions and to control problematic behavior was and remains controversial. Since 1983, the Society for Neuroscience, the largest professional society of neuroscientists in the United States, has held an annual Social Issues Roundtable that has generally focused on the ethical, legal, and social implications of neuroscience research. These sessions have examined such topics as research into possible sex differences in the brain and the applications of that research, brain death, therapeutic and nontherapeutic use of cognitive enhancers, the use of fetal tissue to treat neurological diseases, neurotoxicity of food additives, and the role of neuroscience research into drug addiction in the development of health and public policy. Also in 1983, the US Office of Technology Assessment commissioned a report that included the potential societal impacts of neuroscience research. Among the topics covered were the implications of neuroscience findings for the criminal justice system, the workplace, and education.

Attention to neuroethics in research practice has primarily emphasized issues associated with the humane treatment of research subjects, especially laboratory animals. Debate pitting the scientific advances and societal benefits of animal experimentation against concerns about animal cruelty has raged since the early 1800s. This controversy was brought directly to neuroscience research in the 1980s, initially when police raided a laboratory in Silver Spring, Maryland, where neuroregeneration was being studied in primates and later as a result of highly publicized exposures of controversial practices in neuroscience research. In the United States, neuroscience research with laboratory animals, like animal experimentation in general, is regulated by the Department of Agriculture and the Department of Health and Human Services, primarily through required institutional animal care and use committees (IACUCs). The neuroscience community is actively engaged in discussing the ethical issues associated with animal experimentation in neuroscience.

Features of the Nervous System with Ethical Implications

The complexity of the nervous system, its plasticity, the dynamic, interactive quality of its elements, and the remarkable variation in structure and function from one individual to the next are qualities of the nervous system with important implications for neuroethics. Because there are more uniquely expressed genes and more types of cells in the nervous system than in the rest of the body, more interconnections throughout the nervous system (and connections with the immune, endocrine, and other physiological systems as well), and because dynamic interactive communication between and among cells can result in short-term and long-term changes in the elements of

the system, it can be difficult, if not impossible, to reliably predict short- and long-term effects of the manipulation of brain structure or function. Moreover, individual differences in neuroanatomy, physiology, and biochemistry defy easy generalization regarding the impact of various types of experience or interventions from one individual to the next.

Although there is great interest and appeal in applying neuroscience to address societal concerns, and technological advances provide ever more detailed information about molecular and cellular mechanisms underlying emotion, perception, cognitive function, and behavior, scientific understanding of brain structure and function continues to evolve. Although much is known, much remains to be discovered. The relevance and importance of these characteristics of the nervous system to ethical analysis become increasingly clear in the context of particular research practices and applications of research findings. They are fundamental to accuracy and reliability and therefore to the interpretation and appropriate application of neuroscience findings and techniques.

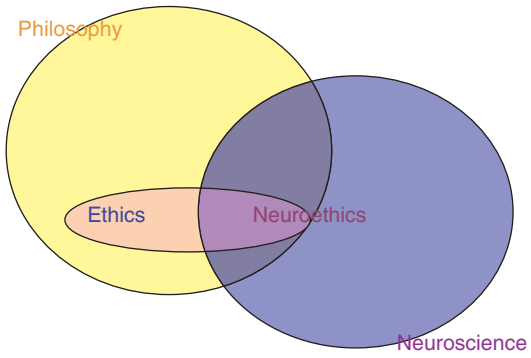

Figure 1 Neuroethics includes both the ethics of neuroscience and the neurobiology of ethics.

Structure of Neuroethics

Broadly speaking, there are three branches of neuroethics: (1) ethical issues that are uniquely or especially associated with the conduct of neuroscience research; (2) issues associated with the application of neuroscience research findings or technologies to human needs, interests, public policies, or societal concerns; and (3) the neurobiological aspects of ethical and moral thinking, judgment, decision making, and behavior (Figure 2).

Conduct of Neuroscience Research

In general, ethical issues associated with conducting neuroscience research are not unique to neuroscience and are the same as those encountered in any area of research in the natural and social sciences. They fall under the umbrella of research ethics, and especially the responsible conduct of research (RCR). The Office of Research Integrity in the Office of the Secretary of the US Department of Health and Human Services identifies the areas of RCR as follows:

- Data acquisition, management, sharing, and ownership
- Conflict of interest and commitment
- Human subjects
- Animal welfare
- Research misconduct
- Publication practices and responsible authorship
- Mentor and trainee responsibilities
- Peer review
- Collaborative science

Appropriate recognition for contributions to a research project, the responsibilities of authorship, and confidentiality and impartiality in the peer review process are a few of the many aspects of research

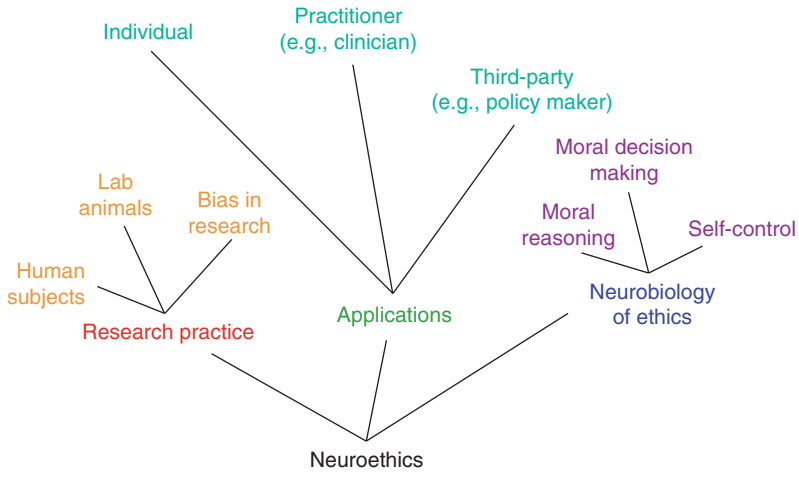

Figure 2 The three branches of neuroethics.

ethics that are not unique to neuroscience and are not, *per se*, part of neuroethics. However, some components of RCR, such as the humane treatment of research subjects and potential for bias in research, raise questions of particular relevance to neuroscience research.

Clinical research A core ethical principle central to research involving humans is respect for persons. It is reflected in the requirement that participation in a research study be a voluntary and informed decision, and that research subjects can discontinue their involvement in the study at any time. Whether one can make a truly informed decision to be a research subject is widely debated no matter what the nature of the research, but it is especially problematic when the mental capacity of a research subject is compromised, for example, a patient with dementia or schizophrenia. A patient with Alzheimer's disease who initially agrees to participate in an extended study of the efficacy of a particular treatment may lose the capacity to comprehend the broader value of the study as the disease progresses and his or her experience as a research subject becomes a series of confusing, painful encounters with clinical research staff. Although proxy or surrogate consent allows family members or other legal guardians to give consent for the patient, to some degree it inevitably undermines the notion of research participation as a fully informed choice. Even for fully functioning research participants, given the complexity, plasticity, and interconnectivity of the nervous system, research that involves psychoactive substances or other treatments that may directly affect brain function could impair the capacity to understand and evaluate the risks and benefits of participation and affect the participant's ongoing assessment of his or her desire to continue in the study.

Basic research Required for both the ethical and the scientific justification of the use of laboratory animals in research is that the work has the potential to reveal useful insights into biological structure and/or function that lay a foundation for the understanding and, ultimately, the treatment, prevention, and/or cure of disease, dysfunction, or injury. At the same time, it is expected that research can be carried out so that the animals will experience minimal or no pain, suffering, or distress. Any research with animals potentially raises ethical concerns: research that is redundant, poorly designed, of dubious merit, or that inflicts unnecessary or extreme suffering is ethically unacceptable. Whether the procedure involves maternal–pup separation, implementation of starvation or dietary manipulations, surgical interventions, or any other painful or stressful components, IACUCs, the research community, and ultimately the society which funds the research and in whose name the research is carried out must determine whether the potential benefits of the expected research findings outweigh the suffering borne by the research subjects. Given that suffering is a subjective experience, it is challenging if not impossible to carry out this calculus objectively or reliably.

Some areas of neuroscience research pose particular concerns. For example, when research focuses on mental conditions such as schizophrenia or elements of cognition such as intentionality, investigators must make assumptions about the similarity of brain function across species. The reliability of those assumptions and their implications for the understanding of human brain function and disease can be questioned. Moreover, when the focus of research is pain, anxiety, or other forms of stress, some degree of pain and/or stress are elements of the research. Paradoxically, the more like humans an animal is – that is, the more it possesses attributes such as intelligence, sensitivity to pain, emotional complexity, or other similar qualities – the more the research is informative but also ethically problematic if the justification for conducting the research in animals is that it is unethical to investigate the phenomenon in humans because the benefit does not outweigh the harm to the research subject.

Bias in research Attention to bias in the various components of research, from research design to data collection, selection, and interpretation and the reporting of study findings, is always an important aspect of research ethics. Bias in research design is of particular relevance and concern in neuroscience research, and thus an aspect of neuroethics, because of the extent, nature, and implications of findings in this field. Assumptions that underlie research questions are often not adequately investigated. Yet they are likely to reflect conscious or unconscious bias that arises from long-standing socially determined beliefs. For example, it is widely assumed that some differences in male and female behavior reflect anatomical and physiological differences in the brains of males and females. Although this may be true, it is not clear whether biological differences relevant to behavior result from the presence of different sex-related genes or molecules, from differences in the many external factors that shape interactions with others from birth, or from a combination of both. Whatever the basis of sex differences in behavior, the extent to which they are linked to biology and perceived as predetermined and immutable can have far-reaching ramifications for education, employment, health care, and other areas of social and public policy.

Application of Research Findings

Although neuroscience research findings have application in veterinary medicine (e.g., psychoactive medication to treat anxiety), in animal control (e.g., the design of slaughter houses), and in the development of more humane animal research methods, the literature in neuroethics has most commonly focused on the use of neuroscience findings or technology to address human needs, interests, and public policy issues and concerns. The ethical issues associated with the use of research findings are linked to the particular application: who uses the information, for what purpose (e.g., therapy and enhancement), and how it will be used (e.g., passively to monitor brain activity or actively to manipulate behavior). In addition, whether information is about the general population or about a particular individual, and whatever its intended application, its accuracy and reliability are always important considerations, as is accurate presentation of its limits and potential for misuse and abuse, because these considerations directly affect decisions about whether and how best to use the information.

Who, why, and how If individuals use research findings for their own purposes, to inform their own decisions about therapy, for example, or for self-understanding, ethical concerns are not the same as if the information is being used by a second or third party to decide how to interact with another individual (e.g., a patient or client) or to develop policies for a particular group (e.g., sex offenders or women students). For an individual, if the information is general and benign with noninvasive applications (e.g., mnemonic techniques for remembering names), the accuracy and reliability of the research findings are less critical than if the information may expose an individual to risk (e.g., research that suggests a particular dietary supplement is an effective sleep aid although it has the potential for inducing heart arrhythmias). When research findings provide information specific to a particular individual, the accuracy and reliability of the information can be critically important depending on the nature of the information and the purpose for which it is being gathered. For example, the reliability of predictions of a test for a debilitating hereditary neurological or mental illness is key. If test findings are perceived to be consistent markers for the disease (i.e., individuals with a positive test result inevitably get the disease), then the actual reliability and limits of the test (and the research on which the test is founded) are critically important if individuals are to make important medical and personal decisions (e.g., whether to have children) based on test results.

On the other hand, if the test is an indication of a predisposition for a mental illness, then additional ethical concerns arise. In particular, given the dynamic and interactive nature of the human mind, knowledge of the identification of a biological element that is neither necessary nor sufficient for a mental illness but, rather, indicates a predisposition for that condition can become a contributing factor in its own right and a self-fulfilling prophecy. Thus, ethical concerns regarding information about predispositions to disease are related not only to the accuracy and reliability of the test but also to the nature of the nervous system and the independent power of the information. In addition, given the continuing social stigma associated with mental illness, provision of test results to third parties, whether health insurance providers, employers, family members, or others, may also contribute to stress and the development, expression, and manifestation of disease. As a result, information about mental function poses risks as well as benefits because it is provided in a personal and social context with which it interacts. Technological advances can improve the accuracy of the information but may not have much impact on the contexts in which it is provided.

With increased understanding of brain chemistry, pharmacology, and physiology has come the development of invasive techniques such as pharmaceuticals, electroconvulsive shock therapy, stereotaxic surgery, and deep brain stimulation to address mental conditions. These therapeutic strategies are prescribed to treat an individual's self-report of dysfunction and, as with any treatment, raise issues of benefit versus risk, reversibility, patient expectations and informed decision making, and allocation of resources. However, because brain dysfunction and mental illness are often at the extreme ends of normal brain function, some therapeutic agents may be able to enhance normal function. For example, some treatments for Alzheimer's disease or other forms of dementia may be able to enhance normal cognitive function. The use of pharmaceuticals for artificial enhancement of performance, whether mental or physical, is highly controversial (e.g., the use of steroids or erythropoietin to enhance performance in sports). The potential development of cognitive and/or emotional enhancers to gain personal advantage not only changes the benefit versus risk analysis and alters discussions of the fair allocation of scarce resources but also raises issues of respect for both the self and others, fair competition, and informed decision making. Enhancement also raises questions regarding who is being enhanced, by whom, and for what purpose. Psychotherapeutic agents that modify brain chemistry to treat mental conditions

(e.g., anxiety, dementia, depression, and schizophrenia) are prescribed and are taken by individuals to modify brain function, mental activity, and behavior. Those who take these agents may report feeling 'more like themselves' or conversely 'not themselves.' This not only prompts the philosophically interesting question of how one defines and recognizes 'the self' but also raises ethical concerns regarding the extent to which peer and/or societal pressures may lead an individual to modify his or her mental processes, behavior, or other elements of the self in order to conform to the expectations of others or to internalized social norms.

When second-party practitioners (e.g., clinicians, lawyers, and educators) or third parties (e.g., corporations and government agencies) apply the results of neuroscience research to others, accuracy and reliability continue to be fundamental, but concerns about the purpose for applying the information or technology, and the way in which it is used, take on even greater significance. For example, the purpose may be to determine brain death, assess awareness, ascertain dangerousness, or evaluate the capacity to learn. An obvious clinical application is in assessing awareness in comatose patients or patients in a vegetative state. Currently, clinical evaluation depends on subjective interpretation of a patient's response to stimuli. Disturbingly, a number of studies have shown that a diminished response has commonly led clinicians to underestimate the degree of a severely disabled patient's awareness and to incorrectly diagnose a persistent vegetative state. Technological developments that make possible physically noninvasive access to brain activity seem to offer the possibility of an objective measure of awareness and potentially promise a clearer determination of brain death or more reliable prediction of recovery. At the same time, this raises questions about the relationship between brain activity and awareness and the extent to which brain activity can predict future loss or recovery of mental function, and therefore how the information will be used.

The possibility of monitoring the mental activity of others also holds the possibility of detecting deception or determining the intentions of others. This application challenges notions of individual integrity and privacy. In general, respect for the individual includes the right to privacy, and exceptions are allowed only when the health, safety, or welfare of that individual, or others, is threatened. A key consideration is whether the individual whose mental activity is being examined will benefit from the process. In addition, the extent of the invasion of privacy (and attendant harm to the principle of respect for persons and potential harm to that individual) is balanced against the seriousness and certainty of the harm or threat to be averted. A physician is likely to be willing to invade privacy to determine if a patient is suicidal. In the criminal justice system, the primary beneficiary may not be the individual examined. It is well established that eyewitness testimony is unreliable, and the potential for conflicting interests among experts as well as the concerns of a hostile or threatened witness can also call into question the reliability of courtroom testimony. Thus, if and when it is possible and in the putative interest of justice, authorities might seek to access directly the memories of a witness or an accused to determine what actually occurred. Similarly, they might seek access to the mental activity of a perpetrator in order to determine the individual's intentions.

The notion of monitoring the mental activity of others again raises questions regarding the actual and perceived correlation between brain activity and mental activity. This topic is dominated by presumption and uncertainty. Although the consensus of the neuroscience community is that, at least in humans, brain and mind are two sides of the same coin, there is considerable controversy and disagreement regarding the degree to which mental activity can be correlated with, identified as, and reduced to brain activity. The view that patterns of brain activity detectable with imaging technologies or by monitoring electrical changes can be identified with specific cognitive functions is not universally accepted. Yet the actual and perceived nature and reliability of this correlation is central to many of the ethical concerns associated with the application of scientific findings and technical developments in neuroscience.

Brain–computer interfaces Studies are currently exploring the possibilities of the brain–computer interface (BCI). Initially, BCIs have been designed to be therapeutic (e.g., to overcome physical limitations or visual deficits). Some are implanted (e.g., cochlear implants) and others respond to transcranial signals. Again, beyond the ethical concerns associated with any therapy, BCIs emphasize the importance of who is applying the technology and for what purpose. There are distinct and important differences between the BCI that makes the brain of a quadriplegic a 'transmitter' that can manipulate the external environment (e.g., move a cursor on a computer screen), a cochlear implant that makes the brain a 'receiver' for information about the outside world (and again thereby overcomes a physical limitation), and an implant that makes the brain a receiver in order to alter brain function in some other way (e.g., to treat obsessive–compulsive disorder).

Although manipulation and control of others are always ethically problematic because they violate the basic bioethical principle of respect for individuals

and their autonomy, two primary considerations are (1) the degree of invasiveness and (2) the extent to which the individual being controlled is aware of, and consents to, the control. Invasiveness need not be physical: education and subliminal suggestion can permeate one's thinking with long-term, widespread effect (e.g., educational programs that include evolutionary theory and/or creationism or that exclude acknowledgment of the Jewish holocaust and/or Chinese 'comfort women'). Moreover, the conscious intent of manipulation or control may well be in the eye of the beholder. Thus, education, although not physically invasive, is potentially manipulative; subliminal suggestion is not physically invasive but is designed to be manipulative; and psychoactive agents and BCIs may be more or less physically invasive and can be perceived as manipulative or not. Scientific and technological advances that reflect new or refined understanding of brain structure and function have the potential for making possible more specifically targeted monitoring and manipulation of individual or group perceptions and function, but the extent to which individuals are aware of and consent to the manipulation makes the ethical concerns akin to those raised by con artists, rabble rousers, propaganda, and deceptive advertising.

Neurobiology of Ethics

The third branch of neuroethics considers the neurobiological underpinnings of ethical thought and practice. This rich and complex area of study is contributing to an intellectual understanding of moral development, the cognitive and emotional elements of moral reasoning and decision making, and the mechanisms by which ethical decisions are expressed in behavior. How society understands notions of free will, moral agency, and responsibility will be influenced by the findings of neuroscience research, and of necessity, this understanding will reflect recognition of the limits of human capabilities.

For example, notions of responsibility require intentionality, and moral decision making consists of a number of elements, including (1) awareness of ethical problems, which consists of both perception and understanding; (2) analysis of a situation, which entails identification of ambiguities, development of alternative responses, and envisioning the implications of various courses of action; (3) evaluation of the advantages and disadvantages of the alternative responses; (4) deciding how to proceed; and (5) carrying out the action. A variety of factors affect these elements of decision making, such as education, intelligence, peer influence, and empathy. At the same time, a number of endogenous and exogenous biological factors are also known to influence one or more of the elements of decision making, ranging from reflexes, brain tumors, and hormones to neurotoxins and other psychoactive substances including those in food. In addition, current research is investigating a number of relevant topics, such as the biological basis of empathy, the impact of drugs of addiction on brain function such as emotional and motor control, and the behavioral effects of interactions between genetic makeup, environment, and experience.

Responsibility is often framed simplistically in terms of the presence or absence of free will (free will vs. determinism), despite both the fact that moral decision making is composed of a large number of stages, steps, and interacting elements (most of which cannot be readily dissected from each other) and the fact that a host of simple and complex biological and environmental factors influence these stages, steps, and elements through complicated, dynamic, and variable neurobiological mechanisms. Only some of these factors (e.g., education) are under the control of the individual. It is not clear how these findings can or should be used in judgments of moral and legal responsibility or culpability.

In addition, as with research in neuroscience in general, conscious and unconscious assumptions may introduce an inappropriate bias into research design, analysis, or reporting. For example, it is widely assumed that moral reasoning is a rational rather than emotional process. This assumes a potentially false dichotomy in brain processing. Thus, the ethical issues that are likely to be raised by future investigations of the neurobiological basis of ethics will be complex and dynamic like the nervous system.

Conclusion

During the past several decades, research in the neurosciences has expanded dramatically and the applications and implications of this research are becoming more apparent. Like other areas of science and technology, ethical values are both intentionally and unintentionally embedded in neuroscience. Neuroethics, like other areas of ethics in science and technology, is identifying and examining the practices and ethical, legal, and social impacts of the field. Unlike other areas of science and technology ethics, the dynamic, reflexive nature of the object of study of the neurosciences raises unique questions. Neuroscience research seeks to investigate, understand, and explain to scientists and nonscientists alike the nature of the structure and function of the nervous system and, by extension, higher brain function, cognition, and behavior. This inevitably introduces ethical issues that must

be recognized, acknowledged, and addressed as the field and the knowledge it creates continue to expand and evolve.

See also: Brain-Computer Interface; Cognitive Neuroscience: An Overview; Deep Brain Stimulation; Deep Brain Stimulation and Parkinson's Disease.

Further Reading

Bird SJ (2005) Neuroethics. In: Mitcham C (ed.) *Encyclopedia of Science, Technology, and Ethics*, pp. 1310–1316. Farmington Hills, MI: Macmillan.

Birke L, Arluke A, and Michaels M (2007) *The Sacrifice: How Scientific Experiments Transform Animals and People*. West Lafayette, IN: Purdue University Press.

Blass T (1999) The Milgram paradigm after 35 years: Some things we now know about obedience to authority. *Journal of Applied Social Psychology* 29(5): 955–978.

Chorover SL (1979) *From Genesis to Genocide*. Cambridge, MA: MIT Press.

Dworkin G (1976) Autonomy and behavior control. *Hastings Center Report* 6(2): 23–28.

Farah MJ and Wolpe PR (2004) Monitoring and manipulating brain function: New neuroscience technologies and their ethical implications. *Hastings Center Report* 34: 35–45.

Gaylin W, Macklin R, and Powledge TM (eds.) (1981) *Violence and the Politics of Research*. New York: Plenum.

Gray HM, Gray K, and Wegner DM (2007) Dimensions of mind perception. *Science* 315(5812): 619.

Illes J (ed.) (2006) *Neuroethics: Defining the Issues in Theory, Practice, and Policy*. New York: Oxford University Press.

Illes J and Bird SJ (2006) Neuroethics: A modern context for ethics in neuroscience. *Trends in Neurosciences* 29: 511–517.

Longino HE (1990) *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Longino HE and Doell R (1983) Body, bias and behavior: A comparative analysis of reasoning in two areas of biological science. *Signs* 9: 206.

Marcus SJ (ed.) (2002) *Neuroethics: Mapping the Field: Conference Proceedings*. New York: Dana Press.

Mitcham C (ed.) (2005) *Encyclopedia of Science, Technology, and Ethics*. Farmington Hills, MI: Macmillan.

Office of Technology Assessment (1984) *Impacts of Neuroscience*. Washington, DC: U.S. Congress.

Pfaff DW (ed.) (1983) *Ethical Questions in Brain and Behavior: Problems and Opportunities*. New York: Springer.

Raichle ME (2006) The brain's dark energy. *Science* 314: 1249–1250.

Roskies A (2002) Neuroethics for the new millennium. *Neuron* 35: 21–23.

Valenstein ES (ed.) (1980) *The Psychosurgery Debate*. San Francisco: Freeman.

Zeki S and Goodenough O (eds.) (2004) *Law and the Brain*. New York: Oxford University Press.

Neuroethological Perspective

K K Watson and M L Platt, Duke University, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The unifying goal of ethology, as well as of newer fields of behavioral ecology and sociobiology, is to provide an evolutionary explanation for behavior. This approach proposes that both natural and sexual selection favor those behaviors that maximize the reproductive success of individuals within the context of their physical and social environments. Ethologically, rewards can be considered proximate goals that, when acquired, tend to enhance survival and mating success. Similarly, avoiding punishment is a proximate goal that ultimately serves to augment the long-term likelihood of survival and reproduction. These definitions thus extend the traditional psychological and neurobiological notions of reward and punishment, which are typically defined by the quality of eliciting approach and avoidance, respectively. As detailed herein, the assumption of evolutionary adaptation within ethology, behavioral ecology, and sociobiology has promoted the development of mathematical models that formally define rewards and punishments within specific behavioral contexts. Such models suggest that full understanding of the neurobiology of reward and decision making will require consideration of naturally occurring behaviors in the specific ecological and social contexts in which they are normally expressed.

The Economics of Natural Behavior

In one of the first direct applications of economic modeling to animal foraging behavior, MacArthur and Pianka proposed in 1966 a model that defined criteria for the consumption or rejection of prey by a foraging animal. Using feeding efficiency as a proxy measure for evolutionary fitness, the benefits of consuming a particular prey type were modeled as the ratio of the energetic benefit afforded to the animal relative to the time costs of seeking and handling that prey type:

$$R = E/(T_h + T_s)$$

R is the net benefit gained by the predator for consuming a particular prey type, E is the amount of energy gained, T_h is the handling time, and T_s is the search time. The model is solved to maximize R, which determines the diet offering the greatest net energetic return and thus maximizing evolutionary success.

One prediction of this model is that the greater the abundance of higher quality foods, the less an animal's diet will consist of lower quality foods (the 'independence of inclusion from encounter' rule). In the 1970s, Goss-Custard found evidence in support of this prediction in his studies of redshank, small wading birds found in estuarine habitats in Great Britain. Redshanks feed on crustaceans and worms; crustaceans offer higher energetic returns relative to worms, which offer lower energetic returns. As predicted, the birds did not indiscriminately consume every worm or crustacean encountered; instead, they exclusively ate crustaceans when their density was high, but included worms in the diet when crustacean density declined. This result makes intuitive sense because the cost of eating worms includes 'missed opportunities' to search for and eat more nutritionally profitable crustaceans. When crustaceans are rare, however, it is more profitable to focus on small but abundant worms rather than wasting time searching for higher value foods.

The general prediction of the prey model – that inclusion of a particular food in the diet is independent of its encounter rate – has since been replicated in a variety of species, including birds, insects, fish, mammals, and even humans. However, as is often the case in economics, MacArthur and Pianka's simple model does not perfectly describe behavior in the real world. In Goss-Custard's study, as well as subsequent studies of different species using different methods, the animals did not display the steplike preference shift predicted by the prey model. Instead, foragers exhibited 'partial preferences' for different types of prey (e.g., preferring one type 75% of the time) when prey density changed. Such partial preferences might reflect the implementation of sampling behavior, which allows the animal to acquire improved information about the statistics of the local environment. Alternatively, partial preference could reflect sensory or memory-related cognitive limitations.

Early models, including MacArthur and Pianka's, described behavioral optimization in a generic sense, without regard to the specific physiological or cognitive constraints of a particular animal. Prediction of a precise outcome in an individual species, on the other hand, requires consideration of the specific set of physiological and environmental constraints relevant to that animal. For example, moose foraging for terrestrial and aquatic plants must satisfy both energetic needs and sodium requirements within the limitations

imposed by gut capacity. Terrestrial plants are richer in energy than aquatic plants are, and take up less room in the gut. However, aquatic plants contain more sodium than terrestrial plants do, and, because aquatic plants are buried under ice during the winter, moose must consume enough of them during the summer to satisfy their sodium requirements for the rest of the year. Belovsky developed a model of diet choice in moose, satisfying these constraints using linear programming. According to this model, if moose forage in a manner that maximizes energetic return while simultaneously satisfying sodium needs and rumen constraints, the model predicts a diet of 18% aquatic plants, in precise agreement with actual field observation data (Figure 1).

Early field studies using optimal foraging models, such as Goss-Custard's redshank study and Belovsky's moose study, demonstrated the strengths of the economic approach: formulation of models allows for clear and precise predictions that can be tested empirically, and provides a quantitative tool around which to organize explanations of behavior. Importantly for this discussion, such models make clear that defining rewards and punishments requires careful consideration of the behavioral and physiological capacities of a given species and the specific physical and social environments in which the species normally acts.

Neurobiology of Reward and Decision Making

Ultimately, the nervous systems of humans and other animals have evolved to promote behaviors that enhance fitness, such as acquiring food and shelter, attracting mates, avoiding predators, and prevailing over competitors. To achieve these goals, animal brains have become exquisitely specialized to attend to important features of the environment, extract their predictive value for success or failure, and then use this information to compute the evolutionarily optimal course of action. Traditionally, these brain mechanisms have been studied with regard to their roles in acquiring rewards and avoiding punishments.

As noted, rewards are traditionally defined as stimuli that elicit approach behavior, while punishments can be defined as stimuli that elicit avoidance. Recent studies have revealed elementary properties of the neural systems that process rewards and punishments as traditionally defined. Specifically, the circuit connecting midbrain dopamine neurons to the ventral striatum and prefrontal cortex appears to be crucial for processing information about rewards. For example, animals will work to receive current delivered via electrodes implanted in the dopaminergic ventral tegmental area (VTA) or medial forebrain bundle, which connects the VTA to the ventral striatum. In fact,

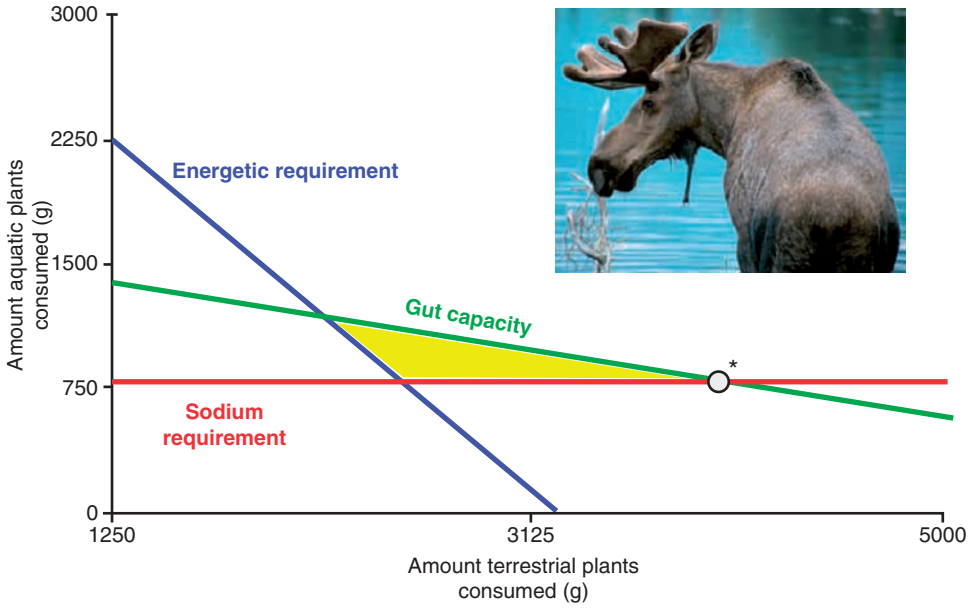

Figure 1 Optimal diet choice in moose (*Alces alces*). The daily ratio of aquatic plants to terrestrial plants consumed by moose must satisfy three constraints: the diet must meet energetic (blue line) and sodium requirements (red line), subject to digestive limitations (green line). Those ratios that meet these constraints are contained in the yellow shaded area. The vertex marked with an asterisk indicates the aquatic-to-terrestrial plant ratio that maximizes energetic intake while also satisfying all constraints. Adapted from Stephens PW and Krebs JR (1987) *Foraging Theory*, p. 120. Princeton, NJ: Princeton University Press. Photo courtesy of the National Park Service.

animals will preferentially work for such intracranial self-stimulation (ICSS), to the exclusion of acquiring food or mating opportunities, which are normally reinforcing.

Electrophysiological recordings from dopaminergic neurons show that these cells respond to primary rewards, such as food and water, as well as to conditioned stimuli that predict such rewards (Figure 2). Moreover, dopamine neuron responses scale with both reward magnitude and reward probability. Dopamine neurons do not, however, merely signal rewards and the stimuli that predict them. Current evidence suggests that phasic bursts by dopamine neurons may correspond to the reward prediction error term initially proposed in purely behavioral models of learning. According to this view, such phasic dopamine responses provide a mechanism for updating predicted valuation functions, which can be used both to learn about stimuli in the environment and to select profitable courses of action.

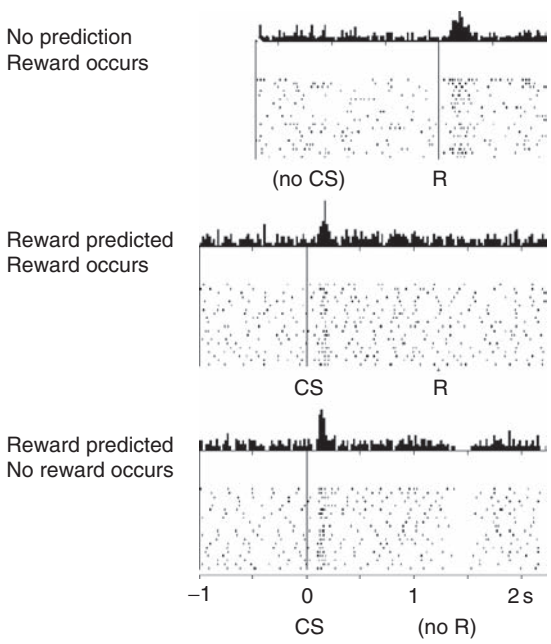

Figure 2 Reward-related responses by a single dopamine neuron recorded in a macaque monkey. (Top) When the animal is still learning the task, the fruit juice reward is unexpected, and the neuron responds at the time of reward delivery (R). (Center) After the monkey learns the relationship between a conditioned stimulus (such as a light or tone) and a reward, the neuron responds to the conditioned stimulus that predicts reward delivery (CS), but not to the reward itself. (Bottom) If the reward is omitted after the predictive stimulus, dopamine neuron activity is suppressed during the time of expected reward delivery. Each raster indicates the time of neuron spiking, and each row corresponds to a single trial for that neuron. The histograms summate the spikes over all the trials. Adapted from Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

These valuation functions can be thought of as the neural implementation of the optimization functions assumed to guide behavior in economic models of behavior developed in behavioral ecology and ethology.

Signals from the dopaminergic midbrain neurons influence processing within cortical decision-making areas, primarily in prefrontal and parietal cortex, that assign value to sensory stimuli and eventually transform that information into motor output. Platt and Glimcher probed the impact of expected value on sensorimotor processing in the lateral intraparietal area (LIP), a region of the brain previously linked to visual attention and motor preparation. In that study, monkeys were cued to shift gaze from a central light to one of two peripheral lights to receive a fruit juice reward. In separate blocks of trials, the authors varied the expected value of orienting to each light, either by varying reward size or by varying the probability of each cue onset. Platt and Glimcher found that LIP neurons signaled expected target value, the product of reward size, and saccade likelihood, prior to cue onset (Figure 3). In a second experiment, monkeys were permitted to choose freely between the two targets, and both neuronal activity in LIP and the probability of target choice were correlated with expected target value.

Sugrue, Corrado, and Newsome extended these observations by probing the dynamics of decision-related activity in LIP using a virtual foraging task. In this experiment, the rewards associated with each of two targets fluctuated over time. Under these conditions, monkeys tended to match the rate of choosing each target to its relative rate of reinforcement. Moreover, the responses of individual LIP neurons to a particular target corresponded to the relative rate of reward gained from choosing it on recent trials, with the greatest weight placed on the most recent trials. Together, these and other studies suggest that simple behavioral decisions may be computed by scaling neuronal responses associated with a particular stimulus or movement by its value, thus modifying the likelihood of reaching the threshold for eliciting a specific motor action.

Uncertainty and Decision Making

Early ethological models of behavior assumed that animals maintained complete knowledge of the environment and that reward contingencies were deterministic. In practice, however, uncertainty about environmental contingencies places strong constraints on behavior. The impact of uncertainty on choice has long been acknowledged in economics, which defines the spread of an outcome's known probability as risk. In the eighteenth century, Daniel Bernoulli proposed

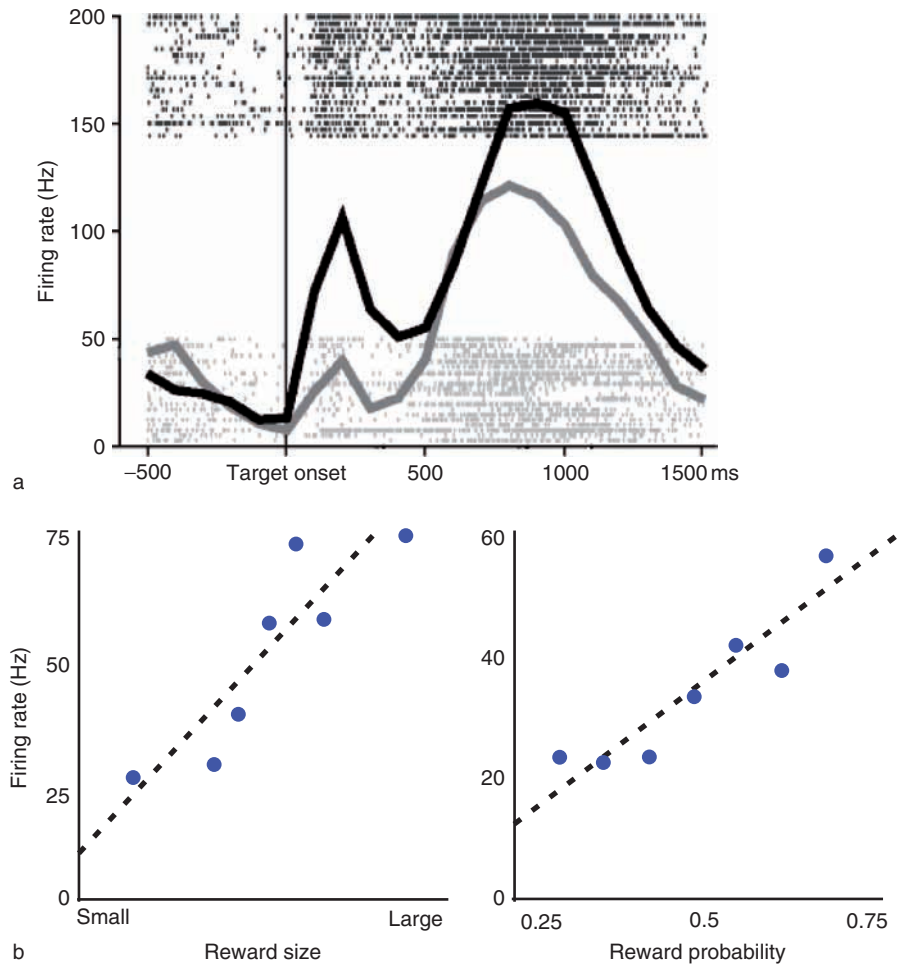

Figure 3 Lateral intraparietal area neurons encode expected target value. (a) Neuronal firing is greater during trials when the expected reward is large (black line) than when the expected reward is small (gray line). Black and gray rasters indicate the time of individual spikes for large- and small-reward trials, respectively; each line of rasters corresponds to a single trial. Curves represent the summation of activity over all the trials. (b) Firing rate of a single lateral intraparietal area neuron increases linearly with reward size (left) and reward probability (right). Adapted from Platt ML and Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400(6741): 233–238.

that the expected values of monetary transactions, particularly risky financial ventures, differ from their corresponding subjective utilities (as determined by the economic agent). This idea, which would eventually revolutionize the field of economics, challenged the traditional notion that people value outcomes strictly according to their financial returns.

Initial economic models applied to animal behavior explicitly ignored variance in reward outcomes. For example, Charnov's marginal value theorem, devised to predict when a foraging animal should leave a particular food patch, was based purely on the average distribution of resources. Although this model described behavior in simple contexts fairly well, it failed to account for behavioral sensitivity to variability within patches even when the average value of each patch was identical.

In a seminal ethological study probing the impact of risk on decision making, Caraco observed the behavior of yellow-eyed juncos, a species of small bird. These birds were given the option of choosing a tray with a fixed number of millet seeds or a tray with a probabilistically varying number of seeds, with the same mean as the fixed option. Preferences depended on the ambient temperature. At 19 °C, juncos preferred the fixed option, but at 1 °C, they preferred the variable option. The proposed explanation for this switch from risk aversion to risk seeking is that, at the higher temperature, the rate of gain from the fixed option was sufficient to maintain the bird on a positive energy budget. At the lower temperature, however, energy expenditures were elevated, so the fixed option was no longer adequate to meet the animal's energy needs. When cold, the bird's best

chance for survival was to gamble on the risky option, since it might yield a higher rate of return, compared to the fixed option.

The relationship between energy budget and risk taking has been found to be generally true across a variety of animal species, including insects, fish, birds, and mammals. Moreover, this principle also describes human decisions in experiments using money as a reward. The convergence of behavioral results from animals and humans endorses the notion that studying risk-sensitive decision making in animals can illuminate the mechanisms underlying risk sensitivity in humans. These observations strongly suggest that sensitivity to risk is an evolutionarily ancient neural adaptation supporting adaptive decision making.

Neuroimaging studies in humans have revealed that preference for a risky option is associated with increases in neuronal activity in the ventral striatum and posterior parietal cortex. Moreover, choosing a risky option activates the dorsal striatum, precuneus, and premotor cortex. A recent electrophysiological study in monkeys probed how such risk-related activity might be translated into action. Monkeys were given a choice between juice rewards of fixed or variable sizes, with the same mean reward rate. Under these conditions, monkeys showed a strong preference for the risky option. Simultaneous recordings from single neurons in the posterior cingulate cortex, a region of the brain associated with spatial attention, visual orienting, and reward processing, revealed responses that were correlated with subjective preferences for the risky option. The foregoing discussion makes plain that the discrepancy between observable outcomes and subjective preferences in decision making under risk offers a powerful paradigm for investigating the neural mechanisms underlying adaptive decision making.

Social Rewards in Primates

In most neurobiological studies of decision making in nonhuman animals, food or water is delivered following performance of a particular action. Such direct and immediate reinforcers are typically referred to as primary rewards, and, as reviewed earlier, are associated with activation of midbrain dopamine neurons, as well as neurons in the ventral striatum and orbitofrontal cortex. In humans, a varied assortment of hedonically positive experiences can evoke activity in these regions, including eating chocolate, hearing pleasant music, or even reading a funny cartoon.

Although outcomes such as food consumption or the opportunity to mate clearly motivate behavior, abstract goals such as information gathering or social interaction can also motivate approach or orienting behavior in the absence of hedonic experience. For humans and other primates, in particular, many decisions are motivated by competitive and cooperative interactions with others in a social group. The adaptive significance of navigating a complex social environment explains why social stimuli and interactions, such as an attractive smiling face, a cooperative transaction, or the opportunity to punish a traitor, evoke activity in neural circuits that overlap with those activated by primary rewards. Physical features of the face, for example, provide information about genetic quality and thus can be useful in determining whether to pursue mating. In addition to attractiveness, people apparently also use information apparent in faces to assess the expected value of cooperation. Together, these observations implicate the operation of a neural system dedicated to linking social stimuli, such as faces, to the valuation functions guiding behavioral decision making.

Nonhuman primates also use social information to evaluate their behavioral options. For example, field studies of primate social behavior have revealed that monkeys preferentially invest in relationships with dominant individuals. Moreover, male primates often use visual cues to predict female mating receptivity. These observations suggest that primates maintain valuation functions for specific social and reproductive stimuli that guide behavior. Deaner, Khera, and Platt explored this hypothesis quantitatively in the laboratory using a pay-per-view task in which male rhesus macaques were given a choice between two targets. Orienting to one target yielded fruit juice but orienting to the other yielded fruit juice and the picture of a familiar monkey. By systematically changing the juice payoffs for each target and the pools of images revealed, the authors estimated the value of different types of social and reproductive stimuli in a liquid currency.

Their work revealed that male monkeys forego larger juice rewards to view female sexual signals or the faces of high-ranking males, but need overpayment to view the faces of low-ranking males (**Figure 4**). In contrast to the valuation functions governing target choice, the patterns of gaze associated with each class of image hint at the affective complexity associated with social stimuli. Specifically, monkeys looked at female sexual signals for longer times than they looked at either high-ranking or low-ranking male faces, perhaps reflecting differences in the hedonic qualities of these stimuli (**Figure 4**). Presumably, male monkeys value the opportunity to view the faces of high-ranking males not because they are hedonically pleasing, but rather because they are potentially threatening and thus highly relevant for guiding behavior. Thus, social stimuli and behaviors that are strongly motivating are not necessarily hedonically

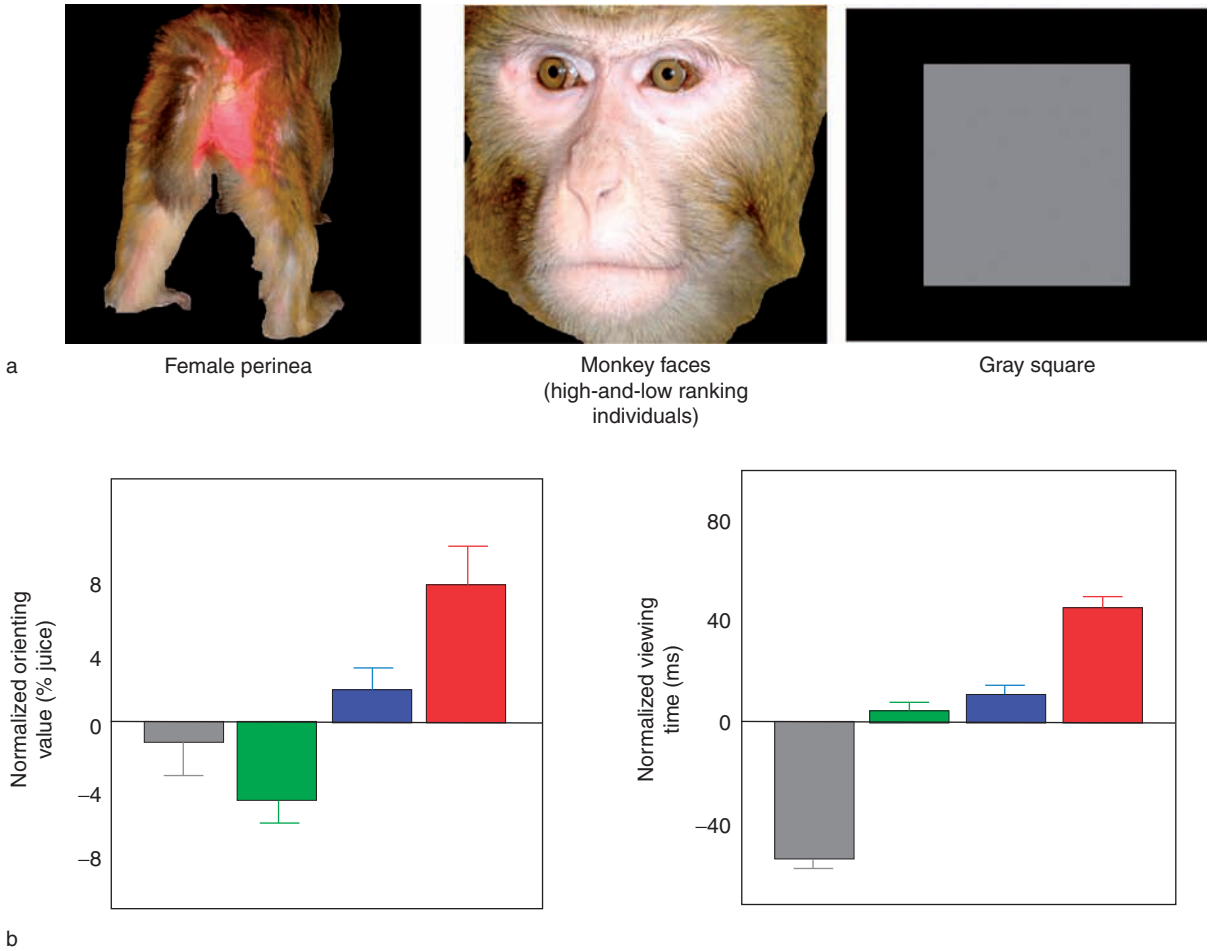

Figure 4 Monkeys value social and reproductive information. (a) Example images shown to monkeys during a 'pay-per-view' task used to assess valuation of socially relevant visual images. (b) Mean normalized orienting values (left) and looking times (right) for various image classes. Orienting values are significantly higher for both the perinea (red bar) and high-status faces (blue bar), in contrast to either the low-status faces (green) or gray square (gray). Although the monkeys choose to orient more frequently to the high-status faces than the low-status faces, the lengths of time they gaze at either of these image classes are both shorter than the time they spend viewing the perinea. Adapted from Deaner RO, Khera AV, and Platt ML (2005) Monkeys pay per view: Adaptive valuation of social images by rhesus macaques. *Current Biology* 15: 543–548.

pleasing, and it is likely that they engage distinct neural circuitry – a hypothesis yet to be tested empirically.

Social Games

One of the results of the dialogue between biology and economics was the development of evolutionary game theory. As a conceptual framework, game theory can be used to describe the ways in which behavior is influenced by the behavior of other animals when they compete for limited resources, such as mates and food, in order to survive and reproduce. A classical game describes the interaction of two or more agents with conflicting interests, both trying to maximize some gain. Each game makes precise the number of agents involved, the actions available to those agents, and the payoff that will result from all possible interactions. In economics, the participants in the game identify the costs and benefits available to each player, and are generally expected to adopt a 'rational' behavioral strategy. Typically, these behavioral strategies comprise a probabilistic distribution of responses for all players, often called the Nash equilibrium, invulnerable to penetration by other behavioral strategies. In biological applications of game theory, the economic assumptions of self-interest and rationality are replaced by the evolutionary assumptions of Darwinian fitness and population stability.

Lee and colleagues explored frequency-dependent decision making in monkeys while recording from neurons in the dorsolateral prefrontal cortex (DLPFC). Monkeys played an analog of matching pennies against a computer opponent. In this game, the animal is rewarded for choosing the target not chosen

by the computer. By manipulating the algorithm governing the computer's choices, the experimenters were able to simulate social opponents implementing various strategies. When confronted with an opponent that tracked both the history of choices and the rewards received, monkeys' choice frequencies approached the optimal random solution. Neurons in the DLPFC were sensitive to the history of choices, rewards, and current movement plans. These signals could in theory be used to update the values of each alternative action, a computation necessary for the animal to choose optimally. Obviously, the ability to make such strategic changes in behavior would be highly favored in the natural social environment. However, the neural mechanisms underlying strategic interaction with other individuals, rather than a computer opponent, have just begun to be explored in humans using neuroimaging techniques, and have yet to be probed in animals.

Conclusion

Although still in the early stages, the union of ethology, economics, and neuroscience – the emerging field of neuroeconomics – offers a potentially powerful way to study the neural mechanisms underlying decision making and behavioral allocation. Just as in other animals, natural selection has shaped human behavior and its neural substrate. Thus, the behavior we display today may more strongly reflect the operation of a nervous system which evolved over eons to optimize hunting and gathering behavior in small groups, rather than being economically rational. These considerations predict that neuroethological studies will be crucial for understanding the neurobiology of reward and decision making in humans and other animals.

See also: Computational Neuroethology; Decision-Making in Financial Markets; Decision-Making and Neuroeconomics; Delayed Reinforcement: Economics; Game Theory and the Economics of Animal Communication; Games in Monkeys: Neurophysiology and Motor Decision-Making; Neuroeconomics: History; Reward and Learning; Reward Systems: Human; Reward Decision-Making; Social Interaction; Social Cognition.

Further Reading

Barraclough DJ, Conroy ML, and Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience* 7(4): 404–410.

Deaner RO, Khera AV, and Platt ML (2005) Monkeys pay per view: Adaptive valuation of social images by rhesus macaques. *Current Biology* 15: 543–548.

Glimcher PW (2003) *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. Cambridge, MA: MIT Press.

Goss-Custard JD (1977) Responses of Redshank, *Tringa totanus*, to the absolute and relative densities of two prey species. *Journal of Animal Ecology* 46(3): 867–874.

Kacelnik A and Bateson M (1997) Risk-sensitivity: Crossroads for theories of decision-making. *Trends in Cognitive Science* 1(8): 304–309.

MacArthur RH and Pianka ER (1966) On optimal use of a patchy environment. *American Naturalist* 100(916): 603–609.

McCoy AN and Platt ML (2005) Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature Neuroscience* 8(9): 1220–1227.

Platt ML and Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400(6741): 233–238.

Sanfey AG, Loewenstein G, McClure SM, et al. (2006) Neuroeconomics: Cross-currents in research on decision-making. *Trends in Cognitive Science* 10(3): 108–116.

Schultz WW (2002) Getting formal with dopamine and reward. *Neuron* 36(2): 241–263.

Schultz W, Dayanz P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

Smith JM (1982) *Evolution and the Theory of Games*. Cambridge, MA: Cambridge University Press.

Stephens PW and Krebs JR (1987) *Foraging Theory, p.120*. Princeton, NJ: Princeton University Press.

Neurofibromatosis Type 1

A Theos and B R Korf, University of Alabama at Birmingham, Birmingham, AL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term phakomatoses is derived from the Greek root, *phakos*, meaning birthmark, and refers to a group of inherited disorders that affect the skin and nervous system. Neurofibromatosis 1 (NF1) is the most common phakomatoses, with an estimated prevalence of 1 in 3500 individuals worldwide. It is an autosomal dominant disorder with complete penetrance but highly variable expression, even among family members. NF1 is characterized by a propensity to develop benign tumors of the nerve sheath (neurofibromas) along with other benign and malignant tumors of a variety of neural and nonneural tissues. Other features of NF1 include hyperpigmented skin lesions, skeletal dysplasias, learning disabilities, structural cerebral abnormalities, and vascular anomalies. In 1990, the NF1 gene was cloned, and its protein product, neurofibromin, identified. The *NF1* gene functions as a tumor suppressor, which explains the susceptibility for individuals with NF1 to develop tumors; the pathogenesis of nontumor manifestations is not known.

NF1 Phenotype

Diagnostic Criteria

Clinical diagnostic criteria for NF1 were established at the National Institutes of Health (NIH) Consensus Development Conference on Neurofibromatoses in 1987 and the diagnosis of NF1 is still made on the basis of these criteria (Table 1). Radiographic studies and molecular genetic testing are not necessary to establish a diagnosis of NF1.

NF1 is a progressive disorder and the incidence of diagnostic clinical signs increases with age; therefore, a definitive diagnosis is not always possible in young children. The diagnosis of NF1 can be reliably established in only 70% of children by age 1, but increases to 97% by age 8 and 100% by age 20. Café-au-lait spots are the earliest feature and are present in 99% of infants with NF1. Skin-fold freckling is the next most common feature to appear and is present in 90% of children by age 7 and often confirms a suspected diagnosis of NF1 in a sporadically affected child. Iris hamartomas (Lisch nodules) are present in 70% of affected individuals by the age of 10 and are best viewed by slit lamp examination. Cutaneous neurofibromas do not start to appear until early adolescence and are present in almost all adults with NF1. In contrast, plexiform neurofibromas are often congenital, but may not be noticed until later in life and occur in 25–30% of affected individuals. The incidence of osseous lesions, particularly tibial dysplasia, and optic pathway gliomas is less common (2–3% and 15%, respectively), but if present are typically recognized in early childhood. Table 2 lists the likely age of onset of other clinical features and complications of NF1.

Tumors

NF1 is characterized by a propensity to develop benign and malignant tumors of a variety of neural and nonneural tissues. The hallmark tumor of NF1 is the neurofibroma, which arises from cells of the nerve sheath. Neurofibromas can arise anywhere along any nerve, from the dorsal root ganglion to the nerve ending. Cutaneous neurofibromas appear either as discrete soft pink or brown papules above the skin surface or as more subtle bluish lesions below the skin surface (Figure 1). The visibility and number of lesions accounts for much of their morbidity. Subcutaneous neurofibromas run along the course of the nerve and present as discrete, firm, often painful, nodules appreciable only by palpation. Plexiform neurofibromas involve multiple nerve fascicles, including multiple nerve branches, and can invade surrounding tissues. Involvement can be superficial or deep. Some plexiform neurofibromas can be seen on a surface examination, and manifest as ill-defined cord-like masses often with overlying hyperpigmentation, hypertrichosis, and/or increased vascular markings (Figure 2). Others that do not involve superficial tissues may not be apparent on clinical examination. Complications of plexiform neurofibromas include disfigurement, pain, signs, or symptoms related to compression of nerves or other vital structures, and malignant transformation.

Malignant peripheral nerve sheath tumors (MPNSTs) arise predominantly in preexisting plexiform neurofibromas; the lifetime risk in individuals with NF1 is 10%. Signs of malignancy are pain and sudden growth. The prognosis of MPNST is very poor and is due to the frequent delay in diagnosis, the infiltrative growth of these tumors, and a lack of response to traditional chemotherapy and radiation.

In addition to peripheral nerve sheath tumors, tumors of the central nervous system may occur in affected individuals. Optic pathway and brain stem gliomas are the most common intracranial

Table 1 NIH diagnostic criteria for NF1 (2 or more needed for definitive diagnosis)

Clinical criterion	Comments
Café-au-lait spots	≥6 larger than 5 mm before puberty or 15 mm after puberty; 10% of population has one café-au-lait spot
Skin fold freckling	Axilla, inguinal folds, neck, inframmary region; highly specific for NF1
Lisch nodules	≥2; best viewed on slit lamp examination; highly specific for NF1
Neurofibromas	≥2 dermal neurofibromas or one plexiform neurofibroma
Optic glioma	Bilateral lesions highly suggestive of NF1
Tibial or sphenoid wing dysplasia	With or without pseudoarthrosis
Affected 1st degree relative	Severity may vary widely among family members

Table 2 Clinical features and complications of NF1 and the typical age of onset

Age of onset (years)	*Clinical features and complications*
Congenital or infancy (0–2)	Café-au-lait spots, plexiform neurofibromas, macrocephaly, tibial or orbital dysplasia, pulmonic stenosis, aqueductal stenosis, hypotonia
Childhood (2–12)	Skin fold freckles, iris hamartomas, optic pathway gliomas, learning disabilities, ADHD,[a] scoliosis, precocious or delayed puberty, short stature, renal artery stenosis, JMML,[b] rhabdomyosarcoma, migraines, abdominal pain, vascular dysplasia
Adolescence and adulthood (>12)	Cutaneous neurofibromas, pheochromocytoma, MPNST,[c] hypertension, vascular dysplasia

[a]ADHD, attention-deficit/hyperactivity disorder.
[b]JMML, juvenile myelomonocytic leukemia.
[c]MPNST, malignant peripheral nerve sheath tumor.

Figure 1 Cutaneous neurofibromas: (a) discrete soft pink papules; (b) more subtle bluish lesions (indicated by arrows).

neoplasms. Optic pathway gliomas are low-grade pilocytic astrocytomas that occur in 15% of children with NF1. In NF1, these tumors are characterized by slow growth and a usually benign course. Signs and symptoms occur in approximately half of affected children and are a result of tumor growth with compression or invasion of contiguous structures. These symptomatic tumors present with visual abnormalities, most commonly decreased visual acuity, or hypothalamic dysfunction. Hypothalamic dysfunction most commonly presents as development of signs of puberty, including breast development and pubic hair, at an early age (5 years or less). Such precocious puberty is indicative of chiasmatic involvement. Brain stem gliomas behave much like optic gliomas and follow a relatively indolent course. Hydrocephalus and cranial neuropathies are complications of these tumors.

Other malignancies reported to occur at a higher frequency in the NF1 population include leukemia, particularly juvenile myelomonocytic leukemia (JMML), rhabdomyosarcoma, and pheochromocytoma. Malignancies are a major cause of premature death in patients with NF1.

Nontumor Manifestations

The NF1 phenotype also includes many nontumor manifestations. The most common of these are

Figure 2 Superficial plexiform neurofibroma with overlying hyperpigmentation and hypertrichosis.

hyperpigmented skin lesions, skeletal abnormalities, learning disabilities, structural cerebral abnormalities, and vascular anomalies. Café-au-lait spots and skin fold freckles are important diagnostic criteria but are of no medical significance.

Skeletal abnormalities range in severity and most commonly involve the sphenoid wing, vertebrae, and long bones. Potentially serious skeletal lesions include thinning of long bone cortices, particularly the tibia, which can lead to pathological fractures and pseudoarthrosis and dystrophic scoliosis, which can progress to paraplegia. Short stature is also common.

Learning disabilities are very common, occurring in more than one half of children with NF1. Learning disabilities are diverse and include a global decline in general intellectual functioning as evidenced by lower IQ scores compared to a control population, visuospatial and visuoperceptual deficits, difficulty with planning and abstract concept formation, lower expressive and receptive language skills, and deficits in attention (Table 3). One-third of children with NF1 meet the diagnostic criteria for attention-deficit/hyperactivity disorder (ADHD). Verbal and visual memory seem to be unaffected in NF1 children. The learning deficits are static and persist into adulthood, although compensatory mechanisms are developed and performance seems to improve.

Macrocephaly is frequent in children with NF1 and relates to a significantly larger brain size. The ratio of gray to white matter volume is different when compared to controls and accounts for the larger brain size. White matter volume is increased in the frontal lobe whereas gray matter volume is increased in the temporal, parietal, and occipital lobes. In addition, high signal intensities (unidentified bright objects or UBOs) have been noted on T2-weighted magnetic resonance imaging (MRI) scans in 60–70% of children with NF1. These abnormal signals are found most commonly in the basal ganglia, brain stem, and cerebellum. The pathogenesis of these signal abnormalities is unknown and they seem to disappear over time. There is conflicting evidence on whether or not the presence of UBOs correlates with cognitive deficits. These lesions do not cause focal defects and are not necessary to establish the diagnosis of NF1; therefore, routine imaging searching for these lesions is currently not recommended.

The vascular complications of NF1 are uncommon but are an important cause of premature death and consist of vasculopathy and hypertension. The vasculopathy causes arterial stenoses, due to proliferation of cells in the intima. Involvement of the renal arteries can lead to hypertension, which most often presents

Table 3 Types of cognitive deficits in NF1

Global cognitive deficits (6% are mentally retarded with an IQ[a] < 70)
Specific learning disabilities
Attention deficits and ADHD[b]
Executive deficits (planning and abstract concept formation)
Visuospatial deficits
Fine motor coordination deficits
Slowing of motor speed

[a]IQ, intelligence quotient.
[b]ADHD, attention-deficit/hyperactivity disorder.

in childhood or pregnancy. Cerebral artery stenosis can lead to strokes or to moya moya syndrome, where collateral vessels develop around the stenotic areas and appear as a 'puff of smoke' on cerebral angiogram. Arterial dissection and hemorrhage is another complication of the vascular involvement of NF1.

Genetics of NF1

Genetic Transmission

NF1 is a classical autosomal dominant single gene disorder with complete penetrance but variable expression. NF1 has a high rate of new mutations and one half of NF1 patients are a result of a *de novo* mutation of the *NF1* gene. Affected individuals have a 50% chance of passing this disorder to any of their offspring. Parents of a sporadically affected child have a low chance of having another affected child, barring the rare occurrence of germ line mosaicism.

Segmental NF1 and Mosaicism

Segmental NF1 refers to the confinement of typical features of NF1 to a localized region or segment of the body. Café-au-lait spots, skin fold freckling, and neurofibromas are the most common features, but iris hamartomas, plexiform neurofibromas, and skeletal abnormalities have been reported in segmental NF1. At least some cases of segmental NF1 are due to somatic mosaicism of the *NF1* gene. Mosaicism refers to the presence of two genetically distinct cell populations and results from mutations occurring during mitotic cell division in the embryo with subsequent clonal expansion of the affected somatic cells. If the mutation occurs early in embryonic development, germ cells can also be affected and that individual is at risk of passing generalized NF1 to any offspring. Individuals with generalized features of NF1 have also been shown to result from somatic mosaicism and this may explain the occurrence of a subset of patients with sporadic NF1. It is thought that these patients may display a milder phenotype. There is currently no easy way to identify if an individual with segmental NF1 or sporadic generalized NF1 has a somatic or germ line mosaicism for an *NF1* gene mutation.

NF1 Gene Mutations

The *NF1* gene is approximately 350 kb in size, maps to chromosome 17q11.2, and encodes a 2818-amino-acid protein product designated neurofibromin. The large size of the gene, the diverse types of mutations, and equal distribution of mutations along the gene has made mutation detection difficult and laborious. The highest detection rate of pathogenic mutations has been reported by Messiaen et al. Using a multistep mutation detection protocol, 95% of *NF1* gene mutations were detected in a cohort of patients with NF1. Pathogenic mutations in the *NF1* gene have been found in most of the 60 exons and comprise a wide diversity of mutation types. These include complete gene deletions, chromosome rearrangements that disrupt the gene, smaller deletions or insertions, stop mutations, amino acid substitutions, and splicing mutations. The latter are particularly prevalent given the large number of exons. The majority of *NF1* gene mutations result in premature termination of translation and a truncated protein product. Molecular genetic testing is available on a clinical basis for individuals who do not yet meet diagnostic criteria for NF1 and for prenatal testing in offspring. Identification of the *NF1* gene mutation does not alter management or predict severity or progression of NF1.

Genotype–Phenotype Correlations

Genotype–phenotype correlations have been difficult to establish due to the complexity of both the phenotype and the gene. Complete gene deletions, consisting of about 1.5 Mb, including the *NF1* gene and multiple contiguous genes, are associated with a severe presentation, including dysmorphic facial features, mental retardation, cardiac anomalies, a large neurofibroma burden, and a substantially higher lifetime risk of developing MPNST. Another distinct phenotype, familial spinal neurofibromatosis, characterized by café-au-lait spots, paucity of dermal neurofibromas, and numerous dumbbell-shaped neurofibromas along the spinal roots, may also be associated with a distinct mutational spectrum, with enrichment for amino acid substitutions or stop mutations near the 3′ end of the gene. It is possible that other genotype–phenotype correlations will emerge as a larger number of carefully phenotyped patients are studied.

Variable Expression

NF1 is known for its highly variable expression. Families with multiple affected individuals demonstrate a wide range in the presence and severity of phenotypic features and complications. A parent with a mild presentation can have a child with a severe presentation, and vice versa. Mechanisms of variable expressivity in NF1 include alleleic heterogeneity, modifying genes, somatic mosaicism, haploinsufficiency, environmental factors, and stochastic factors.

Alleleic heterogeneity plays at least some role, as noted above, though the full extent of phenotypic variability attributable to allelic heterogeneity is unknown. Modifying genes may play an important

role. Concordance studies with identical twins have shown correlation between some features (e.g., café-au-lait spots, neurofibromas) but not others (e.g., plexiform neurofibromas). Identification of genetic modifiers is an area of active study. Somatic mosaicism, as explained above, may account for some of the cases of sporadic NF1 and possibly milder phenotypic expression. Haploinsufficiency has been used to explain the occurrence of some nontumor features of NF1. To date, no environmental factors have been identified that alter the expression of NF1, although tumors often grow during times of hormonal change, such as puberty and pregnancy. Lastly, stochastic factors have been postulated as a mechanism to explain the often random progression of an individual's NF1 phenotype.

Pathophysiology of NF1

Function of Neurofibromin

Neurofibromin is expressed in a variety of cell types but is most abundant in cells of the peripheral and central nervous system. Neurofibromin shares sequence homology with members of the Ras GTPase-activating protein (GAP) family. This central GAP-related domain (GRD) is the only part of the protein whose function is known. Neurofibromin functions as a GAP protein and inactivates the intracellular signal transduction protein Ras by converting the active GTP-bound form into its inactive GDP-bound form. This in turn leads to the downregulation of Ras activity. Loss of neurofibromin activity increases Ras activity, which in turn promotes the transcription of a number of genes required for cell growth and proliferation.

NF1 Is a Tumor Suppressor Gene

Substantial evidence has accumulated that the *NF1* gene functions as a tumor suppressor and that this accounts for the propensity of tumor development in patients with NF1. Loss of heterozygosity (LOH) or mutation of both *NF1* alleles was initially demonstrated in some MPNSTs and subsequently in the Schwann cells of dermal neurofibromas. JMML cells in NF1 patients also have mutation of both *NF1* alleles. In addition, patients with JMML without NF1 may also have *NF1* mutations, activating mutations of Ras, or of the Ras activating gene PTPN11, suggesting that activation of the Ras pathway is critical to the development of JMML.

Mouse models have generated additional support for the tumor suppressor hypothesis. Mice rendered heterozygous for *Nf1*, the murine equivalent of the *NF1* gene, do not develop neurofibromas, although they eventually succumb to leukemias or sarcoma. Homozygously mutated mice die of cardiac anomalies *in utero*, but chimeras with homozygously mutated *Nf1* alleles survive and develop plexiform neurofibromas. More recently, conditional knockout mice have been created with one *Nf1* allele mutated in the germ line, and the other allele subject to mutation in developing Schwann cells. These animals develop neurofibromas, providing powerful support for the tumor suppressor hypothesis. There is also a mouse model for optic glioma, created by conditional knockout of the *Nf1* gene with an astrocyte-specific promoter.

The pathogenesis of nontumor manifestations of NF1 is not as well understood. Evidence suggests that the pigmentary disturbances and learning deficits do not result directly from a tumor suppressor mechanism. Melanocytes from café-au-lait spots have thus far been shown to be heterozygous for an *NF1* mutation. Similarly, mice that are heterozygous for an *Nf1* (Nf1 +/−) mutation have cognitive deficits, specifically visuospatial deficits. These learning deficits have been shown to be Ras dependent and reversible with the inhibition of Ras. It is not known whether homozygous *NF1* mutations occur within focal skeletal lesions.

Management of NF1

The mainstay of care in NF1 is anticipatory guidance, surveillance for treatable complications, and genetic counseling. Optimum management of NF1 involves a multidisciplinary approach.

Surveillance

Once the diagnosis of NF1 is confirmed by clinical criteria or is suspected, for instance in a young child with multiple café-au-lait spots without an affected parent, the patient should be evaluated on an annual (or more often, if indicated) basis, preferably by a physician familiar with NF1. The history should focus on symptoms associated with NF1, such as cognitive or psychomotor deficits, pain, visual complaints, change in bowel or bladder function, weakness, seizures, headaches, abdominal pain, and in children a thorough review of developmental history. The physical examination is directed at identifying potential medical complications of NF1, such as hypertension, scoliosis, ocular abnormalities, focal neurologic deficits, developmental disabilities, proptosis, short stature, macrocephaly, and signs of precocious puberty. Routine evaluations should include a yearly ophthalmologic evaluation in children under the age of 7, developmental assessment, and blood

pressure monitoring. Routine screening with computed tomography, magnetic resonance imaging, electroencephalogram, and visual evoked potentials is not advocated unless indicated by signs or symptoms detected on history or physical examination.

Treatment of Manifestations

Café-au-lait spots cause no medical problems but can be located in cosmetically sensitive areas. Lasers for pigmented lesions exist and have been used to lighten or remove café-au-lait spots with variable results.

Cutaneous neurofibromas can cause pruritus, pain, or cosmetic disfigurement. Surgical removal or other forms of destructive therapy (e.g., carbon dioxide laser) is the only available option for removing symptomatic lesions. Associated risks include scarring and the possibility of recurrence.

Plexiform neurofibromas can cause significant deformity as well as compression of nerves or other vital structures and erosion of bones. Surgical excision is the mainstay for symptomatic or progressive lesions, but due to the infiltrative nature of many of these tumors complete surgical excision is frequently not possible and recurrence is common. The use of ketotifen, a mast cell stabilizer, retinoic acid, interferon alpha, and thalidomide have been reported as potential treatments for progressive or symptomatic plexiform neurofibromas. No objective change in the volume of the tumors was reported, but a minority of patients treated with some of these agents experienced either symptomatic improvement or a slight decrease in the surface dimensions of the tumor. There are several ongoing clinical trials evaluating the use of other medical agents for the treatment of plexiform neurofibromas. Those under investigation include tipifarnib, a farnesyl protein transferase inhibitor, pirfenidone, an antifibrotic agent, pegylated interferon, and combination chemotherapy with methotrexate and vinblastine. Clinical trials for NF1 are listed on the Children's Tumor Foundation website.

Vigilance for malignant transformation into an MPNST is vital in patients with plexiform neurofibromas. The best clinical indicators of malignant transformation are persistent pain, sudden growth or swelling, and the onset of new neurological deficits. If present, an MRI with contrast should be obtained and the presence of inhomogeneous contrast enhancement is suggestive of malignancy. Fluorodeoxyglucose positron emission tomography (PET) has also shown promise in distinguishing benign from malignant lesions and deserves further study. A tissue biopsy is necessary for definitive diagnosis but false negatives can occur due to occurrence of MPNSTs in preexisting plexiform neurofibromas and sampling error.

MPNSTs have a poor prognosis with an estimated 5-year survival rate of less than 20%. Early diagnosis along with complete surgical excision portends the best prognosis. It is not clear that radiation or chemotherapy markedly improves outcome. There are ongoing clinical trials evaluating the use of other agents for MPNSTs.

Optic pathway gliomas occur in 15% of patients with NF1 and approximately half of affected individuals will develop signs or symptoms associated with their tumor, generally during the first decade. Asymptomatic children with known or suspected NF1 should have an annual ophthalmologic evaluation. If signs or symptoms of visual disturbance or neuroendocrine dysfunction are noted, an MRI is indicated. Symptomatic lesions should be followed more frequently with examination and imaging. Treatment is reserved for lesions that are symptomatic and progressive. Treatment options include surgical resection, radiation, and chemotherapy. Carboplatin in combination with vincristine has been reported for the treatment of progressive tumors, usually with a good outcome.

Learning disabilities are one of the most common features in children with NF1. Any child with NF1 suspected of having learning disabilities should be examined early and intervention begun promptly. Children with NF1 and ADHD have been shown to respond to stimulant medication with improved cognitive, academic, and social performance. Lovastatin, an HMG-CoA reductase inhibitor that also inhibits Ras, was shown to reverse the learning deficits in *Nf1*+/− mice and is currently being investigated in clinical trials in individuals with NF1 and learning disabilities.

Management of skeletal dysplasias has been symptomatic, with the tibial and orbital dysplasias proving to be the most problematic. Affected individuals should be referred to an experienced orthopedic surgeon. Prevention of fracture in children with long bone dysplasia is of paramount importance because the fractures tend to heal poorly and can result in pseudoarthrosis.

The importance of counseling the patient and family should not be underestimated. Education should include an explanation of the signs and symptoms of NF1, possible complications, the effects of puberty and pregnancy on certain features of NF1, and the importance of yearly evaluations to monitor for these complications. Individuals with NF1 should be provided genetic counseling regarding the 50% recurrence risk. Prenatal testing is available by linkage

analysis or direct mutational analysis if the mutation can be identified. It is not possible to predict severity in an individual based on family history or mutation identification.

See also: Neurofibromatosis Type 1: Molecular and Cellular Biology; Neurofibromatosis Type 2: Further Questions and Answers; Neurofibromatosis Type 2; Tumors of the Brain and the Spinal Cord.

Further Reading

DeSchepper S, Boucneau J, Lambert J, Messiaen L, and Naeyaert JM (2005) Pigment cell-related manifestations in neurofibromatosis type 1: An overview. *Pigment Cell Research* 18: 13–24.

Friedman JM, Gutmann DH, MacCollin M, and Riccardi VM (eds.) (1999) *Neurofibromatosis: Phenotype, Natural History, and Pathogenesis*. Baltimore, MD: Johns Hopkins University Press.

Gutmann DH, Aylsworth A, Carey JC, et al. (1997) The diagnostic evaluation and multidisciplinary management of neurofibromatosis 1 and neurofibromatosis 2. *The Journal of the American Medical Association* 278: 51–57.

Hyman SL, Shores A, and North KN (2005) The nature and frequency of cognitive deficits in children with neurofibromatosis type 1. *Neurology* 65: 1037–1044.

Listernick R, Louis DN, Packer RJ, and Gutmann DH (1997) Optic pathway gliomas in children with neurofibromatosis 1: Consensus statement from the NF1 Optic Pathway Glioma Task Force. *Annals of Neurology* 41: 143–149.

Maria BL (2002) Neurobiology of disease in children: Neurofibromatosis 1. *Journal of Child Neurology* 17: 547–653.

Messiaen LM, Callens T, Mortier G, et al. (2000) Exhaustive mutation analysis of the NF1 gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. *Human Mutation* 15: 541–555.

North KN, Riccardi V, Samango-Sprouse C, et al. (1997) Cognitive function and academic performance in neurofibromatosis 1: Consensus statement from the NF1 Cognitive Disorders Task Force. *Neurology* 48: 1121–1127.

Packer RJ, Gutmann DH, Rubenstein A, et al. (2002) Plexiform neurofibromas in NF1: Toward biologic-based therapy. *Neurology* 58: 1461–1470.

Xu GF, O'Connell P, Viskochil D, et al. (1990) The neurofibromatosis type 1 gene encodes a protein related to GAP. *Cell* 62: 599–608.

Zhu Y and Parada LF (2001) Neurofibromin, a tumor suppressor in the nervous system. *Experimental Cell Research* 264: 19–28.

Relevant Websites

http://www.nfinc.org – Neurofibromatosis, Inc.
http://www.ctf.org – The Children's Tumor Foundation.
http://www.understandingnf1.org – Understanding NF1.

Neurofibromatosis Type 1: Molecular and Cellular Biology

D H Viskochil and D A Stevenson, University of Utah, Salt Lake City, UT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

As one of the most common heritable conditions, neurofibromatosis type 1 (NF1) displays a number of genetic principles that capture the molecular biology of human disease. Although a tremendous amount of knowledge has been gained since the gene was cloned, therapeutic options for the clinical presentations are limited. This article reviews the clinical aspects of NF1, details the molecular biology of the gene and its encoded product, neurofibromin, and concludes by examining the pathogenesis of this genetic condition.

NF1 is an autosomal dominant condition that afflicts approximately 1 in 3500 individuals worldwide. The physical deformities of Joseph Merrick, born in 1862 in England and documented in film and literature as *The Elephant Man*, due to affliction with a disfiguring disease causing tissue and vessel enlargements and grayish skin tone, were originally thought to be secondary to NF1. NF1 was thus termed in years past as the 'elephant man's disease.' This was an erroneous label, as further investigation showed that Joseph Merrick's physical deformities more closely resembled the features of Proteus syndrome, not NF1. Before the age of 20 years, individuals with NF1 typically develop cutaneous manifestations, neurofibromas, and café-au-lait spots. Approximately two-thirds of NF1 individuals will have only these cutaneous features. The remaining one-third will display a varied array of medical complications that affect multiple organ systems. Even though the condition was first described in the medical literature in the nineteenth century by Dr. von Recklinghausen, the biomedical characterization of the underlying defect was hindered by the variability of clinical expression and pleiotropy of NF1. NF1 captures many of the classical principles of human genetics, including expressivity, penetrance, pleiotropy, and sporadic occurrence.

A number of factors made this condition an early candidate for the 'reverse genetics' research approach, which applies molecular biological techniques to evaluate the potential function of a disease-causing protein by examination of its gene template. The prevalence of NF1, its straightforward clinical recognition, the availability of autosomal dominant pedigrees, and the minimal progress made through standard biochemical techniques all suggested feasibility of applying gene cloning by the mapping approach to the task of understanding the basic biology of the *NF1* gene product.

Clinical Aspects

Diagnostic Criteria

The diagnosis of NF1 is classically made on clinical grounds through a defined set of diagnostic criteria. Genomic and cDNA-based testing of the *NF1* gene is available clinically, but usually is not necessary for a clinical diagnosis in the majority of individuals. The typical clinical features generally enable practitioners to easily make the diagnosis in adults based on a simple physical examination. The diagnosis generally can be made in approximately 90% of NF1 individuals by the time they are 6–7 years old.

There are seven diagnostic features of NF1 that were developed in a consensus conference sponsored by the U.S. National Institutes of Health (NIH). In order to fulfill the diagnostic criteria for NF1, one must have two or more of the following features:

1. Six or more café-au-lait macules (1.5 cm or larger in postpubertal patients; 0.5 cm or larger in prepubertal patients).
2. Two or more neurofibromas or one or more plexiform neurofibromas.
3. Freckling in the axillary or groin region.
4. Optic pathway tumor.
5. Two or more Lisch nodules (benign hamartomas).
6. A distinctive osseous lesion (long bone dysplasia/pseudarthrosis or sphenoid wing dysplasia).
7. A first-degree relative with NF1.

Table 1 shows approximate percentages of NF1 individuals who have the typical features; the significant variation depends on the mode of ascertainment and the clinical setting of the evaluators. The Baylor study served a referral-based population, whereas the Southeast Wales study is a more population-based ascertainment. Other common manifestations of NF1, not included in the diagnostic criteria, include but are not limited to educational difficulty (40–60%), macrocephaly (38–45%), short stature (18–34%), scoliosis (10–30%), headache (22%), seizures (5–7%), neoplasms (5%), hemangiomas (5–10%), sensorineural hearing loss (5%), precocious puberty (4–5%), hypertension (4%), hydrocephalus (3–4%), and congenital heart disease (2–4%).

Table 1 Seven diagnostic criteria for neurofibromatosis type 1 and the prevalence of specific feature at the time of diagnosis according to two studies

Clinical feature	*Prevalence (%) in studies*	
	Baylor[a]	*SE Wales*[b]
Café-au-lait spots, >5 in number	99	85
>0.5 cm for prepubertal children	–	98
>1.5 cm for postpubertal adults	–	78
Neurofibromas: two of any type, or one plexiform	67	78
	–	35
Freckling in the axillary/inguinal regions	64	70
Optic nerve glioma	14	2
Two or more Lisch nodules (iris hamartomas)	84	85
Distinctive skeletal dysplasia	15	1
First-degree relative with NF1	47	30

[a]Study reported by Ricciardi (1992) involved 970 patients at a university-based clinic in a major city in the United States.
[b]Study reported by Huson and Hughes (1994) involved 135 patients ascertained through hospital admissions, relatives of probands, and general practitioners.

Café-au-lait spots Café-au-lait spots (pigmented macules resembling 'birthmarks') are the hallmark of this condition, and rightly so. Individuals without NF1 rarely have more than two birthmarks, whereas NF1 patients typically have more than five. These lesions are not pathological. Other disorders have café-au-lait spots as a feature, and since café-au-lait spots present as the first manifestation of NF1 without other clinical manifestations initially, other conditions need to be considered. A partial list of other syndromes that have café-au-lait spots or similar pigmentary findings includes Russell–Silver, Bloom, Noonan, Watson, LEOPARD, Sotos, Dubowitz, Klippel–Trenaunay–Weber, Proteus, and McCune–Albright syndromes.

Neurofibromas The presence of cutaneous neurofibromas is the other hallmark of NF1; these benign tumors generally comprise Schwann cells, extracellular matrix, fibroblasts, perineural cells, and mast cells. The neurofibromas typically appear after the development of the café-au-lait macules, beginning in late childhood and adolescence. Neurofibromas can remain under the skin or take on a pedunculated appearance, and they generally are not painful, though they are frequently noted to itch. Plexiform neurofibromas tend to be more aggressive, albeit nonmalignant, tumors that have a propensity to infiltrate surrounding tissue and cause pain. Plexiform neurofibromas can transform into malignant peripheral nerve sheath tumors. Particular attention to unexplained pain and rapid tumor growth will help identify patients with malignant peripheral nerve sheath tumors.

Intertriginous freckling Axillary and groin freckling (Crowe's sign) are common NF1-related findings not usually seen in the general population. Freckling patterns around the neck are also quite characteristic in the NF1 population.

Optic pathway tumors Optic nerve gliomas are low-grade tumors that arise anywhere along the optic nerve tract in the brain; approximately half of the tumors cause some symptoms, even though for unknown reasons they are much less aggressive than optic nerve gliomas are in individuals who do not have NF1. Optic pathway tumors often can be picked up by clinical exam by an experienced ophthalmologist (pale optic nerve) and then confirmed by magnetic resonance imaging (MRI) scan. These tumors affect approximately 15% of the NF1 population.

Lisch nodules Lisch nodules are harmless hamartomatous growths on the iris that are easily detected by ophthalmologic slit lamp exam; they serve as a reliable diagnostic marker for this condition.

Distinctive osseous lesion Helpful diagnostic signs also include the rare, NF1-specific skeletal dysplasias, which include sphenoid wing dysplasia and tibial pseudarthrosis. Other osseous findings include scoliosis, short stature (compared to familial background), bone cysts, and relative macrocephaly. Of NF1 individuals, approximately one-third have one or more orthopedic findings.

Long bone dysplasia is seen in approximately 5% of NF1 individuals and classically involves the tibia and fibula, although involvement of other long bones has been reported. The typical presentation is anterior lateral bowing leading to fracture and nonunion or pseudarthrosis. Tibial pseudarthrosis is strongly associated with NF1 – 50% of individuals with tibial pseudarthrosis have NF1. The bowing usually

presents before 4 years of age and fractures have been seen *in utero*. Approximately half of NF1 individuals with long bone bowing will experience a fracture before 2 years of age.

Both idiopathic and dystrophic forms of scoliosis are seen in individuals with NF1. The dystrophic form is usually a sharp, short-angle scoliosis. The nondystrophic form of scoliosis is more common in children with NF1, but dystrophic scoliosis is progressively debilitating and requires a more aggressive therapeutic approach.

Sphenoid wing dysplasia is a rare manifestation, usually unilateral and associated with a plexiform neurofibroma of the temporal-orbital region in about 50% of cases. Exophthalmosis is a rare complication that may require intervention, but the sphenoid wing dysplasia usually does not cause significant clinical complications or require therapeutic management.

Family history of NF1 Finally, the diagnostic criteria incorporate the well-documented dominant inheritance pattern in NF1: first-degree relatives with NF1 carry as much diagnostic weight as do multiple café-au-lait spots. The recognition that this condition is fully penetrant in older adolescents enables one to confidently provide reproductive risk counseling for families with members who have NF1.

Medical Complications

NF1 patients are prone to a number of medical complications that are quite varied, although one individual seldom is affected with more than one complication. They are also more likely to have a learning disability; approximately 40–50% of diagnosed patients have some form of learning disorder, which can be effectively managed if discovered early in a child's formative education. Short stature, macrocephaly, and hypertension are some of the more common complications. Seizures are also seen at a slightly increased frequency in NF1 individuals.

There is an association between NF1 and cancer, leading to a small but significantly increased relative risk to develop specific malignancies. Malignant peripheral nerve sheath tumors (MPNSTs), embryonal rhabdomyosarcomas, myelogenous leukemias, and pheochromocytomas are all found at a higher incidence in the NF1 population. Likewise, low-grade astrocytomas are relatively common in the NF1 population.

Although there is an increased incidence of intracranial tumors, controversy exists over the routine use of neuroimaging studies in NF1 individuals. Some groups perform an initial MRI upon diagnosis. Hyperintense T_2-weighted signals are commonly seen in NF1 individuals, but have unknown clinical consequences. Other groups advise against routine neuroimaging, as these and other incidental findings can cause anxiety when there is no clinical indication or subsequent approach to management, and therefore suggest relying on clinical symptoms to direct imaging protocol.

Genetic Principles and *NF1*

The clinical aspects of NF1 encompass many genetic principles, including Mendelian inheritance, variable expressivity, age-related penetrance, pleiotropy, mosaicism, and sporadic occurrence. Before we explore the molecular biology aspects that might explain the pathophysiology of NF1, it will be beneficial to review such principles in order to gain a better understanding of the clinical features of this condition.

Variable Expressivity

Variable expressivity of NF1 manifests in many ways, from the diagnostic criteria to the unpredictable multiple medical complications that arise. The presence of severe manifestations involving one organ system in one patient does not correlate with severity in other clinical aspects. For example, an NF1 patient who has more than 30 café-au-lait spots with a high degree of freckling may have few, small neurofibromas. Furthermore, as a medical complication, renal artery stenosis does not predict that an NF1 patient is more prone to other vasculopathies. It is difficult to explain the intrafamilial variability of this dominant disorder; affected siblings typically display remarkable differences in their clinical expression, as different as if they were not related. Historically, practitioners could not predict the natural progression of an individual's clinical course based on another family member's degree of involvement. However, larger sampling and more detailed clinical profiles of the NF1 population have revealed more similarity among first-degree relatives than had been appreciated, suggesting that other unlinked genetic loci play some role in the expression of NF1 features. The identification and characterization of such loci, coupled with knowledge of the *NF1* genotype, may lead to better prediction of the natural history of NF1 in any given individual.

Penetrance

Even though NF1 is 100% penetrant, there are a number of issues regarding the age-related penetrance of NF1 features, and perhaps the most important pertains to diagnosis. Typically, children present at an age less than 5 years with multiple café-au-lait spots as their

only feature, thus they do not fulfill the criteria for NF1. Axillary and groin freckling, and Lisch nodules, tend to arise by the age of 10 years, whereas neurofibromas are noticed sometime during adolescence. Thus there is a time period during which families must deal with the possibility of a diagnosis of NF1 while keeping in perspective the variability of clinical expression of NF1, which includes potential medical complications that are not part of the diagnostic criteria. For example, a child not fulfilling the diagnostic criteria may display learning disabilities prior to the onset of a second sign of NF1.

Some features of NF1, such as optic nerve glioma, sphenoid wing dysplasia, tibial pseudarthrosis, and plexiform neurofibroma, tend to be congenital and are thought to represent true malformations of embryological development. On the other hand, cutaneous neurofibromas that initially arise during adolescence will persist, and additional tumors tend to arise throughout the patient's lifetime in a progressive fashion. Additionally, women with NF1 may also experience an exacerbation of neurofibroma numbers and size during pregnancy. It is not understood how age and/or hormones alter the clinical expression of this condition, and one aspect of future research is to clearly delineate the natural history of NF1 in attempts to correlate it with the molecular biology discoveries.

Pleiotropy/Mosaicism/Mutation

Pleiotropy is a term used to describe the multiple organ systems affected by this condition: peripheral nervous system (neurofibromas), vasculature (renal arteries), brain (learning disabilities), endocrine (pheochromocytoma), skin (café-au-lait spots), and skeleton (dysplasias and scoliosis). Mosaicism also plays some role in NF1. There are a number of patients who have one or two NF1 manifestations occurring in regional parts of their body; this is termed 'segmental NF1.' Usually segmental cases of NF1 are restricted to cutaneous features, although solitary and deep plexiform neurofibromas have also been detected in individuals who do not fulfill the criteria of NF1. There has been one case of two affected offspring whose father had no clinical features of NF1, although approximately 10% of his sperm carried an identical disease-causing *NF1* mutation. The high mutation rate (1/10 000 gametes) involving the germ line, as determined by the number of *de novo* cases, suggests that multiple cells in normal individuals could conceivably carry nonmanifesting *NF1* mutations. Approximately half of patients who present to medical clinics have no family history of NF1, and, in sporadic individuals who have had testing to identify the chromosome that carries the disease-causing allele, it has been determined that in more than 90% of the cases the mutation arises in the paternal chromosome. The etiology of this observation has not been established: pedigree analysis does not support the hypothesis of parent-of-origin effects in NF1, selection against maternal oocytes with *NF1* mutations is not reflected in altered fertility or miscarriage rates, and paternal age effect is not apparent, as it is in the case of achondroplasia. As applicable to other aspects of NF1, the cloning and characterization of the *NF1* gene allow investigators to explore mechanisms that could explain the somewhat unusual genetic principles of NF1.

Molecular Biology of NF1

Gene Cloning

The *NF1* gene was identified by application of the paradigm of gene cloning by the mapping approach. As an autosomal dominant condition, genetic linkage analysis of combined pedigrees mapped *NF1* to the centromeric region of the long arm of chromosome 17, band 11.2. The identification of two NF1 patients who had balanced translocations provided researchers with the opportunity to pinpoint the *NF1* locus within this large region of chromosome 17 (see Figure 1). The *NF1* locus has turned out to be one of the most complex loci described. Screening various cDNA libraries with genomic DNA probes from the region encompassing the two *NF1* translocation breakpoints identified four genes: three small genes, *EVI2A*, *EVI2B*, and *OMGP*, mapping between the breakpoints, and *NF1*, a large gene that spans both breakpoints.

Gene Structure

NF1 spans approximately 335 kb of genomic DNA and is divided into 57 exons (see Figure 1). Three exons, 9A, 23A, and 48A, are alternatively spliced as insertion exons, but their respective expression pattern in various tissue has not been completely elucidated. Most cDNAs used in assembling the *NF1* coding sequence were isolated from fetal brain cDNA libraries; screening of other libraries yields few clones. A shorter *NF1* product with a poly(A) tail that terminates 14 bp downstream of a polyadenylation signal in intron 10C has also been identified from a placental cDNA library. Its role in the NF1 phenotype has not been examined, however.

NF1 contains two large introns greater than 40 kb, introns 1 and 27B. Intron 27B harbors the t(1;17) breakpoint, two short terminal repeats useful in linkage studies, and the three other genes that map

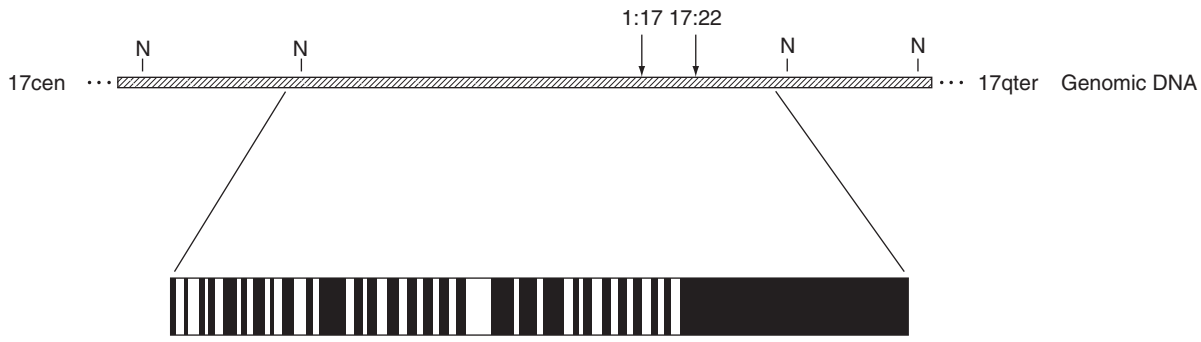

Figure 1 Physical map of the *NF1* gene. The genomic DNA is represented as a hatched bar. The cDNA is represented as a bar code, reflecting the sizes of each of the exons. The 5′-untranslated and first 30 codons are in the first exon, which maps just centromeric to a NotI site embedded in a CpG island. The stop codon is in exon 49 and is continuous with 3.5 kb of 3′-untranslated sequence that maps approximately 15 kb centromeric to a NotI site. The primary *NF1* transcript spans approximately 335 kb of genomic DNA. Translocation breakpoints map to introns 27B and 31.

between the translocation breakpoints. These genes are oriented in head-to-tail fashion, with the 3′ end of *OMGP* lying approximately 2.5 kb telomeric of the t (1;17) breakpoint and 20 kb centromeric of the 5′ end of *EVI2A*. Transcription of all three genes is in the orientation opposite to that of *NF1;* however, their influence on *NF1* expression has not been studied in detail. Intron–exon boundaries for *NF1* have been established, and assays are now available to amplify each exon independently, using genomic DNA as template.

The promoter region of *NF1* is embedded in a CpG island, and, similar to other 'housekeeping' genes, it does not harbor the typical regulatory TATA or CCAAT boxes. A number of DNA sequence motifs suggest binding sites for transcription factors, including SP1 and AP2, and, of note, there is a cyclic AMP response element. Functional studies with various *NF1* promoter sequences have not been reported, and therefore important motifs for *NF1* expression have yet to be elucidated.

The *NF1* mRNA encodes a peptide 2818 amino acids long, excluding the insertion exon-encoded segments of 10, 21, and 18 amino acids for exons 9A, 23A, and 48A, respectively. Exons 21–26 encode one domain lying between amino acids 1240 and 1430 that has approximately 30% homology to the catalytic domain of p120GAP, a protein that activates the intrinsic GTPase of p21Ras (see Figure 2). In addition to the shared catalytic domain, adjacent amino acid sequence extending in both directions from this core homologous region also has homology to two yeast proteins, IRA1 and IRA2, so named because they play an inhibitory role in the yeast RAS activation pathway. The 3′-untranslated region of *NF1* is 3500 bp (relatively long for a mammalian mRNA),

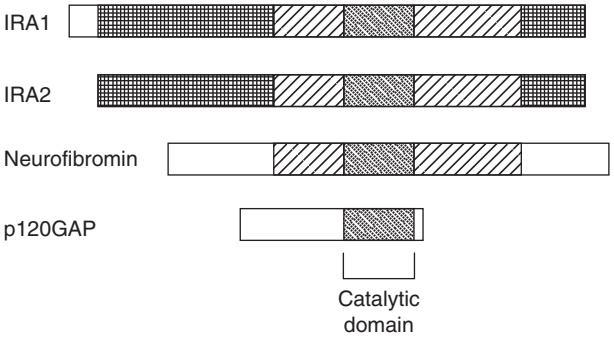

Figure 2 Schematic depicting regions of amino acid homology between yeast proteins IRA1 and IRA2, neurofibromin, and p120GAP. The p21Ras GTPase-activating catalytic domain is represented by stippling. IRA homology is depicted by hatching.

and the polyadenylation signal, lying 11 to 16 bases upstream of the poly(A) tail, is AGUAAA, instead of AUAAA, the typical sequence that most efficiently specifies polyadenylation.

Neurofibromin Function

The homology between a domain of the *NF1*-encoded peptide, neurofibromin, and the catalytic domain of p120GAP suggested that neurofibromin may also activate the GTPase of p21Ras. To prove this, the GAP-related domain of neurofibromin was expressed and shown to stimulate the GTPase activity of H-Ras, N-Ras, and yeast RAS1 and RAS2 *in vitro*. This observation clearly placed neurofibromin in the Ras signal transduction pathway, and it acts as a 'negative regulator' of p21Ras (see Figure 3). Even though p21Ras has been studied for many years, there is much still to be learned about the Ras signal transduction pathway, and, as more data accumulate,

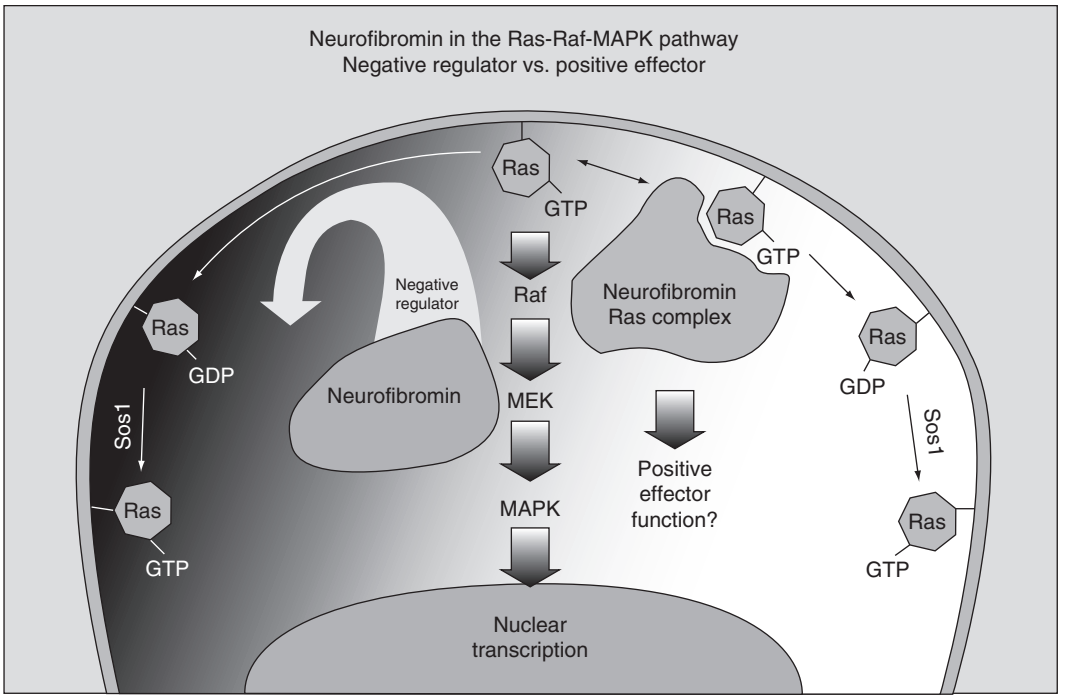

Figure 3 Diagram depicting the potential roles of neurofibromin in the p21Ras signal transduction pathway. One potential role is the 'negative regulation' of p21Ras–GTP simply by stimulating the intrinsic GTPase of p21Ras. A second possibility is a 'positive effector' function, whereby interaction of neurofibromin with p21Ras could send a separate downstream signal independent of the Raf/MEK/MAPK pathway. Presently, a specific role for neurofibromin has not been elucidated. Sos1, Ras GTP exchange factor; MAP, mitogen-activated protein; MEK, MAP kinase.

the role of neurofibromin should be clarified. The GAP-related domain of neurofibromin comprises less than one-fifth of the entire peptide. Homology with the yeast IRA proteins suggests that the remainder of neurofibromin could play a significant role in modifying its interaction with p21Ras.

Gene Expression

NF1 expression has been evaluated at both the mRNA and protein levels. mRNA studies include both Northern analysis and reverse transcriptase (RT)-polymerase chain reaction (PCR), whereas protein studies include both immunohistochemistry and Western analysis using antibodies generated against various predicted peptide domains, and various *in vitro* p21Ras GTPase activation assays. When the most sensitive method of RT-PCR is used, *NF1* appears to be ubiquitously expressed in numerous adult tissues. However, the functional role that neurofibromin may play in different tissues is not so easy to demonstrate. Antibody studies do not demonstrate the same degree of expression. It is clear from Western analysis, however, that neurofibromin antibodies are not very sensitive, and the absence of staining may not reflect absence of neurofibromin activity in that tissue. Nevertheless, the highest levels of immunodetectable neurofibromin in adult rodents have been demonstrated in neurons and oligodendrocytes of the central nervous system, and in nonmyelinating Schwann cells in the peripheral nervous system.

In contrast to the limited immunodetection in adult tissue, neurofibromin is widely detected in numerous rodent embryonic tissues. *NF1* expression studies have not been performed in human embryos or fetuses, although the gene is presumed to be expressed in higher levels in developmental stages than in the adult, since the *NF1* cDNAs have been most readily isolated from fetal retina and fetal brain cDNA libraries. Functional assays for neurofibromin activity have shown that various tissues contain p21Ras GAP activity that is not due to p120GAP, which suggests the presence of neurofibromin-dependent p21Ras GTPase activation. However, these assays are dependent on the lipid inhibition of NF1–GAP activity in cell lysates, and the specificity of this assay for NF1–GAP has not been demonstrated as additional RasGAPs have been identified.

Pathophysiology of NF1

It is clear that mutations in the *NF*1 gene cause the phenotype that is clinically diagnosed as NF1. The mechanisms involved are not understood. Standard clinical management of NF1 presently involves

symptomatic treatment after medical complications have arisen, yet, until the basic biology relating to the pathogenesis of NF1 is unraveled, clinicians are powerless to rationally intervene in a proactive manner. Important considerations in delineating the processes leading to expression of the clinical features of NF1 include genotype–phenotype correlations, the role of neurofibromin in tumorigenesis, and the evaluation of animal models.

Genotype–Phenotype Relationships

The *NF1* mutational spectrum that enabled investigators to identify the *NF1* gene included multiple-exon deletions, interrupting translocation, a frame-shifting insertion, and both missense and nonsense base pair substitutions. Mutation screening has included analysis of both mRNA and DNA by various techniques, including single-strand conformational polymorphism, chemical cleavage of mismatches, size shift of PCR products, Southern analysis, and detection of two mRNA alleles as a check for mutations involving the transcriptional regulation. Once verified, *NF1* mutations can be deposited in a mutation database sponsored as a consortium by the Children's Tumor Foundation. One group screened 67 NF1 patients and found 10 recurrent mutations (approximately 3% of the germ line mutations). Mutations seem to be spread throughout the *NF1* gene, with few apparent 'hot spots' for mutation. Examples of exons with recurrent mutations previously reported include exons 4b, 7, 10a–10c, 13, 22, 23.2, 27a, 29, 31, 37, and 39.

Approximately two-thirds of NF1 patients have mutations resulting in truncated neurofibromin peptides. Notably, some mutations lead to stable mRNAs that can translate smaller peptides, including the GTPase-associated protein (GAP)-related domain (GRD). This implies that domains outside the *NF1* GRD are important either as modifiers of the interaction with p21Ras or as effectors. The observation that many *NF1* mutations are truncating has been extended to the application of screening techniques using transcriptional/translational systems that detect premature translational stops in PCR products from segments of *NF1* mRNA.

The evaluation of various *NF1* mutations with respect to the respective clinical expression of NF1 in each individual has not demonstrated a clear genotype–phenotype correlation. Different patterns of associations of some clinical features have been reported in NF1 families, but further investigation is needed. Whole-gene microdeletions in *NF1* have been associated with mental retardation, minor facial abnormalities, plexiform neurofibromas, and MPNSTs. The phenotype in patients with microdeletions that involve the entire *NF1* gene potentially have contributions from contiguous genes affecting this phenotype.

There appears to be some randomness to number and location of café-au-lait spots and neurofibromas, to the presence of multiple diagnostic features, and to the severity of medical complications. One possible explanation is simply the nature of stochastic processes, whereby the balance of signal transduction is altered as a result of variable levels of expression of both *NF1* and modifying genes involved in a complex biochemical pathway. Having a constitutional *NF1* mutant allele could conceivably lower the 'threshold' at which such seemingly random events could alter the normal signal. For example, one mechanism to explain the age-related penetrance of neurofibromas could be the random somatic mutation of either the normal *NF1* allele or other modifying loci resulting in 'loss of signal' within the cell, hence unregulated growth. This scenario fits the paradigm of a 'tumor suppressor' model, and there is some evidence supporting the role of *NF1* as a tumor suppressor in certain tissue.

The Role of *NF1* in Tumorigenesis

The association of cancer and NF1 was an important motivating factor in the cloning and characterization of the *NF1* gene. Since NF1 is a familial cancer predisposition syndrome, it was hoped that identifying the biochemical pathway involving the *NF1* gene product would provide researchers with insight into early genetic events in tumorigenesis.

The *NF1* mutations are generally inactivating. This would suggest that the cellular biochemical processes are operating with half of the normal amount of neurofibromin. If neurofibromin were acting as an enzyme, this would not be important, as evidenced by the many examples of recessive inheritance of disorders involving inborn errors of metabolism; however, the dominant inheritance pattern of *NF1* suggests that a critical level of neurofibromin is necessary to prevent the varied clinical expression of many features. The 'tumor suppressor' model would predict that any process that lowers the function of neurofibromin below the threshold of signal transduction would culminate in a resultant altered cellular phenotype (i.e., growth). One process to account for this is the loss of the normal allele, and there is accumulating evidence that many tumors associated with NF1 involve a deletion of the *NF1* locus as a 'second hit.'

The best evidence supporting the role of *NF1* as a tumor suppressor is the observation that the *NF1* allele inherited from the unaffected parent is deleted in the clonal expansion of myelogenous leukemia cells in NF1 patients whose other parent has NF1.

Presumably, the *NF1* allele inherited from the affected parent carries an inactivating mutation, and the subsequent somatic deletion of the normal allele in a bone marrow precursor cell allows unregulated cell replication. The identification of the 'second hit' in a MPNST from an NF1 patient, and the demonstration of loss of immunoprecipitable neurofibromin with concomitant altered p21Ras–GTP/p21Ras–GDP levels in an MPNST cell line derived from a different NF1 patient, also support the tumor suppressor function of neurofibromin. As compelling as these observations are, the accumulation of other somatic mutations, unrelated to *NF1,* may also play a role in the development of malignancies in NF1. The inactivation of the normal *NF1* allele may be necessary but not sufficient to initiate the process of tumorigenesis in NF1-related sarcomas.

The common tumors of NF1, neurofibromas and gliomas, do not conform to the model of tumor progression that is apparent with the malignant tumors. As benign growths, these tumors are very different from the relatively rare malignancies that fit the tumor suppressor role of *NF1*. Even though the benign tumors are unpredictable in timing and location of presentation, the gliomas and plexiform neurofibromas usually present in childhood, whereas cutaneous neurofibromas tend to arise during puberty and continue to present throughout adulthood. Unlike the case of benign colonic adenoma, whose paradigm comprises progression to adenocarcinoma in colon cancer via the accumulation of mutations in multiple genes, there is no definitive evidence demonstrating transformation of these benign NF1-associated tumors into malignancies. The multicellular composition of neurofibromas makes it difficult to identify specific genetic changes in these tumors, and the status of the 'normal' *NF1* allele in all cells within the tumors is difficult to ascertain. In some studies of human neurofibromas, a subpopulation of Schwann cells undergoes double inactivation of NF1, which then induces adjacent NF1 haploinsufficient cells of various lineages to participate in tumor progression. This paracrine action is also demonstrated in transgenic mouse models. Presently, there is no accepted therapy aimed at limiting either the numbers or continued growth of neurofibromas in afflicted individuals.

Animal Models for NF1

The amino acid homology between human, rodent, and chicken neurofibromin predicts that this protein plays a significant role in many species. However, there are no recognizable animal models depicting NF1 features that occur naturally in other species. Apparent neurofibromas are reported in multiple members of one bovine pedigree, and tumors that are similar to neurofibromas have been propagated in damsel fish via a presumed viral infection. In laboratory experiments, hamsters injected during pregnancy with *N*-nitroso-*N*-ethylurea have pups that develop sarcomas associated with large sensory nerves, and ecotropic viral insertions at the mouse *Nf1* locus result in myelogenous leukemias in one strain of mice. These models have yet to provide information applicable to human NF1. A more promising model for the evaluation of NF1, however, is found among transgenic mice.

Identification of the *NF1* gene enabled investigators to 'knock out' neurofibromin production in strains of mice, and both heterozygous and homozygous animals have been propagated. The homozygous mice die *in utero* no later than 14.5 days postconception, and death is tentatively attributed to cardiac dysfunction. Other features include overgrowth of the para- and prevertebral sympathetic ganglia, hypoplastic skeletal muscle, reduced number of renal glomeruli, and hepatocellular hypoplasia. The heterozygous mice do not display an NF1 phenotype; in age-related fashion, however, they have a higher incidence of malignancies, including lymphomas, adrenal tumors, myeloid leukemias, and one case of neurofibrosarcoma. Thus, heterozygous mice provide some insight into tumor progression, whereas homozygous mice yield information on the possible roles of neurofibromin in embryogenesis. Given the possibility that modifying genes likely play a role in the variability of clinical expression of human NF1, the genetic background of the transgenic mice may come to play a significant role in the establishment of the murine NF1 phenotype. Thus, crosses between different animal strains could provide important revelations in the future.

Conclusions

The ongoing research involving the basic biology of NF1 has become possible with the cloning and characterization of the gene. It is only by understanding the pathogenesis of NF1 that rational approaches to meaningful therapeutic intervention can be developed. The identification of proteins that interact with neurofibromin and/or modifying genes will further delineate the pathway involving neurofibromin. Little is known about the regulation of *NF1* gene expression in various cell types and the timing of its expression during embryogenesis. Knowledge of transcriptional factors important in *NF1* expression

could link the signaling pathways involving neurofibromin with other pathways, to provide a clearer understanding of cascade networks within a given cell type.

The identification of more *NF1* mutations in patients who have sound clinical information may reveal genotype–phenotype correlations previously overlooked for lack of disease phenotype. In conjunction with the identification of modifying loci, such information may enable practitioners to provide better counsel regarding the natural history for an individual with a specific genotype. Finally, knowledge of the pathogenesis of NF1 is prerequisite to the development of rational clinical management regimens for those afflicted with this disorder. It is hoped that areas of research touched upon here will provide the foundation for a more directed approach to the development of treatment protocols.

See also: Neurofibromatosis Type 2: Further Questions and Answers; Neurofibromatosis Type 1; Neurofibromatosis Type 2; Tumors of the Brain and the Spinal Cord.

Further Reading

Bernards A (1995) Neurofibromatosis type 1 and Ras-mediated signaling: Filling in the GAPs. *Biochimica et Biophysica Acta* 1242: 43–59.

Friedman JM and Birch PH (1997) Type 1 neurofibromatosis: A descriptive analysis of the disorder in 1728 patients. *American Journal of Medical Genetics* 70: 138–143.

Gutmann DH, Aylsworth A, Carey JC, et al. (1997) The diagnostic evaluation and multidisciplinary management of neurofibromatosis 1 and neurofibromatosis 2. *Journal of the American Medical Association* 278: 51–57.

Gutmann DH and Collins FS (1994) von Recklinghausen neurofibromatosis. In: *The Metabolic and Molecular Basis of Inherited Disease*, 7th edn., pp. 1–19. New York: McGraw-Hill.

Huson SM and Hughes RAC (1994) *The Neurofibromatoses: A Pathogenetic and Clinical Overview.* London: Chapman & Hall.

Messiaen LM, Callens T, Mortier G, et al. (2000) Exhaustive mutation analysis of the *NF1* gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. *Human Mutation* 15: 541–555.

Riccardi VM (1992) *Neurofibromatosis: Phenotype, Natural History, and Pathogenesis*, 2nd edn. Baltimore: Johns Hopkins University Press.

Shaw D (ed.) (1994) *Molecular Genetics of Human Inherited Disease.* London: Wiley.

Viskochil D (2001) Neurofibromatosis type 1. In: Cassidy SB and Allanson JE (eds.) *Management of Genetic Syndromes*, 1st edn., pp. 29–251. New York: Wiley-Liss.

Viskochil D, Cawthon R, and White R (1993) The neurofibromatosis type 1 gene. *Annual Review of Neuroscience* 16: 183–205.

Young H, Hyman S, and North K (2002) Neurofibromatosis 1: Clinical review and exceptions to the rules. *Journal of Child Neurology* 17: 613–621.

Relevant Websites

http://ctf.org – Children's Tumor Foundation.

http://nfmutation.org – Children's Tumor Foundation NF1 International Mutation Database.

http://www.geneclinics.org – GeneTests (a medical genetics information resource funded by the National Institutes of Health).

http://www.nfinc.org – Neurofibromatosis, Inc.

Neurofibromatosis Type 2

S R Plotkin, Massachusetts General Hospital, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurofibromatosis type 2 (NF2, MIM 101000) is an autosomal dominant neurogenetic disorder characterized by the presence of schwannomas, meningiomas, ependymomas, and ocular abnormalities. The first clinical description of NF2 dates to 1822 when the Scottish surgeon JH Wishart described a 21-year-old man with amblyopia and macrocephaly-who became deaf at age 19. At autopsy, multiple tumors arising from the skull base were identified. For many years, NF2 was confounded with the more common syndrome NF1 from which it derives its name. In the 1980s, these two disorders were finally differentiated when tumor studies and linkage analysis localized the genes to different chromosomes. The *NF2* gene was cloned in 1993 by two groups. As expected by Knudson's two-hit hypothesis, patients with NF2 have a germline mutation in the *NF2* gene and loss of the second allele in tumors. Subsequent studies on the NF2 protein have established its role in linking the cell membrane to the cytoskeleton, and progress on the nature of tumorigenesis continues. Transgenic mouse models of NF2 have advanced the understanding of NF2 and hold promise for accelerating the discovery of active drugs to treat NF2.

Epidemiology and Diagnostic Criteria

The current knowledge about the prevalence and incidence of NF2 comes from large population-based studies in the United Kingdom and Finland. In an early study, the birth incidence was estimated to be 1 in 40 000 with a prevalence of 0.86 per 100 000 persons. In a more recent study from the same group, the birth incidence was estimated to be 1 in 25 000 with a prevalence of 1.14 per 100 000 persons.

Clinical criteria for the diagnosis of NF2 were first formulated at the National Institutes of Health (NIH) Consensus Conference on NF1 and NF2 in 1987 and revised in 1991. Under NIH criteria, a diagnosis of NF2 is based on either (1) the presence of bilateral vestibular schwannomas or (2) a family history of NF2 and either a unilateral vestibular schwannoma or any two other tumors typically associated with NF2 (Table 1). Thus, only patients with bilateral vestibular schwannomas or a family history can qualify for a diagnosis under NIH criteria. Patients who lack these criteria but have multiple features associated with NF2 represent a diagnostic dilemma. For this reason, revised criteria were proposed by the Manchester group in 1992 and by the National Neurofibromatosis Foundation in 1997 (Table 1). The relative merits of these criteria continue to be debated by researchers.

Clinical Features

NF2 is characterized by the predisposition to develop multiple tumors, including schwannomas, meningiomas, and spinal cord gliomas. However, nontumor abnormalities such as cataracts and peripheral neuropathy are also common in NF2. The average age of onset of symptoms is between 17 and 21 years and typically precedes a formal diagnosis by 5–8 years. Signs of eighth nerve dysfunction (deafness, tinnitus, or imbalance) are the most common presenting symptoms in adults but occur in only a minority of pediatric patients. In younger patients, presenting signs include cranial nerve dysfunction, peripheral nerve dysfunction, myelopathy, seizures, skin tumors, café-au-lait macules, and juvenile cataracts (Table 2).

Tumor Manifestations of NF2

Schwannomas Vestibular schwannomas are the hallmark of NF2 and invariably develop in patients with the disorder (Figure 1). Schwannomas of nonvestibular cranial nerves and spinal nerves (Figure 2) are common. For example, in one series of 27 patients, 88% of 201 spinal tumors were consistent with schwannomas, with an average of 6.8 tumors per patient. In some NF2 patients, tiny nodular expansions of nerve fibers are present, most commonly in the cauda equina. For many years, it was debated whether these small nodules represented Schwann cell hyperplasia or an initial step in schwannoma formation. Loss of heterozygosity analysis and immunohistochemistry of these nodules confirmed the loss of both *NF2* alleles, suggesting they represented an early phase of tumorigenesis. Schwannomas in patients with NF2 are histologically benign. Rarely, malignant peripheral nerve sheath tumors occur in patients who have received radiation therapy.

Meningiomas About 50% of all patients with NF2 develop intracranial meningiomas (Figure 3) and 50% develop spinal meningiomas. In the previously mentioned series of 27 patients, 12% of 201 spinal tumors were consistent with meningiomas, with an average of 2.6 tumors per patient. Optic sheath meningiomas occur in 4–8% of patients with

Table 1 Clinical criteria for diagnosis of NF2

National Institutes of Health criteria
Bilateral vestibular schwannoma
or
First-degree family relative with NF2 and either unilateral VS or any one of the following: meningioma, schwannoma, glioma, neurofibroma, juvenile posterior subcapsular lens opacity

Manchester criteria
Bilateral vestibular schwannoma
or
First-degree family relative with NF2 and either unilateral VS or any two individual manifestations including meningioma, schwannoma, glioma, neurofibroma, posterior subcapsular lens opacity
or
Unilateral VS and any two individual manifestations including meningioma, schwannoma, glioma, neurofibroma, posterior subcapsular lens opacity
or
Multiple meningiomas ($\geq$2) and either unilateral VS or any two individual manifestations including schwannoma, glioma, neurofibroma, cataract

National Neurofibromatosis Foundation (Children's Tumor Foundation) criteria
Definite NF2
Bilateral vestibular schwannoma
or
First-degree family relative with NF2 and either unilateral VS at <30 years of age or any two of the following: meningioma, schwannoma, glioma, juvenile lens opacity
Presumptive or probable NF2
Unilateral VS <30 years of age and at least one of the following: meningioma, schwannoma, glioma, juvenile lens opacity
or
Multiple meningiomas and either unilateral VS <30 years of age or at least one of the following: schwannoma, glioma, juvenile lens opacity

Table 2 Prevalence of clinical findings in individuals with NF2

	All subjects	*Pediatric subjects*
Mean age at onset (years)	17–21	5.5–6.5
Cranial nerve schwannoma (%)	24–48	33–90
Intracranial meningiomas (%)	45–58	22–83
Spinal tumors (%)	26–90	75–78
Skin tumors (%)	64–68	56–92
>10 skin tumors (%)	10–15	
Optic sheath meningioma (%)	4–8	
Lens opacities (%)	38–81	36–75
Retinal hamartoma (%)	9	0–42
Astrocytoma (%)	2–16	0–24
Ependymoma (%)	3–6	11–25
Peripheral neuropathy (%)	2–6	0–42
Any CAL macules (%)	42–48	0–100
Lisch nodules (%)	3–4	0–22

NF2 and are a disproportionate cause of decreased visual acuity. Compression of the optic nerve can produce optic atrophy. Meningiomas associated with NF2 are almost universally benign histologically.

Spinal cord gliomas Spinal ependymomas and astrocytomas in patients with NF2 present as intramedullary spinal cord lesions and occur in up to 53% of patients. Two-thirds of patients with ependymomas have multiple tumors. The cervicomedullary junction or cervical spine is most commonly involved (63–82%), followed by the thoracic spine (36–44%). The brain and lumbar spine, common sites for sporadic tumors, are rarely involved. Radiographic evidence of tumor progression occurs in less than 10% of patients, and progressive neurologic dysfunction requiring surgical intervention occurs in 12–20% of patients.

Nontumor Manifestations of NF2

Ophthalmic abnormalities Lens opacities occur in 70–90% of patients with NF2. Only a minority of patients experience diminished visual acuity (20/40 or less) or increased sensitivity to glare. Surgical intervention for lens opacities is rarely necessary. Epiretinal membranes have been detected in 80% of patients undergoing comprehensive ophthalmologic testing and in 100% of eyes examined during autopsy. Pigment epithelial or retinal hamartomas are benign growths that are located in the peripapillary, macular, or peripheral retinal areas. These hamartomas are derivatives of glial, vascular, and/or melanocytic tissue within the retina and occur in up to 25% of patients with NF2. In a minority of patients, retinal or pigment epithelial hamartomas can cause decreased visual acuity, amblyopia, or strabismus in the affected eye. Epiretinal membranes and pigment epithelial/retinal hamartomas are associated with inactivation of chromosome 22 and loss of expression of merlin.

Neuromuscular abnormalities Peripheral neuropathy and monomelic atrophy have been reported in patients with NF2. Estimates of the prevalence of peripheral neuropathy vary from 1.3% to 47%, which likely reflects the technique and definitions used in the studies. Regardless, it is likely that this manifestation is underrecognized. MRI scans revealed no evidence of tumors that could explain the clinical or electrophyiological findings.

Focal amyotrophy has been documented in patients with NF2. Although the prevalence of this finding has not been firmly established, wasting of a single limb was noted in 4% of patients in a registry of NF2 patients in the United Kingdom. Detailed examination of four different patients by MRI and electrodiagnostic studies excluded compression of proximal nerves or roots by tumors as a cause. The etiology of focal amyotrophy in NF2 is not known but is thought to arise from neurofibromatous changes in peripheral nerves.

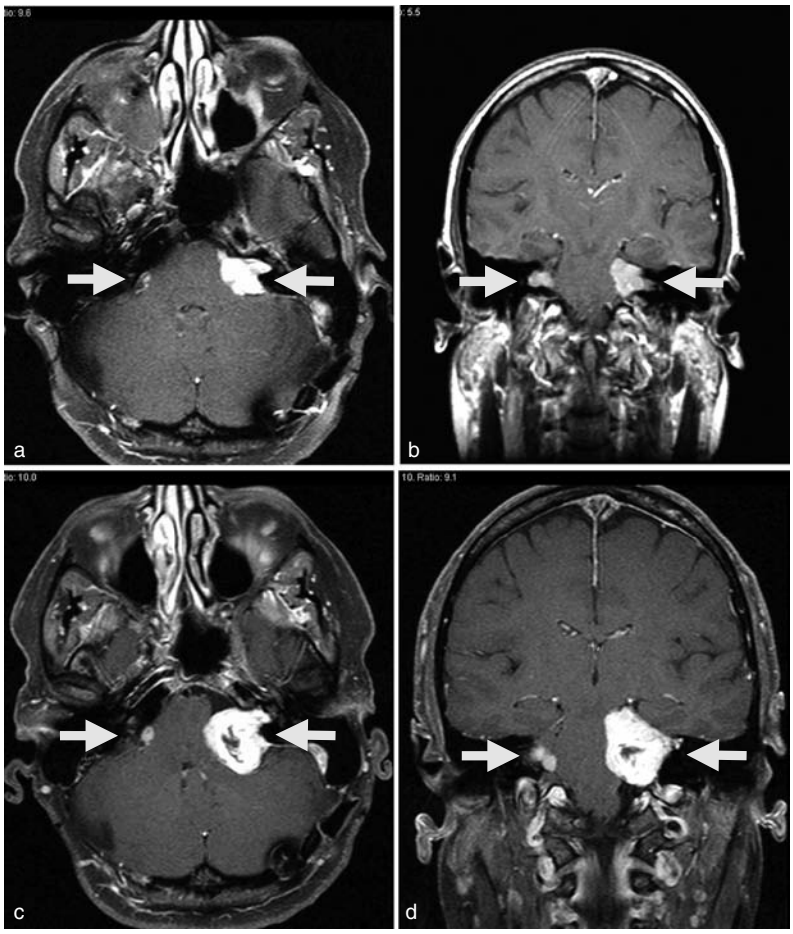

Figure 1 Contrast-enhanced MRI scans from a patient with NF2. The left vestibular schwannomas (arrows) noted in 2002 (a, axial T1 post-contrast image; b,coronal T1 post-contrast image) have enlarged in size by 2004 (c, axial T1 post-contrast image; d, coronal T1 post-contrast image).

Delineation of NF2 from NF1

Historically, patients with skin and spinal cord tumors were diagnosed with neurofibromatosis. The term 'central' neurofibromatosis was introduced to distinguish patients with tumors involving the eighth cranial nerves but not the skin from those with the more common 'peripheral' neurofibromatosis described by von Recklinhausen. Unfortunately, the distinction between central and peripheral neurofibromatosis was confounded by Harvey Cushing in 1917. Subsequently, NF2 was seen as a manifestation of von Recklinghausen's disease rather than as an independent entity. The disorders were conclusively separated in the 1980s when the genes were localized to different chromosomes.

In the clinic, patients with multiple nerve sheath tumors are recognized to have neurofibromatosis but the distinction between NF1 and NF2 is not always emphasized. A detailed history and physical examination (including the skin and eyes) can reliably distinguish between the two disorders in almost all patients. Occasionally, MRI scan of the brain (including the internal auditory canals) is required for diagnosis. Rarely, biopsy of a tumor is required. The main similarities and differences between NF1 and NF2 are shown in Table 3.

Molecular Genetics of NF2

Linkage Studies and Positional Cloning of *the NF2* Gene

The first clues to the location of the *NF2* gene came from tumor studies. Cytogenetic analysis of sporadic and NF2-related meningiomas revealed frequent deletions of chromosome 22. Subsequently, Seizinger et al. scanned sporadic and NF2-related vestibular schwannomas for chromosomal deletions using polymorphic DNA markers. They confirmed loss of heterozygosity for D22S1, suggesting that loss of genes on chromosome 22 was important for vestibular schwannoma formation. Subsequent linkage analysis of D22S1 in a large NF2 kindred established that the defect causing NF2 was located in the immediate vicinity of D22S1

on chromosome 22. Taken together, these tumor and linkage studies suggested that the *NF2* locus encoded for a tumor suppressor gene whose inactivation leads to tumor formation.

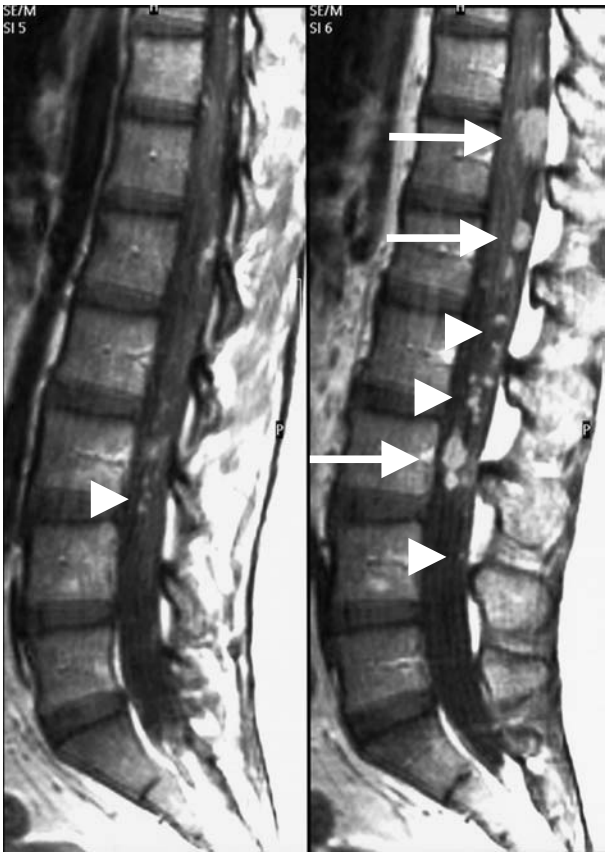

Figure 2 Sagittal T1 post-contrast MRI scan of the lumbar spine in a patient with NF2. Note innumerable tiny contrast-enhancing lesions (arrowheads) consistent with schwannoma tumorlets and larger contrast-enhancing lesions consistent with schwannomas (arrows).

The region of interest was narrowed to the interval between D22S212 and D22S32 using segregation studies in affected families and tumor deletion mapping. In 1993, two groups independently published the identity of the *NF2* gene. One group cloned the gene after delineating nonoverlapping germline deletions in two unrelated NF2 families. This group named the encoded protein 'merlin' (moezin, ezrin, radixin-like protein) to emphasize its relationship to various cytoskeletal proteins. The other group screened the region of interest for chromosome rearrangements by hybridization of single copy probes to DNA. This technique identified two probes that detected abnormal fragment sizes on two unrelated patients and the group named the protein 'schwannomin' to emphasize its role in suppression of schwannoma formation.

Table 3 Comparison of clinical features of NF1 and NF2[a]

	NF1	*NF2*
Mean age at diagnosis	**First decade**	**Third decade**
Vestibular schwannomas	**No**	**Yes**
Hearing loss	No	Yes
Nerve sheath tumors	Neurofibromas	Schwannomas
Glial tumors	Optic pathway gliomas	Spinal ependymomas
Meningiomas	No	Yes
Opthalmologic examination	**Lisch nodules**	**Cataracts, epiretinal membranes**
Skin tumors	**Many**	**Few**
Skin fold freckling	**Yes**	**No**
≥6 CAL macules	**Yes**	**No**

[a]Features that help distinguish between the two diseases are shown in bold.

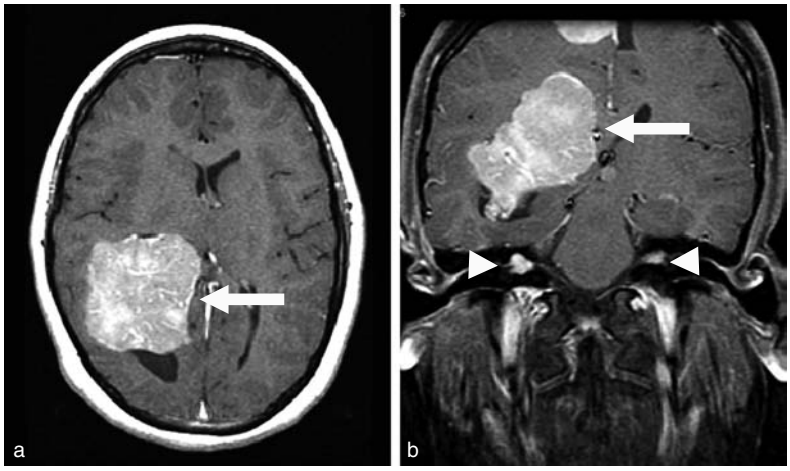

Figure 3 Axial (a) and coronal (b) T1 post-contrast images from an cranial MRI of a patient with NF2. Note ventricular-enhancing lesions consistent with meningioma (arrow) and bilateral vertibular schwannomas (arrowheads).

NF2 Gene and Its Protein Product

The *NF2* gene is composed of 17 exons spanning 110 kb. The NF2 gene is highly conserved between species, including mouse, rat, pig, chicken, hamster, cow, turkey, and monkey. There are three alternative messenger RNA (mRNA) species (7, 4.4, and 2.6 kb) in humans due to variable length of the 3′ untranslated region. In mice, there is a single mRNA species of 4.5 kb. The predominant *NF2* gene product is a 595-amino-acid protein and is a member of the 4.1 family of cytoskeletal proteins that includes ezrin, radixin, and moesin (ERM proteins). The NF2 protein shares approximately 45–50% amino acid identity with these proteins and lacks an extended hydrophobic stretch associated with transmembrane proteins. Merlin is composed of an N-terminal four point one ezrin radixin moesin (FERM) domain that has 59–62% homology with moesin and radixin, an α-helix domain, and a C-terminal domain that has 40–42% homology for moesin and radixin. The moesin FERM domain is composed of three lobes that bind the C-terminal tail, thereby masking a large portion of the FERM domain surface. This intramolecular association plays a key role in regulation of ERM proteins. Although the C-terminal tail of merlin has important differences from the ERM proteins, the overall conservation of critical residues suggests that merlin is also regulated by FERM–tail domain interactions.

Tissue specificity Studies of mRNA using Northern blot and reverse-transcriptase polymerase chain reaction have revealed ubiquitous expression in human and mouse tissues. Similarly, immunoblotting has confirmed the widespread expression of merlin in all tissues. Attempts to localize mRNA and protein expression to specific cells within tissues using *in situ* hybridization and immunohistochemistry have produced conflicting results. mRNA expression is reliably seen within neurons of the brain and spinal cord and within epithelial cells of the developing lens. Protein expression is noted in neurons (particularly Purkinje cells of cerebellum and motor neurons of spinal cord), Schwann cells, and blood vessels. The inability to demonstrate mRNA and protein expression using *in situ* hybridization and immunohistochemistry in some tissues is likely due to the relatively low sensitivity of the techniques used in these studies.

Regulation and function of merlin Despite more than a decade of intense scientific investigation, the function of the merlin remains elusive. In part, this has been attributed to the technical difficulty in isolating and localizing the protein. Localization and functional studies suggest that merlin links membrane-associated proteins and the actin cytoskeleton, thereby acting as an interface with the extracellular environment. Merlin is regulated, in part, through phosphorylation of serine and threonine residues, which accounts for the two major isoforms seen by Western blotting. The best studied stimulus for merlin phosphorylation is activation of Rac1, which is mediated by p21-activated kinase and by protein kinase A. In addition, merlin has been shown to interact with many different proteins, such as magicin, N-WASP, and betaII-spectrin, but functional studies to confirm these findings have been lacking. Understanding the molecular basis of merlin's function remains one of the most pressing issues faced by the research community.

Merlin represents a novel class of tumor suppressor. Unlike p53 or retinoblastoma protein, merlin does not act directly on the cell cycle machinery but instead mediates and interprets signals from the extracellular milieu. The tumor suppressor function of merlin is manifested by its ability to maintain contact-dependent inhibition of proliferation. The mechanistic basis for contact inhibition is poorly understood and is the subject of intense investigation. Evidence suggests that merlin is involved in trafficking of growth factor receptors on the cell surface such as the epidermal growth factor receptor. This provides a plausible mechanism by which merlin could regulate cellular proliferation.

Model organisms Germline inactivation of both *NF2* alleles is embryonically lethal due to the absence of extraembryonic structures and failure of gastrulation. Germline inactivation of one *NF2* allele leads to a high incidence of malignant tumors such as osteosarcoma, lymphoma, lung adenocarcinoma, hepatocellular carcinoma, and fibrosarcoma, suggesting a broad role in tumorigenesis. Unfortunately, these transgenic mice do not develop tumors seen in human patients with NF2, which limits the use of these animals in clinically related studies. However, mice with targeted inactivation of *NF2* in Schwann cells do develop multiple Schwann cell tumors, and those with targeted inactivation in arachnoid cells develop tumors that resemble meningioma.

Tumor studies Merlin acts as a true tumor suppressor since loss of both copies of the gene results in tumor formation. In NF2, loss of the first copy occurs in the egg or sperm and usually involves a coding sequence mutation or deletion. Loss of the second copy occurs as a somatic event, often as loss of part or all of the *trans* chromosome. In sporadic tumors,

losses of both copies occur as somatic events. Molecularly, somatic events can be visualized as loss of heterozygosity of nearby polymorphic markers. Inactivation of the *NF2* gene can be detected in the vast majority of sporadic schwannoma, in 50–60% of sporadic meningiomas, and in a minority of sporadic ependymomas. *NF2* gene mutations have been found in unrelated tumors such as melanoma, malignant mesothelioma, thyroid carcinoma, and perineural cell tumors, suggesting a more general role for the protein in tumor suppression for certain cancers.

Mutational Analysis of the *NF2* Gene

All families with affected members in more than one generation show linkage to the *NF2* gene and there is no evidence for locus heterogeneity. Using exon scanning and single-strand conformation polymorphism, detection rates for germline mutation in patients with bilateral vestibular schwannomas vary between 40% and 66%. The presence of large deletions, mutations in promoter or intronic regions, and somatic mosaicism may explain the difficulty in identifying a causative mutation in all patients. Use of multiplex ligation-dependent probe amplification assays has improved the ability to detect deletions of one to a few exons. The vast majority of alternations are 'private' – unique to a family or individual. Nonsense and frameshift mutations are most common, followed by splice site and nontruncating mutations. CpG dinucleotides within CGA codons (in exons 2, 6, 8, and 11) represent mutation hot spots since C>T transitions result in a stop codon TGA through deamination of 5-methylcytosine. The majority of these mutations are predicted to truncate (nonsense or frameshift mutations) or shorten (splicing mutation) the gene product. Loss of function due to foreshortening of the C-terminus or instability of the protein product leads the tumor suppressor phenotype.

Mosaicism Mutations in the *NF2* gene have been detected in up to 70% of affected subjects. The detection rate for mutations is 20–30% lower for founders (i.e., patients without a family history of NF2). In part, the low rate of detection for mutations in sporadic cases is due to somatic mosaicism. Somatic mosaicism results from a mutation that occurs during postzygotic embryonic development. In these patients, only a subpopulation of cells carry the constitutional mutation and analysis of blood leukocytes is often unremarkable.

Indirect evidence suggests that mosaicism exerts a mitigating effect on the NF2 phenotype such that patients may experience a later time of onset and milder course of disease. Historically, a precise estimate of the rate of mosaicism in these patients has been difficult to establish. The most definitive numbers come from a study of 233 NF2 founders with bilateral vestibular schwannomas. In this study, the authors estimated the rate of mosaicism to be 16.7–24.8%. Recurrence risk to offspring of mosaic patients is significantly lower than the 50% expected rate for an autosomal dominant disorder and is probably on the order of 5–15%.

Genotype/phenotype correlations Genotype/phenotype correlations have been published by multiple groups. All studies are limited by the small number of patients with mutations identified. In a German study using number of tumors identified by MRI scan as a measure of disease severity, frameshift and nonsense mutations were associated with severe disease, whereas missense mutations were associated with mild disease. Splice site mutations were associated with both mild and severe phenotype, suggesting the possible importance of additional determinants such as stochastic, epigenetic, or environmental factors. In a US study, frameshift or nonsense mutations were associated with a younger age at onset and diagnosis and with a larger mean number of tumors than were splice site mutations. Interestingly, an association between nonsense mutations and retinal hamartomas and/or epiretinal membranes has been found. Finally, in a UK series, truncating mutations, compared with other mutations, were associated with an early age at onset and diagnosis, with symptomatic central nervous system tumors besides vestibular schwannomas, and with spinal tumors. Point mutations are rare in patients with NF2 and have been associated with mild, moderate, and severe disease. Genotype/phenotype correlations suggest that patients with mosaic disease and those with large deletions have a significantly reduced risk of developing cataracts compared to patients with classic NF2.

Detailed clinical studies of patients with NF2 suggest that phenotypic variability is smaller within families than between families. This is in contrast to patients with NF1, where there is significant variability within families. Mathematical modeling of intrafamilial correlation supports these clinical impressions. In a study of 390 patients from 153 families in the UK NF2 registry, age at onset of diagnosis, age at onset of hearing loss, and number of intracranial meningiomas were correlated within families. A significant intrafamiliar correlation was noted for age at onset (correlation coefficient, 0.35), age at onset of hearing loss (correlation coefficient, 0.51), and number of meningiomas (correlation coefficient, 0.29). This study supports the notion of familial homogeneity but also demonstrates the importance of other uncharacterized factors.

Future Directions

Our understanding of NF2 has significantly improved since the gene was cloned in 1993. The near future will witness an intense collaboration among basic scientists and clinicians. As scientists examine the role of merlin in contact-dependent inhibition of growth and generate animal models that recapitulate the human disease, clinicians will begin to mount clinical trials for NF2-related tumors. Advances in mutational analysis will lead to better family planning and more accurate presymptomatic diagnosis of family members. Finally, patients with NF2 will play an increasingly important role in expanding research support for NF2.

See also: Neurofibromatosis Type 1: Molecular and Cellular Biology; Neurofibromatosis Type 2: Further Questions and Answers; Neurofibromatosis Type 1; Tumors of the Brain and the Spinal Cord.

Further Reading

Baser ME (2006) The distribution of constitutional and somatic mutations in the neurofibromatosis 2 gene. *Human Mutation* 27(4): 297–306.

Evans DG, Baser ME, O'Reilly B, et al. (2005) Management of the patient and family with neurofibromatosis 2: A consensus conference statement. *British Journal of Neurosurgery* 19(1): 5–12.

Gutmann DH, Aylsworth A, Carey JC, et al. (1997) The diagnostic evaluation and multidisciplinary management of neurofibromatosis 1 and neurofibromatosis 2. *JAMA* 278(1): 51–57.

Gutmann DH and Giovannini M (2002) Mouse models of neurofibromatosis 1 and 2. *Neoplasia* 4(4): 279–290.

Kluwe L, Mautner V, Heinrich B, et al. (2003) Molecular study of frequency of mosaicism in neurofibromatosis 2 patients with bilateral vestibular schwannomas. *Journal of Medical Genetics* 40: 109–114.

McClatchey AI and Giovannini M (2005) Membrane organization and tumorigenesis – the NF2 tumor suppressor, merlin. *Genes & Development* 19(19): 2265–2277.

Parry DM, Eldridge R, Kaiser-Kupfer MI, et al. (1994) Neurofibromatosis 2 (NF2): Clinical characteristics of 63 affected individuals and clinical evidence for heterogeneity. *American Journal of Medical Genetics* 52: 450–461.

Roleau GA, Merel P, Lutchman M, et al. (1993) Alteration in a new gene encoding a putative membrane-organizing protein causes neuro-fibromatosis type 2. *Nature* 363(6429): 515–521.

Ruggieri M, Iannetti P, Polizzi A, et al. (2005) Earliest clinical manifestations and natural history of neurofibromatosis type 2 (NF2) in childhood: A study of 24 patients. *Neuropediatrics* 36(1): 21–34.

Trofatter JA, MacCollin MM, Rutter JL, et al. (1993) A novel moesin-, ezrin-, radisin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. *Cell* 75(4): 826.

Relevant Websites

http://neurosurgery.mgh.harvard.edu – Neurosurgical Service. Graphical representation of mutations in the *NF2* gene.

http://www.interscience.wiley.com – Wiley Inter Science. Listing of constitutional mutations in NF2.

http://www.ctf.org – Children's Tumor Foundation.

Neurofibromatosis Type 2: Further Questions and Answers

M H Ruttledge, National Hospital for Neurology and Neurosurgery, London, UK
G A Rouleau, Notre Dame Hospital, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The neurofibromatosis type 2 (NF2) tumor suppressor gene, *NF2*, was identified in 1993, and was variably named *Schwannomin* or *Merlin* by the two groups involved. The name *Schwannomin* was proposed because *NF2* is inactivated in schwannomas. *Merlin* was suggested as a name (and Merlin, for the protein) because of the homology of *NF2* to the ERM family of proteins (moesin, ezrin, and radixin). This dual nomenclature, both in the literature and at conferences, creates confusion and it would be preferable to use a single designation. Therefore, in an attempt to simplify matters, but also to recognize the valuable contributions of both research groups who first identified the gene, we suggest that the *NF2* gene/protein be called *Schmerlin*/Schmerlin.

Germ line mutation of one copy of *Schmerlin*, with somatic mutation of the remaining allele, results in the inherited syndrome, NF2. Patients frequently develop multiple tumors, with the severity determined by several factors, including the type of mutation in *Schmerlin* (see the section titled 'Genotype–phenotype correlation in NF2'). Early reports describing NF2 are difficult to find in the literature, as the condition was historically grouped together with neurofibromatosis type 1 (NF1; also termed von Recklinghausen's disease, or peripheral neurofibromatosis). However, it became clear during the late twentieth century that NF2 was a distinct clinical entity. The original descriptions of NF2 referred to the disease as central neurofibromatosis or bilateral acoustic neurofibromatosis, but over the past two decades the term 'NF2' has become almost universally accepted. Clinically, the hallmark of NF2 is the development of bilateral vestibular schwannoma (BVS) together with solitary or multiple meningiomas (and occasionally ependymomas). However, bilateral eighth-nerve tumors are not compulsory for a diagnosis of NF2. For example, a solitary vestibular schwannoma combined with a positive family history for NF2, or a solitary meningioma, is sufficient (see the section titled 'Epidemiology, clinical features, and diagnostic criteria').

Biallelic somatic mutations in *Schmerlin* predispose to the corresponding sporadic nervous system tumors commonly found in NF2, including schwannomas and meningiomas (and less commonly ependymomas). Somatic *Schmerlin* mutations are also found in malignant mesothelioma, but these tumors are not typically associated with NF2. *Schmerlin* is a tumor suppressor gene (TSG) and thus follows the 'two-hit' model of tumorigenesis. This model was first proposed by Knudson in his landmark paper on retinoblastoma in 1971. The 'first hit' is usually a protein-inactivating point mutation in a single somatic cell (point mutations comprise single nucleotide substitutions, leading to missense, nonsense, and splice-site alterations, and small insertions/deletions). A 'second hit' to knock out the remaining copy of the gene's locus is then required in order to give this cell a growth advantage, which eventually leads to tumorigenesis. This second hit is often monosomy 22 (loss of one entire copy of chromosome 22) or deletion of a large proportion of the chromosome. Monosomy 22 is usually the result of nondisjunction during mitosis in the cells containing the original point mutation. The term 'loss of heterozygosity' (LOH) describes the experimental findings in tumors that are missing one copy of a particular chromosomal region. Numerous groups have performed LOH studies in sporadic NF2-associated tumors (including meningiomas and schwannomas) and have shown a high frequency (usually 50–80%) of loss of one copy of chromosome 22. These high frequencies of LOH would suggest that inactivation of *Schmerlin* is a prerequisite for the development of these tumors.

The biological role of *Schmerlin* is not completely understood. Reports published soon after the identification of the gene demonstrated that it was located intracellularly at the plasma membrane and colocalized with elements of the actin cytoskeleton. It was believed that Schmerlin, like the other members of the ERM family, acts to stabilize the plasma membrane cytoskeleton junction. It is difficult to rationalize how this role alone results in inhibition of cell growth and proliferation. Subsequently, Schmerlin was found to interact with numerous intracellular components – for example, p21-activated kinase (Pak1) and phosphatidylinositol 3-kinase enhancer (PIKE) – many of which could conceivably be involved in tumor suppression. To date, a unifying tumor suppressor mechanism for the NF2 protein has not been conclusively demonstrated. However, several distinct lines of investigation are now under way and these are discussed in detail in the following section.

Epidemiology, Clinical Features, and Diagnostic Criteria

Mutations in *Schmerlin* occur frequently in human disease. It is helpful to divide clinical cases which

involve NF2 mutations into the two principal modes of presentation. Patients can either present with a sporadic NF2-associated tumor (schwannoma, meningioma, or ependymoma) which occurs as a result of biallelic somatic gene inactivation or, alternatively, patients have an inherited mutation and thus develop the full clinical syndrome of NF2. The prognosis for patients in the first group (a solitary meningioma or schwannoma) is usually very good, as often these tumors grow slowly and rarely display aggressive or malignant features. These patients do not have an inherited mutation and are thus very unlikely to develop further tumors. As already discussed, these sporadic tumors develop as the result of inactivation of both copies of *Schmerlin* in a single somatic cell, the first event usually being a point mutation and the second being loss of the remaining wild-type allele by deletion of the entire gene (complete or partial loss of chromosome 22, demonstrated by LOH studies). The prognosis for patients with NF2 is usually considerably worse than for patients with sporadic NF2-associated tumors. However, there are some rare exceptions to this rule and the prognosis for NF2 patients may be very variable. Some patients with NF2 have mild disease and only develop BVS at a late age and often lead relatively normal lives. The opposite end of the NF2 spectrum is severe disease in which multiple intracranial and intraspinal tumors cause symptoms at an early age, resulting in significant disability and death in the third or fourth decade. There are a number of factors which determine the clinical severity of NF2. These include the severity in other family members (see the section titled 'Genotype–phenotype correlation in NF2'), the location of tumors within the central nervous system (CNS; tumors in close proximity to sensitive structures will carry a poorer prognosis), the histological classification and growth rate of each tumor.

Constitutional (inherited) mutations of *Schmerlin* occur with a birth incidence of approximately one in every 25 000 to 40 000 individuals. In England, Wales, and the Republic of Ireland (with a combined population of more than 57 million people), there are approximately 700 000 births per year. Thus, between 18 and 28 people are born with NF2 in these countries each year. At first glance, it would therefore appear that *Schmerlin* mutations are not very common. However, sporadic mutations causing isolated solitary tumors are much more frequent. Sporadic meningiomas and schwannomas together constitute the second largest group of primary CNS neoplasms. The classical schwannoma described in NF2 is located on the vestibular branch of the eighth cranial nerve; such tumors are termed vestibular schwannomas (previously known as acoustic neuromas). Vestibular schwannomas occur in greater than 90% of NF2 patients, but NF2 patients account for only a small proportion (approximately 5%) of all vestibular schwannomas found in a given population. The incidence of sporadic unilateral vestibular schwannomas (UVSs) is approximately 1.4 per 100 000 individuals per year, and it is estimated that 1 in every 1000 people will be diagnosed with a vestibular schwannoma in their lifetime. Meningiomas (often multiple) are also found in patients with NF2 in frequencies greater than 50%. Again, their sporadic counterparts occur much more frequently in the general population. Sporadic meningiomas comprise 20–25% of all primary intracranial neoplasms in the United States, with an annual diagnostic incidence of 8 per 100 000. Moreover, asymptomatic meningiomas are believed to be 10 times more common, the majority of which are identified incidentally at autopsy.

The diagnostic criteria for NF2 are well established and have been previously documented. There have, however, been some slight modifications of these criteria in recent years. In most large series published to date, approximately 85–90% of patients develop BVS. A further 5% of NF2 cases develop a UVS. Thus, up to 95% of patients develop at least one vestibular schwannoma. Schwannomas at other sites (including other cranial nerves, within the spinal cord, and on peripheral nerves) are also very common, and are found in more than half of all patients in most studies. Approximately 60–70% of NF2 cases develop at least one meningioma (intracranial and spinal). More than half of all NF2 patients also develop posterior subcapsular lens opacities (PSLO). Ependymomas are seen in less than 5% of all NF2 patients. Less than 2% of NF2 patients have six or more café-au-lait macules, one of the diagnostic criteria for neurofibromatosis type 1 and sometimes a source of confusion between the two entities. The revised National Institutes of Health (NIH) diagnostic criteria for NF2 (published in 1991) are as follows: a patient has NF2 if there is evidence of BVS or if a patient has a first-degree relative with NF2 and evidence of a UVS or any one of meningioma, schwannoma, glioma, neurofibroma, or PSLO. The Manchester criteria involve further subdivisions, in an attempt to pick up more cases. A patient has NF2 if they have (1) BVS, (2) a first-degree relative with NF2 and UVS or any two of meningioma, schwannoma, glioma, neurofibroma, or PSLO, (3) UVS and any two of meningioma, schwannoma, glioma, neurofibroma, or PSLO, or (4) multiple meningiomas (two or more) and UVS or any two of schwannoma, glioma, neurofibroma, or PSLO. A recent report suggests that neither set of diagnostic criteria is sufficient and still further refined criteria are needed.

Identification of the NF2 Gene

The NF2 gene was originally localized to chromosome 22 employing cytogenetic studies in sporadic NF2-associated tumors (predominantly meningiomas). Confirmation of this localization came in 1987 from genetic linkage studies in NF2 families. A number of studies further refined the localization to 22q12.2, and concluded that NF2 appeared to be a genetically homogeneous disease. *Schmerlin* was cloned simultaneously by two groups in 1993 and found to be mutated in NF2 patients and in the majority of sporadic meningiomas and schwannomas. It comprises 17 exons and codes for a protein of 595 amino acids. The protein is closely related to the ERM proteins ezrin, radixin, and moesin. There are no hotspots for mutation in *Schmerlin* in NF2 patients, in sporadic meningiomas, or in schwannomas. Furthermore, all possible types of mutations (missense, nonsense, frameshift, splice-site, and various deletions/insertions) are found. Interestingly, mutations have not been found in exon 16 or 17, suggesting that these regions are not necessary for *Schmerlin* to perform as a tumor suppressor.

There is now overwhelming evidence that *Schmerlin* behaves as a TSG. Furthermore, it is becoming clear that it is the principal TSG for sporadic meningiomas and schwannomas. This is analogous to the retinoblastoma TSG for retinoblastoma and many other cancer predisposition genes. This is not to say that *Schmerlin* is the only gene required for meningioma and schwannoma development. Mutations in other, as yet, unknown genes are necessary before these tumors develop. One such candidate is the tumor protein p53 suppressor gene (*TP53*), which has recently been shown to act synergistically with the *Schmerlin* gene (in conditional NF2 mutant mice) in the development of malignant tumors of neural crest origin. Furthermore, mutation/loss of *Schmerlin* is probably an early event in tumor development, with aberrations of genes on other chromosomes resulting in a more aggressive phenotype. One such candidate gene is *DAL-1/4.1B* (approved symbol, *EPB41L3*), which is located on chromosome 18. A recent study of sporadic meningiomas has shown LOH at the *DAL-1/4.1B* locus in approximately one-fifth of the tumors studied (practically all of which also had LOH at the *Schmerlin* locus on chromosome 22). These findings would support a secondary progressive role of *DAL-1/4.1B* in meningioma tumorigenesis.

Evidence to support a TSG role for *Schmerlin* comes from various quarters. First, inactivating mutations in *Schmerlin* are found in the majority of NF2 patients and in the sporadic tumors associated with NF2 (including meningiomas, schwannomas, and ependymomas). Second, mice with targeted inactivation of *Schmerlin* specifically in Schwann cells develop schwannomas (the commonest tumors found in NF2) and Schwann cell hyperplasia. Third, mice heterozygous for an NF2 mutation develop numerous malignant tumors, including osteosarcomas at high frequency, and fibrosarcoma and hepatocellular carcinoma at a lower frequency. There is also a growing body of evidence from cell biology experiments that *Schmerlin* functions normally as a negative regulator of cell growth and proliferation (see the section titled 'Biological role of the NF2 protein'). Loss of function experiments, in a number of different biological systems, appear to confirm these results.

Genotype–Phenotype Correlation in NF2

Initial descriptions of NF2 showed that the severity of disease was relatively preserved within individual families. Affected members from the same family generally tended to develop symptoms of the disease around the same age, and had a similar rate of progression with equivalent disability and similar numbers of tumors. Therefore, it became clear clinically at a very early stage that a strong genotype–phenotype correlation existed for NF2. This was in stark contrast to neurofibromatosis type 1, in which there is often marked variation with regard to severity of disease within a given family. It is only since the identification of *Schmerlin* that it has been possible to provide the molecular evidence for a clear correlation. NF2 has been historically divided into severe (Wishart) and mild (Gardner) disease. However, it is now apparent that a more refined system is needed. A milder NF2 phenotype (with a later age of onset and fewer tumors) is associated with missense mutations, certain splice-site mutations, and certain large deletions. In addition, mosaics for any particular type of mutation tend to display a milder phenotype. One could further hypothesize that mosaics harboring missense and splice-site mutations should have an even milder phenotype, although this would require a very large series of patients to confirm. Severe disease is usually caused by protein-truncating alterations (such as frameshifts and nonsense mutations). A number of studies have shown that individuals with splice-site mutations in exons 1–5 of *Schmerlin* had more severe disease than did those with splice-site mutations in exons 11–15. Phenotype–genotype correlations have now been documented for overall disease severity, for vestibular schwannomas, for cataracts, and for noneighth cranial nerve tumors (including intracranial meningiomas, spinal tumors, and peripheral nerve tumors). However, there are exceptions to this rule and even in monozygotic twins the clinical course may not be identical.

Mosaicism and Schwannomatosis

Mosaicism for NF2 mutations is a relatively common occurrence. These patients often appear to have a relatively mild phenotype or may just present with multiple schwannomas or meningiomas, either diffusely or in a specific anatomical distribution (single limb, defined segment of the spinal cord, etc.). The common understanding of mosaicism is that there is a mutation in a proportion of cells or tissue types in a particular individual. The mutation is not present in either gamete from the individual's parents, but it occurred during early development of the embryo. The number and location of tumors are dependant on precisely when the mutation occurred. Patients may have very mild disease, whereby a small proportion of Schwann cells or meningeal cells are harboring a single copy of mutated *Schmerlin*. Alternatively, patients may present with clinically typical NF2, but fail to show a constitutional *Schmerlin* mutation in circulating peripheral white blood cells, and also do not have a family history of the disease.

Schwannomatosis is the term used to describe the disease in patients who develop multiple schwannomas in the absence of BVS, the hallmark of NF2. There is an ongoing debate as to whether schwannomatosis is a form of NF2 mosaicism or whether it is a distinct genetic entity. There have been studies to suggest that secondary loci on chromosome 22 are involved, and that both *Schmerlin* and this secondary gene are necessary for the development of this separate disease. The data to support this argument are, however, not convincing, and we would suggest that schwannomatosis is an unusual presentation of mosaicism for NF2.

Biological Role of the NF2 Protein

The exact role of Schmerlin in tumorigenesis is not clear. However, there are some very interesting clues as to the possible mechanisms whereby loss of NF2 function may promote uncontrolled growth and proliferation. Various cytoskeletal and signaling proteins which interact with and bind to the NF2 protein have been identified and are being closely characterized. The principal candidates are described in the following discussion.

The Pak1 protein has been shown to be inhibited by Schmerlin and this has been suggested as a mechanism for tumor suppression. Pak1 belongs to a subgroup of serine/threonine kinases (the group 1 Paks) that are known to be downstream effectors of Rac/cdc42. Pak1 function is required for activation of stress-activated protein kinases by cdc42 and Rac1, as well as for transformation by activated forms of these GTPases. Furthermore, activated forms of Pak1 induce rapid formation of membrane ruffles and focal complexes. Pak1 is essential for both Ras transformation and neurofibromatosis type 1. The proposed mechanism by which Pak1 may be involved in tumorigenesis is as follows: Schmerlin normally functions to inhibit Pak1 and therefore downregulates Rac/cdc42. This in turn would result in inhibition of Ras-induced malignant transformation. Cell culture experiments have demonstrated that mutated Schmerlin, which lacks the Pak1-inhibiting domain, fails to suppress Ras transformation. Thus, Paks are probably bifunctional proteins, affecting both gene transcription (through a kinase cascade) and actin dynamics (via an unknown mechanism). Recently, it has been demonstrated that the NF2 protein is normally involved in arresting cell growth at the G1 phase, with decreased expression of cyclin D1, inhibition of cyclin-dependent kinase 4 (CDK4) activity, and dephosphorylation of the retinoblastoma protein, pRB.

A second candidate pathway for Schmerlin-mediated tumor suppression is via the related FERM (four-point-one, ezrin, radixin, moesin) domain protein, named 'Expanded.' In *Drosophila*, Schmerlin and Expanded are components of the Hippo signaling pathway (a system which is believed to play a role in tumor suppression). In this system, both Schmerlin and Expanded are necessary for proliferation arrest and apoptosis in developing imaginal discs. Thus, this new signaling pathway appears to restrict cell proliferation in differentiating epithelia. The Schmerlin/Expanded proteins control the activation of the Hippo/Salvador kinase complex, which in turn activates the Warts/Mats kinase complex. Warts/Mats kinase phosphorylates and inhibits Yorkie, a transcriptional activator that positively regulates cell growth and proliferation. A further associated possible mechanism of Schmerlin/Expanded tumor suppression, again in *Drosophila*, involves regulating the abundance, localization, and turnover of certain cell surface receptors. One report suggests that misregulation of these processes may be a key component in tumorigenesis. Further studies will be necessary to prove these hypotheses.

There have been a number of reports describing an interaction between Schmerlin and PIKE. PIKE is a brain-specific GTPase that binds to phosphatidylinositol 3-kinase (PI3K) and stimulates its lipid kinase activity. PIKE exists as three isoforms: PIKE-S (short form), PIKE-L (long form), and PIKE-A. PIKE-S and PIKE-L result from differential splicing whereas PIKE-A results from the use of a second transcription initiation site. Schmerlin binds to PIKE-L and abolishes its stimulatory effect on PI3K. PI3K is a lipid kinase that generates phosphatidylinositol 3,4,5-trisphosphate (PI(3,4,5)P3), a second messenger that is

essential for the translocation of Akt to the plasma membrane, where it is phosphorylated and activated. The PI3K/Akt pathway is fundamentally involved in several key cellular functions and plays a central role in cell proliferation. There are numerous examples of alterations in the PI3K/Akt pathway in a variety of human cancers. The implications of Schmerlin interaction with this pathway for tumor suppressor function are thus quite significant.

The interaction of the NF2 protein with hepatocyte growth factor-regulated tyrosine kinase substrate (HRS) is also intriguing. The NF2 protein binds (via its C-terminal end) to HRS. It has been suggested that this interaction may facilitate its ability to function as a tumor suppressor. Experiments in schwannoma cell lines have shown that both the NF2 protein and HRS inhibit activation of signal transducers and activators of transcription (STATs). Moreover, rat schwannoma cell lines experiments have demonstrated that HRS can reduce the amount of total and active epidermal growth factor receptor (EGFR). Therefore, a potential direct consequence of Schmerlin inactivation may be the downregulation of EGFR.

Numerous other cellular proteins/systems have been shown to interact/influence Schmerlin. These include CD44, BII-spectrin, SCHIP-1, HE110, NHE-RF, eukaryotic initiation factor 3 (eIF3c), and BI-integrin. Moreover, Schmerlin co-localizes and interacts with adherens junction components in cell culture experiments. Knocking out Schmerlin in this environment results in an inability to undergo contact-dependent growth arrest, and to form stable cadherin-containing adhesion junctions. Again, this may be an important mechanism whereby tumor suppressor function is inactivated.

The study of animal models often provides valuable clues as to the function of specific proteins in humans. A variety of mouse models have been developed in an attempt to further elucidate the molecular biology of Schmerlin. Heterozygous NF2 mutant mice develop various neoplasms which show inactivation of the wild-type copy of *Schmerlin* (in keeping with its function as a TSG). However, these mice do not develop the characteristic tumors seen in human NF2. Furthermore, homozygous NF2 mutant mice (homozygous *Schmerlin* knockouts) are embryonic lethal and are therefore unhelpful as a model for human disease. In an attempt to overcome the deficiencies of the heterozygous and homozygous NF2 mutant mice, conditional mutant mice were generated which have restricted homozygous *Schmerlin* mutations in predominantly Schwann cell populations. These 'conditional' (Schwann cell) NF2 knockout mice show many similarities with human NF2 disease – for example, Schwann cell hyperplasia, Schwann cell tumors, cataracts, and cerebral calcification. They do not, however, develop vestibular schwannomas or meningiomas, two of the hallmark features of NF2. The conditional knockout mice have already provided important details regarding certain domains within *Schmerlin* (exon 2) and insight into how its protein product may function in cell migration and differentiation. Conditional NF2 knockout mice do aid in the understanding of the molecular mechanisms involved. However, one needs to be cautious about extrapolating the findings in rodents to human disease.

Conclusions

The neurofibromatoses are a complex group of genetic diseases which have been the principal research subject of numerous independent groups for more than three decades. This has resulted in the accumulation of a vast array of information during this period. A novel phase in molecular biology is now beginning which will characterize how genetic alterations in *Schmerlin* result in the disruption of the regulation of different proteins and signaling pathways, culminating in tumor formation. The biological interactions that have recently been elucidated may prove even more exciting than the findings of the last century. Judging by the pace of publications during the past 6 years, it may not be long before therapeutic strategies for NF2 and its tumors become a reality.

See also: Demyelination and Demyelinating Antibodies; Neurofibromatosis Type 1: Molecular and Cellular Biology; Neurofibromatosis Type 1; Neurofibromatosis Type 2; Tumors of the Brain and the Spinal Cord.

Further Reading

Evans DGR, Huson SM, Donnai D, et al. (1992) A genetic study of type 2 neurofibromatosis in the United Kingdom. I: Prevalence, mutation rate, fitness, and confirmation of maternal transmission effect on severity. *Journal of Medical Genetics* 29: 841–846.

Feiling A and Ward E (1920) A familial form of acoustic tumor. *British Medical Journal* 10: 496–497.

Gardner WJ and Frazier CH (1930) Bilateral acoustic neuromas: A clinical study and field survey of a family of five generations with bilateral deafness in thirty-eight members. *Archives of Neurology and Psychiatry* 23: 266–302.

Kanter WR, Eldridge R, Fabricant R, et al. (1980) Central neurofibromatosis with bilateral acoustic neuroma. Genetic, clinical and biochemical distinctions from peripheral neurofibromatosis. *Neurology* 30: 851–859.

Knudson AG (1971) Mutation and cancer: Statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* 68(4): 820–823.

National Institutes of Health (1987) National Institutes of Health Consensus Development Conference Statement on Neurofibromatosis. *Archives of Neurology* 45: 575–579.

Rouleau G, Seizinger BR, Wertelecki W, et al. (1987) Genetic linkage analysis of bilateral acoustic neurofibromatosis to a DNA marker on chromosome 22. *Nature* 329: 246–248.

Rouleau GA, Merel P, Lutchman M, et al. (1993) Alteration in a new gene encoding a putative membrane-organizing protein causes neurofibromatosis type 2. *Nature* 363: 515–521.

Ruttledge MH, Sarrazin J, Rangaratnam S, et al. (1994) Evidence for the complete inactivation of the NF2 gene in the majority of sporadic meningiomas. *Nature Genetics* 6(2): 180–184.

Trofatter JA, MacCollin MM, Rutter JL, et al. (1993) A novel moesin-ezrin-radixin like gene is a candidate for the neurofibromatosis 2 tumor suppressor. *Cell* 72: 791–800.

Wertelecki W, Rouleau GA, Superneau DW, et al. (1988) Neurofibromatosis 2: Clinical and DNA linkage studies of a large kindred. *New England Journal of Medicine* 319(5): 278–283.

Wishart JH (1822) Case of tumors in the skull, dura mater, and brain. *Edinburgh Medical and Surgical Journal* 18: 393–397.

Zankl H, Weiss AF, and Zang KD (1975) Cytological and cytogenetical studies on brain tumors. VI: No evidence for a translocation in 22-monosomic meningiomas. *Humangenetik* 30(4): 343–348.

Neurofilaments: Organization and Function in Neurons

C S Lobsiger and D W Cleveland, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Neurofilaments in Neuronal Function

Introduction to Neurofilaments

A neuron is composed of the neuronal cell body (perikaryon) and several processes (dendrites and axon), of which the axon is usually the longest one. The axon is comparable to an electrical wire that the neuronal cell body extends to other neurons or muscles to transmit its signal. Assembled neurofilaments are an essential part of the cytoarchitecture of mature axons and are particularly important in establishing the diameter of the largest axons, where they are the most abundant structural component. The first description of this fibrous axonal network dates back to the late nineteenth century, when newly developed silver staining methods allowed anatomists to visualize so-called neurofibrils within axons. Neurofilaments form a very dense network of discontinuous, cablelike filaments (oriented along the axon) which are interconnected by small extensions. This reversibly cross-linked meshwork of structural elements provides mechanical stability to the axon (Figure 1).

Structure of the Neurofilament Proteins

The subunits of neurofilaments belong to one of five classes of proteins and assemble into long protein polymers named intermediate filaments. There are more than 50 different kinds of these proteins in humans. Intermediate filaments are involved in cell architecture of most eukaryotic cells. The name 'intermediate' derives from the diameter of the assembled filaments (8–10 nm), which distinguishes them from the two other major structural polymers of the cytoskeleton, the actin microfilaments (6 nm) and the microtubules (24 nm). In a large myelinated axon, the neurofilaments make up the bulk of the intra-axonal volume, with long polymers interlinked to each other and to the thicker and structurally more rigid microtubules, which are outnumbered (by 10- to 30-fold) by neurofilaments. The actin microfilaments form a spiderweb-like structure underneath the axonal membrane. Both neurofilaments and microtubules are aligned in the same orientation along the axon, and all three cytoskeletal structures (actin microfilaments, neurofilaments, and microtubules) are interconnected by a separate family of proteins, the plakins, to produce a reversibly cross-linked filament meshwork that proves a flexible, deformable array that provides mechanical strength.

Neurofilaments are obligate heteropolymers requiring their smallest subunit, neurofilament light (NF-L; 60 kDa), and the two large subunits, neurofilament medium (NF-M; 90 kDa) and neurofilament heavy (NF-H; 115 kDa). Like all other intermediate filament subunits, the molecular structure of each of the three neurofilament subunits contains a central helical rod domain of approximately 310 amino acids and a subtype-specific globular domain at the head. The rod domain is highly hydrophobic and can form stable coiled-coil dimers with two protein subunits wrapped around each other within the rod domain, forming an initial assembly unit around 50 nm in length. The actual neurofilament is built up by assembling the coiled-coil dimers and then aligning both longitudinally and laterally in a staggered fashion, ultimately to form a ropelike structure 10 nm in diameter and of indeterminate length. NF-M and NF-H are unique among the intermediate filament subunits in that each contains an extended tail domain (about 400 and 600 amino acids long, respectively) which protrudes from the axis of the assembled filament. The NF-M and NF-H tail domains protrude up to 80 nm from the assembled filaments and interact with neighboring neurofilaments and microtubules (Figure 1). The stoichiometry of the neurofilament subunit composition varies but is approximately 5:3:1 for NF-L:NF-M:NF-H.

Peripherin and α-internexin represent two additional components of neuronal intermediate filaments present in a subset of neurons. They share the structural features of intermediate filament subunits, and both can co-assemble with neurofilaments. In the mature nervous system, peripherin is expressed predominantly in small-caliber sensory neurons of the peripheral nervous system and also to some extent in motor neurons of the central nervous system (CNS). α-Internexin is a component of certain types of CNS neurons, mainly the cerebellar granule cells.

Function of Neurofilaments

In the human nervous system, the largest neurons and the ones with the longest axons are the lumbar spinal cord motor neurons, which mainly innervate the leg muscles. There are thousands of these motor neurons, each with a cell body around 50 μm in diameter and each extending a single axon to its target muscle. These are the most asymmetric cells in nature, with axons in humans up to a meter in length (the full length of the leg), but even the largest have a maximal

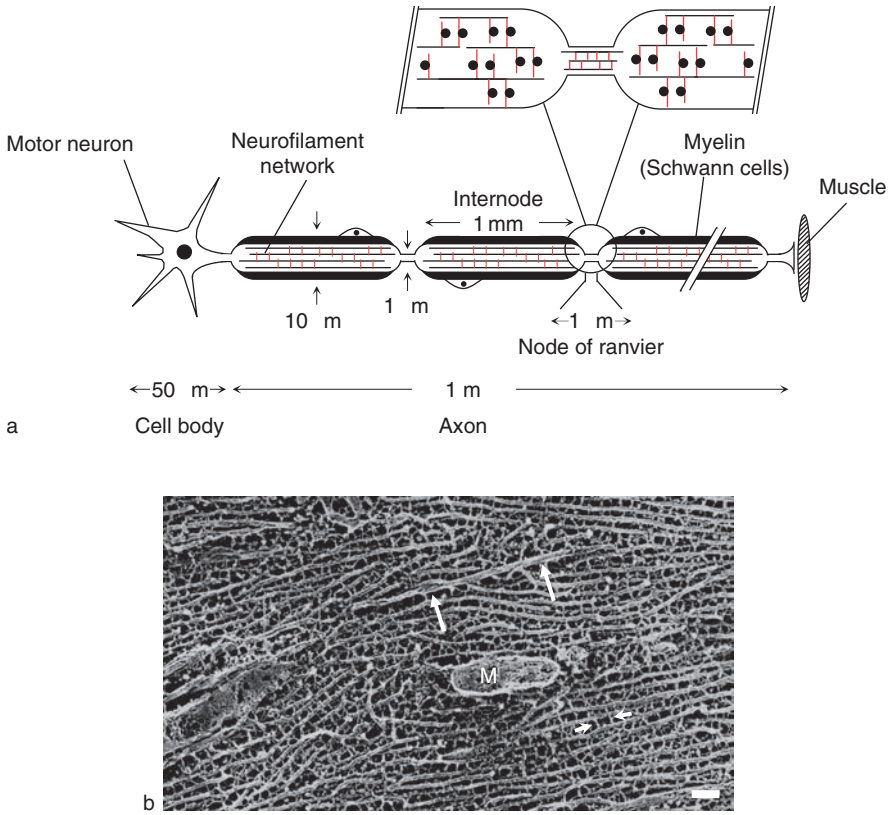

Figure 1 Neurofilaments are the major component of the cytoskeleton of large myelinated axons. (a) Schematic representation of the structural features of a human spinal cord motor neuron innervating muscle with the myelinating Schwann cells along the axon. The neurofilament network is represented by black lines, with the cross-linking neurofilament tail domains in red. The enlargement shows that the tail domains of the neurofilaments are heavily phosphorylated (black circles) in the internodes (large axonal caliber), whereas at the Node of Ranvier they are not (small axonal caliber). (b) Neurofilaments in the axon. Quick-freeze deep-etch view of the axonal cytoskeleton from a large myelinated axon, showing the neurofilament network. Arrowheads point to cross-bridges between the 10-nm-diameter neurofilaments, and arrows point to the single microtubule in the field. M, mitochondria. Scale bar = 100 nm. Reproduced from Hirokawa N (1982) Cross-linker system between neurofilaments, microtubules, and membranous organelles in frog axons revealed by the quick-freeze, deep-etching method. *Journal of Cell Biology* 94(1): 129–142, by copyright permission of The Rockefeller University Press.

axonal diameter (14 μm in humans) that is only 1/70 000th of their length. A real-life example of these cell dimensions is a soccer ball (with a diameter of 22 cm) as the cell body and a tube 6 cm in diameter extending over the length of 44 soccer fields, each 100 m long, as its axon. A large motor axon is more than 2000 times the volume of its cell body.

Neurofilaments are central determinants in generating normal axonal diameters. Mature axonal caliber is of importance for normal neuronal function because the speed of transmission of the electrical signal along the axon (the nerve conduction velocity) is directly proportional to the diameter (the bigger the diameter, the faster the speed). Complete loss of neurofilaments caused by loss of the major neurofilament subunit NF-L (either by gene deletion or by mutation that precludes synthesis of NF-L) blocks growth in axonal diameter that normally is contemporaneous with myelination. In mice, loss of neurofilaments is asymptomatic, but in larger animals (as seen in the quiver quail), absence of neurofilaments and reduced conduction velocity are accompanied by constant quivering and behavioral abnormalities.

The extensive, multiphosphorylated tail domains of NF-M and NF-H subunits (in human with 43 or 44 sites on NF-H and at least seven sites on NF-M) have, for a long time, been proposed to mediate neurofilament-dependent growth of axonal diameter that takes place within myelinated axonal segments. Although it has been proposed that space-filling properties of neurofilaments are achieved by electrostatic repulsion between these highly charged domains on adjacent neurofilaments, this is not the case. Deletion of the entire NF-H tail domain has little effect on radial growth, with organization of the remaining neurofilament array largely unaffected. Removal of the NF-M tail, on the other hand, dramatically limits acquisition of correct axonal diameter, producing a

50% reduction of caliber (fourfold reduction in overall axonal volume) and concomitant lower nerve conduction velocities. Replacement by transplantation of normal Schwann cells with mutant ones that are unable to produce myelin blocks this growth in diameter, but only within the unmyelinated segment, precluding neurofilament tail phosphorylation in the transplanted axonal region, whereas in the neighboring nerve regions with normal myelination, caliber and phosphorylation state remain normal. Thus, growth in diameter is limited to myelinated portions of the axon, evidence supporting the current model that correct diameter is achieved through an 'outside-in' signal generated by the enwrapping myelinating Schwann cells (Figure 1).

The Involvement of Neurofilaments in Neurodegenerative Diseases

Abnormal assembly and/or accumulations of neurofilaments are thought to be involved in the pathology of several human neurodegenerative diseases, including amyotrophic lateral sclerosis (ALS), infantile spinal muscular atrophy, and hereditary sensory-motor neuropathy. This is especially so in ALS (known familiarly in the United States as Lou Gehrig's disease), the most prominent human adult motor neuron disease. ALS is typically fatal, with onset between 50 and 60 years of age and complete paralysis within 1–5 years after onset, mainly resulting from premature death of upper (brain) and lower (spinal cord) motor neurons, especially the subset with the largest axons and the highest neurofilament contents. Aberrant neurofilament accumulations both in the neuronal cell bodies and in axons are a major pathological hallmark in both sporadic and familial ALS cases (Figure 2).

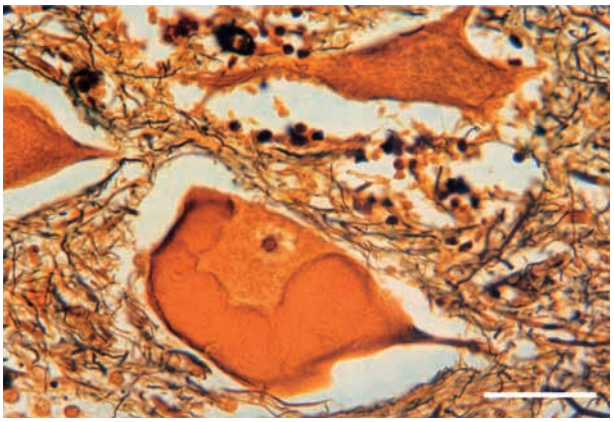

Figure 2 Accumulations of neurofilaments in motor neurons in the spinal cord of a patient with amyotrophic lateral sclerosis. Most of the perinuclear area is occupied by abnormal aggregates of 10 nm neurofilaments. Scale bar = 40 μm. Reproduced with permission from Hirano A (1988) *Color Atlas of Pathology of the Nervous System*, 2nd edn. Tokyo: Igaku-Shoin Medical Press, with permission from Igaku-Shoin Ltd.

Disorganized Neurofilament Arrays as a Cause for Motor Neuron Disease

Altered neurofilament organization can be a primary cause of motor neuron disease. Expression of a mutant NF-L subunit whose incorporation into neurofilaments disrupts their continued assembly provokes an aggressive, fatal, early onset motor neuron disease producing paralysis caused by massive neurodegeneration and death of motor neurons. Even increased synthesis of normal NF-L or NF-H subunits has been shown to produce abnormal accumulations of neurofilaments in motor neuron cell bodies and proximal axons, accompanied by motor neuron dysfunction and atrophy of the target muscles, very similar to the pathological signs of ALS.

Mutations in Neurofilament Genes Linked to Human Motor Neuron Diseases

Genetic evidence in humans has suggested mutations in neurofilament genes as contributors or direct causes for human motor neuron diseases. The most compelling evidence comes from missense mutations in the *NF-L* gene that are strongly associated with human Charcot-Marie-Tooth Disease Type 2E (CMT2E). CMT2E is a milder motor neuron disease than ALS, with an earlier onset and a slower progression and typified by muscle weakness and (depending on the severity) partial paralysis. With respect to ALS, several mutations have been found in the tail domain of the *NF-H* gene in a small group of sporadic ALS patients but not in control individuals. A large-scale sequencing of DNAs from ALS and control individuals has demonstrated that neurofilament mutations are not a significant primary cause of ALS, albeit they may be risk factors for ALS. These findings suggest, but fall short of proving, linkage of neurofilament sequence variants as contributors to human disease.

Peripherin is another intermediate filament subunit of mature motor neurons, and an abnormal version of it has also been suggested to be involved in ALS. Peripherin is normally encoded by two differentially spliced mRNAs transcribed from the same gene and encoding proteins with slightly different molecular weights (58 and 56 kDa). An aberrantly spliced form that generates a 61 kDa protein and that is both assembly incompetent and toxic *in vitro* in neuronal cells has been detected in motor neurons of a few human ALS patients and in ALS mouse models, suggesting a linkage of aberrantly spliced peripherin and human motor neuron disease.

Axonal Neurofilaments as Determinants of Neurodegeneration in ALS

Most incidences of ALS are sporadic, that is, they are without evidence for a genetic origin. Approximately 10% of ALS is inherited in an autosomal dominant manner, and a proportion of those are caused by missense mutations in the gene encoding the ubiquitously expressed Cu/Zn-superoxide dismutase 1 (SOD1). Mice that accumulate such mutations develop a fatal-progressive motor neuron disease comparable to ALS in humans, accompanied by massive motor neuron death and paralysis. As in ALS in humans, it is specifically the largest motor neurons with the biggest axons and highest neurofilament content that are most affected.

Neurofilament accumulations in spinal motor neurons are prominent hallmarks of all ALS disease forms, including mutant SOD1-induced ALS, both in humans and in transgenic mouse models. Elimination of assembled neurofilaments from axons (by deleting NF-L) or trapping most neurofilaments in motor neuron cell bodies (by overexpressing NF-H) both greatly slow ALS-like disease. Although the disease-modifying effect of altering neurofilament content could be achieved through increased accumulation of neurofilaments in the neuronal cell bodies or their removal from axons, removal of the NF-M and NF-H tail domains (by gene replacement in mice) provides a comparable benefit without altering axonal neurofilament content. This outcome eliminates the possibility that the phosphorylation site-rich tail domains of NF-M and NF-H can serve as phosphorylation sinks for buffering a detrimental mutant SOD1-induced hyperactivation of kinases, including cyclin-dependent kinase 5. Thus, neurofilaments are a modifier of ALS caused by mutant SOD1, probably through their influence on slowing transport of other cargoes transported along the axon.

Conclusion

Neurofilaments are essential for establishing a flexible, deformable, reversibly cross-linked array that supports growth in axonal diameter. Absence of neurofilaments reduces conduction velocity, resulting in behavioral abnormalities. Changes or dysregulations of the neuronal neurofilament network can induce pathological signs of neurodegeneration and even neuronal death. Certain human neurodegenerative diseases, such as CMT2E, are directly caused by mutations in a neurofilament gene. Other human neurodegenerative diseases, such as the motor neuron disease ALS, show pathological neurofilament accumulations, although mutations in any of the three neurofilament genes represent disease modifiers rather than direct causes. Changes in the neurofilament network can strongly influence the ALS disease course.

See also: Amyotrophic Lateral Sclerosis (ALS); Axonal Transport and ALS; Axonal Transport and Neurodegenerative Diseases; Intermediate Filaments.

Further Reading

Bruijn LI, Miller TM, and Cleveland DW (2004) Unraveling the mechanisms involved in motor neuron degeneration in ALS. *Annual Review of Neuroscience* 27: 723–749.

DeWaegh SM, Lee VM, and Brady ST (1992) Local modulation of neurofilament phosphorylation, axonal caliber, and slow axonal transport by myelinating Schwann cells. *Cell* 68: 451–463.

Fuchs E and Cleveland DW (1998) A structural scaffolding of intermediate filaments in health and disease. *Science* 279: 514–519.

Garcia ML, Lobsiger CS, Shah SB, et al. (2003) NF-M is an essential target for the myelin-directed "outside-in" signaling cascade that mediates radial axonal growth. *Journal of Cell Biology* 163: 1011–1020.

Hirano A (1988) *Color Atlas of Pathology of the Nervous System*, 2nd edn. Tokyo: Igaku-Shoin Medical Press.

Hirokawa N (1982) Cross-linker system between neurofilaments, microtubules, and membranous organelles in frog axons revealed by the quick-freeze, deep-etching method. *Journal of Cell Biology* 94(1): 129–142.

Julien JP and Kriz J (2006) Transgenic mouse models of amyotrophic lateral sclerosis. *Biochimica et Biophysica Acta* 1762 (11–12): 1013–1024.

Lariviere RC and Julien JP (2004) Functions of intermediate filaments in neuronal development and disease. *Journal of Neurobiology* 58: 131–148.

Lee MK and Cleveland DW (1996) Neuronal intermediate filaments. *Annual Review of Neuroscience* 19: 187–217.

Lobsiger CS, Garcia ML, Ward CM, and Cleveland DW (2005) Altered axonal architecture by removal of the heavily phosphorylated neurofilament tail domains strongly slows superoxide dismutase 1 mutant-mediated ALS. *Proceedings of the National Academy of Sciences of the United States of America* 102: 10351–10356.

Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain

J G Emsley*, **U S Sohur***, **and J D Macklis**,
MGH–HMS Center for Nervous System Repair, and
Harvard University, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Until very recently, it was thought that the adult brain was incapable of generating new neurons or having neurons added to its complex circuitry. However, the development of new techniques has resulted in a major increase in research demonstrating that the adult central nervous system (CNS) contains neural progenitors, precursors, and stem cells that are capable of generating new neurons, astroglia, and oligodendroglia. It has also been shown that neurogenesis constitutively occurs in two specific regions of the adult mammalian brain (olfactory bulb and hippocampal dentate gyrus), and that a significant number of multipotent neural precursors, or 'stem cells,' reside in many parts of the adult CNS. While challenging previous dogma that no new neurons are born in the adult mammalian CNS, these findings bring with them future possibilities for the development of novel neural repair strategies.

A substantial body of research regarding constitutively occurring neurogenesis provides insight into the potential for and limitations of neuronal replacement therapies based on manipulation of neural precursors. Recent work has partially elucidated the normal behavior of endogenous adult precursors, including their ability to migrate to select brain regions, differentiate into neurons, integrate into normal neural circuitry, and, finally, functionally integrate into the adult brain. Research is also beginning to identify and describe molecular and activity-related controls over constitutively occurring neurogenesis.

Defining Neural Stem Cells, Progenitors, and Precursors

Rigorously defined, adult CNS stem cells exhibit three cardinal features: (1) self-renewal, with the theoretically unlimited ability to produce progeny indistinguishable from themselves; (2) they are proliferative, continuing to undergo mitosis (although perhaps with quite long cell cycles); and (3) they are multipotent for the different neuroectodermal lineages of the CNS, including the multitude of different neuronal and glial subtypes. Multipotent progenitors of the adult brain are proliferative cells with only limited self-renewal that can differentiate into at least two different cell lineages (multipotency). Lineage-specific precursors or progenitors are cells with restriction to one distinct lineage (e.g., neuronal, astroglial, or oligodendroglial). Together, CNS stem cells and all precursor/progenitor types are, broadly defined, 'precursors.' We favor reserving the term stem cell as rigorously defined previously and using 'precursor' or 'progenitor' for most forms of multipotent or lineage-restricted mitotic cells. We use 'precursor cell' as a generic term encompassing both stem and progenitor cells.

Adult neurogenesis comprises all the events of neuronal development, beginning with the division of a precursor cell and ending with the survival of a mature, integrated, functioning new neuron. Neurogenesis is sometimes incorrectly used only in the sense of 'precursor cell proliferation,' but this definition clearly falls short. Precursor cell proliferation is not predictive of net neurogenesis. True neuronal integration depends on many complex variables and progressive events.

Even in the absence of a fully detailed molecular characterization, neural precursor cells of the adult brain clearly possess features that justify a view of these cells as extremely promising for the goals of CNS cellular repair. They are undifferentiated, often highly mobile cells, relatively resistant to hypoxia and other injury, proliferatively active, and able to produce mature neurons and glia. The most critical caveat, however, is that precursor cells in both the developing and the adult mammalian brain are heterogeneous, and although all of these criteria might apply in many cases, it is not clear how many of the sometimes site-specific and condition-specific observations in neural precursor biology can be broadly generalized and how many reflect particulars of a certain brain region or a certain condition.

Although the identity of 'true' CNS stem cells is unknown, limiting some conclusions, the heterogeneity of neural precursor cells is also an important aspect of the promise of such cells. Their variable specialization and differentiation competence provides great promise for future clinical applications. The idea that 'one cell fits all' might turn out to be as unrealistic as it is unnecessary. Adult neural precursors with partial fate restriction may, in some cases, allow for more efficient production of desired cell types. The brain may provide the environmental determinants for the different forms and sequelae of adult neurogenesis and gliogenesis, and it may powerfully enable and direct specific lineage development.

*These authors contributed equally to this work.

The Discovery of Constitutively Occurring Adult Mammalian Neurogenesis

Ramon y Cajal has been widely quoted as writing that "in the adult centres the nerve paths are something fixed, ended and immutable. Everything may die, nothing may be regenerated." The lack of recovery from CNS injury and neurodegenerative disease in the adult mammalian CNS is in marked contrast to the enormous regenerative potential of the CNS in non-mammalian vertebrates. Indeed, the consensus in the field for some time was that neurogenesis does not occur in the adult mammalian brain. Joseph Altman was the first to use techniques sensitive enough to detect the ongoing cell division that occurs in the adult brain. Using tritiated thymidine as a mitotic label, he published evidence that neurogenesis constitutively occurs in the mammalian hippocampus and olfactory bulb. These results were later replicated using tritiated thymidine labeling followed by electron microscopy. However, the absence of neuron-specific immunocytochemical markers at the time resulted in identification of putatively newborn neurons being made on purely morphological criteria. These limitations led to a widespread lack of acceptance of these results and made research in the field difficult.

The field of adult neurogenesis was rekindled in 1992 with demonstrations that precursor cells isolated from the forebrain can differentiate into neurons *in vitro*. These results, and technical advances such as the development of immunocytochemical reagents that could more easily and accurately identify the phenotype of various neural cells, led to an explosion of research in the field. Normally occurring neurogenesis in the subventricular zone (SVZ)/olfactory bulb and the dentate gyrus has now been well characterized in the adult CNS.

Neurogenic versus Nonneurogenic Regions in the Adult Brain

Neurogenesis constitutively occurs only in the olfactory bulb and the dentate gyrus of the hippocampus. In the olfactory system, precursor cells reside in the anterior aspect of the SVZ (sometimes also called the subependymal zone) in the walls of the lateral ventricles, migrate via 'chain migration' through a structure surrounding the remnants of the olfactory ventricle (the rostral migratory stream (RMS)) into the olfactory bulb, and differentiate into granule or periglomerular interneurons. In the dentate gyrus of the hippocampus, precursor cells are found in the subgranular zone (SGZ). Here, new hippocampal granule cell neurons are produced. These two brain regions are thus referred to as 'neurogenic regions.' As discussed later in more detail, a report of extremely low-level constitutive neurogenesis in the normal neocortex has been disputed on important methodological grounds, leaving this reported result quite controversial. In addition, the same group later modified its interpretation by stating that the existence of these neurons was only transient. There have also been reports of extremely low-level constitutive adult neurogenesis in the amygdala, area cornu ammonis field1 (CA1) of the hippocampus, the dorsal vagal complex of the adult brain stem, the spinal cord, the substantia nigra, and the hypothalamus, although some of these results have been disputed. Overall, these isolated data remain controversial, but not because the possibility of extremely low-level neurogenesis occurring in essentially nonneurogenic regions is remote. Instead, isolated reports of such a phenomenon have not sufficiently overcome many of the important methodological issues mentioned previously.

Transplantation studies support the principle of defining neurogenic and nonneurogenic regions, and they provide evidence about the role of the microenvironment in influencing the potential of neural precursors. If precursor cells are transplanted into neurogenic regions, they can differentiate into neurons in a region-specific manner. SVZ precursor cells generate hippocampal neurons when transplanted into the hippocampus, and SGZ precursor cells generate olfactory interneurons after transplantation into the RMS. When implanted outside the neurogenic regions, both types of precursor cells generate only glia. These data further support the interpretation that neurogenesis is dependent on a permissive microenvironment rather than on the regionally different properties of precursor cells. Local microenvironmental influences on the behavior of precursors and their ability to differentiate into neurons support the argument that it will be critical for the goal of neuronal repopulation to understand the cellular and molecular controls over cell type-specific precursor cell differentiation in the adult CNS.

It came as a surprise to many in the field that neurogenic regions are defined by the neurogenic permissiveness of the local microenvironment rather than by the presence of neural precursor cells. Neural precursor cells have been found in a large number of brain regions, including white matter tracts. It is possible that the entire brain contains neural precursor cells (albeit at very low density), just as it contains partially restricted glial precursors that act as progenitor cells. These widely distributed neural precursor cells do not seem to differ fundamentally from one another, although they are far from homogeneous, with notably different growth kinetics and likely at

least partially restricted differentiation potential. Their function outside the classical neurogenic regions is not known (e.g., whether evolutionary remnants or occasionally activated in certain circumstances), nor is their relation to the precursor cells of the neurogenic regions. For example, neural precursor cells from the spinal cord behave *in vitro* much like their counterparts from the SGZ and SVZ, and they are multipotent after implantation into the hippocampus, producing granule cell neurons, but only generate glial cells *in situ* or when transplanted back to their original site in the spinal cord. *In vivo*, no adult neurogenesis can be found in the rodent spinal cord. In summary, the extent of neurogenesis in a region is a function of the local microenvironment on precursor cells with (varying) neurogenic potential and differentiation restriction.

Therefore, there is an important conceptual distinction between neurogenic and nonneurogenic regions, but it might turn out that this distinction reflects complex molecular and functional states rather than a fixed cellular environment. Can nonneurogenic regions be molecularly modified to promote neurogenesis? Could this change in neurogenic permissiveness be induced under certain pathological conditions? There are strong natural precedents in other vertebrate systems, such as the avian telencephalon and the olfactory epithelium, for the concept that selective neuronal death can alter the differentiation of precursors, thereby inducing or increasing neurogenesis. Although not yet proved, it is hypothesized that this change in neurogenesis occurs via alterations in the local molecular microenvironment and intercellular interactions with partner glial and interneuron populations that resemble conditions seen during nervous system development and in adult neurogenic regions.

Several groups have demonstrated the induction of modest levels of neurogenesis in normally nonneurogenic regions in response to selective neuronal death or degeneration. As described in more detail later, we found that endogenous multipotent precursors normally located in the adult brain can be induced to differentiate into neurons in the normally nonneurogenic adult mammalian neocortex – a result extended to corticospinal motor neurons. Other groups have reported similar and complementary results in a normally nonneurogenic region of the hippocampus and in the striatum, confirming and further supporting this direction of research. These reports of seemingly regenerative neurogenesis in the vicinity of targeted cortical neuron degeneration and cerebral ischemia suggest that the lack of normal neurogenic permissiveness in these regions is not unalterable. The controversial examples of reportedly extremely low-level constitutive neurogenesis outside the classical neurogenic regions might also support this possibility, potentially as 'leakage' from normally tightly inhibited systems. Precursor cells resident in nonneurogenic regions appear at least partially competent to respond to either instructive neurogenic signals or release from inhibition to induce them to develop into new, potentially function-restoring neurons.

The vast majority of studies investigating potential neurogenesis in the neocortex of the well-studied rodent brain do not report normally occurring adult cortical neurogenesis. Rigorous studies employing serial-section analysis and three-dimensional confocal reconstruction demonstrate a complete absence of constitutively occurring neurogenesis in the murine neocortex. These experiments provide evidence that satellite glial cells, closely apposed to neurons in normal neocortex, could be mistakenly interpreted as adult-born neurons when, in fact, a newborn glial cell overlies a preexistent neuronal nucleus. However, a few studies report low-level, constitutively occurring neurogenesis in specific regions of the neocortex of adult primates and in the visual cortex of adult rat. In one such study, neurogenesis of two or three new neurons per cubic millimeter was reported in prefrontal, inferior temporal, and posterior parietal cortex of the adult macaque (in one case, after likely damage from prior electrophysiological recordings in the same regions) but not in striate cortex, which was interpreted by the authors as being due to the fact that this is a simpler primary sensory area. These authors later reported that these cells are transient and do not survive. In contrast, other reports analyzed with more rigorous methods, as noted previously, have been unable to reproduce these findings and report a complete absence of constitutive cortical neurogenesis in both rodents and primates. There exists a single report of neurogenesis in the visual cortex of the adult rat, but this study used tritiated thymidine and purely morphological cell type identification, and it has not been confirmed by any other group (the satellite glia issue casts substantial doubt on this study performed prior to confocal microscopy availability). Work using ^{14}C counts from human tissue, based on the peri-1960 aboveground nuclear testing contamination of the atmosphere, confirms the absence of any substantial adult neocortical neurogenesis. It seems highly unlikely that even very low-level neurogenesis occurs normally in the neocortex of any mammals (and most data indicate none in rodents or primates), but further examination of the potential for constitutive neurogenesis in what are considered normally nonneurogenic regions will be required.

The situation in the literature is far from clear, but it is important to remain critical and cautious in interpretation. After all, the brain does regenerate poorly. Although some of the newer results fundamentally change our view on the possibility of pathology-induced neuroplasticity, it should not be forgotten that there is little evidence that neurogenesis outside normally neurogenic regions by itself significantly contributes to brain repair. Attempts toward endogenous repair will almost certainly require exogenous activation and/or release of inhibition of the processes involved, control over directed subtype differentiation, and support of both neuron survival and axon outgrowth and pathfinding. Restorative neurobiology focusing on these approaches aims at activating the potential that lies in precursor cells in the neurogenic and nonneurogenic regions of the adult brain.

Induction of Neurogenesis in Nonneurogenic Regions of the Adult Brain

Induced neurogenesis in nonneurogenic regions can be envisioned in two basic forms: (1) Local parenchymal precursor cells might be activated to generate new neurons upon a stimulus associated either with a pathological event or with a set of directed molecular and/or genetic interventions, or (2) precursor cells from normally neurogenic regions (e.g., the SVZ) might respond to either type of stimulus and migrate into the perenchyma. There are already examples for both possibilities.

The first evidence that regenerative neurogenesis is possible was reported by our lab for the neocortex; this work served as a 'proof of concept' for the induction of adult mammalian neurogenesis. Work by our lab and other groups has extended this general concept importantly in new directions – in the striatum, hippocampus, hypothalamus, and corticospinal system. Whereas multipotent precursors are concentrated in the anterior SVZ and the dentate gyrus, they can be found in lower densities in several other regions of the adult brain. As discussed previously, these precursors have broad potential; they can differentiate into astroglia, oligodendroglia, and neurons; receive afferents; and extend axons to their targets, given an appropriate *in vitro* or *in vivo* environment. Prior results demonstrated that cortex undergoing targeted apoptotic degeneration could direct multipotent precursors to integrate into adult cortical microcircuitry. The population specificity of neuronal death was found to be essential, allowing for maintenance of local circuitry and survival of surrounding cells that alter gene expression to direct precursors to their final location. These properties of endogenous multipotent precursors led us to investigate the fate of these precursors in an adult cortical environment manipulated in a manner previously demonstrated to support neurogenesis.

We found that endogenous multipotent precursors could be induced to differentiate into neurons in the adult mouse neocortex, without transplantation. These experiments induced synchronous apoptotic degeneration of corticothalamic projection neurons and examined the fates of newborn cells within cortex using markers of progressive neuronal differentiation. The results demonstrated that endogenous neural precursors (mostly from the SVZ and possibly some from the parenchyma of the cortex) could be induced *in situ* to differentiate into cortical neurons, survive for many months, and form appropriate long-distance connections in the adult mammalian brain. These results indicated that the normal absence of constitutive cortical neurogenesis does not reflect an intrinsic limitation of the endogenous neural precursors' potential but more likely results from a lack of appropriate microenvironmental signals necessary for neuronal differentiation or survival. Elucidation of these signals could enable CNS repair.

Other groups have reported similar and complementary results in hippocampus and striatum, confirming and further supporting this direction of research. Nakatomi and colleagues used an *in vivo* ischemia model by which it was found that repopulation of neurons in the CA1 region of the hippocampus is possible following the elimination of CA1 neurons. It was found that the overwhelming majority of adult-born neurons repopulating the damaged CA1 region originated from a proliferative response in the posterior periventricle, and that this proliferative response could be augmented by infusion of high levels of epidermal growth factor (EGF) and fibroblast growth factor-2 (FGF-2), dramatically increasing the number of neurons able to migrate into and repopulate the damaged CA1 region. Although the levels of EGF and FGF-2 substantially exceeded those reasonable for human application, these experiments serve as a proof of principle for enhancement of an endogenous neurogenic response. We demonstrated induction of low-level neurogenesis of corticospinal motor neurons, with progressive extension of axons into the cervical spinal cord over 3 or 4 months. Flier and colleagues found that ciliary neurotrophic factor (CNTF) seemingly supports the survival of a very small number of new neurons of the hypothalamus involved in energy balance, rescuing them from what appears to be normal death following a possibly very low level of constitutive but abortive neurogenesis. Ultimately, a considerable and critical challenge will be to assess true functional integration of recruited adult-born neurons.

Taken together, these results demonstrate that endogenous neural precursors can be induced to differentiate into CNS neurons in a region-specific manner and at least partially replace neuronal populations that do not normally undergo neurogenesis. It appears that these results are generalizable to at least several classes of projection neurons, including the clinically relevant striatal medium spiny neurons and corticospinal motor neurons. Microenvironmental factors, importantly including signaling from local interneuron populations that normally interact during the development of their partner projection neuron populations, appear capable of supporting and instructing the neuronal differentiation and axon extension of adult-born neurons derived from endogenous precursors.

Conclusions and Future Prospects

Following the demonstration of the existence of neurogenesis in discrete regions of the adult mammalian CNS, a vast amount has been learned about the function of and controls over this set of processes. New knowledge is emerging regarding the lineage-specific differentiation of specific neuronal subtypes, particularly spinal motor neurons and corticospinal motor neurons and other forebrain projection neuron populations. Looking forward, it might be possible to induce the cellular repopulation of the diseased brain and spinal cord via specific activation and differentiation of endogenous neural precursors along desired neuronal or glial lineages. Even if multipotent precursors are not located in large numbers outside of normally neurogenic regions of the brain, it might be feasible to induce them to proliferate and differentiate from small populations that are more widely distributed throughout the neuraxis. There is accumulating evidence that endogenous neural precursors are regionally and even locally quite heterogeneous, and that subsets of partially fate-restricted precursors might exist local to regions in which neuronal replacement is desired. Such precursors might offer advantages in directing lineage-specific differentiation for cellular repair. However, this field is just at the beginning of understanding the complex interplay between neural precursors' potential and signals in their local microenvironment; much can be learned about precursor heterogeneity and how to take advantage of what may be partial cell type restriction, permissive and instructive developmental signals, and modulation of specific aspects of neuronal differentiation and survival. Progress during the past decade has been great, and the coming decades promise to offer significant insight into potential strategies for directed CNS repair.

See also: Axonal Injury: Neuronal Responses; Neural Stem Cells and CNS Diseases; Neural Stem Cells: Adult Neurogenesis; Neurogenesis in the Intact Adult Brain; Neuronal Plasticity after Cortical Damage; Olfactory Bulb Anatomy; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Altman J and Das GD (1965) Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *Journal of Comparative Neurology* 124: 319–335.

Arlotta P, Molyneaux BJ, Chen J, Inoue J, Kominami R, and Macklis JD (2005) Neuronal subtype-specific genes that control corticospinal motor neuron development *in vivo*. *Neuron* 45: 207–221.

Arvidsson A, Collin T, Kirik D, Kokaia Z, and Lindvall O (2002) Neuronal replacement from endogenous precursors in the adult brain after stroke. *Nature Medicine* 8: 963–970.

Emsley JG, Mitchell BD, Kempermann G, and Macklis JD (2005) Adult neurogenesis and repair of the adult CNS with neural progenitors, precursors, and stem cells. *Progress in Neurobiology* 75: 321–341.

Gage FH (2000) Mammalian neural stem cells. *Science* 287: 1433–1438.

McKay R (1997) Stem cells in the central nervous system. *Science* 276: 66–71.

Molyneaux BJ, Arlotta P, Hirata T, Hibi M, and Macklis JD (2005) Fezl is required for the birth and specification of corticospinal motor neurons. *Neuron* 47: 817–831.

Nakatomi H, Kuriu T, Okabe S, et al. (2002) Regeneration of hippocampal pyramidal neurons after ischemic brain injury by recruitment of endogenous neural progenitors. *Cell* 110: 429–441.

Reynolds BA and Weiss S (1992) Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science* 255: 1707–1710.

Richards LJ, Kilpatrick TJ, and Bartlett PF (1992) *De novo* generation of neuronal cells from the adult mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* 89: 8591–8595.

Sohur US, Emsley JG, Mitchell BD, Kempermann G, and Macklis JD (2006) Adult neurogenesis and cellular brain repair with neural progenitors, precursors, and stem cells. *Philosophical Transactions of the Royal Society Series B* 361: 1477–1497.

van der Kooy D and Weiss S (2000) Why stem cells? *Science* 287: 1439–1441.

Neurogenesis in the Intact Adult Brain

G Kempermann, Max Delbrück Center for Molecular Medicine (MDC) Berlin-Buch, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Adult neurogenesis is the development of new neurons from resident neural precursor cells in the adult brain. The term 'adult neurogenesis' encompasses the entire process of neuronal development from the division of the precursor cell to the existence of a functionally integrated new neuron. Adult neurogenesis is rare and locally restricted in mammals but frequent and widespread in lower vertebrates; it seems that over the course of evolution with increasing brain complexity adult neurogenesis became more and more limited. Quantitatively, most adult neurogenesis occurs relatively early in life, but adult neurogenesis persists lifelong on a very low level. There is no evidence that adult neurogenesis would primarily contribute to a neuronal turnover in that it would replace damage or lost neurons. Rather, the new neurons are added to the persisting networks.

In mammals, physiological adult neurogenesis appears to be restricted to two canonical neurogenic regions, the dentate gyrus of the hippocampus and the olfactory bulb (**Figure 1**). Adult hippocampal neurogenesis generates new excitatory granule cells, the principal cells of the dentate gyrus, whose axons form the mossy fiber tract to hippocampal subregion CA3. Adult olfactory neurogenesis produces two types of interneurons in the granule cell layer and the periglomerular regions of the olfactory bulb.

Adult neurogenesis in other brain regions, most notably the neocortex, remains controversial, and if it exists will range on a minute scale. Initial reports about the production of large numbers of cortical neurons were not confirmed by others. In contrast, the question of whether small numbers of cortical interneurons, primarily in the deep cortical layers, might be generated in adulthood is more difficult to prove or disprove. In addition, under particular circumstances (e.g., in the face of locally limited pathology), regenerative or targeted neurogenesis in physiologically nonneurogenic regions seems to be possible. Such exceptions might question a rigid conceptual distinction between neurogenic and nonneurogenic regions. For most practical purposes, the distinction, however, is valid. In contrast to the nonneurogenic areas, the canonical neurogenic regions contain a morphologically distinct neurogenic or stem cell niche that provides a microenvironment permissive for precursor cell function and neuronal development. The niche consists of the precursor cells and their progeny, supporting cells, immune cells, vasculature, and a specialized extracellular matrix, whose exact composition is still unknown. The structure of the precursor cell niches of the adult brain resemble similar niches, for example, in bone marrow, testes, and olfactory epithelium.

Precursor cells can also be found outside the neurogenic regions, albeit at very low density. No structural evidence of a niche is found in these regions and physiologically no (or extremely limited) neurogenesis appears to take place. The function of precursor cells outside the neurogenic regions is unknown. It is also not clear how homogenous the population of neural precursor cells in the adult brain is. Concepts of neuroanatomists like Wilhelm His, Alfred Schaper, and Wilder Penfield, who proposed the existence of a 'spongioblast' in the brain parenchyma, are thereby revived. Currently, the best candidates for the parenchymal precursor cells outside the neurogenic regions are slowly proliferative cells expressing the proteoglycan NG2. The neurogenic precursor cells in the hippocampus and olfactory system, in contrast, are NG2-negative but express markers including nestin, doublecortin, brain lipid binding protein, glial fibriallary acidic protein, Sox2, and others, but neither sensitivity nor specificity are sufficiently clear.

Meaning and relevance of markers that are associated with neurogenesis in the neurogenic regions (nestin, doublecortin, calretinin, etc.) are not known for the nonneurogenic regions. The presence of these markers alone should not be taken as evidence of a neurogenic potential or even ongoing neurogenesis.

Adult neurogenesis also takes place in the olfactory epithelium (as part of the peripheral nervous system), where olfactory receptor neurons are continuously replaced from local precursor cells.

History

Adult hippocampal neurogenesis was first described by Joseph Altman, then at the Massachusetts Institute of Technology, in the early 1960s. Altman also described postnatal neurogenesis in the olfactory bulb in 1969. The response by the scientific community was skeptic, largely for methodological reasons but also because of the fact that no stem cell population was known in the brain and the origin of new neurons thus remained speculative. In the 1980s, Fernando Nottebohm and colleagues described adult neurogenesis in those brain nuclei of canary birds that are responsible for song learning. These studies provided the first link between adult neurogenesis

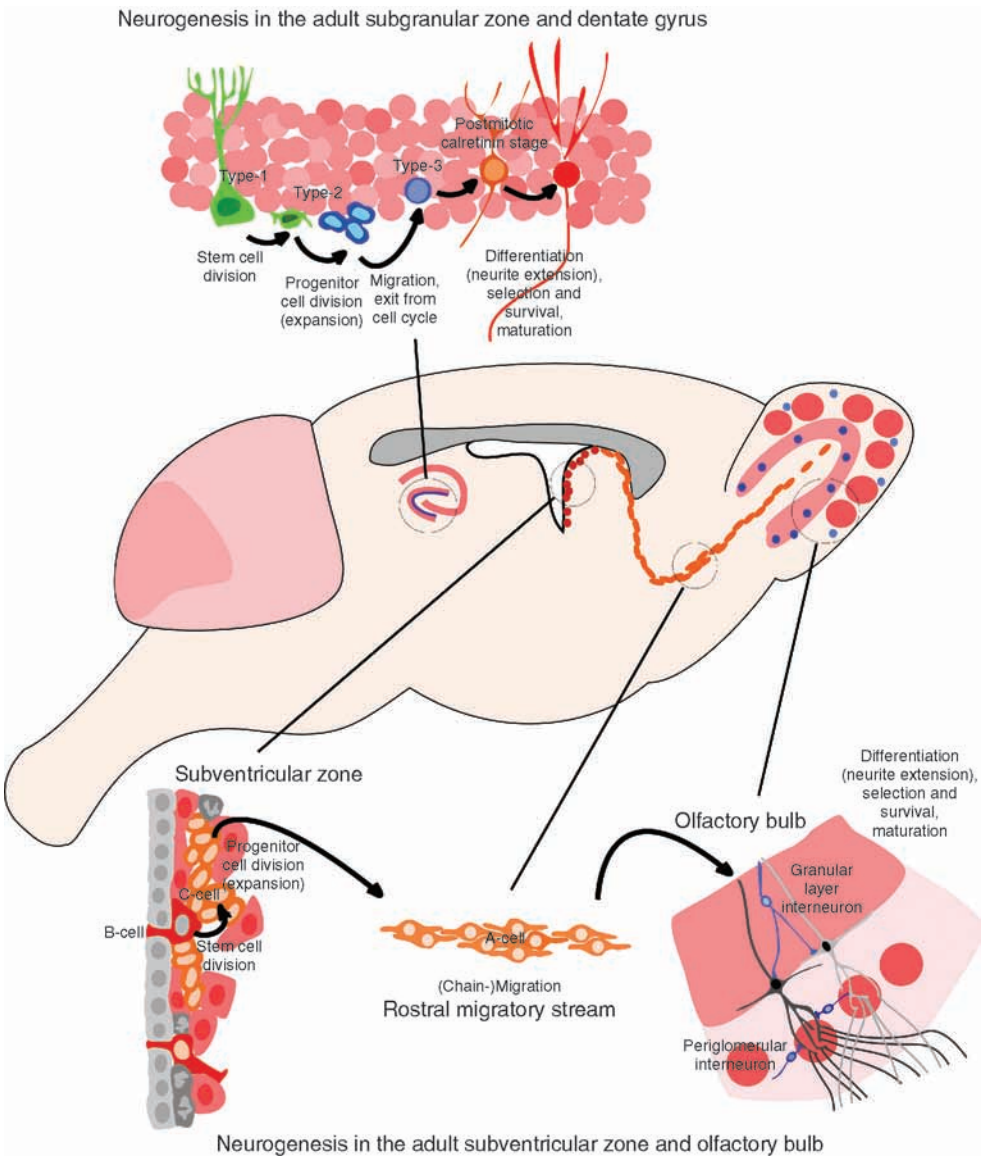

Figure 1 Adult neurogenesis occurs in two canonical neurogenic regions: the olfactory system and the hippocampus.

and behavior. After the discovery of neural stem cells in the adult mammalian brain in 1992, research on adult neurogenesis gained momentum. The isolation of stem and progenitor cells in the olfactory system was first reported by Brent Reynolds and Sam Weiss. Stem cells in the adult hippocampus followed in 1995 with the work of Theo D Palmer, Jasodarah Ray, and Fred H Gage. In the early 1990s, Elizabeth Could and co-workers reported that corticosteroids negatively affect adult neurogenesis. In 1997 and 1998, Fred H Gage's group showed that environmental enrichment and physical exercise positively influence adult hippocampal neurogenesis. The occurrence of adult neurogenesis in the human dentate gyrus was confirmed in 1998 by Eriksson and colleagues.

Adult Hippocampal Neurogenesis

The precursor cells driving adult hippocampal neurogenesis reside in the subgranular zone (SGZ) of the adult dentate gyrus. An astrocyte-like stem cell with radial morphology (type 1 cell) gives rise to highly proliferative intermediate progenitor cells (type 2 cells) and lineage-determined neuroblast-like cells (type 3 cells) that exit from the cell cycle and go through an immature postmitotic stage. During this stage, which coincides with the transient expression of calretinin, the new neurons fully extend their neurites and go through a phase of increased synaptic plasticity. The distinction of the cell types by numbers is preliminary and should one day be exchanged to a functional nomenclature.

Regulation of adult neurogenesis mainly occurs in two stages, an expansion phase on the level of the precursor cells and a phase of selective survival on a postmitotic stage. A transient expression of doublecortin lasts from the type 2 cell level to the calretinin phase. Doublecortin has become a surrogate marker of adult neurogenesis that is even used for quantification. The equalization of doublecortin expression with hippocampal neurogenesis, however, is not without problems because doublecortin expression ends before the cells are fully integrated. The newly generated cells receive first input through GABAergic synapses, which are excitatory. This input is involved in the induction of full maturation. The new cells receive glutamatergic input within days to weeks and go through a transient phase of increased synaptic plasticity (i.e., a lowered threshold for the induction of long-term potentiation). After approximately 7 weeks, the cells have become largely indistinguishable from the older granule cells and remain stably integrated into the network.

Adult Olfactory Neurogenesis

The precursor cells for adult olfactory neurogenesis are found in the subventricular zone (SVZ) of the lateral ventricle. The stem cell of this region has a cilia-bearing process touching the ventricular surface between the ependymal cells. The exact relationship between these stem cells (B cells) and the ependymal cells (E cells) is not clear; in particular, it remains controversial whether under certain conditions ependymal cells can act as stem cells. Undoubtedly, however, the astrocyte-like B cells do serve as stem cells and give rise to highly proliferative progenitor cells (C cells) that generate migratory neuroblasts (A cells). The A cells migrate in a specialized form of migration, homonymic chain migration, through the rostral migratory stream (RMS) to the core of the olfactory bulb, where they change for a radial migratory pattern and reach the granule cell layer and the glomerular and periglomerular regions. Precursor cells with a neurogenic potential can be isolated not only from the SVZ but also from the RMS and the olfactory bulb. All new neurons in the olfactory bulb are inhibitory interneurons and thus GABAergic. A subset of new periglomerular neurons co-expresses dopamine as neurotransmitter. Maturation of the new cells is delayed until they have reached their final position. The new cells in the granular layer express calretinin. The exact lineage relationship between the different types of new olfactory bulb neurons is not clear.

Regulation

Regulation of adult neurogenesis mainly falls into two categories. Numerous relatively nonspecific stimuli can affect the proliferation of precursor cells. Acutely this increase can lead to an increase in net neurogenesis, but this does not necessarily have to be the case. For many factors, the question of whether a particular stimulus has long-term effects on net neurogenesis has not yet been specifically addressed. Many of the nonspecific stimuli are associated with 'activity,' and many examples of pathology (ischemia, trauma, etc.) can upregulate neurogenesis. Other stimuli such as stress and inflammation downregulate precursor cell proliferation and neurogenesis. For many such regulators, complex dose–response relationships might exist so that generalization is difficult. Age can be considered a strong negative regulator early in life but apparently loses this strong impact, because with increasing age adult neurogenesis settles at a very low level.

Stimuli that appear to be of more specific relevance to the hippocampus or the olfactory system act on the selective survival of the new cells. Under physiological conditions, these stimuli tend to be associated with the function of the hippocampus (memory, learning, etc.) or the olfactory system (olfaction). Genetical studies indicate that regulation on the survival and differentiation level has a larger quantitative (and presumably qualitative) impact on the control of adult neurogenesis than the regulation of cell proliferation.

On a mechanistic level, regulation involves several neurotransmitter systems that provide input to the SGZ and SVZ. For the SGZ, the research focus has been on the GABAergic, the serotonergic, and the glutamatergic system. High levels of glucocorticoids acutely suppress neurogenesis. Presumably because the expression of corticosteroid receptor expression changes in the course of neuronal development in the adult, the net effect under prolonged exposure to glucocorticoids is complex. Many situations with elevated levels of glucocorticoids are thus associated with increased levels of neurogenesis. The details of this regulation are unknown. The role of other hormone systems is even more ambiguous.

Growth factors and neurotrophic factors, most notably fibroblast growth factor 2 (FGF2) and epidermal growth factor (EGF), strongly affect adult neurogenesis in a complex way. Insulin-like growth factor 1 (IGF1) and vascular endothelial growth factor (VEGF) have been discussed as centrally involved in the activity-dependent regulation of adult neurogenesis. Brain-derived neurotrophic factor (BDNF) is widely proposed to play a major role in controlling adult neurogenesis but experimental evidence supporting this assumption so far remains limited.

Among the short-range-acting signaling molecules with effects on adult neurogenesis contributions of the ephrins, bone morphogenic proteins (BMPs) and

the Wnt system have been described. More are likely to follow. It is assumed that the general and intrinsic developmental principles regulating neuronal development in the adult more or less recapitulate the processes prevalent in the fetus. Extrinsic and activity-dependent regulation, in contrast, are likely to set adult neurogenesis apart from fetal neurogenesis. In addition, differences between the two neurogenic regions might exist with regard to the contribution of individual regulatory factors.

Function

The fact that adult neurogenesis is the exception and not the rule has raised the question of why dentate gyrus and olfactory bulb in contrast to all other brain regions not only accommodate the integration of new neurons but actually call for it in a function-dependent manner. In both systems, adult neurogenesis does not add bulk numbers of new neurons. Rather, the contribution of adult neurogenesis might be qualitative and allow a possibly subtle modification of the existing neuronal network. In the hippocampus, adult neurogenesis allows the activity-dependent optimization of the mossy fiber connection between the dentate gyrus and CA3. Integration is triggered during a phase of increased synaptic plasticity and leads to a cumulative adaptation of the network to a level of complexity and novelty. Alternative hypotheses that mostly suggest a transient functional contribution of the immature neurons in the sense of a direct contribution to memory formation are questioned by the lacking evidence for substantial neuronal turnover and the scarcity of new neurons. Mechanistically, it has been proposed that the new neurons help the hippocampal network to avoid catastrophic interference in the dentate gyrus, that is, the destruction of codings for older information by incoming new information.

Similarly, adult olfactory neurogenesis has been brought into connection with novelty detection, here the coding of novel odors. It seems that new neurons are more sensitive to novel odors and might help to adjust the network to the processing of a new range of olfactory stimuli.

A general principle might be that adult neurogenesis is used in bottleneck situations, where the insertion of new neurons is strategic and allows the incremental optimization of a network that in general tends to remain as lean as possible. Networks in nonneurogenic areas might not exist under this priority of size limitation.

Medical Relevance

Adult neurogenesis might be of medical relevance either because a failure of adult neurogenesis might contribute to brain disease or because a reactive increase in adult neurogenesis might be part of endogenous regenerative attempts. A role for adult neurogenesis in brain pathology, among others, is discussed for temporal lobe epilepsy, major depression, and dementias. In temporal lobe epilepsy, cell proliferation is massively increased, many new neurons are found in ectopic locations, and aberrant projections from the new neurons might support the chronification of the disease. In the context of major depression, a reduction in hippocampal volume has been noted (that is, however, much larger to be explained by a lacking contribution of adult neurogenesis) and all antidepressants have been found to increase adult hippocampal neurogenesis. As the most extreme position in the context, it has been suggested that antidepressants might even require neurogenesis for their action. Dementias might at least partly be explained by a reduction in the contribution that new neurons make to hippocampal function – a concept that is complicated by the fact that the functional relevance of adult neurogenesis itself is not yet fully understood. As yet, no similar concrete links to pathology have been proposed for neurogenesis in the olfactory bulb.

Numerous pathological stimuli, most notably trauma and ischemia, induce the proliferation of precursor cells and in many cases net adult neurogenesis. These effects appear to be indirect because they are found even in the absence of direct damage to the dentate gyrus or SVZ. In addition, after middle cerebral artery occlusion in rats, precursor cells migrated from the SVZ into the striatum and differentiated into local neurons. The massive infusion of growth factors supported the replacement of ischemic CA1 neurons by SVZ precursor cells. Phototoxically induced circumscript cell death in the cortex also led to the replacement of neurons.

Envisioned therapeutic strategies that are based on the transplantation of neurons that have been generated from neural precursor cells benefit from the knowledge accumulating from research on adult neurogenesis, because here nature exemplifies how neuronal development is possible under the conditions of the otherwise nonneurogenic adult brain.

Medical relevance also depends on the occurrence of adult neurogenesis in humans. Adult hippocampal neurogenesis has been demonstrated for humans and nonhuman primates. There is evidence for neurogenesis in the adult olfactory bulb of nonhuman primates but not yet unambiguously of humans. Precursor cells, however, have been isolated from ventricle biopsies and resection specimens from the adult human hippocampus, as well as postmortem brain tissue of infants.

See also: Axonal Injury: Neuronal Responses; Cell Replacement Therapy: Parkinson's Disease; Cell Replacement Therapy for Huntington's Disease; Exercise: Optimizing Function and Survival at the Cellular Level; Lipids and Membranes in Brain Aging; Neural Stem Cells and CNS Diseases; Neural Stem Cells: Adult Neurogenesis; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Neuronal Plasticity after Cortical Damage; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Glomeruli: Structure and Circuitry; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Altman J and Das GD (1965) Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *Journal of Comparative Neurology* 124: 319–335.

Alvarez-Buylla A, Garcia-Verdugo JM, and Tramontin AD (2001) A unified hypothesis on the lineage of neural stem cells. *Nature Reviews Neuroscience* 2: 287–293.

Alvarez-Buylla A and Lim DA (2004) For the long run: Maintaining germinal niches in the adult brain. *Neuron* 4: 683–686.

Cameron HA, Woolley CS, McEwen BS, and Gould E (1993) Differentiation of newly born neurons and glia in the dentate gyrus of the adult rat. *Neuroscience* 56: 337–344.

Emsley JG, Mitchell BD, Kempermann G, and Macklis JD (2005) Adult neurogenesis and repair of the adult CNS with neural progenitors, precursors, and stem cells. *Progress in Neurobiology* 75: 321–341.

Eriksson PS, Perfilieva E, Bjork-Eriksson T, et al. (1998) Neurogenesis in the adult human hippocampus. *Nature Medicine* 4: 1313–1317.

Gould E and Tanapat P (1999) Stress and hippocampal neurogenesis. *Biological Psychiatry* 46: 1472–1479.

Kempermann G (2006) *Adult Neurogenesis – Stem Cells and Neuronal Development in the Adult Brain*. New York: Oxford University Press.

Kempermann G, Jessberger S, Steiner B, and Kronenberg G (2004) Milestones of neuronal development in the adult hippocampus. *Trends in Neurosciences* 27: 447–452.

Kempermann G, Kuhn HG, and Gage FH (1997) More hippocampal neurons in adult mice living in an enriched environment. *Nature* 386: 493–495.

Lie DC, Song H, Colamarino SA, Ming GL, and Gage FH (2004) Neurogenesis in the adult brain: New strategies for central nervous system diseases. *Annual Review of Pharmacology and Toxicology* 44: 399–421.

Lledo PM, Alonso M, and Grubb MS (2006) Adult neurogenesis and functional plasticity in neuronal circuits. *Nature Reviews Neuroscience* 7: 179–193.

Lledo PM and Saghatelyan A (2005) Integrating new neurons into the adult olfactory bulb: Joining the network, life–death decisions, and the effects of sensory experience. *Trends in Neurosciences* 28: 248–254.

Magavi SS, Leavitt BR, and Macklis JD (2000) Induction of neurogenesis in the neocortex of adult mice. *Nature* 405: 951–955.

Nottebohm F (2004) The road we travelled: Discovery, choreography, and significance of brain replaceable neurons. *Annals of the New York Academy of Sciences* 1016: 628–658.

Palmer TD, Ray J, and Gage FH (1995) FGF-2-responsive neuronal progenitors reside in proliferative and quiescent regions of the adult rodent brain. *Molecular and Cellular Neurosciences* 6: 474–486.

Neurohypophyseal System

C D Sladek, University of Colorado at Denver and Health Science Center, Aurora, CO, USA
W E Armstrong, University of Tennessee Health Science Center, Memphis, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Hormone Structure, Biosynthesis, and Secretion

Peptide and Prohormone Structure and Processing

Vasopressin (VP) and oxytocin (OT) are nine-amino-acid peptides that differ from each other by only two amino acids (Figure 1). Each includes a disulfide ring formed between cysteine residues. They are synthesized as part of larger prohormone molecules (Figure 1). Following translation, the prohormone molecule undergoes initial glycosylation in the endoplasmic reticulum and is packaged into large (150–200 nm) secretory granules in the Golgi complex. During transport of these secretory granules to the axon terminal in the neurohypophysis, the prohormone molecule is enzymatically cleaved into the peptide, its associated neurophysin, and, in the case of VP, a C-terminus glycopeptide molecule. Secretion of the hormones occurs by calcium-dependent exocytosis induced by depolarization of the axon terminal. During exocytosis, the contents of the secretory granule are discharged into the extracellular fluid surrounding the nerve terminal and diffuse into the surrounding capillaries for distribution by the bloodstream to distant target organs. Although neurophysin and the glycopeptide (as well as VP and OT) are detectable in the blood, no hormonal action of these molecules has been identified.

Genes and mRNA

VP and OT are produced only by mammals, but they differ by one amino acid from the closely related peptide, vasotocin, which is produced in nonmammalian vertebrates. The genes for VP and OT evolved from the vasotocin gene by a gene duplication event. In all species studied, the two genes are linked on the same chromosome in opposite transcriptional directions. The intergenic sequence is important for regulation of gene expression. Each gene is composed of three exons. Exon 1 codes for the N-terminal portion of the preprohormone molecule that includes the signal peptide that directs the molecule from the ribosome to the endoplasmic reticulum, the VP or OT peptide, and the N-terminal portion of the neurophysin molecule. Exon 2 codes the central component of neurophysin and shows a high degree of homology between VP and OT. Exon 3 codes the C-terminal portion of neurophysin and, in the case of VP, the glycopeptide. Mutations in the VP gene with resultant loss of VP production result in neurogenic diabetes insipidus (DI). A rat with a single base deletion that disrupts the translation stop codon for VP mRNA was discovered in Brattleboro, Vermont, in 1962. This animal exhibits polydipsia and polyuria, the hallmark symptoms of VP deficiency, and it has been a very useful animal model for studying VP secretion and action. Multiple mutations in the human gene have also been identified in families with hereditary neurogenic DI. These mutations generally lead to abnormalities in folding or packaging of the prohormone.

Structure of the Neurohypophyseal System

Hypothalamic Nuclei (Supraoptic/Paraventricular/Accessory)

VP and OT are synthesized by neurons localized in the supraoptic (SON) and paraventricular nuclei (PVN) of the hypothalamus as well as in several small groups of accessory magnocellular neurons (Figures 2 and 3). The axons from these neurons form the hypothalamo-neurohypophyseal tract (Figure 3(c)) and terminate in the posterior lobe of the pituitary or neurohypophysis. Due to dilution of the hormones in the blood prior to their delivery to their target tissues, they must be synthesized and released in large quantities and this is reflected by the large size of the VP and OT neurons (25–30 μm in diameter). The hormones are synthesized in the cell bodies of these 'magnocellular' neurons and the secretory granules are transported by fast axonal transport to the nerve terminal in posterior pituitary. Large clusters of secretory vesicles are visible as Herring bodies (5–15 μm) in the nerve terminals (Figure 4(a)). These contain large stores of hormone sufficient, in the case of VP in the rat posterior pituitary, to maintain hormone release for 5–7 days of water deprivation.

Although all mammals, including humans, have a mix of VP and OT neurons in both the PVN and the SON, within each nucleus the VP and OT neurons cluster in specific regions. In rat SON, OT neurons predominate in the dorsal, lateral, and rostral regions of the nucleus, and in many species a large cluster of magnocellular VP neurons in the lateral wing characterizes the PVN (Figure 3). Although in most cases VP and OT are produced in separate neurons, many neurons can synthesize both peptides, and under conditions of enhanced hormone release, co-localization of

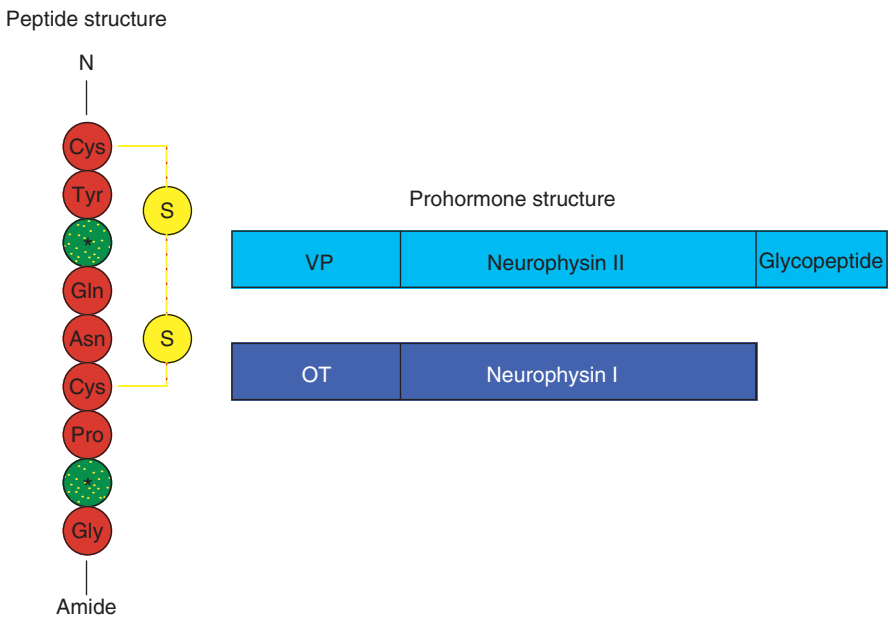

Figure 1 The peptide and prohormone structures of the mammalian neurohypophyseal hormones vasopressin (VP) and oxytocin (OT). They are similar to the nonmammalian hormone vasotocin in that all three are nine-amino-acid peptides with a disulfide ring differing only by the amino acids in positions 3 and 8 (asterisk). Position 3 is Ile in vasotocin and oxytocin but Phe in vasopressin, whereas position 8 is Arg in vasotocin and vasopressin but Leu in oxytocin. An exception is the pig family, in which vasopressin has Lys at position 8. The prohormones for VP and OT both include the peptide sequence at the N-terminus followed by neurophysin. After packaging into the secretory granule, the prohormone is cleaved at the positions indicated by the vertical bars.

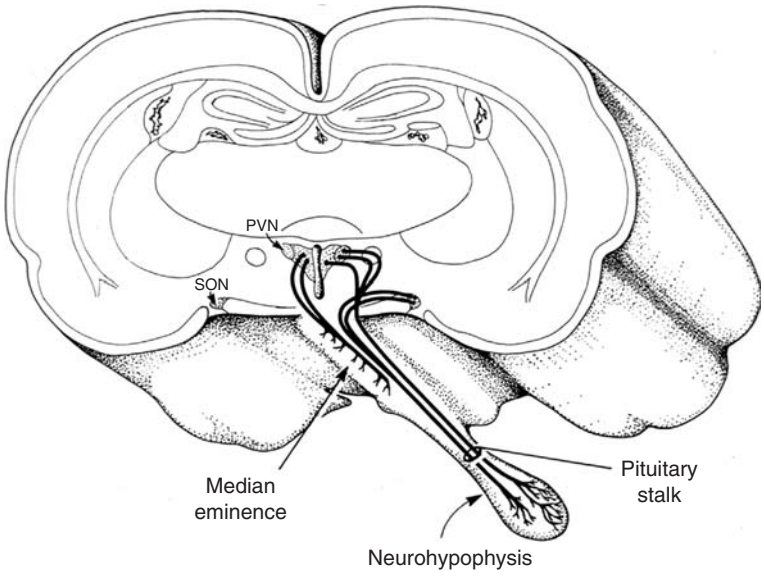

Figure 2 Diagram of the hypothalamo-neurohypophyseal system in the rat brain. The rat brain is cut away at the level of the optic chiasm to show the location of the SON and PVN. On the right side is shown the axonal projections from the magnocellular VP and OT neurons to the neurohypophysis. On the left side is shown the axonal projections from the parvocellular VP/corticotropin neurons in the PVN to the portal vessels in the median eminence that carry hormones to the anterior pituitary (not shown). Note that these latter axon terminals are visible in the external zone of the median eminence in **Figure 3(c)**. From Hatton GI (2002) Glial–neuronal interactions in the mammalian brain. *Advances in Physiology Education* 26: 225–237.

the peptides increases (Figure 3(b)). Both VP and OT are also synthesized in small (parvocellular) neurons located in PVN and, in the case of VP, other brain regions, such as the suprachiasmatic nucleus of the hypothalamus and the medial amygdala. These parvocellular neurons project to regions of the brain other than the neurohypophysis and are not involved in producing VP and OT for release

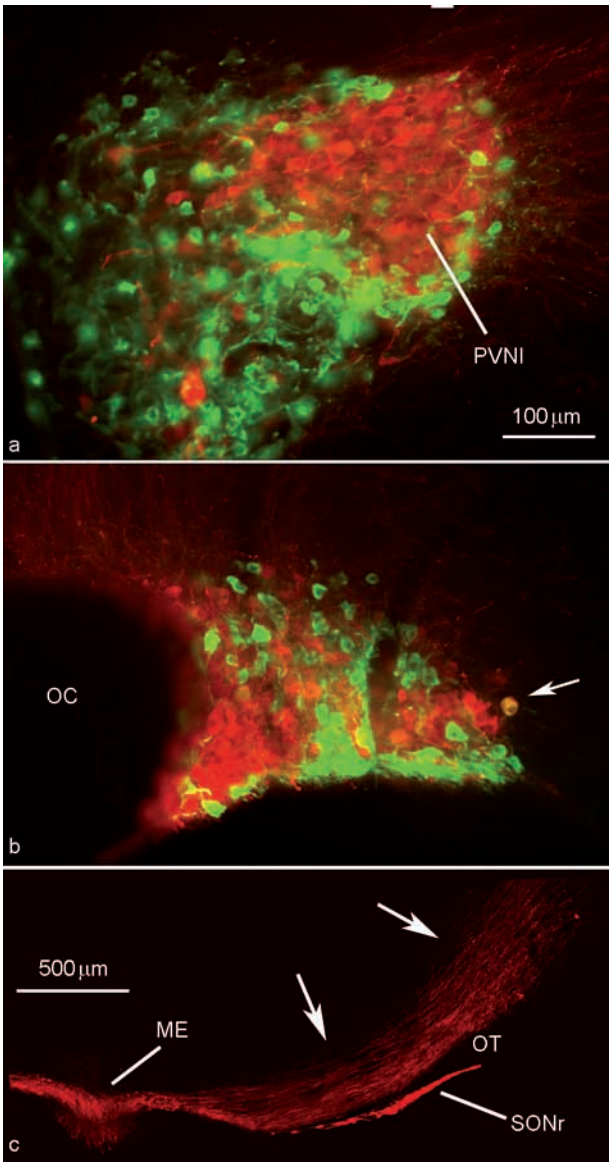

Figure 3 Location of optic tract (OT) and VP neurons and the hypothalamo-neurohypophyseal tract in coronal sections of rat brain. (a) Immunofluorescence of VP- and OT-associated neurophysins (NPs) in the PVN. The VP–NP neurons (red) are clustered in the lateral subnucleus (PVNl), whereas the OT–NP neurons (green) are more medial. (b) VP–NP (red) and OT–NP neurons (green) in the principal portion of the SON, adjacent to the optic chiasm (OC). Note the orange neuron (arrow), indicating double labeling for both OT–NP and VP–NP. Scale as in (a). (c) VP–NP-positive axons in the hypothalamo-neurohypophyseal tract in the hypothalamus, coronal section. The axons make a wide lateral arch (arrows) before turning medially to enter the median eminence (ME). SONr, retrochiasmatic portion of the SON. Antibodies courtesy of Dr. Harold Gainer, National Institutes of Health.

into the general circulation. Their involvement in generation of circadian rhythms, responses to stress, and maternal behavior is discussed elsewhere in this encyclopedia.

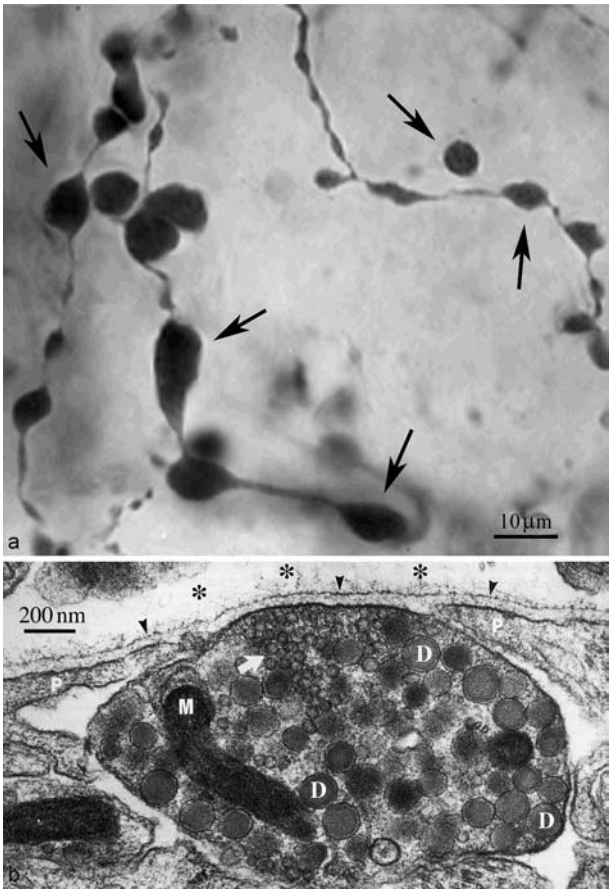

Figure 4 Neurosecretory axons and terminals. (a) Axons in the neural lobe anterogradely labeled with biotinylated dextran applied in the SON and reacted using the avidin–biotin peroxidase method. Note the varicose, beaded morphology, with many large swellings (arrows) indicative of Herring bodies. (b) Electron micrograph from a rapidly frozen neural lobe showing a neurosecretory terminal. The terminal is directly apposed to the basal lamina (arrowheads) of the perivascular space (asterisks). The terminal is filled with dense core vesicles (D), which house VP and OT but also contain many microvesicles (white arrow). Pituicyte processes are also visible near the terminal. M, mitochondria. (a) Reproduced from Armstrong WE (2004) Hypothalamic supraoptic and paraventricular nuclei. In: Paxinos G (ed.) *The Rat Nervous System: A Handbook for Neuroscientists,* 3rd edn., pp. 369–388. Sydney, NSW: Academic Press, with permission from Elsevier. (b) From Tian M, Reger JF, and Armstrong WE (1991) Electron microscopic and immunocytochemical study of rapidly frozen, freeze-substituted neural lobes of rats. *Journal of Neurocytology* 20: 79–96.

Glial/Neuronal Interactions

The VP and OT neurons are closely associated with glial cells, and plastic changes in the neuronal/glial relationship associated with physiological stimulation of VP and OT release have been well documented in both SON and the neurohypophysis. A ventral glial lamina separates the SON from the ventral surface of the brain. This region contains the cell bodies of astrocytes that under basal conditions send long processes dorsally into SON, separating adjacent

neurons from each other and wrapping around the dendrites. During lactation or following dehydration (e.g., increased OT and VP secretion, respectively), the glial processes withdraw, allowing increased apposition of somatic membranes and formation of multiple synapses between dendrites. These conditions allow for enhanced neuronal activity by removing inhibitory agents released by glia as well as enhancing the potential for synchronous activation of neurons by afferents synapsing on multiple dendrites. The astrocytes in the posterior pituitary (called pituicytes) demonstrate similar plasticity. Under conditions stimulating hormone release, withdrawal of pituicyte processes is observed, resulting in increased exposure of the nerve terminals to the basal lamina of the perivascular space. This morphological change improves access of released hormone to the capillaries and may remove inhibitory agents such as taurine released by the pituicytes.

Regulation of Hormone Secretion

Vasopressin

Due to its vasoconstrictive actions on peripheral vessels, and its ability to cause water reabsorption from the renal filtrate, VP directly modifies the osmolality of the extracellular fluid (ECF) as well as blood volume and pressure. These parameters in turn regulate its secretion from the neurohypophysis.

Osmotic regulation VP release is exquisitely sensitive to changes in osmolality of the plasma or ECF. A 1% increase in osmolality increases plasma VP to a level sufficient to elicit detectable antidiuresis. This is the result of an elaborate series of osmoreceptive elements strategically situated to monitor ECF osmolality at different sites in the body. Osmoreceptive elements are located in the mesenteric and hepatic portal vasculature. These receptors monitor osmolality of the blood draining the intestine and provide an early indication of water, electrolyte, and nutrient intake. Osmoreceptors are also located in several regions of the brain that lack a blood–brain barrier (i.e., subfornical organ (SFO) and organum vasculosum of lamina terminalis (OVLT)). These receptors are positioned to monitor plasma osmolality unaffected by transport limitations of the blood–brain barrier. Finally, the magnocellular neurons of the SON and PVN are also osmosensitive, providing a final site for regulation of VP secretion by osmolality.

In 1947, Verney postulated the existence of 'osmoreceptors.' His pioneering work and subsequent studies by many others have led to the consensus that neurons in the SON and PVN as well as in the OVLT and SFO function as osmoreceptors by monitoring changes in their cell volume induced by fluctuations in the concentration of nonpermeable solutes (e.g., Na^+ and sucrose, but not urea) in the surrounding ECF. Stretch-inactivated, nonselective ion channels have been functionally demonstrated in these regions and shown to underlie the osmosensitivity of SON and OVLT neurons. Modulation of these channels by angiotensin II, cholecystokinin, and neurotensin underlies the ability of these neuropeptides to modulate osmotic regulation of VP secretion. These channels also contribute to the observation that increases in ECF Na^+ can induce VP release independent of changes in ECF osmolality because the permeability of these channels to Na^+ is enhanced by increases in ECF Na^+. The ability of gadolinium to block these channels suggested that they are mechanosensitive ion channels, and evidence that an N-terminal variant of the Trpv1 channel is expressed in SON neurons and is required for osmosensation suggests a molecular identity of the osmosensor.

Although the magnocellular neurons are osmoreceptive, several additional mechanisms contribute to the ability of the intact organism to tightly regulate ECF osmolality. First, excitatory afferents from osmosensitive neurons in the SFO and OVLT innervate the magnocellular neurons, enhancing their endogenous responses to changes in osmolality. Second, similarly, afferents carrying information from the peripheral osmosensors also influence VP release. Third, taurine release from astrocytes in SON may be regulated by cell volume-sensitive mechanisms. Taurine released by glial cells under hypotonic conditions activates inhibitory glycine receptors on magnocellular neurons and thus may contribute to inhibition of VP and OT neurons. Fourth, the withdrawal of astrocytic processes during hypertonicity removes this inhibitory influence and increases exposure of the neurons to excitatory neurotransmitters from osmosensitive afferents. Finally, additional sodium sensors may participate as well. SON neurons have been shown to express Na_x, a sodium channel that is sensitive to the ECF sodium concentration. This channel is also expressed in SFO and OVLT, and mice deficient in the gene ingest greater quantities of salt.

Pressure/volume regulation As a potent vasoconstrictor agent, VP is important for preventing decreases in blood pressure, and as a potent antidiuretic agent, it is important for maintaining blood volume. Thus, it is appropriate that information about blood pressure and volume are important regulators of VP secretion. Although larger percentage decreases in blood pressure or volume are required to achieve comparable increases in plasma VP compared to increases in

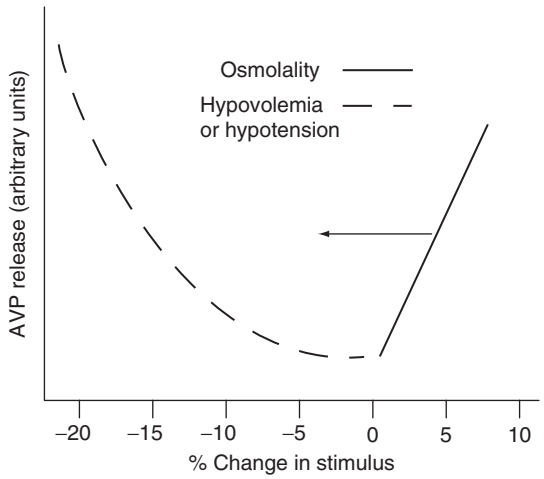

Figure 5 Comparison of the VP response to decreases in blood pressure or volume versus increases in plasma osmolality. VP release is linearly related to increases in osmotic pressure once a threshold is exceeded while it is exponentially stimulated by decreases in blood volume or pressure. Thus, it is more sensitive to changes in osmolality (e.g., a smaller percentage change from baseline is required to induce detectable increase in plasma VP) but greater overall increases in plasma VP are induced by decreases in blood pressure or volume. The arrow indicates that the osmotic relationship is displaced to the left in the presence of hypovolemia/hypotension such that VP release in response to osmotic stimulation is enhanced. Based on original data published in Dunn FL, Brennan TJ, Nelson AE, and Robertson GL (1973) The role of blood osmolality and volume in regulating vasopressin secretion in the rat. *Journal of Clinical Investigation* 52: 3212–3219, and modified with permission from Baylis PH (1987) Osmoregulation and control of vasopressin secretion in healthy humans. *American Journal of Physiology* 253: R671–R678.

osmolality, decreases in blood pressure result in much greater increases in plasma VP (Figure 5).

Arterial blood pressure is monitored by stretch receptors in the carotid sinus and blood volume is monitored by stretch receptors in the atria and pulmonary vessels. This information is transmitted to neurons in the nucleus tractus solitarius of the brain stem via cranial nerves IX and X (glossopharyngeal and vagus, respectively) and in turn to the SON and PVN via multisynaptic pathways. The A1 catecholamine pathway is activated in response to a decrease in blood pressure or volume. This pathway directly innervates the SON and PVN, and it uses ATP, norepinephrine, and neuropeptide Y as excitatory neurotransmitters. Increases in blood pressure activate a different and only partially characterized pathway involving the parabrachial nucleus and the diagonal band of Broca, with ultimate inhibition of VP neurons by GABAergic neurons located in the perinuclear zone adjacent to SON.

Other physiological regulators A large number of hormones and cytokines alter VP release. Since many peripherally generated hormones or cytokines cannot cross the blood–brain barrier, these agents may act indirectly by activating receptors on circumventricular organs (e.g., angiotensin II and interleukin-1 and -6). In contrast, specific transport mechanisms allow insulin to access receptors in the brain, and the lipid-soluble steroid hormones can freely permeate the blood–brain barrier. The VP neurons express insulin, estrogen, and, during chronic hyponatremia, glucocorticoid receptors. Understanding of the role of these agents in regulating VP release is incomplete, but insulin may facilitate glucose uptake by the VP neurons since glucose becomes an effective osmotic agent for stimulation of VP release in the insulin deficiency of uncontrolled diabetes mellitus.

Electrophysiology of VP neurons The pioneering observations of Harris and Douglas revealed that Ca^{2+}-dependent, stimulus-secretion coupling in the neurohypophysis resembled that observed at the neuromuscular junction. Electrophysiological studies later affirmed that it is the pattern and rate of action potential activity originating in SON and PVN neurons that is the primary determinant for peripheral VP and OT release. Action potentials propagate toward and depolarize neurohypophyseal terminals, increasing Ca^{2+} influx through voltage-gated channels to trigger the exocytosis of hormone-containing vesicles near fenestrated capillaries. A common theme for both cell types is that the patterns of electrical activity adopted lead to efficient hormone release.

Hypotension, hypovolemia, and hyperosmolality all trigger an electrical response from VP neurons referred to as phasic bursting (Figure 6). With even moderate stimulation, such as 6 h of dehydration, the majority of VP neurons adopt this pattern. Further stimulation of phasic neurons increases firing rate within the burst, and often bursts lengthen. VP neurons fire asynchronously, leading to smooth increases in VP release.

The efficiency (hormone release per action potential) of phasic bursting for VP release can be directly demonstrated using the action potential distribution of recorded bursts to time electrical stimulation of neurohypophysial axons. Phasic bursting stimuli at the same mean frequency as continuous trains release far more VP, with a maximal efficiency of ~13 Hz. This rate corresponds to that achieved within VP bursts during stimulation. Over extended periods, efficiency is related to both the high frequency of discharge at the beginning of each burst (facilitation) and recovery from fatigue during interburst intervals. Recovery during interburst intervals maximizes Ca^{2+} uptake into the terminals for successive bursts.

Although activity is instigated and modified by peripheral signals promoting VP release, phasic bursting is

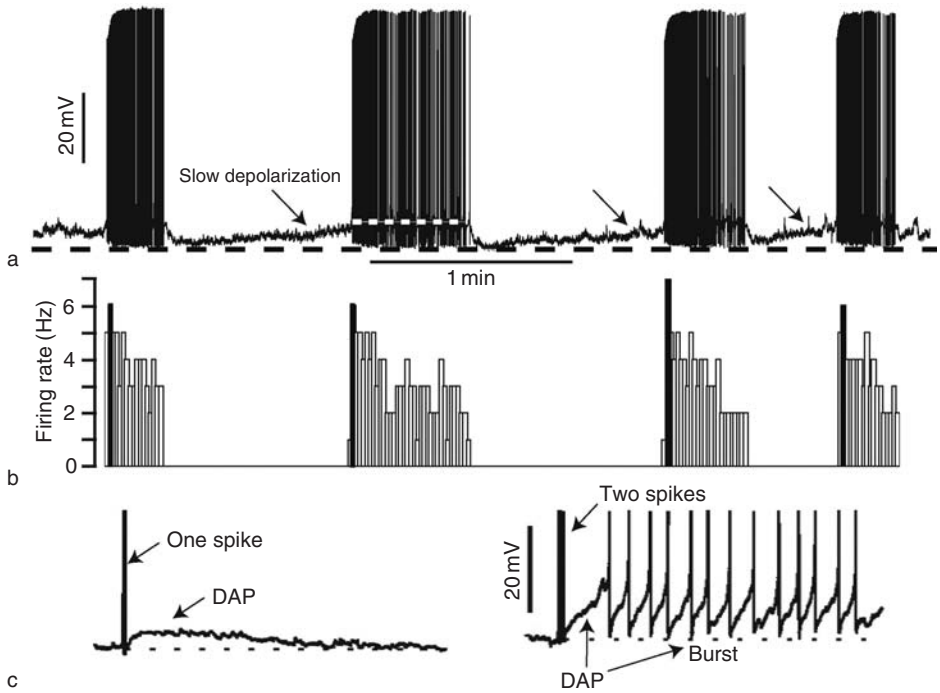

Figure 6 Phasic bursting VP neuron. (a) Whole cell recording of a phasically bursting VP neuron recorded from a rat hypothalamic slice *in vitro*. The firing pattern is similar to that observed *in vivo*. Each burst of action potentials rides on a plateau potential (white dotted line) and is preceded by a slow depolarization (arrows). Black dotted line indicates the hyperpolarization of the membrane potential following each burst. (b) Rate meter histogram of the fire rate of the neuron shown in (a). The peak firing rate (solid black vertical line) occurs near burst onset. (c) The spike depolarizing afterpotential (DAP) of a VP neuron. (Left) The DAP following a single spike. (Right) After two spikes, the DAP summates and underlies a burst as the plateau potential. Adapted from Roper P, Callaway J, and Armstrong W (2004) Burst initiation and termination in phasic vasopressin cells of the rat supraoptic nucleus: A combined mathematical, electrical, and calcium fluorescence study. *Journal of Neuroscience* 24(20): 4818–4831.

an intrinsic, voltage-dependent property of the VP neuron. Thus, synaptic or osmotic stimuli must bring the VP neuron to a membrane potential of −60 to −50 mV before bursting can commence. With this depolarization, the Ca^{2+} influx from voltage-gated Ca^{2+} channels opened by action potentials triggers a Ca^{2+}- and voltage-dependent depolarizing afterpotential (DAP) (Figure 6(c)). The DAP summates with repetitive firing to form a plateau potential on which a long (10–60 s) burst discharge is maintained. The current underlying the DAP may result from either the reduction in a resting K^+ conductance or an increased conductance through nonselective cationic channels. SON neurons contain many other conductances that shape the activity of the burst, the best studied of which are voltage-gated Ca^{2+} currents; Ca^{2+}-dependent K^+ currents, which gate firing rate and shape burst length; voltage-gated K^+ currents, which influence firing rate and spike width; and a hyperpolarization activated, nonselective cation current (I_H) that contributes to plateau level.

VP neurons also synthesize the opiate peptide dynorphin and package it within VP dense core vesicles. Both dynorphin and VP are released locally (i.e., from the soma or dendrites) during neuronal activity, and both modify burst parameters through feedback. Dynorphin acts on kappa opioid receptors on VP neurons and serves to terminate bursts by direct inhibition of the DAP/plateau potential. VP gates burst firing rate through unknown mechanisms. VP receptors V1a and possibly V1b (or related variant) are located on the membrane of VP neurons, and their activation increases $[Ca^{2+}]_i$ and modulates VP neuronal activity. The action of VP depends on the neuron's current activity and serves to reinforce the most efficient firing rate for VP release.

Oxytocin

OT causes contraction of smooth muscle in the uterus and mammary glands, resulting in expulsion of the fetus during birth and milk ejection during lactation. Its release from the neurohypophysis is regulated by nerve impulses initiated by stretch of the cervix during parturition and suckling of the nipples by the infant. It also participates in fluid and electrolyte homeostasis, ejaculation, and possibly appetite regulation in males and nonpregnant or nonlactating females.

Pregnancy and parturition During pregnancy, both uterine and hypothalamic mechanisms participate in preventing premature, OT-induced, uterine contractions. Several of these mechanisms are driven by the high circulating concentrations of pregnancy-associated hormones, including estrogen, progesterone, and relaxin. Just prior to parturition, this brake on OT stimulation of uterine contractions is reversed by the regression of progesterone and relaxin secretion. In the uterus, progesterone withdrawal during late pregnancy allows estrogen to rapidly induce upregulation of OT receptor expression and tissue responsiveness. In the hypothalamus, progesterone is converted to allopregnanalone, which acts on $GABA_A$ receptors to potentiate GABAergic inhibition of OT neurons. Progesterone withdrawal at the end of pregnancy reduces this inhibitory influence on OT secretion. Progesterone-independent, opioid inhibitory mechanisms also restrain OT release during pregnancy. The site and source of this opioid inhibition remain to be completely delineated, but multiple mechanisms may participate. Activation of kappa opioid receptors on OT nerve terminals in the neurohypophysis suppresses activity-induced OT release, and activation of either mu or kappa opioid receptors inhibits the firing of OT neurons. Neurohypophyseal opioid inhibition is upregulated in midpregnancy and then downregulated near parturition. This change may reflect relaxin secretion by the corpus luteum because relaxin stimulates opioid release from the adrenal gland, which in turn can access the OT nerve in the neurohypophysis to inhibit activity-induced OT release. Activation of opioid receptors in the hypothalamus may reflect autorelease of enkephalin and/or dynorphin by the OT neurons, release of beta-endorphin by afferents from the arcuate nucleus, and/or release of enkephalin by brain stem afferents.

At parturition, stretch of the cervix activates vagal and spinal afferents to the NTS in the brain stem. In turn, the A_2 noradrenergic neurons in NTS as well as non-noradrenergic NTS neurons activate the OT neurons, inducing high-frequency bursts of action potentials that efficiently release OT. Since active OT receptors are present in the uterus at this stage of pregnancy, OT release sets up a positive feedback situation called the Ferguson reflex, in which cervical stretch stimulates OT release, which in turn stimulates uterine contraction and results in additional cervical stimulation.

Norepinephrine excites OT neurons via activation of α_1 noradrenergic receptors. Since 60% of the brain stem neurons activated during parturition are noradrenergic, this is undoubtedly an important component of the system. However, other neurotransmitters (perhaps including cotransmitters and/or neuropeptides in the catecholamine neurons such as ATP) can take over in the absence of norepinephrine since α_1 adrenergic antagonists injected into SON do not delay or impede spontaneous parturition.

Suckling OT is essential for successful lactation because milk secreted by the mammary glands must be pushed into the duct system by contraction of the smooth muscles surrounding the mammary glands. This process is called milk 'let-down' or ejection and is induced by OT. The pups of OT-deficient mice die of malnutrition because milk is not available in the mammary ducts for suckling. Suckling in turn induces OT secretion from the neurohypophysis by activating spinal afferents that synapse on noradrenergic neurons in the NTS. Suckling increases norepinephrine release in SON, and disruption of hypothalamic noradrenergic transmission prevents systemic OT release during lactation. Additional neurotransmitters required for suckling-induced OT release include histamine and glutamate. Both of these agents may be important mediators of noradrenergic stimulation. Histaminergic innervation of the OT neurons arises from the tuberomammillary nucleus. Histamine alters the sensitivity of SON neurons to norepinephrine, and blockade of H1 histamine receptors prevents suckling-induced release of OT. Glutamatergic afferents to the OT neurons include the osmoreceptive pathways described previously as well as local glutamatergic neurons in the perinuclear regions surrounding SON. Central blockade of either glutamate (AMPA) receptors or α_1 adrenergic receptors blocks milk ejection, but the precise location of antagonist actions is unknown. Norepinephrine has been shown to activate magnocellular neurons directly as well as via presynaptic actions on glutamatergic terminals. Thus, glutamate may relay suckling-induced noradrenergic signals to the OT neurons.

Milk let-down can also occur in response to the sight, sound, and/or smell of the infant. It is likely that this effect reflects innervation of the OT neurons by afferents from the amygdala and olfactory bulb. Reciprocal oxytocinergic innervation of these brain regions has been implicated in development of maternal behaviors required for successful nurturing of the young.

Homeostatic regulation of OT release In addition to the specialized reflexes that result in bolus release of OT during parturition and suckling, OT is released with VP in response to hypertonicity and hypovolemia. Thus, the pathways described previously for osmotic and hemodynamic regulation of VP secretion also innervate the magnocellular OT neurons in SON and PVN. This is physiologically relevant because OT

induces natriuresis (renal sodium excretion) and therefore contributes to maintenance of fluid and electrolyte homeostasis. The natriuretic action of OT reflects both direct actions of OT on the kidney and its ability to stimulate release of atrial natriuretic peptide from the heart. During pregnancy, relaxin activates osmosensitive neurons in SFO and OVLT. This, in conjunction with high circulating estrogen, is thought to result in resetting of the threshold for osmotic stimulation of thirst as well as VP and OT secretion. The resultant increase in OT-induced natriuresis and VP-induced water retention contributes to the plasma volume expansion required to meet the increased vascular needs of the fetus.

OT release is also stimulated by peripheral administration of cholecystokinin (CCK) or CCK released by gastric distension. This action of peripheral CCK is mediated via activation of brain stem pathways perhaps involving the A_2 noradrenergic pathways and is highly selective for the OT neurons. It has been used as a tool for experimental identification of OT neurons but may also be indicative of a role for OT in appetite regulation.

During pregnancy, OT secretion is relatively insensitive to the homeostatic signals (e.g., osmolality and gastric distention) that stimulate its secretion in males and in nonpregnant females. During lactation, the system is able to respond effectively to both types of signals due to the difference in the response pattern: Homeostatic signals increase the firing rate of OT neurons but maintain a relatively uniform firing frequency which induces a sustained elevation in OT secretion, whereas suckling evokes synchronized, high-frequency burst activity inducing the bolus pulses of OT secretion required for milk ejection.

Ejaculation Plasma OT is increased during sexual arousal and orgasm in both men and women. In males, OT contracts the reproductive tracts, which appears to aid in sperm release and transport. The central pathways for this stimulus for OT release have not been fully elucidated, although it is known that electrical stimulation of the rat dorsal penile nerve excites PVN and SON OT cells, and penile erection is attenuated by lesions of the catecholaminergic inputs to the hypothalamus.

Electrophysiology of OT neurons OT neurons show two different modes of enhanced activity, depending on the stimulus. In response to many of the same stimuli releasing VP (e.g., hypotension and hyperosmolality), OT release is associated with a continuous firing pattern of OT neurons, the rate of which reflects stimulus intensity. To these sustained stimuli related to fluid balance, OT's capacity as a natriuretic agent may be more relevant than its actions on smooth muscle.

In most mammalian species, OT release during suckling is episodic, and it is during lactation in the rat that OT neuronal activity has been most intensely studied. In response to suckling, OT neurons fire in a bursting mode that is quite different from the phasic bursting of VP neurons (Figure 7). First, OT bursts are synchronous across OT neurons in the PVN and SON, and this firing pattern leads to pulsatile OT release that maximizes discrete milk ejections at the mammary gland. Second, OT bursts are much shorter and more intense than those of VP neurons, firing for ~4 s at rates averaging 50 Hz. Third, the interburst intervals are long, from 5 to 15 min. These intervals are relatively constant within a period of suckling.

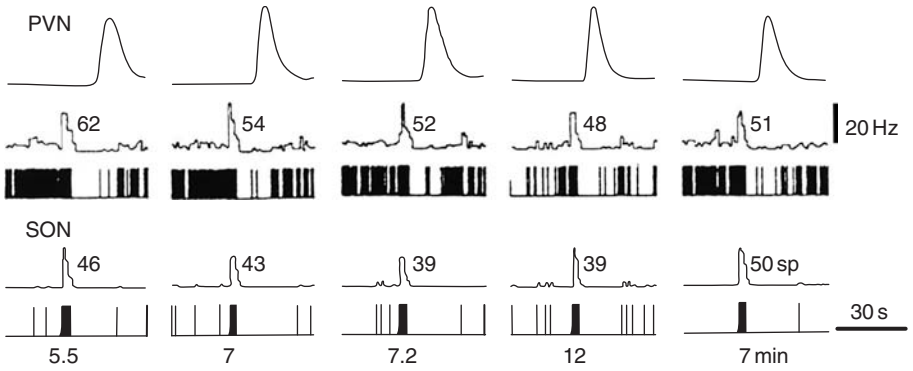

Figure 7 OT neurons firing during lactation. The top trace shows the intramammary pressure changes associated with synchronous bursts of OT neurons in a lactating female rat, and beneath this are two sets of traces for neuronal firing recorded extracellularly – one cell from the PVN and one from a simultaneously recorded SON neuron. For each neuron, the rate meter (in Hz) and a pen recording registering each action potential are shown. The numbers associated with the rate meters indicate the number of spikes in a milk-ejection burst. The numbers at the bottom of the figure indicate the delay in minutes since the previous burst. Adapted from Richard P, Moos F, and Freund-Mercier M-J (1988) Bursting activity in oxytocin cells. In: Leng G (ed.) *Pulsatility in Neuroendocrine Systems*, pp. 75–97. Boca Raton, FL: CRC Press.

The bolus release of OT strongly increases intramammary pressure, signaling to the attached pups that milk is being extruded, and thereby stimulating a stereotypic 'stretch reflex' as each pup sucks vigorously to remove milk. Although recording studies are uncommon, bursting can appear during parturition, when OT promotes uterine contraction. However, these bursts may be triggered by irregular periods of cervical dilation, a strong afferent stimulus for OT release.

Studies *in vivo* and *in vitro* suggest that excitatory glutamatergic input is critical for milk ejection and thus burst firing. As mentioned previously, blockade of AMPA receptors blocks milk ejection *in vivo*. *In vitro*, OT cell bursting reminiscent of that observed during milk ejection is facilitated by OT, an effect that is totally dependent on glutamatergic excitatory neurotransmission.

Of great interest is the role that OT plays in regulating its own release. OT neurons possess OT receptors such as those found in the myoepithelium of mammary gland and in uterine smooth muscle. These receptors sense somatodendritically released OT within the SON and PVN in response to suckling. This local release recruits additional neurons into the synchronous burst discharges and increases burst amplitude. Blockade of brain OT receptors severely disrupts milk ejection, decreasing burst amplitude and frequency. Although these autoreceptors are undoubtedly important, the level at which OT produces these effects is not known. The autoreceptors are linked to increases in intracellular Ca^{2+} through internal stores. However, although activation of OT receptors on other neuronal types (e.g., vagal motor neurons) results in membrane depolarization through a persistent Na^{+} current, such an action has yet to be identified in the SON and PVN. This action is G-protein coupled but independent of intracellular Ca^{2+} mobilization, and it may work through cAMP and phospholipase C second messengers. *In vitro* studies further suggest that OT modulates excitatory inputs to OT neurons since OT applications can increase the frequency of bursts driven by volleys of excitatory postsynaptic potentials as well as burst amplitude. Presynaptic inhibitory actions of OT have also been described, but data suggest these may be mediated through OT-induced cannabinoid release subsequent to autoreceptor activation on SON and PVN neurons.

See also: Angiotensin II; Blood Pressure: Baroreceptors; Circumventricular Organs; Energy Homeostasis: Hypothalamic Development; Fluid and Electrolyte Homeostasis: Clinical Disease; Hypothalamic Structure–Function Relationships; Osmoregulation; Oxytocin (Peripheral/Central Actions and their Regulation); Salt Appetite; Thirst; Vasopressin/Oxytocin and Receptors.

Further Reading

Armstrong WE (2004) Hypothalamic supraoptic and paraventricular nuclei. In: Paxinos G (ed.) *The Rat Nervous System: A Handbook for Neuroscientists*, 3rd edn., pp. 369–388. Sydney, NSW: Academic Press.

Armstrong WE, Stern JE, and Teruyama R (2002) Plasticity in the electrophysiological properties of oxytocin neurons. *Microscopy Research and Technique* 56: 73–80.

Baylis PH (1987) Osmoregulation and control of vasopressin secretion in healthy humans. *American Journal of Physiology* 253: R671–R678.

Cazalis M, Dayanithi G, and Nordmann JJ (1985) The role of patterned burst and interburst interval on the excitation-coupling mechanism in the isolated rat neural lobe. *Journal of Physiology* 369: 45–60.

Crowley WR and Armstrong WE (1992) Neurochemical regulation of oxytocin secretion in lactation. *Endocrine Reviews* 13: 33–65.

Dunn FL, Brennan TJ, Nelson AE, and Robertson GL (1973) The role of blood osmolality and volume in regulating vasopressin secretion in the rat. *Journal of Clinical Investigation* 52: 3212–3219.

Hatton GI (2002) Glial–neuronal interactions in the mammalian brain. *Advances in Physiology Education* 26: 225–237.

Lipschitz DL, Crowley WR, Armstrong WE, and Bealer SL (2005) Neurochemical bases of plasticity in the magnocellular oxytocin system during gestation. *Experimental Neurology* 196: 210–223.

Naeini RS, Witty MF, Seguela P, and Bourque CW (2006) An N-terminal variant of Trpv1 channel is required for osmosensory transduction. *Nature Neuroscience* 9: 93–98.

Poulain DA and Wakerley JB (1982) Electrophysiology of hypothalamic magnocellular neurones secreting oxytocin and vasopressin. *Neuroscience* 7: 773–808.

Renaud LP and Bourque CW (1991) Neurophysiology and neuropharmacology of hypothalamic magnocellular neurons secreting vasopressin and oxytocin. *Progress in Neurobiology* 36: 131–169.

Richard P, Moos F, and Freund-Mercier M-J (1988) Bursting activity in oxytocin cells. In: Leng G (ed.) *Pulsatility in Neuroendocrine Systems*, pp. 75–97. Boca Raton, FL: CRC Press.

Roper P, Callaway J, and Armstrong W (2004) Burst initiation and termination in phasic vasopressin cells of the rat supraoptic nucleus: A combined mathematical, electrical, and calcium fluorescence study. *Journal of Neuroscience* 24(20): 4818–4831.

Russell JA, Leng G, and Douglas AJ (2003) The magnocellular oxytocin system, the fount of maternity: Adaptations in pregnancy. *Frontiers in Neuroendocrinology* 24: 27–61.

Sladek CD (1999) Antidiuretic hormone: Synthesis and release. In: Fray JCS (ed.) *Handbook of Physiology, Section 7: Endocrinology, Vol. III: Hormonal Regulation of Water and Electrolyte Balance*, pp. 436–495. Oxford: Oxford University Press.

Stricker EM, Huang W, and Sved AF (2002) Early osmoregulatory signals in the control of water intake and neurohypophyseal hormone secretion. *Physiology & Behavior* 76: 415–421.

Theodosis DT, Trailin A, and Poulain DA (2006) Remodeling of astrocytes, a prerequisite for synapse turnover in the adult brain? Insights from the oxytocin system of the hypothalamus.

American Journal of Physiology – Regulatory, Integrative, and Comparative Physiology 290: R1175–R1182.

Tian M, Reger JF, and Armstrong WE (1991) Electron microscopic and immunocytochemical study of rapidly frozen, freeze-substituted neural lobes of rats. *Journal of Neurocytology* 20: 79–96.

Verney EB (1947) The antidiuretic hormone and the factors which determine its release. *Proceedings of the Royal Society of London* 135: 25–106.

Voisin DL and Bourque CW (2002) Integration of sodium and osmosensory signals in vasopressin neurons. *Trends in Neurosciences* 25: 199–205.

Neuroimaging

M J Fulham, Royal Prince Alfred Hospital, Camperdown, NSW, Australia

Crown Copyright © 2009 Published by Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

Introduction

Neuroimaging is the discipline that deals with the *in vivo* depiction of anatomy and function of the central nervous system (CNS) in health and disease. The term 'neuroimaging' is more comprehensive than the older term of neuroradiology; although in the past, neuroimaging was considered the domain of radiologists with a specific interest in the nervous system, the participants in this rapidly changing field now come from diverse backgrounds. Contributors exist in all areas in the neurosciences, including basic scientists, neurologists, neurosurgeons, psychologists, psychiatrists, physicists, chemists, and clinicians in radiology and nuclear medicine. It is important to realize that neuroimaging is a dynamic and evolving field. The developments that have occurred in the past decade could not have been predicted when the first edition of this Encyclopedia was published, and we can expect that in the future some of what we currently regard as essential elements of neuroimaging will be surpassed.

In 1895, a year after Roentgen discovered the X-ray and thus the medical specialty of radiology, the penetrating ability of this new ray was used to localize bullets in the calvarium. It therefore indirectly provided the first insights into the morphology of the living brain. However, the skull, vertebral column, and theca remained formidable barriers to direct visualization of the brain and spinal cord for over 75 years. Instead, the anatomy of the CNS was inferred from indirect observations, such as bony erosion on plain skull radiograph, displacement and distortion of the ventricular system (via pneumoencephalography) and the arteries and veins (via angiography), and the outline of the spinal cord (via myelography). Confirmation of the neuroscientists' diagnosis during life was only possible at neuropathologic examination of the brain or spinal cord after death. It was not until 1972 when Hounsfield introduced X-ray computed tomography (CT) that a true representation of the brain itself was realized. CT revolutionized the neurosciences, because for the first time a noninvasive modality could demonstrate the ventricular system, separate gray from white matter, outline the deep nuclei, and recognize pathologic changes in the cerebral parenchyma. CT was remarkable not just for its ability to display anatomic detail, but more generally, because it introduced the fundamentally new technique of 'true' tomography – the imaging of slices – through mathematic reconstruction of data obtained from multiple projections. Computers enabled this advance and also aided the development of the other major three-dimensional imaging modalities in neuroimaging – emission computed tomography (ECT) in its two forms (positron emission tomography [PET] and single photon emission computed tomography [SPECT]) and magnetic resonance (MR) imaging. Although these modalities provide different imaging data, they have the common unifying thread of a computed mathematic reconstruction of the final image (Figure 1).

The basis of all reconstruction techniques is the synthesis of a three-dimensional image from a series of two-dimensional images (projections) taken at various angles around the subject. Thus although the image value of each small volume element within the object cannot usually be determined directly, it can be computed by making a series of measurements at many angles with detectors placed outside the patient. In CT, for example, a slab of tissue is penetrated by a narrow X-ray beam, which is itself composed of photons, that originates from a large number of angles. Antipodal detectors record the multiple transmissions so that an array of point-by-point relative absorption coefficients can be calculated and techniques such as filtered back projection and Fourier analysis are used to make the computation. In MR imaging a somewhat different approach is used, as gradients of the magnetic field permit a more direct evaluation of the Fourier coefficient of each volume element; however, Fourier techniques are still used to reconstruct the full image. Regardless of the algorithm, the reconstructed image is a quantitative measure of some property of the tissue. In CT, the tissue property is the X-ray attenuation coefficient. PET and SPECT techniques measure the tissue concentration of a radioisotope (tracer) in raw counts or nanocuries (nCi) per milliliter after the tracer has been injected into the patient in trace quantities and for MR imaging the property being measured is the transverse magnetization density.

Computed Tomography

A CT scanner measures the linear attenuation or reduction in the X-ray beam as the beam passes

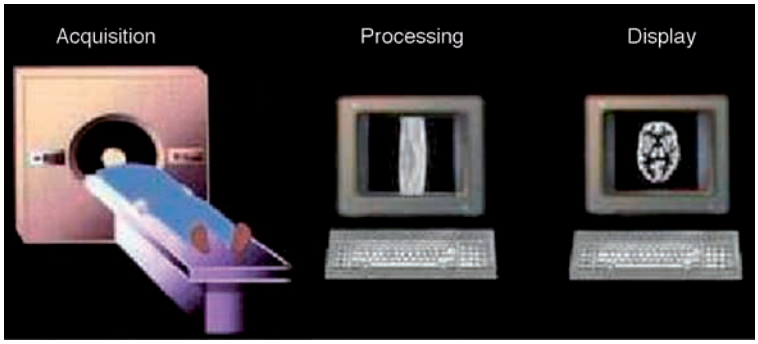

Figure 1 Sequence of events in neuroimaging.

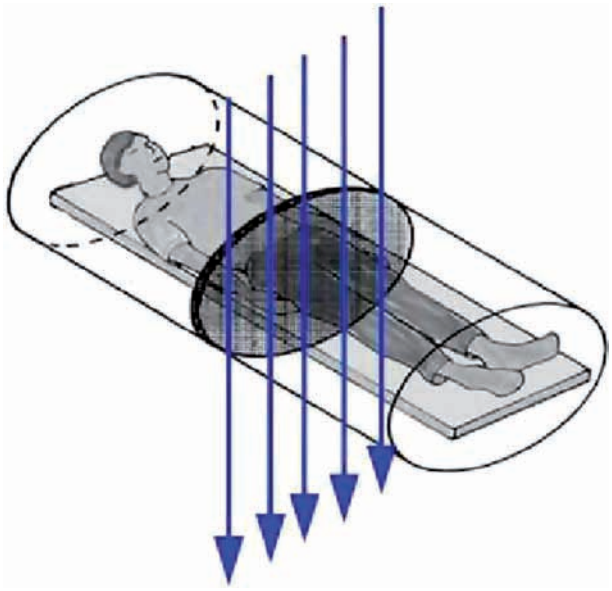

Figure 2 Relationship of X-ray beam, the plane of tissue imaged, and the subject in computed tomography.

through the body (Figure 2) X-ray attenuation is made up of two physical components: (1) Compton scattering, which depends on electron density, and (2) photoelectric absorption, which depends on atomic number and density. Collimators (devices for producing a beam of parallel X-rays) are used to reduce the amount of scattered photons. The unit of X-ray attenuation coefficient is the Hounsfield unit and it is defined as 1/1000 of the attenuation coefficient of water, assuming the same X-ray energy. The intensity of each exiting X-ray beam is measured and transformed into an electronic signal that is then quantified, digitized, reconstructed, and displayed. To a first approximation, a CT scan may be thought of a density image. Limitations of CT include the associated patient radiation exposure, because X-rays are ionizing radiation, artifacts caused by beam hardening, patient motion and partial volume effects, and inferior anatomic detail for gray and white matter when compared with MR imaging. Beam hardening occurs particularly in regions adjacent to thick dense bone such as the skull base or in the oropharynx caused by dental fillings. It occurs because average energy of the exiting X-ray is greater than that entering the patient as a result of absorption of the lower energy X-ray beam by bone. Artifacts induced by beam hardening render the interpretation of structures in the middle and posterior cranial fossae problematic. Helical or spiral CT involves translation of the patient through the scanning gantry while the X-ray source rotates simultaneously so that continuous data acquisition exists throughout the volume of interest and has now become the standard in most medical imaging departments (Figure 3). The X-ray traces a helix, akin to a spiral staircase, on the patient's surface. Spiral CT has distinct advantages in whole-body scanning, where respiratory misregistration artifacts are minimized and the overlapping image planes offer the potential for true three-dimensional image displays. In the brain the newer multi-slice tomographs reduce movement artifacts because of shorter acquisition times and allow for dynamic data acquisition (CT angiography) and three-dimensional data reconstruction.

Cerebral CT is still widely used for day-to-day management of patients with neurologic and neurosurgical conditions. It can rapidly identify cerebral edema/atrophy, hydrocephalus, cerebral infarction, mass lesions, and extravasated blood from sub-arachnoid and intra-cerebral hemorrhage. The integrity blood–brain barrier (BBB) is evaluated after the intravenous injection of iodinated contrast media. Although the patterns of enhancement that are associated with tumors, cerebral infarction, infection, and inflammation differ slightly, CT contrast enhancement is relatively nonspecific. However, CT is without peer in the evaluation of the bony skull. Further, in the acute stroke setting, the wider availability of helical CT

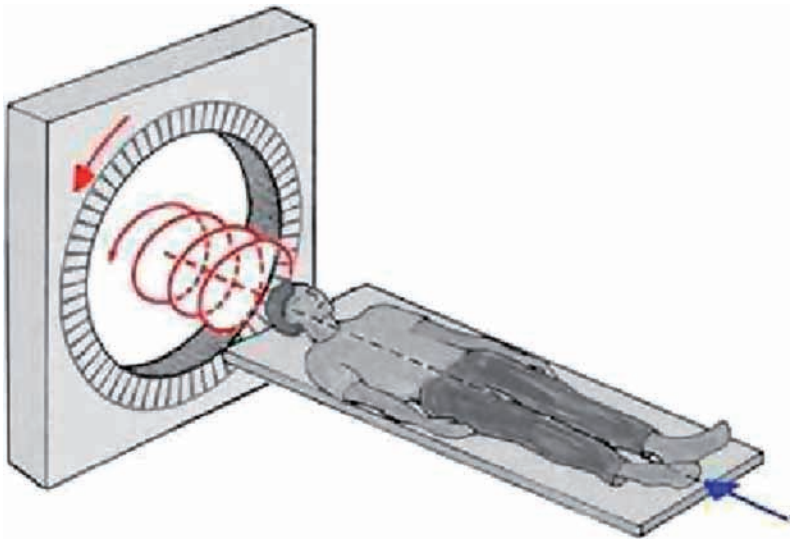

Figure 3 Helical computed tomography shows relationship between patient translation (purple arrow) and simultaneous movement of X-ray source (red arrow).

has enabled the acute assessment of patency of the major cerebral vessels – the extracranial and intracranial carotid, distal vertebral, basilar, and proximal middle cerebral arteries – through dynamic data acquisition after the rapid intravenous injection of contrast (CT angiography). However, to a great extent, if MR imaging is available, it is the method of choice used by neuroscience clinicians because of its superior depiction of anatomy and physiology in the supra- and infratentorial compartments, its ability to evaluate the integrity of the BBB with para-magnetic contrast agents, and its lack of ionizing radiation.

Emission Computed Tomography

Emission computed tomography (i.e., PET, SPECT) is the term for functional imaging methods that use injected radioisotopes and combine the principles of the tracer kinetic method and tomographic image reconstruction. In the tracer method, a radiolabeled, biologically active compound (the tracer) is injected into a subject in minute or tracer quantities. The tracer then participates in, without perturbing, a biochemical process. The tomography or scanner detects the tissue concentration of the tracer. A mathematic model can then be derived to describe the kinetics of the process, and in some instances it is possible to convert the raw counts to a unit of physiologic function such as the cerebral metabolic rate for glucose or cerebral blood flow. A number of fundamental differences exist between PET and SPECT. The instrumentation and also the radioisotopes that are used differ. Table 1 lists a number of common PET and SPECT tracers, the half-lives of the radioisotopes, and the biologic parameters that can be measured. The list is not exhaustive but noteworthy in that the radioactive isotopes of some of the most important elements in biology happen to be positron emitters (^{15}O, ^{13}N, ^{11}C). These PET isotopes have very short half-lives and are produced in cyclotrons, which generally means that for all but ^{18}F the cyclotron must be in close proximity to the PET scanner. SPECT isotopes are much longer lived and may be transported long distances. Most commonly used SPECT radioisotopes are produced in reactors, but ^{123}I, ^{68}Ga, and ^{201}Tl are produced by cyclotrons.

In PET the unstable positron emitters decay to a stable state by emitting a positron (positive electron) from the nucleus. The positron travels a small distance (1–2 mm) in tissue before colliding with a surrounding electron and annihilation takes place. The masses of the positron and electron are converted to energy in the form of two high-energy photons or gamma rays that are emitted 180 degrees apart (Figure 4). The simultaneous production of annihilation photons in opposite directions and their coincident detection by the PET scanner are the fundamental bases for the localization of the positron emitter in tissue. The PET scanner has multiple rings of detectors that are composed of inorganic crystals (e.g., bismuth germanate), which scintillate when struck by a photon. The detectors are connected to photomultiplier tubes and the light energy is converted into a digital signal to provide a three-dimensional distribution of the tracer. The physics of PET allow the measurement of scattered and attenuated photons

Table 1 Some PET and SPECT radiotracers in the neurosciences

Technique	*Radiotracer/neuroreceptor ligand*	*Radioisotope/half-life*	*Biologic parameter/receptor*
PET			
	[^{18}F]fluorodeoxyglucose (FDG)	^{18}F ~ 108 min	Glucose metabolism
	[^{18}F]-6-flurodopa		Dopamine
	[^{11}C]-methionine	^{11}C / 20 min	Amino acid uptake
	[^{11}C]-raclopride		Dopamine
	[^{11}C]-nomifensine		Catecholamine re-uptake
	[^{11}C]-flumazenil		Central benzodiazepine
	[^{11}C]-PK 11195		Peripheral benzodiazepine
	^{15}O-water	^{15}O/~ 2 min	Cerebral blood flow
SPECT			
	99 m Tc-hexamethyl propylene amine oxime (HMPAO)	99 mTc / 6 h	Cerebral blood flow
	^{201}Thallium	^{201}T1 / 73.1 h	+ analog-cellularity
		^{133}Xe xt-raise: 3.0pt'	Cerebral blood flow
	^{123}I-iodoamphetamine (IMP)	^{123}I ~ 13 h	Cerebral blood flow
	^{123}I-methyl tyrosine (IMT)		Amino acid transport
	^{123}I-iodoamphetamine (IMP)		Cerebral blood flow
	^{123}I-QNB		Muscarinic cholinergic
	^{123}I-iodobenzamide (IBZM)		Dopamine
	^{123}I-iomazenil		Central benzodiazepine
	^{123}I-iododexetimide (IDEX)		Muscarinic cholinergic

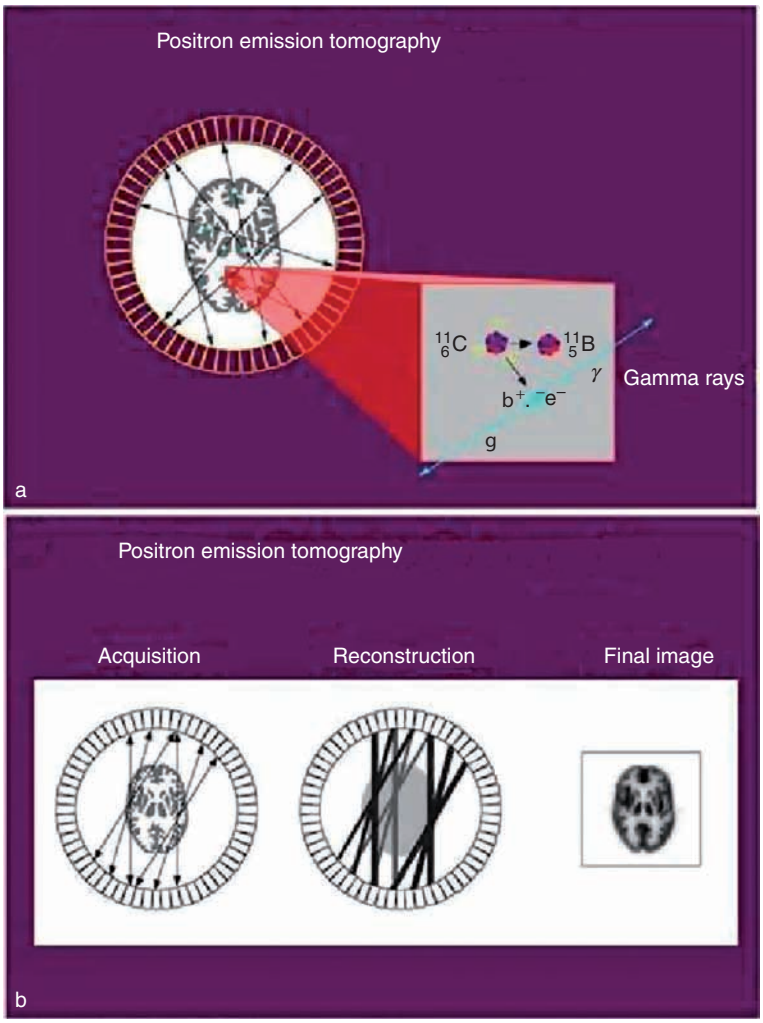

Figure 4 (a) Depiction of positron emission and co-incidence detection in positron emission tomography (PET). (b) Steps in image production in PET. Center panel shows line of response used to produce image.

so that quantitative measurement of a particular physiologic process is possible. In SPECT, single photons are detected by rotating gamma cameras with a single large crystal detector (Figure 5). Multi-detector gamma cameras (double- and triple-headed cameras) have enabled improved resolution through the detection of a greater proportion of unscattered photons. However, scatter and attenuation correction are not routine in SPECT, which limits the provision of quantitative data. Instead, ratios of regional counts within a study are used to compare data. The resolution in PET is also superior to that in SPECT. The latest generation PET scanners have resolution in the order of 2–3 mm full-width at half-maximum (FWHM), whereas for triple-headed SPECT cameras, resolution approximates 8–10 mm FWHM compared with 14–17 mm FWHM for single-head systems and 7–8 mm FWHM for ring-detector imaging devices, similar to PET scanners. However, these devices are not widely available. Resolution is determined by a number of factors, including (1) the type of detectors, (2) the distance a positron (for PET) travels in tissue before annihilation, (3) random and scatter coincidences, and (4) reconstruction algorithms. In an exciting new development, a new and more efficient detector crystal called lutetium oxyorthosilicate (LSO) with a shorter fluorescence time has been incorporated into PET scanners. This detector material has significantly shortened the acquisition time for neurologic PET scans, but in addition it has been placed in PET scanners that have been coupled to CT scanners, so called PET-CT. These devices are now replacing stand-alone PET scanners as the norm for clinical PET devices. They offer the added advantage of perfectly registered anatomic (CT) and functional (PET) data. The CT scanners are helical devices that

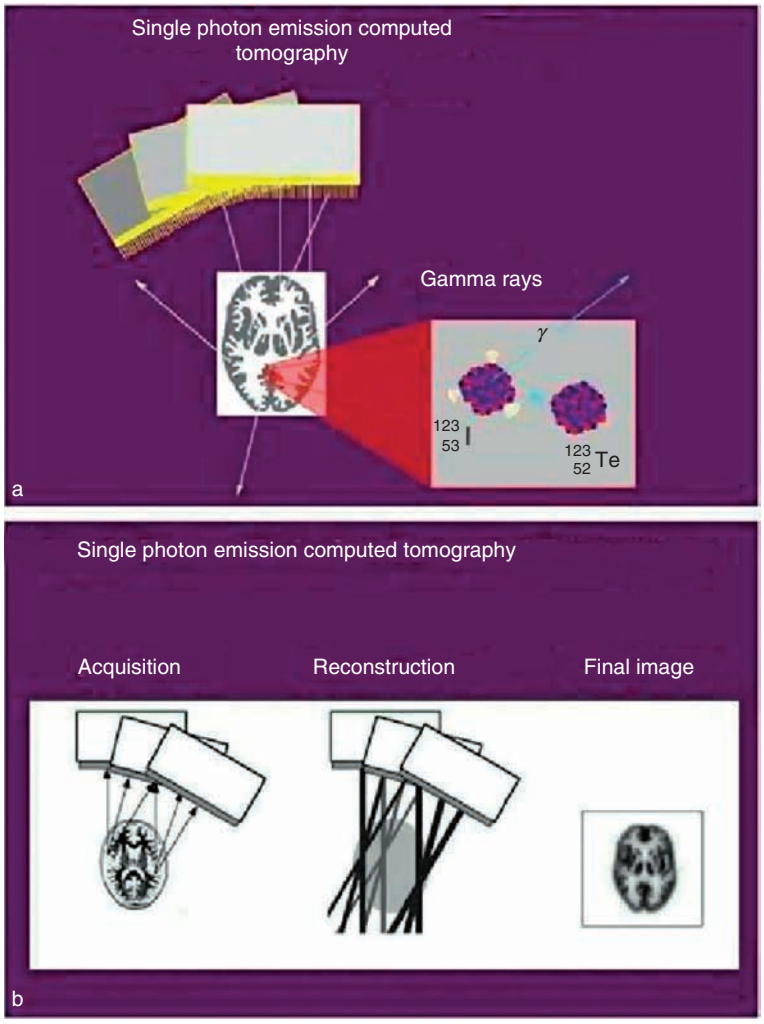

Figure 5 (a) Depiction of single-photon emission and detection in single-photon emission computed tomography (SPECT). (b) Steps in image production in SPECT. Center panel shows line of response used to produce image.

provide the means of attenuation correction for the PET, as well as the anatomic information. These combined devices offer advantages for selective PET receptor experiments. It is anticipated that the next generation of PET scanners will approach the theoretic finite in plane spatial resolution in PET of less than 2 mm.

Overall, PET has proved more fruitful than SPECT in the neurosciences by virtue of its superior resolution and ability to quantify *in vivo* biochemical processes. PET has been used in the evaluation of cerebral blood flow, blood volume, oxygen use and metabolism, glucose metabolism, amino acid transport, and metabolism and protein synthesis. Neuroreceptor ligands for the dopaminergic, opiate, benzodiazepine, cholinergic, and serotonergic systems have been developed and used in humans. In recent years much effort has been directed to brain mapping with ^{15}O-water, where localization of cerebral function is deduced from changes in local cerebral blood flow that are induced by the presentation of a paradigm (e.g., a complex image) to a normal subject. In some instances these data have confirmed findings in nonhuman primates, but in other instances these data have provided unique insights into the functioning of the brain. In the clinical arena the most commonly used PET isotope is FDG (see Table 1), and through the evaluation of cerebral glucose metabolism PET is used to localize epileptic foci in patients with refractory epilepsy, separate low-grade from high-grade brain tumors, differentiate the effects of treatment from recurrent tumor, and distinguish Alzheimer's disease from other types of dementia and neurodegenerative conditions (such as multiple system atrophy, corticobasal degeneration, Huntington's disease, acquired immunodeficiency syndrome (AIDS), dementia, and progressive supranuclear palsy). A reduction in parietal lobe glucose metabolism has been described in individuals at high risk for developing familial dementia before the onset of symptoms. In this scenario, PET will provide a method to both identify those at risk and evaluate new therapies in the dementias. Early observations also indicate that PET neuroreceptor studies will assist us in the management of extrapyramidal disorders such as Parkinson's disease, other neurodegenerative conditions including dementia, and the refractory epilepsies.

Magnetic Resonance Imaging

The nuclear magnetic resonance phenomenon was originally described in 1946. It results from the magnetic properties of some atomic nuclei, particularly the hydrogen nuclei in water molecules. Hydrogen nuclei in the body behave like tiny bar magnets when placed in a strong magnetic field. Just as bar magnets with north and south poles align themselves with the external magnetic field, so too do the hydrogen nuclei in the body when placed in the strong magnetic field of an MR scanner. This orientation can be perturbed by the application of electromagnetic energy in the form of radiofrequency (RF) pulses. An RF pulse applied at right angles to the static field around a subject changes the orientation of the nuclear spins of the hydrogen molecules and increases their energy state. When the field is turned off, the nuclei return to their original state with the emission of energy at the same frequency as that of the applied RF (resonance or Larmor frequency) pulses. The Larmor frequency varies significantly from element to element and even for different isotopes of the same element. The emitted radiofrequency signal can be detected in a receiver coil placed around the subject's head. The frequency of the signal depends upon the strength of the magnetic field, the physicochemical properties of the tissue, and the manipulation of the applied RF pulses. By varying the strength of the field spatially, the resulting signal reflects the magnetization of water molecules at each point in space, and with proper spatial encoding these signals can be reconstructed to yield images with high contrast and spatial resolution and orientated in axial, coronal, or sagittal planes.

Many factors affect the MR image: excited proton density, the longitudinal (spin-lattice) relaxation time T1, the transverse (spin-spin) relaxation time T2, diffusion, and flow. Differences in tissue relaxation times are the main determinants of image contrast. Therefore *a priori* knowledge of the relaxation behavior of different tissues is important in the selection of the proper RF pulse sequence for optimal images production. RF pulses, which tilt the nuclear spins 90 degrees or 180 degrees, may be injected at various times, with various sequences and frequencies. The terms *TE-echo time*, TR-repetition time, and T1-inversion time are used to indicate some of these parameters. In addition, the signal-to-noise ratio is favorably affected by the field strength and surface coils. It was apparent, very early in the development of MR imaging, that the image quality surpassed that of CT (Figure 6). MR imaging was able to demonstrate small structures such as the substantia nigra, the red nucleus and, in selected cases, the various thalamic nuclei separated by strips of myelin. In the spinal cord, the H-shaped central distribution of gray matter was delineated. Pathologic lesions were also much more easily identified, because they are often characterized by long T1 and T2. In multiple sclerosis, MR imaging provided a marked increase in

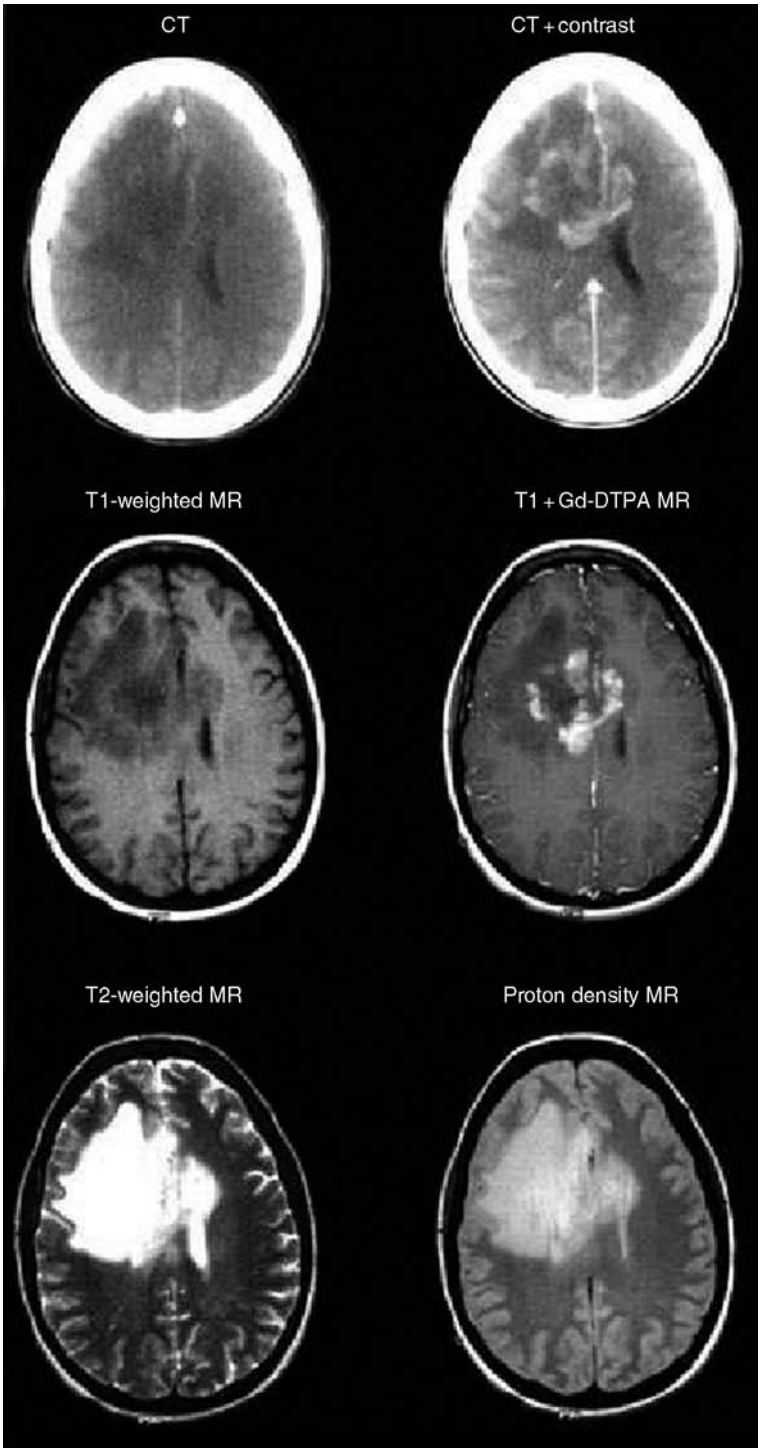

Figure 6 Patient study. Thirty-five-year-old woman with multiple sclerosis and massive plaque of demyelination in corpus callosum imaged with computed tomography and magnetic resonance (MR) imaging. MR images demonstrate superior anatomic definition but also the inability to separate edema from inflammation on T2-weighted and proton density images.

sensitivity for identifying plaques of demyelination and insights into the activity of the disease by the demonstration of rapid appearance and disappearance of asymptomatic plaques. Mass lesions with their associated edema were easily shown, and the natural history of small cavernous hemangiomas was documented. Examination of the spinal cord was simpler – tumors, syrinxes, cord compression from discs and tumors, and atrophy, as well as congenital abnormalities such as tethered cord were easily

recognized. Imaging the posterior and middle cranial fossae was easier because of the absence of beam-hardening artifacts.

However, although MR imaging offered greater sensitivity when compared with CT, it had poor specificity. Tumors, cerebral infarctions, some hematomas, foci of demyelination, and regions of edema all gave similar signals on T2-weighted images. Therefore, in an effort to improve this deficiency together with a desire to explore and push the technology to its limits, a bewildering number of advances have been made in the past few years. There have been improvements in (1) instrumentation: surface coils and high magnetic field strength (4.0 Tesla) devices; (2) software: new pulse sequences/protocols with acronyms such as FLASH, FLAIR, MPRAGE, magnetization transfer ratios, transverse magnetization decay; (3) contrast media; (4) pulse sequences to image blood flow and hence MR angiography (MRA); and (5) MR techniques known as functional MR imaging. However, before we discuss functional MR imaging methods, it is important to mention that MRA, although it has not yet replaced digital subtraction angiography for the assessment of the intracranial and extracranial arterial and venous circulation, is a useful clinical adjunct and can provide valuable information for clinical management.

Functional Magnetic Resonance Imaging

Functional MR imaging refers to techniques that provide data about aspects of metabolism, be it tissue metabolite concentrations, perfusion/diffusion, and blood flow rather than pure anatomy (which forms the basis of routine clinical MR imaging). The term encompasses MR spectroscopy and those techniques that rely on the change in blood oxygenation levels to provide an indirect image of neuronal activity that is referred to in some circles as *functional MR imaging* (fMRI). This is an exciting area of development, because it means that functional data, formerly the exclusive domain of PET and SPECT, can now be obtained with MR scanners without the associated ionizing radiation.

In magnetic resonance spectroscopy (MRS), signals are obtained from nuclei that are constituents of molecules other than water. These signals are generated by nonradioactive nuclei that are parts of other solutes or metabolites, but these signals are much weaker than the signals arising from water. The solutes that produce them occur in concentrations of 0.5 mM. Non-water ^{1}H, ^{31}P, and ^{13}C are the most important nuclei in MRS. The signals are analyzed and the radio frequencies emitted by the sample of interest appear as peaks on a plot of signal intensity against signal frequency, thus producing an NMR spectrum. The horizontal or frequency axes are normalized to the frequency of a strong signal of known origin and the location of peaks are given as parts per million (ppm) relative to the known signal (Figure 7). The signal strength (the vertical axis) is proportional to a number of factors, including the number of nuclei in the tissue volume being studied. Signals from different isotopes are distinguishable, because they occur at different frequencies in any given magnetic field, nuclear signals from different solutes can also be separated from one another by the property of chemical shift. 'Chemical shift' refers to the fact that within a molecule containing magnetic nuclei, the electron clouds in the chemical bonds surrounding the nuclei change the local magnetic field so that nuclei of the same species spin at different frequencies.

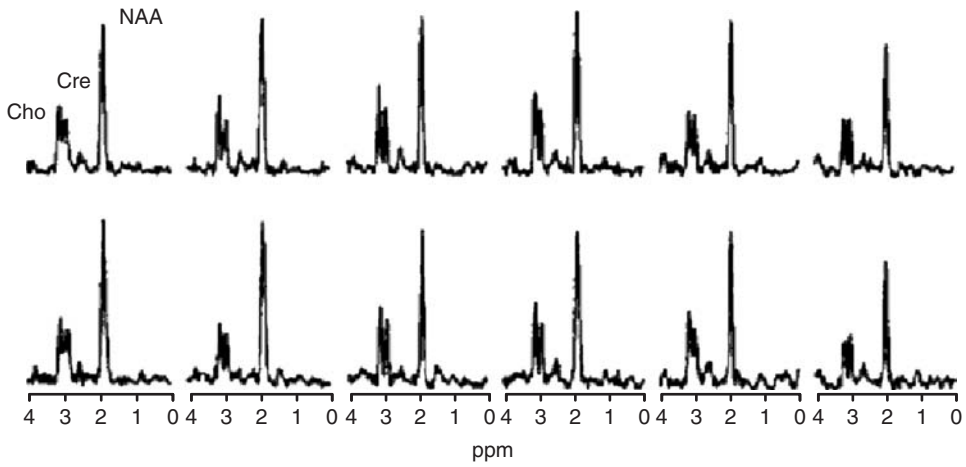

Figure 7 ^{1}H-magnetic resonance spectroscopy imaging spectra show distribution of metabolites (in ppm) from a slab of tissue.

The influence of these weak signals is that (1) MRS has poorer sensitivity and resolution when compared with clinical MR imaging, (2) some solutes cannot be measured with MRS (e.g., virtually all neurotransmitters with the exception of glutamate, glutamine, and GABA are beyond detection). Further, molecular motion is required for a solute to be detected with MRS, and most macromolecules (myelin, proteins, nucleosides, phospholipids) are inaccessible because of their limited mobility. Both ^{31}P and ^{1}H-MRS have been used in the neurosciences, but only the application of ^{1}H-MRS (or proton-MRS), which has an inherently stronger signal, better resolution, and greater sensitivity than ^{31}P, is discussed here. In proton-MRS, four metabolites are of interest: (1) *N*-acetyl-aspartate (NAA), (2) choline-containing compounds (Cho), (3) creatine-containing compounds (Cre), and (4) lactate (Lac) (Table 2). Initial ^{1}H-MRS studies involved acquisition of data from single voxels of varying sizes, so effectively only a small region of the brain could be sampled at once. Gradient-based phase-encoding techniques allowed the simultaneous acquisition of data from multiple voxels in a single tissue slab; therefore, instead of interpreting an NMR spectrum, an image of the topographic distribution of metabolites, or spectroscopic imaging, was produced. Further pulse sequence enhancements led to the development of multi-slice spectroscopic imaging (Figure 8).

Preliminary data suggest that ^{1}H-MRS imaging can be used for the grading of brain tumors, the detection of malignant degeneration in low-grade tumors, the localization of seizure foci in refractory epilepsy, the evaluation of the dementias, and the differentiation of cerebral abscesses from other enhancing lesions. However, proton MRS has some limitations. It places more stringent demands upon the quality and strength of the magnetic field than does clinical imaging. Clinical imaging can be performed quite satisfactorily at 0.5 Tesla (T), but stronger fields are needed for MRS. Most current MRS work is performed at 1.5 to 2.0 T, but increasingly 3.0 and 4.0 T devices that provide better metabolite delineation are being used. Good magnetic field homogeneity is critical for MRS, and poor field homogeneity manifests as difficulty in resolving signals from one another and from noise. Most current MR scanners are sold with spectroscopy capabilities; although this capability is for single voxel examinations, proton spectroscopic imaging should be available on clinical MR scanners within a few years.

'fMRI' refers to a number of methods where gradient-recalled echo (GRE) techniques and sub-second echoplanar imaging (EPI) are used to identify aspects of physiologic function in the nervous system. For instance, these MR techniques can identify diffusion of water in the ischemic brain and neuronal activity in normal subjects akin to the ^{15}O-water activation methodologies used with PET. The advantage of these techniques in the experimental setting is that no ionizing radiation is involved; therefore, experiments can be repeated indefinitely in the one subject; the high spatial resolution of MR can be used so that no need exists for subsequent registration of the functional data to anatomy, because it is already provided by the MR scan.

'Diffusion MR imaging' refers to the ability of MR to detect and measure water diffusion in tissue and hence information about tissue composition, microstructure, and architectural organization. A term such as the apparent diffusion coefficient (ADC) is used to describe diffusion in complex media such as the brain and is defined as the ratio of mean-squared displacement and the diffusion time. Diffusion-weighted images (DWIs) are obtained by incorporating diffusion gradient pulses within a conventional MR sequence. A variation of this technique to allow for the measurement of diffusion in tissues oriented in a variety of direction such as cerebral white matter is called diffusion tensor imaging. The major clinical application of DWI at present is in the evaluation of hyperacute (minutes to hours after the ictus) cerebral ischemia. It can separate acute ischemia from old

Table 2 Metabolites seen with ^{1}H-MRS at TE 272 ms

Metabolite	*Frequency (ppm)*	*Site and function*
N-acetyl-aspartate (NAA)	2.02	Neuronal; possible storage form of aspartate; possibly involved in lipid synthesis and regulation of protein synthesis
Choline (Cho)	3.20	Cytosol; includes phosphorylcholine, glycerophosphoryl choline, and choline; components of phospholipid metabolism; role in cell membrane turnover
Creatine (Cre)	3.93, 3.04	Cytosol; reserves of high- and low-energy phosphates and used for cellular synthetic and transport functions
Lactate (Lac)	1.33	Cytosol; terminal metabolite of glycolysis and initial substrate for mitochondrial tricarboxylic acid cycle

infarction but also identify regions of ischemia that are not yet apparent on conventional T2-weighted images that offers tremendous potential for active intervention to limit the extent of cerebral infarction. fMRI techniques are based on the magnetic difference between deoxyhemoglobin and oxyhemoglobin and the small increase in signal intensity, which has been ascribed to a local change in blood oxygenation that results from neuronal activity. During neuronal activity an increase takes place in local cerebral blood flow and metabolism, a local decrease in the amount of deoxyhemoglobin, and an increase in signal intensity. Although other factors exist, such as inflow, effects from draining venules, hematocrit, oxygenation saturation, and cerebral blood volume (which may also contribute to the signal change a number of comparative studies with ^{15}O-PET), they suggest that both techniques image the same process. Increasingly fMRI is being used in the clinical context to localize eloquent cortex adjacent to brain lesions to help guide the neurosurgeon.

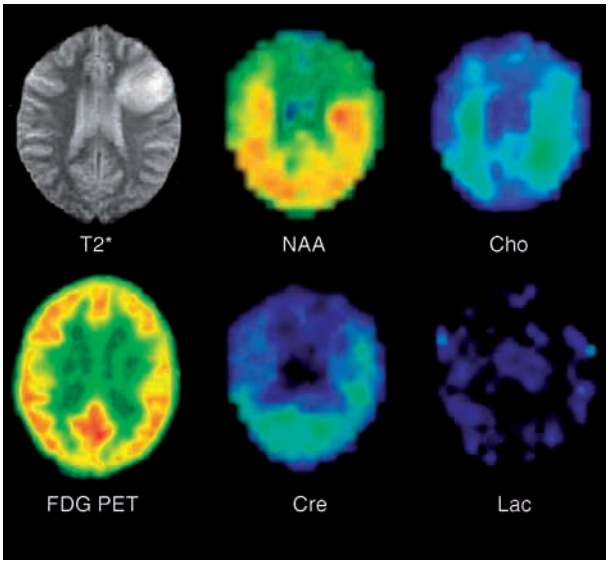

Figure 8 Single slice from ^{1}H-magnetic resonance spectroscopy imaging study in patient with a brain tumor. Individual images indicate maps of metabolite distribution in the one slice plus the anatomic magnetic resonance and positron emission tomography (PET) images of glucose metabolism.

Summary

The advances in neuroimaging in the past decade could not have been predicted in the late 1980s, and major strides will also be made in the next few years. Apart from the rapid pace of technologic developments there has also been the realization that many of the current imaging methods are complementary. Although MR imaging is the single most versatile modality, PET, SPECT, and CT still provide important unique data that are indispensable to the neurosciences for research and patient management. Just as algorithms exist to merge and use these data (Figure 9), contributions to neuroimaging come

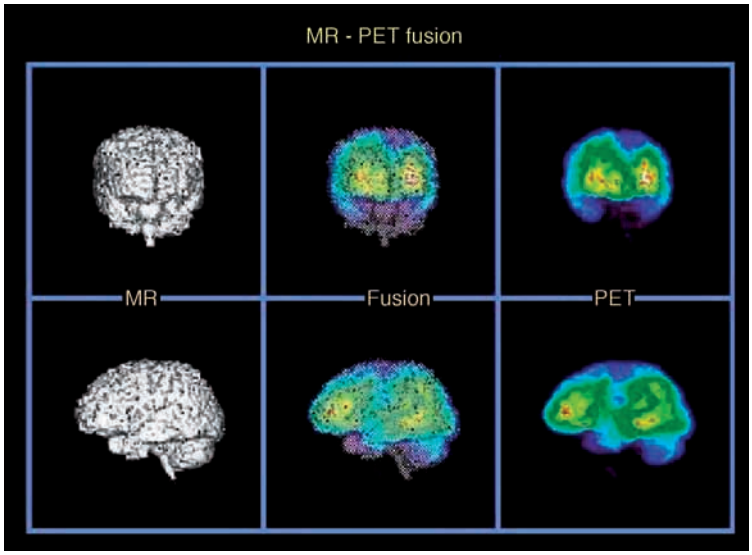

Figure 9 Twelve-year-old boy with refractory frontal lobe epilepsy. Volume rendered magnetic resonance (MR), positron emission tomography (PET), and fused magnetic resonance spectroscopy (MRS)-PET data. Images show a large area of reduced glucose metabolism in region of mesial frontal and dorsolateral prefrontal cortex that is responsible for seizure generation. Minimal change seen in MR image. In addition, crossed cerebellar hypometabolism (PET) and atrophy (MR) from the longstanding frontal lobe heterotopia are seen.

from diverse investigators. PET and CT are being combined in the clinical setting; in the author's opinion, continued collaboration across the imaging specialties will increase our understanding of the nervous system and improve the care of those afflicted with neurologic disorders. The PET-CT and the PET-MR animal scanners provide a window to the next phase of imaging development where unique functional and anatomic datasets are obtained simultaneously.

See also: Cognition: An Overview of Neuroimaging Techniques; Cognitive Neuroscience: An Overview; Evoked Potentials: Recording Methods; Executive Function and Higher-Order Cognition: Neuroimaging; fMRI: BOLD Contrast; Magnetic Resonance Spectroscopy; Magnetoencephalography; Perfusion MRI; Positron Emission Tomography (PET).

Further Reading

Atlas SW (2002) *Magnetic Resonance Imaging of the Brain and Spine*, 3rd edn. Philadelphia: Lippincott Williams & Wilkins.

Fulham MJ and Di Chiro G (1996) Neurological PET and SPECT. In: Harbert JC, Eckelman WC, and Neumann RD (eds.) *Nuclear Medicine: Diagnosis and Therapy*, pp. 361–386. New York: Thieme Medical Publishers.

Mazziotta JC and Gilman S (1992) *Clinical Brain Imaging: Principles and Applications*. Philadelphia: FA Davis.

Prichard JW and Brass LM (1992) New anatomical and functional imaging methods. *Annals of Neurology* 32: 395–440.

Wahl RL (2002) *Principles and Practice of Positron Emission Tomography*. Philadelphia: Lippincott Williams & Wilkins.

Neuroimmune System: Aging

E Bellavista and C Franceschi, University of Bologna, Bologna, Italy

© 2009 Elsevier Ltd. All rights reserved.

The Immune-Neuroendocrine System: The Evolutionary Perspective

The immune system, consisting of innate and adaptive branches, acts to recognize and cope with a variety of stressors, including physical, emotional, chemical, and biological agents. The evolutionary perspective indicates that this system is a part of a larger network involving the neuroendocrine circuit, termed the immune-neuroendocrine suprasystem, which is able to integrate immunity, stress response, and inflammation. The major characteristic is that this suprasystem is conserved with variations and increased complexity, from invertebrates to vertebrates, and a pool of well-known molecules (pro-opiomelanocortin-derived peptides, cytokines, biogenic amines, glucocorticoids, and nitric oxide) are used for intra- and intersystem communication. The main cellular actor of this network is the macrophage, which produces all of the previously mentioned mediators.

In a number of low eukaryotes, such as yeast and worm, the ability to cope with stressors is provided by a defense network including superoxide dismutase and catalase (free radical scavenger enzymes), heat shock proteins and other chaperones, DNA repair genes, and intracellular signaling molecules (e.g., H-RAS). A study on *Caenorhabditis elegans*, a favorite model for the genetic investigation of the physiology and developmental biology of multicellular organisms, showed that the resistance to pathogen is mediated by a Toll/interleukin-1 receptor domain adaptor protein and a conserved p38 mitogen-activated protein kinase family (PMK)-1 pathway. Moreover, this circuit is involved in mediating resistance to osmotic stress, in addition to mediating immune signaling.

In invertebrates (mollusks, insects, and annelids), immune and neuroendocrine functions are performed by a cell type called immunocyte, the ancestor of macrophage, that can be considered an 'immune-mobile brain.' This phagocyting cell expresses molecules such as adrenocorticotropin hormone (ACTH), corticotropin-releasing hormone (CRH), interleukin-like IL-1α, -1β, -2, and -6, and tumor necrosis factor (TNF)-α that are able to affect migration and increase phagocyte activity against pathogens. Moreover, immunocytes express a NO synthase-like molecule producing NO, which provokes the clumping of bacteria around them and their subsequent killing. Pathogen recognition in *Drosophila melanogaster* is mediated by different molecules, including a member of the immunoglobulin superfamily which is diversified by alternate splicing. This molecule is called Dscam, the homolog of Down syndrome cell adhesion molecule, and its isoforms are expressed in immune tissues as membrane and soluble proteins.

During evolution, the old molecular and cellular effectors merged with the new sophisticated recognition system, the adaptive clonotypical immunity, to face the increasing complexity of the organisms, which required a more sophisticated type of circuitry and recognition. Specialized function for the defense of the organism occurred in jawed fish, in which the thymus, the B and T lymphocytes, and the major histocompatibility complex (MHC) appeared in order to improve recognition of external molecules, on the one hand, and avoid self-destruction, on the other hand. The thymus, a neuroendocrine organ, was built up with cellular and molecular components present in invertebrates, and it is able to perform a local immune and stress response. Accordingly, the appearance of a new organ such as the thymus can be envisaged as an example of tinkering – that is, the combinatorial reutilization of preexisting materials, a strategy largely used in evolution.

In mammalian, immune and neuroendocrine responses are bidirectionally regulated through more specialized tissues and organs, where neuropeptides and cytokines represent the signaling molecules responsible for chemical information. In particular, in the central nervous system (CNS), the hypothalamus integrates information from physical (e.g., trauma and infection) and emotional stimuli (e.g., bereavement) and, in response to stress, releases CRH from the paraventricular nucleus, which acts on the pituitary gland to induce the release of ACTH to stimulate the adrenal gland and the production of glucocorticoids (cortisol) and dehydroepiandrosterone (DHEA) from the cortex. The adrenal gland is also regulated by the sympathetic nervous system, secreting catecholamines in response to acetylcholine release from the fibers innervating adrenal medulla. Catecholamines and glucocorticoids mediate the 'fight-or-flight' response to stress but at the same time induce suppression of the majority of immune functions. The principal mechanism of glucocorticoids is the downregulation of pro-inflammatory cytokine synthesis and the inhibition of enzymes such as cyclooxegenase-2 (COX-2). Moreover, they affect cell-mediated immune function,

suppressing natural killer (NK) cell cytotoxicity and neutrophil chemotaxis. By contrast, DHEA is primarily immune enhancing and counterbalances the potentially detrimental effect of long-term stimulation of the hypothalamic–pituitary–adrenal (HPA) gland axis. In particular, DHEA improves NK cell number and cytotoxicity and induces superoxide generation in neutrophils. As an example of this connection between the CNS and the immune system, we studied the effect of an emotional stimuli, such as bereavement, on the immune system and found functional alterations of immune parameters, such as responsiveness of peripheral blood lymphocytes to mitogens. In particular, despite a normal number of circulating lymphocyte subsets, the functional activity of NK cells was markedly reduced. Moreover, changes in the intracellular concentration of β-endorphin in peripheral blood mononuclear cells correlated with anxiety and depression scores. Other distress-generating events (e.g., divorce, caring for a relative affected by Alzheimer's disease, and unemployment) are associated with alterations in both cellular and humoral immune response and enhanced risk of developing infectious disease.

Evidence also supports the idea that information can flow from the immune system to the CNS – for example, the innervation of the immune system and organs, the production of neuropeptides by cells of the immune system, and the presence of receptors for both cytokines, as well as neurotransmitters on immune and CNS cells. An example of such communication is that, upon peritoneal infection with bacteria, resident macrophages produce IL-1, which acts on the vagus nerve to cause behavioral changes and illness symptoms, and on hypothalamus and pituitary gland to produce CRH and ACTH. Additionally, lymphocytes produce β-endorphin to modulate pain sensations, acting on peripheral sensory nerves. Moreover, evidence from our laboratory indicates that a pure immunological stimulus is able to impact on the brain inducing neuroendocrine changes and to modify cognitive performance. Indeed, healthy volunteers were vaccinated against hepatitis B and tested during a 6-month period on a simple reaction times task and the Stroop task. Although we did not find improvement in the general efficiency of the CNS as indexed by simple task reaction, when we considered higher cognitive function (Stroop task), the performance of vaccinated subjects was faster and more accurate. Additionally, these effects were not acute, being evident 4 weeks after antigenic stimulation, and they persisted even after 25 weeks.

Thus, the available data in the literature suggest that a bidirectional circuitry is present between the immune–endocrine and the nervous system, where both systems continuously alert and inform each other.

Immunosenescence

The aging of the immune system (immunosenescence) is a recent phenomenon related to the linear improvement in survival and life span not predicted by evolutionary force. Indeed, the immune system has probably been selected to serve individuals living until reproduction, as supported by the thymic ontogenesis and involution. Nowadays, the immune system supports the soma of individuals living 80–120 years; during this increased life span, elderly people have to cope with long-lasting antigenic burden (clinical and subclinical infection, as well as continuous exposure to other types of antigens), which causes a chronic activation of the immune system. In order to adapt to and counteract the effect of chronic stress, the immune system undergoes reshaping and remodeling, in which some parameters decrease with age while others increase or remain unchanged. The adaptive immunity is the most affected by aging, being characterized by expanded clones of memory and effector T cells, mostly prominent in $CD8^+$ and specific for viral antigens. In particular, we showed that Epstein–Barr virus (EBV) and cytomegalovirus (CMV) infections are responsible for the expansion of memory $CD8^+$ T cell clones specific for a limited number of viral peptide epitopes which accumulate in the immune system of older people, thus filling the 'immunological space.' We have shown that $CD4^+$ T cells are also affected by viral chronic infections such as CMV. Indeed, when cells from CMV^+ $CD4^+$ T cell are stimulated *in vitro* with specific peptides of the CMV, they produce seven or eight times more inflammatory cytokines such as interferon-γ (IFN-γ) and TNF-α, thus contributing to the inflammatory status present in elderly people. Moreover, aging is associated with a shrinkage of T cell repertoire and a decrease in thymic emigrant T cells and virgin T cells. The result is the filling of the immunological space (the physical space in which all possible interactions among immune cells and/or cell subsets must occur) with indolent cells with memory markers which could exert a suppressor/negative effect on other bystander T cells, including the few naive (virgin or antigen-nonexperienced) cells. Considering the innate immunity, different studies have shown that age-associated alterations are also present in this branch of immunity, but overall it is well preserved. Our attention was particularly focused on NK cells, macrophages, and cytokine circuits. NK cells are large granular lymphocytes expressing

surface markers such as CD16 and CD56 and lacking the CD3 T cell receptor. They display two alternative mechanisms of cytotoxicity activity: the MHC unrestricted cytotoxicity against neoplastic and virus-infected cells and antibody-dependent cell-mediated cytotoxicity. In elderly people, an increase in the absolute and relative number of NK cells, with a mature phenotype ($CD56^{dim}$), is seen, with no variation in their killing activity as a whole, despite the presence of subtle impairment of killing activity on a per cell basis. Moreover, a feedback loop between NK cells and macrophages has been described. Since NK cells secrete factors such as IFN-γ which activate macrophages to eliminate pathogens, some of the age-associated macrophage dysfunctions may be secondary to alteration in cytokine production of NK cells. Additionally, an age-related reduction in hydrogen peroxide and nitric oxide synthesis by macrophage has been described, paralleled by an induction of suppressive substances (e.g., prostaglandin E) which have an inhibitory effect on dendritic cells, the main antigen presenting cells. *In vitro* studies on cytokine production by peripheral mononuclear cells from healthy young and aged subjects identified a complex scenario. Significant increases in levels of IL-6, TNF-α, and IL-1β were found in mitogen-stimulated cultures from aged donors despite a similar amount of cytokines in unstimulated cultures from young and aged subjects. Moreover, chemokines such as MCP-1 and RANTES, soluble TNF-α receptor I and II, and IL-1 receptor antagonist increased in the elderly. For one of the most important proinflammatory cytokines, IL-6, synthesis in macrophage cells is controlled by the CRH–DHEA circuit. Whereas production of cortisol remains constant or even increases with age, levels of DHEA decrease gradually from the third decade, reaching levels that are only 10–20% of their maximum by the eight decade; aged-related decrease of DHEA correlates with IL-6 overproduction. The elderly thus gradually approach a status of relative glucocorticoid excess with age, with a potential negative impact on baseline immune function and an exaggerated response to stressors. The correlation between DHEA and IL-6 has been reported in several studies. In particular, it has been demonstrated that in normal subjects, serum IL-6 levels increased with age, whereas serum DHEA concentrations decreased. Furthermore, serum IL-6 levels were inversely correlated with serum DHEA levels. In addition, in *in vitro* experiments, DHEA inhibited the lipopolysaccharide-stimulated IL-6 production from monocytes and peripheral blood mononucleated cells, even in a narrow concentration range.

Inflamm-Aging

As previously described, the major consequence of chronic exposure to antigenic stress is the progressive activation of macrophages and related cells in most organs and tissues of the body, leading to a progressive chronic, low-grade proinflammatory status which is the principal characteristic of the aging process. We called this phenomenon 'inflamm-aging,' and we postulated that it contributes substantially to the pathogenesis of all major age-related diseases, such as cardiovascular diseases, diabetes type 2, dementia, and cancer, and also to complex pathological conditions very common and characteristic of elderly people, such as frailty and sarcopenia. Thus, directly or indirectly, inflamm-aging is a major determinant of morbidity and mortality in the elderly. However, inflammatory status *per se* is compatible with extreme longevity in good health, and paradoxically proinflammatory characteristics have been documented in healthy centenarians, suggesting that inflammation is a physiologic phenomenon and a beneficial response to cope with chronic stressors. We also found evidence that centenarians activate a strong and effective anti-inflammatory response mediated by cytokines such as IL-10 and T cell transforming growth factor-β capable of counteracting their inflammatory status. Moreover, the rate at which the threshold of proinflammatory status is reached and over which disease/disabilities emerge and the individual's ability to adapt to stressors can be considered a complex trait with high interindividual variability and a strong genetic component. We propose that the persistence over time of inflammatory stimuli, such as CMV chronic infection, represents the biological background favoring susceptibility to a variety of diseases. The absence of robust gene variants and/or the presence of frail gene variants is necessary to develop overt organ-specific, age-related disease. We predicted that the concomitant presence of high inflammatory status and inherited variants of frailty genes or the absence of robust genes can explain why some individuals, but not others, are more susceptible to certain types of age-related disorders.

Inflammation and the Neuroendocrine System

Major Depression

The proinflammatory status in elderly people is associated with several inflammatory-based pathologies, such us cardiovascular disease, sarcopenia, and Alzheimer's disease. Moreover, studies suggest that proinflammatory cytokines play an important role

in the pathophysiology of major depression. This psychiatric disorder, at a neurobiological level, is associated with alterations in neurotransmitter function involving serotonin, norepinephrine, and dopamine. Hypersecretion of CRH and alteration of its regulation by glucocorticoids are also described. Regarding the emerging role of the immunological pathway in major depression, several studies have shown that patients with this pathology, who are otherwise medically healthy, have increased serum and plasma concentrations of IL-6 and/or C-reactive protein and elevated IL-1β and TNF-α in peripheral blood and cerebrospinal fluid. Additionally, increased levels of chemokines and adhesion molecules, such as E-selectin, have been described. A positive correlation between plasma concentrations of various inflammatory mediators and depressive symptom severity has been reported. Finally, a nascent literature suggests that functional allelic variants of genes for IL-1β and TNF-α increase the risk for depression and are associated with reduced responsiveness to antidepressant therapy.

Given the primary factors that are associated with the pathophysiology of depression, it is likely that peripherally released inflammatory cytokines can access the brain. Because of the large size of cytokines and their resultant inability to readily penetrate the blood–brain barrier, attention has focused on routes by which they access the brain, including through leaks in the blood–brain barrier, the binding to specific transport, or the activation of vagal afferent fibers which transmit cytokine signals to specific brain nuclei. Once in the brain, there is a CNS cytokine network made up of cells that not only produce cytokines and express cytokine receptors but also amplify cytokine signals, which in turn can have profound effects on neurotransmitter and CRH function, as well as behavior. These cytokine-induced behavioral changes are associated with alterations in the metabolism of serotonin and dopamine in brain regions essential for the regulation of emotion and psychomotor function. Moreover, these cytokines induce the activation of the HPA axis hormone, as well as CRH in both the hypothalamus and the amygdala, regions that have an important role in fear and anxiety. Finally, NF-κB induction in the brain may also contribute to alterations in neuronal growth and survival, especially through the induction of nitric oxide production and, ultimately, oxidative stress. The discovery that depression shares common features with inflammation and stress system activation led to a novel perspective to treatment. Indeed, patients with a history of nonresponse to antidepressants have been found to exhibit increased plasma concentration of IL-6 compared to treatment-responsive patients. Moreover, antidepressants may only address select symptoms (e.g., depressed mood and anxiety), whereas others are relatively unaffected. Thus, inhibition of proinflammatory cytokine signaling represents a viable strategy for the treatment of depression, especially in patients with evidence of increased inflammatory activity who might be less likely to respond to convectional agents.

Anorexia Nervosa

Prolonged elevation of proinflammatory cytokines may induce a cascade of biochemical effects in the CNS, involving secretion of neurotransmitters, neuropeptides, neurohormones, and peripheral hormones, which may be etiopathogenetic for the development of anorexia nervosa. Indeed, in experimental animal models, IL-1β, IL-6, and TNF-α introduced directly in the CNS induced anorexia, fatigue, apathy, depressed mood, and impaired secretion of catecholamines, serotonin, and HPA axis hormones. Although there are no reports regarding the cerebrospinal fluid concentration of such cytokines in patients affected by anorexia nervosa, we performed a study on a group of anorectic subjects (restricted and bingeing–purging) measuring plasma concentrations of IL-1β, IL-6, TNF-α, and their soluble receptors. Surprisingly, we found that the concentrations of the three cytokines were normal in our patients, as well as the concentrations of the TNF-α soluble receptors. Interestingly, the levels of the IL-6 soluble receptor were lower than normal in our patients. Since there was a positive correlation between the levels of cytokine and its soluble receptors, the low level of receptors contrasts with the normal levels of IL-6. The biological significance of the soluble IL-6 receptor is not clear, but an impairment of the proinflammatory cytokine pathway might be involved in the development of anorexia nervosa.

Alzheimer's Disease

One of the most important age-associated pathologies is Alzheimer's disease (AD), a heterogeneous and progressive neurodegenerative form of dementia. Neuropathological hallmarks of AD are neuronal and synapsis loss, extracellular β-amyloid deposition (neuritic/senile plaques), and intracellular deposition of hyperphosphorylated tau-based paired helical filaments in hippocampus, temporal and frontal cortex. Inflammation clearly occurs in this pathologically vulnerable region of the AD brain, where the principal proinflammatory stimuli are highly insoluble Aβ42 and Aβ40. Indeed, these peptides are recognized by astrocytes and microglia, which in turn produce many cytokines such as IL-1, IL-6, and TNF-α. In particular, TNF-α acts as a proapoptotic agent,

which contributes to neuronal loss in AD. Moreover, the senile plaques are potent activators of the complement cascade; during this process, small fragments, known as anaphylotoxins, are produced and stimulate inflammation. Chemokines and acute phase proteins are also co-localized as secondary components in neuritic plaques and are overproduced in brain of the AD patient. Other important mediators of inflammation in AD are prostaglandins, a group of fatty acids derived from the precursor arachidonic acid, with COX catalyzing the initial step. Both the constitutive COX-1 and the inducible COX-2 mRNA, as well as their protein products, are upregulated in affected brain areas. COX-1 is mainly expressed in microglia, whereas COX-2 is abundantly expressed in pyramidal neurons and seems to play a role in synaptic plasticity. Epidemiological studies have shown that levels of proinflammatory mediators, particularly cytokines such as IL-6, IL-10, and TNF-α, in blood and brain may be under genetic control and the risk of AD is influenced by several polymorphisms in the noncoding region of these genes. Thus, the overall chances of an individual developing AD might be profoundly affected by a susceptible proinflammatory profile, reflecting the combined influence of inheriting multiple high-risk alleles. The identification of such a profile might in the future lead to strategies for therapeutic intervention in the very early stages of the disorders. Indeed, one possible strategy to suppress inflammatory response would be the administration of nonsteroidal anti-inflammatory drugs (NSAIDs). Epidemiological investigations indicate that the use of NSAIDs decreases the risk of AD by approximately 50% and delay the onset of AD by 5–7 years. (R)-flurbiprofen, the R enantiomer of racemic flurbiprofen, is under license for the potential treatment of AD, modulating the gamma secretase and the production of Aβ42. Indeed, in a murine model of AD, it lowered the level of this toxic peptide, and chronic administration reduced brain amyloid pathology and prevented defects in learning and memory. In a phase II clinical trial in AD, patients demonstrated a benefit in cognitive and behavioral performance. Two pivotal phase III clinical trials in AD are under way.

Frailty Syndrome

Another aged-related disorder, involving innate immunity and inflammation, is frailty and sarcopenia. Frailty has been defined as a decline in lean body mass and decreased muscle strength, endurance, balance, and walking performance, accompanied by a high risk of disability, falls, hospitalization, and mortality. This syndrome reflects a profound metabolic imbalance due to overproduction of catabolic cytokines (likely TNF-α) and the diminished availability of anabolic hormones such as insulin-like growth factor-1 (IGF-1), resulting from aging and the presence of associated chronic conditions. Sarcopenia, the loss of muscle mass and strength that occurs with normal aging, is a central part of the frailty syndrome, probably due the decline of neural, hormonal, and environmental trophic signals to muscle. The most

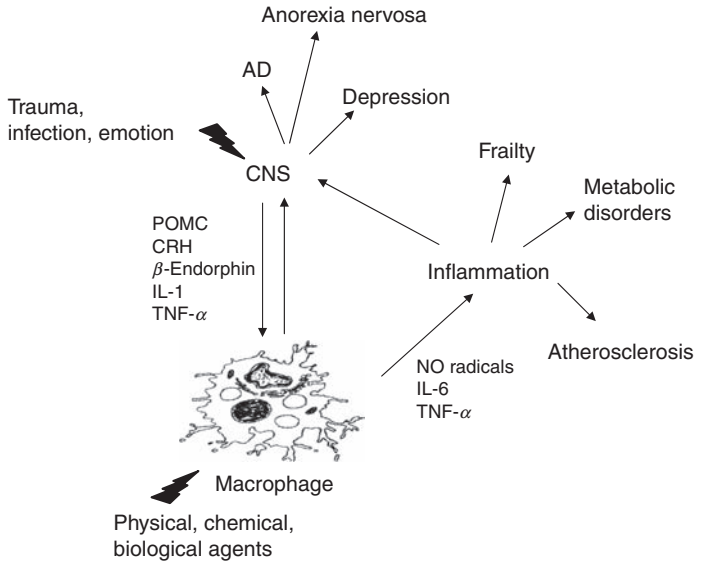

Figure 1 The central role of the macrophage in stress response, inflammation, and age-related diseases. The immune and neuroendocrine systems share common mediators to exchange information and they are affected by different stressors. The macrophage is the most important actor in integrating information, and its chronic activation in aging supports the development of inflammatory diseases.

important endogenous cause may be the irreversible age-related loss of α-motor units in the central nervous system. TNF-α has effects that may contribute directly to sarcopenia, including increased basal energy expenditure, and anorexia, and it is associated with wasting/cachexia in chronic inflammation disorders. In the CNS, TNF-α can drive neuronal apoptosis, as previously described for AD. The role of IL-6 in sarcopenia is not clear, even though the effect of IL-6 in the regulation of IGF-1 action has been widely documented. To better understand the involvement of this circuit in this complex condition, we evaluated the relationship of plasma concentrations of IGF-1 and IL-6 with muscle function in apopulation-based sample of elderly people. We found that higher IL-6 and lower IGF-1 plasma levels are associated with lower muscle strength and power, suggesting a multifactorial relationship between these two mediators. Thus, insulin resistance in the elderly and loss of insulin anabolic action on muscle are factors favoring the onset and progression of this pathology.

Conclusion

In this article, we discussed the most important link between the immune system and the neuroendocrinal circuit. These three systems form a complex network, conserved during evolution, in which cytokines and neuropeptides are the mediators. The most important actor which regulates this network is the macrophage, a very ancestral cell that is able to integrate immunity, stress response, and inflammation (**Figure 1**). Aging is characterized by a proinflammatory status which supports the most important aged-related pathologies.

See also: Hypothalamic–Pituitary–Adrenal (HPA) Axis; Immune System–Neuroendocrine Interactions; Neuroendocrine Aging: Pituitary Metabolism; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Neuroendocrine Aging: Pituitary-Adrenal Axis; Stress, Cytokines and Depressive Illness; Thermoregulation: Autonomic, Age-Related Changes.

Further Reading

Blalock JB (2005) The immune system as the sixth sense. *Journal of Internal Medicine* 257: 126–138.

Butcher SK and Lord JM (2004) Stress responses and innate immunity: Aging as contributory factor. *Aging Cell* 4: 151–160.

De Martinis M, Franceschi C, Monti D, and Ginaldi L (2005) Inflamm-aging and lifelong antigenic load as major determinants of ageing rate and longevity. *FEBS Letters* 579: 2035–2039.

Franceschi C and Bonafè M (2003) Centenarians as model of healthy aging. *Biochemical Society Transactions* 31: 457–461.

Franceschi C, Bonafè M, and Valensin S (2000) Human immunosenescence: The prevailing of innate immunity, the failing of clonotypic immunity, and the filling of immunological space. *Vaccine* 18: 1717–1720.

Franceschi C, Bonafè M, Valensin S, et al. (2000) Inflamm-aging. An evolutionary perspective of immunosenescence. *Annals of the New York Academy of Sciences* 908: 208–218.

Franceschi C, Valensin S, Bonafè M, et al. (2000) The network and the remodelling theories of aging: Historical background and new perspectives. *Experimental Gerontology* 35: 879–896.

Glaser R and Kiecolt-Glaser JK (2005) Stress-induced immune dysfunction: Implications for health. *Nature Reviews Immunology* 5: 243–251.

Harding G, Mak YT, Evans B, et al. (2006) The effects of dexamethasone and dehydroepiandrosterone (DHEA) on cytokines and receptor expression in a human osteoblastic cell line: Potential steroid-sparing role for DHEA. *Cytokine* 36(1–2): 57–68.

Jacob F (1977) Evolution and tinkering. *Science* 196(4295): 1161–1166.

Licastro F, Candore G, Lio D, et al. (2005) Innate immunity and inflammation in ageing: A key role for understanding age-related diseases. *Immunity & Aging* 18: 2–8.

Ottaviani E and Franceschi C (1996) The neuroimmunology of stress from invertebrates to man. *Progress in Neurobiology* 48(4–5): 421–440.

Ottaviani E and Franceschi C (1997) The invertebrate phagocytic immunocyte: Clues to a common evolution of immune and neuroendocrine systems. *Immunology Today* 18(4): 169–174.

Ottaviani E, Franchini A, and Franceschi C (1997) Pro-opiomelanocortin-derived peptides, cytokines, and nitric oxide in immune responses and stress: An evolutionary approach. *International Review of Cytology* 170: 79–141.

Ottaviani E, Valensin S, and Franceschi C (1998) The neuroimmunological interface in an evolutionary perspective: The dynamic relationship between effector and recognition system. *Frontiers in Bioscience* 3: 431–435.

Pawelec G, Solana R, Remarque E, and Mariani M (1998) Impact of aging on innate immunity. *Journal of Leukocyte Biology* 64: 703–712.

Neuroinformatics

G A Ascoli and M Halavi, George Mason University, Fairfax, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Successes and Challenges of Modern Neuroscience

The foundations of neuroinformatics as a new multidisciplinary field of science were laid in the late 1980s as advances in both neuroscience and information science made their incorporation possible. Relating the complex structures and functions of the nervous system requires coordination among diverse domains of knowledge, integration across multiple levels of investigation, and fusion of seemingly disparate technical approaches, from molecules to behavior. The challenge of neuroinformatics is to provide a unified computational information framework to enable, facilitate, and foster such an enterprise. In practice, major advances in our understanding of the brain require the development and application of suitable electronic tools to handle, represent, transform, analyze, and synthesize digital neuroscience data. In turn, such a challenge is the key solution to preventing, diagnosing, and treating brain diseases.

Neuroinformatics is still in its infancy, and it may be premature to attempt a historical perspective on a field in evolving and dynamic development. In this article, we present an overview of the status of neuroinformatics mostly by means of focused examples. We first state the main areas of neuroinformatics research in broad terms. Next, we describe some of the principal institutional initiative and resources, exemplifying both organizations and funding programs. We then offer several examples of ongoing efforts. This selection is not meant to (and could not possibly) provide a comprehensive account of current neuroinformatics; on the contrary, we wish to illustrate the diversity and breadth of this rapidly growing branch of science. To offer a sense of the depth involved in these programs, we present one specific case our own laboratory is contributing. Finally, we close by venturing a forecast of the future research landscape after the full maturation and blossoming of neuroinformatics.

By its own nature, much neuroinformatics content exists on the World Wide Web and in fact mostly consists of its very online presence. It is, thus, impossible (and would be counterproductive) for this review to be devoid of direct references to uniform resource locators (URLs) corresponding to the described information; we therefore include an extensive list of websites. Internet content evolves dynamically, and a static text such as this can provide only a temporary snapshot of journalistic value, destined to age quickly. The same applies to the illustrations, which are all derived from material accessed in April 2007.

Neuroinformatics: Premise and Promise

Neuroscience studies all aspects of the nervous system and its development to understand the biological principles of the mind and behavior. Modern neuroscience encompasses many diverse areas of investigation: from the brain's shape and cellular structure (neuroanatomy) to molecular manipulations and chemical composition (neurochemistry) and from bioelectrical plasticity and circuit dynamics (neurophysiology) to the personal and social organization of cognitive systems (psychology). Heterogeneous and complex results of such specialized domains require proper integration and analysis. How can firing patterns enable slow, inaccurate computational elements (synapses and nerve processes) to perform fast and accurate computation (logical reasoning)? Questions such as this call for the application of proper mathematical methods, computational models, analytical tools, and numerical simulations.

Completion of the Human Genome Project has clearly evidenced the usefulness of structured databases in biomedical researches. However, compared to genome sequences, neuroscience data are generally more interconnected, detailed, and multidimensional. Naturally, such complexity makes neuroscience data very difficult to deal with. These same challenging features, however, make the application of theoretical and computational models to neuroscience data particularly effective and powerful.

Neuroinformatics efforts can be broadly ascribed to four frontiers:

- Defining consistent knowledge environments and constructing web-based information management systems.
- Promoting and enabling data sharing and providing network-accessible, interoperable databases.
- Developing analysis, visualization, and simulation tools for and across all major neuroscience data types.
- Implementing accurate computational models for subcellular components, neurons, networks, and neuromorphic artifacts.

It is important to stress that informatics tools are already commonly and widely employed in neuroscience and that increasing data complexity requires correspondingly more sophisticated tools. In neuroimaging, the only available approach to the study of the intact human brain in both physiological and pathological conditions, for example, information technology plays crucial roles in all stages of data manipulation. Informatics resources are used to design the experiments; control imaging instruments; minimize noise; reconstruct 3D images; quantify, analyze, and compare the results; and organize, store, query, retrieve, and share the data.

Complementary to the technological aspects lies the social facet of neuroinformatics. One of the main goals is to create an optimal environment for researchers to collaborate, communicate more effectively, and share their data. Neuroinformatics brings together investigators from a wide variety of basic and clinical disciplines, along with computational scientists and engineers. Despite a common interest in understanding brain development and function, such diverse cultural backgrounds make a productive dialog one of the most critical elements for the acceptance of common standards.

Institutional Resources and Initiatives

The US National Institutes of Health (NIH) formally established the Human Brain Project (HBP) in 1993 to develop and support the new science of neuroinformatics (Figure 1(a)). All approaches and technologies under this initiative were to be generalizable, scalable, extensible, and interoperable. The success of the HBP is evidenced by the range of digital resources it supported, from molecular biology to functional magnetic resonance imaging (fMRI) databases (Figure 1(b)). The HBP was also instrumental in nurturing the growth of the neuroinformatics community through annual public research conferences, regular meetings of the funded principal investigators, and sponsored satellite symposia and social gatherings at major scientific events. Several of the resources and initiatives described in this section, as well as the projects and databases exemplified in the next section, resulted directly or indirectly from the HBP.

The Society for Neuroscience (SfN), the largest worldwide scientific organization devoted to studying the nervous system, assembled the Brain Information Group (BIG) in 2003. One of the BIG recommendations was to establish and maintain an active role in informing the neuroscience community about available digital tools and resources. The SfN established a standing Neuroinformatics Committee and, recognizing the growing importance and number of electronic archives and repositories, launched the Neuroscience Database Gateway (NDG). The NDG provides a one-stop resource for neuroscientists, increasing awareness of and facilitating access

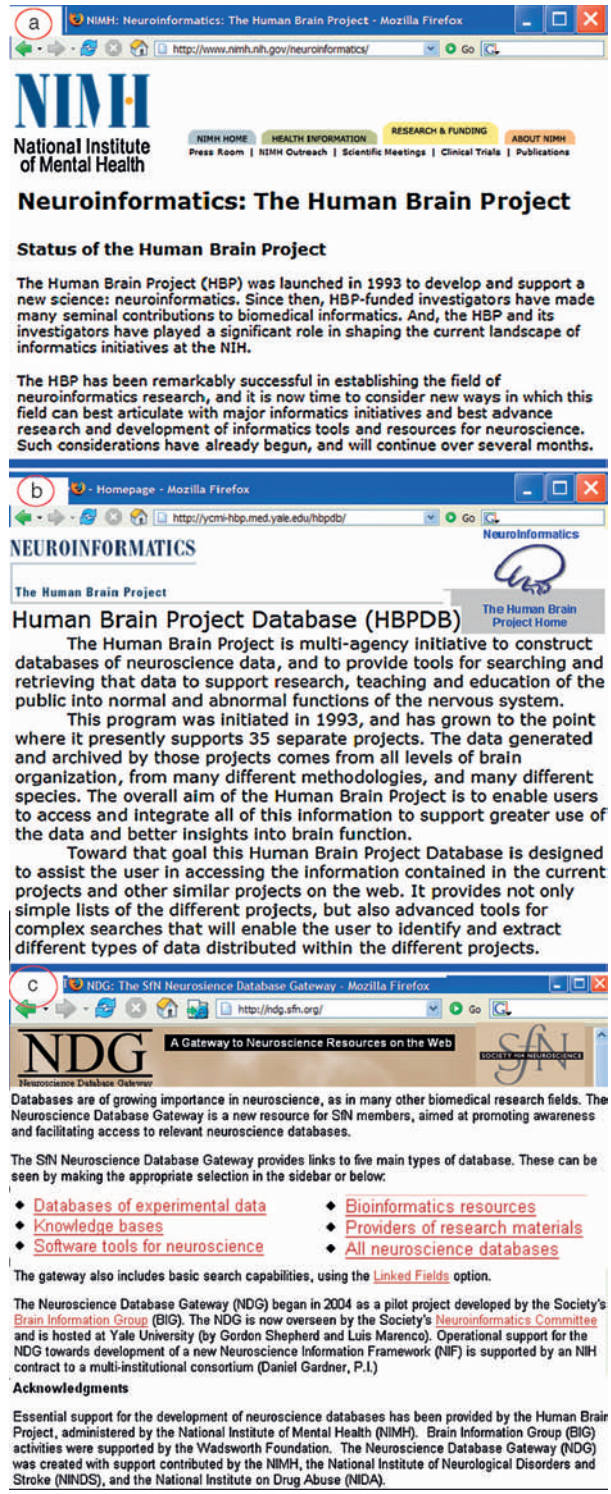

Figure 1 Institutional resources: (a) National Institute for Mental Health (NIMH) Neuroinformatics: Human Brain Project; (b) Human Brain Project database (HBPDB) at Yale; (c) Society for Neuroscience (SfN) Neuroscience Database Gateway (NDG). Collated from material available in April 2007.

to free tools and available databases (Figure 1(c)). In early 2007, the NDG listed more than 175 entries, including databases of experimental data, knowledge bases, software tools for neuroscience, bioinformatics resources, and providers of research materials. Other online indices include Neuroguide, Neuroinf.org, and the Internet Analysis Tools Registry.

The NIH Blueprint for Neuroscience Research was established in 2004 to develop new tools, resources, and training opportunities to facilitate discovery in brain science. One of the first projects sponsored under the Blueprint was the Neuroscience Information Framework (NIF), which was endorsed by the SfN as an expansion and continuation of the NDG. The goal of NIF is to integrate, provide access to, and facilitate the utilization of neuroinformatics resources. In addition to the partnership of five academic institutions (California Institute of Technology; George Mason University; University of California at San Diego; Yale University; and the Weill Medical College of Cornell University, which is leading the project), the NIF consortium relies on a network of nine additional sites as volunteer consultants and collaborators. The project NeuroMorpho.org, highlighted as a case example in a later section, is part of the NIF program.

In 2000, the Biomedical Information Science and Technology Initiative Consortium (BISTIC) was established at NIH as the focus for making optimal use of computer science and technology in biology and medicine. The BISTIC is composed of senior-level representatives from NIH centers and institutes as well as from other federal agencies concerned with bioinformatics and computational applications. Notable initiatives in this regard are also constituted by the National Centers for Biomedical Computing (NCBC), created to build the computational infrastructure of biomedical computing, and the Biomedical Informatics Research Network (BIRN), established by the NIH Center for Research Resources (NCRR). A geographically distributed virtual community of shared resources, BIRN, aims to enhance communication and promote collaboration across research disciplines.

Two NIH institutes, the Institute of Mental Health and the Institute of Neurological Disorders and Stroke, jointly sponsored the Neuroimaging Informatics Technology Initiative (NIfTI) to provide coordinated and targeted service, training, and research to speed the development and enhance the use of informatics tools related to neuroimaging. Among the practical results of NIfTI is the standardization of imaging file formats (under the same name), which was adopted by academia and the corporate world alike.

The International Neuroinformatics Coordinating Facility (INCF) was formed in 2005 on the recommendations of the Organisation for Economic Cooperation and Development (OECD) Mega Science Forum Working Group on Biological Informatics (1999) and the OECD Global Science Forum Working Group on Neuroinformatics (2002). Endorsed by 17 countries (Australia, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, India, Italy, Japan, the Netherlands, Norway, Sweden, Switzerland, the United Kingdom and the United States), as well as by the European Commission, the INCF aims to coordinate key networking activities and to provide services and tools for identifying, analyzing, and drawing on knowledge from main neuroscience information resources all around the world.

Finally, the journal *Neuroinformatics*, now in its fifth year of publication, is the first scientific periodical dedicated entirely to the information science infrastructure of neuroscience.

A Sampler of Individual Projects and Databases

Neuroscience occupies a major and growing position in many general bioinformatics resources, including (but not limited to) genetic sequence databases, protein structure repositories, metabolic and regulatory networks, and microrray and expression pattern archives. Several bioinformatics developments, however, are dedicated entirely to or explicitly emphasize brain science. For instance, neuromolecular efforts to investigate the biochemical basis of neurological disorders have a relatively long-standing online presence, such as Neuromuscular information on ion channels, transmitters, receptors, and related diseases; the Ligand-Gated Ion Channel Database of nucleic and proteic sequences of channel subunits; and the Inherited Peripheral Neuropathies Mutation Database, which was made public in May 1999. A few selected examples of mature neuroinformatics resources in other domains are collected here.

The National Institute of Neurological Disorders and Stroke (NINDS) sponsors the Gene Expression Nervous System Atlas (GENSAT) for genetic mappings of the mouse central nervous system in different stages of development. Data in GENSAT are generated using radiolabeled riboprobes, *in situ* hybridization, and transgenic mouse techniques. In a similar vein, the Allen Institute developed the Brain Atlas as a freely available scientific resource to provide expression maps of the approximately 20 000 genes in the mouse brain (Figure 2(a)). After the completion of the sequencing of the mouse genome, and through the systematic automation of high-throughput techniques to probe gene expression, the goal of creating a detailed cellular-resolution, genomewide map of gene expression in the mouse brain was achieved.

The HBP sponsored several projects in related scientific domains, three of which we mention here. The Mouse Brain Library (MBL), a database of brains and high-resolution images from many genetically characterized rodent strains, aims to systematically map and characterize the key genes for the architecture of the mammalian central nervous system (Figure 3(a)). BrainMaps.org is an online high-resolution digital brain atlas implementing interactive zoomable technology (virtual microscope). Queriable and retrievable data include complete sets of scanned images from serial sections of both primate and non-primate brains (Figure 3(b)). BrainInfo helps identify brain structures and related information (Figure 3(c)) based on three distinct knowledge bases that provide the index to, relations among, and location of brain regions in the primate brain (NeuroNames, the Template Atlas, and NeuroMaps, respectively). Other knowledge bases include CoCoDat, a microcircuitry database with bibliographic references and parameter values from published experimental reports, and Textpresso, an NIF project to extract and process information from the neuroscience literature.

One of the efforts aimed at integrating disparate information at the molecular, cellular, and system level, the Blue Brain Project (BBP) was jointly established in 2005 by IBM and the Ecoles Polytechniques Fédérales de Lausanne in Switzerland (Figure 2(b)). By using the Blue Gene supercomputer, the BBP uses a strategy of reverse engineering the brain by modeling more details than ever before attempted, ultimately with large-scale bottom-up simulations. In particular, the BBP is directed at creating a microcircuit of the neocortical column, with accurate biological properties

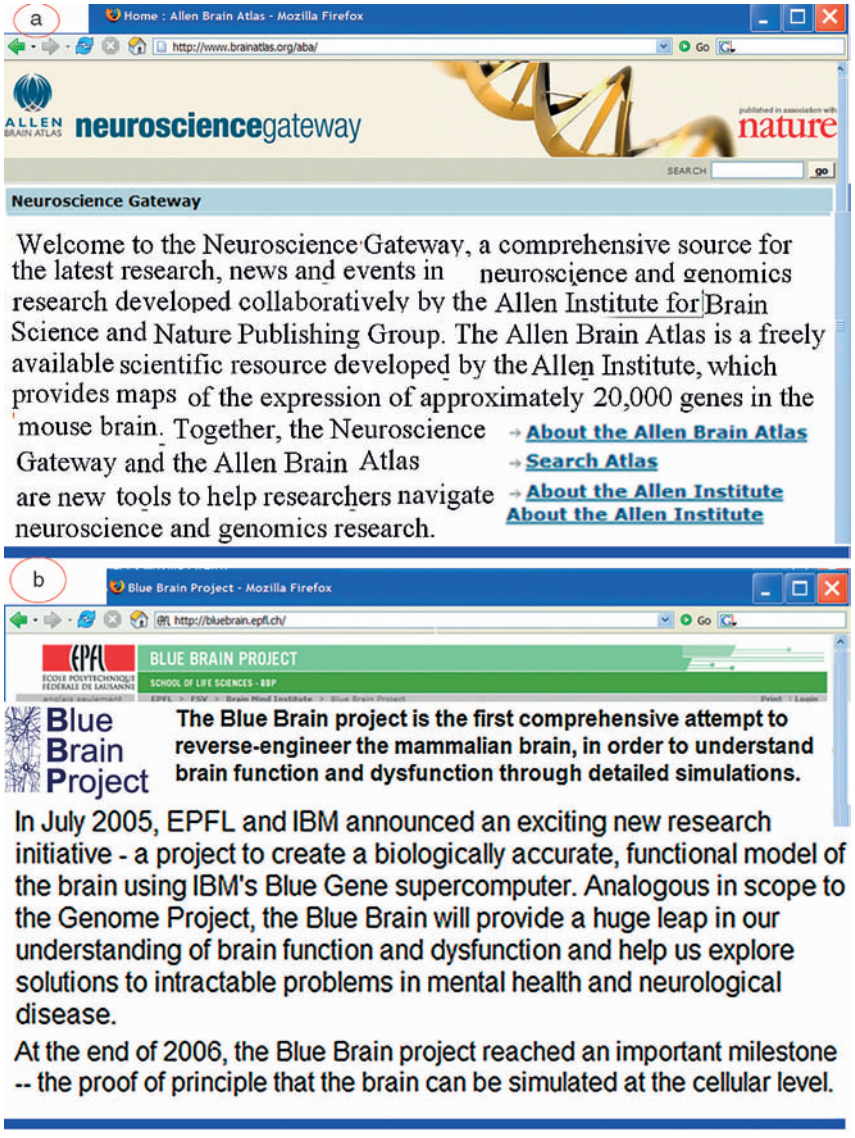

Figure 2 Projects with corporate support: (a) Allen Brain Atlas; (b) Blue Brain Project. Collated from material available in April 2007.

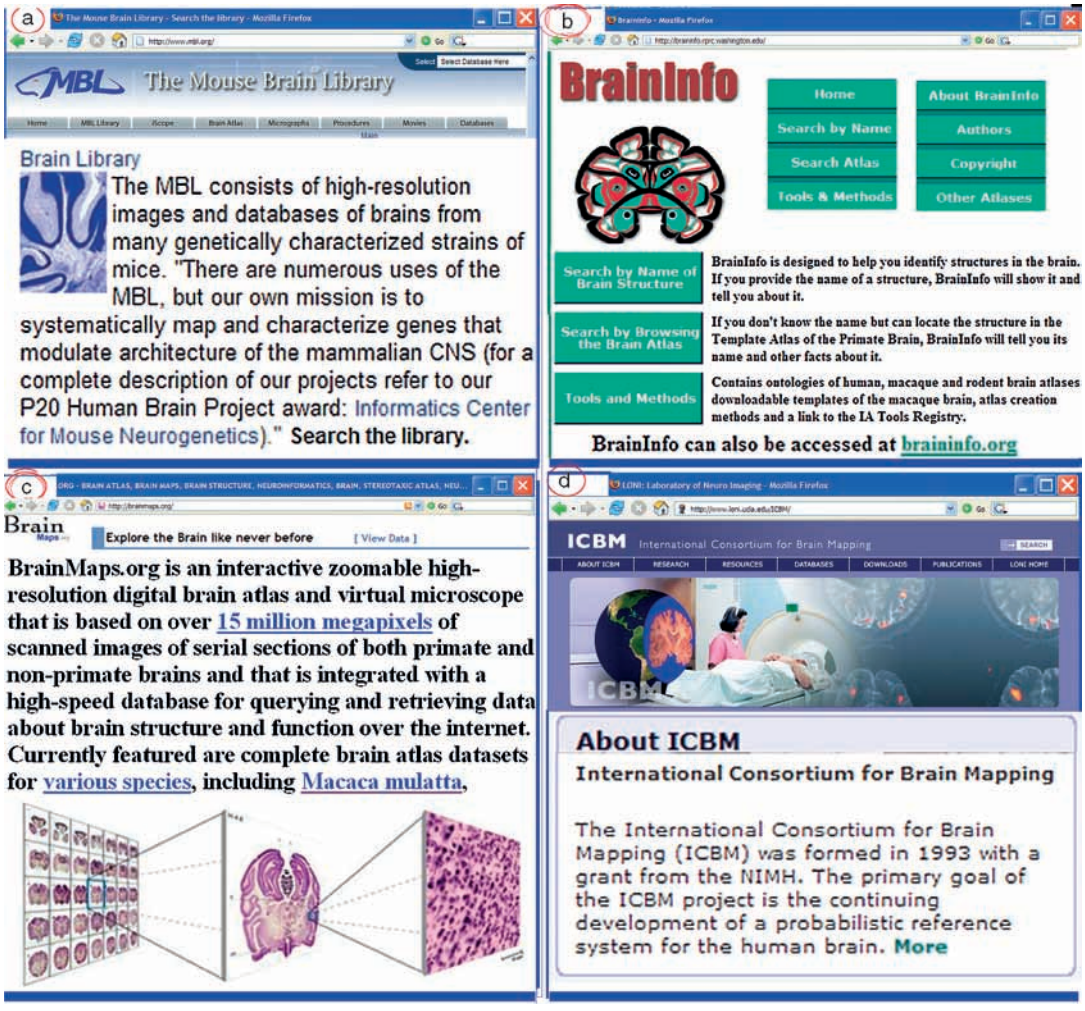

Figure 3 Academic initiatives: (a) Mouse Brain Library; (b) BrainInfo; (c) BrainMaps; (d) International Consortium for Brain Mapping. Collated from material available in April 2007.

(including membrane biophysics and synaptic connectivity) of individual neurons from the mammalian brain classified based on their morphology, electrophysiology, and gene expression. A longer-standing effort funded by the HBP, SenseLab consists of a collection of six databases supporting experimental and theoretical research on membrane properties and the olfactory pathway. For example, NeuronDB provides information regarding voltage-gated conductances and neurotransmitter receptors, whereas ModelDB collects computational neuroscience models which users can search and reuse. Mature simulation environments for individual neurons, their subcellular components and biochemical reactions or neural networks, and systems-level models include NEURON, GENESIS, and PDP++.

The HBP funded several neuroimaging projects. The International Consortium for Brain Mapping (ICBM), formed in 1993, is composed of four core research sites – the University of California at Los Angeles, Montreal Neurologic Institute, University of Texas at San Antonio, and the Institute of Medicine of Juelich/Heinrich Heine University (Germany) – with data acquisition sites from Asia and Europe also contributing. The main ICBM goal is the continued development of probabilistic reference systems for the human brain (Figure 3(d)). ICBM built and expanded neuroinformatics tools for data sharing and to map structural and functional anatomy both macroscopically (*in vivo*) and microscopically (post-mortem). An allied resource, the Alzheimer's Disease Neuroimaging Initiative (ADNI), was started in 2004 to define the rate of progress of mild cognitive impairment and improve the current methods for clinical and treatment trials in Alzheimer's disease by building a large neuroimaging database.

Another important neuroimaging project is the Functional Magnetic Resonance Imaging Data Center (fMRI-DC), started in 1999 with grants from the National Science Foundation (NSF) and the

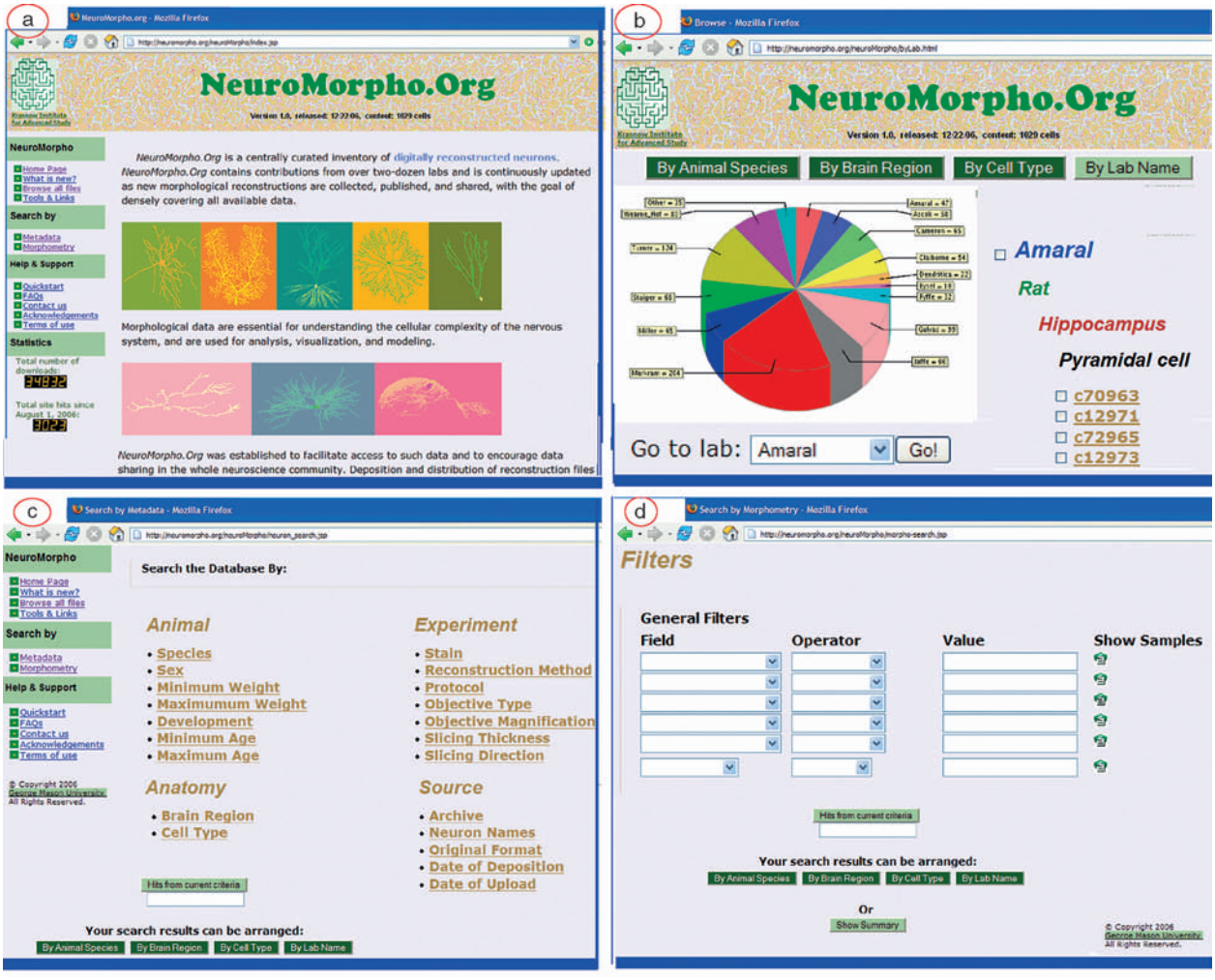

Figure 4 NeuroMorpho.org: (a) home page; (b) browse all cells screen; (c) search by metadata screen; (d) search by morphometry screen.

W. M. Keck Foundation. The fMRI-DC archive serves as a public repository for data from studies published in the peer-reviewed literature. Uniquely in the brain sciences, the *Journal of Cognitive Neuroscience*, one of the leading venues for scientific reports based on fMRI, mandates that the data underlying manuscripts accepted for publication be made available for deposition in the fMRI-DC. The establishment of similar policies in other domains of biology (e.g., protein structures and gene sequences) has greatly benefited the explosive success of bioinformatics, and these standards also apply to other fields of science such as physics (e.g., astronomical and particle collision experiments).

NeuroMorpho.org: Opportunities and Solutions for Cellular Structure

NeuroMorpho.org is one of the projects piloted within the NIF (Figure 4(a)). A centrally curated web-based inventory of digitally reconstructed neurons, NeuroMorpho.org, aims at densely archiving and organizing all publicly available 3D reconstructions of neuronal morphology. This effort started in 2006 with contributions from over two dozen labs and is continuously updated. Dendritic and axonal arborizations influence neuronal integration and excitability and therefore overall brain activity. Morphological studies and neuronal reconstructions are increasingly popular, both for quantitative modeling of neuronal electrophysiology or network dynamics and in response to advances in genetic and physiological studies, providing a rapidly expanding pool of knowledge. Management of this exploding flow of information requires appropriate informatics tools, and effective data sharing is indispensable. In this sense, NeuroMorpho.org has significant promise both as a new resource and as a tool, which could become a model for other neuroscience data types. Archival in NeuroMorpho.org protects hard-earned

valuable data files from being lost without leveraging their full scientific and educational potential. The provided morphometric and stereological measures also enable a cascade of anatomical analyses and cellular classifications.

With over 1000 entries, NeuroMorpho.org is to date the largest online collection of digital reconstructions in cellular neuroscience. Data sets were gathered by mining available public databases and by direct peer-to-peer requests to researchers willing to share their data. Neurons are included from different species and strains, ages and developmental stages, experimental protocols, brain regions, and so on, thus providing a one-stop resource for neuronal morphology regardless of the individual investigator's interest (Figure 4(b)). Such diversity could foster comparative studies between neurons, for example, in various brain areas or animals or reconstructed from different histological and microscopic techniques. Users also gain the opportunity to familiarize themselves with the works and approaches of other researchers.

Data in NeuroMorpho.org are extensively annotated to allow efficient browsing and searching (Figure 4(c)). Fast navigation through a rich data selection has important scientific consequences; for example, it aids computational modelers in determining and accessing the most appropriate morphological data available to implement more accurate and precise biophysical simulations of neural activity and function. Morphometric parameters (including arbor size, branching angles, tree topology, and fractal dimension) are also precomputed to facilitate stereological studies, and the investigation of the structure–activity relationship or the effect of morphology on electrophysiology and plasticity (Figure 4(d)).

Each data file is classified with respect to brain regions and cell types based on the corresponding scientific report. NeuroMorpho.org adopts three hierarchical levels for brain region, starting with the main area (e.g., the neocortex, hippocampus, and cerebellum), further refining an intermediate level (the somatosensory vs. prefrontal cortex, CA3 vs. CA1, etc.), and finishing with the detailed cytoarchitectonic layer (such as layers I–VI or alveus through lacunosum-moleculare). Cell types are similarly divided into principal neurons and interneurons, then organized into common types (pyramidal, granule, stellate, basket, chandelier, etc.), and finally categorized into more specialized subtypes as described in the corresponding scientific reports. Reconstructions are also accompanied by useful information extracted from the articles' Materials and Methods (or equivalent) sections, including the species, strain, age, sex, weight, and development of the animal; reconstruction method; slicing thickness and direction; staining; imaging magnification and objective type; experiment protocol; and other experimental details. A PubMed link is also provided to enable users to look for other specifications and to identify all co-authors contributing to the reconstruction files.

In addition to a 2D image of the cell, NeuroMorpho.org renders neurons growing and revolving as a dynamic 3D display. A user-friendly neuron viewer has been integrated, allowing interactive visualization and manipulation of the reconstruction file (zooming, panning, turning, etc.) as well as conversion to common electrophysiology simulation environments such as NEURON and GENESIS. Users can download both the source reconstruction files (as provided by the original contributors) and a standardized integrity-checked and -corrected version in a common nonproprietary format. All conversion changes as well as notes of recognized residual irregularities are listed in additional log files, also optionally bundled for data retrieval.

With the expected future improvement of algorithms for automated reconstruction, the number of available digital morphologies will rapidly increase, exacerbating the need for central curation and ease of access. An important advantage of NeuroMorpho.org is its simplification of the data-sharing process, whereby any researchers willing to make their files available can simply send them to the database administrator without any technical hardship or storage-space investment while maintaining (and in fact gaining) recognition and full credit for their work.

The Future: Explosive Potential of Digital Neuroscience

Databases, visualization and analysis tools, and simulation environments have already impacted all fields of brain science, from molecular to cognitive, in both basic research and clinical applications. Computational skills have become an essential component of the neuroscientist's expertise. Nevertheless, neuroinformatics is still in its early stages. Ongoing progress, which will mature in the near future, includes the development of formal neuroinformatics vocabulary and effective neurontologies, that is, data models defining and interrelating neuronal concepts to maximize communication and interoperability. Technical advances will lead to the fall of the current social barrier to open and pervasive data sharing.

What will the typical neuroscience investigation of tomorrow look like? A massive amount of available data will be accompanied by powerful computational tools for analysis, integration, and modeling. In this scenario, the complexity of neuroscience will be an

opportunity for discovery and an asset for invention. Brain researchers will be able to design and execute sophisticated experiments at a computer terminal, developing, testing, and refining hypotheses, mechanisms, and theories. The ability to bridge diverse subdomains of neuroscience, transcending techniques and integrating heterogeneous data types, will become more important than specialized wet-bench expertise in one or few instruments and protocols. Neuroscientists will be one step closer to the direct acquisition, manipulation, and creation of knowledge.

Of course, neuroinformatics will never substitute for, but will, rather, integrate, experimental neuroscience. Primary data will continue to be needed, and newly synthesized or isolated chemicals will have to be tested *in vivo*. Neuroinformatics will certainly aid in the design of molecules and primary experiments alike, and the extracted data will, in turn, fuel and stimulate further neuroinformatics progress. In the life science, bioinformatics has already grown to occupy a prominent role in the spectrum of biological approaches. In neuroscience, the same complexity that constitutes a challenge today will lead to a future in which the vast majority of research will be carried out on archived, rather than primary, data.

See also: Cognitive Neuroscience: An Overview; Computational Approaches to Motor Control; Computational Methods; Computational Neuroethology; Consciousness: Theoretical and Computational Neuroscience; Executive Function and Higher-Order Cognition: Computational Models; Hippocampus: Computational Models; Memory: Computational Models.

Further Reading

Arbib MA and Grethe JS (2001) *Computing the Brain: A Guide to Neuroinformatics*. San Diego, CA: Academic Press.

Ascoli GA, Schutter ED, and Kennedy DN (2003) An information science infrastructure for neuroscience. *Neuroinformatics* 1(1): 1–2.

Gardner D and Shepherd GM (2004) A gateway to the future of neuroinformatics. *Neuroinformatics* 2(3): 271–274.

Gardner D, Toga AW, Ascoli GA, et al. (2003) Toward effective and rewarding data sharing. *Neuroinformatics* 1(3): 289–295.

Koslow SH and Huerta MF (1997) *Neuroinformatics: An Overview of the Human Brain Project*. Mahwah, NJ: Lawrence Erlbaum.

Koslow SH and Subramaniam S (2005) *Databasing the Brain: From Data to Knowledge* (*Neuroinformatics*). New York: Wiley.

Kötter R (ed.) (2003) *Neuroscience Databases: A Practical Guide*. Norwell, MA: Kluwer Academic Publishers.

Richmond B (2004) Editorial commentary. *Journal of Computational Neuroscience* 17(1): 5.

Shephered GM (2002) Supporting databases for neuroscience research. *Journal of Neuroscience* 22(5): 1497.

Relevant Websites

http://www.brainatlas.org – Allen Brain Atlas.
http://www.nbirn.net – Biomedical Informatics Research Network.
http://www.bisti.nih.gov – Biomedical Information Science and Technology Initiative Consortium.
http://bluebrain.epfl.ch – Blue Brain Project.
http://braininfo.rprc.washington.edu – BrainInfo.
http://www.sfn.org – Brain Information Group.
http://brainmaps.org – BrainMaps.org.
http://www.cocomac.org – CoCoDat.
http://www.fmridc.org – Functional Magnetic Resonance Imaging Data Center (fMRI-DC).
http://www.genesis-sim.org – GENESIS.
http://www.ncbi.nlm.nih.gov – GENSAT.
http://www.nimh.nih.gov – Human Brain Project.
http://www.molgen.ua.ac.be – Inherited Peripheral Neuropathies Mutation Database.
http://www.loni.ucla.edu – International Consortium for Brain Mapping.
http://www.incf.org – International Neuroinformatics Coordinating Facility.
http://www.cma.mgh.harvard.edu – Internet Analysis Tools Registry.
http://www.ebi.ac.uk – Ligand-Gated Ion Channel Database.
http://www.mbl.org – Mouse Brain Library.
http://www.nigms.nih.gov – National Centers for Biomedical Computing (NCBC).
http://www.neuroguide.com – Neuroguide.
http://nifti.nimh.nih.gov – Neuroimaging Informatics Technology Initiative.
http://www.neuroinf.org – Neuroinf.org.
http://neuromorpho.org – NeuroMorpho.org.
http://www.neuro.wustl.edu – Neuromuscular.
http://www.neuron.yale.edu – NEURON.
http://ndg.sfn.org – Neuroscience Database Gateway.
http://neurogateway.org – Neuroscience Information Framework.
www.neuroguide.com – Neurosciences on the Internet.
http://www.nia.nih.gov – NIH Alzheimer's Disease Neuroimaging Initiative.
http://www.ncrr.nih.gov – NIH Center for Research Resources.
http://www.cnbc.cmu.edu – PDP++.
http://senselab.med.yale.edu – Senselab.
http://www.textpresso.org – Textpresso.

Neuroleptics

J M Stone and L S Pilowsky, King's College London Institute of Psychiatry, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Schizophrenia is a devastating mental illness, usually striking individuals in their late teens or early twenties. It is characterized by a significant change in normal perception, with affected individuals suffering from hallucinations (most commonly auditory but sometimes affecting other sensory modalities), delusions (fixed false beliefs, often concerning elaborate plots in which the sufferer is being persecuted or that their actions are being controlled by an external entity), and impairment of cognitive ability, as well as the so-called negative symptoms of apathy, avolition, loss of emotional reactivity, and withdrawal from social contact. These symptoms are termed 'psychotic,' being qualitatively different from normal experience (cf. neurotic symptoms). Until the 1950s, no effective treatments for schizophrenia existed. The discovery that the drug chlorpromazine was effective in reducing the psychotic symptoms of schizophrenia heralded the start of a new era in the management of patients with the illness.

Chlorpromazine was not originally developed as an antipsychotic. It was first marketed as an antiemetic in the early 1950s. It's effects, leading to sedation without anesthesia, were first noted by the surgeon Henri Laborit. When given to patients with schizophrenia, as well as causing mild sedation, chlorpromazine also led to reductions in hallucinations and delusions, allowing individuals to become more able to function on a normal level. Following treatment with chlorpromazine, patients who previously had to be housed in mental institutions were able to move out into less restrictive settings.

The success of chlorpromazine in treating schizophrenia led to a huge interest in finding other drugs with a similar action against psychotic symptoms. Between 1952 and the mid-1970s, more than 40 first-generation antipsychotics (FGAs) had been developed, but it was some time before the specific mechanism of action of antipsychotic drugs began to emerge.

Antipsychotic Affinity for the Dopamine Receptor

In the mid-1970s, it was discovered that all drugs with antipsychotic effects shared the common property of high affinity for the dopamine D2 receptors (including D_2, D_3, and D_4 subtypes), and that, furthermore, the affinity of the drugs for the D2 receptors directly correlated with clinical potency (i.e., the higher their affinity for dopamine D2 receptors, the lower the dose of drug was required to have a clinical effect).

One consequence of this, thought to be unavoidable, was that all antipsychotic drugs led to extrapyramidal (Parkinsonian) side effects (EPSEs). The primary deficit in Parkinson's disease is a loss of dopamine in the striatum (caudate/putamen; see Figure 1). Antipsychotic drugs mirrored this effect by blocking dopamine receptors in the striatum. As a result, patients with schizophrenia, despite partial resolution of their psychotic symptoms, were often left with severe problems with abnormal movements, shuffling gait, and reduced facial expression. Furthermore, antipsychotic drugs were not helpful to all patients, with approximately one-third showing no improvement in symptoms despite treatment. Many psychiatrists concluded that patients who did not respond well to antipsychotic treatment might be taking insufficient doses of antipsychotic drug, causing a therapeutic occupancy of dopamine D2 receptors in the brain. As a result, patients who did not respond to antipsychotic drugs were often prescribed very high doses, with severe EPSEs as a common outcome. This practice was not universal, however, and some suggested that very high doses of antipsychotic drugs might not be the best method for tackling treatment-resistance.

Studies of Antipsychotic Action in Living Patients

The development and use of single-photon emission tomography (SPET) and positron emission tomography (PET) as research tools in psychiatry led to a reevaluation of the role of D2 receptors in antipsychotic drug action. SPET and PET are two methods by which neuroreceptor density can be measured in living humans. Both SPET and PET work by measuring the regional distribution of an injected radioactive tracer (radioligand) in the body. If the radioligand has been designed to bind to specific neuroreceptors, the regional density of these receptors can then be estimated.

Patients with schizophrenia who were good or poor responders to antipsychotic drugs were scanned using labeled (*S*)-*N*-[(1-ethyl-2-pyrrolidinyl)]methyl-2-hydroxy-3-iodo-6-methoxybenzamide ([^{123}I]IBZM), a SPET ligand for striatal D2 receptors, and [^{18}F] *N*-methylspiroperidol, the equivalent PET ligand.

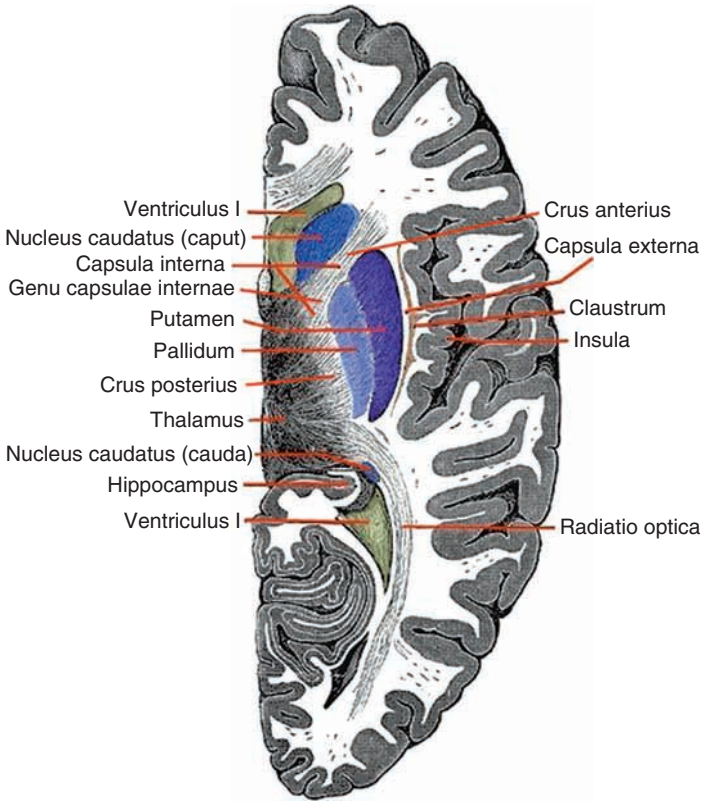

Figure 1 Transverse section of one hemisphere of human brain showing three main subregions of striatum: caudate nucleus (nucleus caudatus), putamen, and globus pallidus (pallidum). Reproduced from *Grays Anatomy of the Human Body*, 20th edn. (1918).

No difference was found, in terms of striatal D2 receptor occupancy, between patients with a good or a poor clinical response to antispychotic treatment. Indeed, some patients who had not responded to antipsychotic treatment had close to 100% occupancy of striatal dopamine receptors. This suggested that striatal occupancy was not sufficient to explain antipsychotic response in all cases. It was also found that that 90% striatal dopamine receptor occupancy could be obtained at very low doses of FGAs. This meant that there was no logical reason to continue giving treatment-resistant patients very high doses of FGAs, since the same receptor occupancy could be obtained with lower doses, causing less severe EPSEs, and leading to a real benefit in the quality of life for such patients.

More recently, there have been studies of patients who were good responders to FGAs, or patients with first-episode schizophrenia (who generally show a good initial response to antipsychotic drugs). These studies found that greater striatal dopamine D2 receptor occupancy by FGAs predicted a better response to antispychotic treatment. This suggests that there is a subset of patients with schizophrenia in whom striatal dopamine D2 receptor occupancy is an important aspect of antipsychotic efficacy. It should be noted, however, that striatal D2 occupancy in these patients might simply be a marker for dopamine receptor occupancy in other parts of the brain.

Second-Generation Antipsychotics

Further progress in antipsychotic development after chlorpromazine did not occur until the reintroduction of clozapine in the early 1990s. Clozapine was first marketed in the 1970s but was withdrawn from widespread use because it led to a serious reduction in white cells (agranulocytosis) in a small proportion of patients, which caused several deaths. Despite this, several psychiatrists in Europe continued using it in patients who did not develop agranulocytosis, as it appeared to have greater efficacy than did other antipsychotic drugs, and furthermore, it did not have as much propensity to cause EPSEs. In 1988, Kane and colleagues performed a double-blind study of clozapine compared to chlorpromazine in treatment-resistant patients with schizophrenia. These patients had been treated with other existing antipsychotic drugs and had shown no significant response. In the study, one-third of these patients showed a marked response to clozapine and did not develop EPSEs. As a result of this study, clozapine was reintroduced as

an antipsychotic drug, with compulsory blood monitoring requirements. It was found that if patients had regular blood monitoring, and were permitted only to continue on clozapine while their white blood cell count remained within a healthy range, the risk due to agranulocytosis could be virtually eliminated.

Following on from the success of clozapine, a number of 'second-generation' antipsychotic drugs (SGAs) were developed. These sought to emulate the enhanced efficacy of clozapine in treatment-resistant patients, and also to minimize the extrapyramidal side effects traditionally associated with FGA treatment. In general, they were successful in having lower EPSE potiential, compared to FGAs, but none had the same enhanced efficacy in treatment resistance as seen with clozapine, with most sharing equal clinical effectiveness with the FGAs. Because of their lower propensity to cause EPSEs as compared to FGAs, the SGAs have also been termed 'atypical' antipsychotics. Recently there has been evidence from meta-analysis of available literature that some SGAs may have improved efficacy compared to FGAs, although clozapine was still found to be superior to any other available antipsychotic drug.

Mechanisms of Atypicality

The development of the SGAs led to interest in how they might obtain their beneficial antipsychotic effect without EPSEs. Initial studies used SPET and PET to try to answer this question. Pilowsky and colleagues studied striatal occupancy of dopamine D2 receptors by clozapine using [^{123}I]IBZM SPET to compare dopamine D2 receptor binding in patients treated with clozapine and patients treated with FGAs. Farde and colleagues performed the same study using [^{18}F]raclopride (a PET ligand for striatal D2 receptors). Both groups found that clozapine achieved an antipsychotic effect with a much lower occupancy of striatal dopamine D2 receptors than seen with FGAs. This unexpected result conflicted with conventional wisdom that striatal dopamine receptor occupancy was required for antipsychotic efficacy, although it gave some explanation as to why clozapine had lower propensity than FGAs did to cause EPSEs. This led to speculation as to what the mechanism of antipsychotic action of SGAs might be, and how they might avoid EPSEs.

5HT2A Occupancy

The finding that many SGAs shared the characteristic of a high affinity for 5-hydroxytryptamine (serotonin) 2A (5HT2A) receptors led to speculation that the ratio between 5HT2A and dopamine D2 receptor occupancy could be a key feature leading to lower propensity to EPSEs. Initial PET and SPET studies confirmed that many SGAs showed very high 5HT2A receptor occupancy at all doses. A later analysis revealed that SGAs could induce EPSEs with striatal D2 receptor occupancies greater than 78%, suggesting that it was the lower striatal D2 occupancy by SGAs, rather than any interaction with 5HT2A receptors, that was the cause of their lower EPSE potential. The fact that amisulpride, an SGA with no affinity for the 5HT2A receptor, also had low EPSE potential further suggested that 5HT2A receptor occupancy was not necessary for an atypical antipsychotic profile.

Extrastriatal Selectivity

Due to low D2 receptor densities in regions outside the striatum, it was not possible, with the first PET and SPET D2 receptor ligands, to determine whether SGA occupancy of extrastriatal dopamine receptors differed from their striatal occupancy, although it was hypothesized that dopamine D2 receptor occupancy in limbic regions of the cortex (temporal lobes) could be more important in antipsychotic action than was blockade of striatal dopamine D2 receptors. It was not until the development of high-affinity ligands for D2 receptor that this question could be addressed. The first SPET study comparing clozapine with haloperidol binding in striatal and temporal cortex revealed that clozapine appeared to preferentially bind cortical over striatal D2 receptors, while haloperidol showed equal occupancy of both cortical and striatal D2 receptors (**Figure 2**). Subsequent studies examined the binding characteristics of other SGAs and found that other SGAs, including olanzapine, quetiapine, risperidone, amisulpride, and sertindole, appeared to share the property of cortical over striatal D2 receptor selectivity, while FGAs did not show any preferential occupancy (**Figure 3**). Importantly, drugs with modest affinity for dopamine D2 receptors exhibited this effect more robustly across their whole dose range, where atypical drugs with higher affinity for D2 receptors (e.g., risperidone, amisulpride) displayed dose-dependent limbic selectivity. There is a nonreplication of these data in the literature and full *in vivo* PET or SPET confirmation awaits further study. Nevertheless, regionally selective dopaminergic action of atypical antipsychotic drugs is supported by both electrophysiological and animal studies, and the finding fits well with the theory that cortical, rather than striatal, dopamine receptors are the primary site for antipsychotic efficacy.

Intrastriatal Selectivity

The striatum has roles beyond its involvement in motor function. It may be divided into three distinct

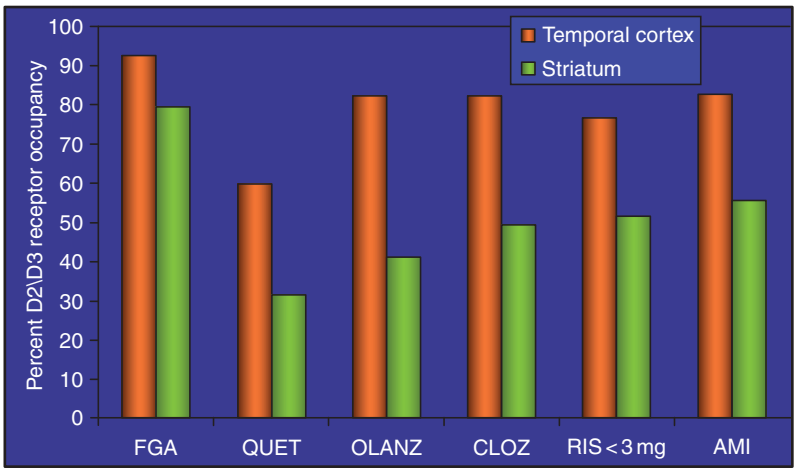

Figure 2 Average percent occupancy of striatum and temporal cortex in patients with schizophrenia treated with first-generation antipsychotic (FGA) drugs, quetiapine (QUET), olanzapine (OLANZ), clozapine (CLOZ), low-dose risperidone (RIS < 3 mg) and amisulpride (AMI).

functions, which are approximately defined by anatomical boundaries. The ventral parts of the caudate and putamen are limbic in function, being involved in emotional processing and salience. The caudate nucleus (particularly the anterior and lateral portions) is involved in associative learning, and the dorsomedial caudate and putmen are involved in the integration of motor function. These functional distinctions can be observed in the differences in symptomatology in two neurological conditions affecting different striatal regions. Parkinson's disease primarily affects the putamen, and individuals with the disease have a marked difficulty in initiating movements, as well as having the classic tremor and masklike facies that accompany the illness. Huntington's disease, on the other hand, shows greater deterioration of the caudate nucleus and affected individuals have more pronounced problems with associative learning than do persons with Parkinson's disease. Stone and colleagues studied the occupancy of two pharmacologically distinct SGAs in substriatal regions and found that both showed preferential occupancy of head of caudate nucleus over striatum (Figure 3). This finding has been replicated by Grunder and colleagues, who reported similar caudate selectivity of clozapine and amisulpride. Intrastriatal selectivity, although a relatively recent discovery, may also contribute toward the lack of EPSEs propensity in SGAs. Further work is required to confirm this finding in other SGAs and to determine whether FGAs also show caudate selectivity.

How Might Atypical Drugs Show Extrastriatal and Intrastriatal Selectivity?

Kapur and Seeman noted that the strongest single predictor of a drug's status as an SGA was a high dissociation constant (k_{off}). They hypothesized that, in regions where there was low competition for receptors by endogenous dopamine (such as cortical brain regions), drugs with a low k_{off} would bind readily. In striatal brain regions, where the highest levels of dopamine in the brain are found, they suggested that SGAs would not bind so readily, as they would compete poorly with endogenous dopamine. They hypothesized that this would lead SGAs to show preferential binding to cortical brain regions over the striatum. It is interesting to note that, within the striatum, anterolateral caudate and ventral regions are reported to have lower dopamine concentrations than are found in the putamen, suggesting that SGAs would preferentially bind to caudate over putamen, as has been shown using SPET and PET.

Dopamine Abnormalities in Schizophrenia

Because of the striking relationship between antipsychotic drug dopamine D2 receptor affinity and clinical potency, it was hypothesized that the major neurochemical abnormality in schizophrenia might be an elevation in dopamine transmission. Postmortem studies appeared to support this hypothesis, with the demonstration of elevated dopamine D2 receptor density in striatum in patients with chronic schizophrenia. These findings were potentially confounded by the long-term use of antipsychotic medication, which may have altered dopamine receptor expression. The development of PET and SPET allowed the study of dopamine D2 receptor density in unmedicated living patients with schizophrenia. The first study appeared to support the hypothesis of elevated D2 receptor density in the striatum in patients with schizophrenia, but the methodology of this study was

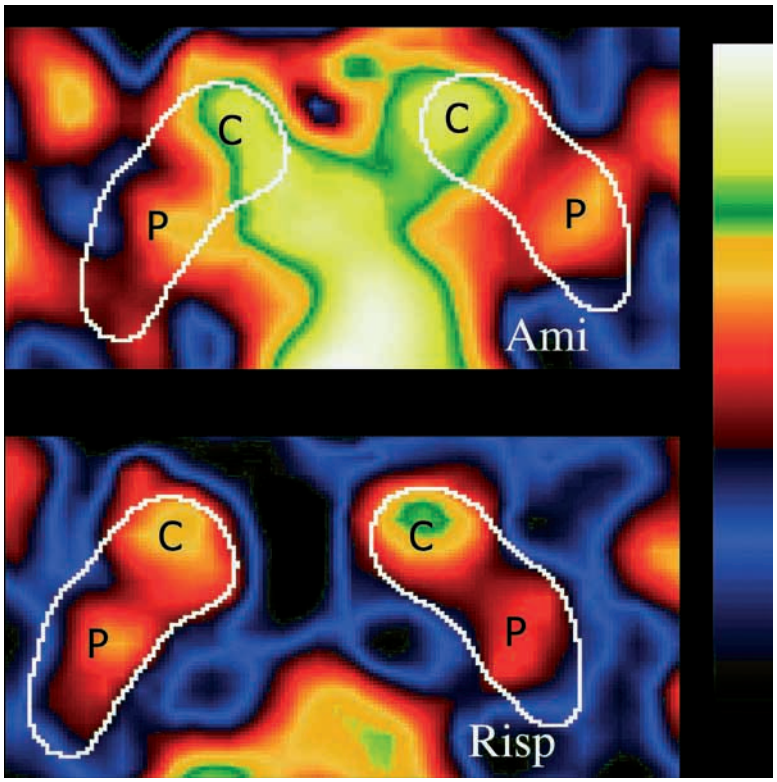

Figure 3 Occupancy map of dopamine D2 receptors for patients with schizophrenia treated with amisulpride (Ami) and low-dose risperidone (Risp). Lighter colors represent greater antipsychotic drug occupancy of dopamine receptors. An outline of the striatum has been overlaid and heads of caudate (C) and putamen (P) have been labeled. Both amisulpride and risperidone appear to preferentially bind to dopamine D2 receptors in head of caudate over head of putamen. Reproduced from Stone JM, Bressan RA, Erlandsson K, et al. (2005) Non-uniform blockade of intrastriatal D2/D3 receptors by risperidone and amisulpride. *Psychopharmacology (Berlin)* 180(4): 664–669, with kind permission of Springer Science and Business Media.

criticized. Later studies generally failed to show any difference in D2 receptor density in patients with schizophrenia and in controls. Two recent meta-analyses suggest that there may be a small elevation in D2 receptor density in schizophrenia, although there is significant overlap with controls.

The basically negative finding with regard to striatal dopamine D2 receptors in schizophrenia led groups to look for other evidence of dopamine dysregulation in the illness. Laruelle and colleagues used IBZM SPET to study dopamine release following amphetamine administration. Since endogenous dopamine displaces the binding of IBZM, it was possible to estimate the magnitude of dopamine release by comparing IBZM binding before and after amphetamine was given. Patients with schizophrenia were found to have increased dopamine release following amphetamine challenge compared to healthy controls. In subsequent studies, it was found that this dopamine release correlated with increase in positive psychotic symptoms.

Recent studies have extended analysis of dopamine receptors to regions outside the striatum. These have reported either low or normal D2 receptor density in patients compared to controls. Correlations between dopamine receptor binding and psychotic symptoms have also been found. Symptoms were found to correlate directly with D2 receptor binding in frontal cortex, and to correlate negatively with D2 receptor binding in temporal cortex and thalamus, suggesting that extrastriatal dopamine receptors may play an important modulatory role in psychotic symptom formation.

Negative Symptoms

Although positive psychotic symptoms of schizophrenia (hallucinations, delusions, disorders in flow of thought) generally show a response to antipsychotic drug treatment, the so-called negative symptoms (avolition, apathy, emotional blunting, self-neglect) do not appear so amenable to treatment. Some studies have shown that negative symptoms improve, to a certain extent, in line with positive symptoms, but others have not found any effect of conventional antipsychotic drugs on negative symptoms and, in most cases, there appear to be residual negative symptoms that are refractory to current drug treatment.

There are several theories about the basis for negative symptoms in schizophrenia, but the generally accepted theory is that they relate to hypofunction of the frontal cortex that is found in schizophrenia. There is some evidence that this frontal hypofunction may be driven by a reduction in dopaminergic transmission (compared to the elevation of striatal dopamine transmission found following amphetamine administration in patients with schizophrenia). Since antipsychotic drugs act primarily as dopamine antagonists, it can be seen that they would be expected to worsen negative symptoms. In fact, it has been shown that although patients with schizophrenia suffer an exacerbation of positive symptoms when taking amphetamine, their negative symptoms may even improve, possibly due to increased frontal dopamine release, with amphetamine.

Dopamine-Stabilizing Drugs

The newest development in currently available antipsychotic drugs in schizophrenia is the release of first 'dopamine-stabilizing' antipsychotic, aripiprazole. Aripiprazole has a very high affinity for the dopamine D2 receptor but, in contrast to other antipsychotic drugs, which are inverse agonists at the dopamine D2 receptor, aripiprazole is considered a weak partial agonist. The theory behind the use of partial agonists for the dopamine receptor as antipsychotic drugs is that they should be able to treat both positive and negative symptoms, by reducing activation of dopamine receptors where endogenous dopamine levels are high (through competition for the binding site), with the aim of treating positive symptoms, and increasing activation of dopamine receptors where dopamine levels are low (through intrinsic agonist activity), theoretically improving negative symptoms. So far, aripiprazole is the only partial dopamine D2 agonist that has been marketed as an antipsychotic drug. Other partial agonists showed initial antipsychotic efficacy, which was lost after repeated dosing, and recent work suggests that aripiprazole shows functional selectivity at the D2 receptor, with different levels of agonism depending on the signaling pathway activated, rather than acting as a simple partial agonist. Disappointingly, aripiprazole has not proved to be markedly more effective than other antipsychotic drugs are, but an extensive study on its effects in patients with severe negative symptoms has yet to be performed.

The Future of Antipsychotic Treatment in Schizophrenia

Recently, the search for effective antipsychotic drugs has extended beyond the dopamine system. The finding that drugs blocking the *N*-methyl-D-aspartate (NMDA) system, such as ketamine and phencyclidine (PCP), induce symptoms closely resembling those in schizophrenia (including both positive and negative symptoms) led to the suggestion that NMDA receptor function could be abnormal in schizophrenia. Drugs aiming to enhance NMDA receptor function, given in tandem with SGAs or FGAs, have generally shown modest improvements in negative symptoms, with little effect on reducing positive psychotic symptoms.

Acetylcholine transmission at the muscarinic receptor has also been implicated as being abnormal in schizophrenia. A SPET study of unmedicated first-episode patients with schizophrenia found a significant decline in binding to muscarinic receptors compared to healthy controls. Furthermore, both clozapine and its main active metabolite, *N*-desmethylclozapine, are partial agonists with very high affinity for M1 and M4 muscarinic receptors. Muscarinic partial agonists have also shown some promise as potential antipsychotic agents in preclinical studies.

Last, there has been interest in developing drugs acting at non-D_2 subtypes of the dopamine receptor. D1 receptor agonists have been suggested as being helpful for treating cognitive deficits and negative symptoms in schizophrenia, and preferential antagonists at the D_3 dopamine receptor subtype have been put forward as potential antipsychotic agents with fewer side effects compared to drugs with greater affinity for the D_2 subtype.

Conclusions

Despite recent research suggesting novel targets for antipsychotic drugs, D2 receptor antagonism remains an essential aspect of any medication for schizophrenic symptoms. Little progress in terms of antipsychotic efficacy has been made since the discovery of chlorpromazine, although SGAs have a much lower propensity to cause EPSEs than do FGAs. Clozapine still remains the most effective antipsychotic available, but the basis for its therapeutic superiority remains an enigma. Unraveling its actions will provide leads for novel therapies, with the hope of more effective, less toxic treatments for schizophrenia.

See also: Antipsychotic Drugs; Basal Ganglia: Functional Models of Normal and Disease States; Cognition: An Overview of Neuroimaging Techniques; Cognitive Deficits in Schizophrenia; Cognitive Dysfunction in Psychiatric Disorders; Dopamine; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Parkinsonian

Syndromes; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin-Related Psychedelic Drugs.

Further Reading

Abi-Dargham A, Gil R, Krystal J, et al. (1998) Increased striatal dopamine transmission in schizophrenia: Confirmation in a second cohort. *American Journal of Psychiatry* 155: 761–767.

Abi-Dargham A and Moore H (2003) Prefrontal DA transmission at D1 receptors and the pathology of schizophrenia. *Neuroscientist* 9(5): 404–416.

Farde L, Wiesel FA, Stone-Elander S, et al. (1990) D2 dopamine receptors in neuroleptic-naive schizophrenic patients. A positron emission tomography study with [^{11}C]raclopride. *Archives of General Psychiatry* 47: 213–239.

Grunder G, Landvogt C, Vernaleken I, et al. (2005) The striatal and extrastriatal D2/D3 receptor-binding profile of clozapine in patients with schizophrenia. *Neuropsychopharmacology* 31: 1027–1035.

Kane J, Honigfeld G, Singer J, et al. (1988) Clozapine for the treatment-resistant schizophrenic. A double-blind comparison with chlorpromazine. *Archives of General Psychiatry* 45(9): 789–796.

Kapur S, Remington G, Jones C, et al. (1996) High levels of dopamine D2 receptor occupancy with low-dose haloperidol treatment: A PET study. *American Journal of Psychiatry* 153(7): 948–950.

Kapur S and Seeman P (2001) Does fast dissociation from the dopamine d(2) receptor explain the action of atypical antipsychotics?: A new hypothesis. *American Journal of Psychiatry* 158: 360–369.

Kercuin and Pilocosky LS (1995) Traditional receptor theory and its application to neuroreceptor measurements in functional imaging. European journal of Nuclear Medicine 22(7): 699–710.

Laruelle M (1998) Imaging dopamine transmission in schizophrenia. A review and meta-analysis. *Quarterly Journal of Nuclear Medicine* 42: 211–221.

Lidow MS, Williams GV, and Goldman-Rakic PS (1998) The cerebral cortex: A case for a common site of action of antipsychotics. *Trends in Pharmacological Sciences* 19(4): 136–140.

Pilowsky LS, Costa DC, Ell PJ, et al. (1994) D2 dopamine receptor binding in the basal ganglia of antipsychotic-free schizophrenic patients. An ^{123}I-IBZM single photon emission computerised tomography study. *British Journal of Psychiatry* 164: 16–26.

Pilowsky LS, Mulligan RS, Acton PD, et al. (1997) Limbic selectivity of clozapine. *Lancet* 350: 490–491.

Seeman P and Lee T (1975) Antipsychotic drugs: Direct correlation between clinical potency and presynaptic action on dopamine neurons. *Science* 188(4194): 1217–1219.

Stone JM, Bressan RA, Erlandsson K, et al. (2005) Non-uniform blockade of intrastriatal D2/D3 receptors by risperidone and amisulpride. *Psychopharmacology (Berlin)* 180(4): 664–669.

Tandon R (1999) Cholinergic aspects of schizophrenia. *British Journal of Psychiatry* (supplement 37): 7–11.

Neuroligins

P Taylor, A De Jaco, and D Comoletti, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Neuroligin Genes and Structures

Individual neuroligins have extremely high sequence conservation across mammalian species (~99%), whereas the amino acid identity among individual neuroligins within the family is approximately 52% (Figure 1). Their globular, esterase domain has the highest sequence conservation (~72%) and the intracellular domain the lowest (~31%). Neuroligins are also related to a series of tactin proteins found in *Drosophila*: neurotactin, gliotactin, and glutactin.

Structurally, neuroligins contain four distinctive domains: (1) an N-linked glycosylated dimeric extracellular domain of approximately 630 amino acids; this domain belongs to the α,β-hydrolase fold and, as such, is homologous to many serine hydrolases, such as the cholinesterases; (2) a heavily O-linked glycosylated stalk region of approximately 50 amino acids that resembles a bottlebrush-like structure and links the above globular domain to the membrane; (3) a single transmembrane span; and (4) a cytoplasmic domain lacking a defined structure with a known fold but containing a PDZ recognition domain enabling interaction with scaffolding proteins on the intracellular side of the membrane. Neuroligin binds to the PDZ domain of several scaffolding proteins, including PSD95, SAP102, S-SCAM, Magi1 and -3, Shank1 and -3, Pick1, and GOPC. A dendritic-targeting motif is present in the cytoplasmic domain of neuroligin-1, distinct from the classical PDZ recognition motif.

With the four previously discussed domains, neuroligins have the capacity to serve as a heterophilic adhesion protein to dock and anchor proteins localized within the synaptic gap and, in turn, link with scaffolding molecules on the intracellular side of the cell membrane. The primary identified partner extending from the presynaptic membrane, neurexin, also has its synaptic recognition domain connected through an extendable stalk through the membrane, where its PDZ recognition domain interacts with PDZ motif proteins in the cytoplasm of the nerve ending.

The structure of the dimeric extracellular domain of the neuroligin molecule has been described at intermediate resolution through a combination of homology modeling with the crystal structures of the cholinesterases, low-angle X-ray and neutron scattering, and hydrodynamic analyses (Figures 2 and 3). The core structure is the neuroligin dimer stabilized through a four-helix bundle characteristic of many α,β-hydrolase fold proteins. Whereas hydrodynamic data indicate a 2:2 stoichiometry for the neuroligin–neurexin association, small-angle scattering studies establish the two neurexin sites to be on opposite sides of the long axis of the dimer near the neuroligin subunit interface. The O-glycosylated, linking sequence is also detected, extending tangentially from the dimer structure. The site of neurexin association appears distinct from the active center gorge of the cholinesterases, revealing that the adhesion and catalytic functions of the α,β-hydrolase fold proteins take place in discrete domains (Figure 3).

Recognition of the Neurexins

The neuroligin–neurexin system has built in a hierarchy for the molecular recognition code originating in the biosynthesis and processing steps. Neuroligin possesses as many as five genes (neuroligin-1 to -5), and whereas neuroligin-1, with its two alternative splice inserts (A and B), can encode for four splice variants, neuroligins-2 to -5 only show alternative splicing at position A (Figure 1). Neuroligins, encoded by the different genes and splice variants of neuroligin-1, show a differential affinity for the neurexins. Although the β-neurexins possess a higher affinity for the neuroligins, it has been shown that α-neurexin exhibits significant and graded interactions with the neuroligins. Finally, glycosylation and its processing also affect neuroligin affinity for neurexin in a negative manner. One of the influential glycosylation sites in neuroligin-1 is on an alternatively spliced region B. Given that both neuroligin and neurexin are encoded by multiple, alternatively spliced genes and they may be differentially expressed in various tissues, a complex code likely exists for the maintenance and formation of synapses. Moreover, it has been shown that neuroligin-1, -3, and -4 are associated with excitatory amino acid synapses in the central nervous system (CNS), whereas neuroligin-2 is associated with inhibitory synapses.

Neuroligins in Development and Disease

The importance of neuroligin in the formation and development of synapses was dramatically demonstrated in cell-based studies in which both neuroligins and neurexins presented on the surface of nonneuronal cells induced the formation of functional pre- and postsynaptic components involving partnering neuronal cells. Ultrastructure and electrophysiology show transfected neuroligin and neurexin appearing at the synaptic gap in a functional synapse. Further *in vitro*

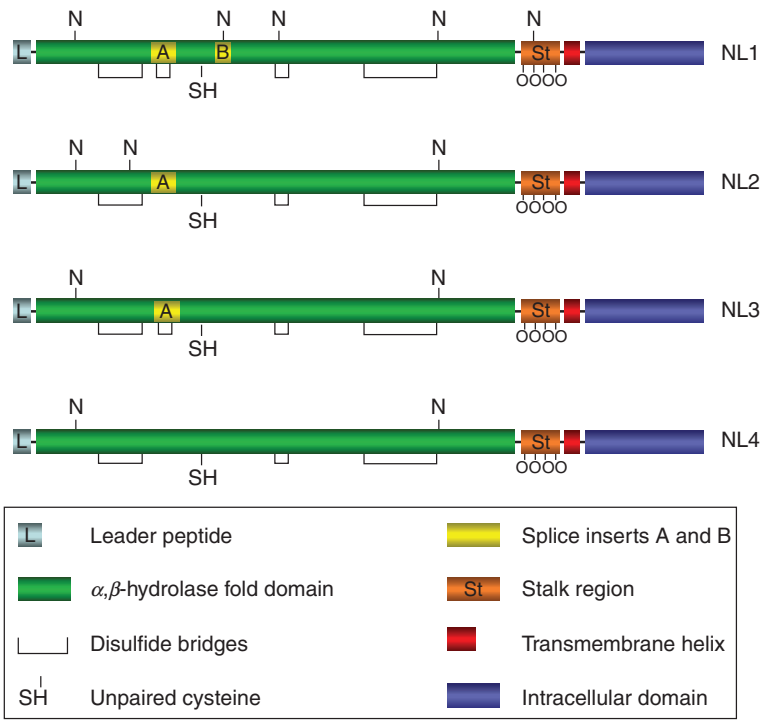

Figure 1 Schematic representation of the domain organization of the mammalian neuroligins 1–4. The leader peptide is absent in the mature protein. N and O, consensus sequences for N- and O-linked glycosylation sites, respectively. For neuroligin-1, disulfide bonds were assigned, and glycosylation sites were identified by mass spectrometry. A fifth neuroligin has been described and located on the Y chromosome rather than the X chromosome for NL4. NL4 and NL5 have close sequence identity.

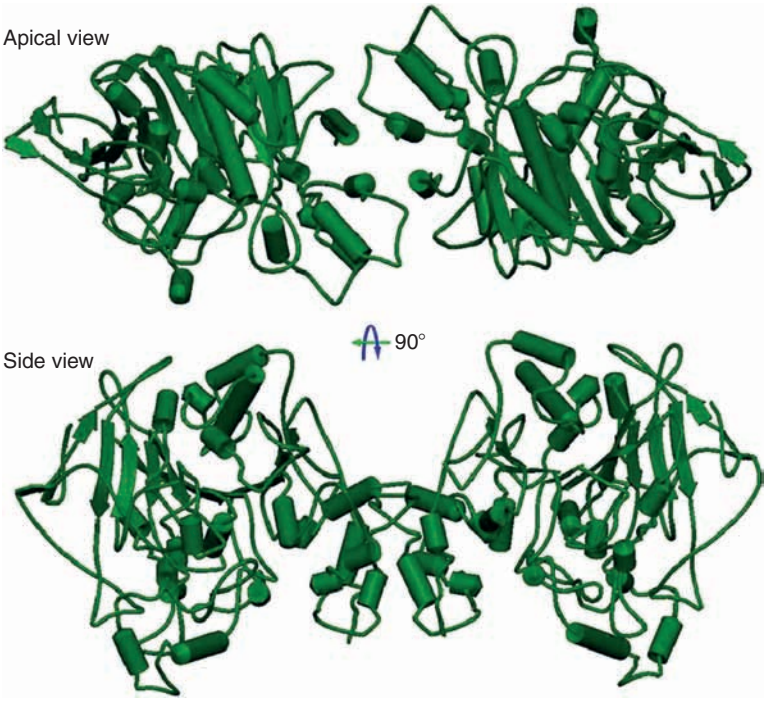

Figure 2 Ribbon representation of dimeric neuroligin-1 homology model. The structural model of the extracellular domain of the neuroligin-1 was generated using the crystal structure of mouse acetylcholinesterase as a template. Two monomeric neuroligin-1s are connected at the center of the image through a four-helix bundle typical of the α,β-hydrolase fold proteins. In the side view, the O-linked glycosylation region would extend toward the bottom of the page as visualized in **Figure 3**. The arrow indicates the direction and degree of rotation.

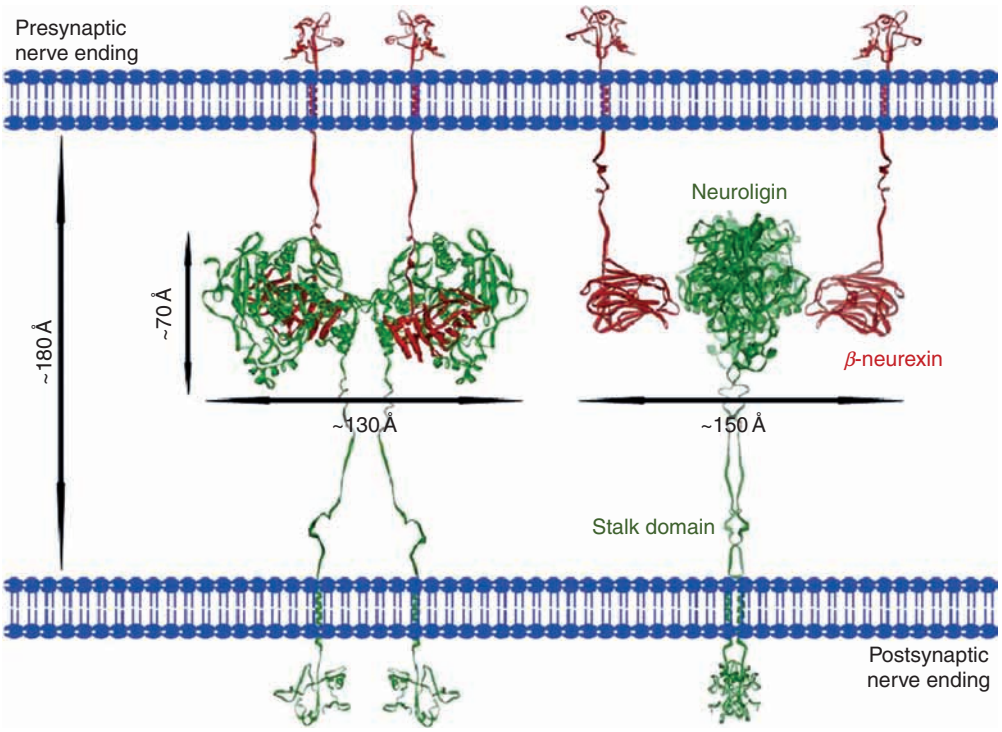

Figure 3 Assembly of the neuroligin–neurexin complex at the synapse. Shown here is a proposed model based on sedimentation analysis, ligand affinities, small-angle X-ray, and neutron diffraction. The two views differ by a 90° rotation.

studies, using overexpression and knockdown of neuroligins, suggest that these molecules control the balance of inhibitory/excitatory synapses in the CNS. *In vitro* work has shown that the binding between the scaffolding protein PSD-95 and neuroligin can modulate the release probability of synaptic vesicles in a retrograde manner, suggesting a role for β-neurexin as a presynaptic mediator. *In vivo* studies have demonstrated that individual neuroligin knockout mice survive and are fertile, whereas the *NL1*, *NL2*, and *NL3* triple knockout mice die soon after birth because of respiratory failure. Ultrastructure of the synapses of these animals appears normal, but marked functional defects in synaptic transmission have been noted. This study indicates that neuroligins are necessary for synapse stabilization and maturation but not for the initial formation of synaptic contacts.

Genetic investigations have revealed a linkage in twin sets and siblings between the incidence of the autism spectral disorders and mutations in the neuroligin gene. Although the frequency of neuroligin polymorphism in the overall autistic population is low, this key finding, substantiated by an increasing number of studies, points to adhesion proteins or their associating partners (neurexin-1 and Shank-3) as being linked to the autism spectral disorders. Although some mutations result in a truncated and inactive protein, other nonsynonymous coding region single nucleotide polymorphisms give rise to neuroligins that are deficient in association properties or processing after biosynthesis. Mutants of the neuroligin-3 gene with defective processing, when transfected into mammalian cells, give rise to abnormal expression, trafficking, and localization as evidenced by altered patterns of clustering of the neuroligin. Hence, a cellular phenotype can be defined based on cell culture studies of the modified neuroligin.

Neuroligin in the Peripheral Nervous System and Secretory Tissues

In addition to their expression in the CNS, neuroligins are found in several other peripheral tissues such as the uterus, in which neuroligin-2 expression appears to be under the influence of both estrogens and progesterone, and in heart, liver, skeletal muscle, and pancreas, indicating that different neuroligins may have discrete functions in neurosecretory cells in the central and peripheral nervous systems.

Conclusions

Neuroligins are a family of closely related gene products of highly conserved sequences between species. Their extracellular domain utilizes a α,β-hydrolase fold similar to acetylcholinesterase. In the CNS, they

are found on postsynaptic surfaces and interact in a selective manner in synaptic space with presynaptic α and β neurexins. Their C-terminal recognition domains associate with several PDZ-containing synaptic proteins. *In vivo*, the neuroligins seem to be necessary for maintaining synapse integrity. Mutations in various neuroligin and neurexin genes appear to be linked to the autistic spectral disorders and mental retardation.

See also: Autism; Cell Adhesion Molecules at Synapses; Neurexins.

Further Reading

Autism Genome Project Consortium (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genetics* 39: 319–328.

Boucard AA, Chubykin AA, Comoletti D, Taylor P, and Sudhof TC (2005) A splice code for trans-synaptic cell adhesion mediated by binding of neuroligin 1 to alpha- and beta-neurexins. *Neuron* 48: 229–236.

Chih B, Engelman H, and Scheiffele P (2005) Control of excitatory and inhibitory synapse formation by neuroligins. *Science* 307: 1324–1328.

Comoletti D, De Jaco A, Jennings LL, et al. (2004) The Arg451-Cys-neuroligin-3 mutation associated with autism reveals a defect in protein processing. *Journal of Neuroscience* 24: 4889–4893.

Cowan WM, Südhof TC, and Stevens CF (2001) *Synapses*. Baltimore: John Hopkins University Press.

De Jaco A, Comoletti D, Kovarik Z, et al. (2006) A mutation linked with autism reveals a common mechanism of endoplasmic reticulum retention for the alpha, beta-hydrolase fold protein family. *Journal of Biological Chemistry* 281: 9667–9676.

Graf ER, Zhang X, Jin SX, Linhoff MW, and Craig AM (2004) Neurexins induce differentiation of GABA and glutamate postsynaptic specializations via neuroligins. *Cell* 119: 1013–1026.

Ichtchenko K, Hata Y, Nguyen T, et al. (1995) Neuroligin 1: A splice site-specific ligand for beta-neurexins. *Cell* 81: 435–443.

Ichtchenko K, Nguyen T, and Sudhof TC (1996) Structures, alternative splicing, and neurexin binding of multiple neuroligins. *Journal of Biological Chemistry* 271: 2676–2682.

Irie M, Hata Y, Takeuchi M, et al. (1997) Binding of neuroligins to PSD-95. *Science* 277: 1511–1515.

Jamain S, Quach H, Betancur C, et al. (2003) Paris Autism Research International Sibpair Study. Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nature Genetics* 34: 27–29.

Laumonnier F, Bonnet-Brilhault F, Gomot M, et al. (2004) X-linked mental retardation and autism are associated with a mutation in the NLGN4 gene, a member of the neuroligin family. *American Journal of Human Genetics* 74: 552–557.

Meyer G, Varoqueaux F, Neeb A, Oschlies M, and Brose N (2004) The complexity of PDZ domain-mediated interactions at glutamatergic synapses: A case study on neuroligin. *Neuropharmacology* 47: 724–733.

Scheiffele P, Fan J, Choih J, Fetter R, and Serafini T (2000) Neuroligin expressed in nonneuronal cells triggers presynaptic development in contacting axons. *Cell* 101: 657–669.

Varoqueaux F, Aramuni G, Rawson RL, et al. (2006) Neuroligins determine synapse maturation and function. *Neuron* 51: 741–754.

Yan J, Oliveira G, Coutinho A, et al. (2005) Analysis of the neuroligin 3 and 4 genes in autism and other neuropsychiatric patients. *Molecular Psychiatry* 10: 329–332.

Neuromodulation

P S Katz, Georgia State University, Atlanta, GA, USA
R J Calin-Jageman, Dominican University, River Forest, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The role of the nervous system is to control the body. This involves responding to sensory information from outside of the body and acting on information about internal drives or states. There is a tendency to think of information flow within the nervous system as an integration of excitatory and inhibitory input. However, the information processing performed by the nervous system is much more subtle and multifaceted.

Unfortunately, descriptions of complex phenomena tend to be reduced to simple dichotomies. In the case of information flow in the nervous system, the dichotomy is often presented as neurotransmission versus neuromodulation, where neurotransmission is taken to mean classical synaptic transmission (excitatory postsynaptic potentials and inhibitory postsynaptic potentials) and neuromodulation is everything else. Neuromodulation can also be defined as the alteration of cellular or synaptic properties by a neuron or a substance released by neurons. Thus, neuromodulatory actions do not necessarily convey excitation or inhibition, but they can lead to a change in the excitability of a neuron or a change in the strength of a synapse. In this way, neuromodulatory actions can be contingent upon the activity of the neurons being modulated. Neuromodulation can cause neurons and synapses to undergo state changes, effectively switching into a different processing mode.

It is useful to distinguish neuromodulation from homosynaptic plasticity such as frequency facilitation, synaptic depression, posttetannic potentiation (PTP), and long-term potentiation (LTP). In homosynaptic plasticity, the properties of a neuron or a synapse are altered by its own pattern of activity. This is different, in principle, from neuromodulation, which is the alteration of cellular and synaptic properties through the activity of another neuron and can thus be called heterosynaptic plasticity. Whereas homosynaptic plasticity changes how a neuron integrates information over time based solely on its own history, neuromodulation or heterosynaptic plasticity is imposed upon a neuron and, although it can depend on a neuron's own pattern of activity, provides new information to a neuron.

Neuromodulation is produced through a variety of mechanisms and plays diverse roles in information processing, from sensory adaptation to motor pattern selection and learning. Understanding the basics of neuromodulation is essential for understanding how the nervous system produces behavior.

Mechanisms of Neuromodulation

One of the primary mechanisms of neuromodulation is activation of G-protein-coupled receptors (GPCRs). GPCRs are membrane-bound neurotransmitter receptors that do not form ion channel pores; instead, they are linked via G-proteins to important intracellular signaling pathways. The association between neuromodulation and GPCRs is so strong that the two concepts have become practically synonymous. However, neuromodulation also occurs through other mechanisms that also produce changes in cellular and synaptic properties. These include influx of Ca^{2+} through ligand-gated ion channels and activation of intracellular receptors by membrane-permeable messengers such as nitric oxide (NO).

Most neurotransmitters act at both ligand-gated ion channels and GPCRs. Moreover, most postsynaptic membranes express a combination of ligand-gated ion channels and GPCR receptors. From this, several important conclusions can be drawn. First, neuromodulatory transmitters are not a specific chemical subgroup; nearly all neurotransmitters can have neuromodulatory effects. Second, neuromodulation is not particularly distinct from classic inhibitory/excitatory actions; both processes are likely to occur in parallel at most synapses. Finally, neuromodulation is not a 'rare' or 'specialized' form of neural communication; it is very common in the nervous system. In fact, the GPCRs of the endocannabinoid system are expressed at densities near the ligand-gated $GABA_A$ receptors that mediate most classical inhibition. In this light, it is a mistake to think of neuromodulation as secondary to neurotransmission; neuromodulation is a fundamental means by which information is processed in the nervous system.

The activation of G-proteins by GPCRs produces effects either through direct interactions of G-protein subunits with membrane proteins or by regulation of an enzyme that generates a second messenger. Several enzyme–second messenger systems have been identified, including, most notably, adenylyl cyclase–cAMP and phospholipase C–IP3/DAG. The second messengers can have direct actions on ion channels

or they can activate an effector enzyme such as a kinase or a phosphatase. For example, cAMP activates cyclic nucleotide-gated (CNG) channels and the enzyme protein kinase A. Similarly, IP3 gates IP3 receptors, which allow Ca^{2+} to enter the cytoplasm from intracellular stores. The elevation of intracellular Ca^{2+} together with DAG activates protein kinase C. Phosphorylation by kinases or dephosphorylation by phosphatases changes the conformation of proteins such as ion channels and thereby changes properties such as gating kinetics.

Neuromodulatory actions can also be mediated by other second messenger – effector systems that abound in neurons. For example, endocannabinoids and gaseous messengers such as NO evoke neuromodulatory actions. These messengers are cell-permeable metabolic by-products, not vesicularly released substances. As such, they do not act at anatomically defined synaptic contacts; rather, they can diffuse from the point of release. This type of local broadcast of a signal has been called volume transmission, in contrast to point-to-point transmission at defined synaptic contacts.

Neuromodulation of Neuronal Properties

Regardless of whether neuromodulation occurs via GPCR activation or another mechanism, the end result is an alteration of neural function. There are three possible targets: neuronal properties, presynaptic release, or postsynaptic responsiveness (Figure 1). The site of action implies a particular type of neuromodulatory effect with its own implications for signal processing.

The primary mechanism by which neuromodulation alters neural properties is through the modification of ion channel function and expression. This is a powerful target of action because the passive and active properties of a neuron are determined largely by the relative densities of ion channels expressed and their functional properties. By altering ion channel function, neuromodulators can shape the complex and highly nonlinear passive and active properties of a neuron (Table 1).

To understand the role that neuromodulation of ion channels can play in shaping neuronal properties, consider a simple Hodgkin–Huxley-style neuron that has just two types of channels: a voltage-gated sodium conductance (G_{Na}) and a delayed rectifier potassium conductance (G_K) (Figure 2). If neuromodulatory

Table 1 Partial list of neuronal properties that are subject to neuromodulation

Spike frequency adaptation
Postinhibitory rebound
Bistability/plateau potentials
Bursting pacemaker potentials
Spike duration
Delayed firing
Accelerating firing
Spike propagation

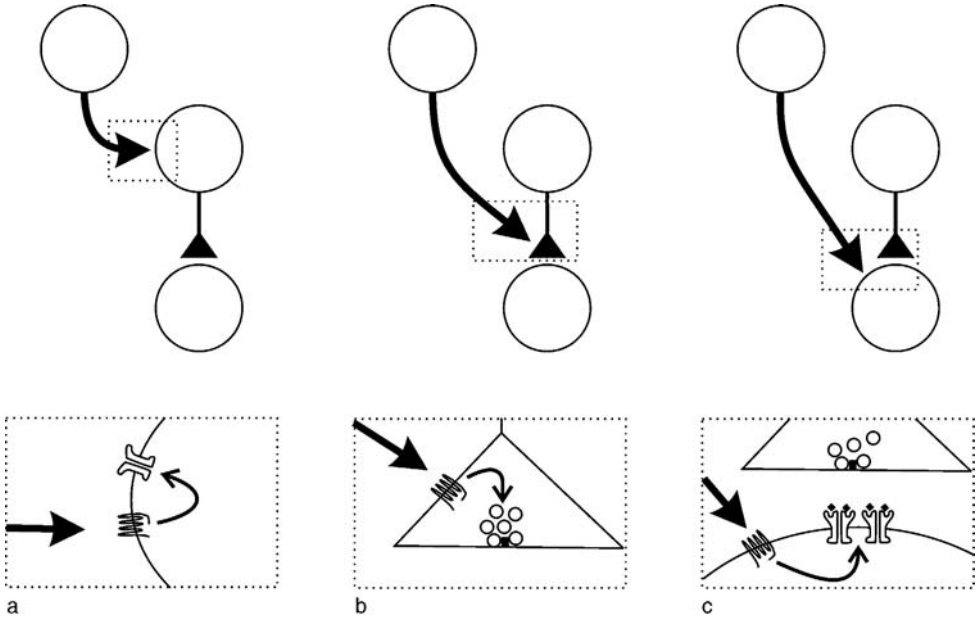

Figure 1 Neuromodulatory inputs (arrows) acting at G-protein-coupled receptors can affect neurons at three distinct sites. (a) Neuronal properties can be affected by changes in the properties of membrane ion channels. (b) Synaptic release can be affected by changes in the properties of synaptic release proteins. (c) Synaptic responses can be affected by changes in neurotransmitter receptors.

inputs alter the maximal conductance of G_K, the properties of the neuron will vary tremendously. For example, if the neuron is tonically active at rest (Figure 2 (b)), then decreasing G_K will increase the spontaneous firing rate of the neuron (Figure 2(a)), and increasing G_K can cause the neuron to become silent at rest (Figure 2(c)). Thus, neuromodulation of one type of ion channel in this simple example changes the neuron's basal firing properties.

Neuromodulation of ion channel properties can do more than alter a neuron's tonic firing properties. Changes can also occur in the way a neuron integrates synaptic input, the location of action potential initiation, the speed and shape of the action potential, and the success of action potential propagation. One type of modulatory change that has attracted particular interest is the regulation of firing mode. For example, neuromodulation of ion channel properties can shift a neuron from tonic activity into an oscillatory state, in which it fires rhythmic bursts of action potentials. Alternatively, neuromodulation can regulate membrane potential bistability. Bistable neurons exhibit two stable states: a quiescent state below threshold and an active state above threshold in which a long-lasting plateau potential is produced. Neuromodulation can shift neurons between these states and/or change the probability of shifting between these states. The regulation of firing mode indicates that neuromodulation can exert significant influence on a neural circuit even in the absence of ongoing synaptic input.

Although ion channel function serves as an important target for neuromodulation, neural properties can also be altered by changes in neural morphology. This type of change may be particularly important for expressing modulatory changes at long timescales.

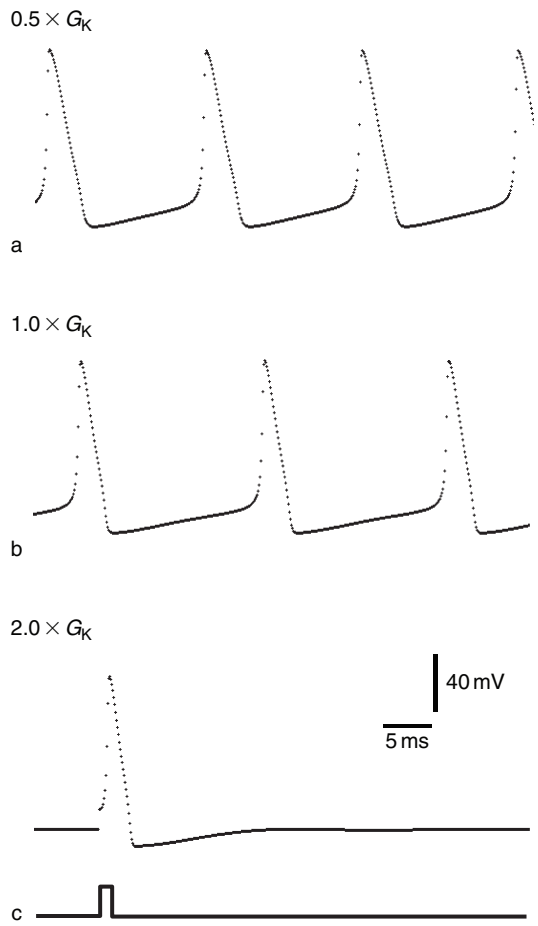

Figure 2 The firing properties of a simple Hodgkin–Huxley model neuron change by altering the maximal potassium conductance. The conductance was varied by multiplying the number of potassium channels by 0.5 (a), 1.0 (b), or 2.0 (c). With the potassium conductance halved (a), the neuron exhibits a higher rate of spontaneous firing. When the conductance is doubled (c), the neuron is not spontaneously active but responds to injected current (bottom trace). Stimulation produced using http://www.afodor.net/HHModel.htm.

Neuromodulation of Synaptic Properties

Neuromodulation not only affects the firing properties of neurons but can affect the synaptic properties of neurons as well (Table 2). This can occur through presynaptic modulation of release (Figure 1(b)) or postsynaptic sensitivity (Figure 1(c)).

On the presynaptic side, a number of modulatory targets are available. One common mode of action is again the modification of ion channel function. For example, increasing the maximal conductance of voltage-gated Ca^{2+} channels increases the Ca^{2+} influx for each action potential and thereby increases synaptic strength. If the targeted ion channels are localized to synaptic regions, this type of neuromodulation can occur independently of changes in neuronal firing properties. In addition to ion channels, presynaptic terminals expose a number of additional targets for neuromodulation. For example, changes in the extrusion of Ca^{2+} from intracellular stores affect synaptic release. Also, phosphorylation of proteins related to the synaptic release mechanism alters the amount of neurotransmitter released in response to Ca^{2+} influx.

Table 2 Partial list of synaptic properties that are subject to neuromodulation

Synaptic strength
Synaptic depression
Synaptic facilitation
Synaptic plasticity
Voltage sensitivity
Ion selectivity

Postsynaptic modification is also an important aspect of neuromodulation (**Figure 1(c)**). As on the presynaptic side, a diversity of targets are available for regulation. For example, phosphorylation of neurotransmitter receptors can alter their conductances or even their rates of desensitization, thereby affecting synaptic strength. Altering the relative contributions of different postsynaptic receptors can also have important implications for the information conveyed by a synapse. For example, at glutamatergic synapses, neuromodulatory changes that enhance the role of NMDA receptors over AMPA/kainate receptors will change the extent of Ca^{2+} influx into the synapse.

Finally, neuromodulation can also produce more global changes in synapse structure and function. This includes changes in synapse morphology, changes in spine morphology, and/or changes in the total number of synaptic contacts between neurons. These changes can serve to either increase or decrease the synaptic coupling between neurons.

Regardless of the site of synaptic change, it is important to consider that neuromodulation can result in more than a simple alteration in synaptic efficacy (stronger or weaker synapse): It can extend to a change in synaptic dynamics. For example, homosynaptic depression is a decrease of synaptic strength with the repeated activation of a synapse. Similarly, homosynaptic facilitation is a frequency-dependent increase in synaptic strength. Neuromodulatory actions can change the rates of depression and facilitation at synapses, allowing synaptic dynamics as well as strength to vary.

Neuromodulation can also change the susceptibility of a synapse to other forms of plasticity. This is known as metaplasticity. Neuromodulation can raise or lower the threshold for inducing several forms of synaptic plasticity, including PTP to LTP and long-term depression (LTD). Because LTP and LTD are believed to be particularly important for learning and memory, it seems that modulatory metaplasticity may be an important regulator of these processes. Specifically, neuromodulatory inputs may act as a gate for permitting synaptic changes that result in learning and memory.

By analogy with metaplasticity, neuromodulatory actions can be changed by neuromodulation, a situation that has been termed metamodulation. In some cases, one neuromodulatory action gates another. In other cases, it has been shown that neuromodulatory inputs converge on the same target, producing nonlinear summation of actions. The modulation of a modulation may play an important role in the control of neural circuits.

Time Course of Neuromodulation

Neuromodulation was traditionally thought of as having a gain-setting function, producing a static alteration in neural or synaptic function. It is now becoming clear, however, that neuromodulation is highly dynamic, operating on multiple timescales ranging from milliseconds to days.

Part of the early perception of neuromodulation as a static process probably stemmed from limitations in methodology. It is often difficult to selectively activate neuromodulatory neurons and record from appropriate targets, so neuromodulation is typically studied by direct application of a neuromodulatory substance. This static application tends to produce static effects, reflecting more on the mode of application than on the true nature of neuromodulation.

A clearer view of the dynamic nature of neuromodulation comes from studies in which it has been possible to directly evoke neuromodulatory action via selective activation of neuromodulatory neurons. In these cases, it has been shown that modulation can have a very rapid onset and decay, sometimes lasting only a few seconds. Moreover, a single modulator can evoke multiple processes operating on different timescales. This produces a highly dynamic pattern of neuronal and synaptic change.

The recognition that neuromodulation can occur on multiple timescales raises the question of how the time course of modulation is controlled. One important parameter seems to be the pattern of exposure to the modulator. Specifically, short pulses of the modulator can activate different intracellular pathways than a sustained exposure, leading to very different patterns of expression over time.

Another important aspect of the time course of modulation is the mechanism of action. Unlike the fast excitatory/inhibitory actions of neurotransmitters, neuromodulatory actions occur primarily via activation of GPCRs, which activate intracellular signaling cascades. Thus, neuromodulatory actions can greatly outlast the presence of the modulatory neurotransmitter. In fact, neuromodulation can have extremely long-lasting effects on neural function. In these cases, it seems that the primary target of modification is the transcriptional machinery of the neuron, enabling relatively permanent changes in neural and synaptic function. Thus, neuromodulators can act over a tremendous array of timescales, sometimes with multiple processes having different time constants.

The Complexity of Neuromodulation

From this overview, it should be clear that neuromodulation can alter nearly every aspect of neuronal and synaptic function. Moreover, these alterations can be highly dynamic, involving parallel signaling pathways operating over vastly different timescales. Unfortunately, the situation becomes even more

complicated once it is considered that neuromodulation never occurs within a vacuum but, rather, as part of the function of a complex and dynamic neural network.

One consequence is that neuromodulatory effects will have complex reverberations both at the direct sites of modulation and throughout the network. For example, in the mollusk *Aplysia californica*, serotonin produces neuromodulation that alters the function of voltage-gated potassium channels in some sensory neurons. The primary effect of this modulation is a narrowing of the action potential. Spike narrowing alters the active properties of the sensory neurons, making it possible to fire action potentials more rapidly. In addition, narrower spikes reduce synaptic release by limiting the duration of Ca^{2+} influx during each action potential. The decrease in initial release also slows the depression that occurs during a sustained activation, thus altering synaptic dynamics. Presumably, alterations in sensory neuron synaptic dynamics alter the ability to activate postsynaptic targets, but predicting exactly how faster, smaller, less depressing excitatory postsynaptic potential trains will influence postsynaptic activity is clearly quite challenging. This is especially true given that serotonin has several parallel effects on *Aplysia* sensory neurons.

Sources of Neuromodulation

In the mammalian brain, there are nuclei that are known to provide diffuse neuromodulatory input to other areas of the brain. Serotonergic axons arising from neurons in the raphe nuclei project to other brain areas. It is thought that practically every neuron in the brain receives some serotonergic input. Similarly, dopamine projects to other areas of the brain from the substantia nigra and the ventral tegmental area. Noradrenergic fibers originate in the locus coeruleus. Cholinergic axons originate in the basal forebrain region. Thus, these neuromodulatory substances arise from neurons whose cell bodies are clustered in discrete nuclei and project to diverse areas of the nervous system. Clearly, these neurons are extrinsic to the local circuits that they affect; thus, they are considered extrinsic neuromodulation.

Neuromodulation does not arise only from brain nuclei. Sensory input can also evoke neuromodulatory actions. This was noted with respect to muscle receptors in the crab stomatogastric nervous system. Neuromodulation is also present within sensory structures such as the retina. For example, some dopaminergic interplexiform neurons modulate the strength of electrical coupling between horizontal cells in the retina.

In contrast to extrinsic neuromodulation, neurons intrinsic to a local circuit can also exert neuromodulatory actions on other neurons within that circuit. For example, in the central pattern generator (CPG) for swimming in the mollusk *Tritonia*, some serotonergic neurons are rhythmically active and necessary for producing the swim motor pattern. They also alter the strength of synapses made by other members of the swim CPG. Thus, these neurons represent an example of intrinsic neuromodulation, in which the neuromodulatory elements are part of the local circuit. There are many examples in which neurons evoke neuromodulatory actions in addition to their classical neurotransmitting actions. It is likely that it will turn out to be a rule rather than an exception that neurons alter the properties of other neurons at the same time that they convey excitation or inhibition.

Functions of Neuromodulation

Neuromodulation has been implicated in a variety of functions at all levels of the nervous system. By changing cellular and synaptic properties, neuromodulation can adjust the gain of sensory input to the nervous system, the activity of interneurons, the output of motor neurons, and even the responses of muscles. Neuromodulation therefore provides the nervous system with the ability to extend the dynamic range of neurons and effectors as well as change the state of operation of systems.

The dynamic range of sensory neurons (i.e., the range of input values that elicit linear changes in the response) can be adjusted through neuromodulatory mechanisms. This can contribute to sensory adaptation or it can be used for dynamic filtering of sensory input. In central circuits, the receptive field properties of neurons can be adjusted by neuromodulation. For example, serotonin alters the sound frequency selectivity of auditory-responsive neurons in the inferior colliculus of echolocating bats.

For neuromuscular systems, neuromodulation can adjust the amplitude of contraction as well as the relaxation rates of muscles. This can alter how muscles respond to motor neuron input. For example, if the time constant of relaxation of a muscle is longer than the interval between motor neuron bursts, then contractions will summate. Increasing the muscle relaxation rate will allow the muscle to follow the motor neuron input with discrete contractions. Neuromodulation of amplitude and relaxation rate plays a role in the control of feeding movements in the mollusk *Aplysia*.

Neural circuits process information or generate activity based on the properties of their component

neurons and synapses. Since neuromodulation can change neuronal and synaptic properties, it can change how neural circuits process information or generate activity. In this way, neuromodulation can cause an ensemble of neurons to produce different patterns of activity. In motor systems, this results in the generation of different movements by the same motor circuitry.

A classic example of the ability of neuromodulation to alter the generation of motor activity comes from the stomatogastric ganglion of lobsters and crabs. This small ganglion contains just 30 neurons that comprise two semiautonomous CPG circuits, which control the rhythmic movements of the animal's foregut. The synapses among these neuronal circuits comprise an anatomically defined neural network. The motor patterns produced by this neural network depend on the neuromodulatory input that the ganglion receives; neuromodulatory inputs to the stomatogastric ganglion cause neurons to change their phase relations and their levels of activity and even switch their participation from one CPG to another. Thus, neuromodulatory inputs cause this anatomically defined network of neurons to be configured into different functional networks, which correspond to different modes of activity.

Homeostasis and Neuromodulation

The ability of neuromodulation to change cellular properties and thereby cause neural circuits to produce different outputs implies that the nervous system might be infinitely flexible. However, this does not appear to be the case. The anatomical pattern of synaptic connectivity constrains the neurons in a neural circuit to a limited range of possible firing patterns. Furthermore, not all combinations of ion channel properties result in adaptive cellular properties. For example, if potassium conductance overwhelms sodium conductance, then the neuron will not be able to fire an action potential, which would clearly be outside of the range of functionally relevant states.

Neurons have homeostatic corrective mechanisms that keep their cellular and synaptic properties within a functional range that is appropriate for each type of neuron. Thus, if the firing rate of a neuron decreases, the neuron will compensate by downregulating potassium channels or upregulating excitatory inputs.

The function of neuromodulation must be viewed within the context of homeostatic regulation. That is, neuromodulation does not determine cellular and synaptic properties; it merely perturbs them temporarily from their current state. In the presence of persistent neuromodulation, a neuron may homeostatically adapt to restore its basal properties. In this way, neuromodulation expands the dynamic range of neurons by allowing their properties to be up- and downregulated.

Neuromodulation and the Evolution of Behavior

Neural circuits, under the control of neuromodulation, can be multifunctional. Therefore, it follows that a mechanism underlying the evolution of species-specific behavior could involve alterations of neuromodulation. There is evidence that this has occurred. For example, in the stomatogastric system of decapod crustaceans, there are species differences in the neuromodulatory inputs even though the basic interconnectivity of the CPG circuits remains conserved across species.

Changes in the expression of GPCRs can alter the function of neuronal circuits. This can have far-reaching effects for the behavior of the animal. For example, it has been found that pair bonding by male prairie voles with their female partners is dependent on vasopressin receptor activation in the nucleus accumbens. However, the closely related montane voles lack these vasopressin receptors in that area of the brain and do not exhibit pair bonding. Thus, evolution of a complex social behavior may have involved a shift in the localization of neuropeptide receptors.

Summary

Neuromodulation plays an important, if underappreciated, role in information processing by the nervous system. By changing neuronal and synaptic properties, neuromodulatory actions reconfigure neuronal circuits and enable a wider variety of sensory responses and motor patterns to be produced by neural circuits. Without neuromodulation, the nervous system would act more like a hardwired device. Neuromodulatory actions bring the brain to life by endowing neural circuits with tremendous flexibility. Any mechanistic theory of how the brain produces behavior must take into account neuromodulatory signaling.

See also: Acetylcholine Neurotransmission in CNS; Adenosine Triphosphate (ATP); Calcium and Signal Transduction; Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Dopamine; Gamma-Aminobutyric Acid (GABA); Glutamate; Invertebrate Neurohormone GPCRs; Learning and Memory in Invertebrates: *Aplysia*; Learning and Memory in Invertebrate Models: Tritonia; Learning, Action, Inference and Neuromodulation; Neuromodulation of Sodium Channels; Neuromodulation of Calcium

Channels; Neutrotransmission and Neuromodulation: Acetylcholine; Nitric Oxide; Noradrenaline; Pattern Generation; Potassium Channel Regulation; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation.

Further Reading

Abraham WC and Bear MF (1996) Metaplasticity: The plasticity of synaptic plasticity. *Trends in Neurosciences* 19: 126–130.

Bao SW, Chan WT, and Merzenich MM (2001) Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature* 412: 79–83.

Brezina V and Weiss KR (1997) Analyzing the functional consequences of transmitter complexity. *Trends in Neurosciences* 20: 538–543.

Byrne JH and Kandel ER (1996) Presynaptic facilitation revisited: State and time dependence. *Journal of Neuroscience* 16: 425–435.

Davis GW (2006) Homeostatic control of neural activity: From phenomenology to molecular design. *Annual Review of Neuroscience* 29: 307–323.

Getting PA (1989) Emerging principles governing the operation of neural networks. *Annual Review of Neuroscience* 12: 185–204.

Harris-Warrick RM and Marder E (1991) Modulation of neural networks for behavior. *Annual Review of Neuroscience* 14: 39–57.

Harris-Warrick RM, Marder E, Selverston AI, and Moulins M (1992) *Dynamic Biological Networks: The Stomatogastric Nervous System*. Cambridge, MA: MIT Press.

Kaczmarek LK and Levitan IB (1987) *Neuromodulation: The Biochemical Control of Neuronal Excitability*. New York: Oxford University Press.

Katz PS (1999) *Beyond Neurotransmission: Neuromodulation and Its Importance for Information Processing*. Oxford: Oxford University Press.

Katz PS and Frost WN (1996) Intrinsic neuromodulation: Altering neuronal circuits from within. *Trends in Neurosciences* 19: 54–61.

Katz PS and Harris-Warrick RM (1990) Actions of identified neuromodulatory neurons in a simple motor system. *Trends in Neurosciences* 13: 367–373.

Levitan IB (1988) Modulation of ion channels in neurons and other cells. *Annual Review of Neuroscience* 11: 119–136.

Marder E and Goaillard JM (2006) Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience* 7: 563–574.

Powis DA and Bunn SJ (eds.) (1995) *Neurotransmitter Release and Its Modulation: Biochemical Mechanisms, Physiological Function and Clinical Relevance*. Cambridge: Cambridge University Press.

Turrigiano GG (1999) Homeostatic plasticity in neuronal networks: The more things change, the more they stay the same. *Trends in Neurosciences* 22: 221–227.

Relevant Websites

http://www.gpcr.org – Molecular Class-Specific Information System (MCSIS) Project.

http://stke.sciencemag.org – *Science* magazine, Signal Transduction Knowledge Environment.

http://www.afodor.net – The Hodgkin–Huxley Model of the Action Potential.

Neuromodulation of Calcium Channels

A Lee, Emory University School of Medicine, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Voltage-gated (Ca_v) Ca^{2+} channels play dual roles in controlling membrane excitability and activating Ca^{2+} signaling events in neurons. Ca_v channels contribute a depolarizing inward current, which influences the generation and duration of action potentials in some neurons. Ca_v channels also act as transduction elements, converting depolarizing stimuli into intracellular Ca^{2+} signals that activate gene transcription, protein phosphorylation, and neurotransmitter release. Therefore, factors that modulate Ca_v channel properties can powerfully influence cellular processes governing the proper development and function of the nervous system.

Multiple mechanisms have been identified that directly or indirectly adjust various parameters of neuronal Ca_v function. These include neurotransmitters acting through G-protein-coupled receptors, alternative splicing of the pore-forming Ca_v α_1 subunit, and direct interactions between Ca_v channels and various neuronal proteins. Neuromodulators can increase or inhibit the ability of Ca_v channels to open, close, or inactivate in response to membrane depolarization. They can also alter the localization or targeting of Ca_v channels to the plasma membrane and to discrete pre- or postsynaptic microdomains of neurons. The detailed molecular characterization of Ca_v channels has allowed the identification and analysis of factors that modulate Ca_v channels. Due to the relatively recent identification of the molecular sequences for Ca_v3 channels, considerably greater progress has been made to date in understanding the regulatory influences for Ca_v1 and Ca_v2 and the physiological consequences of these forms of modulation in the nervous system.

Neurotransmitter Modulation of Ca_v2 Channels

Mechanisms of G-Protein Modulation

$Ca_v2.2$ channels (which conduct N-type currents) and $Ca_v2.1$ channels (which conduct P/Q-type currents) are inhibited by neurotransmitters such as norepinephrine and γ-aminobutyric acid (GABA), which are agonists for receptors coupled to heterotrimeric G-proteins. Ca_v2 channels are largely expressed in presynaptic nerve terminals, where they mediate Ca^{2+} signals that trigger neurotransmitter release. G-protein-dependent inhibition of Ca_v2 is blocked by pertussis toxin, which inactivates $G\alpha_i$ and $G\alpha_o$, and can occur via two mechanisms.

Voltage-dependent G-protein inhibition Ca_v2 channel inhibition by G-proteins can occur by a voltage-dependent pathway, where inhibition is strong at negative, but not at very positive, membrane potential, such as during intense neuronal activity. This mechanism is fast (milliseconds) and membrane delimited in that it does not require the generation of second messengers: receptor activation causes G-protein $\beta\gamma$ subunits to dissociate from $G\alpha$ subunits and bind directly to the Ca_v2 α_1 subunit ($\alpha_1 2.2$). In electrophysiological recordings, voltage-dependent G-protein inhibition manifests as (1) slowing of Ca_v activation kinetics during sustained depolarizations, (2) depolarizing shift in activation voltage, and (3) relief of current inhibition (facilitation) by strong depolarizing prepulses. $G\beta\gamma$ binds to sites in the cytoplasmic N- and C-terminal domains of $\alpha_1 2.2$, and to two sites in the loop connecting domains I and II (Figure 1). These interactions coordinate conformational changes that cause channels to be 'reluctant' to open. Strong depolarization relieves the inhibitory conformations favored by $G\beta\gamma$, restoring channels to a 'willing' mode of opening. Inhibition by the membrane-delimited pathway is generally more prominent for $Ca_v2.2$ than for $Ca_v2.1$ and allows for rapid presynaptic inhibition by neuromodulators.

Voltage-independent G-protein inhibition A second route by which G-proteins inhibit Ca_v2 does not depend on membrane potential and can involve cytosolic second messengers. As a consequence, this voltage-independent inhibition is relatively slow (seconds) and can result from multiple mechanisms. In sympathetic neurons, $Ca_v2.2$ channels are inhibited by agonists of M1 muscarinic acetylcholine receptors, which are linked via $G\alpha_{q/11}$ to the activation of phospholipase $C\beta$. This enzyme hydrolyzes phosphatidylinositol 4,5-bisphosphate (PIP_2) to the second messengers inositol 1,4,5-trisphosphate and diacylglycerol. PIP_2 depletion can itself cause slow, voltage-independent suppression of $Ca_v2.2$ in neurons and also $Ca_v2.1$ expressed in *Xenopus* oocytes. As has been shown for other ion channels, PIP_2 can interact directly with the $Ca_v2.1$ α_1 subunit ($\alpha_1 2.1$) and support channel activity, which would be reversed upon PIP_2 depletion by the $G\alpha_{q/11}$ pathway. However, phosphoinositide hydrolysis can liberate arachidonic acid, which has also been

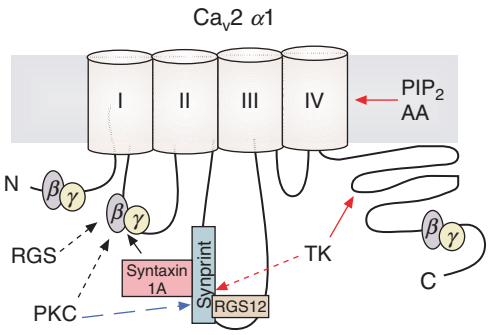

	Modulator	Effect on Ca$_v$2
Voltage-dependent regulation	Gβγ	Slow activation kinetics Depolarizing shift in activation voltage Facilitation by depolarized prepulses
	RGS	Terminate G-protein inhibition
	Syntaxin 1A	Facilitate G-protein inhibition
	PKC	Antagonize G-protein inhibition
Voltage-independent regulation	TK AA	Inhibit current amplitude
	PIP$_2$	Stablize channel activity

Figure 1 G-protein modulation of Ca$_v$2 channels. (Top) Schematic of Ca$_v$2 α subunit with Gβγ binding sites indicated. Factors supporting and opposing G-protein inhibition are shown as solid and dotted arrows, respectively. Regulators of G-protein signaling (RGS) proteins, protein kinase C (PKC), and syntaxin 1A affect voltage-dependent inhibition (black), while tyrosine kinases (TK) and phosphatidylinositol 4,5-bisphosphate/arachidonic acid (PIP$_2$/AA) are involved in voltage-independent pathways (red). PKC also affects soluble *N*-ethylmaleimide-sensitive attachment factor receptor (SNARE) binding to the synaptic protein interaction (synprint) site (blue dashed arrow). (Bottom) Major factors influencing voltage-dependent and voltage-independent G-protein regulation of Ca$_v$2.

implicated in the slow inhibition of Ca$_v$2.2. Therefore, voltage-independent inhibition of Ca$_v$2 channels via Gα$_{q/11}$ signaling is complex and may involve multiple lipid mediators.

A second mechanism for voltage-independent Ca$_v$2 inhibition occurs on a longer timescale (minutes) and is mediated by receptors coupled to Gα$_{i/o}$. Activation of these receptors by GABA promotes phosphorylation of α$_1$2.2 by tyrosine kinases and/or removal of the channel from the plasma membrane. The latter mechanism can involve direct receptor/Ca^{2+} channel interactions and internalization of the entire G-protein-coupled receptor/Ca$_v$2.2 complex. Ca$_v$2.2 channel internalization may allow for short-term desensitization and subsequent resensitization of activity-dependent presynaptic Ca^{2+} signals.

Modulators of G-Protein Inhibition

Ca$_v$2 inhibition by G-proteins can be controlled by a number of regulatory factors: proteins that are regulators of G-protein signaling (RGS), synaptic proteins, protein kinase C, and Ca$_v$2 α$_1$ splice variation.

RGS proteins RGS proteins cause ~100-fold stimulation of the GTPase activity of Gα, which promotes reassociation with Gβγ and restoration of the inactive G-protein heterotrimer. By limiting levels of active Gβγ, RGS proteins can terminate voltage-dependent inhibition of Ca$_v$2 channels. However, RGS proteins can also act through direct interactions with α$_1$2.2 to reverse voltage-independent modulation by G-proteins. For example, the GABA-B receptor-dependent phosphorylation of a tyrosine residue in the domain II–III linker initially promotes voltage-independent Ca$_v$2.2 inhibition that is eventually reversed by recruitment of RGS12 to this site (Figure 1). Greater than 20 RGS family members have been identified; they show specificity not only in how they turn off G-protein inhibition, but also in which G-protein signaling pathways are affected. Therefore, RGS proteins may

significantly influence the level and specificity of Ca_v2 channel inhibition by G-protein-coupled receptors.

Synaptic proteins $Ca_v2.1$ and $Ca_v2.2$ interact directly with soluble *N*-ethylmaleimide-sensitive factor (NSF)-attachment factor receptor (SNARE) proteins involved in the exocytotic release of neurotransmitter from presynaptic nerve terminals. SNAREs, which include syntaxin 1 and SNAP-25, interact with the synaptic vesicle Ca^{2+}-binding protein, synaptotagmin, during exocytosis. Biochemical studies indicate that SNAREs bind to a synaptic protein interaction ('synprint') site in the domain II–III linker of $Ca_v2\alpha_1$ (Figure 1). These interactions are generally thought to facilitate coupling between presynaptic Ca^{2+} entry and the neurotransmitter release machinery, but can also directly regulate the gating and G-protein modulation of Ca_v2 channels.

Syntaxin 1A is a membrane-associated protein that limits the functional availability of $Ca_v2.2$ by causing a decrease in channel expression levels and a hyperpolarizing shift in the steady-state inactivation curve. The effects on inactivation can be reversed by botulinum toxin C1, which cleaves syntaxin 1 close to its membrane anchor. In presynaptic nerve terminals, botulinum toxin C1 also prevents Ca_v2 inhibition by G-proteins. Biochemical studies indicate that syntaxin 1A binds to G-protein β subunits. When co-expressed with $Ca_v2.2$ in mammalian cells, syntaxin 1A facilitates tonic voltage-dependent G-protein inhibition of $Ca_v2.2$, perhaps by recruiting $G\beta$ into the channel complex. The syntaxin 1B isoform, which is ~80% identical to syntaxin 1A but is encoded by a separate gene, binds $\alpha_1 2.2$ and $G\beta$, but does not support G-protein modulation. This difference may arise from conformational distinctions in how syntaxin 1B versus syntaxin 1A interacts with $\alpha_1 2.2$ and/or $G\beta$. Variations in syntaxin 1A and syntaxin 1B expression, which occur between neurons and during development, may therefore determine the extent to which neuronal $Ca_v2.2$ channels are tonically regulated by G-proteins.

Protein kinase C Receptors coupled to $G\alpha_q$ activate protein kinase C (PKC), which can antagonize G-protein inhibition of $Ca_v2.2$. When activated pharmacologically by phorbol esters, PKC phosphorylates a threonine residue located in the $G\beta\gamma$ binding site of the $\alpha_1 2.2$ loop I–II linker (Figure 1), which opposes voltage-dependent inhibition by $G\beta\gamma$. PKC can also phosphorylate multiple sites in the synprint region of the $\alpha_1 2.2$ loop II–III linker. Phosphorylation of specific residues inhibits the ability of syntaxin 1A to cause negative shifts in $Ca_v2.2$ steady-state inactivation but not effects of syntaxin 1A on tonic G-protein regulation. Complex regulation of $Ca_v2.2$ channels by protein kinase C and $G\beta\gamma$ may permit dynamic control of presynaptic Ca^{2+} signals by 'cross-talk' of receptors coupled to functionally distinct G-proteins.

Ca_v2 α_1 splice variation Ca_v α_1-subunit genes are composed of multiple exons. Variable inclusion of particular exon sequences or exclusion of intron sequences during pre-mRNA processing can directly affect voltage-dependent Ca_v2 inhibition by $G\beta\gamma$. An $\alpha_1 2.1b$ variant containing a valine insertion in one of the two $G\beta\gamma$ binding sites in domain I–II linker shows stronger but slower inhibition by agonists of the μ-opioid G-protein-coupled receptor, compared to channels with $\alpha_{1A\text{-}a}$, which lacks the valine insertion. Expression of $\alpha_1 2.1b$ in some nerve terminals may therefore alter the efficiency with which Ca^{2+}-induced neurotransmitter release is presynaptically suppressed by $G_{i/o}$-linked receptors.

Cell-type-specific alternative splicing can also alter voltage-independent G-protein regulation of Ca_v2 channels. An $\alpha_1 2.2$ variant containing the alternatively spliced C-terminal exon 37a is mainly expressed in pain-transmitting nociceptive neurons that synapse in the dorsal horn of the spinal cord. Inclusion of exon 37a maintains voltage-dependent G-protein inhibition, but also strengthens voltage-independent inhibition by opiates and GABA, which depends on phosphorylation of a unique tyrosine residue in the alternatively spliced region (Figure 1). Since high-frequency stimulation relieves voltage-dependent inhibition of $Ca_v2.2$ by G-proteins, $\alpha_1 2.2$ variants containing exon 37a may allow for maintained suppression of presynaptic $Ca_v2.2$ channels and neurotransmission even during periods of heightened neuronal activity. This mechanism may be important for the analgesic effects of neurotransmitters and drugs on pain transmission in the spinal cord.

Neurotransmitter Modulation of Ca_v1 Channels

Ca_v1 (L-type) channels are also subject to modulation by G-protein-coupled receptor agonists (Figure 2). $Ca_v1.2$ channels are strongly potentiated by activation of β-adrenergic receptors coupled to G_s proteins. Initially characterized in the heart, this regulation involves a signal transduction pathway resulting in phosphorylation of the $Ca_v1.2$ α_1 subunit ($\alpha_1 1.2$) by cAMP-dependent protein kinase, which increases the duration and frequency of channel openings. Facilitation of $Ca_v1.2$ Ca^{2+} signals in response to β-adrenergic receptor agonists also occurs in individual dendritic spines of hippocampal neurons, suggesting that β receptors, $Ca_v1.2$, and signaling components must be highly

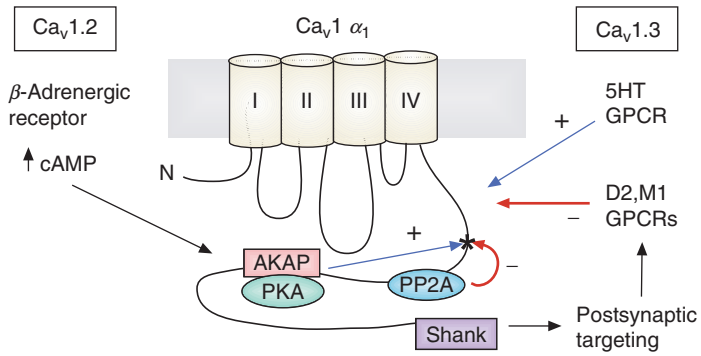

	GPCR	Effect on Ca_v1 currents
$Ca_v1.2$	β-Adrenergic	Facilitation
$Ca_v1.3$	D1 dopamine M1 Ach	Inhibition
	5HT serotonin	Facilitation

Figure 2 G-protein modulation of Ca_v1 channels. Major G-protein-coupled receptor (GPCR) regulatory pathways for $Ca_v1.2$ and $Ca_v1.3$. Increases in cAMP caused by β-adrenergic receptor activation promote $\alpha_1 1.2$ phosphorylation by cAMP-dependent protein kinase A (PKA), which is recruited to the channel by A-kinase anchoring proteins (AKAPs). The resulting facilitation of $Ca_v1.2$ is reversed by $\alpha_1 1.2$ dephosphorylation via associated protein phosphatase 2A (PP2A). $Ca_v1.3$ channels are potentiated by serotonin (5-hydroxytryptamine, 5HT) receptors and inhibited by dopamine (D2) and muscarinic acetylcholine (M1) receptors. The latter effect requires targeting of $Ca_v1.3$ to postsynaptic sites via the scaffolding protein, Shank, which binds to the $\alpha_1 1.3$ C-terminus.

compartmentalized. Biochemical studies confirmed the existence of an entire signaling complex (β-adrenergic receptor, $Ca_v1.2$, G_s-protein, adenylyl cyclase, cAMP-dependent protein kinase) that allows for rapid and localized signaling from receptor to Ca^{2+} channel. Efficient modulation of postsynaptic $Ca_v1.2$ channels may be required for the central role of these channels in Ca^{2+}-dependent synaptic plasticity.

Although expressed at lower levels than $Ca_v1.2$ in the nervous system, $Ca_v1.3$ is a second class of L-type channel that is regulated by G-protein-coupled receptors. $Ca_v1.3$ channels activate in a more negative voltage range than do $Ca_v1.2$ channels and therefore can facilitate attainment of action potential thresholds and regulate neuronal firing patterns. In spinal motoneurons and interneurons, $Ca_v1.3$ channels contribute to plateau potentials important for bistability and plasticity of postsynaptic depolarizations. $Ca_v1.3$ currents in these neurons are facilitated in response to activation of $G\alpha_q$-linked serotonin or metabotropic glutamate receptors that promote intracellular Ca^{2+} release. $Ca_v1.3$ facilitation is Ca^{2+} dependent in that it is blocked with high intracellular concentrations of Ca^{2+} chelators such as ethylene glycol tetraacetic acid (EGTA), and serves to modify the excitability and selective recruitment of motoneurons for the coordinated control of muscles during movement.

$Ca_v1.3$ channels are also inhibited by neurotransmitters such as dopamine and acetylcholine. In the striatum, $Ca_v1.3$ Ca^{2+} currents promote glutamatergic receptor-driven 'upstates' in medium spiny neurons, which are characterized by increased excitability. Activation of dopamine receptors and muscarinic acetylcholine receptors suppresses these upstate transitions by inhibiting $Ca_v1.3$. This inhibition requires a sequence in the distal C-terminal domain of the $Ca_v1.3$ α_1 subunit ($\alpha_1 1.3$), which is a consensus site for binding postsynaptic density-95 (PSD-95)/discs/large/zona occludens-1 (PDZ) domains, which are protein interaction modules present in a large variety of proteins. $Ca_v1.3$ channels interact with postsynaptic PDZ proteins, which act as molecular scaffolds and target these channels to particular synapses in medium spiny neurons, where they can be modulated by dopamine or

acetylcholine receptors. Thus, the formation of receptor/Ca^{2+} channel complexes is an important means for maintaining the specificity and efficiency of $Ca_v1.3$ channel regulation by neuromodulators (Figure 2).

Neurotransmitter Modulation of Ca_v3 Channels

Ca_v3 (T-type) channels are subject to both inhibitory and stimulatory regulation by neurotransmitters, peptides, and hormones acting through G-protein-coupled receptors. Ca_v3 channels are activated more negative voltages than are Ca_v1 and Ca_v2 channels, and are important for action potential bursting, transmission of noxious stimuli, and hormone secretion. There is considerable heterogeneity in how different Ca_v3 channels ($Ca_v3.1$–$Ca_v3.3$) respond to various stimuli and their underlying molecular mechanisms. $Ca_v3.2$ channels are inhibited by G_i-linked receptors in a membrane-delimited pathway mediated by direct binding of $G\beta\gamma$ to sites in the domain II–III linker. This effect differs from voltage-dependent inhibition of Ca_v2 channels in that it does not involve positive shifts in activation voltage or reversal by strong depolarizations. The II–III linker of $Ca_v3.1$ α_1 ($\alpha_1 3.1$) cannot bind $G\beta\gamma$ and therefore does not support similar G-protein inhibition of $Ca_v3.1$. The II–III linker of $Ca_v3.2$ α_1 ($\alpha_1 3.2$) is also the locus for the direct tethering and stimulatory effects of calmodulin-dependent protein kinase II, and so may coordinate modulatory cross-talk between receptors coupled to G_q and G_i.

Ca^{2+}-Dependent Modulation of Ca_v Channels

Although Ca_v channels readily conduct both Ca^{2+} and Ba^{2+} ions, L-type and P/Q-type Ca^{2+} currents decay (inactivate) strikingly faster than do Ba^{2+} currents. This process, Ca^{2+}-dependent inactivation, efficiently prevents Ca^{2+} overloads and refines activity-dependent Ca^{2+} signals in excitable cells. In cardiac myocytes, Ca^{2+}-dependent inactivation of $Ca_v1.2$ helps terminate action potentials and is surprisingly insensitive to high intracellular concentrations of Ca^{2+} chelators such as EGTA or 1,2-bis (*o*-aminophenoxy)ethane-*N*,*N*,*N*′,*N*′-tetraacetic acid (BAPTA). This latter result indicates that Ca^{2+}-dependent inactivation for $Ca_v1.2$ depends on very rapid, local increases in Ca^{2+} either in or near the channel pore.

Ca^{2+}-dependent inactivation was long considered specific for L-type currents, in part because most whole-cell patch clamp recordings employed high concentrations of EGTA or BAPTA (5–10 mM) in the intracellular recording solution, levels which are permissive for Ca^{2+}-dependent inactivation of L-type currents. However, with EGTA reduced to 0.5 mM, Ca^{2+}-dependent inactivation for $Ca_v2.1$ channels is robust, similar to presynaptic P/Q-type currents at the calyx of Held, a giant synapse in the brain stem. During trains of depolarizations, P/Q-type Ca^{2+} currents inactivate faster than Ba^{2+} currents do and undergo an initial increase (Ca^{2+}-dependent facilitation) prior to onset of Ca^{2+}-dependent inactivation. The latter, but not the former, is prevented by high intracellular concentrations of EGTA (10 mM), indicating a difference in the Ca^{2+} sensitivities of the two processes. Given that P/Q-type Ca^{2+} currents mediate neurotransmitter release at many central synapses, this dual feedback regulation may permit activity-dependent sharpening of presynaptic Ca^{2+} signals underlying synaptic plasticity. With low concentrations of EGTA (0.5 mM), Ca^{2+}-dependent inactivation also characterizes $Ca_v2.2$ and $Ca_v2.3$. Therefore, Ca_v2 channels are generally distinguished from Ca_v1 in showing Ca^{2+} feedback regulation that is sensitive to the Ca^{2+} buffering conditions of the cell.

Calmodulin as a Ca_v Channel Subunit

Ca^{2+} feedback for Ca_v1 and Ca_v2 involves the ubiquitous Ca^{2+}-binding protein calmodulin. For $Ca_v1.2$, calmodulin binds to a consensus site (IQ domain) in the C-terminal domain of $\alpha_1 1.2$ (Figure 3). Ca^{2+} ions permeating the channel bind EF-hand motifs (composed of two helixes, E and F) in the N- and C-terminal lobes of calmodulin, causing conformational changes that favor inactivation. Ba^{2+} ions bind poorly to calmodulin, and so cannot support rapid inactivation for Ca_v Ba^{2+} currents. Calmodulin interactions with the $\alpha_1 1.2$ IQ domain are functionally complex: mutation of the initial isoleucine to glutamate abolishes Ca^{2+}-dependent inactivation, but when substituted with alanine reveals Ca^{2+}-dependent facilitation, which is normally masked by intense Ca^{2+}-dependent inactivation for $Ca_v1.2$ in recombinant systems. Structural and functional studies demonstrate that positive and negative Ca^{2+} feedback depend on the N- and C-terminal lobes of calmodulin, respectively, which make contacts with different residues in the IQ domain. Among Ca_v1 channels, the IQ domain and other sequences (pre-IQ and EF-hand motifs; Figure 3) required for Ca^{2+}-dependent inactivation are nearly identical, illustrating the essential nature of these determinants for Ca_v function.

In the C-terminal domain of the $\alpha_1 2.1$, calmodulin interacts with a less well-conserved IQ domain and a second site (calmodulin binding domain; CBD), both of which are required for Ca^{2+} feedback (Figure 4). Calmodulin mutants unable to bind Ca^{2+} at the

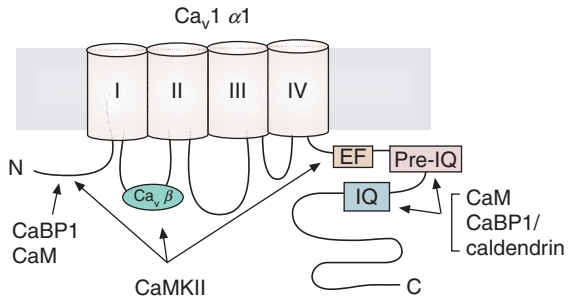

Modulator	Ca^{2+}-dependent inactivation	Ca^{2+}-dependent facilitation	Voltage-dependent activation
CaM	↑	↑	–
CaBP1	↓↓	↑	–
Caldendrin	↓	–	–
CaMKII	–	↑	↑ ($Ca_v1.3$)

Figure 3 Ca^{2+}-dependent modulation of Ca_v1. (Top) Schematic of Ca_v1 α subunit, showing binding sites for calmodulin (CaM), a Ca^{2+}-binding protein subfamily (CaBP1), and caldendrin, and calmodulin-dependent kinase II (CaMKII) phosphorylation sites (EF, calcium-binding motif composed of two helixes, E and F; the IQ domain is a consensus site). (Bottom) Effects of Ca^{2+} signaling proteins on Ca_v1 function. Positive (↑) and negative (↓) effects on Ca^{2+}-dependent inactivation, facilitation, and voltage-dependent activation are indicated.

N-terminal EF-hands cannot transduce Ca^{2+}-dependent inactivation, while C-terminal EF-hands are more critical for Ca^{2+}-dependent facilitation. Since the C-terminal lobe of calmodulin binds Ca^{2+} with higher affinity than does the N-terminal lobe, the calmodulin lobe dependence of Ca^{2+}-dependent facilitation and inactivation may underlie the distinct lobe sensitivities to Ca^{2+} buffering, and the increased Ca^{2+}-buffer sensitivity of Ca^{2+}-dependent inactivation for $Ca_v2.1$, compared to $Ca_v1.2$.

Calmodulin-Like Ca^{2+}-Binding Proteins as Ca_v Modulators

Calmodulin is but one member of a large family of EF-hand Ca^{2+}-binding proteins (CaBPs), some of which are expressed mainly in neurons. Like calmodulin, these CaBPs possess four EF-hand Ca^{2+} binding sites, at least one of which is nonfunctional. CaBPs are similar enough to displace calmodulin from binding sites in the Ca_v α_1 subunit, but different enough to confer distinct forms of regulation.

CaBP regulation of $Ca_v2.1$ channels CaBP1 represents a subfamily of CaBPs highly expressed in the brain and retina and is co-localized to some extent with presynaptic $Ca_v2.1$ channels. Like calmodulin, CaBP1 binds to the CBD of $\alpha_12.1$, but does not support Ca^{2+}-dependent facilitation and causes intense inactivation independent of Ca^{2+} (Figure 4). $Ca_v2.1$ also can interact with other CaBPs, such as neuronal Ca^{2+} sensor-1 (NCS-1), which has inhibitory effects on P/Q-type current amplitude in *Xenopus* oocytes and adrenal chromaffin cells. However, injection of purified NCS-1 protein into presynaptic nerve terminals at the calyx of Held synapse promotes facilitation of P/Q-type currents, and activity-dependent facilitation P/Q-type currents at this synapse can be prevented by injection of NCS-1-inhibitor peptides. Therefore, the effects of NCS-1 on $Ca_v2.1$ may depend on the cellular and subcellular context in which they are co-expressed. A third CaBP that modulates $Ca_v2.1$ is visinin-like protein-2 (VILIP-2), which is highly expressed in the neocortex and hippocampus and undergoes Ca^{2+}-dependent association with the plasma membrane in neurons and other cell types. When cotransfected with $Ca_v2.1$ in mammalian cells, VILIP-2 does not affect Ca^{2+}-dependent facilitation, but inhibits Ca^{2+}-dependent inactivation. These effects of VILIP-2 may involve displacement of calmodulin from the CBD, although both the CBD and IQ-like domain of $\alpha_12.1$ are required for the association of VILIP-2 with the channel. How

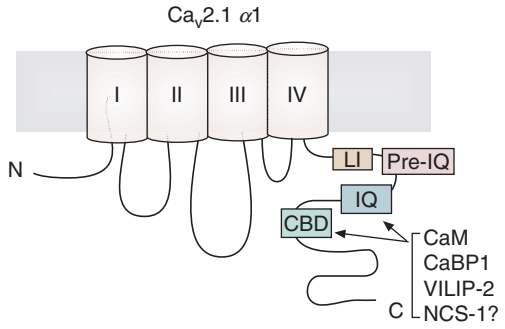

Modulator	Ca^{2+}-dependent inactivation	Ca^{2+}-dependent facilitation
CaM	↑	↑
CaBP1	↑	↓
VILIP-2	↓	–
NCS-1	–	↑

Figure 4 Ca^{2+}-dependent modulation of $Ca_v2.1$. $\alpha_1 2.1$ binding sites for calmodulin (CaM), a Ca^{2+} binding protein subfamily (CaBP1), and visinin-like protein-2 (VILIP-2), and potentially neuronal Ca^{2+} sensor-1 (NCS-1), and positive (↑) and negative (↓) effects on Ca^{2+}-dependent inactivation and facilitation are indicated.

VILIP-2 and CaBP1 can have such opposing effects on $Ca_v2.1$ function is not entirely clear, but may involve structural distinctions between the two CaBPs and/or how they interact with $Ca_v2.1$. The remarkably divergent actions of CaBPs on $Ca_v2.1$ may increase the range of presynaptic P/Q-type Ca^{2+} signals, thus contributing to heterogeneous forms of short-term plasticity at different synapses.

CaBP regulation of Ca_v1 channels Ca_v1 channels are also differentially regulated by CaBPs. Unlike its inhibitory effects on $Ca_v2.1$, CaBP1 prolongs $Ca_v1.2$ Ca^{2+} currents and completely abolishes calmodulin-dependent Ca^{2+}-dependent inactivation. The opposing actions of calmodulin and CaBP1 on $Ca_v1.2$ depend on differences in how the two interact with the channel. CaBP1 binds to the IQ-domain as well as to a site in the N-terminal domain of the $\alpha_1 1.2$ (Figure 3). Deletion of the N-terminal site prevents the effects of CaBP1 but not calmodulin on $Ca_v1.2$ inactivation. In contrast, IQ-EE substitutions in the IQ-domain prevent calmodulin-dependent inactivation, but partially spare the effects of CaBP1. Interestingly, a CaBP1 variant, caldendrin, causes more modest suppression of $Ca_v1.2$ Ca^{2+}-dependent inactivation than CaBP1 through interactions solely with the IQ-domain, and not with the N-terminal site in $\alpha_1 1.2$. CaBP1 and caldendrin associate and co-localize with $Ca_v1.2$ in somatodendritic domains of neurons, and so may help fine-tune and boost postsynaptic $Ca_v2.1$ Ca^{2+} signals that regulate gene transcription and neuronal excitability.

Calmodulin-Dependent Protein Kinase II Regulation of Ca_v1 Channels

Ca^{2+} binding to calmodulin causes autophosphorylation of the calmodulin-stimulated protein kinase, CaM kinase II, which can also promote Ca^{2+}-dependent facilitation of Ca_v1 (Figure 3). CaM kinase II binds to and phosphorylates both $\alpha_1 1.2$ and the auxiliary Ca_v β subunit, both of which cause Ca^{2+}-dependent facilitation during repetitive stimuli. Although characterized mainly for cardiac $Ca_v1.2$ channels, that neuronal $Ca_v1.2$ channels undergo a similar form of regulation is supported by the prevalence of CaM kinase II and $Ca_v1.2$ channels at postsynaptic sites in the brain. Such interactions may be important for boosting activity-dependent postsynaptic Ca^{2+} signals that trigger long-term changes in synaptic efficacy.

For $Ca_v1.3$, CaM kinase II causes a distinct form of facilitation that relies on intracellular Ca^{2+} release rather than Ca^{2+} influx. Insulin-like growth factor-1 (IGF-1) promotes neuronal survival and dendritic arborization in part by potentiating Ca_v1 channel activity. IGF-1 binds to a receptor tyrosine kinase, which causes Ca^{2+} release from intracellular stores by activation of the phospholipase Cγ signaling pathway. Activated CaM kinase II phosphorylates a serine residue in the EF-hand sequence in the C-terminal domain of $\alpha_1 1.3$, which causes faster activation kinetics and a negative shift in activation voltage for $Ca_v1.3$ expressed in a neuronal cell line. Since the serine that is phosphorylated by CaM kinase II in $\alpha_1 1.3$ is absent in $\alpha_1 1.2$, these effects of CaM kinase II may differentiate $Ca_v1.2$ and $Ca_v1.3$ modulation by IGF-1 and the respective roles of these L-type channels in Ca^{2+} signaling events required for neuronal development.

Conclusions

Due to their importance in controlling Ca^{2+} entry in neurons, Ca_v channels are subject to numerous and diverse forms of modulation. The intrinsic biophysical properties of Ca_v channels characterized in recombinant systems may therefore not be representative of how these channels behave in neurons. Characterizing the dynamic regulation of Ca_v channels by neurotransmitters, protein interactions, and alternative

splicing will further illuminate how Ca_v channels help orchestrate the normal development, function, and plasticity of the nervous system.

See also: Action Potential Initiation and Conduction in Axons; Calcium Channels; Calcium Channel Subtypes Involved in Neurotransmitter Release; Calcium Channels and SNARE Proteins; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Dendritic Signal Integration; Invertebrate Neurohormone GPCRs; Ion Channel Localization in Axons; Ion Channel Localization in Cell Bodies and Dendrites; Voltage-Gated Calcium Channels.

Further Reading

Budde T, Meuth S, and Pape HC (2002) Calcium-dependent inactivation of neuronal calcium channels. *Nature Reviews Neuroscience* 3: 873–883.

Catterall WA (2000) Structure and regulation of voltage-gated Ca^{2+} channels. *Annual Review of Cell and Developmental Biology* 16: 521–555.

Chemin J, Traboulsie A, and Lory P (2006) Molecular pathways underlying the modulation of T-type calcium channels by neurotransmitters and hormones. *Cell Calcium* 40: 121–134.

Delmas P, Coste B, Gamper N, et al. (2005) Phosphoinositide lipid second messengers: New paradigms for calcium channel modulation. *Neuron* 47: 179–182.

Dolphin AC (2003) G protein modulation of voltage-gated calcium channels. *Pharmacological Reviews* 55: 607–627.

Evans RM and Zamponi GW (2006) Presynaptic Ca^{2+} channels – integration centers for neuronal signaling pathways. *Trends in Neurosciences* 29: 617–624.

Lipscombe D, Pan JQ, and Gray AC (2002) Functional diversity in neuronal voltage-gated calcium channels by alternative splicing of Ca_v $\alpha 1$. *Molecular Neurobiology* 26: 21–44.

Moosmang S, Kleppisch T, Wegener J, et al. (2007) Analysis of calcium channels by conditional mutagenesis. *Handbook of Experimental Pharmacology* 178: 469–490.

Strock J and Diverse-Pierluissi MA (2004) Ca^{2+} channels as integrators of G protein-mediated signaling in neurons. *Molecular Pharmacology* 66: 1071–1076.

Tsien RW, Lipscombe D, Madison DV, et al. (1988) Multiple types of neuronal calcium channels and their selective modulation. *Trends in Neurosciences* 11: 431–438.

Welsby PJ, Wang H, Wolfe JT, et al. (2003) A mechanism for the direct regulation of T-type calcium channels by Ca^{2+}/calmodulin-dependent kinase II. *Journal of Neuroscience* 23: 10116–10121.

Neuromodulation of Sodium Channels

W A Catterall, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuromodulation of electrical excitability and synaptic transmission is the fundamental basis for many aspects of learning, memory, and physiological regulation. Synaptic plasticity allows regulation of the input–output relationships of individual synapses in response to their level of activity and other influences. Long-term potentiation of synaptic transmission at glutamatergic synapses on central neurons is the most well-studied example of this type of regulation. In addition to this synapse-specific form of plasticity, there is increasing evidence of the importance of cellular plasticity, a form of plasticity that modifies the input–output relationships of the entire neuron. One can imagine many different mechanisms by which the input–output relationships of a neuron could be modified, but neuromodulation of voltage-gated Na^+ channels is one of the simplest. Voltage-gated Na^+ channels are responsible for the initiation and propagation of action potentials in neurons. Because generation of action potentials is the final common pathway for neural output, Na^+ channels are ideally positioned to serve as an important point of regulation of neuronal signal integration and transduction.

Na^+ Channel Protein Phosphorylation and Dephosphorylation

Na^+ channel proteins in mammalian brain consist of an α subunit of approximately 260 kDa in association with one or two auxiliary subunits of 33–36 kDa (β_1–β_4; Figure 1). Early biochemical studies of purified brain Na^+ channels showed that the α subunits are exceptionally good substrates for phosphorylation by cAMP-dependent protein kinase A (PKA). In contrast, no phosphorylation of the β_1 or β_2 subunits was detected. Na^+ channels were rapidly phosphorylated on multiple sites following activation of PKA or protein kinase C (PKC) in synaptosomes, and measurements of neurotoxin-activated Na^+ influx showed that cAMP-dependent phosphorylation is correlated with a reduction of approximately 26% in the maximum neurotoxin-activated Na^+ conductance. Therefore, by this *in vitro* biochemical measurement of Na^+ channel function, cAMP-dependent phosphorylation of the α subunit is correlated with a reduction in activity of brain Na^+ channels.

Na^+ channels in intact brain neurons in cell culture are rapidly phosphorylated by PKA on at least four sites, which are located in the large intracellular loop connecting domains I and II of the α subunits (Figure 1). This complex pattern of phosphorylation of sites in $L_{I–II}$ is responsible for the reduction of Na^+ channel activity by activation of the cAMP pathway.

Purified brain Na^+ channels are dephosphorylated by multiple phosphoprotein phosphatases. The calcium-regulated phosphatase calcineurin is more effective than the catalytic subunits of phosphatase 1 or phosphatase 2A in dephosphorylating purified Na^+ channels, and it is as efficient in dephosphorylating Na^+ channels as in dephosphorylating its known substrates. Purification of Na^+ channel phosphatases from brain also demonstrated the importance of calcineurin and phosphatase 2A in Na^+ channel dephosphorylation. Thus, biochemical experiments provide clear evidence for both second messenger-regulated phosphorylation and dephosphorylation of Na^+ channels in brain neurons.

Functional Modulation of Cloned and Expressed Na^+ Channels

Functional modulation of voltage-gated Na^+ channels by PKC- and PKA-dependent phosphorylation has been observed for cloned $Na_v1.2$ channels expressed in mammalian cells and in *Xenopus* oocytes. Activation of PKC with phorbol esters or diacylglycerols such as 1-oleoyl-2-acetyl-*sn*-glycerol (OAG) reduces peak Na^+ currents and slows the time course of inactivation (Figure 2). The observed effects of diacylglycerol activators of PKC are directly attributable to PKC activity because they were blocked by specific inhibitors and mimicked by application of purified rat brain PKC to the cytoplasmic surface of the channel in excised membrane patches.

Activation of PKA also modulates the Na^+ current mediated by cloned $Na_v1.2$ channels expressed in *Xenopus* oocytes and in cultured mammalian cells. In both cell types, peak Na^+ current is reduced 25–40% without change in the kinetics or voltage dependence of activation or fast inactivation (Figure 2). PKA-dependent inhibition has also been reported for $Na_v1.1$ channels expressed in *Xenopus* oocytes. This is not surprising since $L_{I–II}$ of $Na_v1.1$ contains five consensus PKA phosphorylation sites in similar positions to those described for $Na_v1.2$.

Sites of Neuromodulation by PKA and PKC

The molecular mechanism underlying the regulation of $Na_v1.2$ channels by PKC and PKA is complex and

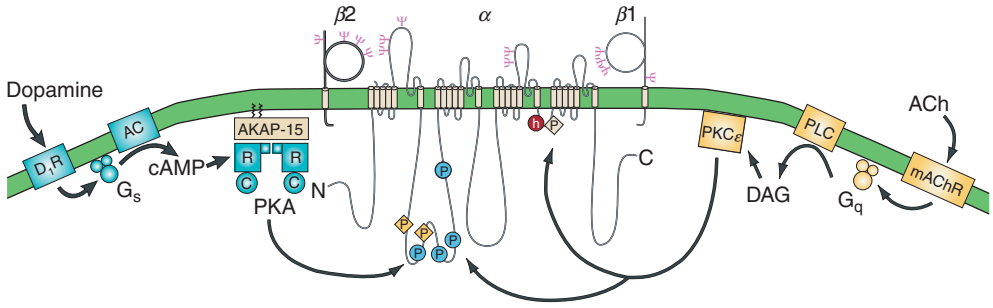

Figure 1 Neurotransmitter pathways that modulate Na^+ channels in hippocampal neurons. (Middle) The Na^+ channel α, β_1, and β_2 subunits are illustrated as transmembrane folding diagrams. Sites of PKA phosphorylation are indicated by the letter P blue circles. Sites of PKC phosphorylation are indicated by the letter P in yellow diamonds. (Left) The pathway for modulation of Na^+ channels by dopamine is illustrated. Dopamine binds to D1-like receptors (D_1R), which activate G_s and adenylyl cyclase (AC). The increase in cAMP activates PKA. PKA tethered to AKAP-15 phosphorylates the Na^+ channel on four sites in $L_{I–II}$ (serines 573, 610, 623, and 687). Two of these sites have been shown to be required for maximal PKA modulation of Na^+ channels: serine 573 and serine 687. (Right) The pathway for modulation by acetylcholine is illustrated. Acetylcholine binds to muscarinic receptors, which activate G_q and phospholipase C (PLC). The increase in diacylglycerol (DAG) activates protein kinase C, which phosphorylates Na^+ channels on serine 1506 in the inactivation gate and serines 554, 573, and 576 in $L_{I–II}$.

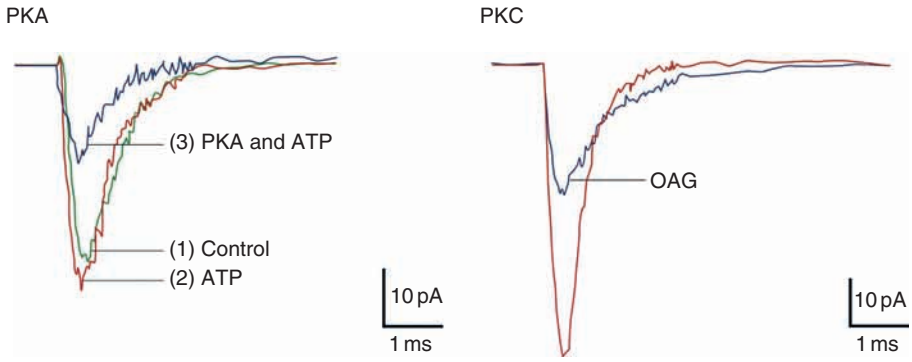

Figure 2 Regulation of Na^+ currents by PKA and PKC. (PKA) Macroscopic Na^+ currents in an excised membrane patch from a cell transfected with $Na_v1.2a$ under control conditions, after addition of ATP, and, finally, after addition of PKA in the presence of ATP. Note the reduction in peak Na^+ current without change in the kinetics of the current. (PKC) Whole cell Na^+ currents in a cell transfected with $Na_v1.2a$ before and after activation of PKC with oleylacetylglycerol (OAG). Note the reduction in peak Na^+ current and slowing of inactivation.

involves phosphorylation at multiple sites on the α subunit, as identified using a combination of biochemistry, mutagenesis, and electrophysiology. PKA-mediated reduction in peak Na^+ current at typical resting membrane potentials of −70 to −80 mV is primarily mediated by phosphorylation at a single consensus phosphorylation site centered at serine 573 in $L_{I–II}$ of $Na_v1.2$ channels (Figure 1).

The dual effects of PKC activation – reduction of peak Na^+ current and slowing of fast inactivation – are mediated by PKC-dependent phosphorylation of multiple sites on the α subunit. Reduction of the rate of inactivation requires phosphorylation of serine 1506 in the intracellular loop connecting homologous domains III and IV (Figure 1). Effects on fast inactivation due to phosphorylation of this site are expected since this region of the channel is proposed to be the inactivation gate. PKC-mediated reduction in peak Na^+ current requires additional phosphorylation of $L_{I–II}$ at serine residues 554 and 573 (Figure 1). However, a double mutation in which both of these serine residues have been converted to alanine does not completely prevent PKC-mediated reduction in peak Na^+ current amplitude, suggesting the existence of one or more additional PKC phosphorylation sites. The involvement of serine 573 in PKC-mediated reduction of peak current is interesting since this is an important PKA phosphorylation site, raising the possibility of dual kinase regulation at a single site.

Membrane Potential-Dependent, Convergent Regulation by PKA and PKC

Surprisingly, PKA-dependent reduction of peak Na^+ currents requires concomitant activation of PKC and phosphorylation of serine 1506 on the cytoplasmic

linker connecting homologous domains III and IV. PKA-dependent modulation of $Na_v1.2$ channels is voltage dependent in the range of −110 to −70 mV. At a holding potential of −110 mV, little effect of PKA is observed on Na^+ channels expressed in non-neuronal cells. Modulation at this negative membrane potential was recovered following concurrent activation of PKC. At more positive membrane potentials, the requirement for PKC activation is progressively reduced and substantial PKA modulation is observed at the holding potential of −75 mV without PKC activation. The voltage dependence of these effects of PKA phosphorylation provides an additional element of specificity and sensitivity to modulation of brain Na^+ channels. Evidently, the Na^+ channel is a point of convergence of regulation by membrane potential and neurotransmitter receptors acting through second messengers, PKA, and PKC.

Enhancement of the regulatory effects of PKA on $Na_v1.2$ channels by PKC requires both PKC- and PKA-dependent phosphorylation at additional sites on the α subunit. The increase in strength of the PKA modulation results in part from recruitment of regulation by PKA-dependent phosphorylation at serine 687 (Figure 1). No PKA-dependent modulation was observed for mutant S573A under control conditions. However, PKA modulation was recovered for this mutant during concurrent PKC activation, suggesting that PKA phosphorylation may be recruited at one or more additional sites in these circumstances. Systematic mutagenesis of PKA phosphorylation sites in $L_{I–II}$ revealed the requirement for PKA phosphorylation at serine 687. A double mutation at positions 573 and 687 (SS573/687AA) eliminated the effect of concurrent PKC activation at both the permissive and the nonpermissive membrane potentials. The potentiating effects of PKC were blocked in the presence of the PKA inhibitor peptide, demonstrating the requirement for PKA for phosphorylation of this site.

How does PKC activation recruit regulation by PKA-dependent phosphorylation at serine 687? Systematic mutation of known phosphorylation sites showed that PKC must also phosphorylate serine 576 in $L_{I–II}$ in order to enhance modulation by PKA (Figure 1). A single mutation at this position (S576A) eliminates the effects of concurrent PKC activation on PKA modulation of $Na_v1.2a$ channels at all membrane potentials tested. Thus, maximal modulation of $Na_v1.2a$ channels by PKA requires phosphorylation at three distinct sites – serine residues 573, 576, and 687 – and concurrent activation of both PKC and PKA. One possible explanation for this synergistic regulation is that PKC-dependent phosphorylation of serine 576 might bring about a conformational change in $L_{I–II}$ that increases the accessibility of serine 687 to PKA. This type of convergent regulation provides a novel mechanism by which information from multiple signaling pathways may be integrated depending on the state of the cell and its prior history of electrical activity.

Enhancement of Slow Inactivation by Protein Phosphorylation

In addition to their well-known fast inactivation process, which acts on the timescale of milliseconds and involves the fast inactivation gate in the intracellular loop connecting domains III and IV, Na^+ channels enter a distinct slow-inactivated state upon depolarization for hundreds of milliseconds to seconds. Recovery from this slow-inactivated state is also very slow, requiring seconds at a repolarized membrane potential for completion. Slow inactivation therefore makes Na^+ channels unavailable for activation by depolarizing stimuli, and this intrinsic, voltage-dependent slow-inactivation process of Na^+ channels is known to modulate the threshold and firing pattern of brain neurons.

Analysis of single channel recordings indicated that neither the single channel currents nor the number of available Na^+ channels are altered by PKA. Instead, the decrease in peak current results from an increase in null sweeps during which no Na^+ channel openings are observed after depolarization, consistent with slow inactivation. Studies show that the biophysical mechanism underlying this form of modulation is indeed slow inactivation. The voltage dependence and kinetics of reduction of Na^+ current by activation of PKA and PKC correlate quantitatively with the kinetics and voltage dependence of slow inactivation. Similarly, the kinetics of recovery from slow inactivation are precisely correlated with the kinetics of recovery from inhibition by PKA and PKC phosphorylation. Moreover, mutations of Na^+ channels that specifically impair slow inactivation also prevent the reduction of Na^+ current caused by activation of PKA and PKC in transfected nonneuronal cells expressing $Na_v1.2$ channels. Together, these results provide strong support for the conclusion that modulation of Na^+ channels by PKA and PKC is caused by enhancement of the intrinsic slow inactivation process.

The functional impact of enhanced slow inactivation on the firing properties of a cortical pyramidal cell is illustrated in Figure 3. Under control conditions, depolarization by steady injection of inward current into the cell produces trains of action potentials faithfully. In contrast, after activation of serotonin receptors ($5\text{-}HT_{2a/c}$) with the agonist 2,5-dimethyloxy-4-iodoamphetamine HCl (DOI; 10 μM), stronger depolarizing stimuli produce trains of action potentials

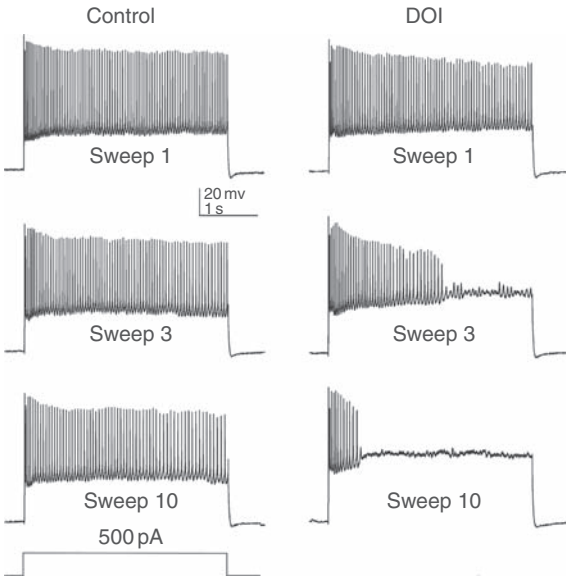

Figure 3 Change in firing properties of prefrontal cortical neurons accompanying neuromodulation of Na^+ channels. Action current recordings obtained under current clamp in the whole cell configuration from single prefrontal cortical neurons in slices of mouse brain. Repetitive stimulation of control neurons with depolarizing current yielded continuous trains of spikes, whereas spike failures were observed for neurons treated with DOI (10 μM) to activate serotonin receptors and PKC signaling. From Carr DB, Day M, Cantrell AR et al. (2003) Transmitter modulation of slow, activity-dependent alterations in sodium channel availability endows neurons with a novel form of cellular plasticity. *Neuron* 39: 793–806.

that terminate due to depolarization block. This block of firing occurs because the treated neurons have progressively fewer Na^+ channels that are available to sustain firing as the train continues at more depolarized potentials due to enhanced slow inactivation during the train. In this way, phosphorylation of Na^+ channels controls the input–output relationships of neurons.

Modulation of Na^+ Channels in Neurons

How widepsread is this complex regulation of Na^+ channel activity in neurons? This critical question has been analyzed in cultured brain neurons and in acutely dissociated neurons from hippocampus, striatum, cerebral cortex, and dorsal root ganglion.

Modulation of Na^+ channels by direct activation of PKA and PKC is observed in primary cultures of embryonic rat brain neurons, as in transfected cells. Activation of PKC reduces peak Na^+ current and slows inactivation, whereas activation of PKA decreases Na^+ channel activity without change in kinetics.

Acetylcholine, serotonin, and dopamine regulate the functional properties of the voltage-gated Na^+ channel in brain neurons via activation of PKC and PKA, respectively (Figure 1), and one can anticipate that other neurotransmitters that activate these kinases will have similar effects. Activation of muscarinic acetylcholine receptors coupled to phospholipase C exerts a PKC-dependent inhibitory effect on Na^+ current in acutely isolated hippocampal pyramidal neurons. Both peak transient and persistent Na^+ currents are reduced without significant change in the voltage dependence of activation or fast inactivation (Figure 2). The rate of inactivation is also appreciably slowed. These effects of muscarinic stimulation require the activation of PKC because they were eliminated in the presence of the PKC specific inhibitor $PKCI_{19–36}$ and were mimicked by the extracellular application of the membrane permeant PKC activator OAG.

The hippocampus receives dopaminergic innervation from the mesocorticolimbic dopamine system. The effects of dopamine are mediated by a family of five G-protein-coupled dopamine receptors that are broadly classified as D1-like (D1 and D5) or D2-like (D2–D4) based on pharmacology and coupling to intracellular second-messenger systems. Activation of D1-like dopamine receptors, which couple to the activation of adenylyl cyclase, decreases Na^+ current in acutely isolated hippocampal pyramidal neurons without altering the voltage dependence or kinetics of activation or fast inactivation. These effects require PKA activity because they were eliminated in the presence of the specific PKA inhibitor $PKAI_{5–24}$ and were mimicked by a PKA activator. Na^+ channels in transfected cells as for (Figure 2). D1/PKA modulation in hippocampal neurons is voltage dependent in the range of −110 to −70 mV and is enhanced by concurrent activation of PKC at all membrane potentials tested. Thus, Na^+ channels in hippocampal pyramidal neurons are modulated via a voltage-dependent, convergent mechanism by neurotransmitter activation of the PKA- and PKC-signaling pathways as predicted from studies of transfected cells.

Dopamine-mediated regulation of neuronal excitability is also critically important in the neostriatum, in which the dopaminergic projection from the substantia nigra controls the firing of neostriatal neurons, and loss of this regulatory system results in the movement disorders associated with Parkinson's disease. Activation of D1-like dopamine receptors on medium spiny neostriatal neurons inhibits peak Na^+ current in a reversible, dose-dependent manner without significant effects on voltage- and time-dependent properties of the current. Persistent Na^+ current is also inhibited. This effect requires activation of PKA since it is mimicked by intracellular loading of PKA and prevented in the presence of a PKA inhibitor peptide, $PKAI_{5–24}$. Conversely, activation of D2 receptors, which are coupled to the inhibition of adenylyl cyclase and consequent reduction of PKA activity, increases the amplitude of Na^+ current in medium spiny neurons.

Phosphorylation-dependent neuromodulation of voltage-gated Na^+ channels has been demonstrated in neocortical neurons as well. Activation of D1-like dopamine receptors reduces peak Na^+ currents without a change in kinetics or voltage dependence of activation or fast inactivation in dissociated cortical neurons under voltage clamp. This effect is mimicked by cAMP analogs, implicating the PKA pathway. Activation of D1 receptors had a smaller effect on the persistent Na^+ current (I_{NaP}) remaining after fast inactivation. As in hippocampal neurons, muscarinic receptor activation in rat neocortical neurons inhibited the transient and persistent Na^+ currents evoked by slow depolarizing voltage ramps or by long depolarizing voltage steps. In cortical neurons, persistent Na^+ current is active near critical subthreshold voltages where other ionic currents are small, and modulation of its functional properties is thought to contribute to synaptic integration, regulation of intrinsic firing patterns, and active back-propagation of action potentials from the soma into the dendrites.

Sensory neurons express multiple Na^+ channel isoforms, including $Na_v1.8$ and $Na_v1.9$, which mediate slowly gated Na^+ currents that are tetrodotoxin (TTX) resistant. Neuromodulation of TTX-resistant Na^+ current (TTX-R I_{Na}) in small-diameter, nociceptor-like rat sensory neurons is mediated by proinflammatory, hyperalgesic agents, including serotonin, prostaglandin E_2 (PGE_2), and adenosine. These agents increase the amplitude of the TTX-resistant Na^+ current, accompanied by a negative shift in the voltage dependence of activation and increase in the rates of activation, deactivation, and inactivation. Direct activation of adenylyl cyclase or PKA mimicked the effects of PGE_2 and intracellular perfusion with a PKA inhibitor blocked the effects of PGE_2, demonstrating the requirement for PKA in this form of Na^+ channel modulation. Mutation of consensus PKA phosphorylation sites in $L_{I–II}$ greatly reduced PKA modulation compared to wild-type channels, leading to the conclusion that the major substrates involved in PKA modulation of the TTX-resistant channel are located in $L_{I–II}$ as for $Na_v1.2$. PKC plays a permissive role in the PGE_2-mediated modulation of TTX-resistant Na^+ current, similar to that reported for dopamine modulation of Na^+ channels in hippocampal neurons. However, in contrast to $Na_v1.2$ channels, activation of either PKA or PKC increases peak TTX-resistant Na^+ current. PKA-mediated modulation of TTX-resistant Na^+ current was partially attenuated by inhibition of PKC, suggesting a requirement for PKC in the PGE_2-mediated modulation of TTX-resistant Na^+ current. The sites at which PKC exerts its effects on TTX-resistant Na^+ current are unknown. However, it is intriguing that PKA and PKC act synergistically to increase the activity of $Na_v1.8$ channels, at least in part through phosphorylation of sites in $L_{I–II}$, whereas the same two kinases reduce the activity of $Na_v1.1$ and $Na_v1.2$ channels by a similar, synergistic mechanism based on phosphorylation of sites in $L_{I–II}$. Evidently, the molecular mechanisms of regulation of these different channel types by protein phosphorylation are similar but the effects on channel activity are opposite.

Role of Kinase Anchoring in Na^+ Channel Modulation in Neurons

Specificity and speed of kinase signaling are often achieved through association with anchoring or adaptor proteins that target kinases to subcellular organelles or to specific substrates via protein–protein interactions. PKA is localized by protein–protein interactions between the regulatory (R) subunit dimer of PKA and A-kinase anchoring proteins (AKAPs), a family of functionally related proteins defined by their ability to bind R subunits of PKA and target PKA to specific compartments or substrates. AKAPs contain an amphipathic helix that inserts into a hydrophobic pocket in the R subunit dimer. Consistent with this theme, modulation of the Na^+ channel by dopamine receptors in hippocampal neurons requires targeted localization of PKA near the channel via AKAP15 (Figure 1), which was characterized initially in studies of calcium channel modulation. This novel 81-amino acid AKAP has an N-terminal lipid anchor for membrane targeting, an amphipathic helix for PKA binding, followed by a site for interaction with Na^+ channels via a modified leucine zipper motif. PKA, AKAP15, and the Na^+ channel therefore form a signaling complex that controls channel activity. The site of binding of AKAP15 on the Na^+ channel is located at the beginning of the large intracellular loop connecting domains I and II ($L_{I–II}$), which also contains the sites of PKA phosphorylation and regulation (Figure 1). Thus, AKAP15 targets PKA to the Na^+ channel near its substrate sites.

Functional Consequences of Na^+ Channel Modulation

Neuromodulation of Na^+ channels by protein phosphorylation is implicated in diverse aspects of control of firing properties of neurons in the brain and sensory ganglia. Several examples are described next.

Neostriatal neurons can function in two states with different input–output relationships, designated upstate and downstate. For a cell at rest (approximately-85 mV) or in the downstate, Na^+ channels are largely

inactive. D1-mediated inhibition of peak Na^+ current via PKA-dependent phosphorylation increases the threshold for action potential firing and favors the downstate. Consistent with this mechanism, dopamine acting at D1 receptors inhibits action potential firing of neostriatal neurons. D2-mediated augmentation of whole-cell Na^+ current via inhibition of cAMP synthesis speeds depolarization and increases the rate of entry into the upstate. The reciprocal effects of these dopamine receptors on Na^+ channels contribute to the stability of the downstate and upstate of striatal neurons and to the transitions between them.

The nucleus accumbens is a primary site mediating reward behavior, and it is thought to be directly involved in reinforcing and addictive behaviors in response to drug use. Cocaine exerts its pharmacological effects by inhibition of dopamine-reuptake transporters, thereby increasing the extracellular concentration of dopamine in the vicinity of dopamine-responsive neurons. Na^+ currents of medium spiny neurons in the nucleus accumbens are inhibited by dopamine acting through D1 receptors, as in other striatal neurons. Remarkably, chronic (2-week) treatment with cocaine sensitizes these neurons to the effects of dopamine and causes a marked decrease in Na^+ currents that persists for 3 days after drug withdrawal. These results demonstrate a temporal correlation between persistent inhibition of Na^+ currents and the acute withdrawal effects of anhedonia and craving that accompany cessation of cocaine treatment. Neuromodulation of Na^+ channels via the dopamine/PKA pathway may be critically involved in the complex behavioral responses to chronic treatment with drugs of abuse.

The effects of dopamine released from the mesocorticolimbic system in the hippocampus are primarily inhibitory, resulting in elevation of spike threshold, slower firing rates, and, in some cases, membrane hyperpolarization. Although it is believed that these effects are primarily a consequence of activation of slow Ca^{2+}-dependent K^+ channels, they may actually be mediated by a concerted action of the transmitter on several ion channels, including voltage-gated Na^+ channels. The inhibitory effects of activation of D1-like receptors are consistent with a role for modulation of Na^+ currents as a contributing factor to the inhibitory effects of dopamine. The hippocampus also receives a rich cholinergic innervation from neurons of the basal forebrain, whose effects are mediated primarily by muscarinic receptors. Slowing of fast inactivation of Na^+ currents by muscarinic receptor action via the PKC pathway would subtly increase the duration of the action potential and possibly alter the pattern of action potential firing. Reduction of peak Na^+ current by this pathway would more strongly influence the functional properties of target neurons by shifting the threshold for action potential generation toward more depolarized potentials and reducing the frequency at which the cell is capable of generating action potentials. Consistent with these expectations, cholinergic input to the hippocampus acting through muscarinic acetylcholine receptors and PKC inhibits intrinsic bursting activity of CA1 neurons, converting their firing pattern from phasic bursting separated by quiescent periods to tonic firing of single spikes. These effects are correlated with a reduction of persistent Na^+ current in CA1 neurons, which is likely to be accompanied by a similar reduction in peak Na^+ current. Since generation of bursts is a major mechanism of information transmission, these effects are likely to strongly affect input–output relationships in hippocampal neurons.

Similar modulatory effects of dopamine and acetylcholine are observed in cortical pyramidal neurons. Cholinergic input to the cortex from the nucleus basalis causes a shift in spike generation from phasic bursts to tonic spiking. Dopaminergic input to the prefrontal cortex reduces Na^+ currents and increases the threshold for action potential generation. Thus, neuromodulation of Na^+ channels may contribute to cellular plasticity in both hippocampus and cerebral cortex.

Hyperalgesia is associated with abnormal bursting activity in primary sensory neurons involved in nociception. Heightened excitability in these neurons occurs within moments of an inflammatory stimulus and involves a negative shift in response threshold and an increase in spontaneous activity. Hyperexcitability is correlated with PKA/PKC-dependent upregulation of the TTX-resistant Na^+ current. Application of PGE_2 increases the excitability of rat dorsal root ganglion neurons by reducing firing threshold and altering the response to stimulation so that a single evoked action potential is converted to a train of action potentials. Expression of the TTX-resistant Na^+ channels is restricted to a subpopulation of sensory neurons that are likely to be involved in nociception, suggesting a specialized role for TTX-resistant Na^+ channels in mediating pain responses. Deletion of the gene encoding $Na_v1.8$ causes specific (but mild) deficits in nociception. The biophysical properties of TTX-resistant Na^+ channels make them ideal candidates for mediating repetitive firing behavior in response to prolonged membrane depolarization. Moreover, inflammatory mediators, including PGE_2, serotonin, and adenosine, decrease the activation threshold, increase the rates of activation and inactivation, and increase the magnitude of the TTX-resistant Na^+ current, consistent with the effects observed for $Na_v1.8$ expressed in transfected cells. This modulation is inhibited by

activation of μ-opioid receptors that block inflammatory hyperalgesia. Together, these results provide strong support for the idea that upregulation of Na^+ channel function is an essential component of the hyperalgesic response, and $Na_v1.8$ and $Na_v1.9$ are now considered to be prime therapeutic targets for treatment of intractable pain.

In the retina, the bipolar neurons control the dynamic response to light intensity. In dim conditions, Na^+ currents amplify the inputs to the bipolar cells and increase their activity. In bright light, dopamine released from amacrine cells inactivates Na^+ channels and reduces bipolar cell responses. These results show that neuromodulation of Na^+ channels by dopamine contributes significantly to light adaptation.

Other Regulatory Pathways

Neuromodulation of Na^+ channels by other cellular kinase signaling pathways may also have an important effect on neuronal development and function. Phosphorylation and dephosphorylation of Na^+ channels on tyrosine residues by specifically bound Fyn kinase modulates Na^+ channel inactivation, and receptor tyrosine phosphatase-β is physically associated with Na^+ channels and may regulate their function in response to cell–cell interactions. Neuromodulation of voltage-gated Na^+ channels may be important in neuroprotection during anoxia, and the NO pathway may regulate Na^+ channels under physiological and pathological conditions. Direct interactions of Na^+ channels with G-proteins may regulate voltage-dependent gating and persistent Na^+ currents. It is likely that we are only beginning to understand the broad influence of Na^+ channel regulation in mediating functional responses in the central nervous system. A greater understanding of the regulatory systems that target voltage-gated Na^+ channels and mediate different forms of cellular plasticity may provide insight into the regulation of electrical excitability and the mechanisms of many neurological disorders and may help to identify new therapeutic interventions.

See also: Addiction: Neurobiological Mechanism; Cyclic AMP (cAMP) Role in Learning and Memory; Dendritic Signal Integration; Dorsal Root Ganglion Neurons; Drug Addiction: Behavioral Neurophysiology; Drugs Addiction: Actions; Ion Channel Localization in Cell Bodies and Dendrites.

Further Reading

Calabresi P, Misgeld U, and Dodt HU (1987) Intrinsic membrane properties of neostriatal neurons can account for their low level of spontaneous activity. *Neuroscience* 20: 293–303.

Cantrell AC and Catterall WA (2001) Neuromodulation of sodium channels: An unexpected form of cellular plasticity. *Nature Reviews Neuroscience* 2: 397–407.

Cardenas CG, Del Mar LP, Cooper BY, and Scroggs RS (1997) $5HT_4$ receptors couple positively to tetrodotoxin-insensitive sodium channels in a subpopulation of capsaicin-sensitive rat sensory neurons. *Journal of Neuroscience* 17: 7181–7189.

Carr DB, Day M, Cantrell AR, et al. (2003) Transmitter modulation of slow, activity-dependent alterations in sodium channel availability endows neurons with a novel form of cellular plasticity. *Neuron* 39: 793–806.

Catterall WA (2000) From ionic currents to molecular mechanisms: The structure and function of voltage-gated sodium channels. *Neuron* 26: 13–25.

Chen Y, Yu FH, Surmeier DJ, Scheuer T, and Catterall WA (2005) Neuromodulation of sodium channel slow inactivation via cAMP-dependent protein kinase and protein kinase C. *Neuron* 49: 409–420.

England S, Bevan S, and Docherty RJ (1996) PGE_2 modulates the tetrodotoxin-resistant sodium current in neonatal rat dorsal root ganglion neurones via the cyclic AMP–protein kinase A cascade. *Journal of Physiology* 495: 429–440.

Fitzgerald EM, Okuse K, Wood JD, Dolphin AC, and Moss SJ (1999) cAMP-dependent phosphorylation of the tetrodotoxin-resistant voltage-dependent sodium channel SNS. *Journal of Physiology* 516: 433–446.

Gold MS, Levine JD, and Correa AM (1998) Modulation of TTX-R I_{Na} by PKC and PKA and their role in PGE_2-induced sensitization of rat sensory neurons *in vitro*. *Journal of Neuroscience* 18: 10345–10355.

Gold MS, Reichling DB, Shuster MJ, and Levine JD (1996) Hyperalgesic agents increase a tetrodotoxin-resistant Na^+ current in nociceptors. *Proceedings of the National Academy of Sciences of the United States of America* 93: 1108–1112.

Goldin AL, Barchi R, Caldwell J, et al. (2000) Nomenclature of voltage-gated sodium channels. *Neuron* 28: 365–368.

Numa S and Noda M (1986) Molecular structure of sodium channels. *Annals of the New York Academy of Sciences* 479: 338–355.

Surmeier DJ and Kitai ST (1997) State-dependent regulation of neuronal excitability by dopamine. *Japanese Journal of Clinical Psychopharmacology* 17: 105–110.

Zhang X, Hu X, and White FJ (1998) Whole-cell plasticity in cocaine withdrawal: Reduced sodium currents in nucleus accumbens neurons. *Journal of Neuroscience* 18: 488–498.

Neuromorphic Systems

G Indiveri, S-C Liu, T Delbrück, and R Douglas, University of Zurich, Zurich, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Neuromorphic Engineering

Neuromorphic engineering is concerned with the design and fabrication of artificial neural systems whose architecture and design principles are based on those of biological nervous systems. Neuromorphic systems of neurons and synapses can be implemented in the electronic medium CMOS (complementary metal oxide semiconductor) using hybrid analog/digital VLSI (very large-scale integrated) technology.

The notion of neuromorphic engineering has appeared in various forms during the last few centuries: From the French and Swiss mechanical automata of the eighteenth to nineteenth centuries (e.g., Vaucanson's duck) through the biological robots of the 1950s (e.g., Grey Walter's turtle) to the early neural electrical circuits of the 1960s and 1970s (e.g., Fukushima in Japan, Mueller at the University of Pennsylvania, and Lettvin at MIT). The concept took root again at Caltech during the mid-1980s, this time in the research of Carver Mead, who had already made major conceptual contributions to the design and construction of digital VLSI systems. He recognized that the use of transistors for computation had changed very little from the time when John Von Neumann first proposed the architecture for the programmable serial computer. The bit-perfect, strictly synchronous, and largely serial processing strategy required by present computers is untenable in the long run because it does not scale. Moreover, this design implies an exponentially greater power consumption for larger systems and exponentially higher costs of constructing, testing, and maintaining elements that must be fault-free in operation.

The design of biological neural computation is very different from that of modern computers. Neuronal networks process information using energy-efficient, asynchronous, event-based methods. Biology uses self-construction, self-repair, and self-programming, and it has learned how to flexibly compose complex behaviors from simpler elements. Of course, these biological abilities are not yet understood. But they offer an attractive alternative to conventional technology and have enormous consequences for future artificial information processing and behavior systems.

The challenge for neuromorphic engineering is to explore the methods of biological information processing in a practical electrical engineering context. Can existing CMOS VLSI technology be deployed in a novel way that reflects more closely the neuronal approach to computation? And can we thereby reap some of the computational facility, speed, and efficiency that biological nervous systems bring to the solution of real-world problems? Conversely, can we, by engineering artifacts, gain new insights into how biology organizes computation?

Digital and Analog in Neuromorphic VLSI Systems

The majority of integrated circuits represent numbers as binary digits. Binary digits are used because it is possible to standardize the behavior of transistors so that their state can be determined reliably to a single bit of accuracy. The reliable bits can then be combined to encode variables to an arbitrarily high precision. It is the combination of this precision and the synchronous, serial model of processing that has been the foundation of the long boom in digital electronics. One unnecessary consequence of this exuberance is that modern computers use deterministic high precision to deal even with real-world tasks whose variables can often be encoded with only a few bits.

For many such problems, particularly those in which the input data are ill-conditioned and the computation can be specified in a relative matter, biological solutions are many orders of magnitude more effective than those we have been able to implement using digital methods. This advantage is due principally to biology's use of elementary physical phenomena as computational primitives and to the representation of information by the relative values of analog signals rather than by the absolute values of digital signals. Typically, it is this style of processing that neuromorphic engineers explore. Their systems are large collections of communicating computational primitives implemented either in analog or, more commonly, in hybrid analog–digital circuits.

Analog computing has a number of interesting features. First, analog processors are inherently more dense than digital ones because the individual electrical nodes of analog circuits can represent multiple bits of information. Second, unlike conventional digital circuits, whose coordination is governed by a global clock, analog circuits are naturally synchronized in physical time (as opposed to being synchronized against clock time). This property is particularly useful for systems that model and interact with real-world processes because, if the time

constants of the analog circuits are appropriately set, their processing can be scaled to match those of the real-world processes with which they interact.

Unfortunately, an analog process is much more difficult to implement than its digital counterpart. Usually an analog processor must be purpose-built and does not have the general character of digital systems. The analog circuits are difficult to implement because the physics of the material used to construct the circuits plays an important role in the solution of the problem. It is also difficult to control the physical properties of micron-size devices such that their analog characteristics are well matched. This matching of analog device characteristics is a major difficulty facing an analog designer. And digital machines have an advantage over analog ones when high precision is required. The analog approach usually requires adaptive techniques to mitigate the effects of component differences. However, this kind of adaptation leads naturally to systems that learn about their environment.

Digital systems depend on a small set of primitives that can support universal logical computation. For example, any logic function can be implemented using only NAND gates. In principle, a similar mathematical universality is possible in analog circuits by using primitive circuits for addition and multiplication. However, the neuromorphic engineer is usually more concerned with identifying the primitives needed to emulate neuronal processing than with mathematics *per se*.

For many crucial neuronlike operations, dedicated analog processors can be constructed much more compactly than their digital counterparts. Some neuromorphic primitives are obvious, such as the simple addition of currents by joining their conductors. Others provide useful generic functions, such as logistic and tanh circuits. Yet others are simple but sophisticated processors (e.g., Mahowald and Mead's elementary contrast reporting pixel and Delbrück's correlation circuit) (Figure 1).

These circuits are usually based on CMOS field effect transistors (FETs), whose conductivity can be altered by an applied electric field. A FET is a device in which the flow of current between its source and drain terminals is controlled by a voltage applied to its gate terminal. The relationship between the gate voltage and the current that flows across the transistor's channel is rather subtle. There are two primary regimes of behavior. In the subthreshold regime, in which small voltages are applied between the gate and the source, the transistor current grows exponentially with the gate voltage. In above-threshold regime, in which larger voltages are applied, the current grows more slowly. Conventionally, transistors

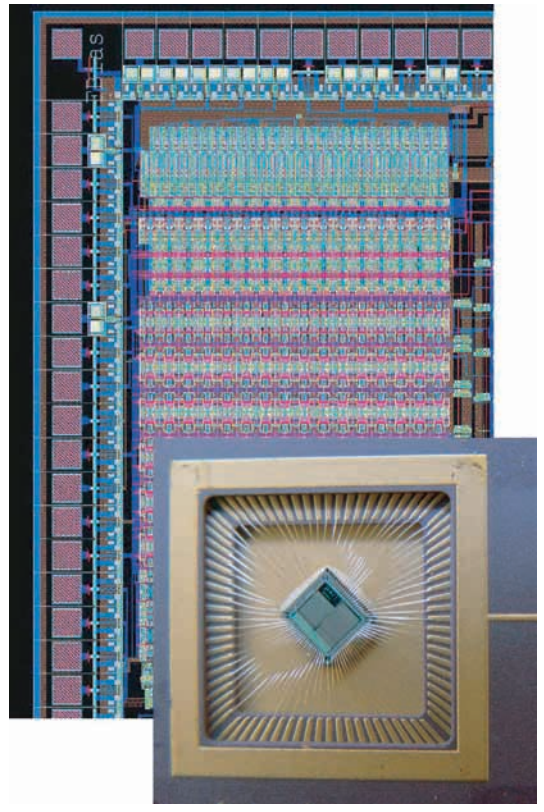

Figure 1 A neuromorphic VLSI chip. The background shows a 1×1 mm section of the chip layout, comprising an array of silicon synapses, a row of silicon neurons, and peripheral interfacing circuits. The foreground shows a picture of the 3×3 mm chip bonded to its package. VLSI, very large-scale integrated.

are operated in the above-threshold regime, and in the extreme case of digital circuits only very large gate voltages are used so that the transistors are either fully off or fully conducting.

By contrast, neuromorphic circuits typically exploit the otherwise unpopular subthreshold regime because this regime offers some striking advantages for nondigital circuits. The exponential increase in transistor current with gate voltage means that the gain (loosely defined as the change in current for a given change in voltage) of the subthreshold transistor is very high, and gain is the lever of information processing. The exponential property can be used to compose interesting analog circuits with natural exponential and logarithmic properties that are useful for emulating neuronal properties. A further feature of subthreshold circuits is that the absolute currents flowing through the transistors are very small, so these circuits operate at very low power. The downside is that the signals are very small, so the circuits are susceptible to noise and fabrication variations.

All these interesting properties offer the possibility of constructing very low-power brainlike electronic

circuits that operate in real time. An important strategic consideration in this quest is that the basic CMOS VLSI technology does not have to be developed by the neuromorphic engineering community. The development of VLSI technology is fueled by the digital information technology (IT) market, so those developments need only be applied to the needs of subthreshold analog neuromorphic engineering. The very same technology used to construct digital computers can also be used (in subthreshold) to construct computational primitives for systems of neuronlike processors. Unlike digital systems, which have a relatively small number of general-purpose processors that process a command stream sequentially, the performance of neuromorphic systems depends on the parallel configuration of large numbers of these special-purpose primitives.

Emulation versus Simulation

What advantages do these neuromorphic circuits have? One advantage is that they offer a medium in which neuronal networks can be emulated directly in hardware rather than simply simulated on a general-purpose computer. However, it should be noted that the uses of emulation are not the same as those of simulation. Digital simulation is generally used to explore the quantitative behavior of neuronal systems. Because they are composed of large numbers of nonlinear elements and have a wide range of time constants, their mathematical behavior can rarely be solved analytically. But the speed of simulation is limited by the shortest time constant of the problem, so simulation performance slows dramatically as the number and degree of coupling of elements increase.

By contrast, neuromorphic emulations operate in real time, and the speed of the network is independent of the number of neurons or their coupling. However, analog circuits provide only a good qualitative approximation of the exact performance of simulated neurons. Moreover, the design of special-purpose hardware is a significant investment, particularly if it is analog hardware because analog VLSI (aVLSI) design remains very much an art form. Designing robust novel circuits depends on engineers with considerable experience, preferably working in small groups with sufficient collective knowledge to monitor and criticize one another's design assumptions and circuit layouts. So, in their present form, neuromorphic circuits are not suitable for quantitative simulation.

On the other hand, where the electronic neural circuits do provide a tangible advantage is in the investigation of questions concerning the strict real-time interaction of the system with its environment, possibly taking sensory input from neuromorphic sensors such as silicon retinas or silicon cochleas. Neuromorphic retina and cochlea sensors represent examples of extremely successful systems built with these design principles. Neuromorphic retinas are vision sensors that can adapt in real time to local changes in illumination over a wide range of intensities and that reproduce many of the functionalities of the retina's outer plexiform layer. Thanks to their adaptation and real-time response properties, they are currently being applied to machine vision tasks that are difficult to solve using conventional approaches. These include, for example, traffic monitoring in open uncontrolled environments such as freeways or fast robotics, in which the low latency and sparsity of address–event representation (AER) sensor data are very advantageous. Similarly, silicon cochleas convert auditory signals into sequences of spikes generated by populations of silicon neurons arranged tonotopically along the frequency axis, similar to the inner-ear hair cells located along the basilar membrane of real cochleas. These devices have promising application domains, both as alternative low-power solutions for cochlear implants and as computational devices for engineering applications.

Neurons in Silicon

One of the central goals of neuromorphic engineering is to capture the computational principles of neurons and their networks in hardware. A cornerstone of this quest has been the development over the last decade of hybrid analog–digital VLSI neurons, together an infrastructure for composing networks of these neurons. It is now possible to assemble quite complex systems of such neurons.

Integrate-and-Fire Models

Neurons communicate by pulses (or spikes) that propagate along electrically lossy point-to-point wires that are the axons. Real neurons have a complex morphology and even more complex biophysics, whose full emulation is beyond the reach of present electronic technology. Nevertheless, the integrate-and-fire neuron (I&F), which is a bold simplification of real neurons, has proved to have significant explanatory power in understanding the behavior of neuronal networks both in theory and simulation. So far, it is this simplified model that has received the most attention in neuromorphic circles.

I&F neurons integrate presynaptic input currents and generate a voltage pulse analogous to an action potential when the integrated voltage reaches a threshold. Many variants of these circuits were built during the 1950s and 1960s using discrete electronic components. The first simple VLSI version was probably the

Axon Hillock circuit, built by Mead in the late 1980s. In this circuit, a capacitor that represents the somatic membrane capacitance integrates current input to the neuron. When the capacitor voltage crosses a threshold, it is reset by a positive feedback loop. More recent models include additional neural characteristics, such as spike-frequency adaptation properties and refractory period mechanisms. These features can be combined in a compact low-power circuit such as the one shown in Figure 2.

Conductance-Based Models

These VLSI I&F neurons provide convenient approximations of the behavior of neuronal somata without committing to the overhead of emulating the plethora of voltage-dependent conductances and currents present in real neurons. But, if necessary, these conductances can be emulated using subthreshold CMOS circuits. For example, the silicon neurons of Mahowald and Douglas are composed of connected compartments, each of which is populated by modular subcircuits that emulate particular ionic conductances. The dynamics of these types of circuits is qualitatively similar to the Hodgkin–Huxley mechanism without implementing their specific equations. An example of this type of silicon neuron circuit is shown in Figure 3.

In this circuit, the membrane capacitance C_{mem} is connected to a transconductance amplifier that implements a conductance term, whose magnitude is modulated by the bias voltage G_{leak}. This passive leak conductance couples the membrane potential to the potential of the ions to which the membrane is permeable (E_{leak}). Similar strategies are used to implement the active sodium and potassium conductance circuits. In these cases, transconductance amplifiers configured as simple first-order low-pass filters are used to emulate the kinetics of the conductances. A current mirror is used to subtract the sodium activation and inactivation variables (I_{Naon} and I_{Naoff}) rather than multiply them, as in the Hodgkin–Huxley formalism. Additional current mirrors half-wave rectify the sodium and potassium conductance signals, so that they are never negative.

Several other conductance modules have been implemented using these principles. For example,

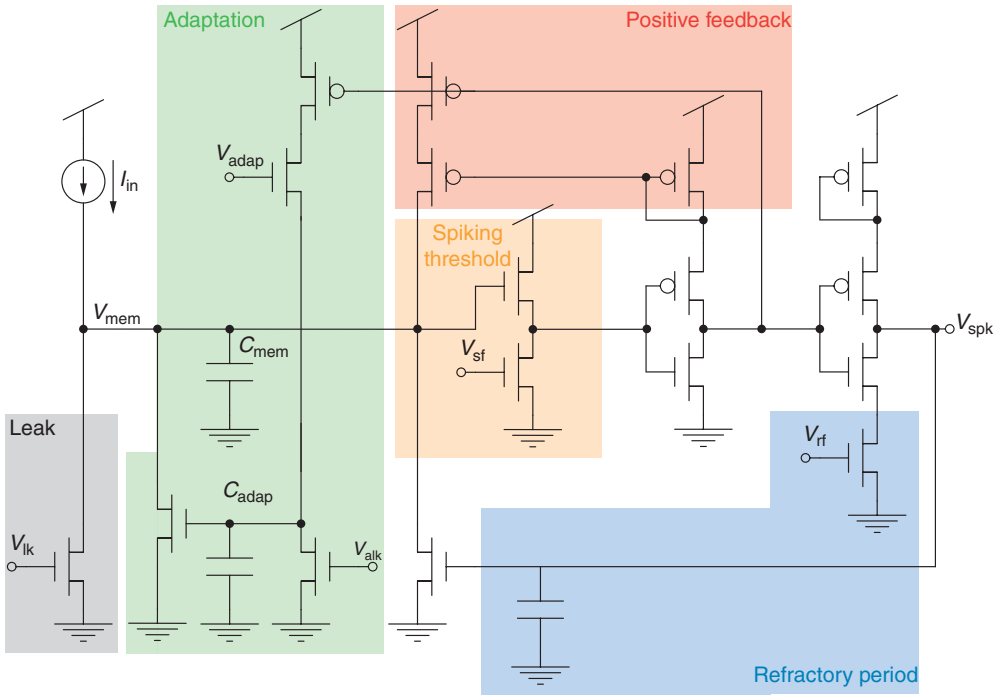

Figure 2 Diagram of an integrate-and-fire neuron. The input current I_{in} is integrated on to the neuron's membrane capacitor C_{mem} until the spiking threshold is reached. At that point the output signal V_{spk} goes from zero to the power supply rail, signaling the occurrence of a spike. Then the membrane capacitor is reset to zero, and the input current starts to be integrated again. The leak module implements a current leak on the membrane. The spiking threshold module controls the voltage at which the neuron spikes. The adaptation module subtracts a firing rate dependent on current from the input node. The amplitude of this current increases with each output spike and decreases exponentially with time. The refractory period module sets a maximum firing rate for the neuron. The positive feedback module is activated when the neuron begins to spike and is used to reduce the transition period in which the inverters switch polarity, dramatically reducing power consumption. The circuit's biases (V_{lk}, V_{adap}, V_{alk}, V_{sf}, and V_{rf}) are all subthreshold voltages that determine the neuron's properties.

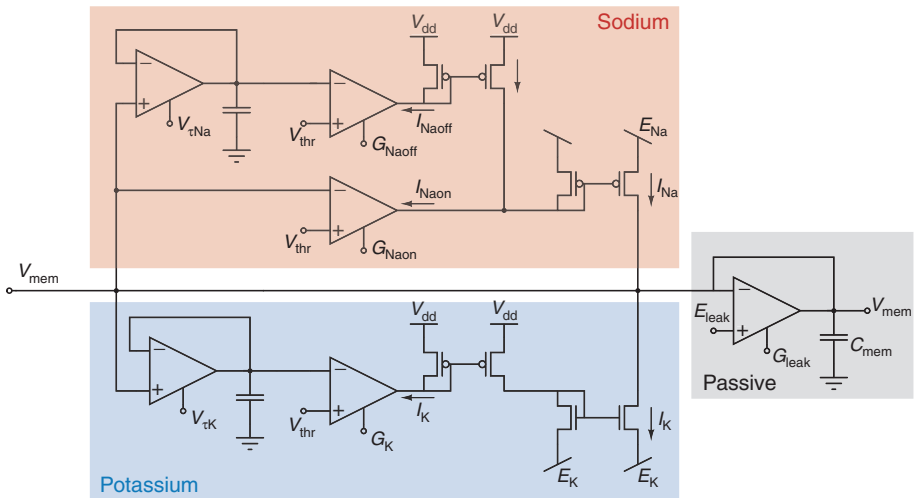

Figure 3 A conductance-based silicon neuron. The passive module implements a conductance term that models the passive leak behavior of a neuron; in the absence of stimulation, the membrane potential V_{mem} leaks to E_{leak} following first-order low-pass filter dynamics. The sodium module implements the sodium activation and inactivation circuits that reproduce the sodium conductance dynamics observed in real neurons. The potassium module implements the circuits that reproduce the potassium conductance dynamics. The bias voltages G_{leak}, $V_{\tau Na}$, and $V_{\tau K}$ determine the neuron's dynamic properties, whereas G_{Naon}, G_{Naoff}, G_K, and V_{thr} are used to set the silicon neuron's action potential characteristics.

there are modules for the persistent sodium current, various Ca^{2+}currents, Ca^{2+}-dependent potassium current, potassium A current, nonspecific leak current, and an exogenous (electrode) current source. The prototypical circuits can be modified in various ways to emulate the particular properties of a desired ion conductance. For example, some conductances are sensitive to Ca^{2+} concentration rather than membrane voltage and require a separate voltage variable representing free Ca^{2+} concentration.

Synaptic conductance circuits have also been developed using similar principles. But synaptic conductances are sensitive to ligand concentrations rather than membrane voltage, so these circuits require a voltage variable that represents neurotransmitter concentration. This set of ionic conductances, with their different dependencies and time constants, gives rise to state-dependent dynamics when embedded in the resistive capacitive network that represents the passive electrical structure of a neuron. These circuits can be composed in this way to model in detail the electrophysiological behavior of several types of neurons, for example, pyramidal cells.

Axons, Action Potentials, and the Address–Event Representation

Biological neurons communicate with one another using dedicated point-to-point axons. The all-or-nothing action potential can be translated into a discrete level signal, that is robust against noise and interchip variability, and can be conveniently transmitted between chips and easily interfaced to standard logic and computer systems. In the AER method developed by Mahowald and others, the action potentials generated by a particular neuron are transformed into an address that identifies the source neuron and then broadcast on a common data bus. Many silicon neurons can share the same bus because switching times in CMOS and on the bus are much faster than the switching times of neurons. Events generated by silicon neurons can be broadcast and removed from a data bus at frequencies greater than a megahertz. Therefore, more than 1000 address events could be transmitted in the time it takes one neuron to complete a single action potential. The addresses are detected by the target synapses, which then initiate their local synaptic action (see Figure 4). In this way, synapses that should be connected to the source neuron initiate a synaptic input whenever the source neuron spikes. These multiplexing strategies are most effective if, like their biological counterparts, the neurons of a network are only sparsely active.

Event-based digital encoding methods are convenient for configuring large networks because the network connectivity can be implemented by a programmable digital address mapper. The mapper receives an address event from a source neuron and translates it into one or more target addresses that are transmitted on the presynaptic bus. Because the mapping table is programmable, arbitrary network topologies can be set and modified dynamically. This method greatly facilitates the configuration and testing of large multichip systems.

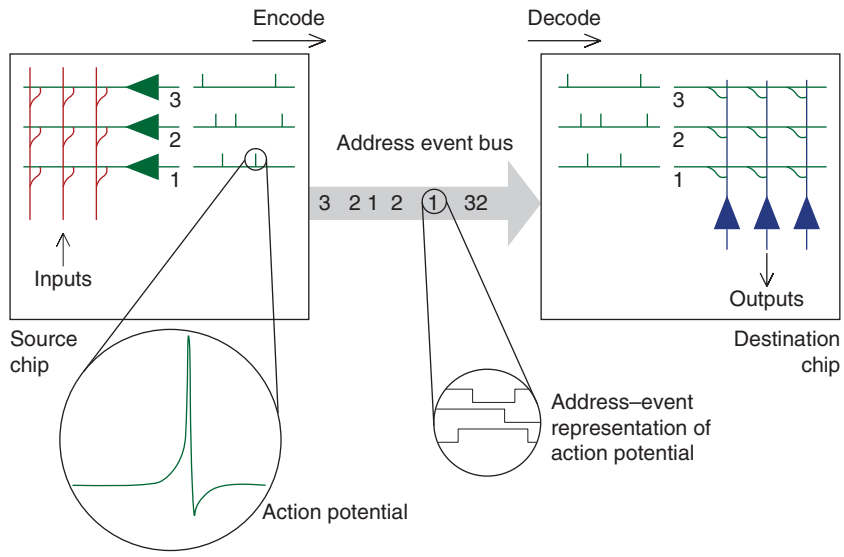

Figure 4 Asynchronous communication scheme between two chips using the address–event representation (AER). When a neuron on the source chip generates an action potential, its address is placed on a common digital bus. The receiving chip decodes the address events and routes them to the appropriate synapses.

Synapses in Silicon

Silicon synapses come in many forms, depending on the underlying model and the circuit implementation. The models range from simple current source models to more complicated conductance-based models including approximations of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and *N*-methyl-D-aspartate (NMDA) excitatory synapses, and potassium-mediated and chloride-mediated (shunting) inhibitory synapses. An example of a synaptic circuit that can reproduce the exponential dynamics observed in real synapses is shown in Figure 5.

Plasticity and Learning

One of the key properties of biological synapses is their short- and long-term plasticity. Short-term plasticity produces a dynamic modulation of the synaptic strength by the timing of the input spikes alone, whereas long-term plasticity produces sustained changes in synaptic strength that are induced by the correlations in the spiking activity of the pre- and postsynaptic processes.

Circuits have been developed that emulate the short-timescale synaptic depression; also, various learning circuits that implement long-term plasticity of synapses use spike-timing information and/or rate information. Spike-timing-dependent plasticity (STDP) mechanisms are particularly well suited to the VLSI neuromorphic networks described here because these networks process AER signals. In STDP, the precise timing of spikes generated by the pre- and postsynaptic neurons have an important role in shaping the synaptic efficacy. If a presynaptic spike arrives at the synaptic terminal before a postsynaptic spike is emitted, during a window of causality, the synaptic efficacy is increased. Conversely, if the postsynaptic spike is emitted soon, before the presynaptic one arrives, the synaptic efficacy is decreased. Several modeling studies have developed learning algorithms based on STDP and have demonstrated how systems that use these types of algorithms can carry out complex information processing tasks. The excitatory synaptic circuit of Indiveri and colleagues implements both long- and short-term plasticity.

Memory and Synaptic Weight Storage

There are a number of practical problems that slow down the development of large-scale, distributed, massively parallel networks of VLSI I&F neurons. Three of the most important ones are (1) how to access the individual synapses of the network to provide input signals, and how to read from each neuron to generate output signals; (2) how to set and/or store the weights of individual synapses in the network; and (3) how to (re)configure the network topology on the same chip. Of course, these problems arise directly from the fundamental difference in architecture between conventional computers and neural networks. Conventional computers have a single processor, or small number of processors, connected to a random access memory. This global access

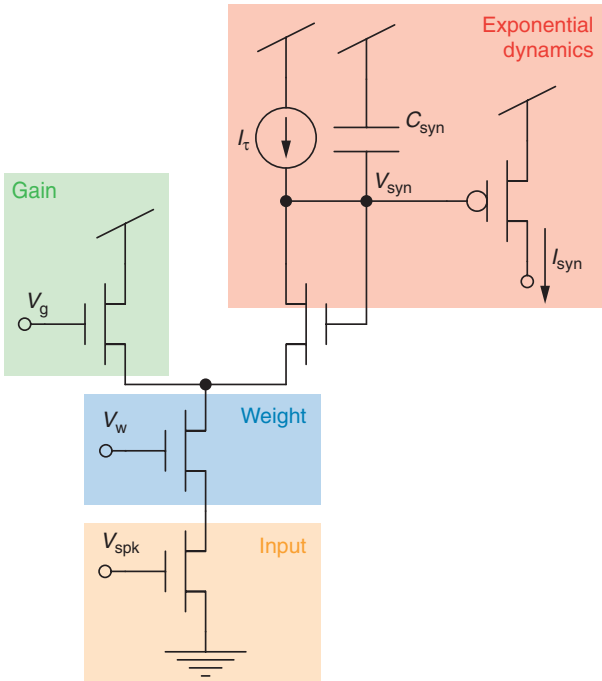

Figure 5 A compact synaptic circuit that exhibits the exponential dynamics observed in real synapses. The circuit's analog output current I_{syn} encodes the frequency of the input spikes, represented as digital voltage pulses arriving at the V_{spk} node. The circuit's time constant is set by adjusting the value of the I_τ current. The V_g bias in the gain module can be used to set the circuit's gain; lower values increase the gain, whereas higher ones decrease it. Similarly, the synaptic efficacy is controlled by the V_w bias voltage in the weight module. In this case, higher values of V_w lead to higher output currents. Additional circuits can be connected to the V_w node to implement short-and/or long-term plasticity mechanisms and to locally store or refresh the value of the weight.

means that the state of the machine can be conveniently loaded and examined. By contrast, in neuronal networks the memory and processing is massively distributed and is co-localized at the synapses. In this case, direct loading and inspection are not possible, except at the huge cost of providing duplicate access lines. Long-term plasticity circuits alleviate these problems slightly (for both biology and neuromorphic VLSI) by allowing the weights of synapses to be set automatically, without requiring dedicated access to individual synapses. But the difficulty of experimental observation and control remains for these circuits.

Basic considerations about the relation between precision and required silicon area suggest that biologically realistic low-resolution weights could also be stored as analog values as voltages across capacitors. But conventional CMOS capacitors are subject to leakage, so their voltages decay over seconds if not refreshed. Alternatively, the analog values can be stored using nonvolatile technology similar to that used in electrically erasable programmable memory (EEPROM). In this case, the analog value is stored as charge on a floating gate, that is the gate of an FET sandwiched between two layers of (perfect) oxide insulator. Charge is added or removed from the floating gate by Fowler–Nordheim tunneling and impactionized hot-electron injection. The uncontrolled decay of charge from floating gates is negligible, so learned synaptic weights can be retained for decades, even when the power to the circuits is off. This exciting technology is still under development, but simple structures such as single transistor synapses and examples of networks that implement learning algorithms based on these circuits have been demonstrated. The problem of synaptic weight storage would be simplified considerably if only one binary bit, rather than an analog value, need be stored. In fact, it has been demonstrated that networks of sufficiently large numbers of binary synapses are adequate for any memory task. Circuits for such binary synapses have been proposed that bypass the need to have specialized structures for nonvolatile analog memory within each synapse.

Multichip Neural Networks

Several examples of successful multichip networks of spiking neurons have been demonstrated during recent years. Using present CMOS technology, it is possible to implement on the order of hundreds of neurons and thousands synapses per square millimeter of silicon. In principle, networks of this type can be scaled up to any arbitrary size, but in practice the network size is limited by the maximum silicon area and AER bandwidth available. Given the current speed and specifications of the AER-interfacing circuits and the availability of present silicon VLSI technology, the network size could be increased by at least two orders of magnitude. It is likely that large neuronal networks, such as those of the neocortex, are dominated by local connectivity, with only a relatively small fraction of long-range connections. In this case, it may be possible to make similarly large-scale VLSI networks in which multiple regional AER buses with the same address spaces carry local-event traffic and are interconnected by sparser long-range traffic between local domains. However, before testing those connectivity limits, there is more immediate interesting work to be done with these multichip networks. For example, existing methods can be used for investigating complex spike-based learning algorithms in real time. And these studies are all the more

interesting when we consider problems of adaptation and learning in neuronal networks interfaced to neuromorphic AER sensors, such as silicon retinas and silicon cochleas.

Impact of Neurobiology on Computer Engineering

Conventional engineering computing systems are beginning to face challenges that have many points in common with those faced by neuromorphic and biological neural systems. As VLSI technology scales to the submicron feature size, power consumption per square micron starts to become a serious limiting factor and single transistors start to behave like unreliable stochastic devices. Under these conditions, insights obtained from building low-power neuromorphic circuits and biologically inspired methods for computing reliably with unreliable components can bring crucial knowledge to advanced computing technologies. Similarly, the current trend toward multicore digital processors in the IT industry will continue and will require the understanding of massively parallel computation, which is also a central question of neuroscience. The principles of unclocked, real-time asynchronous event-driven computation are still far from clear, but it seems likely that this form of communication and processing lies at the heart of biology's ability to achieve effective collective behavior from massively distributed, weakly connected, and relatively slow neuronal processors. We expect that the exploration of neuromorphic systems will continue to contribute strongly to solving this intriguing puzzle.

See also: Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Neuroplasticity: Computational Approaches; NMDA Receptors and Development; Retinal Models; Spike-Timing-Dependent Plasticity Models.

Further Reading

Bartolozzi C and Indiveri G (2007) Synaptic dynamics in analog VLSI. *Neural Computation* 19: 2581–2603.

Boahen K (1998) Communicating neuronal ensembles between neuromorphic chips. In: Lande TS (ed.) *Neuromorphic Systems Engineering*, pp. 229–259. Norwell, MA: Kluwer Academic.

Boahen K (2005) Neuromorphic microchips. *Scientific American* 292: 56–63.

Cauwenberghs G and Bayoumi MA (eds.) (1999) *Learning on Silicon: Adaptive VLSI Neural Systems*. Boston: Kluwer.

Chan V, Liu SC, and van Schaik A (2006) AER EAR: A matched silicon cochlea pair with address event representation interface. *Special Issue: Sensors. IEEE Transactions on Circuits and Systems I* 54(1): 48–59.

Delbrück T (1991) "Bump" circuits for computing similarity and dissimilarity of analog voltages. In: *Proceedings of the International Joint Conference on Neural Networks,* vol. 1, pp. 475–479. Seattle,WA, 8–12 July. IEEE.

Fusi S, Annunziato M, Badoni D, Salamon A, and Amit DJ (2000) Spike–driven synaptic plasticity: Theory, simulation, VLSI implementation. *Neural Computation* 12: 2227–2258.

Indiveri G, Chicca E, and Douglas R (2006) A VLSI array of low-power spiking neurons and bistable synapses with spike–timing dependent plasticity. *IEEE Transactions on Neural Networks* 17(1): 211–221.

Koch C and Segev I (2001) *Methods in Neural Modeling*, 2nd edn. Cambridge, MA: MIT Press.

Lichtsteiner P, Posch C, and Delbrück T (2006) A 128 × 128 120dB 30mW asynchronous vision sensor that responds to relative intensity change. In: *2006 IEEE ISSCC Digest of Technical Papers*, pp. 508–509. San Francisco, CA: IEEE.

Liu SC, Kramer J, Indiveri G, Delbrück T, and Douglas R (2002) *Analog VLSI: Circuits and Principles*. Cambridge, MA: MIT Press.

Mahowald M (1994) *An Analog VLSI System for Stereoscopic Vision*. Boston: Kluwer.

Mahowald M and Douglas R (1991) A silicon neuron. *Nature* 354: 515–518.

Mead C (1989) *Analog VLSI and Neural Systems*. Reading, MA: Addison Wesley.

Mead C (1990) Neuromorphic electronic systems. *Proceedings of the IEEE* 78(10): 1629–1636.

Serrano-Gotarredona R, Oster M, Lichtsteiner P, et al. (2005) AER building blocks for multilayer multichip neuromorphic vision systems. In: Weiss Y, Scholkopf B, and Platt J (eds.) *Advances in Neural Information Processing Systems,* vol. 18, pp. 1217–1224. Cambridge, MA: MIT Press.

Neuromuscular Connections: Vertebrate Patterns of

C R Slater, University of Newcastle upon Tyne, Newcastle upon Tyne, UK

© 2009 Elsevier Ltd. All rights reserved.

Neural Control of Muscles

Movement is a feature of all animals, and a central function of all nervous systems is to control movement in a way that promotes survival. In most animals, movement results from the neural activation of specialized contractile cells, the muscle fibers. This activation occurs at the neuromuscular junction (NMJ), the interface between nerve and muscle. While many features of muscle innervation and neuromuscular transmission are similar in all vertebrates, there are also many important differences. This article describes this diversity and how patterns of neuromuscular connectivity have changed during vertebrate evolution.

There are several distinct types of vertebrate muscles. In all of them, contraction arises from the interaction of the motor protein myosin with the cytoskeletal protein actin. However, myosin and actin are organized in different ways in different types of muscles. In striated muscles, they are present in extended arrays of overlapping filaments that give rise to periodic cross-banding, visible in the large cylindrical cells known as muscle fibers. Striated muscles that are attached to skeletal elements, and are used to move body parts (e.g., limbs, digits, eyes) relative to the external environment, are known as skeletal muscles. They are innervated by motor neurons at highly specialized NMJs.

The muscles of the internal organs are controlled by the autonomic nervous system. The cardiac muscle that causes the forceful pumping action of the heart is striated but has a very different cellular organization from skeletal muscle. In the smooth muscles of most internal organs, the actin and myosin arrays are less highly organized and no striations are present. Many cardiac and smooth muscles are intrinsically active, producing myogenic cycles of contraction and relaxation. The effect of the autonomic innervation is to alter the frequency and strength of these cycles. This involves both excitatory and inhibitory effects, mediated by different transmitters. These transmitters are released from beadlike swellings, or varicosities, along the autonomic motor axons. Although the molecular basis of release and transmitter action is similar to that in skeletal muscle, the elaborate NMJs that mediate rapid neuromuscular transmission in skeletal muscles are not present.

The rest of this article is limited to a consideration the innervation of skeletal muscles in vertebrates.

Diversity of Vertebrate Skeletal Muscles

The properties of individual skeletal muscles, and of the fibers within them, vary greatly both between and within vertebrate species. A particularly important distinction exists between muscle fibers that are electrically excitable, that is, that generate propagating action potentials when adequately depolarized, and those that are not. Muscle contraction is regulated by the concentration of free calcium in the cytoplasm of the muscle fiber and this, in turn, is controlled by the electric potential across the muscle fiber membrane. In lower vertebrates, as in most invertebrates, inexcitable muscle fibers are relatively common. They are innervated at multiple sites along the muscle fibers where nerve impulses give rise to local depolarizations (see later). The summed effect of these local depolarizations causes what may be very finely graded contractions of the muscle fiber. In general, the speed of contraction of these fibers is relatively slow (see later). This class of muscle fiber has been variously referred to as slow, tonic, and red, depending on the context. An example is found in the frog, where the properties of inexcitable fibers in vertebrates were first studied in detail, and where they were referred to as slow. Unfortunately, slow is also used to distinguish different subclasses of excitable fibers (see later). In this article, electrically inexcitable, multiply innervated muscle fibers are referred to as 'nonEx-MIF.'

Most muscle fibers, even in lower vertebrates, are electrically excitable. Each action potential in the muscle fiber gives rise to a brief contraction of relatively constant amplitude known as a twitch; fibers excited in this way are called twitch fibers. Within this broad class, there is considerable diversity of functional properties. This has led to definition of fast and slow subtypes of twitch muscle fibers on the basis of contraction speed. In addition, the situation is further complicated by classification of muscle fibers into red and white, based on the appearance of the live muscles. Red fibers tend to contract slowly and to be able to maintain their contraction for long periods of time. Associated with this ability are a predominantly oxidative energy metabolism, a rich vascular supply, and a high concentration of the oxygen-carrying heme protein myoglobin, which gives the fibers their red color. White fibers tend to contract more rapidly but also to fatigue more rapidly. They tend to utilize anaerobic metabolism, to have a less dense vascular supply, and

to have a much lower concentration of myoglobin, making them uncolored, or 'white.'

In addition to these two broad classes of muscle fibers, nonEx-MIF and twitch, there are numerous intermediate forms. In several animal groups, fibers exist that are multiply innervated but can also generate action potentials. While white fibers are generally twitch in character, red is used in the literature to describe both nonEx-MIF and slow twitch muscles. This makes for a confusing nomenclature, but it also highlights the important underlying natural diversity of muscle properties. Not surprisingly, this is reflected in a similar diversity in the patterns of muscle innervation.

General Features of the Motor Innervation of Skeletal Muscles

Skeletal muscles are innervated by specialized motor neurons. The cell bodies of most vertebrate motor neurons are located in the ventral horn of the spinal cord and in corresponding regions of the brain stem. Each spinal motor neuron has a single axon that leaves the central nervous system via the ventral spinal roots. It then extends unbranched in one of the peripheral nerves until it reaches its one target muscle. There, it branches extensively to make contact with many muscle fibers.

In the twitch muscles of higher vertebrates, it is generally the case that every action potential in a motor neuron causes contraction of every muscle fiber that neuron innervates. As a result, each motor neuron, together with the set of muscle fibers it innervates, operates as a functional 'motor unit.' The muscle fibers within a motor unit tend to have very similar functional properties, so that the motor unit as a whole is, for example, fast or slow. In lower vertebrates, as in most invertebrates, muscle fibers are often innervated by more than one motor neuron. They may thus be viewed as parts of more than one motor unit. It is likely that these muscle fibers are used in different ways, depending on which of their neuronal inputs is active.

The force of contraction of a muscle can be controlled by the nervous system in several ways. One is by regulating the frequency of activation of a given motor unit. Skeletal muscle fibers respond to a single nerve action potential with an all-or-none twitch that is brief, but nonetheless much longer than an action potential. If a second nerve action potential arrives before the first twitch has fully decayed, the second twitch will sum with it and the peak force achieved will be greater than that in a single, isolated twitch. Within a range that differs for different muscles, the peak force developed by a muscle increases with increasing frequency of neural activation. Many fast motor neurons fire in stereotyped high-frequency bursts, during which the force developed by the muscle increases rapidly to a maximum. In some situations the duration of these bursts, and with it the peak force, are varied as an additional means of regulating force output.

A second, very different, type of control can be achieved by varying the number of active motor units. In mammalian muscles, the force generated by activation of a small set of motor units is nearly linearly related to the sum of the forces of the motor units when they are stimulated independently. The fineness with which contraction can be graded by this form of control is determined by the number of motor units in a muscle.

Phylogenetic Variation in Patterns of Skeletal Muscle Innervation

During the course of vertebrate evolution, there have been major changes in the organization of muscles and the patterns of their innervation. The appearance of limbs and the increasingly sophisticated control of digits are reflected in the increasingly specific control over small groups of muscle fibers by the central nervous system. This is most developed in humans where pyramidal cells of the motor cortex make direct synaptic input to spinal motor neurons. In what follows, some of the main features of muscles and their innervation in the major groups of vertebrates are described. For each group, the species for which most information is available is indicated.

Agnatha (Hagfish, Lamprey)

In primitive vertebrates of the class Agnatha, the arrangement of muscles and their innervation differ from that in other vertebrates. Within the segmental myotomes, the muscle fibers are arranged into 'muscle units' in which large-diameter 'white' fibers are placed centrally. They are surrounded by a superficial layer of much smaller 'red' fibers.

Within the muscle units, the large central fibers are typical twitch fibers. Those nearest the abdominal cavity are innervated at their tendinous ends, sometimes by more than one axon. Stimulation of the nerve evokes an all-or-none action potential that propagates along the full length of the muscle fiber and evokes contraction. In lampreys, though probably not in hagfish, the remaining central fibers (those nearest the skin) are apparently not directly innervated but are electrically connected to innervated fibers by gap junctions. This ensures rapidly synchronized contraction

of the central fibers in each unit and an economy of innervation. However, it appears to be a relatively inflexible arrangement designed for speed, rather than subtlety, of action.

The most superficial fibers are not electrically excitable. Both anatomical and electrophysiological observations indicate that they are innervated at multiple sites along their length, often by the axons of more than one motor neuron. Thus they are similar to the nonEx-MIF fibers of other vertebrates.

Cartilaginous Fish (Dogfish)

In the myotomes of both cartilaginous and bony fish, two distinct types of muscle fibers, often referred to as white (or fast) and red (or slow), are present. In both types of fish, the white fibers form the central bulk of the muscles, while the red fibers form a superficial layer (Figure 1, dogfish). As in Agnatha, the white fibers in dogfish appear to be innervated mainly at their ends, while the red fibers are innervated at multiple sites along their length. However, so-called aberrant fibers are also present in which innervation occurs at a single centrally placed site, as in many higher vertebrates.

Bony Fish (Tench, Zebra Fish)

In many teleosts (bony fish), as in sharks, fast and slow muscle systems exist, with the slow fibers present in a superficial layer. The slow fibers are innervated at multiple sites by branches of several fine motor axons. There is considerable variation in the patterns of innervation of the fast white fibers in different species. In the catfish and eel, the pattern is similar to that in sharks, with nerve–muscle contacts made at the ends of the fast fibers. In most other teleosts that have been examined, the pattern of innervation of the white fibers is similar to that of the red fibers, with several sites of nerve–muscle contact, originating from more than one motor neuron, along the fibers (Figure 1, tench). Nonetheless, the white muscle fibers generate action potentials and twitch contractions, so are not typical nonEx-MIF fibers.

In the zebra fish, two distinct populations of motor neurons innervating the fast white muscles have been identified in the spinal cord. Primary motor neurons have large cell bodies and axons and are the first to appear during development. There are three primary motor neurons innervating each segmental myotome on each side of the fish. The primary motor neurons innervate a nonoverlapping set of muscle fibers within the myotome. Their axonal arbors do not cross the boundaries between segments. These motor neurons can thus be recognized unambiguously from one individual to the next.

In addition to the primary motor neurons, there is an additional class of secondary motor neurons. These are smaller than the primaries and have peripheral fields of innervation that are smaller and less focused than those of the primary motor neurons, often innervating muscle fibers in more than one segment. Individual white muscle fibers receive input from one primary and several secondary motor neurons. These synaptic inputs differ in the strength of their effect on the muscle fibers. Stimulation of the primary neurons evokes a large, short-latency endplate potential, suggesting that they may be used to in escape behavior. In contrast, secondary neurons evoke smaller, longer latency endplate potentials. Because white muscles of fish are both multiply innervated and generate action potentials, they form a group with properties intermediate between nonEx-MIF and twitch fibers as described previously.

Amphibia (Salamanders, Frogs)

The evolution of limbs made important new demands on the organization of skeletal muscles and their innervation. In what was to prove to be the beginning of a sustained evolutionary trend, the slow fiber system that plays an important part in normal fish behavior appears to be reduced in importance in amphibia, and greater reliance is placed on focused innervation of individual fast muscle fibers.

Urodeles In salamanders and newts, both twitch and nonEx-MIF muscle fibers are present. The twitch muscle fibers are generally innervated by several motor axons (typically one to three), as in fish. The contractile properties of individual motor units vary greatly. Interestingly, individual muscle fibers may belong to motor units with quite different contractile properties.

Anura In frogs and toads, most muscle fibers are twitch fibers, and are innervated by a single motor neuron (Figure 1, frog – fast). These twitch fibers vary in their functional and molecular properties, and two or three different types have been defined. NonEx-MIF fibers are also present in many anuran muscles (Figure 1, frog – slow). However, in contrast to fish, they are generally not present as contiguous groups. In addition, some muscle fibers of an intermediate type have been found in frogs (submaxillary muscles) that are electrically excitable, but are innervated by several motor axons.

Reptiles (Garter Snake, Lizard, Turtle)

Reptiles, like fish and amphibia, have both nonEx-MIF and twitch muscle fibers. In most reptiles, however, the

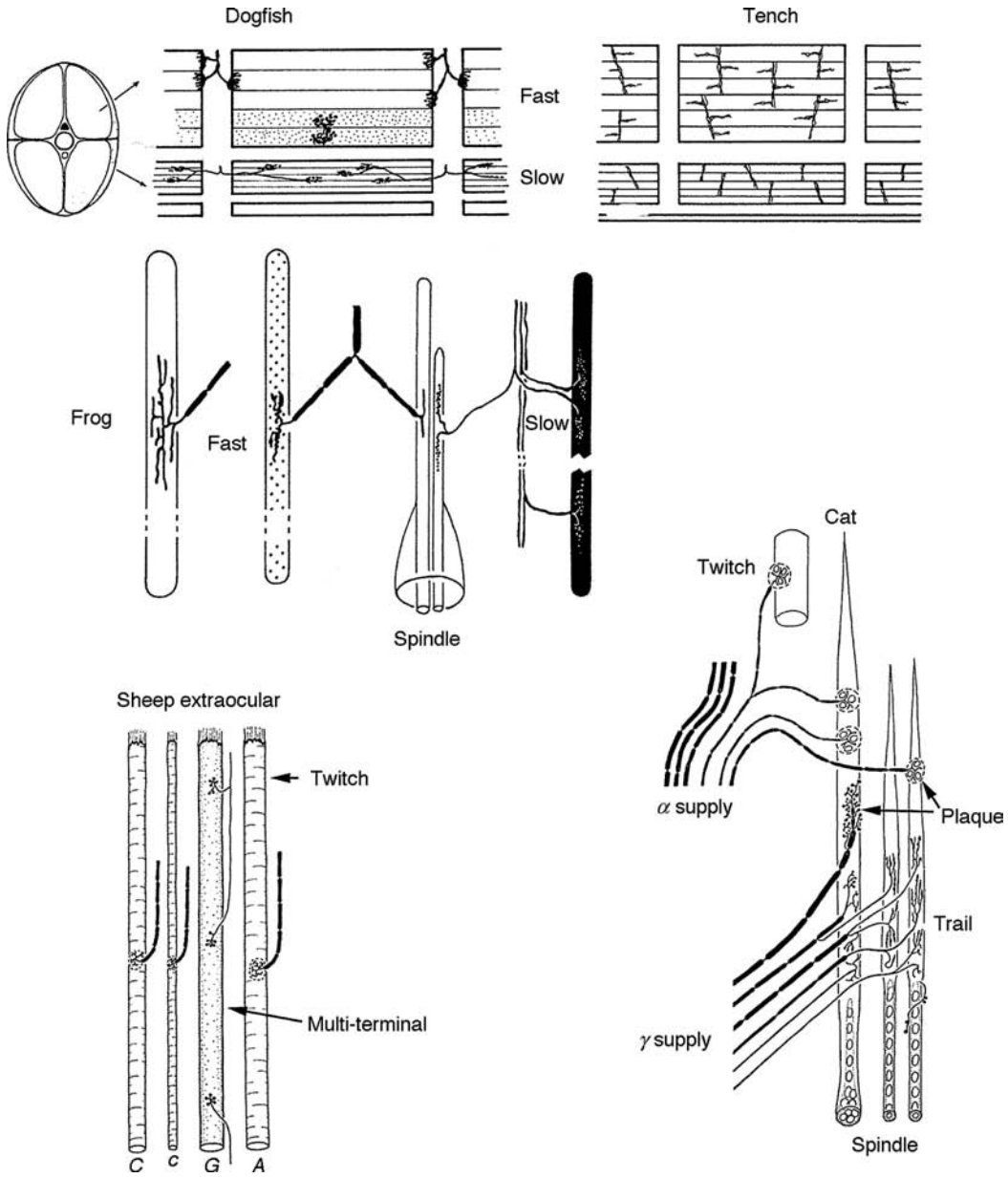

Figure 1 Some common patterns of innervation of vertebrate muscles. In dogfish, multiply innervated slow muscle fibers form a superficial layer surrounding the central fast muscle fibers. The latter are mainly innervated by basketlike endings at the ends of the fibers. Some intermediate forms, with centrally located plaquelike endings, are also present. In the tench, as in many bony fish, both fast and slow muscle fibers are multiply innervated. Reproduced from Bone Q (1964) Patterns of muscular innervation in the lower chordates. *International Review of Neurobiology* 33: 99–147, with permission from Elsevier. In frogs, fast muscle fibers are generally innervated at a single site by nerve endings whose several branches extend along the muscle fiber. In addition, electrically inexcitable, multiply innervated muscle (nonEx-MIF) fibers are present which are innervated at multiple sites by more diffuse endings. The intrafusal muscle fibers present in the sensory muscle spindles are innervated by branches of the motor neurons that innervate both twitch and nonEx-MIF extrafusal fibers. From Barker D (1968) L'innervation motrice du muscle strié de vertebrés. *Actualités Neurophysiologiques* 8: 23–71. In mammals, most twitch muscle fibers are innervated at a single site, but in some, multiply innervated muscle fibers are present in the extraocular muscles that control the position of the eye (sheep extraocular). Reproduced from Harker DW (1972) The structure and innervation of sheep superior rectus and levator palpebrae extraocular muscle. *Investigative Opthamology and Visual Science* 11: 956–969, with permission. The contractile intrafusal fibers of the mammalian muscle spindle (cat) receive multiple motor innervation derived from both α- and γ-motor neurons. The latter terminate in both compact plaque endings and more diffuse trail endings. Reproduced from Barker D (1970) Fusimotor innervation in the cat. *Philosophical Transactions of the Royal Society B* 8: 336, with permission from Philosophical Transactions of the Royal Society B.

twitch fibers are generally innervated by a single motor neuron. Different functional types (fast and slow) of twitch fibers have been identified. In the sheetlike transversus abdominis muscle of the garter snake, fast and slow twitch fibers alternate with nonEx-MIF fibers in a highly regular pattern. Available evidence suggests that this pattern is not determined by the nerve and is therefore assumed to be determined by inherent differences in the muscle fibers. In turtles, nonEx-MIF fibers are present and of particular importance in the neck muscles that support the head.

Birds (Chicken)

Chickens also have both multiply innervated nonEx-MIF and singly innervated muscle fibers. In contrast to nonEx-MIF fibers in frogs, those in chickens are capable of generating regenerative action potentials, at least in laboratory conditions. Whether this occurs during normal use is unclear. One often studied chicken muscle, the anterior latissimus dorsi, is composed predominantly of slow, multiply innervated nonEx-MIF muscle fibers.

Mammals (Human, Rat, Mouse, Cat)

In mammals, virtually all muscle fibers are twitch fibers (but see later). Each fiber receives innervation at a single site from a single motor neuron. The number of muscle fibers in a mammalian motor unit varies from less than 10 to 1000 or more. The number of motor units in an individual muscle ranges from about 25 to several hundred, and sets limits on how finely contractions can be graded.

There is considerable variability in the properties of mammalian twitch fibers. In particular, they vary in their speed of contraction and in their ability to maintain contraction without fatiguing. Mammalian twitch fibers are generally recognized as being slow twitch (or type I) or fast twitch (or type II). However, the fast-twitch fibers are further subdivided into subtypes (IIA, IIB, IIX) based on their contractile properties and the main subtype of myosin expressed in them. The muscle fibers in a single motor unit are normally all of the same subtype.

Like muscle fibers, mammalian motor neurons vary in their properties. For example, within the set of motor neurons innervating an individual muscle, some motor neurons have larger cell bodies than others. Axon diameter, and hence the speed of action potential conduction, are closely correlated with cell body size. In this sense, large motor neurons are also fast. Fast motor axons branch more extensively within the muscle and innervate more muscle fibers than do smaller, slower, ones. As a result, fast motor units tend to generate more force than do slow ones. Finally, the muscle fibers innervated by fast motor neurons are themselves functionally fast. Thus the functional properties of the nerve and muscle fibers are closely matched in each motor unit.

In contrast to lower vertebrates, there are very few nonEx-MIF muscle fibers in mammals. The extraocular muscles are a notable exception (**Figure 1**, sheep extraocular). Some of these have a substantial population of nonEx-MIF fibers that are capable of generating finely graded contractions in response to varying frequencies of neural input. It has been suggested that the pattern of innervation of these fibers may be a relic of a primitive anterior 'vegetative' region of the vertebrate body.

NonEx-MIF fibers of a second group are contained within the sensory muscle spindles. These so-called intrafusal fibers (within the spindle, as distinct from the majority of extrafusal fibers) are primarily innervated by a distinct class of γ-motor neurons (**Figure 1**, cat-spindle). Some of these neurons terminate in plaque endings, similar to those of the α-motor neurons on extrafusal twitch fibers, and others end in more diffuse trail endings, reminiscent of endings on the nonEx-MIF fibers of lower vertebrates. Indeed, in frogs, the motor axons ending in trail endings sometimes also innervate extrafusal nonEx-MIF fibers (**Figure 1**, frog – spindle). Some intrafusal fibers are of an intermediate type in that they are capable of generating action potentials.

Variations in NMJ Properties

Vertebrate NMJs are as diverse as the motor neurons and muscle fibers of which they are composed. At most vertebrate NMJs, the sites of quantal release are small domains 2–5 μm in diameter, known as boutons. Separate boutons are generally connected by fine unmyelinated axon branches, as in fish, snakes, and humans. However, in some species, they may be variably fused into more linear arrays, as in the frog and turtle (**Figure 2**). In the great majority of mammalian muscles, each muscle fiber is innervated at a single point by the dense plaquelike ('en plaque') terminal of a single motor neuron. By contrast in lower vertebrates, and in fish in particular, multiterminal polyaxonal innervation is more common, and individual terminals often consist of a dispersed set of boutons, often described as like a bunch of grapes ('en grappe').

Associated with the differing patterns of innervation in different muscle types and in different animal groups is a range of the efficacy of neuromuscular transmission. In mammals, at one extreme, each motor nerve impulse leads to a depolarization of the muscle fiber that is several times greater than required

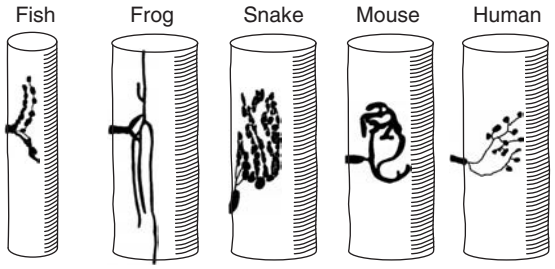

Figure 2 Variations in neuromuscular junction structure in different vertebrates. The top row of drawings shows the typical appearance of individual NMJs in twitch muscle fibers of the species indicated. Many of these consist of distinct boutons which are the sites of transmitter release. In some species these are variably merged so that the nerve terminal becomes a continuous band. In fish, each NMJ has two types of endings, one more continuous than the other. Frogs are one of few species in which the NMJ is greatly elongated in the long axis of the muscle fiber. In snakes and humans, the individual boutons are distinct but much more numerous in the snake.

to trigger an action potential. Two main factors contribute to this high safety factor. On the one hand, the nerve routinely releases more transmitter than is needed to reach threshold. On the other, the threshold for action potential generation in the muscle is lowered in the region of the NMJ. Each of these factors is associated with distinctive structural features of the NMJ.

Many studies have shown that the amount of transmitter released from a motor nerve terminal is roughly proportional to its area of contact with the muscle. The transmitter at the NMJ, acetylcholine (ACh), is released in multimolecular packets known as 'quanta.' Each quantum corresponds to the ACh contained in a single presynaptic vesicle. The number of quanta released by a single nerve impulse is known as the 'quantal content' of release. During low-frequency stimulation (1 Hz) *in vitro*, many NMJs release 0.15–0.25 quanta μm^{-2}. In contrast, there are substantial variations in quantal content at NMJs in different species (e.g., frog, 200; human, 20). In many cases, the differences in quantal content are associated with equivalent differences in NMJ size (Figure 3(a)).

The effect of the ACh released from the nerve on the membrane potential of the muscle fiber is also influenced by structural factors. For example, the smaller the fiber, the higher its electrical 'input' resistance and the greater the depolarization caused by the current flowing through the ion channels opened by a quantum of ACh. In order for the depolarization caused by ACh to trigger an action potential, it must open an adequate number of the voltage-gated sodium channels that underlie the action potential. These channels are concentrated in the postsynaptic membrane, so a smaller fraction of them needs to open to trigger an action potential than if their density was lower. In many species, particularly of higher vertebrates, the postsynaptic membrane is highly folded (Figure 3(b)). This enhances the effect of the increased concentration of sodium channels in the postsynaptic membrane and has led to the suggestion that an important function of the folds is to act as amplifiers of the effect of the ACh released from the nerve.

The NMJs in different animal groups use different balances between pre- and postsynaptic effects to achieve the necessary reliability of neuromuscular transmission (Figure 3(c)). In humans the terminals are small and release relatively few quanta. However, human NMJs have very extensive postsynaptic folds which amplify the effect of the modest amount of ACh released. It is likely that this plays an important part in ensuring an adequate safety factor of transmission. In frogs, the terminals are large and release many quanta but the folding is less extensive than in humans. Rats and mice have intermediate degrees of folding. In fish and chickens, postsynaptic folds are generally lacking and it is possible that the safety factor is relatively low in these species. Finally, in electrically inexcitable fibers, postsynaptic folds are absent. Thus there is a clear correlation between the extent of postsynaptic folding and the tendency to generate an action potential in the muscle cell.

Patterns of Motor Unit Use in Vertebrate Muscles

The diversity of functional properties of vertebrate muscle fibers only makes physiological sense because fibers with different properties can be specifically activated by motor neurons which themselves have distinctive properties and patterns of activity. However, the selective innervation of individual muscle fibers with specific properties is a much more prominent feature of higher vertebrates such as mammals than of lower vertebrates such as fish.

In fish, the fast white muscle fibers are innervated by both primary and secondary motor neurons (see earlier). The secondary motor neurons activate the relatively low-force, and possibly more finely regulated, contractions needed to power smooth movement. The primary motor neurons trigger the rapid movements that give rise to escape behavior. Since each primary motor neuron innervates a large block of muscle fibers in each segment (see earlier), it is clear that all these fibers are likely to fire simultaneously during rapid escape movements. As most white fibers are innervated by both primary and secondary motor neurons, it is likely that they are used in the generation of both slow and fast movements.

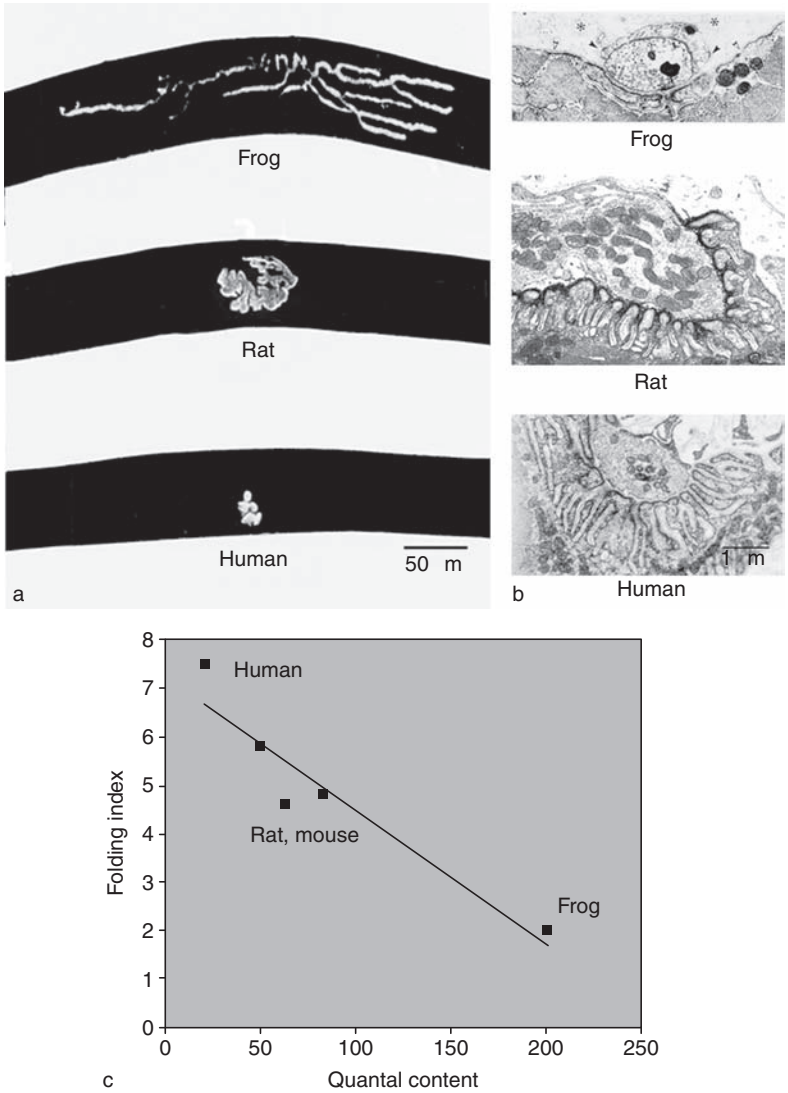

Figure 3 Neuromuscular junctions vary in size and extent of postsynaptic folding. (a) Light microscope views of NMJs in frog, rat, and human, showing region of NMJ in white, muscle fiber in black. (b) Electron microscope views of NMJs in the same species, showing much more extensive folding in humans. (c) Inverse relationship between extent of folding ('folding index') and amount of transmitter released ('quantal content').

In mammals, individual muscle fibers are innervated by a single motor neuron, and are only active when that neuron fires. Studies of rat leg muscles show that motor neurons innervating type I muscle fibers (slow twitch) fire at about 10–20 Hz, in prolonged trains, which are associated with sustained postural contractions of anti-gravity muscles. In contrast, motor neurons innervating type II fibers (fast twitch) fire in brief stereotyped bursts of 5–10 action potentials at a frequency of about 50–100 Hz. In the type I fibers, as found in the rat soleus muscle, the natural frequency of firing is the lowest that maintains maximal force output. By contrast type II motor units, which are often used when a rapid acceleration of a body part is required, fire in short bursts of closely spaced impulses that are optimal for the rapid development of force.

The distinctive patterns of activity in different mammalian motor neurons are largely determined by the intrinsic properties of their membranes rather than by the temporal patterns of inputs to them. Following an action potential in a motor neuron, there is generally a period of increased negativity during which generation of another action potential is suppressed. The duration of this afterhyperpolarization (AHP) is greater in motor neurons innervating type I muscle fibers than in those innervating type II muscle fibers. There is a close correlation between the time course of the AHP and of the twitch of the muscle fibers in the motor unit. This illustrates further the effect of matching of the properties of motor neurons to those of the muscle fibers they innervate.

A second aspect of motor unit use is revealed during graded voluntary contractions in mammals when there is an orderly increase in the number of active motor units. For relatively weak contractions, such as are used to maintain posture, the motor units activated are small and slow. As the force output increases, as during a rapid movement of a limb, increasingly large and fast motor units are recruited. An important effect of this orderly recruitment is that the greater the preexisting force being generated, the greater is the extra force added by each newly recruited motor unit. Thus the fractional increase in force per recruited unit stays roughly constant. This tendency for an orderly recruitment of motor units of increasing size and speed was first emphasized by Henneman and his colleagues and is known as the 'size principle.'

A simple biophysical feature of motor neurons appears to underlie the size principle. Thus, consistent with Ohm's law, small motor neurons have a relatively high 'input resistance.' As a result, a synaptic current of a given intensity causes a larger depolarization and is more likely to trigger an action potential than it would in a larger, lower resistance, neuron. As the synaptic input, or 'drive,' to a pool of motor neurons innervating a muscle increases, the first neurons to be recruited are therefore likely to be those with the highest input resistance and, hence, the smallest size and slowest conducting axons. This relationship assumes that the intrinsic membrane properties of motor neurons of different sizes are similar. Important exceptions to the size principle are known to exist. Thus, during constant activation of the motor neuron pool, some units may abruptly cease firing while others of similar strength begin to fire. This allows individual units to have periods of rest without loss of overall force.

There is relatively little information about the natural activity patterns of motor units in freely moving lower vertebrates. It is therefore difficult to interpret the functional consequences of the varied arrangements of muscle fiber types and arrangements seen in those animals. It seems very likely that motor units composed of nonEx-MIF muscle fibers, whether they are distributed throughout a mixed muscle or concentrated into a single muscle, are used to generate slow, maintained contractions that can be very finely graded in strength. A good example may be the extraocular muscles that must position the eye with great accuracy. The 'stepwise' increase in tension associated with the recruitment of additional motor units may simply be too coarse for this purpose.

Conclusions

It is clear that vertebrate muscles and their innervation are specialized to carry out distinctive sets of tasks. As the vertebrate body plan has evolved, muscles and their innervation have become adapted to new functional requirements. Several general trends can be identified in this process. One of these is the increasing specificity of control of individual muscle fibers by the nervous system. In Agnatha, many muscle fibers are not even directly innervated, and depend on electrical coupling to innervated fibers for their control. In this situation the nerves exert their control on blocks of muscle fibers. This results in speed of response but limited flexibility of how muscles are used. In fish and some amphibia, while all muscle fibers are innervated, many are innervated by more than one motor neuron. This means that individual muscle fibers are part of several motor units, and are presumably used in somewhat different ways depending on the circumstances. Elucidating the way in which the several inputs to each muscle fiber interact to control overall behavior in these animals remains an important task for the future. By contrast, in higher vertebrates, including mammals, virtually all muscle fibers are innervated at a single point by a single motor neuron whose properties are closely matched to those of the muscle fibers it innervates. This gives the nervous system unambiguous access to highly specialized sets of muscle fibers, allowing appropriate ones to be activated for particular tasks. As with many phylogenetic trends, the loss of polyneuronal innervation of muscles is repeated in the ontogeny of higher vertebrates.

A second closely related trend is the loss of importance of the system of multiply innervated muscle fibers. These are present in most muscles of most vertebrates up to birds, but are largely absent in mammals. Thus the pattern of innervation that is predominant in higher vertebrates may be seen to derive from the fast system associated with escape behavior in lower vertebrates. It may be that the fine of control of force associated with the graded activation of nonEx-MIF motor units in lower vertebrates is achieved in other ways in mammals, possibly involving more subtle regulation of antagonistic muscles under the control of the sensory muscle spindles.

A third trend is the gradual increase in the efficacy of neuromuscular transmission. In Agnatha and fish, most individual motor nerve terminals are small and release relatively few quanta of ACh. Thus activation of the muscle requires either repetitive firing of the nerve or activation of more than one axon or both. In most vertebrates higher than fish, every nerve impulse causes a sufficient depolarization of the muscle fiber membrane to trigger an action potential. The high safety factor of neuromuscular transmission in these species is either associated with large neuromuscular junctions, a highly folded postsynaptic membrane rich in voltage-gated sodium channels, or both.

See also: Motor Control of Feeding and Drinking; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction: Synapse Elimination; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Non-Quantal Release of Transmitter; Presynaptic Events in Neuromuscular Transmission.

Further Reading

Banks RW (1994) The motor innervation of mammalian muscle spindles. *Progress in Neurobiology* 43: 323–362.

Barker D (1968) L'innervation motrice du muscle strié de vertebres. *Actualités Neurphysiologiques* 8: 23–7.

Barker D (1970) Fusimotor innervation in the cat. *Philosophical Transactions of the Royal Society B* 8: 336.

Bone Q (1964) Patterns of muscular innervation in the lower chordates. *International Review of Neurobiology* 33: 99–147.

Bone Q (1972) The dogfish neuromuscular junction: Dual innervation of vertebrate striated muscle fibres? *Journal of Cell Science* 10: 657–665.

Harker DW (1972) The structure and innervation of sheep superior rectus and levator palpebrae extraocular muscle. *Investigative Opthamology and Visual Science* 11: 956–969.

Hennig R and Lomo T (1985) Firing patterns of motor units in normal rats. *Nature* 314: 164–166.

Lichtman JW and Wilkinson RS (1987) Properties of motor units in the transversus abdominis muscle of the garter snake. *Journal of Physiology* 393: 355–374.

Mark RF, von Campenhausen G, and Lischinsky DJ (1966) Nerve–muscle relations in the salamander: Possible relevance to nerve regeneration and muscle specificity. *Experimental Neurology* 16: 438–449.

Morgan DL and Proske U (1984) Vertebrate slow muscle: Its structure, pattern of innervation, and mechanical properties. *Physiological Reviews* 64: 103–169.

Ogata T (1988) Structure of motor endplates in the different fiber types of vertebrate skeletal muscles. *Archives of Histology & Cytology* 51: 385–424.

Rome LC (2002) The design of vertebrate muscular systems: Comparative and integrative approaches. *Clinical Orthopaedics and Related Research* 403(supplement): S59–S76.

Slater CR (2003) Structural determinants of the reliability of synaptic transmission at the vertebrate neuromuscular junction. *Journal of Neurocytology* 32: 505–522.

Walrond JP and Reese TS (1985) Structure of axon terminals and active zones at synapses on lizard twitch and tonic muscle fibers. *Journal of Neuroscience* 5: 1118–1131.

Westerfield M, McMurray JV, and Eisen JS (1986) Identified motoneurons and their innervation of axial muscles in the zebrafish. *Journal of Neuroscience* 6: 2267–2277.

Wilkinson RS and Teng H (2003) The nerve-muscle synapse of the garter snake. *Journal of Neurocytology* 32: 523–538.

Wood SJ and Slater CR (2001) Safety factor at the neuromuscular junction. *Progress in Neurobiology* 64: 393–429.

Neuromuscular Junction (NMJ): A Target for Natural and Environmental Toxins in Humans

J Harris, Newcastle University, Newcastle upon Tyne, UK

© 2009 Elsevier Ltd. All rights reserved.

Natural Toxins and the Neuromuscular Junction

The neuromuscular junction (NMJ) is the ultimate relay between the motor cortex, which controls the excitation of skeletal muscle, and the mechanical response of the muscle, which results in physical movement. Anything that perturbs the functioning of the neuromuscular system clearly threatens the survival of the organism. It should be no surprise, therefore, that numerous animals – plants, fungi, bacteria, and protists (Figure 1) – have evolved highly specific toxins that target specific components of the NMJ. The toxins may be used defensively to ward off potential predators or offensively to restrain prey. In some cases, it is not clear why the toxins are of any benefit to the organism; for example, some soft-bodied anthozoans produce the highly toxic polyether palytoxin (*qv*), one of the most dangerous of all toxins to mammals, and it is assumed that the poison is used by the anthozoans to protect their growing space from competitors. All components of the NMJ may be targets of natural toxins. These include voltage-gated ion channels of all kinds located on the axon, the nerve terminal, and the postjunctional membrane of the muscle fiber; membrane-bound ATPases; synaptic vesicle-docking proteins (e.g., synaptobrevin, SNAP-25, and syntaxin); junctional acetylcholine esterase; and the postjunctional acetylcholine receptor.

Natural toxins are structurally highly diverse. Some are simple amino acids, some are small polypeptides of between 15 and 150 amino acids, and some are very large and complex polypeptides such as the clostridial toxins. Other common toxins such as palytoxin are large polyethers and toxical kaloids are produced by numerous plants and algae. The diversity of structures is illustrated in Figure 2.

It is difficult to identify a single coherent system for the classification of natural toxins. Classification according to chemical structure has an obvious logic, but there is little relationship between chemical structure and pharmacological activity. For example, among the toxins that block the acetylcholine (ACh) receptors at the vertebrate NMJ are alkaloids (e.g., tubocurarine); polypeptides (e.g., α-bungarotoxin); and toxins that block the ion pore of the voltage-gated sodium channel, including cyclical guanidiums (e.g., tetrodoxin and saxitoxin) and small polypeptides (e.g., the μ-conotoxins). Classification according to the source (i.e., phylum or class) is suitable in some circumstances, but a discussion, for example, of the venom toxins of the squamata would require considerable subclassification according to the function of the toxins. Because this article concerns the NMJ, natural toxins have been considered in terms of their molecular or subcellular target.

Whichever method of classification is used, two important complicating factors need to be borne in mind. The first is that some toxins may modify function by a direct action on a recognized molecular target and others act by comprising the structural integrity of the target. For example, many of the presynaptically active phospholipases of snake venoms cause enhanced exocytosis, the depletion of synaptic vesicles, and the eventual degeneration of the nerve terminal, but these changes are initiated by the hydrolysis of the nerve terminal phospholipids and nerve-terminal depolarization. The α-latrotoxins bind to specific nerve-terminal proteins neurexin and synaptophilin to cause enhanced exocytosis and the eventual depletion of the synaptic vesicles without tissue degeneration. The second factor is that many animals, plants, and microorganisms elaborate or accumulate multiple isoforms or derivatives of major toxins and often use a battery of toxins of different types as part of their armory. For example, snakes of the genus *Dendroaspis* (the mambas) elaborate *inter alia* in their venom glands two types of toxins that block the nicotinic ACh receptor: toxins that inhibit acetylcholinesterase activity and bind to muscarinic ACh receptors, and toxins that block voltage-gated potassium channels. Similarly, the fish-eating cone snails (e.g., *Conus magus* and *Conus geographus*) produce toxins that block ACh receptors and voltage-gated Na^+ and Ca^{2+} channels. Filter-feeding shellfish and detritus-grazing Xanthid crabs accumulate a range of neurotoxic toxins, including tetrodotoxin, saxitoxin, and palytoxin, all of which are produced by marine microorganisms.

Voltage-Gated Ion Channels

Voltage-gated sodium, calcium, and potassium channels are the molecular targets for a wide range of very potent natural toxins. The toxins interact with voltage-gated ion channels in three ways: occluding the ion pore, thus disrupting function by physically preventing the

Figure 1 Organisms that produce major toxins that are active on the nervous system: (a) *Gonyaulax,* a dinoflagellate that produces saxitotoxin and a range of related toxins that block voltage-gated Na^+ channels; (b) *Palythoa caesia*, a zoanthid that is a source of palytoxin, a complex toxin that converts Na^+/K^+ adenosine triphosphate (ATP)ase to nonselective cation channels in numerous cells; (c) a tarantula and a (much smaller) black widow spider, which produce toxins resulting in the accelerated release of neuromuscular transmitter; (d) an Australian tiger snake, whose venom contains toxins that block acetylcholine receptors, damage skeletal muscle, and destroy nerve terminals.

movement of ions across the cell membrane; or modifying the gating process by binding to extracellular sites and affecting the voltage sensitivity of the channel; or modifying the kinetics of the opening and/or closing the ion channels. Lipid soluble toxins may enter the channels and bind to stabilize the channel in one particular state (open, closed, or inactivated). The extensive molecular diversity within the superfamily of voltage-gated ion channels, their respective distribution between different excitable tissues (e.g., muscle and nerve), and the highly selective predatory behavior of venomous animals suggest that most toxins targeting voltage-gated ion channels are selective for both ion channel subtype and animal species. Although detailed information on these issues is not routinely collected, there are good examples of such selectivity. For example, the voltage-gated sodium ion channels can be differentiated on the basis of their speed of inactivation and their sensitivity to tetrodotoxin and saxitoxin. The sodium channel subtypes $Na_v1.1$, 1.2, 1.3, 1.4, 1.6, and 1.7 are widely distributed in tissues of the central nervous system (CNS), peripheral nervous system (PNS), and skeletal muscle; and regardless of location, all are blocked by saxitoxin and tetrodotoxin at concentrations between 5 and 15 nM. Subtypes 1.5, 1.8, and 1.9 are also widely distributed, but are blocked at concentrations between 2 and 60 μM. This distinction led to the observation that in skeletal muscle the immature form of the voltage-gated ion channel is the tetrodotoxin-resistant $Na_v1.5$ and that its expression is suppressed by imposed or neurally regulated activity, at which time the adult $Na_v1.1$ subtype is upregulated and sensitivity to tetrodotoxin appears. The $Na_v1.5$ subtype reappears after denervation and during the early stages of regeneration of skeletal muscle and is suppressed during reinnervation (see Figure 3). Species sensitivity is particularly marked in the toxins of spider and scorpion venoms, and more than 30 different toxins selectively targeting insect voltage-gated sodium ion channels have been identified in scorpion venoms alone.

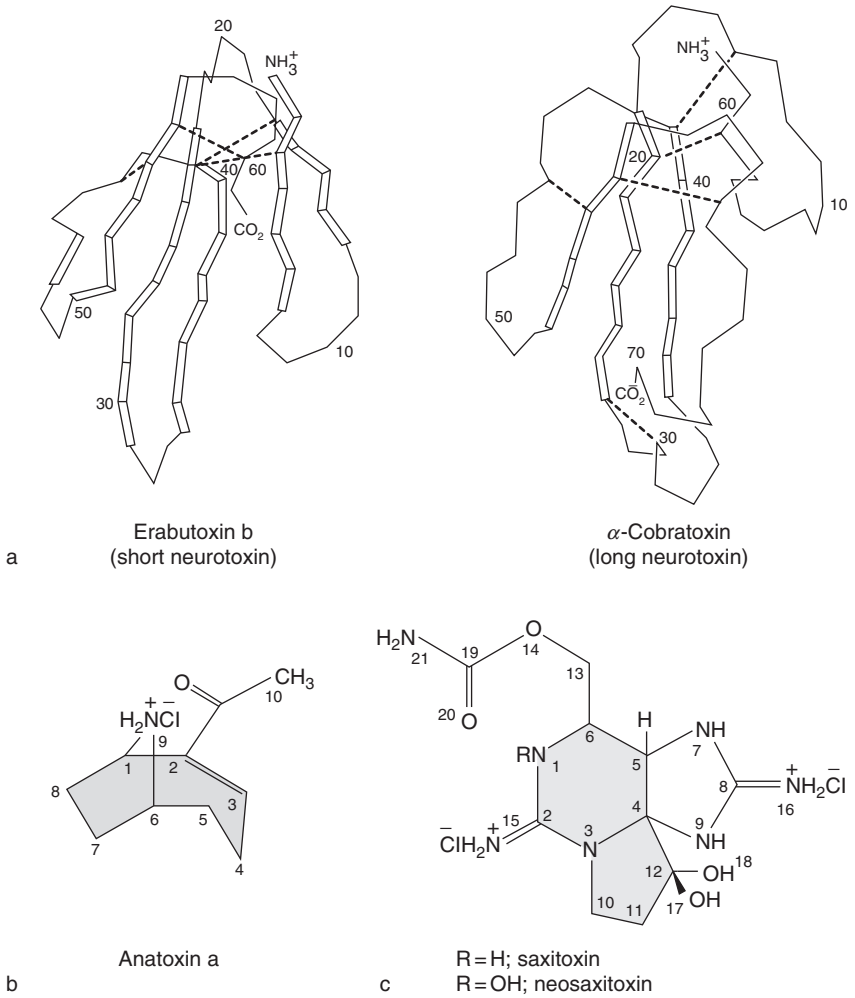

Figure 2 Diverse structures of natural toxins: (a) long and short α-neurotoxins from snake venoms, which block acetylcholine receptors at the neuromuscular junction and are small, highly folded polypeptides; (b) anatoxin, which is a natural anticholinesterase produced by blue-green algae; (c) saxitoxin, which is produced by marine dinoflagellates and blocks voltage-gated Na^+ channels. In (a), β-pleated sheet formations are indicated by plates and disulfide bridges are indicated by dotted lines.

Sodium Channels

The voltage-gated sodium channels comprise a single protein complex containing the α subunit, which contains both the pore-forming and voltage-sensing elements, and one or two subsidiary β subunits. The α subunit has four transmembrane regions, each of six hydrophobic segments. A hydrophobic loop between the fifth and sixth segment forms the neck of the ion channel (Figure 4). Activation of the sodium channel results in the reorientation of the fourth segments in each of the four transmembrane regions and the opening of the channel to allow the passage of sodium ions down their electrochemical gradient into the cell. Toxins binding with high degrees of specificity to voltage-gated sodium channels have been isolated from the venoms of snakes, scorpions, cone snails, and sea anemones; from the skin secretions of tropical frogs; from marine bacteria and dinoflagellates; from blue-green algae; and from plants. Six distinct binding sites have been identified in the voltage-gated sodium channel.

Toxins active at site 1 occlude the sodium channel to produce a complete block of neurotransmission and a potentially fatal neuromuscular paralysis (Figure 4). The toxins most commonly involved are tetrodotoxin, a heterocyclic guanidinium produced by symbiotic marine bacteria and isolated originally from puffer and sun fish; saxitoxin, a structurally related toxin produced by marine dinoflagellates of the genus *Gonyaulax* and usually isolated from filter-feeding shellfish; and the μ-conotoxins, small polypeptides of approximately 22-amino-acid residues produced by a number of marine snails of the genus *Conus*. Site 2 on segment 6 is the binding site for a number of

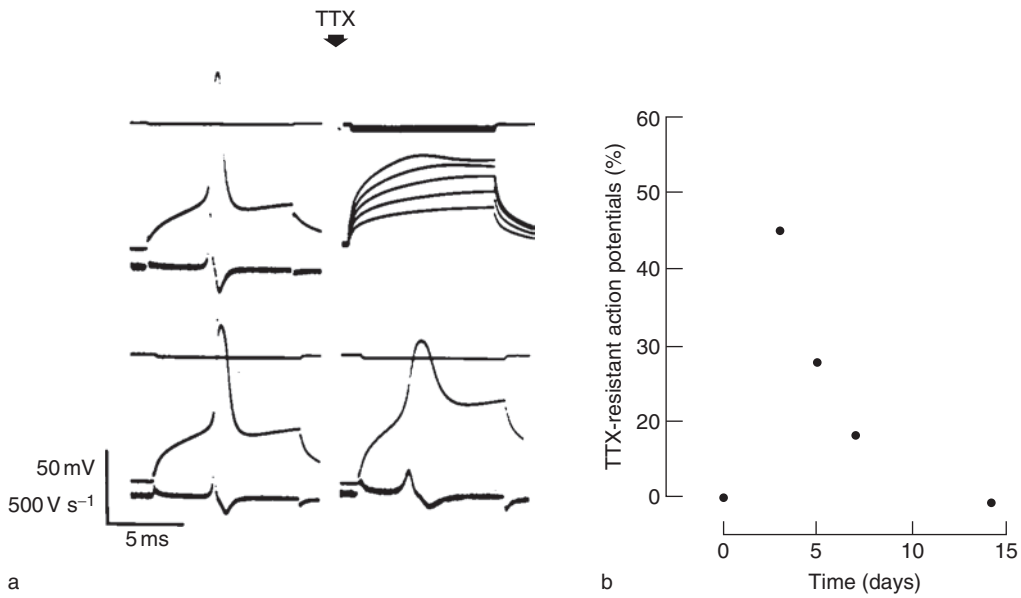

Figure 3 Tetrodotoxin (TTX) and voltage-gated Na^+ channels: (a) in controls and after denervation; (b) during reinerveraton. In (a), TTX (10^{-6} M) blocks voltage-gated Na^+ channels in mature mammalian skeletal muscle and this prevents action potential generation (top traces). After denervation, during development, or during regeneration, the gene for the TTX-resistant $Na_V1.5$ channel is upregulated and the toxin no longer blocks action potential generation (bottom traces). In (b), the generation of TTX-resistant action potentials in muscle fibers declines as reinnervation returns in denervated muscles.

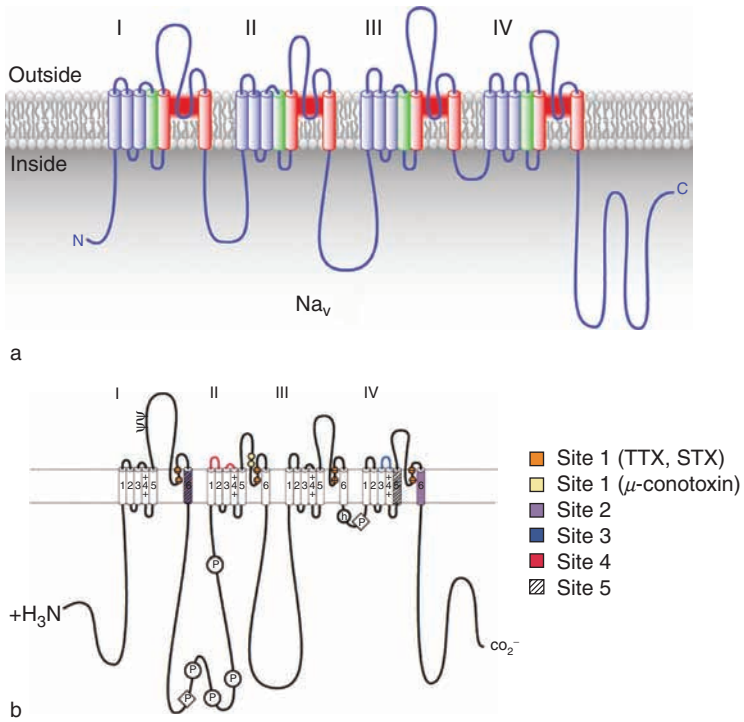

Figure 4 Voltage-gated Na^+ channel: (a) organization; (b) major binding sites for natural toxins. Modified from Catterall WA, Cestèle S, Yarov-Yarovoy V, Yu F, Konoki K, and Scheuer T (2007) Voltage-gated ion channels and gating modifier toxins. *Toxicon* 49(2): 124–141.

lipophilic alkaloids such as batrachotoxin, isolated from the skin secretions of Dendrobatid frogs of the genus *Phyllobates*, and veratridine, isolated from plants of the genus *Schoenocaulen*. Some of the toxic alkaloids are preferentially active on the nervous system of insects and therefore have potential use as biological insecticides. Toxins acting on site 2 cause the activation of the sodium channel and reduce

sodium channel inactivation, leading to a prologation of the action potential and repetitive firing at the NMJ.

Site 3 is an extracellular site occupied by a number of small polypeptides isolated from the venoms of sea anemones, the α-toxins of scorpions and the μ-agatoxins isolated from the venom of spiders of the genus *Agelenopsis*. Although the α-toxins of scorpion venoms and the sea anemone toxins all delay sodium channel inactivation, and thereby prolong action potential duration, they are structurally distinct. The scorpion toxins are small poypeptides of 60–70 residues; the sea anemone toxins are only 45–50 residues. The prolongation of the action potential in the peripheral motor system results in hyperexcitability and repetitive firing at the NMJ (Figure 5).

Site 4 is occupied by β-toxins of scorpion venoms and some spider toxins, which act to enhance activation by reducing the voltage dependence of activation. Site 5 is occupied by brevetoxin, a toxin produced by marine dinoflagellates of genus *Ptychodiscus* to cause persistent activation of the sodium channel. Site 6 is the binding site for a specific group of δ-conotoxins. These toxins cause a slow inactivation of voltage-gated sodium channels in marine molluscs.

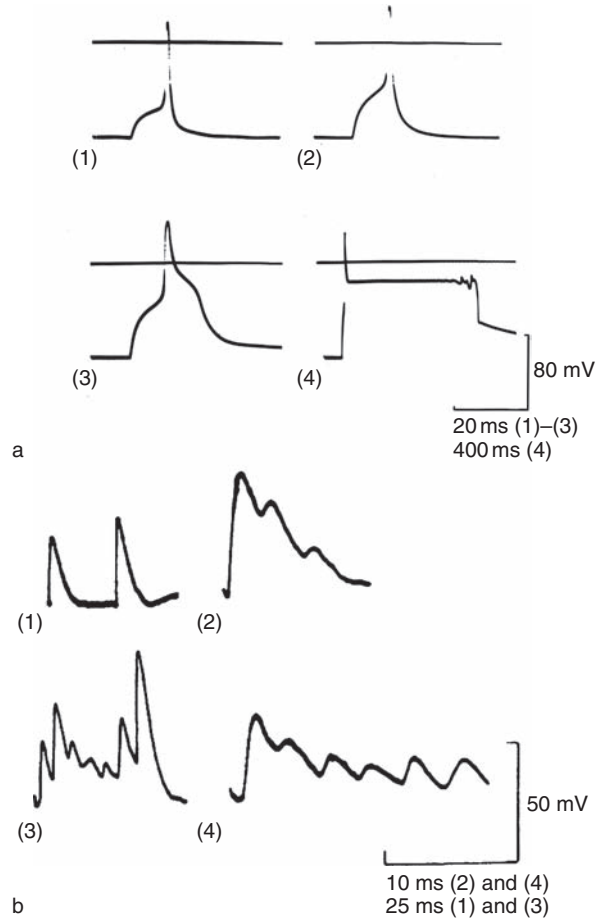

Figure 5 Effects of ATX-II, a toxin from the sea anemone *Anemonia sulcata*: (a) on Na^+ channel and action potentials; (b) on neuromuscular junction. (1) control; (2) first action potential after ATX-II; (3) tenth action potential; (4) fiftieth action potential. In (a), the Na^+ channel is inactivated (2), and this results in the prolongation of the action potential in the skeletal muscle and axons, the duration of the action potential increasing with time and activity. In (b), the prolongation of the action potential in motor axons leads to repetitive firing at the neuromuscular junction.

Calcium Channels

The voltage-gated calcium channels are major signaling transducers, converting the depolarization of a cell membrane into the influx of Ca^{2+}. The structural organization of the voltage-gated Ca^{2+} channel is virtually identical to that of the voltage-gated sodium channel, comprising an α subunit, which contains those structures that form the ion channel and the signaling domains, and one or more auxiliary subunits. As in the sodium channel, the α subunit of the calcium channel is composed of four domains, each containing six hydrophobic sequences, the pore being formed when the four domains form a pseudo-symmetrical array across the cell membrane. The known voltage-gated calcium channels fall into one of three major types, defined by their role. Ca_V1 voltage-gated calcium channels, also known as type L, are involved in the activation of intracellular signaling processes, including gene regulation, endocrine secretion, and the contraction of smooth, cardiac, and skeletal muscle. Ca_V2, also known as types N, P/Q, and R channels, regulate fast synaptic transmission in the nervous system; Ca_V3, also known as T-type channels, are primarily involved with the regulation of repetitive firing in the nervous system. There is a great deal of diversity of function within the individual types of voltage-gated calcium ion channels because there are at least ten recognizable isoforms of the α subunit, further complicated by extensive alternative splicing. A similar diversity among the auxiliary units gives rise to a high degree of combinatorial heterogeneity.

Natural toxins targeting voltage-gated calcium channels are produced by large numbers of predatory animals. The toxins involved include the ω-conotoxins, toxins from the venoms of spiders of the genera *Agenelopsis, Grammostola, Hololena,* and *Plectreurys* and of snakes of the genera *Dendroaspis*. The ω-toxins of Conus (i.e., the ω-conotoxins) and *Agelenopsis* (i.e., the ω-agatoxins) are by far the most widely used toxins in calcium channel pharmacology. Despite the very high specificity of the ω-conotoxins and-agatoxins for voltage-gated calcium channels, they are structurally unrelated. The ω-conotoxins are small polypeptides of 24–30 residues cross-linked by three disulfide bonds; the ω-agatoxins are small

polypeptides of 45–75 residues cross-linked by four disulfide bonds. With care, the pharmacological differentiation between the various types of calcium channel may be made on the basis of sensitivity to a small number of agents. L-type channels are very sensitive to the dihydropyridines and are resistant to the majority of ω-conotoxins and-agatoxins. P/Q channels are particularly sensitive to ω-conotoxin MV11C and ω-agatoxin 1VA, and N channels are particularly sensitive to ω-conotoxin GV1A. This approach to the classification of calcium channels is, however, fraught with difficulty, especially when a novel calcium channel is being classified, because the majority of ω-toxins are selective for voltage-gated calcium channels but are not particularly specific for an individual channel type.

Other toxins directly affecting voltage-gated calcium channels include ω-grammotoxin SIA, isolated from the venom of the Chilean tarantula *Grammostola spatulata,* which blocks N-, P-, and Q-type channels but not L-type; hololena toxin from the venom of the spider *Hololena curta;* and PLTX11 from the venom of the spider *Plectreurus tristes,* which block voltage-gated calcium channels in insects. Calciseptine, isolated from the venom of the mamba *Dendroaspis polylepis*, inhibits L-type Ca^{2+} channels in the aorta, ventricular muscle, and dorsal root ganglia; calcicludine, from the venom of *Dendroaspis augusticeps,* blocks L-, N-, and P-type channels in the CNS. The typical response to the blockade of voltage-gated calcium channels at the NMJ is the failure of evoked transmission and the packing of the nerve terminal with synaptic vesicles.

Potassium Channels

Voltage-gated potassium channels regulate the movement of potassium ions across cell membranes. Activation leads to an increase in conductance and the termination of action potentials, hyperpolarization, and a reduction in excitability. Conversely, a block of the channels leads to depolarization, prolongation of action potentials, repetitive firing, and increases in transmitter release and endocrine activity. The channels are homo- or heterotetramers, comprising four subunits, each composed of six transmembrane segments (Figure 6). Thus, each subunit is analogous to a single domain in the α subunit of a voltage-gated sodium or calcium channel.

The toxins known to block potassium channels fall into five major groups. The apamin group comprises two homologous small polypeptides isolated from the venom of the honey bee, *Apis mellifera*. The toxins are 18 and 22 residues in length, respectively, and each is cross-linked by two disulfide bridges. Apamin binding sites have been identified in both the CNS and a number of peripheral sites, both neuronal and non-neuronal, but it is particularly toxic when administered by directly into the CNS, where it is a potent inhibitor of small conductance Ca^{2+}-activated potassium

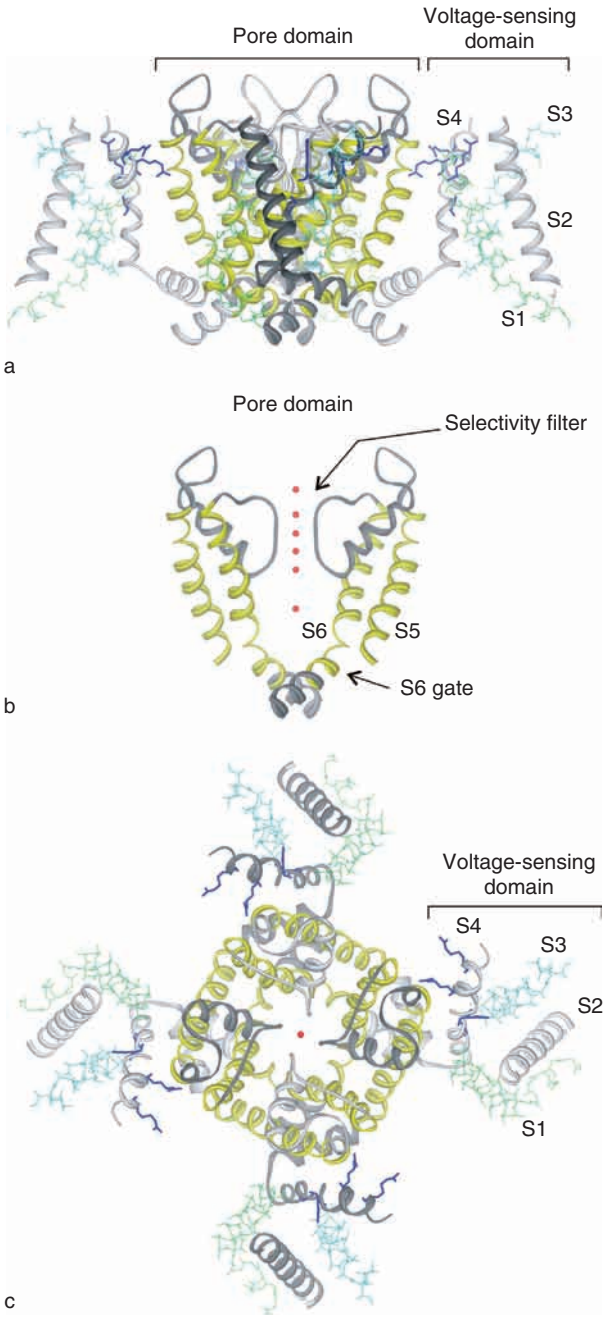

Figure 6 Organization of the tetrameric voltage-gated K^+ channel: (a) X-ray structure, extracellular domain to top; (b) side view of two lateral subunits; (c) view through pore from extracellular side. Red spheres are K^+ ions. From Swartz KJ (2007) Tarantula toxins interacting with voltage sensors in potassium channels. *Toxicon* 49(2): 213–230.

channels. MCDP blocks delayed noninactivating outward rectifier potassium channels in the CNS to cause arousal and epileptiform discharges.

The second group of toxins targeting potassium channels are members of the charybtoxin group. All toxins in this group are isolated from the venoms of scorpions. They are small homologous polypeptides, comprising 35–40 residues cross-linked by three disulfide bridges. The toxins vary greatly in their selectivity for the different types of potassium ion channel. Charybtoxin (isolated from the venom of *Leirus quinqestriatus*) is not particularly selective, blocking both $K_v1.3$ maxi-potassium channels and small conductance Ca^{2+}-activated potassium channels. Iberiotoxin (isolated from the venom of *Buthus tamulus*) appears to be highly specific for maxi-potassium channels. The ubiquitous distribution of potassium ion channels means that most of the charybtoxin group of toxins have high-affinity binding sites similarly widely spread throughout the body of both vertebrates and invertebrates. Neural targets are most widely distributed within the CNS, and as a result the toxins tend to be most toxic when administered directly into the CNS. Peripherally, the toxins tend to cause an instability of both the heart rate and blood pressure and to cause an increase in both somatic and autonomic activity – signs and symptoms generally ascribed to enhanced transmitter release from all accessible nerve terminals. Myogenic activity of smooth muscle is common. There is some evidence that at motor nerve terminals maxi-potassium channels co-localize with voltage-gated calcium channels and thus play a role in the regulation of both spontaneous and evoked transmitter release. The scorpion toxins are generally considered to block the potassium ion pore, but there is little detailed understanding of their molecular targets.

The third group of toxins targeting potassium ion channels are the dendrotoxins isolated from the venoms of three species of African mamba (*Dendroaspis augusticeps, D. viridis, and D. polylepis*). The toxins are small basic polypeptides of 57–60 residues cross-linked by three disulfide bridges. The dendrotoxins were first studied as a result of their ability to cause a facilitation of transmitter release and repetitive firing at the vertebrate NMJ. The toxins block voltage-gated potassium channels of types $K_v1.2$, $K_v1.1$, and $K_v1.6$. Because these channel types are widespread throughout the nervous system, it is not surprising that, like other toxin types blocking potassium channels, the dendrotoxins are more toxic following injection into the CNS than following intraperitoneal or intravenous injection. Several natural variants of the dendrotoxins have been identified and it is presumed that individual potassium channels exhibit different susceptibilities to the respective variants.

The fourth group of toxins comprise a distinct group of small (35–37 residues) basic polypeptides isolated from the venom of sea anemones and a group of dendrotoxin-like toxins of 58–59 residues. The precise details of binding of these toxin groups have not been fully characterized.

Presynaptically Active Neurotoxins

Snake venoms are rich sources of Ca^{2+}-dependent phospholipases A_2 (PLA_2). Although the primary role of these enzymes is probably digestive, many of the enzymes are potent, presynaptically active neurotoxins. Snakes of the family Elapidae and their close relatives the sea snakes produce enzymes homologous with pancreatic phospholipases; the viperid and crotalid snakes produce enzymes that are homologous with intracellular PLA_2. These two groups of enzyme are classified as Class 1 and Class 2 enzymes, respectively. A number of viperid snakes also produce a toxic PLA_2homolog in which the Ca^{2+}-binding Asp-49 has been substituted for a Lys residue, resulting in a reduction of hydrolytic activity. The snake venom PLA_2 toxins are polypeptides of between 115 and 125 residues. Other species producing toxic phopholipases include bees and the lizards *Heloderma suspectum* and *H. horridum,* but the enzymes are glycoproteins with little similarity with those of snake venoms, and conopidene, a dimer of two small polypeptides of 72 and 42 residues, respectively, linked by a single disulfide bond (recently isolated from the venom of *Conus magus*) appears to be a ligand for L-type calcium channels. Class 1 and Class 2 toxins may be single-chain polypeptides (e.g., notexin from the venom of the Australian tiger snakes, *Notechis* sp.), heterodimers with no covalent linkage between the two chains (e.g., crotoxin from the venom of the South American rattlesnake, *Crotalus durissus terrificus*), heterodimers of covalently bound subunits (e.g., β-bungarotoxin from the venom of kraits, *Bungarus* sp., of Southeast Asia), or multimers of up to five subunits (e.g., textilotoxin from the venom of *Pseudonaja textiles*, the Eastern brown snake).

Although the neurotoxic PLA_2 are very selective for motor nerve terminals of vertebrates, they do not have a specific identifiable receptor. Rather, they appear to bind one or more integral membrane lipids or glycoproteins. The hydrolysis of membrane phospholipids is their primary mode of action. This results in a complex sequence of events including a short-lived reduction in spontaneous and evoked reduction in transmitter release, followed by a transient recovery

before a progressive and irreversible complete failure of neuromuscular transmission. At the point of failure, neither axonal conduction nor postsynaptic function is impaired.

There is no general agreement on the biological basis of the resulting neuromuscular paralysis. However, there is agreement that the hydrolytic activity of the toxins is important and that the products, free fatty acids and lysophophatides, play a role in the expression of toxicity. One view is that the hydrolytic activity of the toxins causes a depolarization of the nerve terminal, a loss of ion homeostasis, a loss of mitochondrial metabolic capacity, the depletion of synaptic vesicles, and the degeneration of the nerve terminal (Figure 7). Another is that the depolarization is associated with the internalization of the toxins, the partial hydrolysis of internal membranes, and a failure of synaptic vesicle recycling without degeneration of the nerve terminal.

There is physiological and morphological evidence for both points of view. It was suggested that the two points of view can be resolved if it is imagined that under circumstances in which the neurotoxic potential of the toxin was high (e.g., as a result of the inherent activity of the enzyme, its concentration, the choice of tissue, and the technical aspects of the experiment), neurodegeneration is most likely outcome, and under circumstances in which the neurotoxic potential was lower, the outcome is less severe. The highly selective susceptibility of the motor nerve

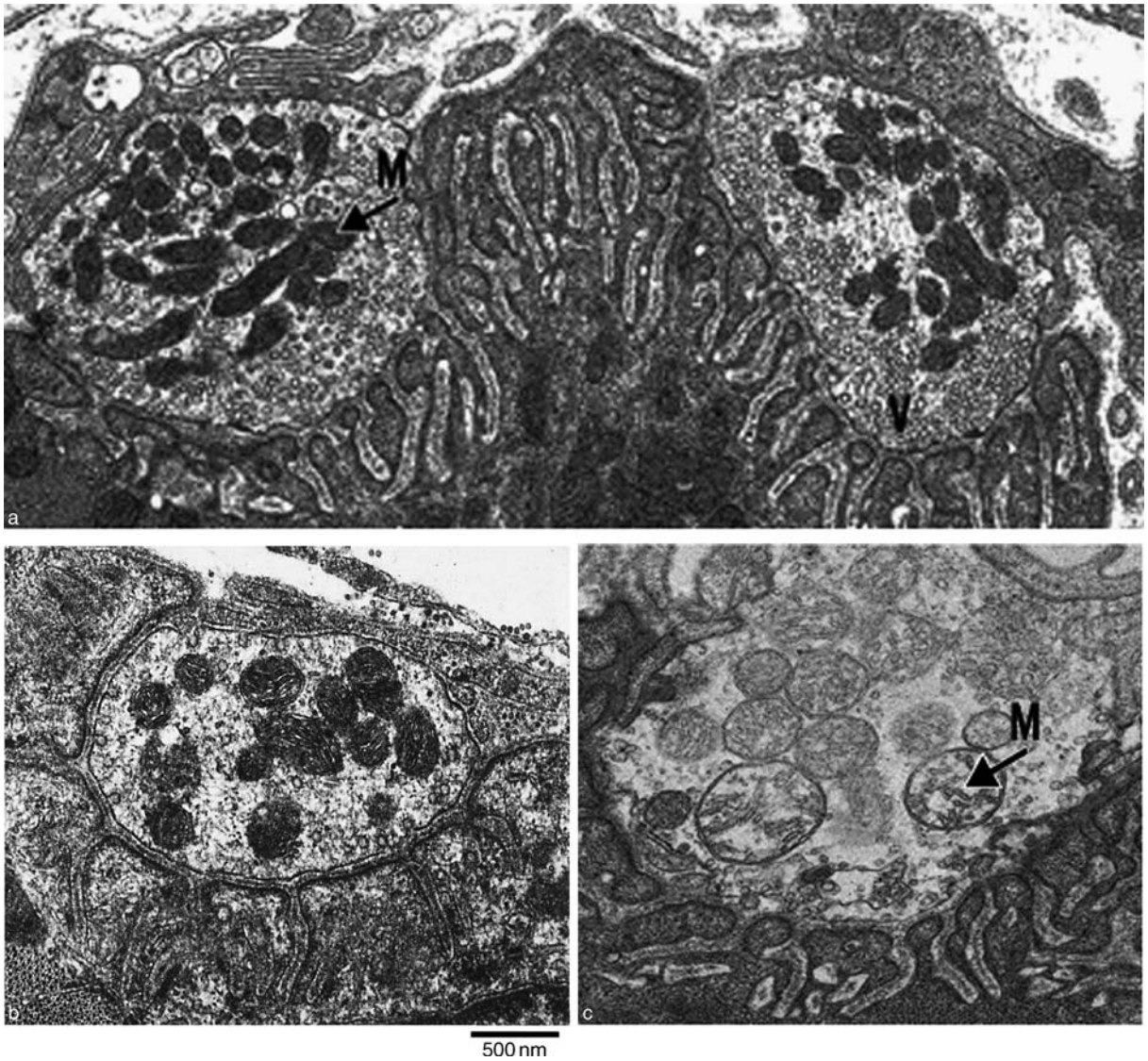

Figure 7 Effect of neurotoxic phospholipases on mammalian neuromuscular junctions: (a) control; (b) exposure to phospholipases, causing an initial depletion of synaptic vesicles; (c) the subsequent degeneration of the terminal. In (a), the nerve terminals sit in a deeply folded depression on the muscle fiber. The terminals are filled with mitochondria (M) and synaptic vesicles (V).

terminal is unexplained. Many of the neurotoxic PLA_2 are also myotoxic and thus nerve terminal and muscle pathology may coexist (see the section titled 'Myotoxic activity').

The black widow spiders, *Lactrodectus* sp., produce presynaptically active toxins typified by α-latrotoxin. This is a single-chain, mildly acidic polypeptide of 1401 residues. α-Latrotoxin causes a massive increase in the spontaneous release of transmitter from all tested peripheral and central synapses, causing a well-characterized syndrome that includes acute pain, fasciculation and fibrillation, fluctuating blood pressure and heart rate, and anxiety. The mechanism of action of the toxin is not well understood. The enhanced release of transmitter is partially Ca^{2+}-dependent, mediated by membrane-bound neurexin 1a, but can occur in the absence of external calcium when the activity is mediated by synaptophilin. Depolarization and a rise in internal calcium are also involved. The depletion of synaptic vesicles is the end result of severe latrodectism. The insect-specific neurotoxin α-latroinsectotoxin is homologous with the mammalian-specific α-latrotoxin, but its membrane-bound receptors have not been identified.

The clostridial toxins tetanospasmin and botulinum toxins are extremely important, presynaptically active neurotoxins targeting major soluble *N*-ethylmaleimide-sensitive factor attached protein receptor (SNARE) proteins syntaxin, synaptosomal-associated protein of 25 kDa (SNAP-25), and synaptotagmin.

Postsynaptically Active Toxins

The ACh receptors involved in neuromuscular transmission at the mammalian NMJ are located at the top of the synaptic folds of the muscle-fiber plasma membrane. The receptor is a heteropentameter with a stoichiometry of α2, β1, δ1, and ε1. The ε subunit is developmentally regulated, replacing the immature γ subunit during NMJ formation. The five subunits are arranged to form a central pore in the circular sequence α, β, α, ε, and δ. There are two binding sites for ACh, one at the δ–α interface and one at the α–ε interface. The binding of ACh results in a conformational change in the receptor and the opening of the ion channel. The channel is not particularly selective, and there is an increase in conductance for sodium, potassium, and calcium ions.

Postsynaptically active neurotoxins (conventionally referred to as α-neurotoxins) are elaborated in the venom glands of elapid snakes, their close relatives the sea snakes, and the fish-eating cone snails. The toxins are used by both groups of animals to immobilize their prey prior to ingestion. It is another interesting example of convergent evolution in the sense that completely different groups of animals have evolved structurally distinct polypeptides to target very specifically the postjunctional ACh receptor.

The α-neurotoxins of the elapid and sea snakes are single-chain polypetides of either 60–62 residues (the short neurotoxins) or 66–74 residues (the long neurotoxins), cross-linked by four or five disulfide bridges The most widely used α-neurotoxin is α-bungarotoxin, isolated from the venom of the krait *Bungarus multicinctus*. The toxin is a single chain of 74 residues folded into three major loops cross-linked by five disulfide bridges. Its very high affinity for ACh receptors (K_d 10^{-12}–10^{-10} M) allows it to be used very extensively in neurobiology. The precise details of binding between the ACh receptor and the α-toxins are not known, but it is clear that the toxins engage in multipoint interactions. The α-neurotoxins of cone snails are typically 13–15 residues in length with two disulfide bridges.

All the α-neurotoxins toxins bind selectively to the postjunctional ACh receptors at the vertebrate NMJ. It is often stated that the snake toxins bind preferentially to the δ–α binding site and the conotoxins bind to the α–ε site, but this is misleading because the precise affinity for the two binding sites largely depends on the primary sequences of the respective α, δ, and ε subunits. For example, sequence changes around the binding site on the α subunit are the reason for the remarkable resistance of the mongoose to the venom of the cobra and probably underlie many other apparently anomalous data on interspecies variation in sensitivity to α-neurotoxins. Toxins that target directly other sites on the ACh receptor are rather uncommon, but they are well established. A toxin isolated from the viper *Tropidolaemus wagleri* binds selectively to the ε subunit and can cause a potentially fatal failure of neuromuscular transmission; a toxin from the cone snail *Conus obscurus* specifically inhibits the immature ACh receptor. Natural toxins that target ACh receptor populations in the CNS and muscarinic ACh receptors are not considered here.

Myotoxic Activity

Most of the neurotoxic phospholipases A_2 are also myotoxic. Details of the myotoxic process have been studied only for notexin, a neurotoxic phospholipase isolated from the venom of the Australian tiger snake, *Notechis scutatus*. The toxins bind to and hydrolyze the sarcolemma. The hydrolysis of the membrane leads to the breakdown of the internal structure of the muscle fiber (Figure 8). The postsynaptic membrane (including its integral ACh receptors) appears to be relatively resistant to the process of degeneration, even when both muscle fiber and nerve terminal have been destroyed. The basal lamina of the muscle fiber,

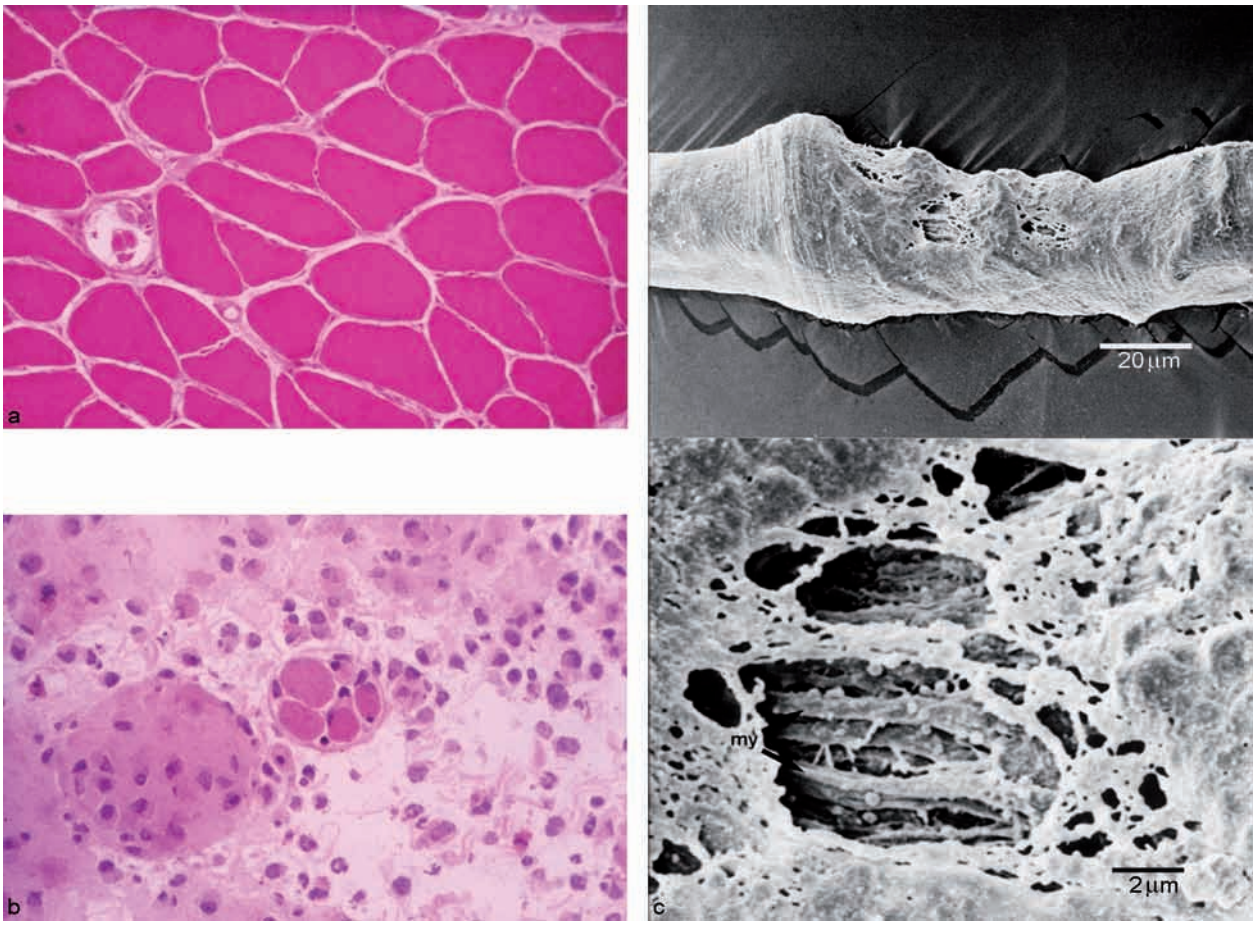

Figure 8 Muscle damage caused by myotoxic phospholipases from snake venoms: (a) control muscle; (b) 24-h after exposure to the toxin notexin; (c) scanning electron micrographs of single muscle fibers, showing hypercontraction and membrane lesions in the damaged fibers. In (b), note the enlarged pale muscle fibers and the survival of the intrafusal fibers of the muscle spindle.

the satellite cells, and the microcirculation are preserved in muscle fibers damaged by these toxins, and as a result regeneration begins within 3 days. The myotoxic activity does not extend to smooth muscle, but cardiac muscle may be susceptible.

The Anatoxins

The anatoxins are toxic constituents of blue-green algae produced by *Anabaena flos-aquae* and a number of other freshwater species of cyanobacteria. The toxins are water-soluble alkaloids that have been responsible for a number of incidents of human poisoning, some of which have been fatal. Anatoxin a and homoanatoxin are powerful agonists of the ACh receptor, causing prolonged excitation at the NMJ. Anatoxin a (s) is a natural organophosphate that causes a prolonged inhibition of acetylcholinesterase. The anatoxins may be expressed in the absence or presence of the microcystins. Anatoxin-producing cyanobacteria have been identified in freshwater lakes and reservoirs in the United Kingdom, Japan, and North America.

Miscellaneous

A number of toxins commonly identified as neurotoxins cause neurotoxic signs as an indirect consequence of a more general pharmacological effect. For example, the pore-forming toxins such as equinotoxins, produced by a large number of sea anemones of the genus *Actinaria* and pardaxin from the skin secretion of the fish *Pardachirus marmoratus* are primarily cytotoxins, causing *inter alia* an increase in exocytosis from a number of synapses and palytoxin that was first isolated from soft anthozoans *Palythoa*. The toxin converts Na^+/K^+ ATPase from a selective Na^+/K^+ exchanger into a nonselective open cation pore. It, too, causes an enhanced exocytosis.

Summary

This article has clearly been limited, but natural neurotoxins that have evolved in a very large number of species are discussed. Most have arisen in the predatory animals that use venoms as tools to subdue prey and to defend themselves against predators.

The polypeptide toxins are very common, and many animals use a battery of such toxins as part of their armory; for example, the fish-eating cone snails include toxins that bind to the ACh receptor and block voltage-gated Na^+ and Ca^{2+} channels, along with numerous other constituents. Often the natural toxins are very selective in terms of their molecular targets, and this allows the toxins to be used as biological tools in the study of synaptic activity and receptor and ion channel pharmacology. There is also growing interest, as well, in the use of these highly selective ligands as potential new therapeutic agents.

See also: Botulinum and Tetanus Toxins; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction Plasticity in Mammals and Botulinum Toxins; Neurotoxins; Neurotoxins and their Neurotoxicology.

Further Reading

Alexander SPH, Mathie A, and Peters JA (2006) *Special Issue: Guide to Receptors and Channels,* 2nd edn. *British Journal of Pharmacology* 147(supplement 3).

Catterall WA and Cestèle S, Yarov-Yarovoy V, Yu F, Konoki K and Scheuer T (2007) Voltage-gated ion channels and gating modifier toxins. *Toxicon* 49(2): 124–141.

Garcia ML and Possani LD (2007) Gating modifier peptides of ion channels. *Toxicon* 49(2): 123.

Harvey AL (ed.) (1991) *Snake Toxins*. New York: Pergamon Press.

Rappuoli R and Montecucco C (1997) *Protein Toxins and Their Use in Cell Biology.* Oxford: Oxford University Press.

Spencer PS, Schaumburg HH, and Ludolph AC (2000) *Experimental and Clinical Neurotoxicology,* 2nd edn. New York: Oxford University Press.

Swartz KJ (2007) Tarantula toxins interacting with voltage sensors in potassium channels. *Toxicon* 49(2): 213–230.

Neuromuscular Junction (NMJ): Acetylcholinesterases

R L Rotundo, University of Miami Miller School of Medicine, Miami, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Acetylcholinesterase (AChE) was first described by Sir Henry Dale in 1914 as an "esteratic activity" present in mammalian tissues that could hydrolyze certain esters of choline. He later went on to share the Nobel Prize in Physiology and Medicine in 1936 with Otto Loewi for their discovery of acetylcholine, the first neurotransmitter, and for their work demonstrating chemical neurotransmission at synapses. Thus began the study of AChE, the first identified synaptic component, almost 100 years ago. Today, this enzyme is still being studied as a marker for cholinergic synapses in the central nervous system (CNS) and peripheral nervous system, some of which degenerate in certain diseases such as Alzheimer's and some forms of myasthenias; as a target for therapeutic drugs in humans and animals; as part of the mechanism of action for several important pesticides used worldwide; and unfortunately as the major target for many highly toxic nerve gases. Although this enzyme has been well studied for many years, only within the past 25 has the biochemistry, cell, and molecular biology of AChE been worked out in detail. Modern methods of visualizing AChE at synapses, such as the use of fluorescently tagged toxin molecules that bind specifically to the enzyme, are used to observe the presence of the molecules at the neuromuscular synapse, where they co-localize with those of the nicotinic acetylcholine receptor (Figure 1). This article focuses specifically on the biogenesis and localization of AChE at the neuromuscular synapse.

The major if not unique function of AChE, in fact the only indisputable function, is to terminate neurotransmission at cholinergic synapses in the CNS and peripheral nervous system. This function is accomplished through the hydrolysis of acetylcholine into choline and acetic acid. Once hydrolyzed, the products are free to diffuse away from the synapse or are taken up by the presynaptic terminals. AChE is essential for normal function of the nervous system, and the irreversible inhibition of its activity by specific organophosphate compounds such as nerve gases and certain pesticides are often fatal. A transgenic mouse has been developed that is AChE null and survives weakly in captivity, most likely because the phenotype has been rescued by another enzyme called butyrylcholinesterase. Inhibition of butyrylcholinesterase, which normally has no effect on the organism, results in rapid death of the AChE-null mice. Otherwise, the mice appear morphologically and physiologically normal.

AChE was also the first synaptic component to be studied at the structural level; it was ultimately crystallized in the laboratory of Joel Sussman and Israel Silman. Hence, we probably know more about the details of this molecule than any other at the synapse. This has greatly facilitated our understanding of the mechanisms of catalysis as well as the mechanisms of localization. However, one surprise was that the catalytic site of the enzyme was not near the surface, as was widely believed because of its very high turnover number, but rather deeply embedded within the center molecule and accessible only by diffusion through a long narrow space called the gorge. It is still not clear how the enzymatic reaction products leave the active site of the enzyme while new substrate molecules enter. One working hypothesis is that there is in fact a back door through which choline and acetate can diffuse out while the new substrate molecules enter through the front door near the peripheral anionic site.

AChE Detection and Assay

With one of the highest turnover numbers of any enzyme known, AChE is one of the easiest and most sensitive to measure. Many different assays based on catalytic activity (based on the hydrolysis of acetylcholine to choline and acetic acid) have been developed over the years, with almost as many detection methods. The products of acetylcholine hydrolysis are easily measured; acetic acid reacts with bicarbonate to produce CO_2, whose rate of production can be quantified, and the acid can also be measured by acid–base titration, potentiometrically, and colorimetrically. Choline production can also be assayed, and radioisotopically tagged versions of the substrate can be measured quantitatively. Chemically modified substrates, such as acetylthiocholine, can be used as well because they can be measured colorimetrically. Finally, radioactive and fluorescently tagged AChE ligands, such as many organophosphate compounds and toxins, have been used to visualize the enzyme *in situ* at the light- and electron-microscopic levels. Thus, AChE can be detected, visualized, and quantified at the biochemical, cellular, tissue, and organismic levels with ease, making it one of the best-studied enzymes and an important tool for biologists in many disciplines.

Structure of AChE

In all vertebrate species, AChE is encoded by a single gene that gives rise to multiple forms of the enzyme in a cell type-specific manner (Figure 2). Depending on the species, this gene can give rise to either one or two alternatively spliced variants and, in a least one species, to a third form referred to as the readthrough form. Although more than 90% of the amino acid sequence is shared by all AChE forms in each species, the two major alternative splicing variants encode different C-termini. The first is capable of assembling with one of two noncatalytic subunits that specify cellular location, the variant expressed in the skeletal muscles and neurons of higher vertebrates, and is called the t form of the AChE catalytic subunit ($AChE_t$) after its original association with a collagen-like tail. The other, called the h form for hydrophobic ($AChE_h$), signals the addition of a glycophosphoinositol (GPI) lipid membrane anchor. This latter form is expressed only in cells of the hematopoietic lineages, red and white blood cells.

The overall structure of AChE is virtually identical in all species examined. The catalytic monomer consists of a single polypeptide chain of approximately 600 amino acids in most species and contains three to four asparagine-linked oligosaccharide side chains that are processed differently depending on the cell type in which they are expressed. However, all contain the core high mannose residues as well as *N*-acetylglucosamine and *N*-acetylglactosamine residues. Variable residues include the presence of and extent of a terminal sialic acid addition as well as the presence of as yet incompletely characterized sugar residues. Thus, AChE is approximately 15% carbohydrate by weight. No other posttranslational modifications have been described other than the addition of the GPI anchors, except for possible palmitoylation in one instance and the addition of HNK epitopes in another.

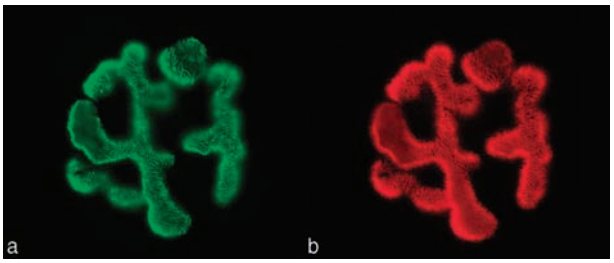

Figure 1 Comparison of the distribution of acetylcholine receptors (AChRs) and acetylcholinesterase (AChE): (a) distribution of the nicotinic AChRs (labeled with oregon green conjugated α-bungarotoxin); (b) distribution of AChE at the mouse neuromuscular junction (labeled with Rhodamine-fasciculin 2). Notice the near perfect alignment of the two molecules at the synapse involved in initiating and terminating neurotransmission.

AChE Oligomeric Forms

The different AChE forms arising from these gene products are illustrated in Figure 2. The molecular mass of this protein is approximately 75 kDa in most species; however, in birds it is closer to 100–110 kDa for each catalytic subunit. The soluble globular $AChE_t$ subunits can associate to form homomeric disulfide-bonded dimers and tetramers. These tetramers, in turn, can associate with a 154-amino-acid

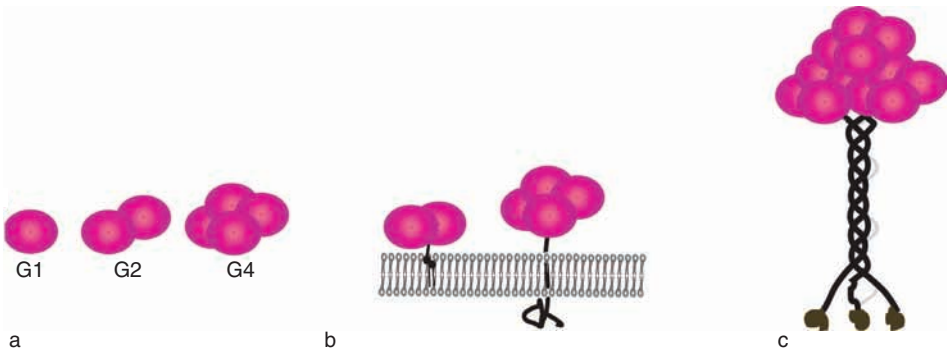

Figure 2 Structure of the acetylcholinesterase (AChE) oligomeric forms: (a) globular forms; (b) membrane-bound forms; (c) extracellular matrix (ECM)-associated form. The AChE forms expressed in nerves and muscle can assemble into several different oligomeric forms depending on which alternative splicing variants or which noncatalytic subunits are expressed by the cells. These forms include soluble globular forms in (a), referred to as the G forms, and the membrane-bound dimers and tetramers (b), the former being a glycolipid-anchored molecule encoded by an alternative spliced mRNA expressed in many species and the tetramer that is attached to a 154-amino-acid transmembrane polypeptide chain. The complex hetero-oligomer associated with the ECM forming the synaptic basal lamina at the neuromuscular junction (c) is the collagen-tailed form ColQ-AChE, which consists of three tetramers covalently attached to a three-stranded collagen-like tail.

transmembrane peptide called the p peptide or proline-rich membrane anchor (PRiMA) that anchors it in the membranes of neurons and maybe some types of muscle. Alternatively, the tetramers can associate with the approximately 150 kDa three-stranded collagen-like tail (each of three strands approximately 500 amino acids long), called ColQ, which attaches it to the synaptic basal lamina of skeletal muscle. This appears to be the unique form associated with the neuromuscular junction.

The basic functional unit of AChE is the catalytically active monomer, an exportable glycoprotein that acquires its activity shortly after synthesis and rapidly assembles with another monomer to form the dimeric enzyme. This applies to both the $AChE_t$ and $AChE_h$ types of subunits. The $AChE_r$, what little there is, always appears to be a monomer. These AChE dimers can form covalent interchain disulfide bonds via a C-terminal cystein. For the $AChE_t$ molecules, the dimers can, in turn, associate with other dimers to form the tetrameric AChE molecules. Both the dimers and tetramers can be secreted unmodified or further assembled with one of the noncatalytic subunits described next. For the $AChE_h$ the monomers can form disulfide-bonded dimers, both of which are glycolipid-anchored, but cannot further associate with other subunits. These, as already mentioned, are destined for the cell surface of erythrocytes and certain types of lymphocytes, where their function is yet to be determined.

Noncatalytic AChE Subunits

Two types of noncatalytic AChE subunits have been described. The first, discovered in the late 1960s in Jean Massoulié's lab in Paris, is ColQ. This three-stranded molecule, a trimer consisting of three identical 50 kDa polypeptide chains, can attach through a specific region near the N-terminus to the C-terminus of the AChE catalytic subunit (see Figure 2(c)). This subunit is encoded by a separate gene, in fact on a separate chromosome, and exists in a least two alternative splicing variants expressed in skeletal muscle. These two variants differ in their N-terminus primarily in the region encoding the leader peptide. The function of this variation is not known.

The primary and secondary structure of ColQ is well established. This almost 50 kDa subunit consists of approximately 460 amino acids with no known posttranslational modifications, divided into roughly three regions. The first of these is an N-terminus consisting of the first 95 amino acids including the cleavable leader peptide and containing the proline-rich attachment domain or PRAD, the 17-amino-acid proline-rich sequence that interacts with the α-helix-rich C-terminal domain of AChE to form the tetramers. The second region comprises a glycine- and proline-rich stretch approximately 200 amino acids long containing hydroxyproline and that is virtually indistinguishable from a regular collagen molecule. This domain forms a triple helical structure after synthesis and assembly and contains two binding sites for interactions with heparin. The third main region is the C-terminal 160 amino acids and consists of a stretch of hydrophilic amino acids followed by a cystein-rich region with ten cystein residues that most likely forms a tight globular structure. This region is essential for attachment to the extracellular matrix of the synaptic basal lamina, and some single amino acid mutations in this region are capable of preventing the attachment of AChE to the neuromuscular junction.

The second type of noncatalytic subunit is the 154-amino-acid transmembrane polypeptide PRiMA. This relatively small molecule has a roughly 24-amino-acid leader sequence, a 67-amino-acid extracellular domain followed by a 22-residue transmembrane segment, and a 41-amino-acid intracellular domain. This polypeptide is also encoded by a separate gene and exists in at least two distinct splicing variants. Again, the functional significance of these variants has not yet been determined. Like ColQ, this subunit also has a PRAD domain at the N-terminus that interacts with the AChE catalytic subunits to form tetramers. These tetramers are the major form expressed in the nervous system of higher vertebrates, where they localize the enzyme to specialized regions of the neuronal cell surface.

Biogenesis, Processing, and Assembly of AChE

All AChE forms are glycoproteins destined for secretion or localization on the cell surface and, hence, are synthesized on the rough endoplasmic reticulum and co-translationally glycosylated. It appears that only asparagine-linked oligosaccharides are associated with the AChE catalytic subunit, whereas no posttranslational modifications has been found on either of the noncatalytic subunits which are also synthesized on the rough endoplasmic reticulum. The catalytic subunit appears to be synthesized as an initially catalytically inactive molecule which then undergoes the necessary folding and processing to produce the catalytically active enzyme. The active molecules can then associate to form active dimers and tetramers that, in turn, transit the Golgi apparatus and undergo additional carbohydrate processing to yield the fully mature forms. These mature and immature molecules can be distinguished by their resistance to specific endoglycosidases as well as structural stability. The inactive

AChE forms, consisting mostly of monomers and some dimers, are rapidly degraded by the endoplasmic reticulum-associated protein degradation system.

The intracellular assembly sites of the different AChE oligomeric forms are still incompletely understood. The dimeric and tetrameric forms are clearly assembled in the endoplasmic reticulum, as is the cleavage of the $AChE_h$ C-terminal amino acids and the addition of the glycolipid anchor. The addition of the transmembrane polypeptide anchor most likely occurs in the endoplasmic reticulum as well, although this has not yet been proven. Only one study has addressed the site of assembly of the ColQ AChE form, and it appears that assembly occurs in the Golgi apparatus approximately 90 min following synthesis of the globular forms. This is based on pulse–chase studies of catalytically active enzyme as well as the timing of the addition of the terminal carbohydrate residues which are known to occur in the Golgi apparatus. However, it appears that we cannot easily distinguish between an AChE tetramer and a tetramer associated with a single strand of the ColQ subunit. Thus, additional studies will need to be done to fully understand the assembly of the synaptic ColQ AChE form.

Skeletal muscle fibers are long multinucleated cells that can contain well over 1000 individual nuclei. Yet the motor nerve terminal contacts the muscle fiber on only a very small portion of the total surface membrane, usually less than 1% (Figure 3). Beneath this small region of contact a number of muscle cell nuclei accumulate, referred to as the subsynaptic or junctional nuclei, and their role appears to be to express many if not all of the genes encoding the proteins involved in forming the muscle's portion of the synapse. Thus, the genes encoding AChE, the acetylcholine receptor subunits, and at least several other synapse specific proteins have been shown to be selectively expressed by the subsynaptic nuclei. Moreover, once synthesized, these molecules appear to be co-packaged and co-transported to the plasma membrane, where they are jointly inserted into the synaptic domain. This highly localized expression of AChE ensures that the synapse has sufficient numbers of enzyme molecules to terminate neurotransmission.

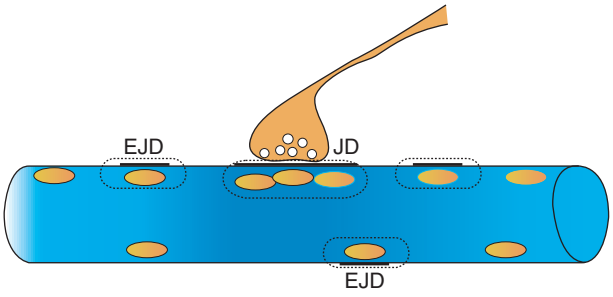

Figure 3 Organization of the nerve terminal in relation to the muscle fiber. The acetylcholinesterase (AChE) forms expressed in nerves and muscle can be regulated at several different levels depending on where they are in the muscle fiber. Although the extrajunctional nuclei are distributed evenly throughout the length of the muscle fiber, a small number (usually 5–10) of specialized nuclei accumulated beneath the site of nerve muscle contact. The junctional nuclei express the genes encoding synaptic components. EJD, extrajunctional domain; JD, junctional domain.Adapted from Rotundo RL, Rossi SG, Kimbell LM, Ruiz C, and Marrero E (2005) Targeting acetylcholinesterase to the NMJ. *Chemico-Biological Interactions* 157–158(1): 15–21, with permission from Elsevier.

Mechanism of AChE Localization at the Neuromuscular Junction

We have only recently begun to understand the molecular mechanisms involved in localizing AChE at the neuromuscular synapse. Depending on the species, a density of between approximately 500 and 2500 AChE catalytic subunits μm^{-2} are required to maintain normal transmission at the neuromuscular synapse. This includes an approximately fivefold safety factor because cholinergic transmission is not significantly impaired until approximately 80% of the AChE molecules at the synapse are inactivated. More than 25 years ago, ColQ AChE was shown to be highly concentrated at the neuromuscular junction, and its appearance was demonstrated to be under neural control. In addition, a strong basement membrane surrounds each muscle fiber, enveloping it like a cocoon, with a specialized region being interposed between the nerve terminal and the postsynaptic acetylcholine receptor-rich membrane. This specialized region is referred to as the synaptic basal lamina, and it contains many molecules unique to sites of nerve–muscle contact. The synaptic AChE was shown to be tightly associated with the synaptic basal lamina, where it could persist for weeks or months after denervation and could interact directly with several components of the synaptic basal lamina. The synaptic ColQ AChE form was shown to bind directly to several different extracellular matrix molecules *in vitro*, including fibronectin, chondroitin sulfate, and heparan sulfate proteoglycans, but it was not until direct binding of ColQ AChE to the synaptic basal lamina was demonstrated that the idea that a specific acceptor molecule was present at the synapse was entertained.

The molecule that ColQ AChE appeared to be binding to was subsequently identified as perlecan, a multifunctional heparan sulfate proteoglycans present in many tissues and highly concentrated at the neuromuscular junction. Direct binding of ColQ AChE was shown *in vitro* and *in vivo*, and both perlecan- and dystroglycan-null mice (dystroglycan is a transmembrane molecule involved in tethering perlecan to the

muscle cell surface) were shown to be devoid of synaptic AChE. Thus, a reasonable case can be made for perlecan being the major binding site for AChE at sites of nerve–muscle contact.

Putting together many of the studies described here, it now appears that local expression of the AChE gene, local translation of the mRNA, and local assembly and transport of the molecules to the cell surface underlie the specific accumulation of the enzyme at the neuromuscular synapse. This all occurs in what is referred to as the junctional domain (Figure 3). The role of perlecan, then, appears to be to retain the ColQ AChE on the synaptic basal lamina, thus preventing its diffusion into the circulation. However, what ultimately determines the numbers of AChE molecules present on the synaptic basal lamina at any one time remains to be determined. Also, how AChE is turned over at the synapse and the mechanisms that determine the density and distribution of AChE at the sites of nerve muscle contact remain to be understood.

Appearance of AChE during Development of Skeletal Muscle

The motor neurons form their initial contacts with the muscle fibers very early during embryonic development, shortly after they differentiate and grow out of the spinal chord into the muscle cells mass. In fact, the spinal chord motor neurons are the first neurons to differentiate in vertebrates, and the neuromuscular connections are the first functional synapses to form in the nervous system. Thus, all the molecules necessary for initiating and terminating neurotransmission must appear very early during development. And that is exactly what has been observed by many investigators. The nicotinic acetylcholine receptors begin to accumulate at sites of nerve–muscle contact within minutes after the first contact has formed, and the esterase molecules follow shortly behind.

However, how quickly the AChE molecules appear is not yet precisely known because of the large differences in the numbers of receptor versus esterase molecules that populate the neuromuscular synapse. The number of AChE molecules at the synapse is approximately 5–20% of the number of receptors, and thus it is more difficult to detect synaptic AChE during the earliest stages of development. Moreover, early studies on the developmental appearance of AChE relied entirely on the colorimetric histochemical localization techniques which require the reaction products to bind to molecules localized at the synapse themselves. It is quite likely that the immature synapse does not bind many of these reaction products, thus preventing the existing AChE molecules from being detected.

The current view on the appearance of AChE at the developing neuromuscular junction is that it appears some time after the initial appearance of the nicotinic acetylcholine receptors. How much time depends on the method of detection, but it can be between 8 and 48 h. This model holds that the nerve signals the muscle cell to start aggregating synaptic components at sites of nerve–muscle contact, producing surface clusters of these molecules, with components like the receptor moving in first. Subsequently, other components accumulate to stabilize the synapse, including the appearance of extracellular matrix molecules that, in turn, anchor the ColQ AChE to the synaptic basal lamina.

Regulation of AChE Expression in Muscle

It is essential that the correct number of AChE molecules be maintained at the neuromuscular synapse for synaptic transmission to occur. Too few enzyme molecules interposed between the nerve terminal and the postsynaptic membrane allow too many acetylcholine molecules to pass through and overstimulate the muscle, causing abnormal and prolonged contraction. Conversely, too many AChE molecules hydrolyze the acetylcholine before it can reach the postsynaptic membrane, thereby effectively preventing neurotransmission. In both cases, the result is muscle failure and death because the diaphragm muscle depends on rhythmic acetylcholine release to contract so that breathing to occur.

Regulation of AChE expression occurs at many levels. The regulation of AChE is very complex, occurring at the transcriptional, posttranscription, translational, and posttranslational levels. Even the assembly process itself is highly controlled. First, during muscle development the proliferating mononucleated myoblasts do not express AChE transcripts. It's not until after the proliferating cells have withdrawn from the cell cycle and undergone terminal division that the AChE genes are expressed. However, the newly expressed mRNAs must also be localized and regulated, and here several important RNA-binding proteins associate with the transcript to regulate its translation. Once translated, the newly synthesized polypeptides must undergo conformational maturation to acquire catalytic activity and associate with other catalytic as well as the noncatalytic subunits. Improperly folded and inactive enzyme molecules are rapidly degraded intracellularly, whereas the correctly assembled molecules transit the cell and are allowed to reach the cell surface.

In skeletal muscle, AChE is regulated by the activity state of the fiber, that is, the extent to which it is contracting and the membrane depolarized, as well as specific signals from the nerve. All these signals and events are integrated by the muscle fiber to ultimately control the rate of AChE production, determine which oligomeric forms are expresses, and determine how many reach the synapse to be integrated into the molecular signaling machinery necessary to elicit muscle contraction.

Consequences of AChE Inactivation in Muscle, Pesticides, Nerve Gases, Congenital Myasthenia, and Alzheimer's Disease

As discussed earlier, the function of AChE is to terminate neuromuscular transmission by rapidly hydrolyzing the acetylcholine released by at the motor nerve terminal. This hydrolysis of acetylcholine occurs extremely rapidly, with a half-life of approximately 1 ms, and is absolutely necessary for normal synaptic transmission. Imagine a hummingbird beating its wings at over 100 times s^{-1}; each cycle of muscle contraction and relaxation of the pectoral muscles must occur within less than 10 ms if the flight pattern is to be maintained. If the half-life of acetylcholine in the synaptic space is 1 ms, then it requires approximately 4 ms to reduce the local neurotransmitter concentration to less than 90% of its initial amount, allowing for a small safety margin. Although this may be an extreme example, there are many rapidly moving muscles in humans and other mammals, such as the extraoccular muscles of the eyes that need to move to rapidly in order to track moving objects or even limb-muscle contracting in response to a noxious stimulus.

The behavior of the acetylcholine molecules released by the motor nerve at the synapse and the effects of inhibiting AChE are depicted in Figure 4. These beautiful images are supercomputer simulations modeling the time course of the distribution of acetylcholine molecules once released from a point source at the presynaptic terminal. Figure 4(a) shows the time course of acetylcholine hydrolysis under normal conditions and following inactivation by an irreversible inhibitor, such as a nerve gas or organophosphate pesticide. The several-fold longer time course of acetylcholine receptor activation following AChE inhibition allows for continued membrane depolarization and a protracted time of muscle contraction. In fact, depending on the rate of nerve stimulation, this can be so severe that the muscles do not have time to relax, resulting in muscle paralysis. When this affects the muscles of the diaphragm, death can result within minutes.

Another consequence of the overactivation of the acetylcholine receptors at the neuromuscular synapse is the excessive influx of Ca^{2+} ions that can occur through both the nicotinic receptor itself and the voltage-dependent Ca^{2+} channels in the muscle membrane responsible for initiating muscle contraction. This excessive Ca^{2+} can activate specific proteases such as calpain that, in turn, can partially digest the muscle-specific proteins making up the contractile apparatus. Thus, excessive stimulation of muscle can result in severe muscle degeneration and a destruction of the synaptic region of the muscle from the inside.

Again, in most animals, the complete inactivation of AChE is invariably fatal. For this reason, many of the most successful and lethal pesticides and nerve gases are anticholinesterase compounds. Pesticides, such as malathion and parathion, are organophosphates structurally similar to nerve gases that have a greater specificity for insect AChE than for vertebrate AChE. However, their specificity is dose-dependent, and at high enough doses they can also irreversibly inhibit the vertebrate enzyme. In underdeveloped countries, for example, where the pesticides are not as carefully controlled and training in their use is minimal, tens of thousands of people die annually from exposure to pesticides. Some estimates range as high as in the low hundreds of thousands. Likewise, nerve gases pose a continuous problem in many parts of the world. Although many of the developed countries are destroying their stocks, many more countries are keeping large stockpiles as a deterrent to land invasions. Sadly, nerve agents are also serious potential threats from nefarious groups, as occurred in the 1995 Tokyo subway disaster in which thousands of people were seriously poisoned and almost a dozen killed when the nerve agent sarin was released by a fringe cult. Unfortunately, these compounds are all too easily made by someone who barely has taken a first-year college chemistry course.

Because of its essential role in neuromuscular transmission, disorders involving the loss of AChE are extremely rare. A congenital form of myasthenia with synaptic AChE deficiency was discovered by the neurologist Andrew Engel in which the patients exhibit normal expression of the AChE catalytic subunit but are either lacking or express mutated forms of the collagenic tail. The disorder is so devastating that only a few dozen cases have been described worldwide, and those that have survived it most likely have done so because of other compensatory changes, such as possibly the expression of high levels of butyrylcholinesterase. It is likely that the majority of individuals with this rare and devastating disorder are either still born or die shortly after birth due to respiratory failure.

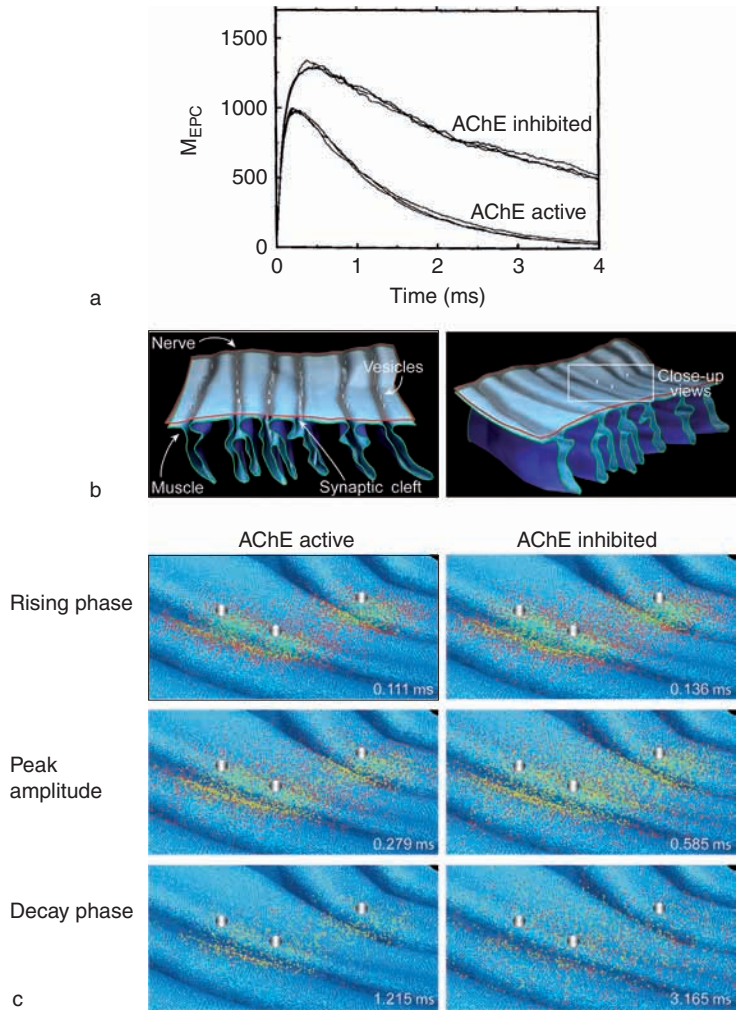

Figure 4 Effects of inhibiting acetylcholinesterase (AChE) on the amplitude and time course of the synaptic currents: (a) measured experimental postsynaptic responses (M_{EPC}s), the number of open acetylcholine receptor (AChR) channels; (b) realistic low-magnification view of the model neuromuscular junction (NMJ) used in the simulation, with folded postsynaptic muscle membrane and synaptic vesicles aligned for release from the nerve; (c) higher-magnification visualization of the neurotransmission at the vertebrate NMJ (the rectangle outlined in (b)) using the Monte Carlo simulation in the MCell/DReAMM environment for spatially realistic micro-physiological simulations. The M_{EPC}s in (a) and results of the simulation (b–c) completely overlap, showing that AChE inhibition causes an increase in amplitude and a prolongation of the decay time course. In (c), each image shows an oblique view of the folded muscle membrane, and three synaptic vesicles are indicated by gray spheres (50 nm in diameter). The nerve terminal membrane is transparent for clarity, as are AChE active sites within the cleft space. Unbound muscle AChRs are visible as blue stippling on the membrane surface. AChRs with a single bound ACh are red. Double-bound AChRs in the closed ion-channel conformation are green, and those in the open conformation are yellow. A comparison of the right and left images illustrates the spatiotemporal effects of AChE inhibition: the additional and prolonged spread of diffusing ACh molecules released from the vesicles and the overlap of the resulting AChR activation during the current rising phase (top; at 50% of peak amplitude), peak amplitude (middle), and falling phase (bottom; at 50% of peak amplitude). All these increase the rise time, peak amplitude, and, most dramatically, the falling phase. Simulations and visualizations by Jacob Czech and Joel Stiles, Center for Quantitative Biological Simulation, Pittsburgh Supercomputing Center, Carnegie Mellon University. Courtesy of Joel R Stiles and Lili Anglister; adapted following Anglister L, Stiles JR, and Salpeter MM (1994) Acetylcholinesterase density and turnover number at frog neuromuscular junctions, with modeling of their role in synaptic function. *Neuron* 12: 783–794.

On a more positive note, most of the important current treatments for Alzheimer's disease are selective AChE inhibitors. Their actions are important in the CNS, however, in which degeneration of the major cholinergic inputs to centers involved in memory and learning occur. Here, the inhibition of AChE is beneficial insofar as it prolongs the actions of acetylcholine at certain CNS synapses and enhances the persistence of short-term memory. In the peripheral nervous system, AChE inhibitors such as neostigmine and pyridostigmine that do not cross the blood–brain barrier, and that hence have no CNS effects, are used for the treatment of myasthenia gravis, an autoimmune disorder affecting the numbers of acetylcholine receptors at the neuromuscular synapse.

Epilogue

AChE has been studied for almost 100 years, during which time it has served as a model molecule for understanding enzyme catalysis; cellular differentiation; nerve–target interactions; regulation of gene expression; and many other aspects of biochemistry, pharmacology, cell biology, physiology, and molecular biology. While we do know quite a bit about AChE at the neuromuscular synapse, much more remains to be learned regarding the regulation of this molecule in both the CNS and peripheral nervous system. Thus, there is still much more to be gleaned from investigations at the cellular, molecular, and organismic levels, and many more applications of this work for the solution of biomedical problems.

See also: Acetylcholine Neurotransmission in CNS; Acetylcholinesterase Inhibitors and Alzheimer's Disease; Acetylcholinesterase; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Neuromuscular Junction (NMJ): Postsynaptic Basal Lamina; Neuromuscular Junction (NMJ): Postsynaptic Events in Neuromuscular Transmission.

Further Reading

Anglister L, Stiles JR, and Salpeter MM (1994) Acetylcholinesterase density and turnover number at frog neuromuscular junctions, with modeling of their role in synaptic function. *Neuron* 12: 783–794.

Couteaux R and Nachmansohn D (1938) Cholinesterase at the end-plates of voluntary muscle after nerve degeneration. *Nature* 142: 481.

Hall ZW (1973) Multiple forms of acetylcholinesterase and their distribution in endplate and non-endplate regions of rat diaphragm muscle. *Journal of Neurobiology* 4: 343–361.

Jasmin BJ, Lee RK, and Rotundo RL (1993) Compartmentalization of acetylcholinesterase mRNA and enzyme at the vertebrate neuromuscular junction. *Neuron* 11: 467–477.

Krejci E, Coussen F, Duval N, et al. (1991) Primary structure of a collagenic tail peptide of torpedo acetylcholinesterase: Co-expression with catalytic subunit induces the production of collagen-tailed forms in transfected cells. *EMBO Journal* 10: 1285–1293.

Legay C (2000) Why so many forms of acetylcholinesterase? *Microscopy Research and Technique* 49: 56–72.

Lomo T and Slater CR (1980) Control of junctional acetylcholinesterase by neural and muscular influences in the rat. *Journal of Physiology* 303: 191–202.

Lwebuga-Mukasa JS, Lappi S, and Taylor P (1976) Molecular forms of acetylcholinesterase from *Torpedo californica*: Their relationship to synaptic membranes. *Biochemistry* 15: 1425–1434.

Massoulié J and Rieger F (1969) L'Acetylcholinesterase des organes electriques de poissons (torpille et gymnote): Complexes membranaires. *European Journal of Biochemistry* 11: 441–455.

McMahan UJ, Sanes JR, and Marshall LM (1978) Cholinesterase is associated with the basal lamina at the neuromuscular junction. *Nature* 271: 172–174.

Peng HB, Xie H, Rossi SG, and Rotundo RL (1999) Acetylcholinesterase clustering at the neuromuscular junction involves perlecan and dystroglycan. *Journal of Cell Biology* 145: 911–921.

Perrier AL, Massoulié J, and Krejci E (2002) PRiMA: The membrane anchor of acetylcholinesterase in the brain. *Neuron* 33: 275–285.

Rotundo RL (2003) Expression and localization of acetylcholinesterase at the neuromuscular junction. *Journal of Neurocytology* 32: 743–766.

Rotundo RL, Rossal SG, Kimbell LM, Ruiz C, and Marrero E (2005) Targeting acetycholinesterase to the NMJ. *Chemico-Biological Interactions* 157–158(1): 15–21.

Salpeter MM, Rogers AW, Kasprzak H, and McHenry FA (1978) Acetylcholinesterase in the fast extraocular muscle of the mouse by light and electron microscope autoradiography. *Journal of Cell Biology* 78: 274–285.

Schumacher M, Camp SJ, Maulet Y, et al. (1978) Primary structure of *Torpedo californica* acetylcholinesterase deduced from Its cDNA sequence. *Nature* 319: 407–409.

Sussman JL, Harel M, Frolow F, et al. (1978) Atomic structure of acetylcholinesterase from *Torpedo californica*: A prototypic acetylcholine-binding protein. *Science* 253: 872–879.

Relevant Website

http://www.mcell.psc.edu – MCell/DReAMM environment.

Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation

T Lømo, University of Oslo, Oslo, Norway

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Motor neurons act as the 'final common path' for the entire central motor system. Only motor neurons can directly drive the skeletal muscles to produce movements, perform work, or stabilize joints. In higher vertebrates, the focus of this article, the peripheral motor system is organized into units, each motor unit consisting of a motor neuron and all of the muscle fibers that it exclusively innervates. Motor units display striking differences in the firing patterns of their motor neurons, the properties of their neuromuscular junctions (NMJs), and the contractile speed and resistance to fatigue of their muscle fibers, properties that are well matched to the different types of work that different types of motor units do.

The functional tasks of individual motor units vary throughout life – for example, with changes in body size, occupational workloads, amount and type of physical exercise, or as a consequence of aging, disease, or injury. To function efficiently under such changing conditions, the motor units must have the property of plasticity and must be capable of adaptation. Plasticity comes in two forms: one form is observed in the brain, when brief synaptic inputs lead to long-lasting changes in synaptic efficacy through processes that are important for learning and memory; the other form is observed when structures such as synapses or muscle cells respond to new conditions by adaptive changes, which last only as long as the conditions inducing the changes last. While plasticity is a striking property of NMJs and skeletal muscle cells, it is only of the adaptive kind. In this article, the properties of different types of motor units and the adaptations of their NMJs and muscle fibers to different patterns of activity are described mainly for mammals (e.g., the rat), together with some discussion of underlying mechanisms.

Motor Neurons Form Motor Units to Drive the Muscles

Motor neurons drive their muscle fibers by generating action potentials that are conducted as nerve impulses along their axons in the peripheral nerves to their terminals in the muscles, where the impulses are transmitted across NMJs to the muscle fibers. In mature mammals, only one NMJ is formed by one motor neuron on each muscle fiber. But each motor neuron has many terminal branches in the muscle and forms NMJs with tens or hundreds of muscle fibers, creating a motor unit. All of the muscle fibers in one motor unit have essentially the same properties. In rat limb muscles, the fibers express either type 1, 2A, 2X, or 2B myosin heavy chains (MyHCs), giving rise to correspondingly named fiber types which have properties that range from slow to fast, with respect to contractile speed, and from sensitive to resistant, with respect to fatigue. Reflecting such differences, motor units, including the motor neurons, are often called fast fatigable (FF), fast intermediate (FI), fast fatigue resistant (FR), or slow fatigue resistant (S). However, the concept of divisions into types of motor units often obscures the fact that many properties within one type can be quite graded and not well separated from those of another.

The Firing Patterns of Motor Neurons

The function of the NMJ of higher vertebrates is to transform every nerve impulse arriving at its presynaptic nerve terminals into a postsynaptic muscle impulse, which then travels to each end of the muscle fiber, eliciting intracellular calcium release and fiber contraction along the way. Consequently, it is possible to obtain an exact replica of the firing pattern of any motor neuron by implanting fine electrodes into a muscle and recording the action potentials from the muscle fibers that it innervates. Such single motor unit recordings have been made continuously over 24 h from soleus (SOL) and extensor digitorum longus (EDL) muscles of adult rats that move unrestrained in their cages.

Three distinct firing patterns are observed. In the SOL, which contains >95% slowly contracting, fatigue-resistant type 1 fibers, the amount of single motor unit impulse activity over 24 h is high, the trains of impulses are often very long, and the frequency of firing is relatively low (~20 Hz; Figure 1). In contrast, the EDL contains >95% rapidly contracting type II fibers, many of which are very sensitive to fatigue, although others are relatively resistant. In the EDL, two different firing patterns are revealed, each of which differs markedly from that in the SOL. One pattern, EDL-1, is characterized by relatively few impulses over 24 h and isolated episodes of either single impulses or brief, high-frequency bursts of impulses (Figures 1 and 2(a)). Typically, the first interpulse interval is particularly short (~5 ms; Figure 2(b)) so that each burst begins with an 'initial doublet'

Firing patterns	Median firing frequency (Hz)	Longest train duration (s)	Impulses per 24 h (#)	Time on per 24 h
	69–91	0.8–3.9	2600–11 200	0.5–3 min
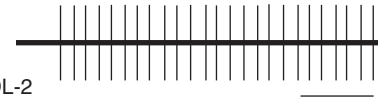	48–83	59–141	89 500–243 100	23–72 min
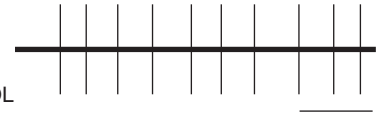	18–21	290–548	309 500–495 800	5.3–8.4 h

Figure 1 Firing patterns of motor neurons innervating extensor digitorum longus (EDL) and soleus (SOL) muscles of adult rats. The numbers are based on continuous recordings over 24 h from single motor units during spontaneous motor behavior of unrestrained adult rats in large cages. They represent a range of median values from five EDL-1, five EDL-2, and six SOL motor units. Drawn patterns to the left are based on segments of recorded activity at the onset of a burst of EDL-1 and a train of EDL-2 or SOL unit activity, as indicated. Inset shows another burst of an EDL-1 unit, as recorded. Note initial doublet and short duration of the EDL-1 bursts. The trains of EDL-2 and SOL units continued beyond the segments shown here. Adapted from Hennig R and Lømo T (1985) Firing patterns of motor units in normal rats. *Nature* 314: 164–166.

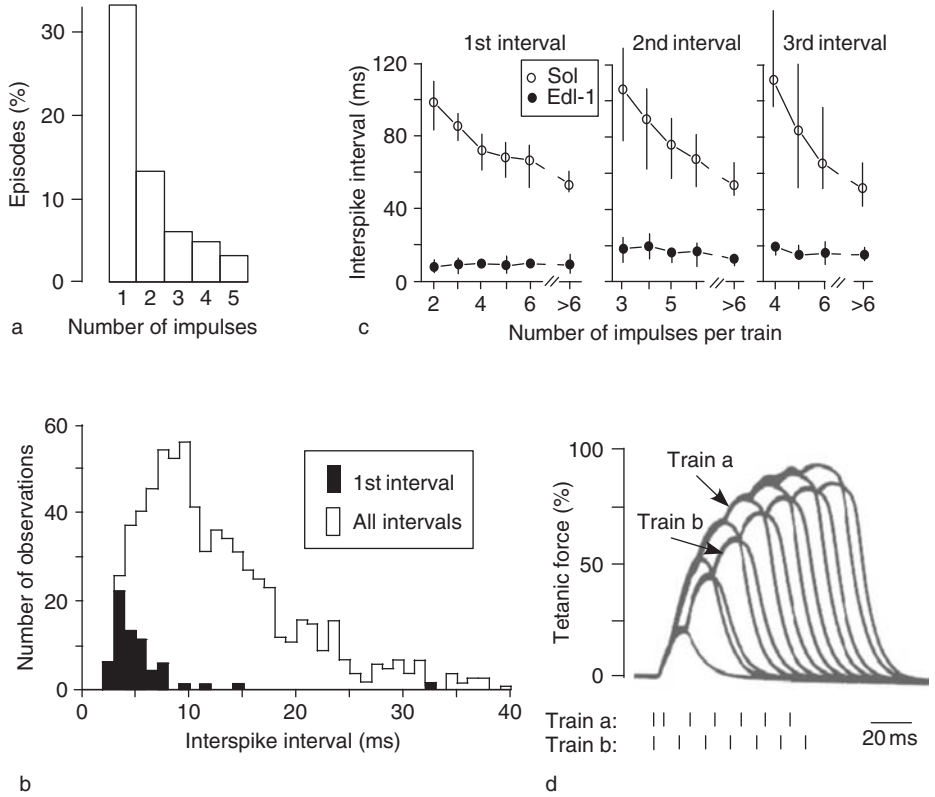

Figure 2 Firing patterns of extensor digitorum longus (EDL) and soleus (Sol) motor neurons and functional consequences. The underlying data are the same as for **Figure 1**. (a) The most common firing event for EDL-1 units over 24 h is one impulse, the next common event is two impulses, then three, and so on. (b) The first interspike interval is the shortest in bursts of impulses in EDL-1 units. (c) For EDL-1 units, interspike intervals stay the same, whereas for EDL-2 units (not shown) and Sol units, they become shorter, with increased number of impulses per train of impulses. (d) Muscle force output of EDL increases as the number of stimuli delivered to EDL axons increases from one to six. This increase in force output is further speeded up and augmented by making the first pulse interval short (~5 ms, initial doublet). (a–d) From Hennig R and Lømo T (1985) Firing patterns of motor units in normal rats. *Nature* 314: 164–166; and Hennig R and Lømo T (1987) Gradation of force output in normal fast and slow muscles of the rat. *Acta Physiologica Scandinavica* 130: 133–142, with permission from Blackwell Publishing.

(see later). Most likely, this pattern reflects the activity of motor neurons innervating fast glycolytic type 2B fibers to form fast-fatigable motor units. The other pattern, EDL-2, also consists of relatively high-frequency impulses, but the amount of activity over 24 h is greater, the trains last longer, and the median frequency is lower (Figure 1). Most likely, this pattern reflects the activity of motor neurons innervating type 2A fibers to form fast fatigue-resistant motor units. However, the two categories, EDL-1 and EDL-2, might also include intermediate units containing type 2X fibers.

Summation of all the times that an individual unit is active over 24 h reveals that SOL units are on duty up to 8 h day^{-1}, whereas the corresponding duty time for EDL-1 units is only about 1 min (Figure 1). The most common episode of EDL-1 activity is a single isolated impulse; two impulses are less common, three impulses still less common, and so on (Figure 2(a)). Together, isolated episodes of one to five impulses of EDL-1 activity make up ~60% of all episodes of EDL-1 activity over 24 h (Figure 2(a)). Moreover, only EDL-1 units display initial doublets. EDL-2 and SOL units frequently generate single impulses or brief impulse trains as well as long episodes of tonic activity. Thus, in contrast to EDL-1 units, EDL-2 and SOL units fire both phasically and tonically, but given their long durations, the periods of tonic activity are responsible for more than 90% of the total amount of activity over 24 h. During tonic activity at around 20 Hz, SOL units generate more than 80% of maximal tetanic tension owing to their long contraction times and effective twitch summation at this frequency. Differences in firing pattern, such as those just described, pose different demands on motor units to which both their NMJs and muscle fibers must adapt.

Plateau Potentials Generate Tonic Firing

Tonic firing in motor neurons does not require continuous excitatory synaptic drive. A brief excitatory synaptic input can start it and, at different times afterwards, a brief inhibitory synaptic input can stop it. The excitatory input activates a self-sustained persistent inward current carried mainly by calcium ions. This generates a plateau potential, which, in turn, elicits the tonic firing. In rat SOL motor neurons, such firing occurs at a remarkably stable rate of about 20 Hz. The generation of self-sustained plateau potentials in motor neurons depends on and is controlled by monoaminergic inputs descending from the brain stem, particularly serotonergic fibers acting on serotonin (5-hydroxytryptamine; $5HT_{2A}$) receptors ($5HT_{2A}Rs$) on cell bodies and dendrites of the motor neurons. Rat SOL lumbar motor neurons begin to display detectable $5HT_{2A}Rs$ and to fire tonically in the second week after birth. Thereafter, both the intensity of $5HT_{2A}R$ labeling and the amount of tonic firing increase gradually as the rat's legs support an increasing body weight. During this period, a substantial number of the SOL muscle fibers are transformed from a phenotype of fast type 2 fibers soon after birth to that of slow type 1 fibers in the adult, reflecting adaptation to the postnatal increase in weight bearing and accompanying postural tonic activity (see later).

There are apparent advantages in controlling tonic firing in this way. With a mechanism for turning on or off self-sustained tonic firing, more costly sustained synaptic drives can be dispensed with. The mechanism favors steady and predictable firing suited for supporting the relatively constant load of the body weight. As demonstrated experimentally, it also allows for cycling of activity between motor units from one tonic period to another, in contrast to the fixed order of recruitment and derecruitment of motor units occurring during voluntary contractions against increasing or decreasing loads, also known as the 'size principle.' In addition, we have unpublished evidence that the tonic activity described here can be used independently of motor control to produce heat to fine-tune body temperature around its set point when ambient temperature varies.

Motor Neuron Firing Regulates Muscle Force Output

Regulation of muscle force output by motor neurons is commonly presented as relying simply on motor unit recruitment and firing rate modulation. However, given the actual firing patterns during natural spontaneous motor behavior as described for the rat, this classic view needs to be expanded. Two results stand out: the predominance of single episodes of one or a few impulses and, for EDL-1 units, the constancy of interpulse intervals regardless of the number of impulses in the trains (Figure 2(c)). Consequently, the force output of EDL-1 units is regulated not by modulating the firing rate but by varying the number of impulses in the burst and hence the amount of twitch summation obtained (Figure 2(d)). Thus, it is the duration and not the intensity of the excitatory synaptic drive on the motor neuron that appears to be regulated. Similar regulation by impulse number has been described for ballistic movements in humans. A further consequence of the firing pattern of EDL-1 units is that the presence of initial doublets speeds up the contraction by exploiting a catchlike property in the muscle fibers whereby force output develops at maximum speed and then remains at a higher level than in the absence of such doublets (Figure 2(d)).

Comparable regulation by impulse number also occurs for EDL-2 and SOL units during brief episodes of activity, but for these units the pulse interval decreases with increasing number of impulses per train (Figure 2(c)). Presumably, this gradual decrease reflects an increasing excitatory synaptic drive on the motor neurons. Accordingly, force regulation for such units occurs by a combination of variable impulse number (for short episodes of activity) and variable excitatory synaptic drive on the motor neurons. In addition, of course, there is regulation by motor unit recruitment for all unit types.

Differences in motor neuron firing pattern thus play an important role in the execution of diverse motor tasks. This conclusion is reinforced by experiments showing that muscle fibers and NMJs continually adapt their properties to the pattern of activity.

The Neuromuscular Junction

To transform every nerve impulse into a muscle impulse, transmission across vertebrate NMJs must be highly efficient and reliable. Many specializations at the NMJ contribute to this efficiency. For example, each nerve impulse releases tens of quanta of the transmitter acetylcholine (ACh); these act on a high number and density of ACh receptors (AChRs) in the postsynaptic membrane of the muscle fiber. The postsynaptic membrane is generally highly infolded, which increases the efficiency of transmission. In the depths of these folds, voltage-sensitive sodium channels are present at high density, which further facilitates the generation of muscle impulses. There is no afterhyperpolarization (AHP) following each muscle impulse and no persistence of released ACh because of highly active ACh esterase (AChE) in the synaptic cleft. As a result, the membrane potential returns quickly to its resting level after an impulse and is ready to effectively transmit the next impulse even at the high frequencies that characterize FF motor units.

Motor neurons induce formation of the NMJs during development by releasing neural agrin and acetylcholine from their nerve terminals in the muscles. Neural agrin causes the aggregation of AChRs and other synapse-specific proteins underneath the terminals and ensures the development and maintenance of the NMJ by rendering the synthesis and organization of synapse-specific proteins resistant to the inhibitory effects of nerve-evoked muscle impulse activity. Such evoked activity then regulates the size and other properties of the developing NMJ, while blocking the expression of AChRs along the rest of the muscle fiber and preventing synapse formation there.

In the adult, numerous features affect the efficacy of transmission at NMJs, among them the number and lifetime of AChRs in the postsynaptic membrane. After inactivation of NMJs by denervation or chronic applications of tetrodotoxin (TTX) or botulinum toxin, the postjunctional membrane gradually loses more and more AChRs, while those that remain, or are subsequently inserted, turn over much more rapidly than before. Direct stimulation of the denervated muscle prevents these effects, but only if the stimulation starts within the first day after denervation. In that case, the number and stability of postjunctional AChRs stay normal for at least 2 months. If the stimulation starts progressively later, up to 2 weeks after denervation, its effectiveness gradually declines so that muscles stimulated from the end of that time go on to lose as many receptors as unstimulated muscles do. These results suggest that muscle impulse activity is necessary to preserve a nerve-derived 'trace' at the denervated NMJ that influences the number and lifetime of AChRs in the postsynaptic membrane. This trace may be neural agrin itself or any of its downstream postsynaptic effects, since neural agrin both clusters AChRs and regulates their lifetime in the membrane.

NMJs Differ in Structure and Function

NMJs on fast and slow muscle fibers, such as those on the fast EDL and slow SOL of the rat, differ both in structure and function. Presynaptically, fast nerve endings and slow nerve endings arborize differently, and fast nerve endings contain more terminal swellings (varicosities). The fast endings release more acetylcholine at the onset of impulse trains, but since they have smaller vesicle pools, they fatigue more easily as release runs down faster during continued stimulation. Postsynaptically, relative to slow NMJs, fast NMJs have denser, deeper, and more branched infoldings. They also respond to given doses of ACh with larger postsynaptic currents. In their synaptic clefts, fast NMJs contain predominantly the large A12 form of AChE, whereas slow NMJs contain more of the smaller A8 and A4 forms. The AChE molecules are anchored to the synaptic basal lamina by one of two types of collagen-tailed proteins encoded by different transcripts from the same *ColQ* gene in fast and slow muscle fibers. Two distinct promoters for the *ColQ* gene have been identified; these are differentially expressed in fast and slow muscle fibers. The promoter present in slow fibers is activated by the transcription factor called nuclear factor of activated T cells (NFAT) at a 'slow upstream regulatory element' (SURE), whereas the promoter expressed in fast fibers contains a 'fast intronic

regulatory element' (FIRE). Assuming that different compositions of AChE isoforms reflect differences in function, such findings suggest that the fast and slow impulse patterns transmitted by fast and slow NMJs pose different demands on AChE function.

Plasticity of NMJs

The appearance of individual NMJs in adult mice or rats is remarkably stable when observed repeatedly over weeks or months *in vivo*. But this is probably as it should be in animals that live under relatively constant animal housing conditions. Whenever such conditions change – for example, with the seasons in lower vertebrates or with hormonal interventions, exercise, muscle unloading (as in space flights or hind limb suspension), or any other manipulation that alters the impulse activity in motor units or the size of the muscle fibers – signs of synaptic remodeling appear. For example, procedures that cause muscle fibers to increase or decrease in size also cause the NMJs on them to increase or decrease in extent, like an imprint on a balloon that inflates or deflates. But changes in nerve terminal dimensions and branching patterns may also occur independently of muscle fiber size. Moreover, such remodeling is reversible, as expected from purely adaptive processes.

Striking functional changes have been observed at NMJs in adult rat EDL and SOL muscles subjected to altered patterns of synaptic inputs. In these experiments, complete control of the amount and pattern of impulses arriving at the NMJs was obtained by applying TTX locally to the sciatic nerve for 3–4 weeks and stimulating the axons to EDL or SOL with slow (20 Hz for 20 s every 30 s) or fast (150 Hz for 1 s every 60 s) stimulus patterns through electrodes implanted below the TTX block. In this way, only the applied slow pattern of impulses arrived at the fast NMJs in the EDL, and only the applied fast pattern of impulses arrived at the slow NMJs in the SOL. At the end of such treatments, run down of end plate potential amplitude was much less than normal in EDL and much greater than normal in SOL, reflecting >70% transformation of sensitivity to fatigue in either direction (**Figures 3(a) and 3(b)**). In addition, the ratio of quantal content to total vesicle pool size became much lower in EDL than in SOL, whereas normally it is much higher in EDL than in SOL (**Figures 3(c) and 3(d)**). Vesicle recycle time was not affected by the chronic stimulation, nor was any effect expected since NMJs in normal EDL and SOL are similar in this respect.

Similar transformations between fast and slow phenotypes occur with respect to AChE. Thus, fast and slow patterns of stimulation affect AChE expression differentially in rat SOL and EDL muscles, as the distribution of AChE isoforms becomes more like that of fast NMJs under the influence of fast pattern stimulation and more like that of slow NMJs under the influence of slow pattern stimulation. Moreover, since these stimulations were applied to denervated muscles, the different forms of AChE almost certainly derived from the postsynaptic muscle fibers and, presumably, as a consequence of muscle impulse pattern-dependent activation of SURE or FIRE promoter elements of the *ColQ* gene (as discussed previously). Consistent with this interpretation, similar SURE and FIRE elements have been described for several genes encoding contractile proteins; these are differentially regulated by slow or fast activity patterns. Therefore, adaptation to different motor unit firing patterns appears to occur by differential postsynaptic gene expressions affecting both synaptic and contractile properties.

Mechanisms of Plasticity

The firing patterns of motor neurons play a dominant role in the regulation of plastic changes at NMJs. Such changes may be initiated in the cell body of the motor neuron, at the NMJ itself, or in the muscle fiber. An example of initiation in the cell body comes from crayfish, where chronic slow pattern stimulation of the fast axon central to a local TTX block makes the NMJs distal to the block slow through mechanisms that are blocked by local application of calcium channel blockers to the cell body. An example of initiation in the muscle comes from cats, in which chronic fast pattern stimulation of the sciatic nerve distal to a TTX block shortens the afterhyperpolarization after each action potential in SOL motor neurons, an effect that is not observed after chronic slow pattern stimulation. A reasonable interpretation of this result is that the fast stimulation transformed the SOL muscle into a fast muscle (as described later), and that this switch in phenotype gave rise to factors that fed back across the NMJs and retrogradely shortened the afterhyperpolarization in the SOL motor neurons, with the result that the SOL motor neurons would now fire at higher than normal frequencies. In this way, the firing characteristics of the SOL motor neurons would become better matched to their newly acquired rapidly contracting muscle fibers for rate modulation of muscle force output, while their NMJs would adapt to the faster impulse patterns. At the level of the NMJ, there is evidence that growth factors and other signaling molecules released locally from nerve terminals and

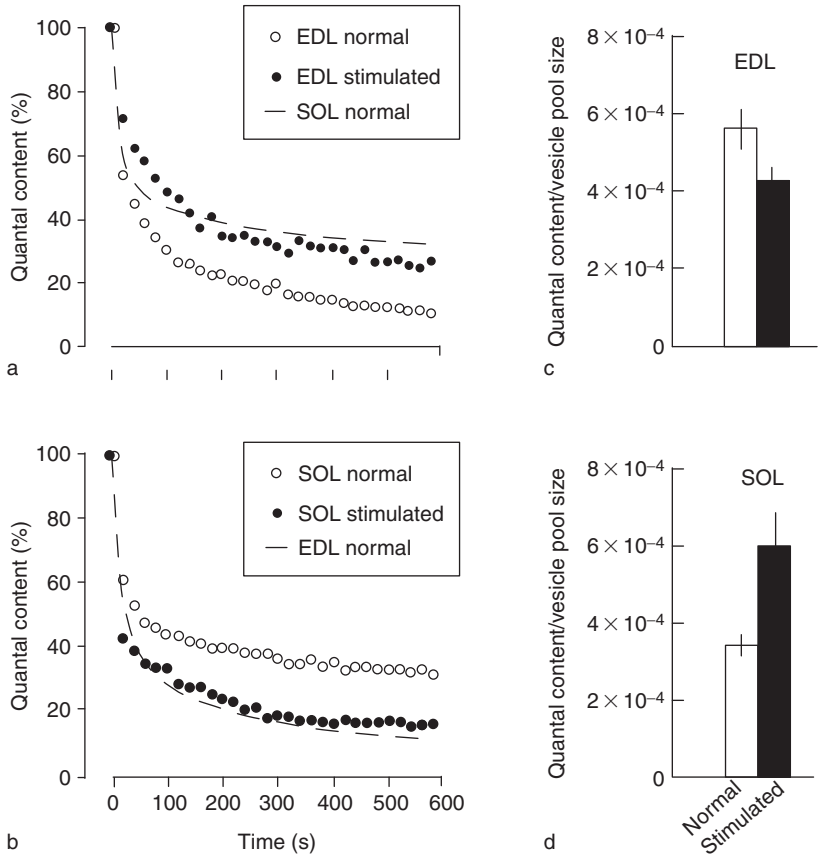

Figure 3 Impulse pattern-dependent fast-to-slow and slow-to-fast transformations of transmitter release characteristics. (a) Decrease in quantal content during 10 min of 20 Hz acute nerve stimulation is fast at normal extensor digitorum longus (EDL) neuromuscular junctions but becomes slower after 3–4 weeks of chronic slow pattern stimulation of EDL axons (20 Hz for 20 s every 30 s). (b) Conversely, the decrease in quantal content is slow at normal SOL neuromuscular junctions but becomes faster after chronic fast pattern stimulation of SOL axons (150 Hz for 1 s every 60 s). (c, d) Quantal content is relatively large and total vesicle pool size is relatively small at normal EDL neuromuscular junctions, whereas the opposite is true at normal SOL neuromuscular junctions. Therefore, the ratio quantal content:vesicle pool size is relatively large at normal EDL neuromuscular junctions (c) and relatively small at normal SOL neuromuscular junctions (d). Following chronic slow pattern stimulation, EDL neuromuscular junctions become like normal slow SOL neuromuscular junctions in this respect (c), and following chronic fast pattern stimulation SOL neuromuscular junctions become more like normal EDL NMJs (d). Adapted from Reid B, Martinov VN, Njå A, et al. (2003) Activity-dependent plasticity of transmitter release from nerve terminals in rat fast and muscles. *Journal of Neuroscience* 23: 9340–9348.

muscle fibers, as well as terminal Schwann cells, can affect the properties of NMJs.

The Muscle Fibers

The work muscle fibers do varies. Fibers belonging to FF motor units bring about brief, rapid, and forceful contractions that are relatively rare, partly because their motor neurons have high thresholds of activation. In contrast, fibers belonging to S motor units are usually tonically active over long periods of time, as they are frequently used to stabilize joints or counteract gravity. Consequently, muscle fibers must be specialized – for example, for high speed and force output or high resistance to fatigue. But such specializations come at a price, as the fastest units are also the least fatigue resistant, and the most fatigue-resistant units are also the slowest. Many properties of motor units may therefore be placed on a scale from the slowest and most fatigue-resistant units to the fastest and most fatigue-sensitive units.

Specializations for different tasks are revealed at many levels and in interesting ways. For example, in tonically active, fatigue-resistant postural muscles such as the SOL, numerous capillaries supply the oxygen that the numerous mitochondria and well-developed oxidative enzyme systems in such muscles need to produce the necessary ATP for sustained work. At another level, selective expression of slow types of myosins and a slow transporter of calcium ions into sarcoplasmic reticulum conserves energy by prolonging the contractions, slowing the breakdown of ATP, and decreasing the utilization of ATP per unit work. In contrast, in fast fatigue-sensitive muscles,

such as the rat EDL, the blood perfusion rate is much lower and energy mobilization is much faster, owing to rapid breakdown of ATP by glycolysis. Selective expression of fast types of myosin speeds up ATP utilization and sarcomere shortening, while selective expression of a fast transporter for bringing calcium back into the sarcoplasmic reticulum further speeds up the twitch so that it can match the high firing rates of their motor neurons.

Plasticity of Muscle Fibers

Muscle fibers are plastic. They adapt their properties to altered use as dictated by the firing patterns of their motor neurons. Whether motor nerve terminals also release so-called trophic factors to regulate contractile and other nonjunctional properties of muscle has been the subject of numerous studies and much debate. So far, however, no one has been able to identify such factors. Moreover, results that appear to support such a mechanism have other and more plausible interpretations. Therefore, in the continued absence of direct and incontrovertible evidence for trophic control of this kind, this particular hypothesis can be laid to rest. While muscle load and hormones, such as thyroxin, may clearly modulate the expression of contractile proteins, it is the firing patterns of the motor neurons that play the overriding role in the day-to-day regulation of the properties of muscle fibers and NMJs.

Adaptation to Fast and Slow Impulse Patterns

Different exercise regimes may induce marked changes in muscle strength, endurance, and, to a lesser degree, in fiber types. Mechanisms underlying such changes are revealed in muscles that have been stimulated over days and weeks with stimulus patterns that mimic to some extent the natural firing patterns of fast and slow motor neurons (described earlier). Muscles have been stimulated directly after denervation, in which case the imposed pattern is exactly known, or indirectly through the nerve, in which case an unknown pattern of endogenous activity from the motor neuron is added to the imposed pattern. The general outcome of such studies is that increasing amounts of activity cause fast fibers to become more fatigue resistant as more capillaries appear around them and they acquire more mitochondria and well-developed oxidative enzyme systems. They also become slower contracting as the expression of slow proteins is turned on and that of fast proteins suppressed. When the stimulation is interrupted, the muscle fibers return to their previous state, as expected from purely adaptive processes. Conversely, slow fibers become faster and less fatigue resistant during stimulation with a fast stimulus pattern.

Matching Impulse Patterns and Contraction Speed

In denervated rat slow SOL muscles, 3 weeks of direct fast pattern stimulation is sufficient to make the twitch almost as fast as that of the normal EDL, while slow pattern stimulation maintains the normal slow twitch (Figure 4(a)). Such impulse pattern-dependent changes in twitch speed have interesting functional implications. As the impulse frequency increases, the SOL twitch becomes faster (shorter in duration) (Figure 4(c)). Higher impulse frequencies are therefore needed for twitch summation. As a result, the steep part of the tension frequency curve moves toward the imposed frequency (Figure 4(b)) and the SOL twitch becomes better matched to the imposed frequency for frequency gradation of force output. But in SOL, the amount of impulse activity also affects twitch speed, as increased amounts slow down, and decreased amounts speed up, the twitch (Figure 4(c)). In EDL and other fast muscles, an altered amount rather than altered frequency of impulses affects twitch speed. The fastest fibers normally receive low amounts of activity at high frequency and are as fast as they can become. Any change in twitch speed will therefore be in the slow direction. Nevertheless, it remains an open question whether changes in impulse frequency contribute to the fine-tuning of contractile properties among the different types of fast units.

Adaptation within Adaptive Ranges

It has not been possible by cross-reinnervation or by electrical stimulation of adult rat EDL and SOL muscles by fast or slow stimulus patterns, either indirectly through the nerve or directly after denervation, to completely transform the properties of either muscle into those of the other. Figure 4(d) illustrates schematically results obtained after ~2 months of such stimulations. Under fast pattern stimulation, SOL muscles become only about half as fast as the normal EDL with respect to intrinsic shortening velocity (V_{max}), which is consistent with the finding that such transformed SOL muscles express fast types 2A and 2X MyHCs but not type 2B. Similarly, under slow pattern stimulation, the EDL twitch becomes as slow as that of slow type 1 muscle fibers in EDL but only about half as slow as that of slow type 1 fibers in normal SOL muscles. Such changes occur within the adaptive range of a particular property in a particular muscle, as illustrated by the arrows in Figure 4(d). Within such a range the changes are graded and, as shown in Figure 4(c) for the rat SOL, depend on both the rate and amount of imposed impulse activity. However, other features of motor

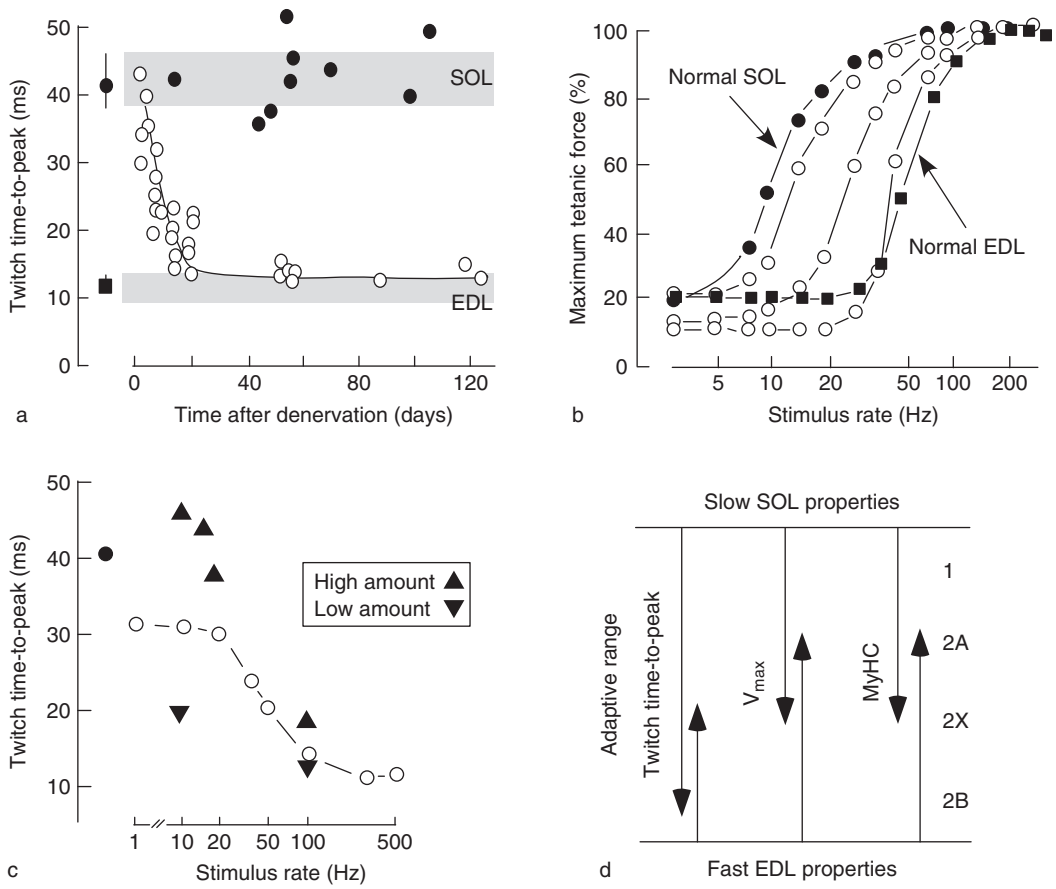

Figure 4 Adaptive range of muscle fiber properties. (a) The twitch of denervated soleus (SOL) muscles becomes almost as fast as that of extensor digitorum longus (EDL) muscles during direct fast pattern stimulation (60 pulses at 100 Hz every 60 s; ○), but remains slow during slow pattern stimulation (150 pulses at 15 Hz every 15 s; ●). (b) The SOL was denervated and stimulated directly with the fast pattern for 7, 14, or 130 days (○). The steep part of the tension–frequency curve of the normal SOL (●) moved progressively toward that of the normal EDL (■). (c) Denervated SOL muscles were stimulated directly for 28–106 days (mean 55 days), with 60 pulses at the frequencies indicated along the abscissa every 60 s (same number of pulses, mean frequency 1 Hz). For each increase in frequency, the twitch speed (ordinate) increased (○). At some frequencies (10, 15, 20, or 100 Hz), the amount of activity was increased (▲) or decreased (▼) by increasing or decreasing the number of pulse trains, which resulted in slower or faster twitches. (d) Scheme to illustrate that EDL and SOL respond differently to fast and slow impulse patterns, as these patterns fail to completely transform several properties of one muscle into those of another. Although overlapping, the adaptive ranges of diverse contractile properties are different in adult rat EDL and SOL muscles owing to intrinsic differences between them. The present scheme is based on results obtained after ~2 months of chronic stimulations. Longer periods of stimulation may expand the ranges shown here. (a, b) Adapted from Gorza L, Gundersen K, Lømo T, et al. (1988) Slow-to fast-transformation of denervated soleus muscles by chronic high-frequency stimulation in the rat. *Journal of Physiology* 402: 627–649. (c) From Westgaard RH and Lømo T (1988) Control of contractile properties within adaptive ranges by patterns of impulse activity in the rat. *Journal of Neuroscience* 8: 4415–4426.

neurons' firing patterns, such as initial doublets, also affect the position along its adaptive range that a given property may attain. Interestingly, the differences in firing pattern achievable in experimental settings, and the resulting adaptations of the muscles, are much more extreme than can be achieved voluntarily in an intact animal or human.

That muscles such as rat EDL and SOL muscles respond so differently to a given impulse pattern and display such different adaptive ranges indicates the existence of intrinsic differences between these muscles. Recently, intrinsic differences between the satellite cells in EDL and SOL that give rise to regenerated fibers after injury have also been reported. Nevertheless, the capacity for adaptation in diverse muscle fibers is substantial, so that their properties may be adjusted in a graded manner to the prevailing firing patterns of their motor neurons and the types of work that are imposed on them.

Cellular Mechanisms of Adaptation

By what mechanisms do fast and slow impulse patterns turn on or off fast and slow gene programs in muscle fibers? Probably, they involve calcium

signaling, because fast and slow impulse patterns raise free intracellular calcium concentration differently with respect to rate, magnitude, and duration and because such factors differentially affect gene expression in other cells. While it is still unclear how time- and space-dependent changes in free calcium levels work in muscle cells, their downstream effects are being clarified. For example, only the slow impulse pattern activates the pathway calcium/calmodulin → calcineurin → NFAT, which causes depophosphorylated NFAT to enter the muscle nuclei and activate genes for many different slow contractile proteins, and presumably, the forms of junctional AChE that dominate at slow NMJs (see earlier). This entry into the muscle nuclei occurs within minutes of starting slow pattern stimulation, as recently shown *in vivo* in muscle fibers transiently expressing an isoform of NFAT (NFAT c1) that had been coupled to green fluorescent protein (GFP) to make NFAT and its movements in the cell visible. Conversely, upon cessation of stimulation, NFAT moves equally rapidly out of the nuclei. For NFAT transcriptional activity to be effective, however, not only nuclear translocation but also other conditions must be met.

A different signaling pathway includes the sequence phosphoinositide 3-kinase (PI3K) → protein kinase B (also known as AKT) → mammalian target of rapamycin (mTOR). Activation of this pathway, which can occur by innervation, muscle stimulation, or insulin-like growth factor 1 (IGF-1), regulates muscle size but not fiber type. Recent studies have thus revealed several intracellular signaling pathways that independently or in combination regulate the phenotype of muscle fibers.

Studies have shown that muscle fibers adapt to virtually any change in the pattern of impulse activity imposed on them, whether it is the number or frequency of pulses, the duration or spacing of pulse trains, or the presence or absence of initial doublets. Therefore, it appears that muscle fibers continually fine-tune their contractile properties to the impulse patterns imposed on them at any time, and that this tuning is mediated by intracellular signaling pathways that are sensitive to subtle changes in these patterns.

Conclusions

Motor neurons to leg muscles in adult rats generate markedly different impulse patterns. Such firing patterns reflect the different types of work that motor neurons evoke from the muscle fibers in their motor units. The properties of NMJs and muscle fibers in motor units are well matched to the firing patterns of their motor neurons for effective neuromuscular transmission and muscle work. Thus, NMJs and muscle fibers belonging to fast motor units are specialized for effective transmission of brief, high-frequency bursts of impulses and rapid, forceful contractions at the cost of fatigue resistance, while NMJs and muscle fibers belonging to slow motor units are specialized for effective transmission of prolonged trains of impulses and slow, fatigue-resistant contractions.

NMJs and muscle fibers display substantial plasticity when new impulse patterns are imposed on them. This plasticity is of the adaptive kind. It lets NMJs and muscle fibers fine-tune their properties to effectively transmit impulses and do muscle work under changing circumstances. The adaptations occur within adaptive ranges that may overlap but are intrinsically different in different types of motor units. Intracellular signaling pathways that transform the different impulse patterns into particular gene expression patterns and phenotypes are presently being identified and clarified.

Much greater changes in muscle speed can be obtained by electrical stimulation of muscles using fast or slow pulse patterns than by physical training, presumably because the firing pattern of individual motor neurons in the intact organism cannot be switched to similar degrees. Thus, there seems to be a large potential for adaptation in muscle that can be tapped as more extreme forms of training and motivation expand the range of firing patterns that individual motor neurons are capable of. Perhaps this is one reason why athletes continue to set new world records.

See also: Motor Control of Feeding and Drinking; Motor Neuron Specification in Vertebrates; Neuromuscular Junction (NMJ): Aging; Neuromuscular Junction (NMJ): A Target for Natural and Environmental Toxins in Humans; Presynaptic Events in Neuromuscular Transmission; Schwann Cells and Plasticity of the Neuromuscular Junction.

Further Reading

Eken T and Gundersen K (1988) Electrical stimulation resembling normal motor-unit activity: Effects on denervated fast and slow rat muscles. *Journal of Physiology* 402: 651–669.

Eken T and Kiehn O (1989) Bistable firing properties of soleus motor units in unrestrained rats. *Acta Physiologica Scandinavica* 136: 383–394.

Gorza L, Gundersen K, Lømo T, et al. (1988) Slow-to fast-transformation of denervated soleus muscles by chronic high-frequency stimulation in the rat. *Journal of Physiology* 402: 627–649.

Hennig R and Lømo T (1985) Firing patterns of motor units in normal rats. *Nature* 314: 164–166.

Hennig R and Lømo T (1987) Gradation of force output in normal fast and slow muscles of the rat. *Acta Physiologica Scandinavica* 130: 133–142.

Kernell D (2006) *The Motoneurone and Its Muscle Fibres*. Oxford: Oxford University Press.

Ko C-P and Thompson WJ (eds.) (2003) The neuromuscular junction, a special issue in tribute to B Katz. *Journal of Neurocytology* 32: 421–1037.

Lee HH, Choi RC, Ting AK, et al. (2004) ColQ: Differential expression in fast and slow muscle fibers is driven by distinct promoters. *Journal of Biological Chemistry* 279: 27098–27107.

Lømo T (2003) Clinical neurophysiology of disorders of muscle and neuromuscular junction, including fatigue. In: Daube JR and Mauguière F (eds.) *Handbook of Clinical Neurophysiology,* vol. 2, pp. 47–65. Amsterdam: Elsevier.

Pette D (2002) The adaptive potential of skeletal muscle fibers. *Canadian Journal of Applied Physiology* 27: 423–448.

Reid B, Martinov VN, Njå A, et al. (2003) Activity-dependent plasticity of transmitter release from nerve terminals in rat fast and slow muscles. *Journal of Neuroscience* 23: 9340–9348.

Schiaffino S and Reggiani C (1996) Molecular diversity of myofibrillar proteins: Gene regulation and functional significance. *Physiological Reviews* 76: 371–423.

Tothova J, Blaauw B, Pallafacchina G, et al. (2006) NFATc1 nucleocytoplasmic shuttling is controlled by nerve activity in skeletal muscle. *Journal of Cell Science* 119: 1604–1611.

Westgaard RH and Lømo T (1988) Control of contractile properties within adaptive ranges by patterns of impulse activity in the rat. *Journal of Neuroscience* 8: 4415–4426.

Wood SJ and Slater CR (2001) Safety factor at the neuromuscular junction. *Progress in Neurobiology* 64: 393–429.

Neuromuscular Junction (NMJ): Aging

G C Sieck and C B Mantilla, Mayo Clinic College of Medicine, Rochester, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

The neuromuscular junction (NMJ) is the final effector for neural control of muscle contraction. Mammalian NMJs serve as the sole communicative link between a motor neuron and the muscle fibers it innervates, which together constitute a motor unit (Figure 1). Mechanical properties of motor units are critically important in determining the functional performance of a skeletal muscle across a range of physiological behaviors. In fact, motor units exhibit considerable diversity in terms of their morphological, mechanical, and fatigue properties, and recruitment of specific types of motor units is a major mechanism in neural control of muscle force generation. Despite the diversity among motor unit types, there is remarkable homogeneity in the mechanical and biochemical properties of muscle fibers comprising an individual motor unit. Thus, the properties of muscle fibers and motor neurons appear to be precisely matched to facilitate neuromotor control.

Synaptic plasticity is a hallmark of the nervous system's adaptive response to both intrinsic and extrinsic stimuli and forms the basis for changes in the efficacy of synaptic transmission. With aging, there are a number of changes that could affect plasticity of the NMJ: (1) neuromuscular activity is generally reduced with concomitant unloading of limb muscle fibers; (2) the number of motor neurons is reduced, resulting in denervation of some muscle fibers, which are subsequently reinnervated by the sprouting of axons from remaining motor neurons; and (3) muscle fiber size is reduced in a fiber-type-dependent manner. In the late 1940s, Donald Hebb introduced the concept of activity-dependent synaptic plasticity, which addresses the structural and functional changes that occur at synapses in response to altered use (either an increase or a decrease in activity). According to the Hebbian theory, synaptic efficacy is enhanced when the extent of correlation between pre- and postsynaptic activity increases. With an age-related decrease in neuromuscular activity, synaptic plasticity may occur, but only if there are alterations in the fidelity of neuromuscular transmission (i.e., the extent of correlation between pre- and postsynaptic activity). Clearly, conditions other than changes in activity can also drive synaptic plasticity. For example, with aging there is a loss of motor neurons, which leads to denervation. Following denervation, axonal regeneration and reinnervation of muscle fibers can occur, but this is affected by activity and trophic influences. With aging, there is also a decrease in muscle fiber size (i.e., age-related sarcopenia). A number of studies have reported changes in NMJ morphology and neuromuscular transmission in conditions in which muscle fiber size either increases or decreases, which may relate to common trophic influences. Each of these conditions, in which long-term plasticity of the NMJ occurs, varies with motor unit and/or muscle fiber type.

Although motor unit heterogeneity in mammalian skeletal muscles reflects a continuum of mechanical and biochemical properties of muscle fibers, motor units are generally categorized into different types based on mechanical and fatigue properties of motor unit fibers. Accordingly, motor units are classified into four types: (1) slow-twitch, fatigue-resistant (type S); (2) fast-twitch, fatigue-resistant (type FR); (3) fast-twitch, fatigue-intermediate (type FInt); and (4) fast-twitch, fatigable (type FF) (Figure 2). Muscle fiber type classification follows a corresponding scheme, whether based on histochemistry or myosin heavy chain (MHC) isoform expression. Thus, muscle fibers are classified into four types: type I (MHC_{Slow}), type IIa (MHC_{2A}), type IIx (MHC_{2X}), and type IIb (MHC_{2B}). The innervation ratio (i.e., the number of muscle fibers innervated by a motor neuron) varies across muscles, ranging from 10 or less in small intrinsic hand and eye muscles to several hundred in larger limb and postural muscles. Within a muscle, the innervation ratio is greater at type FInt and FF motor units compared to type S and FR units. In muscles of mixed fiber type composition, differences in fiber size also exist across motor unit types, with type IIx and/or IIb fibers having greater cross-sectional area than type I or IIa fibers. Across muscles the cross-sectional areas of a given fiber type may vary. For example, type I fibers in the rat soleus muscle are much larger than type I fibers in muscles with mixed fiber type composition such as the diaphragm. Although controversial, some studies suggest that the specific force (i.e., force per unit cross-sectional area) of type IIx and/or IIb fibers is greater than that of type I and IIa fibers. Together, the greater innervation ratio, larger fiber size, and greater specific force contribute to the greater forces generated by type FInt and FF motor units compared to type S and FR units. The recruitment of motor units is generally matched to these mechanical and fatigue properties, such that type S and FR motor units are recruited first, followed by type FInt and FF units. In fact, forces generated during most sustained motor behaviors

require the recruitment of only type S and FR motor units, and only during high-force, short-duration motor behaviors would it be necessary to recruit more fatigable FInt and FF units.

Aging Effects on Motor Units

The basis for classification of motor unit types does not change with aging. However, age-related changes can affect the proportion of motor unit types, their innervation ratio, and their mechanical properties. For example, with aging there is a decrease in the number of motor neurons innervating skeletal muscles, and this loss of motor neurons appears to be selective for FInt and FF motor unit types. Thus, the proportion of these motor unit types decreases with age. It is possible that the selective loss of more fatigable motor units reflects age-related changes in neuromuscular activity, but this has not been clearly established. With the loss of motor neurons, some type IIx and/or IIb fibers are denervated. As a result, remaining motor axons sprout and reinnervate these muscle fibers, leading to an increase in innervation ratio of the remaining motor units. In addition, the size of type IIx and/or IIb fibers decreases with age, possibly the result of motor neuron loss and denervation, decreased neuromuscular activity, unloading, or other underlying mechanisms. Finally, specific force also decreases with age, being more pronounced at type IIx and/or IIb fibers. Together, these age-related changes tend to decrease the diversity among motor unit types, with a lower proportion of FInt and FF motor unit types that generate greater forces. The age-related decrease in the force generated by FInt and FF motor units (due to decreased fiber cross-sectional area and specific force) may be partially offset by a concomitant increase in the force generated by type S and FR motor units (due to an increased innervation ratio). Certainly, these changes

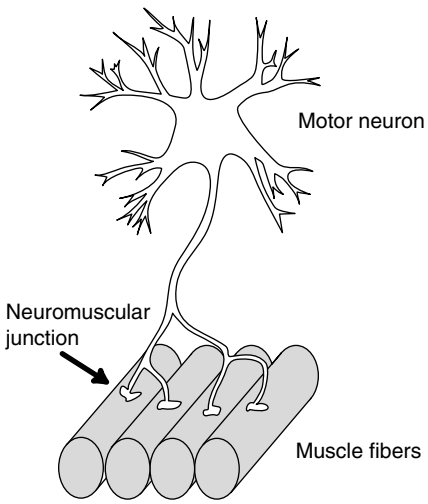

Figure 1 Schematic representation of the elements of a motor unit. The neuromuscular junction (NMJ) serves as the sole communicative link between the motor neuron and the muscle fibers it innervates.

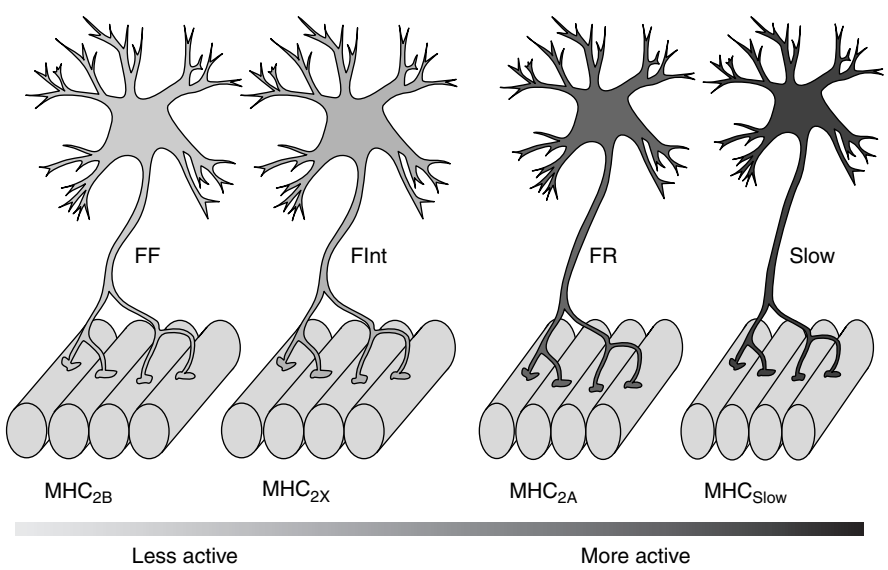

Figure 2 Motor units are classified based on their mechanical and fatigue properties. Four types are commonly described: (1) slow-twitch, fatigue-resistant (type S); (2) fast-twitch, fatigue-resistant (type FR); (3) fast-twitch, fatigue-intermediate (type FInt); and (4) fast-twitch, fatigable (type FF). These generally correspond to the expression of specific myosin heavy chain (MHC) isoforms in the muscle fibers: type I, MHC_{Slow}; type IIa, MHC_{2A}; type IIx, MHC_{2X}; and type IIb, MHC_{2B}. Activation history of motor units also differs, with type S and FR motor units (innervating type I and IIa muscle fibers, respectively) being more frequently recruited than type FInt and FF motor units (innervating type IIx and/or IIb muscle fibers).

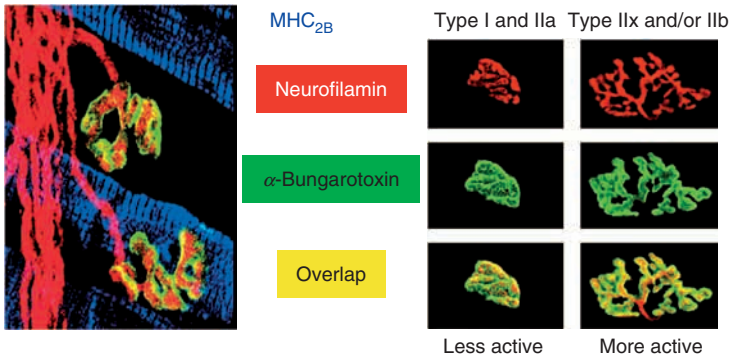

Figure 3 The structure of neuromuscular junctions (NMJs) varies across muscle fiber types. Rat diaphragm muscle fibers were labeled with antibodies specific to the MHC_{2B} isoform (blue). Note that other muscle fibers remain unlabeled and appear black. The motor nerve endings were labeled with antibodies specific to neurofilamin (red), whereas the motor end-plates were stained with rhodamine-conjugated α-bungarotoxin (green). Note the differences in size and complexity (number and length of branches) across fiber types, with NMJs present at type I or IIa fibers being smaller and less complex than those at type IIx and/or IIb fibers (magnification, ×400).

in mechanical properties might affect recruitment of motor units, particularly during high-force, short-duration motor behaviors. With aging, there appears to be a preservation of mechanical properties of motor units required for low-force, sustained motor behaviors. In some cases, the advantage of such preservation is quite obvious – for example, recruitment of type S and FR motor units in the diaphragm muscle to sustain ventilation or a similar recruitment of motor units in antigravity muscles to sustain posture.

Structure of Neuromuscular Junctions

The structural properties of NMJs also vary considerably across different motor unit types (Figure 3). In rat muscles of mixed fiber type composition, NMJs at type I and IIa fibers are smaller, more compact, and show less complexity in branching patterns than NMJs at type IIx and/or IIb fibers. However, in the soleus muscle, NMJs are larger compared to NMJs at type IIx and/or IIb fibers in the extensor digitorum longus muscle, which is also a muscle with fairly homogeneous fiber type composition. It is possible that fiber size is an important determinant of NMJ size regardless of fiber type. Other factors, such as age, species, and activation history, may play important roles in determining NMJ size. Differences also exist across motor unit types in the ultrastructure of pre- and postsynaptic elements of NMJs. The presynaptic terminal surface area is greater at type IIx and/or IIb fibers compared to type I and IIa fibers. The density of active zones at presynaptic terminals does not appear to differ across fiber types; however, because of the difference in total surface area, the total number of active zones is greater at type IIx and/or IIb fibers compared to type I and IIa fibers. The number of docked vesicles at each active zone

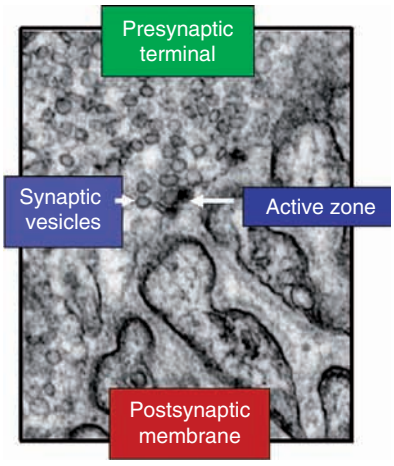

Figure 4 Electron micrograph showing the ultrastructure of neuromuscular junctions.

(i.e., the readily releasable pool, defined morphologically as those vesicles that are fused to the terminal membrane and thus immediately available for release) is comparable across fiber types (Figure 4). However, given the greater number of active zones at type IIx and/or IIb fibers, the total number of synaptic vesicles in the readily releasable pool is greater at these fibers compared to type I and IIa fibers. Synaptic vesicles at presynaptic terminals are also separated into a pool of vesicles that are immediately adjacent to active zones (within 200 nm) and those that are more distant (reserve pool). The densities of synaptic vesicles in both the immediately adjacent pool and the reserve pool are greater at presynaptic terminals of type I and IIa fibers compared to type IIx and/or IIb fibers. Mitochondrial volume density is also greater at presynaptic terminals innervating type I and IIa fibers compared to those innervating type IIx and/or

IIb fibers, possibly reflecting the metabolic requirements of increased activation of these presynaptic terminals.

At motor end-plates of type I and IIa fibers, postsynaptic folding is less complex compared to those at IIx and/or IIb fibers. In addition, at type I and IIa fiber motor end-plates, cellular organelles including mitochondria, rough endoplasmic reticulum, free polysomes, and nuclei are frequently interposed between postsynaptic specializations and myofibrils. Indeed, muscle fiber types can be classified based on these morphological differences at the NMJ. There is no evidence for fiber type differences in the density of cholinergic receptors at the postsynaptic membrane. However, the density of sodium channels near the motor end-plate is much higher at type IIx and/or IIb fibers than at type I fibers.

Aging Effects on NMJ Structure

Aging is associated with an increased number of nerve terminal branches at type IIx and/or IIb diaphragm muscle fibers and a greater fragmentation of nerve terminals. This has also been reported as an age-related increase in complexity of nerve terminal arborization at type IIx and/or IIb fibers. The increased number of terminal branches may reflect sprouting as motor neurons degenerate with aging. The increased branching and fragmentation (complexity) of presynaptic terminals with aging does not lead to an increase in the actual surface area of presynaptic terminals. Indeed, presynaptic terminal area has been reported to decrease with aging. Thus, although it has not been directly assessed, it may be that the total number of active zones at presynaptic terminals decreases with aging and, as a result, the number of synaptic vesicles in the readily releasable pool may decrease. Such changes may reflect alterations in the cyclical remodeling of NMJs reflecting outgrowth and retraction of pre- and postsynaptic elements. It is possible that trophic influences that affect this normal remodeling are altered with aging. Note that an increase in the size of NMJs does not necessarily correlate with improved synaptic efficacy. A possible reflection of instability of NMJ remodeling may be the increased expression of neural cell adhesion molecule that occurs with aging. There is an age-related gradual decrease in the number of cholinergic receptors at the motor end-plate and extrajunctional receptors appear. With aging, there is no change in the numbers of sodium channels at skeletal muscle fibers. However, it is not known whether the relative proportion of junctional, perijunctional, or extrajunctional channels changes with age.

Neuromuscular Transmission

Neuromuscular transmission also varies across motor unit types as a result of differences in safety factor, especially during repeated activation. Safety factor is determined by the ratio of end-plate potential (EPP) amplitude to the threshold for muscle fiber action potential generation. The amplitude of EPPs is greater at type IIx and/or IIb fibers, whereas the threshold for action potential generation is lower at type I and IIa fibers. Overall, the safety factor for neuromuscular transmission during a single evoked response is higher at type IIx and/or IIb fibers compared to type I or IIa fibers. However, during repetitive stimulation the EPP amplitude progressively declines across all fiber types, but the decline is greater at type IIx and/or IIb fibers. Accordingly, type IIx and/or IIb fibers are more susceptible to neuromuscular transmission failure.

Across muscles, whether of mixed or homogeneous fiber type composition, the total number of vesicles in the readily releasable pool is greater at type IIx and/or IIb fibers compared to type I and IIa fibers. The total number of synaptic vesicles released during a single nerve impulse (i.e., quantal content) is higher at type IIx and/or IIb fibers than at type I and IIa fibers. However, there appears to be no fiber type difference in the probability of synaptic vesicle release during a single nerve impulse. With repetitive nerve stimulation, quantal content decreases across all fiber types. Initially, there is a rapid decline (within the first 500 ms) in quantal content that is dependent on the rate of stimulation. This rapid decline in quantal content reflects both a depletion of the readily releasable pool of synaptic vesicles and a decrease in the probability of release. Quantal content continues to decline during repetitive stimulation but at a much slower rate, reflecting a balance between depletion and repletion of the readily releasable pool of synaptic vesicles. Two mechanisms underlie replenishment of synaptic vesicles in the readily releasable pool: (1) recruitment from the immediately adjacent (or available) pool of vesicles and (2) recycling of released vesicles. As mentioned previously, the density of vesicles in the immediately available pool is greater at type I and IIa fibers compared to type IIx and/or IIb fibers. This is consistent with a higher rate of recruitment of vesicles into the readily releasable pool at type I and IIa fibers. It also appears that the extent of synaptic vesicle recycling (e.g., as measured by the uptake of styryl dyes such as FM4–64 or FM1–43) is greater at type I and IIa fibers compared to type IIx and/or IIb fibers. A greater extent of replenishment of the readily releasable pool at type I and IIa fibers during repetitive stimulation may underlie their reduced susceptibility to neuromuscular transmission

failure. Clearly this is important since these motor units are more frequently recruited and are responsible for sustained motor behaviors (e.g., breathing and posture).

Aging Effects on Neuromuscular Transmission

With aging, it has been reported that the safety factor for neuromuscular transmission decreases progressively. Unfortunately, there is little, if any, information regarding the effect of aging on safety factor across different fiber types. As mentioned previously, the safety factor for neuromuscular transmission depends on both EPP amplitude and the threshold for generating an action potential in muscle fibers. Quantal content and EPP amplitude decrease with aging, especially at type IIx and/or IIb fibers. However, the cross-sectional area of type IIx and/or IIb fibers also decreases with aging, which would be associated with an increase in input resistance and, thus, a lower threshold for action potential generation. Therefore, although it has not been directly examined, the safety factor for neuromuscular transmission at type IIx and/or IIb fibers may not be affected by aging. In contrast, the cross-sectional area of type I and IIa fibers changes very little with aging. Accordingly, the threshold for action potential generation would remain unaffected, and an age-related decrease in EPP amplitude would be associated with a decrease in safety factor at these fibers. Such an effect would be especially pronounced during the rapid decline in quantal content that occurs with repetitive stimulation.

Conclusions

There is surprisingly little direct information about the effects of aging on the long-term plasticity of NMJs. Yet aging is clearly associated with changes that could affect plasticity of the NMJ. Physical activity is reduced with aging, and the resulting unloading of limb muscles clearly has an effect on NMJ structure and function that is more pronounced than in younger animals. There is a loss of motor neurons with aging, resulting in the functional denervation of muscle fibers, which are subsequently reinnervated by the sprouting of axons from remaining motor neurons. This process leads to loss of motor unit diversity, with greater loss of those motor units that are able to generate greater forces. Concomitantly, there is a reduction in muscle fiber size at these same motor units. Thus, with aging, there appears to be preservation of motor units required for low-force, sustained motor behaviors, which may be advantageous, for example, in the maintenance of adequate ventilation or posture.

See also: Hebbian Plasticity; Motor Control of Feeding and Drinking; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): A Target for Natural and Environmental Toxins in Humans; Neuromuscular Junction Plasticity in Mammals and Botulinum Toxins; NMDA Receptor Function and Physiological Modulation; Schwann Cells and Plasticity of the Neuromuscular Junction.

Further Reading

Balice-Gordon RJ (1997) Age-related changes in neuromuscular innervation. *Muscle & Nerve* 5: S83–S87.

Courtney J and Steinbach JH (1981) Age changes in neuromuscular junction morphology and acetylcholine receptor distribution on rat skeletal muscle fibres. *Journal of Physiology* 320: 435–447.

Fahim MA, Holley JA, and Robbins N (1983) Scanning and light microscopic study of age changes at a neuromuscular junction in the mouse. *Journal of Neurocytology* 12: 13–25.

Kelly SS and Robbins N (1987) Statistics of neuromuscular transmitter release in young and old mouse muscle. *Journal of Physiology* 385: 507–516.

Mantilla CB, Rowley KL, Fahim MA, Zhan WZ, and Sieck GC (2004) Synaptic vesicle cycling at type-identified diaphragm neuromuscular junctions. *Muscle & Nerve* 30: 774–783.

Prakash YS, Miller SM, Huang M, and Sieck GC (1996) Morphology of diaphragm neuromuscular junctions on different fibre types. *Journal of Neurocytology* 25: 88–100.

Prakash YS and Sieck GC (1998) Age-related remodeling of neuromuscular junctions on type-identified diaphragm fibers. *Muscle & Nerve* 21: 887–895.

Robbins N (1992) Compensatory plasticity of aging at the neuromuscular junction. *Experimental Gerontology* 27: 75–81.

Rosenheimer JL and Smith DO (1985) Differential changes in the endplate architecture of functionally diverse muscles during aging. *Journal of Neurophysiology* 53: 1567–1581.

Ruff RL (1996) Sodium channel slow inactivation and the distribution of sodium channels on skeletal muscle fibres enable the performance properties of different skeletal muscle fibre types. *Acta Physiologica Scandinavica* 156: 159–168.

Sieck GC and Fournier M (1989) Diaphragm motor unit recruitment during ventilatory and nonventilatory behaviors. *Journal of Applied Physiology* 66: 2539–2545.

Smith DO (1988) Muscle-specific decrease in presynaptic calcium dependence and clearance during neuromuscular transmission in aged rats. *Journal of Neurophysiology* 59: 1069–1082.

Smith DO and Weiler MH (1987) Acetylcholine metabolism and choline availability at the neuromuscular junction of mature adult and aged rats. *Journal of Physiology* 383: 693–709.

Wokke JH, Jennekens FG, van den Oord CJ, Veldman H, Smit LM, and Leppink GJ (1990) Morphological changes in the human end plate with age. *Journal of the Neurological Sciences* 95: 291–310.

Wood SJ and Slater CR (1997) Safety factor at the neuromuscular junction. *Journal of Physiology* 500: 165–176.

Neuromuscular Junction (NMJ): Inherited and Acquired Disorders

A Vincent, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Background

The neuromuscular junction is a prototype synapse between the terminal arborizations of the motor nerve and the postsynaptic specialization of the muscle surface. The chemical nature of neuromuscular transmission and the main anatomical features have been known for decades and are described in detail elsewhere. Over the last 30 years the main functional molecules involved in synaptic function have been identified, their genes cloned, and function examined by many different approaches. In parallel, it has become clear that there are a variety of inherited or acquired diseases that can affect neuromuscular transmission resulting in either muscle weakness, or less often muscle hyperactivity. At the same time, a number of different neurotoxins, from a variety of species, have been identified which bind strongly to these molecules. This article discusses the genetic and autoimmune diseases that affect neuromuscular transmission and includes mention of some of the neurotoxins that have helped to make this knowledge possible.

Investigating Neuromuscular Transmission in Human Disease

Much of what we know about the diseases of neuromuscular transmission results from *in vivo* and *in vitro* examination of the neuromuscular junction. Typically, patients with a defect in neuromuscular transmission will complain of muscle weakness or excessive fatigue during repeated effort. The nature of this weakness can be identified by use of electromyography. This involves stimulating a peripheral nerve (e.g., in the forearm) while recording from the muscle it innervates (e.g., the abductor of the thumb). When the nerve is stimulated at low frequency (3 Hz), the amplitude of the compound muscle action potential should remain within 10% of the initial amplitude; any defect in transmission will tend to cause a decrease ('decrement') in the amplitude. An alternative and more sensitive technique is to use single-fiber electromyography in which the time intervals between the responses of two endplates from one motor unit are measured during voluntary activity; defects in transmission at the two endplates will cause excessive variation in this interval ('increased jitter').

In vitro techniques to assess the response of the acetylcholine receptors (AChRs) to released acetylcholine (ACh) include microelectrode recordings from the muscle fibers to look at the amplitude of the miniature endplate potentials and of the endplate potentials (under artificial conditions where the muscle does not fire action potentials). The amount of ACh released per impulse can be assessed by measuring the quantal content (the number of packets released) or by biochemical measurements from isolated muscles. Binding of ^{125}I-α bungarotoxin can be used to measure the numbers of AChRs on the muscle surface. It is also possible to record extracellular potentials from the motor nerve terminals or to place a voltage-clamp pipette on the muscle surface to record the number and characteristics of single AChR openings.

In addition, histochemistry can be used to demonstrate the localization and distribution of acetylcholinesterase (AChE), and immunolocalization studies to identify other autoantibodies and complement bound to the neuromuscular junction in the autoimmune disorders. Finally, electron microscopy and freeze-fractures microscopy have been performed on the neuromuscular junctions to quantify abnormalities in the synaptic vesicles, the active zone particles, or the postsynaptic membrane and secondary folds.

Treatment of Neuromuscular Junction Disorders

In the majority of cases in which the defect is a reduction in neuromuscular transmission, the symptoms can be improved substantially by use of two types of drugs that increase the amount of ACh available to the AChRs. AChE inhibitors are the mainstay of treatment, and diamino-pyridine, which blocks potassium channels and causes prolonged and increased ACh release from the motor nerve terminal, can also be helpful.

In the rare conditions in which neuromuscular transmission is enhanced, notably in neuromyotonia, sodium channel blockers such as the anti-epileptic drugs can help to stabilize the motor nerve and prevent hyperexcitability. In the slow-channel syndrome, quinidine and the serotonin reuptake inhibitor, fluoxetine, both of which block the AChR channel when open, have been found to be helpful.

Congenital Myasthenic Syndromes

The congenital myasthenic syndromes (CMS) are a group of heterogenous inherited disorders that are

Table 1 Targets for autoimmune and genetic disorders of neuromuscular transmission

Target	*Disease*	*Pathogenetic mechanism*
Genetic		
AChR	Acetylcholine receptor deficiency	Recessive mutations in AChR genes causing reduced AChR expression
	Acetylcholine receptor deficiency	Recessive mutations in RAPsyn genes causing reduced anchoring of AChR on the postsynaptic membrane
	AChR kinetic abnormalities	Dominant or recessive mutations in AChR genes causing kinetic defects – 'slow' and 'fast' channel syndromes
AChE	Acetylcholine esterase deficiency	Recessive mutations in the collagen tail that anchors AChE at the NMJ causes absence of AChE
Choline acetylase	Choline acetylase deficiency	Recessive mutations in the gene for choline acetylase that synthesizes ACh causes reduced Ach release
Autoimmune		
AChR	Myasthenia gravis	Antibodies to AChR in 85% reduce AChR numbers
MuSK	Myasthenia gravis	Antibodies to MuSK in 5–10% – mechanism not known
AChRs, MuSK	Transient neonatal myasthenia	Maternal antibodies cause transient disease in neonate; not seen commonly if mother receiving treatment
Fetal AChR	Arthrogryposis (multiple joint contractures in babies)	Maternal antibodies that inhibit function of fetal AChR paralyse baby *in utero* leading to joint contractures; very rare
VGCCs	Lambert Eaton myasthenic syndrome	Antibodies to VGCC in 90% reduce VGCC numbers and ACh release and EPP amplitude
VGKCs	Acquried neuromyotonia	Antibodies to VGKCs in 40% probably lead to spontaneous ACh release

caused by mutations in the genes coding for different molecules at the neuromuscular junction. Most of the conditions are recessively inherited, but one is due to a dominant mutation that causes a toxic gain of function. The main conditions are listed in Table 1 and the targets illustrated in Figure 1.

AChR Deficiency

In this condition the patient is normally born with mild weakness or develops weakness within the first 2 years of life, although in a few cases the diagnosis might not be made for many years. The AChRs at the neuromuscular junction are reduced in number. The result is that the miniature endplate potentials and endplate potentials are reduced in amplitude and neuromuscular transmission fails making the patient weak. The area of neuromuscular contact is typically elongated, possibly as the result of a compensatory sprouting of the motor nerve terminals.

There are two main causes of this syndrome. In one case there are mutations in the AChR epsilon gene (see Figure 2 for the subunit structure of the AChR) which generally cause complete loss of this subunit. This means that the adult form of the AChR cannot be made, and the patient survives by persistent expression of the fetal AChR. In the other case, which is increasingly recognized, the loss of AChRs is due to a defect in the receptor aggregating protein at the synapse (RAPsyn) molecule that is essential for clustering the AChRs during development and for maintaining these clusters in mature muscle. These conditions are usually treated with AChE inhibitors and sometimes also with 3,4-diaminopyridine.

Kinetic Abnormalities of the AChR

There are two disorders in which the AChR is present in normal or near-normal amounts but functions abnormally. In the slow-channel syndrome, dominant mutations in any of the AChR subunit genes lead to abnormally prolonged AChR channel openings. This means that a single nerve impulse results in a grossly prolonged endplate potential. If this endplate potential persists after the refractory period of the muscle action potential, a second action potential will occur and typically this can be recorded *in vivo* by electromyography as a double response to a single impulse. The patients complain of weakness and the condition is progressive often leading to marked wasting of some muscles. It is thought that the prolonged openings allow excess calcium to enter the muscle resulting in 'excitotoxic' damage. This can be prevented by using quinidine or fluoxetine, which block the open AChR channel and help to reduce calcium entry.

The converse situation exists in the AChR fast-channel syndrome, in which mutations lead to reduced duration of each channel opening. The miniature endplate potential amplitudes are very small and the symptoms are very similar to AChR-deficiency syndromes; the treatments are also the same.

AChE Deficiency

This is a recessive disorder and the patients suffer from a progressive disease, not unlike the slow-channel syndrome; this is because loss of the endplate form of the AChE leads to prolonged exposure of the postsynaptic membrane to released ACh. The endplate potentials

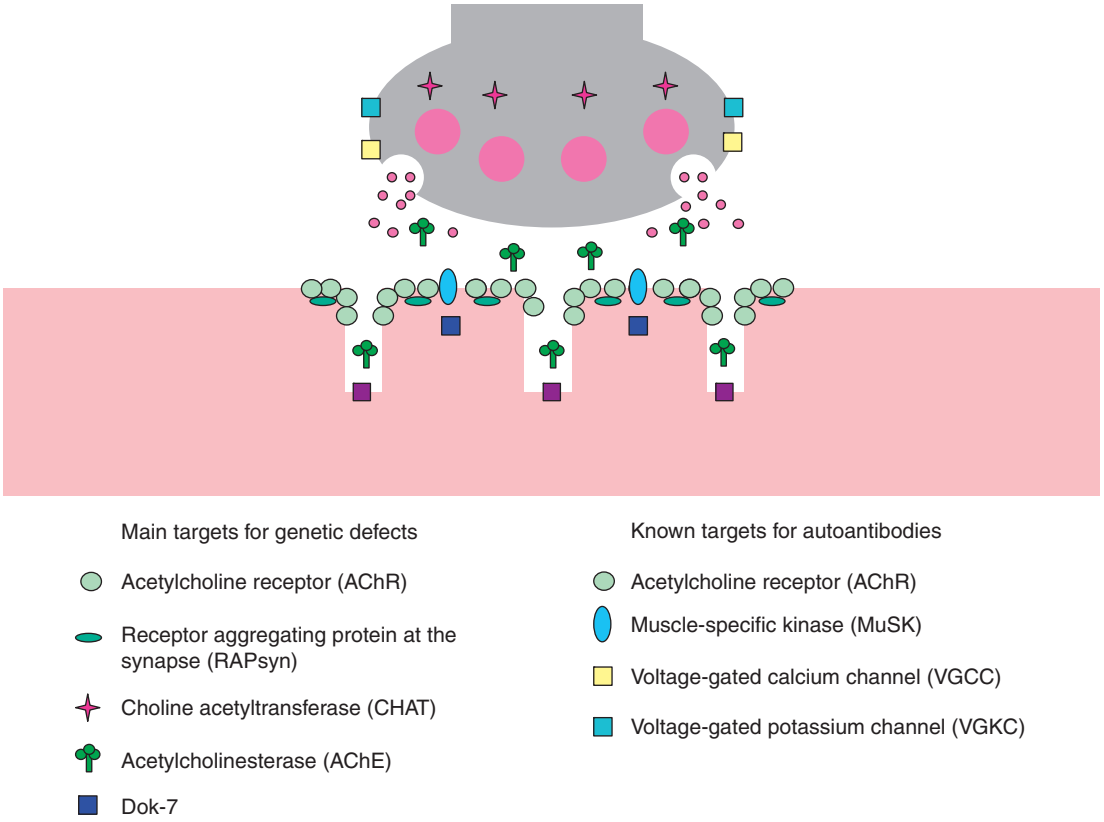

Figure 1 Diagram of the neuromuscular junction illustrating the main enzymes, ion channels, and anchoring proteins involved in genetic (left) or autoimmune (right) diseases.

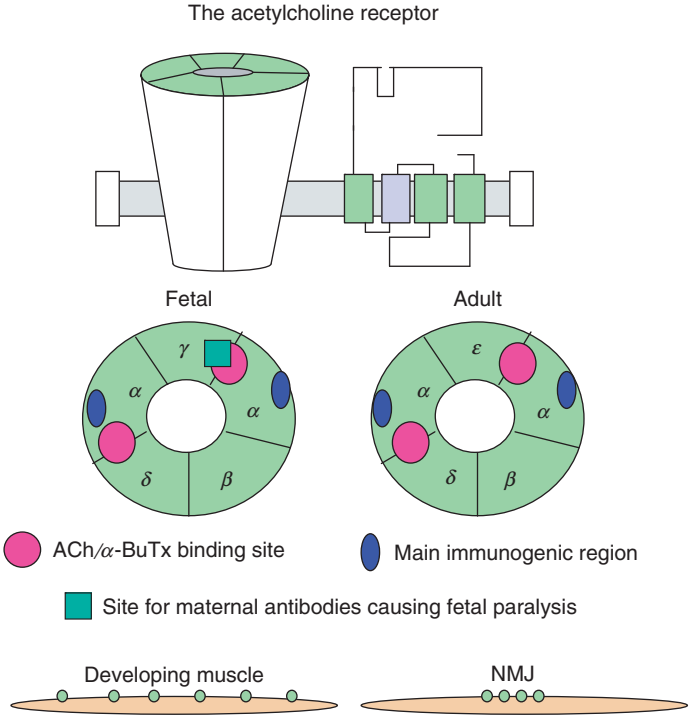

Figure 2 General structure of the AChR in the membrane (top left) showing the topology of each subunit (top right). From above, the AChR consists of five subunits, including two α subunits which are the main sites of antibody binding (main immunogenic regions). The ACh-binding sites are at the interface of the α subunits with their adjacent subunits. There is a site on the γ subunit which binds maternal antibodies that can paralyze the fetus *in utero*.

are prolonged and a double compound muscle action potential is common on *in vivo* testing. This is the most severe of the CMS, and no useful treatments are available at present. Interestingly, the genetic defect is not in the AChE gene itself, but in the collagen molecule, ColQ, that anchors AChE at the neuromuscular junction.

Choline Acetylase Deficiency

A few patients with CMS have reduced ACh release from the motor nerve terminal that may only be evident during repetitive high-frequency nerve stimulation. The patients often present during neonatal life with respiratory problems or choking, but if the syndrome is recognized these can be adequately treated with small doses of AChE inhibitors. The gene defect is in the gene for choline acetylase, the enzyme that synthesizes ACh from acetate and coenzyme A.

Limb Girdle Myasthenia

A recently recognized syndrome is a form of myasthenia that may occur in families and causes weakness of the trunk and limb muscles, with some facial involvement. The neuromuscular junctions in this condition are reduced in size with small motor nerve terminals, reduced area of AChE staining, and reduced AChR numbers; as a result the miniature endplate potentials are only marginally reduced in amplitude but the endplate potentials are small due to reduced numbers of packets of ACh released. It is not clear why the neuromuscular junction does not increase in size to try to compensate for the defective transmission as it appears to do in some of the other syndromes. Mutations in a postsynaptic cytoskeletal protein Dok-7 have recently been identified in this syndrome, but how these affect both pre- and postsynaptic morphology and function is not yet clear.

Autoimmune Disorders

The autoimmune disorders are caused by the immune system reacting against self-proteins at the neuromuscular junction. The humoral immune system has the capacity to react against foreign microbes or parasites, producing immunoglobulin antibodies that bind specifically to the foreign proteins. These immunoglobulin (antibody) molecules are produced by specific B (bone-marrow-derived) lymphocytes. For these antibodies to be made, other lymphocytes, called T (thymus-derived) helper cells, have to recognize the antigen and help the antibody response. New antibodies are being made all the time in response to different environmental, or internal, stimuli. Sometimes the antibodies are able to bind to self-proteins resulting in autoimmune disease. Thus, these diseases are not inherited but are acquired during life. The environmental, or possibly internal stimuli, that initiate the self-reactive immune response are only known in a few diseases. Although they are often chronic conditions, they can be treated by removing the antibodies from the circulation or by suppressing the immune system so that it produces less of the pathogenic antibodies.

The reason that we understand so well the diseases of the neuromuscular junction depends, in large part, on its accessibility to circulating factors. Quite large molecules like antibodies (MW 160 000) can diffuse out of the blood and find their way to the neuromuscular junction which is not protected by a 'blood–nerve barrier.' As a result, many venomous animals have taken advantage of this accessibility to develop neurotoxins that, when absorbed into the circulation, can access the neuromuscular junction, block the crucial ion channels, and paralyze their prey. These neurotoxins have been of great value in the study of the diseases described.

Myasthenia Gravis

Myasthenia gravis (MG) is an acquired autoimmune disease caused in most cases by autoantibodies to the AChR. It was probably first described by Thomas Lewis in 1676. The patients first notice the disease at any age, but childhood cases are rare, young female adults are common, and older patients (after 50 years) are increasingly being identified. They complain of muscle weakness and fatigue. The eye muscles are often involved first causing double vision and drooping of the eyelids (ptosis). There is usually limb weakness and it is common to have weakness of facial and pharyngeal and respiratory muscles (bulbar weakness), which can be life threatening when breathing is impaired. However, some patients remain with only ocular muscle weakness (ocular MG).

The story of how MG was shown to be autoimmune and due to antibodies to the AChR is one of the best examples of how basic science can lead to understanding of clinical disease. It was known for many years that the defect in MG was at the neuromuscular junction, and in the 1960s it was shown that the miniature endplate potential amplitudes were very small – often almost undetectable. However, it was not clear whether the defect was in the amount of ACh in each packet or the number of AChRs. Two findings in the 1960s and 1970s were crucial. First, Lee and his collaborators working on the venom of *Bungarus multicinctus*, the banded krait, found that it contained a neurotoxin, α-bungarotoxin, that blocked neuromuscular transmission completely and irreversibly. They showed that this toxin binds highly

specifically to the endplate region of normal and denervated muscles (when AChRs appear over the entire muscle surface), and prevented the response of the muscles to ACh. This indicated that the toxin was blocking the AChRs themselves rather than interfering with release of ACh. Drachman and his colleagues at Johns Hopkins University were then able to use radioactively tagged α-bungarotoxin to identify and measure AChRs in human muscle biopsies, and they found that the number of binding sites was reduced in MG biopsies. Almost concurrently with these observations, Patrick and Lindstrom working at the Salk Institute, purified AChRs from the electric eel and immunized rabbits against AChR in order to raise antibodies for use in immunolocalization studies. The rabbits developed antibodies to the AChRs and also became very weak. These two observations together strongly suggested that MG could be the result of autoimmune response to self-AChRs in the patients, as was shown by Lindstrom a couple of years later.

For the last 30 years the diagnosis of MG has been confirmed by measuring the level of antibodies to AChR in the patients' serum or plasma. The diagnostic test relies on immunoprecipitation of human AChRs that have been radioactively tagged with ^{125}I-α-bungarotoxin. Normal healthy individuals do not have antibodies to the AChR and their serum antibodies do not therefore immunoprecipitate the AChRs; in 85% of patients with MG, the AChR antibody levels are much higher, although quite variable.

Autoantibodies that bind to self-protein may not always be pathogenic This is the case if the protein is not accessible to the circulating antibodies, or the antibodies do not bind to the native conformation of the protein. It was therefore important to demonstrate that these antibodies were causing the disease, rather than perhaps developing as a secondary immune response resulting from other abnormalities at the neuromuscular junction. This was done in three main ways: first, the antibodies to the AChR in MG were shown to be directed against the extracellular, exposed, domains of the AChR and they were bound to the AChRs in animal models of MG and in the patients' muscle biopsies. Second, if the patient is treated by plasma exchange, which removes large amounts of the patients' own antibodies and replaces them with substitute healthy antibodies, they almost always improve substantially. The improvement is only maintained for as long as the AChR antibody levels remain low – when the levels rise again, due to resynthesis, the patients' symptoms return. Third, signs and experimental evidence for myasthenia can be induced in experimental animals by injecting them with the serum or purified immunoglobulin antibodies of the patients. Thus, the AChR antibodies in MG are pathogenic.

In general, the serum antibodies in MG patients do not block directly the function of the AChR, although such antibodies are present in some patients. The antibodies, particularly those directed toward the main immunogenic region (see Figure 2), can bind divalently to the AChRs at the neuromuscular junction, and this causes increased internalization and degradation of the AChRs. In addition, and perhaps most importantly, serum complement components can bind to the antibodies, activate the membrane attack complex, and lead to lysis and morphological damage to the postsynaptic membrane (Figure 3). The reduction in the postsynaptic folds may enhance the defect in transmission by reducing the number of VGKCs. On the other hand, there is some evidence that both the motor nerve and the muscle try to compensate for the loss of AChRs by increasing ACh release, and increasing AChR synthesis, respectively.

An obvious question is why do the patients develop antibodies to their own AChRs? There are a few clues. One is the presence of abnormalities in the thymus gland. This gland is responsible for educating the cells of the immune system, and it expresses many self-proteins so that those cells of the immune system that might react against 'self' can be deleted. Small muscle-like cells called myoid cells express AChRs and other muscle proteins. In some patients at least, an immune response toward these myoid cells seems to initiate myasthenia. This is supported by the fact that lymphocytes cultured from the thymus gland produce AChR antibody, and if the gland is removed the patients tend to improve over the following year with decreases in their AChR antibody levels. However, this applies mainly to patients under the age of 50 years. A proportion of slightly older patients (usually between 30 and 60 years) will have a thymic tumor, which is also likely to be involved in sensitizing the immune system to the AChR protein, but the cause of the disease in those who develop symptoms later in life without a thymic tumor is not at all clear.

Ocular MG

Only about 50% of the patients whose symptoms do not progress beyond the ocular muscles (causing ptosis and double vision) have AChR antibodies. It is still a mystery as to why the disease is so restricted. One possible explanation is that neuromuscular transmission in the extraocular muscles that control eye movements is subtly different from that in other muscle groups; this is partly the case as some of the muscle fibers are multiply innervated and muscle contraction depends directly on the endplate potentials generated at each junction, rather than on a propagated action potential. It has also been suggested that the extraocular muscles express more of the fetal

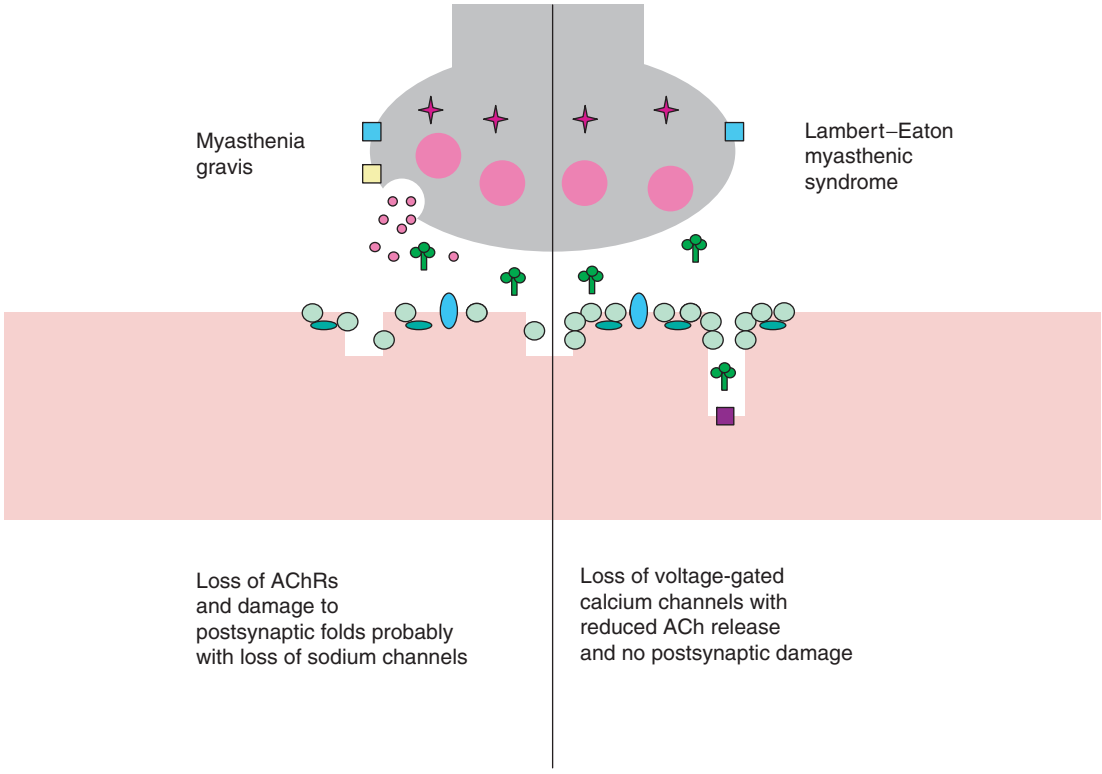

Figure 3 The neuromuscular junction in myasthenia gravis (left) and the Lambert–Eaton myasthenic syndrome (right).

AChR, but this has been disputed by subsequent studies. Another possibility is that these muscles do not express sufficient complement inhibitors. Whatever the explanation, it should be realized that these muscles are often the first to be affected in systemic neurotoxicity such as with botulinum or snake envenomation. Thus, the basis of their susceptibility to fatigue must lie partly in the physiology of their neuromuscular junctions.

MG without AChR Antibodies

Only about 85% of patients with generalized MG have AChR antibodies, and for many years there was controversy about whether the remaining 15% had a different disease. It turns out that some of the 15% have antibodies to another protein at the neuromuscular junction. This is the muscle-specific kinase (MuSK). MuSK is expressed in muscle during development and is essential for orchestrating the RAPsyn-induced clustering of the AChRs at the developing neuromuscular junction. It is not yet clear what role(s) MuSK plays in mature muscle when the AChRs are already clustered, and when the high density of AChR is maintained by endplate-specific synthesis.

Only up to 50% of the patients without AChR antibodies have MuSK antibodies, and it is not clear how these antibodies cause a neuromuscular junction defect. However, these patients often have mild limb weakness with marked facial and bulbar symptoms, relatively frequently leading to muscle atrophy or wasting which is not a marked feature of typical MG. It is possible that the MuSK antibodies interfere not only with MuSK's function at the neuromuscular junction, leading to the myasthenic symptoms, but also to some downstream functions involved in the trophic support of the muscle itself.

Maternal Antibodies and Development

It has long been recognized that some mothers with MG give birth to babies who have a transient form of the disease. This results from the transfer of the AChR antibodies across the placenta, which is a normal event that occurs during the latter half of pregnancy. In very rare cases, a mother has high levels of antibodies that are mainly directed against the fetal form of the AChR (Figure 2). When these transfer to the fetus they can lead to neuromuscular junction failure and fetal paralysis. The baby may survive but has multiple joint contractures and other abnormalities as a result of lack of movement *in utero*.

The Lambert–Eaton Myasthenic Syndrome

In 1957 Lambert and Eaton described a myasthenic syndrome that was distinct from MG. The distribution

of weakness usually affects the legs first with relatively little ocular muscle involvement. Also, muscle reflexes which are normal in MG are absent or depressed but may become stronger after voluntary contraction. Autonomic symptoms (dry mouth, constipation, impotence) are present in many patients, suggesting that the defect may be one that is common to autonomic neurons.

The difference from MG was most obvious on electromyography. When the motor nerve was stimulated *in vivo*, the compound muscle action potential (CMAP) amplitude was small and became smaller at low rates of stimulation ($<10\,s^{-1}$), but during stimulation at higher rates, or following a few seconds of voluntary contraction, the amplitude increased substantially ('increment'). These findings contrasted with those in MG where the CMAP fails to show an increase during high-frequency stimulation (see above).

The basis for these changes was examined in patients' muscle biopsies. These showed that the miniature endplate potentials were normal but the endplate potentials were very small, indicating a reduction in the number of packets of ACh released per impulse. The endplate potentials increased during repetitive stimulation, and also in response to increases in extracellular calcium concentration. It seems that increasing the buildup of intracellular calcium, either by increasing the amount of extracellular calcium or by high-frequency nerve stimulation, increases ACh release and improves neuromuscular transmission. Moreover, freeze-fracture electron microscopic studies showed that the active zone particles, which are thought to be the voltage-gated calcium channels (VGCCs) (Figure 1), were reduced in numbers and the normal double parallel arrays of these particles were mainly absent.

The approaches used to prove that Lambert–Eaton myasthenic syndrome (LEMS) was autoimmune, relied on those established in MG. Plasma exchange was found to lead to clinical improvement and the patients also improved clinically when given long-term immunosuppressive drugs. Moreover, daily injection of LEMS plasmas, or IgG fractions, into mice led to reduced endplate potentials and quantal content in the mouse muscles, with changes in the active-zone particles similar to those found in the patients' biopsies.

These observations suggested VGCC as a target for antibodies in LEMS. The VGCCs are a family of transmembrane proteins consisting of $\alpha1$, β, and $\alpha2/\delta$ subunits. The VGCC at the neuromuscular junction is the $\alpha1$A (P/Q-type) and binds a particular cone snail-derived toxin, ω-Conotoxin (ω-CmTx) MVIIC. It was relatively easy therefore to use ^{125}1-ω-Conotoxin (ω-CmTx) MVIIC to tag solubilized VGCCs and test the patients' sera for antibodies to the VGCC (in a method analogous to that used for AChR antibodies in MG). Further studies showed that the antibodies were highly specific for this VGCC. Indeed, in mice injected with the patients' antibodies, there is upregulation of another VGCC, the $\alpha1$B (N-type) which seems to be an attempt of the motor nerve to compensate for loss of the $\alpha1$A (P/Q-type) VGCC.

Similarly, the $\alpha1$A (P/Q-type) VGCCs play a role at autonomic synapses, although this role is much more subtle. For instance, multiple subtypes of VGCC are involved in the release of neurotransmitter at mouse bladder and vas deferens smooth muscles, including P-, Q-, N-, and R-types. Injection of LEMS IgG into mice causes a reduction in the proportion of release that is dependent on the $\alpha1$A (P/Q-type) without in general interfering with the release dependent on the other subtypes.

Paraneoplastic Neurological Disorders

The cause of LEMS in half the patients is unknown, but about 50% of the patients have a small cell lung cancer (SCLC). This tumor is aggressive and generally has a poor prognosis. Only 3% of the patients with SCLC develop LEMS, but it is highly likely that the tumor is responsible because SCLC cells express VGCC on their surface. Therefore, the VGCC antibodies are a byproduct of the body's attempt to get rid of the tumor, but the antibodies bind to the VGCCs on the motor nerve terminals, causing the neurological symptoms. From this it follows that if the tumor is successfully removed or treated, the stimulus decreases and the patient tends to improve. Moreover, the antibody response against the SCLC VGCCs may inhibit tumor growth, so that patients with the autoimmune disease may have a better prognosis than those who do not have LEMS.

There is a family of diseases in which the tumor initiates a neurological condition. The recognition of these 'paraneoplastic neurological syndromes' is important because in many cases the onset of the neurological disease ensures early detection of the relevant tumor. However, not all of these conditions are due to autoantibodies and in many the patients do not improve with immunological or tumor treatments.

Acquired Neuromyotonia

Neuromyotonia (NMT), or Isaacs' syndrome, is a syndrome of muscle hyperactivity with muscle stiffness and cramps most prominent in the limbs and trunk. Increased sweating is also very frequent. In this condition, the electromyograph (EMG) shows spontaneous motor activity. Muscle biopsies have not been studied in NMT, but the muscle hyperactivity is initiated in the motor nerves because it can be

blocked by curare which is an AChR antagonist (given locally in the arm).

NMT may be associated with other autoimmune diseases or other autoantibodies suggesting an autoimmune basis, and cerebrospinal fluid analysis may show an increase in IgG or 'oligoclonal bands', suggesting that there is an immune response to neuronal antigens. Like MG, NMT can be associated with a thymoma. Moreover, plasma exchange results in clinical improvement, and passive transfer of the patients' serum antibodies to mice results in increased motor nerve and sensory nerve activity and an increase in the number of AChR packets released. These results were very similar to those obtained with low concentration of the potassium channel blockers, 4 amino-pyridine or 3,4-diaminopyridine.

The most likely target for the autoantibodies, therefore, was a potassium channel. Aminopyridines or specific toxins such as α dendrotoxin block certain types of VGKCs and cause the nerve to generate repetitive or spontaneous bursts of action potentials A functional VGKC consists of four transmembrane α subunits that combine as homomultimeric and heteromultimeric tetramers. α-Dendrotoxin binds to VGKCs of the subtypes Kv1.1, 1.2, and 1.6: 1.1 and 1.2 are highly expressed at nodes of Ranvier and are therefore important for controlling motor nerve excitability. The subtype expressed at the motor nerve terminal, however, is not known, and dendrotoxin has only quite modest effects on neuromuscular transmission *in vitro*.

Antibodies to VGKCs can be detected in about 40% of NMT patients by immunoprecipitation of ^{125}I α-dendrotoxin-labeled VGKCs. Antibodies to Kv1.2 probably dominate in most patients, but antibodies to Kv11 or Kv16 may be important in determining some of the clinical effects. The antibodies, or IgG fractions, can reduce potassium channel currents in a neuroblastoma cell line, Nb-1, and this does not require complement but does require divalent (Fab)2 antibodies. Thus the antibodies probably act mainly by cross-linking the VGKCs in the membrane.

Some patients with NMT also have autonomic and central nervous system disease with cardiac and gastrointestinal problems, memory loss, and sleep disorder (Morvan's syndrome). Stemming from studies on these very rare patients, high levels of antibodies to VGKCs have also been found in individuals who present with memory loss and seizures without (for unknown reasons) the peripheral or autonomic disturbance (usually called nonparaneoplastic limbic encephalitis). Both these diseases improve with immunological treatments that reduce the VGKC antibody levels.

Miller–Fisher Syndrome

There is another condition that can be considered with disorders of the neuromuscular junction, although it is strictly classified as an immune-mediated neuropathy. Miller–Fisher Syndrome (MFS) is an acute, monophasic disease in which the patients suffer from ophthalmoplegia (ocular muscle paralysis), ataxia (incoordination of movements), and areflexia (loss of muscle reflexes). It usually gets better spontaneously and is not life threatening although sometimes there is overlap with the Guillain–Barre syndrome. The condition often follows an infection with a bacterium, *Campylobacter jejunii*, and the patients' sera almost always contain antibodies to a ganglioside (glycolipid molecule) called GQ1b. It appears that these antibodies are formed primarily against similar structures on the bacterial cell wall, and that this is a prime example of molecular mimicry in which autoimmunity arises as a result of cross-reactivity between microbial and self-antigens.

The relevance to the neuromuscular junction was recognized when serum from patients with this condition was applied to mouse neuromuscular preparations *in vitro*. The MFS serum produced a massive release of ACh from the motor nerve terminals which began about 10 min after application, and continued until transmission failed altogether. This has been shown to result from the binding of the GQ1b antibodies to the glycolipid on the motor nerve terminal, and activation of complement leading to influx of calcium and release of the transmitter. How these *in vitro* findings relate to the patients only neuromuscular junctions *in vivo* is not yet clear.

See also: Autonomic Neuroimmunology; Calcium Channels; Excitotoxicity in Neurodegenerative Disease; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction (NMJ): A Target for Natural and Environmental Toxins in Humans; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Postsynaptic Events in Neuromuscular Transmission; Presynaptic Events in Neuromuscular Transmission; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Engel AG, Ohno K, and Sine SM (2003) Sleuthing molecular targets for neurological diseases at the neuromuscular junction. *Nature Reviews Neuroscience* 4(5): 339–352.

Vincent A (2002) Unravelling the pathogenesis of myasthenia gravis. *Nature Reviews Immunology* 2(10): 797–804.

Vincent A, Bowen J, Newsom-Davis J, and McConville J (2003) Seronegative generalised myasthenia gravis: Clinical features, antibodies, and their targets. *Lancet Neurology* 2 (2): 99–106.

Newsom-Davis J, Buckley C, Clover L, et al. (2003) Autoimmune disorders of neuronal potassium channels. *Annals of the New York Academy of Sciences* 998: 202–210.

Buckley C and Vincent A (2005) Autoimmune channelopathies. *Nature Clinical Practice Neurology* 1: 22–33.

Relevant Websites

http://www.myasthenia.org – Myasthenia Gravis Foundation of America, Inc.

http://en.wikipedia.org – The Free Encyclopedia.

http://www.mgauk.org – The Myasthenia Gravis Association.

http://www.pnseuronet.org – Concerted Action on Paraneoplastic Neurological Syndromes.

http://www.neurowustl.edu – Washington University in St. Louis.

Neuromuscular Junction (NMJ): Mammalian Development

C R Slater, University of Newcastle, Newcastle upon Tyne, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The neuromuscular junction (NMJ) is an example of a highly specialized, asymmetric, cell-cell junction. The speed with which fast chemical transmission occurs requires that the specializations of the postsynaptic muscle cell surface that allow it to respond to the transmitter released from the nerve are less than a micrometer from the presynaptic release sites. An important question concerning the development of the NMJ is therefore, what are the mechanisms that account for the alignment and proximity of the distinctive specializations of its pre- and postsynaptic components? A second important question concerns the specificity of the mature nerve-muscle contacts. What part does the process of NMJ formation play in achieving functional matching of individual motor neurons to the muscle fibers they innervate?

This article gives an overview of the way mammalian NMJs develop. Most of the information has come from rats and mice, in which the most detailed studies of NMJ formation in vertebrates have been made. The article starts with the origin of motor neurons and skeletal muscle fibers and then considers how the NMJ itself is formed.

Development of Motor Neurons

Birth

Motor neurons are among the earliest nerve cells to be born, that is, to complete their final round of DNA synthesis. Soon after cell birth, the motor neurons begin to extend an axon that leaves the spinal cord in the nascent ventral roots. In mammals, the motor neurons that innervate an individual muscle are usually grouped into a longitudinally oriented column that extends over two to three spinal segments.

Axon Outgrowth and Motor Neuron Identity

The immature motor axons leave the spinal cord even before their target muscles have formed. As they grow, the axons select paths that lead to the muscles they are destined to innervate. The ability of an immature motor neuron to make such decisions indicates that it has some knowledge of its identity, and that different motor neurons therefore have different identities. In adults, motor neurons that innervate slowly contracting nonfatigable muscle fibers tend to be rich in oxidative enzymes that can be visualized by appropriate histochemical techniques. Well before the first NMJs are formed, embryonic motor neurons already differ in their oxidative enzyme profiles. This supports the idea that the motor neurons that innervate a single muscle differ in their properties and that those differences arise before any interaction with the muscle occurs.

Once contact with the appropriate premuscle mass has been established, but not before, the axons branch extensively. In rats and mice, functional contacts with newly formed limb muscles are first present around embryonic day 14 (E14), a week before birth. A similar stage occurs in humans at about week 9 of gestation.

Release of Acetylcholine from Growth Cones

The terminals of cultured motor neurons can release acetylcholine (ACh), the chemical transmitter at neuromuscular junctions, even before they make contact with muscle. This suggests that motor neurons *in vivo* also synthesize ACh and have the necessary specializations for its activity-dependent release at an early stage of their development.

Development of Skeletal Muscle Fibers

Myotube Formation

Most vertebrate skeletal muscle fibers arise during development from the fusion of many mononucleated postmitotic myoblasts. These spindle-shaped cells line up and fuse to form the primitive myotubes. As the myotubes grow and mature into muscle fibers, they incorporate additional myoblasts. In adult muscle fibers, there is typically one nucleus for every 10 μm of length. Thus, a single fiber in a large human muscle, such as *vastus lateralis*, which has fibers up to 20 cm long, has up to 20 000 nuclei.

Kinetics of Primary and Secondary Myotube Formation

In mammalian muscles, myotube formation occurs in two phases. The first involves the formation of an initial cohort of relatively few primary myotubes. As the muscle elongates, further myoblasts line up along the primary myotubes and eventually fuse to form secondary myotubes. These form initially roughly midway between the ends of the primary myotubes and then grow rapidly in length as they add new myoblasts. The two populations of myotubes, and the mature muscle fibers that form from them, have

somewhat different functional properties. The great majority of fibers in the adult muscles are derived from secondary myotubes.

Origin of Muscle Fiber Types

Most adult mammalian muscles contain fibers of different functional types. Some are specialized for relatively slow, sustained contraction whereas others are specialized for fast, high-powered contractions. At birth, the muscle fibers in rats and mice contract uniformly slowly. Differentiation into faster and slower contracting fibers begins soon after birth and is well established 2–3 weeks later. Most of the muscle fibers that develop from primary myotubes end up as slow in the adult, while secondary myotubes give rise to both fast and slow fibers. These distinctive properties arise from the pattern of genes expressed by the fibers. Most of the nuclei in each fiber express the same set of genes, raising the as yet unanswered question of how that homogeneity of gene expression comes about.

Early Appearance of Delocalized Postsynaptic Properties

Even before they are innervated, immature muscle fibers begin to express the receptor proteins that allow them to respond to ACh (acetylcholine receptors, or AChRs). AChRs of the fetal form, containing $\alpha_2\beta\delta\gamma$ subunits, are initially present over the whole fiber surface at a density of about 500 μm^{-2}. Other molecules that play roles in AChR localization and gene expression, such as rapsyn and muscle specific kinase (MuSK; see 'AChR aggregation' below), have a similar pattern of expression. As the muscle fibers become increasingly active, their activity suppresses the expression of all these proteins away from the NMJ.

Early Development of the NMJ

Initial Nerve–Muscle Encounters

Muscle fibers become innervated very soon after they first form. In rats and mice, signs of functional innervation can be detected within a day or two of the earliest myotube formation at about E12–14, depending on the muscle. The early nerve-muscle contacts lack many of the structural features of mature NMJs but are characterized by a high density of AChRs in the muscle fiber membrane. Whether the AChR clusters form before or after nerve contact is a matter of continuing debate (see below). Indeed, there is evidence that NMJs in different muscles in the same species form by different sequences of events.

Polyaxonal Innervation

A distinctive feature of the early motor innervation of vertebrates is that several motor neurons initially innervate each muscle fiber. This innervation occurs at a single postsynaptic site that is thus contacted by the terminal axons of several motor neurons (Figure 1). Myotubes cultured *in vitro* can acquire polyneuronal innervation, but such multiple inputs are normally distributed over the myotube surface rather than focused on a single site. The forces that initially restrict the immature nerve terminals to this single site *in vivo* are not known.

AChR Accumulation

A high density of AChRs is a hallmark of the postsynaptic membrane of the vertebrate NMJ. A distinct cluster of AChRs, detectable after labeling with fluorescent conjugates of the snake toxin α-bungarotoxin, is present from a very early stage of NMJ formation (Figure 2). Within the immature cluster, AChRs are often gathered into microclusters less than 1 μm across. The mean density of AChRs within the plaque

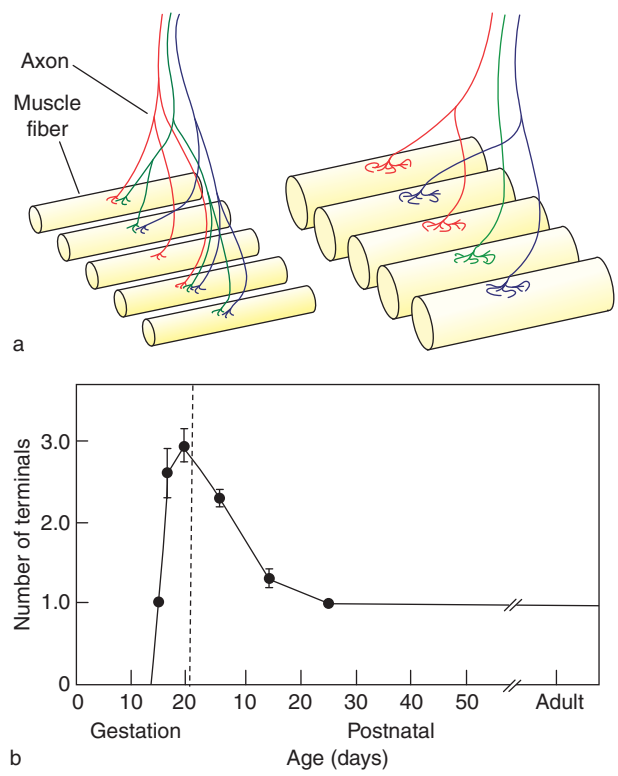

Figure 1 Development of muscle innervation. (a) Comparison of polyneuronal innervation of muscles in newborn rats or mice (left) with the adult state (right). (b) Changes in the number of axon terminals innervating neuromuscular junctions in rat diaphragm with age. (b) Reproduced from Bennett MR and Pettigrew AG (1974) The formation of synapses in striated muscle during development. *Journal of Physiology* 241: 515–545, with permission from Blackwell publishing.

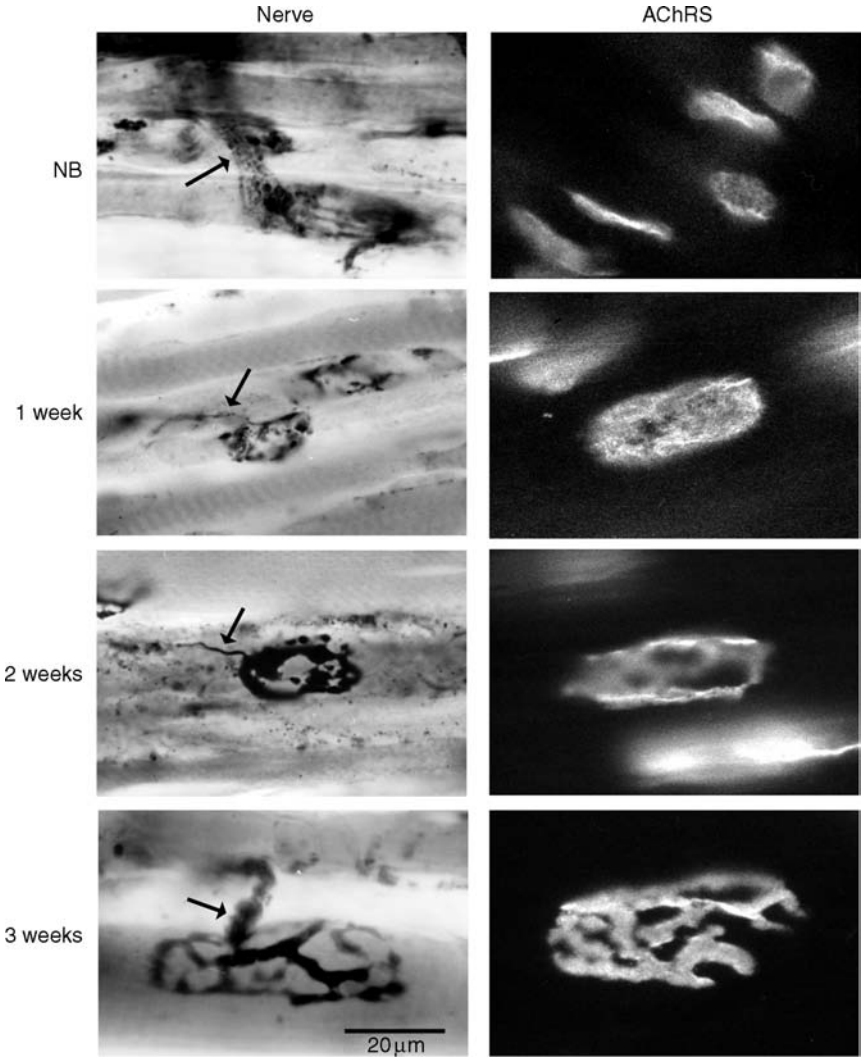

Figure 2 Structural maturation of mouse neuromuscular junctions (NMJs). Nerve terminals in postnatal mouse extensor digitorum longus muscles were labeled with zinc iodide-osmium, and acetylcholine receptors (AChRs) were labeled with R-α-bungarotoxin. Arrows point to the preterminal motor axon(s). Several axons are present at birth, but between 1 and 2 weeks after birth, all but one are withdrawn. The surviving axon becomes myelinated 2–3 weeks after birth. As the axon terminal enlarges, the initially uniform granular distribution of AChRs breaks up, leaving areas with low AChR density that come to match the branches of the nerve terminals. NB, newborn.

is about $3000\,\mu m^{-2}$. How the AChR cluster forms initially and is subsequently maintained are topics of extensive current research. There appear to be at least two components of this process: the aggregation of AChRs within the membrane of the muscle fiber and the enhanced expression of the genes encoding the AChR subunits by the myonuclei closest to the site of nerve contact.

AChR aggregation AChRs can exist in a form that is mobile in the plane of the cell membrane. In the case of cultured immature frog muscle precursor cells, such mobile AChRs aggregate at sites of contact with motor axons, suggesting that the nerve induces aggregation of AChRs in the muscle cell. Much evidence suggests that it is a protein, agrin, which is made by and released from motor axons, that triggers AChR aggregation. Agrin function is associated with its activation of a muscle specific kinase (MuSK), which triggers downstream signaling cascades in the muscle. A second key player in the process of agrin-induced AChR aggregation is the 43 kDa protein rapsyn. Rapsyn links AChRs to the plasma membrane and to components of the membrane skeleton. In the absence of rapsyn, no aggregation of AChRs occurs.

There is also much evidence that AChRs can self-aggregate, and on suitable substrates they can form clusters similar to those at mature NMJs. Some recent

studies have shown that in some species and muscles, AChR clusters form in the region of presumptive NMJs before nerve-muscle contacts are made. In these cases, neuromuscular contacts may form selectively on the presumptive postsynaptic sites. On balance, it seems likely that while the muscle fiber surface has an inherent tendency to form AChR clusters, agrin released from the nerve can either trigger clustering *de novo* or stabilize preexisting clusters.

Upregulation of postsynaptic gene expression The second process that contributes to AChR accumulation at the mature NMJ is the enhanced expression of the genes that encode the AChR subunits. This process is discussed below.

Within a few days of nerve contact, acetylcholinesterase (AChE), the enzyme that terminates ACh-action, accumulates in the synaptic basal lamina that forms between the nerve and muscle cells. As for AChRs, this accumulation is associated with upregulation of the genes encoding AChE.

Schwann Cells: The Motor Neurons' Companion

Throughout most of their life, motor axons are closely accompanied by Schwann cells. There is increasing evidence of a mutual dependency of motor axons and their companion Schwann cells during NMJ development. Thus, if the motor axons are cut at birth, the Schwann cells associated with their degenerating distal portion die. This can be prevented by addition of neuregulin, a growth factor normally produced by the motor neuron. Consistent with a role for neuregulin in promoting Schwann cell survival is a reduction of Schwann cell numbers in mice in which neuregulin expression is reduced by genetic manipulation.

NMJ Maturation: Presynaptic

At birth, both the pre- and postsynaptic components of the NMJ of rats and mice are immature in structure, function, and molecular makeup. Nonetheless, neuromuscular transmission is adequate for the simple movements of the neonate. The adult form of the NMJ arises in concert with increasing motor activity during the next 3–4 weeks as the result of a program of coordinated events affecting the presynaptic nerve terminal and the postsynaptic surface of the muscle fiber.

Myelination of the Axon

In rats and mice, myelination of the peripheral nerves begins a day or two after birth. Within a further week or so, the extramuscular parts of the nerves are well myelinated. By contrast, although the fine intramuscular branches leading to individual muscle fibers remain in close contact with Schwann cells, they remain unmyelinated until the elimination of supernumerary innervation is complete.

Synapse Elimination

During the first 2 weeks or so after birth, all but one of the axons that initially innervate each muscle fiber withdraws, leaving a sole survivor (Figure 1). A similar process occurs at most of the vertebrate NMJs that have been investigated. In humans, synapse elimination is complete by about 14 weeks of gestation. This important process has been extensively investigated, both at the NMJ and in the CNS.

A very different pattern of muscle innervation occurs in the muscle of fish, some muscles in other lower vertebrates, and in a few muscles of mammals. In these muscles, the mature muscle fibers are innervated at multiple sites by numerous axons. Although the development of the innervation pattern in these multiply innervated muscles has not been studied in detail, it is clear that the immature nerve–muscle contacts are not restricted to a single site and that the local competition that occurs in most mammalian muscles does not occur. A common feature of these muscle fibers is that instead of generating action potentials, the contraction is triggered by the summed effects of the local depolarizations at the multiple sites of innervation.

Maturation of the Nerve Terminal

Once synapse elimination is complete, the terminal of the sole surviving motor axon expands significantly. This involves a broadening of the regions of contact with the muscle and an overall increase in the area of synaptic contact (Figure 2(a)). As the nerve terminal expands, it retains its close contact with the terminal Schwann cells, which increase in number as a result of continuing cell divisions.

The factors that limit or determine the size of the mature terminal are poorly understood. Zones of postsynaptic specialization can be induced in a variety of experimental conditions and may adopt a size and appearance remarkably similar to those at mature NMJs. This suggests that factors within the muscle play an important role in determining NMJ size and its well-established correlation with muscle fiber caliber. This view has found recent support from the discovery that a class of mutations of the gene encoding Dok-7, a postsynaptic protein that regulates MuSK activity, causes a reduction in human NMJ size without significant alteration in muscle fiber size or the local density of AChRs. It remains to be seen whether neuromuscular synaptic size–strength homoestasis in vertebrates is regulated by mechanisms

similar to those recently shown to operate in *Drosophila* larvae.

Transmitter release While growing motor axons may be able to release ACh in response to depolarization (see above), evidence from *in vitro* studies indicates that contact of the nerve terminal with muscle cells greatly increases the efficacy of quantal release. The structural basis of this enhanced release is unclear since active zones, the sites of nerve-induced quantal release at mature NMJs, have not been described until the NMJs acquire their mature form.

Even at immature NMJs, stimulation of any of the several axons innervating each muscle fiber is often adequate to trigger muscle contraction. However, the quantal content (the number of ACh quanta released by a single nerve impulse) is very low, initially about 5% of the adult value. During the 2–3 week period of synapse elimination, quantal release increases at some of the terminals at each NMJ while it declines at others. In general, greater quantal release is associated with a greater chance of a terminal's surviving the competitive process. Since transmitter release correlates strongly with synaptic size, and it is the smallest synapses that are ultimately eliminated, this association is perhaps not surprising. After synapse elimination is complete, the quantal content of the sole surviving input increases progressively, reaching its adult value at about 2 months of age. Part of this increase reflects the increasing size of the nerve terminal. In addition, there is an increase in the quantal release per unit area of membrane. This may be related to the maturation of the population of active zones.

Ca^{2+} Channel Switch

The changes in the effectiveness of quantal release are accompanied by a change in the type of calcium channels that trigger release. At adult mammalian NMJs, release is mediated by P/Q-type channels, as can be shown using selective blockers. In contrast, at the NMJs of newborn mice, both P/Q-type and N-type channels contribute to release. The contribution of the N-type channels is lost within 1–2 weeks after birth.

NMJ Maturation: Postsynaptic

Maturation of the nerve terminal is paralleled by numerous changes in the postsynaptic region of the muscle fiber that contribute to an increase in the speed and efficacy of neuromuscular transmission. This occurs in the context of a substantial increase in muscle fiber diameter. Immature muscle fibers typically have a diameter of 8–10 μm and relatively high electrical input resistance. As a result, relatively little ACh-induced current is required to depolarize them to the action potential threshold. This helps ensure effective neuromuscular transmission even when ACh output from the nerve is still low. As the muscle grows, the diameter of the fibers also grows, and their input resistance falls. A variety of changes in the postsynaptic region help to ensure and enhance the efficacy of transmission in spite of this.

Remodeling of the Postsynaptic Zone

As synapse elimination proceeds, important changes take place in the distribution of key synaptic molecules and in the conformation of the postsynaptic membrane.

Redistribution of AChRs and AChE At birth, the AChRs and AChE occupy more or less uniformly the oval plaque that defines the NMJ. With time, holes in this plaque appear where the density of postsynaptic molecules is relatively low. As synapse elimination is completed and the one surviving terminal begins to expand, the regions with a high density of postsynaptic molecules become closely associated with the nerve terminals, mirroring their pretzel-like shape (Figure 2(b)). In the case of AChRs, this redistribution is associated with a substantial increase in the local density, rising to about 10 000 μm^{-2}. A number of molecules associated with the AChRs, including rapsyn, utrophin, and syntrophin, undergo a similar redistribution. This suggests that there are close links between them and that these molecules may help to stabilize the AChR cluster by cross-linking to the cytoskeleton.

Appearance of voltage-gated sodium channels channels and postsynaptic folds At the mature NMJ, the postsynaptic membrane is highly folded, with the folds extending into the muscle fiber. The voltage-gated sodium channels that account for the action potential, of a type termed Na_V1, are highly concentrated in the depths of these folds and in a perijunctional zone a few micrometers wide (Figure 3(a)). During development in rats, these channels are first detectable around the time of birth, about a week after the AChRs. At this time, the Na_V1 channels are present in highest density in a diffuse band some 200–300 μm^2 wide, centered on the AChR cluster but extending well beyond it and lacking its distinct boundary (Figure 3(b)). Thus, the two types of ion channel that are central to the postsynaptic response to the nerve have very different developmental patterns of expression.

Formation of the folds The postsynaptic folds begin to develop soon after birth (Figure 4). The factors that

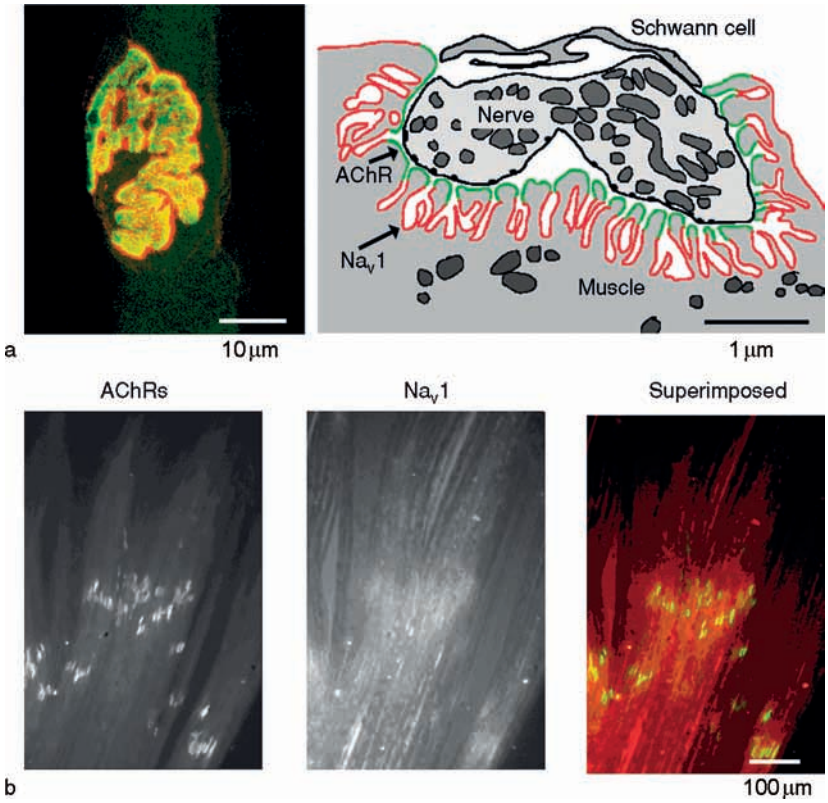

Figure 3 Voltage-gated sodium channels (Na$_V$1) at the mammalian neuromuscular junction (NMJ). (a) Adult rat. Na$_V$1s, shown in red, occupy the depths of the folded postsynaptic membrane and in the light microscope (left panel) appear as a red fringe surrounding the acetylcholine receptors (AChRs), shown in green. (b) Newborn rat. AChRs are concentrated in discrete spots at the immature NMJs while the Na$_V$1s, which are first detected at this age, have a much more diffuse distribution. (a) Reproduced from Slater CR (2003) Structural determinants of the reliability of synaptic transmission at the vertebrate neuromuscular junction. *Journal of Neurocytology* 32: 505–522, with the permission from Springer.

control their distribution and growth are poorly understood. It has been suggested that the opening of the fold represents a site of reduced nerve-muscle adhesion, possibly related to the presence of the active zones in the nerve terminal. However, folds also form in a variety of situations in which AChRs form clusters in the absence of the nerve. These include both denervated and regenerating muscle and muscle exposed to exogenously applied agrin, where AChR clusters associated with folds form at sites away from the NMJ.

From an early stage in fold formation during normal development, while AChRs are concentrated at the crests of the folds, nearest the nerve, the Na$_V$1 channels occupy the depths of the folds. This is likely to result from a high concentration of Na$_V$1 channels in the new membrane added during the process of fold formation combined with a barrier to diffusion of Na$_V$1 channels into the region of high AChR density.

Molecular Differentiation of the Postsynaptic Zone

The remodeling of the postsynaptic region is accompanied by changes in the patterns of isoforms of the ion channels expressed within it. These come about as a result of the combined effects of increasing muscle activity and signaling molecules released from the nerve on gene transcription in the subsynaptic myonuclei.

Changes in AChR expression After birth, the density of AChRs away from the NMJ declines so that in the adult it is generally undetectable. This is the result of suppression of AChR subunit gene transcription by muscle activity, mediated by the binding of myogenic regulatory factors to an E-box sequence in the genes of several AChR subunits. It raises a central question: How is it that AChRs remain at the NMJ itself? In brief, the answer is that agrin released from the nerve and bound to the synaptic basal lamina, through its activation of MuSK, induces activity-resistant transcription of the genes encoding both AChR subunits and a number of other postsynaptic proteins, by the myonuclei in the vicinity of the NMJ (see below).

In addition to the activity-mediated suppression of AChR expression away from the NMJ, there is a change in the type of AChRs expressed at the NMJ

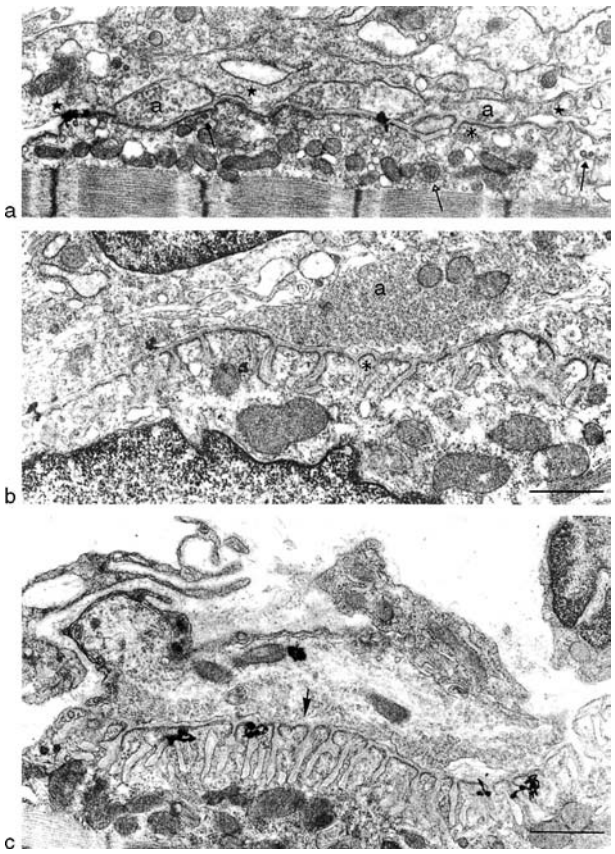

Figure 4 Ultrastructural maturation of mouse neuromuscular junctions (NMJs); development of postsynaptic folds at NMJs in postnatal mouse extensor digitorum longus muscles. (a) In the newborn, few folds are present. Solid arrows, subjunctional cytoplasm rich in coated vesicle and coated pits; open arrows, multivesicular bodies. (b) At 2 weeks, folds are present but less well developed than in adult. (c) In the adult, folds are well formed and closely packed. Solid arrow, junctional folds. a, axon terminal; star, Schwann cell cytoplasm; asterisk, thickened postsynaptic membranes, reflecting accumulation of AChRs and associated protiens. Scale bars = 1 μm. Reproduced from Matthews-Bellinger JA and Salpeter MM (1983) Fine structural distribution of acetylcholine receptors at developing mouse neuromuscular junctions. *Journal of Neuroscience* 3: 644–657. Copyright 1983 by the Society for Neuroscience, with permission.

itself. At birth the predominant fetal form of the AChR has a subunit composition of $\alpha_2\beta\delta\gamma$. During the first 2 weeks after birth, the γ subunit is replaced by an ε subunit to produce the adult form. Expression of the ε subunit is not suppressed by muscle activity. The two forms of AChR differ in their channel properties: The immature form has a longer mean open time (typically 7–8 ms) and a lower conductance (40–50 pS) than the adult form (1–2 ms; 50–60 pS). As a result, a single opening of average duration of a fetal AChR channel allows 3–4 times as much charge to enter the cell as an adult channel. This makes the fetal channels more efficient at converting bound ACh to charge entry. However, the slower kinetics of the currents mediated by fetal AChRs mean that immature NMJs are poorly suited to transmit the high-frequency repetitive activity that occurs in the mature animal. It is thus relevant that the changes in AChR properties occur at about the same time as the speeding up of the firing patterns of the motor neuron (see below) and the myelination of the most distal axons that allows those patterns to reach the NMJ.

Changes in Na_V1 expression An analogous change in isoform expression occurs for the Na_V1 channels. The first Na_V1 channels to appear are of an immature form, designated $Na_V1.5$. These differ from the adult form ($Na_V1.4$) in that they open at more-negative membrane potentials. As a result, less depolarization is required to trigger an action potential in an immature muscle than in a mature one. Expression of the gene encoding $Na_V1.5$, like that of the γ-AChR subunit gene, is suppressed by muscle activity, both at the NMJ and away from it. As a result, expression of $Na_V1.5$ declines to an undetectable level during the first 2–3 weeks after birth. Expression of the gene encoding $Na_V1.4$ is not sensitive to activity. Its expression first becomes detectable at birth and increases during the next few weeks, both away from the NMJ and at a higher level at the NMJ. The factors that control the onset of $Na_V1.4$ expression and its enhancement at the NMJ are not known.

mRNA accumulation The developmental increases in the concentration of AChR and Na_V1 at the NMJ are accompanied by localized increases in the levels of the mRNAs that encode them. At the mature NMJ, there is an increase in the concentration of mRNA encoding a number of critical postsynaptic proteins, including AChR subunits, AChE, and $Na_V1.4$, as well as supporting proteins including rapsyn and utrophin. These increases are apparent soon after birth and are probably induced by agrin acting via MuSK.

Accumulation of Myonuclei

The upregulation of expression at the NMJ of the genes encoding important components of the postsynaptic membrane is now believed to be induced by agrin acting via MuSK. At the NMJ, agrin is fixed in place by association with the synaptic basal lamina. This ensures that the effects of agrin are confined to the vicinity of the NMJ. The effective sphere of influence of this immobilized agrin is not more than about 100 μm, so myonuclei further than this from the NMJ do not have a synaptic profile of gene expression. At many NMJs, there is an accumulation of 5–10 myonuclei within this sphere of influence.

This amplifies the effect of the agrin–MuSK signaling system. The clustering of myonuclei begins soon after birth and is complete by about 1 month of age. The events leading to myonuclear accumulation are not yet understood.

Mammalian Motor Unit Maturation

The formation and maturation of the mammalian NMJ occurs in the context of the development of the motor unit as a whole. The maturation of the motor neuron, and in particular its pattern of firing, has an effect on the development of the NMJ itself. In turn, the increasing ability of the NMJ to transmit faithfully the patterns of activity arising in the motor neuron to the muscle fibers plays an important part in the acquisition of mature muscle properties. It may also lead to retrograde regulation of motor neuron properties.

Firing Patterns

In adult mammals, motor neurons vary considerably in their functional properties. Most notably, some fire action potentials in long trains at a frequency of 10–20 Hz while others fire in short bursts of 5–10 action potentials at frequencies of up to 100 Hz. Motor neurons in neonatal rats and mice fire at a uniformly low frequency and are often synchronized as a result of electrical coupling by gap junctions. The first evidence of faster, more adultlike firing patterns, and of differences between the firing patterns of different motor neurons, is seen about 2 weeks after birth, by which time electrical coupling is lost. The onset of these more adult firing properties coincides with the time of myelination of the most distal intramuscular nerve branches. It is likely that before this happens, these small branches would be unable to fire at high frequencies and could therefore not transmit the features of activity that distinguish different motor neurons in the adult.

Efficacy of Neuromuscular Transmission

During the early stage of polyaxonal innervation, more than one input to the muscle fiber is often capable of triggering contraction. As the competitive process unfolds, some inputs, and ultimately only one, come to have a much stronger impact on the muscle fiber than others. There is evidence that each branch of a motor axon competes locally at a given NMJ and that axons with the strongest input to a given NMJ are more likely to survive the competitive process than weaker ones. However, the evidence on this point is not clear cut, and neither of these findings fully explains the final outcome of the competition, in which a single motor axon, with appropriately matched properties, innervates each muscle fiber.

As the NMJ matures, both the pre- and postsynaptic components increase in size. For the muscle fiber, this results in a decrease in input resistance and therefore the need for a greater synaptic current to reach the action potential threshold reliably. This is achieved by the parallel increase in the size of the motor nerve terminal and with it, the quantal content. This, together with the folding of the postsynaptic membrane, results in the great reliability of neuromuscular transmission in the adult. The nature of the feedback mechanisms that regulate synaptic size and strength is intriguing, however, because one hallmark of the normal adult neuromuscular junction is a three- to fivefold excess of transmitter release over that required to trigger an action potential in the muscle fiber. This high safety factor ensures that every action potential entering a motor nerve terminal will normally trigger an action potential in all the muscle fiber it innervates.

Development of Muscle Fiber Homogeneity

An important consequence of the process of synapse elimination is that each motor neuron ends up innervating muscle fibers with very similar properties. Clear signs of increasing functional homogeneity of the muscle fibers within motor units are seen in mice 1–2 weeks after birth, as synapse elimination nears completion but before the distinctive patterns of activity of different motor units are well developed. It therefore seems unlikely that differences in activity patterns between motor neurons play a decisive role either in selecting which input survives at a given NMJ or in matching the properties of motor neurons to the muscle fibers they innervate. A possible alternative is that the matching of nerve and muscle cells is achieved by a molecular recognition system that involves activity-dependent expression of surface and/or diffusible molecules that interact so that during the process of synapse elimination, the most compatible nerve-muscle pairs survive at each developing NMJ. Such a mechanism could depend on activity as a driving force without the pattern of activity determining the specific outcome of the competition. As yet, however, no likely candidates for such a molecular recognition scheme have been identified.

It is also not yet clear how the properties of a mature muscle fiber are determined. There is good evidence for at least two very different, though not mutually exclusive, mechanisms. On the one hand, there is evidence that myoblasts are predetermined to make fast and slow muscle fibers even before myotube formation occurs. If correct, this mean that myoblasts of the same predisposition may fuse more or less selectively to make myotubes containing nuclei with intrinsically similar properties. On the other

hand, there is very strong evidence, particularly in mammals, that the properties of adult muscle fibers are sensitive to the patterns of activity they experience. This is consistent with the idea that the properties of a muscle fiber are modified after innervation by the pattern of activity of the motor neuron that innervates it. Both schemes raise important questions about how the muscle fibers innervated by a single motor neuron come to have the same properties.

Conclusions

The many changes in the structural, functional, and molecular properties of the NMJ that occur during its maturation adapt it for reliable high-frequency activation of mature muscle fibers. These changes are matched to complementary changes in the nerve and the muscle. Their overall effect is the conversion of an immature system that is good at generating slow muscle contractions in response to low-frequency activity in the nerve to a much faster system, adapted to the needs of a freely moving and increasingly independent animal.

The events that give rise to the mature NMJ are part of a coherent developmental program that defines the patterns of expression of a number of proteins, such as ion channels, that play central roles in neuromuscular transmission. In addition, it determines the size and conformation of the NMJ. Both aspects of the program have important consequences for the efficacy and reliability of the mature NMJ.

See also: Neuromuscular Junction: Neuronal Regulation of Gene Transcription at the Vertebrate; Neuromuscular Junction: Synapse Elimination; Neuromuscular Junction Plasticity in Mammals and Botulinum Toxins; Neuromuscular Junction (NMJ): Acetylcholinesterases; Neuromuscular Junction (NMJ): Postsynaptic Basal Lamina; Neuromuscular Connections: Vertebrate Patterns of; Presynaptic Events in Neuromuscular Transmission; Schwann Cells and Plasticity of the Neuromuscular Junction.

Further Reading

Banks GB, Fuhrer C, Adams ME, et al. (2003) The postsynaptic submembrane machinery at the neuromuscular junction: Requirement for rapsyn and the utrophin/dystrophin-associated complex. *Journal of Neurocytology* 32: 709–726.

Bellinger JA and Salpeter MM (1983) Fine structural distribution of acetylcholine receptors at developing mouse neuromuscular junctions. Journal of Neuroscience 3: 644–657.

Bennett MR and Pettigrew AG (1974) The formation of synapses in striated muscle during development. *Journal of Physiology* 241: 515–545.

Bewick GS, Reid B, Jawaid S, et al. (2004) Postnatal emergence of mature release properties in terminals of rat fast- and slow-twitch muscles. *European Journal of Neuroscience* 19: 2967–2976.

Jansen JK and Fladby T (1990) The perinatal reorganization of the innervation of skeletal muscle in mammals. *Progress in Neurobiology* 34: 39–90.

Landmesser LT (2001) The acquisition of motoneuron subtype identity and motor circuit formation. *International Journal of Developmental Neuroscience* 19: 175–182.

Lichtman JW and Sanes JR (2003) Watching the neuromuscular junction. *Journal of Neurocytology* 32: 767–775.

Marques MJ, Conchello JA, and Lichtman JW (2000) From plaque to pretzel: Fold formation and acetylcholine receptor loss at the developing neuromuscular junction. *Journal of Neuroscience* 20: 3663–3675.

Matthews-Bellinger JA and Salpeter MM (1983) Fine structural distribution of acetylcholine receptors at developing mouse neuromuscular junctions. *Journal of Neuroscience* 3: 644–657.

McMahan UJ (1990) The agrin hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology* 55: 407–418.

Sanes JR and Lichtman JW (1999) Development of the vertebrate neuromuscular junction. *Annual Review of Neuroscience* 22: 389–442.

Sheard PW and Duxson MJ (1996) Composition of newly forming motor units in prenatal rat intercostal muscle. *Developmental Dynamics* 205: 196–212.

Slater CR (1982) Postnatal maturation of nerve-muscle junctions in hindlimb muscles of the mouse. *Developmental Biology* 94: 11–22.

Slater CR (2003) Structural determinants of the reliability of synaptic transmission at the vertebrate neuromuscular junction. *Journal of Neurocytology* 32: 505–522.

Urbano FJ, Rosato-Siri MD, and Uchitel OD (2002) Calcium channels involved in neurotransmitter release at adult, neonatal and P/Q-type deficient neuromuscular junctions. *Molecular Membrane Biology* 19: 293–300.

Neuromuscular Junction (NMJ): Postsynaptic Basal Lamina

M J Werle, University of Kansas Medical Center, Kansas City, KS, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The primary function of skeletal muscle is to produce force. This force is controlled by the motor neuron that contacts the cell. This cell-to-cell contact was best described by Ramón y Cajal as a 'protoplasmic kiss.' The synapse between the motor neuron and the skeletal muscle cell has many names, including motor endplate, neuromuscular synapse, and neuromuscular junction. We will refer to this synapse as the neuromuscular junction. There are three cellular components of the neuromuscular junction – the skeletal muscle cell, the motor neuron terminal (NT), and the terminal Schwann cell (SC) (Figure 1).

Surrounding the muscle cell is an extracellular matrix. The extracellular matrix also covers the Schwann cell. The focus of this article is the specialized extracellular matrix that is found between the nerve terminal and the skeletal muscle cell; this specialized extracellular matrix is called the synaptic basal lamina. The synaptic basal lamina contains many molecules that are commonly found in all basal laminae throughout the organism, but it also contains many molecules that are unique to the neuromuscular junction.

The extracellular matrix is a vital and regulated extension of the cell. The fact that the extracellular matrix is part of the cell is easily understood by the fact that the matrix attaches to a number of membrane-bound structures. These membrane-bound structures in turn attach to the cytoskeleton of the cell. The matrix thus provides mechanical support for the cell. This is particularly important when considering the fact that the muscle fiber is twitching and moving large distances. The motor neuron is very well attached to the surface of the muscle cell that it is innervating. For example, in Figure 1 the motor nerve terminal innervates the muscle cell seen at the bottom of the image. On top of the motor neuron sits the terminal Schwann cell. On top of the Schwann cell is another muscle cell that is regulated by another motor neuron. Thus, the two muscle cells in this image will twitch and move at different times. The extracellular matrix keeps the motor neuron precisely aligned with the muscle cell it innervates.

The precision at which the extracellular matrix aligns the motor neuron and muscle cell is astounding. The active zone is the site on the motor nerve terminal, where the synaptic vesicles fuse with the cell membrane to release the contents of the synaptic vesicle. This active zone is also the site where the voltage-gated calcium channels are concentrated in the membrane of the nerve terminal. The active zones are found precisely opposite the junctional folds in the skeletal muscle cell. The junctional folds are the invaginations in the membrane of the muscle cell. The acetylcholine receptors (AChRs) are concentrated on the crests of the junctional folds, immediately adjacent to the active zones from which ACh is released. The molecules in the synaptic basal lamina lie between these two membranes, and undoubtedly are intimately involved in both the mechanical linking of the two cells, and the cell-to-cell signaling that is essential for maintaining the structure and function of this synapse.

To provide a basis for understanding the synaptic basal lamina, in the following sections we first discuss the molecules that are found in all basal lamina. An important point to emphasize is that in a functional sense, the synaptic basal lamina extends across the nerve and muscle cell membranes to make contact with the cytoskeletons of both cells, thus accounting for the strong mechanical coupling between them. After considering all the extracellular molecules, we therefore discuss the multimolecular complexes that are found in the cell membrane and how they, in turn, are attached to the intracellular cytoskeleton. Finally, we discuss how the synaptic basal lamina is altered by the activities of proteases that cleave the matrix components. Altogether, these features reveal that the synaptic basal lamina is a dynamic cellular component that plays a critical role in the structure and function of the synapse.

Synaptic Basal Lamina Components

Collagen IV

The type IV collagens are critical components of basal laminae throughout the organism, and particularly of the synaptic basal lamina. There are features of collagen IV that are common to all basal laminae, and also specializations that are unique to the synaptic basal lamina.

The α chains of collagen IV have a long collagenous tail that has a glycine at every third amino acid. Characteristic of other collagen proteins there is a non-collagenous domain at the C-terminus; this region is called the NC1 domain. There are six genes that code for collagen IV α chains (*COL4A1* to *COL4A6*). These α chains can form either homo- or heterotrimers.

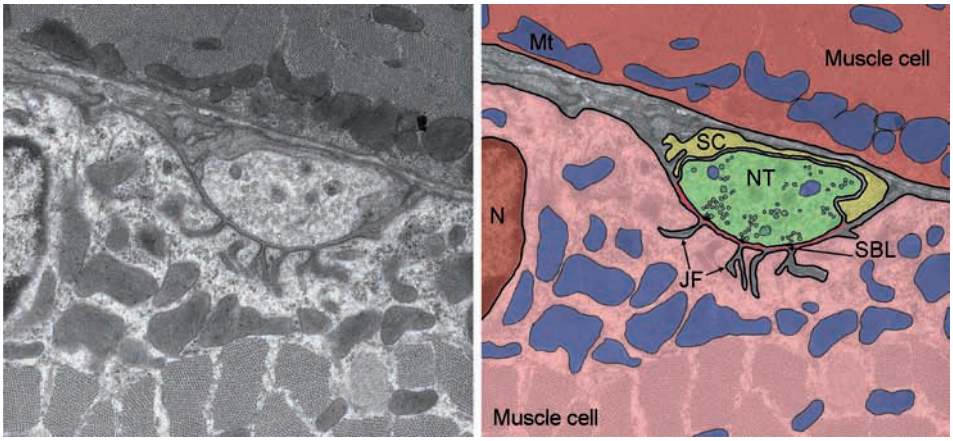

Figure 1 Electron micrograph of a neuromuscular junction from a mouse diaphragm. The left panel is an electron micrograph of a normal mouse neuromuscular junction. The right panel is the same image with the structures colorized and labeled. There are two muscle cells (red and light red). The muscle cell on the bottom is the cell that is innervated by the nerve terminal (green; NT). There is a separate muscle cell that sits on top of the nerve terminal; on top of the nerve terminal is a Schwann cell (yellow; SC). Also seen are the subsynaptic nucleus of the muscle cell (dark red; N) and the mitochondria (blue) in the muscles and in the nerve terminal. Also within the nerve terminal are clusters of synaptic vesicles (dark green). These synaptic vesicles are focused on the active zones (dark black). The active zones are found opposite the mouths of the junctional folds (JF). The synaptic basal lamina (SBL; red) is found in the area between the nerve terminal and the muscle cell on the bottom.

There could therefore be a large number of different combinations, but the only identified groupings are (two α1 and one α2); (α3, α, α5), and (two α5 and one α6). α1 and α2 are found everywhere in the basal laminae, but α3, α4, α5, and α6 are concentrated at the synaptic basal lamina.

The type IV collagens form a meshwork that is critical to the structure of the basal lamina. Three individual collagen IV α chains will combine to form a flexible trimer. The trimer is organized such that the NC1 domains are bundled together at one end, and the N-terminal 7S domains are bundled at the other end. These collagen IV trimers can further associate into larger structures. The N-terminal 7S domains from four trimers will bind to form an X-shaped structure, with the NC1 domains pointed outward (Figure 2). The NC1 domains on collagen IV trimers can then bind head to head with the NC1 domains on a separate collagen IV trimer. The resulting structure is a well-organized and regular meshwork of collagen IV.

The type IV collagens can also associate with other components of the synaptic basal lamina, and with proteins on the surface of the muscle, nerve terminal, and Schwann cell. In the synaptic basal lamina, the most common binding partners are the nidogens, which link the collagen IV meshwork to the laminins. On the cell surfaces, the NC1 domains can interact with integrins, proteoglycans, and the dystroglycan complex. Matrix metalloproteinases can cleave the type IV collagen, and this cleavage results in the release of the NC1 domains. The soluble NC1 domains are known to have potent effects on vascular formation, but the role of the NC1 domain at the neuromuscular junction is still largely unknown.

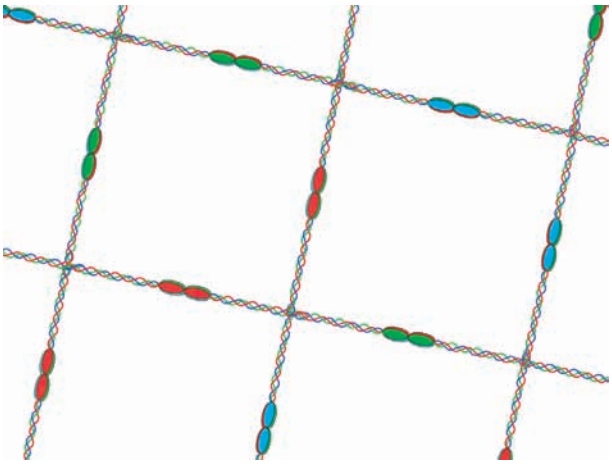

Figure 2 Highly schematic diagram of the type IV collagen meshwork in the basal lamina. Each collagen IV protein has an NC1 domain that is depicted as an oval, and a long collagenous tail. Three subunits assemble into a trimer, with the NC1 domains grouped together, and the collagenous tails intertwined. The NC1 domains from two trimers will interact head to head, and the 7S domains from four trimers will interact to form an X-shaped structure. The individual collagen IV proteins are depicted as either red, green, or blue. This diagram is overly simplified to show a single sheet of the collagen IV matrix.

Laminins

Laminins are heterotrimeric glycoprotein structures. Each laminin is composed of one α subunit, one β subunit, and one γ subunit. There are five genes

that code for α subunits, four genes that code for β subunits, and three genes that code for γ subunits. Of the myriad of possible combinations, currently there are 15 identified laminins. The laminins are important components of basal laminae throughout the organism, and play a special role at the neuromuscular junction.

The skeletal muscle cell produces the laminins found surrounding the skeletal muscle cell in the extrajunctional regions, and also the laminins found in the synaptic cleft. There are strong differences in the distribution of the various laminin subunits, revealing that the release of the laminins by the muscle cell is controlled. The subunits produced by the skeletal muscle cell are predominantly the $\alpha2$, $\alpha4$, or $\alpha5$, plus the $\beta1$, $\beta2$, and $\gamma1$ subunits. Of particular importance to the neuromuscular junction is the $\beta2$ subunit. The $\beta2$-containing laminins are highly concentrated at the synaptic cleft. Since the skeletal muscle cell produces the laminins, the production and release of the $\beta2$ subunit must be tightly controlled. The mechanism is still unknown. The $\beta2$-containing laminins do play an important role in the structure and function of the synapse. Mice that lack the $\beta2$ subunit have defects in synaptic release. The mice die at an early age due to a combination of problems associated with the loss of synaptic function, and also the loss of kidney function. One of the main features observed is the infiltration of the Schwann cell into the synaptic cleft. Thus one of the cellular mechanisms performed by the $\beta2$ subunit may be to stop the Schwann cell from wrapping the nerve terminal. In addition, the $\beta2$ laminin null mutants have disrupted active zones. Thus, the $\beta2$ subunit plays an important role in both the structure and function of the synapse.

The distribution of the α subunit is also controlled by the skeletal muscle cell. The $\alpha2$ subunit is predominantly found in the extrasynaptic basal lamina. The $\alpha5$ subunit is found concentrated at the neuromuscular junction, and it extends into the junctional folds. The $\alpha4$ subunit is found concentrated in a small, specialized region of the synaptic basal lamina. Of particular interest are recent findings that the $\alpha4$ subunit will bind to the voltage-gated calcium channels on the nerve terminal. The distribution and binding of the $\alpha4$ subunit reveal that this laminin chain is likely to play an important role in the communication between the muscle and the nerve terminal.

Nidogens (Entactins)

There are two nidogens in mammalian species. Nidogen-1 is a 150-kDa glycoprotein. It is a common component of all basal laminae, where it acts as a link by binding to both type IV collagens and laminin. Nidogen-2 is found in the basal lamina of skeletal muscle and its distribution mirrors that of nidogen-1. Nidogen-2 will bind to perlecan and collagen IV. Nidogen-2 also weakly binds to laminin-1. Interestingly, mice deficient in nidogen-1 have an apparently normal phenotype, and nidogen-2 is upregulated in the skeletal muscles of these animals. At the neuromuscular junction there is a uniquely glycosylated form of entactin. The size (150 kDa) and the activity of this uniquely glycosylated form of entactin are consistent with its identification as nidogen-1. There is also recent evidence revealing that nidogen can play a role in central nervous system (CNS) synaptic release, thus it is likely that the nidogens play a role at the neuromuscular junction.

Fibronectin

Fibronectin is a large glycoprotein that plays an important role in linking integrins to the extracellular matrix. Fibronectin will bind to a large number of matrix components, and substrates. The important binding partners in the synaptic basal lamina are the collagens and heparan sulfate proteoglycans (predominantly agrin). Fibronectin will then bind to the integrins on the surface of either the muscle cell or the nerve terminal. Fibronectin binding has been shown to alter synaptic release from the motor nerve terminal in culture, and to influence AChR aggregation in cultured myotubes. This activity has been shown to be dependent upon the activation of protein kinases in the cells. Thus, fibronectin acts as a mechanical link between matrix proteins in the extracellular matrix. It also acts as a mechanical link between the matrix and the cell cytoskeleton via the integrin. The binding to integrin also serves as a cell-signaling ligand that increases protein kinase activity. The protein kinase activity has been shown to work predominantly through protein kinases A and C, but it is likely that other protein kinases are also implicated in this signaling.

Heparan Sulfate Proteoglycans

Heparan sulfate proteoglycans (HSPGs) are a family of glycoproteins that share unique features. The molecules contain a protein core with large carbohydrate chains. The carbohydrate chains are sensitive to heparitinase. HSPGs are negatively charged. In their native forms the HSPGs run in bands as an extremely large smear (>200 kDa) when separated by apparent molecular mass in polyacrylamide gels. After heparitinase treatment the core protein is much smaller in size. HSPGs are found in all basal laminae and also associated with the cell surfaces. The three

major HSPGs found in skeletal muscle extracellular matrix are perlecan, agrin, and collagen XVIII.

Perlecan Perlecan has a widespread distribution. It is found in all basal laminae, and plays an important role in the storage and release of growth factors. There are three glycosaminoglycan (GAG) side chains extending form the N-terminal region.

Collagen XVIII Collagen XVIII has been recently shown to be a HSPG and to be present in the basal lamina of skeletal muscle. Loss of collagen XVIII in *Caenorhabditis elegans* has been shown to disrupt neuromuscular junction formation. There are no reports showing that collagen XVIII is concentrated at the vertebrate neuromuscular junction, so it remains unclear whether it has a function there.

Agrin Agrin is the most extensively studied HSPG at the neuromuscular junction, and it plays a pivotal role in the structure and function of the neuromuscular junction. Early experiments performed by UJ McMahan and colleagues were instrumental in revealing the fact that cell-signaling molecules were found in the extracellular matrix. When muscle fibers were damaged, in a way that left the extracellular matrix intact, the damaged portions of the muscle cells would be removed. In addition, the nerve terminal could be removed, leaving behind an empty basal lamina sheath. When the muscle cell regenerated by the proliferation and fusion of skeletal muscle satellite cells, the skeletal muscle would reform within the basal lamina sheath. The regenerated muscle cells would aggregate AChRs on their surfaces precisely at the spot that they contacted the previous synaptic site on the basal lamina. These experiments clearly showed that molecules stably bound to the synaptic basal lamina could direct the aggregation of AChRs on the surface of the muscle cell. Subsequent experiments revealed that the protein agrin could direct the aggregation of AChRs in culture. Mice that lack agrin do not develop neuromuscular junctions, and these mice die at birth.

Agrin is a multifunctional component of the synaptic basal lamina. The C-terminal region of agrin has been most extensively studied, since this is the region that is sufficient to induce the aggregation of AChRs. The number of other molecules that interact with agrin at the neuromuscular junction is extensive. The N-terminal region of agrin has been shown to bind to laminin. Agrin will also bind to α-dystroglycan and thus helps to link the dystroglycan complex to the extracellular matrix. Agrin will also bind to integrins, and to neural cell adhesion molecules (N-CAMs) on the surface of the motor neuron and muscle cell. In addition to these interactions, agrin will also bind to nidogen and collagen. Agrin clearly plays an important role in the structure and function of the neuromuscular junction, and this role is both structural (linking other proteins in the intracellular matrix together) and functional (acting as a cell-signaling molecule).

AChE

The main form of acetylcholinesterase (AChE) at the neuromuscular junction and the synaptic basal lamina is the collagen-tailed form of AChE, ColQ-AChE. ColQ-AChE is bound to the heparan sulfate proteoglycan perlecan and will also interact with the muscle-specific tyrosine kinase receptor. Its main function is to remove ACh from the synaptic cleft. In fact, the ACh released from the nerve terminal must first run the gauntlet of AChE that is bound to the synaptic basal lamina, before it can bind to the AChR concentrated on the crests of the junctional folds.

Cell Surface and Membrane Receptors

Dystroglycan Complex

The main components of the dystroglycan complex are α- and β-dystroglycan. These two forms are from a single gene, but have very different sizes and properties. The α-dystroglycan is 156 kDa in size; it is an extracellular protein that binds noncovalently to β-dystroglycan in the dystroglycan complex, on the one hand, and to laminin and/or agrin in the extracellular matrix, on the other. β-Dystroglycan is a 43 kDa glycoprotein and is an integral membrane protein. Numerous other proteins associate with the dystroglycans and together they form a large structure called the dystroglycan complex. The other members of the dystroglycan complex are the sarcoglycans (α-, β-, δ-, and γ-sarcoglycans), α-dystrobrevin, neuronal NO synthase (nNOS), and the syntrophins (α, β1, and β2). In addition, and of critical importance, the dystroglycan complex will bind to dystrophin or utrophin.

On the intracellular side, the dystroglycan complex will bind to the proteins dystrophin or utrophin. In turn, dystrophin and utrophin will bind to the actin cytoskeleton. The protein utrophin is shorter than dystrophin, and utrophin is concentrated at the neuromuscular junction. The dystroglycan complex is found throughout the entire length of the skeletal muscle cell and is particularly concentrated at the neuromuscular junction and the myotendinous junction. The dystroglycan complex is greatly reduced in Duchenne muscular dystrophy patients, who suffer from muscle fiber breakdown and degeneration. Thus, the dystroglycan complex plays an important

role in linking the intracellular actin cytoskeleton to the extracellular matrix. This link is undoubtedly important in the structural integrity of the muscle cell membrane and is probably critical to protect the membrane from tears that would result from the contractions of the skeletal muscle cell. At the neuromuscular junction the dystroglycan complex may also play an important role in cell-to-cell signaling. The binding of agrin to dystroglycan may be an important feature of the processes that regulate the induction and maintenance of the postsynaptic apparatus in muscle. The exact role of the agrin/dystroglycan complex interaction in postsynaptic signaling is still unknown.

MuSK

The aggregation of AChRs induced by agrin depends absolutely on the presence of a muscle-specific tyrosine kinase (MuSK). MuSK is clearly a downstream mediator of agrin, and MuSK knockout mice have virtually the same phenotype as those lacking agrin. However, the direct binding of agrin to MuSK has not been demonstrated, and the exact choreography of the events that start with agrin and end with the activation of MuSK has not been determined. MuSK has a large extracellular domain and an intracellular kinase domain. More recently it has been found that MuSK is required for the anchoring of the ColQ-AChE complex to the synaptic site. It is also been shown that the Abl kinases (Abl1 and Abl2) are downstream of activated MuSK. Thus, MuSK plays a central role in the events leading to the proper formation of the neuromuscular junction.

Integrins

Integrins are transmembrane protein complexes that form from heterodimers of α and β chains. Both chains are important for ligand binding. There are 18 α subunits and eight β subunits known, and currently 24 unique combinations of these have been identified. In addition, many of the subunits are alternatively spliced so that there are multiple variants of many subunits. Of particular interest is the $\alpha7$ subunit, since it is found concentrated at the neuromuscular junction. The $\alpha7$ subunit interacts with the $\beta1$ to form the $\alpha7\beta1$ integrin. This integrin binds to laminin via the RGD (arginine-glycine-aspartate) domain on the $\beta2$ arm of the synaptic laminin. The $\alpha7$ subunit is alternatively spliced, and there are at least six variants of the $\alpha7$ integrin.

While the binding of the $\alpha7\beta1$ integrin to laminin is a structural link between the matrix and the cell, the integrins also participate in intracellular signaling. On the intracellular side the integrins play a key role in linking to the actin cytoskeleton. This linkage is very similar to the molecular linkage found in a focal adhesion. Namely, the integrin binds to vinculin, α-actinin, and talin. All of these components are found aggregated at the postsynaptic site. In addition, integrins play an important role in regulating the activities of kinases. Of particular interest are the Src kinase family members Abl and Fyn, which have been shown to play a role in the assembly and maintenance of the postsynaptic apparatus. Thus, the integrins play an important structural and signaling role in the postsynaptic apparatus.

Integrins also are likely to play a presynaptic role, since treatment of frog muscles with RGD-containing peptides will reduce the increase in miniature endplate potential release upon muscle stretch. In addition to the $\alpha7\beta1$ integrins, the $\alpha3\beta1$ integrins are found at the presynaptic active zones. The $\alpha3\beta1$ integrins also interact with laminins in the synaptic basal lamina and can activate intracellular kinases. The precise role of the $\alpha3\beta1$ integrin is still largely unknown, but it is clear, at least in frogs, that the binding of integrin to matrix also has an important presynaptic cell-signaling function.

Cadherins

Cadherins are membrane-bound proteins that require calcium for their activity. They play an important role in the early development of the skeletal muscle cell. There are many cadherins, but the main ones produced in adult skeletal muscle are neural (N)-cadherin and muscle (M)-cadherin. Immunolabeling studies have revealed that both N-cadherin and M-cadherin are concentrated at the neuromuscular junction. The role of cadherins at the neuromuscular junction has been largely uninvestigated. N-Cadherin promotes neurite outgrowth *in vitro* and it is of interest that this effect is blocked by agrin. Both N-cadherin and M-cadherin play an important role in the formation of multinucleated muscle cells. They also may play a role at the synapse. The cadherins will form calcium-dependant bonds with cadherins on adjacent cells. Cadherins also interact on the intracellular surface with catenins. The α-, β-, and δ-catenins have been detected at the neuromuscular junction. In other systems it is known that the α- and β-catenins form a link with the actin cytoskeleton. In addition, the liberation of β-catenin from cadherins can result in nucleocytoplasmic shuttling to alter gene transcription via the interaction of β-catenin with T cell factor (TCF). Similarly, δ-catenin has also been shown to be present at the neuromuscular junction and interacts with the promoter region of rapsyn. Altogether, these results argue that the cadherin system functions at the neuromuscular junction to regulate cell structure and function.

Proteases at the Neuromuscular Junction

Matrix Metalloproteinases

The matrix metalloproteinases (MMPs) are a large family of enzymes whose main function is to degrade the extracellular matrix. There are over 25 metalloproteinases, each with preferred matrix substrates. Most MMPs are released from cells in a pro form and must be first activated by other proteases. While it is clear that the MMPs play a role in matrix development and remodeling, it is clear that they also have an influence on cell-to-cell signaling. For example, cleavage of the NC1 domain from type IV collagens is known to play a major role in the signaling of vascular development and remodeling. The strongest evidence for the involvement of MMPs at the neuromuscular junction is the fact that mice that lack MMP3 have increased junctional folds and have AChR receptors on their surface. This likely results from an accumulation of matrix molecules that would normally be removed by MMP3, particularly agrin. MMP2 and MMP9 have been shown to play an important role in the reinnervation of muscle following nerve damage. The fact that MMPs are present at the neuromuscular junction is not a surprise, since it is clear that the matrix must be remodeled to allow synaptic growth. The mechanisms that control MMP activation at the neuromuscular junction are still unknown.

Tissue Inhibitors of Metalloproteinases

Balancing the activity of the MMPs are the tissue inhibitors of metalloproteinases (TIMPs). There are four TIMPs, and one of these, TIMP2, has been shown to be present at the neuromuscular junction. Mice that lack TIMP2 have altered neuromuscular junctions. TIMPs inhibit MMPs, but are also required for the activation of some MMPs. Thus, there is a delicate balance between the activity of MMPs and TIMPs. The control of matrix structure and function is likely to be central to the mechanisms that control synaptic structure and function.

Conclusion

The synaptic basal lamina contains molecules that are common to all basal laminae throughout the body, and also a number of unique molecules. The synaptic basal lamina has both a structural role that links the cytoskeletons of the synaptic components, and a cell-to-cell signaling role. The organization of the synaptic basal lamina is extremely precise, and the maintenance of this structure is undoubtedly important in the organization of the synapse. The synaptic basal lamina is a dynamic structure, and proteases are constantly sculpting this complex matrix. The control of matrix structure and function is likely to be central to the mechanisms that control synaptic structure and function.

See also: Axonal Pathfinding: Extracellular Matrix Role; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction (NMJ): Acetylcholinesterases; Neuromuscular Connections: Vertebrate Patterns of; Neuromuscular Junction (NMJ): Postsynaptic Events in Neuromuscular Transmission.

Further Reading

Barresi R and Campbell KP (2006) Dystroglycan: From biosynthesis to pathogenesis of human disease. *Journal of Cell Science* 119(Pt. 2): 199–207.

Bixby JL, Baerwald-De la Torre K, Wang C, et al. (2002) A neuronal inhibitory domain in the *N*-terminal half of agrin. *Journal of Neurobiology* 50(2): 164–179.

Burkin DJ and Kaufman SJ (1999) The alpha7beta1 integrin in muscle development and disease. *Cell Tissue Research* 296(1): 183–190.

Fox MA and Umemori H (2006) Seeking long-term relationship: Axon and target communicate to organize synaptic differentiation. *Journal of Neurochemistry* 97(5): 1215–1231.

Kummer TT, Misgeld T, and Sanes JR (2006) Assembly of the postsynaptic membrane at the neuromuscular junction: Paradigm lost. *Current Opinion in Neurobiology* 16(1): 74–82.

Miner JH and Yurchenco PD (2004) Laminin functions in tissue morphogenesis. *Annual Review in Cell and Developmental Biology* 20: 255–284.

Ortega N and Werb Z (2002) New functional roles for non-collagenous domains of basement membrane collagens. *Journal of Cell Science* 115(22): 4201–4214.

Patton BL (2003) Basal lamina and the organization of neuromuscular synapses. *Journal of Neurocytology* 32(5–8): 883–903.

Rotundo RL, Rossi SG, Kimbell LM, et al. (2005) Targeting acetylcholinesterase to the neuromuscular synapse. *Chemico–Biological Interactions* 157–158: 15–21.

Strochlic L, Cartaud A, and Cartaud J (2005) The synaptic muscle-specific kinase (MuSK) complex: New partners, new functions. *BioEssays* 27(11): 1129–1135.

Werle MJ and VanSaun M (2003) Activity dependent removal of agrin from synaptic basal lamina by matrix metalloproteinase 3. *Journal of Neurocytology* 32(5–8): 905–913.

Yurchenco PD, Amenta PS, and Patton BL (2004) Basement membrane assembly, stability and activities observed through a developmental lens. *Matrix Biology* 22(7): 521–538.

Neuromuscular Junction (NMJ): Postsynaptic Events in Neuromuscular Transmission

C R Slater, University of Newcastle upon Tyne, Newcastle upon Tyne, UK

© 2009 Elsevier Ltd. All rights reserved.

Essentials of Neuromuscular Transmission in Vertebrates

The neuromuscular junction (NMJ) is the interface between the central nervous system and the muscles it controls. The contraction of muscle fibers is activated by depolarization of the muscle fiber membrane. When motor neuron firing is triggered, action potentials (APs) are initiated near the nerve cell body and then travel along the motor axon to the NMJ. There, the AP causes the release of neurotransmitter molecules that act locally to depolarize the postsynaptic muscle fiber. The ultimate effect of the transmitter is to make the muscle contract. In most vertebrate muscle fibers, an essential intermediate step in the process of neuromuscular transmission is the generation of an AP in the muscle fiber. This AP travels rapidly along the muscle fiber, activating contraction of the whole fiber at nearly the same time. This article describes how vertebrate muscles respond to transmitter released from the nerve and how the APs that activate muscle contraction are generated. In the invertebrates, most muscles do not generate APs and the details of muscle activation are therefore somewhat different. This is also true of a distinct class of electrically inexcitable muscle fibers in vertebrates, discussed briefly at the end of this article.

Most vertebrate twitch muscle fibers are innervated at a single site, often roughly midway along the length of the fiber, by a single motor axon. Following a nerve AP, transmitter released from the nerve causes a local reduction in membrane potential, or depolarization, of the muscle fiber. This brief event (~10 ms) is referred to as the endplate potential (EPP; **Figure 1**), since the NMJ is sometimes called the motor endplate. The EPP is a local event that has its maximum amplitude at the NMJ and declines to a low level within 1–2 mm along the fiber (see later). Since muscle fibers may be more than 10 cm long, some other event must account for the depolarization that activates contraction. This event is the AP (**Figure 1(a)**), which can propagate rapidly in an all-or-none fashion from the NMJ to the ends of the muscle fiber.

APs are generally triggered when the membrane potential of the muscle fiber becomes less negative than about −60 to −55 mV. Thus, from a resting potential of −90 to −70 mV, the membrane must be depolarized by at least 25 mV. In most vertebrates, the EPP is substantially larger than required to reach the AP threshold. This results in very reliable and complete activation of the contractile apparatus, leading to an 'all-or-none' twitch contraction, in each muscle fiber, by every motor nerve impulse.

The postsynaptic events of neuromuscular transmission occur within the highly specialized region of the muscle fiber known as the postsynaptic apparatus. This region contains very high concentrations of the ion channels that mediate the response of the muscle fiber to the transmitter released from the nerve. In many vertebrates, there are also striking structural features of this region, including a highly folded surface membrane and an accumulation of specialized myonuclei with distinctive patterns of gene expression.

AChRs and the Generation of the Endplate Current

The neurotransmitter acetylcholine (ACh) is released from the vertebrate motor axon terminal in multimolecular quanta. The ACh causes a depolarization by binding to, and thus opening, cation-selective ion channels in the muscle fiber membrane known as ACh receptors (AChRs; **Figures 2(a) and 2(b)**). When these channels open, the internal negativity of the muscle fiber drives the net entry of positive ions into the cell. It is this effect that gives rise to the depolarizing phase of the EPP.

The AChRs present at the vertebrate NMJ are pentameric complexes of four types of subunits (**Figure 2(a)**). While the structures of these subunits are broadly similar there are also important differences and each subunit is encoded by a different gene. In the muscles of some lower vertebrates, and in immature or paralyzed mammalian muscles, each AChR molecule has two α-subunits and one each of β-, δ-, and γ-subunits. In many vertebrates, including mammals, the γ-subunit is largely replaced by an ε-subunit. Each subunit has a molecular mass of about 50 kDa, so the entire complex has a molecular mass of about 250 kDa. Every AChR molecule has two ACh binding sites, each of which includes part of one α-subunit.

When no ACh is bound, the AChR channel is closed and no current flows through it. To open the channel, two ACh molecules must bind, one to each binding site. When this happens the channel opens rapidly (**Figure 2 (b)**). The open channel has a conductance of about 30–50 pS. In the ionic conditions found in an intact mammal, the entry of positive charge through this

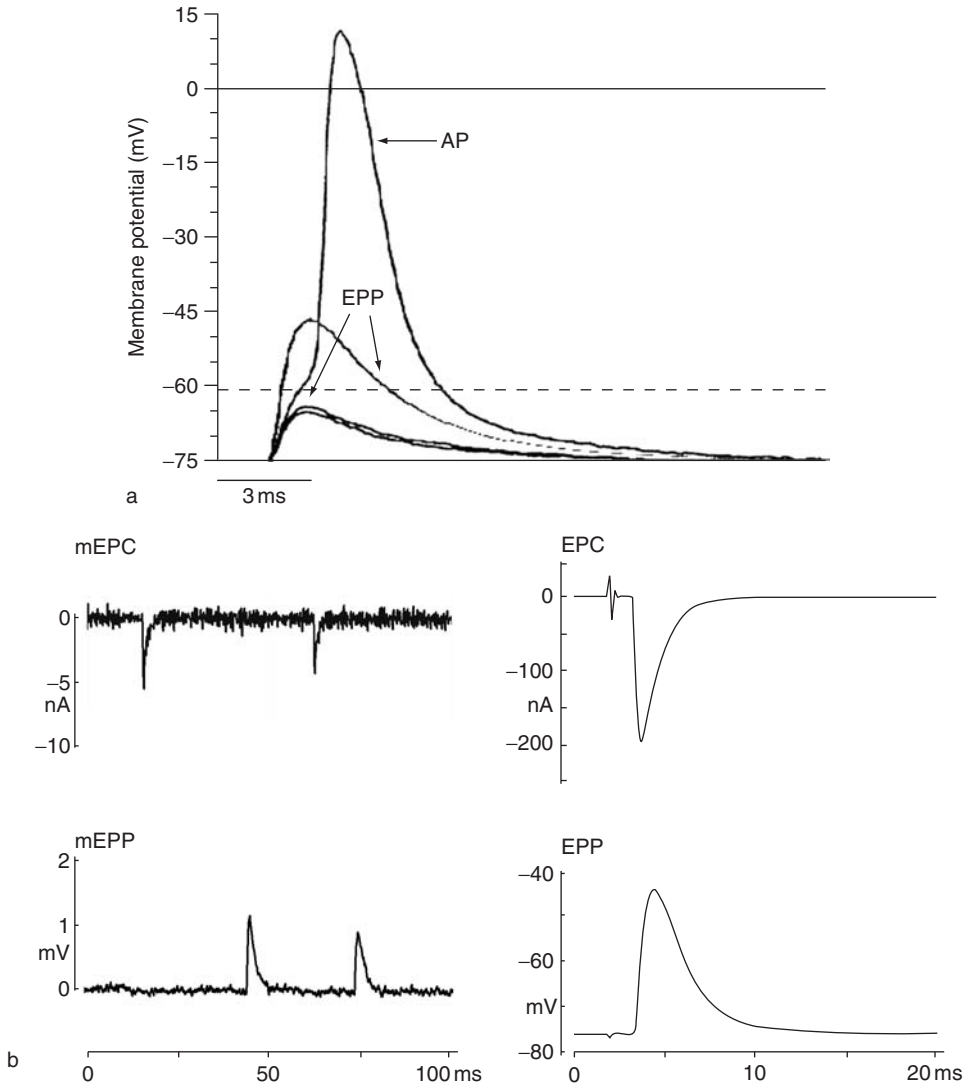

Figure 1 Electrical events associated with transmitter action at the vertebrate NMJ. (a) Intracellular recordings of endplate potentials (EPP) and action potentials (AP). When the EPP fails to reach the AP threshold (horizontal dashed line) no AP is generated. The smaller EPPs illustrated were recorded during partial block of AChRs by curare. The larger EPPs illustrated were recorded during block of Na_V1 channels in the muscle fiber by a cone snail toxin. These EPPs would normally trigger an AP. (b) Miniature endplate currents (mEPC) result from the flow of current into the muscle fiber induced by a single quantum of ACh. The nerve impulse causes the release and action of many quanta, resulting in the much larger endplate current (EPC). The electrical properties of the muscle fiber convert these currents into equivalent changes in membrane potential, the miniature endplate potential (mEPP) and the larger endplate potential (EPP). Note that the potential changes are slower than the currents that give rise to them, a result of the membrane capacitance.

conductance represents an 'inward' current of about 2.5–4 pA (Figure 2(c)). After being opened, each channel closes spontaneously and abruptly, on average after about 1 ms (for AChRs containing an ε-subunit). However, the duration of open times, even of the same channel on different occasions, varies considerably and randomly, with brief openings being the most common and long open times the least common.

At the vertebrate NMJ, each ACh quantum consists of 5000–10 000 molecules. Each quantum causes the opening of 1000–2000 ACh-gated channels, resulting in a brief influx of a current of about 3–5 nA into the muscle fiber. The current induced by a single quantum is known as a miniature endplate current (mEPC; Figure 1(b)). Although all the channels activated by a single quantum of ACh open almost simultaneously, the variability of the open times means that some close very quickly and others remain open for several milliseconds. As a result, the shape of the mEPC differs from that of the individual channels, having an abrupt rise to a maximum value and a slower exponential decay with a time constant of 1–1.5 ms (Figures 1(b) and 2(d)). In the case of AChRs containing a γ-subunit (see earlier), the time constant is severalfold longer.

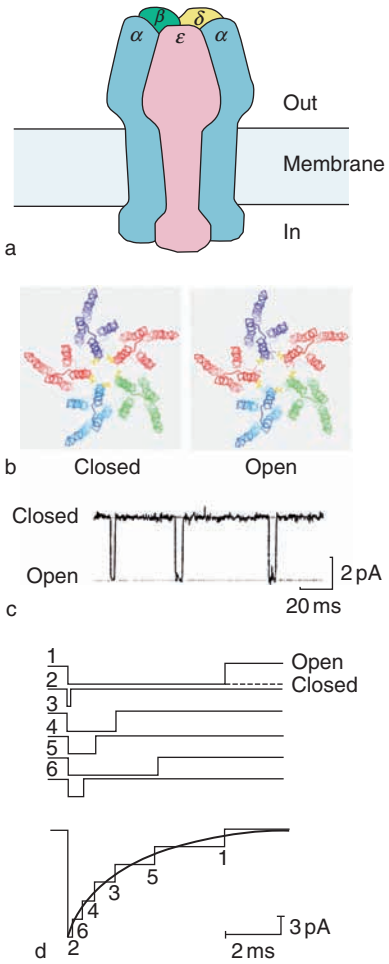

Figure 2 Acetylcholine receptors (AChRs) mediate permeability changes in the postsynaptic membrane. (a) The five subunits of the AChR associate to form a pore that spans the plasma membrane. (b) A model of the open and closed states of the AChR as viewed looking along the pore axis. Small relative movements of the subunits, induced by ACh binding, result in changes in the diameter of the ion pore that have a big impact on permeability. (c) Record of current flowing through a single AChR channel. Each event has an abrupt start and finish. Although the duration of the open states varies, the intensity of the current in different events is very similar. (d) Scheme of how single channel openings sum to give rise to a mEPC. In the upper part, each of six channels is shown to open simultaneously. Each channel stays open for a time that varies randomly about a mean value. Thus the six channels close at different times. In the lower part is shown a model mEPC consisting of the sum of the currents flowing through the six channels. It decays exponentially from the initial peak with a time constant equal to the duration of the mean open time of the individual channels. (b) Diagram provided by Nigel Unwin. (c) Reprinted by permission from Macmillan Publishers Ltd: [Nature] Mishina M, Takai T, Imoto K, et al. (1986) Molecular distinction between fetal and adult forms of muscle acetylcholine receptor. *Nature* 321: 406–411, copyright (1986).

At the NMJ of vertebrate twitch muscles, between 20 and 200 quanta are typically released by a single nerve AP. In general, the sites of release of these quanta are distributed randomly over the entire NMJ. The AChR channels opened by a single quantum do so within less than 100 μs. During this time ACh molecules, on average, can diffuse about 0.3 μm. Each quantum thus acts on about 0.25 μm^2 of postsynaptic membrane. For most vertebrate NMJs, this is less than 0.1% of the membrane containing AChRs. As a result, the sites of action of individual quanta on the muscle fiber membrane are effectively separated from each other, allowing the effects of the individual quanta to sum to produce the much larger endplate current (EPC; Figure 1(b)). In different vertebrate species, the EPC may vary in amplitude from 200 to 2000 nA. Since the quanta released by a single nerve AP act almost simultaneously on the muscle, the time course of the EPC is usually very similar to that of the mEPCs.

From EPCs to EPPs

The passive electrical properties of the muscle fiber determine how the current of the EPC is converted into the depolarization that is the EPP. A brief account of these properties is given in the following sections.

Passive Cable Properties of Muscle Fibers

The flow of current in muscle fibers is governed by the so-called cable properties of the cell. These derive from the relatively high electrical conductivity of the cytoplasm and the low conductivity (or high resistivity) of the plasma membrane. Current entering the cell at a point, such as the NMJ, may flow longitudinally in the cytoplasm or transversely through the membrane (Figure 3(a)). Only the latter pathway leads to a change in the membrane potential. The fraction of the current that takes each route depends on the relative resistance each pathway presents. In practice, the conductivity of the cytoplasm is roughly the same in most cells and the resistivity of the membrane is similar in many (but not all) skeletal muscle fibers. As a result, the dominant factor determining the distribution of current flow in different muscle fibers is their diameter. When the diameter is large, current can flow much more easily along the muscle fiber than across the membrane, and there is relatively little voltage change across the membrane at the site of current entry. In contrast, if the diameter is small, the resistance to longitudinal current flow is high and more current crosses the membrane close to the site of entry, giving rise to a larger local change in voltage.

The peak amplitude of the potential change caused by a given intensity of current flow entering the cell at a single point is determined by the so-called input resistance of the muscle fiber. Typical vertebrate skeletal muscle fibers have 'input resistances' of

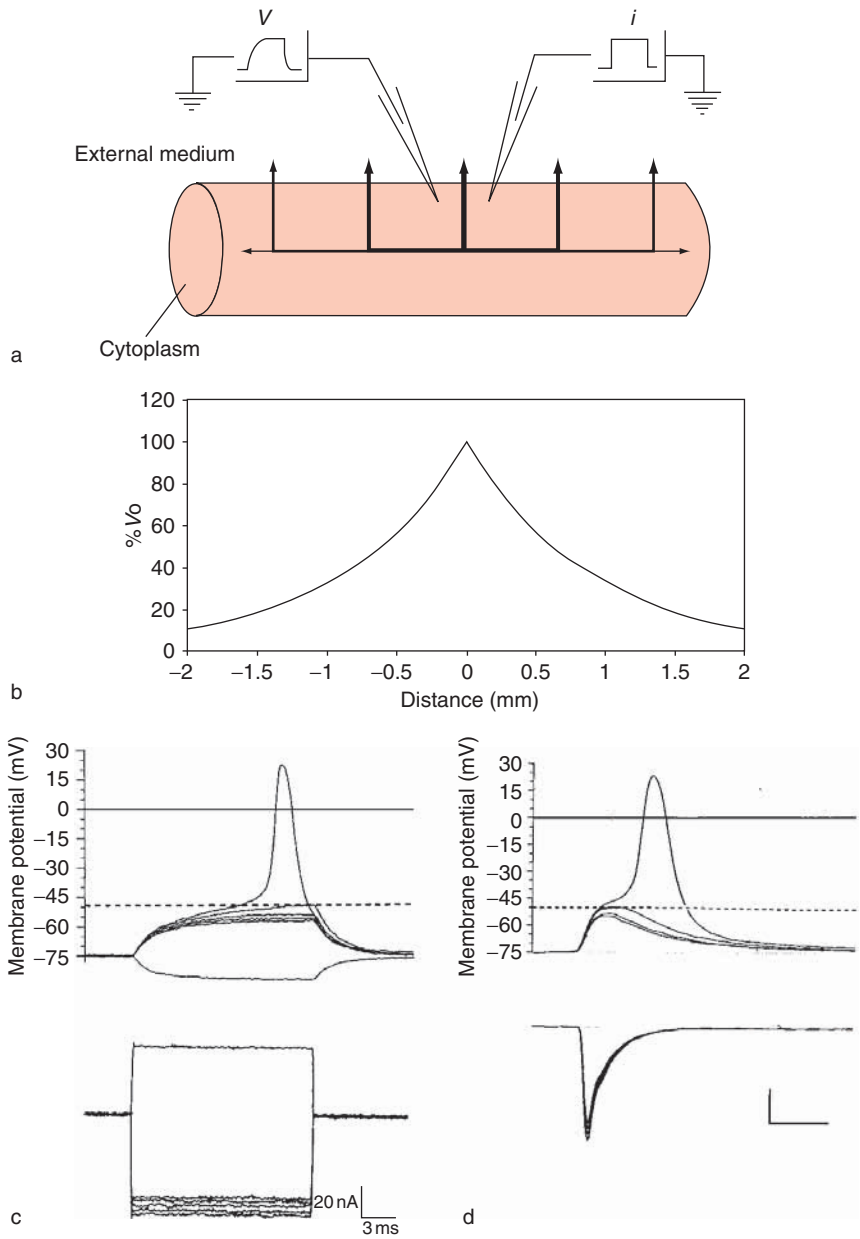

Figure 3 The passive electrical properties of the muscle fiber shape the voltage change caused by current flowing across the membrane. (a) Diagram of a typical experiment. Two electrodes are placed within the muscle fiber, one to pass current into the cell (*i*) and the other to measure the resulting change in transmembrane potential (*V*). Arrows show flow of current. The current across the membrane is greatest near the site of current injection and decreases with distance from that site. (b) Decay of transmembrane potential (*V*m) with distance from the site of injection. The space constant (λ) is the distance at which the potential has fallen to 36% (1/*e*) of its maximum (*V*o). (c) Changes in membrane potential (upper part) in response to rectangular current pulses. Note the slower change in voltage due to the capacitance of the membrane. Horizontal dashed line shows AP threshold: depolarizations greater than this trigger an AP. (d) Potential changes in response to injection of current with a time course similar to that of the postsynaptic mEPC and EPC that occur during neuromuscular transmission. Note that the passive properties of the membrane convert these currents into potential changes very similar to mEPP and EPP. The action potential threshold is very similar to that observed with rectangular pulses. (c) Reproduced from Wood SJ and Slater CR (1995) Action potential generation in rat slow- and fast-twitch muscles. *Journal of Physiology* 486: 401–410, with permission from Blackwell Publishing. (d) Reproduced from Wood SJ and Slater CR (1997) The contribution of postsynaptic folds to the safety factor for neuromuscular transmission in rat fast- and slow-twitch muscles. *Journal of Physiology* 500: 165–176, with permission from Blackwell Publishing.

0.2–1 MΩ (see later). If the full amplitude of the EPC (e.g., 250 nA) were instantly converted into a depolarization, its value (for a resistance of 0.4 MΩ) would be 100 mV.

Spatial factors The decay of potential from a site of current injection (e.g., the NMJ) can usually be approximately described by a single exponential (Figure 3(b)). For many vertebrate muscle fibers, the

space constant of the decay – that is, the distance required for the potential to drop to 37% (1/*e*) of its value at the origin – is 0.5–2 mm. It is because of this that the ACh released from the nerve has little direct effect on the membrane potential more than a few millimeters away from the NMJ.

Effects of membrane capacitance Like all cell membranes, the muscle fiber membrane has electrical capacitance as well as resistance. When positive charge enters the cell through the opened AChRs, it must bring about a reorientation of the charge stored in the molecular dipoles in the membrane before a potential difference develops across the membrane. This takes time. For an instantaneous change in current flow across the membrane of a mammalian muscle fiber, the voltage typically takes 3–5 ms to reach 84% of that, a measure of the so-called time constant of the muscle fiber (Figure 3(c)).

For a flow of current which is brief relative to the membrane time constant, such as the mEPC, the capacitance distorts the time course of the resulting voltage change (Figure 3(d)). In addition, because the mEPC is short relative to the time constant of the membrane, the potential never reaches the maximum value that corresponds to the current at the peak of the mEPC. In practice, the distorting effect of the membrane capacitance means that the observed peak amplitude of the EPP is typically about 30–40 mV.

Initiation of the Muscle AP

Most vertebrate skeletal muscle fibers, like neurons, have voltage-gated sodium channels in their membranes. The specific forms of these channels present in vertebrate skeletal muscle are of the class designated as Na_V1 and are very similar in their structure and function to those in other excitable tissues. However, they differ in detail and are encoded by different genes which are normally expressed only in muscle cells. Each Na_V1 channel contains a major pore-forming α-subunit with a molecular mass of about 250 kDa and a smaller accessory β-subunit. Like most voltage-gated ion channels, Na_V1 channels are largely closed at the normal resting membrane potential and open when the membrane potential becomes sufficiently less negative. In the case of Na_V1 channels in mammalian muscle fibers, opening begins at a membrane potential of about −60 mV and is maximal at about −30 mV.

In a vertebrate twitch muscle fiber, the action potential is normally initiated at a single point along its length, the NMJ. In order for this to happen, the charge entering at that point, and the depolarization it gives rise to, must open sufficient nearby Na_V1 channels so that the inward current entering through them outweighs the flow of longitudinal current away from the NMJ and thus brings the surrounding membrane to threshold. In muscle fibers, this threshold is conventionally measured away from the NMJ using two closely spaced (<100 μm) intracellular electrodes, one to pass current and the other to record the resulting change in membrane potential (Figure 3(a)). In rat skeletal muscle fibers in which the resting potential is maintained at −80 mV, the threshold recorded in this way, using a current pulse that is long relative to the AP, is about −55 mV (Figure 3(c)). In other words, a depolarization of about 25 mV is required to initiate an AP. A similar threshold depolarization is observed when a current pulse which has a time course and amplitude similar to that of an EPC is used (Figure 3(d)).

A number of observations in mammalian muscle suggest that the effective threshold at the NMJ is significantly less than in the extrajunctional region. For example, when the amplitudes of the EPP and EPC are progressively decreased by adding the AChR blocker curare, EPPs as small as 12 mV (starting from a resting potential of −80 mV) can just trigger an AP. To understand the basis of this effect, it is necessary first to consider the structure and molecular organization of the postsynaptic region.

Structure of the Postsynaptic Apparatus

A striking feature of the postsynaptic region at many, but by no means all, vertebrate NMJs is the presence of infoldings of the muscle fiber membrane (Figure 4). These are typically 0.5–1 μm deep, 0.1 μm wide, and spaced 0.5–2 μm apart. Detailed investigation of mammalian NMJs, where the folds are particularly prominent, has shown that there is a remarkable degree of segregation of ion channels within the postsynaptic membrane.

At the crests of the folds, close to the overlying motor nerve terminal, AChRs are clustered at a density of approximately 10 000 μm^{-2}. This means that the individual AChR molecules are virtually in contact with each other in a near-crystalline array. The region of high AChR density typically extends about one-third of the way down the fold, where it ends abruptly. Outside this region, the density of AChRs is generally no more than a few per square micrometer.

In contrast to the membrane at the crests of the folds, that in the depths of the folds, and in a perijunctional zone extending about 20–30 μm away from the NMJ, contains a high density of Na_V1 channels. While the absolute density of these channels is not known accurately, it is estimated to be about

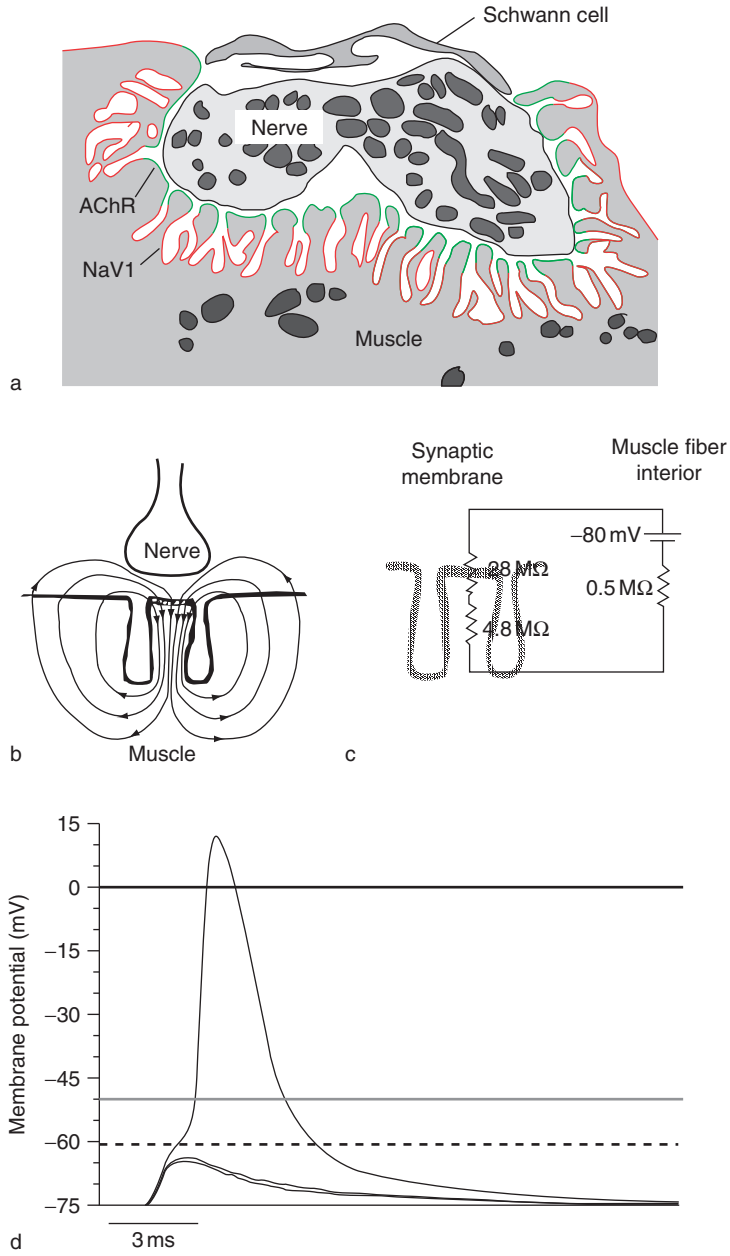

Figure 4 The postsynaptic apparatus and its amplifying effect. (a) A schematic diagram of a mouse NMJ. Note particularly the highly folded postsynaptic membrane and the nonoverlapping distributions of acetylcholine receptors (AChRs) and Na_V1 molecules. AChRs are concentrated near the sites of transmitter release from the nerve while the Na_V1 channels are concentrated in the depths of the folds. (b) Current flow in the muscle fiber induced by the action of ACh. Positive charge enters the muscle fiber through opened AChR channels at the crest of the folds. The resulting current flows in the narrow space between the folds into the cytoplasm and then out across the nonsynaptic membrane. (c) Equivalent circuit of an active fold at a rat NMJ. Note that the resistance of the interfold cytoplasm (4.8 MΩ) is an order of magnitude greater than that of the input resistance of the fiber (0.5 MΩ). (d) AP threshold during neural activation of a rat muscle fiber. The EPP was reduced in amplitude by partially blocking AChRs with curare so that only occasional EPP are big enough to trigger an AP. The AP threshold determined this way (dashed line, about −60 mV) is substantially less than when determined when current is injected away from the NMJ (solid line, about −50 mV; see **Figure 2(a)**). This is a result of the combined effects of the high concentration of Na_V1 channels and the high interfold resistance.

$2000\,\mu m^{-2}$. This is some 10 times greater than in the remainder of the muscle fiber membrane.

These features of the postsynaptic membrane of the mammalian NMJ have an important impact on the current flows initiated by ACh released from the nerve. The current flowing through the opened AChRs enters the cell at the crest of the folds (**Figure 4(b)**). Because the effective resistance of the folded membrane

is very high, the path of least resistance for this current is through the thin sheet of cytoplasm between the folds. However, the resistance of this pathway is some 10 times greater than the input resistance of the muscle fiber as a whole (Figure 4(c)). As a result, there is a steep potential gradient within the cytoplasm between the top of the folds and that in the bulk of the muscle fiber. This results in a similar spatial gradient of the transmembrane potential. At NMJs in the rat, it has been estimated that a single quantum of ACh depolarizes the membrane at the crest of a fold by more than 10 mV, while in the bulk of the fiber the depolarization recorded by an intracellular electrode is only 0.5–1 mV.

The functional significance of this local amplification of the depolarizing effect of ACh is enhanced by the high density of Na_V1 channels in the depths of the folds. To initiate an AP, a minimum number of Na_V1 channels at the NMJ must be opened. As a result of the high density of these channels in the postsynaptic membrane, a smaller fraction of the channels needs to open to generate an AP than if the density were as low as it is in the extrajunctional region of the fiber. In addition, the amplifying effect of the geometry of the folds means that a larger fraction of those channels is opened by a quantum of ACh than if the membrane was not folded. As a result, the effective threshold for AP generation at the NMJ is about half of that in the extrajunctional region (Figure 4(d)).

Safety Factor for AP Generation

The pattern of muscle innervation varies in different species and muscles. In some cases, multiple axons innervate each muscle fiber and no AP is generated (see later). However, in most higher vertebrates, and in mammals in particular, an important principle of the functional organization of muscle innervation is that each AP in a motor neuron causes contraction of every muscle fiber that its axon innervates. This requires that neuromuscular transmission is very reliable, even during intense high-frequency activity when fewer quanta are released per impulse. This reliability results from the release of substantially more quanta of ACh than are necessary to reach the threshold for AP generation in the muscle. Neuromuscular transmission is thus said to have a high 'safety factor.' In rat limb muscles, where it has been studied in most detail, the safety factor is about 4.

It is clear that the postsynaptic specializations just described make an important contribution to the safety factor in mammalian muscle fibers. By lowering the effective threshold they reduce the number of quanta that must be released from the nerve to ensure reliable transmission.

Termination of ACh Action

The action of ACh is terminated by the enzyme acetylcholinesterase (AChE), which is bound to the basal lamina located in the space between the nerve and the muscle. AChE cleaves the ACh into choline and acetate. Although the AChE lies on the diffusion path taken by ACh molecules on the way to their targets, there are many fewer AChE molecules than ACh or AChRs. As a result, most ACh molecules released from the nerve reach the postsynaptic membrane and bind to AChRs. However, when they subsequently dissociate from the AChRs, the ACh molecules are bound and cleaved by the AChE, thus preventing them from rebinding to a second AChR. AChE thus acts to limit both the spatial extent and temporal duration of ACh action.

How Different Species Achieve Reliability

Not all vertebrate species achieve an adequate safety factor in the same way. While mammals rely on the amplifying effect of the postsynaptic folds, many fish and birds have few or no folds. Comparison of the quantal content and intensity of folding in several species suggest that there is often a reciprocal relationship between these two properties of the NMJ (Figure 5). In frogs, the nerve terminals are very big and release many quanta, but the folding is less prominent than in mammals. At the other extreme are human NMJs, which are relatively small and release relatively few ACh quanta (20–50 vs. 100–200 in the frog) but have very extensive folding. The NMJs of rats and mice are of intermediate size and intensity of folding. There is little reliable information about the number of quanta released at avian NMJs. It is also possible that the safety factor is not as high in birds as in some other species.

It is unclear why some species rely more on presynaptic release and others on postsynaptic amplification to achieve an adequate safety factor. In large animals such as humans, each motor neuron innervates many more muscle fibers than in small animals. It may be that cellular economy dictates that the total volume of nerve terminals supported by a single motor neuron should be kept to a minimum. The postsynaptic specializations that increase the efficiency of transmitter action may have evolved in response to such pressures.

Some Special Cases

In the great majority of vertebrate muscle fibers, APs are generated as part of the process by which contraction is activated. However, there are some important exceptions to this rule. In many species, muscle fibers

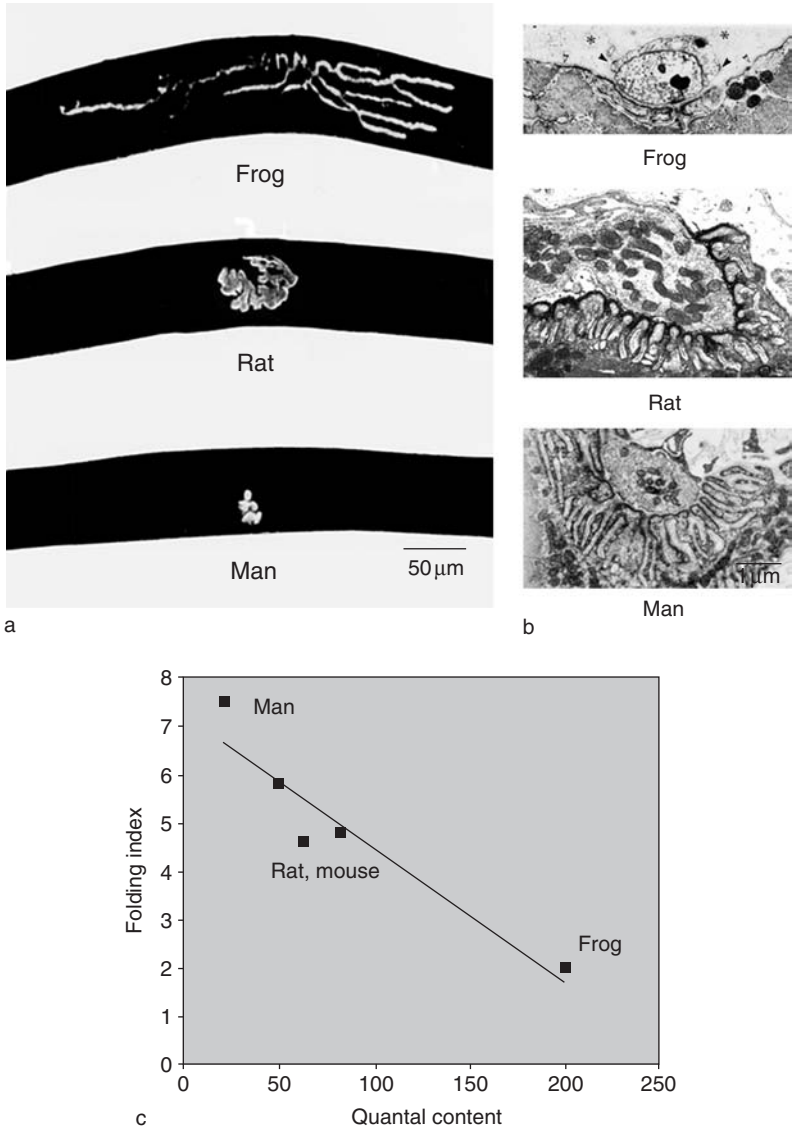

Figure 5 Inverse relationship between the extent of folding and the amount of transmitter release. (a) Light micrographs of NMJs in the indicated species. The muscle fiber is shown in black and the NMJ in white. Note small size of the human NMJs relative to those in frogs. (b) Electron micrographs of NMJs in the same species. Note that human NMJs have much greater postsynaptic folding than do frog NMJs. (c) In these species, the extent of folding (reflected by the folding index) is inversely related to the number of transmitter quanta released from the nerve. This provides the structural basis for the postsynaptic amplification of ACh action that is particularly prominent in humans.

are present that appear to be incapable of generating APs. These fibers tend to have small diameters and very high membrane resistance, and to contract relatively slowly. A common feature of these fibers is that they are innervated at multiple sites along their length, typically at intervals of 1–2 mm. Contraction is activated by the summed effect of the local depolarizations caused by transmitter quanta acting at these sites. These fibers are referred to by a number of terms, the most common being 'slow' and 'tonic.' Unfortunately, both of these terms are used in a number of different ways, leading to much confusion in the literature. The term 'nonEx-MIF' (nonexcitable, multiply innervated fibers), though cumbersome, is more precise.

Frog 'Slow' Fibers

In lower vertebrates, nonEx-MIFs are found in many skeletal muscles. In frogs, the vertebrates in which this type of NMJ was first studied in detail, stimulation of the nerve elicits junction potentials, analogous to EPP. Analysis of the properties of these potentials with intracellular electrodes is greatly complicated by the presence of multiple sites of innervation of individual muscle fibers, and the contamination of recordings at one site by events at others. Nonetheless, it seems that AChRs with kinetics and pharmacology similar to those in twitch fibers are concentrated at these junctions and account for the increased

conductance that underlies the junction potentials. A common feature of electrically inexcitable muscle fibers, both in frogs and in mammals (see later), is that they have a very high input resistance. This is probably a result of a low resting permeability to chloride ions. An important consequence of this is that the depolarization caused by the action of a quantum of ACh is much greater than it would be in a typical twitch fiber. This is of obvious importance in the activation of contraction in the absence of APs.

Mammalian Extraocular Fibers

In higher vertebrates, including mammals, non-Ex-MIFs are mainly restricted to a few specialized types of muscles. The extraocular muscles include fibers that receive innervation from several motor neurons. In contrast to the analogous fibers in frogs, these fibers contract very rapidly in response to the graded stimulus provided by their multiple innervations. Many of these fibers appear unable to generate APs, although a subclass shows some degree of excitability *in vitro*. In these intermediate fibers, the level of excitability seems to vary along the length of the fiber, being greatest in the central region. Very little is known in detail about the quantitative aspects of ACh action or voltage-gated sodium channel activation in these very small diameter muscle fibers.

Intrafusal Fibers

The second main class of non-Ex-MIF in higher vertebrates is the intrafusal fibers of the muscle spindles. The motor innervation of these fibers is complex, arising mainly from specialized γ-motor neurons, but in some cases from collateral branches of the α-motor neurons that also innervate extrafusal fibers. Very little is known in detail about neuromuscular transmission in these fibers, which are even more challenging to investigate than those of the extraocular muscles.

Conclusions

At vertebrate NMJs, the postsynaptic response to ACh quanta released from the nerve is determined primarily by two key ion channel species and the geometry of the postsynaptic region. AChRs mediate the initial influx of excitatory depolarizing current in response to ACh binding. This current is converted into a depolarization by the passive electrical properties of the muscle fiber. In the case of those species in which there are prominent postsynaptic folds, these focus the depolarizing effect on the folded membrane, in a manner analogous to the isolating effect of the neck of dendritic spines on central neurons. The depolarization acts to open Na_V1 channels that are concentrated in the postsynaptic region. When enough of these channels open, a regenerative AP is initiated.

See also: Acetylcholine Neurotransmission in CNS; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction (NMJ): Acetylcholinesterases; Neuromuscular Junction (NMJ): Postsynaptic Basal Lamina; Neuromuscular Connections: Vertebrate Patterns of.

Further Reading

Chiarandini DJ and Davidowitz J (1979) Structure and function of extraocular muscle fibers. *Current Topics in Eye Research* 1: 91–142.

Martin AR (1994) Amplification of neuromuscular transmission by postjunctional folds. *Proceedings of the Royal Society of London, Series B* 258: 179–187.

Mishina M, Takai T, Imoto K, et al. (1986) Molecular distinction between fetal and adult forms of muscle acetylcholine receptor. *Nature* 321: 406–411.

Morgan DL and Proske U (1984) Vertebrate slow muscle: Its structure, pattern of innervation, and mechanical properties. *Physiological Reviews* 64: 103–169.

Salpeter MM (1987) Vertebrate neuromuscular junctions: General morphology, molecular organization, and functional consequences. In: Salpeter MM (ed.) *The Vertebrate Neuromuscular Junction*, pp. 1–54. New York: Alan R. Liss, Inc.

Slater CR (2003) Structural determinants of the reliability of synaptic transmission at the vertebrate neuromuscular junction. *Journal of Neurocytology* 32: 505–522.

Slater CR, Lyons PR, Walls TJ, et al. (1992) Structure and function of neuromuscular junctions in the vastus lateralis of man. A motor point biopsy study of two groups of patients. *Brain* 115(Pt. 2): 451–478.

Vautrin J and Mambrini J (1989) Synaptic current between neuromuscular junction folds. *Journal of Theoretical Biology* 140: 479–498.

Wood SJ and Slater CR (1995) Action potential generation in rat slow- and fast-twitch muscles. *Journal of Physiology* 486: 401–410.

Wood SJ and Slater CR (1997) The contribution of postsynaptic folds to the safety factor for neuromuscular transmission in rat fast- and slow-twitch muscles. *Journal of Physiology* 500: 165–176.

Wood SJ and Slater CR (2001) Safety factor at the neuromuscular junction. *Progress in Neurobiology* 64: 393–429.

Neuromuscular Junction (NMJ): Presynaptic Non-Quantal Release of Transmitter

F Vyskočil, Charles University, Prague, Czech Republic

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During neuromuscular transmission in vertebrates, acetylcholine (ACh) is released from the nerve in multimolecular 'quanta.' Each quantum is believed to represent the 5000–10 000 ACh molecules contained within a single synaptic vesicle. In addition to this quantal release, there is considerable evidence for a form of continuous molecular leakage of ACh from the nerve terminal. The evidence for this nonquantal release (NQR), and for its possible physiological significance as a regulator of muscle fiber excitability at the neuromuscular junction (NMJ), is the subject of this article.

Early Observations

The first electrophysiological evidence for NQR was obtained from cat muscles. A depolarization of the endplate zone of several millivolts was observed after injection of an anticholinesterase (anti-AChE) into the jugular vein. This depolarization developed as soon as 70–80 s after anti-AChE treatment and extended most of the way along the muscle (Figure 1(a)), with peaks in the regions of the NMJs. It was abolished after denervation, suggesting that it was of neural origin. The depolarization was not accompanied by muscle twitching, which seems to exclude the participation of action potentials in the nerve terminal or increased quantal release. Similar observations were subsequently made by us on isolated strips of mouse diaphragm (Figure 1(b)), where the depolarization arises in the region of the muscle corresponding to the single narrow band of NMJs. Subsequent studies found that the amount of ACh released into the bath by an isolated muscle that had been pretreated with an anti-AChE far exceeded the amount that could be accounted for by the quantal leakage due to spontaneous miniature endplate potentials (MEPPs). In mammals, NQR represents most (50–98%) of the total ACh release at rest.

Depolarization of the nerve and muscle by high K^+ increased both total and quantal release, but in different degrees. Further, nerve stimulation increased the total release 1.4-fold but the quantal release 45-fold. This implied that the major part of the resting release results from leakage from the axon, the nerve terminal, or the muscle. Experiments with denervated muscles supported the nerve terminal as the main, but not the only, source of the ACh. These early data were thus largely consistent with the proposal that there are two principal mechanisms of release at resting motor nerve terminal: quantal and nonquantal, or vesicular and nonvesicular, and that the former is only a small fraction of the total.

The H-Effect

The first direct evidence for NQR was obtained by using intracellular microelectrodes to record the muscle fiber membrane potential. This approach revealed a small hyperpolarization (0.04 mV in the frog and 2 mV in the mouse) in the region of the endplate area after applying the ACh antagonist (+)tubocurarine (TC) to muscles previously treated with an anti-AChE (Figure 2). This hyperpolarization, the so-called H-effect, is believed to be due to inhibition of the depolarizing effect of ACh released by NQR. In muscles in which the NQR is absent, for example, several hours after denervation, a similar H-effect can be produced by approximately 10^{-7} M ACh. This suggests that the level of ACh in the synaptic cleft of the mouse or rat muscles is on the order of hundreds of nanomoles. It is interesting to note that in frog muscles, the magnitude of the much smaller H-effect is matched by about 10^{-9}–10^{-8} M ACh.

To learn more about the factors controlling the proposed NQR, the effects of various modifications of the ionic contents of the bath on the NQR have been investigated.

Increasing the K^+ concentration in the bath from 0.5 to 15 mM and exposure to hypertonic solutions, both procedures that increase the frequency of MEPPs at normal NMJs, has little influence on the H-effect. In contrast, lowering the Ca^{2+} from its normal value of 2 mM reduces the H-effect, which becomes undetectable in preparations incubated in Ca^{2+}-free solution for more than 60 min. The H-effect also decreases if the Ca^{2+} level is increased from 2 mM and disappears in 15 mM Ca^{2+}, although the frequency of MEPPs shows the expected dependence on the Ca^{2+} level in parallel recordings. A similar dependence on Ca^{2+} is observed at embryonic neuromuscular synapses in the frog, *Xenopus*. The H-effect has an opposite dependency on the external level of Mg^{2+}. In the absence of Mg^{2+}, the H-effect is maximal; 1 mM Mg^{2+} reduces the H-effect by one-third, and it is virtually absent at 3 mM Mg^{2+}.

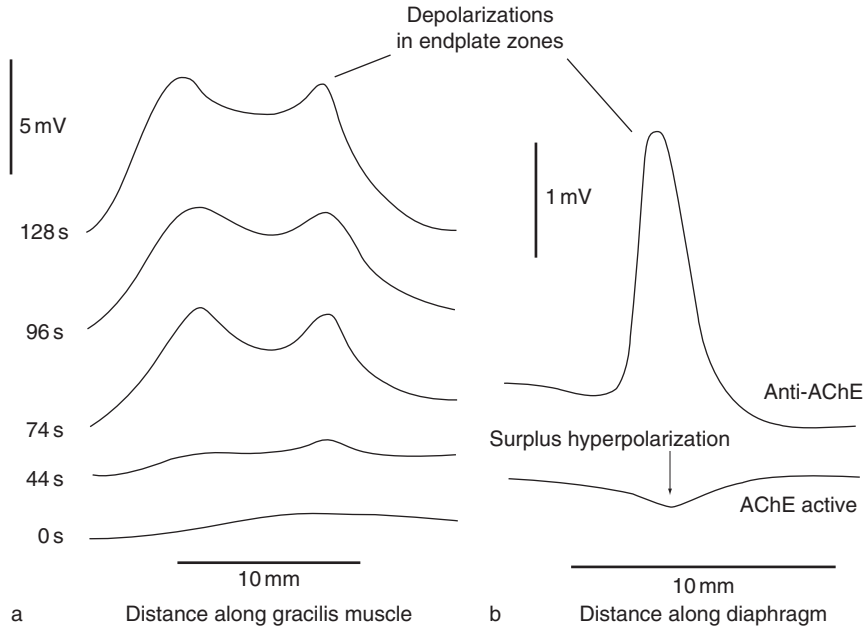

Figure 1 Distribution of potential along the cat gracilis muscle (a) and mouse hemidiaphragm (b). In (a), the control tracing was taken at zero time, after which the AChE tetraethylpyrophosphate (560 mg kg^{-1}) was injected into the jugular vein. The distribution of the electrical potential along the length of the muscle was measured by traversing two endplate regions of a muscle in a paraffin bath with a scanning electrode In (b), a mouse hemidiaphragm was hung vertically and immersed slowly in a vessel in which the liquid-air interface served as a recording electrode. The lower trace shows a small positivity in the endplate area, reflecting the slight hyperpolarisation regularly observed in muscles with intact synaptic cholinesterase (surplus hyperpolarization). After treatment of the same muscle with the irreversible AChE paraoxon, there was a local depolarization (upward inflection) apparently due to nonquantal leakage of the ACh (upper trace). The reference electrode (Ag/AgCl wire) was placed on the top of muscle. (a) Adapted from Douglas WW and Paton WMD (1954). The mechanism of motor end-plat depolarization due to a cholinesterase-inhibiting drug. *Journal of Physiology* 124(2): 325–344.

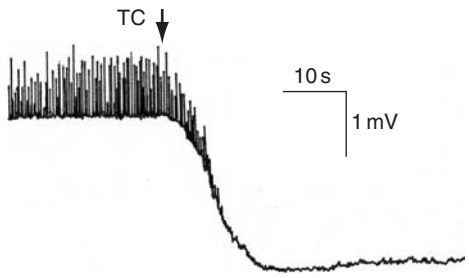

Figure 2 The effect of curare on the membrane potential in the mouse diaphragm endplate with AChE inhibited by paraoxon. At the time of the arrow 0.1 mM TC in saline was applied by pressure from a pipette with a tip diameter 10 μm. It caused the MEPPs (rapid upward deflections, left part of the record) to disappear and endplate to hyperpolarize (slow downward deflection, the H-effect). Intracellular recording by standard glass microelectrode filled with 1M KCl.

A number of compounds abolish the H-effect. These include Botulinum toxin type A; aconitin; and toosendanin, a triterpenoid derivative extracted from the bark of *Melia toosendan*. All abolished the H-effect, suggesting that the quantal and nonquantal components of the ACh release are regulated differently. This is also supported by differing temperature dependencies of quantal release and NQR. The maximal H-effect is observed at 20 °C while the frequency of MEPPs continues to increase with temperature above 20 °C.

The VAChT Hypothesis

It was proposed originally by Charles Edwards that the vesicular ACh transporters (VAChTs) may play a role in NQR of ACh. Synaptic vesicles have been shown to take up ACh by an active process. The VAChT responsible for this uptake is incorporated into the plasma membrane during exocytosis and removed by endocytosis. If these transporter molecules were to retain their orientation following the incorporation of the vesicle membrane into the axon membrane, they would then continue to move ACh out of the axoplasm, but now into the extracellular space instead of the vesicle lumen.

To test the VAChT hypothesis, the effects on the H-response of agents and treatments which act on the transport of ACh into vesicles have been examined. The most potent inhibitors of ACh transport into synaptic vesicles are vesamicol, quinacrine, and

tetraphenylborate. These compounds inhibit the H-effect at concentrations similar to those that block ACh transport in vesicles. The local application of vesamicol from a micropipette hyperpolarizes the muscle fiber membrane within seconds without changes in the amplitude of the MEPPs. This indicates that the reduction of the H-effect is of presynaptic origin.

The transport of ACh into the vesicles is coupled with a proton concentration gradient, in which the vesicle interior is more acidic. If the NQR requires a similar proton gradient, an increase in the bath pH should inhibit the release. Consistent with this, the H-effect is largely abolished at pH higher than 8. Furthermore, the H-effect is potentiated by ouabain, a blocker of the Na^+-K^+ ATPase. This block could remove a source of protons within the nerve terminal that would tend to make the cytoplasm more alkaline. Such a change in the proton gradient across the membrane might facilitate the ACh–proton exchange via the VAChT transport system. Another possibility is based on the finding that ouabain interferes with vesicle endocytosis and can thus prolong the time for which the VAChTs are incorporated into the nerve membrane. Other interpretations are also possible, however.

The H-effect, measured just after the addition and washout of an anti-AChE, is usually in the range of 7–9 mV. This value declines with time, presumably as the VAChTs are taken up by endocytosis. Consistent with this, stimulation of the motor nerve at 50 Hz, which causes extensive exocytosis, can restore the NQR.

There is thus considerable evidence, albeit indirect, that is consistent with the hypothesis that the VAChT is a source of NQR.

Modulation of NQR

A number of factors can influence the size of NQR. These include the activity of AChE in the synaptic cleft as well as substances released from the nerve, such as ACh itself and ATP and its breakdown product adenosine.

AChE Activity and NQR

An immediate question raised by the early studies of the H-effect was, How could ACh released from the nerve by NQR at normal NMJs survive breakdown by AChE in the synaptic cleft? One possible answer is that the AChE is not sufficiently activated by the low concentration of ACh resulting from NQR. This idea is supported by the demonstration that the velocity of ACh hydrolysis decreases sharply when the concentration of ACh is less than 10^{-4} M. As noted earlier, the H-effect can be produced by ACh concentrations as low as 5×10^{-8} M.

Purinergic Regulation of NQR

The role of ATP as a neurotransmitter or cotransmitter at the NMJ is now generally accepted. ATP is present in the synaptic vesicles at the neuromuscular junction in concentrations only several times smaller than that of ACh. Stimulation of the motor nerve releases ATP, together with ACh in a Ca^{2+}-dependent manner. ATP at the relatively low concentration of 10^{-5} M, but not adenosine, depresses NQR whereas both purines reduce the frequency of MEPPs. ATP reduces the amplitude of the H-effect substantially, but not completely, unless Rp-cAMP (an inhibitor of protein kinase) or a guanylyl cyclase inhibitor such as ^{1}H-[1,2,4]oxadiazolol[4,3-a]quinoxalin-1-one (ODQ) is also present. This suggests that more than one pharmacologically distinct mechanism contributes to NQR. The intracellular metabolic route or routes involved in the inhibition of NQR mediated by ATP are unknown. Metabotropic P2Y receptors might be coupled to multiple intracellular cascades. Pretreatment by staurosporine completely prevents the depressant action of ATP on the NQR, suggesting involvement of protein kinase C in the cascade.

The inhibitory influence of a continuous supply of ATP on a substantial fraction of NQR is consistent with the finding that hypoxia increases both MEPP frequency and NQR in solutions with normal extracellular Ca^{2+} (2.0 mM). The effect on MEPP frequency is virtually absent in low Ca^{2+} solution ($(Ca^{2+})_o = 0.4$ mM) whereas there is still a significant increment of NQR. This suggests that these two processes are influenced by different oxygen-sensitive mechanisms. The rise of MEPP frequency during the onset of hypoxia apparently requires the Ca^{2+} entry into the nerve terminal, whereas the NQR can be increased by other factors, such as the lower level of the ATP when its oxidative production is hindered.

The differing effects of ATP and adenosine allow a further test of the VAChT hypothesis. The reduction of NQR is prevented by the P2-receptor inhibitor suramin and by the protein kinase C inhibitor staurosporine. This further supports the ideas that there are two pharmacologically and functionally distinct types of ATP-binding receptors at murine motor nerve terminals and that there are different pathways for regulation of the two types of release.

An intriguing question is whether NQR is itself accompanied by release of ATP, as occurs during quantal release. The finding that a P2 antagonist increases the level of nonquantal ACh release favors such a tonic action of endogenous ATP. Apparently there is a

feedback inhibition of NQR mediated by accumulation of endogenous purines.

Biological Aspects of NQR

Studies of the changes in NQR and quantal release during degeneration and regeneration of the nerve and during the relative inactivity associated with mammalian hibernation have further supported the notion that although these two forms of ACh release are functionally related through the insertion of VAChT into the nerve membrane, they are regulated in somewhat different ways.

Denervation

After motor nerve section, the biochemically measured reduction in the total amount of ACh released from a mammalian muscle precedes the loss of MEPPs, suggesting that NQR declines before quantal release. Biochemical estimates, however, cannot distinguish between ACh of axonal, nerve terminal, or muscle origin. Direct measurements of the decline and recovery of MEPPs and the H-effect in the mouse diaphragm after nerve crush and during regeneration showed that the H-effect drops to 50% within 1 h of denervation and is undetectable 4 h later. In contrast, MEPPs are still present for another 5–10 h. The preferential disappearance of the NQR might be due to rapid inhibition of ACh synthesis in the terminal and exhaustion of cytosolic ACh. Even more rapid disappearance of the H-effect (within minutes) occurs after inhibition of the fast choline uptake by hemicholinium-3 or when Li^+ is substituted for Na^+.

Reinnervation

During reinnervation, the H-effect is detectable 3 days before the reappearance of MEPPs. The H-effect develops first on day 8 in the proximal endplates and then with delays of 3 and 6 days as the nerve reaches the central and distal areas of the diaphragm. Early MEPPs are less than 0.2 mV in amplitude and often slow and K^+ independent. Full recovery of the MEPP frequency and K^+ and Ca^{2+} dependency is achieved 20 days after the nerve crush in proximal parts, and 30 days in distal parts, of the diaphragm. These observations show that NQR can exist in the absence of quantal release. This might mean either that empty vesicles have fused, adding VAChT to the membrane without ACh release (invisible MEPPs) or that quantal release and VAChT incorporation are not absolutely necessary for NQR. Alternatively, early in reinnervation, VAChTs might be synthesized and incorporated into the terminal membrane but not incorporated into vesicles. They might then form a pathway for NQR which precedes vesicle maturation and quantal release. Whatever the explanation, NQR closely follows the morphological restoration of neuromuscular contact and can thus be considered as an early manifestation of the ability of the nerve terminal to synthesize and release ACh during axonal regeneration.

Hibernation

If NQR depends on the quantal release, then the H-effect should be related to the frequency of MEPPs. In muscles from golden hamsters, the frequency of quantal MEPPs is lower than in mouse or rat, particularly in hibernating animals. The H-effect is substantially lower in awake hamsters than in mice and is very small in hibernating hamsters (Figure 3) and is proportional to the rate of spontaneous quantal release. This supports the idea that NQR depends on the number of vesicles fused with the presynaptic membrane during quantal release.

Physiological Implications

What, if any, might be the effects of NQR on neuromuscular transmission in normal muscles with intact AChE activity? One possibility is that it contributes to the maintenance of the negative resting membrane potential (RMP) in the vicinity of the NMJ in the face of continuous inward leakage of Na^+. This, in turn, may influence the excitability of the postsynaptic membrane and hence the reliability with which action potentials can be generated in the muscle fiber.

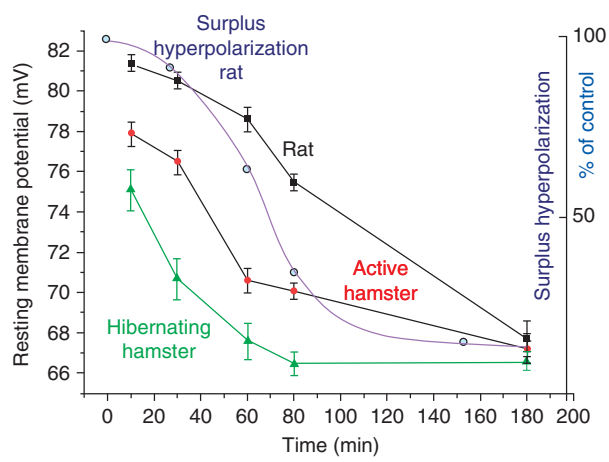

Figure 3 Time course of early postdenervation depolarization in diaphragms of rat, active and hibernating hamster, and decline of the surplus hyperpolarization in rat.

Surplus Polarization of Muscle Fibers at the NMJ

There is normally a small but consistent difference in the RMP of mammalian muscle fibers between the region of the NMJ and the extrajunctional area. At mouse or rat NMJs with intact cholinesterase, the RMP is more negative (by 2–3 mV at 22 °C, 4–5 mV at 37 °C) than in the extrajunctional area. The hyperpolarization is insensitive to the nicotinic antagonist *d*-tubocurarine chloride (dTC) making a direct role of postsynaptic nicotinic receptors unlikely.

Role of Na^+-K^+ATPase The addition of 10^{-6} M ouabain, an inhibitor of the electrogenic Na^+-K^+ pump, abolishes this local hyperpolarization in less than 5 min. This suggests that it is caused by the activity of the electrogenic Na^+-K^+ pump in the junctional region (Figure 1(b)). This activity can be potentiated by ACh when the muscle fibers are loaded with Na^+ and the pump activity is high. Thus it may be that the Na^+-K^+ pump at the NMJ is modulated by ACh. In addition to direct potentiation by ACh, the electrogenic activity of the pump might also be promoted by a high resting Na^+ influx through spontaneous openings of the sodium channels concentrated in the postsynaptic folds. In the presence of ACh, this Na^+ might be removed from the cell more rapidly, thus hyperpolarizing the subsynaptic membrane.

In muscles excised 4 h after denervation, the H-effect has disappeared and the surplus hyperpolarization is replaced by a local depolarization. It is interesting to note that the addition of 5×10^{-8} M ACh (an amount simulating the NQR) causes a reappearance of the hyperpolarization at the endplate zone.

Early Postdenervation Depolarization (EPD)

Within a few hours of nerve section, mammalian muscle fibers undergo a partial depolarization which affects their excitability and contraction. Muscle fibers of the rat diaphragm maintained *in vitro* are depolarized by about 8–10 mV (10–12% of the control RMP) within 3–4 h of denervation. This EPD coincides with the onset of an inward-directed, furosemide-sensitive Cl^- transport which may trigger other changes in the muscle fiber, such as the loss of its ability to control its volume in hypertonic solutions. The appearance of the EPD also coincides with, and may result from, the loss of NQR. If the ACh receptors (AChRs) are kept active by exposing the denervated muscles to ACh or carbacholine (Carb), the EPD is substantially reduced. This reduction is largely prevented by adding Ca^{2+} inhibitors such as verapamil or Cd^{2+}. Thus activation of AChRs, whether by NQR or exogenous cholinergic agonists, exerts a protective effect on the RMP of denervated muscles that appears to be influenced by Ca^{2+} influx (Figure 4).

In many cell types, the entry of Ca^{2+} into the cytoplasm activates nitric oxide synthase (NOS), leading to the generation and release of NO. NO can activate

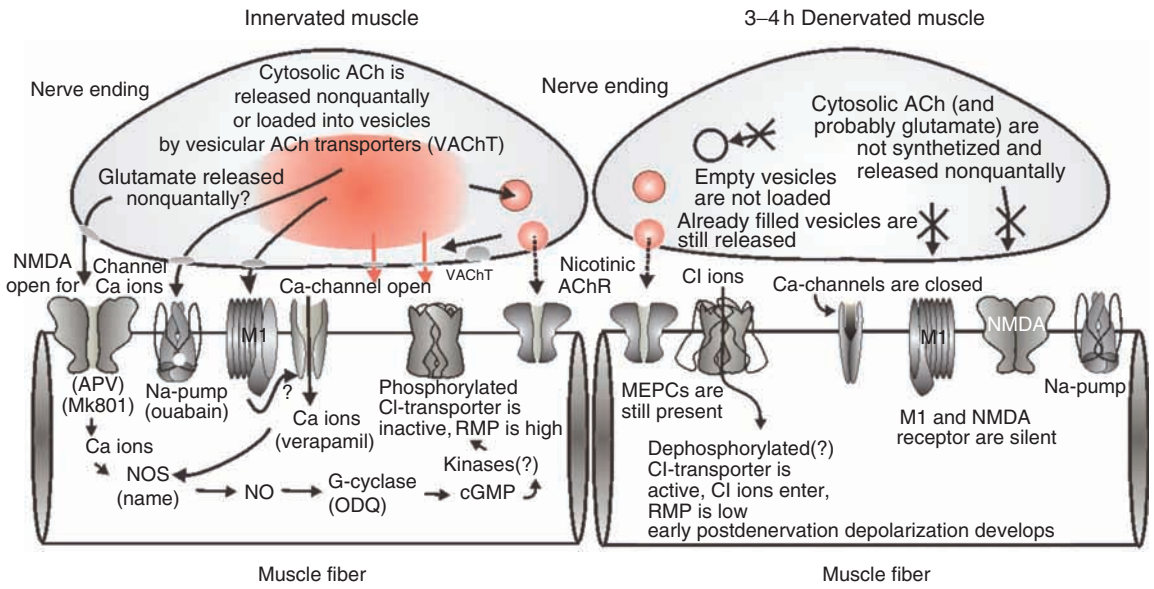

Figure 4 Possible pathways involved in regulation of the resting membrane potential and early postdenervation depolarization in skeletal muscle fibers. Nonquantal and quantal ACh indicated in reddish. M1, muscarinic acetylcholine receptors; ACh, acetylcholine; RMP, resting membrane potential; NOS, nitric oxide synthase; NMDA, *N*-methyl-D-aspartate; ODQ, ^{1}H-[1,2,4]oxadiazolol[4,3-a]Quinoxlin-1-one; cGMP, cyclic guanoline monophosphate; APV, α-(−)-2-amino-5-phosphoxo valeterate, an NMDA receptor antagonist; MEPCs, miniature endplate currents. For more details see text.

soluble guanylyl cyclase to produce cyclic guanosine monophosphate (cGMP), which activates specific protein kinases and results in the phosphorylation of various proteins. Inhibitors of NOS prevent cholinergic agonists from inhibiting EPD, suggesting that ACh released by NQR may enhance Ca^{2+}-dependent NO synthesis and the following phosphorylation cascade.

Muscarinic AChRs (mAChRs) and EPD ACh and Carb appear to delay the EPD by activating M1-cholinergic receptors. Thus while the nicotinic antagonists dTC and α-bungarotoxin have no effect on the action of Carb, oxotremorine (Oxo), a highly specific muscarinic agonist, and atropine both diminish the EPD in a similar way to Carb. This inhibition is present in 5×10^{-9} M atropine and is prevented by M1 antagonists such as pirenzepine. It remains to be determined whether these mAChRs are coupled to Ca^{2+} channels directly or via a NO-cGMP cascade (Figure 4). Consistent with the view that NQR normally suppresses EPD by activating mAChRs, muscarinic antagonists facilitate the development of EPD by protecting fibers from the action of Ach, which is still released for several hours after nerve section. In the presence of both atropine and pirenzepine, the EPD is already detectable within 30 and 60 min.

Hypertonicity and EPD Hypertonicity activates Cl^- transport in muscle fibers. Both sucrose and NaCl (in the presence of tetrodotoxin), when used to increase the osmolarity to 500 mosmol/l, significantly potentiate EPD, whereas Carb completely loses its ability to delay EPD onset. The inhibition of Cl^- transport by furosemide or Cl^-free incubation medium in hypertonic solutions effectively prevents osmotically augmented EPD.

Summary These experiments give some insight into the possible physiological significance of NQR (Figure 4). They support the view that ACh suppresses EPD by activating the furosemide-sensitive inward transport of Cl^- which starts to operate when NQR diminishes after nerve section. The NQR can apparently stimulate the Ca^{2+} influx, NO production and NO cascade-mediated phosphorylation that normally suppress the inward Cl^- current and helps to maintain the RMP at a high level. After nerve section, this protection from depolarization disappears, and the dephosphorylated transporter begins to move Cl^- into the muscle fiber, making the equilibrium potential Cl^- more positive and the muscle membrane more depolarized.

Glutamate

There is some evidence that glutamate, the most common excitatory transmitter in the CNS, may also act on the RMP of muscle fibers at the NMJ. Glutamate is the most common amino acid in most cells, and glutamate-like immunoreactivity has been reported at rat motor nerve terminals. Glutamate is secreted and taken up at the frog neuromuscular junction, and mRNA coding for a glutamate transporter is present in rabbit motor neuron perikarya. On the postsynaptic side, an accumulation of *N*-methyl-D-aspartate receptor-1 (NMDAR-1) glutamate receptor subunits has been demonstrated by immunolabeling in the rat diaphragm.

It has been suggested that glutamate participates in the regulation of muscle fiber RMP through a pathway similar to that activated by ACh. Like ACh and Carb, glutamate and *N*-methyl-D-aspartate (NMDA) induce Ca^{2+}-dependent synthesis of NO in the sarcoplasm, which delays the EPD. The influence of glutamate on RMP appears to be mediated by the NMDA subtype of Glu receptors which are enriched at NMJ. There is thus potential for interaction between the effects of ACh released by NQR and glutamate in the regulation of the RMP, and hence excitability, at the NMJ.

Conclusions

There is considerable evidence for a continuous release of ACh from motor nerve terminals associated with the activity of VAChT inserted into the nerve membrane during exocytosis. Early experiments implied the existence of NQR from the hyperpolarization of the muscle (the H-effect) at the NMJ in response to block of mAChRs in muscles in which the AChE activity had also been blocked. Normal AChE activity would prevent this effect but appears to allow enough ACh to reach the muscle to activate mAChRs. This in turn appears to trigger a postsynaptic phosphorylation cascade that involves both Ca^{2+} and NO as second messengers. This cascade suppresses the activity of an inward Cl^- transporter and activates the electrogenic sodium pump. This supports the local hyperpolarization of the muscle in the region of the NMJ, thus helping maintain muscle excitability. A number of factors including AChE activity and purines released from the nerve appear to influence the magnitude of NQR and the effects resulting from it.

Although the full physiological significance of NQR is not yet clear, it seems that it should not be considered simply a by-product of the quantal

exocytosis of ACh but rather a fundamental aspect of the release process that may play a significant role in regulating synaptic efficacy at the NMJ.

See also: Adenosine Triphosphate (ATP) as a Neurotransmitter; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction (NMJ): Acetylcholinesterases; Neuromuscular Connections: Vertebrate Patterns of; Presynaptic Events in Neuromuscular Transmission.

Further Reading

Douglas WW and Paton WDM (1954) The mechanism of motor end-plat depolarization due to a cholinesterase-inhibiting drug. *Journal of Physiology* 124(2): 325–344.

Edwards C, Doležal V, Tuček S, et al. (1985) Is an acetylcholine transport system responsible for non-quantal release of acetylcholine at the rodent myoneural junction? *Proceedings of the National Academy of Sciences of the United States of America* 82: 3514–3518.

Ferguson SM, Savchenko V, Apparsundaram S, et al. (2003) Vesicular localization and activity-dependent trafficking of presynaptic choline transporters. *Journal of Neuroscience* 23: 9697–9709.

Galkin AV, Giniatullin RA, Mukhtarov MR, et al. (2001) ATP but not adenosine inhibits nonquantal acetylcholine release at the mouse neuromuscular junction. *European Journal of Neurosciences* 13: 2047–2053.

Katz B and Miledi R (1977) Transmitter leakage from motor nerve endings. *Proceedings of the Royal Society of London B: Biological Sciences* 196: 59–72.

Malomouzh A, Nikolsky EE, Lieberman EM, et al. (2005) Effect of N-acetylaspartylglutamate (NAAG) on non-quantal and spontaneous quantal release of acetylcholine at the neuromuscular synapse of rat. *Journal of Neurochemistry* 94: 257–265.

Moravec J and Vyskočil F (2005) Early postdenervation depolarization develops faster at endplates of hibernating golden hamsters where spontaneous quantal and non-quantal acetylcholine release is very small. *Neuroscience Research* 51: 25–29.

Nikolsky EE, Oranska TI, and Vyskočil F (1996) Non-quantal acetylcholine release in the mouse diaphragm after phrenic nerve crush and during recovery. *Experimetal Physiology* 81: 341–348.

Nikolsky EE, Zemkova´ H, Voronin VA, et al. (1994) Role of non-quantal acetylcholine release in surplus polarization of mouse diaphragm fibres at the endplate zone. *Journal of Physiology* 477: 497–502.

Sun Y and Poo M (1985) Non-quantal release of acetylcholine at a developing neuromuscular synapse in culture. *Journal of Neuroscience* 5: 634–642.

Vyskočil F (2003) Early postdenervation depolarization is controlled by acetylcholine and glutamate via nitric oxide regulation of the chloride transporter. *Neurochemical Research* 28: 575–585.

Vyskočil F and Illes P (1977) Non-quantal release of transmitter at mouse neuromuscular junction and its dependence on the activity of Na^+-K^+ ATP-ase. *Pfluegers Archiv* 370: 295–297.

Vyskočil F, Nikolsky E, and Edwards C (1983) An analysis of the mechanisms underlying the non-quantal release of acetylcholine at the mouse neuromuscular junction. *Neuroscience* 9: 429–435.

Vyskočil F and Vrbova´ G (1993) Non-quantal release of acetylcholine affects polyneuronal innervation on developing rat muscle fibres. *European Journal of Neurosciences* 5: 1677–1683.

Zemkova´ H and Vyskočil F (1989) Effect of Mg^{2+} on nonquantal acetylcholine release at the mouse neuromuscular junction. *Neuroscience Letters* 103: 293–297.

Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission

E-L Bélair and R Robitaille, Université de Montréal, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The concept of the neuromuscular junction (NMJ) as a tripartite structure, consisting of a nerve terminal (presynaptic component), a muscle fiber and its endplate (postsynaptic component), and a few perisynaptic Schwann cells (PSCs; glial component), has gained general acceptance over the past decade. Indeed, Schwann cells that were first described as passive cells, implicated in homeostasis functions and NMJ stability, are now believed to actively modulate synaptic transmission. The discoveries that led to the evolution of this concept, mainly evidence that PSCs can detect and respond to synaptic activity and, in return, have the ability to increase or decrease transmitter release, are discussed in this article.

An Advantageous Position

PSCs lie on top of nerve terminals and extend processes that run along and wrap around them in the vicinity of transmitter release sites (active zones). Although this is true for all vertebrates, differences exist among species. At amphibian NMJs, for example, these processes intrude into the synaptic cleft. In contrast, at mammalian NMJs, PSC processes remain at the edge of the synaptic cleft. Interposition of Schwann cell processes between presynaptic and postsynaptic membranes appears to be negatively correlated with synaptic efficacy, suggesting that glial coverage of synaptic clefts may regulate information processing. Moreover, because of their proximity to active zones, PSCs are well positioned to sense synaptic activity and exert a modulating role on pre- and/or postsynaptic elements.

Detectors of Synaptic Transmission

A Cocktail of Presynaptically Released Molecules

During presynaptic activity, many neurotransmitters and neuromodulators can be released from nerve terminals. In fact, at some NMJs, adenosine triphosphate (ATP) is released synchronously together with the neurotransmitter acetylcholine (ACh) following a single nerve impulse and is subsequently degraded by enzymatic activity into ADP, AMP, and adenosine. In addition, immunoreactivity for peptides (calcitonin gene-related peptide and substance P) was found in large dense core vesicles of amphibian and mammalian motor nerve terminals and the release of these peptides is elicited by high-frequency nerve stimulation. Synaptic activity also regulates the release of diverse neurotrophins – namely, brain-derived neurotrophic factor (BDNF), neurotrophin-3 (NT-3), and neurotrophin-4 (NT-4) – that, in turn, act as synaptic regulators. Therefore, according to the level of synaptic activity, a cocktail of molecules can be released into the synaptic environment and the message received by the three synaptic partners varies as a function of the receptors they express.

A Variety of Receptors

Using calcium imaging and immunocytochemical techniques, functional receptors for all of the aforementioned molecules have been found on PSCs. The presence of functional receptors on PSCs *in situ* was first revealed at amphibian NMJs, where PSCs responded to nerve stimulation with an elevation of intracellular Ca^{2+}. This Ca^{2+} rise was greatly reduced by blockade of presynaptic Ca^{2+} entry, was independent of the presence of muscle fibers, and could be reproduced by direct application of ATP, adenosine, and ACh. This suggests that neurotransmitters released from presynaptic nerve terminals mediated the glial Ca^{2+} responses. Also, because Ca^{2+} responses still occurred in the absence of extracellular Ca^{2+}, intracellular Ca^{2+} stores were involved. The use of specific antagonists later allowed the identification of the receptors mediating ATP and adenosine responses on frog PSCs. ATP triggered Ca^{2+} rises via P2Y and P2X receptors; P2YR activation led to the release of Ca^{2+} from internal stores through recruitment of a pertussis toxin-sensitive G-protein and P2XR activated L-type Ca^{2+} channels. Adenosine acted on A1 receptors also coupled to a pertussis toxin-sensitive G-protein. The blockade of A1 receptors did not alter calcium rises evoked by nerve stimulation, suggesting that they are not involved in neuron–glia communication during normal neurotransmission. Conversely, blockade of ATP receptors led to delayed Ca^{2+} elevations with smaller amplitude, indicating that endogenous ATP participates in synapse–glia signaling. The identity of the receptors mediating cholinergic responses in PSCs is, however, less clear. An unconventional muscarinic receptor, insensitive to the general antagonist atropine but inhibited by the nonspecific blocker gallamine, has

been identified. These receptors are coupled to internal Ca^{2+} stores via a pertussis toxin-insensitive G-protein mechanism. Hence, owing to the presence of receptors for the neurotransmitters normally released during presynaptic activity, PSCs are well adapted to sense this activity and ultimately to respond to it.

An Evolutionarily Conserved Property

The ability of PSCs to respond to synaptic activity is well conserved among vertebrates, although some evolutionary differences exist. For example, at the skate neuroelectrocyte junction, which has a neuromuscular embryonic origin, receptors for ATP and ADP, but not for adenosine and AMP, were identified on PSCs. This preparation also receives a cholinergic innervation where ATP is co-released with ACh. Application of ATP was found to trigger a biphasic calcium elevation in PSCs, involving P2Y receptor stimulation and, most likely, activation of phospholipase C, generation of inositol 1,4,5-trisphosphate (IP_3), and subsequent Ca^{2+} release from IP_3-sensitive stores. In mammals, ATP-mediated Ca^{2+} responses do not involve P2 receptors, suggesting that ATP acts through its metabolites, mainly adenosine. Moreover, as opposed to frog PSCs, A1 receptors for adenosine participate in nerve-evoked Ca^{2+} rises, but to a lesser extent than receptors for ACh. In addition, the latter are of a conventional muscarinic type, since atropine blocks ACh-evoked Ca^{2+} responses. Regardless of the interspecies variability in receptor expression, which might contribute to physiological differences between species, the ability of PSCs to respond to the release of transmitters is common to all, reflecting their functional importance. Also, the variety of receptors for a given neurotransmitter and their relative sensitivity will undoubtedly color the signal received by PSCs and adapt their response as a function of synaptic activity.

Adapted Ca^{2+} Responses

In addition to their receptors for neurotransmitters, PSCs possess receptors for molecules that are preferentially released during elevated synaptic activity. For example, substance P induces the release of Ca^{2+} from intracellular stores of frog PSCs, most likely through its metabolites. These Ca^{2+} elevations are mediated by neurokinin-1 receptors (NK-1Rs) coupled to pertussis toxin (PTX)-sensitive G-proteins. NK-1R activity seems to be involved in the modulation of ATP and ACh signaling in PSCs, since their previous activation reduces the amplitude of Ca^{2+} responses evoked by nerve stimulation or local application of muscarine and ATP. Some neurotrophin receptors may play similar roles in specific conditions. In fact, NT-3 and BDNF trigger Ca^{2+} elevations in mouse PSCs, through tyrosine kinase B (BDNF) and tyrosine kinase C (NT-3) receptors. An application of BDNF prior to nerve stimulation increases the duration of the evoked Ca^{2+} response. These molecules may be important for the fine-tuning of PSC Ca^{2+} responses and may enlarge their scale of responsiveness to ACh and ATP, allowing them to discriminate different degrees of synaptic activity.

Presence of Voltage-Gated Ion Channels

In addition to the various receptors for neurotransmitters and neuromodulators, PSCs possess voltage-gated ion channels that also allow them to sense synaptic activity. Depolarization-activated L-type Ca^{2+} channels are found on frog PSC processes, close to the synaptic cleft, while N-type Ca^{2+} channels are presumed to be present on mammalian PSCs. However, the persistence of PSC calcium responses in the absence of extracellular Ca^{2+} suggests that only a small proportion of the response can be accounted for by the activation of Ca^{2+} channels. Hence, PSCs are kept aware of synaptic activity mainly through G-protein-coupled receptor activation and intracellular Ca^{2+} signaling, with Ca^{2+} being released from intracellular stores and, to a lesser extent, entering the cell via voltage-gated ion channels. Therefore, a direct manipulation of G-protein or Ca^{2+} signaling would lead to a modulation of the message received by PSCs, which could, in turn, modulate synaptic transmission.

Modulators of Synaptic Transmission

G-Protein Signaling

The first direct evidence of PSCs modulating synaptic transmission came from a study of the frog NMJ, where PSC G-proteins were specifically activated or inactivated via intracellular injection of guanosine 5′-3-*O*-(thio)triphosphate (GTPγS) and guanosine 5′-*O*-(2-thiodiphosphate) (GDPβS), respectively. This method avoids the complication of bath-applied drugs acting on cells other than PSCs. GTPγS injection into PSCs decreased the amplitude of nerve-evoked endplate currents, but had no effect on the frequency, amplitude, or kinetics of spontaneous events (Figure 1). Furthermore, paired-pulse facilitation was not altered by the injection, suggesting that the modulation of synaptic efficacy by PSCs does not affect the basic mechanisms required for vesicular fusion but rather involves the synchronization mechanisms that are necessary for quantal release. Moreover, the effect of

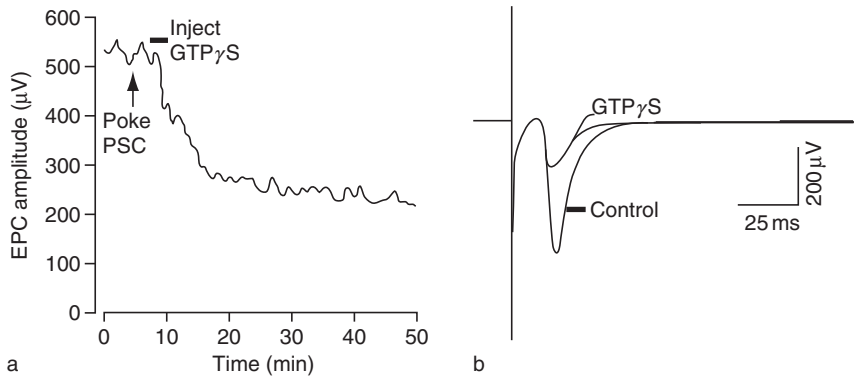

Figure 1 Injection of guanosine 5′-3-*O*-(thio)triphosphate (GTPγS) in perisynaptic Schwann cells depresses synaptic transmission. (a) Endplate current (EPC) amplitude recorded focally from a nerve terminal branch and evoked by a single stimulus at 0.2 Hz before, during, and after injection of GTPγS in the PSC covering the same terminal branch. The injection of GTPγS into the PSC produced a reduction in the amplitude of EPC. (b) Average of four EPCs recorded before (control) and 20 min after the injection of GTPγS in the PSC. Reproduced from Robitaille R (1998) Modulation of synaptic efficacy and synaptic depression by glial cells at the frog neuromuscular junction. *Neuron* 21(4): 847–855, with permission from Elsevier.

GTPγS injection into PSCs was only partially occluded by pretreatment with pertussis toxin, revealing a dual effect of PTX-sensitive and -insensitive G-protein activation on the modulation of synaptic efficacy. On the other hand, inhibition of G-proteins, via GDPβS injection, had no effect on transmitter release evoked by 0.2 Hz stimulation. While an increase of synaptic efficacy would have been expected if PSC G-proteins were tonically active, the observed lack of effect is suggestive of an absence of tonic activity. This is consistent with the absence of detectable Ca^{2+} signaling in PSCs with low-frequency nerve stimulation and reflects an activity-dependent modulation of transmitter release by PSCs.

In addition to their modulation of transmitter release, PSCs also participate in short-term plasticity of synaptic transmission. Indeed, inhibition of PSC G-proteins decreased synaptic depression in response to high-frequency stimulation. Moreover, the depression induced by PSC G-protein activation occluded 75% of high-frequency-triggered depression. These results suggest that PSC G-protein might modulate the presynaptic mechanisms causing the depression of transmitter release (Figure 2).

Calcium Signaling

The effect of PSC Ca^{2+} elevations on transmitter release was later investigated in the same system, using thapsigargin, an inhibitor of Ca^{2+}-ATPase pumps of internal stores. Bath application of thapsigargin induced a transient rise of Ca^{2+} in PSCs as a result of the gradual leak of Ca^{2+} from the internal stores and the incapacity to replenish them. Thapsigargin also prolonged Ca^{2+} responses in presynaptic nerve terminals following single action potentials and caused Ca^{2+} accumulation during high-frequency stimulation without altering resting levels of Ca^{2+}. Therefore, presynaptic internal stores were assumed to be empty at rest and to participate to Ca^{2+} clearance during synaptic activity. Correlated with the prolonged Ca^{2+} signal in presynaptic nerve terminals was a rapid increase in transmitter release. After a delay, transmitter release was further increased, transiently and with a slower rate. This delayed and transient increase of transmitter release was associated with PSC Ca^{2+} rise, since it was occluded by specific injection of the Ca^{2+} chelator 1,2-bis(2-aminophenoxy)ethane-*N*,*N*,*N*′,*N*′-tetraacetic acid (BAPTA) into PSCs. Moreover, the release of Ca^{2+} from PSC internal stores, triggered by the injection of inositol 1,4,5-trisphosphate, mimicked thapsigargin-induced potentiation of transmitter release (Figure 3). This potentiation was associated with an increase in the frequency of miniature endplate potentials (mEPPs). This, together with the lack of effects of thapsigargin on mEPP amplitude and time course, implies that this potentiation is presynaptic in origin.

Because Ca^{2+} signaling is undetectable in PSCs at low frequencies of stimulation, their Ca^{2+}-dependent potentiation of transmitter release is likely to occur mainly during elevated synaptic activity. Decreasing the availability of free Ca^{2+} by increasing the Ca^{2+} buffering capacity of PSCs should inhibit this potentiation. As expected, specific BAPTA injection into PSCs, prior to high-frequency stimulation, enhanced the depression of synaptic transmission, confirming the potentiating role of PSC Ca^{2+} responses (Figure 4). Therefore, PSCs can modulate

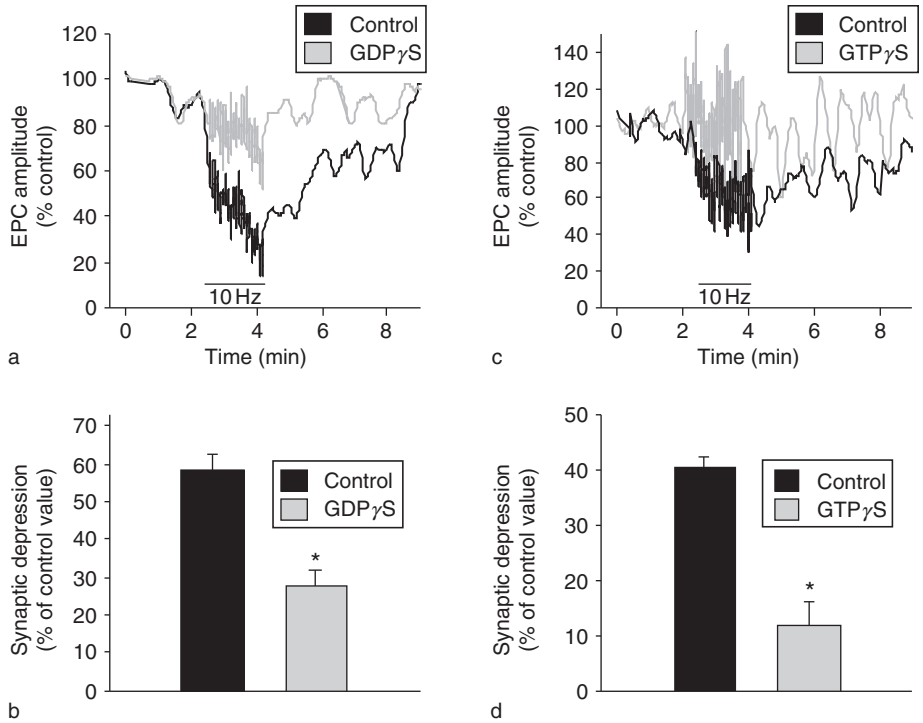

Figure 2 Perisynaptic Schwann cells modulate endogenous high-frequency depression. (a) Endplate current (EPC) amplitude expressed as a percentage of control value before, during, and after a train of stimuli (10 Hz, 80 s) recorded before (black trace) and 25 min following guanosine 5′-*O*-(2-thiodiphosphate) (GDPβS) injection in the PSC covering the nerve terminal branch recorded by the focal electrode (gray trace). EPCs were evoked at a frequency of 0.2 Hz before and after the train of stimuli. (b) Average depression in control (black) and following GDPβS injection in PSCs (gray) ($n = 5$). Note that the amount of synaptic depression was significantly reduced (asterisk, $p < 0.05$; Student's *t* test). (c) EPC amplitude expressed as a percentage of control value before, during, and after a train of stimuli (10 Hz, 80 s) recorded before (black trace) and 35 min following GTPγS injection (gray) in the PSC covering the nerve terminal branch recorded by the focal electrode. EPCs were evoked at a frequency of 0.2 Hz before and after the train of stimuli. (d) Average depression in control (black) and following GTPγS injection (gray) in PSCs ($n = 4$). Note that the amount of synaptic depression was significantly reduced (asterisk, $p = 0.02$; paired Student's *t* test), suggesting that nerve-evoked synaptic depression was occluded by the depression induced by PSC G-protein activation. Adapted from Robitaille R (1998) Modulation of synaptic efficacy and synaptic depression by glial cells at the frog neuromuscular junction. *Neuron* 21(4): 847–855, with permission from Elsevier.

short-term plasticity either by increasing or decreasing transmitter release, via Ca^{2+} signaling and G-protein activation, respectively. As a result, the balance between potentiation and depression by PSCs may vary according to the state of the synapse and the level of its activity.

Glutamate Carboxypeptidase II

N-Acetylaspartylglutamate (NAAG) is a widely distributed neuropeptide in the mammalian central nervous system (CNS). The presence of its hydrolyzing enzyme on PSC membranes and its neuromodulatory role at the rat NMJ have raised the possibility of its release by motor neuron terminals. Indeed, NAAG was found to inhibit nonquantal release of ACh, through the activation of postsynaptic *N*-methyl-D-aspartate (NMDA) receptors and synthesis of the retrograde messenger nitric oxide (NO). This effect was mimicked by glutamate application and blocked by the inhibition of the PSC membrane-associated ectoenzyme, glutamate carboxypeptidase II (GCP II), which hydrolyzes NAAG into glutamate. Therefore, PSCs could modulate transmitter release indirectly through the metabolism of presynaptically released neuropeptides. This function may be tightly regulated by synaptic activity, as activation of glial type II metabotropic glutamate receptors has been shown to upregulate GCP II activity in crayfish nervous tissue. Thus, previous synaptic activity could provide an important reciprocal modulation of synaptic activity by PSCs.

The variety of mechanisms by which PSCs can modulate neurotransmission strongly suggests that the ultimate effect on the synapse is a balance of potentiation and depression. The equilibrium may be displaced according to the state of the synapse and its previous activity, favoring one or another pathway and leading to the release of appropriate gliotransmitters (Figure 5).

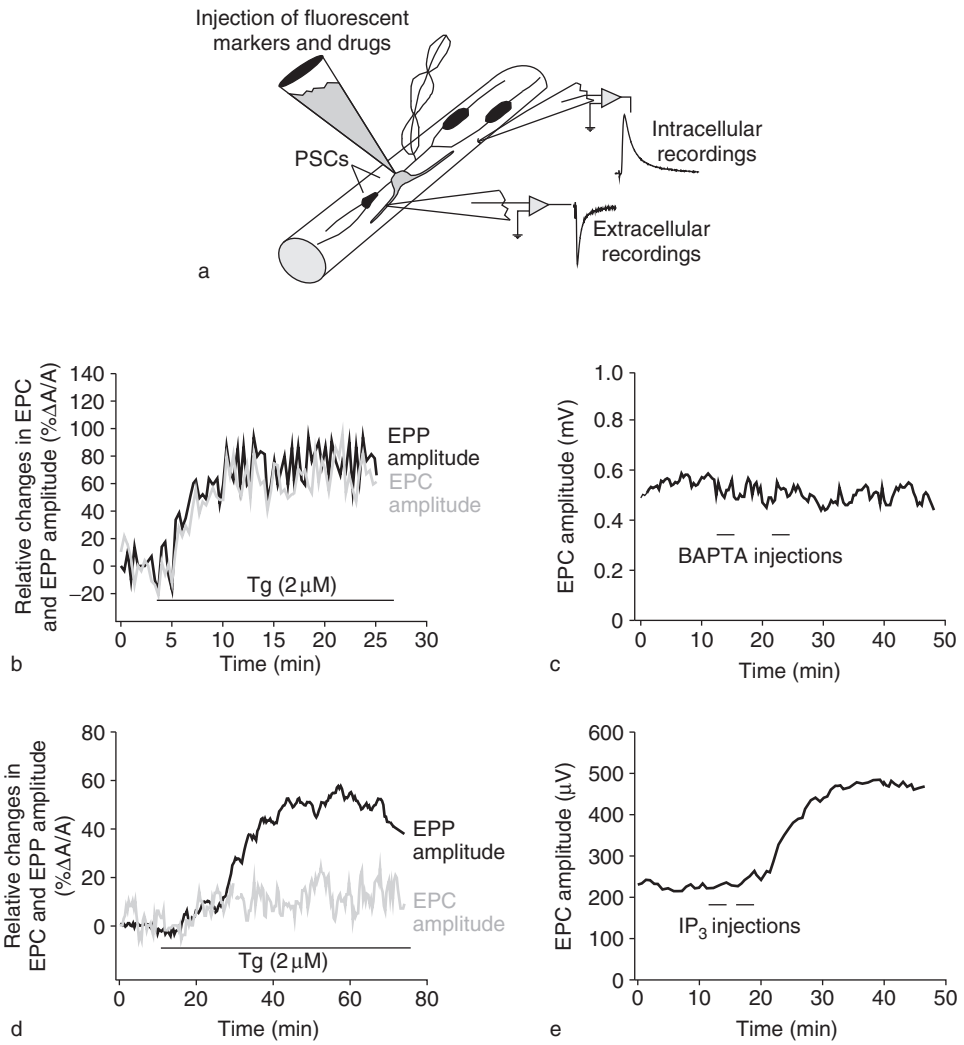

Figure 3 Potentiation of synaptic transmission by perisynaptic Schwann cells (PSCs). (a) Experimental model of the three-electrode recording method. Intracellular recordings of synaptic activity generated by the whole neuromuscular junction were performed simultaneously with extracellular recordings, using a focal electrode, which monitored the portion of the presynaptic terminal covered by a PSC injected via the third electrode. (b) Endplate current (EPC) and endplate potential (EPP) amplitude expressed as a percentage of change relative to control before and during bath application of an inhibitor of the Ca^{2+}-ATPase pumps, thapsigargin (Tg; 2 μM), after the injection of the PSC covering the portion of the nerve terminal recorded by the focal electrode with the vehicle solution (500 mM K^+-acetate and 10 μM Ca^{2+}-green-1). Changes in EPC amplitude represent the activity of the nerve terminal covered by the injected PSC, whereas changes in EPP amplitude represent changes occurring at the whole neuromuscular junction. Note that Tg induced a similar increase in EPP amplitude (whole neuromuscular junction activity) and in EPC amplitude (portion of the terminal covered by the injected PSC). (c) EPC amplitude recorded by a focal electrode located near the portion of a nerve terminal covered by a PSC before, during, and after the injection of the calcium chelator 1,2-bis(2-aminophenoxy)ethane-*N,N,N′,N′*-tetraacetic acid (BAPTA; 10 mM in the electrode) in the PSC. EPCs were evoked by nerve stimulation at 0.2 Hz. Note that BAPTA injection had no effect on synaptic transmission. (d) EPC and EPP amplitude expressed as a percentage of changes relative to control before and during bath application of Tg (2 μM) after the injection of BAPTA (10 mM in the electrode) in the PSC covering the portion of the nerve terminal recorded by the focal electrode. Note that Tg induced a rapid rise, followed by a second, slower increase in EPP amplitude (whole neuromuscular junction activity) but induced only a small and rapid rise in EPC amplitude (portion of the terminal covered by the injected PSC). (e) EPC amplitude before and after the injection of inositol 1,4,5-trisphosphate (IP_3; 10 mM in the electrode) in the PSC covering the portion of the nerve terminal recorded by the focal electrode. EPCs were evoked by motor nerve stimulation at 0.2 Hz. Note that the injection of IP_3 in PSCs potentiated transmitter release. Reprinted from Castonguay A and Robitaille R (2001) Differential regulation of transmitter release by presynaptic and glial Ca^{2+} internal stores at the neuromuscular synapse. *Journal of Neuroscience* 21(6): 1911–1922. Copyright 2001 by the Society for Neuroscience.

Putative Gliotransmitters

In order to achieve reciprocal communication, PSCs need to release some neuroactive substances that can diffuse to, and activate signaling cascades in, its synaptic partners, just as presynaptic nerve terminals do when releasing neurotransmitters and neuropeptides. Many molecules have been cited as putative gliotransmitters, suspected to be

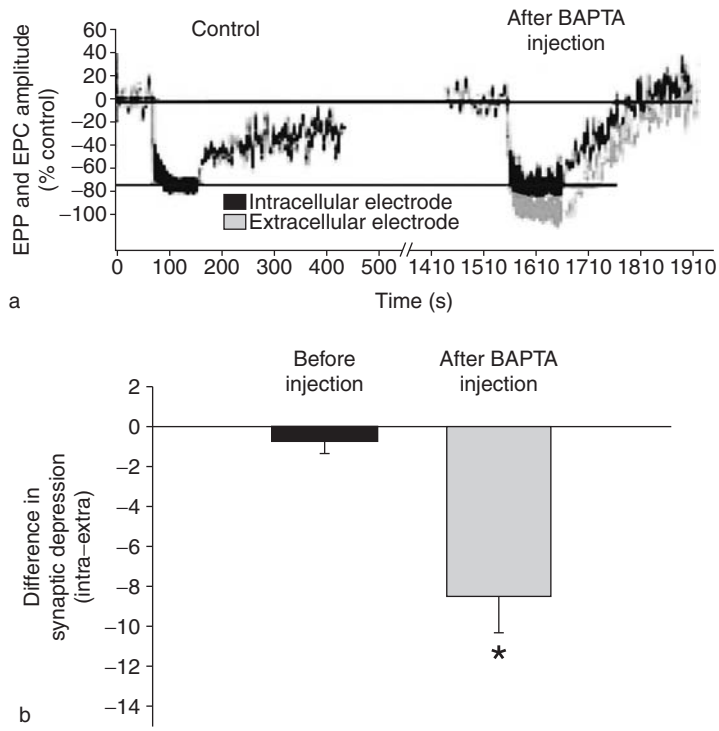

Figure 4 Synaptic depression is modulated by perisynaptic Schwann cells. (a) Changes in endplate current (EPC; gray trace) amplitude and endplate potential (EPP; black trace) amplitude expressed as a percentage of control before, during, and after the induction of high-frequency depression (10 Hz/80 s) at the frog neuromuscular junction. Note that the focal and intracellular recordings show the same amplitude of depression. After the injection of calcium chelator 1,2-bis(2-aminophenoxy)ethane-*N,N,N′,N′*-tetraacetic acid (BAPTA) into a PSC (break in *x*-axis), the focal electrode (gray trace), which records synaptic transmission under the injected PSC, shows a bigger depression than the intracellular on does, thus suggesting that the presence of BAPTA in the PSC blocked the potentiating effect on synaptic transmission. (b) Mean ± SEM of the difference between intracellular and extracellular depression in control and after BAPTA injection into PSCs. Note that, after BAPTA injection, the difference between the amplitude of depression recorded by the intracellular and focal electrodes was significantly greater. Reprinted from Castonguay A and Robitaille R (2001) Differential regulation of transmitter release by presynaptic and glial Ca^{2+} internal stores at the neuromuscular synapse. *Journal of Neuroscience* 21(6): 1911–1922. Copyright 2001 by the Society for Neuroscience.

involved in depression or potentiation of transmitter release.

Possible Mediators of Synaptic Depression

Nitric oxide NO is a ubiquitous neuromodulator of the nervous system. At the frog NMJ, exogenous application of NO causes a reduction of EPP amplitude that is presynaptic in origin. NO is also endogenously produced in a tonic fashion and is implicated in synaptic depression via a cyclic guanosine monophosphate (cGMP)-dependent mechanism. Furthermore, NO is synthesized during elevated synaptic activity and contributes to high-frequency depression independently of soluble guanylate cyclase activation. A form of neuronal NO synthase (NOS) has been found in PSCs of the frog NMJ. Many NOS enzymes are Ca^{2+} dependent, therefore it is likely that activity-dependent Ca^{2+} elevations in PSCs could trigger NO production and mediate synaptic depression. This would, however, be inconsistent with the greater synaptic depression observed with specific injection of BAPTA into PSCs, unless Ca^{2+} activates two different pathways leading either to potentiation or depression, with potentiation prevailing in normal conditions. As an alternative, NOS could be activated through a G-protein pathway, consistent with the effect of G-protein activation in PSCs.

Glutamate Like NO, glutamate and (1*S*,3*R*)-aminocyclopentanedicarboxylic acid (ACPD), a metabotropic glutamate receptor (mGluR) agonist, decrease transmitter release at the frog NMJ. Glutamate also modulates synaptic transmission in an activity-dependent manner, since blockade of mGluRs by (*S*)-α-methyl-4-carboxyphenylglycine (MCPG) during high-frequency stimulation reduces synaptic depression. This frequency dependence of glutamatergic modulation of transmitter release resembles the frequency dependence of glial regulation of synaptic depression. Therefore, glutamate was suggested to be glial in origin. This proposition is consistent with the observation of glutamate-like immunoreactivity in

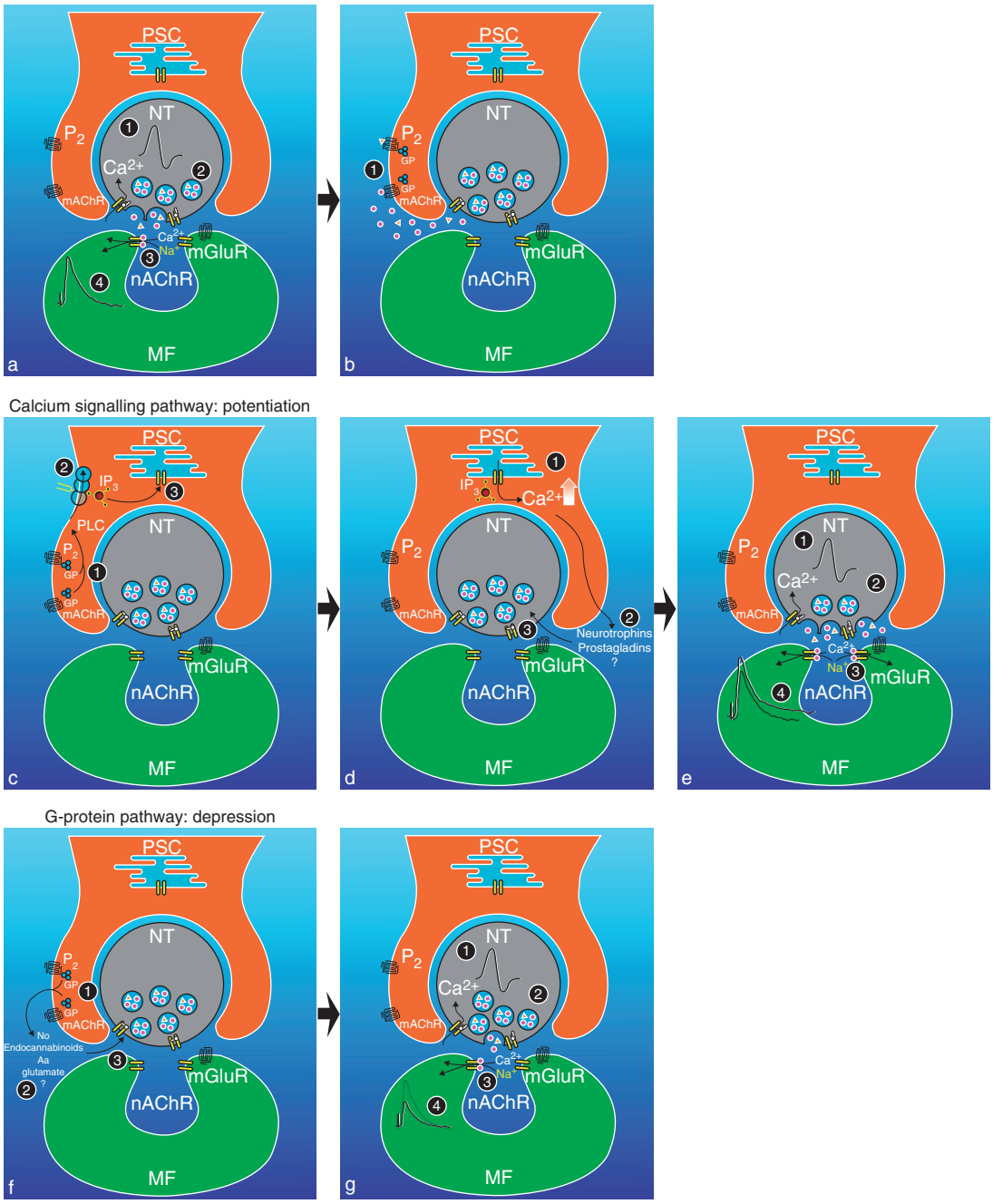

Figure 5 Perisynaptic Schwann cell (PSC) modulation of neurotransmission. Schematic representation of synapse–glia interactions at the neuromuscular junction. (a) Upon arrival of an action potential at the nerve terminals (NT; 1), voltage-gated calcium channels open and Ca^{2+} enter the nerve terminals, triggering the release (2) of neurotransmitters acetylcholine (circles) and ATP (triangles) into the synaptic cleft. (3) Acetylcholine (ACh) binds to nicotinic receptors (nAChR) on muscle fiber (MF) endplate, leading to an influx of Na^{+} and Ca^{2+} (4) that causes a depolarization (endplate potential). (b) Neurotransmitters also bind to PSC metabotropic receptors (1; ACh → mAChR, ATP → P2R), which activate G-proteins. (c) Calcium signaling pathway: potentiation. (1) G-proteins activate phospholipase C (PLC), which (2) cleaves membrane lipid phosphatidylinositol bisphosphate into diacylglycerol and inositol trisphosphate (IP_3). (3) IP_3 then binds to its receptors on endoplasmic reticulum. (d) Activation of IP_3 receptor triggers the release of Ca^{2+} (1) from internal stores. (2) Calcium elevation in PSCs induces the release of gliotransmitters (prostaglandins, neurotrophins, and/or unidentified molecules) into the synaptic cleft. These molecules, in turn, act on nerve terminals (3). (e) At the nerve terminals (1) following the arrival of an action potential, the molecules (2) increase the release of transmitters through unknown mechanisms. (3) This greater transmitter release results in a stronger activation of nAChRs and (4) a larger endplate potential. (f) G-protein pathway: depression. (1) G-protein activity in PSCs also activates a signaling cascade (2) that leads to the release of depressing gliotransmitters (nitric oxide (NO), endocannabinoids, arachidonic acid (AA), glutamate, and/or unidentified molecules) into the synaptic cleft. (3) These molecules, in turn, act on nerve terminals. (g) At the nerve terminals (1) following the arrival of an action potential, the molecules (2) decrease the release of transmitters through unknown mechanisms. (3) This reduction in transmitter release results in a weaker activation of nAChR and (4) a smaller endplate potential.

rat PSCs and the presence of functional glutamate uptake machinery on these cells at the frog NMJ, thus providing an economical mechanism to recycle glutamate. By releasing glutamate, likely in a Ca^{2+}-dependent manner, or by regulating the rate and capacity of glutamate reuptake, PSCs may thus modulate short-term plasticity through postsynaptic mGluR activation and possibly NO synthesis by muscle fibers.

Endocannabinoids Endocannabinoids have been found to decrease cholinergic transmission by reducing the release of ACh in several smooth and skeletal muscle preparations. In addition, in the CNS, glial cells express some enzymes of the endocannabinoid metabolism, both *in vitro* and *in situ*, suggesting that they may synthesize these molecules. Thus, a role of glial cells in the cannabinergic modulation of synaptic transmission at the NMJ could be envisioned.

Arachidonic acid As with endocannabinoids, arachidonic acid (AA), an important component of membrane lipids, and also a degradation product of the endocannabinoid 2-arachidonoylglycerol (2-AG), has been shown to inhibit presynaptic Ca^{2+} currents at the frog NMJ. Moreover, AA has a depressive effect on transmitter release. Cultured Schwann cells present an agonist-dependent regulation of AA release, and therefore AA could be envisaged as a gliotransmitter implicated in synaptic depression. The effects of AA on synaptic transmission at the frog NMJ have been shown to be dependent on cyclooxygenase activity and are believed to be mediated by prostaglandin E_2. Depressive effects on synaptic transmission have also been described for prostaglandins E_1 and $F_{2\alpha}$ at the frog NMJ.

Possible Mediators of Synaptic Potentiation

Prostaglandins In addition to their depressive effect, prostaglandins can also potentiate synaptic transmission. Indeed, at the frog NMJ, inhibition of a prostaglandin-synthesizing enzyme, cyclooxygenase, decreased quantal release. This decrease was partially restored by subsequent exposure to prostaglandin E_2. However, when applied on untreated preparations, prostaglandin E_2 decreased the amplitude of evoked EPPs, suggesting that its potentiating effect depends on the state of the synapse and may be unmasked in synapses that are already depressed. Therefore, one could hypothesize a role of prostaglandin E_2 in the modulation of high-frequency induced synaptic depression, and an appealing possibility would be an activity-dependent release by PSCs. This seems very likely since rat Schwann cells produce different prostanoids *in vitro*. In addition, AA- and prostaglandin E_2-synthesizing enzymes, phospholipase A_2 and cyclooxygenase, are found in frog PSCs.

Neurotrophic factors: BDNF At developing NMJs, acute exposure to BDNF rapidly potentiates spontaneous and evoked synaptic activity. BDNF also enhances synaptic transmission in the adult rat diaphragm muscle. Because cultured Schwann cells produce molecules with BDNF-like activity and possess the mRNA coding for BDNF, this neurotrophic factor may be involved in PSC potentiation of transmitter release.

Thus, many molecules could modulate synaptic transmission and therefore act as gliotransmitters. They may be recruited in different conditions, according to the specific needs of the synapse. For example, activation of a certain pathway following G-protein activation could favor the release of depressing molecules, while the activation of a Ca^{2+} cascade could encourage the release of potentiating molecules. The two pathways may act independently, causing the release of both types of molecules, with the resulting influence on synaptic transmission being a balance of depression and potentiation. As another possibility, the G-protein pathway and the Ca^{2+} pathway may interact such that, downstream of G-protein activation, Ca^{2+} signaling modulates the activity of proteins that are involved in the release of depressing molecules.

Long-Term Modulation by PSCs

In an attempt to elucidate the role of PSCs in synaptic transmission, techniques have been developed to specifically ablate these cells at the NMJ. Using the monoclonal antibody 2A12, specific for PSCs *in vivo*, complement-mediated lysis was performed to eliminate PSCs from adult frog NMJs without damaging nerve terminals, muscle, or myelinating Schwann cells along axons. Acute ablation of PSCs resulted in no obvious defects in evoked synaptic transmission, while the frequency of spontaneous release was slightly increased. Paired-pulse facilitation and synaptic depression of NMJs lacking PSCs were also comparable to these functions in control NMJs. Similar results were obtained in mammals: PSCs were acutely ablated using antibodies specific for disialosyl epitopes on gangliosides (preferentially GD3), which are enriched in the PSC membrane.

In contrast to these acute effects, chronic absence of PSCs leads to long-term morphological and physiological transformations of frog NMJs. After 1 week of PSC ablation, 10% of nerve terminals were retracted, although they were not detached from the muscle. Active zone length, vesicle density, and ACh

receptor distribution were not altered, ruling out degenerative changes in the nerve terminal or postsynaptic fiber. In contrast, the lack of PSCs altered synaptic transmission, decreasing evoked EPP size and average mEPP frequency, without affecting mEPP amplitude. Thus, synaptic efficacy was decreased in PSC-ablated NMJs, but postsynaptic function remained unchanged. Paired-pulse facilitation and synaptic depression were also decreased, supporting a role for PSCs in the maintenance of synaptic function. These results thus raise the possibility of a role for PSCs in long-term synaptic plasticity.

Conclusion

Taken together, studies on PSCs have revealed the importance of synapse–glia interactions on NMJ stability and have proved the necessity of considering these cells as major players in short-term and long-term modulation of neuromuscular activity. The ability of PSCs to detect and respond to synaptic activity and to modulate neurotransmission likely allows them to regulate activity levels with the effect of maintaining optimal nerve–muscle communication. Future work will undoubtedly shed light on the patterns of synaptic activity that most efficiently recruit PSCs in physiological conditions and that prompt glial modulation of synaptic transmission. Further insights on PSC function will also arise from comparative studies linking PSC activation and modulation properties to functional differences between NMJs of mammals and lower vertebrates.

See also: Calcium Channels; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Non-Quantal Release of Transmitter; Presynaptic Events in Neuromuscular Transmission; Schwann Cells and Plasticity of the Neuromuscular Junction; Synaptic Depression.

Further Reading

Auld DS and Robitaille R (2003) Glial cells and neurotransmission: An inclusive view of synaptic function. *Neuron* 40(2): 389–400.

Auld DS, Colomar A, Belair EL, et al. (2003) Modulation of neurotransmission by reciprocal synapse–glial interactions at the neuromuscular junction. *Journal of Neurocytology* 32(5–8): 1003–1015.

Bourque MJ and Robitaille R (1998) Endogenous peptidergic modulation of perisynaptic Schwann cells at the frog neuromuscular junction. *Journal of Physiology* 512(part 1): 197–209.

Castonguay A and Robitaille R (2001) Differential regulation of transmitter release by presynaptic and glial $Ca2^{+}$ internal stores at the neuromuscular synapse. *Journal of Neuroscience* 21(6): 1911–1922.

Colomar A and Robitaille R (2004) Glial modulation of synaptic transmission at the neuromuscular junction. *Glia* 47(3): 284–289.

Green AC, Dowdall MJ, and Richardson CM (1997) ATP acting on P2Y receptors triggers calcium mobilization in Schwann cells at the neuroelectrocyte junction in skate. *Neuroscience* 80(2): 635–651.

Halstead SK, Morrison I, O'Hanlon GM, et al. (2004) Anti-disialosyl antibodies mediate selective neuronal or Schwann cell injury at mouse neuromuscular junctions. *Glia* 52(3): 177–189.

Jahromi BS, Robitaille R, and Charlton MP (1992) Transmitter release increases intracellular calcium in perisynaptic Schwann cells *in situ*. *Neuron* 8(6): 1069–1077.

Malomouzh AI, Nikolsky EE, Lieberman EM, et al. (2005) Effect of *N*-acetylaspartylglutamate (NAAG) on non-quantal and spontaneous quantal release of acetylcholine at the neuromuscular synapse of rat. *Journal of Neurochemistry* 94(1): 257–267.

Reddy LV, Koirala S, Sugiura Y, et al. (2003) Glial cells maintain synaptic structure and function and promote development of the neuromuscular junction *in vivo*. *Neuron* 40(3): 563–580.

Robitaille R (1998) Modulation of synaptic efficacy and synaptic depression by glial cells at the frog neuromuscular junction. *Neuron* 21(4): 847–855.

Robitaille R, Bourque MJ, and Vandaele S (1996) Localization of L-type Ca^{2+} channels at perisynaptic glial cells of the frog neuromuscular junction. *Journal of Neuroscience* 16(1): 148–158.

Silinsky EM and Redman RS (1996) Synchronous release of ATP and neurotransmitter within milliseconds of a motor nerve impulse in the frog. *Journal of Physiology* 492(part 3): 815–822.

Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission

A M Holohean and K L Magleby, University of Miami School of Medicine, Miami, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Short-term stimulation-induced changes in transmitter release have been characterized in terms of various underlying components. Facilitation, augmentation, and potentiation are components of increased transmitter release that differ from one another on the basis of their kinetic properties and time courses of decay. Depression is a component of decreased transmitter release. The properties and possible mechanisms of these components of variable synaptic efficacy are described in the following sections.

Techniques of Study

Variable synaptic efficacy has been studied extensively at many different synapses, including the neuromuscular junction. The frog neuromuscular junction will form the basis for much of the material presented herein. In a typical experiment the sartorius muscle is placed in a small chamber and bathed with a simple salt solution containing 114 mM NaCl, 2 mM KCl, 0.1–2 mM $CaCl_2$, 0–15 mM $MgCl_2$, and a pH buffer. The motor nerve leading to the muscle is stimulated with a train of impulses, and the resulting change in transmitter release after each impulse is determined from changes in the amplitude of successive postsynaptic responses. Postsynaptic responses are determined either from endplate potential (EPP) amplitudes measured using a microelectrode placed in the muscle cell at the region of an endplate, or by measuring the current that flows through the endplate (endplate current) using a two-microelectrode voltage clamp. For experimental conditions in which the EPP amplitudes exceed threshold for generation of action potentials in the muscle, the postsynaptic receptors can be partially blocked with tubocurare to prevent the muscle from contracting. Since the purpose of these studies is to examine stimulation-induced changes in synaptic efficacy, the response to the first impulse in the train serves as a control for the changes that occur during and after the train. Sufficient time is allowed between each train of impulses so that the synapse returns to its unstimulated (control) state.

Stimulation-induced changes in synaptic efficacy have been found to involve a number of different processes or components. To facilitate the study of the individual components, it is advantageous to modify the external media by reducing or increasing the concentration of Ca^{2+} or Mg^{2+}. Reducing Ca^{2+} or increasing Mg^{2+} decreases the amount of transmitter released so that depression due to depletion of synaptic vesicles no longer masks the processes that act to increase transmitter release. An example of variable synaptic efficacy under conditions of decreased transmitter release is shown in Figure 1. Transmitter release was decreased by reducing the amount of extracellular Ca^{2+} available for entry into the nerve terminal and by increasing the amount of extracellular Mg^{2+} which inhibits Ca^{2+} entry. Each record presents EPPs recorded during 5 s of repetitive stimulation at the indicated stimulation rate. Because of the slow time base, each EPP appears as a vertical line. The amplitudes of the EPPs, which are a good measure of transmitter release under the conditions of this experiment, increase during the trains, with the greatest increase at the higher stimulation rates. The enhancement in transmitter release shown in Figure 1 is due to an increase in the number of synaptic vesicles that release their contents of transmitter from the presynaptic nerve terminal into the synaptic cleft for each impulse. This increased release under conditions of low Ca^{2+} can be accounted for in terms of four components: potentiation, augmentation, and the first and second components of facilitation. Each component of increased transmitter release is defined in a similar manner, as the fractional increase in transmitter release over the control level in the absence of the other components, such that

$$y(t) = (\mathrm{EPP}(t) - \mathrm{EPP}_0)/\mathrm{EPP}_0 \qquad [1]$$

where $y(t)$ is the magnitude of the specified component of increased transmitter release at time t, EPP $(t) - \mathrm{EPP}_0$ is the increase in EPP amplitude at time t due to the component, and EPP_0 is the control EPP amplitude when the component equals zero. The components have been distinguished from one another on the basis of widely separated decay rates and differential sensitivities to Sr^{2+} and Ba^{2+}. The time courses of decay of the components are determined by placing test impulses at various intervals after one or more conditioning impulses. Since the decay of each component can be approximated by a single exponential, decay rates are characterized in terms of time constants – the time required to decay to 37% ($1/e$) of the initial magnitude. The properties of each component are summarized in the following sections.

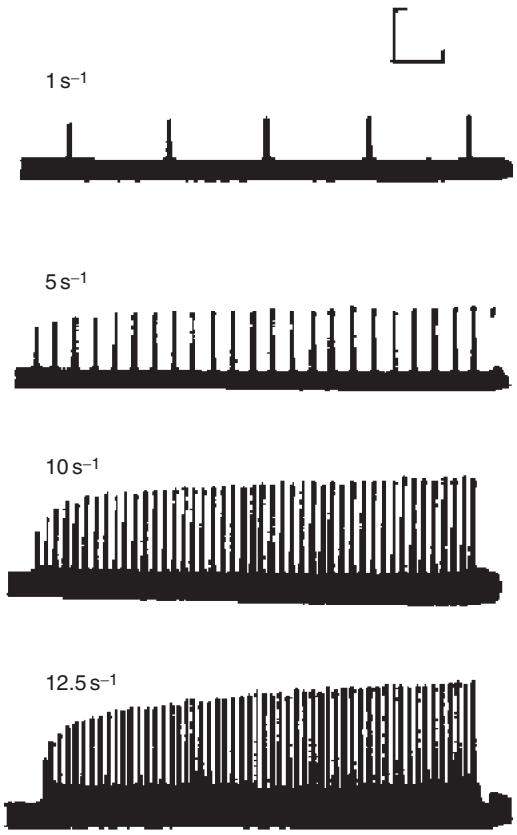

Figure 1 Variable synaptic efficacy at the frog sartorius neuromuscular junction under conditions of low levels of transmitter release. Each record presents endplate potentials recorded with a surface electrode when the nerve was stimulated at the indicated rate. Each vertical line represents an endplate potential. The amplitude of the endplate potentials is proportional to the amount of transmitter released by each nerve impulse. Transmitter release increased during the trains, and the increased release was greater for higher stimulation rates. Vertical bar = 0.5 mV; horizontal bar = 0.5 s. Reproduced from Magleby KL (1973) The effect of repetitive stimulation on facilitation of transmitter release at the frog neuromuscular junction. *Journal of Physiology (London)* 234: 327–352, with permission.

Facilitation of Transmitter Release

Facilitation includes two components of increased transmitter release. Each nerve impulse adds a first and second component of facilitation. The first component increases transmitter release about 80% above the control level and then decays with a time constant of about 50 ms. The second component increases release about 12% and then decays with a time constant of about 300 ms. The facilitation added by a single nerve impulse can be described by

$$f(t) = 0.8e^{-t/50\,\mathrm{ms}} + 0.12e^{-t/300\,\mathrm{ms}} \qquad [2]$$

where $f(t)$ is the fractional increase in transmitter release at time t following the conditioning impulse.

During repetitive stimulation, the facilitation added by each release combines with that remaining from previous impulses, so that facilitation builds up during the train, reaching a steady-state level after approximately 1 s of stimulation. The greater the frequency of stimulation, the greater the combined facilitation. Because of their large magnitudes and rapid time courses of decay, stimulation-induced changes in the two components of facilitation can lead to large and rapid changes in transmitter release following step changes in stimulation rate. Facilitation decays to insignificant levels within about 1 s after a conditioning train ends. Replacing Ca^{2+} in the bathing solution with Sr^{2+} leads to a selective increase in the magnitude and time constant of the second component of facilitation.

Augmentation of Transmitter Release

Each nerve impulse adds an increment of augmentation which increases transmitter release about 1% above the control level. Augmentation then decays with a time constant of about 7 s. The augmentation added by each impulse can be described by

$$a(t) = 0.01\,e^{-t/7\,\mathrm{s}} \qquad [3]$$

where $a(t)$ is augmentation at time t. The increment of augmentation added by each impulse typically increases above the 0.01 level during trains of several hundred impulses. Because of its slow 7 s decay rate and the progressive increase in the increment of augmentation added by each impulse, augmentation can build up to large levels during several hundred impulses in the absence of depression, acting to increase the transmitter released by each nerve impulse several fold in the absence of depression (see later). Addition of small amounts of Ba^{2+} to the bathing solution leads to a selective increase in the magnitude of augmentation while having little or no effect on its time constant of decay.

Potentiation and Posttetanic Potentiation of Transmitter Release

Each nerve impulse adds an increment of potentiation that increases transmitter release about 1% above the control level. Potentiation then decays with a time constant that ranges from tens of seconds to minutes. The time constant of decay, which is about 60 s after the end of a 300-impulse conditioning train, increases with the number and frequency of conditioning impulses. Potentiation after a train can develop with a delay if depression (see later) occurs during the conditioning train. Under these conditions potentiation is often called posttetanic potentiation (PTP), because

it becomes apparent with a delay after the tetanus due to the slow recovery from the masking depression. The potentiation added by each impulse may be approximated by

$$p(t) = 0.01\,\mathrm{e}^{-t/60\,\mathrm{s}} \qquad [4]$$

where $p(t)$ is potentiation at time t. Potentiation builds up during repetitive stimulation, often increasing release for each impulse two times the control level after several hundred impulses.

Relationship between the Components of Increased Transmitter Release

Potentiation, augmentation, and the two components of facilitation retain their kinetic properties during and following repetitive stimulation. The relationship between these four components is not entirely clear, but their effects do not add linearly to one another. Instead, there appears to be a multiplicative relationship among the components. This is expressed as

$$\mathrm{EPP}(t)/\mathrm{EPP}_0 = (F(t)+1)(A(t)+1)(P(t)+1) \qquad [5]$$

where $\mathrm{EPP}(t)/\mathrm{EPP}_0$ is the increase in transmitter release at time t over the control level in the absence of repetitive stimulation, and $F(t)$, $A(t)$, and $P(t)$ are the magnitudes of facilitation, augmentation, and potentiation at time t.

The combined properties of the four components of increased transmitter release are sufficient to account for stimulation-induced changes in transmitter release under conditions of low quantal release (e.g., low extracellular Ca^{2+}). The four components build up during repetitive stimulation and decay thereafter, with their characteristic time constants. Under conditions of low quantal release the combined effects of the four components can increase the transmitter released by each impulse an order of magnitude over the control level after several hundred impulses of repetitive stimulation at 10–20 impulses per second.

Comparison of the Components of Increased Transmitter Release at Three Different Preparations

Even though the time constants of decay of the four components of increased transmitter release are temperature dependent in reptiles and mammals, the decay rates of the components in mammalian sympathetic ganglion at 34 °C and in fascia dentata of rat hippocampus at 35 °C are similar to those stated in the previous sections for the frog neuromuscular junction at 20 °C. The general kinetic properties of facilitation, augmentation and potentiation appear to be similar at frog neuromuscular junction and fascia dentata of rat hippocampus. In contrast, the increments of the second component of facilitation and augmentation added by each impulse in the mammalian sympathetic ganglion are about 6 to 10 times larger than at the frog neuromuscular junction.

Mechanisms for the Components of Increased Transmitter Release

Depolarization of the presynaptic nerve terminal by an action potential opens voltage-sensitive channels in the nerve terminal that are permeable to Ca^{2+}. It is the influx of Ca^{2+} into the nerve terminal through these Ca^{2+} channels that leads to evoked transmitter release. Since Ca^{2+} triggers transmitter release, it might be expected that increases in the amount of Ca^{2+} in the nerve terminal would lead to increases in transmitter release, and it has been suggested that all four components of increased transmitter release are related to increased Ca^{2+} in the nerve terminal. Such a hypothesis is attractive since Ca^{2+} does accumulate in the nerve terminal during repetitive stimulation, and the time course of Ca^{2+} accumulation and removal can be similar to the combined time courses of the components of increased transmitter release.

An increase in intracellular Ca^{2+} during repetitive stimulation could arise from several different mechanisms, including incomplete removal of the Ca^{2+} entering through Ca^{2+} channels with each nerve impulse, saturation of the buffering capacity of the nerve terminal, increased Ca^{2+} entry through the Ca^{2+} channels, or release of Ca^{2+} from intracellular stores. While all of these processes may act to increase Ca^{2+} under some conditions, facilitation appears to be associated with residual Ca^{2+} remaining from preceding impulses. Augmentation follows a similar time course as Ca^{2+} decay in the nerve terminal and therefore may be due to delayed extrusion of residual Ca^{2+} by pumps such as the Na^+/Ca^{2+} exchanger, where a reduction in the Na^+ gradient due to continual stimulation will slow or even reverse the rate of exchange between Na^+ and Ca^{2+}. Potentiation follows a time course similar to the efflux of the Ca^{2+} accumulated in the mitochondria during prolonged stimulation. On this basis, the different time courses of the components of increased release may reflect the time course of diffusion, buffering, and multicompartment uptake and extrusion of Ca^{2+} from the nerve. A residual Ca^{2+} model for the components of increased transmitter release can give an approximate multiplicative relationship between the components if the residual Ca^{2+} is expressed as a power, such that the transmitter release is proportional to the third or fourth

power of Ca^{2+} acting at one or a combination of proteins involved in release.

Additionally, the components of increased Ca^{2+} release could also reflect the time course of binding and release of Ca^{2+} from Ca^{2+} binding sites on proteins involved in synaptic release; an example is synaptotagmin, a vesicle membrane protein which has multiple binding sites for Ca^{2+}. The differential effects of Ba^{2+} and Sr^{2+} could reflect different kinetic properties of these ions on one or more of the multiple proteins involved in vesicle release and/or on the various uptake mechanisms for Ca^{2+}.

There are a number of difficulties with a simple residual (free) Ca^{2+} hypothesis for stimulation-induced increases in transmitter release. The correlation between free Ca^{2+} and facilitation does not always follow the expected properties for such a hypothesis, and the expected concentrations of residual free Ca^{2+} calculated from the observed increases in transmitter release predict that the rate of transmitter release between nerve impulses, measured as the frequency of miniature endplate potentials seen, should be many times higher than is observed.

However, release evoked by an action potential and residual release from miniature endplate potentials may involve different release sites/mechanisms, so such correspondence might not be expected, and such calculations typically ignore the complexity of a multistep release process. It has also been proposed that one reason why residual free Ca^{2+} does not produce the predicted high level of transmitter release is because of a hypothesized voltage component that directly enhances release at the time of the action potential in the presynaptic nerve terminal, although effective release models have been developed without such a voltage-dependent component.

Residual Ca^{2+} could alter a number of factors involved in the transmitter release, giving rise to the components of increased transmitter release. Such factors could include an increase in the number of sites with vesicles available for release and an increase in the probability that a primed site will release the contents of its synaptic vesicle. The mechanism for the multiplicative relationship between the components is less clear, but if release is determined by the joint probability of multiple factors, such as sites and probability, then these factors could have a multiplicative relationship. Release would then be given by the product of these factors.

Studies of the mechanism of transmitter release suggest that a number of Ca^{2+}-dependent steps with different affinities for Ca^{2+} precede exocytosis. On this basis, the time course of decay of some of the components of increased transmitter release may reflect the time course of Ca^{2+} unbinding associated with the various Ca-dependent steps, as well as the time course of removal of Ca^{2+}.

At least nine families of proteins associated with synaptic vesicles and the presynaptic membrane, many of which are expressed in multiple isoforms, have been implicated in the release of transmitter. These proteins are involved in synaptic vesicle turnover, docking of the vesicles at the release sites, priming to enable the vesicles to be released, and the release process itself that leads to the exocytosis of the vesicle contents. When the functions of all these proteins are clearly defined, as well as the properties of the systems involved in the uptake, sequestration, and extrusion of Ca^{2+}, then the precise molecular mechanisms of facilitation, augmentation, and potentiation may also become clear.

Depression of Transmitter Release

Depression is a decrease in postsynaptic potential amplitude due to a decrease in the number of synaptic vesicles that release their contents into the synaptic cleft by each nerve impulse. Depression is typically insignificant under conditions of low levels of transmitter release (low Ca^{2+}) and becomes more pronounced the greater the level of release. Depression typically develops during prolonged or high-frequency repetitive stimulation under physiological conditions. The magnitude of depression, measured as the fractional decrease in transmitter release, increases with higher stimulation rates or increased duration of stimulation. In physiological Ca^{2+}, the higher the stimulation rate, the greater the resulting depression of transmitter release.

Figure 2 presents an example of depression during repetitive stimulation of the rat diaphragm neuromuscular junction (Figure 2(a)) and the frog sartorius neuromuscular junction (Figure 2(b)) under conditions of normal levels of transmitter release. The postsynaptic response recorded in each case was the endplate current (EPC); inward current through the endplate is plotted downward. Shown in Figure 2(a), EPC amplitudes are successively decreased during a short train of 150 stimuli per second. In Figure 2(b), successive EPC amplitudes first increase and then decrease during a train of 100 stimuli per second. The decrease in amplitude, which results from a decrease in transmitter release, is referred to as depression. Depression is typically superimposed on or masks facilitation, augmentation, and potentiation, which also develop during repetitive stimulation, even under conditions of normal or increased quantal release. The increase in EPC amplitudes in Figure 2(b), before depression becomes prominent, is due to the components of increased transmitter release. Depression is pronounced in Figure 2 over what would be

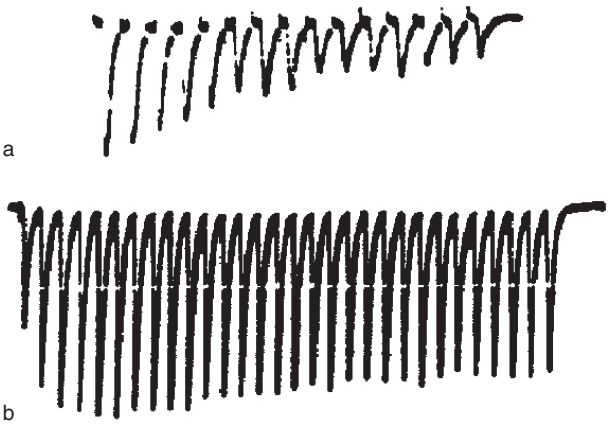

Figure 2 Depression of transmitter release at rat diaphragm (a) and frog sartorius (b) neuromuscular junctions under physiological levels of transmitter release. Each record presents endplate currents recorded under voltage clamp from a neuromuscular junction. The nerve was stimulated at 150 impulses per second in (a) and 100 impulses per second in (b). Inward current through the endplate is plotted downward, and the absolute endplate current amplitude gives a measure of the amount of transmitter released. The decrease in transmitter release that occurs after the first impulses in (a) and after 5 impulses in (b) results from depression. The first endplate current amplitude in (a) is 100 nA and the endplate currents were evoked every 6.67 ms. The first endplate current amplitude in (b) is 140 nA and endplate currents occur every 10 ms. Reproduced from Magleby KL, Pallotta BS, and Terrar DA (1981) The effect of (+)-tubocurarine on neuromuscular transmission during repetitive stimulation in the rat, mouse, and frog. *Journal of Physiology (London)* 312: 97–113, with permission.

observed under physiological stimulation rates because of the high stimulation rates used in this experiment.

Depression can be divided into two components under physiological levels of transmitter release: a fast component depression that decreases transmitter release about 15% below the control level after a single impulse and recovers with a time constant of about 5 s, and a slow component depression that decreases transmitter release about 0.1% for each impulse and recovers with a time constant of tens of seconds to minutes. The slow component depression is only apparent after prolonged stimulation, and its time course of recovery is slowed as the duration of stimulation increases. At the frog neuromuscular junction, the fast component of depression has been shown to conceal an underlying component of augmentation and the slow component of depression masks the onset of potentiation (posttetanic); the components of augmentation and potentiation increase incrementally during successive impulses of stimuli, but under normal levels of transmitter release the components of depression dominate the response.

The relative amount of depression that develops under physiological levels of transmitter release depends on the synapse and stimulation pattern. At some synapses, such as those in fascia dentata of rat hippocampus, depression does not appear to be significant, and facilitation, augmentation, and potentiation dominate the response. On the other hand, depression is more prominent at rat diaphragm neuromuscular junctions, and tends to mask augmentation and the onset of potentiation.

Mechanisms of Depression

The observation that the magnitude of depression is related to the amount of transmitter released from the nerve terminal has led to the suggestion that depression is due to a depletion of the transmitter immediately available for release. On this basis, the fast component depression might reflect a depletion of the synaptic vesicles at the release sites (the readily releasable pool of vesicles), with a 5 s time constant of recovery reflecting the time required for depleted synaptic vesicles to be replaced from the larger pool of synaptic vesicles in the nerve terminal. A number of factors could give rise to the slow component depression: it may reflect a more general depletion of a larger pool of synaptic vesicles; vesicle release and membrane recycling could decrease the number of functional sites available for transmitter release or decrease their release probability; the coupling between Ca^{2+} channels and release sites may decrease with prolonged release. The depression associated with prolonged, slow (0.5–10 Hz) nerve stimulation in the frog occurs without changes in the entry of Ca^{2+} into the nerve terminal. Thus, decreases in Ca^{2+} entry into the nerve terminal with each impulse do not appear to be the major mechanism for depression in the frog. At some synapses, however, a decrease in Ca^{2+} entry through Ca^{2+} channels may also contribute to depression.

Mobilization is often used to refer to the replacement of depleted synaptic vesicles. The rate of mobilization may be modified by the level of intracellular Ca^{2+}, the activation of second messengers acting on the mobilization machinery, the size of the mobilization pool, and also the fractional depletion of vesicles at the release sites. If the mobilization process overshoots the original level, then mobilization might lead to an increase in transmitter release, or, conversely, if mobilization cannot keep up with release, the result would be a depression of transmitter release. Second messengers acting presynaptically may also be involved in depression and recovery from depression. For example, transmitter release at the neuromuscular junction can be inhibited by the ATP that is co-released from the nerve terminal with acetylcholine, and also by adenosine, a by-product of ATP hydrolysis. These agents may act at some presynaptic site to induce

depression without altering Ca^{2+} entry into the nerve terminal. It is not known whether ATP or adenosine induces depression by slowing mobilization of transmitter or by acting through some other mechanism.

Developing Models to Describe Mechanisms Consistent with Short-Term Plasticity

Given the number of components of increased and decreased transmitter release and their overlapping time constants, together with the large numbers of factors that contribute to transmitter release, such as the numbers of release sites, the fraction of release sites with docked vesicles, the fraction of docked vesicles that are primed for release, the changes in release probability based on Ca^{2+} concentrations at multiple sites, possible cooperativity of multiple binding sites for Ca^{2+} on proteins involved in release, the localization of Ca^{2+} channels and their relationship to release sites, the diffusional microdomains of Ca^{2+}, the intracellular Ca^{2+} buffers, the uptake and storage of Ca^{2+} by the mitochondria, the extrusion of Ca^{2+} by Na^{+}/Ca^{2+} exchangers, and the multiple pools of synaptic vesicles with different rates of replenishment and recycling (just to name some of the factors), models of transmitter release must be formulated and tested in terms of solutions of simultaneous differential equations, as the interactions among the various factors make it difficult if not impossible to study a single factor in isolation. Considerable progress has been made using this approach to develop realistic physical models of short-term synaptic plasticity that include many of the complex features of transmitter release.

Short-Term Synaptic Memory

Facilitation, augmentation, potentiation, and depression are different components of short-term synaptic memory (synaptic plasticity). These components act to integrate information about previous synaptic activity over time, and the synapse indicates its memory of the previous activity by releasing a different amount of transmitter to subsequent nerve impulses. This short-term synaptic memory provides a ready means for the nervous system to store information about previous synaptic activity for a period of milliseconds to several minutes. The multiplicative relationship between the components would provide a means for low-frequency input to a synapse to modulate higher frequency input. It is not known, however, to what extent the nervous system makes functional use of this short-term synaptic memory, or whether the nervous system is designed to operate reliably in spite of it, but such short-term synaptic plasticity is present in the central nervous system and has been shown to have profound effects on the efficacy of synaptic transmission.

See also: Calcium Channels; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission; Presynaptic Events in Neuromuscular Transmission; Presynaptic: Mitochondria and Presynaptic Function; Synaptic Plasticity: Short-Term Mechanisms.

Further Reading

Augustine GJ, Santamaria F, and Tanaka K (2003) Local calcium signaling in neurons. *Neuron* 40: 331–346.

Delaney KR, Zucker RS, and Tank DW (1989) Calcium in motor nerve terminals associated with posttetanic potentiation. *Journal of Neuroscience* 9: 3558–3567.

Gandhi SP and Stevens CF (2003) Three modes of synaptic vesicular recycling revealed by single-vesicle imaging. *Nature* 423: 607–613.

Jackson MB and Chapman ER (2006) Fusion pores and fusion machines in Ca^{2+}-triggered exocytosis. *Annual Review of Biophysics and Biomolecular Structure* 35: 135–160.

Kalkstein JM and Magleby KL (2004) Augmentation increases vesicular release probability in the presence of masking depression at the frog neuromuscular junction. *Journal of Neuroscience* 24: 11391–11403.

Magleby KL (1973) The effect of repetitive stimulation junction. *Journal of Physiology* (*London*) 234: 327–352.

Magleby KL (1987) Short-term changes in synaptic efficacy. In: Edelman GM, Gall WE, and Cowan WM (eds.) *Synaptic Function,* 1st edn., pp. 21–56. New York: Wiley.

Magleby KL and Zengel JE (1982) Quantitative description of stimulation-induced changes in transmitter release at the frog neuromuscular junction. *Journal of General Physiology* 80: 613–628.

Magleby KL, Pallota BS, Terrar DA (1981) The effect of (+)-tubocurarine on neuromuscular transmission during repetitive stimulation in the rat, mouse, and frog. *Journal of Physiology* (*London*) 312: 97–113.

Millar AG, Zucker RS, Ellis-Davies GCR, et al. (2005) Calcium sensitivity of neurotransmitter release differs at phasic and tonic synapses. *Journal of Neuroscience* 25: 3113–3125.

Rizzoli SO and Betz WJ (2005) Synaptic vesicle pools. *Nature Reviews Neuroscience* 6: 57–69.

Schneggenburger R and Neher E (2005) Presynaptic calcium and control of vesicle fusion. *Current Opinion in Neurobiology* 15: 266–274.

Schneggenburger R and Forsythe ID (2006) The calyx of held. *Cell and Tissue Research* 326: 311–337.

Stevens CF and Wesseling JF (1999) Augmentation is a potentiation of the exocytotic process. *Neuron* 22: 139–146.

Südhof TC and Scheller RH (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Van Der Kloot Q and Molgo J (1994) Quantal acetylcholine release at the vertebrate neuromuscular junction. *Physiological Reviews* 74: 899–991.

Von Gersdorff H and Borst JGG (2002) Short-term plasticity at the calyx of held. *Nature Reviews* 3: 53–64.

Zucker RS and Regehr WG (2002) Short-term synaptic plasticity. *Annual Review of Physiology* 64: 355–405.

Neuromuscular Junction (NMJ): Presynaptic Stretch Effects on Neuromuscular Transmission

A D Grinnell, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

How Muscle Stretch Influences Neuromuscular Transmission

One of the most robust forms of synaptic plasticity described to date is the modulation of neurotransmitter release from frog motor nerve terminals by muscle stretch. Fatt and Katz, in their classic paper first describing the quantal nature of neurotransmission, noted that "Stretching a muscle 10–15% beyond its resting length reversibly increased the rate by a factor of 2.5–3, and pulling on the motor nerve sometimes caused a similar 'speeding up' of the discharge." Indeed, subsequent research has shown that over the physiological range of muscle length, from about 5–10% below rest length to 15–20% muscle stretch, miniature endplate potential (mEPP) frequency increases by about 10% for each 1% change in length. The effect is fully reversible. Following partial curarization, or at a reduced [Ca^{2+}] where quantal content can be measured, EPP amplitude shows comparable enhancement (Figure 1).

The Kinetics of Stretch-Induced Enhancement of Release

In the 1990s, Grinnell and his associates turned their attention to the kinetics and possible mechanisms of the stretch effect. Using 'floating electrodes' and a computer-driven stretching/shortening apparatus that could change the length of a 30 mm-long frog sartorious muscle linearly by as much as 4 mm at rates up to 1 mm/5 ms, they confirmed that the enhancement of EPP amplitude is linear with length, maintained constant at a new length, and fully reversible. Remarkably, there is virtually no delay between the change in length and the corresponding change in EPP amplitude (Figure 2).

Lack of Dependence on Ca^{2+} Influx or Release of Ca^{2+} from Internal Stores

Given the importance of Ca^{2+} in triggering neurotransmitter release, an obvious hypothesis is that stretch increases Ca^{2+} influx or causes release of Ca^{2+} from internal stores. Neither appears to be the case. The enhancement of spontaneous release is not blocked by ω-conotoxin GVIA, which blocks N-type Ca^{2+} channels and EPPs in frog muscle, and it persists in a Ringer containing only 0.1 mM Ca^{2+} and 10 mM Mg^{2+}. Even after an hour's incubation in a Ringer containing 0 mM Ca^{2+}, 2 mM Mg^{2+} and 1 mM ethylene glycol tetracetic acid (EGTA), which blocked EPPs and reduced the mean resting mEPP frequency by about 50%, the stretch effect was not blocked, although it too was reduced by about 50%. Thus, Ca^{2+} influx through stretch-activated or any other kind of channel is not required for the stretch enhancement of spontaneous release, although some Ca^{2+} in the external environment appears to enhance the effect.

Several experiments also ruled out the possibility that nerve terminal stretch somehow causes the release of Ca^{2+} from internal stores. When junctions were incubated in 25 μM 1,2-bis-(o-aminophenoxy)-ethane-*N*,*N*,*N*′,*N*′-tetraacetic acid tetraacetoxy methyl ester (BAPTA-AM) to clamp the intraterminal [Ca^{2+}] at ~100 nM and further treated with 10 μM thapsigargin to block the endoplasmic reticulum Ca^{2+}-ATPase, disabling a major compartment for sequestering or releasing Ca^{2+}, stretch still produced about half the normal enhancement of release. Moreover, the rapidity and linearity of the modulation, the lack of 'overshoot', and the equally rapid and linear reversal of the phenomenon are inconsistent with internal Ca^{2+} release, or indeed with any mechanism on involving biochemical reactions or a second messenger cascade. Finally, when the effects of stretch on mEPP frequency and EPP amplitude were compared at 13° and 22°, the degree of modulation was the same, that is, the temperature effect (Q_{10}) of the effect is ~1. These findings are compelling evidence that the modulation is a purely mechanical phenomenon.

Several hypotheses have been proposed to explain the modulation on the basis of physical phenomena. Frog motor terminals consist of multiple branches, often up to 100–200 μm in length and running parallel to the muscle fiber, slightly invaginated into gutters in the muscle fiber surface. As a muscle fiber is stretched, the nerve terminal branches are proportionately elongated. This could slightly decrease the terminal diameter, arguably forcing water out and causing an increase in the internal Ca^{2+} concentration. However, calculations suggest that a possible change in [Ca^{2+}] due to decrease in volume per unit length of terminal, if it occurs, would be much too small to explain the large change in release efficacy.

Alternatively, stretch of the terminal might expose release sites that, at the shorter length, are cryptic, for example, hidden by folding of the presynaptic membrane or by Schwann cell processes and exposed by

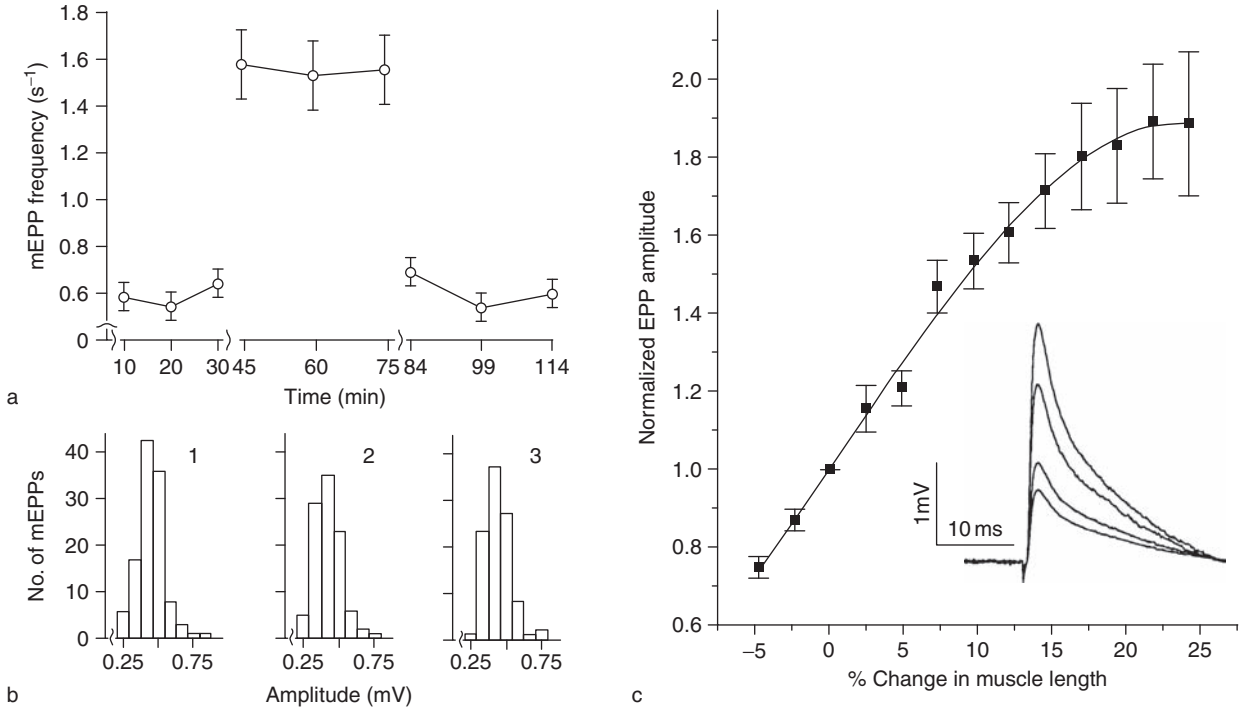

Figure 1 Modulation of neurotransmitter release by stretch. (a) mEPP frequency in a sartorius muscle fiber before, during, and after a 22% stretch of the muscle from a rest length of 2.59 μm/sarcomere. Each point represents the mean ± SE of frequency calculated from the time required for occurrence of 100 mEPPs. (b) Histograms of mEPP amplitude during the prestretch, stretched, and poststretch periods in the same preparation. There was no change in mean mEPP amplitude. (c) Average change in EPP amplitude as a function of changes in muscle length in six sartorius muscle junctions (mean ± SE). The EPP quantal content was reduced in a low-Ca^{2+} frog Ringer containing 0.56 mM Ca^{2+} and 4 mM Mg^{2+}. Inset: EPPs obtained in one preparation at 95%, 100%, 110%, and 120% of rest length. (At rest length, the sarcomere spacing was ~2.25 μm.) (a, b) Reproduced from Turkanis SA (1973) Effects of muscle stretch on transmitter release at end-plates of rat diaphragm and frog sartorius muscle. *Journal of Physiology (London)* 230: 391–403, with permission from Blackwell publishing (c) From Chen B-M and Grinnell AD (1997) Kinetics, Ca^{2+} dependence, and biophysical properties of integrin-mediated mechanical modulation of transmitter release from frog motor nerve terminals. *Journal of Neuroscience* 17(3): 904–916.

stretch. However, many studies of the ultrastructure of frog motor terminals have failed to find any evidence of significant Schwann cell occlusion or folding of the presynaptic membrane that could account for a large increase in exposed active zones with stretch. Thus, the likely mechanism is one that affects the configuration or distance between interacting molecules (or membranes) involved in release, thereby altering release probability.

Integrins as Mediators of the Stretch Modulation

The magnitude and rapidity of the stretch modulation of release suggest that release is influenced by stretch-imposed mechanical stress on the presynaptic membrane near active zones. Because Ca^{2+} influx is not necessary but the effect is reduced in zero-Ca^{2+} Ringer, Ca^{2+}-dependent adhesion interactions with one or more ligands in the extracellular matrix (ECM) are implicated. Obvious candidate adhesion molecules are cadherins and integrins, both of which occur widely at synapses throughout the nervous system. The role of cadherins, if any, has not been investigated, but integrins play a major role in the stretch effect.

Integrins are nearly ubiquitous heterodimeric transmembrane molecules that are the principal receptors connecting the cytoskeleton, intracellularly, to different ligands in the ECM. Ligand bonding can trigger profound changes in cells, including activation of protein kinase C (PKC) and phospholipase C, initiation of tyrosine phosphorylation cascades, induction of nitric oxide (NO) synthase, cytoplasmic alkalinization, and changes in gene expression. Integrins are also capable of transducing mechanical tension in both directions across the cell membrane, potentially bringing reactant molecules closer together or changing reaction equilibria by altering protein–protein conformation or stress on noncovalent bonds.

A large fraction of the extracellular binding by integrins is to ligands containing a specific triplet of amino acids: arginine, glycine, and aspartic acid (RGD). Short peptides that contain the RGD sequence reduce the stretch enhancement of mEPP frequency and EPP amplitude by 40–50%, presumably by displacing

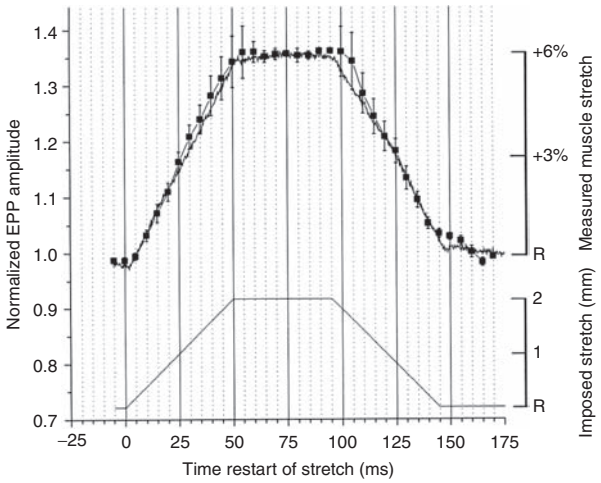

Figure 2 Kinetics of stretch/shortening modulation of EPP amplitude. Average of six experiments in low-Ca^{2+} Ringer containing 0.56 mM Ca^{2+} and 4 mM Mg^{2+}. The muscles were attached at either end to a pair of modified chart recorder arms that moved in opposite directions in response to a command signal from a computer. The stretch began at 0 ms, reached 2 mm (~6%) stretch at 50 ms, was held at that length for 45 ms, and shortened in another 50 ms (thin solid line at bottom). Upper records show the actual measured change in length of this muscle (noisy solid line from phototransistor) and the EPP amplitude evoked by nerve stimulation timed to generate EPPs at 36 different times during the lengthening/shortening cycle. Each point is the average of 50 or more EPPs, each evoked in a single lengthening/shortening cycle. There was virtually no delay between change in muscle length and change in release efficacy. From Chen B-M and Grinnell AD (1997) Kinetics, Ca^{2+} dependence, and biophysical properties of integrin-mediated mechanical modulation of transmitter release from frog motor nerve terminals. *Journal of Neuroscience* 17(3): 904–916.

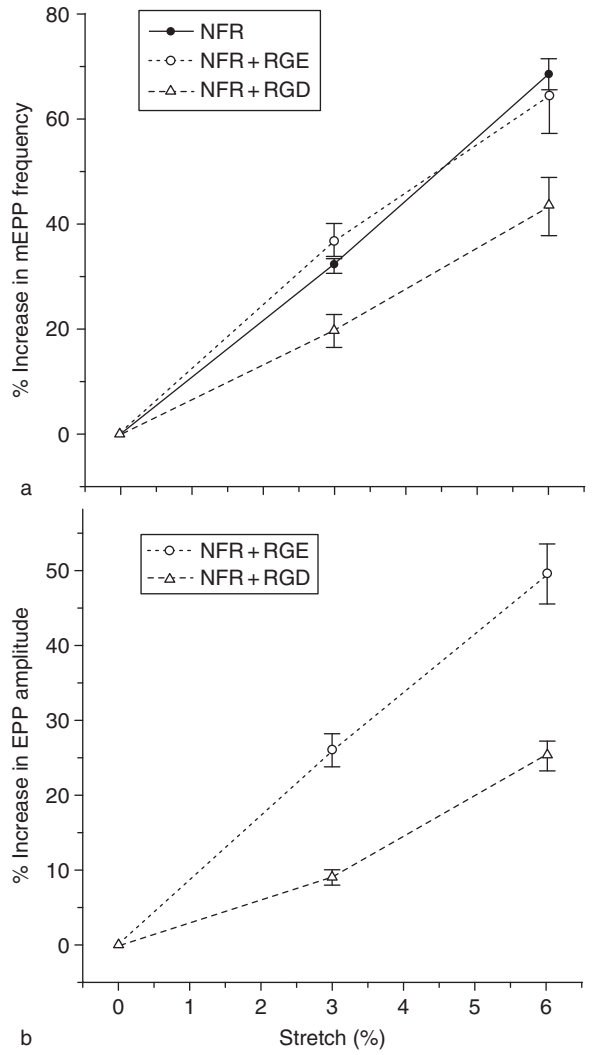

Figure 3 Suppression of the stretch enhancement by RGD peptides, but not the inactive control RGE peptide, at frog sartorius neuromuscular junctions. mEPP frequency (a) was recorded in normal frog Ringer and EPP amplitude in 3–6 μM D-tubocurarine chloride. (b) Muscles were incubated for 1 h or more in 0.2 mM peptide before commencing recording. From Chen B-M and Grinnell AD (1997) Kinetics, Ca^{2+} dependence, and biophysical properties of integrin-mediated mechanical modulation of transmitter release from frog motor nerve terminals. *Journal of Neuroscience* 17(3): 904–916.

many of the integrin–ECM bonds. Inactive control peptides containing the sequence RGE have no effect (Figure 3). Moreover, a function-blocking polyclonal antibody against the $\beta 1$ subunit of *Xenopus* integrin also reduces the stretch enhancement, while the preimmune serum has no effect. The remaining stretch effect in the presence of RGD peptides might indicate that integrins mediate only part of the phenomenon. However, integrin binding is dependent on the presence of Ca^{2+}. By lowering the concentration of Ca^{2+} to 50 μM or less, integrin bonds to native ligands are broken and the suppression of the stretch effect by RGD-containing peptides is much greater.

Although the effect of reducing integrin binding to native ligands with RGD-containing peptides in a low-Ca^{2+} Ringer did not totally block the stretch effect, this can be explained by the probability that many of the native ligand bonds survived the treatment. Other experiments, moreover, support the conclusion that virtually all of the stretch effect can be attributed to integrins. These experiments also help explain the reduction in the stretch enhancement in zero-Ca^{2+} Ringer. When frog nerve–muscle preparations are treated with Ringer containing zero-Ca^{2+} and 0.2 mM Mn^{2+}, the effect of stretch on mEPP frequency is close to that in normal Ringer. Manganese is a potent Ca^{2+} channel blocker but equally potent stabilizer of integrin–ligand bonds. Hence, Mn^{2+} apparently stabilizes integrin bonds that would otherwise be broken in the absence of extracellular Ca^{2+}. On the other hand, in the presence of Mn^{2+} integrins have been shown to bind more avidly to small RGD-containing peptides than to native ligands. Under these conditions,

RGD-containing peptides virtually totally block the stretch modulation of release at the frog neuromuscular junctions (Figure 4). It appears, therefore, that integrin binding to native ligands in the ECM can explain most or all of the stretch effect.

Interfering with integrin–ECM binding does not alter the appearance of frog motor nerve terminals or their elongation with muscle stretch; other connective tissue components are still intact and attach the terminal tightly to the muscle fiber surface. Thus, the effect of reducing integrin–ligand binding appears to be specific to a subset of connections, probably attaching close to the presynaptic release sites, that are not needed for terminal elongation as a muscle is stretched but which profoundly regulate release probability.

The Intracellular Mechanism of Modulation of Release

The nature of the mechanical link between integrins and the release apparatus is still unknown. Integrins attach to the internal cytoskeletal via a number of linking proteins, and tension on integrins can transmit a mechanical stimulus almost instantly over many micrometers distance within the cell. To date, integrins have not been shown in any preparation to be connected, directly or indirectly, to any of the recognized vesicle or membrane soluble NSF attachment protein receptors (SNARES) or other docking/fusion apparatus. It is surely relevant, however, that treatment of frog junctions with okadaic acid, a broad-spectrum

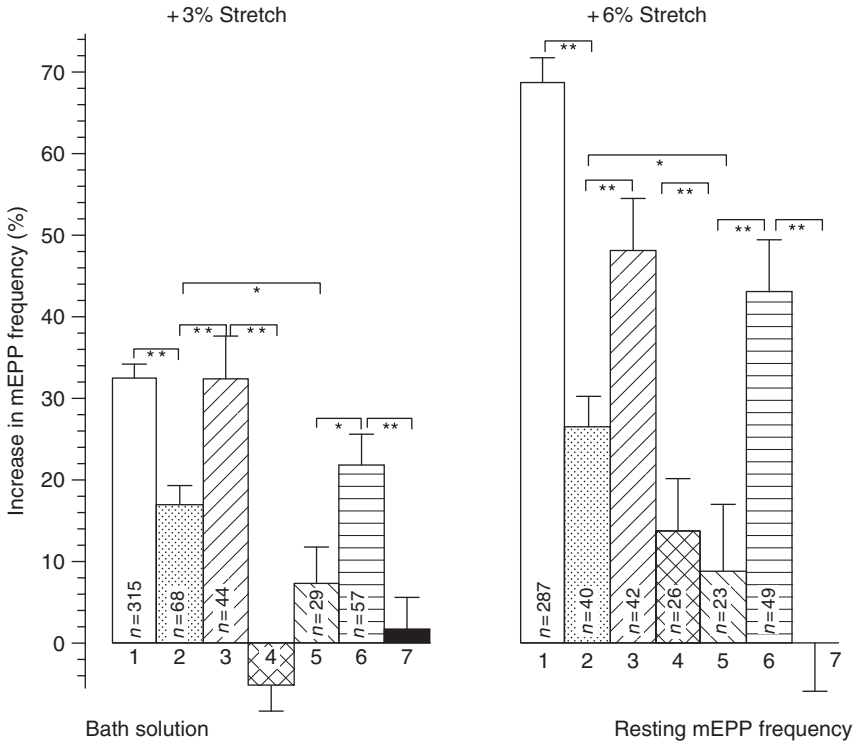

Bath solution	Resting mEPP frequency
1. NFR	1.47 ± 0.06
2. 0 Ca^{2+}	0.69 ± 0.08
3. 0 Ca^{2+} + 0.5 mM Mn^{2+}	1.75 ± 0.20
4. 0.2 mM RGD + 0 Ca^{2+} + 0.5 mM Mn^{2+}, n = 36 (1 mm stretch)	1.63 ± 0.17
5. 0 Ca^{2+} + DM-BAPTA	0.64 ± 0.09
6. 0 Ca^{2+} + DM-BAPTA + 0.5 mM Mn^{2+}	1.64 ± 0.14
7. DM-BAPTA + 0.2 mM RGD + 0 Ca^{2+} + 0.5 mM Mn^{2+} n = 45 (1 mm stretch), n = 36 (2 mm stretch)	0.99 ± 0.14

Figure 4 Stretch enhancement of mEPP frequency at sartorius neuromuscular junctions is not dependent on Ca^{2+} influx and is mostly blocked by 0.2 mM RGD. The stretch enhancement was reduced in 0 Ca^{2+} Ringer (column 2) and severely reduced by loading with BAPTA (column 5). 0.2 mM Mn^{2+}, which stabilizes integrin–ligand bonds, largely restored the stretch effect (columns 3 and 6). Further incubation with 0.2 mM RGD virtually eliminated the stretch effect (columns 4 and 7). From Chen B-M and Grinnell AD (1997) Kinetics, Ca^{2+} dependence, and biophysical properties of integrin-mediated mechanical modulation of transmitter release from frog motor nerve terminals. *Journal of Neuroscience* 17(3): 904–916.

phosphatase inhibitor which causes 'untethering' of synaptic vesicles in the cloud around each active zone, also blocks the stretch effect, and that this effect of okadaic acid can be prevented by preincubation with the kinase inhibitor staurosporine (Figure 5). It would appear that one or more proteins in the mechanical chain between integrins and the release apparatus must be in the unphosphorylated state for the connection to be functional.

Functional Importance and Generality of the Phenomenon

The stretch effect on release probability has been studied almost exclusively in the frog sartorius, in which it is clearly of functional significance. About 30% of the fibers in this muscle, at rest length in normal Ringer, do not generate suprathreshold EPPs to a single stimulation of the nerve. When a muscle is stretched, however, as it would be by contraction of an antagonist muscle, a single nerve impulse can excite most or all of the fibers, that is, the stretch effect is a strong peripheral amplifier of the spinal stretch reflex. Other frog muscles that have been tested, such as the cutaneous pectoris and some toe muscles, also exhibit a stretch effect, even though most or all of their fibers receive suprathreshold input at rest length.

Integrins have been found at a wide variety of synapses and may be present in almost all. In *Xenopus* motor terminals, active zones contain $\alpha 3\ \beta 1$ interins, while at rat neuromuscular junctions, the $\beta 1$ subunit is associated with $\alpha 1$ presynaptically and with α_V and with three different isoforms of $\alpha 7$ postsynaptically. Whether there is stretch enhancement of release in rat junctions is uncertain, however. Studies of rat diaphragm muscle have found no stretch effect, but the motor nerve boutons in most mammalian muscles are on top of distinct postsynaptic pads ('endplates') and might not undergo significant mechanical stretch or distortion when a muscle is stretched. Moreover, EPPs at mammalian neuromuscular junctions normally are suprathreshold at rest length, with a large safety margin, so there would be no obvious advantage in having peripheral amplification of release. Smooth muscle, with multiple innervation and contraction graded in response to EPP amplitude would seem a better candidate for this form of modulation.

The widespread presence of integrins at synapses throughout the central nervous system (CNS) suggests that they may play important roles in nonneuromuscular connections. Evidence is accumulating that this is the case. Integrins are implicated in memory formation and in the consolidation of long-term potentiation (LTP) in hippocampal slices, perhaps stabilizing activity-related change in spine architecture; RGD peptides also cause abnormalities in kindling behavior in hippocampal slices, suggesting changes in

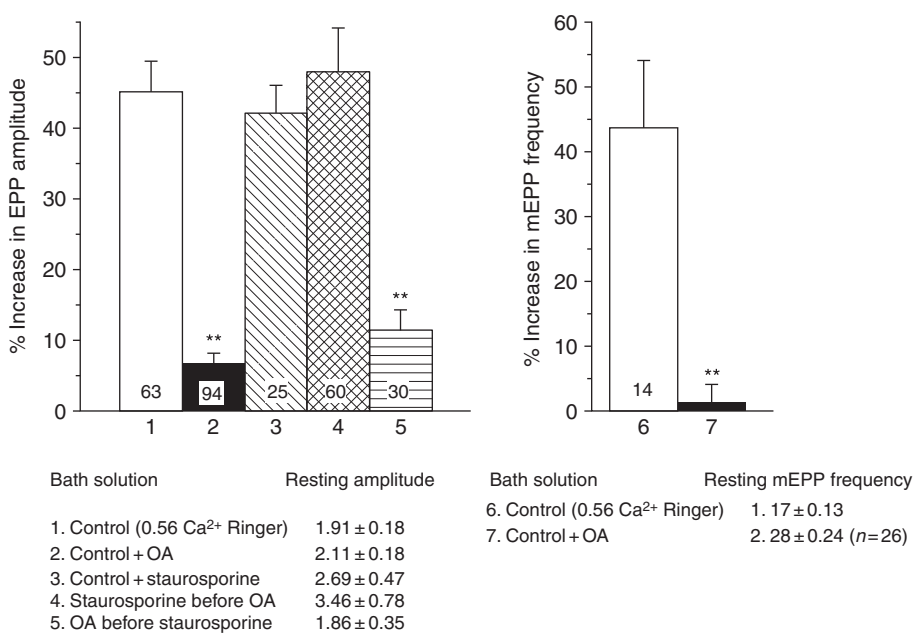

Figure 5 Stretch enhancement of sartorius EPP amplitude and mEPP frequency are severely reduced by ≥ 1 h incubation at 0.3 μM okadaic acid (OA), a broad-spectrum phosphatase inhibitor. This effect of OA was largely prevented by 1–1.5 h preincubation in 0.5 μM staurosporine, a broad-spectrum kinase inhibitor. From Grinnell AD, Chen B-M, Kashani A, Lin J, Suzuki K, and Kidokoro Y (2003) The role of integrins in the modulation of neurotransmitter release from motor nerve terminals by stretch and hypertonicity. *Journal of Neurocytology* 32: 489–503.

excitability. Interference with integrin binding prevents a developmental increase in glutamate release and expression of *N*-methyl-D-aspartate (NMDA)-receptor subunits. In *Drosophila*, mutations in the *volado* gene, which encodes two forms of an α-integrin subunit, cause defective consolidation of olfactory memory. These mutants also exhibit abnormalities in neuromuscular neurotransmitter release and short-term presynaptic plasticity, while mutations of the β-subunit lead to changes in motor nerve branching and neuromuscular morphology.

Thus integrins may play a variety of important roles at synapses – roles that are only beginning to be elucidated. It is noteworthy that RGD peptides that disrupt integrin bonds sharply reduce the enhancement of spontaneous release by hypertonic solutions in both frog and *Drosophila* neuromuscular junctions. Since hypertonicity causes increased spontaneous release in almost all chemical synapses, integrins may play a role in regulating synaptic effectiveness at most synapses.

See also: Calcium Channel Subtypes Involved in Neurotransmitter Release; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Non-Quantal Release of Transmitter; Presynaptic Events in Neuromuscular Transmission; Presynaptic: Mitochondria and Presynaptic Function.

Further Reading

Betz WJ and Henkel AW (1994) Okadaic acid disrupts clusters of synaptic vesicles in frog motor nerve terminals. *Journal of Cell Biology* 124: 843–854.

Chavis P and Westbrook G (2001) Integrins mediate functional pre- and postsynaptic maturation of a hippocampal synapse. *Nature* 411: 317–321.

Chen B-M and Grinnell AD (1995) Integrins and modulation of transmitter release from motor nerve terminals by stretch. *Science* 269: 1578–1580.

Chen B-M and Grinnell AD (1997) Kinetics, Ca^{2+} dependence, and biophysical properties of integrin-mediated mechanical modulation of transmitter release from frog motor nerve terminals. *Journal of Neuroscience* 17(3): 904–916.

Cohen MW, Hoffstrom BG, and Desimone DW (2000) Active zones on motor nerve terminals contain $\alpha 3$ $\beta 1$ integrin. *Journal of Neuroscience* 20: 4912–4921.

Fatt P and Katz B (1952) Spontaneous subthreshold activity at motor nerve endings. *Journal of Physiology (London)* 117: 109–128.

Gailit J and Ruoslahti E (1988) Regulation of the fibronectin receptor affinity by divalent cations. *Journal of Biological Chemistry* 263: 12927–12932.

Greensmith L and Vrbova G (1996) Motoneurone survival: A functional approach. *Trends in Neuroscience* 19: 450–455.

Grinnell AD, Chen B-M, Kashani A, Lin J, Suzuki K, and Kidokoro Y (2003) The role of integrins in the modulation of neurotransmitter release from motor nerve terminals by stretch and hypertonicity. *Journal of Neurocytology* 32: 489–503.

Grotewiel M, Beck C, Wu K, Zhu X, and Davis R (1998) Integrin-mediated short-term memory in *Drosophila*. *Nature* 391: 455–460.

Hutter OF and Trautwein W (1956) Neuromuscular facilitation by stretch of motor nerve endings. *Journal of Physiology (London)* 133: 610–625.

Martin PT, Kaufman SJ, Kramer RH, and Sanes JR (1996) Synaptic integrins in developing, adult, and mutant muscle: Selective association of alpha1, α7A, and α7B integrins with the neuromuscular junction. *Developmental Biology* 174: 125–139.

Martin PT and Sanes JR (1997) Integrins mediate adhesion to agrin and modulate agrin signaling. *Development* 124: 3909–3917.

Staubli U, Chun D, and Lynch G (1998) Time dependent reversal of long-term potentiation by an integrin antagonist. *Journal of Neuroscience* 18: 3460–3469.

Turkanis SA (1973) Effects of muscle stretch on transmitter release at end-plates of rat diaphragm and frog sartorius muscle. *Journal of Physiology (London)* 230: 391–403.

Neuromuscular Junction Plasticity in Mammals and Botulinum Toxins

C R Slater, University of Newcastle, Newcastle upon Tyne, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The highly coordinated activation of skeletal muscles that underlies most animal movement depends on reliable transmission of signals from motor neurons to their target muscle fibers at the neuromuscular junction (NMJ). An important component of this reliability is the release by each nerve impulse of more transmitter than is required to activate muscle contraction. In many species and muscles, the amount of transmitter released from the nerve is approximately proportional to the area of synaptic contact at the NMJ. As a result, an important structural determinant of transmitter output is the size of each junction.

NMJs are distinguished from most synapses in the central nervous system (CNS) by their very large size. This makes physiological sense because the postsynaptic muscle fiber is much larger than most neurons and, accordingly, requires much more synaptic current to activate it. As animals grow, the size of the NMJ keeps up with the increasing dimensions of the muscle fibers to ensure that transmission remains effective. However, in animals such as mammals, in which there is relatively little growth after maturity is reached, the form and size of the NMJs reach a stable state: when repeated observations of the same NMJ are made *in vivo*, there is little change in size or form on a time scale of weeks or months. Although this suggests that the mature mammalian NMJ is relatively static, other evidence belies this impression and reveals an inherent plasticity that plays a vital part in ensuring that effective neuromuscular transmission remains, or is restored, after even severe disruption. The first part of this article describes this plasticity and the evidence for it.

Important insights into plasticity of the NMJ have been gained from observing the response to botulinum toxins, which block transmission by interfering with the release of transmitter from the nerve. These toxins are among the most poisonous substances known. Yet in recent years they have been used increasingly in a clinical context to weaken the contraction of specific muscles in conditions in which the patterns of muscle activation are abnormal, such as the spasticity associated with cerebral palsy. Even more common is the cosmetic use of botulinum toxins to paralyze the facial muscles that cause wrinkling of the skin, in the hope of maintaining a youthful appearance. The latter part of this article deals with a number of issues associated with such uses of these toxins and their impact on NMJ form and function.

Natural Plasticity of the NMJ

Plasticity During Aging

When the NMJs of elderly human subjects are compared with those of younger people, consistent changes in structure are seen. Most obvious is an increase in branching of the terminal motor axon and the presence of numerous small acetylcholine receptor (AChR) clusters associated with the terminals of those branches (Figure 1). In some cases, such branches may come to innervate adjacent muscle fibers. In spite of these structural changes, there is little evidence of any deterioration in the efficacy of neuromuscular transmission. Similar structural changes are seen in old rats and mice in which there may even be an increase in transmitter release. The causes of age-related remodeling of the NMJ are not well established. It is likely that several factors contribute.

The Response to Muscle Damage

One possible trigger of NMJ remodeling is mechanical damage to muscle fibers resulting from mild injury or overexertion. When skeletal muscles are experimentally damaged, for example, by exposure to a myolytic toxin, the muscle fibers may regenerate and become innervated by the surviving nerve terminal. In such cases, the NMJs that reform at the original sites often differ in structural detail from the original ones. As during aging, they tend to consist of numerous small synaptic spots rather than the bands of synaptic contact present normally. An example of this kind of change is seen in muscles of *mdx* mice, which have a form of muscular dystrophy that results in the degeneration and regeneration of most muscle fibers (Figure 2). In spite of the altered form of the restored NMJs, neuromuscular transmission is essentially normal, indicating that changes in overall structure of the NMJ may not necessarily have an important impact on the efficacy of transmission.

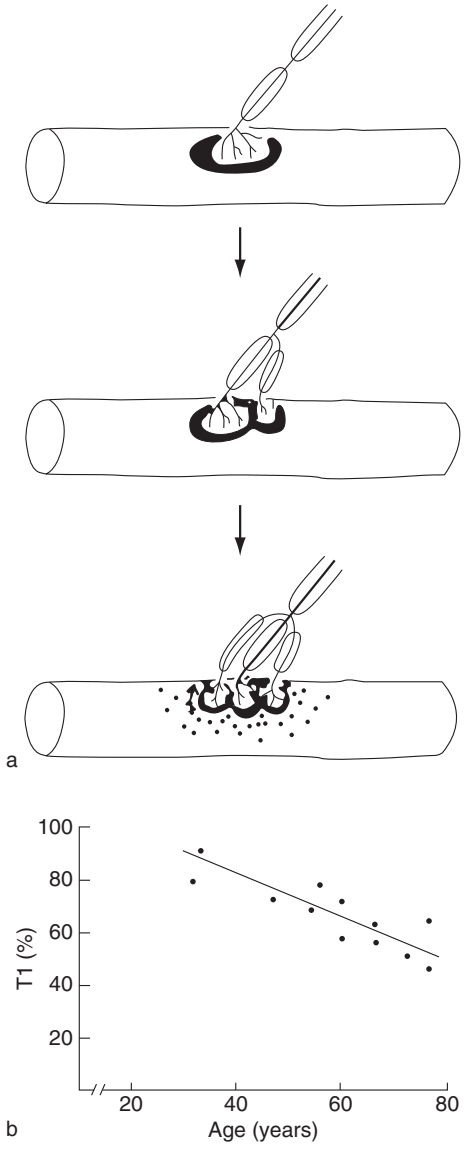

Figure 1 Changes in the structure of human neuromuscular junctions (NMJs) with age: (a) three classes of NMJs based on extent of branching of preterminal motor axon; (b) decrease with age in fraction of NMJs with an unbranched terminal axon (T1 (%)). From Oda K (1984) Age changes of motor innervation and acetylcholine receptor distribution on human skeletal muscle fibres. *Journal of the Neurological Sciences* 66: 327–338.

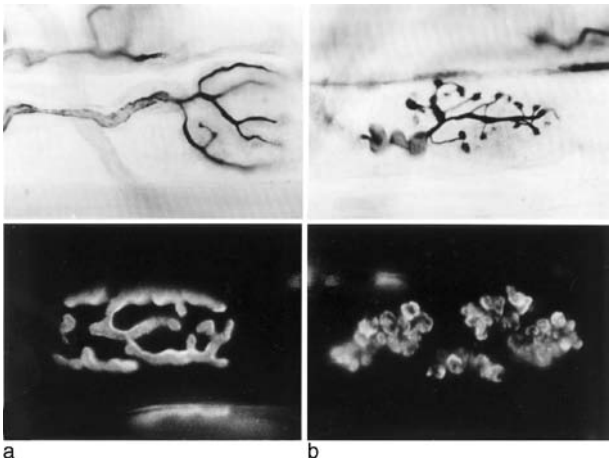

Figure 2 Mouse neuromuscular junction (NMJ) reorganization after muscle damage (a) in control; (b) in *mdx* mouse NMJs following muscle degeneration and regeneration. Upper panels show nerves visualized after silver impregnation. Lower panels show the distribution of acetylcholine receptors (AChRs) visualized by R-α-bungarotoxin labeling. Images are from four different NMJs. Note that following muscle fiber regeneration, the NMJ consists of a cluster of spotlike synaptic regions instead of the continuous bands of contact seen normally. Reproduced from Lyons PR and Slater CR (1991) Structure and function of the neuromuscular junction in young adult mdx mice. *Journal of Neurocytology* 20: 969–981, with permission.

The Response to Nerve Injury

A second likely trigger of NMJ modification with age is the death of some motor neurons. It is well established in humans that the number of functional motor units (a single motor neuron and the muscle fibers it innervates) declines steeply after the age of 60. Death of a motor neuron leaves the muscle fibers it used to innervate in a state of denervation. Experimental studies in which some of the axons innervating a muscle have been destroyed intentionally have shown that such partial denervation is a potent sprouting stimulus for the intact motor axons. This sprouting may originate either from the nerve terminal itself or from the most distal nodes of Ranvier.

Partial denervation When some of the motor axons innervating a rat or mouse muscle are damaged, the muscle fibers so denervated are often reinnervated by newly formed branches of the surviving axons. Within a few days of partial denervation, axonal sprouts grow out from the intact nerve terminals and/or nodes of Ranvier. If such sprouts make contact with denervated NMJs, they are likely to form new synaptic contacts with them. As a result, the size of the surviving motor units gradually increases. In humans, an increase of this kind can be detected by single-fiber electromyography (EMG) as an increase in 'fiber density' (FD). In elderly patients, in whom changes in NMJ structure are common, there is a parallel increase in FD, consistent with the view that some of the changes in NMJ structure may be associated with cycles of partial denervation and reinnervation (Figure 3).

Complete denervation Some forms of injury may lead to the disruption of the entire nerve supply to a muscle. If the nerve is cut through, the axons are likely to grow out from the cut end; however, they may never grow back to the muscle they used to innervate. On the other hand, if the nerve is crushed so that the sheaths of connective tissues and extracellular matrix

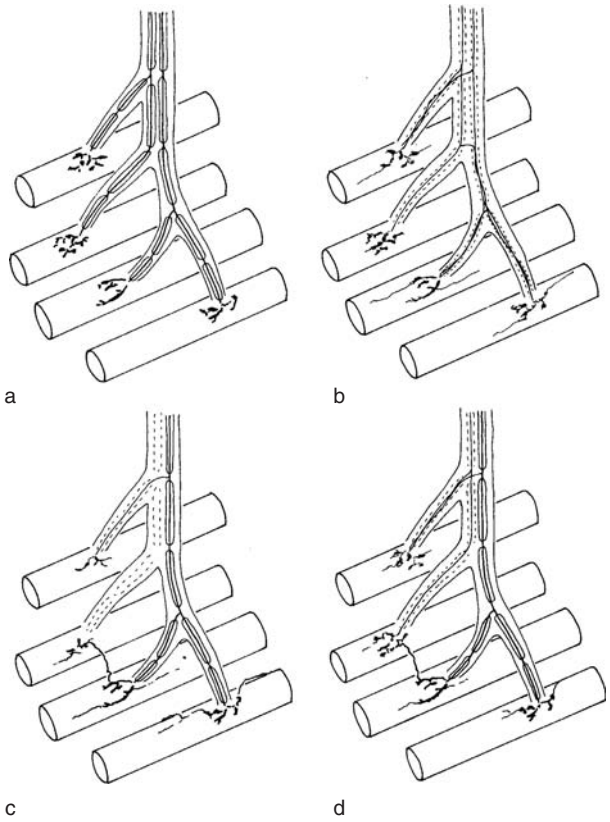

Figure 3 Motor axons forming sprouts after nerve injury: (a) normal; (b) regeneration; (c) sprouting; (d) sprouting and regeneration. In (a), in normal mouse NMJs, the motor axon ends in a highly branched arborization in which each branch remains within the roughly oval region representing the NMJ. In (b), during nerve regeneration, such as after a crush injury of the entire nerve, axons make their way back to the original NMJs, but thin sprouts may also continue to grow beyond the NMJ. In (c), when only some of the axons are damaged, leading to partial denervation, axonal sprouts grow from the intact terminals and may reinnervate denervated muscle fibers. In (d), when damaged axons regenerate after partial denervation, both forms of sprouting may occur. NMJ, neuromuscular junction. Reproduced from Hopkins WG and Brown MC (1985) Growth and remodeling of the neuromuscular synapse. In: Cotman CW (ed) *Synaptic Plasticity*, pp. 169–199. New York: The Guilford Press, with permission from The Guilford Press.

surrounding each axon remain intact, the outcome is more successful. Then many axons may grow back within their original sheaths and eventually reach precisely the same muscle fibers they innervated previously. When the regenerating axons reach the site of the original NMJs, they often reoccupy those sites and effective transmission is rapidly restored.

In some cases, regenerating axons overshoot the original synaptic site. The result is that fine sprouts extend beyond the original NMJ and may even establish some new sites of synaptic contact with the muscle fiber (Figure 3(b)). As a rule, these escaped sprouts regress after a number of weeks. Why the newly made synaptic contacts are less stable than those made at the sites of the original NMJ is not known.

The ability of motor axon sprouts to form new synaptic contacts with adult muscle fibers is a particularly striking aspect of NMJ plasticity. This process of ectopic NMJ formation involves the localized induction by the nerve of all the key components of the postsynaptic apparatus that allow increases in acetylcholine (ACh) concentration to excite the muscle fiber.

The Role of the Schwann Cell

The molecular events associated with the initiation and growth of the axonal sprouts that form the structural basis of NMJ plasticity are generally poorly understood. At a cellular level, however, an important discovery was that within a few days of denervating a mammalian muscle, the Schwann cells that cap the nerve terminal at the NMJ extend sprouts similar to those that grow out from the nerve. These Schwann cell sprouts can become very extensive, forming a dense network over the surface of the muscle. Following reinnervation of the muscle and renewed contractile activity, the Schwann cell sprouts regress and a more or less normal arrangement is restored. When muscles are partially denervated, only those Schwann cells at the denervated NMJs extend sprouts. Thus, the control of Schwann cell sprouting is mediated very locally.

The real significance of the Schwann cell sprouts for NMJ plasticity was revealed by careful examination of the relationship of axon and Schwann cell sprouting in partially denervated muscles (Figure 4). This has shown that axon sprouts generally grow along Schwann cell sprouts, which apparently serve as guides. When an NMJ is denervated, its Schwann cells extend sprouts in all directions but particularly in the direction of intact NMJs on muscle fibers that are still active. If one of those sprouts makes contact with an innervated NMJ, it apparently induces the outgrowth of axonal sprouts that follow the Schwann cell processes back to the denervated NMJ. Thus, the Schwann cell plays a central role in the events that account for the adaptive plasticity of the NMJ.

How Is Nerve Terminal Sprouting Regulated?

The events that account for Schwann cell and axon sprouting during adaptive plasticity of the NMJ are not known with any certainty. It seems likely that both the rearrangement of cytoskeletal proteins and the directed insertion of membrane into growing sprouts are involved. Whatever the mechanisms of

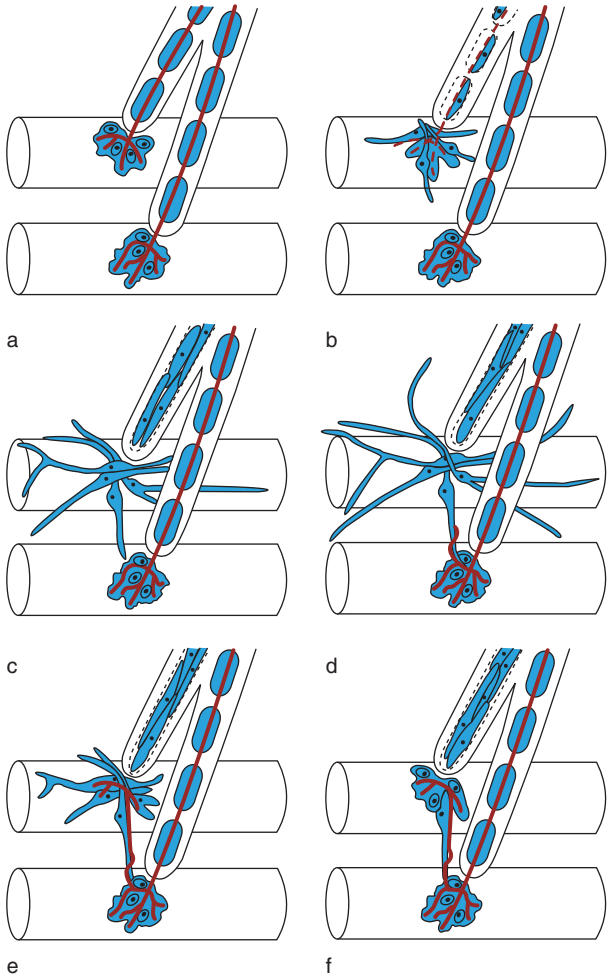

Figure 4 Schwann cells guiding axonal sprouts to denervated neuromuscular junctions (NMJs): (a) normal NMJ (rat); (b) Schwann cells forming sprouts after damage to a motor axon terminal; (c) sprouts growing away from damaged NMJ; (d) sprouts contacting intact NMJs; (e) intact axons growing back along the Schwann cell sprouts; (f) intact axons making contact with the denervated NMJs and reestablishing synaptic contact. Axons are shown in red and Schwann cells are shown in blue in all diagrams. From Son YJ, Trachtenberg JT, and Thompson WJ (1996) Schwann cells induce and guide sprouting and reinnervation of neuromuscular junctions. *Trends in Neuroscience* 19: 280–285.

sprout formation and growth, additional questions remain concerning the factors that trigger these events.

The role of activity The search for factors that induce sprouting has generally started from observations that axonal sprouting is influenced by muscle activity. There is evidence that both increased and reduced levels of activity may induce changes in NMJ structure and function. For example, increased activity associated with exercise training caused an increase in the length and extent of branching of rat NMJs without altering the total area of synaptic contact. In contrast, reduction of activity by unloading muscles caused a reduction in total area without altering nerve branching. The basis of these rather subtle changes is unknown.

In contrast, a complete block of muscle activity, whether by long-term application of local anesthetics to the nerve or by exposure to botulinum toxin to block neuromuscular transmission, induces extensive sprouting. Direct stimulation of muscles in which axonal sprouting has been triggered by exposure to botulinum toxin reduces sprouting from motor axon terminals. It thus seems that the amount of activity can have a long-term regulatory effect on the NMJ plasticity.

Growth factors Many features of muscle fiber biochemistry are influenced by activity. An obvious possibility is that nerve sprouting is triggered by the action of one or more diffusible substances that are released from inactive (but not from active) muscles. A candidate for which there is considerable evidence is insulin-like growth factor (IGF-)1. The expression of IGF-1 by muscle is suppressed by normal electrical activity, but increases following denervation. When muscles are exposed to IGF-1 *in vivo,* it has a sprout-promoting effect. Furthermore, natural sprouting in response to botulinum toxin is reduced if the muscles are exposed to an IGF-1 binding factor IGF-BP4, known to inactivate IGF-1 in *in vitro* assays. The effect of these experimental procedures on the Schwann cells is not known.

Although these observations are intriguing, it is not clear that a single factor could control the entire sprouting response of the nerve at the NMJ in a wide variety of situations. Although various other factors have been reported to have sprout-promoting activity, the evidence that they play a role in the sprouting response *in vivo* is not strong.

Botulinum Toxins and How They Block Neuromuscular Transmission

Natural History of Botulism and Botulinum Neurotoxins

Botulism (from the Latin *botulus,* sausage) is a paralytic disease often caused by eating inadequately preserved food. Since the classic studies of van Ermengem in the 1890s, it has been recognized that the pathogens are neurotoxins synthesized by the bacterium *Clostridium botulinum*. These toxins cause a block of ACh release from motor nerve terminals, leading to muscle weakness or paralysis. It was recently found that the toxins are peptidases that specifically cleave one or more of the proteins that play a central part in synaptic vesicle exocytosis.

Several forms of botulism, varying in the mode of infection, are recognized clinically. In the classic food botulism, the toxins are ingested and reach the NMJs via the bloodstream. In infant botulism, the most common form in some countries, bacterial spores (often present in honey) are ingested and then germinate in the intestinal tract, leading to the proliferation of the pathogenic bacteria and production of the active toxin. The susceptibility of infants to this form of botulism is believed to result from the absence of protective bacterial flora and bile acids that inhibit the proliferation of the bacteria in the adult digestive tract. A third form, wound botulism, arises from the bacteria *C. botulinum* in abscesses and is particularly common in drug abusers.

Clinical features of the disease The main features of botulism are weakness and paralysis. When this involves the respiratory muscles, it is potentially fatal. Indeed, botulinum toxin is one of the most toxic substances known, with an estimated fatal dose of as little as 0.1 μg. However, even in severe cases, recovery is often possible with appropriate management, although it may take weeks before the patient can breathe unassisted, months before ambulation is possible, and years before fully normal movement recovers.

It is clear that clinical botulism results from a block of neuromuscular transmission. Evidence for this comes from both EMG studies and electrophysiological investigations of NMJs in biopsy samples from patients.

Botulinum neurotoxin structure and serotypes There are seven distinct serotypes of botulinum neurotoxin (BoNT/A–G) (Table 1). These are the products of distinct but closely related genes. All are also closely related in structure to tetanus toxin produced by the bacterium *Clostridium tetani,* suggesting that all these toxins arose from a common ancestor. Each toxin has a molecular mass of approximately 150 kDa and consists of a heavy (H) chain of approximately 100 kDa and a light (L) chain of approximately 50 kDa.

Overview of Botulinum Neurotoxin Action

All forms of BoNT are zinc peptidases. Their natural substrates are one or more of the soluble *N*-ethylmaleimide-sensitive factor attached protein receptor (SNARE) proteins that play a central part

Table 1 Botulinum neurotoxin serotypes[a]

Serotype	*Extracellular protein binding target*	*Intracellular enzymatic target*	*Duration of action*		*Sources*[b]
			Mouse/rat NMJ	*Human NMJ*	
A	SV2	SNAP-25	30–90 days	>3 months	[2, 4, 11, 12, 13]
B	Synaptotagmins	Synaptobrevin	21 days	~2 months	[3, 5, 7, 10]
C	?	SNAP-25, syntaxin	~2 months	~2 months	[1, 3]
D	?	Synaptobrevin			[11]
E	?	SNAP-25	4–5 days	< 1 month	[4, 6, 11, 13]
F	?	Synaptobrevin	7 days	1–2 months	[3, 6, 8]
G	Synaptotagmins	Synaptobrevin			[7, 9]

[a]?, unknown; NMJ, neuromuscular junction; SNAP-25, synaptosomal-associated protein of 25 kDa; SV2, synaptic vesicle protein 2.
[b]Data from: 1. Blasi J, Chapman ER, Yamasaki S, Binz T, Niemann H, and Jahn R (1993) Botulinum neurotoxin C1 blocks neurotransmitter release by means of cleaving HPC-1/syntaxin. *EMBO Journal* 12: 4821–4828. 2. Dong M, Yeh F, Tepp WH, et al. (2006) SV2 is the protein receptor for botulinum neurotoxin A. *Science* 312: 592–596. 3. Eleopra R, Tugnoli V, Quatrale R, Rossetto O, and Montecucco C (2004) Different types of botulinum toxin in humans. *Movement Disorders* 19: S53–S509. 4. Eleopra R, Tugnoli V, Rossetto O, De Grandis D, and Montecucco C (1998) Different time courses of recovery after poisoning with botulinum neurotoxin serotypes A and E in humans. *Neuroscience Letters* 256: 135–138. 5. Foran PG, Mohammed N, Lisk GO, et al. (2003) Evaluation of the therapeutic usefulness of botulinum neurotoxin B, C1, E, and F compared with the long lasting type A. *Journal of Biological Chemistry* 278: 1363–1371. 6. Meunier FA, Lisk G, Sesardic D, and Dolly JO (2003) Dynamics of nerve terminal remodeling unveiled using SNARE-cleaving botulinum toxins: The extent and duration are dictated by the sites of SNAP-25 truncation. *Molecular and Cellular Neuroscience* 22: 454–466. 7. Rummel A, Eichner T, Weil T, et al. (2007) Identification of the protein receptor binding site of botulinum neurotoxins B and G proves the double receptor concept. *Proceedings of the National Academy of Sciences of the United States of America* 104: 359–364. 8. Schiavo G (1993) Botulinum neurotoxin serotype F is a zinc endopeptidase specific for VAMP/synaptobrevin. *Journal of Biological Chemistry* 268: 11516–11519. 9. Schiavo G (1994) Botulinum G neurotoxin cleaves VAMP/synaptobrevin at a single Ala-Ala peptide bond. *Journal of Biological Chemistry* 32: 202123–20216. 10. Schiavo G, Benfenati F, Poulain B, et al. (1992) Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin. *Nature (London)* 359: 832–835. 11. Schiavo G, Rossetto O, Catsicas S, et al. (1993) Identification of the nerve terminal targets of botulinum neurotoxin serotypes A, D, and E. *Journal of Biological Chemistry* 268: 23784–23787. 12. Tonge DA (1974) Chronic effects of botulinum toxin on neuromuscular transmission and sensitivity to acetylcholine in slow and fast skeletal muscle of the mouse. *Journal of Physiology* 241: 127–139. 13. Washbourne P, Pellizzari R, Rossetto O, et al. (1998) On the action of botulinum neurotoxins A and E at cholinergic terminals. *Journal of Physiology (Paris)* 92: 135–139.

in the process of quantal release of ACh from the motor nerve terminal (Figure 5). Because these proteins are all intracellular, the toxins can gain access to them only by crossing the plasma membrane to enter the cytoplasm of the nerve terminal.

Binding and Activation of Botulinum Neurotoxins

Binding targets The first step in the action of BoNT at the NMJ is binding to a specific site on the extracellular surface of the nerve terminal. This binding is achieved by the C-terminal region of the H chain. The binding sites on the nerve appear to consist of two parts: a protein and a ganglioside. It was recently found that the protein component of the binding site for BoNT/A is the extracellular domain of SV2, an integral protein of the synaptic vesicle membrane that is exposed on the surface of the nerve terminal by vesicle exocytosis during neuromuscular

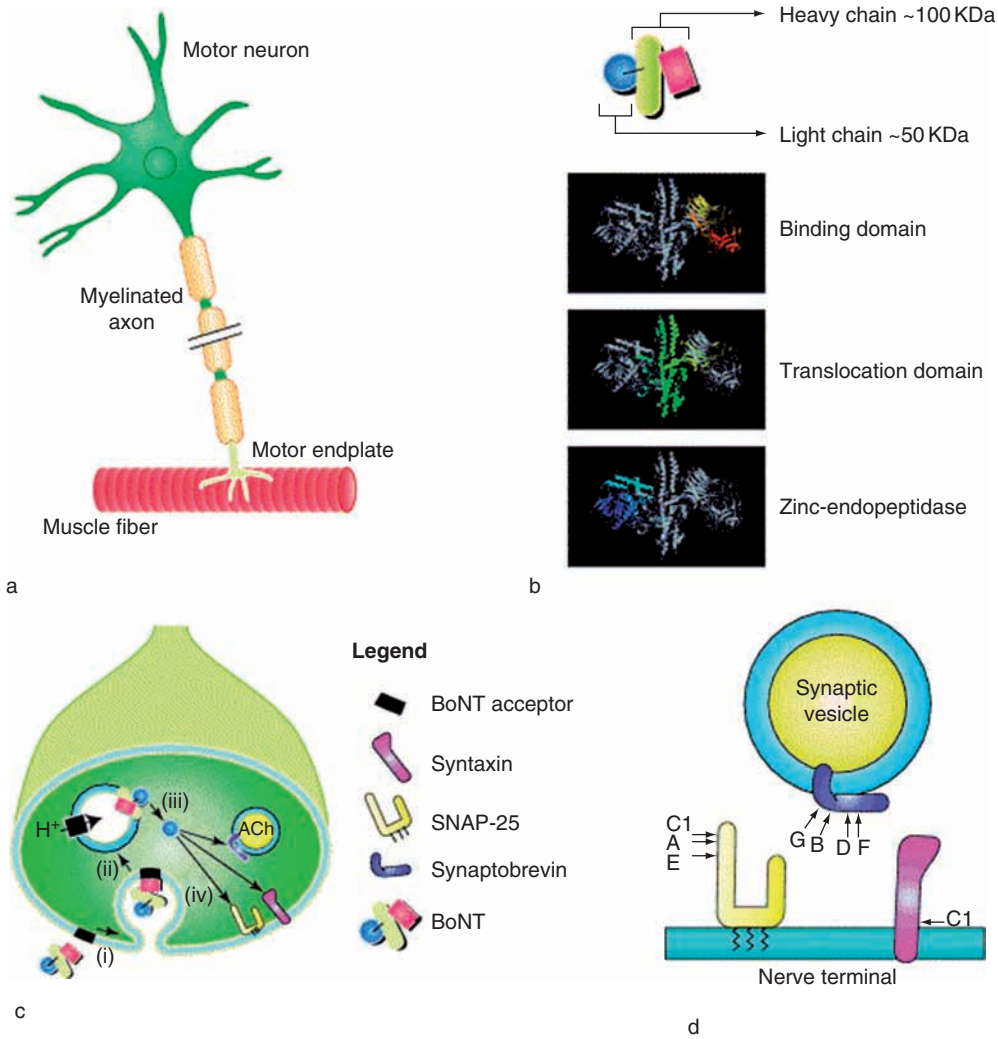

Figure 5 BoNT/A acting on the neuromuscular junction (NMJ): (a) the NMJ target of botulinum neurotoxins (BoNTs); (b) the crystal structure of BoNT serotype A (BoNT/A); (c) the mode of action of BoNTs; (d) cleavage sites of the various BoNT serotypes (A–G) on vesicle-associated membrane protein (VAMP)/synaptobrevin (blue). As shown in (a), the cell body of the motor neuron is located in the spinal cord, and the myelinated axon projects to the muscle fiber, making contact at the motor endplate, the prime target of BoNTs. In (b) and (c), the three distinct domains of BoNT/A are color coded. As shown in (c), the mode of action of BoNTs comprises (i) binding to a proteinaceous BoNT acceptor and gangliosides, (ii) internalization into an endocytic compartment, (iii) translocation of BoNT into the cytosolic compartment following endosomal acidification and the subsequent release of the light chain from the heavy chain by reduction of the disulfide bridge, and (iv) soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) cleavage via zinc-endopeptidase activities possessed by the light chains. As shown in (d), the VAMP)/synaptobrevin protein is located on synaptic vesicles; by contrast, synaptosomal-associated protein of 25 kDa (SNAP-25; yellow) and syntaxin 1 (pink) are situated predominantly on the nerve terminal plasmalemma. ACh, acetylcholine. From Foran PG, Davletov B, and Meunier FA (2003) Getting muscles moving again after botulinum toxin: Novel therapeutic challenges. *Trends in Molecular Medicine* 9: 291–299.

transmission. In contrast, BoNT/B and BoNT/G bind to the extracellular domain of synaptotagmin, a component of the nerve membrane. The ganglioside components of the binding sites have not been fully characterized.

Uptake, translocation, and activation Once bound to the nerve terminal surface, the BoNT is translocated into the cytoplasm by endocytosis. The lumen of the endocytotic vesicles is acidic, and this causes a conformational change in the H chain that enables the toxin to pass through the vesicle membrane into the cytoplasm. There the L chain, and the enzyme activity associated with it, separates from the H chain.

Enzyme activity The L chains of the seven BoNT serotypes are all zinc-dependent metalloproteases, but they differ in their substrate specificities (Table 1). BoNT/A and BoNT/E both attack synaptosomal-associated protein of 25 kDa (SNAP-25) but at different sites; BoNT/B, /D, /F, and /G all attack synaptotagmin but again different cleavage sites are involved. Finally, BoNT/C attacks both syntaxin and SNAP-25. In each case, cleavage of the target protein causes the cessation of Ca^{2+}-dependent vesicular ACh release and hence the failure of neuromuscular transmission. The particular pattern of cleavage associated with each serotype has an important influence on the duration of its toxic effect.

Response of the Neuromuscular Junction to Botulinum Neurotoxin

Block of Transmission

Following an intramuscular injection of BoNT/A, a block of neuromuscular transmission usually occurs within 1–4 days. Both spontaneous miniature endplate potentials (mEPPs) and evoked EPPs are affected. In a clinical context, the decline in the electrical signs of muscle activation may continue for up to 2 weeks. By contrast, when isolated nerve–muscle preparations are exposed to BoNT/A *in vitro*, both spontaneous and evoked quantal release may cease within 1–2 h. It thus seems that the main determinant of the rate of onset of the blocking effect *in vivo* is the time required for an active concentration of BoNT to build up in the vicinity of the NMJs. Clearly in the cases of wound and infant botulism, in which the bacteria must first proliferate in the gut and then produce and release the toxin, the delay between contact with the source of the pathogen and the onset of symptoms is likely to be longer than for food botulism.

Recovery of Transmission after Block by BoNT/A

The recovery of transmission takes place on a time scale of days to months, depending on the BoNT serotype involved and the dose (Table 1). The longest acting serotype is BoNT/A, and for this reason it is used most often in the clinical context. It is also the serotype whose action has been studied in greatest detail.

Axon sprouting Blockade of neuromuscular transmission by BoNT/A in rats and mice is followed within a week or less by the initiation of sprouting by the terminal motor axons (Figure 6). The fine unmyelinated sprouts extend away from the parent NMJ, and in mice grow at a rate that increases from approximately 3 μm day^{-1} at the outset to 15 μm day^{-1} after several weeks. Individual sprouts may ultimately attain lengths of several hundred micrometers. The axonal sprouts are accompanied by Schwann cell sprouts, as following partial denervation.

A poorly understood aspect of the sprouting induced by BoNT/A is the mechanism of insertion of the new membrane that forms the sprouts. Evidence from some studies suggests that the normal activity of SNAP-25, the target of BoNT/A action, is required for axon growth, presumably as part of the process by which new membrane is added to the growing sprouts. During the early stages of motor axon sprouting induced by BoNT/A, the exocytosis associated with the quantal release of ACh is fully blocked, whereas that associated with sprouting is active. One possible explanation is that BoNT/A is inactivated soon after the initial SNAP-25 cleavage. Another is that the SNAP-25 molecules targeted by BoNT/A are not required for sprout formation.

Formation of new synapses The muscle paralysis induced by BoNT/A leads to the generalized expression of numerous postsynaptic molecules within the muscle fiber. Among these are the AChRs. Within approximately a week after sprout formation begins, clusters of AChRs start to form along the growing axons. This process has been followed using fluorescent conjugates of the AChR-binding protein α-bungarotoxin, together with antibody labeling of the axons and their sprouts. Within days of the appearance of the first AChR clusters, lateral branches of the axons form. Individual branches often end in a swollen bouton 1–3 μm in diameter. Most, if not all, new AChR clusters are associated with these boutons. The new synaptic spots also acquire a high density of AChE. Thus, two of the most distinctive components of the NMJ, and probably many more, come to be specifically associated with the newly formed axonal boutons.

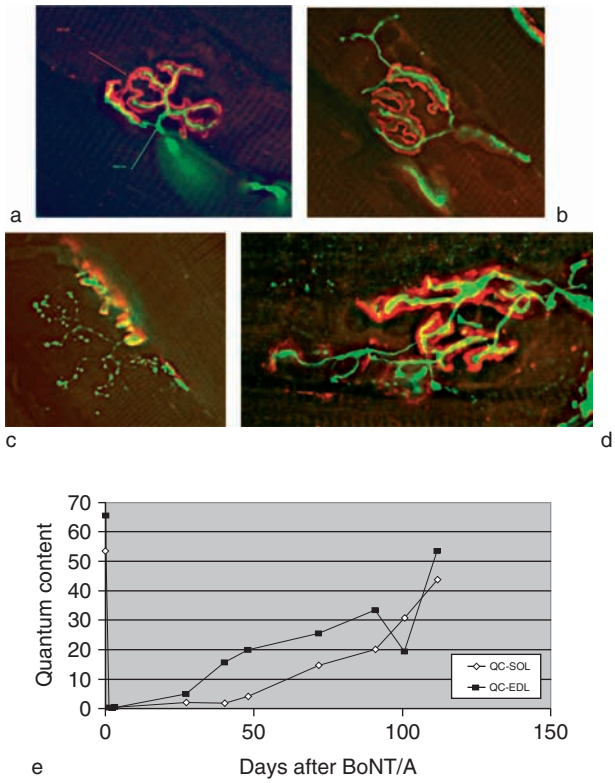

Figure 6 BoNT/A inducing axon sprouting and AChR clustering: (a) normal mouse NMJ; (b) 4 days after a blocking dose of BoNT/A; (c) 8 days after dose; (d) 16 days after dose; (e) graph of quantal content of ACh over time. As shown in (a), at a normal mouse NMJ, the branches of the terminal axon (green) occupy regions of high AChR density (red).As seen in (b), 4 days after a blocking dose fine axonal sprouts can be seen; these sprouts grow and branch over the surface of the muscle. As seen in (c), they often end in swollen boutons. In (d), after approximately 2 weeks, AChRs begin to accumulate in the vicinity of the terminal boutons. ACh, acetylcholine; AChR, acetylcholine receptor; BoNT/A, botulinum neurotoxin serotype A; NMJ, neuromuscular junction; QC-EDL, quantal content of the extensor digitorum longus muscle; QC-SOL, quantal content of the soleus muscle.

Recovery of Quantal Release

The first signs of renewed quantal release of ACh are detected approximately 2 weeks after exposure of the rat/mouse NMJs to a fully blocking dose of BoNT/A. There is then a period of several months during which both spontaneous and evoked release gradually return to near normal values.

It has generally been assumed that the initial recovery of quantal release after BoNT/A intoxication occurs at new synaptic contacts formed by the sprouts. Support for that view comes from studies in which a membrane-soluble fluorescent dye, FM1–43, was used to visualize sites of membrane cycling associated with nerve activity. At normal NMJs, FM1–43 is taken up into the nerve terminal during activity. During the period of complete block following exposure to BoNT/A, this activity-dependent labeling is abolished. The first sites where labeling is again seen during recovery are associated with the sprouts. Eventually, FM1–43 labeling returns in the region of the original NMJ and that associated with the sprouts is lost (Figure 7(b)).

Although these observations are consistent with the view that newly formed synaptic spots account for the initial function recovery after BoNT/A, release was not measured directly. Other studies, made on different muscles of the mouse, suggest that recovery of release at the original NMJ may occur as early as that associated with the sprouts. In both cases, however, it is clear that the new contacts contribute to the initial recovery.

Withdrawal of sprouts In the original description of axonal sprouting following BoNT/A intoxication, it was noted that many of the sprouts were eventually withdrawn (Figure 7(a)). This has been subsequently confirmed by other studies (Figure 7(b)). The withdrawal occurs after neuromuscular transmission has recovered sufficiently to generate appreciable muscle activity. This suggests that, as during nerve regeneration, the sprouts and their newly formed synaptic contacts are less stable than the contacts formed at the original synaptic sites. The basis for this differential stability is not yet known.

It has generally been observed that, even after sprout withdrawal, the appearance of the surviving NMJs differs substantially from normal. Because the reformed NMJs appear to function well, the situation is similar to that following muscle damage in that essentially normal function may be restored even though the structure of the NMJs remains abnormal.

Serotypes other than BoNT/A The rate of recovery from BoNT intoxication depends greatly on the serotype. Although it takes several months to recover from BoNT/A, recovery from BoNT/E is largely complete in less than 1 week. Recovery from other serotypes occurs at intermediate times (Table 1). Here note that the time course of recovery depends on the dose and that in Table 1 typical, but not always directly comparable, values are given. The other factors that determine the duration of functional block associated with the different serotypes are only partly understood. Of particular interest is the comparison between the longest-acting serotype, BoNT/A, and the shortest, BoNT/E, because both cleave SNAP-25. The difference between them is in the site of cleavage. Whereas BoNT/E removes a 26-amino-acid peptide from the C-terminus, the peptide removed by BoNT/A is only 9-amino-acids long. It has been suggested that the removal of 9 amino acids from SNAP-25 is enough to completely inactivate the enzyme activity of the

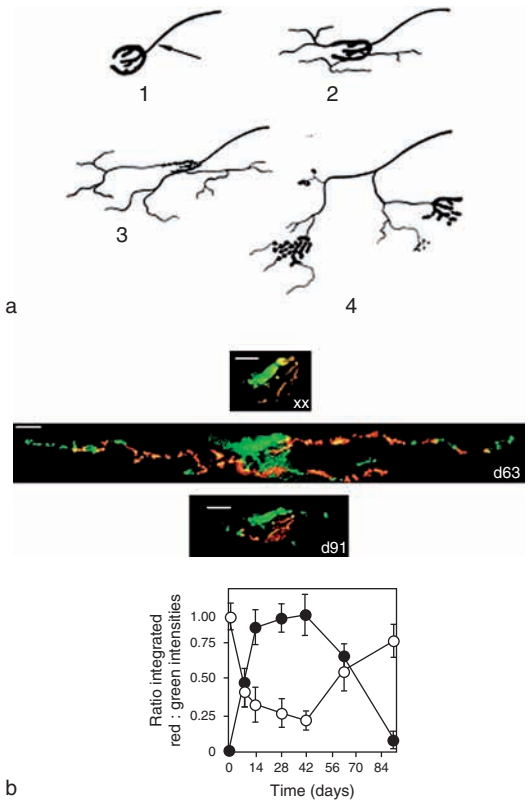

Figure 7 Reorganization of BoNT/A-induced sprouting: (a) drawings of a hypothetical motor axon terminal in a mouse soleus NMJs at different stages in the response to BoNT/A; (b). Experimental results from studies using repeated observation of the same NMJ *in vivo*. In (a), normal NMJ with no fine sprouts present (stage 1). During the early response to BoNT/A, fine sprouts appear (stage 2, 4 days; cf. Figure 6). More extended sprouts as seen during the recovery of functional transmission (stage 3). Enlarged terminals of new synaptic contacts are present along with persisting sprouts when transmission has recovered (stage 4). Note that original NMJ is no longer recognizable. In (b), the nerve terminal is labeled green, representing the mitochondrial dye 4-di-2-asp. The regions of the terminal which take up the membrane dye FM1–43 appear red; these are assumed to be regions of activity-dependent quantal release of ACh. Scale bars = 20 μm. ACh, acetylcholine; BoNT/A, botulinum neurotoxin serotype A; d0, d53, d91, day 0, day 53, day 91; NMJ, neuromuscular junction. Panel (a) Reproduced from Duchen LW (1970) Changes in motor innervation and cholinesterase localization induced by botulinum toxin in skeletal muscle of the mouse: Differences between fast and slow muscles. *Journal of Neurology, Neurosurgery and Psychiatry* 33: 40–54, with permission from BMJ Publishing Group; panel (b) Reproduced from de Paiva A, Meunier FA, Molgó J, Aoki KR, and Dolly JO (1999) Functional repair of motor endplates after botulinum neurotoxin type A poisoning: Biphasic switch of synaptic activity between nerve sprouts and their parent terminals. *Proceedings of the National Academy of Sciences of the United States of America* 96: 3200–3205, with permission from proceeding of the National Academy of Sciences, USA.

toxin while still leaving it in a sufficiently normal state to prevent it from being recognized as defective and hence being rapidly broken down and replaced. Thus, its continued presence as an inactive protein may account for the long duration of the functional blocking effect of BoNT/A.

The rate of recovery has a substantial influence on the extent of axonal sprouting and new synapse formation. Following exposure to BoNT/E, which blocks neuromuscular transmission for less than a week, there is little if any sprouting. Thus functional recovery in this case must occur by recovery of release at the original nerve terminals.

BoNT/F blocks transmission for approximately 1–2 weeks. Sprouting occurs after BoNT/F, but is withdrawn soon after the block is overcome.

Differences Related to Muscle Type and Species

The response to BoNTs also varies both between species and between muscles in the same species. For example, axonal sprouting and functional recovery are much slower in frogs than in mammals. A number of studies on mammalian muscles found that the extent of sprouting is greater in slow-twitch than in fast-twitch muscles. This has sometimes been interpreted to mean that slow motor neurons are more plastic and therefore more successful at recovering from impaired function. Alternatively, it may be that fast motor neurons make new synaptic contacts more rapidly and effectively than slow ones and that the muscle activity resulting from those contacts limits more extensive sprouting in fast muscles.

Effects on muscle When mammalian skeletal muscles are paralyzed for more than a few days, they generally undergo significant atrophy. This loss of volume is part of a complex response to inactivity that has among its effects a facilitation of the restoration of adequate innervation. As the NMJs in a muscle recover from a BoNT-induced block, activity resumes and the muscle fibers regain their normal size. It is a common finding that, during recovery from BoNT intoxication, the variation in muscle fiber diameters is greatly increased. This results from the presence of both abnormally small- and abnormally large-diameter fibers. The former presumably reflects disuse atrophy of the muscle fibers that are functionally denervated, whereas the latter may reflect hypertrophy due to overuse of the fibers. This situation presumably reflects different rates of recovery at different NMJs, possibly related to the functional type of the motor neuron.

Clinical and Cosmetic Uses of BoNT/A

It is a remarkable fact that BoNT/A, one of the most toxic substances known, is currently used clinically to manage patients with abnormal patterns of muscle activation, (e.g., spasticity and dystonias). Even more

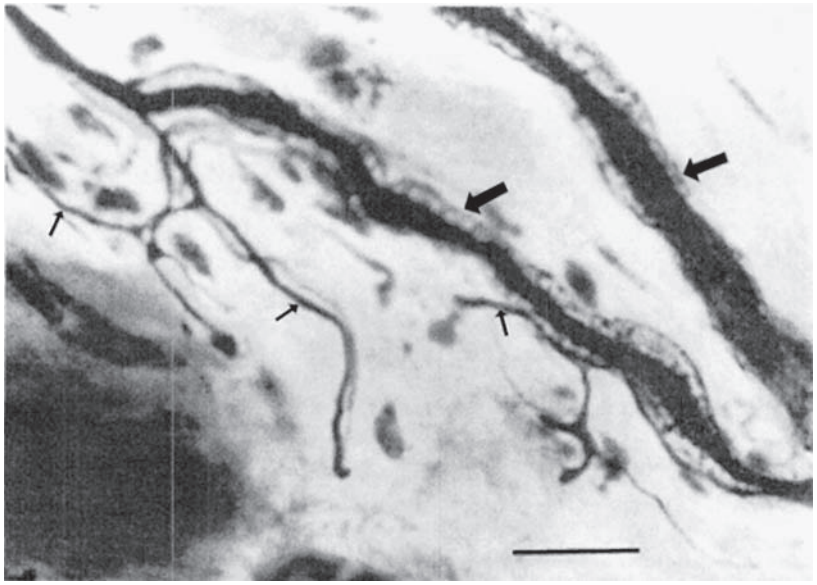

Figure 8 BoNT/A inducing sprouting at human NMJs: axonal sprouts (small arrows), visualized after silver impregnation, can still be seen in the human orbicularis oculi muscle 3 years after treatment with BoNT/A. Reproduced from Holds JB, Alderson K, Fogg SG, et al. (1990) Motor nerve sprouting in human orbicularis muscle after botulinum A injection. *Investigative Ophthalmology & Visual Science* 31: 964–967, with permission from Association for Research in Vision and Ophthalmology.

remarkable is the elective use of BoNT/A as a cosmetic reagent to reduce facial wrinkling by selective weakening of facial muscles. These uses are made possible by the fact that, when injected intramuscularly, BoNT/A binds tightly to the binding sites at the NMJ. Unbound toxin is rapidly broken down. As a result, the diffusion of active toxin away from the site of injection in the extracellular environment may be limited. By keeping doses and injection volumes small, it is therefore possible for experienced and careful operators to use BoNT/A to achieve localized weakening of muscles with minimal risk. At the same time, it is clear that the risks of adverse side effects exist, particularly when repeated injections over a long period of time are used.

The initial use of BoNT/A in a clinical context was to weaken muscles involved in various focal dystonias. These include blepharospasm, strabismus, hemifacial spasm, cervical dystonia, and both juvenile and adult forms of spasticity. More recently, it is also used to treat a variety of autonomic disorders (e.g., hyperhidrosis) and pain syndromes.

Effects of Botulinum Neurotoxins on Human Neuromuscular Junctions and Muscle

Remarkably little is known about the effects of therapeutically administered BoNTs on human NMJs. There are no detailed studies of the short-term effects on NMJ structure and function and only one serious study of the long-term effects on NMJ structure, made on muscles of the eyelid. By far the most common serotype used in clinical practice is BoNT/A because of its long-lasting effects. What follows, therefore, concerns the response to this form.

Time course of action The development of detectable muscle weakness after the administration of BoNT/A is usually reported to take several days. Because the toxin acts within hours on isolated muscles, it is likely that the longer time course seen in practice reflects the low doses used and the spread of the toxin through the extracellular spaces.

Some recovery of muscle strength is usually reported to be detectable 1–3 months after an injection. Here it must be recalled that there is a high safety factor for neuromuscular transmission. As a result, some muscle fibers can be activated when the quantal release of ACh from the nerve is still much lower than normal. Because the aim of the therapeutic use of BoNT/A is to cause muscle weakness, even slight recovery may be considered disadvantageous and viewed as a signal for repeated treatment. It is, therefore, common practice to repeat injections at intervals of several months to maintain the therapeutic effect.

Effects on nerve It is clear from clinical EMG studies that the muscle weakness caused by injections of BoNT/A results from a block of neuromuscular transmission. Almost nothing is known in detail about the process or true extent of recovery of neuromuscular

transmission after BoNT/A injections. Studies of eye muscles from 6 weeks to 3 years after cessation of BoNT/A treatment reveal abundant axonal sprouts (Figure 8). It is therefore likely that the initial recovery, as in animals, is associated with the formation of such sprouts and new synaptic contacts, but there is no direct evidence to support or refute this possibility. What is clear is that, even several years after ending BoNT/A treatment, the nerve terminals may retain strikingly abnormal features.

Effects on muscle The muscle paralysis resulting from BoNT/A treatment generally leads to muscle-fiber atrophy and muscle wasting. Muscles that have been studied at various times after treatment show a striking increase in variability of fiber diameters, similar to that seen in animal muscles. It is not yet known whether normal muscle-fiber diameters are ever fully restored. However, there is some evidence that increased fiber-diameter variation may persist for more than 1 year in human muscles. Whether this reflects a very prolonged period of reinnervation or a stable situation in which some muscle fibers remain permanently functionally denervated is not clear.

Conclusion

The adequate functioning of the NMJ is essential for survival. It is therefore not surprising that very effective mechanisms have evolved to maintain reliable transmission even after severe mechanical trauma or toxic assault. Most notable of these mechanisms are the regeneration of motor axons and effective reinnervation of muscles after damage and the formation of axonal sprouts and new synaptic contacts following partial denervation or BoNT/A intoxication. In these processes, the nerve, the muscle, and the Schwann cell, and interactions among them, all play important parts.

See also: Botulinum and Tetanus Toxins; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction: Neuronal Regulation of Gene Transcription at the Vertebrate; Neuromuscular Junction: Synapse Elimination; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): Aging; Neuromuscular Junction (NMJ): A Target for Natural and Environmental Toxins in Humans; Neuromuscular Connections: Vertebrate Patterns of; Neurotoxins; Neurotoxins and their Neurotoxicology; Presynaptic Events in Neuromuscular Transmission.

Further Reading

Bennett MR, McLachlan EM, and Taylor RS (1973) The formation of synapses in reinnervated mammalian striated muscle. *Journal of Physiology* 233: 481–500.

Brown MC, Holland RL, and Hopkins WG (1981) Motor nerve sprouting. *Annual Review of Neuroscience* 4: 17–42.

Caroni P and Schneider C (1994) Signaling by insulin-like growth factors in paralyzed skeletal muscle: Rapid induction of IGF1 expression in muscle fibers and prevention of interstitial cell proliferation by IGF-BP5 and IGF-BP4. *Journal of Neuroscience* 14: 3378–3388.

de Paiva A, Meunier FA, Molgó J, et al. (1999) Functional repair of motor endplates after botulinum neurotoxin type A poisoning: Biphasic switch of synaptic activity between nerve sprouts and their parent terminals. *Proceedings of the National Academy of Sciences of the United States of America* 96L: 3200–3205.

Duchen LW (1970) Changes in motor innervation and cholinesterase localization induced by botulinum toxin in skeletal muscle of the mouse: Differences between fast and slow muscles. *Journal of Neurology, Neurosurgery & Psychiatry* 33: 40–54.

Foran PG, Davletov B, and Meunier FA (2003) Getting muscles moving again after botulinum toxin: Novel therapeutic challenges. *Trends in Molecular Medicine* 9: 291–299.

Holds JB, Alderson K, Fogg SG, and Anderson RL (1990) Motor nerve sprouting in human orbicularis muscle after botulinum A injection. *Investigative Ophthalmology and Visual Science* 31: 964–967.

Hopkins WG and Brown MC (1985) Growth and remodeling of the neuromuscular synapse. In: Cotman CW (ed.) *Synaptic Plasticity*, pp. 169–199. New York: The Guilford Press.

Humeau Y, Doussau F, Grant N, and Poulin B (2000) How botulinum and tetanus neurotoxins block neurotransmitter release. *Biochemie* 82: 427–446.

Lyons PR and Slater CR (1991) Structure and function of the neuromuscular junction in young adult mdx mice. *Journal of Neurocytology* 20: 969–981.

Maselli RA and Bakshi N (2000) Botulism (American Association of Electrodiagnostic Medicine case report 16). *Muscle & Nerve* 23: 1137–1144.

Oda K (1984) Age changes of motor innervation and acetylcholine receptor distribution on human skeletal muscle fibres. *Journal of the Neurological Sciences* 66: 327–338.

Schiavo G, Matteoli M, and Montecucco C (2000) Neurotoxins affecting neuroexocytosis. *Physiological Reviews* 80: 717–766.

Son YJ, Trachtenberg JT, and Thompson WJ (1996) Schwann cells induce and guide sprouting and reinnervation of neuromuscular junctions. *Trends in Neuroscience* 19: 280–285.

Tonge DA (1974) Chronic effects of botulinum toxin on neuromuscular transmission and sensitivity to acetylcholine in slow and fast skeletal muscle of the mouse. *Journal of Physiology* 241: 127–139.

Wernig A and Herrera AA (1986) Sprouting and remodelling at the nerve-muscle junction. *Progress in Neurobiology* 27: 251–291.

Wilson MH and Deschenes MR (2005) The neuromuscular junction: Anatomical features and adaptations to various forms of increased, or decreased neuromuscular activity. *International Journal of Neuroscience* 115: 803–828.

Neuromuscular Junction: Neuronal Regulation of Gene Transcription at the Vertebrate

H R Brenner, University of Basel, Basel, Switzerland
D L Falls, Emory University School of Medicine, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

The Vertebrate Neuromuscular Junction: An Excellent System for Studying Neurotransmitter Receptor Accumulation in the Postsynaptic Membrane

Synapse formation in the nervous system is an example of highly coordinated subcellular differentiation driven by localized reciprocal interactions between the synapsing cells. These interactions result in the apposition of the molecules that mediate transmitter release in the nerve terminal and the ligand-gated ion channels that are opened by the transmitter in the target cell. To enable synaptic transmission to operate sufficiently rapidly, these two features of the synapse must be located within less than 100 nm of each other. This raises the question, what are the mechanisms that coordinate the appearance and spatial alignment of these crucial aspects of pre- and postsynaptic differentiation?

The vertebrate neuromuscular junction (NMJ) has served as a valuable model system for formulating and testing ideas regarding the cell–cell communication that governs synapse formation. Compared with synapses in the brain, which are small and are made by different neuronal types at different stages of development, the NMJs are large, have qualitatively similar properties, develop at a similar time, and are easily accessible for experimental manipulation. Moreover, new NMJs can be surgically induced to form in adult muscle. Furthermore, the electrical impulse activity which influences the expression of genes involved in synapse formation can be experimentally controlled independently in the pre- and postsynaptic cells forming these *de novo* synapses. Finally, combinations of transgenes can be readily transferred into the myofibers, allowing examination of their function for subsynaptic differentiation.

At NMJs, action potentials in the nerve cause acetylcholine (ACh) to be released from the nerve terminals. This ACh then binds to acetylcholine receptors (AChRs) in the postsynaptic muscle membrane. The muscle AChRs are ligand-gated ion channels which open on binding ACh. The flow of cations through the open channels causes a depolarization that induces an action potential which propagates along the muscle membrane, triggering muscle contraction. So that each action potential in the motor neuron is transmitted to the muscle fiber, AChRs must accumulate in the postsynaptic muscle membrane during development and be maintained there at high concentration throughout life.

Vertebrate skeletal muscle fibers are very large, multinucleated cells. The region of nerve contact at the NMJ occupies about 0.01% of the surface area of a typical muscle fiber. Two complementary processes account for the high density of AChRs in this region. First, in immature muscle, AChRs that are mobile in the plane of the cell membrane aggregate at sites of nerve–muscle contact. This aggregation is mediated by AGRIN, a protein which is released from the nerve terminal. Like all proteins, the clustered AChRs are constantly degraded and need to be continuously replaced. However, the electrical activity associated with muscle activation suppresses transcription of *AChR* subunit genes by muscle fiber nuclei. A second key process therefore exists to ensure that new AChRs continue to be synthesized near the NMJ in active muscles. Signals released by the motor neuron act to maintain a high level of *AChR* subunit gene expression by the small number of muscle nuclei in the immediate vicinity of the synapse. This article discusses the signals and molecular pathways by which the motor nerve maintains *AChR* gene transcription at the synapse despite its suppression in nonsynaptic regions. It should be noted, however, that in addition to *AChR* genes, nuclei at the synapse also express genes coding for other synaptic components at elevated levels (see Table 1).

Recent experiments have dramatically reshaped our understanding of the mechanisms regulating the accumulation of AChRs in the postsynaptic membrane of the vertebrate NMJ. Recent genetic studies *in vivo* have demonstrated convincingly that major conclusions drawn from earlier *in vitro* (muscle culture) studies were incorrect. The data from these *in vitro* studies (some of which are reviewed below) strongly suggested that neuregulin (NRG)/ARIA was the principal molecule stimulating synapse-specific synthesis of AChRs but had little effect on clustering AChRs whereas AGRIN, which clearly induced clustering, had no stimulating effect on synapse-specific synthesis of AChRs. However, the current evidence supports the conclusion that *in vivo*, AGRIN plays an essential role in both clustering of AChRs and their synapse-specific synthesis. In contrast, NRG/ARIA has only a minor role in postsynaptic AChR accumulation (but does

Table 1 Genes for which elevated expression levels by subsynaptic nuclei of the adult neuromuscular junction have been demonstrated

Gene	*Function*
AChRα	AChR subunit
AChRβ	AChR subunit
AChRδ	AChR subunit
AChRε	AChR subunit
Ache	Acetylcholinesterase, rapidly terminating ACh action by hydrolysis
Rapsyn	Membrane associated protein required for synaptic AChR clustering
Utrophin	Cytoskeletal protein involved in synaptic AChR clustering
Musk	Component of Agrin receptor complex, required for all aspects of postsynaptic differentiation
Dok-7	MuSK-interacting cytoplasmic protein, essential for MuSK activation
Na_v1.4	Voltage-activated sodium channel, expressed in adult muscle
Erbb2	NRG receptor
Erbb3	NRG receptor
Laminin–β2	Component of synaptic basal lamina, involved in presynaptic differentiation
Ncam	Neural cell adhesion molecule thought to be involved in neuromuscular adhesion

Data based on elevated transcriptional activity or on synaptic levels of respective mRNAs.

have other essential roles in the development of the neuromuscular system). The reader is asked to keep these points in mind while reading this article.

Expression of AChRs throughout Uninnervated Differentiating Myotubes Is Replaced by Synapse-Specific Expression in Innervated Fibers

The pattern of AChR expression along muscle fibers and cultured myotubes was initially resolved by electrophysiological mapping of ACh sensitivity, that is, changes in membrane potential in response to localized application of short pulses of ACh to the fiber surface. Later, application of labeled α-bungarotoxin (α-BTX), a snake venom protein that binds with very high affinity to AChRs, was used to visualize and quantify AChR distribution. Both methods revealed that AChRs are distributed along the full length of uninnervated and denervated muscle fibers, often grouped in small clusters, but that in mature innervated muscle, almost all surface AChRs are aggregated in the postsynaptic muscle membrane immediately opposite the nerve terminal. These findings indicated that there is a mechanism that tightly traps (clusters) AChRs at the synapse but left open the question as to whether in innervated muscle (1) *AChR* subunit genes are expressed by nuclei throughout the muscle, with the assembled receptors or unassembled subunit proteins or subunit mRNAs subsequently transported to the synapse, or if instead, (2) the *AChR* subunit genes are expressed primarily by subsynaptic nuclei. Molecular biological studies over the period ≈1985–93, applying techniques of increasing sophistication, definitively demonstrated synapse-specific transcription by subsynaptic nuclei.

Muscle AChRs are heteropentameric transmembrane proteins composed of α, β, δ, and either a γ or an ε subunit, each encoded by a separate gene. The AChRs present in immature muscle have the subunit composition $\alpha_2\beta\gamma\delta$ and are sometimes referred to as fetal while those in mature muscles have the composition $\alpha_2\beta\delta\varepsilon$ and are known as adult. The AChR α, β, γ, and δ subunits are expressed by differentiating myoblasts as part of their developmental program. Mammalian myoblasts fuse to form myotubes in two waves. In the first wave, myoblasts fuse to form a relatively small population of 'primary' myotubes that extend to both ends of the developing muscle. In a second wave, a few days after the first, other myoblasts that have undergone further proliferation line up along the primaries, generally in the central innervated region, and then fuse to form a much larger population of 'secondary' myotubes that make up the majority of fibers in adult muscle. During and after myoblast fusion, the expression of fetal AChR subunits along the entire length of the myotube (called global expression) is greatly increased relative to the levels expressed in unfused myoblasts.

It is at the developmental stage when primary myotubes are being formed that nerve processes, later followed by Schwann cells, appear in the developing muscle and begin to form neuromuscular contacts. These are characterized by clusters of fetal AChRs in the muscle membrane. Though the contacts at this stage lack many morphological and physiological characteristics of mature synapses, single quanta of ACh released from the nerve and activating the fetal AChRs are sufficient to generate action potentials in the myotubes, a process that is facilitated by their high input resistance and the long open time of fetal AChR channels. Thus, from the earliest stages of synapse formation, action potentials in the nerve induce muscle action potentials and contraction. Muscle action potentials repress *AChR* subunit gene expression by nuclei along the full length of the muscle fiber, and thus the consequence of innervation is that expression of *AChR* subunit genes by extrasynaptic nuclei becomes almost completely repressed as the muscle matures. An important exception is the

small group of muscle nuclei (termed 'synaptic' nuclei) that underlie the synaptic contact and continue to express high levels of *AChR* α-, β-, and δ-subunit genes in electrically active muscle. In the same nuclei, expression of the γ-subunit gene is replaced by that of the ε-subunit gene. This occurs during the early postnatal period in the rodent. The switch in subunit composition of the AChRs from $\alpha_2\beta\gamma\delta$ to $\alpha_2\beta\varepsilon\delta$ at the synapse results in channels with shorter open times but increased conductance for Na^+, K^+, and Ca^{2+} ions.

Do Motor Neurons Produce Signals That Induce Clustering of AChRs at the Synapse, Stimulate Synapse-Specific AChR Synthesis, or Both?

The central question addressed in this article is, what are the mechanisms that account for the production of AChR subunits at the synapse (synapse-specific expression) in the face of the global repression caused by muscle electrical activity? One possibility is that the synaptic nuclei differentiate independent of (and prior to) innervation to a state in which electrical activity does not suppress their expression of AChR subunits (discussed further in the section titled 'Myocentric model'). However, evidence from two experimental models of NMJ formation, nerve–muscle cocultures and the induction *in vivo* of ectopic synapses by surgical transplantation of foreign motor nerves onto nonsynaptic muscle regions, indicated that motor neurons produce signals that are capable of inducing activity-resistant accumulations of AChRs in the muscle plasma membrane opposed to the nerve terminal. Three hypotheses for the mechanism by which these nerve-derived signals cause AChR accumulation in the postsynaptic membrane were considered: that the signals induce clustering (aggregation) of AChRs, that they induce synapse-specific transcription of *AChR* subunit genes, or both. Using *in vitro* assays of AChR clustering and AChR synthesis rate, investigators set out to identify candidate signaling molecules.

A Nerve-Derived Factor Inducing AChR Aggregation: AGRIN

In a series of elegant experiments, McMahan and colleagues pursued the hypothesis that components of the basal lamina at the NMJ induce aggregation of AChRs. The basal lamina is the layer of extracellular matrix that ensheathes each skeletal muscle fiber and each myelinated axon. At the NMJ, a single layer of 'synaptic' basal lamina is present in the synaptic cleft between the nerve and the muscle. When adult muscles are damaged and allowed to regenerate within the persisting basal lamina sheath, AChRs aggregate at original synaptic sites (marked by the activity of acetylcholinesterase (AchE) that survives muscle fiber breakdown), even if the nerve is cut and its terminals degenerate. Material that could cause AChR aggregation in cultured chick muscle cells was found to be present in the extracellular matrix of the electric organ of the electric fish *Torpedo*, a modified NMJ of great size. This material was purified, found to be based on a protein, and named AGRIN. Antibodies against AGRIN label the synaptic basal lamina at frog and mammalian NMJs and prevent the AChR-aggregating activity of motor neurons. AGRIN is synthesized by motor neurons and released from motor nerve terminals (see also below). In mice in which expression of *Agrin* has been knocked out, nerve-induced AChR clustering fails to occur and nerves wander across the entire muscle, as they do in mice lacking the AGRIN receptor MUSK (Figure 1). There is thus very strong evidence that AGRIN released from the nerve mediates clustering of AChRs at developing NMJs. As pointed out above, AGRIN was not found to affect AChR synthesis rate in cultured myotubes. This is not the case *in vivo*, however, as discussed below.

Candidate Nerve-Derived Factors Inducing *AChR* Gene Transcription

Several molecules have been implicated in inducing the expression of *AChR* subunit genes by synaptic

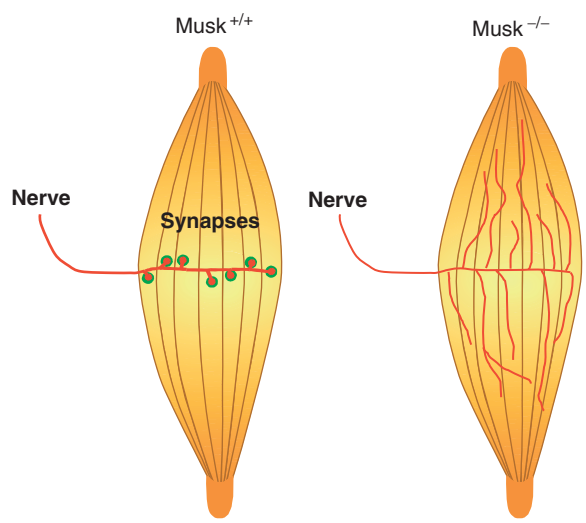

Figure 1 Deletion of AGRIN or of the AGRIN receptor MUSK prevents AChR clustering and neuromuscular synapse formation. Instead, nerve fibers grow along muscle fibers, suggesting that an AGRIN/MUSK-dependent stop signal is required for neuromuscular synapse formation.

nuclei. They include products of four genes, *Nrg1*, *Nrg2, Agrin*, *αCgrp*, as well as ATP. We very briefly discuss αCGRP, and ATP, and then turn to NRG1, NRG2, and AGRIN.

αCGRP and ATP

αCalcitonin gene-related peptide (αCGRP; acting via a G-protein-coupled receptor) and ATP (acting via P2YX receptors enriched in the synaptic membrane) are present in motor nerve terminals and enhance AChR subunit mRNA in cultured myotubes. ATP also affects the rate of AChR desensitization by high concentrations of cholinergic agonist. In contrast to AGRIN and NRG1, however, they have not been shown to be essential for NMJ formation *in vivo*. Specifically, ablation of the *αCgrp* gene in mice does not affect NMJ development, nor does abolishing presynaptic αCGRP and ATP delivery by sectioning the motor nerve abolish synapse-specific *AChR* transcription. However, this does not exclude a redundant role of these factors in this process.

Neuregulins

***In vitro* studies implicating NRG1 as a motor neuron-produced signal regulating postsynaptic *AChR* gene expression** Fischbach and colleagues reasoned that if motor neurons produce a signal that locally stimulates the synthesis of AChRs, an extract of spinal cord (where the cell bodies of motor neurons are located) would increase the rate at which receptors are inserted into the surface membrane of cultured muscle. They found that chick spinal cord and chicken brain extracts did accelerate the rate of new AChR insertion and that this 'acetylcholine receptor-inducing activity' (ARIA) had the biochemical characteristics of a protein. Peptide sequence of ARIA purified from chick brain revealed that this protein is encoded by the gene now designated *Nrg1*.

The NRG1 proteins are ligands for the receptor tyrosine kinases ERBB2, ERBB3, and ERBB4. Several features of NRG1 support the hypothesis that NRG1 is a signal used by motor neurons to induce the synthesis of the muscle postsynaptic specializations. These features include the following: (1) NRG1 is produced by motor neurons and transported along motor axons to the motor nerve terminals; (2) NRG1 receptors are expressed by muscle and become concentrated in the postsynaptic membrane; (3) NRG1 treatment of cultured muscle increases the amount of AChRα, AChRδ, and AChRε mRNAs; and (4) NRG1 activates signal transduction pathways and *cis*-acting DNA sequences implicated in regulating synapse-specific gene transcription (see discussion below of pathways involved in AGRIN-activated gene transcription). It is noteworthy that NRG1 has little effect on AChR clustering. Thus, a widely accepted view in the 1990s was that motor neurons induced postsynaptic differentiation via two signals: NRG1, which stimulated synapse-specific gene transcription, and AGRIN, which caused clustering and anchoring of AChRs and other synapse-specific proteins.

***In vivo* genetic studies have revealed relatively normal NMJ development in mice lacking neuromuscular NRG signaling** A role of NRG/ErbB was difficult to test *in vivo* by genetic approaches because mice homozygous for *Nrg1* or *Erbb* null alleles die due to defects in heart development at an embryonic stage before neuromuscular synapse formation begins. This difficulty has been circumvented through production of mice with tissue-specific mutations of the *Nrg1*, *Erbb2*, and *Erbb4* genes. Synapses begin to develop normally in mutants lacking *Nrg1* selectively in motor and sensory neurons. However, because neural *Nrg1* mutants die at birth (see below), that is, before the developmental time when nerve-induced electrical activity causes marked global suppression of AChR expression, the dispensability of *Nrg1* from neurons for synapse-specific *AChR* gene transcription cannot be tested in the adult in these genetic models. NMJ phenotypes in *Nrg2* mutants have not been analyzed, but mutants are viable, suggesting formation of functional NMJs.

The recent analysis of mice with muscle-specific deletion of both *Erbb2* and *Erbb4* genes has allowed clear assessment of the requirement for NRG-mediated nerve to muscle signal in neuromuscular synapse formation and maintenance. Ablation of both *Erbb2* and *Erbb4* rendered fibers insensitive to NRG, thus abolishing NRG signaling selectively to muscle (ERBB3 homodimers are inactive as they lack a kinase domain) while leaving ERBB2 and ERBB4 expression by Schwann cells and motor neurons unaffected. Synapses developed normally in these mutants (Figure 2). They were viable for months past birth, and synapse-specific *AChRε* and *AChRδ* transcription in the adult was only marginally affected. Therefore, NRG signaling from nerve to muscle is not essential for formation and maintenance of the subsynaptic apparatus. The function of muscle ERBB receptors in the postsynaptic membrane remains unclear.

Essential functions of NRG1 in neuromuscular development While the biological significance of muscle ERBB receptors, and possible *Nrg1*-mediated neuron-muscle, Schwann cell-muscle, and muscle autocrine signaling, remains unclear, there is no uncertainty that NRG1 signaling is essential for neuromuscular system development. First, mice homozygous for a targeted

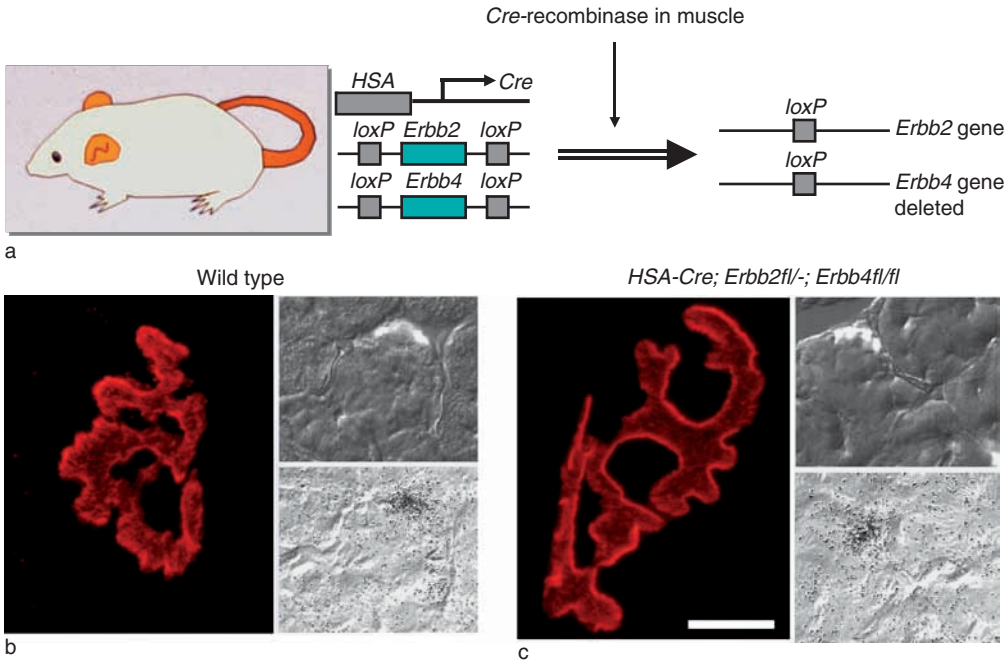

Figure 2 NRG signaling to muscle is not essential for neuromuscular synapse formation. Endplates develop normally in mouse mutants in which NRG/ErbB signaling to muscle is genetically abolished by Cre-induced recombination. (a) Muscle-specific *Erbb* deletion was induced by crossing a mouse line carrying a Cre transgene expressed under the control of a human skeletal actin promoter fragment (HSA-Cre) with a mouse line in which *Erbb* alleles were flanked by loxP sites. Endplates in (b) wild-type and (c) *Erbb*-deficient muscles appear similar (*en face* views (left), bar: 10 μm). Also shown are cross-sections after hybridization with a radiolabeled cRNA probe specific for the detection of *AChR* ε mRNA and development of photo emulsion. Note that accumulations of silver grains reflecting *AChR* ε mRNA co-localize with synaptic AChR clusters marked by fluorescent α-BTX (bar: 20 μm).

mutation that prevents production of Type III NRG1 protein die at birth because of respiratory paralysis and lack neuromuscular synapses because of degeneration of peripheral nerves. Analysis of these and other genetically altered mice, as well as cell culture studies, has demonstrated critical roles of NRG1 in proliferation, survival, migration, and myelination by Schwann cells. Second, targeted ablation of *Nrg1* in motor and sensory neurons (including γ-motor neurons and muscle spindle proprioceptive afferents) or of *Erbb2* in muscle prevents normal development of muscle spindles, causing abnormal gait.

AGRIN

AGRIN molecular variants and the AGRIN receptor MUSK AGRIN is secreted by muscle, nerve, and Schwann cells. As a consequence of differential mRNA splicing, multiple AGRIN proteins are produced by muscle, nerve, and Schwann cells (Figure 3). An N-terminal splice variant, NtA, binds to laminins, thus anchoring AGRIN to the synaptic portion of the muscle fiber's basal lamina. Another variant, TM, is a type II membrane protein with unknown function, expressed throughout the nervous system. NtA–AGRIN in turn is expressed in several C-terminal variants. The isoforms expressed by motor neurons include an 8, 11, or 19 amino acid insert (termed *B*- in chicken and *z*- in mouse) that is required for its function at the NMJ. While the three B/z variants differ somewhat in biological activity, they all act through a receptor tyrosine kinase (RTK) in muscle, MUSK, which is concentrated in the postsynaptic membrane. In contrast, AGRIN proteins lacking the B-/z-insert are expressed by terminal Schwann cells, along the entire length of muscle fibers, as well as in other tissues. The function of this AGRIN is poorly understood. In muscle, it does not activate MUSK, but it is involved in the organization of the muscle fiber's cytoskeleton.

***In vivo* genetic studies of AGRIN'S roles in synapse-specific gene expression** Consistent with the classical function of clustering AChRs, deletion of *Agrin* or of *Musk* genes abolishes synapse formation. It is interesting to note that in fetal muscle of mice lacking MUSK, no nerve-associated elevation of AChR subunit mRNA levels is seen, consistent with the idea that MUSK may be necessary for synapse-specific transcription (see also the section titled 'Models of neuromuscular synapse formation'). In these mice,

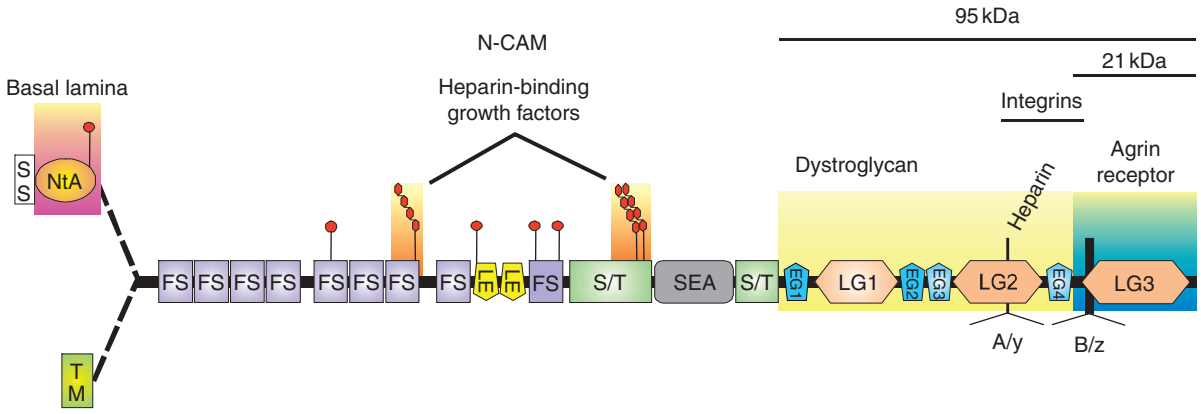

Figure 3 Structural domains of AGRIN and sites of alternative *Agrin* messenger RNA splicing (NtA vs. TM; A/y (0 vs. 4 amino acids); B/z (8 vs. 11 vs. 19 amino acids).) A miniAGRIN generated by fusion of an N-terminal fragment of Agrin (NtA), mediating binding to the basal lamina, to a C-terminal 21 kDa fragment comprising the B/Z-insert and LG3 is sufficient to activate MUSK and can induce ectopic nerve-free postsynaptic membranes, but with less potency than longer C-terminal fragments. SS: signal sequence required for secretion; LG: laminin globular domain. Binding regions to other partners (potentially) involved in NMJ formation are discussed in Bezakova G and Ruegg MA 2003 New insights into the roles of agrin. *Nature Reviews Molecular Cell Biology* 4: 295–308.

the growing motor neurons fail to make terminals but instead grow along the myotubes toward their ends. This suggests that a reciprocal signal from the muscle, in response to the AGRIN/MUSK interaction, is required for the induction of presynaptic terminals (Figure 1).

Positive evidence for a major physiological role of AGRIN in mediating synapse-specific *AChR* gene transcription comes from gain of function experiments in adult muscle. Recombinant neural AGRIN applied to nonsynaptic regions of normally innervated adult muscle fibers *in vivo* can induce the formation of functional ectopic, nerve-free, and Schwann cell-free postsynaptic membranes. These membranes contain accumulations of AChRs and all known molecules involved in AChR aggregation and anchoring and in *AChR* gene induction (muscle-derived NRG1, MUSK, and ERBBs) as well as the type of voltage-gated sodium channels concentrated at adult NMJs, $Na_v1.4$. The AChRs at these ectopic sites have the physiological properties of adult $\alpha_2\beta\varepsilon\delta$channels present at mature NMJs. Finally, the nerve-free ectopic membranes show synaptic folds characteristic of normal NMJs.

A striking feature of this inductive effect of AGRIN is that the ectopic postsynaptic membrane differentiation occurs in regions of the muscle where ongoing activity would normally suppress expression of *AChR* genes. Just as at normal NMJs, ectopically applied AGRIN induces the accumulation of a number of muscle nuclei. These nuclei, but not those elsewhere in the muscle fiber, express the genes *AChRε*, *Musk*, and, as implied from the activity-resistant AChR clusters, *AChRα*-, *β*-, and *δ*-subunits (Figure 4). In addition, these nuclei express higher levels of the gene *$Na_V1.4$* than do the nonsynaptic nuclei. Thus, the ectopic postsynaptic membranes induced by AGRIN are, at least qualitatively, similar to those

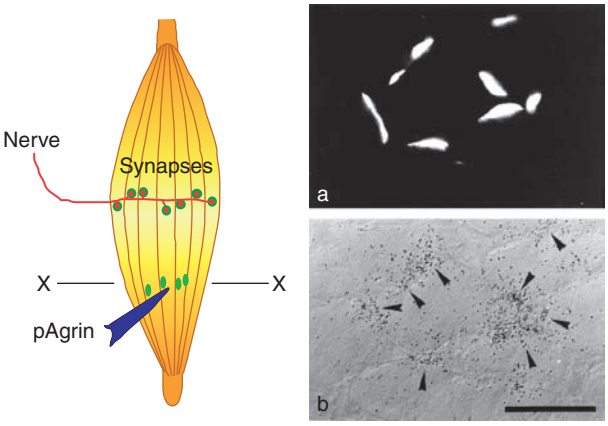

Figure 4 Ectopically applied AGRIN induces nerve-free, ectopic acetylcholine receptor (AChR) clusters and *AChRε*-subunit gene expression. Note that this occurs in spite of electrical impulse activity in the muscle fibers, which is maintained through the original synapses. Left: Ectopic expression of neural AGRIN by muscle fibers is induced by intracellular injection of AGRIN expression plasmid (pAgrin) into single muscle fibers. X- - - X denotes plane of cross-section shown on the right: (a) AGRIN-induced ectopic AChR clusters marked by fluorescent α-BTX; (b) same section after hybridization with a radiolabeled cRNA probe specific for the detection of *AChRε*-mRNA. Note that accumulations of *AChRε*-mRNA (arrowheads) co-localize with ectopic AChR clusters. Scale bar = 70 μm. (a, b) Reproduced from Jones G, Meier T, Lichtsteiner M, Witzemann V, Sakmann B, and Brenner HR (1997) Induction by agrin of ectopic and functional postsynaptic-like membrane in innerrated muscle. *Proceedings of the National Academy of Sciences of the United States of America* 94: 2654–2659, copyright (1997) National Academy of Sciences, USA.

induced by transplanted motor neurons. It is interesting to note that recombinant AGRIN induces nerve-free ectopic AChR clusters and *AChRε* transcription in electrically active muscle, even in *Erbb*-deficient fibers. Therefore, AGRIN alone is sufficient for targeting *AChR* gene transcription to the synapse in electrically active muscle fibers.

Signaling Pathways by Which MuSK Regulates Postsynaptic Gene Expression

***Musk* expression in the postsynaptic membrane: A positive feedback loop?** AGRIN induces synapse-specific *AChR* gene transcription via MUSK. Whereas *Musk* is expressed throughout fetal muscle cells, it is downregulated later in development by electrical muscle activity. But if electrical activity suppresses *Musk* expression, how is MUSK maintained at synapses of active fibers? In adult fibers, forced expression of *Musk* mutants that are constitutively active activates endogenous *Musk*. This implies that MUSK protein is able to autoactivate the expression of the *Musk* gene. Thus synaptic *Musk* expression is stabilized in electrically active fibers by AGRIN/MUSK in a positive feedback signaling loop.

Signaling pathways AGRIN-induced *Musk* transcription is mediated, at least in part, by a regulatory sequence, CCAAGG, termed *N-box*, in the first intron of *Musk*. The same motif is also found in the promoters of *AChRε*- and *δ*-subunit genes, for which it mediates transcription induced by NRG1*β* in cultured myotubes as well as that induced by the nerve at synapses *in vivo*. Thus, the pathways activated by AGRIN, NRG1*β*, and the nerve converge. The AGRIN-dependent pathway has been examined by overexpression of constitutively active and inactive signaling components. AGRIN is known to activate the small GTPases Rac, Cdc42, and Rho. Forced expression of the active mutant RacV12 strongly activates *Musk* and *AChRε*, as does overexpression of active MKK7, an effector of Rac. In contrast, in muscle fibers *in vivo*, induction of *AChRε* and *Musk* by AGRIN is barely affected by overexpression of inactive ERBBs, consistent with the findings on *Erbb*-deficient muscles. However, overexpression of the inactive mutant RacN17, of the MAP kinase JNK, an effector of MKK7, and of inactive GAPB*β*, an effector of JNK, strongly inhibit transcription of *AChRε* and *Musk* both *in vivo* and *in vitro*. In combination, these data are consistent with a pathway at the synapse in which AGRIN/MUSK, activating in turn small GTPases RAC, MKK7, JNK, and GABP and ERM, members of the Ets family of transcription factors, converge on the same N-box in the *AChRε* and *Musk* genes that mediates their activation by the nerve and by NRG/ERBB (Figure 5).

It should be noted that the levels of AChRs and of *Musk* mRNA at ectopic sites induced by AGRIN or constitutively active MUSK are lower than at original synapses. Therefore, full transcriptional activation at the synapse may require synapse-specific signaling components acting downstream of MUSK and essential for its function (e.g., DOK-7), or neural signals in addition to AGRIN, such as NRGs or ATP, which may also target regulatory sequences other than the N-box. For example, NRG1*β* has been found *in vitro* to activate strongly *AChRε* via the transcription factor EGR1, and ATP via Elk-1. Moreover, *in vivo*, forced expression in muscle of a general Ets-dominant negative mutant competing for binding to N-box regulatory elements reduces *AChRδ*- and *ε*-without an obvious effect on *Musk* mRNA levels. Finally, the *AChRα* gene lacks an N-box in its promoter. Alternative pathways are further suggested for another synaptic component, UTROPHIN, which is part of the synaptic cytoskeleton anchoring the AChRs in the synaptic muscle membrane. Specifically, synaptic levels of *Utrophin* mRNA are not regulated by transcriptional activation alone but also by RNA stabilization.

MUSK and synaptic nuclei One factor thought to be important for synapse-specific transcription is the clustering of 5–8 nuclei underneath the synaptic muscle membrane, thus exposing them optimally to signals from the nerve terminal, such as AGRIN. The mechanisms governing this clustering process are not known. Hints come from studies of the control of cell migration, in which crucial roles of small GTPases of the Rho family, in particular of Cdc42, in defining cell polarity with respect to extracellular cues, are well established. Given that (1) recombinant AGRIN and constitutively active MUSK accumulate muscle nuclei in adult fibers, (2) subsynaptic nuclei are embedded in a rich and elaborate microtubular network, (3) efficient and persistent long-range migration requires stabilization of cell polarity, a process achieved through reorganization of the microtubule cytoskeleton, and (4) AGRIN activates Rho GTPases, it is plausible that nuclear translocation to the synapse is mediated through modifications of the cytoskeleton driven by AGRIN-activated Rho GTPases.

A number of proteins in addition to AChRs and MUSK display synapse-specific expression Although the role of neuromuscular NRG/ErbB signaling for synapse formation remains unclear, it is worth mentioning that synapse-specific expression of ERBBs is linked to the AGRIN/MUSK pathway. Both ERBB receptors and their mRNAs as well as *Nrg1β* precursors are observed at the synapse and at ectopic sites induced by AGRIN/MUSK. This indicates that the nerve, at

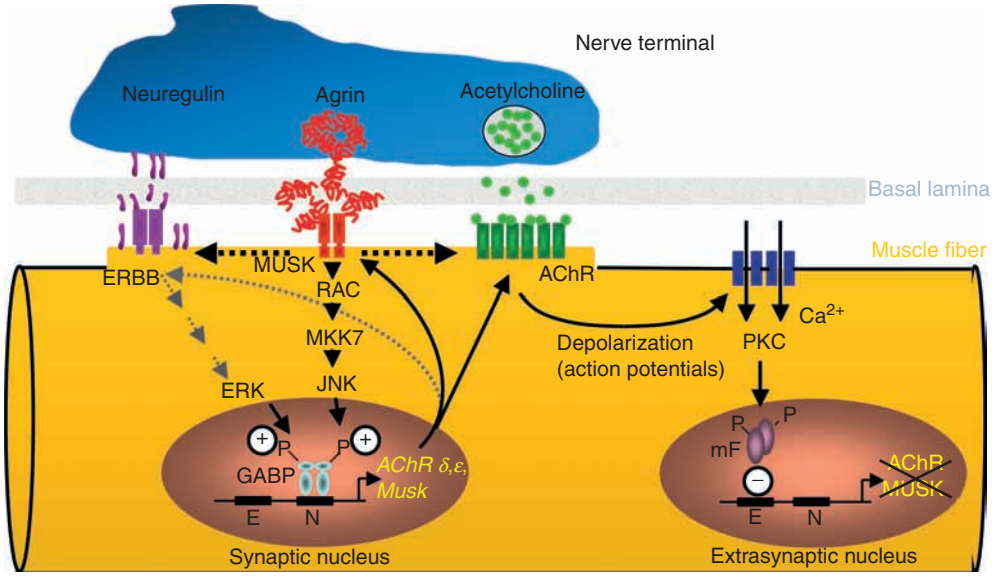

Figure 5 AGRIN from the nerve induces synaptic expression of acetylcholine receptors (AChRs) by targeting the expression of the *Musk* gene to the synaptic muscle region and aggregating AChRs in the synaptic muscle membrane. By activating MUSK in the muscle fiber, AGRIN induces (1) accumulation of muscle nuclei under the synapse via unknown mechanisms and (2) synapse-specific expression of (a) *Musk*, via a positive feedback loop (black solid arrows), as well as of other synaptic genes, including *AChRε*- and δ-subunits (indicated in yellow), all sharing the same regulatory sequence (N-box) in their promoters, and (b) *Erbb*s (broken gray arrows; pathways unknown). The activation of ErbB receptors by neuregulins (NRGs) from the nerve and muscle fiber can also stimulate transcription of genes with N-boxes and thus may strengthen the MUSK-induced expression level. Note that they may be enhanced further by other pathways (not shown). AGRIN also induces synapse-specific expression of (c) *AChR*α- and β-subunit genes (not shown, pathways unknown) and (3) subsynaptic accumulation of AChRs, MUSK, and ERBBs by anchoring them to the synaptic cortical cytoskeleton of the muscle fiber (broken black arrows). PKC is protein kinase C. ERK, RAC, MKK7, JNK are components of the respective signaling pathways; P denotes phosphorylation, mf is myogenic factor, E is promoter sequence binding mf.

least in part by secreting AGRIN, stimulates synaptic nuclei to express *Erbb* genes and to aggregate their products in the synaptic muscle membrane (see also Table 1).

Models of Neuromuscular Synapse Formation

Neurocentric Model

The simplest model to explain the induction and maintenance of synapse-specific gene transcription by the motor neuron in fetal muscle is that *Musk*, expressed constitutively in developing myotubes, is activated locally by AGRIN released from a motor neuron growth cone. This MUSK activation, in turn, stabilizes the expression of *Musk* in myonuclei closely adjacent to the site of contact (subsynaptic nuclei) against the downregulation of *Musk* expression throughout the muscle fiber caused by the electrical muscle activity which begins soon after the neuromuscular contact is established. Via a similar pathway, AChR expression is stabilized by AGRIN/MUSK, ensuring high synaptic ACh sensitivity in electrically active fibers. While NRG/ERBB signaling is not essential for postsynaptic membrane formation and maintenance, induction of genes by AGRIN/MUSK may be augmented by NRG from the nerve and/or the muscle converging on common regulatory elements, as well as by other hitherto unknown alternative pathways, perhaps mediated via NRG2, αCGRP, and/or ATP from motor neurons, muscle fibers, and/or Schwann cells, respectively. As the synapse further differentiates, the number and density of subsynaptic nuclei increase. The accumulation of extrasynaptic nuclei expressing *AChR* genes in response to ectopic AGRIN indicates that any nucleus can be reprogrammed by AGRIN to become a subsynaptic nucleus. Implicit in this model and the supporting data is the idea that during development, motor neuron terminals have the capability of inducing synapse formation anywhere along the length of the muscle fiber.

Myocentric Model

This alternative to the neurocentric model postulates that in fetal muscle, AChRs first cluster independently of neural influence and only later are stabilized by AGRIN from motor neurons contacting them. Conversely, AChR clusters not stabilized by AGRIN would be dispersed. The model is based on the observation that in fetal diaphragms of mouse

mutants lacking phrenic nerves, a central band of AChR clusters and of elevated level of *AChR* α-mRNA is seen in the area where synapses are normally made (although co-localization of nerve-free AChR clusters with focally elevated *AChR* mRNA has not been established at the subcellular level). As noted above, neither the central AChR clusters nor elevated AChR mRNA is seen in *Musk* mutant muscles, indicating an essential role of MUSK in this process. The myocentric model is supported by the observation that in certain muscles of the zebra fish, AChR clusters form prior to innervation and are then preferentially contacted by nerve processes. However, such 'prepatterning' is not generally required for fetal zebra fish neuromuscular synapse formation: AChR clusters follow, rather than precede, the appearance of motor neurons in other zebra fish muscles. A plausible explanation for the 'synaptic zone' in nerve-free mouse muscle is that (1) at high concentrations, MuSK auto-activates, and (2) high MuSK spatially coincides with developing myotubes elongating from the central muscle region and expressing elevated levels of MUSK and, as a consequence of auto-activated MuSK, high AChR levels.

A Unifying Model?

It is possible that the two models represent extreme states of an equilibrium between active and inactive MuSK. MuSK is activated not only by AGRIN but also by itself, depending on its abundance and that of other proteins (e.g., DOK-7) with which it interacts. The number of nerve-free AChR clusters in any particular situation is therefore likely to depend on the levels of expression of MUSK and these other proteins. For example, in myotubes expressing high MUSK levels, the number of nerve-free AChR clusters, and therefore the likelihood of a nerve process to encounter and stabilize a preformed AChR cluster, would be high. Conversely, in myotubes expressing low levels of MUSK, delivery of AGRIN by motor neurons would be required for initial MUSK activation and subsequent AChR cluster formation.

Conclusion

Among the candidates mediating the neural control of neuromuscular synapse formation, AGRIN and NRGs have for more than two decades commanded the widest attention. New genetic studies *in vivo* have dramatically redefined their roles in this process. The key pathways regulating the expression of synaptic AChRs, as well as their aggregation in the synaptic muscle membrane, are activated by AGRIN/MUSK. Synapse-specific transcription of several key synaptic genes appears to be regulated, at least in part, via a common pathway. In contrast, the neuregulins, which had been considered the primary candidates regulating *AChR* gene transcription, have turned out to be dispensable for this process. They are, however, essential for normal development of Schwann cells and muscle spindles.

See also: Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction: Synapse Elimination; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation.

Further Reading

Bezakova G and Ruegg MA (2003) New insights into the roles of agrin. *Nature Reviews Molecular Cell Biology* 4: 295–308.

Brenner HR, Witzemann V, and Sakmann B (1990) Imprinting of acetylcholine receptor messenger RNA accumulation in mammalian neuromuscular synapses. *Nature* 344: 544–547.

Buckingham M, Bajard L, Chang T, et al. (2003) The formation of skeletal muscle: From somite to limb. *Journal of Anatomy* 202: 59–68.

DeChiara TM, Bowen DC, Valenzuela DM, et al. (1996) The receptor tyrosine kinase MuSK is required for neuromuscular junction formation *in vivo*. *Cell* 85: 501–512.

Escher P, Lacazette E, Courtet M, et al. (2005) Synapses form in skeletal muscles lacking neuregulin receptors. *Science* 308: 1920–1923.

Falls DL (2003) Neuregulins and the neuromuscular system: 10 years of answers and questions. *Journal of Neurocytology* 32: 619–647.

Gautam M, Noakes PG, Moscoso L, et al. (1996) Defective neuromuscular synaptogenesis in agrin-deficient mutant mice. *Cell* 85: 525–535.

Jones G, Meier T, Lichtsteiner M, Witzemann V, Sakmann B, and Brenner HR (1997) Induction by agrin of ectopic and functional postsynaptic-like membrane in innervated muscle. *Proceedings of the National Academy of Sciences of the United States of America* 94: 2654–2659.

Lacazette E, Le Calvez S, Gajendran N, et al. (2003) A novel pathway for MuSK to induce key genes in neuromuscular synapse formation. *Journal of Cell Biology* 161: 727–736.

McMahan UJ (1990) The agrin hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology* 55: 407–418.

Merlie JP and Sanes JR (1985) Concentration of acetylcholine receptor mRNA in synaptic regions of adult muscle fibers. *Nature* 317: 66–68.

Mishina M, Takai T, Imoto K, et al. (1986) Molecular distinction between fetal and adult forms of muscle acetylcholine receptor. *Nature* 321: 406–411.

Ruegg MA, Tsim KW, Horton SE, et al. (1992) The agrin gene codes for a family of basal lamina proteins that differ in function and distribution. *Neuron* 8: 691–699.

Sanes JR and Lichtman JW (1999) Development of the vertebrate neuromuscular junction. *Annual Review of Neuroscience* 22: 389–442.

Schaeffer L, de Kerchove d'Exaerde A, and Changeux JP (2001) Targeting transcription to the neuromuscular synapse. *Neuron* 31: 15–22.

Tsim KW, Choi RC, Siow NL, et al. (2003) ATP induces postsynaptic gene expressions in vertebrate skeletal neuromuscular junctions. *Journal of Neurocytology* 32: 603–617.

Neuromuscular Junction: Synapse Elimination

R R Ribchester, University of Edinburgh, Edinburgh, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Motor neurons supply the last outposts of the motor system, where volition finally interfaces with and influences the material world. In higher vertebrates, the axon of an adult somatic motor neuron projects to a single anatomically defined muscle target. The motor axon then divides into several tens to hundreds of intramuscular collateral branches, each terminating on a muscle fiber at a neuromuscular junction (NMJ) to form its sole innervation. A high margin of safety for neuromuscular transmission ensures that every action potential generated in the motor neuron reliably triggers an action potential and contraction of all the muscle fibers it supplies. Thus, a motor neuron and its muscle fibers function together in a unitary fashion: The combined set is referred to as a motor unit. The number of muscle fibers in one unit is termed the motor unit size. Motor units are activated in an orderly fashion during voluntary movements, often sequentially by size. The variation in motor unit sizes (more than tenfold in some muscles and more than 1000-fold between other muscles) confers a remarkable dynamic range on the motor system as a whole.

All historic and contemporary images of adult NMJs reveal a remarkable and stereotyped pattern of the motor nerve supply to skeletal muscle fibers, found throughout the vertebrate subphylum but particularly well represented in mammals: Each well-defined motor junctional area, also called the 'endplate,' is supplied by the axon of one and only one motor neuron. However, this pattern is not formed *de novo*. Rather, the motor endplates of individual muscle fibers are initially supplied by terminals derived from several motor neurons. This state is called polyneuronal innervation. The transition from polyneuronal to mononeuronal innervation during development occurs by a process commonly known as synapse elimination, the subject of this article. The reduction in motor unit convergence (overlap) has another important consequence: Since in most cases the number of muscle fibers does not increase during synapse elimination, the degree of motor divergence also decreases; that is, synapse elimination also brings about a reduction in motor unit size, therefore ultimately limiting the maximum force a motor unit can produce.

Three main physiological factors regulate motor unit force production in vertebrate muscle. In addition to motor unit size, these are the frequency of activation and the functional 'type' of the muscle fiber. In rodents, where these factors have been studied in greatest detail, they all change during development. It is interesting to note that the three properties appear to be linked: Frequency of activation and intrinsic properties such as motor neuron and muscle fiber type play roles in steering the process of synapse elimination to its conclusion, thus determining the motor unit size.

Different rules govern the maturation of innervation patterns in smooth, cardiac, and some specialized types of voluntary striated muscle fibers that contract relatively slowly. Persistent, stable polyneuronal innervation is the norm in these muscle types, and they are therefore not discussed here. Synaptic remodeling also occurs during development of invertebrate muscles, and this plasticity is increasingly studied in the fruit fly, *Drosophila melanogaster*, mainly because of the power and pace of molecular genetic analysis in this species. However, although pruning of motor axon branches occurs during metamorphosis, neuromuscular synapse elimination comparable to that in mammalian muscles does not normally take place in *Drosophila*. The present discussion is therefore mainly restricted to mammalian skeletal muscle.

From π to μ

Neuromuscular synapses form about halfway through gestation in mice and rats. In some muscles, axonal growth cones appear to contact random points on the membrane of the immature muscle fibers (myotubes) and to induce specialized synaptic features in them. In other muscles, it appears that growth cones contact preformed postsynaptic sites. Either way, axons from other motor neurons then add to the first neuromuscular inputs. Thus, by birth, virtually all muscle fibers are innervated by terminals of several motor neurons, that is, polyneuronal innervation (π). The first evidence of polyneuronal innervation in mammalian muscle was obtained in 1916–17 in a study of fetal human tissue by Tello. Contemporaneously, Boeke reported polyneuronal innervation in reinnervated intercostal muscles after nerve injury in hedgehogs. Skeletal muscles of mice and rats are presently the tissues of choice for descriptive and mechanistic studies of neuromuscular synapse elimination both in neonates and in reinnervated muscles in adults.

More-recent interest in synapse elimination began with two electrophysiological exposés, one on neonatal

rat diaphragm muscle and the other on reinnervated adult rat hind limb muscles. Both studies showed that when the muscle nerves were stimulated with electric pulses of gradually increasing strength, discontinuous, stepwise increments occurred in the amplitude of the endplate potentials recorded with intracellular electrodes from individual muscle fibers. This is in contrast to the one-step response normally observed during graded nerve stimulation in adults. The simplest explanation of these observations is that multiple axons converging on the same endplate, and having distinct electrical excitation thresholds, are progressively recruited as the stimulating current is increased.

Polyneuronal innervation was also demonstrated using electrophysiological techniques at about the same time in other immature vertebrates, including tadpoles and chick embryos. These studies were complemented by physiological evidence of polyneuronal innervation in skeletal muscles of neonatal kittens by recording isometric tension. These studies showed that the force produced in response to the simultaneous stimulation of two immature motor axons is usually less than the arithmetic sum of the responses to stimulating both separately. This is also most simply explained by convergent innervation of muscle fibers.

More recently, direct histological evidence of polyneuronal innervation and synapse elimination has been obtained by confocal microscopy analysis of three- and four-dimensional image stacks (X,Y,Z +/− T), for example in preparations immunostained for neurofilament protein in axons (Figure 1) or expressing different fluorescent proteins. Such images have provided unequivocal confirmation of the convergence of several axon terminals on individual motor endplates at birth in rodents and synapse elimination during the first 3 postnatal weeks.

One early study has since achieved iconic status in the field: a comprehensive analysis of polyneuronal innervation and synapse elimination carried out by Jansen, Brown, and Van Essen. These investigators brought to bear a powerful combination of tension measurements, intracellular recordings, and histology. All contemporary studies and analysis of polyneuronal innervation and neuromuscular synapse elimination can be traced back to this study. Its main findings and conclusions were the following:

- Motor unit size is maximal at birth and declines roughly exponentially during the first 3 postnatal weeks, during which time there is no reduction in the total number of motor neurons. Therefore, synapse elimination is due to withdrawal of motor nerve terminals and pruning of axon collateral branches rather than to complete degeneration of individual motor neurons.

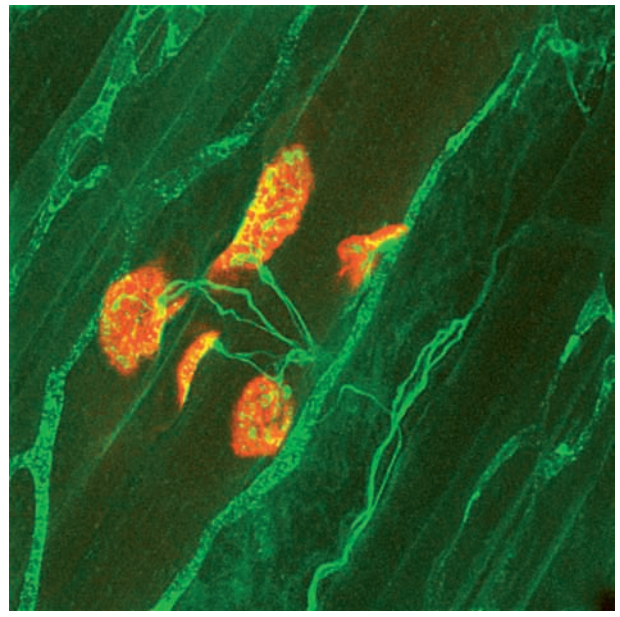

Figure 1 Polyneuronally innervated neuromuscular junctions in an 8-day-old mouse skeletal muscle, stained for neurofilament proteins in axons (green) and acetylcholine receptors on muscle fibers at motor endplates (red). Two axons can be seen projecting in the same motor endplate on several of these muscle fibers.

- The rate of synapse elimination is highest during the first postnatal week. However, virtually all muscle fibers remain innervated by at least two inputs during this period. Mononeuronally innervated muscle fibers emerge rapidly during the second to third postnatal weeks.
- Partial denervation at birth leads to reduced synapse loss in the surviving motor units but does not prevent synapse elimination completely. Mononeuronal innervation therefore arises partly by competition between the terminals of different motor neurons converging on the same endplate and partly by an intrinsic tendency of motor neurons to withdraw a proportion of their terminal branches.
- Implanting the cut end of a nerve that normally innervates one neonatal muscle into another, at some distance away, can result in stable polyneuronal innervation. The distance between converging synaptic inputs therefore mitigates elimination of polyneuronal connections: The further apart two inputs are, the more likely they are to persist and become stable.

A striking feature of the process of synapse elimination at the developing NMJ revealed by these studies is that the great majority of NMJs end up innervated by one and only one motor neuron μ. Although the response to partial denervation at birth suggests that there is a limit to the number of supernumerary branches that can be maintained, it is clear that the

usual result of synapse elimination is not consistent with a random loss of a fraction of the branches formed initially. This would leave some NMJs innervated by several axons and others by none. Rather, the observations point to some sort of local competition at each NMJ that results in a single 'winner.' (Whether synapse formation by motor axon branches is initially random is another unresolved matter.)

Comparable findings were subsequently reported by several groups using similar combinations of histological and electrophysiological methods in studies of reinnervated muscles in adults. Following certain forms of nerve injury, particularly crushing the nerve so that the surrounding connective tissue sheaths remain intact, the damaged motor axons grow back into the target muscle. Regeneration of damaged axons may restore some or all their original neuromuscular connections, in some cases leading to excellent functional recovery. When axons regenerate into partially denervated muscles, they often polyneuronally innervate endplates already occupied by intact motor nerve terminals. In such cases, a form of competition takes place between the original and the regenerated axons. The many similarities between synapse elimination in neonatal muscles and reinnervated adult muscles led to the working hypothesis that the underlying mechanisms are the same.

Synapse Elimination versus Synapse Degeneration

Ultrastructural analysis of NMJs established that the characteristics of synapses and axon collaterals undergoing elimination and withdrawal are quite unlike those of the orthograde (Wallerian) degeneration that occurs when axons are damaged. Wallerian degeneration is rapid and involves radical breakdown of cytoplasmic integrity and organelles within intramuscular axons and motor nerve terminals. By contrast, synapse elimination results in regression of a subset of the collateral branches in a motor neuron's axonal arbor. It is protracted over several days, and for the most part it involves gradual retraction of synapses and resorption of the affected branches into the parent axon trunk with no overt loss of integrity of the cytoplasm and organelles (e.g., mitochondria and synaptic vesicles appear to remain healthy).

Real-Time Visualization of Synapse Elimination *In Vivo*

An improvement in the basic description of the neonatal remodeling of motor units has followed the development of transgenic mice expressing fluorescent proteins in motor neurons. This enables all the connections made by individual motor neurons to be visualized at all stages of development. These transgenic mice have been made by inserting variants of the fluorescent jellyfish protein Green Fluorescent Protein into the mouse genome under the control of a chemically modified promoter, thy1.2, that has the effect of selectively driving protein expression mainly in neurons. Collectively, the members of this family of fluorescent transgenic mice are referred to below as XFP mice (X = R for red, Y for yellow, G for green, or C for cyan). Recently, via Cre-Lox floxed-stop tamoxifen induction, variants have been generated that express variable amounts of YFP. In all these cases, expression of the fluorescent proteins in neurons appears to be completely harmless: The transgenic mice are indistinguishable from their littermates until their brains, spinal cords, and peripheral nerves are viewed in a fluorescence microscope. Use of XFP mice has already revealed interesting new insights into the process of synapse elimination and is likely to continue to do so for some time.

Is competition local? Visualization of XFP-labeled NMJs during postnatal synapse elimination has confirmed and extended previous interpretations based on the more indirect methods applied for the first time by Jansen, Brown, and Van Essen. Studies of motor nerve terminal size, number, and disposition at different stages in the elimination process show that synapses and collateral branches retract into their parent nerve trunks such that during early stages, a dwindling fraction of each NMJ is occupied by the withdrawing synapses. The local competitions experienced by each of the numerous terminals of an immature motor neuron do not occur simultaneously but by an asynchronous process protracted over several days. At any given time, some of the terminals of a single neuron are 'victorious,' fully occupying the endplate, while others show variable fractional occupancies, giving way to or taking over the space occupied by the terminal synapses of other motor neurons. Takeover of endplates during synapse elimination has now been observed directly, by repeated visualization of polyneuronally innervated junctions whose axonal inputs are labeled by selective expression of YFP and CFP in different motor neurons. These direct observations appear to refute the hypothesis previously favored by some, that the main route to mononeuronal innervation was via elimination of acetylcholine receptors and their overlying synaptic boutons, with no takeover by the remainder. The routine takeover of endplates by one terminal, with simultaneous retraction of others, is now a generally accepted account of the normal process leading from polyneuronal to mononeuronal innervation. This view also accords with the progressive recovery

of synaptic area observed in partially denervated muscles following regeneration of injured axons, as suggested by static images of vital staining of regenerating and intact axons.

Expression of different fluorescent proteins (CFP and YFP) in different motor neurons has also been used to study unusual behavior at some NMJ. A few synapses show oscillating 'flip-flop,' in which first one of two inputs occupies most of the endplate at the expense of the other, but then the second gains the upper hand and the most fractional occupancy. It is unknown what determines this remarkable dynamic interaction or its ultimate outcome in favor of one of the two inputs, with complete elimination of the other.

At first glance, it is tempting to conclude that asynchronous retraction and flip-flop behavior suggest that synapse elimination is driven locally, entirely by competitive interactions between converging terminals. However, experiments show that asynchronous synapse loss also occurs when global properties such as axonal transport are interrupted. For example, axonal injury in adult mutant mice with the 'Wallerian degeneration, slow' mutation (Wld^S), in which orthograde degeneration is delayed, also results in protracted, asynchronous synapse retraction. Thus, observations of asynchronous synapse loss constitute insufficient evidence to conclude that local interactions drive synapse elimination (see below).

The fate of the losers Finally, detailed descriptions have been obtained by time-lapse imaging *in vivo* of the last stages of synaptic retraction. These have confirmed that removal of the losing synapse partly involves retraction of a collateral nerve branch into the parent axon trunk, identified by characteristic end bulbs (retraction bulbs) on the withdrawing axon branch. In addition, combined confocal and ultrastructural (electron microscopic) analysis has also shown that some retracting synapses and axon branches undergo fragmentation, with sequestration of the cellular residues, named axosomes, into Schwann cells and other phagocytic cells in the vicinity of the NMJ. Collectively, the retraction and axosomal fragmentation of synapses undergoing elimination appear remarkably similar to the loss of synapses following axotomy in some forms of neuromuscular pathology. This includes neuromuscular synaptic degeneration in Wld^S mutant mice after nerve injury; mouse models of some forms of motor neuron disease, such as in the SOD1G93A transgenic mouse model of amyotrophic lateral sclerosis (ALS); and in *smn* mutant models of spinal muscular atrophy (SMA; see below). However, it remains to be seen whether these remarkable morphologically similarities are caused by similar underlying molecular mechanisms.

Are There Intrinsic Hierarchies Among Motor Neurons?

A recent study by Kasthuri and Lichtman seems set to reignite debate over the forces that establish which synapses persist when two motor nerve terminals belonging to different motor neurons vie for exclusive occupancy of the same motor endplates. Transgenic mice expressing YFP in a small subset of motor neurons and CFP in a different subset were examined. Sometimes only two motor neurons supplying the same muscle were labeled, one yellow and the other cyan. Thus, the investigators were able to identify all the polyneuronally innervated motor endplates where both axons contributed a motor nerve terminal. They found that while both axons innervated variable fractions of the endplates they did not share, the same member of the pair nearly always occupied a greater fraction than the other at all endplates that they did share. This suggests that while competitive interactions take place locally at motor endplates, a pecking order or dominance hierarchy may operate within motor neuron pools, intrinsically biasing the outcomes of all their interactions with others during synapse elimination. An important prediction of this hypothesis is that all the branches of the motor neuron with the highest ranking should be 'winners.' By contrast, the lowest-ranked motor neurons should lose virtually all competitions and have either all or nearly all their terminals eliminated.

The suggestion of a fixed hierarchy among motor neurons needs to be viewed in the context of several previous studies. First, time-lapse observations of flip-flop (made by the same research group) made over several days *in vivo* suggest that if there is a hierarchy, then either it may be a dynamic one or the differences that determine the hierarchical rank could be very small and subtle. Second, indirect evidence based on isometric tension recordings in the watershed study by Jansen et al. suggests that some motor neurons withdraw their terminals from motor endplates in the absence of competitors, albeit this observation has been challenged by other studies and the issue remains unresolved. Third, axons regenerating into partially denervated muscles fail to displace sprouts from most of the terminals they previously innervated. Fourth, high levels of stable, persistent polyneuronal innervation are found following recovery from muscle paralysis. However, all but the first of these caveats are based on indirect evidence that predated the availability of transgenic XFP mice. It is to be hoped that the general availability of these mice will eventually allow clarification of the possibility of an intrinsic hierarchy among motor neurons.

The Role of Activity in Synapse Elimination

One of the questions examined in the immediate wake of the Jansen, Brown, and Van Essen study was whether activity plays an obligatory and decisive role in synapse elimination. In rats and mice, synapse elimination occurs in hind-limb muscles as the animals are first beginning to use those limbs to bear weight. By analogy in the visual system, use of the eye during a critical period when the precision of adult connectivity is being established plays a vital part in determining the outcome of that process. It therefore seemed plausible that something similar might happen at the NMJ. Several studies have explored the effect of changes in neuromuscular activity on both the timing and the outcome of synapse elimination, either in development or during reinnervation of adult muscle. The various studies have led to the following conclusions:

- Reducing activity delays elimination of polyneuronal innervation.
- Increasing activity makes synapse elimination occur sooner.
- Selectively altering activity levels in motor nerve terminals at NMJs supplied by convergent motor neurons frequently biases the outcome of synapse elimination in favor of the more active axons.
- Polyneuronal innervation is not an inherently unstable state: it persists indefinitely in active muscle under some experimental conditions.
- Activity is not strictly necessary for synapse elimination since it can occur at some NMJs even when they are experimentally paralyzed.
- Activity-dependent competitive vigor correlates with the synaptic strength of convergent inputs, that is, the amount of neurotransmitter release per unit area.

The overall conclusion from these studies is that activity is strongly influential but not strictly decisive in determining the outcome of synapse elimination. Other variables likely to mitigate competitiveness include intrinsic limits on the numbers of peripheral connections any one motor neuron can support, sensitivity to neurotrophic factors, and selective recognition or adhesion of motor neurons to specific muscle fibers based on either topographic or histochemical markers.

Muscle Fiber Type Specificity and Selective Synapse Elimination

Skeletal muscle fibers are metabolically heterogeneous. For example, some are specialized for brief, faster and others for prolonged, slower contractions. In addition, many are segmentally or compartmentally organized with axonal inputs constrained to arborize within the segmental or compartmental boundaries. Synapse elimination in development therefore occurs largely within rather than between segments or compartments. Occasionally, however, it appears that mistakes are made and editing of these occurs by axon branch pruning.

In rats and mice, some muscle fiber subtypes are already specified at birth: specifically, type I fibers expressing myosin isoforms characteristic of slow-twitch muscles. In contrast, fast-twitch type II muscle fibers develop mainly postnatally. In adults, these subtypes of fibers are selectively innervated by motor neurons whose activity patterns match the muscle fiber type characteristics. There is some evidence for nonselective innervation of muscle fiber types before synapse elimination occurs. Thus in mice in the first week after birth, individual motor units contain both fast and slow muscle fibers. By the time synapse elimination is complete, the homogeneity of the units has increased to nearly its adult level. This suggests that specific matching of motor neuron type to functionally appropriate muscle fiber types plays a role in synapse elimination. However, as yet there are no selective molecular markers for different motor neuron types, so this hypothesis awaits a stringent test.

Molecular Mechanisms of Synapse Elimination and Axonal Pruning

Progress toward understanding the molecular mechanisms that induce and execute synapse elimination has been exasperatingly slow. Neuromuscular synapses are virtually inaccessible to systematic biochemical analysis, because they comprise such a small fraction of the volume of skeletal muscles (unlike the brain, where synapses are the main constituent). However, several studies have used the incidence of polyneuronal innervation as a bioassay for the effects of pharmacological blockers, growth factors, and other treatments on synapse elimination. Virtually all these studies show only transient effects and have thus far provided no clear understanding at a molecular level of the causal events leading to mononeuronal innervation.

Neurotrophic Influences on the Rate of Synapse Elimination

Neurotrophic factors play a crucial role in maintaining the size of neuronal populations and the number and disposition of their connections. For instance, survival and growth of autonomic and sensory

neurons depend on maintenance of physiological levels of nerve growth factor. Some cortical neurons show similar requirements for the related neurotrophin, brain-derived neurotrophic factor (BDNF). Members of another class, the neurotrophic cytokines, have potent effects on motor neuron survival. The effects of different families of these cell-survival molecules have been tested on neonatal muscles to see whether they may play similar regulatory roles in synapse elimination. The results have been largely equivocal and/or negative. For example, administration of either BDNF or ciliary neuronotrophic factor delays synapse elimination in mice, but only by about a day. Remarkably, however, transgenic expression of glial cell line-derived neurotrophic factor (GDNF) in mouse muscle delays synapse elimination by about 1–2 weeks. Unfortunately, GDNF is unlikely to be a physiological regulator because many muscle fibers in normal mice do not express it at the right stage of development.

Role of Proteases and the Ubiquitin-Proteasome System

Some studies have focused on the possibility that regulated proteolysis of extracellular components of the NMJ might be involved in the elimination of supernumerary nerve terminals. There is some evidence that inhibition of either Ca-activated or serine proteases delays synapse elimination. Likewise, it has been reported that activation of protein kinase C delays synapse loss. At present, however, there is no detailed model of how these effects might operate during normal NMJ development.

Recently, there has been interest in the possible role of the ubiquitin-proteasome system (UPS), the mechanism within cells for targeted breakdown of proteins. Proteins tagged with a chain of ubiquitin monomers are recognized by chaperone proteins that convey them to proteasomes, organelles that execute protein degradation. Interest in this system stems partly from the role of the UPS in several forms of neurodegenerative disease, including ALS. There is evidence that the UPS plays a significant role in both Wallerian degeneration and axon pruning. For instance, synaptic degeneration is delayed in Wld^S mutant mice, in which there is expression of a chimeric protein, part of which contains the N-terminal 70 amino acids of a ubiquitination cofactor, Ube4b. Axotomy induces asynchronous synaptic retraction and fragmentation in these mice. Inhibitors of the UPS also delay axon pruning in *Drosophila* and axon degeneration in the SOD1G93A mouse model of ALS. However, there are as yet no published data on the role of the UPS in neonatal synapse elimination.

Role of Other Nonneural Cell Types at Immature and Adult NMJs

Parts of three cell types are thought to constitute the normal NMJ: muscle fiber, motor nerve terminal, and one or more perisynaptic (terminal) Schwann cells. These structures are bonded to one another by synaptic basal lamina. Nerve injury in adults triggers, after a latent period of about 3 days, proliferation and branching of terminal Schwann cells, which then form a scaffold of bridges between motor endplates. These appear to facilitate reformation of synapses by regenerating motor axons. The average number of Schwann cells per NMJ increases postnatally, but there is no compelling evidence that this process plays a role in selective elimination of inputs during synapse elimination. However, Schwann cells engulf degenerating nerve terminals after axotomy, during Wallerian degeneration, and appear to fulfill a similar function in synapse elimination, with formation of axosomal fragments. Recent availability of transgenic mice expressing fluorescent proteins in these terminal Schwann cells will permit longitudinal studies that should help establish whether they have a role in synapse elimination. NMJs also appear to be a favorable local environment for a population of so-called perisynaptic fibroblasts. These cells have been observed at most NMJs in frogs, chickens, rodents, and human muscle. They lie outside the basal lamina, but although they react by spreading and expression of specific cell adhesion molecules, their precise functions at NMJs remain intriguing but unknown.

Synapse Elimination and Neurodegenerative Disease

Most neurodegenerative diseases show an increased likelihood of onset with age. This applies to most adult forms of motor neuron disease (MND), such as ALS. Most adult forms of MND (more than 90%) are sporadic, the remainder inherited. About 20% of the inherited forms (i.e., less than 2% overall) are due to mutations in genes coding for superoxide dismutase (SOD). However, the availability of good animal models of SOD-dependent ALS has led to this form's being the most widely studied. Overexpression of SOD genes in transgenic mice leads to defective axonal transport and motor neuron degeneration, with clinical signs resembling the human disease. Other forms of MND are neonatal and juvenile. These include SMAs of various types, the most aggressive and earliest onset being type I (Werdnig-Hoffman disease), which is lethal within the first few months of life. In contrast to ALS, most forms of SMA are inherited, and in more

than 95% of cases, the mutations responsible have been pinpointed to the *smn* (survival of motor neurons) gene. There are good animal models of these forms of motor neuron disease as well.

Recent studies of neuromuscular pathology in mouse models of ALS and SMA, as well as other mutants, such as *pmn* and *wasted*, have revealed some remarkable common features. Specifically, neuromuscular synapses within some motor units in all these models degenerate or retract into their parent axons at early stages in the disease process, in advance of axonal or motor neuron cell body degeneration. This has led to three views: first, that these motor neuron disease variants may represent a form of dying-back neuropathy, beginning with primary pathology at neuromuscular synapses; second, that neurodegenerative processes in general may be compartmentalized, with independent mechanisms regulating the degeneration of synapses, axons, and cell bodies; and third, that degeneration of synapses in circumstances like MND may share some common molecular mechanisms with normal forms of synaptic plasticity, including synapse elimination in postnatal development.

These views are not universally held and remain open to stringent experimental validation. For instance, although there may be strong morphological and physiological similarities between developmental synapse elimination and MND pathology in SOD-dependent ALS, recent data suggest that synaptic degeneration in SOD1G93A mouse variants shows loss of synaptic vesicle proteins in advance of degeneration, which synapses undergoing elimination in development do not. Moreover, synapses in SOD1G93A mice degenerate synchronously in some muscles, suggesting a primary axonopathy (consistent with the disruptive effect of the SOD-1 mutation on axonal transport).

Synapses are known to be especially vulnerable to defects in axonal integrity, perhaps partly due to the high metabolic and energetic demands placed on them by the exigencies of synaptic transmission. Thus, it is perhaps not surprising that early loss of synapses could lead to the erroneous conclusion that these foci are the primary loci of disease. On the other hand, primary synaptic pathology has not been conclusively ruled out, either for SOD1-related ALS or for any other form of MND. Synapses contain abundant mitochondria, so these foci of oxidative metabolism may be especially vulnerable to direct, primary effects of oxidative stress. Either way, genetic or pharmacological strategies aimed directly at inhibiting synaptic degeneration could prove highly effective in mitigating disease progression. For instance, *pmn* mutant mice cross-bred with Wld^S mice showed strong evidence of mitigation of disease progression. However, transferring the Wld^S gene to SOD1G93A mice has, at best, only a weak neuroprotective effect. This may be partly explained by the weakening of the protective effect of the Wld^S gene on synapses as mice age. Screens of mutations induced by ethylnitrosourea are currently under way in a search for novel mutations that may protect synapses as effectively as Wld^S protects axons. Such mutations could ultimately be exploited so as to mitigate disease progression in several forms of motor neuron disease.

Summary

We still do not know precisely what genetic, environmental, or even stochastic factors determine whether a synaptic bouton should persist or be removed from a π-junction or indeed whether there is a single deciding factor. Activity levels and sensitivity to neurotrophic factors, selective recognition between motor neurons and muscle fibers, and other intrinsic properties of motor neurons and the muscle fibers they innervate may all help to establish the size of motor units, shaping their organization and orderly recruitment, as we accumulate repertoires of behavior during development that we carry with us into adulthood. Ultimately, all these potential regulators may individually play influential rather than decisive roles. For example, perhaps inactivity induces and sustains polyneuronal innervation merely by stimulating and maintaining nerve branching, enhancing the opportunities for competitive interactions to occur between synaptic terminals. Intrinsic determinants of synaptic strength, differential sensitivity to neurotrophic factors, and differential adhesion may then bias the outcome of the subsequent interplay as motor terminals advance or retreat over endplate territory defined by the boundaries of postsynaptic receptor clusters they require to mediate their primary function. But the final outcome could also be generated by chance, through some fundamentally stochastic process. By analogy, reliable chemical synaptic transmission is produced by an underlying stochastic mechanism in which action potentials elevate the probability of exocytosis rather than determining precisely how many vesicles fuse with release sites in presynaptic active zones. Perhaps chance and necessity, an apt epithet for the forces regulating molecular evolution, may also prove to be a fitting summary of the interplay between environment andgene expression leading to development of functionally appropriate innervation patterns, not only at NMJs but throughout the developing nervous system, thereby individualizing all voluntary behavior.

See also: Comparative Biology of Invertebrate Neuromuscular Junctions; Gap Junctions and Neuronal Oscillations; Motor Control of Feeding and Drinking; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction: Neuronal Regulation of Gene Transcription at the Vertebrate; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Neuromuscular Connections: Vertebrate Patterns of.

Further Reading

Barry JA and Ribchester RR (1995) Persistent polyneuronal innervation in partially denervated rat muscle after reinnervation and recovery from prolonged nerve conduction block. *Journal of Neuroscience* 15: 6327–6339.

Bishop DL, Misgeld T, Walsh MK, et al. (2004) Axon branch removal at developing synapses by axosome shedding. *Neuron* 44: 651–661.

Brown MC, Jansen JK, and Van Essen DC (1976) Polyneuronal innervation of skeletal muscle in new-born rats and its elimination during maturation. *Journal of Physiology* 261: 387–422.

Costanzo EM, Barry JA, and Ribchester RR (2000) Competition at silent synapses in reinnervated skeletal muscle. *Nature Neuroscience* 3: 694–700.

Gillingwater TH and Ribchester RR (2003) The relationship of neuromuscular synapse elimination to synaptic degeneration and pathology: Insights from Wld^S and other mutant mice. *Journal of Neurocytology* 32: 863–881.

Kasthuri N and Lichtman JW (2003) The role of neuronal identity in synaptic competition. *Nature* 424: 426–430.

Keller-Peck CR, Walsh MK, Gan WB, et al. (2001) Asynchronous synapse elimination in neonatal motor units: Studies using GFP transgenic mice. *Neuron* 31: 381–394.

Sanes JR and Lichtman JW (1999) Development of the vertebrate neuromuscular junction. *Annual Reviews of Neuroscience* 22: 389–442.

Schaefer AM, Sanes JR, and Lichtman JW (2005) A compensatory subpopulation of motor neurons in a mouse model of amyotrophic lateral sclerosis. *Journal of Comparative Neurology* 490: 209–219.

Walsh MK and Lichtman JW (2003) *In vivo* time-lapse imaging of synaptic takeover associated with naturally occurring synapse elimination. *Neuron* 37(1): 67–73.

Zuo Y, Lubischer JL, Kang H, et al. (2004) Fluorescent proteins expressed in mouse transgenic lines mark subsets of glia, neurons, macrophages, and dendritic cells for vital examination. *Journal of Neuroscience* 24: 10999–11009.

Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions

H L Atwood and M K Klose, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

General Features of Modulation at Invertebrate Neuromuscular Junctions

Modulation of neurotransmission in its broadest sense refers to alterations in preexisting synaptic efficacy due to engagement of pre- or postsynaptic modifying mechanisms by altered activity at the synapse or by influences of other cells that impinge upon the synapse. Modulation of neuromuscular transmission in muscles subserving locomotion and posture has been investigated in detail in the decapod crustaceans, and to a lesser extent in three orders of insects (Orthoptera, Diptera, and Lepidoptera). Apart from sporadic observations, little is known about mechanisms of modulation in other groups of invertebrates. However, the major invertebrate phyla share features of neuromuscular structure with arthropods: motor neurons form small individual synapses, often very numerous in large muscles. In the present account, we define 'synapse' as the closely apposed and tightly adhering pre- and postsynaptic membranes containing presynaptic structural specializations for transmitter release and postsynaptic arrays of receptors for released transmitter substances. The relatively small and discrete synapses of invertebrate neuromuscular junctions stand in contrast with the large consolidated neuromuscular junctions (end plates) typically found in many vertebrate muscles. In most cases, altered effectiveness of invertebrate neuromuscular transmission involves changes in the presynaptic nerve terminal rather than modification in responsiveness or number of postsynaptic receptors. The probability of transmitter release at individual synapses, or changes in the number of responsive synapses, or both, may change during modulation. Most likely, what has been found in arthropods will apply in other groups of invertebrates and also, at least to some extent, to central and peripheral synapses of vertebrates.

Modulation of neuromuscular transmission includes several different processes:

1. Altered synaptic transmission resulting from changes in steady-state membrane potential in the presynaptic nerve terminal.
2. Modulations of synaptic response properties arising from changes in impulse activity in the motor neuron (activity-dependent modulation).
3. Rapid changes imposed by other neurons synapsing on the neuromuscular junction. The best-known case is presynaptic inhibition in crustaceans.
4. Changes effected by nearby glial cells. This action is known to be significant at vertebrate neuromuscular junctions, but little is known about comparable actions at invertebrate neuromuscular junctions. It should be noted, however, that many invertebrate neuromuscular junctions, especially among dipteran insects, lack accompanying glial cells along most of the presynaptic terminal.
5. Changes in neurotransmission mediated by neurohormones released at a distance and carried to the neuromuscular junction by circulation, or by neuromodulators which diffuse from nearby nerve cells to target neuromuscular junctions. Amines and peptides are known to act as neuromodulators or neurohormones in invertebrates.
6. Reverse (retrograde) transmission or transport from muscle to motor neuron. Processes of this type are thought to underlie certain forms of long-term potentiation at mammalian central synapses. Strong indirect evidence implicates retrograde processes at invertebrate neuromuscular junctions. However, on present evidence most of these processes apparently act over a long period of time during development, and can be more usefully considered as processes of neuromuscular formation and development. The present account focuses on modulation of neurotransmission in fully formed neuromuscular junctions.

Presynaptic Membrane Potential

Steady-state membrane potential of the presynaptic nerve terminal exerts a modulatory effect on the release of neurotransmitter, as illustrated by studies of the single excitatory motor axon supplying the opener muscle of the crayfish leg (**Figures 1(a) and 1(b)**). When depolarizing current is applied to the terminal, asynchronous transmitter release is observed at membrane potentials close to –60 mV. Stimulated action potentials are reduced in amplitude by steady-state depolarization (due to reduction of the electrical driving force for Na^+ entry), yet synaptic transmission is enhanced; the converse scenario is apparent with steady-state hyperpolarization. Similar effects have been reported more recently for the mammalian calyx of Held synapse, and for amphibian retinal rod photoreceptors. In these cases, small changes in background levels of intracellular Ca^{2+} ($[Ca^{2+}]_i$), initiated by an increase in Ca^{2+} entry

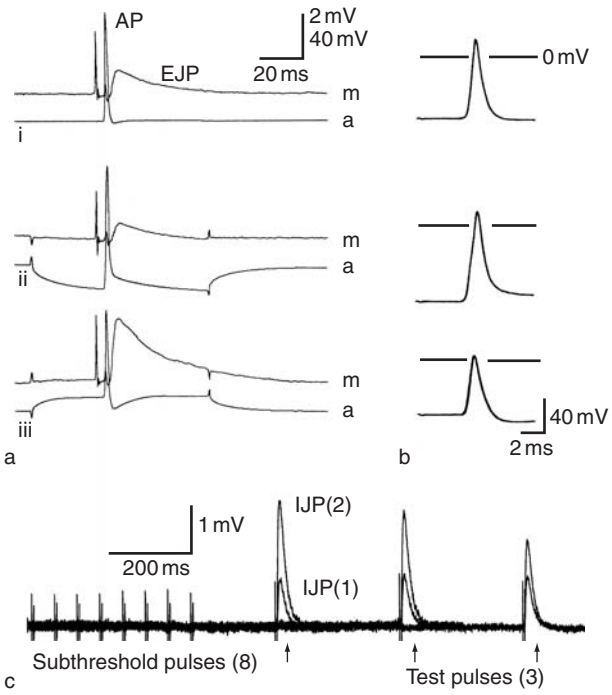

Figure 1 Modulation of neurotransmission by presynaptic membrane potential. (a) Intracellular recordings at a crayfish excitatory neuromuscular junction from presynaptic axon terminal (trace labeled 'a'), showing the presynaptic action potential (AP), and from postsynaptic muscle fiber (m), showing the excitatory junction potential (EJP). Current was injected into the presynaptic axon terminal to hyperpolarize (ii) and depolarize (iii) the presynaptic membrane. Hyperpolarization resulted in a larger AP than seen normally (i), and a smaller EJP; depolarization produced the opposite result. (b) Presynaptic action potentials in panel a shown at higher temporal resolution. (c) Enhancement of synaptic transmission by presynaptic conditioning pulses at a crayfish inhibitory neuromuscular junction. Inhibitory junction potentials (IJP) were recorded in response to three test pulses before conditioning stimulation (IJP(1)) and after eight subthreshold pulses of conditioning stimulation (IJP(2)). Conditioning pulses which do not cause neurotransmitter release enhance the IJPs evoked by test pulses; the enhancement decays with a time constant (737 ms) equivalent to that of the F2 component of short-term facilitation. (b) Reproduced from Wojtowicz JM and Atwood HL (1984) Presynaptic membrane potential and transmitter release at the crayfish neuromuscular junction. *Journal of Neurophysiology* 52: 99–113, with permission from The American Physiological Society. (c) Reproduced from Vyshedskiy A and Lin J-W (1997) Activation and detection of facilitation as studied by presynaptic voltage control at the inhibitor of the crayfish opener muscle. *Journal of Neurophysiology* 77: 2300–2315, with permission from The American Physiological Society.

through voltage-dependent Ca^{2+} channels, were shown to be responsible for modulating neurotransmitter release.

A related finding from a crayfish inhibitory neuromuscular junction is enhancement of neurotransmission by brief subthreshold depolarizing pulses which release no transmitter but augment subsequent release by suprathreshold pulses (Figure 1(c)). The enlarged synaptic response has the decay kinetics of short-term facilitation (F2 phase). Thus, membrane potential changes can activate a process of facilitation without releasing neurotransmitter. It is thought that $[Ca^{2+}]_i$, elevated by these pulses, acts on a Ca^{2+} receptor that has higher binding affinity than do the receptors that activate transmitter release. The most likely interpretation for these collective findings is that membrane potential modulates neurotransmitter release through regulation of the number of open presynaptic Ca^{2+} channels, with possible secondary mobilization of intracellular Ca^{2+} stores at some synapses.

Activity-Dependent Modulation

Short-Term Modulation

Invertebrate neuromuscular junctions, like most other synapses, are sensitive to previous neural activity, which sets in motion processes leading to enhancement or to depression of neurotransmission. Both forms of response are strikingly displayed at neuromuscular junctions of crustaceans, which consequently have been heavily utilized for descriptive and analytical studies. Neuromuscular junctions of tonic motor neurons, which are active during normal locomotion, exhibit dramatic facilitation at many terminals. In contrast, those of phasic neurons, which normally fire briefly to produce fast responses, exhibit synaptic depression.

Synaptic enhancement The features of the major categories of short-term enhancement that have been analyzed at crustacean neuromuscular junctions (NMJs) are summarized in Table 1. The buildup in amplitude of the excitatory junction potential (EJP) proceeds during maintained activity, and is modulated by the frequency of nerve impulses and by the duration of the impulse train. When a train of impulses ends, the EJP declines in amplitude. For analytical purposes, the time course of decline in EJP amplitude is measured; from this, three or four processes with distinct time courses are discerned, as shown in Table 1.

Short-term facilitation (STF), which is often described as having two components – F1, which is over in less than a second, and F2, which decays in 1–10 s – appears after one impulse, but often increases further with more impulses. Crustacean examples are illustrated in Figure 2. These show variable responses often seen at different NMJs of a single motor axon (Figure 2(a)), and an extremely long-lasting enhancement produced by a single impulse (Figure 2(b)).

The slower processes, augmentation and potentiation, require more stimulation to become evident, and decay more slowly. These processes were

Table 1 Activity-dependent synaptic enhancement of transmission in the crayfish opener motor neuron

Process	*Life span*	*Regulators*	*Sites of regulation*
Short-term facilitation (STF)	Up to a few seconds after one impulse	Ca^{2+} residue from previous impulse	Binding site for Ca^{2+} (synaptic vesicles?)
Augmentation (AUG)	Tens of seconds after a train of impulses	Ca^{2+} and Na^{+}	Surface membrane
Tetanic and posttetanic potentiation (POT)	Up to a few minutes after a train of impulses	Ca^{2+} and Na^{+}	Surface membrane, mitochondria
Long-term facilitation (LTF)	A few hours or longer after a train of impulses	Ca^{2+}, Na^{+}, cAMP, protein synthesis	Surface membrane, active zone

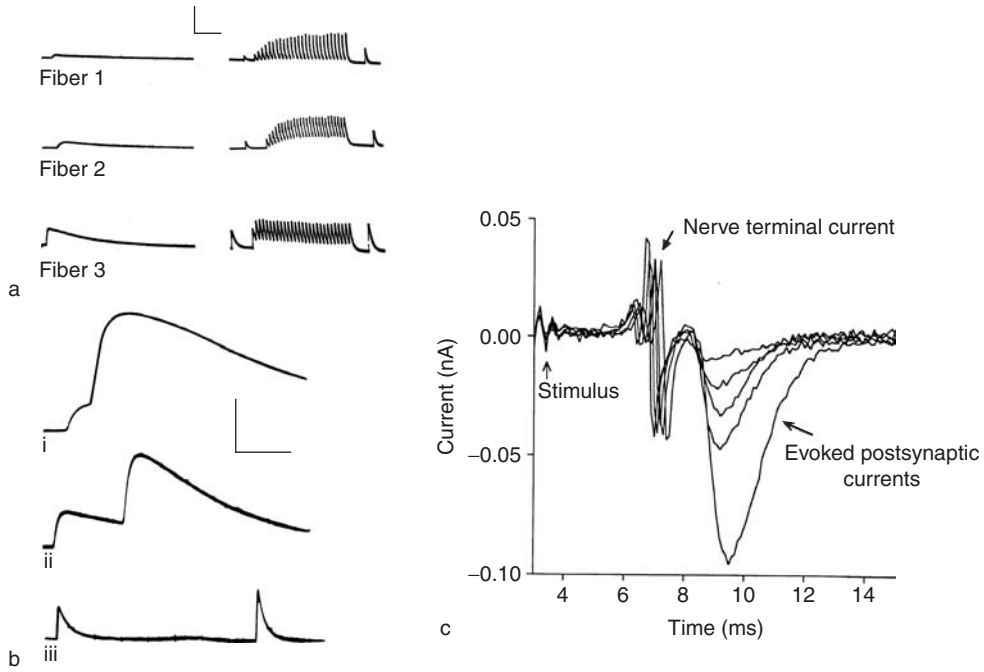

Figure 2 Manifestations of short-term synaptic facilitation at crustacean neuromuscular junctions. (a) Variation at different endings of one neuron. Excitatory junction potentials (EJPs) recorded from three different muscle fibers of the spider crab *Hyas* to show the range in excitatory junction potential amplitude and facilitation that occurs in this singly motor-innervated muscle. Initial stimulation frequency was 1 Hz; it was changed from 1 to 10 Hz and then to 1 Hz again (shown at lower temporal resolution in the right-hand records). During the 10 Hz trains, progressive enhancement of EJPs (frequency facilitation) occurs in fibers 1 and 2, but is not apparent in fiber 3. EJP amplitude and duration were smallest in fiber 1, intermediate in fiber 2, and largest in fiber 3. The extent of facilitation was in the reverse order (i.e., largest in fiber 1 and smallest in fiber 3). Enhancement of evoked EJPs after the 10 Hz train (posttetanic potentiation) is also inversely related to initial EJP amplitude. Vertical calibration: 7.5 mV, fiber 1; 15 mV, fibers 2 and 3. Horizontal calibration: 30 ms, fibers 1 and 2; 75 ms, fiber 3; 750 ms for 10 Hz trains. (b) Twin-pulse facilitation of EJPs at the neuromuscular junction of a stomach muscle in the crab *Callinectes sapidus*. Two pulses were separated by successively longer intervals: 20 ms (i), 250 ms (ii), and 4000 ms (iii). Facilitation of the second EJP persisted for approximately 10 s. Calibration: amplitude, 10 mV; times were 40 ms (i), 100 ms (ii), and 1 s (iii). (c) Frequency facilitation of synaptic currents recorded with an external macro-patch electrode at a crayfish excitatory neuromuscular junction. Averaged responses were obtained at frequencies of 2, 3, 4, 5, and 10 Hz from a recording site in the central region of the crayfish opener muscle. Averages of 200 sweeps are shown for each frequency. Transmitter release increased progressively with frequency of stimulation. (a) Reproduced from Sherman RG and Atwood HL (1972) Correlated electrophysiological and ultrastructural studies of a crustacean motor unit. *Journal of General Physiology* 59: 586–615. Copyright 1972 The Rockeyeller University Press. (b) From Atwood HL, Govind CK, and Jahromi SS (1977) Excitatory synapses of the blue crab gastric mill muscles. *Cell and Tissue Research* 177: 145–158. (c) From Wojtowicz JM, Marin L, and Atwood HL (1994) Activity-induced changes in synaptic release sites at the crayfish neuromuscular junction. *Journal of Neuroscience* 14: 3688–3703.

originally distinguished at the amphibian neuromuscular junction, but were then found at many other synapses. At some crustacean neuromuscular junctions, stimulation for several minutes at frequencies of 10 Hz or more induces a much longer-lasting process referred to as long-term facilitation (LTF; similar to certain forms of long-term potentiation in the vertebrate CNS).

All of these processes are attributed to changes in the amount of neurotransmitter released by the nerve impulse, and not to postsynaptic modulation of receptors. At the crayfish neuromuscular junction, entry of Ca^{2+} during voltage clamp of the presynaptic terminal does not increase during STF; thus, the Ca^{2+} trigger for exocytosis remains unchanged (Figure 3). The size of the quantal units (each representing exocytosis of one synaptic vesicle) remains constant, while the number of quantal units that compose an EJP increases. Recordings with extracellular electrodes at individual varicosities (Figure 2(c)) yield counts of the number of quantal units released by each impulse at several synapses on the varicosity. These data can be fitted to binomial distributions which incorporate two factors contributing to enhancement at the recording site. The binomial model is based upon the principle that quantal content, or quantal units released per impulse (m), is a product of the number of responding units (n) and the probability of response of each unit (p): in simplest form, $m = np$.

Numerous statistical analyses have shown in general that both n and p increase during short-term facilitation. Although several interpretations have been offered to explain the physical identities of n and p, the available evidence best supports the idea that n represents the number of responding synapses, since values for n are always similar to, but lower than, the number of morphologically discerned synapses. The value for p would then reflect an increase in response probability for the responding units. An extension of this model for crustacean neuromuscular junctions allows individual synapses on a varicosity to have different response probabilities, with recruitment of more synapses during higher frequencies of activity. Evidence in support of this view comes from statistical analyses and from experiments in which terminals exposed to the fluorescent vesicle-accumulated dye FM1–43 exhibit more bright spots (putative active synapses) after stimulation at higher frequencies.

Although best-fitted binomial distributions provide a formal description of changes in the presynaptic terminal, they do not in themselves reveal the underlying cellular mechanisms. The most commonly accepted proximate mechanism for STF is a persistent elevation of presynaptic $[Ca^{2+}]_i$. Another possibility, increased entry of Ca^{2+} with successive impulses, has been ruled out at the crustacean neuromuscular junction (Figure 3). On the other hand, persistent elevation of presynaptic $[Ca^{2+}]_i$ after impulse activity has been directly demonstrated in crustaceans with imaging techniques for intracellular Ca^{2+} (Figure 4(a)). Thus, action of Ca^{2+} on specialized receptors is most likely to account for increased probability of transmitter release (p) during STF, and probably also for increased n (additional active synapses), since general enhancement of p could bring initially silent synapses into play. (Direct

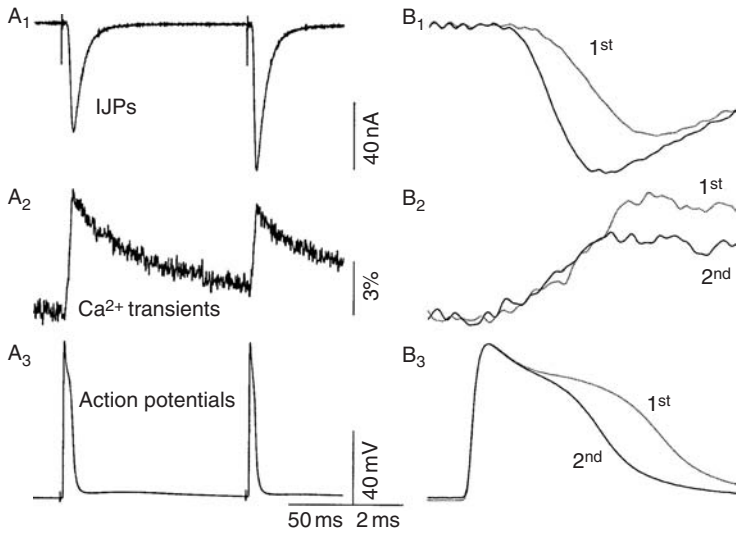

Figure 3 Calcium entry into crayfish inhibitory presynaptic terminal during short-term synaptic facilitation. Twin-pulse facilitation of inhibitory junction potentials (IJPs) occurs without increased entry of Ca^{2+}. Facilitation was induced by paired action potentials, with duration prolonged by potassium channel blockers to enhance Ca^{2+} entry. A_1 A_2, and A_3 illustrate IJPs, presynaptic Ca^{2+} transients, and presynaptic action potentials, respectively. The first (dotted line) and second (solid line) IJPs (B_1), Ca^{2+} transients (B_2), and action potentials (B_3) are superimposed and displayed at a higher time resolution. Although the second IJP is facilitated, the duration of the second action potential is shorter than that of the first one. The second Ca^{2+} transient has a lower amplitude than the first. Each trace represents the average of 100 trials. From Vyshedskiy A, Allana T, and Lin J-W (2000) Analysis of presynaptic Ca^{2+} influx and transmitter release kinetics during facilitation at the inhibitor of the crayfish neuromuscular junction. *Journal of Neuroscience* 20: 6326–6332.

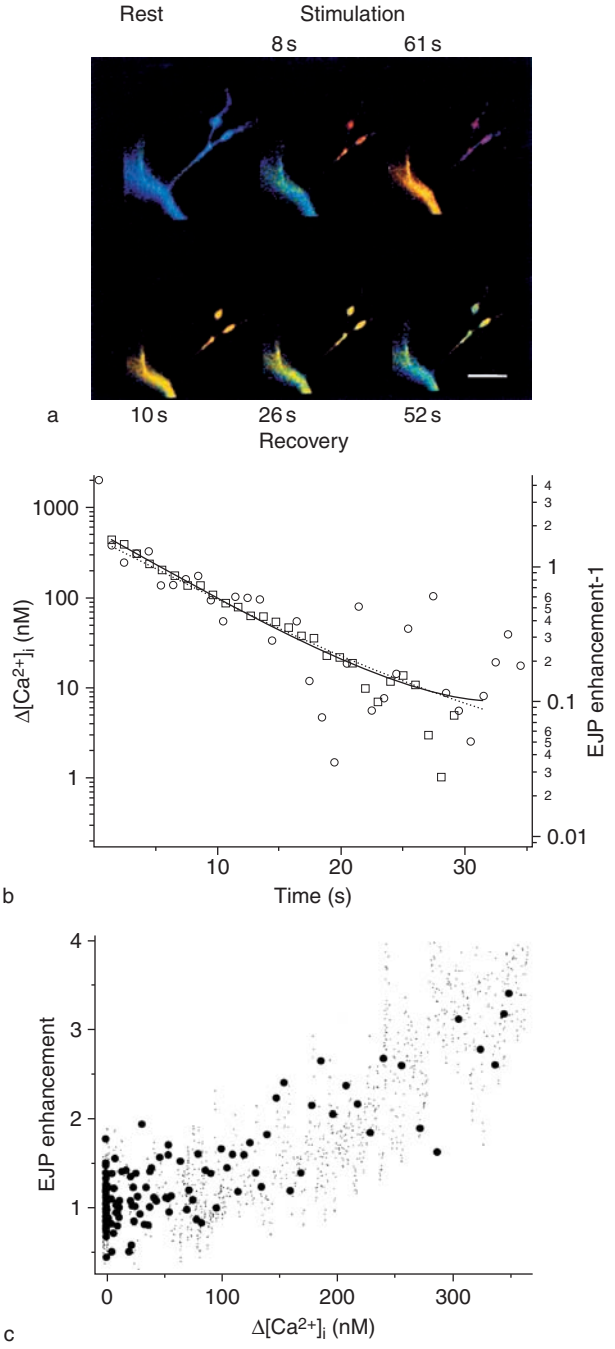

Figure 4 Residual presynaptic calcium and synaptic potentiation at crayfish neuromusclar junctions. (a) $[Ca^{2+}]$ in three synaptic boutons and an axon branch of the excitatory motor axon of the crayfish leg 'opener' muscle, observed with ratiometric fluorescence imaging, using the Ca^{2+}-sensitive dye fura-2. Images were taken at rest, during stimulation at 20 Hz, and during recovery after cessation of stimulation at the times indicated. Colors indicate $[Ca^{2+}]$ values: blue, 50–100 nM; yellow, 300–600 nM; orange, 600–1000 nM; red, 1000–1200 nM; violet, 1200–1600 nM. The small branches and boutons exhibit a larger and more rapid increase in Ca^{2+} than does the larger axon branch. Scale bar = 4 μm. (b) Decay of $[Ca^{2+}]_i$ (squares) and enhancement of synaptic transmission (circles) after tetanic stimulation (50 Hz for 4 s) in crayfish excitatory presynaptic terminals. The solid line is an exponential fit to excitatory junction potential (EJP) data, and the dashed line is an exponential fit to the $\Delta[Ca^{2+}]_i$ data, both beginning about 1.5 s after tetanic stimulation. The kinetic features of the two recovery processes match closely. (c) Relationship between $[Ca^{2+}]_i$ and synaptic enhancement after tetanic stimulation as in panel b (solid circles) and during elevation of $[Ca^{2+}]_i$ (dots) following application of the Ca^{2+} ionophore bromo-A23187 (0.5 μM). In both cases, there is a nearly linear relationship, and the amount of synaptic enhancement associated with $\Delta[Ca^{2+}]_i$ is comparable. EJP enhancement is the ratio of EJP amplitude during treatment to EJP amplitude prior to teatment. (a) After Zucker RS, Delaney KR, Mulkey R, et al. (1991) Presynaptic calcium in transmitter release and posttetanic potentiation. *Annals of the New York Academy of Sciences* 635: 191–207. (b, c) From Delaney KR and Tank DW (1994) A quantitative measurement of the dependence of short-term synaptic enhancement on presynaptic residual calcium. *Journal of Neuroscience* 14: 5885–5902.

evidence for addition of active synapses at crayfish excitatory NMJs has been provided by imaging a fluorescent dye, FM1-43, taken up in synaptic vesicles during exocytotic activity. More fluorescent spots are added with STF. Additional compelling evidence for the importance of residual Ca^{2+} derives from experiments with caged buffers which rapidly remove Ca^{2+} upon photolysis. This procedure abruptly decreases enhanced release of neurotransmitter.)

Short-term enhancement of transmitter release after nerve impulses appears to be linearly related to residual $[Ca^{2+}]_i$ (Figures 4(b) and 4(c)). This linear Ca^{2+}–EJP relationship for enhancement of transmission was initially an embarrassment for the residual Ca^{2+} concept, because the magnitude of the synaptic response to a single nerve impulse is a third- or fourth-order function of intracellular Ca^{2+}. Exocytosis of a single synaptic vesicle is therefore considered to require binding of three to five Ca^{2+} ions, and the vesicle-associated synaptotagmin molecule with five Ca^{2+}-binding sites is currently the preferred candidate for the Ca^{2+} receptor. Thus, it was not possible to account for both evoked release and facilitation with a single Ca^{2+}-binding site leading to release; instead, at least two binding sites, one for exocytosis (low affinity) and one or more for enhancement (high affinity), are implicated. Evidence for a separate high-affinity receptor for enhancement is presented in Figure 1(c). One possibility emerging from recent work is that Ca^{2+}-dependent priming of synaptic vesicles is promoted by residual Ca^{2+}, and then the high Ca^{2+} concentration of the nanodomain which appears at the active zone after a nerve impulse preferentially selects primed vesicles for exocytosis (Figure 5). A version of this model can also explain the dominance of STF at crustacean tonic synapses (more vesicle priming is required) and the depression at phasic synapses (vesicle priming is more complete, and docked vesicles have a higher

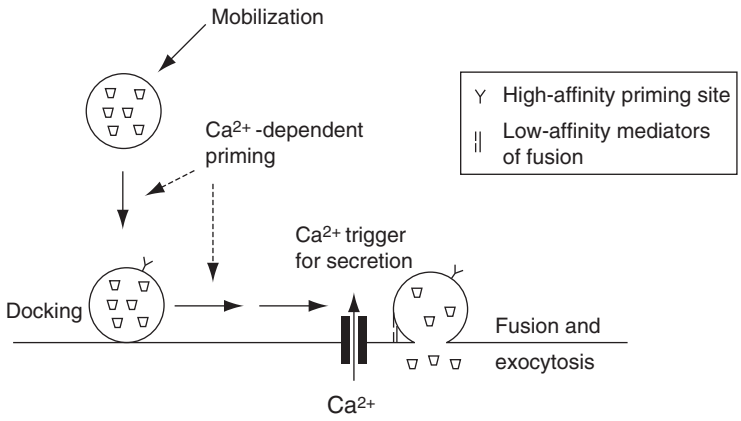

Figure 5 Diagram to illustrate that two separate Ca^{2+}-binding sites are needed to explain short-term facilitation and exocytosis of synaptic vesicles. Low-affinity sites activated by a relatively high Ca^{2+} concentration formed near calcium channels of the active zone are thought to trigger fusion of docked and primed synaptic vesicles, with rapid release of their contained neurotransmitter(s). One or more high-affinity sites that react at low concentrations of Ca^{2+} are thought to be responsible for facilitation, augmentation, and posttetanic potentiaton. The number and molecular identity of these sites, and the sequence of priming reactions, remain unknown at present.

probability of exocytosis). As yet, a good molecular candidate for the priming site (or sites) has not been discovered, although candidate molecules have emerged in recent work on *Drosophila* (Table 2).

Enhancement processes that usually require multiple impulses and longer initiation periods – augmentation and potentiation – also depend on residual Ca^{2+} (Figure 4). Their time courses are affected by impulse-mediated Na^{+} accumulation, which acts to slow the extrusion of Ca^{2+} by the surface membrane's Na^{+}–Ca^{2+} exchange. The plasma membrane Na^{+}–Ca^{2+} exchanger actually operates in reverse mode to accelerate buildup of presynaptic Ca^{2+} during intense tetanic stimulation. In addition, the slower process of potentiation is apparently prolonged by leakage of Ca^{2+} from mitochondria, where it is sequestered during periods of elevated presynaptic Ca^{2+} and released thereafter at a slow rate, sustaining the enhanced transmission. Whether different low-affinity Ca^{2+} receptors are responsible for STF, augmentation, and potentiation remains unresolved.

Synaptic depression On initial stimulation, crustacean phasic neuromuscular junctions typically generate large EJPs which depress dramatically as the stimulus is repeated (Figure 6). Tonic neuromuscular junctions, in contrast, show little depression at normal operating frequencies; depression can be induced only with very high nerve impulse frequencies, or when release is abnormally elevated by pharmacologically broadening the presynaptic action potential. Part of the explanation for the phasic–tonic differences lies in the much higher fraction of the readily releasable vesicles that undergo exocytosis at phasic synapses. This, in turn, is linked to the greater sensitivity of the phasic exocytotic machinery to Ca^{2+} and the higher priming rate postulated for synaptic vesicles in phasic terminals (Figure 5).

Depression occurs with very low frequencies at many phasic synapses (low-frequency depression; Figure 6). This type of depression cannot be attributed to depletion of readily releasable synaptic vesicles. Recent work on this obscure phenomenon has shown that low-frequency depression is relieved by blocking calcineurin and other phosphatases (Figure 6(c)). Phosphorylation of specific proteins attached to synaptic vesicles or the presynaptic membrane appears to counteract depression. The molecular targets of phosphorylation and dephosphorylation are as yet unknown.

With high frequencies of stimulation, a different form of depression appears which is linked to loss of readily releasable vesicles (high-frequency depression). The phasic–tonic disparity is partly explained by the smaller size of the vesicle pool, lower concentrations of cytoplasmic glutamate (GLUT), and higher fractional release of available vesicles per impulse in phasic neurons. All of this is most likely related to the supply of ATP by mitochondria, which is less robust in phasic than in tonic terminals, and is required to counteract high-frequency depression. There is good evidence from various invertebrate studies to support this view. Electron microscopic observations have shown loss of vesicles around synapses when fixation is carried out during high-frequency depression. Inhibiting mitochondrial ATP production accelerates depression in both phasic and tonic neuromuscular junctions. In fact, tonic junctions assume phasic characteristics: enlargement of the EJP (due to elevation of cytosolic Ca^{2+}) occurs soon after incapacitation of mitochondria, and rapid depression ensues. Another line of evidence comes from recent studies of *Drosophila* mutants

which lack mitochondria at the neuromuscular junction. These mutants exhibit rapid synaptic depression when repeatedly stimulated, even though initial EJP amplitude is close to normal; the defect can be rescued by supplying exogenous ATP directly through the motor axon to the neuromuscular junction.

Both forms of depression are presynaptic in origin. However, a small postsynaptic contribution due to desensitization of glutamate receptors has been described at *Drosophila* neuromuscular junctions: quantal size is reduced slightly as a result of prior glutamate exocytosis. Despite this, it is currently thought that the major cause of high-frequency synaptic depression at invertebrate neuromuscular junctions is the inability of presynaptic mechanisms to sustain continuous release of neurotransmitter at the required rate.

Long-Term Modulation

Long-term facilitation Activity-dependent enhancement of synaptic transmission lasting many hours or longer occurs in some facilitating motor neurons of crustacean limb muscles. After several minutes of tetanic stimulation at frequencies of 5–10 Hz or greater, and following the decay of posttetanic potentiation, EJPs tested with low-frequency stimulation remain stably enhanced at 150–200% of normal amplitude (long-term facilitation; LTF). The enhancement is entirely presynaptic – quantal size is unaltered, while quantal content increases – and does not involve changes in amplitude or duration of the presynaptic action potential (Figure 7). In many of its general features, LTF resembles a form of long-term potentiation described in the mammalian central nervous system (CNS) at hippocampal mossy fiber synapses on CA1 pyramidal neurons.

Binomial statistical analyses have indicated that the enhancement is entirely due to an increase in number of responding units (n), and not to a change in mean response probability (p). In this, LTF differs from STF. Evidence from electron microscopy suggests that some individual synapses acquire additional presynaptic densities, indicating the presence of more active zones for these synapses. No detectable change in number or disposition of synaptic vesicles was detected. These observations support a model in which a few synapses become modified structurally or functionally to respond more effectively to low-frequency stimulation (Figure 7). In a parallel case of presynaptic modification, direct evidence for addition of new synapses during long-term potentiation has been obtained from optical studies of dendritic spines of pyramidal cells in CA3 region of the mammalian hippocampus, using cultured slice preparations.

Recent experimental work on induction and expression of LTF has implicated several Ca^{2+}-activated kinases and phosphatases, including calcineurin. In addition, surprisingly, local protein synthesis contributes to establishment of LTF, as its expression is sensitive to translation inhibitors (cycloheximide and anisomycin). Collectively, these processes may be involved in synaptic modification, perhaps in structural modification of active zones; but the proteins and structures involved in synaptic modification are not yet defined.

Induction of LTF has been linked to activation of hyperpolarization-activated and cyclic nucleotide-activated (HCN) channels by hyperpolarization of the presynaptic membrane, which results from electrogenic surface extrusion of accumulating Na^+ by the Na^+/K^+-ATPase (sodium pump). HCN channels may be linked in turn to actin polymerization, which leads to enhancement by an unknown process. There is, however, conflicting evidence on the need for Na^+ accumulation in LTF induction, as earlier experiments demonstrated induction of LTF by depolarizing pulses in neurons treated with tetrodotoxin (TTX) to block action potential-mediated entry of Na^+. Thus, the details of the pathways and mechanisms for LTF remain to be confirmed and further elucidated, and the model for this process is presently incomplete.

The cellular components involved in short-term enhancement and long-term facilitation are diagramed in Figure 8, and candidate molecules (emerging from recent work in *Drosophila*) which are likely components of the mechanisms of synaptic modulation are presented in Table 2.

Experience-dependent strengthening At *Drosophila* larval neuromuscular junctions, a form of long-term plasticity that requires coordinated pre- and postsynaptic modification has been described. When larvae are forced to be more active, their neuromuscular junctions exhibit enhanced synaptic transmission within 2 h. Correlated with this, subsynaptic protein synthesis apparently increases, and soon extra glutamate receptors (DGluR-IIA subunits) start to accumulate around synapses. Additional boutons and synapses with active zones and postsynaptic receptor patches are added to the neuromuscular junctions. Similar changes are induced by genetic manipulation of genes for two glutamate receptor subunits (*dGluR-IIA* and *dGluR-IIB*), showing a link between increased expression of the DGluR-IIA subunit and recruitment of additional presynaptic active zones. Thus, increased activity may initiate a process resembling LTF (presynaptic enhancement). At the *Drosophila* neuromuscular junction, postsynaptic protein synthesis is then induced, which in turn feeds back to the presynaptic neuron,

Table 2 Learning, memory, and synaptic plasticity mutants in *Drosophila*

Gene product/protein	*Mutant or gene name*	*Cytogenetic positions*	*General physiological role*	*Effect of mutation on synaptic transmission and structure at larval NMJ (reduced function unless otherwise stated)*	*Effect of mutation on modulation at larval NMJ (reduced function unless otherwise stated)*	*Learning/memory phase affected*
Transmitters PACAP38	*amnesiac*	18F4–19A2 (X)	Neuropeptide activity: activates AC (*rutabaga*) through G-protein-coupled receptor; activates Ras/Raf pathway; induces slow depolarization during high-frequency stimulation; ↑ K^+ current		Eliminates both ↑ K^+ current and slow depolarization	MTM
Dopa decarboxylase (Ddc)	*Ddc*	37C1 (2L)	Catalysis of dopamine and 5-HT: ↑ dopamine and 5-HT			Associative memory deficit
Receptors G-protein sα	*Gsα60A*	60A12–A13 (2R)	Stimulatory G-protein: activates AC (*rutabaga*); most severe learning and memory mutant (complete abolishment)	↑ Post Gsα: ↑ amplitude mini's	↓ F acilitation; ↓ PTP	Acq
Metabotropic glutamate receptor subunit	*DmGluRA*	102C5–6 (4)	Metabotropic glutamate receptor subunit: modulatory effect requires G-protein β13F subunit	Asynchronous release; ↑ bouton size; ↓ bouton number	↑ Short-term facilitation; ↑ augmentation	
Ionotropic glutamate receptor subunit	*DGluRIIA*	25E6 (2L)	Ionotropic glutamate receptor subunit: involved in CaMKII-dependent retrograde synaptic strengthening pathway	↓ Quantal size; ↑ quantal content; ↑ rate of desensitization		
Notch	*Notch*	3C7–9 (X)	Surface receptor: ultrastructural remodeling			LTM
Channels Shaker cognate b	*shab*	63A1–2 (3L)	Delayed rectifier K^+ channel subunit		↑ Augmentation; ↑ facilitation	
Ether-a-go-go	*eag*	13A2–5 (X)	K^+ channel subunit: modulated by CaMKII; mutants have hyperexcitable nerves and muscles	↑ Mini's after discharge		Nonassociative memory deficit
Slowpoke	*slowpoke*	96A14–17 (3R)	Ca^{2+}-activated K^+ channel: large conductance			ARM
Shaker	*Shaker*	16F3–5 (X)	K^+ channel subunit: involved in shaping action potentials; controls basal transmitter release; possible point of converging pathways activated by PACAP32; adults shake in ether	↓ Presynaptic K^+ current; action potential widening; ↑ EJP amplitude due to ↑ transmitter release	↑ Facilitation; augmentation lost and depression ensues	STM: nonassociative memory deficit
Slowpoke-binding protein	*Slob*	28C1 (2L)	Binding protein: ↑ voltage sensitivity of slowpoke; CaMKII regulates binding to *slowpoke* and *leonardo*	↓ Channel open prob.; ↓ peak current		

Second-messenger kinases cAMP-dependent protein kinase 1 regulatory subunit of PKA	PKA-RI	77E8–F1 (3L)	Protein phosphorylation: inhibits function of kinase when associated with catalytic subunit; mutant causes constitutively active PKA	↑ PKA activity; ↓ EJC amplitude; ↓frequency of mini's; ↓ quantal size	Loss of large vesicle recruitment during periods of high activity; loss of retrograde modulation of transmitter release	MTM
cAMP-dependent protein kinase 1 catalytic subunit of PKA	PKA-CI/DCO	30C5 (2L)	Protein phosphorylation: becomes active when freed from regulatory subunit; phosphorylates proteins in the cAMP/PKA pathway, including CREB, which then induces 'learning and memory' genes; modulates L-type Ca^{2+} channel and Ca^{2+}-dependent K^+ channel; recruits vesicles from reserve pool	↓ PKA activity; ↓ EJC amplitude; no Δ in frequency of mini's; ↑ quantal size (requires DGluRIIA)	Loss of large vesicle recruitment during periods of high activity; loss of retrograde modulation of transmitter release	MTM
PKB	*Akt*	89B3 (3R)	Protein phosphorylation: insulin receptor signal transduction		Disrupts LTD; normal facilitation and PTP	
PKC	*PKC 98E* *PKC 53E* *PKC 51D*	98F6 (3R) 53E2 (2R) 51D5–6 (2R)	Protein phosphorylation: regulates structural formation of synapse			ARM
PKG	*Foraging*	24A2–4 (2L)	Protein phosphorylation: regulates polymorphic foraging behavior (rover vs. sitter)	For(r): ↑ PKG, ↑ K^+ current; for(s): ↓ PKG, ↓ K^+ current; ↑ mini's; ↑ transmitter release		Nonassociative memory deficit
Calcium/calmodulin-dependent kinase	*CamKII*	102D1–3 (4)	Protein phosphorylation: coordinates responses to neuroactive substances; part of retrograde synaptic strengthening pathway requiring bone morphogenetic protein; pre- and postdevelopment/maturation; activates *ether-a-go-go*, *slob*, *naylot*	↓ Post-CaMKII; ↑ bouton number; ↑ presynaptic T bars; ↑ quantal content	↓ Facilitation; ↓ augmentation; ↓ PTP	LTM
Ras oncogene at 85D	*Ras85D/Ras1*	85D19 (3R)	GTPase: converges with AC path to modulate K^+ channels at larval NMJ; part of MAPK/ERK path; PACAP38 effect eliminated in mutant		Eliminates modulatory effect of PACAP38 on K^+ current	Associative memory deficit
Raf/polehole	raf^{HM7}/*phl*	3A1 (X)	Protein phosphorylation: converges with AC path to modulate K^+ channels at larval NMJ; part of MAPK/ERK path; PACAP38 effect eliminated in mutant		Eliminates modulatory effect of PACAP38 on K^+ current	

Continued

Table 2 Continued

Gene product/protein	*Mutant or gene name*	*Cytogenetic positions*	*General physiological role*	*Effect of mutation on synaptic transmission and structure at larval NMJ (reduced function unless otherwise stated)*	*Effect of mutation on modulation at larval NMJ (reduced function unless otherwise stated)*	*Learning/memory phase affected*
Second-messenger pathway components Adenylyl cyclase–Ca^{2+} CaM sensitive	*rutabaga*	12F4 (X)	Converts ATP to cAMP: cAMP activates PKA, recruits vesicles from reserve pool; *rutabaga* mutation → ↓ cAMP and ↓ PKA activity	↓ Bouton number; ↑ bouton size; ↑ T-bars/synapse; ↓ ratio of docked/undocked vesicles; ↓ transmission; ↑ asynchronous release	↓ Facilitation; ↓ PTP; normal LTD	STM
cAMP phosphodiesterase	*Dunce*	3C9–D1 (X)	Degradation of cAMP: *dunce* mutation → ↑ cAMP and ↑PKA activity	Poorly defined synaptic specialization; ↑ T-bars/synapse; ↑ ratio of docked/undocked vesicles; ↑ asynchronous release	↓ Facilitation; ↓ PTP; Ca^{2+} dependence of transmitter release; normal LTD	STM
Calmodulin	*Cam*	48F1 (2R)	Calcium-binding protein: activates adenylyl cyclase (*rutabaga*); altered Ca^{2+} transport, NO regulation, and cytoskeletal regulation; muscular hypercontraction; disrupted locomotion	↑ Transmitter release; deregulation of muscle Ca^{2+} fluxes	Disrupted facilitation and PTP	
Neurofibromin 1	*Nf1*	96F9 (3R)	Guanosine triphosphatase-activating protein: role in PACAP38-induced ↑ K^+ current (through AC/cAMP path)		Eliminates modulatory effect of PACAP38 on K^+ current	STM
14–3-3	*leonardo*	46E6–8 (2R)	Hydroxylase: MAPK and PKA regulation; Ca^{2+}-induced transmitter release; mutants have↓ voltage sensitivity of Ca^{2+}-act. K^+ channel (*slowpoke*)	↓ Transmission	↑ Facilitation; ↓ augmentation; ↓ PTP	Acq
Phospholipase-A2	*radish*	11E1 (X)	Phospholipid metabolism			ARM
Structural α-Integrin subunit	*Volando/scb*	51E10–11 (2R)	Cell adhesion protein	↑ Synaptic arbors; ↑ EJP amplitude; ↓ Ca^{2+} dependence	↓ Short-term facilitation; ↓ augmentation; ↓ PTP	STM
Fasciclin II	*fasII*	4B1–3 (X)	Cell adhesion molecule: regulates synaptic structure; role in cAMP-dependent plasticity	↑ Bouton number; no Δ in quantal content		STM, LTM, Acq

Bruchpilot	*bruchpilot*	45D3–4; 45F2–6 (2R)	Structural protein: associated with cytoskeletal matrix at the active zone	↓ Evoked transmission; loss of T-bars	Loss of depression with transient short-term facilitation; ↑ PTP	
Dystrophin	*dystrophin*	92A5–10 (3R)	Structural protein: link between cytoskeleton and extracellular matrix; ↓ dystrophin; ↑ probability of release	↑ Transmitter release; no Δ in synapse size, T-bar number, vesicle pool size, post morphology, or sensitivity	↓ Facilitation	
Transcription/DNA regulation cAMP response element-binding protein (CREB)	*DCREB*	71E1 (3L)	Transcription factor: phosphorylated by PKA, which then activates transcription of 'learning and memory' genes	Post ↓ CREB: ↓ in transmitter release and ↑ bouton number; pre or post ↑ CREB: also ↓ transmitter release, but no Δ in bouton number	Pre ↑ CREB: impairs short-term facilitation	LTM
Origin recognition complex subunit	*latheo*	49F7–8 (2R)	DNA regulation protein	↓ Ca^{2+} dependence; ↑ EJP amplitude		Acq
Staufen	*staufen/ pumillio*	55B5–7 (2R)	Location and translation of mRNA			LTM
Pumilio	*pumilio*	85C4–DI (3R)	mRNA translation repressor			LTM
EIF-5°C	*extrabases eIF-5°C*	83B4 (3R)	mRNA translation initiation factor			LTM
Oscar	*oscar*	85B7 (3R)	Involved in staufen-dependent mRNA translocation			LTM
Adh transcription factor 1	*Nalyot/ADF-1*	42C3 (2R)	Myb-related transcription factor	↑ Bouton number		LTM
AP1	*AP-1/Kayak/ Dfos*	99B10–C1 (3R)	Immediate-early transcription factor: with CREB and cAMP, ↑ EJC amplitude; without CREB, ↑ bouton number	↑ AP-1: ↑ EJC amplitude, ↑ bouton number; ↓AP-1: ↓ EJC amplitude, ↓ bouton number		
Synaptic vesicle proteins Cysteine string protein	*csp*	79D3 (3L)	Synaptic vesicle protein and chaperone: phosphorylated by PKA; plays a key role in transmitter release	↓ EJP amplitude; ↑ presynaptic resting Ca^{2+} and Ca^{2+} entry; transmission block at moderate temperature	↓ Short-term facilitation; ↓ augmentation; ↓ PTP	
Synapsin	*syn*	85F16–86A1 (3R)	Synaptic vesicle protein: target of PKA, which recruits synaptic vesicle from reserve pool through spectrin			Nonassociative memory deficit STM/MTM

Continued

Table 2 Continued

Gene product/protein	*Mutant or gene name*	*Cytogenetic positions*	*General physiological role*	*Effect of mutation on synaptic transmission and structure at larval NMJ (reduced function unless otherwise stated)*	*Effect of mutation on modulation at larval NMJ (reduced function unless otherwise stated)*	*Learning/memory phase affected*
Syntaxin	*Syntaxin 1A*	95E1–2 (3R)	Synaptic vesicle protein			LTM
Synaptotagmin	*synaptotagmin IV*	84D3 (3R)	Vesicle-associated Ca^{2+} sensor: retrograde synaptic strengthening signal pathway (developmental and acute plasticity); postsynaptic Ca^{2+} ↑ → sensed by sytIV → vesicle fusion → retrograde signal → ↑ cAMP → ↑ PKA → alters cytoskeleton and ↑ transmitter release	↓ Amplitude of evoked current; ↓ varicose synaptic structures	Loss of the high-frequency stimulation-induced ↑ mini's	
Others Cathespin	*crammer*	56D3 (2R)	Cysteine protease inhibitor			LTM
Sarah	*sarah/nebula*	89B7 (3R)	Polypeptide involved in meiosis: required for eag activation; critical for mitochondrial function		↑ Facilitation	LTM
Frequenin	*frq1* *frq2*	16F7–17A1 17A1–2 (X)	Calcium-binding protein	↓ Bouton number; ↓ transmission		

AC, adenylyl cyclase; CaMKII, Ca^{2+}/calmodulin-dependent protein kinase II; eag, ether-a-go-go; STM, short-term memory; MTM, medium-term memory; LTM, long-term memory; ARM, anesthetic-resistant memory; Acq, acquisition of memory; LTD, long-term depression; PACAP, pituitary adenylyl cyclase-activating peptide; PK, protein kinase; PTP, posttetanic potentiation; pre, presynaptic; post, postsynaptic; EJC, excitatory junctional current; EJP, excitatory junctional potential; freq, frequency; mini, miniature synaptic currents (spontaneous events); ↑, increase; ↓, decrease; →, followed by; Δ, change/difference.

The majority of information presented in this table is derived from the following select review articles and research articles: Arredondo et al. (1998) *Genetics* 150: 265; Bogdanik et al. (2004) *Journal of Neuroscience* 24: 9105; Broadie et al. (1997) *Neuron* 19: 391; Cheng et al. (2005) *Nature Neuroscience* 8: 1577; Davis (1996) *Physiological Reviews* 76: 299; Davis (2004) *Neuron* 44: 31; Davis (2005) *Annual Review of Neuroscience* 28: 275; Davis et al. (1998) *Neuron* 20: 305; Dawson-Scully et al. (2000) *Journal of Neuroscience* 20: 6039; DiAntonio et al. (1999) *Journal of Neuroscience* 19: 3023; Dubnau and Tully (1998) *Annual Review of Neuroscience* 21: 407; Engel et al. (2000) *memory phase & Memory* 7: 341; Engel and Wu (1998) *Journal of Neuroscience* 18: 2254; Guo et al. (1997) *Science* 276: 795; Guo et al. (2006) *Journal of Neuroscience* 26: 4004; Haghighi et al. (2003) *Neuron* 39: 255; Horner et al. (2006) *Current Biology* 16: 1441; Kane et al. (1997) *Neuron* 18: 307; Margulies et al. (2005) *Current Biology* 15: R700; McGuire et al. (2005) *Progress in Neurobiology,* 76: 328; Michels et al. (2005) *memory phase & Memory* 12: 224; Renger et al. (2000) *Journal of Neuroscience* 20: 3980; Rohrbough et al. (2000) *Journal of Neuroscience* 20: 6868; Sanyal et al. (2002) *Nature* 416: 270; Steinert et al. (2006) *Neuron* 50: 723; Stewart et al. (2004) *Journal of Neurobiology* 61: 392; Ueda and Wu (2006) *Journal of Neuroscience* 26: 6238; van der Plas et al. (2006) *Journal of Neuroscience* 26: 333; Wadell and Quinn (2001) *Annual Review of Neuroscience* 24: 1283; Wang et al. (1994) *Neuron* 13: 1373; Wolfgang et al. (2004) *Developmental Biology* 268: 295; Yoshihara et al. (2005) *Science* 310: 858; Zhong (1995) *Nature* 375: 588; Zhong and Pena (1995) *Neuron* 14: 527; Zhong and Wu (1991) *Science* 251: 198; Zucker and Regehr (2002) *Annual Review of Physiology* 64: 355.

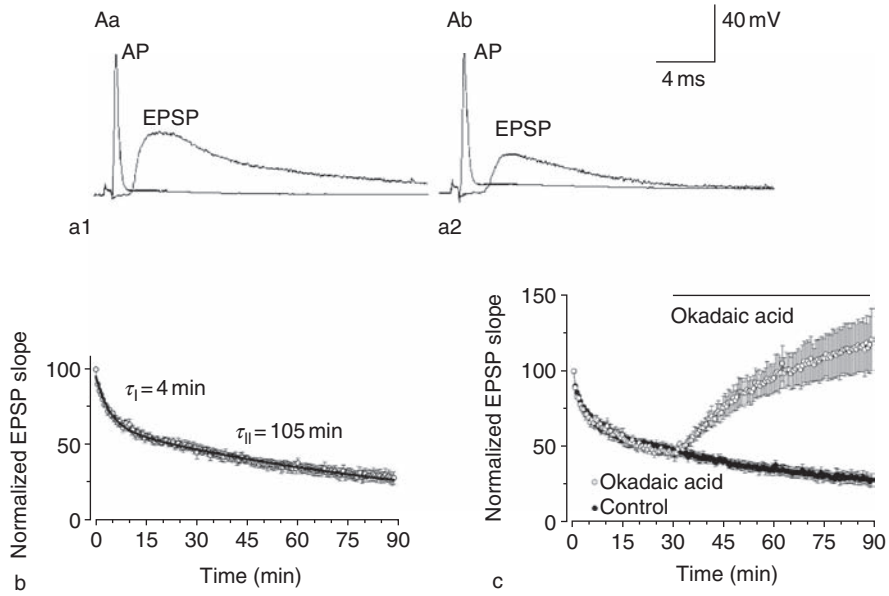

Figure 6 Low-frequency depression and the modulating role of phosphorylation in the presynaptic terminal. (a) Simultaneous recordings of the presynaptic action potential (AP) and the excitatory postsynaptic potential (EPSP) taken before (1) and after (2) induction of low-frequency depression in the phasic excitatory axon of the crayfish leg extensor muscle. Stimuli were delivered at 0.2 Hz for 75 min; the action potential retained its initial amplitude, while the EPSP declined. (b) Time course of decline of transmitter release at the phasic axon's neuromuscular junctions during 0.2 Hz stimulation; measurements were of maximum rate of rise (slope) normalized to the first EPSP. Two phases of depression appeared, with time constants of 4 and 105 min, respectively. (c) Recovery from depression with application of the phosphatase inhibitor, okadaic acid (10 μM), after 30 min of stimulation at 0.2 Hz. From Silverman-Gavrila LB, Orth PMR, and Charlton MP (2005) Phosphorylation-dependent low-frequency depression at phasic synapses of a crayfish motoneuron. *Journal of Neuroscience* 25: 3168–3180.

initiating additional presynaptic modifications. Protein kinase A has been implicated as a mediator of postsynaptic alterations, while an unknown retrograde signal is thought to initiate presynaptic changes. Feedback regulation of structure and synaptic strength have been investigated in several studies of *Drosophila* neuromuscular junctions (the details are beyond the scope of this article). The studies on experience-dependent strengthening illustrate the great advantages of the genetic approaches available in *Drosophila* for investigations of synaptic modification at the neuromuscular junction.

Long-term depression A presynaptically mediated form of long-term depression (LTD) has recently been analyzed at the larval *Drosophila* neuromuscular junction. Induction of this form of LTD requires tetanic activation of the motor axons at 30 Hz for 20 s, and persists for at least an hour. Mutations in the *akt* gene (which encodes protein kinase B/Akt) disrupt LTD expression but do not affect other forms of activity-dependent modulation (STF and potentiation).

Long-term adaptation Chronic modulation of the normal activity of crustacean phasic neurons leads to striking physiological and structural alterations at the neuromuscular junction. These neurons are normally inactive and are recruited sporadically for fast muscle contractions, when they fire one or a few impulses in a brief burst, evoking large, rapidly depressing EJPs. Imposition of tonic activity at 5 Hz for 1 h on two to three successive days initiates a transformation of the neuromuscular junction to a more tonic phenotype: the EJP becomes smaller in amplitude but more resistant to depression when challenged with continuous stimulation. Strikingly, the physiological changes are accompanied by morphological alteration of the neuromuscular junction. Varicosities enlarge and their contained mitochondria increase in volume and metabolic activity (Figure 9). The changes require several days for completion, and once established persist for several weeks if neural activity reverts to the normal phasic pattern. Activity of the motor neuron's central processes is required for this transformation; this is thought to induce synthesis and transport of proteins or aggregates that mediate synaptic modification. These adaptive changes have been demonstrated in normal populations of crayfish and are more readily induced in small, immature individuals, demonstrating greater plasticity of neuromuscular junctions in younger animals.

Partial analysis of the underlying mechanisms has been undertaken. Fatigue resistance of synaptic transmission correlates with mitochondrial volume

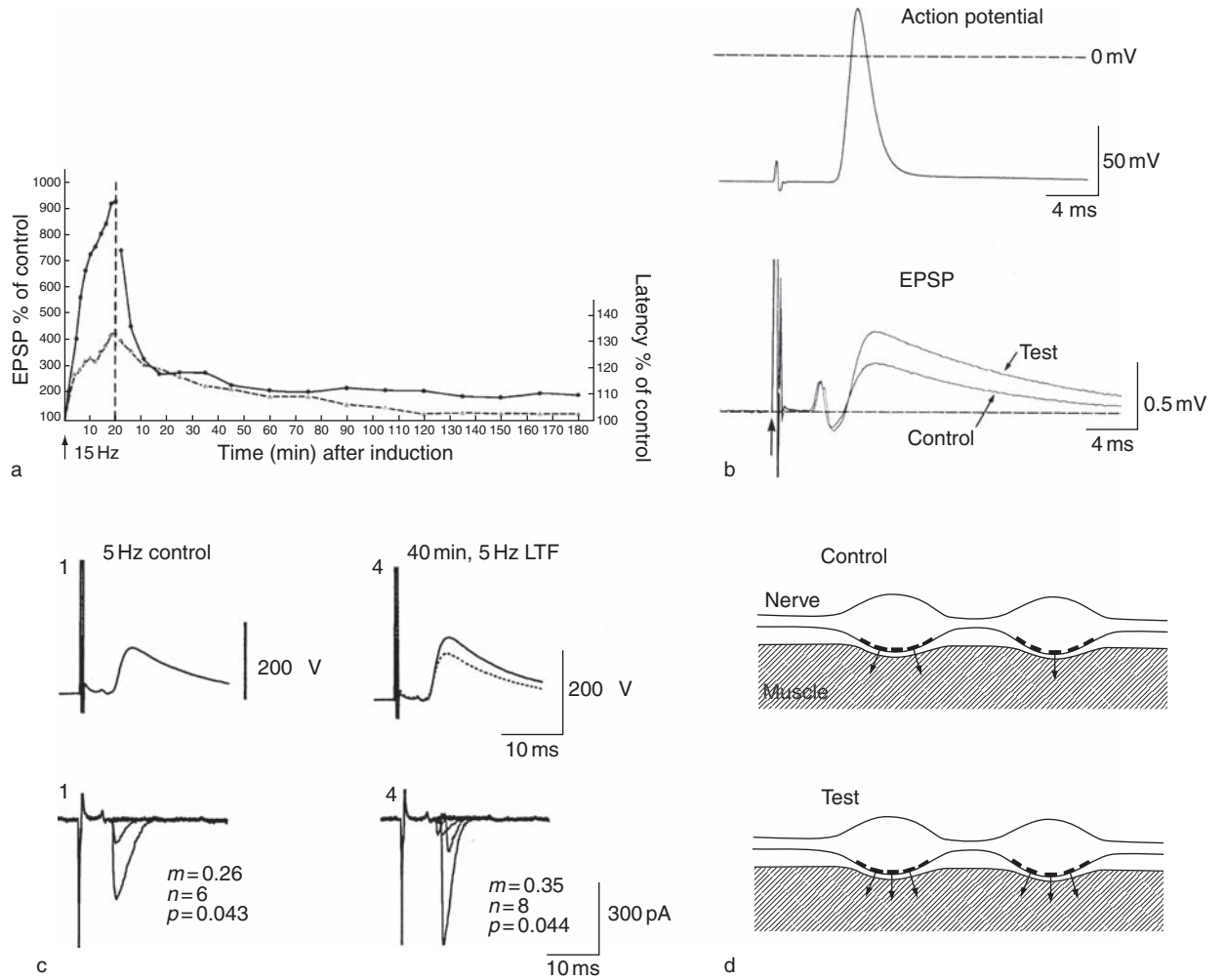

Figure 7 Long-term facilitation at the crayfish neuromuscular junction. (a) Time course of long-term facilitation of excitatory neurotransmission, and of simultaneous changes in conduction speed of the motor axon's action potential monitored by latency of the presynaptic nerve terminal potential (crayfish opener muscle). Growth of the excitatory postsynaptic potential (EPSP) to almost 1000% during stimulation at 15 Hz is shown, together with decay in amplitude of a test response (to 64 stimuli at 10 Hz) following the inducing stimulus train. Changes in response latency reverse more rapidly than does synaptic transmission, which remains elevated for many hours. (b) Simultaneous intracellular recordings of the action potential from a presynaptic nerve terminal, and the excitatory postsynaptic potential from a postsynaptic muscle fiber in the crayfish leg 'opener' muscle. Records were taken during 5 Hz stimulation before (control) and 40 min after (test) tetanic stimulation at 20 Hz for 10 min. The EPSP was persistently enhanced by tetanic stimulation, but there was no corresponding change in the presynaptic acton potential. (c) Quantal events at a single presynaptic bouton (lower traces) associated with persistent enhancement of the EPSP (upper traces). Records were taken at 5 Hz before and 40 min after tetanic stimulation at 20 Hz for 10 min. Mean quantal content (m) at the selected bouton increased by 35%; this change was due to an increase in the number of responding units (n) rather than to increased mean probability of response (p), according to best-fitted binomial distributions for the data. (d) A model for long-term facilitation (LTF) shown in panel c, illustrating recruitment of previously ineffective synapses to the responding population following induction of LTF. (a) After Atwood HL, Swenarchuk LE, and Gruenwald CR (1975) Long-term synaptic facilitation during sodium accumulation in nerve terminals. *Brain Research* 100: 198–204. (b,d) Reproduced from Wojtowicz JM and Atwood HL (1986) Long-term facilitation alters transmitter releasing properties at the crayfish neuromuscular junction. *Journal of Neurophysiology* 55: 484–498, with permission from The American Physiological Society; (c) from Wojtowicz JM, Marin L, and Atwood HL (1994) Activity-induced changes in synaptic release sites at the crayfish neuromuscular junction. *Journal of Neuroscience* 14: 3688–3703.

and mitochondrial membrane potential. Reduced EJP amplitude is linked to smaller quantal emission rates per synapse; however, the basis for reduced exocytosis has not been fully analyzed. The large size of the motor neurons in crustaceans affords an opportunity for single-cell analysis of changes in protein synthesis during long-term adaptation. Similar adaptive manifestations have been reported at other neuromuscular junctions and may also occur at central synapses.

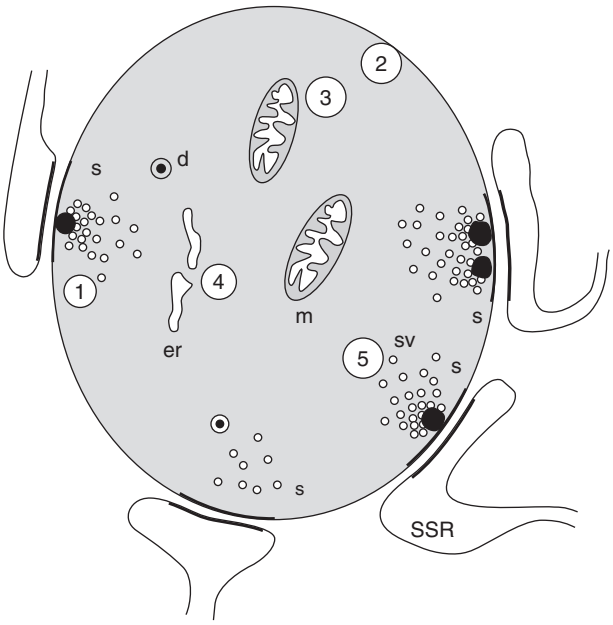

Figure 8 Diagram of the sites of activity-dependent modulation in crustacean presynaptic nerve terminals. Structures shown are mitochondria (m), synaptic vesicles (sv), endoplasmic reticulum (er), synapses (s), dense-cored vesicles (d), and subsynaptic reticulum (SSR). Sites of action are (1) synapse and active zone (modified in long-term facilitation), (2) surface membrane (site of Na^+/K^+–ATPase, Na^+/Ca^{2+} exchanger, and Ca^{2+}-ATPase, all of which participate in regulation of Na^+ and Ca^{2+}, affecting facilitation, augmentation, and posttetanic potentiation), (3) mitochondria (involved in posttetanic potentiation and high-frequency depression), (4) endoplasmic reticulum (a source of intracellular Ca^{2+}), and (5) synaptic vesicles (involved in short-term facilitation, augmentation, and posttetanic potentiation).

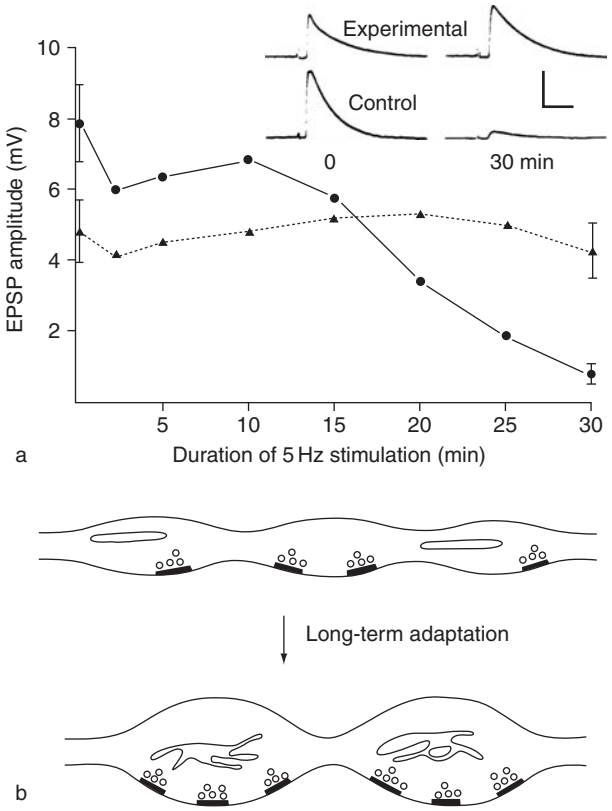

Figure 9 Long-term adaptation of synaptic transmission at a crayfish phasic neuromuscular junction. (a) Comparison of excitatory postsynaptic potentials (EPSPs) at neuromuscular junctions of phasic excitatory motor axons in normal (control) and conditioned (experimental) claws of the crayfish *Procambarus clarkii*. The axons were stimulated at 5 Hz for 30 min, and EPSPs were recorded throughout the period of stimulation. The time course of changes in EPSP amplitude is shown in the graph for representative muscle fibers of a conditioned claw (experimental; solid triangles) and the contralateral unconditioned claw (control; solid circles). Each point on the graph represents the mean value for 10 muscle fibers. Conditioning was achieved by activating the experimental phasic axon at 5 Hz for 2 h a day over a 14-day period. EPSPs in the experimental claw are initially smaller, but decline less with continued stimulation, compared to EPSPs in the control claw. Scale bar = 4 mV, 20 ms. (b) Morphological transformation of nerve terminals of the fast axon after 2 weeks of conditioning, as observed in serial electron micrographs. Varicosities are more pronounced and contain a greater volume of mitochondrial material after conditioning. Reproduced from Atwood HL and Lnenicka GA (1987) Role of activity in determining properties of the neuromuscular system in crustaceans. *American Zoologist* 27: 977–989, with permission from Oxford University Press.

Molecular Approaches

Genetic and molecular studies in the fruit fly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans* have recently provided numerous candidates for molecules that affect synaptic transmission and synaptic plasticity; some of these are also linked to processes of learning and memory. A catalog of significant molecular candidates found in *Drosophila* is provided in Table 2, which illustrates that several different categories are involved in regulation of synaptic performance. Most physiological processes depend on interacting networks of molecules rather than on single players. Thus, establishing the relative importance of individual molecular species for synaptic performance typically requires many independent studies with various techniques.

Inhibitory Modulation

Presynaptic Inhibition

Peripheral inhibitory neurons of crustaceans and insects synapse extensively on muscle fibers of limb and abdominal muscles, producing inhibitory junction potentials (IJPs). This inhibitory action affects the postsynaptic muscle fiber membrane directly (postsynaptic inhibition). An additional indirect effect known as presynaptic inhibition is prominent in limb muscles of some groups of decapod crustaceans, including crayfish and crabs.

Presynaptic inhibition is mediated by axo-axonal synapses formed by inhibitory nerve terminals on

excitatory neuromuscular junctions. When an inhibitory nerve impulse activates axo-axonal synapses 1–4 ms before the arrival of an excitatory impulse, transmitter release at the excitatory synapses is reduced, often drastically (Figure 10). That this effect is presynaptic was first established by demonstrating reduction in number of emitted quantal units comprising the EJP without an effect on quantal size. Intracellular recordings from excitatory axons confirm the occurrence of evoked inhibitory potentials in response to stimulation of the inhibitory axon. These inhibitory potentials have the same reversal potential as that obtained for exogenously applied γ-aminobutyric acid (GABA), the transmitter substance of inhibitory neurons. The amplitude of the excitatory axon action potential is significantly reduced during inhibitory stimulation. It is clear that presynaptic inhibition at crustacean excitatory neuromuscular junctions results from activation of GABA receptors on the excitatory axon, which when opened increase the axonal membrane's conductance for Cl^- and reduce the amplitude of the excitatory action potential. The smaller action potential is less effective in opening calcium channels and triggering exocytosis at affected excitatory synapses. Modeling of the effects of the inhibitory conductance change on conducted action potentials has indicated that these effects are accentuated by the varicose structure of the excitatory neuromuscular junctions.

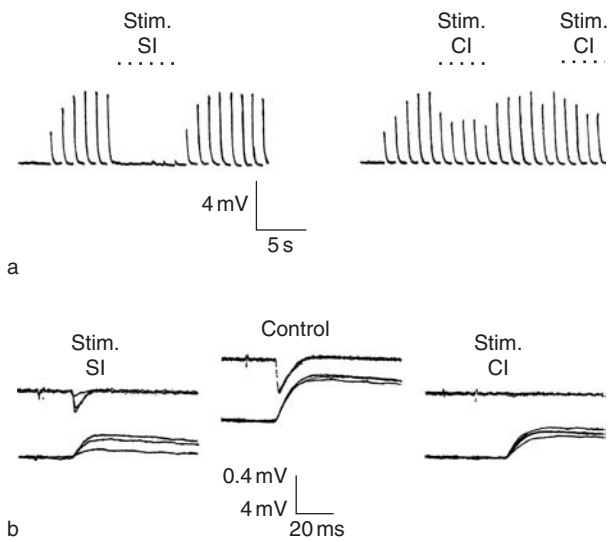

Figure 10 Presynaptic inhibition at a crab neuromuscular junction. (a) Inhibition of excitatory junction potentials at a neuromuscular junction in the leg 'stretcher' muscle of the crab *Hyas araneus*. Stimulation of the excitatory axon at 1 Hz produced large excitatory junction potentials, which were almost completely abolished by optimally timed stimulation (Stim) of the specific inhibitory (SI) axon. The common inhibitory (CI) axon was less effective for comparable stimulation. (b) Reduction of quantal emission at a single bouton shown by extracellular recording during stimulation of SI and CI axons. Upper traces, excitatory junction potentials; lower traces, extracellular recordings from a single bouton. In each record, several superimposed traces are shown, taken at 1 Hz stimulation. The control recording is without inhibition; the other records show the effects of optimally timed SI and CI pairing. Transmission at this bouton is reduced more drastically by the CI axon, although the SI axon is more effective in reducing the excitatory postsynaptic potential of the muscle fiber as a whole. From Wiens TJ and Atwood HL (1975) Dual inhibitory control in crab leg muscles. *Journal of Comparative Physiology* 99: 211–230.

Pharmacological experiments have implicated two types of GABA receptors in crustacean presynaptic inhibition. The predominant $GABA_A$ ionotropic receptors are coupled to Cl^- ionophores, while slower-acting metabotropic $GABA_B$ receptors activate G-proteins, which in turn probably act to open K channels in the membrane of the excitatory axon. The ultimate effect of both receptor types is to reduce excitatory transmission by lowering the peak of the presynaptic action potential. At neuromuscular junctions of crabs and crayfish, $GABA_A$ receptors account for 70–90% of presynaptic inhibition, and $GABA_B$ receptors account for 10–30%. The proportional contributions vary among different crustacean muscles that have been investigated.

The functional significance of presynaptic inhibition is debatable. It occurs in relatively few of the muscles that have been investigated. The requirement for precise timing of inhibitory impulses is not met during normal activity in many crustaceans, although one report suggests that timing is utilized in a crab species. There is no doubt that presynaptic inhibition adds efficiency to rapid suppression of muscle contractions during locomotion, and in some cases allows short-term facilitation of EJPs to be retained during the period of inhibition, thereby promoting faster contractions for the next cycle of movement.

Autoinhibition

At crayfish neuromuscular junctions, autoinhibition of quantal release by micromolar concentrations of L-glutamate has been described. The receptor responsible for this inhibition reportedly activates a glutamate transporter with an associated Cl^- ionophore. Increased Cl^-conductance mediates the autoinhibition. An additional glutamate autoreceptor that facilitates quantal release has also been reported for both crayfish and insect neuromuscular junctions, suggesting the occurrence of multiple autoreceptors on many arthropod neuromuscular junctions. The normal functional roles of these receptors have not been worked out.

Neuromodulators and Neurohormones

General Features

The activity resulting from synaptic communications can be altered by neuroactive substances from adjacent or distant neurons. Neuromodulators are released by neurons at synapses and access their targets by diffusion, modulating the activity of other local cells. Neurohormones enter the circulatory system to alter the cellular activity of cells distant from their release sites. Several types of modulatory substances exist, including amines, peptides, steroids, and lipids/phospholipids. Many neuroactive substances have multiple roles as neurotransmitters, neuromodulators, and neurohormones. Here we summarize the effects on invertebrate neuromuscular transmission of the two most-studied groups of neuroactive modulatory substances, amines and peptides.

The physiological activities regulated by amines and peptides are very diverse and include tissue growth, metabolic pathways, immune system activity, apoptosis, and behavior such as fight or flight responses. Early work in lobsters revealed that several of these neuroactive substances could induce fight- or flightlike behaviors. Amines and peptides affect behavior by acting both centrally, on the cell bodies of motor neurons and the circuitry controlling them, as well as peripherally, on muscles and the nerve endings which excite them. We focus here primarily on the modulation of exoskeletal neuromuscular transmission. The sites of modulation in the presynaptic nerve terminal are indicated in Figure 11.

Several aminergic and peptidergic modulators, including octopamine and pituitary adenylyl cyclase-activating peptide (PACAP)-like peptide, act both locally and distally to their release sites. Thus, the actions of these signaling molecules are determined not only by the receptors they bind to and the signal transduction pathways they activate, but also by the pattern of their release. In addition, activity-dependent modulation of the crayfish opener muscle by octopamine first revealed that the level of modulation depends on the target cell's activity state.

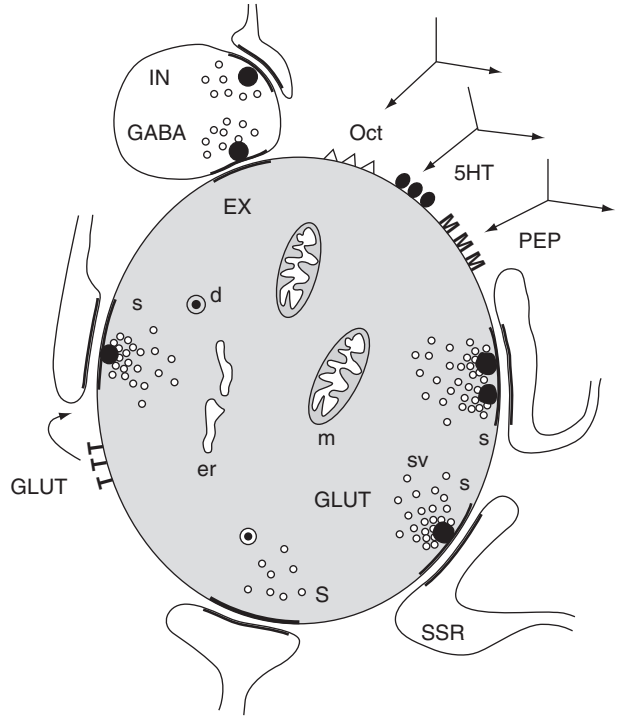

Figure 11 Diagram of the sites of external modulation in a crustacean presynaptic nerve terminal. Excitatory (EX) and inhibitory (IN) nerve terminals are shown, together with an inhibitory axo-axonal synapse mediating presynaptic inhibition. Transmitter substances are γ-aminobutyric acid (GABA) and glutamate (GLUT). Surface receptors on the excitatory terminal are for neurohormones and neuromodulators: octopamine (Oct), serotonin (5HT), and peptides (PEP). In addition, glutamatergic autoreceptors are indicated. Structural features: mitochondria (m), synaptic vesicles (sv), endoplasmic reticulum (er), synapses (s), dense-cored vesicles (d), and subsynaptic reticulum (SSR).

Amines

Amines are derived from amino acids and include indoles such as serotonin, and catecholamines such as dopamine and octopamine.

Octopamine The biogenic monoamine octopamine (OA) is related structurally and functionally to noradrenaline; both act as 'stress hormones' in fight or flight responses. Octopamine has no physiological role in vertebrates, making it the focus of much pesticide research, since antagonists for OA would likely not affect humans but would selectively target insects. The modulatory effects of OA in invertebrates are very broad and include effects on flight muscles, heart muscles, oviducts, endocrine glands, hemocytes, and fat bodies; on sensory receptors involved in taste, vision, olfaction, proprioception, and phermone responses; and on central nervous system activity related to learning and memory, motivation, and rhythmic motor behaviors.

Octopamine (OA) can be released directly into the circulatory system from the pericardial organ of crustaceans, or the copora cardiacum of insects, both of which are neurosecretory structures containing several other neurohormones. OA is found in the dorsal unpaired median (DUM) and ventral unpaired median (VUM) neurons of insects from which it is released into most neuropils of the brain, acting as both neurotransmitter and neuromodulator. DUM neurons are believed to provide the majority of OA

released into the hemolymph. *Drosophila* type II motorneuron terminals, which form small (1–2 μm) varicosities, contain OA and its cotransmitter glutamate. These terminals are formed by three VUM neurons, each covering either the dorsal, ventral, or lateral musculature. Thus, OA can be released either from neurosecretory structures into the circulation or onto neurons and muscles from nerve endings. Octopamine is typically released during bouts of high neural activity, as might be expected when an animal is startled or in an excited state.

Activity-dependent release of OA at the locust extensor tibiae neuromuscular junction increases both muscle tension and the rate of muscle relaxation, along with modest enhancement of EJPs. Thus, OA induces a physiological transition in the state of the muscle from one which favors maintenance of posture to one which favors rapid movements. The opposite occurs in a similar but primitive orthopteran insect, the 'weta' (*Hemideina femorata*) which does not use its legs for jumping but instead exhibits a defensive posture by extending the legs and rotating them toward the threat. Instead of promoting rapid changes in tension, as in locusts, OA induces a very long-lasting plateau of catchlike tension accompanied by large increase in EJP amplitude. The differences in muscle tension responses in orthopterans may be due to postysynaptic (muscle) properties. In the marine gastropod *Rapana thomasian*, OA enhances the radula protractor twitch contraction, while it reduces it in the antagonistic retractor by hyperpolarizing the muscle. This functionally significant coordination of opposite actions on antagonistic muscles is termed 'orchestration.'

A single type of modulator can have widely differing actions, depending upon its target cell. The G-protein-coupled OA receptors have been found in several invertebrate tissues, including locust leg and flight muscles. Three subtypes have been elucidated using the extensor tibiae muscle preparation: type 2A is found presynaptically, type 2B is found postsynaptically, and class 1 is restricted to the myogenic bundle, a small group of muscle fibers that is spontaneously active. A third type of receptor (class 3) is found in the CNS. All of these OA receptors are G-protein coupled. Binding of OA dissociates the $G\beta\gamma$ subunit from the $G\alpha$ subunit. Depending upon the type of receptor, the $G\alpha$ subunit can either stimulate or inhibit adenylyl cyclase activity, leading to either an increase or a decrease in cAMP levels. The second-messenger cAMP regulates protein kinase A (PKA) activity, which has modulatory actions on various ion channels (K^+), ion pumps (Na^+/K^+-ATPase), and transcription factors (cAMP response element-binding protein; CREB). In the locust extensor tibiae preparation, OA increases muscle Cl^- conductance, decreases K^+ conductance, and increases Na^+/K^+ pump activity; presynaptically, it enhances transmitter release. The mechanism for its presynaptic effects is not presently known.

Serotonin Serotonin (5-hydroxytryptamine; 5-HT) is a very important neuroactive substance which acts as a neurotransmitter, neuromodulator, and neurohormone for invertebrates, with widespread and profound effects on motor programs, behavioral repertoires, synaptic transmission, and muscular contraction. Serotonergic neurons of invertebrate central nervous systems typically have wide distributions of their processes within the CNS, and often project in peripheral nerves to neurosecretory organs or zones, where they release serotonin into the circulating hemolymph. Decapod crustaceans (lobsters and crayfish) have provided the most detailed information on serotonin's effects at neuromuscular junctions. In contrast, insects, including *Drosophila* larvae, have not shown the dramatic effects of serotonin at neuromuscular junctions characteristic of crustaceans.

Initial studies in crustaceans established that serotonin, at low concentrations, makes muscles produce more action potentials and greatly enhances synaptic transmission at both excitatory and inhibitory neuromuscular junctions. Enlargement of EJPs is primarily due to increased quantal content. The enhancement occurs rapidly and decays in two phases after removal of serotonin: a rapid phase, which is over in less than 30 min, and a long-lasting phase prolonged for 1–2 h. Significantly, the duration of serotonin enhancement is much less than for LTF, and different cellular mechanisms are evidently utilized.

The two phases of enhancement apparently represent activation of two presynaptic second-messenger cascades, though whether two different serotonin receptors are involved has not been determined. Activation of phospholipase C is implicated by the effects of pharmacological inhibitors directed against it: both phases of enhancement are completely eliminated. The initial phase is imitated by inositol 1,4,5-trisphosphate (IP_3) injections into the presynaptic axon, while the long-lasting enhancement likely involves activation of protein kinase C and adenylyl cyclase, leading to increased intracellular cAMP.

Given elevated cAMP, the logical next step would be increased activity of protein kinase A and phosphorylation of susceptible presynaptic proteins involved in aspects of excitation–secretion coupling. However, recent pharmacological manipulations at the crayfish excitatory neuromuscular junction

present a different scenario. Rather than protein kinase A, the prime targets for cAMP are (a) hyperpolarization and cyclic nucleotide-activated (HCN) ion channels and (b) exchange factor activated by cAMP (Epac). The first target is coupled to polymerization of actin; the second is coupled to presynaptic proteins associated with synaptic vesicles and the exocytotic machinery, including the G-protein Rab3 and its interacting partner, Rim. Collectively, the effects of these intermediate targets for cAMP are responsible for enhancement of transmission. Involvement of Epac in synaptic enhancement has now been reported for other synapses, including the *Drosophila* neuromuscular junction and the mammalian calyx of Held in the auditory pathway.

A striking feature of serotonin's effect is the reduction in requirement for Ca^{2+} in exocytosis. Transmission can take place in very low concentrations of external Ca^{2+} after serotonin treatment. This indicates increased Ca^{2+} sensitivity of exocytotic mechanisms. In addition, serotonin treatment does not lead to a detectable increase in presynaptic Ca^{2+} with action potentials, contradicting expectations that prime effectors might be synaptic Ca^{2+} channels or endoplasmic reticulum Ca^{2+} stores. Synaptic vesicle recycling is much enhanced, suggesting that a larger proportion of synaptic vesicles is in the readily releasable pool. Thus, the most likely targets for serotonin's effect are the synaptic vesicles, which appear to be more fully primed and mobilized for exocytosis, and one or more of the proteins that control exocytosis or endocytosis.

Peptides

FMRFamide-like peptides The tetrapeptide Phe-Met-Arg-Phe-NH_2 (FMRFamide) and homologous peptides containing the FMRFamide or FLRFamide sequence form the largest known group of invertebrate peptides. Several of theses peptides are found in the pericardial organ of crustaceans, where they regulate cardiac function. These peptides are also released into the hemolymph and modulate exoskeletal muscle. Their mechanisms of action include activation of G-protein-coupled receptors, modulation of second messengers (cAMP and IP_3), and direct effects on ion channels (Na^+ and Ca^{2+} channels). Peptides generally have stronger effects on muscle than they do at the neuromuscular junction, but in several cases there is evidence for neuromuscular modulation.

Several members of this peptide group, including DRNFLRFamide (DF_2), increase the amplitude of EJPs recorded in crustacean muscle. In crayfish extensor muscle, DF_2 does not alter muscle-fiber input resistance nor does it have any effect on postsynaptic glutamate receptors, but it does increase the amount of transmitter released from presynaptic terminals. Elevation of either PKA or PKG activity by direct activation, or by inhibition of phosphodiesterases (which break down cyclic nucleotides), increases EJP amplitude, paralleling the effect of DF_2. Inhibition of kinases (PKA, PKC, PKG, or CaMKII) disrupts DF_2 modulation of EJPs. Thus, cyclic nucleotides and their protein kinases are required components of the neuropeptide modulation of transmitter release. Whether the effects of these second-messenger pathways occur directly on presynaptic mechanisms or via a retrograde messenger has not been established.

In the crab (*Eriphia*) closer muscle, DF_2 targets the fast (phasic) closer excitor but not the slow (tonic) closer excitor. This specific action, which enhances neurotransmitter release, is mediated by N-like calcium channels which are found almost exclusively at phasic nerve terminals.

Proctolin The pentapeptide proctolin was first identified in the hindgut of the cockroach (*Periplaneta americana*) and may be unique to arthropods. It can cause contractions in exoskeletal, cardiac, and hindgut muscles in insects and crustaceans. Proctolin acts as neurotransmitter, neuromodulator, and neurohormone.

Proctolin is co-localized with glutamate in three of the five excitatory motorneurons in the tonic flexor postural muscle of the crayfish. Proctolin has no effect by itself, but when released together with glutamate, muscle tension is increased. A major site of proctolin modulation is the muscle fiber membrane, where voltage-dependent calcium channels are activated. Proctolin, in general, enhances excitation–contraction coupling but has no observed presynaptic effect. However, in the leg closer muscle of the crab *Eriphia*, proctolin targets both the phasic and tonic excitor axon terminals. This presynaptic action is mediated by P/Q-like calcium channels.

PACAP-like neuropeptide Recent work on pituitary adenylyl cyclase-activating polypeptide-like neuropeptide is perhaps the most detailed study of the molecular pathways of invertebrate peptidergic modulation. During high-frequency activity the type 1b boutons of *Drosophila* larval muscle release PACAP-like neuropeptide. This leads to a biphasic modulation of postsynaptic responses – an immediate depolarization of the muscle followed by a hyperpolarizing voltage-activated potassium current. Activation of the adenylyl cyclase/cAMP pathway through $Gs\alpha$, and the Ras/Raf pathway through $G\beta\gamma$, is required for potassium

channel activation. Other proteins, such as neurofibromin 1 (NF1) and the Shaker potassium channel subunit, have also been implicated in mediating the modulatory effects of PACAP-like neuropeptide.

Conclusion

Invertebrate neuromuscular junctions are structurally closer to central synapses than to many vertebrate neuromuscular junctions, and have often been used as a convenient model for central neurotransmission. Invertebrate neuromuscular junctions, especially in mollusks, also share some features with vertebrate autonomic innervation of smooth muscles – varicose structure of nerve terminals, and often co-release of neuromodulators – but in most cases differ profoundly in synaptic organization and speed of response, due to the very close apposition of pre- and postsynaptic membranes. Arthropod neuromuscular junctions have been investigated more than those of other invertebrate groups; these junctions have a wide range of modulatory adjustments at their disposal. At the single-neuron level, various forms of activity-dependent short-term plasticity provide adjustments in synaptic transmission to meet the requirements for action on a moment-to-moment basis. Should the demands for altered performance be on a longer time scale, amounting to a change in the neuron's lifestyle, several forms of long-term adjustment are available, exemplified by experience-dependent synaptic strengthening in *Drosophila* and long-term adaptation in crustaceans. Neurohormonal mechanisms produce a more global change in performance, with neuromuscular modulation being only a part. The actions of large families of peptides have barely been explored at the neuromuscular level. Future work on many of the known neuromodulators will be directed to molecular aspects of cellular mechanisms, and, for this, *Drosophila* offers an excellent experimental system.

See also: Comparative Biology of Invertebrate Neuromuscular Junctions; Development of Drosophila Neuromuscular Junctions; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): Aging; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Postsynaptic Events in Neuromuscular Transmission; Presynaptic Events in Neuromuscular Transmission; Voltage-Gated Calcium Channels.

Further Reading

Atwood HL and Tse FW (1993) Physiological aspects of presynaptic inhibition. *Advances in Neural Science* 1: 19–65.

Atwood HL and Wojtowicz JM (1986) Short-term and long-term plasticity and physiological differentiation of crustacean motor synapses. *International Review of Neurobiology* 28: 275–362.

Beaumont V, Zhong N, Fletcher R, et al. (2001) Phosphorylation and local presynaptic protein synthesis in calcium- and calcineurin-dependent induction of crayfish long-term facilitation. *Neuron* 32: 489–501.

Glusman S and Kravitz EA (1982) The action of serotonin on excitatory nerve terminals in lobster nerve-muscle preparations. *Journal of Physiology* 325: 223–241.

Hoyle G (1983) *Muscles and Their Neural Control*. New York: Wiley.

Jorge-Rivera JC, Sen K, Birmingham JT, et al. (1998) Temporal dynamics of convergent modulation at a crustacean neuromuscular junction. *Journal of Neurophysiology* 80: 2559–2570.

Keshishian H, Broadie K, Chiba A, et al. (1996) The *Drosophila* neuromuscular junction: A model system for studying synaptic development and function. *Annual Review of Neuroscience* 19: 545–575.

Kravitz EA, Beltz B, Glusman S, et al. (1985) The well-modulated lobster. The roles of serotonin, octopamine, and proctolin in the lobster nervous system. In: Selverston AI (ed.) *Model Neural Networks and Behavior*, pp. 339–360. New York: Plenum Publishing Corporation.

Millar AG, Zucker RS, Ellis-Davies GCR, et al. (2005) Calcium sensitivity of neurotransmitter release differs at phasic and tonic synapses. *Journal of Neuroscience* 25: 3113–3125.

Rathmayer W and Djokaj S (2000) Presynaptic inhibition and the participation of $GABA_B$ receptors at neuromuscular junctions of the crab *Eriphia spinifrons*. *Journal of Comparative Physiology. A: Sensory, Neural and Behavioral Physiology* 186: 287–298.

Roeder T (1999) Octopamine in invertebrates. *Progress in Neurobiology* 59: 533–561.

Wang C and Zucker RS (1998) Regulation of synaptic vesicle recycling by calcium and serotonin. *Neuron* 21: 155–167.

Schuster CM (2006) Glutamatergic synapses of *Drosophila* neuromuscular junctions: A high-resolution model for the analysis of experience-dependent potentiation. *Cell Tissue Research* 326: 287–299.

Sigrist SJ, Reiff DF, Thiel PR, et al. (2003) Experience-dependent strengthening of *Drosophila* neuromuscular junctions. *Journal of Neuroscience* 23: 6546–6556.

Wiese K (ed.) (2002) *Crustacean Experimental Systems in Neurobiology*. Berlin: Springer.

Wiese K (ed.) (2002) *The Crustacean Nervous System*. Berlin: Springer.

Wu MN and Bellen HJ (1997) Genetic dissection of synaptic transmission in *Drosophila*. *Current Opinion in Neurobiology* 7: 624–630.

Zhong N, Beaumont V, and Zucker RS (2001) Roles for mitochondrial and reverse mode Na^+/Ca^{2+} exchange and the plasmalemma Ca^{2+} ATPase in post-tetanic potentiation at crayfish neuromuscular junctions. *Journal of Neuroscience* 21: 9598–9607.

Zucker RS and Regehr WG (2002) Short-term synaptic plasticity. *Annual Review of Physiology* 64: 355–405.

Neuron Doctrine: Historical Background

G M Shepherd, Yale University School of Medicine, New Haven, CT, USA

© 2009 Published by Elsevier Ltd.

Overview

As a new scientific discipline develops, concepts emerge that help to organize the diverse data into general principles. In experimental biology, a key organizing principle has been the cell, introduced for plants by Matthias Schlieden in 1838 and for animals by Theodor Schwann in 1839. The cell was established by an outpouring of microscopical studies in the 1840s as the basic principle of organization in most organs of the body, but acceptance of this role in the nervous system was delayed by the difficulty of visualizing under the light microscope the branches of cellular processes.

The impregnation method discovered by Camillo Golgi in 1873 finally enabled visualization of individual cells in the central nervous system. With this method, Golgi established a number of firsts. The branching dendrites end freely, an observation which disposed of previous theories of networks involving dendrites. Golgi recognized the different dendritic branching patterns as the basis for classifying different types of nerve cells. He extended Deiters' observation of a single axon arising from a motor neuron to show that this was also true of central neurons on the brain. He classified cells on the basis of their axons into cells of long axon (projecting out of a region to other regions) and cells of short axon (which came to be recognized as interneurons). The discovery of which he was most proud was the extensive collateral branches of the axon. In all these findings Golgi was the pioneer; all who followed him were building on his foundation, a fact probably not adequately recognized at the time and certainly not in our own time.

Beyond these fundamental findings, all confirmed by Santiago Ramón y Cajal and later histologists, Golgi made a further observation and fateful interpretation. In his stained preparations it appeared to him that the finest collaterals were continuous with each other. His interpretation was that these widely branching collaterals, and only these collaterals, form a reticulum which carries out the basic signaling functions of the nervous system.

The story of the discovery of the Golgi technique by an obscure histologist in Spain, Cajal, has been told many times, beginning with Cajal's own memoirs. In his first studies in the late 1880s, Cajal confirmed Golgi's findings that the dendrites end freely and that their different patterns are the basis of different types of nerve cells, and that the axon is usually single and that it characteristically gives off collaterals in different branching patterns. However, he also found persuasive evidence that the axon collaterals, too, end freely. This implied that the collaterals do not form a reticulum, but rather that communication between nerve cells is by contacts rather than by continuity.

Cajal's findings and interpretations were supported by most of the other classical histologists who took up the Golgi method. Cajal in particular worked furiously, producing an outpouring of papers on different cells throughout the nervous system during the 1890s. The consensus emerging from the early work was reviewed in 1891 by Waldeyer:

> [A]ll those authors who do not accept a network formation of the nerve fibers inside the central organ (Ramón y Cajal, Kolliker, His, Nansen, Lenhossek, Retzius) can easily unite points I and II above into one common brief fundamental law of wider range. It would sound thus: "The nervous system consists of numerous nerve units (neurons), anatomically and genetically independent." (Waldeyer, *Deutsche Medizinische Wochenschrift* 17, 1891)

This gave rise to the neuron theory, which soon became known as the neuron doctrine. As commonly understood, it states that each nerve cell is a separate individual, bounded, like all other cells in the body, by its plasma membrane. This conclusion was in line with the interpretation of the relations – by contact, not contiguity – between cells in other tissues and organs throughout the body. It is for this reason that the neuron doctrine was regarded as the extension of the cell theory to the nervous system. The points of contact between nerve cells were assumed to have special properties allowing for transfer of signals. Charles Sherrington, a physiologist, suggested the term 'synapse' for these contacts, and was the first to envisage special functional properties for them.

In addition to his insistence on the individuality of the entire nerve cell and its interactions through contact, Cajal made two other conceptual contributions that were essential for understanding the functional significance of the neuron. The first had to do with the flow of information through the nerve cell. Waldeyer had left this question uncertain. In further work in the early 1890s, Cajal, together with Alfred van Gehuchten of Belgium, inferred that there was a polarization of the flow of activity, from contacts on the dendrites and soma to the axon and out of the axon terminals. He termed this the 'law of dynamic

polarization.' This was a fundamental contribution which gave a functional meaning to the nerve cell as an integrative entity, taking in inputs from different sources and distributing them to different target cells.

The second conceptual contribution built on the first. By applying his tremendous powers of imagination to the functional interpretation of the relations between his stained cells, he deduced the flow of information not only within each cell but also between the cells. He devised diagrams showing how these relations formed the pathways within a region to accomplish multineuronal functions of integrating inputs from different regional sources and distributing them to different target regions. This enabled him to represent the flow of activity through the cells of the olfactory bulb, for example, and compare it with an equivalent flow of activity through the retina. This approach provided persuasive evidence that the nervous system is built of interconnected nerve cells that form the circuits and pathways that carry out the functional operations of the different regions and systems.

Many authors have attested to the importance of the original neuron doctrine and these functional concepts for the rise of modern neuroscience. For example, David Hubel observed that Cajal's textbook of histology of the nervous system, replete with the evidence of nerve cells as separate entities, "is considered to be the most important work ever published in neurobiology." On the physiological side, Sherrington made clear the central importance of the extension of the cell theory to the nervous system. In his great work *The Integrative Activity of the Nervous System*, published in 1906 just after Cajal's *Histologia* in Spanish, the first sentence reads: "Nowhere in physiology does the cell-theory reveal its presence more frequently in the very framework of the argument than at the present time in the study of nervous reactions."

In its simplified set of postulates the neuron doctrine thus provided the conceptual basis for the rise of modern neuroscience in understanding normal brain function, and the rise of the clinical neurological disciplines in understanding neurological diseases. In this respect it can be said to have provided the fundamental organizing principle for modern cellular neuroscience, performing a role similar to that of the Bohr atom for modern physics, the chemical bond for chemistry, the cell for cell biology, and the gene for genetics and molecular biology.

Modern Research First Confirmed, Then Challenged the Classical Neuron Doctrine

Cajal's last publication, in 1934 just before his death, was an exhaustive review of the evidence for the neuron theory against the reticular theory. By then the neuron doctrine had narrowed to one key point: "that the expansions of nerve cells end freely in the gray substance just as they do in the peripheral sense organs."

The advent of modern methods of cell biology in the 1950s, particularly the electron microscope for the fine structure of cells and the recording microeletrode for membrane and cell physiology, brought both support for and challenges to the classical doctrine. When the first electron microscopical investigations were carried out on nerve cell synapses, they focused on whether, at sites of synapses, the plasma membrane was continuous or not. The initial finding of the continuity of the plasma membrane at the chemical synapse appeared to validate the concept of synaptic transmission by contact and not continuity, and the neuron theory was widely assumed to have been validated. However, the discovery soon after of electrical synapses, and their subsequent identification with gap junctions, showed that in fact cells could also be interconnected by continuity as well. It is amusing, and perhaps a little sobering, to imagine that if electrical synapses had been discovered first in electrophysiological recordings or if gap junctions had been seen first under the electron microscope, it would have been taken as initial evidence to vindicate Golgi and the reticular theory over Cajal and the neuron theory.

Chemical versus Electrical Transmission

As the importance of gap junctions and electrical synaptic functions has been increasingly recognized, especially in the mammalian nervous system, it has seemed to some that it supports Golgi's concept of the reticular nature of nervous organization over Cajal's neuron theory. However, it should be clear from the remarks thus far that the situation is not that simple. Let us discuss the reasons.

First, Golgi restricted his continuous network to the finest branches of axon collaterals. Any electrical synapses or gap junctions between dendrites or between axons and dendrites would have been outside his theory. Therefore, the ubiquity of gap junctions and electrical synapses does not prove Golgi right; in fact, it proves him wrong, just as much as it does Cajal. Second, there has been evidence over many years for the action at a distance of different types of neuroactive substances, both classical neurotransmitters and neurohormones. This type of action by diffusion through the extracellular space to distant targets is inconsistent with both Golgi's network of continuous branches and Cajal's transmission by contact. Third, it has been claimed that the tiny branches of Golgi's reticulum would be a substrate for plastic

changes underlying learning, against the rigid connectivity of Cajal's contacts. However, Cajal always thought in terms of function, and assumed that learning would take place at contacts. This was explicit in Manouelian's suggestion of plasticity at contacts, a concept brought forward already in 1897.

Finally (though other points could also be made), the ubiquity of gap junctions is not a recent discovery. It has been recognized for over 30 years that gap junctions are present throughout the body, including the brain. In the body, this has not been taken as evidence against the cell theory; on the contrary, the cell theory has continued to be the basis for cell biology, and has incorporated the molecular biology of the gene as well. It may be that a radical new view of the cell as an anatomical, physiological, and developmental unit may emerge, and as it does, the concept of the nerve cell will undoubtedly evolve with it. In fact, it may even be leading the way, as we shall discuss next.

The Neuron: From Simple to Complex Functional Unit

Cajal's functional concepts of the neuron also come in for criticism. Given that Cajal made all of his interpretations on the basis of only the histology of the neurons, more than a half century before any experiments could be carried out on cells in the brain, it is not surprising that the physiological evidence would give new insights. Theodore Bullock (1959), in reviewing the early electrophysiological studies in the 1950s, argued that these results showed that rather than functioning as a single element, the nerve cell has different properties in its dendrites and axons, so that its function is characterized by 'complexity-within-unity.'

In our own studies we have focused on the mammalian olfactory bulb as a model for probing the properties first revealed in the spinal motor neuron. Using the new approach pioneered by Wilfrid Rall, we combined fine structure, electrophysiology, and computational models of the neurons and their dendrites. We found that the dendrites of mitral and granule cells interact directly with each other through dendrodendritic synapses, providing a pathway for feedback and lateral inhibition of the mitral cells. In this circuit, an action potential in the mitral cell dendrite causes synaptic excitation of granule cell dendritic spines; the excitatory postsynaptic potential (EPSP) in the spines then causes feedback synaptic inhibition of the mitral cell dendrite, and lateral inhibition of neighboring mitral cell dendrites through electrotonic spread. The combined study enabled the computational model to predict the interactions, the fine structure to confirm the synaptic circuit, and the physiological properties to be correlated with the fine structure.

These properties were inconsistent with the functional postulates of the neuron doctrine. Instead, several new principles were exemplified in this circuit. The direction of action potential propagation in the dendrites may be forward or backward. Dendrites can have specific synaptic outputs as well as inputs. The synapses may be activated by action potentials (mitral cells) or by local synaptic potentials (granule cells). The circuit can mediate feedback as well as lateral inhibition. A local part of a neuron (e.g., dendritic spine) can have input–output functions that are semi-independent of the rest of the cell. These local parts could be of differing extents, from a single synapse up to a dendritic branch. These properties have been repeatedly tested and extended and compared with equivalent properties in the retina, thalamus, and many other parts of the nervous system.

A New Theoretical Framework for Brain Organization: Multilevel Functional Units

Each of the parts of a neuron was thus seen to operate as a 'functional unit.' The key was that a functional unit at one level of organization (e.g., a synapse) provided the ensemble that gave rise to the functional unit at the next level (e.g., a synaptic pattern on a dendritic branch). This suggested that an enlarged hierarchical framework was needed for understanding the functional units in the nervous system and how they related to each other:

1. The nervous system is organized in terms of functional units. A functional unit is defined as the morphological substrate for a specific function.
2. Neurons (and functionally related cells) provide the morphological substrate for functional organization. Neurons have different combinations of morphological processes and physiological properties. For a given process of a given neuron, the synaptic position and physiological property depend on the integrative context within the functional units of which it is a part.
3. Functional units are formed at several levels of organization: synaptic terminals, dendritic trees, and multineuronal interconnections. The nervous system is built of overlapping assemblies and hierarchies of such units of increasing extent and complexity.

In the intervening years evidence has been obtained for functional units at different levels of organization in many regions. To review this work would take us deep into current research across a wide range of disciplines in neuroscience. It will suffice to note that a hierarchical framework is eventually necessary

for organizing this evidence, both for the organization of the individual neuron and for the organization of multineuronal circuits.

As an example of the complexity-within-unity of the single neuron, consider that the integrative functions of single dendritic branches form the functional units that are the basis for the integrative functions of multiple branches, which are the basis for the integrative functions of that part of a dendritic tree, which in turn are the basis for the integration in the entire tree. The properties at each of these levels need to be clearly related to their particular level, in order to be able to coordinate the functions in a bottom-up or top-down direction.

A similar problem arises in characterizing multineuronal assemblies. Here the problem is the apparent variety of circuits within any given region. An approach to reducing this complexity is to identify first a minimum pattern of connections and interactions that serves as the backbone, so to speak, of the circuits of a region, termed a 'basic circuit.' The concept is very close to the concept of the basic seven-transmembrane domain structure of G-protein receptors, a 'backbone' which is adapted by local diversification within the binding pocket for different neurotransmitter ligands and different G-protein effectors. It was further suggested that a 'microcircuit is a specific pattern of synaptic interaction within the basic circuit – in other words, at a more local organizational level. An example would be the specific interactions between olfactory mitral and granule cells, within the overall multilayer basic circuit of the olfactory bulb.

Recently, work on the cerebral cortex has focused on a 'canonical microcircuit' underlying cortical organization, the term being equivalent to the concept of a basic circuit. It conveys more strongly the idea of a stereotyped backbone that is adapted in different cortical regions by variations in local synaptic interactions. A key question currently under investigation is whether there is a canonical microcircuit (i.e., 'basic circuit') for all of cerebral cortex. Answering this question will require being specific about the level of organization. For example, one can hypothesize that there is a 'canonical' circuit for the entire cerebral cortex (equivalent to the seven-transmembrane structure of all G-protein-coupled receptors), that is adapted in different cortical areas by local variations in synaptic connections (binding pocket diversity in G-protein-coupled receptors).

These considerations indicate how important it will be to analyze experimental studies of circuits in different brain regions within a theory of multilevel organization. For the case of the cerebral cortex, it may be hypothesized that there will be specific stereotyped microcircuits identified at several levels of organization that, working in coordination, carry out the information processing functions of an overall canonical microcircuit.

Summary

In conclusion, familiarity with the history of the neuron doctrine is interesting not only for a perspective on the rise of modern neuroscience, but also for the challenges that lie ahead. Almost from the moment that the new methods appeared a half-century ago to support the original neuron doctrine, it began to be clear that it was inadequate to account for the extraordinary complexity of the individual neuron and of multineuronal circuits. A new theoretical framework is needed within which to organize the tremendous amounts of data that modern neuroscience is producing. The effort appears to be essential if the complexity of neuronal organization underlying behavior is to be understood. Whether having this new framework will play as important a role in supporting research in the future as the role that the neuron doctrine has played in the past is for the future to decide. This gives a new perspective on the magnitude of the achievement of Cajal and his contemporaries. Thus, in judging the classical neuron doctrine, the question is not whether it can stand up to our criticism today; the question is rather, can we produce a conceptual framework that will equal theirs in practical value, and serve as well in guiding experiments in the century ahead as theirs did for the century that is past?

See also: Brain Development: The Generation of Large Brains; Brain Evolution: The Radiator Theory; Brain Scaling Laws; Cajal's Place in the History of Neuroscience; Comparative Neurobiology: History; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; History of Neuroscience: Early Neuroscience; Neurophysiology: Past and Present.

Further Reading

Bennett MV, Aljure E, Nakajima Y, et al. (1963) Electrotonic junctions between teleost spinal neurons: Electrophysiology and ultrastructure. *Science* 141: 262–264.

Bullock TH (1959) Neuron doctrine and electrophysiology. *Science* 129: 997–1002.

Bullock TH, Bennett MV, Johnston D, et al. (2005) Neuroscience. The neuron doctrine, redux. *Science* 310: 791–793.

de Robertis E and Franchi CM (1956) Electron microscope observations on synaptic vesicles in synapses of the retinal rods and cones. *Journal of Biophysical and Biochemical Cytology* 2: 307–318.

de Robertis E (1959) Submicroscopic morphology of the synapse. *International Review of Cytology* 8: 61–96.

de Felipe J, Alonso-Nanclares L, and Arellano JI (2002) Microstructure of the neocortex: Comparative aspects. *Journal of Neurocytology* 31: 299–316.

Furshpan EJ and Potter DD (1959) Transmission at the giant motor synapses of the crayfish. *Journal of Physiology* 145: 289–325.

Loewenstein WR (1981) Junctional intercellular communication: The cell-to-cell membrane channel. *Physiological Reviews* 61: 829–913.

McCulloch WS and Pitts W (1990) A logical calculus of the ideas immanent in nervous activity. 1943. *Bulletin of Mathematical Biology* 52: 99–115; discussion 73–97.

Palay SL (1956) Synapses in the central nervous system. *Journal of Biophysical and Biochemical Cytology* 2: 193–202.

Rall W, Shepherd GM, Reese TS, et al. (1966) Dendrodendritic synaptic pathway for inhibition in the olfactory bulb. *Experimental Neurology* 14: 44–56.

Rall W and Shepherd GM (1968) Theoretical reconstruction of field potentials and dendrodendritic synaptic interactions in olfactory bulb. *Journal of Neurophysiology* 31: 884–915.

Ramón y Cajal S (1937) *Recollections of my Life*. (trans.) Craigie EH and Cano J. Philadelphia: The American Philosophical Society.

Ramón y Cajal S (1953) *Histologie du Système Nerveux de l'Homme et des Vertébrés*. Madrid: Consejo Superior de Investigaciones Científicas, Instituto Ramón y Cajal.

Revel JP and Karnovsky MJ (1967) Hexagonal array of subunits in intercellular junctions of the mouse heart and liver. *Journal of Cell Biology* 33: C7–C12.

Shepherd GM (1972) The neuron doctrine: A revision of functional concepts. *Yale Journal of Biology and Medicine* 45: 584–599.

Shepherd GM (1978) Microcircuits in the nervous system. *Scientific American* 238: 93–103.

Shepherd GM (1991) *Foundations of the Neuron Doctrine*. New York: Oxford University Press.

Shepherd GM (2004) *The Synaptic Organization of the Brain*. Oxford; NewYork: Oxford University Press.

Sherrington CS (1906) *The Integrative Action of the Nervous System*. New Haven, CT: Yale University Press.

Neuronal Angiotensin

A M Allen, E L O'Callaghan, F A O Mendelsohn, and S-Y Chai, University of Melbourne, Melbourne, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Overview of Systemic Angiotensin Production

Our understanding of the renin–angiotensin system began in the late 1800s with the discovery of a pressor agent, called renin, derived from the kidney. Further work throughout the 1900s defined a humoral system in which the active agent, angiotensin (Ang II), was derived from a protein precursor, angiotensinogen (Aogen), via the sequential action of two enzymes, renin and angiotensin converting enzyme (ACE) (Figure 1). In this humoral system, Aogen is produced by the liver and secreted into the circulation. Renin, an aspartyl protease, is secreted into the plasma from cells of the juxtaglomerular apparatus in the kidney. Renin's only known substrate is Aogen, and the activity of this enzymatic reaction is considered to be the rate-limiting step in the pathway. Cleavage of Aogen yields a large glycoprotein of no known function and an inactive decapeptide, Ang I. Cleavage of Ang I by the zinc-dependent carboxypeptidase, ACE, produces the active octapeptide, Ang II. ACE is located on endothelial cells throughout the vascular system, with a particularly high concentration in the lung. Subsequent N-terminal cleavage of Ang II produces the heptapeptide Ang III, which is equipotent at Ang II receptors and may be an important ligand in some tissues.

Receptors for Ang II were cloned in the 1990s and shown to be G-protein-coupled receptors with seven transmembrane spanning domains. Two receptors, with approximately 30% homology, were cloned and termed AT_1 and AT_2 receptors. The AT_1 receptor, which mediates most of the classical actions of Ang II, was shown to occur in two isoforms in rodents, AT_{1A} and AT_{1B}. Although the number of intracellular signaling pathways employed by the AT_1 receptor continue to expand, many of the principal actions involve coupling to the $\alpha_{q/11}$ G-protein subunit. Activation results in cleavage of phosphatidyl-inositol (4,5) bisphosphate, by phospholipase C_β, to yield diacylglycerol and inositol trisphosphate with subsequent increases in intracellular calcium concentrations and activation of protein kinase C. The angiotensin AT_2 receptor signals via activation of the protein phosphatases PP2A, SH2 domain-containing phosphatase, and mitogen-activated protein kinase phosphatase 1, resulting in the dephosphorylation and inactivation of kinases including growth factor tyrosine kinase receptors and extracellular signal-regulated kinase 1 (ERK-1) and 2 (ERK-2). Antiproliferative and proapoptotic actions mediated by the AT_2 receptor appear to involve the inactivation of ERK-1 and -2 by dephosphorylation. The binding of Ang II to the AT_2 receptor has also been shown to activate nitric oxide and phospholipase A2 in certain tissues. Investigation into the signaling pathways associated with AT_2 receptor activation is further complicated by the complex physiological roles of the receptor, often in pathological states.

Studies have indicated the involvement of other enzymes and receptors in the activity of the renin–angiotensin system (Figure 1). Angiotensin converting enzyme 2 (ACE2) was cloned and found to exhibit 40% homology with ACE but with one catalytic domain. In contrast to ACE, ACE2 cleaves single amino acid residues from the C-terminus of peptides. ACE2 is thought to have a counterregulatory role in the renin–angiotensin system, being involved in the degradation of Ang II and the generation of Ang 1–7, an effector peptide that often shows opposing actions to Ang II. The receptor for Ang 1–7 was found to be the 'orphan' G-protein-coupled receptor that is encoded by the proto-oncogene *mas*. Mice with a deletion of the *mas* proto-oncogene are not responsive to Ang 1–7 and lose the binding site for Ang 1–7 in the kidney.

ACE2 is also the receptor for some coronaviruses, including the virus that causes severe acute respiratory syndrome (SARS-CoV) and another virus associated with respiratory tract infections, HCoV-NL63. ACE2 facilitates viral entry into and infection of target cells by binding specifically and with high affinity to the S protein of the viruses.

Angiotensin IV is generated from Ang III by the cleavage of the N-terminal arginine by aminopeptidase N. Angiotensin IV was initially thought to be biologically inactive because it exhibited very low affinity for the AT_1 and AT_2 receptors. However, injection of Ang IV into the cerebral ventricles enhances performance in memory tasks, and it was shown to bind to a specific, high-affinity site termed the angiotensin AT_4 receptor. Subsequent work identified the AT_4 receptor as an enzyme, insulin-regulated amino peptidase.

Tissue Renin–Angiotensin Systems

In the early 1970s, evidence of local formation of angiotensin peptides in specific tissues began to

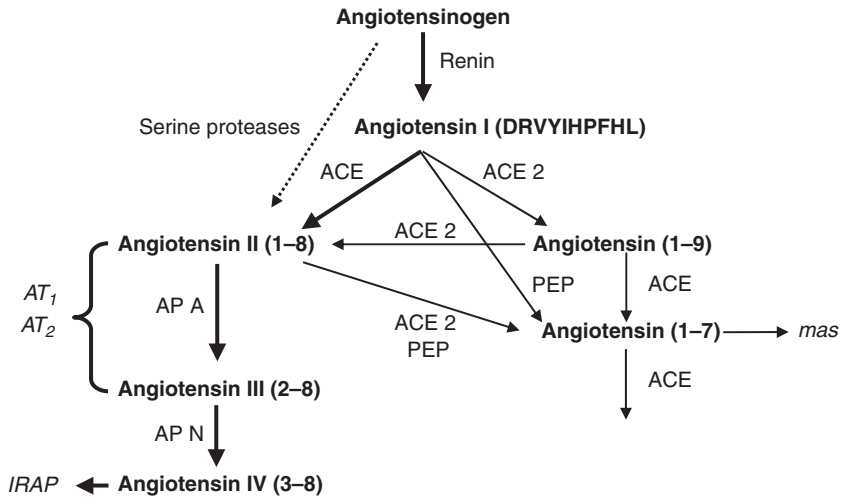

Figure 1 Overview of the enzymes and receptors involved in the production and function of angiotensin peptides. The bold arrows indicate the pathway traditionally ascribed to this system. Proteins and peptides are indicated in bold type. Enzymes are shown in normal type, and receptors are shown in italics. ACE, angiotensin converting enzyme; AP, aminopeptidase; IRAP, insulin-regulated aminopeptidase; PEP, prolyl endopeptidase.

emerge. It has since been demonstrated that many tissues, including kidney, heart, blood vessels, adrenal gland, uterus, testes, and brain, have the potential to produce Ang independently of the circulating renin–angiotensin system. Although the exact mechanism of formation remains controversial for many of these tissues, the existence of independently regulated tissue renin–angiotensin systems is widely accepted. As discussed in the following sections, there is also considerable evidence for the local generation of Ang peptides in the brain which act on receptors within the blood–brain barrier to regulate a variety of functions.

Angiotensinogen in Brain

Angiotensinogen is expressed in relatively high levels by the brain, with mRNA levels reported to be approximately one-third those produced by the liver. Measurements of Aogen concentrations in cerebrospinal fluid by radioimmunoassay indicate that it represents 2% or 3% of the total protein content. This Aogen is identical in amino acid structure to that produced by liver but may be differentially glycosylated. Most Aogen mRNA expression in the brain occurs in ependymal and astroglial cells. Interestingly, there is regional heterogeneity of expression, with highest levels in hypothalamic, midbrain, and medullary regions that are associated with the known actions of Ang peptides. Transgenic mouse lines with reporter protein expression under the control of the Aogen promoter demonstrate neuronal expression of Aogen but only in a very small number of cell groups. Most evidence indicates that Aogen synthesized by glia is constitutively secreted into the extracellular space; however, studies have demonstrated that some Aogen might be directed to the nucleus. The relative proportion of intracellular versus secreted Aogen existing *in vivo* and the role of the intracellular protein have not been examined.

Transgenic rat and mouse models demonstrate the functional importance of brain-derived Aogen. Transgenic rats which produce an Aogen antisense under the regulation of a glial-specific promoter show multiple functional deficits, including decreased blood pressure and impaired fluid and electrolyte balance. Evidence for the converse is available in mice, in which global overexpression of renin and Aogen leads to increased blood pressure that can be reduced by glial-specific ablation of Aogen production. Together, these observations indicate that brain Aogen plays an important role in the regulation of blood pressure, but they do not shed light on how the Aogen is processed to yield Ang II and other active components of the system.

The presence of high concentrations of Aogen in astrocytes may also be suggestive of wider roles for the brain renin–angiotensin system. Investigation of functional alterations in the Aogen knockout mouse revealed impairment of blood–brain barrier function which can be restored by exogenous administration of Ang II or Ang IV. This blood–brain barrier dysfunction is thought to be due to the decreased expression of glial fibrillary acidic protein and laminin in the astrocytes of these animals.

Renin in the Brain

Efforts to demonstrate the existence and localization of renin in brain have proven extremely difficult.

Initial biochemical studies, which made every effort to remove contamination from systemic renin, demonstrated renin-like enzymatic activity in extracts of brain but did not establish the identity of renin. Several other enzymes, including cathepsin D and G, tonin, and kallikrein, have the capacity to form Ang peptides from renin substrate, and it remains a distinct possibility that enzymes other than renin might contribute to the generation of Ang peptides in brain. The use of renin antibodies in immunohistochemical and Western blot analysis to identify brain renin was also fraught with issues pertaining to antibody specificity. Initial attempts to demonstrate the existence of renin mRNA, using *in situ* hybridization or Northern blot analysis, also produced variable results, most likely due to the use of whole brain, which thus diluted any renin present in discrete brain regions to below the level of detectability. Using Northern blot hybridization methods, Dzau and colleagues found evidence for the existence of renin mRNA in brain with levels approximately 10% of those found in kidney.

In the kidney, renin is produced as preprorenin that is sequentially cleaved to form active renin. The renin gene has nine exons, the first of which encodes a signal peptide which targets the nascent protein to the endoplasmic reticulum for glycosylation and subsequent secretion from the cell. Exon 1 has three different transcriptional initiation points, which form different renin isoforms. Experiments have demonstrated two distinct renin isoforms formed by tissue-specific transcriptional initiation in brain. Initiation from exon 1a forms the classic 'renin a' found in the kidneys and systemic circulation, whereas the exon 1b transcription start point forms 'renin b.' This latter isoform lacks the coding for the signal peptide and is likely to be a constitutively active intracellular enzyme. These data raise the possibility of two distinct mechanisms of Ang peptide formation in brain – extracellular and intracellular.

However, these studies did not provide information on the distribution of renin or evidence of the cell type responsible for production. Other efforts have focused on the use of long lengths of renin promoter to drive expression of marker proteins in transgenic mice. Using a 4.1 kbp of the 5′ flanking sequence of the renin gene to drive expression of green fluorescent protein, Gross and colleagues demonstrated tissue-specific expression and regulation of the marker protein. Using this mouse model, Sigmund and colleagues localized renin expression in different regions of the mouse central nervous system. This expression occurs in many brain regions known to be involved in the regulation of fluid and electrolyte balance and autonomic control by Ang II and is predominantly localized to neurons.

Thus, current evidence supports the view that renin is expressed in the brain predominantly by neurons. The presence of renin isoforms indicates the potential for both intracellular and extracellular enzymatic activity.

Questions remain regarding the localization of renin in relation to the precursor, Aogen, and how these interact to produce Ang peptides in brain. The highest levels of renin expression occur in the cerebral and cerebellar cortices in regions where Aogen expression is low or absent. Observations in double transgenic mice, in which long sequences of the renin and Aogen promoters are used to drive expression of different reporter proteins, indicate that coexpression of renin and Aogen is rare; this is in line with observations that they are expressed in different cell types. However, adjacent cellular expression was observed in several nuclei, providing evidence to support the view that secreted Aogen might be cleaved locally in the extracellular space or taken into neighboring neurons for further processing to Ang I.

Angiotensin Converting Enzyme

Within the circulation, Ang I is rapidly converted to Ang II by the action of ACE, a zinc-dependent metallopeptidase that cleaves a dipeptide from the C-terminus of Ang I to form the effector peptide Ang II. Specific inhibitors of ACE are effective in inhibiting the formation of Ang II *in vivo* and attenuating the physiological effects of the peptide. Hence, these inhibitors have been developed very successfully into therapeutic drugs that are widely used to treat hypertension, heart failure, vascular disease, and renal disease. In addition to enzymatic cleavage of Ang I, ACE cleaves a range of other small peptide substrates, many of which are present in brain, including bradykinin, substance P, met-enkephalin, beta-endorphin, dynorphin 1–8, neurotensin, and luteinizing hormone releasing hormone.

Two isoforms of ACE exist in mammals and both are membrane bound, although a secreted form of ACE that is catalytically active is found in the circulation. The more abundant 140 kDa somatic enzyme occurs on the plasma membrane of endothelial and epithelial cells and has a large extracellular domain that contains two homologous regions, each with an active site. The testicular isoenzyme contains only one active site, which shares sequence homology with the C-terminal active site of the somatic enzyme and is the site that has been shown to selectively cleave Ang I.

In the many brain regions, ACE occurs in high concentrations. Many of these sites are also enriched with Ang II-containing cell bodies and fibers and Ang AT_1 receptors, including the circumventricular

organs, several hypothalamic nuclei, and the dorsal vagal complex in the medulla. It is proposed that at these sites, ACE participates in the local production of Ang II, either from circulating Ang I (as is the case for the circumventricular organs) or from brain-derived Ang I. However, ACE is much more abundantly distributed and high concentrations of the enzyme are detected in the basal ganglia, choroid plexus, hippocampus, cerebral cortex, and cerebellum, where its roles remain to be elucidated and may involve cleavage of other neuropeptide substrates. Although the predominant form of ACE in the brain appears to be membrane bound, a low level of the enzyme has been detected in cerebrospinal fluid, at a concentration 100-fold lower than that detected in serum.

Ang II/III Peptide Localization

Because they are cleavage products of a precursor that is secreted into the extracellular space, the cellular localization of Ang peptides in the brain has proven difficult. Immunohistochemistry has been problematic because of uncertain antibody specificity. Detailed maps of the distribution of Ang II-like immunoreactive cells and fibers have been produced but have mostly relied on results obtained with one antibody. Due to this scarcity of reliable antibodies, our knowledge of the neuronal circuitry that may utilize Ang II as a neurotransmitter is limited. However, in support of the veracity of the maps of Ang II distribution, there is a very high concordance in the distributions of Ang II-like immunoreactive nerve terminals and Ang II AT_1 receptors throughout the brain. There are some notable exceptions; for example, the magnocellular nuclei of the hypothalamus and central nucleus of the amygdala contain high levels of Ang II-like immunoreactive nerve terminals but low to undetectable concentrations of receptor.

Angiotensin AT_1 Receptors

As outlined previously, several controversies exist in relation to the production of Ang II in the brain. Some of the most convincing evidence for the existence of a brain renin–angiotensin system is derived from examination of the distribution of the Ang receptors and the physiological actions exerted through them. Unequivocal evidence points to the existence of Ang AT_1 and AT_2 receptor mRNA and binding sites (Figure 2) present in a characteristic distribution throughout the brain. These receptors show identical binding profiles to the peripheral receptor, including antagonism by the nonpeptide AT_1 and AT_2 receptor antagonists. In some cases, brain AT_1 and AT_2 receptors are located outside the blood–brain barrier, where they are accessible to systemic Ang peptides. However, the vast majority of sites are behind the blood–brain barrier, where they are presumably accessed by brain-derived Ang peptides. Administration of Ang peptides into these regions generally results in neuronal excitation and a characteristic range of site-dependent physiological responses. More important, in the context of establishing a role for brain-derived Ang peptides, pharmacological blockade of these receptors with specific and selective antagonists attenuates many physiological responses, such as drinking following dehydration or autonomic activation in response to mental stress. Together, these observations provide strong support for the existence of brain-derived Ang peptides that regulate physiological functions in the central nervous system.

There is an extensive distribution of Ang II AT_1 receptors in the brain. Moderate to high levels of AT_1 receptors occur in the sensory circumventricular organs (subfornical organ and vascular organ of the lamina terminalis and area postrema); piriform cortex; lateral septum; bed nucleus of the stria terminalis; median preoptic nucleus; periventricular, parvocellular paraventricular, and dorsomedial nuclei of the hypothalamus; locus coeruleus; lateral parabrachial nucleus; nucleus of the solitary tract; spinal nucleus of the trigeminal tract; and ventrolateral medulla. Most of these sites also possess high levels of Aogen-expressing astrocytes and ACE.

Actions via AT_1 Receptors

Electrophysiological studies have demonstrated that neurons in many of the sites listed previously are responsive to local application of Ang II. The predominant response is excitation via an AT_1 receptor-mediated inhibition of outward potassium channels. Microinjection studies with exogenous Ang II and/or AT_1 receptor antagonists have identified several actions of Ang II in the brain. In brief, these include the following:

- Alterations in autonomic activity to the cardiovascular system, leading to increases in blood pressure and heart rate, through direct stimulation of efferent pathways and regulation of sensory afferent (e.g., the baroreceptor reflex) information
- Increases in fluid and salt intake
- Stimulation of the secretion of several neuroendocrine hormones, including vasopressin, adrenocortictropic hormone, and luteinizing hormone releasing hormone
- Effects on learning and memory

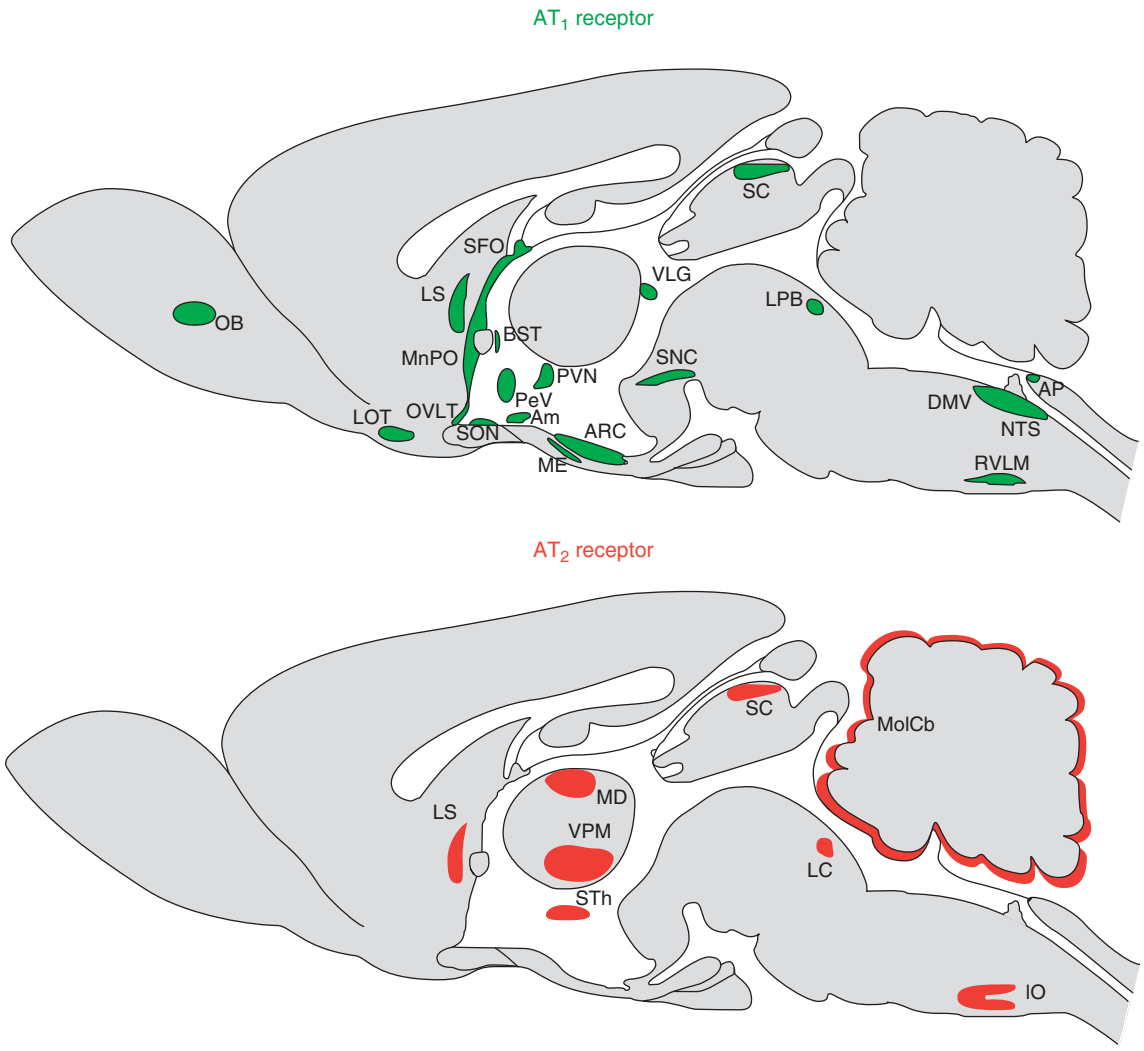

Figure 2 Schematic diagram representing the distribution of angiotensin AT_1 and AT_2 receptors in the brain. The distribution of receptors, determined by either *in vitro* autoradiography or *in situ* hybridization histochemistry, is depicted on midsagittal sections of the rat brain. Am, amygdala; AP, area postrema; ARC, arcuate nucleus; BST, bed nucleus of the stria terminalis; DMV, dorsal motor nucleus of the vagus; IO, inferior olivary nucleus; LC, locus coeruleus; LOT, nucleus of the lateral olfactory tract; LPB, lateral parabrachial nucleus; LS, lateral septum; MD, mediodorsal nucleus of the thalamus; ME, median eminence; MnPO, median preoptic nucleus; MolCb, molecular layer of the cerebellum; NTS, nucleus of the solitary tract; OB, olfactory bulb; OVLT, organum vasculosum of the lamina terminalis; PeV, periventricular nucleus of the hypothalamus; PVN, paraventricular nucleus of the hypothalamus; RVLM, rostral ventrolateral medulla; SC, superior colliculus; SFO, subfornical organ; SON, supraoptic nucleus; SNC, substantia nigra pars compacta; STh, subthalamic nucleus; VLG, ventrolateral geniculate nucleus; VPM, ventral posteromedial nucleus of the thalamus.

Angiotensin AT_2 Receptors in Brain

Angiotensin AT_2 receptors occur in high levels with a wide distribution in the brain during development. In the adult, this distribution becomes restricted to a few sites, including some thalamic nuclei, the cerebellum, and the inferior olivary nucleus. This distribution indicates a role in sensory information processing. The phenotype of the AT_2 receptor knockout mouse also points to some effects of Ang II on these receptors in the brain. These mice show reduced spontaneous movement and exploratory behavior, and there is evidence for effects of Ang II on cognitive function and memory. However, these are often contradictory in regard to potentiation or inhibition of memory formation. It has been proposed that the detrimental effects of Ang II on memory consolidation and retrieval are mediated by the AT_2 receptor. Further support for the angiotensin AT_2 receptor altering cognitive function comes from the detection of mutations in the coding region of this receptor in patients with mental retardation.

Ang 1–7 and Its Binding Site, the *mas* Oncogene

Angiotensin 1–7 may play a role in the central regulation of blood pressure because microinjection of Ang 1–7 into various sites alters arterial pressure and stimulates release of vasopressin. Ang 1–7 has also been reported to enhance long-term potentiation in the lateral nucleus of the amygdala and in the CA1 region of the hippocampus, suggesting a role in memory formation.

Ang IV and Its Binding Site, IRAP

Intracerebroventricular injection of Ang IV in rats enhances performance in several different memory tasks. A binding site that recognizes Ang IV with high affinity, but not Ang I, Ang II, AT_1, or AT_2 antagonists, was named the AT_4 receptor based on its distinct pharmacology. A surprising finding was the discovery that this receptor is the previously known insulin-regulated aminopeptidase (IRAP), a transmembrane zinc metallopeptidase known to traffic with the GLUT-4 glucose transporter in response to insulin. Ang IV, LVV-hemorphin-7, and other AT_4 ligands were all found to be competitive inhibitors of IRAP but were not cleaved to any appreciable extent by the enzyme. The mechanism by which Ang IV and other AT_4 ligands mediate their memory effects is yet to be resolved.

Conclusion

During the past several decades, it has been established that Ang peptides are generated in the brain. However, despite considerable effort, many questions regarding the biochemical pathways involved in the generation of this peptide system remain incompletely resolved. This system appears to have unique properties with, for example, precursor and processing enzymes existing in different cellular compartments and the potential for neurotransmitter, paracrine, and even intracrine effects. Yet despite these deficiencies in our knowledge, it is clear that Ang peptides play important physiological and pathological roles in the central nervous system regulation of a wide variety of functions, including those that complement its peripheral actions (e.g., autonomic regulation and fluid and electrolyte balance) and other more disparate functions, such as modulation of memory and blood–brain barrier formation.

See also: Angiotensin II; Angiotensin Actions on and Within Brain.

Further Reading

Allen AM (2007) Central neural regulation of cardiovascular function by angiotensin – A focus on the rostral ventrolateral medulla. *Neuroendocrinology* (in press).

Baltatu O and Bader M (2003) Brain renin–angiotensin system: Lessons from functional genomics. *Neuroendocrinology* 78: 253–259.

Chai SY, Fernando R, Peck G, et al. (2004) The angiotensin IV/AT4 receptor. *Cellular and Molecular Life Science* 61: 2728–2737.

Paul M, PoyanMehr A, and Kreutz R (2006) Physiology of local renin–angiotensin systems. *Physiological Reviews* 86: 747–803.

Sakai K and Sigmund CD (2005) Molecular evidence of tissue renin–angiotensin systems: A focus on the brain. *Current Hypertension Reports* 7: 135–140.

Santos RA and Ferreira AJ (2007) Angiotensin-(1–7) and the renin–angiotensin system. *Current Opinion in Nephrology and Hypertension* 16: 122–128.

Neuronal Motility and Structure: Cdk5 Pathways

B A Samuels and L-H Tsai, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction to Cdk5

Cdk5 and the Cyclin-Dependent Kinase Family

Cdk5 is an atypical cyclin-dependent kinase Cyclin-dependent kinase 5 (Cdk5) was first isolated based on its homology to other Cdks, specifically human cell division control kinase 2 (Cdc2/Cdk1). The typical role of a Cdk is to phosphorylate substrates that drive progression through the cell cycle. Cdc2, for example, is the catalytic subunit of the maturation promoting factor (MPF) that is necessary for the G2/M phase transition. Cdk5, however, is an atypical Cdk. There is no strong evidence that Cdk5 is functionally involved in promoting the cell cycle. In fact, the principal associated kinase activity of Cdk5 is detected in the central nervous system, where the protein is enriched in postmitotic neurons. Although this signifies a functional distinction from other Cdks, Cdk5 does share a few very basic characteristics with members of the Cdk family that are important for understanding its various roles.

p35 and p39 are essential activators of Cdk5 Monomeric forms of Cdks do not display any kinase activity; instead, association with an activator is required. Most Cdks associate with and are activated by cyclins, a family of unstable proteins whose expression level during the cell cycle is phase dependent. For example, MPF is achieved through the association of Cdc2 with cyclin B, a protein whose expression rises through the cell cycle until M phase, when it is abruptly degraded. Such regulation of activator expression and turnover is a major mechanism underlying the regulation of Cdk activity. Cdk5 activity is achieved through association with one of two activators, p35 or p39. Although p35 and p39 are very similar to each other, they have little sequence similarity to cyclins. However, the tertiary structure of a portion of p35 bears some resemblance in that it adopts a cyclin-like fold. Another similarity to cyclins is that p35 is very unstable. Ubiquitination of p35 is readily detectable, and the protein is rapidly targeted for degradation by the proteasome, retaining only a 20–30 min half-life in primary cultured cortical neurons.

Posttranslational modifications of Cdk5 In addition to the turnover rate of the activators, Cdk5 activity, like that of other Cdks, can be regulated by posttranslational modifications such as phosphorylation. Phosphorylation of Cdk5 on tyrosine 15 enhances Cdk5 activity. Cdk5 associates with the c-Abelson tyrosine kinase (c-Abl) through a linker protein termed Cables, and this association stimulates phosphorylation of Cdk5 on Y15 in a c-Abl-dependent manner. Cdk5 may also be phosphorylated on serine 159 by casein kinase I. Structural studies suggest that this phosphorylation event would inhibit Cdk5 activity by preventing association with its activators.

Consensus site for phosphorylation by Cdk5 The only other similarity to the Cdk family is that Cdk5 shares the consensus phosphorylation site on its substrates. Cyclin-dependent kinases are proline-directed serine/threonine kinases, phosphorylating substrates at the minimal consensus motif S/TP. Cdc2, Cdk2, and Cdk5 also prefer a basic residue in the +3 position, making the consensus sequence for phosphorylation (S/T)PX(K/H/R).

Cdk5 and Its Activators

Cdk5 activity is defined by the distribution of p35 and p39 Whereas Cdk5 protein is enriched in postmitotic neurons, its expression is widespread in nonneuronal tissues. Importantly, Cdk5 activity is primarily restricted to the nervous system because p35 and p39 are detected at high levels exclusively in neurons. The intracellular distribution of the activators is likely also important for regulating Cdk5 activity and restricting the kinase to its physiological substrates. In this regard, p35 contains an N-terminal myristoylation site that helps target kinase activity to membrane fractions. Cdk5 and its activators are also distributed to detergent-insoluble cytoskeleton fractions and both pre- and postsynaptic compartments. The spatial distribution pattern of Cdk5 activity defines its chief physiological roles, which seem to be regulating cytoskeletal elements during developmental neuronal migration and synaptic function in the mature nervous system. A diverse group of identified Cdk5 substrates have corroborated these functions and also have implicated altered regulation of the kinase in the pathology of neurodegeneration.

Cdk5 Regulates Neuronal Migration

Mouse Models of Cdk5 Deficiency

Cerebral cortex The best-understood role of Cdk5 is in the regulation of the cytoarchitecture of the central nervous system. The characteristic phenotype of mice deficient in Cdk5 activity is an inverted laminar organization of pyramidal neurons in the cerebral cortex. Mice deficient in Cdk5 and double knockouts of p35 and p39 display indistinguishable phenotypes, suggesting that p35 and p39 are the only activators of Cdk5 during development. p35 knockout mice display a similar but less severe phenotype and survive into adulthood.

Other brain regions In addition to the inverted neocortex, mice lacking Cdk5 activity display widespread cytoarchitectural abnormalities in many brain regions. The pyramidal neurons of the cornus ammonus in the hippocampus form only diffuse laminae with cell-free rifts between clusters of cells. Furthermore, the dentate gyrus of the hippocampus is imperceptible in Cdk5-deficient mice. The cerebellum of Cdk5-deficient mice is also strikingly abnormal because it completely lacks foliation (the several folds or gyri normally seen on the cerebellar surface), is much smaller in size, and does not have a Purkinje cell layer. There are also major disruptions in the olfactory bulb, thalamus, and brain stem. Furthermore, motor neurons exhibit ballooned perikarya suggestive of chromatolysis, where the cell body swells, Nissl bodies disperse, and the nucleus becomes displaced peripherally in the cell. Chromatolysis, also known as the 'axonal reaction,' is commonly seen upon damage or lesion to an axon.

Brief Introduction to Neocortical Development

Different neurons in the neocortex come from distinct places Pyramidal neurons and interneurons are the principal types of neurons comprising the neocortex, but these different classes of neurons take distinct paths to their final destination. The pyramidal neurons are generated in the dorsal ventricular zone and migrate in a radial manner toward the pial surface before reaching their final cortical position. Interneurons, on the other hand, are produced in the ventral telencephalon and must migrate dorsally by taking a tangential route. Consequently, the distance traversed by the tangentially migrating interneurons is much longer than that of the radially migrating pyramidal neurons (Figure 1).

Formation of cortical laminae Classic studies of corticogenesis have shown that the earliest born cortical neurons align above the ventricular zone to form the preplate, which signifies the first distinct event in the histogenesis of the neocortex. The next group of neurons born in the ventricular zone migrates radially to split the preplate into the superficial marginal zone and the deeper subplate, forming the cortical plate in between. The assembly of the cortical plate proceeds in an 'inside-out' manner, in which each cohort of neurons migrates past its cortical plate predecessors to form a more superficial layer (closer to the eventual pial surface). Ultimately, the six cortical laminae can

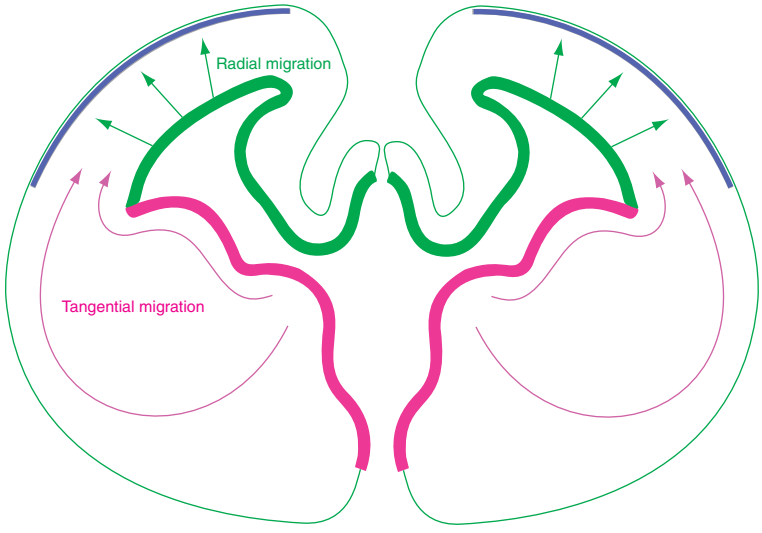

Figure 1 Two distinct routes of neuronal migration. Pyramidal neurons are produced along the ventricular surface in the dorsal telencephalon (heavy green band) and migrate in a radial direction (green arrow) to the outermost layer of the developing cortex (heavy blue band). Interneurons are produced in the ventral telencephalon (light red band) and migrate along a much longer route in a tangential manner (red arrow).

be distinguished by the earliest born neurons residing in the deepest layers of the neocortex and the latest born neurons forming the most superficial layer (Figure 2).

Cdk5 Activity Is Essential for Proper Neocortical Lamination

Formation of cortical laminae in p35 knockout mice Neuronal birth dating and histological experiments demonstrated that in p35-deficient mice early born neurons are capable of splitting the preplate. However, unlike normal inside-out cortical plate assembly, each subsequent cohort of neurons in p35-deficient mice is unable to migrate past its predecessors. Therefore, without proper Cdk5 activity, corticogenesis proceeds in an erroneous 'outside-in' manner, resulting in an inverted laminar structure.

Reelin Signaling Also Regulates Neuronal Migration

The phenotype of the mutant mice has cemented the Cdk5 pathway as a major regulator of neuronal migration during development of the central nervous system. However, to fully grasp the importance of Cdk5 in brain development, it is necessary to also consider another classical neuronal migration pathway that involves Reelin signaling. Upon gross inspection, there are many similarities in the abnormal neuronal cytoarchitecture between Cdk5 and Reelin mutant mice.

Mouse Models of Reelin Deficiency

Multiple mutant mice deficient in Reelin signaling have been described, including naturally occurring mutants such as *reeler*, *scrambler*, and *yotari*, as well as genetically modified mice such as Dab1, VLDLR, and ApoER2. In *reeler* mice, first described approximately 50 years ago for their ataxic 'reeling' gait, the early born neurons intended to form the cortical plate fail to split the preplate. Subsequent waves of migrating neurons are unable to pass their predecessors and accumulate in gradually deeper positions. Therefore, similar to Cdk5 mutants, *reeler* mice have an inverted laminar organization of the neocortex. The defective migration in these mice is attributable to a deficiency in the protein Reelin, a very large secreted extracellular matrix protein. The other naturally occurring or genetically modified mice with identical phenotypes have defined a pathway in which Reelin binds to VLDLR and ApoER2 receptors on the surface of migrating cells, which triggers a signaling cascade mediated by tyrosine phosphorylation of Disabled-1 (Dab1), the gene that is mutated in *scrambler* and *yotari* mice.

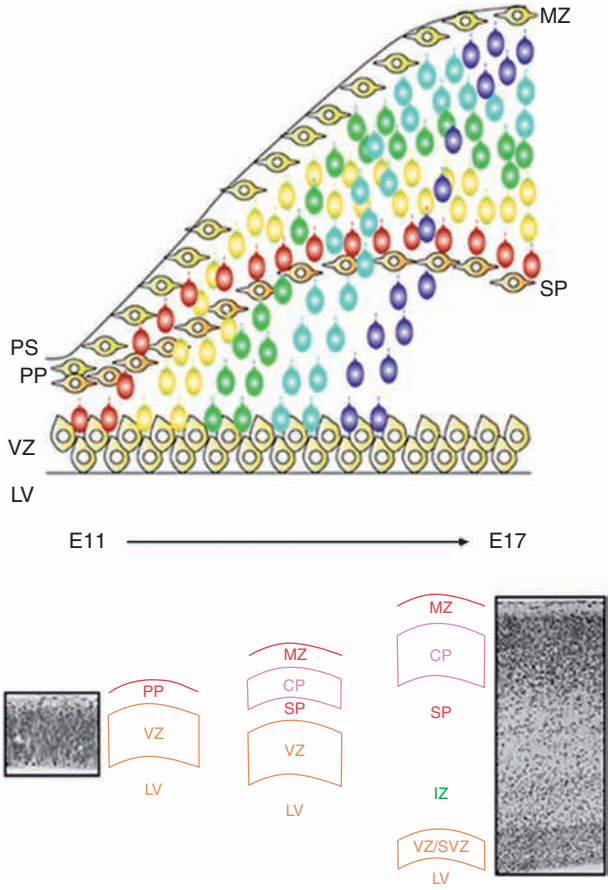

Figure 2 Development of the mouse cerebral cortex via radial migration from E11 to E17. (Top) Corticogenesis proceeds in an 'inside-out' manner. Neuronal progenitors are present in the ventricular zone (VZ) along the surface of the lateral ventricle (LV). The earliest born neurons align above the VZ and form the preplate at the young pial surface (PS). The next wave of neurons (red) to be born splits the preplate into the marginal zone and subplate, forming the cortical plate in between. The following cohort of neurons (yellow) migrates past their cortical plate predecessors (red) to form a more superficial layer. Successive waves of neurons then continue to migrate past their predecessors into six distinct laminae, with the earliest born cortical plate neurons residing in the deepest layer. This is described as inside-out formation of the cerebral cortex. Also labeled are the preplate (PP), the superficial marginal zone (MZ), and the subplate (SP). (Bottom) Distinct cortical zones during development. The VZ exists along the surface of the LV. The first wave of neurons, the PP, aligns above the VZ. The next wave of neurons forms the cortical plate, splitting the PP into the superficial MZ (at the pial surface and composed mostly of Cajal–Retzius cells) and the SP. As the brain matures, neurons must migrate longer distances, including through a cell body-sparse intermediate zone (IZ) composed mostly of an increasing number of projecting axons.

Distinctions between Cdk5 and Reelin Mutant Mice

Formation of the Preplate

The similarities between mice deficient in Reelin and Cdk5 signaling have fueled much research aimed at

determining if there is any potential cross-talk between these two pathways. However, time has demonstrated that it is really the distinctions, not the similarities, between these mutant mice that are more useful in defining the relative importance of the pathways and understanding the developmental neuronal migration program. One major difference is that early born cortical plate neurons in Cdk5-deficient mice migrate fairly normally and split the preplate into the superficial marginal zone and the deeper subplate. However, in Reelin-deficient mice these early born neurons intended to form the cortical plate fail to split the preplate, instead accumulating in a deep position, beneath the preplate.

Migratory Mode

Further distinctions in Cdk5 and Reelin mice have come from analysis of the behavior of migrating neurons at the cellular level *in vivo*. Neurons that migrate radially to form the neocortex (i.e., the pyramidal neurons born in the dorsal ventricular zone) display one of two characteristic migratory modes.

Somal translocation Early born neurons migrate primarily by extending a leading process that attaches to the nearby pial surface. The cell body of the migrating neuron then moves in a rather continuous manner toward the pial surface, with the leading process shortening progressively as the cell traverses. This migratory mode is called somal translocation, and it is likely used by the early born neurons because they have the shortest distance to migrate and are capable of maintaining their leading process and its attachment to the pial surface throughout their journey. Importantly, somal translocation requires only a single step of continuous nuclear movement and consequent shortening of the leading process.

Locomotion Later born neurons that must traverse longer distances display a distinct and more complex migratory behavior described as locomotion. These neurons extend a leading process into the subventricular zone, which they move into and then pause to become multipolar. Some of these neurons then reverse direction and move back into the ventricular zone before reversing polarity, and all of these neurons ultimately use a pia-directed leading process to migrate along their mother radial glia cell scaffold toward their final cortical position. Neurons that are undergoing locomotion display a short leading process that is not attached to the pial surface and migrate in a saltatory manner characterized by short bursts of movement followed by long stationary phases. In the final stage of migration, when their leading process has reached the pial surface and there is only a short distance to traverse, neurons switch from locomotion to somal translocation. This complex migratory behavior of later born neurons, which consists of multiple switches in polarity and repeated bursts of movement, likely requires much more intricate cytoskeletal rearrangement than that of the early born neurons that move by one step of somal translocation.

Migratory modes in Cdk5 and Reelin mutant mice As might have been predicted, the cellular behavior in mice with decreased Cdk5 activity (p35−/− mice) displays mostly normal somal translocation of early born neurons that first form the cortical plate properly by splitting the preplate. However, when the migratory mode should switch to that involving locomotion, p35−/− neurons fail to remain attached to their mother radial glia guide and start to move by an abnormal branched migration mode in which there is repeated extension of a branched leading process and subsequent zigzag movement of the soma to different branch points. This phenotype is in sharp contrast with mice that have a deficiency in the Reelin signaling pathway (scrambler mice). First, early born scrambler neurons fail to attach the endfeet of their leading process within the preplate. More important, however, scrambler neurons seem unable to dissociate from their mother radial glia at the completion of the migratory journey. This rather major distinction between Cdk5- and Reelin-deficient migrating neurons suggests that the two pathways mediate different cellular functions.

Compound Mutant Mice

Further evidence that Cdk5 and Reelin are operating in parallel pathways to regulate neuronal migration comes from studies of compound mutant mice. Mice that are deficient in ApoER2 or Dab1 and p35 exhibit an exacerbated cytoarchitectural phenotype in the hippocampus and neocortex suggestive of synergistic contributions of each pathway to brain development. Taken together, all of these studies suggest that proper lamination of the neocortex requires distinct functions of Cdk5 and Reelin signaling.

Doublecortin and Dynein Pathways in Neuronal Migration

Doublecortin

In addition to those directly involved in Cdk5 and Reelin signaling, several proteins that are also important for proper corticogenesis have been discovered through analysis of mutant mice and human diseases. Most of these proteins are important for cytoskeletal functions or cell adhesion. Missense mutations in

Doublecortin (DCX) can result in an X-linked form of lissencephaly (smooth brain) in humans. Furthermore, silencing of DCX expression with RNAi in rats, or genetic knockout of DCX and DCLK (a related protein) in mice, results in neuronal migration defects. DCX is expressed only in postmitotic neurons and is a microtubule-associated protein containing two tandemly repeated 90-amino acid domains that are capable of binding purified tubulin. Analysis of disease-causing human mutations showed that several of the missense mutations cluster in the tandem repeats. It is believed that DCX normally functions to stabilize microtubules.

Dynein, Lis1, and Ndel1 (Nudel)

Another pathway involved in neuronal migration involves the function of the minus-end microtubule motor dynein and two dynein interacting proteins, Lis1 and Ndel1 (also referred to as Nudel). In addition to DCX, human lissencephaly-causing mutations have been linked with the LIS1 gene. Furthermore, mouse models of Lis1 deficiency show a dosage-dependent effect on neuronal positioning. Ndel1 was first cloned as a Lis1-binding protein, but it can also associate directly with dynein. Further evidence has suggested that both Lis1 and Ndel1 positively regulate dynein function and maintain microtubule networks essential for migration. RNAi-mediated knockdown in embryonic mouse brain of dynein heavy chain, Lis1, or Ndel1 results in severe neuronal positioning defects.

Cdk5 Pathways Controlling Neuronal Migration

Since the phenotype of Cdk5-deficient mice is identical to the double knockout of its activators, p35 and p39, it is Cdk5 kinase activity that is essential for neuronal migration. Therefore, elucidation of relevant Cdk5 substrates will provide the underlying molecular mechanisms by which this kinase regulates neuronal migration. Furthermore, since Cdk5-deficient neurons display irregular migratory behavior that likely requires massive cytoskeletal rearrangement, logic dictates that Cdk5 substrates would regulate cytoskeletal functions during migration.

Cdk5 Phosphorylation of Focal Adhesion Kinase

S732-phosphorylated Focal Adhesion Kinase regulates centrosome-associated microtubules The first Cdk5-dependent phosphorylation event demonstrated to be essential for neuronal migration was phosphorylation of focal adhesion kinase (FAK) at serine 732. FAK was identified as a tyrosine kinase that localizes to focal adhesions, which are structures linking the extracellular matrix to the actin cytoskeleton. Interestingly, Cdk5-dependent phosphorylation of FAK is actually more important for microtubule organization during neuronal migration. S732-phosphorylated FAK, which is abolished in Cdk5-deficient mice, is enriched along a centrosome-associated microtubule fork/cage-like structure that abuts the nucleus in wild-type neurons. In Cdk5-deficient neurons, and in neurons overexpressing the S732A nonphosphorylatable FAK mutant, there is no discernible organization to this microtubule fork/cage structure.

The role of the cytoskeleton in migrating neurons A general hypothesis to explain the role of the cytoskeleton during neuronal migration relies on a simplified model based mainly on *in vitro* experiments in which the movement of a neuron consists of a few steps. First, there is extension of a leading process from the cell soma in the direction of migration, an event that primarily relies on the actin cytoskeleton. At this point, the centrosome, which is the major microtubule-organizing center, is located ahead (in the direction of migration) of the cell soma in the leading process. Second, microtubules linking the centrosome and the nucleus, which occupies most of the cell soma, drive nuclear movement into the leading process. This second step is commonly referred to as nucleokinesis. A third step is retraction of the trailing process. Successive cycles of leading process outgrowth and nucleokinesis will result in a migrating neuron. Importantly, this model suggests that different components of the cytoskeleton have distinct cellular functions during migration (Figure 3).

S732-phosphorylated FAK and nucleokinesis It was hypothesized that the microtubule fork/cage structure decorated with Cdk5-phosphorylated FAK, linking the centrosome to the nucleus, and disorganized in the cases of FAK–S732A expression or Cdk5 deficiency, would be essential for proper nucleokinesis. Indeed, live imaging of neurons migrating *in vitro* showed impaired nuclear movement. When the portion of the nucleus distal to the leading process began to translocate in the direction of migration, the proximal pole of the nucleus failed to move in tandem, resulting in a more rounded nuclear morphology. Furthermore, overexpression of FAK–S732A in embryonic brain by *in utero* electroporation resulted in neuronal migration defects with a characteristic rounded nuclear morphology phenotype that was very similar to what was observed in Cdk5-deficient brains (Figure 4). Therefore, Cdk5 phosphorylation of FAK is important for the maintenance of an essential microtubule network that mediates nucleokinesis. Although the discovery

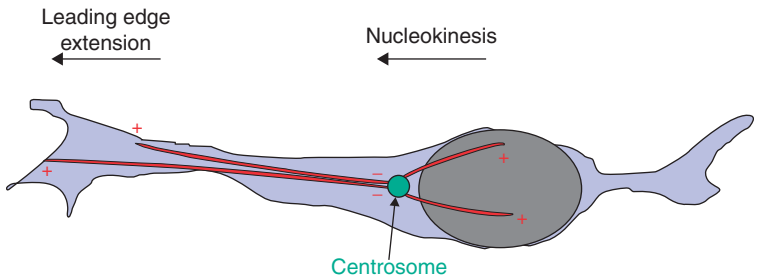

Figure 3 The role of the cytoskeleton in neuronal migration. Neurons first extend a leading process in the direction of migration (leading edge extension), an event primarily dependent on the actin cytoskeleton. Microtubules (red) emanating from the centrosome (green) and abutting the nucleus then drive nucleokinesis, the movement of the nucleus (dark gray) toward and into the leading process.

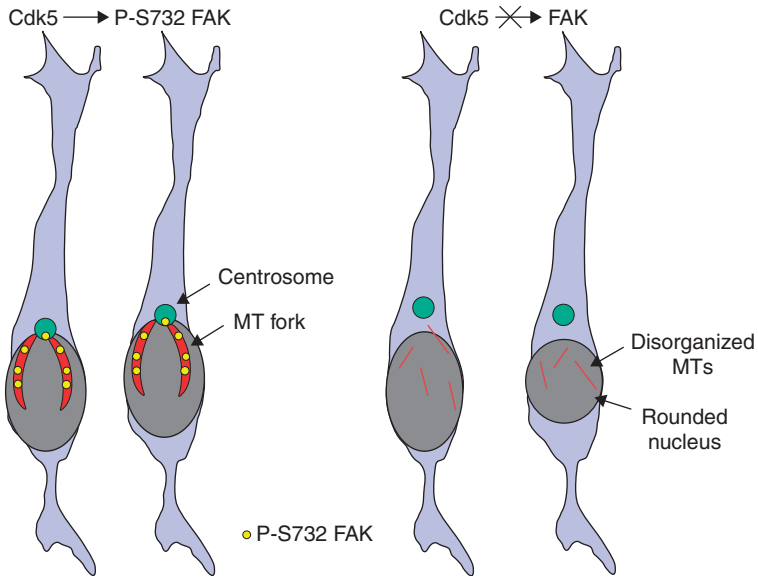

Figure 4 Cdk5 phosphorylation of FAK at S732 regulates nucleokinesis. (Left) Microtubules (red) emanating from the centrosome (green) and abutting the nucleus (dark gray) in a fork-like configuration are decorated with Cdk5-phosphorylated FAK (yellow). Nucleokinesis occurs as the nucleus moves into the leading process (compare nuclear position in the two cells). (Right) In cases of Cdk5 deficiency or expression of a nonphosphorylatable form of FAK, microtubules are severely disorganized. The pole of the nucleus proximal to the leading process fails to move, whereas the distal pole does move. This aberrant coupling of movement results in a more rounded nuclear morphology and a failure of nucleokinesis.

of FAK–S732 phosphorylation represented an important step in understanding the mechanisms that underlie Cdk5-mediated neuronal migration, recent work has suggested that a few other substrates are also important for implementing the complete downstream effects of Cdk5.

Cdk5 Phosphorylation of Doublecortin

An essential neuronal migration protein, DCX, is also a Cdk5 substrate. DCX has been shown to regulate the distance of nuclear-centrosome coupling in neurons, and it is localized to the microtubule fork/cage-like structure linking the centrosome and nucleus. Phosphorylation of DCX at serine 297 by Cdk5 regulates the interaction between DCX and microtubules, and mutation of S297 blocks the effects of DCX in a migration-dependent cellular reaggregation assay in a manner similar to pharmacological inhibition of Cdk5 activity, suggesting a role in nucleokinesis.

Cdk5 Phosphorylation of Ndel1

Cdk5 can also phosphorylate Ndel1, an essential member of the dynein/Lis1/Ndel1 complex. Indications that this complex is involved in nucleokinesis during neuronal migration first came from counterparts in *Aspergillus nidulans*, in which mutations in the orthologs of dynein (NudA, NudC, and NudG), Lis1 (NudF), and Ndel1 (NudE) all result in abnormal nuclear distribution ('Nud'). Furthermore, dynein, Lis1, and Ndel1 all are localized to the microtubule fork/cage that links the centrosome to the nucleus, and RNAi-mediated knockdown of these proteins caused a disruption of microtubule bundles. Live imaging of RNAi transfected cultured neurons demonstrated that

nuclear movement is also disrupted; rather than moving continuously toward the centrosome in the leading process, the movement of the nucleus oscillated back and forth. Cdk5-phosphorylated Ndel1 is capable of interacting with 14–3-3ε, a gene that is deleted along with Lis1 in severe cases of lissencephaly (e.g., Miller–Dieker syndrome). Interestingly, mice deficient in 14–3-3ε display a mislocalization of Ndel1 and Lis1 and a mild cortical distortion. Cdk5-dependent phosphorylation of Ndel1 also facilitates an interaction with the microtubule severing protein katanin/p60, and dominant negative mutants of p60 result in defective migration.

Cdk5 Phosphorylation of p27 (kip1)

Whereas the majority of Cdk5 research aimed at understanding migration has suggested a critical role for the kinase in regulating microtubules, the description of p27 (kip1) as a novel substrate opened up the possibility that Cdk5 can also regulate the actin cytoskeleton in migrating neurons. Cdk5 phosphorylates p27 on serine 10, an event that is important for stabilizing p27 levels. Once stabilized, higher p27 levels result in the activation of cofilin, an actin-severing protein. Although p27-deficient mice do not have any obvious neuronal migration phenotype, RNAi-mediated knockdown of p27 in embryonic brains results in impaired migration of cortical neurons and decreased amounts of F-actin in neuronal processes.

Other Substrates of Cdk5 Potentially Important for Neuronal Migration

One intriguing candidate to play a role in neuronal migration is Pak1, a small GTPase effector that regulates both the microtubule and the actin cytoskeleton. Cdk5 can phosphorylate Pak1 at threonine 212, and expression of a nonphosphorylatable form of Pak1 results in severe neuritic patterning defects. Pak1 phosphorylation at T212 by Cdc2 regulates microtubules in dividing cells, so it is possible that the altered patterning in cortical neurons is due to a microtubule defect. Another microtubule-associated protein involved in neuronal migration that can be phosphorylated by Cdk5 is MAP1B. Stimulation of cultured cortical neurons with reelin or netrin increased MAP1B phosphorylation in a GSK3- and Cdk5-dependent manner, but Cdk5 does not appear to phosphorylate MAP1B in nonstimulated brain slices. *In vitro* experiments have also demonstrated that Cdk5 activity can regulate N-cadherin/β-catenin-mediated complexes, indicating a direct role for Cdk5 in cell adhesion. Finally, Cdk5 is capable of phosphorylating Dab1, a critical mediator of Reelin signaling, at serine 491. Interestingly, although this phosphorylation event occurs *in vivo*, it seems to be completely independent of Reelin signaling. It will ultimately be important to determine if phosphorylation of Dab1 impacts on neuronal migration (Figure 5).

Cdk5 and Axon Guidance

Corpus Callosum Defects in p35-Deficient Mice

In addition to neuronal migration, there is strong evidence that Cdk5 plays an important role in guidance for the correct establishment of axon/target connections. The first indication of this role of Cdk5 came from studies of the corpus callosum in p35-deficient mice, in which callosal axons display an abnormal projection pattern. Tracer injection and axonal staining experiments performed in p35−/− mice demonstrated a general failure of projecting cortical axons to assimilate into the body of the corpus callosum, suggesting a defect in fasciculation between axons.

Cdk5 and Semaphorin 3A

Further experiments have placed Cdk5 into a pathway with semaphorins, a well-established, large family of

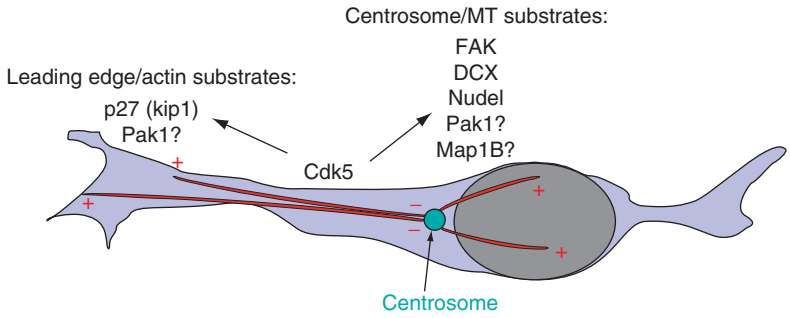

Figure 5 Cdk5 substrates involved in cytoskeletal regulation during neuronal migration. The majority of Cdk5 substrates understood to be involved in migration localize to the centrosome and microtubules, and they are known to regulate microtubule function and nucleokinesis. These include FAK, doublecortin (DCX), and Nudel. MAP1B phosphorylation, if important for migration, likely also regulates microtubule function. The first Cdk5 substrate shown to regulate the actin cytoskeleton during neuronal migration, and likely leading edge extension, was p27 (kip1). Phosphorylation of Pak1 may regulate actin or microtubule function.

secreted and membrane-bound proteins that can function as chemorepulsive/attractive cues. In cell culture, Sema3A normally induces collapse of growth cones through a large signaling pathway that includes functional roles for collapsin response-mediator proteins (CRMP), Rac, and LIM-kinase. Interestingly, dorsal root ganglia growth cones from Cdk5-deficient mice fail to collapse in response to Sema3A. Likewise, a dominant negative Cdk5 mutant attenuates Sema3A-induced collapse. *In vitro* kinase assays demonstrated that Cdk5 is capable of phosphorylating CRMP-2 at serine 522. Importantly, overexpression of a CRMP-2 alanine 522 mutant inhibited Sema3A-induced collapse. Mechanistically, it seems that Cdk5-dependent phosphorylation primes CRMP-2 for phosphorylation by GSK-3β, another event necessary for growth cone collapse. Therefore, the essential role of Cdk5 in Sema3A-induced growth cone collapse is likely mediated in part through phosphorylation of CRMP-2.

Cdk5 and GDNF/GFRα1

A second guidance pathway that seems to be associated with Cdk5 activity involves the function of GFRα1, a GPI-anchored receptor for glial cell line-derived neurotrophic factor (GDNF) that can be secreted or exist as a membrane-bound protein. Treatment of the motor cell line MN1 with GDNF in the presence of exogenous GFRα1 induced an increase in Cdk5 activity. Most interestingly, use of the Cdk inhibitor roscovitine or a dominant negative Cdk5 blocked GDNF/GFRα1-stimulated neurite outgrowth and guidance in explanted sympathetic ganglia.

Cdk5 Pathways Control Neuronal Motility and Structure

Taken together, all these findings suggest that Cdk5 is directly upstream of substrates that regulate the cytoskeleton during neuronal migration, and it specifically maintains the integrity of a microtubule network that links the centrosome with the nucleus and is essential for nucleokinesis. It is possible that Cdk5 activity is a signaling mediator during neuronal migration, relaying extracellular cues about the migration status to the cytoskeleton to permit the necessary remodeling. Multiple types of cues may converge on Cdk5 to translate their message into cytoskeletal rearrangement. Furthermore, Cdk5 is emerging as an important regulator of axon guidance, suggesting that the functions of Cdk5 in orchestrating brain development are manifold.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Cerebral Cortex; Cytoskeletal Interactions in the Neuron; Neocortex: Origins; Neuronal Motility and Structure: MARK and GSK Pathways.

Further Reading

Dhavan R and Tsai LH (2001) A decade of CDK5. *Nature Reviews Molecular Cell Biology* 2: 749–759.

Feng Y and Walsh CA (2001) Protein–protein interactions, cytoskeletal regulation and neuronal migration. *Nature Reviews Neuroscience* 2: 408–416.

Gupta A, Tsai LH, and Wynshaw-Boris A (2002) Life is a journey: A genetic look at neocortical development. *Nature Reviews Genetics* 3: 342–355.

Kawauchi T, Chihama K, Nabeshima Y, and Hoshino M (2006) Cdk5 phosphorylates and stabilizes p27kip1 contributing to actin organization and cortical neuronal migration. *Nature Cell Biology* 8: 17–26.

Kriegstein AR and Noctor SC (2004) Patterns of neuronal migration in the embryonic cortex. *Trends in Neurosciences* 27: 392–399.

Kruger RP, Aurandt J, and Guan KL (2005) Semaphorins command cells to move. *Nature Reviews Molecular Cell Biology* 6: 789–800.

Lambert de Rouvroit C and Goffinet AM (2001) Neuronal migration. *Mechanisms of Development* 105: 47–56.

Ledda F, Paratcha G, and Ibanez CF (2002) Target-derived GFRalpha1 as an attractive guidance signal for developing sensory and sympathetic axons via activation of Cdk5. *Neuron* 36: 387–401.

Marin O and Rubenstein JL (2003) Cell migration in the forebrain. *Annual Review of Neuroscience* 26: 441–483.

Niethammer M, Smith DS, Ayala R, et al. (2000) NUDEL is a novel Cdk5 substrate that associates with LIS1 and cytoplasmic dynein. *Neuron* 28: 697–711.

Rice DS and Curran T (2001) Role of the reelin signaling pathway in central nervous system development. *Annual Review of Neuroscience* 24: 1005–1039.

Ross ME and Walsh CA (2001) Human brain malformations and their lessons for neuronal migration. *Annual Review of Neuroscience* 24: 1041–1070.

Sasaki Y, Cheng C, Uchida Y, et al. (2002) Fyn and Cdk5 mediate semaphorin-3A signaling, which is involved in regulation of dendrite orientation in cerebral cortex. *Neuron* 35: 907–920.

Tanaka T, Serneo FF, Tseng HC, et al. (2004) Cdk5 phosphorylation of doublecortin Ser297 regulates its effect on neuronal migration. *Neuron* 41: 215–227.

Toyo-oka K, Shionoya A, Gambello MJ, et al. (2003) 14–3-3epsilon is important for neuronal migration by binding to NUDEL: A molecular explanation for Miller–Dieker syndrome. *Nature Genetics* 34: 274–285.

Walsh CA and Goffinet AM (2000) Potential mechanisms of mutations that affect neuronal migration in man and mouse. *Current Opinion in Genetics & Development* 10: 270–274.

Xie Z, Samuels BA, and Tsai LH (2006) Cyclin-dependent kinase 5 permits efficient cytoskeletal remodeling – A hypothesis on neuronal migration. *Cerebral Cortex* 16(supplement 1): i64–i68.

Xie Z, Sanada K, Samuels BA, Shih H, and Tsai LH (2003) Serine 732 phosphorylation of FAK by Cdk5 is important for microtubule organization, nuclear movement, and neuronal migration. *Cell* 114: 469–482.

Neuronal Motility and Structure: MARK and GSK Pathways

B Lu, Stanford University School of Medicine, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Neurons are highly polarized cell types that send and receive signals through two structurally and functionally distinct processes – axons and dendrites. The microtubule cytoskeleton serves as the structural core of the neuronal processes and its dynamic behavior is critical for neurite outgrowth, extension, and branching. Microtubule-associated proteins (MAPs) modulate microtubule dynamics by attaching to microtubules and stabilizing them. Phosphorylation plays a key role in regulating the microtubule-binding activity of MAPs. The microtubule affinity regulating kinase/partitioning defective-1 pathway has emerged as a major signaling pathway that regulates MAP–microtubule interaction. Deregulation of this pathway may underlie the pathogenesis of neurodegenerative diseases.

Introduction

Microtubules play an important function in regulating cell division, intracellular transport, and the maintenance of cell shape and polarity. To carry out these tasks, microtubules need to be dynamically unstable. This is best illustrated in neurons, the most polarized cells in the human body. Neurons send out axons and dendrites to communicate with their target cells by forming synapses. The 'outbound' transport of synaptic vesicles and organelles and the 'inbound' transport of neurotrophic factors and signaling complexes, among others, are essential for the proper functioning and survival of neurons. Microtubules provide the tracks for axonal and dendritic transport, whereas microtubule-based motor proteins such as kinesins and dyneins with built-in directionality provide the driving force. Both microtubule-stabilizing and -destabilizing drugs can inhibit the formation of axons and dendrites in neurons. This necessitates the precise control of microtubule dynamics in neurons.

Microtubules and Microtubule-Associated Proteins in Neuronal Motility and Structure

MAPs bind to microtubules along their entire length and modulate their dynamics. The best-studied MAP is the tau protein, which contains a conserved microtubule-binding domain with three or four imperfect repeats of 31-amino-acid motifs. Tau has been shown to be a microtubule stabilizer, a regulator of axonal trafficking, and a linker between the microtubule network and the plasma membrane. Tau promotes neurite outgrowth during neuronal differentiation. Even in nonneuronal cells, tau can promote the formation of cellular processes with characteristics of neurites. Despite these impressive cell culture studies, tau knockout mice are viable and do not show dramatic defects in neuronal differentiation. This is likely due to functional compensation by other MAPs *in vivo*. Consistent with this, tau and MAP1b double-mutant mice show severe defects in axonal elongation and neuronal migration.

The microtubule-binding activity of tau is regulated by phosphorylation. Most of the phosphorylation sites are SP or TP sites that have relatively small impact on the microtubule-binding activity of tau. These sites are likely acted on by some Pro-directed kinases, such as glycogen synthase kinase (GSK)-3β, cyclin-dependent kinase (Cdk)-5, or mitogen-activated protein kinase (MAPK). Other phosphorylation sites are non-Pro-directed. Of these, phosphorylation within the KXGS motifs in the microtubule-binding domains exerts a strong influence on tau–microtubule interaction. Drewes et al. pursued a biochemical purification of the kinase(s) responsible for phosphorylating the KXGS motifs in tau and this led to the identification of microtubule affinity regulating kinase (MARK).

A Conserved Role of MARK/PAR-1 Kinases in Establishing Cell Polarity

MARKs are mammalian homologs of the *Caenorhabditis elegans* Ser/Thr kinase PAR-1 (partitioning defective-1). PAR-1 kinase was initially identified in a genetic screen for genes that control the cell polarity and asymmetric division of *C. elegans* early embryos. PAR-1 is required for the polarized distribution of P granules to the germline lineage. The function of PAR-1 as a cell polarity regulator appears to be conserved. The *Drosophila* homolog of PAR-1 is required for microtubule organization and embryonic axis formation. In the fission yeast, the PAR-1/MARK ortholog Kin1 localizes to new cell ends after cell division and is important for bipolar growth. Mammalian MARK is asymmetrically localized in epithelial cells and may influence epithelial apical–basal polarity. Although it is a formal possibility that PAR-1 may regulate cell polarization through phosphorylation of MAPs in *C. elegans* and other organisms, this has not been proven biochemically or genetically. A total of six *par* genes (*par-1* through *par-6*) were isolated from the original genetic screen

in *C. elegans*. The products encoded by the six PAR genes perform evolutionarily conserved roles in establishing cellular polarization in different cell types, including neurons.

MARK Kinases Play an Important Role in Regulating Axonal Transport and Neurite Outgrowth

In addition to tau, other MAPs, such as MAP2 and MAP4, are also phosphorylated by MARK at the KXGS motifs in their microtubule-binding domains. Such phosphorylation events similarly led to the detachment of MAP2 and MAP4 from the microtubule and increased the dynamic instability of microtubules. These studies thus define a general role of MARK/PAR-1 kinases in regulating microtubule dynamics. Although the binding of MAPs to microtubules can have stabilizing effects, it also competes with microtubule-based molecular motors for microtubule binding. Thus, overexpression of tau can lead to reduced attachment of kinesin to the microtubule surface, inhibition of kinesin and dynein-directed axonal transport, and accumulation of mitochondria and amyloid precursor protein (APP)-containing vesicles at the cell interior, eventually resulting in the degeneration of the axons. Interestingly, moderate expression of MARK2 can rescue the toxic effects of tau on axonal transport, presumably by detaching tau from the microtubule surface. Because the velocity of motor-mediated active transport is not changed by MARK expression, it appears that simply by removing tau and creating a more open path, motor-driven axonal transport is facilitated. It should be noted that in this experimental condition, in which there is excess tau associated with the microtubule, modest coexpression of MARK removes excess tau without removing all tau from the microtubule and affecting the integrity of microtubules. It is possible that when higher levels of MARK kinase activity are induced, or in cell types containing less tau and other MAPs, disintegration of the microtubules and cellular toxicity could occur. Indeed, in cell types with a low endogenous level of MAPs, transient expression of MARK leads to rapid disintegration of microtubules and subsequent cell death. This biphasic effect of MARK/PAR-1 on microtubule function and cell survival depending on gene dosage may explain some divergent results from animal studies in which different levels of PAR-1 activity can have opposite effects on cell survival.

The importance of the MARK pathway in neurite outgrowth has been demonstrated in two different culture systems, although the effects may differ depending on the cell type. In neuroblastoma N2a cells, overexpression of MARK promoted neurite outgrowth, whereas inhibition of MARK2 using either a dominant negative mutant or kinase inhibitor blocked the formation of neurites. This suggests that MARK2-induced plasticity of microtubules is needed for neuronal polarity and the formation of neurites. In another study using cultured primary hippocampal neurons, it was demonstrated that knocking down the expression of MARK2 with small interfering RNAs (siRNAs) induced the formation of multiple axonlike neurites and promoted axon outgrowth, whereas ectopic expression of MARK2 led to loss of axons. It was further shown that the PAR-3/PAR-6/atypical protein kinase C (aPKC) complex negatively regulates the activity of MARK via aPKC-mediated phosphorylation of MARK, leading to less tau phosphorylation and stable microtubules. Thus, MARK2 appears to play a role in axon/dendrite fate specification and neurite outgrowth, but its effects may differ in different cell types. This may reflect a difference in the cytoskeletal composition or organization in different cell types. One caveat of these studies is the heavy reliance on transfection and neuronal culture. More rigorous tests using genetic mutants of MARK/PAR-1, especially in model organisms such as *Drosophila* and *C. elegans*, in which there is only one PAR-1 gene in each organism, will test the exact physiological roles of PAR-1 in axon–dendrite specification and neurite outgrowth. It will also be interesting to define the molecular and cellular mechanisms by which MARK/PAR-1 kinases may regulate the specification of axons. Local destabilization of actin has been proposed as a trigger for axonal specification and PAK5 has been described as an inhibitor of MARK/PAR-1 that can lead to stable microtubules and dynamic actin. It would be interesting to test whether PAK5–MARK interaction is involved in axonal specification by regulating actin dynamics.

Upstream and Downstream of MARK/PAR-1

Since the discovery of tau and MAPs as substrates of MARK, several other MARK/PAR-1 substrates have been identified, including Doublecortin (Dcx), Dishevelled, and several 14-3-3-binding proteins (PTPH1, Cdc25C, KSR1, and Plakophilin 2) whose interaction with 14-3-3 depends on prior phosphorylation by MARK/PAR-1. In the case of Dcx, phosphorylation by MARK regulates its affinity for microtubules in a manner similar to the regulation of tau and MAPs. The MARK phosphorylation site in Dcx is mutated in individuals affected by X-linked lissencephaly, a neurodevelopmental disorder associated with aberrant neuronal migration. This provides further *in vivo* evidence for the importance of the MARK/PAR-1

signaling pathway in regulating neuronal motility and structure. *Drosophila* PAR-1 was isolated in a biochemical purification for kinases that can phosphorylate Dishevelled, a key component of the Wingless/Wnt signaling pathway. Phosphorylation of Dishevelled by PAR-1 leads to activation of β-catenin and inhibition of Jun N-terminal kinase (JNK) signaling. A similar mechanism of Dishevelled regulation by MARK appears to be operating in mammalian systems. Interestingly, Dishevelled was found to be associated with axonal microtubules and to regulate microtubule stability through inhibition of the Ser/Thr kinase GSK-3β. As discussed later, the GSK-3β pathway has also been implicated in the control of neuronal motility and structure. It thus appears that the MARK/PAR-1 pathway may act through Dishevelled to impinge on the GSK-3β pathway in these processes.

The upstream signaling events that regulate MARK/PAR-1 kinase activity are just beginning to be understood. MARK kinases purified from bovine brain have been phosphorylated at multiple sites, including the T208 and S212 sites located in the kinase domain, suggesting that MARK/PAR-1 activity is regulated by phosphorylation *in vivo*. The T208 site is located at a strategic position in the kinase domain consistent with an activating role for T208 site phosphorylation on MARK/PAR-1 activity since a phosphate group at that position would stabilize the kinase domain via electrostatic interactions with other charged residues. Mutagenesis studies have largely confirmed this finding. In contrast, conflicting data have been reported on the role of S212 site phosphorylation. Two kinases, the tumor suppressor LKB1 and MARK kinase (MARKK), have been implicated as the upstream activating kinases that act on the T208 site (Figure 1). LKB1 binds *in vivo* with two ancillary proteins (STRAD and MO25) to form an active kinase complex. LKB1 has been shown to respond to stress signals and MARKK is a member of the Ste20 family of kinases that are also implicated in stress response. It is thus possible that the LKB1 and MARKK pathway may incorporate upstream stress signals to regulate MARK/PAR-1 activation. Supporting this notion, in mammalian cell culture, MARK is activated by osmotic stress. There is also evidence that MARKK may act through MARK to regulate the binding of tau and MAPs to microtubules and influence microtubule stability. Whether LKB1 may exert similar effects on MARK/PAR-1 remains to be determined. GSK-3β was reported to be a kinase that acts on the S212 site of the MARK kinase domain to positively regulate the activity of MARK toward tau (Figure 1). If confirmed, this finding suggests yet another level of interaction between the MARK and GSK-3β pathways.

Both mammalian MARK and *Drosophila* PAR-1 are phosphorylated by aPKC at the C-terminal 'spacer' domain. Although this domain is the least conserved region among the MARK/PAR-1 kinases, the aPKC phosphorylation site is highly conserved. Phosphorylation by aPKC disrupts the peripheral membrane localization of MARK and negatively regulates its activity (Figure 1). A study of mammalian hippocampal neurons confirmed the functional significance of this phosphorylation event in regulating neurite outgrowth. Overexpression of MARK2 or a mutant form of MARK2(T595A) in which the aPKC phosphorylation site is mutated causes failure of axonal specification. Whereas the coexpression of PAR-3/PAR-6/aPKC could rescue the defects caused by wild-type MARK2,

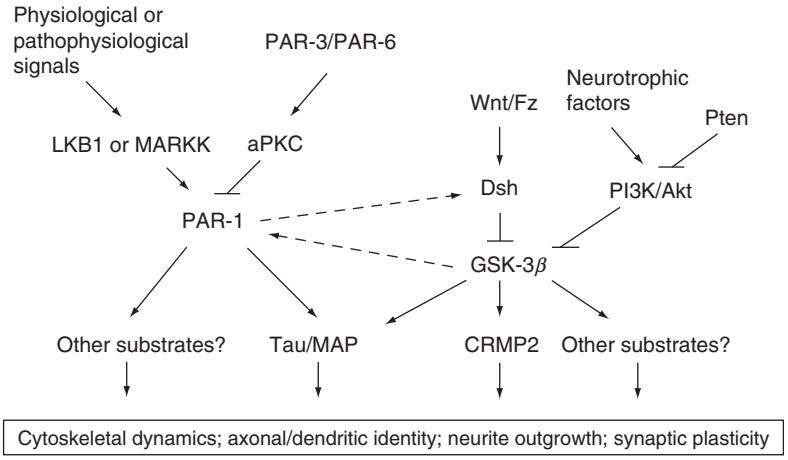

Figure 1 Diagram summarizing the current knowledge of the PAR-1 and GSK-3β signaling pathways that regulate microtubule dynamics, axonal/dendritic identity, neurite outgrowth, and synaptic plasticity. Dashed arrows indicate potential sites of cross-talk between the two pathways.

it had no effect on MARK2(T595A)-induced polarity defects. The involvement of aPKC–MARK interaction in neuronal polarization was further demonstrated by the finding that overexpression of a mutant form of MARK with a deleted kinase domain but an intact aPKC-binding domain abolished neuronal polarization. Moreover, suppression of MARK2 by siRNA rescued the polarity defects induced by a dominant negative aPKC mutant. These results thus establish aPKC as an upstream negative regulator of MARK/PAR-1 in axonal specification (Figure 1). In addition to PAR-3 and PAR-6, the insulin/PI3K/Akt pathway can also regulate aPKC activity, suggesting that diverse physiological signals may impinge on aPKC to regulate the MARK pathway.

GSK-3β in Neuronal Motility and Structure and Its Relationship with MARK/PAR-1

As mentioned previously, the GSK-3β pathway can be connected to the MARK pathway through various mechanisms. Although the exact relationship between the two pathways in determining neuronal motility and structure remains to be determined, given the known important roles of GSK-3β in regulating microtubule dynamics and neuronal polarity, it is necessary to review what is known about the regulation and function of the GSK-3β pathway in these processes, which will shed light on its interaction with the MARK pathway. GSK-3β is a multifunctional kinase highly conserved in eukaryotes. GSK-3β has been implicated in a range of biological processes, including the canonical Wnt signaling pathway, microtubule dynamics and cell migration, and axonal specification. Unlike MARK, which requires activation by upstream activating kinases, GSK-3β generally has a high basal kinase activity and various signaling pathways act by inhibiting GSK-3β activity. Phosphorylation of Ser at position 9 is a major mechanism by which the activity of GSK-3β is inhibited. In cultured hippocampal neurons, there is no significant difference in the distribution of activated GSK-3β in the neurites of nonpolarized neurons. However, after the transition from nonpolarized to polarized neurons, GSK-3β activity is higher in dendrites than in axons. This asymmetry in GSK-3β activity is important for determining axon–dendrite polarity. Akt and Pten act upstream of GSK-3β to determine neuronal polarity, and they play opposite roles in the process, with Akt inhibiting GSK-3β and Pten activating it.

There are multiple downstream targets through which GSK-3β may regulate neuronal polarity (Figure 1). Collapsin response mediator protein-2 (CRMP-2) is a microtubule-binding protein that plays a critical role in specifying axon/dendrite fate, possibly by promoting neurite elongation via microtubule assembly. CRMP-2 is phosphorylated by GSK-3β at Thr514 and this phosphorylation event inactivates CRMP-2's activity in axonal specification. Expression of a nonphosphorylatable form of CRMP-2 induced the formation of multiple axonlike neurites in hippocampal neurons and suppressed the loss of neuronal polarization phenotypes caused by a constitutively active form of GSK-3β. Furthermore, treatment of hippocampal neurons with neurotrophin-3 (NT-3) led to inactivation of GSK-3β, dephosphorylation of CRMP-2, and axonal outgrowth. Knockdown of CRMP-2 inhibited the effects of NT-3. These results thus convincingly demonstrate that GSK-3β regulates neuronal polarity by phosphorylating CRMP-2.

GSK-3β can also phosphorylate other microtubule-binding proteins, such as MAP1B, tau, and adenomatous polyposis coli gene product (APC). Phosphorylation by GSK-3β inhibits the ability of tau and MAP1B to bind microtubules. Hippocampal neurons isolated from mice lacking both tau and MAP1B could form dendrites but not axons, suggesting that tau and MAP1B may also mediate the effect of GSK-3β in regulating axonal specification. Inactivation of GSK-3β in axonal tips would allow tau and MAP1B to promote microtubule assembly and axonal elongation. Phosphorylation of APC and inhibition of its microtubule-binding activity by GSK-3β may also play a role in establishing neuronal polarity. APC and kinesin superfamily (KIF)3A have been shown to interact with mPAR-3. In polarized hippocampal neurons, APC selectively accumulates and co-localizes with mPAR-3 at nascent axon tips. Inhibition of GSK-3β activity at axon tips is required for mPAR-3 targeting and neuronal polarization. Given the role of the PAR-3/PAR-6/aPKC pathway in regulating MARK/PAR-1 activity, as discussed previously, this effect of GSK-3β on PAR-3 localization via APC may indirectly impact the MARK pathway. However, it should be kept in mind that the relationship between the MARK pathway and the GSK-3β pathway may be more complicated and there may not be a simple linear relationship between them. For example, as described previously, MARK/PAR-1 can directly phosphorylate and regulate Dishevelled in both *Drosophila* and mammals. Dishevelled can stabilize microtubules by inhibiting GSK-3β through a transcription- and β-catenin-independent pathway, resulting in a decrease in GSK-3β-mediated phosphorylation of MAP1B. Thus, through Dishevelled, MARK/PAR-1 could also impinge on the GSK-3β pathway.

In addition to a connection between MARK and GSK-3β via cross-talk between the two signaling

pathways, MARK and GSK-3β may cooperate at another level by acting on the same substrate proteins. GSK-3β phosphorylation of its substrates usually requires prior (priming) phosphorylation by other kinases. In *Drosophila*, there is evidence that prior phosphorylation of tau by PAR-1 in the microtubule-binding domain is required for the subsequent phosphorylation of tau by downstream kinases including GSK-3β. Similarly, phosphorylation of tau at a different site by Cdk5 may also serve a priming role for GSK-3β-mediated phosphorylation at a nearby site. It will be interesting to test whether the phosphorylation of other substrates by GSK-3β may also require prior phosphorylation by MARK/PAR-1. In this regard, it would be interesting to test whether GSK-3β and MARK/PAR-1 cooperate in phosphorylating some of the known GSK-3β substrates such as kinesin motors, whose potential phosphorylation by MARK/PAR-1 has not been tested.

MARK and GSK-3β Pathways in Neurodegenerative Diseases

The regulation of tau phosphorylation by kinases and phosphatases has received considerable attention because abnormal phosphorylation of tau is associated with a number of neurodegenerative diseases, such as Alzheimer's disease, frontotemporal dementia, Pick's disease, and supranuclear palsy. In diseased neurons, tau becomes hyperphosphorylated, highly insoluble, and forms aggregates called neurofibrillary tangles. Although a few of these phosphorylation events may be disease specific, the majority of these events, including those at MARK/PAR-1 and GSK-3β target sites, also occur in normal brains, albeit at a reduced frequency compared to diseased brains. This suggests that dysregulation of these tau kinases or the corresponding phosphatases may be integral to disease pathogenesis. Transgenic animal studies are beginning to decipher the functional roles of the different tau kinases in mediating tau toxicity and neurodegeneration. In *Drosophila*, for example, overexpression of PAR-1 can lead to neurodegeneration in photoreceptor neurons in a tau-dependent manner. Overexpression of PAR-1 also promotes the phosphorylation and neurotoxicity of human tau in *Drosophila* photoreceptor neurons. Genetic analysis has demonstrated that PAR-1 kinase activity is required for the phosphorylation of tau in the microtubule-binding domain and for conferring tau toxicity. Importantly, PAR-1 appears to play an initiator role in triggering a sequential phosphorylation event that involves downstream kinases such as GSK-3β and Cdk5. PAR-1 may act as a priming kinase, as discussed previously. Alternatively, phosphorylation of tau in the microtubule-binding domain by PAR-1 may efficiently release tau from the microtubule network, facilitating its subsequent phosphorylation by other kinases such as GSK-3β. Overexpression of GSK-3β can also enhance tau toxicity in *Drosophila*. Interestingly, although tau transgenic flies do not normally exhibit obvious neurofibrillary tangle formation, in GSK-3β and tau coexpression animals, prominent tau aggregates resembling neurofibrillary tangles are induced. Since there is no direct correlation between the extent of neurofibrillary tangle formation and neurotoxicity, the neurofibrillary tangle formation process is generally considered a late step in the disease process or may even be a protective response by the neurons attempting to sequester the smaller neurotoxic tau species. Several transgenic mouse models of GSK-3β have also been established. Whereas some of these models can enhance tau-induced neurodegeneration, others show rescue of tau-induced neurotoxicity. These divergent results may be due to the different promoters and different levels of GSK-3β transgene expression.

The cellular mechanism by which tau abnormality leads to neurodegeneration in disease state is not understood. Although it is possible that hyperphosphorylated tau may acquire some novel toxic function, it is also possible that abnormal phosphorylation of tau and other MAPs may interfere with their normal function in axonal transport and neuronal polarity. Because of their highly polarized morphology, neurons rely heavily on intracellular movement and transport such that even slight perturbation of the transport system will have dire consequences on neuronal physiology and result in neuronal death. There is increasing evidence that neurodegeneration in a number of diseases, such as Alzheimer's disease and Huntington's disease, is accompanied by defects in axonal and dendritic transport. For example, the APP, mutations in which cause familial Alzheimer's disease, may play a role in neurons as a vesicular receptor protein for kinesin I motor proteins. Studies have shown that APP is required *in vivo* for the transport of a particular class of vesicles that may contain the proteolytic processing machinery for APP. These is also evidence that overexpression of tau can inhibit axonal transport, including the transport of APP-containing vesicles. Disruption of axonal transport may therefore lead to increased intracellular cleavage of APP and generation of the neurotoxic Aβ peptides. Loss of axonal–dendritic polarity has long been considered a hallmark of neurodegeneration. It is possible that in Alzheimer's disease and other tauopathies, disease-initiating events may lead to dysregulation of tau kinases such as MARK and GSK-3β,

leading to defects in axon/dendrite fate specification or neurite polarization.

If MARK and/or GSK-3β are indeed aberrantly activated and are the culprits in causing neurodegeneration, then pharmacological inhibition of these kinases may offer therapeutic benefits against a range of neurodegenerative diseases with tau pathology. Although the unwanted side effects of long-term inhibition of these essential kinases, such as potential activation of the oncogenic Wnt pathway by GSK-3β, should not be overlooked, preliminary animal studies have shown that inhibition of GSK-3β can attenuate the toxic effects of tau. It is possible that by adjusting the dosage of kinase inhibitors to partially inhibit the disease-causing elevation of kinase activity but maintain the normal function of these kinases, a desired therapeutic effect may be achieved without the unwanted side effects. The approval of a small molecule inhibitor of a cancer-associated tyrosine kinase raises the hope that MARK- or GSK-3β-targeted drugs for the treatment of neurodegenerative diseases may soon be developed.

Summary

Neurons use axons and dendrites to send and receive signals. Axon/dendrite polarity is important for information processing and for ensuring unidirectional information flow in the nervous system. Furthermore, for neurons to function properly and to survive, axonal and dendritic transport of organelles, vesicles, and trophic factors needs to be precisely regulated. Microtubules and MAPs play essential roles in neuronal transport and axon/dendrite polarity formation. Protein kinases such as MARK and GSK-3β phosphorylate MAPs, regulate the binding affinity of MAPs toward microtubules, and in turn impact neuronal transport and axon/dendrite specification. These kinases respond to physiological and stress signals through various signal transduction pathways, and there is evidence that these signaling pathways may cross-talk. The physiological importance of the MARK and GSK-3β kinase pathways is highlighted by the association of their aberrant activation with neurodegenerative diseases. Understanding the molecular regulation and function of these pathways in regulating neuronal transport and structure may therefore offer novel and rational therapeutic strategies for the treatment of neurodegenerative diseases.

See also: Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and Alzheimer's Disease; Axonal Transport and Neurodegenerative Diseases; Axonal and Dendritic Identity and Structure: Control of; Microtubule Associated Proteins in Neurons; Microtubules: Organization and Function in Neurons; Neuronal Motility and Structure: Cdk5 Pathways.

Further Reading

Biernat J, Wu YZ, Timm T, et al. (2002) Protein kinase MARK/PAR-1 is required for neurite outgrowth and establishment of neuronal polarity. *Molecular Biology of the Cell* 13: 4013–4028.

Chen YM, Wang QJ, Hu HS, et al. (2006) Microtubule affinity-regulating kinase 2 functions downstream of the PAR-3/PAR-6/atypical PKC complex in regulating hippocampal neuronal polarity. *Proceedings of the National Academy of Sciences of the United States of America* 103: 8534–8539.

Drewes G (2004) MARKing tau for tangles and toxicity. *Trends in Biochemical Sciences* 29: 548–555.

Drewes G, Ebneth A, Preuss U, Mandelkow EM, and Mandelkow E (1997) MARK, a novel family of protein kinases that phosphorylate microtubule-associated proteins and trigger microtubule disruption. *Cell* 89: 297–308.

Drubin DG and Kirschner MW (1986) Tau protein function in living cells. *Journal of Cell Biology* 103: 2739–2746.

Goldstein LS (2001) Kinesin molecular motors: Transport pathways, receptors, and human disease. *Proceedings of the National Academy of Sciences of the United States of America* 98: 6999–7003.

Guo S and Kemphues KJ (1995) par-1, a gene required for establishing polarity in *C. elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell* 81: 611–620.

Hirokawa N (1994) Microtubule organization and dynamics dependent on microtubule-associated proteins. *Current Opinion in Cell Biology* 6: 74–81.

Jackson GR, Wiedau-Pazos M, Sang TK, et al. (2002) Human wild-type tau interacts with wingless pathway components and produces neurofibrillary pathology in *Drosophila*. *Neuron* 34: 509–519.

Lee VM, Goedert M, and Trojanowski JQ (2001) Neurodegenerative tauopathies. *Annual Review of Neuroscience* 24: 1121–1159.

Mandelkow E and Mandelkow EM (1995) Microtubules and microtubule-associated proteins. *Current Opinion in Cell Biology* 7: 72–81.

Mandelkow EM, Thies E, Trinczek B, Biernat J, and Mandelkow E (2004) MARK/PAR1 kinase is a regulator of microtubule-dependent transport in axons. *Journal of Cell Biology* 167: 99–110.

Nishimura I, Yang Y, and Lu B (2004) PAR-1 kinase plays an initiator role in a temporally ordered phosphorylation process that confers tau toxicity in *Drosophila*. *Cell* 116: 671–682.

Woodgett JR (2001) Judging a protein by more than its name: GSK-3. *Science STKE* 2001: RE12.

Yoshimura T, Kawano Y, Arimura N, Kawabata S, Kikuchi A, and Kaibuchi K (2005) GSK-3beta regulates phosphorylation of CRMP-2 and neuronal polarity. *Cell* 120: 137–149.

Neuronal Pacemaking

B P Bean, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuronal pacemaking refers to the ability of some neurons to generate spontaneous electrical activity by firing action potentials without being stimulated by synaptic input or by input from sensory receptors (Figure 1). When spontaneous firing of single neurons is rhythmic and regular, it is often called 'pacemaking,' by analogy to the regular pacemaking of action potentials in the sinoatrial node of the heart. However, firing of a pacemaking neuron need not necessarily trigger firing of downstream cells in a manner analogous to the pacemaker region of the heart. In fact, many neurons with pacemaking electrical activity are inhibitory neurons that release GABA.

Types of Pacemaking Neurons

Many different types of neurons in the mammalian brain have intrinsic pacemaking activity. One broad class of pacemaking neurons comprises neurons that release a variety of modulatory neurotransmitters, including acetylcholine, norepinephrine, dopamine, serotonin, and histamine. Spontaneous activity of such neurons is not surprising, since it appears that many such transmitters are often continuously released so as to produce tonic effects at low levels in a manner that is not cell-specific. Neurons releasing modulatory transmitters are often found to be spontaneously active at typical frequencies of 0.5–10 Hz (action potentials per second) if all synaptic input is removed. In a few cases, such as serotonergic neurons in the dorsal raphe nucleus, the neurons are silent in the absence of all synaptic input but fire rhythmically in a pacemaker-like fashion when there is steady excitation by norepinephrine acting on α-adrenergic receptors.

Many neurons involved in controlling sleep and wakefulness are spontaneously active. Neurons in the suprachiasmatic nucleus of the hypothalamus fire spontaneously at frequencies that show circadian rhythmicity, with faster firing daytime than during nighttime. Hypothalamic neurons that release the neuropeptide orexin, which promotes both wakefulness and feeding behavior, also show pacemaking activity, as do neurons in the tuberomammillary nucleus of the hypothalamus, which release histamine, associated with general stimulatory or activating effects throughout the brain.

A subset of interneurons in the cortex and hippocampus show pacemaking activity. Firing of groups of interneurons is often synchronized, with synchrony mediated both by electrical synapses (gap junctions) and by phase-lock of firing by means of chemical γ-aminobutyric acid (GABA)ergic synapses. It is believed that synchronous firing of networks of interneurons is centrally involved in controlling overall oscillatory activity of the cortex, especially at higher frequencies.

Several neuronal types in the thalamus are spontaneously active. Thalamic relay neurons, which receive sensory input and make glutamatergic projections to the cortex, are spontaneously active during sleep, when their spontaneous activity and synaptic input to the cortex (along with thalamocortical loops of synaptically connected neurons) may serve to underlie slow wave oscillations.

Many neurons throughout the motor system show pacemaking activity. Examples include neurons of the inferior olive; cerebellar Purkinje neurons; neurons in the deep cerebellar nuclei; and a number of neuronal types in the basal ganglia, including GABAergic neurons in the globus pallidus and substantia nigra pars reticulata, glutamatergic neurons in the subthalamic nucleus, giant cholinergic interneurons in the striatum, and dopaminergic neurons in the substantia nigra pars compacta.

Mechanisms of Pacemaking

Changes in the voltage across the membrane of excitable cells are produced by ionic currents flowing through ion channels, transmembrane proteins that form pores through the membrane. Neurons contain dozens of different types of ion channels with differing selectivity to various ions. In most cases, ion channels are normally closed but can be opened ('activated') by changes in membrane voltage. Activation is usually transient, either because the membrane voltage returns to a level that causes closing ('deactivation') of the channels or because of a property called 'inactivation,' whereby channels close even though a change in voltage is maintained. Pacemaking depends on opening of ion channels that pass electrical current flowing from outside to inside the cell. Such current produces depolarization of the cell membrane. In most cases, this current is carried by sodium ions, calcium ions, or both. A variety of different ion channels are involved in different types of pacemaking neurons.

Hyperpolarization-Activated Current

The best-known mechanism of pacemaking involves activation of a voltage-dependent current known as

hyperpolarization-activated current (I_h). This current is carried by a family of ion channels called hyperpolarization-activated cyclic nucleotide-sensitive (HCN) channels. I_h is a voltage-gated current with an unusual voltage dependence: it is closed at normal resting potentials and is activated by hyperpolarization, the opposite of the voltage dependence of almost all other voltage-dependent channels, which are opened by depolarization. The HCN channels carrying I_h are permeable to both potassium and sodium, with P_K:$P_{Na} \sim 3$:1. This gives a reversal potential near −30 mV. The channels begin to be activated by hyperpolarization below −60 to −70 mV, and at these potentials the inward current is carried by sodium ions. The flow of sodium ions into the cells produces depolarization toward threshold for action potential formation.

I_h is present in many neurons. However, many of its properties and a role in spontaneous firing of action potentials were first described in cardiac muscle cells in the sinoatrial node of the heart, where the electrical pacemaking of the heart originates. Because of this association, the channels underlying I_h have sometimes been called 'pacemaker' channels. (It is interesting, however, that although I_h is clearly involved in pacemaking in the heart, blocking the channels generally only slows but does not stop the heartbeat, showing that other channels are also involved.) In mammalian central neurons, I_h appears to play a major role in pacemaking in a subset of pacemaking neurons. These include thalamic relay neurons, large cholinergic interneurons in the striatum, globus pallidus neurons, some interneurons, and a subset of dopaminergic neurons in the midbrain. The role of I_h is often assayed by applying a reasonably selective channel blocker called ZD 7288 (Figure 2).

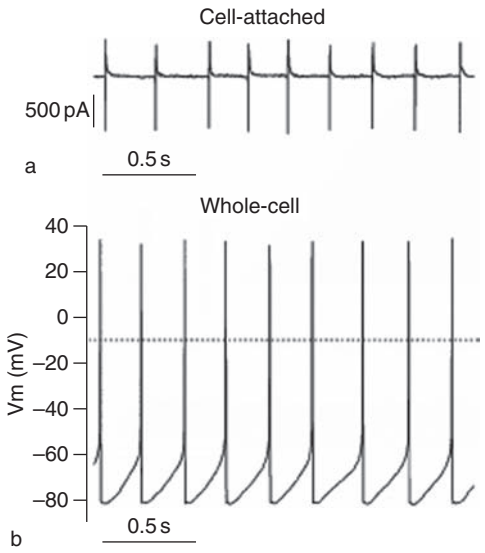

Figure 1 Electrical pacemaking in a suprachiasmatic nucleus neuron. Spontaneous firing of action potentials was recorded first in cell-attached patch mode and then in whole-cell current-clamp mode from an acutely isolated neuron from the suprachiasmatic nucleus of a rat hypothalamus. (a) Action currents from spontaneous firing recorded in cell-attached mode. In this recording configuration, the interior of the cell is undisturbed. (b) Spontaneous firing recorded in whole-cell mode 7 min after breaking through into whole-cell mode. The pipette solution consisted of 123 mmol^{-1} K-methanesulfonate, 9 mmol^{-1} NaCl, 1.8 mmol^{-1} $MgCl_2$, 0.9 mmol^{-1} EGTA, 9 mmol^{-1} HEPES, 14 mmol^{-1} creatine phosphate, 4 mmol^{-1} Mg–adenosine triphosphate, and 0.3 mmol^{-1} guanosine triphosphate, pH 7.3. The external solution included 1.2 mmol^{-1} $CaCl_2$, 3.5 mmol^{-1} KCl, 150 mmol^{-1} NaCl, 1 mmol^{-1} $MgCl_2$, 10 mmol^{-1} glucose, and 10 mmol^{-1} HEPES, pH 7.4. Reproduced from Jackson AC, Yao GL, and Bean BP (2004) Mechanism of spontaneous firing in dorsomedial suprachiasmatic nucleus neurons. *Journal of Neuroscience* 24: 7985–7998.

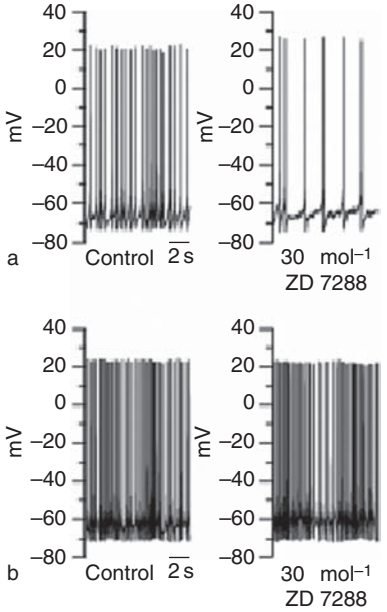

Figure 2 Effect of the hyperpolarization-activated current (I_h) blocker ZD 7288 on pacemaking in two dissociated substantia nigra pars compacta dopaminergic neurons. (a) Left, pacemaking in an acutely dissociated dopaminergic neuron from the substantia nigra par compacta (mouse). To identify dopaminergic neurons, a transgenic mouse line was used in which catecholaminergic neurons in the central nervous system express human placental alkaline phosphatase (PLAP) on the outer surface of the cell membrane. The line was obtained by introducing into the mouse genome PLAP complementary DNA linked to a promoter sequence of the gene for tyrosine hydroxylase. Living dopaminergic neurons were identified before recording by labeling of their membrane with a monoclonal antibody to PLAP directly conjugated to the fluorochrome Cy3 (E6-Cy3). Right, spontaneous firing in the same cell after application of 30 mmol^{-1} ZD 7288 to block I_h. (b) Lack of effect on pacemaking of ZD 7288 applied in a different acutely dissociated neuron (also identified as dopaminergic neuron by antibody to PLAP). Although I_h contributes to pacemaking in the cell in (a), it is not absolutely necessary, and it plays little or no role in the cell in (b). Reproduced from Puopolo M, Raviola E, and Bean BP (2007) Roles of subthreshold calcium current and sodium current in spontaneous firing of mouse midbrain dopamine neurons. *Journal of Neuroscience* 27: 645–656.

Persistent Sodium Current

In many pacemaking neurons, blocking I_h with ZD 7288 has little or no effect on spontaneous electrical activity. In most of these, the spontaneous depolarization activity appears to be due to current through voltage-activated sodium channels that produce steady state, noninactivating sodium current at subthreshold voltages (Figure 3). In such neurons, application of tetrodotoxin (TTX), an alkaloid toxin from puffer fish that selectively blocks voltage-dependent sodium channels, silences the firing and produces a stable resting potential (Figure 3(b)). Current through voltage-dependent sodium channels can be characterized most clearly in voltage-clamp experiments, in which the voltage of the membrane is controlled (using electronic circuitry) and ionic currents are measured. Current through voltage-dependent sodium channels is usually most obvious as a large transient current that activates rapidly (<1 ms) and inactivates (during several milliseconds) when the membrane is suddenly depolarized from rest. However, voltage-dependent sodium channels can also carry smaller, steady, noninactivating currents. This steady-state current is often called 'persistent' sodium current. It activates with a steep voltage dependence, with activation first evident near −70 or −65 mV and with maximal activation near −40 mV. Figures 3(c) and 3(d) show an example, in which TTX inhibited a steady inward current at −60 mV, as well as a bigger noninactivating current that was activated when the voltage was stepped from −60 to −52.5 mV. Figure 3(e) shows current as a function of voltage in the absence and presence of TTX, illustrating a TTX-sensitive inward current (shown subtracted in Figure 3(f)) that increases steeply from −65 to −50 mV. This steep voltage dependence produces regenerative depolarization, so if the membrane voltage is depolarized to −65 mV, inward sodium current flows, which produces further depolarization, which activates more channels, which pass larger inward currents, depolarizing the cell further.

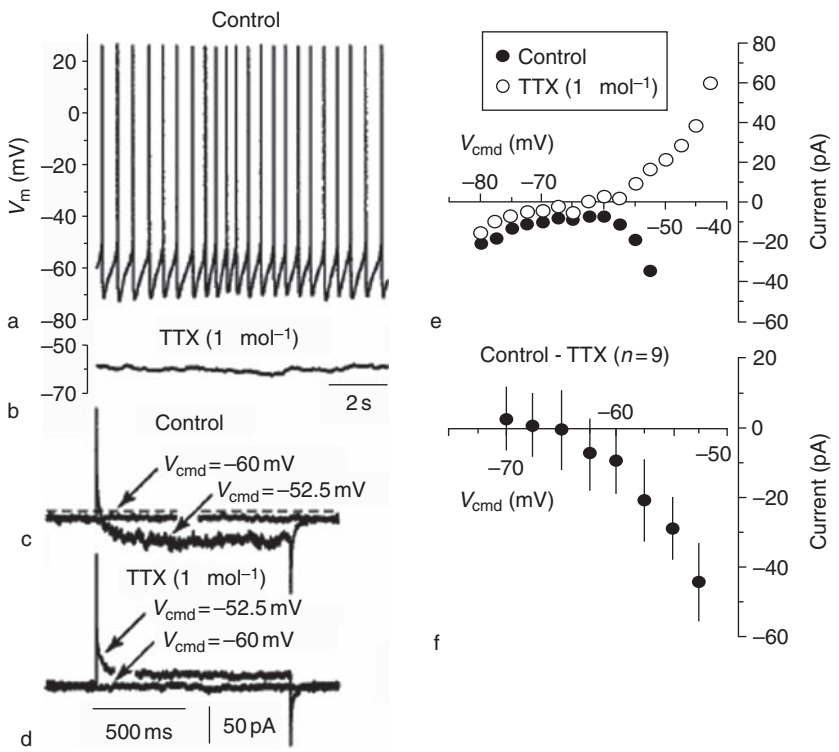

Figure 3 Role of subthreshold sodium current in driving pacemaking in a cholinergic interneuron in the striatum. (a) Spontaneous firing in a striatal cholinergic interneuron. (b) Application of tetrodotoxin (TTX) (1 μmol^{-1}) prevented action potential generation and established a stable subthreshold membrane potential (approximately −60 mV). (c) In voltage clamp, stepping the membrane from a holding potential of −60 to −52.5 mV elicited a slowly developing inward current that persisted throughout the pulse. The dashed line indicates the zero current level. (d) After TTX (1 μmol^{-1}) treatment, the inward current produced by the depolarizing voltage step was absent, and a small outward current was observed. In addition, the zero current level was shifted to −60 mV. (e) The current–voltage plot for a range of voltage steps (V_{cmd}) shows that, under control conditions, there is a region of negative slope conductance at potentials positive to −60 mV and no zero current point in the subthreshold voltage range. The inward current generated by depolarizing steps was blocked by TTX. (f) Voltage dependence of the TTX-sensitive current that was obtained by subtraction for nine neurons and pooled (mean ± standard deviation). V_{cmd}, voltage command. Reproduced from Bennett BD, Callaway JC, and Wilson CJ (2000) Intrinsic membrane properties underlying spontaneous tonic firing in neostriatal cholinergic interneurons. *Journal of Neuroscience* 20: 8493–8503.

This regenerative process continues until the threshold for action potential formation (typically in the range near −55 to −40 mV) is reached.

Evidence suggests that persistent sodium current active at subthreshold voltages is carried by the same population of channels that underlie the transient, inactivating sodium current that underlies action potentials. Models of gating of conventional inactivating sodium channels suggest that with strong depolarization, inactivation of transient sodium channels is ~99.5% complete and that the incomplete nature of inactivation results in steady-state sodium current that activates steeply at voltages between −65 and −40 mV. Virtually all neurons have large transient sodium currents, reaching typical magnitudes of 5–20 nA. Thus, steady-state current at the level of 0.5% would be in the range of 25–100 pA, typical of steady-state persistent sodium currents that are measured. Some neurons, such as pyramidal neurons in the entorhinal cortex, also have specialized sodium channels with more incomplete inactivation that can produce particularly large persistent sodium currents.

The ability of a single population of sodium channels to make both persistent sodium current at subthreshold voltages and transient sodium current at more-depolarized voltage constitutes an intrinsic system of rhythmic pacemaking generated by a single type of ion channel. The presence of a large density of voltage-dependent sodium channels in virtually all neurons implies that virtually all neurons have an intrinsic drive to spontaneous firing. Clearly, many neurons do not fire spontaneously, and pacemaking by this mechanism can be prevented if there is a resting potassium conductance large enough to produce a resting potential negative to −70 mV or so, at which persistent current is first activated. However, the association of persistent sodium current with transient sodium current suggests that pacemaking behavior is an intrinsic propensity of all neurons that use sodium channels for spike formation, requiring the presence of resting potassium channels to prevent or modulate this inherent tendency.

Neurons in which pacemaking appears to be mainly due to persistent sodium current at subthreshold voltages include neurons of the suprachiasmatic nucleus, pedunculopontine nucleus neurons, subthalamic neurons, Purkinje neurons, tuberomammillary nucleus neurons, and neurons of the deep cerebellar nuclei. Even in neurons in which blocking I_h stops pacemaking, it is possible that persistent sodium current plays the major role in generating spontaneous depolarization in the range from −70 mV to threshold, with I_h serving to ensure that the membrane voltage does not fall below −70 mV for very long.

Resurgent Sodium Current

Some pacemaking neurons have an unusual component of TTX-sensitive voltage-dependent sodium current called 'resurgent' sodium current, which activates when the membrane is repolarized to resting potentials following strong depolarizing pulses that have produced maximal inactivation of the sodium channels. Resurgent sodium current appears to arise from an unusual mechanism of inactivation in which channels are plugged by a positively charged intracellular blocking particle after they open. On repolarization, the block is relieved, and the channels carry an inward sodium current transiently before deactivating. This is unlike recovery from the most common mechanism of inactivation in sodium channels, which occurs without any current passing through the channels. Resurgent sodium current gives rise to a component of TTX-sensitive sodium current immediately after a spike. This produces a depolarizing influence and promotes a more rapid approach to threshold than would otherwise occur. Recovery from the mechanism of inactivation associated with resurgent current is faster than recovery from normal inactivation, so as sodium channels pass resurgent sodium current after a spike, they are also recovering from inactivation and able to give transient sodium current on depolarization. This effect also promotes faster firing of a second spike. Resurgent current has been found in several types of neurons that have pacemaking activity at rapid frequencies, notably cerebellar Purkinje neurons (which often fire spontaneously at ~40 Hz), subthalamic nucleus neurons (which often fire at ~20 Hz), and motor neurons of the trigeminal mesencephalic nucleus.

Voltage-Activated Calcium Current

In a few types of neurons, voltage-dependent calcium currents seem to be important for pacemaking. For example, in dopamine neurons in the pars compacta of the substantia nigra, blocking entry of calcium though calcium channels by replacing calcium with cobalt (which is impermeant in most calcium channels) stops spontaneous firing (Figure 4), and cessation of firing is accompanied by hyperpolarization, as if a depolarizing inward current carried by calcium has been blocked. This effect of blocking calcium entry is relatively unusual: in many other pacemaking neurons, such as cerebellar Purkinje neurons, subthalamic nucleus neurons, and suprachiasmatic nucleus neurons, blocking calcium entry has little effect on firing frequency (or may speed it) and often produces depolarization rather than hyperpolarization. These effects likely occur because calcium entry is often strongly coupled to activation of calcium-dependent

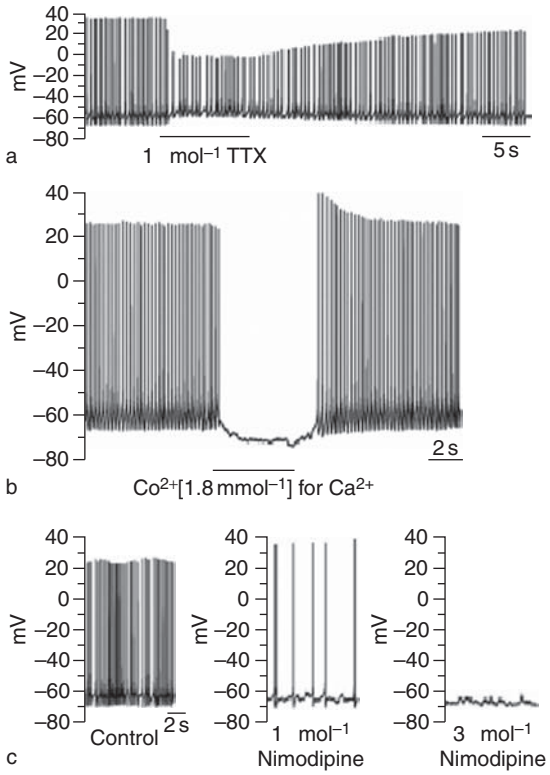

Figure 4 Role of calcium current in pacemaking of dopaminergic neurons from the substantia nigra pars compacta. (a) Effect of tetrodotoxin (TTX) on spontaneous activity in an acutely dissociated dopaminergic neuron from the substantia nigra pars compacta. Spontaneous activity continued, with smaller and broader spikes from a more depolarized potential. (b) Effect of blocking calcium entry (cobalt substitution for extracellular calcium) on spontaneous activity in another dissociated dopaminergic neuron. Immediate hyperpolarization and cessation of pacemaking suggest a role for calcium current. (c) Effect of the L-type calcium channel blocker nimodipine on spontaneous firing in an acutely dissociated dopaminergic neuron from the substantia nigra pars compacta. From Puopolo M, Raviola E, and Bean BP (2007) Roles of subthreshold calcium current and sodium current in spontaneous firing of mouse midbrain dopamine neurons. *Journal of Neuroscience* 27: 645–656.

potassium channels, and the net effect of blocking calcium entry is to inhibit an outward current.

Pacemaking dopaminergic neurons in the pars compacta of the substantia nigra will continue firing action potentials in the presence of TTX (Figure 4). This firing appears to be caused by calcium action potentials since it can be inhibited by removal of Ca or by cobalt, which blocks voltage-dependent calcium channels. The calcium current involved in pacemaking of midbrain dopaminergic neurons appears to be at least partly 'L-type' calcium current, since pacemaking is slowed or stopped by L-type channel blockers such as nifedipine and nimodipine (Figure 4). Although the dominant type of L-type calcium channels in most neurons (Cav1.2 channels) requires suprathreshold depolarizations for activation (positive to ∼−40 mV), L-type current carried by Cav1.3 channels can activate at much more negative voltages, as negative as −70 mV. Evidence suggests that current through Cav1.3 channels may play a major role in pacemaking of dopaminergic neurons in the substantia nigra. It is interesting that Cav1.3 channels are also strongly expressed in the sinoatrial node of the heart and probably contribute significantly to cardiac pacemaking.

Another class of calcium channels that can open at subthreshold voltages and likely contribute to pacemaking in some neurons includes low-threshold, T-type calcium channels (also called low-voltage-activated channels) formed by channels in the Cav3 family. These channels can activate at voltages as negative as −75 mV, but they also inactivate strongly, so that at steady voltages near −60 mV, they are almost completely inactivated. They are likely to contribute to pacemaking primarily when the action potential is followed by a deep afterhyperpolarization lasting long enough to remove inactivation. A significant role for current through T-type channels has been proposed for pacemaking in thalamic relay neurons. However, as there are no blockers that are highly selective for T-type channels, this theory has been difficult to test directly in these or other neurons.

Importance of Resting Potential

A key element regulating the occurrence and frequency of spontaneous firing is the resting potential of a cell. The mixture of conductances that are not voltage-dependent and that determine resting potential of neurons is not well understood. Obviously, a neuron that is continually active does not have a true resting potential. For spontaneously active neurons, the resting potential must be determined when firing is stopped, for example by application of TTX. TTX application usually, but not always, results in a stable resting potential (dopaminergic substantia nigra neurons being one of the exceptions). In general, neurons have resting potentials is the range between −75 and −55 mV, considerably depolarized to the potassium equilibrium potential (near −90 mV for mammalian solutions). Pacemaking neurons frequently have somewhat more depolarized resting potentials than neurons in general, from −60 to −45 mV. The resting potassium conductance of neurons (and all cells) appears to result primarily from expression of 'two-pore domain' family (KCNK) potassium channels, which are not voltage-dependent channels. These channels are highly potassium-selective. In cells with resting potentials considerably positive to E_K, it is unclear whether there is relatively moderate expression of KCNK channels or whether there is additional

conductance with permeability to other ions. Some cells with depolarized resting potentials have very high input resistances (i.e., very low resting conductance), and in these, there may simply be relatively little expression of KCNK channels. In lower-resistance cells, there may be substantial expression of KCNK channels but additional expression of background sodium conductances. Evidence for the latter is that in some neurons, replacement of external sodium by the larger cation *N*-methyl-D-glucamine causes hyperpolarization of the resting potential. In some neurons there may also be background calcium conductances that are not voltage-dependent, but these have not been characterized in any detail.

A very simple mechanism of pacemaking is provided by a cell in which there is expression of a reasonably high density of voltage-dependent sodium channels (which provide both persistent steady-state sodium current positive to −65 mV and also transient sodium current to form spikes) along with background conductances that by themselves would produce a stable resting potential positive to −65 mV. Such a cell is likely to fire spontaneously in a pacemaker-like manner as long as the action potential activates potassium conductances that can repolarize the membrane during the falling phase of the action potential sufficiently to provide recovery from inactivation of the transient sodium channels.

Voltage-Dependent and Calcium-Dependent Potassium Conductances as Regulators of Firing Rate

Most neurons have at least a half-dozen to a dozen distinct types of voltage-dependent or calcium-activated potassium channels, and the frequency of pacemaking can be strongly influenced by these conductances. A particular potassium current strongly implicated in control of pacemaking frequency is a current known as 'A-type' current (I_A), generally carried by channels in the Kv4 family, which both activates and inactivates at subthreshold voltages. I_A enables repetitive firing at low frequencies. During the period of relative hyperpolarization after a spike, some inactivation of I_A is removed. Then, as the membrane depolarizes at potentials in the range of −70 to −55 mV (still below threshold for an action potential), I_A increasingly activates and thereby slows the approach to threshold. In the steady state, I_A inactivates nearly completely at voltages positive to about −65 mV, so it eventually inactivates, allowing depolarization to threshold. Expression of a large density of I_A appears common in pacemaking neurons that fire rhythmically at relatively slow frequencies (0.5–5 Hz), such as midbrain dopaminergic neurons. There are other potassium currents that can be activated at subthreshold voltages, notably a current called 'I_D,' which is activated by subthreshold depolarizations, inactivates slowly (more slowly than I_A), and can similarly prolong approach to threshold. However, subthreshold I_D has been characterized mainly in nonpacemaking neurons such as cortical and hippocampal pyramidal neurons, and it is unclear whether there is prominent expression of I_D in pacemaking neurons.

Most voltage-dependent potassium channels require suprathreshold depolarizations to be activated significantly and are not activated in the lead-up to the action potential but rather during the spike itself. Potassium currents through these channels can still influence the frequency of pacemaking if they deactivate slowly enough after the spike repolarizes, pushing the membrane voltage toward E_K as long as they are activated. Of purely voltage-activated potassium channels, channels in the Kv1 family generally have the slowest deactivation kinetics and might influence pacemaking frequency by producing a period of relative hyperpolarization following the action potential.

Calcium-activated potassium currents may also play a major role in regulating firing frequency. During the action potential, there is generally activation of voltage-dependent calcium channels, especially during the falling phase of the action potential, producing a large increase in intracellular calcium, which activates calcium-activated potassium currents. There are two major types of calcium-activated potassium channels: KCa1 channels (also known as 'BK' channels), which are activated both by intracellular calcium and by depolarization, and KCa2 channels (also known as 'SK' channels), which are activated purely by increases in intracellular calcium. Both KCa1 and KCa2 channels are strongly expressed in most neurons. KCa1 channels respond to relatively large increases in intracellular calcium, and they are activated very quickly during the falling phase of the action potential. KCa2 calcium-activated channels respond to lower levels of calcium than do KCa1 channels but have intrinsically slower activation kinetics. Current through KCa2 channels contributes little to the fast repolarization phase of the action potential but helps shape the afterhyperpolarization that follows the action potential. After a spike, KCa2-mediated conductance can last for many hundreds of milliseconds before decaying, probably reflecting the relatively slow decay of intracellular free calcium. In contrast, KCa1 channels generally deactivate far more quickly. The longer-lasting conductance from KCa2 channels gives them a larger role in controlling pacemaking frequency. In several types of pacemaking neurons, block of KCa2 channels (or block of the

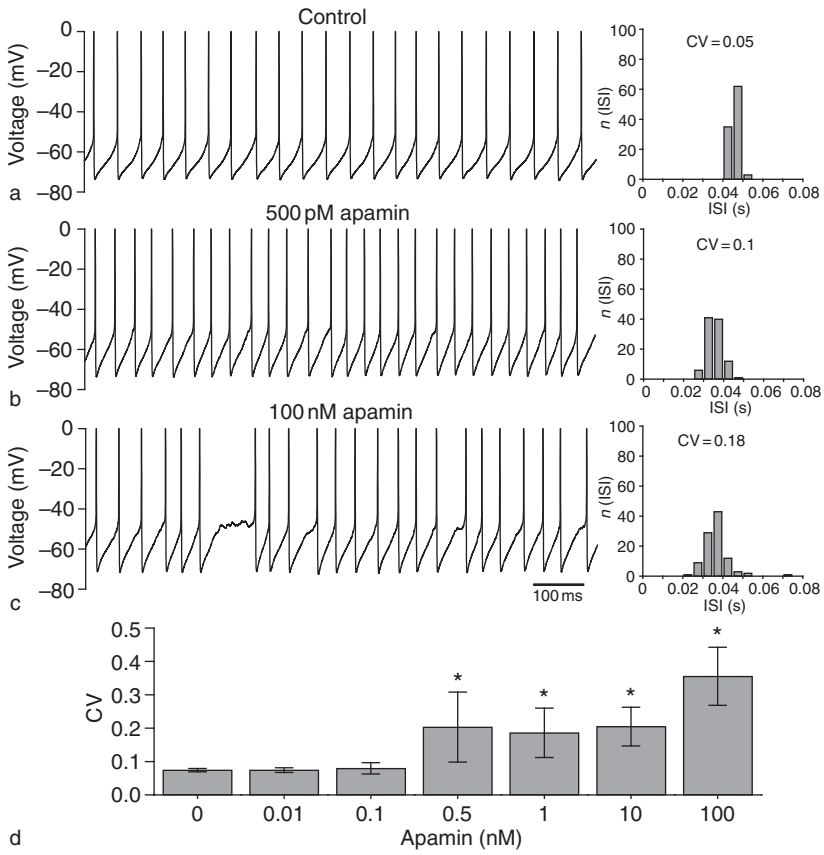

Figure 5 SK(KCa2) calcium-activated potassium channels are critical for the rhythmicity of pacemaking activity in rat subthalamic nucleus neurons. (a) Pacemaking was extremely regular in control conditions. Right, interspike interval histogram generated from 100 interspike intervals. Coefficient of variation (CV) = 0.05; frequency = 21.8 Hz. (b) Partial inhibition of SK channels by 0.5 $nmol^{-1}$ apamin. Firing was less regular (CV = 0.1), and frequency was somewhat faster (28.1 Hz). (c) Complete inhibition of SK channels by 100 $nmol^{-1}$ apamin. Firing was much less regular (CV = 0.18), frequency (27.5 Hz) not much changed from 0.5 $nmol^{-1}$ apamin. Calibration in (c) also applies to (a) and (b). From Hallworth NE, Wilson CJ, and Bevan M (2003) Apamin-sensitive small conductance calcium-activated potassium channels, through their selective coupling to voltage-gated calcium channels, are critical determinants of the precision, pace, and pattern of action potential generation in rat subthalamic nucleus neurons *in vitro*. *Journal of Neuroscience* 23: 7525–7542.

calcium current that activates them) has the striking effect of producing substantially less-regular firing (Figure 5). The reasons for this are not clear.

Plasticity of Pacemaking

A number of cases are known in which the frequency of firing of pacemaking neurons can be altered in a relatively long-lasting manner (many minutes to hours). For example, in spontaneously firing neurons in the medial vestibular nucleus, silencing the neurons for relatively brief periods (~5 min) by sustained inhibitory synaptic input can result in increases in the frequency of pacemaking that can last more than 2 h. These neurons serve to transform information about head movement into eye movements and are central to the vestibulo-ocular reflex that keeps appropriate eye position when the head moves. The plasticity of the pacemaking frequency (which is accompanied by an increase in the excitability of the neurons when responding to excitatory stimulation) is believed to constitute a form of motor learning that can recalibrate the gain of the vestibulo-ocular reflex. The mechanism of plasticity of vestibular neurons is thought to involve, during the period of inhibition, a decrease in intracellular calcium that leads to a decrease in basal Cam-kinase II enzymatic activity and a reduction in BK calcium-activated potassium current.

A remarkable example of plasticity is seen in principal neurons in the entorhinal cortex of the parahippocampal region, a key brain area for acquisition and consolidation of working memory. When the conductances in these neurons are modulated by steady activation of muscarinic acetylcholine receptors (known to be important in memory), the neurons acquire the ability to fire in a pacemaker-like fashion at frequencies that can be stepped up or down in a remarkably stable manner (Figure 6). Specifically, stimulation with a short (4 s) injection of current (which transiently produces high-frequency firing)

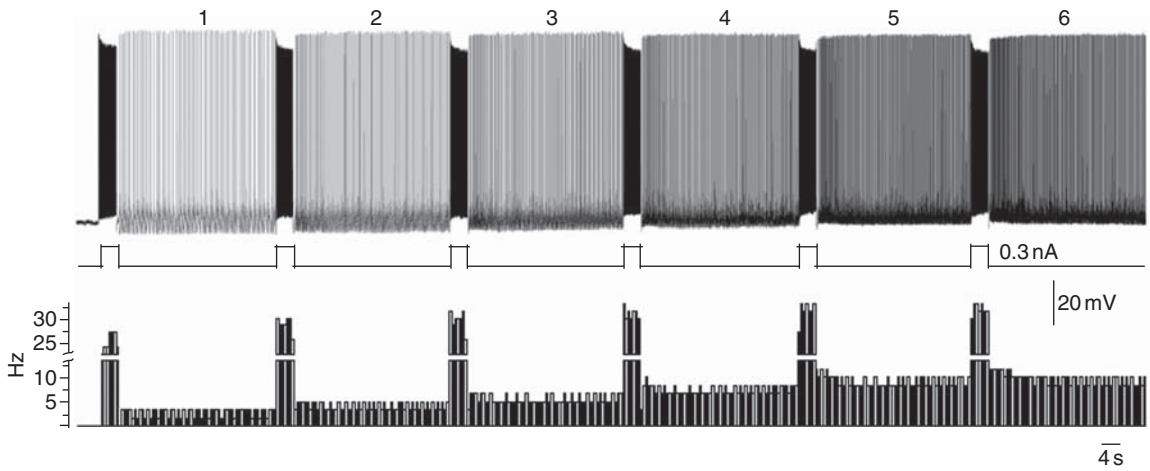

Figure 6 Plasticity of firing frequency in entorhinal cortical neuron exposed to an agonist of muscarinic acetylcholine receptors ($10\,\mu mol^{-1}$ carbamylcholine). Stimulation with a 4 s depolarizing step first induces persistent spontaneous firing (cell was initially quiescent at a stable resting potential of −64 mV). Five more episodes of stimulation each give rise to a further increase in firing frequency, in each case stably expressed until the next stimulation. Top trace shows membrane voltage, middle trace shows injected current, and bottom trace shows frequency of firing. Reproduced from Egorov AV, Hamam BN, Fransen E, Hasselmo ME, and Alonso AA (2002) Graded persistent activity in entorhinal cortex neurons. *Nature* 420: 173–178.

can produce a sustained increase in pacemaking frequency that can last for many minutes. Moreover, this process can be repeated with multiple stimuli, with pacemaking frequency increasing step-wise over a wide range. The increase in frequency can be reversed in a step-wise manner by application of 6 s hyperpolarizing steps that temporarily silence the cell. (Note that the increase in firing frequency following periods of stimulation and the decreases following inhibition are opposite to the plasticity in medial vestibular neurons.) This plasticity constitutes a mechanism of short-term analog memory at a single-cell level. The mechanism is not completely understood but appears to involve increased calcium entry during the period of stimulation that may activate calcium-activated nonselective cation conductances.

Another prominent instance of plasticity of pacemaking behavior occurs in the neurons controlling electric organ discharge of the weakly electric fish *Apteronotus leptorhynchus*. These fish generate an oscillating electric field around their bodies, used for detecting objects and other organisms in muddy water. The oscillatory discharge of electric organ is controlled by pacemaker neurons that fire action potentials at up to 1 kHz, the highest frequency of electrical pacemaking that is known. Because the cells in the pacemaker nucleus are electronically coupled, they have been difficult to study under voltage clamp, and so the ionic conductances underlying pacemaking are not well-characterized. Most likely, pacemaking is generated by TTX-sensitive sodium currents with unusually rapid kinetics, together with rapidly activating and rapidly deactivating potassium channels that keep the spikes short. Pacemaking is extremely regular. However, when two fish using similar frequencies of electric organ discharge share the same area, the one using the higher frequency shifts its frequency higher to avoid jamming of the two signals. This change in pacemaker frequency is normally temporary, lasting only as long as the other signal is present. However, if the exogenous signal is maintained for >30 min, the increase in pacemaking frequency can last for many hours after the other signal is terminated. The mechanism of this plasticity is not understood.

Neurons in the suprachiasmatic nucleus of mammals are the 'master pacemaker' neurons for circadian rhythms in behavior. They fire spontaneously (although sometimes with less regularity than seen in other types of electrical pacemakers), and if the firing of individual neurons is followed, the frequency of spontaneous firing shows a circadian rhythm, higher in the daytime and slower at nighttime (even in nocturnal animals like rats and mice that are less active in the day than in the night). Intensive work over the last decade has identified a molecular clock based on feedback loops of transcription factors and enzymes, involving regulation of translation, dimerization, nuclear translocation, and degradation of multiple elements. However, how the molecular clock is linked to the circadian regulation of electrical pacemaking frequency has been very difficult to determine. Day–night differences in a number of different conductances have been described, including voltage-dependent L-type calcium channels, Kv3-family voltage-dependent potassium channels, BK calcium-activated potassium

channels, and background potassium channels, but it has been difficult to definitively ascribe differences in firing frequency to any of the changes in conductance that have been described.

Significance of Pacemaking

The central nervous system is characterized by overall oscillatory electrical activity, as is evident from the oscillatory signals at many different frequencies present in the electroencephalogram. Oscillatory electrical activity occurs at many different levels, including the pacemaking activity of single neurons, oscillations in local networks, and oscillations involving larger-scale connections between brain regions. Oscillatory activity does not require sensory input but rather appears to occur spontaneously (although it can certainly be modified and influenced by sensory input). Presumably neuronal pacemaking at the single-cell level is central in generating the overall spontaneous activity of the brain, amplified and modified by circuits and networks.

Pathophysiology of Pacemaking

It is striking that so many neurons in the motor system are characterized by rhythmic spontaneous firing. Pacemaking activity seems counterintuitive for a system which must often be inactive and in which very rapid events and processing must occur in almost an all-or-none manner. It seems likely that the underlying pacemaking activity of elements of the motor system plays a major role in the generation of oscillatory tremors that characterize diseases such as Parkinson's disease and Huntington's disease. It is not yet clear whether the pathophysiology of tremors involves changes in the pacemaking activity of single neurons, changes in control and circuit connections, or both. It is interesting that a current treatment for Parkinson's disease involves deep brain stimulation of neurons in the subthalamic nucleus, which show prominent pacemaking activity, but the exact mechanism by which deep brain stimulation works is not yet known.

Altered pacemaking mechanisms may also be involved in some other diseases of the nervous system involving hyperexcitability of neurons, including neuropathic pain (some forms of which might involve spontaneous firing of primary pain-sensing neurons, which normally would require input from sensory receptors to be active) and epilepsy.

Summary

In some neurons, electrical pacemaking activity produces spontaneous activity in the absence of any synaptic or sensory input. Such pacemaking activity at the cellular level probably underlies spontaneous oscillatory activity at the level of networks and brain regions, which appears to be a crucial characteristic of the operation of the mammalian brain. There are several different ionic mechanisms of pacemaking in different types of pacemaking neurons, most commonly depending on activation of ion channels that allow entry of either sodium ions or calcium ions into the cell, thus carrying an electrical current that depolarizes the cell membrane to the threshold for firing an action potential.

See also: Calcium and Signal Transduction; Calcium Channels; Calcium Channel Subtypes Involved in Neurotransmitter Release; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Gamma-Aminobutyric Acid (GABA); Large Conductance Calcium-Activated Potassium Channels; Neuromodulation of Sodium Channels; Neuromodulation; Potassium Channel Regulation; Sodium Channels.

Further Reading

Bennett BD, Callaway JC, and Wilson CJ (2000) Intrinsic membrane properties underlying spontaneous tonic firing in neostriatal cholinergic interneurons. *Journal of Neuroscience* 20: 8493–8503.

Chaplan SR, Guo HQ, Lee DH, et al. (2003) Neuronal hyperpolarization-activated pacemaker channels drive neuropathic pain. *Journal of Neuroscience* 23: 1169–1178.

Do MT and Bean BP (2003) Subthreshold sodium currents and pacemaking of subthalamic neurons: modulation by slow inactivation. *Neuron* 39: 109–120.

Egorov AV, Hamam BN, Fransen E, Hasselmo ME, and Alonso AA (2002) Graded persistent activity in entorhinal cortex neurons. *Nature* 420: 173–178.

Forti L, Cesana E, Mapelli J, and D'Angelo E (2006) Ionic mechanisms of autorhythmic firing in rat cerebellar Golgi cells. *Journal of Physiology* 574: 711–729.

Hallworth NE, Wilson CJ, and Bevan M (2003) Apamin-sensitive small conductance calcium-activated potassium channels, through their selective coupling to voltage-gated calcium channels, are critical determinants of the precision, pace, and pattern of action potential generation in rat subthalamic nucleus neurons *in vitro*. *Journal of Neuroscience* 23: 7525–7542.

Hille B (2001) *Ion Channels of Excitable Membranes*. Sunderland, MA: Sinauer.

Jackson AC, Yao GL, and Bean BP (2004) Mechanism of spontaneous firing in dorsomedial suprachiasmatic nucleus neurons. *Journal of Neuroscience* 24: 7985–7998.

Liss B, Franz O, Sewing S, Bruns R, Neuhoff H, and Roeper J (2001) Tuning pacemaker frequency of individual dopaminergic neurons by Kv4.3L and KChip3.1 transcription. *EMBO Journal* 20: 5715–5724.

Maccaferri G and McBain CJ (1996) The hyperpolarization-activated current (Ih) and its contribution to pacemaker activity in rat CA1 hippocampal stratum oriens-alveus interneurones. *Journal of Physiology* 497(Pt. 1): 119–130.

Meredith AL, Wiler SW, Miller BH, et al. (2006) BK calcium-activated potassium channels regulate circadian behavioral

rhythms and pacemaker output. *Nature Neuroscience* 9: 1041–1049.

Neuhoff H, Neu A, Liss B, and Roeper J (2002) *I*(h) channels contribute to the different functional properties of identified dopaminergic subpopulations in the midbrain. *Journal of Neuroscience* 22: 1290–1302.

Poolos NP (2004) The yin and yang of the H-channel and its role in epilepsy. *Epilepsy Currents* 4: 3–6.

Puopolo M, Raviola E, and Bean BP (2007) Roles of subthreshold calcium current and sodium current in spontaneous firing of mouse midbrain dopamine neurons. *Journal of Neuroscience* 27: 645–656.

Robinson RB and Siegelbaum SA (2003) Hyperpolarization-activated cation currents: From molecules to physiological function. *Annual Review of Physiology* 65: 453–480.

Smith GT and Zakon HH (2000) Pharmacological characterization of ionic currents that regulate the pacemaker rhythm in a weakly electric fish. *Journal of Neurobiology* 42: 270–286.

Surmeier DJ, Mercer JN, and Chan CS (2005) Autonomous pacemakers in the basal ganglia: Who needs excitatory synapses anyway? *Current Opinion in Neurobiology* 15: 312–318.

Taddese A and Bean BP (2002) Subthreshold sodium current from rapidly inactivating sodium channels drives spontaneous firing of tuberomammillary neurons. *Neuron* 33: 587–600.

Wilson CJ and Callaway JC (2000) Coupled oscillator model of the dopaminergic neuron of the substantia nigra. *Journal of Neurophysiology* 83: 3084–3100.

Wolfart J, Neuhoff H, Franz O, and Roeper J (2001) Differential expression of the small-conductance, calcium-activated potassium channel SK3 is critical for pacemaker control in dopaminergic midbrain neurons. *Journal of Neuroscience* 21: 3443–3456.

Neuronal Plasticity after Cortical Damage

B Kolb, University of Lethbridge, Lethbridge, AB, Canada

© 2009 Elsevier Ltd. All rights reserved.

Measuring Brain Plasticity

Although brain plasticity is presumed to reflect changes in the synaptic organization of the brain, it can be inferred from the measurement of change at many levels of analysis. Perhaps the most obvious level is behavior. If an organism's behavior is altered, such as when a new habit or bit of information is learned, we can infer that there must be some persistent change in the nervous system that represents the habit or information. The identification of this plastic change may be seen by utilizing a range of techniques that run from global to molecular analyses of brain organization. One of the most global measures is the analysis of brain maps that can be demonstrated either by recording electrical activity of the cortex in response to external stimulation (e.g., sound, touch, visual patterns), by stimulating the cortex to elicit movements or perceptions, or less invasively by using *in vivo* imaging techniques such as functional magnetic resonance imaging (fMRI). For example, it has been known for over 50 years that stimulation of the cerebral cortex of humans or lab animals with a mild electric current will elicit movements of different body parts such that a map (in humans known as an homonculus) can be identified. Stimulation of specific points will elicit movements of the digits, leg, face, etc. As animals learn new motor skills, the details of the maps change, thus representing the newly learned skill. Thus, if a person learns to play a musical instrument, we can expect that the representation of the digits will expand to reflect the newly learned ability to make fine and rapid digit movements.

Changes in the maps must reflect changes in the synaptic organization of the underlying cerebral tissue. Such synaptic changes can be shown by using specific stains to identify the structure of the neurons in the region of interest. One procedure, known as a Golgi stain, allows the dendritic fields of the neurons to be visualized and measured. Because most synapses are found on the dendrites and because space on dendrites is inevitably filled with synapses, we can measure the amount of dendritic space and infer synapse number. This is much like estimating the number of leaves on a tree by measuring branch length and the typical distance between the leaves. Simple arithmetic would allow an estimate of the total number of leaves. A similar logic allows synapse number to be estimated on the dendrites visualized in Golgi stains. As animals learn new tasks, there are changes in dendritic organization that reflect the experience-dependent behavioral changes. But this method only allows an inference about synaptic change. In order to confirm that synapses really have changed, one must use electron microscopic (EM) techniques. EM analyses make it possible to visualize individual synapses, to identify synaptic type (excitatory vs. inhibitory), and to count the number of each type of synapse. Such analyses are laborious and time consuming, however, and not practical for most behavioral studies.

If new synapses are formed in response to some experience, there must be some molecular mechanism that underlies the change. Thus, if new synapses are formed, proteins must be manufactured by cells, and the instructions for this manufacture will ultimately be controlled by genes. It is possible to measure both protein and gene expression in brain by using molecular probes, which provide yet another measure of brain plasticity. Finally, although thought to be relatively uncommon, brain plasticity can also result from the production of new neurons (known as neurogenesis) that become integrated into the host brain. Specific markers can be used to identify the birth of such new cells. The best studied region for these new cells is the hippocampus, although they may occur in small numbers elsewhere in the human brain.

In sum, brain plasticity can be measured at many levels. Although the ultimate measures of plasticity are molecular, it is not realistic to begin there. The place to begin is with detailed analysis of behavior followed by global measures such as topographic maps, where relevant, and measures of synaptic change by using stains for pre- or postsynaptic processes.

Behavioral Change after Brain Injury

We can identify at least three outcomes that could occur after brain injury. First, there may be compensation resulting from an adaptation to the lost function. Compensation could reflect a change in strategy or it may represent a substitution of a new behavior for the lost one. For example, a person who has a deficit in moving the eyes can be expected to have difficulty in reading. This difficulty can be solved, however, by moving the head, which allows restitution of function (reading) by substitution of one behavior for another. There are unlikely to be many plastic changes in the cortex to produce this behavioral change because the needed circuits are presumably already present. Similarly, if, following a stroke, a person can walk only with the aid of a cane, the

ability to walk has returned but the original walking ability has not returned, just an adaptation to the behavioral difficulty.

A second outcome might be partial restitution of the original behavior. This could reflect recovery from some sort of nonspecific effect of the injury, such as swelling, or it may reflect plastic changes that allow the brain to partly restore the disturbed behavior. Such plastic changes are most clearly seen in people with motor disturbances. Individuals may be unable to make movements of the digits or to organize their movements into purposeful sequences, but over time the capacity partially returns. Mapping studies using *in vivo* imaging have shown that the partial return of function is related to an expansion of the remaining motor cortical regions, as well as the recruitment of more distal regions (especially in the intact hemisphere) to control the movements. Animal studies have shown correlated changes in dendritic organization that presumably underlie the changes in map topography.

Third, there is the possibility that there could be a complete restitution of the original behavior. Although it is at least theoretically possible for lost functions to completely return in a plastic brain, careful behavioral analysis shows that this is rarely, if ever, the case in either adult laboratory animals or human patients. In principle, however, it ought to be possible to reorganize the brain to allow at least some functions to completely recover. This process is most likely in the developing brain, and it has been documented for over 100 years that children with damage to language areas can show almost complete recovery of lost language functions.

Neuronal Adaptations to Brain Injury

When the brain is damaged, there is a loss of neurons in the injured area, but, in addition, many neurons that are not directly injured lose their normal synaptic inputs. The immediate effect of this synaptic loss is (1) neuronal death, (2) atrophy of the dendritic and/or axonal fields that have lost their connections, and/or (3) subsequent reorganization of the connections. In principle, there are three ways that the brain could make changes that might support recovery. First, there could be changes in the organization of the remaining intact circuits in the brain. The general idea is that the brain could reorganize in some way to do 'more with less.' It is unlikely that a complexly integrated structure like the cerebral cortex could undergo a wholesale reorganization of cortical connectivity, but rather, recovery from cortical injury would be most likely to result from a change in the intrinsic organization of local cortical circuits in regions directly or indirectly disrupted by the cortical injury. Although it might be possible to produce significant reorganization of cortical connections in the young brain, the overwhelming evidence in experimental animals is that this is rare and it is just as likely to be associated with abnormal functioning as with recovery.

Second, there could be a generation of new circuitry. Cerebral reorganization can be stimulated either by some sort of behavioral therapy or the application of some sort of pharmacological treatment. In either event, the stimulus could influence reparative processes in the remaining brain or could enhance the production of new circuitry. Once again, it seems most likely that the induced neuronal changes would be in the intrinsic organization of the cortex. One might predict that induced neuronal changes might be more extensive than in the case of endogenous change, in part because the treatment can act upon the whole brain rather than just on affected regions.

Third, there could be a generation of new neurons and glia to replace at least some lost cells. Stem cells that give rise to the neurons and glia of the brain remain active in the brain throughout life. It is thus possible that neurogenesis could be stimulated after injury, especially during development, and that these new neurons could replace those lost to injury or disease. All three of these possibilities for cerebral plasticity do occur, but the likelihood of each type of change varies with age at injury.

Plasticity after Adult Brain Injury

The hippocampal formation of adult rats provided one of the first models to study synaptic replacement. When input to the hippocampus is reduced by damage to the major input pathway via the entorhinal cortex, there is a massive denervation of the hippocampal formation, which leads to atrophy of the dendritic fields of the hippocampal cells in the week after injury. Beginning at about 7–10 days after the injury, there is a slow regrowth of the dendritic fields, although the pattern of regrowth is different than the original pattern. This regrowth reflects a reorganization of the circuitry and it is correlated with a partial recovery of cognitive functions. The restructuring of the synaptic organization of the hippocampus occurs primarily because the remaining inputs to the hippocampus expand to occupy the regions of lost inputs.

An analogous process appears to occur after restricted neocortical injury in adulthood. Frontal lesions in rats lead to atrophy of neurons in adjacent cortical regions, but by 30 days postinjury the atrophy begins to reverse, and by 3 months the affected neurons actually have more synaptic space than

normal. Again, the dendritic change is correlated with partial recovery of cognitive tasks associated with the frontal injury. More recently, there has been much research on the effect of unilateral motor cortex injury on adjacent motor regions, contralateral motor regions, and the striatum. The general finding is that there is an initial atrophy of affected regions on the ipsilateral hemisphere, followed by a compensatory reorganization of neural networks that can be seen in dendritic organization in both the cortex and striatum. Some studies have shown increased dendritic fields in the motor regions of the intact hemisphere as well, which might be predicted from the *in vivo* imaging studies in humans. One unexpected finding of the studies is that lesion etiology appears to interact with the synaptic plasticity. It appears that the way the cortical tissue is killed may somehow affect the molecular cascade that subsequently leads to the synaptic reorganization.

Little is known about the changes in gene or protein expression that give rise to the synaptic changes after cortical injury but there are large changes in the expression of various growth factors, such as fibroblast growth factor-2 (FGF-2), that likely play some role in stimulating the synaptic remodeling.

Plasticity after Infant Brain Injury

Much more is known about plasticity following injury to the developing brain. Damage to various cortical areas in the second week of life in rats has been shown to lead to good functional recovery in adulthood, and this recovery is correlated with expansion of dendritic fields across the cortical mantle. This dendritic expansion takes several weeks to develop, however, and as the dendritic fields reorganize, there is an associated improvement in a variety of behavioral measures. In contrast, damage to cortical regions in the first week of life in rats produces atrophy of cortical neurons and this atrophy does not recover, nor does the lost behavior.

Infant animals show changes not only in local cortical circuits but also in more distant cortical connectivity. Perhaps the most extensive studies of changes in cortical connectivity are those showing that after unilateral motor cortex lesions in infant rats or cats there is a major expansion of the ipsilateral corticospinal pathway from the undamaged hemisphere, which is correlated with partial recovery of skilled forelimb use. These expanded connections typically do not reflect the growth of new pathways, however; rather, they reflect the failure of superfluous pathways to die off during development. That is, normally in development, there is a period during which many pathways are eliminated, but after some types of brain injury these pathways are maintained. The presence of these abnormal pathways may support some types of functional recovery, but it may also interfere with the normal functioning of remaining cortical functions. This possibility has been termed 'crowding' to reflect the idea that the normal functions of a cortical region can be crowded out by the development of abnormal connections. Thus, there may be recovery of some types of behaviors, such as motor functions, at the expense of other functions. This appears to be especially likely if the injury occurs during the period of neuronal migration and unlikely if it occurs during the period of maximal synaptogenesis.

As noted earlier, synaptic change may also occur after injury because the brain generates new neurons. Although this is likely to be rare in the adult, it does occur during development, but the generation of these neurons varies with precise embryonic age and location of injury. If the brain is damaged prenatally during the time of neuronal generation, the brain can compensate by generating rather large numbers of new cells that appear to support at least partial restitution of function. If the developing brain is damaged postnatally, there can also be extensive neurogenesis, although it appears largely limited to injury during the second week of life and is most extensive when the injury is to the certain regions of the midline telecephalon, namely, the olfactory bulb or midline cortex. The newly generated tissue is functional and animals regain at least partial function.

As in adult injury, little is known about the changes in gene or protein expression that give rise to the synaptic changes after cortical injury in infancy. There are increases in FGF-2 that correlate with functional outcome, but the exact role of the growth factors in synaptic reorganization and neurogenesis is poorly understood.

Factors Influencing Recovery and Plasticity after Cortical Injury in Adulthood

Given that a wide range of factors can stimulate synaptic plasticity in the normal brain, it is not surprising that brain plasticity and behavior after cortical injury can be modified by many factors. The most powerful influence appears to be complex housing. Studies have shown consistently that placing experimental animals in complex environments in which there are conspecifics to interact with, novel objects to explore and interact with, and an opportunity for motor activity in complex terrain produces widespread changes in dendritic organization, endogenous growth factor production, and gene expression. Animals with cerebral injuries clearly benefit behaviorally from such experience, and there is a

correlated reorganization of neuronal networks. A key feature of complex housing is that such experience produces extensive changes in the brain of laboratory animals; but what is the equivalent treatment in humans? The best guess is some type of multidisciplinary rehabilitation program that is relatively intense and thus lasts several hours a day. Such treatments may not be practical for many patients who may tire easily or have other ailments that make the intensity difficult. Other treatments thus may provide more viable options. The most common option is rehabilitation training, which is probably the most common treatment in human patients. Experimental studies in lab animals have shown functional improvement correlated with shifts in the organization of cortical motor maps with rehabilitation training, but rehabilitation training is generally only effective with fairly restricted brain injuries.

One type of treatment that has been tried is the use of psychomotor stimulants such as amphetamine and nicotine. These stimulants produce large changes in synaptic organization in the normal frontal cortex, the effects being more widespread with nicotine. The administration of stimulants after brain injury is presumed to stimulate synaptic plasticity that can be co-opted by the damaged brain to facilitate recovery. Because stimulants also increase the endogenous production of neurotrophic factors such as FGF-2, it might be expected that neurotrophic factors might also facilitate reactive synaptic plasticity after brain injury and that is the case. Both FGF-2 and nerve growth factor (NGF) stimulate synaptic plasticity and functional recovery after cortical injury.

There are a few studies looking at other types of factors including (1) inosine, which is a compound that enhances the production of connections from the injured hemisphere across to the damaged one, and (2) antibodies to a protein known as NoGo, which untreated acts to prevent synaptic plasticity. Both treatments are correlated with increased synapse formation in the remaining brain.

Factors Influencing Recovery and Plasticity after Cortical Injury in Infancy

It is generally assumed that the developing brain will show more plasticity after injury than the adult brain and, similarly, that plasticity in the injured developing brain should be influenced more by factors than the adult brain. This is generally true, although the precise age at the time that the treatments are initiated does interact with the extent of observed plasticity. As a general rule of thumb, the sooner the treatments are initiated, the larger the plastic changes (and correlated recovery) will be.

As in adults, complex housing has a very large effect on both synaptic organization and behavior. Another treatment that is nearly as effective is tactile stimulation with a soft brush, typically lasting about 15 min several times a day. As little as 2 weeks of such stimulation after perinatal injury facilitates recovery measured in adulthood and produces large increases in synapse numbers relative to control animals. One reason for this large effect appears to be related to the stimulation of FGF-2 in the skin. This FGF-2 travels through the bloodstream to the brain and acts to enhance plasticity. Direct administration of FGF-2 can act in a similar fashion and under some circumstances can stimulate neurogenesis and the partial regeneration of injured cortical tissue.

But not all treatments are beneficial after early injury. Prenatal stress or prenatal selective serotonin reuptake inhibitors (SSRIs) can interfere with reactive plasticity after perinatal injury and animals show severe behavioral impairments in adulthood. The fact that prenatal experiences can alter later brain plasticity suggests that the experiences alter gene expression, which later interacts with the brain's capacity to generate plastic changes to facilitate functional recovery.

Summary

In summary, the brain is capable of significant synaptic reorganization in response to a wide range of experiences, including injury. Plasticity can be measured at many levels ranging from behavior to genes. As methods of analysis are refined, we can anticipate that the mechanisms underlying brain plasticity and recovery of function after injury will become better understood and will allow the development of better treatments for brain injury throughout the life span.

See also: Adult Cortical Plasticity; Cell Replacement Therapy: Mechanisms of Functional Recovery; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Spinal Cord Regeneration and Functional Recovery: Strategies; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Plasticity: Neuronal Sprouting.

Further Reading

Grossman AW, Churchill JD, Bates KE, Kleim JA, and Greenough WT (2002) A brain adaptation view of plasticity: Is synaptic plasticity an overly limited concept? *Progress in Brain Research* 138: 91–108.

Hensch TK (2004) Critical period regulation. *Annual Review of Neuroscience* 27: 549–579.

Horn G (2004) Pathways of the past: The imprint of memory. *Nature Reviews Neuroscience* 5: 108–120.

Johnston MV (2003) Brain plasticity in paediatric neurology. *European Journal of Paediatric Neurology* 7: 105–113.

Kolb B (1995) *Brain Plasticity and Behavior*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kolb B and Whishaw IQ (1998) Brain plasticity and behavior. *Annual Review of Psychology* 49: 43–64.

Neville H and Bavelier D (2002) Human brain plasticity: Evidence from sensory deprivation and altered language experience. *Progress in Brain Research* 138: 177–188.

Nudo RJ (2003) Adaptive plasticity in motor cortex: Implications for rehabilitation after brain injury. *Journal of Rehabilitation Medicine* 41: 7–10.

Payne BR and Lomber SG (2001) Reconstructing functional systems after lesions of cerebral cortex. *Nature Reviews Neuroscience* 2: 911–919.

Shaw CA and McEachern JC (2003) *Toward a Theory of Neuroplasticity*. New York: Elsevier.

Van der Kooy D and Weiss S (2000) Why stem cells? *Science* 287: 1439–1441.

Wolf ME, Sun X, Mangiavacchi S, and Chao SZ (2004) Psychomotor stimulants and neuronal plasticity. *Neuropharmacology* 47: 61–79.

Woo NH and Lu B (2006) Regulation of cortical interneurons by neurotrophins: From development to cognitive disorders. *Neuroscientist* 12: 43–46.

Neuron–Glia Coupling in Glutathione Metabolism

R Dringen, University of Bremen, Bremen, Germany; and Monash University, Clayton, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The tripeptide glutathione (GSH; γ-L-glutamyl-L-cysteinylglycine) is present in all cells in millimolar concentrations. GSH has essential cellular functions as an important antioxidant. In addition, it serves as cofactor for some enzymes, is a storage and transport form of the amino acid cysteine, and keeps the cellular thiol reduction potential in a strongly reduced form, thus protecting essential thiol groups in enzymes against oxidation. More recently, GSH has been demonstrated to be involved in cellular regulation, since the activity of proteins can be modulated by glutathionylation. Thus, the availability of a sufficiently high concentration of GSH is essential for a variety of cell functions.

The antioxidative function may be the most important activity of GSH, since reactive oxygen species (ROS), such as radicals and peroxides, are continuously generated in cells. ROS generation is counterbalanced by an appropriate antioxidative defense to prevent cell damage. The consequence of a disturbance of this balance of oxidative and antioxidative processes by insufficient antioxidative defense and/or an increased rate of ROS generation is oxidative stress. The human brain utilizes about 20% of the oxygen consumed by the body, indicating that this organ has the potential to generate ROS in high quantities. Thus, a disturbance in the homeostasis and the metabolism of the antioxidant GSH in brain increases the risk for oxidative stress and therefore has been connected with oxidative damage in neurodegenerative diseases such as Parkinson's disease (PD) and Alzheimer's disease.

In the brain, especially the neurons appear to be handicapped regarding GSH metabolism, since the availability of extracellular cysteine as GSH precursor limits the synthesis of the tripeptide. Since neurons cannot use extracellular cystine as GSH precursor, these cells rely on the supply of cysteine from neighboring cells. Astrocytes are well known for their broad range of metabolic pathways and for their function in supplying various essential metabolites, including GSH precursors, to neighboring neurons. In the following sections, the current knowledge on the mechanisms involved in the supply of GSH precursors from astrocytes to neurons is summarized, and consequences of a disturbance of this metabolic cooperation are considered in relationship to the pathology of PD.

Metabolism of GSH in Brain Cells

Synthesis of GSH

GSH synthesis in brain cells follows the pathway known to function in other tissues. GSH is synthesized by consecutive reactions of two enzymes, γ-glutamylcysteine (γGluCys) synthetase and GSH synthetase (Figure 1). γGluCys synthetase generates γGluCys from glutamate and cysteine. Catalyzed by GSH synthetase, this dipeptide reacts with glycine to GSH. ATP is cosubstrate for both enzymes. GSH synthesis is strongly regulated by a feedback inhibition of γGluCys synthetase by the end product GSH.

Cellular GSH synthesis depends on the intracellular availability of the substrates glutamate, cysteine, and glycine (Figure 1). Since glutamate and glycine are neurotransmitters and cysteine in high concentrations is toxic to neurons, the extracellular availability of these amino acids in the brain is limited. Thus, GSH synthesis in brain cells depends on their ability to take up suitable extracellular GSH precursors and/or to synthesize sufficient amounts of the substrates for GSH synthesis (Figure 1). Uptake and metabolism of GSH precursors differ among brain cell types. In contrast to neurons, astrocytes can use a wide range of amino acids and peptides that are taken up by the respective transporters and subsequently converted to the substrate amino acids of the two enzymes of GSH synthesis. In neurons the availability of cysteine limits GSH synthesis. These cells do not use extracellular cystine as source for the GSH substrate cysteine and prefer glutamine as the best extracellular precursor for the glutamate moiety of GSH. In contrast, astrocytes prefer glutamate and cystine as extracellular GSH precursors. This differential preference for extracellular GSH precursors prevents competition between neurons and astrocytes. However, since neurons require cysteine for GSH synthesis, these brain cells depend for their GSH synthesis on the supply of cysteine or a cysteine precursor from neighboring astrocytes.

Metabolism of GSH

During detoxification of peroxides GSH is oxidized to glutathione disulfide (GSSG; Figure 2). This reaction is catalyzed by glutathione peroxidase (GPx). Several GPx isoenzymes have been reported that differ in their tissue expression and in substrate specificity.

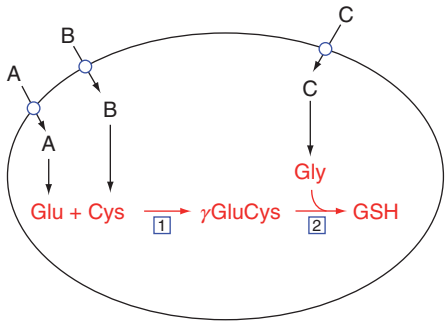

Figure 1 Cellular synthesis of glutathione (GSH). GSH is synthesized in the consecutive reactions of γGluCys synthetase (1) and GSH synthetase (2). Glutamate, cysteine, and glycine are the intracellular substrates for GSH synthesis and may be derived from the extracellular precursors A, B, and C, respectively.

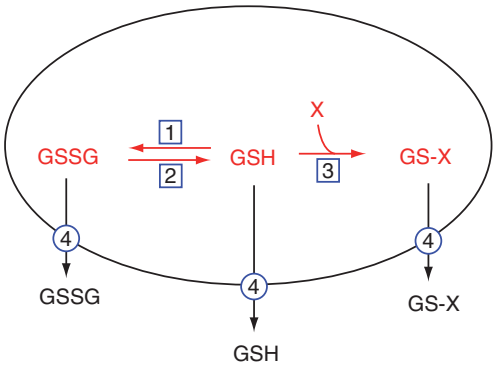

Figure 2 Metabolism of glutathione (GSH). In cells, GSH is oxidized during peroxide disposal by glutathione peroxidase (1) to GSSG (two molecules of disulfide-bonded GSH) and is regenerated from GSSG by glutathione reductase (2). In addition, GSH can react (3) in glutathione-*S*-transferase-catalyzed reactions with xenobiotics and endogenous compounds (X). Members of the transporter family of multidrug-resistance proteins mediate (4) the export of GSH, GSSG, and glutathione-*S* conjugates from the cells.

In addition, GSH is oxidized by chemical reactions with radicals. The product of GSH oxidation is glutathione disulfide. Within cells, GSH is regenerated from GSSG by glutathione reductase (GR) (Figure 2). This enzyme uses NADPH as electron donor. Thus, efficient regeneration of GSH from GSSG requires sufficient supply of the reduced coenzyme NADPH. By the intracellular GSH redox cycling of GPx-mediated GSH oxidation and GR-catalyzed regeneration of GSH from GSSG (Figure 2), the total amount of cellular GSH is not substantially altered. Reduction of GSSG by GR is highly efficient due to the low K_m values of GR for GSSG and NADPH. Therefore, under unstressed conditions GSSG is hardly detectable in cells. In contrast, under oxidative stress GR can become rate limiting for GSH redox cycling, as indicated by accumulation of GSSG and a corresponding shift in the thiol reduction potential of the stressed cells.

In contrast to the GSH redox cycling, cells lose GSH by the formation of glutathione-*S* conjugates (GS-X), a reaction catalyzed by glutathione-*S*-transferases (GSTs), and by export of GSH or GSSG (Figure 2). Such processes participate in the consumption of GSH in brain cells. Members of the family of multidrug-resistance proteins (MRPs) mediate the cellular export of GSH and its conjugates. MRPs are ATP-driven export pumps of organic anions and belong to the family of ATP-binding cassette transporters. In addition to MRPs, other transporters, such as organic anion transporters (OATPs) or the cystic fibrosis transductance regulator (CFTR), have also been connected with GSH transport processes.

Delivery of GSH Precursors from Astrocytes to Neurons

The mechanisms involved in the transport of sulfur-containing amino acids in the brain are still a matter of debate. Nevertheless, within the brain the availability of cysteine is crucial for GSH synthesis in neurons. Since the presence of co-cultured astrocytes maintains and even increases GSH content in neurons, the amino acid cysteine that limits neuronal GSH synthesis is supplied by neighboring astrocytes. The mechanism involved in this cysteine supply involves efficient export of GSH from astrocytes, the cleavage of extracellular GSH by the ectoenzymes γ-glutamyltranspeptidase (γ-GT) and aminopeptidase N, and the uptake of cysteine into neurons (Figure 3).

Export of GSH

Release of GSH from brain cells has been reported so far only for astrocytes. These cells release, under unstressed conditions, substantial amounts of GSH and even protect the GSH exported against oxidation by a superoxide dismutase-like factor that is released into the medium. Cultured astrocytes export, within 1 h, about 10% of their intracellular GSH. A variety of transporters have been connected with cellular GSH transport. Several MRP family members are known to be expressed on the mRNA level in cultured astrocytes from mouse or rat brain. For MRP1, the presence of the protein has also been demonstrated for astrocytes *in vitro* and *in vivo*. This transporter mediates about 60% of GSH export from astrocytes. The transporter responsible for the residual 40% of MRP1-independent GSH export has not been identified so far. However, the use of astrocytes from MRP-deficient mice and inhibitor studies suggest that the MRPs 4 and 5, the OATPs 1 and 2, and CFTR do not contribute to the residual 40% GSH export from MRP1-deficient astrocytes.

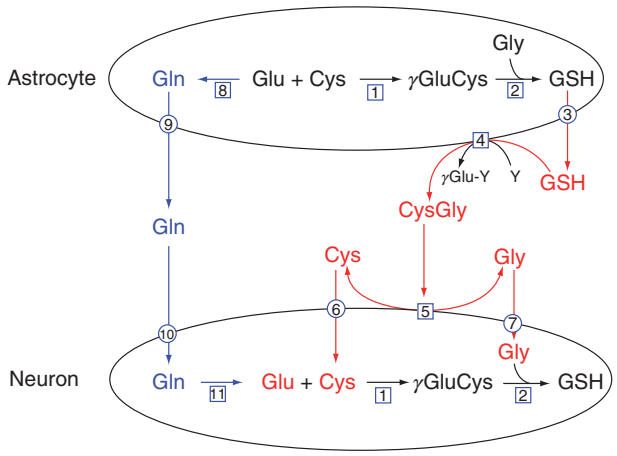

Figure 3 Metabolic interaction between astrocytes and neurons in glutathione (GSH) metabolism. In astrocytes and neurons, GSH is synthesized from glutamate (Glu), cysteine (Cys), and glycine (Gly) in the consecutive reactions of γGluCys synthetase (1) and GSH synthetase (2). The GSH exported from astrocytes (3) by multidrug-resistance protein 1 or other transporters is substrate for the astrocyte ectoenzyme γ-glutamyltranspeptidase (4). This enzyme transfers the γ-glutamyl (γGlu) moiety from GSH onto an acceptor Y. The dipeptide CysGly generated by the γ-glutamyltranspeptidase reaction is hydrolyzed by aminopeptidase N (5), and the generated amino acids cysteine and glycine are taken up into neurons by the excitatory amino acid carrier 1 (6) and sodium-dependent glycine transporters (7), respectively. In addition, glutamine (Gln) is provided by astrocytes and serves as extracellular precursor for the glutamate required for neuronal GSH synthesis. In astrocytes, glutamate is amidated by glutamine synthetase (8) to glutamine, which is released from astrocytes by glutamine transporters (9). Extracellular glutamine is taken up into neurons (10) and subsequently hydrolyzed by phosphate-activated glutaminase (11). Consequently, all three amino acids that are required for GSH synthesis in neurons are provided as extracellular precursors by neighboring astrocytes.

Extracellular Processing of GSH

Extracellular processing of the GSH exported from astrocytes is a key requirement for the metabolic interaction between astrocytes and neurons in GSH metabolism of the brain. The enzymes involved in this process are the astrocytic γ-GT and a neuronal ectopeptidase (Figure 3).

γ-GT catalyzes the transfer of the γ-glutamyl moiety of GSH onto an acceptor amino acid, as well as the hydrolysis of GSH to glutamate and the dipeptide CysGly. Thus, the astrocytic ectoenzyme γ-GT produces the dipeptide CysGly and consumes the substrate GSH in equimolar concentrations (Figure 3). Inhibition of γ-GT by acivicin totally prevents the astrocyte-induced increase of the GSH content of co-cultured neurons, demonstrating that this enzyme plays a key role in the supply of cysteine for neuronal GSH synthesis.

Both γ-GT activity and γ-GT immunoreactivity have been reported for astrocytes *in vitro* and *in vivo*. The activity of γ-GT in cultured astrocytes is increased after treatment of the cells with lipopolysaccharide, nitric oxide (NO) donors, or thyroid hormone. In addition, a dramatic increase in the specific γ-GT activity of astrocytes is observed after application of tumor necrosis factor-α (TNF-α). This cytokine increases by stimulation of the γ-GT expression the specific activity of γ-GT by a factor of 3–4, and thereby improves the ability of the cells to process extracellular GSH.

The γ-GT-product CysGly is efficiently utilized, in micromolar concentrations, as a precursor for neuronal GSH. Inhibition of the neuronal aminopeptidase N (ApN) prevents the use of extracellular CysGly for GSH synthesis, demonstrating that intact CysGly is not taken up into neurons by a peptide transporter, but is hydrolyzed extracellularly to amino acids (Figure 3). This contrasts with the situation for astrocytes, which take up CysGly via the peptide transporter PepT2 and hydrolyze the dipeptide intracellularly.

Cysteine availability for neuronal GSH synthesis depends on astroglial GSH export and extracellular processing of GSH by γ-GT and ApN (Figure 3). With the release of the glutamate precursor glutamine by astrocytes and the extracellular generation of CysGly, all three constituent amino acids of GSH are provided to neurons by astrocytes (Figure 3). This interaction of astrocytes and neurons in GSH metabolism of the brain has been deduced from data obtained on cultures and co-cultures of brain cells. However, γ-GT-dependent generation of cysteine from GSH by brain cells has been confirmed for brain slices and for the brain, strongly suggesting that astrocytes *in vivo* also supply neurons with cysteine for GSH synthesis.

Uptake of Amino Acids into Neurons

A variety of amino acid transporters have been reported to be expressed in neurons. Most important for GSH metabolism is the uptake of the glutamate precursor glutamine, as well as uptake of cysteine and glycine into neurons (Figure 3). Glutamine is the best extracellular precursor for the glutamate used in neurons for GSH synthesis. In neurons, glutamine is deamidated by phosphate-activated glutaminase, yielding glutamate. Glycine and cysteine are substrates of sodium-dependent transporters. Since cysteine availability limits neuronal GSH synthesis, cysteine uptake into neurons is crucial for GSH metabolism of these cells. Recently it was demonstrated that deficiency of the sodium-dependent excitatory amino acid carrier 1 (EAAC1) leads to neuronal GSH deficiency, oxidative stress, and age-dependent neurodegeneration, suggesting that the cysteine for GSH synthesis in neurons is predominately taken up by EAAC1.

Implications for a Disturbed GSH Metabolism in Parkinson's Disease

Oxidative stress has been connected with physiological processes such as aging and with pathological processes in neurological disorders. A reduction in GSH content of given brain areas has been reported for PD as well as for other neurological disorders, such as schizophrenia, Alzheimer's disease, and epilepsy. Currently it is not known whether the decline in the GSH concentration in pathological brains is due to insufficient synthesis of GSH or to an uncompensated elevated consumption of GSH. Nevertheless, disturbance of the supply of cysteine from astrocytes to neurons could contribute to the lowered GSH levels of diseased brain. The best evidence for a connection between a disturbed GSH metabolism and a neurodegenerative disease has been presented for PD.

PD is characterized by a progressive degeneration of dopaminergic neurons in the substantia nigra pars compacta. The GSH content in this brain region is only 40–50% of that seen in normal controls. This and other biochemical analyses of postmortem tissues provide evidence for oxidative stress in the substantia nigra. On the cellular level, a significant loss of GSH in the surviving nigral neurons has been reported. The decline of GSH levels is an early marker for PD, since a lowered GSH level in the substantia nigra is found in incidental Lewy body disease, a presymptomatic form of PD. Nevertheless, a loss of GSH alone is not sufficient to induce nigrostriatal damage in PD, since chronic infusion of an inhibitor of γGluCys synthetase lowers GSH levels substantially but causes no decline in the number of dopaminergic neurons. Thus, GSH depletion may instead enhance the susceptibility of brain cells to harmful events such as an impairment of mitochondrial respiration. Since astrocytes in the substantia nigra also contain substantial amounts of GSH, a loss in GSH content in this brain area in PD is likely to involve astrocytes, suggesting a compromised antioxidative defense of astrocytes as well as an impairment in their supply of cysteine for neuronal GSH synthesis.

Both dopamine (DA) and TNF-α have been connected with pathological alterations observed during the progression of PD. Since both compounds also modulate GSH metabolism of astrocytes, their presence could strongly affect the cooperation between astrocytes and neurons in GSH metabolism.

Dopamine and GSH

The DA metabolism has been closely linked to an alteration in the GSH metabolism of the brain, since injection of DA in the striatum causes a decline in the GSH content in this brain region. This loss of GSH is most likely the consequence of a chemical reaction of the catecholamine DA with GSH, to form 5-glutathionyl dopamine. This reaction of GSH with DA takes also place in cultured astrocytes. Here, DA has two independent effects on the GSH metabolism, depending on the DA concentration applied. At low micromolar concentrations DA reacts with the GSH released from astrocytes in a superoxide-dependent reaction, whereas at high micromolar concentrations H_2O_2 is generated by DA autoxidation, which is followed by intracellular GPx-catalyzed oxidation of GSH and subsequent GSSG export via MRP1. Both effects of DA on the GSH metabolism of astrocytes could contribute to the observed loss of GSH in the striatum after injection of DA into this brain region. However, since the superoxide-dependent reaction of DA with GSH occurs with DA concentrations as low as 1 μM, this reaction appears to be especially important for pathological situations. The high levels of the DA conjugate of GSH found after injection of DA into the brain demonstrates that DA indeed reacts *in vivo* with GSH. Such processes also occur in pathological conditions, as demonstrated by the elevated amounts of DA conjugates in the substantia nigra of PD patients.

Extracellular GSH conjugates are converted to the corresponding cysteinyl conjugates by γ-GT and aminopeptidases. Cysteinyl conjugates of DA and its derivatives are neurotoxic. Some of these compounds inhibit the activity of complex I of the mitochondrial respiratory chain and also α-ketoacid dehydrogenase complexes. These observations are important for PD, since in the substantia nigra of PD patients a lowered activity of complex I and a reduced immunostaining for the α-ketoglutarate dehydrogenase complex are observed. Consequently, besides the impairment of the supply of the extracellular GSH precursor cysteine to neurons, the reaction of GSH with DA could generate neurotoxic compounds from DA thioethers that interfere with the energy metabolism of neurons. Both effects may contribute to the loss of nigral GSH, to the selective vulnerability of the nigrostriatal system, and to the neuronal degeneration in the substantia nigra of PD patients.

TNF-α and γ-GT

The cytokine TNF-α has been considered to participate in the pathogenesis of PD, since elevated levels of TNF-α in the substantia nigra and an increased TNF-α immunoreactivity in glial cells have been reported for this disease. Source of this TNF-α in the PD brain could be infiltrated peripheral cells as well as microglial cells and astrocytes that release, upon activation, cytokines, including TNF-α. TNF-α strongly induces

γ-GT activity in astrocytes. Thus, the elevated levels of this cytokine in PD may cause the increase in γ-GT activity that is observed for the substantia nigra in PD. Since TNF-α treatment of astrocyte cultures improves the extracellular consumption of the GSH exported by the cells, the elevation of γ-GT activity in the substantia nigra of PD brain may be beneficial for neurons by improving the availability of the extracellular GSH precursor cysteine. Nevertheless, the improved consumption of extracellular GSH could also contribute to the lowered total GSH content that is seen in the substantia nigra in postmortem brain tissue of PD patients.

γ-GT is also involved in the extracellular processing of glutathione-*S* conjugates. Thus, a TNF-α-induced elevated γ-GT activity could contribute to a quicker extracellular breakdown of compounds such as 5-glutathionyl dopamine, which is formed by the reaction of DA with GSH. Indeed, compared with controls, the substantia nigra of PD patients contains significantly higher concentrations of 5-cysteinyl dopamine. Since this compound and its derivatives have been reported to act as neurotoxins, the TNF-α-induced increase in astroglial γ-GT activity could improve the formation of toxic compounds and subsequently foster neuronal cell death.

Conclusions

GSH has essential functions in the intracellular defense of brain cells against oxidative stress. Neurons rely on the supply of GSH precursors, especially of cysteine, from neighboring astrocytes. Thus, astrocytes are highly important for the GSH metabolism and GSH homeostasis in the brain. Export of GSH from astrocytes is the first step in the supply of the GSH precursor cysteine to neurons. The extracellular GSH is processed by γ-GT and ApN to generate extracellular cysteine, which limits the synthesis of GSH in neurons. Alterations of the release rate of GSH from astrocytes, reduced activities of the ectoenzymes, and extracellular reactions of GSH, CysGly, or cysteine with catecholamines such as DA substantially disturb the availability of the GSH precursor cysteine for neurons. Such processes may contribute to a lowered antioxidative defense in neurons, to an increased susceptibility to oxidative stress, and to neuronal death in neurodegenerative diseases such as PD. On the other hand, improvement of the ability of astrocytes to synthesize, export, and/or process extracellular GSH would enhance the supply of GSH precursors to neurons and could be beneficial for the antioxidative defense of neurons in brain.

See also: Animal Models of Parkinson's Disease; Deep Brain Stimulation and Parkinson's Disease; Dopamine – CNS Pathways and Neurophysiology; Oxidative Damage in Neurodegeneration and Injury.

Further Reading

Aoyama K, Suh SW, Hamby AM, et al. (2006) Neuronal glutathione deficiency and age-dependent neurodegeneration in the EAAC1 deficient mouse. *Nature Neuroscience* 9: 119–126.

Benarroch EE (2005) Neuron–astrocyte interactions: Partnership for normal function and disease in the central nervous system. *Mayo Clinic Proceedings* 80: 1326–1338.

Bharath S, Hsu M, Kaur D, et al. (2002) Glutathione, iron and Parkinson's disease. *Biochemical Pharmacology* 64: 1037–1048.

Dringen R (2000) Metabolism and functions of glutathione in brain. *Progress in Neurobiology* 62: 649–671.

Dringen R and Hirrlinger J (2003) Glutathione pathways in the brain. *Biological Chemistry* 384: 505–516.

Dringen R, Pfeiffer B, and Hamprecht B (1999) Synthesis of the antioxidant glutathione in neurons: Supply by astrocytes of CysGly as precursor for neuronal glutathione. *Journal of Neuroscience* 19: 562–569.

Dringen R, Gutterer JM, and Hirrlinger J (2000) Glutathione metabolism in brain metabolic interaction between astrocytes and neurons in the defense against reactive oxygen species. *European Journal of Biochemistry* 267: 4912–4916.

Hirrlinger J, Schulz JB, and Dringen R (2002) Effects of dopamine on the glutathione metabolism of cultured astroglial cells: Implications for Parkinson's disease. *Journal of Neurochemistry* 82: 458–467.

Minich T, Riemer J, Schulz JB, et al. (2006) The multidrug resistance protein 1 (Mrp1), but not Mrp5, mediates export of glutathione and glutathione disulfide from brain astrocytes. *Journal of Neurochemistry* 97: 373–384.

Ruedig C and Dringen R (2004) TNFα increases activity of γ-glutamyl transpeptidase in cultured rat astroglial cells. *Journal of Neuroscience Research* 75: 536–543.

Schulz JB, Lindenau J, Seyfried J, et al. (2000) Glutathione, oxidative stress and neurodegeneration. *European Journal of Biochemistry* 267: 4904–4911.

Spencer JP, Whiteman M, Jenner P, et al. (2002) 5-*S*-Cysteinyl-conjugates of catecholamines induce cell damage, extensive DNA base modification and increases in caspase-3 activity in neurons. *Journal of Neurochemistry* 8: 122–129.

Neuron-Glia pH Regulation

J W Deitmer, Technische Universität Kaiserslautern, Kaiserslautern, Germany
M Chesler, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In nerve and glial cells, and in the extracellular spaces of nervous systems, pH changes can be evoked by neuronal activity, neurotransmitters and neuromodulators, active cellular pH regulation, metabolic processes, and secondary transporters carrying acid–base equivalents. In particular, neurotransmission mediated by glutamate, γ-aminobutyric acid (GABA), or glycine as the transmitters is associated with intra- and extracellular pH changes which can be large and rapid, with kinetics not unlike intracellular calcium transients. However, unlike intracellular calcium changes that rise and then recover, pH shifts can occur in two directions, namely in acid and alkaline directions. Indeed, since acid–base fluxes with different time courses may be elicited by a stimulus (e.g., a brief train of action potentials), the pH changes, particularly in the extracellular spaces, may be multiphasic. These multiphasic changes in extracellular pH (pH_e) may in turn affect the electrical and synaptic activity of neurons by modulating voltage- and ligand-gated ion channels and also the transport of metabolic substrates between glial cells and neurons.

The interpretation of acid–base fluxes in tissues is restricted by the methods employed to measure intra- and extracellular pH. The two standard techniques to measure pH in cells and tissues involve pH-sensitive, neutral carrier microelectrodes, which are impaled into larger cells and into extracellular spaces of a tissue and provide an 'online' recording of pH, and fluorescent probes (e.g., 2′,7′-bis (carboxyethyl)-5,6-carboxyfluorescein or seminaphtorhodafluor), which can be loaded into cultured cells or distributed in tissues and measured by fluorescence microscopy. New pH-sensitive fluorescent dyes, such as phosphoethanolamine fluorescein, DHPE, can be used to measure surface pH (i.e., on the extracellular face of the cell membrane). The pH-sensitive microelectrodes are very reliable but relatively slow, and they may cause some damage to the tissue, whereas pH-sensitive dyes are basically noninvasive and fast but more difficult to calibrate, and they bleach, which can produce radicals harmful to cells and tissues.

Rapid pH transients may be signals rather than simply resulting from inadequate homeostatic acid–base regulation. Analogous to the signaling pattern of other ions, such as calcium and potassium, transient shifts of protons and bicarbonate, together with the gas carbon dioxide and carbonic anhydrase activity, may influence or initiate functional processes in the nervous system. These include pH-induced changes in neuronal excitability, synaptic activity, gating of acid-sensing ion channels, the modulation of gap junctions (connexins) forming electrical synapses and the glial syncytium, transport processes in which acid and/or base are involved, and regulation of enzyme activities. Proton signaling in cells and local extracellular domains, particularly the microenvironment of synapses, may contribute to information processing in nervous systems.

Intracellular pH and Buffering

The free H^+ concentration in cells is in the nanomolar range, and the high buffer capacity of cells provides a reservoir of acid equivalents in the millimolar range. In other words, there is a pool of protons in rapid exchange between buffer sites and free solution, with approximately 10^5 protons being buffered for each proton in solution. The buffer capacity β is measured in millimoles of acid or base needed to shift the pH by one unit. The value β in cells is dependent on pH and on the type of buffer present. In experiments on isolated tissues or cells, β is usually measured either in the nominal absence of CO_2/HCO_3^- (the intrinsic buffer capacity, β_i, made up of mostly H^+ binding sites of cellular proteins) or in the presence of CO_2/HCO_3^-, providing the total buffer capacity, β_t. The difference between β_t and β_i gives the Co_2/bicarbonate-dependent buffer capacity β_{CO_2}.

For neurons and glial cells, β_i has been measured to be 10–40 mM, and β_{CO_2}, approximated to be 2.3 times the HCO_3^- concentration, is between 14 mM (corresponding to 6 mM HCO_3^- at a pH_i of 6.8) and 40 mM (corresponding to 17 mM HCO_3 at a pH_i of 7.2) in the presence of 5% CO_2. Hence, β_t is usually between 24 and 80 mM in nerve and glial cells, with the CO_2/HCO_3^- system contributing a large part of the pH buffering in cells and tissues. Its effectiveness, and particularly its kinetics, is dependent on the activity of carbonic anhydrase (CA), which catalyzes the reaction $CO_2 + H_2O \leftrightarrow HCO_3^- + H^+$ and hence speeds up equilibration between CO_2 and H^+/HCO_3^-. In the absence of this enzyme, the reaction is slow, and CA can accelerate the reversible conversion of CO_2 by several 100-fold, depending on the isoform and concentration of the enzyme. Within the vertebrate nervous system, CA is present in all cells, with prominent expression in oligodendrocytes and

astrocytes. Due to the rapid conversion of CO_2 to H^+ and HCO_3^- by CA, glial cells may act as sinks for CO_2, resulting in an increase in H^+ and HCO_3^- and subsequent activation of H^+ and HCO_3^- coupled transporters.

Little is known about the buffer capacity of the extracellular fluid. Nonbicarbonate buffering is believed to occur as a result of the collective contributions of inorganic phosphate and various proteins and amino acids. There is no quantitative measurement for the extracellular intrinsic buffer capacity in mammalian brain; however, in the leech central nervous system, extracellular intrinsic buffer capacity has been estimated to be approximately 5 mM. Given the small volume fraction of the extracellular compartment, the interstitial buffer reservoir per unit volume is low compared to that of the intracellular space. These two factors render the extracellular spaces a volume, in which a given amount of acid/base equivalents results in much larger pH changes than in the cytosol of the cells.

Extracellular pH and Buffering

Baseline Interstitial pH of the Brain

The pH of the brain extracellular fluid (pH_e) is determined primarily by the partial pressure of CO_2 (PCO_2) and the bicarbonate concentration. The contribution of phosphate and titratable groups on proteins to the baseline pH_e is negligible. pH_e over the long term is thus governed by the Henderson–Hasselbach equation:

$$pH = pK_a + \log[HCO_3^-]/s\ PCO_2 \qquad [1]$$

where K_a refers to the apparent dissociation constant of carbonic acid, s is the solubility coefficient of CO_2 expressed in mM/torr, and the PCO_2 is expressed in torr.

The interstitial fluid of the brain is typically more acidic than the blood and cerebral spinal fluid. A variety of mammalian studies have reported a brain pH_e ranging between 7.2 and 7.3. Given a brain PCO_2 of 46 torr, the $[HCO_3^-]_e$ is approximately 22 mM at a pH_e of 7.30. The origin of the low pH_e of the brain may be the metabolic production of CO_2, as well as the process of transporter-mediated acid extrusion from cells.

It is not widely appreciated that the extracellular fluid of a brain slice preparation is typically more acidic than both the bath solution and the corresponding interstitial fluid *in vivo*. pH_e in brain slices is often as low as 7.10–7.20. The extracellular acidosis measured in rat hippocampal slices has been attributed to an elevation of the PCO_2 and lack of non-CO_2/HCO_3^- buffers in the interstitial space. Given a typical bath PCO_2 of 40 torr (i.e., 5% CO_2), the elevation above this level in the interstitial space can be 10–20 torr for slices in an interface chamber and greater than 20 torr in a submersion-style chamber. CO_2 is an end product of aerobic respiration. Its retention in a brain slice can be explained by the absence of blood flow. Egress of the gas from the slice is therefore limited by diffusion, which must occur across both the thickness of the tissue and the unstirred fluid layer at its surface.

Buffering Properties of the Brain Interstitial Fluid

The 'buffering power' or 'buffer capacity' of an aqueous solution (β) is classically expressed as the ratio of the concentration of added strong base (ΔB), or strong acid (ΔA), to the consequent ΔpH. Thus, β is given in units of concentration. For a solution in which only HCO_3^- and CO_2 serve as buffer, and in which PCO_2 is essentially constant (via equilibration with a large, fixed CO_2 source), and the system is at equilibrium, the buffer capacity can be approximated by $\beta = 2.3 \times [HCO_3^-]$. Thus, assuming that PCO_2 is clamped over the long term by the systemic circulation and pulmonary ventilation, the extracellular buffer capacity of the brain (β_e) would be anticipated to be 50–60 mM against an acid or base load. In the face of a CO_2 challenge, however, the interstitial buffering system would have little ability to mitigate the resulting pH change, given the low concentration of phosphate and titratable groups on proteins. A doubling of the PCO_2 would be expected to cause a decrease in pH_e of approximately 0.30, with little change in $[HCO_3^-]_e$. This accounts for the significant extracellular acidosis in a brain slice, and it stands in marked contrast to the buffering properties of intracellular fluid. Within cells, a doubling of PCO_2 would cause a much smaller change in cytosolic pH and a measurable increase in the cytoplasmic $[HCO_3^-]$ due to the sizable intrinsic buffer capacity that is provided by intracellular proteins and phosphate.

Buffer capacity is typically considered for a system at equilibrium, with fixed PCO_2. These assumptions may not be valid over small segments of time, when the brain interstitial fluid may be subject to rapid acid or base loads. In the short term, a disequilibrium can arise due to the slow kinetics of the reversible reactions,

$$CO_2 + H_2O \underset{k-1}{\overset{k_1}{\leftrightarrow}} H_2CO_3 \leftrightarrow HCO_3^- + H^+ \qquad [2]$$

In the face of a rapid base load, as occurs during neural activity, the time for this buffer system to obtain equilibrium is on the order of 10–20 s (varying with temperature). The slow settling is due entirely to the low rate constants for the hydration of CO_2 (k_1) and the dehydration of H_2CO_3 (k_{-1}).

The enzyme CA can speed this process by the reversible, direct formation of HCO_3^- from CO_2 and water. Three extracellular isoforms of CA have been associated with the brain interstitial space: CA IV, CA XIV, and CA XV. Several studies indicate that buffering of this compartment is accelerated by these enzymes, with most of this catalysis attributable to CA IV and CA XIV. The localization of these CA isoforms may be both cell and region dependent. In the hippocampus and cortex, CA IV appears to be the principal isoform on astrocytes, whereas CA XIV is confined to neuronal surfaces. In contrast, CA XIV appears to be the glial isoform in the retina and is absent from the surface membrane of retinal neurons.

Although it is clear that CA speeds extracellular buffering in brain slices and *in vivo*, the total interstitial CA activity is apparently insufficient to maintain equilibrium when faced with a rapid alkaline load. Moreover, studies with CO_2 microelectrodes suggest that the PCO_2 does not remain fixed during sustained synchronous neural activity. Consequently, in some circumstances, interstitial buffering can be expected to be considerably less than the theoretical equilibrium value, despite the presence of extracellular CA.

Disequilibrium during a rapid acid or base challenge implies that the effective buffer capacity of the interstitial fluid will exhibit a time dependence. Depending on the acceleration coefficient (a) of the hydration–dehydration reactions, the time required to attain equilibrium can range from milliseconds to seconds. In hippocampal slices, a CA inhibitor had a limited effect on the rate of onset and recovery from interstitial alkalization. These experiments suggested that endogenous extracellular CA activity accelerated the buffering reactions less than threefold within the general interstitial space. With such limited acceleration, the full equilibrium buffer capacity of the interstitial fluid would not be realized for several seconds. Accordingly, within the time frame of synaptic events, the effective buffer capacity would be extremely low.

Buffer disequilibrium may have functional consequences in that a small acid or base source could have a relatively large effect on pH_e, provided it occurred with sufficient speed. Limited buffering of this sort may be relevant to a number of studies in the retina. Evidence suggests that suppression of presynaptic calcium currents within a few milliseconds of transmitter release is due to a decrease in subsynaptic pH_e, mediated by the release of the acidic contents of synaptic vesicles. Compared with Ca^{2+}channels, *N*-methyl-D-aspartate (NMDA) receptors are more sensitive to pH_e changes in the physiological range.

It may be speculated that the low effective buffer capacity of the extracellular fluid, coupled with the rapid transmembrane acid–base fluxes during neural activity, generated the need for pH_e regulatory mechanisms activated by neural activity. Thus, the evolution of glial transport systems that can respond to changes in membrane potential and quickly 'muffle' transient extracellular alkaline shifts may have been fueled by the time-dependent constraints on extracellular buffering.

pH Regulation in Neurons

Pioneering investigations of intracellular pH (pH_i) regulation in eukaryotic cells were performed on large invertebrate neurons. Studies of the squid giant axon and the nerve cell bodies of the snail established many of the principles, protocols, and norms of analysis subsequently utilized to study pH_i regulation in vertebrate cells. It has been established that for most cells, the membrane potential is considerably more negative than the equilibrium potential for H^+, which is typically in the range of 0 to −20 mV. Thus, given a transmembrane pathway, H^+ will enter a cell and acidify the cytosol. Combining the Nernst and Henderson–Hasselbach equations, it is readily shown that other charged acid or base species (e.g., OH^-, HCO_3^-, and NH_4^+) have the same equilibrium potential as H^+ and, given a flux pathway, will similarly acidify the cytosol. Metabolic processes, such as the aerobic generation of CO_2, may also contribute to acidification. Given a typical membrane potential of −50 to −70 mV, electrochemical equilibrium for a charged acid–base species would occur at a pH_i of approximately 6.4. Thus, to maintain a typical pH_i near 7.0–7.2 requires the continuous, energy-dependent, extrusion of acid.

In most invertebrate and vertebrate nerve cells, the net extrusion of acid is accomplished via secondary active transport, wherein the efflux of H^+ or the influx of HCO_3^- is coupled to an Na^+ influx, utilizing energy stored in the transmembrane Na^+ gradient. In many invertebrate and vertebrate neurons, acid extrusion is accomplished via an Na^+–H^+ exchanger and a Na^+-dependent HCO_3^- cotransporter. In the squid giant axon, snail giant neurons, and leech Retzius neurons, HCO_3^--dependent acid extrusion is mediated by an electroneutral, Na^+-driven, Cl^-–HCO_3^- exchanger (Figure 1).

Cloning and characterization of the Na^+-H^+ exchangers, as well as various Na^+-linked HCO_3^- transporters has been carried out for the mammalian forms of these carriers. The mammalian Na^+–H^+ exchangers comprise the SLC9A gene family, consisting of at least eight functional genes. These include five plasma membrane transporters, SLC9A1–SLC9A5 in human (or SLC9a1–SLC9a5 in mouse), also designated by their respective protein names NHE1–NHE5.

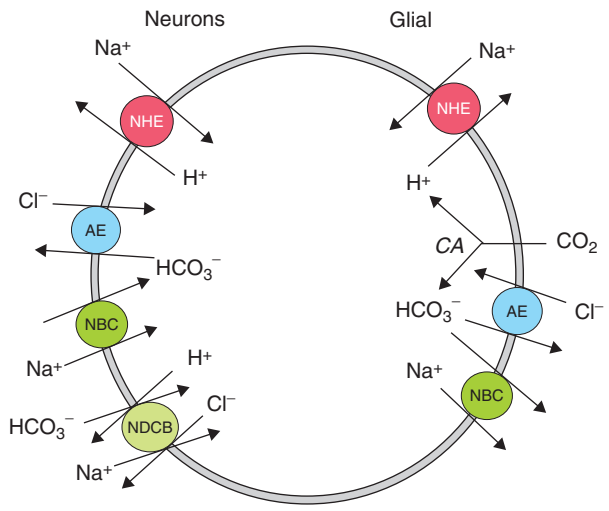

Figure 1 Major membrane acid–base carriers involved in cytosolic pH regulation in neurons (left) and in glial cells (right).

The most extensively characterized is NHE1. This ubiquitous transporter typically exhibits sensitivity to amiloride and its derivatives, and it is considered to serve a housekeeping role in maintaining pH_i. Although NHE1–NHE5 share a 1 Na^+:1 H^+ electroneutral stoichiometry, they exhibit differences in pharmacological sensitivities and in kinetics. All five of these Na^+–H^+ exchangers are expressed in the brain.

The Na^+-coupled HCO_3^- transporters are part of the human SLC4A gene family (SLC4a in mouse). Most abundantly expressed in brain are SLC4A4, SLC4A7, SLC4A8, and SLC4A10. The respective proteins are commonly designated as NBCe1 (from SLC4A4), an electrogenic Na^+–HCO_3^- cotransporter with typical Na^+:HCO_3^- stoichiometry of 1:2 when acting as an acid extruder; NBCn1 (from SLC4A7), an electroneutral Na^+–HCO_3^- cotransporter; NDCBE (from SLC4A8), an Na^+-driven, Cl^-–HCO_3^- exchanger; and NCBE (from SLCA10), an electroneutral Na^+–HCO_3^- cotransporter, with a possible dependence on Cl^-.

Acid transport has been physiologically characterized in a variety of vertebrate neurons. Typically described is a Na^+–H^+ exchanger and one or more Na^+-coupled HCO_3^- transporters. Apart from a few studies, the molecular identification of Na^+–H^+ exchangers and Na^+-coupled HCO_3^- transporters has not been correlated with the various physiological descriptions of acid extrusion in particular neurons. In this respect, the cells of the hippocampus are the best characterized mammalian neurons. Early physiological studies indicated that hippocampal neurons utilize an Na^+–H^+ exchanger that is insensitive to amiloride and its derivatives, as well as a Na^+-driven, Cl^-–HCO_3^- exchanger. Staining with polyclonal antibodies indicated that these cells also express the electrogenic Na^+–HCO_3^- cotransporter NBCe1. These observations suggest that mammalian neurons can make use of a complement of Na^+-linked HCO_3^- transporters. Diversity in the utilization of Na^+–H^+ exchangers in particular neurons may also be expected. For example, the CO_2-responsive or 'chemosensitive' neurons in the medulla appear to recover from acid loads mainly via an Na^+–H^+ exchanger that appears to be the NHE3 isoform.

In addition to mechanisms that extrude acid, neurons and other cells express transporters that import acid and lower pH_i. An 'acid loader' may be used to recover from an increase in pH_i, or it may function concomitantly with acid extrusion to effect tighter control over steady-state pH_i. Acid loading is commonly attributed to passive Cl^-–HCO_3^- exchangers. Anion exchangers of this class belong to the SLC4A gene family comprising the genes SLC4A1, SLC4A2, and SLC4A3, corresponding to anion exchange (AE) proteins AE1, AE2, and AE3, respectively. Both AE2 and AE3 are found in brain, and expression appears to increase with maturation. The AE proteins are transporters that link pH with Cl^- regulation in neurons, which may be significant for synaptic inhibition mediated by $GABA_A$ receptors. Prominent intracellular alkalinization may activate Cl^- influx via Cl^-–HCO_3^- exchange and thus shift the Cl^- equilibrium potential, and the inhibitory reversal potential, to more positive values; this may lead to a membrane depolarization, and hence excitation, following activation of $GABA_A$ receptors.

pH Regulation in Glial Cells

Glial cells have been ascribed a special role for acid–base regulation in the brain, particularly the extracellular spaces. The major pH regulating systems in glial cells are the sodium–bicarbonate cotransporter (NBC) and the sodium–hydrogen exchanger (NHE). The NHE extrudes H^+ out of the cell at the expense of Na^+ influx, and it is electroneutral and operates in the same way as described for neurons and many other cell types. In some glial cells, such as the leech giant glial cell, NHE has a much smaller impact on intracellular pH regulation due to the dominant presence of NBC in these cells. In vertebrate glial cells, however, NHE and NBC both contribute significantly to cellular pH regulation (Figure 1).

NBC has been described in nearly all types of macroglial cells, including astrocytes, oligodendrocytes, Schwann cells, and retinal Müller glial cells. NBC of most glial cells is electrogenic and cotransports one Na^+ with two HCO_3^-. The reversal potential of the NBC lies near the membrane resting

potential and can be adjusted to the actual membrane potential by driving NBC in the inward or outward direction. Activation of NBC in the inward direction results in an intracellular alkalinization and an increase in Na_i^+ and also a membrane outward current leading to hyperpolarization of the membrane, whereas activation of NBC in the outward direction acidifies the cytosol, drives Na^+ against a steep gradient out of the cell, and depolarizes the cell membrane. In addition, with the bicarbonate carried out of the cells, buffering capacity is added to the extracellular spaces. In the Müller cells of the retina, NBC is localized in the end foot processes and has a stoichiometry of 3 HCO_3^-:1 Na^+, specializations that suggest a role in the transport of CO_2/HCO_3^- from the retina to the vitreous humor.

The direction in which NBC operates across the plasma membrane depends on the pH_i and pH_o (and the $[HCO_3^-]_{i/o}$), the $[Na^+]_i$, and the membrane potential. This is also reflected by the pH_i changes induced by slow voltage steps in voltage-clamped leech giant neuropil glial cells. De- and hyperpolarizing voltage steps resulted in an intraglial alkalinization and acidification, respectively, providing an extrapolated change of 1 pH unit/110 mV membrane potential change, supporting a stoichiometry of 2 HCO_3^-:1 Na^+. Expressed in *Xenopus* oocytes, NBC imposed a membrane current, which was voltage and pH dependent (Figure 2(a)). The cotransporter appears to have a remarkably high affinity for HCO_3^- since it was found to be active even in the nominal absence of CO_2/HCO_3^-, when the extracellular bicarbonate concentration is usually less than 0.3 mM corresponding to air–CO_2. Expressed in frog oocytes, NBC significantly contributes to the apparent cytosolic buffer capacity, which becomes voltage dependent, and may enhance the efficacy of other acid–base transport systems. NBC thus contributed 0.7 mM mV^{-1} apparent buffer capacity to the cytosol (Figure 2(b)). Binding of CA to NBC renders the transport activity of NBC significantly more efficient by forming a 'transport metabolon' with CA, as shown when expressed in *Xenopus* oocytes. *In situ* hybridization revealed NBC mRNA expression throughout the rat central nervous system, with particularly high levels in the olfactory bulb, the hippocampus dentate gyrus, and the cerebellum.

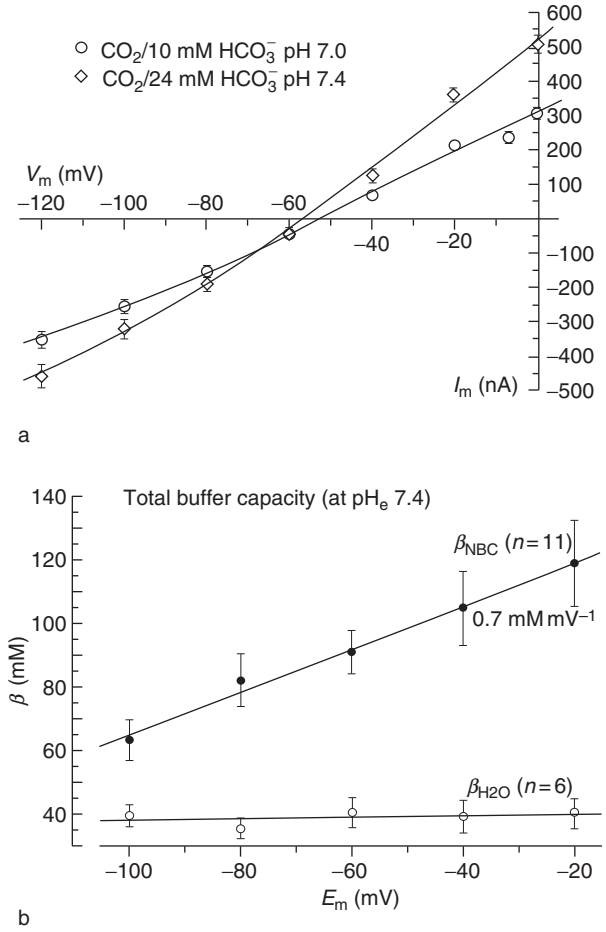

Figure 2 Functional characteristics of the electrogenic sodium–bicarbonate cotransporter (NBC) which contribute to cytosolic pH regulation mainly in glial cells. (a) Current–voltage relationship of the NBCe1 when expressed in *Xenopus* oocytes at different pH in the external solution. (b) Voltage dependence of the buffer capacity of native oocytes (open circles) and oocytes expressing NBCe1 (solid circles), revealing that the buffering increases by 0.7 mM/mV depolarization due to the activity of NBCe1. (a) Reproduced from Becker HM and Deitmer JW, unpublished. (b) Adapted from Becker HM and Deitmer JW (2004) Voltage dependence of H^+ buffering mediated by sodium–bicarbonate cotransport expressed in *Xenopus* oocytes. *Journal of Biological Chemistry* 279: 28057–28062.

Changes in pH Evoked by Neuronal Activity

Electrical activity can give rise to rapid, localized changes in the pH of the brain intracellular and extracellular fluid. These pH shifts are not uniform but, rather, can occur with different magnitude, direction, and time course in the neuronal, glial, and interstitial compartments. These distinct compartmental pH responses are considered in turn.

Activity-Dependent Shifts in Neuronal Intracellular pH

Depolarizing stimuli have been observed to cause a decrease in pH_i in several kinds of neurons. The earliest studies of neuronal pH_i responses were performed on invertebrate cells. Repetitive firing was associated with cytosolic acidification in molluscan and leech neurons, and depolarizing receptor potentials were found to acidify barnacle muscle. In snail neurons, depolarization elicited an acidification that

was Ca^{2+} dependent, and injection of Ca^{2+} into the cytosol was found to decrease pH_i as well, an effect partially explained by mitochondrial Ca^{2+}–H^+ exchange. Later experiments on snail neurons indicated that a large component of the cytosolic acidification was due to the Ca^{2+}–H^+ exchange property of the plasmalemmal Ca^{2+}-ATPase.

Similar Ca^{2+}-dependent acidification of neuronal cytosol has been observed in a wide variety of vertebrate neurons in which a voltage- or ligand-gated Ca^{2+} conductance was activated. Although most of these changes have been relatively slow, fast dendritic acidifications have been described in cerebellar Purkinje cells. Activation of a plasmalemmal Ca^{2+}-ATPase has been suggested as a basis for many of these observations, but other mechanisms may be involved as well. A metabolic source of acid is an additional possibility that has been suggested by some studies.

Activity-Dependent Shifts in Glial Intracellular pH

A rapid alkalosis has been observed in glial cells following adjacent neural activity. This response was observed *in vivo* in astrocytes of rat cortex and was also noted in the giant glial cells of the leech. Its presence in these unrelated species indicates that the underlying mechanism has been well conserved across evolution. Numerous studies suggest that the basis for this response is activation of an electrogenic Na^+–HCO_3^- cotransporter caused by depolarization of the glial cell membrane, as was first described in giant glial cells of the leech. This glial mechanism and its role in the regulation of extracellular pH are discussed in more detail later.

Some neurotransmitters, such as glutamate, can induce an intracellular acidosis via the glutamate uptake carrier, which is not only driven by the inward gradient of Na^+ but also takes up one H^+ per cycle. Intraglial pH decreases of several tens of pH units have been reported during exposure to glutamate or D-aspartate. As glutamate is converted to glutamine in glial cells, a reaction catalyzed by the pH-sensitive glutamine synthetase, this may in turn stimulate glutamine export from glial cells using neutral amino acid transporters which are coupled to H^+, such as SNAT3 and SNAT5 (human genes SLC38A3 and SLC38A5, respectively). Hence, glutamine uptake and glutamine export by glial cells, as described by the glutamate–glutamine shuttle hypothesis, can cause intraglial acidosis.

Activity-Dependent Shifts of Interstitial pH

Changes in interstitial pH that arise from neural activity vary greatly among the regions of the central nervous system. The most widely observed pH_e change consists of an interstitial alkaline shift at the onset of neural activity, followed by a more prolonged acidosis. This alkaline–acid response occurs in the hippocampus, cortex, cerebellum, and lateral geniculate nucleus (**Figure 3(a)**). In the spinal cord, activity elicited by stimulation of the dorsal root was found to produce a similar early alkalosis in neonatal preparations. At approximately 1 week of age, however, only an early acidification was present in the spinal cord (**Figure 3(b)**). A developmental shift from initial alkaline to initial acid transient has also been described in the dentate gyrus of rat hippocampal slices.

The early activity-dependent extracellular alkaline shifts studied in hippocampal slices arise from two distinct mechanisms. One component of the alkalosis has been attributed to the efflux of HCO_3^- across $GABA_A$ anion channels, a feature originally described for these receptors in crayfish muscle. This alkalosis requires extracellular carbonic anhydrase and, accordingly, is blocked by inhibitors of this enzyme. The second form of alkalosis is independent of HCO_3^-, mitigated by extracellular carbonic anhydrase, and becomes markedly amplified when this enzyme is blocked. On the basis of this observed amplification, the underlying mechanism has been referred to as a 'proton sink,' in distinction from the aforementioned 'bicarbonate source' that was attributed to $GABA_A$ receptors. This proton sink was found to be largely dependent on the presence of external Ca^{2+}. It was therefore suggested to arise from the Ca^{2+}–H^+ exchange property of the plasmalemmal Ca^{2+}-ATPase. Thus, entry of Ca^{2+}, by activating this transporter, would generate an extracellular alkalosis accompanied by a cytoplasmic acidification. However, apart from its

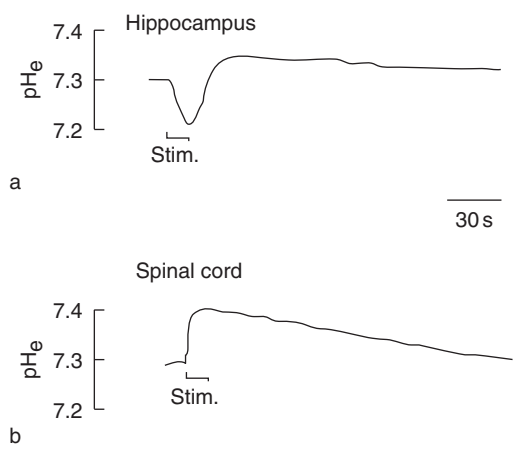

Figure 3 Variation in activity-dependent interstitial pH shifts. (a) Typical alkaline–acid response in rat hippocampal area CA1, elicited by repetitive electrical simulation of the Schaffer collateral fibers. (b) Typical rapid acid response in the dorsal or ventral horn of adult rat spinal cord during repetitive electrical stimulation of the dorsal root. Note that acid responses are upward.

dependence on extracellular Ca^{2+}, there is no additional compelling evidence to support this hypothesis.

Upon synaptic activation of hippocampal pyramidal neurons, inhibition of extracellular CA with benzolamide caused an amplification of the bicarbonate-independent alkaline transient and a simultaneous prolongation of the NMDA receptor-mediated component of the postsynaptic current. This suggested that the amplified pH_e changes had sufficient amplitude and speed to relieve the extracellular proton block of these receptors during the synaptic response. Similarly, benzolamide also increased epileptiform activity. It has been confirmed that these alkaline transients arise within tens of milliseconds, and evidence indicates that they can augment NMDA receptor activity even without amplification by benzolamide.

Other kinds of extracellular alkaline shifts have been described in retinal preparations in response to light. In frog retina, a light-evoked alkalosis was amplified by CA inhibitors and was suggested to arise from a downturn of ongoing metabolic acidosis, consistent with the light-evoked decrease of photoreceptor glycolysis. In the bee retina, light stimulation of the photoreceptors caused an increase in interstitial pH that was accompanied by an acidification in the photoreceptors and the glial cells. This insect response was likely due to enhanced uptake of ammonium into the cells by a chloride-dependent transporter.

The interstitial acid shifts that arise in response to neural activity are categorized as either rapid, short latency responses or slow, persistent acidifications. In rats, stimulation of the dorsal roots elicited a rapid acid shift in spinal cord. The absence of this acidification in x-irradiated pups was correlated with diminished astrocyte proliferation. In addition, the acid response was inhibited by stilbene derivatives such as 4,4-diisothiocyanostilbene-2,2′-disulfonic acid (DIDS) and 4-acetoamido-4′-isothiocyanostilbene-2,2′-disulfonic acid (SITS). These observations are consistent with an extracellular acidosis arising from glial electrogenic Na^+–HCO_3^- cotransport.

Late, slow acid shifts have been associated with metabolic activity. In cerebellum, activity-dependent acidosis was attenuated by fluoride, a glycolytic inhibitor. In rat hippocampal slices, the late acid shift was entirely attributed to elevation of tissue PCO_2, consistent with a transient increase in aerobic respiration.

Functional Aspects of H^+ Homeostasis and H^+ Signaling

The intracellular, cytosolic, pH (pH_i) of neurons and glial cells is actively maintained usually between 7.0 and 7.4, which is similar to the extracellular pH value (pH_o) in nervous tissue (7.1–7.3). Thus, 'resting' intra- and extracellular pH ranges between 40 and 100 nM free H^+ concentration. This means that there is often only a small chemical gradient of H^+ concentration, if any, across the cell membranes of resting neurons and glial cells. The main driving force of H^+ from the outside to the inside of a cell is the negative membrane potential. Any conductance of the cell membrane for H^+ and/or HCO_3^- at cell membrane potentials negative to −30 mV would result in a gain of acid or loss of base and would hence lead to an intracellular acidification.

In isolated cells and tissues, pH_i depends on whether the preparation is bathed in a solution buffered with a standard buffer, such as 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), or with CO_2 and HCO_3^-. In general, the steady-state pH_i is lower in CO_2–HCO_3^- buffered saline by 0.1–0.4 pH units. Neurons are the main consumers of ATP and, therefore, the main producers of CO_2. The link between CO_2 production and pH is the CO_2–HCO_3 buffer system, which plays a major role in the maintenance of pH both intra- and extracellularly. In order to understand the mechanisms of the acid and alkaline transients, it is often necessary to monitor pH in all three compartments of the tissue: neurons, glial cells, and the extracellular spaces. If there is a predominant flux of acid or base in one direction, then extracellular pH shifts will be approximate mirror images of intracellular pH changes. Net acid flux across glial and neuronal membranes may occur in opposite directions in different cells, however. Thus, during neural activity, cytosolic alkalization of glial cells, cytosolic acidification of neurons, and interstitial alkalization may occur simultaneously.

The proton equilibrium potential thus ranges between 0 and −30 mV in most animal cells. At a negative cell membrane potential of brain cells between −40 and −90 mV, the total H^+ gradient is 10–90 mV. Acid must be actively extruded by the cells against this electrochemical gradient, which requires energy. In most cells, this energy is provided by the Na^+ gradient, which drives NHE and other acid–base transport systems. The different isoforms of the NHE are the most prominent and ubiquitous transport proteins for acid extrusion in animal cells, and their malfunction is associated with a variety of diseases. NHE helps the cytoplasm to maintain a more alkaline pH than the acid–base electrochemical equilibrium across the plasma membrane, which is 6.4–6.8 in most animal cells. When cells become alkaline beyond their steady-state pH_i, base equivalents are extruded mainly by the Cl^-–HCO_3^- exchanger.

Active maintenance of the cytosolic pH is required to allow many enzymatic processes in cells to operate

at a normal rate. Most of these processes and reactions there are optimum at a pH of approximately 7.0 or higher. Changes in pH_i can therefore critically modulate molecular processes and cellular activity. Although the proton gradient across animal plasma membranes is not very large, and amounts to only approximately half the size of the Na^+ gradient, it is a significant source of energy that can be used to fuel secondary transport. Thus, membrane acid–base transporters can be fueled by the H^+ gradient, and they may also be modulated on the protein level by intra- and/or extracellular pH changes. In addition, there are a variety of voltage- and ligand-gated ion channels which are pH sensitive. Protons can either block these channels, as is the case for NMDA-sensitive, ionotropic glutamate receptors, or they may modulate gating, inactivation and/or deactivation, as found for some voltage-dependent channels. Changes in intra and extracellular pH can therefore modulate the excitability of neurons and can do so with the temporal dynamics of these pH changes. Thus, H^+ may acquire the function of signaling molecules both in the cytosol and in the extracellular spaces.

There are a number of pathological processes related to the impairment or the failure of acid–base regulation in the brain; in particular, hypoxia, anoxia, glucose deprivation, and conditions associated with ischemia can severely interfere with the normal acid–base homeostasis in the brain as in other tissues. Alkalosis in nervous tissue can contribute to hyperexcitability of neurons and hence ignite epileptiform activity in the brain. Oxidative stress decreases the activity of NHE and hence the intracellular pH as shown in rat solitary complex neurons. Hyperosmotic changes in the brain, which lead to shrinkage of cells, can lead to intracellular alkalinization due to activation of NHE. Changes in cellular and tissue pH are associated with hyperammonemia, which results from liver failure or urea cycle enzyme deficiencies, leading to brain edema and hepatic encephalopathy. The main pathways for acid-linked Ca^{2+} influx are reversed Na^+–Ca^{2+} exchange following increased intracellular Na^+ levels and Ca^{2+}-permeable acid-sensing ion channels (ASICs). Indeed, blocking ASICs or deleting the ASIC1a gene protect the brain from ischemic injury, suggesting that ASICs are responsible for glutamate-independent, acidosis-mediated brain injury during ischemia. Thus, acidosis can injure the brain via membrane receptor-mediated mechanisms with subsequent Ca^{2+} overload. This appears to be a major, glutamate-independent, pathological process underlying ischemic brain damage, directly mediated by extracellular acid shifts.

However, extracellular acidosis has also been associated with neuroprotective mechanisms (i.e., contributing to the survival of neurons during ischemia) by tuning down excitatory voltage- and ligand-gated membrane channels, such as Na^+ and Ca^{2+} channels, NMDA-sensitive glutamate channels, and other ionotropic receptor channels. The protective and harmful effects of long-lasting acidosis or alkalosis in nervous tissue also depend on the origin and context of these pH changes.

Summary

pH regulation by neurons and glial cells involves a variety of membrane acid–base carriers, including sodium–hydrogen exchange, sodium–bicarbonate cotransport, and sodium-dependent and sodium-independent chloride–bicarbonate exchange. Transport of acid–base equivalents across the plasma membranes of neurons and glial cells also results in pH changes in the extracellular spaces. Cytosolic and extracellular buffer capacity and the activity of carbonic anhydrase contribute to pH changes. pH changes can be elicited by neuronal activity, neurotransmitters and neuromodulators, metabolic processes, active cellular pH regulation, and secondary transporters carrying acid–base equivalents. Rapid pH transients may be signals rather than simply resulting from inadequate homeostatic acid–base regulation. Proton signaling in cells and local extracellular domains, particularly in the microenvironment of synapses, may contribute to information processing in nervous systems and to pathological, as well as neuroprotective mechanisms in nervous tissue.

See also: Glial Glutamate and GABA Metabolism; Neuron–Glia Coupling in Glutathione Metabolism.

Further Reading

Becker HM and Deitmer JW (2004) Voltage dependence of H^+ buffering mediated by sodium–bicarbonate cotransport expressed in *Xenopus* oocytes. *Journal of Biological Chemistry* 279: 28057–28062.

Becker HM and Deitmer JW (2007) Carbonic anhydrase II increases the activity of the human electrogenic Na^+/HCO_3^- cotransporter. *Journal of Biological Chemistry* 282: 13508–13521.

Chen JC and Chesler M (1992) pH transients evoked by excitatory synaptic transmission are increased by inhibition of extracellular carbonic anhydrase. *Proceedings of the National Academy of Sciences of the United States of America* 89: 7786–7790.

Chesler M (2003) Regulation and modulation of pH in the brain. *Physiological Reviews* 83: 1183–1221.

Deitmer JW (1991) Electrogenic sodium-dependent bicarbonate secretion by glial cells of the leech central nervous system. *Journal of General Physiology* 98: 637–655.

Deitmer JW (2002) A role for CO_2 and bicarbonate transporters in metabolic exchanges in the brain. *Journal of Neurochemistry* 80: 721–726.

Deitmer JW and Rose CR (1996) pH regulation and proton signalling by glial cells. *Progress in Neurobiology* 48: 73–103.

Kaila K and Voipio J (1987) Postsynaptic fall in intracellular pH induced by GABA-activated bicarbonate conductance. *Nature* 330: 163–165.

Makani S and Chesler M (2007) Endogenous alkaline transients boost postsynaptic NMDA receptor responses in hippocampal CA1 pyramidal neurons. *Journal of Neuroscience* 27: 7438–7446.

Putnam RW (2001) Intracellular pH regulation of neurons in chemosensitive and nonchemosensitive areas of brain slices. *Respiratory Physiology* 129: 37–56.

Romero MF, Fulton CM, and Boron WF (2004) The SLC4 family of HCO_3^- transporters. *Pflugers Archiv* 447: 495–509.

Rose CR and Deitmer JW (1994) Evidence that glial cells modulate extracellular pH transients induced by neuronal activity in the leech central nervous system. *Journal of Physiology* 481(1): 1–5.

Tong CK, Chen K, and Chesler M (2006) Kinetics of activity-evoked pH transients and extracellular buffering in rat hippocampal slices. *Journal of Neurophysiology* 95: 3686–3697.

Xiong ZG, Zhu XM, Chu XP, et al. (2004) Neuroprotection in ischemia: Blocking calcium-permeable acid-sensing ion channels. *Cell* 118: 687–698.

Neuropathic Pain

J Borzan and R A Meyer, The Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Neuropathic pain is a chronic pain condition that is caused by pathology in the nervous system. The pathology can be due to diseases that involve the nervous system (e.g., AIDS and diabetes), to dysfunction of the nervous system (e.g., following stroke), or to a traumatic injury of the nervous system. The pathology can involve the peripheral nervous system, the central nervous system, or both. Clinical characteristics of neuropathic pain depend on the type and the progression of disease but include burning pain, dysesthesias, pain to light stroking of the skin, sensory deficits, and widespread pain.

Etiology of Neuropathic Pain

Nerve Trauma

Nerve trauma can occur in the peripheral or the central nervous system. Peripheral lesions can be caused by nerve compression, limb amputation, or traumatic injuries and commonly result in dysesthesias, paresthesias, or chronic neuropathic pain. After limb amputation, painful sensations may be perceived as originating in the missing part of the body. Unfortunately, this phantom pain is experienced by more than half of amputees.

Pain of central origin may occur following spinal cord or brain trauma due to injury, compression (by tumors or cysts), or hemorrhagic or ischemic stroke. Not all central nervous system traumas result in central pain, but a necessary condition is damage to the spinothalamic tract. Moreover, it appears that spinal cord injury induces pain only if the gray matter is damaged.

A number of animal models of nerve trauma have been developed. The majority of the peripheral nerve injury models involve a lesion of all or part of the sciatic nerve system (Figure 1(a)). A complete sciatic nerve axotomy at the midthigh level leads to the degeneration of the nerve distal to the cut. The animals display autotomy (biting of the injured paw) which some interpret as a sign of pain but others argue is merely an attempt to remove the paw as a 'foreign object' because of the loss of sensation in that paw. In models of partial axotomy, the animals do not display self-injurious behavior but, rather, exhibit guarding behavior and signs of mechanical, heat, and cold hyperalgesia similar to the frequently observed clinical symptoms. Each partial nerve injury model has inherent advantages and disadvantages. For example, some models are easier to produce than others, some models lead to more reproducible injuries than others, and some models allow the relative roles of injured and uninjured afferents to be easily investigated. Animal models of central nervous system neuropathies involve contusion, compression, or partial transection of the spinal cord; electrical or chemical lesions of the spinal cord or various brain regions; and cerebral or spinal cord ischemia by occluding a major blood vessel.

A variety of behavioral assays are used to investigate the pain and hyperalgesia in these models. Behavioral tests measuring the threshold, latency, or duration of a reflexive response to mechanical, heat, or cold stimuli are used most frequently (Figures 1(b) and 1(c)). Several escape/avoidance conditioning paradigms have also been developed to elucidate changes in supraspinal processing due to nerve injury. A more challenging task is measuring spontaneous pain behavior in animals. No ideal model exists, but researchers typically record the number or duration of foot lifts, guarding behavior, and degrees of autotomy.

Disease Induced

Various disease states can lead to the development of neuropathic pain. Diabetes can often lead to a painful peripheral neuropathy. Diabetic neuropathy is a length-dependent neuropathy because it most commonly affects the extremities and progresses from more distal to proximal levels. High intracellular glucose levels can lead to pathological changes, such as vasoconstriction endothelial hyperplasia, and increased oxidative stress. These changes in turn promote neuronal dysfunction through decreased nerve blood flow and consequent neural hypoxia. Chronic alcohol abuse may also result in a peripheral neuropathy in the lower extremities due to direct degenerative effects on both myelinated and unmyelinated nerve fibers, disruption of axonal transport, and enhancement of glutamate neurotoxicity in neurons. Animal models of diabetes include genetic mutations such as in obese, leptin-deficient ob/ob mice or drug-induced diabetes produced by injection of streptozotocin.

Painful neuropathies can also develop as a result of viral infections. For example, after an acute infection of chicken pox, the neurotropic human herpes virus, varicella-zoster, remains dormant in the dorsal sensory or cranial nerve ganglia. A localized reoccurrence of the varicella-zoster virus (herpes zoster or shingles) can occur particularly when the immune

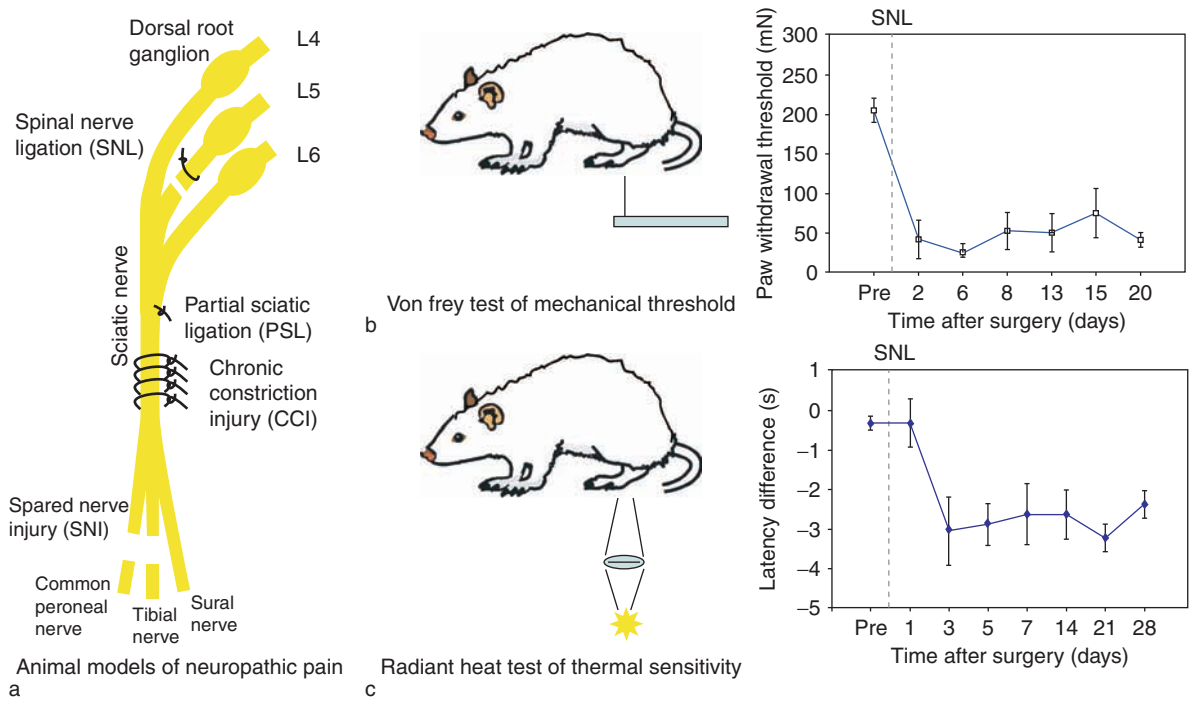

Figure 1 Animal models of neuropathic pain. (a) Peripheral nerve injury models include ligation of one or more spinal nerves (SNL), partial ligation of the sciatic nerve (PSL), loose ligation of the sciatic nerve using chromic gut sutures referred to as the chronic constriction injury (CCI), and ligation of common peroneal and tibial nerves to spare the sural nerve (SNI). (b, c) For each of these nerve injuries, the animals exhibit behavioral signs of hyperalgesia. (b) Animals develop a decreased paw withdrawal threshold to mechanical stimulation of the foot with von Frey monofilaments (mechanical hyperalgesia). (c) The latency to withdraw the paw from a heat stimulus is significantly shorter on the injured paw compared to the contralateral paw after nerve injury (heat hyperalgesia). Reproduced from Campbell JN and Meyer RA (2006) Mechanisms of neuropathic pain. *Neuron* 52(1): 72–92, with permission from Elsevier.

system is suppressed, as is common in older adults or individuals with other immunosuppressive conditions (e.g., Hodgkin's disease or HIV). This is most likely to occur in the dermatomes that had been most severely affected by the primary disease. The acute rash and pain are restricted to the area of one or several dermatomes and most commonly occur in the chest and trunk. After the acute viral infection subsides, approximately one-third of patients develop a chronic pain condition known as postherpetic neuralgia (PHN). The pain is described as burning, tingling, deep aching, itching, or stabbing. The pain in a subset of PHN patients is thought to be due to 'irritable nociceptors.' These patients report severe mechanical hyperalgesia but have minimal sensory loss in the painful area and their responses to thermal stimulation remain normal. Lidocaine gel or patches can be effective in providing at least temporary pain relief. Frequent topical application of capsaicin cream is moderately effective in reducing pain if the patients can tolerate the initial burning sensation capsaicin causes. Varicella-zoster virus is specific to humans, so animal models of PHN had to be developed using a different virus from the same family, namely herpes simplex virus type I. Animals develop lesions in the injected skin and show signs of hyperalgesia within 5 days after inoculation.

HIV infection commonly leads to the development of neuropathy either by immune system and macrophage-mediated axonal degeneration or by toxic effects of antiviral treatment medications on mitochondria. Similar to diabetic neuropathy and amyloidosis, HIV-induced neuropathy characteristically progresses by first affecting the most distal portions of axons. Activation of proinflammatory mediators such as tumor necrosis factor-α (TNF-α), interleukin (IL)-6, and nitric oxide in the dorsal root ganglia (DRG) is thought to contribute to the axonal damage and to the development of a painful neuropathy frequently characterized by bilateral painful paresthesias, dysesthesias, or numbness. Because the most common cause of sensory neuropathy in HIV patients stems from antiretroviral drugs, animal models of HIV-related sensory neuropathy utilize these drugs (e.g., didanosine, 2′,3′-dideoxycytidine) to study the mechanisms underlying the observed hyperalgesia and neuronal degeneration.

Autoimmune diseases such as Guillaine–Barré syndrome and multiple sclerosis (MS) cause demyelination and sensory deficits including pain. Approximately

80% of Guillaine–Barré and 50% of MS patients report pain. The etiology of MS is not well-known, although it is believed that the onset of the disease involves pathology of the immune system, particularly disruption of myelin proteins. Guillaine–Barré syndrome has several subtypes based on the suggested etiology of the disease, namely acute inflammatory demyelinating polyradiculoneuropathy, acute motor axonal neuropathy, and acute motor and sensory axonal neuropathy. T cell disruption of myelin proteins is indicative of the first subtype, whereas antibodies to gangliosides are thought to promote demyelination and injury in the other two subtypes. Bacteria (e.g., *Campylobacter jejuni*) and viruses have also been implicated as disease triggers for Guillaine–Barré syndrome. Animal models of these autoimmune disorders include induction of experimental autoimmune encephalomyelitis or demyelination of primary afferents by lysolecithin.

Chemotherapy Induced

Chemotherapeutic drugs such as vincristine or paclitaxel promote microtubule stabilization, thus inhibiting tumor cell proliferation. However, the effectiveness of chemotherapy agents is often limited by severe neurotoxic side effects. Many patients undergoing paclitaxel chemotherapy develop peripheral neuropathies. Chemotherapeutic agents such as paclitaxel, cisplatin, or vincristine also cause sensory neuropathies in animals and are used to elucidate mechanisms of this type of painful neuropathy. The most common animal model of cancer pain involves injections of sarcoma cells into the medullary cavity of the humerus, calcaneous, or femur bone. Sarcoma cells cause gradual degradation of bone that is accompanied by behavioral signs of hyperalgesia.

Mechanisms of Neuropathic Pain

Neuropathic pain is associated with pathological changes at many sites along the neural axis. For the spinal nerve ligation model of neuropathic pain, numerous abnormalities that may account for neuropathic pain have been identified in both the peripheral and the central nervous system (Figure 2). At the site of nerve injury (Figure 2, no. 1), spontaneous neural activity and sensitivity to mechanical stimuli develops. At the DRG of the injured fibers (Figure 2, no. 2), there is an up-(or down)regulation of numerous molecules in small and large neurons reflecting, at least in part, the loss of trophic support from the periphery. Spontaneous neural activity also develops. Wallerian degeneration of the nerve fibers distal to the injury exposes the neighboring intact fibers to a milieu of cytokines and trophic factors (Figure 2, no. 3). Partial denervation in the target tissues can lead to a sensitization at the primary afferent terminals due to an excess of trophic factors (Figure 2, no. 4). At the DRG of the uninjured fibers (Figure 2, no. 5), there is an up-(or down)regulation of numerous molecules in small neurons reflecting, at least in part, the enhanced trophic support from the periphery. In the central nervous system, sensitization develops in dorsal horn neurons (Figure 2, no. 6) characterized

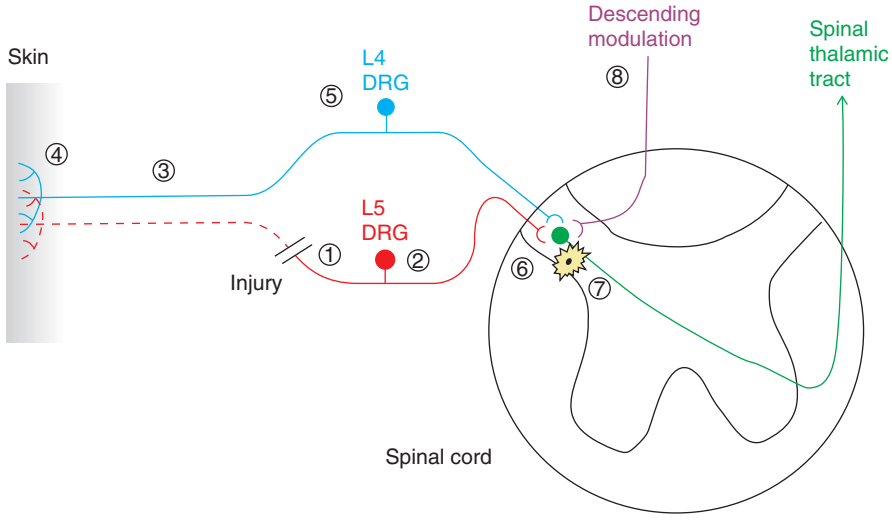

Figure 2 Sites of pathology following peripheral nerve injury include (1) the site of injury, (2) the dorsal root ganglia (DRG) of the injured afferents, (3) the region of Wallerian degeneration distal to the injury site, (4) peripheral terminals of the uninjured afferents, (5) the DRG of the uninjured afferents, (6) dorsal horn neurons, (7) immune cell–neuron interaction, and (8) descending pathways influencing spinal cord neurons. Reproduced from Campbell JN and Meyer RA (2006) Mechanisms of neuropathic pain. *Neuron* 52(1): 77–92, with permission from Elsevier.

by an enhanced responsiveness to cutaneous stimuli that may account for the hyperalgesia/allodynia observed in the animals. This sensitization may be driven by spontaneous activity in nociceptors but also may be due, at least in part, to activation of glial cells (Figure 2, no. 7) and alterations in descending modulation (Figure 2, no. 8).

Role of the Peripheral Nervous System

The importance of the peripheral nervous system in neuropathic pain is supported by the observation that peripherally acting drugs alter neuropathic pain behavior. For example, the cannabinoid receptor 2 (CB2) is selectively expressed in the peripheral nervous system, and selective CB2 agonists were effective in reversing mechanical and heat hyperalgesia in an animal model of neuropathic pain. In addition, peripherally acting opiates have been shown to ameliorate neuropathic pain behavior.

Ectopic activity A hallmark of neuropathic pain is ectopic activity in primary afferents. In the case of nerve axotomy, ectopic activity develops at the site of injury as well as in the nearby spared afferents. Ectopic activity consists of ongoing neural activity as well as sensitivity to mechanical stimuli. The spontaneous neural activity provides a novel impulse barrage to the spinal cord that contributes to spontaneous pain and to the development of sensitization in second-order neurons. The mechanosensitivity may account for the pain to mechanical stimulation of the nerve injury site.

For distal nerve injuries, ectopic activity is seen in injured myelinated and unmyelinated fibers. In contrast, for proximal injuries (e.g., a spinal nerve ligation) the ectopic activity in injured afferents appears to be restricted to myelinated fibers (Figure 3). These spontaneously active A fibers appear to be low-threshold mechanosensitive afferents that normally transmit information about nonnoxious stimuli. As described later, ongoing activity of nociceptive C (but not A) fibers is considered critical in the development of central sensitization and neuropathic pain. In fact, electrical stimulation of primary afferents at A fiber strength does not result in sensitization of dorsal horn neurons. One possible explanation is that large DRG neurons undergo a phenotypic switch after nerve injury such that they express substance P and brain-derived neurotropic factor (BDNF); release of these transmitters in the dorsal horn could lead to central sensitization. A small proportion of large-diameter, myelinated, Aβ afferent neurons are nociceptors. It is possible that if these afferents develop ectopic activity

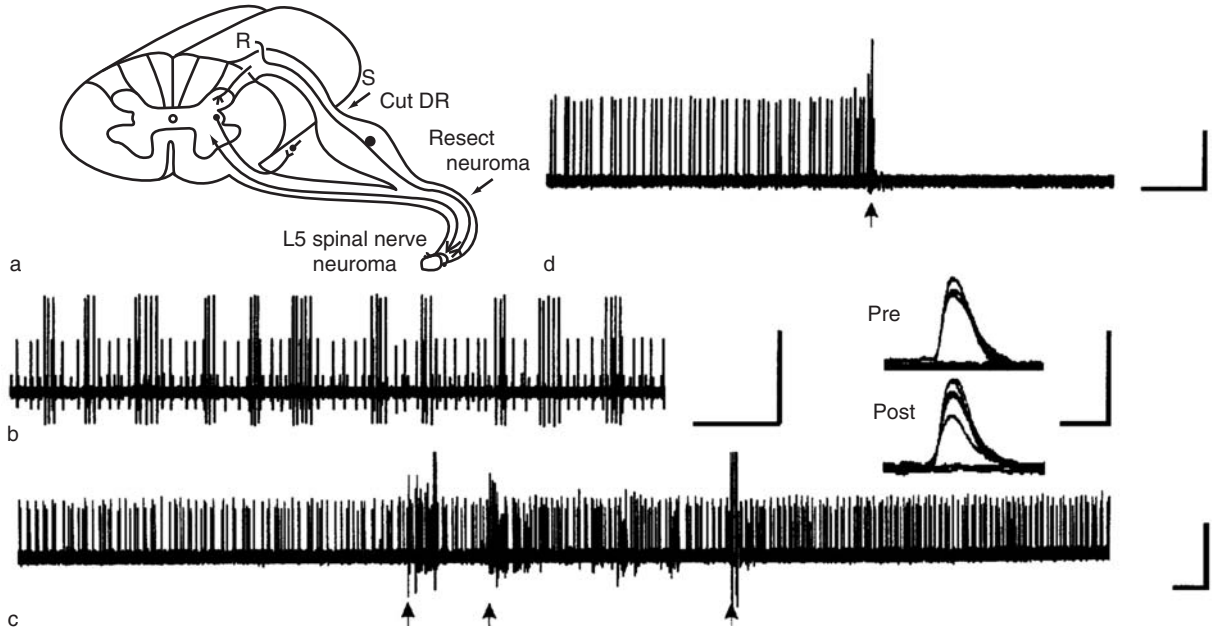

Figure 3 Large-diameter primary afferents develop ectopic activity following a spinal nerve injury. (a) Action potential (AP) recordings were made in the L5 dorsal root after an L5 spinal nerve lesion. (b) The spontaneous activity observed in large myelinated fibers was bursting (large AP shape), at uniform frequency (medium size AP), or irregular (small AP). (c) Nine days after the lesion, spontaneous activity was not eliminated by resecting the neuroma (but an injury discharge was apparent at arrows). (d) The spontaneous activity was eliminated by cutting the dorsal root, indicating that it originated from the dorsal root ganglia. Adapted from Liu CN, Wall PD, Ben-Dor E, Michaelis M, Amir R, and Devor M (2000) Tactile allodynia in the absence of C-fiber activation: Altered firing properties of DRG neurons following spinal nerve injury. *Pain* 85(3): 503–521, with permission.

after nerve injury, they could potentially play a role in neuronal sensitization.

Neuropathic pain is not solely dependent on the changes that occur in the injured nerve. The fact that neuropathic pain still occurs after a ganglionectomy (in which the injured afferent soma are removed) or ventral rhizotomy provides evidence that alterations in uninjured afferent fibers are sufficient to produce hyperalgesia. Both myelinated and unmyelinated uninjured afferents develop ectopic activity. In this case, unmyelinated C fibers in uninjured axons show rates of spontaneous activity sufficient to contribute to sensitization, hyperalgesia, and neuropathic pain.

Alterations of sodium channel expression and activation properties are considered major contributors to the development of ectopic activity. Nerve lesions lead to a reduction of tetrodotoxin (TTX)-resistant currents and their channel transcripts and an increase in rapidly repriming TTX-sensitive current and Nav1.3 channel transcripts. These changes alter cell membrane potential properties and make the cells more excitable. Whereas the TTX-resistant, sensory neuron-specific channels, Nav1.8 and Nav1.9, are reportedly downregulated in injured afferents, Nav1.8 is upregulated in the uninjured neighboring fibers. The Nav1.7 TTX-sensitive sodium channel is also thought to play an important role in nociception and certain forms of neuropathic pain. Nav1.7 is expressed predominantly in the small-diameter primary afferents and is believed to participate in the action potential initiation at the peripheral nerve terminals. Certain mutations of the SCN9A gene encoding the α subunit of the Nav1.7 channel have been found to account for the neuropathic pain associated with erythermalgia and paroxysmal extreme pain disorder. Other mutations of the same gene result in loss of function of Nav1.7 and are the underlying cause of channelopathy-associated insensitivity to pain.

Peripheral sensitization Following many types of tissue injury, including nerve injury, primary afferents develop lower activation threshold, become more responsive to natural stimuli, and develop larger receptive fields. In other words, they become sensitized. A change occurs in the transduction mechanisms that convert external stimuli (heat, touch, etc.) to electrical signals of neurons. After injury, proteins and ion channels in the primary afferents can undergo structural changes (e.g., phosphorylation) or primary afferents can develop altered gene expression that signals the production of new molecules. Consequently, previously nonnoxious stimuli may now be sufficient to activate a given channel or receptor, they may remain active for a longer period of time, or new channels may become activated by different factors within the injury site. These changes are responsible for peripheral sensitization.

It is well known that cyclo-oxygenase enzymes play a role in inflammatory pain conditions by converting arachidonic acid into proinflammatory factors such as prostaglandin E2 (PGE2). Prostaglandins have also been shown to contribute to peripheral sensitization in neuropathic pain by shifting the activation threshold of tetrodotoxin-resistant sodium channels (TTX-R Na^+) Nav1.8 and Nav1.9. PGE2 binds to G-protein-coupled receptors and increases cyclic AMP, which can sensitize TTX-R Na^+ channels by activating protein kinases A and C.

Following peripheral nerve lesion, peripheral sensitization occurs in the uninjured fibers innervating the same territory. Sensitization of uninjured fibers occurs partly because Wallerian degeneration of the injured nerve (**Figure 2**, no. 3; **Figure 4**) results in a release of a number of proinflammatory factors, including cytokines and growth factors, that can alter the activation threshold of the surviving neurons (**Figure 2**, no. 4; **Figure 4**). Several studies have demonstrated increased sensitivity of uninjured afferents to chemical (TNF-α), mechanical, and heat stimuli following peripheral nerve injury. Upregulation of TRPV1 and TRPA1 expression in the dorsal root ganglia of intact nerves that neighbor damaged nerves may contribute to the perception of heat and cold hyperalgesia observed after injury (**Figure 2**, no. 5). Animal models of nerve injury have also demonstrated upregulation of calcitonin gene-related peptide (CGRP), BDNF, $P2X_3$, and substance P mRNA in the uninjured DRG or nerve (**Figure 2**, nos. 3 and 5). Release of the neuromodulator CGRP causes activation of dorsal horn neurons and facilitates sensitization by other stimuli (**Figure 2**, no. 6). Upregulation of primary afferent release of BDNF can result in greater phosphorylation of NMDA receptors on second-order neurons in the dorsal horn and their subsequent sensitization. Adenosine 5′-triphosphate activation of $P2X_3$ receptors on sensory nerve terminals can also contribute to spinal cord sensitization. Increased levels of nerve growth factor (NGF) can sensitize uninjured afferents and contribute to peripheral sensitization and ectopic activity.

Hyperexcitability of peripheral sodium channels is a contributing underlying mechanism in a number of painful peripheral neuropathies, such as PHN. Patients with 'irritable nociceptor'-type PHN obtain pain relief with sodium channel blockers (e.g., anticonvulsant medications). The fact that application of local anesthetics (e.g., 1% lidocaine) or partial denervation by topical application of capsaicin also provides pain relief in these patients is evidence that ectopic peripheral input maintains the hyperalgesic state in these patients.

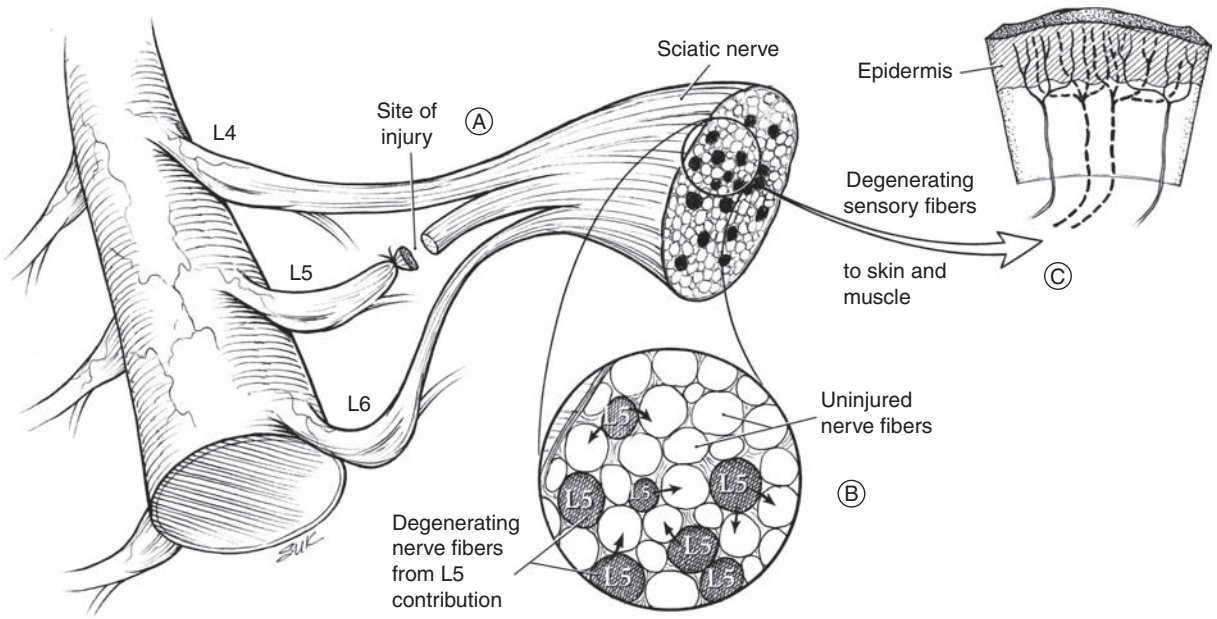

Figure 4 The properties of uninjured afferents may be altered by the neighboring injured fibers. An L5 spinal nerve lesion leads to Wallerian degeneration of the axons distal to the cut. This leads to invasion of the nerve by macrophages and the release of cytokines and chemokines. The adjacent, uninjured fibers from the L4 and L6 spinal nerves are exposed to a pool of reaction products associated with degeneration that could lead to an alteration of their properties including the development of ectopic activity. Spinal nerve injury also leads to a partial denervation in the peripheral target tissue. This could lead to the upregulation of factors in the tissue that could result in a sensitization of the primary afferent nociceptors. Adapted from Ringkamp M and Meyer RA (2005) Injured versus uninjured afferents: Who is to blame for neuropathic pain? *Anesthesiology* 103: 221–223, with permission.

Role of the Central Nervous System

Peripherally directed treatments do not always help in alleviating neuropathic pain, indicating that changes also occur in the central nervous system. Nerve injury can lead to sensitization of central pain-signaling neurons as well as changes in the properties of the inhibitory interneurons and descending inhibition. Enhanced ongoing input from the periphery onto spinal dorsal horn neurons causes central sensitization – an altered response of the central neurons.

Wind-up is a single synapse phenomenon of enhanced responsiveness of a central neuron following continuous low-frequency nociceptor activation from the periphery (Figure 2, no. 6). Central sensitization can result from spatial and temporal summation of wind-up. Wind-up is a short-lasting phenomenon (tens of seconds) largely mediated by glutamatergic activation of NMDA receptors. Hence, blocking NMDA receptors abolishes wind-up to various stimuli in healthy individuals and reduces neuropathic pain in select conditions, such as painful phantom limb and PHN. However, secondary hyperalgesia (pain in the area surrounding the injury and attributed to central sensitization) is generally not altered by inhibiting NMDA receptors. Enhanced responsiveness of dorsal horn neurons after central sensitization lasts for prolonged periods of time, suggesting additional or alternative mechanisms to the wind-up phenomenon. Specifically, the release of substance P from primary afferent neurons can facilitate depolarization of dorsal horn neurons by directly activating NK-1 receptors and by subsequent modulation of NMDA receptor activation.

Low-threshold mechanoreceptors normally provide a central inhibitory influence on the peripheral nociceptive input. Injury results in a sensitization of nociceptive dorsal horn neurons such that light touch or brushing now evokes activity in these neurons (Figure 2, no. 6). Allodynia, or pain evoked by normally nonnoxious stimuli such as brush, is signaled by low-threshold mechanoreceptors.

Some of the spinal cord interneurons rich in GABA, glycine, or opioid peptides have an inhibitory influence on synaptic transmission. Peripheral nerve injury can lead to impaired tonic inhibitory function of these interneurons by reducing the synthesis or release of inhibitory neurotransmitters or to apoptosis of these interneurons due to glutamate neurotoxicity. Decreased expression of potassium chloride exporter (KCC2) in lamina I dorsal horn neurons after injury alters the transmembrane chloride ion gradient. The net result is depolarization of the cell membrane and a phenotypic switch of GABA and glycine neurotransmission from an inhibitory to an excitatory role in the spinal cord. Additionally, disinhibition in the

dorsal horn occurs in part due to direct loss of spinal GABAergic neurons which is sufficient (although not necessary) to produce neuropathic pain following nerve injury. Mice lacking the μ opioid receptors show enhanced nerve injury-induced behavioral hyperalgesia and wind-up compared to wild-type animals. Intrathecal administration of delta opioid agonist reversed mechanical and cold hyperalgesia following spinal nerve ligation. These and similar studies indicate that endogenous opioids serve an antinociceptive role following nerve injury.

Similarly, nerve trauma in the central nervous system along the spinothalamic pathway can result in central pain in part due to the spinal or cortical changes in GABA or glutamate function. Spinal cord hemisection in animals leads to mechanical hyperalgesia reversed by intrathecal administration of antagonists to ionotropic glutamate receptors. Upregulation of metabotropic glutamate receptor group I (mGluR1) but not mGluR5 and downregulation of mGluR2/3 expression in the spinal cord mirrors the development of behavioral hyperalgesia following spinal contusion injury. Group I mGluR antagonist or group II and group III mGluR agonists effectively reversed mechanical but not heat hyperalgesia in this model. As mentioned previously, glutamate receptors are important contributors to the overall excitability of the spinal cord neurons and hence influence the development or maintenance of central sensitization. Serotonin-rich descending neurons from the brain stem or ascending projection neurons in the spinal cord neurons contain serotonin receptors 5-HT1, 5-HT2, and 5-HT3 that can serve an inhibitory or a facilitatory role in modulation of nociceptive input. However, downregulation of serotonin following spinal cord injury parallels development of hyperalgesia, an effect reversible by intrathecal administration of serotonin.

Central pain patients have been shown to have increased activity in the thalamic neurons. In an animal model of central pain, the hyperactivity of thalamic neurons was exaggerated by local application of glutamate and inhibited by application of MK-801 (NMDA antagonist) or GABA agonist. These results support clinical use of GABA agonists or NMDA antagonists in the treatment of central pain. In fact, ketamine, lamotrigine (NMDA antagonist or glutamate inhibitor), baclofen, propofol, and benzodiazapines (GABA agonists) are sometimes successful in providing pain relief in central pain patients. Cortical activation of μ opioid receptors (e.g., in prefrontal ventrolateral orbital cortex or anterior cingulate cortex) attenuates sensory or aversive components of neuropathic pain in animal models of nerve injury.

Nociceptive input from the periphery is also regulated by supraspinal descending neuronal pathways (Figure 2, no. 8). The major brain structures demonstrated to provide inhibitory control in nociception are periaqueductal gray (PAG) and rostral ventromedial medulla (RVM). Neuronal activation in PAG or RVM by electrical stimulation or opioid agonist (morphine) administration leads to analgesia, whereas transaction of their descending axons in dorsolateral funiculus results in hyperalgesia. Additionally, descending serotonergic input from the brain stem can either inhibit excitation of spinal cord neurons by activating 5-HT_1 and 5-HT_2 receptors or facilitate maintenance of neuropathic pain by activating spinal serotonergic 5-HT_3 neurons. Locus coeruleus and locus subcoeruleus participate in descending inhibition of the dorsal horn neurons by noradrenergic release (particularly the α_2 receptor subtype). Certain antidepressant medications are efficacious in neuropathic pain because they act on serotonergic or noradrenergic receptors.

Role of the Sympathetic Nervous System

In some patients, blockade of the sympathetic nervous system leads to pain relief. This sympathetically maintained pain (SMP) is thought to depend on the peripheral release of norepinephrine from sympathetic efferent fibers that activates and sensitizes primary afferents. Animal studies have shown that peripheral nerve injury can result in *de novo* expression of α_1 adrenergic receptors in adjacent uninjured afferents and sensitization to adrenergic agonists, indicating that the coupling site between the sympathetic and pain systems may be in the primary afferent fibers. In human studies, intradermal norepinephrine injections caused pain in patients with SMP but not in healthy controls. Some patients with complex regional pain syndrome have SMP and gain pain relief following sympathetic ganglion block or after intravenous administration of phentolamine, an α-adrenergic antagonist.

The sympathetic nervous system has been implicated in central pain mechanisms as well. Neuropathic pain patients with stroke-induced central pain frequently complain of exacerbated pain under stress and have areas of abnormal blood flow and sweating.

Influence of the Immune System

In addition to the obvious role of the immune system in inflammatory pain, there is accumulating evidence that immune factors are of importance in neuropathic pain. Following nerve injury, the immune system response includes macrophage infiltration, T-cell activation, and increased expression and release of proinflammatory factors such as IL-1β and IL-6, prostanoids such as PGE2, and TNF-α (Figure 2, no. 3;

Figure 5(b)). Damaged or denervated Schwann cells secrete leukemia inhibitory factor (LIF) and monocyte chemoattractant protein-1 (MCP-1). Upregulation of LIF following injury results in increased release of substance P and other molecules that contribute to nociceptor sensitization. MCP-1 is a chemokine cytokine that binds to chemokine receptor CCR2 and has been shown to activate monocytes and depolarize nociceptors following nerve injury either directly or by promoting the release of proinflammatory factors from surrounding nonneuronal cells. Cytokine IL-1β promotes expression and release of NGF that can lead to sensitization of primary afferents. Deletion of IL-1β or overexpression of IL-1β antagonist largely diminished the development of neuropathic pain following spinal nerve ligation or axotomy. These results complement previous evidence that injection of IL-1β or TNF-α in the sciatic nerve causes mechanical and heat hyperalgesia. However, it appears that inhibition of cytokine TNF-α effectively reduces neuropathic pain only when administered preemptively. It is suggested that pretreatment with TNF-α inhibitors prevents phosphorylation of MAP kinase p38 in DRG, which parallels attenuation of behavioral hyperalgesia following spinal nerve ligation.

Immunological factors participate in the pathology of neuropathic pain in the central nervous system as well. Microglia, the immune cells of the central nervous system, become activated following peripheral nerve injury (Figure 5(a)), particularly if the lesion site is distal to the DRG. Painful peripheral neuropathy due to HIV infection may also be caused in part by spinal cord microglia activation. Various approaches to inhibit microglial activation provide direct support that immune cell activation in the central nervous system contributes to neuropathic pain, especially in its development. Immune cells may influence neuronal activation and sensitization through upregulation of purinergic receptors and chemokines, activation of various kinases, and release of various

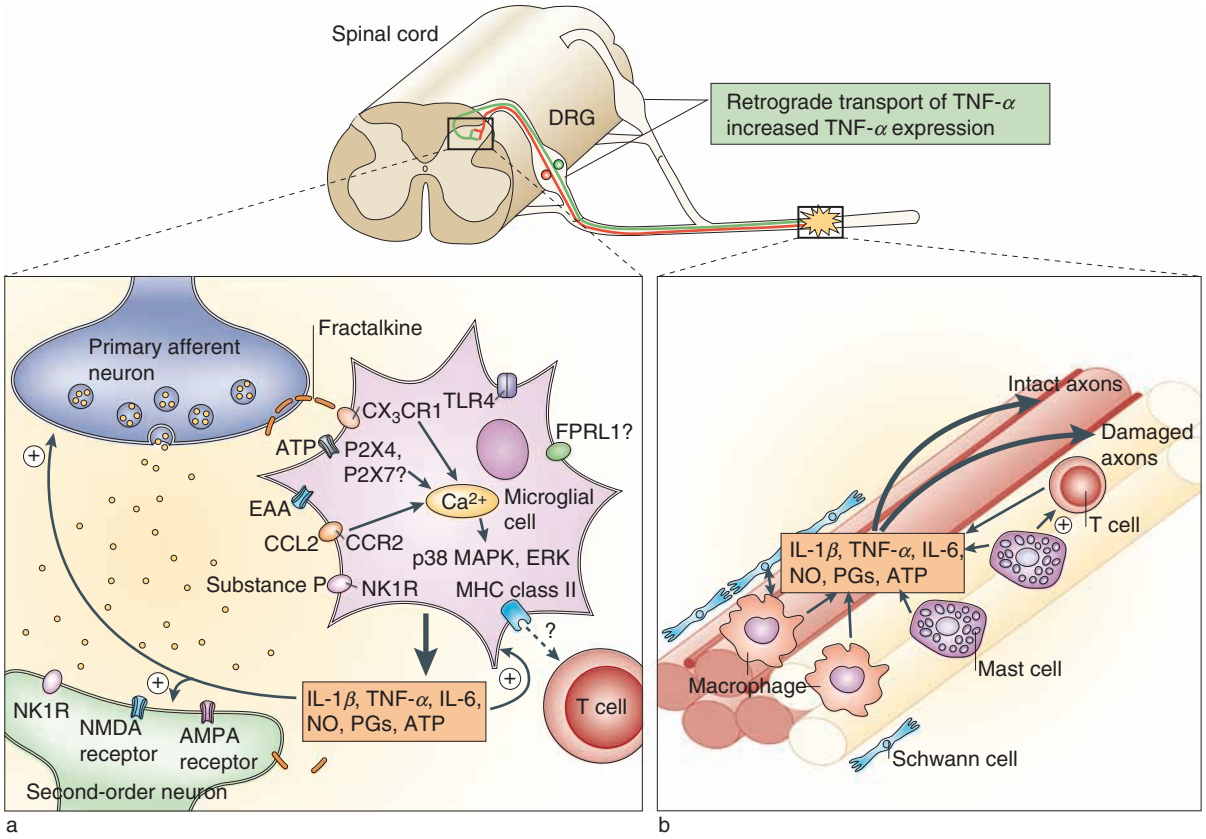

Figure 5 Immune cells present or recruited in the central (a) and peripheral (b) nervous system following nerve injury influence the excitability of primary afferents and second-order neurons. TNF-α, tumor necrosis factor-α; IL-1β, interleukin-1β; IL-6, interleukin-6; NO, nitric oxide; PGs, ATP and prostaglandins; AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; CCR2, CCL2 receptor (CX3CR1), fractalkine receptor; EAA, excitatory amino acids; ERK, extracellular signal-regulated kinase; FPRL1, formyl peptide receptor-like 1; MHC, major histocompatibility complex; NGF, nerve growth factor; NK1R, neurokinin-1 receptor; NMDA, *N*-methyl-D-aspartate; P2X4 and P2X7, ionotropic purinoceptors; p38 MAPK, p38 mitogen-activated protein kinase; TLR4, toll-like receptor 4; CCL2, chemokine (C–C motif) ligand 2. Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Neuroscience* (Marchand F, Perretti M, and McMahon SB (2005) Role of the immune system in chronic pain. *Nature Reviews Neuroscience* 6: 521–532), copyright 2005.

cytokines. Microglia express a number of purinergic receptors. Two have been demonstrated as important in neuropathic pain: P2X4 and P2X7. Inhibition of P2X4 after nerve injury alleviates mechanical allodynia and prevents upregulation of P2X4 in the dorsal horn. Genetic deletion of the P2X7 receptor resulted in decreased neuropathic pain, but it is not clear whether this effect is mediated specifically by microglia or includes involvement of other central or peripheral cell types. Neuronal release of chemokine fractalkine activates spinal microglia via CX3CR1 receptor and causes activation of the MAPK p38 pathway. Nerve injury also leads to activation of Src kinase in microglia but not neurons or astrocytes. Spinal inhibition of Src kinase attenuates mechanical allodynia by inhibiting extracellular signal-regulated protein kinase. Increased levels of proinflammatory cytokines have been observed in the spinal cord tissue following peripheral nerve injury. Additionally, intrathecal delivery of cytokine inhibitors can lead to the reduction of neuropathic pain.

Evidence indicates that the immune system plays an important role in neuropathic pain conditions, both peripherally and centrally. Interestingly, it appears that immune cell responses are more critical in initiating rather than maintaining neuropathic pain.

See also: Autoimmune Autonomic Neuropathy; Diabetic Neuropathy; Neuropathy: Chemically-Induced; Neuropathy: Metabolically-Induced; Neuropathy: Peripheral; Nociceptor Responses; Pain Management; Pain: Central; Sodium Channels; Wallerian Degeneration; Zoster and Postherpetic Neuralgia.

Further Reading

Campbell JN and Meyer RA (2006) Mechanisms of neuropathic pain. *Neuron* 52(1): 77–92.

Cohen S and Abdi S (2002) Central pain. *Current Opinion in Anaesthesiology* 15: 575–581.

De Leo JA, Tawfik VL, and LaCroix-Fralish ML (2006) The tetrapartite synapse: Path to CNS sensitization and chronic pain. *Pain* 122: 17–21.

Devor M (2006) Sodium channels and mechanisms of neuropathic pain. *Journal of Pain* 7(supplement 1): S3–S12.

Gallagher RM (2006) Management of neuropathic pain: Translating mechanistic advances and evidence-based research into clinical practice. *Clinical Journal of Pain* 22(1): S2–S8.

Gold MS (2000) Spinal nerve ligation: What to blame for the pain and why. *Pain* 84: 117–120.

Liu CN, Wall PD, Ben-Dor E, Michaelis M, Amir R, and Devor M (2000) Tactile allodynia in the absence of C-fiber activation: Altered firing properties of DRG neurons following spinal nerve injury. *Pain* 85(3): 503–521.

Marchand F, Perretti M, and McMahon SB (2005) Role of the immune system in chronic pain. *Nature Reviews Neuroscience* 6: 521–532.

Ringkamp M and Meyer RA (2005) Injured versus uninjured afferents: Who is to blame for neuropathic pain? *Anesthesiology* 103: 221–223.

Trang T, Beggs S, and Salter MW (2006) Purinoceptors in microglia and neuropathic pain. *Pflugers Archiv: European Journal of Physiology* 452: 645–652.

Vanegas H and Schaible HG (2004) Descending control of persistent pain: Inhibitory or facilitatory? *Brain Research: Brain Research Review* 46(3): 295–309.

Waxman SG (2006) Neurobiology: A channel sets the gain on pain. *Nature* 444(7121): 831–832.

Woolf CJ and Salter MW (2000) Neuronal plasticity: Increasing the gain in pain. *Science* 288: 1765–1768.

Yaksh TL (2006) Calcium channels as therapeutic targets in neuropathic pain. *Journal of Pain* 7(1, supplement 1): S13–S30.

Neuropathy: Chemically-Induced

M Jokanović, Experta Consulting, Belgrade, Serbia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the nervous system there are four targets for the action of chemicals: the neuron, the axon, the myelinating cell, and the neurotransmission system. As a result, it is possible to identify neurotoxic chemicals which cause neuronopathies, axonopathies, myelinopathies, and neurotransmission-associated toxicity. This is a common classification system and is utilized here to discuss chemically induced neuropathies and mechanisms of action of selected neurotoxic compounds.

Neuronopathies

Certain chemicals selectively act on neurons, causing their injury or, when relatively high doses are taken, their death. The loss of a neuron is irreversible, since it includes degeneration of all neuronal components, such as dendrites and axons, and the myelin. Some typical characteristics of the neuron are a high metabolic rate, a cellular process supported by the cell body, and an excitable membrane that can be rapidly polarized and depolarized.

Neurotoxic compounds share certain common characteristics. Most of these compounds selectively interact with the neuron. The initial injury to the neuron is followed by apoptosis or necrosis, leading to permanent loss of the neuron. Some neurotoxic compounds may act specifically on certain neuronal subpopulations and, as a result, injury and interruption of particular functions may occur.

Methylmercury

The neurotoxicity of mercury compounds has received worldwide attention for centuries. However, it was not until the outbreak of methylmercury poisoning in Minamata, Japan, in the 1950s, that the extreme toxicological consequences of human exposure were fully recognized. During this outbreak of poisoning, thousands of inhabitants were exposed to methylmercury. The source of exposure was found to be industrial effluent that contained large amounts of mercury. The mercury made its way into Minamata Bay, where it was converted to methylmercury by marine biota. This accumulated in fish and shellfish, which were eventually consumed by the local residents. This consumption of contaminated seafood continued for several years and resulted in hundreds of reported cases of methylmercury poisoning. Since that time, methylmercury poisoning has been referred to as 'Minamata disease.'

A second major outbreak of methylmercury poisoning occurred in Iraq during 1971–72. In this case, the source of exposure was the ingestion of seed grain treated with a methylmercury fungicide; the grain had been ground into flour to make bread. Approximately 6500 people were hospitalized due to poisoning, and at least 400 died.

Methylmercury poisoning has several important features: (1) a long latent period, usually lasting several months; (2) damage almost exclusively limited to the central nervous system; (3) areas of damage to the brain are highly localized (focal; e.g., in the visual cortex and the granular layer of the cerebellum, especially in the infolded regions, or sulci); (4) effects in severe cases are irreversible due to destruction of neuronal cells; (5) the earliest effects are nonspecific subjective complaints, such as paresthesia, blurred vision, and malaise, progressing to concentric constriction of the visual field, deafness, dysarthria, ataxia, and, in the worst cases, coma and death.

Observations of methylmercury exposure in both humans and animals indicate that the developing central nervous system is more sensitive to damage from methylmercury than is the adult nervous system. The first indications arose from the Minamata outbreak of methylmercury poisoning when it was found that mothers who were slightly poisoned gave birth to infants with severe cerebral palsy. Subsequent studies on experimental animals confirmed the increased sensitivity of the fetus.

The clinical picture is dose dependent. In those infants who have been exposed to high maternal blood levels of methylmercury, microcephaly, hyperreflexia, and gross motor and mental impairment, sometimes associated with blindness or deafness, are the main pattern. After exposure to lower doses, patients show mainly psychomotor impairment and persistence of pathological reflexes.

Inhibition of protein synthesis in target nerve cells at the initial stage of synthesis, associated with transfer RNA, is well documented and appears before the first clinical signs of intoxication. The mechanisms involved include interaction with sulfhydryl groups of active ribosomes, effects on membrane structure, integrity of microtubules, glucose transport across the blood–brain barrier, axonal transport and inhibition of acetylcholine transmission at the neuromuscular junction, and inhibition of neuronal cell division by causing metaphase arrest, presumably by disruption of the mitotic spindle.

Methanol

Like most hydrocarbon solvents, methanol (methyl alcohol) can produce reversible sensory irritation and narcosis at airborne concentrations below those producing organ system pathology. Serious methanol toxicity is most commonly associated with ingestion. When untreated or treated too late, acute methanol poisoning in humans is characterized by an asymptomatic latent period of 12–24 h followed by formic acidemia, ocular toxicity, coma, fatal cerebral edema with seizures as a preterminal event, and, in extreme cases, death. Visual disturbances generally develop between 18 and 48 h after ingestion and range from mild photophobia and misty or blurred vision to markedly reduced visual acuity and complete blindness. Although there is a considerable variability among individuals in susceptibility to methanol toxicity, a frequently cited lethal oral dose is $1\,\mathrm{ml\,kg^{-1}}$. Blindness and death have been reported with dosages as low as $0.1\,\mathrm{ml\,kg^{-1}}$.

The target of methanol within the eye is the retina, specifically the optic disk and optic nerve. Optic disk edema and hyperemia are seen, along with morphological alterations in the optic nerve head and the intraorbital portion of the optic nerve. Both axons and glial cells exhibit altered morphologies. Evidence is accumulating that Müller cells, neuroglia that function in the maintenance of retinal structure and in intra- and intercellular transport, are early targets of methanol. Rods and cones, the photoreceptors of the retina, are also altered functionally and structurally due to mitochondrial disruption, including ATP reduction in Müller and photoreceptor cells.

Methanol exerts its primary toxicity after undergoing biotransformation by alcohol dehydrogenase to formaldehyde and then to formic acid by the action of aldehyde dehydrogenase. The formic acid is thought to be responsible for both the ocular (blindness) and the acid–base toxicity of methanol.

While there is still much to be learned about mechanisms of methanol toxicity, what is known allows for effective therapies, if they are applied in a timely manner. Sodium bicarbonate is usually given to correct severe acidosis. Metabolic blockade is usually achieved with ethanol or 4-methylpyrazole, both acting as effective competitive inhibitors of alcohol dehydrogenase by virtue of their greater affinity than methanol for the enzyme.

Axonopathies

The neurotoxic disorders termed axonopathies are those in which the primary site of toxicity is the axon itself. The axon degenerates with the myelin surrounding it; however, the neuron cell body remains intact. It is now understood that neurotoxic chemicals cut off the axon at some point along its length, and the axon distal to the transection, biologically separated from the cell, degenerates. Longer axons are more affected in toxic axonopathies. The involvement of long axons of the central nervous system, such as ascending sensory axons in the posterior columns or descending motor axons, along with long sensory and motor axons of the peripheral nervous system, suggested the frequent use of the term 'central peripheral distal axonopathies' for this type of neurotoxic event.

A basic difference exists in the significance of axonal degeneration in the central nervous system compared with that in the peripheral nervous system: peripheral axons can regenerate whereas central axons cannot. In the peripheral nervous system, glial cells and macrophages support axonal regeneration, and Schwann cells maintain this ability. In the central nervous system, release of inhibitory factors from the damaged myelin and astrocyte scarring interfere with regeneration. When this glial interference is removed through transplantation of neurons from the central nervous system to the peripheral nervous system, the neurons are capable of extending neurites. The inability of the central nervous system to regenerate appears to be due to both limiting environmental glial factors and properties of the mature neuron.

Axonopathies can be considered to result from a chemical transection of the axon. The number of neurotoxic chemicals acting on the axon is large, but all of them share the common pathologic loss of distal axons with the survival of the cell body. Axonal transport appears to be the target in many of the toxic axonopathies. When the axons degenerate the result is frequently the clinical condition of peripheral neuropathy, in which sensory and motor pathways are first damaged in the most distal parts, such as the feet and hands. With time and continued injury the damage progresses toward more proximal areas of the body and the long axons in the spinal cord. The potential for regeneration is significant when the damage is limited to peripheral nerves, and complete in axonopathies in which early events are recognized and removed.

Organophosphate-Induced Delayed Polyneuropathy

Organophosphorus compounds (OPCs) are widely used in industry, agriculture, and maintenance of community hygiene. Organophosphate-induced delayed polyneuropathy (OPIDP) is unique in that it is caused by a single exposure to certain OPCs, with effects

appearing after 7–21 days. In addition, the most important reasons historically for interest in OPIDP were the thousands of cases of poisoning with triorthocresyl phosphate that occurred mainly due to beverage and food contamination in the United States in 1930 and in Morocco in 1959. By the end of twentieth century, there were many human cases of OPIDP due to different OPC-based insecticides.

OPIDP is a relatively rare neurodegenerative disorder in humans characterized by loss of function and ataxia of distal parts of sensory and motor axons in peripheral nerves and ascending and descending tracts of the spinal cord. The early neurological symptoms usually are sharp, cramplike pains in the calves, tingling in the feet and hands, followed by distal numbness and paresthesia. Symptoms of pain and weakness in muscles develop rapidly and patients become unsteady and unable to keep their balance. Progressive leg weakness occurs, together with depression of tendon reflexes. Symptoms may also appear in the arms and forearms. Sensory loss may be mild. Muscle tonus of the limbs gradually increases and spasticity appears in the lower limbs. Physical examination reveals distal symmetrical, mainly motor, polyneuropathy, with wasting and flaccid weakness of distal limb muscles, especially in the lower limbs. In severe OPIDP, quadriplegia, with footdrop and wristdrop, is observed, as well as mild pyramidal signs. There may be some functional recovery, but pyramidal and other signs of central neurological involvement may become more evident with time. The recovery affects only sensory nerves, while motor neurons may permanently lose function. The prognosis for functional recovery depends on the degree of pyramidal involvement, with ataxia and paralysis representing a permanent outcome of severe OPIDP. In children up to 12 years old, OPIDP rarely develops, appearing in mild form, and recovery is usually complete within 6 months.

Electrophysiological evaluation has revealed partial denervation of affected muscles with abnormal spontaneous activity (fibrillation potentials, positive sharp waves), increased insertional activity, reduced interference pattern, and large polyphasic motor unit potentials. The muscle action potentials are reduced in amplitude, terminal motor latencies are delayed, and with motor conduction velocity slightly reduced or normal.

Histopathological studies have described Wallerian degeneration affecting myelin sheaths of peripheral nerves and involvement of the central nervous system, with changes in fasciculus gracilis at the cervical level and on the corticospinal tracts at lumbar levels. The extent of involvement of the nervous system is dependent on the chemical structure of the OPCs, the dose administered, the clinical signs of exposure, and the time elapsed until the appearance of signs of OPIDP, and assessment of these features depends on the site and method of tissue sampling. Aggregation, accumulation, and partial condensation of neurofilaments in peripheral nerves prior to onset of OPIDP have been described. After the appearance of OPIDP, mitochondrial accumulation was observed. In the central nervous system there was degeneration of myelin sheets and axons in the cerebrospinal system involving 'swelling of axons.' After the onset of paralysis, the disappearance of muscular tissue and its replacement with connective and lipid tissue was described.

In numerous experimental studies a correlation was established between the clinical signs of OPIDP and activity of a protein, initially named 'neurotoxic esterase' and later 'neuropathy target esterase (NTE).' NTE is an integral membrane protein in vertebrate neurons and its activity depends on lipid content. It is present in all neurons, but is absent from glia. The active site of NTE contains Ser966 and two aspartates, Asp960 and Asp1086, that appear essential for enzymatic activity. It was recently shown that NTE is anchored in the endoplasmic reticulum, interacting with the cytoplasmic face of the endoplasmic reticulum, and that overexpression of NTE causes aggregation of endoplasmic reticulum.

Physiological functions of NTE are presently unknown. It was recently shown that human NTE catalyzes hydrolysis of several naturally occurring membrane-associated lipids, indicating that it may be involved in intracellular membrane trafficking. A possibility that NTE may have an important function during brain development through involvement in a cell-signaling pathway between neurons and glial cells is under investigation. Elucidation of the molecular mechanisms of OPIDP may provide valuable clues to the functioning of the normal adult nervous system and during neurodegenerative disorders.

n-Hexane

n-Hexane is used as a solvent in the extraction of vegetable oils, production of glue and in rubber, and in the shoe industry. *n*-Hexane is metabolized oxidatively to a number of compounds, including 2,5-hexanedione, which is thought to be the ultimate neurotoxic agent. 2,5-Hexanedione is also a toxic metabolite of *n*-butyl ketone and several other γ-diketones. γ-Diketones react with amino groups in all tissues to form pyrroles, and this step is important in pathogenesis of this axonopathy. The cellular changes are

characterized by development of neurofilament aggregates and swelling in the distal parts of the axon.

Peripheral neuropathy has been reported in humans exposed occupationally to *n*-hexane or through solvent abuse of glues or solvents containing *n*-hexane. The first signs of *n*-hexane neuropathy are symmetrical paresthesia and weakness. The lower extremities are affected first. A number of other symptoms, such as headache, anorexia, and dizziness, may precede or coincide with the neuropathy. A 'stocking and glove' sensory impairment to touch, pain, vibration, and temperature develops, with weakness and atrophy affecting proximal muscles of the extremities. On clinical examination, most patients show reduced body weight and diminished or absent reflexes. There is a marked reduction in conduction velocity in sensory and motor nerves. Electromyography shows a neuropathic pattern including fibrillation potentials, positive sharp waves, polyphasic potentials, and reduced interference patterns. Axonal lesions of the large myelinated fibers, with axonal swelling followed by myelin retraction at the node of Ranvier, are found on nerve biopsy. There is only a slight to moderate reduction in the number of fibers, mainly of large myelinated fibers. With continued intoxication, swellings are seen more proximally and there is degeneration of the distal axon along with its myelin. Muscle biopsies show signs of myofibrillar atrophy. Recovery is always gradual and slow and signs of residual neuropathy may persist in severe cases for 3–4 years. Signs of pyramidal tract defects, such as residual hyperreflexia and spasticity, and autonomic defects have been reported. Electrophysiological evidence of central nervous system dysfunction (abnormal evoked potentials) has occasionally been reported in individual cases of peripheral neuropathy.

Carbon Disulfide

Carbon disulfide (CS_2) is a volatile solvent used for a variety of industrial purposes. Many cases of human CS_2 poisoning, characterized by various neurological and behavioral effects, occurred during and after World War II. The psychological effects consisted of personality changes, irritability, memory deficits, insomnia, bad dreams, decreased libido, and constant fatigue. Paresthesia and dysesthesia from CS_2 exposure tend to occur in a 'stocking and glove' distribution characteristic of peripheral nerve injury due to a slowing of motor conduction in peripheral nerves. In addition to this chronic axonopathy, CS_2 can also lead to aberrations in mood and signs of diffuse encephalopathy. Some of these are transient at the beginning and later become long-lasting, a feature that is common in vascular insufficiency in the nervous system. Various sensory alterations have been reported following CS_2 exposure, particularly in vision, such as central field loss, disturbances in color vision, enlargement of the blind spot, and reduction in peripheral vision. Alterations in auditory, vestibular, and olfactory senses have also been observed.

Experimental CS_2 neuropathy consists of weakness or loss of power in the limbs, depressed reflexes, and general behavioral suppression. CS_2 produces filamentous inclusions in axons in both the central and peripheral nervous systems that are similar to those caused by *n*-hexane. However, there is a an important difference between CS_2 and *n*-hexane toxicity: *n*-hexane has to be metabolized to its active metabolite, 2,5-hexanedione, to be toxic, while CS_2 is the ultimate toxicant; it interacts with protein amino groups to form the dithiocarbamate adducts involved, through a series of reactions, in the pathogenesis of CS_2 axonopathy.

Acrylamide

Acrylamide is a vinyl monomer used in the manufacture of paper products, water treatment, crude oil production, and mineral ore and concrete processing. Smaller quantities are used as cosmetic additives and in food processing, and have application in molecular biology for acrylamide gel electrophoresis. Acrylamide is a potent neurotoxin affecting both the central and peripheral nervous systems. Only the acrylamide monomer is toxic.

Most cases of human acrylamide toxicity have occurred in the context of chronic occupational exposure, with predominant routes believed to be combined inhalational and dermal. Excessive sweating of the extremities is an early and universal symptom of chronic acrylamide toxicity. Other initial symptoms include inflammation of the skin of the hands and feet with blistering and desquamation. The acrylamide neuropathy is a toxic distal axonopathy, beginning with degeneration of the nerve terminal. Within nerve terminals early events include decreased numbers of synaptic vesicles and mitochondria and accumulation of neurofilaments with terminal sprouting. Multifocal accumulations of membranous bodies, mitochondria, and neurofilaments are observed in the distal axon, suggesting impaired retrograde axonal transport. The clinical course is characterized by the development of symptoms and signs of a motor and sensory peripheral neuropathy that slowly progress in severity if exposure continues. Symptoms initially involve the hands and feet and may progress to involve the entire upper and lower extremities. Symptoms include paresthesias and numbness, coldness, and difficulty with fine

movements such as writing. Signs include impaired touch or vibration sense in a 'stocking and glove' distribution, impaired joint position sense, absent tendon reflexes, and atrophy of the small muscles of the hand. Features suggesting involvement of the central nervous system include tremor, slurred speech, and hallucinations. Muscle pain and weakness are less common. Complete recovery usually occurs over a period of weeks to months following removal of the person from the toxic environment, preventing further exposure.

Myelinopathies

Myelin provides electrical insulation of neuronal processes and its absence leads to a slowing of conduction and aberrant conduction of impulses between adjacent processes. The diverse, structurally unrelated chemicals that cause toxic myelinopathies have been investigated and can be categorized into two groups. Some demyelinating chemicals seem to leave intact the myelinating cells (oligodendrocytes in the central nervous system and Schwann cells in the peripheral nervous system), while others damage the myelinating cells as well as the myelin. The significance between the two is that with the myelinating cells still intact, repair of the myelin sheath can occur. However, if the myelinating cells are destroyed, repair and reversal of the neuropathy may not occur. Histologically, these chemicals produce an edema of the white matter of the brain. By electron microscopy, vacuoles can be seen in the myelin surrounding axons; these are characterized as fluid-filled separations of myelin lamellae. Vacuoles can further degenerate to full demyelination, affecting conduction through those axons. Regeneration of the myelin layers can occur. Some of these chemicals, however, can increase cerebrospinal fluid pressure in the brain, optic nerve, and spinal cord, and/or intraneuronal pressure in the perineurium surrounding the axons in the peripheral nervous system. This increased pressure has been correlated with decreased conduction capacity through the axon, ischemia to the neuronal tissue from decreased blood flow, and, subsequently, permanent axonal damage.

Hexachlorophene

Hexachlorophene (2,2′-methylenebis-3,4,6-trichlorophenol) induces, in both humans and animals, vacuolation of central and peripheral myelin by the separation (splitting) of myelin lamellae at the intraperiod lines of the oligodendroglial and Schwann cells. Experimental studies with erythrocyte membranes show that hexachlorophene binds to cell membranes, resulting in the loss of ion gradients across the membrane. It is possible that hexachlorophene causes the loss of the ability to exclude ions from between layers of myelin, and that, with ion entry, water also separates the myelin layers, forming edema. Another related effect is uncoupling of mitochondrial oxidative phosphorylation by hexachlorophene, because this process is dependent on a proton gradient. Intramyelinic edema is reversible in the early stages, but with increasing exposure, hexachlorophene causes demyelination. Swelling of the brain causes increased intracranial pressure which may be fatal. The optic nerve fibers are especially vulnerable, and may not recover. Humans exposed acutely to hexachlorophene may have generalized weakness, confusion, and seizures, progressing to coma and death.

Lead

Peripheral neuropathy is a classical manifestation of lead toxicity, particularly the footdrop and the wristdrop that characterized the workers in past centuries who were occupationally exposed to lead for decades. Segmental demyelination and possibly axonal degeneration follow lead-induced Schwann cell degeneration. Wallerian degeneration of the posterior roots of sciatic and tibial nerves is possible, but motor nerves are apparently more sensitive to lead than are sensory nerves.

Toxic effects of lead are different in adults and children. For adults, the major concerns are peripheral neuropathy with injury of long neurons, chronic nephropathy, and hypertension. Other symptoms include abdominal pain, anemia, headache, ataxia, and memory loss. In children the symptoms are lead encephalopathy, lethargy, vomiting, irritability, loss of appetite, dizziness progressing to ataxia, reduced level of consciousness, coma, and death. It appears that lead exposure has an adverse effect on the intellectual abilities of children.

The pathologic findings at autopsy are severe edema of the brain due to extravasation of fluid from capillaries in the brain, accompanied by the loss of neuronal cells and an increase in glial cells. Vacuoles are seen as the result of splitting of both the major dense line and the intraperiod lines of the myelin. It has been shown that there is vascular injury as the initial neuropathological event, followed by endothelial, microglial, and astrocytic proliferation and widespread edema. The changes in the permeability of the blood–brain barrier for lead, possibly as the result of activation of protein kinase C in microvessels, have been suggested as one of the initial pathological events in lead neuropathy. Lead has been shown to accumulate in the endoneurium at potentially toxic levels, resulting in segmental demyelination before development of edema. Movement of lead into the endoneurial compartment is associated with ultrastructural abnormalities of Schwann cells,

which show degenerative cytoplasmic changes, as well as intranuclear inclusions typical of lead poisoning.

A highly significant morphologic effect in the nervous system is the result of lead impairment of timed programming of cell–cell connections, resulting in modification of neuronal circuitry. Lead impairs normal calcium homeostasis and uptake by calcium membrane channels and substitutes for calcium in calcium–sodium ATP pumps. Lead also blocks entry of calcium into nerve terminals; it inhibits calcium uptake in brain mitochondria, with a decrease in energy production to perform brain functions. Lead interferes with calcium receptors that are coupled with second-messenger functions.

Neurotransmission-Associated Neurotoxicity

There are many chemicals that interact with the process of neurotransmission. Among them are natural toxins and synthetic drugs. These chemicals may interrupt the transmission of impulses, block or support transsynaptic communication, block reuptake of neurotransmitters, and interfere with second-messenger systems. In general, the acute toxic effects of these compounds are directly related to the immediate concentration of the chemical at the active site, which is comparable to the level of the chemical in the blood. Toxic effects of these substances are usually short-term interactions that are reversible after removal of the substance from the blood. Their effects may be counteracted by the use of appropriate antagonists. The effects of certain substances may be irreversible. Both reversible single-dose toxicity and the effects of repeated exposure are common characteristics of substances interacting with neurotransmission.

Nicotine

Nicotine is a widely available natural alkaloid having a considerable medical significance because of its toxicity, its presence in tobacco, and its propensity for conferring a dependence on its users. Nicotine exerts its effects by binding to nicotinic receptors located in ganglia, at the neuromuscular junction and within the central nervous system. Tobacco smoking and low doses of nicotine accelerate heart rate, elevate blood pressure, constrict blood vessels in the skin, and facilitate neurotransmission. When larger doses of nicotine are applied, the initial stimulation is quickly followed by a blockade of transmission. Nicotine also possesses a biphasic action on the adrenal medulla; small doses evoke the discharge of catecholamines and larger doses prevent their release in response to splanchnic nerve stimulation.

Poisoning from nicotine may occur from accidental ingestion of insecticide sprays in which nicotine is present as the toxic agent, or, in children, from ingestion of tobacco products. In each of these settings, the rapid rise in circulating levels of nicotine leads to excessive stimulation of nicotine receptors followed by ganglionic paralysis. The onset of symptoms of acute severe nicotine poisoning is rapid – symptoms include nausea, salivation, abdominal pain, vomiting, diarrhea, cold sweat, headache, dizziness, disturbed hearing and vision, mental confusion, and weakness. Faintness and prostration ensue; the blood pressure falls, breathing is difficult, the pulse is weak, rapid and irregular, and collapse may be followed by terminal convulsions. Somnolence, confusion, and coma may occur; if death results it is often the result of paralysis of respiratory muscles.

Exposure to lower doses of nicotine for longer duration is very common, and the health effects are a significant epidemiological problem. In smokers the effects of nicotine are accompanied by the effects of other components of cigarette smoke. The complications include cardiovascular disease, cancers, chronic pulmonary disease, and attention-deficit disorders in children of women who smoke during pregnancy. High levels of nicotine in smokers may be a factor in increased propensity for platelets to aggregate and in myocardial ischemia. In addition, nicotine inhibits apoptosis and may have a direct role in tumor promotion and tobacco-related cancers.

Conclusion

A large number of chemicals act on the nervous system and can influence its functions. Among them are the substances used in medicine as drugs (anesthetics, analgesics, sedatives), those consumed as life-style factors (ethanol, nicotine, caffeine). In addition, humans are exposed to many potent chemicals (pesticides, heavy metals, organic solvents, additives) present in their environment, food, and water. The nervous system is very sensitive and vulnerable to these agents. In the nervous system there are four targets for the action of chemicals: the neuron, the axon, the myelinating cell, and the neurotransmission system; each target has clinical, histopathological, and toxicological features and mechanisms due to the neurotoxic effects of chemicals.

See also: Excitotoxicity in Neurodegenerative Disease; Neuropathy: Metabolically-Induced; Neuropathy: Peripheral; Neurotoxins; Neurotoxins and their Neurotoxicology; Nicotine.

Further Reading

Anthony DC, Montine TJ, and Graham DG (1996) Toxic responses of the nervous system. In: Klaassen CD (ed.) *Casarett and Doull's Toxicology. The Basic Science of Poisons*, 5th edn., pp. 463–486. New York: McGraw-Hill.

Garman RH, Fix AS, Jortner BS, et al. (2001) Methods to identify and characterize developmental neurotoxicity for human health risk assessment. II: Neuropathology. *Environmental Health Perspectives* 109(supplement 1): 93–100.

International Programme on Chemical Safety (1986) Principles and methods for the assessment of neurotoxicity associated with exposure to chemicals. *Environmental Health Criteria 60*. Geneva: World Health Organization.

International Programme on Chemical Safety (1990) Methylmercury. *Environmental Health Criteria 101*. Geneva: World Health Organization.

International Programme on Chemical Safety (1991) *n*-Hexane. *Environmental Health Criteria 122*. Geneva: World Health Organization.

Jokanovic M, Kosanovic M, and Stukalov PV (2004) Organophosphate induced delayed polyneuropathy. *Medicinal Chemistry Reviews Online* 1(2): 123–131.

Jortner BS (2000) Mechanisms of toxic injury in the peripheral nervous system: Neuropathologic considerations. *Toxicological Pathology* 28: 54–69.

Lotti M and Moretto A (2005) Organophosphate-induced delayed polyneuropathy. *Toxicological Reviews* 24(1): 37–49.

Slikker W Jr. (1994) Principles of developmental neurotoxicology. *Neurotoxicology* 15: 11–16.

Spencer PS and Schaumburg HH (eds.) (2000) *Experimental and Clinical Neurotoxicology*. New York: Oxford university Press.

Neuropathy: Metabolically-Induced

R E Schmidt, Washington University School of Medicine, Saint Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Diabetic Neuropathies: Overview

The public health importance of diabetic neuropathy, the most common neuropathy in Europe and North America, will increase substantially as the result of the looming global epidemic of type 2 diabetes. The prevalence of diabetic neuropathy depends on criteria selected, ranging from 10% to 15% of diabetics who are clinically symptomatic to 48–90% of diabetics with electrophysiologically detectable dysfunction. A variety of clinical presentations include sensory loss symmetrically involving the distal extremities (stocking–glove pattern), occasionally complicated by disabling neuropathic pain; autonomic insufficiency involving many modalities; and, asymmetrical presentation of pain and weakness. Patients may develop neuropathy as adults or children, in both type 1 and type 2 diabetes and following years of diabetes or at its initial presentation. Risk factors for diabetic neuropathy include duration of diabetes, age, hypertension and smoking (in type 1 diabetics), patient height, alcohol consumption, and worsening degree of glycemic control. Neuropathy is commonly associated with other diabetic complications including retinopathy and nephropathy. Multiple presentations suggest that diabetic neuropathies may have different causative (pathogenetic) mechanisms and require different therapies.

Diabetic neuropathy is not a single neuropathy with a single pathogenetic mechanism, rather it is a complex spectrum of neuropathies (Table 1). Subclassification of diabetic neuropathies is clinically significant since different types of diabetic neuropathy may require individualized therapy. Some newly recognized forms of diabetic neuropathy (e.g., diabetic lumbosacral radioplexus neuropathy (LRPN)) may reflect a possible immune pathogenesis and the therapeutic possibility of treatment by immune suppression. Furthermore, not all diabetic patients with peripheral nerve dysfunction have diabetic neuropathy; specifically, neuropathy in an individual patient may represent a second superimposed process (e.g., vasculitis (blood vessel inflammation or destruction) or chronic inflammatory demyelinating neuropathy (CIDP)) rather than diabetes itself. Thus, the initial workup (including a detailed history, electrophysiologic testing, determination of the presence of a paraprotein (an abnormal plasma protein) or autoantibodies (antibodies directed against normal tissue constituents) of diabetic patients must address the possibility of additional factors contributing to the neuropathy. An atypical presentation dominated by motor symptoms, severe sensory loss (particularly joint and vibration sense out of proportion to the severity of diabetes), asymmetry of involvement, or an acute, rapidly worsening course represent grounds for nerve biopsy. New studies of cutaneous nerves using multiple skin biopsies will allow the ability to monitor the effect of therapy on autonomic and sensory axons in diabetic patients.

Neuropathy with Impaired Glucose Tolerance

Recently, in patients without a history of frank diabetes, an idiopathic, typically painful neuropathy has been described in the absence of elevated HbA1c or fasting hyperglycemia but in the presence of an impaired glucose tolerance test (GTT). Considering the frequency of abnormal GTT in the adult patient population and the frequency of neuropathy without established cause, some of these cases may represent a collision of both processes. Nonetheless, it has been reported that small axons, including those comprising the innervation of the epidermis, are prominently affected and may be the earliest detectable sign of neuropathy in the presence of glucose dysmetabolism. A syndrome of paresthesias (abnormal skin sensations with no apparent physical cause) and pain in newly diagnosed diabetes, that is, hyperglycemic neuropathy, or increased resistance to experimentally induced ischemia (decreased blood supply) may respond quickly to normalization of blood glucose valueso.

Symmetric Sensorimotor Polyneuropathy

The most common form of diabetic neuropathy is symmetric sensorimotor polyneuropathy which ranges in severity from subclinical electrophysiologic alterations to distal 'stocking–glove' anesthesia. The presence of neuropathy contributes to the increased risk of amputation in diabetes.

The pathologic hallmark of symmetrical sensory polyneuropathy is the variable loss of myelinated and unmyelinated axons (Figure 1), which is particularly severe in the distal aspects of the lower limbs. This form of diabetic neuropathy is characterized by axonal degeneration with the formation of numerous bands of Büngner (collections of Schwann cell processes forming a substrate for axonal regeneration)

and clusters of regenerating axons which may vary with severe or sustained neuropathy. Some patients, particularly those with chronic involvement or mild symptoms, show primary or secondary forms of demyelination, remyelination, and occasional onion bulbs (i.e., imbricated concentric proliferation of collagen and Schwann cell processes around a central axon), although this may reflect coexistence of diabetes and CIDP. Active axonal degeneration, numerous bands of Büngner, and regenerative clusters of small, thinly myelinated axons suggest axonal degeneration and regeneration. The examination of individual axons (osmium-stained to highlight myelin sheaths) by teased fiber analysis may be needed to discriminate axons with myelin sheaths that are too thin for axon caliber (produced as the end stage of axonal regeneration) from true demyelination. Axonal atrophy is not a consistent finding in human diabetic neuropathy. A common finding, thickened basement membranes, involve the endoneurial microvasculature, Schwann cells, and particularly perineurial cells, and may result from failure to degrade glycated basal laminae (i.e., containing nonenzymatic bound glucose or related saccharides). The microvasculature within the nerve fascicles is thickened (Figure 2) by multiple concentric layers of basal laminae and collagen, a pattern which, although commonly seen in diabetic nerve, is not specific for diabetes. Recent studies of diabetic neuropathy early in its clinical course, which may possibly elucidate the initial insult, have reported normal myelinated axon density, axonal area, and *g*-ratio (the ratio of axon diameter to diameter of axon + myelin sheath, one measure of myelinopathy) in the presence of paranodal abnormalities, segmental demyelination (myelin loss with preservation of axonal integrity), and remyelination in the absence of degeneration of myelinated axons.

Although neuropathy develops in both types 1 and 2 diabetes, differences in the pathology of diabetic somatic nerves have been proposed. Type 1 diabetic patients are reported to develop a lesser degree of axonal degeneration and more evidence of myelinopathy than type 2 diabetics, in whom axonal degeneration is more prominent. Anders Sima and colleagues have described axoglial dysjunction, an unusual and contentious ultrastructural and molecular alteration of the perinodal junctional apparatus resulting in paranodal demyelination and conduction, particularly in type 1 diabetes in rats and man. Diabetic BB/Wor rats (a model of type 1 diabetes) develop severe nerve conduction defects, node of Ranvier-based pathology with molecular alterations of nodal proteins, axonal atrophy, defects in axonal regeneration, and, eventually, significant loss of myelinated axons. In comparison, neuropathy in the type 2 diabetic BBZDR/Wor rat model exhibits a lesser degree of atrophy of myelinated

Table 1 Classification of the diabetic neuropathies

Impaired glucose tolerance and hyperglycemic neuropathy
Generalized (symmetrical) neuropathies
Acute painful (including treatment induced)
Sensorimotor
Autonomic
Acute motor
Focal and multifocal neuropathies
Lumbosacral radioplexus (Bruns–Garland syndrome)
Cranial
Thoracolumbar
Superimposed chronic inflammatory demyelinating neuropathy (CIDP)
Hypoglycemic neuropathy

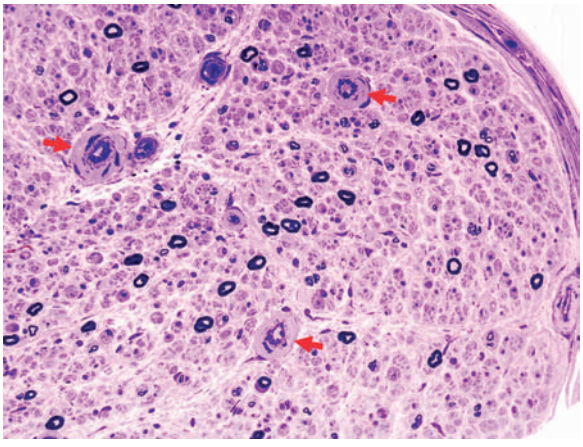

Figure 1 Endoneurial vessels in diabetic symmetrical sensory neuropathy are markedly thickened (arrows), a morphologic feature commonly seen in, but not confined to, diabetic neuropathy (1-μm-thick plastic embedded section, toluidine blue stain).

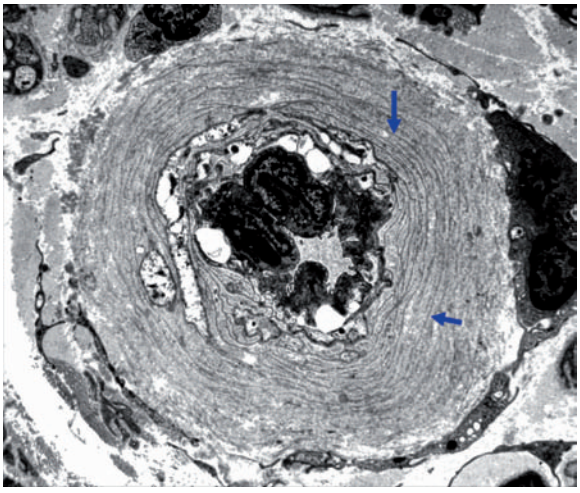

Figure 2 Endoneurial vessels in diabetic symmetrical sensory neuropathy are thickened by reduplication of basement membranes (arrows) (electron micrograph).

axons, minimal nodal pathology with no molecular changes, segmental demyelination, and active axonal degeneration with relative preservation of axon number, suggesting increased axonal regeneration.

Although axon loss in diabetic sensory neuropathy is typically diffuse and symmetrical distally, foci of axon loss suggested that it arises as the result of summation of ischemic foci. However, patchy intrafascicular or fascicle to fascicle variability of axon loss may be seen in CIDP and in inherited nonvasculopathic neuropathies, in which ischemia has no recognized role. Nonetheless, endoneurial microvasculopathy correlates with electrophysiologic dysfunction and may anticipate the development of significant neuropathy. However, the expected deficiency of nerve blood flow in diabetic patients with early polyneuropathy has not consistently been shown, even in the presence of decreased sensory nerve action potential amplitudes and numbers of axons. Therefore, argument continues whether microvasculopathy results in an ischemic pathogenesis in diabetic neuropathy or is unrelated to nerve dysfunction.

Recently, a number of investigators have begun to examine the terminal epidermal cutaneous innervation of the skin in patients with a variety of neuropathies using antibodies to PGP9.5 (which labels axons of all modalities) or selected neurotransmitters. A decrease in and disorganization of epidermal axons have been demonstrated in diabetic patients (Figure 3) and, in some cases, are accompanied by the formation of terminal axonal swellings at the distalmost portions of the cutaneous innervation. The ability to use skin biopsies to detect the early development of neuropathy and to follow the effect of therapeutic regimens holds much promise.

A number of diabetics develop debilitating painful neuropathy characterized by spontaneous pain or exaggerated response to normal stimuli (dysesthesias). Although it was initially suspected to reflect preferential damage to small caliber axons, painful diabetic neuropathy has not been consistently correlated with active fiber degeneration or regeneration in sural nerve biopsies. However, recent studies using skin biopsies suggest a correlation with intraepidermal sensory nerve fiber density in some, but not all, patients. Ectopic discharges related to pain perception have been demonstrated to arise from dorsal root ganglia (DRG) neurons and accompany changes in sensory neuron sodium channels. The presence of dystrophic axonal swellings, often immediately adjacent to normal-appearing cell bodies, containing neurofilaments and calcitonin-gene-related peptides (CGRP), trkA-, and substance P-immunoreactivity are thought to involve are thought to involve potentially dysfunctional intraganglionic sprouts of nociceptive small neurons, although there is no proven association with painful neuropathy.

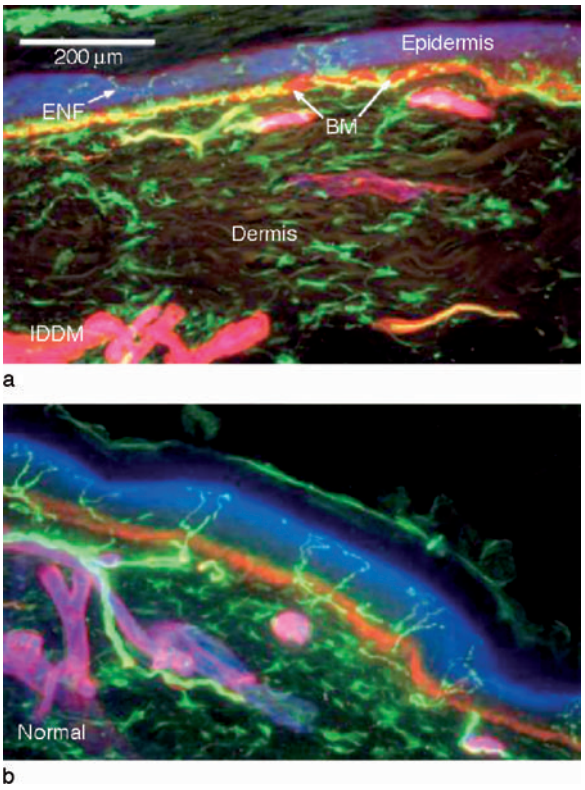

Figure 3 Immunohistochemical studies of skin biopsy of a type 1 diabetic patient with symmetrical sensory neuropathy shows decreased numbers and disorganization of epidermal nerve fibers (ENF, (a)) in comparison with an age-matched nondiabetic control (b). (Immunohistochemical detection of PGP 9.5, which stains all types of cutaneous nerves.) Images courtesy of Dr. William R. Kennedy.

Asymmetric Diabetic Neuropathies

Occasional diabetic patients, typically middle-aged or elderly men with early type 2 diabetes, may develop an unusual asymmetric neuropathy, often following institution of insulin therapy or associated with significant weight loss. Such patients experience sudden or subacute onset of motor dysfunction with muscle wasting, typically involving large nerves to the lower limb or cranial nerves. This unusual category of diabetic neuropathy includes diabetic LRPN (diabetic amyotrophy, truncal radiculopathy, Bruns–Garland syndrome, femoral neuropathy of diabetes, and proximal diabetic neuropathy), upper limb mononeuropathy, and cranial nerve palsies (usually cranial nerves 3, 4, and 6). Rarely, upper limb involvement occurs.

These cases have an unusual pathology which may reflect an autoimmune pathogenesis. An inflammatory (mostly T cell) nonnecrotizing epineurial microvasculitis

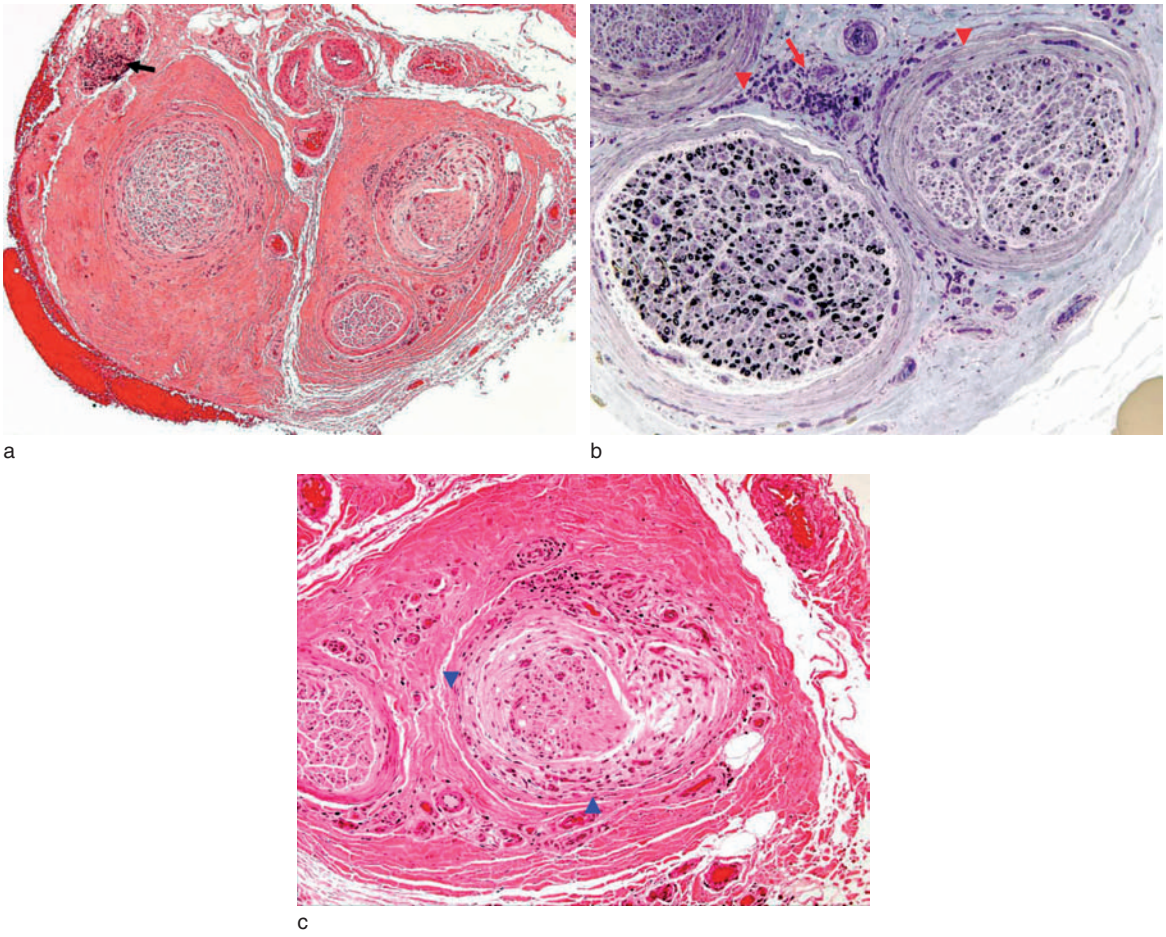

Figure 4 Diabetic LRPN is characterized by microvasculitis (arrows: (a) and (b); hematoxylin and eosin stain and 1 μm plastic section, respectively) and perineuritis (arrowheads: (b) and (c)).

(Figures 4(a) and 4(b)) may be prominent, often accompanied by significant axon loss, which varies from fascicle to fascicle, perineuritis (inflammation of the covering of individual fascicles) (Figure 4(c)), and endoneurial and perineurial hemosiderin deposition indicative of prior hemorrhage, in some cases forming neuromas. The pathologic findings in diabetic LRPN are identical to those in nondiabetic patients with LRPN, although it has been suggested that some nondiabetic patients with LRPN may also have abnormal glucose tolerance. The typical course results in spontaneous recovery over several months.

Diabetic Autonomic Neuropathy

Diabetic patients may develop a constellation of autonomic complaints involving any or several modalities. Although early on parasympathetic dysfunction was thought to be predominantly affected, sympathetic dysfunction occurs in diabetes and may be accompanied and exaggerated by superimposed enteric or visceral sensory dysfunction. Cardiovascular dysfunction may be manifest as abnormal R–R interval, often involved early on; later, cardiac reflexes may be lost and, in a minority of patients, orthostatic hypotension (i.e., significant loss of blood pressure, typically with change in position from recumbent to standing) develops. Genitourinary dysfunction includes erectile defects (in 35–50% of diabetic men) as well as loss of bladder tone (with subsequent distension and infection). Visceral sensory neuropathy may result in bladder distension without a strong voiding stimulus. Dysfunction may involve any level of the alimentary tract resulting in esophageal dysmotility, gastroparesis (slowed gastric emptying), loss of gall bladder contractility, or chronic diarrhea or constipation which may alternate. Patients may have difficulty sweating or sweat excessively in response to various stimuli (e.g., 'gustatory sweating' in anticipation of food). The presence of symptomatic diabetic autonomic neuropathy significantly worsens morbidity and mortality, due to cardiac arrhythmias, kidney failure, stroke, or increased perioperative mortality. Visceral dysfunction in diabetes also involves

subpopulations of cells of the enteric nervous system, the nonadrenergic noncholinergic pathway in the pylorus and the interstitial cells of Cajal. Although symptomatic autonomic neuropathy is considered infrequent, the involvement of autonomic functions in the extremities commonly parallels somatic neuropathy and may have different pathogenetic mechanisms than visceral autonomic neuropathy.

Systematic autopsy studies of the sympathetic autonomic nervous system in diabetic patients and rats have demonstrated the preservation of neurons in prevertebral (located adjacent to main vascular trunks serving viscera) and paravertebral (bilateral chains of ganglia adjacent to vertebral bodies serving somatic and selected visceral functions) sympathetic ganglia. Instead, the pathologic hallmark in the diabetic prevertebral sympathetic ganglia is the development of markedly swollen neurofilament- and membrane-laden axons and synapses, designated neuroaxonal or neuritic dystrophy (Figure 5(a)), accompanied by the loss of normal terminals (which may be the most functionally relevant finding). Marked differences in involvement of subpopulations of terminal axons suggests selective vulnerability (Figures 5(b) and 5(c)) and may account for the targeting of sympathetic prevertebral and lumbosacral chain ganglia while largely sparing superior cervical chain ganglia. The often-quoted classic study of the sympathetic ganglia of diabetic patients with symptomatic alimentary tract autonomic neuropathy demonstrated dystrophic changes accompanied by lymphocytic infiltrates, the latter finding interpreted as evidence for a possible autoimmune pathogenesis. However, inflammatory cells were neither more frequent nor of greater intensity in diabetic sympathetic ganglia in a large series of autopsied patients compared to those of nondiabetic subjects. Loss of myelinated axons in the preganglionic greater splanchnic nerves (i.e., preganglionic axons to prevertebral sympathetic ganglia) and abdominal vagus nerves in diabetic gastroparesis has also been described. Diabetic mouse parasympathetic ganglia experience the

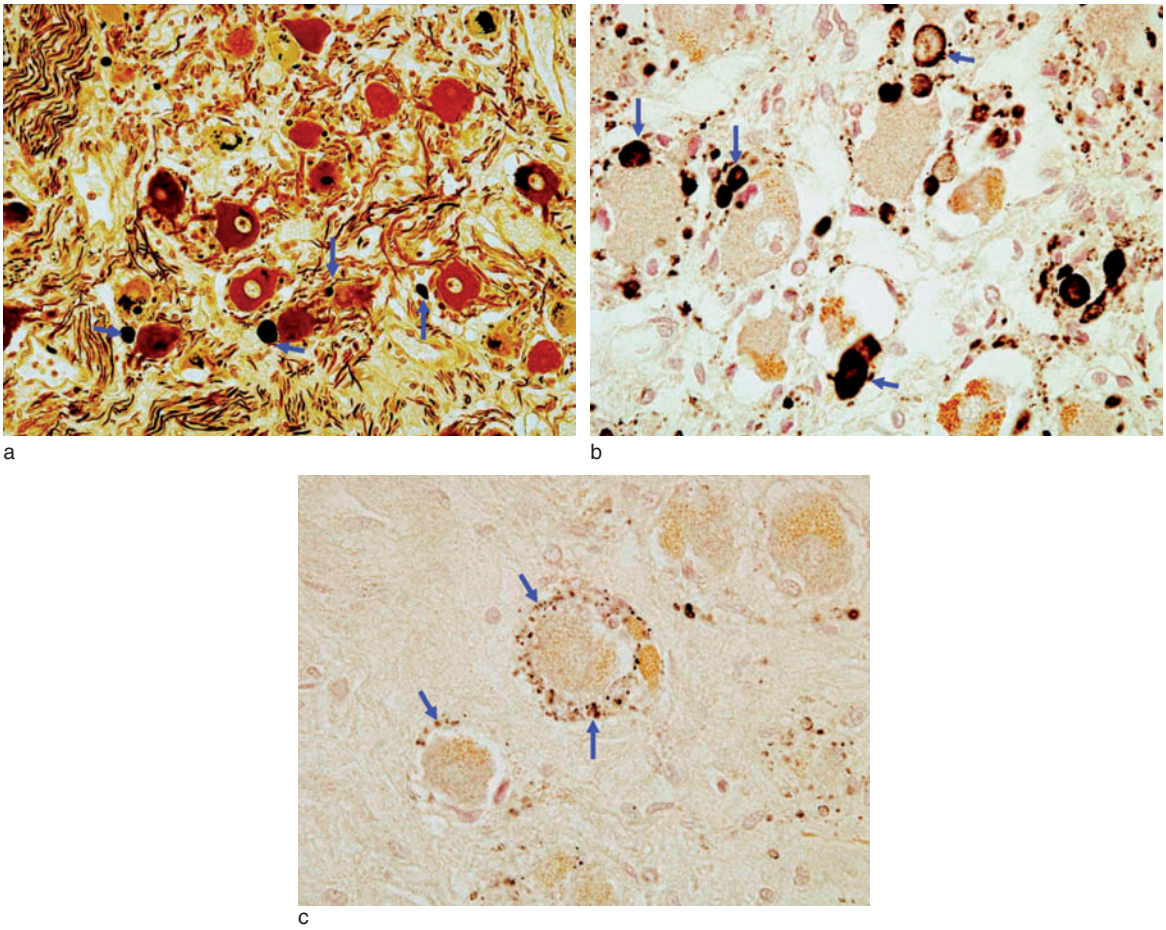

Figure 5 The prevertebral superior mesenteric sympathetic ganglion in a patient with diabetic autonomic neuropathy shows perisomal argyrophilic dystrophic axons (arrows, (a); Bielschowsky axon stain) which are markedly immunoreactive for neuropeptide Y (arrows, (b)) but do not stain with antibody against substance P (arrows indicate normal terminals, (c)).

selective loss of nitrergic (using nitric oxide (NO) as a neurotransmitter) innervation of the stomach and penis; in other studies, subpopulations of autonomic fibers within the gut, bladder, penis, and other end organs may show preferential damage.

Pathogenetic Mechanisms of Diabetic Neuropathy

The pathogenesis of diabetic neuropathy is largely unknown; however, many pathogenetic mechanisms have been advanced with no current consensus. Considering the range of pathologies described in patients with diabetic neuropathy, multiple mechanisms may interact and vary between forms. It is believed that tight glucose control does delay the development of somatic and autonomic neuropathy; however, it does not appear to completely prevent it. Intensive insulin treatment of type 2 diabetics produced minimal improvement in established neuropathy. A variety of mechanisms have been proposed to play a role in its pathogenesis.

Advanced Glycosylation End Products Formation

Glycosylation of various intracellular and extracellular proteins in response to hyperglycemia may be followed by chemical rearrangement into advanced glycosylation end products (AGEs), some of which may result in cross-linking of a variety of proteins and subcellular organelles. Additionally, the binding of AGEs to their receptors (RAGEs), members of an immunoglobulin superfamily, is thought to increase oxidative stress. Aminoguanidine, an inhibitor of AGE formation, has shown a positive effect on a variety of disturbed functions in diabetic somatic nerves.

Apoptosis

A role for apoptosis (a complex, highly regulated pattern of cell death) in a large variety of developmental and pathologic processes has been proposed, including diabetic sensory and autonomic neuropathy. Largely based on *in vitro* studies, a prime role for glucose-induced apoptosis was proposed. However, this hypothesis conflicts with the maintenance of neuron number in dorsal root and sympathetic ganglia and only a minor degree of axon loss in the somatic nerves of the streptozotocin-induced (STZ)-diabetic rat. Experiments with rats diabetic for extended periods have shown caspase-3 activation in DRG neurons in the absence of neuronal degeneration, confirmed by nonbiased counting methods, although others have claimed a small decrease in subpopulations of DRG neurons. In contrast, loss of sensory neurons and axon numbers in diabetic mice suggests that even rodent models may differ.

Autoimmune Processes

A role for an autoimmune cell- or antibody-mediated pathogenesis has been proposed, particularly for diabetic autonomic neuropathy. Antibodies directed against adrenal gland, sympathetic ganglia, and vagus nerve may be associated with an increased incidence of autonomic neuropathy in diabetic patients, although its role in autonomic neuropathy is still contentious.

Axonal Transport

Diminished axonal transport of a variety of materials, initially thought to represent a primary pathogenetic mechanism, may primarily reflect decreased neuronal synthesis or altered production of trophic substances by target tissues, rather than a defect in the transport apparatus itself.

Ischemia

The relevance of ischemia to the pathogenesis of diabetic sensory neuropathy has been controversial in man and animal models. Nonetheless, evidence exists for a role of tissue hypoxia in nerve and ganglia in experimental animal models and humans, perhaps as the result of endothelial damage, red cell deformability, or defects in local oxygen release. The simultaneous presence of tissue hypoxia, alterations of NO metabolism with the formation of peroxynitrite, alterations of other vasoactive substances such as endothelin, and defective reactive hyperemia of nerve microvessels may combine to place diabetic nerves at risk. Although nerve ischemia is associated with dysfunction of diabetic nerves, it may not be a primary or major cause; rather, it may simply accelerate the pathologic process.

Mitogen-Activated Protein Kinases

Neurotrophic substances exert their effect by initiating a signaling cascade, only now being unraveled, which represents a possible pathogenetic mechanism and target for therapeutic intervention. Mitogen-activated protein kinases (MAPKs) are activated by excessive glucose and oxidative stress and may have an important role in the development of diabetic neuropathy possibly by inducing excessive neurofilament phosphorylation.

Neurotrophic Substances

Deficiencies in the amount, regulation, or function of a variety of neurotrophic substances (e.g., nerve growth factor (NGF), neurotrophin-3 (NT-3), insulin-like

growth factors (IGFs), even insulin itself) are thought to contribute to the development of sensory and autonomic neuropathies. Decreased endogenous target-derived and transported NGF may explain the decrease in sensory nerve expression of NGF-sensitive sensory neuropeptides CGRP and substance P in diabetic animal models. Unfortunately, clinical trials of exogenous recombinant NGF treatment have not produced a significant beneficial effect. Nerve brain-derived neurotrophic factor (BDNF) content is upregulated in diabetes and may exert a reparative or protective function. Similarly, NT-3 deficiency in diabetic animal models suggests that multiple neurotrophic elements may participate. IGF-I, decreased in diabetic serum and nerve, has antioxidant, survival-promoting effects. Exogenously administered IGF-I corrects hyperalgesia and deficits in axonal regeneration of diabetic rats *in vivo* and reverses established experimental autonomic neuropathy in diabetic rats. Direct application of insulin to diabetic somatic nerves (at doses that do not affect circulating glucose levels) unilaterally corrects nerve conduction velocity defects. C-peptide, cleaved from the proinsulin molecule during the secretion of insulin, has also been identified as a neurotrophic factor with a possible pathogenetic and therapeutic role in type 1 diabetes. Other trophic substances, including vascular endothelial growth factor (VEGF) and erythropoietin, may represent therapeutic avenues.

Oxidative Stress

Reduction in antioxidant defenses in the face of increased production of oxidant stressor molecules in diabetes may increase production of lipid peroxides, DNA damage, and mitochondrial injury with resultant damage to intracellular and extracellular proteins. Brownlee et al. have proposed a unifying hypothesis by which diabetes-induced increased glucose metabolism exaggerates the mitochondrial proton gradient and results in increased superoxide production and general oxidative stress. Others, however, have reported an insulin- and NT-3-sensitive decrease in mitochondrial membrane potential in adult cultured DRG neurons. Antioxidants have produced a salutary effect on biochemical and physiologic defects of diabetic nerves and have been approved for clinical usage in some countries.

Polyol Pathway

The polyol pathway in diabetic nerve and ganglia (normally a minor pathway of glucose metabolism) converts glucose to sorbitol and then to fructose using aldose reductase and sorbitol dehydrogenase, respectively. Increased flux through the polyol pathway is thought to increase oxidative stress, presumably by diminishing nerve glutathione content. Inhibition of aldose reductase, the first synthetic enzyme of the polyol pathway, has resulted in marked improvement of various neuropathic parameters in experimental animals, although clinical trials have been disappointing.

Protein Kinase C

Increased protein kinase C β II (PKCβ II) isoform activity has been reported to result in alterations of microvascular structure and function and induces transforming growth factor-β (TGF-β) effects on the extracellular matrix. Inhibition of PKC activity is reported to have salutary effects on experimental diabetic neuropathy.

Regeneration and Synaptic Dysplasia

Recently, an interest has developed in the consideration of the distal axon/nerve terminal as an important part of somatic and autonomic neuropathy in diabetes. Skin biopsies have demonstrated the loss of intraepidermal sensory axons in the absence of comparable findings in their more proximal portions in the sural nerve. Marked swelling of axon terminals (neuroaxonal dystrophy) in the sympathetic ganglia and of diabetic rodents and man has been proposed to reflect a diabetes-induced defect in cycles of synaptic turnover or remodeling, which may underlie synaptic plasticity. Synapse-targeted processes (particularly those related to synaptic vesicle function, membrane retrieval, and turnover of subcellular organelles) or failure to degrade transported materials at the nerve terminal are poised to contribute to the pathogenesis of terminal swellings. Impairment of axonal regeneration following sciatic nerve crush injury has also been reported in diabetic rodent models. Delayed and attenuated upregulation of IGF-I, NGF, p75 mRNA, or asynchronous changes in IGF-I receptor mRNA at injury sites in the distal sciatic nerve may significantly worsen plasticity in type 1 diabetic rats and contrasts with significantly milder changes in type 2 models.

Uremia

Chronic renal failure induces a distal sensorimotor neuropathy which is characterized by both axonal degeneration (more prominent in larger than in smaller myelinated axons) and secondary demyelination. In secondary demyelination, analysis shows concentration of demyelinated segments on atrophic axons with complete sparing of other adjacent normal axons, different from the random distribution of demyelinated segments on all myelinated axons. Its precise pathogenesis remains unknown, although a

number of pathogenetic mechanisms (presence of toxic circulating molecules including urea, ischemia, etc.) have been considered.

Vitamin Deficiencies

Beriberi (Thiamine)

Patients suffer from a distal symmetrical sensorimotor neuropathy in which pain, dysesthesias (unpleasant sensations induced by ordinary stimuli), and cutaneous hyperesthesia (an abnormal increase in sensitivity to sensory stimuli) may be prominent, as may dysphonia (e.g., abnormal speech such as hoarseness) from involvement of the recurrent laryngeal nerve. Both 'dry' (neuropathic) and 'wet' (cardiac form with edema) forms of beriberi are the result of nutritional deficiency, although it is not clear that it represents a pure thiamine deficiency disorder. In somatic nerves in the extremities, vagus and phrenic nerves, axonal degeneration is most prominent in distal portions of the nerves and (secondary) demyelination is manifest more proximally. Relative preservation of small myelinated and unmyelinated axons has been described in the presence of intrafascicular edema. Axonal degeneration is also demonstrated in the gracile columns in the spinal cord, in which distribution it represents a central–peripheral distal axonopathy. A distinctive accumulation of flattened membranous sacs accompanied by decreased numbers of cytoskeletal constituent neurofilaments and microtubules have been described in the distal branches of peripheral axons and central sensory projections.

Thiamine deficiency produces decreased nerve thiamine pyrophosphate (TPP), a cofactor in the decarboxylation of α-ketoacids, particularly the reaction of pyruvate and acetyl coenzyme A catalyzed by pyruvate dehydrogenase, and the formation and degradation of α-ketols.

Pellagra (Niacin)

In pellagra, there is a constellation of cutaneous lesions, gastrointestinal disturbances, behavioral manifestations and dementia, spinal cord involvement, and peripheral neuropathy. The peripheral neuropathy in pellagra is characterized by mild axonal degeneration. An epidemic of optic neuropathy, painful sensory neuropathy, hearing loss, and myelopathy developed in Cuba. This pattern of injury resembled Strachan's syndrome (amblyopia, orogenital dermatitis, and painful neuropathy). The precise relationship of multinutritional defects to discrete syndromes is complicated by the complex metabolic milieu existing in some patients.

Cobalamin (Vitamin B_{12})

Patients deficient in vitamin B_{12} develop a symmetrical sensory polyneuropathy characterized by axonal degeneration, not demyelination, which may or may not coexist with subacute combined degeneration of the spinal cord. Experimental studies suggest that the biochemical deficit involves methyl group addition reactions as the result of inadequate methionine and S-adenosyl methionine synthesis, a methyl folate trap, and formate starvation.

Vitamin E

Vitamin E deficiency develops most typically in conditions in which fat malabsorption develops (e.g., abetalipoproteinemia, cystic fibrosis, and congenital biliary atresia). Vitamin E deficiency in man, rat, and monkey results in spinocerebellar degeneration and a sensory neuropathy in which there is loss of large myelinated axons as part of a central–peripheral distal axonopathy and development of numerous dystrophic axons in the gracile nuclei. Vitamin E, a natural antioxidant, in deficient amounts may result in increased oxidative stress.

Hypothyroid Neuropathy

Nerve damage in hypothyroidism frequently results from focal compression and produces, most typically, a carpal tunnel syndrome. In addition, a more generalized sensory-symptom dominant polyneuropathy may also develop which responds to thyroid replacement therapy. Controversy exists whether segmental demyelination or axonal degeneration is the dominant pathology, but there is agreement that nerves show a decrease in large-diameter myelinated axons. Excessive amounts of glycogen accumulate in Schwann cells, axons, capillaries, and perineurial cells and increased numbers of subperineurial Renaut bodies (rounded collections of collagen, other cytoskeletal elements, and fibroblasts/perineurial cells) also are noted.

Acromegaly

Similar to the situation in hypothyroid neuropathy, the most common neuropathy in patients with acromegaly is compression injury resulting in carpal tunnel syndrome. In acromegalic patients, the cause of the neuropathy may be multifaceted since these patients have a metabolic syndrome that may also include diabetes, which complicates the interpretation of neuropathic symptoms and pathology. Nerve biopsies show decreased numbers of myelinated and unmyelinated axons, axonal degeneration, segmental demyelination, and increased fascicular area with

increased subperineurial and endoneurial connective tissue, thought to reflect a growth hormone-driven increase in circulating IGF-I and the presence of IGF-I receptors on Schwann cells and fibroblasts.

Liver Disease

Nerve biopsies in patients with chronic liver disease are reported to show evidence of segmental demyelination and remyelination. As in other metabolic syndromes, the interpretation is complicated by a variety of underlying, often associated, neuropathic diseases (e.g., alcoholism and diabetes). Primary biliary cirrhosis, a sensory neuropathy in which distinctive xanthomatous (lipid-rich collections within perineurial cells) are found in the perineurium, may have an immunological basis.

Alcoholism

Some alcoholics develop a chronic insidious progressive distal symmetrical sensorimotor neuropathy which may occasionally be painful or hyperesthetic and in which weakness of feet, hand, and thigh muscles may be prominent. Although classical studies in alcoholic patients reported segmental demyelination, more recent studies of animal models of alcoholism show that distal axonal degeneration is the predominant pathology. As in other metabolic neuropathies, the complex, deficient nutritional status of patients with severe alcoholism may involve a final pathway reflecting superimposed nutritional deficiency.

Summary

Metabolic neuropathies are common and complicate the clinical management of a variety of diseases, which is particularly true for diabetes, the most common of this group. The specter of an epidemic of diabetes and its complications has accelerated the pace of research which, at times, appears complex and contradictory. Nonetheless, investigators have begun to find common themes (e.g., oxidative stress) which may underlie a number of possible mechanisms (e.g., the polyol pathway, mitochondriopathy, IGF-I deficiency, and formation of AGEs). The many forms of diabetic neuropathy may involve a single mechanism or, more likely, involve several mechanisms which may require a cocktail of therapeutic agents in addition to an ongoing effort to better control blood glucose levels.

See also: Diabetic Neuropathy; Neuropathy: Chemically-Induced; Neuropathy: Peripheral; Oxidative Damage in Neurodegeneration and Injury.

Further Reading

Brownlee M (2001) Biochemistry and molecular cell biology of diabetic complications. *Nature* 414: 813–820.

Duby JJ, Campbell RK, Setter SM, White JR, and Rasmussen KA (2004) Diabetic neuropathy: An intensive review. *American Journal of Health System Pharmacy* 61: 160–173.

Freeman R (2005) Autonomic peripheral neuropathy. *Lancet* 365: 1259–1270.

Granberg V, Ejskjaer N, Peakman M, and Sundkvist G (2005) Autoantibodies to autonomic nerves associated with cardiac and peripheral autonomic neuropathy. *Diabetes Care* 28: 1959–1964.

Lleewelyn JG, Tomlinson DR, and Thomas PK (2005) Diabetic neuropathies. In: Dyck PJ and Thomas PK (eds.) *Peripheral Neuropathy*, vol. 2, ch. 85, pp. 1951–1991. Philadelphia: Saunders.

Malik RA, Tesfaye S, Newrick PG, et al. (2005) Sural nerve pathology in diabetic patients with minimal but progressive neuropathy. *Diabetologia* 48: 578–585.

Schmidt RE (1996) Neuropathology of human sympathetic autonomic ganglia. *Microscopy Research and Technique* 35: 107–121.

Siemionow M and Demir Y (2004) Diabetic neuropathy: Pathogenesis and treatment. A review. *Journal of Reconstructive Microsurgery* 20: 241–251.

Sima AA (2003) New insights into the metabolic and molecular basis for diabetic neuropathy. *Cellular and Molecular Life Sciences* 60: 2445–2464.

Sumner CJ, Sheth S, Griffin JW, Cornblath DR, and Polydefkis M (2003) The spectrum of neuropathy in diabetes and impaired glucose tolerance. *Neurology* 60: 108–111.

Tomlinson DR (ed.) (2002) *Neurobiology of Diabetic Neuropathy*, 463 p. Amsterdam: Academic Press.

Vinik AI, Holland MT, LeBeau JM, Liuzzi FJ, Stansberry KB, and Colen LB (1992) Diabetic neuropathies. *Diabetes Care* 15: 1926–1975.

Zochodne DW (1999) Diabetic neuropathies: Features and mechanisms. *Brain Pathology* 9: 369–391.

Relevant Website

http://www.neuro.wustl.edu – Neuromuscular Division, Washington University.

Neuropathy: Peripheral

C Chalk, Montreal General Hospital, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In its broadest sense, peripheral neuropathy means disease of the peripheral nervous system (PNS). However, PNS disorders are usually divided into disorders of individual peripheral nerves (mononeuropathies) and those that affect the PNS in more widespread fashion (polyneuropathies). In common use, the terms polyneuropathy and peripheral neuropathy are often used interchangeably. The distinction between mononeuropathy and polyneuropathy is made because of important etiological differences and functional implications. Mononeuropathies are most often a result of mechanical injury or compression of the nerve somewhere along its course, and the functional disability is limited to the nerve's motor or sensory distribution. Polyneuropathies, by contrast, have a long list of potential causes, ranging from neurotoxins to systemic metabolic disorders to genetic lesions, often making determination of their etiology difficult. From a functional viewpoint, polyneuropathies can range from trivial to life threatening. The main focus of this article is the polyneuropathies.

The PNS

The PNS comprises cranial nerves 3–12, the spinal roots, sensory and autonomic ganglia, and the somatic nerves. At a cellular level, the PNS is that part of the nervous system which is myelinated by Schwann cells, as opposed to the central nervous system (CNS), which is myelinated by oligodendroglial cells. In humans the transition between the CNS and PNS is quite abrupt, occurring a few millimeters distal to the points of exit of the cranial and spinal roots from the brain stem and spinal cord.

The PNS has three basic functions: (1) conveying motor commands to all voluntary striated muscles in the body; (2) carrying sensory information about the external world and the body to the brain and spinal cord (except visual information: the optic nerves, which convey information from the retina to the brain, are in fact outgrowths from the brain); and (3) regulating autonomic functions such as blood pressure or sweating. Most polyneuropathies impair all three types of PNS function, although disorders selectively affecting motor, sensory, or autonomic functions occur.

In most of the PNS, motor, sensory, and autonomic nerve fibers are intermingled to form mixed nerves, and as might be anticipated, most PNS disorders result in a mixture of motor, sensory, and autonomic dysfunction. However, at the proximal and distal extents of the PNS, functions are anatomically segregated: the ventral spinal roots contain efferent motor and autonomic fibers only, and the dorsal roots comprise only afferent sensory fibers. Likewise, neuronal cell bodies in the dorsal root ganglia, which lie in the intervertebral foramina just proximal to the confluence of the dorsal and ventral spinal roots, are exclusively sensory, and autonomic ganglia contain autonomic neurons only. At the distal extent of their distribution, the motor and sensory axons of mixed nerves are sorted into purely motor or sensory terminal branches. An appreciation of these anatomical arrangements is important in diagnosis as localizing the likely site of the disorder helps narrow the possibilities that must be considered. For example, a disorder affecting sensation and sparing motor and autonomic function can be localized to dorsal roots, dorsal root ganglia, or sensory terminal branches and is unlikely to be arising in mixed somatic or cranial nerves, ventral roots, or autonomic ganglia. The process of localization as the first step in diagnosis, which is axiomatic in brain and spinal cord disorders, is equally important in the PNS, yet it is often neglected.

Throughout the PNS, the same basic histological plan is found. Myelinated and unmyelinated axons, Schwann cells, and interspersed capillary-like blood vessels (collectively referred to as endoneurium) are bundled into fascicles by perineurium, composed of five to ten lamellae of cytoplasm of perineurial cells, a type of modified fibroblast. Most nerves consist of several such fascicles, bound together by a loose connective tissue, the epineurium. Within the epineurium, between the fascicles, are found arterioles and venules of various sizes, structurally similar to those found in other tissues. The epineurial vessels are the major blood supply of the nerve, with the capillary-like endoneurial vessels constituting a parallel system. This dual blood supply ensures that peripheral nerve is relatively well protected from ischemic insult. Cells in both the perineurium and the endoneurial microvessels are interconnected by tight junctions, producing a blood–nerve barrier (or blood–nerve interface). This structural specialization ensures that peripheral nerve axons and Schwann cells occupy an environment whose chemical and cellular constituents are carefully regulated, analogous to the situation for neurons and glial cells in the CNS. Certain parts of the PNS lack this regulated

blood–nerve interface: the sensory and autonomic ganglia, the portion of spinal roots and cranial nerves traversing the subarachnoid space, and the distal terminations of motor axons at neuromuscular junctions.

The relative simple histology and function of the PNS limits its range of expressions of pathology. In fact, PNS disorders can be conveniently divided into two main types: diseases primarily affecting axons and diseases in which derangement of myelin is the primary problem, a distinction that is useful for diagnosis as well as for a rational approach to therapy. Nerve conduction studies and needle electromyography, which are easily performed in patients and which are widely available, are used to help characterize a process as demyelinating or axonal.

Given myelin's physiological role in greatly enhancing the speed of action potential propagation along axons, demyelinating polyneuropathies can be identified by marked slowing of nerve conduction. In genetic disorders of PNS myelination, the nerve conduction slowing affects all peripheral nerves uniformly, whereas acquired causes of PNS demyelination typically affect myelin in a patchy, nonuniform way. A consequence of the latter point is that there may be marked variability in conduction slowing between adjacent nerves, dispersion of compound nerve action potentials, or focal conduction block. In axonal neuropathies, by contrast, nerve conduction velocities are preserved or only minimally slowed, and the loss of functioning motor or sensory axons is reflected in decreased amplitudes of motor or sensory compound nerve action potentials. This dichotomous view of polyneuropathies as either demyelinating or axonal is useful, but it is an oversimplification. Myelin and axon integrity are interdependent, so all types of polyneuropathies ultimately have an impact on both axon and myelin.

The clinical manifestations of a polyneuropathy are determined by the classes of nerve fibers involved. Motor findings are those of the lower motor neuron type, that is, weakness accompanied by decreased muscle tone and decreased or absent reflexes. Axonal loss is also marked by muscle atrophy and, sometimes, fasciculation. The main symptoms of sensory fiber involvement are sensory loss and paraesthesia (abnormal sensations such as prickling or 'pins and needles'), and the latter are often painful. Examination may show relatively selective loss of only some sensory modalities in diseases that affect only certain classes of sensory fibers (e.g., in most acquired demyelinating neuropathies, the main sensory finding is of vibration and joint position sensation, reflecting a disease process mainly involving large-diameter myelinated axons). Most often the motor and sensory impairments are first manifest and remain most pronounced in the distal parts of the legs and arms, a pattern referred to as a glove-and-stocking distribution. Autonomic dysfunction is prominent in some types of polyneuropathy, and usually vasomotor symptoms such as orthostatic hypotension are the most troublesome. However, a host of other symptoms reflecting autonomic failure may be seen, including absent or excessive sweating, impaired lacrimation and salivation, bladder dysfunction, abnormal intestinal motility, and erectile impotence and other impairments of sexual function.

Determining that a patient has a polyneuropathy usually presents little difficulty, but discovering the cause of the neuropathy can be challenging. The list of potential causes is long (at least 100 are known), and the clinical manifestations of many of these are indistinguishable. The usual clinical approach is to sift through the symptoms, neurological examination findings, and nerve conduction data to see whether the patient may have a syndrome with a short list of possible causes. For example, sensory symptoms and impairment of all sensory modalities without motor or autonomic impairment might lead one to diagnose a dorsal root polyganglionopathy, of which there are only a few causes. In a patient with motor and sensory abnormalities in the distal legs and arms, finding slowed nerve conduction velocities greatly limits the potential diagnoses.

Often, however, a distinctive syndrome is not present, and a broad range of possibilities must be considered. One approach to classifying the possibilities is to separate inherited from acquired diseases and to subdivide the acquired causes into four large categories: metabolic (e.g., diabetes mellitus), infectious (e.g., leprosy), toxic/deficiency states (e.g., vinca alkaloids and other antineoplastic drugs), and immune-inflammatory (e.g., vasculitic neuropathy). Worldwide, the three most common types of polyneuropathy are probably leprotic, diabetic, and inherited. The population from which patients are drawn influences the likelihood of certain diagnoses. For example, leprosy is common in developing countries but rare in North America and Europe. In a hospital-based practice, chemotherapeutic drugs are a common cause of neuropathy.

Establishing a clear etiology for a patient's polyneuropathy may lead to specific therapy, but equally important, establishing the cause allows the physician to advise the patient about the long-term prognosis (some types of polyneuropathy can lead to major loss of function, but more often the disability is mild or less and does not worsen substantially over the years). Some therapies can be useful in almost all types of

polyneuropathy, regardless of cause. Pain is common in many types of polyneuropathy, and drugs for neuropathic pain are a mainstay of therapy in many patients. Medications used include amitriptyline (a tricyclic antidepressant) and gabapentin (an anticonvulsant that reduces synaptic release of neurotransmitters by inhibiting voltage-gated calcium channels). In patients with distal leg weakness, the ankle–foot orthosis (an L-shaped plastic brace that fits into a shoe and is strapped to the leg) holds the foot in neutral position and stabilizes the ankle, improving walking and stability, irrespective of the cause.

Pathobiological Mechanisms in Peripheral Neuropathy

The wide range of causes of peripheral neuropathy implies that nerve injury may be produced by a considerable variety of mechanisms at the cellular level. Although little is known about the pathophysiology of many types of peripheral neuropathy, in some there is considerably detailed understanding of the cellular events involved. An exhaustive account of pathophysiological mechanisms in peripheral neuropathy is beyond the scope of this article. However, a sense of current knowledge can be gained by considering four important but disparate varieties of neuropathy: Charcot–Marie–Tooth disease, diabetic neuropathy, Guillain–Barré syndrome, and neuropathy due to exogenous toxins.

Charcot–Marie–Tooth Disease

Charcot–Marie–Tooth disease (CMT) was first described in the nineteenth century as a familial disorder of slowly progressive distal weakness, muscle atrophy, and sensory loss. The term 'hereditary motor and sensory neuropathy' is used interchangeably with CMT. At present, 24 distinct genes have been identified which can produce the basic CMT phenotype. Nerve conduction studies show that CMT can be divided into two major groups, those with slow nerve conduction (type 1) and those in which nerve conduction velocity is normal (type 2). Autosomal dominant inheritance is most frequent, although recessive and X-linked forms also occur.

The identification of causative genes has been most rapid in type 1 CMT, and eight distinct loci producing the CMT type 1 phenotype are now known. The most common is characterized by a 1.4 Mb duplication in the long arm of chromosome 17 containing the gene encoding a 22 kDa peripheral myelin protein, PMP22. Establishing that the human PMP22 gene duplication was indeed the cause of the CMT type 1 phenotype was aided by the demonstration of a homologous gene duplication in the trembler mouse, which was long known to have slowed nerve conduction and nerve pathology closely resembling the human CMT type 1. Transgenic mouse and rat models of CMT type 1 are now available; they exhibit the 1.5-fold increase in PMP22 expression and histological abnormalities characteristic of CMT type 1. Studies in these models and in Schwann cell cultures suggest that PMP22 overexpression leads to endoplasmic reticulum overload, a blockade of normal intracellular protein degradation, and eventually cell death. The resulting impaired nerve myelination is reflected in the slowed nerve conduction. However, functional disability is not correlated with conduction slowing but with the progressive axonal loss which also occurs in all types of CMT. Presumably Schwann cell–axon interactions are necessary for axonal survival, but a detailed understanding of these interactions is lacking. Another puzzle in CMT type 1 is the striking variability in the phenotype severity, both between and within kinships, despite a uniform causative genetic lesion. The nature of environmental or other genetic factors that influence phenotype severity is currently unknown, as are the cellular consequences of other genetic loci for CMT type 1, although disruption in intracellular myelin protein trafficking seems likely.

Progress in the genetics of CMT type 2 (the 'axonal' variety, with preserved nerve conduction velocity) has been slower, although a few genes and chromosomal linkages have been identified. One CMT type 2 gene, mitofusin 2, is a coregulator of mitochondrial fusion and fission. X-linked CMT has a phenotype similar to CMT type 1 and 2 but has nerve conduction slowing intermediate between type 1 and type 2. In most kinships, point mutations in a gene encoding a gap junction protein, connexin 22, have been found. It seems logical that the generation and maintenance of compact myelin around each axon may require the participation of intra- and intercellular gap junctions, but further studies to elucidate mechanisms are awaited.

Diabetic Polyneuropathy

Peripheral neuropathy is a common complication of diabetes mellitus. Estimates of the prevalence of neuropathy in diabetics are in the range of 10–50%, depending on factors such as how neuropathy is defined, methods of detection, and the population studied. The symptoms and signs of diabetic polyneuropathy mainly involve the distal parts of the limbs and are approximately symmetrical. Sensory loss and paresthesia (often painful) are predominant, while muscle weakness is mild in most patients. In addition to disability related to pain and the sensory

and motor deficits, polyneuropathy is an important independent predictor of foot ulceration in diabetics, and it is associated with a twofold to 25-fold increase in the risk of amputation.

The pathogenesis of diabetic polyneuropathy has been the subject of much study and discussion, and there is not universal agreement about its cause. Electrophysiological and nerve biopsy data suggest that the major pathological process is axonal degeneration, though there is evidence of secondary demyelination as well. Potential mechanisms of pathogenesis include an ischemic process secondary to microangiopathy, analogous to diabetic retinopathy and nephropathy, glycosylation of structural proteins consequent to chronic hyperglycemia, or oxidative stress injury stemming from altered glucose metabolism. These mechanisms are not mutually exclusive and may coexist and interact in many or most patients with diabetic polyneuropathy.

In animal models of diabetic neuropathy, such as diabetes produced in rats by the pancreatic islet cell toxin streptozocin, reduced endoneurial blood flow and endoneurial hypoxia can be demonstrated. These are assumed to be consequent to basal lamina thickening and other structural alterations in endoneurial and epineurial vessels that are evident in histological studies of diabetic nerve. Chronic hyperglycemia is hypothesized to have both extracellular and intracellular effects on peripheral nerve. Extracellular effects include various types of glycation reactions, including direct glycosylation of structural proteins, as well as chemical rearrangements resulting in the production of advanced glycosylation end products. Intracellular effects result from increased flux of glucose through the polyol pathway, resulting in various potentially deleterious consequences: (1) a decrease in cellular nicotinamide adenine dinucleotide phosphate levels, resulting in decreased concentrations of glutathione (a free radical scavenger) and nitric oxide (a vasodilator); (2) increased cellular sorbitol levels, leading to decreased levels of myoinositol (necessary for Na-K adenosine triphosphatase function); and (3) increased fructose levels, which enhance formation of advanced glycosylation end products (fructose is tenfold more active than glucose in glycosylation reactions).

Attempts to prevent or reverse human diabetic neuropathy using treatments based on the wealth of experimental data have been disappointing to date. There is some evidence that strict glucose control (best achieved by pancreatic transplantation) is beneficial. Suppressing glucose metabolism via the polyol pathway by inhibiting the pathway's rate-limiting step appears promising in animal models, but in human trials the aldose reductase inhibitors have failed to show a clear benefit. Likewise, free radical scavengers like α-lipoic acid have shown promise experimentally, but clinical trials have been negative. Part of the inability to translate experimental success into therapies for human diabetic neuropathy may be related to issues of trial design such as reliable endpoints and duration of trials.

Guillain–Barré Syndrome

Since the widespread use of poliomyelitis vaccine, the most common cause of acute flaccid paralysis in the West is Guillain–Barré syndrome (GBS), or acute inflammatory demyelinating polyradiculoneuropathy, with an incidence of about 1.5 per 100 000 people per year. Prior to the development of modern intensive care units and mechanical ventilation, about one-third of patients with GBS would die of respiratory failure. Death is now uncommon in this condition, but GBS continues to produce substantial immediate and long-term neurological disability.

The course of GBS is characterized by rapid onset and worsening, with maximum weakness usually attained within 2 weeks of onset. About 25% of patients will require artificial ventilation, and a greater fraction will be bed-bound because of limb weakness. After a period of stability lasting several weeks or more, there is slow recovery of function over several months or more, due to nerve remyelination and regeneration. The recovery in most patients is fairly complete, although about 20% are left with permanent major disability.

The onset of GBS often closely follows minor upper respiratory or gastrointestinal infections. As there is no evidence of direct infection of the PNS in GBS, the concept of GBS as a parainfectious syndrome arose, that is, consequent to immune system activation triggered by an infectious agent. It was subsequently appreciated that infections with certain agents were particularly likely to precede the onset of GBS, in particular the gram-negative enteric bacterium *Campylobacter jejuni* and the Epstein–Barr and cytomegaloviruses. These observations lead to the idea that GBS might be an example of molecular mimicry, in which the immune system recognizes and targets epitopes common to the triggering microorganism and peripheral nerve. This concept is now widely accepted, although the specific epitopes responsible in most patients remain unknown and understanding of the immune effector mechanisms at work is incomplete.

The fundamental pathological process in classical GBS is an attack on myelin by macrophages. Biopsy

material from GBS patients and animal models of the disease shows macrophages engaged in active phagocytosis of myelin, with processes insinuated between myelin lamellae and lysis of myelin fragments in phagocytic vesicles. Macrophage activation appears to depend on both antigen presentation by T cells and the elaboration of cytokines and B cell production of antibodies directed against myelin epitopes. In recent years a variant of classical GBS has been described, acute motor axonal neuropathy, in which the macrophage attack targets the axolemma at nodes of Ranvier rather than the myelin sheath. This axonal form of GBS is particularly common in China, Japan, and Latin America; in northern China, axonal GBS occurs in summertime epidemics and is closely associated with antecedent *C. jejuni* infection.

The main animal model of GBS, experimental allergic neuritis, is produced by immunizing rats with peripheral myelin. Although specific moieties of several myelin proteins are able to induce experimental allergic neuritis, none of these have been convincingly shown to be target antigens in human GBS. Another unresolved issue is the role that antibodies against myelin gangliosides may play in the pathogenesis of GBS. There are many reports of different ganglioside antibodies in patients with classical GBS and its variants, but there is little agreement at present about the pathogenic role of these antibodies. The most convincing association has been found in a cranial nerve variant of GBS, Miller Fisher syndrome, in which 90% of patients have serum antibodies against the ganglioside GQ1b, which is particularly abundant in the myelin of ocular motor nerves affected in this syndrome. GQ1b antibody titers follow the clinical course of Miller Fisher syndrome, and they are not found in other types of inflammatory neuropathy. The case for the role of molecular mimicry in GBS has also been most closely worked out in Miller Fisher syndrome, as the GQ1b antibodies also exhibit a high affinity for certain moieties contained in the lipopolysaccharide capsule of the *C. jejuni* serotypes frequently associated with Miller Fisher syndrome. Similar antigenic homology between myelin components and other triggering microorganisms in other forms of GBS is assumed to exist but remains unproven.

Other than respiratory and other supportive care, treatments in GBS aim to abrogate the immunological cascade that leads to peripheral nerve injury. Current therapies that have been demonstrated to improve the rate of recovery in GBS are plasma exchange and intravenous immunoglobulin infusions, both of which exert widespread but rather nonspecific modulatory effects on the immune system. Corticosteroids have been shown to offer no benefit in human GBS. It is hoped that better understanding of the antigenic target(s) and contributions of the several immune effectors in GBS may allow development of specific, targeted immune therapy to improve outcomes in this disease.

Toxic Polyneuropathy

Peripheral nerve is susceptible to many chemical toxins. Some of these have been identified by careful study of clusters of polyneuropathy in communities or factory workers. Notorious examples include arsenical neuropathy from contaminated beer in nineteenth-century England, tri-orthocresyl phosphate neuropathy ('jake leg paralysis') from contaminated Jamaica ginger extract during Prohibition in the United States, and neuropathy in factory workers handling acrylamide monomer. Once such environmental neurotoxins are identified, further outbreaks of neuropathy can usually be prevented.

A more difficult problem is encountered with drugs whose adverse effects include peripheral neuropathy. This has become particularly important in oncology and in the treatment of HIV infection, in which many effective drugs predictably produce neuropathy, usually in a roughly dose-related fashion. Drug-induced neutropenia, formerly the dose-limiting factor for most antineoplastics, can now be treated with granulocyte colony stimulating factor. Consequently, neuropathy is now the dose-limiting side effect for many of these drugs. Efforts to understand how these drugs injure peripheral nerve have an intrinsic biological interest and are of great practical clinical importance as well. Although toxic neuropathy appears to offer a relatively uncomplicated system for study of the cellular mechanisms of nerve injury, there has been surprising difficulty developing good animal models for many of these neurotoxins.

An important strategy in antineoplastic drug design has been to develop drugs that interfere with microtubule function, which will inhibit cell division by disrupting mitotic spindles. Examples include the vinca alkaloids (vincristine, vinblastine, vinorelbine, and vindesine) and the taxanes (paclitaxel, docitaxel). Most neurons in the PNS must maintain long axons (some one meter or more) and are critically dependent on microtubule-based axonal transport systems to do so. As a consequence, it was anticipated that drugs impairing microtubule function would affect the PNS.

The vinca alkaloids bind with high affinity to tubulin, thereby disrupting microtubule assembly. In animal models, disorganized axonal transport microtubules and accumulation of neurofilaments in neuronal soma can be demonstrated, implying that the

vinca alkaloids do indeed impair anterograde axonal transport. Subsequent axonal degeneration appears to be calcium dependent and mediated by calpain. The taxanes bind to the beta subunit of tubulin and result in abnormally stable microtubules, in which disassembly is inhibited. This is assumed to impair axonal transport, resulting in neuropathy. With both vinca alkaloids and taxanes, preventive treatments that block interactions with tubulin might prevent neuropathy but would also likely limit their antineoplastic properties. Therefore, strategies to minimize the neurotoxicity of these drugs have aimed at enhancing neuronal survival or axonal regeneration, using agents like acetyl-L-carnitine, glutamate, or growth factors such as insulin-like growth factor. Although promising in animal models, these approaches have had limited success in humans.

See also: Autoimmune Autonomic Neuropathy; Demyelinating Diseases; Demyelination and Demyelinating Antibodies; Diabetic Neuropathy; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Neuropathy: Chemically-Induced; Neuropathy: Metabolically-Induced; Peripheral Nerve Regeneration: An Overview.

Further Reading

Argyriou AA, Chroni E, Koutras A, et al. (2005) Vitamin E for prophylaxis against chemotherapy-induced neuropathy: A randomized controlled trial. *Neurology* 64: 26–31.

Dyck PJ and Thomas PK (2005) *Peripheral Neuropathy,* 4th edn. Philadelphia: Elsevier Saunders.

Hughes RAC and Cornblath D (2005) Guillain–Barré syndrome. *Lancet* 366: 1653–1666.

Leinninger GM, Edwards JL, Lipshaw MJ, and Feldman EL (2006) Mechanisms of disease: Mitochondria as new therapeutic targets in diabetic neuropathy. *Nature Clinical Practice Neurology* 2: 620–628.

Meyer zu Hörste G and Nave K-A (2006) Animal models of inherited neuropathies. *Current Opinion in Neurology* 19: 464–473.

Wang MS, Wu Y, Culver DG, and Glass JD (2000) Pathogenesis of axonal degeneration: Parallels between Wallerian degeneration and vincristine neuropathy. *Journal of Neuropathology & Experimental Neurology* 59: 599–606.

Neuropeptide FF and Receptors

P Panula, University of Helsinki, Helsinki, Finland

© 2009 Elsevier Ltd. All rights reserved.

Five Neuropeptide Genes and New Receptor Mechanisms

Discovery of a New Neuropeptide Gene Family

It took a surprisingly long time after the discovery of neuropeptide FF (NPFF) and neuropeptide AF (NPAF) by Hsiu-Ying Yang and co-workers in 1985 to clone the corresponding gene. This was apparently due to the low abundance of the messenger RNA (mRNA) in the brain and the amino-acid composition (many codons) of the common C-terminus of these peptides. The NPFF gene is located in chromosome 12 in humans (Table 1). The well-conserved human, rat, mouse, and bovine genes encode for both NPFF and NPAF (or the corresponding peptide). The gene comprises two short introns and three exons. The second gene, named prolactin-releasing peptide (PrRP), was identified and cloned as an endogenous ligand for a previously known orphan G-protein-coupled receptor (GPCR), GPR10, in 1998. This gene gives rise to one active peptide which can vary in length. In different species, highly similar peptides referred to as $PrRP_{20}$ and $PrRP_{31}$ are produced.

The presence of additional genes was recognized after cloning of the NPFF gene because there was a distinct hypothalamic group of neurons that expressed no NPFF mRNA despite strong NPFF immunoreactivity, suggesting the existence of a closely related immunoreactive peptide. This Arg-Phe-amide-related peptide (RFRP) gene was identified from databases and cloned. It encodes for three distinct active peptides, two of which are Arg-Phe-amide (RFamide) peptides.

The Kiss-1, or metastin, gene was initially identified as a protein that suppresses tumor metastasis in mice. Several groups then showed that the peptide originating from the Kiss-1 precursor activates a G-protein-coupled receptor which is present in many tissues in the periphery and brain.

The P518/26RFa/QRFP peptide gene was identified as an open reading frame for a 124-ti 136-residue protein from gene databases after testing the affinity of predicted peptides to a previously known orphan receptor or by searching for new RFamide peptide genes. The human precursor encodes for two potential RFamide peptides, whereas in other mammals, one of the predicted peptides is an RLamide (Figure 1). The nomenclature of the predicted peptides derived from the two potential precursor regions is unclear at the moment as no thorough analysis has been done for all predicted peptides, and the activity of these peptides has not been studied in detail. They have been referred to as P51, P52, QRFP, or 26RFa. Currently, the nomenclature of the mammalian RFamide peptides is complicated, and no logical, understandable, and generally accepted one exists. The names and chromosomal localization of the genes, peptide names, and most important receptors are shown in Table 1. The peptide sequences of typical active peptides are shown in Figure 1.

Neuropeptide FF, the Prototype Mammalian RFamide Peptide

After the discovery of the clam cardioexcitatory peptide in 1977, it was possible to explore the possible existence of related mammalian peptides using antisera against the molluscan tetrapeptide. Reports on the presence of FMRFamide immunoreactivity in mammalian brain and pituitary were followed by the identification of the first mammalian (bovine) neuropeptides FLFQPGRFamide and AGEGLSSPFWSLAAPQRFamide by Hsiu-Ying Yang and co-workers in 1986. The new peptide was called FLFQPQRFamide by those who discovered it. Other names, such as morphine-modulating peptide, were also used. Investigators were reluctant to name it because the most important functions of the peptide were not obvious, and the name could have become a misnomer. The NPFF gene was difficult to clone, which may have nourished the idea that even the immunoreactivity was due to cross-reactivity to a more abundant neuropeptide Y (NPY). This was indeed supported by the fact that many FMRFamide antisera cross-reacted with NPY, which is much more abundant in the central nervous system (CNS) than NPFF. Despite the lack of evidence of the gene and mRNA, many studies were performed with synthetic peptides and antisera as reagents. The distribution of NPFF immunoreactivity in the spinal cord, medulla, hypothalamus, and posterior pituitary suggested that NPFF may have different functions in the CNS. Many studies have since shown distinct effects on the pain system, including acute, inflammatory, and neuropathic pain. NPFF was reported to induce a vigorous morphine abstinence syndrome in rats, and the NPFF antiserum given intracerebroventricularly (icv) attenuated naloxone-induced abstinence symptoms. Acute but not chronic morphine also upregulates NPFF mRNA expression in the brain stem, but not in the spinal cord. NPFF alone does not affect catecholamine turnover, but it enhances rather than

Table 1 Mammalian RFamide peptide genes, active peptides, human chromosomal localization of the gene, and corresponding G-protein-coupled receptors

Gene	*Peptide*	*Human chromosome*	*Receptor*
NPFF; *farp1*	NPFF NPAF	12	NPFF2/HLWAR77/NPGPR/GPR74 NPFF1/OT7T022/GPR147
PrRP; *farp2*	PrRP31 PrRP20	2	GPR10/UHR-1/hGR3
RFRP; *farp3*	RFRP-1 (RFRP-2)[a] RFRP-3	7	NPFF2/HLWAR77/NPGPR/GPR74 NPFF1/OT7T022/GPR147
Kiss-1; *farp4*	Metastin/ Kiss-1	1	GPR54/OT7T175/AXOR12/Kiss-1R
P518; *farp5*	P518/QRFP/ 26RFa	9	GPR103/AQ27/SP9155

[a]RFRP-2 is in parenthesis, because the peptide is not an RFamide peptide.

```
NPFF
hNPAF                                        PRAGEGLNSQFWSLAAPQRF-amide
mNPAF                                               ARQFWSLAAPQRF-amide
hNPSF                                               GRSQAFLFQPQRF-amide
mNPFF                                                  QAFLFQPQRF-amide
mNPSF                                               RRSPAFLFQPQRF-amide

RFRPs
hRFRP1                      SLNFEELKDWGPKNVIKMSTPAVNKMPHSFANLPLRF-amide
mRFRP1                      SVSFQELKDWGAKNVIKMSPAPANKVPHSAANLPLRF-amide
hRFRP3                            SAGATANLPLRSGRNMEVSLVRRVPNLPQRF-amide
mRFRP3                                         VNMEAGTRSHFPSLPQRF-amide

PrRP
hPrRP20                                    IRTPDINPAWYASRGIRPVGRF-amide
hPrRP31                           SRTHRHSMEIRTPDINPAWYASRGIRPVGRF-amide

KiSS-1/metastin
hKiSS1     GTSLSPPPESSGSPQQPGLSAPHSRQIPAPQGAVLVQREKDLPNYNWNSFGLRF-amide
mKiSS1       SSPCPPVEGPAGRQRPLCASRSRLIPAPRGAVLVQREKDLSTYNWNSFGLRY-amide

P518/26RFa/QRFP
mP51                                     ASGPLGNLAEELNGYSRKKGGFSFRF-amide
hP51                                     TSGPLGNLAEELNGYSRKKGGFSFRF-amide
hP52                                   ASQPQALLVIARGLQTSGREHAGCRFRF-amide
hP52                                     NPQPQALLVVAREQQASHREHTGFRL-amide

NPY
mNPY                          YPSKPDNPGEDAPAEDMARYYSALRHYINLITRQRY-amide
hNPY                          YPSKPDNPGEDAPAEDMARYYSALRHYINLITRQRY-amide
```

Figure 1 C-terminal peptide structures of the typical mammalian RFamide peptides shown in comparison with neuropeptide Y (NPY). Note that the human Kiss-1 peptide has the characteristic RFamide structure whereas the mouse peptide shares the RYamide C-terminus with NPY.

attenuates changes induced by acute morphine on brain monoamines. Inactivation of NPFF by immunoglobulin G from the antiserum affects spatial memory in mice.

Recently, the first dipeptide analog of RFamide was reported as a selective and potent antagonist of NPFF at NPFF receptors. The mammalian RFamide peptides are difficult to study functionally because they share functions and receptor mechanisms. Apparently at least some of them have evolved as results of gene duplication, and some peptides may have lost their functions during evolution in some species. Studies with antisera directed toward the related C-termini of these peptides may also be difficult to interpret because several endogenous peptides derived from different genes may be detected. Application of one peptide at a particular brain site may also mimic the action of another endogenous peptide. Studies with gene-modified animals may also be hampered by activation of compensatory mechanisms. So far, data is available from some receptor knockout animals, and these studies suggest important interactions between the different RFamide peptides. As research on these peptides is still limited, new functions for them are likely to emerge.

Prolactin-Releasing Peptide

Although PrRP was first reported to release prolactin from a pituitary cell line and isolated pituitary cells, several subsequent studies failed to demonstrate a direct prolactin-releasing activity in intact animals. Interstingely, both NPFF and RFRP release prolactin both *in vivo* and *in vitro*, which demonstrates that RFamide peptides do possess this activity. The PrRP gene is expressed in a neuronal population of hypothalamic cells in the dorsomedial nucleus and in two different sites of medulla: the commissural nucleus tractus solitarius (NTS) and the group A1/A2 noradrenergic neurons. The first location would fit with regulation of endocrine functions. Indeed, the PrRPs are implicated in corticotropin-releasing hormone-mediated adrenocorticotropic hormone-release and subsequent regulation of plasma corticosterone levels. Current evidence suggests that PrRP may be involved in stress-mediated regulation of oxytocin release from supraoptic neurons. Several different types of experiments have also supported the involvement of PrRP in regulation of energy metabolism: icv administration of PrRP reduces food intake and increases uncoupling protein-1 (UCP-1) mRNA expression in brown adipose tissue. GPR10 knockout mice are obese and display elevated leptin and insulin levels, and their glucose tolerance is decreased.

Arg-Phe-Amide-Related Peptide

The RFRP gene gives rise to three different active peptides, two of which are RFamides. Their C-terminal structure resembles closely that of NPFF precursor derived peptides, and the RFRP-3 peptide shares the C-terminal tetrapeptide sequence QPRFamide with NPFF. The RFRP gene expression in the brain is limited to a central hypothalamic area, an area that covers parts of the dorsomedial nucleus and the area between the ventromedial and dorsomedial nucleus, and the periventricular nucleus. RFRP peptides have clear endocrinological and sensory effects, whereas their other functions have not been studied in detail.

P518/26RFa/QRFP

The brain distribution of this newly identified peptide gene in the ventromedial nucleus and lateral hypothalamic area suggest roles in regulation of metabolism. In human CNS, the gene is also expressed in the ventromedial nucleus and in the paraventricular nucleus. It has also been reported in human spinal cord, suggesting possible involvement in sensory functions. In mice deprived of food, the peptide increases short-term feeding. Other CNS functions of this peptide are yet to be studied. Activity studies have been done mainly using the C-terminally located active peptide 26RFa/QRFP/P518, whereas the receptors and functions of the second active peptide (which is RFamide only in humans) are less well known. The gene is also expressed in several other tissues, including the prostate, testes, urinary bladder, gut, thyroid, and parathyroid glands, suggesting that the peptides may have several different functions.

Kiss-1

In the mouse brain, Kiss-1, or metastin, mRNA is expressed in the periventricular nucleus, the arcuate and anterodorsal part of the preoptic nucleus, and the medial amygdala and nucleus of the stria terminalis. The neuroendocrine role of Kiss-1 through GPR54 is now well known, but other central functions remain largely unexplored.

GPCR Mechanisms

All mammalian RFamide peptides activate one or several GPCRs (Table 1). The first identified ligand/peptide system was the PrRP/GPR10 system. GPR10 was known as an orphan receptor. The GPR10 receptor couples to at least G_q protein in cell assays. Binding sites for NPFF had been mapped throughout the CNS long before the NPFF receptor identities were revealed. High levels of binding were seen in, for example, dorsal horn of the spinal cord and dorsal medullary regions, suggesting important functions in sensory and autonomic regulation. The first NPFF receptor gene identified was the GPCR referred to as HLWAR77. This receptor was also identified as NPFF2 receptor, together with another GPCR nominated NPFF1 receptor, or OT7T7022, by independent research groups. The NPFF receptors are coupled to at least the $G_{i/o}$ pathway, although coupling to other G-proteins is also possible. The Kiss-1 peptide was identified as a natural ligand for the orphan receptor GPR54, also called OT7T175 or AXOR12. Activation or this receptor activates phospholipase C through G_q protein. The P518/25RFa/QRFP was soon shown to be the natural ligand for the orphan receptor GPR103, also called AQ27 or SP9155. P518/25RFa/QRFP induces a rapid elevation of intracellular Ca^{2+} and inhibits cyclic adenosine monophosphate formation, suggesting coupling to $G_{i/o}$ and G_q proteins in cells that express the receptor.

Thus, all known RFamide peptides are now known to act as natural ligands for at least one GPCR. Although it has not been shown that any of the mammalian RFamide peptides displace opioid ligands from opioid receptors, evidence from cells that express delta-opioid receptors but no NPFF receptors suggests that NPFF can allosterically modify the δ-opioid receptor responses as measured using different signaling assays. This remains a possible mechanism that may contribute

to the modulatory effect of NPFF on opioid-induced analgesia. In addition, a subfamily of GPCRs referred to as maslike receptors present in, for instance, sensory ganglia MrgA1 and MrgC11 have affinity for RFamide peptides. No clearly specific ion channel receptors for mammalian RFamide peptides have been identified so far. NPFF potentiates the H^+-induced currents through acid-sensing ion channels in sensory ganglion neurons at high concentrations.

Multiple Functions of Mammalian RFamide Peptides

Sensory and Autonomic Functions

NPFF was first identified as a peptide that modulates morphine analgesia. Injection of NPFF into the brain ventricles attenuates the analgesic response to morphine in rats. Recent studies suggest that the RFRP peptide is even more potent in antagonizing the opioid effects. A distinctly different effect can be obtained after intrathecal injections of NPFF to the spinal space. A prolonged and enhanced response to morphine has been reported in rats when NPFF has been administered simultaneously. In models of inflammatory pain, like carrageenan-induced paw inflammation, the number of NPFF-producing cells in the superficial dorsal horn of the rat increases significantly, and a concomitant increase in NPFF2R also appears. When applied topically to exposed cells, exogenous NPFF enhances the descending inhibitory effect to spinal wide dynamic range neurons through a direct spinal action, suggesting that the increased expression of NPFF in inflammatory conditions may have functional significance. Intrathecally administered stable NPFF analog (1DMe)NPYF is effective against both thermal hyperalgesia and mechanical allodynia.

Contrary to the inflammatory changes, neuropathy induced by ligation of spinal nerves does not induce significant increase in spinal NPFF expression. However, intrathecal NPFF reduces cold allodynia. Few cells in dorsal root ganglia have also been reported to express NPFF mRNA, especially after nerve ligation.

NPFF-containing nerve fibers are abundant in many brain stem nuclei involved in pain regulation and autonomic functions (Figure 2). Functionally, NPFF attenuates tactile allodynia induced by nerve ligation by a mechanism that is in part sensitive to naloxone. In both normal and neuropathic rats, NPFF reverses the antinociception induced by morphine in the periaqueductal grey. The effect of NPFF in this brain area is thus paradoxical: it both reverses the effect of morphine and displays a clear antiallodynic effect.

In a model of visceral pain (rat acute colonic inflammation), the NPFF and NPFF2R mRNA are not affected in the spinal cord, but the brain stem expression of receptor mRNA is increased, suggesting involvement of this system in visceral pain.

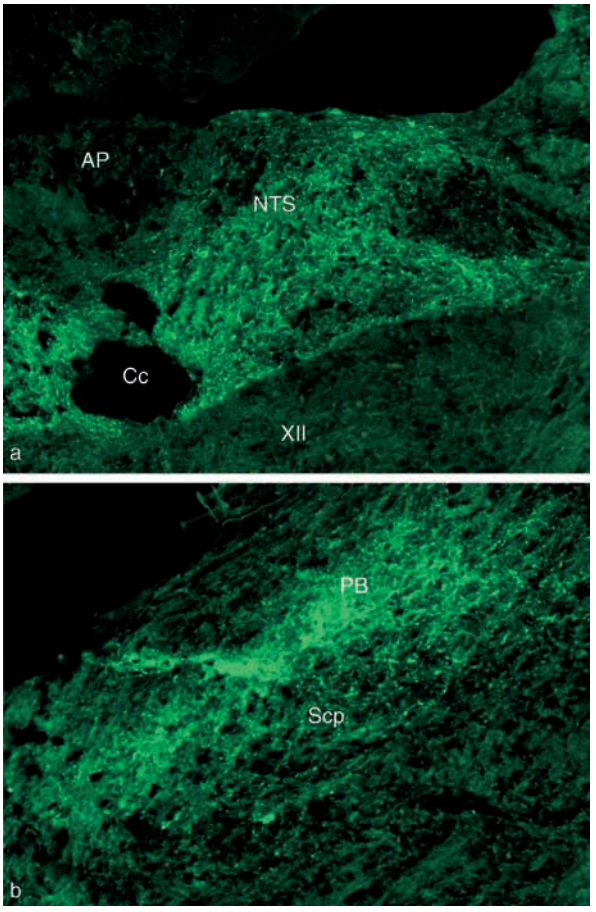

Figure 2 Immunocytochemistry of neuropeptide FF (NPFF) in the mouse brain. (a) A dense network of nerve fibers and small immunoreactive neurons in the nucleus tractus solitarius (NTS) are seen in the vicinity of the nonreactive hypoglossal nucleus (XII) and area postrema (AP). (b) Dense fibers immunoreactive for NPFF but no immunoreactive neurons in the lateral parabrachial nucleus (PB). The antiserum used here shows no staining in NPFF knockout mouse, so that the immunoreactivity shown here is due to NPFF gene products rather than the closely related peptides derived from the other RFamide peptide genes. Cc, central canal; Scp, superior cerebeller peduncle.

Considering the closely related structures of mammalian RFamide peptides, it is not surprising that several of these peptides show significant effects in various models of pain. Although $PrRP_{20}$ and its C-terminal octapeptide fail to produce antinociception on the spinal level, $PrRP_{20}$ injected into the brain stem produces strong mechanical antinociception which is not reversed by naloxone. RFRP-1 is effective particularly in tactile allodynia and also in thermal nociception when administered intrathecally, whereas mechanical hyperalgesia is not affected to the same degree. In NTS, FRRP-1 suppresses mechanical hyperalgesia, like $PrRP_{20}$. Whether these peptides act

through the same or different receptors is not currently known. Recently, it has been shown that the anti-opioid effects of NPFF are absent in mice lacking the PrRP receptor, which suggests that at least some of the central effects may be mediated through the PrRP/GPR10 system, although no significant affinity of NPFF and GPR10 has been demonstrated.

NPFF and NPAF injected into the systemic circulation elevate blood pressure significantly. This effect may be in part mediated by catecholamines as prazocin and guanethidine partially inhibit the elevation in conscious rats. NPFF injected directly into the NTS induces an increase in arterial blood pressure, bradycardia, and inhibition of the cardiac component of the baroreceptor reflex. PrRP20 at this site has been reported to be inactive, but it facilitates heat-evoked blood pressure elevation in the caudal ventrolateral medulla. On this level PrRP thus controls not only somatic but also autonomic responses to pain.

Central Endocrine Regulation

NPFF has been shown to release prolactin in rats, both *in vivo* and from isolated pituitary cells. After icv administration, RFRP-1 also releases prolactin and is more potent that NPFF. Under these circumstances, PrRP did not release prolactin. Thus, the NPFF and RFRP precursor-derived peptides seem to be more potent releasers of prolactin than is PrRP. The effects of PrRP seem to be more pronounced on other hypothalamic and pituitary hormones than is prolactin. PrRP regulates growth hormone release, probably through somatostatin, and luteinizing hormone (LH) surge through gonadotropin-releasing hormone (GnRH). Kiss-1 appears to be a crucial regulator of the hypothalamic GnRH neurons, as intracerebroventricular administration of either full-length 54 aa Kiss-1 or the C-terminal fragment increases both LH and follicle-stimulating hormone secretion. Furthermore, inactivation of GPR54 in mice leads to hypogonadotropic hypogonadism, and mutations in human GPR54 are responsible for the disorder in humans. These findings suggest Kiss-1 and GPR54 as the ligand and receptor in key position in GnRH regulation.

Energy Balance, Feeding, and Water Balance

The presence of very high levels of NPFF in the rat posterior pituitary renders it possible that NPFF regulates vasopressin and/or oxytocin or is co-regulated with these. Indeed, NPFF peptide is practically absent from the pituitary of homozygous Brattleboro rats, which produce no functional vasopressin, although NPFF peptide levels in other parts of the brain are normal. Similarly, salt loading depletes both vasopressin and NPFF storage sites in the pituitary. However, the vasopressin gene expression increases whereas NPFF gene expression decreases after salt loading in rats. Icv NPFF acutely stimulates water intake in rats. NPFF appears to inhibit vasopressin release on the hypothalamic and medullary levels. Since NPFF in at least some species is made in supraoptic and paraventricular neurons and is transported to the posterior pituitary, pituitary nerve terminals are one of the candidate cells that may release NPFF into the blood. NPFF immunoreactivity is detected in human blood in a manner that suggests pulsatile secretion independent of vasopressin. Electroconvulsive therapy also increases blood levels of both vasopressin and NPFF, but the physiological stimulus responsible for the regulation is not known.

NPFF injected into the lateral ventricle of rats suppresses feeding in rats. PrRP injected icv also reduces food intake without affecting drinking, and PrRP mRNA expression in the hypothalamus and brain stem decreases during fasting. PrRP injections also induce UCP-1 mRNA in brown adipose tissue, cause hyperthermia, and increase oxygen consumption. GPR10 knockout mice become hyperphagic and obese with associated endocrine changes in leptin and insulin levels. Peptide 26RFa also increases food intake in mice, suggesting that most mammalian RFamide peptides are involved in regulation of food intake and metabolism, although it is currently unclear to what extent the mechanisms are the same.

NPFF1R and NPFF2R are also found in fat tissue, and the peptides regulate a number of genes in adipocytes, including the β-adrenergic receptor, suggesting that one or several circulating peptides may regulate metabolism. NPFF is secreted in human blood, but the exact identity of the circulating peptide, its origin, and the regulation of its secretion are not known.

See also: Mammalian Neuropeptide Families; Neuropeptide S; Neuropeptides Phylogeny and Evolution; Neuropeptides: Discovery; Neuropeptides: Endocrine Cells; Neuropeptides: Food Intake; Neuropeptides: Pain.

Further Reading

Bonini JA, Jones KA, Adham N, et al. (2000) Identification and characterization of two G protein-coupled receptors for neuropeptide FF. *Journal of Biological Chemistry* 275: 39324–39331.

de Roux N, Genin E, Carel JC, Matsuda F, Chaussain JL, and Milgrom E (2003) Hypogonadotropic hypogonadism due to loss of function of the KiSS1-derived peptide receptor GPR54. *Proceedings of the National Academy of Sciences of the United States of America* 100: 10972–10976.

Elshourbagy NA, Ames RS, Fitzgerald LR, et al. (2000) Receptor for the pain modulatory neuropeptides FF and AF is an orphan G protein-coupled receptor. *Journal of Biological Chemistry* 275: 25965–25971.

Fukusumi S, Fujii R, and Hinuma S (2006) Recent advances in mammalian RFamide peptides: The discovery and functional analyses of PrRP, RFRPs and QRFP. *Peptides* 27: 1073–1086.

Hinuma S, Habata Y, Fujii R, et al. (1998) A prolactin-releasing peptide in the brain. *Nature* 393: 272–276.

Hinuma S, Shintani Y, Fukusumi S, et al. (2000) New neuropeptides containing carboxy-terminal RFamide and their receptor in mammals. *Nature Cell Biology* 2: 703–708.

Laurent P, Becker JA, Valverde O, et al. (2005) The prolactin-releasing peptide antagonizes the opioid system through its receptor GPR10. *Nature Neuroscience* 8: 1735–1741.

Lefrere I, De Coppet P, and Camelin JC (2002) Neuropeptide AF and FF modulation of adipocyte metabolism: Primary insights from functional genomics and effects on beta-adrenergic responsiveness. *Journal of Biological Chemistry* 277: 39169–39178.

Muir AI, Chamberlain L, and Elshourbagy NA (2001) AXOR12, a novel human G protein-coupled receptor, activated by the peptide KiSS-1. *Journal of Biological Chemistry* 276: 28969–28975.

Ohtaki T, Shintani Y, and Honda S (2001) Metastasis suppressor gene KiSS-1 encodes peptide ligand of a G-protein-coupled receptor. *Nature* 411: 613–617.

Panula P, Aarnisalo AA, and Wasowicz K (1996) Neuropeptide FF, a mammalian neuropeptide with multiple functions. *Progress in Neurobiology* 48: 461–487.

Panula P, Kalso E, Nieminen M, et al. (1999) Neuropeptide FF and modulation of pain. *Brain Research* 848: 191–196.

Perry SJ, Yi-Kung HE, Cronk D, et al. (1997) A human gene encoding morphine modulating peptides related to NPFF and FMRFamide. *FEBS Letters* 409: 426–430.

Roumy M and Zajac JM (1998) Neuropeptide FF, pain and analgesia. *European Journal of Pharmacology* 345: 1–11.

Seminara SB, Messager S, Chatzidaki EE, et al. (2003) The GPR54 gene as a regulator of puberty. *New England Journal of Medicine* 349: 1614–1627.

Simonin F, Schmitt M, Laulin JP, et al. (2006) RF9, a potent and selective neuropeptide FF receptor antagonist, prevents opioid-induced tolerance associated with hyperalgesia. *Proceedings of the National Academy of Sciences of the United States of America* 103: 466–471.

Vilim FS, Aarnisalo AA, Nieminen ML, et al. (1999) Gene for pain modulatory neuropeptide NPFF: Induction in spinal cord by noxious stimuli. *Molecular Pharmacology* 55: 804–811.

Yang HYT and Iadarola MJ (2006) Modulatory roles of the NPFF system in pain mechanisms of the Spinal Cord. *Peptides* 27: 943–952.

Neuropeptide Inactivation or Metabolism

B P Roques, Université Paris Descartes, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

More than 100 peptides have already been characterized in mammalian brains, but their physiological roles remain to be clarified. For some of them, this has been achieved by modulating *in vivo* the extraneuronal levels of the secreted neuropeptides because the interruption of the neuropeptidergic transmission occurs mainly by inactivating the ectopeptidases anchored in the membranes of neuron and/or glial cells. Selective inhibition of these enzymes, which belong essentially to the zinc metallopeptidase group, gives us access to the functions of the peptides. The induced effects can be compared to those obtained by using selective agonists or antagonists or invalidation of the genes encoding the components of a neuropeptidergic pathway (neuropeptide precursor, receptor, inactivating peptidases). These approaches are briefly presented here for some important neuropeptides (e.g., enkephalins, angiotensin, substance P, and cholecystokinin) for which all these methods have been used. Finally, the putative clinical interest of this physiological approach is discussed.

Information Transfer by Neuropeptides versus Neurotransmitters

It is now well known that neuropeptides act as messengers in the central nervous system (CNS); they meet most of the criteria characterizing neurotransmitters (e.g., acetylcholine, glutamic acid, dopamine, and serotonin). For example, (1) neuropeptides are localized in neuronal cells but not exclusively at the nerve terminal level; (2) neuropeptides are synthesized from large precursors that are generated by gene expression and then undergo subsequent maturation by well-defined proteases cleaving the precursor at the level of dibasic (Arg, Lys) sequences; (3) neuropeptides are stored in large dense-core (or granular) vesicles, from which they are released following neuron stimulation at high frequency; (4) neuropeptides interact with specific receptors (usually of the G-protein-coupled family) not necessarily located at short distances from the site of release (evidenced by the mismatch between the distribution of many neuropeptides and their targets in brain structures); (5) the neuropeptide–receptor interaction is expected to produce changes in ionic permeability of the targeted neuron and modifications of cellular biological events and gene expression; and (6) the neuropeptidergic transmission is interrupted by well-adapted biological mechanisms, such as peptidases able to cleave the active peptides into inactive metabolites.

Despite this, there are several differences between neuropeptides and neurotransmitters. Neurotransmitters are almost exclusively located at the nerve terminal, where they are synthesized through an enzymatic reaction cascade. They are stored in small vesicles and released by low-impulse-frequency action potentials. Receptors for neurotransmitters are usually located in synapses and therefore at a very short distance from the site of effector release. The interruption of the neurotransmitter action is ensured mainly, although not exclusively, by the secreted neurons through an efficient reuptake process. Thus, the concentration level of neurotransmitters rapidly becomes too low to bind the postsynaptic receptors, and this clearing mechanism could eventually be assisted by presynaptic receptors modulating negatively the release of the neurotransmitter, as is clearly shown in the case of dopamine. Consistent with these features, the cell-to-cell transmission is very rapid. In contrast, neuropeptide receptors are located far from the site of neuropeptide release, the diffusion of these large effectors is slow, and the inactivating peptidases have a rather low efficiency compared to the neurotransmitters reuptake processes; thus, the neuropeptidergic transmission is generally slow (Figure 1).

Due to this (often called volume transmission), we expect that neuropeptides play the role of neuromodulator rather than neurotransmitter, accounting for their critical role in behavioral control. This is reinforced by the frequent coexistence in the same neurons of one (or more) neurotransmitter(s) and neuropeptide(s). Thus, numerous neuropeptides are co-localized in dopaminergic, cholinergic, or GABAergic neurons, most frequently in interneurons able to modulate the firing of long neuronal pathways. Examples are:

- cholecystokinin (CCK): Asp-Tyr(SO_3H)-Met-Gly-Trp-Met-Asp-Phe-NH_2;
- enkephalin (ENK): Tyr-Gly-Gly-Phe-Met(Leu);
- substance P (SP): Arg-Pro-Lys-Pro-Gln-Phe-Phe-Gly-Leu-Met-NH_2; and
- neurotensin (NT): pGlu-Leu-Tyr-Glu-Asn-Lys-Pro-Arg-Arg-Pro-Tyr-Ile-Leu.

The single or co-release of two different effectors provides a large variety of responses at the level of the recruited neurons, located either in the immediate surrounding area or at longer distance from the secreted compounds. This increases the amount and diversity of information required for brain functioning.

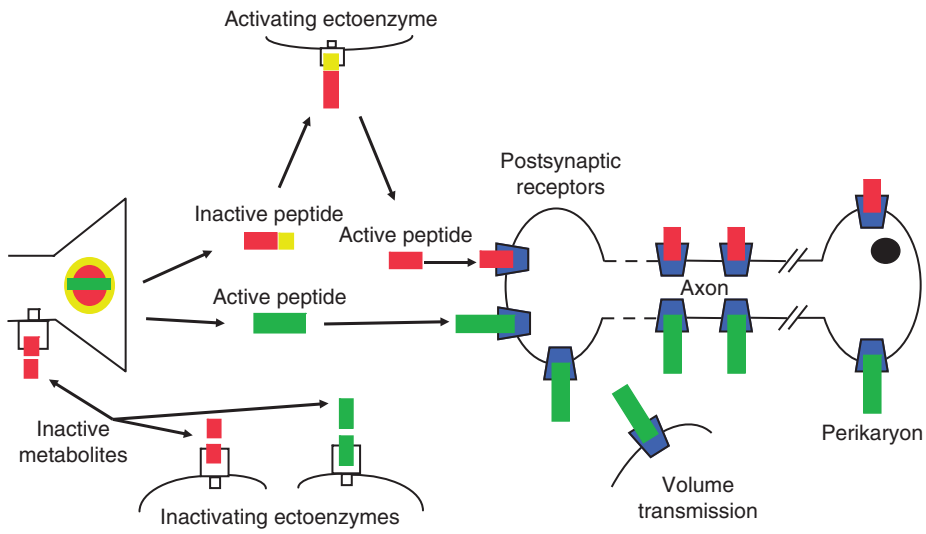

Figure 1 Schematic representation of a neuropeptidergic pathway.

Table 1 Nomenclature of the main peptidases involved in brain metabolism of neuropeptides[a]

Name	*Enzyme commission (EC) number*	*[Abbreviation] and other names*	*Type*[b]
Endopeptidase-24.11	3.4.24.11	[E-24.11] Neutral endopeptidase, neprilysin, enkephalinase, CALLA, CD10	II
Endothelin-converting enzyme	3.4.24.71	$[ECE]^{-1}$	II
Endopeptidase 24–16	3.4.24.16	Neurolysin	Cyt
Angiotensin-converting enzyme	3.4.15.1	[ACE] Peptidyl dipeptidase A	I
Dipeptidyl-peptidase IV	3.4.14.5	[DPP-IV] X-pro-dipeptidyl-aminopeptidase, CD26	II
Thyrotropin-releasing hormone-degrading enzyme	3.4.19.6	[TRH-degrading enzyme] Pyroglutamyl aminopeptidase II	II
Tripeptidyl-peptidase	3.4.14.10	TPP-II, subtilisin-like serine peptidase	Cyt
Aminopeptidase N	3.4.11.2	[APN] Aminopeptidase M, CD13	II
Aminopeptidase A	3.4.11.7	[APA] Gutamyl aminopeptidase, aspartate aminopeptidase, angiotensinase A	II

[a]Except for DPP-IV and TPP-II, which are serine peptidases, all these enzymes belong to the superfamily of zinc metallopeptidases.
[b]I, N-terminus of the protein outside the cell; II, C-terminus of the protein outside the cell; Cyt, enzyme found mainly in the cytosol with a small fraction bound to the internal or external part of the membrane.

Peptidase-Induced Inactivation of Neuropeptides

A large number of peptidases that are able to cleave neuropeptides are present in mammalian brains, but only some of them have been clearly characterized as inactivating enzymes that generate inactive metabolites from an active neuropeptide (Table 1). Most of these enzymes belong to the zinc metallopeptidase group, which can be divided into two subgroups: (1) the exopeptidases, which remove one (aminopeptidase N (APN) and aminopeptidase A (APA)) to three (tripeptidyl-peptidase (TPP)) amino acids from the N- or C-terminus of neuropeptides and (2) the endopeptidases (neutral endopeptidase (NEP), angiotensin-converting enzyme (ACE), and endothelin-converting hormone (ECE)), which can hydrolyze peptide bonds inside the peptide sequence. Another subclassification is by localization inside or outside the neurons. The most important group of inactivating peptidases possess a short intracytosolic domain, followed by a transmembrane sequence anchoring the enzyme in the external membrane of the neuron with the enzyme active site located in the extracellular domain (Figure 1); these peptidases are designated ectoenzymes.

In a limited number of cases, neuropeptides are released in an inactive form and must be activated by the removal of a part of the secreted precursor.

This is the case with big-endothelin, which is processed in the external medium by the ectopeptidase ECE to provide the active endothelin. Angiotensin II and CCK undergo similar processes in brain, with the peptidase-evoked formation of shorter peptides endowed with different pharmacological activities than their peptide precursors (Figure 1).

Similarities and Differences in Zinc Metallopeptidases Involved in Neuropeptides Catabolism

The early studies of brain peptide degradation were essentially devoted to neurohormones and releasing factors. An increasing interest in inactivating peptidases in the brain emerged after the discovery of the two opioid peptides enkephalins. Tyr-Gly-Gly-Phe-Met (Leu) binds opioid receptors with higher *in vitro* affinities than morphine, but induces a very weak and transient analgesia even after intracerebral administration in rodents. This was shown to result from the cleavage of the Tyr-Gly bond by APN and of the Gly-Phe bond by NEP. The metabolites formed by the action of these two membrane-bound zinc metallopeptidases are completely inactive. NEP was shown to be efficiently inhibited by phosphoramidon, a well-known blocker of the bacterial endopeptidase thermolysin (TLN), which contains 315 amino acids.

The structure of TLN has been resolved by X-ray crystallography with and without various types of inhibitors, including phosphoramidon, and a mechanism of zinc-assisted hydrolysis of the peptide bond was proposed as a result of all these studies (Figure 2). The hypothesized similarity in the organization of the catalytic sites of TLN and NEP was supported by the cloning and sequencing of NEP. Indeed, two consensus sequences of the type HEXXH and EXXXD were found in both enzymes. Each of the amino acids present in these short fragments have been shown to participate in the catalytic hydrolysis of the peptide bond in TLN. The same mechanism for NEP was established by several site-directed mutagenesis experiments.

All the critical amino acids present in the active site of TLN were also found in NEP. Unlike TLN, however, NEP possesses an Arg-102 residue located at the edge of the active site, which is probably responsible for the dipeptidyl carboxypeptidase activity of NEP that is observed in the release of Phe-Leu from Leu-enkephalin (Tyr-Gly-Gly-Phe-Leu).

All the zinc metallopeptidases have almost the same mechanism of action. Schematically, hydrolysis occurs through the formation of a penta-coordinated complex of the metal, which includes the three zinc-coordinating amino acids of the peptidase, the oxygen of the scissile bond, and the water molecule that is initially bound to the zinc atom (Figure 2).

The specificity of the zinc metallopeptidases is essentially ensured by van der Waals and ionic interactions between the S_2, S_1, S_1', and S_2' subsites and the lateral chains of the corresponding P_2, P_1, P_1', and P_2' moieties of the substrate (Figure 3). NEP has a broad selectivity and can cleave various short linear or cyclic peptides. The enzyme behaves like an efficient nonspecific endopeptidase, cleaving *in vitro* a large number of linear neuropeptides such as NT, SP, and CCK-8.

The consensus sequences of amino acids present in the catalytic sites of TLN and NEP were used for cloning ACE, ECE, and other zinc metallopeptidases. In the case of aminopeptidases, the second consensus EXXXD is not preserved; but for APN and APA, site-directed mutagenesis studies showed that the third zinc ligand is also a glutamate, as it is in endopeptidases (Figure 2). Moreover, the pure exopeptidase action of these enzymes was shown to be due to the interaction of the free N-terminal amino group of the

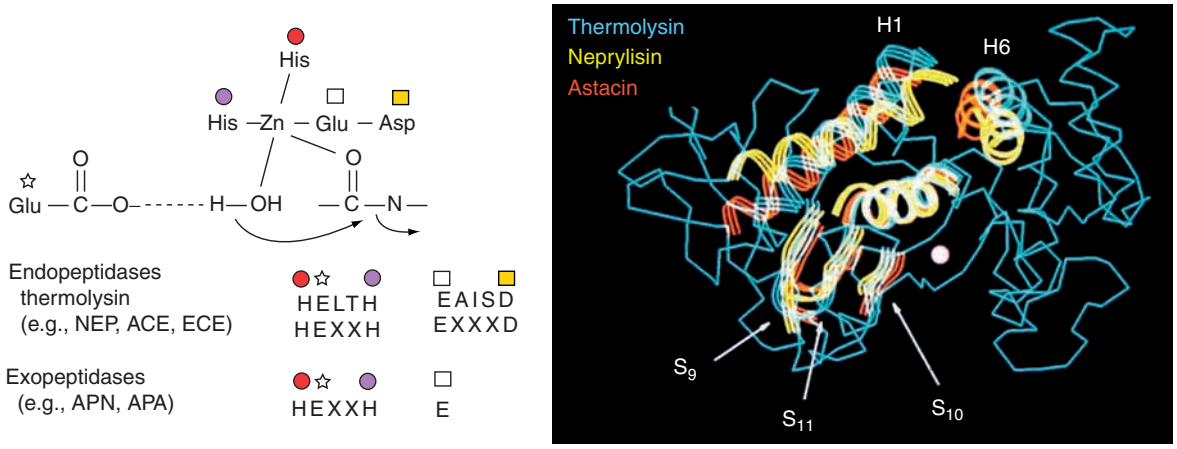

Figure 2 Zinc metallopeptidases: sequence homologies and mechanisms of peptide bond hydrolysis (left) and superposition of active sites of three different enzymes (right).

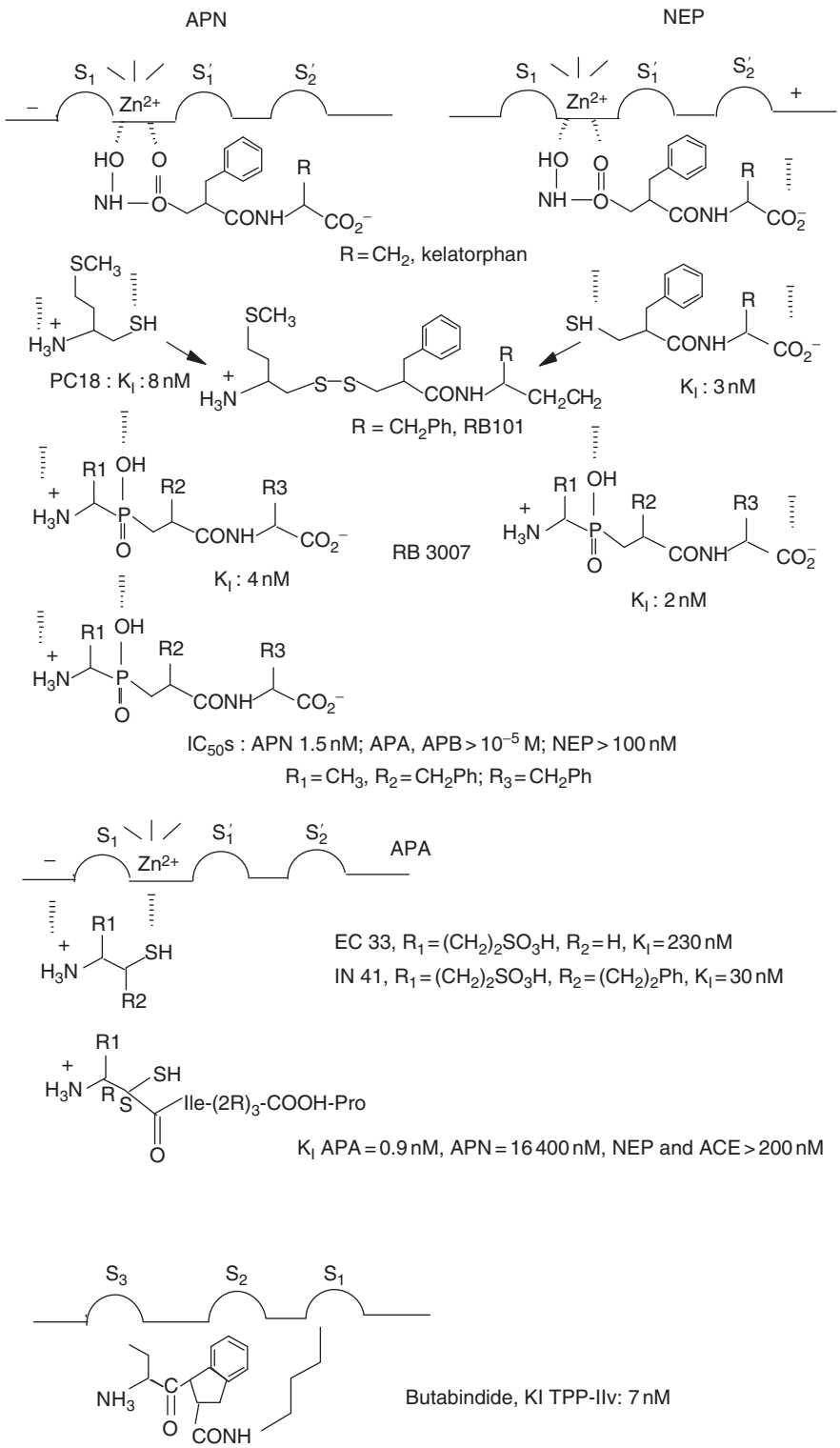

Figure 3 Inhibitors of neuropeptide-inactivating peptidases. ACE, angiotensin-converting enzyme; APA, aminopeptidase A; APN, aminopeptidase N; IC_{50}, median inhibitory concentration; NEP, neutral endopeptidase; TPP, tripeptidyl peptidase.

neuropeptidase substrate with a glutamate, hindering the displacement of the peptides beyond the S_1 subsite.

Another important group of zinc metallopeptidases, designated matrixins, has a slightly different organization of the consensus sequence, which is larger and contains three histidine as zinc ligands – HEXXHXXGXXH – but the mechanisms of peptide-bond hydrolysis is very close to that of TLN

and related enzymes. Several matrixins have been crystallized and co-crystallized, and the structure of their active site was shown to be superimposable on that of TLN and NEP, with the presence of four α-helices and three β-sheets forming both the catalytic site and the S_1' pocket (Figure 2). In 2000, NEP was crystallized, and the structure of the functional part of the protein was found to correspond to a previously proposed model obtained by computer-assisted modeling.

In spite of the large differences in the number of residues in TLN, NEP, neurolysin, and ACE and the very weak sequence homologies among all these enzymes, their crystal structures showed a close arrangement of active sites and an almost identical mechanism of catalysis and inhibitor recognition. The latter parameter is essentially governed by the importance of restriction in access to the active site. The molecular modeling of ECE-1 structure is perfectly consistent with these results.

The use of selective peptidase inhibitors showed that despite the relatively broad specificity of the limited number of brain peptidases, a certain *in vivo* specificity is achieved, at least for some peptides (ENKs, CCK, NT, and angiotensin) governed by the distribution of a given peptidase, its potential neuropeptide substrate, and their specific receptors.

Natural and Synthetic Inhibitors of Neuropeptide Metabolizing Enzymes

The synthesis of selective or dual inhibitors, the latter being able to block different peptidases (e.g., an endopeptidase such as NEP and an exopeptidase such as APN), has been facilitated by their membership of the zinc metallopeptidase group. The development of these inhibitors is based on the introduction into the structure that recognizes the active subsites ($S_1 \rightarrow S_n$ and $S_1' \rightarrow S_n'$) surrounding the catalytic site of groups capable of strongly binding the zinc atom (e.g., —SH, —COOH, —HONH-CO, —P(O)OH; Table 1). Thus, compounds with nanomolar inhibitory potencies and very high selectivity were obtained for ACE/NEP, ECE-1, APN, APA and for the serine protease TPP (Figure 3).

Natural inhibitors generally issued from cultures of bacteria such as bestatin, amastatin, and phosphoramidon are also often used. However, these substances are less potent and selective than inhibitors designed to fit the enzyme active site optimally.

The disadvantage of peptidase inhibitors compared with exogenous agonists or antagonists is their lower pharmacological potencies. However, this disadvantage is compensated by their more specific effects, which correlate with the phasic release of their peptide substrate in brain structures recruited by a particular stimulus (e.g., pain, stress, or emotion) and the absence or low change in either the secretion of the peptide or expression of its targets (e.g., metabolizing enzymes or receptors).

Thus, as expected, the maximal antinociceptive effects of NEP/APN dual inhibitors are in the 40–70% range of those produced by morphine; this is consistent with a local enhancement of endogenous ENKs that is too low to saturate brain opioid receptors. This eliminates or reduces strongly the major side effects of exogenous opiates (respiratory depression, constipation, tolerance, and dependence).

Biochemical Criteria Used to Characterize Neuropeptide-Activating or -Inactivating Peptidases

Anatomical Criteria

Several criteria must be fulfilled for an enzyme to be considered a neuropeptide-metabolizing peptidase. The most frequently claimed is the existence of a co-localization of the three components of the neuropeptidergic pathway (neuropeptide, specific receptors, and metabolizing peptidase). This criterion is often difficult to establish, due to the previously discussed extended distribution of the three proteins in a brain structure. Nevertheless, quantitative autoradiography using highly selective and potent tritiated or radioiodinated inhibitors of NEP and ACE has clearly demonstrated the discrete distribution of these metallopeptidases in brain structures where their physiological role is expected (in the case of ENKs, the spinal cord, the periaqueductal gray matter (PAG), the thalamus for the control of pain, and the mesolimbic and mesocortical pathways and the amygdala for the control of mood). This was completed recently through the discrete autoradiographic visualization of APN, the other ENK-inactivating peptidase showing its presence in brain tissue, although this enzyme was earlier thought to be present only in cerebral microvessels (Figure 4).

Lesion studies of neurons with kainic acid infused in a particular structure and subsequent labeling with radioactive probes showed the preferential localization of peptidases (e.g., NEP) on neurons which was consistent with the visualization of their mRNAs by *in situ* hybridization. In addition, lesions of the caudate-putamen, in the case of NEP, demonstrated a distant and large reduction in NEP and the opioid receptors in the pallidum, indicating a presynaptic co-localization of the enzyme and the receptors on axon terminals. In the spinal cord, unilateral rhizotomy demonstrated that approximately 60% of opioid

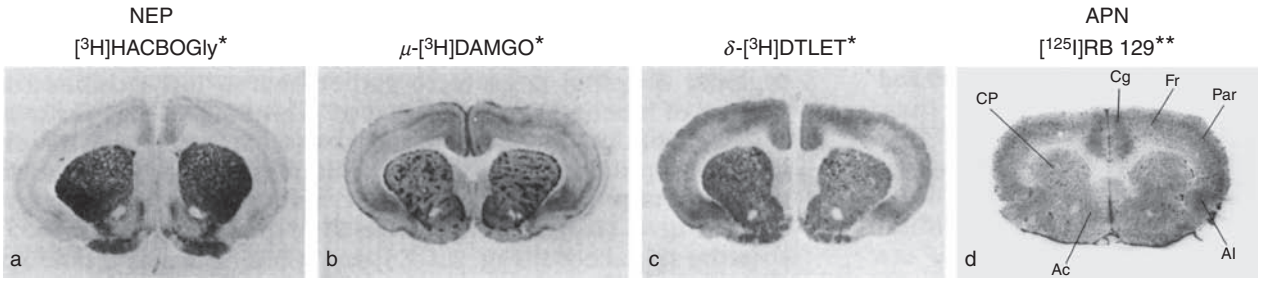

Figure 4 Co-localizations of: (a) NEP; (b) μ-opioid receptors; (c) δ-opioid receptors; (d) APN. APN, aminopeptidase N; NEP, neutral endopeptidase. Statistical analysis $^{*}p < 0.05$ and $^{**}p < 0.01$.

receptors are presynaptically located on afferent fibers and 40% are postsynaptically located. Interestingly, no change was observed regarding the level of NEP, suggesting that the peptidase and opioid peptides are co-localized in interneurons regulating the secretion of efferent neurons.

Most of these results were confirmed by immunohistochemical methods, showing in many cases the ultrastructural localization of the components of a neuropeptidergic pathways. Thus, it has been demonstrated that NEP is not concentrated at synaptic interfaces but, rather, uniformly distributed on the surface of the neuronal perikaryon and dendrites. The vicinal localizations of NEP visualized by the immunogold method and of SP in the substantia nigra, where the density of SP is much higher than that of ENK, suggest that NEP could play a role in the inactivation of SP in this structure.

Analytical Criteria: Microdialysis Experiments

The second method used to categorize an inactivating peptidase as a brain enzyme is the characterization of the metabolites of the peptidase-cleaved neuropeptide by microdialysis. This technique is now currently used in various animals, including mice; it allows the extracellular fluid secreted from a group of neurons in a given structure to be collected. Highly sophisticated methods using capillary electrophoresis coupled with fluorescence detection or online mass spectrometers (matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF)/time-of-flight mass spectrometry (TOF MS)) have now been developed that replace the immunodetection of the peptides. These methods, which can be used in freely moving animals, are preferred to the *in vitro* techniques in which homogenates or even slices of brain structures are employed. In these cases, the diffusion of the studied peptides, generally radiolabeled, across the tissue cannot escape the collision with peptidases not concerned with the physiological inactivation process. For these reasons, conflicting and uninterpretable results have been obtained for several important neuropeptides (e.g., ENK, SP, NT, and CCK). Another advantage of the *in vivo* microdialysis is the possibility of using the method in animals with different pathological conditions (e.g., pain, stress, and anxiety).

Nevertheless, to be entirely conclusive, the microdialysis experiments should be associated with the local or systematic administration of selective inhibitors of the incriminated metabolizing enzymes. This seems to have been achieved systematically only in the case of the ENKs, demonstrating clearly that in the superfused spinal cord of anesthetized rats dual NEP/APN inhibitors, such as kelatorphan, inhibit completely *in vivo* endogenous ENK catabolism. The use of selective NEP or APN inhibitors showed that each peptidase participates about equally in ENK inactivation. Interestingly, the level of extracellularly secreted ENKs is strongly increased by painful stimuli. Likewise, the systemic administration of dual inhibitors in freely moving rats produces a long-lasting four- and twofold increase in ENK levels measured by immunocytochemistry in the nucleus accumbens and the PAG, respectively.

Strikingly, although various metallo-ectoenzymes, in particular NEP, ACE, ECE, and APN, alone or in association, have been suspected of being involved in the *in vivo* inactivation of almost all the neuropeptides, often with conflicting results, the selective inhibitors of these peptidases have not been systemically added to the vehicle of the microdialysis experiments. In these cases, the role of inactivating peptidase cannot be attributed to these enzymes.

Nevertheless, measurements of ENK levels in presence or in absence of inhibitors showed that, for example, in primates an intravenous (i.v.) administration of SCH 32615, a NEP blocker that enters the brain, enhanced the level of ENKs in cerebrospinal fluid ninefold compared to controls (administered the vehicle alone). In addition, several microdialysis studies *in vivo* in rats using the dual NEP/APN inhibitor kelatorphan demonstrated that NEP probably cleaves SP in the substantia nigra but not in the spinal cord, where another specific enzyme appears to be implicated.

Pharmacological Experiments Required to Attribute the Role of Activating or Inactivating Peptidase to a Neuropeptide-Cleaving Enzyme

A rarely reported characteristic of the neuropeptidergic transmission is the frequent lack of direct relationship between the presence of specific receptors for a neuropeptide and its tonic or phasic release in a given structure. This has been studied in detail with the ENKs using selective or dual inhibitors of NEP and APN.

The effects of the inhibitors depend on (1) the magnitude and duration of the ENK release evoked by a particular stimulus, which varies in the different enkephalinergic pathways (this accounts for the higher analgesic responses observed using dual inhibitors such as RB 101 to nociceptive thermal stimuli than to mechanical stimuli) and (2) the efficiency of the inhibition of both peptidases, depending on their levels and localization with regard to the receptors and the rate of clearance of the inhibitor(s) from the site of peptidase(s) action. These considerations could explain why in many pharmacological assays the administration of the rather μ-selective antagonist naloxone does not produce a measurable effect on the response induced by the circulating ENKs. Thus, naloxone fails to affect pain thresholds in normal volunteers, but lowers the thresholds of patients with naturally high pain sensitivity.

The large differences in tonic or phasic release of ENKs in various tissues are nicely illustrated in structures such as the PAG, the locus coeruleus (LC), and the nucleus tractus solitarius (NTS). As shown in Figure 5(b), the systemic administration of RB 3007 induces a long-lasting increase in extracellular levels of Met-enkephalin measured by microdialysis in PAG, a structure critically involved in the control of pain. Accordingly, a significant antinociceptive activity was measured in rodents with the same time-course after the administration of RB 3007 (Figure 5(a)).

The LC contains a densely packed group of noradrenaline-containing neurons which develop a hypersensitivity following chronic morphine treatment, and this is considered to be one of the main causes of withdrawal syndrome and the development of dependence. Interestingly, kelatorphan was able to potentiate strongly the firing of LC neurons but had no significant intrinsic effect, indicating that there is little or no tonic endogenous opioid action in this brain region. This is probably the main reason why chronic treatment with peptidase inhibitors is unable to induce physical and psychic dependence compared to morphine.

The major side effect of opiates is a central respiratory depression, mainly due to the inhibition of the bulbar respiratory neurons. Unlike morphine, dual inhibitors applied locally in this brain structure or administered by systemic routes are unable to produce a reduction of respiratory frequency, suggesting again a weak tonic release of endogenous ENKs in the neural network that generates the motor respiratory activity.

The lack of psychological dependence observed after chronic treatment with dual inhibitors of ENKs used at high doses could be related to a rather low stimulation of the opioid receptors located in the ventral tegmental area (VTA), which are able to activate the dopamine receptors, leading to an increased release of dopamine in the nucleus accumbens. The chronic activation by opiates of this mesolimbic pathway seems to be involved in addiction. The failure of dual inhibitors to establish a conditioned place preference probably results from a lower recruitment of opioid receptors than opiates in the VTA. This hypothesis is supported by the minimal changes in dopamine release in the nucleus accumbens after local administration of kelatorphan into the VTA compared to morphine.

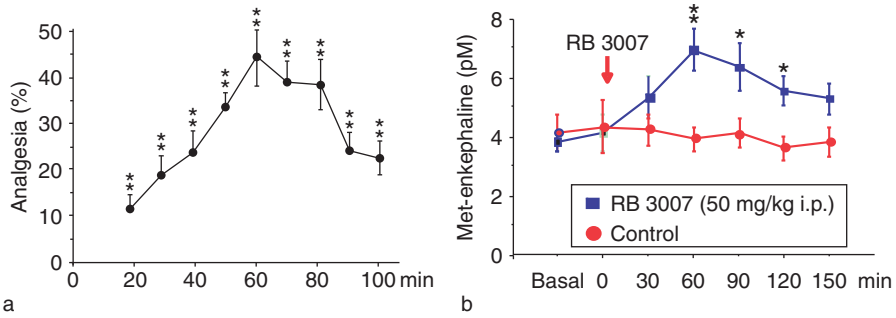

Figure 5 (a) Analgesic effect of dual NEP/APN inhibitor RB 3007, administered IP (50 mg kg^{-1}); (b) correlation with its induced increase in extracellular Met-enkephalin levels in the PAG, as measured by microdialysis in freely moving rats. APN, aminopeptidase N; IP, intraperitoneally; NEP, neutral endopeptidase; PAG, periaqueductal gray. The ↓ indicates the time of RB 3007 administration. Statistical analysis $^{*}p < 0.05$ and $^{**}p < 0.01$.

Investigation of the Physiological Roles of a Given Neuropeptide through Selective Inhibitor-Induced Modulation of Its Extracellular Levels

As previously discussed, peptidases capable of inactivating a given neuropeptide, such as NEP or ACE, are distributed in a very large number of structures, both in the CNS and in peripheral organs, as shown by autoradiography. This could indicate that they contribute to the activation or inactivation of a large number of peptides. However, only some physiological substrates (ENKs and SP in the CNS; atrial natriuretic peptides, bradykinin, and endothelin-1 for NEP in peripheral organs; and angiotensin I and bradykinin for ACE) have been clearly identified for both peptidases. Other peptides may be characterized in the future, but it appears that in numerous structures these peptidases have been recruited only during a phase of development or that the tonic or phasic release of putative unknown substrates by these peptidases is too low to give significant pharmacological responses following enzyme inhibition. The main interest in using selective inhibitors is to allow the investigation of the physiological actions of a (neuro)peptide by modulating its extracellular levels. Nevertheless, this requires to verify that the responses given by the selective inhibitors are really due specifically to the protection of the studied peptide from the targeted peptidase. This is illustrated in the case of ENKs, for which the invalidation of the preproenkephalin gene leads to a decrease in the pain threshold and an enhancement in anxiety and agressivity. These responses are the opposite of those observed by the systemic administration of dual NEP/APN inhibitors. Moreover, we recently showed that a dual inhibitor such as RB 101 loses its antinociceptive property almost completely in preproenkephalin knockout mice, supporting the specific role of ENKs as selective substrates for NEP and APN in brain. Likewise, knockout mice for the gene encoding NEP show a limited increase in the pain threshold, indicating that APN cannot completely compensate for the absence of NEP. Further, the increase in tissues of ENK levels by a feedback mechanism occurs only in certain regions, suggesting, as previously discussed, that the physiological role of NEP is restricted to structures with a sufficient phasic release of these neuropeptides. And, unlike naloxone, selective antagonists of the various neuropeptides (NT, SP, somatostatin, and CCK) that were shown generally *in vitro* to behave as substrates of NEP were unable to modify the responses given by ENKs protected by dual inhibitors.

The inhibition of inactivating peptidases to explore the physiology of a brain peptide has several advantages over the use of a selective receptor antagonist or gene invalidation of its receptors or of its neuropeptide precursor. During the blockade of a peptide receptor by an antagonist, the circulating levels of the peptide messenger remain unchanged or can even increase by a feedback mechanism. This also occurs following the invalidation of neuropeptide receptors. In such cases, depending on the volume of the neuropeptidergic transmission and the broad specificity of the inactivating enzymes, an even more important drawback could result – the possible inhibition of another peptidase by the generated orphan peptide could result in enhanced levels of the natural neuropeptide substrate of the mistakenly targeted peptidase, possibly modifying its physiological actions.

Thus, emotional stressors produce an increase of SP in the medial amygdala and anxiety-related behavior blocked by a specific SP antagonist. The anxiolytic effects of RB 101 suggest that, at least in this structure, SP is probably not inactivated by NEP. Contradictory results (analgesia and hyperalgesia) were observed following the local administration of SP in the spinal cord or the PAG. This was interpreted by a modulation of the endogenous opioid release, a part of which could be evoked by the binding of SP or one of its metabolites SP(1–7) to NEP.

Physiological Interactions between Two Neuropeptidergic Systems Studied by Inhibition of Metabolyzing Enzymes

Among the neuropeptides, CCK-8 (Asp^1-Tyr(SO_3H)-Met-Gly-Trp-Met-Asp-Phe^8NH_2) was shown to be cleaved *in vivo* by APA, leading to CCK-7, and to be cleaved by the ubiquitously distributed tripeptidyl peptidase II (TPP-II, EC 3.4.14.10), mainly enriched in the cytosol but also present as an internal and/or external component of the cell membrane. Butabindide is a highly potent and selective inhibitor of TPP-II (Figure 3). In contrast to various neuropeptides, including ENKs, the degradation of CCK-8 by rat brain membranes is much slower, suggesting that peptidases could be associated with other clearing mechanisms. Thus, an uptake system has been identified for CCK-8 that could function to accumulate CCK-8 and/or CCK-7 for subsequent degradation by TPP-II inside the cell, although we cannot exclude a role also for the dipeptidyl-aminopeptidase (DAP; EC 3.4.13.11).

Interestingly, in many cases, the metabolism of CCK-8 leads either to fragments interacting with

about the same affinity with CCK-1 and CCK-2 receptors, such as CCK-7, or to shorter metabolites such as CCK-5 or CCK-4. These small fragments, shown to be present in brain, interact selectively with high affinities with the CCK-2 receptor and could play a major role in the control of emotion, vigilance, and memory processes. Therefore, the metabolic cascade involved in CCK-8 metabolism warrants further study, especially because the administration in humans of CCK-4 or the CCK-5-related peptide pentagastrin triggers panic attacks.

An antagonism between opiates and CCK has been reported, and the association of a dual NEP/APN inhibitor with a CCK-2 antagonist allowed the mechanism of this process to be investigated. Thus, the analgesic effect of RB 101 in rodents is potentiated (200–800%) by the co-administration of CCK-2 antagonist, suggesting that endogenous CCK counteracts the action of endogenous opioids in all animal models of pain in which RB 101 was found active. Similarly, the antidepressant effects of RB 101, shown to be dependent on δ-opioid receptor stimulation, are enhanced by the CCK-2 antagonist L-365,260. They appear to result from an increase in dopamine in the nucleus accumbens and are almost completely suppressed by D1 and D2 antagonists. The opioid system is able, in turn, to regulate the release of CCK peptides. Thus, the stimulation of the δ-opioid receptor by endogenous ENKs (RB 101) or by a selective δ agonist enhances the release of endogenous CCK-8. These responses are antagonized by the δ-selective antagonist naltrindole (NTI) supporting a critical role of the δ-opioid receptor in the modulation of the CCK-2 system.

Increasing the levels of endogenous ENKs by RB101 was shown to induce antidepressant-like effects in relevant paradigms, such as forced swimming or the conditioned suppression of motility in rats. Behavioral studies showed that blockade of CCK-1 and CCK-2 receptors produces opposite effects. Thus, the antidepressant-like effects of RB 101 were suppressed by the CCK-1 receptor antagonist and enhanced by the CCK-2 receptor antagonist PD-134,308. The effects of RB 101 were potentiated by microinjection of PD-134,308 in the anterior nucleus accumbens and the central amygdala but not in the caudate nucleus, suggesting that the mesolimbic system plays an important role in the interaction between the CCK and opioid systems in the control of these behavioral responses.

Several Zinc Metallopeptidases Produce Two Active Neuropeptides, Angiotensin III and IV, from the Brain Angiotensin II

The physiological role of the brain renin–angiotensin systems was recently clarified, thanks to the use of highly potent and selective inhibitors of APA and APN, two zinc metallopeptidases involved in the degradation of angiotensin II (AII) (Figure 6). This peptide is formed under the action of ACE, both at the periphery and in the CNS. In the circulation, this peptide is responsible of blood pressure (BP) increase, and several studies suggest that AIII could be the peptide involved in the control of BP at the brain level. Selective inhibition of the formation of AIII by APA and blockade of its degradation by APN have confirmed unambiguously the hypothesis about the central role of AIII in the regulation of BP. Thus, EC 33 increased the half-life of [^{3}H]/AII injected intracerebrovascularly (ICV) into conscious mice and completely blocked the formation of [^{3}H]/AIII, whereas the APN inhibitor PC18 (Figure 3) enhanced the half-life of AIII by a factor of four. The central bioavailability of EC 33 was strongly enhanced by the

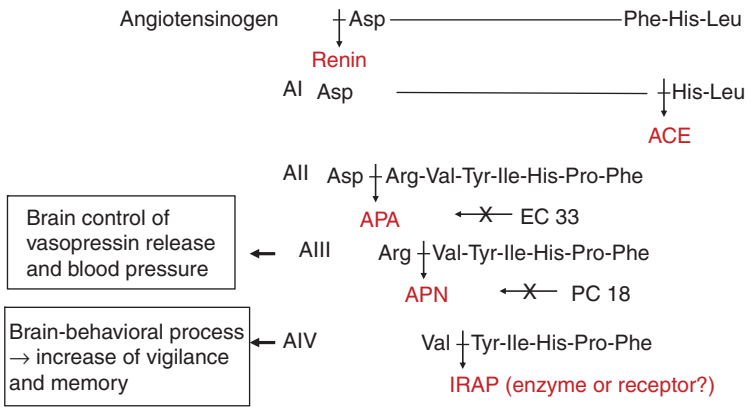

Figure 6 Angiotensinogen-derived regulatory peptides. ACE, angiotensin-converting enzyme; APA, aminopeptidase A; APN, aminopeptidase N; IRAP, insulin-regulated zinc aminopeptidase.

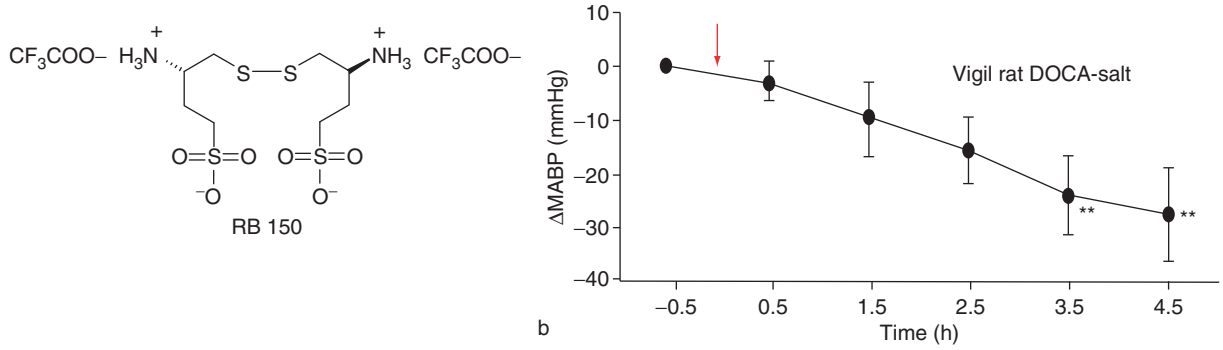

Figure 7 EC 33 disulfide (RB 150): (a) structure; (b) entering the brain after i.v. administration (1 mg kg^{-1}) and inducing blood pressure reduction. ΔMABP, change in mean arterial blood pressure; DOCA, desoxycorticosterone acetate. Statistical analysis $^{**}p < 0.01$.

formation of a disulfide bond in RB 150 ([H_2N-CH ($CH_2CH_2SO_3H$)-CH_2S]$_2$) (Figure 7(a)).

In the brain, RB 150 gives the active form EC 33 and produces a long-lasting (24 h) hypotensive effect in rats following i.v. administration. A possible peripheral effect can be rejected because the i.v. injection of EC 33, which is unable to cross the blood–brain barrier, has no effect on BP.

The action of APN on AIII generates AIV, which was shown to facilitate memory retention and retrieval in rats using various experimental models. This could be related to the long-term potentiation elicited by AIV in the hippocampus. Recent studies showed that the target of AIV, designated AT4 receptor, corresponds to insulin-regulated zinc aminopeptidase (IRAP). The mechanism by which IRAP improves memory function remains unclear, but the use of selective inhibitors is expected to clarify this fascinating problem. Two other neuropeptides, NT and TRH, seem to be almost exclusively inactivated in the CNS by the cleavage of their sequences by EC 24.16 and pyroglutamyl aminopeptidase II, respectively. The problem is more complicated with dipeptidyl-peptidase (DPP-)IV, which has several substrates in the brain and peripheral organs, although the most abundant brain peptide, neuropeptide Y (NPY), appears to be essentially inactivated by this peptidase. In all cases, microdialysis experiments, both with and without selective inhibitors, are required to confirm these hypotheses.

Conclusion

The modification of the extracellular levels of neuropeptides is a well-adapted means of investigating the physiology of a neuropeptidergic pathway. This requires the use of several *in vivo* methods able to delineate the role of the inactivating peptidases in various structures. This has been achieved for some neuropeptides (ENKs, angiotensins, and CCK), leading to promising pharmacological results in analgesia, depression, and BP control that have yet to be evaluated in clinical trials. Approximately 2% of the human genome could correspond to the zinc metallopeptidases, which represent putative pharmaceutically interesting targets. An anticipated, fascinating step in the study of the newly discovered activating/inactivating peptidases will be to achieve their de-orphanization, leading to the characterization of both the peptide substrates and subsequently their receptors.

See also: Angiotensin II; Angiotensin Actions on and Within Brain; Corticotropin-Releasing Hormone and Urocortins: Binding Proteins and Receptors; Hypocretin/Orexin and MCH and Receptors; Invertebrate Neurohormone GPCRs; Mammalian Neuropeptide Families; Melanocortins: Brain Effects; Neuropeptides: Discovery; Peptidergic Receptors.

Further Reading

Barret AJ (1995) *Methods in Enzymology, vol. 248: Proteolytic Enzymes, Aspartic and Metallo Peptidases*. San Diego: Academic Press.

Chen H, Noble F, Roques BP, and Fournié-Zaluski MC (2001) Long lasting antinociceptive properties of enkephalin degrading enzyme (NEP and APN) inhibitor prodrugs. *Journal of Medicinal Chemistry* 44: 3523–3530.

Fournié-Zaluski MC, Fassot C, Valentin B, et al. (2004) Brain renin angiotensin system blockade by systemically active aminopeptidase A inhibitors: A potential treatment of salt-dependent hypertension. *Proceeding of the National Academy of Sciences USA* 101: 7775–7780.

Fournié-Zaluski MC and Roques BP (2002) New selective aminopeptidase N inhibitors as potential therapeutics. In: Langer J and Ansorge S (eds.) *Ectopeptidases: CD13/Aminopeptidase N and CD26/Dipeptidylpeptidase IV in Medicine and Biology*, pp. 51–94. New York: Kluwer Academic/Plenum Publishers.

Hökfelt T, Millhorn D, Seroogy K, et al. (1987) Coexistence of peptides with classical neurotransmitters. *Experientia* 43: 768–780.

Kenny AJ and Boustead CM (eds.) (1997) *Cell-Surface Peptidases in Health and Disease*. Oxford: BIOS Scientific Publishers Limited.

Köning M, Zimmer AM, Steiner H, et al. (1996) Pain responses, anxiety and aggression in mice deficient in pre-proenkephalin. *Nature* 383: 535–538.

Matthews BW (1988) Structural basis of thermolysin and related zinc peptidases. *Accounts of Chemical Research* 21: 333–340.

Noble F and Roques BP (2002) Inhibitors of the enkephalin-catabolism as a new therapeutic tool in opioid dependence. In: Maldonado R (ed.) *Molecular Biology of Drug Addiction*, pp. 61–65. Totowa, NJ: Humana Press.

O'Cuinn G, O'Connor B, Gilmartin L, and Smyth M (1995) Neuropeptide inactivation by peptidases. In: O'Cuinn G (ed.) *Metabolism of Brain Peptides*, pp. 23–68. Boca Raton, FL: CRC Press.

Roques BP (2000) Novel approaches to targeting neuropeptide systems. *Trends in Pharmacological Sciences* 21: 475–483.

Roques BP, Noble F, Daugé V, Fournié-Zaluski MC, and Beaumont A (1993) Neutral endopeptidase 24.11: Structure, inhibition and experimental and clinical pharmacology. *Pharmacological Review* 45: 87–146.

Rose C, Vargas F, Facchinetti P, et al. (1996) Characterization and inhibition of a cholecystokinin-inactivating serine peptidase. *Nature* 380: 403–409.

Saria A, Hauser KF, Traurig HH, et al. (1997) Opioid-related changes in nociceptive threshold and in tissue levels of enkephalins after target disruption of the gene for neutral endopeptidase (EC 3.4.24.11) in mice. *Neuroscience Letters* 234: 27–30.

Valverde O, Maldonado R, Fournié-Zaluski MC, and Roques BP (1994) Cholecystokinin B antagonists strongly potentiate antinociception mediated by endogenous enkephalins. *Journal of Pharmacological Experimental Therapeutics* 270: 77–88.

Neuropeptide Receptors – Drug Development

D Hoyer, Novartis Institutes for Biomedical Research Basel, Basel, Switzerland; and The Scripps Research Institute, La Jolla, CA, USA
T Bartfai, The Scripps Research Institute, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Neuropeptide Ligands

Neuropeptides are 5- to 50-amino-acid-long linear polypeptide gene products, synthesized on ribosomes in the form of preprohormones, which are larger precursors that have to be cleaved and often posttranslationally modified. The first neuropeptide discovered was substance P, with potent pressor effects noted by Euler and Gaddum in 1931. Its sequence was determined more than 40 years later by Leeman and colleagues. The pace of discovery of neuropeptides picked up when Mutt and colleagues started to isolate C-terminally amidated peptides, with the rationale that such peptides are unlikely to be proteolytic degradation products. They discovered glucagon, secretin, vasoactive intestinal polypeptide (VIP), galanin (Gal), neuropeptide Y (NPY), and many other peptides.

The understanding that the neuropeptides are encoded on mRNAs that are first translated into large preprohormones also led to the examination of all the peptide fragments that are formed during the maturation of these precursors. Starting with the case of the proopiomelanocortin (POMC) preprohormone, the discovery that tissue and cell type-specific processing of the peptide precursors occurs in the central and peripheral nervous systems added to our understanding of the complexity and versatility as well as the physiological specificity of the neuropeptidergic signaling.

The next wave of peptide discoveries came when G-protein-coupled receptors (GPCRs) were discovered by homology cloning using the β-adrenergic receptor as a template; because of a large group of these GPCRs whose ligands were not known, the notion of orphan receptors was adopted. Also, a few years later, the Human Genome Project provided a whole set of GPCR sequences without any indication of the nature of their endogenous ligands as well as the sequences of a whole range of putative prepropeptides with their putative dibasic cleavage points.

These orphan receptors (as well as the peptides in search of their receptors) became highly coveted drug targets – based on *in situ* hybridization-supported mapping of their expression patterns – that showed that these are often in areas of the brain that are associated with normal function and with the pathogenesis of higher functions. Some of the orphan receptors indeed have peptide ligands (e.g., nociceptin/orphan FQ (OFQ), melanin-concentrating hormone, and ghrelin), whereas other orphan GPCRs have been shown to bind lipid transmitters such as the endocannbinoids and sphingosine phosphate derivatives or even to act as pH sensors.

Neuropeptide Signaling as Drug Target

The neuropeptidergic transmission is an attractive area for pharmacological intervention for several reasons.

- We have (almost) all but exhausted the receptors for the classical neurotransmitters – the monoamines, glutamate (Glu), γ-aminobutyric acid (GABA), and acetylcholine (ACh) – as drug targets in our efforts to treat diseases such as epilepsy, major depression, and obesity. We have, however, come to realize that neuropeptides coexisting with these classical neurotransmitters have receptors expressed in brain regions identified as important physiologically and pathophysiologically. This suggests that we should target neuropeptide receptors to modify the activity of neurons in these selected brain regions to achieve therapeutic effects.
- The concentration of (most) neuropeptide receptors is much lower than that of receptors for the classical neurotransmitters, and there are seldom large populations of spare receptors. By the same token, the number of cells producing the corresponding neuropeptides may be as low as only a few thousand in the entire brain.
- The neuropeptide receptor type is a GPCR – a class of drug targets traditionally easier to address chemically.
- The synaptic concentrations of the neuropeptide ligand is several orders of magnitude lower than that of the classical transmitters (nanomolar vs. micromolar or even submillimolar); thus, when antagonists are required, the competition with the endogenous ligand is not that difficult to achieve, and when agonists are sought, there are not large numbers of receptors to saturate. Nevertheless, in both the case of agonists and antagonists, high-affinity (K_D in the nanomolar or subnanomolar range) ligands have to be identified.

The synthesis of neuropeptides depends on gene transcription and translation and is subject to multiple posttranslational modification steps; the ready neuropeptide is often loaded into and stored with adenosine triphosphate (ATP) in large dense-core vesicles. These vesicles release their content under high-frequency

stimulation or at bursting activity, which studies of the periphery on norepinephrine (NE)/NPY and ACh/VIP and in the central nervous system (CNS) on ACh/Gal-containing neurons have shown. This frequency-dependent chemical coding permits at low frequency the ACh/galanin-containing neuron in the septum to release only ACh within the ventral hippocampus, whereas at higher frequencies the same neuron also releases Gal in addition to ACh. Gal, then, in multiple ways, can modify the actions of ACh on pre- and postsynaptic targets in the hippocampus. There are therapeutic possibilities hidden in this scenario; when a given transmitter or neuropeptide is only released at higher frequencies or at specific stimulation patterns, it is a very tempting target of pharmacological intervention because such a mechanism would leave normal synaptic function unaffected but would kick in during runaway discharge and trauma-mediated massive release.

The biosynthesis of neuropeptides can be safely modified chronically, as most transgenic animal models with null mutation of neuropeptides demonstrate. These studies show that there is as strong redundancy of the neuropeptide signaling such that a number of other neuropeptides can take over the role of the one not synthesized. This is good in terms of drug safety, but of course also it indicates that not all chronic neuropeptide targeted actions are necessarily robust or remain robust. For example, the hyperpolarizing neuropeptides in the hippocampus that contribute to control of excitability – somatostatin, NPY, Gal, dynorphin, and so on – can take over, in a compensatory manner, one another's roles to some extent in seizure models. Nevertheless, NPY and Gal receptor subtype selective agonists are sought as anticonvulsant agents (Table 1) because subtle levels of receptor modulation may just do the job.

The selective suppression of neuropeptide gene expression with antisense and/or RNA interference will most likely provide us an important entry into the understanding of neuropeptide (receptor) pharmacology.

Similarly, we may envisage inhibiting the enzymes involved in the processing or transcriptional modification of neuropeptides. When this occurs via a specific enzyme that has one or few neuropeptide substrates, such as the biosynthetic angiotensin-converting enzyme (ACE), successful drugs can result from this approach such as the ACE inhibitors. However, when the enzyme such as that involved in the oxidation of the C-terminal glycine to generate C-terminal amide (necessary for the activity of many neuropeptides) is involved, then no pharmacological selectivity can be expected from the inhibition of this very target.

To slow down, decrease, or altogether stop production and maturation of a neuropeptide seems to be difficult, with only few exceptions (e.g., ACE inhibitors). Furthermore, most neuropeptides bind to several receptor subtypes on different cell types, mediating different actions. Thus, inhibiting the synthesis of neuropeptide agonists has to be evaluated in comparison to the generation of receptor subtype-selective antagonists to the GPCR-type neuropeptide receptors – a process the industry is very familiar with.

Neuropeptide Receptors as Drug Targets

The neuropeptide receptors studied so far belong to the GPCR type and transmembrane domain receptors that, when occupied with agonist, can activate intracellular G-protein-mediated signaling. GPCRs are favorite targets of the pharmaceutical industry, which has produced safe and chronically useful drugs such as antagonists to β-adrenergic or angiotensin receptor for the treatment of hypertension and chronic heart failure, H1 or H2 histaminergic antagonists for allergy or digestive ulcer treatment, and D2 dopaminergic antagonists as antipsychotics. Agonists acting at GPCRs such as serotonin (5-HT_{1A}) receptor agonists anxiolytic and antidepressant drugs, 5-HT_{1B} agonists to treat migraine, or μ- and κ-opiate receptor agonists such as morphine and synthetic analogs for the treatment of several types of pain.

The experience with GPCR antagonists has shown that chronic treatment causes only moderate supersensitivity and that, similarly, treatment with agonists has resulted in moderate desensitization, if any, in most systems. The most notable exceptions are the opiate agonist and muscarinic cholinergic agonists, which caused rapid desensitization. In the fastest documented desensitization (based on rapid internalization of the GPCR), neurokinin (NK)1 was shown, on exposure to the neuropeptide substance P, to rapid internalized within a few minutes. However, the somatostatin receptor agonists used in the treatment of acromegaly, for example, do not cause substantial desensitization of somatostatin receptor, even when constant plasma levels are maintained over months with the current slow release formulations and although *in vitro* data suggest that desensitization occurs under such conditions.

Validation of Neuropeptide Receptors as Drug Targets

The past decade saw the extensive use of molecular, biological, and transgenic techniques to validate neuropeptide receptors as drug targets for the treatment of different diseases. This is a very important development because the time it takes to synthesize

Table 1 Neuropeptides and their receptor subtypes and the links to physiological and pathophysiological actions[a]

Receptor	*Receptor subtypes*	*Endogenous ligands*	*Link to disease, physiological actions*	*Comments*
Angiotensin receptors	AT1, AT2	Angiotensin II, III	Hypertension (AT1), cardiac hypertrophy, congestive heart failure	
Bombesin receptors	BB1, BB2 (GRPR), bb3	Neuromedin B, GRP, GRP(18–27) (aka neuromedin C)	Anxiety (GRPR), cancer, circadian rhythms	
Bradykinin receptors	BK1, BK2	Bradykinin, Lys-bradykinin (kallidin)	Pain, hyperalgesia	
Calcitonin gene-related peptide receptors	Calcitonin, CGRP1, AMY1, -2, -3, AM1, -2	α- and β-CGRP, AMY (aka islet amyloid peptide), AM, AM2/ intermedin.	Hypocalcemia (calcitonin)	Need RAMP 1, 2, 3; AMY 1, 2, 3 = calcitonin receptor + RAMP 1, 2, 3, respectively; CGRP1, AM1, AM2 = calcitonin gene-related receptor + RAMP 1, 2, 3, respectively
Cholecystokinin and gastrin receptors	CCK-A, CCK-B	CCK4, CCK8, gastrin I, gastrin II	Anxiety, panic	
Corticotropin-releasing hormone receptors	CRH1, CRH2	CRH, urocortin 1, 2, and 3	Anxiety, depression, stress	Species differences in splice variants of CRH2
Endothelin receptors	Endothelin 1, -2, -3	ET-A, ET-B	Congestive heart failure, pulmonary hypertension, stroke, kidney failure, hypertension, angioplasty-induced restenosis, asthma, chronic obstructive pulmonary disease, pain, cancer	
Galanin receptors	Gal1,-2,-3	Galanin, GALP	Epilepsy, seizure, anxiety, depression, pain, alcoholism	
Ghrelin receptor	GHSR1	Ghrelin	Obesity, feeding disorders, depression	Obestatin is derived from preproghrelin and was described as acting on GPR39; however, these data have been disputed
Leptin receptor	Leptin receptor	Leptin	Obesity, depression	The receptor is a single membrane-spanning receptor most related to the gp130 signal-transducing component of the growth hormone, IL-6, G-CSF, and LIF receptors, which are part of the class I cytokine receptor family
Melanocortin receptors	MC1, -2, -3, -4, -5	ACTH, AGRP, α-MSH, γ-MSH, γ2-MSH;	Endocrine, metabolic, and feeding disorders; obesity	MC6 has not been cloned; agouti and AGRP are two of the very few endogenous antagonists; haploinsufficiency of the MC4-R is responsible for at least 5% of severe human obesity
Melanin-concentrating hormone receptors	MCH1, MCH2	MCH	Obesity	MCH2 may be nonfunctional in rodents
Neuropeptide S receptor	NPSR	NPS	Anxiety, depression, asthma, COPD, vigilance	SNPs within the NPSR have been associated with an increased prevalance of asthma in a variety of populations

Continued

Table 1 Continued

Receptor	*Receptor subtypes*	*Endogenous ligands*	*Link to disease, physiological actions*	*Comments*
Neuropeptide Y receptors	NPY1, -2, -4, -5	NPY, PYY, and PP	Epilepsy, pain, anxiety, depression, obesity	NPY6 is not a functional receptor; NPY3 has been dropped as a true NPY receptor
Neurotensin receptors	NTS1, NTS2	Neurotensin, neuromedin N	Schizophrenia	A third neurotensin receptor, NTS3, was cloned; it is a 100-kDa protein with a single transmembrane domain (i.e., not a GPCR) and is identical to the previously cloned protein sortilin; the status of NT3/sortilin/gp95 as a true neurotensin receptor remains to be established
Neurotropin receptors	TrkA, TrkB, TrkC, p75	BDNF, LNGFR, NGF, NT-4/5	Pain, hyperalgesia, neurodegeneration, depression, schizophrenia	
Opioid receptors	μ, mu, or MOP; δ, delta, or DOP; κ, kappa, or KOP; ORL1	β-Endorphin, dynorphin A, endomorphin 1 and 2, [Leu5]-enkephalin, [Met5]-enkephalin, nociceptin/orphanin FQ	Pain, constipation, autism	
Orexin receptors	OX1, OX2	Orexin A and B (aka hypocretin 1 and 2)	Sleep regulation, obesity	Disruption of prepro-peptide expression or mutations in the gene encoding OX2 receptors result in a narcoleptic phenotype
Somatostatin receptors	sst1, -2, -3, -4, -5	Somatostatin 14 and 28, cortistatin 17 and 29	Acromegaly, GEP tumors, epilepsy, pain, diarrhea, sleep, anxiety, depression, memory	Somatostatins and cortistatins are products of different genes; there no receptor specific for cortistatin
Tachykinin receptors	NK1, -2, -3	SP, NK-A, NK-B	Anxiety, depression, schizophrenia, pain, migraine, COPD, asthma, emesis, IBD, urinary incontinence	
Vasoactive intestinal peptide, pituitary adenylyl cyclase-activating peptide, and related peptide receptors	VPAC1, VPAC2, PAC1	VIP, PACAP(1–27) and (1–38), PHI	Inflammation, injury, circadian rhythms	
Vasopressin and oxytocin receptors	V1a, V1b, V2, OT	Vasopressin and oxytocin	Learning, memory, anxiety, depression, autism, preterm labor, diabetes insipidus, congestive heart failure	

[a]We have not listed following peptide receptors: adiponectin, anaphylatoxin and chemotactic peptide, apelin, chemokine, cytokine family, glucagon and secretin family, glycoprotein hormones, gonadotropin-releasing hormone (GnRH), motilin, neuromedin U, neuropeptides B and W, proteinase-activated, relaxin family, resolvin, thyrotropin-releasing hormone (TRH), and urotensin II.

ACTH, adrenocorticotropic hormone; AGRP, agouti-related protein; AM, adrenomedullin; AMY, amylin; BDNF, brain-derived neurotrophic factor, CCK, cholecystokinin; CGRP, calcitonin gene-related peptide; COPD, chronic obstructive pulmonary disease; CRH, corticotropin-releasing hormone; GALP, galanin-like peptide; G-CSF, granulocyte-colony stimulating factor; GEP, gastroenteropancreatic; GHSR, growth hormone secretagogue receptor; GRP, gastrin-related peptide; IBD, inflammatory bowel disease; IL, interluekin; LIF, leukemia inhibitory factor; LNGFR, low-affinity nerve growth factor receptor; MCH, melanin-concentrating hormone; MSH, melanocyte-stimulating hormone; NGF, nerve growth factor; NK, neurokinin; NT, neurotropin; ORL, opiate receptor-like; PACAP, pituitary adenylyl cyclase-activating peptide; PHI, peptide histidine isoleucine amide; PP, pancreatic polypeptide; PYY, peptide YY; RAMP, receptor activity-modifying proteins; SNP, single nucleotide polymorphism; SP, substance P; VIP, vasoactive intestinal peptide.

adequate agonists to neuropeptide receptors that are able to cross the blood–brain barrier and that are peptidase resistant is measured in years; therefore, the increasing number of examples of transgenic animals that overexpress the neuropeptide in selectively chosen sets of neurons circumvents these problems in studying the pharmacological effects of chronic neuropeptide agonists in disease models. Thus, it was shown that the neuronal overexpression of Gal suppresses seizures in a variety of seizure models. An even faster method is the localized viral vector-mediated expression of a neuropeptide, which informs not only about the effects of an agonist but also about the site of action.

Similarly, null mutations of neuropepitdes have been informative about what antagonists will do in different disease states. The recent surge in the use of antisense oligonucleotides and of small interfering RNA (siRNA) experiments further focuses the answer in terms of the pharmacological effects of an antagonist and the site of action.

The use of receptor knockouts, as in whole pharmaceutical development, has been useful in deciding which receptor subtype is responsible for which effects of the neuropeptide. The usual caveat about compensatory changes occurring in transgenic animals is of particular relevance in the study of neuropeptide systems that are plastic and redundant. Nevertheless, these methods that make use of the fact that neuropeptides, unlike classical transmitters, are gene products and that thus their production can be manipulated by null mutation or overexepression has been very informative and constitutes an important part of the validation process of neuropeptide receptor as drug target.

The Problems of Finding Exogenous Ligands to Neuropeptide Receptors

Agonists

Table 2 summarizes the pharmaceutical development state of numerous neuropeptide receptor-targeting drugs, showing that we seek agonists for several neuropeptide GPCRs. These receptors include: MC4, NPY1, sst2, Gal1, Gal2, OX1, and the search for better μ- and κ-opioid receptor ligands continues.

The problems of finding neuropeptide receptor agonists relate to the nature of the endogenous ligands, which are often large molecules of 20–40 amino acids with possibly large corresponding binding pockets within the receptor. This offers the advantage of high-affinity and high-selectivity binding for the endogenous neuropeptides but makes the discovery of small nonpeptide agonists an almost impossible task. Obviously, the solid phase synthesis of peptides or peptide mimics or the recombinant production of peptides (when large quantities are needed) is not a problem today, but the peptide drugs have a number of inherent problems that make them not well suited as drugs.

1. They have very short half-lives because of peptidolytic degradation by both nonspecific and specific peptidases, although stability can be significantly improved by cyclization.
2. Many of the most sought-after actions for neuropeptides are at receptors expressed in the brain (or in the peripheral nervous system), but the peptide ligands do not cross the blood–brain barrier passively and selective transporters have not been reported (with a very few exceptions, e.g., secretin, thyrotropin-releasing hormone (TRH)) and, in general, oral bioavailability is hardly achievable.
3. The endogenous peptides are pan-ligands to multiple subtypes of receptors, whereas we would prefer to aim at pharmacological actions with receptor subtype selectivity (although such approaches may change again as the number of failed attempts using very selective ligands accumulate).

The synthesis of truncated and L Ala-substituted analogs (called Ala scan) of neuropeptides has shown that the binding epitopes can be well separated from one another in space and thus that it is not easy to reach/mimic these on a small synthetic nonpeptide molecule. Although strong progress has been made in the synthesis of scaffolds such as the β-sheet scaffolds and α-helix mimetics, the design of neuropeptide receptor agonists is not a fully realized task today. The introduction of pseudo-inverse-peptides (e.g., β or β/γ peptides) has addressed some of the peptidolytic stability issues, but it has not fully solved them. Nevertheless, peptidomimetic ligands, particularly for cyclic peptides, have been introduced successfully into clinical practice (cf. somatotstain receptor agonists), and new formulations have been produced that achieve very constant exposure over long periods (weeks).

Nonpeptide agonists to neuropeptide receptors The bright example of morphine keeps us confident that wonderfully important nonpeptidic agonists to neuropeptide receptors can be found. The problem is that drug companies do not work on the timescale of evolution and many have largely stopped or slowed the use of natural product libraries. The other issues, as already listed, are that the large neuropeptides have discontinuous binding epitopes and that, with these, they may also bind to the extracellular domains of the neuropeptide receptor in a polar milieu.

Table 2 Preclinical and clinical development and use of selected neuropeptide receptor ligands in 2006

Receptor	*Receptor subtypes*	*Indications*	*Clinical stage*	*Drugs and drug candidates*
Angiotensin receptors	AT1, AT2	AT1: hypertension, CHF, diabetic retinopathy, anxiety, glaucoma	Marketed AT1 antagonists	Valsartan, Losartan (AT1) launched in hypertension, CHF; ongoing studies in diabetic retinopathy, anxiety, glaucoma
Bombesin receptors	BB1, BB2 (GRPR), bb3	BB2: anxiety, depression BB3: cancer, cardiovascular disease, anxiety	Preclinical	
Bradykinin receptors	BK1, BK2	Pain	BK2 antagonists are developed for chronic pain; no positive data	
Calcitonin gene-related peptide receptors	Calcitonin, CGRP1, AMY1, -2, -3, AM1, -2	Hypocalcemia	Various calcitonins are used for the treatment of osteoporosis	Salmon calcitonin
Cholecystokinin and gastrin receptors	CCK-A, CCK-B	CCK-A: pain, eating disorders, antitumor, small-cell lung cancer, gut motility, schizophrenia CCK-B: anxiety, panic disorder, gastric secretion, sleep	CCK-B antagonists have been tested with negative outcome; however, CCK4 is used in patients/ volunteers to induce panic attacks	CCK-A: Lintitript [SR 27897] for bowel motility and gastric emptying, and biliary tract disease (phase II) CCK-B: GW-150013 in phase II for GAD, but stopped in sleep
Corticotropin-releasing hormone receptors	CRH1, CRH2	CRH1: antagonists developed in pain, depression, epilepsy, Alzheimer's disease, cognition, eating disorders, IBS CRH2: anxiety, eating disorders.	Numerous CRH1 antagonists are being tested in GAD/ depression, one open clinical trial with positive results in depression Combined CRH1/2 or combined CRH1 with SSRIs are considered	CRH1: BMS-562086, SB 723620, NBI-458; NBI-31199; NBI-35583; SSR-125543; also trials of CRH1 with SSRIs CRH1/2: JNJ-19567470 in phase I for anxiety and depression
Endothelin receptors	ET1, ET2, ET3	Pulmonary hypertension	Pulmonary hypertension is a recognized indication for ET antagonist; other indications did not meet proof of efficacy	Bosentan is marketed for pulmonary hypertension
Galanin receptors	Gal1, Gal2, Gal3	Depression, anxiety, pain, epilepsy, Alzheimer's	Gal3 antagonist developed for depression Galanin in open trial had antidepressant effects	Gal3: HT-2157 in depression
Ghrelin receptor	GHSR1	Obesity	A clinical trial with an ghrelin antibody has just been reported to be negative	Cytos ghrelin antibody
Leptin receptor	Leptin receptor	Obesity	No positive human data in obesity so far	
Melanocortin receptors	MC1-5	Mainly obesity MC4: obesity, erectile dysfunction, cachexia, pain, depression, anxiety	Haploinsufficiency of the MC4-R is responsible for at least 5% of severe human obesity	MC4: MCL-0129, preclinical
Melanin-concentrating hormone receptors	MCH1, MCH2	Obesity, anxiety, depression	MCH1 antagonists undergo clinical trials in obesity and possibly anxiety and depression, phase I/II	MCH1: SB-568849 obesity and anxiety, phase I/II; SNAP-7941 preclinical
Neuropeptide S receptor	NPS	Anxiety, sleep	No ligands	
Neuropeptide Y receptors	NPY1, 2, 4, 5	Obesity	Merck reported a very extensive NPY5 antagonists study that is negative, although receptor occupancy was carefully monitored	Ongoing phase II/III

Continued

Table 2 Continued

Receptor	*Receptor subtypes*	*Indications*	*Clinical stage*	*Drugs and drug candidates*
Neurotensin receptors	NTS1, NTS2, NTS3	Schizophrenia	Positive data with NTS3 agonist in schizophrenia phase II	NT-69-L, positive in schizophrenia
Neurotropin receptors	TrkA, TrkB, TrkC, p75			
Opioid receptors	μ, mu, or MOP; δ, delta, or DOP, κ, kappa or KOP, ORL1	μ: pain, diarrhea, anxiety, autism, drug abuse δ: pain, anxiety, depression, cardioprotection κ: pain ORL1: anxiety	Opioid agonists: pain, anxiety, drug abuse Opioid antagonists: opiate poisoning, diarrhea RO 646198 [ipidant] is a nonpeptide orphanin FQ/nociceptin ligand as a potential anxiolytic	μ agonists: morphine, codeine, methadone, fentanyl, and derivatives mixed agonists/ antagonists: buprenorphine, butorphanol, nalbuphine, pentazocine naloxone is a nonselective antagonist; AstraZeneca is developing AZD-2327, a modulator of the enkephalinergic receptor, in anxiety and depression, phase I
Orexin receptors	OX1, OX2	Feeding disorders sleep disorders, narcolepsy	OX1 agonists: feeding disorders, OX2 antagonists: sleep disorders, narcolepsy	Mixed (?) orexin antagonist, phase II (Actelion), sleep disorders
Somatostatin receptors	sst1, sst2, sst3, sst4, sst5	Acromegaly, gastroenteropancreatic tumors, GIT disorders (diarrhea), tumor imaging. Universal agonists: similar indications, transplantation, diabetes, neuropathies, oncology (restenosis), Cushing syndrome.	sst2/sst5 agonists: are marketed for these clinical indication; radioabeled sst2/sst5 agonists for imaging in various cancers; various radiolabeled somatostatin analogs for cancer treatment; universal agonist undergoing clinical trials for the indications listed	sst2/sst5 agonists: octreotide, lanreotide, and vapreotide; [^{111}In] octreoscan; various radiolabeled somatostatin analogs for cancer treatment; SOM230 and other universal agonists
Tachykinin receptors	NK1, NK2, NK3	NK1: pain, migraine, anxiety, GAD, depression, emesis, IBS, overactive bladder, chemotherapy-induced nausea and vomiting and postoperative nausea and vomiting NK2: IBS, COPD, asthma, lower urinary tract disorders, anxiety NK3: schizophrenia, COPD, urinary incontinence	NK1 antagonists registered in emesis, the other indications were eventually dropped; although positive data had been obtained in depression, they were not replicated NK2 and NK3 antagonists still under clinical investigation, phase I/II Combinations: vestipitant (GW-597599), an NK1 receptor antagonist, and paroxetine for depression, phase II	NK1: Aprepitant [MK 869] launched for the treatment of chemotherapy-induced emesis, phase II in GAD, stopped in depression Casopitant (GW-679769) and L 759274 in GAD and depression, phase II/III in other indications NK2: saredutant (SR-48968) and SSR-241586 for IBS, COPD, asthma, lower urinary tract disorders, anxiety, depression NK3: Talnetant (SB-223412) schizophrenia, COPD, urinary incontinence, phase II; SSR 146977 anxiety, depression, psychotic disorders, phase I

Continued

Table 2 Continued

Receptor	*Receptor subtypes*	*Indications*	*Clinical stage*	*Drugs and drug candidates*
Vasoactive intestinal peptide, pituitary adenylyl cyclase activating peptide, and related peptide receptors	VPAC1, VPAC2, PAC1	Inflammation, neurodegeneration, sleep	Under investigation	
Vasopressin and oxytocin receptors	V1a, V1b, V2, OT	V1a, V1b: depression V2: diabetes insipidous OT: anxiety, schizophrenia	V1a: depression and menstrual pain; clinical studies under way V1b: positive data reported in depression	V1a: SRX251, depression, premenstrual syndrome V1b: SSR-149415 is in phase II for anxiety and depression

CHF, congestive heart failure; GAD, generalized anxiety disorder; GHSR, growth hormone secretagogue receptor; GIT, gastrointestinal tract; IBS, irritable bowel syndrome; ORL, opiate receptor-like; SSRI, selective serotonin reuptake inhibitors.

The achievement of agonist action by inhibiting the degradation of the neuropeptides has been tested using the synthesis of enkephalinase inhibitors that show effects in pain models, but the robustness of such effects is strongly dependent on the release of the endogenous peptide which is faltering because the peptide stores can be exhausted easier than the stores of classical neurotransmitters that are synthesized at the presynaptic terminal enzymatically. The transcription, translation, and posttranslational modification-dependent peptide synthesis and release can be outstripped by demand.

The other major problem with neuropeptidase inhibitors is that not only specific enzymes such as enkephalinase, but also nonspecific peptidases and dipeptidases can degrade the peptide of interest, in particular if its concentrations rise as result of the inhibition of the selective degrading pathway.

Nonpeptide Antagonists to Neuropeptide Receptors

High-throughput screening (HTS) for neuropeptide receptor antagonists depends on our ability to engineer reporter systems using promiscuous G-proteins to couple almost any neuropeptide receptor to phospholipase C (PLC) activation and Ca^{2+} release from intracellular store or to the production of cyclic adenosine monophosphate (cAMP) and to cAMP responsive element-binding protein (CREB)-dependent transcription of luciferase. Such advances and the availability of additional reporter systems led to the screening of large chemical libraries for ligands to neuropeptide receptors.

Thus, nonpeptide antagonists and allosteric ligands of agonist and antagonist type have been found for many receptors that were screened, although some screens may have been unsuccessful, in spite of the large size of the libraries used. Of course, the chemical space that has been explored so far is very limited, especially given the nature of some of the targets investigated.

However, even when the HTS has identified hits, the subsequent development of drugs by pharmaceutical chemists follows the same slow and difficult path of engineering selectivity for receptor subtypes and for other relevant targets, proper safety, and determination of pharmacokinetic properties to reach the clinical candidate.

Lack of Structural Information on the G-Protein-Coupled Receptors

The lack of x-ray structure data for the GPCRs and for the GPCR–ligand complexes is a major impediment to the rational design of neuropeptide receptor ligands. The models based on rhodopsin have been somewhat useful in the design of monoaminergic GPCR ligands, but the models based on rhodopsin are far from perfect and have very little value when it comes to large neuropeptides. The long-awaited breakthrough in GPCR X-ray structure should make drug development for neuropeptide receptors easier, especially for straight receptor agonists.

Dynamics of Neuropeptide–G-Protein-Coupled Receptor Functions in Drug Development

The GPCR dynamics that can be used in achieving pharmacological effects or modifying such effects –are:

1. Internalization of GPCRs: modification by β-arrestins and receptor-specific kinases.
2. Dimerization of GPCRs: homodimerization or heterodimerization of receptors of the same, or even

different, families; heteroligomerization with receptor activity-modifying proteins (RAMPs), not forgetting the huge reservoir of activity-modifying proteins that seem to interact with almost every single GPCR (called GPCR-interacting proteins (GIPS), e.g., calmoduline, multiple PDZ domain containing protein (MUPP)1, postsynaptic density protein (PSD-)95, channel-interacting PDZ domain protein (CIPP), autocamtide-2-related inhibitory peptide (ARIP-)1, calmodulin-dependent protein kinase (CaMK), Src-homology domain 3 (SH3), synapse-associated protein (SAP), Homer, and 14-3-3).

The various posttranslational modifications of neuropeptide receptors include glycosylation, myristilation, and palmitoylation, each of which is important for the membrane targeting and function of the GPCR, but (apparently) not specifically relevant to neuropeptide receptors. The function-dependent agonist-dependent modification via receptor kinases such as β-arrestin is also very important, but they are not selective targets to achieve the modulation of neuropeptidergic action.

More selectivity can probably be achieved if the homodimerization of neuropeptide GPCRs (as has been reported for several receptors) can be targeted in such a manner that it affects signaling. This has so far been shown with model peptides that are not druglike but that suffice as a proof of the principle that neuropeptide action can be affected by blocking the dimerization of receptors such as μ-opiate, GalR1 or GalR2, or NPY2. Heterodimerization of GPCRs, in particular, among opiate receptors has been demonstrated to provide additional targets.

The most successful and potentially clinically useful example pertains to the blockade of the interaction of the subunits calcitonin receptor-like receptor (CLR) and RAMP of the calcitonin gene-related peptide (CGRP) receptor. The benzimidazole (BI) compound that binds to RAMP and acts as a CGRP antagonist was shown to be effective in migraine and in pain models, although the compound has bioavailability issues.

Allosteric ligands or modulators at neuropeptide receptors are found most easily when HTS is carried out in the agonist mode. The changes in the activity of the receptor-coupled reporter systems can identify allosteric ligands that can be developed into safe and effective drugs.

Perspective

The localization of neuropeptide receptor in important brain areas controlling many functions makes continued interest in neuropeptide receptor ligands certain. The increasing number of genetic studies showing linkage between disorders and neuropeptidergic signaling further emphasizes the view that we will see many neuropeptide receptor ligands as drugs. Finally, we have exhausted the possibilities of using the classical neurotransmitter pharmacology to achieve clinical effects such as control of excitability and treatment of major depression, obesity, and pain, and we are turning to the neuropeptide receptors as drug targets. For these reasons, we expect that many new drugs will target neuropeptide receptors, although some failures are also expected, as with the more classical GPCR ligands (e.g., NPY5 in obesity or NK1 in pain; Table 2).

Summary

The number of genuine neuropeptides and their corresponding receptors has increased steadily over the last 35 years through isolation of the peptides, through deorphanization of several GPCRs, and through the completion the Human Genome Project. Neuropeptides and their receptors show distinct regionally specific distributions in the central and peripheral nervous systems. Neuropeptide signaling is somewhat more distinct spatially than signaling with classical low-molecular-weight neurotransmitters that are more widely expressed, and therefore we assume that drugs acting at neuropeptide receptors will have more specific pharmacological actions with fewer side effects than drugs acting on glutamatergic, GABAergic, monoaminergic, or cholinergic systems. Neuropeptide receptors, which often have multiple subtypes and also multiple splice variants, belong almost exclusively to the GPCR family (also known as seven-transmembrane receptors (7TM)), a favorite class of drug targets in the pharmaceutical industry. Most neuropeptides are co-stored and co-released with classic neurotransmitters, albeit often only at higher frequencies of stimulation or at bursting activity, restricting the neuropeptide signaling to specific circumstances and increasing the expectation that neuropeptidergic drugs will have fewer side effects. Neuropeptides have been shown to possess a wide spectrum of functional roles, ranging from neurohormone to neurotransmitter to growth factor. Much evidence indicates that neuropeptides play crucial and varied roles when the nervous system is challenged by stress, injury, drug abuse, and other types of derangements, such as neuropsychiatric disorders with genetic, epigenetic, and/or environmental components. The ever-increasing number of true neuropeptides and their cognate receptors provide opportunities for us to identify novel drug targets for the treatment of both central and peripheral nervous

system disorders. In fact, receptor subtype-selective antagonists and agonists have been developed, mainly within the framework of the pharmaceutical industry, as illustrated by the success of somatostatin agonists and angiotensin and endothelin antagonists and also by the expected clinical applications of NK-1/2/3 (substance P) receptor antagonists, CRH, vasopressin, NPY, neurotensin, antagonists or neuropeptide receptor modulators which are reported to have efficacy in preclinical or clinical models of pain and neuropsychiatric diseases, such as migraine, chronic/neuropathic pain, anxiety, depression, and schizophrenia.

See also: Anxiety Disorders; Anxiety: Drug Therapy; Depression and the Brain; Epilepsy; Invertebrate Neurohormone GPCRs.

Further Reading

Bartfai T, Iverfeldt K, Fisone G, and Serfozo P (1988) Regulation of the release of coexisting neurotransmitters. *Annual Review of Pharmacology and Toxicology* 28: 285–310.

Belzung C, Yalcin I, Griebel G, Surget A, and Leman S (2006) Neuropeptides in psychiatric diseases: An overview with a particular focus on depression and anxiety disorders. *CNS & Neurological Disorders – Drug Targets* 5: 135–145.

Butler AA (2006) The melanocortin system and energy balance. *Peptides* 27: 281–290.

Cammalleri M, Cervia D, Langenegger D, et al. (2004) Somatostatin receptors differentially affect spontaneous epileptiform activity in mouse hippocampal slices. *European Journal of Neuroscience* 20: 2711–2721.

Carvajal C, Dumont Y, and Quirion R (2006) Neuropeptide Y: Role in emotion and alcohol dependence. *CNS & Neurological Disorders – Drug Targets* 5: 181–195.

Greengard P (2001) The neurobiology of slow synaptic transmission. *Science* 294: 1024–1030.

Hokfelt T, Bartfai T, and Bloom F (2003) Neuropeptides: Opportunities for drug discovery. *Lancet Neurology* 2: 463–472.

Keck ME, Ohl F, Holsboer F, and Muller MB (2005) Listening to mutant mice: A spotlight on the role of CRF/CRF receptor systems in affective disorders. *Neuroscience and Biobehavioral Reviews* 29: 867–889.

Kokaia M, Holmberg K, Nanobashvili A, et al. (2001) Suppressed kindling epileptogenesis in mice with ectopic overexpression of galanin. *Proceedings of the National Academy of Sciences of the United States of America* 98: 14006–14011.

Landgraf R (2005) Neuropeptides in anxiety modulation. *Handbook of Experimental Pharmacology* 169: 335–369.

Lin EJ, Richichi C, Young D, Baer K, Vezzani A, and During MJ (2003) Recombinant AAV-mediated expression of galanin in rat hippocampus suppresses seizure development. *European Journal of Neuroscience* 18: 2087–2092.

Nunn C, Langenegger D, Hurth K, Schmidt K, Fehlmann D, and Hoyer D (2003) Agonist properties of putative small-molecule somatostatin SST2 receptor-selective antagonists. *European Journal of Pharmacology* 465: 211–218.

Schwartz TW and Holst B (2006) Ago-allosteric modulation and other types of allostery in dimeric 7TM receptors. *Journal of Receptor and Signal Transduction Research* 26: 107–128.

Seebach D, Hook DF, and Glattli A (2006) Helices and other secondary structures of beta- and gamma-peptides. *Biopolymers* 84: 23–37.

Neuropeptide Release

F Bergquist and M Ludwig, University of Edinburgh, Edinburgh, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A main criterion for interneuronal chemical messengers, including neurotransmitters, neuromodulators, and neurohormones, is that their release be regulated and conditioned by neuronal activity. This article discusses the regulated release of neuropeptides, a large class of neuroactive messengers present in both the central nervous system (CNS) and peripheral nervous system. Other characteristics of neuropeptides that are vital for their function as interneuronal messengers, such as receptor effects, synthesis, storage, and breakdown/inactivation are also highlighted here because it is important to understand how neuropeptide release differs from classical neurotransmitter release.

The neuropeptide messenger family is large, and the list of members is still increasing steadily some 70 years after Gaddum and von Euler's discovery of substance P. More than 100 neuropeptides have established messenger properties. For most of these, the mechanisms of release have not been studied in detail, but there seem to be some common themes regarding activity-dependent neuropeptide release that allow some generalizations. The following section discusses some characteristics that are believed to be common to the release of most neuropeptides.

Neuropeptide Release Compared to Classical Neurotransmitter Release

Synthesis

Peptides are synthesized as larger precursors in the cell body but not in axons and axon terminals. The precursors are stored in vesicles where they are broken down by convertases into active peptides, and the vesicles are transported to the release sites. The supply of neuropeptides can, therefore, not be replenished quickly by synthesis at the release site, as is the case with conventional neurotransmitters. Thus there are mechanisms that economize the release of peptidergic vesicles, and transport mechanisms that can deal with changes in demand at particular release sites.

Vesicle Types and Release Sites

Releasable neuropeptides are always stored in and released by exocytosis from large dense-core vesicles (LDCVs). LDCVs differ from the other main class of neurotransmitter vesicles, small synaptic vesicles (SSVs) not only by size and physical density but also by the different composition of membrane proteins involved in the release processes and by a different cellular localization (Figure 1). SSVs are the predominant vehicle for classical neurotransmitter release, although amino acids and acetylcholine are also released from dense-core vesicles to some extent, together with peptides and sometimes amines. As indicated by the name, SSVs are mostly found at synapses, and release from SSVs is restricted to these highly specialized membrane structures (active zones). In contrast, LDCVs are broadly distributed and can release their contents from any part of the neuron, including the cell body and the dendrites. Release of classical transmitters from SSVs is thus restricted spatially, and the action of a classical transmitter is further constrained to the release sites by a highly efficient termination-of-release signal by uptake mechanisms and a fast breakdown. Peptidergic signaling, on the other hand, mostly takes place outside synaptic specializations (Figure 1).

Calcium Dependent Release

All activity-dependent peptide release appears to be triggered by an increase in intracellular Ca^{2+} concentration. The release machinery that mediates neuropeptide release has a higher affinity for Ca^{2+} than that involved in release of classical transmitters. This is necessary because, unlike SSVs, LDCVs are not positioned close to Ca^{2+} channels, and the Ca^{2+} that enters through these channels in response to depolarization is rapidly buffered on its way into the cell. The Ca^{2+} sensor on SSVs is synaptotagmin I, but the identity of the Ca^{2+} sensor on LCDVs is not yet known. When neuropeptides are released from dendrites, the release can be triggered not only by influx of Ca^{2+} via channels in the cell membrane but also by the mobilization of Ca^{2+} from intracellular stores.

Timescale of Neuropeptide Release and the Type of Stimulus Required

The characteristic Ca^{2+} dependence of LDCV fusion influences the speed at which neuropeptides are released. Classical neurotransmitters in SSVs are typically released with latencies of 0.3–1 ms after a Ca^{2+} trigger, whereas the latencies of LDCV release range from 30 to 2000 ms. The long LDCV latencies arise partly from the wider distances between LDCV Ca^{2+} sensors and the sources of Ca^{2+} entry into the cytosol,

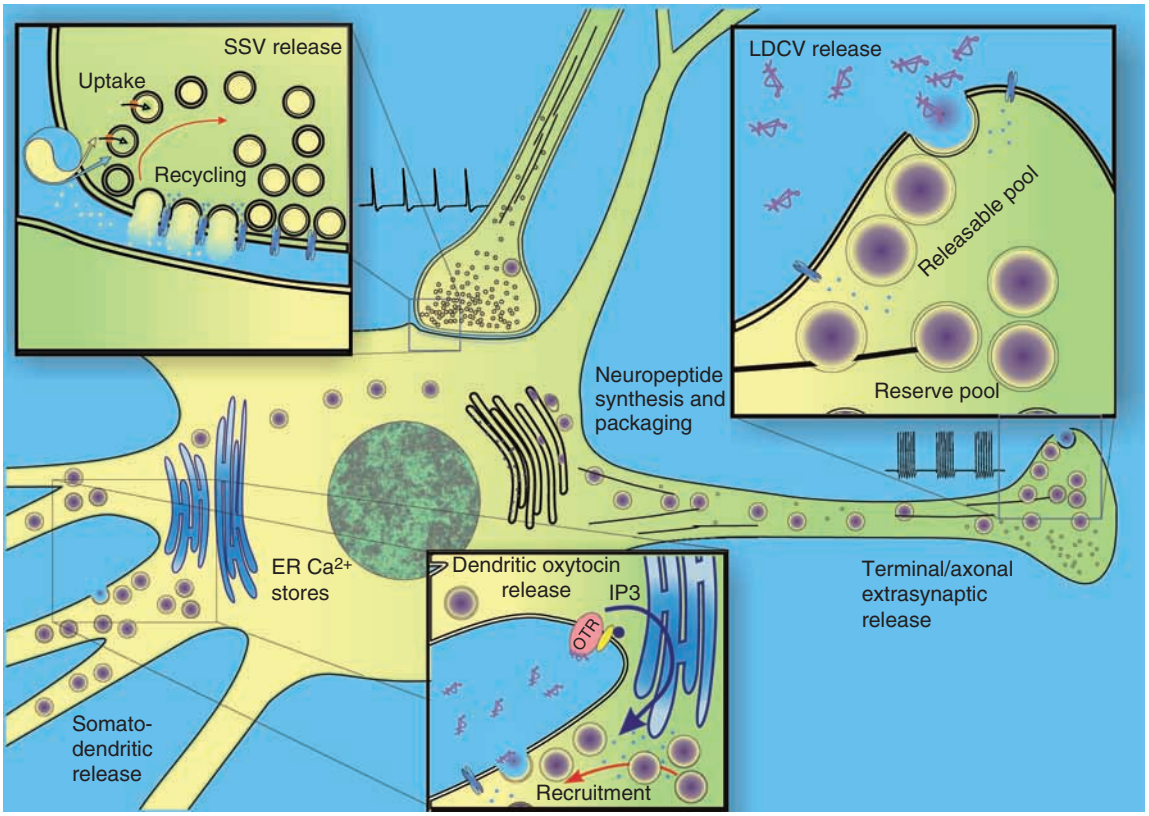

Figure 1 Characteristic features of the small synaptic vesicle (SSV) and large dense-core vesicle (LDCV) release mechanisms. The SSV exocytosis (upper-left insert) can be evoked by low-frequency firing or single action potentials. The vast majority of docked SSVs are primed and therefore readily releasable. When an action potential reaches the axon terminal, voltage-sensitive Ca^{2+} channels open and increase the intracellular Ca^{2+} concentration close to the cell membrane to high micromolar levels that trigger SSV release (intracellular Ca^{2+} marked as blue dots). SSVs are found in close apposition to the Ca^{2+} channels and are released with short latencies after the arrival of an action potential. The SSVs can either be integrated into the terminal membrane for later recycling or release their contents by a kiss-and-run mechanism. SSV neurotransmitters are released into a narrow synaptic cleft, where they act on the postsynaptic receptors. Efficient uptake and/or breakdown processes rapidly terminate the signal, and very few transmitter molecules escape the synaptic cleft to act at a distance from the release site. SSVs are refilled by local neurotransmitter synthesis and/or uptake of already released neurotransmitters. The LDCV release of neuropeptides (upper-right insert) can take place from any part of the neuron. LDCVs typically require high-frequency or burst-patterned firing to be released. Compared to SSVs, a much smaller fraction of the docked LDCVs are primed, and only a small fraction of the vesicles present at a release site are available for release. There is no close relationship between Ca^{2+} channels and LDCVs, and LDCVs have a more sensitive Ca^{2+} sensor that responds to the lower Ca^{2+} concentrations that are attained at a distance from the Ca^{2+} source. The release latencies are long, from tens of milliseconds to seconds. Neuropeptides are released outside synaptic specializations and are not recycled into the releasing neuron but are broken down slowly in the extracellular space. Often the breakdown results in one or more active metabolites that require further breakdown to be inactivated. Consequently neuropeptides have a long range of action. Synthesis and packaging of neuropeptides only take place in the cell body, and all LDCVs need to be transported from there to the release sites. LDCVs in transport are captured into the releasable pool of vesicles or the reserve pool at active release sites. The capture of LDCVs is believed to be upregulated when there is increased demand for release at a particular site. Somatodendritic oxytocin release (lower insert) can be triggered by Ca^{2+} from stores in the endoplasmatic reticulum (ER) in addition to the inflow of extracellular Ca^{2+}. In vasopressin and oxytocin neurons, mobilization of intracellular Ca^{2+} can lead to the sustained enhancement of release by promoting recruitment of LDCVs to the releasable pool. Oxytocin cells can autoregulate the dendritic release by the activation of oxytocin receptors (OTRs) that release intracellular Ca^{2+} via an inositol-3,4,5-phosphate (IP3)-mediated pathway.

and partly from differences in the release machineries of SSVs and LDCVs. Neuropeptides require strong stimulation, such as high-frequency firing or burst-patterned firing, to be released, whereas classical neurotransmitters respond also to low frequencies and even single action potentials (Figure 1). The requirement of more intensive stimulation for neuropeptide release is probably linked to the long latencies to release and reflect a need for repeated depolarizations and accumulation of intracellular Ca^{2+}. The reason why neuropeptides play hard to get in this fashion might be that it is disadvantageous to trigger a prolonged and widespread signal too easily, but also it is time- and energy-demanding to refill the neuropeptide stores. Some neuropeptides act in an autoinhibitory fashion, and the restrictive release of such messengers

could serve as an emergency brake that prevents cellular excitotoxic damage or could contribute to cyclic activity in neuronal networks.

Molecular Release Mechanisms – SNARE Machinery

The molecular mechanisms of exocytosis have been studied in some detail over several decades. Some of the cellular systems that have been most useful for these studies examine mainly LDCV fusion; this is the case in adrenal chromaffin cells, from which release of easily oxidized catecholamines can be monitored with high time resolution using carbon-fiber electrodes, and pancreatic β-cells, in which insulin release is usually monitored as changes in cell capacitance. Detailed information about vesicle kinetics and exocytosis has also been obtained by expressing fluorescent tags directed to vesicle membrane proteins or lumen. Less is known regarding exocytosis in peptidergic neurons, but it is usually assumed that LDCVs in neurons fuse with the cell membrane by a mechanism similar to those of peripheral LDCVs because there are many recurring motifs in peripheral and central exocytosis. Here we only briefly describe the characteristics particular for the release of LDCVs.

LDCV exocytosis from endocrine cells occurs from a small fraction of vesicles present near the cell membrane, the releasable pool (~1% of vesicles present at the release site). Some of these vesicles are immediately releasable, but a larger portion requires a longer time to be released. The release process involves several steps and includes the formation of soluble *N*-ethylmaleimide-sensitive factor attached protein receptor (SNARE) complexes consisting of vesicular proteins (vSNARE), cell membrane-bound proteins (target SNARE or tSNARE), and interactions between SNARE complexes and soluble proteins. Before a vesicle can undergo exocytosis, it must be docked to the cell membrane. However, not all docked vesicles are releasable; the docked vesicle needs further preparation, which might involve the binding of Munc13 or Ca^{2+}-dependent activator protein for secretion (CAPS) in a process called priming. The docking and priming processes ensure that only some of the available vesicles are released and that a reserve pool is retained for subsequent release, minimizing the need for newly synthesized peptides (Figure 1). However, in situations of increased demand, it appears that newly synthesized vesicles are released before the already-present reserve pool. The reason for this is unclear, but it might be a way to update the composition of the vesicle content. SSVs also undergo docking and priming before release, but the priming process is more efficient. For example, nearly all docked vesicles are primed and ready to be released in hippocampal neurons. CAPS appears to be essential for the late stages of LDCV exocytosis, but is not necessary for SSV exocytosis, indicating molecular differences in the priming and release processes of the two types of vesicles.

The fusion process is believed to be similar, but probably not identical, for SSVs and LDCVs. In principle, the fusion of vesicle and cell membrane can range from being total, including a flattening and integration of the vesicle membrane into the plasma membrane, to being restricted to a short-lived fusion pore, which may or may not allow the vesicle content to escape. Overall, the fusion times appear to be longer for LDCVs than for SSVs, and although fast retrieval of intact vesicles (called kiss-and-run events) occur with LDCVs, quick recycling of intact vesicles is functionally more important for classical neurotransmitter release, in which the vesicle can be quickly refilled or topped up (Figure 1).

Termination of the Signal

The termination of a signal is critical for information transfer. The simpler the signal is, the more important this is. If we wanted to use a flashlight to send a signal in the dark, a simple solution might be to switch the flashlight on and off in Morse code. However, if we had a box of matches instead of a flashlight, it would be laborious to create a Morse code signal. A better approach might be to write our message on the wall and illuminate it by lighting a few matches. In the brain, neuropeptides signal like illuminating matches, and the message that they convey is dictated by the pattern of receptor expression, which in this metaphor is the writing on the wall. The signaling mode of classical transmitters resembles the Morse code approach, and to achieve this, the action of classical transmitters is both rapidly turned on and rapidly terminated. Efficient uptake mechanisms and fast breakdowns ensure that the signal can be repeated with high frequencies. By contrast, neuropeptide signals are terminated by the slow breakdown of the peptides by extracellular peptidases. The cerebral half-lives for the neuropeptides vasopressin and oxytocin, for example, are ~20 min, whereas fast-acting classical neurotransmitters typically have half-lives of ~5 ms. Because neuropeptides display high-affinity interactions with their receptors (typically in low nanomolar ranges), the action can be widespread in both time and space. This kind of transmission, in which molecules act at a distance from the release site, is sometimes referred to as paracrine, indicating that the message indiscriminately reaches other cells near the site of release. Neuropeptides can, in this respect, be thought of as paracrine messengers in the brain, and an important function appears to be to convey persistent modulating or tuning effects on

neuronal networks rather than to transmit immediate but fugacious messages.

The lack of peptide uptake mechanisms has an important bearing on release dynamics; unlike amines and amino acids, a released peptide cannot be reused but must be synthesized *de novo* whenever it is released. There might be some exceptions; a high-affinity uptake of cholescystokinin has been reported, for example.

Neuropeptide Release Is Not Uniform

Variations between different neurons and over time Although the release of different neuropeptides shares characteristics that are related to the type of vesicle used, there are many variations. Even the morphological appearances of LDCVs indicate differences. LDCVs vary in diameter from 60 to 500 nm. This variation occurs between and within neurons of a population, as well as between different neuron populations. However, it is not clear whether different sizes are associated with differences in release mechanisms other than the obvious difference in released quanta. The content of individual vesicles varies not only between neurons but also within them. LDCVs often contain more than one neuropeptide, and in fact many neurons release a mixture of neuropeptides, sometimes referred to as a neuropeptide cocktail. For instance, vasopressin coexists with dynorphin and galanin in magnocellular neurons of the supraoptic nucleus and paraventricular nucleus and with corticotropin-releasing hormone (CRH) in parvocellular neurons of the paraventricular nucleus. On the other hand, oxytocin in the supraoptic nucleus coexists with enkephalin, dynorphin, cocaine- and amphetamine-regulated transcript (CART), and cholecystokinin. Neuropeptides can also be co-released with classical neurotransmitters such as acetylcholine and γ-aminobutyric acid (GABA) or neuromodulators such as amines, which provide additional flavors to the cocktail. For example, the neuropeptide Y (NPY) neurons also make the classical inhibitory neurotransmitter GABA.

Moreover, most central neuropeptide-releasing neurons appear to have at least one major classical neurotransmitter that sustains fast neurotransmission via SSV release, in addition to the slower effects of neuropeptide release from LDCVs. The amount of co-released and/or co-expressed classical transmitters can also be subject of variation, contributing to the diversity of neuropeptidergic transmission. A dynamic change in neuropeptide expression over time is another factor. Some peptidergic neurons have a relatively stable neuropeptide expression (e.g., some hypothalamic peptides, such as vasopressin, oxytocin, and galanin), but others express neuropeptides only transiently during development or display pronounced upregulation from low levels in response to appropriate physiological stimuli (e.g., vasoactive intestinal peptide (VIP), galanin, or NPY in primary sensory neurons). It is natural to assume that two or more messengers that are released at the same time from the same axon terminal (or even the same vesicle) interact with one another's effects. There are, indeed, several examples both of synergism between co-released neuropeptides and classical transmitters (e.g., acetylcholine and VIP in salivary glands) and of antagonism between co-released transmitters (e.g., galanin and noradrenalin in the brain stem). However, co-released neuropeptides might not always serve as co-transmitters. Because the physical and time-dependent range of neuropeptide action is so large, it is possible that two neuropeptide transmitters released from the same axon terminal or vesicle have completely independent effects, even involving separate receiving neurons. Neuropeptide release is clearly a multifaceted phenomenon in the brain, and we are only beginning to understand the reasons for this complexity. One suggestion which has been put forward is that the co-release of some neurotransmitters constitutes phylogenetic remnants that are no longer linked to function, and it cannot be ruled out that this is, at least sometimes, the case. However, this is not to suggest that the neurotransmitter in question is without biological action. It might just not have a physiological function in every cell and vesicle where it is present.

Variation within neurons – axon terminal versus somatodendritic release Neuropeptides can be differentially released from separate compartments within a neuron. This is the case in hypothalamic magnocellular oxytocin and vasopressin neurons. These neurons release neuropeptides into the bloodstream from axon terminals in the posterior pituitary but also into the CNS from dendrites that are densely packed with oxytocin- or vasopressin-containing LDCVs. The particular anatomical arrangement of these neurons makes it possible to separate the release from axon terminals from the release originating in somatodendritic elements. The magnocellular neurons have virtually no axon collaterals that project back to the dendritic region. Consequently, all axon-terminal release takes place in the posterior pituitary, and conversely, in the supraoptic nucleus, all vasopressin and oxytocin release comes from magnocellular somata and dendrites.

The peripheral action of vasopressin is to promote water retention and increase blood pressure. When magnocellular neurons are activated by systemically increased osmolarity, they respond by releasing vasopressin (and in the rat, oxytocin) from their terminals in the posterior pituitary but also from the dendrites in the supraoptic nucleus. Interestingly, the release from

dendrites and terminals does not occur simultaneously, as might be expected. The dendritic release *in vivo* is delayed by approximately 1 h and persists longer than terminal release. An opposite order of events is observed in lactating rats when oxytocin release in response to suckling is measured *in vivo*; in this case, dendritic release precedes terminal release. Apparently, axon-terminal and somatodendritic release employ distinct activity-dependent release mechanisms. One way in which dendritic oxytocin release differs from terminal release is by its dependence on mobilization of Ca^{2+} from intracellular stores (Figure 1). This is illustrated by the effects of α-melanocyte-stimulating hormone (α-MSH), which, when administered to oxytocin cells, triggers intracellular Ca^{2+} mobilization by activating MC4 receptors. *In vivo* this leads to dendritic oxytocin release but also to decreased firing of the oxytocinergic neurons and thereby decreased axon-terminal release into the blood.

Somatodendritic vasopressin and oxytocin release, furthermore, displays a kind of priming which is not seen with axon-terminal release. Mobilization of intracellular Ca^{2+} from thapsigargin-sensitive stores of the endoplasmic reticulum leads to enhancement of subsequent release. The consequence is a more pronounced somatodendritic release of the neuropeptides in response to a given stimulus. Interestingly, this enhancement of dendritic release is long-lasting, and it is not a direct effect of the increase in intracellular calcium because it develops some time after the intracellular Ca^{2+} concentrations have returned to normal. As might be expected, the lasting enhancement of somatodendritic neuropeptide release is associated with an increased number of vesicles that are closely associated with the cell membrane, reflecting a change in the recruitment of vesicles to the readily releasable pool (Figure 1). The described enhancement (priming) of somatodendritic oxytocin and vasopressin release can be evoked by drugs such as thapsigargin and cyclopiazonic acid (Ca^{2+} mobilizers) but also by endogenous oxytocin. Oxytocin released from the somata and dendrites therefore promotes its own release by a feed-forward mechanism that facilitates recruitment of LDCVs from the reserve pool.

As demonstrated by these examples, axon-terminal and somatodendritic neuropeptide release can occur independently, indicating that neuropeptides can have separate physiological functions depending on which compartment of the neuron they are released from. Vasopressin and oxytocin neurons differ from most other neurons that release neuropeptides in the brain by the unusually high concentration of peptide-containing LDCVs in the dendritic compartments. It remains to be investigated whether differential dendritic and terminal release is also found in other peptidergic neurons. It is, however, already known that dendritic release is not restricted to peptidergic neurons but is a feature of many, if not all, neurotransmitter systems. Dopaminergic cells also display differences between terminal and somatodendritic release that suggest independent release mechanisms, and accumulating evidence favors the concept that dendritic release is a distinct mode of neurotransmission.

Physiological Effects of Neuropeptide Release

We have touched on the characteristics of neuropeptide release that single out this type of transmission from fast classical neurotransmission. The many neuropeptides in the nervous system can have an equally large number of physiological effects. Neurotransmitters generally affect the excitability of other neurons by depolarizing them or by hyperpolarizing them. Peptides have much more diverse effects; among other things, they can affect gene expression, local blood flow, synapotogenesis, and glial cell morphology. Some examples, which illustrate how the characteristic neuropeptide release mechanisms translate into physiological functions, are summarized next.

Neuropeptides have well-established roles in a number of physiological systems such as CNS control of food intake, pain and anxiety, stress, learning and memory, and social behavior. They may also have a role in the pathophysiology of addiction, depression, and eating disorders. Common to the action of neuropeptides in these examples is the long timescale. For example, the CNS control of food intake is not concerned with matching intake and expenditure on millisecond, second or even daily basis but, instead, provides a balance over a week and longer. The action of hypocretins/orexins, ghrelin, NPY, α-MSH, and other appetite-regulating neuropeptides obviously does not last that long *per se*, but they can evidently tune the drive for food intake over time periods of this magnitude. It is easily recognized that fast and spatially precise signals are not needed for this. Even peptides that counteract epileptic seizures, such as galanin and NPY, act on a relatively long timescale. NPY-deficient mice can, in most electrophysiological aspects, not be separated from wild-type mice. When tested with the pharmacological seizure-inducer kainic acid, wild-type and NPY −/− mice display inseparable latencies to electroencephalographic seizure activity, and partial seizures are also similar in behavioral appearance. However, when seizure activity generalizes, wild-type mice seizures end within 30 s, whereas NPY −/− seizures last more than twice as long and ultimately progress into status

epilepticus and death. Thus, it seems that NPY can serve as a brake to stop epileptic seizures. The fact that the major difference between wild-type mice and NPY −/− is that the latter have problems terminating epileptic activity is consistent with the need for intensive stimulation to evoke neuropeptide release, as well as the long latencies to release.

Somatodendritic releases of vasopressin and oxytocin have autoregulatory effects. In the case of vasopressin, this regulation can be either inhibitory or excitatory, depending on the prevailing activity of the neuron. Autoregulation in vasopressin neurons can be amplified by the insertion of κ-opioid receptors, which are present in the LDCV membrane and inserted into the cell membrane on exocytosis. Through such insertions of opioid receptors or other neuropeptide receptors, neuropeptidergic neurons can use LDCV exocytosis to transmit a signal to surrounding neurons and at the same time change their own responsiveness.

One autoregulatory effect of oxytocin is to promote burst activity. Synchronized burst firing of oxytocin neurons is seen during labor and in response to suckling in lactating females but not otherwise. The transition from the normal continuous unsynchronized behavior of oxytocin neurons into synchronized burst-firing response is mediated by central oxytocinergic transmission. By priming somatodendritic oxytocin release using thapsigargin, it is also possible to generate burst-patterned firing in virgin rats, indicating that this network plasticity is mediated by specific dendritic release mechanisms. This kind of functional rewiring of neural networks through paracrine signaling has recently emerged as a mechanism by which dendritically released neuropeptides can alter behavior.

Neurons that release luteinizing hormone-releasing hormone (LHRH) also display dendritic neuropeptide release, in addition to the terminal release into the hypophysial portal circulation. Unlike vasopressin and oxytocin neurons, the LHRH neurons in the preoptic/septal area are scattered throughout a relatively large area. In spite of this, they release LHRH in a synchronized pulsatile manner, and it has been suggested that LHRH autoregulation plays a role in the generation of this pulsatile behavior, which triggers puberty and is necessary for normal ovulatory menstrual cycles in females. Neuropeptide release clearly has widespread physiological effects that stretch beyond the scope of this chapter.

Summary

A large number of different neuropeptides are released from neurons in the periphery and in the brain. In most cases, little is known about the release mechanisms and the regulation of release, but the available information reveals some common themes that distinguish neuropeptide release from the release of classical neurotransmitters. Neuropeptides are always released from LDCVs and have to be synthesized *de novo* for each release event. An economizing transport and recruitment mechanism is therefore needed to prevent depletion. This is mediated in part by a more restrictive priming process, which only allows a small fraction of the present vesicles to be released in response to depolarization, and in part by a spatial separation of LDCV Ca^{2+} sensors and voltage-sensitive Ca^{2+} channels. Typically, neuropeptides are, therefore, released only in response to high-frequency or burst-patterned firing or in response to intercellular messengers that mobilize intracellular Ca^{2+} stores. On the other hand, the action of neuropeptides is prolonged and extended due to slow inactivation and high-affinity receptors. This slow and wide mode of transmission is reflected in many physiological actions of neuropeptides.

See also: Calcium Channels and SNARE Proteins; Large Dense Core Vesicles (LDCVs); Neuropeptide Inactivation or Metabolism; Neuropeptide Synthesis and Storage; Oxytocin (Peripheral/Central Actions and their Regulation); SNAREs; Synaptic Vesicles; Vasopressin/Oxytocin and Receptors.

Further Reading

An S and Zenisek D (2004) Regulation of exocytosis in neurons and neuroendocrine cells. *Current Opinion in Neurobiology* 14(5): 522–530.

Baraban SC and Tallent MK (2004) Interneuron diversity series: Interneuronal neuropeptides – endogenous regulators of neuronal excitability. *Trends in Neurosciences* 27(3): 135–142.

Burgoyne RD and Morgan A (2003) Secretory granule exocytosis. *Physiological Reviews* 83(2): 581–632.

Harata NC, Aravanis AM, and Tsien RW (2006) Kiss-and-run and full-collapse fusion as modes of exo-endocytosis in neurosecretion. *Journal of Neurochemistry* 97(6): 1546–1570.

Hökfelt T, Bartfai T, and Bloom F (2003) Neuropeptides: Opportunities for drug discovery. *Lancet Neurology* 2(8): 463–472.

Hökfelt T, Broberger C, Xu ZQ, et al. (2000) Neuropeptides – an overview. *Neuropharmacology* 39(8): 1337–1356.

Kits KS and Mansvelder HD (2000) Regulation of exocytosis in neuroendocrine cells: Spatial organization of channels and vesicles, stimulus-secretion coupling, calcium buffers and modulation. *Brain Research Reviews* 33(1): 78–94.

Landgraf R and Neumann ID (2004) Vasopressin and oxytocin release within the brain: A dynamic concept of multiple and variable modes of neuropeptide communication. *Frontiers in Neuroendocrinology* 25(3–4): 150–176.

Ludwig M (ed.) (2005) *Dendritic Neurotransmitter Release*. New York: Springer.

Ludwig M and Leng G (2006) Dendritic peptide release and peptide-dependent behaviours. *Nature Reviews Neuroscience* 7(2): 126–136.

Martin TF (2003) Tuning exocytosis for speed: Fast and slow modes. *Biochimica et Biophysica Acta* 1641(2–3): 157–165.

Salio C, Lossi L, Ferrini F, and Merighi A (2006) Neuropeptides as synaptic transmitters. *Cell and Tissue Research* 326(2): 583–598.

Neuropeptide S

R K Reinscheid, University of California at Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Structure and Expression of Neuropeptide S

Neuropeptide S (NPS) was named for its N-terminal serine residue, conserved across all species in which NPS sequences have been found at the genomic level (Figure 1). NPS was isolated as a ligand of an orphan G-protein-coupled receptor and first disclosed in a patent. Similar to other neuropeptides, NPS is encoded by a precursor protein that contains a pair of basic amino acid residues immediately preceding the immature peptide that will likely serve as processing sites for proteolytic cleavage. The NPS peptide is found at the C-terminal end of the precursor protein. The NPS precursor is divided into at least three exons and is located on human chromosome 10. Thus far, DNA sequences encoding NPS have been found in a number of vertebrate species with the exception of fish. The primary structure of NPS has been well conserved throughout evolution. This degree of conservation is indicative of the importance of the N-terminal part for biological activity (Figure 1).

In the rat, NPS precursor messenger RNA (mRNA) is expressed mainly in the brain. A cluster of cells located between the locus coeruleus and Barrington's nucleus in the pontine brain stem shows high expression levels of NPS precursor mRNA (Figure 2). This previously unidentified group of cells defines a novel nucleus in the brain stem. NPS precursor expression is also detected in the lateral parabrachial nucleus, the principle 5 sensory nucleus of the brain stem, and in a few scattered neurons in the amygdala and hypothalamus. NPS-producing neurons in the locus

SFRNGVGTGMKKTSFQRAKS	*Homo sapiens* (human)	Primates
SFRNGVGTGMKKTSFRRAKS	*Pan troglodytes* (chimpanzee)	Primates
SFRNGVGTGMKKTSFRRAKS	*Pongo pygmaeus* (orangutan)	Primates
SFRNGVGTGMKKTSFRRAKS	*Callithrix jacchus* (marmoset)	Primates
SFRNGVGTGGKKTSFQRAKS	*Otolemur garnettii* (galago)	Primates
SFRNGVGTGMKNTSFRRAKS	*Macaca mulatta* (macaque)	Primates
SFRNGVGTGMKKTSFQRAKS	*Loxodonta africana* (African elephant)	
SFRNGVGTGMKKTSFRRAKS	*Bos taurus* (cattle)	
SFRNGVGTGMKKTSFQRAKS	*Equus caballus* (horse)	
SFRNGVGTGMKKTSFRRAKS	*Canis familiaris* (dog)	
SFRNGVGMGMKKTSFQRAKS	*Sorex araneus* (Eurasian shrew)	
SFRNGVGTGMKKTSFRRAKS	*Tupaia belangeri* (northern tree shrew)	
SFRNGVGTGMKKTSFRRAKL	*Oryctolagus cuniculus* (rabbit)	
SFRNGVGSGVKKTSFRRAKQ	*Rattus norvegicus* (rat)	Rodents
SFRNGVGSGAKKTSFRRAKQ	*Mus musculus* (mouse)	Rodents
SFRNGVGTGMKKTSFPRAKP	*Cavia porcellus* (guinea pig)	Rodents
SFRNGVGTGLKKTPFPRAKS	*Myotis lucifugus* (little brown bat)	
SFRNGVGQGIKKTSFRRAKS	*Ornithorhynchus anatinus* (platypus)	
SFRNGVGSGIKKTSFRRAKS	*Gallus gallus* (chicken)	
SFRNGVGSGIKKNSFRRAKL	*Xenopus tropicalis* (western clawed frog)	
SFRNGVGTGIKKTSFRRAKS	*Macropus eugenii* (wallaby)	Marsupials
SFRNGVGSGMKKTSFRRAKS	*Monodelphis domestica* (opossum)	Marsupials

Figure 1 Alignment of neuropeptide S (NPS) amino-acid sequences from various species. Sequences were deduced from GenBank resources by searching databases containing expressed and genomic DNA sequences. Residues different from the human NPS structure are shown in red.

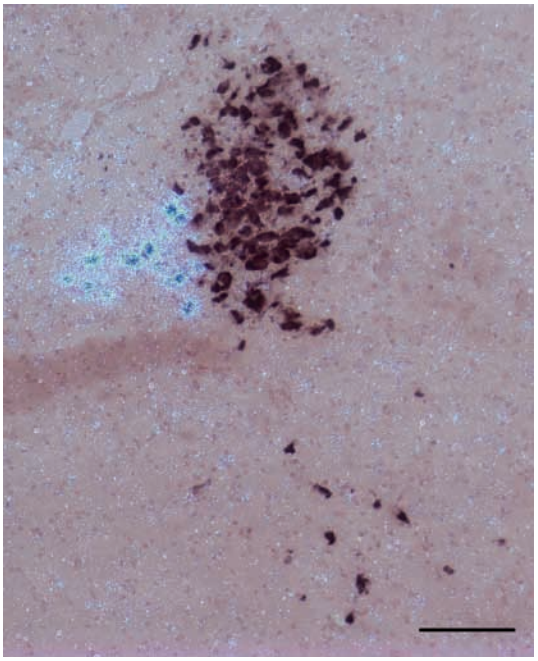

Figure 2 Double *in situ* hybridization of the rat brain stem for tyrosine hydroxylase (TH, brown) and neuropeptide S (NPS) precursor (white grains) in the locus coeruleus area. Antisense riboprobes for TH were labeled with digoxigenin, probes for NPS precursor were labeled with ^{35}S-UTP, and hybridizations were carried out. Pictures for TH signals were taken under brightfield illumination, and a separate darkfield image was made from the same section to capture NPS precursor hybridization signals. Both images were combined in Photoshop and adjusted for brightness and contrast. Medial areas are on the left, and lateral areas are located on the right side in the section shown. Scale bar: 200 μm. Experimental conditions are described in detail in Xu YL, Reinscheid RK, Huitron-Resendiz S, et al. (2004) Neuropeptide S: A neuropeptide promoting arousal and anxiolytic-like effects. *Neuron* 43: 487–497.

coeruleus area and the principle 5 sensory nucleus are mainly glutamatergic; the NPS-expressing neurons in the lateral parabrachial nucleus co-express corticotropin-releasing factor. Overall, the expression of NPS precursor is rather restricted in the brain. Peripheral sites expressing NPS precursor include mostly endocrine tissues such as thyroid, mammary gland, salivary gland, and testis.

NPS Receptor

The NPS receptor (NPSR) is a typical G-protein-coupled receptor, also known as GPR135 and vasopressin receptor-related receptor 1. The receptor gene contains at least nine exons and is located on human chromosome 7p14. NPSR has been found to be expressed mainly in the central nervous system of rats by using *in situ* hybridization. NPSR mRNA is widely distributed in many brain areas, with high expression levels in cortex, hypothalamus, amygdala, and multiple midline thalamic nuclei (Figure 3). Many of these areas have been functionally associated with arousal, learning and memory, or processing of emotional behavior.

Physiological Effects of NPS

Central administration of NPS promotes arousal in mice and is able to suppress all stages of sleep in rats. Nanomolar doses of NPS administered centrally induce hyperlocomotion. In addition, NPS appears to produce anxiolytic-like behavior in rodents. This pharmacological spectrum is quite unique since classical psychostimulants either increase anxiety-like behavior or do not affect emotionality. On the other hand, typical anxiolytics normally reduce locomotion and cause sedation. NPS was also shown to suppress feeding although it is currently unclear whether the anorectic effects might be secondary to the behavioral arousal caused by activation of central NPS receptors.

Genetics of NPSR

In 2004, NPSR was identified as an asthma susceptibility gene in a genomewide screen in Finnish and Canadian patients. This study termed the gene "G protein-coupled receptor for asthma susceptibility" but did not provide functional data explaining the physiological role of NPSR in asthma. The study showed that a number of polymorphic variants of NPSR exist in humans and that particular sets of these variants (haplotypes) are associated with an increased risk of asthma and possibly allergic diseases characterized by high immunoglobulin E serum levels. Only one polymorphic site was identified within the coding region, changing asparagine (Asn) to isoleucine (Ile) at amino acid position 107 of the receptor protein. A C-terminal splice variant of human NPSR was found to be overexpressed in human asthmatic airway tissue. Expression of NPSR mRNA was also found upregulated in a mouse model of airway inflammation. These data indicate that NPSR might have important physiological functions in peripheral tissues and could be involved in the pathophysiology of asthma.

A subsequent pharmacological analysis of NPSR variants showed that the coding polymorphism at amino acid position 107 (Asn-Ile) of NPSR produces a gain of function in the receptor protein. NPSR Ile^{107} displays a tenfold increase in agonist efficacy without changing ligand binding affinity. The C-terminal splice variant of human NPSR did not show altered pharmacology.

The involvement of NPSR in the pathophysiology of asthma has recently been challenged by a report

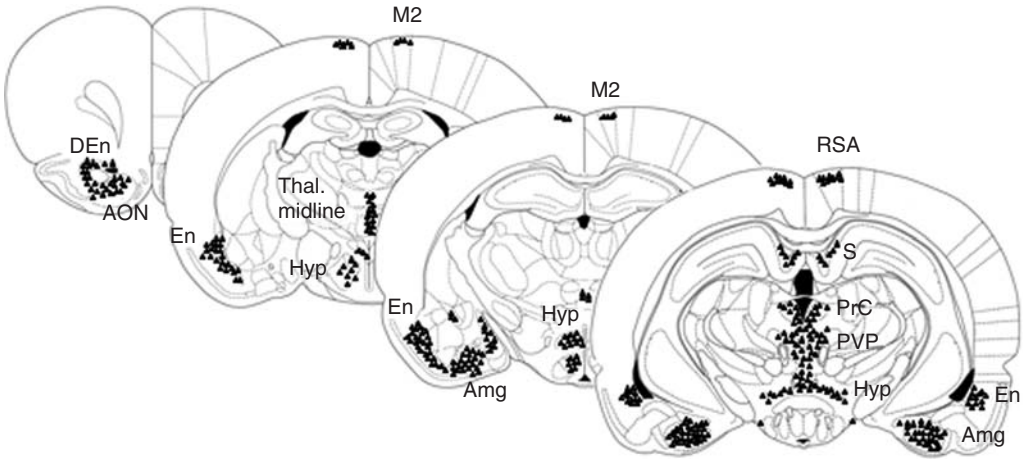

Figure 3 Schematic drawings of neuropeptide S (NPS) receptor messenger RNA (mRNA) expression in the rat brain. Representative regions with high levels of NPS receptor mRNA signals (small triangles) are depicted in the drawings. Strong NPS receptor mRNA expression is found in the anterior olfactory nucleus, endopiriform nucleus, piriform cortex, motor cortex, retrosplenial cortex, and subiculum. Several thalamic nuclei, including the midline nuclei of the thalamus, express high levels of NPSR (thal. midline). Significant expression of NPSR mRNA is also observed in the hypothalamus and the amygdala complex. Amg, amygdala; AON, anterior olfactory nucleus; DEn, dorsal endopiriform nucleus; En, endopiriform nucleus (dorsal and ventral); Hyp, hypothalamus; M2, motor cortex 2; PrC, precommissural nucleus; PVP, paraventricular thalamic nucleus; RSA, retrosplenial agranular cortex; S, subiculum. Adapted from Paxinos G and Watson C (1997) *The Rat Brain in Stereotaxic Coordinates*, 3rd edn. San Diego, CA: Academic Press, with permission from Elsevier.

showing that NPSR-deficient mice did not differ from wild types in their susceptibility to allergic lung disease. Also, the report could not confirm expression of NPSR in human lung or mice with allergic lung disease. The investigators concluded that NPSR might have only indirect influence on lung physiology. This contrasts with a report showing expression of NPSR in human bronchial tissue by *in situ* hybridization. The reason for these discrepancies is currently unclear. Future studies using small molecule agonists and antagonists for NPSR will certainly help define the physiological functions of the NPS system.

See also: Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptide FF and Receptors; Neuropeptides and Receptors in Glia; Neuropeptides: Discovery; Neuropeptides: Food Intake.

Further Reading

Allen IC, Pace AJ, Jania LA, et al. (2006) Expression and function of NPSR1/GPRA in the lung before and after induction of asthma-like disease. *American Journal of Physiology. Lung Cellular and Molecular Physiology* 291(5): L1005–L1017.

Beck B, Fernette B, and Stricker-Krongrad A (2005) Peptide S is a novel potent inhibitor of voluntary and fast-induced food intake in rats. *Biochemical and Biophysical Research Communications* 332: 859–865.

Civelli O, Nothacker HP, Saito Y, Wang Z, Lin SH, and Reinscheid RK (2001) Novel neurotransmitters as natural ligands of orphan G-protein-coupled receptors. *Trends in Neuroscience* 24: 230–237.

Gupte J, Cutler G, Chen JL, and Tian H (2004) Elucidation of signaling properties of vasopressin receptor-related receptor 1 by using the chimeric receptor approach. *Proceedings of the National Academy of Sciences of the United States of America* 101: 1508–1513.

Koob GF and Greenwell TN (2004) Neuropeptide S: A novel activating anxiolytic? *Neuron* 43: 441–442.

Laitinen T, Polvi A, Rydman P, et al. (2004) Characterization of a common susceptibility locus for asthma-related traits. *Science* 304: 300–304.

Paxinos G and Watson C (1997) *The Rat Brain in Stereotaxic Coordinates*, 3rd edn. San Diego, CA: Academic Press.

Reinscheid RK and Xu YL (2005) Neuropeptide S and its receptor: A newly deorphanized G protein-coupled receptor system. *Neuroscientist* 11: 532–538.

Reinscheid RK, Xu YL, Okamura N, et al. (2005) Pharmacological characterization of human and murine neuropeptide s receptor variants. *Journal of Pharmacology and Experimental Therapeutics* 315: 1338–1345.

Sato S, Shintani Y, Miyajima N, and Yoshimura K (2002) Novel G Protein-Coupled Receptor Protein and DNA Thereof. World Patent Application WO 02/31145 A1.

Smith KL, Patterson M, Dhillo WS, et al. (2006) Neuropeptide S stimulates the hypothalamo-pituitary-adrenal axis and inhibits food intake. *Endocrinology* 147: 3510–3518.

Vendelin J, Pulkkinen V, Rehn M, et al. (2005) Characterization of GPRA, a novel G protein-coupled receptor related to asthma. *American Journal of Respiratory Cell and Molecular Biology* 33: 262–270.

Xu YL, Gall CM, Jackson VR, Civelli O, and Reinscheid RK (2007) Distribution of neuropeptide S receptor mRNA and neurochemical characteristics of neuropeptide S expressing neurons in the rat brain. *Journal of Comparative Neurology* 500: 84–102.

Xu YL, Reinscheid RK, Huitron-Resendiz S, et al. (2004) Neuropeptide S: A neuropeptide promoting arousal and anxiolytic-like effects. *Neuron* 43: 487–497.

Neuropeptide Signaling in Invertebrates

D R Nässel, Stockholm University, Stockholm, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Invertebrates constitute the vast majority of all living animals and display an immense variation in nervous system organization and complexity of behavior. However, at all these levels of organization neuropeptides seem to be used for signaling. Thus, basal invertebrates such as cnidarians (e.g., jellyfish and corals) appear to express large numbers of different neuropeptides, and complex peptidergic signaling has been described in the small nematode worm *Caenorhabditis elegans*. This worm, and the fruit fly *Drosophila melanogaster,* have become very useful animal models for the analysis of neuropeptide function. Classically, however, crustaceans (crabs and crayfish), some larger insects (cockroaches, locusts, and moths), and mollusks (e.g., *Aplysia* and *Lymnaea*) were used for endocrinological and molecular studies of neuropeptides and peptide hormones. This article provides a brief and topical overview of neuropeptides and their possible functions in some key groups of invertebrates.

Neuropeptides and Their Precursors in Invertebrates

Invertebrate neuropeptides consist of four to approximately 80-amino-acid residues linked by peptide bonds. Some larger proteins are also known to act as hormones in insects, for example, prothoracicotropic hormone and the heterodimeric cystin knot protein bursicon. Neuropeptides display a great diversity in their molecular structures and display a comparable diversity in their modes of action and functional roles. Invertebrate neuropeptides are known to regulate many aspects of animal development, growth, reproduction, homeostasis, physiology, and behavior. There has recently been a tremendous increase in the number of identified neuropeptides and peptide G-protein-coupled receptors (GPCRs) in several invertebrate species due to the completion of genome-sequencing projects. Thus, we can estimate the number of genes encoding neuropeptide precursors and peptide GPCRs in a given species. For example, in the nematode worm *C. elegans,* there are 98 genes identified that encode putative neuropeptide precursors, including 37 insulin-like genes and approximately 50 genes for predicted peptide GPCRs. In the fruit fly *Drosophila* there are approximately 35 neuropeptide-encoding genes and 48 genes encoding peptide and protein GPCRs.

As in vertebrates, the invertebrate neuropeptides are processed from large precursor proteins with conserved basic features. Some neuropeptides are present in single copies on the precursors, but in invertebrates, more commonly, multiple copies of identical or slightly diversified peptides (isopeptides) can be seen. In mollusks, it has been shown that different peptides that are cleaved from a single precursor can have differential functions and act on different receptors. These peptides are distributed to mutually exclusive populations of neurons by means of alternative splicing of the precursor transcripts. Many of the invertebrate neuropeptides are C-terminally α-amidated and display other posttranslational modifications of structure that generate varying degrees of stability to peptidases.

Invertebrate neuropeptides are localized in neurosecretory cells, interneurons, motor neurons, and sensory neurons. Peptides can also be found in endocrine cells of the intestine and at peripheral locations. Thus, peptides can act as hormones released into the circulation as local neurohormones released nonsynaptically in a paracrine fashion within the nervous system or at muscles or glands. Neuropeptides may also be released at synapses and may act as synaptic modulators or even as co-transmitters, thereby modifying the action of fast-acting neurotransmitters. Each neuropeptide is distributed in a stereotypic pattern in specific neurons, neurosecretory cells, or endocrine cells. Neuropeptides commonly act on different types of GPCRs, but some activate receptor tyrosine kinases or, in the case of molluscan FMRFamide, act on a ligand-gated sodium channel.

Neuropeptides and Their Roles in Some Invertebrates

Cnidarians Provide Information on Basic Neuropeptide Signaling

Cnidarians constitute a phylum of primitive animals, including, for instance, jellyfish, corals, and sea anemones. The basic organization of the cnidarian nervous system is simple, with a plexus of neurons in the basal part of the epithelium that form centralizations in some strategic regions. Analysis of neuropeptide signaling in cnidarians is of evolutionary interest because this phylum has such a basal phylogenitic position.

Neuropeptides are widespread in the cnidarian nervous system and their differential distribution suggests functional specializations (Figure 1). From

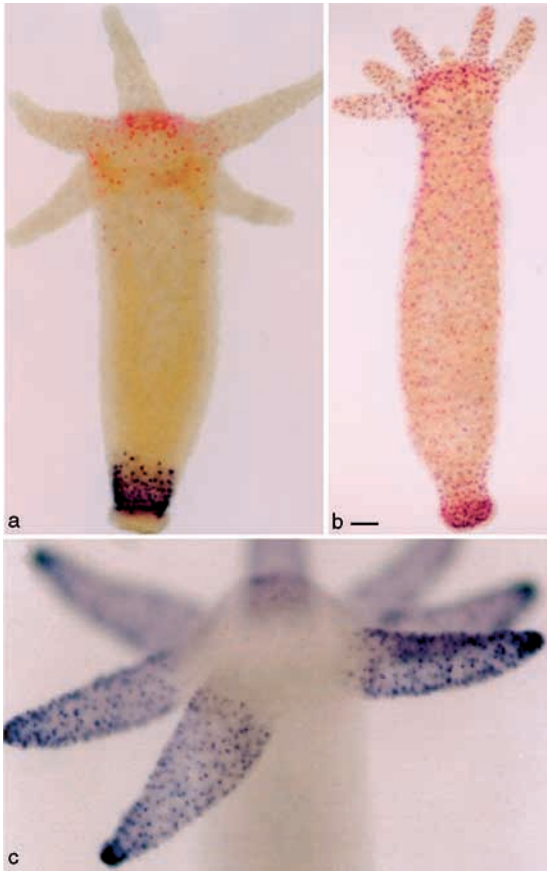

Figure 1 *In situ* hybridization of neuropeptides in *Hydra magnipapillata* (Cnidaria): (a) double-labeling *in situ* hybridization with specific cRNA probes coding for preprohormone-A (stained with Fast Red) and prepro-KVamide (stained with BM purple); (b) double-labeling *in situ* hybridization using a probe coding for the Hydra-RGamide preprohormone (stained with BM purple) and one for the Hydra-LWamide preprohormone (stained with Fast Red); (c) hybridization with a preprohormone-C cRNA probe. *In situ* hybridization reveals the localization of various precursors in specific subsets of neurons, suggesting distributed functions of the encoded peptides. In (a), note that co-expression of the two mRNAs occurs in neurons in the peduncle, which now have obtained a brilliant *violet* color. In (b), the resulting color in the neurons indicates that the two prehormone genes are co-expressed. In (c), preprohormone-C neurons occur only in the tentacles. From Hansen GN, Williamson M, and Grimmelikhuijzen CJ (2000) Two-color double-labeling *in situ* hybridization of whole-mount Hydra using RNA probes for five different Hydra neuropeptide preprohormones: Evidence for colocalization. *Cell and Tissue Research* 301: 245–253; Hansen GN, Williamson M, and Grimmelikhuijzen CJ (2002) A new case of neuropeptide coexpression (RGamide and LWamides) in Hydra, found by whole-mount, two-color double-labeling in situ hybridization. *Cell and Tissue Research* 308: 157–165.

a single species of sea anemone (*Anthopleura elegantissima*), 30 peptides, four- to nine-amino acids long, have been identified. These were classified into seven groups based on their C-terminal dipeptides. Molecular cloning has revealed several peptide precursor genes from various cnidarians. These commonly encode many copies of related or identical peptides. There are, for instance, 36 identical copies of antho-RFamide in the predicted precursor from *Anthopleura*. Many of the cnidarian neuropeptides are myostimulatory or myoinhibitory in different muscle groups. Some neuropeptides are morphogenitic factors. For instance, the metamorphosins or LWamides induce metamorphosis in cnidarian larvae and hydra-RGamide stimulates differentiation of neuronal stem cells into neurons.

Except for the RFamide peptides, which are grouped with the FMRFamide-related peptides of other invertebrates, the cnidarian peptides appear unrelated to known peptides of other phyla. A striking finding is that cnidarians possess peptides with multiple types of posttranslational modifications, some of which are unique. This suggests the presence of multiple peptide-processing enzymes even in a simple nervous system.

The Nematode Worm *C. elegans*

Genome sequencing has revealed 98 *C. elegans* genes encoding neuropeptide precursors, of which 37 encode insulin-like precursors. Not less than 29 genes (*flp-1* to *flp-29*) code for multiple FMRFamide-like peptides (FLPs). Nine of these genes contain a single FLP and the others two to nine FLPs each. Together, the 29 *flp* genes may give rise to at least 68 separate FLPs. In addition to the *flp* genes, the worm genome contains 32 other genes encoding neuropeptide-like precursors (*nlp-1* to *nlp-32*). These predicted precursors may produce a total of 151 neuropeptides. Some of these *nlp*-derived peptides appear to be related to neuropeptides known from other invertebrates, for example, peptides with sequence similarity to molluskan APGWamide and buccalin peptides, crustacean orkokinin peptides, and arthropod allatostatin peptides. It is estimated that the *C. elegans* genome has approximately 50 peptide GPCRs, but only a few of these have been paired with peptide ligands, all of which are *flp*-derived.

Making use of the genomics data, it is possible to approach various aspects of neuropeptide biology by the application of sophisticated molecular methods. As an example, systematic RNA interference (RNAi) screens were conducted to search for functions of specific peptide GPCRs. One study of 60 worm GPCRs, including 38 peptide GPCRs, showed that disruption of several of these produced alterations in worm coordination or the rate of egg laying, especially GPCRs of the neuropeptide Y (NPY) receptor group produced clear phenotypes when disrupted.

Genetic analysis of naturally occurring variation in the feeding behavior of *C. elegans,* seen in isolates from different parts of the world, has also shed light

on how neuropeptide signaling may regulate behavior. The receptor NPR-1, related to mammalian NPY receptors, displays a polymorphism that correlates with a striking behavioral difference in response to bacterial food. NPR1–215F and NPR1–215V are the two natural receptor variants with a single amino acid difference at position 215 (in a region where the G-protein couples). These variants are associated with feeding socially (aggregation) or solitarily (dispersion), respectively. Complete *Npr-1* knockouts display aggregation of worms, indicating that the normal role of this receptor is to inhibit social feeding. The ligands of NPR-1 are derived from two FLP-encoding genes, *flp-21* and *flp-18*. These genes are expressed in specific sets of neurons (Figure 2), including sensory neurons, interneurons, and motor neurons known to be involved in feeding and locomotion. The NPR-1 is expressed on neurons that promote social feeding via cyclic guanosine monophosphate (cGMP) signaling and is dependent on environmental oxygen levels.

Mollusks

Neuropeptides in mollusks have primarily been investigated in the pond snail *Lymnaea stagnalis* and the nudibranch slug *Aplysia californica*. Molluskan peptides act as hormones involved in egg laying and sexual behaviors or in growth, renal function, respiration, and circulation. Furthermore, various peptides isolated from nerve terminals in muscles have provided important leads to the roles of peptides in the modulation of muscle and neurons. Approximately 100 peptides have been identified in *Aplysia* and approximately 75 in *Lymnaea*. The molluskan neuropeptides identified so far are to a large extent restricted to this phylum. Some exceptions are the FMRFamide-related peptides, myomodulins, conopressin-like peptides, neuropeptide Y-like peptides, and insulin-like peptides, which are similar to peptides in other invertebrates or even in mammals.

Egg laying in *Aplysia* is a classic example of complex stereotyped behavior initiated and orchestrated by hormonally and locally released neuropeptides. This behavior lasts more than 1 h and involves the release of eggs, specific head movements, increased respiratory pumping, and inhibition of locomotion and feeding. Egg laying is initiated and controlled by two clusters of bag cells in the abdominal ganglion. The bag cells have processes in synaptic neuropil and in hormonal release sites, and they produce nine or more neuropeptides encoded on a single precursor. Four of these participate in regulation of the egg-laying behavior: egg-laying hormone (ELH), and α-, β-, and γ-bag cell peptides (BCPs). The ELH and all three BCPs act nonsynaptically in the abdominal ganglion, and ELH acts as a hormone to induce ovulation. Thus, the peptides encoded on one gene act at different levels – in the central nervous system (CNS) they induce prolonged modulation of circuits controlling egg-laying behavior leading to the inhibition of feeding and locomotion and the stimulation of respiratory pumping; in the periphery the peptides act on the muscles of the hermaphrodite gland and heart.

Complex modulatory action of multiple neuropeptides has been studied in *Aplysia* at synapses between the cholinergic motor neurons and muscles controlling the radula at consummatory biting. In addition to acetylcholine, motor neurons synthesize different neuropeptides that can be co-released with acetylcholine at higher frequencies of stimulation. These peptides act as modulators that extend the dynamic range over which the muscle can operate efficiently and thus maintain an optimal motor output when the intensity and frequency of feeding behavior (biting) change.

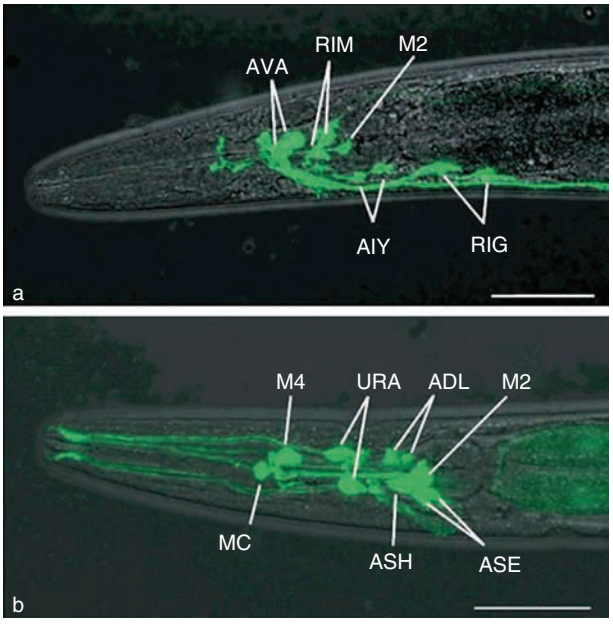

Figure 2 Neuronal expression of peptide genes in *Caenorhabditis elegans*: (a) *flp*-18; (b) *flp*-21. These are seen by expression of GFP under control of *flp*-18 and *flp*-21 promoters. The *flp*-expressing neurons in the anterior part of the worm are indicated the acronyms. GFP, green fluorescent protein. All acronyms represent names of identified neurons of the *C. elegans* nervous system. From Rogers C, Reale V, Kim K, et al. (2003) Inhibition of *Caenorhabditis elegans* social feeding by FMRFamide-related peptide activation of NPR-1. *Nature Neuroscience* 6: 1178–1185.

Crustaceans

Decapod crustaceans have been extensively used in classical studies of the endocrinology of reproduction, molting, color change, carbohydrate metabolism, and water and ion balance. They have also been used as models in the analysis of modulation in

neuronal circuits subserving motor rhythms in feeding. A number of hormonal peptides have been isolated from the sinus gland of the eyestalks of crabs and crayfishes; several myotropic neuropeptides have been found in other tissues. In total, more than 20 distinct types of crustacean peptides have been identified. Some of the sinus-gland peptides are large multifunctional hormones, such as crustacean hyperglycemic hormone and molt-inhibiting hormone, which consist of between 72 and 83 amino acids. The remaining peptides have 5–18 amino acids. Many of the crustacean peptides are similar to insect peptides. Actually, some of the peptides first isolated from crustaceans were later identified in insects and found to play major roles in physiology; they are thus the focus of extensive research in *Drosophila*.

The neuronal circuits of the stomatogastric ganglion (STG) of crustaceans constitute a powerful model for mechanisms in neuromodulation. In crabs, the STG consists of 30 endogenous neurons, most of which are motor neurons innervating different muscles of the foregut and stomach. The hard-wired pattern-generating circuitry of the STG is functionally flexible and can produce sets of dynamic output patterns, subserving a repertoire of chewing and filtering behaviors of the foregut and stomach. STG neurons and their output patterns are modulated by several neuropeptides. The neuropeptides act in these circuits by altering intrinsic membrane properties or the synaptic strength of neurons. Most peptides are excitatory and initiate activity in the quiescent STG or increase the frequency of existing pyloric rhythms; some produce a fusion of two rhythms, thus forming a novel pattern.

Drosophila and Other Insects

Many larger insects such as moths, locusts, and cockroaches have long served in studies of peptide hormone signaling. More recently, *Drosophila* has become a tractable model due to the development of appropriate molecular genetics techniques. In *Drosophila,* 35 genes have been identified that encode neuropeptide precursors (Table 1). These include seven genes of insulin-like peptides and three of protein hormones. In total, there may be 70–90 neuropeptides in *Drosophila*. Of the predicted peptides, many constitute closely related peptide isoforms that may be physiologically redundant. Thus, the total number of functionally distinct neuropeptides may be substantially lower. Approximately, 48 neuropeptide, peptide, and protein hormone GPCRs, have been identified in *Drosophila*.

A variety of peptide actions are known in insects: regulation of development, growth and longevity, control of carbohydrate and lipid metabolism, modulation of contractions of the heart, visceral and skeletal muscle, regulation of water and ion transport, activation of pheromone biosynthesis, regulation of the ovary, and regulation of egg maturation. Peptides are also known to trigger and orchestrate ecdysis behavior, to regulate feeding behaviors, and to act as output factors and modulators of the biological clock. Furthermore, some neuropeptides (e.g., allatotropins, allatostatins, and prothoracicotropic hormone) regulate the release or biosynthesis of other hormones. A likely neuropeptide function that has not been explored carefully in insects is as co-transmitters in the central and peripheral synapses. In insects, neuropeptides are distributed in stereotypical patterns of neurons in the CNS (Figure 3).

One of the very exciting and successful areas in *Drosophila* research is the study of the circadian clock. These clocks are endogenous timekeepers that help organize daily rhythms in behavior, physiology, and metabolism. In *Drosophila,* the master circadian clock is located in the brain and consists of several bilateral sets of neurons (~100) that express specific clock genes and exhibit circadian pacemaker properties. Eighteen of these neurons, in two bilateral groups, express the neuropeptide PDF (pigment-dispersing factor) and are critical for the light-induced and circadian rhythmic behavior of flies. There is evidence that PDF is an output factor of these clock neurons both for rhythm generation and synchronizing pacemaker activity in the clock circuitry. Recent identification and knockout of the *Drosophila* PDF receptor confirmed the importance of PDF signaling in the clock.

A signaling system that appears to be well conserved throughout evolution is that of vertebrate NPY and invertebrate neuropeptide F (NPF). The *Drosophila* NPF, expressed in four brain neurons only, consists of 36 residues. NPF is known to specifically activate an NPF receptor (NPFR1) that displays critical sequence similarities to the mammalian NPY receptors and the NPF receptor of *C. elegans*.

Interference with the expression of NPF or NPFR1 has revealed behavioral phenotypes in *Drosophila* that shed light on peptide signaling in feeding, foraging, and social behavior. NPF signaling in third instar larvae of *Drosophila* is important during the behavioral switch from continuous feeding to wandering, food aversion, clumping, and cooperative burrowing. Younger feeding larvae display a strong expression of NPF mRNA, whereas older nonfeeding third instar larvae display very low levels. Transgenic larvae with a deficiency in NPF signaling precociously display food aversion and social behaviors normally only seen in older nonfeeding larvae. On the other hand, when overexpressing NPF or its receptor, larvae exhibited prolonged feeding activity and suppressed the social behavior of older larvae.

Table 1 Neuropeptide precursor genes identified from *Drosophila melanogaster*

Neuropeptide precursor (peptides)	*Gene*	*CG number*[a]	*Sequence*[b]
Adipokinetic hormone	*akh*	CG1171	pQLTFSPDWa
Amnesiac product(s)	*amn*	CG11937	*c*
Allatostatin-A (drostatin-A1–4)	*ast*	CG13633	VERYAFGLa
Allatostatin-B (drostatin-B1–5)	*mip*	CG6456	AWQSLQSSWa
Allatostatin-C (drostatin-C)	*ast2*	CG14919	pEVRYRQCYFNPISCF
Bursicon	*burs*	CG13419	*c*
Partner of bursicon (bursicon β)	*pburs*	CG15284	*c*
Capability (CAP-1–2, PK-1)	*capa*	CG15520	GANMGLYAFPRVa TGPSASSGLWFGPRLa
Crustacean cardioactive peptide	*ccap*	CG4910	PFCNAFTGCa
Corazonin	*crz*	CG3302	pQTFQYSRGWTNa
Diuretic hormone (CRH-like, DH_{44})	*Dh*	CG8348	*c*
Diuretic hormone (calcitonin-like)	*Dh31*	CG13094	TVDFGLARGYSGTQEAKH-RMGLAAANFAGGPa
dFMRFamides (dFMRFa-1–8)	*fmrfa*	CG2346	DPKQDFMRFa
Drosulfakinins (DSK-1–2)	*dsk*	CG18090	FDDYGHMRFa
Dromyosuppressin (DMS)	*dms*	CG6440	TDVDHVFLRFa
Ecdysis triggering hormone (ETH-1–2)	*eth*	CG18105	DDSSPGFFLKITKNVPRLa
Eclosion hormone	*eh*	CG6400	*c*
Hugin/pyrokinin-2 (hug-γ, PK-2)	*hug*	CG6371	LRQLQSNGEPAYRVRTPRLa SVPFKPRLa
Insulin-like peptide 1	*Dilp-1*	CG14173	*c*
Insulin-like peptide 2	*Dilp-2*	CG8167	*c*
Insulin-like peptide 3	*Dilp-3*	CG14167	*c*
Insulin-like peptide 4	*Dilp-4*	CG6737	*c*
Insulin-like peptide 5	*Dilp-5*	AE003550[d]	*c*
Insulin-like peptide 6	*Dilp-6*	CG14049	*c*
Insulin-like peptide 7	*Dilp-7*	CG13317	*c*
Ion transport peptide (CHH-like)	*itp*	CG13586	*c*
IPNamide (of NPLP-1 precursor)	*nplp1*	CG3441	NVGTLARDFQLPIPNa
Leucokinin-like	*lk*	CG13480	NSVVLGKKQRFHSWGa
Neuropeptide F (long)	*npf*	CG10342	SNSRPPRKNDVNTMADA-YKFLQDLDTYYGDRARVRFa
Neuropeptide F (short) (sNPF-1–4)	*snpf*	GG13968	AQRSPSLRLRFa PQRLRWa
Pigment-dispersing factor	*pdf*	CG6496	NSELINSLLSLPKNMNDAa
Proctolin	*Proct*	CG7105	RYLPT
Prothoracicotropic hormone (PTTH)	*ptth*	CG13687	*c*
SIFamide	*Ifamide*a	CG4681	AYRKPPFNGSIF
Tachykinin-related (DTK-1–6)	*tk*	CG14734	APTSSFIGMRa

[a]Celera Genomics accession numbers.
[b]Sequences given for representative peptides of each precursor.
[c]Sequences too long to be given here.
[d]GenBank accession number for gene cluster.
Data from Hewes RS and Taghert PH (2001) Neuropeptides and neuropeptide receptors in the *Drosophila* genome. *Genome Research* 11: 1126–1142; Nässel DR (2002) Neuropeptides in the nervous system of *Drosophila* and other insects: Multiple roles as neuromodulators and neurohormones. *Progress in Neurobiology* 68: 1–84; Taghert PH and Veenstra JA (2003) *Drosophila* neuropeptide signaling. *Advances in Genetics* 49: 1–65.

There is a convergence of NPF and *Drosophila* insulin-like peptide (DILP) signaling in the regulation of motivated food ingestion. Food-deprived larvae are more likely to feed on noxious or less accessible food. Overexpression of NPFR1 causes nondeprived (fed) larvae to feed on nonfavorable food, whereas the knockdown of the receptor leads to the opposite. The neural activity in NPFR1-expressing neurons seems to be directly regulated by DILP signaling; the overexpression of DILP signaling components in NPFR1 neurons leads to a decreased feeding response to noxious food, whereas downregulation produced the opposite phenotype.

Another peptidergic system has been implicated in feeding in *Drosophila*, the approximately 20 neurons in the subesophageal ganglion that express the neuropeptide precursor gene *hugin* (*hug*) (**Figure 4**). The *hug* gene can give rise to two neuropeptides both of them activate a GPCR with similarities to the neuromedin U type receptor in mammals. The *hug* transcript is downregulated in starved larvae, correlating with increase in food-seeking behavior. Genetic interference with transmission in *hug*-expressing neurons shows that the *hug*-derived peptides are important for the initiation of feeding, depending on food quality, and that they act in circuitry that modulate feeding

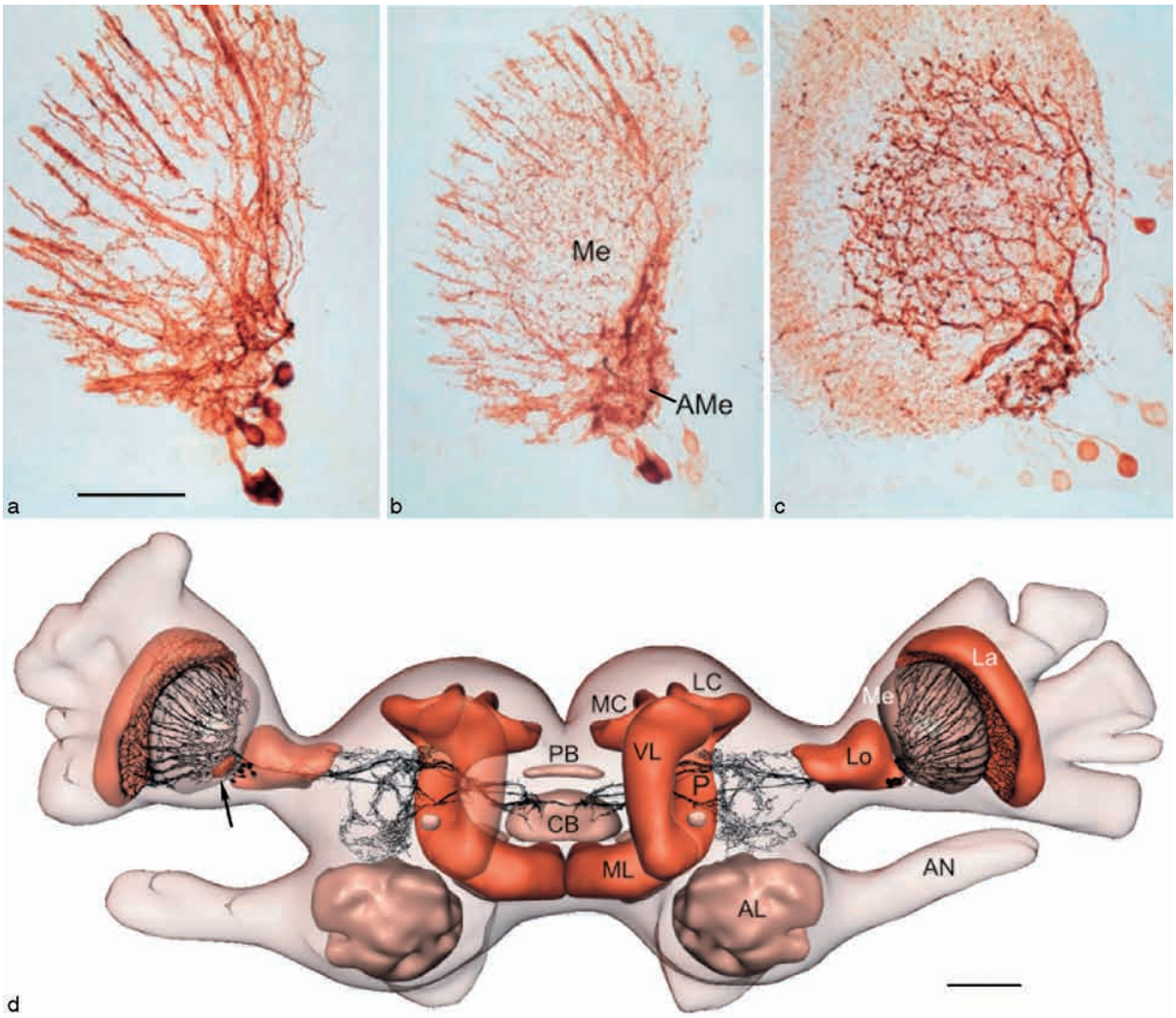

Figure 3 Neuropeptides in the optic lobe and brain of the cockroach *Leucophaea maderae*: (a) neurons immunostained with antisera against the peptide PDH; (b) neurons immunostained with antisera against the peptide FMRFamide; (c) neurons immunostained with antisera against the peptide orcokinin; (d) frontal reconstruction of PDH-immunoreactive neurons with cell bodies near the AMe (arrow) embedded into a three-dimensional reconstruction of the cockroach brain. The images in (a–c) are frontal views. The accessory medulla (AMe) is a circadian pacemaker center in the optic lobe. Acronyms label major neuropils of the brain. AL, antennal lobe; AN, antennal nerve; CB, central body; FMRFa, FMRFamide-like peptide (neuropeptide); La, lamina; LC, lateral calyx; Lo, lobula; MC, median calyx; Me, medulla; ML, median lobe; P, peduncle; PB, protocerebral bridge; PDH, pigment-dispersing hormone; VL, vertical lobe. Scale bar = 100 μm (a–c) and 200 μm (d). From Nässel DR and Homberg U (2006) Neuropeptides in interneurons of the insect brain. *Cell and Tissue Research* 326: 1–34. (d) Originally prepared by T Reischig (University of Marburg, Germany).

behavior based on chemosensory and nutrient signals. Anatomically, the *hug* neurons are well placed in the subesophageal ganglia to integrate chemosensory signals with behavioral output (Figure 4).

Complex interplay between hormonal peptides and neuropeptides in insects has been shown in the cascade of factors controlling molting in *Drosophila* and in the moth *Manduca sexta*. Molting entails the synthesis of new cuticle and the shedding of the old one; the latter requires certain patterns of peristaltic movements, referred to as preecdysis and ecdysis behaviors. Some of the peptides involved, eclosion hormone, corazonin, CCAP, and bursicon, are located in CNS neurons or neurosecretory cells; another ecdysis-triggering hormone is produced by epitracheal cells. These neuropeptide signaling systems act in parallel and in series to coordinate the precise sequence of behaviors that produces ecdysis.

Evolutionary Conservation of Peptide Signaling

Some invertebrate neuropeptides seem to be ancestrally related and can be found in several invertebrate

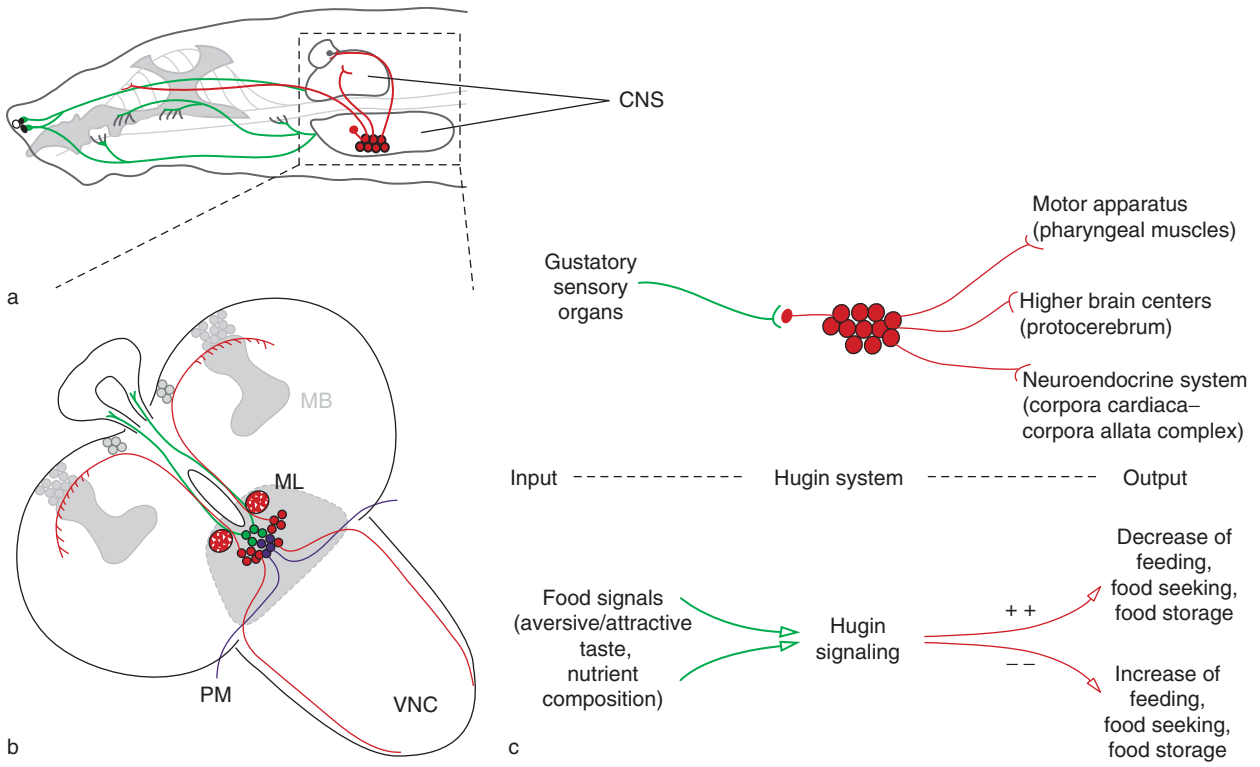

Figure 4 The *hugin*-derived peptides and feeding in *Drosophila*: (a) drawing of the head region of a *Drosophila* larva; (b) diagram of the larval CNS; (c) model of *hug* as a modulator of feeding behavior. In (a) are shown the external and internal chemosensory neurons in the antennomaxillary complex and internal mouth region innervating the larval CNS (depicted in green). *hug* neurons in the SOG (shown in red) project to the ring gland, pharyngeal muscle region, and the protocerebrum. In (b) are shown the distinct subpopulations of *hug*-expressing neurons projecting to the ring gland (neurohemal organ) only (green), pharynx only (blue), and the remaining targets (red). As shown in (c), the *hug* neurons, which express the neuropeptide gene *hug* and which interconnect gustatory sensillae via the SOG to the pharyngeal muscles, the protocerebrum, and the neuroendocrine organ, modulate chemosensory-dependent feeding behavior. Increased *hug* signaling correlates with decreased feeding, whereas decreased *hug* signaling correlates with increased feeding. CNS, central nervous system; MB, mushroom body; ML, subesophageal ganglion; PM, pharyngeal muscles; SOG, subesophageal ganglion; VNC, ventral nerve cord. From Melcher C and Pankratz MJ (2005) Candidate gustatory interneurons modulating feeding behavior in the *Drosophila* brain. *PLoS Biology* 3: 1618–1629, e305.

phyla from nematodes onward. A few may even be related to vertebrate neuropeptides. Peptide receptors appear even more conserved over evolution, and some ligand-receptor pairs may actually subserve related physiological functions from *C. elegans* to mammals – those of NPF-/NPY-like peptides and the insulin-like peptides. Other examples are insect sulfakinins and mammalian gastrin/cholecystokins and members of the vasopressin/oxytocin superfamily in mollusks and vertebrates. In the following, striking similarities in the insulin pathway are outlined.

There are extensive and functionally important similarities in structure of insulin-like peptides and their receptors in *C. elegans*, mollusks, insects, and vertebrates. The insulin superfamily consists of peptides such as vertebrate insulins, mammalian relaxin, and insulin-like growth factors (IGFs), as well as insulin-like peptides in invertebrates. The typical members of this superfamily consist of heterodimers of A and B chains of approximately 20 and 30 amino acids, respectively. In invertebrates, insulin-like peptides are mostly produced by neurons, by sensory neurons in *C. elegans*, and by neurosecretory cells in mollusks and insects.

The *C. elegans*, *Aplysia*, *Lymnaea*, and *Drosophila* insulin receptors share remarkable structural and functional properties with the insulin tyrosine kinase receptors of vertebrates. In addition, the components of the signaling pathway from insulin receptor substrate to glycogen synthase kinase are also highly conserved among *C. elegans*, *Drosophila*, and mammals. Also some of the signaling functions appear to be conserved during evolution. The mollusk insulin-like peptides (MIPs) are involved in the regulation of growth and associated metabolic processes, in the control of glycogen levels and gonadotropic activity, and in the promotion of neurite growth. Thus, MIPs combine the actions of insulin, IGFs, and relaxin.

In *C. elegans*, insulin signal transduction regulates longevity. There are 38 *C. elegans* genes encoding

insulin-like peptides (*ILP*) but only a single ancestral receptor *daf2*. Mutations in *daf2* produce pleiotropic effects, including increased fat storage, stress resistance, and entry into a diapuase-like dauer state. Available evidence suggests that stressful environmental conditions downregulate insulin signaling, allowing the promotion of a program of longevity.

In *Drosophila,* the insulin-like receptors and insulin signaling components are important in development, cell and organism growth, longevity, and female fertility. The role of insulin signaling components in extension of life span has especially been target of numerous genetic studies. Recent studies showed that interference with insulin signaling affects carbohydrate (trehalose) levels in the fly circulation. Thus, as in mollusks and mammals, the fly insulin-like signaling pathway is important for the control of metabolic carbohydrates.

Summary and Conclusion

A wide variety of neuropeptides has been identified from invertebrates of different organizational levels. In more evolved animals, peptides have defined hormonal roles in the initiation or regulation of specific behaviors or in the regulation of physiological functions critical in development, feeding, growth, reproduction, metabolism, and water balance. Neuropeptides also act as modulators of neuron or muscle properties. In rhythm-generating neural networks peptides trigger or modulate rhythmic patterns, and in muscle they change contractile properties to accommodate for changing patterns of rhythmic contractions. As in mammals, the concerted action of several related or unrelated neuropeptides, often in combination with fast neurotransmitters, is likely to be a common phenomenon in the invertebrate nervous system. We can also see that some signaling pathways are highly conserved throughout evolution. Not only are invertebrate neuropeptide signaling systems of evolutionary interest, some of the conserved ones, such as that of insulin-like peptides, may in fact be more tractable to molecular analysis in animals such as *C. elegans* and *Drosophila*.

See also: Comparative Biology of Invertebrate Neuromuscular Junctions; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Invertebrate Neurohormone GPCRs; Mammalian Neuropeptide Families; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Neuropeptide Y (NPY) and its Receptors; Neuropeptides: Food Intake; Olfaction in Invertebrates: Drosophila.

Further Reading

Barbieri M, Bonafe M, Fracheschi C, and Paolisso G (2003) Insulin/IGF-1 signaling pathway: An evolutionarily conserved mechanism of longevity from yeast to humans. *American Journal of Endocrinology* 285: E1064–E1071.

Brezina V and Weiss KR (1997) Analyzing the functional consequences of transmitter complexity. *Trends in Neuroscience* 20: 538–543.

Brogiolo W, Stocker H, Ikeya T, Rintelen F, Fernandez R, and Hafen E (2001) An evolutionary conserved function of the *Drosophila* insulin receptor and insulin-like peptides in growth control. *Current Biology* 11: 213–221.

de Bono M and Maricq AV (2005) Neuronal substrates of complex behaviors in *C. elegans*. *Annual Review of Neuroscience* 28: 451–501.

Grimmelikhuijzen CJP, Williamson M, and Hansen GN (2002) Neuropeptides in cnidarians. *Canadian Journal of Zoology* 80: 1690–1702.

Hewes RS and Taghert PH (2001) Neuropeptides and neuropeptide receptors in the *Drosophila* genome. *Genome Research* 11: 1126–1142.

Keating CD, Kriek N, Daniels M, et al. (2003) Whole-genome analysis of 60 G protein-coupled receptors in *Caenorhabditis elegans* by gene knockout with RNAi. *Current Biology* 13: 1715–1720.

Kits KS, Boer HH, and Joosse J (eds.) (1991) *Molluscan Neurobiology.* Amsterdam: North-Holland.

McVeigh P, Geary TG, Marks NJ, and Maule AG (2006) The FLP-side of nematodes. *Trends in Parasitology* 22: 385–396.

Melcher C and Pankratz MJ (2005) Candidate gustatory interneurons modulating feeding behavior in the *Drosophila* brain. *PLoS Biology* 3: e3051618–1629 [online].

Nässel DR (2002) Neuropeptides in the nervous system of *Drosophila* and other insects: Multiple roles as neuromodulators and neurohormones. *Progress in Neurobiology* 68: 1–84.

Nathoo AN, Moeller RA, Westlund BA, and Hart AC (2001) Identification of neuropeptide-like protein gene families in *Caenorhabditis elegans* and other species. *Proceedings of the National Academy of Sciences of the United States of America* 98: 14000–14005.

Nusbaum MP, Blitz DM, Swensen AM, Wood D, and Marder E (2001) The roles of co-transmission in neural network modulation. *Trends in Neuroscience* 24: 146–154.

Rogers C, Reale V, Kim K, et al. (2003) Inhibition of *Caenorhabditis elegans* social feeding by FMRFamide-related peptide activation of NPR-1. *Nature Neuroscience* 6: 1178–1185.

Taghert PH and Veenstra JA (2003) *Drosophila* neuropeptide signaling. *Advances in Genetics* 49: 1–65.

Wu Q, Wen T, Lee G, Park JH, Cai HN, and Shen P (2003) Developmental control of foraging and social behavior by the *Drosophila* neuropeptide Y-like system. *Neuron* 39: 147–161.

Neuropeptide Synthesis and Storage

J A Sobota, B A Eipper, and R E Mains, University of Connecticut, Farmington CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Small, biologically active peptides are crucial to intercellular communication in most multicellular organisms. Both neuropeptides and classical neurotransmitters are secreted at most nerve terminals. Neurons and endocrine cells have the ability to synthesize bioactive peptides, store them for long periods, and secrete them upon command; many other cells in the body also use peptides for communication, although storage in nonneuroendocrine cells is usually minimal. Peptides are used as neurotransmitters, neuromodulators, neurotoxins, autocrine and paracrine agents acting locally, and growth factors and hormones acting long distance. These peptides range in size from three amino acids (TRH) to more complex, disulfide-linked peptides such as insulin, with its A (21 amino acids) and B (30 amino acids) chains. All bioactive peptides share important common features in their biosynthesis, and these features affect the way in which peptides are stored for secretion in a regulated manner.

Steps in Peptide Biosynthesis

Peptides are synthesized as larger inactive precursors with an NH_2-terminal signal sequence that guides the insertion of the nascent chain into the lumen of the endoplasmic reticulum (ER), after which the signal sequence is removed (Figure 1). Peptide synthesis always begins in the cell body and requires the ER and the Golgi complex. This is a critical difference from the biosynthesis of small-molecule neurotransmitters such as glutamate, γ-aminobutyric acid (GABA), acetylcholine, and norepinephrine, which are synthesized at the nerve terminal and can be replenished locally after depletion during neurotransmission.

Some peptide precursors are glycosylated; N-linked oligosaccharide chains are added co-translationally (Figure 1, ER). The NH_2-terminal regions of many promolecules contain Cys residues that form intramolecular disulfide bonds; this process occurs in the ER. As intact promolecules proceed through the Golgi complex, N-linked oligosaccharides are modified, O-linked sugar chains can be added, and other modifications such as sulfation and phosphorylation can occur (Figure 1, Golgi). The same enzymes that carry out these early modifications of peptide precursors are also responsible for modifying integral membrane proteins and other secreted proteins; they are not unique to neuropeptide biosynthesis.

Depending on the promolecule and the cell type, the first enzymatic step unique to neuropeptide biosynthesis, endoproteolytic cleavage, may begin to occur at the distal, or exit, side of the Golgi complex in the *trans*-Golgi network (TGN) (Figure 1). In neurons and endocrine cells, the promolecules and a number of peptide-processing enzymes enter immature secretory granules, where a series of cell-type specific endoproteolytic cleavages proceed more rapidly. As the secretory granules mature, endoproteolytic cleavages continue, along with additional enzymatic modifications of the smaller peptides: COOH-terminal and NH_2-terminal exoproteolytic removal of basic amino acids, α-amidation at exposed COOH-terminal Gly residues, conversion of NH_2-terminal Gln residues into pyroglutamic acid, and NH_2-terminal acetylation. Many of these seemingly minor modifications are essential for biological activity.

For many neuropeptides, closely related precursors produce families of related product peptides; examples include the oxytocin and vasopressin precursors; the three precursors for opiate peptides, β-endorphin, methionine-enkephalin, and dynorphin; substance P; and related tachykinins. The information content encoded by neuropeptides is magnified by several factors. Nearly every peptide precursor is expressed in more than one tissue (e.g., procholecystokinin and proenkephalin in the gut and brain), and different sets of product peptides are often found in each tissue. For example, cholecystokinin (CCK-4) is produced in the brain, whereas larger molecules such as CCK-8 and CCK-33 are more prevalent in the gut; methionine-enkephalin is produced in the brain, whereas larger molecules such as peptides E and F are found in adrenal chromaffin cells.

To add further complexity, several bioactive peptides can be produced from a single precursor. The most widely studied example is proopiomelanocortin (POMC), which can give rise to corticotropin (which stimulates adrenal glucocorticoid production); melanotropins (which cause skin darkening and affect appetite); and β-endorphin, an opiate-active peptide. The anterior pituitary POMC-producing cells (corticotropes) produce corticotropin, but do not make the endoproteolytic cleavages that yield melanotropins or opiate-active peptides. The intermediate pituitary POMC-producing cells cleave corticotropin to

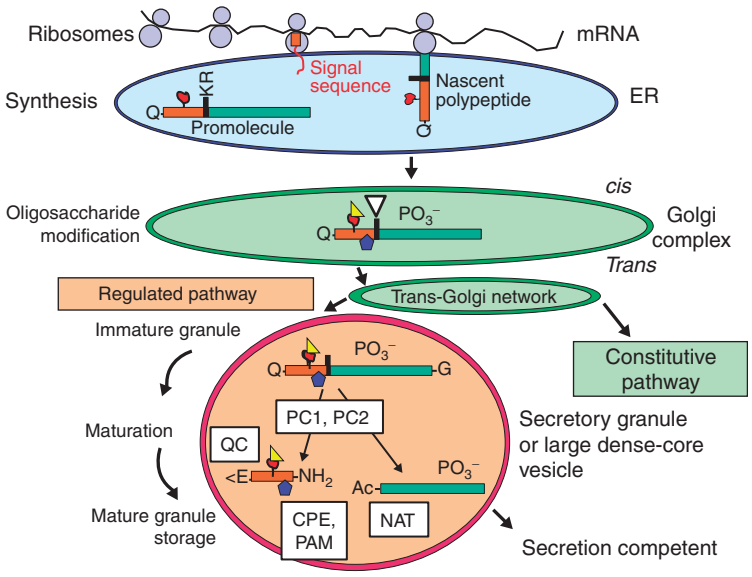

Figure 1 Steps common to the biosynthesis of neuropeptides. The major steps leading from mRNA encoding a preproneuropeptide to its final products are outlined. The subcellular location of step is indicated. <E, pyroglutamic acid; Ac, acetyl; CPE, carboxypeptidase E/H; ER, endoplasmic reticulum; G, glycine; KR, lysine-arginine; NAT, *N*-acetyl-transferase; PAM, peptidylglycine α-amidating monooxygenase; PC, propeptide/prohormone convertase; Q, glutamine; QC, glutaminyl cyclase.

produce melanotropin, and POMC neurons in the arcuate nucleus of the hypothalamus produce both β-endorphin and melanotropins.

Most of the seemingly minor posttranslational alterations of bioactive peptides are essential for full biological activity. Cholecystokinin is inactive on the gallbladder and pancreas unless a specific Tyr residue is sulfated; opiate peptides are inactivated when their NH_2-terminal Tyr residue is N-acetylated. In contrast, NH_2-terminal acetylation activates melanotropin, vasopressin is inactive without its COOH-terminal α-amide group, and gonadotropin-releasing hormone is inactive without its NH_2-terminal pyroglutamic acid residue. Finally, the pattern of precursor processing in a given tissue may be developmentally regulated. Anterior pituitary corticotropes cleave corticotropin to produce melanotropin activity until a time around birth, when adrenocortical function is necessary; at this time, corticotropin cleavage ceases, adrenocorticotropin (ACTH) is produced and glucocorticoid production is stimulated. Transient expression of many peptides and peptide-processing enzymes occurs during development, but the significance of this transient expression is not understood.

Diseases associated with peptide biosynthesis include disorders of protein folding within the ER. Unstable and hypofunctional gene products are produced as a result of mutations impairing prohormone folding. Retention and accumulation of mutant proteins within the ER severely challenges its folding capacity and generates an ER stress response. For example, the $Ins2^{C96Y}$ mutation found in the diabetic Akita mouse prevents the formation of an essential disulfide bond between the A and B chains of insulin, impeding proper folding and the processing of the mutant proinsulin. The misfolded proinsulin is retained in the ER of pancreatic β cells; apoptotic cell death of β cells correlates with the progression of diabetes and is thought to be a consequence of proteotoxicity to the ER. The pathogenesis of the human syndrome hereditary neurogenic diabetes insipidus is similarly thought to be related to ER proteotoxicity, caused by the accumulation of misfolded mutant arginine vasopressin (AVP) in neurons.

Precursor-Processing Enzymes

The enzymes responsible for peptide processing fall into two broad classes: those specific for peptide processing and those that are important in the secretory pathway for many other molecules as well as for bioactive peptides. The latter group includes the enzymes responsible for signal peptide cleavage, oligosaccharide attachment and maturation, O-glycosylation, disulfide bond formation, phosphorylation of luminal Ser and Thr residues, and sulfation of Tyr residues.

Endoproteolytic cleavage of prohormones frequently occurs at pairs of basic amino acids, although single Arg residues are not uncommon as sites of cleavage. The endoproteases responsible for these cleavages include a set of proteins structurally related to the bacterial subtilisins and to the yeast enzyme

Kex2. The members of this propeptide/prohormone convertase (PC) family include PC1/3, PC2, PC4, PC5/6A, PC7/8, paired basic amino acid cleaving enzyme (PACE)4, and furin. The subtilisins are all synthesized as inactive precursors with an NH_2-terminal proregion (Figure 2); the removal of the ~80-residue proregion is necessary for catalytic activity and occurs autocatalytically in several cases. PC1/3 and PC2 are enriched in peptide-containing secretory granules, where they cleave prohormones on the C-terminal side of pairs of basic residues (Lys or Arg). PC1/3 and PC2 cleave secretory granule proteins and are not thought to cleave other proteins in the secretory pathway. There is evidence supporting the participation of aspartyl and cysteinyl endoproteases in the biosynthesis of selected peptides.

Two enzymes expressed in most neurons and endocrine cells mediate the final events in peptide maturation: carboxypeptidase E/H (CPE) removes the basic residues exposed by the PCs; peptidylglycine α-amidating monooxygenase (PAM) converts COOH-terminal Gly residues into α-amide groups. Glutaminyl cyclase (QC), which converts NH_2-terminal Gln residues into pyroglutamic acid residues, modifies a small subset of peptides. Much less is known about the *N*-acetyltransferase (NAT) and aminopeptidase that function in secretory granules. Because they function late in the peptide biosynthetic pathway, these granule enzymes must be active at pH 5–6, the pH of mature secretory granules (Figure 2).

Most of the peptide-processing enzymes require a divalent metal ion for full activity: the PCs are CA^{2+}-dependent; CPE and QC require zinc; bifunctional PAM requires copper, zinc, iron, and calcium. PAM is unusual in that it requires a cofactor, ascorbic acid, which is oxidized as molecular oxygen is consumed in each reaction cycle. Cytochrome b_{561}, an integral membrane protein, shuttles reducing equivalents from the cytosol into the lumen of the secretory granule, replenishing the luminal supply of reduced ascorbate. Tyrosine sulfotransferase requires adenosine 3′-phosphate,5′-phosphosulfate (PAPS) as the active sulfate donor and Mn^{2+}. Neuropeptide expression, which varies in response to many stimuli, is generally controlled by regulating the transcription of mRNA encoding the preprohormone. The expression of peptide precursors is usually tightly coordinated with expression of the relevant peptide-processing enzymes. Thus, when the amount of peptide produced is increased, the pattern of postranslational processing is largely unaltered. Secretory rates can temporarily exceed the synthetic rates for peptides, but mRNA levels usually change in response to changes in secretory demand.

The activity of each PC is controlled by its proregion, by interactions with specific chaperones, and by the pH within different compartments of the secretory pathway (Figure 2). As for bacterial subtilisin, the proregions of PC1/3 and PC2 bind to their active sites. Autocatalytic cleavage of proPC1/3 occurs in

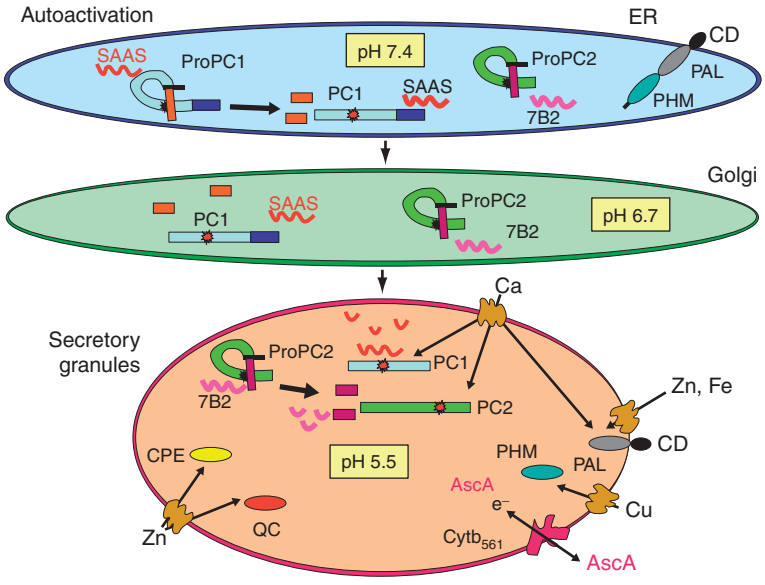

Figure 2 Activation of peptide-processing enzymes. Key steps in the activation of PC1 and PC2 are indicated, as are the various divalent metal ions which much be supplied to the PCs and to CPE, PHM, and PAL. Acidification of the luminal compartment is indicated. 7B2 and SAAS are small proteins that serve a mixed chaperone–inhibitor function for the PCs. AscA, ascorbic acid; CD, cytoplasmic domain; CPE, carboxypeptidase E/H; ER, endoplasmic reticulum; PAL, peptidyl-hydroxyglycine α-amidating lyase; PC, propeptide/prohormone convertase; PHM, peptidylglycine α-hydroxylating monooxygenase; QC, glutaminyl cyclase.

the ER. The proregion of PC2 is not cleaved until proPC2 reaches the immature secretory granules. The activities of PC1/3 and PC2 are also affected by their interactions with naturally occurring inhibitors, pro-SAAS and 7B2, respectively (Figure 2). In the case of PC2, pro7B2 interacts with proPC2 early in the secretory pathway, facilitating its transport from the ER to the TGN. Whereas the NH_2-terminus of 7B2 is essential for PC2 activation, the COOH-terminus of 7B2 inhibits its activity. Like the proregions of PC1 and PC2, the ~10 kDa-propeptide fragment of 7B2 acts as an inhibitor, binding at the active site of PC2. Only when these 8–10 kDa peptides are further cleaved do they lose their affinity for their target PC, allowing it to exhibit activity (Figure 2).

Processing Enzyme Knockouts

Targeted deletion of processing enzymes in the mouse has increased our knowledge of specific PC–substrate pairs. In some cases, redundant activities were identified via the examination of the peptides present in knockout mice. Disruption of the TGN-localized enzymes furin and PACE4 both result in embryonic lethality, with developmental defects associated with the impaired processing of essential inductive signaling molecules (transforming growth factor (TGF)-β, nodal, and bone morphogenetic protein (BMP) family members). In these cases, the mutant phenotypes partially overlap with, but are less severe than the phenotypes of the substrate knockouts, suggesting that a limited degree of processing enzyme redundancy exists. Mice lacking PC1 are viable, but have a severe block in proinsulin cleavage at the B–C chain junction. Although they are hyperproinsulinemic, they do not show abnormalities in glucose tolerance. Dwarfism occurs due to defective cleavage of the growth hormone-releasing hormone (GHRH) precursor. The PC2-null mice also accumulate proinsulin; cleavage at the A–C junction is impaired. Proglucagon processing is completely blocked, with progressive islet hyperplasia and chronic hypoglycemia. Other PC2 substrates accumulate in the brains of mutant animals: enkephalin, CCK, orphanin FQ/nociceptin, and dynorphin. Homozygous mutants for the amidating enzyme PAM die between embryonic day (E)14.5 and E15.5, with severe generalized edema. This phenotype is strikingly similar to the adrenomedullin (AM) knockout phenotype; the AM gene encodes two amidated peptides which function as cardiovascular regulators, and amidation is likely essential for their full bioactivity.

Naturally occurring mutations in processing enzymes are also known. In humans, a syndrome associated with PC1 deficiency has been reported; this syndrome includes extreme childhood obesity, abnormal glucose homeostasis, hypocortisolism, amenorrhea, and impaired gastrointestional (GI) function. In addition, CPE missense mutations have been reported in patients with type 2 diabetes. In mice, another CPE mutation was found (S202P); the mutant CPE ($CPE^{fat/fat}$) is inactive and rapidly degraded. These $CPE^{fat/fat}$ mice are infertile, hyperglycemic, and develop late-onset obesity. The phenotype of the true CPE null mice is similar, although more severe than that of the $CPE^{fat/fat}$ mice. Additional carboxypeptidase family members are probably responsible for the retention of reduced peptide processing that prevents lethality in these animals. Carboxypeptidase D (CPD), which is localized to the TGN and immature granules, probably plays a moderate role because it can only complete reactions initiated at the TGN.

Granule Biogenesis and Storage

Neuroendocrine cells producing proinsulin or POMC generally secrete equimolar amounts of all of the products derived from the promolecule. This means that product peptides with no clearly defined function (e.g., C-peptide derived from proinsulin) are secreted along with the active peptide(s). However, the separation of product peptides into different granules and differential secretionhas been observed in other cell types.

Following the synthesis and translocation of nascent polypeptide chains into the lumen of the ER, secretory proteins are transported through the Golgi complex to the TGN. Once this sorting station is reached, vesicles destined for different subcellular compartments acquire the correct cargo. In all cell types, a constitutive pathway delivers soluble secretory proteins directly to the plasma membrane, where they are released without the need for a stimulus. In addition to the constitutive pathway, a regulated pathway or pathways exist in neurons and endocrine cells; regulated proteins are stored in secretory granules until released in response to an incoming signal. Two models, not mutually exclusive, provide a framework for thinking about the mechanisms by which constitutive and regulated soluble proteins are targeted to their respective pathways for secretion.

One model, known as sorting for entry, suggests that granule content proteins bind selectively to a membrane receptor in the TGN (Figure 3). This could involve either direct binding to a membrane component of the TGN such as cholesterol-rich membrane microdomains (i.e., lipid rafts) or interaction with a specific sorting receptor. Studies on PC2 support a role for lipid rafts in its entry into the regulated

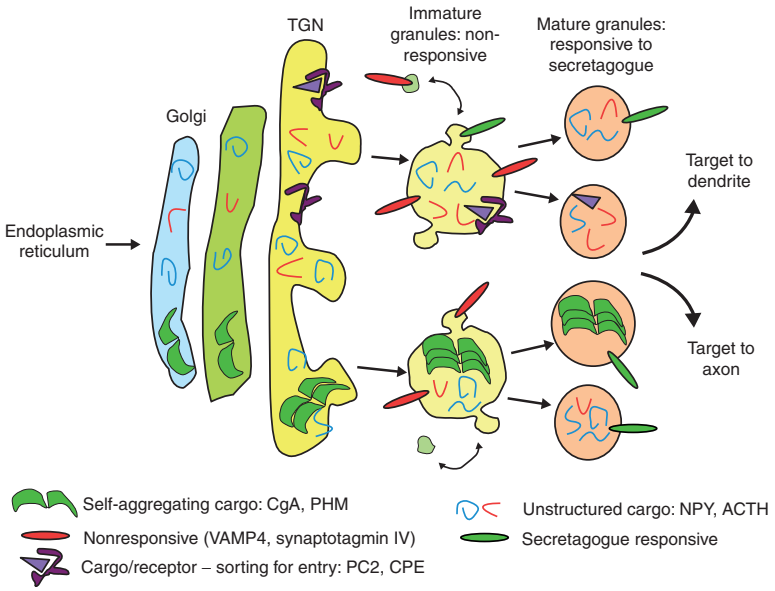

Figure 3 Intracellular sorting of neuropeptides and enzymes to different vesicle populations. Although some sorting begins in the ER/Golgi, most occurs in the TGN and immature secretory granules which undergo significant remodeling. Some cargo is unstructured and mixes with other unstructured cargo; other cargo self-aggregates, excluding other molecules. There is evidence for cargo receptors (proteins and lipid rafts) which steer molecules into regulated granules. A number of proteins present in immature secretory granule membranes must be removed for mature secretory granules to function properly. ACTH, adrenocorticotropin; CgA, chromogranin A; CPE, carboxypeptidase E/H; ER, endoplasmic reticulum; NPY, neuropeptide Y; PC, propeptide/prohormone convertase; PHM, peptide histidine-methionine; TGN, *trans*-Golgi network; VAMP4, vesicle-associated membrane protein 4.

pathway. It has been proposed that CPE binds prohormone aggregates in the TGN and, through its association with lipid rafts, serves as a sorting receptor-mediating prohormone entry into regulated granules. Proteins unable to bind to the sorting receptor enter the constitutive pathway by default.

A second sorting model, known as sorting by retention, is based on the idea that protein entry into immature granules occurs by bulk flow; the retention of regulated cargo is facilitated by the aggregation and condensation of these proteins. Nonaggregating cargo proteins may be retained through membrane interaction. In professional secretory cells, in which over 10% of the total protein synthesis may be one preprohormone, it is difficult to envision a specific receptor with a large enough capacity to accomplish sorting for entry.

Granule trafficking of peptide-processing enzymes with transmembrane domains (PAM, CPD, and PC6) is mediated by signals in their cytoplasmic domains and can be regulated by phosphorylation. Whereas soluble processing enzymes are secreted along with their product peptides, integral membrane-processing enzymes can be retrieved from immature granules during maturation and from the endocytic pathway following exocytosis. Remodeling of the membranes of immature granules is an essential part of the granule maturation process and must occur before secretion can be stimulated. Clathrin is recruited to immature granules by the adaptor protein Golgi-associated γ-ear-containing ADP-ribosylation-factor-binding-protein (GGA), and the selective removal of nongranule proteins occurs via budding of clathrin-coated vesicles. In this manner, the immature granule functions as a more advanced sorting station, succeeding the TGN. Experimental support for this type of remodeling process comes from the progressive removal of transfected exocrine proteins from granules in AtT-20 cells, as well as the transient presence of lysosomal proteins in immature granules.

Although a conserved sorting signal does not exist for regulated secretory proteins, sequence and structural similarities have been found in subsets of proteins. An N-terminal sorting signal comprising a hydrophobic domain, characteristic amino acid content, and amphipathic α-helix has been identified. Structural features of the sorting domain, such as N-terminal disulfide-bonded loops, have been proposed to be important for stabilizing its structure and may contribute to the routing of POMC, chromogranin B, and chromogranin A. Although loop-mediated homodimerization of chromogranin A is required for its sorting into the regulated secretory pathway, there is disagreement on the importance of the disulfide loops of POMC.

A key feature in the sorting of regulated secretory proteins is their ability to selectively aggregate

(Figure 3). Aggregation may begin as early as the ER and may result in homotypic aggregates of secretory proteins. For pancreatic exocrine proteins, the majority of condensation occurs between the ER and *cis*-Golgi; aggregated atrial natriuretic peptide (ANP) has been identified in all layers of the Golgi complex. Other granule proteins, most notably members of the granin family, are known to aggregate when exposed to the slightly acidic (pH <6.5), high calcium (millimolar) conditions in the TGN. Low pH is also important for the aggregation of CPE, insulin, prolactin, POMC, and proPC2; the aggregation of POMC also requires Ca^{2+}, and the aggregation of prolactin requires Zn^{2+}. The aggregation conditions required for each protein may be related to its isoelectric point. Constitutively secreted proteins are excluded from these aggregates. The granins are thought to function as aggregation vehicles and could act by binding or co-aggregating with other secretory proteins. It is in this way that chromogranin A may act as an assembly factor that initiates aggregation. Finally, the ability of some granule content proteins to generate granulelike structures when expressed in cells lacking a regulated secretory pathway is consistent with an important role for aggregation.

The mechanisms by which integral membrane-processing enzymes are targeted to the regulated pathway are fundamentally different from those governing soluble proteins. The cytosolic domains of granule membrane proteins contain information essential for routing. For example, sorting of the vesicular monoamine transporter (VMAT) to secretory granules requires its COOH-terminal extended dileucine motif, suggesting an interaction with adaptor proteins. PAM proteins with a truncated cytoplasmic domain localize to the plasma membrane instead of granules. Cytosolic proteins that interact with PAM and regulate its trafficking include Rho guanosine diphosphate (GDP)/guanosine triphosphate (GTP) exchange factors (kalirin and trio) and a Ser/Thr kinase (P-CIP2). The cytoplasmic domain of PAM is multiply phosphorylated and cycles of phosphorylation/dephosphorylation are essential for normal trafficking within both the biosynthetic and endocytic pathways. Furin enters immature secretory granules, but is excluded from mature granules, returning to the TGN; furin also cycles between the plasma membrane and endosomes. Both the TGN localization of furin and its retrieval from endosomes are mediated by phosphofurin acidic cluster-sorting protein (PACS)-1, which recognizes furin that has been phosphorylated by casein kinase (CK)II. This interaction facilitates binding of adaptor protein (AP-)1, which mediates the entry of furin into clathrin-coated vesicles. PACS-1 also controls the endosome to TGN trafficking of CPD. Like furin, CPD cycles between the TGN and cell surface; the binding of adaptor proteins to its 58-residue cytosolic tail is required for trafficking, which is dependent on the phosphorylation of its CKII sites.

The dynamic regulation of secretory granule membrane composition also plays a role in the acquisition of regulatability. Secretagogue application does not stimulate the exocytosis of immature granules. The maturation of immature secretory granules involves the removal of vesicle-associated membrane protein (VAMP)4 and synaptotagmin IV; removal is an ADP-ribosylation factor (ARF)-dependent process. Remodeling of granule membranes is inhibited by brefeldin A, causing granules to remain in a stimulus-insensitive state. Mature granules acquire the ability to be stimulus responsive. Removal of synaptotagmin IV, a potential inhibitor of Ca^{2+} triggered exocytosis, may be the key to establishing competence to undergo regulated exocytosis. Retrieval of VAMP4 requires phosphorylation-dependent recruitment of AP-1 to immature granules, a process mediated by PACS-1, suggesting that the removal of these inhibitory proteins from immature granules is likely to occur via clathrin-coated vesicles.

Peptides in the Nervous System

Although most work on the synthesis and storage of neuropeptides has been performed in endocrine and neuroendocrine primary cell cultures and cell lines, the general principles apply to primary neurons. In both the central and peripheral nervous systems, neuropeptides have a widespread distribution and are processed primarily by the endoproteases PC1/3 and PC2. The majority of neurons coexpress peptides along with classical low-molecular-weight neurotransmitters such as glutamate, GABA, acetycholine, and biogenic amines. The two types of neurotransmitter are stored in different subcellular compartments within the same neuron. For neuropeptides, secretory granules (also known as large dense-core vesicles) are the only site of storage. These granules are localized to the neuronal cell soma, axons, and dendrites. In contrast to neuropeptides, small-molecule neurotransmitters are localized to small synaptic vesicles, which are concentrated at axon terminals (Figure 4).

One major functional implication of this differential distribution of peptides compared to small-molecule transmitters is that release is selective. Whereas the release of small-molecule neurotransmitters is triggered by a small localized increase in cytoplasmic Ca^{2+}, the release of neurotransmitter from large dense-core vesicles requires a greater increase in Ca^{2+} in the terminal. Low-frequency stimulation may

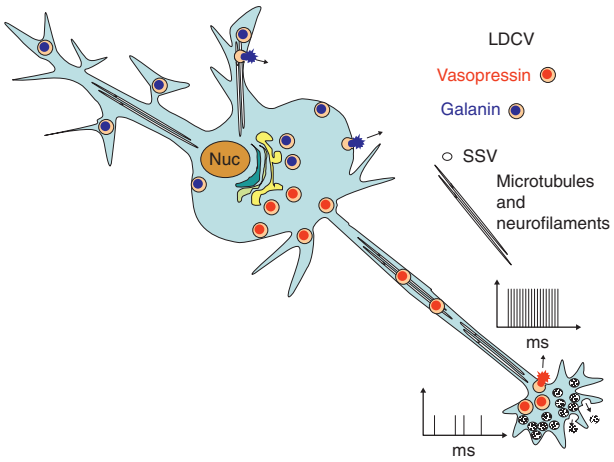

Figure 4 Peptidergic transmission in neurons. Most neurons contain large dense-core vesicles (secretory granules) and small synaptic vesicles (SSVs). The SSVs can be emptied and refilled at the nerve terminal while the large dense-core vesicles, which contain peptides, cannot be refilled. When two peptide precursors are expressed in one cell, the products can be packaged in the same granules or packaged into separate granules that are trafficked to different parts of the cell. LDCV, large dense-core vesicles; Nuc, nucleus. Adapted from Landry M, Vila-Porcile E, Hokfelt T, and Calas A (2003) Differential routing of coexisting neuropeptides in vasopressin neurons. *European Journal of Neuroscience* 17: 579–589.

release only small synaptic vesicles, whereas high-frequency stimulation releases a combination of peptides and small-molecule neurotransmitters. Consistent with the widespread distribution of peptide receptors, peptide release can occur at a significant distance from any synapse. Peptides can diffuse a significant distance after release and often mediate longer-lasting responses than classical neurotransmitters. Furthermore, depending on the stimulus, peptide release from secretory granules may be axonal or somatodendritic and may occur in either a coordinated or independent manner.

Secretion from axons versus dendrites provides neuropeptides to different target tissues and produces different physiological effects. For example, thyrotropin-releasing hormone (TRH) neurons in the hypothalamic paraventricular nucleus (PVN) project axons to the median eminence, where TRH entry into the hypophysial-portal system stimulates TSH release from the anterior pituitary. Intra-PVN release of TRH from the same neurons may contribute to the regulation of sympathetic outflow during cold stress. In the magnocellular neurons of the supraoptic nucleus (SON), local oxytocin release is stimulated by suckling before a significant peripheral secretion of oxytocin occurs. After osmotic stimulation, the release of oxytocin and AVP within the SON lags behind peripheral secretion.

Neurons often express more than one neuropeptide precursor. The products of the two precursors may be stored in separate granules or may be stored in the same granules; co-storage usually implies co-release and co-transmission. One such example is the primary sensory neurons of the dorsal root and trigeminal ganglia; tachykinins co-localize with calcitonin gene-related peptide (CGRP) in neuronal cell bodies and at the central and peripheral terminals. In this case, there does not appear to be a selective transport mechanism to form different granules to be shipped to functionally distinct processes. CCK and corticotropin-releasing hormone (CRH) are similarly co-stored in the periportal nerve terminals in the rat median eminence. In contrast, in hypothalamic magnocellular neurons, galanin and vasopressin, are only partially co-stored; galanin-containing granules are preferentially targeted to dendrites whereas vasopressin-containing granules are targeted to axons that extend into the posterior pituitary. These neurons have regions of their ER devoted to the synthesis of vasopressin and other regions devoted to galanin synthesis; separation at this early stage may promote packaging into separate granules, which would allow targeting to dendrites versus axons. The bag cell neurons of *Aplysia* contain granules filled with peptides derived from the N-terminal part of the egg-laying hormone precursor and granules filled with peptides derived from its C-terminal region; cleavage of the egg-laying hormone precursor occurs before exit from the Golgi complex. Granules containing N-terminal-derived bag cell peptides are released locally, whereas granules containing C-terminal-derived egg-laying hormone are released into the circulation.

Genes Controlling Expression of Granules and Granule Proteins

That a master gene may control secretory granule biogenesis and the expression of granule proteins has been proposed as a mechanism by which these processes are coordinately regulated. Although the idea is controversial, the acidic glycoprotein chromogranin A (CgA) has been put forth as a candidate master molecule responsible for regulating the initiation of secretory granule biogenesis. Antisense RNA-mediated reductions of CgA expression in PC12 cells result in proportional granule depletion, with a loss of regulation of other content proteins. Consistent with this, a transgenic mouse line with antisense-mediated reduction in CgA expression resulted in aberrant adrenal chromaffin cell granules. In addition, the expression of CgA in nonendocrine cells induced the formation of granulelike structures. However, analysis of one line of CgA knockout mice revealed no

structural abnormalities in adrenal chromaffin cells; vesicle size and number were normal, with upregulation of other granin proteins probably compensating for the CgA deficiency. In an independent line of CgA knockout mice, there were significant changes in chromaffin granule number and appearance, along with marked hypertension, but no general endocrine problems that would be expected if CgA were crucial to the function of all large dense-core vesicles. The fact that not all cells with a regulated secretory pathway contain CgA (e.g., pituitary sommatotropes and mammotropes, which have CgB) also argues against an essential role for CgA for all peptide-secreting cells. CgA may serve as an assembly factor; low pH and Ca^{2+} clearly induce CgA aggregation in the TGN, and aggregated CgA may specifically bind to or co-aggregate with cargo proteins. An additional role for chromogranin A in granule biogenesis may be to protect regulated secretory proteins from degradation by upregulating the expression of protease nexin-1, an endogenous serine protease inhibitor.

In *Drosophila*, the neuropeptidergic phenotype is under the control of the basic helix-loop-helix protein DIMM, a member of the atonal family. DIMM is related to similar molecules in vertebrates such as neuroD, neurogenin, and olig. Recently, dual functionality has been proposed for DIMM. In addition to being a critical regulator of neuropeptide gene expression in differentiating cells, spatiotemporal manipulation of the *dimm* gene revealed that DIMM regulates the expression of neuropeptides and peptide histidine-methionine (PHM) in mature neurons. This ability to regulate secretory protein expression suggests an effective mechanism by which peptidergic signaling may be maintained and dynamically adjusted in fully differentiated cells.

See also: Large Dense Core Vesicles (LDCVs); Mammalian Neuropeptide Families; Neuropeptide Inactivation or Metabolism; Neuropeptide Release; Neuropeptides in Autonomic Neurons; Neuropeptides: Enteric Nervous System.

Further Reading

Arvan P (2004) Secretory protein trafficking: Genetic and biochemical analysis. *Cell Biochemistry and Biophysics* 40: 169–178.

Borgonovo B, Ouwendijk J, and Solimena M (2006) Biogenesis of secretory granules. *Current Opinion in Cell Biology* 18: 365–370.

Czyzyk TA, Morgan DJ, Peng B, et al. (2003) Targeted mutagenesis of processing enzymes and regulators: Implications for development and physiology. *Journal of Neuroscience Research* 74: 446–455.

Day R and Gorr SU (2003) Secretory granule biogenesis and chromogranin A: Master gene, on/off switch or assembly factor? *Trends in Endocrinology and Metabolism* 14: 10–13.

Eaton BA, Haugwitz M, Lau D, and Moore HP (2000) Biogenesis of regulated exocytotic carriers in neuroendocrine cells. *Journal of Neuroscience* 20: 7334–7344.

Fisher JM, Sossin W, Newcomb R, and Scheller RH (1988) Multiple neuropeptides derived from a common precursor are differentially packaged and transported. *Cell* 54: 813–822.

Fortenberry Y, Hwang JR, Apletalina EV, and Lindberg I (2002) Functional characterization of ProSAAS: Similarities and differences with 7B2. *Journal of Biological Chemistry* 277: 5175–5186.

Fricker LD and Leiter EH (1999) Peptides, enzymes and obesity: New insights from a "dead" enzyme. *Trends in Biochemical Sciences* 24: 390–393.

Hewes RS, Gu T, Brewster JA, Qu C, and Zhao T (2006) Regulation of secretory protein expression in mature cells by DIMM, a basic helix-loop-helix neuroendocrine differentiation factor. *Journal of Neuroscience* 26: 7860–7869.

Landry M, Vila-Porcile E, Hokfelt T, and Calas A (2003) Differential routing of coexisting neuropeptides in vasopressin neurons. *European Journal of Neuroscience* 17: 579–589.

Merighi A (2002) Costorage and coexistence of neuropeptides in the mammalian CNS. *Progress in Neurobiology* 66: 161–190.

Prigge ST, Mains RE, Eipper BA, and Amzel LM (2000) New insights into copper monooxygenases and peptide amidation: Structure, mechanism and function. *Cellular and Molecular Life Sciences* 57: 1236–1259.

Ron D (2002) Proteotoxicity in the endoplasmic reticulum: Lessons from the Akita diabetic mouse. *Journal of Clinical Investigations* 109: 443–445.

Seidah NG and Prat A (2002) Precursor convertases in the secretory pathway, cytosol and extracellular milieu. *Essays in Biochemistry* 38: 79–94.

Taylor NA, Van de Ven WJ, and Creemers JW (2003) Curbing activation: Proprotein convertases in homeostasis and pathology. *FASEB Journal* 17: 1215–1227.

Thiele C and Huttner WB (1998) Protein and lipid sorting from the trans-Golgi network to secretory granules-recent developments. *Seminars in Cell and Developmental Biology* 9: 511–516.

Thomas G (2002) Furin at the cutting edge: From protein traffic to embryogenesis and disease. *Nature Reviews Molecular Cell Biology* 3: 753–766.

Zhu YL, Conway-Campbell B, Waters MJ, and Dannies PS (2002) Prolonged retention after aggregation into secretory granules of human R183H-growth hormone (GH) a mutant that causes autosomal dominant GH deficiency type II. *Endocrinology* 143: 4243–4248.

Neuropeptide Y (NP Y) and its Receptors

D R Gehlert, Eli Lilly and Company, Indianapolis, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Pancreatic polypeptide (PP) was originally isolated as a contaminant observed during the preparation of insulin. Subsequently, neuropeptide Y (NPY) was isolated from pig brain and named for the presence of C- and N-terminal tyrosines, while peptide YY (PYY) was isolated from pig intestine. While NPY is primarily localized to neurons and PP is found mainly in endocrine pancreas, PYY is found in both endocrine cells and neurons and, therefore, may exhibit neurotransmitter functions as well. While these peptides exhibit diversity in their expression, they share a common three-dimensional structure. Amino acid residues 1–8 form a type II proline helix followed by a loop, after which residues 15–32 form an α-helix, and the four C-terminal residues are in a flexible loop conformation. When comparing the amino acid sequences from a variety of mammals, only 2 of the 36 amino acids of NPY are variable. As such, NPY is one of the most evolutionary conserved peptides known. On the other hand, the PYY sequence exhibits eight variable amino acids between different orders of mammals. PP is one of the least conserved peptides known and appears to have undergone rapid and recent evolution. Despite the limited conservation observed in the PP and PYY amino acid sequences, the general three-dimensional structure (Figure 1) is maintained across a wide variety of species, indicating its potential importance in function.

Expression and Regulation of NPY

The NPY gene is composed of four exons and results in the synthesis of a 97-amino-acid preproNPY and is located on human chromosome 7 at the locus 7p15.1. PreproNPY is proteolytically processed by a prohormone convertase into the C-terminal peptide of NPY (CPON) and NPY1-39. This is further processed by carboxypeptidase to NPY1-37. The processing and C-terminal amidation of NPY is accomplished by peptidylglycine α-amidating monooxygenase and is essential for biological activity. NPY can be further processed to NPY3-36 and NPY2-36 by two enzymes, dipeptidyl peptidase IV and aminopeptidase P, respectively. While the functional significance of CPON is unknown, the C-terminal fragments of NPY have some selectivity for the Y2 receptor subtype.

In the rat brain, NPY is found in numerous brain regions including the hypothalamus, amygdala, hippocampus, nucleus of the solitary tract, locus coeruleus, nucleus acumbens, and cerebral cortex. NPY is colocalized with norepinephrine, γ-aminobutyric acid (GABA), and somatostatin in agouti-related protein (AGRP) containing neurons of the hypothalamus. In addition, NPY is found in peripheral neurons including sympathetic neurons where it is co-stored and co-released with norepinephrine. The adrenal medulla is the primary source of circulating NPY, though it is also expressed in liver, heart, kidney, spleen, pancreas, platelets, and endothelial cells of blood vessels.

The human PYY and PP genes are found 10 kb apart on chromosome 17q21.1. Like NPY, the PYY and PP genes are organized into four exons with three introns. The rat preproPYY sequence is 98 amino acids, consisting of a preceding signal peptide sequence and followed by a 31-amino-acid C-terminal extension peptide called CPOP. PYY is found primarily in the L cells of the gastrointestinal tract, with the highest concentrations found in the rectum, followed by the ileum and the colon. Like NPY, PYY is cleaved by dipeptidyl peptidase to the 3-36 fragment that is thought to be the principal active circulating peptide. PYY and PYY3-36 are stored in the gut mucosal endocrine cells, and both are found in the circulation. Rat preproPP has a signal peptide sequence and a 30-residue C-terminal extension peptide. PP is found primarily in the pancreas and is localized to the islets of Langerhans.

NPY synthesis and secretion are regulated by a number of hormones and neurotransmitters. At the cellular level, activation of either protein kinase A or protein kinase C will result in increased NPY synthesis and secretion. Catecholamines and muscarinic agonists will increase NPY, while GABA agonists will decrease it. Glucocorticoids, growth hormone, ghrelin, and gonadal steroids are hormones that increase NPY synthesis and secretion in the hypothalamus, while insulin, leptin, and retinoic acid will decrease it. NPY synthesis is also increased by a variety of growth factors such as nerve growth factor, insulin-like growth factor-I and brain-derived neurotrophic factor in several brain regions. Of particular interest is the increase in brain NPY observed following treatment of experimental animals with antidepressants, amphetamines, and lithium, indicating that it may play a role in the clinical response to these agents.

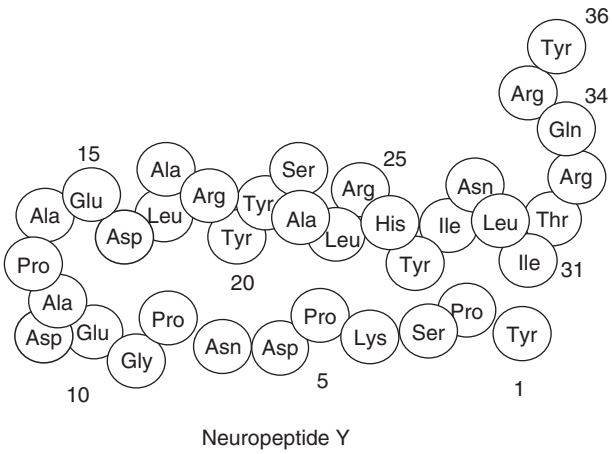

Figure 1 Representation of the three-dimensional structure of NPY.

Functional Activities of NPY

When administered directly into the brain, NPY produces marked increases in food intake and a preference for carbohydrate-rich food. This effect has been seen in a number of species ranging from fish to rhesus monkey. Exogenous NPY is the most potent and efficacious orexigenic peptide known. Chronic central administration to rats results in increased adiposity and the symptomology of the metabolic syndrome. Centrally administered NPY also decreases energy expenditure to promote adiposity. Less is understood about the role of endogenous NPY in the regulation of feeding. NPY levels are changed in experimental animals exhibiting a disturbed energy balance such as that seen in anorexia, bulimia neurosa, and diabetes. Starvation increases the concentrations of NPY and NPY mRNA in the hypothalamus, and, when administered centrally to rats, antisense oligonucleotides decrease feeding. NPY produces a number of other centrally mediated effects including anticonvulsant activity, modulation of arousal, anxiolysis, and modulation of cognition. Central administration of NPY produces antidepressant-like effects in animal models, and chronic administration of antidepressant drugs affects hippocampal NPY expression. NPY also acts with the central and peripheral nervous systems to decrease neuronal excitability. NPY has been reported to decrease excitability *in vitro* using hippocampal and dorsal root ganglion preparations. Mice lacking NPY are more susceptible to seizures, and people with temporal lobe epilepsy have increased NPY expression in CA3 regions as well as prominent rearrangements in receptor density and distribution. Furthermore, NPY-deficient mice were found to have an impaired development of neurons in the olfactory epithelium, suggesting a role in neuronal development. Central NPY also regulates ethanol consumption in rodents. In transgenic mice, an inverse correlation between NPY levels and drinking has been shown. Furthermore, a point mutation (leucine[7] to proline) in the preproNPY gene has been associated with higher alcohol consumption in humans and linkage analysis suggesting that dysfunction in the NPY gene locus contributes to ethanol preference in rats.

In addition to its role in feeding, hypothalamic NPY is important in the regulation of reproductive function. NPY acts at both the hypothalamic and pituitary levels to regulate reproductive hormone release. In the presence of estrogen, NPY stimulates leutinizing hormone (LH), while with low levels of estrogen, NPY will inhibit LH release. Continuous central infusion of NPY inhibits LH release in both male and female rats. NPY also exhibits a steroid-dependent regulation of gonadotropin-releasing hormone (GnRH) release. As hypothalamic NPY release is regulated by a vast number of neurotransmitters and hormones, NPY may be an important integrator for the multiple signals that regulate reproductive status.

NPY is frequently co-localized with noradenaline (NA) in sympathetic nerves and in the peripheral nervous system. Exogenous NPY enhances NA-mediated vasoconstriction, especially upon strong stimulation. In addition, NPY appears to stimulate smooth muscle proliferation and is important in angiogenesis and revascularization of ischemic tissues. Centrally administered NPY reduces arterial blood pressure and heart tone in both rats and dogs. Several cardiovascular dysfunctions as well as some tumor diseases are associated with increased plasma levels of NPY. In addition, NPY can act as an antinociceptive peptide, probably through inhibition of substance P release in the dorsal horn of the spinal cord.

PYY and PYY3-36 Function

PYY is the found primarily in the periphery and is released from the gastrointestinal (GI) tract in response to meals. Many of the GI effects described for PP, including inhibition of gall bladder secretion, gut motility, and pancreatic secretion, can also be produced by PYY. Other peripheral effects produced by PYY include the inhibition of fluid and electrolyte secretion in the intestinal tract and direct vasoconstriction. When administered centrally, PYY increases food intake with potency and efficacy similar to that seen with NPY. Via dipeptidyl peptidase IV, PYY is posttranslationally processed to PYY3-36, which appears to be the primary circulating form of the peptide. Peripherally administered PYY3-36 produces a potent decrease in food intake in both human and rodents.

PP Function

PP is almost exclusively expressed in endocrine pancreas and is released in response to meals. PP produces its biological effects mainly in the GI tract, where it inhibits pancreatic secretion, gall bladder activity, intestinal motility, ileum contractions, and gastric emptying and stimulates colon contractions. In addition, PP affects metabolic functions including glycogenolysis and decreases fatty acid levels. However, binding sites for PP have been found in several rat brain regions, including the interpeduncular nucleus, hypothalamus, and brain stem, suggesting that PP may also have direct effects of brain function. This may explain why intracerebroventricular (ICV) injection of PP has been found to stimulate feeding in several different species of experimental animals. In contrast, peripherally administered PP inhibits food intake in humans and rodents.

Receptors for the NPY Family of Peptides

Four functional G-protein-coupled receptors (GPCRs) have been identified that bind the NPY family of peptides, termed Y1, Y2, Y4, and Y5 (Table 1). An additional y6 receptor has been identified as a functional receptor in some species including mice and rabbits; however, it does not encode a functional protein in humans and is absent in rats. Y1, Y2, and Y5 bind NPY and PYY with similar high affinities. PP binds to Y4 with higher affinity than the other peptides. Y1 and Y2 are the predominant brain receptor populations (Figure 2). All the NPY receptors have been shown to couple to Gi and thus mediate inhibition of cAMP synthesis. When expressed in Chinese hamster ovary (CHO) cells, intracellular signaling involves pertussis toxin-sensitive phosphorylation of extracellular signal-regulated protein kinase (ERK)-1 and -2, confirming that these receptors couple to Gi and/or Go. In addition, Y1 activates mitogen-activated protein kinase in gut epithelial cells (IEC-6). Y1, Y2, Y4, and Y5 receptors can couple to phospholipase C to provoke release of Ca^{2+} from intracellular stores.

Table 1 Y receptors and their proposed functions

Receptor	*Agonist actions*
Y1	Angiogenesis, anxiolysis, cellular proliferation, circadian rhythm regulation, endocrine regulation, increase feeding, sedative, vasoconstriction
Y2	Angiogenesis, anticonvulsant, anxiogenesis, cellular proliferation decrease feeding, gastrointestinal motility, and secretion
Y4	Decrease feeding, gastrointestinal regulation
Y5	Anticonvulsant, circadian rhythm regulation, increase feeding

Y1 Receptor

The gene for Y1 is located in a cluster together with Y2 and Y5 on human chromosome 4q31. Y1 is primarily localized to the brain, peripheral nervous system, and peripheral blood vessels. The highest Y1 levels in the brain are found in the anterior thalamus, cerebral cortex, medial geniculate, and amygdala. In the peripheral nervous system, Y1 mRNA is found in the superior cervical and dorsal root ganglion. In the vasculature, Y1 is localized to intramyocardial, colonic, and renal blood vessels with the expression being localized to the intima as well as the media. Y1 mRNA has been detected in a number of peripheral tissues including the colon, kidney, adrenal gland, heart, and placenta. Within the kidney, Y1 mRNA is detected in the renal collecting ducts, loop of Henle, and the juxtaglomerular apparatus.

Y1 receptor is important in the feeding response produced by exogenous NPY. Central injection of Y1-selective agonists increase feeding, while Y1 antagonists can inhibit NPY-induced feeding. Most of the vasoconstriction produced by NPY and PYY is transduced via the Y1 receptor, as shown by the inhibitory affects of various Y1-selective antagonists on NPY-induced vasoconstriction and the absence of effects in Y1-deficient mice. Many of the other central effects of NPY have been attributed to the Y1 receptor, including reduction of blood pressure and heart rate, anxiolysis, and antidepressant activity. Stimulation of brain Y1 receptors inhibits the consumption of ethanol by rodents, inhibits neurogenic inflammation, and is antinociceptive.

Y2 Receptor

The Y2 gene is located near the Y1 gene on chromosome 4 and encodes a 381-amino-acid protein. Y2 has more than 90% amino acid identity within mammals, although Y2 is only about 30% identical to the Y1 and Y4 receptors. Y2 is present in the central and peripheral nervous system, the intestine, and certain blood vessels. Brain Y2 is pronounced in the hypothalamus, lateral septum, hippocampus, and amygdala. In the peripheral nervous system, Y2 is localized to the superior cervical ganglion and dorsal root ganglia. Using polymerase chain reaction (PCR) detection, mRNA encoding the Y2 receptor has been found in a variety of regions of the GI tract, including the crypt, villus, colonic epithelium, and nonepithelial jejunum.

Y2 is predominantly a presynaptic autoreceptor inhibiting further release of NPY and cotransmitters.

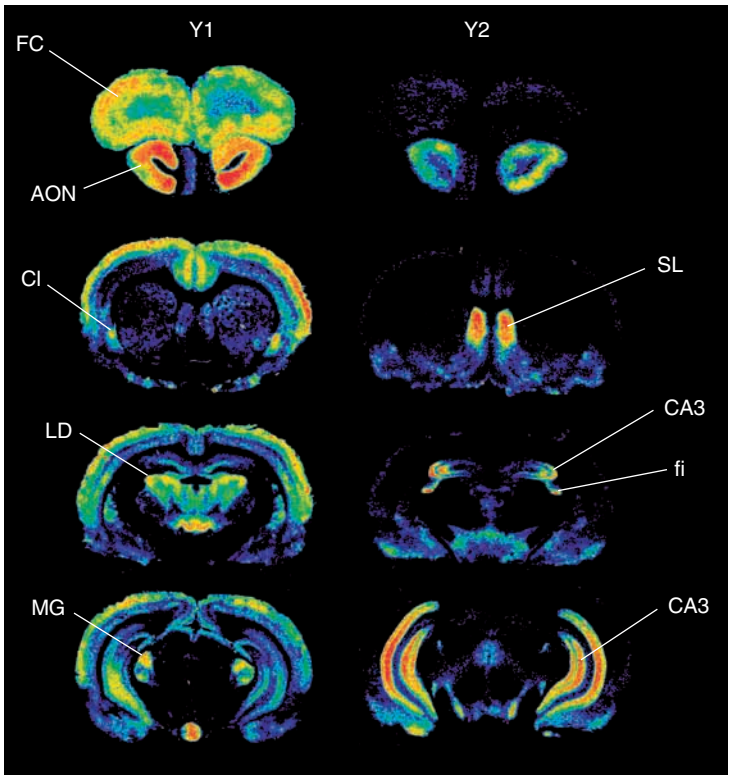

Figure 2 Differential autoradiographic localization of Y1 and Y2 receptors in the rat brain. Rat brain sections were labeled with ^{125}I-Leu^{31}-Pro^{34}-PYY (Y1) and ^{125}I-PYY3–36 (Y2) and pseudocolor representations of the receptor distributions obtained. Abbreviations are frontal cortex (FC), accessory olfactory nucleus (AON), claustrum (Cl), lateral septum (SL), lateral dorsal thalamus (LD), CA3 subregion of the hippocampus (CA3), fimbria (fi), and medial geniculate (MG).

An opposing relationship between Y1 and Y2 is evident for the central effects of NPY on blood pressure, as centrally administered Y2 agonists increase blood pressure while central Y1 stimulation decreases it. Given into the brain, Y2 agonists produce anxiogenic responses, presumably by inhibiting the release of NPY and/or GABA from neurons within the amygdala to limit their postsynaptic actions. In the pig spleen, Y2 stimulation evokes potent vasoconstriction that is inhibited by a Y2-selective antagonist. In addition, Y2 is involved in NPY-induced angiogenesis and NPY-induced effects on circadian rhythms. The phenotype of Y2-deficient mice indicates the potential importance of Y2 in feeding responses and bone formation. Peripherally administered PYY3–36 produces a decrease in food consumption by rodents and humans, presumably by activating Y2 receptors in the hypothalamus.

Y4 Receptor

The human Y4 gene is found on chromosome 10 and encodes a protein of 375 amino acids. Of the Y receptors, Y4 is the least conserved across species and is thought to have evolved rather recently. Considerable differences in peptide pharmacology between the rat and human receptors have been noted. For instance, when position 34 of NPY or PYY is replaced by proline, there is an increase in rat Y4 affinity while the human Y4 affinity is unaffected by this change. In contrast to the Y1 and Y2 receptors, the Y4 receptor exhibits a rather restricted distribution within the brain. While mRNA encoding this receptor exhibits a rather broad distribution, studies using radiolabeled PP indicate a high level expression in the interpeduncular nucleus and locus coeruleus in the nucleus of the solitary tract. A lower level of expression is seen within the hypothalamus and thalamus as well as a number of other brain regions. Interestingly, the binding site labeled by ^{125}I-PP in the interpeduncular nucleus does not pharmacologically match the Y4 receptor, indicating that this may be an additional PP binding site. Y4 mRNA is also abundant in the human colon with somewhat lower levels in the prostate and pancreas. In addition, mRNA encoding Y4 has been found in skeletal muscle, thyroid gland, heart, stomach, small intestine, adrenal medulla, cerebral cortex, and nasal mucosa.

Since PP is likely to be the preferred endogenous ligand for Y4, this receptor may mediate many of the GI effects produced by PP, including rabbit ileum contractions. Intravenously administered PP can also reduce food consumption, though the role of endogenous peptide and the Y4 receptor in appetite and body weight regulation has not been established.

Y5 Receptor

Y5 is encoded by a gene located on chromosome 4 encoding a 446-amino-acid protein. The Y5 receptor is highly conserved (88–90% overall amino acid identity) between orders of mammals. While Y5 mRNA receptor has a broad distribution in the central nervous system, only low levels of Y5 receptor protein are found in the brain using receptor autoradiography and immunohistochemistry. In the periphery, Y5 mRNA is been reported to be present in the intestine, ovary, testes, prostate, spleen, pancreas, kidney, skeletal muscle, liver, placenta, and heart.

The importance of Y5 in NPY-induced feeding has been confirmed by studies involving antisense knockdown, knockout animals, and Y5-selective agonists. Other effects of NPY that are mediated by the Y5 receptor include reproduction through inhibition of LH release and regulation of brain excitability and seizures. Y5 receptor stimulation inhibits neuronal activity in the suprachiasmatic nucleus without generating a phase shift, indicating that Y5 may be indirectly involved in regulation of circadian rhythms.

Regulation of Y Receptors

Internalization and desensitization of NPY family receptors Cell surface expression of Y receptors is regulated primarily by agonist-induced desensitization. This is an important mechanism *in vivo* since NPY knockout mice have a dramatic increase in Y receptor binding when compared to wild-type controls. After agonist stimulation, approximately 20–30% of Y1 expressed in the neuroblastoma cell line SK-N-MC was internalized after stimulation. Y1 internalization follows an endosome pathway followed by rapid recycling to the cell surface. Y4 and Y5 also exhibit rapid internalization after agonist stimulation, while theY2 receptor appears to internalize at a very slow rate. β-Arrestin 2 is an important protein involved in the initial desensitization and internalization of GPCRs. After agonist stimulation, Y1 and Y5 exhibited a rapid recruitment of β-arrestin 2, while Y2 showed a very weak and slow increase in β-arrestin 2 association. Y4 interacts at a rate intermediate to the other receptors.

Dimerization among NPY family receptors It has been recently established that GPCRs form homo- and heterodimers and these complexes have functional implications by altering the signaling, the pharmacology, or the internalization. GPCR dimers can be covalently linked as described for the metabotrophic glutamate receptors or held together by noncovalent interaction like that observed for many of the rhodopsin-like receptors (family A). The Y receptors homodimerize when expressed in HEK293 cells. Y1, Y2, and Y5 homodimers were not affected by agonist treatment, whereas Y4 appears to be assembled as a dimer that dissociates upon agonist stimulation. More recently, Y1 and Y5 were found to heterodimerize when co-expressed. This heterodimer has similar binding properties to what would be expected with a Y1 and Y5 mixture; however, functional coupling and internalization are dramatically altered.

The NPY System and Medication Development

Given the robust effects of the NPY family of peptides on feeding behavior and body weight, there has been considerable interest in the discovery and development of molecules that exploit these effects. In the early 1990s there were a number of efforts by pharmaceutical companies to discover centrally active Y1 and Y5 antagonists. In general, it was difficult to obtain Y1 compounds with suitable properties for clinical development. Adequate oral bioavailability and brain penetration were the consistent challenges observed with a number of chemical series. Several compounds were evaluated preclinically and were able to reduce food consumption and body weight. With the discovery of the Y5 receptor in the mid-1990s, a number of efforts targeted this receptor for drug discovery. While a number of very nice compounds were identified, the highly selective Y5 antagonists produced very modest, if any, changes to food consumption and body weight in preclinical testing and, with one compound, no effect in clinical testing. To date, alternate clinical applications for this class of compounds have not been identified.

In the early 2000s, peripherally injected PYY3–36 was found to decrease food intake through activation of the hypothalamic Y2 receptor. The efficacy was observed in preclinical models as well as several clinical trials. Subsequent efforts have found the potential for improved efficacy using peptides with affinity for both the Y2 and Y4 receptors. Several of these peptides are under clinical development, and an improved understanding of the potential of these peptides to

treat obesity and the metabolic syndrome should be forthcoming.

In addition to body weight disorders, there are substantial opportunities to discover and develop agonists or antagonists for the treatment of a variety of cardiovascular disorders, epilepsy, immune disorders, psychiatric disorders, and substance abuse disorders.

See also: Mammalian Neuropeptide Families; Neuropeptide Receptors – Drug Development; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Food Intake.

Further Reading

Balasubramaniam A (2003) Neuropeptide Y (NPY) family of hormones: progress in the development of receptor selective agonists and antagonists. *Current Pharmaceutical Design* 9: 1165–1175.

Berglund MM, Hipskind PA, and Gehlert DR (2003) Recent developments in our understanding of the physiological role of PP-fold peptide receptor subtypes. *Experimental Biology & Medicine* 228: 217–244.

Colmers WF and El Bahh B (2003) Neuropeptide Y and epilepsy. *Epilepsy Currents* 3: 53–58.

Gehlert DR and Crawley JN (eds.) (2004) Special Issue on Neuropeptide Y. *Neuropeptides* 38: 135–267.

Heilig M and Thorsell A (2002) Brain neuropeptide Y (NPY) in stress and alcohol dependence. *Reviews in Neuroscience* 13: 85–94.

Jacques D, Sader S, Perreault C, and Abdel-Samad D (2006) NPY and NPY receptors: Presence, distribution and roles in the regulation of the endocardial endothelium and cardiac function. *EXS* 95: 77–87.

Kalra SP and Kalra PS (2004) NPY – an endearing journey in search of a neurochemical on/off switch for appetite, sex and reproduction. *Peptides* 25: 465–471.

Magni P (2003) Hormonal control of the neuropeptide system. *Current Protein and Peptide Science* 4: 45–57.

Morton GJ and Schwartz MW (2001) The NPY/AgRP neuron and energy homeostasis. *International Journal of Obesity and Related Metabolic Disorders* 25(Supplement 5): S56–62.

Obuchowicz E, Krysiak R, and Herman ZS (2004) Does neuropeptide Y (NPY) mediate the effects of psychotropic drugs? *Neuroscience and Biobehavioral Reviews* 28: 595–610.

Redrobe JP, Dumont Y, and Quirion R (2002) Neuropeptide Y (NPY) and depression: From animal studies to the human condition. *Life Sciences* 71: 2921–2937.

Ubink R, Calza L, and Hokfelt T (2003) 'Neuro'-peptides in glia: Focus on NPY and galanin. *Trends in Neuroscience* 26: 604–609.

Zukowska Z and Feuerstein G (eds.) (2005) *The NPY Family of Peptides in Immune Disorders, Inflammation, Angiogenesis and Cancer.* Basel: Birkhäuser Verlag.

Neuropeptides and Coexistence

A Merighi, University of Turin, Turin, Italy

© 2009 Elsevier Ltd. All rights reserved.

General Concepts and Definitions

The term 'neuropeptide' was originally coined to indicate small protein molecules that are contained in neurons and that are composed of 3–100 amino-acid residues. These molecules are usually produced as large, inactive precursors, which are enzymatically cleaved to yield the biologically active peptide. Commonly, precursors contain several molecules of the same neuropeptide and/or more or less structurally related compounds.

Neuropeptides exert several different biological effects, such the regulation of gene transcription, local blood flow, synaptogenesis, and glial cell architecture. Notably, most neuropeptides also influence membrane excitability, and, although this should perhaps not be thought of as their main biological action, this has made them very attractive to neurobiologists. When thinking of neuropeptides as being involved in cell-to-cell communication, the first consideration to be made is that they are about 50 times larger than low-molecular-weight 'classical' neurotransmitters, such as acetylcholine, biogenic amines, and amino acids. As a consequence, neuropeptides possess several more recognition sites for receptors than do smaller neurotransmitters, and thus a higher receptor-binding affinity (about 1000 times greater, with values in nanomoles/liter vs. micromoles/liter) and selectivity. For these reasons neuropeptides can elicit their biological effects even when released at lower quantities. Although large neuropeptide molecules can diffuse and bind slowly to receptors, their half-lives in the brain extracellular space are remarkably long: for example, the half-lives of oxytocin and vasopressin have been estimated to be in the order of about 20 min.

Neuropeptide receptors are usually seven-transmembrane-region G-protein-coupled receptors (GPCRs), which are internalized upon activation as a desensitization mechanism. For example, the neurokinin-1 receptor (the preferred substance P receptor) is internalized 5 min after agonist activation, and is then recycled and restored to the cell membrane within 30 min.

Cellular and Subcellular Sites of Neuropeptide Storage

Localization within Large Granular Vesicles

Identification of the subcellular site of storage of neuropeptides remains crucial to the understanding of their mechanism of action as synaptic (or nonsynaptic) transmitters and of their interaction with other molecules involved in chemical neurotransmission. Neuropeptides are commonly found to be present together with low-molecular-weight classical neurotransmitters in individual nerve cells. However, initial observations indicated that the two classes of transmitter molecules were stored in different subcellular compartments: in large granular vesicles (LGVs) for the neuropeptides, and in small clear synaptic vesicles (SSVs) for the classical transmitters (Figure 1).

In the past, the debate regarding such a compartmentalization was left open, since it could not be excluded that, at least in some cases, LGVs also contained some of the more conventional transmitters, although peptides never seemed to be found in small clear vesicles. Nowadays, the concept that LGVs are the sole site of neuropeptide storage is widely established. The histological demonstration of differential subcellular sites of storage for neuropeptides and low-molecular-weight neurotransmitters is consistent with the possibility that they are selectively released upon specific stimuli. On the other hand, coexisting neuropeptides, even when synthesized from different mRNAs, are usually stored together in LGVs. This has a series of functional implications, since co-stored neuropeptides are not differentially released, and a modulation of biological effects is most readily accomplished by regulating their relative proportions of synthesis.

Synthesis, Storage, and Targeting to Processes

As occurs for all peptide molecules, neuropeptides are synthesized in precursor form within the rough endoplasmic reticulum (RER) and subsequently move to the Golgi apparatus, where they are packaged into LGVs. The primary sensory neurons are a unique class of pseudounipolar neurons; they are grouped outside the central nervous system (CNS) in the sensory ganglia, associated with certain cranial nerves and in dorsal root ganglia (DRGs), and display an extremely abundant and variegated neuropeptide content. Immediately after emerging from the cell body the single process of these neurons divides into a central branch and a peripheral branch that function, respectively, as an axon and a dendrite. Pioneering work carried out with the use of multiple immunogold labeling methods about 20 years ago led to the localization of different tachykinins and calcitonin gene-related peptide (CGRP) to individual LGVs in neuronal cell bodies and central peripheral and branches of DRG neurons. In these cells, virtually

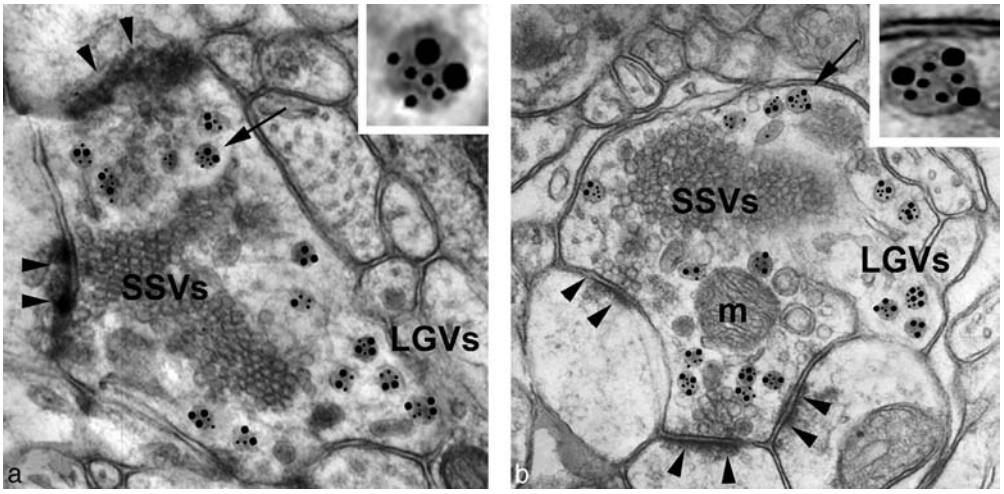

Figure 1 Subcellular localization of neuropeptides and presence within large granular vesicles at central synapses. Calcitonin gene-related factor/substance P (a) and brain-derived neurotrophic peptide/substance P (b) dual immunogold labeling (10 and 20 nm gold particles) in central terminals of a dorsal root ganglion neuron within the substantia gelatinosa of spinal cord. A multisynaptic complex (glomerulus) contains small synaptic vesicles (SSVs) and large granular vesicles (LGVs) with a dense core. Note in both images that SSVs are consistently unlabeled, whereas the majority of LGVs are double labeled. Note also that LGVs are not in proximity of synaptic specializations (arrowheads), unlike SSVs (m, mitochondrion). The arrows indicate the LGV that is shown at higher magnification in the corresponding insets at top right of each panel.

all LGVs are multiple/dual-labeled in both central and peripheral branches, whereas double (multiple)-labeled LGVs in the cell body relatively rare. These observations show that when multiple peptides are produced by one neuron, they are not selectively packaged into different LGV subpopulations, but rather a cocktail of neuropeptides is consistently found within individual LGVs. Moreover, given that LGVs (and therefore the neuropeptides packed therein) are detected at both central and peripheral branches of DRG neurons, it seems reasonable to hold that no selective transport to functionally different neuronal processes (the axons and the dendrites) and/or to different branches of the same process occurs. In neuronal cell bodies, LGVs containing just one component of the cocktail can be regarded as immature vesicles that will probably incorporate the other peptide(s) before being transported to terminals. Studies on nonneuronal cell regulated secretion suggest that peptides are packed into immature LGVs budding from the *trans*-Golgi network, and ultimately transform into mature LGVs. Fusion between immature secretory granules has been demonstrated, and it is thought that after granule–granule fusion, soluble contents condense, and excess membrane are removed by vesicle budding.

Very recent work has shown that an analogous pattern of co-storage is also apparent in central neurons, and that co-stored neuropeptides are quantitatively detected in LGVs with remarkably constant ratios. The ratio might be regulated at the level of neuropeptide synthesis, before packaging in the *trans*-Golgi network (Figure 2).

Coexistence

Coexistence, the concurrent presence of two or more transmitters in a single neuron, is now regarded as a common feature of central and peripheral neurons. Neuropeptides coexist with other neuropeptides, low-molecular-weight fast-acting neurotransmitters, the gaseous transmitter NO, and certain neurotrophins. When multiple neuropeptides are present in neurons, coexistence equals co-localization, since, as discussed earlier, coexisting peptides are co-stored in LGVs.

As a rule, neurons produce a combination of one (or more) low-molecular-weight transmitter(s) and one (or more) high-molecular-weight neuropeptide(s). One remarkable exception seems to be represented by the oxytocin/vasopressin magnocellular neurons that contain a complex cocktail of peptides, but apparently no low-molecular-weight transmitters. When a neuropeptide coexists with a classical transmitter, the latter is generally believed to be the principal messenger, whereas the neuropeptide appears to modulate neuronal response by acting on pre- and/or postsynaptic GPCRs (see the section titled, 'Functional implications of neuropeptide coexistence').

Neuropeptide Co-Localization

Several areas of the CNS are particularly enriched with neuropeptides. In most circumstances, it is common that individual neurons in these areas produce and store more than a single peptide. Relevant examples are given in the following sections.

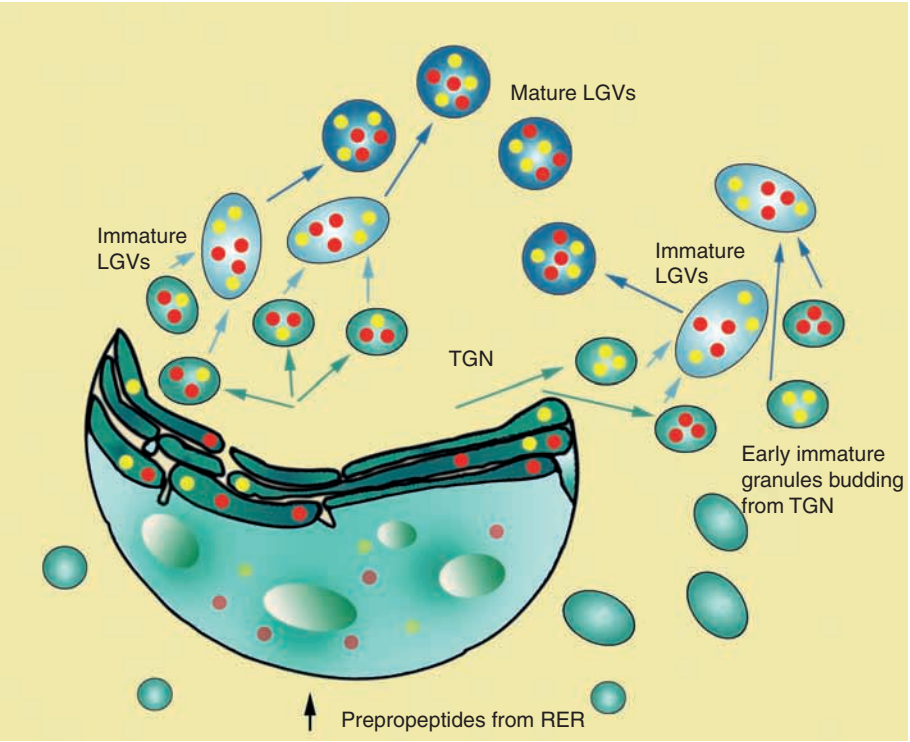

Figure 2 Packaging of neuropeptides in large granular vesicles (LGVs). For simplicity, only two peptides (indicated by the red and yellow circles) are represented, although there is direct evidence that more neuropeptides can be stored in individual LGVs. Prepropeptides reach the Golgi apparatus from the rough endoplasmic reticulum (RER). They are packaged in granules budding from the *trans*-Golgi network (TGN); the granules subsequently fuse together to give rise to immature LGVs. Immature LGVs are larger and display a more irregular shape compared to their mature counterparts. As maturation of LGVs proceeds, there is a condensation of the granin content, so that the vesicles assume the typical size, shape, and electron density of mature LGVs. When a neuron produces more than a single neuropeptide, it is possible that immature granules budding from the TGN (left) already contain a mixture of cargo molecules. Alternatively, each peptide is singularly packed into individual immature granules (right), which then fuse together to give rise to the mature LGVs. In both cases, the latter contain the cocktail of neuropeptides that are eventually transported to terminals.

Spinal cord The co-localization of tachykinins and CGRP is one of the most widely studied example of peptide co-localization (Figure 1(a)). The cell bodies of the neurons containing these peptides are located in the peripheral nervous system in the DRGs and trigeminal ganglia, and in several autonomic ganglia and intramural plexuses. Galanin is co-localized mainly with enkephalin, but sometimes also with neuropeptide Y (NPY) in some local neurons of the substantia gelatinosa (lamina II) of the dorsal horn. In the same location, the co-localization of tachykinins and enkephalin is also reported. Galanin and substance P are co-localized in neurons of the deeper layers of the dorsal horn. Co-localization of these neuropeptides and neurotensin with galanin has been demonstrated also at the ultrastructural level.

Hypothalamus The hypothalamic hormones oxytocin and vasopressin were the first neuropeptides discovered, and an extensive literature exists on the localization, biological actions, and plasticity of these peptides. The oxytocin- and vasopressin-containing neurons are located in both large (magnocellular) and small (parvocellular) hypothalamic neurons. In the supraoptic nucleus the oxytocin/vasopressin system undergoes a remarkable plasticity under normal and experimental conditions. Although oxytocin and vasopressin neurons usually constitute two separate populations, there are some plasticity inducing conditions, such as parturition and lactation, in which co-localization of the two peptides has been hypothesized on the basis of *in situ* hybridization and single-cell reverse transcription–polymerase chain reaction studies.

Considerable evidence has accumulated that approximately 50% of the corticotropin-releasing hormone (CRH) neurons in the paraventricular nucleus contain vasopressin or its precursors. Moreover, about 30% of the CRH parvocellular neurons in the median eminence contain vasopressin. These neurons are also highly plastic, and the relative proportion of released vasopressin and CRH changes under different physiological and experimental conditions.

Magnocellular oxytocin neurons contain several opioid peptides of the enkephalin family, and magnocellular vasopressin neurons contain dynorphin. In addition, some oxytocin neurons contain either cholecystokinin (CCK) or CRH. Co-localization of CCK and CRH has been demonstrated in periportal nerve terminals of the rat median eminence. It has also been shown that galanin is co-produced in a subpopulation of luteinizing hormone-releasing hormone (LHRH) neurons. The incidence of LGVs immunolabeled for both neuropeptides is high in the female and low in the male rat. Ovariectomy results in a dramatic decline in the number of LHRH/galanin-co-expressing LGVs, which is reversed by the administration of estradiol.

Retina A large number of neuropeptides occur in amacrine cells, and to a lesser extent, in the ganglion cells. These include avian pancreatic polypeptide (APP), enkephalin, neurotensin, NPY, pituitary adenylyl cyclase-activating peptide (PACAP), pancreatic polypeptide (PP), somatostatin, and substance P. Many of these neuropeptides are co-localized within individual neurons.

Coexistence of Neuropeptides and Fast-Acting Neurotransmitters

Acetylcholine Neuropeptides coexist with acetylcholine in numerous areas of the brain. For example, galanin is found to coexist with acetylcholine in the basal forebrain and hippocampus, where the coexistence of somatostatin and acetylcholine has also been described. Coexistence of acetylcholine and CGRP has been reported in neurons of the vestibular nuclei.

Amino acids The coexistence of peptides and aspartate, glutamate, γ-aminobutyric acid (GABA), or glycine is a widespread finding in the CNS. Aspartate coexists with substance P in the spinal cord, and with NPY or galanin in the arcuate nucleus. The coexistence of glutamate and peptides is by far more common. Coexistence of enkephalin and glutamate has been shown to occur in the locus coeruleus. Glutamate and PACAP are localized within the same neurons in the retinohypothalamic tract. Glutamate and substance P coexist in the spinal cord. GABA has been found to be present with several different neuropeptides in various areas of the brain: with enkephalin in the cochlea, striatum, and spinal cord; with somatostatin in the visual and entorhinal cortex, suprachiasmatic nucleus, and basolateral amygdala; with cholecystokinin-8 (CCK-8) and/or vasoactive intestinal peptide (VIP) in the hippocampus dentate gyrus, basolateral amygdala, and suprachiasmatic nucleus; with CGRP in the cerebellum; with NPY in the arcuate nucleus, basolateral amygdala, and dorsal horn; with substance P in the hypothalamus and retinal ganglion cells; with galanin in the dorsal horn; with somatostatin in the entorhinal cortex; with CRH in the hippocampus; and with vasopressin in the suprachiasmatic nucleus.

Biogenic amines Several different neuropeptides have been found to coexist with biogenic amines such as serotonin (5-hydroxytryptamine; 5-HT), and with noradrenaline. The most widely reported combinations are the coexistence of substance P and 5-HT in caudal medullary neurons projecting to the spinal ventral horn, and in dorsal raphe nuclei neurons, and the coexistence of enkephalin or galanin and 5-HT in the dorsal raphe nuclei.

Other combinations have been found with noradrenaline. These include substance P and noradrenaline in the medulla oblongata, NPY and noradrenaline in the hypothalamus, and galanin and noradrenaline in the cerebral cortex and hippocampus. Multiple combinations are possible, and there is strong evidence that more than one neuropeptide may be contained within biogenic amine-synthesizing neurons. Also, these neurons can produce other classical transmitters, such as transmitter amino acids.

Nitric oxide Certain peptides have been found to be present in nitric oxide-producing neurons (detected after NADPH diaphorase histochemistry or nitric oxide synthase immunocytochemistry). Examples include somatostatin in the hippocampus, vasopressin in the hypothalamus, and galanin in the dorsal raphe nucleus.

Coexistence of Neuropeptides and Brain-Derived Neurotrophic Factor

A direct demonstration for the selective storage of brain-derived neurotrophic factor (BDNF) in LGVs has recently been provided by ultrastructural immunocytochemistry (Figure 1(b)), in terminals within the central nucleus of amygdala and spinal cord dorsal horn. These terminals originate from central and peripheral neurons that are capable of synthesizing and anterogradely transporting the neurotrophin. Individual LGVs have been found to contain a cocktail of BDNF, CGRP, and substance P. Such a pattern of cellular storage is compatible with a transmitter function of BDNF in these neuronal populations.

Neurotransmitter Function of Neuropeptides

Despite the enormous amount of literature on neuropeptide distribution in mammals, much of our

knowledge about neuropeptide function comes from studies in invertebrates. Indeed, it is far easier to demonstrate the presence of multiple messengers in neurons than it is to establish their physiological role, or even to show that they have any kind of biological activity at all. Moreover, although an organic framework to describe the function(s) of individual neuropeptides at synapses and/or nonsynaptic sites is now available, relatively little is known about the functional interactions and the control of release of co-stored neuropeptides at central synapses. The existence of synapses can be unequivocally demonstrated only by electron microscopy. Ironically, however, ultrastructural studies have also shown the existence of nonsynaptic transmission. The identification of gases as interneuronal signals or the modulation of neuronal function by lipophilic substances has left no doubt regarding the existence of nonsynaptic information transfer in the CNS. Nonetheless, it is more difficult to accept that vesicle-stored transmitters can be released following quantal mechanisms and operate at distant nonsynaptic sites. This notion applies not only to neuropeptides, but also to certain low-molecular-weight classical neurotransmitters.

Neuropeptides as Slow-Acting Neurotransmitters

Neurotransmitters are consistently packaged in two different morphological types of vesicles at synapses. The most abundant SSVs accumulate fast-acting, low-molecular-weight neurotransmitters. The less frequently observed LGVs store the neuropeptides (Figure 1(a)). The distribution of SSVs and LGVs in typical axodendritic central synapses follows a rather stereotyped pattern. SSVs occupy a variable, but usually large, area of the axon terminal. Some are docked at the presynaptic grid, being ready to be released at the active synaptic zone. LGVs, on the other hand, are localized away from the presynaptic membrane, singularly or in clusters. In addition, electron microscopy has shown that release of LGVs can occur at plasma membranes even in the absence of synaptic specializations. The major functional implication of such an arrangement is that fast- and slow-acting transmitters may be selectively released upon activation of different cellular pathways. Early evidence was obtained showing that the release of coexisting peptides and classical neurotransmitters could be differential and dependent on the frequency and pattern of firing. In general terms, neuropeptide release is triggered by a small increase in the intracellular Ca^{2+} concentration, whereas release of transmitter amino acids from SSVs requires a rise of intracellular Ca^{2+} concentration in the proximity of the Ca^{2+} channels at synapses. Therefore, in terminals with both types of vesicles, a focal increase in Ca^{2+} at the synaptic membrane leads to a preferential discharge from SSVs, whereas a more general elevation of Ca^{2+} inside the terminal favors the release of LGVs. The remarkably short delay in response to fast-acting transmitters (about 1–3 ms) is a close reflection of this mode of discharge from SSVs. On the other hand, the spatial independence from Ca^{2+} channels clustered at synapses explains why neuropeptide release can occur independently from synapses. Among the consequences of the existence of selective mechanisms of release for coexisting peptides and classical transmitters is the possibility that long-lasting intracellular Ca^{2+} elevation may cause the release of neuropeptides to outlast the duration of electrical activity, thus uncoupling release from spiking.

Mode(s) of Release

Even upon prolonged stimulation, not all vesicles at synapses unload their transmitter content. A variable fraction of SSVs and LGVs is readily releasable, but the remaining ones, forming the reserve pool, need further steps to become competent. Two different mechanisms of transmitter emptying have been shown to occur in SSVs and LGVs. These include the slower classical exocytosis, with complete fusion of the vesicle to the plasma membrane, or a faster mechanism whereby vesicles come in close proximity to the membrane and, with the formation of a transient pore, release part of their transmitter content by a process described as 'kiss and run.' The transient pore mechanism allows the quick, simultaneous passage of amine transmitters (and perhaps other small molecules, which may be present in LGVs together with neuropeptides) from LGVs into the extracellular fluid. On the other hand, neuropeptides contained within LGVs remain trapped inside the retrievable vesicle, as a consequence of their higher molecular weight. Therefore, release of most neuropeptides from the LGVs is unlikely to occur through kiss and run, for several reasons. These include the larger size of the peptides relative to the transient pore, and the slow emptying of peptide content from LGVs upon exocytosis. In support, simultaneous capacitance measurements and confocal imaging have shown that peptide release by kiss and run is negligible. On the other hand, complete vesicle fusion is usually required, through a mechanism involving a priming step followed by retrieval of the vesicle as a coated vesicle. A divergence in this respect between peptide-containing LGVs and amine-containing LGVs may merely be additional

to the several differences between these two classes of neurotransmitters. Indeed, neuropeptides do not have a known re-uptake mechanism, as opposed to biogenic amines, so that there is no way locally to refill the peptide-containing LGVs after emptying. Moreover, neuropeptides are synthesized at the rough endoplasmic reticulum in the neuronal perikarya, but not in axon terminals, whereas amines can also be synthesized inside LGVs.

When a neuron produces more than a single neuropeptide (as appears to be the rule), several theoretical possibilities arise regarding their release. It is possible that all co-stored neuropeptides can be released all at once, at all processes. Alternatively, or in addition, individual neuropeptides can be theoretically liberated singularly or in different combinations at different processes. In *Aplysia*, the various neuropeptides derived from a common prohormone seem to be targeted to different neuronal processes. The situation is far less clear in mammals, considering that most studies have been carried out on isolated neurons. Nonetheless, all co-stored neuropeptides can be reasonably assumed to be released together upon exocytosis. For example, the co-release of CGRP and substance P (and other tachykinins) has been demonstrated, and this occurs at both central and peripheral endings of the DRG neurons. Moreover, if the co-release of co-stored neuropeptides is indeed the rule, then it should occur from any neuronal process containing LGVs, although this latter issue also needs further clarification. The major functional implication is that co-released peptides probably act together in determining the response of target cells. A differential release of co-stored peptides (if indeed this occurs *in vivo*) would more likely rely on mechanisms different from those that apply to co-stored biogenic amines and/or coexisting low-molecular-weight neurotransmitters. From this perspective, the relative rate of peptide dissolution from the LGV core might be of primary relevance, since this appears to be the critical determinant of the speed of peptide secretion *in vitro*.

Functional Implications of Neuropeptide Coexistence

For an evolutionary perspective, coexistence of multiple neurotransmitters in neurons has led to substantial advantages. In particular, neuropeptides have a wide diversity of direct or modulatory effects on the electrical responses of target cells, in addition to trophic effects. When they are co-released with other neurotransmitters, the wealth of responses of target neurons increases dramatically. Neurotransmitters produced and released by a single neuron are often defined as cotransmitters. However, it is probably unsafe to consider that cotransmitters must display some kind of interaction, simply because they are co-released, even though such a co-release occurs under physiological conditions. The common existence of a combination of neuropeptides and classical neurotransmitters in neurons enables fast (2–5 ms) and slow (100–500 ms) synaptic communication to take place. Fast- and slow-acting cotransmitters can act on completely independent targets and, therefore, do not interact at all. However, there is a general consensus that when multiple neurotransmitters are released within the extracellular space, they usually display at least some type of interactive actions, irrespective of the finding that such a release occurs from the same neuron – that is, they are true cotransmitters, or from separate sources.

The simplest mode of interaction of two (or more) neurotransmitters occurs when two (or more) distinct receptor complexes are present in the (postsynaptic) membrane of target cells, and a receptor–receptor interaction occurs (**Figure 3**). When neuropeptides coexist with low-molecular-weight neurotransmitters, the neuropeptide usually acts on GPCRs, whereas the low-molecular-weight transmitter generally opens a ligand-gated ion channel. The low-molecular-weight transmitter is generally the principal messenger, and the neuropeptide interacts with it by altering the ion channel gating properties or its response to further signals. This occurs by direct operation on the receptor complex or by the activation of second-messenger systems that, in turn, act on the receptor complex. Hence, one neurotransmitter may alter the number of receptors or the affinity of the receptor to the other(s) simultaneously released. Interestingly, receptor recruitment from the interior of the cell to the plasma membrane may be an additional and ubiquitous mechanism of modulation of signal transduction, leading to receptor sensitization. The interaction of cotransmitters also occurs through presynaptic regulation. This implies the existence of presynaptic receptors for one or more messengers. In this case, one of the neurotransmitters feeds back on presynaptic receptors and thus affects its own release or the release of the cotransmitter(s). This latter possibility has been demonstrated, for example, in the striatum in which tachykinins presynaptically stimulate the release of dopamine, and in the locus coeruleus in which noradrenergic neurons can be activated by the stimulation of neurokinin-3 receptors.

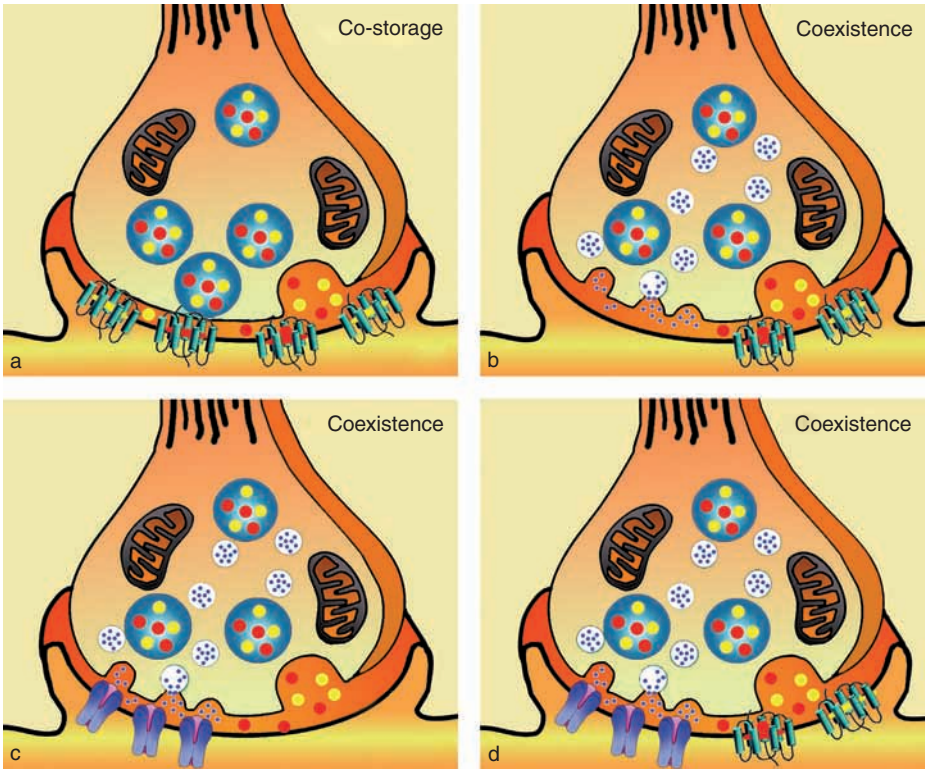

Figure 3 Functional implications of neuropeptide coexistence. For simplicity, only two peptides (indicated by the red and yellow circles) are represented, and only postsynaptic interactions are considered. In purely peptidergic neurons (a), neuropeptides co-stored within large granular vesicles are released together and affect the activity of their target neurons upon binding to their cognate G-protein-coupled receptors. In neurons producing and releasing a combination of neuropeptides and a fast-acting neurotransmitter (b–d), the coexistence of these neuroactive molecules does not necessarily imply that they are acting as cotransmitters. As depicted, in panel (b) only the neuropeptides affect the target neuron that expresses specific G-protein-coupled receptors, whereas in panel (c) the target neuron binds receptors (ion channels) only for the low-molecular-weight transmitter. Only in panel (d) are the neuropeptides and the low-molecular-weight transmitter true cotransmitters. In this case the general rule is that the low-molecular-weight transmitter acts as the principal messenger, whereas the neuropeptides subserve a modulatory role.

See also: Calcitonin Gene-Related Peptide (CGRP) and Receptors; Cotransmission; Mammalian Neuropeptide Families; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Discovery; Neuropeptides: Electrophysiology; Substance P/Tachykinins and its/their Receptors.

Further Reading

Hökfelt T, Johansson O, Ljungdahl Å, et al. (1980) Peptidergic neurons. *Nature* 284: 515–521.

Hökfelt T, Broberger C, Xu Z-QD, et al. (2000) Neuropeptides – an overview. *Neuropharmacology* 39: 1337–1356.

Hökfelt T, Pernow B, and Wahren J (2001) Substance P: A pioneer amongst neuropeptides. *Journal of Internal Medicine* 249: 27–40.

Meldolesi J, Chieragatti E, and Malosio ML (2004) Requirements for the identification of dense-core granules. *Trends in Cell Biology* 14: 13–19.

Merighi A (2002) Costorage and coexistence of neuropeptides in the mammalian CNS. *Progress in Neurobiology* 66: 161–190.

Nusbaum MP, Blitz DM, Swensen AM, et al. (2001) The roles of co-transmission in neural network modulation. *Trends in Neuroscience* 24: 146–154.

Salio C, Lossi L, Ferrini F, et al. (2006) Neuropeptides as synaptic transmitters. *Cell and Tissue Research* 326(2): 583–598.

Salio C, Averill S, Priestley JV, et al. (2007) Costorage of BDNF and neuropeptides within individual dense-core vesicles in central and peripheral neurons. *Developmental Neurobiology* 67: 326–338.

Neuropeptides and Receptors in Glia

L Calzà, University of Bologna, Bologna, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuropeptides were originally defined as chemically distinct small proteins (3–100-amino-acid residues) widely distributed in the central and peripheral nervous system and acting as neurotransmitters, cotransmitters, and/or neuromodulators. Initially isolated from different tissues and organs, thanks to their biological properties or by searching for C-terminally amidated fragments, their already large number has further increased over the past 15 years with the advent of molecular cloning and related techniques. New neuropeptides have been found on the basis of messenger RNA (mRNA) sequence, genomic data searching for structural features expected to be intrinsic to neuropeptides, and homology screen approaches applied to G-protein-coupled receptors (GPCRs) (because all the receptors activated by neuropeptides are GPCRs). Today, several tens of small peptide molecules have been identified. They are involved in synaptic transmission, regulation of neurotransmitter recycling and neurotransmitter receptor synthesis (vasoactive intestinal peptide/pituitary adenylate cyclase-activating polypeptide (VIP/PACAP)); ion spatial segregation and, thus, neuronal excitability (NPY, neurotensin); and the complex regulation of energy metabolism coupled to neuronal activation and local vasodilation (VIP/PACAP).

At the same time, the concept of neuropeptides such as neurotransmitters being uniquely synthesized in nerve cells and involved only in synaptic transmission has been severely challenged since the discovery that many neuropeptides can be detected in all main types of glial cells, including astrocyte, oligodendrocyte, NG2-positive oligodendrocyte precursors, microglia, and Schwann cells. Moreover, glial cells are no longer merely likened to the connective cell type of nervous tissue acting as supporting cells: Glial cells play active roles in the development, maintenance, function, and pathology of the nervous system.

It is now clear that glial cells sense and respond to neuronal activity, as indicated by changes in intracellular Ca^{2+} levels that can trigger 'gliotransmitter' release. They possess the molecular machinery necessary for regulated exocytosis, including fusion protein synaptosome-associated protein (SNAP) receptor (SNARE). Astrocytes express synaptobrevin 2, cellubrevin, synthaxin 1, and SNAP 23, associated with electron-lucent (clear), possibly containing classical neurotransmitters, and some dense core vesicular structures, possibly storing peptides. In fact, as soon as astrocytes' ability to release 'neuroligands' was proved, it became clear that not only classical transmitters (i.e., glutamate) but also neuropeptides, such as atrial natriuretic peptides, and aminoacids, such as D-serine, could be released using an exocytotic mechanism from astrocytes, although the lack of evidence in support of an 'active zone' on the plasma membrane makes the site of release uncertain.

Glial cells not only synthesize neuropeptides but are also regulated by them. The early evidence of classic neurotransmitter receptors on glial cells was in fact followed by the demonstration of binding sites and receptor mRNA for many neuropeptides in purified cultures of all major classes of glial cells, and *in situ* hybridization studies on tissue sections have confirmed the presence of neuropeptide receptor mRNA in glial cells in different brain areas. In terms of membrane effects, both hyperpolarization (cholecystokinin (CCK), galanin, neurotensin, VIP/PACAP) and depolarization (neuropeptide Y (NPY)) have been observed. The presence of second messenger systems has also been documented, including generation of cyclic adenosine monophosphate (cAMP), cyclic guanosine monophosphate (cGMP), inositol triphosphate, and intracellular Ca^{2+} transients, and activation of mitogen-activated protein kinase (MAPK), phospholipase A2 and D, and specific G-proteins, as well as transcription factor regulation (induction of immediate early genes and nuclear factor-kappa B (NF-κB)).

Peptides have been shown to participate in the regulation of major processes in glial cells, such as proliferation (e.g., angiotensin, calcitonin gene-related peptide (CGRP), opioids, VIP/PACAP), development and maturation (opioids), energy metabolism (angiotensin), survival (opioids), and in the regulation of the proper function of each glial cell class (e.g., VIP and opioids regulate cytokine production in microglial cells; opioids regulate myelin production in oligodendrocytes; neurotensin and NPY modulate astrocyte reaction during inflammation and after lesions), as is also indicated by regulation of receptor expression, binding properties, and intracellular effects after different stimuli. A completely new chapter has been recently opened in light of the possible role of neuropeptides in gliogenesis and *in vitro* assays based on neural precursor cell technology (*in vitro* neurosphere assay).

Glia expression of peptide and peptide receptors, together with *in situ* detection of molecular markers

Table 1 Peptide expression by different glial cells

Neuropeptide	*Detection level*	*Cell*	*System*
Angiotensin II/renin–angiotensin system (RAS)	mRNA and peptide	Astrocytes	*In vivo*/cell culture
Atrial natriuretic peptide	Peptide	Astrocytes	*In vivo*
	mRNA	Microglia	Cell culture
Enkephalin	Peptide	Astrocytes	Slice culture
	mRNA	Astrocytes	Cell culture
	mRNA and peptide	Astrocytes	Cell culture
	Peptide	Oligodendrocyte precursor cells	Cell culture
Dynorphin	Peptide	Oligodendrocyte precursor cells	Cell culture
Galanin	peptide mRNA and peptide	Microglia oligodendrocyte precursor cells	*In vivo* after colchicine; *in vivo* after colchicine and CSD
Neuropeptide Y	mRNA and peptide	Astrocytes	Cell culture
	mRNA and peptide	Olfactory ensheathing cells	*In vivo*
	mRNA and peptide	Schwann cell precursors	*In vivo*
Nociceptin/orphanin	mRNA and peptide	Astrocytes	Cell culture
Somatostatin	mRNA and peptide	Astrocytes	Cell culture
Substance-P	Peptide	Astrocytes	*In vivo*
	mRNA	Astrocytes	Cell culture
	mRNA	Microglia	Cell culture, human
Vasoactive intestinal peptide	Peptide	Astrocyte	*In vivo*

CSD, cortical spreading depression; mRNA, messenger RNA.

for cell activation and the well-documented role of glial cells in the production and action of growth factors, has defined a new scenario in which neuron-derived neuropeptide regulates glial function and glial-derived neuropeptides regulate neural function and also act in an autocrine-paracrine fashion, thus opening up completely new perspectives for reciprocal neuron–glia interactions in information processing in the nervous system, in brain metabolism, and in the biology of neurodegenerative diseases, migraine, and chronic pain.

This article describes peptide and peptide receptor expression by glial cells, as demonstrated in tissue section, primary cultures, and cell lines using morphological, biochemical, functional, and pharmacological approaches. Available data are listed in Table 1 (neuropeptides) and Table 2 (neuropeptide receptors), and some illustrative peptides are described in the following sections. Just as neuropeptides are heterogeneously distributed among neuronal populations and brain regions, their expression in glia is also highly regional. Several of these neuropeptides are under developmental regulation, and after the postnatal downregulation of glial peptide transcripts, injury in adulthood can cause upregulation again. Thus, data reported in the tables are based on resting conditions of well-characterized *in vivo* (mouse, rat, human) and *in vitro* systems (primary culture and cell lines). Retina is not considered in this article. Different expression profiles, as found studying tissue sections derived from experimental and spontaneous pathologies, mixed cultures, and tumor cell lines, are discussed.

Illustrative Neuropeptides

Angiotensin II

Angiotensin II consists of eight amino acid residues and is the main active substance of the rennin–angiotensin system. It derives from the precursor decapeptide molecule angiotensinogen, a glycoprotein of molecular weight 55 000–58 000, that is converted by renin into angiotensin I. Angiotensin I is converted by angiotensin converting enzyme to angiotensin II. The brain contains an autonomous renin–angiotensin system with all the enzymes and peptides necessary for the biosynthesis of these angiotensins, and astrocytes appear to be the chief source of angiotensinogen in the brain. Large amounts of angiotensinogen mRNA are found in the forebrain medulla; medial thalamus; and hypothalamus, including anterior hypothalamus and the paraventricular, supraoptic, and suprachiasmatic nuclei, as also confirmed by transgenic technology. The presence of immunoreactivity to angiotensin I and II around most brain vessels, presumably in perivascular astrocytes at the surface of astroglial boutons, has also been described in mouse brain tissue.

Angiotensin receptors have been classified into two subtypes, AT1 and AT2. AT1 receptors are expressed on perivascular astrocytes, suggesting that angiotensin

Table 2 Peptide receptor expression by different glial cells

Peptide	*Cell*	*Receptor subtype*[a]	*Effects and effectors*
Angiotensin	Astrocyte	AT1, AT2	Reduction of membrane conductance activation of Na-K-ATPase via PKC delta, activation of PLC e PG release stimulation of PI turnover and reduction of cGMP
			Early genes, glucose uptake and transporter transcription stimulation
			mitogenic
	Microglia	UD	Non-AT1/AT2 type?
	Oligodendrocyte	UD	
	Schwann cell	UD	Regulation of protease nexin-1 expression
Atrial natriuretic peptide	Astrocyte	NPR-A, B, C	Decrease of proliferation suppressing DNA synthesis via MAPK inhibition
	Microglia	NPR-A, B, C	Inhibition of LPS-induces activation through the inhibition of transcription factors NF-κB and AP-1
Calcitonin gene-related peptide	Astrocyte	CALCRL/RAMP1, 2	UD
	Microglia	UD	cAMP accumulation
			Early genes transcription stimulation
	Oligodendrocyte	CALCRL/RAMP1, 2	cAMP accumulation
	Schwann cell	UD	Increase proliferation (synergic with PDGF)
Cholecystokinin	Astrocyte	CCK2	Hyperpolarization
			Increase of intracellular free Ca^{2+} concentration
Galanin[b]	Astrocyte	UD[b]	Hyperpolarization
			Stimulation of early genes transcription
	Microglia	GAL2	Inhibition of LPS-stimulated TNF-α release
Neuropeptide Y	Astrocyte	Y1 (Y2,Y4[c])	Depolarization
			Increase of intracellular free Ca^{2+} concentration
Neurotensin[b]	Astrocyte	NTS2 (NTS1[d])	Hyperpolarization
			Increase of intracellular free Ca^{2+} concentration Up regulated by lesion stimuli
	Microglia	NTS3 (sortilin)	Phosphorylation of ERK1/2 and Akt Elicits migration of the microglial cell by a modification of the F-actin cytoskeleton; induces the gene expression of cytokines and chemokines
	Oligodendrocyte	NTS3 (sortilin)	UD
Opioids	Astrocyte	MOP	Inhibition of Ca^{2+} release
			Decrease of proliferation suppressing DNA synthesis
			Protection from NO-donor-mediated apoptosis
		DOP	Upregulated during mitosis
		KOP	Intracellular Ca^{2+} mobilization
			Cell proliferation inhibition
	Microglia	MOP	apoptosis induction
		KOP	Inhibits cytokines' effects on superoxide generation
	Oligodendrocyte	MOP[e]	Developmentally regulated synthesis
			DNA synthesis enhancement
		DOP	UD
		KOP	Developmentally regulated synthesis increases myelin-like membranes
Somatostatin	Astrocyte	sst1, 2, 4, (3)	Inhibition of cAMP formation
			Inhibition of endozepin release
	Microglia	UD	UD
	Oligodendrocyte	UD	UD
Tachykinins[b]	Astrocyte	NK1	Depolarization
			PI turnover stimulation
			IP formation potentiate IL-6 and PGE2-mediated effects
	Microglia	NK1	NF-κB activation
	Oligodendrocyte	NK2	IP formation
	Schwann cell	UD	UD

Continued

Table 2 Continued

Peptide	*Cell*	*Receptor subtype*[a]	*Effects and effectors*
VIP/PACAP	Astrocyte	PAC1, VPAC1, VPAC2	cAMP formation Increase of intracellular free Ca^{2+} concentration MAPK, cAMP stimulation Glutamate uptake increase
	Microglia	UD	Outward current induction Inhibition of TNF-α production (cAMP-dependent pathway)
	Oligodendrocyte	UD	DNA synthesis and proliferation enhancement Delay of maturation and myelinogenesis

[a]Receptor nomenclature according to the International Union of Pharmacology (IUPHAR), www.iuphar-db.org.
[b]IUPHAR: receptor family under development.
[c]Biochemical but not pharmacological data.
[d]Anatomical but not functional data.
[e]Immature, not mature, oligodendrocytes.
[f]Occasionally. Akt, protein kinase B; AP-1, activating protein-1; ATPase, adenosine triphosphatase; CALCRL, calcitonin receptor-like receptor; cAMP, cyclic adenosine monophosphate; cGMP, cyclic guanosine monophosphate; DOP, delta receptor; ERK, extracellular signal-regulated kinase; IP, inositol phosphate; KOP, kappa receptor; LPS, lipopolysaccharide; MAPK, mitogen-activated protein kinase; MOP, mu receptor; NF-κB, nuclear factor-kappa B; NPR, natriuretic peptide receptor; PACAP, pituitary adenylate cyclase-activating polypeptide; PDGF, platelet-derived growth factor; PGE2, prostaglandin E2; PI, phosphatidylinositol; PKC, protein kinase C; PLC, phospholipase C; RAMP, receptor activity modifying protein; TNF-α, tumor necrosis factor-α; UD, undefined; VIP, vasoactive intestinal peptide.

may act on astrocytes in an autocrine or paracrine manner. Angiotensin receptors on distinct populations of neonatal rat astrocytes may activate unique signaling pathways and produce different cellular responses. Three different populations have been described: astrocytes containing AT1 subtype coupled to phospholipase C (PLC) activation and prostacyclin release (in the medulla and cerebellum); astrocytes containing high-affinity AT2 receptors which do not activate PLC or release prostacyclin but may be coupled to an additional signaling pathway (in the cerebral cortex); and astrocytes which do not contain high-affinity plasma membrane angiotensin II receptors and do not activate PLC or release prostacyclin (in the thalamic/hypothalamic area).

Most experimental evidence supports the view that brain angiotensin II mediates several classic physiological effects, including blood–brain barrier, body water balance, blood pressure, sodium homeostasis, and blood volume. Beside these classic functions, the brain renin–angiotensin system has more-subtle functions involving complex mechanisms, such as learning, memory, reaction to stress, and control of cerebral circulation.

Atrial Natriuretic Peptide

The family of natriuretic peptides (NPs) comprises atrial, brain, and C-type NPs (ANPs, BNPs, and CNPs), and three NP receptors have been cloned. NP receptor-A (NPR-A) and NP receptor-B (NPR-B) are guanylyl cyclase linked, and they utilize cGMP as intracellular messenger. Both ANP and BNP bind preferentially to NPR-A, whereas CNP binds preferentially to NPR-B. All three NPs bind to the third receptor, known as NP receptor-C (NPR-C). NPR-C is not linked to guanylyl cyclase but appears to act to clear the NPs from circulation. NPs are cardiac and vascular peptides with vasodilatory, diuretic, natriuretic, anti-inflammatory, antifibrotic, and antimitogenic actions. NPs and their mRNA have been localized in forebrain and brain stem regions associated with olfaction, blood pressure regulation, and fluid/electrolytic balance.

ANP has been detected in different types of astrocytes, including glia limitans, and in proximity to microvessels, where the astrocytic bouton processes envelop the vessel wall and play a major role in blood flow regulation. ANP is released from astrocytes through a vesicular exocytotic mechanism in response to cytosolic Ca^{2+} oscillations, as demonstrated by chimeric pro-ANP fused with emerald green fluorescence protein. ANP has a potent vasodilating activity, suppresses the production and signaling of vascular endothelial growth factor, and preserves the functional morphology of endothelial tight junctions. It has been suggested that astrocyte-released ANP might be involved in the development and maintenance of the integrity of the blood–brain barrier, in the regulation of the gliovascular interface and of water and ion homeostasis, and in keeping the brain osmotic microenvironment stable. In pathological conditions, glia-released ANP acts as an anti-vasopermeability factor, which inhibits the development of vascular edema, resulting in protective effects under hypoxic and ischemic conditions.

Astrocytes also express all three types of ANP receptors. In type I astrocytes, it has been shown that ANP decreases Cl^- uptake in a dose-dependent manner. Moreover, astrocyte proliferation is inhibited by NP's binding to C-type receptors, through the inhibition of DNA synthesis associated with inhibition of MAPK activity.

Also microglial cells mRNA coding for ANP receptors (types A, B, and C) and the ANP molecule, thus supporting a possible local paracrine–autocrine regulation. Stimulation of such cells with lipopolysaccharide induces NO and IL-1β production, morphological change in the cells, and the activation of NF-κB and activating protein-1 (AP-1) transcription factors, and all these effects are inhibited by ANP, suggesting that either exogenous or endogenous ANP binds to its receptors on LPS-stimulated microglial cells and thereby inhibits their activities through inhibitions of the transcription factors NF-κB and AP-1.

Calcitonin Gene-Related Peptide

CGRP is a member of the calcitonin peptide family that also includes adrenomedullin (ADM) and two newly isolated peptides, intermedin and calcitonin receptor-stimulating peptide. CGRP and ADM are potent, long-lasting vasodilators, and all major cerebral arteries, small intraparenchymal microvessels, and pial vessels are richly innervated by CGRP-containing nerve fibers. CGRP and ADM have been implicated in cerebrovascular disorders, including migraine headache, vasospasm after subarachnoid hemorrhage, and stroke evolution. While there is no evidence for endogenous synthesis of CGRP, all types of glial cells express CGRP binding sites that function as receptors, forming a complex with the accessory protein from the receptor activity modifying protein family.

In human as well as in rat astrocytes, molecular and pharmacological studies have indicated the presence of type 1 and type 2 receptors and an additional, undefined type. CGRP receptors are also found in the human cerebromicrovascular bed, whereas brain astrocytes, but not microvascular tissue and cells, express functional ADM-R. In view of the suggested importance of astrocytes in the regulation of brain metabolic and ionic homeostasis, local cerebral blood flow, and blood–brain barrier function, CGRP most likely has the ability to modulate some of these various astroglial functions.

CGRP receptors are also expressed by microglia. It has been suggested that neural-derived CGRP from the presynaptic terminals of the neurons, such as motor neuron or primary sensory neurons, activates microglia, leading to immediate early gene and NFκB expression, inducing them to produce proinflammatory cytokines, ATP, and other neuroactive factors, which may further alter neuronal excitability or act on tissue remodeling after lesion. This mechanism has been implicated in, for instance, sustaining hyperalgesia, in which such glial activation could be the driving force, maintaining the pain sensation even after the original injury has healed.

Human oligodendrocytes are also sensitive to calcitonin peptide family stimulation, as indicated by increased cAMP production. ADM and amylin are released by oligodendrocytes, thus suggesting that both peptides act via autocrine/paracrine mechanisms to play functional roles through the elevation of cAMP. It has been reported that some neurons, astrocytes, and brain microvessels have receptors responsive to ADM. ADM released from oligodendrocytes may therefore also affect neighboring cells, such as neurons or astrocytes in the central nervous system (CNS), as a paracrine modulator. The function of CGRP is uncertain in oligodendrocytes, although CGRP has been shown to be an endogenous regulator for modulating cell proliferation in Schwann cells.

Galanin

The expression of galanin in gliallike cells was first reported in colchicine-treated rat brain in NG2- (a chondroitin sulfate proteoglycan) positive cells (Figure 1). NG2-positive cells express properties similar to the oligodendrocyte–type 2 astrocyte progenitor, which can differentiate into oligodendrocyte or type 2 astrocytes and are then classified as oligodendrocyte precursor cells (OPCs). Colchicine-induced galanin expression in OPCs was strongly reduced in hypothyroid rats, suggesting that galanin expression is regulated by the thyroid hormone, possibly via the thyroid hormone-responsive element in the galanin promoter. OPCs express thyroid hormone receptors, and thyroid hormone regulates OPC proliferation and maturation – so galanin may also be involved in these processes. Not only chemical but also electrical stimulation able to activate NG2-positive cells, such as cortical spreading depression (CSD) induced galanin production (mRNA and peptide) in 'activated' NG2-positive cells throughout the entire neocortex, where few or no galanin-positive cells are normally observed. CSD-induced expression of galanin in cerebral cortex may be related to the proliferation of NG2-positive OPCs, as observed after experimental focal ischemia.

Galanin is also expressed by other glial cell types. The peptide and type 2 receptor mRNAs have been found in olfactory ensheathing cells (OECs). OECs are located exclusively in the olfactory bulb, olfactory epithelium, and olfactory nerve. They are specialized glial cells which share some phenotypic characteristics with both the Schwann cells of the peripheral nervous system and the astrocytes of the CNS, serving as a guide to the continuous renewal of olfactory

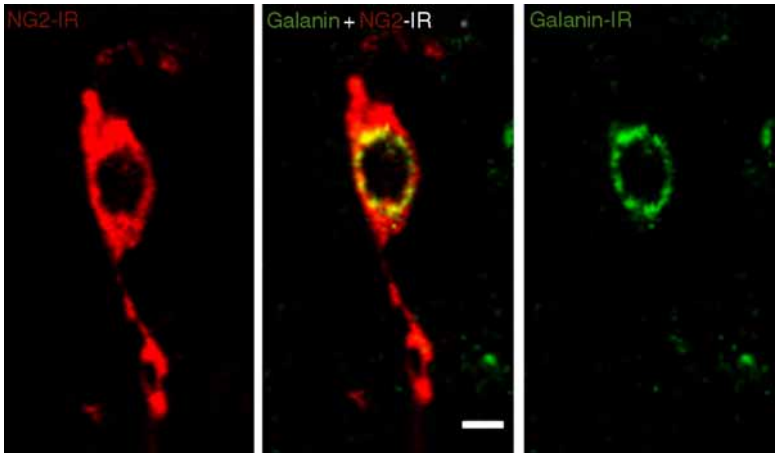

Figure 1 Galanin may be expressed by NG2-positive cells in the grey and white matter after different lesion stimuli, including intraparenchimal colchicine injection. The micrograph illustrates galanin immunoreactivity (IR; green) in NG2-positive cells (red) in the corpus callosum 24 h after colchicine injection in rat. NG2, a membrane sulfated proteoglycan. From Ubink R, Calza L, and Hokfelt T (2003) 'Neuro'-peptides in glia: Focus on NPY and galanin. *Trends in Neuroscience* 26: 604–609.

axons from the olfactory epithelium to the glomerular layer of the olfactory bulb. GAL and galanin agonists can also significantly inhibit the proliferation of OECs without cytotoxic effects.

Galanin, GalR1 and GalR2 mRNA, and galanin (and GalR1) immunoreactivity are strongly expressed in the subventricular zone and in the rostrally migratory stream of normal adult rat. The observed colocalization of galanin with the neuroblast markers polysialic acid–neural cell adhesion molecule and glial fibrillary acidic protein in these areas suggests that galanin is expressed by the progenitor cells behaving as stem cells *in vitro*. Due to the GalR2 mRNA expression in the same areas, galanin is presumed to regulate stem/progenitor cell activity by autocrine and/or paracrine mechanisms.

Galanin in fact inhibits proliferation of poorly differentiated glial precursors, such as OECs, thus suggesting a direct participation in cell cycle control. Galanin is also expressed in neuroblastomas, gangliocytomas, and pituitary neoplasms, and some of these tumors also express galanin receptors, thus suggesting a possible role of galanin in regulating tumor growth; galanin has indeed been included in triple antitumor therapy in experimental cancer, together with the long-acting synthetic somatostatin analog octeotride and serotonin.

Neuropeptide Y

NPY, peptide YY (PYY), and pancreatic polypeptide (PP) are 36 closely related amino acid polypeptides. Receptors for NPY, PYY, and PP are classified together as 'NPY receptors.' At present, five distinct NPY receptors have been established by receptor cloning studies. With regard to endogenous agonists, the receptors Y1, Y2, and Y5 preferentially bind NPY. Y5 receptor mRNA expression is largely restricted to the CNS.

NPY and NPY mRNA are expressed in astrocytes, and synthesis is highly regulated. For example, phorbol 12-myristate-13-acetate increases NPY protein and mRNA levels, suggesting regulation via the protein kinase C pathway. Interleukin-1b induces NPY expression in human, but not rat, astrocytes; this induction involves early genes. Basic fibroblast growth factor has a similar effect on NPY expression in both rat and human astrocytes, whereas forskolin increases NPY levels only in human astrocytes. These data raise the possibility that growth factors might be endogenous regulators of glial peptide expression, a view supported by their temporal, spatial, and developmental expression. In the nucleus accumbens and median eminence, Y1R was found at astrocyte gap junctions, suggesting that it may influence gap junction permeability. Since NPY takes part in a loop system responsible for feedback mechanism in the CNS control of energy homeostasis, this raises the possibility that NPY, acting on Y1R in astrocytes of the arcuate nucleus and median eminence, may regulate astrocyte coupling and hence influence signaling involved in body mass regulation.

NPY and related receptors are also expressed by glial cells different from astrocytes, microglia, oligodendrocytes, and Schwann cells. NPY has been found in OECs and transiently in Schwann cell precursors, which are present at a time during development when many peripheral axons are growing toward their target. The finding that NPY (both peptide and transcript) is expressed close to the CNS, overlapping with the zone in which incoming axons are

sorted before they enter the CNS, is indicative of a functional role related to growth, pathfinding, and/or trophism.

Opioids

After the biochemical purification and characterization of the three endogenous opioid peptides (the enkephalins, the endorphins, and the dynorphins), a fourth member of the family was isolated and termed simultaneously nociceptin, because of its ability to lower the perception threshold for painful stimuli, and orphanin FQ, as endogenous ligand of the orphan Gi/Go-coupled opioid receptor-like 1 receptor. Pharmacological studies have shown that multiple types of opioid receptors exist, today denoted as mu receptor (MOP), delta receptor (DOP), kappa receptor (KOP), and NOP (for nociceptin or orphanin FQ).

Glial-derived opioid peptides and related receptors play a critical role during development and maturation of oligodendrocyte precursors into myelinating oligodendrocytes, the cell which provides the myelin sheath of the axons in the CNS (Figure 2). Oligodendrocytes arise from OPCs in the ventral regions of the germinal neuroepithelium of the ventricular zones at very early stages of embryonic life. The progression of OPCs to mature oligodendrocytes represents a continuum during which the cells exhibit morphological and biochemical diversity. According to developmental stage-related markers, early oligodendrocyte progenitors, prooligodendroblasts, immature oligodendrocytes, and mature myelinating oligodendrocytes can be identified. Developmentally generated OPCs, as identified by the expression of the surface marker NG2, are also present in the mature CNS, where they account for 3–9% of total cell number in the adult CNS, representing an abundant, slowly cycling population of cells which reenter mitosis after injury. Although these cells have a modest rate of cell division, they represent the major dividing population of the adult CNS and are considered responsible for remyelination, a remarkable repair capability of the mature CNS. It is generally accepted that the process of remyelination represents a recapitulation of myelination during development, so molecular regulators active during development are also likely to guide remyelination.

Opioids participate in most of the critical steps of oligodendrocyte maturation. In culture, β-endorphin acts on the MOP and DOP receptors of hippocampal-derived neuroglial precursors in an autocrine manner, increasing proliferation rate through phosphorylation of extracellular signal-regulated kinase 1/2. Beta-endorphin results in a threefold increase in oligodendrogenesis but no significant change in astrogliogenesis.

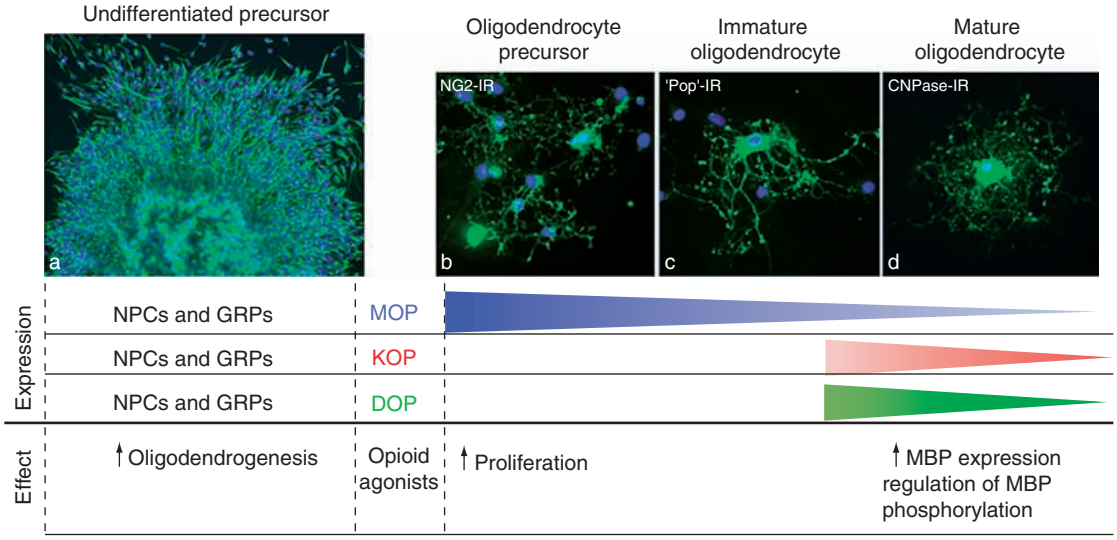

Figure 2 Expression of opiate receptors in oligodendrocytes is dependent on developmental stage and cellular system. *In vitro*, primary cultures of oligodendrocytes express mu receptors (MOPs), coupled to mitogenic signaling, in a very early developmental stage (oligodendrocyte precursor cells, as identified by immunoreactive (IR) staining for the membrane sulfated proteoglycan NG2). MOP expression progressively decreases with maturation (see (c), in which 'Rip'-immunostaining recognizes noncompact myelin protein), and kappa receptors (KOPs) predominate in late developmental stages (see (d), in which 2′,3′-cyclic nucleotide 3′-phosphodiesterase (CNPase) positivity identifies mature oligodendrocytes), influencing differentiation. Also, delta receptor (DOP) seems to increase with maturation, but data are conflicting. When oligodendrogenesis is investigated in cultures derived from undifferentiated precursors (as identified in the neurosphere culture system by the immunostaining for nestin, an intermediate filament expressed by undifferentiated precursors of neuroectodermal origin (a)), all receptor subclasses are expressed by Sox2-(self-renewing, pluripotent neural cells, NPCs) and Nkx2.2-immunoreactive glial restricted precursor (GRP), and opioid agonists seems to induce oligodendrogenesis.

Blockade of opioids reduces proliferation, increases neurogenesis, and reduces gliogenesis. Complementary DNA arrays reveal lowered levels of mRNAs expressed in oligodendrocytes. Very immature oligodendrocytes express MOP receptors, whereas KOP receptors become expressed during late development. Selective opioid agonists and antagonists alter the rate of DNA synthesis and the size of myelin-like membranes in developing oligodendrocytes. This time-dependent development of opioid receptors on precursors reflects a possible role of opioids in determining the pool of oligodendrocytes available for myelin repair during adulthood and also myelin production in demyelinating conditions. It is interesting that leucoencephalopathy has been reported to occur during heroin abuse and to be partially reversible after withdrawal, thus supporting the hypothesis that opioids inhibit proliferation and maturation toward myelinating oligodendrocytes. Moreover, β-endorphin increases the mRNA levels of myelin basic protein, and morphine regulates phosphorylation of myelin basic protein in the rat brain.

Substance P

Tachykinins are a family of closely related peptides whose best-known members are substance P (SP), neurokinin A (NKA), and neurokinin B (NKB). NKB is present in the CNS and the spinal cord, and SP and NKA are found in the CNS and in primary afferent sensory neurons supplying a number of peripheral tissues. The three types of tachykinin receptors, denoted as neurokinin-(NK) 1, NK2, and NK3 receptors, are heterogeneously distributed within the tissues in each species.

Tachykinin synthesis by glial cells is not an ancillary phenomenon. SP has also been found to be expressed by human fetal brain microglia, and quantitative analysis of the transcript has indicated that brain microglia express higher levels than do peripheral immune cells, that is, monocyte-derived macrophages and lymphocytes, thus suggesting that brain microglia can be an important source of SP in the CNS.

Emerging evidence indicates that tachykinins regulate the immune functions of CNS glial cells. NK1 receptors have a high affinity for SP and are expressed by a number of resident CNS cell types, including astrocytes and microglia. In the human postmortem brain, the NK3 receptor immunoreactivity was found in the glia limitans at the pial surface, where astrocytes and beaded fibers stained intensely. In the white matter, astrocytes and beaded fibers displayed NK3 receptor immunoreactivity, particularly in areas surrounding blood vessels. SP–NK1 receptor interactions elicit activation of signal transduction pathways in both cell types and can initiate, or augment, inflammatory responses by astrocytes and microglia.

Tachykinins also clearly illustrate how glial cells can participate in pathological events, such as persistent pain resulting from either neuroinflammation or neuroimmune activation. Astrocytes and microglia occurring in the vicinity of the terminals in the dorsal horns of the spinal cord are activated in pain states. Such activation could arise from 'pain transmitters' such as SP, CGRP, glutamate, and/or ATP. It is well known that astrocytes and microglial cells are the main sources of ATP. Glutamate can stimulate ATP release from spinal cord astrocytes, and this release is greatly enhanced when SP is present.

Vasoactive Intestinal Peptide/Pituitary Adenylate Cyclase-Activating Polypeptide

VIP and PACAP are members of a superfamily of structurally related peptide hormones that includes glucagon, glucagon-like peptide, secretin, and growth hormone-releasing hormone. At least three receptors for PACAP exist in mammals, two of which are also high-affinity receptors for VIP, all widely distributed in the nervous tissue: VPAC1 receptor, which responds to VIP and PACAP with comparable affinity; VPAC2 receptor, first cloned from rat olfactory bulb, which also responds to VIP and PACAP with comparable affinity; and PAC1 receptor, which recognizes PACAP with much higher potency than it does VIP.

The classical roles attributed to astrocytes in brain function include cerebrovascular regulation through neurovascular coupling. The active neurons increase energy demand, and this is linked to the local increases in blood flow in the area of elevated neuronal activity. Astrocytes, which are much more numerous than neurons, send processes to synapses and blood vessels and communicate with other astrocytes through gap junctions and by release of ATP. Because of their polarized anatomical structure and of the vicinity of their boutons to contractile elements of blood vessels, such as smooth muscle cells in arterioles and pericytes in capillaries, astrocytes have long been suggested to contribute to the regulation of blood flow during neuronal activity.

In this context, VIP/PACAP peptides offer an example of peptidergic regulation of glia/neuron communication. Originally, VIP was thought to act predominantly by directly modulating neuronal activity. More recently, it has become clear that VIP influences neurons in the cortex through activation of VIP receptors located on astrocytes. VIP and PACAP are also key regulators in glycogen metabolism in

astrocytes, where they regulate expression of a number of genes encoding for enzymes of energy metabolism regulation.

In vivo, astrocytes envelop synapses so intimately that they are thought to compartmentalize the space between cells. One astrocyte may envelop several of the synapses of several neurons, and one neuron may have its synapse enveloped by several astrocytes. Neuron-derived VIP and PACAP are important in regulating astrocyte participation in glutamatergic synaptic function. PACAP increases glutamate uptake by promoting expression of glutamate transporters in astrocytes, and VIP also enhances glutamate uptake by astrocytes, thus decreasing glutamate toxicity.

Moreover, VIP and PACAP participate in the regulation of astroglial precursor proliferation and neural stem cell differentiation and in microglia biology during activation, acting as anti-inflammatory agents.

Future Directions

Neuropeptides represent a highly active area of research in neurobiology and neuropharmacology, and much of the data listed above probably need to be updated in the light of the emerging roles of glial cells in the physiology and pathology of the nervous system. Moreover, there is increasing evidence that GPCRs are present in nuclear or perinuclear membranes. Nuclear localization of NK1, CGRP, and SP has been reported in a subpopulation of dorsal root ganglia cells in rat; AT1 in the kidney, vascular smooth muscle cells, and neural cell lines; and NPY/Y1 in endocardial endothelial cells and opioid receptors in embryonic stem cells. Finally, neurotensin receptor 3, originally defined as sortilin, is a protein for the intracellular trafficking of this receptor, which leads to the internalization of the NT/NT3 complex via the endosomal pathway in different cell types. Potential roles for nuclear signaling pathways activated by neuropeptides/neuropeptide receptors include the regulation of nuclear transport, gene expression, nuclear envelope formation, and cell cycle and cell differentiation regulation.

See also: Angiotensin II; Atrial Natriuretic Peptide: Fluid/Mineral Balance; Calcitonin Gene-Related Peptide (CGRP) and Receptors; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Galanin and Receptors; Invertebrate Neurohormone GPCRs; Mammalian Neuropeptide Families; Natriuretic Peptides; Neuropeptide Y (NPY) and its Receptors; Neuropeptides: Discovery; Opioid Peptides and Receptors; Substance P/Tachykinins and its/their Receptors; Vasoactive Intestinal Peptide and Pituitary Adenylate Cyclase Activating Peptide Receptors.

Further Reading

Brain SD and Cox HM (2006) Neuropeptides and their receptors: Innovative science providing novel therapeutic targets. *British Journal of Pharmacology* 14: S202–S211.

Buch SK, Khurdayan VK, Lutz SE, Knapp PE, El-Hage N, and Hauser KF (2007) Glial-restricted precursors: Patterns of expression of opioid receptors and relationship to human immunodeficiency virus-1 Tat and morphine susceptibility in vitro. *Neuroscience* 146(4): 1546–1554.

Deitmer JW, McCarthy KD, Scemes E, and Giaume CB (2006) *Special Issue: Calcium Signaling in Glial Cells and Mechanisms of Transmitter Release from Astrocytes*. *Glia* 54(7).

Deschepper CF (1998) Peptide receptors on astrocytes. *Frontiers in Neuroendocrinology* 19: 20–46.

Haydon PG and Camignoto G (2006) Astrocyte control of synaptic transmission and neurovascular coupling. *Physiological Review* 86: 1009–1031.

Hokfelt T, Bartfai T, and Bloom F (2003) Neuropeptides: Opportunities for drug discovery. *Lancet Neurology* 2: 463–472.

Hokfelt T, Johansson O, Ljungdahl A, Lundberg JM, and Schultzberg M (1980) Peptidergic neurons. *Nature* 284: 515–521.

Hosli E and Hosli L (1993) Receptor for neurotrasmitters on astrocytes in the mammalian central nervous system. *Progress in Neurobiology* 40: 477–506.

Jabs R, Pivneva T, Huttmann K, et al. (2005) Synaptic transmission onto hippocampal glial cells with hGFAP promoter activity. *Journal of Cell Science* 118: 3791–3803.

Káradóttir R and Attwell D (2006) Neurotransmitter receptors in the life and death of oligodendrocytes. *Neuroscience* 145(4): 1426–1438.

Kastin A-J (2006) *Handbook of Biologically Active Peptides*. Amsterdam: Academic Press.

Ubink R, Calza L, and Hokfelt T (2003) 'Neuro'-peptides in glia: Focus on NPY and galanin. *Trends in Neuroscience* 26: 604–609.

Von Rossum D and Uwe-Karsten H (2004) Microglia. *Metabolic Brain Disease* 19: 393–411.

Relevant Website

http://www.iuphar-db.org – International Union of Pharmacology (IUPHAR) receptor database.

Neuropeptides in Autonomic Neurons

L Klimaschewski, Medical University of Innsbruck, Innsbruck, Austria

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Autonomic ganglia are widely used as models for studying the establishment and plasticity of neurochemical organization in the autonomic nervous system. They contain well-defined and rather homogenous neuron populations which express a variety of neurotransmitters and neuromodulators (i.e., classical transmitters and neuropeptides in different combinations). The multipolar postganglionic neurons (PGNs) extend 10–15 dendrites and a single, mostly unbranched axon as revealed by silver studies and intracellular dye injection.

The PGNs are classified morphologically and neurochemically by the target they innervate. They regenerate readily, making it possible to study changes in neuropeptide content, for example, in response to axonal lesion. Above all, PGNs may be maintained *in vitro* to investigate effects of neurotrophic factors, cytokines and genetic manipulations on transmitter and neuropeptide expression. In addition, autonomic ganglia contain a variable number of small paraganglionic cells, called small, intensely fluorescent (SIF) cells because of their formaldehyde-induced catecholamine fluorescence. These are mainly located in clusters and ultrastructurally characterized by abundant large granular vesicles containing catecholamines and various neuropeptides.

Neuropeptides differ from classical transmitters with respect to their molecular weight, synthesis, storage, release, action, and metabolism. Peptide precursors are synthesized at the rough endoplasmic reticulum and, after enzymatic cleavage, sorted into large dense-core vesicles (LDV; ~ 100 nm in diameter). This step determines specificity with regard to the neurochemical phenotype and takes place at the level of the *trans*-Golgi network.

Different vesicle populations may be sorted into distinct terminals of the same neuron, suggesting that nerve endings of a particular neuron exhibit different neurochemical codes. Transmitters are stored in small synaptic vesivcles at the presynaptic membrane (active zone) and released rapidly after low-frequency stimulated exocytosis. On the other hand, neuropeptide-containing LDVs are randomly distributed in the nerve terminal, which supports the hypothesis that parasynaptic peptide release requires high frequency stimulation and proceeds in a significantly slower time course. Neuropeptide receptors represent guanine nucleotide-binding protein (G-protein)-linked molecules, consisting of seven membrane-spanning hydrophobic sequences. Neuropeptides are inactivated by proteolysis or by diffusion, whereas the action of classical transmitters is terminated by enzymatic cleavage, specific reuptake mechanisms or by diffusion.

Already at early stages of development, catecholamines and neuropeptides are detectable in autonomic ganglia, with decreasing cell numbers until birth. It appears likely that the restriction of neuropeptide expression to specific subpopulations of autonomic neurons is controlled by extracellular molecules encountered by the growth cone during axon outgrowth and finally by target-derived signals retrogradely transported after organ contact is established. Target tissues, therefore, not only play an important role in regulating the soma size, number, and distribution of neurons but also in determining the species-dependent neurochemical code of individual PGNs. However, we emphasize that no exclusive end organ–mediator relationship exists in autonomic ganglia. Rather, specific tissues (muscle or secretory cells) within the target appear to be the main regulator of neuropeptide synthesis.

Functional subpopulations of pre- and postganglionic neurons (e.g., vasomotor, pilomotor, or secretomotor circuits) are specified by their neuropeptide content (Table 1). During development and adult neuroplasticity, intraganglionic soluble factors derived from satellite cells, macrophages, or neighboring PGNs, as well as electrical activity and cell contact, alter the phenotype of PGNs. It becomes clear that the synthesis of each neuropeptide is regulated by more than one instructive factor or experimental condition and that, vice versa, neuronal differentiation factors induce not just one particular neuromodulator but affect a set of transmitter/peptide genes. In this article, the neuropeptides expressed by pre- and postganglionic autonomic neurons are discussed in alphabetical order regarding their localization, regulation, and function.

Calcitonin Gene-Related Peptide

The 37-amino-acid calcitonin gene-related peptide (CGRP) was shown to be generated by alternative processing of RNA transcribed from the calcitonin gene. CGRP is detected in a population of noncatecholaminergic, mostly vasoactive intestinal polypeptide (VIP)-immunoreactive, and sometimes also substance P

Table 1 The most common neuropeptides expressed in pre- and postganglionic neurons of functionally identified autonomic circuits in various mammals

	Vasoconstrictor	*Vasodilator*	*Sudomotor*	*Pilomotor*	*Muscle vasoconstrictor*	*Muscle vasodilator*
CGRP			+ (post)		+ (pre)	
ENK/DYN		+ (pre)	+ (pre)	+ (pre, post)	+ (pre)	+ (pre)
GAL	+ (post)			+ (post)	+ (post)	
NPY	+ (post)				+ (post)	+ (post)
SP	+ (pre)		+ (post)			+ (pre)
VIP/PACAP		+ (post)	+ (post)			+ (post)

(SP)-immunoreactive PGNs in different species. Axodendritic synaptic contacts between CGRP-immunoreactive varicosities and PGNs represent links between the pre- and postganglionic efferent pathway. Consequently, CGRP-positive axons are absent from sympathetic ganglia after transection of the sympathetic trunk. Moreover, CGRP-containing nerve terminals provide the morphological correlate for reflex loops between sensory axon collaterals and PGNs. Activin A, a member of the transforming growth factor superfamily, but not leukemia inhibitory factor (LIF) or related cytokines, induce CGRP in sympathetic neurons.

Enkephalins

Two short pentapeptides, Met-enkephalin (ENK) and Leu-ENK, are derived from three different precursors; Met-ENK is included in proenkephalin A and in proopiomelanocortin, whereas Leu-ENK is derived from proenkephalin A or proenkephalin B. The latter is known as prodynorphin. The heptapeptide Met-ENK-Arg-Phe (MEAP) and the octapeptide Met-ENK-Arg-Gly-Leu (MEAGL) are derived from proenkephalin A and converted to Met-ENK. ENK- and MEAP-positive PGN are present in autonomic ganglia. ENK-immunoreactive PGN are catecholaminergic but neuropeptide Y (NPY)- and VIP-negative. Dynorphin is localized in most NPY-immunoreactive and in some NPY-nonreactive PGNs that project to arterioles and piloerector muscles. ENK-immunoreactive axons form pericellular baskets around PGN and establish synaptic contacts with dendrites of PGN. The ENK content increases in response to axotomy but not after decentralization of sympathetic ganglia. However, denervation of the adrenal gland upregulates preproenkephalin mRNA in chromaffin cells. Chronic treatment with the opiate receptor antagonist naloxone elevates ENK levels in the superior cervical ganglion by blocking ENK release from preganglionic nerves.

Mainly inhibitory functions have been assigned to ENK and its derivatives by the inhibition of voltage-dependent Ca^{2+} influx or by opening potassium channels. Physiological studies suggest that the release of acetylcholine from preganglionic nerves is inhibited by MEAP. Compared to the depressant effects of catecholamines on preganglionic acetylcholine release, ENK appears to be up to 100 times more potent. The amplitude of the fast excitatory postsynaptic potential (EPSP) is depressed after application of Leu- or Met-ENK onto PGNs. This effect is prevented by the addition of the opioid antagonist naloxone. ENK does not affect the cholinergic potential induced by the iontophoretic application of acetylcholine onto these neurons. In addition, endogenous opioids mediate inhibition by the activation of postsynaptic opiate receptors. Whereas κ-opioid receptor stimulation inhibits Ca^{2+} uptake through voltage-sensitive channels, δ- and μ-opioid receptors signaling inhibits adenylate cyclase and the activation of neuronal K^+ channels.

Galanin

The 124-amino-acid precursor includes the 29-amino-acid peptide galanin (GAL; the name is derived from the N-terminal glycine and C-terminal alanine residues) and a 60-amino-acid GAL mRNA-associated peptide (GMAP). GAL-immunolabeling is observed in catecholaminergic PGNs which, to a large extent, coexpress NPY. Surgical and chemical denervation (applying acetylcholinesterase (AChE) antibodies) produces a small increase in GAL- and GMAP-immunostaining. Major changes in GAL and GAL mRNA have been found in response to axotomy of autonomic neurons. The signal mainly responsible for increased GAL production in PGNs appears to be LIF because LIF-deficient mice exhibit only a small increase in GAL levels after axotomy. GAL-positive PGNs project to blood vessels of the muscle and skin as well as to the piloarrector muscles. However, compared to the number of positive perikarya, the frequency of GAL-immunoreactive nerve endings in peripheral target tissues is low.

GAL inhibits the release of neurotransmitters by activating a pertussis toxin-sensitive G-protein. Specifically, GAL decreases vagal activity via presynaptic

inhibition of acetylcholine release from parasympathetic nerve terminals, and, in cardiac ganglia, GAL induces a hyperpolarizing potential in the majority of neurons probably due to the activation of a potassium conductance.

Neuropeptide Y

The 36-amino-acid neuropeptide tyrosine (NPY) belongs to a family of structurally related pancreatic polypeptides that include peptide YY. NPY is highly concentrated in large LDVs of sympathetic postganglionic neurons. Scattered, mainly small PGNs exhibit immunolabeling for NPY, and a majority belong to the catecholaminergic neuron population. Some muscle-vasodilator NPY-immunoreactive PGNs in paravertebral ganglia contain co-localized VIP and lack tyrosine hydroxylase (TH)-immunoreactivity. NPY is co-released with noradrenaline upon preganglionic activity to induce long-lasting vasoconstriction (Figure1). Treatment with reserpine or 6-OH-dopamine results in a parallel depletion of both NPY and catecholamines at nerve terminals in the target tissues of PGNs, whereas NPY mRNA and peptide levels increase in the corresponding ganglia. NPY levels decrease in the superior cervical ganglion 1 week after transection of the cervical sympathetic trunk, whereas after axotomy the peptide is elevated but NPY mRNA is downregulated. In genetically hypertensive rats, NPY levels are reduced in the superior cervical ganglion, whereas increased circulating NPY levels are detected in patients with essential hypertension or with pheochromocytoma.

Compared to noradrenaline, NPY-release mainly depends on high-frequency stimulation of PGNs and is facilitated by presynaptic β-receptor activation. NPY apparently is responsible for the α-adrenoceptor antagonist-resistant vasoconstriction in response to stimulation of sympathetic nerves. The large amidated C-terminal part of NPY is necessary for receptor binding, inhibition of cyclic adenosine monophosphate (cAMP) formation, and vasoconstriction caused by stimulation of Y_1 receptors. The presynaptic NPY receptor is of the Y_2 type and mediates the suppression of noradrenaline, serotonin, and NPY release. NPY-mediated suppression of transmitter release from the sympathetic nerve endings by reducing Ca^{2+} influx is due to the selective inhibition of neuronal N-type Ca^{2+} channels. Finally, NPY directly inhibits the catalytic activity of the catecholamine-synthesizing enzymes dopamine β-hydroxylase (DBH) and TH *in vitro*.

Somatostatin

Somatostatin (SOM), the most important inhibitor of pituitary growth hormone secretion, is a cyclic 14-amino-acid peptide isolated from the bovine

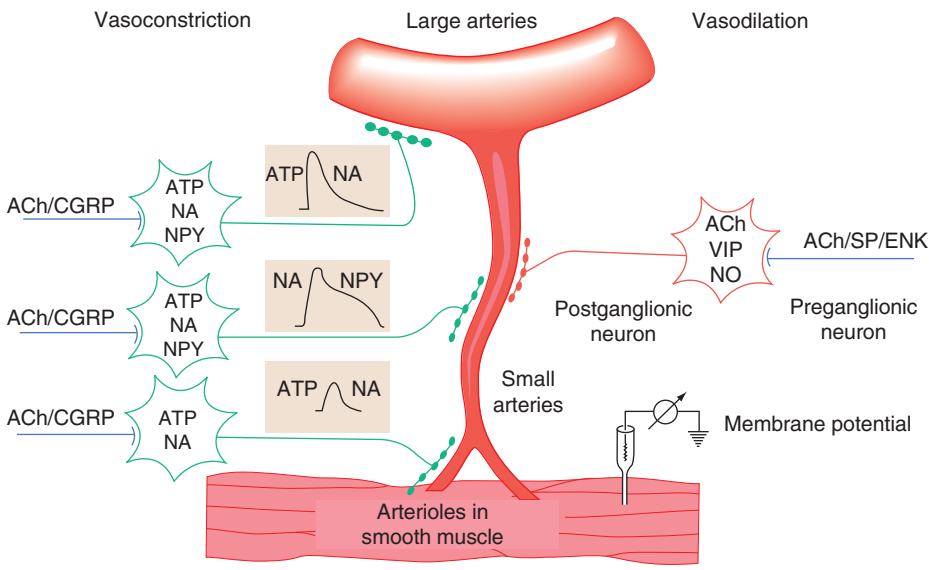

Figure 1 Neuropeptides released from autonomic neurons regulate tissue perfusion. Sympathetic vasoconstrictor neurons contain noradrenaline (NA), adenosine triphosphate (ATP), and neuropeptide Y (NPY). They are contacted by cholinergic preganglionic neurons expressing calcitonin gene-related peptide (CGRP). NPY is predominantly acting at small arteries, whereas ATP modulates noradrenergic transmission at larger arteries and fine arterioles within the smooth muscle. Vasodilation is induced by parasympathetic neurons synthesizing acetylcholine (ACh), nitric oxide (NO), and vasoactive intestinal peptide (VIP). Substance P (SP) and enkephalins (ENKs) are neuromodulators expressed in preganglionic neurons innervating them. Adapted from Gibbins IL and Morris JL (2000) Pathway specific expression of neuropeptides and autonomic control of the vasculature. *Regulatory Peptides* 93: 93–107.

hypothalamus. SOM is mainly detected in autonomic neurons throughout development. In adults, only few TH-positive PGNs contain SOM-immunoreactivity. In prevertebral ganglia, for example, these neurons project to the secretomotor neurons in the submucosal ganglia. SOM-immunoreactive axons may be of preganglionic origin because neurons in the nucleus intermediolateralis have been shown to contain SOM-immunoreactivity. SOM serves as a good example of the differential regulation of neuropeptides on mRNA and peptide levels because co-culture of autonomic neurons with ganglionic nonneuronal cells increases prepro-SOM mRNA without altering SOM peptide level. Two different signals derived from satellite cells appear to be responsible for this observation: a soluble factor stimulating the transcription of the prepro-SOM gene and a membrane-associated factor inhibiting peptide synthesis. SOM synthesis is induced by LIF because LIF treatment results in higher SOM mRNA levels. This effect is reversible and appears to be selective because SOM is not increased by this cytokine in sensory neurons. Inhibitory effects of SOM on peripheral noradrenergic neurons include the blockade of a Ca^{2+} current and inhibition of the K^{+}-induced release of acetylcholine.

Substance P

SP belongs to a family of closely related neuropeptides, called tachykinins or neurokinins, which include neurokinin A (NKA), neurokinin B (NKB), neuropeptide K (NPK), and neuropeptide γ (NPγ). Multiple preprotachykinin (PPT) mRNAs arise from a single gene through alternate splicing of the primary transcripts. SP-immunoreactivity has been demonstrated in few sympathetic PGNs, sometimes co-localized with ENK. Whereas the co-localization of SP and CGRP is characteristic of primary sensory neurons, sympathetic PGNs containing both SP and CGRP rarely occur. However, SP/CGRP-positive axons are observed frequently in the autonomic ganglia. They are diminished after postganglionic nerve transection, suggesting that these fibers may represent axon collaterals of sensory neurons. Regarding the origin, distribution, and connectivity of SP-immunoreactive intraganglionic fibers, SP may therefore act as a modulator in local reflex arcs involving PGNs, interconnected SIF cells, or short loop circuits between sensory axon collaterals and PGNs. SP peptide and mRNA are elevated in response to decentralization or axotomy of autonomic neurons. As with VIP, the elevation of SP after axotomy is prevented by dexamethasone, which blocks SP expression induced by interleukin (IL-)1β. IL-1β is synthesized in the sympathetic ganglia and released in response to tumor necrosis factor, which by itself stimulates SP expression in co-cultures of PGNs with nonneuronal cells. Because SP is not elevated in sensory neurons, the effects of IL-1β on SP expression are likely to be selective and, at least in part, mediated by LIF. This suggests that SP expression may be linked to cholinergic development because LIF is identical to cholinergic differentiation factor (CDF). In fact, similarities exist in the regulation of both traits.

SP released from the sympathetic ganglia by depolarization potentiates the amplitude of fast EPSPs, resulting in spike discharge. Physiological studies suggest a role for SP in evoking a slow long-lasting membrane depolarization of PGNs that is influenced by enkephalinergic presynaptic inhibition. SP specifically inhibits the increase in TH activity observed after treatment with nicotinic agonists. However, in cultures of dissociated superior cervical ganglia (SCGs) exogenously applied, SP, like SOM, modulates catecholamine metabolism by increasing TH activity. At the spinal cord level, SP increases sympathetic activity and arterial pressure by activating preganglionic neurons via neurokinin (NK-)1 receptors.

Vasoactive Intestinal Polypeptide and Pituitary Adenylate Cyclase-Activating Polypeptide

VIP, a linear 28-amino-acid peptide, was isolated from porcine gut and is structurally related to the peptides of the secretin–glucagon family. VIP is derived from a larger precursor molecule that contains another biologically active peptide, peptide histidine isoleucin or, in humans, histidine methionine-27. Only few PGNs contain VIP. In contrast, a large proportion of chick lumbosacral sympathetic neurons and 10–15% of mainly cholinergic PGNs in the stellate ganglion exhibit VIP-immunoreactivity. Strongly VIP-positive PGNs occurring in clusters are often localized in the ganglion periphery. They are distinguished from scattered cells exhibiting moderate VIP-immunostaining. The latter cell bodies co-localize CGRP or SP and probably project to the sweat glands. VIP is found in some PGNs containing NPY and projecting to smooth muscle of the airways. Clustered CGRP-nonreactive, but VIP-immunoreactive, PGNs are preferentially apposed by SOM- and CGRP-immunoreactive fibers and probably project to the smooth muscle.

An increase in the number of VIP-immunoreactive cell bodies and nerve fibers is observed after axonal lesion of SCG neurons and in ganglia maintained in organ culture. IL-1β has been found to potentiate

this effect, whereas dexamethason reduces the increase in VIP-immunoreactivity. The increase in peptide content is accompanied by an upregulation of VIP mRNA. As indicated by LIF knockout and antibody studies, LIF/CDF induces VIP mRNA in response to axotomy of the sympathetic ganglia. However, because the induction of VIP after axotomy is reduced only by 60–70% in LIF-deficient mice compared to the lesioned wild-type SCG, other factors are likely to be involved. Although VIP-immunoreactive nerve fibers in the sympathetic ganglia, in part, are derived from preganglionic neurons, an increase of VIP-immunoreactivity is observed after decentralization (i.e., transection of the preganglionic fibers of the SCG).

The cooccurrence of acetylcholine and VIP in sympathetic neurons suggests a cooperative function in the regulation of blood vessels (Figure 1). Indeed, VIP has been found to act as a potent vasodilator. VIP mediates the atropine-resistant vasodilation and increases the secretory response to acetylcholine in the cat submandibular salivary gland. With respect to its possible neuronal functions, VIP induces a voltage-dependent inhibition of N-type Ca^{2+} channels in rat sympathetic neurons. VIP enhances synaptic transmission, probably by increasing acetylcholine release from presynaptic terminals and facilitates the electrophysiological response to muscarinic agonists in the SCG. In parasympathetic ganglia, VIP produces depolarization associated with spontaneously firing action potentials and decreased membrane conductance. Furthermore, the noncholinergic component of the transsynaptic stimulation of TH activity may be due to VIP released from preganglionic fibers. VIP rapidly increases TH activity by elevating intracellular cAMP levels as well as TH mRNA synthesis, followed by an increase in TH activity 2 days later. This long-term neuromodulatory role of VIP may be responsible for the described time course of TH activity after axotomy. Because a significant increase of the intraganglionic VIP content has been observed, a further decrease of TH activity after an initial fall may be prevented by VIP, possibly mediated by a rise in intracellular inositol phosphate. With regard to axonal regeneration after lesion, VIP has been found to induce neurite outgrowth, to stimulate mitosis, and to enhance the survival of sympathetic neuroblasts *in vitro*. Moreover, VIP delays neuronal death for approximately 6 h after nerve growth factor (NGF) deprivation of explanted rat neonatal SCG neurons.

Similar effects on autonomic neurons are mediated by the related pituitary adenylate cyclase-activating peptide (PACAP), which is found in two amidated forms (PACAP38 and PACAP27) and shows a 68% sequence homology with VIP. There is increasing evidence that PACAP acts as a neuromodulator in the autonomic nervous system. In peripheral ganglia, PACAP expression is detected in sensory, vagal, and cranial parasympathetic neurons. In otic and sphenopalatine ganglia, many neurons are positive for both PACAP and VIP. PACAP is upregulated in a subpopulation of PGNs after axotomy and probably participates in neuronal development and nerve regeneration.

The effects of PACAP are mediated by receptors that interact with PACAP and VIP with similar high affinity (VPAC1 and VPAC2), whereas PAC1 is highly specific for PACAP. PACAP/VIP receptors are coupled to adenylate cyclase via G_s and activate downstream signaling molecules, including mitogen-activated protein kinase (MAPK), phosphatidylinositol 3-kinase, and phospholipase C.

Other Neuropeptides

Atrial natriuretic polypeptide is detected in PGNs of the sympathetic ganglia and participates in the inhibition of the carbachol-induced increase in catecholamine synthesis. Bombesin/gastrin-releasing peptide has been demonstrated in axons but not in PGNs or SIF cells of the sympathetic ganglia.

Neurotensin (NT)-immunoreactive nerve fibers occur around CGRP/VIP-immunoreactive PGNs of the SCG and stellate ganglion. Such fibers are probably derived from preganglionic NT-immunoreactive neurons which are found in the spinal intermediolateral nucleus. NT appears to participate in the noncholinergic excitation of PGN.

Vasopressin-immunoreactivity is present in sympathetic PGNs, and a novel neuropeptide, secretoneurin, a cleavage product of secretogranin II, is highly expressed in autonomic neurons and in SIF cells.

See also: Calcitonin Gene-Related Peptide (CGRP) and Receptors; Galanin and Receptors; Invertebrate Neurohormone GPCRs; Mammalian Neuropeptide Families; Neuropeptide Signaling in Invertebrates; Neuropeptide Inactivation or Metabolism; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptide Y (NPY) and its Receptors; Neuropeptides: Electrophysiology; Neuropeptides: Endocrine Cells; Neuropeptides: Enteric Nervous System; Somatostatin and Receptors.

Further Reading

Benarroch EE (1994) Neuropeptides in the sympathetic system: Presence, plasticity, modulation, and implications. *Annals of Neurology* 36: 6–13.

Elfvin L-G, Lindh B, and Hökfelt T (1993) The chemical neuroanatomy of sympathetic ganglia. *Annual Review of Neuroscience* 16: 471–507.

Gibbins IL (1994) Comparative anatomy and evolution of the autonomic nervous system. In: Nilsson S and Holmgren S (eds.) *Comparative Physiology and Evolution of the Autonomic Nervous System*, pp. 1–67. Chur, Switzerland: Harwood Academic Publishers.

Gibbins IL and Morris JL (2000) Pathway specific expression of neuropeptides and autonomic control of the vasculature. *Regulatory Peptides* 93: 93–107.

Hökfelt T (1991) Neuropeptides in perspective: The last ten years. *Neuron* 7: 867–879.

Klimaschewski L, Kummer W, and Heym C (1996) Localization, regulation and functions of neurotransmitters and neuromodulators in cervical sympathetic ganglia. *Microscopy Research and Technique* 35: 44–68.

Lundberg JM and Hökfelt T (1983) Coexistence of peptides and classical neurotransmitters. *Trends in Neuroscience* 6: 325–333.

Morris JL and Gibbins IL (1989) Co-localization and plasticity of transmitters in peripheral autonomic and sensory neurons. *International Journal of Developmental Neuroscience* 7: 521–531.

Schmidt RF and Unsicker K (eds.) (2003) *Lehrbuch Vorklinik*. Cologne: Deutscher Aerzte-Verlag.

Schultzberg M and Lindh B (1988) Transmitters and peptides in autonomic ganglia. In: Björklund A, Hökfelt T, and Owman C (eds.) *Handbook of Chemical Neuroanatomy, Vol. 6. The Peripheral Nervous System*, pp. 297–315. Amsterdam: Elsevier Science.

Neuropeptides Internalization

J M Jimenez-Andrade and P W Mantyh, University of Minnesota, Minneapolis, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Over the past few decades, numerous neuropeptides and their receptors have been discovered in various regions of the peripheral and central nervous systems. Likewise, an increasing number of studies have revealed that neuropeptides play a significant role in the regulation of nervous system function in normal and pathological conditions.

The biological actions of neuropeptides are mainly mediated by activating either G-protein-coupled receptors (GPCRs), ligand-gated ion channels, cytokine receptors, or receptor tyrosine kinases. In this article, we focus on GPCRs, as the majority of neuropeptides bind to this class of receptor. The GPCRs, once activated, may regulate the function of various protein kinases and enzymatic effectors and modulate the transcription of numerous target genes. An important feature of GPCRs is that the activation of these receptors also triggers 'feedback' mechanisms that protect against both acute and chronic receptor overstimulation. According to SG Ferguson, these mechanisms involve (1) uncoupling of the receptor from heterotrimeric G-proteins (desensitization), (2) the internalization of cell surface receptors to intracellular membranous compartments, and (3) the downregulation of the total cellular complement of receptors due to reduced receptor mRNA and protein synthesis, the lysosomal degradation of preexisting receptors, or recycling of receptors to the cell surface.

A wide variety of neuropeptides and their receptors have been shown to undergo internalization (endocytosis), including angiotensin, neurotensin, bradykinin, corticotropin-releasing factor, somatostatin, vasoactive intestinal peptide, and substance P. In the following discussions we focus on the substance P and neurokinin-1 (SP–NK-1) system to discuss some of the features, functional consequences and the potential therapeutic usefulness of neuropeptide receptor-induced internalization.

Neurobiology of Substance P and NK-1 Receptor

Substance P belongs to the tachykinin neuropeptide family which also includes neurokinin-A (NK-A) and neurokinin-B (NK-B). All tachykinins are characterized by the C-terminal sequence-Phe-Xaa-Gly-Leu-Met-NH_2. Substance P was first identified in 1931 as a component of brain and intestinal extracts which possessed smooth muscle contractile activity. Early studies have shown that substance P is present primarily in the gray matter of the central nervous system (CNS), with the highest levels seen in the hypothalamus and substantia nigra. Substance P was also found to be expressed in the spinal cord, mostly in dorsal areas, as well as in the peripheral nervous system (PNS), particularly in the autonomic nerves, spinal ganglia, and sympathetic trunk.

Comparative analysis of pharmacologic properties of substance P, NK-A, and NK-B has led to the recognition of three distinct NK receptors, each with a preferred ligand: substance P preferentially binds to the NK-1 receptor, whereas NK-A and NK-B show preference for the NK-2 and NK-3 receptors, respectively. While each NK possesses agonist properties at all three receptor types, the biological activity of these systems appears to be governed by both co-localization and affinity of various NKs for different NK receptors. The NK-1 receptor is a G-protein-coupled receptor which, when activated, induces inositol phospholipid hydrolysis and, in some cases, adenylate cyclase.

The SP–NK-1 receptor pathway is an abundant and extensively studied neuropeptide system that is present in the CNS, PNS, and nonneuronal tissues and that regulates a wide variety of functions. Some of the most-studied functions of the SP–NK-1 system include modulation of the signaling of nociception, regulation of chemotherapy-induced nausea and vomiting, and control of respiration.

Localization of Substance P and NK-1 Receptors in the Brain and Spinal Cord

Autoradiographic, immunohistochemical, and messenger RNA (mRNA) expression studies have documented broad distribution of substance P and NK-1 receptors in the CNS. Intense substance P staining has been observed in amygdala, locus coeruleus, hypothalamus, substantia nigra, and peduncular nuclei, whereas moderate labeling has been detected in caudate putamen, nucleus accumbens, and raphe nuclei, as well as in the lamina I of the spinal cord. Relatively low, but detectable, levels of substance P are present in the cerebral cortex, cerebellum, and hippocampus. Substance P has been shown to be expressed in both neuronal cell bodies and dendrites, to be localized to synaptic vesicles, and to be released in a calcium-dependent manner. Taken together, these findings suggest that substance P may be a neurotransmitter involved in regulation of nociceptive transmission as

well as emotional and stress responses. In contrast to receptors for other known neurotransmitters (e.g., serotonin, norepinephrine), which are expressed by virtually all neurons in a given CNS region, NK-1 receptors are expressed by a small percentage (5–7%) of neurons in various regions of the brain and the spinal cord (Figure 1).

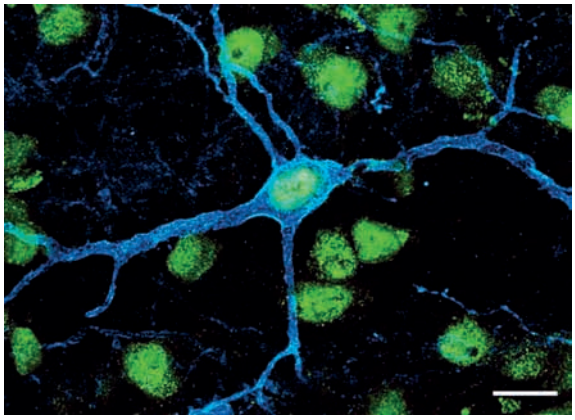

Figure 1 Selective expression of the neurokinin-1 receptor in the central nervous system. Confocal photomicrograph showing that only a small proportion of spinal cord neurons (green, neuronal nuclei-positive cells) express the neurokinin-1 receptor (blue). Scale bar = 25 μm.

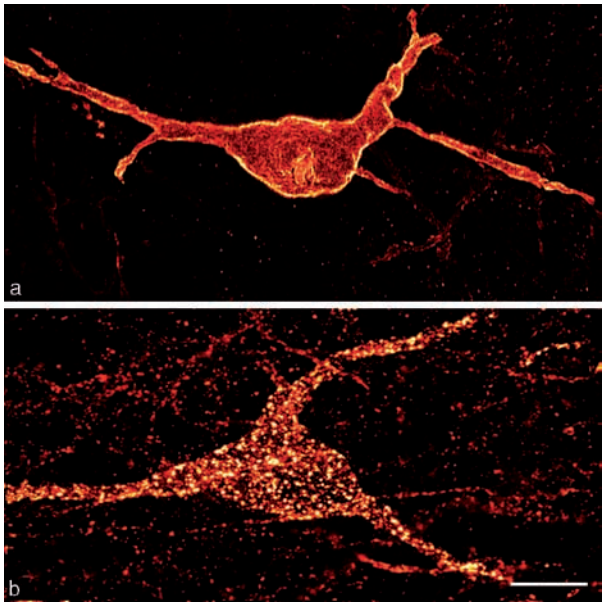

Figure 3 Internalization of neurokinin-1 receptors following noxious stimulation. Confocal photomicrographs showing the localization of neurokinin-1 receptors in different conditions. (a) On unstimulated neurons, neurokinin-1 receptors are uniformly distributed in the plasma membrane of the cell body and dendrites. (b) Five minutes after noxious somatosensory stimulation, which induces substance P release, neurokinin-1 receptors are internalized into the cytoplasm. Scale bar = 20 μm. Reprinted from Mantyh PW, DeMaster E, Malhotra A, et al. (1995) Receptor endocytosis and dendrite reshaping in spinal neurons after somatosensory stimulation. *Science* 268: 1629–1632.

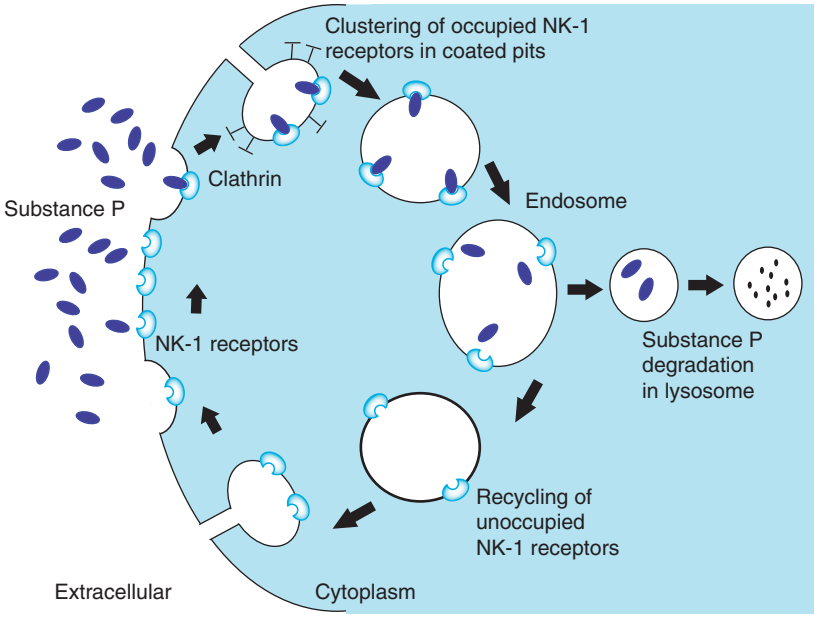

Figure 2 Model of substance P (SP)-induced endocytosis of neurokinin-1 (NK-1) receptor. NK-1 receptors are localized on the plasma membrane of both cell body and dendrites. Following noxious stimulation, substance P is released from presynaptic terminals and binds to NK-1 receptors present in the postsynaptic terminal. This is followed by a rapid internalization of SP and NK-1 into the cytoplasm in a clathrin-dependent manner via endosomes. Substance P is degraded in the lysosomes and NK-1 receptors are recycled back to the plasma membrane.

Substance P–NK-1 Internalization following Noxious Stimuli

Activation of substance P-containing neurons in response to acute noxious stimuli results in the rapid synaptic release of substance P, which then binds to the closely apposed NK-1 receptors in the immediate vicinity (Figure 2). On unstimulated neurons, NK-1 receptors are localized on the plasma membrane of both the cell body and dendrites (Figure 3(a)), but after stimulation and substance P binding, they are rapidly internalized (within 5 min) into the cytoplasm via endosomes (Figures 2 and 3(b)). This NK-1 receptor internalization in response to substance P binding is reversible, with complete return of internalized receptors to the surface after 30 min, although the kinetics of the internalization and recycling of NK-1 (Figure 2) and other GPCRs may depend on the specific phenotype of the cell expressing the receptor.

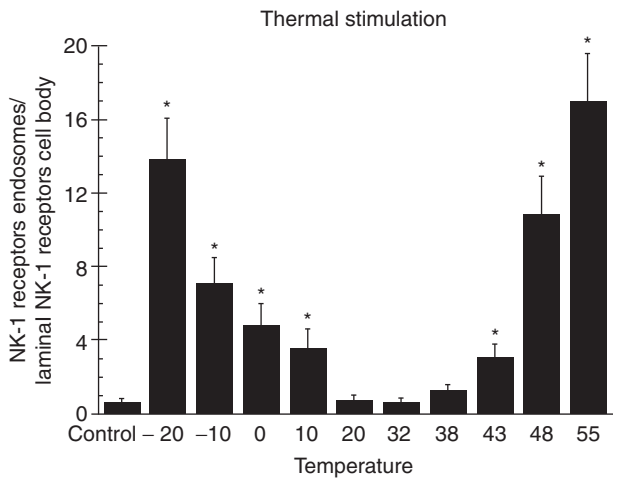

Figure 4 Internalization of substance P and neurokinin-1 (NK-1) receptors is proportional to the intensity of noxious stimulus. Adapted from Allen BJ, Rogers SD, Ghilardi JR, et al. (1997) Noxious cutaneous thermal stimuli induce a graded release of endogenous substance P in the spinal cord: Imaging peptide action *in vivo*. *Journal of Neuroscience* 17: 5921–5927.

At least in the spinal cord, the degree of substance P release and NK-1 receptor internalization is proportional to the intensity and frequency of noxious stimuli. More potent stimuli (e.g., either noxious cold or heat) result in a substantially greater release of substance P and internalization of the NK-1 receptor (Figure 4). Similarly, previous exposure to stress amplifies the activity of the SP–NK-1 receptor system, as documented in the rat model of irritant-induced hind paw inflammation, in which the initial decrease in NK-1 receptor activity is followed by a prolonged period ($\geq$8 days) of increased expression. Greater substance P release in response to higher intensity or repeated administration of stressful stimuli not only leads to more potent activation of adjacent NK-1 receptor-expressing neurons, but also permits diffusion of substance P away from the site of release and therefore stimulation of more distant neurons (Figure 5). It has been estimated that approximately three to five times more neurons are activated in response to more potent or more frequent stress stimuli (Figure 5).

Using Receptor Internalization to Deliver Therapies That Target Specific Classes of NK-1-Expressing Cells

A separate line of research has focused on identifying the functional consequences of inactivating the NK-1 receptor by either pharmacological (substance P receptor antagonists, cytotoxic substance P–saporin conjugates) or genetic (NK-1 receptor knockout

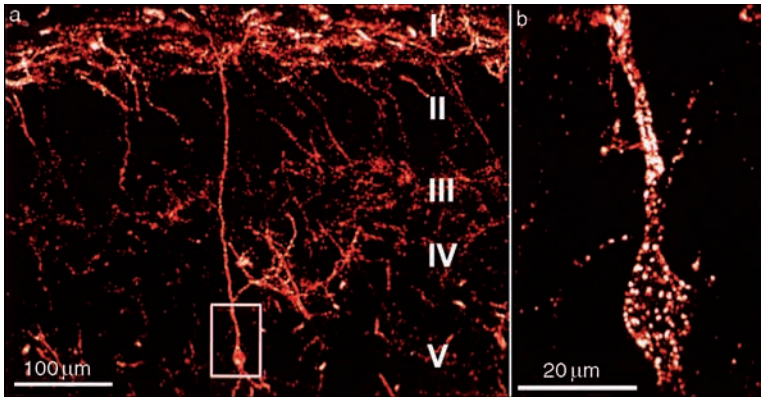

Figure 5 Substance P released from sensitized neurons diffuses and activates relatively distant neurokinin-1 receptor-expressing neurons. (a) Confocal photomicrograph showing the internalization of substance P–neurokinin-1 in neurons with cell bodies in lamina V of the spinal cord. (b) Higher power magnification. Adapted from Abbadie C, Trafton J, Liu H, Mantyh PW, and Basbaum AI (1997) Inflammation increases the distribution of dorsal horn neurons that internalize the neurokinin-1 receptor in response to noxious and non-noxious stimulation. *Journal of Neuroscience* 17: 8049–8060.

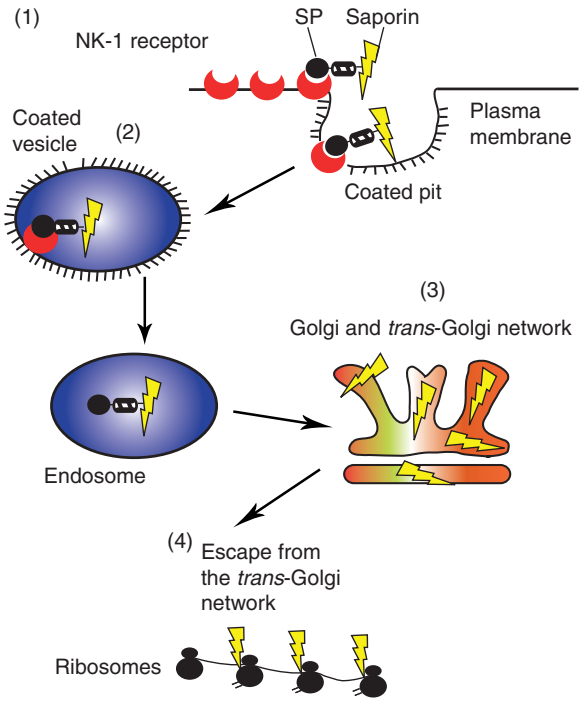

Figure 6 Mechanism of cytotoxic action of the substance P (SP)–saporin complex. The SP–saporin conjugate is internalized via the neurokinin-1 receptor endosomal pathway, and saporin subsequently escapes from *trans*-Golgi network to inhibit ribosomal protein synthesis and thereby induce cell death. Localized application of SP–saporin effectively leads to ablation of neurons containing the neurokinin-1 receptor.

mice) means. The substance P–saporin (SP–SAP) conjugate exploits the internalization of the NK-1 receptor to achieve selective killing of NK-1 receptor-expressing neurons by SAP, which inhibits ribosomal protein synthesis (Figure 6). Localized application of SP–SAP promotes selective pharmacologic ablation of neurons expressing the NK-1 receptor (Figure 7) and thereby allows the examination of the role of the SP–NK-1 receptor pathway in various regions of the brain and spinal cord. Injection of the SP–SAP complex into the dorsal horn of the spinal cord results in a marked attenuation of responses to highly noxious stimuli, mechanical and thermal hyperalgesia, and chronic neuropathy and inflammation, suggesting that the SP–NK-1 receptor pathway in the spinal cord may be involved in nociception.

Summary

The SP–NK-1 receptor pathway is involved in the neural processing of a range of noxious and stressful stimuli, and it represents the most abundant and the most thoroughly studied neuropeptide system in both the CNS and PNS. To ensure that extracellular stimuli are translated into intracellular signals of appropriate magnitude and specificity, most signaling cascades are tightly regulated. One of the major mechanisms involved in the regulation of GPCRs involves their

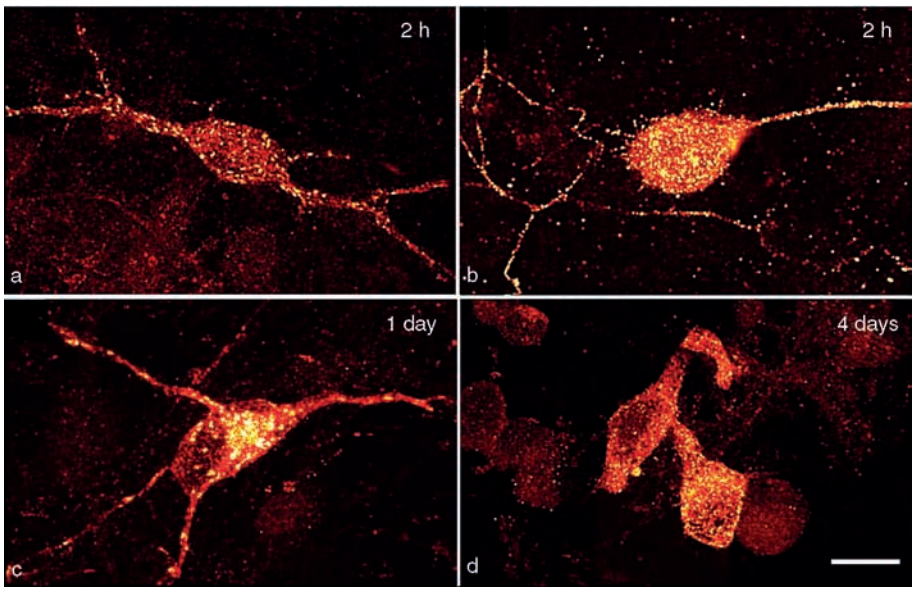

Figure 7 Internalization and cytotoxicity of substance P–saporin in primary cultures of neonatal spinal cord neurons. Confocal photomicrograph showing neurokinin-1 receptor immunofluorescence at different time points (a, c, d) following substance P–saporin treatment. Areas of concentrated neurokinin-1 receptor immunofluorescence appear yellow. Two hours after substance P–saporin, neurokinin-1 internalization was visualized with an antibody against neurokinin-1 (a), and intracellular accumulation of saporin was observed with an antibody against saporin (b). While neurokinin-1 receptor is recycled back to the plasma membrane following substance P alone (data not shown), neurokinin-1 receptor immunofluorescence is localized within intracellular endosomes 1 day after substance P–saporin (c). At day four posttreatment, neurokinin-1 receptor-positive neurons show shrunken cell bodies and diffuse neurokinin-1 immunoreactivity throughout the cytoplasm (d). Scale bar = 25 μm. Reprinted from Mantyh PW, Rogers SD, Honore P, et al. (1997) Inhibition of hyperalgesia by ablation of lamina I spinal neurons expressing the substance P receptor. *Science* 278: 275–279.

endocytic trafficking. GPCR endocytic trafficking entails the targeting of receptors to discrete endocytic sites at the plasma membrane, followed by receptor internalization and intracellular sorting. Thus, internalization of neuropeptide-receptors can regulate the level of cell surface receptors, the sorting of receptors to degradative or recycling pathways, and in some cases the specific signaling pathways.

See also: Hypocretin/Orexin and MCH and Receptors; Mammalian Neuropeptide Families; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neuropeptide Inactivation or Metabolism; Neuropeptide Receptors – Drug Development; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Discovery; Neuropeptides: Endocrine Cells; Neuropeptides: Mental Disease; Neuropeptides: Pain; Substance P/Tachykinins and its/their Receptors.

Further Reading

Abbadie C, Trafton J, Liu H, Mantyh PW, and Basbaum AI (1997) Inflammation increases the distribution of dorsal horn neurons that internalize the neurokinin-1 receptor in response to noxious and non-noxious stimulation. *Journal of Neuroscience* 17: 8049–8060.

Allen BJ, Rogers SD, Ghilardi JR, et al. (1997) Noxious cutaneous thermal stimuli induce a graded release of endogenous substance P in the spinal cord: Imaging peptide action *in vivo*. *Journal of Neuroscience* 17: 5921–5927.

Ferguson SG (2001) Evolving Concepts in G protein-coupled receptor endocytosis: The role in receptor desensitization and signaling. *Pharmacological Reviews* 53: 1–24.

Hofland LJ and Lamberts SW (2003) The pathophysiological consequences of somatostatin receptor internalization and resistance. *Endocrinology Reviews* 24: 28–47.

Hokfelt T, Bartfai T, and Bloom F (2003) Neuropeptides: Opportunities for drug discovery. *Lancet Neurology* 2: 463–472.

Hokfelt T, Pernow B, and Wahren J (2001) Substance P: A pioneer amongst neuropeptides. *Journal of Internal Medicine* 249: 27–40.

Mantyh PW, DeMaster E, Malhotra A, et al. (1995) Receptor endocytosis and dendrite reshaping in spinal neurons after somatosensory stimulation. *Science* 268: 1629–1632.

Mantyh PW, Rogers SD, Honore P, et al. (1997) Inhibition of hyperalgesia by ablation of lamina I spinal neurons expressing the substance P receptor. *Science* 278: 275–279.

Mazella J and Vincent JP (2006) Internalization and recycling properties of neurotensin receptors. *Peptides* 27: 2488–2492.

Neuropeptides Phylogeny and Evolution

D Larhammar, Uppsala University, Uppsala, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuropeptides comprise a large and diverse set of intercellular signaling molecules in the nervous system of most, if not all, animal phyla. Many of the neuropeptides also function as hormones or have evolutionarily related peptides that do so. Due to the small size of many neuropeptides, often below 15 amino acids with the shortest consisting of only three amino acids, their phylogenetic relationships have been difficult to study across distantly related animal groups.

Neuropeptides are generated by proteolytic cleavage of larger precursor peptides (prepro-peptides) that consist of a signal peptide and one or several spacer peptides, in addition to the biologically active mature neuropeptides. These spacer peptides usually diverge much more rapidly than the neuropeptides, thereby obscuring evolutionary relationships. Several neuropeptides show similarities between vertebrates and insects or mollusks (Table 1), thus demonstrating their origin prior to the divergence of deuterostomes and protostomes. In addition to sequence comparisons, some genetic criteria are considered here that have proven useful in determining gene orthology, most significantly conserved synteny (i.e., the same location of the gene, relative to other genes, across species).

Vertebrate genomes may encode well over 100 distinct neuropeptides and possibly considerably more. Analyses of the presently available sequenced genomes are probably far from exhaustive because of the difficulties in identifying short peptides and their precursors. Several additional genes encoding neuropeptides or endocrine peptides are expected to be discovered in these genomes because many orphan receptors exist that are likely to have peptide ligands. The genome of the fruit fly *Drosophila melanogaster* has been reported to contain about 35 genes encoding neuropeptides, and the nematode worm *Caenorhabditis elegans* has at least 109 genes encoding a total of more than 250 potential neuropeptides.

Many neuropeptides make up families of related members (see Table 1) that have arisen by tandem duplications within a single gene or by duplications of entire genes, either locally within a chromosomal region, by retrotranscription and reinsertion into a chromosome, or by duplication of a chromosomal segment. Neuropeptides usually act through G-protein-coupled receptors (GPCRs), and the diversity and relationships of neuropeptides are reflected in the multiplicity and relationships of their receptors. A few examples are known in which neuropeptides and their corresponding receptors have coevolved structurally and functionally. This article describes in general terms the evolution and phylogeny of neuropeptides, with several examples taken from well-studied families in vertebrates and invertebrates.

Duplications within Genes

Some neuropeptides are generated from simple prepro-peptides that contain only a single mature peptide, for instance neuropeptide Y (NPY) and galanin, whereas others may harbor a few or several biologically active peptides, such as the famous precursor proopiomelanocortin (POMC), which encodes corticotropin (ACTH) and β-endorphin as well as a few related copies of melanocyte-stimulating hormone (MSH). Additional copies of a peptide within a single precursor may arise by duplication of the DNA segment that encodes the mature neuropeptide.

One common type of duplication is tandem duplication within the same exon of a gene, as exemplified by thyrotropin-releasing hormone (TRH), the first neuropeptide to be purified and sequenced. TRH is present in vertebrates and has the sequence (pyro) Glu-His-Pro-amide, generated from the sequence Gln-His-Pro-Gly flanked by proteolytic cleavage sites in the precursor. TRH is present in six copies within the same exon in the human prepro-TRH gene (Figure 1(a)), seven copies in the frog *Xenopus tropicalis,* and eight copies in goldfish. Chicken prepro-TRH contains four copies flanked by canonical cleavage sites and includes three more TRH-like motifs that may no longer be excised. Several other prepro-peptides contain multiple copies within the same exon, for instance, six identical copies of Met-enkepahlin (Tyr-Gly-Gly-Phe-Met) and one copy of Leu-enkephalin in prepro-enkephalin (Leu-enkephalin differs only in the last amino acid from Met-enkephalin), and two similar copies of orexin in prepro-orexin orexin A and B (also called hypocretin-1 and -2), share approximately 50% of their amino acid positions.

The most impressive examples of tandem duplications in the same exon of a prepro-peptide gene are probably the precursors of FMRF-amide-like peptides in invertebrates. *D. melanogaster* has three precursors for FMRF-amide-like peptides, one of which contains information for 13 such peptides encoded by a single exon. Even more extreme is the FMRF-amide

Table 1 Selected families of vertebrate/chordate neuropeptides and endocrine peptides[a]

Neuropeptide family	*Proposed protostome homologs*
Neuropeptide Y (NPY) Peptide YY (PYY) Pancreatic polypeptide (PP)	Neuropeptide F (NPF)
Cholecystokinin (CCK) Gastrin (cionin in the tunicate *Ciona intestinalis*)	Sulfakinins
Oxytocin Vasopressin (isotocin, mesotocin, vasotocin in nonmammalian vertebrates)	Cephalotocin, octopressin, Lys-conopressin
Insulin Insulin-like peptides Insulin-like growth factors I and II Relaxins	Insulin-like
Somatostatin Cortistatin (somatostatin 2) Urotensin II	
Urotensin II-related peptide	
Tachykinins: Substance P Neurokinins A and B Hemokinin-1 (physalaemin, kassinin in frogs)	Tachykinin-related
Opioid peptides: Met-enkephalin Leu-enkephalin dynorphins A and B α-neoendorphin β-endorphin orphanin FQ (= nociceptin)	
Galanin Galanin-like peptide	
Glucagon Glucagon-like peptides 1 and 2 (GLP-1, GLP-2) Vasoactive intestinal peptide (VIP) Peptide histidine-methionine (PHM) Secretin Gastric inhibitory peptide (GIP) Growth hormone-releasing hormone (GHRH = GRF) Pituitary adenylate cyclase-activating peptide (PACAP) PACAP-related peptide (PRP)	
Melanocyte-stimulating hormones (MSH): $\alpha, \beta, \gamma, \delta$	
Corticotropin-releasing hormone (CRH) Urocortins II and III	Diuretic hormone (DH44)
Gonadotropin-releasing hormones (GnRH = LHRH)	
Neuropeptide FF Neuropeptide AF	
Orexins A and B	
Prolactin-releasing peptide (PrRP)	
Calcitonin Calcitonin gene-related peptide (CGRP) Amylin Adrenomedullins Intermedin Calcitonin receptor-stimulating peptides (CRSPs)	Diuretic hormone 31 (DH31)
Neuromedin U	Hugin
Neuromedin B (NM-B) Gastrin-releasing peptide (GRP) (bombesin in frogs)	
Galanin Galanin-like peptide (GALP)	

[a]Some family members function as neuropeptides, others function as endocrine peptides, and some serve dual roles.

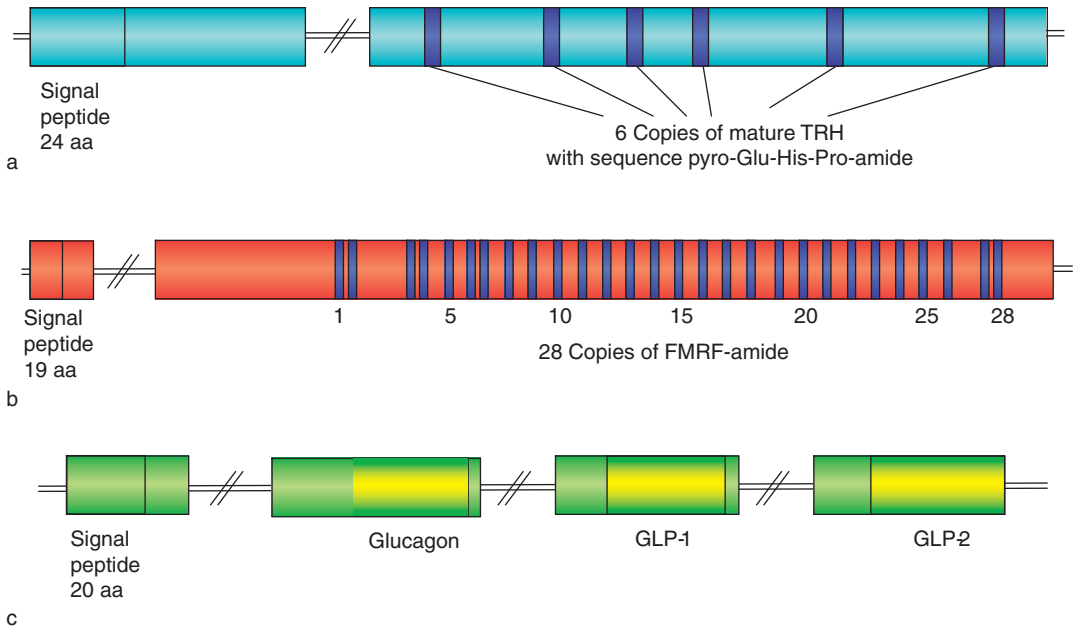

Figure 1 Schematic outline of the genes encoding three peptide precursors: (a) human prepro-TRH (242 amino acids); (b) *Aplysia californica* FMRF-amide precursor (597 amino acids); (c) human precursor for glucagon, GLp-1, and GLP-2 (179 amino acids). The precursors for TRH (a) and FMRF-amide (b) contain all copies of the mature peptide within the same exon, whereas the precursor for the related peptides glucagon, GLP-1, and GLP-2 (c) has the three peptides on separate exons as a result of exon duplications. Note that the FMRF-amide precursor in (b) has a different scale. The intron sizes are not drawn to scale, as indicated by the parallel diagonal lines. aa, amino acid(s); GLP, glucogon-like peptide; TRH, thyrotropin-releasing hormone.

precursor from the marine snail *Aplysia californica*, with no less than 28 copies of the same FMRF-amide motif (Figure 1(b)). The nematode *C. elegans* has 29 genes encoding more than 70 FMRF-amide-like peptides.

On numerous occasions, multiplication of a peptide- encoding segment has occurred by duplication of an entire exon that contains the coding information for a mature neuropeptide. This difference in duplication mechanism relative to the within-exon duplications already described is worth noting because exon duplication need not occur with the same precision as a duplication within a coding segment that must maintain the translational reading frame; this is because the splicing machinery connects duplicated exons to generate a continuous coding sequence as long as their splice junctions are located in the same phase of the codons. One example of a gene with exon duplication is the gene for the glucagon precursor, which also encodes, on separate exons, the glucogon-like peptide (GLP)-1 and GLP-2 (Figure 1(c)). Likewise, the precursor for tachykinin-1, encoding substance P and neurokinin A, and the precursor for vasoactive intestinal peptide (VIP), encoding VIP and peptide histidine-methionine (PHM), have expanded by exon duplication to encode additional structurally related peptides.

Over time, the duplicated peptides may accumulate amino acid differences, leading to more or less different functional roles. Examples of this are the peptides GLP-1 and GLP-2, which have found separate receptors from that of glucagon (due to duplications of the glucagon receptor gene), and prepro-tachykinin-1, whose peptides substance P and neurokinin A also-have separate receptors due to a receptor gene duplication.

A more complicated result of an exon duplication is displayed by the gene encoding calcitonin and calcitonin gene-related peptide (CGRP). These two structurally related peptides arose by an exon duplication, and the primary gene transcript is alternatively spliced to use either of these exons. Whereas endocrine C cells in the thyroid gland splice the pre-mRNA to include the calcitonin exon, neurons generate a splice variant that contains the CGRP exon and an additional exon (Figure 2).

Gene Duplications

A common form of duplication involves entire genes. Such duplicates are often located adjacent to one another in the same chromosomal region, either in tandem (i.e., in the same orientation) or with opposite orientation relative to one another. One example of tandem duplication of a prepro-peptide gene is the duplication of the peptide YY gene (expressed in endocrine cells and a few neurons in mammals) to generate the pancreatic polypeptide (PP) gene

(see Figure 3), a duplication that seems to have taken place early in tetrapod evolution. The genes are located only 10 kb apart in the human genome. Another example is the duplication that generated the oxytocin and vasopressin precursor genes from a common ancestral gene, probably at an early stage in vertebrate evolution. These two peptide genes are also located close to one another and seem to differ in their orientation relative to one another due to gene inversion in some lineages (the frog *Xenopus laevis* has the genes in tandem; mammals have them tail-to-tail). In these two examples, the exon–intron organization is the same for the two copies, indicating that the genes were duplicated at the DNA level, possibly by unequal crossing-over due to repeated DNA elements flanking the ancestral gene.

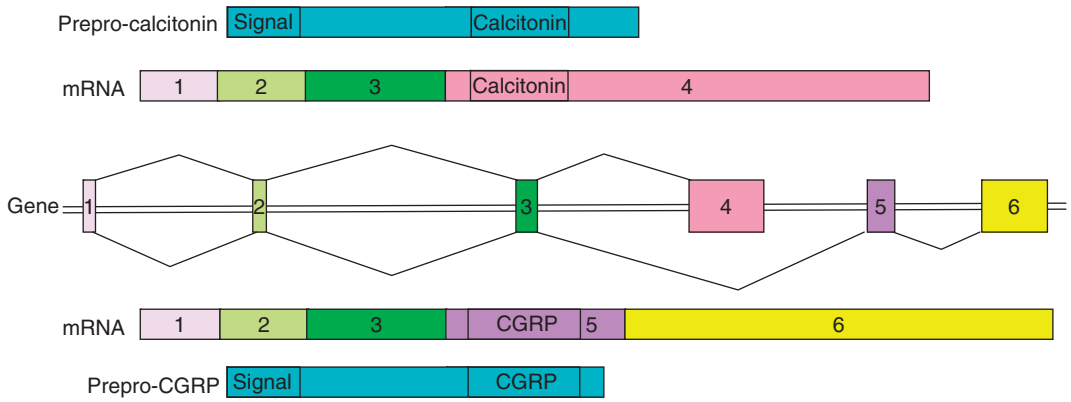

Figure 2 Human gene for the calcitonin–CGRP precursor. Alternative splicing of the Pre-mRNA leads to the production of either calcitonin (endocrine cells in the thyroid) or CGRP (neurons). CGRP, calcitonin gene-related peptide.

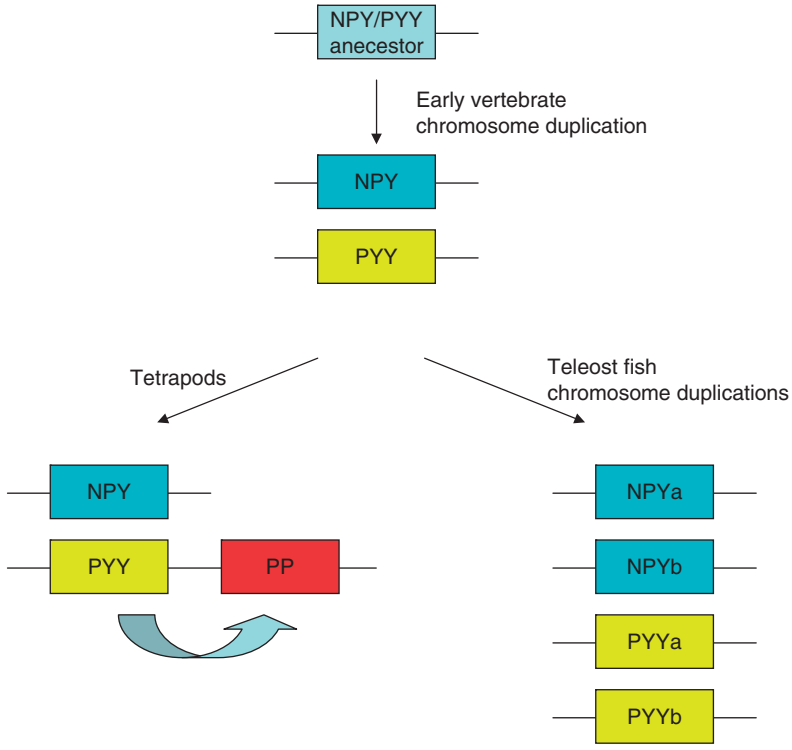

Figure 3 Gene duplication scheme for the vertebrate NPY family. The gene duplication that gave rise NPY and PYY from a common ancestral gene occurred early in vertebrate evolution through the duplication of a chromosomal region or an entire chromosome (probably in a tetraploidization event), which preceded the origin of jawed vertebrates (gnathostomes). Subsequently, the PYY gene was duplicated locally to generate PP, probably in an early tetrapod ancestor because PP is present in amphibians, reptiles/birds, and mammals. After these gene duplications, PP accumulated considerably more amino acid changes than NPY and PYY. NPY, neuropeptide Y; PP, pancreatic polypeptide; PYY, peptide YY.

In other instances, the introns are missing from one of the copies, suggesting that duplication occurred via mRNA that was retrotranscribed to DNA and then reinserted into the genome. This seems to be the case for the second insulin gene found in the rat. This duplicate retains one of the two introns present in the original gene, suggesting that the DNA copy was retrotranscribed from a partially spliced mRNA intermediate. Retrotranscription also may explain the origin of the gonadotropin-releasing hormone (GnRH)-1 gene in the tunicate *Ciona intestinalis,* which lacks the two introns present in the GnRH-2 gene.

Proposed Vertebrate Chromosome or Genome Duplications

Many families of neuropeptide genes in vertebrate genomes have their family members distributed on different chromosomes. It has been proposed that this is due to duplications of large chromsomal regions or entire chromosomes, or even of the entire genome (tetraploidization). This has most convincingly been demonstrated for the fish lineage, which underwent a tetraploidization before the radiation of teleost fishes (which make up more than 99% of the extant fish species). At least one and possibly two such tetraploidizations also seem to have occurred before the divergence of gnathostomes (jawed vertebrates). More recently, additional tetraploidizations have occurred independently in several vertebrate groups, for instance *xenopus* frogs, salmonid fishes, cyprinid fishes, and sturgeons. The tetraploidizations have generated additional gene copies for a huge number of gene families, including several families of neuropeptides, endocrine peptides, and their receptors. This mechanism of duplication means that new peptides and receptors arose concomitantly for several peptide–receptor systems, thereby probably facilitating novel peptide–receptor partnerships.

One well-studied example of family expansion by chromosome duplication is the NPY family, in which NPY and peptide YY (PYY) arose from a common ancestral gene (Figure 3). Subsequently, PYY was duplicated locally to generate PP on the same chromosome, as already mentioned. The teleost fish tetraploidization generated duplicates of both NPY and PYY. All four peptide genes are present in the genome of the pufferfish fugu (*Takifugu rubripes*), but one of the NPY copies seems to have been lost from the genome of the zebra fish (*Danio rerio*). Assignment of the fish peptides as either NPY or PYY has been difficult to achieve by sequence analyses only, but has been possible by comparison of the chromosomal location of the genes. Thus, gene neighborhood (conserved synteny or location in the same chromosomal region) is a useful criterion for the identification of gene orthology (i.e., species homologs).

Also, the receptor family for NPY has expanded by chromosome duplications, although here a local gene triplication preceded the chromosome duplications. Another peptide family that has grown by this mechanism is the tachykinin family, whose three prepro-peptide genes are located in three related chromosome regions. Likewise, the receptors for the tachykinins have also received additional copies by chromosome doubling.

Another peptide family that seems to involve both a local duplication and a chromosome-segment duplication is the family consisting of somatostatin 1, cortistatin (= somatostatin 2), urotensin II, and urotensin II-related peptide. A combination of local and chromosomal duplications also explains the complex insulin/relaxin family, which has 10 members in humans, some of which are expressed in the nervous system. Detailed gene-linkage information for distantly related vertebrates is required to determine whether additional peptide families have expanded by chromosome duplication.

Phylogenetic Sequence Analyses

The primary criterion for assignment of different peptides to a family is, of course, that they display some degree of amino acid sequence identity (shared positions). If the degree of identity is significantly higher than can be expected from chance events, it is presumed to be due to common ancestry. However, the shortness of neuropeptides and endocrine peptides often make detailed sequence-based phylogenetic analyses ambiguous. Furthermore, some positions in a neuropeptide may be crucial for function and may, therefore, have been maintained constant by conservative selection pressures over long evolutionary periods (e.g., cysteine residues involved in disulfide bonds), whereas other positions are less important and therefore tolerate evolutionary changes to a considerable extent. For these reasons, it has been pointed out that short peptides are risky to use for phylogenetic studies. However, to understand the evolution of the peptides themselves, sequence comparisons are, of course, of crucial importance.

Even more problematic for phylogenetic analyses are positions that may switch back and forth between a few amino acid alternatives, often involving residues with similar properties such as positively charged Lys-Arg or negatively charged Asp-Glu. Spacer peptides may evolve so rapidly that they sometimes cannot even be aligned when comparing animal groups as distantly related as tetrapods and bony fishes or insects and crustaceans, not to speak of deuterostomes

versus protostomes. Nevertheless, some neuropeptides show similarities between, for instance, vertebrates and insects, thus demonstrating their origin prior to the divergence of protostomes and deuterostomes. Such examples seem to include the vertebrate peptide families for NPY, cholecystokinin (CCK)/gastrin, oxytocin/vasopressin, tachykinins, insulin, corticotropin-releasing hormone (CRH), calcitonin, and neuromedin U (see Table 1).

Among vertebrates, sequence comparisons usually allow the unambiguous assignment of a peptide to a family, although detailed relationships within a peptide family may be more complicated to resolve across species. A particularly problematic example is the family of peptides comprising oxytocin and vasopressin in mammals, which usually differ in only two of the nine positions. Similar peptides have been identified in all gnathostome groups investigated and have been given names such as isotocin, mesotocin, and vasotocin, along with several additional variants identified in cartilaginous fishes (sharks, rays, skates, and chimaeras). It is not obvious if these are orthologs (species homologs) to mammalian oxytocin or vasopressin. In situations such as these, additional comparative criteria may be helpful, such as chromosomal location and conserved synteny.

Some peptides have dramatically uneven rates of evolutionary change across animal phyla. Also, rates of change may differ considerably among peptide members of the same peptide family. Some of the most intriguing examples are provided by peptide/protein hormones because information is available for a very large number of species, but sequence information is emerging also for several neuropeptides. For instance, the protein hormones prolactin and growth hormone (sizes just under 200 amino acids) differ greatly in their evolutionary rate across orders of mammals. Both of these hormones have evolved extremely rapidly in primates. Growth hormone has remained almost constant in, for instance, the pig, dog, horse, and elephant since their respective lineages diverged from the common mammalian ancestor shared with primates. In contrast, prolactin has changed extensively in the elephant. It is not yet obvious how the rates of change correlate with any physiological functions. Another peptide hormone with a rapid rate of change is PP. Interestingly, the replacement rate of PP seems to have been highest shortly after its origin (i.e., its duplication from PYY; Figure 3), after which it slowed down in the different tetrapod lineages.

Short peptides tend to be highly conserved, probably because each position has a relatively greater impact on the peptide's structure and interactions with receptors. Strongly conserved short peptides include, for instance, GnRH, somatostatin, oxytocin, and vasopressin. Among the longer peptides, some of the most highly conserved during vertebrate evolution are NPY (36 amino acids) and CRH (41 amino acids). But even for such long and well-conserved neuropeptides, the phylogenetic trees diverge from the established species trees due to uneven rates, back mutations, and extremely strong conservative selection pressure.

Different parts of the same peptide may also display considerable differences in the rate of change. One such peptide is galanin (29 amino acids), whose N-terminal 14 residues are well conserved, whereas the C-terminus has accumulated several changes. These internal-rate differences seem to correlate well with the conservative selection pressure imposed by receptor interactions, such that the conserved part of the peptide is most important. For the NPY family peptides, the C-terminal part is much more strongly conserved than the N-terminal part.

One particularly challenging neuropeptide family from the evolutionary perspective is the GnRH family, which consists of two to three highly similar copies in most vertebrates. These decapeptides share four identical positions and have another three highly conserved positions across all vertebrates investigated as well as tunicates. The remaining three positions, on the other hand, are quite variable. This has made it difficult to determine the orthology relationships across species. The terminology is further complicated by the fact that a particular GnRH sequence is often named after the species in which it was first discovered. Thus, chicken GnRH-II is used to refer to a particular GnRH sequence that can be found in many different species.

Naturally, the accuracy of the phylogenetic trees will improve as the list of species for which neuropeptide sequences are known grows, particularly when information becomes available for species that represent distantly related and crucial evolutionary branches. The genome projects offer the great advantage of being able to provide sequences for most or all of the members of each neuropeptide and receptor family within a species, thereby facilitating a more accurate assignment of orthologs across species.

Evolutionary Change of Neuropeptide Functions

The functions of a particular neuropeptide may shift during evolution, but the definitive assignment of ancestral versus derived functions is risky until detailed information has been collected for several species from different evolutionary lineages. Neuropeptides that have clearly distinct roles in different animal groups include some of the peptides that exist both in protostomes and deuterostomes.

Functional differences have been observed also among more closely related lineages, although some of these differences are still in the process of being explored. For instance, the role of prolactin-releasing hormone is thought to differ between mammals and birds (chicken), as is of course the role of the protein hormone prolactin itself. It may also be that the role of oxytocin for olfactory reproductive memory in rodents is not as prominent in humans.

Likewise, changes in function after a gene duplication are difficult to determine and require functional data for species that diverged prior to the gene duplication as well as for species that have the duplication. It is expected that neuropeptide duplicates may undergo either of the same fates that have been reported for other gene families: a new gene copy may find new functions (neo-functionalization), it may be severely mutated and be lost (nonfunctionalization), or it may split the mother gene's functions with the other daughter gene (subfunctionalization).

Numerous examples exist in which structurally related neuropeptides have quite different functional roles in some respects, yet partially overlap to some extent. In mammals, oxytocin and vasopressin both seem to be involved in neuronal aspects of reproductive and social behavior, but have more distinct roles in mammary and renal endocrinology. Even duplicated peptides generated from the same precursor may evolve different functional roles, as shown by substance P and neurokinin A from the prepro-tachykinin-1 precursor.

In mammals, the two related neuropeptides pituitary adenylate cyclase-activating polypeptide (PACAP) and VIP influence the circadian pacemaker system in the suprachiasmatic nucleus (SCN) of the hypothalamus. Whereas PACAP acts on two receptor subtypes, PAC1 and VPAC2, VIP acts only on the latter. The difference in receptor preference along with differences in anatomical expression of these components within the SCN explains how the two peptides have evolved distinct roles in the regulation of the mammalian circadian clock. Whereas VPAC2 (responding to both neuropeptides) is involved in both maintenance and synchronization of the clock, PAC1 seems to be involved in light-induced phase shift, as suggested by experiments with knockout mice.

Another example of rather distinct functions for gene duplicates, not least with regard to sites of production, is the NPY family, which in mammals involves NPY produced in neurons and PYY made in intestinal endocrine cells. PYY is also produced in some brain stem neurons in the rat and more extensively in the nervous system of the zebra fish and lamprey, in addition to endocrine cells. Possibly, the ancestral NPY/PYY gene was expressed in both neurons and endocrine cells.

An intriguing functional difference between NPY and PYY is that NPY is released before meals in the hypothalamus and stimulates appetite, whereas PYY is released postprandially (after meals) from the intestine and reduces appetite via the hypothalamus. Some evidence suggests that the function of the ancestral peptide was appetite stimulation and that the appetite inhibition induced by PYY is a result of receptor Y2 expression in a different cell type than those that stimulate appetite (subtypes Y1 and Y5). The time for this functional switch in evolution is not yet known.

One type of evolutionary change concerns how a peptide and a receptor subtype may evolve preference for one another leading to the (partial) exclusion of other related peptides and receptors. This can also be exemplified by the NPY family, in which PP and receptor Y4 have stronger preference for one another in mammals, particularly in rodents, than they have in the chicken. This preference is a derived feature because PP is a copy of PYY and therefore must once have bound to the same receptors as PYY. One function of PP in humans seems to be the inhibition of appetite after meals (like PYY), albeit the two peptides exert this effect via separate receptor subtypes.

Additional mechanisms that may lead to altered peptide structures during evolution are posttranslational modifications and changes in peptide processing and degradation. Although detailed functional data are more scarce for these types of evolutionary changes (because they usually cannot be predicted from gene sequences or predicted protein sequences), some observations suggest that, for instance, somatostatin, peptide YY, and bradykinin may be processed differently among classes of vertebrates. Sequence differences in prepro-peptides suggest that precursor processing may differ among species for several additional peptides.

Summary and Conclusion

The evolutionary history of neuropeptides and related endocrine peptides is replete with gene duplications, involving parts of genes or entire genes by a variety of mechanisms, that have resulted in a plethora of peptides. The large number of neuropeptides and their corresponding GPCRs influence a broad range of functions and have presumably contributed greatly to the diversification and refined regulation of neuronal and endocrine processes such as reproduction, food intake, and circadian rhythm, as well as many other physiological processes and behaviors. The duplications have paved the way for more specific functions for some neuropeptides and novel functions for

others. The phylogenetic relationships of neuropeptides have in some cases been difficult to resolve due to their small size and uneven rates of change. Novel types of information from the many genome projects such as gene location (gene neighborhood), along with information from additional species representing important evolutionary branches, will make it possible to sort out neuropeptide and receptor phylogenies with much greater precision.

See also: Invertebrate Neurohormone GPCRs; Mammalian Neuropeptide Families; Neuropeptide Signaling in Invertebrates; Neuropeptides: Discovery; Opioid Peptides and Receptors; Somatostatin and Receptors; Substance P/Tachykinins and its/their Receptors; Vasopressin/Oxytocin and Receptors.

Further Reading

Adams BA, Tello JA, Erchegyi J, et al. (2003) Six novel gonadotropin-releasing hormones are encoded as triplets on each of the two genes in the protochordate, *Ciona intestinalis*. *Endocrinology* 144(5): 1907–1919.

onlon JM and Larhammar D (2005) The evolution of neuroendocrine peptides. *General Comparative Endocrinology* 142: 53–59.

Dores RM, Rubin DA, and Quinn TW (1996) Is it possible to construct phylogenetic trees using polypeptide hormone sequences? *General Comparative Endocrinology* 103: 1–12.

Larhammar D, Fredriksson R, Larson ET, and Salaneck EP (2004) Phylogeny of NPY-family peptides and their receptors. In: Michel MC (ed.) *Handbook of Experimental Pharmacology, Vol. 162: Neuropeptide Y and Related Peptides*, pp. 75–100. Berlin-Heidelberg: Springer.

Lethimonier C, Madigou T, Munoz-Cueeto JA, Lareyre JJ, and Kah O (2004) Evolutionary aspects of GnRHs, GnRH neuronal systems and GnRH receptors in teleost fish. *General Comparative Endocrinology* 135: 1–16.

Li C (2005) The ever-expanding neuropeptide gene families in the nematode. *Caenorhabditis elegans*. *Parasitology* 131(supplement): S109–S127.

Maywood ES, Reddy AB, Wong GK, et al. (2006) Synchronization and maintenance of timekeeping in suprachiasmatic circadian clock cells by neuropeptidergic signaling. *Current Biology* 21: 599–605.

Olinski RP, Lundin LG, and Hallbook F (2006) Conserved synteny between the Ciona genome and human paralogons identifies large duplication events in the molecular evolution of the insulin-relaxin gene family. *Molecular Biology and Evolution* 23: 10–22.

Sherwood NM, Krueckl SL, and McRory JE (2000) The origin and function of the pituitary adenylate cyclase-activating polypeptide (PACAP)/glucagon superfamily. *Endocrine Reviews* 21: 619–670.

Strand F (1999) *Neuropeptides – Regulators of Physiological Processes*. Cambridge, MA: MIT Press.

Taussig R and Scheller R (1986) The Aplysia FMRFamide gene encodes sequences related to mammalian brain peptides. *DNA* 5: 453–461.

Tostivint H, Joly L, Lihrmann I, et al. (2006) Comparative genomics provides evidence for close evolutionary relationships between the urotensin II and somatostatin gene families. *Proceedings of the National Academy of Sciences of the United States of America* 103: 2237–2242.

Wallace M (2001) Episodic evolution of protein hormones in mammals. *Journal of Molecular Evolution* 53: 10–18.

Neuropeptides: Discovery

O Civelli, University of California at Irvine, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Chemically, neuropeptides are molecules composed of short stretches of covalently linked amino acids. Neuropeptides are messenger molecules that transfer information from one cell to another in the central nervous system (CNS). They accomplish this by being released from a cell as 'first messengers' and by interacting with specific receptors found on the membrane of a recipient cell. This induces a second messenger response that will change the chemistry of the recipient cells and that, after a cascade of events, results in a physiological or behavioral change. All neuropeptide receptors are G-protein-coupled receptors (GPCRs). Consequently, the expression of a particular peptide GPCR in the CNS indicates that this peptide should be classified as a neuropeptide.

Neuropeptides are numerous and, in general, structurally diverse; they form the largest family of first messenger molecules in mammals. They are small in size when compared to proteins, ranging from 3 to less than 100 amino acids. They are all synthesized in the form of large, inactive precursor polypeptides, which need to be cleaved by processing enzymes to yield the final peptides. These are often enzymatically modified to be rendered bioactive.

These characteristics, which today specify the nature of neuropeptides, were not found easily. They have resulted from years of work searching for different neuropeptides, their modes of synthesis, and their modes of action. It all started with the discovery of neuropeptides.

The Discovery of Neuropeptides

At first neuropeptides were discovered via their functions. A particular physiological response would be observed to be regulated by some tissue extract. Upon further analysis, which often took years, it would be found that a peptide was responsible for this regulation. Later, when more had been learned about them, neuropeptides were also discovered on the basis of similarities expected to exist in their structures. Recently, this line of research has involved genomic analyses. Finally, neuropeptides have been discovered as the natural ligands of orphan GPCRs (see Table 1).

The Organ Bath-Based Approaches

One approach relies on the use of the 'organ bath' introduced by Rudolph Magnus in Utrecht at the beginning of the twentieth century. A freshly harvested organ (e.g., intestine) is bathed *ex vivo* in a physiological salt solution linked to a force transducer that records changes in muscle activity. Using such a bath, it is possible to monitor smooth muscle contractions and vasodilator properties. This approach led to the discovery of many neurotransmitters, in particular, the biogenic amines and acetylcholine. It was also used to test synthetic drug effects, and that led to the idea that it could be used for testing extracts of natural origin. This approach opened the door to the discovery of the first neuropeptides.

The first neuropeptide and the enkephalins In the 1930s, von Euler and Gaddum characterized 'preparation P' from extracts of horse brain and intestine that stimulated the contraction of intestinal tissues *in vitro*. This activity was not inhibited by atropine, evidence that the responsible agent was not acetylcholine. One notable aspect of this activity was revealed when it was shown to be protease sensitive. This was the first evidence for the existence of peptides active in the CNS and gave birth to the concept of neuropeptides. This active peptide was later named substance P; its structure was finally resolved in 1971.

But the discovery that, possibly, more than any other established the role of neuropeptides in neurotransmission was that of the enkephalins. They were discovered, in brain extracts, as agents able to mimic the action of morphine in a bath assay. Because morphine is the standard analgesic and an addictive molecule, these peptides created an immediate interest in research fields dealing with pain management or drug abuse. Their names (literally, endogenous brain peptides) indeed helped prove the validity of the neuropeptide concept.

Two families of hormones that became neuropeptides At the beginning of the 1950s, du Vigneaud et al. isolated, from posterior pituitary lobes, oxytocin and vasopressin on the basis of their uterus-contracting and pressor activities, respectively. These peptides act primarily on peripheral tissues, namely, the uterus for oxytocin and the kidney for vasopressin. We know now that their receptors are expressed in the CNS and that some of their activities are brain related. They are therefore neuropeptides.

Table 1 Dates in the discovery of neuropeptides

Date	*Discovery*
1902	Secretin, discovered as a hormone, now classified as a neuropeptide
1930	Substance P, the first neuropeptide, sequenced in 1971
1950–51	Oxytocin and vasopressin, found as hormones, now classified as neuropeptides
1969	TRH and the other hypothalamic releasing factors
1975	The enkephalins, proof of concept for the neuropeptides
1960s–1990s	The major schools of neuropeptide discovery:
	Mutt and Tatemoto: Discovery of C-amidated neuropeptides
	Erspamer and Melchiorri: Discovery of neuropeptides in amphibian skin
	Matsuo, Minamino, and Kangawa: Discovery of neuropeptides involved in muscle contraction or relaxation
1979	γ-MSH, the first neuropeptide found not on the basis of its activity but by sequence analysis
1983	CGRP and neurokinins, found as products of alternative splicing
1995	OFQ/N, first neuropeptide found as ligand of an orphan GPCR
1995–97	Urocortin and cortistatin neuropeptides found by homology screening and RNA subtraction analyses
1995–2000s	The orphan receptor strategy leads to the discovery of 12 neuropeptides
	A major school: Takeda chemical industry
2000	hRFRP-1–3, the first neuropeptides found by bioinformatics

TRH, thyrotropin-releasing hormone; MSH, melanocyte-stimulating hormone; CGRP, calcitonin gene-related peptide; OFQ/N: orphanin FQ/nociceptin; hRFRP, RFamide-related peptides.

The releasing factors are also among the first neuropeptide discoveries. Work by Harris had specified that there exist hypothalamic substances that regulate pituitary function. The concept was validated with the isolation and characterization, from tens of thousands of hypothalami, of one of these messenger substances, thyrotropin-releasing factor (TRH). This was followed by the identification of releasing factors for each of the pituitary glycoprotein hormones. Although isolated as hormones, these releasing factors can be now classified as neuropeptides since their receptors have also been shown to be expressed in the CNS.

Three schools of neuropeptide discovery Some biological activities related to intestinal motility and pancreatic secretion had been detected since the beginning of the twentieth century, and some of them had been linked to the action of 'hormones,' which proved to be peptides. Indeed, the term hormone was invented as a consequence of the discovery of such a bioactive peptide, secretin.

From the 1950s to the 1980s, Victor Mutt and coworkers determined the sequence of several of these bioactive peptides, in particular, secretin, cholecystokinin, and vasoactive intestinal peptide. They also realized that C-terminal amidation is a hallmark of many bioactive peptides and used this chemical principle to isolate randomly potentially active peptides from large quantities of intestines or brains. This search ultimately led to the discoveries of novel bioactive peptides such as peptide histidine isoleucine (PHI), peptide tyrosine tyrosine (PYY), neuropeptide tyrosine (NPY), and galanin. Again, we can classify these peptides as neuropeptides since their receptors are expressed in the CNS.

Another school of neuropeptide discovery, under the direction of Erspamer and Melchiorri, took advantage of the fact that amphibian skin proved to be a nearly inexhaustible mine of bioactive peptides. While the structures of the frog skin peptides are specific to that species, related representatives can often be found in the brain or the gut of mammals. This is how bombesin and its counterparts in mammals were found.

While researchers from these two schools of neuropeptide discovery went about finding neuropeptides in a random manner, under the supervision of Matsuo, Minamino, and Kangawa, those from a third school used biological assays to find peptides that exhibit tachykinin-like effects. At the end of the 1970s and in the 1980s they discovered the neoendorphins, adrenomedullin, and the different neuromedins.

The Nucleic Acid-Based Approaches

The identification of novel neuropeptides has also resulted from experiments involving not peptide extractions but instead mRNA sequence analyses.

Neuropeptides discovered as part of precursors With the cloning of neuropeptide precursors at the end of the 1970s and in the 1980s, it became clear that most neuropeptide sequences are flanked by pairs of basic amino acid residues in the nonprocessed

precursors. In-depth studies on the processing of proopiomelanocortin (POMC) showed the importance of trypsin-like enzymes in the maturation of neuropeptides. Indeed, the first neuropeptide identified solely on the basis of its genomic sequence was γ-melanocyte-stimulating hormone (γ-MSH), found in the POMC precursor in 1979; this may be the first success of 'bioinformatics,' years before the introduction of the personal computer. The existence of several neuropeptides has been postulated on the basis of this principle.

Neuropeptides resulting from alternative splicing Because many neuropeptide precursors are transcribed from multiexonic genes, alternative splicing can give rise to new peptide sequences from a single gene. In 1983, calcitonin gene-related peptides (CGRPs) and the neurokinins were discovered as products from the alternative splicing of the calcitonin and substance P genes, respectively. Other neuropeptides have also been detected as alternatively spliced products of a known neuropeptide precursor; some, however, remain only putative neuropeptides as they have not been matched to a specific receptor.

Neuropeptides found by subtractive hybridization Another approach, which has its roots in the search for brain-specific mRNAs, uses subtractive hybridization to enrich for tissue-specific mRNAs. Applied to the search for mRNAs specific to different CNS tissues, it has led to the discovery of cortistatin and the hypocretins.

Neuropeptides found by homology screening Taking advantage of possible sequence similarities between different neuropeptide precursors is another way to discover new neuropeptides. In this case, genes or mRNAs are screened in nonstringent hybridization conditions with a nucleic acid probe derived from the sequence of a known precursor. Identified clones may be sequentially related to the probe and thus may encode a protein similar to the original precursor, that is, possibly a new neuropeptide precursor. This approach is applicable only to products that belong to gene families. While there are precursors that show an evolutionary relationship, the vast majority of them are not closely sequence related (especially when compared to the neuropeptide receptors). Urocortin, however, is an example of a neuropeptide discovered on the basis of its sequence similarities to the urotensin I precursor.

Neuropeptides found in silico With the genome sequence, bioinformatics has also been applied to the discovery of neuropeptides. Taking advantage of the facts that neuropeptides are often C-terminally amidated, that some contain conserved sequence, such as the RFamide sequence, and that they must be encoded by a precursor protein harboring a signal sequence, it is possible to design computer programs that can fish new neuropeptides from data banks. The hRFRP-1–3 peptides and the pyroglutamylated RFamide peptide (QRFP) were discovered using this approach.

The Orphan GPCR-Based Approach

Because neuropeptides exert their actions by activating GPCRs, some GPCRs that have not been matched to any known first messenger molecules, the so-called orphan GPCRs, have been used as targets to discover new neuropeptides. Orphan GPCRs are expressed in eukaryotic cells which are tested against tissue extracts, often brain extracts. Receptor reactivity is monitored by quantifying changes in second messenger levels. The tissue extracts showing reactivity are fractionated until the active component is purified to homogeneity.

The first success of the orphan receptor strategy was the discovery of a novel neuropeptide, named orphanin FQ or nociceptin (OFQ/N). This strategy has since led to the discovery of 13 neuropeptides: the orexins/hypocretins, prolactin-releasing peptide, apelin, ghrelin, metastin, neuropeptide B/W, the prokineticins, relaxin 3, neuropeptide S, and neuromedin S.

The advantage of this approach is that because it is based on the use of receptors, it ensures biological activity of the discovered peptides. It offers the most promise for finding new neuropeptides since several dozen GPCRs that may bind peptides are still orphan receptors.

How Many Neuropeptides Are Yet to Be Found?

By the mid-1990s, some 65 neuropeptides had been discovered. Other peptides, potentially neuropeptides, had also been found, in particular as processing products of known precursors. Some exhibit regulatory roles on particular responses. But according to the definition stating that a neuropeptide is a peptide that acts on a specific receptor expressed in the CNS, as long as these peptides are not matched to specific receptors they fall outside of the neuropeptide classification. There is a caveat associated with this neuropeptide definition, however, which has to do with receptor specificity. A GPCR can exhibit different affinities for structurally distinct ligands. Two unrelated peptides may bind the same GPCR because

they share some topological motif that permits activation. Of course, these two peptides will exhibit different affinities for the GPCR. As a rule of thumb, one can expect neuropeptides to activate their receptors with affinities in the low nanomolar range or lower. However, it is possible that low-affinity peptides activate the GPCR *in vivo* under particular circumstances, and thus the number of bioactive neuropeptides may be significantly larger than can be currently evaluated.

The surest way to go about predicting how many neuropeptides are yet to be found is to count the potential receptors. The application of the orphan receptor approach has led to the discovery of a dozen neuropeptides. As defined previously, if a neuropeptide is a peptide whose receptor is expressed in the CNS, then one can turn to the remaining orphan GPCRs to ask how many more may be found. There exist some 100 orphan GPCRs that, on the basis of their sequences, may bind ligands that are not chemosensory (olfactory, gustatory, chemoattractants, or chemokines). There is thus far no algorithm that can ascertain whether these orphan receptors may bind peptides or ligands of other chemical classes. Yet if one considers that about half of the first messengers that direct interneuronal signaling are neuropeptides, it can be extrapolated that half of the remaining nonchemosensory orphan GPCRs may bind peptides and thus that some 50 neuropeptides may still remain to be discovered.

See also: Invertebrate Neurohormone GPCRs; Neuropeptide Release; Neuropeptide Synthesis and Storage; Substance P/Tachykinins and its/their Receptors; Vasopressin/Oxytocin and Receptors.

Further Reading

Civelli O, Nothacker HP, Saito Y, Wang Z, Lin SH, and Reinscheid RK (2001) Novel neurotransmitters as natural ligands of orphan G-protein-coupled receptors. *Trends in Neuroscience* 24: 230–237.

de Lecea L, Criado JR, Prospero O, et al. (1996) A cortical neuropeptide with neuronal depressant and sleep-modulating properties. *Nature* 381: 242–245.

Erspamer V, Melchiorri P, Erspamer CF, and Negri L (1978) Polypeptides of the amphibian skin active on the gut and their mammalian counterparts. *Advanced Experimental Medicine Biology* 106: 51–64.

Guillemin R (2005) Hypothalamic hormones a.k.a. hypothalamic releasing factors. *Journal of Endocrinology* 184: 11–28.

Hughes J, Smith TW, Kosterlitz HW, Fothergill LA, Morgan BA, and Morris HR (1975) Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature* 258: 577–580.

Minamino N, Sudoh T, Kangawa K, and Matsuo H (1985) Neuromedins: Novel smooth-muscle stimulating peptides identified in porcine spinal cord. *Peptides* 6: 245–248.

Nakanishi S, Inoue A, Kita T, et al. (1979) Nucleotide sequence of cloned cDNA for bovine corticotropin-beta-lipotropin precursor. *Nature* 278: 423–427.

Pierce JG and du Vigneaud V (1950) Studies on high potency oxytocic material from beef posterior pituitary lobes. *Journal of Biological Chemistry* 186: 77–84.

Reinscheid RK, Nothacker HP, Bourson A, et al. (1995) Orphanin FQ: A neuropeptide that activates an opioidlike G protein-coupled receptor. *Science* 270: 792–794.

Rosenfeld MG, Mermod JJ, Amara SG, et al. (1983) Production of a novel neuropeptide encoded by the calcitonin gene via tissue-specific RNA processing. *Nature* 304: 129–135.

Tatemoto K and Mutt V (1980) Isolation of two novel candidate hormones using a chemical method for finding naturally occurring polypeptides. *Nature* 285: 417–418.

Von Euler US and Gaddum JH (1931) An identified depressor substance in certain tissue extracts. *Journal of Physiology* 72: 74–87.

Neuropeptides: Electrophysiology

W Zieglgänsberger, Max Planck Institute of Psychiatry, Munich, Germany

© 2009 Elsevier Ltd. All rights reserved.

Neuropeptides versus Classical Neurotransmitters

There are several main differences between neuropeptides (3–100 amino acid residues) and classical neurotransmitter molecules such as glutamate, γ-aminobutyric acid (GABA), dopamine, noradrenaline, glycine, and acetylcholine with respect to their synthesis, storage, and signaling function. Because classical transmitters can be rapidly synthesized in nerve terminals and stored in small synaptic vesicles they maintain a fairly constant level, whereas neuropeptide transmitters, which are synthesized in the cell body and stored in large dense-core vesicles transported to the terminal, may be more rapidly depleted. Neuropeptides, which are up to 50 times larger than classical neurotransmitters, display a complex three-dimensional structure, and some possess several recognition sites with nanomolar affinity for their receptors.

Neuropeptides are typically well conserved across species and outnumber the classical neurotransmitters. They are present in numerous neurons, together with one or even two small classical neurotransmitters (coexistence), which may affect their target neurons differentially. In cases of co-localization the peptide transmitters are often released at higher firing rates and particularly under burst-firing patterns from afferent fibers (Figure 1). In general, it is unlikely that neuropeptides participate in rapid integrative functions in a neuronal assembly.

Volume Transmission

In contrast to classical neurotransmitters, which interact with receptors linked to elaborate pre- and postsynaptic specializations, neuropeptides, once released, diffuse and may activate receptors far from their release site, a mechanism termed volume transmission (Figure 2). Thus, in contrast to conventional synaptic transmission neuropeptides may influence large ensembles of neurons in a divergent gating function rather than mediating rapid local point-to-point communication. The effective concentrations of neuropeptides needed for this mode of signaling are much lower than for synaptic transmission mediated by classical neurotransmitters, which reach much higher concentrations at their receptors, due to the small volume of the synaptic cleft. The diffusion of the neuropetide depends on the size of the extracellular space, its tortuosity, and the degradation encountered. After diffusing, the neuropeptide binds more slowly, but with much higher binding affinity and selectivity, to its receptor than classical transmitters confined to the synaptic cleft. Peptide-induced receptor trafficking and redistribution of heterogeneous extrasynaptic receptors (Figure 3), spontaneous release and specific regional inactivation mechanisms, and information transfer via glia cells add to the complexity of this mode of delocalized signaling.

Neuropeptides: Synthesis, Release, and Degradation

Neuropeptides occur in a bewildering range of shapes and sizes in the animal kingdom. Neuropeptides are found in all types of neurons and glia cells of the mammalian central nervous system (CNS). Genes that code for a neuropeptide family are generally not clustered and are spread out all over the mammalian genome. Duplications and point mutations in old primordial genes are probably the origin of the presently known members of the family of a neuropeptide. Because the precursors for neuropeptides are encoded in the genome, their functions can be detailed in transgenic animal models that do not produce or that overexpress the neuropeptide. However, the deletion of one neuropeptide gene often shows no noticeable effect because the action of the coded neuropeptide may be too subtle or is compensated by related neuropeptides.

Significantly, even if the gene that encodes the precursor is identified, it is not clear what particular peptide it will create because neuropeptide precursors undergo extensive posttranslational processing. The proneuropeptides are processed into active forms in large dense-core vesicle precursors by the concurrently packaged enzymes. Although certain peptidases may produce a variety of biologically active fragments, most commonly the peptide that is stored in vesicles and then released is considered the transmitter.

After trafficking into the presynaptic compartment, they may dock and fuse with the plasma membrane after a specific priming process at sites different from the synaptic vesicles containing classical neurotransmitters. The release from dense-core vesicles is probably regulated by specific protein kinases, and they are replenished after exocytosis by anterograde

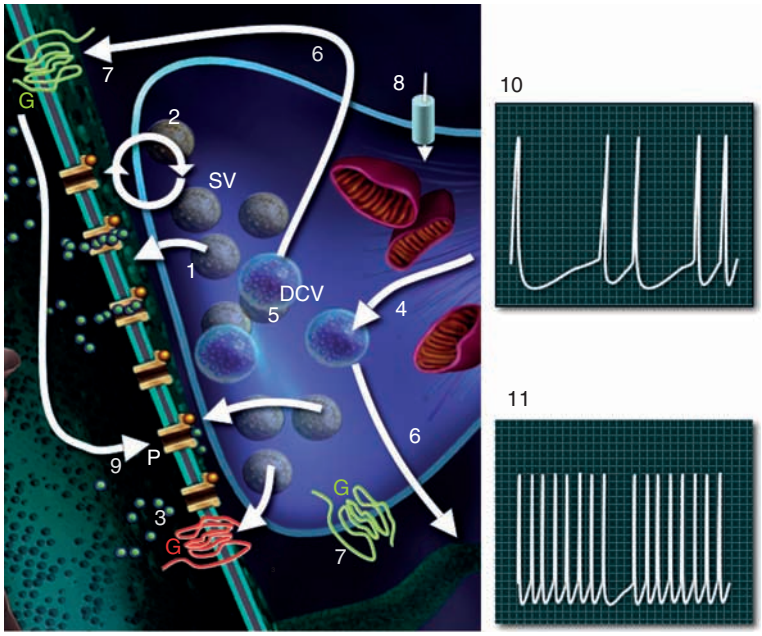

Figure 1 Co-localization and co-release of classical neurotransmitters and neuropeptides. Classical transmitters are synthesized in nerve terminals and stored in small synaptic vesicles (SV), released, and recycled (1 and 2). They interact with subsynaptic receptors, including metabotropic receptors (3). Neuropeptides are synthesized in the cell body and transported to the terminal (4), stored in large dense-core vesicles (DCV; 5) and released at nonsynaptic sites (6). Neuropeptides interact with pre- and postsynaptic receptors (7). Voltage-gated Ca^{2+} channels control the release of classical transmitter and neuropeptide (8). Activation of a neuropeptide receptor induces, e.g., the phosphorylation (P) of an ionotropic receptor (9). Peptide transmitters are often released at higher firing rates, particularly under burst-firing patterns (10 and 11).

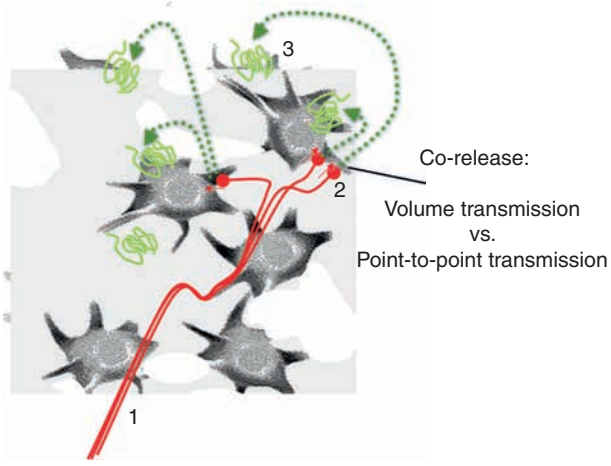

Figure 2 Point-to-point vs. volume transmission. Neuropeptides, once released (1 and 2), diffuse and activate receptors (3) far from their release site (volume transmission). Classical neurotransmitters mediate local point-to-point communication (2). Co-released neuropeptides participate in signaling at nonsynaptic and synaptic somatic and dendritic sites.

transport of new secretory granules from the *trans*-Golgi network. Classical neurotransmitters are packaged into small clear synaptic vesicles that are clustered near release sites at synaptic terminals where they undergo repeated rounds of exocytosis, endocytosis, and refilling with neurotransmitter.

The function of neuropeptide-encoding mRNAs in neuronal processes beyond the perikaryon still remains to be elucidated. Neuropeptides participate in signaling at nonsynaptic and synaptic somatic and dendritic sites of the target neuron, and there is evidence that they also exert trophic actions. The dendritic release of neuropeptides is partially independent from axon terminal release, and the exocytosis machinery at the two sites seems to be different (Figure 4). Action potentials backpropagating from the axon hillock into the dendritic tree probably open high-voltage-activated Ca^{2+} channels to trigger and modulate dendritic release. Synapses located on dendritic spines are probably important sites for the plastic changes produced by such a retrograde signaling, which questions the classical point of view of synaptic transmission and short-term modifications of excitatory and inhibitory synaptic efficacy. The invasion of the dendritic tree by the backpropagation of action potentials is subtly controlled by, for example, dendritic voltage-sensitive Na^{+}- and K^{+}-conductances, and it appears to be highly dynamic. It remains to be shown whether the dendritic release of neuropeptides and classical transmitters is, in addition, differentially

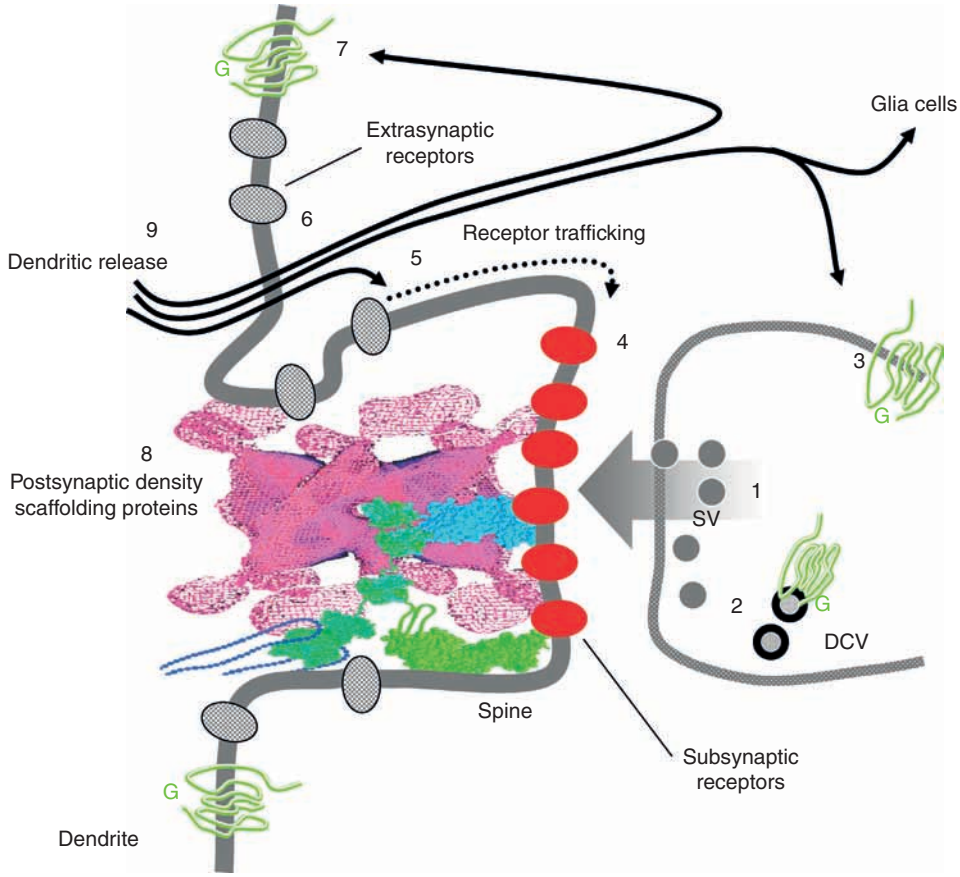

Figure 3 Receptor localization and trafficking on a dendritic spine. Transmitter release from synaptic vesicles (SV; 1); some dense-core vesicles (DCV) carrying receptors for the neuropeptides (2); presynaptic receptors (3) and postsynaptic receptors (7). Receptors moving from extrasynaptic sites to subsynaptic sites to join subsynaptic receptors (receptor trafficking; 4, 5, and 6); postsynaptic density in a dendritic spine (8); dendritic release (9).

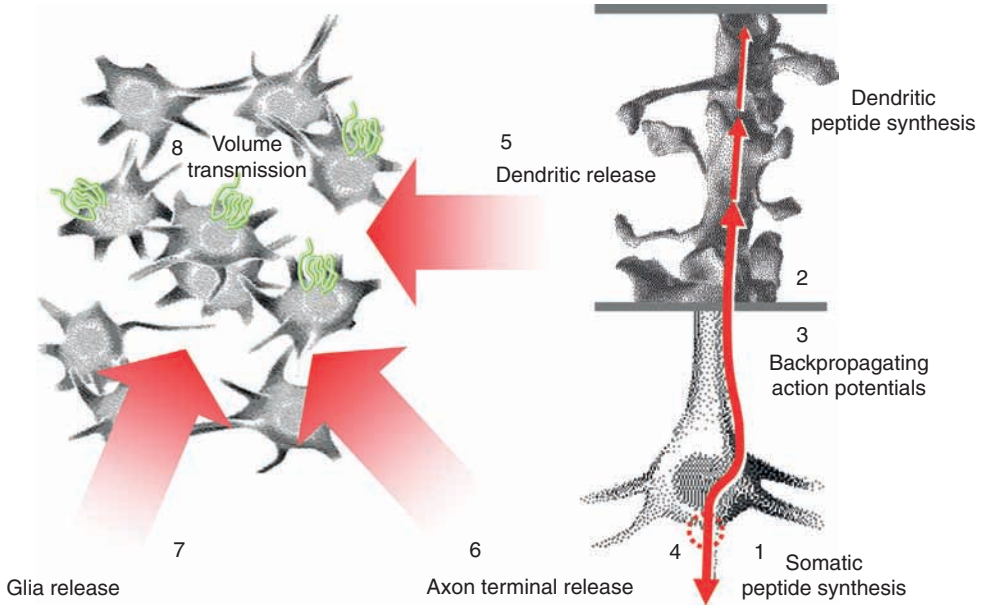

Figure 4 Release sites of neuropeptides. Neuropeptides are synthesized somatically (1) and dendritically (2). Dendritic release (5) is triggered by action potentials backpropagating (3) from the axon hillock (4) into the dendritic tree. Peptides are also released from presynaptic terminals (6) and glia cells (7) and may act in concert on remote receptors (volume transmission; 8).

regulated by the release of Ca^{2+} from intracellular pools. New experimental approaches employing novel techniques will soon lead to a better understanding of the dynamics of dendritic computation and its contribution to neuronal signal processing.

The release of a neuropeptide is, like that of classical transmitters, Ca^{2+}-dependent, but requires higher frequencies of discharge in afferent fibers and is considerably slower because neuropeptides cannot diffuse through the diffusion pore. There is evidence that neuropeptides may come into play as messengers mainly when the nervous system is trying to respond to a pathological challenge such as nerve injury, seizure, or stress. Under these conditions, neurons and glia cells change their phenotype in regard to neuropeptides and their endowment with neuropeptide receptors. This phenotypic switch may be seen as an attempt to promote survival and regenerative mechanisms in a neuronal system.

After release, peptide transmitters appear to be inactivated enzymatically or by diffusion. With few exceptions (e.g., cholecystokinin), they lack a high-affinity reuptake process that removes the transmitter from the extracellular space. The introduction of *in vivo* push–pull cannulae and microdialysis has been useful for the measurement of extracellular concentrations of neuropeptides. Another successful *in vivo* approach has been to insert electrodes coated with antibodies into the specific area of the brain for a short period. The removed probes are then exposed to a radiolabeled peptide, which binds to receptors that are not already occupied by the peptide that was released in the brain tissue. Without these techniques, the release of some neuropeptides may not have been detected because they were present at too low a concentration or only during particular developmental stages.

Neuropeptide Receptors

Neuropeptides bind more tightly than classical neurotransmitters to their receptors, which are with few exceptions G-protein-coupled receptors (GPCRs). The receptors for neuropeptides belong to different families (rhodopsin-like, glucagon-vasoactive intestinal peptide/calcitonin). Several GPCRs form homodimers or heterodimers, and most commonly several receptor subtypes exist for a given peptide ligand. Furthermore, there is evidence that so-called receptor activity-modifying proteins can alter the pharmacological profiles of a single receptor.

Most neuropeptide receptors are rapidly ($\sim$5 min) desensitized after agonist activation. The receptor–neuropeptide complexes are internalized, and the receptors are then recycled and restored to the cell membrane within $\sim$30 min. It remains to be shown to what extent this recycling process is brought about by metalloendopeptidases affecting intracellular degradation of signaling peptides. Conjugation of the toxin saporin to the neuropeptide (e.g., substance P) can be used to selectively destroy neurons involved in pain processing in the dorsal horn of the spinal cord. Animal models show that this selective targeting technology prevents the formation of hyperalgesia and can reverse established neuropathic pain behavior.

The action of an endogenously released neuropeptide may depend on the local circuitry as well as the strategic location of the receptor subtypes. In addition to actions at neighboring neurons, neuropeptides can regulate the excitability of the neuron from which they are released via an autoreceptor function. The release of some neuropeptides seems to be regulated in a reciprocal fashion by autoreceptors of coexisting classical neurotransmitters such as dopamine.

Peptidergic Interneurons

The interneuronal system, often classified according to neuropeptide content, is crucial for maintaining appropriate brain function under normal and pathophysiological conditions. It is widely accepted that the effects of neuropeptides can be very distinct at the network level and that neuropeptides released from interneurons can trigger diverse K^+ currents such as the voltage-sensitive K^+ current I_M or the voltage-insensitive K^+ leak current I_{LK}. Both currents hyperpolarize the neuron and shunt excitatory synaptic potentials and can markedly affect neuronal excitability. The augmentation of the nonselective cation current I_H and the reduction of voltage-sensitive N-type Ca^{2+} channels have been suggested as mechanisms underlying pre- and postsynaptic peptidergic signaling.

Electron-microscopy studies have shown neuropeptide-immunoreactive terminals localized to the dendritic shafts and spines of principal neurons. There is evidence from recent studies that dendritic spines undergo an activity-dependent structural remodeling and that memories are created by alterations in glutamate-dependent excitatory synaptic transmission on dendritic spines. In the past, electrophysiological recordings were often limited to short-lasting impalements of the soma of only vaguely identified neurons. The advent of novel recording techniques for synaptic potentials and single-channel activity from dendritic sites remote from the soma under visual control and its combination with extremely targeted stimulation techniques in which the active transmitter is uncaged from a larger molecule by an ultraviolet (UV) laser has opened the possibility that we can detail these processes. Electrophysiological investigations into the acute effects of neuropeptides suggest that

they interfere not only with short-term alterations in synaptic efficacy but can also affect neuronal plasticity in neuronal circuits in a long-lasting way through indirect effects on release mechanisms of excitatory or inhibitory classical transmitters. The inhibition of glutamate release decreases the frequency of spontaneous glutamate-mediated events and the amplitude of evoked excitatory postsynaptic responses. Such a reduction in glutamate release and co-released GABA could act in concert to reduce the likelihood of neuronal firing and the generation of long-term potentiation. It has been shown for central neurons that the early and the late components of a GABAergic inhibitory postsynaptic potential (IPSP) are mediated by $GABA_A$ and $GABA_B$ receptors, respectively.

Synaptic strength at glutamatergic synapses shows a remarkable degree of use-dependent plasticity. This dynamic regulation of synaptic efficacy plays a crucial role in experience-dependent modification as well as in the formation of neuronal connections in neural circuitry as physiological correlates to learning and memory.

Some neuropeptides may be considered as endogenous antiepileptic agents because they reduce excitatory synaptic transmission in the CNS. Their application elicits a long-lasting decrease in evoked excitatory postsynaptic current (EPSC) amplitude and a delayed, long-lasting increase in the amplitude of an evoked monosynaptic inhibitory postsynaptic current (IPSC). Like opioid peptides, substance P and other neuropeptides of the tachykinin family also can indirectly affect pyramidal neurons via an effect on neighboring hippocampal CA1 interneurons. Neurokinin (NK)-1 receptors are located both on presynaptic terminals and at somatodendritic sites of GABAergic neurons. By increasing the excitability of CA1 GABAergic interneurons, tachykinins can powerfully facilitate the inhibitory synaptic input to pyramidal neurons. This indirect inhibition could play a role in regulating short-term and long-term synaptic plasticity, promoting neuronal circuit synchronization or influencing epileptogenesis. Because tachykinins do not alter the frequency or amplitude of miniature inhibitory synaptic currents and are without effect on evoked inhibitory synaptic currents, these neuropeptides probably act at the somatodendritic membrane of GABAergic interneurons rather than at their axon terminals. In the hippocampus, opioid peptides increase the excitability of CA1 pyramidal neurons via a reduction of GABAergic inhibition from neighboring interneurons (disinhibition).

Peptidergic Synaptic Potentials

Most of our knowledge about peptidergic transmission and most therapeutically used drugs is based on exogenous application in which substances may diffuse in the extracellular space and evoke their actions, for example, via the simultaneous activation and/or inhibition of ensembles of neurons or glia cells. Even when highly selective agonists and antagonists are available or transgenic mice expressing antibodies to the neuropeptide under study are combined with state-of-the-art single-unit recordings in *in vitro* preparations, these studies are still notoriously hard to interpret.

Most of the information that has been gathered so far about the actions of neuropeptides in the CNS *in vivo* derives from extracellular recording of single-unit activity in combination with iontophoretic or pneumatic administration from multibarreled micropipettes in regions containing a high density of specific binding sites. Most of these *in vivo* studies do not attempt to differentiate a presynaptic site of action from overlapping postsynaptic effects or to characterize the role of single neurons in an ensemble of neurons.

There is evidence that endogenously released neuropeptides shift membrane permeability and can directly cause postsynaptic inhibition and elicit peptide-mediated IPSPs. These findings indicate that neuropeptides may not just alter the response of a target neuron to classical neurotransmitters without interfering directly with neuronal excitability. However, such consistent postsynaptic effects on neuronal excitability are documented only for few neuropeptides. Such a shift of the membrane to a more hyperpolarized level directly reduces the voltage-dependent glutamatergic transmission mediated by *N*-methyl-D-aspartate (NMDA) receptors involved, for example, in cognition or epileptogenesis and shift glutamatergic transmission preferentially to α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor-mediated mechanisms in the whole neuron. The generation of excitatory postsynaptic potentials (EPSPs) by neuropeptides has been reported in neurons of the peripheral nervous system and a few neurons in the CNS.

Growth Factors

Some of the known growth factors such as brain-derived nerve growth factor (BDNF), which are secreted from neurons and glia cells, fulfill the neuropetide criteria and may act like neurotransmitters or get indirectly involved in the interneuronal communication by affecting, for example, ion transporters, which control the intracellular concentration of chloride in neurons. It has recently been shown that BDNF released from activated glia cells reduces the chloride transport out of the target neurons. The increase in intracellular chloride concentration shifts the commonly inhibitory effects of $GABA_A$ receptor activation toward excitation in these GABAergic

neurons, and thus dramatically affects acute functioning and plasticity in these neuronal circuits.

Conclusion

Considerable progress has been made in the field of neuropeptide research, and fascinating detailed insights into novel signaling modes in the nervous system have been gained through the advent of new technologies. New neuropeptides were described, and a multitude of receptors for neuropeptides have been characterized and cloned. Nevertheless, there is still only little information available on their role in humans. Because numerous neuropeptides are involved in mammalian brain functions under normal and pathological conditions, it was predicted that many agonistic and antagonistic drugs, as well as peptidase inhibitors preventing peptide breakdown and thus strengthening the peptidergic transmission, would rapidly be developed for therapeutic use. To date, this prediction has not yet materialized. Only substance P has secured a position in the field of psychopharmacology. It was shown that substance P (NK-1) antagonists have an antidepressant activity comparable to the most commonly prescribed serotonin reuptake inhibitors but are associated with less pronounced side effects. There is no doubt that peptidergic signaling mechanisms in the human brain will remain important targets for drug development and novel therapeutic algorithms.

See also: Neuropeptide Inactivation or Metabolism; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides in Autonomic Neurons; Neuropeptides: Discovery; Neuropeptides: Endocrine Cells; Neuropeptides: Enteric Nervous System.

Further Reading

Baraban SC and Tallent MK (2004) Interneuronal neuropeptides – endogenous regulators of neuronal excitability. *Trends in Neuroscience* 27: 135–142.

deWied D and Burbach JPH (1999) Neuropeptides and behaviour. In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience,* vol. 2, pp. 1431–1436. Amsterdam: Elsevier.

Dodt HU, Eder M, Schierloh A, and Zieglgänsberger W (2002) Infrared-guided laser stimulation of neurons in brain slices. *Science Signal Transduction Knowledge Environment* 120: PL2 [online].

Hökfelt T, Bartfai T, and Bloom F (2003) Neuropeptides: Opportunities for drug discovery. *Lancet Neurology* 2: 463–472.

Ubink R, Calza L, and Hökfelt T (2003) "Neuro"-peptides in glia: Focus on NPY and galanin. *Trends in Neuroscience* 26: 904–909.

Zieglgänsberger W, Dodt H-U, Deisz R, and Pawelzik H (1991) Peptides in neurotransmission. In: Fuxe K and Agnati L (eds.) *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission,* pp. 371–380. New York: Raven Press.

Neuropeptides: Endocrine Cells

N Wierup and F Sundler, Lund University, Lund, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuropeptides can be defined as peptides that are expressed in neurons and have a transmitter or cotransmitter function. However, it is common knowledge that many neuropeptides are not restricted to neurons, but occur also in other cells, endocrine cells in particular (Table 1). It should be mentioned that the opposite is also true; several peptides first discovered as hormones produced by endocrine cells were later shown to be neuropeptides, since they were expressed in neurons as well. The close relationship between peptide hormone-producing endocrine cells and neurons, and also between central and peripheral neurons, with regard to peptide expression has fostered an array of concepts and terms, such as neurohormonal peptides, brain–gut (or gut–brain) peptides, and the more neutral term regulatory peptides. There are several other histochemical and morphological characteristics that endocrine cells and neurons have in common (Table 2). Based on these similarities it was long held that endocrine cells in the gut, the pancreas, and the airways, together with the thyroid C cells and the adrenal medullary cells, all were embryologically of neuroectodermal origin. However, it is now quite firmly established that endocrine cells in the gut, pancreas, and airways originate from the endoderm (i.e., of the same embryonic origin as the epithelial cells that surround them), while C cells and adrenomedullary cells do have neuroectodermal origin. The following discussions represent a survey of the main lines of development in this field, focusing on those peptide genes that display a dual expression in neurons and endocrine cells.

Hormones Later Becoming Neuropeptides

The very first hormone described, thereby giving rise to the concept of hormones and thus forming the origin of endocrinology, was secretin. More than 100 years ago this substance was functionally defined in duodenal extracts by virtue of its stimulating effect on pancreatic fluid secretion. Around 70 years later secretin was chemically identified as a peptide. Still later the secretin-producing endocrine cells could be identified by immunocytochemistry (ICC) in the epithelium of the upper small intestine. However, not until quite recently was secretin gene expression described in the brain, and, in fact, much of current secretin research centers around the possible role of the peptide in the pathogenesis of brain disease, autism in particular.

Other peptides have a longer history as brain–gut peptides. One example is cholecystokinin (CCK), also long known as a gut hormone. However, rather soon after its chemical characterization, and the ensuing development of antibodies for radioimmunoassay (RIA) and ICC, it was found that CCK was also a neuropeptide with a wide distribution in both central and peripheral neurons.

The inevitable trend is that, as new techniques are developed, and as those already available become more and more sensitive, increasing numbers of peptide hormones turn out to have a neuronal localization as well. One recent example is the intestinal hormone GIP, initially named gastric inhibitory peptide but later renamed with the physiologically more appropriate descriptor, glucose-dependent insulin-releasing peptide; GIP recently has been found to be discretely expressed also in the brain. Correspondingly, certain pituitary hormones, such as proopiomelanocortin-derived hormones (adrenocorticotropic hormone, α-melanocyte-stimulating hormone, β-endorphin), and also growth hormone, have become to be viewed as neuropeptides in that they have been demonstrated in neuronal systems in the brain.

Neuropeptides Entering Endocrine Cells

Over the years a steadily increasing number of newly discovered peptides have first been localized to neurons in the central and/or peripheral nervous systems, and have later been found to exist also in endocrine cells. Thus, calcitonin gene-related peptide (CGRP), a product of an alternative calcitonin pre-mRNA splicing, was, apart from a low level of co-expression with the parent mRNA in C cells, initially thought to be expressed more or less exclusively in central and peripheral neurons: However, a few years after its discovery, CGRP was detected also in islet endocrine cells and in endocrine cells of the airways. Another example is gastrin-releasing peptide (GRP), a bombesin-like peptide first established as a neuropeptide with a wide distribution in the enteric nervous system (ENS), but soon found to occur also in endocrine cells in the airways and elsewhere. Neuropeptide Y (NPY), a major peptide in both central and peripheral neurons, was quite rapidly localized also to endocrine cells of the adrenal medulla, and later to islet endocrine cells. One more recent example is cocaine- and

Table 1 Peptides expressed in both neurons and endocrine cells in mammals

Peptides regarded as principal hormones, with their own cell type:
Propiomelanocortin derivatives (adrenocorticotropic hormone, melanocyte-stimulating hormone, β-endorphin, and other processing products)
Growth hormone (some evidence of neuronal expression)
Ghrelin (some evidence of neuronal expression)
Secretin (some evidence of neuronal expression)
Cholecystokinin
Glucose-dependent insulin-releasing peptide (some evidence of neuronal expression)
Motilin (some evidence of neuronal expression)
Somatostatin
Proglucagon derivatives (glucagon-like peptides GLP-1 and GLP-2 and other processing products)
Neurotensin
Co-expressed peptides in endocrine cells:
Enkephalins (adrenal medulla and gut)
Dynorphins (islets)
Neuropeptide Y (adrenal medulla and islets)
Peptide YY (gut and islets)
Calcitonin gene-related peptide (thyroid, islets, and airways)
Islet amyloid polypeptide (islets and stomach)
Thyrotropin-releasing hormone (islets during development)
Substance P and other tachykinins (gut enterochromaffin cells and carcinoid tumors)
Pituitary adenylate cyclase-activating peptide (adrenal medulla)
Galanin (adrenal medulla)
Chromogranins, and their processing products – e.g., pancreastatin from chromogranin A, secretoneurin from secretogranin II (many different endocrine cell types)
Cocaine- and amphetamine-regulated transcript (islets, gut, and adrenal medulla)
Gastrin-releasing peptide/bombesin (airways)
Orexin[a] (islets and gut)
Vasoactive intestinal peptide (endocrine tumors)
Growth hormone-releasing hormone (endocrine tumors)

[a]Claimed but not verified.

Table 2 Features common to peptide-producing endocrine cells and neurons

Endogenous amines and other 'classical' neurotransmitters stored in secretory granules/dense-cored vesicles
Capacity to form amines from endogenous or exogenous precursor amino acids, so-called APUD (amine precursor uptake and decarboxylation) mechanism
Vesicular monoamine transporters (V-MATs)
Certain 'neuron-specific' enzymes and other marker proteins (e.g., cholinesterases, neuron-specific enolase, protein gene product 9.5, neurophysin)
Granule-stored chromogranins and related proteins and their processing products
Granule-stored peptide-processing enzymes

amphetamine-regulated transcript (CART) peptides, which were first identified as abundantly expressed peptides in both central and peripheral neurons. CART was subsequently found to be expressed in endocrine cells of the adrenal medulla, the pancreatic islets, and the intestine, and, at least in rats, also in antral gastrin cells of the stomach.

Messenger Coexistence in Endocrine Cells and Neurons

One feature that is characteristic of a peptide, whether occurring in endocrine cells or in neurons, is that it is not the one and only hormone/transmitter in these cells, but rather is co-localized with other peptides and/or 'classical' neurotransmitters (Table 3). This feature of co-localization is sometimes referred to as 'chemical coding.' In fact, the chemical coding, which can be extensive in certain neurons of the ENS, is quite remarkable in its extent also in certain endocrine cells, such as in insulin-producing beta cells in pancreatic islets. The frequent co-existence of peptides indicates an even higher complexity to the way the target cells are regulated. Unraveling of the functional significance of co-expression of peptides and of peptides and other messengers in endocrine cells is still in its infancy, but some interesting advances have been made in the nervous system. Thus, the concept

Table 3 Releasable substances in granules of both endocrine cells and neurons

Low-molecular-weight messengers:
Amines (e.g., serotonin, melatonin, dopamine, noradrenaline, adrenaline, histamine)
Amino acids (e.g., γ-aminobutyric acid, glutamic acid)
Adenosine triphosphate
Nitric oxide
Carbon monoxide
Peptides and proteins:
Messenger peptides
Other propeptide processing (silent?) fragments
Acidic proteins (chromogranins and related proteins)
Peptide-processing enzymes

Table 4 Possible advantages of multiple messengers in endocrine cells

Coordinated effects at close range (paracrine or autocrine actions) and at a distance (endocrine actions)
Coordinated effects over time: immediate effects and delayed (e.g., trophic) effects
Precise and detailed instructions to target cells
Flexibility of command by changes in the combination of messengers expressed

Table 5 Plasticity of messenger expression

Species dependent
Age dependent
Activity dependent

of cotransmission (i.e., cooperation at the target cells of co-stored and simultaneously released messengers) was fueled by a series of observations on sympathetic nerves co-storing NPY and noradrenaline, in regulating vascular tone and blood flow, and on parasympathetic nerves co-storing vasoactive intestinal peptide (VIP) and acetylcholine, in regulating secretion from exocrine glands. It is not inconceivable that similar mechanisms operate also in the endocrine system; co-stored and co-released peptides (hormones and other regulatory peptides) may cooperate at the target cells. Some of them may, alternatively or additionally, have paracrine or autocrine functions by targeting neighboring cells or the cell of origin.

To be able to understand the full physiological significance of messenger coexistence in endocrine cells, we need to address a number of questions:

- What is the messenger mix (chemical coding) of each individual population of endocrine cells?
- How is the expression of the co-stored messengers controlled?
- What advantages do co-storage and co-release offer?

Detailed answers to these questions cannot yet be given. Possible functional advantages of the coexistence and co-release of messengers are listed in Table 4. It is clear, however, that the number of messengers – including peptides – in certain endocrine cells is high, and that some peptides are not permanent residents, but are rather transiently expressed during development and in pathophysiological conditions, as illustrated by islet beta cells (see later). One notable observation regarding coexisting peptides is the variation among species in the peptide constellation in a given cell type. Consequently, data obtained in common laboratory animals, such as rats and mice, may not in all details be valid for other mammals. Hence, caution is needed when extrapolating data from experimental animals to humans. The main aspects of peptide plasticity in endocrine cells are summarized in Table 5. An emerging issue in this context is that in tissues harboring both nerves and endocrine cells, a peptide may in one species be located preferentially in nerve fibers, while in another species the primary site of expression is in endocrine cells. It may then be asked whether a peptide can play the same paracrine role regardless if it is released from a nerve fiber or from an endocrine cell. Possibly, a neuronal localization of the messenger gives the potential of a more rapid release – and response of the target cell.

Adrenal Medulla

As more and more information has been gathered, the catecholamine-producing endocrine cells of the adrenal medulla have turned out to be a virtual storehouse of neuropeptides. Some of these peptides may occur exclusively in endocrine cells, whereas others are, in addition, present in neuronal elements. Many of the peptides are known to participate in the regulation of the secretory activities of the gland, presumably by paracrine or autocrine mechanisms. It cannot be excluded, however, that some peptides reach the circulation to join the battery of amines and peptides that function as hormones. Among peptides known to be produced in adrenomedullary cells are enkephalins, NPY, galanin, somatostatin, CGRP, VIP, pituitary adenylate cyclase-activating peptide (PACAP), and CART. Although not studied in detail, it is evident that many peptides coexist in the same cells. It should also be mentioned that the exact cellular localization (adrenaline- vs. noradrenaline-producing cells), the extent of expression (all cells of a given type or a subpopulation of them), and the presence in endocrine cells versus in neuronal elements are to some extent species dependent.

Islet Beta Cells

Beta cells, in addition to insulin, and the second proinsulin-derived C-peptide, invariably produce islet amyloid polypeptide (IAPP, or amylin). Insulin and IAPP are known to be co-stored in the same secretory granules and co-released at secretion. However, the two encoding genes may not always be co-regulated. Beta cells, in addition, express a plethora of neuropeptides during development and under pathophysiological conditions. Such peptides include CART, NPY, and thyrotropin-releasing hormone (TRH), as studied mostly in rodents to date. Thus, CART, NPY, and TRH are strongly upregulated pre- and neonatally, but are then gradually downregulated to very low levels of expression, or no expression, in the adult beta cell. It is notable that the upregulation of CART during development is not limited to beta cells but occurs in several of the other islet endocrine cell types (Figures 1(a)–1(d)). Interestingly, upon severe and long-standing challenge of the beta cells in adults, as in states of insulin resistance (i.e., in animal models of type 2 diabetes), expression of CART (Figures 1(m)–1(p)) and NPY is again heavily upregulated, particularly during the phase of islet expansion when beta cells are capable of compensating for the increased insulin demand by cell hypertrophy and hyperplasia.

Intestinal L Cells

An important population of L cells (defined by their comparatively large secretory granules) express the glucagon gene. However, in these cells proglucagon processing yields glucagon-like peptides (GLP-1 and GLP-2) rather than glucagon, which remains embedded in a larger peptide, glicentin, lacking glucagon

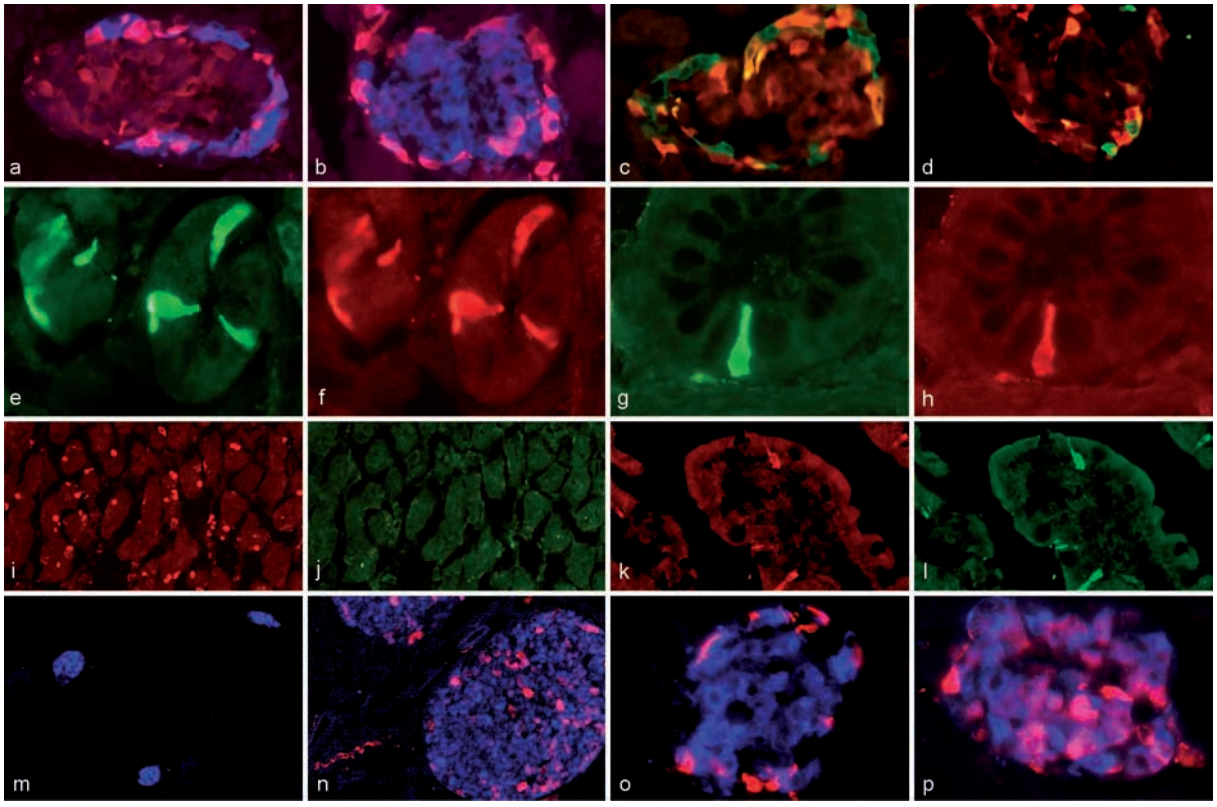

Figure 1 (a–d) Fetal (embryonic day 20) rat islets. Double immunostaining for cocaine- and amphetamine-regulated transcript (CART; red in all images) and each of the four major islet hormones, illustrating that CART is expressed in all of these islet cell types during a period around birth. (a) Double staining for CART and glucagon (blue). (b) Double staining for CART and insulin (blue). (c) Double staining for CART and somatostatin (green). (d) Double staining for CART and pancreatic peptide (green). CART is to a varying degree co-expressed (violet or yellow) with all four islet hormones in a subpopulation of each cell type. (e–h) Distal ileum of mouse (e, f) and pig (g, h). Double immunostaining for proglucagon (e, g) and peptide YY (f, h). The two peptides coexist in the L cells. (i–l) Human gastric fundus – oxyntic mucosa (i, j) and duodenal villi (k, l) – double immunostained for ghrelin (i, k) and motilin (j, l). Ghrelin and motilin are co-localized in the same cells in the duodenum (k, l), while gastric ghrelin cells lack motilin (i, j). (m–p) Rodent islets double immunostained for CART (red) and insulin (blue). (m) Low magnification of normal mouse islets and islets of an ob/ob mouse (n). CART is robustly upregulated in many beta cells of the ob/ob mice, whereas the normal islets have no detectable CART in islet cells. Note also the bundle of CART immunoreactive nerve fibers in panel n. (o) Higher magnification of normal rat islet and diabetic GK rat islet (p). CART is restricted to delta cells in the normal islet, while it is upregulated in many beta cells of the GK rat islet. Co-expression is represented by violet staining.

bioactivity. These L cells predominate in the distal intestine and, in addition to proglucagon, they express peptide YY (PYY), as documented in several mammals (Figures 1(e)–1(h)).

Coexistence of Peptides with Overlapping Functions

Evidence is emerging that certain peptides that have, at least in part, overlapping functions also overlap in their cellular localization. Thus, it has been shown in the human intestine that a significant population of the GIP-producing endocrine cells, the K cells, also express GLP-1. Both of these peptides are known to stimulate insulin secretion and belong to the group of incretin hormones. Furthermore, ghrelin and motilin, both of which stimulate gastric emptying and are implicated in the regulation of gut motility via the generation of migrating motor complexes (MMCs), also partly overlap in their cellular localization. Although a majority of gastrointestinal ghrelin cells are located in the stomach in human adults, a significant population is present also in the mucosa of the upper intestine, and a vast majority of these latter cells are, in fact, identical with motilin cells (Figures 1(k) and 1(l)). It is notable that gastric ghrelin cells do not express motilin (Figures 1(i) and 1(j)), nor do the ghrelin cells present in the pancreatic islets. This partial coexistence arrangement not only indicates phenotypical differences, but suggests also functional differences among subpopulations of ghrelin-producing cells, adding further complexity to the functional characterization of endocrine cells and their hormones.

The peptides coexisting in the islet beta cells may also have overlapping functions. IAPP and NPY are known as inhibitors of insulin secretion. In rats, CART is also an inhibitor of basal insulin secretion, but differs from the other two peptides by its potent augmenting effect on secretion stimulated by glucose and cAMP activating agents such as GLP-1.

Do peptides GLP-1 and PYY, coexisting in intestinal L cells, also fit into this concept (keeping in mind that PYY1–36 is known to be orexigenic, whereas GLP-1 is an acknowledged anorexigenic peptide)? It is not unlikely that they do, since PYY1–36 is rapidly converted to PYY3–36, which is also anorexigenic. This property of PYY3–36 is presumably due to its high selectivity for Y-2 receptors known to inhibit the release of the orexigenic peptide NPY in the hypothalamus.

Peptides with Dual Localization in Endocrine Cells and Neurons – Ontogeny and Phylogeny

It has gradually been realized that the distinction between true neuropeptides(i.e., peptides restricted to neurons) and peptides with a dual localization in neurons and endocrine cells is not crystal clear. Some peptides may be exclusively expressed in neurons or endocrine cells in adults under normal conditions, but have a dual localization during ontogeny and in pathophysiological conditions.

Phylogenetic studies have complicated things even further. Thus, it seems that certain peptides expressed exclusively in endocrine cells in mammals are produced in neurons, or have a dual localization in nerves and endocrine cells in lower vertebrates. The opposite seems also to be true; a neuronally localized peptide in mammals may be produced by endocrine cells in submammalian vertebrates. Thus, for instance, VIP, a well known neuronally localized peptide in mammals, is present in a large population of endocrine cells in the distal intestine of cartilaginous fish.

Role as Principal Hormone or as Co-Expressed Peptide

One note of consideration is the actual status of a neuropeptide when it is expressed in endocrine cells. It may be the principal hormone, like secretin and CCK. More often, however, it is co-produced with the principal hormone. In this latter instance the coexisting peptides as a rule are co-stored within the same secretory granules, as revealed by, for example, immunogold staining, at the electron microscopic level.

Another note of consideration is the propeptide processing and the size of the resulting fragments. This has been quite extensively studied for pro-CCK, for example, and, in general it seems that the precursor processing is more advanced – that is, the final processed fragments are smaller – in neurons than in endocrine cells. Thus, the short fragment, CCK-8, predominates in neurons, whereas the longer fragments, CCK-58 and CCK-33, predominate in endocrine cells.

The chromogranins and related peptides form, in this context, a special group. The first of these acidic proteins, chromogranin A, was originally described in the secretory granules of endocrine cells in the adrenal medulla. Chromogranins were soon found to form a considerable proportion of the granule matrix of very many endocrine cells, and of many neurons as well. It was also demonstrated that chromogranins are subjected to proteolytic processing at specific cleavage sites, to yield fragments of different size, and that this processing mirrors that of hormone and neuropeptide precursor processing. Furthermore, it has turned out that many of the chromogranin fragments display biological activity, and that they,

being stored in the secretory granules, are released together with the main hormone or neurotransmitter upon exocytosis.

Pathophysiological Aspects

The islet beta cells have been studied quite extensively with respect to coexisting peptides and the plasticity of peptide expression during development of experimental type 2 diabetes, particularly in rat models. In the early phase of this process the beta cells accelerate insulin (and IAPP) expression in order to compensate for the increased need of hormone, to avoid hyperglycemia. At the same time the beta cells upregulate their CART and NPY expression severalfold, from very low, or no expression, to high levels. This is suggested to reflect, on one hand, an attempt of the cell to boost the rate of insulin secretion, since CART is capable of augmenting an already stimulated secretion. On the other hand, there is, at the same time, a need to prevent cell exhaustion, and eventual cell damage, by upregulating also inhibitors of insulin secretion, such as IAPP and NPY, to act as local breaks. It may be added in this context that the islet parasympathetic innervation, containing CART and VIP in addition to acetylcholine, becomes denser, as studied in rats, during diabetes development, thereby providing more insulin secretagogues to the beta cell environment.

Tumors arising from peptide hormone-producing cells in the gastro–entero–pancreatic region, the airways, or the thymus (adenomas if benign, or carcinoids if malignant) are on the whole infrequent, but may give rise to severe symptoms due to hormone overproduction, even if the tumor is benign. It is not uncommon that endocrine tumors are composed of several cell types, producing a number of different peptides (multihormonal tumors). Nevertheless, it is usually the predominating cell type that produces the hormone causing the main symptoms. Occasionally, the tumor releases counteracting hormones, or hormone release inhibitors, which may result in a delay of symptoms, and thereby a delay in diagnosis and in treatment as well. Further, it is not uncommon that metastases produce a different spectrum of peptides than the primary tumor. These general features also apply to endocrine tumors arising in the adrenal medulla (pheochromocytoma) or from the thyroid C cells (medullary thyroid carcinoma). Pheochromocytomas may host a great number of peptides, but are usually diagnosed on the basis of the overproduction of catecholamines.

Finally, one feature that is pertinent to mention is that endocrine tumors may occasionally produce peptides that are not normally expressed in endocrine cells in mammals, at least not in adulthood. Such peptides include VIP, which causes the watery diarrhea, hypokalemia, and anacidity (WDHA) syndrome, and growth hormone-releasing hormone (GHRH), which may give rise to acromegaly. Further, the tachykinin substance P, which in the human gut is expressed in but a small subpopulation of the serotonin-producing enterochromaffin (EC) cells in the ileum, is present in almost all ileal carcinoids, by far the most common intestinal endocrine tumor. The appearance of unexpected peptides in endocrine tumors may be related to a broader ontogenetic and phylogenetic messenger programming of certain peptide hormone-producing cells.

Concluding Remarks and Future Directions

Studies of peptides as hormones and neuropeptides have witnessed both ups and downs over the years. These types of peptides were the basis of the very start of endocrinology (secretin), and constitute by now a dominating category of hormonal and neuronal messengers. Their functions cover a wide spectrum; it has, in this context, become an intriguing fact that many hormonal peptides expressed in the gastro–entero–pancreatic region are engaged in the regulation of feeding, in addition to having other more specified functions (Table 6). However, since many of these peptides are expressed also in the brain, the exact origin of the orexigenic and anorexigenic signals is in some instances not yet resolved. When gene technology provided the possibility to knock out such peptides, or their receptor(s), one by one, the resulting generally meager phenotype, or total lack of it, represented one 'down.' Redundancy became a reality in this research area, and is one conceivable explanation. In retrospect, we have to admit that it is certainly a good strategy for survival not to be dependent on a single messenger to handle vital functions, such as feeding.

It is to be expected that, with techniques becoming more and more sophisticated and sensitive, more neuropeptides, including those expressed at low levels, will be detected in endocrine cells, and, conversely, more hormonal peptides will be detected in neurons. Furthermore, new peptides with a cell-to-cell messenger function are still regularly discovered (e.g., as ligands of orphan G-protein-coupled receptors). It is therefore to be expected that new peptides will continue to appear for quite some time in the future. To this end, there are still endocrine cells (e.g., along the gastrointestinal tract) that lack a peptide hormone, and a host of orphan receptors that need a ligand.

Table 6 Hormones and regulatory peptides in gastro–entero–pancreatic endocrine cells thought to be involved in feeding regulation

Peptide	*Sites of production*	*Actions*
Islet amyloid polypeptide/ amylin	Islet beta cells and delta cells, gastric endocrine cells	Inhibit insulin secretion, inhibit feeding
Cocaine- and amphetamine-regulated transcript peptides	Islet delta cells and antral G cells, many islet cell types neonatally(rat), islet beta cells in diabetes models	Regulate islet hormone secretion, inhibit feeding
Cholecystokinin	I cells, upper small intestine	Stimulate pancreatic enzyme secretion and gallbladder contraction, inhibit feeding
Enterostatin	Endocrine cells (mostly enterochromaffin cells) in upper gut	Inhibit feeding
Ghrelin	A-like or P/D1 cells in stomach, Mo cells in upper intestine and islet ghrelin cells	Stimulate feeding and gastric emptying, inhibit insulin secretion
Glucagon-like peptide-1	L cells (major population), distal small and large intestine	Stimulate insulin secretion glucose dependently, inhibit feeding
Motilin	Mo cells, upper small intestine	Stimulate interdigestive gut motility, stimulate feeding
Neurotensin	Neurotensin (subpopulation of L) cells, distal small intestine	Inhibit gut motility, acid secretion, and feeding
Neuropeptide Y	Islet beta cells during development, adrenal medulla	Inhibit insulin secretion, gut motility, and secretion; stimulate feeding
Oxyntomodulin	L cells, distal small and large intestine	Inhibit gastric acid secretion and feeding
Pancreatic peptide	Islet pancreatic peptide cells, gut (few)	Inhibit exocrine pancreatic secretion, inhibit feeding
Peptide YY (PYY3–36)	L cells, distal small intestine and large intestine, islet alpha and delta cells	Inhibit feeding, gut motility, exocrine pancreatic secretion
Somatostatin	D cells in islets and along the entire gut	Inhibit numerous gut and islet functions, inhibit feeding
Thyrotropin-releasing hormone	Islet beta cells during development	Stimulate gastric secretion and motility, inhibit feeding

See also: Calcitonin Gene-Related Peptide (CGRP) and Receptors; Neuropeptide Inactivation or Metabolism; Neuropeptide Release; Neuropeptide Y (NPY) and its Receptors; Neuropeptides Internalization; Neuropeptides and Coexistence; Neuropeptides: Food Intake.

Further Reading

Ahrén B, Wierup N, and Sundler F (2006) Neuropeptides and the regulation of islet function. *Diabetes* 55(supplement 2): S98–S107.

Buchan AMJ (1999) Structure and function of gastrointestinal endocrine cells. In: Greeley GH (ed.) *Gastrointestinal Endocrinology*, pp. 1–30. Totowa, NJ: Humana Press.

Chaudhri O, Small C, and Bloom S (2006) Gastrointestinal hormones regulating appetite. *Philosophical Transactions of the Royal Society London, Series B: Biological Sciences* 361(1471): 1187–1209.

Ekblad E, Håkanson R, and Sundler F (1991) Microanatomy and chemical coding of peptide-containing neurons in the digestive tract. In: Daniel EE (ed.) *Neuropeptide Function in the Gastrointestinal Tract*, pp. 132–179. Boca Raton, FL: CRC Press.

Håkanson R, Böttcher G, Ekblad E, et al. (1990) Functional implications of messenger coexpression in neurons and endocrine cells. In: Schwartz TW, Hilsted LM and Rehfeld JF (eds.) *Neuropeptides and Their Receptors, Alfred Benzon Symposium 29*, pp. 211–236. Copenhagen: Munksgaard.

Harmar AJ (2004) Receptors for gut peptides. Clinical endocrinology and metabolism. *Best Practice & Research Clinical Endocrinology & Metabolism* 18(4): 463–475.

Hökfelt T, Millhorn D, Seroogy K, et al. (1987) Coexistence of peptides with classical neurotransmitters. *Experientia* 43(7): 768–780.

Hökfelt T, Bartfai T, and Bloom F (2003) Neuropeptides: Opportunities for drug discovery. *Lancet Neurology* 2(8): 463–472.

Korbonits M, Goldstone AP, Gueorguiev M, et al. (2004) Ghrelin – a hormone with multiple functions. *Frontiers in Neuroendocrinology* 25(1): 27–68.

Sundler F and Böttcher G (1991) Islet innervation, with special reference to neuropeptides. In: Samols E (ed.) *The Endocrine Pancreas*, pp. 29–52. New York: Raven Press.

Sundler F, Ekblad E, Böttcher G, et al. (1985) Coexistence of peptides in the neuroendocrine system. In: Håkanson R and Thorell J (eds.) *Biogenetics of Neurohormonal Peptides*, pp. 213–243. London: Academic Press.

Sundler F, Böttcher G, Ekblad E, et al. (1993) PP, PYY, and NPY. In: Colmers WF and Wahlestedt C (eds.) *The Biology of Neuropeptide Y*, pp. 157–196. Totowa, NJ: Humana Press.

Taupenot L, Harper KL, and O'Connor DT (2003) The chromogranin-secretogranin family. *New England Journal of Medicine* 348(12): 1134–1149.

Wierup N and Sundler F (2006) CART is a novel islet regulatory peptide. *Peptides* 27(8): 2031–2036.

Neuropeptides: Enteric Nervous System

M Costa and S Brookes, Flinders University, Adelaide, SA, Australia

© 2009 Elsevier Ltd. All rights reserved.

Neuropeptides in the Enteric Nervous System

The enteric nervous system (ENS) comprises all neurons embedded within the digestive tract and their connections with extrinsic neurons originating in sympathetic, parasympathetic, and sensory ganglia. The ENS is involved in controlling movements of the digestive tube, in mucosal transport, in controlling the local blood supply and in modulating cellular defense mechanisms. The process of unraveling the neural circuits of the ENS has progressed significantly over the past few decades, with the major classes of enteric neurons being identified in some experimental animals. A relatively small number of functional classes have been characterized on the basis of their neurochemistry, projections, biophysical properties, and effects of their activation.

Enteric circuits mediating the different functions usually work in concert. For instance, neurally mediated changes in motility are usually associated with appropriate alterations of secretion and blood flow. Although extensive analysis of enteric neural circuits is well advanced, how the circuits give rise to integrated enteric motor, circulatory, and secretory behaviors is less well understood. Consequently, the role of the neuropeptides contained in individual classes of enteric neurons, or in specific circuits, remains largely uncertain. We present here a critical perspective of the role of neuropeptides in the neural circuits underlying the major functions controlled by the ENS.

Neuropeptides are relatively short sequences of amino acids, synthesized in the nerve cell body by cleavage from larger precursor peptides and transported to the nerve endings, where they are stored and released from vesicles. Neuropeptides cleaved from the same or similar precursors have been classified into families. All neuropeptides give rise to effects mediated via G-protein-coupled receptors (GPCRs). Neuropeptide receptors also form subgroups with a range of molecular forms and with varying affinities for the native (endogenous) or synthetic ligands.

The discovery of neuropeptides in enteric neurons was reported nearly 30 years ago. The list of peptides expressed in the enteric nervous system has grown in parallel with the burgeoning list of new peptides identified in other parts of the nervous system. There seems to be a fundamental agreement in the literature that nearly all classes of enteric neurons contain one or more neuropeptides. Despite the vast amount of information that has been accumulated on the nature and distribution of peptides in the ENS, relatively little progress has been made in achieving a real understanding of their neuronal role.

In the gastrointestinal mucosa there are many classes of enteroendocrine cells containing a wealth of peptides, similar to those contained in enteric neurons. This adds to the difficulty of investigating the actual role of neuropeptides in the enteric nervous system. Some of the peptides released from enteroendocrine cells play a hormonal role, but may also act as mediators for transducing luminal signals to intrinsic and extrinsic afferent neurons. Peptides are also present and released from other nonneuronal sources – for example, calcitonin gene-related peptide (CGRP) can be released by mast cells.

Since the 1970s, almost every class of biologically active peptides discovered in mammalian brain has also been found in enteric neurons. Indeed, some peptides were first discovered in the gut. At the time, it was expected that neuropeptides could be used to distinguish different functional classes of enteric neurons. This rapidly created an inflated number of potential functional classes. The discovery that several peptides could coexist in the same enteric neuron, and that neurons with similar shapes (axonal projections) contained unique combinations of peptides and other neurochemicals, led to the idea of 'chemical coding.' It was also proposed that neurons with different 'chemical coding' are likely to be functionally distinct. This concept opened a new era in investigations of the ENS. It has helped identify the major classes of enteric neurons, making it possible to place these classes within functional circuits and to identify the neurons in functional studies. Nearly 20 functional classes of enteric neurons have been identified using 'chemical coding' principles, including excitatory and inhibitory motor neurons, ascending and descending interneurons, secretomotor neurons, vasomotor neurons, and intestinofugal and intrinsic afferent neurons.

An important corollary of these studies was recognition that the same peptide may be present in very different classes of neurons. The concept of 'one enteric peptide, one function' had to be abandoned. The discovery that enteric neurons utilize multiple transmitter mechanisms also meant that the idea that a neuron can be named after one transmitter also had to be abandoned. Thus terms such as 'peptidergic,' 'VIPergic,' and 'tachykinergic,' often used to describe classes of enteric neurons, must be regarded

as inaccurate and inappropriate. First, the mere presence of a peptide in a neuron does not prove that it is used as transmitter. Second, even if this is the case, neurons usually contain and utilize several other transmitters. To avoid confusion, enteric neurons are mentioned here according to their functional classes (e.g., inhibitory motor neurons) and as 'peptide containing' – for example, vasoactive intestinal peptide (VIP)-containing inhibitory motor neurons. The classes of enteric neurons and their circuitries should be regarded as scaffolds on which different messenger substances play their relative roles as modulators or cotransmitters.

Another implication of the chemical coding concept is that each class of enteric neurons may utilize multiple transmitter substances. This led to the concept of plurichemical transmission, whereby individual classes of enteric neurons use different cotransmitters affecting different mechanisms of recipient cells, over a range of time courses. Detailed knowledge of chemical coding of enteric neurons has enabled more direct and meaningful comparisons of enteric circuitries in different species.

Identification, Quantification, Localization, and Biological Actions of Neuropeptides

The discovery of the presence of a neuropeptide in the ENS usually involved extraction and purification, testing for biological activity, followed by sequencing of the peptide and its eventual localization by immunohistochemistry. For example, the original extraction of the tachykinin, substance P, from horse intestine preceded by decades its sequencing in central tissue. It led to the identification of different members of the tachykinin family in mammalian gut, with detailed localization in different classes of enteric neurons. More recently, new peptides and their receptors are being increasingly discovered through bioinformatics and by molecular biological methods.

The main technique for localizing neuropeptides over the past three decades has been by histological staining using antibodies raised against a part of or the whole neuropeptide sequence or its precursor. Although immunohistochemical techniques are relatively simple, the interpretation of the results need some attention. For example, only if tissue labeling is prevented by inactivating the antibody with its synthetic antigen ('preabsorption') can the localization be said to show 'peptide-like' immunoreactivity. Further identification of the actual molecular form requires additional analytical techniques, including sequencing and Western blots.

The effort to study localization of neuropeptides in the gut has blossomed since the early studies of immunohistochemistry in the 1970s. Somatostatin was the first neuropeptide visualized in enteric neurons, in 1977 (Figure 1). Since then, methods to visualize entire enteric circuits in whole-mount preparations has permitted characterization of the shapes, projections, and proportions of enteric neurons containing particular peptides. In 1979, Tomas Hokfelt and colleagues published the first evidence of co-co-localization of peptides in enteric neurons. Since then, a large number of reports have revealed the presence of new peptides in enteric neural tissue, providing a large and sometimes contradictory database on the diversity of enteric neurons. With a patient systematic study of the multiple localization of peptides and other markers, much of

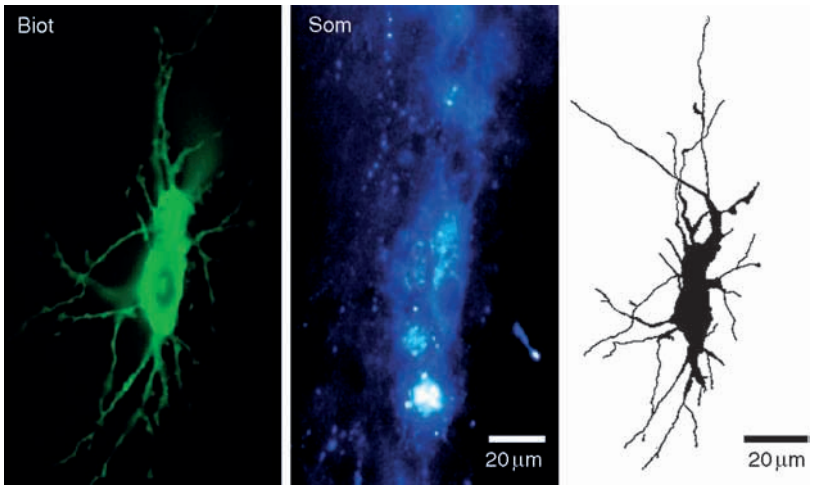

Figure 1 Descending cholinergic interneuron in the myenteric plexus of the guinea pig small intestine was filled with biotinamide (green) after intracellular recording. The same neuron was immunoreactive for the peptide, somatostatin (blue), and shows distinctive, filamentous dendrites in the camera lucida drawing on the right.

this information was crystallized into a comprehensive and simplified account for the enteric circuits in the guinea pig small intestine, by the mid-1990s. It is fair to say similar attempts to identify all classes of enteric neurons have been attempted in only a few other regions of gut or species. The literature is still surprisingly full of isolated and unrelated observations of single immunohistochemical labeling of one or other neuropeptide, which cannot yet be put into a meaningful functional context.

Establishing Physiological Functions of Enteric Neuropeptides

Investigations on the role of neuropeptides in enteric neurons is part of the general problem of demonstrating a transmitter role for any proposed substance. A classic list of prerequisites that a putative transmitter must fulfill includes transmitter synthesis, storage, and release, and a demonstration of an effect mediated via specific receptors that mimics a biological mechanism. However, too often publications merely demonstrate the presence of a peptide in the gut by histological means, present evidence that the exogenous peptide (or an agonist) produces an effect, and from this incomplete evidence conclude too eagerly that the peptide is an enteric transmitter.

A critical piece of evidence for identifying a transmitter role is to show that interfering with the transmission mechanisms of the endogenous peptide results in some change of function, usually through the use of a specific and selective antagonist. This has only been achieved in a very few instances. Partly this is due to the relative lack of effective antagonists to peptide receptors. Even when good antagonists are available, as for the tachykinin receptors, the multiplicity of receptors involved demands the use of cocktails of antagonists. For example, it was less than a decade ago that the role of tachykinins in motor behavior of the intestine (peristalsis in the distal colon of the guinea pig) was established. This was long after antagonists for the three classes of tachykinin receptor (NK-1, NK-2, and NK-3) became available. Marcello Tonini and co-workers used simultaneously antagonists to all three tachykinin receptors, to demonstrate that the propulsive contractions of the guinea pig colon are mediated by tachykinins, probably released from excitatory motor neurons, which are also cholinergic. The nature of inhibitory transmission from enteric inhibitory motor neurons has been subject of extensive research and long debate. The original descriptive term 'nonadrenergic noncholinergic' (NANC) gave way to the more specific 'purinergic,' 'VIPergic,' and 'nitrergic' to describe ATP, VIP, or nitric oxide (NO) as the transmitter substance, respectively. It is now well accepted that most inhibitory motor neurons contain and utilize all three classes of substances (see later), with variations in their relative contributions.

The localization of peptide receptors has assisted the understanding of the role of neuropeptides in the gastrointestinal tract. By and large, receptors were localized where they were expected. The visualization of opioid receptors on some enteric neurons by immunohistochemistry confirmed that this was the main site at which opiates inhibited intestinal motility (Figure 2). Before immunohistochemical localization of receptor proteins, the effects of agonists comprised the main way to establish the potential action if the endogenous peptides were to be released. For example, endogenous tachykinins act on enteric neurons via NK-3 receptors. Senktide, a specific agonist, evokes contraction of intestinal muscle solely by an action on enteric neurons. The use of selective

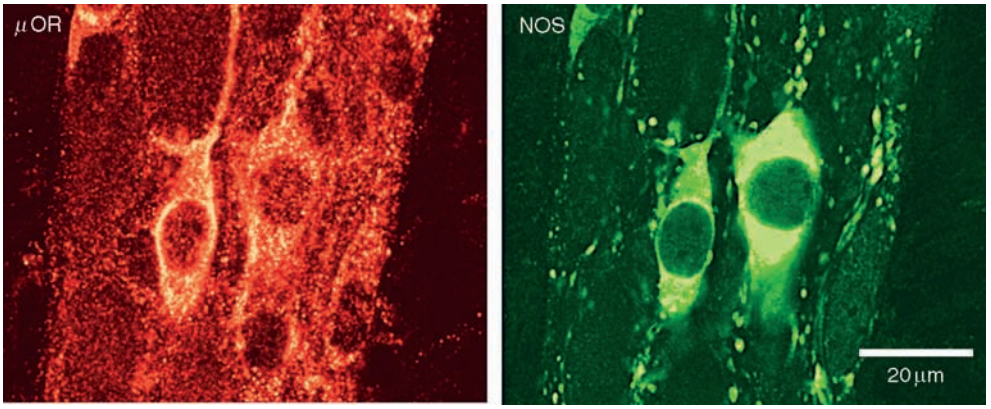

Figure 2 Single confocal section through a myenteric ganglion of guinea pig ileum shows several neurons with immunoreactivity for the mu opioid receptor (μOR; red), which is concentrated at the cell membrane, with some granular background labeling throughout the ganglion. Two of the neurons are also immunoreactive for nitric oxide synthase (NOS; green), a marker for inhibitory interneurons and descending interneurons.

antagonists to NK-3 receptors confirmed that endogenous tachykinins, probably released by enteric primary afferent enteric neurons, act on other neurons via slow synaptic mechanisms involving NK-3 receptors.

When receptors are found at cellular sites not predicted by previous pharmacological studies, new sites of action may be suggested. Fluorescently labeled agonists have proved useful to follow the dynamics of peptide receptors during their activation. Recycling of opioid and tachykinin receptors in functional studies has provided evidence of the release of the endogenous peptides. For example, a fluorescently labeled NK-3 agonist supported previous evidence that endogenous tachykinins mediate slow synaptic transmission between enteric primary afferent neurons, interneurons, and motor neurons.

Targeted mutations in the gene encoding a peptide or its receptor (yielding 'knockout' animals) have also been useful, although compensatory mechanisms during development may complicate interpretation of the results. There are, however, relatively few experiments of this type that have shed light on the enteric peptides. Mice lacking the VIP gene (−/−) show significant alterations in intestinal morphology and structure as seen by light microscopy. The length of the bowel is decreased compared to wild-type (WT) controls and the smooth muscle layers are thicker, with impaired intestinal transit. A significant percentage of VIP(−/−) mice die in the first postnatal year, with overt stenosis of the gut.

Another way to investigate transmitter function, which has been useful for many nonpeptide transmitters, is to experimentally alter the endogenous mechanisms of inactivation. However, this has been rather less successful for neuropeptides. For example. the enzyme aminopeptidase A is known to inactivate cholecystokinin (CCK-8) and selective inhibitors are available, but other peptides are also substrates of the same enzyme. The neutral endopeptidase inhibitor, thiorphan, has been shown to enhance and prolong inhibitory transmission to the smooth muscle, and this has been taken as indication for inhibitory peptides involvement.

Functional Role of Peptides in Enteric Circuits

The restriction of neuropeptides to specific functional populations of enteric neurons has made it easier to identify their roles in transmission. A few families of neuropeptides appear to play major roles, but the roles of most other peptides still remain to be established.

Tachykinins are a family of peptides; members include substance P (SP), neurokinin-A (NK-A), and neuropeptide-γ (which lacks residues 3–17 of neuropeptide K). They are cleaved from preprotachykinin A (PPT-A), which occurs in several forms (preprotachykinin mRNA species, α, β, and γ). Preprotachykinin B (PPT-B) encodes neurokinin-B (NK-B) and peptide 2 (P2), a nontachykinin peptide sequence. The discovery of a third preprotachykinin gene (*TAC4*) has more than doubled the number of tachykinins through the discovery of several new species-divergent peptides. These include hemokinin-1 (HK-1) in mouse and rat, endokinin-1 (EK-1) in rabbit, and EK-A, EK-B, human HK-1 (hHK-1), and hHK(4–11) in humans. Each exhibits a remarkable selectivity and potency for the NK-1 receptor, similar to SP. These recently cloned mammalian tachykinins are primarily expressed in nonneuronal cells. The biological actions of all of the tachykinins are mediated through three types G-protein-coupled receptors denoted NK(1), NK(2), and NK(3).

Tachykinins are present in excitatory motor neurons both to the circular muscle and longitudinal muscle layers. They are released with the major transmitter, acetylcholine, and act on muscle cells via NK-2 receptors and on interstitial cells of Cajal via NK-1 receptors. Tachykinins appear to be released by higher frequency firing of motor neurons, compared to acetylcholine, and elicit more prolonged responses. Tachykinins are also contained in the enteric primary afferent neurons and are released to provide the main slow synaptic transmission to neighboring primary afferent neurons, interneurons, and motor neurons, mostly via NK-3 receptors. Tachykinins are also co-transmitters with acetylcholine in secretomotor neurons.

In 1970, SI Said and V Mutt isolated a novel peptide from porcine intestinal extracts; the peptide, which had powerful vasoactive properties, was named vasoactive intestinal peptide, and it is now known to belong to the secretin/glucagon family, along with peptide histidine isoleucine (PHI), peptide histidine valine (PHV), and pituitary adenylate cyclase-activating polypeptide (PACAP). These are all cleaved from different precursors from two genes. Behavioral and neurological tests on transgenic mice reveal some functional distinctions between the two genes. These findings suggest that PACAP and VIP evolved to possess distinct biological activities and that they do not compensate each other effectively when one is absent in a knockout mouse. Molecular cloning of PACAP receptors has shown three distinct receptor subtypes: the PACAP-specific PAC1 receptor, which is coupled to several transduction systems, and the two PACAP/VIP-indifferent VPAC1 and VPAC2 receptors, which are primarily coupled to adenylyl cyclase.

In addition to VIP (**Figure 3**), enteric inhibitory motor neurons use PACAP as a transmitter to relax intestinal smooth muscle by eliciting slow hyperpolarization. They act via receptors located on both smooth muscle cells and interstitial cells of Cajal;

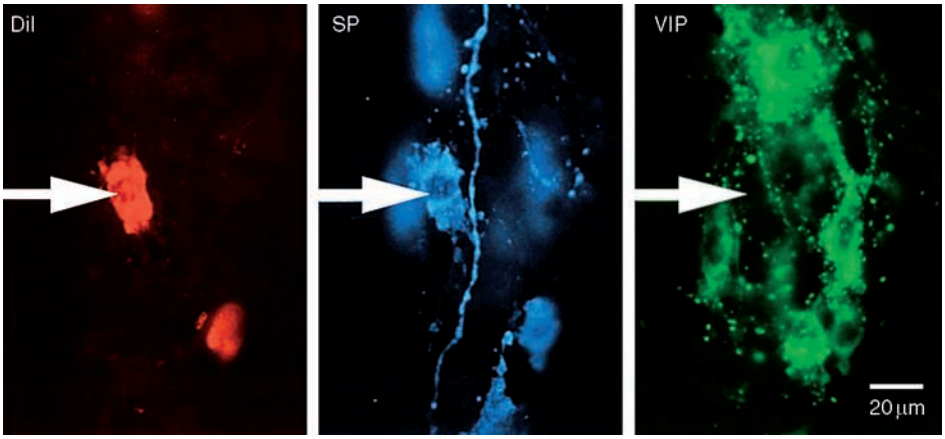

Figure 3 Motor neurons to the circular muscle of the guinea pig ileum were retrogradely labeled with the carbocyanine dye, DiI (red). One motor neuron close to the DiI application site was immunoreactive for substance P (SP) but not for vasoactive intestinal peptide (VIP). The other motor neuron (lower red profile, slightly out of focus) was one of a very sparse population of enteric neurons, immunoreactive for both SP and VIP.

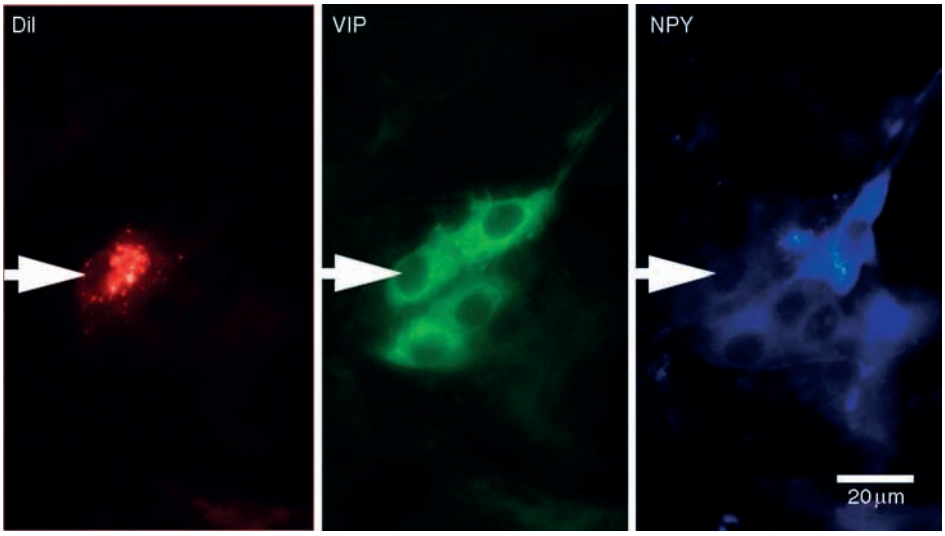

Figure 4 Neurons in a submucous ganglion of guinea pig ileum, showing typical chemical coding. A single neuron was retrogradely labeled by the carbocyanine dye, DiI, applied to a myenteric ganglion (red). The same neuron, probably a noncholinergic secretomotor neuron, was immunoreactive for the peptide vasoactive intestinal polypeptide (VIP; green) but not for neuropeptide Y (NPY; blue). In fact, there is no coexistence of VIP and NPY; these peptides are found in separate populations in submucous neurons of the small intestine in this species.

a receptor antagonist for both VIP and PACAP reduces inhibitory transmission. The same inhibitory motor neurons also utilize nitric oxide as transmitter, relaxing muscle cells via the guanylate cyclase/3′,5′-cyclic guanosine monophosphate pathway. Nitric oxide and vasoactive intestinal polypeptide also interact; NO enhances the release of VIP from nerve terminals and vice versa. ATP co-released from these same neurons also inhibits smooth muscle over a shorter time course. In some parts of the gastrointestinal tract, there is evidence that other inhibitory peptides may contribute to inhibitory transmission, in some species. For example, a neurotensin receptor antagonist has been shown to reduce inhibitory transmission in jejunal and ileal segments of rats.

VIP is also a major transmitter of secretomotor neurons (Figure 4), where it may also interact with NO; each potentiates the other's effect in stimulating intestinal fluid secretion. The role of VIP in some descending interneurons is less clear; it may partially mediate slow

synaptic transmission from these cells. Similarly, VIP contained in intestinofugal neurons may contribute to slow transmission in prevertebral sympathetic ganglia.

The discovery of the presence of opioid peptides in enteric neurons led to extensive studies of their roles. These peptides include leu-enkephalin (leu-ENK) and met-enkephalin (met-ENK), cleaved from preproenkephalin; dynorphin (DYN), cleaved from predynorphin; and the recently discovered heptadecapeptide, nociceptin, also known as orphanin FQ, which is the endogenous ligand for the 'orphan' opioid receptor-like $(ORL)_1$ receptor. They all inhibit enteric neurons directly, with little or no effect on smooth muscle or secretory cells. Opioid peptides are present in enteric excitatory motor neurons, where they are co-released with acetylcholine and tachykinins. Under physiological conditions, they appear to exert little effect. However, when neural activity is intense, such as during peristaltic work against an obstruction, the opioids play a braking role by inhibiting the neurons from which they are released. Opioid peptides are also present in inhibitory motor neurons, where their role is less certain. Nevertheless, as the nerve endings of inhibitory and excitatory neurons run in the same bundles, opioids released from inhibitory neurons may inhibit the release of excitatory transmitters from excitatory motor neurons. The role of opioid peptides in interneurons is likely to be similar. This suggests a widespread phenomenon in which transmission involves autoregulatory mechanisms that limit excitability during sustained or excessive activation. The action of opioid peptides is mediated by all three known opioid receptors: the mu, kappa, and delta receptors. Both tolerance to and dependence on opiates have been demonstrated in enteric neurons, leading to significant interest by neuropharmacologists on the cellular bases of drug addiction.

Neuropeptide Y (NPY) and galanin are also contained in some interneurons and inhibitory motor neurons, where they probably act as 'brakes,' similar to opioid peptides, rather than as transmitters in their own right. NPY and peptide YY (PYY) are structurally related peptides that mediate inhibitory actions on gastrointestinal motility, secretion, and blood flow. Several receptor subtypes have been identified and the Y1, Y2, Y4/PP1, Y5, and Y5/PP2/Y2b receptors have been cloned. The selective Y1 receptor antagonist BIBP 3226 augments ascending reflex contractions, suggesting that endogenous NPY is co-released to autoregulate excitatory transmission. Galanin's effects are mediated by three distinct G-protein-coupled receptors, (galanin receptor 1 (GalR1), GalR2, and GalR3), which activate different intracellular signaling pathways. All three Gal receptor mRNAs have been detected in the gut. Galanin inhibits transmitter release from myenteric neurons through inhibition of voltage-dependent Ca^{2+} channels via a $G_{(i/o)}$-coupled GalR. Evidence for a physiological role of Gal is still lacking, as receptor antagonists do not affect peristalsis either in normal conditions or when efficiency is reduced by partial nicotinic transmission blockade with hexamethonium.

Similar to NPY and galanin, somatostatin is likely to act by inhibiting neural activity either presynaptically or postsynaptically. It has been localized in some descending cholinergic neurons as well as in secretomotor pathways. An excellent example of the functional relevance of chemical coding is the differential expression of NPY in postganglionic sympathetic vasoconstrictor neurons to the intestine. Somatostatin is present in the sympathetic neurons that inhibit secretion, through actions on submucosal secretomotor neurons. In contrast, the sympathetic neurons that inhibit motility by acting on myenteric neurons contain neither NPY nor somatostatin. These three classes of sympathetic neurons are activated independently by physiological stimuli.

Although CGRP is widely distributed in several classes of enteric neurons, its role in these neurons is unclear. However, CGRP is also a sensory peptide that is released, together with tachykinins, from the collaterals of extrinsic spinal sensory neurons, and its actions on blood vessels, on muscle, and on enteric neurons are well established. Other peptides present in enteric interneurons and motor neurons have been shown to be able to excite other enteric neurons. These include CCK, gastrin-releasing peptide (GRP), corticotropin-releasing factor (CRF), neurotensin, neuromedins, ghrelin, endothelins, and atrial natriuretic factor. CCK antagonists reduce slow excitatory synaptic potentials in myenteric neurons, suggesting that neurally released CCK is involved in synaptic transmission in some enteric pathways. However, its role in physiological conditions remains to be established. GRP is a 27-amino-acid peptide and is closely related to bombesin, sharing all but one of the last ten amino acids at the C-terminal end. Endogenous GRP is released during motor activity, and antagonists to the bombesin receptor (BB_2) inhibit propulsion of artificial fecal pellets through isolated segments of guinea pig distal colon.

None of the other peptides have been shown directly to mediate synaptic transmission and thus their role remains obscure. The more recently discovered peptide, cocaine- and amphetamine-regulated transcript peptide (CART), has been shown to be present in some enteric motor neurons. However, this peptide has not been shown to have a detectable role, apart from a possible modulatory effect on NO transmission. Similarly, the more recently discovered peptides, ghrelin and orexins (orexins

A and B), are present in some enteric neurons and are capable of stimulating motor neurons, interneurons, and secretomotor neurons, but their role remains to be determined. A hint that orexin may serve as an excitatory transmitter in ascending excitatory reflexes comes from a study in which antagonists to the orexin receptors reduced ascending reflex contractions in the mouse small intestine.

Changes in Peptide Levels and Their Receptors

Measurements of neuropeptide transmitters and their receptors have accompanied studies of their localization. Radioimmunoassay (RIA) and enzyme-linked immunoassy (ELISA) tests have been widely used to quantify peptides in the intestine, largely replacing the quantitative bioassays used in earlier times. Immunoblots have been used to determine protein levels in tissues, and expression of messenger RNA indicates the potential production of neuropeptides or their receptors. Messenger RNA can be used for cellular localization, when detected with *in situ* hybridization. Messenger RNA for neuropeptides and their precursors and receptors can be assessed by real-time quantitative reverse transcription–polymerase chain reaction (qRT-PCR).

Using such methods (often with immunohistochemical methods), the changes in tissue contents of precursors, peptides, and their receptors have been explored in a range of physiopathological conditions. Changes in gene expression of neuropeptide precursors or receptors are taken as indications of up- or downregulation – as either adaptive or maladaptive responses to the disease state. A major problem with such studies is that it is not clear if changes are a primary cause of the correlated symptom or if they are an effect, or whether they are the result of plastic changes of tissues. For example, inflammation of the human ileum results in changes in substance P and NK-1 expression, even at distant sites in the body. Whether this contributes to the symptoms or is an adaptive response to minimize damage is not entirely clear. What is becoming apparent is that the chemical coding of enteric neurons can be readily modified in physiopathological conditions, with neuropeptides likely to be primarily or secondarily involved in plastic adaptive changes.

New Advances

The effects of neuropeptides on less well-studied functions, such as modulation of gene expression, cell proliferation, cell migration, differentiation, and (generically) morphogenesis and maintenance of tissue morphology, are all likely to become more important in future studies. Neuropeptides are also likely to play roles in wound healing, antiapoptosis, and responses to injury and inflammation, and, in ontogenetic functions, on proliferation/differentiation from early stages.

In disease states, such as inflammatory bowel disease, inhibition of myocyte proliferation by VIP and PACAP may serve to control smooth muscle hyperplasia. CART may be neuroprotective, playing a role in the survival and maintenance of enteric neurons. Undoubtedly, systematic analysis of plasticity in neuropeptides and their receptors will be enhanced by the application of microarray techniques and targeted molecular approaches. The study of transgenic animals and knockdown techniques, using RNA interference or synthetic oligonucleotides, also has enormous potential to clarify the roles of neuropeptides in gastrointestinal function under normal or abnormal conditions. We can also foresee the discovery of new peptides, possibly new families, and certainly the identification of new receptors. A large number of 'orphan' G-protein-coupled receptors exist: these comprise a large class of receptors with unknown cognate ligands. Finding the known neuropeptides that may act on them, or identifying novel neuropeptides that fulfill this role, will be a productive endeavor for years to come. There is no shortage of intestinal functions and malfunctions that may involve subtle roles of old and new enteric neuropeptides. The challenge will remain how to match the increasing pharmacopoeia of peptides expressed by a finite population of enteric neurons with their range of highly subtle roles and functions.

The wealth of neuropeptides in enteric neurons is still puzzling, three decades after they were first described. For most of them, there is currently little indication that they play much more than a minor, secondary role. Yet evolutionary theory tells us that the preservation of such molecular mechanisms is likely to indicate that they play a significant role in the organism's life. The experimental conditions under which they have been studied so far may not simulate adequately the complex range of environmental conditions in which specific peptides are likely to play their important evolutionary role. It remains for the investigators to use their inventiveness to identify the physiological conditions under which this can be demonstrated.

See also: Enteric Nervous System: Sensory Pathways; Enteric Nervous System: Neural Circuits and Chemical Coding; Enteric Nervous System Development; Enteric Nervous System: Disorders; Enteric Nervous System: Neurotrophic Factors; Enteric Nervous System: Glial

Cells and Interstitial Cells of Cajal; Enteric Nervous System: Physiology; Invertebrate Neurohormone GPCRs; Neuropeptide Y (NPY) and its Receptors; Neuropeptides: Discovery; Neuropeptides: Food Intake; Vasoactive Intestinal Peptide and Pituitary Adenylate Cyclase Activating Peptide Receptors.

Further Reading

Bornstein JC, Costa M, and Grider JR (2004) Enteric motor and interneuronal circuits controlling motility. *Neurogastroenterology & Motility* 1: 34–38.

Bornstein JC and Furness JB (1988) Correlated electrophysiological and histochemical studies of submucous neurons and their contribution to understanding enteric neural circuits. *Journal of the Autonomic Nervous System* 25(1): 1–13.

Brookes SJH (2001) Retrograde tracing of enteric neuronal pathways [Review]. *Neurogastroenterology & Motility* 13(1): 1–18.

Brookes SJH and Costa M (2004) Functional histoanatomy of the enteric nervous system. In: Johnson LR (ed.) *Physiology of the Gastrointestinal Tract*, 4th edn., Amsterdam: Elsevier.

Brookes SJH and Costa M (2002) Cellular organisation of the mammalian enteric nervous system. In: Brookes SJH and Costa M (eds.) *Innervation of the Gastrointestinal Tract: The Autonomic Nervous System*, pp. 393–467. London, New York: Taylor and Francis.

Costa M and Furness J (1982) Neuronal peptides in the intestine. *British Medical Bulletin* 38: 247–252.

Costa M, Furness JB, and Gibbins IL (1986) Chemical coding of enteric neurons. *Progress in Brain Research* 68: 217–219.

Costa M, Brookes SJ, Steele PA, et al. (1996) Neurochemical classification of myenteric neurons in the guinea-pig ileum. *Neuroscience* 75(3): 949–967.

Hokfelt T, Johansson O, Ljungdhal A, et al. (1980) Peptidergic neurones. *Nature* 284: 515–521.

Holzer P and Holzer-Petsche U (1997) Tachykinins in the gut. Part II. Roles in neural excitation, secretion and inflammation. *Pharmacology & Therapeutics* 73(3): 219–263.

Holzer P and Holzer-Petsche U (1997) Tachykinins in the gut. Part I. Expression, release and motor function. *Pharmacology & Therapeutics* 73(3): 173–217.

Minnis JG, Patierno S, Kohlmeier SE, et al. (2003) Ligand-induced mu opioid receptor endocytosis and recycling in enteric neurons. *Neuroscience* 119(1): 33–42.

Neuropeptides: Epilepsy

A Mazarati, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Dale's Principle versus the Concept of Coexistence: Implications for Epilepsy

The pathophysiology of epilepsy is typically viewed as the shift in the balance between the inhibitory (γ-aminobutyric acid (GABA)) and the excitatory (glutamate) neurotransmission, in favor of the latter. This shift occurs due to both selective loss of inhibitory GABA-ergic neurons after precipitating epileptogenic insults (e.g., status epilepticus, stroke, and traumatic brain injury) and the reorganization of neuronal circuits that favor hypersynchrony of neuronal populations (e.g., aberrant connections formed by the axons of dentate granule cells of the dentate gyrus, known as mossy fiber sprouting). Deficit in GABA-mediated signaling and augmentation of glutamatergic transmission that have been documented in many (although not all) types of epilepsy represent the basis for the pharmacotherapy of the disease.

The glutamate–GABA hypothesis of epilepsy is based on Dale's principle, the assumption that the neuropharmacological profile of every single neuron is defined by certain single neurotransmitter. However, Dale's principle has been proven wrong, so it has become obvious that the neuropharmacological basis of epilepsy is far more complex and multifaceted.

In contrast to Dale's principle, the concept of coexistence postulates that individual neurons produce, store, and release more than one messenger molecule. Indeed, a diverse group of biologically active substances has been shown to coexist with classic neurotransmitters in the same neuron and to regulate many aspects of neuronal functioning. Originally, the functional role of these substances was not clear; hence, in order to emphasize their putatively secondary, supporting role they were named neuromodulators. Today, however, the difference between the two classes of neuronal messengers is merely semantic; it has been well established that many neuromodulators are as important in shaping neuronal activity and in defining the biochemical profile of neurons and neuronal networks as classic neurotransmitters and, in fact, often play primary, rather than secondary, role in the pathophysiology of brain diseases.

Neuromodulator neuropeptides constitute a large and diverse group. Along with classic neurotransmitters, neuropeptides have been implicated in the mechanisms of many brain disorders, including epilepsy.

The basic properties of neuropeptides make them particularly important players in the mechanisms of epilepsy. As established during early stages of peptide research, neuropeptide release from neurons is activity dependent; particularly, neuropeptides, in contrast to classic transmitters, are released only at high frequencies of neuronal firing. Because on single-neuron-level epilepsy is characterized by an abnormally high firing rate, it is conceivable that neuropeptide release occurs in epileptic, rather than in normal, neurons. This assumption, if true, has several important ramifications. Seizure-induced increase in neuropeptide release might contribute both to the further progression and to the arrest of ongoing seizure activity; that is, it may represent either a maladaptive or an adaptive component of epileptic process, depending on the peptide involved. This, in turn, suggests that neuropeptide receptors may represent targets for antiepileptic therapy in addition to or in lieu of targeting classic neurotransmitters.

Temporal Lobe Epilepsy

Epilepsy is not a single disease but rather is a diverse group of diseases, each of which has distinct etiology, pathophysiology, and clinical characteristics. Temporal lobe epilepsy (TLE), which is primarily associated with hippocampal pathology, is one of the most commonly occurring forms of the disease. TLE bears several features that make it a particularly important object of epilepsy research. TLE develops after a variety of precipitating insults, such as status epilepticus, traumatic brain injury, infection, prolonged febrile seizures in the childhood, and stroke; it is not linked to any definite genetic trait and thus might affect practically anyone. Most important, TLE is frequently refractory, that is, resistant to currently available antiepileptic drugs; drug resistance represents a major challenge in managing the disease in at least one-third of patients.

Along with classic neurotransmitters, the hippocampus, which represents the neuroanatomical and neurophysiological substrates of TLE, expresses multiple peptides and peptide receptors. To date, a significant body of evidence has been accumulated to prove that peptidergic modulation represents both an important mechanism of TLE and a promising target for the development of novel antiepileptic drugs. Although both clinical and experimental studies have implicated neuropeptides in the mechanisms of

several forms of epilepsy, this article focuses on peptidergic modulation of TLE. On the one hand, basic principles of the regulation of limbic seizures by neuropeptides are applicable to other seizure types; on the other hand, the hippocampus, due to its anatomical organization, neurochemical profile, and a variety of neurodegenerative and neuronal plastic changes that mark epileptic process, represents a convenient model for examining the role of neuropeptides in the pathophysiology of the disease.

Most of the data on the role of neuropeptides in TLE comes from experimental studies that employed animal models. Among experimental models of TLE (also referred to as limbic epilepsy), two have been most widely used. Poststatus epilepticus chronic epilepsy closely resembles the evolution of human TLE, starting with precipitating insult (status epilepticus induced by chemical convulsants or electrical stimulation of the brain); followed by the silent period, which is characterized by neuronal degeneration and synaptic reorganization; and finishing with epilepsy *per se*, that is the occurrence of spontaneous recurrent seizures. Kindling is another model of limbic epilepsy which is characterized by the progressive development of seizures in response to repeatedly applied initially subconvulsive epileptogenic stimulus (e.g., electrical stimulation of the brain). In contrast to status epilepticus, neuronal degeneration is not a major factor in kindling process; instead, the chronic remodeling of neuronal circuits plays a primary role. The two models therefore allow us to focus on different mechanisms that underlie epileptic process, with an emphasis on neurodegeneration and neuronal plasticity, respectively.

Several Lines of Evidence Implicate Neuropeptides in the Mechanisms of Epilepsy

Expression of Peptides and Their Receptors Are Altered in Epileptic Brain

Immunohistochemistry, Western blot, *in situ* hybridization, polymerase chain reaction, and ligand binding assays revealed both the increase and the reduction of neuropeptide and/or peptide receptor expression in epileptic hippocampus. Increased expression was reported for enkephalins, orphanin FQ (nociceptin), neuropeptide Y, substance P, and neurokinin B; the expression of dynorphin and galanin was reduced. Changes in somatostatin depended on the epilepsy model used. Thus, somatostatin expression was elevated in the hippocampus of kindled animals, whereas status epilepticus-induced epilepsy led to the loss of somatostatin-containing neurons (which is also the case in human TLE). In addition to a mere increase or reduction of peptide expression, seizures can lead to qualitative changes; for example, status epilepticus induced *de novo* expression of neuropeptide Y in the inhibitory interneurons of the hippocampus and of substance P in all hippocampal subfields, that is, in neuronal populations that normally contain neither of these peptides.

Neuropeptide receptors also undergo plastic changes in epileptic brain. For example, status epilepticus increased the expression of μ-opioid receptors and Y2 receptors of neuropeptide Y; it also reduced the expression of orphanin FQ receptors, galanin type 2 receptors, Y1 receptors of neuropeptide Y, and somatostatin receptors sst2, 3, and 4.

The observed changes in peptide and peptide receptor expression are difficult to interpret on their own. Do they represent merely surrogate markers of epilepsy, or do they play certain role in epileptic process? If the latter is true, do these changes promote or counteract seizures?

Exogenously Administered Neuropeptides Modify Seizures

The answers to these questions come, in part, from experiments that examined how the exogenous administration of neuropeptide receptor ligands (both agonists and antagonists) modify seizures.

Anticonvulsant activity, both in *in vivo* (status epilepticus, kindling) and *in vitro* models of epilepsy, was shown for agonists of κ-opioid receptors (dynorphin), galanin type 1 and 2 receptors (galanin (1–29), galanin(2–11), and M617), orphanin FQ, neuropeptide Y receptors Y2 ([ahx5–24]NPY) and Y5 (Ala-Aib-NPY), somatostatin sst2 (octreotide and seglitide), antagonists of neurokinin 1 receptor (spantide II and RP 67,580), and the blocker of receptor Y1 BIBP3226. At the same time, κ-antagonist norbinaltorphimine and galanin antagonists chimeric peptides M35 and M40 facilitated seizures. Further, agonists of SST4 (L-803087) μ-(enkephalins) and neurokinin 3 (neurokinin B) receptors exerted proconvulsant effects. The agonist of neurokinin 1 receptors, substance P, exacerbated status epilepticus *in vivo*, but was shown to have anticonvulsant effect in an *in vitro* seizure model.

Although the exogenous administration of neuropeptides can provide certain answers about their role in seizures, such an approach has its drawbacks. Particularly, it is not known what final concentration of the peptide reaches the receptor; the amounts of peptides used are high compared to their endogenous concentrations and, thus, may not relevantly reproduce their physiological effects. At the same time,

saturating amounts of peptide receptor antagonists can hardly be achieved. Progress in molecular biology and the related introduction of transgenic animals into the routine of research laboratories has helped to overcome some of these and other possible pitfalls.

Animals with Altered Peptide/Peptide Receptor Expression Have Altered Seizure Phenotypes

Enhanced susceptibility to and increased severity of seizure responses have been reported for neuropeptide Y knockout mice and for the animals lacking the neuropeptideY receptor Y5. Mice lacking somatostatin exhibited exacerbated seizure responses; however, sst2 receptor knockout mice showed enhanced seizure resistance. The elimination of the preprotachykinin A gene, which encodes both substance P and neurokinin A, mitigated both status epilepticus-associated seizures and hippocampal neuronal injury. In contrast, both seizures (status epilepticus and kindling) and neuronal injury (status epilepticus) were exacerbated in galanin knockout and in galanin receptor type 1 and 2 knockout mice; at the same time, overexpression of galanin under both β-hydroxylase and platelet-derived growth factor B promoters decreased the severity of kindled- and status epilepticus-induced convulsions. Orphanin FQ knockout animals showed increased resistance to seizures, despite the fact that exogenously administered orphanin FQ was anticonvulsant.

The use of animals with peptide or peptide receptor mutations has its limitations as well. For example, the studies employ mice, but the patterns of human TLE are better reproduced in rats; the approach cannot be applied to all peptides because functional deletion of some of them might be lethal. Furthermore, a high degree of neuronal plasticity might result in the compensatory changes (e.g., overexpression of receptor sst2 in somatostatin knockout mice), which in turn may mask the effects of the peptide or peptide receptor in question.

Other Evidence

Other approaches have been employed to study the role of neuropeptides in epilepsy. For example, chronic adeno-associated virus vector overexpression of both neuropeptide Y and galanin in the hippocampus inhibited limbic seizures. Knockdown of galanin receptors 1 and 2 using peptide nucleic acid antisense targeting respective RNA exacerbated status epilepticus-associated seizures and neuronal injury.

Clinical Data

Studies in TLE patients documented polymorphism of prepro-dynorphin gene; hippocampal overexpression of substance P, dynorphin, and enkephalins; and the loss of somatostatin-positive neurons.

Table 1 represents a compilation of data on the role of several neuropeptides in epilepsy based on alterations in peptide and/or peptide receptor expression; the effects of exogenously administered ligands; and seizure phenotype in mutant mice.

Other Neuropeptides

Regulation of seizures is not limited to the neuropeptides listed in Table 1 and discussed above. Virtually every single neuropeptide known to date has been examined with regards to its role in epilepsy. These peptides include thyrotropin-releasing hormone, cholecystokinin, corticotropin-releasing factor, bombesin, δ-sleep-inducing peptide, neurotensin, kyotorphin, atrial natriuretic peptide, angiotensin, and vasopressin, although in many cases reports on their effects in epilepsy lack systematic approach and mechanistic studies.

Seizure-Induced Neuropeptide Plasticity: Maladaptive and Adaptive Changes in Epilepsy

For obvious reasons, research reports generally focus on the role of a single peptide in epilepsy; however, as we see from the data presented so far, seizures lead to multiple modifications in peptide and peptide receptor expression, which occur both simultaneously and sequentially. A synthesis of observations accumulated from experimental studies allows us to define a broader role of peptidergic modulation of epileptic process. Based on the pharmacological effects and seizure phenotype in mutant animals, peptides can be divided in two groups: anticonvulsant, that is, those that counteract epilepsy, and proconvulsant, that is, those that facilitate seizures (it should be emphasized that such a division is not completely accurate, keeping in mind that in certain cases the same peptide may exhibit opposite effects on seizures in different models and that the same peptide acting on different receptor subtypes may have the opposite effects). Therefore, some changes in the peptide and/or peptide receptors expression observed in epileptic brain can be interpreted as adaptive, that is, those that counteract seizures and that include the seizure-induced overexpression of anticonvulsant peptides or the downregulation of proconvulsant peptides. At the same time, other alterations are apparently maladaptive because they favor seizure progression; the latter include the downregulation of anticonvulsant peptides or overexpression of proconvulsant peptides (Figure 1). When applied to

Table 1 Role and possible mechanisms of seizure modulation by neuropeptides

Peptide	*Changes in peptide expression (hippocampus)*	*Changes in peptide receptor expression (hippocampus)*	*Role in epilepsy based on the effects of peptide receptor ligands*	*Seizure phenotype in mutant mice*	*Mechanism of action*
Dynorphin	Decrease (status epilepticus; kindling)	No data available	Anticonvulsant (kappa)	No data available	G-protein-coupled inward rectifier K^+ channels (GIRK) – activation; Sub-threshold voltage-gated K^+ channels (KCNQ2/KCNQ3) – activation
Enkephalins	Increase (poststatus epilepticus; chronic epilepsy)	Increase (Mu-opioid receptors, poststatus epilepticus)	Proconvulsant	No data available	Ca^{2+} influx – facilitation KCNQ2/KCNQ3 channels – inhibition (through delta receptor) GIRK – activation
Orphanin FQ	Increase (kindling; status epilepticus)	Decrease (status epilepticus)	Anticonvulsant (0 Mg^{2+} in hippocampal slices; kindling)	Increased resistance (peptide knockout, kindling; status epilepticus)	KCNQ2/KCNQ3 channels – activation
Galanin	Decrease (poststatus epilepticus)	No change (GalR1, poststatus epilepticus)	Anticonvulsant (GalR1/Gal2)	Increased susceptibility (peptide knockout); increased resistance (peptide overexpression)	ATP-sensitive K^+ channels (KATP) – activation
		Decrease (GalR2, poststatus epilepticus)	Proconvulsant (GalR2)	Spontaneous seizures (Galanin receptor 1 knockout); increased susceptibility (galanin receptor 1, galanin receptor 2 knockouts)	GIRK – activation
Neuropeptide Y	Increase (kindling; poststatus epilepticus; chronic epilepsy); ectopic expression	Increase (Y2, poststatus epilepticus and kindling, TLE)	Anticonvulsant (Y2; Y5)	Increased susceptibility (peptide knockout)	GIRK – activation
		Decrease (Y1, poststatus epilepticus and kindling)	Proconvulsant (Y1)	Increased susceptibility (Y5 receptor knockout)	
Somatostatin	Increase (kindling)	Decrease (SST2, kindling; SST3, SST4, poststatus epilepticus)	Anticonvulsant-SST2 (rats)	Increased susceptibility (somatostatin knockout, SST2 knockout, kindling, status epilepticus);	KCNQ2/KCNQ3 channels – activation
	Reduced release; loss of SOM interneurons (poststatus epilepticus, human TLE)		Proconvulsant-SST4 (mice)	Increased resistance (SST2 receptor knockout, status epilepticus in mice)	
Substance P	Increase (poststatus epilepticus)	No data available	Proconvulsant (status epilepticus);	Increased resistance (pre-protachykinin knockout	Facilitation of NMDA-mediated transmission
			Anticonvulsant (kindling; entorhinal cortex *in vitro*)		Facilitation of GABA-mediated transmission KCNQ2/KCNQ3 – inhibition
Neurokinin B	Increase (kindling; status epilepticus)	No data available	Proconvulsant	No data available	GABA-A receptors internalization

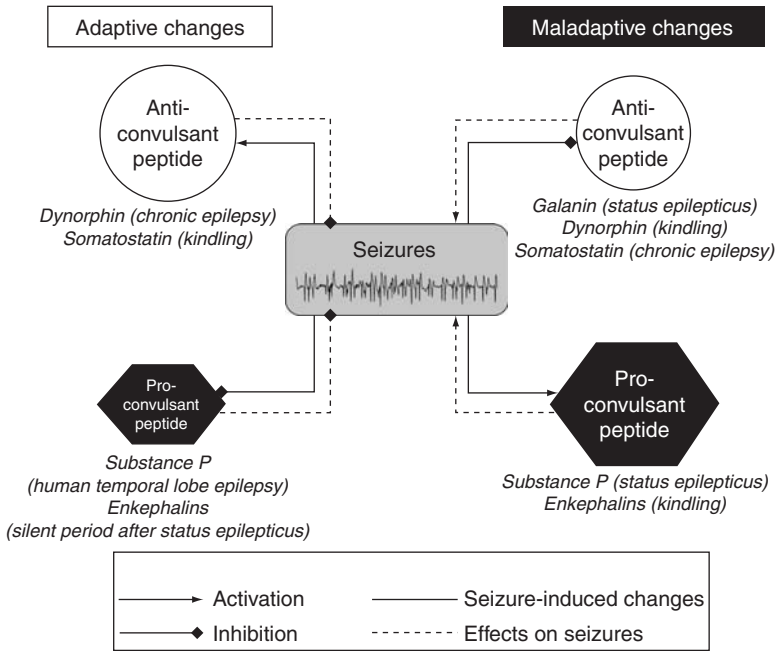

Figure 1 The role of seizure-induced plasticity of neuropeptides in seizure regulation. Based on the effects of exogenously administered neuropeptides and/or changes in seizure susceptibility in neuropeptide/peptide receptor mutant mice, neuropeptides can be described as either anticonvulsant or proconvulsant. Hence, seizures-induced alterations in the expression of certain neuropeptides and/or their receptors can be interpreted as adaptive, that is those that counteract seizure activity (e.g., upregulation of anticonvulsant and downregulation of proconvulsant peptides), or maladaptive, that is those that favor seizure progression (e.g., downregulation of anticonvulsant and upregulation of proconvulsant peptides). Examples of adaptive and maladaptive changes for selected peptides are provided in italics; more information is given in Table 1.

neuropeptides, the concept of adaptive and maladaptive changes might have important clinical implications, as it would become a central principle for the development of peptide-based anticonvulsant therapies.

Same Neuropeptide Can Be Both Anticonvulsant and Proconvulsant

Apparently controversial results obtained in various experimental studies revealed that the same peptide may both inhibit and facilitate seizures. Although there are many reasons for such opposite effects, the following three scenarios are most important in evaluating the overall effects of neuropeptide of interest.

In the Same Brain Area, the Neuropeptide Acts at Different Receptor Subtypes

The effect of neuropeptide Y in the hippocampus exemplifies this scenario (Figure 2). In the hippocampus, neuropeptide Y is expressed in a subset of inhibitory interneurons, where it coexists with the classic neurotransmitter GABA. Three types of receptors for neuropeptide Y are expressed in the hippocampus: Y5, Y2, and Y1. Y5 receptor is located on the axons projecting from the dentate granule cells into the CA3 area (mossy fibers); the activation of Y5 by neuropeptide Y inhibits the propagation of epileptic discharge from the dentate gyrus further into the hippocampus and beyond. Y2 receptors are expressed on excitatory interneurons, mossy cells, that are driven by and project back to dentate granule cells, thus providing feedback excitation; the activation of Y2 disrupts reverberation of the excitation between the dentate granule and mossy cells. Y1 receptors are located on the neuropeptide Y-containing interneurons, which are also driven by and project back to the dentate granule cells, thus providing feedback inhibition; the activation of Y1 by neuropeptide Y released from the same neuron inhibits further neuropeptide release.

The differences in the localization of neuropeptide Y receptors explain why both Y5 and Y2 agonists acted as anticonvulsant (through the interference with feed-forward and feedback excitation, respectively), whereas the Y1 agonist facilitated seizures (attenuation of the effects of the peptide on receptors Y5 and Y2 due to the inhibition of peptide release). Furthermore, differential localization of receptors Y5 and Y2 implies that the correspondent agonists will inhibit epileptic activity through different modes – whereas the former primarily interfere with the seizure spread, the latter are particularly important in suppressing seizure activity within the epileptic focus.

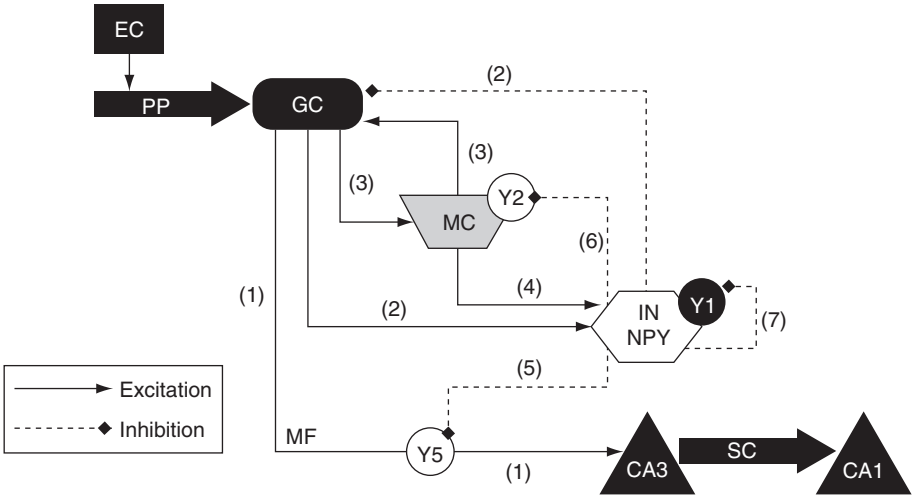

Figure 2 Regulation of hippocampal excitability and seizures by neuropeptide Y (NPY) and its receptors. Seizure discharge originating from the entorhinal cortex (EC) spreads through perforant path (PP) to the dentate granule cell of the dentate gyrus (GC). Excitation of GC involves three pathways: (1) feed-forward excitation through mossy fibers (MF) to the CA3 pyramidal cells and further through Schaffer collaterals (SC) to the CA1 pyramidal cells; (2) feedback inhibition through the excitation of GABAergic inhibitory interneurons (IN), including those that contain NPY; and (3) feedback excitation through glutamatergic excitatory interneurons mossy cells (MC). In addition to the direct projection of granule cells, inhibitory interneurons receive positive drive from mossy cells (4). Positive drive to NPY-positive interneurons is accompanied by the release of NPY that acts on three types of receptors: Y5 receptors (5), located on mossy fibers (the activation of Y5 interferes with the spread of seizure activity from granule cells into the CA3); Y2 receptors (6), expressed on mossy cells (the activation of Y2 blocks reverberation of seizure discharge between mossy cells and granule cells); and Y1 autoreceptors (7), located on NPY-positive interneurons (the activation of Y1 inhibits NPY release).

In Different Brain Areas, Activation of the Same Receptor Subtype May Have Opposite Effects on Seizures

In this scenario, galanin acting at galanin receptor type 1 (GalR1) may both inhibit and facilitate seizures of hippocampal origin (Figure 3). The activation of GalR1 in the hippocampus, achieved by the local infusion of GalR1 agonist, inhibits the release of the major excitatory neurotransmitter glutamate and exerts the end anticonvulsant effect. At the same time, local activation of GalR1 in the dorsal raphe inhibits serotonin release and blocks serotonergic projection from this area into the hippocampus. Because hippocampal serotonin itself inhibits seizures through the activation of postsynaptic receptors 5HT1A, which in turn interferes with glutamate release, the end effect of GalR1 agonist applied locally into the dorsal raphe is to facilitates seizure response.

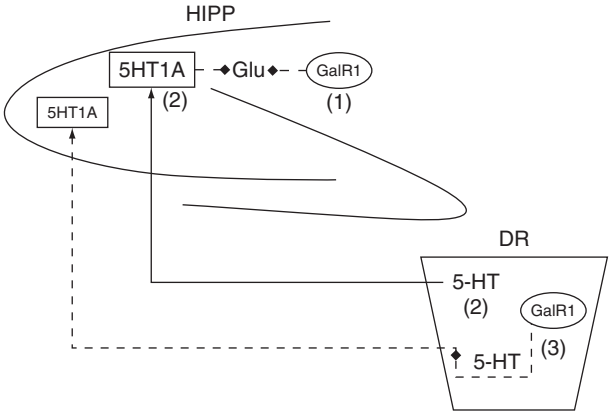

Figure 3 Regulation of limbic seizures by galanin type 1 receptors (GalR1s). Activation of GalR1 in the hippocampus inhibits seizures through presynaptic inhibition of glutamate (Glu) release from principal hippocampal cells (1). Another mechanism that counteracts seizures is an ascending serotonergic pathway originating from the dorsal raphe (DR) and projecting into the hippocampus (HIPP) (2); it inhibits glutamate release through hippocampal 5HT1A receptors. The activation of GalR1s in the DR (3) inhibits serotonin (5-HT) release and facilitates seizures through attenuating the serotonergic inhibition of excitatory transmission in the hippocampus.

Same Peptide Acting at the Same Receptor Subtype Has Different Species-Dependent Effects

In rats, somatostatin has been shown to inhibit seizures through receptor sst2. At the same time, sst2 agonists failed to modify seizures in wild-type mice. Moreover, sst2 knockout mice exhibited a surprising attenuation of seizure severity compared to wild-type animals. This phenomenon has been explained by the fact that sst4 receptors, which by themselves facilitate seizures, are coupled to sst2; thus, in sst2 knockout mice the loss of function of the proconvulsant somatostatin receptor subtype leads to the reduced seizure susceptibility.

Mechanisms of Seizure Modulation by Neuropeptides

Neuropeptide receptors belong to the family of G-protein-coupled receptors (GPCRs), with a variety of downstream mechanisms; some of these mechanisms play an important role in epilepsy and thus represent possible targets for both the pro- and anticonvulsant effects of neuropeptides (Table 1).

G-Protein-Coupled Inward Rectifier K^+ Channels

G-protein-coupled inward rectifier K^+ channels (GIRKs) represent a target for the anticonvulsant effects of such neuropeptides as dynorphin, galanin, and neuropeptide Y. The activation (opening) of GIRKs depends on the coupling of GPCRs to $G_{\beta\gamma?}$ subunit of G_i-protein. Depending on whether the GIRK is pre- or postsynaptic, the resulting inwardly rectifying K^+ current interferes with either neurotransmitter release on the presynaptic level or with the postsynaptic effects of transmitter through neuronal hyperpolarization.

Proconvulsant effects of enkephalins, which also have been shown to activate GIRKs, might be explained by the hyperpolarization of inhibitory interneurons in the hippocampus and with the ultimate interference with the release of GABA (a phenomenon known as the inhibition of inhibition).

Adenosine Triphosphate-Sensitive K^+ Channels

Adenosine triphosphate-sensitive K^+ channels (KATPs) are another type of inward potassium rectifier and represent an apparent target for galanin. The function of KATPs is mainly controlled by the metabolic status of the neurons, in which a decrease of the adenosine triphosphate (ATP) : adenosine diphosphate (ADP) ratio leads to the channel opening. In addition, the opening of KATPs occurs through an active G_α subunit of G_i protein and results in inward rectifying K^+ current, which interferes with both pre- and postsynaptic aspects of neurotransmission. Thus, the end effects of KATPs activation is similar to that of GIRKs.

Subthreshold Voltage-Gated K^+ Channels

The subthreshold voltage-gated K^+ channels KCNQ2 and KCNQ3represent a substrate for a noninactivating, slowly deactivating, subthreshold current known as M-current and are activated in the presence of such neuropeptides as orphanin FQ, somatostatin, and dynorphin. In contrast, it has been suggested that substance P and δ-opioid receptor agonists inhibit M-current. M-currents play a critical role in controlling neuronal excitability and have been implicated in the mechanisms of epilepsy. The novel, clinically available antiepileptic drug retigabine inhibits seizures through enhancing KCNQ2/KCNQ3 currents.

Other Mechanisms

In addition to the mechanisms discussed so far, facilitation of seizures by enkephalins may occur through the stimulation of Ca^{2+} influx and the consequent increase of neuronal excitability. Substance P and neurokinin B exhibit a unique effect on classic neurotransmission – they stimulate trafficking of *N*-methyl-D-aspartate (NMDA) receptor from inside the cell into the membrane and thus facilitate excitatory neurotransmission via the increase in the availability of target receptor for glutamate (at the same time, electrophysiological data found facilitation of GABA-ergic transmission by substance P, which may explain its anticonvulsant effect *in vitro*).

Can Neuropeptides Be Turned into Antiepileptic Drugs?

The ability of neuropeptides to modulate seizures raises the question of whether peptide receptor ligands can be used as antiepileptic drugs. Physical-chemical and biological properties of native neuropeptides determine their low bioavailability and thus prohibit their use in clinical practice. Such properties include digestion in the gut, poor permeability through the blood–brain barrier, and a very short half-life due to quick inactivation by peptidases. Furthermore, reported opposite effects on seizures exhibited by the same peptide acting on different receptor subtypes (e.g., neuropeptide Y acting at Y1 vs. Y5 and Y2) implies that a peptide-based antiepileptic drug may have to possess receptor-subtype selectivity; at the same time, native peptides do not discriminate among their receptor subtypes.

We consider three approaches have been used to modify peptide receptor ligands to optimize their pharmacokinetic and pharmacodynamic properties: nonpeptide ligands, peptide modification, and alternative routes of delivery.

Nonpeptide Ligands for Peptide Receptors

This approach pursues three major goals: reducing the size of the active molecule and increasing its permeability through the blood–brain barrier, reducing biodegradation, and enhancing receptor subtype preference.

Several nonpeptide low-molecular-weight ligands of peptide receptors have been synthesized and examined for their anticonvulsant efficacy. Particularly, a

selective κ-opioid agonist U-69,593, mixed galanin type 1 and type 2 receptor agonist galnon, preferential galanin receptor type 1 agonist galmic, sst2 agonists BIM 23120 and L-779976Y1, and Y1 antagonist BIBP3226 exhibited anticonvulsant effects upon systemic administration under conditions of experimental seizure models. Ligands for other neuropeptides receptors are available and, based on the effects of native peptides, have antiepileptic potential (e.g., neurokinin 1 antagonist M869), but have not been yet studied for anticonvulsant efficacy.

Peptide Modification

Nonpeptide approaches worked well for some peptides, but it was not as successful for others. For example, currently available nonpeptide galanin agonists have only moderate affinity toward galanin receptors, compared to native peptide, and one of them (galnon) has interaction sites outside the peptide receptor. An alternative approach in developing peptide-based drugs consists of preserving peptide structure but modifying it in such a way as to enhance bioavailability. For example, the 8 amino acid synthetic peptide octreotide exhibits high affinity toward receptor sst2 and exerts potent anticonvulsant effects when given systemically.

Alternative Routes of Delivery

The ubiquity of neuropeptide targets both in the brain and in peripheral tissues and organs, such as the gut, heart, and pancreas, may lead to side effects. Thus, it represents a substantial challenge for using peptides as antiepileptic drugs.

The nasal route of delivery of neuropeptides has the potential to avoid peripheral side effects along with other important advantages, such as bypassing the peptide digestion in the gut and the liver first-pass effect. This approach has been applied to thyrotropin-releasing hormone (TRH) to inhibit experimentally induced convulsions.

Viral vector-mediated transduction and expression of anticonvulsant factors in select neuronal populations affords direct delivery and continuous production of anticonvulsant peptides in the epileptic brain. This strategy has been applied to neuropeptide Y and galanin in experimental epilepsy models.

Presently, peptide-based anticonvulsant therapy is at its early preclinical stage, although several peptide receptor ligands with proven or prospective anticonvulsant potential have been approved for clinical use, for example, the sst2 agonist octreotide for the therapy of certain hormone-secreting tumors and the substance P antagonist MK869 for treating chemotherapy-induced emesis.

See also: Epilepsy; Epilepsy: Channelopathies; Galanin and Receptors; Neuropeptide Receptors – Drug Development; Neuropeptide Release; Neuropeptide Y (NPY) and its Receptors; Neuropeptides and Coexistence; Opioid Peptides and Receptors; Somatostatin and Receptors; Substance P/Tachykinins and its/their Receptors.

Further Reading

Baraban SC and Tallent MK (2004) Interneuron diversity series: Interneuronal neuropeptides – endogenous regulators of neuronal excitability. *Trends in Neuroscience* 27: 135–142.

Charles AC and Hales TG (2004) From inhibition to excitation: Functional effects of interaction between opioid receptors. *Life Sciences* 76: 479–485.

Hokfelt T, Bartfai T, and Bloom F (2003) Neuropeptides: Opportunities for drug discovery. *Lancet Neurology* 2: 463–472.

Mazarati AM (2004) Galanin and galanin receptors in epilepsy. *Neuropeptides* 38: 331–343.

McCown TJ (2004) The clinical potential of antiepileptic gene therapy. *Expert Opinion in Biological Therapy* 4: 1771–1776.

Morimoto K, Fahnestock M, and Racine RJ (2004) Kindling and status epilepticus models of epilepsy: Rewiring the brain. *Progress in Neurobiology* 73: 1–60.

Schwarzer C, Sperk G, Samanin R, et al. (1996) Neuropeptides-immunoreactivity and their mRNA expression in kindling: Functional implications for limbic epileptogenesis. *Brain Research Brain Research Reviews* 22: 27–50.

Simonato M and Romualdi P (1996) Dynorphin and epilepsy. *Progress in Neurobiology* 50: 557–583.

Vezzani A and Hoyer D (1999) Brain somatostatin: A candidate inhibitory role in seizures and epileptogenesis. *European Journal of Neuroscience* 11: 3767–3776.

Vezzani A, Sperk G, and Colmers WF (1999) Neuropeptide Y: Emerging evidence for a functional role in seizure modulation. *Trends in Neuroscience* 22: 25–30.

Wasterlain CG, Mazarati AM, Naylor D, et al. (2002) Short-term plasticity of hippocampal neuropeptides and neuronal circuitry in experimental status epilepticus. *Epilepsia* 43(supplement 5): 20–29.

Wickenden AD (2002) Potassium channels as anti-epileptic drug targets. *Neuropharmacology* 43: 1055–1060.

Zupanc GK (1996) Peptidergic transmission: From morphological correlates to functional implications. *Micron* 27: 35–91.

Neuropeptides: Food Intake

S F Leibowitz, Rockefeller University, New York, NY, USA
K E Wortley, Regeneron Pharmaceuticals, Tarrytown, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

There is a rich network of interrelated factors controlling feeding behavior and body weight. These include the status of energy stores and metabolic demand, the availability of diet and circulating nutrients, and the palatability of ingested foods. This network functions through a wide variety of neurochemical systems in the brain, which are either stimulatory or inhibitory in nature. Those that increase feeding and weight gain include the peptides neuropeptide Y (NPY), agouti-related protein (AgRP), melanin-concentrating hormone (MCH), galanin (GAL), orexins (ORX), and the opioids enkephalin (ENK) and dynorphin (DYN). Those that inhibit feeding are peptides such as proopiomelanocortin, cocaine-activated regulatory transcript, and corticotropin-releasing hormone, in addition to the monoamine serotonin. This article focuses on the seven peptides that increase feeding. It summarizes a large body of information related to the anatomical distribution of these peptide systems, the endocrine signals of energy abundance or insufficiency or diet and nutrient signals that control their activity, and the behavioral and physiological effects produced by peptide injections, antagonists, or mutations of the peptide and receptor genes (see Table 1). In addition to their considerable redundancy, these peptide systems show great diversity and specialization in their actions that allow them to maintain energy and nutrient homeostasis under a variety of physiological states and environmental conditions that occur over a lifetime.

Neuropeptide Y

NPY is expressed in neurons concentrated in the arcuate nucleus (ARC), which send their projections to different areas of the hypothalamus. These include the paraventricular nucleus (PVN) and perifornical area/lateral hypothalamus (PFLH), where NPY is believed to act through different receptor subtypes (Y1R, Y5R and possibly Y2R) to control energy homeostasis.

This peptide in the ARC and PVN is highly responsive to states of negative energy balance and increased metabolic demand. It is markedly stimulated by food deprivation and the adrenal steroid corticosterone (CORT), which is released in states of energy deficiency to maintain glucose availability. Both NPY and this glucocorticoid exhibit a strong circadian peak at the onset of the natural feeding period, when gluconeogenesis and glycogenolysis are elevated. The gut hormone ghrelin, also an indicator of energy insufficiency, rises before fasting or meal initiation and activates NPY neurons in the ARC. These positive relationships between NPY and both CORT and ghrelin contrast with this peptide's inverse relation with the adiposity-related hormones insulin and leptin, which provide a signal of energy abundance and have a potent inhibitory effect on NPY. These findings illustrate the homeostatic principle of negative feedback, whereby a rise in insulin or leptin and/or decline in CORT or ghrelin with food consumption and increased body weight suppresses anabolic peptides, like NPY, that promote further eating and weight gain.

In addition to these hormones, dietary nutrients as well as circulating and intracellular nutrients can affect endogenous NPY in the ARC and PVN. This peptide is consistently higher in subjects consuming a low-energy, carbohydrate-rich diet. Together with CORT, NPY peaks at the onset of the active feeding cycle or around the time of weaning, when carbohydrate is the strongly preferred macronutrient and energy stores are low. This relationship to carbohydrate intake confirms that NPY is positively linked to signals reflecting a deficiency of glucose availability or utilization. This is supported by evidence that NPY is increased after pharmacological blockade of glucose utilization and is inversely related to glucose levels, declining immediately after the ingestion or injection of glucose when circulating glucose is high and rising subsequently as glucose levels drop toward or below the baseline and CORT rises.

Injection studies demonstrate that NPY functions in states of negative energy balance to stimulate the ingestion and metabolism of carbohydrate to maintain glucose homeostasis and promote fat synthesis. Exogenous NPY has a potent, stimulatory effect on feeding, with a preferential increase in the ingestion of carbohydrate or sucrose, and it stimulates body fat accrual. Its endocrine and metabolic effects include a rise in circulating levels of both CORT and insulin, which together promote *de novo* lipogenesis, and a reduction in energy expenditure, sympathetic nervous system activity, and fat utilization in favor of carbohydrate metabolism. Thus, by suppressing inhibitory postingestive feedback, NPY both replenishes carbohydrate stores and gears excess carbohydrate towards fat storage.

Mice models with a deletion of the *npy* or receptor genes show little change or even an increase in daily

Table 1 Summary of peptides

Orexigenic peptides	*Responsive to negative energy balance?*	*Responsive to dietary fat?*	*Primary functions*
Neuropeptide Y	Yes	No	Stimulate food intake during periods of negative energy balance and increased metabolic demand
Agouti-related protein	Yes	No	Reduce signaling through catabolic pathways during periods of negative energy balance, thereby stimulating food intake
Melanin-concentrating hormone	Yes	Yes	Maximize intake of energy-rich foods and storage of excess calories as fat
Galanin	No	Yes	Restore carbohydrate balance under conditions of excess dietary fat intake
Orexins	Yes	Yes	Coordinate motivated behaviors with sleep/wake cycle
Enkephalin Dynorphin	No	Yes	Stimulate intake of palatable, fat-rich diets

food consumption or body weight. However, closer examination of these genetic mutants reveals a phenotype consistent with the proposed function of NPY in states of negative energy balance. This phenotype includes a decrease in fasting-induced refeeding, diabetes-induced hyperphagia, weight gain in leptin-deficient mice, and utilization of carbohydrate relative to lipids. Further, transgenic mice with NPY overexpression exhibit an obese phenotype, specifically on a sucrose-loaded diet, which is accompanied by overeating, hyperglycemia, hyperinsulinemia, and increased NPY receptor expression.

In summary, these investigations demonstrate that NPY neurons in the ARC are stimulated under conditions of negative energy balance and increased metabolic demand. This peptide rises in response to changes in various hormones, including elevated levels of CORT and ghrelin and reduced levels of leptin and insulin, and it responds positively to signals related to a need for carbohydrate, decline in glucose levels or storage, or disturbance in glucose utilization. The anabolic functions of endogenous NPY are to promote food intake, replenish carbohydrate stores, reduce fat oxidation in favor of carbohydrate, and stimulate the utilization of excess glucose to promote *de novo* lipogenesis. These NPY effects involve reciprocal interactions with neural systems that inhibit feeding, such as the melanocortins in the ARC, corticotropin-releasing hormone in the PVN, and serotonin in the midbrain; and this peptide may be a downstream mediator of the feeding-stimulatory effects of orexigenic peptides, such as ORX and MCH.

Agouti-Related Protein

AgRP is expressed in the ARC in neurons that also produce NPY. It forms part of the central melanocortin system, which comprises α-melanocyte stimulating hormone, an agonist of both melanocortin receptor 3 (MC3-R) and receptor 4 (MC4-R), in addition to AgRP, an endogenous competitive antagonist of these receptors. Whereas AgRP cell bodies are restricted to the ARC, AgRP fibers project to multiple areas of the hypothalamus and brain, and all of these AgRP terminals contain NPY.

Studies indicate a close functional relationship between AgRP and NPY because the expression of these peptides is similarly modulated under various physiological conditions. This is evident in states of negative energy balance or increased energy demand, such as food deprivation or lactation, which have reduced leptin and insulin levels, increased ghrelin or CORT levels, and elevated AgRP together with NPY. Like NPY, AgRP shows a strong circadian rhythm, rising with CORT at the onset of the nocturnal feeding cycle, and it is suppressed under conditions of positive energy balance, when leptin and insulin levels are elevated and ghrelin levels are reduced. The activity of NPY/AgRP neurons is excited by orexigens, such as ghrelin and ORX, but decreased by leptin.

Dietary studies reveal further similarities between these two peptides. As described for NPY, AgRP gene expression is higher in rodents consuming a high-carbohydrate low-fat diet, and it decreases as the fat content of the diet increases. It peaks in rat pups at weaning and is associated with peak circulating CORT and a strong preference for dietary carbohydrate. Thus, AgRP is stimulated by low energy status and by periods of reduced fat consumption, such as the transition to weaning, when glucose, insulin, and leptin levels are falling.

As with NPY, AgRP has a potent stimulatory effect on food intake. Chronic administration of AgRP increases food intake and decreases metabolic rate, leading to increased body fat accrual. Both peptides appear to mediate the orexigenic effect of ghrelin and

suppress the hypothalamic–pituitary–thyroid axis. They are additive in their effects on feeding, suggesting that they represent redundant components of a parallel circuit.

A large number of genetic studies have been devoted to identifying the role of AgRP in physiological settings. Overexpression of the *Agrp* gene in transgenic mice produces hyperphagia and obesity. However, young AgRP-deficient mice, similar to NPY-deficient mice, are normophagic with normal body weight under *ad libitum* feeding conditions, indicating that these peptides do not function tonically when food is freely available. However, following brief food deprivation, AgRP knockout mice, like the NPY mutants, exhibit impaired reflexive hyperphagia. This supports the physiological data implicating AgRP in processes geared toward restoring energy homeostasis when food sources are low. Interestingly, as AgRP knockout mice age, they maintain a leaner body composition, in contrast to the progressive adiposity seen in aging wild-type mice, and this is associated with increased metabolic rate, possibly due to the increased thyroid hormone seen in these mice.

Taken together, this evidence demonstrates that AgRP in the ARC is similar to NPY in being stimulated by conditions of energy deficiency. Under these conditions, AgRP is required to stimulate hunger, reduce lipid oxidation in favor of carbohydrate, and, when food deprivation is prolonged, reduce locomotor activity and metabolic rate via inhibition of the hypothalamic–pituitary–thyroid axis in order to conserve energy. These anabolic effects are mediated, in part, via inhibitory projections to catabolic melanocortin cells in the ARC. These two parallel systems, AgRP and NPY, probably evolved to ensure the signaling of hunger during food scarcity and to enable the body to endure long periods of negative energy balance.

Melanin-Concentrating Hormone

MCH is an orexigenic peptide which acts via a specific receptor, MCHR1. This peptide is heavily expressed in the hypothalamus, predominantly in the PFLH and zona incerta but also in the ARC and ventromedial hypothalamus (VMH), suggesting a role in feeding and energy homeostasis. It is even more heavily expressed in areas of the forebrain which are involved in the rewarding aspects of motivated behaviors.

Studies of MCH provide support for this peptide's role in energy balance. Fasting, which reduces leptin, stimulates the expression of MCH, and although leptin injection suppresses MCH, this fasting-induced increase in MCH can be seen in leptin-deficient mice. Further, rats that become obese on a fat-rich diet and have markedly elevated leptin levels show no difference in MCH expression. This indicates that MCH is regulated by other hormonal factors, such as the glucocorticoid CORT and gonadal steroids.

There are limited data regarding the sensitivity of MCH neurons to circulating nutrients. The firing frequency of MCH neurons increases in response to rising concentrations of glucose. Also, MCH expression in the PFLH is stimulated by the administration of an inhibitor of fatty acid oxidation or fatty acid synthase. This evidence indicates that MCH responds to a deficiency in fat oxidation or fat synthesis.

Central administration of MCH stimulates food intake but also increases the consumption of alcohol and sucrose. Further, this peptide impacts metabolic processes that contribute to weight gain. It enhances the secretion of CORT and suppresses the release of thyroid stimulating hormone. Chronic administration of MCH or an MCHR1 agonist induces hyperphagia and increases body fat mass, particularly on a highly palatable diet, and these effects occur in association with reduced fatty acid oxidation and thermogenesis. Conversely, MCHR1 antagonists decrease body weight and adiposity, due to both a reduction in food intake and an increase in energy expenditure. Together, these findings indicate that MCH acts through metabolic as well as behavioral mechanisms to promote weight gain.

Genetic studies support a physiological role for MCH in energy homeostasis. Transgenic mice that overexpress MCH exhibit hyperphagia and obesity on a fat-rich diet. Targeted ablation of the MCH gene leads to a lean phenotype associated with hypophagia and an increased metabolic rate, and it attenuates the marked obesity of leptin-deficient mice through increased energy expenditure and locomotor activity, effects that become more apparent on a high-fat diet. Similarly, mice lacking functional MCHR1 gain less weight than wild-type littermates on a high-fat diet, due to an increase in locomotor activity that stimulates metabolic rate and thermogenesis. These mice become leaner despite a significant increase in food intake, suggesting that MCH plays a more prominent role in regulating energy expenditure and activity than feeding behavior.

Together, these results indicate that MCH has an important role in energy homeostasis. Whereas MCH responds to negative energy balance and is anatomically and functionally linked to NPY neurons in the ARC, this peptide can function independently of leptin and is unresponsive to nutrient-sensing signals indicative of reduced glucose metabolism. The evidence that the orexigenic effect of MCH is most prominent with palatable diets and that MCH neurons project to reward centers in the brain suggests that this peptide recognizes energy-rich foods and functions to maximize the intake of the energy source

and its storage as fat. These effects should allow the efficient storage of energy during periods of food excess and better equip animals for ensuing periods of food deprivation.

Galanin

GAL is expressed in a number of neuronal populations within the hypothalamus, including the PVN and PFLH, as well as ARC. These neurons, which send projections throughout the hypothalamus where the GAL receptor subtypes, GALR1 and GALR2, exist, are believed to have a role in feeding behavior and energy homeostasis. Investigations of endogenous GAL or GAL injection reveal marked differences between this peptide system and the orexigenic peptides, NPY and AgRP, already described.

In the different hypothalamic nuclei, GAL is found to be relatively unresponsive to signals related to energy deficiency, including food restriction, CORT, and the blockade of glucose oxidation. Further, leptin is relatively weak in its effects on endogenous GAL, with GAL expression in the PVN consistently elevated in obese rats on a high-fat diet. Also, GAL is particularly sensitive to the stimulatory effects of the gonadal steroids, which are increased by dietary fat. Thus, GAL in the PVN functions through distinct mechanisms and in different physiological states or conditions than NPY and AgRP in the ARC.

Endogenous GAL also responds very differently from NPY to manipulations of diet and nutrient metabolism. In the PVN but not the ARC, GAL in rats on pure macronutrient diets is positively related to the amount of fat but not carbohydrate consumed. It is stimulated by acute and chronic periods of high-fat diet consumption, and it peaks toward the middle of the natural feeding cycle, when fat is the preferred macronutrient. Also, it is strongly, positively related to circulating levels of triglycerides, in addition to measures of fat uptake and oxidation in muscle, and it is reduced by pharmacological blockade of fat oxidation, which suppresses fat intake. Thus, GAL, unlike NPY, appears to function specifically under dietary conditions rich in fat and respond to signals related to the metabolism of fat rather than carbohydrate.

The effects of GAL injection also differ from NPY, indicating that it functions under different conditions and physiological states. The GAL-induced feeding response is smaller in magnitude and shorter in duration, but it is stronger and more prolonged, leading to increased weight gain, specifically in animals that are on a mixed high-fat diet or that naturally prefer fat when given a choice of macronutrients. It is greatly reduced when fat is removed from the diet or after injection of the pentapeptide enterostatin and GAL antagonist, both of which reduce fat consumption. In some cases, GAL's physiological effects are diametrically opposite to those of NPY, including a suppression of insulin, CORT, and vasopressin release and an increase in gastric acid secretion. Also, GAL, unlike NPY, has no effect on respiratory quotient and increases the capacity of muscle to metabolize carbohydrate. Thus, when stimulated on a high-fat diet, GAL's primary function may be to restore carbohydrate balance, compensating for or reversing the diet-induced drop in carbohydrate metabolism while contributing to the hyperphagia that provides more dietary carbohydrate to replenish liver glycogen stores.

Mice with mutations of the *gal* gene or GAL receptor genes or that overexpress GAL are found to normally defend their food intake and body weight when fasted or fed *ad libitum* on a low-fat chow diet. Thus, in contrast to NPY, endogenous GAL is not a critical component in the metabolic response to starvation and has little function under dietary conditions low in fat and high in carbohydrate. There is evidence, however, to support the idea that GAL functions specifically under conditions when dietary fat is in excess. Repeated PVN injections of antisense oligonucleotides to GAL mRNA suppress fat consumption in association with a reduction in endogenous GAL, and the deletion of the *gal* gene causes a marked reduction in fat consumption. Further, GAL knockout mice exhibit increased sensitivity to the inhibitory effects of leptin on body weight, indicating that endogenous GAL plays a role in counteracting the actions of leptin as it rises on a fat-rich diet.

In summary, GAL-expressing neurons in the PVN have very different properties and functions from the other orexigenic peptides described so far. This peptide is relatively unresponsive to states of energy deficiency and conditions where food is scarce or rich in carbohydrate. In contrast, GAL functions in close relation to dietary fat, mediating specific functions that allow animals to adapt to these conditions of positive energy balance and abundant fat intake. Through behavioral and metabolic actions, it restores carbohydrate balance under conditions of excess dietary fat that compromise the utilization, storage, and intake of carbohydrate. To perform these functions, GAL may act through other peptide systems in the brain, including the opioid peptides, which show a similar relationship to dietary fat.

Orexins

The orexin peptides, ORX A and ORX B (also referred to as hypocretin 1 and 2), share 46% identity,

are coded by the same gene, and are localized in neurons of the PFLH. These ORX neurons project within the hypothalamus and throughout the brain, and the two ORX receptors, OX1R and OX2R, display similar widespread distribution patterns.

The ORX neurons are controlled by changes in circulating nutrients that link feeding behavior to arousal and physical activity. For example, ORX expression is stimulated by a high-fat diet, and this increase in ORX expression is positively related to levels of circulating triglycerides. Consistent with this relationship is the circadian rhythm of ORX peptides, which, like GAL, show peak levels several hours after onset of the feeding cycle, when rats exhibit a rise in their preference for dietary fat. This association between ORX neurons and circulating lipids suggests a possible contribution of this peptide system to the acute hyperphagia associated with consumption of a high-fat diet.

The ORX neurons are also controlled by changes in circulating glucose. They exhibit tonic spontaneous firing, which is inhibited by signals related to food ingestion and by physiological or experimentally induced increases in glucose, and they are stimulated when plasma glucose falls and food is withheld. Based on evidence linking the ORX peptides to the sleep–wake cycle, these findings led to the proposal that this peptide system is a key player in coupling nutrient homeostasis to states of consciousness.

Central administration of the ORX peptides, particularly ORX A, stimulates food intake. This peptide produces a stronger feeding response on a high-fat diet compared to a low-fat diet, supporting the idea that ORX contributes to the hyperphagia induced by a high-fat diet. The ORX peptide system is also implicated in the sleep–wake cycle and is thought to play a role in maintaining periods of wakefulness. Further, ORX A stimulates spontaneous physical activity independently of its effects on food intake. These effects of ORX substantiate the proposed role of this peptide system in regulating energy balance and suggest that a primary role for ORX may be synchronizing feeding and activity with the sleep–wake cycle.

Genetic studies support a role for the ORX peptides in energy balance. Orexin-deficient mice show hypophagia and inactivity compared to control littermates, which may contribute to their unexpected late-onset obesity, and they fail to exhibit the increase in arousal induced by food deprivation. Their phenotypes are similar to narcolepsy, a human sleep disorder. These findings support the idea that the obese phenotype in ORX-deficient mice reflects an underlying decrease in energy expenditure, due to reduced locomotor activity, and in foraging behaviors. They indicate that this peptide system regulates both food intake and physical activity, but in the absence of ORX, the effect on activity predominates. Thus, rather than arising from greater hunger *per se*, the feeding-stimulatory effect of ORX may be attributed to an increase in wakefulness and locomotor activity that is optimal for food-seeking behavior.

In conclusion, these results indicate that ORX neurons are highly responsive to changes in circulating nutrients. The ORX peptides sense negative energy balance, predominantly through a decrease in circulating glucose, and they act to stimulate feeding by increasing arousal and the motivation to seek food. They also sense positive energy balance induced by a high-fat diet and rise with circulating lipids, possibly contributing to the hyperphagia and activity-based energy expenditure associated with a fat-rich diet. Through their interactions with NPY/AgRP and melanocortin neurons in the ARC, the ORX peptides may provide a critical link between energy balance and brain mechanisms coordinating sleep–wakefulness and motivated behaviors.

Opioid Peptides

Opioid peptides have long been known to have a role in stimulating feeding behavior. This article focuses on the peptides ENK, which acts through μ and δ receptors, and DYN, which acts on κ receptors. These two peptides and their receptors are heavily expressed in the different hypothalamic nuclei and also in various forebrain structures, such as the nucleus accumbens (NAc), which have an important role in mediating the rewarding properties of food.

Similar to GAL, these opioid peptides are very different from NPY in their response to energy deficiency. In the ARC as well as NAc, ENK and DYN are either unaffected or reduced by food deprivation, and DYN in the ARC is actually stimulated by leptin. Whereas there are few studies of these opioids in relation to the adrenal and gonadal steroids, those that do exist show little effect of estradiol on DYN in the hypothalamus, in contrast to its potent stimulatory effect on GAL.

Palatability and sensory attributes of preferred foods that signal high energy density are particularly important in driving caloric consumption in excess of metabolic demand, thus promoting obesity. Opioid signaling in the brain is believed to control food intake, specifically through the modulation of neural processes related to palatability and hedonistic impact of energy-dense foods. Whereas the effects of sucrose drinking on the endogenous opioids and their receptors remain unclear, acute or chronic consumption of palatable,

fat-rich foods (compared to less palatable, low-fat foods) increases ENK and DYN in the hypothalamus, and this effect is strongest and most consistent in the PVN but opposite in the NAc. This occurs independently of the palatability and caloric density of the diet and is also seen after injection of a lipid emulsion, which bypasses the ingestion process and thus avoids stimuli related to palatability. These findings suggest that postingestive factors, such as circulating lipids, may be important in determining the expression level and functions of these peptides in the PVN. Consumption or injection of a fat-rich diet invariably produces a marked increase in circulating levels of triglycerides, and these lipids, in turn, are positively related to ENK and DYN and produce an increase in neuronal activity in the PVN, specifically in fat-responsive neurons that express ORX peptides.

Although there is little evidence supporting a role for these opioid peptides in glucose or lipid metabolism and body fat accrual, pharmacological studies with behavioral measures consistently demonstrate a potent stimulatory effect of the opioids on food intake, with the opposite pattern observed with opiate antagonists. The effect of the agonists is particularly strong after injection into the PVN and NAc, and it is largest with palatable diets rich in fat or sugar. In addition to their close relation to signals of diet palatability, the opioids appear to function in relation to the specific macronutrients of the palatable diets, switching preferences from a high-sucrose to a high-fat diet. Evidence suggests that the opioids in the forebrain are more closely related to signals of diet palatability, whereas those in the hypothalamus have a closer relation to the specific macronutrient in the palatable diet.

Mutant mice with a deletion of the opioid genes or their receptors fail to show consistent changes in daily food intake and body weight. However, more subtle changes in behavior are evident. These include a reduced preference for saccharine in opioid receptor-deficient mice and a reduction in willingness to work for palatable food in ENK knockout mice.

These results suggest that these orexigenic opioid peptides play a key role in mediating excess feeding induced by palatable food. Although relatively unresponsive to states of energy deficiency, these opioids, similar to GAL and ORX, are closely, positively related to fat intake and circulating lipids, which promote hyperphagia. These peptides, in turn, potently stimulate the consumption of palatable energy-dense diets, having a stronger effect on a high-fat than a low-fat diet. Their functions are mediated by a broadly distributed network of structures within the hypothalamus and forebrain, with the former related more to macronutrient content of the palatable diet and the latter related to the palatability itself. These opioids are found to be involved in mediating the feeding-stimulatory actions of the other orexigenic systems, including NPY, AgRP, MCH, GAL, and ORX.

Conclusion

The study of eating behavior and obesity from the perspective of the brain is clearly a burgeoning field, with many technological advances that are allowing investigators to make new and exciting discoveries while defining specific functions of the peptide systems in energy homeostasis. These orexigenic systems exhibit considerable diversity and redundancy in their actions, which involve not only the stimulation of food intake but effects on more intricate behavioral and physiological processes related, for example, to the circadian rhythm of meals, arousal of food-seeking behavior, diet palatability, energy expenditure, and nutrient metabolism in different tissues. The studies reviewed here have clearly advanced our understanding, at the level of the hypothalamus, of the complex systems involved in energy homeostasis. Such complex regulation is made possible by the large number of circulating hormones and nutrients that manipulate neuronal function and the diverse anatomical and functional relationships existing among the different peptides within the hypothalamus. It is also made possible by neural mechanisms that extend from the gut and other peripheral organs to the hindbrain and forebrain. A major miscalculation at any step within this complex circuit can lead to overeating and obesity. Whereas the unfolding of the beauty and intricacy of the neural and humoral systems involved is a thing of awe, its application in modern times is impossible without a parallel appreciation of the enormous impact of environment, in terms of food availability and diet composition. The complex systems of energy homeostasis evolved in an environment of scarcity. Thus, in a society where food is superabundant and diets are rich in fat or carbohydrate, we need sociopolitical help in coping with problems of body-weight regulation. An understanding of the neuroscience of eating behavior and obesity, although critical to the discovery and development of future anti-obesity agents, must bring with it a sensible food distribution and education system. Otherwise, it will just be science for the sake of treatment, without all the benefits that could accrue from the prevention of overeating.

See also: Energy Homeostasis: Hypothalamic Development; Food and Water Intake: Regulation; Galanin and Receptors; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Hypocretin/Orexin and MCH and Receptors; Hypothalamic Structure–Function Relationships; Neuroendocrine Control of Energy Balance (Central Circuits/Mechanisms); Neuropeptide Y (NPY) and its Receptors; Opioid Peptides and Receptors.

Further Reading

Bodnar RJ and Klein GE (2005) Endogenous opiates and behavior 2004. *Peptides* 26: 2629–2711.

Flier JS (2006) AgRP in energy balance: Will the real AgRP please stand up? *Cell Metabolism* 3: 83–85.

Hervieu GJ (2006) Further insights into the neurobiology of melanin-concentrating hormone in energy and mood balances. *Expert Opinions on Therapeutic Targets* 10: 211–229.

Kalra SP and Kalra PS (2004) NPY and cohorts in regulating appetite, obesity and metabolic syndrome: Beneficial effects of gene therapy. *Neuropeptides* 38: 201–211.

Kelley AE (2005) Corticostriatal-hypothalamic circuitry and food motivation: Integration of energy, action and reward. *Physiology & Behavior* 86: 773–795.

King PJ (2005) The hypothalamus and obesity. *Current Drug Targets* 6: 225–240.

Leibowitz SF (2005) Regulation and effects of hypothalamic galanin: Relation to dietary fat, alcohol ingestion, circulating lipids and energy homeostasis. *Neuropeptides* 39: 327–332.

Leibowitz SF and Wortley KE (2004) Hypothalamic control of energy balance: Different peptides, different functions. *Peptides* 25: 473–504.

Levine AS (2006) The animal model in food intake regulation: Examples from the opioid literature. *Physiology & Behavior* 89: 92–96.

Rokosz LL and Hobbs DW (2006) Biological examination of melanin concentrating hormone receptor 1: Multi-tasking from the hypothalamus. *Drug News & Perspectives* 19: 273–286.

Sakurai T (2006) Roles of orexins and orexin receptors in central regulation of feeding behavior and energy homeostasis. *CNS & Neurological Disorders – Drug Targets* 5: 313–325.

Williams G, Cai XJ, Elliott JC, and Harrold JA (2004) Anabolic neuropeptides. *Physiology & Behavior* 81: 211–222.

Wilson KR, Todorovic A, Proneth B, and Haskell-Luevano C (2006) Overview of endogenous and synthetic melanocortin peptides. *Cellular and Molecular Biology* 52: 3–20.

Neuropeptides: Mental Disease

F Holsboer, Max Planck Institute of Psychiatry, Munich, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the 1970s, David de Wied formulated the neuropeptide concept, which states that these molecules synthesized and released from nerve cells in the brain exert behavioral effects. Before then, it was well established that central neuropeptides are involved in pituitary hormone secretion and are, in turn, regulated by peripheral hormones via feedback mechanisms, mostly at the hypothalamic nuclei. It was also documented that neuropeptide-containing hypothalamic cells project to extra-hypothalamic targets and, further, that neuropeptides are copackaged with neurotransmitters belonging to the class of biogenic amines. It were Tomas Hökfelt and Jan Lundberg who spearheaded the view that neuropeptides are in contrast to biogenic amines released under conditions in which homeostatic control needs to be established (Figure 1). This view was based on the finding that, in contrast to biogenic amines, neuropeptides are released only at higher neuronal firing rates. This opened the possibility of expanding the time range at which synaptic communication can occur.

In the context of mental disorder, the hypersecretion of neuropeptides can be viewed as the attempt of the brain to adapt to a specific demand, such as coping with a stressor or stimulating hunger in the case of energy loss. If adaptation fails, the continuous neuropeptide hyperdrive may ultimately result in a clinical condition. These neuropeptides are particularly suited for carrying out complex tasks, such as behavioral adaptation because they contain more chemical information and feature more recognition sites for the receptors to which they bind with much higher affinity (10^{-6} vs. 10^{-3} M) than biogenic amine neurotransmitters. Along with the rapidly evolving refinements in measuring neuropeptides and their molecular control mechanisms, we also see a rapidly developing pharmacology that offers small chemical molecules suited for interfering with neuropeptide signaling pathways. In particular, neuropeptide receptors are preferred drug discovery targets because their cognate ligands bind selectively and tightly and they promise to exhibit fewer side effects.

In this article, I take some examples from psychiatry research that demonstrate how our advanced understanding of neuropeptide physiology can be translated into the development of new drugs and where the field is now.

Substance P and Neurokinin Receptor 1 Antagonists in Depression

Substance P, the most abundant neuropeptide in the brain, is a C-terminal amidated, 11-amino-acid residue. When injected into a rodent's brain, substance P induces anxiety-like behavior, which is counteracted by blocking neurokinin receptor type 1 (NK1), to which the neuropeptide binds. This corresponds with studies in mice in which the *Tac1* gene was deleted. These mouse mutants do not produce substance P and show decreased anxiety- and depression-like behavior. Neuroanatomical studies revealed that substance P is present in brain circuits thought to be involved in the processing of anxiety. These include the amygdala, septum, mid-brain, periaqueductal gray, hippocampus, and hypothalamus. There are also connections between biogenic amines and substance P. Central administration of substance P increases neuronal firing in the locus coeruleus, an area in the brain stem reticular formation on the floor of the fourth ventricle, on which norepinephrine fibers project into all areas of the cortex. Thus, substance P-induced increase of neuronal firing in the locus coeruleus results in the activation of many brain areas involved in arousal, affect, and anxiety. In turn, stress exposure results in increased locus coeruleus firing. A direct effect of substance P in the brain stem area is also suggested by the presence of NK1s on tyrosine hydroxylase – positive cell bodies in the locus coeruleus. Based on these and on corresponding findings, NK1 antagonists were developed as drugs to be used when failed stress adaptation is implicated in pathogenesis.

Studies relating pharmacological interventions at NK1s revealed that blocking substance P signaling prevents the stress-induced activation of the ascending noradrenergic projections. Also, the serotoninergic system is intertwined with substance P because, in contrast to rodents, many serotoninergic neurons in the dorsal raphe express substance P. In this brain stem area, most cell bodies of serotoninergic neurons are located. Administration of NK1 antagonists in mice results in decreased anxiety-like behavior in parallel with a selective desensitization of serotonin inhibitory autoreceptors (5-HT_{1A}), which resembles the effect of sustained antidepressant treatment.

Some studies exist that suggest an association between the central release of substance P and depression.

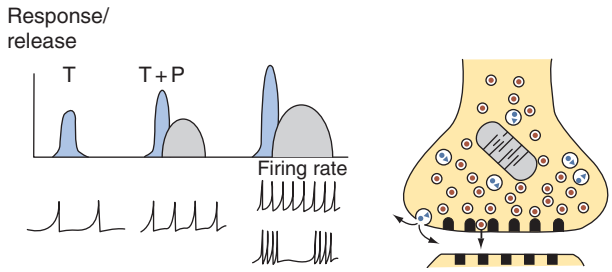

Figure 1 Tomas Hökfelts's hypothesis submits that in synaptic transmission (right), unlike biogenic amine and amino acid neurotransmitters (T), neuropeptides (P) are secreted only at a high firing rate or bursting activity (left).

Some studies found elevated substance P concentration in the cerebrospinal fluid (CSF), and one study reported decreased NK1 binding. Because continuous overexposure of NK1s to substance P results in their internalization, the decreased NK1 binding can be viewed as reflecting a desensitizing mechanism secondary to overabundant substance P exposure.

After the first report of clinical efficacy of NK1 antagonist NK869 in depression, several studies attempted to confirm this novel pharmacological principle but achieved mixed results. Although, there is currently some disillusionment; the jury is still out on whether NK1s will become targets for future antidepressant drugs.

Growth Hormone-Releasing Secretagogues – Multitasking Neuropeptides

There is literally no mental disorder, including dementia, in which sleep is not affected. Thus, it is of interest to explore to which degree changes of neuropeptide signaling are involved. As shown by many investigators, nocturnal sleep is a period of high endocrine activity, and in turn several hormones have the capacity to affect sleep. For example, the onset of sleep is characterized by a steep increase of growth hormone (GH) secretion. During this part of sleep, called slow-wave sleep, low frequencies and high amplitude in the sleep electroencephalogram (EEG) predominate. GH secreted from the pituitary is controlled by a number of neuropeptides:, growth hormone-releasing hormone (GHRH) and ghrelin, which stimulate GH secretion, and somatostatin, which inhibits GH secretion.

GHRH, produced in the hypothalamus, not only governs GH secretion from somatotrophs but also provides the subcortical mechanism synchronizing sleep and GH secretion. Because the GH secretion is highest when we sleep most deeply, it was suggested that GHRH is a key promoter of deep (slow-wave) sleep in humans, as well as the sleep physiological equivalent in animals. This is often termed nonrapid eye movement (non-REM) sleep, sleep episodes during which eye movements are absent. As first demonstrated by Axel Steiger, repeated pulsatile injections of GHRH are important for observing a GHRH-induced effect on the sleep EEG in humans. When GHRH is systemically administered in a pulsatile mode, slow-wave sleep is enhanced in young male subjects; however, women do not respond in this way. Elderly people suffer from fragmental sleep with many awakenings and exhibit much less slow-wave sleep compared to younger people. Therefore, activating GHRH-mediated signaling pathways promised the restoration of sleep in people of advanced age. Indeed, GHRH promotes sleep continuity, but it fails to increase slow-wave sleep in this age group. When we block GH secretion by administering somatostatin, which acts at the somatostatin receptor 2 (sst2), we would expect an effect on sleep opposite to that of GHRH. In fact, when the longer-acting somatostatin analog octreotide, which stimulates sst2 (together with the sst5), is given to animals, both non-REM sleep and GH secretion are suppressed. It is of note that somatostatin receptor-mediated effects fulfill many tasks in the brain. For example, angiotensin-like response, such as prompt drinking, blood pressure rises, and vasopressin surges, were also elicited by octreotide. The effects of sst2 stimulation, however, seem to be independent of octreotides actions on angiotensin because pretreatment with angiotensin-converting enzyme inhibitor did not interfere with octreotide-induced sleep (Figure 2).

A different situation emerged with a third GH-regulating peptide, ghrelin, which exerts its action via the GH secretagogue receptor type 1a (GHS1a) Ghrelin is an acylated peptide predominantly produced by the stomach and has received considerable interest as a molecule with orexigenic effects. Originally, ghrelin, produced in the gastrointestinal tract, was thought to regulate food intake and energy homeostasis by reaching the GHS1a via blood circulation. In that capacity, ghrelin can be viewed as a factor complementary to leptin, which is produced and released from adipocytes and signals satiety. More recently, ghrelin has also been found to be expressed in a previously uncharacterized group of neurons adjacent to the third ventricle between the dorsal, ventral, paraventricular, and arcuate hypothalamic nuclei, which send efferents to a number of nuclei known to be involved in the control of appetite and energy balance. In accordance with this, it was found that ghrelin enhanced appetite and increases food intake in humans.

The ghrelin receptor was recently shown to exhibit a high constitutive activity, probably providing a set

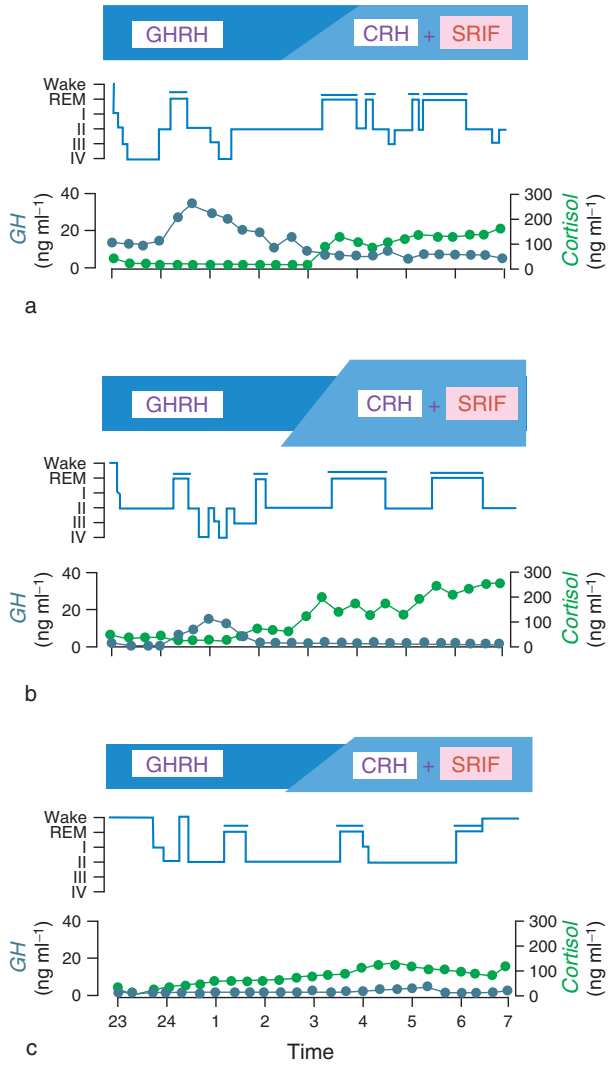

Figure 2 Balance of GHRH, somatostatin (SRIF), and CRH determines sleep regulation and hormonal activity during sleep: (a) young normal control; (b) patient with depression; (c) senior normal control. CRH, corticotropin-releasing hormone; GH, growth hormone; GHRH, growth hormone-releasing hormone; REM, rapid-eye movement sleep.

point for signal transduction activity against which all the inhibitory mechanisms act. Leptin, for example, when lowered is a key signal for an increased food intake and decreased energy expenditure, probably mediated through the decreased inhibition of the constitutive ghrelin receptor signaling. Unsurprisingly, the ghrelin receptor is viewed as an intriguing target for developing drugs that regulate eating behavior. More specifically, ghrelin receptor antagonists are long sought-after antiobesity drug candidates. The induction of sleep by GHRH in males and the impairment of sleep by somatostatin receptor activation raised interest in how ghrelin may act on sleep. Pulsatile injections of ghrelin to young healthy volunteers showed that this peptide enhances not only GH secretion but also slow-wave sleep. From this exploratory study, it was surmised that in addition to its effect on GH secretion, food intake, and weight gain, ghrelin is a sleep-promoting neuropeptide. Whether stimulating the GHS receptor is a pharmacological route for future sleep medicines is not yet easy to predict because via the same receptor hunger is induced, which is an undesired side effect for a sleep-promoting drug.

Modulating the Vasopressin System – A Yet Unexploited Therapeutic Strategy for Treating Anxiety Disorders and Depression

Arginine vasopressin (AVP) is a nonapeptide, and this system serves two purposes:

1. Parvocellular neurons of the hypothalamic paraventricular nucleus (PVN) secrete AVP into the portal blood, where it synergizes the actions of corticotropin-releasing hormone (CRH) on the anterior pituitary to control secretion of corticotropin (ACTH) and corticosteroids.
2. Magnocellular neurons of the nucleus supraopticus and PVN project to the posterior pituitary, where they enter the blood circulation and contribute to the regulation of blood pressure (via renal water reabsorbtion) and hepatic glycogenolysis.

In this article, the focus is on behavioral effects of AVP, particularly those related to coping with stress and to anxiety and depression. Abundant literature exists showing that centrally released AVP has many effects on memory processing, social affiliation, and affective conditions in rodents. The possibility that AVP plays also a role in human clinical conditions was hinted at by neuroendocrine studies in which depressed dexamethasone-pretreated patients responded to intravenous injections of CRH with increased ACTH and cortisol secretion. The hypothesized underlying mechanism submits that depressives,when pretreated with dexamethasone, have suppressed ACTH, and subsequently cortisol, concentrations. Thus, their brains are deprived of cortisol, the negative feedback signal, which cannot be compensated by dexamethasone because of its poor blood–brain barrier penetration. In response to this cortisol depletion, central CRH-stimulating neuropeptides are activated. The most powerful of them is AVP, which is supposed to be overexpressed in depression and which synergizes at the pituitary level with injected CRH, overriding the dexamethasone suppression at pituitary corticotrophs. The ACTH and cortisol hypersecretion as seen in depressives following the administration of the combined

dexamethsone–CRH test gradually normalizes prior to the resolution of depressive psychopathology. Clinical studies measuring AVP in the PVN of suicide victims and basic science experiments using rats with innate high anxiety confirm this view. The latter animal model generated by Rainer Landgraf by selectively breeding animals according to the degree of their anxiety-like behavior, was of particular interest because it showed that the rats' AVP hypersecretion is associated with abnormal test results in the combined dexamethasone–CRH test and because treatment with an antidepressant normalized this neuroendocrine abnormality. The mechanism causing central AVP hypersecretion in these high-anxiety rats was shown to be a polymorphism in the AVP gene promoter, which disrupts the cognate binding of a transcriptional repressor leading to unrestrained AVP overexpression (see Figure 3).

Whether a comparable mechanism is also causally related to depression and anxiety in humans is unclear. Measurements of AVP in the CSF and plasma provided mixed results, and whether plasma and CSF concentrations of AVP reflect the release of the neuropeptide at specific brain sites is arguable. Nevertheless, animal studies in combination with neuroendocrine and postmortem neuropathological studies provided an argument that AVP receptors (V_{1A} and V_{1B}) are potentially interesting drug targets. The V_{1B} antagonist SSR-149415 was shown to induce antidepressant- and anxiolytic-like effects in rodents. This compound also reversed the stress-induced suppression of neurogenesis in the dentate gyrus in a mouse model of depression. Neurogenesis is currently viewed as a surrogate marker for antidepressant-like activity and is suppressed by corticosteroid hypersecretion. A number of other V_{1A} and V_{1B} antagonists are currently under consideration for clinical development.

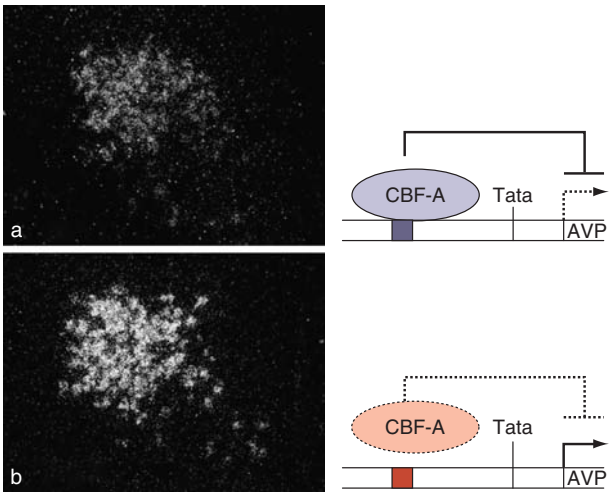

Figure 3 Hypersecretion of AVP: (a) low-anxiety; (b) rats with high innate anxiety. A mutation in the promoter of the AVP-Gen in prevents the binding of the repressor (right), resulting in increased AVP expression in the hypothalamic PVN (left). AVP, arginine vasopressin; CBF-A, CArG binding factor A; PVN, paraventricular nucleus.

The Role of Neurotensin in Causality and Treatment of Schizophrenia

A number of neuropeptide and other neurotransmitter systems have been implicated in the pathophysiology of schizophrenia and antipsychotic mechanisms of drug action. Aside from the hypothesized mesolimbic hyperdrive of dopamine, the strongest case for a neuropeptide has been made for neurotensin, a tridecapeptide.

1. Pathophysiological approaches suggest that the primary disturbance in schizophrenia lies in neurocircuits that converge in the dopaminergic projections from the ventral tegmental area to the nucleus accumbens. Many afferents implicated in schizophrenia, including glutamatergic afferents from the prefrontal cortex, hippocampal areas, or the amygdala terminate in this nucleus from which projections to cortical areas emerge. All these circuits are closely associated with neurotensin systems.
2. It was repeatedly confirmed that patients with schizophrenia had lower CSF neurotensin concentrations and, further, that these changes are secondary to lowered neurotensin secretion in the brain and do not stem from peripheral sources.
3. When antipsychotic drugs are given and the psychopathology is resolved, CSF neurotensin concentrations increase. This accords with preclinical studies showing that antipsychotic drugs increase neurotensin mRNA and neuropeptide concentration and secretion at the nucleus accumbens.

These findings indicate that the antipsychotic effect of drugs given for the treatment of schizophrenia work via enhanced neurotransmission with neurotensin as messenger. Specifically, it was demonstrated in two animal models of antipsychotic drug activity that the blockade of neurotensin signaling by neurotensin receptor antagonists prevented the haloperidol-elicited behavioral effect. Haloperidol is an antipsychotic drug (APD) that primarily acts at dopamine type 2 receptors (DA2s). Mapping APD actions in the brain using immediate early gene expression allowed the identification of brain regions and neurotransmitter systems that are potentially involved in mediating the actions of these drugs. In fact, this technique corroborated the predominant role of the nucleus accumbens because all APDs increased *FOS* expression in this locus. Currently, the first-line treatment of schizophrenia is the use of atypical antipsychotics,

which are different from typical APDs in that they actively block DA2 less but also antagonize $5HT_{2A}$. It is remarkable that only typical (e.g., haloperidol), but not atypical, APDs activate *FOS* expression in the dorsolateral caudate/putamen; this is probably related to the development of extrapyramidal side effects produced by typical, but not atypical, APDs. In contrast, atypical APDs, which are superior to typical APDs with regard to the improvement of negative symptoms of schizophrenia (e.g., deficiency in emotional response, hedonia, and spontaneous speech and motivation), increase *FOS* in the medial prefrontal cortex, specifically in the deep layers of the infralimbic and ventral prelimbic cortices. Many data point to the possibility that neurotensin may be involved in the antipsychotic mechanism of action of APDs. Increased neurotensin neurotransmission following APD administration may also contribute to APD-induced *FOS* expression. A recent extension of this approach also studied the interaction of haloperidol and a neurotensin antagonist (SR14 2948A) and measured APD-induced *FOS* expression as endpoint. In essence, this study did not find a consistent effect on APD-induced *FOS* expression in all brain areas investigated. However, on the behavioral level there was a clear effect in that the neurotensin antagonist was able to prevent the APD-induced restoration of prepulse inhibition of the acoustic startle reflex in isolation-reared rats. This paradigm is taken as a predictive test for antipsychotic-like activity. The reported effect of the neurotensin antagonist is believed to be mediated by inhibited neurotensin neurotransmission in the nucleus accumbens. Based on these findings, it had been suggested that neurotensin receptor agonists may be an interesting opportunity for a new class of antipsychotic drugs.

Corticotropin-Releasing Hormone Signaling – New Targets for Drugs for the Treatment of Stress-Related Disorders

The impaired regulation of plasma stress-hormone secretion is frequently observed in major depression, depressive episodes in the context of bipolar disorder, and psychotic depression. This finding links stress and traumatic experiences to genetic vulnerability to affective disorders. Stressors of all kinds, and particularly those that occur early in life during the critical phase of brain development, can facilitate the development of depression among those carrying a genetic disposition. This relationship between external factors (stressors) and genetic disposition has advanced research exploring how the stress hormone system is regulated. The original finding among depressive patients was hyperactivity of the hypothalamic–pituitary–adrenocortical (HPA) system as reflected by elevations of the stress hormones (ACTH and cortisol), whose secretary pattern was characterized by a flattening of the diurnal rhythm, and increased cortisol release at any secretory burst. In fact, although only a fraction of depressives shows this HPA hyperactivity, disturbed regulation also becomes apparent when refined neuroendocrine function tests are administered. Today, the most sensitive test to identify impaired HPA regulation is the combined dexamethasone–CRH test, which administers a low (1.5 mg) dose of dexamethasone at night (23.00 h) and injects a dose of 100 μg CRH intravenously the next day at 15.00 h. The elicited plasma ACTH and cortisol concentrations are strikingly different between normal controls and depressives (Figure 4).

Because the exaggerated response to this test normalizes weeks before depressive symptomatology resolves, the combined dexamthesone–CRH test is now considered a biomarker of treatment outcome. However, all these results reflect the activity of the HPA system and do not necessarily inform us about the central stress hormone system. In other words, the absence of a peripherally documented stress hormone abnormality does not rule out a still-existing central stress hormone disturbance. Whereas CRH released from the neurons of the hypothalamus may produce hypercortisolism via the HPA system, CRH released from other brain loci and binding at the prefrontal cortex, amygdala, hippocampus, locus coeruleus, and so on can induce behavioral changes unrelated to HPA changes. A key role for CRH, a 41-amino-acid neuropeptide, in mediating the signs and symptoms of human depression came from studies showing an elevation of CRH in the CSF of depressives along with decreased CRH binding sites in the prefrontal cortex of depressed suicide victims and from neuroendocrine test results such as blunted ACTH response to an intravenous dose of CRH. A decrease of CRH in patients responding to antidepressant treatment is well documented.

A large body of evidence from basic research corroborated the view that CRH is a key mediator of anxiety and depression (see Figure 5). For example, intracerebroventricular administration of CRH produced behavioral changes in rats that are analogous to the cardinal symptoms of depression. This was also the case for mice which transgenically over- or underexpress CRH. Transgenic mice overexpressing CRH show behavioral changes that include increased anxiety-like behavior, decreased sexual activity, deficits in learning, and impairment in sleep. In contrast, mouse mutants in which the CRH gene has been silenced do not exhibit a behavioral phenotype,

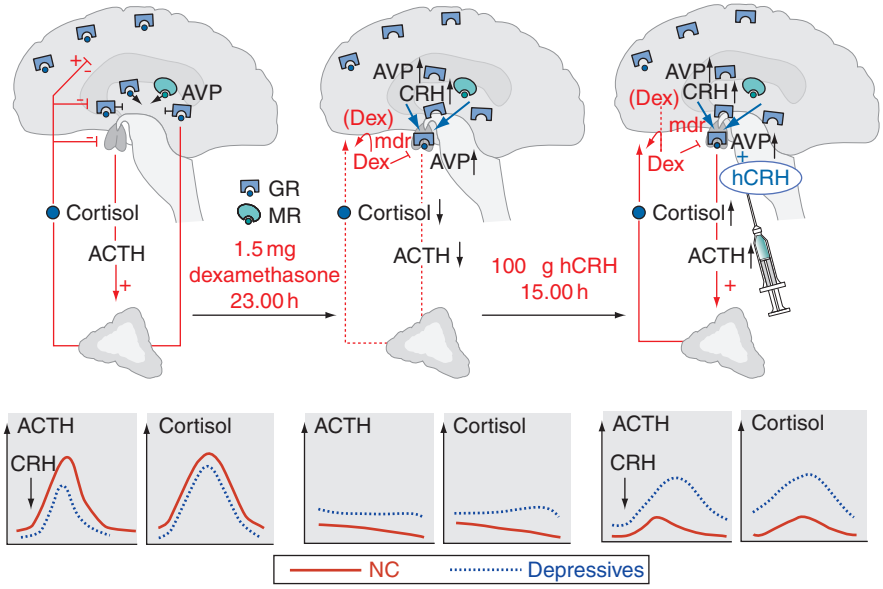

Figure 4 Among depressives (red), the combined dexamethasone–CRH test results in increased HPA activity because AVP is increased at many CNS sites and also at the pituitary (controls = black). ↑, increased; ↓, decreased; +, stimulates; −, inhibits; ACTH, corticotropin; AVP, arginine vasopressin; CNS, central nervous system, CRH, corticotropin-releasing hormone; Dex, dexamethasome; GR, glucocorticoid receptor; h, human; HPA, hypothalamic–pituitary–adrenal; mdr, multidrug resistance gene; MR, mineralocorticoid receptor.

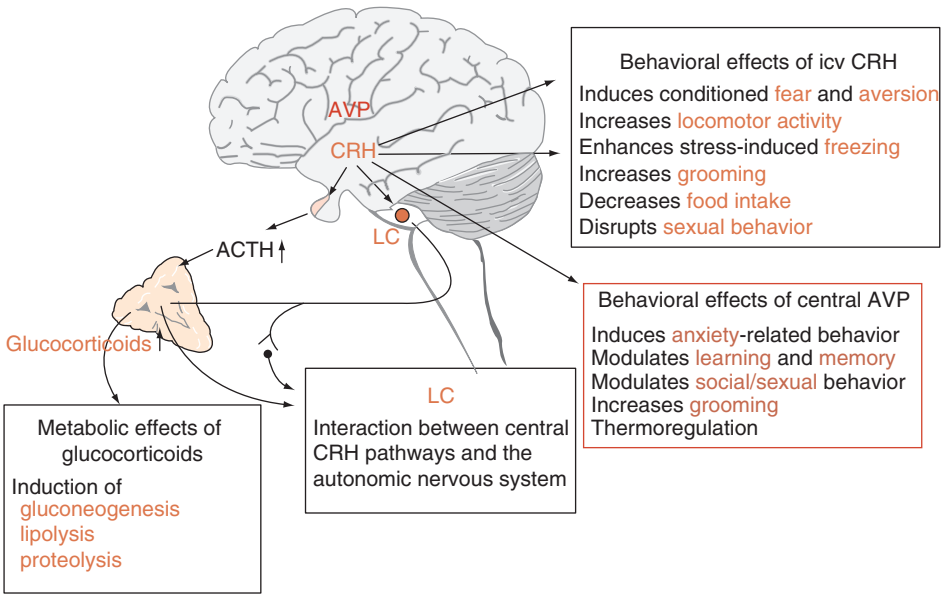

Figure 5 Signs and symptoms induced by CRH and AVP resemble many features common in affective disorders. ↑, increased; ACTH, adrenoscorticotropin; AVP, arginine vasopressin; CRH, corticotropin-releasing hormone; icv, intracerebroventricular; LC, locus coeruleus.

indicating that CRH is only important when needed to adapt to a stressor. Also, the effect of early trauma has been reproduced in animal models. Prolonged separation in neonatal rats, for example, produced depression-like symptoms along with increased CRH mRNA and CRH expression in the hypothalamus, locus coeruleus, and amygdala; decreased CRH receptor binding in the pituitary; and increased CRH concentrations in the median eminence, portal blood, and CSF. Similarly, chronic stress in adult mice produces persistent behavioral changes reminiscent of depression. It is of note, however, that not everyone exposed to a stressor or trauma develops depression, neither acutely nor later in life, inasmuch as not every patient who suffers from a depressive episode has experienced a major stressor that would explain his

or her clinical condition. Genetic disposition seems to be an important determinant, and stressful experiences, which via epigenetic modulation have often longstanding effects on gene activity, act in concert on liability or resilience to depression.

All these findings led to the question: through which pathway is the CRH signal conveyed? Until now there are two CRH receptor subtypes identified, which display distinct anatomical localizations and pharmacology. Both subtypes of G-protein-coupled receptors (GPCRs) belong to the class B family, and they have been termed CRH_1 and CRH_2, with CRH_2 existing in three identified splice variant isoforms: $CRH_{2(a)}$, $CRH_{2(b)}$, and $CRH_{2(c)}$. Both receptor subtypes are found in the human pituitary and throughout the neocortex of primates. In the rodent brain, only the CRH_1 is expressed at the pituitary and neocortex. Not only CRH but also the urocortins activate the CRH receptors. Urocortin (UCN-) I has equal affinity for both receptor subtypes, whereas-UCN II (often termed stress copin-related peptide) and UCN III (often termed stresscopin) have higher affinity and selectivity for CRH_2. In the light of these differences and because human CRH binds at the CRH_1 with 15-fold higher affinity over CRH_2, the drug discovery process targeting CRH signaling in the brain focused on CRH_1.

This development was supported by studies in rodents, in which suppressed CRH_1 gene expression via antisense or interference RNA directed against CRH mRNA revealed decreased anxiety-like behavior when the rodents were exposed to the elevated-plus maze test. Also, genetically engineered disruption of CRH signaling in CRH_1-null mutants showed decreased anxiety-like behavior. Over the last 15 years, a number of small-molecule CRH_1 antagonists have been developed, but until now none of them has entered clinical practice. In an animal model in which rats were selectively bred for high anxiety, a clinical candidate drug (R121919) reduced anxiety-like behavior, whereas in nonanxious rats with the same genetic background the drug failed to alter their level of anxiety. The anxiolytic effects of R121919 in rats were independent of the dose-dependent attenuation of stress hormone secretion. This corroborates the hypothesis that behavioral alterations and peripheral levels of stress hormones are not always closely linked. The first series of clinical studies used R121919 as CRH_1 antagonist and demonstrated a clinical profile consistent with antidepressant activity. In an exploratory small open-label study, patients responded to R121919 in a similar way as treated with paroxetine, a selective serotonin reuptake inhibitor. As in animals, in depressed patients no direct relationship between hormonal effects and changes in psychopathology was observed. Because CRH hyperactivity in the brain has been implicated in sleep disturbances associated with depression, it is remarkable that R121919 improved sleep in these patients. Although R121919 had been discontinued because of safety concerns, some other CRH_1 antagonists are now under study. One of these is NBI-34041, an orally available heterocyclic compound with a chemical structure distinct from other related compounds. When higher dosages were given to rats, stress-elicited ACTH levels were suppressed. This compound was also tested clinically in normal volunteers who underwent a psychosocial stress test, called the Trier Social Stress Test (TSST). In the TSST, the hormonal response to a public-speaking and a mental-arithmetic stress situation is measured. Although the CRH antagonists failed to change hormonal response at baseline or after an intravenous stimulation with CRH, markedly decreased hormonal responses to the TSST were observed. This finding confirms that neither central stress hormone regulation at baseline nor CRH-elicited responsivity is attenuated by the CRH_1 antagonist. However, the administration of this drug does protect the central stress hormone system from overshooting in response to an emotional challenge. Blocking all CRH_1 that fulfill many diverse functions in the brain and the periphery may, nevertheless, be harmful. Therefore, the discovery that CRH, when administered to the brain induces a specific pattern of signaling pathway activation is of interest. Beyond CRH_1 receptor, stimulation of mitogen-activated protein kinase (MAPK) occurs. This pathway depends on protein kinase A (PKA) activity involving phosphorylated extracellular-signal-regulated kinase (ERK)1/ 2. These kinases are only activated in limbic areas, such as CA_3–CA_1 hippocampal cell layers and the basolateral complex of the amygdala, two brain areas involved in the neuroanatomy of depression and anxiety. Because ERK1/2 are pleiotrophic kinases, their brain-area-specific activation by CRH_1 signaling suggests that they can be used as bait for finding downstream partners that, through their specific cellular context, can be explored as targets for drugs with restricted effects to relevant brain loci.

Conclusion

It is obvious that neuropeptides have functions in the brain that are distinct from their roles as regulators of peripheral homeostasis under stress exposure. Significantly, peptides may play only a minor role under normal circumstances and the absence of defined phenotypes of many neuropeptide-null mutant mice points in this direction. Therefore, neuropeptide receptors are interesting future targets because blocking

their function may not result in major effects as long as there is no potentially pathological condition. Thus, neuropeptide signaling pathways provide an opportunity for the discovery of drugs to treat mental illness specifically with a low risk of side effects.

See also: Anxiety Disorders; Anxiety: Drug Therapy; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Corticotropin-Releasing Hormone and Urocortins: Binding Proteins and Receptors; Depression and the Brain; Growth Hormone ; Neuropeptide Receptors – Drug Development; Neurotensin and Receptors; Psychiatric Disorders Associated with Disturbed Sleep and Circadian Rhythms; Sleep and Sleep States: Cytokines and Neuromodulation; Sleep Deprivation: Neurobehavioral Changes; Stress, the HPA Axis and Depressive Illness; Substance P/ Tachykinins and its/their Receptors; Vasopressin/ Oxytocin and Receptors.

Further Reading

Arzt E and Holsboer F (2006) CRF signaling: Molecular specificity for drug targeting in the CNS. *Trends is Pharmacological Sciences* 27: 531–538.

Binder EB, Kinkead B, Owens MJ, and Nemeroff CB (2001) The role of neurotensin in the pathophysiology of schizophrenia and the mechanism of action of antipsychotic drugs. *Biological Psychiatry* 50: 856–872.

de Kloet RE, Joëls M, and Holsboer F (2005) Stress and the brain: From adaptation to disease. *Nature Reviews Neuroscience* 6: 463–475.

de Wied D (1999) Behavioral pharmacology of neuropeptides related to melanocortins and the neurohypophyseal hormones. *European Journal of Pharmacology* 375: 1–11.

Ghigo E, Broglio F, Arvat E, Maccario M, Papotti M, and Muccioli G (2005) Ghrelin: More than a natural GH secretagogue and/or an orexigenic factor. *Clinical Endocrinology* 62:1–17.

Hökfelt T (1991) Neuropeptides in perspective: The last ten years. *Neuron* 7: 867–879.

Hökfelt T, Johansson O, Ljungdahl A, Lundberg JM, and Schultzberg M (1980) Peptidergic neurones. *Nature* 284: 515–521.

Holsboer F (1999) The rationale for corticotropin-releasing hormone receptor (CRH-R) antagonists to treat depression and anxiety. *Journal of Psychiatric Research* 33: 181–214.

Holst B and Schwartz TW (2004) Constitutive ghrelin receptor activity as a signaling set-point in appetite regulation. *Trends in Pharmacological Sciences* 25: 113–117.

Insel TR and Young LJ (2000) Neuropeptides and the evolution of social behaviour. *Current Opinion in Neurobiology* 10: 784–789.

Ising M, Horstmann S, Kloiber S, et al. (2007) Combined Dex/ CRH test predicts treatment response in major depression – a potential biomaker? *Biological Psychiatry* 62: 47–54.

Landgraf R (2005) Neuropeptides in anxiety modulation. *Handbuch für Experimentelle Pharmakologie* 169: 335–369.

Nemeroff CB and Vale WW (2005) The neurobiology of depression: Inroads to treatment and new drug discovery. *Journal of Clinical Psychiatry* 66: 5–13.

Rupniak NM and Kramer MS (1999) Discovery of the antidepressant and anti-emetic efficacy of substance P receptor (NK1) antagonists. *Trends in the Pharmacological Sciences* 20: 485–490.

Sakurai T, Amemiya A, and Ishii M (1998) Orexins and orexin receptors: A family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behaviour. *Cell* 92: 573–585.

Steiger A (2006) Neurochemical regulation of sleep. *Journal of Psychiatric Research* 41: 537–552.

Van der Lely AJ, Tschöp M, Heiman M, and Ghigo E (2004) Biological, physiological, pathophysiological, and pharmacological aspects of ghrelin. *Endocrine Review* 25: 426–457.

Neuropeptides: Pain

X-J Xu and Z Wiesenfeld-Hallin, Karolinska Institutet, Stockholm, Sweden

© 2009 Published by Elsevier Ltd.

General Aspects of Neuropeptides and Pain

The possibility that substance P (SP) could be a nociceptive transmitter was raised by Lembeck in the early 1950s based on the finding that there is higher concentration of SP in the dorsal roots and dorsal horn than in the ventral roots and ventral horn. But it was not until the 1970s that the role of neuropeptides as pain modulators, both in excitatory and inhibitory capacities, began to be seriously considered. This was sparked mainly by the identification of SP as an undecapeptide and the localization of SP in small-diameter, nonmyelinated afferent fibers. Subsequently, several other neuropeptides were also found in dorsal root ganglia (DRG) neurons, including neurokinin A (NKA), calcitonin gene-related peptide (CGRP), somatostatin, vasoactive intestinal peptide (VIP), pituitary cyclase-activating peptide (PACAP), neuropeptide Y (NPY), and galanin. The receptors for these peptides have been localized in the dorsal horn and, in many cases, in the terminals of afferents and on dorsal horn interneurons, which s also express a large number of neuropeptides. This has made the dorsal horn of the spinal cord an interesting area to study the role of peptides in pain transmission.

Like in other areas of the nervous system, peptides often coexist with each other and with classical neurotransmitters in the DRG and dorsal horn. Neuropeptides are generally stored in large dense-core vesicles and are released from these vesicles optimally by higher frequency of stimulation than that required for the release of classical neurotransmitters. The excitatory effect of neuropeptides such as SP and CGRP on postsynaptic neurons is often delayed in onset, but with a prolonged duration of action. This suggests that neuropeptides, in contrast to classical transmitters, play more of a modulatory role in sensory transmission. Further, expression of peptides and peptide receptor undergoes marked changes in the DRG and spinal cord after injury to the nervous system or inflammation, suggesting that some of the neuropeptides may also be involved in adaptive changes and plasticity and may have trophic effects.

Tachykinins

Distribution, Receptors, and Plasticity

The mammalian tachykinin SP is the first neuropeptide to be implicated in pain transmission. SP, together with tachykinin NKA, is present in a subpopulation of DRG cells, primarily with small somata. In addition, SP-like immunoreactivity has also been localized in the fiber network of the dorsal horn, mainly in the outer laminae, but also in bulbospinal tract fibers and in dorsal horn interneurons. In the DRG, SP and NKA are also co-localized with other peptides, such as CGRP and galanin, and with classical neurotransmitters, such as glutamate. Neurokinin B (NKB), on the other hand, appears to be present mainly in dorsal horn interneurons, but not in the DRG.

Three subtypes of tachykinin receptors have been identified and cloned. The NK-1 receptor is concentrated in the superficial laminae of the dorsal horn, particularly in lamina I, where internalization of this receptor has been shown to occur after noxious stimulation. NK-2 receptor-like immunoreactivity has also been observed in the outer part of lamina I and in the area around the central canal. These receptors are, however, exclusively astrocytic. NK-3 receptor-like immunoreactivity has been localized in laminae I, II, and X.

SP is markedly downregulated in the DRG after axotomy. Such downregulation occurs primarily in small-sized DRG neurons, and the loss of target-derived nerve growth factors is believed to be responsible for this effect. Paradoxically, it has been reported that nerve injury triggers a *de novo* synthesis of SP in large DRG neurons. Moreover, NK-1 receptors are upregulated in the dorsal horn after nerve injury. In contrast to nerve injury, inflammation appears to be associated with a moderate increase in the synthesis of SP and NKA in the DRG. The synthesis of NK-1 and NK-3 receptors is also increased in the dorsal horn after peripheral inflammation.

Release and Function

In vitro or *in vivo* release of SP in the spinal cord upon noxious peripheral stimulation has been repeatedly shown by a number of methods. Release occurs primarily at the superficial laminae of the dorsal horn. NKA is also released in the spinal cord following noxious stimulation, and – unlike SP, the release of which is confined in the superficial laminae – released

NKA appears to spread into deeper laminae and persists much longer than does SP. This has lead to the suggestion that NKA may be responsible for more prolonged increase in spinal excitability, compared to SP.

SP excites dorsal horn neurons and produces behavioral reactions indicative of pain upon intrathecal administration in rodents. Intrathecal SP has also been consistently shown to produce hyperalgesia in many nociceptive tests and to increase the excitatory effect and the release of other excitatory transmitters, such as glutamate. However, NK-1 receptor antagonists are not antinociceptive in classical pain tests, such as the hot plate or tail flick tests, and these compounds also have minimal effect on baseline responses of dorsal horn neurons to noxious stimulation. Thus, SP is unlikely to be involved in transmission of basal nociception. On the other hand, activation of NK-1 receptors has been shown to be a critical component in the activity-dependent increase (central sensitization) in spinal cord excitability after intense nociceptive stimulation. Furthermore, NK-1 receptor antagonists are also able to reduce hyperalgesia in rodent models of inflammatory pain. This activity profile of NK-1 receptor antagonists has been confirmed in studies with knockout mice lacking genes for the NK-1 receptor or preprotachykinin.

Possible Clinical Application

Several clinical trials have been conducted to examine the analgesic efficacy of NK-1 receptor antagonists in humans in various pain conditions, such as postoperative dental pain, osteoarthritis, diabetic neuropathy, postherpetic neuralgia, and migraine, and the results have been largely disappointing. There may be several reasons for this apparent discrepancy between animal and human studies, including species differences in the distribution and function of the NK-1 receptors, the relevance of the human clinical pain model used, or the redundancy of pain involving large numbers of transmitters and modulators and the complex interactions among them.

More recently it has been suggested that SP may also be involved in neuropathic pain after peripheral nerve injury. This is primarily based on the finding that nerve injury causes increased synthesis and release of SP in and from large-diameter sensory neurons. It is, however, important to note that nerve injury is associated with downregulation of SP in small-diameter C-afferents, and the role of SP in mediating central sensitization was diminished after nerve injury. Functional studies examining the effect of NK-1 antagonists in rodent models of neuropathic pain have produced mixed results, and NK-1 antagonists are ineffective in human neuropathic pain conditions. Depleting the NK-1 receptor does not influence the development of neuropathic painlike behaviors. Overall, it is unlikely that NK-1 receptor antagonists could be useful as analgesics in neuropathic pain.

Calcitonin Gene-Related Peptide

CGRP is the most abundant peptide in mammalian primary afferents, where it is present in sensory neurons of all sizes. CGRP is released from primary afferents at their peripheral and central terminals. The role of CGRP in peripheral neurogenic inflammation is well established, whereas in the spinal cord CGRP is also believed to have an excitatory role. CGRP depolarizes dorsal horn neurons with a very slow onset and long duration of action. In both behavioral and electrophysiological studies, CGRP has been found to markedly enhance the excitatory effect of SP and other neuropeptides. Mice lacking CGRP genes fail to develop heat hyperalgesia after inflammation and have reduced spinal sensitization and diminished responses to pain-causing agents. Similar to SP, the synthesis of CGRP is downregulated after nerve injury and upregulated after inflammation.

Recently, a direct role for CGRP in heat pain sensitivity has been established using a genetic approach. It was shown in mice that robust strain differences in behavioral sensitivity to noxious heat are due to genetic differences in CGRP expression and sensitivity. Such CGRP dependence is also reflected in differential responsiveness of primary afferent thermal nociceptors to heat stimuli.

Somatostatin

Somatostatin is present in a subset of small-diameter sensory neurons; these neurons do not express SP, but coexist with CGRP. Several subtypes of somatostatin receptors are also expressed in the DRG and dorsal horn. The synthesis of somatostatin is downregulated after nerve injury.

Somatostatin and its analogs have been shown to induce pain relief in a number of clinical conditions. This effect of somatostatin is believed to be related mostly to its peripheral effect and/or to its direct effect on the disease process. In the spinal cord, somatostatin produces complex effects on neuronal activity; a previously reported antinociceptive effect following intrathecal administration appears to be due to a neurotoxic action of this peptide in the spinal cord.

VIP, PACAP-27, and PACAP-38

Three structurally related peptides, VIP, PACAP-27, and PACAP-38, have limited presence in small to

medium-sized DRG neurons under normal conditions, but are markedly upregulated after nerve injury. Some reports also suggest an upregulation of PACAPs following peripheral inflammation.

Upon application to the spinal cord, VIP and PACAP produce primarily an excitatory effect on dorsal horn neurons, increase spinal nociceptive reflexes, and produce behavioral hyperalgesia. The relative physiological roles played by these peptides are increased after nerve injury, replacing tachykinins as important mediators of central sensitization. Furthermore, antagonists of receptors for these peptides inhibit neuronal responses to noxious stimulation in the dorsal horn, particularly after nerve injury. Thus, the receptors for VIP-related peptides may be a drug target for development of analgesics in neuropathic pain.

Galanin

Like the VIP family of peptides, galanin is expressed in few sensory neurons in rat DRG under normal circumstances, although the number of neurons that express galanin is higher in monkeys. The galanin-positive sensory neurons are mainly of small size and also contain SP and CGRP. Galanin-like immunoreactivity has also been localized in dorsal horn neurons, mainly in lamina II, where it coexists with γ-aminobutyric acid (GABA), enkephalin, and neuropeptide Y.

Behavioral and electrophysiological studies have shown that galanin produces complex effects with a predominantly inhibitory action on spinal nociception. Electrophysiological studies using the flexor reflex model suggest that galanin may be particularly effective in reducing central sensitization through neuronal hyperpolarization and in blocking the facilitatory effect of SP and CGRP. However, spinal administration of galanin also produces an excitatory effect on nociception under some circumstances. Recent evidence suggests that the complex effects of galanin may be mediated by subtypes of galanin receptors, with R1 receptor being inhibitory and R2 receptor being excitatory.

The inhibitory effect of galanin on dorsal horn neuronal activity and nociceptive reflexes is increased after nerve injury, and intrathecal application of galanin alleviated neuropathic painlike behaviors in rodents. Galanin may also be tonically active in suppressing painful input from injured sensory fibers. Chronic intrathecal infusion of galanin receptor antagonist increases autotomy behaviors in axotomized rats, representing a putative animal model of neuropathic pain. Furthermore, application of galanin antisense oligonucleotide to the distal stump of the transected sciatic nerve effectively reduces the upregulation of galanin in sensory neurons and exaggerates autotomy behavior. In the Bennett nerve constriction model, it has been reported that the extent of galanin upregulation is inversely correlated to the severity of painlike behavior among individual rats. An inhibitory effect of galanin on nociception after nerve injured has also been supported by studies in mice overexpressing galanin or lacking the R1 galanin receptors. Thus, agonists of galanin receptors, possibly of the R1 subtype, deserve to be further evaluated for their analgesic potential for treating neuropathic pain.

Neuropeptide Y

In normal conditions, NPY is rarely present in DRG neurons in rats, whereas in the dorsal horn it is found in interneurons in the superficial laminae, where it coexists with GABA. Peripheral nerve lesion induces a dramatic increase in the expression of NPY in the DRG, mainly in the medium/large-sized neurons.

Similar to galanin, NPY also produces a complex effect on spinal pain mechanisms, with predominantly inhibitory action. It has been shown that the inhibitory effect of NPY is mainly mediated by the activation of the Y1 receptor in normal rats, but by both Y1 and Y2 receptors after nerve injury. On the other hand, NPY has been shown at the level of the nucleus gracilis to mediate tactile hypersensitivity after nerve injury, in accordance with an upregulation of NPY in large fibers.

Opioid Peptides

The opioid peptides (endorphins, enkephalins, dynorphins, endomorphins, and nociceptin) are present mainly in dorsal horn interneurons, but not in sensory neurons. Enkephalins and endorphins are well-established mediators of the endogenous antinociceptive system activated by placebo treatment, chronic pain, stress, acupuncture, etc. The action of endomorphins and nociceptin, by activation of μ and opioid receptor-like (ORL) receptors, is also primarily inhibitory in the spinal cord, whereas dynorphins appear to produce a more complex effect in pain modulation.

Cholecystokinin and Other Antiopioid Peptides

It is well established that the antinociceptive and/or analgesic effect of opioids can be reduced by a range of nonopioid peptides, including cholecystokinin (CCK), Phe-Met-Arg-Phe-NH_2 (FMRF-amide)-related peptides, angiotensin II, and melanocyte-inhibiting factor (MIF)-related peptides. Among these, the antiopioid effect of CCK is best studied. In general, there is an overlap in the anatomical distribution of CCK,

opioids, and their receptors in the spinal cord and brain; this overlap may underlie the documented interaction between the opioid and CCK systems. In the spinal cord, CCK is found in dorsal horn interneurons but not in sensory neurons in the DRG, and only a small number of rat DRG neurons express CCK-B receptor mRNA under normal conditions. After nerve injury, the level of CCK and CCK-B receptor mRNA is upregulated in the DRG.

A large body of literature has accumulated that indicates that upon local, systemic, intrathecal, or intracerebral injection, CCK reduces the effect of exogenous as well as endogenous opioids. The inhibition by CCK on opioid antinociception appears to be tonic, as blockade of endogenous CCK by receptor antagonists results in an enhancement of opioid-induced antinociception. Importantly, CCK may be involved in determining opioid sensitivity under pathological conditions. Thus, the antinociceptive effect of opiates is enhanced in animals with acute inflammation, which is associated with a decreased release of CCK. In contrast, opiates have reduced effect after nerve injury, which has been shown to be related to enhanced activity of the endogenous CCK system. Furthermore, repeated administration of opiates induces a gradual reduction in their ability to induce analgesia, a condition known as tolerance, which is also related to the CCK system. Hence, blockade of CCK receptors may have potential clinical applications in pain management by increasing the effect of opioids and reducing tolerance.

See also: Calcitonin Gene-Related Peptide (CGRP) and Receptors; Galanin and Receptors; Neuropeptide Y (NPY) and its Receptors; Neuropeptides: Sensory Systems; Nociceptor Responses; Opioid Peptides and Receptors; Somatostatin and Receptors; Substance P/Tachykinins and its/their Receptors; Vasoactive Intestinal Peptide and Pituitary Adenylate Cyclase Activating Peptide Receptors.

Further Reading

Dickenson T and Fleetwood-Walker SM (1999) VIP and PACAP, very important in pain. *Trends in Pharmacological Sciences* 20: 324–329.

Drey A and Urban L (1996) New pharmacological strategies for pain relief. *Annual Review of Pharmacology and Toxicology* 36: 253–280.

Hill RG (2001) NK1 (substance P) receptor antagonists – why are they not analgesic in humans? *Trends in Pharmacological Sciences* 21: 244–246.

Hökfelt T, Zhang X, and Wiesenfeld-Hallin Z (1994) Messenger plasticity in primary sensory neurons following axotomy and its functional implications. *Trends in Neuroscience* 17: 22–30.

Hökfelt T, Pernow B, and Wahren J (2001) Substance P: A pioneer amongst neuropeptides. *Journal of Internal Medicine* 249: 27–40.

Hökfelt T, Zhang X, Xu X-J, et al. (2006) Central consequences of peripheral nerve damage. In: McMahon S and Koltzenburg M (eds.) *Wall and Melzack's Textbook of Pain*, pp. 947–960. London: Elsevier/Churchill Livingstone.

McMahon SB and Koltzenburg M (eds.) (2006) *Wall and Melzack's Textbook of Pain,* 5th edn. London: Elsevier/Churchill Livingstone.

Millan MJ (1999) The induction of pain: An integrative review. *Progress in Neurobiology* 57: 1–164.

Mogil JS, Miermeister F, Seifert F, et al. (2005) Variable sensitivity to noxious heat is mediated by differential expression of the CGRP gene. *Proceedings of the National Academy of Sciences of the United States of America* 102: 12938–12943.

Sawynok J and Cowan A (eds.) (1999) *Novel Aspects of Pain Management: Opioids and Beyond*. New York: Wiley-Liss.

Willis WD and Coggeshall RE (2004) *Sensory Mechanisms of the Spinal Cord*. Berlin: Springer.

Xu X-J, Hökfelt T, Bartfai T, et al. (2000) Galanin and spinal nociceptive mechanisms: Recent advances and therapeutic implications. *Neuropeptides* 34: 137–147.

Neuropeptides: Sensory Systems

J V Priestley, Queen Mary University of London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The somatosensory system carries information from the body surface (skin), viscera, muscles, and joints. Specialized neurons, referred to as primary afferent or primary sensory, respond to events that arise from outside the body or inside the body and convey information to second-order neurons within the spinal cord and brain stem. All primary sensory neurons use the fast-acting transmitter glutamate, but, remarkably, many also contain multiple neuropeptide transmitters. This is particularly true of the subpopulation of neurons that respond to tissue damage, known as nociceptors. These neuropeptides function as neurotransmitters within the spinal cord, conveying information about both acute pain and the chronic pain associated with nerve injury and inflammation. In addition, they act as neurohumeral factors in the periphery, regulating blood flow and the tissue response to injury. This article briefly reviews the organization and subtypes of primary sensory neurons, their regulation by growth factors, the changes that take place after nerve injury and inflammation, and the functional role of primary afferent neuropeptides.

Biochemical Characterization and Anatomy of Primary Sensory Neurons

The majority of the body surface and most muscles and joints are supplied by primary sensory neurons whose cell bodies are located in ganglia just lateral to the spinal cord, referred to as spinal or dorsal root ganglia (DRGs). The face, and the muscles of facial expression, is supplied by primary sensory neurons in the trigeminal and petrosal (glossopharyngeal) cranial ganglia; however, their properties are generally similar to those of DRG neurons. This article therefore focuses mainly on the DRG neurons. However, many viscera receive innervation not only from DRG neurons but also from specialized primary sensory neurons located in the nodose and jugular ganglia. Those neurons have distinct properties and are not considered in detail. Most information comes from studies in rat, but similar principles seem to apply to other species, although there are many differences in detail.

DRG neurons do not possess dendrites, but have an axon that divides into a peripheral branch which has receptor, and possibly also effector, properties and a central branch that synapses with second-order neurons within the spinal cord (Figure 1). They can be divided into various subpopulations based on their neurochemistry (Figure 2), anatomy, and physiology. Large and medium-size neurons (25–60 μm in diameter) form a group that can be distinguished by their high content of phosphorylated heavy-chain (200 kDa) neurofilament and that make up approximately 40% of the lumbar DRG neurons. These cells have mostly myelinated axons that conduct in the Aα/β range (12–31 m s^{-1}, rat lumbar DRG neurons), and their peripheral axons have specialized encapsulated endings and are mainly low-threshold mechanoreceptors. Many subtypes exist, defined according to their stimulation and adaptation properties, such as Pacinian corpuscles and Merkel touch domes in the skin, and Ia spindle afferents in muscle. Aβ high-threshold mechanoreceptors (nociceptors) also exist, but are less common than Aδ and C fiber nociceptors.

Nociceptive Dorsal Root Ganglion Neurons

The remaining DRG neurons are small in size (10–30 μm), express the intermediate filament peripherin, have unmyelinated axons (C fibers, <1.3 m s^{-1} conduction velocity), and have unencapsulated nerve endings that function as nociceptors and/or thermoreceptors. Several subtypes exist and include low-threshold mechanoreceptors (e.g., 12% of rat saphenous C fibers), warming receptors and cooling receptors, and silent receptors (11%, insensitive or unexcitable). However, most C fibers (70%) are polymodal nociceptors and respond to noxious mechanical, thermal, and chemical stimuli. This property is largely due to the fact that their peripheral terminals express transient receptor potential (TRPV)1, a nonselective cation channel that is a member of the vanilloid subfamily of TRPV ion channels. TRPV1 has been described as a molecular integrator of pain-producing stimuli. It is activated by temperatures above 43 °C; by molecules that possess a vanilloid moiety (e.g., capsaicin, the pungent component of capsicum peppers); and by a range of factors associated with tissue damage, including the endocannabinoid anandamide, the leukotriene hydroperoxyeicosatetraenoic acid (HPETE), and low pH (<5.9). Another TRPV member, TRPV2, responds to noxious heat (>52 °C) and is an important receptor expressed by finely myelinated Aδ mechano-heat nociceptors (type I AMHs, also known as Aδ high-threshold mechanoreceptors). In addition, nociceptive DRG neurons express a range of receptors for

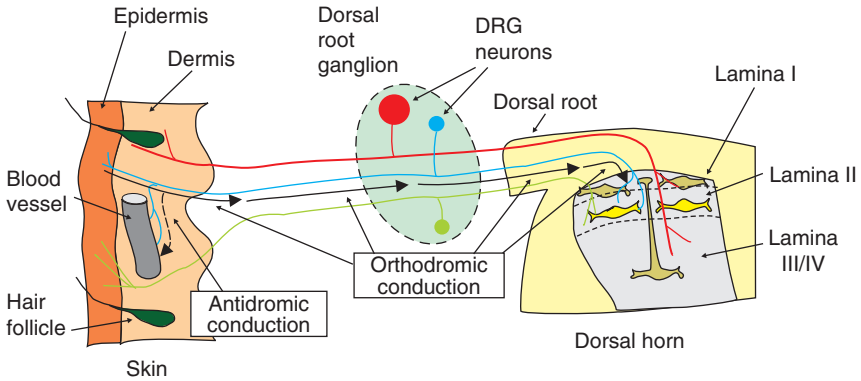

Figure 1 Diagram showing afferent and efferent properties of a typical cutaneous DRG neuron. The cell bodies of DRG neurons are located in dorsal root ganglia, and each neuron has an axon which divides to form a peripheral and a central branch. The peripheral branch runs in a peripheral nerve and forms specialized receptors in the dermis and epidermis of the skin. The central branch runs in a dorsal root and synapses with second-order neurons in the dorsal horn of the spinal cord. The normal flow of information and direction of action potential propagation is from the periphery to the spinal cord (orthodromic) and, in the case of DRG neurons that use neuropeptides, the neuropeptide is released at the central synapses and acts as a neurotransmitter. However, some DRG neurons (shown in blue) have collateral branches that innervate blood vessels. In their case, action potentials can also be conducted antidromically and neuropeptides may be released from their peripheral terminals to control vasculature. DRG, dorsal root ganglion.

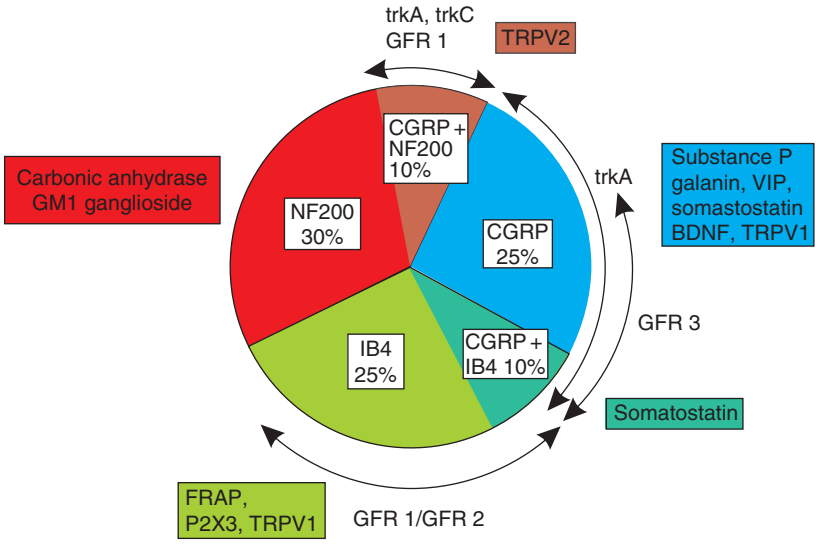

Figure 2 Pie chart summarizing the main neurochemical populations of DRG neurons. One population comprises large cells which give rise to myelinated axons and have high levels of neurofilament (NF200). Small cells, which have mainly unmyelinated axons and are predominantly nociceptors, comprise two populations. One population consists of cells that constitutively synthesize the neuropeptide CGRP. The other population is characterized by cells that bind the lectin *Griffonia simplicifolia* IB4. An overlap between the NF200 and CGRP populations corresponds to Aδ nociceptors. There is also some overlap between the IB4 and CGRP populations. Neurotrophin receptors and GDNF receptors expressed by small and medium-size cells (predominantly nociceptors) are indicated around the periphery of the chart, together with molecules found in each major population. The percentages reflect the situation in lumbar (L4 and L5) ganglia of the rat. Similar populations exist at other segmental levels (cervical, thoracic, and sacral), but the exact percentages vary. BDNF, brain-derived neurotrophic factor; CGRP, calcitonin gene-related peptide; DRG, dorsal root ganglion; FRAP, fluoride-resistant acid phosphatase; GDNF, glial cell line-derived neurotrophic factor; GFR, glial cell line-derived neurotrophic factor receptor; IB4, isolectin B4; TRPV, transient receptor potential. Adapted from McMahon SB and Priestley JV (2005) Nociceptor plasticity. In: Hunt SP and Koltzenburg M (eds.) *The Neurobiology of Pain*, pp. 35–64. Oxford: Oxford University Press.

inflammatory mediators and/or products of tissue damage, including purinoceptors, bradykinin receptors, serotonin (5-hydroxytryptamine (5-HT)) receptors, acid-sensing ion channels (ASICs), the CB1 cannabinoid receptor, the H1 histamine receptor, the transient receptor potential A1 (TRPA1) receptor, and prostanoid receptors. Some of these receptors are, like TRPV1, ion channels that lead to the direct activation of nociceptors by causing membrane depolarization and the firing of action potentials. Others, such as the

bradykinin B2 receptor, are G-protein-coupled receptors (GPCRs) that modulate TRPV1 and other ion channels via second messengers. Nociceptive DRG neurons also possess specialized voltage-gated sodium channels, which generate action potentials with distinctive propagation and firing properties. Such channels include $Na_V1.7$ (tetrodotoxin (TTX)-sensitive sodium current), $Na_V1.8$ (sensory neuron specific (SNS), TTX-resistant) and $Na_V1.9$ (sodium channel NaN, TTX-resistant).

Peptidergic and Nonpeptidergic Nociceptors

In addition to the functional diversity outlined so far, neurochemical studies have revealed that small DRG neurons can be divided into two main subgroups depending on whether or not they constitutively express neuropeptides. Just over half of all small DRG neurons (50–60%) constitutively express the neuropeptide calcitonin gene-related peptide (CGRP) and may also express other neuropeptides, including substance P (a tachykinin derived from preprotachykinin-A), somatostatin, vasoactive intestinal polypeptide (VIP) and galanin (Table 1), and the neurotrophin brain-derived neurotrophic factor (BDNF). Many of these neuropeptides coexist not only within single DRG neurons but also within individual dense-core vesicles. Histochemical and functional studies have shown that the majority of CGRP neurons (59–65%) possess TRPV1 and are likely to be polymodal nociceptors. Neuropeptides are generally not found in the large-diameter mechanoreceptive cells, but CGRP is present in many medium-size 200 kDa neurofilament-rich neurons which are also mainly nociceptors (AMHs). The other half of the small DRG neurons either do not contain neuropeptides or contain only low levels that can be identified using a variety of markers. These include staining by the lectin *Griffonia simplicifolia* isolectin B4 (IB4), expression of a nonlysosomal fluoride-resistant acid phosphatase (FRAP), expression of the purinoceptor P2X3, and expression of a subtype of sensory neuron-specific GPCR referred to as Mas-related G-protein-coupled receptor subtype d (Mrgprd). The majority of these cells also express TRPV1 (75% of IB4 cells), and thus polymodal nociceptors appear to have two subtypes: one which constitutively expresses neuropeptides and one which does not. There is good evidence that the neuropeptides serve as neurotransmitters and/or neuromodulators (see the section titled 'Function of sensory neuropeptides'), and so I refer to these two populations as peptidergic and nonpeptidergic. Note that the thermoreceptors have not been characterized as thoroughly as the nociceptors and that it is not known whether they also comprise peptidergic and nonpeptidergic populations. It is also not known whether such a dichotomy applies to visceral afferents. Many nodose and jugular ganglion neurons are peptidergic and express TRPV1. In addition, there are many P2X3/IB4 neurons in the nodose ganglion which are nonpeptidergic. However, at least in the lung, these appear to be myelinated mechanoreceptors and not capsaicin-sensitive.

Table 1 Neuropeptides synthesized by small and medium-size DRG neurons[a]

	DRG cells (%)
Calcitonin gene-related peptide (CGRP)	32–50
Substance P and neurokinin A	18–38
Cholecystokinin (CCK)	0–3
Galanin	0–5
Somatostatin	5–15
Vasoactive intestinal polypeptide (VIP) and Peptide histidine isoleucine (PHI)	0–5 (mainly in visceral afferents)
Neuropeptide Y	0
Bombesin/gastrin-releasing peptide	0–8
Dynorphin	0–3
Enkephalin	0–3

[a]DRG, dorsal root ganglion.

From the material reviewed so far, it can be seen that neuropeptides in sensory neurons are predominantly associated with nociceptors but that not all nociceptors express neuropeptides. Nociceptive DRG neurons can basically be divided into three main populations (Figure 3): (1) medium-diameter TRPV2-expressing peptidergic (CGRP) cells that have finely myelinated (Aδ) axons; (2) small-diameter TRPV1-expressing peptidergic (CGRP and substance P) cells that give rise to unmyelinated axons (C fibers); and (3) small-diameter TRPV1- and P2X3-expressing and IB4-binding nonpeptidergic cells that give rise to unmyelinated axons (C fibers). Among cutaneous afferents, all three nociceptive populations have peripheral processes that form free nerve endings in the epidermis, but peptidergic C fibers (population 2) terminate more superficially than nonpeptidergic axons and also give rise to perivascular endings in epidermis and dermis. This population also has proportionally more visceral projections than the nonpeptidergic DRG neurons (population 3). The central projections of these three populations have slightly different termination patterns. The peptidergic C fibers terminate predominantly in the lamina I and outer lamina II (IIo), nonpeptidergic axons terminate mainly in the inner lamina II (IIi), and Aδ fibers terminate predominantly in lamina I and laminae III–V.

The significance of having two populations of polymodal nociceptors (peptidergic and nonpeptidergic) is not known, but one possibility is that they signal different types of pain (e.g., acute and chronic). It is also possible that each population targets distinct central pathways. For example, recent studies

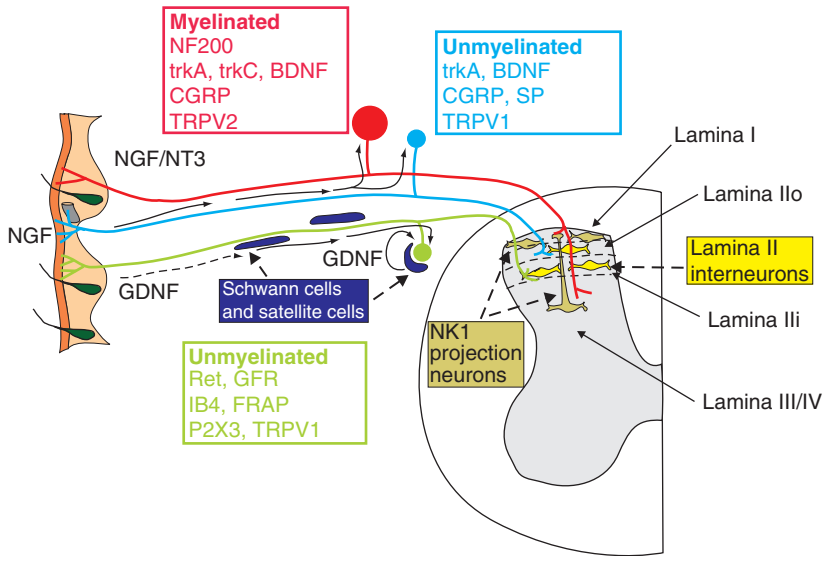

Figure 3 Diagram showing the three main populations of nociceptors: (1) medium-diameter peptidergic (CGRP) cells (red) that have finely myelinated axons that function as mechano-heat nociceptors (AMHs), respond to NGF and NT3, and innervate predominantly laminae I and III/IV; (2) small-diameter peptidergic (CGRP and substance P) cells (blue) that function as polymodal nociceptors, respond to NGF, and innervate predominantly laminae I and II outer (IIo); and (3) small-diameter IB4-binding nonpeptidergic cells (green) that function as polymodal nociceptors, respond to GDNF, and innervate predominantly lamina II inner (IIi). BDNF, brain-derived neurotrophic factor; CGRP, calcitonin gene-related peptide; FRAP, fluoride-resistant acid phosphatase; GDNF, glial cell line-derived neurotrophic factor; GFR, glial cell line-derived neurotrophic factor receptor; IB4, isolectin B4; NF200, 200 kDa neurofilament; NGF, nerve growth factor; NK1, neurokinin 1; NT3, neurotrophin 3; SP, substance P; TRPV, transient receptor potential. Adapted from McMahon SB and Priestley JV (2005) Nociceptor plasticity. In: Hunt SP and Koltzenburg M (eds.) *The Neurobiology of Pain*, pp. 35–64. Oxford: Oxford University Press.

indicate that peptidergic C fibers innervate dorsal horn neurons that project to traditional central nervous system (CNS) pain relay sites, such as the parabrachial nucleus and the ventrolateral thalamus, whereas nonpeptidergic C fibers may activate neurons that project to limbic and striatal sites including the bed nucleus of the stria terminalis and the globus pallidus. Thus, the nonpeptidergic cells may innervate pathways that convey primarily the affective component of pain rather than the sensory-discriminatory component and may also activate brain centers that control motor responses to pain.

Target-Derived and Injury-Derived Growth Factors

The organization reviewed so far is true of the normal, healthy nervous system, but major changes take place after tissue or nerve injury, particularly as regards neuropeptide expression. It appears that both the normal phenotypes of DRG cells and the changes after injury are regulated by growth factors that arise either from peripheral tissues (target-derived) or from the local glial and immune response to injury (injury-derived). This regulation will now be briefly reviewed because it helps to explain the neurochemical subtypes (see the section titled 'Biochemical characterization and anatomy of primary sensory neurons') and the changes after tissue or nerve injury (see the section titled 'Neuropeptide plasticity following inflammation or nerve injury').

Target-Derived Neurotrophins

During development, the survival of DRG neurons depends on nerve growth factor (NGF), which binds to a receptor tyrosine kinase called trkA, and neurotrophin (NT)-3, which binds to trkC; and virtually all DRG cells express trkA or trkC. In the adult, DRG neurons no longer require neurotrophins for survival, and there are postnatal changes that establish a new pattern of receptor expression (Figure 2). The peptidergic population of polymodal nociceptors expresses predominantly trkA, and low-threshold mechanoreceptors express either trkB the receptor for BDNF or trkC. AMH nociceptors express both trkA and trkC. All DRG neurons that express trk receptors also express p75, a nonselective neurotrophin receptor that is a member of the tumor necrosis factor (TNF) receptor superfamily. p75 acts as a trk coreceptor, but can also signal directly by recruitment of various adaptor proteins. The expression of trkA by the peptidergic population of nociceptors suggests that this population responds to NGF, and this is borne out by experimental studies in which animals were treated

with exogenous NGF. In healthy animals, NGF has a range of direct effects that are mediated by trkA, including an acute sensitization leading to thermal hyperalgesia and a longer-lasting effect on gene expression. The chronic effects are mediated by the internalization of trkA and retrograde transport to the cell body, where the activation of the mitogen-activated protein kinase (MAPK) pathway leads to induction of transcription factors such as c-Myc, Elk-1, c-Fos, and c-Jun and either positive or negative regulation of a large number of different target genes. Positively regulated genes have their expression increased by NGF and include substance P, CGRP, BDNF, TRPV1, and P2X3. Negatively regulated genes have their expression suppressed by NGF and include galanin and activating transcription factor (ATF)3. Regulation by endogenous NGF is most evident when there are large changes in NGF levels, such as occurs in nerve injury or inflammation (see the section titled 'Neuropeptide plasticity following inflammation or nerve injury'). However, regulation can also be seen in normal healthy animals, indicating that the adult phenotype of the peptidergic nociceptors is maintained by NGF derived from cells in peripheral targets, such as keratinocytes in the epidermis of the skin. This interpretation is supported by studies in which NGF has been sequestered or blocked in normal animals. This produces a thermal and chemical hypoalgesia; a reduction in the sensitivity of individual nociceptors; a downregulation of substance P, CGRP, and $Na_V1.8$; and a reduction in the innervation density of the epidermis. NGF blockade also results in an upregulation of injury-associated markers such as c-Jun, VIP, and galanin (see the section titled 'Inflammation').

Glial Cell Line-Derived Neurotrophic Factor and Cytokines

NGF is the most thoroughly studied growth factor that regulates neuropeptide expression in DRG neurons, but it is not the only one. Two other important families are the glial cell line-derived neurotrophic factor (GDNF) family and the neuropoietic cytokines. Low-threshold mechanoreceptors and the nonpeptidergic population of nociceptors respond to GDNF and express GDNF receptor components such as the tyrosine kinase RET and the accessory subunits GFRα1, GFRα2, and GFRα3. GDNF treatment *in vivo* has acute effects on DRG neurons that are due to sensitization of TRPV1. However, GDNF also increases the expression of proteins such as P2X3 and somatostatin in nonpeptidergic cells and protects nonpeptidergic cells from some of the effects of axotomy (see the section titled 'Nerve injury'). The peptidergic C fiber nociceptors do not express RET, but some do express GFRα3 and this expression increases after nerve injury. GDNF is thought to regulate nonpeptidergic cells as a targetderived factor, but a more important role is probably as a local injury-derived factor. GDNF is synthesized by Schwann cells and by DRG satellite glial cells, and synthesis is increased further by axotomy. The neuropoietic cytokines can also be classed as injury-derived factors, produced by resident and infiltrating inflammatory cells. Their members include ciliary neurotrophic factor (CNTF), oncostatin M (OM), leukemia inhibitory factor (LIF), and interleukin (IL)-6. Galanin, which is strongly upregulated after nerve injury, is regulated by LIF and IL-6.

Neuropeptide Plasticity Following Inflammation or Nerve Injury

The organization and neurochemistry of DRG neurons are not stable, and major changes take place in response to alterations in levels of growth factors and cytokines. This is seen most dramatically following tissue inflammation and nerve injury.

Inflammation

In inflammatory conditions, such as rheumatoid arthritis or cystitis, the peripheral terminals of DRG neurons become exposed to a complex cocktail of algesic chemicals released from injured cells, from resident mast cells, and from infiltrating macrophages and lymphocytes (Figure 4). This inflammatory milieu activates ligand-gated ion channels (e.g., TRPV1, activated by a fall in tissue pH) and GPCRs (e.g., histamine H1, bradykinin B1 and B2, and P2X3) and leads to an acute activation and sensitization that is evident as a thermal hyperalgesia. However, growth factors and cytokines are also released from Schwann cells, keratinocytes, and inflammatory cells and act on the receptors we reviewed previously to produce both acute and chronic effects. For example, NGF contributes to the acute sensitization of peptidergic nociceptors by a pathway that involves the activation of phospholipase C by trkA and the activation of multiple downstream pathways, including the phosphorylation of TRPV1 by protein kinase C. In addition, internalization and retrograde transport of trkA drives long-term changes in genes related to many aspects of DRG function, including neurotransmitters and neuropeptides, receptors, ion channels, and cytoskeletal proteins. Increased expression of TRPV1 and bradykinin B1/B2 receptors, for example, maintain the sensitization of peptidergic nociceptors, whereas increases in substance P and CGRP expression augment transmission in the dorsal horn (see the section titled 'The role of sensory neuropeptides in the spinal

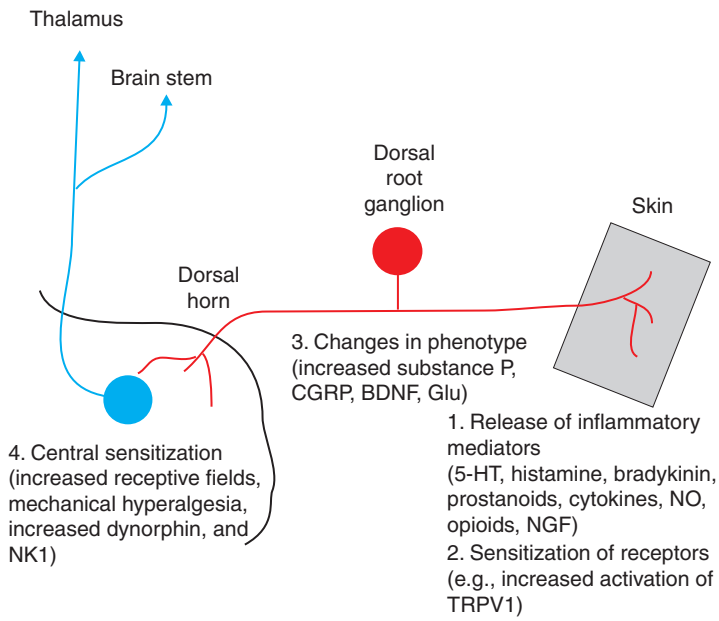

Figure 4 Diagram summarizing the changes that occur in nociceptors as a consequence of tissue inflammation. Inflammatory mediators act on the peripheral endings of primary sensory neurons to produce both acute sensitization and chronic changes in the phenotype of DRG neurons. These chronic changes include increased expression of neuropeptides (substance P and CGRP) and of the neurotrophin BDNF. The changes in the primary sensory neurons lead to secondary changes in the dorsal horn (central sensitization). 5-HT, serotonin; BDNF, brain-derived neurotrophic factor; CGRP, calcitonin gene-related peptide; DRG, dorsal root ganglion; Glu, glutamate; NGF, nerve growth factor; NK1, neurokinin 1; NO, nitric oxide; TRPV, transient receptor potential.

cord') and contribute to vasodilatation and plasma extravasation via their peripheral release (see the section titled 'The role of peripherally released sensory neuropeptides').

Nerve Injury

Following nerve injury, there are even greater changes in gene expression by DRG neurons than are seen in inflammation. For example, microarray studies showed that the expression of 240 different genes change after axotomy. Many of the changes that take place are opposite to those seen in inflammation and due not to the increased production of growth factors in the periphery but to the axotomy causing a disconnection of DRG neurons from their peripheral targets and consequently a reduction in supply of growth factors. Here I focus on changes in such axotomized neurons (Figure 5). However, note that nerve injury often leaves some DRG neurons whose axons remain intact but which intermingle with those degenerating in the distal nerve and target tissue. These intact axons may actually be exposed to increased levels of growth factors and show changes that are opposite to those of their axotomized neighbors. This is partly because Schwann cells and inflammatory cells respond to the axotomy by increasing growth factor production; it is also because target-derived trophic factors are usually made in limiting amounts. When a significant number of sensory neurons lose contact with their target, the remaining innervation has less competition for these factors. The overall effects of a nerve injury therefore depends on the proportions of axotomized versus intact axons.

Many of the neuropeptides constitutively expressed by peptidergic nociceptors, such as substance P, CGRP, somatostatin, and BDNF, are downregulated in axotomized DRG neurons (Figure 5). These changes develop over a few days and persist for many months, and they appear to be due to loss of the NGF that normally positively regulates neuropeptide expression. Nonpeptidergic cells are also affected by axotomy and, in response to the loss of GDNF, downregulate IB4 binding sites and FRAP. Treatment of the damaged nerves with NGF prevents the downregulation in the peptidergic population, and GDNF prevents it in the nonpeptidergic population. In addition, there are increases in gene expression that appear to be due either to loss of a target-derived factor that negatively regulates neuropeptide expression or to the local production of an injury-derived factor. For example, galanin is dramatically upregulated in a large number of axotomized DRG neurons, including both the peptidergic and nonpeptidergic nociceptive populations and many of the large cells.

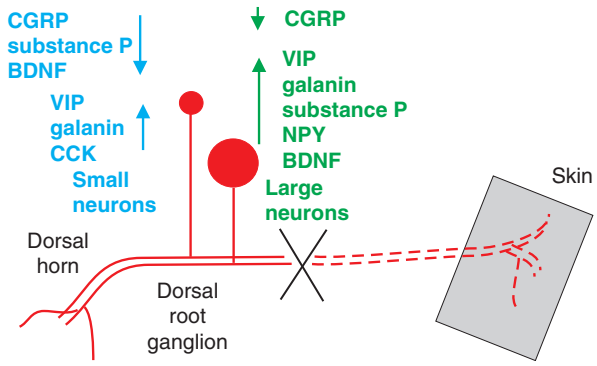

Figure 5 Diagram summarizing the range of phenotypic changes that take place in DRG neurons that are axotomized as a consequence of a peripheral nerve injury. The neuropeptides substance P and CGRP and the neurotrophin BDNF are downregulated in small-diameter (predominantly nociceptive) DRG neurons, whereas CCK is upregulated in small-diameter DRG neurons and substance P, NPY, and BDNF are upregulated in large-diameter DRG neurons. The neuropeptides VIP and galanin are upregulated in both small and large neurons. Note that a different set of changes occurs in DRG neurons, whose axons remain intact after a partial nerve injury. BDNF, brain-derived neurotrophic factor; CCK, cholecystokinin; CGRP, calcitonin gene-related peptide; DRG, dorsal root ganglion; NPY, neuropeptide Y; VIP, vasoactive intestinal peptide.

This upregulation starts very rapidly, within a day and well before the downregulation of neuropeptides can be detected. Experimental studies indicate that the galanin upregulation is partly due to the loss of target-derived NGF and partly due to gain of IL-6 or LIF. Thus, NGF treatment can partially offset galanin upregulation, and NGF deprivation has been reported to induce it. In addition, the application of LIF or IL-6 induces galanin expression.

Function of Sensory Neuropeptides

DRG neurons are capable of releasing neuropeptides from both their central and peripheral processes (Figure 1), and at the two sites the neuropeptides have quite different roles. There is also evidence that neuropeptides can be released from DRG cell bodies, but this has been much less studied and its significance is unclear.

The Role of Sensory Neuropeptides in the Spinal Cord

As already reviewed, the central axons of peptide-containing C fibers terminate mainly in laminae I and II of the dorsal horn, where they make synapses on second-order neurons. In this location, neuropeptides are thought to act as neurotransmitters or neuromodulators, coexisting with the fast-acting neurotransmitter glutamate. Thus, they are released from terminals in response to depolarization by an action potential, and they act on GPCRs to cause slow excitatory postsynaptic potentials (EPSPs) or inhibitory postsynaptic potentials (IPSPs). However, unlike glutamate, which is stored in agranular vesicles and released exclusively into the synaptic cleft to act predominantly on *N*-methyl-D-aspartate (NMDA) and α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors concentrated at the postsynaptic density, neuropeptides are thought to be released also from extrasynaptic sites and to diffuse locally (referred to as paracrine or volume transmission). Thus, their site of action is determined more by the distribution of their receptors than by the synaptic circuitry, and the anatomical range of their action is determined by the quantity of neuropeptide released and the presence of extracellular peptidases (which degrade the neuropeptide and hence limit diffusion). In addition, neuropeptides differ from glutamate in requiring a higher frequency of action potentials to evoke their release.

The main neuropeptides that are constitutively expressed by C fiber nociceptors are CGRP and substance P, and of these the most fully characterized as a dorsal horn neurotransmitter is substance P. Thus, many spinothalamic and spinobulbar projection neurons in lamina I are innervated by substance P-containing primary afferents and possess the substance P (neurokinin (NK)1) receptor. Additional projection neurons are located in laminae IV and V, but have dendrites that extend ventrally into laminae I and IIo, where they are innervated b substance P-containing primary afferents (Figure 3). Substance P is released on high-frequency C fiber stimulation, binds to and causes internalization of NK1 within projection neurons, and leads to the activation of inositol triphosphate (IP_3)- and protein kinase A-mediated signaling pathways. This leads to the generation of slow EPSPs, possibly via phosphorylation and inhibition of potassium channels, but it can also activate the transcription factor c-fos and lead to changes in gene transcription. Although primary afferent-evoked and substance P-mediated EPSPs can be detected in the normal dorsal horn, it appears that substance P does not make a significant contribution to transmission in pain pathways until the substance P levels are increased and/or the primary afferent firing is increased, such as occurs with inflammation (see the section titled 'Inflammation'). The strongest evidence for this comes from studies of preprotachykinin-A and NK1 knockout mice, both of which show normal thresholds for acute pain but reduced responses to intense noxious stimuli. In inflammation, increased peripheral NGF sensitizes peptidergic nociceptors

and increases substance P synthesis, both of which lead to increased substance P release from central terminals. This, in turn, leads to acute and chronic changes in the activation of second-order neurons and contributes to central sensitization, a state in which spinal cord neurons become more excitable (Figure 4). Acute changes include membrane depolarization, which leads to further depolarization via the release of Mg^{2+} block of NMDA receptors and the activation of voltage-gated Ca^{2+} channels. Chronic changes include transcriptionally mediated increases in the expression of the NK1 receptor and of neuropeptides expressed by dorsal horn neurons, such as dynorphin. Dynorphin is thought to either suppress or facilitate the activity of nociceptive dorsal horn neurons, depending on the type of cells that express it. The neurotrophin BDNF, which is constitutively expressed in nociceptive C fibers and positively regulated by NGF, is thought to be co-released with substance P and CGRP and to also contribute to central sensitization. Thus BDNF, via its receptor trkB, causes protein kinase-dependent phosphorylation of NMDA and AMPA receptors, as well as long-term changes in gene transcription.

The changes that take place after nerve injury also lead to profound changes in neurotransmission within the dorsal horn. Upregulation of substance P, CGRP, and BDNF in spared afferents can lead to central sensitization by similar mechanisms to those already reviewed. In addition, the dramatic changes in peptide expression within damaged primary afferents (Figure 5) mean that neuropeptides such as VIP, neuropeptide Y, and galanin take over from substance P and CGRP as the neurotransmitters used by injured afferents. Galanin, which is the neuropeptide that is most dramatically increased (approximately 120-fold), has predominantly inhibitory effects on spinal cord processing (mediated via receptors GalR1), and this has been proposed to reduce nociceptive transmission after nerve injury. Consistent with this interpretation, galanin overexpressing mice show reduced pain responses. However, galanin also has neurotrophic actions, and this may explain why galanin knockout mice, which have a loss of peptidergic nociceptors, show reduced pain in nerve-injury models. In addition, in some situations galanin may act at presynaptic receptors GalR2 to promote the release of primary afferent transmitters and increase pain transmission. Cholecystokinin (CCK) is also upregulated after nerve injury and, because CCK acts as a functional opioid antagonist, may contribute to the relative refractoriness of neuropathic pain to morphine analgesia.

The Role of Peripherally Released Sensory Neuropeptides

In addition to release from central terminals, CGRP and substance P are released from peripheral terminals, where they play a key role in the tissue response to injury in a process referred to as neurogenic inflammation. This can be seen, for example, in the triple response that accompanies injuries such as insect bites or nettle stings: a wheal (due to edema), local reddening (due to increased blood flow), and a flare (visible for several centimeters around the injury site due to a spread of the increased blood flow). CGRP, released from sensory endings, is thought to be the major vasodilator contributing to the increased blood flow (acting via receptor CGRP1), and substance P, released from sensory endings, is one of several mediators that increase vascular permeability and lead to plasma extravasation and edema. The substance P effects are both direct (via receptors NK1 on endothelial cells) and indirect (via the activation of mast cells and subsequent release of edema-inducing amines). Classically the triple response has been regarded as an axon reflex, due to the activation of nociceptors and antidromic propagation of action potentials leading to the release of neuropeptides from perivascular endings (Figure 1). However, there is evidence that neurogenic inflammation does not require propagation of action potentials, which suggests that sensory receptor activation and local peptide release are closely linked.

In some situations, peripherally released neuropeptides may contribute to tissue repair and the resolution of inflammation. Substance P acts as a mitogen for fibroblasts and endothelial cells and has been proposed as playing a role in angiogenesis. However, it is clear that sensitization of peptidergic primary afferents, and the associated neurogenic inflammation, can also contribute to chronic inflammatory conditions. This is the case, for example, in asthma. Sensory fibers participate in regulatory reflexes within the airways, and their sensitization increases the activation of parasympathetic efferent pathways, which in turn contribute to coughing, mucus hypersecretion, and bronchoconstriction. This is exacerbated by the antidromic release of substance P and neurokinin A (another tachykinin derived from preprotachykinin-A), which are potent inducers of airway smooth-muscle contraction, vasodilatation, bronchial edema, and mucus hypersecretion. Neuropeptides can also contribute to local inflammation by direct effects on inflammatory cells, in addition to promoting neutrophil and lymphocyte recruitment indirectly by their

effects on vasculature. For example, substance P enhances lymphocyte proliferation and immunoglobulin production, and it enhances cytokine secretion from lymphocytes, macrophages, and mast cells.

See also: Axonal Injury: Neuronal Responses; Axonal Injury in Demyelinating Disease and CNS Injury; Dorsal Root Ganglion Neurons; Inflammation in Neurodegenerative Disease and Injury; Neural Circuitry in the Somatosensory System; Neuropathic Pain; Neuropeptides: Pain; Nociceptor Responses.

Further Reading

Agnati LF, Zoli M, Stromberg I, and Fuxe K (1995) Intercellular communication in the brain: Wiring versus volume transmission. *Neuroscience* 69: 711–726.

Belvisi MG (2003) Airway sensory innervation as a target for novel therapies: An outdated concept? *Current Opinion in Pharmacology* 3: 239–243.

Brain SD (1997) Sensory neuropeptides: Their role in inflammation and wound healing. *Immunopharmacology* 37: 133–152.

Costigan M, Befort K, Karchewski L, et al. (2002) Replicate high-density rat genome oligonucleotide microarrays reveal hundreds of regulated genes in the dorsal root ganglion after peripheral nerve injury. *BMC Neuroscience* 3: 16.

Farinas I, Cano-Jaimez M, Bellmunt E, and Soriano M (2002) Regulation of neurogenesis by neurotrophins in developing spinal sensory ganglia. *Brain Research Bulletin* 57: 809–816.

Harrison S and Geppetti P (2001) Substance P. *International Journal of Biochemistry and Cell Biology* 33: 555–576.

Hokfelt T, Zhang X, Xu X, and Wiesenfeld-Hallin Z (2006) Central consequences of peripheral nerve damage. In: McMahon SB and Koltzenburg M (eds.) *Wall and Melzack's Textbook of Pain*, pp. 947–959. Edinburgh: Elsevier Churchill Livingstone.

Holmes FE, Mahoney SA, and Wynick D (2005) Use of genetically engineered transgenic mice to investigate the role of galanin in the peripheral nervous system after injury. *Neuropeptides* 39: 191–199.

Lawson SN (1992) Morphological and biochemical cell types of sensory neurons. In: Scott SA (ed.) *Sensory Neurons: Diversity, Development, and Plasticity*, pp. 27–59. New York: Oxford University Press.

McMahon SB and Priestley JV (2005) Nociceptor plasticity. In: Hunt SP and Koltzenburg M (eds.) *The Neurobiology of Pain*, pp. 35–64. Oxford: Oxford University Press.

O'Connor TM, O'Connell J, O'Brien DI, Goode T, Bredin CP, and Shanahan F (2004) The role of substance P in inflammatory disease. *Journal of Cell Physiology* 201: 167–180.

Priestley JV, Michael GJ, Averill S, Liu M, and Willmott NJ (2002) Regulation of nociceptive neurons by NGF and GDNF. *Canadian Journal of Physiology and Pharmacology* 80: 495–505.

Szolcsanyi J (2004) Forty years in capsaicin research for sensory pharmacology and physiology. *Neuropeptides* 38: 377–384.

Woolf CJ and Salter MW (2006) Plasticity and pain: Role of the dorsal horn. In: McMahon SB and Koltzenburg M (eds.) *Wall and Melzack's Textbook of Pain*, pp. 91–105. Edinburgh: Elsevier Churchill Livingstone.

Wynick D, Thompson SW, and McMahon SB (2001) The role of galanin as a multi-functional neuropeptide in the nervous system. *Current Opinion in Pharmacology* 1: 73–77.

Neurophysiology: Past and Present

S Jacyna, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term 'neurophysiology' is essentially a twentieth-century coinage. It was only after 1900, moreover, that the discipline acquired a distinct institutional identity, with appointments, university departments, and journals dedicated to its cultivation. Examples of the hyphenated version 'neuro-physiology' and the cognate term 'nerve physiology' can, however, be found in the nineteenth century. Despite the recent provenance of the name, it is nevertheless commonplace for those wishing to delineate the history of neurophysiology to venture far into the past.

Thus the American physiologist and medical historian John Farquhar Fulton (1899–1960), in 1933, declared that: "It is not difficult to trace the history of neurophysiology prior to 1885 for fundamental discoveries were few and can be enumerated on the fingers of two hands." Nonetheless, he then went on to enumerate the efforts of the ancient medical writer Galen (129–200 AD) in the field. Galen had, Fulton claimed, promulgated a theory of nervous conduction that was to last largely unmodified until the end of the eighteenth century. According to Fulton, there was then something of a lengthy hiatus in the history of neurophysiology until the early nineteenth century, when the doctrine of reflex action was formulated.

Underlying histories of this kind is the assumption that a thread of common endeavor is discernible in the history of Western science and medicine. Since antiquity individuals have – largely unwittingly – made contributions to a field that ultimately came to be known as neurophysiology. The fact that they might not have recognized this characterization of their activities, or would, more probably, have classified what they were doing under some other rubric, is seen as irrelevant. This 'discernible thread' assumption permits a curious kind of collegiality to arise between 'neurophysiologists' past and present. The efforts of earlier workers are to be judged in relation to currently accepted understandings of nervous function. They may be deemed to have contributed in some positive way, to have been mistaken, or even to have retarded the advance of knowledge. Thus the seventeenth-century English physician Thomas Willis' (1621–73) 'concepts of neurophysiology' may receive a mixed appraisal. His ideas on the localization of function in the brain are dismissed as 'fanciful,' and based on clumsy experimental technique and anatomical ignorance. On the other hand, Willis is acknowledged as "a pioneer in the development of our knowledge of autonomic function." He was, in fact, as Meyer and Hierons wrote, "very close to our modern ideas of the cerebral control of the autonomic system." One might note in passing that these comparisons are made in language that would have been alien to Willis. The term 'autonomic nervous system' was coined only in 1898 by JN Langley.

Modern historians of science are apt to dismiss historiography of this kind as naive or 'whigish.' On the other hand, imposing this kind of retrospective unity on a disparate body of writings from different periods and places does provide a rationale of sorts for historical investigation, especially for those who see themselves as heirs to this heritage. In order to understand fully our current knowledge of the workings of the nervous system, it is necessary to grasp how we attained it. The perceived failures, as well as successes, of previous generations of investigators may be instructive to those seeking to advance knowledge in the present. It should be noted that what might be characterized as 'practitioners' history' has, moreover, often served a range of polemical goals. Fulton, for instance, was, through historical exposition, keen to demonstrate the intimate relations between clinical and experimental work in advancing the field of neurophysiology. Others have written the history of a particular aspect of the subject with the view of denigrating certain theories, approaches, and conclusions while highlighting the advantages of others. Henry Head's highly partisan history of aphasia studies, published in 1926, comes to mind.

On the other hand, a 'presentist' approach to the history of neurophysiology – as of any other science – is subject to serious criticism. In particular, this perspective on the past runs the risk of seriously distorting the views of the authors under consideration by detaching them from their context. Thus it might be argued that the political and religious circumstances in which Thomas Willis worked, along with the clinical goals that he pursued, are of more relevance to understanding his work than the correctness or otherwise of his doctrines from a modern perspective. More generally, the endeavor to translate the terms in which the body has been understood in the past into a modern idiom runs the risk of obscuring the profound differences involved. Even when similar terms are in use, the meaning is often utterly different. This is true even of as basic a term as 'nerve.'

It is possible to strive for a middle way – that is, to try to chart the development of a modern neurophysiological concept while remaining sensitive to shifts of meaning and to context. An excellent example of such history is Roger Smith's account of the emergence of the concept of physiological inhibition over the past two centuries. In particular, Smith casts important light upon the interdependence of the various settings – both technical and lay – in which the meanings of inhibition evolved during this period.

In view of these issues, any attempt to establish the most salient points in the history of neurophysiology must be undertaken with great caution. This article is mostly confined to the period after 1800, on the grounds that the nineteenth and early twentieth centuries saw the emergence of new ways of understanding the structure and function of the nervous system that differ from those of previous epochs and yet appear to us as distinctively modern. With this in mind, the following discussion is organized under three headings: the reflex, microstructure, and localization of function.

The Reflex

The predominant theory of nervous action from the time of Galen until at least the middle of the eighteenth century revolved around the doctrine of 'animal spirits.' These subtle liquids were elaborated in the brain and then transmitted through the nerves. Their centripetal motion produced sensation; their centrifugal action caused muscular movement. One of the implications of the doctrine of animal spirits was thus that the brain was the sole central organ of the nervous system: it alone received impulses from the outside world, and it was only from the brain that bodily movement was initiated.

The observation that movements, often of a complex and coordinated kind, took place in decapitated animals led a number of authors to reconsider this doctrine. Some speculated about the existence of a 'spinal soul' that would continue to operate after the cerebral soul was detached. While the doctrine of the spinal soul retained advocates even in the nineteenth century, it was challenged by an alternative, purely mechanistic, explanation for the evident capacity of the spinal cord for functional autonomy. The figure most closely associated with this development was the English physician and experimentalist Marshall Hall (1790–1857). In a series of papers Hall argued for the existence of a separate "excito-motor system of nerves" located in the spinal marrow, which subserved reflex action. Hall originated the term 'reflex arc' to designate this mechanism. Hall insisted that the operations of this system, although apparently purposive, were entirely mechanical and involved no psychic component. Although at first controversial, Hall's views on the existence of reflex action were eventually to gain general recognition.

Hall insisted that reflex action was confined to the spinal cord and basal ganglia. However, in 1863 the Russian physiologist Ivan Mikhailov Sechenov (1829–1905) argued that the functions of the cerebral hemispheres could also be understood in terms of reflex action. This was at the time a bold assertion (attempts were made to censor Sechenov's publication), but one that laid the foundation for a unified conception of nervous function in its simplest as well as in its most complex manifestations. Sechenov's work inspired, among others, Ivan Pavlov (1849–1936), who maintained that reflexes could be acquired or conditioned as well as innate.

At the turn of the twentieth century another English investigator, Charles Scott Sherrington (1857–1952), was to elaborate the theory of the reflex in important ways. Sherrington placed the concept of the reflex within a more general biological context. In his monograph, *The Integrative Action of the Nervous System* (1906), he maintained that the simple reflex represented the most primitive instance of the general property of the nervous system to coordinate the body's responses to external and internal stimuli. More complex reactions were to be understood as compounds of such basic reflexes.

Microstructure and Function

The 1830s saw the promulgation of the theory that all bodily structures were composed of 'cells' and their processes. Jan Evangelista Purkyně (1787–1869) and his collaborator Gustav Gabriel Valentin had by 1836 declared that this theory applied to the nervous system, which was, they claimed, composed entirely of globules and fibers. In 1838 Robert Remak (1815–65) announced that in nervous tissues cells and fibers were continuous, the latter originating from the body of the former. By analogy with other organs it might be concluded that the cell body was the active portion of the nervous system, while the fibers performed a commissural function.

These insights, however, left open the issue of how the different nerve cells communicated. This question was to be the source of much controversy throughout the later nineteenth century. On the one hand there was the so-called reticularist theory, which maintained that the nerve cells formed a continuous network. The best known reticularist was the Italian histologist Camillo Golgi (1843–1926), who devised new staining techniques designed to reveal the microscopic structure of the nervous system with

unprecedented clarity. He nonetheless maintained that the fibers that emerged from cell bodies possessed processes that ultimately merged with those of other cells.

Golgi's staining techniques were adopted and refined by the Spaniard Santiago Ramón y Cajal (1852–1934). Working in relative isolation from the main schools of scientific research, he was less influenced by prevailing orthodoxies, including the reticularist theory. Whereas Golgi insisted to the end of his life on the continuity of the nervous system, when Cajal looked through the microscope he saw the nerve cells as discrete anatomical entities. Rather than studying specimens derived from adult organisms, Cajal made use of embryonic tissue in which the structure was less complicated and the nerve fibers remained unmyelinated.

Cajal found it difficult to gain acceptance for his claims. A crucial moment came at the 1889 meeting of the German Anatomical Society in Berlin. Cajal was, through direct demonstration of his specimens, able to convince a number of scientists, including the highly influential Albrecht von Kölliker – previously a reticularist – of the truth of his claim that the processes of nerve cells approximated to but never merged with those of others. In 1891 another convert, Wilhelm von Waldeyer (1836–1921), proposed the name 'neuron' for this functional unit. As a highly regarded professor in the University of Berlin, Waldeyer was able to propagate and gain acceptance for this new conception of the microstructure of the nervous system.

Sherrington came to know of Cajal's work and relied upon these morphological assumptions when formulating his theory of the integrative action of the nervous system. The notion of the neuron as anatomically autonomous made it possible to conceptualize the orderly and unidirectional transmission of impulses through the nervous system in a way that was impossible in the reticularist hypothesis. Sherrington in 1897 coined the term 'synapse' (derived from the Greek word 'to clasp') to describe the junction between one nerve cell and another. He was, however, careful to insist that the existence of this junction, as well as the mode of transmission, remained hypothetical.

Electron microscopy in the second half of the twentieth century made it possible to visualize synaptic junctions in far greater detail than was permitted by the light microscope. Chemical theories of transmission had been suggested as early as 1905. An alternative electrical theory was, however, upheld by some, including for a time Sherrington's student, John Carew Eccles (1903–97). The researches of Bernard Katz (1911–2003) in the 1950s, however, demonstrated the existence of a synaptic delay in neuromuscular transmission, which could not be accounted for in the electrical hypothesis but which implied the action of a chemical neurotransmitter. The agent concerned was identified as acetylcholine. Reversing his earlier position, Eccles in 1964 showed the existence of a comparable delay in transmission between afferent and motor neuron, suggesting that chemical transmission at synaptic junctions was a general phenomenon.

Localization of Function

For much of the nineteenth century, physiology remained closely connected to anatomy. This bias was especially marked in Britain and has been explained in terms of a national aversion to conducting experiments on living, conscious beings. A leading exponent of this anatomical approach was the Scottish surgeon Charles Bell (1774–1830), who became involved after 1822 in a controversy with the French physiologist François Magendie (1774–1842) over priority of discovery in the matter of the functions of the spinal nerve roots. Magendie worked in a context where far less stigma was attached to extensive use of animal experimentation. His exhibition of his techniques on a visit to Britain, however, caused outrage.

The convoluted and somewhat sordid details of the controversy between Bell and Magendie need not be detailed here. What should be noted is that the so-called Bell–Magendie law that was ultimately adopted recognized that the posterior and anterior roots of the spinal nerves possessed distinct functions: while the former were sensory the latter were motor in nature. This simple dichotomy was called into question by later research that identified some sensory fibers in the anterior root. From an historical perspective, however, this 'law' was a significant moment in the movement that gathered momentum in the first half of the nineteenth century, to assign particular functions to distinct portions of the nervous system.

Another important contribution to this tendency came from the work of Franz Josef Gall (1758–1828). Gall was a skilled anatomist who pioneered new techniques for the dissection of the brain and spinal cord. He was, however, the author of an ambitious project, which Gall called 'organology,' but which became better known as 'phrenology.' According to this system, the brain was not, as had been generally assumed, a unitary entity, but a congeries of organs, each with its distinct function. Gall further surmised that the size of an organ was directly related to the strength of a particular trait in an individual. These characteristics could be read by means of a

craniological examination, thus providing insights into character and aptitude.

While Gall's system excited considerable interest and support, it never gained general acceptance. His work was, moreover, subjected to what was seen as a devastating critique by the French experimentalist Pierre Jean Marie Flourens (1794–1867), who advocated a more holistic understanding of cerebral function. Flourens maintained that the cerebral hemispheres, cerebellum, and medulla each possessed distinct functions. The hemispheres served as the center for volition and cognition, the cerebellum regulated equilibrium and coordination, and the medulla controlled the vital functions. Flourens, however, insisted that each of these areas of the brain performed its operations as a whole. It was therefore impossible to assign particular offices to circumscribed portions of the cerebrum. Flourens based these conclusions on ablation experiments performed on birds as well as on small mammals. Flourens' views on cerebral function enjoyed the status of orthodoxy until the second half of the nineteenth century.

Gall's suggestion that the cerebral cortex might be composed of functionally distinct areas was not, however, entirely forgotten. When after 1860 new theories of cerebral localization began to circulate, his name was often invoked as an important, if flawed, precursor. As Herbert Spencer (1820–1903) declared in 1870:

> [I]t seems to me that most physiologists have not sufficiently recognized the general truth of which Phrenology is an adumbration. Whoever calmly considers the question, cannot long resist the conviction that different parts of the cerebrum must, *in some way or other*, subserve different kinds of mental action. Localization of function is the law of all organization whatever; and it would be marvellous were there here an exception.

This was just a particular instance of the general law that Spencer called the physiological division of labor.

The 'New Phrenology' that developed after 1860 relied on two principal sources of evidence: clinical observation and animal experiment. The former source derived inferences about normal brain function from the symptoms presented by patients suffering from brain lesions. A particularly important event occurred in 1861 in the course of a debate at the *Société d'Anthropologie* of Paris. The *Société* was known as a center for advanced materialist ideas and included a number of unreconstructed supporters of Gall's views. The members of the *Société* were engaged in a debate on the relationship, if any, between cranial size and intelligence when, in April 1861, a surgeon and anatomist named Pierre Paul Broca (1824–80) made a particularly momentous intervention.

Broca presented the case of a patient named Leborgne who had for several years been unable to utter any words save for the syllable 'tan.' At autopsy Broca had discovered a lesion of the third frontal convolution of Leborgne's brain. On the basis of this and subsequent observations, Broca proposed that the posterior part of this convolution was the seat of articulate language. In 1863 Broca tentatively advanced the view that the speech center was located in the 'dominant' hemisphere of the brain and that there was thus a correlation with handedness.

These claims provoked widespread interest throughout the medical world as practitioners sought out cases of acquired language deficit and attempted to correlate these symptoms with injury to determinate parts of the cortex and its underlying structures. An important contribution to this literature was made in 1874 by the German psychiatrist Carl Wernicke (1848–1905). Wernicke described a form of 'sensory' aphasia in which the ability to comprehend words was impaired, and correlated this condition with injury to the temporal lobe. His observations again supported the hypothesis of hemispherical dominance.

Some scientists, such as Friedrich Goltz (1834–1902), continued to maintain the view that the cerebral hemispheres performed their functions as a unitary organ. There was, moreover, much controversy about the validity of particular localizations. Nonetheless, by the last decades of the nineteenth century the principle of cerebral organization had, as a result of extensive clinical evidence gathered in the aftermath of Broca's initial announcement, received widespread acceptance. If the various aspects of language could be assigned a seat on the cerebral cortex, then there seemed every reason to suppose that the other higher functions of the brain might also be localized.

These clinical findings were supplemented by a series of experimental results that emerged in the later nineteenth century. In 1870, the psychiatrist Eduard Hitzig (1838–1908) and the anatomist Gustav Fritsch (1838–1927) published the results of experiments in which they had applied an electric current to various parts of the cerebral cortex of a dog. These investigations appeared to show that stimulation of specific sites upon the cortex elicited movements of particular parts of the body; there was thus a 'motor cortex.' This result overthrew Flourens' doctrine that the cerebral hemispheres were not motor in function, as well as supplying further confirmation of the reality of cerebral localization.

Fritsch and Hitzig's results were corroborated by the British physician David Ferrier (1843–1928), who, working on the brain of monkeys, produced more detailed maps of the motor cortex. Ferrier also proposed cortical sites for some of the special senses.

The doctrine of cerebral localization received further support from the aforementioned developments in the histology of the nervous system. If each neuron was a distinct unit – an "absolutely autonomous canton" (Cajal) – then it became much easier to conceive the structural basis of functional specialization. Developments in cytoarchitectonics, which appeared to show the existence of histologically distinct areas in the cerebral cortex, were also seen to provide further support for localization.

By the end of the nineteenth century a 'mosaicist' view of cerebral function was thus in the ascendant. Around 1900, however, the contention that particular functions could be assigned to ever more specific zones within the brain was challenged. A number of neurologists, such as Henry Head (1861–1940) and Kurt Goldstein (1878–1965), argued that a proper study of clinical cases revealed evidence that was incompatible with any strict notion of cerebral localization. They cited the fact that cortical injury led to global behavioral disorders and pointed to the apparent ability of the brain to recover functional integrity despite the destruction of portions of the cortex. Shepherd Ivory Franz (1874–1933) came to similar conclusions on the basis of both clinical and experimental work. The American psychologist Karl Lashley (1890–1958) derived the doctrine of cortical 'equipotentiality' from his own experiments. According to this theory, any part of the cortex could substitute for the functions of the other in case of local injury. Lashley was, however, willing to accept a limited form of localization for sensory and motor functions.

The middle decades of the twentieth century saw, however, a reassertion of the reality of cortical localization based upon the use of increasingly sophisticated experimental techniques. Working at Johns Hopkins University, Clinton Nathan Woolsey (1904–93) employed electrophysiological methods to show that specific areas of the cortex responded to the stimulation of small cutaneous surfaces. The British physiologist Edgar Douglas Adrian (1889–1977) also developed techniques that made it possible to record the discharge of single nerve fibers. This methodology was used, in conjunction with refined ablation techniques, to map the functional divisions of the mammalian cortex with far greater precision than had previously been possible. More recently, brain-imaging techniques, such as positron emission tomography, have apparently provided striking visual evidence that specific areas of the brain are active during certain cognitive states. Thanks to the wide dissemination of such images through the media, the doctrine of cerebral localization has now entered popular culture.

See also: Cajal's Place in the History of Neuroscience; Cognitive Neuroscience: An Overview; History of Neuroscience: Early Neuroscience; Neuron Doctrine: Historical Background; Neuropsychology: Theoretical Basis.

Further Reading

Brazier MAB (1984) *A History of Neurophysiology in the 17th and 18th Centuries*. New York: Raven Press.

Brazier MAB (1988) *A History of Neurophysiology in the 19th Century*. New York: Raven Press.

Clarke E and Jacyna LS (1987) *Nineteenth-Century Origins of Neuroscientific Concepts*. Berkeley: University of California Press.

Cranefield PF (1974) *The Way in and the Way Out, François Magendie, Charles Bell and the Roots of the Spinal Nerves*. Mount Kisco, NY: Futura Publishing.

Fearing F (1930) *Reflex Action: A Study in the History of Physiological Psychology*. Baltimore: Williams & Wilkins.

Finger S (1994) *Origins of Neuroscience: A History of Explorations into Brain Function*. New York: Oxford University Press.

Fulton JF (1933) Science in the clinic as exemplified by the life and work of Joseph Babinski. *Journal of Nervous and Mental Disease* 77: 121–133.

Hagner M (1997) *Homo Cerebralis: Der Wandel vom Seelenorgan zum Gehirn*. Berlin: Berlin Verlag.

Harrington A (1987) *Medicine, Mind, and the Double Brain*. Princeton: Princeton University Press.

Harrington A (1990) *A Feeling for the 'Whole': the Holistic Reaction in Neurology from the Fin de Siècle to the Interwar Years*. Princeton: Princeton University Press.

Head H (1926) *Aphasia and Kindred Disorders of Speech*. Cambridge: Cambridge University Press.

Jacyna LS (2000) *Lost Words: Narratives of Language and the Brain, 1825–1926*. Princeton: Princeton University Press.

Liddell EGT (1960) *The Discovery of Reflexes*. Oxford: Oxford University Press.

Marshall LH and Magoun HW (1998) *Discoveries in the Human Brain: Neuroscience Prehistory, Brain Structure, and Function*. Totowa, NJ: Humana Press.

Martensen RL (2004) *The Brain Takes Shape: An Early History*. Oxford: Oxford University Press.

Meyer A and Hierons R (1965) On Thomas Willis's concepts of neurophysiology. *Medical History* 9: 1–15.

Ochs S (2004) *A History of Nerve Functions: From Animal Spirits to Molecular Mechanisms*. Cambridge: Cambridge University Press.

Shepherd GM (1991) *Foundations of the Neuron Doctrine*. New York: Oxford University Press.

Smith R (1992) *Inhibition: History and Meaning in the Sciences of Mind and Brain*. London: Free Association Books.

Young RM (1990) *Mind, Brain, and Adaptation in the Nineteenth Century: Cerebral Localization and Its Biological Context from Gall to Ferrier*. New York: Oxford University Press.

Neuroplasticity: Computational Approaches

H Z Shouval, University of Texas Medical School at Houston, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Experience-dependent changes in the efficacy of synaptic connections between different neurons is believed to be a basis for learning memory and various forms of development. In 1949 Donald Hebb postulated that when one neuron persistently drives the activity of another neuron, the synaptic connection between these neurons is potentiated. This activity-dependent synaptic potentiation could account for various forms of learning. Theoretical analysis indicates that not only Hebbian-like synaptic potentiation is necessary but also depression between two neurons that are not sufficiently coactive. If depression did not exist, all synapses would tend to saturate at their maximal allowed values, neurons would not be selective, and they would respond to different stimuli in a similar manner. The experimental correlates of these theoretically proposed forms of synaptic plasticity are called long-term potentiation (LTP) and long-term depression (LTD). Significant experimental evidence has accumulated to support the hypothesis that synaptic plasticity is indeed a basis for learning memory and various forms of development.

The cellular and molecular machinery responsible for synaptic plasticity is complex. Hundreds of different molecules are involved in the induction and maintenance of synaptic plasticity, and this complexity has raised doubts about the ability to truly understand these mechanisms. The role of computational modeling is to start from a set of assumptions about underlying mechanisms, in this case about the underlying physiology and molecular biology, formalize them as a set of equations, and use simulations to obtain the unambiguous consequences of this set of assumptions. Therefore, for complex systems, computational modeling is an indispensable tool for gaining a true understanding of how they operate.

Two broad classes of synaptic plasticity models can be described: (1) Phenomenological models are very simple models that are typically based on an input–output relationship between neuronal activity and synaptic plasticity. Phenomenological models are typically used in simulations to account for higher-level phenomena such as the formation of memory or the development of neuronal selectivity. (2) Biophysical models are more-detailed models that incorporate more of the cellular and synaptic biophysics of neurons and are typically used to account for controlled synaptic plasticity experiments. This article focuses on biophysical models.

Synaptic plasticity has an early induction phase in which external signals rapidly cause a biochemical cascade that results in a change in synaptic efficacy. This change in efficacy is expressed by changing either pre or postsynaptic variables that control synaptic transmission, such as the number of postsynaptic receptors or the probability of presynaptic release. There is also a late phase of synaptic plasticity in which these efficacy changes are maintained over long periods of time. These two phases can be partially separated physiologically and pharmacologically. The section titled 'Biophysical models for the induction of synaptic plasticity' reviews models describing the induction of synaptic plasticity, and the section titled 'Models for maintaining synaptic plasticity' reviews models accounting for the maintenance of synaptic plasticity.

Different Induction Protocols for Bidirectional Synaptic Plasticity

LTP was first described in 1973 by Bliss and Lømo, and since that time many different induction paradigms have been used to induce LTP and LTD. The most common method for inducing LTP has been to use extracellular stimulation together with a recording of extracellular field excitatory postsynaptic potentials (fEPSPs). The fEPSP, although recorded extracellularly, is believed to represent an average of the EPSPs of many synapses in the vicinity of the recording electrode. To induce LTP, first a baseline is established (Figure 1(a)). This means that a very low-frequency stimulus is applied (typically once every 30–60 s), the fEPSP is recorded, and over the baseline period, the magnitude or slope of the fEPSP is required to remain constant. After the baseline period, several bursts of a high-frequency stimulus are applied (theta burst stimulation (TBS), typically two to five bursts of four stimuli at 100–200 Hz, with interburst period of 200 ms). After the TBS, the magnitude and slope of the baseline are usually increased (Figures 1(a)–1(c)). Using this method of extracellular stimulation, LTD can also be induced using low-frequency stimulation (typically 1–5 Hz) with a large number of stimuli (≈1000).

Pairing is the name given to another commonly used induction protocol. In a pairing protocol, the postsynaptic cell is recorded and stimulated by means of a patch clamp electrode. After baseline, LTP is induced by pairing a low-frequency extracellular stimulus

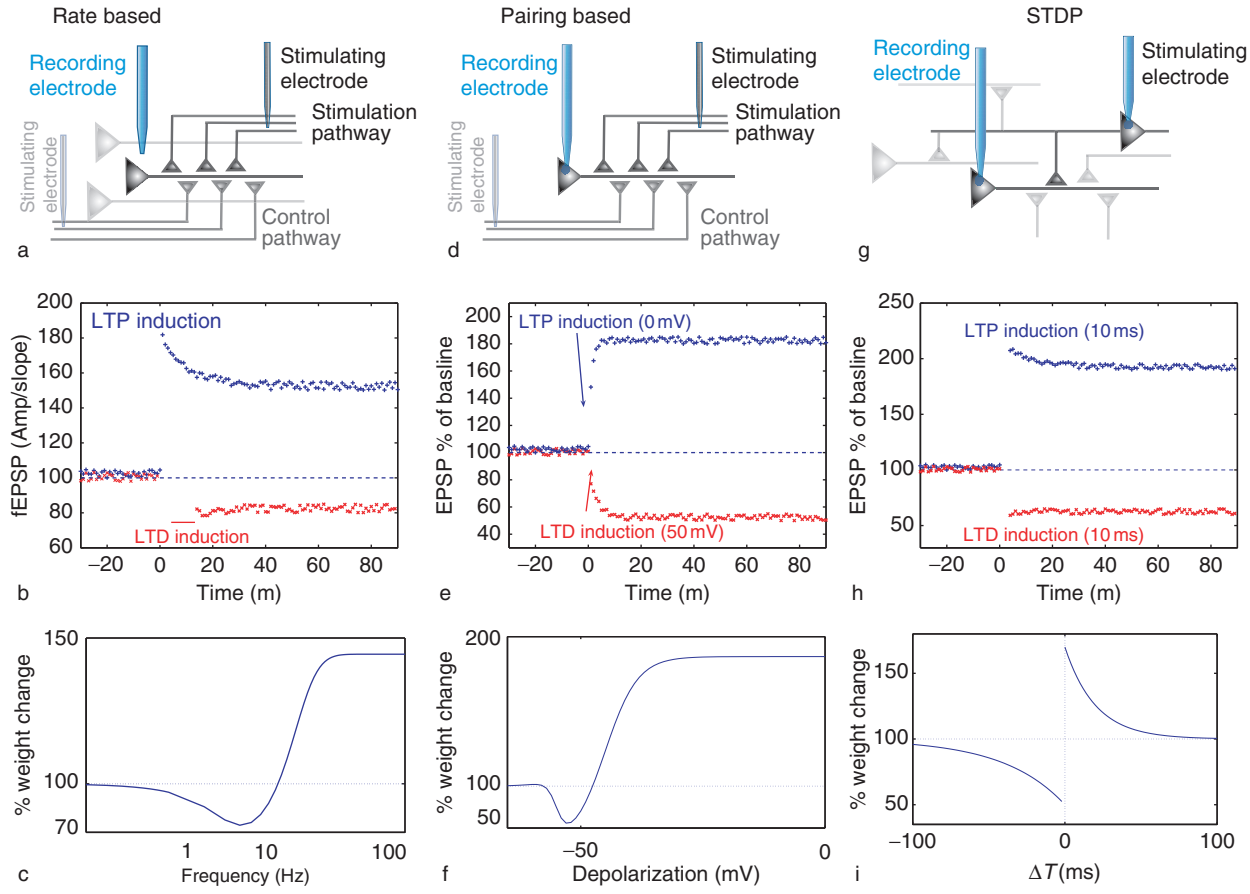

Figure 1 Different protocols for inducing synaptic plasticity in slices. (a) In rate-based induction, extracellular stimulating and recording electrodes are used. (b) First a baseline field excitatory postsynaptic potential (fEPSP) is established using a very low-frequency stimulus. Next a low-frequency stimulation (hundreds of pulses at 1–10 Hz) is used to elicit long-term depression (LTD) or a high-frequency stimulus (30–200 Hz) is used to elicit long-term potentiation (LTP). (c) A schematic diagram of plasticity as a function of frequency. (d) In pairing-based plasticity, an extracellular stimulation is paired with an intracellular voltage clamp-controlled depolarization. (e) If the cell is slightly depolarized to −50 mV, LTD is induced, whereas strong depolarization to 0 mV produces LTP. (f) Plasticity curve as a function of postsynaptic depolarization. (g) In spike timing-dependent plasticity (STDP) induction protocols, one presynaptic and one postsynaptic cell are patch clamped. (h) If the presynaptic cell is induced to repetitively fire 10 ms before the postsynaptic cell, LTP is induced. If the firing order is reversed, LTD is induced. (i) A schematic plasticity curve as a function of Δt. (All diagrams are schematic illustrations.)

with a strong depolarization of the postsynaptic cell (Figures 1(d)–1(f)). A moderate depolarization of the postsynaptic cell, paired with a low-frequency stimulus, results in LTD (Figures 1(d)–1(f)).

A more recent induction protocol is spike timing-dependent plasticity (STDP). This form of plasticity is induced by a repeated delivery of pre- and postsynaptic spikes with a given time interval Δt between the pre- and postsynaptic spikes. In an STDP protocol, the precise value of Δt determines the size and magnitude of the resulting synaptic plasticity. Typically, if the presynaptic spike precedes the postsynaptic spike ($\Delta t > 0$), LTP is obtained (Figures 1(g)–1(i)), and if the postsynaptic spike comes before the presynaptic spike ($\Delta t < 0$), LTD is obtained. However, other parameters, such as frequency of stimulation, magnitude of presynaptic EPSP, and the location of the stimulated synapse along the dendritic tree also affect the induction of STDP.

Biophysical Models for the Induction of Synaptic Plasticity

A biophysical model is defined by a set of mathematical equations that represent the biochemical and physiological processes that lead to the induction and expression of synaptic plasticity. However, because it is not possible to implement precisely every portion of the physiological and biochemical networks leading to synaptic plasticity, these

biophysical models must rely on many simplifications and abstractions.

Calcium-Dependent Models of Bidirectional Synaptic Plasticity

Calcium transients in the spine determine the sign and magnitude of many forms of synaptic plasticity. The source of the major portion of postsynaptic calcium is influx through the *N*-methyl-D-aspartate (NMDA) receptors, and therefore blocking them pharmacologically can eliminate bidirectional synaptic plasticity. A partial block of NMDA receptors can convert LTP to LTD. Moreover, strong and brief postsynaptic calcium transient, in the absence of a presynaptic stimulus, can produce LTP, whereas prolonged calcium transients with moderate magnitudes result in LTD.

A key component of most biophysical models of synaptic plasticity is the NMDA receptor. Since the NMDA receptor depends both on binding of neurotransmitter and on postsynaptic voltage, it is a coincidence detector, thus providing Hebbian properties to NMDA receptor-dependent synaptic plasticity. Early models of LTP simulated calcium influx through NMDA receptors during LTP induction protocols. In these models, a large calcium transient is taken to be the correlate of LTP. These models simulated stimulation protocols that activate multiple synapses on the same dendritic branch at various frequencies and examined resulting calcium transients in a single spine. A common conclusion of these different studies was that to generate large calcium transients during LTP, the calcium buffers, which absorb free calcium, must be saturated, thus amplifying the small signal observed in the calcium currents. However, these early studies did not take into account back-propagating action potentials (BPAP) or active dendritic conductances.

The calcium control hypothesis, an influential concept in the field, postulates that a large calcium transient produces LTP whereas a moderate increase in calcium results in LTD. This hypothesis has subsequently received significant experimental support. Several different models for the induction of synaptic plasticity are based, either explicitly or implicitly, on this hypothesis.

To simulate calcium-dependent synaptic plasticity, it is necessary to define rules that translate calcium dynamics in postsynaptic spines to changes in synaptic strength. A simple choice is to assume that the peak of the calcium transients determines the sign and magnitude of the synaptic weight change. This simplified assumption might have some unrealistic consequences, such as that the width of calcium transients has no effect on the magnitude of plasticity. A simple dynamical system that implements the calcium control hypothesis has the following form:

$$\mathrm{d}W_i/\mathrm{d}t = \eta(\mathrm{Ca})[\Omega(\mathrm{Ca}) - \lambda W_i] \qquad [1]$$

Here W_i is the synaptic efficacy of synapse i, Ω (Figure 2(a)) determines the sign magnitude of synaptic plasticity as a function of Ca levels, η (Figure 2(b)) is a calcium-dependent learning rate, which is typically a monotonically increasing function of calcium, and λ is decay constant. If $\lambda = 1$, for sustained calcium elevation, the synaptic efficacy converges to the value of Ω, and the rate of convergence depends on η.

It is very easy to simulate pairing-induced plasticity (Figure 1(d)–(f)) on the basis of the calcium control hypothesis and mathematical models of NMDA receptors. Because of the voltage dependence of NMDA receptors, pairing a presynaptic stimulus with a moderate postsynaptic depolarization (Figure 3(a)) produces a significantly smaller calcium transient than if the presynaptic stimulus is paired with a strong depolarization. Consequently (Figure 3(b)), the change in the simulated synaptic weights, after 100 presynaptic stimuli, is a function of the level of the postsynaptic voltage during induction, in qualitative agreement with experimental results. In the simplest models, all calcium influx is assumed to originate from NMDA receptors. Models that take into account calcium influx through voltage-gated calcium channels (VGCC) as well produce results that are qualitatively similar.

Rate-dependent induction protocols have also been simulated using similar assumptions. Modeling such

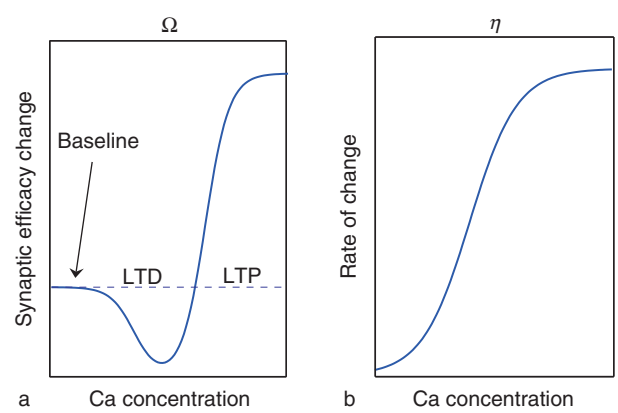

Figure 2 Calcium-dependent functions that control synaptic plasticity. (a) The Ω function controls the sign and magnitude of synaptic plasticity. At low calcium concentrations corresponding to baseline activation, no change in synaptic weight occurs. At intermediate calcium concentration, long-term depression (LTD) is induced, and at high calcium, long-term potentiation (LTP) is induced. (b) The η function, which controls the rate of convergence to the steady state, is a monotonically increasing function of calcium.

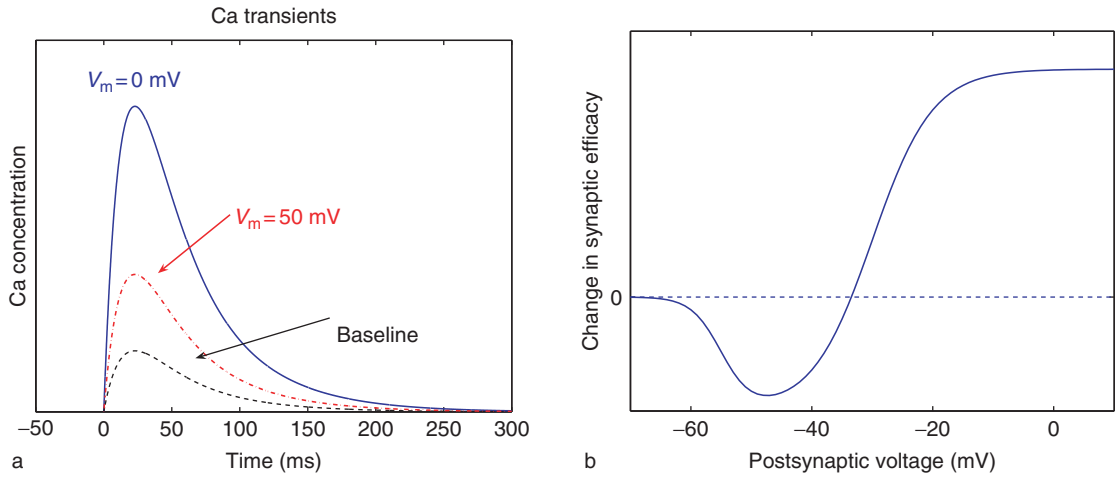

Figure 3 Pairing-induced plasticity (schematic diagrams). (a) Calcium transients induced by a single presynaptic stimulation at resting membrane potential (black dashes), at slightly depolarized potential (−50 mV, red dashes), and for highly depolarized potential (0 mV, blue). (b) The plasticity curve shows the magnitude of plasticity as a function of postsynaptic voltage; V_m, membrane potential.

rate-dependent protocols, in which there is no direct experimental control over the postsynaptic neuron, requires a model of a spiking postsynaptic neuron.

In a calcium-based model of synaptic plasticity, the presynaptic spike naturally effects calcium influx because glutamate released due to the presynaptic spike binds postsyanaptic NMDA receptors and causes the receptors to open. The postsynaptic spike can alter calcium influx through NMDA receptors because the BPAP, which depolarizes the postsynaptic cell, causes an increased conductance through the NMDA receptors. Given the assumptions of the calcium control hypothesis, what is required for induction of STDP is that calcium transients during post–pre ($\Delta t < 0$) conditions be larger than transients due to unpaired presynaptic or postsynaptic stimulation but smaller than transients due to pre–post stimuli ($\Delta t > 0$). Typically, neuronal spikes in cortex are assumed to be of very short duration, ≈1 ms. Figure 4(d) shows calcium transients for different conditions: pre alone (blue), $\Delta t = -10$ ms (red), and $\Delta t = 10$ ms (black), while assuming a BPAP that decays within 2 ms. There is very little difference between pre alone and post–pre, and in the pre–post condition, calcium transients are only slightly larger. There is no difference between post–pre and pre only conditions because by the time of the presynaptic spike, the postsynaptic voltage generated by the BPAP entirely disappears, and the extra calcium influx in the pre–post condition is small because of the short duration of the BPAP. These two problems can be overcome if the BPAP, as commonly observed in dendrites, has a slow repolarization tail potential (Figure 4, bottom). This tail potential is still present when the presynaptic spike occurs during a post–pre protocol (Figures 4(f) and 4(h)), producing a difference between post–pre and baseline. The long tail of the BPAP also increases the total calcium influx during pre–post (Figures 4(g) and 4(h)) protocols. A consequence of this set of assumptions is STDP curves as presented in Figure 5(b). For $\Delta t < 0$ they produce LTD, as observed experimentally; for small values of $\Delta t > 0$, they result in LTP; but for larger values of Δt, they produce LTD again.

The pre–post form of LTD predicted by such models is different from the cartoon image of STDP presented in Figure 5(a). Does this pre–post form of LTD exist experimentally? In hippocampal slices, this form of LTD exists, is robust, and has a similar magnitude to the post–pre form of LTD. In contrast, in experiments performed in cortical slices or hippocampal cultures, this form of LTD is either absent or at least significantly smaller than the post–pre form of LTD. This raises the possibility that different mechanisms are at work in these different cortical regions or types of experimental preparation.

Another explanation for the absence of the second LTD window in some preparations is that calcium transients show a significant variability due to stochastic synaptic transmission. Much of this stochastic variability might arise from the stochastic opening and closing of the small number of postsynaptic NMDA receptors. If this variability is large enough, it might result in a very significant reduction in the magnitude of the second LTD window.

Two-Coincidence Models and Other Pathways for Induction

Most models of synaptic plasticity based on the calcium control hypothesis predict an STDP curve in

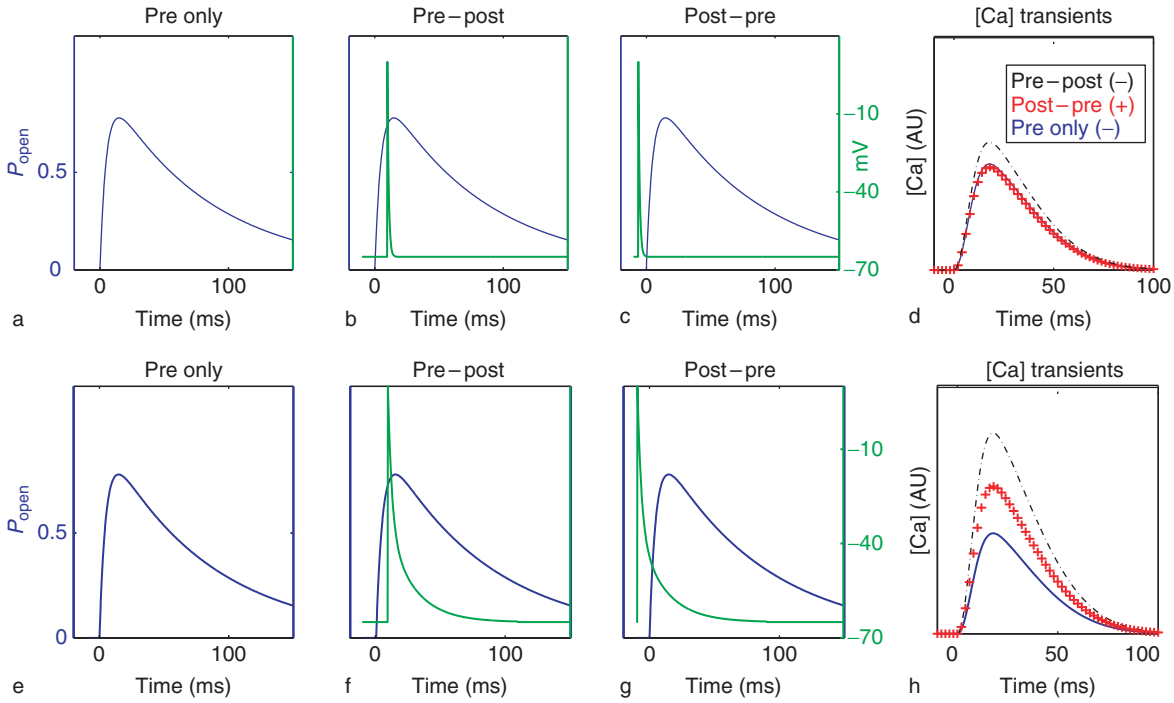

Figure 4 Calcium transients during induction of spike timing-dependent plasticity, with a narrow back-propagating action potential (BPAP; (a)–(d)) and a BPAP with a wide tail ((e)–(h)). Here all currents are assumed to arise from *N*-methyl-D-aspartate (NMDA) receptors, and these diagrams are schematic. The probability of NMDA receptors in the open state (P_{open}) due to glutamate binding (blue) and the postsynaptic voltage due to BPAP (green) (a) pre only; (b) pre before post; (c) post before pre. (d) Shows calcium transients, in arbitrary units (AU), for the three cases: pre only (blue), post–pre (red), pre–post (black). Post–pre does not differ from baseline, and pre–post increased only moderately. Calcium transients are now different for the three different cases (h).

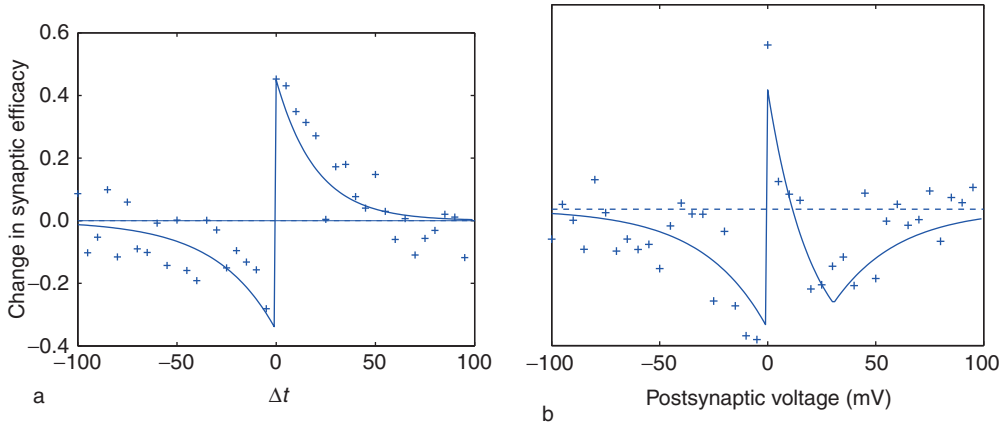

Figure 5 Spike timing-dependent plasticity (STDP) curves (schematic diagrams). (a) Some experimental results indicate STDP curves with a single long-term depression (LTD) region at $\Delta t < 0$ and a single long-term potentiation (LTP) region at $\Delta t > 0$. (b) Other experiments find an additional LTD region at larger values of $\Delta t > 0$. Note that in all cases, results of single experiments (+) diverge significantly from the simple fit curves.

which there is a pre–post form of LTD in addition to the standard post–pre form of LTD. An alternative to the single-coincidence models described above are two-coincidence models, which postulate that LTP is triggered by calcium influx through NMDA receptors, whereas LTD is triggered by a second coincidence detector that is sensitive to post-before-pre activation. One option is that the second coincidence detector is implemented by calcium influx through VGCCs concurrent with the activity of metabotropic glutamate receptors. Experimental work has indeed shown that postsynaptic NMDA receptors might not be necessary for spike timing-dependent LTD, and that metabotropic glutamate receptors and VGCCs

are necessary for spike timing-dependent LTD. Some experiments have also found that the activation of cannabinoid receptors is necessary for this LTD. Although elements of the two-coincidence model have received support, the biophysical details of this second coincidence detector have yet to be elucidated.

Alternatives to the idea that the second coincidence detector leads to LTD have been proposed. One idea is that calcium influx through VGCC suppresses LTD or that specific calcium temporal profiles invoke a veto mechanism to suppress LTD, eliminating the pre–post form of LTD.

Modeling the Biochemical Pathways Leading from Calcium to Change in Synaptic Efficacy

The question of how the same molecule, calcium, can trigger both LTP and LTD was first addressed in a model proposed by John Lisman in 1989. Lisman's model postulates that moderate calcium preferentially activates phosphatases, which dephosphorylate calcium/calmodulin-dependent protein kinase II (CaMKII), whereas high calcium levels cause a net phosphorylation of CaMKII. The correlate of synaptic efficacy in this model is the phosphorylation level of CaMKII, and this mechanism could therefore account for bidirectional synaptic plasticity.

Recent experimental evidence has shown that a correlate of synaptic plasticity is the phosphorylation state of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors: LTP is correlated with phosphorylation at s831, a CaMKII site, and LTD is correlated with dephosphorylation at s845, a protein kinase A (PKA) site. Models by Castellani and colleagues simulated some of the kinases and phosphatases involved in the signal transduction pathways leading from calcium transients to synaptic plasticity. This approach was that synaptic plasticity is first expressed in posttranslational changes in AMPA receptors. Therefore, they constructed a minimal biochemical induction pathway that links these specific changes in AMPA receptors with the original signal, the change in postsynaptic calcium levels.

Figure 6 shows the elements of this induction model, which is an extension of Lisman's work. Plasticity protocols cause calcium to enter the spine, which then binds to calmodulin (CaM) to form Ca-CaM. Ca-CaM can bind and activate, either directly or indirectly, different kinases and phosphatases. At low levels, Ca-CaM is assumed to primarily bind protein phosphatase 2b, which indirectly activates protein phosphatase 1 (PP1) (Figure 6(a)). PP1 then dephosphorylates AMPA receptors at both s831 and s845 sites (Figure 6(b)). A higher postsynaptic calcium concentration causes direct activation of CaMKII and an indirect activation of PKA through the cyclic adenosine monophosphate pathway. CaMKII phosphorylates AMPA receptors at s831 and PKA at s845. PKA also indirectly inactivates PP1, thus decreasing phosphorylation at both sites. The phosphorylation of both these sites increases the conductance of the AMPA receptors and controls receptor trafficking.

To construct this model, various forward and backward kinetic coefficients are needed. However, only a small portion of them are known. It is possible to choose constants, in a way that is consistent with experimental data, such that in this model, at moderate calcium, LTD is induced and at higher calcium, LTP is induced. However, different sets of kinetic coefficients can be extracted from the data. Although the different sets of coefficients produce results that are qualitatively

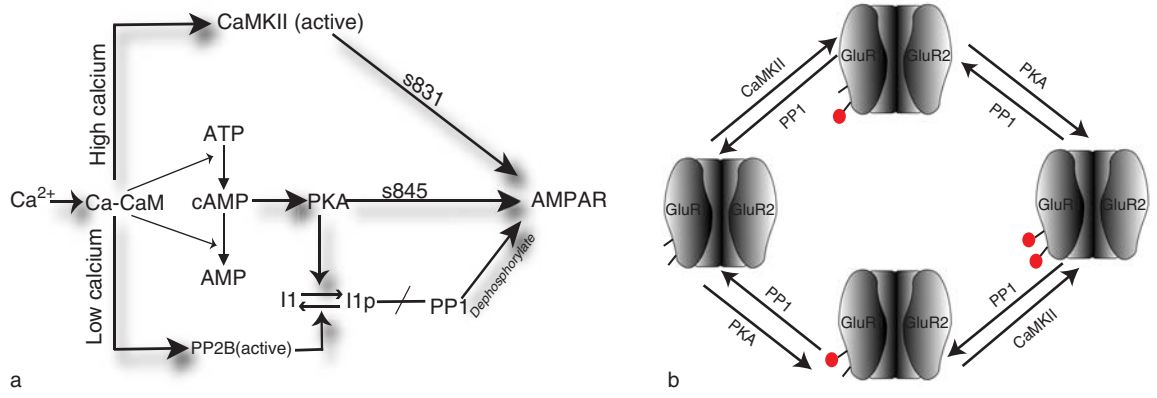

Figure 6 Schematic diagrams of biochemical pathways involved in the induction of bidirectional synaptic plasticity. (a) Kinases and phosphatases that are activated by calcium and that modify the conformational state of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors (AMPARs). (b) Biochemical pathways for the phosphorylation and dephosphorylation of AMPA at two phosphorylation sites. I1 is the regulatory subunit of PP1 that binds to PP1, and inhibits its activity when in the phosphorylated state (I1p). AMP, adenosine monophosphate; ATP, adenosine triphosphate; CaM, calmodulin; CaMKII, calcium/calmodulin-dependent protein kinase II; cAMP, cyclic adenosine monophosphate; GluR1 and GluR2, metabotropic glutamate receptors; PKA, protein kinase A; PP, protein phosphatase; s831, a CaMKII site; s845, a PKA site.

similar to each other and to experiments, these different parameters produce results that are quantitatively significantly different. Under certain conditions, such enzymatic models can be approximated by the simpler calcium-dependent models described by eqn [1].

One problem with these enzymatic models is that they require many unmeasured kinetic coefficients. However, a far more difficult problem is that these models are based on a mass action approach, an approach that is valid for large, well-mixed systems. However, synapses are neither large nor well mixed, and therefore these mass action simulations might produce results with significant errors. An alternative approach is to simulate biochemical reactions using detailed stochastic simulations that are not based on the mass action approach. Such simulations must take into account both stochastic fluctuations due to the limited synaptic volume and the three-dimensional structure and organization in the spine. Consequently, such simulations are very difficult and computationally intensive; however, they might be necessary in modeling biochemical processes that occur in spines. It is also important to know which of the biochemical processes in the spine require this detailed form of modeling and which ones are suitable for the simpler mass action approach.

Models for Maintaining Synaptic Plasticity

Synaptic plasticity is expressed by changes to the synaptic transmission machinery. Such changes include receptor phosphorylation, change in receptor number, and presynaptic changes in probability of release. Changes in conformational states of receptors through phosphorylation are counteracted by dephosphorylation by phosphatases and by protein turnover and trafficking and therefore are likely to have a limited lifetime, much shorter than the lifetime of memory. Similarly, changes in number of receptors are likely to be transient due to trafficking and turnover of receptor proteins. Similarly, presynaptic changes, although their mechanism is less explored, are likely expressed in changes in proteins, changes that are also likely to be transient. However, although changes in protein function or number are transient, the memories they encode can last a lifetime.

The inherent transience of synaptic efficacies was noted in 1984 by Francis Crick, who suggested that cooperative interactions among proteins can overcome this problem. Crick suggested a specific but quite abstract model of such cooperativity. Some form of cooperativity underlies most models that address this problem.

In the absence of a correcting mechanism, protein turnover sets an upper limit to the stability of synaptic plasticity since it will erase both conformational changes to proteins and change in protein number. Similarly, protein trafficking also limits the lifetime of memories because proteins that are altered in one synapse move away, removing the synapse-specific trace laid down by plasticity. However, there is no agreed on estimate about the rates of protein turnover and trafficking. While recent estimates of protein turnover in slices are on the order of 2–20 h, previous studies produced much longer estimates. Similarly, different methods produce significantly differing estimates regarding the timescale of receptor trafficking. However, even longer estimates of receptor dwell time are much shorter than the lifetime of some memories.

Induction protocols of synaptic plasticity (see the section titled 'Different induction protocols for bidirectional synaptic plasticity' above) last between seconds and minutes. However, different types of induction protocols can produce either a transient change, one that returns to baseline within 2–4 h, or a long-lasting change. For example, LTP induced by a single theta burst produces a transient LTP (E-LTP), but four theta bursts produce a long-lasting LTP (L-LTP). Similarly, different induction protocols can produce short LTD (E-LTD) or long-lasting LTD (L-LTD). It is interesting that the induction of L-LTP and L-LTD depends on the synthesis of new proteins during the induction phase. The timescales of E-LTP and E-LTD are roughly consistent with temporal estimates for protein turnover and trafficking. The much longer lifetimes of L-LTP and L-LTD and their protein synthesis dependence can lead to the hypothesis that protein synthesis can somehow counteract protein turnover and trafficking.

A Molecular Switch

One possible mechanism that can stabilize synaptic efficacies is a bistable molecular switch. Such a switch can be generated by a kinase that is active when phosphorylated and that can phosphorylate itself. This type of a positive feedback loop often has two stable steady states: one is unphosphorylated, and another one is nearly totally phosphorylated. If the system is in the phosphorylated stable state, when a phosphorylated molecule is degraded and replaced by an unphosphorylated molecule, it will quickly be phosphorylated by its active neighbors. This autophosphorylating kinase can act directly on receptors to determine synaptic efficacy.

The most widely investigated mechanism proposed for such a bistable switch is the autophosphorylation of CaMKII molecules. CaMKII molecules are embedded as subunits in a holoenzyme composed of approximately ten other CaMKII molecules. CaMKII molecules are activated by a Ca-CaM signal and can move to a relatively stable active phosphorylated state

by interacting with other CaMKII subunits. This is a more complex model than the simple phosphorylation loop described above. It is apparent that a model of CaMKII requires some level of a Ca-CaM signal so that the phosphorylated state is stable in face of protein turnover. Different models of CaMKII autophosphorylation have been proposed; some exhibit bi-stability, while other models do not. However, similar to biophysical models of induction, many of the kinetic parameters of the system are unknown, and the exact details of the appropriate model are not yet fully agreed on. Therefore, it remains to be determined whether a realistic representation of CaMKII autophosphorylation allows bi-stability. Moreover, experimental studies have not found direct support for bi-stability of the autophosphorylating CaMKII system.

Although the importance of CaMKII for the induction of LTP is not disputed, there is less unequivocal evidence regarding its importance for the maintenance of L-LTP or long-term memory. Recent results have shown prolonged phosphorylation of CaMKII during the maintenance of L-LTP; however, the fraction of phosphorylated kinase after LTP in such experiments seems far from saturation, unlike the predictions of the computational models. It is interesting to note that the putative targets of CaMKII, such the AMPA receptors, seem to be only transiently phosphorylated after protocols that induce long-term memory. Blocking of CaMKII activity during the maintenance phase of L-LTP in most experiments, but not all, did not block the maintenance phase of LTP.

The Cluster Model

The cluster theory of synaptic stability offers an alternative to molecular switch approaches to synaptic stability. In this theory it is assumed that receptors in the postsynaptic density are typically clustered (Figure 7(a)) and that receptors within the cluster interact so that they influence the trafficking of receptors in and out of the cluster in such a way that the lifetime of the cluster exceeds by order of magnitude the dwell time of each receptor within the cluster. This model is based on several assumptions: (1) Synaptic efficacy is proportional to the number of postsynaptic receptors, such as AMPA receptors; (2) receptors in the postsynaptic density are clustered; (3) the insertion rate of a receptor in the vicinity of other receptors in the cluster is much higher than for an isolated receptor; and (4) the rate of receptor removal from the cluster is independent of interactions with other receptors in the cluster. Assumptions (1) through (3) are essential assumptions of this model, while assumption (4) could be altered while preserving the main features of the model.

Simulations of such networks (Figures 7(b) and 7(c)) show that although individual receptors within a cluster are rapidly trafficked in and out of the cluster, the cluster itself remains stable for long periods. Although the exact instantaneous number of receptors fluctuates, it fluctuates around a mean that remains stable for periods that are much longer than the dwell time of a single receptor. These receptor clusters are not stable fixed points; instead they are in a metastable state and will eventually decay.

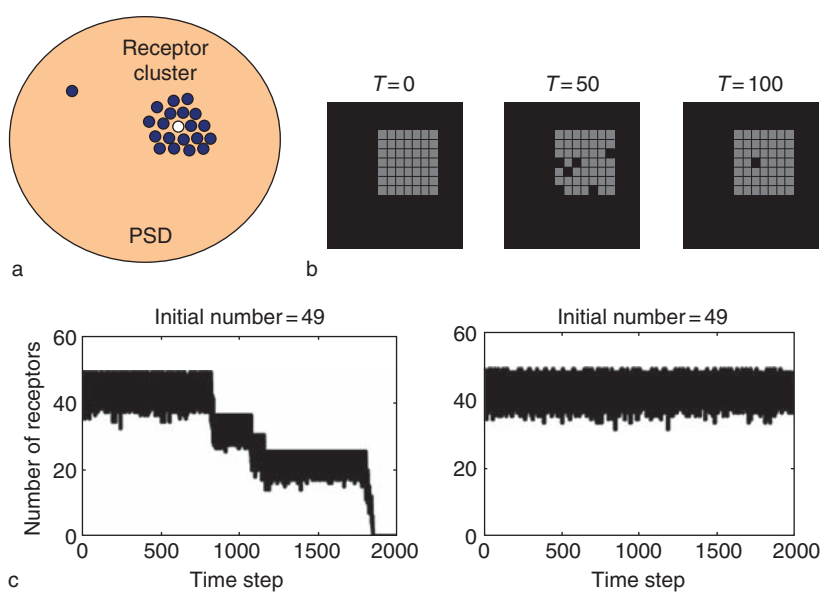

Figure 7 The cluster theory of synaptic stability. (a) Receptors in the postsynaptic density (PSD) tend to cluster. When a receptor is removed because of turnover or trafficking out of the receptor, it is rapidly replaced by another receptor. (b) A simulation of a 7×7 cluster of receptors shown at time $T = 0$, 50, and 100, where 1 corresponds to the average dwell time of a receptor. Although receptors are rapidly removed and replaced, the cluster remains stable. (c) The number of receptors as a function of time in two independent simulations.

The lifetime of clusters depends linearly on the dwell time of receptors and increases steeply with the initial size of the cluster. For example, if we assume a receptor dwell time of 20 min, then the median lifetime for a 9×9 cluster of receptors can be more than 1 year.

Additional Mechanisms

Other possibilities have been raised to account for the stability of learning and memory in the face of molecular instability. One such idea is that synaptic efficacies themselves are not stable and that in the absence of neuronal activity, they decay. However, network activity might be able to stabilize efficacies. According to this idea, the interaction between synaptic plasticity and network activity acts to stabilize synaptic weights. If a pattern is embedded in a network by changing appropriate synaptic weights, such patterns are more likely to be spontaneously activated, and when they are activated, they are further potentiated, counteracting the synaptic decay. Such spontaneous activation could be manifest during periods of dreams or rest. However, a study by Wittenberg and colleagues has shown that although one embedded activity pattern can be stabilized with this mechanism, multiple patterns cannot since one of the embedded patterns takes over the network and eliminates the other patterns. It remains to be seen whether this is a general property of this proposed mechanism or whether it can be corrected with an appropriate choice of plasticity mechanism.

Another idea, motivated by experimental findings regarding long-term facilitation in *Aplysia*, is that memory can be preserved by a self-sustaining prionlike state. This theory assumes that a specific translation factor (cytoplasmic polyadenylation element binding protein (CPEB)), which is associated with the persistence of long-term facilitation in *Aplysia*, has prionlike properties. The prionlike behavior implies that this protein can be folded in different ways which have different functional properties and that this prionlike state is self-sustaining in the sense that new proteins that arrive to the vicinity of prions take on the prion form. The prionlike isoform of CPEB in *Aplysia* neurons is more active than the nonprion form. The concept of memories being encoded by a prion state therefore assumes that synaptic facilitation causes transformation of a key synaptic protein to its prionlike state, which is self-sustaining. It is not clear how such memory can be reversed or whether the prionlike state will not spread to nearby synapses. Although CPEB is also important for L-LTP in mammals, its structure in mammals is significantly different, it has phosphorylation sites that are absent in *Aplysia*, and prionlike properties have not been demonstrated in mammalian CPEB.

A detailed model for the protein synthesis-dependent induction of L-LTP was proposed by Smolen and colleagues. However, this model does not propose a mechanism for how L-LTP can persist beyond the characteristic timescales of the different molecules within the network. Observations regarding the importance of the synthesis of new proteins on L-LTP and L-LTD are confusing because the inhibition of protein synthesis has an impact on the induction, but not the maintenance, of L-LTP. However, an intriguing possibility that bi-stability at the level of protein synthesis can contribute to synaptic stability is yet to be examined in mathematical models.

Discussion

Various models for the induction and maintenance of synaptic plasticity have been proposed. Such biophysical models rely on known components of the physiology and biochemistry that lead to synaptic plasticity. These models are also required to account for known induction protocols of synaptic plasticity and for the impact of different pharmacological inhibitors. However, all models must at this point rely on approximations, as well as on not fully substantiated assumptions. In this sense all models are to a certain extent phenomenological. One problem with most existing models is that they rely on the mass action approach, which is valid only for well-mixed systems with a large number of molecules. Alternative approaches to simulations in structured, small-volume molecules are in their infancy. Therefore, current models of synaptic plasticity are not truly quantitative, but rather they are qualitative models that could be used as proof of concept. Like quantitative models, they can have robust consequences that can be tested experimentally, and we can use such models to test whether a specific mechanism can indeed accomplish what it was proposed to accomplish and whether it has additional unforeseen consequences. In the long run, and despite current limitations, computational models are the only systematic methodology that can connect the biophysical mechanisms of synaptic plasticity to experimental results observed on the cellular and system levels.

See also: Cortical Plasticity and Learning: Mechanisms and Models; Cytoskeleton in Plasticity; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Hebbian Plasticity; Long-Term Depression: Cerebellum; Long-Term Depression (LTD):

Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); NMDA Receptors and Development; Spike-Timing-Dependent Plasticity Models; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Abraham WC and Williams JM (2003) Properties and mechanisms of LTP maintenance. *Neuroscientist* 9: 463–474.

Bi GQ and Poo MM (1998) Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience* 18: 10464–10472.

Holmes WR and Levy WB (1990) Insights into associative long-term potentiation from computational models of NMDA receptor-mediated calcium influx and intracellular calcium concentration changes. *Journal of Neurophysiology* 63: 1148–1168.

Karmarkar UR and Buonomano DV (2002) A model of spike-timing dependent plasticity: One or two coincidence detectors? *Journal of Neurophysiology* 88: 507–513.

Lisman JA (1989) A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proceedings of the National Academy of Sciences of the United States of America* 86: 9574–9578.

Lisman JE, Schulman H, and Cline H (2002) The molecular basis of CaMKII function in synaptic and behavioural memory. *Nature Reviews Neuroscience* 3: 175–190.

Malenka RC and Bear MF (2004) LTP and LTD: An embarrassment of riches. *Neuron* 44: 5–21.

Markram H, Lubke J, Frotscher M, and Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275: 213–215.

Martin SJ and Morris RG (2002) New life in an old idea: The synaptic plasticity and memory hypothesis revisited. *Hippocampus* 12: 609–636.

Shouval HZ (2005) Clusters of interacting receptors can stabilize synaptic efficacies. *Proceedings of the National Academy of Sciences of the United States of America* 102: 14440–14445.

Shouval HZ, Bear MF, and Cooper LN (2002) A unified theory of NMDA receptor-dependent bidirectional synaptic plasticity. *Proceedings of the National Academy of Sciences of the United States of America* 99: 10831–10836.

Si K, Lindquist S, and Kandel ER (2003) A neuronal isoform of the Aplysia CPEB has prionlike properties. *Cell* 115: 879–891.

Sjostrom PJ, Turrigiano GG, and Nelson SB (2003) Neocortical LTD via coincident activation of presynaptic NMDA and cannabinoid receptors. *Neuron* 39: 641–654.

Wittenberg GM and Wang SS (2006) Malleability of spike-timing-dependent plasticity at the CA3-CA1 synapse. *Journal of Neuroscience* 26: 6610–6617.

Yang S-N, Tang Y-G, and Zucker R (1999) Selective induction of LTP and LTD by postsynaptic $[Ca^{2+}]_i$ elevation. *Journal of Neurophysiology* 81: 781–787.

Neuroprotection: Endogenous Mechanisms

M A Perez-Pinzon, K R Dave, R A DeFazio, A P Raval, E J Kim, and D Della Morte, University of Miami Miller School of Medicine, Miami, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

This is an article on endogenous neuroprotective strategies found in nature against anoxic/ischemic insults. Cells require oxygen and glucose to produce the energy required for function. In vertebrates, blood flow confers these necessary components. Loss of blood flow is referred to as ischemia (the deprivation of oxygen and glucose). Ischemia commonly leads to a rapid loss of cell function and the potential for cell death. The phenomenon of endogenous protection against ischemia has been observed throughout the animal kingdom. Special attention has been placed on these protective adaptations for organs that are highly susceptible to energy deprivation, such as the heart and brain. The brain has the highest energy requirements of vertebrate organs and may be the most vulnerable to energy failure. In mammals, brain is approximately 3% of the body mass, yet it receives approximately 15% of the total cardiac output and consumes 20% of the body oxygen. In mammals, anoxia (lack of oxygen), ischemia, and hypoglycemia (low glucose) each lead to the cessation of electrical activity in the brain within minutes. If these insults are prolonged, the loss of electrical activity is followed by depletion of high-energy intermediates (i.e., adenosine triphosphate (ATP), phosphocreatine) and the loss of the calcium, potassium, sodium, and chloride ion gradients across cell membranes. The loss of these gradients results in membrane depolarization. As ionic depolarization occurs, as for example in ischemia, neurons release the excitatory neurotransmitter glutamate, which in turn may amplify or directly cause irreversible pathologies during energy failure. This glutamate-mediated toxicity is termed excitotoxicity and is a major contributor to neurodegeneration following ischemia.

The pathophysiological mechanisms leading to neuronal cell death after cerebral ischemia are highly complex. Neurons continue to die over days and months after a stroke (an ischemic event occurring in the brain). This phenomenon is referred to as delayed cell death. Although not clearly defined, neuronal cell death may result from apoptosis, necrosis, or a cell death mechanism that is a mixture of these processes.

Reperfusion refers to the restoration of normal blood flow after an ischemic event. It is now understood that the reperfusion phase following cerebral ischemia contributes substantially to ischemic injury and that mitochondrial dysfunction plays a central role in this type of injury.

Although mammalian brains and hearts are highly sensitive to the loss of energy supply, as during a stroke or heart attack, these organs have an intrinsic metabolic adaptation that can confer resistance to ischemia. This state is referred to as ischemic tolerance or ischemic preconditioning (IPC). Ischemic tolerance can be induced using a brief ('sublethal') ischemic episode. After a period of recovery, the organ is resistant to a subsequent stronger ischemic insult. This induction of tolerance against ischemia has gained attention as a powerful neuroprotective metabolic state against conditions of stress such as anoxia/ischemia in heart and brain. Because stroke remains the leading cause of disability and the third leading cause of death in the Western world, the intrinsic mechanisms of ischemic tolerance are important considerations for the development of therapeutics.

We consider that the induction of tolerance by ischemic preconditioning has three distinct phases: (1) the triggering phase, (2) the signal transduction phase, and (3) the genomic phase.

Triggering Phase

Multiple possible induction pathways (triggering phase) (Figure 1) have been implicated in the induction of ischemic preconditioning: neuroactive cytokines, activation of glutamate receptors, adenosine receptors, the ATP-sensitive potassium channels, nitric oxide, oxidative stress, and hypothermia.

Signal Transduction Phase

Studies of these putative triggering factors and putative links among them have centered on the roles of several signal transduction pathways (Figure 1). For example, increases in cytosolic Ca^{2+} may play a key role, since Ca^{2+} can activate several important enzymes, with subsequent initiation of downstream signal transduction pathways for IPC-induced tolerance. Research has identified some members of the signaling pathways: protein kinase C (PKC), inducible nitric oxide synthase, and mitogen-activated protein kinases (MAPKs). PKC activation might be a common signal transduction factor in heart and brain, since Ca^{2+} can directly or indirectly activate selective PKC isozymes. PKC is also the target of the adenosine A1 receptor, which promotes PKC activation via

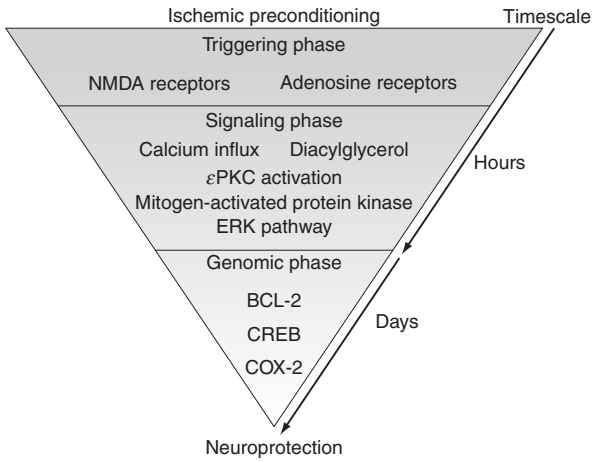

Figure 1 Schematic representation of the different phases of ischemic preconditioning with a hypothetical timescale. NMDA, *N*-methyl-D-aspartate; εPKC, protein kinase C epsilon; ERK, extracellular signal-regulated kinase; BCL-2, B cell leukemia/lymphoma-2; CREB, cAMP response element binding protein; COX-2, cyclooxygenase-2.

production of diacylglycerol (DAG) and inositol trisphosphate (Figure 1).

Support for the possible role of different PKC isozymes during IPC emerged from recent studies showing that preconditioning could be simulated with anactivator of εPKC (ψε receptor of activated C kinase (ψεRACK)). This phenomenon is referred to as pharmacological preconditioning (PPC). Blockade of ischemic preconditioning with a specific antagonist of εPKC (εV1-2) provided further evidence for the role of εPKC in ischemic preconditioning. These results demonstrated that the preconditioning-induced neuroprotective effect requires activation of εPKC. εPKC also activates the extracellular signal-regulated kinase (ERK)1/2 pathways after ischemic preconditioning in brain, and adenosine A1 receptor initiated this signaling pathway (Figure 1).

An alternative signaling pathway for IPC is mediated by MAPKs. These are components of a common signaling cascade found in the brain. These enzymes can be activated by receptor tyrosine kinases, PKC, G-protein-coupled receptors, and diverse cellular stresses (note that the mild ischemic insult used for preconditioning is also a cellular stress). The three major MAPK pathways identified are the 42 and 44 kDa ERKs and the two stress-activated MAPK families – the 46 and 54 kDa c-Jun N-terminal kinases (JNKs) and the 38 kDa p38/reactivating kinases (Figure 1).

Another study reported that *in vitro* oxygen/glucose deprivation (OGD) preconditioning induces p21ras (Ras) activation in an *N*-methyl-D-aspartate receptor-dependent and NO-dependent manner. Ras activates the (MAP or ERK kinase) MEK1/2–ERK1/2 pathways, but can also activate other alternative signal transduction pathways. Ras activity is necessary and sufficient for OGD tolerance in neurons. Pharmacological inhibition of Ras, as well as a dominant negative mutation of Ras, blocked OGD preconditioning, whereas a constitutively active form of Ras promoted neuroprotection against lethal OGD insults.

Genomic Phase

A major emphasis in this field has been placed on the role of the expression of neuroprotective genes after IPC (Figure 1). It is plausible that the increase in expression of these stress-related, neurotrophic, and immediate early genes (IEGs) in response to a mild preconditioning insult may help mediate the protection of vulnerable neurons to subsequent lethal ischemic insults.

Growth and neurotrophic factors have been implicated in protection of vulnerable regions with IPC. Neurotrophins are best known as survival factors that are capable of sustaining neurons during critical periods. For example, preconditioning by cortical spreading depression (CSD) induced brain-derived neurotrophic factor (BDNF) expression.

IEGs were shown to be activated by preconditioning. IEGs are transcriptional regulators, and members of the fos and jun families have been found to bind to the activator protein-1 (AP-1) binding site of downstream genes. AP-1 binding activity was reported to increase in a rapid linear fashion in the CA1 and CA3 regions destined for neuronal survival.

Heat shock proteins (HSPs) were also induced by preconditioning. Two minutes of ischemia (i.e., IPC) caused inhibition of protein synthesis at 1 h of recirculation, but recovery of translation was present at 5–24 h later. Induction of HSP70 was found to enhance recovery of protein synthesis, which is normally blocked during heat shock and other metabolic stresses. HSP70 expression was induced by 2 min of ischemia in gerbils, and this protected vulnerable neurons in the CA1 hippocampus from a later 5 min ischemic insult. HSPs are thought to act as molecular chaperones, preventing proteins from aggregating or denaturing under conditions of physiological stress.

Cellular Targets of Protection

As described previously, it is believed that reperfusion following cerebral ischemia contributes to ischemic injury and that synaptic and mitochondrial dysfunction play a central role. The next two sections discuss the roles of synaptic and mitochondrial dysfunction.

Synaptic Dysfunction

The link between synaptic dysfunction and brain ischemia is evident by the fact that about 50% of long-term survivors of cardiac arrest or transient ischemic attack (TIA) exhibit impaired mental abilities, manifested as learning impairment, memory disturbance, lack of concrete thinking, impaired attention and concentration, and reduced capacity to plan, initiate, and carry out mental activities.

Experimental cortical infarcts in focal cerebral ischemia models lead to an enhanced mode of synaptic plasticity, long-term potentiation (LTP) in the area surrounding the infarct. In addition, previous studies demonstrated that γ-aminobutyric acid A ($GABA_A$) receptor-mediated decreases in electrophysiological activities precede the morphologic deterioration in post-ischemic CA1 hippocampal neurons. Seizures and altered synaptic plasticity (LTP) linked to cognitive deficits strongly suggest that alterations of synapses *per se* may be involved in these post-ischemic complications.

Anoxic insults can also directly induce a form of LTP. Anoxic LTP involves post-ischemic overactivation of *N*-methyl-D-aspartate (NMDA) receptor-mediated synaptic responses. Since it has been postulated that overactivation of NMDA receptors may promote cell death, it was proposed that this type of synaptic derangement was key in delayed neuronal death.

Many groups have demonstrated that IPC promotes tolerance by ameliorating different aspects of neuronal excitotoxicity. For example, glutamate release during ischemia was ameliorated by two forms of preconditioning: CSD and by hypoxia/hypoglycemia. Decreases in glutamate release were accompanied by increases in GABA release (GABA is the major inhibitory neurotransmitter in the brain), and were linked to downregulation of the glial glutamate transporter isoforms EAAT2 and EAAT1. IPC also promoted an increase in GABA release and upregulation of $GABA_A$ receptors in the hippocampus.

In a recent study, it was found that IPC promoted a robust increase in the release of GABA after lethal ischemia (Figure 2). It was also observed that the activity of glutamate decarboxylase (the predominant pathway of GABA synthesis in the brain) was higher in the IPC group compared with control and ischemic groups.

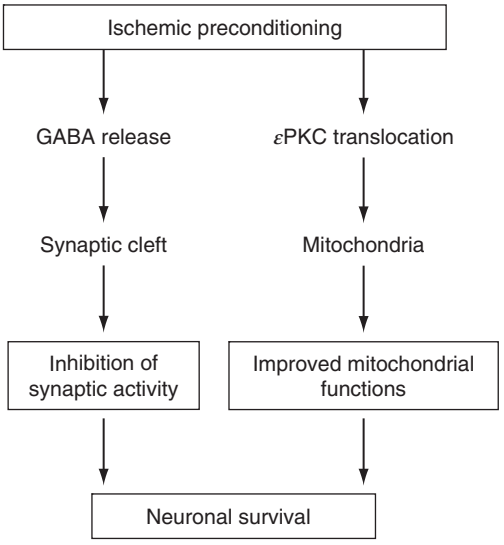

Figure 2 Chart flow depicting potential cellular targets for neuroprotection after ischemic preconditioning. GABA, γ-aminobutyric acid; εPKC, protein kinase C epsilon.

Mitochondrial Dysfunction

In the past 10 years a specific organelle, namely, the mitochondrion, has received special attention as a key player in cell death pathways. This is evidenced by numerous studies that support the hypothesis that mitochondrial dysfunction plays a central role in reperfusion injury after cerebral ischemia.

Mitochondria make ATP using oxidative phosphorylation via the mitochondrial respiratory chain. ATP is the primary source of energy for active cellular processes. Following cerebral ischemia, a prominent change in redox activity of mitochondrial respiratory chain components has been observed. Previous studies suggested that this hyperoxidation may result from loss of electron carriers from mitochondria following cerebral ischemia, such as cytochrome *c*. The loss of cytochrome *c* from mitochondria might affect respiratory chain activity and/or trigger the apoptotic cascade. The release of cytochrome *c* has been linked to apoptosis (programmed cell death).

Mitochondria respiratory deficits have also been observed after ischemia. Mitochondria isolated from ischemic brain exhibited decreases in 'state 3' respiratory rates of approximately 70% with nicotinamide adenine dinucleotide (NAD)-linked respiratory substrates. Nonsynaptosomal mitochondria were insensitive to ischemia itself; however, mitochondria became dysfunctional in the late reperfusion phase. Mitochondria from synaptic terminals were greatly affected by ischemia but partially recovered during reperfusion. The rate of oxygen consumption decreased in the CA1, CA3, and CA4 regions in the late reperfusion phase. Mitochondrial dysfunctions are reported in core and penumbra regions after focal cerebral ischemia. Isolated mitochondria exhibited decrease in adenosine diphosphate (ADP)-stimulated respiration during the early reperfusion period.

However, activity of mitochondrial electron transport chain (ETC) complex I, II–III, and IV was preserved in penumbra as well as in the core region during early reperfusion following middle cerebral artery occlusion (MCAO). During the late reperfusion phase following MCAO, the activity of mitochondrial ETC complex I is decreased. In summary, there are two phases of mitochondrial dysfunctions. In the first phase (early reperfusion), mitochondrial dysfunctions are mild (decreased activity of complex IV). However, in the second phase (late reperfusion), mitochondrial dysfunction is more pronounced.

IPC significantly protected against delayed cell death, and this protection was correlated with mitochondrial protection against the deficits in respiration affecting complexes I–IV (Figure 2). As described previously, many studies have demonstrated that reactive oxygen species (ROS) and the resulting oxidative stress play a pivotal role in neuronal cell death. There are two major regions in the ETC where ROS are produced. One is complex I and the other is complex III. Since oxidative stress is implicated in the pathophysiology that ensues after cerebral and cardiac ischemia, one can surmise that a key mechanism by which IPC protects hippocampus against delayed neuronal cell death is by protecting mitochondrial oxidative phosphorylation, diminishing ROS production, and limiting pro-apoptotic molecules to be released from mitochondria. In fact, all three mechanisms have been observed to occur in different models of IPC. In this model, the activity of the antioxidant enzymes glutathione peroxidase, glutathione reductase, and Mn superoxide dismutase was significantly increased. Furthermore, superoxide and hydrogen peroxide concentrations following OGD were significantly lower in the IPC group. Finally, cytochrome *c* release from mitochondria to the cytosol was suppressed in the ischemia tolerance-induced hippocampal CA1 region.

Cerebral Ischemic Preconditioning: Clinical Relevance

The clinical relevance of ischemic preconditioning was first described in heart. As in heart, TIA has been suggested to represent a clinical equivalent of ischemic preconditioning in brain. TIA is defined as a brief episode of neurologic dysfunction caused by focal brain or retinal ischemia, with clinical symptoms typically lasting less than 1 h, and without evidence of acute infarction. Weih et al. added support to the hypothesis that TIA may provide protection from a subsequent stroke when examining stroke outcome from 148 patients with and without a prior TIA. Furthermore, of 2490 screened stroke patients, 293 patients with ipsilateral TIA a few days before the main stroke had favorable outcomes.

Conclusions

The value of effective neuroprotective therapies for stroke patients as well as cerebrovascular surgery candidates is considerable. In the United States, stroke constitutes the third leading cause of death, and there are nearly 3 million people who have survived a stroke but live with disabilities. By limiting the extent of ischemic damage, neuroprotective therapies may reduce not only the number of deaths, but also the resulting degree of disability incurred with stroke and cerebrovascular surgery. The definition of ischemic preconditioning will definitely provide an important step toward such therapies.

See also: Mitochondrial Dysfunction in Nervous System Injury; Neuropathy: Metabolically-Induced; Neuroprotection: Pharmacological Approaches; Stroke; Stroke: Injury Mechanisms.

Further Reading

Chance B and Williams G (1956) The respiratory chain and oxidative phosphorylation. *Advances in Enzymology* 17: 65–134.

Choi DW (1987) Ionic dependence of glutamate neurotoxicity. *Journal of Neuroscience* 7(2): 369–379.

Dave KR, Lange-Asschenfeldt C, Raval AP, et al. (2005) Ischemic preconditioning ameliorates excitotoxicity by shifting glutamate/gamma-aminobutyric acid release and biosynthesis. *Journal of Neuroscience Research* 82(5): 665–673.

Dirnagl U, Simon RP, and Hallenbeck JM (2003) Ischemic tolerance and endogenous neuroprotection. *Trends in Neuroscience* 26(5): 248–254.

Fiskum G, Murphy AN, and Beal MF (1999) Mitochondria in neurodegeneration: acute ischemia and chronic neurodegenerative diseases. *Journal of Cerebral Blood Flow and Metabolism* 19(4): 351–369.

Gidday JM (2006) Cerebral preconditioning and ischaemic tolerance. *Nature Reviews Neuroscience* 7(6): 437–448.

Gonzalez-Zulueta M, Feldman AB, Klesse LJ, et al. (2000) Requirement for nitric oxide activation of p21(ras)/extracellular regulated kinase in neuronal ischemic preconditioning. *Proceedings of the National Academy of Sciences of the United States of America* 97(1): 436–441.

Liou AK, Clark RS, Henshall DC, Yin XM, and Chen J (2003) To die or not to die for neurons in ischemia, traumatic brain injury and epilepsy: A review on the stress-activated signaling pathways and apoptotic pathways. *Progress in Neurobiology* 69(2): 103–142.

Perez-Pinzon MA (2004) Neuroprotective effects of ischemic preconditioning in brain mitochondria following cerebral ischemia. *Journal of Bioenergy and Biomembranes* 36(4): 323–327.

Perez-Pinzon MA, Dave KR, and Raval AP (2005) Role of reactive oxygen species and protein kinase C in ischemic tolerance in the brain. *Antioxidants and Redox Signaling* 7(9-10): 1150–1157.

Perez-Pinzon MA, Raval AP, and Dave KR (2005) Protein kinase C and synaptic dysfunction after cardiac arrest. *Pathophysiology* 12(1): 29–34.

Perez-Pinzon MA, Xu GP, Born J, et al. (1999) Cytochrome c is released from mitochondria into the cytosol following cerebral anoxia or ischemia. *Journal of Cerebral Blood Flow and Metabolism* 19(1): 39–43.

Sick TJ, Perez-Pinzon M, Lutz PL, and Rosenthal M (1993) Maintaining coupled metabolism and membrane function in anoxic brain: A comparison between the turtle and rat. In: Hochachka PW, Lutz PL, Sick TJ, Rosenthal M, and van den Thillart G (eds.) *Surviving Hypoxia: Mechanisms of Control and Adaptation*, pp. 351–363. Boca Raton, FL: CRC Press.

Sokoloff L (1989) Circulation and energy metabolism of the brain. In: Siegel G, et al. (eds.) *Basic Neurochemistry: Molecular, Cellular, and Medical Aspects*. New York: Raven Press, Ltd.

Weih M, Prass K, Ruscher K, et al. (2001) Ischemia tolerance; model for research, hope for clinical practice? *Der Nervenarzt* 72(4): 255–260.

Neuroprotection: Pharmacological Approaches

M Singh, University of North Texas Health Science Center, Fort Worth, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuroprotective strategies vary according to the particular aspect of disease pathology that they are intended to affect. Alzheimer's disease (AD) is an age-associated, progressive neurodegenerative disorder, the etiology of which is not completely understood but for which several mechanisms have been proposed. In the following sections, in an effort to review the different pharmacological approaches used to promote neuroprotection, the example of AD will be used, whereby a description of specific pathogenic mechanisms will be followed by a discussion of what pharmacological approaches have been developed or proposed.

The Amyloid Hypothesis

The amyloid hypothesis states that the excess production, aggregation, and/or accumulation of amyloid beta (Aβ) contributes to neurotoxicity, resulting in one of the histopathological hallmarks of AD, the amyloid plaque. Aβ, a protein first described in meningeal vasculature of Alzheimer's patients, is generated through two sequential proteolytic cleavages of its precursor, a large type I membrane protein called amyloid precursor protein (APP). The first cleavage step involves the protease, beta amyloid cleavage enzyme (β-secretase; BACE1), also known as Asp2 and memapsin 2. The second proteolytic complex, termed γ-secretase, consists of at least four integral membrane proteins, presenilin, nicastrin, APH1, and presenilin enhancer 2 (PEN2).

Based on the hypothesis that the generation and/or inappropriate accumulation of Aβ results in neurodegeneration, it is easy to see how pharmacological approaches that target the synthesis, aggregation, and clearance of this protein may be advantageous. BACE inhibitors (or inhibitors of their expression) represent one class of therapeutics that is currently being developed. As an example, the nonpeptidic BACE inhibitor, GSK188909, has been found to decrease the level of Aβ produced both *in vitro* and *in vivo* and thus may have promise in reducing the neurotoxic consequences of Aβ. Alternatively, neprilysin is an Aβ-degrading enzyme. Thus, identifying pharmacological agents that regulate neprilysin expression/activity may influence Aβ levels in the brain. Antioxidants can upregulate neprilysin activity, as can somatostatin. In fact, somatostatin reduces the 42-amino-acid form of Aβ in the culture medium of primary neurons. Cerebrolysin, currently in phase III clinical trials in the United States, may also be an effective agent in reducing amyloid burden, as it has been shown to improve cognition (assessed by the Alzheimer's Disease Assessment Scale – Cognitive) and global function in AD patients. One of the mechanisms underlying this effect is believed to be through its influence on the maturation and transport of APP.

Other agents that are being considered as neuroprotectants include activity-dependent neurotrophic factor (ADNF) and humanin, both of which have been shown to be capable of suppressing amyloid toxicity. However, these compounds are not without their limitations. For example, ADNF appears to lose its protective effects at higher concentration, and humanin has a relatively low bioavailability, given its vulnerability to degradation. However, a recently described 'hybrid peptide,' colivelin, appears to hold significant promise. Colivelin is a synthetic peptide resulting from the fusion of ADNF to a derivative of humanin; it has been found to exert extremely potent protective effects against amyloid toxicity (at concentrations as low as 100 fM), protects against memory deficits, and has the advantage of being able to readily cross the blood–brain barrier.

Another approach that has been considered for use to alleviate amyloid burden is administration of an anti-Aβ vaccine (AN1792). Though several animal models have suggested its likelihood for success, the initial trial to assess the efficacy of this strategy was terminated due to the incidence of meningeoencephalitis in approximately 6% of the patients. However, alternative approaches in vaccinating against Aβ involving modified immunogens and modes of administration are still being considered and could hold promise.

Cholinergic Deficits

Cholinergic neurons, particularly in the ventral forebrain, are severely compromised in AD. This observation, when coupled with the finding that reduced cholinergic function in the cerebral cortex and hippocampus correlates with dementia severity, serves as an important basis for the cholinergic hypothesis of AD and has prompted the development of anticholinesterases. Anticholinesterases serve to reduce the breakdown of the neurotransmitter, acetylcholine, and thus enhance cholinergic neurotransmission. Interestingly,

however, these compounds not only influence the sustainability of cholinergic neurotransmission, but also appear to have neuroprotective effects through other mechanisms. For example, the cholinesterase inhibitors donepazil and galanthamine protect neuronal cells from glutamate-induced neurotoxicity. This protective effect is mediated through α4- and α7-containing nicotinic receptors, and subsequent activation of the phosphatidylinositol 3-kinase pathway. Thus, the most common and Food and Drug Administration (FDA)-approved pharmacological treatment for AD appears to be protective by two complementary systems, one that involves the maintenance of neurotransmission and another that involves activation of neuroprotective signaling pathways initiated by nicotinic cholinergic receptors.

Oxidative Stress

There is increasing evidence that oxidative stress is a prevalent component of the pathobiology of AD. Accumulating free radicals and the resulting oxidized proteins, lipids, and nucleic acids are believed to contribute to the dysfunction and death of brain cells, consistent with the free radical theory of aging. As such, antioxidants have been proposed to have neuroprotective effects that may help reduce the progression of neurodegenerative disorders, including AD. Antioxidants with neuroprotective potential can be classified as those that directly scavenge free radicals, those that inhibit the generation of free radicals, or those that work indirectly by enhancing the cellular defense systems that serve to detoxify the free radicals produced.

Examples of antioxidants are numerous and include tocopherols (e.g., vitamin E) and certain estrogens (e.g., 17*β*-estradiol). These compounds, by virtue of their chemical composition (i.e., the presence of a phenolic group), can donate a proton to saturate the unpaired electron of highly reactive oxygen species such as the hydroxyl radical. With respect to its potential benefit in protecting against neurodegeneration in AD, vitamin E has been shown to reduce oxidative stress elicited by the pro-oxidant stressor, A*β*, and to supports its potential benefit as a pharmacological treatment of the disease. However, clinical studies that have addressed the efficacy of vitamin E have been equivocal. Some epidemiological studies suggest that vitamin E may delay the onset of AD while others fail to support the benefit of this pharmacological strategy.

Excitotoxicity

Glutamate, the most abundant excitatory neurotransmitter in the brain, can activate its *N*-methyl-D-aspartate (NMDA) receptor, resulting in long-term potentiation, which in turn, is thought to be an important cellular mechanism underlying learning and memory. However, persistent activation of the NMDA receptor results in a chronic influx and accumulation of calcium, leading to neuronal dysfunction and death, a phenomenon referred to as excitotoxicity. This excitotoxicity has been proposed to play a major role in the neurodegenerative changes seen in AD. As such, an obvious therapeutic approach to reduce the neurodegeneration in AD would be to reduce this overactivation of the NMDA receptor. In fact, the most recent FDA-approved drug for the treatment of AD is the NMDA receptor antagonist, memantine. An important feature of this compound is its ability to cause an 'open channel block,' whereby the ion channel is blocked principally under conditions when it is excessively open and thus spares normal, physiological function of the NMDA receptor.

The therapeutic potential for memantine has been demonstrated in clinical trials; however, it appears that memantine does not alter the disease progression, but rather provides symptomatic benefits. This may argue that glutamate receptor overactivation or dysregulation may be an important component of the pathological cascade of events, but, in and of itself, is not the pathogenic mechanism underlying AD. Nonetheless, the use of memantine together with other pharmacological approaches may yield the necessary, multitarget strategy of treating a complex syndrome such as AD.

Loss of Trophic Support

Neurotrophins are a family of growth factors that include nerve growth factor (NGF) and brain-derived neurotrophic factor (BDNF). These trophic factors have been found to protect against cell death induced by a variety of insults that mimic specific pathologic aspects of aging and AD. Coupled with the observation that neurotrophin expression is altered in AD brains, it is not surprising that efforts to promote the restoration of neurotrophin expression and/or function have been suggested as a means of reducing neuronal dysfunction and loss in AD.

Neurotrophins function through the activation of their cognate receptors, the Trk family of tyrosine kinases and the p75 neurotrophin receptor. Typically, NGF binds preferentially to TrkA, and BDNF to TrkB, while both bind with approximately equal affinity to the p75 neurotrophin receptor. The p75 receptor has historically been ascribed a supportive role in neurotrophin signaling, whereby it helps promote high-affinity binding and selectivity of the neurotrophin to its cognate Trk receptor. More recently, however, it has become apparent that the p75 receptor may play an active role in modulating

cell survival, although the precise function of the p75 receptor may depend heavily on cell context. In cells in which p75 is co-expressed with Trk receptors, p75 may serve to enhance responses to the neurotrophins (as stated previously). However, in cells that have diminished or low levels of Trk receptors, the p75 receptor may activate death-promoting signaling cascades in a ligand-dependent manner, resulting in apoptosis.

This 'death-inducing' influence of p75 receptors has attained further significance with the recent finding that the neurotrophins are synthesized as propeptides (proneurotrophins), which, in contrast to the mature neurotrophins that have higher affinities for their respective Trk receptors, have higher affinities for the p75 neurotrophin receptor instead. Thus, neuronal dysfunction or atrophy consequent to aging or age-associated diseases may result not only from decreases in mature neurotrophin expression or function, but also potentially from increased accumulation of the proneurotrophins. In fact, Fahnestock and colleagues have described that the proneurotrophin for NGF (pro-NGF) is increased in AD brains.

However, given that neurotrophins, such as NGF, do not readily cross the blood–brain barrier, achieving therapeutic concentrations of neurotrophins in the central nervous system (CNS) may be problematic, and would require rather invasive methods to deliver the neurotrophin. Coupled with the fact that neurotrophins have pleiotropic effects (some of which may be beneficial while some may be detrimental) and have potentially poor pharmacological profiles (e.g., stability issues), the alternative pharmacological strategies have been considered. One such alternative is the development of small-molecule mimetics, which not only have the advantage of crossing the blood–brain barrier readily, but are also being designed to exhibit functional selectivity based on preferential binding to either the Trk or p75 receptors. For example, some small-molecule mimetics are aimed at mimicking the protective effects of neurotrophins by eliciting activation of protective signaling pathways triggered by activation of the Trk receptors. Others are designed to bind to the p75 receptor and can prevent the damage-promoting effects associated with binding of Aβ to the p75 receptor.

Selective Pleiotropism as a Consideration in Developing Pharmacological Neuroprotectants

Given that AD is a complex disorder that may have multiple etiologies, it is possible that the lack of sustainable efficacy of currently approved pharmacological interventions for the disease is due to having targeted only one of multiple systems going awry. Thus, it is possible that the most effective neuroprotective treatment for a disorder such as AD is one that exerts pleiotropic effects – that is, the treatment aids in reducing oxidative stress, amyloid toxicity, and excitotoxicity, and helps restore trophic support. However, pleiotropy may also have unwanted consequences. For example, it has been suggested that the significant side effects (e.g., pain) associated with neurotrophin administration to human patients are a direct result of activation of both receptor mechanisms (p75 and Trk). Thus, by developing pharmacological agents that have 'selective pleiotropy,' we may be able to take advantage of the beneficial effects while limiting the negative consequences of a particular system. The neurotrophin system serves as a useful model to explain this concept. Through a better understanding of structure–function determinants, specific ligands have been developed that take advantage of the cell survival-promoting effects of Trk receptors, while minimizing (or even antagonizing) the cell death-promoting effects (i.e., preventing Aβ from promoting toxicity through the p75 receptor).

Closing Remarks

Although there are relatively few FDA-approved pharmacological interventions to reduce neuronal dysfunction and death in diseases such as AD, there are a variety of newly recognized and promising candidates. The themes by which they are being developed are based on our current understanding of the pathobiology by which neuronal dysfunction occurs, and emphasize the principle that safe and effective pharmacological approaches for such diseases can only be realized through a thorough understanding of both the normal biology and the pathobiology of the disease.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Alzheimer's Disease: Neurodegeneration; Amyloid: Vascular and Parenchymal; Cholinergic System; Cholinergic Pathways in CNS; Enteric Nervous System: Neurotrophic Factors; Excitotoxicity in Neurodegenerative Disease; Growth Factors: Neuronal Atrophy; Inflammation in Neurodegenerative Disease and Injury; Neuropathy: Metabolically-Induced; Neuroprotection: Endogenous Mechanisms; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Oxidative Damage in Neurodegeneration and Injury.

Further Reading

Alvarez XA, Cacabelos R, Laredo M, et al. (2006) A 24-week, double-blind, placebo-controlled study of three dosages of Cerebrolysin in patients with mild to moderate Alzheimer's disease. *European Journal of Neurology* 13: 43–54.

Barker PA (1998) p75NTR: A study in contrasts. *Cell Death & Differentiation* 5: 346–356.

Beal MF (1995) Aging, energy, and oxidative stress in neurodegenerative diseases. *Annals of Neurology* 38: 357–366.

Behl C (1999) Alzheimer's disease and oxidative stress: Implications for novel therapeutic approaches. *Progress in Neurobiology* 57: 301–323.

Chiba T, Yamada M, Hashimoto Y, et al. (2005) Development of a femtomolar-acting humanin derivative named colivelin by attaching activity-dependent neurotrophic factor to its N terminus: Characterization of colivelin-mediated neuroprotection against Alzheimer's disease-relevant insults *in vitro* and *in vivo*. *Journal of Neuroscience* 25: 10252–10261.

Dasilva KA, Aubert I, and McLaurin J (2006) Vaccine development for Alzheimer's disease. *Current Pharmaceutical Design* 12: 4283–4293.

Fahnestock M, Michalski B, Xu B, et al. (2001) The precursor pro-nerve growth factor is the predominant form of nerve growth factor in brain and is increased in Alzheimer's disease. *Molecular and Cellular Neuroscience* 18: 210–220.

Hardy J and Selkoe DJ (2002) The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics. *Science* 297: 353–356.

Lee R, Kermani P, Teng KK, et al. (2001) Regulation of cell survival by secreted proneurotrophins. *Science* 294: 1945–1948.

Longo FM and Massa SM (2005) Neurotrophin receptor-based strategies for Alzheimer's disease. *Current Alzheimer Research* 2: 167–169.

Markesbery WR (1999) The role of oxidative stress in Alzheimer disease. *Archives of Neurology* 56: 1449–1452.

Reisberg B, Doody R, Stoffler A, et al. (2003) Memantine in moderate-to-severe Alzheimer's disease. *New England Journal of Medicine* 348: 1333–1341.

Vassar R and Citron M (2000) A beta-generating enzymes: recent advances in beta- and gamma-secretase research. *Neuron* 27: 419–422.

Watkins JC and Jane DE (2006) The glutamate story. *British Journal of Pharmacology* 147(supplement 1): S100–S108.

Neuroproteomics

M O Collins and S G N Grant, Wellcome Trust Sanger Institute, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The ultimate goal of studying proteomes is to achieve a comprehensive view of the characteristics and function of every protein present within a biological system. This has only recently become feasible since the completion of genome projects which have defined the basic set of genetic instructions which can code for a multitude of proteomes. The complexity of the genome is overshadowed by that of the transcriptome, which describes the full complement of mRNA transcripts transcribed from the genome of the cell. This, in turn, gives rise to yet another level of complexity seen in that of the cellular proteome. The complexity of proteomes is controlled by a variety of transcriptional, translational, and posttranslational regulatory points, each of which can modify the expression of genetic information in response to the dynamic cellular state (Figure 1). The covalent modification of mature proteins is one of the most important points of dynamic regulation of proteomes, and of the over 300 or more different protein modifications which have been reported, phosphorylation and glycosylation are prototypic.

Comprehensive proteome analysis also seeks to elucidate protein subcellular distribution and protein–protein interactions, which are crucial in determining the integration of large numbers of proteins into interaction networks, and functional molecular complexes that ultimately perform a myriad of cellular functions (Figure 2). The composition of such multiprotein complexes or organelles is variable among cell types, organs, regions of organs (such as different brain regions), and species. In addition, the proteome is highly dynamic and temporal and thus describes the complement of proteins expressed in a cell at a single time.

Understanding proteomes at many of these levels of complexity is undoubtedly necessary for dissecting the molecular basis for many complex neurological diseases. The starting point for neuroproteomics has been to define and map synaptic proteomes in model organisms as well as humans. The synapse is a subcellular specialization that transmits and converts patterns of electrical activity into cellular memory. This processing of electrical information is mediated by the protein components of the synapse and the phosphorylation of these components is an important part of the regulation of this process. This specialized proteome can be biochemically isolated and the protein components can be identified by mass spectrometry (MS). Establishment of such a molecular map permits insight into the organization and functioning of the synaptic proteome.

There are many approaches used to study large numbers of proteins in the brain, from the comparative analysis of protein expression to the mapping of molecular complexes. A number of examples of such approaches are reviewed here to highlight the diversity of proteomic applications in neuroscience. Herein, we review how these key proteomic methodologies are being applied to the study of neurobiology and describe the recent paradigm shifts in the understanding of synaptic biology.

Proteomic Mass Spectrometry

MS permits the identification of proteins by the determination of the exact mass of peptides and the fragmentation of these peptides to determine the sequence (and thus the identification) of the peptide and the parent protein. Protein samples, such as purified organelles or protein complexes are usually separated by their relative molecular weight on sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) gels. The protein bands are visualized by dyes such as Coomassie brilliant blue and then are excised from the gel. The protein-containing gel slices are washed and incubated with proteolytic enzymes such as trypsin, and the resultant peptides extracted from the gel slices (Figure 3).

A mass spectrometer consists of an ion source, a mass analyzer that measures the mass-to-charge ratio (m/z) of peptides, and a detector that counts the numbers of ions at each m/z value. Electrospray ionization (ESI) mass spectrometers use a heated capillary needle through which the peptide solution elutes and is subsequently ionized and introduced into the mass spectrometer. This type of ion source is readily compatible with liquid chromatographic (LC) separations, resulting in an LC-MS/MS (liquid chromatography-coupled mass spectrometry) setup in which thousands of peptides can be identified in a single experiment. In a typical LC-MS/MS experiment, a peptide mixture is separated by analytical reversed-phase chromatography in a high-performance liquid chromatography (HPLC) system and is introduced into the mass spectrometer via the ESI source. Multiply protonated peptides give rise to the first mass spectrum (MS1), and then the computer generates a prioritized list of the most intense precursor ions detected, which are

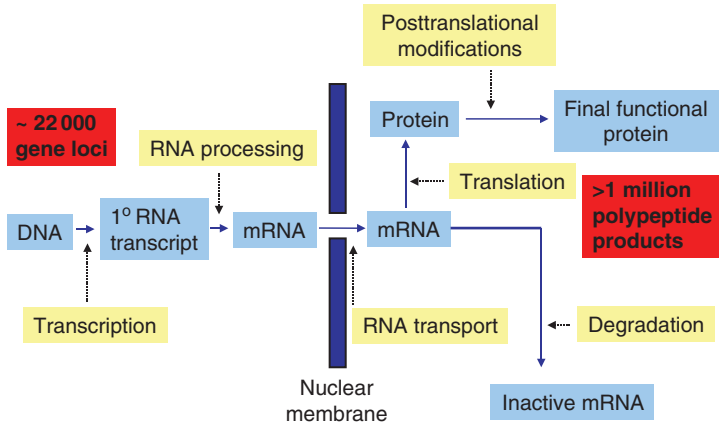

Figure 1 Generation of molecular complexity. There are at least six points at which the complexity of the information encoded in the genome is increased by transcriptional, translational, and posttranslational control. At the posttranslational level, more than 300 different types of modifications, including protein phosphorylation, exist. These processes can easily transform the 22 000 gene loci in the human genome into an estimated 1 million different polypeptide forms, greatly increasing genome-to-proteome complexity.

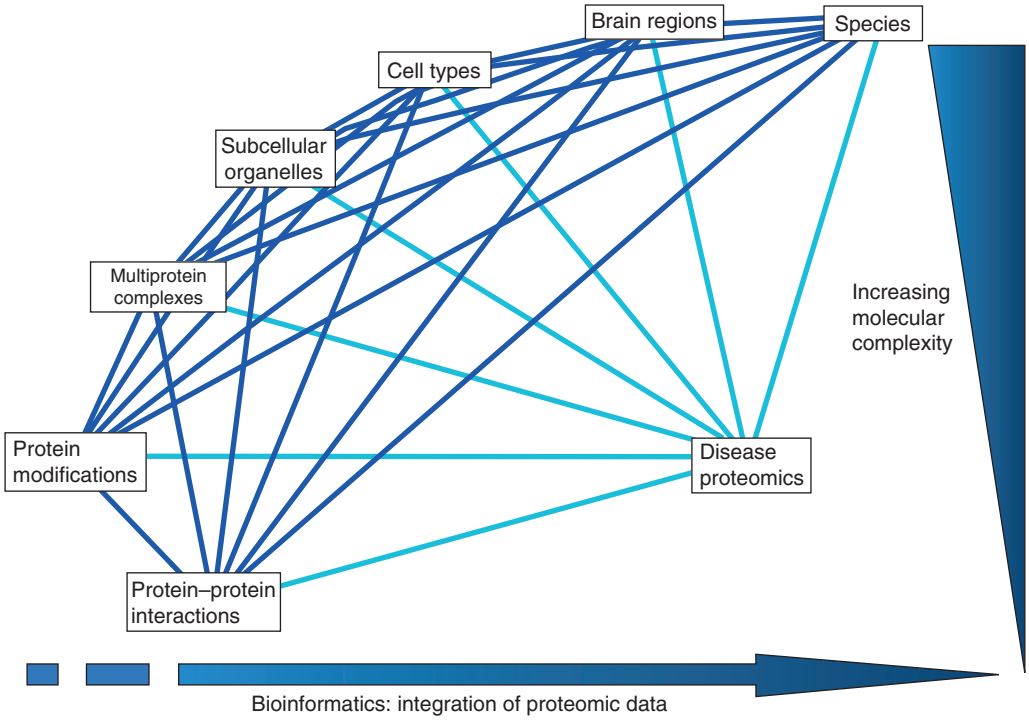

Figure 2 Complexity of proteomes. Once a polypeptide has undergone splicing, posttranslational modifications, and so forth, there are further levels of proteome complexity relating to how proteins exist in a cell. Proteomes from different species, tissues, or cell types differ not only in their composition but also in their protein interactions, protein complexes, and organelles. Complex diseases are usually caused by changes in more than one level in this complexity, and therefore disease proteomics should take this into account. An essential point about this complexity is the need for adequate bioinformatics approaches to integrate such proteome data.

subsequently subjected to fragmentation by energetic collision with gas (collision-induced dissociation (CID)) in the collision cell. This MS/MS spectrum contains data relating to the amino acid composition of the peptides and the pattern of fragmentation that occurred. This process of tandem MS provides *m/z* values of peptides and fragmentation data for each peptide. These data are then used by sequence database-searching programs such as MASCOT or SEQUEST to provide probability-based peptide matches and therefore protein identifications.

Subcellular Proteomics

The observation of cells under an electron microscope for the first time in the 1940s revealed a previously

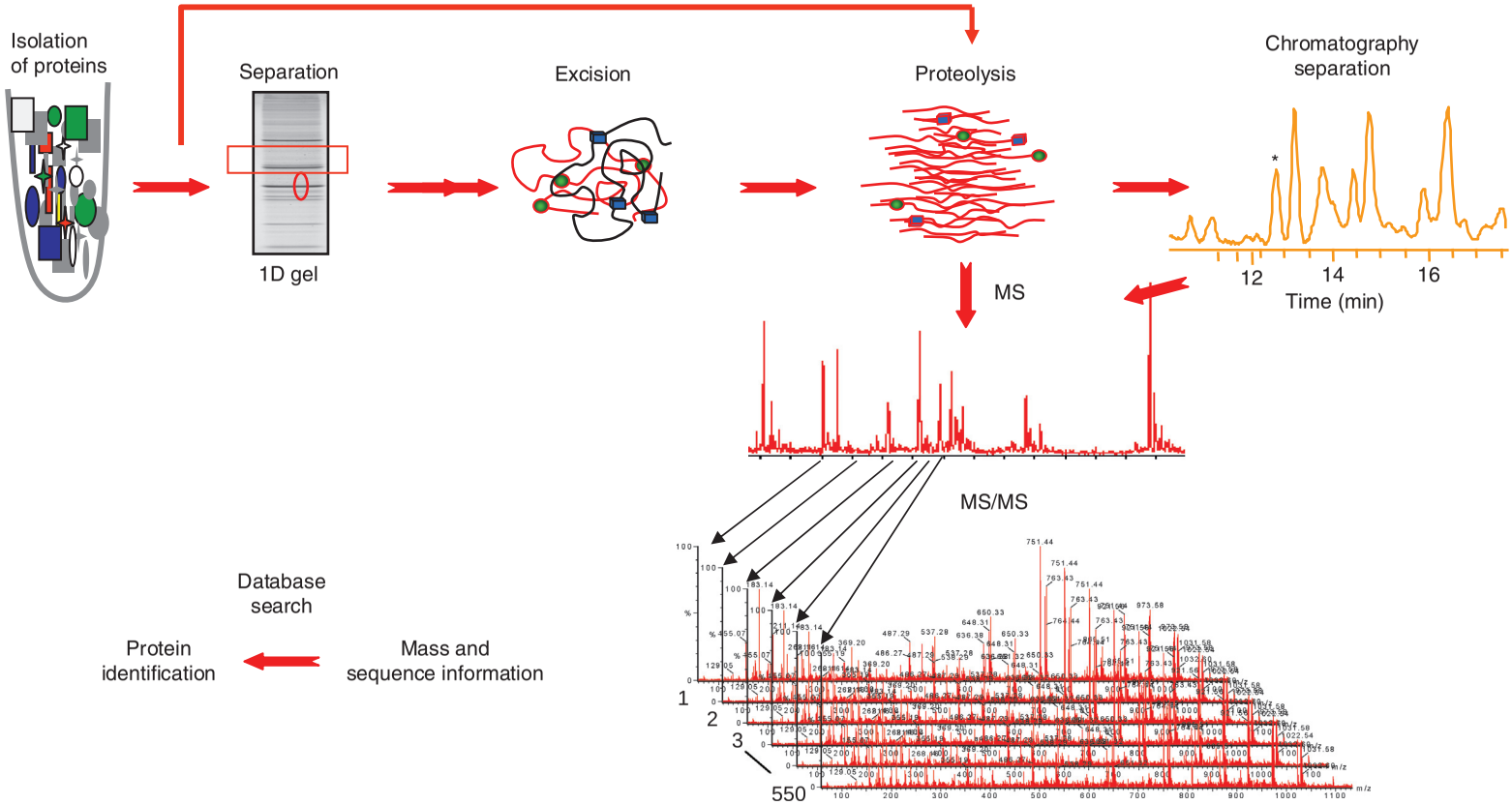

Figure 3 Workflow of an LC-MS/MS experiment. A mixture of peptides from a protein sample digest is separated by reversed-phase chromatography on a nanoflow HPLC. The peptides elute from the RP column and are ionized by an electrospray source. In the first stage of mass spectrometry, *m/z* values and charge states for each precursor ion are determined and the most abundant precursor ions are selected for analysis in the second stage. The ions are then fragmented with by collision-induced dissociation (CID) with a gas to produce fragment ions which are detected. Using the mass (from MS-1) and sequence information (from MS-2), protein sequence databases are searched to provide peptide identifications and protein matches. The analysis of phosphopeptides is similar except that specific scanning approaches can be used in the mass spectrometer to detect the loss of phosphate during CID. Assignment of the exact position of phosphorylation is accomplished by manual interpretation of the mass spectra. HPLC, high-performance liquid chromatography; LC, liquid chromatography; *m/z,* mass-to-charge ratio; MS, mass spectrometry; RP, reversed-phase.

unexpected level of complexity. Extensive compartmentalization was observed and was found to be achieved by intracellular membranes. This realization prompted intensive investigation of subcellular organelles and in the following decades led to the development of a number of methodologies to physically isolate cellular organelles. The identification of the protein composition of subcellular organelles has been a significant challenge for many years and is only now becoming routine with sophisticated MS-based platforms (Table 1). Most subcellular organelles are isolated using density-gradient centrifugation, in which a gradient of sucrose solutions is used to partition different organelles according to their density. A limitation of this approach is the difficulty in the assessment of the degree of purity of the isolated fractions. Usually, the presence and absence of subcellular markers is used to indicate the level of contamination of cosendimenting cellular debris.

Recently, with the advent of quantitative MS-based approaches, it is possible to assess the relative amounts of each protein present in a fraction compared to the starting material of the isolation. This approach was successfully applied to the identification of proteins in the centrosome. In this study, researchers quantitatively profiled hundreds of proteins across several centrifugation fractions and identified proteins which were quantitatively enriched in centrosomes, thereby identifying bona fide centrosomal proteins.

Traditionally, the protein content of purified organelles are separated by either one-dimensional (1-D) or two-dimensional (2-D) SDS-PAGE, and protein-containing bands or spots are excised, digested, and subjected to analysis by MS. 1-D SDS-PAGE has the advantage of being particularly well suited for the separation of membrane proteins, whereas 2-D SDS-PAGE provides better separation of protein species. However, both of these approaches require band or spot excision, which is very labor intensive, is not very reproducible, and causes significant sample loss due to incomplete peptide extraction from the gel slices/spots. These difficulties have been overcome, in part, by gel-free separations of peptides such as multidimensional protein identification technology (MudPIT) pioneered by John Yates, in which the protein sample is denatured and digested in solution and the resultant peptides are separated in the first dimension by ion-exchange chromatography and in the second dimension by reverse-phase chromatography. The first application of this methodology was the profiling of the yeast proteome, and this facilitated identification of 1484 proteins, far exceeding the numbers of protein identifications from traditional methods.

Table 1 Summary of proteomic data from subcellular organelles studies[a]

Organelle	*Number of proteins*
Postsynaptic proteome	1124
Mitochondria	847
Phagosome	525
Nucleolus	489
Exosome	295
Spliceosome	292
Neutrophil granules	286
Lipid raft	241
Lysosome	215
Clathrin-coated vesicles	209
Nuclear pore	174
Golgi	136
Centrosome	64

[a]The large numbers of proteins in the postsynaptic proteome probably reflects the fact that this cellular specialization contains elements of other organelles, such as mitochondria and Golgi, as well as integral postsynaptic membrane proteins.

Profiling Synaptic Proteomes

The postsynaptic density (PSD) is a structural component of the synapse that can be observed under the electron microscope, beneath the postsynaptic membrane. Recent proteomic efforts have provided an insight into the complexity of the PSD and its subcomponents (Figure 4). The last 5 years have seen tremendous growth in our knowledge of the protein components of mammalian synapses. The first investigation of the synapse proteome appeared in the year 2000, with the identification of the components of the *N*-methyl-D-aspartate (NMDA) receptor complex (77 proteins) (Figure 4). The large number of proteins in this complex was at first controversial because it exceeded the total number of known PSD proteins. This report was soon followed by the first MS-based characterization of the PSD, which revealed just 31 proteins in the PSD, most of which were already known.

Since then, a number of proteomic analyses of isolated PSDs have been reported and the most common approach used was as follows. Synaptic fractions were first isolated from mouse forebrains by differential centrifugation on sucrose gradients. PSD fractions were isolated from the crude synaptic fraction by the relative insolubility of PSDs in Triton X-100. PSD proteins were separated by SDS–PAGE (Figure 5), and in one example, 37 gel slices were analyzed by LC–MS/MS. Some 7402 peptides were identified, corresponding to 620 nonredundant protein identifications, with an average of 34 proteins identified per gel slice (corresponding to an average of 16.7 unique proteins per gel slice).

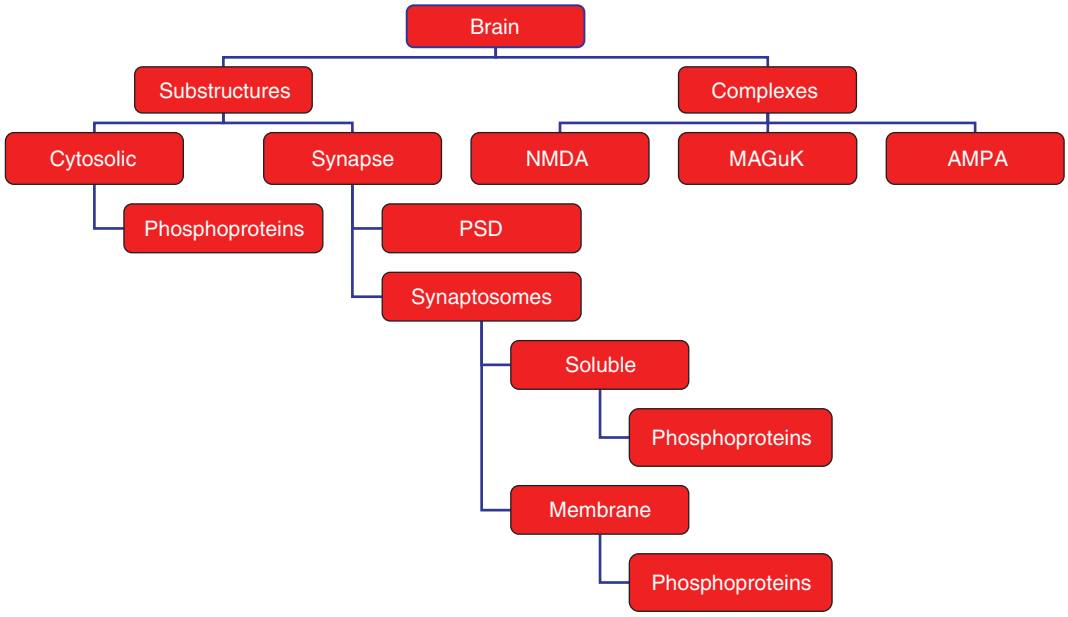

Figure 4 Examples of brain proteomes. A number of different types of brain proteomes with different levels of complexity have been studied. These can be classified into many categories, ranging from subcellular organelles such as the postsynaptic density, to multiprotein complexes such as glutamate receptor complexes (NMDA, MAGUK, and AMPA) to subproteomes which comprise phosphorylated proteins. All these proteomes are very likely to be heterogeneous in different cell types and in different brain regions. AMPA, α-amino-3-hyroxy-5-methyklisoxazole-4-proprionic acid; MAGUK, membrane-associated guanylate kinase; NMDA, *N*-methyl-D-aspartate; PSD, postsynaptic density.

When data from seven of these PSD analyses (and data from glutamate receptor (GluR) complexes) are compiled into one postsynaptic data set, a nonredundant list of 1124 postsynaptic proteome (PSP) proteins is obtained. In order to define a set of higher confidence, proteins that were identified two or more times (466 proteins) can be placed in a group termed the consensus PSD (cPSD). The majority of these proteins are abundant and so are more likely to be detected in multiple analyses, and the set of proteins detected once is perhaps of lower abundance or may reflect differences in sample preparation. The data integration performed on these PSD data sets has resulted in a confidence-ranked database of postsynaptic proteins which totals 1124. This tenfold increase in known postsynaptic components since the year 2000 reflects the technological advances of MS-based protein identification compared to traditional methods.

The initial approach to characterize the organization of the PSP was to classify the proteins into categories that describe their known functional properties (Table 2). Within the overall PSP, representatives of many classes of proteins representing a broad range of cell biological functions were observed: membrane-bound receptors; adhesion proteins and channels; a plethora of signaling proteins and adaptors; and proteins involved with regulation of transport, RNA metabolism, translation, and transcription. The cPSD contains similar numbers of proteins in most classes compared to the total PSD data set, with the exception of channels and receptors, signaling proteins, and synaptic vesicle proteins. This is most probably due to the lower abundance of these proteins, resulting in a decreased likelihood of detection. This supports the idea that the cPSD represents a consensus of protein identifications which, on the one hand, decreases individual data set bias for contaminating proteins (nuclear or mitochondrial) and, on the other hand, has limited coverage of some protein classes because of abundance issues.

Synaptic Multiprotein Complexes

Proteomic approaches for the molecular dissection of multiprotein complexes usually involve purification of the entire complex by a variety of affinity methods, and protein identification by Western blotting or MS-based approaches. The first proteomic analysis of multiprotein complexes relevant to the brain was the purification and identification of the molecular constituents of the NMDA receptor–adhesion protein signaling complexes (NRC). The NMDA receptor has a low abundance in neurons compared with the overall protein content. Therefore, it is important to generate highly enriched samples containing the target molecule. This can be achieved using a variety

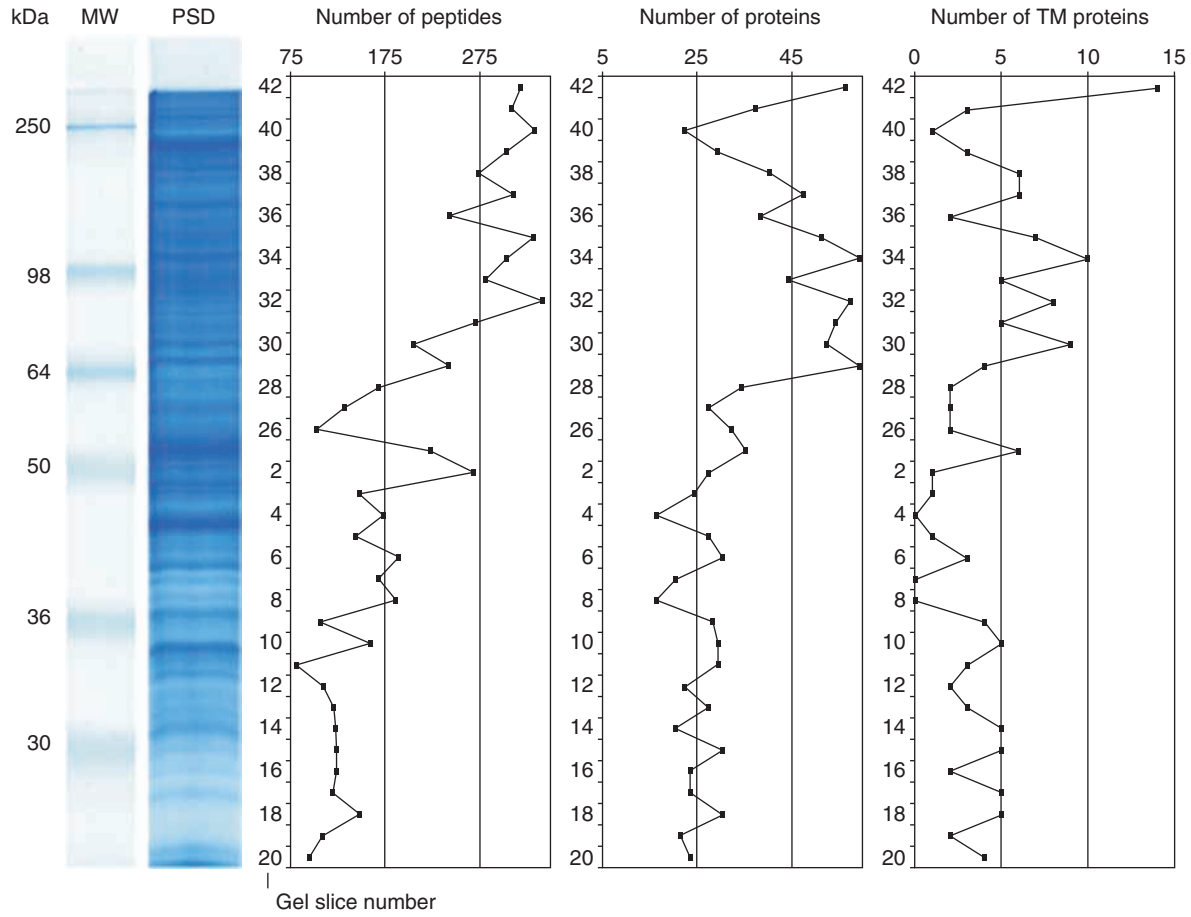

Figure 5 Identification of postsynaptic density (PSD) proteins by MS analysis of gel slices using a coomassie-stained SDS-PAGE gel lane containing purified mouse forebrain PSD proteins. The resulting numbers of peptides, proteins, and transmembrane (TM) proteins identified in each gel slice by LC-MS/MS analysis are shown. 1D SDS-PAGE separation offers effective protein separation of complex mixtures with an average of 34 proteins identified per gel slice and an average of 12 peptides characterized for each protein. The number of peptides detected per gel slice increases proportionally with molecular weight. 1D SDS-PAGE is very compatible with membrane samples, and the distribution of proteins with TM regions is shown. LC, liquid chromatography; MS, mass spectrometry; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis. Adapted from Collins MO, Husi H, Yu L, et al. (2006) Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *Journal of Neurochemistry* 97(supplement 1): 16–23.

Table 2 Summary of PSD protein classifications[a]

	Total PDS	*cPSD*
Channels and receptors	80	39
MAGUKs/adaptors/scaffolders	54	33
Ser/Thr kinases	46	20
Tyr kinase	3	2
Protein phosphatases	18	11
G-proteins and modulators	77	34
Signaling molecules and enzymes	278	98
Transcription and translation	119	41
Cytoskeletal and cell adhesion molecules	153	97
Synaptic vesicles and protein transport	159	76
Novel	107	11
Other	30	4
Total	1124	466

[a]Classified to show protein class enrichment. cPSD, consensus PSD; MAGUKs, membrane-associated guanylate kinases; PSD, postsynaptic density.

of methods, generally involving affinity chromatography steps. These can range from ligand affinity (drugs, substrates, or co-factors) to the more widely used approach of immunoprecipitation, involving specific antibodies to the molecule of interest. A major drawback in using antibody-based approaches for proteomic analysis is the introduction of the IgG itself, which interferes with the analysis by overshadowing proteins of interest that co-migrate during the gel separation. All these ligands, groups, or proteins are then coupled to a resin either irreversibly (usually in the case of small molecules) or through a secondary affinity linkage specific for the bait molecule. Neuronal extracts are then incubated with the resins bearing the bait, and bound proteins separated from one another as well as from the matrix by conventional gel analysis.

Identification of isolated proteins can be readily achieved using Western blotting or MS analysis such as that described for the analysis of PSDs. Using an antibody-based approach targeting the NR1 subunit of the NMDA receptor, the NRC was shown to comprise 77 proteins organized into receptor, adaptor, signaling, cytoskeletal, and novel proteins, of which 30 are implicated in binding studies and another 19 participate in NMDA receptor signaling.

NMDA receptor and metabotropic glutamate receptor (mGluR) subtypes were linked to cadherins and L1 cell-adhesion molecules in complexes lacking α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors. These neurotransmitter–adhesion receptor complexes were bound to kinases, phosphatases, guanosine triphosphatase (GTPase)-activating proteins and Ras with effectors, including mitogen-activated protein kinase (MAPK) pathway components. A striking feature of the composition is that there appears to be modules or sets of signaling proteins that are known to make up key components of signal transduction pathways that can be distinctly regulated. For example, all the molecules necessary to induce the phosphorylation of MAPK following NMDA receptor stimulation are present in the NRC (including calmodulin-dependent protein kinase II (CaMKII), SynGAP, Ras, MAPK kinase (MEK), and extracellular signal-regulated kinase (ERK)). Their modules may allow the NMDA receptor and mGluR to integrate signals within the complex and then to couple downstream cellular effector mechanisms, such as trafficking of AMPA receptors and cytoskeletal changes that mediate structural and physiological plasticity. Furthermore, at least 18 NRC constituents are regulated by synaptic activity, indicating that the composition of the complex is dynamic. In the hippocampus, these activity-dependent genes are known to undergo specific temporal changes following the induction of plasticity.

This NRC data set was further expanded from 77 to 186 proteins using a combination of peptide-affinity chromatography and MS identification of new components of the complex. Purification of membrane-associated guanylate kinase (MAGUK)-associated signaling complexes (MASCs) was achieved by peptide-affinity chromatography using a hexapeptide corresponding to the C-terminus of the NMDA receptor NR2B subunit. This approach extended the NRC data set to comprise 186 proteins, which constitute the MASC. The importance of this complex is highlighted by the fact that mutations or polymorphisms in 47 MASC genes are associated with 183 human disorders, of which 54 are nervous system disorders. In addition, 43 MASC proteins have been reported to be important in synaptic plasticity, and 40 have been associated with rodent behavior.

AMPA receptor complexes have been purified using immunoaffinity chromatography, resulting in the identification of ten unique proteins. The mGluR5 receptor complex has been characterized in a similar fashion and was found to comprise 76 proteins, including the NMDA receptor subunit NR2A. The majority of proteins in all three GluR complexes have been independently identified as PSD proteins (Figure 6). These complexes also contain proteins which have not yet been detected in PSD analyses, presumably due to abundance issues in the analyses of PSD preparations compared to enriched protein complexes. It is important to note that the majority of the components of these complexes have been validated as PSD proteins and that the overlap between these complexes occurs in the cPSD (Figure 6), supporting the notion that the cPSD is an important subset of the total PSD data set. This is consistent with a postsynaptic organization in which the MASC signaling complex is connected to multiple cell biological effector mechan-

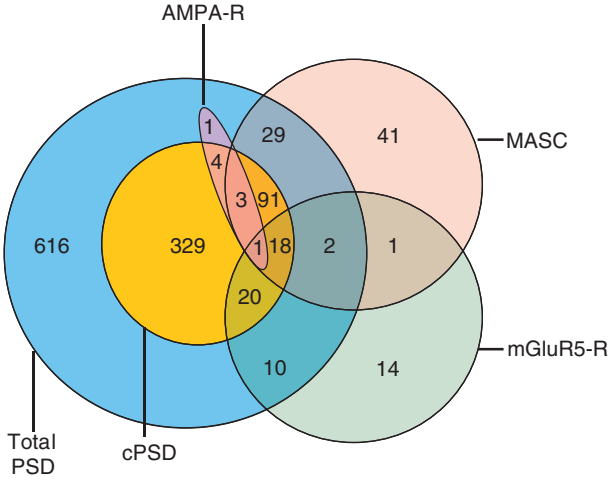

Figure 6 Multiprotein complexes in the postsynaptic proteome. Venn diagram illustrating the overlap of three glutamate receptor complexes (AMPA, MASC, and mGluR5) with the total and cPSD datasets. It can be seen that the majority of overlap between components of these complexes and the PSD occurs in the high-confidence cPSD dataset. Proteins detected in these multiprotein complexes, which were not found in any of the PSD datasets, are generally of low abundance and are enriched in immunopurifications of complexes compared to whole PSD analyses. AMPA, (α-amino-3-hyroxy-5-methyklisoxazole-4-proprionic acid; AMPA-R, AMPA receptor; cPSD, consensus PSD; MAGUK, membrane-associated guanylate kinase; MASC, MAGUK-associated signaling complex; mGluR5, metabotropic glutamate receptor 5; mGluR5-R, metabotropic mGluR5 receptor complex; PSD, postsynaptic density. Reproduced from Collins MO, Husi H, Yu L, et al. (2006) Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *Journal of Neurochemistry* 97(supplement 1): 16–23, with permission from Blackwell Publishing.

isms organized into their respective complexes. In addition to MASC and AMPA complexes, components of other complexes such as cell adhesion, growth factor, cytoskeletal, transport, and ribosomal complexes were found in the PSP. These three receptor protein complex data sets, together with PSD data sets, provide for the first time a general composition of the PSP, and further mapping of other receptor protein complexes will undoubtedly refine our view of the molecular architecture of the synapse.

Synapse Phosphoproteomics

Approaches to Probe the Phosphoproteome

Traditional biochemical approaches to analyzing phosphorylation usually involve focused study of one phosphoprotein and often of one particular phosphorylation site. Antibodies that specifically recognize the phosphorylated epitope can be made and used to purify the phosphoprotein and to detect the phosphoprotein in Western blots or for imunohistochemical or immunocytochemical analyses. Generation and validation of such phospho-specific antibodies is a time-consuming process, taking up to 1 year to obtain an antibody that can be used confidently. The use of ^{32}P-labeled γ-adenosine triphosphate (ATP) for *in vitro* kinase reactions has been employed widely, and separation of labeled phosphorylated peptides by thin-layer chromatography (TLC) provides information about the number phosphopeptides and the intensity of phosphorylation. Analysis of these phosphopeptides from TLC plates has been performed using MS (matrix-associated laser desorption ionization time-of-flight (MALDI-TOF)) and Edman sequencing. However, this methodology is suitable for single-phosphoprotein studies, and the use of ^{32}P is hazardous.

Phosphoproteome enrichment strategies Development of MS and in particular LC-MS/MS in the late 1990s permitted the identification of large numbers of peptides in proteome-scale experiments. The use of MS for the characterization of phosphorylation on a proteome scale was hampered by the characteristics of phosphorylation itself. It has been estimated that one-third of any given proteome is phosphorylated and the stoichiometry of phosphorylation is often low; so to study phosphorylation on a global level, phospho-enrichment is a prerequisite.

Digestion of a proteome or even a portion of a proteome can result in the generation of tens of thousands of tryptic fragments. This number of peptides is overwhelming in an LC-MS/MS experiment, and because the mass spectrometer is usually operated in a mode which selects only the most intense ions for sequencing, only the most abundant portion of the proteome is sampled. Many signaling proteins are present at low abundance in cells, and the portion of these proteins that are phosphorylated at any one time may also be quite low. Therefore, in a standard LC-MS/MS experiment, the chance of detecting and sequencing a phosphopeptide is very low indeed. A number of strategies to overcome these problems were developed, involving either physically enriching for phosphorylated species or isolating phosphopeptide-specific signals in the mass spectrometer.

A number of chemical modification strategies have been developed in which the phosphate group has been replaced with a moiety that is chemically more stable than phosphate. One such method employs β-elimination of the phosphate from phospho-threonine or phospho-serine and results in the formation of dehydroaminobutyric acid or dehydroalanine, respectively. These products can be detected using tandem MS or in another method, Michael addition is used to add a reactive thiol group after β-elimination to provide a linker for the attachment of an affinity tag. This affinity tag is usually biotin, and it permits purification of the chemically modified (previously phosphorylated) peptides. This type of chemical modification of phosphate is not applicable to phosphotyrosine residues and suffers from side-chain reactions in which cysteine and methionine can be modified. Other chemical modification protocols, which are applicable to all phosphorylated residues, have been developed, but they involve many chemical reactions and have not been used successfully on real proteomic samples.

These chemical modification-based strategies were challenged in 2002 with the modification of the traditional immobilized metal-affinity chromatography (IMAC) methodology. IMAC purification of phosphopeptides is based on the attraction of negatively charged phosphate groups to positively charged immobilized (on a resin) metals. The main problem associated with this protocol is that acidic groups of aspartic and glutamic acid and electron donors (histidine) can bind to the positively charged metal, thereby contaminating the eluted sample. This nonspecific binding was addressed by chemical esterification of acidic residues before the IMAC enrichment. Application of this modified procedure to the yeast proteome permitted the characterization of 216 phosphopeptides.

Mass spectrometry-based strategies for phosphorylation analysis The use of MS for phosphorylation-site identification has increased greatly over the last decade, superseding traditional approaches of 2-D phosphopeptide mapping and Edman sequencing. In the period from 1992 to 2003, 1281 phosphorylation

sites were characterized by MS in 203 publications. In the latter years of this period, advances in sample preparation enabled hundreds of phosphorylation sites to be determined and reported in a single publication. In addition, the concomitant advances in technology of mass spectrometers have also facilitated the transition from single-phosphoprotein analysis to the analysis of phosphoproteomes. Identification of phosphopeptides and the characterization of phosphorylation sites by LC-MS/MS analysis is more difficult than standard protein identification experiments and can be achieved by special phosphoscanning functions in the mass spectrometer. Neutral loss scanning is a popular approach used to detect phosphopeptides, and it involves monitoring for the loss of ions characteristic of phosphorylated residues. When peptides containing phosphoserine or phosphothreonine are subjected to CID, they usually undergo gas-phase β-elimination, resulting in a neutral loss of phosphoric acid (-98 Da) or are dephosphorylated (-80 Da). The resulting mass spectrum indicates the position of the loss of phosphate by a spacing of 69 Da (dehydroalanine) or 83 Da (dehydroaminobutyric acid). This type of analysis requires prior knowledge of the charge state of the precursor ion before CID and thus is highly compatible with quadrupole time-of-flight (Q-TOF) mass spectrometers operated in data-dependent acquisition mode. This approach is not as efficient for detecting phosphotyrosine because phosphotyrosine is more resistant to neutral loss. However, another scanning approach is suitable for the detection of phosphotyrosine and is based on the production of reporter ions characteristic for the loss of phosphotyrosine. Immonium ions of phosphotyrosine have a mass-to-charge ratio (m/z) value of 216.043 and are generated by the double cleavage of the peptide backbone. These can be detected with good sensitivity, with detection possible using subpicomolar amounts of gel-separated protein. In addition, when phospho-Ser/Thr/Tyr-containing peptides are fragmented by CID in the negative ion mode, a characteristic PO^{3-} ion ($m/z = -79$) can be detected by precursor ion scanning. Once this ion is detected, the mass spectrometer can be set to automatically switch to positive ion mode to conduct MS/MS for sequencing.

Challenges of phosphoproteome analysis The main challenge for phosphoproteome analysis is the enrichment of phosphorylated proteins and peptides to sufficient quantities so that they can be characterized by MS. Preservation of phosphorylation during enrichment is essential because any loss of phosphorylation due to phosphatase activity or by chemical dephosphorylation significantly reduces the likelihood of detection of low levels of phosphorylation. Phosphoproteome-enrichment methods should also be as straightforward as possible to minimize sample losses and to enhance the reproducibility of the protocol. Another important challenge of proteomic or phosphoproteomic experiments is the analysis of the large volumes of data that can be produced. After identification of the phosphorylation sites, comprehensive bioinformatic analysis is necessary to annotate the phosphorylation sites. Extensive literature mining for functional information pertaining to phosphoproteins, phosphorylation sites, and also the sequence surrounding the phosphorylation sites should be performed. The use of prediction algorithms to annotate domains, amino acid sequence characteristics, and potential kinases and binding sites is necessary. Without this kind of postanalysis annotation, lists of phosphopeptides are of limited initial value. In a similar manner, the generation of long lists of protein identifications in proteome profiling experiments presents the problem of data analysis. It is necessary to systematically annotate such lists with as much relevant data as possible to extract the value and biological significance of the data.

Application to the Synapse Phosphoproteome

The complexity of the synapse in terms of its components and the processes it carries out necessitates a complex and robust regulatory system. Phosphorylation at the synapse has been studied for over 35 years, and the identification of phosphorylation sites has been an arduous process. The study of synaptic phosphorylation on a protein-by-protein level was inherently biased, with attention focused on channels and receptors, which were the earliest class of proteins characterized at the synapse by pharmacological methods. Undoubtedly, phosphorylation of this class of proteins is important, but the identification of phosphorylation sites in a global unbiased manner is necessary to expand our understanding of synaptic signaling beyond local regulation of channel function.

Recently, a number of studies have focused on the identification of synaptic phosphorylation sites by MS. One of the largest of these studies, which employed a sequential protein and peptide IMAC enrichment, resulted in the detection of over 650 phosphorylation events, corresponding to 331 sites (92% of which are novel) on 79 phosphoproteins (**Figure 7**). An additional 149 candidate phosphoproteins were identified by profiling the composition of the protein IMAC enrichment. This degree of novelty in terms of newly identified phosphoproteins as well as phosphorylation sites highlighted a previously unrecognized complexity in synaptic phosphorylation.

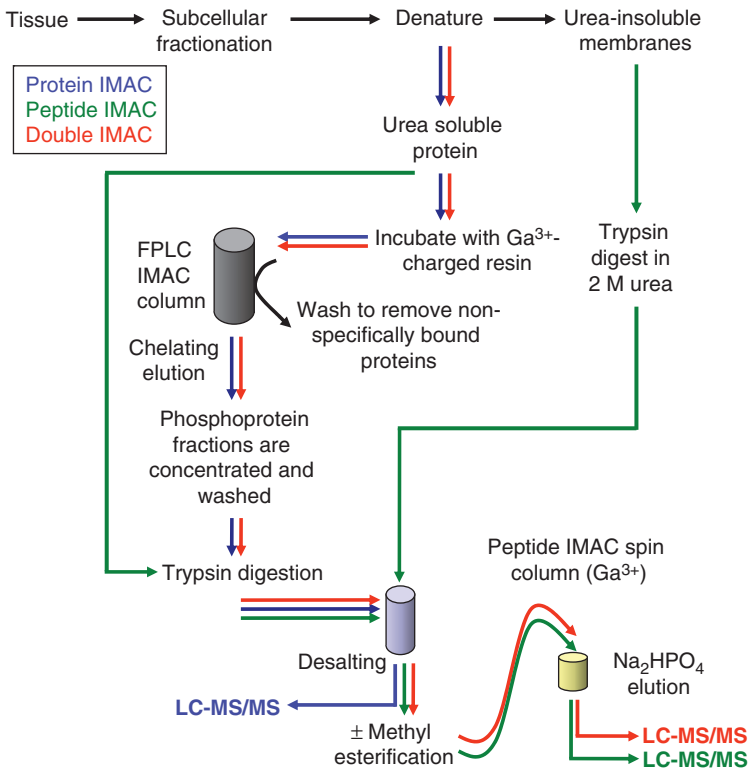

Figure 7 Schematic overview of methodology. The target proteome, in this case a subcellular fraction, is phase partitioned by its differential solubility in 6 M urea. The soluble fraction can be processed in two ways. Traditionally, this fraction is digested in solution, desalted, and applied to a peptide IMAC column (green arrows) with or without methyl esterification and analyzed by LC-MS/MS. In addition, we have developed top-down enrichment step in which the soluble proteins are applied to a protein IMAC column (blue arrows) and the eluted phosphoproteins are digested and also analyzed by LC-MS/MS. A novel combination of these two approaches (red arrow) in which a two-step enrichment of phosphoproteins and phosphopeptides and analysis by LC-MS/MS results in improved identification of phosphorylation sites. The urea-insoluble fraction, which represents tightly bound or integral membrane proteins, can be analyzed by direct digestion in 2 M urea to release peptides from cytoplasmic domains of these proteins, which can be applied to a peptide IMAC column to isolate phosphopeptides. These combined strategies, when applied to the synapse phosphoproteome, yielded the identification of over 650 phosphorylation events corresponding to 331 phosphorylation sites. This combined strategy of phase partitioning and sequential enrichment of phosphoresidues allows the comprehensive analysis of a complex sample. FPLC, fast-protein liquid chromatography; IMAC, immobilized metal-affinity chromatography; LC, liquid chromatography; MS, mass spectrometry.

To examine the significance of this complexity, the representation of phosphorylation data across synaptic proteins and important synaptic subcomponents and organelles was examined. This data set contained postsynaptic components including the PSD (134 proteins) and NMDA receptor-associated proteins (44 proteins) and presynaptic components, of which the most striking examples were Piccolo-Bassoon transport vesicles. A prominent feature of the synapse phosphoproteome was the appearance of many phosphorylated postsynaptic scaffold proteins. Major components of the NMDA receptor complex, such as PSD-95, Chapsyn-110, and SynGAP, were all phosphorylated, indicating that the scaffolding functions of these proteins may be regulated by phosphorylation. This suggested that the PSD scaffold, which provides a molecular matrix that supports and connects synaptic membrane receptors with intercellular signaling pathways, is itself subject to regulation.

A prominent theme which emerged from the study of the synapse phosphoproteome was that of kinase multiplicity. Data derived from *in silico* predictions of cognate kinases for identified phosphorylation sites and *in vitro* phosphorylation of peptide arrays suggested that a small number of kinases could phosphorylate many proteins and that each substrate is phosphorylated by many kinases. This phenomenon may represent signal integration (i.e., when signaling pathways converge and result in multisite phosphorylation of substrate). Another possibility that is not mutually exclusive is that kinase multiplicity represents robustness in signaling networks. The existence of a number of variant signaling pathways involving different kinases/phosphatases or substrates, which have the same downstream target (e.g., transcription

factor activation), would ensure signal transduction robustness. Finally, the complexity of synaptic signaling may require subtle fine-tuning, which may be mediated by combinations of different kinases working on the same substrates, resulting in slightly altered activity or localization of the substrate.

Neuroproteomics Perspective

The strength of MS-based proteomics lies in its ability to generate vast amounts of unbiased data that permit rapid characterization of proteomes and phosphoproteomes, which would not be possible using conventional techniques. The complete characterization of the protein inhabitants of every organelle in a cell type would provide valuable information on protein localization, provide annotation for the large numbers of proteins of unknown function, and provide a starting point for large-scale functional investigations of the biology of each organelle. Such data would be invaluable for neuroscience research, and the application of this approach to synapses would certainly provide an unprecedented parts list and platform for the study of synaptic biology.

The most important and striking point to come out of the research reviewed here is that the PSP is incredibly complex. The large number of proteins in the PSP presents a daunting challenge to systematically attribute synaptic function to each protein. However, for the first time we are in a position where we actually know the identity of the majority of proteins in the PSP and are able to ask questions concerning the entire PSP, as opposed to discrete subsets of proteins, as in the past.

Now that the field of proteomics is coming to the end of its developmental phase, the application of proteomic methodologies to biological problems should provide even more leaps of knowledge in the next 5 years. Analysis of the phosphoproteome is not currently routine; however, phosphoproteomic methodologies are maturing. It is anticipated that full phosphoproteome coverage is not achievable from analyses of basal states in tissue because some phosphorylation sites are of such low abundance in the absence of a specific stimulation. Therefore, the real challenge is to move from a discovery phase of phosphorylation-site identification to a profiling phase of phosphorylation in response to stimulation of cells or tissues with drugs and inhibitors. Similarly, the recent development of quantitative MS will facilitate the profiling of dynamic proteome states. The use of bioinformatics in the analysis of these data will be of the utmost importance, and this is an area that will need much attention if the value of such data is to be fully appreciated. We can envisage that the application of these new approaches using well-defined chemical (NMDA) or electrical stimulation paradigms (long-term potentiation (LTP)) in cultured neurons or tissue slices will undoubtedly reveal novel and interesting aspects of the complexity of synaptic plasticity.

See also: Mass Spectroscopy of Proteins; Postsynaptic Density/Architecture at Excitatory Synapses; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Aebersold R and Goodlett DR (2001) Mass spectrometry in proteomics. *Chemical Reviews* 101: 269–295.

Carlin RK, Grab DJ, Cohen RS, and Siekevitz P (1980) Isolation and characterization of postsynaptic densities from various brain regions: Enrichment of different types of postsynaptic densities. *Journal of Cell Biology* 86: 831–845.

Collins MO, Husi H, Yu L, et al. (2006) Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *Journal of Neurochemistry* 97(supplement 1): 16–23.

Collins MO, Yu L, Coba MP, et al. (2005) Proteomic analysis of *in vivo* phosphorylated synaptic proteins. *Journal of Biological Chemistry* 280: 5972–5982.

Collins MO, Yu L, Husi H, Blackstock WP, Choudhary JS, and Grant SG (2005) Robust enrichment of phosphorylated species in complex mixtures by sequential protein and peptide metal-affinity chromatography and analysis by tandem mass spectrometry. *Science Signal Transduction Knowledge Environment* 2005: p16 [online].

Farr CD, Gafken PR, Norbeck AD, et al. (2004) Proteomic analysis of native metabotropic glutamate receptor 5 protein complexes reveals novel molecular constituents. *Journal of Neurochemistry* 91: 438–450.

Ficarro SB, McCleland ML, Stukenberg PT, et al. (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnology* 20: 301–305.

Gavin AC and Superti-Furga G (2003) Protein complexes and proteome organization from yeast to man. *Current Opinion in Chemical Biology* 7: 21–27.

Husi H and Grant SG (2001) Isolation of 2000-kDa complexes of *N*-methyl-D-aspartate receptor and postsynaptic density 95 from mouse brain. *Journal of Neurochemistry* 77: 281–291.

Husi H and Grant SG (2001) Proteomics of the nervous system. *Trends in Neurosciences* 24: 259–266.

Husi H, Ward MA, Choudhary JS, Blackstock WP, and Grant SG (2000) Proteomic analysis of NMDA receptor–adhesion protein signaling complexes. *Nature Neuroscience* 3: 661–669.

Jordan BA, Fernholz BD, Boussac M, et al. (2004) Identification and verification of novel rodent postsynaptic density proteins. *Molecular and Cellular Proteomics* 3: 857–871.

Kim M, Jiang LH, Wilson HL, North RA, and Surprenant A (2001) Proteomic and functional evidence for a P2X7 receptor signaling complex. *EMBO Journal* 20: 6347–6358.

Manning G, Whyte DB, Martinez R, Hunter T, and Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298: 1912–1934.

Sheng M and Kim MJ (2002) Postsynaptic signaling and plasticity mechanisms. *Science* 298: 776–780.

Neuropsychological Testing

D C Delis, University of California at San Diego, La Jolla, CA, USA; and San Diego veterans Affairs Healthcare System, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The psychometric assessment of cognitive and motor functions has proved to have considerable utility in the diagnosis of neurological and psychiatric disorders. Many insidious disorders that evade early identification using the most advanced medical technologies (e.g., neuroimaging; electroencephalography, EEG) can often be effectively diagnosed using neuropsychological testing. Following are examples of neurological conditions that depend heavily on neuropsychological assessment for accurate diagnosis:

- Individuals with early Alzheimer's disease often have normal neurological examinations, magnetic resonance imaging (MRI) and computed tomography (CT) brain scans, and EEG findings. However, studies have found that neuropsychological testing has proved to be one of the most effective procedures for the earliest diagnosis of this pervasive disorder.
- Some elderly patients with severe depression present with memory complaints and low scores on mental status exams that are comparable to those of patients with early Alzheimer's disease, thereby making it difficult to distinguish between these two conditions. However, neuropsychological tests can differentiate between the qualitatively different memory profiles of depressed individuals versus those of early Alzheimer's patients with approximately 95% accuracy.
- Children with fetal alcohol syndrome (FAS), who often have significant deficits in cognitive development, are typically first identified by their facial dysmorphic features. However, many children who are exposed to alcohol *in utero* have normal facial features but go on to develop levels of cognitive impairment that are comparable to those of children with FAS. Neuropsychological testing is important for identifying children who were prenatally exposed to alcohol but who do not have the facial dysmorphic features in order to determine if they should receive early educational interventions for cognitive impairments.
- Some individuals who have had serious head traumas suffer brain damage and cognitive impairments and yet have normal findings on neuroimaging. Other individuals can have head trauma without brain damage, but exaggerate cognitive deficits for external incentive. Neuropsychologists have developed empirically based methods for distinguishing between head-injured individuals with bona fide cognitive deficits secondary to brain damage versus head-injured individuals without permanent brain damage who attempt to feign cognitive impairment on cognitive testing.

Historical Perspective

The field of clinical neuropsychology evolved in the 1940s through the 1960s primarily as a tool to aid in addressing the very broad question of whether a patient had organicity or brain damage. The need to address this general question was important at that time, because CT and MRI brain scans had not yet been developed. In addition, the primary focus of the neurological examination has always involved testing of reflexes and senses, both of which are mediated by the sensorimotor cortex and connecting pathways. This focus leaves vast regions of the brain that can be damaged but without altering sensorimotor functions. Instead, neuropathology in these other regions often affects cognitive skills. When brain damage is subtle to moderate, the decline in cognitive functions can also be mild to moderate. Neurologists and other physicians often conduct a 'mental status exam,' which is a cursory assessment of cognitive skills, such as asking the patient to remember three words. Such brief, nonstandardized exams usually are capable of identifying patients with severe brain damage, but they frequently fail to identify patients with mild to moderate neurocognitive dysfunction.

The need to identify individuals with mild to moderate brain damage outside the sensorimotor circuitry was one of the primary reasons why psychologists began studying neurological patients. Since the early 1900s, psychologists have been armed with powerful methodologies for rigorously measuring subtle differences in mental abilities. The fields of psychometrics and IQ testing, two of the most important contributions in the history of psychology, have long provided clinicians with the means for reliably measuring the development of intellectual and cognitive functions in children and adolescents. In the hands of the neuropsychologist, these methodologies offer the ideal procedures for gauging even subtle declines in intellectual and cognitive functions in individuals with brain dysfunction.

The pioneers in this area were a handful of psychologists during the 1940s through the 1960s; working at a time before there was a field of clinical neuropsychology, these psychologists had the vision to see the

tremendous utility of adapting psychometric tests to the assessment of cognitive decline in brain-damaged patients. These early leaders included, among others, Arthur Benton, Kurt Goldstein, Harold Goodglass, Ward Halstead, Donald Hebb, Edith Kaplan, Alexander Luria, Brenda Milner, Ralph Reitan, Andre Rey, Hans-Lukas Teuber, Elizabeth Warrington, and David Wechsler. In administering a wide variety of cognitive tests to patients with focal brain damage, psychologists soon discovered another vital contribution of their methodologies: more accurate characterizations of neuropsychological syndromes with different profiles of cognitive strengths and weaknesses. This work resurrected the importance of the early writings (late 1800s and early 1900s) of neurologists on the aphasias, agnosias, apraxias, and other syndromes, and brought a psychometric rigor to the descriptions of these syndromes. In addition, the emergence of cognitive science in the 1960s also brought more sophisticated experimental procedures for isolating and studying the integrity of specific cognitive processes in patients with focal brain lesions.

In modern times, the need for neuropsychologists to identify organicity or brain damage *per se* has decreased with the development of more sensitive structural and functional neuroimaging techniques. Although there continues to be some neurological disorders (e.g., Alzheimer's disease) that begin with such microscopic lesions that neuropsychological testing is still required for early detection, neuroimaging will likely soon be capable of detecting these insidious disorders as well. However, with advances in the detection of more subtle neurological conditions, and with such diverse individual differences in how similar levels of brain damage can affect cognitive functions, there continues to be growing need for clinical neuropsychologists to document each patient's profile of cognitive and motor strengths and weaknesses. In this context, clinical neuropsychology has become a flourishing field of practice.

The Basic Premise of Neuropsychological Testing

As a general rule, cognitive and motor functions tend to fall within the normal range of the bell curve in an individual with a healthy brain, with some relative strengths and weaknesses. However, when one or more regions of the brain have been damaged by injury or disease, the cognitive and motor functions that are mediated at least in part by those regions often will become significantly impaired. Accordingly, the clinical neuropsychologist will typically administer to a patient with known or suspected brain damage 20–30 different tests assessing attention, language skills, math abilities, visual–spatial functions, new learning and memory, problem solving, abstract thinking, and other cognitive and motor skills. In past research, these tests were administered to large numbers of neurologically intact individuals (the normative group) and to patients with different types of brain damage; the neuropsychologist uses these normative and clinical data to draw conclusions about whether a patient's raw scores on the tests fall within the expected or impaired ranges. This determination is not an exact science, since many neurologically normal individuals will have some test scores that fall in the impaired range, and some brain-damaged patients, especially if they have relatively high premorbid cognitive skills, may have few scores in the impaired range. Nevertheless, by examining the patient's profile of cognitive and motor strengths and weaknesses, the neuropsychologist can make inferences about the integrity of different brain regions and can begin to formulate hypotheses about the presence or absence of different neurological disorders.

What Comprises a Neuropsychological Evaluation?

The typical neuropsychological evaluation takes 6–8 h. For most teenagers and adults, the examination can be completed in a day, with ample breaks. For younger children, older adults, or patients whose endurance has been significantly compromised by a neurological, medical, or psychiatric condition, the examination can be conducted across two or three sessions.

In the morning and early afternoon hours of a 1-day exam, patients are typically administered the cognitive and motor tests when they are usually more alert. Some neuropsychologists administer the tests themselves, whereas others employ a trained psychological technician to administer the tests; both approaches are acceptable in the field, since the tests have standardized administration procedures. In the afternoon, patients who have adequate cognitive skills are often administered a psychological inventory designed to assess their self-reported emotional functioning. In addition, a comprehensive clinical interview is also typically conducted. In this era of managed care, many patients are first given neuropsychological screening exams, which usually vary from 30 min to 3 h in duration. If a patient exhibits some evidence of cognitive impairment on the screening exam, then more extensive neuropsychological testing may be conducted.

Selection of Tests

Neuropsychologists typically select tests that have empirically documented reliability and validity;

adequate, updated normative data; and empirically demonstrated utility in identifying neurocognitive deficits in brain-damaged patients. Modern neuropsychological tests often contain improvements over their predecessors by the incorporation of constructs from cognitive science, thereby making the tests more sensitive to the assessment of specific cognitive processes. Some neuropsychologists use a 'battery approach,' meaning that they will administer all tests or subtests that comprise a particular battery or scale (e.g., all subtests of the Wechsler Memory Scale-III). Other neuropsychologists use an 'eclectic approach,' meaning that they will pick and choose only certain tests or subtests from different batteries or scales (e.g., many psychologists administer only two subtests of the Wechsler Memory Scale-III: the Logical Memory subtest and the Visual Reproduction subtest). In addition, some neuropsychologists adopt a 'fixed approach,' meaning that they try to give the same tests to all patients, whereas other neuropsychologists use a 'flexible approach,' meaning that they will tailor the selection of tests administered to each patient based on the patient's presenting problems and referral question. These different approaches are all acceptable in the field, provided that the major domains of cognitive, motor, and behavioral functions are adequately covered by the tests selected. Table 1 lists a number of tests that are commonly used by neuropsychologists to assess cognitive and motor functions in children and adults.

The Clinical Interview

A critical part of the neuropsychological evaluation is the clinical interview. A comprehensive survey of possible risk factors for brain damage should be explored in the interview. In addition, many factors beyond brain damage *per se* can affect performance on rigorous psychometric tests of cognitive and motor skills, and these factors should also be covered in the interview. Important areas to explore in the interview include early developmental problems (e.g., pregnancy and delivery complications; prenatal exposure to alcohol, drugs, or other teratogenic agents; delays in achieving developmental milestones as an infant or toddler); cultural factors (e.g., English as a second language, being raised in another country); educational history (e.g., level and type of education obtained, history of learning disability or attention problems); occupational history (e.g., the degree to which an individual's work requires higher level thinking skills); history of medical illnesses that are risk factors for brain dysfunction (e.g., hypertension, diabetes); history of head injury or loss of consciousness; past and current psychiatric disorders; past and current alcohol and drug use; current medications (e.g., narcotic analgesics); current emotional state; current physical state (e.g., headaches and other pain symptoms, which can affect performance on psychometric tests); and family history of medical and psychiatric disorders (e.g., a family history of Alzheimer's disease is a risk factor for this disorder). Based on information gathered in the clinical interview, the neuropsychologist attempts to identify which risk factors may be present in an individual and how they may affect the person's performances on the cognitive and motor tests.

Another step in the clinical interview is to ask the patient, and family members if they are available, about their perceptions of the patient's cognitive complaints, the onset of the problems, and whether the problems have changed over time. Sometimes this information can be helpful in evaluating, for instance, whether a patient's cognitive difficulties started abruptly, as with a stroke, or gradually, as in Alzheimer's disease. However, the self-reported cognitive difficulties of patients and their family members often fail to correlate with objective test results. For example, depressed patients sometimes report severe memory difficulties when their actual scores on objective tests of memory may be normal or near normal.

Finally, the clinical interview allows the neuropsychologist to observe directly patients' emotional demeanor and behavioral functioning at a time when they are discussing their problems and symptoms. These observations can be valuable for identifying whether a psychiatric component is contributing to a patient's presentation.

Test-Taking Effort

Scores on difficult cognitive and motor tests are valid only if the examinee exerts adequate effort in taking them. An important part of the neuropsychological evaluation is the assessment of the individual's motivation to perform well on the tests. Some individuals exert minimal effort on neuropsychologist tests for secondary gain. For example, an inmate may intentionally perform poorly on memory tests to support his or her claim of amnesia for the accused crime. As another example, an individual involved in litigation related to a car accident may claim that he or she suffered a brain injury in the accident. When this individual undergoes a neuropsychological evaluation, he or she may fail to exert adequate effort on the tests, with the hope of looking 'brain damaged' and obtaining a large settlement. The frequency with which individuals have been found to exaggerate cognitive problems on neuropsychological tests has been reported to be as high as 30% of cases referred in the context of forensic evaluations.

Table 1 Examples of commonly used neuropsychological tests for adults and children

Domain	*Representative test*
Adults	
Intellectual functions	Wechsler Adult Intelligence Scale-III
Learning and memory	California Verbal Learning Test-II
	Rey Auditory Verbal Learning Test
	Rey Osterrieth Complex Figure Test
	Warrington Recognition Memory Test
	Wechsler Memory Scale-III
Visuospatial abilities	Block Design Subtest (WAIS-III)
	Hooper Visual Organizational Test
	Judgment of Line Orientation
	Rey Osterrieth Complex Figure Test
Language	Boston Diagnostic Aphasia Exam
	Boston Naming Test
	Controlled Oral Word Association Test
	Token Test
	Vocabulary Subtest (WAIS-III)
Executive functions	Category Test
Abstraction ability	Delis–Kaplan Executive Function System
	Similarities Subtest (WAIS-III)
	Trail Making Test
	Verbal Fluency Tests
	Wisconsin Card Sorting Test
Attention	Digit Span Subtest (WAIS-III)
	Digit Vigilance Test
	Spatial Span (WMS-III)
Motor Functions	Finger Tapping Test
	Grooved Pegboard Test
	Grip Strength (Hand Dynamometer)
Emotional/behavioral functions	Beck Depression Inventory
	Hamilton Depression Scale
	Millon Clinical Personality Inventory
	Minnesota Multiphasic Personality Inventory-II
Screening instruments	Kaplan Baycrest Neurocognitive Assessment
	Mattis Dementia Rating Scale
	MicroCog
	Repeatable Battery for the Assessment of Neuropsychological Status
Academic achievement	Wechsler Individual Achievement Test
	Wide Range Achievement Test-III
Children	
Intellectual functions	Cognitive Assessment System
	Wechsler Intelligence Scale for Children-IV
	Woodcock–Johnson Tests of Cognitive Ability
Specific neuropsychological domains	California Verbal Learning Test–Children's Version
	Children's Memory Scale
	Delis-Kaplan Executive Function System
	NEPSY Neuropsychological Battery
	Peabody Picture Vocabulary Test-III
	Wide Range Assessment of Memory and Learning
	WISC-III as a Neuropsychological Instrument
Emotional/behavioral functions	Child Behavior Checklist
	Personality Inventory for Children

Neuropsychologists have developed a number of tools for evaluating an individual's test-taking effort, including specific cognitive 'malingering' tests, analysis of typical and atypical profiles of neurocognitive dysfunction, consistency of test findings and profiles across repeat evaluations, and examination of whether a patient's low scores on neuropsychological tests are consistent with how he or she is functioning in everyday life. Even with these procedures, however, it is sometimes difficult to identify a clever patient who is feigning cognitive difficulties, especially if he or she does so to a mild degree (more

blatant forms of exaggeration are easier to detect). As reported in the literature, some individuals have admitted to having been 'coached' by their attorneys as to which tests are designed to detect malingering, prior to their neuropsychological evaluations. In addition, individuals can learn about malingering tests on the Internet. For these reasons, neuropsychologists strive to use different methods for the detection of inadequate effort and to stay one step ahead of public knowledge by developing new effort-testing techniques.

Premorbid Level of Cognitive Functioning

The most common reason that patients are referred for a neuropsychological evaluation is to determine whether they have acquired brain damage from some neurological insult, and, if so, whether they have experienced a decline in their level of cognitive functioning from a premorbid (preexisting) state. However, inferences about declines in cognitive functioning must include estimates of the individual's level of cognitive functioning before the onset of the insult. Neuropsychologists employ several techniques for estimating or correcting for premorbid level of cognitive functioning, including the following approaches:

- Testing overlearned cognitive skills that tend to be resilient to the effects of more diffuse brain damage and that correlate significantly with preexisting IQ (e.g., tests of individual word reading or vocabulary level are good predictors of premorbid verbal IQ).
- Developing normative test data that are corrected not only for age but also for education level, a strong predictor of premorbid IQ.
- Obtaining prior scholastic test scores from the individual's school records, since these scores correlate robustly with an individual's verbal IQ as an adult.
- Using regression analyses based on key demographic variables, such as education and occupation levels, to derive formulas for estimating premorbid IQ.

The estimation of premorbid level of cognitive functioning is a difficult task. Each method should be considered as a tool that may or may not be useful or even appropriate for a particular patient. For instance, the use of education-corrected norms can be invaluable for one patient and misleading for another. Following are three case examples that illustrate this point:

- A physician with 20 years of education and a premorbid IQ of 130 suffers a left hemisphere stroke. While her postinjury scores on language tests are within the average (normal) range for her normative age group, these scores fall in the impaired range for her education group (i.e., compared to age-matched individuals with approximately 20 years of education). In this case, the education-corrected norms are warranted because they provide an empirically based method for documenting evidence of acquired cognitive decline in this individual.
- A self-made businessman has a premorbid IQ of 130 but left high school after his sophomore year because his father was killed in a car accident and he had to begin working to help support his family. As an adult, this individual suffers a left hemisphere stroke and obtains scores on language tests that fall in the average (normal) range using both age- and education-corrected norms (i.e., compared to age-matched individuals with approximately 10 years of education). However, these standardized scores are likely misleading, because they are obtained from a normative reference group that likely has a lower mean premorbid verbal IQ than that of the patient. For this individual, it would have been more appropriate to use one or more of the other methods to estimate the patient's premorbid level of cognitive functioning (e.g., obtain school records to see if scholastic test scores are available for this individual; administer a single-word reading test).
- At birth, an individual suffers a brain insult secondary to delivery complications. In school, he is found to have an IQ that falls in the borderline mentally retarded range. He completes only 8 years of education. As an adult, he is evaluated by a neuropsychologist who scores the test findings using education-corrected norms (i.e., his raw scores are compared to age-matched individuals with approximately 8 years of education). The patient's standardized scores using these norms fall in the low-average range and the neuropsychologist claims that the patient does not have brain damage. However, the use of education-corrected norms is inappropriate for this case. The patient's relatively low education level was likely one of the 'effects' of his early brain insult (not the cause), and to correct his scores for a low level of education essentially negates one of the main effects of his brain damage. For this individual, age-corrected norms would be more appropriate to document that the patient shows evidence of cognitive impairment (i.e., borderline mental retardation) relative to his age-matched peers.

As a general rule, prior scholastic or IQ test scores that are obtained from a patient's school records when he or she was around 10 years of age or older are often the best method for estimating premorbid level of intellectual functioning, because achievement and IQ scores tend to be relatively stable across the life span. In addition, each person's premorbid

functioning should be estimated individually, given that there are significant individual differences in IQ level among people with similar levels of education. Other factors that contribute to discrepancies in IQ among people with similar levels of education include motivation to achieve in school (e.g., a person with average IQ and high motivation may achieve a high level of education, whereas a person with superior IQ and low motivation may drop out of school early); differences in the quality of the education (e.g., university-based graduate school versus private professional graduate school); and English as a second language (e.g., a highly educated person may have learned English as a second language and, as a result, obtain only average scores on English-based verbal tests, despite achieving a relatively high level of education). Unfortunately, for the majority of patients evaluated, school test results are not accessible, necessitating the use of one of the other methods for estimating premorbid functioning. Finally, it is important to note that the different methods for estimating premorbid mental abilities may apply to certain domains of cognitive functioning but not to others. For example, the ability to read individual words of varying difficulty has been found to be a significant predictor of verbal IQ. However, verbal IQ tends to have relatively low correlations with executive functions and memory skills (accounting for 0–16% of the variance). Thus, while current reading ability can be a useful predictor of premorbid verbal IQ, it tends to be less effective in predicting premorbid level of executive functions and memory skills.

Analysis of Test Performance

After the neuropsychological tests have been administered, raw scores are summed for each test and converted to standardized scores. In addition to these quantitative scores, many neuropsychologists also examine test performance for qualitative features that may be pathognomonic of brain dysfunction. For instance, a patient may speak primarily in content words and omit the little grammatical words of language (e.g., the patient says, "Hospital . . . Monday" when trying to say that he came into the hospital on Monday). This qualitative feature of the patient's language, when seen consistently, is pathognomonic of a type of language disorder called Broca's aphasia, and likely reflects a brain injury in the posterior region of the left frontal lobe. As another example, a patient may omit most details and features from the left side of his drawings (see Figure 1); this neurobehavioral syndrome, known as left hemi-inattention, typically occurs following brain damage to posterior regions of the right hemisphere.

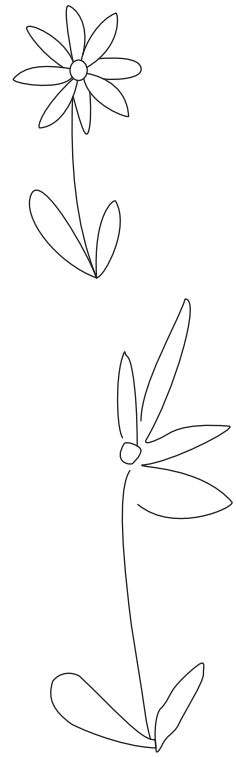

Figure 1 An individual who suffered a cerebral vascular accident in the right parietal lobe was asked to copy a model drawing of a daisy (top). The patient's drawing (bottom) illustrates left hemi-inattention.

Emotional, Behavioral, and Personality Testing

The neuropsychological evaluation typically goes beyond the assessment of cognitive and motor skills, and also includes psychometric testing of emotional, behavioral, and personality functioning. Many neuropsychologists administer tests such as the Minnesota Mutiphasic Personality Inventory-2 (MMPI-2) or the Beck Depression Inventory as part of their standard evaluation, given that the patient has the cognitive capacity to take them. These tests assist the neuropsychologist in determining whether a patient has a psychiatric disturbance, which, in turn, may be affecting his or her performance on the neuropsychological tests. Most practicing neuropsychologists obtain their doctoral degrees in clinical psychology before receiving specialized training in neuropsychology, and this training enables them to conduct general psychodiagnostic assessments as part of their neuropsychological evaluations.

Final Interpretations and Recommendations

The heart of the neuropsychological evaluation, and the most difficult part, is drawing final conclusions

about the presence, nature, extent, and causes of a patient's cognitive, motor, and behavioral difficulties and, if necessary, formulating recommendations for helping the patient. Just because a patient obtains test scores that fall in the impaired range does not mean that he or she has brain damage.

The first step in the interpretation process is to determine whether a patient's test scores are valid (using techniques discussed earlier) and internally consistent (e.g., if a patient has a word-finding deficit, then he or she should exhibit difficulty on any task that places extensive demands on precise naming ability). If the test scores appear valid and consistent, the neuropsychologist then attempts to characterize the patient's profile of cognitive–motor strengths and weaknesses. Based on an estimate of the patient's premorbid level of cognitive functioning, the neuropsychologist opines whether the patient's weaknesses represent acquired deficits. Next, the neuropsychologist, drawing upon his or her knowledge of brain–behavior relationships, tries to decide whether the patient's profile of cognitive–motor strengths and weaknesses is typical of particular types and locations of brain dysfunction. The neuropsychologist then must weigh all of the patient's risk factors for acquired brain damage, such as a recent head trauma or stroke, history of medical illnesses that may compromise brain functioning, history of serious psychiatric disorder, past alcohol or drug abuse, or a family history of a genetically based neurological disorder (e.g., Alzheimer's disease). In considering these different risk factors, the neuropsychologist tries to determine whether the patient's cognitive–motor profile is more likely to be associated with one risk factor than another, or with multiple risk factors.

Before a final conclusion can be reached about the presence of neurocognitive dysfunction, the neuropsychologist must rule out all possible explanations for the patient's low test scores, going beyond brain damage *per se*, such as educational and cultural factors, current medications, current emotional and physical state, and level of motivation to perform well on the tests. In addition, the neuropsychologist must conduct a psychodiagnostic evaluation to determine whether the patient has a psychiatric disorder and, if so, whether this disorder is contributing to the patient's cognitive difficulties. After integrating, analyzing, and synthesizing this extensive and varied body of information, the neuropsychologist is then in a position to proffer his final impressions about the presence or absence of brain dysfunction and its sequelae. If the neuropsychological data strongly implicate a neurological disorder, such as Alzheimer's disease, then it falls within the neuropsychologist's purview to make those diagnoses.

At the end of the report, the neuropsychologist will typically offer recommendations for the patient. Common areas of recommendations include the following approaches:

- Additional neurological, medical, or psychiatric examinations, consultations, or procedures (e.g., MRI brain scan; psychotropic medication consultation).
- Speech therapy or cognitive rehabilitation.
- Individual or family psychotherapy or other behavioral interventions.
- Educational interventions and programs (especially for children and young adults).
- Vocational testing and counseling.
- Social work consultation to address the patient's living situation and the need for various levels of assistance and supervision with activities of daily living.
- Repeat neuropsychological evaluation in the future to determine the status of the patient's cognitive functioning over time.

When feasible, the neuropsychologist often will meet with the patient and his or her family to discuss the findings of the evaluation and the need for the various recommended interventions.

Repeat Neuropsychological Testing

The readministration of the same or similar neuropsychological tests to a patient at a later time can serve several clinically useful purposes. As an example, for patients suspected of having a progressive dementia such as Alzheimer's disease, the first neuropsychological testing provides baseline data. The repeat testing, which usually is done 9–12 months later, affords an objective measure of whether the patient exhibits a decline in cognitive functions typical of a progressive disease. Some neuropsychological tests may show practice effects on repeat testing, which is an improvement in performance from having taken the same or similar tasks at an earlier time. However, the inexorable progression of a disorder such as Alzheimer's disease typically overrides any improvement in scores related to practice effects.

Repeat testing also is helpful for charting the recovery of neurocognitive functions over time following a brain insult, such as a stroke or brain injury. Recovery of brain functions tends to occur relatively slowly, with a 1-year time period often used as the benchmark for the vast majority of the recovery to have occurred. By comparing a patient's test scores obtained a few months after an injury with those obtained a year or more postinjury, the neuropsychologist, after factoring in practice effects, can provide objective data about the

patient's recovery process. Other important reasons for conducting repeat testing include monitoring improvement in cognitive functions following various interventions, such as cognitive rehabilitation and pharmacological treatment, and determining the validity of an individual's test scores by examining the consistency of his or her profile of cognitive and motor strengths and weaknesses over time.

Common Referral Questions for Adults

Adult patients typically are referred for a neuropsychological evaluation by neurologists, neurosurgeons, clinical psychologists, psychiatrists, primary care physicians, speech pathologists, other health professionals, and attorneys. The following examples of consult questions illustrate some of the reasons why adult patients undergo neuropsychological testing:

- To assist in evaluating whether a patient with increasing cognitive difficulties, as reported by the patient, family members, or health professionals, has an insidious neurological disorder. Examples of these disorders include multiple sclerosis, brain tumor, idiopathic hydrocephalus, white-matter ischemic changes, and Alzheimer's disease.
- To evaluate whether a patient who has had a risk factor for brain damage does, in fact, exhibit cognitive–motor deficits or emotional–behavioral changes consistent with brain dysfunction. Examples of these risk factors include head trauma, chronic alcohol or drug abuse, certain medical illnesses (e.g., hypertension, diabetes, kidney failure, HIV infection), certain psychiatric disorders (e.g., schizophrenia), and exposure to neurotoxic agents.
- To assess the presence, nature, and extent of cognitive–motor deficits and emotional–behavioral changes in patients with known brain pathology. These patients are referred after it has been determined, usually from the neurological examination and neuroimaging, that they have had, for instance, a cerebral vascular accident or have a brain tumor, Parkinson's disease, probable Alzheimer's disease, or MRI- or CT scan-confirmed brain damage following a head trauma. The question for these patients is not whether they have brain damage, but rather the presence, nature, and extent of cognitive, motor, and behavioral sequelae of their brain dysfunction.
- To assist neurosurgeons in several ways, such as in evaluating whether a patient with intractable seizures and in need of resection of the epileptogenic focus has language skills localized in this general brain region (this testing usually involves a Wada procedure, where one cerebral hemisphere at a time is anesthetized during cognitive screening); in conducting pre- and postsurgery evaluations to determine whether a neurosurgical procedure (e.g., insertion of a shunt tube to relieve hydrocephalus; resection of a brain tumor) alters cognitive–motor functions for better or worse; and in reevaluating cognitive functions over a period of several years in a hydrocephalus patient who has received a shunt in order determine if the shunt has continued to function normally.
- To evaluate the cognitive functions of individuals with more serious psychiatric disorders that often have a neurological component (e.g., schizophrenia).
- To provide recommendations regarding the patient's capacity to return to work or to receive vocational retraining in a new occupation that would be best suited for the patient, given his or her profile of cognitive strengths and weaknesses.
- To offer recommendations for cognitive rehabilitation (e.g., teaching the patient memory compensatory strategies).
- To assist in determining whether a patient with brain dysfunction has the cognitive, motor, and behavioral capacity to perform important activities of daily living (e.g., ability to drive a car, handle finances, prepare meals, live independently, comply with medication regimens, care for self).
- To make recommendations regarding the best living situation for the patient in light of his or her cognitive, motor, and behavioral deficits (e.g., independent, assisted-living, or nursing-home placement).

Common Referral Questions for Children

Neuropsychological evaluations are generally more difficult to conduct for children than for adults. Just as children differ in their physical growth rates, they also differ in the development of their cognitive and motor functions. It is sometimes difficult to determine whether a 'cognitive deficit' in a child represents a permanent impairment or simply a lag or delay in the development of that ability. In addition, the classic brain–behavior relationships that have been documented in adults do not always apply to children, making it difficult to localize cognitive deficits to particular brain regions. For example, a lesion in the left superior temporal gyrus can result in a Wernicke's (fluent) aphasia in an adult and in a nonfluent aphasia in a child. Finally, some children present with pronounced attention problems, hyperactivity, or other behavioral disturbances that make it difficult for the child to focus on the psychometric tests and provide valid test responses.

Despite the challenges inherent in child neuropsychology, this area of practice has become one of the fastest growing disciplines within the field. Referrals for child neuropsychological evaluations usually are made by pediatricians, pediatric neurologists, clinical

psychologists, school psychologists, speech therapists, educators, and sometimes parents. Many of the referral questions for children are similar to those for adults, such as assessing whether the apparent onset of new cognitive difficulties reflects an insidious neurological disorder (e.g., brain tumor); evaluating whether a known risk factor for brain damage (e.g., birth complications, head trauma, lead exposure) has, in fact, resulted in cognitive, motor, or behavioral dysfunction; and assessing the presence, nature, and extent of cognitive, motor, and behavioral changes in patients with known brain pathology (e.g., childhood seizure disorder). Following are examples of other specific reasons why children are referred for neuropsychological testing:

- To determine whether a child who is struggling academically in school has a learning disability. Neuropsychological research has documented different types of learning disabilities (e.g., verbal vs. nonverbal) that may reflect neurodevelopmental abnormalities in distinct brain regions.
- To assess whether a child whose mother abused alcohol or drugs or consumed other teratogenic agents during her pregnancy (e.g., certain medications) exhibits cognitive, motor, or behavioral deficits suggestive of prenatal brain damage (e.g., FAS).
- To assist educators in developing the best educational program for a child based on the child's profile of cognitive and behavioral strengths and weaknesses.
- To assess recovery of function in a child who has suffered a documented traumatic brain injury and to provide recommendations for cognitive remediation and educational assistance.

Conclusions

Clinical neuropsychology has become one of the fastest growing disciplines in psychology. Neuropsychological testing not only provides an empirically rigorous method for assisting in the diagnosis of neurological, medical, and psychiatric patients, it also enhances our scientific understanding of one of the great mysteries of life, namely, the generation of mental processes from brain structures.

See also: Aging of the Brain and Alzheimer's Disease; Cognition: An Overview of Neuroimaging Techniques; Cognitive Neuroscience: An Overview; Human Methods: Psychophysics; Magnetic Resonance Spectroscopy; Neuroimaging; Neuropsychology: Theoretical Basis; Psychophysics of Attention.

Further Reading

Benton AL (1980) Psychological testing for brain damage. In: Kaplan HI, Freedman AM, and Badock BJ (eds.) *Comprehensive Textbook of Psychiatry,* vol. 1, ch. 12. Baltimore, MD: Williams & Wilkins.

Delis DC, Kaplan E, and Kramer JH (2001) *The Delis-Kaplan Executive Function System.* San Antonio, TX: The Psychological Corporation.

Delis DC, Kramer JH, Kaplan E, et al. (2000) *The California Verbal Learning Test,* 2nd edn. San Antonio, TX: The Psychological Corporation.

Goodglass H and Kaplan E (1983) *The Assessment of Aphasia and Related Disorders.* Philadelphia, PA: Lea & Febiger.

Grant I and Adams K (eds.) (1996) *Neuropsychological Assessment of Neuropsychiatric Disorders,* 2nd edn. New York: Oxford University Press.

Heaton RK, Grant I, and Matthew CG (1991) *Comprehensive Norms for an Expanded Halstead–Reitan Battery.* Odessa, FL: Psychological Assessment Resources.

Heilman K and Valenstein E (eds.) (1993) *Clinical Neuropsychology,* 3rd edn. New York: Oxford University Press.

Kaplan E, Fein D, Morris R, et al. (1991) *WAIS-R as a Neuropsychological Instrument.* New York: Psychological Corporation.

Larrabee GL (ed.) (2005) *Forensic Neuropsychology.* New York: Oxford University Press.

Lezak MD, Howieson DB, Loring DW, et al. (2005) *Neuropsychological Assessment,* 4th edn. New York: Oxford University Press.

Luria AR (1980) *Higher Cortical Functions in Man,* 2nd edn. New York: Basic Books.

Massman PJ, Delis DC, Butters N, et al. (1992) The subcortical dysfunction hypothesis of memory deficits in depression: Neuropsychological validation in a subgroup of patients. *Journal of Clinical and Experimental Neuropsychology* 14: 687–706.

Mattson SN, Riley EP, Gramling L, et al. (1997) Heavy prenatal alcohol exposure leads to IQ deficits with or without all of the features of fetal alcohol syndrome. *Journal of Pediatrics* 131: 718–721.

Spreen O and Strauss E (1998) *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary,* 2nd edn. New York: Oxford University Press.

Squire L and Butters N (eds.) (1992) *Neuropsychology of Memory,* 2nd edn. New York: Guilford Press.

Neuropsychology of Primate Reward Processes

E A Murray, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

What Is Reward and Where Is It Processed?

Reward is operationally defined as anything an animal will work to obtain or will systematically approach. This article uses the term reward interchangeably with positive reinforcement, notwithstanding their somewhat different origins and meanings.

A key problem in understanding reward processing in primates (or in any other group of animals) is that reward is a surprisingly complex and nuanced concept. Rewards have both hedonic and incentive value, which correspond to the concepts of liking and either wanting or seeking, respectively. In addition, rewards may vary in terms of probability, timing, quantity, quality, context, and effort required to obtain them. Context alone has many facets. A reward could be available all the time (i.e., it could be context independent). More often, however, reward is available conditionally. For example, a reward might usually be available but not as much recently (i.e., it could be dependent on temporal context); it might be available in one location but not another (i.e., it could be dependent on spatial context); and, in human primates, at least, it could be both wanted and not wanted to be wanted (i.e., it could be dependent on first-order and second-order intentionality). The complexity of reward does not end there. Rewards have both a general aspect and have a wide diversity of specific aspects (e.g., olfactory, gustatory, tactile, visceral, and visual features of multiple dimensions). Naturally occurring rewards include food, in general, and specific kinds of food, in particular. They also include water, salt, warmth (and sometimes cold), and sex. In addition, initially neutral stimuli or places associated with these types of reward become reinforcing under certain circumstances.

Many parts of the brain allow animals to learn about rewards. Indeed, we might argue that every part of the brain plays a role in reward processing; each part is either directly or indirectly involved in acquiring information about the stimuli, places, events, and actions associated with reward. Dopaminergic neurons, which detect rewards and deviations from predicted rewards to promote learning, play an especially crucial role in reward processing in primates, as they also do in other animals. This article explores, however, a more narrow topic. It deals with those neural structures known to play both an essential and a general role in processing rewards in primates, namely the amygdala, orbital and medial prefrontal cortices, and basal ganglia.

What Is Special about Reward Processing in Primates?

In primates, the advent of the granular prefrontal cortex, which interacts with the amygdala, the basal ganglia, and other structures involved in reward processing, makes the primate way of dealing with rewards different than that of other animals, including other mammals. The nature of this difference is not yet well understood. Nevertheless, I assume in this article that the granular prefrontal cortex underlies a knowledge of the physical world, the self, and social groups that is unique to primates. This view of primate neuropsychology does not disparage other animals. Many other clades have evolved unique ways of knowing about the world and thereby producing advantageous outcomes. But we are primates and so have an intensified interest in our way of doing things.

Several areas that are called prefrontal cortex in rodents and other nonprimate mammals, such as the orbital and medial prefrontal areas, have clear homologs in primate brains. The common mammalian areas include the infralimbic and prelimbic cortices, the anterior cingulate cortex, and the agranular insular cortex, among others. However, according to the comparative neuroanatomical evidence, rodents lack a homolog of the granular prefrontal areas of primates. These areas include the dorsolateral prefrontal cortex (approximately area 46), the dorsal prefrontal cortex (area 9), the ventral (or ventrolateral) prefrontal cortex (area 12 or 47/12, and areas 44 and 45), the frontal pole cortex (area 10), and the rostral portion of the orbital prefrontal cortex (parts of areas 11 and 13). In addition, the most rostral parts of other areas, such as the prelimbic, infralimbic and cingulate cortices, might also be unique to primates, depending on how they are defined.

Collectively, neuroanatomists often refer to these uniquely primate areas as the granular prefrontal cortex, despite the fact that they do not have a granular cytoarchitecture. As applied to the remainder of the neocortex, the term granular refers to the hypertrophied layer 4 found in primary sensory areas. Instead, the so-called granular prefrontal areas have a homotypical or dysgranular architecture, a kind of cortex absent in the frontal cortex of other mammals.

Clearly, if these granular prefrontal areas appeared during the course of primate evolution, the interaction of cortex as a whole with the other brain areas involved in reward processing must have changed accordingly. This idea applies to both frontal cortex–amygdala interactions and those between the frontal cortex and basal ganglia. Accordingly, the next two sections take up the role in reward processing of the amygdala and amygdala–frontal interaction, respectively. Subsequently, the role of the orbital and medial prefrontal cortices in reward processing is addressed; the final section discusses the basal ganglia.

Amygdala

Reward versus Punishment

A role for the amygdala in negative reinforcement – punishment, as opposed to reward – has dominated work on the amygdala. The amygdala plays a central role in conditioned fear, including fear-potentiated startle, by linking initially neutral stimuli to the innately programmed triggers of defensive responses such as freezing, reflex potentiation, and many aspects of autonomic arousal. This fact has led to the mistaken idea that the amygdala serves primarily to process negative reinforcement, as opposed to rewards. Following this line of thought, some experts have referred to the primate amygdala as a protection device.

Recent neuropsychological studies demonstrated, however, that the amygdala's role in reward processes is as important as its role in punishment and fear. For instance, when sensory inputs occur in close temporal proximity to reward, the amygdala appears to link those sensory inputs with the neural circuits mediating instinctive behaviors, such as autonomic reflexes and orienting responses. This kind of learning promotes the processing of stimuli that predict biologically significant events. In other instances, the amygdala acts to enhance the processing of unexpected (surprising) sensory events. And in yet other instances, the amygdala acts to assign positive value to neural representations of sensory inputs. Thus, the amygdala plays a key role in connecting reward-linked sensory information to the instinctive processes that underlie some of the most fundamental aspects of vertebrate behavior.

The basolateral group of nuclei (including the lateral nucleus, the basal nucleus, and the accessory basal nucleus) and the central nucleus of the amygdala contribute differently and in parallel. The basolateral amygdala links neutral stimuli with the specifics of reinforcement (e.g., the tastes, smells, sights and visceral sensory signals associated with rewards), whereas the central nucleus links stimuli to more general aspects of reward, such as its affective significance.

Neuropsychology in Nonhuman Primates

Unlike correlative techniques, such as those used in neuroimaging and neurophysiology, neuropsychological methods enable a rigorous assessment of whether a particular structure makes an essential contribution to a given cognitive capacity. There are several ways to measure reward-processing capabilities in nonhuman primates, and two commonly used approaches have contributed to understanding the amygdala's function in reward processes: In one, monkeys learn to link objects with food value; in the other, monkeys learn to link the presentation of a picture with food.

The first approach is carried out in two stages. First, monkeys are familiarized with a large number of objects, half of which are associated with one kind of food, designated Food 1, and half of which are associated with a different food, designated Food 2. Second, the value of one food is decreased by feeding it to the monkey, and the effect of this selective satiety on choices between Food 1- and Food 2-associated objects is compared with a baseline condition in which there was no prefeeding.

Intact monkeys avoid choosing objects associated with a devalued food, as indicated by the high devaluation score for the unoperated control group (Figure 1, (**1**)). In these experiments, sated monkeys shift their choices away from objects associated with the devalued food by about half, yielding a score of ~17 out of a possible 30. In contrast, monkeys with selective, bilateral amygdala lesions continue to choose much as they did before any food had been devalued. This deficit is reflected in the low devaluation score after bilateral amygdala removal (Figure 1, (**4**)), in which the monkeys make, on average, only approximately four fewer choices of objects associated with a devalued food. Control procedures have shown that changes in visual perceptual abilities, food preferences, or satiety mechanisms cannot account for the impairment. Thus, monkeys with amygdala lesions cannot efficiently link objects with the current value of a food reward.

Precisely how objects are linked to food value and how this information translates to shifts in response selection remain largely unknown. The test outlined here comprises several components: forming object representations, linking those representations with the incentive value of the associated food, registering and encoding a change in the reward value due to selective satiation, linking object representations with those updated values, and using these changed representations to choose between objects.

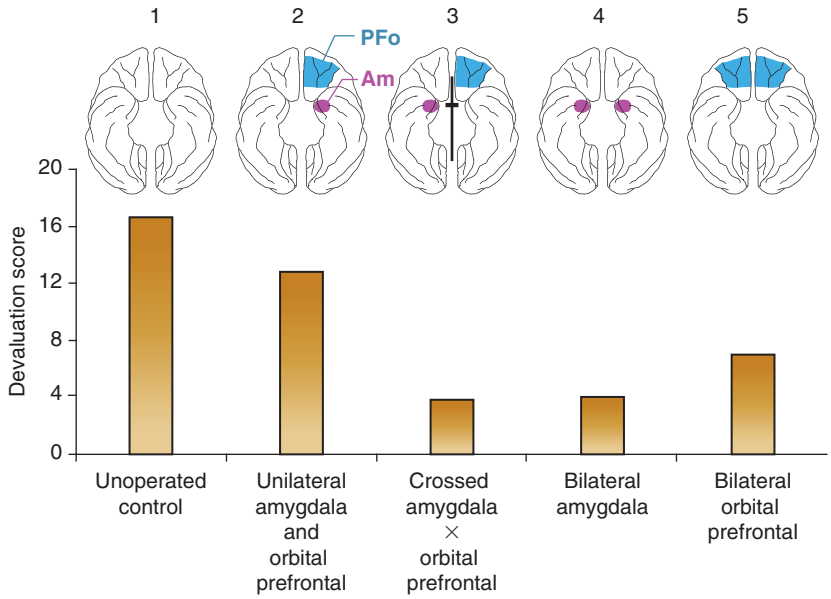

Figure 1 Contribution of the amygdala (Am) and orbital prefrontal cortex (PFo) to object–food value associations. The top row shows ventral views of the macaque brain showing the lesion for the five groups (rostral is up). The black line and bar for the crossed lesion group (3) represents the transection of the corpus callosum and anterior commissure. All operated monkeys underwent two stages of surgery, involving lesion of either the Am (pink shading) or PFo (blue shading). The graph shows the devaluation scores. Higher devaluation scores indicate a greater ability to avoid choosing objects after decreases in the value of their associated food reward. The highest possible score is 30 choices, which would indicate complete avoidance of objects associated with devalued food. When tested after the second-stage surgery, each of the operated groups was significantly impaired relative to controls.

Transient inactivation of the amygdala by focal infusion of the GABA agonist muscimol has clarified the mechanisms of stimulus valuation in primates. In these experiments, monkeys received infusions of either saline or muscimol, bilaterally, via cannulae lowered to the basolateral amygdala, either before or after satiation. As expected, saline infusions had no effect. Nearly identical results were obtained when the basolateral amygdala was inactivated after the selective satiation procedure. By contrast, inactivation of the basolateral amygdala immediately before and during the satiation procedure prevented the shift in object choices that normally occurs with selective satiation. Thus, the basolateral amygdala needs to be functionally intact for registration of a change in the incentive value of the food reward and for the subsequent adjustment in the monkey's choices. Apparently, once the value of the food reward has been updated, no further contribution of the basolateral amygdala is required for these functions.

The contribution of the primate amygdala to processing rewards has also been shown in another way. This second approach relies on the fact that animals will work to obtain the presentation of a picture that has been linked with food, called a conditioned reinforcer. (In general terms, any stimulus that has been linked to reward, and for which animals will work, is called a secondary, or conditioned, reinforcer.) Once a conditioned reinforcer has been established, intact monkeys will put forth a great deal of effort to earn the presentation of the conditioned reinforcer, even if the food reward is available only infrequently. Marmoset monkeys with bilateral amygdala damage fail to put forth this effort. In addition, they fail to show the autonomic responses, such as increases in heart rate and blood pressure, that usually accompany the sight of reinforcers. Presumably the amygdala damage disrupts the same process as previously described for the selective satiation test – linking stimuli with current value – in addition to disrupting the activation of autonomic centers. However, conditioned reinforcers probably act through at least one additional mechanism independently of incentive value, as evidenced by the ability of a conditioned reinforcer to support new learning even after the devaluation of the food with which it was originally paired.

The neuropsychological findings presented here show that the primate amygdala performs a key aspect of reward processing; it is necessary for linking initially neutral stimulus representations with current, updated biological values. Furthermore, different parts of the amygdala contribute to reward processes in different ways. The basolateral amygdala plays a crucial role in updating the representations of value for specific aspects of each reinforcer or reinforcer type (e.g., visual features, taste, and smell),

whereas the central nucleus of the amygdala performs a comparable function for general aspects of reinforcement (e.g., food reward, in general, rather than a particular kind of food).

What Is Different about Primates?

What about this is specific to primates? The answer is that the amygdala is likely to provide the same function – linking neutral representations with updated value – for the highest aspects of cognition.

The amygdala has a long evolutionary history, stretching back to the earliest land animals and perhaps beyond, to the earliest vertebrates. It has relatively direct connections with the hypothalamus, as well as reciprocal anatomical relations with high-order cortical areas such as those that compose the granular prefrontal cortex. Amygdala relations with the granular orbital and medial prefrontal regions may hold the key to the amygdala's contribution to the primate way of processing rewards; in addition to assigning value to objects, the amygdala may also assign value to the kinds of information stored in and processed by the granular prefrontal cortex.

In nonhuman primates, this information includes rules, strategies, categories, and aspects of social knowledge, among other cognitive functions, processed via the extensive and reciprocal connections of the amygdala with the frontal cortex. In humans, likewise, the amygdala assigns values that guide social interactions, and this capacity is also likely to involve interactions with the frontal cortex. Amygdala damage reveals deficits in recognizing facial expressions of fear and in judging from facial expressions whether others are trustworthy. Indeed, the amygdala is thought to be important for acquiring social (moral) conventions, such as learning to avoid actions that will harm others. Other evidence mirrors the findings already described regarding valuation in nonhuman primates. Studies in humans have shown that the amygdala plays a role in assigning positive values to pictures associated with food reward. Significantly, the value assignments mediated by the amygdala may be inaccessible to conscious awareness, at least in any direct way. For example, in experimental settings, when people view pictures and at the same time receive reward, they acquire picture preferences, although they remain unaware of the relationship between the picture presentation and the probability of reward. Patients with amygdala damage, however, do not display such preferences; that is, they do not acquire the unconscious bias, or preference, for particular objects.

The findings described indicate that the amygdala is necessary for linking updated biological value with the representations of objects. There is evidence that this is the case in all mammals and probably in other animals, as well. In nonhuman primates, however, this function probably extends to other representations, such as those for members of a social group, abstract behavior-guiding rules and problem-solving strategies, and knowledge about how to achieve goals during long and highly variable event sequences. Primates may not be unique in all of these capabilities, but we are uniquely good at all of them. In humans, this function probably extends even further, to representations of ideas, abstract concepts and beliefs, and aesthetic values, including self-worth.

Prefrontal–Amygdala Interactions

Although interactions between the amygdala and the primate prefrontal cortex lie at the heart of attaching value to abstractions such as ideas, rules, strategies, and concepts, little is known about the mechanisms of amygdala–frontal cortex interactions in primates. Fortunately, research on rodents may shed light on how value-to-representation linkages are forged in mammals generally, including primates.

Neuronal activity in the orbital frontal cortex of rodents, in areas that are probably homologous with the more posterior parts of orbital frontal cortex in primates, reflects the value of expected outcomes. Significantly, this activity is evident during the presentation of a stimulus that predicts the reward before the reward is given, and the emergence of this activity depends on the integrity of the amygdala. Consequently, stimulus–value associations mediated by the basolateral amygdala appear to be established in the orbital frontal cortex of rodents, and these representations are thought to serve as the basis of reward expectancy. Extending this finding to the granular prefrontal cortex of primates, its more abstract representations of the world are probably linked to current value through similar mechanisms.

The findings from the selective satiation procedures described earlier accord with this general view. According to one model, the sight of an object elicits activity in the orbital prefrontal cortex that reflects the current value of the expected outcome and may therefore underlie the ability of monkeys to choose objects appropriately in the face of changes in the value of the goal. Strong support for this model comes from the effects of crossed surgical disconnection of the amygdala from the orbital prefrontal cortex, which disrupts adaptive response selection after selective satiation (Figure 1, (3)).

Orbital and Medial Prefrontal Cortex

The selective satiation and conditioned reinforcer tests of reward processing have also been applied to the study of orbital prefrontal cortex in primates. This research has revealed that the orbital prefrontal cortex, like the amygdala, is essential for linking objects with current food value (Figure 1, (5)). In addition, this region of the cortex is essential for learning a new response when that learning is supported only by presentation of a conditioned reinforcer. Thus, damage to the orbital prefrontal cortex produces some of the same effects on reward processing as does damage to the amygdala, although these regions probably make selective contributions to reward processing as well.

How do the amygdala and orbital prefrontal contributions to reward processing differ? Some evidence suggests that, whereas the amygdala is primarily important for linking sensory inputs to value, the orbital and medial prefrontal cortical regions are preferentially involved in using that information to guide choices and actions. It has been suggested that the orbital prefrontal cortex, in particular, provides a common currency for the valuation of competing goals, a function critical to adaptive response selection.

The medial and orbital prefrontal cortex both function in reward processes – both have been identified as regions linking motivation, cognition, and action – but they contribute differently. For example, damage to the medial prefrontal cortex, unlike damage to the orbital cortex, spares the acquisition of a new response for conditioned reinforcement. One possibility is that the orbital prefrontal cortex links objects with current reward value to guide choices among objects, whereas the medial prefrontal cortex links actions with current reward value to guide choices among actions. Recent evidence supports that idea for the medial prefrontal areas. Removal of part of the anterior cingulate cortex disrupts monkeys' abilities to link particular actions with reward (see Figure 2) and to choose optimally under conditions in which different actions take on different values, based on the probability of reward delivery. Medial prefrontal cortex damage also disrupts monkeys' valuations of social stimuli. Finally, the medial prefrontal cortex is anatomically linked to visceromotor structures and may therefore be important for adaptive autonomic and affective responses.

Basal Ganglia

If the amygdala, through its interactions with the uniquely primate granular prefrontal cortex, provides the linkage between initially neutral representations and their updated biological value, what then of the basal ganglia's role? The basal ganglia receives a signal from dopaminergic neurons in the midbrain that convey information about reward. At a minimum, dopaminergic neurons signal the occurrence of unexpected rewards and deviations from predicted rewards. Like amygdala–frontal cortex interactions, basal ganglia–frontal cortex interactions are affected by the existence of the uniquely primate parts of the frontal lobe – the granular prefrontal cortex.

In rodents, the dorsal basal ganglia is essential for at least two different types of learning, both involving reward processing: stimulus–response learning and action–outcome learning. In stimulus–response learning, the ingestion of food or fluid reward in the past has acted to enhance the association between a particular stimulus and a particular response. The reward itself does not enter into the stimulus–response association, at least not at the time of the behavior; given the appearance of the stimulus, the animal performs the associated response. Because the animal performs the response without respect to the value of the predicted outcome, stimulus–response behavior is often called a habit. Action–outcome associations, unlike stimulus–response associations, are sensitive to the value of the outcome. For that reason, this latter kind of learning is thought to play a major role in goal-directed behavior. Although some researchers maintain that the basal ganglia functions only in stimulus–response learning, recent work has overturned this idea. Instead, according to research in rodents, different portions of the dorsal basal ganglia are essential for stimulus–response and action–outcome learning.

Although neuropsychological data from non-human primates remain scarce, damage to the dorsal basal ganglia (but not damage to the amygdala or orbital prefrontal cortex) produces deficits in visual discrimination learning. In these tasks, monkeys must learn and remember which of two objects to choose in order to produce a reward. Monkeys with damage to the portions of the basal ganglia that receive highly processed sensory input from visual cortex (such as the head and tail of the caudate nucleus and the posterior ventral putamen) solve visual discrimination problems more slowly than do intact monkeys. Because other already-learned visual tasks can be performed efficiently, this impairment does not depend on alterations in visual perceptual processing. Instead, although discrimination deficits are notoriously difficult to interpret, it seems likely that problems with stimulus–response and action–outcome learning account for the discrimination deficits that follow lesions of the dorsal basal ganglia.

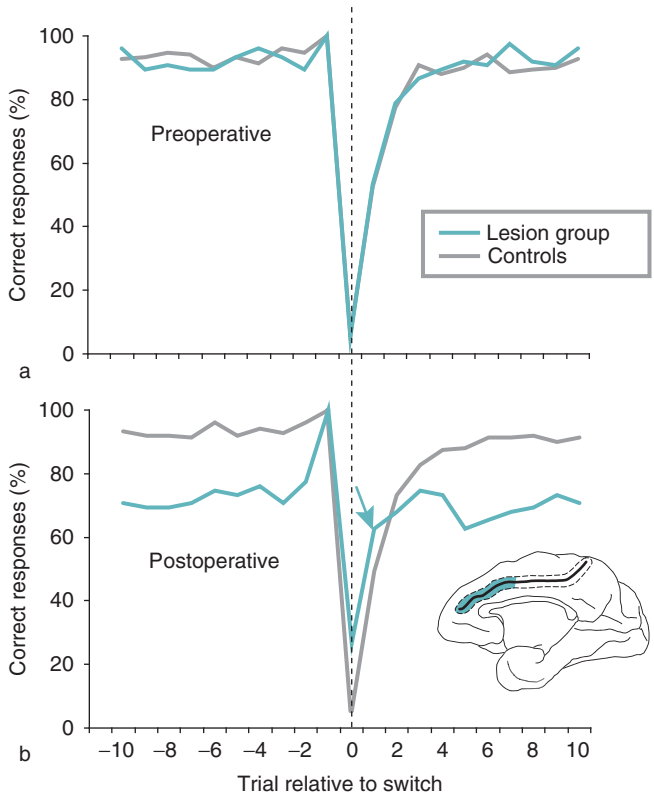

Figure 2 Contribution of the anterior cingulate cortex in monkeys to action–reward associations: (a) presurgery performance; (b) postsurgery performance of controls and lesion group. Monkeys were trained to perform two different actions on a joystick and to switch between them as needed to obtain a food reward. Only one action was rewarded at a time. After monkeys had obtained 25 rewards for performing one action, only the other action was rewarded for a series of trials. Curves show the mean percentage of correct responses for trials before and after a switch trial (trial 0, dashed vertical line), averaged over several such switches. Although both groups performed comparably before surgery (a), monkeys with lesions of the anterior cingulate cortex were impaired after surgery (b). Note that the lesion group detected errors, as indicated by the rapid change in the action performed immediately after the switch (blue arrow). Thus, monkeys with anterior cingulate cortex lesions had a specific difficulty linking actions to reward. Also in (b) is a medial view of the macaque brain (rostral to the left, dorsal up) showing the location of lesion (blue shading), with the dashed lines indicating the banks of the cingulate sulcus. Adapted by permission from Macmillan Publishers Ltd: *Nature Neuroscience*, Kennerley SW, Walton ME, Behrens TEJ, Buckley MJ, and Rushworth MES (2006) Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience* 9: 940–947. Copyright 2006.

In addition to the dorsal basal ganglia, parts of the ventral basal ganglia play a crucial role in reward processing. The ventral basal ganglia, in concert with the amygdala, link stimuli with reward outcome, a type of learning termed stimulus–outcome learning. One study in monkeys showed that bilateral damage to the nucleus accumbens, a part of the ventral striatum, caused impairments during a food hoarding task in which subjects had to remove peanuts from their shells and store them for future consumption. The same monkeys efficiently hoarded peanuts without shells, suggesting that their difficulty might lie in recognizing the incentive value of the shell-covered peanut. A deficit in stimulus–outcome learning might account for this finding.

A theoretical account of basal ganglia function posits that the basal ganglia recognizes contexts for action. According to this view, the cortex and thalamus provide the basal ganglia with a vast array of inputs, some of which predict biologically significant events. When a particular pattern of inputs produces reward delivery, the basal ganglia encodes that context, retaining the relevant inputs and discarding the irrelevant ones. In this way, the basal ganglia could extract the dimensions of input that have the most relevance for either stimulus–outcome or action–outcome learning and for various combinations of these kinds of learning.

Another role for the basal ganglia may be in acquiring performance rules involving reward. In some neuropsychological tests, a reward can be used much like any other sensory signal to trigger a performance rule. According to this idea, the occurrence (or nonoccurrence) of food, for example, guides the

selection of a rule, and the role of food depends on its informational value, as opposed to its reward or reinforcement value. Support for this idea is provided by the finding that intact monkeys can learn to reject their most recent reward history as well and as fast as they can learn to follow it. For example, in one test, monkeys learned a performance rule of the win-stay/lose-shift type, in which they must choose a previously rewarded object to gain an additional reward. Conversely, other monkeys learned the opposite performance rule, one of the win-shift/lose-stay type, in which they must reject a previously rewarded object and choose an unrewarded one instead. Despite the fact that one group of monkeys learned to reject its most recent reward history, both groups learned at the same rate. Some tasks commonly used to assess reward learning in primates, such as object reversal learning, may also depend on this rule-based type of learning, at least in experienced monkeys. Evidence for a role of basal ganglia in this informational (value-independent) aspect of reward processes comes from the finding that damage to the basal ganglia severely disrupts object reversal learning. So does damage to the orbital prefrontal cortex and to areas such as perirhinal cortex and entorhinal cortex, which are anatomically connected to it; however, damage to the amygdala has no effect. These deficits thus involve frontal cortex–basal ganglia interactions more than frontal cortex–amygdala ones.

Summary

Through its role in linking representations in the granular prefrontal cortex with updated valuations, the primate amygdala provides the neural basis not only for choosing among objects based on their current biological value but also for esthetics and the valuation of abstract goals, knowledge, and ideas; for attaching value to a particular course of action; and for social and self-referential valuations. Some evidence suggests that the amygdala is primarily involved in linking sensory inputs to value, whereas the orbital and medial prefrontal cortical regions use that information to guide choices and actions. Like amygdala–frontal interactions, aspects of basal ganglia–frontal interactions in primates reflect the contributions of the granular prefrontal cortex. The cortex and basal ganglia together subserve the kinds of reward-mediated behaviors that depend on performance rules, as well as on stimulus–response and action–outcome associations and, in conjunction with the amygdala and frontal cortex, stimulus–outcome associations.

See also: Connectivity of Primate Reward Centers; Goal-Directed Behavior Theories; Prefrontal Contributions to Reward Encoding; Representation of Reward; Reward and Learning; Reward Neurophysiology and Primate Cerebral Cortex; Social Interaction Effects on Reward and Cognitive Abilities in Monkeys.

Further Reading

Adolphs R (2003) Cognitive neuroscience of human social behavior. *Nature Reviews Neuroscience* 4: 165–178.

Balleine BW (2005) Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology & Behavior* 86: 717–730.

Balleine BW and Killcross S (2006) Parallel incentive processing: An integrated view of amygdala function. *Trends in Neurosciences* 29: 272–279.

Baxter MG and Murray EA (2002) The amygdala and reward. *Nature Reviews Neuroscience* 3: 563–573.

Braesicke K, Parkinson JA, Reekie Y, et al. (2005) Autonomic arousal in an appetitive context in primates, a behavioural and neural analysis. *European Journal of Neuroscience* 21: 1733–1740.

Kennerley SW, Walton ME, Behrens TEJ, Buckley MJ, and Rushworth MFS (2006) Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience* 9: 940–947.

Malkova L, Gaffan D, and Murray EA (1997) Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys. *Journal of Neuroscience* 17: 6011–6020.

Morris JS, Frith CD, Perrett DI, et al. (1996) A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* 383: 812–815.

Izquierdo A and Murray EA (2007) Selective bilateral amygdala lesions in rhesus monkeys fail to disrupt object reversal learning. *Journal of Neuroscience* 25: 1054–1062.

Parkinson JA, Crofts HS, McGuigan M, Tomic DL, Everitt BJ, and Roberts AC (2001) The role of the primate amygdala in conditioned reinforcement. *Journal of Neuroscience* 21: 7770–7780.

Preuss TM (1995) Do rats have prefrontal cortex? The Rose–Woolsey–Akert program reconsidered. *Journal of Cognitive Neuroscience* 7: 1–24.

Roberts AC (2006) Primate orbitofrontal cortex and adaptive behaviour. *Trends in Cognitive Sciences* 10: 83–90.

Rudebeck PH, Buckley MJ, Walton ME, and Rushworth MF (2006) A role for the macaque anterior cingulate gyrus in social valuation. *Science* 313: 1310–1312.

Rushworth MFS, Behrens TEJ, Rudebeck PH, and Walton ME (2007) Contrasting roles for cingulate and orbitofrontal cortex in decision and social behaviour. *Trends in Cognitive Sciences* 11: 167–176.

Stern CE and Passingham RE (1994) The nucleus accumbens in monkeys (*Macaca fascicularis*), I: The organization of behaviour. *Behavioural Brain Research* 61: 9–21.

Wellman LL, Gale K, and Malkova L (2005) $GABA_A$-mediated inhibition of basolateral amygdala blocks reward devaluation in macaques. *Journal of Neuroscience* 25: 4577–4586.

Yin HH and Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* 7: 464–476.

Neuropsychology: Theoretical Basis

G Berlucchi, Università di Verona, Verona, Italy

© 2009 Elsevier Ltd. All rights reserved.

Birth of Neuropsychology

Broadly conceived as an attempt to understand the relationships between brain, mind, and behavior, neuropsychology has a long history, but it was only in the 1960s that an autonomous field of scientific inquiry with that name was formally established. Although there is some evidence for occasional appearances of the term 'neuropsychology' and its derivatives in medical writings of the nineteenth century and perhaps even earlier, the term began to acquire some visibility in the 1940s and 1950s because the influential North American psychologists Lashley, Hebb, and Teuber used it in their efforts to advocate the interaction, if not the coalescence, of neurology and psychology. (Hebb gave the subtitle *A Neuropsychological Theory* to his celebrated book, *The Organization of Behavior*, and Lashley and Teuber mentioned neuropsychology as part of the classical field of physiological psychology.)

The official birth of neuropsychology as an independent scientific discipline can be dated to 1963, when an international specialty journal titled *Neuropsychologia* started its publication on the initiative of a small group of neurologists, psychologists, and psychiatrists partaking in an informal discussion forum called the International Neuropsychology Symposium. The scope and aims of neuropsychology were spelled out in the first issue of the journal in an unsigned manifesto attributable to the Editor-in-Chief, the French neurologist Henry Hécaen. According to the manifesto, the foundation of a new journal and a new distinctive discipline, neuropsychology, was intended to meet the needs of a number of investigators in a number of more traditional fields who had more in common with one another than with the members of their original specialties. More specifically, the term 'neuropsychology' was understood to mean a particular area of neurology of common concern to neurologists, psychiatrists, psychologists, and neurophysiologists, with an interest focused mainly, but not exclusively, on the cerebral cortex, and on disorders of language, perception, and action. It was recognized that although certain of these disorders could be studied only in adult humans, information of great value to the understanding of basic cerebral mechanisms and to human pathology was expected to come as well from experiments on animals, from ethological studies of learning and conditioning, and from analyses of normal development in children.

That the field was ripe for more than one specialized publication was very soon made clear by the appearance, in the following year, of another journal devoted to the study of the nervous system and behavior; this journal was entitled *Cortex*, and its frank neuropsychological character was attested by an opening article by Aleksandr Luria on the role of neuropsychology in the diagnosis of local brain damage. The same year, 1964, saw the publication of a reader, edited by R L Isaacson; this publication, entitled *Basic Readings in Neuropsychology*, collected 16 papers from scientific journals, 14 of which reported animal experiments, in confirmation of the perceived importance of this approach for the new field of neuropsychology.

The Anatomo-Clinical Approach as a Main Root of Neuropsychology

The concept of neuropsychology explicitly or implicitly set forth in the aforementioned publications extended the boundaries of the discipline well beyond those of the outgrowth of one of its main roots, the anatomo-clinical approach applied in the nineteenth and early twentieth centuries to the study of psychological dysfunctions in brain-injured patients. On the inspiration of phrenology, but with a much greater scientific respectability, the anatomo-clinical approach had worked on the belief that mental functions are localizable in the brain and that cerebral organization is based on an orderly system of functionally specialized centers linked by discrete and selective anatomical connections. Functional localization in the brain was assumed to be demonstrated when an autoptically ascertained lesion in a particular cerebral region could be shown to stand in a systematic relation with a particular psychological ability, insofar as that ability was lost, and others were preserved, in patients carrying that lesion. Based on that approach Broca, Wernicke, Liepmann, Lichtheim, Bianchi, Dejerine, and others had been able to localize some specific psychological functions, particularly in the domains of language and voluntary motor control, in specific regions of the cerebral cortex, strengthening the view that the brain is made up by a set of centers, akin in a sense to the cerebral organs of phrenology, each with its own functional specialization.

The physiological connectionism of the anatomo-clinical approach assumed that interactions between functionally allied centers occurred through the discrete

fiber pathways linking the appropriate cortical regions, hence clinical symptoms and syndromes were divided into those resulting from lesions directly encroaching upon putative specialized centers, and those caused by the interruption of fiber pathways and the attendant disconnection of allied centers. This localization–connectionist view received support not only from existing knowledge of brain anatomy and human neuropathology, but also from the study of the effects of circumscribed brain lesions and electrical stimulations in experimental animals, although such effects were of course restricted to basic sensory and motor functions.

Decline of the Anatomo-Clinical Approach in the First Half of the Twentieth Century

In spite of a number of undisputable achievements, in the first decades of the twentieth century the localizationist–connectionist approach came under severe criticism from different angles, and especially as a result of detailed clinical examinations of young soldiers who had suffered from missile wounds of the brain in World War I. Influential clinical neurologists such as Monakow, Head, Marie, and Goldstein, dissatisfied with the rather crude anatomical and behavioral analyses of their predecessors (caricatured as brain cartographers and diagram-makers), shared Hughlings Jackson's aversion to the attribution of normal psychological functions to cortical areas based on behavioral effects of their lesions. They saw any attempt to localize language and other higher nervous functions in the brain as a futile exercise, and convinced themselves that disorders of higher nervous functions had better be imputed to diffuse alterations of brain organization, in contrast with elementary motor and sensory defects which could instead be precisely related to specific sites of brain damage, as in the case of discrete, topographically organized visual field losses resulting from different bullet trajectories within the occipital lobes. At the same time, general theories of cerebral organization entertained by holistically oriented schools of psychology seemed largely incompatible with the localization of psychological functions in brain centers connected by select fiber paths. Radical learning theorists tended to depict the developing nervous system as a random, unstructured, and essentially equipotential network to be molded and shaped into a functionally adaptive system by use, practice, and conditioning.

Based on his lesion experiments in rats, Lashley, perhaps the most important physiological psychologist of the time, contrived a general theory of the brain based on mass action and functional equipotentiality of the cortex, whereby the severity of functional deficits correlated with the amount of cortex removed, independent of lesion site, and most perceptual and learning functions could be subserved by virtually any part of the cortex. In a different vein, but again in sharp contrast with localization–connectionist views, the Gestalt doctrine of psychoneural isomorphism equated brain activities underlying psychological functions with diffuse electrical fields spreading throughout the cortex as a volume conductor rather than with distinct patterns of nervous impulses traveling along orthodox neuronal circuits. Finally, interest in brain–behavior relationships was discouraged by behavioristic extremists who advocated the restriction of behavior analysis to directly observable input–output relationships of the whole organism, with total disregard for any independent neurophysiological evidence.

Return of Localization and Connectionism

These viewpoints against localization and connectionism were neither completely illogical nor totally devoid of some factual evidence. It is true, for example, that during development there is an excessive production of neurons and neuronal connections, many of which are functionally inappropriate and must be eliminated by the action of natural environmental influences at later stages of maturation. It is also true that the brain is not a mosaic of centers and that functionally specialized populations of neurons tend to be distributed in the brain rather than being concentrated in circumscribed centers. But developments in the neurosciences beginning around the middle of the twentieth century proved beyond doubt that anatomical and physiological specificity and selectivity are the hallmarks of a cerebral organization in which there is no room for either mass action or functional equipotentiality. Work in neuroembryology and neurophysiology showed that well before birth the genes and developmental processes determine highly specific neural connections, which are then maintained and perfected by environmental influences.

Cytoarchitectural and myeloarchitectural subdivisions of the cortex, as described by the early neuroanatomists, were given specific functional meaning by the demonstration that most of these subdivisions contain orderly maps of the sensory or motor peripheries, as well as populations of neurons with quite distinctive physiological properties. An important distinction surfaced between neuronal systems conveying specific sensory or motor information, and deep-brain systems with diffuse projections to the thalamus and cortex which can modulate the activity of the entire nervous system in the sleep–wake cycle

as well as in attentional and motivational regulations. The identification of different synaptic transmitters subsequently furnished evidence for the existence of several of these diffusely projecting systems, each marked by a single transmitter specific for that system and for its action on its cortical and subcortical targets. Better controlled lesion experiments in animals and more accurate and systematic analyses of neurological patients made it clear that, depending on its locus, focal brain damage can bring about unique combinations of impairments and spared psychological capabilities, allowing reasonable inferences about the cerebral localization and organization of normal psychological processes. On the basis of such a renewed localizationist approach, it became feasible both to use the pattern of losses and preservations of functional high-order capabilities for inferring the site of the causative brain lesion, and in turn to employ theories about normal psychological functioning in order to predict the psychological outcome of a particular cerebral injury. The logic of the interpretation of the psychological effects of brain lesions benefited considerably from the distinction between a simple dissociation (lesion X produces impairment of performance on task A but not on task B), which may at least in part depend on task difficulty rather than functional specialization, and the more revealing double dissociation (lesion X produces impairment of performance on task A but not on task B, while lesion Y produces impairment of performance on task B but not on task A), an effective methodological weapon for strengthening the reliability of inferences about functional localization.

Technological Progress in the Aid of Neuropsychology

Starting from the 1960s and 1970s, the advances of the newly born discipline of neuropsychology were helped considerably by technological progress. In the old times neuropsychologists had to wait for postmortem examinations or to rely on entry and exit sites of bullets or other objects in the skull in order to obtain direct anatomical evidence for the locus of brain damage. In the case of closed head brain injuries, it was often the neuropsychologist who inferred the locus of damage on the basis of a specific pattern of psychological deficits. This situation changed drastically when the invention of X-ray computed tomography (CT) scanning in the 1970s allowed one to see the injured structures of the living brain in a noninvasive manner, making it possible to obtain independent parallel evidence on psychological deficits and brain pathology, and to follow their evolution over time. In the following decades, further spectacular advances in the noninvasive imaging of the normal working human brain have been obtained with positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), both of which monitor changes in local blood flow related to metabolism and chemical and electrical activities of neurons and glia. These techniques have shown that diverse cognitive operations are associated with rather distinctive combinations of activated cortical regions interspersed among other cortical regions with unchanged activity. It has therefore become feasible to match the deficits in specific cognitive operations resulting from a localized cortical lesion against the activation of the same cortical region in the normal brain during the execution of the same cognitive operations.

Complementary information to the preceding evidence can now be obtained with transcranial magnetic simulation (TMS), which allows temporary functional inactivation of select cortical areas in normal observers during the execution of cognitive tasks. Further, the perfection of the classical electroencephalographic and evoked-potentials techniques and the introduction of magnetoencephalography have furnished the possibility to obtain global or local brain correlates of behavioral and mental states with a temporal resolution far superior to that of the X-ray or PET or fMRI imaging techniques. Digital computers have enormously facilitated the collection and analysis of behavioral and brain data, and computer science and artificial intelligence have contributed to neuropsychological progress both practically and theoretically. In these sciences, a movement called connectionism attempts to explain human cognition using artificial neural networks, composed of large numbers of units linked by connections with variable strength. The units are the analogs of neurons, and the variable connections mimic synaptic connections. Although artificial neural networks have been proved capable to learn typical neuropsychological tasks (e.g., reading or face recognition) with stimulation-induced changes in the strength of connections between the units, the resemblance between the artificial intelligence connectionism and the anatomophysiological connectionism of neuropsychology is vague. The genes and the forces of development predispose a prenatal, extremely selective pattern of connections between the brain neurons, which artificial neural networks, with their initially random connections, cannot simulate in a veridical manner. Finally, thanks to techniques for recording from single neurons in experimental animals, particularly in freely behaving monkeys, an enormous amount of information has been obtained on the organization of the material substrates of the higher nervous functions in those animals. In many cases this information has provided a decisive

guidance for probing the corresponding substrates of cognition in man.

Some Major Achievements of Neuropsychology

As implied in its original program, neuropsychology has mostly succeeded when progress was made in a convergent fashion on the two fronts – the brain on one front and the mind or, to use a more fashionable term, cognition on the other. Since its start in the early 1960s, neuropsychology experienced profound changes in the understanding of cerebral organization, especially as regards the role of long corticocortical connections in the transfer of sensory, motor, and higher order information. In support of the validity of the concept of anatomo-functional connectionism, Geschwind and others reestablished the reality of aphasic disconnection syndromes contingent on lesions of the long association connections within the cortex of the left hemisphere, and the split-brain studies of Sperry and co-workers demolished one of the alleged cornerstones of Lashley's view of the long corticocortical connections as mere 'skeletal structures,' devoid of any demonstrable mental or behavioral function.

Far from being a mere mechanical link between the two hemispheres, the most developed of these connections, the corpus callosum, was shown to be the essential route for the interhemispheric exchange of information underlying mentation and behavior control. In addition, the independent cognitive abilities of the two separate hemispheres proved to be considerably different, in general keeping with much new evidence from the analysis of the deficits caused by right and left hemispheric lesions. The adoption of appropriate statistical tests in studies of large cohorts of patients with brain lesions in approximately corresponding hemispheric locations led to a definitive rejection of the classical concept of an overall left cerebral dominance in favor of the notion of a complementary functional specializations of the two hemispheres.

The unique specialization of the left hemisphere for all language functions was confirmed not only in right-handers but also in most left-handers, and the right hemisphere was in turn found to be predominant in various cognitive tasks that cannot be aided by verbal mediation, such as the exploration and cognition of space, the discrimination and recognition of unfamiliar and familiar faces, and the perception and memory of colors and complex visual shapes. Lesions of the right hemisphere were attributed a special role in the unilateral hemineglect syndrome as well as in emotional and affective disorders such as anosognosia (e.g., the denial of illness) and the flattening of affect. Discoveries on hemispheric asymmetry of functions in brain-injured patients stimulated a multitude of studies on normal observers tested with different sensory inputs channeled into single hemispheres while responses were measured as simple or choice reaction times, in order to assess various aspects of transfer of cognitive information between the hemispheres. The time-honored methodology of mental chronometry was also applied to the analysis of attentional modulation of cognitive processes, allowing important distinctions between automatic and voluntary orienting mechanisms, as well as between the allocation of attention to spatial locations and that to objects.

The field of memory was revolutionized by the work of Brenda Milner on the famous patient H.M., who had been submitted to a bilateral medial temporal removal for epilepsy control, with a resulting extremely severe anterograde amnesia but normal short-term memory and fair-to-good long-term memory for preoperatory facts and events. The demonstration that this patient could learn at a normal rate and retain skills of which he was totally unaware made a definitive case for a fractionation of memory into components, a case which had been postulated mainly on theoretical grounds but never revealed with such compelling evidence. This groundbreaking discovery started many very successful lines of investigation which led to a fractionation of both short-term and long-term memory into multiple components. The subdivision of long-term memory into the declarative and nondeclarative categories, in their turn subdivided into different subcategories, has been greatly fruitful for neuropsychology because convincing evidence is now available that forms of memory distinguishable at the cognitive level are also associated with distinct neural substrates at cortical and subcortical sites. The discovery of strikingly selective agnosic disorders which affect the information processing of some categories of objects or living beings or their parts, while sparing the handling of information from other categories, has strongly suggested that different knowledge categories are represented in different regions of the cortex, a suggestion strongly supported by data from fMRI and TMS studies in normal observers. Finally, possible approaches to the neurology of consciousness have been furnished by the discovery of dissociations between conscious and unconscious forms of vision, as occur in blindsight and other situations in both brain-damaged patients and normal observers, and the proposed distinction of an occipitotemporal brain system for conscious visual perception, and an occipitoparietal brain system for visually guided action not necessarily accompanied by awareness.

Cognitive Neuropsychology: A New Neuropsychology?

The strong development of neuropsychology in the last decades of the twentieth century is attested by a strong increase in the number of publications specializing in various aspects of the field. In the year 2006 there were about 20 international scientific journals dealing specifically with neuropsychology, 12 of which included in their titles the term 'neuropsychology' or related terms (Table 1). In most cases the adjectives which accompany the noun 'neuropsychology' in the titles of these journals, as well as in those of neuropsychological textbooks and other related publications, are merely indicative of obvious disciplinary subfields and specializations: experimental, clinical, human, animal, comparative, developmental, applied, rehabilitative, and so forth. Unlike these self-explanatory qualifications, the adjective 'cognitive' deserves special attention, because in the past two decades the phrase 'cognitive neuropsychology' has come to denote a new conception of neuropsychology and a more restricted agenda for neuropsychological research. These have been clearly explicated in 1984 in an editorial (unsigned, but attributable to the Editor-in-Chief Max Coltheart) in the first issue of the journal *Cognitive Neuropsychology*, and reiterated in an editorial by Coltheart and Alfonso Caramazza in the same journal 22 years later. Since its inception, neuropsychology had indisputably been cognitive, insofar as it had amply concerned itself with cognition and the brain, where the emphasis is on the 'and,' and the word 'cognition' applies broadly to all aspects of mentation and action control. But in the newly proposed scenario it is argued that cognitive neuropsychology, as a branch of cognitive psychology, should primarily aim at understanding how the human mind works, while considering the question of how the brain works as a secondary if not irrelevant problem. Cognitive processes are seen as computational programs instantiated on a special computer, the human brain, the physical structure of which belongs to a level of reality different from that of the mind and does not necessarily provide insights into the mind computational programs. The mission assigned to the newly conceived cognitive neuropsychology is that of testing on brain-damaged patients theories which assume that cognitive tasks are normally carried out by sets of discrete, separable, and relatively autonomous information-processing components (sometimes called modules, not to be confused with the cortical neuronal modules known to neurophysiology). In particularly informative single cases, the cognitive effects of a local brain lesion are such as to suggest that a postulated single processing component of a cognitive system has been selectively removed, while the remaining parts of the system have remained largely or totally functional. What is important in these data is not the location or nature of the brain lesion, but rather the demonstration that a multicomponent cognitive system, *a priori* represented as a box-and-arrow diagram, has been selectively deprived of one of its elements, either a box or an arrow. While the modularity assumption and the reliance on single cases are common to the classical anatomo-clinical approach, the distinguishing features of cognitive neuropsychology are the priority of the theoretical framework in which the cognitive effects of brain lesions are interpreted, and the conviction that explanation of human cognition must take place at a level of analysis totally separate from the brain level. Understanding of brain functioning is neither a foundation nor a goal of the new cognitive neuropsychology, although psychological theories about normal cognition which are tested against the losses of

Table 1 Titles of scientific journals which contain the noun 'neuropsychology' or its derivatives beginning publication date

Neuropsychologia (1963)
Journal of Clinical Neuropsychology (1979); becomes *Journal of Clinical and Experimental Neuropsychology*, official journal of the International Neuropsychology Society in 1985
Cognitive Neuropsychology (1984)
Developmental Neuropsychology (1985)
Archives of Clinical Neuropsychology; the official journal of the National Academy of Neuropsychologists (1986)
The Clinical Neuropsychologist (1987)
Neuropsychology Review (1990)
Applied Neuropsychology (1994)
Child Neuropsychology. A Journal on Normal and Abnormal Development in Childhood and Adolescence (1995)
Neurocase: *Case Studies in Neuropsychology, Neuropsychiatry and Behavioral Neurology* (1995)
Neuropsychology, Development and Cognition. Section B, Aging, Neuropsychology and Cognition (1996)

From the journal database of the PubMed service of the US National Library of Medicine. There are other journals which do not mention neuropsychology in their titles but are nevertheless neuropsychological in character. These include, for example, *Cortex*, *Behavioral Neuroscience*, *Brain and Cognition*, *Brain and Language*, and *Cognitive Brain Research*.

cognitive capabilities following focal brain injuries may in some cases be inspired and constrained by existing neurological knowledge. As a result of these developments, there exists today a wide range of scientists working in neuropsychology, all of whom call themselves cognitive neuropsychologists, but who entertain quite different views of their study subject and its specific methodologies. At one extreme, closer to the pristine neuropsychological tradition, there are those who address such questions as "what are the cognitive functions of the prefrontal cortex?" "what is the role of the hippocampus in memory?" etc., and use a variety of techniques in animals and humans to obtain proper answers. At the other extreme, radical cognitive neuropsychologists, dubbed 'ultracognitive,' seem to pride themselves on an intentional policy of eschewing any attempt to correlate the pattern of cognitive deficits under examination with the causative changes in cerebral organization. For the latter group, traditional anatomo-clinical studies of cognitive defects from brain lesions, along with all other studies of brain–cognition relationships based on more modern technologies, should be regarded as belonging to cognitive neuroscience rather than to cognitive neuropsychology. There is at present some justified resistance to this rather arbitrary distinction and to the restriction of the usage of the phrase 'cognitive neuropsychology' to research conditions in which the prefix neuro-is justified solely by the presence of a *per se* insignificant brain lesion. Perhaps the increasingly fashionable term 'cognitive neuroscience' will eventually prevail and drive out the term 'neuropsychology' for good.

The Future of Neuropsychology

The term 'neuropsychology' may well be destined to disappear from the neuroscientific vocabulary, but the pristine target of neuropsychology – to understand the relationships between the mind and behavior on one hand, and the brain on the other – will still remain the ultimate goal of the neurosciences of the future. Studies of cognitive alterations from brain damage will perhaps continue to occupy a central place on the scene, but progress will inevitably depend on many other experimental and theoretical approaches and on combined contributions from all biological sciences, from molecular biology to paleoanthropology, as well as from psychology, from cognitive and computer science, and from philosophy. The as yet poorly understood neuropsychology of complex cognitive and emotional derangements as occur in major psychoses (e.g., bipolar disorder, schizophrenia, clinical depression, obsessive–compulsive disorder) will progress to a point where the already blurred border toward neuropsychiatry will eventually vanish. There is still a long, effortful way to go, but the endeavor is worth the effort. To know how the human brain furnishes the medium through which all members of our species cognize, emote, and act would be tantamount to solve the perennial mind–brain problem, an achievement which, in William James' words, would make all previous achievements pale.

See also: Cognitive Neuroscience: An Overview; History of Neuroscience: Early Neuroscience; Neurophysiology: Past and Present; Neuropsychological Testing; Neuropsychology of Primate Reward Processes.

Further Reading

Albright TD, Jessell TM, Kandel ER, et al. (2000) Neural science: A century of progress and the mysteries that remain. *Neuron* 25(supplement): S1–S55.

Anonymous (1963) Editorial. *Neuropsychologia* 1: 1–6.

Anonymous (1984) Editorial. *Cognitive Neuropsychology* 1: 1–8.

Benton AL (1988) Neuropsychology: Past, present and future. In: Boller F and Grafman J (eds.) *Handbook of Neuropsychology,* vol. 1, pp. 1–27. Amsterdam: Elsevier.

Boeglin JA and Thomas RK (1996) Where does the term "neuropsychology" come from? *International Journal of Psychology* 21: 163.

Bruce D (1985) On the origin of the term "neuropsychology." *Neuropsychologia* 23: 813–814.

Caramazza A and Coltheart M (2006) Cognitive Neuropsychology twenty years on. *Cognitive Neuropsychology* 23: 3–12.

Humphreys GW (2002) Cognitive neuroscience. In: Pashler H and Medin D (eds.) *Stevens' Handbook of Experimental Psychology,* vol. l.2, pp. 77–112. New York: Wiley.

Isaacson RL (ed.) (1964) *Basic Readings in Neuropsychology.* New York: Harper and Row.

Kolb B and Whishaw IQ (1996) *Fundamentals of Human Neuropsychology.* New York: Freeman.

McCarthy RA and Warrington EK (1990) *Cognitive Neuropsychology.* San Diego: Academic Press.

Milner B, Squire RL, and Kandel ER (1998) Cognitive neuroscience and the study of memory. *Neuron* 20: 445–468.

Shallice T (1988) *From Neuropsychology to Mental Structure.* Cambridge: Cambridge University Press.

Vallar G (2000) The methodological foundations of human neuropsychology: Studies in brain-damaged patients. In: Boller F and Grafman J (eds.) *Handbook of Neuropsychology,* 2nd edn., vol. 1, pp. 305–344. Amsterdam: Elsevier.

Neurosecretion (Regulated Exocytosis in Neuroendocrine Cells)

J F Morris, University of Oxford, Oxford, UK

Published by Elsevier Ltd.

Introduction

Formulation of the Basic Concept

The concept of neurosecretion originated in the 1930s when Ernst Scharrer reported neurons with secretory cell characteristics in certain nuclei in the hypothalamus of mammals, birds, reptiles, fish, and invertebrates. At the time it was known that the posterior pituitary secreted vasopressin and oxytocin. However, Scharrer's concept of neurosecretion was initially derided because it was widely believed that secretion from the posterior pituitary (neurohypophysis) originated from the glia-like pituicytes characterized by large lipid droplets and that the release was stimulated by axons derived from the hypothalamus and running in the neural lobe stalk. It was only after Wolfgang Bargmann and others performed key experiments involving histochemically staining the neurosecretory material throughout the neurons and then separating the neurohypophysis from the hypothalamus that the hypothalamic origin of posterior pituitary hormones and the concept of neurosecretion were established. Later, Harris, Guilleman, Schally, and others demonstrated that the secretion of anterior pituitary (adenohypophysis) hormones is controlled by peptides and amines released from neurons ending on hypothalamo-pituitary portal capillaries and carried by them to the anterior pituitary. This extended the neurosecretion concept to the parvocellular neurosecretory neurons.

As in most other systems secreting peptides and amines, the secretory products of neurosecretory neurons are present in dense core neurosecretory vesicles (NSVs). Release of the products by neurosecretion from the axonal terminals is therefore the result of exocytosis of the NSVs into the extracellular space and then into the bloodstream and distant targets (neuroendocrine secretion). However, it is clear that neurosecretory neurons also release their products into the brain extracellular space to exert widespread effects on brain structural and functional organization.

The terminals of neurosecretory axons in the neural lobe, median eminence, and other neuroendocrine sites resemble synaptic boutons, and it is clear that the basic mechanism of exocytosis of vesicular contents is similar in both. The classical synapse therefore probably evolved from peptidergic neurosecretory terminals because networks of peptidergic signaling cells are present in primitive organisms which evolved long before the first synapse-containing nervous system.

Model Systems

For many years, the mammalian magnocellular neurosecretory system, the neurons of which release oxytocin or vasopressin (antidiuretic hormone) into the systemic circulation to act, respectively, on uterus and breast or kidney and blood vessels, has remained the model system of choice for generic experiments on neurosecretion (Figure 1). The parvocellular system, which releases growth hormone-releasing hormone, somatostatin, gonadotrophin-releasing hormone (GnRH), thyrotrophin-releasing hormone (TRH), corticotrophin-releasing hormone (CRH), and dopamine, is essential for the control of the anterior pituitary, but the magnocellular system has a number of advantages for experimental purposes: more peptide is produced, the neurohypophysis is easily separated from the rest of the brain, and many of the magnocellular neurons are gathered together into the supraoptic (SON) and paraventricular (PVN) nuclei, one of which (SON) is readily accessible from the ventral surface of the brain for recording and peptide collection. Because many of the basic cellular processes are similar, much of the study of vesicular exocytosis in neurosecretion has also been based on studies on chromaffin or PC12 cells and on brain synaptosomes.

Extension of the Basic Concept: Secretion from All Parts of Neurosecretory Neurons

For many years, it was accepted that the secretory products were released only from axonal terminals. Indeed, it was even suggested that the secretory material known to be present in both the axonal dilations (Herring bodies) and dendrites of the neurons was 'parked' there for storage or destruction because it was not required by the axon terminals (destruction of NSV by smooth endoplasmic reticulum (ER)/lysosome systems does occur in both compartments). Studies using tannic acid capture of exocytosis first showed clearly that any part of the axon within the neurohypophysis, and not just the terminals, is capable of exocytosis of the dense core vesicles. Experiments also showed that peptide could be released in the hypothalamus, but this was assumed to come from axon collaterals. However, in 1989, tannic acid capture of exocytosis visualized by electron microscopy showed unequivocally that the dendrites of magnocellular

neurons can release oxytocin and vasopressin by exocytosis. This has since been extended by sophisticated assay studies to many other peptidergic neurons.

Compartments of Neurosecretory Neurons

Perikarya

The perikarya of magnocellular neurosecretory neurons are large, with the following organelles: extensive rough endoplasmic reticulum, smooth endoplasmic reticulum including Golgi apparatus, variable numbers of NSVs, lysosomes, and mitochondria. Numerous synaptic boutons abut the surface, which is otherwise covered with astrocyte lamellae. Perikarya of parvocellular neurosecretory neurons are smaller but contain the same complement of organelles, although usually many fewer NSVs.

Axons and Their Specializations

Neurosecretory axons leave either the perikarya or the initial segments of the dendrites. As they pass toward the neural lobe, they have dilatations which are not dependent on the presence of clusters of NSVs. As the axons enter the neural lobe, each has many large axonal 'swellings' (the largest of which are called Herring bodies), which do not contact the perivascular basal lamina, and many nerve 'endings' (terminals), which abut the basal lamina. Both contain substantial numbers of NSVs and mitochondria, but the nerve endings in both the neurohypophysis and the median eminence are characterized by the presence of clusters of electron-lucent microvesicles (ELMVs), with a few larger electron-lucent vacuoles which are the recaptured membrane of exocytosed NSVs. Occasional clusters of ELMVs are also found in the axonal swellings, oriented against membrane contacting pituicyte cytoplasm (Figure 2). The ELMVs are known to contain calcium and to be able to accumulate it with high affinity. In neurosecretory terminals in the neurohypophysis, median eminence, and organum vasculosum of the lamina terminalis, ELMVs have been shown to contain significant amounts of the vGLUT2 glutamate transporter and glutamate, but their function is unclear. NSVs are transported down the axons, maturing as they proceed. In the neurohypophysis, they preferentially pass first into the nerve endings and then accumulate in the swellings if not released from the nerve endings or if there is no space for them in the nerve endings. This preferential release of newly

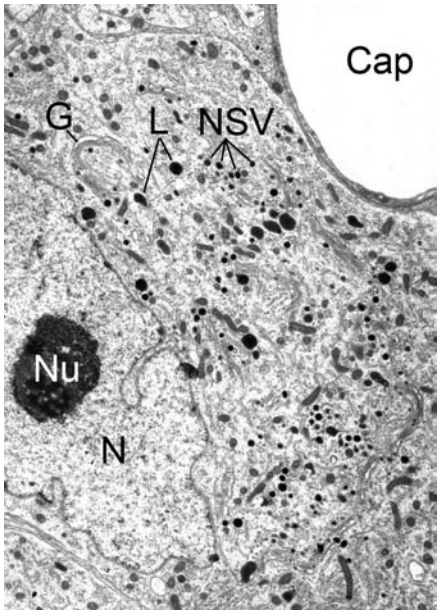

Figure 1 Low-magnification electron micrograph of part of a vasopressin-secreting magnocellular neuron in the supraoptic nucleus of a mouse adjacent to a capillary (Cap). In the cytoplasm, there are many round dense core neurosecretory vesicles (NSV) approximately 160 nm in diameter in addition to larger irregular dense lysosomes (L), mitochondria, and endoplasmic reticulum. Cores of neurosecretory vesicles can be seen being packaged in the numerous Golgi stacks (G). The nucleus (N) contains a prominent nucleolus (Nu). Reproduced from Morris JF and Gainer H (2004) Neurosecretory vesicles. In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience*, 3rd edn. Amsterdam: Elsevier.

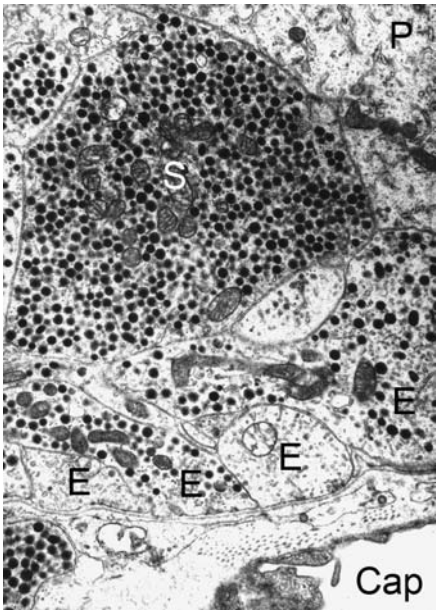

Figure 2 Low-magnification electron micrograph showing the specializations of neurosecretory axons in the posterior pituitary. Undilated parts of the axons give rise to numerous large 'swellings' (S) and 'endings' (E). Only the endings contact the perivascular basal lamina and contain substantial populations of electron-lucent microvesicles. Pituicyte cytoplasm (P) is interspersed between the neurosecretory processes. Cap, capillary.

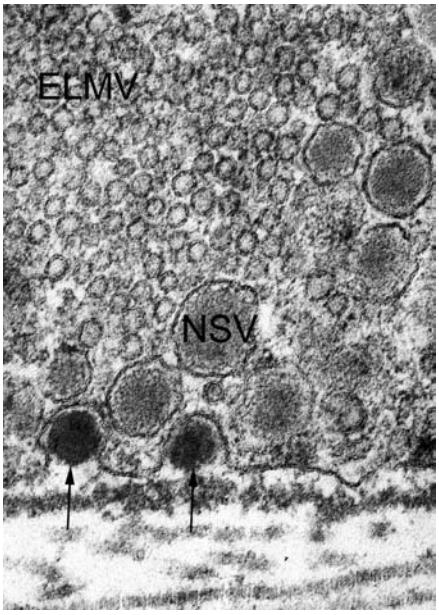

Figure 3 High-power electron micrograph showing part of a neurosecretory nerve ending abutting the basement membrane and perivascular space in the posterior pituitary of a rat. Two neurosecretory vesicle cores (arrows) have been captured in the process of exocytosis by the tannic acid method (which has made their cores more electron-dense). Other neurosecretory vesicles lie at varying distances from the plasmalemma. The nerve ending also contains numerous electron-lucent microvesicles (ELMV) which contain glutamate and can accumulate calcium. Reproduced from Morris JF and Gainer H (2004) Neurosecretory vesicles. In: Adelman G and Smith BH (eds.) *Encyclopedia of Neuroscience*, 3rd edn. Amsterdam: Elsevier.

Figure 4 Low-magnification electron micrograph showing a dendrite of a magnocellular neuron filled with many neurosecretory vesicles and abutted by two synaptic boutons (arrowheads). Reproduced from Kawata M and Leng G (eds.) (2004) Neurohypophysial hormones. *Journal of Neuroendocrinology* 16, p. 404, by permission of Blackwell Publishing.

mature vesicles is similar to that demonstrated for bovine adrenal chromaffin cells. The largest axonal dilatations retain the most aged NSVs and degrade them.

Magnocellular axons appear to have the ability to continue to grow. If the neural lobe is removed or the pituitary stalk damaged, some magnocellular neurons die, but the surviving axons form a mini-neural lobe and function is restored. In part, the restoration of function is due to axons sprouting and making contact with capillaries in the median eminence. In the neurohypophysis, some axons grow through the basement membrane and into the perivascular space (Figure 3). When this happens, perivascular microglia engulf and digest the aberrant axonal processes.

Dendrites

Magnocellular neurosecretory neurons, like most peptidergic neurosecretory neurons, have long but relatively simple dendritic trees, largely devoid of dendritic spines but contacted by many different synaptic boutons (Figure 4). The dendrites emerge from the perikarya with no sharp junction, emphasizing the functional similarity of the two compartments. Dendrites of SON neurons form a dense network beneath the pia just lateral to the nucleus; those of the PVN pass medially toward the third ventricle. Like the axons, many dendrites have large dilatations often filled with NSVs. Substantial amounts of rough ER are present in the dendrites, particularly in the region of synaptic contacts, but there are no recognizable Golgi stacks. However, there are some clusters of smooth ER membranes and small vesicles close to synapses. There is no evidence that the dendrites can package NSVs. This occurs in the perikarya, and the NSVs are transported into the dendrites along the substantial bundles of microtubules in their cores (Figure 5). Dendrites do not, however, contain the clusters of ELMVs which characterize the nerve endings. How NSVs are directed into either the axon or the dendrites is uncertain. The kinesin motor transports NSVs into axons and dendrites, and the glutamate receptor interacting protein (GRIP1) steers kinesin into dendrites. In some invertebrate systems, NSVs are distributed differently according to their contents.

Neurosecretory Vesicles

Neurosecretory vesicles are produced in the cell body of neurosecretory neurons. Their proteins are synthesized in the rough ER, modified and packaged in the Golgi apparatus usually as a larger precursor, and then mature as they pass into the axons or dendrites.

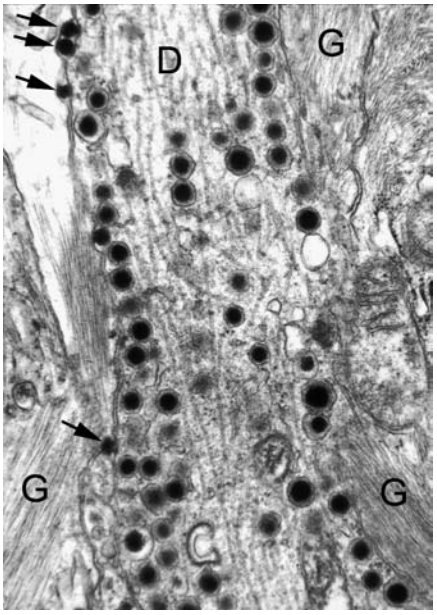

Figure 5 Electron micrograph showing the exocytosis of three neurosecretory vesicle cores (arrows) from a magnocellular dendrite (D) containing many secretory vesicles and a prominent core of neurotubules and surrounded by glial processes (G). Reproduced from Kawata M and Leng G (eds.) (2004) Neurohypophysial hormones. *Journal of Neuroendocrinology* 16, p. 404, by permission of Blackwell Publishing.

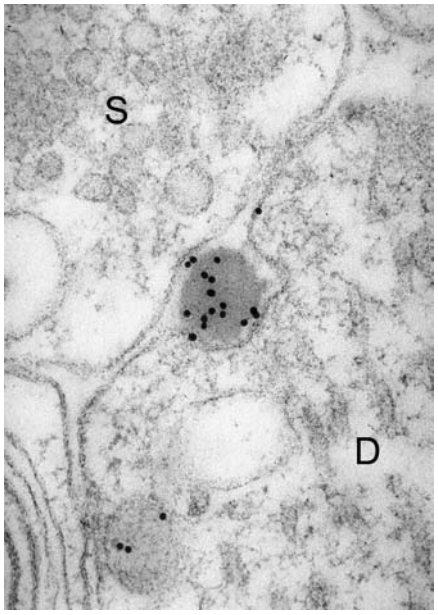

Figure 6 High-power electron micrograph showings the exocytosis of a vasopressin-containing vesicle (immunogold labeled for vasopressin neurophysin) from a dendrite (D). The secretory vesicle core is exocytosed immediately beneath a synaptic bouton (S) contacting the dendrite.

Neurosecretory vesicles vary in size, shape (generally spherical), and the number that are present in the neurons; their peptide/protein or amine content renders them dense core by electron microscopy. They usually contain one major secretory product (e.g., vasopressin or oxytocin) and numerous other compounds. Some of these (e.g., the chromogranins and prohormone convertases) have a role in the formation of the vesicles and their contents, but many others are bioactive signaling molecules present in much smaller amounts than the main hormone. For example, vasopressin vesicles also contain dynorphin, enkephalin, galanin, TRH, PHI, and cholecystokinin; oxytocin vesicles also contain these, but they contain CRH rather than PHI. It has also been shown that salusins, TOR-related peptides thought to be involved in blood pressure regulation, are co-packaged only with vasopressin, and transforming growth factor-β_1 has been added to the ever expanding list of bioactive molecules packaged in magnocellular NSVs. The amounts of the co-stored products vary with physiological status. They are released with the major product and usually have local modulatory actions either on the compartment from which they are released (autocrine) or on surrounding structures (neurons, synaptic boutons, and glia). In certain circumstances, a proportion of magnocellular vesicles contain more than one of the major products (e.g., lactation – vasopressin and oxytocin). The number of neurosecretory vesicles in any system has evolved to match the need. Where there are large fluctuations in demand (e.g., for stress hormones and vasopressin), stores are large; where fluctuations are small (e.g., TRH), stores are much smaller. The size of the vesicles varies in a way that is not obviously related to their content, although presumably the quantal size has evolved to an optimum for that system. Vasopressin vesicles in magnocellular dense core vesicles which secrete into the systemic circulation are approximately 160 nm in diameter, but they are approximately 110 nm in suprachiasmatic neurons which release into the local extracellular space; similarly, the parvocellular peptidergic neurons which release into the very small portal circulation have smaller (~100 nm) dense core vesicles.

In addition to their core contents, vasopressin-containing NSVs (at least) carry vasopressin V_{1a} and V_{1b} receptors in their membranes so that, upon exocytosis, the receptors can be incorporated into the surface membrane and act as autoreceptors (Figure 6). This incorporation is clearly important in the dendrites, but its role in the axonal terminals is unclear. By contrast, the membrane of pituicytes surrounding the axonal terminals is richly supplied with vasopressin receptors. Oxytocin neurons also express oxytocin receptors, but their transport to the membrane via oxytocin-containing NSVs remains to be demonstrated.

Neurosecretion from Axons in the Posterior Pituitary and Median Eminence

Electrical Activation of Release

Release by exocytosis of neurosecretory vesicles from the neurohypophysis is stimulated by the electrical activity of the neurons. Basal release is poorly understood, but it is increased by taurine from the pituicytes acting via $GABA_A$ and glycine receptors. Many neurosecretory neurons fire in burst patterns. Moderately activated vasopressin cells fire phasically, interspersing bursts of action potentials with quiescent periods; these are not coordinated among the numerous vasopressin cells. By contrast, although oxytocin neurons have a variable frequency but steady firing pattern in virgin animals, during lactation they fire in high-frequency bursts 1 or 2 s long and recurring every 5–10 min during the suckling stimulus. These firing bursts are coordinated to approximately 0.5 s not only within a nucleus but also between the oxytocin neurons of the supraoptic and paraventricular nuclei. Action potentials in bursts release proportionally more hormone than do action potentials that are not in bursts. The need for bursting may reflect the electrical properties of the very highly branched and dilated axons within the neural lobe and is known to affect the calcium transients within the nerve terminals.

The Role of Calcium Transients in Neurosecretory Axons

An increase in intraterminal calcium is an essential prerequisite for the exocytosis of NSVs. Spikes invading the terminal arborization can open voltage-dependent calcium channels. Neurosecretory terminals contain a number of these: L- and N-type are present in both oxytocin and vasopressin terminals; Q-type preferentially control vasopressin terminals, whereas R-type control oxytocin terminals. These channels are sensitive not only to voltage but also to opioids which are co-packaged in the NSVs and therefore act as autocrine inhibitors. Release of calcium from intracellular stores of calcium is also involved in stimulated release, but its precise role is uncertain because removal of extracellular calcium blocks depolarization-induced release. Calcium is accumulated and stored in mitochondria, NSVs, and ELMVs, which are a prominent feature of the nerve endings. The bursts of action potentials which characterize neurosecretory neurons are essential not only for generating the increases in intracellular calcium and secretion but also for terminating the activity. High-frequency stimulation causes changes in calcium influx which directly decrease the population spike amplitude via calcium-activated potassium conductances, causing a cumulative increase in spike latency. The calcium extrusion pump plays a major role in decreasing intraterminal calcium; neither sodium–calcium exchange nor calcium storage appears to be involved, despite the observation that a potassium-dependent sodium–calcium exchanger is localized in the neurohypophysial terminals. Caldendrin, a protein associated with the soma–dendritic compartment of brain neurons, is present in variant form in neurohypophysial axons. Its association with Q-type calcium channels suggests a function in triggering calcium transients in vasopressin terminals.

Interestingly, it has been shown that CD38, a transmembrane ADP-ribosyl cyclase which catalyzes the formation of a number of calcium signaling molecules (cyclic ADP-ribose and nicotinic acid adenine dinucleotide phosphate) is important in the control of oxytocin (but not vasopressin release). Lack of it is associated with changes in social behavior reminiscent of autism – a condition for which defects in oxytocin-induced pair bonding are strongly implicated.

Mechanism of Exocytosis in Neurosecretory Axons

Studies on the mechanism of exocytosis in a number of systems, primarily brain synaptosomes and chromaffin cells, have increased our understanding of the mechanisms and molecules involved in vesicle docking, activation, and fusion (exocytosis). Many of the molecules are common to both constitutive and regulated secretory pathways in yeast, invertebrate, and vertebrate systems. The vesicles in 'classical' synapses and in neurosecretory systems have many proteins in common. However, classical synapses contain higher levels of the proteins presumably because of their much faster release. Molecules shared by the two systems include the vesicle membrane proteins synaptotagmin (p65) and synaptobrevin (VAMP), vesicle-associated proteins, rab-3, and munc18; the plasma membrane proteins syntaxin, neurexins, and SNAP-25 (synaptosome-associated protein); and soluble proteins such as NSF (*N*-ethylmaleimide-sensitive factor) and α-, β-, and γ-SNAP (soluble NSF-attachment protein). Some proteins known to be involved in exocytosis appear to be restricted to synaptic vesicles (e.g., synapsin 1 and synaptophysin), whereas others are specific for NSVs (e.g., cytochrome b_{561}). Knockout experiments using chromaffin cells indicate that rabphilin brings together the membrane SNAP-25 and the vesicle protein Rab3A and, in so doing, controls the number of releasable vesicles docked to the plasma membrane. Vesicles docked to the plasma membrane still need 'priming' to make them ready for release. Priming involves the assembly of a SNARE complex in which the vesicular SNARE protein synaptobrevin interacts with SNAP25 and

syntaxin; calcium regulation is provided by the vesicle membrane calcium-sensor protein synaptotagmin. Coupling of synaptotagmin to the SNARE complex involves the vesicle membrane protein snapin, which stabilizes the release-ready vesicles. Calcium-dependent exocytosis in chromaffin cells is significantly reduced in snapin−/− mice.

Readily Releasable Pool of Neurosecretory Material

In all neurosecretory organs, endocrine cells, and probably most secretory cells, even a maximum electrical or depolarization stimulus releases only a small portion of the total hormone or other secretory product available. This protects large stores from being needlessly depleted by a supramaximal stimulus, which is an important mechanism due to the time required to replenish stores by protein synthesis, vesicle packaging, prohormone breakdown, and axonal transport. By contrast, some cell types, such as mast cells, appear to have evolved to be able to secrete all their product upon stimulation. In neurosecretory systems, some vesicles are probably already docked to the membrane and primed for release, whereas others have to be transported to the membrane. In some secretory systems (e.g., chromaffin cells), it appears that partial exocytosis ('kiss-and-run') occurs, but there is no evidence for this in mammalian classical neurosecretory systems. Control of intracellular calcium levels is clearly important in determining what is released, but the precise mechanisms controlling the readily releasable pool are poorly understood. In synapses the RAL_A-GTPases have been shown to be involved in protein kinase C-stimulated release of transmitter from neurons; protein kinase C increases the readily releasable pool of synaptic vesicles docked to the membrane. Certainly, the readily releasable pool of neurosecretory product is not a fixed fraction of the total but, rather, varies with the strength of the stimulus. This is to be expected because there must be a lag between release and the action of the released neurohormone. Secretion of an amount of product determined by the strength of the stimulus allows the feedback control systems fine control without excessive variation about the desired endpoint.

Modulation of Secretion from Neurosecretory Axons

Role of Pituicytes: Neuronal–Glial Plasticity

Pituicytes and their processes surround the neurosecretory axons in the neurohypophysis and occupy part of the perivascular basal lamina on which the nerve endings abut. Pituicytes share some characteristics with astrocytes, and they are probably important in taking up potassium and glutamate released during secretory activity. However, in distinction to astrocytes, they are also characterized by large cytoplasmic lipid droplets, the function of which is unclear. The extent to which the pituicytes envelop the axons varies with the functional state and is reduced when secretion is increased either by osmotic stimulation or in parturition and lactation. Noradrenaline released from nonneurosecretory axons stimulates this pituicyte plasticity. Pituicytes contain significant amounts of taurine; its release is controlled by osmotic conditions, being decreased in hyperosmolar and increased in hypo-osmolar conditions. Taurine acts via strychnine-sensitive glycine receptors to inhibit the depolarization-induced rise in intracellular calcium and subsequent release of vasopressin from the terminals. In this way, the osmotic status controls vasopressin release not only via central osmoreceptors which stimulate synthetic and electrical activity in vasopressin neurons but also locally in the neurohypophysis.

Role of Locally Secreted Factors

Numerous local factors modulate secretion from neurohypophysial terminals. The opioid peptides dynorphin and enkephalin, co-packaged in the NSV, inhibit secretion; the μ-opioid receptor preferentially inhibits oxytocin secretion acting in part via the R-type calcium channels. Similarly, galanin, co-packaged with both oxytocin and vasopressin, inhibits their osmotically stimulated secretion. By contrast, transforming growth factor-β, which coexists in some vasopressin neurons, has not been shown to modulate oxytocin or vasopressin secretion but exerts its effects in the anterior pituitary. Other molecules, not co-packaged, also influence neurosecretion. ATP, which accumulates extracellularly in the stimulated neurohypophysis, stimulates the secretion of vasopressin via a P2X receptor on the terminals, but the metabolism of ATP to adenosine changes the effect because adenosine inhibits terminal calcium and releases via inhibitory adenosine A1 receptors. The ATP/adenosine mechanism therefore strengthens spike/calcium mechanisms to ensure at first an increase in release, but as cellular metabolism generates adenosine, activity in that terminal is switched off and another can take over.

The neural lobe contains scattered noradrenaline terminals, and noradrenaline, via α_1 receptors, stimulates the secretion of vasopressin. It also causes changes in pituicytes such that they occupy less of the perivascular basal lamina, and the neurosecretory terminals can occupy more. Similarly, histamine released from scattered terminals in the neural lobe

increases secretion of both vasopressin and oxytocin via H1 and H2 receptors. α-Melanocyte-stimulating hormone (α-MSH) inhibits release of oxytocin from nerve terminals but has the opposite effect on dendrites. Similarly, the endothelins ET1 and ET3, acting via ETA receptors, stimulate the release of vasopressin in the isolated neurohypophysis but inhibit it in the hypothalamus.

Neurosecretion from Soma and Dendrites

The dendrites of all neurosecretory neurons contain substantial amounts of the secretory product, packaged in NSVs. The amount varies physiologically, as demonstrated by the reduction in dendritic vasopressin during osmotic stimulation and changes in dendritic GnRH during the breeding cycle of starlings. Dendritic secretion of neuropeptides and amines is an important element in the control of neurosecretory neuronal activity. In most systems, release from dendrites is inferred from the release of secretory product into the extracellular space in the region of dendrites and far from the axon terminals. The proximity of the supraoptic nucleus dendrites to the ventral surface of the brain and the large amounts of vasopressin or oxytocin contained in those dendrites have made this a prime experimental model system for the study of dendritic release, which appears to be a universal phenomenon at least among peptidergic and aminergic neurons.

Oxytocin neurons have dendritic oxytocin receptors and vasopressin neurons vasopressin receptors, and both have receptors for opioid peptides in their NSVs. Dendritically released neuropeptide can therefore act in an autocrine and paracrine manner. Oxytocin is also known to act at presynaptic receptors to inhibit transmitter release. This effect is caused by the dendritic secretion of endocannabinoids, aided by an action of oxytocin on local autoreceptors to stimulate the secretion of sufficient endocannabinoid to inhibit the release of glutamate from presynaptic terminals.

There are many more peptides than classical synaptic transmitters, many of which are co-packaged in dense core vesicles. All – whether neuroendocrine or not – are released centrally and act on specific receptors, of which there are many subtypes for each peptide. In many cases, these receptors are abundant in areas where there are few axons containing their peptide ligand. Ludwig and Leng argued that dendritic peptide release has a function strikingly distinct from that of conventional neurotransmitters. Dendrites provide a major source of bioactive peptides with relatively long half-lives which are able not only to act in an autocrine and juxta/paracrine manner on local neurons or glia but also to diffuse to different and distant targets where they can functionally reorganize complex behaviors. Moreover, dendritic-released peptides achieve a much higher concentration (100- to 1000-fold) in the small extracellular space of the brain compared to that of the rest of the body. This allows considerable diffusion from the release site while providing the nanomolar concentrations needed to activate their receptors.

Dissociation of Axonal from Soma–dendritic Release

Given their different functions, it is not surprising that axonal and dendritic release do not occur in parallel and therefore can be differentially regulated. Vasopressin and oxytocin release from axons occurs almost immediately after hyperosmotic stimulation, but their release from dendrites occurs approximately 1 h later and lasts longer. In lactation, oxytocin release in the hypothalamus increases before there is any detectable secretion into the bloodstream. α-MSH, which inhibits the firing of oxytocin neurons and systemic release of oxytocin, acts via MC4 receptors to release intracellular calcium in oxytocin dendrites and stimulate dendritic release.

Mechanism of Soma–Dendritic Release: Role of Calcium and Priming of Dendritic Release

As in the neurohypophysis, an increase in intracellular calcium is essential to stimulate the exocytosis of the dendritic peptide-containing vesicles. That increase in calcium can be stimulated by electrical activity via voltage-gated calcium channels, but release of oxytocin and vasopressin from magnocellular dendrites is stimulated, in the absence of electrical activity, by oxytocin and vasopressin, respectively, via their receptors on the dendrites. The critical source of the intracellular calcium differs in the two systems. In oxytocin cells, an autocrine action of oxytocin mobilizes calcium from thapsigargin-sensitive intracellular stores. By contrast, vasopressin-induced release of vasopressin depends on the entry of calcium from voltage-gated calcium channels and some intracellular release. Because the vasopressin-containing vesicles will not only release vasopressin but also carry vasopressin receptors to the membrane, vasopressin release from the dendrites can be long-lasting. The proteins involved in the exocytosis of dendritic NSVs have not been much studied. It appears that dendrites contain many proteins involved in exocytosis from the nerve terminals, but some, such as synaptobrevin, appear to differ somewhat.

The effect of sex steroids on secretion from magnocellular neurons has long been controversial. Whereas estradiol is known to facilitate oxytocin release, estrogen receptor-β appears to be preferentially

localized to vasopressin neurons. There is morphological evidence that estradiol can induce dendritic exocytosis, apparently acting via NMDA receptors. By contrast, assay or slice preparations indicate that estradiol, progesterone, and the neurosteroid allopregnanolone all induce release of oxytocin and vasopressin from the supraoptic nuclei of young rats by a mechanism that involves $GABA_A$ receptors, and that the effect is much reduced in older animals.

In addition, although spike activity in magnocellular neurons often by itself causes negligible release of peptide, the increase in calcium released from intracellular stores and stimulated by the action of the peptide (or by experimental application of thapsigargin) can 'prime' the dendrites such that subsequent spike activity can, for a long period after the priming, induce substantial dendritic release of peptide.

Role of Soma–Dendritic Release on Neural Glial Plasticity

One of the most striking transitions in adult hypothalamic, if not brain, activity is the change that occurs in oxytocin neurons at the time of parturition in preparation for lactation. Prior to parturition, oxytocin perikarya and dendrites contact one another only at limited, isolated points; the remainder of their surface is separated by lamellae of astrocytic glia. However, during parturition and throughout lactation, the oxytocin cells enlarge, and there are substantial areas of contact between adjacent somata, adjacent dendrites, and between somata and dendrites. The dendrites become grouped as opposed bundles surrounded by astrocyte cytoplasm. Furthermore, whereas in preparturient animals synaptic boutons contact a single postsynaptic cell body or dendrite, double synapses form which contact elements of more than one cell. Associated with these morphological changes is a profound change in electrical activity. Prior to parturition, oxytocin neurons fire with no obvious pattern and no correlation between adjacent cells. During suckling, they fire in coordinated bursts every 5–10 min to generate the bolus of oxytocin secretion required to cause milk ejection from the mammary glands. These profound changes are brought about by the local action of oxytocin in the presence of an appropriate sex steroid milieu.

See also: Exocytosis: Ca^{2+}-Sensitivity; Magnocellular Neurosecretory System: Organization, Plasticity, Model Peptidergic Neurons; Neuroendocrine Peptide Processing; Neuroendocrinology; Peptidergic Receptors; Vasopressin/Oxytocin and Receptors.

Further Reading

Becherer U and Rettig J (2006) Vesicle pools, docking, priming and release. *Cell Tissue Research* 326: 393–407.

Douglas A, Leng G, Ludwig M, and Russell JA (2000) Oxytocin and vasopressin – From molecules to function. *Experimental Physiology* 85(supplement): 1S–267S.

Hirasawa M, Schwab Y, Natah S, et al. (2004) Dendritically released transmitters cooperate via autocrine and retrograde actions to inhibit afferent excitation in rat brain. *Journal of Physiology* 559: 611–624.

Hurbin A, Orcel H, Alonso G, et al. (2002) The vasopressin receptors colocalize with vasopressin in the magnocellular neurons of the rat supraoptic nucleus and are modulated by water balance. *Endocrinology* 143: 456–466.

Hussy N, Brès V, Rochette M, et al. (2001) Osmoregulation of vasopressin secretion via activation of neurohypophysial nerve terminals glycine receptors by glial taurine. *Journal of Neuroscience* 21: 7110–7116.

Jin D, Liu H-X, Hirai H, et al. (2007) CD38 is critical for social behaviour by regulating oxytocin secretion. *Nature* 446: 41–45.

Kawata M and Leng G (eds.) (2004) Neurohypophysial hormones. *Journal of Neuroendocrinology* 16: 291–408.

Leng G and Ludwig M (2006) Information processing in the hypothalamus: Peptides and analogue computation. *Journal of Neuroendocrinology* 18: 379–392.

Ludwig M and Leng G (2006) Dendritic peptide release and peptide-dependent behaviors. *Nature Reviews Neuroscience* 7: 126–136.

Ludwig M and Pittman QJ (2003) Talking back: Dendritic neurotransmitter release. *Trends in Neurosciences* 26: 255–261.

Morris JF (2006) Vasopressin and oxytocin. In: Kastin AJ (ed.) *Handbook of Biologically Active Peptides*, pp. 621–627. Burlington, MA: Academic Press.

Rettig J and Neher E (2002) Emerging roles of presynaptic proteins in Ca^{++}-triggered exocytosis. *Science* 298: 781–785.

Sabatier N and Leng G (2007) Bistability with hysteresis in the activity of vasopressin cells. *Journal of Neuroendocrinology* 19: 95–101.

Theodosis DT (2002) Oxytocin-secreting neurons: A physiological model or morphological neuronal and glial plasticity in the adult hypothalamus. *Frontiers in Neuroendocrinology* 23: 101–135.

Widmer H, Ludwig M, Bancel F, et al. (2003) Neurosteroid regulation of oxytocin and vasopressin release from rat supraoptic nucleus. *Journal of Physiology B* 548: 233–244.

Neurosteroids

M Schumacher, P Robel, and E-E Baulieu, UMR 788 INSERM and University Paris 11, Kremlin-Bicêtre, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Approximately 25 years ago, the observation that some steroids persist in the brain long after removal of the steroidogenic endocrine glands led to the formulation of the neurosteroid concept. This term designates those steroids which are locally synthesized from cholesterol within the nervous system. The study of the biosynthesis, mechanisms of action, and functions of neurosteroids has become a very active field of research. Novel therapeutic perspectives for the diseased, lesioned, and aged brain are offered by the possibility to stimulate the synthesis of neurosteroids with neurotrophic or neuroprotective actions.

Terminology

According to their chemical definition, steroids are small lipids characterized by the cyclopentanoperhydrophenanthrene (sterane) skeleton of four fused rings. This class of molecules includes cholesterol, its precursors starting from lanosterol, the oxysterols, the vitamin D family, the ecdysteroids in insects, the brassinosteroids in plants, and the steroids resulting from cholesterol side chain cleavage (scc) by cytochrome P450scc, which include those commonly referred to as steroid hormones.

The term 'neurosteroid' does not designate a particular class of steroids but, rather, refers to their site of synthesis, the nervous system. Initially, the term was applied to those steroids which are synthesized within the nervous system *de novo* from cholesterol after cleavage of its side chain between carbons 20 and 22. According to this definition, two conditions are mandatory for a steroid to qualify as a neurosteroid: (1) its persistence in nervous tissues after removal of the steroidogenic endocrine glands and (2) the expression and activity within the nervous system of the enzymes involved in its synthesis. The original definition has sometimes been extended to include steroid metabolites formed within the nervous system from circulating precursors. However, their synthesis depends on the presence of peripheral endocrine glands. Estradiol, which is formed in the brain from testis-derived testosterone, is an example. The term neurosteroids is also frequently used to simply designate steroids present and active within the brain. This is very unfortunate because such a broadening of the definition bypasses the original concept: the local synthesis of steroids by neural cells. Steroids with biological effects in the nervous system could be qualified as neuroactive, an extension of the term initially proposed to designate only those steroids which rapidly alter the excitability of neurons by modulating neurotransmitter receptors.

It is well established that almost all the cholesterol present in brain and peripheral nerves is locally synthesized by glial cells and, to a much smaller extent, by neurons. Moreover, cholesterol is not just a structural component of the cell membranes and of the myelin sheaths; rather, it is also an important signaling molecule. One of the oxysterols, 24S-hydroxycholesterol, is almost exclusively formed in the brain, and for this reason it was named cerebrosterol. Thus, cholesterol and oxysterols may qualify as neurosterols.

Historical Background

In the late 1970s, the concept of neuropeptide evolved from the observation that specific peptides are common products of the endocrine glands and of the nervous system. Synthesized from large precursor molecules by nerve cells, neuropeptides were shown to act as neuromodulators or neurotransmitters. In 1981, Etienne-Emile Baulieu and collaborators reported the first evidence that some steroids are locally produced within the brain and designated them as neurosteroids. They were first detected by radioimmunological assays in the rodent brain, where they continue to be present after removal of the steroidogenic endocrine glands by castration and adrenalectomy. The next important step was the demonstration that neurosteroids are indeed synthesized by glial cells and neurons – that is, that functional enzymes involved in their synthesis and metabolism are present, particularly cytochrome P450scc, which catalyzes the first step in steroid hormone synthesis and is only present in steroidogenic mitochondria. From an evolutionary standpoint, the synthesis of neurosteroids has been shown to take place in the brain of all vertebrate phylae, including fish, amphibians, birds, and mammals. Enzymes involved in neurosteroid synthesis are also expressed in the human brain, and the activity of some of them has been demonstrated in biopsies from patients with epilepsy.

Neurosteroid Synthesis and Metabolism

The biosynthesis and metabolism of neurosteroids has been extensively studied in rats and mice. Which steroids can be locally synthesized within the rodent

nervous system? The question needs to be answered cautiously, and results of related studies should be interpreted within their precise context. In fact, the expression and activity of steroidogenic enzymes are subject to complex regulations, and they may be expressed and functional in a defined compartment of the nervous system only under particular conditions, depending on environmental influences, cellular interactions, the presence of neurotransmitters or neuropeptides, the developmental stage, or the integrity of the nervous tissue. It is also important to be aware of some technical constraints. Thus, the relevance of studies on the formation of neurosteroids by cultured neural cells isolated from embryonic or newborn animals needs to be examined *in vivo* because of the phenotypic plasticity these cells exhibit *in vitro*. Also, studies of the expression of steroidogenic enzymes by *in situ* hybridization, reverse transcriptase-polymerase chain reaction, or immunocytochemistry, although they provide very important information, do not demonstrate their functionality.

The synthesis of pregnenolone, progesterone, and its reduced metabolites has been demonstrated within the central and peripheral nervous systems. The conversion of cholesterol to pregnenolone by cytochrome P450scc takes place in the mitochondria, where the enzyme forms a ternary complex with adrenodoxin reductase and adrenodoxin (the cholesterol desmolase complex). The presence of cytochrome P450scc was first detected in the rat and human brain by immunocytochemistry, and then its activity was demonstrated in cultured glial cells. Studies from different laboratories have shown that the enzyme is expressed at low levels throughout the rodent brain and in both glial cells and neurons. Studies have demonstrated the presence and activity of cytochrome P450scc in dorsal root ganglia, the dorsal horns of the spinal cord, and the somatosensory cortex, suggesting an important role of neurosteroid synthesis in the regulation of pain.

Pregnenolone is converted to progesterone by the 3β-hydroxysteroid dehydrogenases (3β-HSDs), present in mitochondria and cytoplasm. 3β-HSD isoenzymes have been cloned from several species: two in human, four in rat, and six in mouse. Unfortunately, the different 3β-HSD isoforms have been numbered according to the chronology of their discovery; consequently, the isoforms of different species do not correspond by number. Most studies do not distinguish isoforms; however, human type 2 and rat type 1 3β-HSDs seem to be the major isoforms expressed in brain. Detailed *in situ* hybridization studies revealed widespread 3β-HSD expression in rat brain and spinal cord during both development and adulthood. Expression and activity of 3β-HSD occur in both neurons and glial cells. After administration of the 3β-HSD inhibitor trilostane, pregnenolone concentrations were markedly increased in the brain of castrated and adrenalectomized rats, thus showing that the conversion of pregnenolone to progesterone is a very active metabolic pathway in the rat brain.

Progesterone is further converted to 5α-dihydroprogesterone (5α-pregnane-3,20-dione) by steroid 5α-reductases. These NADPH-dependent enzymes convert a number of Δ4–3-ketosteroids, including progestagens, glucocorticoids, mineralocorticoids, and androgens, into their 5α-reduced metabolites. Two 5α-reductase isozymes are encoded by distinct genes and designated as type 1 and type 2. The type 1 isoform is expressed throughout the rat brain at all stages of development, whereas the type 2 isoform shows a more restricted distribution and is mainly present in the gray matter of the spinal cord. 5α-Dihydroprogesterone is further converted into allopregnanolone (3α-hydroxy-5α-pregnan-20-one or 3α,5α-tetrahydroprogesterone) by 3α-hydroxysteroid oxidoreductase (3α-HSOR). Only one 3α-HSOR isoform has been identified in rat and mouse, whereas four human isoforms have been cloned. These enzymes are also frequently referred to as 3α-hydroxysteroid dehydrogenases (3α-HSDs), but we prefer the 3α-HSOR nomenclature since they catalyze bidirectional reactions.

A key question regards the presence and activity of cytochrome P450c17 (17α-hydroxylase/17,20-lyase) in the nervous system, since it ensures the formation of androgen and estrogen precursors by catalyzing the conversion of pregnenolone into dehydroepiandrosterone and that of progesterone into androstenedione. Indeed, the other enzymes involved in the formation of androgens and estrogens are expressed in the brain, namely the 17β-hydroxysteroid oxidoreductase, which converts androstenedione into testosterone, and the aromatase, which converts either testosterone into estradiol or androstenedione into estrone. Hence, whether androgens and estrogens are synthesized *de novo* from cholesterol within the brain, or whether they are only formed from circulating precursor steroids, depends on the presence of cytochrome P450c17 in brain. Whereas it is the case for the brain of lower vertebrates, particularly frogs, in which cytochrome P450c17 expression and activity have been unequivocally demonstrated, its presence in mammalian brain long remained controversial until neurons and astrocytes were shown to convert pregnenolone into dehydroepiandrosterone under precise conditions of culture. Cytochrome P450c17 has been detected in the principal neurons of the hippocampal formation, and its expression and activity were demonstrated in the rat spinal cord by a combination of molecular, anatomical, and neurochemical approaches. Local

de novo synthesis of estrogens and glucocorticoids was also evidenced in the rat brain, at least within specific regions. The significance of these observations needs clarification.

Mass spectrometry is a valuable technique for the analysis of steroids and has provided new opportunities for the study of neurosteroid synthesis, metabolism, and role in nervous tissues. The identification and assay of neurosteroids by gas chromatography/mass spectrometry or by liquid chromatography/mass spectrometry has provided reference data for neurosteroid presence and fluctuations within specific brain regions under defined physiopathological conditions. As a consequence, these specific technologies have challenged some well-established concepts, such as the sulfoconjugation of 3β-OH steroids, particularly of pregnenolone and dehydroepiandrosterone, in the rodent brain.

Neural Cells Involved in Neurosteroid Synthesis

According to current knowledge, neurosteroids are synthesized throughout the central and peripheral nervous systems by both glial cells and neurons. It is important to determine in which part of the nervous system and in which cell type a particular neurosteroid is synthesized. However, regarding the question of which steroids can be synthesized within the nervous system, it is necessary to be careful of making excessive generalizations because of the regulated expression of steroidogenic enzymes. Some examples illustrate the complexity of the problem. First, in the adult rat brain, aromatase is exclusively expressed in neurons. However, in response to an excitotoxic or traumatic injury, the enzyme becomes strongly expressed in astrocytes. Second, when cultured at low density, astrocytes preferentially convert pregnenolone into progesterone, whereas pregnenolone is nearly exclusively metabolized to 7α-OH-pregnenolone at confluency. Third, cell culture experiments have provided evidence that the neuroactive progesterone metabolite allopregnanolone is mainly synthesized by glial cells and acts on $GABA_A$ receptors of nearby neurons by a paracrine mechanism. This finding is supported by *in vivo* experiments showing that 3α-HSOR activity in the rat olfactory bulbs is unchanged after the infusion of kainic acid, which destroys neurons but lets glia intact. Conversely, other results have shown that 3α-HSOR is present in the principal glutamatergic neurons of the cerebral cortex, the hippocampus, and the olfactory bulbs and is co-localized with 5α-reductase type 1. Fourth, in peripheral nerves, Schwann cells only express 3β-HSD and synthesize progesterone when in contact with axons. Therefore, these examples indicate that the description of neurosteroid biosynthetic pathways needs to be interpreted according to the experimental context (Figure 1).

Regulation of Neurosteroid Synthesis

The synthesis of steroids in the endocrine glands is regulated by trophic hormones and growth factors. Steroidogenesis involves a basic transcription of the genes encoding steroidogenic enzymes, on which an acute regulation is superimposed, governing the movement of cholesterol into the mitochondria and the modulation of enzyme activities. Very little is known about the regulation of neurosteroid synthesis and metabolism in the nervous system.

The regulatory mechanisms of steroid synthesis in neural cells involve at least in part factors distinct from those operating in the gonads and adrenal glands. Indeed, complex interactions between neural cells obviously play an important role in the regulation of neurosteroid synthesis: specific neurotransmitters and neuropeptides exert strong influences on the expression and activity of steroidogenic enzymes. The role of neuropeptides has been explored mainly in the amphibian brain, and the relevance of these findings to mammalian species needs further investigation. Steroids are also involved in the regulation of neurosteroid biosynthetic pathways. Namely, the *de novo* synthesis of neuroprogesterone is induced by estradiol in the hypothalamus and in cortical astrocytes of ovariectomized and adrenalectomized female rats.

The transport of cholesterol inside the mitochondria, where cytochrome P450scc is located, is a rate-limiting step in the biosynthesis of steroids, and its stimulation results in increased synthesis of neuroactive steroids. Both *in vitro* and *in vivo* studies have demonstrated that the peripheral benzodiazepine receptor (PBR) (also called mitochondrial benzodiazepine receptor or translocator protein) transports cholesterol from the outer to the inner mitochondrial membrane. Hence, PBR agonists stimulate the synthesis of neurosteroids, as shown in cultured glial cells and in the brain of castrated and adrenalectomized rats. Interestingly, strong upregulation of PBR expression is observed in response to injury and during neurodegenerative and demyelinating diseases. These observations suggest that PBR ligands may ensure neuroprotection and myelin repair.

Drugs other than those interfering with the intramitochondrial cholesterol transport also directly affect the biosynthesis and metabolism of neurosteroids. For example, the selective serotonin reuptake inhibitors, used as antidepressants and including fluoxetine, sertraline, and paroxetine, stimulate the formation of

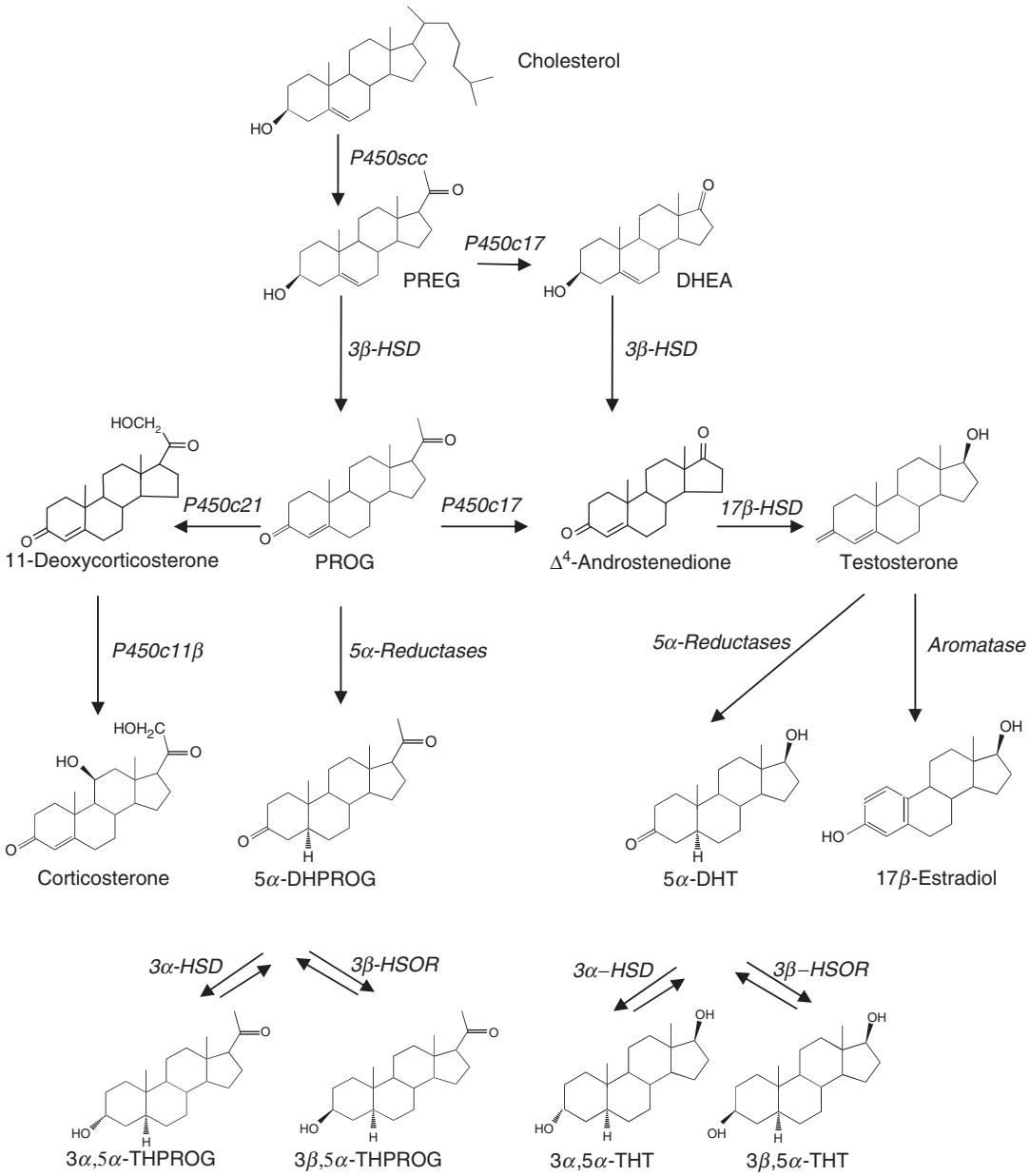

Figure 1 Major biosynthetic pathways of unconjugated neurosteroids and steroid hormones. DHEA, dehydroepiandrosterone; DHPROG, dihydroprogesterone; DHT, dihydrotestosterone; HSD, 3β-hydroxysteroid dehydrogenases; HSOR, hydroxysteroid oxidoreductase; P450scc, cytochrome P450 cholesterol side chain cleavage; PREG, pregnenolone; PROG, progesterone; THPROG, tetrahydroprogesterone; THT, tetradydrotestosterone.

allopregnanolone by increasing the affinity of human 3α-HSD type III for 5α-dihydroprogesterone. Sertraline and fluoxetine also block the reverse oxidative reaction – that is, the conversion of allopregnanolone into 5α-dihydroprogesterone.

Mechanisms of Neurosteroid Actions

Contrary to steroid hormones provided by the endocrine glands, locally synthesized neurosteroids likely act in an autocrine/paracrine manner in the nervous system. It is often claimed that most neurosteroid actions are mediated by the modulation of neurotransmitter receptor activities. However, there is no evidence that the molecular targets of neurosteroids differ from those of steroid hormones reaching the brain via the circulation. Whatever their origin, steroids may influence both gene expression by binding to intracellular receptors and neuronal excitability by directly acting on the cell membrane. The idea of a preferential action of locally synthesized neurosteroids on the neuronal membrane stems from the observations

that the activities of some neurotransmitter receptors, such as *N*-methyl-D-aspartate and nicotinic acetylcholine receptors, are only influenced by high micromolar concentrations of steroids, which are supposedly not reached by the accumulation of circulating hormones. This is mere speculation since it is not technologically feasible to measure the local concentration of a steroid at the level of a few cells or synapses.

During the past few years, there has been tremendous progress in our understanding of the multiple mechanisms of steroid actions in the nervous system. They comprise the regulation of gene transcription, the modulation of neurotransmitter receptors, and the activation of signaling cascades via specific membrane receptors. The recognition of the significance of nuclear coregulator proteins in determining the transcriptional activities of steroid receptors, as well as the cloning of new membrane steroid receptors, is expected to completely change our understanding of steroid action in the near future. Pregnenolone, previously considered as a precursor for other biologically active steroids, has been shown to bind in a selective and saturable manner to microtubule-associated protein (MAP-2), which belongs to a group of proteins involved in the assembly and stabilization of microtubules. By binding to MAP-2, pregnenolone stimulates the polymerization of microtubules and enhances neurite outgrowth of PC12 cells. Based on these findings, it is hoped that new synthetic pregnenolone derivatives with neuroprotective and neurotrophic actions can be developed.

Among the multiple neurotransmitter receptors identified as potential targets for the actions of steroids, the most well studied are the $GABA_A$ receptors. At physiological nanomolar concentrations, the progesterone metabolite allopregnanolone is a potent positive modulator of these receptors. The enantioselectivity of the effects of allopregnanolone on $GABA_A$ receptors and the development of a selective inhibitor of the GABA modulatory effects of 5α-pregnane steroids support the existence of specific binding sites. However, recombinant $GABA_A$ receptor subunit expression coupled with site-directed mutagenesis has not allowed the identification of a steroid-binding pocket on the $GABA_A$ receptor complex, although structural motifs involved in steroid modulation have been described. Thus, the precise mechanisms by which pregnane steroids interact with $GABA_A$ receptors remain undefined. It is important to note that allopregnanolone does not indiscriminately enhance neuronal inhibition via $GABA_A$ receptors throughout the nervous system. Indeed, accumulating evidence indicates that steroid–$GABA_A$ receptor interactions are specific to individual brain regions and different neuronal populations. More important, local neurosteroid metabolism has been demonstrated to influence $GABA_A$ receptor activity.

Biological Significance

The biomolecules have been classified into functional groups. This is also the case for the steroids, which have been grouped into various categories, such as prohormones, sex hormones, and stress hormones. Once such restrictive labels are attributed, it is difficult to get rid of them. However, research during the past few years has revealed the pleiotropic effects of steroids in the nervous system. Indeed, pregnenolone and dehydroepiandrosterone, previously considered as mere precursor molecules, have been shown to exert neurotrophic and neuroprotective effects, respectively, by directly acting on microtubules or on the neuronal membrane. Progesterone and estradiol, usually referred to as female sex hormones, are important regulators of vital neuron and glial functions in the male and female nervous system. Namely, progesterone has been shown to play an important role in the viability and plasticity of neurons and in the formation of myelin sheaths, which insulate and protect the axons. As previously mentioned, the progesterone metabolite allopregnanolone is a potent stimulator of $GABA_A$ receptors, an activity which explains some of its psychopharmacological actions, particularly its anesthetic, analgesic, and anxiolytic effects, as well as its role in stress, depression, memory, seizure susceptibility, and alcohol dependence. Significantly, the modulation of $GABA_A$ receptors by allopregnanolone also plays an important role in neuroprotection and myelin formation, although neuroprotective effects of allopregnanolone may also involve its binding to the pregnane X receptor.

It is within the frame of this conceptual evolution that the search for the biological significance of neurosteroid synthesis has to be placed. In fact, we still know very little concerning their specific functions. A major difficulty is due to the fact that the steroids present in nervous tissues are either locally produced or derived from the endocrine glands. Indeed, most steroids, except the conjugated ones, easily cross the blood–brain and blood–nerve barriers. One strategy to study the specific functions of locally synthesized neurosteroids is to get rid of the peripheral sources of steroid hormones by castration and adrenalectomy. This approach has, for example, allowed the demonstration of an inhibitory influence of neuroprogesterone on $\sigma 1$ receptor function and on memory processes *in vivo*.

The disturbance of vital functions by the removal of the steroidogenic endocrine glands, particularly the

adrenal gland, is a serious drawback, making the study of the physiopathological significance of neurosteroids in long-term processes such as development, regeneration, and aging difficult. For this reason, a number of studies have been limited to the local administration of steroids, receptor antagonists, or enzyme inhibitors in animals with intact gonads and adrenal glands. For example, the local application of the 3β-HSD inhibitor trilostane to the lesioned mouse sciatic nerve has provided evidence for a role of locally synthesized progesterone in peripheral nerve regeneration and myelination. The stereotaxic infusion of neurosteroids into specific brain regions has provided evidence for their involvement in memory processes.

Other studies have established correlations between changes in tissue levels of steroids and particular functions or phenotypic characteristics. Expression of 3β-HSD and levels of progesterone were found to be strongly upregulated in the brains of dysmyelinated mouse mutants and in the spinal cord of streptozotocin-treated diabetic rats. Neuropathic pain was also shown to increase the synthesis of pregnenolone and allopregnanolone in the rat spinal cord. Levels of pregnenolone and progesterone were also significantly increased within the spinal cord of castrated and adrenalectomized rats after injury. All these observations provide strong evidence for the concept that an increase in the synthesis of neurosteroids is involved in the mechanisms by which nerve cells cope with neurodegeneration.

On the contrary, in the brain of the NP-C mouse, an animal model of Niemann–Pick type C disease characterized by neurological deficits and Purkinje cell loss, levels of 5α-dihydroprogesterone and allopregnanolone are decreased, as is the expression of 5α-reductase and 3α-HSOR, especially in the cerebellum. Most important, the systemic administration of allopregnanolone significantly delays the onset of neurological symptoms and prolongs Purkinje cell survival.

With the synthesis of neurosteroids well established, the study of their regulation and their biological functions remains an important challenge. The discovery of neurosteroids has offered new perspectives for neurobiology and therapeutic applications. It is certainly not an easy field of investigation, but the difficulties are similar to those encountered for the neuropeptides.

See also: Adrenal Steroids: Biphasic Effects on Neurons; GABA-A Receptors: Molecular Biology, Cell Biology and Pharmacology; Gene Expression Regulation: Steroid Hormone Effects; Gonadal Steroid Actions on Brain; Sexual Behavior: Neuroendocrine Control.

Further Reading

Baulieu EE (1997) Neurosteroids: Of the nervous system, by the nervous system, for the nervous system. *Recent Progress in Hormone Research* 52: 1–32.

Baulieu EE, Robel P, and Schumacher M (eds.) (1999) *Neurosteroids. A New Regulatory Function in the Nervous System.* Totowa, NJ: Humana Press.

Baulieu EE, Robel P, and Schumacher M (2001) Neurosteroids: Beginning of the story. *International Review of Neurobiology* 46: 1–32.

Belelli D and Lambert JJ (2005) Neurosteroids: Endogenous regulators of the GABA(A) receptor. *Nature Reviews Neuroscience* 6: 565–575.

Hojo Y, Hattori TA, Enami T, et al. (2004) Adult male rat hippocampus synthesizes estradiol from pregnenolone by cytochromes P45017alpha and P450 aromatase localized in neurons. *Proceedings of the National Academy of Sciences of the United States of America* 101: 865–870.

Langmade SJ, Gale SE, Frolov A, et al. (2006) Pregnane X receptor (PXR) activation: A mechanism for neuroprotection in a mouse model of Niemann–Pick C disease. *Proceedings of the National Academy of Sciences of the United States of America* 103: 13807–13812.

Le Goascogne C, Robel P, Gouézou M, et al. (1987) Neurosteroids: Cytochrome P450scc in rat brain. *Science* 237: 1212–1215.

Liere P, Pianos A, Eychenne B, et al. (2004) Novel lipoidal derivatives of pregnenolone and dehydroepiandrosterone and absence of their sulfated counterparts in rodent brain. *Journal of Lipid Research* 45: 2287–2302.

Mellon SH and Vaudry H (2001) Biosynthesis of neurosteroids and regulation of their synthesis. *International Review of Neurobiology* 46: 33–78.

Papadopoulos V, Baraldi M, Guilarte TR, et al. (2006) Translocator protein (18 kDa): New nomenclature for the peripheral-type benzodiazepine receptor based on its structure and molecular function. *Trends in Pharmacological Science* 27: 402–409.

Patte-Mensah C, Kibaly C, Boudard D, et al. (2006) Neurogenic pain and steroid synthesis in the spinal cord. *Journal of Molecular Neurosciences* 28: 17–31.

Paul SM and Purdy RH (1992) Neuroactive steroids. *FASEB Journal* 6: 2311–2322.

Schumacher M, Guennoun R, Robert F, et al. (2004) Local synthesis and dual actions of progesterone in the nervous system: Neuroprotection and myelination. *Growth Hormone & IGF Research* 14: S18–S33.

Stoffel-Wagner B (2001) Neurosteroid metabolism in the human brain. *European Journal of Endocrinology* 145: 669–679.

Tsutsui K, Ukena K, Usui M, et al. (2000) Novel brain function: Biosynthesis and actions of neurosteroids in neurons. *Neuroscience Research* 36: 261–273.

Neurotensin and Receptors

P Sarret, Université de Sherbrooke, Sherbrooke, QC, Canada
P Kitabgi, Institut National de la Santé et de la Recherche Médicale U732, Université Pierre et Marie Curie, Hopital St-Antoine, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurotensin (NT) is a tridecapeptide (Figure 1) that was isolated in 1973 from bovine hypothalamus and shortly after from bovine small intestine. In the brain, NT is predominantly expressed in neurons within discrete brain areas and nuclei, where it acts as a neuromodulator. The effects of centrally administered NT include the well-documented interaction of the peptide with dopaminergic (DA) systems, the ability to induce opioid-independent analgesia, inhibition of food intake, and modulation of pituitary hormone release. In the periphery, NT is primarily produced in endocrine N cells scattered throughout the jejuno-ileal mucosa. The peptide is released on food ingestion and regulates a number of digestive processes, including gastrointestinal motility and pancreatic and biliary secretion. In addition, NT exerts growth-promoting effects on normal gastrointestinal tissues and on cancer cells derived from a variety of organs. Other organs that produce NT include the heart and adrenals.

Like all neuropeptides, NT is synthesized as part of a larger precursor that also contains neuromedin N (NN), a six-amino-acid, NT-like peptide. Pro-NT/NN is differentially processed in the tissues, where it is expressed to give rise to different combinations of active peptides. NT and NN are released from neuroendocrine cells in a Ca^{2+}-dependent manner and exert their effect by interacting with NT receptors located on target cells, of which three have been cloned to date. Two, the NTS1 and NTS2, belong to the family of G-protein-coupled receptors with seven transmembrane domains. The third one, NTS3, is identical to sortilin, a protein that belongs to a family of receptors with one transmembrane domain and is involved in the sorting of proteins to various cell compartments. NT and NN actions are terminated by desensitization of the receptors they activate and by degradation mechanisms that take place in the extracellular space.

This article will review data pertaining mostly to the brain functions of NT.

Biosynthesis, Processing, and Degradation of NT

Biosynthesis

Cloning of canine and bovine pro-NT/NN complementary DNA revealed that NT and NN are part of the same precursor protein. Subsequently, cloning of the rat precursor gene showed it to comprise four exons and three introns, with NT and NN both being encoded in the fourth exon. Studies of pro-NT/NN gene induction in PC12 cells, a neuroendocrine cell line that synthesizes high amounts of pro-NT/NN in response to hormonal stimuli, showed that induction was under the coordinated control of several regulatory sequences, including activating protein-1, cyclic adenosine monophosphate (cAMP), and glucocorticoid response elements.

Rat pre-pro-NT/NN is a protein of 169–170 amino acids that starts with a signal peptide that is quickly removed on translocation into the endoplasmic reticulum to yield pro-NT/NN, whose organization is depicted in Figure 1. The precursor then travels through the regulated secretory pathway, where it undergoes several posttranslational transformations that end with the storage of the final maturation products in secretory vesicles of neuroendocrine cells. Rat pro-NT/NN contains four Lys-Arg dibasic sites (numbered 1 to 4 from N- to C-terminus in Figure 1). Site 1 is located in the middle of the precursor whereas sites 2, 3, and 4 flank and separate NT and NN in the C-terminal portion of the precursor. In tissues that express pro-NT/NN, site 1 is generally unprocessed whereas sites 2, 3, and 4 are differentially cleaved, leading to the production of different combinations of bioactive products represented in Figure 1. Thus, in the brain, pro-NT/NN gives rise to NT and NN. In the gut, precursor cleavage leads mainly to the formation of NT and a large peptide ending with the NN sequence at its C-terminus (large NN). In the adrenals, NT, large NN, and a large peptide ending with the NT sequence (large NT) are the major products.

Processing

Dibasics are the consensus sites recognized and cleaved by processing endoproteases that belong to the recently identified family of proprotein convertases (PCs). PC1 and PC2 are exclusively expressed in the regulated secretory pathway of neurons and endocrine cells whereas PC5-A is expressed in both

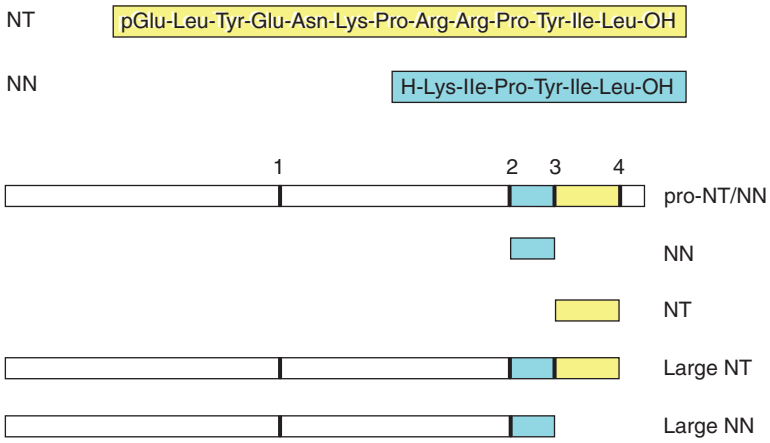

Figure 1 Amino-acid sequence of NT and NN, organization of pro-NT/NN, and major bioactive products generated in tissues by pro-NT/NN processing. The dibasic sites are represented by thick vertical bars and numbered from 1 to 4 from N- to C-terminal.

endocrine and nonendocrine cells. The cleavage specificity of the three PCs toward pro-NT/NN was studied in the neuroendocrine PC12 cell line that lacks PC1 and PC2 expression and presents very little PC5-A. As expected, wild-type PC12 cells poorly processed pro-NT/NN. In contrast, when stably transfected with either of the three PCs, each resulting transfectant readily processed the precursor, albeit with different patterns. Thus, in PC1–PC12 cells, cleavages at sites 3 and 4 proceeded almost to the same extent and were more efficient than cleavage at site 2. This predominantly generated NT and large NN and produced only small amounts of NN (Figure 2). In PC2–PC12 cells, pro-NT/NN was cleaved equally at sites 2, 3, and 4, thereby producing equivalent amounts of NT and NN (Figure 2). In PC5-A–PC12 cells, site 4 was cleaved to a greater extent than site 3 whereas site 2 was very poorly processed, thus resulting in the production of NT, large NT and large NN in comparable amounts and of virtually no NN (Figure 2). It thus appears that PC1, PC2, and PC5-A process pro-NT/NN with patterns that are similar to those observed in the gut, brain, and adrenals, respectively.

Several pieces of evidence indicate that PC2 is indeed the major brain pro-NT/NN convertase. Thus, PC2 is the most abundant prohormone convertase in the brain and therefore, not unexpectedly, co-localizes extensively with immunoreactive NT in neuronal populations that express pro-NT/NN. It is interesting that subsets of NT/PC2-immunopositive neurons are also positive for PC1 and PC5-A, but in general, those neurons expressing PC1 do not contain PC5-A and vice versa. That PC2 plays a major role in processing brain pro-NT/NN is further demonstrated in PC2$^{-/-}$ mice in which brain levels of NN are decreased by more than 50% with a compensatory

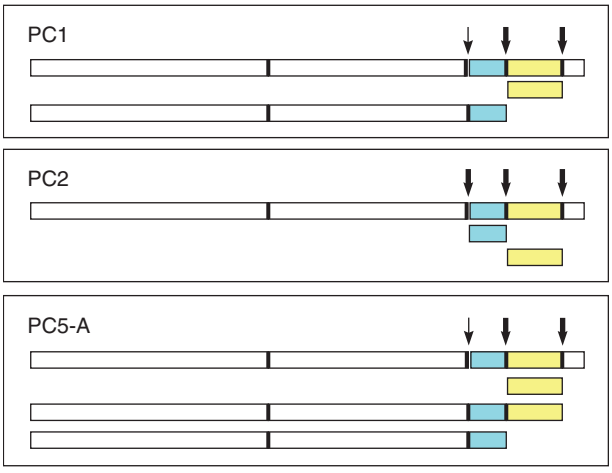

Figure 2 Cleavage specificity of pro-NT/NN and major products generated by proprotein convertase (PC) 1, PC2, and PC5-A. The same representation as in Figure 1 is used. Arrow sizes are indicative of the extent to which each dibasic is cleaved by the PCs.

increase in the levels of large NN, indicating that processing at site 2 is impaired, as would be expected from the cleavage specificity of PC2. There is also a 20% reduction in NT levels in PC2$^{-/-}$ mice, which is less than would be expected, had PC2 been the sole pro-NT/NN convertase in brain, supporting the idea that PC1 and PC5-A are also active in processing brain pro-NT/NN.

Degradation

Whereas classical neurotransmitter action in the synapse is often ended through specific reuptake of the transmitter into nerve terminals, the favored mechanism for terminating neuropeptide action entails proteolytic degradation. NT degradation was initially studied *in vitro* with brain membranes, which identified primary cleavage sites in the NT

sequence: Arg8–Arg8, Pro10–Tyr11, and Tyr11–Ile12 peptide bonds. The use of general and specific peptidase inhibitors showed that the Arg8–Arg9 cleavage was due to endopeptidase EC 3.4.24.15 (also known as thimet oligopeptidase) and that EC 3.4.24.11 (also referred to as enkephalinase, neutral endopeptidase, neprilysin, CD10, or CALLA) was entirely responsible for the Tyr11–Ile12 cleavage and only partly responsible for that at the Pro10–Tyr11 bond. Finally, an unidentified metallopeptidase was shown to account for the remainder of the Pro10–Tyr11 bond-cleaving activity. This enzyme, endopeptidase EC 3.4.24.16 (also designated neurolysin), was later purified and cloned. These data are summarized in Figure 3. All three peptidases belong to the class of Zn-metalloendopeptidases and are specialized in cleaving small bioactive peptides in the extracellular space. They have a wide distribution in the brain and peripheral tissues, and it was indeed demonstrated in all the tissues examined that NT was degraded by a combination of the three metallopeptidases. Their action on NT effectively leads to complete biological inactivation of the peptide as its ability to activate its receptors is entirely borne by the C-terminal hexapeptide sequence which is precisely the target of the metalloendopeptidases. Other peptidases are involved in further degradation of the NT breakdown products generated by the three metallopeptidases (Figure 3). They do not act as primary inactivating enzymes and are therefore less relevant to NT inactivation.

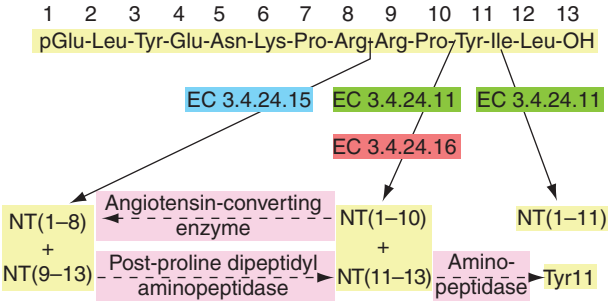

Figure 3 Scheme of NT degradation in brain. Primary inactivating cleavages are indicated by solid arrows, and further degradation of breakdown products by dashed arrows. The primary inactivating metallopeptidases, EC 3.4.24.11, 15, and 16, are highlighted in green, blue, and red, respectively. The secondary cleaving enzymes are boxed in light violet.

Brain Distribution of NT

Originally isolated from bovine hypothalamus, the neuropeptide NT was further localized throughout the central nervous system of several mammalian species, such as rat, calf, monkey, and human. Initial radioimmunoassay, carried out on rat brain extracts, revealed that the highest concentrations of NT immunoreactivity are detected in the hypothalamus, amygdala, spinal cord, and brain stem structures such as periaqueductal gray, dorsal raphe nucleus, substantia nigra, and ventral tegmental area (VTA). Subsequently, high densities of NT-immunoreactive nerve cell bodies and processes were detected throughout the brain and spinal cord by immunohistochemistry (Figure 4). In particular, NT-containing axonal fibers and/or NT-positive cell bodies are abundant in the

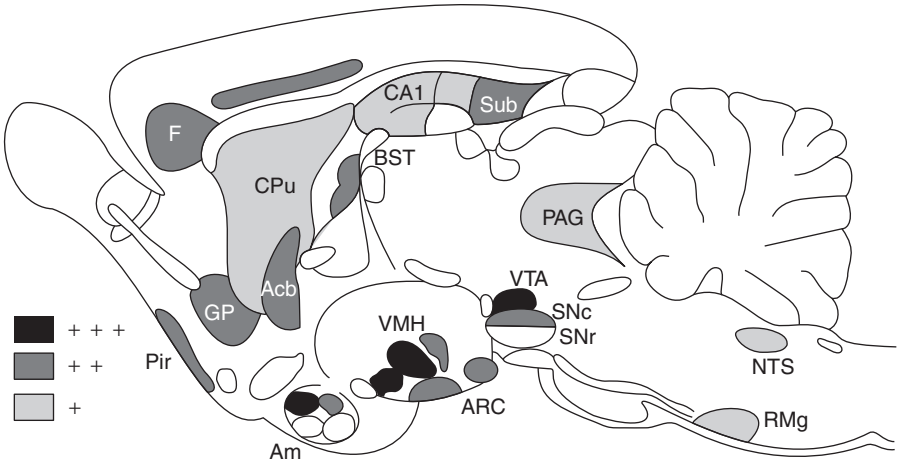

Figure 4 Topographic distribution of NT-containing neurons in the rat central nervous system. Schematic midsagittal section of the rat brain provides an overview of the distribution of NT immunoreactivity throughout the neuraxis. Acb, accumbens nucleus; Am, amygdaloid nucleus; ARC, arcuate hypothalamic nucleus; BST, bed nucleus of the stria terminalis; CA1, field CA1 of Ammon's horn; CPu, caudate-putamen; F, frontal cortex; GP, globus pallidus; NTS, nucleus of the solitary tract; PAG, periaqueductal gray; Pir, piriform cortex; RMg, raphe magnus; SNc, substantia nigra, compacta part; SNr, substantia nigra, reticulata part; Sub, subiculum; VMH, ventromedial hypothalamic nucleus; VTA, ventral tegmental area.

nucleus accumbens; bed nucleus of the stria terminalis; medial and lateral septal nuclei; central, basolateral, and medial amygdaloid nuclei; subparafascicular and gustatory nuclei of the thalamus; paraventricular, and arcuate hypothalamic nuclei; lateral hypothalamus; substantia nigra; VTA; periaqueductal gray; raphe nuclei; locus coeruleus; nucleus of the solitary tract; and substantia gelatinosa of the spinal trigeminal nucleus and dorsal horn of the spinal cord. The topographic distribution of NT immunoreactivity overlapped extensively with that of NT-expressing cells as determined by *in situ* hybridization histochemistry. There were a few regions, however, in which NT immunolabeling appeared more restricted than from *in situ* hybridization data. For instance, the subiculum and CA1 region of the hippocampal formation, which were reported to contain high levels of NT messenger RNA (mRNA), were barely immunoreactive. These discrepancies could be due to selective targeting of NT peptides to neuronal processes.

In the developing rat brain, NT was found, by radioimmunoassay, immunohistochemistry, and *in situ* hybridization, to be expressed very abundantly during the first two postnatal weeks, after which its concentration rapidly decreased to adult levels. This overexpression is regionally selective, affecting predominantly cells located in the retrosplenial and cingulated cortices, the pyramidal cell layer of the subiculum and hippocampal CA1, and the mitral cell layer of the olfactory bulb. The transient overexpression of NT during brain development suggests that this neuropeptide may be involved in the establishment of neuronal circuitry.

NT Receptors and Signaling

Central effects of NT and its analogs are mediated by the activation of at least three different receptor subtypes, identified as NTS1, NTS2, and NTS3 (also called gp95/sortilin). A summary of the biochemical, biological, and pharmacological properties of these receptors is provided in Table 1.

NTS1 and NTS2 belong to the G-protein-coupled receptor family and were initially distinguished pharmacologically on the basis of their affinities for NT and their differential sensitivity to the antihistamine-1 drug levocabastine. Referred to as the high-affinity ($K_{dv} = 0.1$–$0.3\,\text{nmol}\,\text{l}^{-1}$) levocabastine-insensitive binding site, NTS1 has been linked, through Gi/o, Gq, and Gs protein coupling, to a variety of signaling cascades, including formation of cyclic guanosine monophosphate and cAMP, phospholipase C, and inositol phosphate signaling cascades; arachidonic acid production; and activation/inhibition of mitogen-activated protein kinases (ERK1/2 and JNK) and serine/threonine protein kinase Akt. Association of NT to NTS1 induces internalization of receptor–ligand complexes into NTS1-expressing cells via clathrin-coated pits. The endocytic process is dependent on the integrity of the C-terminal extremity and requires the recruitment of dynamin, intersectin, and both isoforms of β-arrestin 1 and 2. NTS1 internalization is accompanied by NT endocytosis. However, the intracellular trafficking patterns of NTS1 and of NT may be distinct, depending on the time period of agonist exposure. Biochemical and cell imaging studies revealed that under acute agonist exposure, internalized NTS1 receptors are targeted to lysosomes for degradation whereas under prolonged agonist exposure, NTS1 is routed toward the perinuclear *trans*-Golgi recycling compartment. NTS1 gene activation and *de novo* receptor synthesis therefore contribute to cell resensitization.

The levocabastine-sensitive, low-affinity NT receptor (NTS2) was cloned from mouse, rat, and human brain by a strategy based on sequence homology with the known NTS1 receptor. NTS2 shares approximately 60% amino-acid homology with NTS1 and binds NT and levocabastine with similar affinity ($K_{dv} = 2$–$10\,\text{nmol}\,\text{l}^{-1}$). In contrast to NTS1, the signaling properties of NTS2 are still controversial, exhibiting cell-type expression- and species-dependent pharmacological properties (Table 1). For instance, depending on the heterologous expression cell system (oocytes, CHO, COS-7, and HEK293 cells) used to characterize the signal transduction mechanisms of the NTS2, levocabastine and NT analogs were reported to act as agonists, inverse agonists, or competitive antagonists at NTS2 sites. The most convincing data supporting an agonist role for endogenous NT on NTS2 receptors are, however, the ability of NT to induce a sustained mitogen-activated protein kinase (MAPK) p42/p44 activation in rat cerebellar granule cells expressing endogenous NTS2. Different results have also been reported on the constitutive activity of the human NTS2 (hNTS2) transfected in COS and CHO cells.

Although the agonistic properties of NT on the NTS2 receptor remain a debate, NTS2 receptors were found to efficiently internalize on NT binding via a clathrin-coated pit and dynamin-dependent mechanism, a property usually associated with an activation by agonists. Note that ligand–receptor complexes were found to form clusters at the cell surface but not to internalize in cultured astrocytes following NT stimulation, suggesting that NTS2 internalization capacities are likely cell type dependent. In contrast to NTS1, sequestrated mouse NTS2 (mNTS2) receptors are efficiently recycled to the plasma membrane. The third intracellular loop and

Table 1 Biochemical, pharmacological, and physiological properties of NT receptor subtypes NTS1, NTS2, and NTS3

NT receptor subtypes	*NTS1 (NTRH)*	*NTS2 (NTRL)*	*NTS3 (gp95/sortilin)*
Receptor classification	G-protein-coupled receptor (7 TM)	G-protein-coupled receptor (7 TM)	Type I receptor (1 TM)
Cloned	rNTS1 mNTS1 hNTS1	rNTS2 mNTS2 hNTS2	rNTS3 mNTS3 hNTS3
Receptor size	rNTS1 and mNTS1 = 424 aa hNTS1 = 418 aa	rNTS2 and mNTS2 = 416 aa Variant isoform (281 aa, 5 TM) hNTS2 = 410 aa	mNTS3 = 825 aa hNTS3 = 833 aa
Molecular weight	50 and 60 kDa	45 kDa	100 kDa
Gene characteristics	hNTS1 gene localized to the long arm (20q13) of chromosome 20 Tetranucleotide repeat polymorphism located <3 kb from the gene	mNTS2 gene mapped at 6cM from the centromere on chromosome 12 Two splice variants derived from a single gene	hNTS3 gene mapped to the proximal short arm of chromosome 1 Two sortilin mRNA transcripts 3.5 and 8.0 kb expressed in human CNS
K_{dv} = 0.1–0.3 nmol l^{-1} (rat brain synaptic membranes)	K_{dv} = 0.1–0.3 nmol l^{-1}	K_{dv} = 2–4 nmol l^{-1}	K_{dv} = 0.3 nmol l^{-1} (CHAPS-solubilized human brain receptors) K_{dv} = 17 nmol l^{-1} (CHAPS-solubilized cloned hNTS3 in COS-7 transfected c)
Maximal binding capacity (Bmax) (adult rat brain homogenate)	Bmax = 13 fmol/mg	Bmax = 170 fmol/mg	Bmax = 150 fmol/mg (CHAPS solubilized human brain receptors)
Levocabastine sensitivity	–	+	–
GTP sensitivity	+	–	?
Na^+ sensitivity	IC50 = 15 mmol l^{-1} (Asp-113)	IC50 = 200 mmol l^{-1}	?
Internalization	65–70% of total binding $t_{1/2}$ = 7 min C-terminal tail (TLY motif)	60% of total binding $t_{1/2}$ = 10 min Intracellular loop 3 (I3)	75% of total binding $t_{1/2}$ = 5 min ?
Recycling	–	+ Y237	+
Cellular localization	Mainly neurons	Neurons and astrocytes	Neurons and glia
G-protein coupling	Gi/o, Gs, Gq	ND	ND
Signal transduction	(+) (−)	(+) (−) (*)	
Human All systems	Phospholipase C NT SR48692 ↑ cytosolic [Ca^{2+}] NN SR142948A ↑ IP Stimulation of cGMP, cAMP MAPK activation Krox-24 gene induction	Phospholipase C NT NT NT ↑ cytosolic [Ca^{2+}] Levo Levo Levo ↑ IP SR48692 JMV 431 MAPK activation SR142948A Arachidonic acid release Constitutive activity ↑ IP	Modulation of NTS1-mediated IP turnover and MAPKs Desensitization of calcium mobilizing effects of NT Induction of PI3 kinases and MAPK pathways

Continued

Table 1 Continued

NT receptor subtypes	*NTS1 (NTRH)*			*NTS2 (NTRL)*		*NTS3 (gp95/sortilin)*
Rat						
All systems	Inhibition of cAMP production	NT	SR48692	↑ cytosolic [Ca^{2+}]	JMV449	
		NN	SR142948A	MAPK activation	NT	
	Arachidonic acid production				SR48692	
					Levo	
Primary cultured cerebellar granule ¢				↑ cytosolic [Ca^{2+}]	SR48692	
				MAPK activation	NT	
					Levo	
Mouse				Activation of Ca^{2+} dependant Cl^{-} conductance	NT	Activation of Akt kinases and ERK 1/2
All systems					Levo	
					SR48692	
Physiological implications	NT-induced turning behavior			NT-induced analgesia		Elicits migration of glial ¢
	Modulation of dopamine transmission			NT-induced hypothermia?		NT-induced c growth
	NT-induced ¢ growth			Effects of NT not blocked by SR48692		NT-induced c growth blocked by SR48692
	Effects of NT blocked by SR48692 and SR142948A			Effects of NT blocked by SR142948A		Sorting and internalization of NT to the TGN
						Inhibitory effect of NT on the proapoptotic effect of NTS3 and p75NTR interaction

aa, amino acids; cAMP, cyclic adenosine monophosphate; cGMP, cyclic guanosine monophosphate; CNS, central nervous system; ERK, extracellular signal-regulated kinase; h, human; K_{dv}, dissociation constant; m, mouse; MAPK, mitogen-activated protein kinase; ND, nondefined; NN, neuromedin neurotensin; NT, neurotensin; TGN, *trans*-Golgi network; TM, transmembrane domains; ¢, cells; +, agonist;-, antagonist; *, inverse agonist; Levo, Levocabastine; ↑, increase

in particular the tyrosine-237 residue are crucial for mNTS2 receptor internalization and recycling, respectively. Recently, the neuron-enriched endosomal protein of 21 kDa (NEEP21) was shown to be involved in mNTS2 recycling. However, species discrepancies are here again detected. In opposition to mNTS2, the hNTS2, in which a cysteine residue naturally replaces the tyrosine in position 237, does not undergo recycling following endocytosis. The underlying mechanisms responsible for NTS2 receptor preservation at the cell surface seem to be different depending on the length of agonist exposure. Indeed, whereas under acute agonist exposure, maintenance of cell surface receptors seems to be related to NTS2 recycling, under persistent exposure to NT, this preservation of cell plasma membrane NTS2 receptor densities appears to involve recruitment of spare receptors from internal stores, thereby preventing functional desensitization.

Originally identified as the intracellular sorting protein sortilin, the NTS3 receptor is unique among neuropeptide receptors in that it is a single transmembrane domain protein. NTS3 exhibits striking sequence similarities with proteins involved in intracellular transport and trafficking. Accordingly, several studies have demonstrated a predominant association of NTS3 with intracellular sorting organelles such as the Golgi apparatus, endosomes, and Glut4 vesicles. Although only 5–10% of this receptor is expressed at the cell surface, NTS3 efficiently internalizes NT. NTS3 has the ability to bind several ligands in addition to NT, such as the lipoprotein lipase, the endoplasmic reticulum resident receptor-associated protein, and its own propeptide released after maturation of the precursor form by furin, suggesting that its role probably exceeds that of an NT signaling receptor. Binding of NT to NTS3 was reported to elicit proliferative responses in a variety of cancer cell lines and to alter agonist-induced cellular trafficking and signaling of NTS1 in HT29 human colon cancer cells. However, the strongest evidence that NTS3 participates in NT signaling comes from the demonstration that NTS3 mediates the NT-induced migration of human microglial cells by a mechanism dependent on the stimulation of both MAPK and phosphoinositide-3 kinase pathways. Recently, sortilin has also been shown to interact with the p75 neurotrophin receptor (p75NTR) in the induction of apoptosis on binding unprocessed nerve growth factor (NGF) proNGF. However, the occupancy of NTS3 by NT prevents the proNGF-induced formation of the sortilin–p75NTR receptor complex and thereby counteracts the p75NTR-induced neuronal cell death. Taken together, these studies suggest that NT binding to NTS3/sortilin may play an important role in inflammation and neurotoxicity.

Localization of NT Receptor Subtypes

Initial evidence for the existence of NT receptors was derived from radioligand-binding experiments on membranes prepared from brain and gastrointestinal tissues. However, it is only with the application of autoradiographic-binding techniques, using [^{3}H]NT or [^{125}I]NT as radioligand, that a comprehensive picture of brain NT receptor distribution began to emerge. Subsequently, the cloning of NT receptors allowed the generation of mRNA probes and antibodies to investigate their distribution by Northern blotting, *in situ* hybridization, and immunohistochemistry at light and electron microscopic levels.

NTS1 Receptors

Autoradiographic studies using [^{125}I]NT in the presence of blocking concentrations of levocabastine (to prevent labeling of the low-affinity site) have revealed a selective association of NTS1-binding sites with selective neuronal populations throughout the hypothalamus, basal forebrain, and limbic system (Figure 5). High concentrations of NTS1-binding sites were distributed in the perikarya and dendrites of subsets of chemospecific neurons in a variety of brain regions, including DA neurons in the ventral midbrain, cholinergic neurons in the basal forebrain, and vasoactive intestinal peptidergic neurons in the suprachiasmatic nucleus of the hypothalamus. This distribution is consistent with reported NTS1-mediated NT effects on locomotion, memory, cognition, circadian rhythms, and sleep–waking cycle, as well as on the regulation of hypothalamopituitary functions. Accordingly, all these regions were later found to contain numerous nerve cell bodies expressing NTS1 mRNA and/or NTS1 receptor proteins by *in situ* hybridization and immunohistochemistry, respectively. The highest levels of staining were observed in the islands of Calleja, diagonal band of Broca, magnocellular preoptic nucleus, pre- and parasubiculum, suprachiasmatic nucleus, anterodorsal nucleus of the thalamus, substantia nigra, and VTA.

At the cellular level, NTS1 receptors were predominantly associated with perikarya and dendrites in some regions (e.g., in the basal forebrain, ventral midbrain, pons, and rostral medulla) and with axons and axon terminals in others (e.g., in the lateral septum, bed nucleus of the stria terminalis, neostriatum, paraventricular nucleus of the thalamus, and nucleus of the solitary tract). These labeling patterns

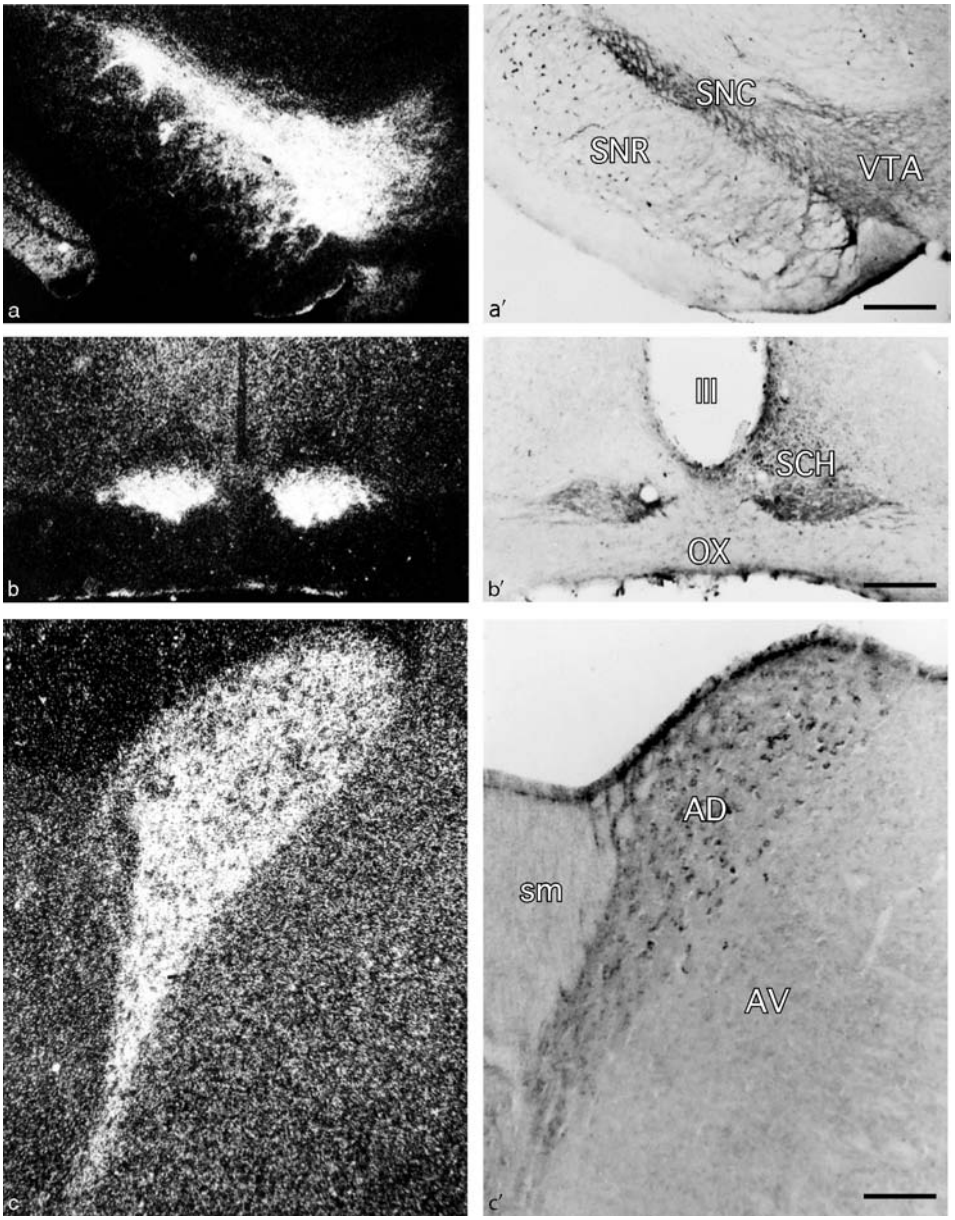

Figure 5 Comparative distribution of [125I]NT-binding sites (dark-field, left) and NTS1 immunoreactivity (bright-field, right) in the substantia nigra (SN; (a), (a′)), the supraschiasmatic nucleus (SCH; (b), (b′)), III: third ventricle; OX: opticchiasm and the anterodorsal thalamic nucleus (AD; (c), (c′)), sm: stria medullaris; AV: anteroventral thalamic nucleus). In all regions, the labeling patterns are strikingly similar. In the substantia nigra, compacta part (SNC) and the ventral tegmental area (VTA), NTS1 immunoreactivity is detected within both perikarya and processes, whereas in the substantia nigra, compacta reticulata (SNR), the labeling is mainly observed over dendritic arbors. Within the SCH and AD, the immunoreactivity is also apparent within perikarya and dendrites in the ventral tier of the nucleus. Scale bars = 300 μm (a, a′); 200 μm (b, b′, c, c′). Reprinted Boudin H, Pélaprat D, Rosténe W, and Beaudet A (1996) Cellular distribution of neurotensin receptors in rat brain: Immunohistochemical study using an antipeptide antibody against the cloned high affinity receptor. *Journal of Compartive Neurology* 373(1): 76–89, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Ltd.

suggest that neurons expressing NTS1 receptors target these receptors to both their somatodendritic and axonal elements, providing evidence for pre- and postsynaptic effects of NT in different brain regions. At the subcellular level, in all areas examined, a significant proportion of both immunoreactive and radiolabeled NTS1 was found in association with neuronal plasma membranes. In brief, in the nucleus accumbens, NTS1 receptors were mainly associated with the plasma membrane of axon terminals, whereas they were mostly observed over the membrane of DA neurons in the substantia nigra and VTA. By contrast,

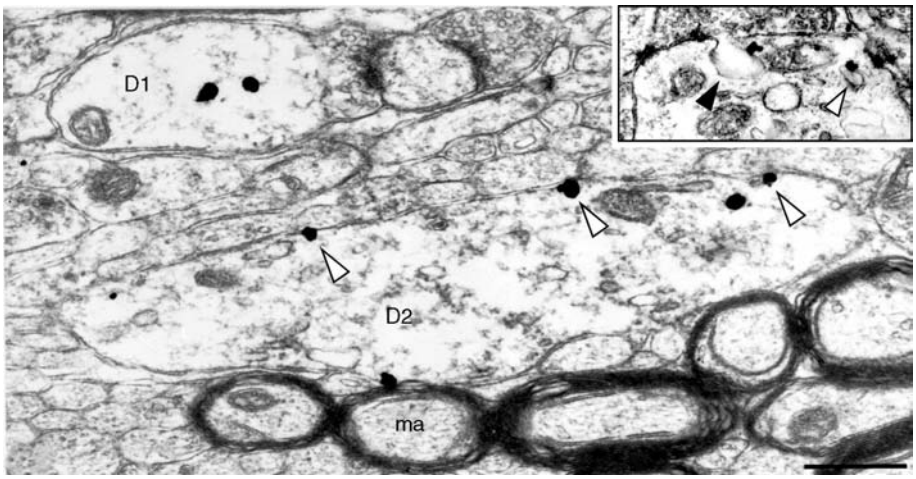

Figure 6 Immunolabeled NTS1 receptor proteins associated with dendrites (D) in rat substantia nigra. One of the labeled dendrites (D1) shows only intracellular immunogold particles whereas the other (D2) is predominantly labeled on its plasma membrane (white arrowheads). Inset: Note the presence of NTS1 proteins in association with the outer membrane of an endosomal vesicle (white arrowheads) and of an endocytic profile (black arrowheads). ma; myelinated axon. Scale bars = 0.5 ìm and 0.2 ìm (inset). Reproduced from Boudin H, Pélaprat D, Rosténe W, Pickel VM, and Beaudet A (1998) Correlative ultrastructural distribution of neurotensin receptor proteins and binding sites in the rat substantia nigra. *Journal of Neuroscience* 18(20): 8473–884. Copyright (1998) by the Society for Neuroscience.

dendrites showed a higher concentration of intracellular receptors, suggesting that neurons expressing NTS1 receptors maintain large receptor reserves (Figure 6). An important fraction of these intracellular receptors are associated with the Golgi apparatus and endoplasmic reticulum, indicating that newly synthesized receptor proteins are in the process of synthesis and posttranslational modifications. Electron microscopy immunocytochemistry also revealed NTS1 receptors in endosomal vesicles, suggesting that NTS1 receptors internalize on ligand stimulation.

NTS2 Receptors

Early autoradiographic-binding studies, based on the displacement of specific [^{3}H]NT, [^{125}I]NT, or [^{3}H] SR142948A (a nonpeptidic NT antagonist) binding by levocabastine or on the direct labeling of NTS2 receptors using [^{3}H]-levocabastine as radioligand, reported a widespread distribution of levocabastine-sensitive NTS2 receptor sites in adult rat brain. In conformity with the results obtained by autoradiographic approaches, *in situ* hybridization histochemistry revealed that mRNA encoding NTS2 was diffusely distributed throughout the brain. High NTS2 expression levels were detected in the olfactory system, cerebral and cerebellar cortices, hippocampal formation, and selective hypothalamic nuclei. A striking feature of NTS2 distribution was its association with every single sensory system in the brain. Moderate to dense hybridization signal and NTS2 immunoreactivity were observed in a variety of olfactory (e.g., olfactory bulb, anterior olfactory nucleus, and piriform cortex), visual (e.g., suprachiasmatic nucleus, superior colliculi, lateral geniculate body), auditory (e.g., superior olivary nuclei, cochlear nuclei, and temporal cortex), and gustatory cerebral structures (e.g., nucleus of the solitary tract), suggesting that the NTS2 might be involved in the modulation of primary afferent pathways (Figure 7). Of particular interest from a functional perspective is the intense labeling of brain stem structures involved in pain control. Indeed, NTS2 mRNA and receptor proteins were highly expressed in brain stem structures implicated in the descending control of nociceptive input, such as the periaqueductal gray, dorsal raphe nucleus, raphe magnus and pallidus, gigantocellular reticular nucleus, pars alpha, and lateral paragigantocellular nucleus. The strong NTS2 immunolabeling also found in the dorsal root ganglia and in the superficial layers of the dorsal horn of the spinal cord suggests that NTS2 receptors are ideally poised to regulate NT's spinal antinociceptive actions (Figure 8).

Biochemical, *in situ* hybridization, receptor autoradiographic, and immunohistochemical studies all concur in demonstrating that in the rodent brain, the developmental pattern of NTS2 receptors is markedly different from that reported for the NTS1s. Indeed, NTS2 receptors are absent at birth, appear late during development, and do not reach their maximal levels until adulthood. In contrast to NTS2, NTS1 is expressed before birth, peaks at

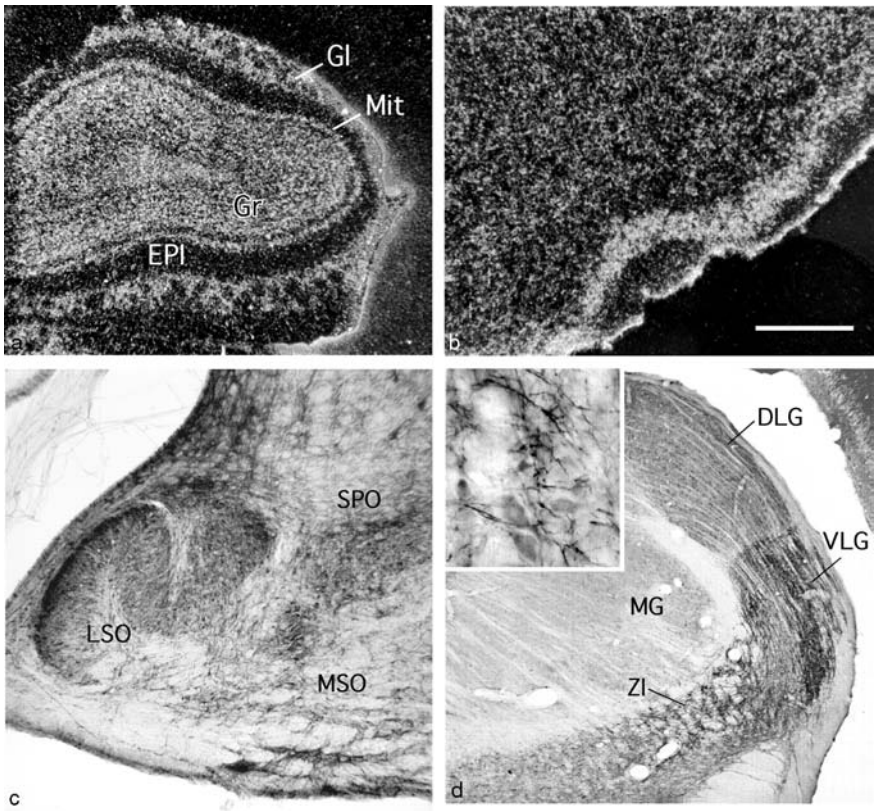

Figure 7 NTS2 receptor mRNA (a, b) and proteins (c, d), as detected by immunohistochemistry and *in situ* hybridization, respectively, in the olfactory (a, b), auditory (c), and visual (d) sensory systems. Dark-field views of sections of the olfactory bulb (a) and piriform cortex (b). Within the olfactory bulb, intense hybridization is apparent over mitral and granular cells whereas only weak labeling is visible over plexiform and glomeruli layers. In the piriform cortex, silver grains were almost exclusively confined to layer II. Distribution of NTS2 immunoreactivity in the superior olivary complex (c) and lateral geniculate nucleus (d). In the rostral pons, the superior olivary complex exhibits a mix of labeled somata and processes. Intense labeling is observed over the dorsal and ventral tiers of the lateral geniculate nucleus as well as in the adjoining zona incerta. As shown in the inset, the NTS2 staining accounts for dendritic processes. DLG, dorsal lateral geniculate nucleus; EPl, external plexiform layer; Gl, glomeruli layer; Gr, granule cell layer; LSO, lateral superior olive; MG, medial geniculate nucleus; Mit, mitral cell layer; MSO, medial superior olive; SPO, superior paraolivary nucleus; VLG, ventral lateral geniculate nucleus; Zl, zona incerta. Scale bars = 1 mm (a, b, d); 300 μm (c); 80 μm inset. (a, b) Reprinted from Sarret P, Beaudet A, Vincent JP, and Mazella J (1998) Regional and cellular distribution of low affinity neurotensin receptor mRNA in adult and developing mouse brain. *Journal of Comparative Neurology* 394(3): 344–356, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Ltd. (c, d) Reprinted from Sarret P, Perron A, Stroh T, and Beaudet A (2003) Immunohistochemical distribution of NTS2 neurotensin receptors in the rat central nervous. *Journal of Comparative Neurology*. 461(4): 520–538. with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Ltd.

around postnatal day 10, and progressively decreases thereafter, reaching adult levels between 30 and 40 days. On the basis of their ubiquitous distribution, as well as their pattern of ontogenic development, NTS2 receptors were originally surmised to be mainly associated with glial cells. Accordingly, NTS2 binding and/or mRNA were detected in association with astrocytes both *in vitro* and *in vivo*. This NTS2 glial expression was low under basal conditions but was massively upregulated in reactive astrocytes surrounding a cortical stab wound, suggesting that NTS2 receptors may play a role in the regulation of inflammatory responses and neuronal regeneration. Despite its expression in glial cells, the cellular localization of NTS2 is predominantly neuronal. However, unlike NTS1 receptors, which were reportedly distributed over perikarya, dendrites, and axons of neurons, NTS2 receptors were predominantly, although not exclusively, associated with dendritic processes.

NTS3 Receptors

Both NTS3/sortilin mRNA and immunoreactive proteins displayed a broad distribution throughout the brain. High levels of NTS3/sortilin expression and immunoreactivity were found in neuronal cell bodies and dendrites of cortical areas such as the piriform cortex and hippocampus, the islands of Calleja, medial and lateral septal nuclei, amygdaloid nuclei, thalamic nuclei, the substantia nigra, and the Purkinje cell layer of the cerebellar cortex

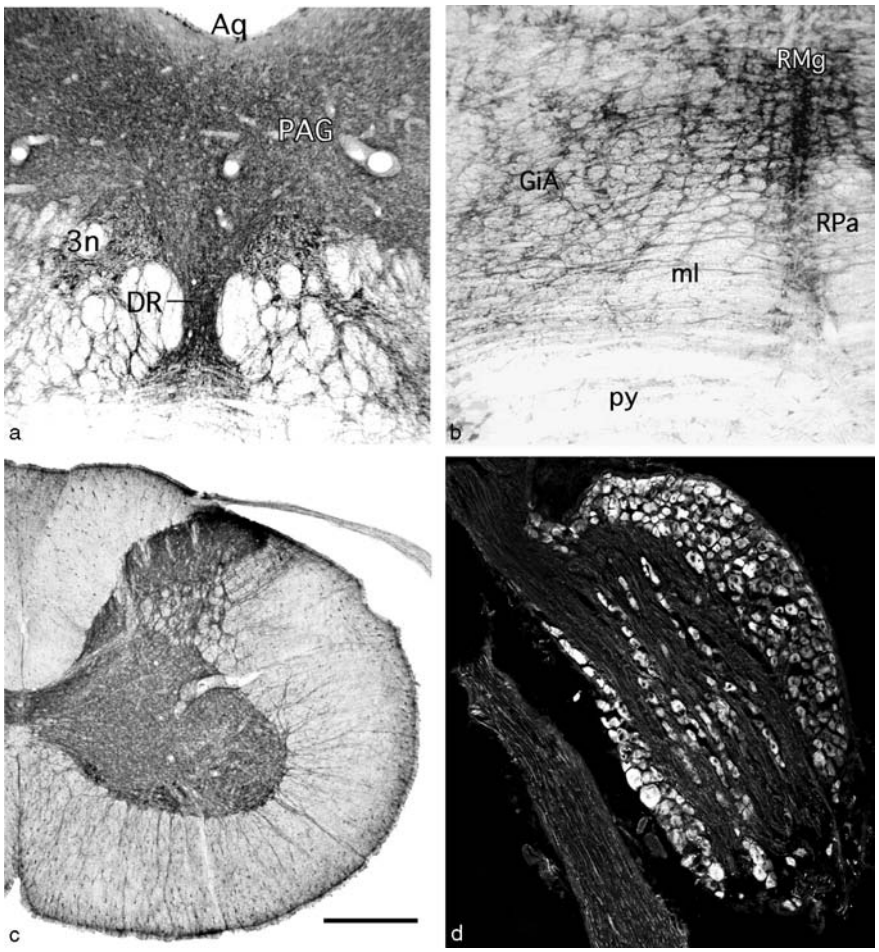

Figure 8 Distribution of levocabastine-sensitive NTS2 receptor immunoreactivity in spinal and supraspinal cerebral structures. (a) The ventral periaqueductal gray (PAG) and dorsal raphe nucleus (DR) both exhibit a dense network of immunopositive processes. (b) The gigantocellular reticular nucleus (GiA), pars alpha, exhibits numerous labeled perikarya as well as processes. Medially, the nucleus raphe magnus (RMg) is strongly stained, and the raphe pallidus (RPa) displays moderate immunoreactivity. (c) Localization of NTS2 receptors in lumbar spinal cord. Immunoperoxidase staining reveals the presence of intense NTS2 immunoreactivity within the superficial layers of the dorsal horn. A dense plexus of NTS2-immunoreactive neuronal processes is observed in laminae I and II. (d) Cellular distribution of NTS2 receptors in primary afferent neurons. Strong NTS2 labeling is evident in subpopulations of small and large dorsal root ganglion cells. Aq, aqueduct; 3n, oculomotor nucleus; py, pyramidal tract; ml, medial lemniscus. Scale bars = 500 μm (a), 350 μm (b), 600 μm (c), 200 μm (d). (a, b) Reprinted from Sarret P, Perron A, Stroh T, and Beaudet A (2003) Immunohistochemical distribution NTS2 neurotension receptors in the rat central nervous system. *Journal of Comparative Neurology* 461(4): 520–536, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Ltd. (c,d) Reprinted from Sarret P, Esdaile MJ, Perron A, Martinez J, Stroh T, and Beaudet A (2005) Potent spinal analgesia elicited through stimulation of NTS2 neurotensin receptors. *Journal of Neuroscience* 25(36): 8188–8196.

(Figure 9). Furthermore, in the brain stem, all cranial nerve motor nuclei were strongly labeled. However, even if NTS3 mRNA and immunoreactivity patterns were mainly associated with neuronal elements, NTS3 receptors were also found in association with glial cells, corresponding, in morphology, to oligodendrocytes.

Many of the regions enriched in NTS3 mRNA and immunoreactive neurons had previously been documented as highly innervated by neurotensinergic axon terminals, thereby supporting the concept that NTS3 may act as a *bona fide* NT receptor. As observed for the NTS2 receptor, the distribution of NTS3/sortilin did, however, largely exceed that of NT terminal fields. Several regions previously described as NT-negative, such as the hippocampal CA fields, the cerebral cortex, and the cerebellum, displayed very high levels of NTS2 and NTS3 mRNA and protein, suggesting that NT may not be the exclusive endogenous ligand to act on NTS2 and NTS3 receptors. Moreover, the superimposition of NTS3/sortilin with NTS1 and NTS2 receptor subtypes suggests that it might interact with these receptors to mediate central NT effects. At the electron microscopic level,

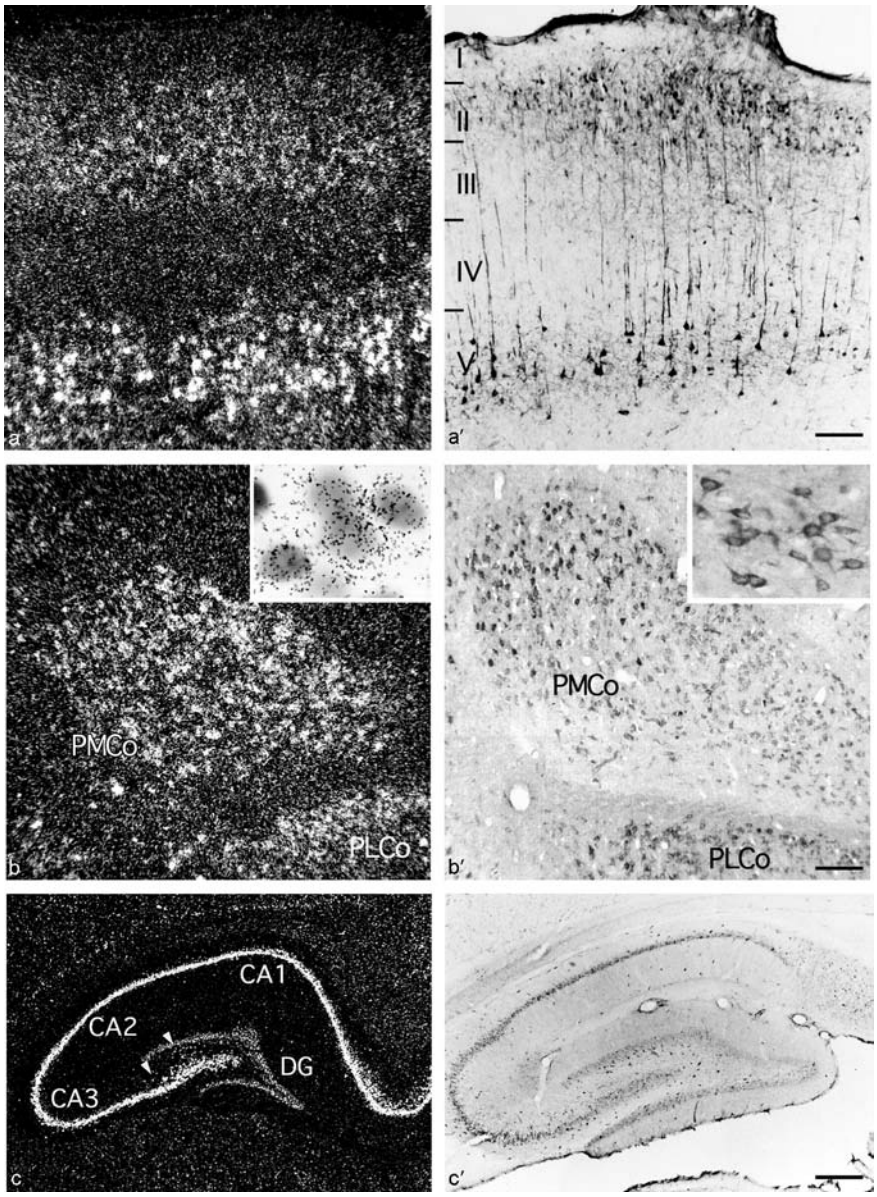

Figure 9 NTS3/sortilin *in situ* hybridization (a–c) and immunoreactivity (a′–c′) in the parietal cortex (a, a′), amygdaloid complex (b, b′), and hippocampal formation (c, c′). In the parietal cortex, dark-field micrograph of NTS3/sortilin antisense probe-hybridized section and immunocytochemical analysis reveal a large number of cell bodies in layers II/III and V. In the amygdaloid complex, sortilin immunoreactivity displays a distribution identical to the pattern of mRNA expression in both nuclei (PMCO, posteromedial cortical amydaloid nucleus; PLCo, posterolateral cortical amydaloid nucleus). At higher magnification (inset), the labeling is clearly over large neuronal cell bodies in the PMCo. In the hippocampal formation, NTS3 mRNA expression and immunoreactivity are both very high and selective. All subfields of Ammon's horn (CA1-3) display extremely high levels of hybridization signal. Some labeled cells are also evident in the hilus of the dentate gyrus (DG, arrowheads), whereas the DG granule cell layer itself is only faintly labeled. Scale bars = 200 μm (a, a′), 150 μm (b, b′), 60 μm (c, c′), 12 μm (inset to b), 40 μm (inset to b′). Reprinted from Sarret P, Krzywkowski P, Segal L, Nielsen MS, Petersen CM, and Mazella J (2003) Distribution of NTS3 receptor/sortilin mRNA and protein in the rat central nervous system. *Journal of Comparative Neurology* 461(4): 483–505, with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Ltd.

NTS3 was predominantly concentrated over intracytoplasmic membrane-bound organelles, such as the Golgi apparatus, saccules of endoplasmic reticulum, and vesicles. This subcellular distribution conforms to the purported role of this protein in the sorting and trafficking of intracellular ligands.

NT/Dopamine Interactions

It has long been recognized that centrally administered NT exerts a number of effects similar to those of neuroleptic drugs. This observation led to a considerable number of studies aimed at defining

the interactions that might exist between NT and DA systems in the brain. There is at present unquestionable neuroanatomical, electrophysiological, neurochemical, and behavioral evidence that NT modulates DA transmission in the nigrostriatal and mesocorticolimbic pathways. Actually, NT behaves either as a psychostimulant or as a neuroleptic, depending on what part of the mesocorticolimbic DA circuitry is targeted by the peptide. *In vitro* and *in vivo* application of NT to DA neurons of the VTA stimulates DA neuronal activity. In contrast, administration of NT in the nucleus accumbens, a projection area of the VTA, inhibits DA transmission. The stimulatory, psychostimulant-like effects are mediated through NTS1 receptors that are abundantly expressed by VTA DA neuronal cell bodies and dendrites. The inhibitory, neuroleptic-like effects are also mediated via NTS1 receptors located postsynaptically to DA nerve terminals in the VTA. In addition, both antipsychotic and psychoactive drugs modify NT expression and transmission in these circuits, suggesting that the peptide mediates the effects of both types of drugs. Recent studies with either the NTS1 antagonist SR 48692 or with NT and NTS1 knockout mice further support the notion that NT participates, via the NTS1, in the response to psychostimulants as well as to that of both typical and atypical neuroleptic drugs.

Since clinical and postmortem data indicate that central NT systems are altered in schizophrenic patients, there is an ongoing debate to know whether an NTS1 agonist or antagonist would be accurate as an antipsychotic drug in the treatment of schizophrenia. Recent evidence points to NT agonists as offering promising leads for developing an active neuroleptic-like drug. Thus, SR48692 failed in clinical trials to reveal antipsychotic properties whereas NT agonist mimetics, crossing the blood–brain barrier, proved to exert neuroleptic-like effects in the rat following peripheral administration. On the other hand, there are reasons to believe that an NT antagonist might find useful application in treating drug addicts since amphetamine and cocaine sensitization are attenuated by SR48692 and SR142948A in the rat.

NT and Pain Modulation

Considerable breakthroughs have been made in the past 5 years in understanding the role of NT in pain responses. Originally, intrathecal and intracerebroventricular administrations of NT were shown to produce a profound analgesia in a variety of pain paradigms, including hotplate, tail-flick, and writhing tests. It is important to note that NT's antinociceptive effects were also reported to be naloxone insensitive, implying that they were not exerted through endogenous opioid systems. However, site-specific microinjection experiments have provided evidence that NT may actually induce hyperalgesia instead of analgesia following administration in the periaqueductal gray and rostroventral medulla (RVM) and brain regions projecting to these structures. Indeed, studies performed using lower, close-to-physiological doses suggest that NT produces facilitatory effects on nociception. The biphasic effect of NT may be explained by a dose-dependent activation of distinct populations of pain modulatory neurons or by the recruitment of different receptor subtypes. Nevertheless, the recent demonstration that NT knockout mice display defects in stress-induced analgesia may also suggest that the transition from pain facilitation to analgesia involves stress-induced increases in NT signaling in pain modulatory regions.

Both NTS1 and NTS2 appear to be required for different aspects of NT-induced analgesia. Knockdown of the NTS1 receptor, using either knockout mice or peptidic nucleic acids, as an antisense strategy, revealed that NT fails to increase response latencies in the hotplate test. Furthermore, intrathecal or intra-RVM administration of the relatively selective NTS1 antagonist SR48692 also partly abolished NT-induced analgesia in the tail-flick assay. Although a number of studies have implicated NTS1 in NT modulation of pain responses, several lines of evidence point to the levocabastine-sensitive low-affinity NTS2 subtype as a major antinociceptor. Indeed, the NT antagonist SR142948A, which recognizes both high- and low-affinity NT receptors, blocked NT-induced analgesia, whereas the NTS1 antagonist SR48692 did not. Accordingly, NTS2-preferring NT analogs were considerably more potent in eliciting antinociception than were NTS1-preferring analogs. The most convincing evidence comes from antisense inhibition experiments and NTS2-deficient mice, in which NT-induced analgesia was markedly inhibited. The future development of tissue-specific knockdown of both NTS1 and NTS2 will help determine their respective role in NT pain modulation. Further research to design NTS1- and NTS2-selective novel analogs that cross the blood–brain barrier may also offer new avenues for the treatment of pain without the side effects associated with the use of opioids.

Feeding Behavior and Obesity

It is well documented that food intake and body weight are regulated by a number of hypothalamic orexigenic and anorexigenic neuropeptides. These peptides are thought to coordinately regulate body weight through a complex, partially elucidated

mechanism orchestrated in part by leptin, an adipose tissue-released hormone that acts centrally in the hypothalamus to control body weight and energy expenditure. The anorexigenic effect of centrally administered NT has been known for 20 years. Further support for a role of NT in food intake came from the observation that hypothalamic expression levels of NT are decreased in genetic models of obese rats and mice including the Zucker fa/fa rat and the ob/ob mouse that are deficient in leptin receptor and leptin, respectively. More recently, a number of reports pointed to NT as a possible mediator of leptin-induced reduction in food intake. It has been shown that leptin receptors are present on and activate NT-expressing neurons in the arcuate nucleus and the dorsomedial area of the rat hypothalamus. Furthermore, central administration of leptin for 3 days increases NT gene expression whereas chronic intraperitoneal administration for 7 days decreases NT levels in the lateral hypothalamus. Leptin and NT reciprocally potentiate each other's inhibitory effect on food intake. Immunoneutralization of central NT with an NT antiserum or administration of the NT antagonist SR48692 completely blocks the inhibitory effect of leptin on food intake in food-deprived rats. Finally, in contrast to NT knockout mice that did not differ in their body weight from their wild-type littermates, NTS1 knockout mice exhibited increased food consumption, body weight, and white adipose tissue mass compared with wild-type littermates. Together, these data strongly suggest that NT is a major link in the complex neuropeptide network that controls body weight and that the peptide effects on feeding are exerted through the NTS1 receptor. NT agonists could therefore be useful in the treatment of obesity. In this context, it is of interest that an NT analog that crosses the blood–brain barrier was found to reduce food intake and body weight following intraperitoneal administration in normal and genetically obese Zucker fa/fa rats.

See also: Dopamine; Mammalian Neuropeptide Families; Neuropeptide Release; Neuropeptide Synthesis and Storage; Neuropeptides: Discovery; Neuropeptides: Enteric Nervous System; Neuropeptides: Food Intake; Neuropeptides: Pain.

Further Reading

Beck B (2000) Neuropeptides and obesity. *Nutrition* 16: 916–923.

Binder EB, Kinkead B, Owens MJ, and Nemeroff CB (2001) Neurotensin and dopamine interactions. *Pharmacological Reviews* 53(4): 453–486.

Boudin H, Pélaprat D, Rostène W, and Beaudet A (1996) Cellular distribution of neurotensin receptors in rat brain: Immunohistochemical study using an antipeptide antibody against the cloned high affinity receptor. *Journal of Comparative Neurology* 373(1): 76–89.

Boudin H, Pélaprat D, Rostène W, Pickel VM, and Beaudet A (1998) Correlative ultrastructural distribution of neurotensin receptor proteins and binding sites in the rat substantia nigra. *Journal of Neuroscience* 18(20): 8473–884.

Boules M, Fredrickson P, and Richelson E (2003) Current topics: Brain penetrating neurotensin analog. *Life Sciences* 73: 2785–2792.

Brunetti L, DiNisio C, Orlando G, Ferrante C, and Vacca M (2005) The regulation of feeding: A cross talk between peripheral and central signalling. *International Journal of Immunopathology and Pharmacology* 18: 201–212.

Caceda R, Kinkead B, and Nemeroff CB (2003) Do neurotensin receptor agonists represent a novel class of antipsychotic drugs? *Seminars in Clinical Neuropsychiatry* 8: 94–108.

Dobner PR (2005) Multitasking with neurotensin in the central nervous system. *Cellular and Molecular Life Sciences* 62: 1946–1963.

Dobner PR, Deutch AY, and Fadel J (2003) Neurotensin: Dual roles in psychostimulant and antipsychotic drug responses. *Life Sciences* 73: 801–811.

Emson PC, Goedert M, and Mantyh PW (1985) Neurotensin-containing neurons. In: Björklund A and Hökfelt T (eds.) *Handbook of Chemical Neuroanatomy, Vol. 4: GABA and Neuropeptides in the CNS, Part I*, pp. 355–405. Amsterdam: Elsevier.

Emson PC, Horsfield PM, Goedert M, Rossor MN, and Hawkes CH (1985) Neurotensin in human brain: Regional distribution and effects of neurological illness. *Brain Research* 347(2): 239–244.

Kinkead B and Nemeroff CB (2004) Neurotensin, schizophrenia, and antipsychotic drug action. *International Review of Neurobiology* 59: 327–349.

Kitabgi P (2002) Targeting neurotensin receptors with agonists and antagonists for therapeutic purposes. *Current Opinion in Drug Discovery & Development* 5: 764–776.

Kitabgi P (2006) Inactivation of neurotensin and neuromedin N by Zn metallopeptidases. *Peptides* 27(10): 2515–2522.

Kitabgi P (2006) Prohormone convertases differentially process pro-neurotensin/neuromedin N in tissues and cell lines. *Journal of Molecular Medicine* 84(8): 628–634.

Rostene WH and Alexander MJ (1997) Neurotensin and neuroendocrine regulation. *Frontiers in Neuroendocrinology* 18(2): 115–173.

Sarret P, Beaudet A, Vincent Jp, and Mazelle J (1998) Regional and cellular distribution of low affinity neurotensin receptor mRNA in adult and developing mouse brain. *Journal of Comparative Neurology* 394(3): 344–356.

Sarret P, Esdaile MJ, Perron A, Martinez J, Stroh T, and Beaudet A (2005) Potent spinal analgesia elicited through stimulation of NTS2 neurotensin receptors. *Journal of Neuroscience* 25(36): 8188–8196.

Sarret P and Beaudet A (2003) Neurotensin receptors in the central nervous system. In: *Handbook of Chemical Neuroanatomy, Vol. 20: Peptide Receptors, Part II*, pp. 323–364. Amsterdam: Elsevier.

Sarret P, Krzywkowski P, Segal L, Nielsen MS, Petersen CM, and Mazella J (2003) Distribution of NTS3 receptor/sortilin mRNA and protein in the rat central nervous system. *Journal of Comparative Neurology* 461(4): 483–505.

Sarret P, Perron A, Stroh T, and Beaudet A (2003) Immunohistochemical distribution of NTS2 neurotensin receptors in the rat central nervous system. *Journal of Comparative Neurology* 461(4): 520–538.

Uhl GR (1982) Distribution of neurotensin and its receptor in the central nervous system. *Annals of the New York Academy of Sciences* 400: 132–149.

Vincent JP, Mazella J, and Kitabgi P (1999) Neurotensin and neurotensin receptors. *Trends in Pharmacological Sciences* 20: 302–309.

Neurotoxins

R M Kostrzewa, East Tennessee State University, Johnson City, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The noted physiologist, Walter B Canon, promoted the precept that the function of a nerve could be determined by cataloging the absence-deficits when that nerve was removed. In the first half of the twentieth century that removal entailed either axonal transection or chemical or electrolytic ablation of the nucleus of origin (perikaryon). This experimental approach was common at the time, and it proved valuable in uncovering the functions subserved by particular nerve tracts. Nevertheless, there were obvious limitations to that experimental approach:

1. Not all nerve tracts are amenable to surgical transection or ablation, because of their physical location or axonal length (which might be measured in micrometers). In the brain there were only a small number of defined nerve tracts in that era, so the approach was impractical in the central nervous system (NS) (i.e., one must recognize the presence of a tract in order to be able to transect it.).
2. Particularly in spinal cord and brain, defined transection of a single tract was virtually impossible, because of the multiplicity of coordinate and crossing nerve bundles and other tracts. Inadvertently, many tracts would be transected.
3. A nerve tract often includes several different neuronal phenotypes, so that even a defined transection routinely would equate with the ablation of many different kinds of nerves within that tract.
4. Finally, as is obvious from the above-noted items, surgical and electrolytic lesions cannot specifically remove a single neuronal phenotype (i.e., cholinergic nerves only).

Selective neurotoxins provide a means to circumvent all of the above-noted limitations of a surgical approach. There are many examples of selective neurotoxins capably destroying a specific neuronal phenotype deep in the brain (or elsewhere), while leaving the surrounding neuropil virtually intact. This article is intended to (1) review the nature of selective neurotoxins, (2) describe the multiple mechanisms by which they might act, (3) discuss the practicality of using neurotoxins and the possible outcomes, (4) briefly touch on the breadth and scope of neurotoxins, and (5) assign limitations to neurotoxins in experimental investigations.

Historical Background

Selective neurotoxins were first discovered in the last half of the twentieth century. In the 1950s, Rita Levi-Montalcini, in her work with Victor Hamburger and Stanley Cohen, initiated the experiments that defined, explored, and uncovered the nature of neurotrophins – showing that the substance nerve growth factor (NGF), in particular, was necessary for development, growth, and sustenance of sympathetic and sensory ganglia. In an elegant series of studies this group ultimately developed an antibody to NGF and demonstrated that sympathetic nerves in whole animals could be destroyed by anti-NGF, leaving an animal 'immunosympathectomized.' This represented the first highly selective neurotoxin. Botulinum neurotoxin (BoNT) was long known as a neurotoxin, but its mechanism of action was realized only in the past 50–60 years, and its usage in medicine and in experimental physiology took hold only recently. BoNT is now regarded as a selective neurotoxin, but it was categorized as a poison (nonselective) in history. New selective neurotoxins were discovered in the late 1960s, and even more afterwards – with known neurotoxins still growing in number.

Criterion for Definition as a Selective Neurotoxin

The definition of 'selective neurotoxin' is a moving target. Anti-NGF was selective for sensory and sympathetic nerves, which represent different neuronal phenotypes. Anti-NGF is still regarded as a selective neurotoxin. However, for virtually all other neurotoxins there is a loose requirement that they have phenotypic or receptor selectivity. For example, 6-hydroxydopamine (6-OHDA) destroys catecholaminergic nerves (i.e., dopaminergic, noradrenergic, and potential for adrenergic) with relative selectivity. 1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) and its metabolite 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridinium ion (MPP^+) are selective for dopaminergic nerves (see Figure 1). Substituted amphetamines (AMPHs) are relatively selective for monoaminergic nerves (i.e., dopaminergic and serotoninergic). Excitatory amino acids have selectivity for subtypes of glutamate receptors, while there is no phenotypic selectivity.

Therefore, the term 'selective neurotoxin' is reserved for a chemical that has relative but not necessarily absolute selectivity for either a specific neuronal phenotype, or for a population of phenotypically diverse nerves with a particular type of receptor. Although

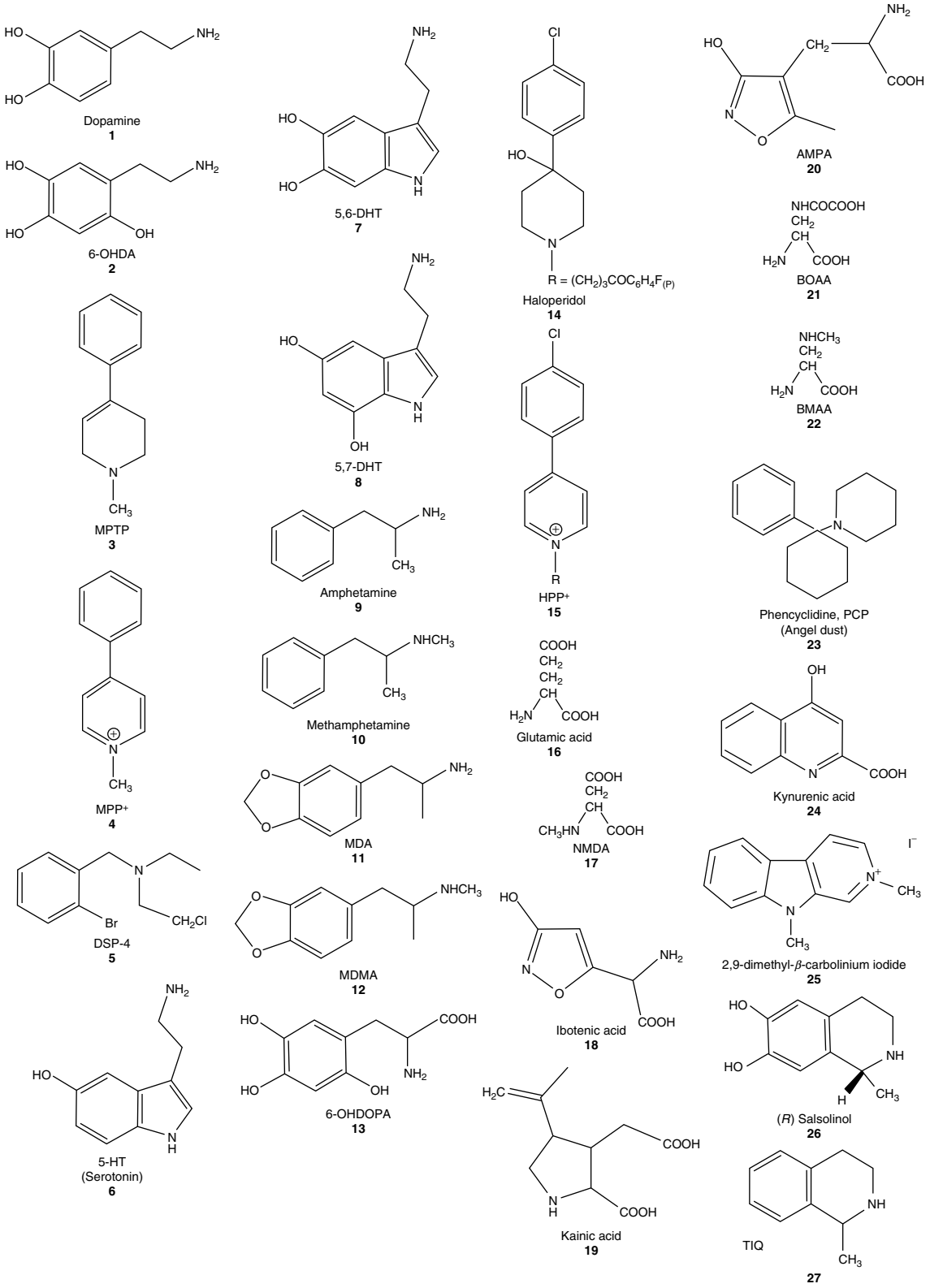

Figure 1 9w?>Structures of neurotransmitters/neurotoxins. 6-OHDA, 6-hydroxydopamine; MPTP, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine; MPP^+, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridinium ion; DSP-4, *N*-2-chloroethyl-*N*-ethyl-2-bromobenzylamine; 5-HT (serotonin), 5-hydroxytryptamine; 5,6- and 5,7-DHT, 5,6- and 5,7-dihydroxytryptamine; MDA, methylenedioxyamphetamine; MDMA, 3,4-methylenedioxymethamphetamine (ecstasy); 6-OHDOPA, 6-hydroxydopa; HPP^+, haloperidol metabolite; NMDA, *N*-methyl-D-aspartate; AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; BOAA, β-*N*-oxalylamino-L-alanine; BMAA, β-*N*-methylamino-L-alanine; TIQ, 1,2,3,4-tetrahydroisoquinoline.

there are cytoplasmic poisons (cyanide, carbon monoxide), heavy metals, and noxious substances that will destroy nerves and other cells in general, by virtue of their nonspecificity they generally are not classified as selective neurotoxins, although rotenone (discussed later) is an exception.

Nature of Neurotoxin Selectivity

Only by examples can one perceive the nature of a selective neurotoxin, and only with this background can generalizations be made. Several neurotoxins are selectivity because of their affinity to the transporter for the neurotransmitter in question. For example, 6-OHDA has high affinity to the dopamine transporter (DAT) and norepinephrine transporter (NET) of dopaminergic and noradrenergic nerves, respectively. As such, 6-OHDA is accumulated in high amount only within dopaminergic and noradrenergic nerves, attaining high cytoplasmic concentration, sufficient to inactivate vital enzymes or cytochromes and thus able to initiate necrotic or apoptotic processes. Several other neurotoxins act in much the same manner as 6-OHDA: MPP^+ has high affinity for the DAT; 5,6- and 5,7-dihydroxytryptamine (5,6-DHT; 5,7-DHT) have high affinity for the serotonin (5-HT) transporter (SERT); and substituted AMPHs have high affinity for DAT, NET and SERT. These respective neurotoxins act primarily in the nerves so associated.

Receptor-selective neurotoxins include glutamate *per se* and aspartate *per se*, which act at multiple glutamate receptors, but nonetheless produce overexcitation (i.e., excitotoxicity) of the postsynaptic nerves on which the receptors reside. *N*-Methyl-D-aspartate (NMDA), quinolinate, and ibotenate produce excitotoxicity via NMDA receptors; kainate and domoic acid are excitotoxins acting at the kainate subtype of glutamate receptor; α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), as well as the unnatural amino acid *β-N*-oxalylamino-L-alanine (BOAA), act at the AMPA receptor; the unnatural amino acid *β-N*-methylamino-L-alanine (BMAA) acts at metabotropic group I receptors. Notably, many of the glutamate receptor agonists, while being neurotoxic because of their action at one particular receptor subtype, act at multiple glutamate receptor subtypes. In this instance, usage defines a 'selective neurotoxin.' In other words, there is an experimental application whereby an excitotoxin fulfilled the criterion needed to obtain the experimental goal.

Clostridial neurotoxins (e.g., botulinum neurotoxin and tetanus toxin) are so-called Trojan horse toxins. The C-terminal end of the heavy chain of BoNT binds to a specific protein on cholinergic nerves; next, BoNT is translocated inside the cholinergic nerve via the N-terminal domain; and inside the nerve ending, the light chain of BoNT catalytically cleaves proteins involved in transmitter exocytosis (i.e., V-SNARE and T-SNARE proteins). Cholinergic nerves thus become functionally inactive for a long period (weeks to months) – but the nerves are not overtly destroyed. This usage thus expands the definition of a neurotoxin to extend to long-term inactivation, not limiting 'neurotoxicity' to an ability to necessarily destroy a nerve.

Mechanisms of Neurotoxin Action

Neurotoxins tend to impede vital cytoplasmic processes, and this can be accomplished by an action at one of many different sites. Identified neurotoxins are known to (1) disrupt energy production, (2) dysregulate ionic gradients, (3) impair neurotransmitter synthesis/storage/release, (4) interfere with transcription, translation, and formation of requisite cytochromes, enzymes, and other proteins, (5) promote formation of abnormal proteins or other products, or (6) eliminate the necessary neurotrophic stimulation (e.g., anti-NGF). Other sites can also be affected. A number of these points highlighted here are outlined and discussed, in part, in the following sections.

Oxidative Inactivation of Vital Cellular Elements

Some neurotoxins act through the inactivation of a vital cellular element or series of elements, which include enzymes, proteins, lipids, cytochromes, and DNA. For example, 6-OHDA and MPP^+ act on mitochondrial cytochromes and ultimately impair energy (ATP) production. In the process, the mitochondrial transition pore is opened, and apoptosis is activated. However, if the damage is sufficiently severe, necrosis can occur before apoptosis or necroptosis runs its course.

In the preceding scenario, the neurotoxin action may be direct or indirect. For example, MPP^+ *per se* potently inhibits mitochondrial cytochrome I. By contrast, 6-OHDA autooxidizes to quinones and semiquinones, each of which can mediate the 6-OHDA effect via their actions at sites on the mitochondria and other cellular elements. Additionally, the quinones and semiquinones promote the formation of peroxides, superoxide radical, and hydroxyl radical, each of which can act at multiple cytoplasmic sites to inactivate vital cellular elements.

The neurotoxins 5,6- and 5,7-DHT act in much the same manner as 6-OHDA, except that 5,6- and 5,7-DHT act primarily in serotoninergic nerves (owing to their high affinity to SERT), while 6-OHDA acts primarily in catecholaminergic nerves (owing to its high affinity to DAT and NET).

Alkylation of Vital Cellular Elements

The 2-chloroethylamine neurotoxins *N*-2-chloroethyl-*N*-ethyl-2-bromobenzylamine (DSP-4) and xylamine have high affinity to the NET, and are thus accumulated in high amount in NE nerve terminals. Inside the nerve ending, DSP-4 randomly alkylates reactive sites, the most notable being unsaturated lipids, sulfur moieties on proteins, including mitochondrial cytochromes. Consequently, the nerve is functionally inactivated for a long duration (months to years), as reflected by long-term reduction of the neurotransmitter NE (also by long-term reduction in the NET-synthesizing enzyme, dopamine β-hydroxylase). Many other undefined sites would also be alkylated and inactivated inside the nerve.

Ethylcholine aziridinium ion (AF64A), having high affinity for the choline transporter, is thus accumulated by cholinergic nerves. In addition to covalently bonding to the choline transporter, once inside the nerve AF64A interacts with reactive molecular sites, like DSP-4, and produces functional inactivation and neurotoxicity to the cholinergic nerve. AF64A binds to and inactivates choline acetyltransferase, choline kinase, choline dehydrogenase, and acetylcholine esterase (AChE). AF64A-generated reactive oxygen species produce oxidative stress; and alkylation by AF64A disrupts the integrity of DNA, gene transcription, and RNA transcription. Initiation of apoptosis leads to cell death.

Trojan Horse Neurotoxins

As indicated earlier, BoNT and tetanus toxin are translocated into the nerves, at which time the toxins are bound. Inside the nerve, these Trojan horse neurotoxins catalytically cleave those proteins essential for exocytosis and thereby produce long-term inactivation of the nerve. In one sense, other neurotoxins that are carried into nerves can be classified as Trojan horse neurotoxins. For example, it can be said that 6-OHDA masquerades as DA and is carried into the DA terminal via the DAT; once in the nerve, 6-OHDA disrupts cellular function and destroys the cell. MPTP/MPP^+, substituted AMPHs, 5,6- and 5,7-DHT, and a number of other neurotoxins would act similarly. However, 6-OHDA is distinctly different molecularly from DA, while BoNT carries its own translocation device (protein), making its own 'door' into the cell. So, a Trojan horse neurotoxin can take different forms, having different categorizations.

Potential for Trojan horse neurotoxins By re-engineering the BoNT molecule, it becomes possible to target proteins on specific types of nerves. A re-engineered BoNT could target only sensory neurons, for example, and produce long-term analgesia in people with chronic pain. Alternatively, even nonneuronal cells (e.g., epithelial cells) can be targeted, and the endopeptidase portion of the molecule can be replaced with a recombinant protein (e.g., epidermal growth factor) to alter release of cellular products (e.g., mucin). A BoNT molecule reformulated with epidermal growth factor (or analog) might impair the exocytosis of mucin. Re-engineered BoNT molecules could be a useful means, and a new drug approach, for the treatment of human disorders such as cystic fibrosis, asthma, and chronic obstructive pulmonary disease. Alternatively, a noxious drug could be substituted for the catalytic domain of BoNT, coupled with an altered binding domain that might target a site on a cancerous cell, to derive an anticancer drug with high potency and little adverse effect. Many other permutations for the BoNT or tetanus toxin molecule are possible.

Neurotoxicity Produced by Excitotoxicity

Agonists at ionotropic glutamate receptors potentially are capable of overstimulation of a postsynaptic neuron. The neurotransmitters glutamate and aspartate are imbued with this potential, as are agonists with preferential activity at kainate receptors (kainate, domoic acid), AMPA receptors (AMPA, kainate, quisqualate, BOAA), or NMDA receptors (endogenous quinolinate, NMDA, ibotenate, BMAA). Excitotoxicity is accompanied by increased Na^+ ion fluxes at each of these receptor subtypes as well as increased Ca^{2+} ion flux at the NMDA receptor. In low amount, the amino acid BMAA additionally potentiates neuronal injury produced by other insults.

Interference with Neurotrophin Stimulation or Ontogenetic Nerve Activity

Anti-NGF, discussed earlier, is a targeted antibody that inactivates NGF, a neurotrophin required for ontogenetic development of sympathetic ganglia. In mice treated with NGF, sympathetic nerves, deriving from sympathetic paraganglia, fail to develop and remain absent throughout life.

During development redundant and unsuccessful neurons undergo apoptosis when lacking adequate physiologic contact and excitation from neuronal cross-talk. Thus, during ontogeny and particularly during the brain growth spurt, substances that interfere with or block neuronal cross-talk are apt to produce neuronal apoptosis. Normally cytoprotective NMDA receptor blockers (in adulthood) become neurotoxins when administered during ontogeny, by blocking neuronal cross-talk and thereby signaling that the soon-to-be apoptotic nerve is quiescent and

therefore 'redundant.' In ontogeny, NMDA blockers are neurotoxins, as are γ-aminobutyric acid (GABA) mimetics, including alcohol. Thus, all anesthetics, being either NMDA blockers or GABA mimetics, are neurotoxic when administered in ontogeny (including in humans). The CNS deficits in fetal alcohol syndrome are attributable to the apoptosis induced in the manner described here.

Targeted Antibody – Production of Neurotoxicity

An immunotoxin, 192 IgG–SAP – immunoglobulin G (IgG) conjugated to a ribosome-inactivating protein (i.e., saporin; SAP) – binds to a membrane-bound p75 neurotrophin receptor on cholinergic nerve terminals; after its internalization from receptor endocytosis it is transported to the cell body, where it irreversibly impairs protein synthesis and causes ultimate neuronal cell death.

Nonselective Neurotoxin(s) with Specialized Use to Model a Disease State

When the mitochondrial complex I inhibitor rotenone is administered to rats over a period of months, a partial loss of pars compacta substantia nigra dopaminergic neurons occurs along with a hypolocomotor syndrome that in some ways models Parkinson's disease (PD). There are pros and cons to the model, including the fact that the extent of complex I inhibition with rotenone is more widespread and more homogeneous in brain than is seen in PD, the extent of nigral degeneration is much less with rotenone than in PD, and the hypolocomotor effect with rotenone is not correlated with DA cell loss, as it is in PD; also, the rotenone effects are not consistent (in replication) among investigators. Rotenone, long used as a rat poison, is a lipophilic substance that enters virtually all cells and has notable nonselectivity. Nevertheless, because rotenone is now commonly used to model PD, it might be considered in the realm of a 'selective' neurotoxin.

Neurotoxins for the Future

Ideal Properties of a Neurotoxin

One can envision an ideal neurotoxin as having a high degree of selectivity coupled with high potency for a specific neuronal target. These properties enable a neurotoxin to destroy one specific set of neurons (possibly a single phenotypic group, or possibly even a particular neuronal nucleus in brain) without damaging adjacent neurons or glia. It thus becomes more reasonable to ascribe particular functions to the missing set of neurons.

Alternatively, one can envision an ideal neurotoxin as one that mimics most or all of the histological/pathological details of a neurodegenerative or neuropsychiatric disorder as it presents in humans – possibly with deficits in many neuronal nuclei. For example, the multidimensional Parkinson syndrome in humans cannot be attributed solely to loss of pars compacta dopaminergic neurons. Rotenone, described earlier as a means for producing an animal model of PD is one attempt at this latter type of selectivity.

Validation of a Neurotoxin

For a neurotoxin that produces overt necrosis, apoptosis, or necroptosis of a single neuronal phenotype, one can validate neurotoxicity by assessing one or more of a host of neuronal functions or viability: (1) loss of neuronal perikarya and terminal innervation, (2) loss of activity of the rate-limiting neurotransmitter synthesizing enzyme, (3) loss of neurotransmitter stores, (4) loss of membranous transporter sites or activity, or (5) altered behavioral expression. For example, for a neurotoxin that destroys dopaminergic neurons in substantia nigra pars compacta, one would see a relative absence of nigral cells and relative absence of dopaminergic innervation in basal ganglia sites, loss of tyrosine hydroxylase activity, DA depletion from tissues and a decrease in microdialysate levels of DA, loss of DAT in striatal homogenates, and a relative absence of spontaneous motor activity. Biochemical, histological, immunocytochemical, physiological, and behavioral techniques might be engaged, according to the expertise of a laboratory and the degree of confirmation (of neurotoxicity) desired or required.

For a neurotoxin that produces long-term functional impairment without induction of necrosis or apoptosis, one would focus on the nature of the impairment, whether it be loss of neurotransmitter stores or loss of response to nerve stimulation, or other parameter relating to function.

Neurotoxins as Experimental Tools, and Limitations

As neuroscience ever evolves to a greater degree of sophistication in methodology, one can envision that neurotoxins will be uniquely synthesized for individualized experiments. For example, identification of a protein or protein fraction will provide a means to develop a unique binding site or antibody to the neuron under study, so that unique neuronal lesioning is produced with a high degree of selectivity. A limitation to this approach is that the clinical presentation of a neurodegenerative disorder often relates to the involvement of many different tissues or even brain sites. A targeted neurotoxin cannot

replicate all elements of a disorder. Nevertheless, the desired lesion will provide a better means of characterizing the functional role of the targeted set of neurons.

Usual limitations of a neurotoxin relate to nonselectivity of action. When there is collateral damage to nontargeted neurons (and/or glia), it becomes problematic in concluding that a particular action or function was attributable to the nerve in question. Also, if potency is low it is not possible to effectively lesion a large proportion of the target neurons or the nerve terminal tree.

Neurotoxins as Therapeutic Agents

Re-engineered clostridial neurotoxins represent a promising therapeutic strategy for targeted treatment of a variety of clinical disorders, including cancer chemotherapy. Already, this project has been initiated. A reconfigured or substituted binding domain provides selectivity of the molecule for the desired cell group; a reconfigured or substituted catalytic domain (which could feasibly be a drug or cell poison) provides the desired cellular effect. Ideally, both are accomplished with minimal collateral undesired effects.

Vanilloid receptor agonists such as capsaicin are already in use for long-term elimination of pain – initially producing excitation, eventually producing receptor desensitization, and ultimately producing toxicity to sensory neurons conducting pain impulses.

Despite these examples of clinical utility for neurotoxins, in nearly all instances preservation of cellular activity is the norm – not destruction of a cell. Accordingly, most classical neurotoxins find greatest utility as experimental research tools to model disease states and to better elucidate mechanisms involved in cell death.

Endogenous Neurotoxins and Neuroprotectants

Quinolinate, another excitatory amino acid at the NMDA receptor, is a putative endogenous neurotoxin. Perhaps as a balance, kynurenic acid is produced endogenously by glia, acting as a broad-spectrum antagonist for all ionotropic glutamate receptors, and as an antagonist with high affinity for the glycine site of the NMDA receptor complex. As such, kynurenic acid is regarded as an endogenous neuroprotectant.

Endogenously produced β-carbolines (e.g., 2,9-dimethyl-β-norharmanium), formed from tryptophan or derived indolealkylamines, and particularly when dimethylated, are neurotoxic species with the potential to damage nerves and give rise to neurodegenerative disorders. Another chemical class, tetrahydroisoquinolines, are metabolites of dopamine (e.g., *N*-methyl(*R*) salsolinol) that have been proposed as a neurotoxic species also giving rise to neurodegenerative disorders. In the case of 1-methyl-1,2,3,4-tetrahydroisoquinoline and certain β-carbolines, a neuroprotective effect is found.

More investigation is needed into purported endogenous neurotoxicants and neuroprotectants, in regard to sites of production, sites of action, threshold amounts for physiological effects, and relative susceptibility of targets.

Selective Neurotoxins in the Twenty-First Century

Discovered only in the past 40 years, selective neurotoxins are now an integral component of neuroscience research. Exogenous neurotoxins, such as MPTP and rotenone, are regularly used in PD research studies. MPTP, as a detrimental xenobiotic, was uncovered consequent to its producing Parkinson syndrome in humans. As such, MPTP has re-enlivened the debate as to whether such substances in the environment are producing nonevident nerve damage and the so-called diseases of aging. This view is further reinforced by the discovery of unnatural amino acids that similarly produce neurotoxicity and neurodegenerative disorders in humans (e.g., BOAA, BMAA). Glutamate is implicated in Alzheimer's disease, in stroke, and in ischemic injury. Glutamate receptor antagonists are under development by pharmaceutical companies, as prophylaxis against Alzheimer's disease. Huntington's disease is now viewed as a neurodegenerative disorder arising from glutamate excitotoxic neuronal damage. Glutamate antagonists and substituted AMPHs produce behaviors and lesions associated with psychosis and schizophrenia. Haloperidol and other 'typical' antipsychotics produce tardive dyskinesia, possibly via neurotoxic metabolites resembling MPP^+. It is evident that there are numerous clinical and behavioral entities which neurotoxins replicate. Many neurotoxins are yet to be discovered, and we see only the tip of the iceberg now, in terms of experimental potential and in reference to clinical applicability of neurotoxins.

Neurotoxic substituted AMPHs (e.g., methylenedioxyamphetamine, methylenedioxymethamphetamine (ecstasy)) are drugs of abuse, commonly used at all-night dance parties (i.e., 'raves'). Users take the risk of serotoninergic neuronal degeneration in brain, along with the sometimes fatal hyperthermia that occurs as an acute effect of the drug. Cocaine, another widely abused drug, is fatal because of its propensity to produce arrhythmia and convulsions. Cocaine neurotoxicity in animals is associated with chronic treatments.

The concept of endogenous neurotoxins and identifiable neurotransmitter-related neuroprotectants is a new one. Studies to date have opened the door

into what is becoming a vast arena for scientific exploration.

See also: Botulinum and Tetanus Toxins; Neuroprotection: Endogenous Mechanisms; Neuroprotection: Pharmacological Approaches; Neurotoxins and their Neurotoxicology.

Further Reading

Degterev A, Huang Z, Boyce M, et al. (2005) Chemical inhibitor of nonapoptotic cell death with therapeutic potential for ischemic brain injury. *Nature Chemical Biology* 1: 68–69.

Factor SA and Weiner WJ (2002) *Parkinson's Disease: Diagnosis and Clinical Management*. New York: Demos Medical Publishing.

Kostrzewa RM (1999) Highly selective neurotoxins. *Neurotoxin Research* 1: 3–25.

Kostrzewa RM (ed.) (1999) *Highly Selective Neurotoxins: Basic and Clinical Applications*, pp. 195–210. Totowa, NJ: Humana Press.

Levi-Montalcini R and Angeletti PU (1966) Second symposium on catecholamines. Modification of sympathetic function. Immunosympathectomy. *Pharmacological Reviews* 18: 619–628.

Lodge D, Danysz W, and Parsons CG (2002) *Ionotropic Glutamate Receptors as Therapeutic Targets*. Mountain Home, TN: F.P. Graham Publishing Co.

Segura-Aguilar J (ed.) (2001) *Mechanisms of Degeneration and Protection of the Dopaminergic System*. Mountain Home, TN: F.P. Graham Publishing Co.

Singh BR (2006) Botulinum neurotoxin structure, engineering, and novel cellular trafficking and targeting. *Neurotoxin Research* 9: 78–92.

Spencer PS, Schaumburg HH, and Ludolph AC (eds.) (2000) *Experimental and Clinical Neurotoxicology,* 2nd edn. Oxford: Oxford University Press.

Relevant Websites

http://www.NeurotoxicityResearch.com – F.P. Graham Publishing Company.

http://www.nida.nih.gov/Infofacts/ecstasy.html – National Institute on Drug Abuse.

Neurotoxins and their Neurotoxicology

M I Sabri and P S Spencer, Oregon Health and Science University, Portland, OR, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurotoxicity of drugs and other chemicals is an important and challenging public health problem. Substances that produce a primary toxic effect on the nervous system are referred to as neurotoxins or neurotoxicants. Agents such as carbon monoxide and cyanide that secondarily render damaging effects on the nervous system are also considered to be neurotoxic agents. Chemicals with neurotoxic potential occur naturally as the products of bacteria (e.g., diphtheria toxin), plants (*β*-*N*-oxalylamino-L-alanine [L-BOAA]), and animals (palytoxin); these agents often have high specificity and discrete biological actions resulting in neurotoxicity. Certain purified natural agents, such as the compounds of arsenic, lead, and manganese, have widespread toxic potential but predominantly affect the nervous system. Naturally occurring and synthetic chemicals used as therapeutic drugs (nitrofurantoin), vitamins (pyridoxine), recreational substances (ethanol, cocaine), domestic products (zinc pyridinethione), pesticides (chlorpyrifos), occupational chemicals (acrylamide, hydrocarbon solvents), and food contaminants (domoic acid) can also induce human neurological diseases. Serious efforts are therefore required to identify, eliminate, and develop a public policy for neurotoxic substances, similar to that for carcinogens. Neurotoxic agents can be divided into various categories (Table 1).

Neurotoxicology is an area of neuroscience that deals with chemicals that produce adverse or undesired effects on the nervous system and behavior of humans and animals. Although the subject of neurotoxicology is solidly based on principles of neuroscience and toxicology, the discipline also encompasses relevant aspects of many other specialized fields, including behavioral toxicology, biochemical toxicology, cancer chemotherapy and other areas of clinical pharmacology (notably the side effects of therapeutic drugs), forensic toxicology, defense medicine (predominantly chemical warfare agents), mycotoxicology, neurology, nutritional toxicology, occupational and environmental medicine, ophthalmology and otorhinolaryngology, pesticide toxicology, psychopharmacology, psychiatry, clinical psychology, teratology, toxicology, tropical medicine, veterinary medicine, and zoology (notably entomology and marine biology). Many experimental neuroscientists employ selected neurotoxic agents as discrete tools to probe the nervous system or to model human neurological or psychiatric disorders. Although the actions of exogenous chemical substances represent the primary focus of neurotoxicology, the metabolic disorders associated with elevated levels of endogenous agents (e.g., copper in Wilson's disease) are also relevant to the discipline.

Each year millions of people are occupationally or otherwise exposed to neurotoxicants. Indeed, many epidemics of neurotoxic diseases have occurred throughout the world (Table 2). Ataxia and blindness, caused by organic mercury compounds in fish in Japan and in fungicide-treated grain in Iraq; spinal cord degeneration and peripheral neuropathy caused by tri-*o*-cresylphosphate (TOCP) in cooking oil in Morocco; toxic oil syndrome in Spain; and tremors, motor disturbances, and anxiety caused by the pesticide Kepone (chlordecone) in the United States are some of the examples of human neurotoxic exposure (Table 2).

Factors Governing Neurotoxicity

Factors that determine vulnerability of the nervous system to chemical agents may be different during development, at maturity, and in late life. The developing nervous system is by far the most sensitive to the actions of neurotoxic agents. One of the most important factors that pose a threat to the nervous system is the immaturity of the blood–brain barrier (BBB) during development. The large majority of the nervous system is protected from chemical exposure by the mature BBB that separates the microenvironment of the brain parenchyma from circulating toxic substances. Regulation of the BBB is a key function of capillaries coursing through nervous tissue. Structural specializations of capillary walls in the form of tight junctions between adjacent endothelial cells constitute a diffusion barrier and allow the endothelium to regulate the selective transport of substances from blood to brain and vice versa. However, gases (oxygen, carbon monoxide, nitrous oxide), cross capillary walls with ease. The structure of chemicals is of critical importance in determining the neurotoxic response. While access to nervous tissue may dictate a direct neurotoxic effect, neurotoxicity ultimately depends on the ability of the agent to bind to the neural target and interfere with functional or structural integrity. Structure–activity relationships are therefore of cardinal importance for the expression of neurotoxicity. Recent studies have shown that 1,2-diacetylbenzene (1,2-DAB) but

Table 1 Types of neurotoxic agents

Category	*Example*
Chemical warfare agents	Sarin, soman, tabun
Drugs of abuse	Ethanol, heroin, ecstasy, nitrous oxide
Environmental pollutants	Carbon monoxide, lead, mercury
Experimental probes	Dinitrophenol, iodoacetate, pyrithiamine
Food additives/fragrances	Acetyl ethyl tetramethyltetralin, dinitro-methoxybutyltoluene
Industrial/occupational	*n*-Hexane, 1,2-diethylbenzene, acrylamide, manganese
Natural toxins	Gossypol, saxitoxin, tetanospasmin, BOAA, BMAA, cyanide, cycasin
Drugs	Cisplatin, taxol, vincristine

Table 2 Selected major neurotoxicity incidents

Year(s)	*Location*	*Agent*	*Comments*
400 BC	Rome	Lead	Hippocrates recognizes lead toxicity in mining industry
1930s	United States	TOCP	Compound often added to lubricating oil, contaminates Ginger-Jake, an alcoholic beverage; ~5000 paralyzed and ~100 000 affected
1930s	Europe	Apiol	Abortion-inducing drug containing TOCP causes neuropathy, 60 cases
1932	United States	Thallium	Barley laced with thallium used as a rodenticide used to make tortillas; 13 hospitalized with neurological symptoms, 6 die
1937	South Africa	TOCP	60 South Africans develop paralysis after using contaminated cooking oil
1946		Lead	More than 25 individuals suffer neurological effects after cleaning gasoline tanks
1950s	Japan	Mercury	Hundreds ingest fish and shellfish contaminated with mercury,121 poisoned, 46 deaths. Many infants with birth defects
1950s	France	Organotin	Contamination with triethyltin results in more than 100 deaths
1950s	Morocco	Manganese	150 miners suffer chronic manganese intoxication involving severe neurobehavioral problems
1950s–70s	United States	AETT	Neurotoxic component of fragrances; withdrawn in 1978
1956		Endrin	49 persons become ill after eating foods prepared from flour contaminated with the insecticide endrin; results in convulsions
1956	Turkey	HCB	Hexachlorobenzene, a seed grain fungicide, leads to poisoning of ~4000; 10% mortality rate
1956/1977	Japan	Clioquinol	Used to treat diarrhea found to cause neuropathy; ~10 000 affected over two decades
1959	Morocco	TOCP	Cooking oil contaminated with lubricating oil affects ~10 000 individuals
1960	Iraq	Mercury	Mercury used as fungicide to treat grain used in bread; affects more than 1000 people
1964	Japan	Mercury	Methyl mercury affects 646 individuals
1968	Japan	PCBs	Polychlorinated biphenyls leak into ice oil; affects 1665 people
1969	Japan	*n*-Hexane	93 cases of neuropathy following exposure to *n*-hexane, used to make vinyl sandals
1971	United States	Hexachlorophene	After years of bathing infants in 3% hexachlorophene, the disinfectant is found to be toxic to the nervous system
1971	Iraq	Mercury	Mercury used as fungicide to treat seed grain is used in bread; more than 5000 severe poisonings, 450 hospital deaths
1973	United States	MnBK	Fabric production plant employees exposed to solvent; more than 80 suffer polyneuropathy
1974–75	United States	Chlordecone	Chemical plant employees exposed to insecticide (Kepone); more than 20 suffer severe neurological problems
1976	United States	Leptophos	At least nine employees suffer serious neurological problems following exposure to insecticide (Phosvel)
1977	United States	Dichloropropene	24 individuals hospitalized after exposure to pesticide Telone
1979–80	United States	BHMH	Seven employees at plastic bathtub manufacturing (Lucel-7) plant experience serious neurological problems following exposure to BHMH
1980s	United States	MPTP	Impurity in synthesis of illicit drug found to cause symptoms identical to those of Parkinson disease
1981	Spain	Toxic oil	20 000 persons poisoned by toxic substance in oil, resulting in more 500 deaths; many suffer from neuropathy

Continued

Table 2 Continued

Year(s)	*Location*	*Agent*	*Comments*
1985	United States	Aldicarb	More than 1000 individuals in California, other western states, and Canada experience neuromuscular and cardiac problems following ingestion of melons contaminated with the pesticide (Aldicarb)
1987	Canada	Domoic acid	Ingestion of mussels contaminated with domoic acid causes 129 illnesses and two deaths. Symptoms include memory loss, disorientation, and seizures
1989–91	United States	L-Tryptophan dimer	1543 cases of eosinophilia-myalgia syndrome, with 28 deaths
1995	Japan	Sarin	4000 sickened and ten deaths from terrorist attack on Tokyo subway
2002	Worldwide	Acrylamide	Identification of acrylamide in fried food (French fries)

TOCP, tri-*o*-cresylphosphate; AETT, acetyl ethyl tetramethyl tetralin; MnBK, methyl-*n*-butyl ketone; BHMH, 2-*t*-butylazo-2-hydroxy-5-methylhexane; MPTP, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine.
Adapted from *Neurotoxicity: Identifying and Controlling Poisons of the Nervous System* (1990), US Congress, Office of Technology Assessment, Washington, DC.

not 1,3-DAB induces peripheral neuropathy because only 1,2-DAB forms adducts with neuroproteins. Similarly 2,5-hexanedione but not 2,4-hexanedione induces hindlimb weakness and causes distal axonal degeneration because only the former reacts with epsilon amino group of lysine moiety and cross-links neural proteins.

The complexity of the nervous system also greatly contributes to its vulnerability to toxic agents. The nervous system is composed of a number of cell types (e.g., neurons and glia) that mature at different times during development. Neurons are classified and identified on the basis of their anatomical locations (hippocampal, striatal), neurotransmitters (cholinergic, dopaminergic), cellular structures (granular, pyramidal), or functions (motor, sensory). Oligodendrocytes, astrocytes, microglia, and Schwann cells are classes of neuroglia that are responsible for a broad range of functions, including neuronal support, maintenance, phagocytosis, and myelin production in the central and peripheral nervous system. Preservation of the normal relationship between neurons and glia is very important for the maintenance of nervous system function. Neurotoxins can easily disrupt this relationship by selectively damaging either cell type. Neurotoxicity is therefore one of the most complex toxic responses in humans and animals. The range of human neurotoxic disorders associated with overexposure to exogenous substances is given in Table 3. Some individuals may be more susceptible to certain neurotoxic agents because of age, malnutrition, genetic type, or disease (Table 4).

The major factors that govern the effect of circulating neurotoxic agents include the ability to gain access to neural tissue, the specific target within the nervous system, the concentration of the proximate toxin at the target site, and the chemical reactivity of the agent and its duration of exposure to the target site. All these factors are heavily influenced by circumstances surrounding exposure to the chemical, its port of entry into the body, its distribution within tissues, the rate of metabolism, the toxicity of its metabolic products, and the efficacy of their excretion. Dose of the chemical substance and duration of exposure are pivotal factors in determining whether an agent will produce a neurotoxic effect. Substances that are harmless at low levels may become neurotoxic as the dose and duration of exposure increase. Others, such as pyridoxine (vitamin B6), may be required in small amounts for normal nervous system function but become neurotoxic when the dose exceeds a certain level. Different types of neurotoxic responses often appear as a function of dose and duration of exposure. For example, ethanol in large quantities over a short period of time induces a reversible state of inebriation; prolonged consumption of ethanol in large quantities not only elicits the acute response, but also causes the development of chronic degenerative disease of brain and peripheral nerves.

Targets of Neurotoxic Chemicals

Neurotoxic substances act directly or indirectly to interfere with the structure and/or function of neural elements. Indirect mechanisms of action include chemical induction of hypoxia (carbon monoxide), hypokalemia (barium), or a vitamin deficiency state (ethanol). During neural development, chemicals that perturb mitosis or cell migration are especially likely to induce malformations (e.g., spinal bifida). Other compounds, such as ethanol, can induce behavioral abnormalities in offspring unassociated with major structural deficits. Although most chemical substances have unknown mechanisms of action, target sites for

Table 3 Illustrative manifestations of adult human neurotoxicity

Function affected	*Manifestation*	*Example*
Sensorium	Irritability	Carbon disulfide
	apathy/lethargy	carbon monoxide
	attention difficulty	anticholinesterases
	illusions, delusions	ergot
	depression, euphoria	dicyclopentadiene
	stupor, coma	3-nitropropionic acid
Sensory, special	Abnormalities of:	
	smell	cadmium
	vision	chloroquine
	taste	selenium
	audition	toluene
	balance	methyl nitrite
Somatosensory	Skin senses	Dichloroacetylene
	(e.g., numbness, pain)	acrylamide, *n*-hexane
Motor	Muscle weakness	Organophosphates
	paralysis	*Lathyrus sativus*
	spasticity	MPTP
	tremor	manganese
	dystonia	organomercury
	incoordination	lead
	fasciculation	anticholinesterases
	cramps	acetonitrile
	seizures, convulsions	carbon monoxide
Autonomic	Abnormalities of:	
	sweating	acrylamide
	temperature control	chlordane
	gastrointestinal function	lead dinitrobenzene
	appetite/body weight	Vacor
	cardiovascular control	thallium
	urination	dimethylaminopropionitrile
	sexual function	β-chloroprene
Immune system	myositis, vasculitis	L-Tryptophan dimer

Adapted from World Health Organization (1985) *Environmental Health Criteria: Nervous System Toxicity.* WHO, Geneva.

Table 4 Recognized susceptibility factors for human neurotoxic disorders

1. Age
2. Genetic abnormality (slow acetylators and hydroxylators)
3. Malnutrition (vitamin deficiency)
4. Preexisting hepatic disorders (impaired metabolism/ inactivation of neurotoxic factors)
5. Preexisting neurological and psychiatric disorders (epilepsy, schizophrenia)
6. Preexisting renal disorders (impaired excretion of neurotoxic agents and their metabolites)

some substances have been determined at the molecular and cellular level (Figures 1 and 2). Because many agents produce different effects as a function of dose and duration of exposure, their classification is problematic. Many agents inhibit axonal transport system and thereby induce axonal degeneration. Some neurotoxic substances disrupt the anterograde transport of materials from sites of synthesis (perikarya) to sites of utilization (e.g., β,β'-iminodipropionitrile). Certain neurotoxic agents interfere with retrograde transport (e.g., acrylamide). Some chemicals (e.g., vinca alkaloids) bind to tubulin and thereby inhibit microtubule-based bidirectional transport of axonal materials and elicit neurotoxicity. Presently, the most satisfactory approach is to group neurotoxins by their apparent target, recognizing that more than one site may be involved (Table 5). This provides a rational approach to understanding the clinical consequences of exposure and the starting point of an inquiry directed at mechanisms of toxicity. This approach also demonstrates that chemical substances of diverse origins may have remarkably similar cellular and functional effects. For example, chemicals that interfere with the opening or closing of sodium channels all produce symptoms of circumoral and distal-extremity dysesthesias in overexposed humans. Such agents can be found in algae (ciguatoxin), higher plants (veratrum alkaloids), invertebrates (tetrodotoxin), vertebrates (batrachotoxin), or manufactured chemicals (pyrethroids, dichloro diphenyl tetrachloroethane (DDT).

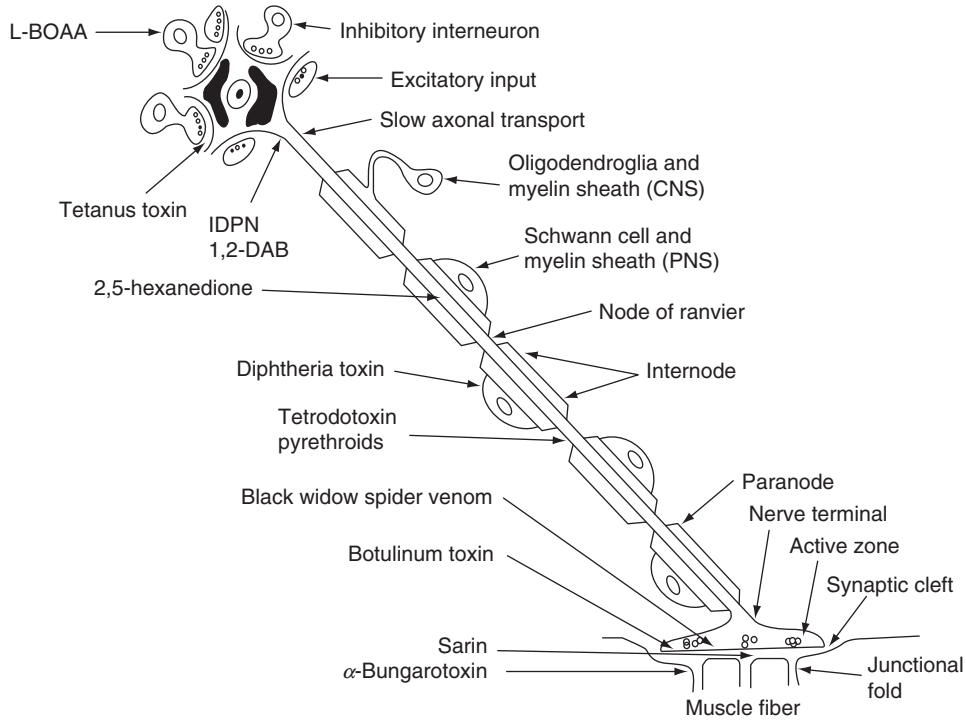

Figure 1 Diagram of a motor neuron showing cytological landmarks (right) and sites of action of certain neurotoxins (left). Tetanus toxin is taken up at nerve terminals and carried by retrograde transport to the neuronal perikaryon from which it passes transsynaptically to bind to, and block, axosomatic inhibitory inputs. 1,2-Diacetylbenzene (1,2-DAB), β,β'-iminodipropionitrile (IDPN), and 2,5-hexanedione block axonal transport of slow- and fast-moving materials synthesized in the cell body. β-*N*-Oxalylamino-L-alanine (L-BOAA) activates non-*N*-methyl-D-aspartate (non-NMDA) receptors. Diphtheria toxin binds to Schwann cells and arrests protein synthesis. Tetrodotoxin and pyrethroids interfere with nerve conduction by disrupting the function of sodium channels at nodes of Ranvier. Black widow spider venom causes massive release of neurotransmitter by promoting fusion of synaptic vesicles at sites outside the active zones. Botulinum toxin impairs release of acetylcholine by interfering with exocytosis at active zones. α-Bungarotoxin blocks neurotransmission by binding to the acetylcholine receptor. Sarin promotes excessive synaptic activity by inhibiting acetylcholinesterase, the enzyme responsible for terminating the action of the neurotransmitter, acetylcholine. Adapted from Adelman G and Smith BH (eds.) (2004) *Encyclopedia of Neuroscience,* 3rd rev. edn. New York: Elsevier, with permission from Elsevier.

Occurrence of Neurotoxic Disorders

Neurobehavioral dysfunction caused by toxic substances is widely recognized in humans. Veterinary neurotoxic disorders are common in animals grazing on poisonous plants (e.g., cycad) and in domestic animals exposed to pesticides. Many human neurotoxic syndromes are associated with the side effects of prolonged treatment with therapeutic drugs, notably anticonvulsants, anticholinergics, neuroleptics, and antiparkinsonian and antineoplastic drugs. Chinese traditional medicine may incorporate the products of dangerous plants (podophyllin) or contaminants (thallium). Neurological and psychiatric dysfunction also commonly occur with abuse of ethanol, narcotics, hallucinogens, central stimulants, solvents, and nitrous oxide. In the occupational setting, the acute cholinergic crisis associated with pesticide poisoning is commonplace and responsible for numerous deaths. Many other substances found in the workplace have been noted to induce short-term or long-lasting disorders of the nervous system. Environmental pollutants with neurotoxic potential (carbon monoxide, lead, mercury) are pervasive. Numerous animals secrete or harbor potent neurotoxins (conotoxin) that sometimes affect humans. Neurotoxic agents present as intrinsic components of foodstuffs (*Manihot esculenta, Lathyrus sativus*) or as contaminants (ergot in *Claviceps* spp.) are also responsible for outbreaks of neurological disease, especially in certain undeveloped regions of the world. Other biological neurotoxins in these regions are those associated with botulism, tetanus, and diphtheria. Finally, chemicals with experimentally proven neurotoxic potential also exist as food additives (monosodium glutamate, aspartate), flavors and fragrances (2,6-dinitro-3-methoxy-4-*t*-butyltoluene), and certain over-the-counter antiseborrheic agents (zinc pyridinethione). Substances with neurotoxic properties are therefore encountered in food, air, water, the general environment, and domestic and occupational settings.

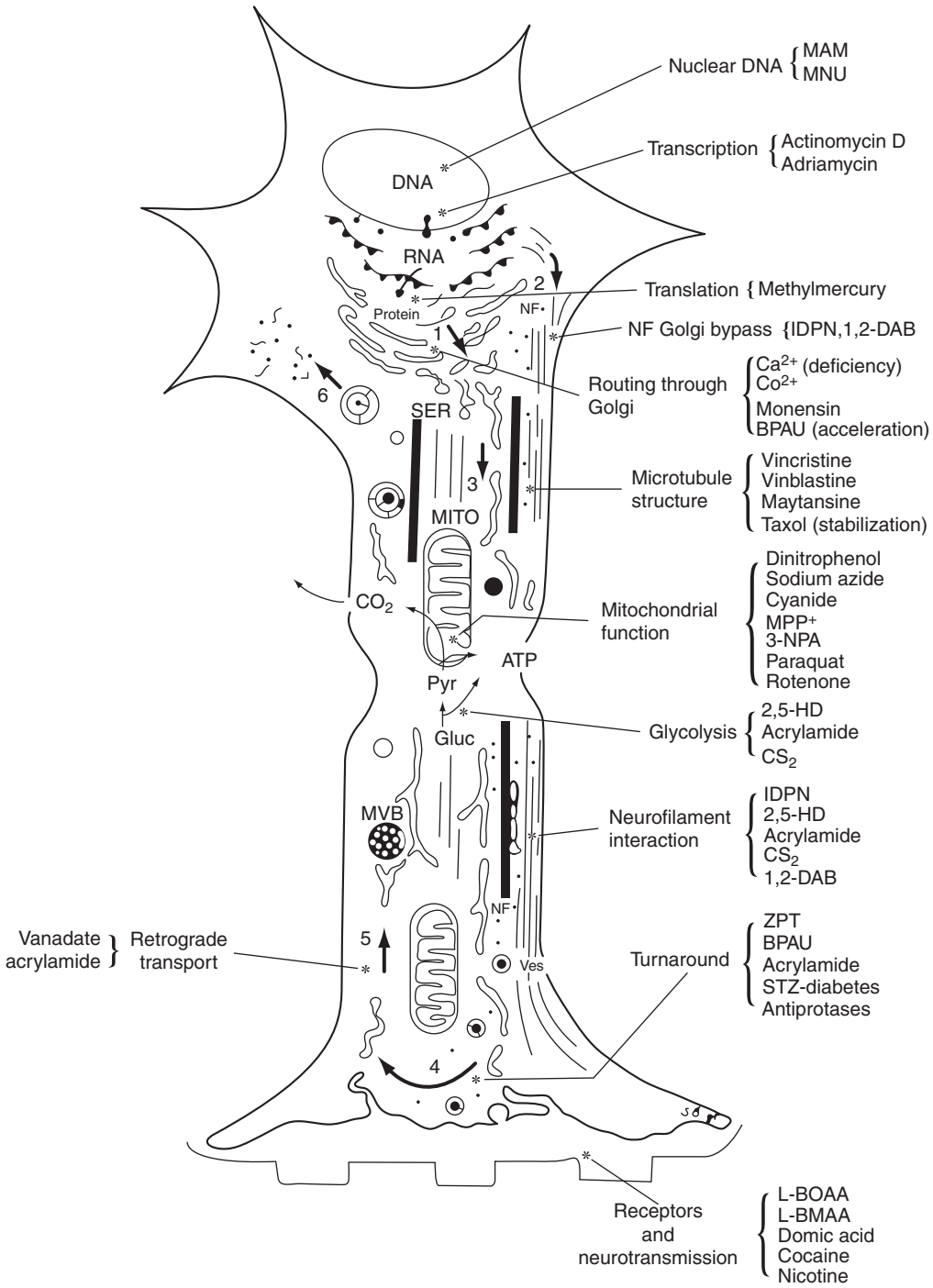

Figure 2 Diagram of a neuron showing suspected and/or established intracellular sites of action of selected neurotoxins that induce peripheral neuropathy and other neurodegenerative diseases. Although certain agents form adducts with nuclear DNA and interfere with transcription or translation in the cell body, some neurotoxins disrupt intracellular mechanisms that are indicated by the numbered arrows: (1) routing of synthesized proteins via the Golgi apparatus; (2) proteins that bypass the Golgi apparatus; (3) microtubule-based fast anterograde transport (kinesin-driven); (4) turnaround; (5) retrograde transport (dynein-driven); or (6) lysosomal digestion. Some neurotoxins disrupt mitochondrial phosphorylation, glucose metabolism, or interact with neurofilaments. NF, neurofilament; SER, smooth endoplasmic reticulum; MITO, mitochondria; ATP, adenosine triphosphate; Pyr, pyruvate; Gluc, glucose; Ves, vesicle (synaptic); MVB, multivesicular body; MAM, methylazoxymethanol; NMU, N-methylnitrosourea; IDPN, β,β'-iminodipropionitrile; 1,2-DAB, 1,2-diacetylbenzene; BPAU, *p*-bromophenylacetylurea; MPP^+, 1-methyl-4-phenylpyridinium; 3-NPA, 3-nitropropionic acid; 2,5-HD, 2,5-hexanedione; CS_2, carbon disulfide; ZPT, zinc pyridinethione; L-BOAA, β-*N*-oxalylamino-L-alanine; L-BMAA, β-*N*-methylamino-L-alanine; STZ, strepto zotocin. Adapted from Adelman G and Smith BH (eds.) (2004) *Encyclopedia of Neuroscience,* 3rd rev. edn. New York: Elsevier, with permission from Elsevier.

Table 5 Classification of neurotoxic actions by target

I. Prenatal period
  A. Neural tube (hypervitaminosis A)
  B. Neurogenesis/organogenesis (ethanol, methylazoxymethanol)
II. Postnatal period
  A. Neurons
    1. Excitable membrane (channel agents including pyrethroids)
    2. Neurotransmitter systems, for example, cholinergic agonists (anticholinesterases); antagonists (tricyclic antidepressants)
    3. Structural integrity: neuronopathy (doxorubicin, mercury, trimethyltin); axonopathy (acrylamide, carbon disulfide, *n*-hexane, 1,2-diethylbenzene); dendritopathy (glutamate excitotoxins)
  B. Glial cells
    1. Myelinating cells and myelin: oligodendrocytes (CNS) (isoniazid, triethyltin); Schwann cells (PNS) (lead, diphtheria toxin)
    2. Astrocytes (6-aminonicotinamide)
    3. Ependymal cells (amoscoline)
  C. Special sense organs
    1. Olfactory/gustatory (penicillamine, thiouracil)
    2. Visual (methanol, chloroquine)
    3. Auditory (noise, toluene)
    4. Vestibular (hydoquinone)
  D. Muscle: Striated (anticholinesterases)
  E. Neural vasculature (cadmium)
  F. Neuroendocrine system: Hypothalamus/hypophysis (chlordecone)
  G. Immune system with secondary effects on nervous system or musculature (L-tryptophan dimer)
  H. Various cell types: malignant transformation with primary or secondary neoplasms affecting nervous system (alkyl nitrosoureas)

Adapted from World Health Organization (1985) *Environmental Health Criteria: Nervous System Toxicity*. WHO, Geneva.

Neurotoxins as Experimental Tools

Certain chemical neurotoxins have proven to be powerful probes to investigate the normal functioning of the nervous system. Many of these are highly specialized biological toxins with discrete sites of action. For example, sodium channel agents, such as tetrodotoxin and saxitoxin, have played a major role in elucidating molecular mechanisms underlying the electrical properties of the neuronal membrane. Snake and spider venoms have proved useful for studying the action of acetylcholine at the neuromuscular junction. Kainic acid has been extensively employed to explore excitatory mechanisms involving glutamate receptors. Mechanisms underlying axonal transport have been dissected with agents such as the spindle inhibitors (colchicine, vincristine), β,β'-iminodipropionitrile, and other neurotoxins (Table 2). Many other substances have been employed to create and explore animal models of human neuromuscular disease. Repeated exposure to acrylamide, 2,5-hexanedione, cisplatin, taxol, or TOCP produces models of peripheral neuropathy. Experimental parkinsonism has been induced in primates by 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) and corticospinal dysfunction by L-BOAA. Paraquat, a herbicide, and rotenone, a pesticide, exert their toxic effects by inhibiting mitochondrial functions, causing oxidative stress and neurodegeneration. The pathogenesis of myasthenia gravis as a disorder of acetylcholine receptors was worked out with the aid of α-bungarotoxin and other elapid toxins. Mechanisms underlying certain types of demyelinating disease have been illuminated with agents such as hexachlorophene, lysolecithin, and acetyl ethyl tetramethyl tetralin (AETT). 1,2-DAB, the putative diketo metabolite of the neurotoxic aromatic solvent 1,2-diethylbenzene, damages motor neurons in rat spinal cord and induces giant proximal axonal swellings filled with 10-nm neurofilament, a prominent early feature of neuron degeneration in amyotrophic lateral sclerosis.

See also: Botulinum and Tetanus Toxins; Neuromuscular Junction (NMJ): A Target for Natural and Environmental Toxins in Humans; Neuromuscular Junction Plasticity in Mammals and Botulinum Toxins; Neuropathy: Peripheral; Neurotoxins.

Further Reading

Abou-Donia MB (1992) *Neurotoxicology*. Boca Raton, FL: CRC Press.

Adelman G and Smith BH (eds.) (2004) *Encyclopedia of Neuroscience*, 3rd rev. edn. New York: Elsevier.

Anthony DC, Montine TJ, and Graham DG (1996) Toxic responses of the nervous system. In: Klaassen CD (ed.) *Casarett & Doull's Toxicology. The Basic Science of Poisons*, 5th edn. New York: McGraw-Hill.

Chang LW and Dyer RS (1995) *Handbook of Neurotoxicity*. New York: Marcel Dekker.

D'Mello JPF (2003) *Food Safety, Contaminants and Toxins*. Oxford: CAB Publishing.

Federal Register (1994) *Principles of Neurotoxicity Risk Assessment. Environmental Protection Agency Final Report*. Washington, DC: US Government Printing Office.

Harry GJ (1994) *Developmental Neurotoxicology*. Boca Raton, FL: CRC Press.

Herken H and Hucho F (1992) *Selective Neurotoxicity*. Berlin: Springer.

Hucho F and Ovchinnikov YA (1983) *Toxins as Tools in Neurochemistry*. Berlin: Walter de Gruyter.

Massaro EJ (2002) *Handbook of Neurotoxicology*. Totowa, NJ: Humana Press.

National Research Council (1992) *Environmental Neurotoxicology*. Washington, DC: National Academy Press.

Ochs S and Brimijoin WS (1993) Axonal transport. In: Dyck PJ, Thomas PK, Griffin JW, et al. (eds.) *Peripheral Neuropathy*, 3rd edn., vol. 1. Philadelphia: Saunders.

O'Donoghue J (1985) *Neurotoxicity of Industrial Agents*. Boca Raton, FL: CRC Press.

Price DL and Griffin JW (1980) Neurons and ensheathing cells as targets of disease processes. In: Spencer PS and Schaumburg HH (eds.) *Experimental and Clinical Neurotoxicology*. Baltimore: Williams & Wilkins.

Spencer PS and Schaumburg HH (2000) *Experimental and Clinical Neurotoxicology*, 2nd edn. New York: Oxford University Press.

US Congress (1990) *Neurotoxicity: Identifying and Controlling Poisons of the Nervous System*. Washington, DC: Office of Technology Assessment.

Wexler P (ed.) (2005) *Encyclopedia of Toxicology*, 2nd edn. Oxford: Elsevier/Academic Press.

World Health Organization (WHO) (1985) *Environmental Health Criteria: Nervous System Toxicity*. WHO, Geneva.

Zavia NH (2004) *Molecular Neurotoxicology*. Boca Raton, FL: CRC Press.

Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells

V Gallo and L-J Chew, Children's National Medical Center, Washington, DC, USA

© 2009 Elsevier Ltd. All rights reserved.

Oligodendrocytes

Neurotransmitter Receptors and Development

Oligodendrocytes (OLs) originate from oligodendrocyte progenitor cells (OPCs) that can be identified both *in vivo* and in culture based on the expression of specific cellular markers, including the platelet-derived growth factor receptor-α (PDGFR-α), the NG2 proteoglycan, and the transcription factor Olig2. OPC lineage progression to fully differentiated, myelinating oligodendrocytes can also be monitored by the expression of developmentally regulated antigens (O4 and O1) and myelin components, including myelin basic protein (MBP), 2′,3′-cyclic nucleotide 3′-phosphohydrolase (CNPase), proteolipid protein (PLP), and myelin-associated glycoprotein (MAG). One of the crucial series of events that occur in OPC development is arrest in cell proliferation, permanent cell cycle withdrawal, and terminal differentiation. It is therefore of critical importance to fully understand how OPC proliferation and cell cycle arrest are regulated both intrinsically and by extracellular signals.

OPCs proliferate in the postnatal subventricular zone (SVZ), and migrate out of the SVZ to developing white-matter regions, where they differentiate to mature OLs and myelinate axons. OPCs develop postnatally, after embryonic neurogenesis and formation of synaptic contacts are largely completed. Therefore neuronally released signals are likely to affect OPC development in the postnatal brain. Neurotransmitters are now recognized as developmental signals that regulate many aspects of neural cell development, including OPC maturation and myelination. For example, glutamate affects gene expression in OPCs through a Ca^{2+}-dependent process. OPCs express all neurotransmitter receptors that have also been found in neurons, although with some molecular and functional differences. Table 1 provides a complete list of these receptors.

Cell proliferation The control of OPC proliferation is one of the mechanisms that determine the final number of oligodendrocytes in the brain. Neurotransmitter amino acids, including γ-aminobutyric acid (GABA), glutamate, and glycine, exert different effects on OPC proliferation by acting on selective receptors that are expressed at different stages of the OL lineage.

Electrophysiological patch clamp studies in cultured brain OPCs demonstrate that these cells display GABA-elicited current responses. The pharmacological properties of these GABA-activated channels were similar to their neuronal counterparts, in terms of sensitivity to different receptor agonists and antagonists. However, differently from neurons, the benzodiazepine agonists and inverse agonists did not modulate the GABA responses in OPCs. Also in contrast to neurons, the rate of deactivation produced by brief agonist pulses in OPCs was voltage independent.

In the subventricular zone (SVZ), GABA inhibits cell proliferation of neural progenitors, including OPCs, acting through $GABA_A$ receptors. $GABA_A$ receptor agonists also inhibit proliferation of perinatal OPCs in cerebellar slice cultures. These inhibitory effects are most likely due to a depolarizing effect of GABA in immature cells. Interestingly, these effects of GABA persist in the adult brain, as demonstrated by the finding that progenitors of the SVZ/rostral migratory stream (RMS)/olfactory bulb (OB) pathway, including cells that generate oligodendrocytes in the OB, are depolarized by activation of extrasynaptic $GABA_A$ receptors to inhibit their proliferation.

OPCs co-express ionotropic glutamate receptors of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and kainate subtypes. Electrophysiological analysis shows rapidly desensitizing responses to kainate and to AMPA. Analysis in single cells shows that two receptor populations are co-expressed and can be distinguished based on affinity for kainate, and on differential sensitivity for potentiation by concanavalin A and for block of desensitization by cylothiazide. Glutamate regulates OPC proliferation through the activation of AMPA receptors. Glutamate inhibits OPC proliferation in purified cultures and in cerebellar tissue slices. Furthermore, glutamate inhibits SVZ progenitor cell proliferation in cortical slices. The antiproliferative effects of glutamate are Ca^{2+} independent and arise from an increase in intracellular Na^+ and subsequent block of outward K^+ currents. AMPA/kainate receptor agonists cause a reversible G1 cell cycle arrest in OPCs, which is associated with accumulation of the cyclin kinase inhibitors $p27^{Kip1}$ and $p21^{CIP1}$, and a selective decrease in cdk2 activity and cdk2/cyclin E complex formation.

Table 1 Neurotransmitter receptors in oligodendrocytes and Schwann cells

Receptor	*Effects in oligodendrocytes*	*Effects in Schwann cells*
Acetylcholine		
Muscarinic M1–M4	Activation of the PI3K/Akt pathway	Increase in IP3
	Stimulation of cell proliferation	Decrease in $[cAMP]_i$
	Stimulation of cell survival	Block GFAP induction
Nicotinic	Increase in $[Ca^{2+}]_i$	
α3, α4, α5, α7; β2, β4		
Dopamine		
D3 receptor	Inhibition of cell differentiation	
GABA		
$GABA_A$ receptors	Depolarizing chloride currents	Mediation of neurosteroid (THP) action
	Increase in $[Ca^{2+}]_i$	Stimulation of PMP22 expression
	Lack of sensitivity to benzodiazepines	
	Inhibition of cell proliferation	
	Stimulation of MBP expression through THP	
$GABA_B$ receptors		Decrease in $[cAMP]_i$
		Inhibition of cell proliferation
		Inhibition of P_0, PMP_{22}, MAG expression
Glycine		
α2- and β-subunits	Depolarizing chloride currents	
	Increase in $[Ca^{2+}]_i$	
Glutamate		
AMPA and kainate: GluR2, 3, 4, 6, 7; KA1, 2	Rapidly desensitizing glutamate, kainate and AMPA currents	Stimulation of ATP release
	Increase in $[Ca^{2+}]_i$	
	Ca^{2+}-dependent induction of gene expression	
	Blockage of K^+ channels	
	Inhibition of cell proliferation	
	Inhibition of cell differentiation	
	Stimulation of integrin-mediated cell migration	
	Involvement in excitotoxic damage in hypoxia/ischemia and periventricular leukomalacia	
NMDA: NR1, NR2A–C, and NR3A	Depolarizing currents in immature and mature oligodendrocytes	
	Involvement in oligodendrocyte damage in hypoxia/ischemia	
Norepinephrine		
α,β-Adrenergic	Increase in $[cAMP]_i$ through β receptors	Increase in $[cAMP]_i$
	Inhibition of cell proliferation through β receptors	
	Stimulation of cell differentiation through β receptors	
Purinergic	Increase in $[Ca^{2+}]_i$	Increase in $[Ca^{2+}]_i$
P2X and P2Y receptors	Inhibition of cell proliferation	Inhibition of cell proliferation
		Inhibition of differentiation
P2X7 receptor		Regulation of IL-1β release
		Stimulation of inward current
Adenosine A2	Inhibition of cell proliferation	Increase in $[cAMP]_i$
Serotonin		
5HT1B, 2A, 2C		Increase in $[Ca^{2+}]_i$
Neuropeptides		
Calcitonin gene-related peptide		Stimulation of cell proliferation
Vasoactive intestinal peptide		Increase in $[cAMP]_i$

Studies in developing white matter have demonstrated that OPCs receive direct synaptic contacts from callaterals of unmyelinated axons in corpus callosum glutamate release from these axons on to NG2+ progenitors is actively dependent and causes activation of AMPA receptors in OPCs. The functional consequences of this phenomenon are unknown, but glutamate regulate OPC proliferation and differentiation through these synapses. It is also unknown whether release of other neurotransmitters also occurs through direct synaptic contacts on OPCs in white matter regions.

Purinergic receptors of the P2X and P2Y subtypes are expressed in OPCs in culture and *in vivo*. Adenosine

increases intracellular Ca^{2+} in OPCs and inhibits their proliferation. Similar effects are also observed by stimulating dorsal root ganglia (DRG) neurons in neuronal–glial co-cultures and in cerebellar slice cultures. The effects of adenosine are reproduced by ATP and ADP, which increase intracellular Ca^{2+} and inhibit OPC proliferation by activating metabotropic P2 receptors. P2Y receptors are localized in white-matter $NG2^+$ OPCs *in vivo* during postnatal development, suggesting that receptor signaling might also regulate proliferation *in vivo*. Inhibition of OPC proliferation is also observed with the β-adrenergic receptor agonist isoproterenol, which increases cyclic adenosine monophosphate (cAMP) levels in OPCs.

Opposite effects on OPC proliferation are detected with muscarinic receptor agonists. Carbachol stimulates OPC proliferation in culture through phosphatidylinositol (PI)-coupled muscarinic receptors in the absence of other mitogenic factors.

Cell migration OPC migration is a highly regulated process that is critical for proper white-matter development. Glutamate acting at G-protein-coupled AMPA receptors stimulates α-v integrin-mediated OPC migration by increasing the frequency of Ca^{2+} transients and by promoting cell motility. Interestingly, AMPA receptor subunits were found to be associated with the proteolipid protein (PLP)/α-v integrin complex, indicating that interactions among several membrane proteins regulate OPC migration. ATP and ADP also stimulate OPC migration, but the mechanism is still undefined.

Cell differentiation and survival OPC differentiation can be monitored by using specific antibodies that bind to developmentally regulated antigens. The possibility of monitoring OPC lineage progression has provided unique opportunities to investigate the effects of neurotransmitters on oligodendrocyte development and myelination. Glutamate prevents OPC lineage progression to the $O4^+$ and $O1^+$ stage by maintaining OPCs in cell cycle arrest, whereas isoproterenol exerts opposite effects and promotes cell differentiation, as determined by a large increase in the percentage of $O4^+$ preoligodendrocytes and $O1^+$ oligodendrocytes. Adenosine is a potent stimulator of OPC differentiation and myelination in DRG–OPC cultures, and ATP and ADP also promote OPC differentiation.

Dopamine D3 receptor expression was detected in OPC cultures and in the corpus callosum at the peak of myelination, but was not found in mature oligodendrocytes. Receptor activation with the agonist quinpirole led to reduced numbers of $O1^+$ cells, along with an increase in the number of precursors, indicating that dopamine binding delays or inhibits OPC differentiation.

OL programmed cell death is one of the crucial mechanisms that regulate the final OL number in the mature brain. The phosphatidylinositol 3-kinase (PI3K)/Akt pathway plays a central role in OPC survival, as demonstrated by effects of both muscarinic and cannabinoid receptor agonists, which promote cell survival through direct activation of this pathway.

Hormone Receptors and Development

Expression of hormone receptors in OL lineage cells is summarized in Table 2. Similar to neurotransmitter receptors, hormone receptors are also expressed through different stages of OL development, suggesting that distinct aspects of OL maturation are under hormonal influence.

Cell proliferation Insulin is a potent stimulator of oligodendrocyte generation and maturation. Insulin-like growth factor 1 (IGF-1) receptors are highly expressed in OPCs and in more mature OLs. Exposure of OPCs to IGF-1 increases cell proliferation through IGF-1 receptors, and IGF-1 synergizes with fibroblast growth factor 2 (FGF-2) to promote OPC entry into the cell cycle. This effect is due to enhanced cyclin D1 in G1 phase.

OLs express mu opioid receptor immunoreactivity early during development, and treatment with receptor agonists causes a large increase in $O4^+$ cell proliferation. Opioid receptors are also highly expressed in neural progenitors of the SVZ, although physiological effects different from those observed *in vitro* were associated with these receptors *in vivo*. In fact, proliferation of glial fibrillary acidic protein (GFAP)-negative cells in the SVZ, which presumably include OPCs, was decreased by morphine treatment.

Retinoic acid (RA) exerts profound effects on brain development, and RA nuclear α receptors (RAR-α) are expressed in OPCs. Interestingly, RAR-α expression is transiently stimulated by RA in the presence or in the absence of basic fibroblast growth factor (bFGF). Activation of these receptors inhibits OPC proliferation, indicating that retinoids participate, in concert with growth factors, in the control of oligodendrocyte development. Similar to RA, other hydrophobic signals, including thyroid hormone and glucocorticoids, also cause cell cycle arrest in OPCs.

Cell differentiation and survival OPCs express thyroid hormone receptors (TRs), and thyroid hormone is considered a major positive regulator of OPC differentiation. Oligodendrocyte subpopulations co-express the TRα and TRβ1 epitopes at specific developmental

Table 2 Hormone receptors in oligodendrocytes and Schwann cells

Receptor	*Effects in oligodendrocytes*	*Effects in Schwann cells*
IGF-1	Promotion of cell cycle entry Promotion of cell proliferation Increase in cellular myelin content	Stimulation of PI3K Promotion of cell survival Stimulation of cell proliferation Stimulation of differentiation Stimulation of focal adhesion kinase Stimulation of cell motility
Thyroid hormone (α1, β1)	Regulation of cell cycle exit Regulation of initiation of differentiation Stimulation of MBP expression	Stimulation of immediate early genes Stimulation of P_0 expression Stimulation of MBP expression
Retinoic acid	Regulation of cell proliferation Regulation of cell differentiation	Decrease in ciliary neurotrophic factor
Estrogen	Delay of cell cycle exit Promotion of membrane sheet formation	Stimulation of cell proliferation with cAMP Induction of progesterone receptor
Progesterone	Enhancement of cell branching Promotion of myelination Stimulation of MBP expression	Stimulation of P_0 and PMP22 expression Regulation of $GABA_B$ receptor expression
Glucocorticoid	Inhibition of cell proliferation	Stimulation of P_0 and PMP22 expression
Opioids (mu and kappa)	Regulation of cell survival Promotion of cell differentiation	
Parathyroid hormone-related protein		Promotion of cell proliferation Promotion of cell migration along axons
Erythropoietin		Promotion of cell proliferation

stages. Both TRα and TRβ1 isoforms are co-localized in oligodendrocytes during early development, before CNPase is expressed. Expression of TRβ1 varies during maturation, whereas TRα decreases during terminal differentiation. Thyroid hormone promotes OPC maturation in culture by (1) timing cell cycle exit and initiation of the differentiation process and (2) by enhancing MBP gene expression and myelination. Consistent with these findings, MBP mRNAs were significantly downregulated in the brain of hypothyroid rats.

IGF-1 strongly promotes OL generation and differentiation *in vitro*, and transgenic mice that overexpress IGF-1 display a higher content of myelin per single OL than do wild-type littermates, although the total OL number in the brain was unchanged. These findings demonstrate that endogenous IGF receptor activation promotes myelin synthesis *in vivo*.

RA has been reported to display opposite effects on OPC differentiation in different cellular systems. In cultured optic nerve OPCs, RA promotes cell differentiation, whereas in brain OPCs, RA inhibits cell differentiation. The opposite effects might be due to regional differences, or to distinct culture conditions used.

OLs possess a considerable 3α-hydroxysteroid dehydrogenase activity, the enzyme involved in progesterone and testosterone metabolism in the brain. The neurosteroid progesterone increases the expression of myelin proteins (MBP and CNPase) in rat OLs in primary culture. Consistent with these findings, this neurosteroid also enhances myelination in cerebellar organotypic slice cultures, as determined by MBP expression. The effect of progesterone on MBP expression is mediated by the classical intracellular progesterone (PROG) receptor (PR), as demonstrated by studies with specific receptor agonists and antagonists. Furthermore, consistent with these findings, no PROG effects on myelination are observed in organotypic cultures obtained from a PR knockout mouse. Progesterone is metabolized in the cerebellar cultures to 5α-dihydroprogesterone (5α-DHP) and to the $GABA_A$ receptor-active metabolite 3α,5α-tetrahydroprogesterone (3α,5α-THP; allopregnanolone). The 5α-reductase inhibitor L685-273 partially inhibits the effect of PROG, whereas 3α,5α-THP significantly stimulates MBP expression. This increase in MBP expression by 3α,5α-THP involves $GABA_A$ receptors, as demonstrated by inhibition with the selective $GABA_A$ receptor antagonist bicuculline. These results indicate enhanced MBP expression and myelination through two signaling systems, the intracellular PR and membrane $GABA_A$ receptors, and they indicate a new role for $GABA_A$ receptors in myelination. Gonadal steroids also elicit differential responses in OPCs: 17β-estradiol, but not PROG or testosterone, delays the exit of OPCs from the cell cycle and enhances membrane sheet formation, whereas PROG promotes cellular branching. Cytoplasmic estrogen receptor β (ERβ) is believed to mediate the increase in membrane formation induced by estradiol, suggesting that sex steroids could contribute to differences in repair.

OLs express proenkephalin and prodynorphin peptides, as well as both mu and kappa opioid receptors. A kappa receptor antagonist inhibits OL differentiation, indicating that endogenous opioids synthesized by OLs modulate cell survival via autocrine and/or paracrine mechanisms.

Oligodendrocyte Receptor Signaling in Injury and Disease

OLs are the major target of a variety of diseases and neurological disorders. For some of these disorders a neurotransmitter or hormone receptor-mediated mechanism is implicated. In other cases, enhanced OL survival under pathological conditions depends on selective hormone receptor activation.

It is well established that AMPA receptor antagonists attenuate excitotoxic injury in the developing white matter that occurs in hypoxia/ischemia. In particular, Ca^{2+}-permeable AMPA receptors are involved in glutamate-mediated death of immature and mature OLs. Interestingly, activation of metabotropic glutamate receptors (mGluRs) in oligodendrocyte lineage cells attenuates AMPA receptor-induced excitotoxicity. A similar effect is also observed with IGF-1, which protects pre-OLs from glutamate-mediated toxicity. Oligodendroccyte degeneration occurs in a variety of white-matter disorders, and glutamate-mediated excitotoxicity is thought to play an important role in many of these, including periventricular leukomalacia and multiple sclerosis.

OL *N*-methyl-D-aspartate (NMDA) receptor (NR) channels are formed by the NR1, NR2A–NR2C, and NR3A subunits, although based on their pharmacological and electrophysiological properties, channels that include the NR1, NR2C, and NR3A subunits are likely to be the most abundantly expressed. Importantly, differently from AMPA and kainate receptors, which are predominantly found on the cell somata, NMDA receptor channels in oligodendrocytes are mostly clustered on the cellular processes. NMDA receptors are found in OLs from different brain regions, indicating that their expression in this class of cells occurs regardless of their origin. NMDA receptors in OLs are less sensitive to Mg^{2+} block than are neuronal NMDA receptors, therefore, differently from neurons, activation of these channels by glutamate can occur at resting membrane potential and independently of simultaneous AMPA receptor activation.

The physiological function of OL NMDA receptors in the normal brain is still unknown, but their localization on distal domains of cellular processes and in myelin sheets suggests that they might be involved in OL differentiation and/or myelin formation and maintenance. Conversely, NMDA receptors have been associated with glutamate-induced damage in OLs. Ischemia induces NMDA-mediated inward currents in OPCs and in mature OLs, and causes Ca^{2+}-dependent damage to myelinating oligodendrocytes. Interestingly, AMPA and kainate receptor antagonists prevent damage of OL cell bodies, but do not influence deterioration of cellular processes, whereas NMDA receptor antagonists prevent their damage. These findings are also reproduced in chemically induced ischemia in the optic nerve, where a Ca^{2+} overload is observed in myelin. This is partially inhibited by NMDA receptor antagonists, which also attenuate structural changes in the myelin sheets.

Schwann Cells

Based on their morphology, biochemical composition, and specific axonal association, Schwann cells in the mature nervous system can be classified into four groups: myelinating Schwann cells (MSCs), nonmyelinating Schwann cells (NMSCs), perisynaptic Schwann cells (PSCs, or terminal Schwann cells), and satellite cells of peripheral ganglia. These diverse cell types are nonetheless generated from a single neural crest cell. Schwann cell formation involves two transient cell populations, Schwann cell precursors (SCPs) and immature Schwann cells. SCPs are distinguished from migrating neural crest cells by the presence of brain fatty-acid-binding protein (BFABP) and protein zero (P_0), and can be distinguished from immature SCs by the lack of glial fibrillary acidic protein, by cytoplasmic S-100 protein, and by the very distinct feature of axon dependence for their survival. Postnatally, GFAP is suppressed in myelin-forming cells but is retained in NMSCs. Mature MSCs maintain the potential to express GFAP, which can be derepressed once axonal contact is lost. Galactocerebroside (GC), the glycosphingolipid of the myelin sheath, is not only expressed in mature oligodendrocytes and MSCs, but is also found in adult NMSCs *in situ*, although its expression is eventually lost from SCs in culture due to lack of axonal signaling.

Axonal signaling determines fates of immature SCs developing into myelinating, nonmyelinating, and perisynaptic cells. Association with large-diameter axons derives cues to generate myelinating Schwann cells, whereas nonmyelinating SCs ensheathe small-diameter axons. The best characterized SCs remain the MSCs; these can be distinguished from MSCs by lower levels of GFAP, although specific markers are not known for PSCs. Survival and migration of Schwann cells are regulated by axon-derived neuregulin through SC ErbB receptor tyrosine kinases, whereas other factors, such as endothelins, regulate the timing of precursor maturation and Schwann cell generation. Unlike SCPs, SCs can survive in the absence of axons, indicating that

survival regulation changes with the development of autocrine loops in SCs that are absent from SCPs. These signals include factors such as IGF, neurotrophins, and platelet derived growth factor-BB.

Neurotransmitter Receptors and Development

Extracellular ATP is an activity-dependent axonal signal that inhibits the proliferation and differentiation of premyelinating SCs. ATP can activate a number of receptors, broadly divided into two families: the ligand-gated ionotropic (P2X) receptors and the G-protein-coupled metabotropic (P2Y) receptors. ATP released from sensory axons (DRGs) elicits in SCs increased intracellular calcium levels ($[Ca^{2+}]_i$), cAMP response element-binding protein (CREB) phosphorylation, and c-fos and Krox24 induction, but reduces mitogen-stimulated cell proliferation and expression of galactocerebroside and the O4 antigen. ATP hydrolysis ultimately generates adenosine, which also inhibits cell proliferation and which can activate several subtypes of metabotropic adenosine (P1) receptors – A1, A2, and A3. The A1 subgroup of adenosine receptors inhibits adenylyl cyclase while A2 receptors facilitate it. Calcium-linked A1 adenosine receptors are present in PSCs that surround the neuromuscular junction, but PSCs are distinct from SCs associated with axons. Surprisingly, in premyelinating SCs, ATP fails to regulate the extracellular signal-regulated kinase/mitogen-activated protein kinase (ERK/MAPK) cascade, which is involved in regulating cell proliferation and differentiation through changes in either calcium or cAMP. ERK activation is instead specifically mediated by adenosine through $A2_A$ receptors which signal through cAMP, although $A2_B$ is also detectable in premyelinating SCs. Thus, adenosine and ATP inhibit SC proliferation through A2 and P2Y1 receptors, respectively, but only ATP arrests SC differentiation. In contrast, adenosine promotes the differentiation of progenitor cells into myelinating oligodendrocytes in the central nervous system.

Both $GABA_A$ and $GABA_B$ receptors have been found in sciatic nerve SCs. Activation of $GABA_B$ receptors which are negatively coupled to adenylyl cyclase inhibits SC proliferation and reduces RNA and protein levels of myelin proteins P_0, myelin-associated glycoprotein (MAG), and peripheral myelin protein 22 (PMP22). $GABA_A$ receptors can also mediate modulatory activities of neurosteroids (see later, 'Hormone receptors and development'), which can in turn modulate $GABA_B$ receptor expression, with important consequences on SC development.

Immature Schwann cells express M1–M4 muscarinic receptor subtypes, inducing inositol 1,4,5-trisphosphate (IP_3) levels and decreasing cyclic AMP, without evoking calcium release. The same subtypes continue to be expressed in MSCs and NMSCs, although their developmental regulation suggests function during proliferation and differentiation. Furthermore, only quiescent, P_0^+-myelinating SCs responded to 5-hydroxytryptamine (5HT; serotonin) with calcium transients, indicating that postmitotic SCs express serotonin receptors, in particular the $5\text{-}HT_{2A}$ receptor, during differentiation, but their function in development remains uncharacterized.

Neurotransmitter Receptors and SC Signaling

Expression of neurotransmitter receptors in the SC lineage is summarized in Table 1. In response to cAMP analogs, SCs from the sciatic nerve express P2Y receptors which modulate intracellular calcium levels. At the neuromuscular junction (NMJ), PSCs express both ATP and adenosine receptors, and signaling via calcium occurs through P2X, P2Y, and A_1 receptors. However, despite excitatory corelease of ATP with acetylcholine (ACh) in the peripheral nervous system, a role for purinergic receptors in neuromuscular PSCs has not been established. Instead, PSCs are known to respond to acetylcholine with release of calcium from intracellular stores. PSCs in turn secrete neuroactive substances to modulate neurotransmitter release at the NMJ. Because muscarine blocks the induction of GFAP expression in PSCs upon blockade of synaptic activity, whereas ATP is ineffective, it is believed that muscarinic ACh receptors, rather than purinergic receptors, control PSC functions of growth, remodeling, and synapse modulation.

Although glutamate receptors of the NMDA, AMPA, and kainate subtypes have been found in adult guinea pig SCs *in vitro*, other studies describing the application of glutamate agonists in rat sciatic nerve report only partial or inconsistent calcium responses. Interestingly, SCs are one of several sources of purine nucleotides besides synapses and axons: ATP release from neonatal Schwann cell cultures occurs in response to glutamate and AMPA, and is specifically blocked by the AMPA receptor antagonist GYKI 52466 and competitive non-NMDA antagonist 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX).

In addition to glutamate receptors, α1-, α2-, and β-adrenergic receptors are also expressed in guinea pig SC. Isoproterenol, norepinephrine, and epinephrine increase intracellular cAMP levels, which can be synergistically enhanced by vasoactive intestinal peptide (VIP) or secretin. At the neuromuscular junction, substance P induces calcium release from internal SC stores, but when applied as a pretreatment, can also block or reduce muscarine- or ATP-induced calcium responses, indicating an involvement of

endogenous SP in PSC regulation and calcium signaling. The neuropeptides VIP, substance P, and somatostatin also hyperpolarize the SC membrane potential through mechanisms believed to be independent of cAMP.

The role of cAMP elevation in SC development remains controversial. The use of forskolin or cAMP analogs has been found to induce GC, O4 sulfatide, and P_0 expression, accompanied by the downregulation of GFAP and nerve growth factor (NGF) receptor in sciatic nerve preparations. In contrast to GC and P_0, the expression of which is inversely correlated with cell proliferation, cAMP induces O4 expression equally in dividing and nondividing SCs. However, in other studies, forskolin and calcitonin gene-related peptide (CGRP), a potent stimulator of adenylate cyclase in SCs, instead were found to promote SC proliferation in the presence of PDGF. These findings nonetheless led to important applications for cAMP-inducing neuropeptides in injury.

Neurotransmitter Receptors and Immune Function

In addition to its myelinating function, SCs also produce proinflammatory cytokines such as interleukin-1β (IL-1β). While ATP alone has no effect on the synthesis of pro-IL-1β, the posttranslational processing and secretion of IL-1β in stimulated SCs are dependent on the P2X7 receptor and its associated chloride conductance, implicating a function for purinergic receptors in neuroinflammatory disorders.

Hormone Receptors and Development

Insulin-like growth factor-1 promotes the proliferation of SCs through IGF-1 receptors. IGF-1 also induces morphological changes and enhances SC differentiation by inducing P_0 protein levels, even in the absence of cAMP. In co-cultures with DRG, SCs treated with IGF-1 show substantial myelination whereas those deprived of IGF-1 still survive, but fail to myelinate. IGF-1 appears multifaceted; not only does it regulate the proliferation and survival of developing SCs, it also promotes SC motility via the phosphatidylinositol 3-kinase pathway.

Thyroid hormones, which are important in oligodendrocytes, are also required for SC development. Triiodothyronine(T3) increases the incorporation of [^{3}H]thymidine in cultured rat SCs. Nuclear T3 receptor (TR) and T3 binding are found in peripheral glia only during the perinatal period, but adult Schwann cells can also exhibit TR immunoreactivity after nerve injury when they are deprived of axonal contact. TRα1 is the predominant isoform in SCs during sciatic nerve development. The subtypes TRα1 and TRβ1 have been localized in the proximal, peripheral part of the cochlear nerve, where reduced P_0 RNA levels are associated with a hypothyroid state. In addition to its genomic effect on the MBP promoter, T3 also produces rapid effects on immediate-early genes such as *Krox-20/Egr-2*, and this is crucial for peripheral myelin formation, Jun B, and c-fos. Thus, T3 regulates SC function and differentiation through the activation of specific transcription factors.

Neonatal SC cultured from sciatic nerves express intracellular receptors for estrogen and progestins, and the induction of progesterone receptors by estrogens is dependent on neuronal signals. The classical progesterone receptor is expressed in SCs. In primary SC culture, estrogen promotes proliferation only in the presence of elevated intracellular cAMP, while progesterone has no effect on proliferation but does enhance myelin formation. However, when SCs are kept in the integrated environment of the nerve, such as in nerve segments, both estrogen and PROG stimulate SC proliferation without forskolin.

SCs are not only targets for circulating steroids such as estrogens, but they can also synthesize neurosteroids, including pregnenolone, PROG, and their reduced metabolites. The enzymatic complex that converts families of steroid hormones into their 5α- and 3α,5α-reduced metabolites is present in nerves and SCs of the peripheral nervous system (PNS). The concentration of PROG in rodent sciatic nerves is 10 times higher than in plasma. SCs in co-culture with DRG produce significant amounts of PROG and its metabolites, some of which may regulate SC functions through autocrine actions.

Both *in vivo* and *in vitro*, the synthesis of the peripheral myelin proteins P_0 and PMP22 is increased by the treatment with PROG and its derivatives. The testosterone metabolite, dihydrotestosterone (DHT) is also capable of increasing P_0 mRNA levels, as are classical glucocorticoids. Since the androgen receptor (AR) is not found in SCs, studies have shown that DHT could exert a transcriptional activity through the PR instead, and also suggest that effects could additionally be mediated indirectly through neuronal AR. PROG, 5α-DHP, and 3α,5α-THP increase mRNA levels of P_0, whereas only THP can affect PMP22. The regulation of P_0 by THP is due to PR activation, because it is abolished by the PR antagonist mifepristone (RU486), but the stimulation of PMP22 expression by THP can be modulated only by agonists and antagonists of the $GABA_A$ receptor. This indicates the existence of distinct mechanisms for myelin synthesis. Interestingly, PROG, DHP, and THP have also been found to influence the expression of $GABA_B$ subunits in SCs, revealing complex cross-reaction

between the GABAergic receptor system with neuroactive steroids. The expression of hormone receptors is summarized in Table 2.

Schwann Cell Receptors in Disease and Repair

Peripheral nerves possess remarkable regenerative capacity. Following injury, axons and their myelin sheaths distal to the lesion deteriorate by a process known as Wallerian degeneration, leaving behind dividing Schwann cells, which produce, and respond to, a large number of factors crucial for regeneration of nerve fibers. IGF-1 from infiltrating macrophages interacts with the SC IGF-1 receptor, resulting in increased SC mitogenesis, which is important for regrowth following nerve transection. Erythropoietin (Epo) and its receptor are upregulated in nerves and SCs, respectively, during Wallerian degeneration, and recombinant Epo has been found to promote SC proliferation. VIP, CGRP, and peptide histidine isoleucine (PHI) are upregulated in many sensory neurons after injury; CGRP stimulates SC proliferation, while local administration of VIP after nerve transection accelerates early myelination and regrowth of axons, a process which may involve VIP-induced laminin synthesis in SCs. Parathyroid hormone-related peptide (PTHrP) upregulated in DRG and injured sciatic nerve stimulates SC proliferation and migration along the axonal membrane, suggesting that PTHrP promotes SC dedifferentiation, an important requirement for successful nerve regeneration.

PROG and pregnenolone increase remyelination of lesioned peripheral nerves, a process that is inhibited by trilostane (which blocks the conversion of pregnenolone to PROG) or by RU486 (which blocks the PR). Transected nerves reexpress TRα1 in the degenerating part of the nerve. Thyroid hormone is likely to act both on axotomized neurons and SCs during regeneration: local treatment with T3 following sciatic nerve transection results in increased numbers of myelinated axons of larger diameters and which possess thicker myelin sheaths.

Although inflammatory responses are important for SC dedifferentiation and removal of myelin debris, evidence also suggests that prolonged inflammation can inhibit SC proliferation and repair. Delivery of antiinflammatory proopiomelanocortin (POMC) peptides *in vivo* improves nerve regeneration after PNS injury. Stimulation of melanocortin-1 receptors with α-melanocyte-stimulating hormone (α-MSH) inhibits cytokine-induced nuclear factor-kappa B (NF-κB) activity in SCs, and may improve healing after injury. In addition, focal inflammatory reactions alone can cause neuropathic pain despite minor structural damage to axons or glia. The G-protein-coupled receptor GPR7, a member of the somatostatin-3 receptor family resembling the mu, delta, and kappa opioid receptors, is restricted to MSCs that contribute to fibers conveying mechanical perceptions and pain. GPR7 is significantly upregulated in SCs in immune-mediated neuropathies, and may be directly involved in modulating pain. GPR7 thus represents one of many potential targets in the development of therapeutic strategies for inflammatory PNS disorders.

Summary

Oligodendrocytes and Schwann cells are well-characterized glial lineages expressing numerous receptors for neurotransmitters and hormones; these receptors enable oligodendrocytes and Schwann cells to respond to developmental, pathological, and regenerative cues from neuronal synapses, axons, circulating immune cells, and even glial cells. The majority of signals which regulate lineage progression and myelin gene expression during development are also important modulators of disease and repair. An understanding of signal transduction mechanisms utilized by these receptors, and others yet uncharacterized, on these cells remains central to the design of effective therapeutic strategies in injury and myelin-associated disorders.

See also: Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Glial Growth Factors; Neuropeptides: Electrophysiology; Oligodendrocyte Morphology; Oligodendrocyte and Schwann Cell Identification Methods; Schwann Cells and Axon Relationship; Schwann Cell Development; Stem Cells and CNS Repair.

Further Reading

Baker MD (2002) Electrophysiology of mammalian Schwann cells. *Progress in Biophysics & Molecular Biology* 78: 83–103.

Baulieu E-E and Schumacher M (2000) Progesterone as a neuroactive neurosteroid, with special reference to the effect of progesterone on myelination. *Steroids* 65: 605–612.

Belachew S and Gallo V (2004) Synaptic and extrasynaptic neurotransmitter receptors in glial precursors' quest for identity. *Glia* 48: 185–196.

Berger T, Muller T, and Kettenmann H (1995) Developmental regulation of ion channels and receptors on glial cells. *Perspectives on Developmental Neurobiology* 2: 347–356.

Bozzali M and Wrabetz L (2004) Axonal signals and oligodendrocyte differentiation. *Neurochemical Research* 29: 979–988.

Fields RD and Stevens-Graham B (2002) New insights into neuron–glia communication. *Science* 298: 556–562.

Gallo V and Ghiani CA (2000) Glutamate receptors in glia: New cells, new inputs and new functions. *Trends in Pharmacological Sciences* 21: 252–258.

Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

Jung-Testas I and Baulieu EE (1998) Steroid hormone receptors and steroid action in rat glial cells of the central and peripheral nervous system. *The Journal of Steroid Biochemistry and Molecular Biology* 65: 243–251.

Kukley M, Capetillo-Zarate E, and Dietrich D (2007) Vesicular glutamate release from axons in white matter. *Nature Neuroscience* 10: 311–320.

Rodriquez-Waitkus PM, Lafollette AJ, Ng BK, et al. (2003) Steroid hormone signaling between Schwann cells and neurons regulates the rate of myelin synthesis. *Annals of the New York Academy of Sciences* 1007: 340–348.

Schumacher M, Guennoun R, Robert F, et al. (2004) Local synthesis and dual actions of progesterone in the nervous system: Neuroprotection and myelination. *Growth Hormone & IGF Research* 14(supplement A): S18–S33.

Ziskin JL, Nishiyama A, Rubio M, Fukaya M, and Bergles DE (2007) Vesicular release of glutamate from unmyelinate axons in white matter. *Nature Neuroscience* 10: 321–330.

Neurotransmitter Release from Astrocytes

A Volterra, University of Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Astrocyte Excitation Leads to Release of Gliotransmitters

Astrocytes are 'excitable' in the sense that, when activated by internal or external signals, they deliver specific messages to neighboring cells – an activity that has been dubbed 'gliotransmission.' However, astrocytes cannot generate action potentials. Their excitation, which is chemically encoded, can be revealed not by electrophysiology, as in neurons, but by assays of $[Ca^{2+}]i$ transients and oscillations. Two main forms of astrocyte excitation are well documented: one that is generated by chemical signals in neuronal circuits (neuron-dependent excitation) and one that occurs independently of neuronal input (spontaneous excitation). Neuron-dependent excitation of astrocytes has been reported in many brain circuits following nerve fiber stimulation and the release of various transmitters and factors such as glutamate, γ-aminobutyric acid (GABA), acetylcholine, noradrenalin, dopamine, adenosine 5′-triphosphate (ATP), nitric oxide, and brain-derived neurotrophic factor (BDNF). The transfer of information from neurons to glia occurs through the spill-over of synaptically released transmitters, and probably also by direct synaptic-like communication in which astrocytes represent atypical postsynaptic cells. The binding of neurotransmitters to specific receptors in astrocytes generates $[Ca^{2+}]i$ elevations whose properties, including amplitude, frequency, and propagation, are governed by the intrinsic properties of both neuronal inputs and astrocytes. Spontaneous excitation is an unexpected property of astrocytes that has been observed both in acute brain slices and *in vivo*. It occurs in most astrocytes during development and decreases considerably during the first two postnatal weeks, when synaptic circuit formation occurs. However, spontaneous astrocytic $[Ca^{2+}]i$ oscillations seem not to fully disappear with adulthood. They are generated by Ca^{2+} release from internal stores when inositol 1,4,5-trisphosphate (Ins(1,4,5)P3) receptors are activated, with additional influx of extracellular Ca^{2+}.

An important response of astrocytes to their excitation, both by neuronal input and by self-generated stimuli, is the release of gliotransmitters – chemicals that act on adjacent neurons, glial cells, and blood vessels (Figure 1).

What Is a Gliotransmitter?

Gliotransmission was first revealed in 1994 when increases in $[Ca^{2+}]i$ in cultured astrocytes were shown to induce glutamate release followed by neuronal activation. However, a precise definition of gliotransmitters is missing. Here, we propose the following criteria for molecules released by astrocytes: (1) synthesis by and/or storage in the astrocytes; (2) regulated release triggered by physiological stimuli; (3) activation of rapid (milliseconds to seconds) responses in neighboring cells; and (4) a role in physiological processes. Over the years, the number of proposed gliotransmitters meeting these criteria has increased, and their properties have, in part, been unraveled (Table 1). In the following we focus on the mechanisms of their release.

Regulated Gliotransmitter Exocytosis

Exocytosis – the rapid form of transmitter release that is typical of neurons and neurosecretory cells – consists of the fusion of specific, membrane-bound vesicles with the plasma membrane, followed by the quantal discharge of their content. Whether a process of this type also occurs in astrocytes and accounts, at least in part, for gliotransmission has long been debated. Over the years, three main lines of indirect evidence have accumulated in support of exocytotic release:

1. The identification in astrocytes of crucial molecular components of the exocytotic machinery of secretory cells, in particular of the three proteins forming the core complex for membrane fusion, the soluble NSF-associated protein receptor (SNARE) proteins.

In neurons these are (a) the vesicle-associated membrane protein 2 (VAMP2; also called synaptobrevin II), exposed at the vesicle surface and defined as a v-SNARE, (b) the synaptosome-associated protein of 25 kDa (SNAP25), and (c) syntaxin1, with the latter two protruding from the plasmalemma into the cytoplasm and defined as t-(target-)SNAREs.

2. The observation that Ca^{2+}-dependent transmitter release in astrocytes – notably glutamate release – is sensitive to blockers of neuronal exocytosis. Such blockers include (a) the clostridial toxins, tetanus, and botulinum neurotoxin, which are Zn^{2+}-dependent endopeptidases that inactivate specifically the SNARE proteins and (b) the macrolide antibiotic Bafilomycin A1, a selective inhibitor of the vesicular proton pump (v-ATPase), which provides the driving

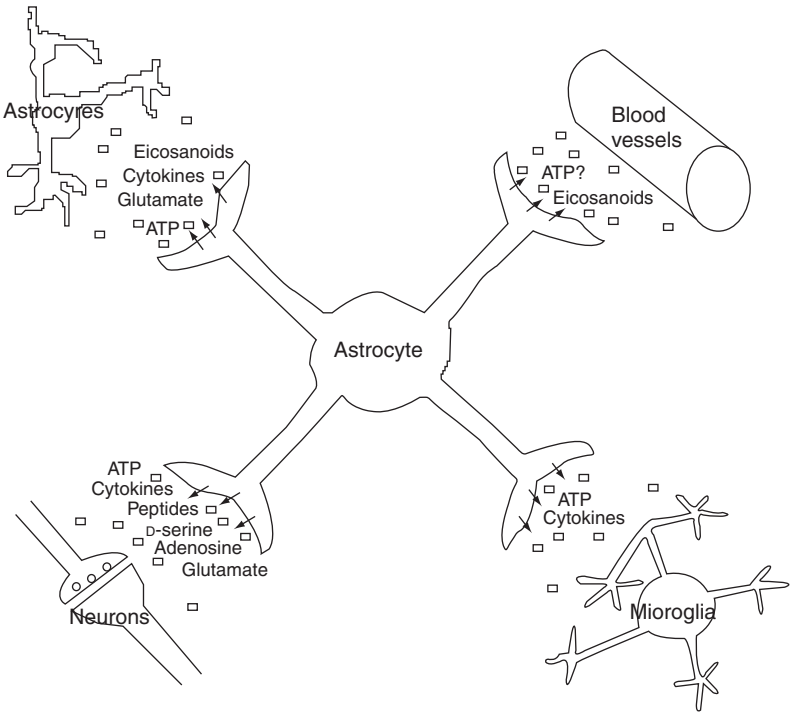

Figure 1 Chemical gliotransmission. Schematic representation of an astrocyte sending chemical signals to various types of cell sitting in its local territory of influence (gliotransmission). As depicted, astrocytes can release a variety of chemical gliotransmitters, some of which act preferentially on a specific cell type. Moreover, astrocytes can release gliotransmitters via several different mechanisms (see text) that probably represent distinct modes of action. Some such modes have been proposed to occur only in pathological conditions, whereas others would underlie physiological regulatory functions of astrocytes, for example, the control of synaptic transmission and of local blood flow.

force for uptake and accumulation of transmitters into synaptic vesicles.

3. The observation that a stimulant of exocytosis from synapses and neurosecretory cells such as PC12, the black widow spider toxin, α-latrotoxin, stimulates glutamate release from the astrocytes.

For a long time the preceding indirect evidence was counterbalanced by the lack of direct evidence for a specific vesicular population competent for regulated exocytosis. Recently, however, a clear synaptic-like microvesicle (SLMV) compartment, which is equipped for the uptake, storage, and release of glutamate, has been identified in adult hippocampal astrocytes by immuno-electron microscopy, and Ca^{2+}-dependent exocytotic fusion events have been documented in astrocyte cultures by several independent approaches, including total internal reflection and confocal fluorescence imaging, membrane capacitance, and electron carbon fiber measurements. Morphologically, astrocytic SLMVs *in situ* appear as clear vesicles resembling in size (30–50 nm) and rounded shape the synaptic vesicles present in glutamatergic nerve terminals. However, astrocytic SLMVs have a much less dense and orderly arrangement. In perisynaptic cell expansions, small, loosely arranged SLMV groups are often distributed just beneath the plasma membrane, opposite neuronal terminals or dendrites, and sometimes within synaptic distance of neuronal *N*-methyl-D-aspartate (NMDA) receptors (NMDARs). These astrocytic SLMVs express proteins that govern glutamate exocytosis – that is, the v-SNARE vesicle-associated membrane protein 3 (VAMP3, also known as cellubrevin), an analog of VAMP2/synaptobrevin II, together with vesicular glutamate transporters (VGLUTs). The three known VGLUT isoforms (VGLUT1–3) are all expressed by astrocytes, although not by all astrocytes. This indicates either that there are VGLUT-positive and VGLUT-negative astrocytes or that various VGLUT isoforms are distributed over distinct subpopulations of astrocytes, as has been observed in neurons. VGLUT-bearing SLMVs have also been identified at the electron microscopic level in cultured astrocytes, where they present a more heterogeneous size distribution (between 30 and 80 nm) and express VAMP2/synaptobrevin II together with VAMP3/cellubrevin.

Functionally, astrocytic exocytosis has peculiar properties when compared to exocytosis at neuronal synapses. Although an accurate analysis has not been yet performed, the stimulus-secretion coupling

Table 1 The repertoire of gliotransmitters

Gliotransmitter	*Cellular storage site*	*Regulated release mechanism*	*Release stimulants and modulators*	*Site of action (receptors)*	*Cell targets and effects*
Glutamate	SLMV	Ca^{2+}-dependent exocytosis	Glutamate, GABA, ATP, PG, TNFα SDF-1α, spontaneous astrocytic excitation	mGluR AMPAR KainateR NMDAR	Astrocytes, neurons (mostly stimulation)
	Cytosol	(activation of channels and/or transporters)			
ATP	DCG?	Ca^{2+}-dependent exocytosis?	ATP, glutamate, dopamine, LPA, thrombin	P2Y; P2X	Astrocytes, microglia, neurons, blood vessel cells? (mostly stimulation)
	Cytosol	(activation of channels and/or transporters)			
Adenosine		Ectonuclotidase-mediated ATP dephosphorylation	ATP, glutamate, dopamine, LPA, thrombin	A1, A2	Neurons (mostly inhibition)
	Cytosol	(activation of channels and/or transporters?)			
D-serine	SLMV?	Ca^{2+}-dependent exocytosis	Glutamate	NMDAR (glycine site)	Neurons (stimulation)
Eicosanoids (PG; HETE)	Not known to be stored	Ca^{2+}-dependent synthesis followed by rapid release	Glutamate, ATP, TNFα, SDF-1α, noradrenaline	Eicosanoid receptors	Astrocytes, blood vessel cells (stimulation, vasodilation/contraction)
Cytokines (TNFα)	Cell surface	Ca^{2+}-dependent, TACE-mediated surface proteolysis	ATP, SDF-1α	TNFα receptors	Astrocytes, neurons (stimulation)
Proteins and peptides (AchBP; ANP; others?)	DCG	Ca^{2+}-dependent exocytosis	Ach (for AchBP)	Ach binding; ANP and other peptide receptors	Neurons (AchBP: inhibition)

Only the agents responding to the four criteria discussed in the text have been included in this table. Several other agents, including taurine and homocysteic acid, are also considered to be gliotransmitters; however, for them the evidence is less complete.
PG, prostaglandin; HETE, 20-hydroxyeicosatetraenoic acid; TNFα, tumor necrosis factor α; AchBP, acetylcholine-binding protein; ANP, atrial natriuretic peptide; SLMV, synaptic-like microvesicle; DCG, dense-core granule; TACE, TNFα-converting enzyme; mGlu-R, metabotropic glutamate receptor; AMPA-R, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid glutamate receptor; NMDA-R, *N*-methyl-D-aspartate glutamate receptor; P2X, P2Y, purinergic 2X and 2Y receptors; SDF-1α, stromal-derived factor 1α, a chemokine; LPA, lysophosphatidic acid.
Adapted from Volterra and Meldolesi (2005) Astrocytes, from brain glue to communication elements: The revolution continues. *Nature Reviews Neuroscience* 6: 626–640.

mechanism in astrocytes seems to be significantly slower (tens of milliseconds) and the Ca^{2+} affinity of the release machinery higher (in the nM range). Gliotransmitter release is triggered not by the action potential-induced opening of specific presynaptic Ca^{2+} channels, but by the activation of G-protein-coupled receptors (GPCRs), with ensuing Ins(1,4,5) P3-induced Ca^{2+} release from the stores of the endoplasmic reticulum. Among GPCR-evoked processes in various cell types, exocytosis of SLMVs in astrocytes appears to be the fastest, which suggests that endoplasmic reticulum cisternae are close to the vesicular release sites. Differences between gliotransmitter and neurotransmitter release might also arise from differences in the molecular components of the exocytotic machinery expressed by neurons and astrocytes, including different isoforms of SNARE proteins. Single-cell polymerase chain reaction after reverse transcription (RT-PCR) studies, performed *in situ* and in freshly isolated astrocytes, and high-resolution immunocytochemistry have documented the expression of four exocytotic proteins: SNAP23, a synaptic t-SNARE, complexin 2, Munc18a, and synaptotagmin IV. By contrast, SNAP25 and the synaptic vesicle proteins synaptotagmin I, synaptophysin, and synaptic vesicle glycoprotein 2 (SV2) were not found. Functionally, VAMP3 can be a substitute for VAMP2, whose expression in tissue astrocytes remains controversial. SNAP23, instead of SNAP25, could enhance the Ca^{2+} sensitivity of

exocytosis. The role of synaptotagmin IV instead of synaptotagmin I, the Ca^{2+} sensor for exocytosis of synaptic vesicles in nerve terminals, cannot be easily deciphered. The contribution of synaptotagmin IV to the Ca^{2+}-sensing mechanism has been recently questioned, but it has been reported that this protein favors a kiss-and-run mechanism – that is, the incomplete exocytotic fusion associated with fast vesicle recycling. Indeed, electron carbon fiber measurements of dopamine-loaded astrocytes suggest that kiss-and-run rather than full fusion could be the main mechanism of astrocytic vesicle fusion in response to physiological stimuli. Each kiss-and-run event would release only a small fraction of the total vesicle content, permitting multiple rounds of evoked releases without vesicle refilling.

Together with SLMVs, larger (100–700 nM), heterogeneous organelles are present in cultured astrocytes, some of which contain secretogranin II, a typical marker of neuronal dense-core secretory granules (DCGs). They also contain ATP. Interestingly, genetic manipulations aimed at preventing formation of the SNARE fusion complex in astrocytes *in vivo* led to reduced extracellular levels of adenosine, a metabolic product of ATP, consistent with the possibility that ATP is released from astrocytes via exocytosis of DCGs and then is rapidly transformed into adenosine by the action of specific extracellular enzymes, the ectonucleotidases. Lysosomes would be a further source of ATP release. Not only nucleotides but also peptides, such as the atrial natriuretic peptide (ANP), could be released from organelles that are distinct and regulated differentially from glutamatergic SLMVs. By contrast, D-serine, a potent co-agonist of glutamate at the NMDAR glycine-binding sites thought to be produced selectively in astrocytes, was recently reported to co-localize with markers of SLMVs, and could, therefore, be co-released from the vesicle population that is responsible for glutamate release. Functionally, this combination could be highly effective in activating NMDARs. These observations, although potentially important in view of the coordinate or antagonistic functional roles of the different gliotransmitters, should be interpreted at the present stage with caution because most of the observations have been made in cell cultures *in vitro* that only partially reproduce the specific gliotransmission properties of astrocytes in their native tissue environment. For example, VGLUT heterogeneity disappears *in vitro*, and all cultured astrocytes express multiple isoforms of the transporter together with various secretion-related proteins that are absent from tissue astrocytes. In addition, the levels of some of these proteins change with time in culture. More studies utilizing preparations *in situ* will therefore be necessary to confirm and clarify the heterogeneity and specific properties of the regulated exocytotic pathways in astrocytes.

Modulation of Gliotransmitter Exocytosis

Exocytotic gliotransmission is probably not a simple consequence of increases in $[Ca^{2+}]i$. Indeed, several distinct stimuli that all induce large $[Ca^{2+}]i$ responses evoke different transmitter release responses. Important regulatory mechanisms seem to exist, and, in some cases, GPCR-induced glutamate release was found to be strongly attenuated by blocking the synthesis of prostaglandins, tumor necrosis factor-α (TNFα), or both.

The role of these mediators is unclear. They could contribute to one or several of the following processes: (1) Ca^{2+} release from the internal stores triggering vesicle fusion; (2) priming of the secretory process itself; and (3) amplification of the whole cascade by autocrine/paracrine loops. Astrocytes lacking TNFα signaling display alterations of GPCR-evoked $[Ca^{2+}]i$ elevations, including reduced and slower Ca^{2+} peaks, suggesting that TNFα contributes to the shaping of GPCR-dependent $[Ca^{2+}]i$ elevations, although it is not strictly indispensable for their generation. It remains to be defined whether the release of prostaglandins and TNFα is part of the stimulus-secretion coupling mechanism triggered by GPCR stimulation or whether these agents are released tonically and set the appropriate basal conditions for efficient $[Ca^{2+}]i$ elevations and glutamate release. Evidence for both modes exists.

Interestingly, gliotransmitter release can be modified by concomitant activation of distinct receptors, with ensuing positive or negative cooperativity. For example, activation of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors, which does not stimulate glutamate release *per se*, was found to potentiate metabotropic glutamate receptor (mGluR)-evoked release and strengthen mGluR-dependent prostaglandin E2 (PGE2) production, possibly through transactivation of epidermal growth factor receptors (ErbB receptors). By contrast, the interaction between mGluRs and α1-noradrenergic receptors is an example of negative cooperativity. The $[Ca^{2+}]i$ elevations induced by the latter fail to stimulate glutamate release, but instead suppress a coincident mGluR-evoked release.

Nonexocytotic Gliotransmitter Release

Molecular transport across the plasma membrane through specialized proteins such as channels and transporters represents a second modality by which hydrophilic agents can be released from cells. Three types of large plasma membrane channel have been

claimed to participate in the nonexocytotic release of gliotransmitters from astrocytes, in particular of ATP and glutamate: (1) volume-regulated anion channels (VRACs), (2) gap junction hemichannels, and (3) purinergic P2X7 receptors.

1. The existence of specific VRACs was deduced from the observation that anionic amino acids, such as glutamate, aspartate, and taurine, are released in a nonvesicular, Ca^{2+}-independent fashion when astrocytes swell, notably in hypotonic media. Recent work raises the possibility that GPCR-evoked $[Ca^{2+}]i$ elevations may also trigger release of anionic amino acids including glutamate through VRACs. One important caveat concerning this release pathway is the fact that the molecular identity of VRACs is still unknown and the existence of such channels is generally inferred via pharmacological studies. However, drugs used as VRAC antagonists are not specific; for instance, they inhibit gap-junction hemichannels (see next listed item). In addition, these drugs affect anion channels in synaptic vesicles and prevent loading of transmitter into the vesicles. Therefore, identification of VRACs based on pure pharmacological criteria is inadequate. An additional, more solid criterion consists of the direct electrophysiological recording of the ion current generated by the passage of anionic gliotransmitters such as glutamate through the plasma membrane channels. Such currents have been recorded, however, only when high glutamate concentrations were added to the patch pipette. Therefore, it remains to be established whether the channels would drive significant glutamate release at the physiological cytosolic concentrations of the amino acid, about 50-fold lower than those used so far.

2. Gap-junction hemichannels, also known as connexons, are large (1–1.2 kDa), nonselective ion channels formed by hexamers of connexin 43 (CX43) that assemble in astrocytes and might function as autonomous permeation pathways instead of being coupled with a hemichannel of an adjacent cell to form a gap junction. The opening probability of the hemichannels, which is very low in astrocytes bathed in physiological media, increases significantly by reducing the concentrations of divalent cations, notably of Ca^{2+}, and was reported to be accompanied by the release of both ATP and glutamate. The involvement of hemichannels is consistent with the suppression of the release of these gliotransmitters by gap junction channel blockers (but see the following) and, in the case of ATP, by the absence of release in a glial cell line that lacks CX43, with restoration after transfection of the connexin.

3. The purinergic P2X7 receptor, which has been proposed to be an additional source of glutamate and ATP release, shares several properties with hemichannels, including an increased open probability at low extracellular Ca^{2+} concentrations. Moreover, recent work has shown that P2X7 receptors (900 Da) are blocked by the so-called gap-junction channel blockers, indicating that use of such agents cannot discriminate between gliotransmitter release via hemichannels and that via P2X7 receptors. This observation calls for a reconsideration of the interpretation of several previous studies. However, the pharmacology of P2X7 receptors has some additional and distinctive features with respect to those of connexin hemichannels: P2X7 is activated by ATP at high concentrations (EC50 $= 300\,\mu M$) and by 3′-*O*-(4-benzoylbenzoyl) ATP (BzATP) at lower concentrations and can be blocked by oxidized ATP but not by suramin. It is not clear whether the high ATP concentrations needed for P2X7 gating can be reached extracellularly in the healthy brain. Alternatively, prolonged activation of P2X7 channels in pathological conditions could lead to the release of cytosolic proteins and cell death.

Plasma membrane transporters may represent additional sources of gliotransmitter release. Among them are: (1) the ATP-binding cassettes (otherwise known as multidrug resistance transporters); (2) the high-affinity plasma membrane glutamate carriers, working in reverse mode; (3) the cystine/glutamate exchanger; and (4) the equilibrative nucleoside transporters (ENTs).

1. The ATP-binding cassettes have been proposed to account for the swelling-induced release of ATP; however, the evidence is still preliminary.

2. Reversal of high-affinity plasma membrane glutamate carriers is unlikely to occur during normal brain function, but probably is a main source of excitotoxic extracellular glutamate accumulation under ischemic conditions. Although most of the glutamate carriers are expressed in astrocytes, the neuronal carriers seem to strongly contribute to ischemic glutamate release: glutamate levels are higher in the neuronal than in the astrocytic cytosol, and this condition favors functioning in reverse mode.

3. Another source of glutamate release could be by the cystine/glutamate exchanger, a candidate for the cytosolic accumulation of cystine that is necessary for the synthesis of glutathione, one of the main endogenous antioxidants. Whether the extracellular concentrations of endogenous cystine are high enough to trigger the exchanger activity and whether its glutamate release occurs from astrocytes are unclear.

4. ENTs passively transport adenosine through cell membranes in either direction depending on the intra- and extracellular adenosine concentration through facilitated diffusion. Isoforms of these transporters

are expressed in astrocytes. Although under physiological conditions adenosine is mainly formed extracellularly, upon release of ATP from the astrocytes followed by ecto-nucleotidase-dependent catabolism, and ENTs mainly act as reuptake system for the nucleoside, under hypoxic conditions adenosine is formed inside the astrocytes, probably via the action of cytosolic nucleotidases, and released via a nonexocytotic process that is inversely controlled by $[Ca^{2+}]i$, that is, is suppressed by $[Ca^{2+}]i$ elevation. ENTs might play a role in such a release, although no direct evidence exists at present.

In conclusion, current evidence indicates that some gliotransmission might occur through channels and transporters; however, many aspects of these processes need to be clarified. Most importantly, it should be determined whether these processes occur under physiological conditions and, in this case, how relevant they are, especially because no mechanisms that account for their specificity and regulation have been identified.

Alternatively, several of these mechanisms could play roles in a variety of pathological conditions. At present, the results of numerous studies in the field are still inconclusive because of poor pharmacology and/or because the available information about the participating transport proteins is insufficient, preventing accurate molecular analysis of localization and function *in situ*.

See also: Adenosine; Adenosine Triphosphate (ATP); Astrocyte: Neurotransmitter and Hormone Receptors; D-Serine: From its Synthesis in Glial Cell to its Action on Synaptic Transmission and Plasticity; Exocytosis: Ca^{2+}-Sensitivity; Gap Junctions and Hemichannels in Glia; Glutamate.

Further Reading

Bezzi P, Carmignoto G, Pasti L, et al. (1998) Prostaglandins stimulate calcium-dependent glutamate release in astrocytes. *Nature* 391: 281–285.

Bezzi P, Gundersen V, Galbete JL, et al. (2004) Astrocytes contain a vesicular compartment that is competent for regulated exocytosis of glutamate. *Nature Neuroscience* 7: 613–620.

Calegari F, Coco S, Taverna E, et al. (1999) A regulated secretory pathway in cultured hippocampal astrocytes. *Journal of Biological Chemistry* 274: 22539–22547.

Chen X, Wang L, Zhou Y, Zheng LH, and Zhou Z (2005) "Kiss-and-run" glutamate secretion in cultured and freshly isolated rat hippocampal astrocytes. *Journal of Neuroscience* 25: 9236–9243.

Coco S, Calegari F, Pravettoni E, et al. (2003) Storage and release of ATP from astrocytes in culture. *Journal of Biological Chemistry* 278: 1354–1362.

Cuénod M, Do KQ, Grandes P, Morino P, and Streit P (1990) Localization and release of homocysteic acid, an excitatory sulfur-containing amino acid. *Journal of Histochemistry and Cytochemistry* 38: 1713–1715.

Domercq M, Brambilla L, Pilati E, et al. (2006) P2Y1 receptor-evoked glutamate exocytosis from astrocytes: Control by tumor necrosis factor alpha and prostaglandins. *Journal of Biological Chemistry* 281: 30684–30696.

Duan S, Anderson CM, Keung EC, et al. (2003) P2X7 receptor-mediated release of excitatory amino acids from astrocytes. *Journal of Neuroscience* 23: 1320–1328.

Jourdain P, Bergersen LH, Bhaukaurally K, et al. (2007) Glutamate exocytosis from astrocytes controls synaptic strength. *Nature Neuroscience* 10: 331–339.

Kimelberg HK, Goderie SK, Higman S, Pang S, and Waniewski RA (1990) Swelling-induced release of glutamate, aspartate, and taurine from astrocyte cultures. *Journal of Neuroscience* 10: 1583–1591.

Kreft M, Stenovec M, Rupnik M, et al. (2004) Properties of Ca^{2+}-dependent exocytosis in cultured astrocytes. *Glia* 46: 437–445.

Montana V, Ni Y, Sunjara V, Hua X, and Parpura V (2004) Vesicular glutamate transporter-dependent glutamate release from astrocytes. *Journal of Neuroscience* 24: 2633–2642.

Mothet JP, Pollegioni L, Ouanounou G, et al. (2005) Glutamate receptor activation triggers a calcium-dependent and SNARE protein-dependent release of the gliotransmitter D-serine. *Proceedings of the National Academy of Sciences of the United States of America.* 102: 5606–5611.

Parpura V, Basarsky TA, Liu F, et al. (1994) Glutamate-mediated astrocyte-neuron signalling. *Nature* 369: 744–747.

Parpura V, Fang Y, Basarsky T, Jahn R, and Haydon PG (1995) Expression of synaptobrevin II, cellubrevin and syntaxin but not SNAP-25 in cultured astrocytes. *FEBS Letters* 377: 489–492.

Szatkovski M, Barbour B, and Attwell D (1990) Non-vesicular release of glutamate from glial cells by reversed electrogenic glutamate uptake. *Nature* 348: 443–446.

Volterra A and Meldolesi J (2005) Astrocytes, from brain glue to communication elements: The revolution continues. *Nature Reviews Neuroscience* 6: 626–640.

Ye ZC, Wyeth MS, Baltan-Tekkok S, and Ransom BR (2002) Functional hemichannels in astrocytes: A novel mechanism of glutamate release. *Journal of Neuroscience* 23: 3588–3596.

Zhang Q, Pangrsic T, Kreft M, et al. (2004) Fusion-related release of glutamate from astrocytes. *Journal of Biological Chemistry* 279: 12724–12733.

Zhang Z, Chen G, Zhou W, et al. (2007) Regulated ATP release from astrocytes through lysosome exocytosis. *Natural cell Biology* 9: 945–953.

Neurotransmitter Release: Synchronous and Asynchronous

R Grantyn and S Kirischuk, Humboldt University Medical School (Charité), Berlin, Germany

© 2009 Published by Elsevier Ltd.

Introduction

More than 50 years ago, Fatt and Katz showed that an action potential (AP) arriving at the frog neuromuscular junction releases the chemical neurotransmitter ACh in multimolecular packages which they called quanta. By that time, the quantum was detected on the basis of the conductance change induced in the follower cell – that is, the postsynaptic response which either occurred as an AP evoked response or as miniature postsynaptic activity ('mini'). Minis were recorded when AP generation was prevented by blocking voltage-activated sodium channels with tetrodotoxin (TTX). Fluctuating evoked responses were observed when lowering the ratio of extracellular Ca^{2+}/Mg^{2+} concentration. In this way, the AP-induced multiquantal end plate potential could gradually be decomposed to postsynaptic potentials containing increasingly fewer quanta, down to one quantum which equaled the mini.

Based on these observations, del Castillo and Katz assumed that the spontaneous asynchronous mini activity is the product of the same release machinery as the quantum in the AP-induced evoked postsynaptic response. In other words, evoked transmitter release was attributed to the synchronous occurrence of the same quanta that constitute mini activity. Thus, the evoked release would merely reflect a massive acceleration of spontaneous release. Whether or not this assumption is correct is the main topic of this article.

Definitions

In the central nervous system, one presynaptic nerve cell typically forms hundreds or thousands of synapses with multiple follower cells. The structural definition of a synapse commonly refers to one specialized contact comprising, on the presynaptic side, between one (some hippocampal inhibitory neurons) and hundreds (calyx of Held) of active zones and, on the postsynaptic side, a corresponding number of postsynaptic specializations that contain the transmitter receptors at high density. The set of synapses formed between two neurons is commonly referred to as a synaptic connection. The number of vesicles released at one synaptic connection varies, similar to the neuromuscular junction, in a quantal manner, with the variance of the release process being smaller the larger the number of active zones contained in the connection between two cells. The total number of quanta released by one presynaptic cell to one follower cell is called the quantum content of a synaptic response, and the total amount of transmitter released is given by the product of the quantum content and the number of transmitter molecules per quantum. Thus, the quantal content can be viewed as the presynaptic indicator of synaptic strength. When it is not possible to record unitary postsynaptic responses (i.e., the response to activation of just one synaptic connection), the frequency of minis is often taken as an indirect estimate of synaptic strength. This again rises the question to what extent observations made with minis are attributable to mechanisms of evoked release.

However, before going into greater detail, it is necessary to agree on a stricter use of the respective terminology because terms such as 'asynchronous release,' 'mini activity,' 'tonic release,' and 'spontaneous release' are often used interchangeably. Furthermore, it is not always clear whether the terms 'asynchronous' or 'spontaneous' refer to what is to be regarded as constitutively active quantal release (i.e., a form of release that can be distinguished by being Ca^{2+} independent).

Because Ca^{2+} concentration in presynaptic terminals and Ca^{2+} sensitivity of the exocytotic machinery are the two most critical determinants of the speed of fusion, Ca^{2+} sensitivity of postsynaptic activity may serve to distinguish between four different modes of the underlying release process: the highly synchronous AP-evoked response (with a K_D of $\sim 10\,\mu M$; data from calyx of Held); the asynchronous delayed response, as induced by high-frequency AP generation (or prolonged presynaptic depolarization) ($K_D \sim 1.5\,\mu M$; data from GABAergic synapses in culture); mini activity ($K_D < 1\,\mu M$); and constitutively active release (no Ca sensitivity). The release underlying the first three modes can, in a broad sense, be referred to as 'regulated,' whereas the constitutively active release is 'unregulated.'

The capacity for constitutively active release in the form of quanta might be a property of many, if not all, cells and is to be expected if the cytosolic concentration of a molecule reaches a critical level. For instance, by using appropriate cellular sensors, Poo and colleagues demonstrated that muscle cells loaded with the neurotransmitter acetylcholine soon start secreting this foreign transmitter in a pulsatile manner reminiscent of minis.

Minis are still to be defined as synaptic responses to transmitter release in the absence of AP generation. As in the original experiments on the frog neuromuscular junction, they are recorded in the presence of TTX. The term 'spontaneous synaptic activity' is commonly reserved for postsynaptic events recorded in the absence of TTX. This activity may therefore reflect all four forms of release.

An interesting question is whether or not minis always reflect a completely stochastic release process. By applying nonlinear methods to quantify any detected determinism in spontaneous release, Walmsley and colleagues presented evidence for the nonrandom nature of spontaneous inhibitory postsynaptic currents in mouse auditory neurons, which suggests that minis might not be very 'spontaneous' in their origin.

In any case, minis should not be equaled with basal constitutively active transmitter release, because in most neurons their frequency displays clear sensitivity to changes in extra- and intracellular Ca^{2+} concentration and to activation of presynaptic membrane receptors with access to intracellular Ca^{2+} signaling pathways at 'resting' conditions. However, mini activity recorded from single synapses under physiological Ca^{2+} concentration may include a fraction of exocytotic events that are Ca^{2+} independent and therefore fall into the category of constitutively active secretion. To isolate this type of mini activity, it might be necessary to inactivate some of the molecular elements that provide Ca^{2+} sensitivity to the fusion process. Also, as an approximation, one may acquire the minis which persist in Ca^{2+}-free extracellular salt solutions. In cortical cultures, only a very small percentage of synapses (~1% according to Murphy and colleagues) are able to sustain spontaneous release in the absence of Ca^{2+}. The very scarce Ca^{2+}-independent spontaneous activity can, however, be enhanced by manipulations such as application of hyperosmolar solutions or α-latrotoxin.

Possible Function of Asynchronous Mini Activity

The prevailing viewpoint is that synchronous evoked and asynchronous mini activity serve distinct functions and have a different origin. It is obvious that the synchronous evoked release is the major means to transmit a precise temporal pattern of presynaptic activity to the follower cell. Therefore, the underlying fusion activity is the most tightly controlled form of release. This tight control generally results in a rather low probability of vesicle release (~0.1–0.3), so that upregulation and also some downregulation can occur. In the course of synapse development, this form of release is achieved with some delay after an initial period during which release is more asynchronous.

Concerning minis, data are accumulating in favor of the idea that this form of asynchronous synaptic activity has its specific function in the maintenance of synapses. The group of Gähwiler and Thompson compared the effect of TTX and clostridial toxins on developing hippocampal neurons in slice cultures from the rodent brain. Because TTX prevented evoked release, whereas treatment with BoNT prohibited any form of release, they were able to examine the relative impact of synchronous evoked versus asynchronous mini release on synapse formation and maintenance. They concluded that maintenance of dendritic spines critically depends on the generation of mini activity, whereas evoked release is not required. Other studies supported the idea that mini activity may have a 'trophic' function, even though in intact brain preparations the mini frequency is typically rather low (typically 1–5 s^{-1}). Considering that one hippocampal neuron carries approximately 50 000 synapses, one has to conclude that only one quantum every 3 h can keep a synapse in place. On the other hand, it might suffice that a dendritic branch or an entire dendrite 'sees' an active presynaptic terminal. Schuman and colleagues showed that NMDA receptor-mediated minis stabilize synaptic function through active suppression of dendritic protein synthesis.

The notion that mini activity represents a specific form of synaptic communication is further supported by the observation that appropriate activation of presynaptic nicotic acetylcholine receptors, metabotropic glutamate receptors, or TrkB receptors (the receptors for the brain-derived neurotrophic factor), among others, can influence the activity of the follower neuron. Of course, the conductance change associated with the release of a single vesicle is quite small (105–180 pS for the AMPA receptor response to a glutamate-containing vesicle and 105–135 pS for a GABA-containing vesicle), but there are synaptic connections in the central nervous system that solely operate by releasing at the most one quantum per AP. Also, experiments have shown that this may be enough to modify the activity of the target cell, especially if the latter has a high input resistance. Regehr and colleagues showed that in cerebellar stellate cells single inhibitory quanta suppressed firing, whereas coincident excitatory quanta reliably and rapidly triggered firing. However, even in the relatively low resistance neurons of the cerebral cortex, nicotinic input (smoking) was shown to influence mini activity to a degree that cortical networks could be affected.

Together, these data justify a closer examination of the origin of minis as opposed to evoked transmitter release.

Different Vesicle Pools for Synchronous and Asynchronous Release?

Since the times of Katz, enormous effort has been focused on the characterization of distinct vesicle pools since it has become obvious that the individual vesicles vary with regard to their distance to the active zones, the site of presynaptic Ca^{2+} channel clusters, the release kinetics of directly monitored individual fusion events, and the incidence of labeling of endocytosed vesicles with styryl dyes. Most important, some vesicles are more readily released than others. The question arose of whether or not minis are preferentially generated by a distinct, more 'reluctant' vesicle pool. This pool would be replenished at rest and release vesicles more slowly than the readily releasable pool of vesicles. However, when vesicles were labeled with different styryl dyes during rest and activity, it became clear that there is a significant amount of cross-depletion of both vesicle pools, which supports the idea that synchronous AP-mediated and asynchronous mini release draw upon the same vesicle pool.

Differential Molecular Control of Synchronous versus Asynchronous Release?

In 1971, Quastel et al. suggested the existence of a final step in exocytosis that is not dependent on Ca^{2+}. Today, the experimental means exist to uncover the basal constitutively active spontaneous release by molecular manipulations. As a rule, evoked release is prohibited by deletion of individual components of the SNARE complex, such as VAMP2, syntaxin 1, or SNAP-25. All this spares some fraction of mini activity. Typically, the remaining mini activity exhibits low Ca^{2+} sensitivity. For instance, experiments with cultured hippocampal neurons revealed that in SNAP-25 knockouts the residual spontaneous release is Ca^{2+} independent, but it can be stimulated by hypertonic solution. It is also known that the key regulator of the SNARE complex, synaptotagmin, is required for Ca^{2+}-dependent synchronous release but not for mini activity. Of particular interest are experiments by Maximov and Südhof showing that synaptotagmin is also essential for the asynchronous delayed release after a stimulus train. This clearly indicates that asynchronous release activity is associated with the synchronous evoked release (Figure 1).

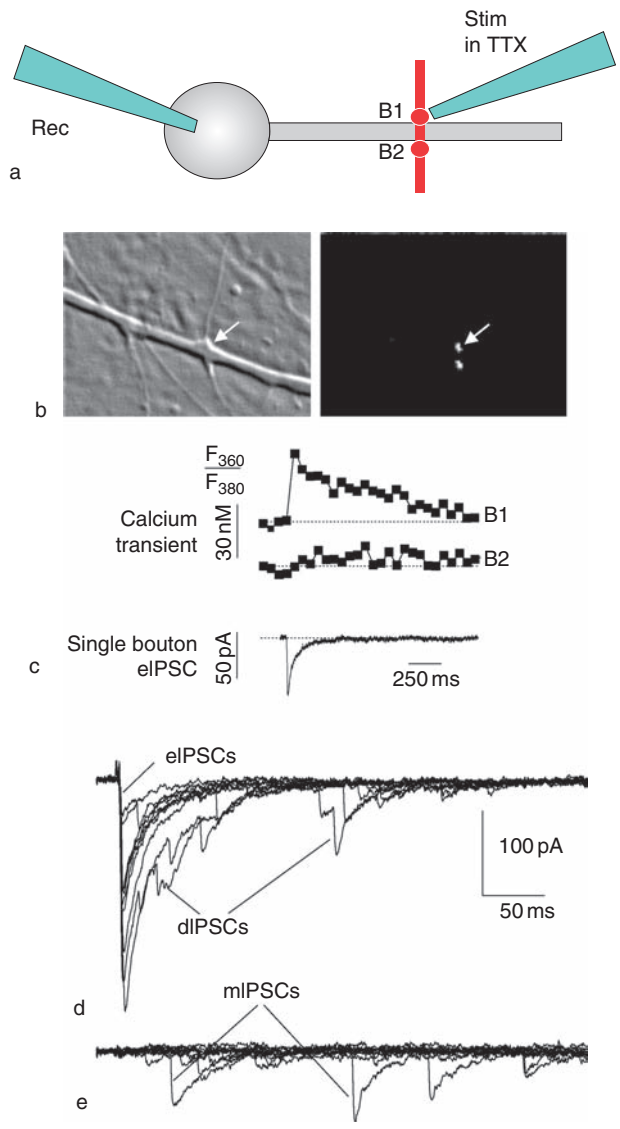

Figure 1 Synchronous and asynchronous forms of synaptic responses to activation of a single FM1-43-labeled synaptic terminal in culture. (a) Experimental scheme. One of two synaptic varicosities formed *en passant* by an axon crossing a dendrite is selected for direct electrical depolarization via a closely apposed patch pipette. The postsynaptic responses are recorded in the whole cell configuration from the nearby soma. (b) Phase contrast and FM1-43 fluorescence images of the respective view field ($10 \times 13.5\,\mu m$). (c) (Top) Presynaptic Ca^{2+} transient as acquired from the respective FM1-43 areas of the stimulated (B1) and nonstimulated (B2) boutons. In this case, Fura-2/AM was loaded for subsequent recording of the bulk Ca^{2+} signal as a measure of presynaptic activation. (Bottom) Corresponding single bouton-activated evoked inhibitory postsynaptic current (IPSC) in response to a very short (0.3 s) depolarizing pulse. (d, e) Superimposed traces of synchronous and asynchronous synaptic activity. The synchronous evoked IPSCs (eIPSCs) and asynchronous delayed IPSCs (dIPSCs), as well as the mini IPSCs (mIPSCs), were recorded in TTX. To elicit both eIPSCs and dIPSCs, the duration of the depolarizing pulse was prolonged to 3 ms.

However, all these data must not necessarily be interpreted as evidence for a totally distinct population of vesicles exclusively dedicated to asynchronous release as opposed to synchronous release. Whether a given vesicle is primed for evoked release or allowed to contribute to miniature activity may depend on the actual state of the fusion switch impacting this vesicle. Results obtained by the groups of Südhof and Rothman suggest that Ca^{2+} elevation enables synaptotagmin to remove a potent fusion-suppressive action of complexin on the SNARE complex. Complexins together with synaptotagmin 1 can be regarded as a fusion clamp that arrests vesicle secretion at rest but promotes synchronous fusion upon activity. According to this model, removal of the synaptotagmin/complexin break should increase the rate of constitutive exocytosis and, vice versa, overexpression of complexins should decrease Ca^{2+}-independent mini activity. The available experimental data are in accord with this proposal. Thus, the actual intracellular Ca^{2+} concentration remains the decisive determinant of the probability of release of any given vesicle. To what extent variable endowment/access of individual vesicles with various isoforms of the SNARE protein VAMP2 or regulatory proteins such as Rab3, RIM-1, and Munc-13 determines whether or not a vesicle is prone to contribute to mini activity remains to be elucidated.

Asynchronous Delayed Release as a Special Form of Evoked Release

After focusing on asynchronous mini release as opposed to synchronous evoked release, we may now regard an 'intermediate' form of release, namely the asynchronous delayed liberation of presynaptic vesicles in response to prolonged or repetitive presynaptic depolarization. Under physiological test conditions, asynchronous delayed release occurs during (or soon after) a burst of presynaptic APs. It is driven by an AP-induced presynaptic Ca^{2+} elevation that usually lasts much longer than an AP. Provided that the level of this Ca^{2+} increase is still high enough to initiate exocytosis (residual Ca^{2+} hypothesis), individual vesicles will be liberated even in the absence of a preceding presynaptic AP. Thus, similar to synchronous evoked release, asynchronous delayed release requires presynaptic activity. However, in contrast to evoked release, the latencies between the arriving APs and vesicle release are (1) in general much longer than a typical synaptic delay and (2) highly variable. The latter means that asynchronous delayed events display a rather stochastic behavior (i.e., they show similarity with mini activity). Indeed, the distribution of inter-event intervals in both cases is well fitted by a single exponential function (an approximate sign of the stochastic nature of the release process). Because the frequency of spontaneous postsynaptic responses is usually quite low, asynchronous delayed instead of truly spontaneous mini responses have often been used to estimate the quantal amplitude of a given postsynaptic response.

Typically, a train of APs is activated/applied to induce asynchronous delayed release. This stimulation protocol depletes the readily releasable pool of presynaptic vesicles and induces large and long-lasting presynaptic Ca^{2+} elevations. Consequently, only newly recruited vesicles mediate asynchronous delayed release. Because vesicle recruitment is a random process, asynchronous delayed activity is composed of monoquantal responses due to the release of single vesicles. Therefore, they are often called 'miniature-like' responses. However, and despite the fact that this approach can considerably speed up the estimation of quantal size, one has to consider and, if possible, avoid stimulation-induced modifications of presynaptic as well as postsynaptic characteristics of transmission (i.e., various expressions of long-term plasticity or postsynaptic receptor desensitization).

The physiological role of asynchronous delayed release is not fully understood. The observation that asynchronous delayed release at immature synapses is usually much stronger than in their mature counterparts led to the suggestion that this form of vesicle fusion may reflect immaturity of the presynaptic release machinery. For instance, the duration of asynchronous delayed currents elicited by 50 Hz trains at GABAergic synapses decreased in the superior colliculus from 2.5 s on postnatal day 0 to 1 s on postnatal day 22. Sensory deprivation in congenitally deaf mice also resulted in much higher asynchronous delayed release than in normally developing counterparts. However, in disagreement with the 'immaturity hypothesis,' many mature synapses display asynchronous delayed release as well, even after a single presynaptic AP. For instance, cerebellar granule cell synapses on Purkinje cells or stellate cells generate after a single spike asynchronous delayed release lasting hundreds of milliseconds. Synaptic terminals in hippocampal cultures release after a single AP vesicle with a time constant of decay of 200 ms.

It should also be considered that the physiological discharge pattern of many mature neurons is not a series of solitary APs, but it often comprises bursts of APs – a condition both necessary and sufficient to initiate asynchronous delayed release. At an interneuron-to-pyramidal cell synapse of the hippocampus, a single AP produces an asynchronous delayed response with a time constant of decay of 59 ms, whereas a train of 10 APs delivered at 50 Hz prolongs

the time constant of asynchronous delayed release to 230 ms. Auditory synapses in the chick brain stem display almost no asynchronous release at frequencies lower than 100 Hz, but at higher frequencies synaptic transmission during AP trains is mainly mediated by asynchronous release. Thus, one may conclude that, independent of the experimental preparation, high-frequency trains of 2–5 presynaptic APs have a good chance to induce an asynchronous delayed response.

Together, all these observations argue against the simple suggestion that asynchronous delayed release results from the imperfectness of release machinery. Instead, it is a complex, functionally relevant phenomenon that depends on a variety of presynaptic factors, including the availability of synaptic vesicles, Ca^{2+} sensitivity of the release machinery, and presynaptic Ca^{2+} signaling. Therefore, an individual presynaptic bouton appears not to function as a single compartment that probabilistically converts individual presynaptic APs into postsynaptic responses but instead as an integrator of many environmental and retrograde signals. This integration would affect the probability of not only evoked release but also all forms of asynchronous release. In addition, asynchronous delayed release provides a mechanism to influence follower cells for a prolonged period, after the presynaptic AP activity has subsided, and the strength of this influence would be determined both by the previous activity period and by newly arriving (retrograde or paracrine) signals.

See also: Action Potential Initiation and Conduction in Axons; Exocytosis: Ca^{2+}-Sensitivity; Fusion Pore; Ion Channel Localization in Axons; Synaptic Vesicles; Vesicle Pools.

Further Reading

Bronk P, Deak F, Wilson MC, Liu X, Sudhof TC, and Kavalali ET (2007) Differential effects of SNAP-25 deletion on Ca^{2+}-dependent and Ca^{2+}-independent neurotransmission. *Journal of Neurophysiology* 98: 794–806.

Carter AG and Regehr WG (2002) Quantal events shape cerebellar interneuron firing. *Nature Neuroscience* 5: 1309–1318.

Dan Y and Poo MM (1992) Quantal transmitter secretion from myocytes loaded with acetylcholine. *Nature* 359: 733–736.

DelCastillo J and Katz B (1954) Quantal components of the end-plate potential. *Journal of Physiology* 124: 560–573.

Giraudo CG, Eng WS, Melia TJ, and Rothman JE (2006) A clamping mechanism involved in SNARE-dependent exocytosis. *Science* 313: 676–680.

Groemer TW and Klingauf J (2007) Synaptic vesicles recycling spontaneously and during activity belong to the same vesicle pool. *Nature Neuroscience* 10: 145–147.

Kirischuk S and Grantyn R (2003) Intraterminal Ca^{2+} concentration and asynchronous transmitter release at single GABAergic boutons in rat collicular cultures. *Journal of Physiology* 548: 753–764.

Korn H and Faber DS (1987) Regulation and significance of probabilistic release mechanisms at central synapses. In: Edelman GM, Gall WE, and Cowan WM (eds.) *Synaptic Function*, pp. 57–108. New York: Wiley.

Leao RN, Leao FN, and Walmsley B (2005) Non-random nature of spontaneous mIPSCs in mouse auditory brainstem neurons revealed by recurrence quantification analysis. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 272: 2551–2559.

Lu T and Trussell LO (2000) Inhibitory transmission mediated by asynchronous transmitter release. *Neuron* 26: 683–694.

Maximov A and Sudhof TC (2005) Autonomous function of synaptotagmin 1 in triggering synchronous release independent of asynchronous release. *Neuron* 48: 547–554.

McKinney RA, Capogna M, Dürr J, Gähwiler BH, and Thompson SM (1999) Miniature synaptic events maintain dendritic spines via AMPA receptor activation. *Nature Neuroscience* 2: 44–49.

Mozhayeva MG, Sara Y, Liu X, and Kavalali ET (2002) Development of vesicle pools during maturation of hippocampal synapses. *Journal of Neuroscience* 22: 654–665.

Otsu Y, Shahrezaei V, Li B, Raymond LA, Delaney KR, and Murphy TH (2004) Competition between phasic and asynchronous release for recovered synaptic vesicles at developing hippocampal autaptic synapses. *Journal of Neuroscience* 24: 420–433.

Prange O and Murphy TH (1999) Correlation of miniature synaptic activity and evoked release probability in cultures of cortical neurons. *Journal of Neuroscience* 19: 6427–6438.

Rahamimoff R and Yaari Y (1973) Delayed release of transmitter at the frog neuromuscular junction. *Journal of Physiology* 228: 241–257.

Sara Y, Virmani T, Deak F, Liu X, and Kavalali ET (2005) An isolated pool of vesicles recycles at rest and drives spontaneous neurotransmission. *Neuron* 45: 563–573.

Serulle Y, Sugimori M, and Llinas RR (2007) Imaging synaptosomal calcium concentration microdomains and vesicle fusion by using total internal reflection fluorescent microscopy. *Proceedings of the National Academy of Sciences of the United States of America* 104: 1697–1702.

Sharma G and Vijayaraghavan S (2003) Modulation of presynaptic store calcium induces release of glutamate and postsynaptic firing. *Neuron* 38: 929–939.

Sutton MA, Ito HT, Cressy P, Kempf C, Woo JC, and Schuman EM (2006) Miniature neurotransmission stabilizes synaptic function via tonic suppression of local dendritic protein synthesis. *Cell* 125: 785–799.

Tang J, Maximov A, Shin OH, Dai H, Rizo J, and Sudhof TC (2006) A complexin/synaptotagmin 1 switch controls fast synaptic vesicle exocytosis. *Cell* 126: 1175–1187.

Washbourne P, Thompson PM, Carta M, et al. (2002) Genetic ablation of the t-SNARE SNAP-25 distinguishes mechanisms of neuroexocytosis. *Nature Neuroscience* 5: 19–26.

Neurotransmitters and Growth Factors: Overview

E Ling, Columbia University, New York, NY, USA
R A Shrikhande, University of California, San Diego, CA, USA
G M Pasinetti, Mount Sinai School of Medicine and the James J. Peters Veteran Affairs Medical Center, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Basics of Neurotransmitters

Neurological research interests over the past 50 years have shifted from the action potential activity of neurons to neuropharmacology. The study of neurotransmitters has been critical in understanding the molecular aspects of neuronal communication and disease. As the chemical messengers of the nervous system, neurotransmitters relay information across synapses via excitation or inhibition of the next neuron or effector tissue. Neurotransmitters can be classified into two broad categories, small-molecule neurotransmitters and neuropeptides. Small-molecule neurotransmitters, such as acetylcholine and the monoamines, are synthesized in the axon terminal of the neuron, and larger neuropeptides, such as somatostatin and vasopressin, are synthesized in the neuron's cell body. Following synthesis, neurotransmitters are transported to the axon terminal, where they are packaged into synaptic vesicles and stored.

An action potential is an electrical signal that is triggered by a stimulus and propagates along an axon. On the arrival of an action potential at the axon terminal, voltage-gated Ca^{2+} channels open to permit an influx of calcium ions. This depolarization of the axon terminal causes the synaptic vesicles to fuse with the presynaptic membrane. The density of voltage-gated Ca^{2+} channels increases with age; the neurotoxicity of the resulting excessive influx of calcium ions has been implicated in the loss of hippocampal neuron viability.

The neurotransmitters are released from the synaptic vesicles by exocytosis and diffuse across the synapse. Some are removed from the synapse by enzyme degradation or transported back to the presynaptic axon terminal, but others reach their specific neurotransmitter receptors on the postsynaptic membrane.

Types and Functions of Major Neurotransmitters with Respect to Aging and Disease

Catecholaminergic System

Dopamine Since Carlsson's research in the 1950s demonstrated that dopamine was a neurotransmitter, this catecholamine has been associated with multiple emotional and age-related behaviors, as well as neurodegenerative diseases. Neurotransmitters of the catecholaminergic system, which are synthesized through a common pathway starting with the amino acid tyrosine, belong to a wider class of substances called biogenic amines. After tyrosine is transported from the liver to catecholaminergic neurons and adrenal medullary cells, it is converted to L-dihydroxyphenylalanine (L-DOPA) by the enzyme tyrosine hydroxylase. Decarboxylation of L-DOPA produces dopamine, a neurotransmitter whose critical roles in movement are impaired in aging and neurodegenerative diseases.

Dopamine employs four major pathways distributed throughout the central and peripheral nervous systems. The nigrostriatal pathway originates in the substantia nigra in the midbrain and regulates the region of the basal ganglia responsible for controlling voluntary movement. The aging brain exhibits decreasing levels of dopamine, with the most reduction found in the putamen of the striatum in brains of patients with Parkinson's disease (PD). Damage to the substantia nigra also occurs in diseased brains, inhibiting the neurons of the thalamocortical circuit and producing motor symptoms such as tremors, bradykinesia, and rigidity. Because PD is a disease of aging, more cases are likely to arise in the future as the older population increases.

Dopamine also has implications in Alzheimer's disease (AD), though its exact roles have not been clearly established. About a third of AD patients eventually develop parkinsonian symptoms and exhibit a reduction in D_2, one of the two G-protein-coupled dopamine receptors, in the caudate and putamen. This is not observed in AD or PD alone, suggesting the existence of different regulatory mechanisms for dopamine receptors in these three conditions. Thus, although the functions of dopamine and the loss of dopaminergic receptors in PD have been well characterized, the implications of dopamine in other age-related diseases remain to be investigated.

Norepinephrine and epinephrine Another neurotransmitter of the catecholaminergic system is produced in synaptic vesicles when dopamine is converted to norepinephrine by dopamine β-hydroxylase. Noradrenergic cells in the central nervous system are most concentrated in the locus ceruleus of the brain stem. However, norepinephrine is more widely used in the sympathetic autonomic nervous system, where its release by postganglionic neurons is mediated by stress. In the adrenal medulla, epinephrine is derived from norepinephrine via phenylethanolamine *N*-methyltransferase. Though epinephrine is primarily released

into the blood as a stress hormone, it can also be used as a neurotransmitter. Norepinephrine and epinephrine's excitatory and inhibitory effects result in increased heart rate, dilation of the bronchi and pupils, and contraction of blood vessels in the skin.

PD often results in failure of the sympathetic nervous system, making patients vulnerable to falls. A common symptom is orthostatic hypotension, caused by impaired constriction of blood vessels due to sympathetic denervation in one or several regions of the heart. Sympathetic denervation occurs in most PD patients with or without accompaniment of orthostatic hypotension. Recent studies show that patients also experience sympathetic denervation in the kidneys and thyroid. Because norepinephrine and epinephrine are derived from dopamine, research on the origins of sympathetic denervation in PD may help in understanding the causes of dopaminergic neuron death.

The adrenergic systems of the central nervous system also have implications in AD pathophysiology. Degeneration of adrenergic neurons in the locus ceruleus expressing C1 adrenergic receptors causes the adrenergic system to compensate for the loss by increasing production of epinephrine in the cerebrospinal fluid (CSF). Such increases in CSF epinephrine levels reportedly correlate with severity of dementia in AD patients. In addition, increased density of α- and β-adrenergic receptors found in postmortem brain tissues of AD patients with symptoms of aggression indicates that aggressive behaviors in AD can be attributed to increased activity of the central adrenergic system.

Aging increases sympathetic activity and reduces norepinephrine uptake due to changes in the morphology of adrenergic receptors. Furthermore, stress causes activation of β1-adrenergic receptors that mediate endothelial cell injury, leading to atherosclerosis. Such chronic events in which stress responses are enhanced make age a risk factor for cardiovascular conditions such as hypertension, ventricular arrhythmias, and heart failure.

Serotonergic System

Loss of serotonin has profound implications in the emotional changes that accompany aging. Serotonin, or 5-hydroxytryptamine (5-HT), is synthesized from the amino acid tryptophan by the enzymes tryptophan hydroxylase and 5-hydroxytryptophan decarboxylase. In the central nervous system, serotonergic neurons are found in the raphe nuclei of the brain stem and project to regions of the hypothalamus and spinal cord; they are also located in the enterochromaffin cells of the gastrointestinal tract. Serotonin monitors a wide range of functions, such as sleep, anxiety, digestion, memory, learning, sexual function, and depression.

Aging is associated with decreases in levels of serotonin type 2A receptor (5-HT_{2A}) binding sites in the hippocampus and frontal and occipital lobes. In postmortem brains, serotonin type 1A receptor (5-HT_{1A}) binding sites were also reduced in the hippocampus and frontal and temporal lobes. A decrease in receptor density is correlated with a higher incidence of depression in older people. Depression is an early symptom of neurodegenerative diseases such as AD and also exacerbates cognitive impairment. Since serotonin transporters remove serotonin from the synapse by facilitating uptake into the presynaptic neuron, many antidepressants block these transporters to maintain endogenous serotonin activity.

Reduction in serotonin levels in the central nervous system and CSF is observed in postmortem AD brains. In addition, neurofibrillary tangles commonly form, and neuronal death occurs in the raphe nuclei. There is also a decrease in the density of cortical 5-HT receptors, which associate with cholinergic nerve terminals in the hippocampus and cerebral cortex, as well as denervation of 5-HT receptors in the neocortex. These factors suggest impairment of the serotonergic system in this disorder, which occurs later than defects of the cholinergic system.

Almost half of all PD patients exhibit mild symptoms of depression, which in turn exacerbates anxiety and contributes to cognitive decline in patients without dementia. Depression possibly also occurs as a result of dopaminergic neuronal death and association of serotonergic circuits with dopaminergic circuits. A product of serotonin catabolism, 5-hydroxyindoleacetic acid, is not only depleted in PD patients' CSF but is also more greatly reduced in PD patients exhibiting symptoms of depression, indicating that serotonin loss causes depression in PD.

Cholinergic System

Acetylcholine is the main neurotransmitter implicated in age-associated memory impairment and AD, the most common neurodegenerative disorder, with both familial and sporadic cases. Synthesis of acetylcholine, the primary neurotransmitter of the cholinergic system, occurs by the action of choline acetyltransferase on choline and acetyl coenzyme A. At the synapse, acetylcholinesterase catabolizes acetylcholine to its metabolites, choline and acetate. Acetylcholine is a major neurotransmitter in the central nervous system, where it is crucial in synaptic plasticity and cognitive functions such as memory and learning; in the peripheral nervous system, it produces excitatory effects in the neuromuscular junction and inhibitory effects in cardiac muscle.

Acetylcholine receptors are divided into two major classes, nicotinic and muscarinic. The ionotropic nicotinic receptors are commonly found in the

neuromuscular junction, where they mediate sodium entry and regulate voluntary muscle contraction. Nicotinic receptors are also active in postganglionic neurons of autonomic ganglia. The metabotropic muscarinic receptors have their activity concentrated in the parasympathetic autonomic nervous system and cardiac muscle, regulating physiological functions such as heart rate, blood pressure, and pupil dilation.

Reduced cholinergic function and long-term potentiation (LTP), caused by decreased acetylcholine production or the impairment of both nicotinic and muscarinic M2 receptors, occur with aging and to a greater degree in dementia. The cholinergic hypothesis of AD states that these defects contribute to cognitive decline and neuronal death in the early stages of the disease; however, it is still undetermined whether reduced cholinergic activity contributes to neuronal death or merely reflects the neurodegenerative processes of dementia.

There is evidence that smoking, during which the cholinergic agonist nicotine stimulates nicotinic receptors in the brain, is neuroprotective against AD. The neuroprotective properties of nicotinic receptors are most associated with those containing the α7-subunit, which protect the spinal cord and hippocampus, and the high-affinity receptors containing α4/β2-subunits, which protect the cortex and striatum. Activation of nicotinic receptors also increases levels of fibroblast growth factor in the striatum, substantia nigra, and hippocampus; brain-derived neurotrophic factor (BDNF) in the striatum; and nerve growth factor (NGF) in the hippocampus. Postmortem studies in older people without AD indicate that smokers have a lower density of senile plaques in hippocampal and cortical regions than do nonsmokers. Senile plaques are aggregates of the 39–43-amino-acid protein β-amyloid (Aβ); it is produced via abnormal processing of the transmembrane amyloid precursor protein (APP), which is normally processed to produce the nonamyloidogenic soluble APP (Figure 1). This connection between cholinergic function and misregulation of APP metabolism has been verified by a reduction in Aβ production on activation of muscarinic M1 and M3 receptors.

PD also leads to impaired cholinergic function in the basal forebrain and cortex and reduced nicotine binding in the substantia nigra. Nicotinic acetylcholine receptor stimulation has been consistently shown

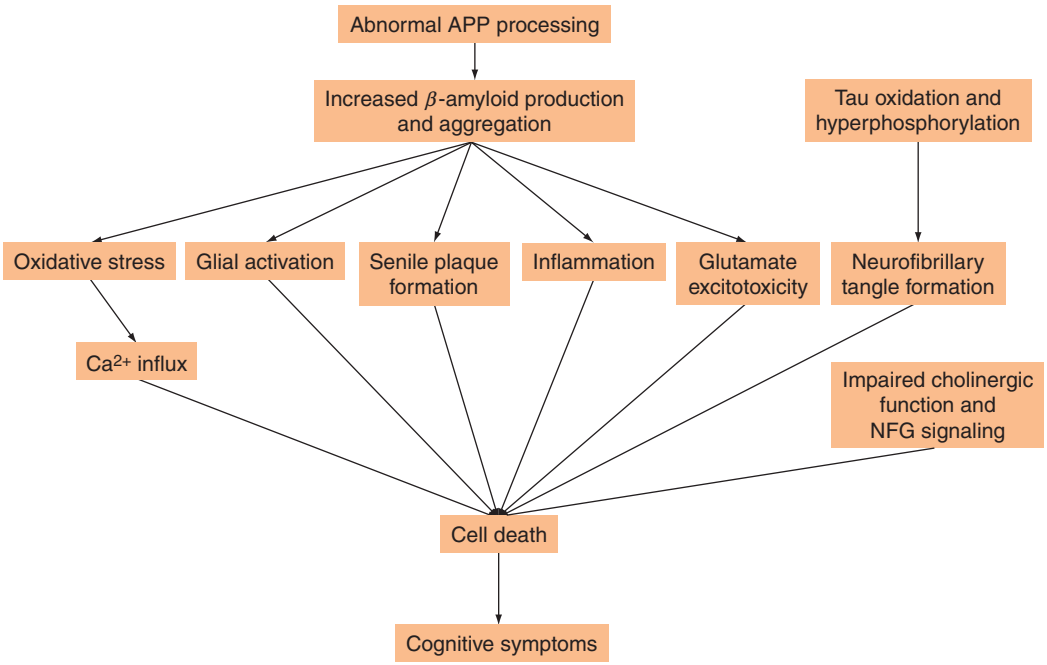

Figure 1 A diagram of the age-associated neuropathology of sporadic Alzheimer's disease (AD). Although neuropathologies associated with the familial forms of AD are not shown, familial and sporadic AD share similar features, including the formation of neurofibrillary tangles and senile plaques. APP, amyloid precursor protein; NGF, nerve growth factor. From Cummings JL (2004) Alzheimer's disease. *New England Journal of Medicine* 351: 56–67; Joyce JN, Murray AM, Hurtig HI, Gottlieb GL, and Trojanowski JQ (1998) Loss of dopamine D2 receptors in Alzheimer's disease with parkinsonism but not Parkinson's or Alzheimer's disease. *Neuropsychopharmacology* 19: 472–480; Mattson MP (2004) Pathways towards and away from Alzheimer's disease. *Nature* 430: 631–639; Meltzer CC, Smith G, DeKosky ST, et al. (1998) Serotonin in aging, late-life depression, and Alzheimer's disease: The emerging role of functional imaging. *Neuropsychopharmacology* 18: 407–430; and Peskind ER, Elrod R, Dobie DJ, et al. (1998) Cerebrospinal fluid epinephrine in Alzheimer's disease and normal aging. *Neuropsychopharmacology* 19: 465–471.

to protect against the disease's cognitive impairments. Administration of muscarinic receptor antagonists produces negative effects on memory.

Amino Acid Neurotransmitters

Glutamate The glutamate excitotoxicity hypothesis has gained considerable interest due to its possible implication in the progression of aging and multiple age-related diseases. Glutamate is the primary excitatory amino acid neurotransmitter secreted by neurons in the central nervous system. Synthesis requires the citric acid cycle intermediate α-ketoglutarate; once released, synaptic glutamate can be converted by glia into glutamine to be metabolized by glutamatergic neurons, be transported back to presynaptic neurons, or bind to postsynaptic receptors. Glutamatergic neurotransmitters most frequently bind to the *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), and kainate receptors, all of which are ionotropic and pentameric; the metabotropic glutamate receptors are classified into groups I, II, and III on the basis of their pharmacological properties. Glutamate's roles in the synaptic plasticity, LTP, and long-term depression of neurons make it a crucial transmitter in learning, memory, and neuronal development.

Glutamate excitotoxicity is primarily attributed to the excessive activation of postsynaptic glutamate receptors, particularly the ionotropic receptors that mediate calcium entry, by ischemic and hyperglycemic events in the hippocampus, striatum, and cortex. Excess intracellular calcium triggers the reduction of adenosine triphosphate and generation of reactive oxygen species that attack DNA and lipid membranes. The glutamate excitotoxicity hypothesis, in which these molecular stresses on the aging neuron accelerate neuronal death, has been implicated in a wide range of neurodegenerative diseases. A probable mechanism of glutamate-associated cell death in AD brains involves cyclooxygenase, an enzyme involved in prostaglandin synthesis and an intermediate in one of the signal transduction pathways initiated by the activation of NMDA receptors. In particular, the inducible isoform cyclooxygenase-2 can potentiate cell cycle abnormalities that promote apoptotic cell death. Lipid peroxidation of Aβ, the key component of senile plaques in AD, is also believed to generate free radicals that inhibit glutamate removal mechanisms, such as the transport of synaptic glutamate back to presynaptic neurons as well as the conversion of glutamate to glutamine.

Impairments in glutamatergic circuits also have implications in the pathogenesis of PD. Studies on primates and rodents with parkinsonism reveal excessive glutamatergic transmission at synapses in the subthalamic nucleus and the substantia nigra. Administering selective group II metabotropic glutamatergic receptor antagonists to the models alleviates akinetic symptoms by reducing glutamate transmission in the subthalamic nucleus. Glutamate has also been suggested to inhibit dopamine release in the striatum through the action of ionotropic glutamate receptors mediated by hydrogen peroxide.

Excessive glutamate has been observed in the CSF of patients with amyotrophic lateral sclerosis (ALS), or Lou Gehrig's disease, of the nonfamilial type. The AMPA receptors on the spinal motor neurons of ALS patients have GluR2 subunits produced from messenger RNA that has been incorrectly processed. The receptors become more permeable to calcium, trigger excessive calcium influx, and trigger motor neuron degeneration. Alternatively, reduced numbers of glutamate transporters on glia could contribute to glutamate excitotoxicity. Glutamate transporter 1 is oxidized and inhibited by a defective superoxide dismutase (SOD1) in the familial type of ALS, preventing glutamate reuptake and prompting excessive activation of calcium-permeable glutamate receptors.

GABA Most of the inhibitory action by the interneurons of the human central nervous system, specifically occurring between Purkinje cells and the vestibular nuclei complex, is performed by γ-aminobutyric acid (GABA), which opposes the excitatory action of dopamine. GABA is synthesized from glutamate via glutamic acid decarboxylase; GABA transaminase (GABA-T) reverses the synthesis. GABA binds to the chloride-permeable ionotropic $GABA_A$ and $GABA_C$ receptors and the metabotropic $GABA_B$ receptors.

Research on GABA activity in senescent animals has presented conflicting reports. In studies on aged macaque monkeys, $GABA_A$ receptor agonists were found to improve visual cortex performance, indicating reduced GABA synthesis, GABAergic receptor density, or GABAergic receptor binding with aging in this region. Other studies showed no difference in the GABAergic system between aged and young rats; however, neurons in the lateral vestibular nucleus of senescent humans and rodents become more receptive to benzodiazepines. GABAergic neurons are also known to interact with glutamatergic and dopaminergic neurons in the nucleus accumbens and the striatum to associate the limbic system with movement. Extracellular GABA increases with aging in these regions, contributing to reduced motor activity; however, these increases may be localized, as neurons in the nucleus accumbens experience greater increases in GABA than do those in the striatum.

Neurotrophic Factors and Neuron Atrophy

Neurotrophic factors are crucial for the survival of neurons, for neurons would be unable to maintain synapses with target tissue without them and would ultimately die via suicide mechanisms, also known as apoptotic cell death. According to the neurotrophic hypothesis, after these proteins are released from a postsynaptic neuron's target cell, they assume the role of signaling molecules that bind to and activate receptors on the presynaptic nerve terminal, producing signals that travel from the axon to the nucleus and induce synthesis of proteins that promote neuronal survival. The limited amount of neurotrophic factors released by target cells drives neurons to compete for these factors. Various neurotrophic factor families influence the differentiation, growth, and development of nerve cells. Ongoing research on neurotrophic factors shows they have possible applications in the treatment of AD, PD, and other age-related neurodegenerative diseases, as well as in the delay of the neuronal death and reduction in brain volume that accompany normal aging.

Importance of Neurotrophic Factors in the Repair of Neurons Damaged by Age-Related Factors or Disease

NGF The first identified neurotrophic factor was NGF, a neurotrophin composed of three subunits (α, β, and γ). It is essential to maintain the survival of basal forebrain cholinergic neurons, sympathetic neurons, and nociceptive sensory neurons. NGF selectively binds to and activates trkA, a high-affinity tyrosine kinase receptor that promotes the expression of choline acetyltransferase and regulates NGF's effects of neuron differentiation and maintenance. The resulting NGF–trkA complex is internalized and transported to the soma by retrograde axonal transport. In addition, NGF and all other neurotrophic factors of the neurotrophin family bind to the low-affinity receptor $p75^{NTR}$, which facilitates binding of NGF to trkA but promotes the death of neurons and mature oligodendrocytes in the absence of trkA. Indeed, expression of cortical trkA receptors is reduced by 50% in patients with mild to moderate and severe AD dementia, while expression of $p75^{NTR}$ receptors remains unchanged; the resulting shift to $p75^{NTR}$ receptor-mediated NGF signaling in the absence of trkA receptors most likely predisposes these cholinergic neurons to apoptosis and possibly contributes to the death of cholinergic neurons during AD.

In vitro treatment of neuron cultures and the central nervous system of rodents with NGF significantly increases the number of neurons, dendrites, and axons, while addition of antibodies to NGF in the same cultures triggers neuronal death. Infusion of recombinant human NGF into the basal forebrain of aged rodents with cognitive impairments prevents neuronal death and stimulates the repair of cholinergic neurons that are damaged in AD. This consequently improves the rodents' performance on memory tasks, making it comparable to that of healthy younger rodents and indicating possible applications of NGF for the treatment of cognitive symptoms in AD. The promising results of recent studies employing gene therapy to treat AD with NGF are discussed further in the section titled 'Gene therapy.' However, since multiple neurotransmitter systems in widespread areas of the brain are affected in AD, the long-term benefits of NGF in its treatment remains to be established.

BDNF BDNF, found in the brain and periphery, promotes neuronal and synaptic growth and development in addition to the maintenance of existing neurons in the cortex, basal forebrain, and hippocampus. Because it aids in neuronal plasticity and regulates hippocampal LTP, an electrophysiological feature of memory, it plays crucial roles in cognition, advanced thinking, learning, and neurogenesis. Similar to NGF, BDNF selectively binds to the trkB receptor and forms the BDNF–trkB complex that is transported as a signal to the soma. In mice, learning impairments caused by reduced trkB-mediated signaling can be reversed with BDNF infusion; exercise also causes upregulation of BDNF.

Reduced levels of BDNF are observed in multiple neurodegenerative disorders, particularly Huntington's disease (HD). HD is an autosomal dominant neurodegenerative disorder in which the mutant huntingtin gene causes motor disabilities and cognitive issues by causing striatal neuron damage. Wild-type huntingtin ordinarily binds to a complex of the repressor element-1 transcription factor (REST) and the neuron restrictive silencer factor (NRSF) in the cytoplasm to prevent their entry into the nucleus and suppression of BDNF gene expression. Mutant huntingtin fails to associate with REST–NRSF, causing accumulation of REST–NRSF in the nucleus and reducing BDNF gene transcription in the cortex, thus reducing levels of BDNF in the basal ganglia.

Glial cell line-derived neurotrophic factor Glial cell line-derived neurotrophic factor (GDNF) belongs to the transforming growth factor-β family of neurotrophic factors. It associates with a receptor consisting of two parts: the GFRα1 subunit is glycosyl–phosphatidylinositol-linked and serves as a binding site, while the c-ret subunit is a receptor tyrosine kinase and participates in signaling. GDNF is critical in the

maintenance of nociceptive neurons as well as dopaminergic and motor neurons, which die during the process of PD. Localized delivery of GDNF to the substantia nigra of rodents with 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridire or 6-hydroxydopamine induces dopaminergic neuron degeneration, protects cell bodies from retrograde cell death, but is unable to preserve axons that project into the striatum. However, GDNF infusion into the striatum protects the nigrostriatal circuit and preserves motor functions affected by the insults.

Ciliary neurotrophic factor Ciliary neurotrophic factor (CNTF) is a cytokine of the interleukin-6 family that is sequestered in adult glia and whose production increases during injury. It is essential in the maintenance of dopaminergic neurons of the substantia nigra and spinal cord motor neurons; when administered to mice with neuromuscular degeneration, it produces neuroprotective effects. Ligand binding of CNTF to the CNTF receptor-α (CNTFRα) subunit triggers signals via the Janus kinase–signal transducers and activators of transcription pathway by the formation of a complex with the subunits glycoprotein 130 and leukemia inhibitory factor receptor-β.

The mutant allele which deters these factors in some humans is related to ALS, characterized by motor neuron degeneration. Patients with a homozygous null mutation in the CNTF gene gain increased susceptibility to the familial form of ALS, which constitutes 5–10% of all cases and which develops approximately 10 years earlier than does ALS in patients without the mutation. This increased susceptibility has also been confirmed in mice that develop motor neuron loss due to the same mutation. In addition, reduced CNTF in glia is observed in the familial form of ALS. On the other hand, in studies of mice exhibiting motor neuron degeneration characteristic of sporadic ALS, which constitutes approximately 90% of cases, inactivation of the CNTF protein has been shown to produce few changes in adult motor neurons; however, mice lacking the gene for CNTFRα have half as many motor neurons and frequently die after birth.

Pharmacology of Age-Related Neurodegenerative Disease Treatment

Increasing knowledge of neurotransmitters and their mechanisms, their pathways, and the actions of agonists and antagonists on specific receptors has led to the development of drugs for treatment of the debilitating symptoms of aging and neurogenerative disorders. L-DOPA, the metabolic precursor of dopamine, is able to pass through the blood–brain barrier although dopamine cannot. Administration of oral L-DOPA to PD patients temporarily delays the progression of parkinsonism. Striatal dopaminergic neurons subsequently use the L-DOPA to produce the dopamine that is deficient in PD brains. To avoid its premature conversion to dopamine in the gut and periphery, L-DOPA is now administered with carbidopa, significantly reducing the dosage of L-DOPA prescribed since its first use in the 1960s. Although L-DOPA successfully reduces parkinsonian symptoms in most instances, some patients experience side effects such as dyskinesia, hypotension, and nausea.

Other antiparkinsonian drugs include dopamine agonists such as pergolide, pramipexole, and ropinirole, which act on dopamine receptors of the substantia nigra. Use of these agonists during the early phases of the disease may reduce the dyskinesia that accompanies prolonged L-DOPA treatment; studies both *in vitro* and *in vivo* have also indicated that these agonists possess neuroprotective properties. However, dopamine agonists and L-DOPA are both suitable initial treatments for PD. Monoamine oxidase B (MAO-B) inhibitors are an alternative form of treatment that prevents the breakdown of dopamine in the brain. Until the recent approval of rasagiline by the Food and Drug Administration, selegiline was the only approved MAO-B inhibitor available for use in the United States. In studies on cultured neurons, rasagiline appears to have a greater neuroprotective effect than does selegiline (Table 1, Figure 2).

Because acetylcholinesterase exacerbates AD by hydrolyzing acetylcholine at the synapse, cholinesterase inhibitors are used to increase acetylcholine levels and stimulate cholinergic activity in the brains of AD patients with the aim of reducing or delaying the onset of psychiatric symptoms. As tacrine, the first cholinesterase inhibitor, was found to be hepatoxic, the three currently used inhibitors are donepezil, galantamine, and rivastigmine. It is rare for more than one cholinesterase inhibitor to be administered at a time; side effects range from nausea to fatigue.

The incidence of PD patients with dementia has prompted research on bifunctional drugs that inhibit acetylcholinesterase without compromising the effects of antiparkinsonian drugs. Research on the compound ladostigil has revealed that it is an effective acetylcholinesterase inhibitor as well as an inhibitor of both isoforms of monoamine oxidase, MAO-A and MAO-B. Ladostigil is currently undergoing clinical trials for AD and PD.

Because of the implications of glutamate excitotoxicity in the etiology of dementia, NMDA receptor antagonists are also used in the treatment of moderate to severe AD. Patients treated with memantine have

Table 1 Approaches to treatment of age-related neurodegenerative diseases

	Alzheimer's disease[a]	*Parkinson's disease*[b]	*Amyotrophic lateral sclerosis*[c]
Pharmacological treatments	Cholinesterase inhibitors (donepezil, galantamine, rivastigmine, ladostigil) NMDA receptor antagonists (memantine)	L-DOPA and carbidopa Dopamine receptor agonists (pergolide, pramipexole, ropinirole) MAO-B inhibitors (selegiline, rasagiline) NMDA receptor antagonists (memantine)	NMDA receptor antagonists (memantine) Glutamate transport activity enhancers (riluzole)
Gene therapy	Enhanced NGF production	Enhanced GDNF production Silencing genes for α-synuclein	Enhanced VEGF production
Transplantation	Neural stem cells to produce cholinergic neurons	Neural stem cells to produce dopaminergic neurons	ES-cell derived glia to produce depleted trophic factors

NMDA, *N*-methyl-D-aspartate; L-DOPA, L-dihydroxyphenylalanine; MAO-B, monoamine oxidase B; NGF, nerve growth factor; GDNF, glial cell line-derived neurotrophic factor; VEGF, vascular endothelial growth factor; ES, embryonic stem.

[a]Data from Cenciarelli C, Budoni M, Mercanti D, et al. (2006) *In vitro* analysis of mouse neural stem cells genetically modified to stably express human NGF by a novel multigenic viral expression system. *Neurological Research* 28: 505–512; Chauhan NB and Siegel GJ (2003) Effect of PPF and ALCAR on the induction of NGF- and p75-mRNA and on APP processing in Tg2576 brain. *Neurochemistry International* 43: 225–233; Cummings JL (2004) Alzheimer's disease. *New England Journal of Medicine* 351: 56–67; Tuszynski MH, Thal L, Pay M, et al. (2005) A phase 1 clinical trial of nerve growth factor gene therapy for Alzheimer disease. *Nature Medicine* 11: 551–555; Youdim MB and Bakhle YS (2006) Monoamine oxidase: Isoforms and inhibitors in Parkinson's disease and depressive illness. *British Journal of Pharmacology* 147: S287–S296.

[b]Data from Ahlskog J E (2003) Slowing Parkinson's disease progression: Recent dopamine agonist trials. *Neurology* 60: 381–389; Bjorklund LM, Sanchez-Pernaute R, Chung S, et al. (2002) Embryonic stem cells develop into functional dopaminergic neurons after transplantation in a Parkinson rat model. *Proceedings of the National Academy of Sciences of the United States of America* 99: 2344–2349; Bodner RA, Outeiro TF, Altmann S, et al. (2006) Pharmacological promotion of inclusion formation: A therapeutic approach for Huntington's and Parkinson's diseases. *Proceedings of the National Academy of Sciences of the United States of America* 103: 4246–4251; Jim JH, Auerbach JM, Rodriguez-Gomez JA, et al. (2002) Dopamine neurons derived from embryonic stem cells function in an animal model of Parkinson's disease. *Nature* 418: 50–56; Kirik D, Georgievska B, and Bjorklund A (2004) Localized striatal delivery of GDNF as a treatment for Parkinson disease. *Nature Neuroscience* 7: 105–110; Schapira AH and Olanow CW (2004) Neuroprotection in Parkinson disease: Mysteries, myths, and misconceptions. *Journal of the American Medical Association* 291: 358–364; Svendsen CN and Langston JW (2004) Stem cells for Parkinson disease and ALS: Replacement or protection? *Nature Medicine* 10: 224–225; Youdim MB and Bakhle YS (2006) Monoamine oxidase: Isoforms and inhibitors in Parkinson's disease and depressive illness. *British Journal of Pharmacology* 147: S287–S296.

[c]Data from Azzouz M, Ralph GS, Storkebaum E, et al. (2004) VEGF delivery with retrogradely transported lentivector prolongs survival in a mouse ALS model. *Nature* 429: 413–417; Raoul C, Abbas-Terki T, Bensadoun JC, et al. (2005) Lentiviral-mediated silencing of SOD1 through RNA interference retards disease onset and progression in a mouse model of ALS. *Nature Medicine* 11: 423–428; Svendsen CN and Langston JW (2004) Stem cells for Parkinson disease and ALS: Replacement or protection? *Nature Medicine* 10: 224–225.

exhibited moderate cognitive improvements on the 'severe impairment battery' and the 'activities of daily living inventory' and had better performance on daily functions. Memantine may also be beneficial in treating vascular dementia as well as less severe forms of AD. Current research focuses on the effectiveness of adding NO groups to memantine to prevent glutamate excitotoxicity (Table 1).

Drugs that interfere with glutamate transmission are also used to treat ALS. Riluzole reduces motor neuron degradation by enhancing glutamate transport activity in the spinal cord; however, this effect seems to be reduced as the disease progresses into its late stages, possibly due to the impaired glutamate uptake capacity of glia in severe ALS. Memantine has also been demonstrated to have protective effects on mouse models of ALS (Table 1, Figure 3).

Gene Therapy

Recent advances in gene therapy have led to possible treatments that help slow the progression of certain diseases associated with the brain and aging. Studies on primate models as well as recent phase I clinical trials on AD patients have shown positive results for *ex vivo* NGF gene delivery in the treatment of AD. Genetically altered fibroblasts that express NGF were injected into the cholinergic nucleus basalis of eight patients with familial AD. Increased cholinergic axon projections into the cortex and increased glucose uptake by these neurons were observed in the patients' brains. During the next 2 years, the patients' average rate of cognitive decline was reduced by as much as half their rate before treatment, based on their performance on the mini-mental state

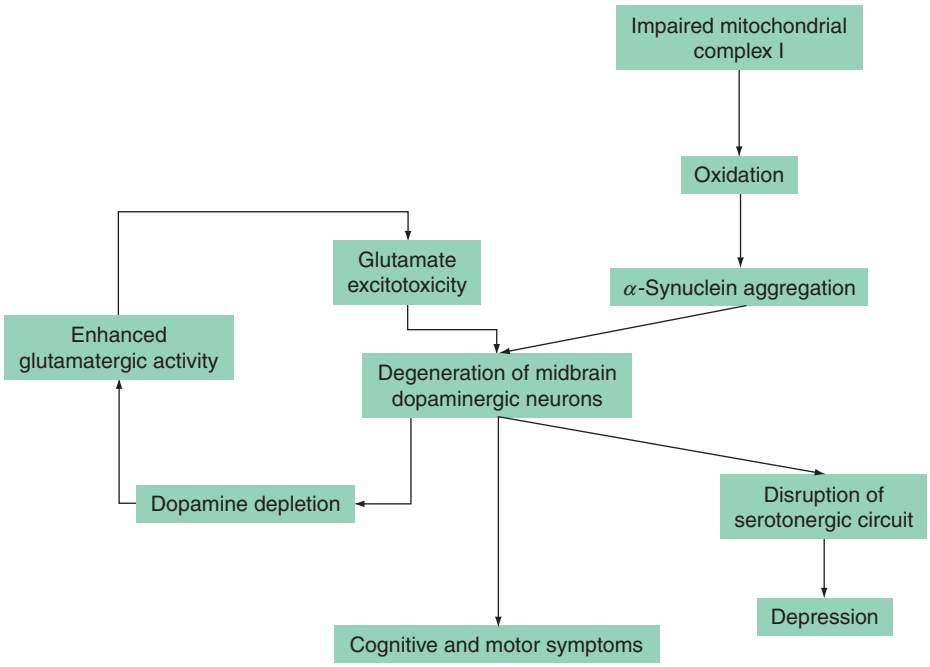

Figure 2 A diagram of the age-associated neuropathology of sporadic Parkinson's disease (PD). Although the causes and neuropathologies of the familial forms of PD are not shown, including mutations in α-synuclein, studies on genetic causes have enabled researchers to produce genetic models of PD. From Cummings JL and Masterman DL (1999) Depression in patients with Parkinson's disease. *International Journal of Geriatric Psychiatry* 4: 711–718; Dawson TM and Dawson VL (2003) Molecular pathways of neurodegeneration in Parkinson's disease. *Science* 302: 819–822; Lange KW, Kornhuber J, and Riederer P (1997) Dopamine/glutamate interactions in Parkinson's disease. *Neuroscience and Biobehavioral Reviews* 21: 393–400; and Sapru MK, Yates JW, Hogan S, et al. (2006) Silencing of human alpha-synuclein *in vitro* and in rat brain using lentiviral-mediated RNAi. *Experimental Neurology* 198: 382–390.

examination and the cognitive subcomponent of the AD assessment scale.

A parallel to the gene therapy described above for PD involves GDNF gene delivery, which helps maintain dopaminergic function in the striatum. Delivery of the gene for glutamic acid decarboxylase into the brains of animal models of PD has also been found to stimulate GABAergic neurons and reduce activity in the subthalamic nucleus, helping attenuate motor symptoms; clinical trials are under way in humans. A recently developed technique involving RNA interference-based gene therapy was able to silence the gene for α-synuclein, the primary component of Lewy bodies, in dopaminergic neurons of rat models. Future research will investigate the effects of gene silencing therapy on cognition in animal models of PD.

In SOD1 transgenic mice as mouse models for ALS, vascular endothelial growth factor gene therapy using a lentiviral vector has produced protective effects in motor neurons of the spine and brain stem, slowing the progression of ALS. RNA interference enables silencing of copies of the mutated SOD1 gene and delivery of functional copies of the gene, indicating that this technique holds much promise in the treatments of various neurodegenerative diseases. Other techniques aim to increase levels of functional dystrophin via injection of antisense oligoribonucleotides into muscle cells of *mdx* transgenic mice, enabling a nonsense mutation in exon 23 of the dystrophin gene to be skipped.

Mouse models of HD that were injected with viral vectors containing the GDNF gene had reduced neuronal death and showed improved performance on tests measuring coordination and balance. A phase I clinical trial has also demonstrated the effectiveness of delivering a baby hamster kidney cell line modified to produce CNTF into striatal neurons.

Stem Cells and Their Application in the Treatment of Neurodegenerative Diseases

Stem cells have potential in the treatment of diseases associated with aging and the brain. Cell replacement therapies seek to generate viable neurons that are able to produce neurotransmitters depleted in various neurodegenerative diseases. Human neural stem cells grafted into rodent models differentiate into cholinergic neurons in the spinal cord but have not been observed in the cortex or the hippocampus. However, because more neural stem cells differentiate into glia and neurons, further research is required to determine

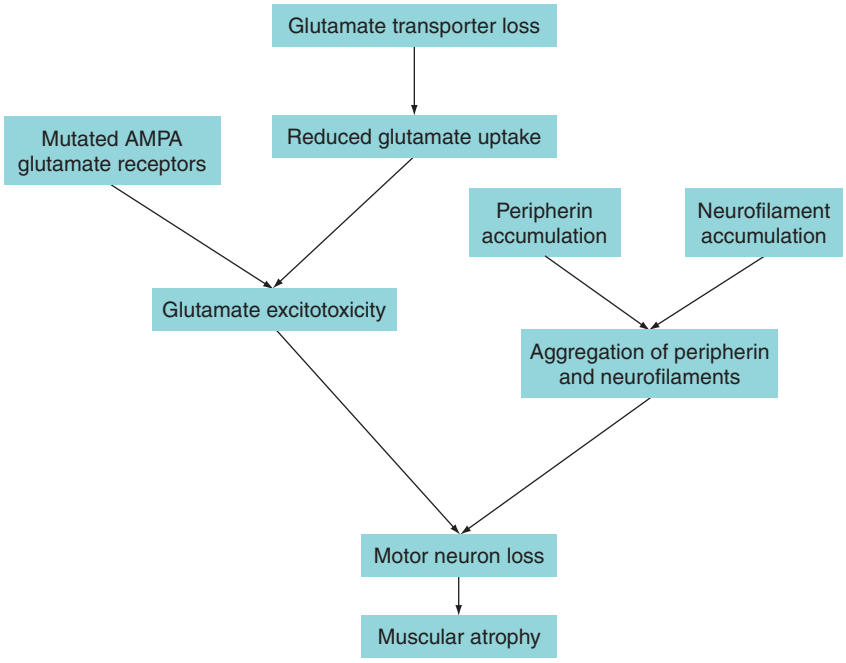

Figure 3 A diagram of the age-associated neuropathology of sporadic amyotrophic lateral sclerosis (ALS). Although neuropathologies associated with the familial forms of ALS are not shown, familial and sporadic ALS share similar features, in particular, motor neuron degeneration. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid. From Dunlop J, Beal McIlvain H, She Y, and Howland DS (2003) Impaired spinal cord glutamate transport capacity and reduced sensitivity to riluzole in a transgenic superoxide dismutase mutant rat model of amyotrophic lateral sclerosis. *Journal of Neuroscience* 23: 1688–1696; Rowland LP and Shneider NA (2001) Amyotrophic lateral sclerosis. *New England Journal of Medicine* 344: 1688–1700; and Trotti D, Rolfs A, Danbolt NC, Brown RH Jr., and Hediger MA (1999) SOD1 mutants linked to amyotrophic lateral sclerosis selectively inactivate a glial glutamate transporter. *Nature Neuroscience* 2: 427–433.

the efficacy of neural stem cells in the treatment of AD. An interesting application of embryonic stem (ES) cells in familial AD research is the generation of postmitotic AD cells for *in vitro* study without the need for primary culture neurons.

Current nonpharmacological alternatives to L-DOPA in PD treatment include transplantation of dopaminergic neurons from the fetal midbrain. Although this method attenuates parkinsonian symptoms, the effect is temporary, and development of more efficient techniques is necessary. In studies on rodent models, midbrain dopaminergic neurons derived from ES cells have been shown to survive following grafting into the striatum. The recent identification of genes that promote ES cell maturation to midbrain dopaminergic neurons, including Nurr1 and Pitx3, paves the way for further studies of ES cells and their long-term effectiveness in treating PD.

Research on rodent models of ALS has shown that axons of motor neurons derived from ES cells can extend to the ventral root, though the axons' interaction with muscle tissue has yet to be observed. Additional studies on these models have revealed the benefits of ES cell-derived glial cells, which produce trophic factors that protect motor neurons from the effects of mutations in the SOD1 gene. Therefore, applications of ES cells in the treatment of ALS range from generating replacement cells to generating glia that promote the growth and repair of motor neurons, providing a wide range of options for study.

However, the ethical implications of ES cell research for the possible treatment of neurodegenerative diseases, particularly those associated with age, are widely debated. Though stem cells hold the promise of replacing damaged or dead neurons by differentiating into desired cell types, considerable time, funding, and adequate cell lines – their source is another issue – must be devoted to research to produce safe and effective treatments. Both the well-being of the aging baby boomer population and the resources invested in ES cell research are to be considered as research probes deeper into the underlying mechanisms of these disorders.

Implications of Lifestyle Factors for Aging and Neurodegenerative Disorders

Although genetic causes for several disorders related to aging and the brain, including HD, ALS, and the familial forms of AD and PD, have been identified, the sporadic and more common forms of disease may be attributed to multiple factors. Lifestyle factors such as nutrition and exercise could notably influence

one's risk for dementia and AD as our neurons are particularly receptive to our daily activities. Studies involving caloric intake on Tg2576 mouse models of AD revealed that generation of Aβ protein is enhanced in diets high in saturated fat but attenuated in caloric restriction diets low in carbohydrates. Furthermore, epidemiological evidence indicates that type 2 diabetes and obesity escalate the risk of AD more than fourfold, suggesting a strong relationship between diet and brain disorders. As a result, rising incidences of obesity in the American population in the twenty-first century could drastically increase the number of people diagnosed with AD. A study of type 2 diabetes in Tg2576 mice has also revealed that the elevated insulin levels in the blood coincided with increased Aβ production and senile plaque formation. Increases in levels of BDNF in the striatum, cortex, and hippocampus – areas of the brain affected in AD and PD – that have been observed in mouse models may explain the beneficial effects of caloric restriction.

See also: Acetylcholinesterase Inhibitors and Alzheimer's Disease; Acetylcholinesterase; Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Brain Volume: Age-Related Changes; Brain Adrenergic Neurons; Cholinergic System; Cholinergic Pathways in CNS; Dopamine; Dopamine – CNS Pathways and Neurophysiology; Dopamine Receptors and Antipsychotic Drugs in Health and Disease; Dopaminergic Agonists and L-DOPA; Excitotoxicity in Neurodegenerative Disease; Exercise: Optimizing Function and Survival at the Cellular Level; GABA-A Receptors and Disease; Gamma-Aminobutyric Acid (GABA); Gene Therapy: Genetically Modified Cells; Gene Therapy: Direct Viral Delivery; Gene Therapy and Protection from Stress-Induced Brain Damage; Glutamatergic and Gabaergic Systems; Growth Factors: Neuronal Atrophy; Monoamines; Neural Stem Cells and CNS Diseases; Neural Stem Cells: Adult Neurogenesis; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Nicotinic Acetylcholine Receptors; NMDA Receptors and Development; NMDA Receptors and Disease + C464; Norepinephrine: Adrenergic Receptors; Parasitic Diseases: Nervous System Effects; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors; Stem Cells and CNS Repair; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Ahlskog JE (2003) Slowing Parkinson's disease progression: Recent dopamine agonist trials. *Neurology* 60: 381–389.

Azzouz M, Ralph GS, Storkebaum E, et al. (2004) VEGF delivery with retrogradely transported lentivector prolongs survival in a mouse ALS model. *Nature* 429: 413–417.

Bjorklund LM, Sanchez-Pernaute R, Chung S, et al. (2002) Embryonic stem cells develop into functional dopaminergic neurons after transplantation in a Parkinson rat model. *Proceedings of the National Academy of Sciences of the United States of America* 99: 2344–2349.

Bodner RA, Outeiro TF, Altmann S, et al. (2006) Pharmacological promotion of inclusion formation: A therapeutic approach for Huntington's and Parkinson's diseases. *Proceedings of the National Academy of Sciences of the United States of America* 103: 4246–4251.

Cenciarelli C, Budoni M, Mercanti D, et al. (2006) *In vitro* analysis of mouse neural stem cells genetically modified to stably express human NGF by a novel multigenic viral expression system. *Neurological Research* 28: 505–512.

Chauhan NB and Siegel GJ (2003) Effect of PPF and ALCAR on the induction of NGF- and p75-mRNA and on APP processing in Tg2576 brain. *Neurochemistry International* 43: 225–233.

Cummings JL (2004) Alzheimer's disease. *New England Journal of Medicine* 351: 56–67.

Cummings JL and Masterman DL (1999) Depression in patients with Parkinson's disease. *International Journal of Geriatric Psychiatry* 4: 711–718.

Dawson TM and Dawson VL (2003) Molecular pathways of neurodegeneration in Parkinson's disease. *Science* 302: 819–822.

Dunlop J, Beal McIlvain H, She Y, and Howland DS (2003) Impaired spinal cord glutamate transport capacity and reduced sensitivity to riluzole in a transgenic superoxide dismutase mutant rat model of amyotrophic lateral sclerosis. *Journal of Neuroscience* 23: 1688–1696.

Esch T, Stefano GB, Fricchione GL, and Benson H (2002) Stress in cardiovascular diseases. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* 8: 93–101.

Goldstein DS, Holmes CS, Dendi R, Bruce SR, and Li ST (2002) Orthostatic hypotension from sympathetic denervation in Parkinson's disease. *Neurology* 58: 1247–1255.

Jim JH, Auerbach JM, Rodriguez-Gomez JA, et al. (2002) Dopamine neurons derived from embryonic stem cells function in an animal model of Parkinson's disease. *Nature* 418: 50–56.

Joyce JN, Murray AM, Hurtig HI, Gottlieb GL, and Trojanowski JQ (1998) Loss of dopamine D2 receptors in Alzheimer's disease with parkinsonism but not Parkinson's or Alzheimer's disease. *Neuropsychopharmacology* 19: 472–480.

Kandel ER, Schwartz JH, and Jessell TM (eds.) (2000) *Principles of Neural Science,* 4th edn. New York: McGraw-Hill.

Kihara T, Shimohama S, Sawada H, et al. (1997) Nicotinic receptor stimulation protects neurons against beta-amyloid toxicity. *Annals of Neurology* 42: 159–163.

Kirik D, Georgievska B, and Bjorklund A (2004) Localized striatal delivery of GDNF as a treatment for Parkinson disease. *Nature Neuroscience* 7: 105–110.

Kunst CB (2004) Complex genetics of amyotrophic lateral sclerosis. *American Journal of Human Genetics* 75: 933–947.

Lange KW, Kornhuber J, and Riederer P (1997) Dopamine/glutamate interactions in Parkinson's disease. *Neuroscience and Biobehavioral Reviews* 21: 393–400.

Mattson MP (2004) Pathways towards and away from Alzheimer's disease. *Nature* 430: 631–639.

Meltzer CC, Smith G, DeKosky ST, et al. (1998) Serotonin in aging, late-life depression, and Alzheimer's disease: The emerging role of functional imaging. *Neuropsychopharmacology* 18: 407–430.

Mirjany M, Ho L, and Pasinetti GM (2002) Role of cyclooxygenase-2 in neuronal cell cycle activity and glutamate-mediated excitotoxicity. *Journal of Pharmacology and Experimental Therapeutics* 301: 494–500.

Pang TY, Stam NC, Nithianantharajah J, Howard ML, and Hannan AJ (2006) Differential effects of voluntary physical exercise on behavioral and brain-derived neurotrophic factor expression deficits in Huntington's disease transgenic mice. *Neuroscience* 141: 569–584.

Pasinetti GM, Zhao Z, Qin W, et al. (2006) Caloric intake and Alzheimer's disease: Experimental approaches and therapeutic implications. *Journal of Biological Chemistry* 281: 21745–21754.

Peskind ER, Elrod R, Dobie DJ, et al. (1998) Cerebrospinal fluid epinephrine in Alzheimer's disease and normal aging. *Neuropsychopharmacology* 19: 465–471.

Picciotto MR and Zoli M (2002) Nicotinic receptors in aging and dementia. *Journal of Neurobiology* 53: 641–655.

Raoul C, Abbas-Terki T, Bensadoun JC, et al. (2005) Lentiviral-mediated silencing of SOD1 through RNA interference retards disease onset and progression in a mouse model of ALS. *Nature Medicine* 11: 423–428.

Rowland LP and Shneider NA (2001) Amyotrophic lateral sclerosis. *New England Journal of Medicine* 344: 1688–1700.

Sapru MK, Yates JW, Hogan S, et al. (2006) Silencing of human alpha-synuclein in vitro and in rat brain using lentiviral-mediated RNAi. *Experimental Neurology* 198: 382–390.

Schapira AH and Olanow CW (2004) Neuroprotection in Parkinson disease: Mysteries, myths, and misconceptions. *Journal of the American Medical Association* 291: 358–364.

Svendsen CN and Langston JW (2004) Stem cells for Parkinson disease and ALS: Replacement or protection? *Nature Medicine* 10: 224–225.

Trotti D, Rolfs A, Danbolt NC, Brown RH Jr., and Hediger MA (1999) SOD1 mutants linked to amyotrophic lateral sclerosis selectively inactivate a glial glutamate transporter. *Nature Neuroscience* 2: 427–433.

Tuszynski MH, Thal L, Pay M, et al. (2005) A phase 1 clinical trial of nerve growth factor gene therapy for Alzheimer disease. *Nature Medicine* 11: 551–555.

Youdim MB and Bakhle YS (2006) Monoamine oxidase: Isoforms and inhibitors in Parkinson's disease and depressive illness. *British Journal of Pharmacology* 147: S287–S296.

Zuccato C, Tartari M, Crotti A, et al. (2003) Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nature Genetics* 35: 76–83.

Relevant Website

http://www.nia.nih.gov – National Institute on Aging.

Neurotrophic Factor Therapy: GDNF and CNTF

E-Y Chen, D Fox, and J H Kordower, Rush University Medical Center, Chicago, IL, USA

© 2009 Published by Elsevier Ltd.

Introduction

During the past decade, there has been a major increase in research associated with neurotrophic factors (NTFs), molecules that provide survival-promoting effects to developing and adult central nervous system (CNS) neurons. From a clinical standpoint, it has been hoped that research would ultimately lead to the development of novel therapeutic approaches for the treatment of various neurodegenerative disorders, such as Alzheimer's disease, Parkinson's disease (PD), Huntington's disease (HD), and amyotrophic lateral sclerosis (ALS). To date, such clinical promise has not been realized, but optimism remains high due to the potency of the NTFs to protect and augment the function of vulnerable neuronal systems, combined with novel methods for their delivery.

Glial Cell Line-Derived Neurotrophic Factor

Identification and Role

The glial cell-derived neurotrophic factor (GDNF) family is distantly related to the transforming growth factor superfamily and includes GDNF and three structurally related members named neurturin, persephin, and artemin. GDNF, the first-identified member of this family, was originally isolated from the rat B49 glial cell line. GDNF binds to a multisubunit receptor system composed of the GFRα1 subunit and the RET subunit, which function as the ligand binding and signaling components, respectively. Ret is a transmembrane receptor tyrosine kinase (encoded by the proto-oncogene c-ret) that triggers many intracellular signaling pathways, including the Ras-MAPK, the phosphoinositol 3-kinase, and Jun N-terminal kinase- and PLCγ-dependent pathways. After the initial discovery that GDNF protects embryonic nigral neurons in culture, a plethora of *in vitro* and *in vivo* studies demonstrated that this factor prevents the degeneration and augments the phenotypic expression of dopaminergic midbrain neurons. With regard to GDNF and dopaminergic systems, there have been a series of pivotal experiments. A single intraventricular injection of GDNF to a normal rat induces a long-term increase in striatal dopamine. Most critically, GDNF has been able to prevent the structural and/or functional consequences that are engendered in virtually every rodent model of PD, including those with lesions induced by 6-hydroxydopamine, methamphetamine, or axotomy, as well as the degenerative changes associated with aging. Furthermore, CNS infusions of GDNF augment motor function and reverse nigrostriatal degeneration in parkinsonian and aged rhesus monkeys, and lentiviral delivery of GDNF prevents nigrostriatal degeneration in 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridire (MPTP)-treated monkeys and augments dopaminergic function in aged monkeys. An exception to this rule is provided by one study in which GDNF failed to prevent the structural and functional consequences that result from overexpression of α-synuclein in a rat model of PD. GDNF also has well-documented effects on the survival and function of cranial and spinal cord motor neurons, brain stem noradrenergic neurons, basal forebrain cholinergic neurons, Purkinje cells, and specific groups of dorsal ganglia and sympathetic neurons.

Distribution of GDNF in the CNS

Using polymerase chain reaction (PCR) and *in situ* hybridization techniques, the regional distribution and cellular localization of GDNF have been examined in the rat and human CNS. GDNF transcripts have been identified in all major regions of the rat CNS, such as the striatum, hippocampus, cortex, cerebellum, and spinal cord. Reverse-transcriptase polymerase chain reaction (RT-PCR) analysis revealed that GDNF mRNA is also observed in the human CNS in regions such as the striatum, hippocampus, cortex, and spinal cord. Using immunohistochemical techniques, it has been shown that GDNF immunoreactive neurons are widely distributed in the normal human brain and spinal cord, including cerebral cortex, hippocampus, basal forebrain, hypothalamus, thalamus, cerebellum, and brain stem. GDNF and its signaling receptor RET are also found in human peripheral nerve axons and DRG neurons, and GDNF is found in Schwann cells. GDNF immunostaining has also been found in melanin-containing neuronal perikarya of the substantia nigra pars compacta (SNpc) in normal human brain. This region undergoes extensive degeneration in PD. Experimental studies performed using RT-PCR techniques indicate that GDNF and its receptor components are broadly expressed in the rat and human hippocampus from early embryonic age to adult life. Available quantitative data show that in the rat hippocampus, levels of GDNF mRNA and protein steadily increase from the early stages to birth, reaching a maximum during the first week after

birth and gradually decreasing to prenatal levels in adult. The occurrence and pattern of overall distribution of GDNF-positive neurons in human hippocampus appear consistently similar throughout life stages from 29 weeks of gestation to adulthood. However, the relative numerical density of labeled neurons changes with age.

GDNF and PD

PD is a neurodegenerative disorder that affects more than 1 million people in North America. PD is characterized by the selective degeneration of nigral dopaminergic neurons in the SNpc and subsequent depletion of striatal dopamine which leads to motor dysfunction. PD patients can present with a variety of symptoms (bradykinesia, muscle rigidity, resting tremor, and loss of postural reflexes), two of which are required for diagnosis of the disease. The diagnosis of PD is made on the basis of clinical criteria. Confirmation of the clinical diagnosis occurs upon neuropathologic examination. Available pharmacological treatments do not stop the progressive loss of dopamine neurons observed during the course of PD. The progressive nature of the disease provides the possibility of therapeutic interventions that slow, or completely halt, the ongoing degenerative process. The use of trophic factors has emerged as a potential therapeutic approach because of their ability to regulate the survival of specific neuronal populations in the CNS. GDNF, a member of the transforming growth factor-β superfamily, has been shown to be one of the most potent neuroprotective factors for dopaminergic neurons. Many studies have demonstrated that GDNF controls the survival and physiologic properties of mesencephalic dopaminergic neurons both *in vitro* and *in vivo*. Therefore, reduction of GDNF expression in the SNpc might lead to neuronal loss in PD, although this concept is controversial. In one early *in situ* hybridization study, GDNF mRNA expression failed to show any labeling in the mesencephalon and striatum of control subjects or patients with PD. RT-PCR analysis revealed the highest expression of GDNF mRNA in the human caudate, with low levels in the putamen and no detectable mRNA in the SN, suggesting that GDNF is synthesized and secreted by striatal neurons and then retrogradely transported to the SN in a target-derived retrograde manner. Receptor studies found that tissue sections from the SNpc of control subjects and PD patients localized RET immunoreactivity to the dopaminergic neurons. RET mRNA expression was also observed in many surviving neurons in all PD patients. GDNF is present in the adult striatum, and the dopaminergic neurons of the SNpc express mRNAs encoding the GDNF receptors, RET and GFRα1. GDNF levels were significantly reduced (approximately 20%) in all sampled surviving melanin-containing neurons in the SNpc of PD patients compared with age-matched controls, supporting a role for decreased availability of GDNF as part of the process of degeneration of SNc neurons in PD. Together, these findings indicate that GDNF may play a role in the pathology of PD and may be the most potent neuroprotective factor for PD.

These and other studies prompted researchers to investigate putative protective effects in different PD models as well as PD patients. As discussed previously, direct injection of the recombinant GDNF protein has potent neuroprotective effects in rodent and primate models of PD. In polymer-encapsulated GDNF-secreting cells, continuous release of low levels of GDNF protects the nigral dopaminergic neurons from an axotomy-induced lesion and improves dopamine graft survival in 6-OHDA-treated rats. Gene therapy is another method that has great potential for delivering trophic factors continually to focal regions of the CNS. In this regard, viral vectors are attractive candidates for delivering GDNF to the PD brain given the nature of the disease. There are numerous studies using gene therapy in animal models of PD. Overexpression of GDNF to the striatum and/or SN has been achieved primarily using adeno-associated viruses or lentiviruses. One of the most promising studies demonstrated that delivery of GDNF using a primate-derived lentiviral vector provided almost complete neuroprotection against MPTP-mediated toxicity in monkeys and augmented dopaminergic function in aged monkeys (**Figure 1**). These encouraging preclinical observations led to a series of pilot human trials in which GDNF was used in PD patients. A double-blind, placebo-controlled, sequential cohort study performed for 8 months compared the effects of monthly intracerebroventricular administration of placebo and GDNF in 50 moderately advanced PD patients. These trials were suspended because of lack of observed efficacy and frequent occurrence of nausea, anorexia, tingling (Lhermitte sign), hallucinations, and depression. GDNF did not improve parkinsonism, possibly because GDNF did not reach the target tissues – putamen and SN. A phase I safety trial reported that 1 year after long-term infusion of GDNF directly via a pump and catheter into the posterodorsal putamen of 5 PD patients, there were no serious clinical side effects, and there was a 39% improvement in the off-medication motor subscore of the Unified Parkinson's Disease Rating Scale (UPDRS) and a 61% improvement in the activities of daily living subscore. Furthermore, levodopa-induced dyskinesias were reduced by 64%. This clinical improvement was accompanied by a 28% increase in putamen PET

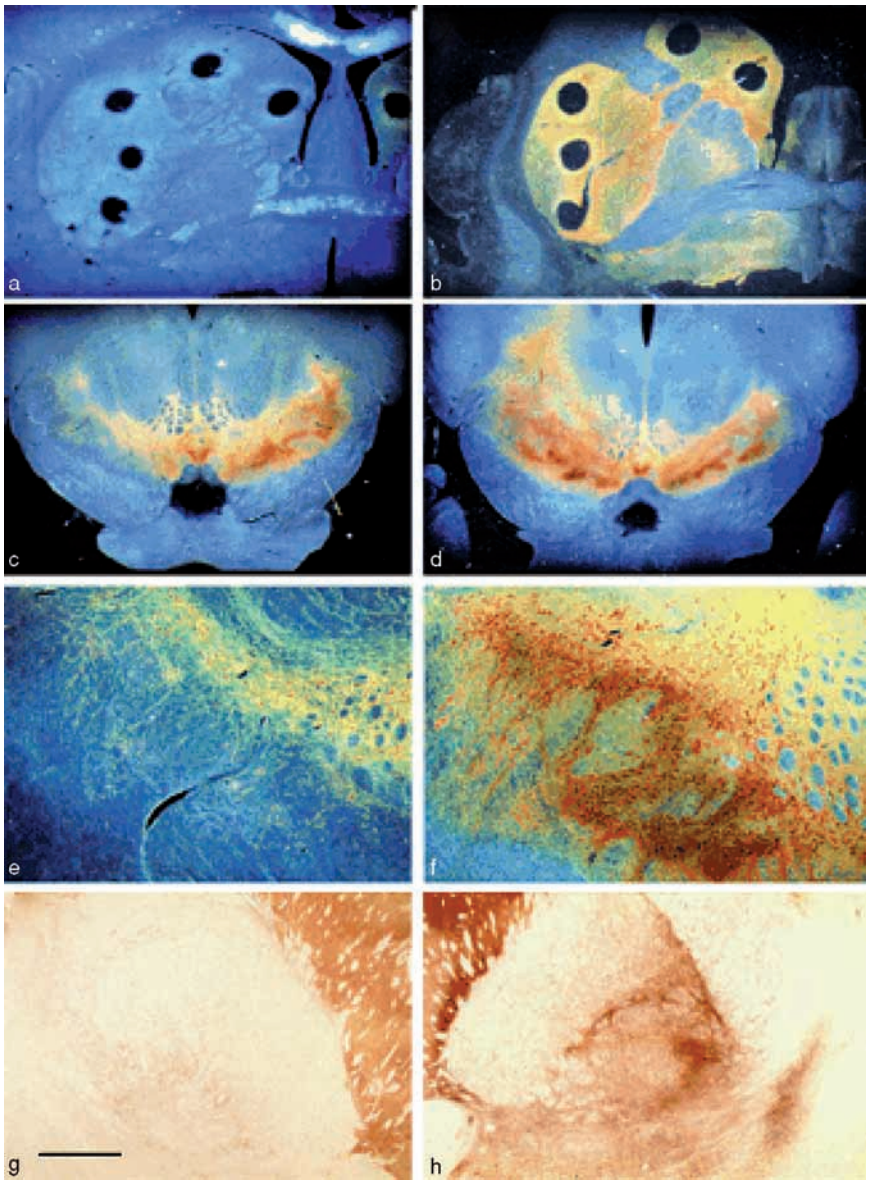

Figure 1 (a and b) Low-power dark field photomicrographs through the right striatum of TH immunostained sections of MPTP-treated monkeys treated with (a) lenti-βGal or (b) lenti-GDNF. (a) There was a comprehensive loss of TH immunoreactivity in the caudate and putamen of lenti-βGal-treated animal. In contrast, a near normal level of TH immunoreactivity is seen in lenti-GDNF-treated animals. Low-power (c, d) and medium-power (e, f) photomicrographs of TH immunostained section through the substantia nigra of animals treated with lenti-βGal (c, e) and lenti-GDNF (d, f). Note the loss of TH immunoreactive neurons in the lenti-βGal-treated animals on the side of the MPTP injection. TH immunoreactive sprouting fibers, as well as a supranormal number of TH immunoreactive nigral perikarya, are seen in lenti-GDNF-treated animals on the side of the MPTP injection. (g, h) Bright field low-power photomicrographs of a TH immunostained section from a lenti-GDNF-treated monkey. (g) Note the normal TH immunoreactive fiber density through the globus pallidus on the intact side that was not treated with lenti-GDNF. (h) In contrast, an enhanced network of TH immunoreactive fibers is seen on the side treated with both MPTP and lenti-GDNF. Scale bar in (g) represents the following: 3500 μm (a–d); 1150 μm (e–h).

F-DOPA uptake and evidence of neuronal sprouting in an autopsied brain. After 2 years of continuous GDNF infusion, there were still no serious clinical side effects and no significant detrimental effects on cognition. Patients showed a 57% and 63% improvement in their off-medication motor and activities of daily living subscores of the UPDRS, respectively, and health-related quality-of-life measures showed general improvement over time. A multicenter trial, however, showed no benefit from GDNF, and further studies have been suspended because of evidence of the development of GDNF antibodies in some patients. In addition, cerebellar neurodegeneration involving both Purkinjie cells and the granular cell layer in monkeys

exposed to high doses of GDNF has been found. Based on these data, clinical trials delivering GDNF protein to the putamen have been stopped.

GDNF and HD

Another potential therapeutic role of GDNF is in HD. HD is an autosomal dominant neurodegenerative disorder resulting from an expanded trinucleotide (CAG) repeat at the IT15 locus on chromosome 4 within the huntingtin gene. The abnormal DNA is translated into mutant huntingtin with an expanded glutamine stretch at the N-terminus of the protein. The excessive number of glutamine repeats is responsible for the misfolding of huntingtin and the subsequent formation of neuronal inclusions, degeneration of striatal and cortical neurons, and a triad of symptoms including severe motor, cognitive, and psychological disturbances that are ultimately fatal. To date, HD remains incurable. Toward this end, GDNF has shown promise in animal models of several different neurodegenerative disorders. In addition to its potent trophic effects on dopaminergic midbrain neurons, GDNF has also been shown to protect striatal, medium-sized spiny GABA projection neurons, the neuronal population most vulnerable in HD. Moreover, the expression of GDNF's receptors (GFRα1 and Ret) is upregulated in striatal neurons and astrocytes after injury, supporting its role as a potential trophic factor for HD. Intraventricularly administered GDNF protected striatal neurons from both excitotoxic and metabolic insult. Intrastriatal grafting of a GDNF-producing cell line or GDNF-secreting stem cells prevented the degeneration of striatal neurons in the excitotoxic rodent model of HD. Injection of adeno-associated virus (AAV) vector-mediated gene transfer of GDNF into the striatum improves behavior and provides neuronal protection in both rat and transgenic HD mice, the gold-standard animal model for HD.

GDNF and Cerebral Ischemia

Ischemic brain injury is a major cause of death and permanent disability worldwide. Cell injury and death after cerebral vascular occlusion have been postulated to result from a number of interacting pathophysiologic factors, including depolarization-induced calcium entry via *N*-methyl-D-aspartate receptors, intracellular free radical generation, damage to mitochondrial respiratory enzymes, and induction of programmed cell death. GDNF was initially considered to be a specific trophic factor and neuroprotective agent for dopaminergic neurons. It was later found that GDNF has effects on other systems. GDNF diminishes damage produced by kainate-induced tonic–clonic convulsions and axotomy of brain stem motor neurons. Since GDNF has an antiapoptotic effect in dopamine and nondopaminergic neurons after injury, it was hypothesized that GDNF could attenuate ischemic brain injury through the same mechanism. Exogenous administration of GDNF reduces the toxic effects of excitatory amino acids, attenuates nitric oxide production, and lowers apoptosis/cell death in stroke animals. Intracerebroventricular and intraparenchymal administration of GDNF potently protected the cerebral hemispheres from damage induced by middle cerebral artery (MCA) occlusion in rats. A reduction of ischemic brain injury by topical application of GDNF after permanent MCA occlusion in rats has also been reported. Intravenous application of a blood–brain barrier-penetrating form of GDNF, TAT-GDNF, reduces infarct size after ischemia in mice. Topically applied to the ischemic brain after MCA ligation, GDNF-fibrin glue reduces ischemic brain injury in chronic focal cerebral ischemia in conscious rats. Both recombinant AAV vector and herpes simplex virus amplicon-based vector expressing GDNF reduce ischemia-induced damage. These data support the concept that GDNF may play a therapeutic role in the treatment of ischemia. However, a remaining challenge is to determine how to deliver this trophic molecule in a manner in which it can be clinically useful.

Ciliary Neurotrophic Factor

Identification and Role

Ciliary neurotrophic factor (CNTF) is structurally and functionally related to members of a family of cytokines that includes leukemia inhibitory factor (LIF), interleukin-6, interleukin-11, and oncostatin M. CNTF was initially identified due to its ability to support the survival of parasympathetic neurons of the chick ciliary ganglion *in vitro* and subsequently purified to homogeneity from sciatic nerves. CNTF is synthesized by astrocytes and is generally thought to provide autocrine and paracrine signals from astrocytes which induce cellular protection and hypertrophy in response to injury. CNTF exerts its biological effects by binding to a ligand-specific α subunit, CNTF receptor α (CNTFR α), which leads to the subsequent recruitment of two signal-transducing β components, LIFRβ and gp130. Following dimerization of the β components, preassociated cytoplasmic Jak and Tyk kinases are activated, resulting in tyrosine phosphorylation of the β components and other cellular proteins. CNTF plays an important role in the maintenance and regeneration of the adult nervous system.

CNTF prolongs neuronal survival of various culture systems and enhances the survival of sensory neurons, motor neurons, cerebral neurons, and hippocampal neurons. CNTF also promotes cholinergic differentiation of noradrenergic sympathetic neurons, the differentiation of glial progenitors into astrocytes, and the maturation and survival of oligodendrocytes. Some peripheral tissues, notably skeletal muscle, are also known to respond to CNTF. Many studies have demonstrated that CNTF administration prevents injury-induced cell death and degeneration of cholinergic, dopaminergic, GABAergic, and thalamocortical neurons in various lesion models of CNS diseases. However, the majority of studies have linked CNTF with motor function in animal models of ALS, HD, and sensory function such as retinopathies.

Distribution of CNTF in the CNS

CNTF was originally identified as a survival-promoting activity for ciliary ganglion neurons in chick eye tissues. Subsequently, it was cloned from mammalian nervous tissue. The protein is mainly expressed in glial cells of both the peripheral nerve system and the CNS and has been implicated in many developmental processes in the nervous system, including cell fate determination of neural stem cells, the proliferation and transmitter choice of neuronal precursor cells, and the survival and differentiation of developing neurons as well as glial cells. In addition, it acts on a variety of mature peripheral and central neurons preventing injury-induced cell death and degeneration. Demonstrations of retrograde axonal transport of radiolabeled CNTF in the sensory neurons of the sciatic nerve suggest that CNTF may function by target synthesis and retrograde transport in central neurons. CNTF was observed only after birth and found in both the neurons and the neuroglia of the SN, mesencephalon, cerebellum, and the spinal cord from postnatal day 1 to postnatal week 3 in developing Mongolian gerbils. CNTF has been localized primarily in glial cells, with highest concentrations in Schwann cells, and in white matter tract astrocytes of the mouse CNS. The expression of CNTFR α is widely distributed in both rat and primate CNS motor, sensory, and autonomic neurons. There are no published data on the distribution of CNTF in human brain.

CNTF and Amyotrophic Lateral Sclerosis

ALS is the most common adult form of motor neuron disease. The clinical hallmarks of ALS are progressive muscle weakness, atrophy, and spasticity, which reflect the degeneration and death of upper and lower motor neurons in the cortex, brain stem, and spinal cord. Generally fatal within 1–5 years of onset, ALS has a prevalence of 1 or 2 per 100 000 people. In 90% of cases, there is no apparent genetic linkage (sporadic ALS), but in the remaining 10% the disease is inherited (familial ALS), often in an autosomal-dominant manner. Among familial cases of ALS, approximately 20% are caused by mutations in the gene encoding Cu, Zn-superoxide dismutase (SOD1). The role of CNTF in the etiology and pathogenesis of ALS is not clear. Using a sensitive enzyme-linked immunoassay, CNTF protein in the lateral corticospinal tract was markedly lower in ALS patients than in controls. In the spinal cord of ALS patients, there is a marked increase in the hybridization signal for CNTFR α subunit mRNA overlying motor neurons. CNTF improves survival in the progressive motor neuronopathy (pmn) and wobbler mutant mouse models of motor neuron disease, even though these mice do not have an obvious defect in CNTF or its receptor. Conversely, mice bred for the null mutation of CNTF or its receptor do develop motor neuron degeneration. CNTF also appears to be effective for retarding cellular and functional losses in the mouse models of ALS. CNTF significantly reduced motor neuron loss, improved motor function, and increased survival time. These observations led to the unsuccessful use of CNTF in a therapeutic clinical trial in ALS. This trial was designed to evaluate the safety, tolerability, and efficacy of subcutaneous administration of recombinant human CNTF. Unfortunately, it failed to ameliorate symptoms of ALS while producing adverse effects, such as injection site reactions, cough, asthenia, nausea, anorexia, weight loss, and increased salivation.

CNTF and HD

CNTF has also been tested in multiple animal models of HD. CNTF has been shown to protect striatal output neurons in a rat model of HD. In one study, encapsulated cells containing baby hamster kidney cells genetically modified to deliver hCNTF were implanted into the lateral ventricle and provided dose-dependent neurochemical and functional protection in a rat model of HD. Human neural stem cell transplants improved motor function in a rat model of HD, and recovery was superior in rats receiving CNTF-treated human cells compared to rats receiving transplants of untreated cells. Intrastriatal injections of adenovirus-mediated CNTF provided neuroprotection for corticostriatopallidal neurons in a rat model of progressive striatal degeneration. In nonhuman primate models of HD, intrastriatal implantation encapsulated CNTF-producing cells prevented the degeneration of vulnerable striatal populations and cortical–striatal basal ganglia circuits, and in a

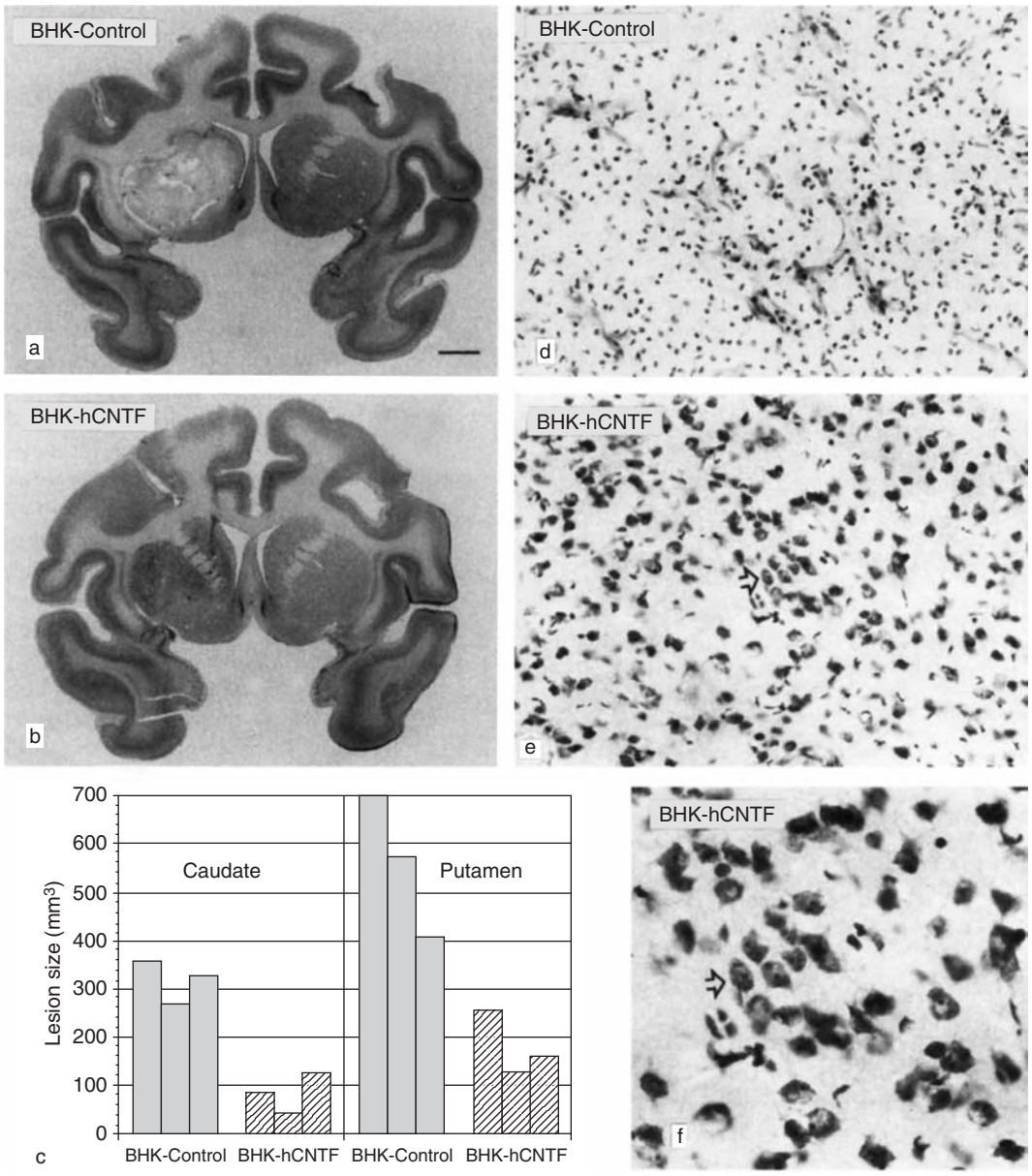

Figure 2 Low-power photomicrographs of Nissl stained sections through the forebrain of monkeys receiving unilateral intracaudate and intraputamenal lesions with QA and encapsulated (a) BHK-control or (b) BHK-hCNTF implants. Note the extensive lesion area on the left side which encompasses much of the striatum in a. In contrast, the damage in b is negligible for a monkey treated with hCNTF. (c) The size of the lesion in both the caudate nucleus and putamen as quantified on Nissl stained sections was significantly decreased in BHK-hCNTF implanted monkeys (striped bars) relative to BHK-control implanted animals (gray bars). Each bar represents data from an individual animal. (d, e) Medium- and high-power (f) photomicrographs of Nissl stained sections through the striatum of BHK-control (d) and BHK-hCNTF implanted (e, f) monkeys. (d) A paucity of neurons is seen in the striatum of control implanted monkeys. (e,f) In contrast, numerous healthy neurons can be seen near a QA injection site in BHK-hCNTF implanted animals. For reference, the arrows in (e) and (f) point to the same location. Scale bar in (a) represents the following: 4000 μm (a, b); 50 μm (d, e); 25 μm (f).

primate model of HD they restored cognitive and motor functions (Figure 2). In 2004, a human clinical trial of CNTF for HD was reported. This phase I trial evaluated the safety of intracerebral administration of CNTF protein in stage 1 or 2 HD patients using a device formed by a semipermeable membrane encapsulating a BHK cell line engineered to synthesize CNTF. No sign of CNTF-induced toxicity was reported, and improvements in electrophysiological results were observed.

CNTF and Retinal Degenerative Diseases

Retinal degeneration in its various forms is responsible for vision loss in a large number of young

adults. Retinitis pigmentosa (RP), for example, affects 50 000–100 000 people in the United States and an estimated 1.5 million people worldwide. RP characterized by a progressive loss of photoreceptors through-mechanisms is not fully understood. Subsequent to photoreceptor loss, there is both vascular and neural remodeling. The pathogenesis of retinal degeneration is complicated. With the exception of vitamin A nutritional supplementation, no treatments have been shown to be effective across the range of these disorders. Regardless of the initial causative genetic defect, the end result is photoreceptor cell death. The multiplicity of mechanisms stimulated a search for therapeutic agents that are effective in slowing photoreceptor death regardless of the causative genetic mutation. Intervention studies have indicated the possibility of using neurotrophic factors as therapeutic agents for RP. CNTF is a cytokine with neurotrophic activity across a broad spectrum of peripheral and CNS cells. CNTF was first identified as a survival factor in studies involving ciliary ganglion neurons in the chick eye. CNTF mRNA expression in the retina during postnatal development increased with postnatal age. CNTF upregulation plays a role in retinal protection against light-induced damage after optic nerve section. CNTF has been shown to modulate the survival of retinal neuronal cells *in vivo* and *in vitro*. Repeated injections of a CNTF analog lead to long-term photoreceptor survival in hereditary retinal degeneration. Encapsulated cell-based delivery of CNTF reduces photoreceptor degeneration in canine models of RP, protects photoreceptors in a canine model of retinal degeneration, and does not adversely affect either rod or cone ERG function of normal rabbit retina. AAV-mediated delivery of CNTF prolongs photoreceptor survival in the rhodopsin knockout mouse and rat models of RP. Intraocular gene transfer of CNTF prevents photoreceptor degeneration in Royal College of Surgeons rats and mice. However, intraocular gene delivery of CNTF results in significant loss of retinal function in both normal mice and Prph2Rd2/Rd2 retinal degeneration mice. A phase I human clinical trial of CNTF for retinal neurodegeneration indicates that CNTF is safe for the human retina even with severely compromised photoreceptors. The approach to delivering therapeutic CNTF proteins to degenerating retinas using encapsulated cell implants may have application beyond disease caused by genetic mutations.

Conclusion

Trophic factors such as GDNF and CTNF hold great promise for the treatment of degenerative disease. However, significant challenges remain with regard to establishing biomarkers for these diseases that will allow for interventions much earlier in the disease process than are currently available. Furthermore, additional studies aimed at developing optimal delivery methods for these potent compounds would ultimately benefit their clinical use.

See also: Amyotrophic Lateral Sclerosis (ALS); Animal Models of Parkinson's Disease; Cell Replacement Therapy: Parkinson's Disease; Cell Replacement Therapy: Mechanisms of Functional Recovery; Cell Replacement Therapy for Huntington's Disease; Enteric Nervous System: Neurotrophic Factors; Gene Therapy: Genetically Modified Cells; Gene Therapy: Direct Viral Delivery; GFL Neurotrophic Factors: Physiology and Pharmacology; Huntington's Disease; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Aebischer P, Schluep M, Deglon N, et al. (1996) Intrathecal delivery of CNTF using encapsulated genetically modified xenogeneic cells in amyotrophic lateral sclerosis patients. *Nature Medicine* 2: 696–699.

Cayouette M, Behn D, Sendtner M, Lachapelle P, and Gravel C (1998) Intraocular gene transfer of ciliary neurotrophic factor prevents death and increases responsiveness of rod photoreceptors in the retinal degeneration slow mouse. *Journal of Neuroscience* 18: 9282–9293.

Clatterbuck RE, Price DL, and Koliatsos VE (1993) Ciliary neurotrophic factor prevents retrograde neuronal death in the adult central nervous system. *Proceedings of the National Academy of Sciences of the United States of America* 90: 2222–2226.

Emerich DF, Winn SR, Hantraye PM, et al. (1997) Protective effect of encapsulated cells producing neurotrophic factor CNTF in a monkey model of Huntington's disease. *Nature* 386: 395–399.

Gash DM, Zhang Z, Ovadia A, et al. (1996) Functional recovery in parkinsonian monkeys treated with GDNF. *Nature* 380: 252–255.

Gill SS, Patel NK, Hotton GR, et al. (2003) Direct brain infusion of glial cell line-derived neurotrophic factor in Parkinson disease. *Nature Medicine* 9: 589–595.

Henderson CE, Phillips HS, Pollock RA, et al. (1994) GDNF: A potent survival factor for motoneurons present in peripheral nerve and muscle. *Science* 266: 1062–1064.

Kordower JH, Emborg ME, Bloch J, et al. (2000) Neurodegeneration prevented by lentiviral vector delivery of GDNF in primate models of Parkinson's disease. *Science* 290: 767–773.

Lang AE, Gill S, Patel NK, et al. (2006) Randomized controlled trial of intraputamenal glial cell line-derived neurotrophic factor infusion in Parkinson disease. *Annals of Neurology* 59: 459–466.

Lin LF, Doherty DH, Lile JD, Bektesh S, and Collins F (1993) GDNF: A glial cell line-derived neurotrophic factor for midbrain dopaminergic neurons. *Science* 260: 1130–1132.

McBride JL, Ramaswamy S, Gasmi M, et al. (2006) Viral delivery of glial cell line-derived neurotrophic factor improves behavior and protects striatal neurons in a mouse model of Huntington's disease. *Proceedings of the National Academy of Sciences of the United States of America* 103: 9345–9350.

Mittoux V, Joseph JM, Conde F, et al. (2000) Restoration of cognitive and motor functions by ciliary neurotrophic factor in a primate model of Huntington's disease. *Human Gene Therapy* 11: 1177–1187.

Patel NK, Bunnage M, Plaha P, et al. (2005) Intraputamenal infusion of glial cell line-derived neurotrophic factor in PD: A two-year outcome study. *Annals of Neurology* 57: 298–302.

Wang Y, Lin SZ, Chiou AL, Williams LR, and Hoffer BJ (1997) Glial cell line-derived neurotrophic factor protects against ischemia-induced injury in the cerebral cortex. *Journal of Neuroscience* 17: 4341–4348.

Yan Q, Matheson C, and Lopez OT (1995) *In vivo* neurotrophic effects of GDNF on neonatal and adult facial motor neurons. *Nature* 373: 341–344.

Neurotrophic Factor Therapy: NGF, BDNF and NT-3

A Blesch and M H Tuszynski, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The discovery of the first nervous system growth factor in 1951, nerve growth factor (NGF), began an era of exploring potential therapeutic benefits of neurotrophins in animal models of disease. This effort gained considerable momentum with reports beginning in the mid-1980s that growth factors prevent neuronal death in the adult brain and then reports in the mid-1990s that growth factors promote extensive axonal growth in the adult central nervous system (CNS).

Numerous studies have confirmed that delivery of neurotrophins (NGF, brain-derived neurotrophic factor (BDNF), neurotrophin-3 (NT-3), or NT-4) by intermittent injection, continuous infusion, cellular grafting, or gene therapy can augment neuronal survival in animal models of neurodegenerative diseases, increase the number of regenerating axons, and amplify the speed and distance of axonal regeneration. However, identifying a clinically practical means of neurotrophin delivery has proven to be one of the most important obstacles in translational human testing of these findings. To be tested in human trials, neurotrophins need to reach targeted regions in effective concentrations but cannot spread beyond targeted regions if adverse affects are to be avoided. Furthermore, ideally the amount and duration of growth factor delivery would be highly controllable in human testing.

This article reviews some of the data regarding physiological and therapeutic roles of neurotrophins in animal models of neurodegenerative diseases and axonal regeneration. These findings have led to potentially promising clinical trials with neurotrophins. The discussion is focused on NGF, BDNF, and NT-3.

Neurotrophins in Neurodegenerative Disorders

NGF and Alzheimer's Disease

NGF is expressed in the neocortex and the hippocampus during development and in adulthood, supporting the survival and function of basal forebrain cholinergic neurons in the nucleus basalis of Meynert and the medial septum. Beginning in 1986, the groups of Franz Hefti, Larry Kromer, and Rusty Gage separately reported that intracerebroventricular infusions of NGF prevent the lesion-induced degeneration of basal forebrain cholinergic neurons in the medial septum, elucidating for the first time the powerful effects of growth factors in preventing cell death in the brain. Because cholinergic neuronal decline is a prominent component of cell loss in Alzheimer's disease (AD), it was suggested that NGF could potentially slow or prevent the loss of this important cell population and thereby improve cognition in AD patients.

Several subsequent studies confirmed the neuroprotective effects of NGF after lesions. Furthermore, NGF was found to improve learning and memory in aged and lesioned rats. Additional effects of NGF were reported in primate models of cholinergic cell loss and aging. This collective body of work led to a small clinical trial of NGF intracerebroventricular infusion into three patients with AD in 1997, but side effects of broad dissemination of NGF throughout the neuraxis required a halt to the trial. Animal studies confirmed that nontargeted growth factor administration led to a variety of adverse effects, including weight loss, sensory axon sprouting (causing pain), sympathetic axon sprouting, and Schwann cell migration into the CNS. These adverse effects of nontargeted NGF delivery highlighted the dual need to not only achieve adequate concentrations of NGF in the brain to stimulate neurons but also spatially restrict growth factor delivery to regions of degenerating neurons to avoid adverse effects of NGF stimulation of intact neurons of other systems.

Subsequent studies attempted to achieve both effective and safe local delivery of NGF using techniques of gene therapy. The effectiveness and safety of NGF gene delivery were established in several preclinical studies using both rodent and primate models of AD (Figure 1), establishing the basis for a human clinical trial.

Based on these data, in 2001 the first human clinical trial of growth factor gene delivery was performed in eight humans with early stage AD. Patients were enrolled based on the rationale that early intervention was necessary both to prevent ongoing cholinergic neuronal loss and to allow research subjects to provide informed consent for this initial clinical trial. Gene therapy was accomplished using retroviral vectors that resulted in NGF secretion from the subjects' own skin fibroblasts, obtained several months earlier from skin biopsies. These NGF-secreting cells were then implanted into the brains of the AD patients adjacent to the cholinergic nucleus basalis of Meynert. Safety, cognitive, and neuroimaging outcomes were subsequently assessed.

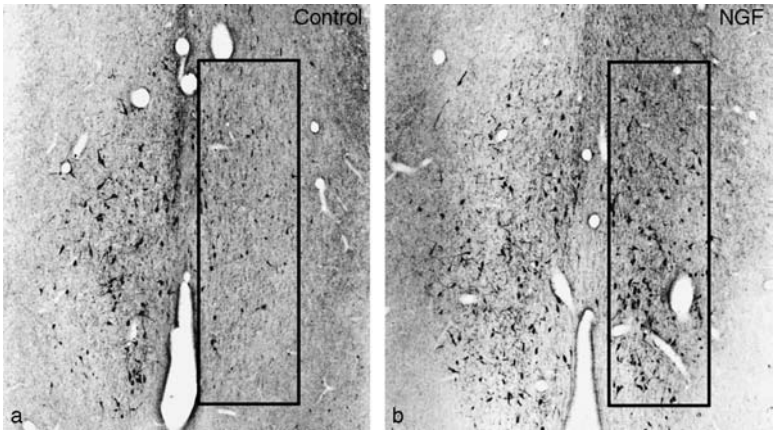

Figure 1 NGF delivery by genetically modified fibroblasts rescues degenerating medial septum cholinergic neurons from lesion-induced degeneration in the primate basal forebrain. (a) Animals treated with control fibroblasts show a loss of cholinergic neurons on the right, lesioned side (boxed area) after unilateral fornix lesions that transect cholinergic axons projecting to the hippocampus (the natural source of NGF for these neurons). (b) In contrast, cholinergic neurons are preserved on the right, lesioned side of the brain (boxed area) after treatment with NGF-secreting cells.

Definitive conclusions regarding the efficacy of NGF delivery cannot be drawn from this small clinical trial in the absence of blinded assessments and placebo controls. Nonetheless, a potential benefit of NGF was identified in that the mean rate of cognitive decline in subjects was reduced by approximately 50% during the 2-year period of the trial compared to their own pretreatment rates of decline. Furthermore, comparison of cortical metabolic activity by positron emission tomography (PET) scan in NGF-treated subjects showed significant improvements in cortical glucose metabolism 6–8 months after treatment (Figure 2). No adverse effects of NGF gene delivery were detected during the 2-year period of the study. However, the trial was initially performed with subjects in a wakened state, and two subjects moved during cell injections, resulting in brain hemorrhages. Subsequent procedures were performed with subjects under general anesthesia and no further complications occurred. One subject came to autopsy 5 weeks after NGF gene transfer; examination of this subject's brain demonstrated robust trophic responses from cholinergic axons adjacent to the site of NGF delivery (axonal sprouting), definitively establishing that degenerating cholinergic neurons of the human AD brain can respond to NGF.

Although conclusions regarding efficacy cannot be drawn from this small study, if effects of this magnitude occur in proper blind and controlled trials, this would represent a significant improvement over current symptomatic therapies for AD and would provide the first long-lasting neuroprotective therapy. The findings of the phase I NGF gene therapy trial were of sufficient interest to support testing in expanded and controlled phase II clinical trials. These trials will begin following completion of a second phase I trial of NGF delivery in AD patients using newer, more efficient adeno-associated viral (AAV) vectors that is in progress.

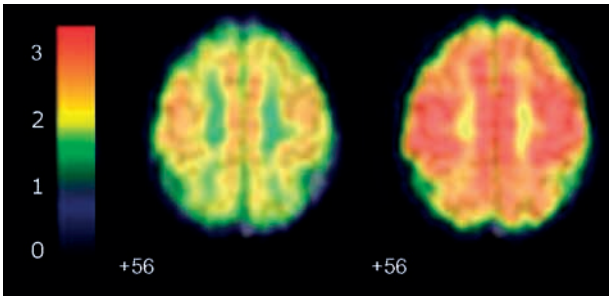

Figure 2 Average fluor-deoxyglucose PET scans in patients grafted with NGF secreting fibroblasts to the nucleus basalis overlaid on standardized MRI templates. Comparison of the first scan (left) and the second scan (right) obtained 6–8 months later indicates a widespread increase in cortical metabolic activity after NGF delivery. Adapted from Tuszynski MH, Thal L, Pay M, et al. (2005) A phase 1 clinical trial of nerve growth factor gene therapy for Alzheimer disease. *Nature Medicine* 11: 551–555.

Since the initiation of the trial described previously, significant progress has been made in the development of viral vectors for *in vivo* gene therapy, making NGF gene delivery simpler and potentially more effective. Studies have shown that *in vivo* NGF delivery using AAV and lentivirus can effectively protect basal forebrain cholinergic neurons. Based on the results of *ex vivo* NGF gene delivery in patients with AD, and extensive preclinical safety, tolerability, and efficacy studies using AAV in rodents and primates, a phase I clinical trial evaluating AAV-mediated NGF gene delivery to the nucleus basalis in patients with

AD is currently under way. Following successful completion of current clinical studies, larger phase II/III trials will be needed to provide clear and objective data regarding the efficacy of NGF in AD. These findings will establish whether NGF can prevent or delay cholinergic degeneration, and whether NGF acting on only one of many neuronal systems undergoing degeneration in AD can improve memory and cognition or delay further cognitive deterioration.

Neurotrophins in Parkinson's Disease, Huntington's Disease, and Amyotrophic Lateral Sclerosis

Several neurotrophic factors, including BDNF, NT-3, and NT-4, have been shown to augment survival of midbrain dopaminergic neurons in the substantia nigra. This key neuronal population is the primary system resulting in the clinical signs and symptoms of Parkinson's disease. Although some promising animal studies demonstrated that BDNF can prevent the degeneration of nigral dopaminergic neurons in animal models, the majority of work has focused on glial cell line-derived neurotrophic factor (GDNF) and other members of this family of neurotrophic factors, discovered in the early 1990s.

NGF, BDNF, NT-3, and NT-4 have also been evaluated as neuroprotective agents of striatal neuronal degeneration in correlational animal models of Huntington's disease (HD). A neuroprotective strategy is particularly attractive in HD because susceptible individuals can be identified prior to disease onset by genetic testing. The excitotoxic lesion models of HD, consisting of quinolinic acid, kainic acid, or 3-nitropropionic acid injections into striatum, do not fully replicate the pathological changes in HD associated with the accumulation of polyglutamine-containing huntingtin inclusions. However, evidence does indicate a role of huntingtin in influencing BDNF transport and BDNF expression, providing additional rationale to investigate BDNF as a potential therapeutic molecule in this devastating polyglutamine disease. All the neurotrophins have in fact been reported to exert some degree of neuroprotection in animal models of HD, although the mechanism underlying these effects (rescue of medium spiny neurons) may be indirect.

BDNF has potent effects on motor neuron survival *in vitro*, during development, in adult animals, and after injuries to motor systems. These findings indicate that BDNF could be developed as a treatment for amyotrophic lateral sclerosis (ALS), in which ventral motor neuron degeneration is extensive. Indeed, several clinical trials evaluated BDNF as a neuroprotective agent in ALS. These trials were negative, however, because they encountered the conundrum of effective and safe delivery of growth factors to targeted neurons. Some trials administered BDNF via subcutaneous injection, likely leading to subtherapeutic levels in the CNS and broad distribution peripherally that would induce side effects. Other trials infused BDNF into the spinal fluid, again resulting in likely subtherapeutic concentrations in motor neurons of the spinal cord (due to low degrees of BDNF diffusion across the spinal cord surface) while inducing adverse effects from CNS-wide exposure. To test true efficacy, a means of targeting growth factors to affected motor neurons will need to be developed. Retrograde transport of gene therapy vectors after peripheral injection may be a means of achieving this goal, but further studies are required to optimize vector serotypes and envelope proteins to enhance retrograde transport.

Neurotrophins and Axonal Regeneration

Role of Neurotrophins in Peripheral Nervous System Regeneration

Axon regeneration after a nerve crush in the peripheral nervous system (PNS) is superior to regeneration observed in the CNS. Some of the factors contributing to the superiority of PNS regeneration include (1) formation of a supportive or 'bridging' extracellular matrix in sites of PNS injury, supporting new axonal growth; (2) secretion of several neurotrophic factors by Schwann cells around the lesion site, thereby attracting and guiding regenerating axons; (3) guidance of injured axons to motor end plate targets via the intrinsic detailed architecture of the peripheral nerve, including bands of Bungner that encompass the neurilemmal sheath, the injured axon, and its associated Schwann cell; and (4) an absence of inhibitors to regeneration (in contrast to the CNS, in which molecular inhibitors associated with myelin and the extracellular matrix impede axonal regeneration).

Of the factors contributing to peripheral regeneration, the production of growth factors in proper spatial and temporal gradients is essential for successful regeneration. That is, growth factors are produced in the right place and at the right time to attract regenerating axons and help guide them to targets often located very long distances distally (more than 0.5 m in humans). Neurotrophins are expressed by both Schwann cells and invading macrophages after PNS injury. Schwann cells upregulate expression of NGF, BDNF, NT-4, and other neurotrophic factors such as GDNF at the site of nerve injury and within the distal stump. In contrast, levels of NT-3 expression decrease, at least transiently. The increases in neurotrophin levels are very rapid and can be observed within a single day. At the same time that neurotrophin levels increase, Schwann cells upregulate expression

of the low-affinity neurotrophin receptor p75 on their cell surface. p75 may act as a receptor to both increase Schwann cell proliferation, and present neurotrophins to regenerating axons to achieve high localized concentrations of these growth factors.

Neurons with afferent and efferent projections in the PNS, including dorsal root ganglion (DRG) sensory neurons, spinal motor neurons, and sympathetic neurons, bind and retrogradely transport NGF, BDNF, NT-3, and NT-4. As axons regenerate, neurotrophin expression decreases at the injury site and subsequently in the distal stump. BDNF and NT-4 appear to play particularly important roles in peripheral nerve regeneration: Axon regeneration is impaired after neutralization of BDNF signaling in peripheral nerve by infusion of BDNF antibodies or in heterozygous trkB knockout mice. Consistent with these findings, both BDNF and NT-4 are upregulated in the nerve distal to an injury, and injections of BDNF enhance axonal regeneration. In addition to changes in neurotrophin levels observed in the injured nerve, increased levels of NGF and BDNF can be detected in the muscle after sciatic nerve transection, potentially contributing to target reinnervation.

These observations indicate that neurotrophins potentially initiate growth programs in injured axons and directly serve as chemotropic attractants for injured axons to stimulate localized axonal extension. Increased BDNF synthesis is also observed in DRG neurons, potentially preventing the death of these cells after injury. The exact mechanisms underlying increased neurotrophin expression by Schwann cells after PNS injury are incompletely understood and might be partially regulated by the loss of axonal contact or by the loss of a diffusible signal released by axons.

Given the importance of neurotrophins in supporting regeneration after nerve crush, their therapeutic application has been studied in models of long-distance nerve transection and root avulsion. Peripheral nerve injuries that result in gaps in the nerve generally have a poor prognosis, and the success of spontaneous regeneration declines with increased lengths of the nerve gap. To improve axonal regeneration in these injuries, autologous nerve transplants and silicone guidance tubes are currently employed to connect the nerve stumps. These strategies are relatively successful for short-distance gaps of less than approximately 0.5 cm, but they become increasingly ineffective as the bridging distance increases. Current studies are therefore investigating whether combinations of Schwann cell transplants, guidance tubes, and neurotrophin delivery will improve the speed of axonal regeneration and the distance and number of regenerating axons. Slow-release neurotrophin delivery methods such as fibrin matrices, genetically modified cells, or localized infusions in combination with nerve conduits made from biocompatible materials may provide superior regeneration compared to current techniques. Given the importance of neurotrophins for peripheral nerve regeneration, neurotrophin support in sufficient quantities for sufficient time periods will likely improve the otherwise poor degree of regeneration across long gaps in peripheral nerves.

In addition to the difficulties observed in PNS regeneration across nerve gaps, avulsion injuries disconnecting the peripheral nerve from the spinal cord have very poor clinical outcomes. This is in part due to motor neuron degeneration and cell death and in part due to the inability of dorsal root sensory axons to cross through the dorsal root entry zone and into the spinal cord as the environment transitions from the permissive regeneration milieu of the peripheral nerve to the nonpermissive regeneration milieu of the CNS. Poor axonal growth from the PNS to the CNS environment appears to be due to several factors, including myelin-based growth inhibitors present in the CNS; inhibitory extracellular matrix that forms after injury at the PNS/CNS interface, including the deposition of chondroitin sulfate proteoglycan family members; and the formation of a glial 'scar' at the dorsal root entry zone. Several studies have shown that delivery of NT-3 or NGF either by infusions or by *in vivo* gene transfer can promote regeneration across the dorsal root entry zone and reinnervation of the spinal cord in rodents. Regenerated axons can form functional synapses indicated by electrophysiological and functional recovery. Although viral NGF delivery has been shown to improve axonal regeneration after dorsal root avulsion injuries, NGF also results in sprouting of uninjured sensory axons from adjacent spinal segments, which can cause hyperalgesia and chronic pain. In fact, anti-NGF therapy might be one means of preventing aberrant sprouting of sensory axons in the spinal cord, which often leads to pain after PNS or spinal cord injury. Thus, the timing, amount, and spatial distribution of neurotrophins are likely to be the most important factors in the successful translation of neurotrophin therapy in PNS regeneration.

Neurotrophins and Regeneration in the CNS

The importance of neurotrophins in neuronal survival and axonal pathfinding during development, and peripheral nerve regeneration in the adult, has made these naturally occurring factors potentially promising targets to enhance regeneration in the adult CNS. Indeed, experiments during the past 10 years have demonstrated that neurotrophins delivered to the injured rodent or primate spinal cord can

substantially augment axonal plasticity and regeneration. Different neurotrophins influence the growth of distinct populations of axons in the spinal cord, based on the distribution of high-affinity neurotrophin receptors for trkA, -B, and -C. trkA is expressed on primary nociceptive afferent axons penetrating the spinal cord from DRG. In contrast, trkB and trkC are widely expressed in neurons with motor projections to the spinal cord, including corticospinal, red nucleus, locus coeruleus, vestibular, raphe, and reticular formation neurons. Trk receptors are actively trafficked into the axonal compartment of neurons, leading to enhanced axonal plasticity or regeneration when neurotrophins are administered to the spinal cord.

Although differences in the number, distance, and phenotype of extending axons have been reported among various studies, virtually all studies conducted to date demonstrate that axonal growth is enhanced by neurotrophins after spinal cord injury. The magnitude of this augmented axonal growth depends on the means, types, and amounts of neurotrophins applied and also on the spinal level of injury (Table 1). In particular, BDNF induces axonal growth from a diversity of descending motor systems, consistent with expression of trkB in the corresponding axonal populations. Responses to NT-4 are generally similar to those to BDNF because both growth factors signal through the same tyrosine kinase receptor, trkB. In addition, corticospinal axon growth in response to cellular or viral NT-3 delivery has been demonstrated in several studies. Ascending, proprioceptive dorsal column sensory axons also exhibit sensitivity to BDNF and NT-3, consistent with trkB and trkC expression in this group of large and medium-sized DRG neurons.

The inability of neurotrophins to cross the blood–brain barrier requires that they be delivered directly into the CNS to stimulate axonal growth. As noted previously, the administration of growth factors in models of axonal regeneration must also be restricted to specific target regions to prevent adverse effects of nontargeted delivery. These dual objectives of site-specific and site-limited growth factor delivery can be accomplished either by local neurotrophin infusions or by direct neurotrophin gene delivery using AAV or lentivirus vectors. If the objective of neurotrophin delivery is to promote axonal regrowth through a lesion cavity, neurotrophin delivery must be accompanied by the placement of a permissive extracellular matrix, or a 'bridge,' in the lesion cavity over which injured axons can extend. A common means of providing both a bridge in the lesion site and growth factors is to graft cells into the lesion site that have been genetically modified to secrete growth factors. This approach provides both a bridge (the cell graft) and the neurotrophin (via gene delivery). Cells used in such studies have included fibroblasts, Schwann cells, pre-degenerated peripheral nerve bridges, bone marrow stromal cells, and neural stem cells or progenitor cells.

The strongest growth has generally been observed following cellular or viral neurotrophin delivery due to a highly localized, long-lasting source of trophic supply. For example, fibroblasts genetically modified to express BDNF or NT-4 and grafted to the lesion site induce robust growth of serotonergic raphaespinal axons, rubrospinal axons, reticulospinal axons, coerulospinal axons, cholinergic spinal motor axons, and subtypes of sensory axons (Figure 3). Although corticospinal neuron cell bodies are known to express the high-affinity BDNF receptor trkB, and BDNF infusions can prevent the degeneration of corticospinal motor neurons after lesions close to the cell body, these axons do not respond to spinal BDNF delivery. The reasons underlying the lack of corticospinal axons growth in response to BDNF remain to be determined. BDNF might serve solely as a survival factor and not an axonal growth factor for corticospinal motor neurons, the distribution of trkB receptors might be restricted to the cortical cell body thereby preventing axonal responses, or myelin-based growth inhibitors and other inhibitory axonal inhibitors might outweigh the growth-promoting effects of BDNF.

Although neurotrophins can promote robust axonal growth into neurotrophin-producing cell grafts placed in spinal cord lesion sites, axonal bridging beyond the lesion sites has been much more difficult to achieve. The lack of extensive long-distance axonal growth can be attributed to a number of different factors, including the growth-inhibitory environment of the host spinal cord distal to the lesion site and a lack of factors attracting axons out of the favorable environment of a cellular graft. Localized neurotrophin delivery at the site of spinal cord injury is likely

Table 1 Growth of ascending and descending tracts after neurotrophin delivery to the injured spinal cord

Factor	*Response*
NGF	Nociceptive sensory, coerulospinal
BDNF	Nociceptive sensory, rubrospinal, reticulospinal, raphaespinal, coerulospinal, spinal motor axons
NT-3	Proprioceptive sensory, corticospinal
NT-4	Nociceptive sensory, propriospinal, reticulospinal, raphaespinal, coerulospinal, spinal motor axons

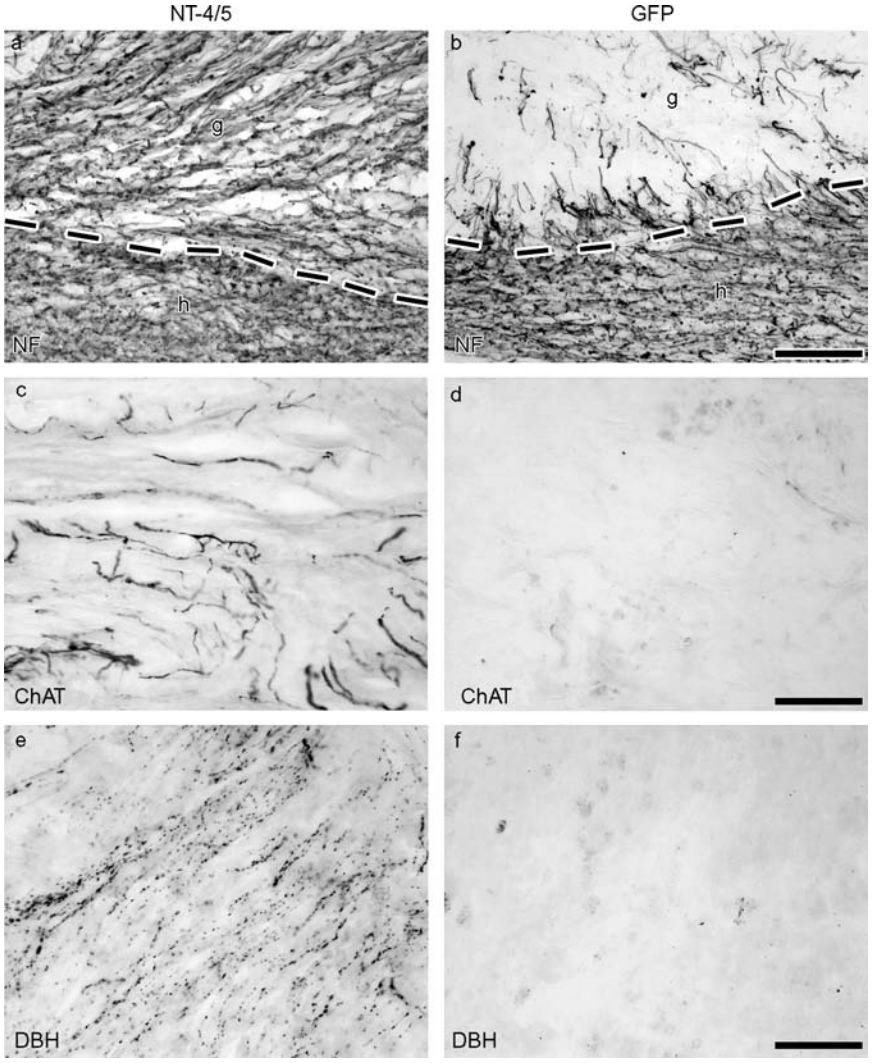

Figure 3 Neurotrophin-secreting fibroblasts grafted to the injured rat spinal cord enhance axonal growth. (a) Fibroblasts genetically modified to express NT-4 induce robust axonal growth into the graft (g), whereas (b) few neurofilament (NF)-labeled axons penetrate control grafts expressing green fluorescent protein (GFP) 3 months postgrafting. (c) Motor axons identified by choline acetyl transferase (ChAT) labeling and (e) numerous coerulospinal axons identified by dopamine β-hydroxylase (DBH) labeling densely penetrate NT-4 secreting cellular grafts, whereas only a few axons are found in control grafts (d, f). Scale bar = 108 μm (a, b); 88 μm (c, d); 64 μm (e, f). Adapted from Blesch A, Yang H, Weidner N, Hoang A, and Otero D (2004) Axonal responses to cellularly delivered NT-4/5 after spinal cord injury. *Molecular and Cellular Neuroscience* 27: 190–201, with permission from Elsevier.

to lead to a concentration gradient where the highest concentration of neurotrophic factors are encountered at the lesion center. Axons are therefore chemotropically attracted to the lesion center, where there is the highest concentration of neurotrophic factors, and may have no incentive for further extension beyond the lesion site. Strategies to enhance axonal regeneration by neurotrophin delivery in the injured spinal cord might therefore require neurotrophin gradients extending beyond the lesion site, combinations with other treatments to overcome axon growth inhibitors present in the CNS, or means of initiating an efficient cellular growth program to activate genes necessary for axonal regeneration (Figure 4).

As mentioned previously, injuries in the PNS result in increased BDNF expression at the injury site and the cell body. Increased BDNF expression in the cell body might contribute to the successful initiation of a genetic regeneration program in DRG neurons. Thus, in addition to the studies described previously that have primarily focused on the site of axonal transection for growth factor delivery, neurotrophin application at the site of the cell soma has been considered as a means to enhance axonal regeneration. In particular, BDNF administration to rubrospinal neuronal cell bodies in the midbrain has been shown to not only prevent lesion-induced cell atrophy but also initiate an increase in the expression of regeneration-associated

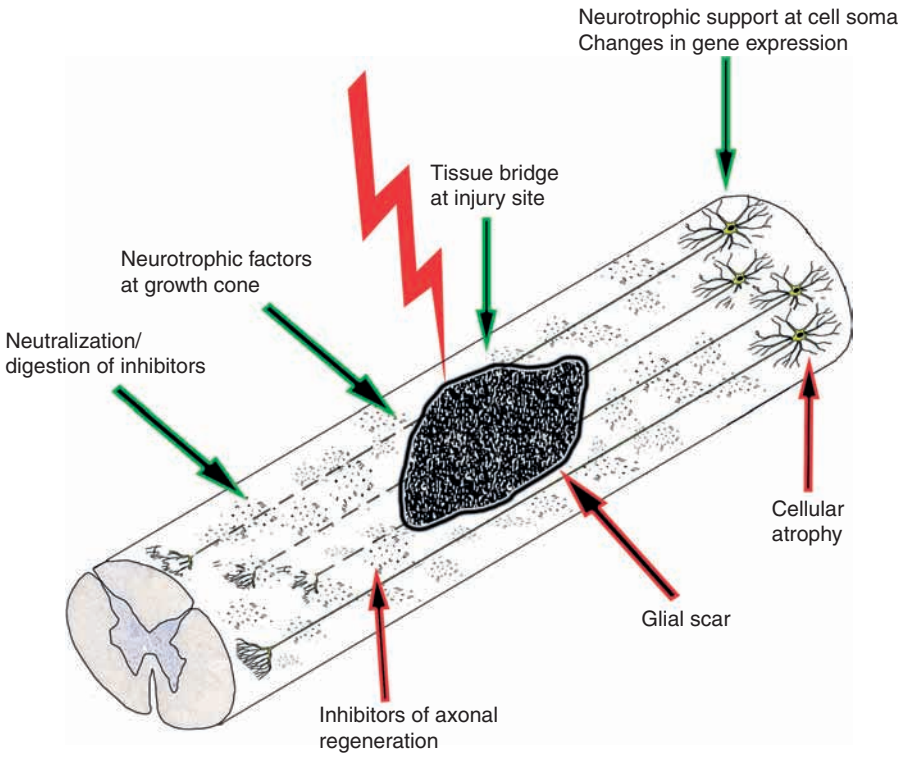

Figure 4 Schematic of positive growth-promoting stimuli (green arrows) and negative growth-inhibiting influences (red arrows) after spinal cord injury.

genes, including GAP-43 and β-tubulin. BDNF delivery at the cell soma also increases axonal growth of rubrospinal axons into a peripheral nerve bridge grafted to the lesioned spinal cord. The induction of a genetic program for axonal regeneration by neurotrophins or by other means is likely to be an important factor or even a prerequisite for optimizing the extent of axonal regeneration.

Optimization of axonal regeneration in the CNS will likely require combinatorial treatments that address several mechanisms that ordinarily limit plasticity and regeneration. Nonetheless, partial functional recovery after neurotrophin administration has been reported in several studies after targeted delivery of BDNF or NT-3 in rodent models of spinal cord injury. For example, NT-3 delivered to sites of midthoracic hemisection spinal cord injury, using genetically modified fibroblasts, induced growth of corticospinal axons for limited distances around the lesion site that was subsequently associated with partial functional recovery. Cellular BDNF delivery after a cervical transection of the dorsolateral funiculus also induced growth of rubrospinal axons for short distances beyond the lesion site, resulting in partial recovery of forelimb function. Combined delivery of both of these growth factors has also been reported to induce functional improvement. Of note, however, mechanisms of functional improvement remain incompletely understood: whereas axonal sprouting or regeneration may account for some neurotrophin-induced recovery, it is also possible that stimulation of motor programs for walking (spinal locomotor pattern generators) may be induced by growth factors independent of inputs from systems above the injury. Furthermore, neurotrophins could also influence myelination or synaptic strength as alternative means of supporting functional improvement. Identification of mechanisms underlying neurotrophin-induced recovery after spinal cord injury requires greater clarification and objective study.

Whereas the majority of studies conducted to date have focused on neurotrophin-mediated effects in acute axonal injury paradigms, patterns of neurotrophin-induced axonal growth after chronic spinal cord injury have generally been found to be similar to those observed after acute injury. Chronically injured neurons appear to remain responsive to neurotrophin delivery for prolonged time periods after the initial injury, when injured axons remain in close proximity to the lesion site. Indeed, most projecting motor systems to the spinal cord survive chronic periods of injury and remain in proximity to the lesion site, representing an axonal pool that can be recruited to potentially undergo regeneration.

Primate studies of axonal plasticity and regeneration in response to neurotrophin delivery have been undertaken. The development of a primate model is

an important step in determining whether findings to date in rodent models are likely to be of relevance to the potential treatment of human disorders.

Conclusion

At least two mechanisms appear to contribute to the regeneration-enhancing properties of neurotrophins: Growth factors attract and stimulate axon growth (a 'neurotropic' property), and growth factors appear to initiate a survival and regeneration program (a 'neurotrophic' property). Spontaneous regeneration in the PNS is accompanied by a tightly regulated program of spatial and temporal changes in neurotrophin expression. Effective therapeutic treatments in CNS regeneration will ultimately require approaches that allow a replication of these temporal and spatial changes in neurotrophin production.

See also: Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Cell Replacement Therapy: Parkinson's Disease; Enteric Nervous System: Neurotrophic Factors; Gene Therapy: Genetically Modified Cells; Gene Therapy: Direct Viral Delivery; Huntington's Disease: Neurodegeneration; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophins: Physiology and Pharmacology; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Peripheral Nerve Regeneration: An Overview; Spinal Cord Regeneration and Functional Recovery: Strategies; Spinal Cord Injuries.

Further Reading

Blesch A and Tuszynski MH (2002) Spontaneous and neurotrophin-induced axonal plasticity after spinal cord injury. *Progress in Brain Research* 137: 415–423.

Hendriks WT, Ruitenberg MJ, Blits B, Boer GJ, and Verhaagen J (2004) Viral vector-mediated gene transfer of neurotrophins to promote regeneration of the injured spinal cord. *Progress in Brain Research* 146: 451–476.

Huang EJ and Reichardt LF (2003) Trk receptors: Roles in neuronal signal transduction. *Annual Review of Biochemistry* 72: 609–642.

Lewin GR and Barde YA (1996) Physiology of the neurotrophins. *Annual Review of Neuroscience* 19: 289–317.

Plunet W, Kwon BK, and Tetzlaff W (2002) Promoting axonal regeneration in the central nervous system by enhancing the cell body response to axotomy. *Journal of Neuroscience Research* 68: 1–6.

Ramer MS, Priestley JV, and McMahon SB (2000) Functional regeneration of sensory axons into the adult spinal cord. *Nature* 403: 312–316.

Ross CA (2004) Huntington's disease: New paths to pathogenesis. *Cell* 118: 4–7.

Sofroniew MV, Howe CL, and Mobley WC (2001) Nerve growth factor signaling, neuroprotection, and neural repair. *Annual Review of Neuroscience* 24: 1217–1281.

Tuszynski MH and Blesch A (2004) Nerve growth factor: From animal models of cholinergic neuronal degeneration to gene therapy in Alzheimer's disease. *Progress in Brain Research* 146: 441–449.

Tuszynski MH, Thal L, Pay M, et al. (2005) A phase 1 clinical trial of nerve growth factor gene therapy for Alzheimer disease. *Nature Medicine* 11: 551–555.

Zweifel LS, Kuruvilla R, and Ginty DD (2005) Functions and mechanisms of retrograde neurotrophin signalling. *Nature Reviews Neuroscience* 6: 615–625.

Neurotrophins: Physiology and Pharmacology

J M Conner and M H Tuszynski, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Nerve growth factor (NGF) was the first neurotrophin family member and the first nervous system growth factor to be identified. NGF was originally isolated from a mouse sarcoma based upon its ability to promote the hypertrophy of, and fiber outgrowth from, peripheral sensory neurons. Subsequently, other neurotrophin family members were identified, including brain-derived neurotrophic factor (BDNF), neurotrophin-3 (NT3), and neurotrophin-4 (NT4). Although originally described as a potentially novel growth factor, neutrotrophin-5 was found to be identical to NT4. Two other neurotrophin members, neurotrophin-6 and-7, were identified in fish but do not appear to have corresponding homologs in mammals.

Neurotrophin Production and Release

Neurotrophin proteins are synthesized as precursors of approximately 270 amino acids (30–35 kDa) and are subsequently cleaved to yield mature proteins of ~120 amino acids (13 kDa) in length. While most physiological actions of neurotrophins have been ascribed to the mature proteins that generally function as soluble, non-covalently linked homodimers, recent studies have identified potentially critical roles for the unprocessed neurotrophin precursors (see later). Elements within the precursor sequence of newly generated neurotrophins presumably direct the intracellular sorting of neurotrophin molecules into one of two distinct pathways, resulting in either constitutive or regulated secretion. This process of intracellular sorting plays a critical role in determining the possible spectrum of physiological actions mediated by a given neurotrophin molecule by determining when and where neurotrophin secretion will take place. While neurotrophins were originally believed to function exclusively as target-derived factors that mediate signaling in a retrograde manner, recent studies have provided substantial evidence that some neurotrophins function partially, or exclusively, as anterograde molecular signals.

Neurotrophin Actions Are Mediated by Distinct Cell Surface Receptors

Once neurotrophins are released from the cell of origin, their actions are mediated by binding to distinct cell surface receptors. The specificity of neurotrophin action is controlled primarily by the pattern of expression of receptors in distinct populations of neurons located at the site of neurotrophin release.

Two distinct classes of neurotrophin receptors have been identified. The p75 neurotrophin receptor (p75NTR) is a member of the tumor necrosis receptor family that binds all of the mature neurotrophins with a similar, nanomolar affinity. Recent studies have indicated that the p75NTR can also bind unprocessed neurotrophin precursors with a high affinity. The second class of neurotrophin receptors is made up of three distinct members of the tropomyosin-related kinase (Trk) receptor family. The various neurotrophins show some degree of specificity with respect to their binding of distinct Trk receptors as shown in Figure 1. Binding of neurotrophins to their respective Trk receptors occurs with an affinity 100-fold greater than binding of mature neurotrophins to the p75NTR. Moreover, the various Trk family members have highly homologous intracellular domains, suggesting they use similar signaling pathways to mediate their effects. Differential splicing of TrkB and TrkC genes can also produce neurotrophin receptors that lack the intracellular kinase domains. The functions of these truncated receptors are unknown, but studies have suggested that they may play a role in regulating the availability or presentation of secreted neurotrophins to full-length and active receptors.

The physiological role of the p75NTR has been a subject of considerable study. The p75NTR was initially thought to act as a coreceptor, conferring high-affinity binding of neurotrophins to their respective Trk receptors and promoting retrograde transport of neurotrophins from their targets. However, in the absence of Trk receptor expression, the p75NTR may promote apoptosis via multiple pathways, including an increase in nuclear factor-kappa B (NF-κB) activity and by inducing sphingomyelin hydrolysis into ceramide (Figure 2). Neurotrophin binding to the p75NTR also leads to increased activation of c-Jun N-terminal kinase (JNK), a stress-activated protein kinase that has also been implicated in cell death. Binding of the precursor for NGF to the p75NTR has recently been demonstrated to increase apoptotic death signaling; the consequence of binding other neurotrophin precursor molecules to this receptor remains to be determined. A key factor in determining the fate of neurotrophin signaling through the p75NTR appears to be the extent to which Trk receptors are co-expressed; in the presence of Trk receptors, neurotrophins primarily support cell survival, whereas

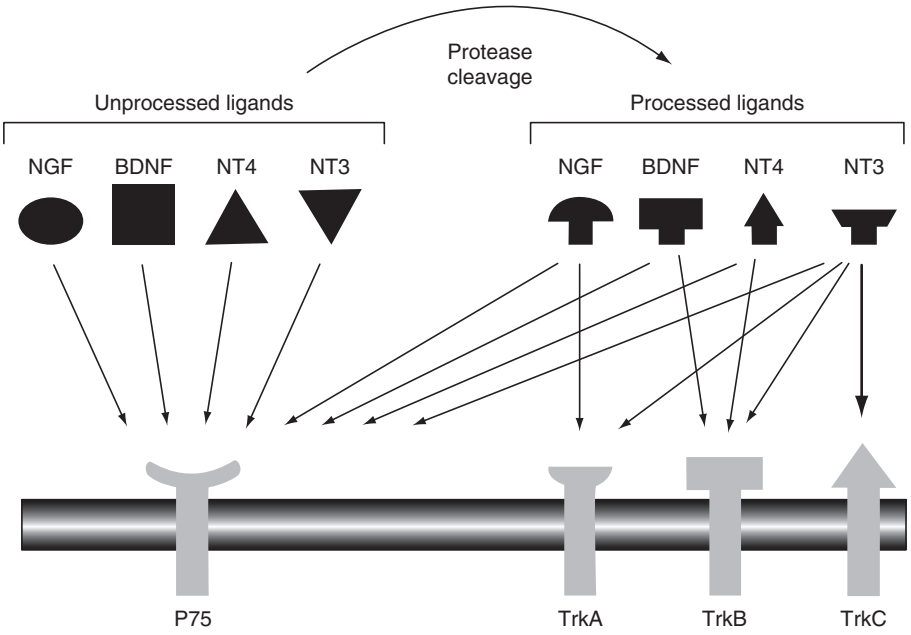

Figure 1 Neurotrophins binding to their receptors. All neurotrophin precursors associate with the p75NTR with similar high-affinity binding. Processed (mature) neurotrophin proteins also bind to the p75NTR with a 100-fold lower affinity. Mature neurotrophins bind with high affinity to members of the tropomyosin-related kinase (Trk) family. Interactions with Trk receptors show a high degree of ligand specificity, with nerve growth factor (NGF) binding selectively to TrkA, brain-derived neurotrophic factor (BDNF) and neurotrophin-4 (NT4) binding to TrkB, and NT3 binding primarily to TrkC. From Segal RA (2003) Selectivity in neurotrophin signaling: Theme and variations. *Annual Review of Neuroscience* 26: 299–330.

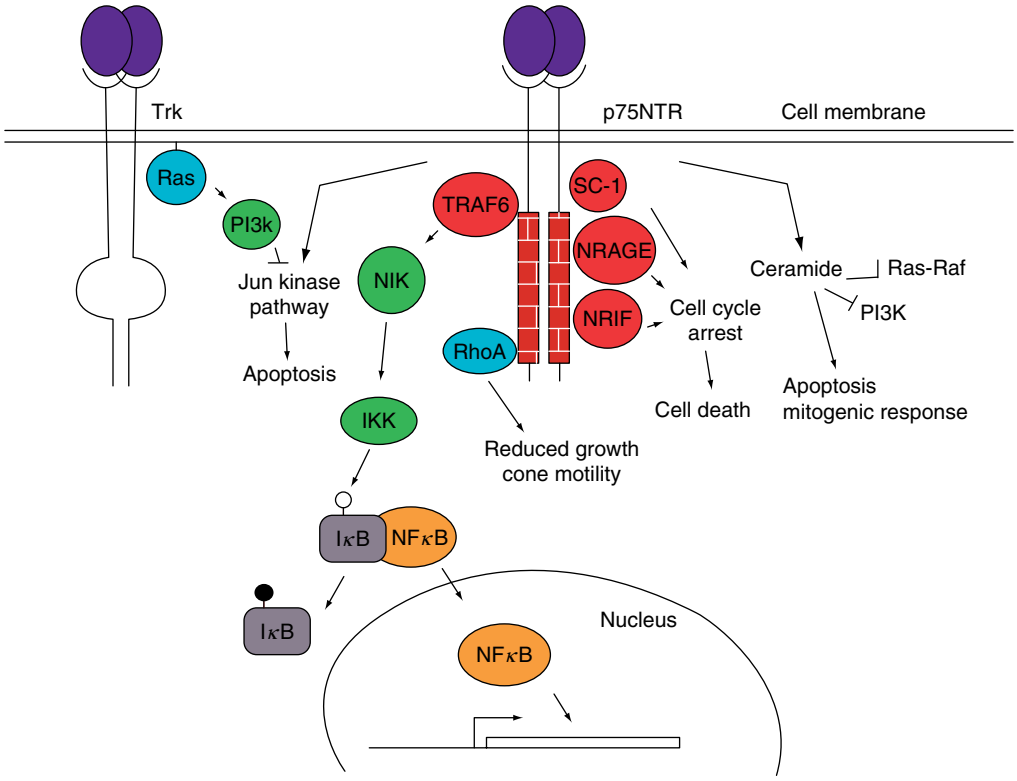

Figure 2 Neurotrophin signaling through the p75 neurotrophin receptor (p75NTR). Binding of neurotrophins to the p75NTR leads to alterations in the function of many associated proteins, leading in turn to changes in gene expression and apoptosis. Adapter proteins are shown in red, kinase proteins in green, small G-proteins in blue, and transcription factors in brown. Trk, tropomyosin-related kinase; PI3K, phosphatidylinositol 3-kinase; NIK, NF-κB-inducing kinase; IKK, inhibitor of κB kinase; IκB, inhibitor of κB; NF-κB, nuclear factor-kappa B; TRAF6, tumor necrosis factor receptor-associated kinase-6; SC-1, Schwann cell factor-1; NRAGE, neurotrophin receptor-interacting MAGE (from melanoma antigen); NRIF, neurotrophin receptor-interacting factor. From Huang EJ and Reichardt LF (2001) Neurotrophins: Roles in neuronal development and function. *Annual Review of Neuroscience* 24: 677–736.

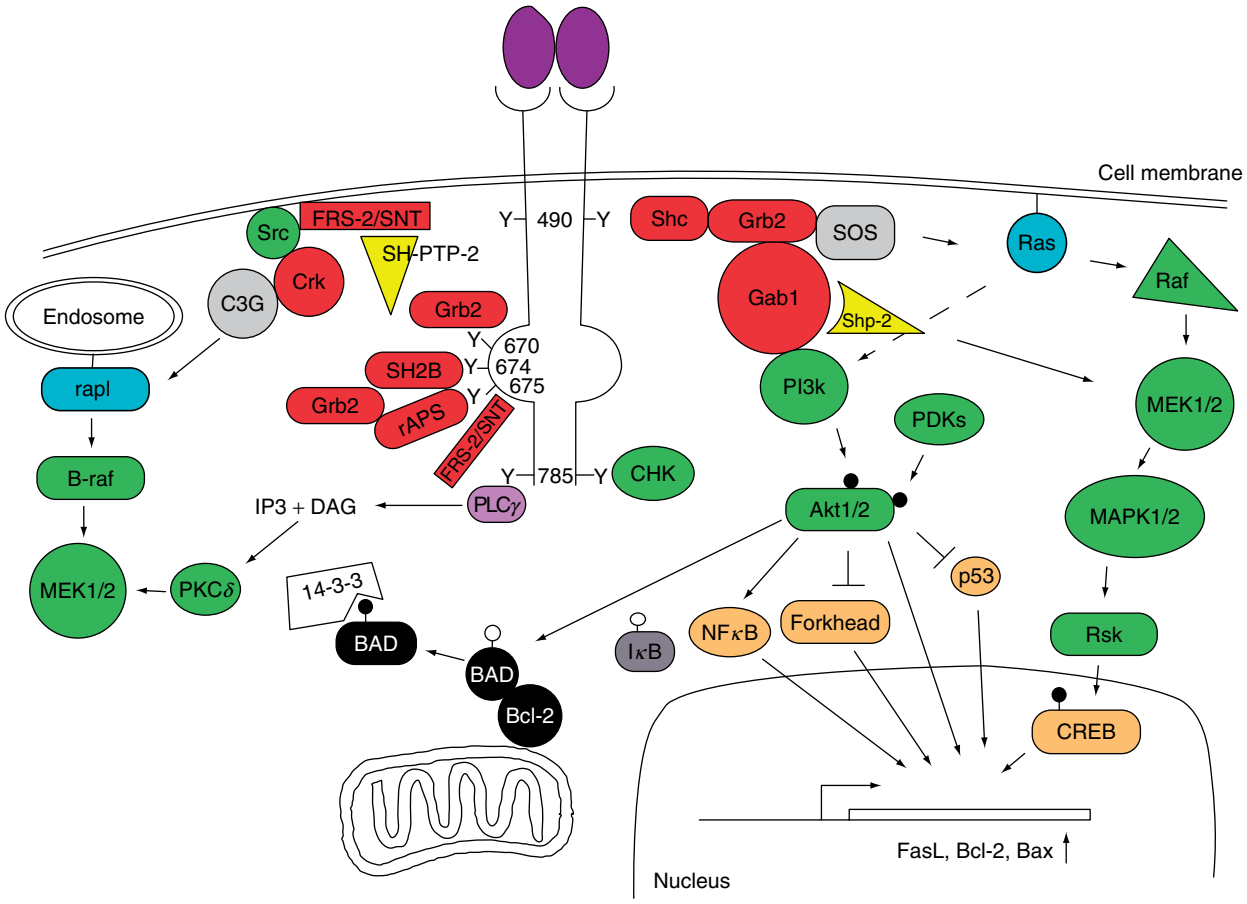

Figure 3 Neurotrophin signaling through Trk receptors. Binding of mature neurotrophins to Trk receptors leads to the phosphorylation of a wide variety of associated proteins, which ultimately induce alterations in gene expression. Neurotrophin actions through Trk receptors affect neuronal survival, neurite outgrowth, cell morphology, and synaptic efficacy. Adapter proteins are shown in red, kinase proteins in green, small G-proteins in blue, and transcription factors in brown. From Huang EJ and Reichardt LF (2001) Neurotrophins: Roles in neuronal development and function. *Annual Review of Neuroscience* 24: 677–736.

in the absence of Trk receptors, neurotrophin signaling through p75NTR primarily mediates cell death.

Binding of mature neurotrophins to their respective Trk receptors induces dimerization of the receptors and the activation of kinase activity. Increased phosphorylation activity mediated through activated Trk receptors drives cellular responses through three distinct molecular pathways as indicated in Figure 3, resulting ultimately in a variety of transcriptional changes. Neurotrophin signaling through Trk receptors can influence many aspects of cell function, including cell survival, neurite outgrowth, and cellular differentiation.

Neurotrophin Physiology

Cell Survival

One of the classic physiological functions attributed to neurotrophins is the regulation of cell survival, specifically within neuronal populations expressing a particular type of neurotrophin receptor. Within the peripheral nervous system, tightly regulated neurotrophin expression is thought to be responsible for controlling neuron number through the process of programmed cell death, whereby a precisely timed and limited expression of target-derived neurotrophins limits the survival of responsive neurons projecting to a given peripheral target. Data from genetically mutant mice with targeted deletions of specific neurotrophins have provided evidence supporting a role for neurotrophins as target-derived regulators of neuronal survival in the periphery. Consistent with known patterns of neurotrophin and receptor expression, targeted deletion of either specific neurotrophins or their receptors results in profound, yet specific, losses of distinct populations of peripheral neurons during development. For instance, targeted deletion of NGF results in a nearly complete loss of peripheral sympathetic and nociceptive sensory neurons that express TrkA receptors. Targeted NGF deletion, however, does not significantly reduce the number of TrkC-expressing peripheral sensory neurons.

Based upon observations made within the peripheral nervous system during development, it was initially postulated that neurons within the central nervous system (CNS) may also rely upon neurotrophin support for their survival, not only during development but also in the adult. In spite of this prediction, there is strikingly little evidence suggesting that neurotrophins play a significant role in promoting cell survival within the CNS, either during development or in the adult. For instance, basal forebrain cholinergic neurons, which possess both p75NTR and TrkA receptors throughout life and show numerous physiological responses to NGF, show minimal deficits in developmental cell survival following targeted deletions of either NGF or its corresponding TrkA receptor. Moreover, deletion of the cellular source of neurotrophins in the adult animal also fails to cause cell death. However, neurotrophins appear to have many other effects upon neurons in the CNS, especially pertaining to cellular differentiation, target innervation, and synaptic plasticity.

Cell Differentiation and Morphology

In addition to their well-defined actions on cell survival, neurotrophins have been implicated in numerous aspects of cellular differentiation. Some of the earliest studies with NGF demonstrated that this neurotrophin induced dramatic outgrowth of neurites from developing peripheral neurons. Subsequently, the various members of the neurotrophin family have been implicated in mediating a variety of morphological changes, in a multitude of distinct neuronal populations, including the extension and guidance of axons, hypertrophy of the cell body, and changes in dendritic complexity and dendritic spine density. For example, respective expression of BDNF or NT3 in cortical target regions of extending axons during brain development appears to define zones of axon growth termination in strikingly laminar patterns, helping to establish the detailed topography of mature innervation. In addition to controlling the initial outgrowth of neuronal processes, neurotrophins have been strongly implicated in the process of maintaining target innervation in both central and peripheral targets. One of the most striking effects observed in NGF knockout animals is the failure of NGF-sensitive cholinergic neurons to retain appropriate innervation of their central targets.

Neurotrophins also play an important role in other aspects of cellular differentiation, helping to determine the ultimate fate and function of certain cells. For example, NGF promotes the differentiation of sympathoadrenal precursors into sympathetic neurons, as opposed to adrenal chromaffin cells. Neurotrophins also play more subtle roles in neuronal function by regulating the expression of neurotransmitters, ion channels, and receptors, both during development and, in many cases, throughout adulthood.

Synaptic Function and Synaptic Plasticity

In addition to regulating neuronal function, neurotrophins have direct actions on synaptic function and synaptic plasticity. Application of neurotrophins at developing and mature synapses can stimulate or modulate neurotransmitter release under many circumstances. The induction of long-term potentiation (LTP), a possible electrophysiological substrate for long-term storage of information in the nervous system, is potently regulated by BDNF. *In vitro* and *in vivo* applications of BDNF to the hippocampal formation are capable of inducing long-lasting synaptic potentiation in the absence of any additional excitatory stimulus. Moreover, the ability to induce LTP in the hippocampus with an electrical stimulus is greatly attenuated when BDNF signaling is disrupted. Together, these results suggest that BDNF may play a critical and necessary role in regulating synaptic plasticity in certain populations of neurons. Whether or not the regulation of synaptic plasticity is a selective property of BDNF is unknown at present, but NGF and NT3 have not been demonstrated to influence synaptic function and LTP in a similar manner.

Neurotrophin Pharmacology

The potent survival effects of neurotrophins on neurons of the peripheral nervous system originally discovered 50 years ago fostered speculation for many years that neurotrophins may also promote the survival of CNS neurons, thereby representing pharmacological candidates for treating neurodegenerative diseases. The degeneration of specific neuronal populations is a hallmark of the most common neurodegenerative disorders, including losses of midbrain dopaminergic neurons in Parkinson's disease and the loss of basal forebrain cholinergic neurons as one component of multisystem cell loss in Alzheimer's disease. The hypothesis that neurotrophins might represent a new class of therapeutic molecules was substantiated by studies published in the mid-1980s, demonstrating that NGF prevented the injury-induced death of basal forebrain cholinergic neurons in adult rats. NGF neuroprotection of adult cholinergic neurons was also reported after lesions in the brains of adult primates. Other studies reported that spontaneous atrophy and dysfunction of basal forebrain cholinergic neurons in aged rats and primates could also be reversed by NGF administration. Further, NGF treatment significantly ameliorated

Table 1 Potential clinical uses for neurotrophins

Neurotrophin	*Targeted cell population*	*Disease*
Nerve growth factor	Basal forebrain cholinergic	Alzheimer's disease
	Sensory nociceptive	Peripheral neuropathy
Brain-derived neurotrophic factor	Cortical neurons	Alzheimer's disease
	Spinal motor neurons	Amyotrophic lateral sclerosis
	Midbrain dopaminergic	Parkinson's disease
	Cortical neurons	Acute brain injury/stroke
Neurotrophin-3	Sensory proprioceptive	Peripheral neuropathy

age-related memory impairments. Collectively, these findings established a theoretical basis for pursuing clinical trials targeting NGF therapy for the treatment of Alzheimer's disease, whereby a significant loss of basal forebrain cholinergic neurons is postulated to contribute to ongoing cognitive decline.

Analogous experimental paradigms explored the possibility that the neurotrophins could protect cell populations affected in a diverse spectrum of neurodegenerative diseases, including nigral neurons, motor neurons of the spinal cord, and dorsal root ganglion sensory neurons (see Table 1).

Clinical Trials with Neurotrophins

The potent actions of neurotrophins in preventing neuronal degeneration and augmenting cell function made them intriguing candidates for clinical testing in human neurological disorders. To date, neurotrophins and other growth factors have been tested in clinical trials in Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis, stroke, and peripheral neuropathy. Trials are also under consideration in treating Huntington's disease and spinal cord injury.

In spite of the extensive animal data supporting the potential use of neurotrophins in treating a variety of neurological diseases, no studies to date have yielded clear evidence of efficacy. One of the most important limiting factors in testing the clinical potential of neurotrophins has been the availability of an effective, safe, and sustained method of delivering growth factors to the brain. Because neurotrophins are large and charged proteins, they do not cross the blood–brain barrier after peripheral administration. When infused into the cerebrospinal fluid space of the brain (intraventricular or intrathecal infusions), growth factors usually fail to diffuse to neuronal targets. In the case of NGF, intraventricular or intrathecal infusions broadly disseminate the neurotrophin throughout the subarachnoid space of the brain and spinal cord, inducing the migration of Schwann cells into the CNS, stimulating nociceptive axons of the dorsal root ganglia (DRG) (resulting in pain), eliciting sympathetic axon sprouting around the cerebral vasculature, and causing weight loss. Thus, clinical testing of neurotrophins requires a delivery method that both achieves adequate growth factor concentrations at target neurons that may lie deep within the brain, and restriction of growth factor delivery to target regions in order to avoid adverse effects of broad growth factor dissemination in the cerebrospinal fluid space. Chronic intraparenchymal growth factor infusion into the brain, or growth factor gene delivery into the brain, may be a means of achieving these goals.

Indeed, recent signs of possible growth factor activity in Alzheimer's disease and Parkinson's disease have been reported using targeted intraparenchymal delivery methods. In Alzheimer's disease, a recent phase I clinical trial of NGF gene delivery to cholinergic neurons was reported to increase glucose utilization in the cortex by positron emission scanning, and to elicit new growth of axons toward the NGF source. Potential effects on cognition are being explored in phase II clinical trials of NGF gene delivery in Alzheimer's disease. A clinical trial of growth factor gene delivery in Parkinson's disease is also under way (using neurturin, a glial-derived neuronal growth factor). Three recent clinical trials using intraparenchymal infusion of glial-derived neuronal factor (GDNF) have also been conducted in Parkinson's disease. The development of these targeted, restricted, and sustained growth factor delivery methods raises the possibility that the next few years will definitively establish whether growth factors will prove useful for the treatment of CNS disorders.

Neurotrophins have also undergone clinical testing for peripheral neuropathy: NGF is a tropic factor for a subpopulation of nociceptive sensory neurons with cell bodies in the DRG, and NT3 is a tropic factor for large-diameter proprioceptive neurons of the DRG. NGF was tested in patients with diabetes using intermittent peripheral subcutaneous injections, but failed to show efficacy. Once again, problems with potentially subtherapeutic doses, inadequate targeting, and rapid peripheral degradation of NGF limited the

interpretability of the study. NT3 is also being tested in clinical trials for drug-induced peripheral neuropathy. Solutions to the challenge of targeted, sustained, and safe growth factor delivery for treatment of peripheral nervous system disorders await identification, to determine whether early potential signals of growth factor activity in the CNS may also be achievable in the peripheral nervous system.

See also: Enteric Nervous System: Neurotrophic Factors; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Proteasome Role in Neurodegeneration.

Further Reading

Chao MV (2003) Neurotrophins and their receptors: A convergence point for many signalling pathways. *Nature Reviews Neuroscience* 4: 299–309.

Dechant G and Barde YA (2002) The neurotrophin receptor p75 (NTR): Novel functions and implications for diseases of the nervous system. *Nature Neuroscience* 5: 1131–1136.

Hefti F (1997) Pharmacology of neurotrophic factors. *Annual Review of Pharmacology and Toxicology* 37: 239–267.

Hempstead BL (2006) Dissecting the diverse actions of pro- and mature neurotrophins. *Current Alzheimer Research* 3: 19–24.

Huang EJ and Reichardt LF (2001) Neurotrophins: Roles in neuronal development and function. *Annual Review of Neuroscience* 24: 677–736.

Huang EJ and Reichardt LF (2003) Trk receptors: Roles in neuronal signal transduction. *Annual Review of Biochemistry* 72: 609–642.

Lessmann V, Gottmann K, and Malcangio M (2003) Neurotrophin secretion: Current facts and future prospects. *Progress in Neurobiology* 69: 341–374.

McAllister AK, Katz LC, and Lo DC (1999) Neurotrophins and synaptic plasticity. *Annual Review of Neuroscience* 22: 295–318.

Reichardt LF and Mobley WC (2004) Going the distance, or not, with neurotrophin signals. *Cell* 118: 141–143.

Segal RA (2003) Selectivity in neurotrophin signaling: Theme and variations. *Annual Review of Neuroscience* 26: 299–330.

Thoenen H and Sendtner M (2002) Neurotrophins: From enthusiastic expectations through sobering experiences to rational therapeutic approaches. *Nature Neuroscience* 5(supplement): 1046–1050.

Tuszynski MH, Thal L, Pay M, et al. (2005) A phase I clinical trial of nerve growth factor gene therapy for Alzheimer's disease. *Nature Medicine* 11: 551–555.

Neurulation

A Lawson, University of Ghana, Accra, West Africa
G C Schoenwolf, University of Utah School of Medicine, Salt Lake City, UT, USA

© 2009 Elsevier Ltd. All rights reserved.

Phases and Stages of Neurulation

The central nervous system is one of the first organ systems to develop in vertebrate embryos. The process by which this occurs is called neurulation. Neurulation is a developmental event that results in formation of the neural tube, the early hollow rudiment of the adult central nervous system. Once formed, the neural tube differentiates into a rostral portion, the future brain, and a caudal portion, the future spinal cord. The future brain further divides into three main swellings, the forebrain, midbrain, and hindbrain, and further subdivision occurs throughout the rostrocaudal extent of the neural tube with the formation of segmental swellings called neuromeres.

Neurulation occurs in two sequential phases termed primary and secondary neurulation. Primary neurulation occurs in four clearly defined but overlapping steps. In the first step, the neural plate is formed from epiblast, one of two primitive layers composing the early embryo. In the second step, the neural plate undergoes shaping, extending its length and narrowing its width. In the third step, the neural plate bends to form a neural groove, flanked by bilateral neural folds. In the final step, the tips of the neural folds meet and fuse in the dorsal midline to generate the neural tube. Secondary neurulation also involves a series of steps. In the first step, a solid mass of cells, derived from the tail bud, forms. This solid mass is called the medullary cord. In the second step, the medullary cord cavitates to form multiple lumina. In the final step, all lumina coalesce into a single, central lumen bounded by the walls of the so-called secondary neural tube. The processes of primary and secondary neurulation give rise to different rostrocaudal levels of the neural tube. Primary neurulation forms the entire brain and most of the length of the spinal cord (down to about the lumbosacral level), whereas secondary neurulation forms the tailmost portion of the spinal cord. Because much more is known about the cellular and molecular mechanisms of primary neurulation, compared to secondary neurulation, and because most of the neural tube forms during primary neurulation, further discussion herein focuses only on primary neurulation. Also, the following discussion of the tissue and cellular bases of primary neurulation is based principally on the chick, because this model system has been the main one used to unravel the morphogenetic mechanisms underlying neurulation.

Tissue and Cellular Events Underlying Primary Neurulation

The neural plate is formed during gastrulation when epiblast cells located rostral to and beside Hensen's node and the cranial portion of the primitive streak respond to signals from the node by a process known as neural induction. The epiblast cells undergo apicobasal thickening to generate a neural plate, which at this initial stage of development is broad mediolaterally and short rostrocaudally. Once the epiblast is committed to such a neural fate, formation of the neural plate becomes autonomous to the committed cells and does not require the presence of either Hensen's node or nonneural ectodermal cells.

Neural plate shaping is the process by which the broad and short neural plate becomes narrowed transversely and elongated rostrocaudally (Figures 1(a)–1(c)). This is achieved through a convergent-extension movement of the neural plate. A convergent-extension movement consists of a simultaneous rostrocaudal lengthening and mediolateral (transverse) narrowing of the neural plate (Figure 1(d)). The neural plate also undergoes further apicobasal thickening during its shaping.

The neural plate begins to bend while neural plate shaping is underway. This commences in the cranial region first, and in the spinal cord region later. The process resembles the closing of a door as it rotates around its hinges. The midline of the neural plate remains fixed during bending, creating a median hinge point (MHP) along essentially the entire rostrocaudal extent of the neuraxis, whereas the lateral aspects of the neural plate are elevated gradually, forming the neural groove and incipient neural folds (i.e., the initial folds at the lateral margins of the neural plate) (Figures 2(a)–2(c)). Subsequently, additional bending occurs in the neural plate close to its junction with the epidermal ectoderm, resulting in the formation of two dorsolateral hinge points (DLHPs) (Figures 3(a)–3(c)). This is seen at prospective cranial levels of the neuraxis, as well as the extreme caudal end of the future spinal cord region. The morphological hinge points are localized regions where the neural plate is anchored to adjacent tissues – the MHP to the prechordal plate mesoderm and notochord and the DLHPs to the adjacent surface ectoderm of the

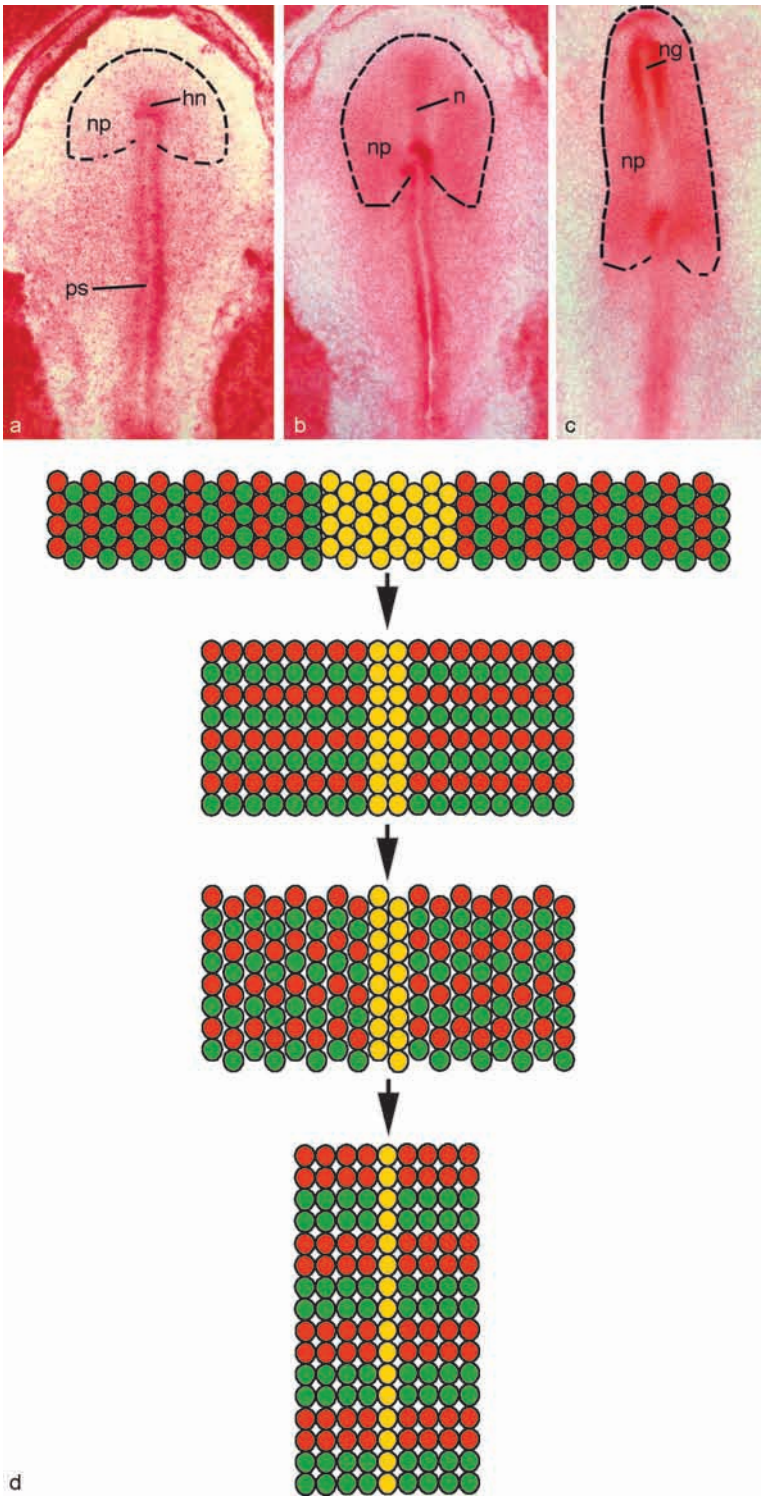

Figure 1 Whole-mounts of chick embryos undergoing primary neurulation, viewed from the dorsal surface (a–c), and a model of neuroepithelial cell rearrangement during neural plate shaping (d). (a) The neural plate (np) has just formed; its approximate borders are outlined (hn, Hensen's node; ps, primitive streak). (b) The neural plate is undergoing shaping; its approximate borders are outlined (n, notochord). (c) The neural plate is initiating bending, establishing a neural groove (ng), while still undergoing shaping. (d) Cell rearrangement occurs during shaping of the neural plate, thereby increasing its length while decreasing its width; median hinge point neural plate cells, yellow; lateral neural plate cells, red and green.

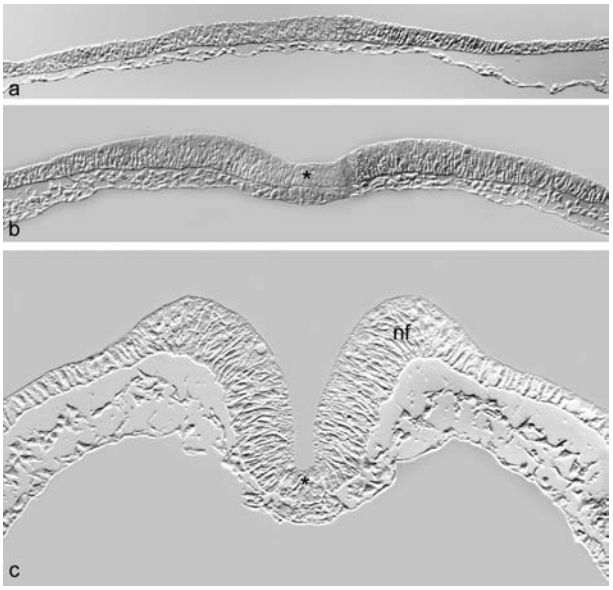

Figure 2 (a–c) Cross-sectional views showing formation of the median hinge point (*) and elevation of the incipient neural fold (nf) at the future midbrain level of the chick. The light micrographs are transverse plastic sections at three sequential stages.

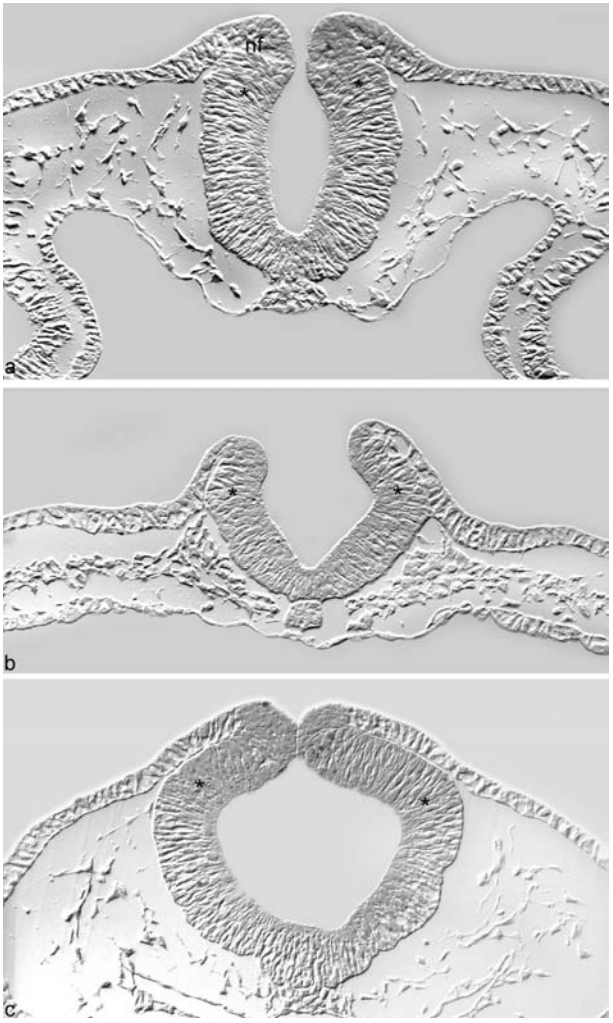

Figure 3 Cross-sectional views showing formation of the dorsolateral hinge points (*), formation and convergence of the definitive neural folds (nf), and closure of the neural groove in the chick. The light micrographs are plastic transverse sections through the future midbrain (a) and hindbrain (b) levels at one stage, and through the future forebrain level (c) at a later stage.

neural folds. Thus hinge points stabilize the neural plate during bending. Bending about the DLHPs results in convergence of the definitive neural folds, directing their tips medially until they meet in the dorsal midline in readiness for fusion (Figure 3(c)). Elsewhere (i.e., throughout most of the length of the spinal cord), where DLHPs are not formed, bending occurs about the MHP only, bringing the apical surfaces of the neuroepithelial cells into apposition with each other (Figure 4). This results in a temporary occlusion of the spinal canal. The absence of the DLHPs, and therefore neural fold convergence, in the spinal cord is one of the major differences between neurulation at cranial and spinal cord levels of the neuraxis.

The neural folds are located at the lateral margins of the neural groove and include the junction of the neural plate with the adjacent epidermal ectoderm. The incipient neural folds first appear as the neural plate folds about the MHP. Later, as folding occurs about the DLHPs, the definitive neural folds become evident. Each neural fold is double-layered, composed of an inner neuroepithelial layer and an outer epidermal ectodermal layer. Four key events occur in sequence during formation and morphogenesis of the neural folds. They are epithelial ridging, kinking, delamination, and apposition (Figure 5). Epithelial ridging is characterized by the formation of a ridge at the prospective outer epidermal–inner neuroepithelial transition zone (Figure 5(a)). It results from the difference in heights of the epidermal ectoderm (ee) and

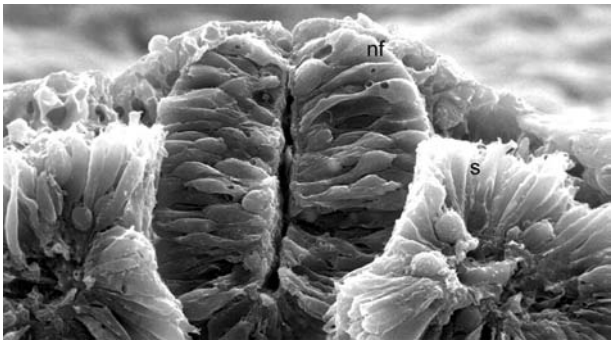

Figure 4 Scanning electron micrograph showing apposition of the neural folds (nf) in the chick at the spinal cord level (level of the seventh somites, S) of the neuraxis. Fusion of the neural folds is in progress and the neurocele is occluded.

neuroepithelial (NE) cells, and it initiates the formation of the neural folds. The NE cells increase their heights, whereas the ee cells decrease theirs. The next step is epithelial kinking (Figure 5(a)), defined as the formation of a concave curvature centered at the prospective ee–NE interface. The main feature here is a change in cell shape within the incipient neural folds; the cells here have constricted bases and expanded apical ends, and are described as inverted wedge-shaped, in sharp contrast to the morphology of the cells at the hinge points, which are described as wedge-shaped. The third event is epithelial delamination (Figure 5(b)), defined as the splitting of a single epithelium into two epithelial layers. The process results in the formation of a linear interface at the ee–NE transition zone between the two epithelial layers. Delamination occurs by cavitation and the deposition of extracellular matrix at the interface. The cells then reorient radially with respect to the interface and acquire intercellular junctions between either adjacent ee cells or adjacent NE cells, but never across the interface between ee and NE cells. The final event of neural fold morphogenesis is epithelial apposition (Figure 5(c)). This is the increase in width of the interface formed during epithelial delamination. The hallmark of this stage of neural fold morphogenesis is a mediolateral expansion of the NE layer, coupled with an apicobasal flattening of the ee layer. Regional differences exist in the degree of epithelial apposition along the rostrocaudal extent of the neuraxis. Generally, where DLHPs are formed and convergence of the neural folds occurs, as seen at forebrain and midbrain levels, apposition is extreme. However, where true DLHPs do not form, as at rostral spinal cord levels, epithelial apposition is lacking.

The final stage of primary neurulation involves the closure of the neural groove. Of all of the steps of avian neurulation, this is the step that has been least explored. Following neural fold convergence, the tips of the neural folds are brought into apposition with each other in the dorsal midline where fusion occurs. In the chick embryo, neural groove closure commences in the prospective midbrain region before progressing rostrally to involve the forebrain, and caudally to involve the hindbrain and spinal cord.

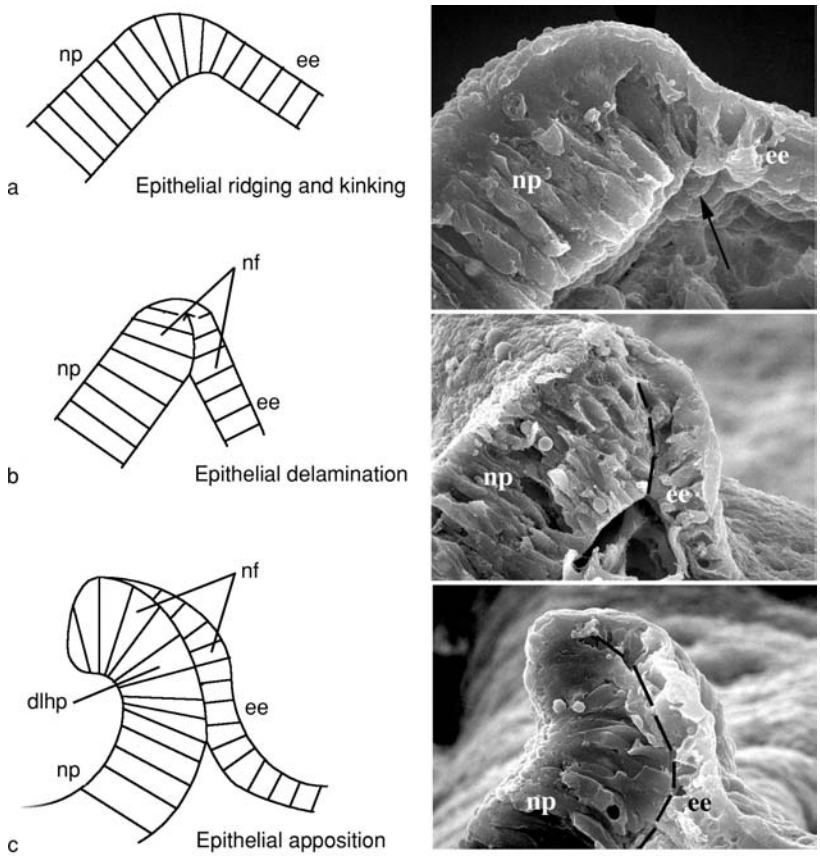

Figure 5 Drawings and scanning electron micrographs showing the four key events of neural fold formation and morphogenesis at the future brain level in the chick embryo: epithelial ridging and kinking (a, b), epithelial delamination (b), and epithelial apposition (c). dlhp, dorsolateral hinge point; ee, epidermal ectoderm; nf, neural fold; np, neural plate. Arrow on micrograph indicates the neural ridge and point of kinking (a). Dashed lines on micrographs indicate the neural fold interface (b, c).

This leaves the rostral and caudal ends of the neural tube temporally open; these openings are known as the rostral and caudal neuropores, respectively. At the tip of each neural fold at most rostrocaudal levels, the epidermal ectoderm caps the neuroepithelium. Apposition of the neural folds, therefore, brings the two epidermal ectodermal layers in contact with each other. Hence, fusion of the neural fold commences with the epidermal ectoderm. As fusion progresses, the newly formed epidermal ectoderm layer delaminates from the neuroepithelium. This brings the neuroepithelium layers of each fold into apposition for fusion of the neuroepithelium to commence. Neural fold fusion is, therefore, described as a double fusion involving two epithelial layers.

Cell Behaviors Generate Tissue Forces in Primary Neurulation

Formation of the neural plate in the chick embryo occurs through cell palisading, which is apicobasal thickening owing to cell elongation. The neural plate so formed consists of a pseudostratified columnar epithelium with two principal cell types, namely, spindle-shaped or fusiform, and wedge-shaped or flask-shaped, as well as other cell types (spherical and inverted wedge-shaped) present in far fewer numbers. The increase in the heights of neuroepithelial cells during neural plate formation is attributed largely to the activity of paraxial microtubules aligned along the long axes of the cells. But other factors such as cell packing and changes in cell–cell adhesion may also contribute to the increase in thickness of the neural plate.

Shaping of the neural plate is largely the result of convergent-extension movements generated by cells within the neural plate. The cell behaviors that account for this movement during shaping of the neural plate in the chick embryo include neuroepithelial cell rearrangement and oriented or nonrandomized cell division. Neuroepithelial cells undergo approximately two rounds of rearrangements (i.e., cell–cell intercalation) during neurulation, narrowing the width of the neural plate by about half during each round, while simultaneously doubling its length with each round (Figure 1(d)). The process is believed to be driven by interkinetic nuclear migration, as well as by cell protrusive activity. Additionally, lengthening of the neural plate is achieved by oriented neuroepithelial cell division. On average, chick neuroepithelial cells undergo two to three rounds of cell division over a 24 h period, the plane of cleavage being at right angles to the orientation of the mitotic spindles. In about half of these divisions, daughter cells are placed in the long axis of the neural plate, thereby contributing to its rostrocaudal extension.

Bending of the neural plate involves two main events, neural plate furrowing and neural plate folding. Furrowing of the neural plate involves the formation of longitudinal furrows at the three morphological hinge points (one MHP and two DLHPs) (Figure 6). Furrowing is driven by changes in neuroepithelial cell shapes at the hinge points. A significant number of neuroepithelial cells at the MHP (70%) and DLHPs (55%) become wedge-shaped during furrowing (Figure 5). In contrast, less than 35% of neuroepithelial cells outside the hinge points become wedge-shaped during neurulation, with a majority of these cells remaining spindle-shaped. Furrowing at the hinge points, therefore, is generated through neuroepithelial cell wedging. Neuroepithelial cell wedging is achieved through both apical constriction and basal expansion of neuroepithelial cells. These processes are likely mediated by contraction of apical bands of microfilaments and translocation of cell nuclei to the bases of the cells during interkinetic nuclear migration, respectively. The signal for furrowing at the MHP comes from the underlying notochord.

Unlike furrowing, folding of the neural plate is mediated by the lateral nonneuroepithelial tissues, principally the epidermal ectoderm, assisted by the mesoderm, endoderm, and the extracellular matrix underlying the neural plate. The cell behaviors in the epidermal ectoderm responsible for folding include cell flattening, oriented cell division, and cell–cell intercalation.

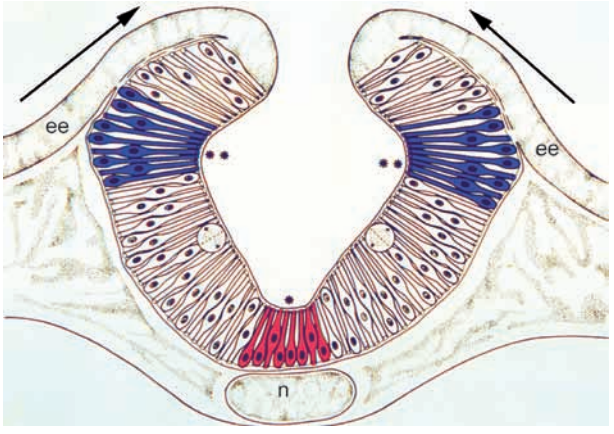

Figure 6 Drawing illustrating the cooperative (hinge point) model of bending of the chick neural plate. Neuroepithelial cell wedging within the hinge points is indicated by red (median hinge point) and blue (dorsolateral hinge points). Arrows indicate mediolateral expansion of the epidermal ectoderm. Single asterisk indicates furrowing associated with the median hinge point; double asterisks indicate furrowing associated with the dorsolateral hinge points; ee, epidermal ectoderm; n, notochord.

Both Intrinsic and Extrinsic Tissue Forces Drive Primary Neurulation

Several studies demonstrate that the forces driving neurulation reside both within the neural plate (intrinsic), as well as outside of this structure (extrinsic). Intrinsic forces are responsible for neural plate shaping and furrowing, and they are generated by behaviors of neuroepithelial cells, such as changes in cell shape, position, and number. In contrast, extrinsic forces generated outside the neuroepithelium account for neural plate folding and neural groove closure. Experimental evidence implicates the epidermal ectoderm in generating a major extrinsic motive force for folding of the neural plate. First, separation of the neural plate and underlying layers from the epidermal ectoderm and underlying layers results in shaping and furrowing of the neural plate, but not folding. Second, removal of the neural plate and underlying layers, leaving intact epidermal ectoderm and underlying layers, results in medial expansion of the epidermal ectoderm. This provides direct evidence of a medially directed epidermal expansion. Third, removal of the epidermal ectoderm, leaving the underlying mesoderm and endoderm intact, stops folding, whereas folding occurs normally after removal of the mesoderm and endoderm beneath the epidermal ectoderm while leaving the latter intact. Finally, when epidermal ectodermal cells at the ee–NE transition zone are tagged with a fluorescent marker and their movement is followed over time, the labeled cells thin and spread medially during neural fold morphogenesis. Primary neurulation thus requires both intrinsic and extrinsic forces acting in concert.

Toward a Molecular Understanding of Primary Neurulation

Progress in understanding the molecular basis of primary neurulation has occurred on two fronts: neural induction and shaping and bending of the neural plate. In regard to the former, the neural plate is now known to be the default state of the ectoderm, and induction of the neural plate actually involves suppressing the formation of the epidermal ectoderm (neural induction is not considered further herein). In regard to shaping and bending of the neural plate, about 100 mutations in the mouse have been identified that result in defective shaping and/or bending and, consequently, result in neural tube defects (NTDs); thus, these mutations provide insight into which genes are involved in both normal and abnormal neurulation. Because these form-shaping events of neurulation are driven by changes in cell behavior, as discussed earlier, it is not surprising that mutation of cytoskeletal, extracellular matrix/cell adhesion, cell cycle, and cell death genes results in NTDs. Neurulation is a highly choreographed morphogenetic series of events that must be precisely timed and coordinated across multiple tissues. This presumably involves signaling among tissues. It is the hope of studies using mouse mutations that such signaling pathways will be identified, ultimately leading to an understanding of the molecular basis of neurulation and the formation of NTDs in both animal models and ultimately in humans. In the following sections, we discuss what has been learned about the molecular basis of shaping and bending of the neural plate.

Shaping and Bending of the Neural Plate

Planar-Cell Polarity Pathway and Convergent Extension

As discussed earlier, convergent extension plays a major role in neurulation. Recent studies have revealed that convergent extension is regulated by the Wnt (wingless) signaling pathway. During development, epithelial sheets become polarized not only apicobasally but also within the plane of the epithelium. In *Drosophila*, the planar-cell polarity (PCP) pathway functions in this latter polarization of the epithelium. Thus, for example, the orientation of wing hairs is established by the PCP pathway. In vertebrates, the PCP pathway is required for proper orientation of stereociliary bundles in the outer hair cells of the mouse inner ear, and for convergent extension during gastrulation and neurulation. How are the PCP and Wnt signaling pathways related?

The *Drosophila* PCP pathway consists of several core proteins that collectively act to convert an extracellular polarity cue into specific changes in the cytoskeleton. These core proteins are now known to be components of the Wnt signaling pathway, and orthologs of several of the *Drosophila* components are conserved in vertebrates. Thus, convergent extension during gastrulation and neurulation is blocked in loss-of-function mutations of the cytoplasmic protein dishevelled in *Xenopus* and its two orthologs in mouse (dishevelled 1 and 2). Wnt signaling involves both a so-called canonical Wnt pathway and a noncanonical Wnt pathway. The PCP pathway utilizes the noncanonical pathway in which certain Wnts, such as Wnt11, bind to their receptors (known as frizzleds). Several other proteins, including dishevelled, must interact in this pathway for proper signaling (and, consequently, for proper convergent extension) to occur. In addition to double dishevelled 1 and 2 mutants, four other mouse mutants exhibit convergent-extension defects: circletail, crash, spin cycle, and loop-tail.

Loop-tail mice have a mutation in the ortholog of the strabismus/van gogh-like gene (*Vangl1*, *Vangl2*), which encodes a transmembrane protein that interacts with dishevelled. Both crash and spin cycle mice have a mutation in the ortholog of the *Drosophila* protocadherin flamingo gene called *Celsr1*. In *Drosophila*, flamingo is required for PCP signaling. Circletail mice have a mutation in the ortholog of the *Drosophila* scribble gene (*Scrib*). Scribble interacts with strabismus. Thus, obtaining an understanding of the PCP pathway in *Drosophila* has had a surprising result – a better understanding also of vertebrate gastrulation and neurulation, and, potentially, a better understanding of how NTDs form in humans.

Actin-Binding Proteins and Apical Constriction

Several actin-associated proteins, when genetically ablated in mice, result in NTDs. One of these, the actin-binding protein shroom, has received considerable study. Overexpression of shroom in cultured epithelial cells is sufficient to cause apical constriction. Shroom causes apical constriction by altering the distribution of F-actin to the apical side of epithelial cells and regulating the formation of a contractile actomyosin network associated with apical intercellular junctions. When shroom is inactivated in *Xenopus* embryos, hinge point formation is drastically altered and neural tube closure fails to occur, providing further evidence for a role of cell shape changes in generating intrinsic forces important for neurulation.

Toward the Prevention of Neural Tube Defects

Neurulation is a complex developmental process that often goes awry in human embryos, leading to neural tube defects, serious congenital anomalies that are severely debilitating and sometimes life threatening. Although the tissue and cellular bases of neurulation are now well understood, the challenge for future studies will be to dissect the molecular basis and provide molecular candidates for mutational analysis in humans. Once such candidate genes are discovered, early and accurate detection of NTDs will be a realty, and their total prevention may be on the horizon. Although the underlying mechanisms remain unclear, periconceptional folic acid supplementation has been shown to reduce the incidence of NTDs in human infants dramatically. Thus, by understanding the process of neurulation better, the hope remains that some day NTDs, severely debilitating birth defects, may no longer occur.

See also: Forebrain Development: Prosomere Model; Forebrain Development: Holoprosencephaly (HPE); Midbrain Patterning; Neural Crest; Neural Crest Cell Diversification and Specification: Melanocytes; Neural Crest Cell Diversification and Specification: ErbB Role; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation; Neural Patterning: Arealization of the Cortex.

Further Reading

Colas J-F and Schoenwolf GC (2001) Towards a cellular and molecular understanding of neurulation. *Developmental Dynamics* 221: 117–145.

Copp AJ, Greene ND, and Murdoch JN (2003) Dishevelled: Linking convergent extension with neural tube closure. *Trends in Neuroscience* 26: 453–455.

Greene ND and Copp AJ (2005) Mouse models of neural tube defects: Investigating preventive mechanisms. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* 135: 31–41.

Haigo SL, Hildebrand JD, Harland RM, et al. (2003) Shroom induces apical constriction and is required for hingepoint formation during neural tube closure. *Current Biology* 13: 2125–2137.

Harland R (2000) Neural induction. *Current Opinion in Genetic Development* 10: 357–362.

Kiefer JC (2005) Planar cell polarity: Heading in the right direction. *Developmental Dynamics* 233: 695–700.

Ladher R and Schoenwolf GC (2005) Making a neural tube: Neural induction and neurulation. In: Rao MS and Jacobson M (eds.) *Developmental Neurobiology*, pp. 1–39. Kluwer Academic/Plenum: New York.

Sasai Y (2001) Regulation of neural determination by evolutionarily conserved signals: Anti-BMP factors and what next. *Current Opinion in Neurobiology* 11: 20–26.

Schoenwolf GC (1994) Formation and patterning of the avian neuraxis: One dozen hypotheses. In: *Neural Tube Defects*, pp. 25–50. Chichester: Wiley.

Schoenwolf GC and Smith JL (1990) Mechanisms of neurulation: Traditional viewpoint and recent advances. *Development* 109: 243–270.

Schoenwolf GC and Smith JL (1990) Epithelial cell wedging: A fundamental cell behavior contributing to hinge point formation during epithelial morphogenesis. In: Keller RE and Fristrom D (eds.) *Control of Morphogenesis by Specific Cell Behaviors. Seminars in Developmental Biology*, pp. 325–334. London: W.B. Saunders.

Smith JL and Schoenwolf GC (1997) Neurulation: Coming to closure. *Trends in Neuroscience* 20: 510–517.

Torban E, Kor C, and Gros P (2004) Van Gogh-like2 (Strabismus) and its role in planar cell polarity and convergent extension in vertebrates. *Trends in Genetics* 20: 70–577.

Wilson SI and Edlund T (2001) Neural induction: Toward a unifying mechanism. *Nature Neuroscience* 4(supplement): 1161–1168.

Neutrotransmission and Neuromodulation: Acetylcholine

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

Acetylcholine is a major neurotransmitter in the peripheral nervous system, with release sites at synapses on all preganglionic and postganglionic parasympathetic neurons, on all preganglionic and some postganglionic sympathetic neurons, and at neuromuscular junctions of skeletal muscle fibers. It is also a neurotransmitter and neuromodulator in the brain, with particularly important roles in learning and memory. Acetylcholine is synthesized from choline and acetyl-coenzyme A (CoA), and is broken down by a dedicated enzyme rather than relying solely on reuptake to eliminate its activity, thus allowing the rapid initiation and termination of signaling when required. There are two distinct classes of acetylcholine receptor, one for rapid responses and one involving second messengers and facilitating longer term effects. Each class of receptor has multiple forms and complex expression patterns in the nervous system and target organs.

Discovery and History

The acetylcholine molecule (Figure 1) was discovered in 1914 by Henry Hallett Dale, who first suggested its possible role as a neurotransmitter. However, it was experiments carried out by Otto Loewi in 1921 that provided definitive proof. Loewi showed that an isolated frog's heart with an intact vagus nerve would continue to beat if bathed in a simple saline solution, and that electrical stimulation of the vagus nerve led to a reduction in the heart rate. If the solution bathing the first heart was then transferred to a second heart that lacked a vagus nerve, that heart too would slow down. Loewi deduced correctly that a diffusible chemical signal released from the first heart was causing the second heart to slow down, and that stimulation of the vagus nerve was triggering its release. He named the molecule 'vagusstoff' because its effects mimicked stimulation of the vagus nerve, but it was later shown to be identical to Dale's acetylcholine. Dale and Loewi were jointly awarded the 1936 Nobel prize in physiology or medicine for their work.

Synthesis, Release, and Recycling

Acetylcholine is synthesized from choline and acetyl-CoA in a reaction catalyzed by the enzyme choline acetyltransferase (Figure 2). This reaction takes place in the cytoplasm of the nerve terminal, to which both reactants are transported independently. Choline comes originally from phosphatidylcholine in the diet (this is the major phospholipid component of animal and plant cell membranes). However, choline is also scavenged from the synaptic cleft, and this is the rate-limiting process in acetylcholine synthesis. The import of choline from the extracellular fluid at the nerve terminal is facilitated by a sodium-dependent carrier. In contrast, acetyl-CoA is available in abundance since it is not only transported to the synapse but also synthesized in mitochondria which are present at the synapse in their thousands. Acetylcholine is stored and transported in synaptic vesicles which may each hold up to 50 000 molecules.

The release of acetylcholine is triggered by an influx of calcium ions, brought about by action potentials stimulating voltage-gated calcium channels in the synaptic membrane. The calcium ions in turn bind to a vesicle-associated protein called synaptotagmin, which promotes the association of a second vesicle protein called synaptobrevin with proteins lodged in the synaptic plasma membrane. This culminates in fusion between the vesicle and plasma membranes and the release of acetylcholine by exocytosis.

Once released, acetylcholine is free to interact with receptors on the postsynaptic cell surface (see the section titled 'Acetylcholine receptors'). However, it is rapidly broken down into choline and acetate by the enzyme acetylcholinesterase (Figure 3), which is present in large amounts in the synaptic cleft. The activity of acetylcholine is therefore short lived. Over 50% of the choline formed in this reaction is taken up by endocytosis at the nerve terminal and reused for acetylcholine synthesis. The empty synaptic vesicle is also recaptured by endocytosis and its components reused.

Acetylcholine Receptors

There are two main classes of acetylcholine receptor, known as nicotinic and muscarinic receptors. Nicotinic receptors are ionotropic, that is, they are ligand-gated ion channels, which are permeable to sodium, potassium, and chloride ions. They are fast acting (capable of responding in microseconds) and are activated by nicotine as well as acetylcholine, hence their name. They are selectively blocked by curare. Nicotinic receptors are made up of five subunits arranged symmetrically around the central pore. In mammals, 17 different types of subunit have been identified thus far, some expressed specifically in muscles and active at the neuromuscular junction, and some expressed specifically

in neurons. The neuronal receptors are very heterogeneous, with a wide range of possible subunit combinations and a complex distribution throughout the central and peripheral nervous systems.

The muscarinic receptors are G-protein-coupled receptors. They are metabotropic and are slower acting, responding in milliseconds but with longer-lasting effects. They work by stimulating the release of second messengers (either cyclic AMP or inositol triphosphate (IP_3), depending on the receptor class) which have downstream effects on other ion channels. The muscarinic receptors are capable of plasticity and play a significant role in learning and memory (see the section titled 'Role and distribution in the central nervous system'). They are so named because they bind selectively to the toxin muscarine, which is found in poisonous mushrooms. In mammals there are five receptors encoded by five genes, and four of these have also been identified through the use of selective radioactively labeled agonists and antagonists. They have distinct and complex distribution patterns throughout the central and peripheral nervous systems and on target muscles and glands.

Role and Distribution in the Peripheral Nervous System

Acetylcholine is the principal neurotransmitter at the neuromuscular junctions of skeletal, cardiac, and smooth muscle fibers. In the autonomic nervous system, all preganglionic and postganglionic parasympathetic neurons are cholinergic, as are all preganglionic sympathetic neurons. Most postganglionic sympathetic neurons are not cholinergic, but secretory fibers to sweat glands and vasodilator fibers to the blood vessels of skeletal muscles are exceptional in this regard. Acetylcholine is also used in preganglionic sympathetic fibers to the suprarenal medulla, a modified sympathetic ganglion. Although the distribution of receptors is complex, the preganglionic fibers of sympathetic and parasympathetic nerves are nicotinic, while the short postganglionic fibers of the parasympathetic nervous system are muscarinic. The neuromuscular junctions of skeletal muscles are nicotinic to facilitate the rapid response to voluntary impulses.

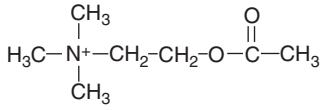

Figure 1 The molecular structure of acetylcholine.

Role and Distribution in the Central Nervous System

There are relatively few cholinergic neurons in the brain, and the most prominent sources of acetylcholine are found in the basal forebrain, particularly in the striatum and in basal nuclei such as the septal nuclei, the nucleus basalis, and the interpeduncular nucleus (Figure 4). All striatal interneurons in the primate brain are cholinergic, although this represents only about 1% of the neurons in total. These are distributed throughout the caudate nucleus, putamen, and nucleus accumbens. Despite their relative scarcity, cholinergic neurons innervate most parts of the brain including the basolateral amygdala, cortex, hippocampus, and thalamus.

Nicotinic receptors are widely distributed in the brain, but they are present in relatively low numbers. Even so, a large proportion of the world's population habitually stimulates these receptors with nicotine through tobacco use (and to a lesser extent through the chewing of betal nuts, which contain another nicotinic receptor agonist, arecoline). The euphoric addiction resulting from this practice is well documented, but the natural synaptic functions of these receptors are unclear.

Most cholinergic receptors in the brain are muscarinic, and the second messenger systems thus deployed provide scope for neuromodulatory activity and the development of synaptic plasticity. Accordingly, studies in animals and humans have suggested that the cholinergic neurons in the brain are important in cognition, learning, and memory (see the section titled 'Acetylcholine as a neuromodulator in learning and memory').

Acetylcholine as a Neuromodulator in Learning and Memory

The role of acetylcholine in cognition, learning, and memory has been established by experiments in humans and animals involving drugs that either stimulate or inhibit muscarinic receptors, and animal experiments involving lesions of basal nuclei. The central

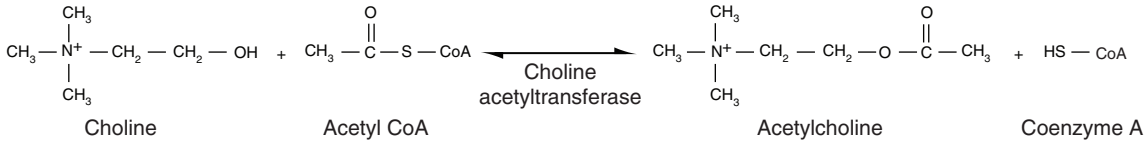

Figure 2 Acetylcholine synthesis.

administration of drugs such as physostigmine, a reversible inhibitor of acetylcholinesterase, significantly improves the performance of humans and animals in learning and cognition tasks. In contrast, the central administration of muscarinic antagonists such as atropine and scopalomine leads to memory loss and impairment of performance in learning tasks. Similar effects can be brought about by surgical lesions of the nucleus basalis, but these effects can be partially ameliorated by treatment with physostigmine. Taken together, these data suggest that acetylcholine plays an important but not exclusive role in learning and memory through the activation of muscarinic receptors, probably by providing neuromodulatory input to cortical and hippocampal neurons.

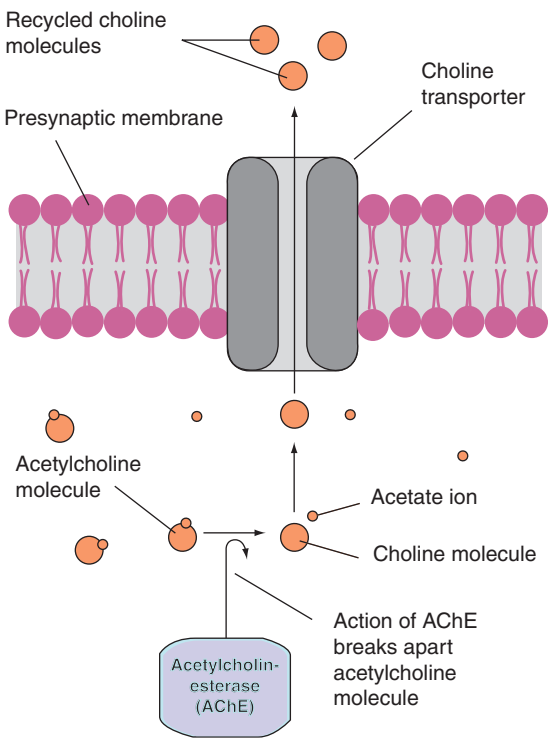

Figure 3 Degradation of acetylcholine and reuptake of choline at the nerve terminal.

The mechanism of cortical neuromodulation by acetylcholine has been studied in brain slices and *in vivo* in the cat motor cortex. Acetylcholine enhances the activity of many cortical neurons by suppressing membrane potassium currents and hence causing depolarization and suppression of adaptation. Acetylcholine has been shown to enhance the amplitude of synaptic potentials following long-term potentiation in the neocortex and piriform cortex. It is thought that direct enhancement of currents through *N*-methyl-D-aspartate (NMDA) receptors is augmented by the indirect suppression of adaptation. Acetylcholine also causes slow depolarization by blocking tonically active potassium currents, thus increasing neuronal excitability. However, the increased spiking activity in inhibitory interneurons leads to weaker synaptic transmission from those cells. This paradoxical combination can be explained by selectivity. The suppression of synaptic transmission occurs at intrinsic synapses within the cortex, but not at the afferent synapses connecting external neurons to the cortex.

The neuromodulatory effects of acetylcholine in the hippocampus involve muscarinic receptors acting upon potassium, calcium, and mixed cation channels, decreasing the slow calcium-activated potassium

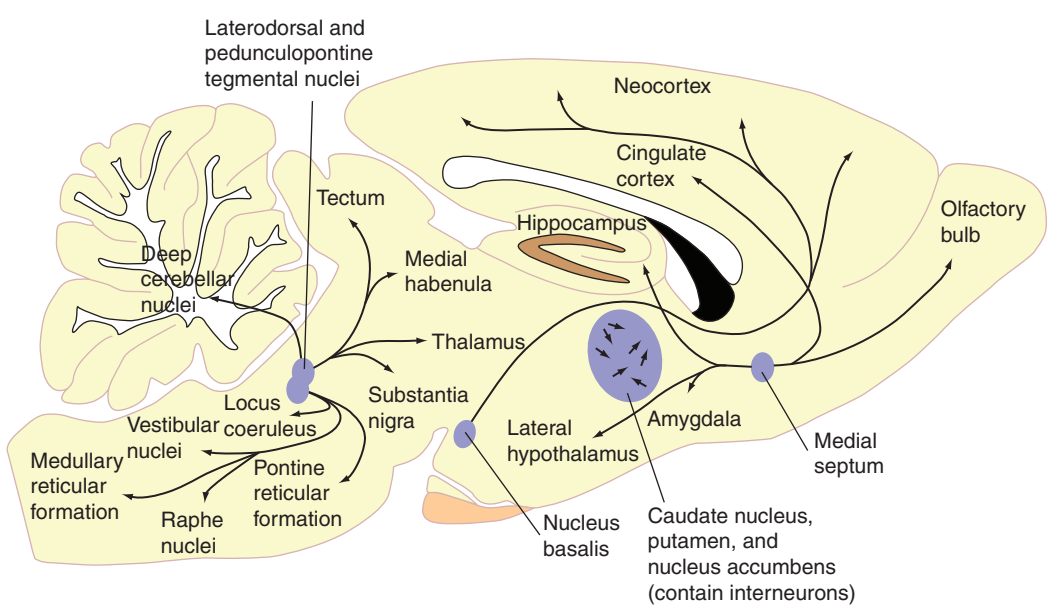

Figure 4 Midsagittal section of rat brain showing locations of important groups of cholinergic neurons and their projections.

current, the fast transient repolarizing potassium current IA, and the resting potassium current. In addition to these ion-selective currents, the hyperpolarization-activated mixed cationic current Ih is potentiated by muscarinic agonists. In pyramidal cells, acetylcholine causes a slow depolarization due to the suppression of the previously mentioned tonically active potassium current, suppresses spike frequency adaptation in pyramidal cells due to decreases in calcium-dependent potassium currents, and causes rapid excitation of hippocampal interneurons by modulating potassium conductance. Acetylcholine can also cause the presynaptic suppression of excitatory synaptic transmission, suggesting a role for the molecule in switching between learning and recall in hippocampal and olfactory networks.

Agonists, Antagonists, and Role in Disease

Given the extensive and diverse functions of acetylcholine in the central and peripheral nervous systems, it is not surprising that chemicals with the ability to enhance or inhibit the effects of this molecule are very useful in medicine and as toxins. Notably, several diseases are known to result either partially or fully from defects in acetylcholine production or signaling. Acetylcholine agonists include agents that mimic the structure of acetylcholine and activate either nicotinic or muscarinic receptors or both, and agents that act indirectly by inhibiting acetylcholinesterase and therefore increasing the availability of acetylcholine. Acetylcholine antagonists can be divided into those which selectively block nicotinic and muscarinic receptors and those which block both.

Reversible acetylcholinesterase inhibitors include galantamine, neostigmine, physostigmine, rivastigmine, and tacrine. They are clinically useful for the treatment of Alzheimer's disease and myasthenia gravis, and they are also used to reverse the effect of muscle relaxants. Given the role of acetylcholine in learning and memory, it is interesting to note that the natural decline of memory and learning ability that accompanies aging correlates with decreased synthesis of acetylcholine in the brain and the loss of cholinergic neurons. In Alzheimer's disease, patients may show a striking reduction in the number of cholinergic neurons in the nucleus basalis, and the effects can be partially reversed using anticholinesterase drugs such as tacrine. However, many other neurons are affected in Alzheimer's disease, and some patients show normal cholinergic cell numbers. There is therefore great variation in the response to anticholinesterase therapy. Myasthenia gravis is an autoimmune disease characterized by profound muscle fatigue, resulting from the production of antibodies against nicotinic receptors, eventually resulting in irreversible degeneration of the neuromuscular junction. Anticholinesterases are useful because they prolong the half-life of acetylcholine, allowing more time for it to interact with the limited number of receptors.

As well as the preceding reversible inhibitors, which work by competing with endogenous acetylcholine for the active site of the enzyme, there are also inhibitors which bind irreversibly to the enzyme and permanently block its activity. These have no medical uses but are often used as toxins. Some are used as pesticides (e.g., the organophosphate Parathion) and others as potent toxins (e.g., Sarin and VX nerve gas are based on this principle).

Centrally acting anticholinesterases worsen the tremors seen in Parkinson's disease because they affect cholinergic interneurons in the striatum. These effects can be reversed by muscarinic antagonists such as atropine, which is also the preferred treatment for anticholinesterase poisoning. Atropine has been used historically to dilate the pupils and is still used for that purpose today, although its effects can last for many days so alternative mydriatics such as tropicamide and phenylephrine are preferred. The major medical use of atropine is resuscitation by blocking the vagus nerve, thus preventing the heart rate from slowing. Nicotinic antagonists are used as muscle relaxants during therapy, and in higher doses can be used to paralyze. The plant-derived toxin curare works by blocking nicotinic receptors, as do a number of snake venoms.

See also: Acetylcholine Neurotransmission in CNS; Acetylcholinesterase Inhibitors and Alzheimer's Disease; Acetylcholinesterase; Basal Ganglia: Acetylcholine Interactions and Behavior; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Muscarinic Receptors: Autonomic Neurons; Neuromodulation; Nicotinic Receptors: Autonomic Neurons; Nicotinic Acetylcholine Receptors; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Cascio M (2004) Structure and function of the glycine receptor and related nicotinicoid receptors. *Journal of Biological Chemistry* 279: 19383–19386.

Eglen RM, Choppin A, and Watson N (2001) Therapeutic opportunities from muscarinic receptor research. *Trends in Pharmacological Science* 22: 409–414.

Giniatullin R, Nistri A, and Yakel JL (2005) Desensitization of nicotinic ACh receptors: Shaping cholinergic signaling. *Trends in Neurosciences* 28: 371–378.

Hasselmo ME (1995) Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behavioral Brain Research* 67: 1–27.

Itier V and Bertrand D (2001) Neuronal nicotinic receptors: From protein structure to function. *FEBS Letters* 504: 118–125.

Kimura F (2000) Cholinergic modulation of cortical function: A hypothetical role in shifting the dynamics in cortical network. *Neuroscience Research* 38: 19–26.

Loewi O and Dale H (2000) The discovery of neurotransmitters. In: Finger S (ed.) *Minds Behind the Brain: A History of the Pioneers and Their Discoveries*, pp. 259–279. New York: Oxford University Press.

Simon MI, Strathmann M, and Gautam N (1991) Diversity of G-proteins in signal transduction. *Science* 252: 802–808.

Watling KJ (ed.) (2006) *The Sigma-RBI Handbook of Receptor Classification and Signal Transduction,* 5th edn. St. Louis, MD: Sigma-RBI.

Yu AJ and Dayan P (2005) Uncertainty, neuromodulation, and attention. *Neuron* 46: 681–692.

Nicotine

D J K Balfour and C S Birch, University of Dundee Medical School, Dundee, UK

© 2009 Elsevier Ltd. All rights reserved.

Nicotine is the principal pharmacologically active component of the tobacco plant *Nicotiana tabacum*, the primary source of smoking tobacco. It is now widely accepted that humans inhale tobacco smoke in order to experience the psychopharmacological properties of the nicotine present in the smoke and that a majority of habitual smokers become addicted to the drug. Nicotine has reinforcing or rewarding properties which are thought to be central to its ability to cause dependence. Additionally, it enhances arousal and sustained attention. It also increases information processing, especially of visual information, and shortens reaction times. Furthermore, nicotine might have anxiety-reducing (anxiolytic) properties and might heighten mood.

The evidence that people smoke in order to experience the pharmacological effects of nicotine is compelling. For example, if the nicotine content of tobacco is manipulated artificially, most smokers adapt their smoking behavior in order to self-administer the dose of the drug which they find most reinforcing. This may be achieved either by smoking more cigarettes or, more commonly, by inhaling the smoke more deeply in order to extract the nicotine more efficiently. Nicotine exerts its effects in the brain by stimulating nicotinic cholinergic receptors located on target neurons. If smokers are given drugs which block these receptors, they tend to adjust their smoking behavior to overcome this blockade. Conversely, the prior administration of nicotine diminishes the desire to smoke and attenuates the withdrawal symptoms experienced by many smokers when they first try to quit the habit. This provides the pharmacological basis for using nicotine replacement therapy as an aid to smoking cessation. It is assumed that the addiction to tobacco depends principally upon the rewarding or reinforcing properties of nicotine, although it seems likely that other factors such as the irritant nature of the smoke might also contribute to the development of addiction. Although studies with human subjects have provided clear evidence that nicotine is the principal addictive component of tobacco smoke, research into the mechanisms underlying nicotine addiction is more easily undertaken using *in vitro* or *in vivo* techniques in experimental animals.

Evidence that Nicotine Has the Properties of a Drug of Addiction

In animal models, nicotine exhibits the principal behavioral characteristics of a drug of dependence. It has reinforcing or rewarding properties which can be detected using, for example, the place preference paradigm or the intravenous self-administration (IVSA) paradigm. The precipitous withdrawal of the drug following a period of chronic treatment also elicits behavioral effects which are thought to reflect a nicotine withdrawal syndrome.

The place preference paradigm utilizes an apparatus which has two compartments that can be clearly distinguished by their appearance and texture. On alternate training days, animals are either injected with nicotine and restricted to one of the compartments or injected with saline and restricted to the other compartment. During the test session, the animals are not given an injection, but are allowed free access to both compartments. The rewarding or reinforcing properties of the drug are demonstrated by the fact that the animals tend to spend more time in the compartment which had previously been paired with the drug.

In the technically more difficult IVSA paradigm, animals are trained to perform a task such as lever-pressing or nose-poking in order to receive small, intravenous injections of a drug. Importantly, such intravenous doses reach the brain rapidly, and, as a result, the animal readily associates performance of the task with delivery of the drug. Similarly, the inhalation of nicotine in tobacco smoke also results in a rapid transfer of nicotine to brain. Thus, although the experimental paradigm does not model completely the inhalation of nicotine in tobacco smoke, it provides a good model for the pharmacokinetic and pharmacodynamic factors which are likely to be important to the development of dependence. Nicotine IVSA is, therefore, now widely accepted as a reliable measure of the reinforcing properties of nicotine which are thought to underpin the addiction to tobacco. However, it has often proved difficult to train animals to self-administer nicotine intravenously, and, indeed, when compared with the IVSA of other psychostimulant drugs of dependence, such as amphetamine or cocaine, the reinforcing properties of nicotine as assessed using this paradigm seem to be relatively weak. Thus, the reinforcing properties of nicotine alone do not seem to explain the powerful addictive properties of the drug when inhaled in

tobacco smoke. The self-administration of addictive drugs is, however, often enhanced significantly if their administration is associated with the presentation of another stimulus which, alone, has little or no efficacy as a reinforcer. When a stimulus of this type (generally a light and/or a tone) is used, it can acquire the properties of a conditioned stimulus (CS) by being closely associated with delivery of the primary rewarding stimulus, in this case an addictive drug. This is the case with nicotine IVSA; when a CS is incorporated into the protocol, the behavior is acquired more readily and reproducibly. As a result, a substantial majority of the studies which have sought to investigate the reinforcing properties of nicotine using an IVSA protocol have employed a CS. Once the relationship between a CS and nicotine has been established, responding can also be maintained for a significant number of days by presentation of the CS alone, indicating that the CS has acquired the ability to act as a reinforcer in its own right. In human smokers, the tobacco smoke vehicle in which nicotine is usually delivered is a rich source of stimuli which could serve as conditioned stimuli. It has been suggested, therefore, that the powerful addiction to tobacco reflects the development of a close association between the pharmacological effects of nicotine in the brain and the sensory stimuli present in tobacco smoke which irritate the mouth, throat, and bronchi of the smoker. This hypothesis gains powerful support from studies reported by Rose and colleagues, which show that smokers find denicotinized cigarettes, which retain a majority of the sensory properties of the smoke inhaled from nicotine-containing cigarettes, satisfy their craving to smoke. Moreover, these sensory stimuli seem to be particularly important to the most heavily addicted smokers.

The other primary characteristic of a drug of addiction shown by nicotine is a withdrawal effect when it is precipitously withdrawn or its effects are antagonized by the administration of a nicotinic receptor antagonist. These effects are best shown in animals which have been given nicotine by constant infusion from a minipump for a number of days, and include teeth chattering, chewing, gasping, writhing, head shakes, body shakes, and tremors. Such withdrawal effects are attenuated or abolished by the administration of nicotine and by some, but not all, of the drugs which are used to aid smoking cessation in humans, and are therefore thought to model the abstinence syndrome experienced by many smokers when they first quit smoking. Some of the behavioral changes observed when nicotine is precipitously withdrawn are thought to be mediated by changes within the brain, whereas others may reflect physiological and biochemical adaptations within the peripheral nervous systems. One behavioral change which clearly depends upon neuroadaptation within the brain has been demonstrated using intracranial self-stimulation (ICSS), which can be used to measure the effects of nicotine withdrawal on reward thresholds. In this paradigm, animals are trained to perform a task to stimulate electrodes, commonly implanted into the lateral hypothalamus, one of the 'reward' pathways of the brain. The current required to elicit this behavioral response is seen as a physiological measure of the reward threshold. Nicotine injections reduce the brain reward threshold, suggesting that the intensity of reward required to evoke a reward-seeking behavioral response is decreased. By contrast, the abrupt withdrawal of nicotine, following a period of chronic treatment, increases the brain reward threshold, suggesting that the intensity of reward required to evoke a reward-seeking behavioral response is increased. This increase in the brain reward threshold is thought to be a behavioral correlate of the anhedonia experienced by many newly abstinent smokers. The increase in reward threshold, evoked by nicotine withdrawal, is attenuated by the administration of nicotine.

Neurobiological Mechanisms Underlying the Reinforcing Properties of Nicotine

Nicotine exerts its effects in the brain by acting on neuronal nicotinic acetylcholine receptors which, physiologically, are stimulated by acetylcholine (ACh) released from adjacent neurons. These receptors are pentameric structures commonly composed of two alpha subunits and three beta subunits, although some receptors composed entirely of alpha subunits (e.g., those containing $\alpha 7$ subunits) are also thought to exist. There is evidence that a number of different subtypes of the receptor, composed of different receptor subunits, are expressed on neurons within the brain. The effects of nicotine on each subtype of the receptor reflect both the affinity of the receptor for nicotine and the physiological effects that the drug exerts at that receptor. Transgenic methods, employing mice in which the expression of different nicotinic receptor subunits has been knocked out or manipulated in some other way, have been used increasingly to investigate the subunit composition of the receptors which mediate the effects of nicotine in the brain. Studies with transgenic mice have shown that deletion of the $\beta 2$ subunit of the nicotinic receptor results in the generation of a mouse strain in which the expression of $\alpha 4\beta 2$ nicotinic receptors is abolished

and nicotine no longer has reinforcing properties. These results have been used as circumstantial evidence that the reinforcing properties of nicotine depend upon stimulation of this subtype of the nicotinic receptor.

The administration of nicotine to experimental animals results in complex changes in the release of various neurotransmitters in many areas of the brain. Many of these changes potentially could contribute to the mechanisms underlying nicotine dependence. However, there is a considerable body of evidence to support the conclusion that the mesolimbic dopamine (DA)-secreting neurons, which innervate the nucleus accumbens from the ventral tegmental area (VTA), play a pivotal role in nicotine-seeking behavior. For example, selective 6-hydroxydopamine lesions of these neurons significantly and substantially attenuate both responding for nicotine in rats trained to lever-press for the drug and the locomotor stimulant response (the increase in spontaneous locomotion) to nicotine. Studies with human subjects have also shown that blocking DA receptors within the brain elicits changes in smoking behavior, which reinforces the idea that brain DA pathways play an important part in the reinforcing properties of tobacco smoke. These observations imply that systemic injections of nicotine stimulate DA release in the nucleus accumbens. This has been confirmed in a number of laboratories, principally using *in vivo* microdialysis which permits the measurement of DA overflow in discrete areas of the brain of conscious freely moving rats. More recently, positron emission tomography (PET) studies have confirmed that the inhalation of tobacco smoke also stimulates DA overflow in the human brain. A majority of the studies which have sought to explore the role of neuronal nicotinic receptors in the mesolimbic DA response to nicotine *in vivo* have found that the DA responses to nicotine are mediated, to a significant extent at least, by nicotinic receptors in the VTA. The results of these studies imply that nicotine stimulates DA overflow in the nucleus accumbens by influencing impulse flow to the terminal field. Furthermore, behavioral responses such as locomotor stimulation and nicotine-seeking behavior, both of which are dependent upon increased DA release from mesolimbic neurons, are also preferentially attenuated by the administration of nicotinic receptor antagonists into the VTA. More recent microdialysis studies employing transgenic strains of mice in which the $\beta 2$ subunit of the nicotinic receptor has been deleted have shown that the effects of nicotine on DA release in the nucleus accumbens depends upon the expression of these $\beta 2$ subunits. Since nicotine also has no reinforcing properties in these mice, the studies provide further evidence that mesolimbic DA neurons play a central role in mediating the reinforcing properties of nicotine. The nucleus accumbens is composed of two principal anatomically and morphologically distinct subdivisions, the shell and the core, which are thought to subserve different functions. These two subdivisions have been shown to respond differentially to several drugs including nicotine: acute administration of nicotine to drug-naive animals selectively stimulates DA overflow in the nucleus accumbens shell whereas administration of a nicotine challenge to animals that have been pretreated with the drug for some days also results in increased DA overflow in the nucleus accumbens core. The accumbal core sends major projections to areas of the brain related to motor function. Initially, therefore, it seemed reasonable to conclude that enhanced DA release in the accumbal core explained the increased locomotor response, observed in rats challenged with nicotine following a period of chronic treatment. However, further research using a variety of experimental approaches has demonstrated a clear dissociation between the enhanced DA response in the accumbal core and the expression of the increased locomotor response to the drug observed in nicotine-pretreated rats. For example, drugs which antagonize *N*-methyl-D-aspartate (NMDA) glutamatergic receptors in the brain attenuate the effects of nicotine on DA overflow in the accumbal core but have no effect on the enhanced locomotor activity seen in nicotine-pretreated rats. Selective sensitization of the DA projection to the accumbal core is a property which nicotine shares with other drugs of dependence such as amphetamine, cocaine, and morphine, and it now seems more likely that the effect is implicated in the neurobiology underlying dependence. Other evidence suggests that the sensitized locomotor response to nicotine, observed in pretreated animals, is associated with sensitization of postsynaptic DA receptors in the shell subdivision of the accumbens.

More recently, it has been suggested that the increased DA overflow in the shell and core subdivisions of the accumbens may play complementary roles in the development of drug addiction, with increased DA overflow in the accumbal shell being thought to facilitate or promote responding for the drug. Di Chiara argues that increased DA overflow in this subdivision of the nucleus accumbens enhances incentive or habit learning of the behaviors that deliver addictive drugs. The presentation of natural reinforcers, such as food or water, has been shown to increase DA overflow in the shell subdivision of the nucleus accumbens. However, the effects of addictive

drugs, including nicotine, on DA overflow in this area of the brain are substantially larger and more prolonged than those normally evoked by natural rewards. Thus, Di Chiara hypothesized that the unphysiological magnitude and duration of the effect results in behaviors associated with the acquisition of drug-seeking behavior which come to dominate the behavioral repertoire of the individual in the manner that is characteristic of dependence. Robinson and Berridge have long argued that the sensitized DA responses, observed in animals pretreated with psychostimulant drugs of dependence, play an important role in mediating the way in which conditioned stimuli influence drug-seeking behavior. More recently, Balfour has summarized evidence which suggests that the sensitized responses to nicotine, observed in the accumbal core of nicotine-pretreated rats, serves this specific function. He hypothesized that the increase in extracellular DA, which can be measured in the accumbal core and shell in response to a nicotine challenge, reflects increased burst firing on the neurons through a process which is dependent upon the co-stimulation of NMDA glutamatergic receptors in the VTA. This causes the substantial and sustained increases in DA overflow which can be measured using *in vivo* microdialysis. Balfour further suggested that the DA released as a consequence of burst firing of these neurons functions as a local hormone which serves to influence the intensity of events mediated by the accumbal shell and the core (Figure 1).

It is proposed that increased DA overflow in the shell increases the probability that an individual will repeat a behavioral response, thereby facilitating acquisition of behaviors that generate rewards. Increased DA overflow in the core is hypothesized to increase the probability that a CS will elicit reward-seeking behavior. The mechanisms which mediated the sustained effects of nicotine on these neurons remain a matter for debate, although Mansvelder and colleagues have suggested that the effect is maintained by a sequence of events within the VTA in which $\alpha 4\beta 2$ subunit-containing nicotinic receptors located on both the DA neurons and γ-aminobutryic acid (GABA) terminals are initially stimulated and then desensitized by exposure to nicotine. The increased activity of the neurons is maintained, however, by the persistent stimulation of $\alpha 7$ subunit-containing nicotinic receptors located on glutamate terminals which facilitate the release of glutamate within the VTA. This, the authors argue, maintains increased activity of DAergic neurons and, thus, increased DA overflow in the nucleus accumbens.

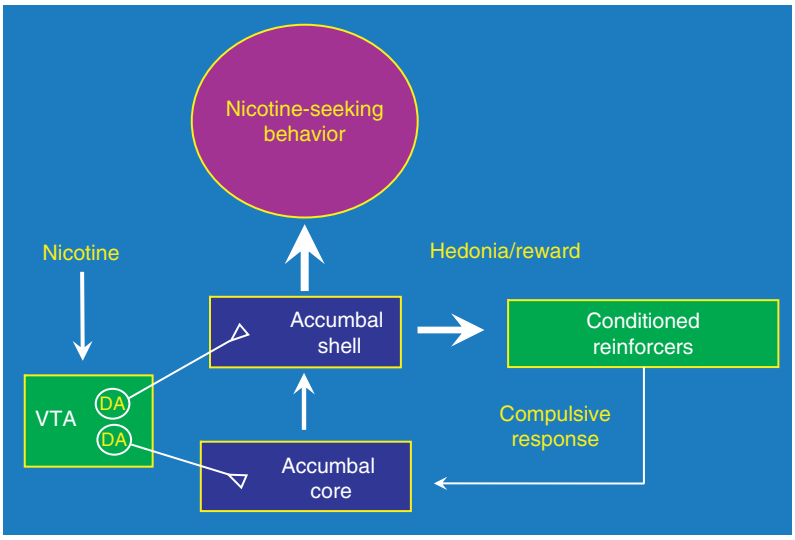

Figure 1 This figure illustrates the DA pathways that are hypothesized to contribute to the development of addiction to nicotine and tobacco. In nicotine-pretreated individuals, injections of nicotine increase DA overflow in the both the shell and core subterritories of the nucleus accumbens. The increased DA overflow in the accumbal shell is hypothesized to increase the 'pleasure' or 'hedonic impact' derived from behaviors associated with the delivery of the drug. As a result, the probability that the individual will repeat the behavior is greatly enhanced. Furthermore, stimuli that are paired with drug delivery acquire pleasurable reinforcing properties. Once the association between the delivery of nicotine and the presentation of the conditioned stimuli has been established, presentation of stimuli alone can reinforce nicotine-seeking behaviors. The hypothesis posits that the increase in DA overflow in the accumbal core elicited by a nicotine injection greatly increases the probability that presentation of the CS will elicit nicotine-seeking behavior; thus, nicotine evokes the compulsive drug-seeking behavior which is characteristic of addiction. Reproduced from Balfour DJK (2006) Complementary roles for the accumbal shell and core in nicotine dependence. In: Bock G and Goode JA (eds.) *Understanding Nicotine and Tobacco Addiction,* Novartis Foundation Symposium 275, pp. 96–115. Chichester and New York: Wiley, with permission from Novartis Foundation.

The Neurobiology of Nicotine Withdrawal

The other important behavioral response evoked by sustained exposure to nicotine is an abstinence syndrome which is evoked by the abrupt withdrawal of the drug. The mechanisms which mediate this syndrome remain to be established, although Malin and colleagues have suggested that they might be related to the stimulation of brain opioid receptors since administration of the opiate receptor antagonist naloxone will also evoke the syndrome. Recent studies with transgenic mice have shown that expression of these nicotine-withdrawal effects is not altered by deletion of neuronal nicotinic receptors which contain $\beta 2$ subunits. Thus, since the reinforcing properties of the drug depend upon stimulation of these receptors (see above), it is likely that different populations of neuronal nicotinic receptor mediate the two principal components of nicotine dependence, reinforcement and withdrawal. It is important to remember, however, that the withdrawal effects described by Malin and colleagues are likely to reflect changes in both the brain and the peripheral nervous system. The nature of the neuronal nicotinic receptors which mediate this effect remain to be established.

Other studies have shown that nicotine can inhibit the release of serotonin (5-hydroxytryptamine; 5-HT) in the rat hippocampus and that chronic exposure to nicotine is associated with a regionally selective decrease in the concentration and formation of 5-HT in this area of the brain. Changes in 5-HT release in this area of the brain have been implicated in the psychopathology and treatment of anxiety and depression. Smoking is significantly more prevalent in people with an underlying depressive disorder, and these smokers tend to be more dependent upon nicotine than smokers in the general population. It seems reasonable to propose that these effects may be implicated in the neurobiology underlying nicotine dependence, because other studies have shown that chronic exposure to the nicotine present in tobacco smoke also elicits regionally selective changes in the concentrations of 5-HT and its principal metabolite, 5-hydroxyindoleacetic acid (5-HIAA), and the density of 5-HT_{1A} receptors in human brain. These data imply that inhaling nicotine in tobacco smoke exerts the same effects on the 5-HT pathways which innervate human hippocampus as those observed in response to nicotine in the rat brain. A more recent study has shown that the increase in brain reward threshold, evoked by nicotine withdrawal in rats, can be attenuated by the co-administration of the 5-HT-selective reuptake inhibitor fluoxetine (more widely known as Prozac), with a 5-HT_{1A} receptor antagonist. This observation implies that the neuroadaptive changes in serotonergic function, evoked by chronic nicotine, may also mediate components of the nicotine abstinence syndrome. This conclusion is consistent with report of Seth and colleagues, which also concluded that elements of the nicotine abstinence syndrome are mediated by stimulation of 5-HT_{1A} receptors.

Effects of Nicotine on the Responses to Stressful and Anxiogenic Stimuli

There are consistent reports that tobacco smoking is more prevalent among individuals who suffer from stress-related psychiatric disorders, especially depression, and that these individuals often smoke more heavily and are more dependent upon tobacco than the general population. This has led Markou and colleagues to suggest that smokers with an underlying depressive disorder might use nicotine as a form of self-medication for their condition. In studies with experimental animals, nicotine has been shown to have properties similar to those of an anxiolytic drug in a number of behavioral models including the elevated plus-maze test and the social interaction test. These results are inconsistent, however, being influenced to a significant extent by the experimental protocol employed, the dose of nicotine used, the duration of treatment, and the route of administration. Other studies have shown that nicotine can exhibit the properties of an antidepressant drug in paradigms employing the forced swim test or learned helplessness. Again, however, the results of these studies are not always consistent and seem to depend upon the protocol used.

The neurobiological mechanisms underlying the putative anxiolytic and antidepressant properties of nicotine remain to be established. One hypothesis which may explain its antidepressant properties posits that the stimulation of DA overflow in the nucleus accumbens, evoked by nicotine administration, attenuates the anhedonia experienced by people who suffer from the condition. Certainly, as mentioned above, there is evidence that nicotine injections reduce brain reward thresholds, an effect which is thought to model the 'hedonic' response to nicotine inhaled in tobacco smoke. Thus, it seems reasonable to conclude that this effect could contribute to the antidepressant properties of the drug. However, it seems unlikely that this provides a complete explanation for the antidepressant-like effects of nicotine because, in some studies, the response was observed in animals given the drug by constant infusion, a protocol that is unlikely to maintain raised extracellular DA in the nucleus accumbens. Nicotine has also been shown to have anxiogenic properties, and

both the anxiogenic and anxiolytic properties have been related most clearly to its effects on 5-HT release in the hippocampus. The anxiogenic properties seem to be mediated by stimulation of 5-HT_{1A} receptors in the dorsal hippocampus and lateral septum while the anxiolytic properties have been related to stimulation of somatodendritic 5-HT_{1A} receptors in the dorsal raphe nucleus which, presumably, serves to inhibit 5-HT release predominantly in the ventral hippocampus.

The Effects of Nicotine on Attention and Cognition

It is widely accepted that cholinergic pathways play an important role in cognition and attention and that important components of these effects are mediate by neuronal nicotinic receptors. In humans, the administration of nicotine is reported to enhance attention and cognitive performance, and it has been suggested that this property of the drug may contribute to the 'beneficial' effects sought by the smoker. Behavioral studies with experimental animals have also shown that the administration of nicotine and other nicotinic receptor agonists can improve performance in a range of cognitive tasks such as the working memory component of radial arm maze tasks, passive avoidance, and water-maze tasks. The nature of the nicotinic receptors which mediate these responses remains to be established with certainty, although there is evidence that $\alpha4\beta2$, $\alpha3\beta2$, and $\alpha7$ receptors located in the hippocampus play an important role in the effects of nicotine on working memory.

The effects of nicotine on spatial learning in a water-maze are most apparent in animals in which learning has been impaired by septal lesions. The administration of neuronal nicotinic receptor antagonists impairs performance in many of these tasks. The neurobiological and neuroanatomical substrates for these effects of nicotine have not been established with certainty, although it seems likely they are mediated, in part, by the increase in ACh release, evoked in the brain, by nicotine, in areas of the brain such as the hippocampus and frontal cortex. This ACh then acts upon muscarinic receptors. More recent studies suggest that nicotine, through stimulation of nicotinic receptors, may also influence the expression and the processing of amyloid precursor protein (APP), increasing the secretion of neuroprotective soluble APP. These data, when taken together, provide support for the possibility that nicotinic agonists may have value as a treatment which slows the progression of Alzheimer's disease and alleviates some of the cognitive deficits observed during the early stages of the condition.

Another mechanism by which nicotine may bring about improvements in tests of cognition is by enhancing attention. This property of the drug seems to contribute significantly to its beneficial effects on cognition and information processing. Studies in experimental models have demonstrated clear effects of the drug on attention in the five-choice serial reaction time task. The response is blocked by the prior administration of nicotinic receptor antagonists, confirming the role of these receptors.

See also: Addiction: Neurobiological Mechanism; Anxiety Disorders; Drug Addiction: Behavioral Neurophysiology; Drugs Addiction: Actions; Nicotinic Receptors: Autonomic Neurons; Nicotinic Acetylcholine Receptors; Stress, Sex and Adolescent Nicotine Response.

Further Reading

Balfour DJK (2004) The neurobiology of tobacco dependence: A preclinical perspective on the role of the nucleus accumbens. *Nicotine and Tobacco Research* 6: 899–912.

Balfour DJK (2006) Complementary roles for the accumbal shell and core in nicotine dependence. In: Bock G and Goode JA (eds.) *Understanding Nicotine and Tobacco Addiction,* Novartis Foundation Symposium 275, pp. 96–115. Chichester and New York: Wiley.

Balfour DJK and Ridley DL (2000) The effects of nicotine on neural pathways implicated in depression: A factor in nicotine addiction? *Pharmacology Biochemistry and Behavior* 66: 79–85.

Balfour DJK, Wright AE, Benwell MEM, and Birrell CE (2000) The putative role of extra-synaptic mesolimbic dopamine in the neurobiology of nicotine dependence. *Behavioural Brain Research* 113: 73–83.

Caggiula AR, Donny EC, White AR, et al. (2001) Cue dependency of nicotine self-administration and smoking. *Pharmacology Biochemistry and Behavior* 70: 515–530.

Corrigall WA and Coen KM (1989) Nicotine maintains robust self-administration in rats on a limited access schedule. *Psychopharmacology* 99: 473–478.

Di Chiara G (2000) Behavioural pharmacology and neurobiology of nicotine reward and dependence. In: Clementi C, Fornasari D, and Gotti C (eds.) *Handbook of Experimental Pharmacology,* vol. 14, pp. 603–750. Berlin: Springer.

Di Chiara G (2002) Nucleus accumbens medial shell and core dopamine: Differential role in behavior and addiction. *Behavioural Brain Research* 137: 75–114.

Malin DH (2001) Nicotine dependence studies with a laboratory model. *Pharmacology Biochemistry and Behavior* 70: 551–559.

Mansvelder HD, DeRover M, McGehee DS, and Brussaard AB (2003) Cholinergic modulation of dopaminergic reward areas: Upstream and downstream targets for nicotine addiction. *European Journal Pharmacology* 480: 117–123.

Markou A, Kosten TR, and Koob GF (1998) Neurobiological similarities in depression and drug dependence: A self-medication hypothesis. *Neuropsychopharmacology* 18: 135–174.

Newhouse PA, Singh A, and Potter A (2004) Nicotine and nicotinic receptor involvement in neuropsychiatric disorders. *Current Topics in Medicinal Chemistry* 4: 267–282.

Picciotto MR, Brunzell DH, and Caldarone BJ (2002) Effect of nicotine and nicotinic receptors on anxiety and depression. *NeuroReport* 13: 1097–1106.

Robinson TE and Berridge KC (2003) Addiction. *Annual Review of Psychology* 54: 25–53.

Rose JE, Behm FM, Westman EC, and Johnson M (2000) Dissociating nicotine and nonnicotine components of cigarette smoking. *Pharmacology Biochemistry and Behavior* 67: 71–81.

Seth P, Cheeta S, Tucci S, and File SE (2002) Nicotinic–serotonergic interactions in brain and behaviour. *Pharmacology Biochemistry and Behavior* 71: 795–805.

Relevant Website

http://www.nida.nih.gov – NIDA-The Science of Drug Abuse & Addition See 'NDA Research Report Series on Tobacco', National Institute of Drug Abuse.

Nicotinic Acetylcholine Receptors

J-P Changeux, Institut Pasteur, Paris, France
Y Paas, Bar-Ilan University, Ramat-Gan, Israel

© 2009 Elsevier Ltd. All rights reserved.

Following Claude Bernard's investigations on the effect of curare, Langley proposed in 1905 that a 'receptive substance,' presently designated as a receptor, "receives the stimulus from the nerve and transmits it to the effector cell." On the basis of the resemblance of acetylcholine (ACh) effects to those of naturally occurring plant alkaloids, Sir Henry Dale distinguished in 1914 between two main classes of receptors for ACh, the muscarinic and the nicotinic receptors. Typically, muscarinic receptors are activated by muscarine and blocked by atropine, whereas nicotinic receptors are activated by nicotine and blocked by curare. Subclasses have also been described. Only nicotinic ACh receptors (nAChRs) are present at the motor endplate of vertebrate skeletal muscles. Muscarinic receptors are found on smooth muscles and gland cells and, together with nicotinic receptors, on autonomic ganglion cells and neurons from the central nervous system.

The nAChR Has an Intrinsic Cationic Channel

Binding of ACh to the nAChR causes the opening of a cationic channel that is physically linked to the receptor site (Figure 1). This intrinsic channel is permeable to Na^+, K^+, and, in many nAChR-subunit combinations, Ca^{2+} ions. By contrast, muscarinic ACh receptors transduce extracellular signals into the cell by interacting with a G-protein to regulate indirectly effector systems such as adenylate cyclase and ion channels.

In 1970, the nAChR became the first neurotransmitter receptor to be isolated and purified and, in 1983, it was the first to be chemically defined by molecular genetic methods. Two factors played a decisive role in this fundamental progress: (1) the presence of very high concentrations of nAChR in the electric organs of the electric fish *Electrophorus electricus* and *Torpedo*, and (2) the availability of small polypeptide α-toxins from snake venoms that bind with very high affinity and selectivity to the nAChR from electric organs.

The nAChR is an oligomeric glycoprotein of about 300 000 Da. In side views obtained by cryo-electron microscopy, it appears as a transmembrane cylinder of ~8 nm diameter and ~16 nm length; the long axis being perpendicular to the membrane plane (Figure 1(a)). In a top view, it looks like a rosette of five subunits organized around a symmetry axis (Figure 1(b)).

Combinatorial Assembly of Subunits Encoded by Homologous Genes Results in a Wide Diversity of nAChRs

In the electric organ and muscle, the nAChR oligomer is composed of four different polypeptide chains associated into a pentamer with the α_1, γ, α_1, δ, β_1 subunits organized in a counterclockwise manner (Figure 1(b)). In nerve cells, combinatorial assembly of up to nine different α subunits (α_2 to α_{10}) and three β subunits (β_2 to β_4) contributes to a wide diversity of pentamers, which may include one, two, or three different types of subunits and possess distinct pharmacological, electrophysiological, and kinetic properties. Each subunit traverses the cell membrane four times and all subunits are in contact with the lipid bilayer.

The complete amino acid sequences of the electric organ, muscle, and neuronal subunits in vertebrates and invertebrates have been deduced based on the nucleotide sequence of their cloned cDNA. The primary structure of the nAChR subunits is highly conserved throughout the vertebrate phylum from *Torpedo* to humans. For instance, the α_1 subunit of *Torpedo* shares 77% sequence identity with the human α_1 subunit. Likewise, the human α_1 subunit shares 36%, 33%, 30%, and 34% identity with the human β_1, γ, ε, and δ subunits, respectively (in mammals, an ε subunit replaces the fetal γ subunit early in postnatal life). These striking homologies suggest that the various subunits evolved by gene duplication from a common ancestral gene. Tissue-specific expression accounts for the muscle and neuronal subunit subfamilies.

Three Major Domains Arise from the Transmembrane Topology of the nAChR Subunits

Each subunit is a polypeptide of 450–700 amino acids that consists of a long extracellular N-terminal segment, four transmembrane hydrophobic segments of about 20 amino acids each (termed M1–M4), a long cytoplasmic segment, and a short extracellular C-terminal tail. The five long extracellular segments fold and assemble into a large extracellular hydrophilic ACh-binding domain (Figure 1). The transmembrane segments form a membrane-embedded domain whose elements are closely organized around

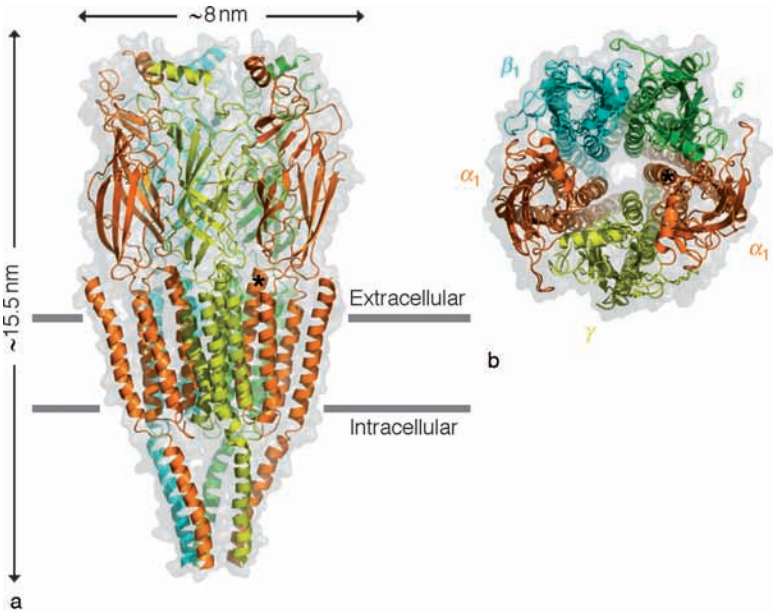

Figure 1 Structural model of the nACHR from the *Torpedo* electric ray. Ribbon representations were prepared using coordinates deposited in the research collaboratory for structural bioinformatics (RCSB) protein data bank (PDB) under ID# 2BG9. The ribbon representations are shown on the background of the gray-colored molecular surface. The two facing α_1 subunits are colored in orange while the β_1, γ, and δ subunits are shown in cyan, yellow, and green, respectively. (a) Side view, from within the membrane. The thick gray lines delineate the membrane borders. (b) Top view, from the extracellular side. The M2 segments are five α helices closely organized around the axis of ion conduction that is at the center of the molecule and is perpendicular to the viewer. In both panels, the extrcellular end of M2 of the right α_1 subunit is labeled with a black asterisk. Note that most of the cytoplasmic domain is missing due to the lack of 3-D structural information.

an axis perpendicular to the membrane plane, delineating thereby an aqueous pore. The five cytoplasmic segments fold into a hydrophilic domain that is thought to associate with components that anchor the receptor to the cytoskeleton, for example, the 43-kDa-Rapsyn protein. The cytoplasmic domain also contains structural elements that modulate channel activity. Notably, the nAChR-subunit topology is common for all other pentameric neurotransmitter receptor channels that are gated by serotonin, glycine, γ-aminobutyric acid (GABA), glutamate, or histamine. These ionic channels constitute the Cys-loop receptor superfamily whose members share a conserved disulfide bridge in their extracellular ligand-binding domain.

The binding pockets for ACh are formed by the interface between each α subunit and the adjacent subunit. The α subunit contributes, to the ACh-binding pocket, the so-called principal component, which consists of a series of loops carrying highly conserved ACh-binding residues, mainly aromatic but also a cysteine pair. The neighboring subunit contributes a series of complementary ACh-binding residues. Affinity labeling, binding studies, and electrophysiological recordings performed with wild-type (WT) and mutated nAChRs identified ACh-binding pockets at the α/γ and α/δ interfaces in the case of the electric organ and muscle nAChRs (**Figure 1(b)**), α/β interfaces in the case of neuronal heteromeric nAChRs, and α/α interfaces in the case of homomeric neuronal nAChRs.

The acetylcholine-binding protein (AChBP) is a glial pentameric water-soluble protein sharing structural homology with the extracellular neurotransmitter-binding domain of the nAChRs' α subunits. It does not have the transmembrane channel and the cytoplasmic domains (**Figure 2**). The AChBP is secreted to the synaptic cleft between neurons of the mollusk *Lymnaea stagnalis* and of *Aplysia*, where it apparently sequesters ACh, regulating thereby cholinergic transmission. The high-resolution X-ray crystal structure of the AChBP also supports the concept of intersubunit ligand-binding pockets (**Figure 2**).

The five M2 helical segments of the full-length nAChRs line the ion-channel pore (**Figure 1**), which was initially identified with noncompetitive pore blockers such as chlorpromazine or triphenylmethylphosphonium. Residues belonging to the M2 segments define the activation gate, which acts as a barrier obstructing the flow of ions when the receptor is at rest (i.e., displaying a closed pore). The location of the activation gate is still controversial. On the one hand, three-dimensional (3-D) structural models built on the basis of cryo-electron microscopy suggested that the activation gate is a 'hydrophobic girdle'

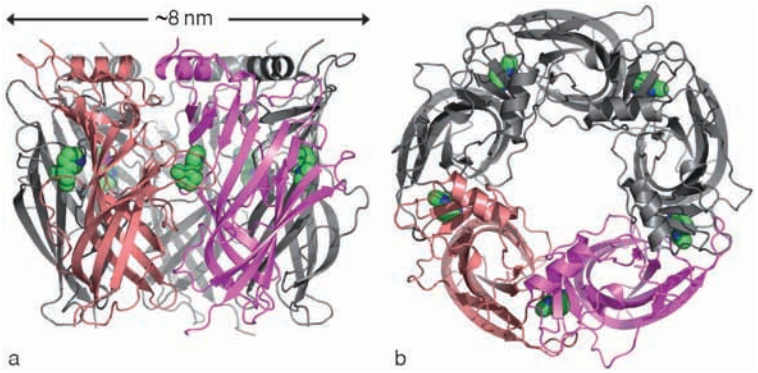

Figure 2 X-ray crystal structure of the AChBP from the mollusk *Lymnaea stagnalis*. Ribbon representations were prepared using coordinates deposited in the PDB under ID# 1UW6. Nicotine molecules are shown at the interface between each two neighboring subunits, as green and blue spheres (carbon and nitrogen atoms, respectively). Note that the five subunits share identical primary, secondary, and tertiary structures. (a) Side view; (b) top view.

located midway between the extracellular and intracellular pore's vestibules. It was further suggested that the 'hydrophobic girdle' breaks apart when the M2 segments rotate, around their own longitudinal axis, upon activation (i.e., upon channel opening). On the other hand, state-dependent accessibility of methanethiosulfonates or zinc ions to cysteines or histidines introduced along the pore revealed that the activation gate is a constriction located close to the intracellular end of the M2 segments, that is, close to the bottom of the pore. These and other structure–function relation studies also indicate that (1) the resting and active states share similar patterns of amino acid accessibility to the pore's solvent and (2) channel gating predominantly involves rigid tilting motions of the M2 segments, which widen (open) or narrow (close) the bottom-pore constriction. Amino acids located within the M1–M2 connecting segment, close to the intracellular end of the M2 segments, contribute to the cationic (versus anionic) selectivity of the channel. As these residues are also part of the bottom-pore constriction, it is concluded that components of the activation gate act as a selectivity filter upon channel opening.

The nAChR is an Allosteric Protein

Cryo-electron microscopy, affinity labeling, binding kinetics of various nicotinic ligands, and rapid kinetic measurements of ion-channel opening and closing events showed that the ACh receptor exhibits properties typical of allosteric proteins. That is, first, the ACh-binding site and the ion channel are spaced far apart (~35 Å). Second, the nAChR can undergo reversible transitions between distinct allosteric conformations even in the absence of an agonist, including rare but detectable spontaneous channel openings. The various allosteric states, which preexist in reversible equilibrium already before ligand binding, are stabilized not only by agonists or competitive antagonists that bind to the ACh-binding sites, but also by 'allosteric' effectors (e.g., Ca^{2+} ions) that bind to topographically distinct sites.

The ion channel of cloned nAChRs expressed in heterologous cell systems opens at relatively high concentration of ACh (K_D in the 50–100 μM range), which is close to the concentration of ACh found in the synaptic cleft during transmission. On the other hand, prolonged exposure of the nAChR to the neurotransmitter leads to stabilization of states that display (1) higher affinity for the agonists ($K_D \approx 1$ μM or 5 nM for ACh in case of neuronal or muscle nAChR, respectively) and (2) a closed pore conformation that does not conduct ionic currents. These states are referred to as 'desensitized,' and they are stabilized by ACh at equilibrium. Hence, the high-affinity binding states do not take part in the transduction of the physiological signal at the synapse; rather, they play a role in the regulation of the efficacy of signal transmission.

Compartmentalized Expression of Muscle nAChRs Depends on Electric Activity and Release of Neurotrophic Factors

In the adult neuromuscular junction, nAChR is localized almost exclusively in the postsynaptic (muscle) membrane of the motor endplate, with a density of 10 000 to 25 000 molecules μm^{-2} at the crest of the folded membrane. In the embryonic muscle fiber, nAChR molecules are evenly distributed extrasynaptically on the surface of the muscle cell, freely mobile and metabolically labile (half-life of about 18 h). During the formation of the endplate, nAChR molecules aggregate on postsynaptic membrane regions,

opposite to the motor-nerve ending, where they become immobile and metabolically stable (half-life of about 11 days). Around birth, in some species, an ε subunit replaces the fetal γ subunit, giving rise to an adult form with changed channel properties. Meanwhile, the extrasynaptic receptors disappear owing to repression of their synthesis, which takes place as soon as the muscle electric activity starts. In the adult, preventing stimulation of the muscle by sectioning of the motor nerve reactivates the biosynthesis of extrasynaptic nAChR, leading to the so-called phenomenon of 'denervation hypersensitivity.' Electrical stimulation of the muscle or its reinnervation by neurons restores the original postsynaptic distribution. Such compartmentalized expression of nAChR genes results, in part, from the regulation of transcription, which is active in most sarcoplasmic nuclei of developing myotubes. In the adult, this process becomes restricted to the subsynaptic 'fundamental' nuclei. In the promoter of the receptor genes, distinct elements (N Box vs. E Box) control subsynaptic transcription versus activity-dependent extrasynaptic repression. Posttranslational mechanisms include the conformational maturation of the receptor protein, its transit via a specialized Golgi apparatus (in the mature endplate), and its targeting, aggregation, metabolic stabilization, and immobilization in the postsynaptic membrane.

Factors of neural origin that are involved in compartmentalizing nAChR gene expression include acetylcholine receptor-inducing activity (ARIA), a factor homologous to human heregulin, and glial growth factor, which binds to tyrosine kinase receptors of the erbB family. Another important component, the acetylcholine receptor-aggregating factor, referred to as AGRIN, elicits formation of nAChR clusters. A cytoskeletal protein 43-kDa-Rapsyn, which is considered to be associated with the nAChR, contributes to AGRIN immobilization and stabilization in the postsynaptic membrane.

In the nematode *Caenorhabditis elegans*, a transmembrane protein, LEV-10, was found to be required for clustering of nAChR molecules in cholinergic neuromuscular junctions. Another transmembrane protein of *C. elegans*, RIC-3, is involved in enhancing the maturation (subunit folding and assembly) of nAChRs.

nAChRs Play Decisive Physiological and Pathological Roles

Activation of nAChRs on the muscle fibers initiates a cascade leading to muscle contraction. Hence, elimination of nAChR molecules from the neuromuscular junction, as occurs in myasthenia gravis, results in severe muscle weakness. The latter is an autoimmune disease induced by anti-nAChR autoantibodies and aggravated by activation of autoimmune T lymphocytes. Congenital myasthenic syndromes are human disorders resulting from a range of mutations in the muscle nAChR that impair channel activity and may lead to degeneration of the endplate.

In the brain, nicotinic receptor genes are expressed exclusively in neurons and display different patterns of expression, from highly restricted to a few nerve cells (e.g., the α_2 subunit) to widespread ones (e.g., the β_2 subunit). Inactivation of the β_2-subunit gene by homologous recombination in mice interferes with their passive avoidance learning and alters a cognitive behavior referred to as 'exploratory,' while a more automatic 'navigatory' behavior is preserved.

nAChRs are responsible for autosomal dominant nocturnal frontal lobe epilepsy and modulate pain transmission, and are also considered to be involved in autism and schizophrenia. Correlations have been reported between cigarette smoking and protection against ulcerative colitis and Parkinson's disease. Addiction to nicotine involves high-affinity nicotinic receptors associated with midbrain dopaminergic neurons, as demonstrated in laboratory mice lacking the nAChR $\beta 2$ or $\alpha 4$ subunit. These mice are also useful animal models of attention-deficit activity disorders and sudden infant death syndrome. Nicotinic receptor-binding drugs are considered as potential therapeutic agents in Alzheimer's disease, Tourette's syndrome, and anxiety disorders.

See also: Cholinergic Pathways in CNS; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Neuromuscular Junction (NMJ): Mammalian Development; Nicotine; Nicotinic Receptors: Autonomic Neurons.

Further Reading

Brejc K, van Dijk WJ, Klaassen RV, et al. (2001) Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature* 411: 269–276.

Bourne Y, Talley TT, Hansen SB, et al. (2005) Crystal structure of a Cbtx–AChBP complex reveals essential interactions between snake alpha-neurotoxins and nicotinic receptors. *EMBO Journal* 24: 1512–1522.

Celie PH, van Rossum-Fikkert SE, van Dijk WJ, et al. (2004) Nicotine and carbamylcholine binding to nicotinic acetylcholine receptors as studied in AChBP crystal structures. *Neuron* 41: 907–914.

Changeux JP and Edelstein SJ (2005) *Nicotinic Acetylcholine Receptors*. New York: Odile Jacob Publishing Corporation.

Cymes GD, Ni Y, and Grosman C (2005) Probing ion-channel pores one proton at a time. *Nature* 438: 975–980.

Engel AG, Ohno K, and Sine SM (2003) Neurological diseases: Sleuthing molecular targets for neurological diseases at the neuromuscular junction. *Nature Reviews Neuroscience* 4: 339–352.

Gally C, Eimer S, Richmond JE, et al. (2004) A transmembrane protein required for acetylcholine receptor clustering in *Caenorhabditis elegans*. *Nature* 431: 578–582.

Halevi S, McKay J, Palfreyman M, et al. (2002) The *C. elegans* ric-3 gene is required for maturation of nicotinic acetylcholine receptors. *EMBO Journal* 21: 1012–1020.

Hansen SB, Sulzenbacher G, Huxford T, et al. (2005) Structures of *Aplysia* AChBP complexes with nicotinic agonists and antagonists reveal distinctive binding interfaces and conformations. *EMBO Journal* 24: 3635–3646.

Karlin A (2002) Emerging structure of the nicotinic acetylcholine receptors. *Nature Reviews Neuroscience* 3: 102–114.

Keramidas A, Moorhouse AJ, Schofield PR, et al. (2004) Ligand-gated ion channels: Mechanisms underlying ion selectivity. *Progress in Biophysics & Molecular Biology* 86: 161–204.

Lester HA, Dibas MI, Dahan DS, et al. (2004) Cys-loop receptors: New twists and turns. *Trends in Neuroscience* 27: 329–336.

Lindstrom J (1997) Nicotinic acetylcholine receptors in health and disease. *Molecular Neurobiology* 15: 193–222.

Maskos U, Molles BE, Pons S, et al. (2005) Nicotine reinforcement and cognition restored by targeted expression of nicotinic receptors. *Nature* 436: 103–107.

Paas Y, Gibor G, Grailhe R, et al. (2005) Pore conformations and gating mechanism of a Cys-loop receptor. *Proceedings of the National Academy of Sciences of the United States of America* 102: 15877–15882.

Sine SM and Engel AG (2006) Recent advances in Cys-loop receptor. *Nature* 440: 448–455.

Sunesen M and Changeux JP (2003) Transcription in neuromuscular junction formation: Who turns on who? *Journal of Neurocytology* 32: 677–684.

Tapper AR, McKinney SL, Nashmi R, et al. (2004) Nicotine activation of alpha4 receptors: Sufficient for reward, tolerance, and sensitization. *Science* 306: 1029–1032.

Taylor P, Osaka H, Molles B, et al. (2000) Contributions of studies of the nicotinic receptor from muscle to defining structural and functional properties of ligand-gated ion channels. In: Clementi F, Fornasari D, and Gotti C (eds.) *Handbook of Experimental Pharmacology: Neuronal Nicotinic Receptors,* vol. 144. Berlin: Springer.

Unwin N (2005) Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution. *Journal of Molecular Biology* 346: 967–989.

Wilson GG and Karlin A (1998) The location of the gate in the acetylcholine receptor channel. *Neuron* 20: 1269–1281.

Nicotinic Receptors: Autonomic Neurons

M De Biasi, Baylor College of Medicine, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

The Role of Nicotinic Acetylcholine Receptors in the Autonomic Nervous System

In vertebrates, the autonomic nervous system (ANS) is responsible for maintaining homeostasis under changing physiological demands. Autonomic ganglia are the peripheral sites where preganglionic neurons arising within the brain stem and spinal cord form connections with postganglionic neurons that send their axons to visceral and vascular targets. Fast synaptic transmission, which allows communication between pre- and postganglionic neurons, mainly relies on the release of acetylcholine (ACh) from preganglionic neurons. ACh binds to neuronal nicotinic acetylcholine receptors (nAChRs) located on postganglionic neurons. Nicotinic AChRs are the main mediators of ganglionic fast synaptic transmission, and, in this respect, they function as an interface between nervous system and peripheral organs. The key role of nAChRs in the ANS was established in the 1800s by Langley, who found that autonomic ganglia are first activated and then blocked by nicotine. The necessary role of nAChRs for synaptic transmission in the peripheral nervous system (PNS) contrasts with the role of the receptors in the central nervous system (CNS), where the most commonly observed nAChR role is to act presynaptically to enhance neurotransmitter release.

Neuronal nAChRs are cation-selective, ligand-gated ion channels expressed in both the CNS and the PNS. They are members of a protein superfamily that includes neuronal and muscle nAChRs, and receptors for γ-aminobutyric acid (GABA), glycine, and serotonin $(5HT)_3$. So far, eleven genes have been identified that encode neuronal nAChR subunits: eight α subunits ($\alpha2$–$\alpha10$) and three β subunits ($\beta2$–$\beta4$). The alpha subunits contain the ligand binding sites, whereas beta subunits are structural subunits that contribute to the function of the receptor. The mammalian nervous system expresses transcripts for all subunits except $\alpha8$, which is found only in avians. In the rat there are two functional splice variants for the $\alpha7$ subunit. The $\alpha7$–2 isoform contains an 87-base-pair insert that represents the incorporation of a novel exon between exons 4 and 5 of the conventional $\alpha7$ gene product ($\alpha7$–1). The two isoforms are found in both the CNS and PNS.

Neuronal nAChRs are formed by the assembly of five subunits, each consisting of four membrane-spanning domains. The majority of functional neuronal nAChRs are composed of two α and three β subunits, with 'duplex' (α/β) or 'triplex' combinations ($\alpha_x\alpha_y\beta$ or $\alpha\beta_x\beta_y$). Heteromeric nAChR channels are composed of $\alpha2$, $\alpha3$, $\alpha4$, $\alpha5$, $\alpha6$, $\beta2$, $\beta3$, and $\beta4$ subunits. $\alpha7$ can form either homopentamers or α/β heteropentamers. $\alpha9$ and $\alpha10$ are believed to assemble as α-only heteropentamers. The $\alpha5$ subunit participates in nAChR receptors with $\alpha_x\alpha_y\beta$ combinations but cannot yield functional receptors when expressed alone or in combination with β subunits only. The $\beta3$ subunit participates in $\alpha\beta_x\beta_y$ combinations, and, similar to $\alpha5$, it cannot yield functional receptors unless coexpressed with another β subunit.

Autonomic Ganglia Integrate and Transfer Neural Signals

Autonomic ganglia, once considered mere relay stations for the transfer of information from the CNS to the periphery, contain circuits capable of integrating and modulating neural signals before they reach the neuroeffector junction. The modulation of the neural signal occurs at preganglionic, postganglionic, and extraganglionic sites by either excitatory or inhibitory mechanisms. Differences in morphology, biochemical profile, and electrophysiological properties among ganglion cells, as well as sensory inputs from the viscera, account for the modulation of ganglionic function. The expression of multiple nAChR subtypes on individual ganglionic neurons also contributes to such complexity, as these receptors are likely to have different roles in the processing and integration of neural signals. For example, the subunit composition affects the speed of activation, the ionic current, the rates of desensitization and recovery from desensitization, and the pharmacology of nAChRs. Another important factor is the anatomical location of the receptors. nAChRs can have presynaptic, axonal, somal, and postsynaptic locations. At presynaptic locations nAChRs can enhance neurotransmitter release, while at postsynaptic locations they mediate excitatory synaptic transmission. Furthermore, even if expressed in the same neuron, nAChRs are likely to mediate different cellular mechanisms.

nAChR Subtypes in Autonomic Ganglia

Virtually every autonomic neuron contains transcripts for nAChR subunits. The $\alpha3$ nAChR subunit is highly expressed in both sympathetic and parasympathetic ganglia and is likely to participate in the

majority of nAChR subtypes expressed by autonomic neurons. α3 is found in a phylogenetically conserved gene cluster that also comprises α5 and β4. All three nAChR subunit genes are coexpressed in neurons of the sympathetic and parasympathetic nervous systems where they may form α3β4 and α3α5β4 nAChRs. The α7 nAChR subunit is also expressed in autonomic ganglia. The subunit has been shown to form homomeric channels in heterologous expression systems, but the literature suggests that it might also form heteromeric channels in native ganglionic neurons. In addition, autonomic ganglia contain transcripts for the α2, α4, β2, and β3 nAChR subunits. Pharmacological analyses suggest that the β2 and β4 subunits may co-assemble into the same receptor. The lack of nAChR subtype-selective ligands has so far delayed the study of the role of α2- and β3-containing nAChRs.

Some insight into the contribution of various nAChR subunits to autonomic function has been gained by the creation of mice lacking one or combinations of nAChR subunits. Those studies revealed that synaptic transmission is abolished in ganglia when the α3 nAChR subunit is deleted via genetic engineering techniques. The absence of α3-containing nAChRs *in vivo* produces a severe multi-organ dysfunction that leads to lack of pupillary response to light, impaired bladder contractility, and cardiac rhythm abnormalities. Mice lacking the α3 nAChR subunit develop distended, urine-filled bladders due to absence of nAChR-mediated ACh release onto the bladder smooth muscle and consequent abnormal parasympathetic control of bladder voiding. Interestingly, lack or malfunction of α3-containing nAChRs in peripheral neurons has been suggested as the possible cause of a rare human disorder, megacystis, microcolon intestinal hypoperistalsis syndrome (MMIHS).

Absence of α3-containing nAChR also alters sympathetic function and causes bradycardia and the appearance of several electrocardiographic abnormalities.

While the absence of α3 is fatal, the lack of β4 does not affect survival, despite the fact that nAChR-mediated whole cell currents are significantly reduced in ganglionic neurons with a β4 null mutation. So far, the role of β4-containing nAChRs has been explored in heart and urinary bladder. Although not evident in basal conditions, the effects of a null mutation on various organ systems can be unmasked under physiological or pharmacological manipulations. Lack of β4 does not affect heart rate in unchallenged, baseline conditions but attenuates the bradycardic response observed during high-frequency stimulation of the vagus nerve. The influence of β4 on cardiac parasympathetic innervation may also contribute to the greater-than-normal heart rate increases measured when mice lacking β4 are exposed to isolation-induced stress. Absence of β4 significantly impairs the bladder's contractile responses but, unlike the absence of α3, does not cause bladder distension and urine retention. This phenomenon could be explained by partial redundancy between the β4 and the β2 subunits.

Similarly to β4, lack of the α5 nAChR subunit does not affect survival rates. Its influence on ganglionic function is subtle and depends on the location of the receptors. While the presence of the α5 subunit in somatic nAChRs does not dramatically affect the biophysical and pharmacological properties of the channels, its presence in prejunctional nAChRs modulates the magnitude of nAChR-mediated neurotransmitter release. *In vivo*, lack of α5 does not alter basal heart rates but mitigates the bradycardic response to high-frequency vagal stimulation. No other α5-associated phenotype has been described.

nAChRs in the Enteric Nervous System

The enteric nervous system (ENS) was first established by Langley as a separate division of the ANS. A complex integrated circuit composed of sensory neurons, modulatory interneurons, and motor neurons, the ENS controls motility, exocrine and endocrine secretions, and microcirculation of the gastrointestinal tract. It is also involved in the regulation of gut immune and inflammatory processes. Enteric neurons are found in two networks or plexuses of neurons, embedded in the wall of the digestive tract: the myenteric plexus, located between the circular and longitudinal gut muscles, and the submucous plexus, embedded in the submucosa. Excitatory neurotransmission in the ENS is predominantly mediated by ACh acting at nAChRs, although adenosine triphosphate (ATP) and other neurotransmitters can also mediate fast excitatory postsynaptic potentials (fEPSPs). The contribution of nAChR-mediated neurotransmission varies depending on the anatomical localization. In the submucous plexus, fEPSPs are almost entirely inhibited by nAChR antagonists, but in the myenteric plexus only about 25% of neurons studied exhibit fEPSPs that are completely blocked by nAChR antagonists.

Immunohistochemical studies indicate that mRNAs for the α3, α5, α7, β2, and β4, but not the α4 nAChR subunit, are expressed in many myenteric and submucosal neurons. α3 is expressed at high levels in enteric neurons where it likely combines with β2 and/or β4 to yield heteromeric nAChRs with pharmacological and biophysical properties similar to those of ganglionic neurons. Absence of α3 impairs gastrointestinal motility in mice and might underlie the

gut-related pathophysiology of MMIHS. Gut motility disorders are also observed in patients with autonomic neuropathies produced by auto-antibodies against α3.

Information is also available on the role of α5 in gut physiology. The nAChR subunit might participate in the control of intestinal smooth muscle contractile responses. In addition to being expressed in enteric neurons, α5 is expressed in the epithelial cells of the gastrointestinal tract where it might participate in ACh-mediated anti-inflammatory responses. Colonic epithelial cells are the main target of ulcerative colitis, a bowel inflammatory condition caused by several factors, including the release of proinflammatory cytokines and prostaglandins, altered gut motility, and reduced colonic mucin synthesis. The influence of nAChR on bowel inflammatory disease is also suggested by epidemiological studies indicating that ulcerative colitis becomes more severe upon smoking cessation, while it ameliorates during smoking. Colonic epithelial cells express α5, and, interestingly, absence of α5 exacerbates an experimental model of colitis. At the same time, the mutation seems to increase nicotine's beneficial effects on colitis, probably due to a change in nAChR affinity.

Nonneuronal nAChRs

As in the previous colonic epithelial cell example, a number of nonneuronal cells express nAChRs. Those receptors might be located at the neuroeffector junction or might be present on cells that are not directly innervated by autonomic nerve terminals. Several nonexcitable, nonneuronal cells have the ability to synthesize and release ACh. In those cells ACh might function as a local hormone to influence both the cells that synthesize the neurotransmitter (autocrine function) and those located in the proximity of the ACh-synthesizing cells (paracrine function). Nonneuronal ACh appears to be involved in the regulation of several cellular processes, including mitosis, cell growth, automaticity, locomotion, ciliary activity, cell–cell contact, cytoskeletal organization, and barrier and immune functions. Among the cells that release ACh are lymphocytes, macrophages, dendritic cells, adipocytes, keratinocytes, endothelial cells, and epithelial cells of intestine, bladder, and lung. These cells express the ACh-synthesizing enzyme choline acetyltranferase and the high-affinity choline transporter but seem to have a different ACh release machinery than neuronal cells. As opposed to the cholinergic nerve endings, where ACh is stored in synaptic vesicles and released following an action potential or cell membrane depolarization, in nonneuronal cells ACh can be released under pharmacological inhibition of the vesicular ACh transporter or the vesicular exocytotic machinery. How ACh is released across the plasma membrane is not yet clear. The activity of the ACh released by nonneuronal cells is mediated by both muscarinic (metabotropic) and nicotinic (ionotropic) receptors, which, due to different coupling mechanisms, will yield intracellular responses on different timescales.

The discovery of nAChRs in nonneuronal cells brings a new perspective on the effects of smoking on the whole organism, as the nicotine contained in tobacco, besides acting at central and peripheral nAChRs, can impact several peripheral functions by binding to nonneuronal nAChRs.

Lung

Messenger RNA for the α3, α5, α7, α9, β2, and β4 but not the α2, α4, α6, or β3 nAChR subunits is found in bronchial epithelial cells. Functional bronchial epithelial nAChRs have been reported to have pharmacological and biophysical properties similar to those of ganglionic receptors and are involved in maintaining the flat shape necessary for those cells to line the bronchial surface. Receptors composed of the α3, α5, and β2 subunits are present in migrating epithelial cells during wound repair in the bronchial epithelium. The presence of mRNA for muscle-type AChRs has also been described in bronchial epithelial cells, but the role of those receptors in lung physiology is yet to be determined.

Airway fibroblasts also express mRNA and protein for nAChRs. α7-containing nAChRs are likely the most abundant neuronal-type receptors in this cell type. Those receptors seem to modulate the synthesis of the extracellular matrix glycoprotein fibronectin. Fibronectin is implicated in tissue injury and repair and is highly expressed in tobacco-related lung diseases. Interestingly, mRNA for muscle α1, β1, δ and ε AChR subunits is highly expressed in fibroblasts as well, but no function for muscle-type nAChRs has been described.

Blood Vessels

The endothelial cells that line blood vessels express functional nAChRs. α3 is the predominant subunit, but mRNA and proteins for α4, α5, α7, β2, and β4 are also detected. Endothelial nAChRs demonstrate ion-gating properties similar to those on ganglionic neurons and might be involved in several ACh-mediated actions, including mitosis, differentiation, organization of the cytoskeleton, cell–cell contact, locomotion, and migration. It is also known that α7-containing nAChRs participate in the mechanisms underlying the migration of endothelial cells induced by

angiogenic growth factors such as vaccinia virus growth factor (VGF).

Vascular smooth muscle cells that lie beneath the endothelial cell layer express the α2, α3, α4, α5, α7, and α10 nAChR subunits. The distribution of those subunits depends on the localization of the vessels. The α7 subunit was found in all vessels examined except those in the kidney. α3 and α5 are widespread but are not present in intrapulmonary or kidney vessels. The role of the various nAChR subtypes in vascular smooth muscle physiology remains for the most part elusive. α7 nAChRs might influence the migration of the vascular smooth muscle cells. nAChRs other than α7 might modulate the synthesis of vascular endothelial growth factor.

The presence of nAChRs on endothelial and smooth muscle cells of the vasculature can explain some of the effects of nicotine on the vascular system, including its ability to increase endothelial cell proliferation, expand capillary networks, and enhance the angiogenic response to inflammation, ischemia, atherosclerosis, and neoplasia.

Urothelium

The ACh released from autonomic nerves regulates urinary bladder function by affecting the activity of the detrusor muscle and the nerve terminals at the neuroeffector junction. Another bladder component affected by ACh is the urothelium. Urothelial cells line the bladder cavity to provide a diffusion barrier. At the same time they can actively release signaling molecules that act in an autocrine or paracrine fashion. ACh is one of those signaling molecules. Because ACh is released when the urothelium stretches, it has been suggested to influence the extent of bladder filling. ACh actions on the urothelium are mediated by both muscarinic and nicotinic receptors. mRNA for the α3, α5, α7, β2, and β4 nAChR subunits has been identified in rat urothelial cells using reverse transcription polymerase chain reaction (RT-PCR). Functional studies conducted in rodents suggest that α3- and α7-containing nAChRs might have opposite roles in controlling bladder filling and voiding by promoting the release of mediators with antithetic functions. Investigations on the human urothelium have so far indicated the presence of the α7, α9, and α10 subunits with a rank order of expression of $\alpha7 \gg \alpha10 > \alpha9$, although ganglionic-type nAChRs are likely to be expressed in the urothelium as well. Interestingly, in human bladders the levels of expression of nAChR subunits seems to be layer specific, with a gradient decreasing in intensity from luminal to basal. The significance of this anatomical arrangement for bladder physiology remains unknown.

Skin

The cholinergic system controls basic skin functions such as terminal differentiation and barrier formation, sweat and sebum secretion, and microcirculation. Keratinocytes are stratified epithelial cells that adhere to each other to provide the skin's first line of protection from the external environment. The phosphorylation of adhesion molecules such as cadherins is one of the factors necessary for optimal keratinocyte adhesion. ACh has been shown to influence this process, as cholinergic antagonists disrupt cell–cell adhesion in keratinocyte monolayers while ACh promotes cell adhesion. Keratinocytes express mRNA for muscarinic receptors as well as that for α3, α4, α5, α7, α9, α10, β2, and β4 nAChR subunits. The epidermis of mice lacking the α3 nAChR subunit presents with loosely attached keratinocytes. In addition, the α3 null mutation decreases E-cadherin and γ-catenin levels while increasing those of β-catenin. Although absence of the α9 nAChR subunit did not alter keratinocyte adhesion, it led to comparable changes in β- and γ-catenin levels, suggesting that the α3 and the α9 nAChR subunit might be regulating either the expression or the posttranslational modification of certain adhesion molecules.

Absence of α7 produces changes that are characteristic of delayed epidermal turnover, and α3- and α7-containing nAChRs influence keratinocyte chemokinesis and chemotaxis, respectively, an effect that has interesting clinical implications for wound healing and spread of cancer metastases.

Dermal fibroblasts, the primal cellular component of the dermis, also express nAChRs. In addition to α3, they express the α5, α7, β2, and β4 nAChR subunits. Lack of α3 produces alterations in fibroblast growth and function, suggesting a role for nAChRs in dermis homeostasis. Finally, the α3, α5, α7, α9, α10, β2, and β4 nAChR subunits are also expressed in hair follicles and sebaceous and sweat glands.

Tobacco is known to aggravate diseases such as psoriasis, atopic dermatitis, acne, and lupus and cause aberrant cutaneous tissue remodeling, such as precocious skin aging and impaired wound healing. Because of those effects of nicotine, the nAChRs found in the various components of the skin are obvious candidates for the study of the mechanisms of tobacco-induced toxicity.

Blood Cells

ACh is found in the blood of several species of mammals, including humans, and it has been suggested that it may originate mainly from T lymphocytes, one of the cell types in the blood. The blood contains two different functional categories of cells:

erythrocytes and leucocytes. Erythrocytes have not been reported to express nAChRs, but leucocytes have. Leukocytes, responsible for the defense of the organism, are subdivided into two categories: granulocytes and lymphoid cells. Granulocytes are further divided into neutrophils, eosinophils, and basophils. nAChR agonists can inhibit stimulated histamine release from human basophils, which have been shown to express α7-containing nAChRs. Nicotine can also prolong neutrophil survival by suppression of apoptosis. This effect could be nAChR mediated, although there are no reports of the expression of nAChRs in neutrophils.

Lymphoid cells, the second category of blood leukocytes, are subdivided into lymphocytes and monocytes. Lymphocytes are the main constituents of the immune system and mount a defense against the attack of pathogenic microorganisms. Lymphocytes B, responsible for the humoral antibody response, express α4(α5)β2 and α7(α5β4) nAChR subtypes, whose levels of expression and roles may vary with the maturation stage of the cells. For example, those two nAChR subtypes may support lymphocyte B survival during differentiation/selection, but limit their activation upon maturation.

Monocytes are the precursors of macrophages, which, after attaining maturity in the bone marrow, enter the blood circulation where they stay for 24–36 h. Afterward they migrate into the connective tissue, where they become macrophages and move within the tissues. In the presence of an inflammation site, monocytes quickly migrate from the blood vessel and start an intense phagocytic activity accompanied by the release of a number of cytokines. The balance between the levels of pro- and anti-inflammatory cytokines and their sequential release may influence the degree of inflammatory responses. ACh released from the vagus nerve is thought to modulate monocyte/macrophage activity by inhibiting proinflammatory cytokine production and systemic inflammation. A central role in these mechanisms is played by α7-containing nAChRs found at the plasma membrane of monocytes/macrophages. Indeed, vagus nerve stimulation attenuates tumor necrosis factor (TNF) release from macrophages, and this effect is abolished in α7 null mice. In addition to inhibiting the release of TNF, α7-containing nAChRs may suppress the release of high-mobility group box 1 (HMGB1), a cytokine involved in severe inflammatory responses such as those observed during sepsis. Receptors containing α4 and β2 may also be involved in controlling the levels of TNF and certain interleukins. Finally, transcripts for the α2, α6, and α9 subunits have also been found in macrophages, but the role of the nAChR subunits that contain them has not been elucidated.

See also: Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autonomic Nervous System; Nicotine; Nicotinic Acetylcholine Receptors; Stress, Sex and Adolescent Nicotine Response.

Further Reading

Galligan JJ and North RA (2004) Pharmacology and function of nicotinic acetylcholine and P2X receptors. *Neurogastroenterology and Motility* 1: 64–70.

Grando SA (2006) Adrenergic and cholinergic control in the biology of epidermis: Physiological and clinical significance. *Journal of Investigative Dermatology* 126: 1948–1965.

Jänig W (2006) *The Integrative Action of the Autonomic Nervous System: Neurobiology of Homeostasis*. Cambridge, UK: Cambridge University Press.

Kawashima K, Kajiyama K, Suzuki T, and Fujimoto K (1993) Presence of acetylcholine in blood and its localization in circulating mononuclear leukocytes of humans. *Biogenic Amines* 9: 251–258.

Racke K, Juergens UR, and Matthiensen S (2006) Control by cholinergic mechanisms. *European Journal of Pharmacology* 533: 57–68.

Ulloa L (2005) The vagus nerve and the nicotinic anti-inflammatory pathway. *Nature Reviews Drug Discover* 4: 673–684

Valenstein ES (2002) The discovery of chemical neurotransmitters. *Brain and Cognition* 49: 73–95.

Wood JD, Alpers DH, and Andrews PL (1999) Fundamentals of neurogastroenterology. *Gut* 45(supplement 2): II6–II16.

Niemann–Pick Disease

J K Fink, University of Otago, Christchurch, New Zealand

Published by Elsevier Ltd.

This article is reproduced from the previous edition, © 2004, Elsevier B.V.

History

Niemann–Pick disease refers to autosomal recessive lysosomal storage disorders due to acid sphingomyelinase deficiency (Niemann–Pick disease, types A and B) or defective cholesterol esterification (Niemann–Pick disease, types C and D). Early descriptions date to 1914 when Niemann, a German pediatrician, described an 18-month-old child with hepatosplenomegaly, progressive neurologic deterioration, and postmortem evidence of lipid accumulation. Over the next 10 years, Pick analyzed similar patients and distinguished the disorder from Gaucher disease on the basis of distinctive cytologic appearance of lipid-laden macrophages. In 1958, Crocker clinically classified Niemann–Pick disease into type A (NPA: acute neuropathic), type B (NPB: chronic nonneuropathic), type C (NPC: chronic neuropathic), and type D (NPD: chronic neuropathic, Nova Scotia variant). Following identification of sphingomyelin and to a lesser extent cholesterol and other phospholipids as the accumulating substances, Brady described marked deficiency of the lysosomal enzyme sphingomyelin phosphodiesterase (acid sphingomyelinase) in NPA and NPB. The acid sphingomyelinase gene was subsequently cloned and mutations identified in NPA and NPB. NPC patients were considered initially to be clinical variants of NPA because they had partial reduction in sphingomyelinase and variable accumulation of sphingomyelin. Subsequent work led by Pentchev et al. showed that the underlying defect in NPC was not sphingomyelinase deficiency but rather defective esterification of endogenous and exogenous cholesterol. Although the underlying biochemical abnormality (or abnormalities) for defective cholesterol esterification in NPC is not known, NPC is genetically linked to a locus on chromosome 18p that is clearly distinct from the acid sphingomyelinase gene on chromosome 11p15.4–15.1. Continued use of terms Niemann–Pick variants or Niemann–Pick disease type C for these patients serves historical purposes and should not be interpreted to imply that these patients share the underlying etiology (primary deficiency of acid sphingomyelinase) of NPA and NPB.

Sphingomyelinase Deficient Forms: NPA, NPB, and NPF

There are marked clinical differences between Niemann–Pick disease types A through F. NPA is a subacutely progressive neurodegenerative disorder of infancy. Developmental milestones may be attained normally for the first several months or occasionally for the first year. Failure to thrive, progressive hepatosplenomegaly, and progressive psychomotor deterioration begin in the first year of life and lead to death in early childhood. Clinical diagnosis, based on the presence of infantile onset, progressive neurodegeneration, accompanied by organomegaly, macular cherry-red spots (in 50% of patients), and bone marrow evidence of lipid-laden macrophages is confirmed by demonstrating acid sphingomyelinase deficiency (less than 5% of normal control subjects) in circulating white blood cells. The recurrence risk of NPA in subsequent progeny (25%) is typical for an autosomal recessive disorder. Prenatal diagnosis can be performed by acid sphingomyelinase assay of amniocyte or chorionic villus samples.

Beginning in infancy or childhood, NPB patients develop progressive hepatosplenomegaly and progressive, diffuse, and pulmonary infiltration. Morbidity is related primarily to pulmonary, hepatic, and skeletal involvement. NPB patients often live to adulthood, but obstructive and restrictive lung disease associated with dyspnea, asthma, recurrent pulmonary infections, and cor pulmonale may limit survival. Accumulation of sphingomyelin-laden macrophages in liver may result in portal hypertension complicated by esophageal varices. Skeletal involvement includes osteopenia and may lead to vertebral body and other skeletal fractures. With rare exception, subjects with NPB do not exhibit progressive neurologic disturbance. Nervous system involvement in NPB is limited to lipid storage and degeneration of retinal ganglion neurons manifested as asymptomatic retinal discoloration (macula halo or granular pigmentation) in a subset of patients. Clinical diagnosis of NPB is based on childhood-onset, progressive organomegaly and pulmonary infiltration, asymptomatic macular abnormalities (in a subset of patients), and bone marrow evidence of lipid-laden macrophages and is confirmed by demonstrating marked deficiency of sphingomyelinase in circulating white blood cells. Like NPA, prenatal testing of NPB is possible by acid sphingomyelinase assay of amniocyte or chorionic villus samples.

Both NPA and NPB are caused by mutations in the acid sphingomyelinase gene, yet only NPA has

prominent neurologic involvement. The specific sphingomyelinase gene mutations responsible for NPA are different than those for NPB. Sphingomyelinase gene mutations associated with NPB generally have a higher amount of residual enzyme activity (~10% sphingomyelinase activity relative to normal control subjects) compared to mutations that cause NPA (<5% activity compared to normal control subjects). Slightly greater residual sphingomyelinase activity in NPB compared to NPA is apparently sufficient to protect central nervous system (CNS) neurons in NPB patients from sphingomyelin storage and degeneration.

Niemann–Pick disease type F (NPF) is a rare benign, nonneuropathic form associated with thermolabile-deficient acid sphingomyelinase, moderate splenomegaly, and mild thrombocytopenia. NPF has been described in two families. It is not known if these patients have mutations in the sphingomyelinase gene.

Cholesterol Esterification Deficiency: NPC, NPD, and NPE

NPC refers to clinically heterogeneous patients with hepatosplenomegaly in whom progressive neurologic deterioration becomes evident in infancy, childhood, adolescence, or adulthood. Separate designations including down gaze paralysis, ataxia foam cell (DAF syndrome), Niemann–Pick variants, neurovisceral storage disease with vertical supranuclear ophthalmoplegia, and juvenile dystonic lipidosis are useful to highlight the phenotype of these patients and emphasize that they do not represent the primary sphingomyelin storage process seen in Niemann–Pick disease types A, B, and F. Three general clinical patterns have been observed in NPC. The first pattern is an early-onset, rapidly progressive form associated with severe hepatic dysfunction and psychomotor delay during infancy and later with supranuclear vertical gaze paresis, ataxia, marked spasticity, and dementia. This infantile-onset, rapidly progressive form is quite rare and clinically similar (although etiologically unrelated) to NPA. The second NPC phenotype is the most common. This delayed-onset, slowly progressive form is heralded by the appearance, usually in early childhood, of mild intellectual impairment, supranuclear vertical gaze paresis, and ataxia, and later associated with dementia and, variably, seizures, and extra pyramidal deficits. Gelastic cataplexy, manifested as momentary episodes of laughter-induced loss of muscle tone with maintenance of consciousness, is occasionally noted in this group of NPC patients. The term juvenile dystonic lipidosis may be appropriate for NPC patients with prominent generalized dystonia. The existence of the first (infantile-onset, rapidly progressive) and second (child through adolescent onset, slowly progressive) NPC phenotypes within the same sibship suggests that they are variant expressions of the same clinicopathologic disorder. The third NPC pattern is a late-onset, slowly progressive form and is distinguished from the second pattern by later age of onset (adolescence or adulthood), much slower rate of progression, and variable presence of organomegaly. Some of these later patients present with dementia or psychiatric symptoms including psychosis. It is important to recognize that virtually all NPC subjects exhibit variable degrees of vertical supranuclear ophthalmoplegia. This important clinical sign is nearly pathognomic for NPC (although it may also be present in olivopontocerebellar degeneration, Huntington's disease, Parkinson's disease, Whipple's disease, and progressive supranuclear palsy). Although subjects with vertical supranuclear gaze paresis are able to maintain fixation and move their eyes passively in response to head movement through all ranges of motion, voluntary vertical eye movements are either very slow or absent entirely. This is readily demonstrated as follows: with the subject visually fixating on a central object, the subject's head is rotated forward and the subject's eyes passively move upward. The subject is then asked to look down. Downward eye movements, if accomplished at all, will be performed by slow, horizontally looping saccades. It is noteworthy that organomegaly, a hallmark of NPA, NPB, and childhood-onset forms of NPC, may be absent or clinically unapparent in adolescent and adult onset NPC. NPC should be considered not only in infants and children with typical signs of organomegaly, progressive ataxia, and vertical supranuclear gaze disturbance, but also in adolescents and adults whose insidiously progressive neurologic dysfunction includes variable degrees of ataxia, dystonia, dementia, supranuclear vertical gaze paresis, or psychosis and in whom liver and spleen size is normal or only slightly enlarged.

Type D (Nova Scotia variant), occurring in French-Canadian descendants of Nova Scotia, is phenotypically and biochemically indistinguishable from NPC and is not considered to represent a separate disease entity.

Lipid accumulation in reticuloendothelial macrophages and CNS neurons in NPC and D is related to impaired esterification of intracellular cholesterol. The molecular basis of impaired cholesterol esterification is not known, although there is evidence that the primary defect involves impaired intracellular translocation of cholesterol leading to delayed induction of cholesterol metabolism and storage as nonesterified cholesterol. The existence of the first and second NPC phenotypes within the same sibship suggests that they

are variant expressions of the same clinico-pathologic disorder. The ability of cultured skin fibroblasts to esterify exogenous-derived cholesterol is markedly impaired (<10% activity compared to cells from control subjects) in each of the NPC phenotypes described above and in phenotypically indistinguishable Nova Scotia variant (NPD). Clinical diagnosis of NPC is based on a constellation of clinical signs including insidiously progressive ataxia, supranuclear vertical gaze paresis, cognitive impairment, and variable splenomegaly and is confirmed by demonstrating deficient cholesterol esterification in cultured skin fibroblasts. This magnitude of impaired ability to esterify cholesterol is associated uniquely with NPC and is not present in NPA, NPB, neuronal ceroid lipofuscinosis, Gaucher disease type III, Friedrich's ataxia, cystinosis, Tay-Sach's disease, Wolman's disease, or Fabry disease. Sphingomyelinase activity in NPC may be normal or mildly reduced. The recurrence risk of NPC in subsequent progeny (25%), like NPA and NPB is typical for autosomal recessive disorders. Assay of cholesterol esterification of cultured amniocytes has been used in prenatal diagnosis of NPC.

The historical term Niemann–Pick disease type E (NPE) referred to adults with chronic lipid storage including sphingomyelin. Individuals with NPE have presented in adulthood with splenomegaly, respiratory problems, or sea-blue histocytes detected in bone marrow during the course of other medical investigations. Other NPE patients have had nervous system involvement including ataxia and cherry red spots. Clearly, the term NPE includes subjects with chronic, nonneuropathic (probable mild NPB) as well as subjects with probable late-onset NPC. Given the availability of acid sphingomyelinase and cholesterol esterification assays, these subjects should now be diagnosed definitively as NPA, NPB, or NPC rather than grouped together as NPE.

Biochemical and Pathologic Features

Patients with Niemann–Pick disease contain lipid-laden macrophages throughout the reticuloendothelial system, particularly in the spleen, liver, lymphoid tissues, and bone marrow. Distension of lysosomes by lipid accumulation and increased numbers of lysosomes create a foamy appearance of these macrophages. Although Giemsa staining may reveal 'sea-blue histiocytes', such cells are not associated specifically with Niemann–Pick disease. The nature of the accumulating lipid varies according to the underlying metabolic defect. Sphingomyelin, an important membrane component of erythrocytes and other cells, accumulates in NPA and NPB because acid sphingomyelinase deficiency prevents cleavage of sphingomyelin into ceramide and phosphorylcholine. In NPC, defective cholesterol esterification causes accumulation of nonesterified cholesterol. Accumulation of other lipids in NPA, NPB, and NPC includes cholesterol (in NPA and NPB), bis(monoacylglycero)phosphate, and glycospingolipids (notably glucosylceramide and lactosylceramide), and G_{M3} ganglioside is attributed to secondary disturbance in other lysosomal enzymes as a result of abnormal lysosomal structure and function due to primary storage of sphingomyelin (in NPA and NPB) or nonesterified cholesterol (in NPC). It is important to recognize the storage of these additional lipids in NPC since the toxic intermediates responsible for regional neuronal degeneration (affecting particularly the cerebellum, for example) have not been defined.

Accumulation of lipid-laden macrophages leads to marrow crowding, hepatomegaly, splenomegaly, and lymphadenopathy. In the liver, storage is primarily in Kupffer cells, leading to organomegaly, portal hypertension, and variable hepatic cirrhosis. Skeletal involvement, particularly in NPB, includes osteoporosis, widening of metaphyses, and pathologic fractures. Accumulation of alveolar macrophages in the lungs (particularly in NPB) leads to diffuse pulmonary infiltration. Involvement of retinal ganglia neurons is evidenced as macular cherry-red spots in 50% of NPA patients. Infiltration of the intestinal mucosa with lipid-laden macrophages leads to chronic diarrhea, malabsorption, and failure to thrive in NPA. Widespread, profound CNS involvement in NPA and NPC is characterized by neurons swollen with vacuolated cytoplasm that is lipophilic in NPA and PAS positive in NPC. Marked loss of cerebral neurons may occur. Electron microscopic examination of brain neurons in NPA and NPC shows multilaminated, membranous cytoplasmic inclusions with electron-dense cores and amorphous electron-dense material similar to lipofuscin. Neuroaxonal dystrophy and accumulation of paired helical filaments (similar to those seen in Alzheimer's disease) have also been observed. Demyelination within the brain is quite variable. Pathologic involvement in NPA and NPC is not limited to the CNS; similar involvement of spinal roots, autonomic plexus, and peripheral nerves has also been observed.

Treatment

Presently, there is no specific treatment to retard or reverse NPA, NPB, or NPC. Cloning the acid sphingomyelinase gene and the feasibility of producing abundant quantities of recombinant acid sphingomyelinase raise the possibility that enzyme replacement therapy, similar to that successfully applied to

Gaucher disease, will benefit patients with NPA and NPB. Alternatively, somatic cell gene therapy may be appropriate for NPA and NPB. Indeed, retroviral-mediated transfer of acid sphingomyelinase cDNA has been shown to correct the biochemical abnormality of NPA fibroblasts *in vitro*. NPB is particularly a good candidate for this approach since the disorder is largely due to lipid accumulation in marrow-derived cells, the precursors of which may be transduced readily by vectors containing the acid sphingomyelinase cDNA. Gene therapy for NPA is more problematic and would require correction of the defect *in utero* (since substantial lipid accumulates prior to birth) and correcting the enzyme deficiency in CNS neurons. In the future, enzyme replacement therapy targeting acid sphingomyelinase to CNS neurons may be efficacious for NPA. The clinical effectiveness of reducing circulating low-density lipoprotein (LDL)-bound cholesterol and suppressing *de novo* cholesterol synthesis is being explored as potential therapy for NPC. Administration of dimethyl sulfoxide (DMSO) in NPC has been tried following observations that cultured DMSO addition to NPC fibroblasts reduced storage of nonesterified cholesterol. Bone marrow and liver transplantation have also been attempted with limited clinical benefit in NPA and NPB. Bone marrow transplantation in NPA and NPC faces the potential limitation that although marrow-derived macrophages may be genetically and biochemically normal, clinical benefit depends on correcting sphingomyelin and cholesterol metabolism in CNS neurons. Recently, subcutaneous implantation of amniotic epithelial cells in NPB yielded promising results.

Future Directions

The greatest importance is placed on developing treatments for these disorders. Enzyme replacement and somatic gene replacement therapies hold significant promise for NPA and NPB. Such treatments for NPC will depend on identification of the underlying genetic defect (the search for which is under intense investigation). Clearly, use of naturally occurring murine and feline models of NPC will provide insight into the underlying biochemical abnormality and facilitate treatment strategies for this disorder. In addition, it is important to understand the molecular basis for clinical heterogeneity in NPC. Finally, it is important to identify among the various accumulating substances, the precise factors that mediate neurotoxicity in NPC. Such information will help explain regional neuronal degeneration in these disorders and hopefully provide insight into the biochemical basis of other degenerative neurologic disorders (including olivopontocerebellar atrophies) that show similar patterns of regional CNS degeneration.

See also: Lysosomal System; Lysosome and Endosome Organization and Transport in Neurons.

Further Reading

Bayever E, Kamani N, Ferreira P, and Machin GA (1992) Bone marrow transplantation for Niemann–Pick Type IA disease. *Journal of Inherited Metabolic Disease* 15: 919–928.

Bembi B, Comelli M, Scaggiante B, and Pineschi A (1992) Treatment of sphingomyelinase deficiency by repeated implantations of amniotic epithelial cells. *American Journal of Medical Genetics* 44: 527–533.

Brady RO, Kanfer JN, Mock MB, and Frederickson DS (1966) The metabolism of sphingomyelin. II. Evidence of an enzymatic deficiency in Niemann–Pick disease. *Proceedings of the National Academy of Sciences of the United States of America* 55: 367.

Brown DE, Thrall MA, Walkley SU, and Wenger DA (1994) Feline Niemann–Pick disease type C. *American Journal of Pathology* 144: 1412–1415.

Carstea ED, Polymeropoulos MH, Parker CC, and Detera-Wadleigh SD (1993) Linkage of Niemann–Pick disease type C to human chromosome 18. *Proceedings of the National Academy of Sciences of the United States of America* 90: 2002–2004.

Crocker AC and Farber S (1958) Niemann–Pick discase: A review of eighteen patients. *Medicine* 37: 1–95.

Higashi Y, Murayama S, Pentchev PG, and Suzuki K (1993) Cerebellar degeneration in the Niemann–Pick type C mouse. *Acta Neuropathology* 85: 175–184.

Patterson MC, DiBisceglie AM, Higgins JJ, and Abe RB (1993) The effect of cholesterol-lowering agents on hepatic and plasma cholesterol in Niemann-Pick disease type C. *Neurology* 43: 61–64.

Pentchev PG, Comly ME, and Kruth HS (1985) A defect in cholesterol esterification in Niemann–Pick disease (type C) patients. *Proceedings of the National Academy of Sciences of the United States of America* 82: 8247–8251.

Quintern L, Schuchman EH, Levran O, Suchi M, Sandhoff K, and Desnick RJ (1989) Isolation of cDNA clones encoding human acid sphingomyelinase. *EMBO Journal* 8: 2469–2473.

Schuchman EH (1995) Two new mutations in the acid sphingomyelinase gene causing type A Niemann–Pick disease: N389T and R441X. *Human Mutation* 16: 352–354.

Spence MW and Callahan JW (1989) Sphingomyelin-cholesterol lipidoses: The Niemann–Pick group of diseases. In: Scriver CR, Beaudet AL, Sly WS, and Valle D (eds.) *The Metabolic Basis of Inherited Disease II*, pp. 1655–1676. New York: McGraw-Hill.

Suchi M, Dinur T, Desnick RJ, and Gatt S (1992) Retroviral-mediated transfer of the human acid sphingomyelinase cDNA: Correction of the metabolic defect in cultured Niemann–Pick disease cells. *Proceedings of the National Academy of Sciences of the United States of America* 89: 3227–3231.

Nightmares

M Schredl, Central Institute of Mental Health, Mannheim, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Nightmares are very common; 70–90% of young adults have reported that they experienced a nightmare at least once in their lifetime. In this article, the definition and etiology of nightmares are reviewed, and a brief and simple treatment technique is presented.

Definition

Nightmares are usually defined as frightening dreams that awaken the dreamer. However, there are two problems with this restricted definition. First, several researchers have demonstrated that very negatively toned dreams that do not wake the dreamer can be as disturbing as nightmares and have coined the term 'bad dreams' for this class of dreams. On average, the emotional intensity of nightmares is higher than that of bad dreams, but there is a large overlap. Whether or not the sleeper is able to judge if the dream emotion was really the cause for waking is still an open question. Some characteristic dreams, such as being chased and waking up just before the attacker grabs you or falling dreams with waking up just before hitting the ground, are obvious examples of dreams that awaken the dreamer, but other dream themes may not be that clear. Second, detailed studies revealed that up to 30% of the nightmares are not dominated by fear but, rather, by other emotions such as extreme sadness, anger, disgust, and confusion.

Nightmares are rapid eye movement (REM) sleep phenomena and often occur in the second half of the night (Table 1). Upon awakening, the person is oriented about her or his surrounding and can give a detailed description of the dream action. Nightmares should be differentiated from night terrors and posttraumatic reenactments. Night terrors are associated with slow-wave sleep and occur predominantly in the first part of the night (usually 1 h after sleep onset). The attack often begins with a piercing cry and – although the eyes are open – the person is not fully awake, experiences intense anxiety (with scarce or no mental content), and does not recall the event in the morning. Sleepwalking often follows a night terror attack. Posttraumatic reenactments also differ from nightmares in that they can occur during REM sleep as well as non-REM sleep (and also during the day – known as 'flashbacks'). Regarding content, there is a continuum from exact replays of the trauma to the nightmares without any obvious reference to the trauma.

Typical nightmare contents are being chased (50%), death or injury (20%), death or injury of close persons (15%), and falling (10%).

Prevalence

The prevalence of children experiencing nightmares once a week or more often is estimated to be approximately 5%. Although nightmare frequency decreases with age (peak prevalence rates occur at ages 6–10 years), large-scale representative studies indicate that approximately 5% of the adult population suffers from nightmares (presumably nightmare frequency is less than once a week in many people). Nightmare frequency is higher in girls and women (with medium effect sizes for the gender difference).

Etiology

Throughout the centuries, many hypotheses regarding the causes of nightmares have been formulated, ranging from devils and demons having sexual intercourse with the dreamer ('succubus' for men and 'incubus' for women) to modern disposition-stress models (Table 2). For the factors depicted in Table 2, empirical evidence of their role in nightmare etiology has been compiled. A large-scale Finnish twin study clearly demonstrated that genetic factors play a role in the etiology of nightmares. Whereas trait anxiety was often not related to nightmare frequency in adults, studies in children and adolescents often found an association between the two. Ernest Hartmann and others have shown that the personality dimension called 'thin boundaries' is associated with nightmares. People who are creative, sensitive, and have intense but problematic relationships, unusual sensory perceptions, and extraordinary occupations suffer from nightmares more often. With respect to the Big Five Factor model of personality, the neuroticism factor is strongly related to nightmare frequency.

Regarding state factors, the presence of a current stressor increases nightmare frequency in children, in nonclinical student samples, and also in nightmare sufferers. Interestingly, one study was able to demonstrate that the effect of personality (neuroticism and thin boundaries) is mediated by the occurrence of stress; that is, people with high neuroticism scores

Table 1 Nocturnal awakenings associated with anxiety

	Nightmare	*Night terror (Pavor nocturnus)*	*Posttraumatic reenactment*
Time of the night	Predominantly in the second part of the night	Predominantly in the first part of the night	Both parts of the night
Sleep stage	REM sleep	Slow-wave sleep	REM sleep or sleep stage 2
Physiological anxiety measures	Moderate	Very strong	Strong to very strong
Dream content	Detailed dream	Rarely a threatening image	Replay of the trauma
Orientation upon awakening	Often fully oriented, although the dream anxiety persists	Not oriented, does not respond properly if addressed	Often fully oriented, strong fear persists
Recall in the morning	Good recall of the dream story	No recall	Good recall
Highest prevalence	6–10 years old	3–7 years old	Event related

Table 2 Etiology of nightmares

Disposition-stress model
Genetic factors
Trait factors ('thin boundaries,' trait anxiety, neuroticism)
State factors (stress)
Trauma
Drugs
Maintaining factors (cognitive avoidance)

experience more stress in their waking life and, therefore, experience nightmares more often. Experiencing a trauma (war experiences, sexual abuse, natural disasters, and severe accidents) can result in a full-blown posttraumatic stress disorder (PTSD) with nightmares as one of the core symptoms, but often nightmare frequency is elevated even if no other PTSD symptoms are present. Within a detailed nightmare history the question about current medication should always be included since quite a few compounds, such as antidepressants (e.g., selective serotonin reuptake inhibitors), acetylcholinesterase inhibitors, or antihypertensive medication, can cause nightmares.

In addition to disposition and triggering factors, one has to keep in mind that nightmares are anxiety phenomena like phobias which persist if the person avoids an active confrontation with the fear. In the case of nightmares, the largest percentage of an adolescent sample reported that trying to forget the nightmare was their primary coping strategy (Table 3). Statements such as "This was only a dream" can be considered as cognitive strategy to avoid confronting the anxiety experienced within the dream. Therefore, the avoidance behavior maintains the nightmares even if the original trigger is no longer active.

Nightmares and Sleep Quality

Several researchers conceptualize nightmares as a primary sleep disorder, although they are categorized into the group of parasomnias and not into the group of dyssomnias such as primary insomnia. Empirical evidence has clearly demonstrated low subjective sleep quality in nightmare sufferers, presumably due to direct effects (waking up from the nightmare and prolonged periods of wakefulness after the nightmare before falling asleep again) and indirect effects (fear of falling asleep in the evening because of the reoccurrence of nightmares). Since stress is associated with poor sleep quality and nightmares, partial correlational analyses were carried out to determine whether nightmares are associated with poor sleep quality even if the stress level is statistically controlled. The results clearly indicate that nightmare frequency and severity of insomnia complaints were associated. The therapy studies performed by Barry Krakow and co-workers also showed that a reduction of nightmare frequency is paralleled by an increase in sleep quality.

Nightmare Frequency versus Nightmare Distress

The hypothesis has been put forward that poorly adjusted people (i.e., those with high neuroticism scores) are more distressed by their nightmares – regardless of frequency – than well-adapted people. This would explain the inhomogeneous findings regarding the relationship between waking-life psychopathology and nightmare frequency and the higher correlation coefficients for the relationship between nightmare distress and psychopathology compared to the correlation coefficients between nightmare frequency and psychopathology. However, most studies have relied on the Belicki scale for measuring nightmare frequency, but this measure is strongly confounded with nightmare frequency (correlation coefficients ranging from $r = 0.24$ to 0.47). To avoid this methodological issue, we developed a scale that measures the distressing effect of a single nightmare and correlated this measure with neuroticism. As expected, neuroticism was correlated with

Table 3 Coping strategies used for bad dreams and nightmares

Category	*Total (N = 606) (%)*	*Boys (n = 329) (%)*	*Girls (n = 277) (%)*
No bad dreams	13.5	19.5	6.5
Tell the dream	22.9	14.3	33.2
Draw a picture	2.0	2.4	1.4
Record (diary)	2.5	0.9	4.3
Try to forget the dream	39.9	41.0	38.6
Image a positive ending	10.1	11.6	8.3
Other	9.1	10.3	7.6

nightmare frequency ($r = 0.39$) and with the sum score of the Belicki scale ($r = 0.42$) but not with the single nightmare distress scale ($r = 0.19$) or nightmare intensity ($r = 0.04$). The hypothesis that people with high neuroticism scores overestimate nightmare distress can be rejected; nightmare frequency is the major factor in determining distress attributed to nightmares.

Nightmares and Sleep Physiology

In the nineteenth century, several scientists held the view that nightmares are caused by a shortage of oxygen. It was sometimes said that a living creature was sitting on the chest ('Nachtmahr' or 'Alpdrücken' in German) or choking the dreamer. Based on experience with his own nightmares, Boerner concluded that blocking the mouth and nose by the bedspread or by the pillow (sleeping in the prone position) caused nightmares. Using a cloth to block the mouth and nose, Boerner was able to induce nightmares in three different sleepers. It seems obvious that studying patients with nocturnal breathing disorders, such as the sleep apnea syndrome, is a promising approach to test the oxygen hypothesis. The sleep apnea syndrome is characterized by apneas (minimal duration of 10 s) which can be accompanied by severe oxygen desaturation. An alarm reaction with an arousal terminates the apnea by increasing muscle tone and reopening the upper airways. In pilot studies, a few examples of breathing-related nightmares have been collected. A 39-year-old male patient with sleep apnea (respiratory disturbance index, 68.1 apneas per hour; maximal decrease in blood oxygen saturation, 43%) stated the following:

> During the dream I felt tied up or chained. I saw thick ropes around my arms and was not able to move. I experienced the fear of suffocation without being able to cope with the situation. Powerlessness and also resignation came up.

One patient reported having recurrent nightmares of the following type: He was lying in a coffin – dead. He cried, "My god, please don't let me die, I am still young and I have small children," and woke up, gasping for breath.

Despite the illustrative examples, a large-scale study with 323 apnea patients did not find any correlation between nightmare frequency measured retrospectively via questionnaire and respiratory parameters (respiratory disturbance index and nadir of oxygen desaturation). Overall, nightmare frequency was not elevated in comparison to that of healthy controls. That is, the oxygen hypothesis is not supported by modern research findings.

Anne Germain and Tore Nielsen carried out a polysomnographic study in 11 nightmare sufferers. They reported a nightmare frequency of one or more nightmares per week. Although they reported considerable nightmare distress, their sleep parameters recorded in the sleep laboratory did not differ from those of healthy controls (e.g., sleep efficiency (92.1 vs. 92.7%), sleep onset latency (14 vs. 11 min), and REM sleep percentage (21.2 vs. 21.5%)). However, they found an elevated index of periodic limb movements during sleep which might be an effect of intense negative dreaming. The sleep laboratory findings (very few studies with small numbers have been carried out) contrast with the subjective reports of the patients. This might be an effect of the laboratory surroundings: many sleep researchers have reported that nightmares occur less often in the lab than at home. This effect might be attributed to the 'caring' environment – the presence of a sleep technician, someone interested in their history, listening to them empathically and taking them seriously. Similar effects (discrepancy between subjective sleep quality and polysomnographic findings) have been reported for other patient groups, such as patients with borderline personality disorder, patients with panic disorders, and adult patients with attention deficit/hyperactivity syndrome.

Charles Fisher and co-workers carried out an extensive laboratory study to investigate the physiology of nightmares. They included 38 patients suffering from three or more nightmares per week and studied

their sleep over 162 nights. Although they recorded a fair number of night terror attacks ($n = 50$), only 22 nightmares occurred in the laboratory. From the patients' evaluations, the attacks in the lab were much less severe than those at home – again the lab effect. Only in three very intense nightmares were marked increases in heart rate (an increase of 20 beats per minute), respiration rate (an increase from 18 to 30 breaths per minute), and the number of eye movements measured in the 8 min prior to the awakening. This is in contrast with the night terror findings: documented increases in heart rate from 64 to approximately 120–150 beats per minute during the attacks. Despite the intense anxiety experienced during a nightmare, the autonomic reaction is moderate (Table 1).

Therapy of Nightmares

The question arises as to when nightmares should be treated. Similar to other mental disorders, it is important to evaluate whether the person suffers from her or his nightmares in a considerable way. Treatment should be offered to patients reporting negative effects on daytime mood and functioning, worries at sleep onset (fear that a nightmare might occur), and insomnia complaints. If a cutoff is needed, for example, for evaluating nightmares in children, a frequency of one or more nightmares per week is reasonable.

Whereas drugs such as REM sleep-suppressing tricyclic antidepressants are often ineffective in the treatment of nightmares, several different psychological interventions seem promising. Although several case report series indicate that lucid dreaming is an effective tool in coping with nightmares, the major problem with this technique is the fact that induction of lucid dreams (learning to have dreams when you know you are dreaming) is a highly expensive and protracted process. Systematic desensibilization was shown to be effective in the treatment of nightmares in two carefully performed studies. The most effective and simplest treatment strategy was developed by Barry Krakow and co-workers and is called imagery rehearsal treatment (IRT). Based on my clinical experience, a reduced form of IRT including the basic steps necessary for the effective treatment of nightmares is depicted in Table 4. For most patients, it is helpful to point out the analogy of coping with fear and anxiety in the waking life.

The first step consists of confrontation, which is simply writing down the dream or, for children, the drawing of the most important dream scene. In the second step, the person is asked to imagine a new ending to the dream. Children were asked to draw something in the picture which reduced anxiety. Ideally, the therapist should encourage the person to develop his or her own personal coping strategy for the nightmare situation and refrain from giving suggestions. The coping strategy should include active behavior. Flying away or hiding is often not helpful (because these reflect avoidance), so one simply asks whether other strategies might be applicable as well. The 'solution' of the dream must not be realistic; most psychologists advise not to increase violence (e.g., killing the opponent) while coping with the dream situation. The 'new dream' should be written down. The last step consists of internalizing the new strategy by imagining the new ending 5–10 min every day during a 2-week period. The repetition during waking life affects subsequent dreams. Interestingly, the effect of the intervention often generalizes; that is, nightmares with other topics are affected and include more adequate coping strategies.

Table 4 Therapeutic principles

Step 1: Confrontation
Record or draw the dream
Step 2: Coping with the situation in the nightmare
Construct a new dream ending or add something to the drawing that reduces the fear
Step 3: Training of the coping strategy
Approximately 5–10 min per day over a 2-week period

For IRT, several randomized controlled trials have been carried out including life-long nightmare sufferers (idiopathic nightmares) and sexually assaulted women with PTSD. In one of the studies by Barry Krakow and co-workers, a single $2\frac{1}{2}$ h group session was effective in reducing nightmare frequency in 39 patients with idiopathic nightmares from six nightmares per week to two. In addition, sleep quality and daytime mood improved. To illustrate the procedure, two case reports from my clinical practice are presented here.

Case Report – Adult

A 22-year-old woman consulted our sleep center due to nightmares that started several months ago. The anamnesis also indicated that several night terror attacks also occurred. Two polysomnographic recordings (including breathing parameters and tibialis EMGs) were normal, with the exception of low sleep efficiency (first night, 67.8%; second night, 76.7%). Nightmares and night terrors did not occur in the sleep laboratory. The psychometric measurement of sleep behavior and sleep quality revealed that sleep quality and the feeling of being refreshed in the morning were not reduced but daytime sleepiness

score was elevated (1.6 standard deviations above the mean of a comparable student sample).

In the first session, the patient reported current stressors. A severe conflict in the core family resulted in a complete cessation of her relationship with her parents. The therapeutic principles, including confronting, coping, and training, were explained to the patient. It was easy for her to understand this approach because it was similar to the treatment of anxiety. In the nightmare of the previous night, the patient was in the dressing room of a fitness club. At first, her complete family was present. Then, her mother was in the center and verbally attacked and criticized the dreamer. She felt very helpless with respect to the accusations. Answering the question as to how the dream could be altered, she came up with several suggestions: ignoring the mother and active acting, such as saying something to her mother and expressing her needs. The 'new dream' included the active confronting of the mother and the sentence, "I can manage my life myself."

In the second session 2 weeks later, the patient reported that she trained the new dream ending regularly during the first week and that no new nightmares including her mother occurred. However, several other negatively toned dreams occurred, one of which was a recurring nightmare. In it, the dreamer stays at the house of her grandmother and is now in the kitchen. At first, the overall emotional tone is positive. Then she senses that there is something threatening outside the house that might come inside and threaten her. The scene changes: the dreamer finds herself in a white room with several fantasy characters (pleasant characters). Two of the characters encourage her to step outside the house and confront the threatening something alone. At this point, she usually wakes up. The awakening from the last dream was at 5 a.m., and the anxiety she felt in the dream was still present after waking up. After talking about the different dream elements (e.g., the relationship to her grandmother, to whom she has a positive connection), the patient was again asked to imagine a new ending to the dream. The patient visualized how she goes outside and confronts the unknown threat with the help of the friendly fantasy characters.

In the follow-up session 2 weeks later, the patient reported a complete cessation of nightmares and that the dreams she recalled incorporate new and active behavior patterns. The dreamer felt more confident in dealing with other dream characters and was able to express her needs. She also reported a marked lessening of daytime sleepiness. Although the stressors did not change over the treatment period, the intervention did reduce nightmare frequency by strengthening the dream ego – that is, the negative effect of the stressor on dreaming and sleep was significantly reduced.

Case Report – Child

A mother consulted our sleep center because her son (5 years old) suffered from nightmares every night during the past month. Before that, nightmares rarely occurred. Common topics are ghosts, shadows, and horrid monsters. At night, he calls out for his mother and needs comfort for up to 2 h because after the nightmare he cannot fall asleep easily. The mother was not aware of any incidents triggering the period of intense nightmares. The boy was somewhat shy but free from any psychopathology or somatic problems.

After taking the nightmare history, the boy was asked to draw a recent nightmare (**Figure 1**). The dreamer is walking within a castle (**Figure 1**, lower left) and carrying a book. He is very frightened by the sight of two enormous ghosts (the differences in size between the figures and himself clearly reflect the fear involved). After he was asked what might help reduce his anxiety, the boy drew a large spider between him and the ghosts, facing the ghosts and protecting him.

After the first session, the boy continued this work with the help of his mother by drawing several pictures of other nightmares and concentrating on one of them each day for approximately 10–15 min. Within 2 weeks, nightmare frequency was drastically reduced. Although the topics of the dreams did not change much, the fear associated with these pictures was almost completely gone. Sleep behavior and quality were back to normal. A follow-up session 1 year later revealed that nightmares were still rare (approximately once a month) and the boy did not feel distressed by them. For several nightmares, he invented new solutions (one involving a magic spray). He has developed a positive attitude toward his dreams which can be seen in the fact that he occasionally asks his mother to record a dream in the booklet which she had kept during the study period.

Conclusion and Future Directions

Nightmares are common and approximately 5% of the population report that they suffer from nightmares. Fortunately, nightmares can be treated effectively with a brief and simple intervention method which can be classified as a cognitive technique. The beneficial effect has been demonstrated for idiopathic nightmares as well as posttraumatic nightmares in children and in adults. The basic principle of confronting and coping with the nightmare situation is easily understandable to patients and can be done

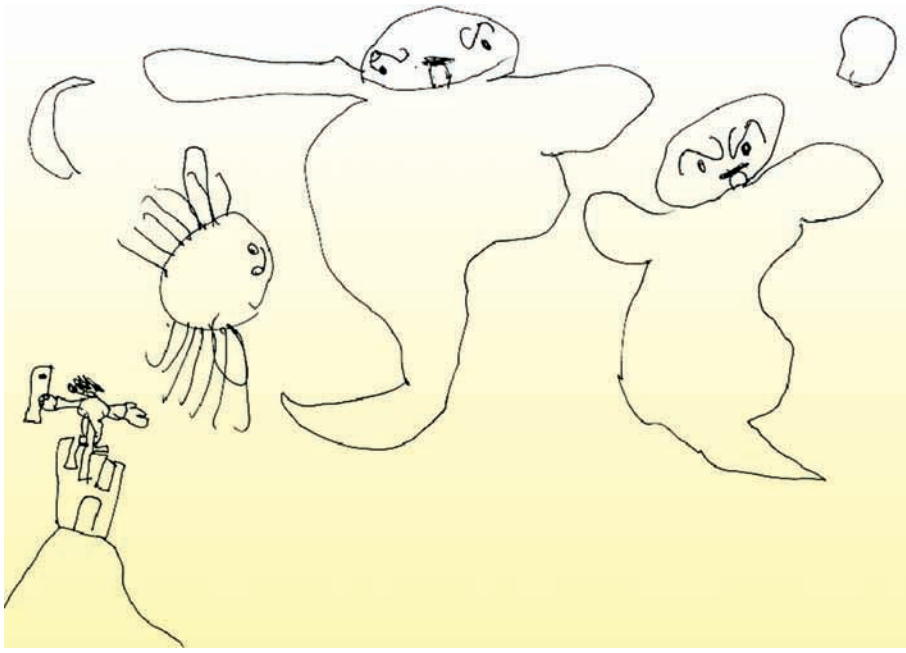

Figure 1 Drawing of a nightmare with a coping strategy included.

without professional assistance, thus encouraging self-help.

In addition to the clinical aspect, nightmares are also an interesting topic for basic sleep and dream researchers. The body–mind interaction can be studied – for example, the correlation between emotional intensity and brain activation related to emotions in waking life (limbic system). Techniques such as magnetic resonance imaging will be applicable to sleep if the scanner noise can be markedly reduced. Since nightmares occur quite rarely in the sleep laboratory, large-scale studies with ambulatory measurement units should be conducted to shed light on the physiological parameters during REM sleep before awakening from a nightmare. The lack of consistent results from sleep laboratory studies may be attributable solely to the 'lab effect' (reducing emotional dream intensity due to the 'caring' environment). Since nightmares reflect stressors and PTSD-related nightmares often replay the trauma experienced by a person, they may help in elucidating the role of dreams in general in sleep-related memory consolidation. Research has shown that sleep facilitates consolidation of declarative and procedural memory tasks. Whether dreams play a crucial role in that relationship remains to be determined.

See also: Dream Function; Dreams and Nightmares in PTSD; Dreams, Dreaming Theories and Correlates of Nightmares; Posttraumatic Stress Disorder: Neurobiology; Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep Apnea; Sleep in Adolescents; Sleep in Aging.

Further Reading

Germain A and Nielsen TA (2003) Sleep pathophysiology in posttraumatic stress disorder and idiopathic nightmare sufferers. *Biological Psychiatry* 54: 1092–1098.

Hublin C, Kaprio J, Partinen M, and Koskenvuo M (1999) Nightmares: Familial aggregation and association with psychiatric disorders in a nationwide twin cohort. *American Journal of Medical Genetics (Neuropsychiatric Genetics)* 88: 329–336.

Krakow B and Zadra A (2006) Clinical management of chronic nightmares: Imagery rehearsal therapy. *Behavioral Sleep Medicine* 4: 45–70.

Schredl M (2003) Effects of state and trait factors on nightmare frequency. *European Archives of Psychiatry and Clinical Neuroscience* 253: 241–247.

Schredl M, Blomeyer D, and Görlinger M (2000) Nightmares in children: Influencing factors. *Somnologie* 4: 145–149.

Schredl M, Landgraf C, and Zeiler O (2003) Nightmare frequency, nightmare distress and neuroticism. *North American Journal of Psychology* 5: 345–350.

Spoormaker VI, Schredl M, and Van den Bout J (2006) Nightmares: From anxiety symptom to sleep disorder. *Sleep Medicine Reviews* 10: 19–31.

Zadra A and Donderi DC (2000) Nightmares and bad dreams: Their prevalence and relationship to well-being. *Journal of Abnormal Psychology* 109: 273–281.

Nitric Oxide

G Wilson and J Garthwaite, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Nitric oxide (NO) is a unique type of cell-to-cell signaling molecule found in almost all tissues of the body and throughout the central nervous system. NO rose to biological fame in 1987 when it was identified as the factor released from endothelial cells lining blood vessels to cause relaxation of the underlying smooth muscle, thereby influencing blood flow and blood pressure. The following year, it was found to be the mysterious substance known to be produced following activation of the *N*-methyl-D-aspartate (NMDA) class of glutamate receptors on brain neurons to relay signals to neighboring cells, and also a molecule generated by activated macrophages as part of their repertoire of defense against foreign organisms. These three diverse functions – smooth muscle relaxation, neural communication, and immune defense – remain at the core of NO biology and are subserved largely by specialized NO-generating proteins known, respectively, as endothelial, neuronal, and inducible NO synthases (eNOS, nNOS, and iNOS, respectively). As well as being a key physiological signaling molecule, NO has been invoked as an important player in many different disorders affecting humans, including neurodegenerative diseases, pain, atherosclerosis, and septic shock.

The physicochemical properties of NO set it apart from conventional signaling molecules, such as neurotransmitters. First, like oxygen and carbon dioxide, NO lacks chemical specialization. It does, however, possess an extra (unpaired) electron, making it a radical. Whereas some other radicals are chemically reactive and can cause damage to cells, NO is stable in physiological concentrations, which are thought to be approximately 1 nM or less. Second, also like oxygen and carbon dioxide, NO diffuses very quickly through membranes, obviating the need for a specialized release mechanism and giving it the ability to act on neighboring cells within microseconds of its manufacture. This property gives the recipient cell temporal information about the events causing NO to be formed. Finally, NO binds very avidly to heme groups possessing vacant coordination sites. Indeed, the capacity of hemoglobin to bind and inactivate NO was well known long before NO was recognized as a biological messenger, and this property proved to be an important one in developing the hypothesis that the biological messenger was NO and also in testing that hypothesis. Although many cellular heme groups do not possess the necessary vacant binding site, this property has been exploited to provide highly sensitive NO detectors within cells, initiating physiological NO signal transduction.

How Is NO Made?

The NO synthase family of enzymes catalyzes the stereoselective conversion of the amino acid L-arginine to NO. The reaction requires molecular oxygen and L-citrulline is formed as a by-product. The enzymatic conversion of L-arginine into NO was first shown in 1989 and, later, the three different isoforms were identified in molecular terms (chronologically, nNOS, eNOS, and then iNOS).

The enzymes exist as homodimers, with each constituent subunit containing an N-terminal oxygenase domain and a C-terminal reductase domain. The former contains binding motifs for heme, tetrahydrobiopterin (BH_4, a cofactor), and the substrate, L-arginine. The reductase domain possesses binding sites for NADPH, FAD, and FMN. Functionally, electrons are donated by NADPH to the reductase domain and proceed via the FAD and FMN redox carriers to the oxygenase domain, whereupon interaction with the heme iron and BH_4 at the active site catalyzes the generation of NO. The two domains are linked by a binding site for calmodulin and the binding here of calmodulin complexed with calcium facilitates electron flow through the enzyme (Figure 1). The calcium dependence of NO synthesis varies with the NOS isoforms, with nNOS and eNOS having a much higher requirement than iNOS. The calcium insensitivity of iNOS is the result of the lack of an autoinhibitory loop insert within its FMN-binding subdomain that, in the other two isoforms, acts by destabilizing calmodulin binding at low calcium concentrations. Loss of eNOS activity at low calcium concentrations is circumvented when the protein is phosphorylated, typically by the kinase known as Akt, or protein kinase B. Physiologically, eNOS phosphorylation may be the most important device for maintaining NO output in blood vessels.

Knowledge of the enzymatic mechanism for NO synthesis led to the identification of inhibitors, which have become invaluable tools for investigating the biological roles of NO. First-generation inhibitors were L-arginine derivatives with substituents on the guanidino nitrogen normally used for NO synthesis,

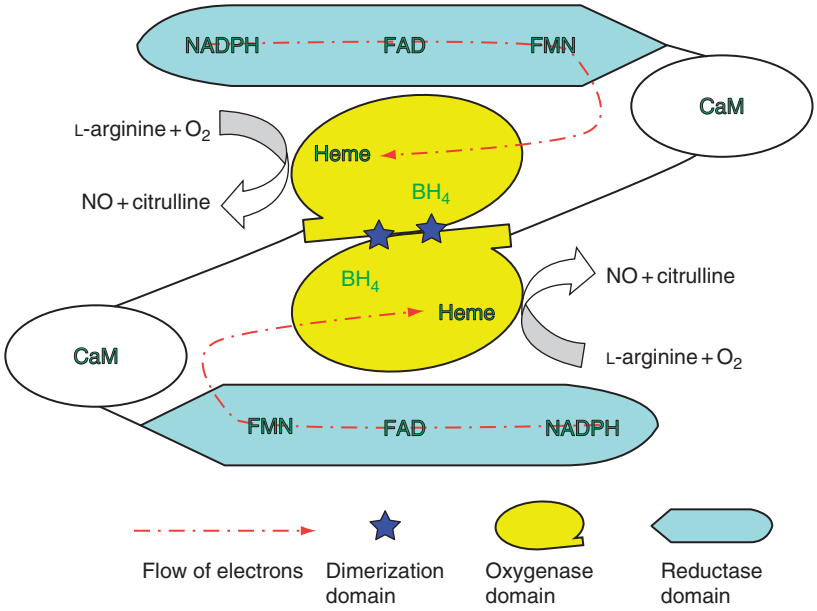

Figure 1 NO biosynthesis. The NO synthases are homodimers, with each subunit containing a reductase and oxygenase domain. Electrons donated by NADPH proceed through the reductase domain via FAD and FMN redox carriers. On reaching the oxygenase domain, electrons interact with the heme group and BH_4 cofactor at the active site to catalyze the reaction of L-arginine with molecular oxygen, yielding NO. CaM, calmodulin.

compounds such as N^G-monomethyl-L-arginine and N^G-nitro-L-arginine. These inhibitors are able to inhibit all NOS isoforms, but compounds showing good selectivity for nNOS or iNOS over the other isoforms are now available.

NO Generation in the Nervous System

The nNOS isoform is found predominantly in neurons and has a distribution in the brain and spinal cord as wide as that of the major neurotransmitters glutamate and γ-aminobutyric acid, although the amounts vary from region to region (Figure 2(a)). In some brain areas, such as the cerebellum, nNOS is found in virtually all cells, whereas in others, such as the striatum and cerebral cortex, it is found in a subpopulation of interneurons comprising only a few percentage of the total neuronal number. Nevertheless, even in those regions that are relatively sparsely endowed with nNOS-containing neurons, dense nNOS-containing fiber networks are to be seen ramifying throughout the neuropil, suggesting that the majority of neurons receive NO signals.

In many of these brain regions, NO formation is coupled to the stimulation of the NMDA class of glutamate receptor, which is found in almost all excitatory synapses. This special relationship is the result of the NMDA-receptor-associated ion channels having a high permeability to calcium ions and of the compartmentalization of the nNOS and NMDA receptor proteins into a complex at synaptic sites. The scaffold for this complex is postsynaptic density-95 (PSD-95) protein, which contains several PDZ domains – the modular protein–protein interaction motifs. The N-terminus of nNOS contains a PDZ domain which interacts with the second PDZ domain on PSD-95. PSD-95 also binds the C-terminal PDZ domain of NMDA receptors, leading to the colocalization of the two proteins. In general, NMDA receptors play only a minor role in fast synaptic transmission but are brought into action when there is a concomitant depolarization of the postsynaptic membrane which, by removing a block of the NMDA receptor channel by magnesium ions, allows calcium ions to pass into the cell. One result of NMDA receptor activation in many synapses is the induction of long-term increases or decreases in synaptic strength, phenomena considered relevant to learning and memory formation and ones in which NO is also commonly involved.

In addition to the coupling with NMDA receptors postsynaptically, there are other instances in which NO is formed presynaptically. The best examples of this neurotransmitter-like mode of action are to be found in the peripheral nervous system, where NO is released from nonadrenergic, noncholinergic nerves to act on smooth muscle (e.g., in the intestine) and relax it. The trigger for NO release here is the activation of voltage-dependent calcium channels by invading action potentials. Also, in parts of the brain, nNOS is concentrated strongly in axons (e.g., in the parallel fibers in the cerebellum) and it may perform

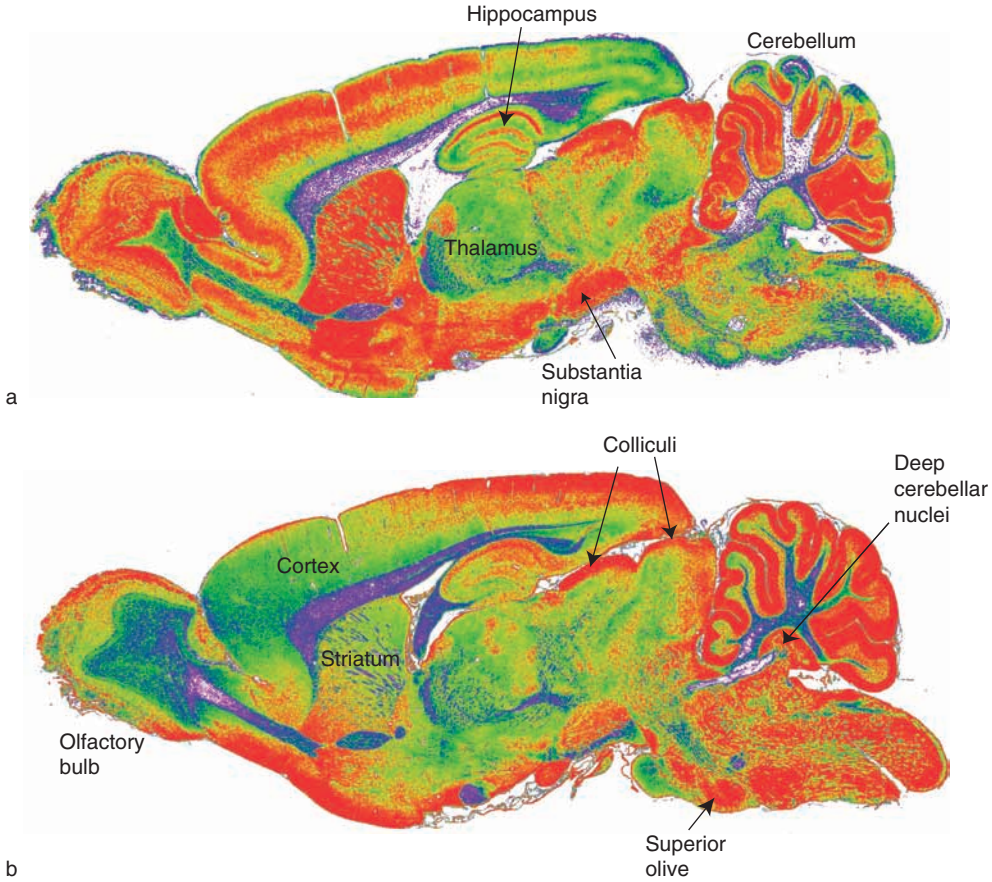

Figure 2 The NO–cGMP signaling pathway in rat brain. Shown are the distributions of (a) NO synthase as indicated by NADPH diaphorase histochemical staining and (b) NO-stimulated cGMP accumulation after perfusion *in vivo* of the NO donor, sodium nitroprusside. The images are of adjacent parasagittal sections and are displayed as heat maps, with red indicating strong staining intensity and blue indicating weak staining intensity. Generally, there is a good match between the presence of NO synthase and cGMP production. Pictures of unmodified sections and experimental details can be found in Southam E and Garthwaite J (1993) The nitric oxide–cyclic GMP signalling pathway in rat brain. *Neuropharmacology* 32: 1267–1277.

an analogous 'orthograde' transmitter function at these locations.

Because it is such an unconventional transmitter, developing a conceptual framework for synaptic NO signaling has been problematic, not least because NO is very difficult to measure at physiological concentrations and in discrete synaptic locations. By analogy with the numbers of NMDA receptors, there may be only 10–100 nNOS molecules per synapse. Assuming that the nNOS has a similar activity to that exhibited by the purified enzyme and that, once made, it is dispersed by diffusion, active NO concentrations (in the 0.1–1 nM range) would only be found very locally, within a submicrometer radius. Hence, NO from nNOS may only be able to function in subsynaptic dimensions, implying that it signals between closely juxtaposed elements, such as the pre- and postsynaptic specializations.

Although nNOS is the main isoform in the nervous system, emerging evidence suggests that NO from eNOS in the microvasculature can also influence neurons. This represents a conceptual departure from the traditional view of brain function, in which signal processing is perceived as being performed by neurons with the support of nearby glial cells. On the other hand, any point in the brain is maximally only approximately a cell diameter (approximately 25 μm) away from a capillary and the geometry of the capillary network is ideal for distributing a molecule such as NO within a tissue volume, just as it is for oxygen. At least *in vitro*, the global NO signal deriving from the endothelium can influence the membrane potential of axons in the optic nerve and the capacity for synaptic plasticity in the hippocampus.

How Does NO Act?

Curiously, the answer to this question preceded the identification of NO as a biological messenger. In the 1970s, it was found that NO was a powerful

activator of guanylyl cyclase (GC) enzymes, causing the synthesis of cyclic GMP (cGMP). This observation provided crucial evidence for the hypothesis that endothelium-derived relaxing factor was NO because the factor was found to cause relaxation through cGMP generation. Because of its location in the soluble fractions in tissue homogenates, the enzyme activity became known as 'soluble' GC, but it is now clear that the proteins constitute major physiological receptors for NO and we refer to them here as NO(GC) receptors. In the rodent brain, NO-evoked cGMP accumulation has a similarly broad distribution to that of the NO-forming enzyme (**Figures 2(a) and 2(b)**).

The NO(GC) receptors are designed to capture and transduce low-level NO signals, even very transient ones. The ligand-binding site is a specialized heme group that excludes oxygen, allowing NO to bind without becoming oxidized, in contrast to its reaction with oxyhemoglobin when nitrate ions are formed. The proteins to which the heme is attached are heterodimers of α- and β-subunits and exist in two main isoforms. The β_1 subunit appears to be the common subunit and it is partnered by either α_1 or α_2. Both isoforms are found abundantly in brain tissue but have an uneven distribution, with some areas or cells showing prominently $\alpha_1\beta_1$ and others showing $\alpha_2\beta_1$. A special feature of the $\alpha_2\beta_1$ isoform is that it binds to pre- or postsynaptic scaffold proteins, which would be convenient for transducing localized NO signals from nNOS, whereas the $\alpha_1\beta_1$ may have a less localized distribution in neurons and can also be found in some glial cells. A β_2 subunit is present in low levels at the mRNA level, but there is no evidence that a β_2 protein is expressed.

The binding of NO to the vacant coordination site on the heme group of the NO(GC) receptor is extremely rapid, being nearly diffusion limited. Initially, a six-coordinate nitrosyl–heme complex is formed, but the subsequent snapping of a bond between the heme and a nearby histidine residue results in a five-coordinate species. This bond breakage causes a conformational change that propagates to the catalytic domain, accelerating conversion of GTP into cGMP by 1000-fold or more. Although both subunits contribute to the catalytic domain, there is only a single active site. The second 'pseudo-symmetric' site, which is thought to be analogous to the forskolin binding site in adenylyl cyclase, may have a regulatory function.

Binding of NO to other heme groups is so tight that it often takes hours or days for it to unbind, a property that would preclude a dynamic signaling role for NO. Although of high affinity (dissociation constant of approximately 1 nM), binding of NO to the NO(GC) receptor heme is readily reversible so that, on removal of NO, the associated GC activity switches off in the 100 ms timescale. Overall, the properties of the receptors mean that they can sense NO concentrations of approximately 0.1 nM or more, convert them into approximately 1000-fold higher concentration of cGMP within 1 s, and then cease activity when NO production stops, again within 1 s. Thus, although many details of their functioning remain to be explained, the NO(GC) receptors are highly effective proteins for transducing NO signals into greatly amplified levels of cGMP with good temporal fidelity. Whether or not other NO receptors exist remains to be determined.

Inactivation of NO

Brain tissue avidly consumes NO such that its half-life at physiological concentrations is estimated to be in the 10 ms timescale. Although rapid, this process is unlikely to contribute much to localized NO signaling because diffusion alone ensures rapid dissipation of the molecules from small volumes. Active NO inactivation is more likely to be of importance in shaping the NO concentrations when there are many nearby sites of simultaneous NO synthesis. Despite much research, the means by which the NO signal is actively quenched remains unclear. Reaction with oxyhemoglobin in red blood cells in the vasculature probably plays some role. Chemically, NO can also react with oxygen, but the rate is too slow to be of physiological importance. NO can, however, react very rapidly with lipid peroxyl radicals, although this process is likely to be significant mainly in pathological conditions associated with oxidative stress. The major mechanism of NO consumption in normal brain tissue remains to be identified.

Downstream Targets

On being generated in response to NO, cGMP may act on protein kinases to initiate phosphorylation cascades, on phosphodiesterases responsible for the hydrolysis of cGMP and/or cAMP, or on cyclic nucleotide-gated/modulated ion channels located in cell membranes.

cGMP-dependent protein kinases (cGKs) are activated by submicromolar cGMP concentrations and then phosphorylate serine or threonine residues. There are three types: cGKIα and -1β (splice variants) and cGKII. Activation appears to involve a conformation change that removes the autoinhibitory domain from the catalytic site. Autophosphorylation can also occur, causing a more persistent kinase activity that could outlast the original cGMP signal. In brain, cGKIα is concentrated in Purkinje cells in the cerebellum,

neurons that also express a 23 kDa target protein known as G-substrate which functions as a protein phosphatase inhibitor. By engaging this cascade in Purkinje cells, NO contributes to synaptic plasticity and motor learning in the cerebellum. The hippocampus is enriched in the cGK1β isoform, which is important in the NO-dependent initiation of synaptic plasticity in this region, whereas cGKII is more widespread, being found particularly abundantly in the cerebral cortex, thalamus, olfactory bulb, and superior colliculi. One role of cGKII is in the regulation of circadian rhythms.

Phosphodiesterases (PDEs) catalyze the hydrolysis of cyclic nucleotides to the corresponding noncyclized monophosphates. In mammals, there are 11 PDE families encoded by genes with several splice variants that may exhibit different subcellular locations or tissue distributions. Most of the PDEs can hydrolyze cGMP; those with the greatest affinity are PDE1–3, -5, -6, and -9 through -11. The first family, PDE 1, is a widely expressed, calcium/calmodulin-dependent isoform that hydrolyzes both cGMP and cAMP. Competition between the two cyclic nucleotides could occur, potentially allowing a rise in one to inhibit breakdown of the other. This scenario is more probable with PDE3, however, because this isoform binds both cGMP and cAMP with similar affinities but hydrolyzes cGMP at a much lower maximal rate. In effect, therefore, cGMP can act as a PDE3 inhibitor and raise the levels of cAMP. Although there is little information available on PDE3 in brain, there is evidence that cGMP signals are transduced through this mechanism in terminals of the vagus nerve in the heart, with the increased level of cAMP ultimately causing increased release of acetylcholine. Another cGMP-regulated PDE, PDE2, is a dual-substrate enzyme equipped with a noncatalytic cGMP binding site (called a GAF domain) whose occupation stimulates enzyme activity. PDE2 is found widely in the brain and plays a major role in hydrolyzing NO-evoked cGMP responses in regions such as the cortex, hippocampus, and striatum. In some peripheral tissues, a rise in cGMP causes a lowering of cAMP levels through PDE2 activation. In the adrenal gland, for example, engagement of this mechanism results in a decrease in aldosterone secretion. However, there is no clear evidence for a similar cross-talk mediated by PDE2 in the brain.

The final category of known receptors for cGMP are members of the family of cyclic nucleotide-gated/modulated ion channels. Cyclic nucleotide-gated (CNG) channels are heteromeric proteins that are directly opened in response to the binding of cAMP or cGMP. They were first discovered in retinal rod photoreceptors and then in olfactory receptor cells and retinal cone cells, but there is now evidence for a more widespread distribution, both within the brain and outside. Molecular cloning has identified two subfamilies of subunits, with one (CNGA1–4) being the core subunits (with all except CNGA4 being channel-forming subunits) and the other (CNGB1 and -B3) having a modulatory effect on ion selectivity, ligand sensitivity, and gating properties. cGMP is the preferred agonist for rod- and cone-type channels, whereas olfactory-type channels are relatively nonselective. Under physiological conditions, CNG channels carry inward currents of sodium and calcium ions and are relatively voltage independent. Unfortunately, the pharmacology of CNG channels is poorly developed, making studies of their participation in NO signal transduction difficult. Nevertheless, there is evidence that in response to activation of the NO–cGMP pathway, the channels can promote neurotransmitter release from presynaptic terminals and increase neuronal firing postsynaptically.

Related structurally to CNG channels are the hyperpolarization-activated cyclic nucleotide-modulated (HCN) channels. As the name suggests, HCN channels are activated at more hyperpolarized membrane potentials (typically negative to approximately -60 mV) and convey a depolarizing, cationic current, known as I_h, largely carried by sodium ions. The channels have an important role in cells that fire rhythmically, such as thalamocortical neurons. Here, the hyperpolarization that follows action potential firing activates HCN channels, which provides a depolarizing drive leading to the activation of other voltage-dependent conductances and further action potential generation. This hyperpolarization-dampening role is also important in setting the resting membrane potential of neurons. Two of the four known HCN channel isoforms (HCN2 and HCN4) are particularly sensitive to regulation by cyclic nucleotides, which shift activation to more depolarized membrane potentials, speeding up rhythmic firing or making the resting membrane potential more depolarized. Being tenfold more potent, cAMP is generally regarded as the natural ligand, but there are several examples in which the NO–cGMP pathway brings about changes in neuronal excitability using HCN channels.

What Does NO Do in the Central Nervous System?

NO has been implicated in numerous behaviors in mammals, including learning and memory formation, feeding, sleeping, reproductive behavior, aggression, and anxiety. Interestingly, some of these roles appear conserved across many millions of years of evolution. For example, the jellyfish *Aglantha digitale* (phylum Cnidaria) first appeared approximately half a billion

years ago and contains a very primitive nervous system consisting of a nerve net, photoreceptors, and mechanosensitive nerve cells. NOS-containing neurons are found in the tentacles and NO plays a key role in swimming behavior associated with feeding, with cGMP being the effector molecule. The NO–cGMP pathway is also important in the feeding behavior of mollusks and honeybees, in which it contributes to the formation of long-lasting associations between scent and food reward. In mammals, NO regulates both food intake and the passage of food through the intestine.

At a more cellular level, NO and cGMP have frequently been found to participate in synaptic plasticity, which refers to the capability of synapses to adjust their strength enduringly upward or downward in response to brief periods of altered input, and which is commonly regarded as a cellular correlate of learning and memory formation. Long-term potentiation in the hippocampus, cerebral cortex, cerebellum, amygdala, and spinal cord has been reported to involve the NO–cGMP pathway, as has long-term depression (LTD) in the cerebellum and striatum. The precise roles of NO remain to be elucidated. In some cases, it appears to act as a retrograde transsynaptic messenger, conveying information about postsynaptic NMDA receptor activity to the presynaptic terminal and influencing neurotransmitter release. In others (e.g., in cerebellar LTD), it appears to be formed presynaptically and act on the postsynaptic neuron. Later phases of the plasticity may require cGMP-dependent alterations in gene expression.

In addition to its abundant physiological roles, NO is important in brain pathology. The emphasis of much research has been on the damaging effect of NO, which is usually considered to be caused by the reaction between NO and superoxide ions, forming peroxynitrite. Peroxynitrite can react with proteins, lipids, or DNA and also break down to yield the noxious hydroxyl and nitrogen dioxide radicals. An alternative mechanism contributing to damage by NO is inhibition of mitochondrial respiration. This occurs in competition with oxygen at the level of cytochrome c oxidase, the terminal complex of the respiratory chain. There has been a shift toward the appreciation of the important protective roles of NO. Protection by NO may be multifaceted. First, NO is a very potent inhibitor of lipid peroxidation which causes membrane damage in many neuropathological conditions. It does so by reacting directly with lipid peroxyl radicals, preventing propagation of the oxidative damage. Second, NO release from endothelial cells is important in maintaining blood flow and preventing adhesion of platelets and leukocytes to vessel walls. Third, NO exerts antiapoptotic effects on many different cell types through multiple mechanisms, including the inhibition of caspases, the release of neurotrophic factors, and the stimulation of the expression of cytoprotective genes. Finally, brain repair may be augmented by the NO–cGMP pathway enhancing the generation of new neurons (neurogenesis) and the formation of new blood vessels (angiogenesis).

See also: Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; NMDA Receptors and Development; NMDA Receptors and Disease + C464.

Further Reading

Alderton WK, Cooper CE, and Knowles RG (2001) Nitric oxide synthases: Structure, function and inhibition. *Biochemical Journal* 357: 593–615.

Bender AT and Beavo JA (2006) Cyclic nucleotide phosphodiesterases: Molecular regulation to clinical use. *Pharmacological Reviews* 58: 488–520.

Bicker G (2005) STOP and GO with NO: Nitric oxide as a regulator of cell motility in simple brains. *Bioessays* 27: 495–505.

Craven KB and Zagotta WN (2006) CNG and HCN channels: Two peas, one pod. *Annual Review of Physiology* 68: 375–401.

Duncan AJ and Heales SJ (2005) Nitric oxide and neurological disorders. *Molecular Aspects of Medicine* 26: 67–96.

Garthwaite G, Bartus K, Malcolm D, et al. (2006) Signaling from blood vessels to CNS axons through nitric oxide. *Journal of Neuroscience* 26: 7730–7740.

Garthwaite J (2005) Dynamics of cellular NO-cGMP signaling. *Frontiers in Bioscience* 10: 1868–1880.

Gibbs SM (2003) Regulation of neuronal proliferation and differentiation by nitric oxide. *Molecular Neurobiology* 27: 107–120.

Hofmann F, Feil R, Kleppisch T, and Schlossmann J (2006) Function of cGMP-dependent protein kinases as revealed by gene deletion. *Physiological Reviews* 86: 1–23.

Hopper RA and Garthwaite J (2006) Tonic and phasic nitric oxide signals in hippocampal long-term potentiation. *Journal of Neuroscience* 26: 11513–11521.

Keynes RG and Garthwaite J (2004) Nitric oxide and its role in ischaemic brain injury. *Current Molecular Medicine* 4: 179–191.

Pilz RB and Broderick KE (2005) Role of cyclic GMP in gene regulation. *Frontiers in Bioscience* 10: 1239–1268.

Prast H and Philippu A (2001) Nitric oxide as modulator of neuronal function. *Progress in Neurobiology* 64: 51–68.

Susswein AJ, Katzoff A, Miller N, and Hurwitz I (2004) Nitric oxide and memory. *Neuroscientist* 10: 153–162.

NMDA Receptor Function and Physiological Modulation

K Zito, University of California at Davis, Davis, CA, USA
V Scheuss, Max-Planck-Institute for Neurobiology, Martinsried, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

With key roles in essential brain functions ranging from the basics of excitatory neurotransmission to the complexities of learning and memory, the *N*-methyl-D-aspartate (NMDA) receptor can be considered one of the fundamental neurotransmitter receptors in the brain. Named for its most potent exogenous agonist, the NMDA receptor has been thoroughly characterized via electrophysiological and pharmacological techniques, as well as through molecular manipulations and transgenic knockout strategies. The NMDA receptor belongs to a family of ionotropic receptors for the excitatory amino acid glutamate and is characterized by high affinity for glutamate, a high unitary conductance, high calcium permeability, and a voltage-dependent block by magnesium ions. this article focuses on the basic biophysical properties and physiological functions of NMDA receptors and how these are modulated by various signaling molecules and biochemical cascades under physiological conditions.

Biophysical Properties of NMDA Receptors

During excitatory neurotransmission, presynaptic release of glutamate activates glutamate receptors in the postsynaptic membrane, resulting in the generation of an excitatory postsynaptic potential (EPSP). Contributing to the EPSP are two classes of glutamate receptors, the non-NMDA receptors (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)/kainate receptors) and the NMDA receptors (Figure 1). The rise times for NMDA receptor currents (10–50 ms) are much slower than those of non-NMDA receptors (0.2–0.4 ms). NMDA receptors also deactivate with a slower time course (~50–500 ms vs. ~2 ms for non-NMDA glutamate receptors), much longer than the time course of glutamate in most synaptic clefts (~1.2 ms). Therefore, during synaptic transmission, the non-NMDA glutamate receptors provide rapid depolarization in response to neurotransmitter release, and the NMDA receptor kinetics determine the duration of the synaptic current.

Activation of NMDA receptors by the neurotransmitter glutamate requires glycine as an essential coagonist. Glutamate and glycine molecules bind different subunits of the receptor; two of each are thought to be required for maximum activation of the receptor. Recent experiments have shown that D-serine can also bind at the glycine site and might represent another physiological coagonist. Compared with glutamate receptors of the non-NMDA type, which have low-affinity binding sites for glutamate (EC_{50} ~500 μM), NMDA receptors bind glutamate with high affinity (EC_{50} ~1 μmol l^{-1}). Despite this high affinity for glutamate, NMDA receptors are not saturated during synaptic transmission at synapses of cortical pyramidal neurons.

NMDA receptor activation leads to opening of an ion channel that is selective for cations, resulting in the influx of Na^+ and Ca^{2+} ions and efflux of K^+ ions. Although most glutamate receptors are cation selective, few are permeable to calcium ions. Its exceptional calcium permeability is the first of two key properties of the NMDA receptor that form the basis for its regulatory role in synaptic plasticity. Entry of calcium into the postsynapse via the NMDA receptor (Figure 2) permits coupling of electrical synaptic activity to biochemical signaling via activation of Ca^{2+}-dependent enzymes and downstream signaling pathways. In this way, calcium influx through the NMDA receptor can lead to long-term changes in synaptic strength and other cellular modifications, including alterations in synaptic structure or connectivity.

The NMDA receptor has a high single channel conductance (~30–50 pS) compared with that of other glutamate receptor types (~4–15 pS). The open probability of agonist-bound receptors has been estimated to range between 0.04 and 0.3, and open times can vary from 0.1 to 8 ms. The molecular events that underlie the opening of NMDA receptors are predicted to include two independent agonist-binding steps preceding a single, concerted conformational change that results in channel opening. Channel closing is thought to be controlled both by the unbinding rate of glutamate and by receptor desensitization. In fact, NMDA receptors can enter long-lived desensitized conformations in which glutamate is bound but the channel remains closed.

The second key biophysical property of the NMDA receptor, and the one that bestows its proposed role in learning and memory, is that it is blocked by magnesium ions in a voltage-dependent manner. At resting membrane potential, NMDA receptors are blocked by magnesium ions; however, if excitation by synaptic inputs causes sufficient depolarization of the neuron, the Mg^{2+} block is relieved and those NMDA receptors which have glutamate bound will open (Figure 3).

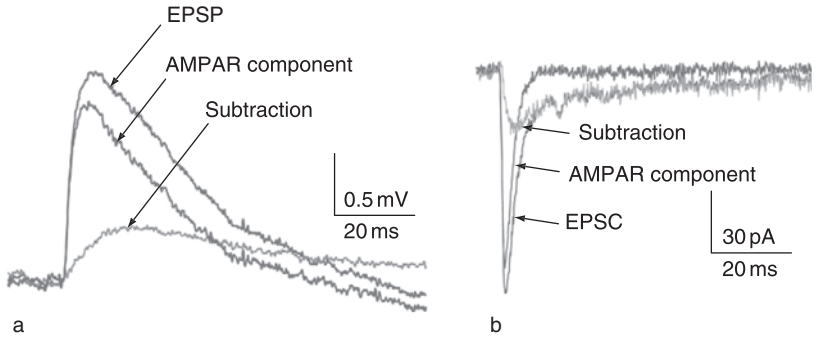

Figure 1 NMDA receptor component of excitatory postsynaptic potentials (EPSPs) and excitatory postsynaptic currents (EPSCs). (a) Unitary EPSP ('EPSC') recorded in a L2/3 cortical pyramidal neuron. The NMDA receptor-mediated component (subtraction) was obtained by subtracting the non-NMDA receptor-mediated component (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPAR) component) when NMDA receptors were blocked with the NMDA receptor antagonist AP5. (b) Unitary EPSC ('EPSC') recorded in a L2/3 cortical pyramidal neuron. The NMDA receptor-mediated component (subtraction) and the non-NMDA receptor mediated component (AMPAR component) were determined as in (a). Reproduced from Blackwell Publishing: figure 6A and B in Feldmeyer D, Lübke J, Silver RA, et al. (2002) Synaptic connections between layer 4 spiny neuron-layer 2/3 pyramidal cell pairs in juvenile rat barrel cortex: Physiology and anatomy of interlaminar signaling within a cortical column. *Journal of Physiology* 538(3): 803–822.

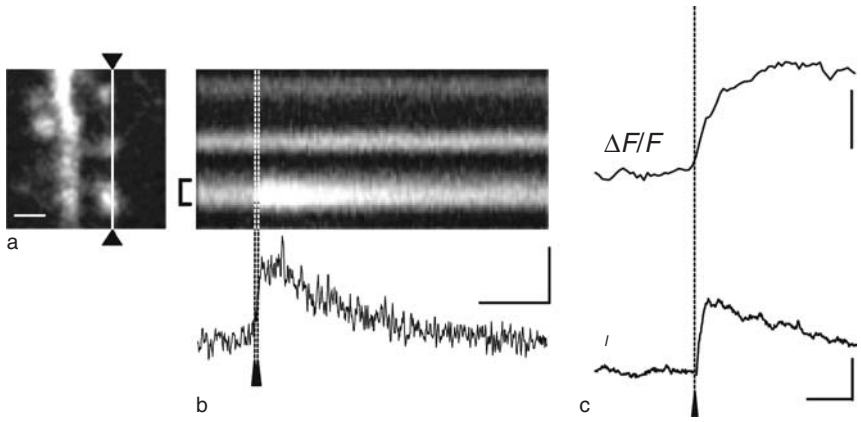

Figure 2 Synaptic activation of NMDA receptors causes calcium entry into spines. (a) Two-photon fluorescence image of a dendritic branch of a hippocampal CA1 pyramidal neuron loaded with the calcium indicator Oregon Green BAPTA. Scale bar = 1 μm. (b) Sequences of line scans in which the vertical dimension corresponds to the line indicated in (a) and the horizontal dimension represents time. Below: Relative change in fluorescence ($\Delta F/F(t)$) at the bottom spine (averaged over window indicated by bracket) during calcium entry evoked by synaptic stimulation (time indicated by arrowhead). The neuron was voltage clamped at positive potential to isolate NMDA receptor-mediated Ca^{2+} influx. Scale bars = 250 ms, 50% $\Delta F/F_{max}$. (c) Average EPSC (bottom) and corresponding $\Delta F/F(t)$ (top) at the same spine as in (b). Scale bars = 25% $\Delta F/F_{max}$, 100 pA, 25 ms. Reprinted by permission from MacMillan Publishers Ltd.: *Nature* (Mainen ZF, Malinow R, and Svoboda K (1999) Synaptic calcium transients in single spines indicate that NMDA receptors are not saturated. *Nature* 399: 151–155) copyright 1999.

Since opening of the NMDA receptor requires simultaneous activation by glutamate and depolarization to relieve the magnesium block, the NMDA receptor can act as a coincidence detector for pre- and postsynaptic activity.

Glutamate and glycine affinities, calcium permeability, ion conductance, channel kinetics, and sensitivity to Mg^{2+} block of the NMDA receptor are all determined in part molecularly, by alternative splicing and subunit composition. Most of these biophysical properties of NMDA receptors can also be modulated by various ionic and molecular interactions and signaling pathways (described below).

NMDA Receptor Physiological Function

The outstanding physiological function of NMDA receptors is the coupling of electrical to biochemical signaling in neurons by mediating calcium influx in response to synaptic activity. However, NMDA receptors also serve important functions in electrical neurotransmission alone. Despite the voltage-dependent magnesium block, NMDA receptors can contribute significantly to the amplitude of unitary evoked postsynaptic potentials. The slow kinetics of NMDA receptor-mediated excitatory postsynaptic currents (EPSCs) facilitates temporal summation and reduces

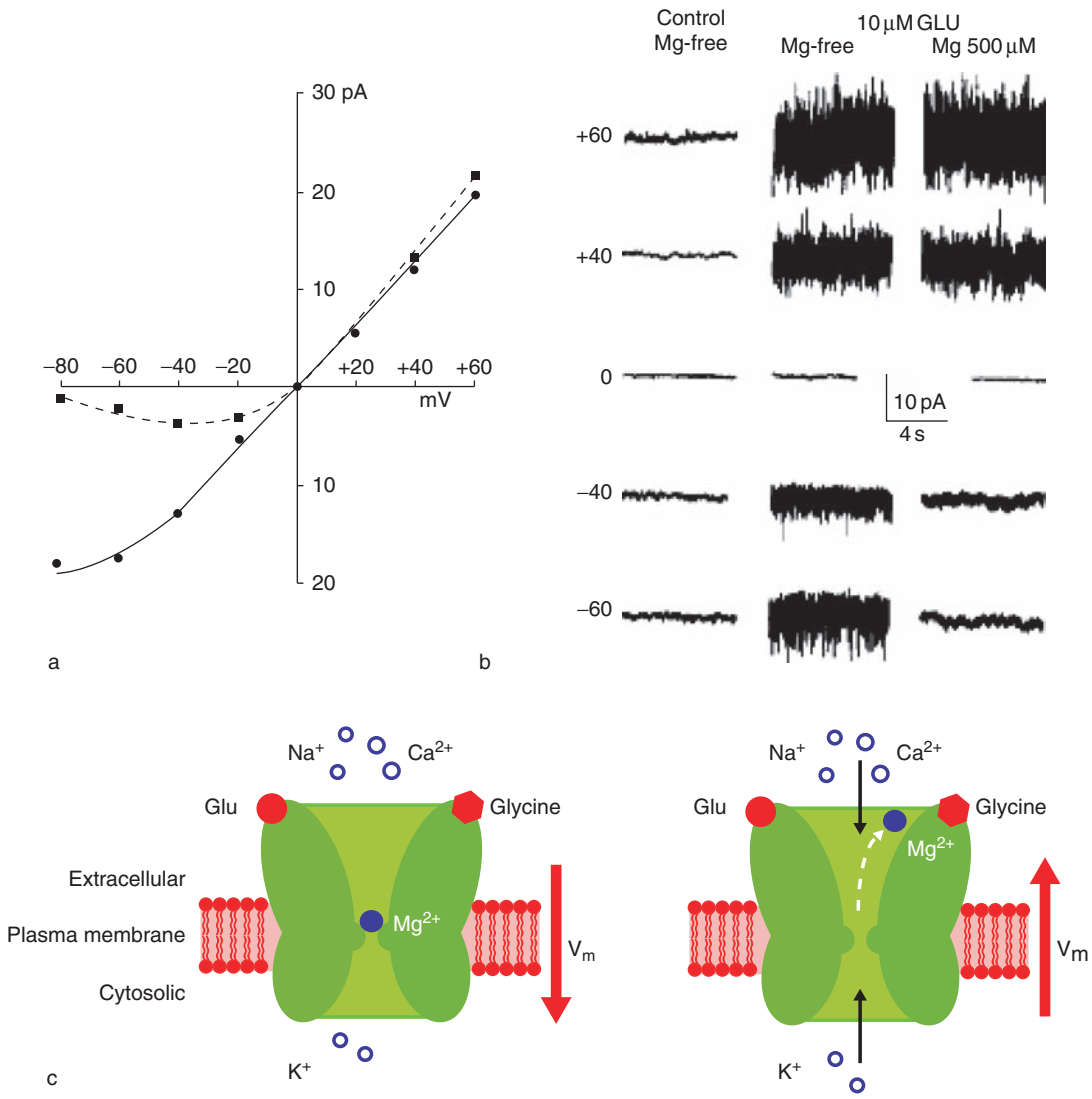

Figure 3 Magnesium block of NMDA receptors. (a) Voltage dependence of glutamate-induced currents in Mg^{2+}-free (circles) and Mg^{2+}-containing solution (500 μM; squares). (b) Voltage and Mg^{2+}-dependence of NMDA receptor current noise. Reprinted by permission from MacMillan Publishers Ltd: *Nature* (Nowak L, Bregestovski P, Ascher P, et al. (1984) Magnesium gates glutamate-activated channels in mouse central neurons. *Nature* 307: 462–465) copyright 1984. (c) Schematic of the mechanism of the magnesium block. At resting potential, the pore of the NMDA receptor channel is blocked by magnesium ions. Upon depolarization, the magnesium ions are removed from the pore, and the channel can pass current.

dendritic filtering of synaptic inputs. In addition, coactivation of multiple synapses can trigger NMDA receptor-dependent dendritic spikes by generating sufficient depolarization to overcome the magnesium block. Such dendritic spikes cause nonlinear synaptic integration, which enhances the computational power of neurons and may allow neurons to detect the synchrony of inputs (Figure 4).

The NMDA receptor plays an essential role in brain plasticity, that is, the ability of the brain to change in response to external stimuli, such as during learning and memory. First investigated pharmacologically, chronic blockade of the NMDA receptor by infusion of an antagonist into the ventricles of rats was shown to impair spatial learning. Since that time, similar behavioral experiments have been performed using knockout mice lacking NMDA receptor subunits from specific subregions of the hippocampus. These experiments have repeatedly provided evidence for the importance of the NMDA receptor in learning and memory processes.

The role of the NMDA receptor in learning and memory is attributed to its ability to regulate excitatory synaptic transmission. The NMDA receptor has been shown to play an essential role in both the strengthening of synapses, through long-term

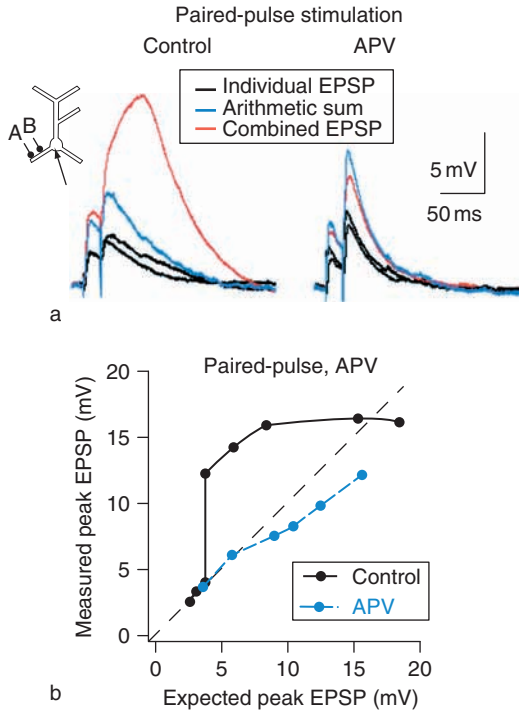

Figure 4 NMDA receptors in nonlinear synaptic integration. (a) Somatic voltage responses of a layer 5 pyramidal neuron to stimuli from two electrodes (A and B) placed 30 μm apart on a single dendritic branch display nonlinear summation (left panel). With application of the NMDA receptor antagonist, APV, the summation is linear. EPSP, excitatory postsynaptic potential. (b) Expected vs. actual peak somatic responses plotted for different stimulus intensities before (black) and after (blue) application of APV. NMDA receptor blockade linearizes synaptic summation. Reprinted by permission from MacMillan Publishers Ltd: *Nature Neuroscience* (Polsky A, Mel BW, and Schiller J (1994) Computational subunits in thin dendrites of pyramidal cells. *Nature Neuroscience* 7: 621–627), copyright 1994.

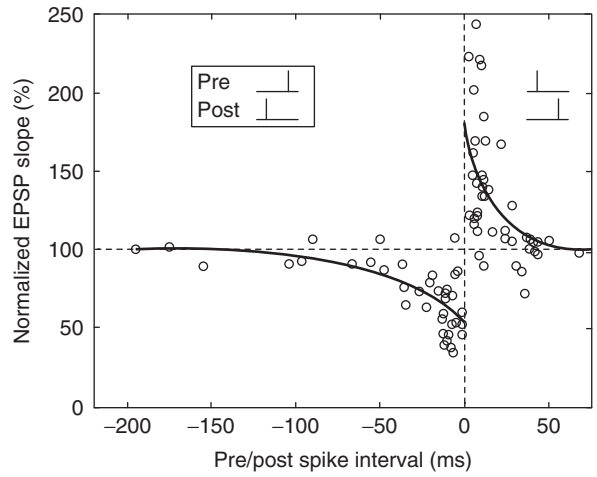

Figure 5 Spike-timing-dependent synaptic plasticity. Change in excitatory postsynaptic potential (EPSP) slope induced by repetitively paired pre- and postsynaptic spikes in layer 2/3 of rat visual cortex. Insets depict the sequence of spiking. Pre before postsynaptic spiking leads to potentiation while post before presynaptic spiking leads to depression. Used from The American Physiological Society: figure 1 in Dan Y and Poo MM (2006) Spike timing-dependent plasticity: From synapse to perception. *Physiological Reviews* 86: 1033–1048.

potentiation (LTP), and the weakening of synapses, through long-term depression (LTD). LTP and LTD are proposed to be cellular mechanisms underlying learning and memory. Via its role as a coincidence detector, the NMDA receptor is capable of signaling coincident pre- and postsynaptic activity. Detailed studies of the relationship between the temporal correlation of pre- and postsynaptic activity and the resulting type of plasticity have led to the discovery of spike-timing dependent plasticity. If the presynaptic action potential (AP; and thus synaptic activity) repetitively precedes the generation of the postsynaptic AP by less than ~50 ms, the synapse is potentiated; however, if the postsynaptic AP occurs within ~50 ms before the presynaptic AP (and thus independently of presynaptic activity), the synapse is depressed (Figure 5). With respect to the more complex and irregular patterns of neuronal activity that are likely found *in vivo*, it has been suggested that both rate and timing determine the direction of neuronal plasticity.

The NMDA receptor also acts as a coincidence detector during development, where it plays a critical role in the maturation of synapses and the activity-dependent establishment of topographic maps in the brain. By acting to strengthen neighboring connections of similar activity patterns, the NMDA receptor enforces the principle that 'cells that fire together, wire together.' In addition, early in development, many excitatory synapses appear to contain only NMDA receptors, and therefore these synapses are thought to be at first silent (no EPSP in response to synaptic stimulation). It is hypothesized that only through coincident activity sufficiently robust to depolarize the postsynaptic cell and relieve the voltage-dependent Mg^{2+} block of the NMDA receptors does the insertion of non-NMDA receptors at these synapses occur.

While presynaptic NMDA receptors have been detected by immuno-electron microscopy, their function remains far less established compared with their postsynaptic counterparts. As autoreceptors, presynaptic NMDA receptors have been shown to modulate synaptic transmission in the spinal cord, cerebellum, and entorhinal cortex and to play a role in cerebellar and cortical LTD.

Calcium influx through the NMDA receptor is thought to be responsible for its roles in LTP and LTD and for both neuroprotective and neurotoxic effects. Although these various effects may at first seem contradictory, the explanation lies in the spatial

and temporal patterns of calcium influx. Calcium influx at synaptic sites is thought to activate signaling pathways that lead to LTP or LTD, depending on the temporal pattern and amplitude of the calcium transient. Synaptic calcium influx also leads to activation of cyclic adenosine monophosphate response element binding protein (CREB), a transcription factor, resulting in activation of gene expression. One of the genes upregulated in response to activated CREB is brain-derived neurotrophic factor (BDNF), which acts to promote prosurvival programs. In contrast, calcium influx through extrasynaptic NMDA receptors is thought to lead to CREB deactivation and inhibition of BDNF expression, which can have deleterious effects on cell health. In fact, overstimulation of NMDA receptors results in neuronal excitotoxicity, which has been implicated in the loss of neurons associated with ischemic stroke; Alzheimer's, Parkinson's, and Huntington's disease; and amyotrophic lateral sclerosis.

Physiological Modulation of NMDA Receptors

Although glycine and/or D-serine are essential co-agonists of the NMDA receptor, and thus, strictly speaking, not modulators, their cytosolic levels nevertheless regulate NMDA receptor activation. One of two amino acids appears to predominate, dependent on the brain region. The cytosolic concentrations are controlled by release and uptake by both neurons and glia. Stimulated release of glycine and D-serine can be induced by depolarization or non-NMDA glutamate receptor activation. In the spinal cord, glycine spillover from inhibitory synapses has been shown to enhance NMDA receptor currents.

Desensitization

As the NMDA receptor is a ligand-gated ion channel, it can display a decrease in conductance despite the continuous presence of agonist, a process referred to as desensitization. Three different types of desensitization have been reported for NMDA receptors, a glycine-sensitive, a glycine-insensitive, and a calcium-dependent type. The glycine-sensitive desensitization refers to the transition of the NMDA receptor into a glutamate-bound closed state, reflecting a negative allosteric interaction between the glutamate and the glycine binding sites, which results in a reduced glycine affinity in the presence of high glutamate concentrations. High glycine concentrations can overcome this type of desensitization. Glycine-insensitive desensitization has been observed only in dialyzed cells and excised membrane patches and thus does not appear to be physiologically relevant. The calcium-dependent desensitization (also referred to as calcium-dependent inactivation) provides a negative feedback loop by which calcium entering the cell via NMDA receptors in turn leads to the desensitization of the receptor (Figure 6), although calcium from other sources (voltage-gated calcium channels or release from intracellular stores activated by second messenger cascades) has the same effect. In a current model, calcium influx activates calmodulin to displace α-actinin2 from the NMDA receptor. The resulting dissociation of the NMDA receptor from the cytoskeleton is proposed to cause a conformational change in the receptor, which reduces its open probability.

Endogenous Allosteric Modulators

NMDA receptor function is fine-tuned by various forms of allosteric modulation involving endogenous extracellular substances such as zinc ions, protons, polyamines,

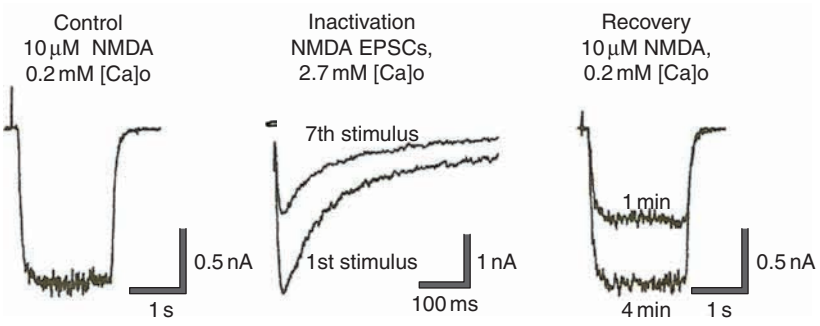

Figure 6 Calcium-dependent desensitization of NMDA receptors. Whole cell currents of cultured rat hippocampal neurons (left and right panels) were evoked in low calcium (0.5 μM) solutions by application of NMDA. During a train of seven excitatory postsynaptic currents (EPSCs) evoked in 2.7 μM calcium, the NMDA receptor mediated EPSC was inactivated to ~50% of control (middle). The whole cell current was inactivated to a similar degree and recovered within 4 min. Used from The American Physiological Society: figure 3 in Rosenmund C, Feltz A, and Westbrook GL (1995) Calcium-dependent inactivation of synaptic NMDA receptors in hippocampal neurons. *Journal of Neurophysiology* 73: 427–430.

and reducing and oxidizing agents (Figure 7). Zinc and polyamines mediate a voltage-dependent block of NMDA receptors, which is weaker but appears to involve the same intrapore residues as block by magnesium. Some of the voltage-independent effects of allosteric modulators are mutually interdependent, and structural evidence suggests convergence by overlapping binding sites or common downstream structural modifications. Zinc, protons, and oxidizing agents inhibit NMDA receptor function, while polyamines and reducing agents cause potentiation by modifying channel open frequency and time or agonist or modulator affinities. These modulations are considered to be physiologically relevant during synaptic transmission, when synaptic activity can cause shifts in pH, co-release of Zn^{2+} occurs at certain synaptic terminals (e.g., mossy fiber-CA3), and the amine histamine is released from modulatory afferents (e.g., those arising from the anterior hypothalamus). An important aspect of the negative regulation of NMDA receptor function by pH and oxidizing agents is that it can limit or delay cell damage under pathological conditions such as stroke and ischemia.

Phosphorylation

NMDA receptors can be phosphorylated by the serine/threonine kinases protein kinase C (PKC), protein kinase A (PKA) and calcium/calmodulin-dependent protein kinase II (CaMKII), as well as by the tyrosine kinases Src and Fyn. In general, phosphorylation enhances NMDA receptor function (Figure 7).

A significant percentage of NMDA receptor subunits in the brain are estimated to be phosphorylated by PKC or PKA at one or more sites (10–70%, depending on the region of the brain). Phosphorylation by PKC reduces the affinity for extracellular magnesium ions and increases the open probability. Calcium influx through the NMDA receptor itself can enhance the potentiation mediated by PKC. However, in some preparations, phosphorylation by PKC reduced NMDA receptor-mediated responses by preventing NMDA receptor subunit clustering. Less is known about the regulation of NMDA receptors by PKA and CaMKII. Active CaMKII directly associates with the NMDA receptor, and this association is required for some forms of activity-driven synaptic potentiation. PKA activation has been shown to increase the fractional Ca^{2+} influx through the NMDA receptor; however, some of the effects of PKA activation on NMDA function appear to be indirect.

In contrast, fewer NMDA receptor subunits on neural membranes are phosphorylated on tyrosine residues (2–4%). Phosphorylation by Src enhances NMDA receptor function (Figure 8) by reducing the potency of Zn^{2+} block of recombinant NMDA receptors expressed heterologously; however, evidence is lacking for this mechanism in hippocampal or spinal cord neurons. An alternative view proposes that phosphorylation of NMDA receptor subunits by Src-family kinases could affect downstream signaling proteins, either by recruiting them to the NMDA receptor, by

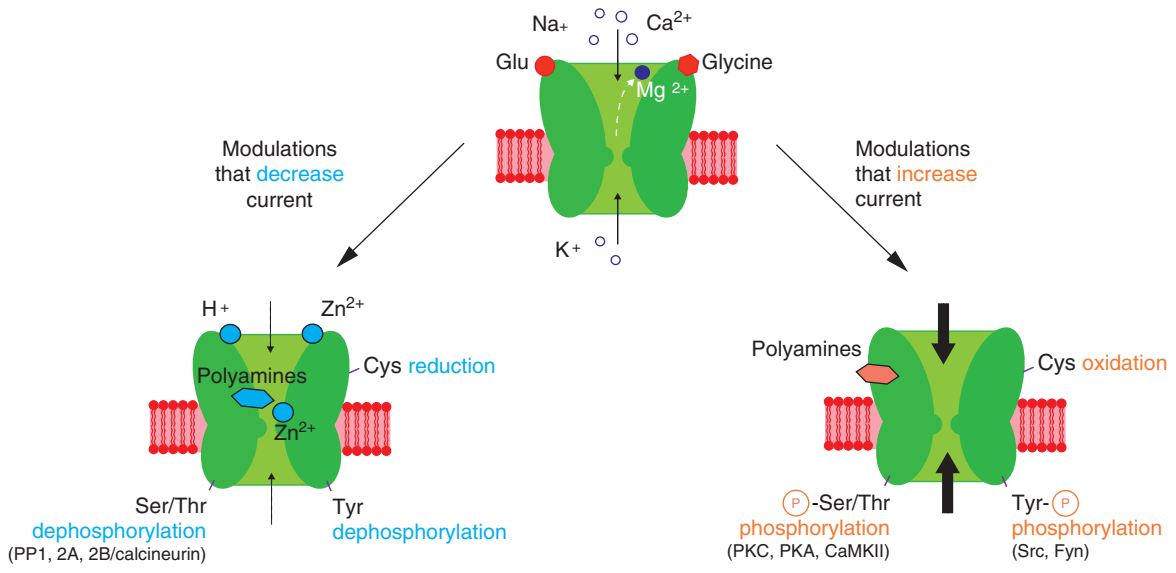

Figure 7 Schematic overview of modulators of NMDA receptor functions. Activation of the NMDA receptor requires binding of glutamate (Glu) and glycine (or serine) together with membrane depolarization to release the magnesium block (center). Modulators causing depression of NMDA receptor currents (left) include H^+, Zn^{2+}, polyamines, reducing agents, and the protein phosphatases PP1, PP2A, and PP2B/calcineurin. Polyamines and Zn^{2+} cause a voltage-dependent block similar to that caused by Mg^{2+}. Modulators causing potentiation of NMDA receptor currents (right) include external polyamines, oxidizing agents, and the protein kinases PKC, PKA, CaMKII, Src, and Fyn. For details see text.

protecting them against degradation, or through blocking the assembly of signaling complexes. Src-mediated phosphorylation appears to play a role in LTP induction and seems to be modulated by BDNF. Phosphorylation by Fyn may target the NMDA receptor to the plasma membrane by antagonizing its interaction with spectrin.

In some cases, Src activity is required for PKC-mediated activation of NMDA receptor currents, placing Src downstream of PKC on the same pathway. PKC stimulation leads to the activation of the nonreceptor protein tyrosine kinase, cell adhesion kinase-β, which activates Src by disrupting the intramolecular interactions that maintain Src in a low-activity state. Activation of Src then leads to the potentiation of NMDA receptor currents. The effects of Src on NMDA receptors can be modulated by H-Ras, a small guanosine triphosphate (GTP)-binding protein, which binds Src and inhibits its activity. Hippocampal pyramidal neurons from H-ras homozygous null mice displayed enhanced NMDA receptor synaptic responses.

The activation of NMDA receptors by phosphorylation can be reversed by serine and threonine phosphatases 1, 2A, and 2B (calcineurin) and endogenous tyrosine phosphatases. Inhibition of endogenous protein tyrosine phosphatase activity led to potentiation of NMDA receptor currents, indicating that these phosphatases participate in determining the basal phosphorylation level of the NMDA receptor.

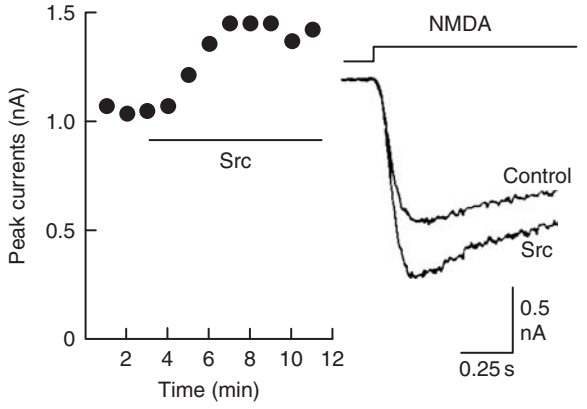

Figure 8 The protein tyrosine kinase pp60$^{c\text{-}src}$ potentiates NMDA currents. Whole cell recordings were made from mouse cultured hippocampal neurons. Peak currents evoked by rapid application of NMDA (100 μM) are plotted in the graph on the left. The tyrosine kinase pp60$^{c\text{-}src}$ (30 U per milliliter) was actively perfused through the recording electrode. The period of the perfusion is indicated by the horizontal bar on the graph. On the right, representative currents recorded before (Control) and after (Src) intracellular perfusion with pp60$^{c\text{-}src}$ are shown. NMDA was applied as indicated by the trace above the currents. Reprinted by permission from MacMillan Publishers Ltd: *Nature* (Wang YT and Salter MW (1994) Regulation of NMDA receptors by tyrosine kinases and phosphatases. *Nature* 369: 233–235), copyright 1994.

Calcineurin can be activated by calcium influx through the NMDA receptor itself.

Modulatory Signaling Receptors

Modulation of NMDA receptors by kinases and phosphatases occurs downstream of a number of membrane receptors, the largest group of which are the G-protein-coupled receptors (GPCRs). GPCRs are integral membrane proteins which transmit extracellular signals to the cytoplasm by coupling to heterotrimeric GTP-binding proteins. Activation of a subset of GPCRs, including metabotropic glutamate receptors (mGluRs), muscarinic acetylcholine receptors, μ opioid receptors, lysophosphatidic acid (LPA) receptors, and the protease-activated receptor PAR1, enhances NMDA receptor currents. Those GPCRs that are coupled to Gq proteins, such as mGluR5, m1 muscarinic receptor, and the LPA receptor, are thought to act through a PKC-Src signaling pathway, and mGluR1 also activates Src, but through a PKC-independent pathway. A second set of GPCRs, those coupled through Gs, are thought to enhance activity of the NMDA receptor through activation of Src-family kinases via PKA and receptor for activated C kinase 1 (RACK1). One such GPCR, the pituitary adenylate cyclase activating polypeptide (PACAP) receptor, is activated by the neuropeptide PACAP(1–38), which has been shown to potentiate NMDA receptor-mediated excitatory postsynaptic field potentials in hippocampal slices. Finally, NMDA receptors have been shown to interact directly with G-protein-coupled dopamine receptors. Depending on the class of dopamine receptor, activation of dopamine receptors has been reported both to enhance and to inhibit NMDA receptor currents.

NMDA receptor currents are also influenced by downstream signaling from receptor protein tyrosine kinases, including the EphB receptors, platelet-derived growth factor (PDGF) receptors, and insulin receptors. EphB receptors interact directly with NMDA receptors in cultured neurons, and their activation by ephrinB2 increases NMDA receptor-dependent calcium responses via the activation of Src. Insulin has been shown to enhance NMDA receptor activity in hippocampal slices via a mechanism that requires PKC and tyrosine kinase activity. In contrast, application of PDGF depresses NMDA receptor currents in hippocampal pyramidal neurons via a phospholipase C-IP_3 receptor-cyclic AMP-PKA pathway that inhibits Src activity.

Cytokine receptor activation has been shown to enhance NMDA receptor currents. One group of cytokines, leptins, enhances NMDA receptor-mediated EPSCs in hippocampal slices via activation of the leptin receptor Ob-R_b. The leptin-mediated potentiation was

reduced by inhibitors of PI3-kinase, mitogen-activated protein kinase, and Src-family kinases. Another cytokine, interleukin (IL)-1β, enhanced the rise in intracellular calcium following application of NMDA to cultured hippocampal neurons, a response mediated by the IL-1RI receptor.

NMDA Receptor Complex

The ability of modulators to influence NMDA receptor currents is dependent on their abundance and proximity to the receptor. Proteomic characterization identified synaptic NMDA receptors as part of a remarkably large macromolecular signaling complex called the NMDA receptor complex. NMDA receptors were found linked to receptors, adhesion molecules, scaffolding proteins, signaling molecules, cytoskeletal proteins, and various novel proteins, in complexes lacking non-NMDA receptors. Among the signaling proteins were kinases, phosphatases, GTPases, and GTPase-activating proteins. Physical linkage of these receptors, signaling molecules, and the cytoskeleton to the NMDA receptor is an important way to facilitate rapid modulation of the receptor in response to synaptic activity.

See also: D-Serine: From its Synthesis in Glial Cell to its Action on Synaptic Transmission and Plasticity; Kainate Receptor Functions; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; NMDA Receptors and Development; NMDA Receptors and Disease + C464; NMDA Receptors, Cell Biology and Trafficking.

Further Reading

Cull-Candy SG and Leszkiewicz DN (2004) Role of distinct NMDA receptor subtypes at central synapses. *Science's STKE: Signal Transduction Knowledge Environment* 255: re16.

Dan Y and Poo MM (2006) Spike timing-dependent plasticity: From synapse to perception. *Physiological Reviews* 86: 1033–1048.

Dingledine R, Borges K, Bowie D, et al. (1999) The glutamate receptor ion channels. *Pharmacological Reviews* 51: 7–61.

Edmonds B, Gibb AJ, and Colquhoun D (1995) Mechanisms of activation of glutamate receptors and the time course of excitatory synaptic currents. *Annual Review of Physiology* 57: 495–519.

Feldmeyer D, Lübke J, Silver RA, et al. (2002) Synaptic connections between layer 4 spiny neuron-layer 2/3 pyramidal cell pairs in juvenile rat barrel cortex: Physiology and anatomy of interlaminar signaling within a cortical column. *Journal of Physiology* 538(3): 803–822.

Husi H, Ward MA, Choudhary JS, et al. (2000) Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. *Nature Neuroscience* 3: 661–669.

Kotecha SA and MacDonald JF (2003) Signaling molecules and receptor transduction cascades that regulate NMDA receptor-mediated synaptic transmission. *International Review of Neurobiology* 54: 51–106.

Lester RA and Jahr CE (1992) NMDA channel behavior depends on agonist affinity. *Journal of Neuroscience* 12: 635–643.

Mainen ZF, Malinow R, and Svoboda K (1999) Synaptic calcium transients in single spines indicate that NMDA receptors are not saturated. *Nature* 399: 151–155.

Nowak L, Bregestovski P, Ascher P, et al. (1984) Magnesium gates glutamate-activated channels in mouse central neurons. *Nature* 307: 462–465.

Polsky A, Mel BW, and Schiller J (2004) Computational subunits in thin dendrites of pyramidal cells. *Nature Neuroscience* 7: 621–627.

Rosenmund C, Feltz A, and Westbrook GL (1995) Calcium-dependent inactivation of synaptic NMDA receptors in hippocampal neurons. *Journal of Neurophysiology* 73: 427–430.

Salter MW and Kalia LV (2004) Src kinases: A hub for NMDA receptor regulation. *Nature Reviews Neuroscience* 5: 317–328.

Schneggenburger R, Zhou Z, Konnerth A, et al. (1993) Fractional contribution of calcium to the cation current through glutamate receptor channels. *Neuron* 11: 133–143.

Vanhoutte P and Bading H (2003) Opposing roles of synaptic and extrasynaptic NMDA receptors in neuronal calcium signalling and BDNF gene regulation. *Current Opinion in Neurobiology* 13: 366–371.

Wang YT and Salter MW Regulation of NMDA receptors by tyrosine kinases and phosphatases. *Nature* 369: 233–235.

NMDA Receptors and Development

M A Phillips and M Constantine-Paton, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Developmental Expression of Diverse NMDAR Subunits

The *N*-methyl-D-aspartate (NMDA) receptors (NMDARs) are tetrameric ion channels that are ligand-gated by glutamate and glycine and voltage-gated by postsynaptic depolarization. These properties make it an excellent coincidence detector of simultaneous presynaptic glutamate release and postsynaptic depolarization. The main pore-forming units of the NMDAR convey on it different kinetic and conducting properties that are changed through a maturational program of gene expression and synaptic activity. The channel pore is cation-selective and highly permeable to calcium, an important second messenger and the most intensely studied player in investigations of NMDAR function. Compared to other glutamate receptors, the time course of NMDAR activation and deactivation is relatively slow. These properties make the receptor a poor candidate for mediating high-frequency spike transmission but an excellent candidate for mediating long-term structural changes in response to the recent activity history of inputs to particular synapses.

Most NMDARs contain two obligate NR1 subunits, and any two of the four NR2 subunits, NR2A–2D. There are up to eight known splice variants of the NR1 subunit that are expressed heterogeneously in the brain and during development; however, virtually every brain region expresses at least one NR1 subunit by E19. Genes encoding the NR2 family of subunits can be regulated by a developmental program, activity, or both. Although detailed expression patterns of individual subunits vary greatly, consistent patterns have emerged (Figure 1). In most regions, NR2B and NR2D are expressed at highest levels in the embryo and early postnatal period, and they yield receptor currents with slow deactivation kinetics (ratios of deactivation time constants *in vitro* for NR2A : 2B : 2C : 2D are 1 : 3 : 3 : 40). NR2B expression usually persists in the adult brain, although at somewhat lower levels, whereas NR2D subunits are virtually eliminated. NR2A and NR2C are present in mature receptors, with NR2C expression largely confined to the cerebellum and thalamus.

The fact that immature NMDARs in most parts of the brain have much slower deactivation kinetics than mature NMDARs is thought to have important implications for receptor function in the young brain, where lower frequency activity patterns predominate. These patterns tend to correlate pre- and postsynaptic firing on a slow timescale (hundreds of milliseconds). Thus, a receptor with a correspondingly slow deactivation permits coincidence detection even at low frequencies. Additional mechanisms appear to exist in the immature brain to facilitate NMDAR activation because immature receptors in some regions lack voltage-dependent gating *in vivo*, a property that may arise from the presence of NR2D subunits. This could potentially eliminate coincidence detection by the NMDAR, and could contribute to the different effects of NMDAR activation observed in the immature brain, compared to the mature brain. Faster kinetics of receptors containing NR2A subunits could allow tighter temporal control in the detection of coincident activity in mature animals, where high-frequency activity predominates.

Two unusual, recently cloned subunits, NR3A and NR3B, are thought to be important in dampening NR2-containing NMDAR currents. They inhibit channel opening and calcium flux in response to glutamate when co-assembled with NR2 subunits both *in vitro* and *in vivo*. Whereas NR3A expression is widespread, NR3B subunits are restricted to motor nuclei of the brain stem and spinal cord.

In these regions and others with high fidelity, nonplastic synapses, NR2 and NR3A subunits are embryonically expressed and postnatally downregulated, often to undetectable levels. As they disappear, inhibitory NR3B subunits appear and persist into adulthood, accompanying the effective loss of detectable NMDA currents in these structures. This downregulation over development may be critical for the prevention of activity-dependent rearrangements in critical hardwired circuits, but it does not preclude the possibility that the more plastic, immature forms of the receptor can return when circuit conditions are changed by trauma, disease, pharmacology, or behavior. Thus, the impetus for biomedical scientists to study the mechanisms of NMDAR regulation and the pathways it uses to effect synaptic change remains strong.

NMDARs in Early Brain Development

NMDARs have diverse but poorly understood roles in the immature brain. At embryonic stages, NMDAR

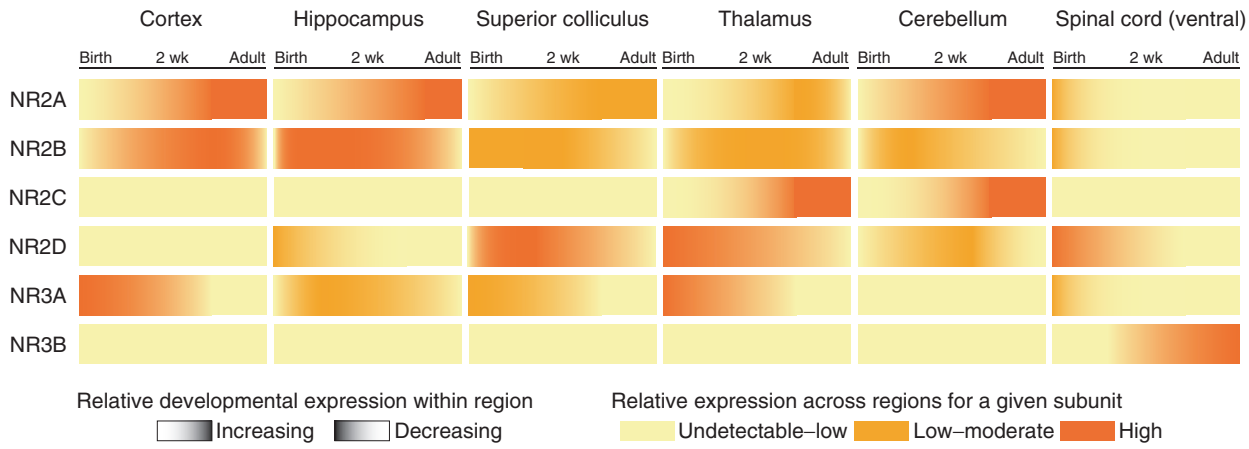

Figure 1 Developmental patterns of NMDAR subunit expression in the rodent brain. Summary of approximate expression levels of each of the known NMDA receptor subunits NR2A–2D and newly cloned subunits NR3A and NR3B over postnatal development. Where possible data from quantitative protein expression studies were used. Note the similar expression patterns of NR2B, NR2D, and NR3A, expressed early and downregulated in approximately the second postnatal week, and NR2A, NR2C, and NR3B, which appear after the first postnatal week. However, there is significant heterogeneity across different brain regions and even within brain regions, where subunits can be expressed in a laminar- and cell-specific manner. In addition, detection methods with lower sensitivity and spatial resolution (e.g., western blotting and histoblots) may not detect weak or sparsely expressed subunits in a given region.Data compiled from Dunah AW, Yasuda RP, Wang Y, et al. (1996) Regional and ontogenic expression of the NMDA receptor subunit NR2D protein in rat brain using a subunit-specific antibody. *Journal of Neurochemistry* 67: 2335–2345; Fagiolini M, Katagiri H, Miyamoto H, et al. (2003) Separable features of visual cortical plasticity revealed by *N*-methyl-D-aspartate receptor 2A signaling. *Proceedings of the National Academy of Sciences of the United States of America* 100: 2854–2859; Fukaya M, Hayashi Y, and Watanabe M (2005) NR2 to NR3B subunit switchover of NMDA receptors in early postnatal motoneurons. *European Journal of Neuroscience* 21: 1432–1436; Monyer H, Burnashev N, Laurie DJ, Sakmann B, and Seeburg PH (1994) Developmental and regional expression in the rat brain and functional properties of four NMDA receptors. *Neuron* 12: 529–540; Nishi M, Hinds H, Lu H, Kawata M, and Hayashi Y (2001) Motoneuron-specific expression of NR3B, a novel NMDA-type glutamate receptor subunit that works in a dominant-negative manner. *Journal of Neuroscience* 21: RC185; Quinlan EM, Olstein DH, and Bear MF (1999) Bidirectional, experience-dependent regulation of *N*-methyl-D-aspartate receptor subunit composition in the rat visual cortex during postnatal development. *Proceedings of the National Academy of Sciences of the United States of America* 96: 12876–12880; Sans N, Petralia RS, Wang Y, Blahos J, Hell JW, and Wenthold RJ (2000) A developmental change in NMDA receptor-associated proteins at hippocampal synapses. *Journal of Neuroscience* 20: 1260–1271; Shi J, Aamodt SM, and Constantine-Paton M (1997) Temporal correlations between functional and molecular changes in NMDA receptors and GABA neurotransmission in the superior colliculus. *Journal of Neuroscience* 17: 6264–6276; Sucher NJ, Akbarian S, Chi CL, et al. (1995) Developmental and regional expression pattern of a novel NMDA receptor-like subunit (NMDAR-L) in the rodent brain. *Journal of Neuroscience* 15: 6509–6520; Wang Y, Bosy TZ, Yasuda RP, et al. (1995) Characteristics of NMDA receptor subunit-specific antibodies: Distribution of MR2A and MR2B receptor subunits in rat brain and ontogenetic profile in the cerebellum. *Journal of Neurochemistry* 65: 176–183; Wenzel A, Villa M, Mohler H, and Benke D (1996) Developmental and regional expression of NMDA receptor subtypes containing the NR2D subunit in rat brain. *Journal of Neurochemistry* 66: 1240–1248; Wenzel A, Fritschy JM, Mohler H, and Benke D (1997). NMDA receptor heterogeneity during postnatal development of the rat brain: Differential expression of the NR2A, NR2B, and NR2C subunit proteins. *Journal of Neurochemistry* 68: 469–478; Wong H, Liu X, Matos MF, et al. (2002) Temporal and regional expression of NMDA receptor subunit NR3A in the mammalian brain. *Journal of Comparative Neurology* 450: 303–317.

function is probably independent of synaptic communication because most glutamatergic synapses form after birth. It is possible that in embryonic structures, ambient glutamate is secreted from glial processes or immature axons in a paracrine manner, and this is how early NMDARs are activated.

Migration

NMDAR activation stimulates neuronal migration in vertebrate embryos. Functional NMDARs are expressed on neurons in the embryonic cortical plate (prior to synapse formation) and on some tangentially migrating neurons during the formation of cortical layers. NMDA agonist stimulates chemotaxis in cultured embryonic cortical neurons. Cerebellar granule cells express functional NMDARs before they have begun migration through the molecular layer and prior to synapse formation with mossy fibers. The migratory movements of these cells are slowed up to 50% by NMDAR antagonists and are not affected by blockade of other glutamate receptors.

Differentiation

NMDARs are also involved in the tuning of local circuitry once the migrating neurons reach their targets. Prior to synapse formation, electrical and chemical signals among groups of neurons are coordinated by gap junction connections between them. In neonatal

mice, cortical neurons are spontaneously active and calcium responses are synchronized among groups of adjacent neurons. Mice deficient in NR2B, and hence most early NMDAR current, have fewer coupled groups of neurons in the cortex. Gap junctions have been investigated closely in the hypothalamus, where NMDAR-mediated activity downregulates the Connexin36 gap junction protein. Both NMDAR blockade and elimination of the NR1 gene extend the period of gap junction coupling in the developing hypothalamus; similar results have also been reported for spinal motor neurons. The data suggest NMDARs might be able to regulate the switch from gap junction-connected cells to synaptically connected circuits in multiple brain regions, and can either speed or slow the rate of that development depending on the region and, potentially, the form of the receptor complex that is expressed.

Sculpting Neural Circuits

After neurons have reached their final destinations and begun to differentiate, they extend axonal and dendritic processes and begin to form synapses. This early neurite extension and synaptogenesis proceeds in the absence of synaptic activity. However, subsequent neuronal survival and the refinement of axonal and dendritic connections are dependent on neuronal activation and synaptic release. Early activity is spontaneous, often in the form of slow bursts. Guidance cues initially set up connections between sensory organs and target structures, where incoming axons are organized in a topographical map. As the map is forming, grossly ectopic axons are eliminated, and competition between axons with different activity patterns within the same topographic location results in the further segregation of axons within the target tissue. With the onset of high-frequency sensory-driven stimulation comes the more precise organization of synapses between cells that enables high-resolution sensory detection in the young adult.

NMDARs are important during afferent development for both the removal of incorrectly positioned axons during the formation of topographically organized connections and for the later stages of input-specific segregation and refinement. NMDAR involvement in these activity-dependent processes has been most extensively studied by pharmacological blockade during the development of visual pathways and by targeted genetic manipulation in somatosensory pathways: vertebrate retino-recipient nuclei (the optic tectum of amphibians and the superior colliculus (SC) and dorsolateral geniculate nucleus (dLGN) of mammals), and somatosensory whisker afferent pathways in rodents.

Competition and Segregation – Visual Pathways

NMDAR involvement in the segregation of competing inputs was first described in the amphibian optic tectum. In a normal tadpole, retinal axons from each eye connect to the contralateral tectum, where they each form synapses with multiple tectal neurons (Figure 2(a)). Retinal axons are initially targeted to approximate anteroposterior topographic locations in the tectum using ephrin and Eph receptor

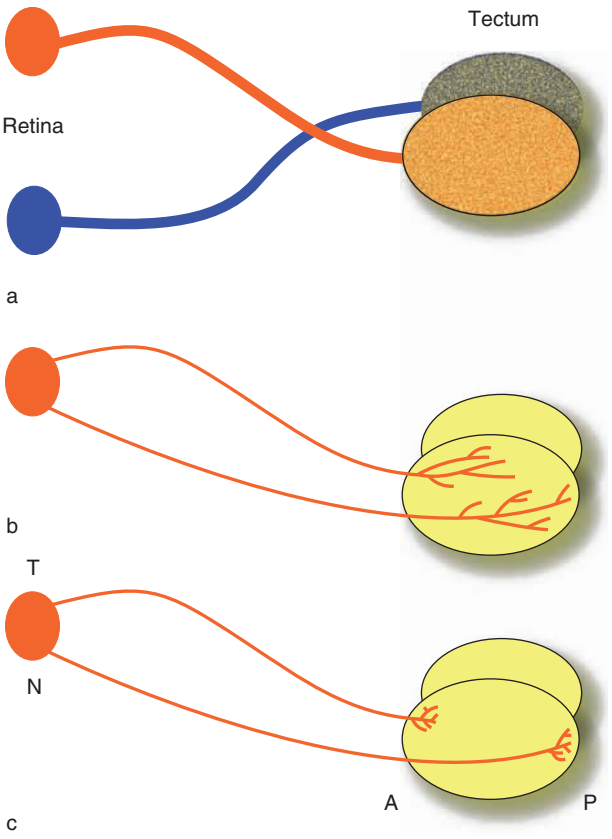

Figure 2 Development of the retinotectal pathway: (a) retinotectal pathway; (b) initial stage of topographical map formation; (c) final stage of topographical map formation. In (a), retinal ganglion cell axons from the retina of each eye travel to the brain as the optic nerve, cross to the other side of the brain at the optic chiasm, and make terminal arbors in the contralateral optic tectum of the brain. In (b), axons initially overshoot their target location in the tectum. Eph receptors expressed on retinal axons and ephrin ligands on postsynaptic membranes are required for this initial stage of map formation. Axons from the temporal side of the retina, destined for the anterior portion of the tectum, are repulsed from posterior tectum; axons from nasal retina are not as repulsed and terminate in posterior tectum. This process sets up a topographical map of the retina (and hence visual space) in the tectum. Ephrin-dependent cues must be present for the elimination of axon branches that are inappropriately topographically placed in the chick optic tectum, but do not affect the elaboration and refinement of branches in the appropriate location. In (c), eventually individual axonal branches become limited in both their mediolateral and anteroposterior extent.

guidance cues. They develop dense arbors in their target zone, through the maintenance of correctly placed branches and/or the elimination of ectopic ones (Figure 2(b)). In three-eyed tadpoles, produced by the embryonic addition of an extra eye primordium, retinal axons from two eyes innervate the same optic tectum, and the terminals of both sets of axons sort out from one another within a topographically appropriate location in a dynamic process called eye-specific segregation (Figure 3). Clusters of axons from the same regions of each retina have more similar activity patterns, and these axons sort within topographically appropriate tectal locations to optimally reproduce the point-to-point order of their cell bodies in each retina. In a tectum exposed to the NMDAR blocker AP5, axons from each eye desegregate and overlap within topographically appropriate regions (Figure 4). When AP5 is removed, the same axons resegregate.

Competitive NMDAR-dependent arbor rearrangement in the optic tectum is a form of the refinement process that keeps similarly active inputs together during normal development of many axonal projections within the brain. In the superior colliculus, NMDAR blockade disrupts the elimination of ectopic arbors during the formation of the retinocollicular map (Figure 5). Clearly, NMDARs are involved in the refinement of retinal axons, although they are not the sole mechanism by which the map refines. NMDARs perform a similarly modulatory function during retinal axon segregation in the developing lateral geniculate nucleus (LGN). In the ferret LGN, retinal axons sort within topographically appropriate regions according to their response properties to light (ON responds to light onset, OFF to light offset) within eye-specific laminae. Without NMDAR activity in the LGN during the refinement period, some axonal arbors spread across both laminae within an eye-specific layer, whereas others refine normally.

In vivo imaging of axonal dynamics during competitive arbor rearrangement provides some insight as to how this process might occur. Branches of a single axon growing into a tectal area dominated by the other eye are eliminated, and branches that grow

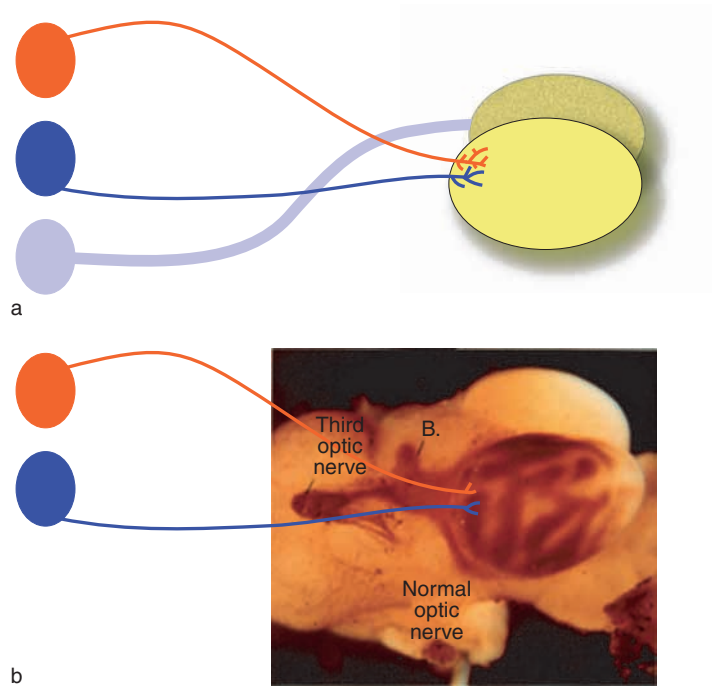

Figure 3 Eye-specific segregation in the optic tectum: (a) growth of axons when third eye is implanted; (b) pattern of axon terminals in the contralateral tectum. (a) When a third eye (dark blue) is implanted in the developing embryo, axons from this eye grow into one of the optic tecta at the same time as axons from the endogenous eye (red). Axons from both eyes undergo normal ephrin-dependent topographical refinement. Within a topographical location, however, individual axons sort out from one another, so that they occupy slightly different territory. In (b), when all the axons from one eye are labeled (here, the blue eye), the pattern of their axon terminals in the contralateral tectum can be observed. The image on the right is a whole mount showing the midbrain of a three-eyed frog in which one eye is labeled with the enzyme HRP, which is taken up and transported by retinal axons and catalyzes a colorimetric reaction in the tissue. Groups of axons from the two eyes segregate into dense stripes of clumped terminal arbors within appropriate topographic locations in the tectum. Retinal axon segregation can also be observed in the neuropil of Bellonci (B.), an amphibian structure most reminiscent of the dorsal lateral geniculate nucleus of mammals. Adapted from Constantine-Paton M and Law MI (1978) Anatomy and physiology of experimentally produced striped tecta. *Science* 202: 741–759, with permission from AAAS.

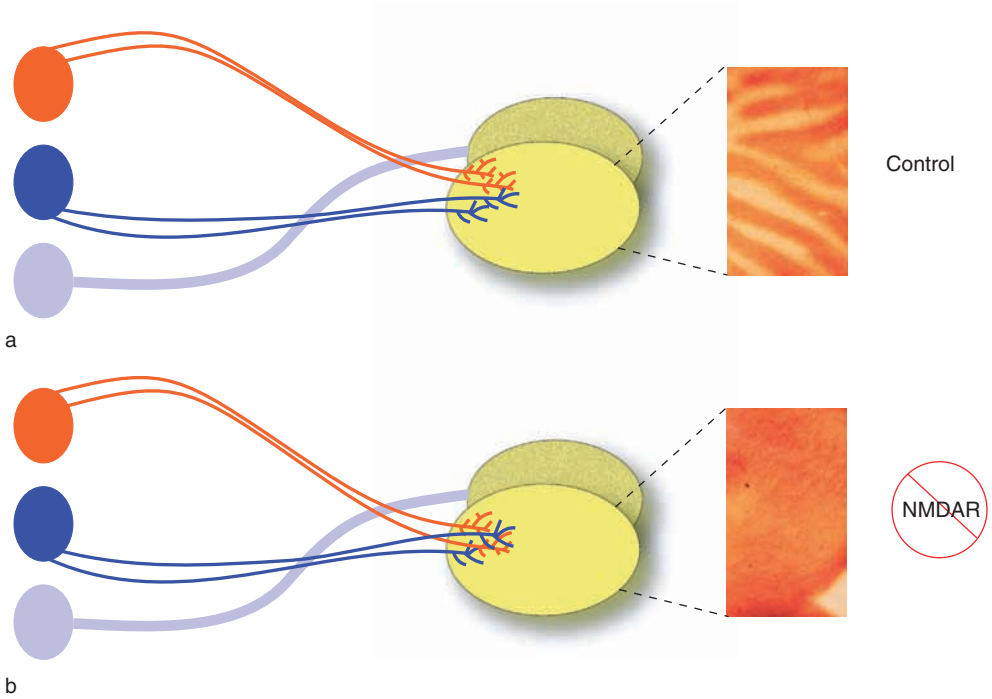

Figure 4 NMDAR-dependence of axon segregation in the optic tectum of a three-eyed frog: (a) control; (b) with NMDAR antagonist. In (a), eye-specific groups of axon terminals segregate into stripes in the three-eyed frog. In (b), when the NMDAR antagonist AP5 is applied chronically over the tectum, eye-specific axon terminals desegregate and are found in overlapping, although topographically appropriate, locations. The application of an agonist sharpens the stripe borders and the removal of the antagonist allows stripe recovery (not shown). NMDAR, *N*-methyl-D-aspartate receptor. Adapted from Cline HT, Debski EA, and Constantine-Paton M (1987) *N*-methyl-D-aspartate receptor antagonist desegregates eye-specific stripes. *Proceedings of the National Academy of Sciences of the United States of America* 84: 4342–4345.

into the area dominated by its own eye are elongated and stabilized. NMDAR blockade prevents the specific elimination of axon branches from areas dominated by the other eye (Figure 6). Thus, when an axon from a disparate retinal locus strays into a region where axons from a similar locus are concentrated, its inability to effectively stimulate tectal neurons results in its elimination. Clusters of axons from nearby retinal sites can cooperate to fire tectal neurons, and they continue to generate and stabilize axons in those areas in the absence of NMDAR activity. Thus, the specific elimination of these less efficient or incorrectly placed axons is NMDAR-dependent, after axons have already formed and elaborated using other (non-NMDAR) mechanisms.

The evidence suggesting that cooperative near-neighbor input activity is critical for activity-dependent refinement during development is strong. In the developing retina, slow cholinergic waves of activity sweep across local populations of ganglion cells and synchronize the firing of neighboring cells. Mice with a deletion of the $\beta 2$ nicotinic acetylcholine receptor subunit lack these spontaneous waves, and their ganglion cells fire in uncorrelated bursts. As a result, neighboring axons in the SC cannot cooperate to fire collicular neurons, and all axons are similarly inefficient at inducing their postsynaptic partners to fire. Retinal axons in the SC of these mice become abnormally dispersed across each collicular lobe. Postsynaptic glutamate receptors are probably involved in this refinement because retinal axons are glutamatergic; however, this has not been directly investigated.

NMDAR-induced synaptic weakening in the absence of correlated pre- and postsynaptic activity has been observed in the visual cortex at later stages of development. Axons of LGN neurons projecting to the cortex segregate into eye-specific domains early in the development of binocular animals, but remain plastic and sensitive to monocular deprivation by lid suture during a critical period. As a result, most cortical cells become responsive to the nondeprived eye, and responsiveness to the deprived eye is lost. When correlations between pre- and postsynaptic activity are disrupted by the suppression of postsynaptic action potentials, NMDAR-dependent depression of the connections between the open eye and the silenced postsynaptic neurons results. Thus, NMDAR-induced synapse weakening is an active process, in which high levels of synaptic activation are detrimental when the postsynaptic neuron cannot spike.

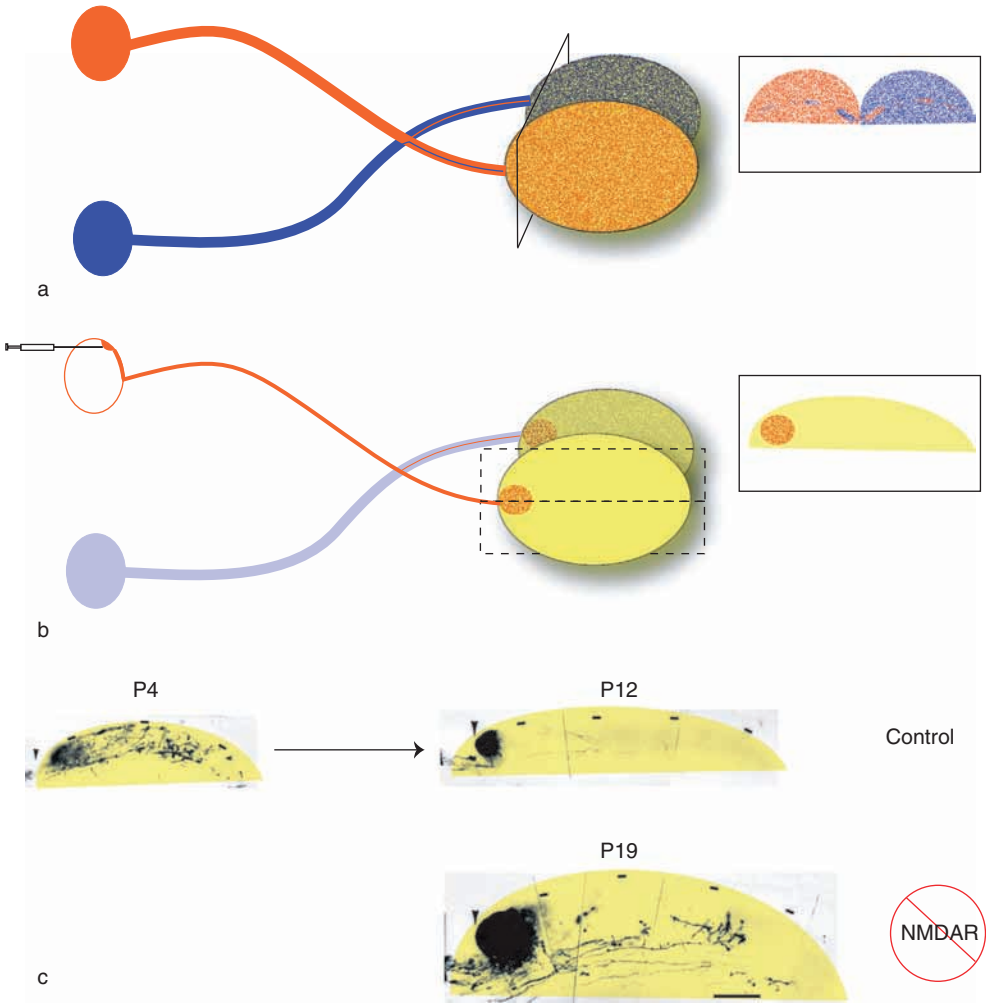

Figure 5 NMDAR-dependence of the elimination of ectopic axons in the mammalian superior colliculus (SC): (a) retinocollicular pathway; (b) dye labeling of a small population of axons; (c) in the control and with NMDAR antagonist. In (a), retinal axons project to the contralateral SC, with a minor sparse projection to the ipsilateral SC; the inset shows a coronal section of rat SC (only the superficial, retino-recipient layers of the SC are shown). In (b), dye is injected into a small portion of temporal retina labels retinal axons in the anterior SC; the inset shows a parasagittal section through the rat SC. In (c), dye injections are made at various times during the postnatal refinement of the topographical retinal map, which is largely complete by eye opening (postnatal day 13 (P13)). When NMDARs are blocked by chronic implantation of the NMDAR antagonist AP5 in a slow-release plastic implanted over the developing SC, some axons are found in inappropriate topographical locations, even at P19. The crude topography of most of the retinal axons is unaffected. NMDAR, *N*-methyl-D-aspartate receptor. Adapted from Simon DK, Prusky GT, O'Leary DD, and Constantine-Paton M (1992) *N*-methyl-D-aspartate receptor antagonists disrupt the formation of a mammalian neural map. *Proceedings of the National Academy of Sciences of the United States of America* 89: 10593–10597.

Competition and Segregation – Somatosensory Pathways

NMDARs are also crucial for the establishment of appropriate axonal and dendritic arbors in the developing somatosensory system, where genetic manipulations of the receptor have been applied. The patterning of rodent whisker-driven afferents in the pathways from the whisker pad to trigeminal nucleus to somatosensory thalamus to cortex have been extensively studied in genetically engineered mice with absent or reduced NMDAR activity. Mice lacking either the NR1 or the NR2B subunit of the NMDAR lack whisker-related patterns of sensory axons in the spinal trigeminal nucleus at birth. These mice die soon after birth from respiratory failure (NR1) and an inability to suckle (NR2B); thus, later stages of development cannot be studied. The addition of low levels of NMDARs to these mice with transgenesis permits their longer survival. In these NR1 knockdown mice, whisker patterns fail to form at all levels of the somatosensory pathway. They lack barrellettes in the trigeminal brain-stem nucleus, barreloids in the somatosensory thalamus, and whisker barrels in the somatosensory cortex (Figure 7).

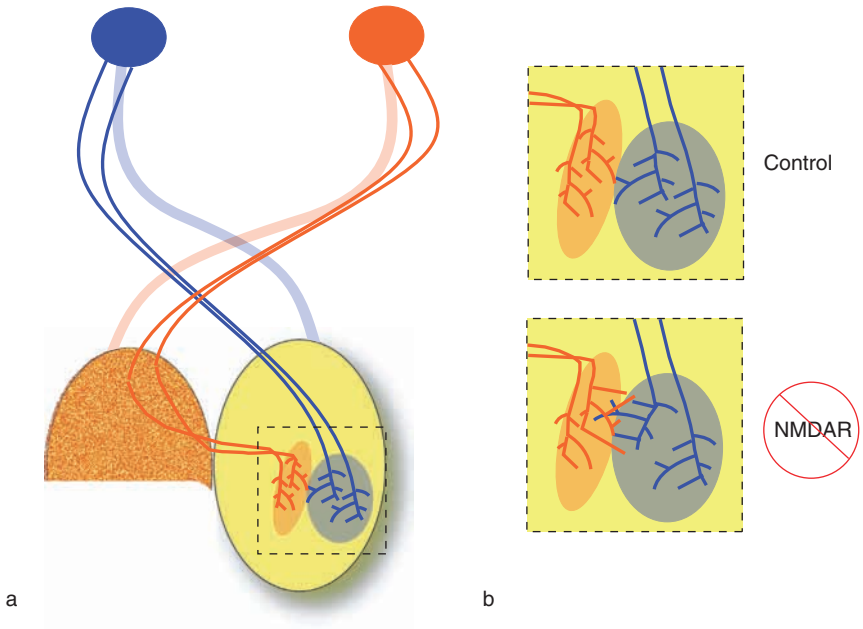

Figure 6 Ocular dominance: (a) partial lesion of one optic tectum inducing intereye competition; (b) NMDAR-dependence of the elimination of minority axons in areas dominated by other inputs. In (a), axons normally destined for a lesioned region of tectum (left) grow into the adjacent tectum and terminate in the appropriate topographical region. Ocular dominance patterns are set up by the two eyes in the doubly innervated region of the unlesioned tectum. In (b), the clear tadpole skull and proximity of the tectum at the top of the head make it possible to follow single axons over time in a living animal. Single axons were labeled with a fluorescent dye, and the pattern of branch additions and retractions over many hours was imaged by fluorescent microscopy. At the end of the experiment, whole retinal projections were labeled to determine the pattern of ocular dominance in the imaged tectum. This combined information revealed that axon branches are preferentially eliminated from territory dominated by the opposite eye. When the tadpole is raised with the NMDAR blocker in the growing media, axon branches are equally eliminated from same and opposite eye territories. These data are consistent with a role for the NMDAR in correlation-based branch elimination. NMDAR, *N*-methyl-D-aspartate receptor. (a) From Constantine-Paton M (1981) Induced ocular dominance zones in tectal cortex. In: Schmidt FO, Worden FG, and Denis SG (eds.) *The Organization of the Cerebral Cortex*, pp. 47–67. Cambridge, MA: MIT Press. Data from Ruthazer ES, Akerman CJ, and Cline HT (2003) Control of axon branch dynamics by correlated activity *in vivo*. *Science* 301: 66–70.

NR1 deletion restricted to excitatory cortical neurons has more subtle effects on cortical barrel patterns. In these mutant mice, gross whisker-related patterns form normally at all levels of the somatosensory pathway and there is a normal critical period for barrel fusion after one or several whiskers are removed. However, cortical barrel patterns are smaller and not well defined, suggesting later stages of activity-dependent refinement have been disrupted (Figure 7). Indeed, careful study of individual thalamocortical axonal afferents to the barrel shows that axonal arbors are not properly restricted to whisker boundaries (Figure 7). As in the visual system, NMDARs in the somatosensory pathways are required for the elimination of branches that are inappropriately positioned. In addition, in cortical barrels NMDAR activity also sculpts postsynaptic dendrite structure to orient dendrites within the appropriate target zone. In the cortex of the same mutant mouse, dendrites of cortical layer IV spiny stellate cells are improperly oriented and lack a normal inwardly polarized orientation with respect to the barrel walls (Figure 7).

Structural Refinement – Synapses

The exact mechanisms through which NMDAR stimulation is coupled to synaptic change and the eventual synaptic rearrangement that takes place are not yet determined. Disruption of NMDAR activity affects the development of excitatory synapses and dendritic spines, and also affects the overall axonal and dendritic structure. It is not known whether the proximal cause of these rearrangements is synaptic or extrasynaptic NMDAR activation or whether changes in synaptic strength that precede structural rearrangements are causal. However, in the immature brain there is growing evidence that the dominant effect of NMDAR activation is to weaken synapses, resulting in a net reduction of contacts and continued sprouting.

In developing pyramidal neurons of somatosensory cortex, refinement occurs via the elimination of dendritic spines, the sites of excitatory synapses on these neurons. Sensory deprivation slows this refinement in single neurons imaged over time *in vivo*, and NMDAR blockade mimics the effects of sensory

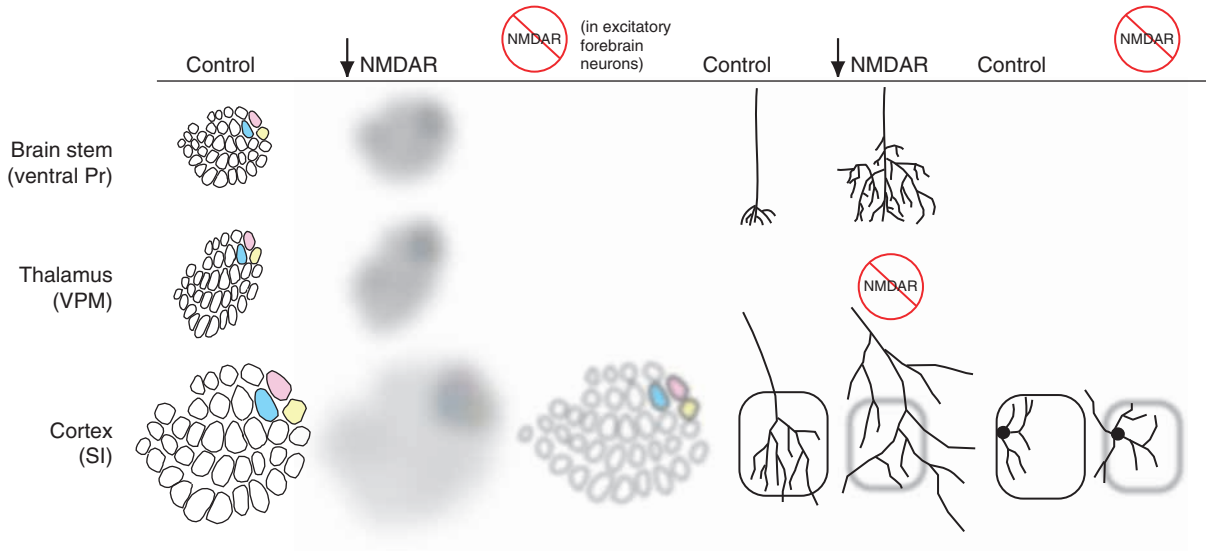

Figure 7 NMDARs as necessary for normal refinement of whisker afferents to somatosensory nuclei. Summary of generalized results from the genetic deletion of NMDARs in mutant mice. Three somatosensory nuclei are described: brain-stem nuclei (ventral Pr), thalamic nuclei (VPM), and somatosensory cortex (SI). Each barrel-shaped structure represents the organization of afferents from a single whisker. Barrels visualized with multiple anatomical methods yielded similar results, and a representative case has been drawn here. The Control column shows whisker maps in an adult, wild-type mouse, shown in similar orientation. There are sharp borders between adjacent whisker regions (first column), and whisker afferents to the brain stem have refined terminal arbors (fourth column). Thalamo cortical axons have terminal arbors that are contained within a single barrel structure in SI (fourth column), and cortical stellate cell dendrites are oriented toward the center of the barrel (sixth column). The ↓NMDAR column shows transgenic mice generated in a background strain containing an NR1 deletion, in which minimal NR1 expression has been restored by transgenesis. This permits survival beyond the first few days after birth (up to ~8 days). Whisker maps with distinct whisker regions are not detected in any somatosensory nuclei within a week after birth (second column). Reconstruction of single whisker afferent axons in the brain stem reveals that axonal terminals are larger than controls (fifth column). The knocked-out NMDAR column shows mice that had normal levels of NMDAR expression in the embryo and early postnatal life (columns three and seven). Postnatal deletion of NR1 that is restricted to excitatory forebrain neurons permits barrel formation in the SI of adult mice. However, the barrels are smaller than normal, and the borders are ill-defined (third column). Individual thalamocortical axons are not contained within the barrel walls (fifth column), and cortical stellate cell dendrites are not as polarized in orientation as cells from wild-type controls (seventh column). ↓, decreased; NMDAR, *N*-methyl-D-aspartate receptor; Pr, principal sensory nucleus of the trigeminal nerve; SI, primary somatosensory cortex; VPM, ventral posterior medial nucleus of the thalamus. Data from Iwasato T, Erzurumlu, RS, Huerta PT, et al. (1997) NMDA receptor-dependent refinement of somatotopic maps. *Neuron* 19: 1201–1210; Iwasato T, Datwani A, Wolf AM, et al. (2000) Cortex-restricted disruption of NMDAR1 impairs neuronal patterns in the barrel cortex. *Nature* 406: 726–731; Datwani A, Iwasato T, Itohara S, and Erzurumlu RS (2002) NMDA receptor-dependent pattern transfer from afferents to postsynaptic cells and dendritic differentiation in the barrel cortex. *Molecular and Cellular Neuroscience* 21: 477–492; Lee L, Iwasato T, Itohara S, and Erzurumlu RS (2005) Exuberant thalamocortical axon arborization in cortex-specific NMDAR1 knockout mice. *Journal of Comparative Neurology* 485: 280–292; Lee L, Lo F, and Erzurumlu RS (2005) NMDA receptor-dependent regulation of axonal and dendritic branching. *Journal of Neuroscience* 25: 2304–2311.

deprivation. Similarly, NMDAR blockade slows the elimination of spines from immature retinal ganglion cells. Thus, in retina and somatosensory cortex, NMDARs are required for developmental and sensory-dependent spine elimination.

Studies in the SC have found that NMDARs eliminate synapses and axonal branches only during a transient developmental period while retinal axons are refining and synaptic depression dominates. Chronic NMDAR blockade in the SC from birth increases the density of glutamatergic synapses and glutamate release sites on incoming retinal axons, until the middle of the second postnatal week (Figure 8(c)). NMDAR blockade has no net effect on synapse number or axonal sprouting in older colliculus, where there is balanced potentiation and depression, and the neuropil has a high synaptic density. Pushing the balance back toward depression with the application of an NMDA agonist pushes synapses toward functional depression in older colliculus and reduces the number of retino-collicular synapses (Figure 8(c)).

If, early in development, NMDAR-mediated synapse elimination is the primary mechanism by which axonal arbors retract, the expectation is that axonal branch elimination will follow synapse loss. Indeed, in the zebra fish and amphibian tecta, in the absence of a mature presynaptic site (identified as clusters of vesicle-release proteins), axon branches retract to the nearest synaptic site on the terminal, where they become stabilized. This is evidence that NMDAR-dependent

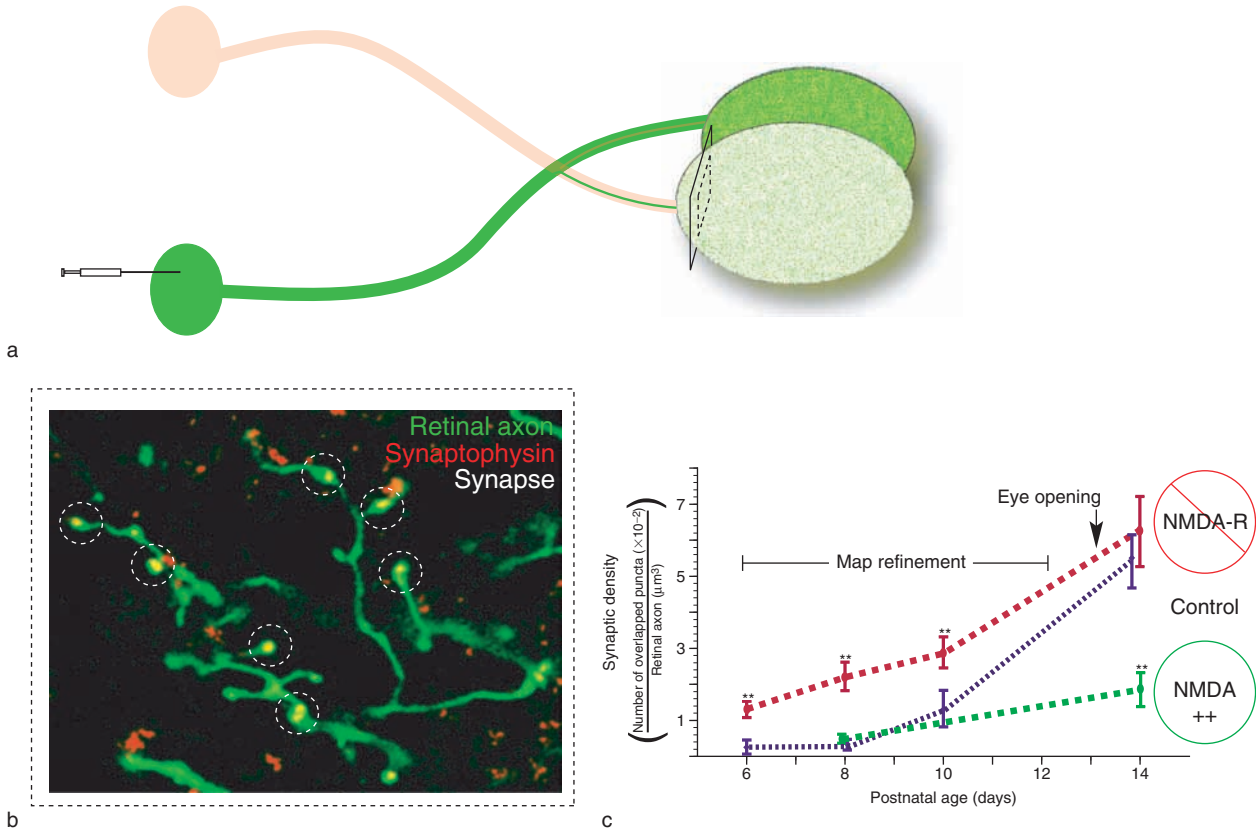

Figure 8 NMDA receptor activity eliminating synapses in the young SC: (a) dye labeling of contralateral and ipsilateral retinal axons; (b) identification of synapses on retinal axons; (c) NMDAR-dependence of synapse elimination during refinement. In (a), complete labeling of all retinal ganglion cells by intravitreally injected cholera toxin subunit B conjugated to fluorophores allows the visualization of both the ipsilateral and contralateral retinal projection from one eye. In (b), confocal microscopy and immunohistochemical localization of clustered presynaptic (or postsynaptic) markers and quantitative overlap analysis allows visualization of synapses along labeled retinal terminals. In accord with observations from electron microscopy, presynaptic release sites occur frequently at swellings or terminal bulges on the axon. In (c), the chronic blockade of NMDARs from birth increases the synaptic density along retinal axons during the first ~10 days. Synaptic density is already low at this time, and treatment with the agonist (NMDA) has no effect. By postnatal day (P)14, treatment with agonist reduces synaptic density and the antagonist ceases to have an effect. NMDAR, *N*-methyl-D-aspartate receptor; SC, superior colliculus. Adapted from Colonnese MT and Constantine-Paton M (2006) Developmental period for N-methyl-D-aspartate (NMDA) receptor-dependent synapse elimination correlated with visuotopic map refinement. *Journal of Comparative Neurology* 494: 738–7351.

sculpting of axon and dendrite structure may result from its proximal effect on synapse formation/elimination. Cytoskeletal organizing pathways involving Rho-family GTPases translate NMDAR activity into rearrangements of the cytoskeleton in dendritic processes; however, it is unknown whether this involves or is preceded by NMDAR-driven synaptic elimination.

Refinement of receptive fields and response properties

Given the importance of NMDA receptors in structural and synaptic refinement, modifications of receptive fields and response properties would also be expected when NMDAR activity is manipulated. Indeed, NMDAR activity is required for the compression of collicular receptive fields after a partial lesion limits the amount of target space available for retinal axon arborization. Similarly, enlarged receptive fields result when NMDARs are blocked during auditory map formation. The map of auditory space does not refine normally, and more neurons are broadly tuned. Finally, visual cortical neurons in NR2A knockout mice show dramatically reduced orientation preferences. Due to the complicated contribution of afferent innervation, dendritic structure, and local inhibition to receptive field responses, the locus and precise nature of these deficits are unknown. All are consistent, however, with a model in which NMDAR activity refines circuits by modifications of synaptic strength, followed by synaptic elimination or stabilization.

NMDAR Subunits and Developmental Long-Term Depression and Long-Term Potentiation

Near-simultaneous activation of pre- and postsynaptic neurons is believed to modify synaptic strength via NMDARs through mechanisms such as long-term potentiation (LTP), long-term depression (LTD), and spike-timing-dependent plasticity (STDP). These are important models of NMDAR function in the developing nervous system because nascent synapses are thought to be stabilized or eliminated by repeated synaptic potentiation or depression, respectively. STDP is a more specific form of LTP/LTD that also requires NMDARs. STDP causes potentiation if the postsynaptic cell responds within $\sim<40$ ms of a presynaptic stimulation (effective inputs), and causes depression if the postsynaptic cell responds between $\sim$40 and 100 ms after stimulation (ineffective inputs). STDP can be induced by sensory stimuli, and may therefore represent the natural mechanism by which long-lasting synaptic changes are induced via NMDAR activation. However, in developing systems the effects of STDP can be rapidly reversed by spontaneous activity, indicating that a stabilizing process must occur before STDP permanently shapes connectivity.

NMDAR-driven synaptic modifications are bidirectional, in that either LTP or LTD of synaptic strength can be induced by receptor activation. LTD tends to be easier to induce in the young brain, and as the brain matures, the induction of LTP becomes possible. This switch from LTD to LTP induction correlates with the developmental progression of NMDAR subunit expression from slowly deactivating NR2B-rich receptors to mixed slowly and rapidly deactivating receptors and with other subunit-specific changes in the gating properties of these receptors that occur with development. Given this correlative evidence, it has been tempting to functionally link NR2B-rich receptor complexes to LTD and NR2A-rich complexes to LTP. At certain ages, a carefully controlled blockade of receptors containing NR2B eliminates hippocampal LTD, without affecting LTP. Mice lacking NR2A show a progressive deficit in LTP in the hippocampus with age that parallels the normal increases in hippocampal NR2A expression in wild-type mice.

However, there are many regions of the brain where tight correlations between LTD/NR2B and LTP/NR2A do not exist. It is unknown how mixed NR2A/NR2B receptors contribute to LTP and LTD. It seems reasonable that the dynamics of the calcium signal through these receptors may differ at different synaptic sites, due to the inherent subunit-specific differences, location, and association with different kinases and phosphatases that regulate calcium flux. Thus the NMDAR is likely sending mixed signals through the same synapses, and the net effect is a result of this diverse network of signals.

Changes in NMDAR subunit composition are also correlated with changes in the expression of scaffold proteins associated with the NMDAR at postsynaptic sites. These proteins may prove to be critically important in linking calcium influx through NMDARs to specific signaling pathways by tethering the necessary molecules close to the channel pore. In this way, both the electrophysiological responses of the NMDAR channel and the effector proteins it binds might interact to control the direction of synaptic strength changes during development.

A Developing NMDA Receptor Complex

Membrane-associated guanylate kinases (MAGUKs) are important components of the NMDAR protein complex in the postsynaptic density (PSD). Recent studies have begun to explore their contribution to the divergent functions of the NMDAR during synaptogenesis and developmental plasticity. These intracellular NMDAR-binding proteins contain multiple protein–protein interaction domains and bind to both scaffolding and signaling molecules at the PSD. In this way, NMDARs are connected to other transmembrane receptors in the PSD through a dynamic horizontal lattice of protein interactions. This complex links the receptor to the cytoskeleton, to receptors such as α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), G-protein-coupled glutamate, and inositol triphosphate receptors (which release calcium from intracellular stores). Signaling molecules that drive cytoskeletal rearrangements and other calcium-dependent processes are also tethered near the NMDAR channel pore by MAGUKs.

Two MAGUKs in particular, synapse-associated protein (SAP-)102 and postsynaptic density protein (PSD-95, are differentially expressed during development. SAP-102 is the only MAGUK present at the young ages when immature NMDARs predominate. Like NR2B, SAP-102 expression does not disappear, but continues to be expressed in the adult, along with the mature MAGUK, PSD-95. They appear to have some specificity for NMDAR complexes containing NR2B or NR2A, respectively. PSD-95 has recently been shown to produce synapse potentiation by anchoring stargazin AMPA receptor (AMPAR) complexes in the PSD, and the activation of NMDARs containing NR2A have been reported to promote insertion of AMPARs containing GluR1 (a critical step in LTP induction). With eye-opening in the developing visual pathway of rodents, PSD-95 is rapidly inserted

into visual synapses, and synaptic potentiation and a second wave of synapse refinement follows soon after.

Calcium-sensitive proteins couple NMDAR activation to downstream transcription/translation, cytoskeletal rearrangements, changes in synaptic strength, and cell death pathways. The exact mechanism linking calcium influx through NMDARs with the activation of these pathways is not well defined. Guanine exchange factors (GEFs) and small GTPases are known to be downstream of NMDAR activation and to drive bidirectional changes in synaptic strength and cytoskeletal rearrangements. MAGUK scaffold proteins may be critically important in localizing these calcium-responsive proteins and their effector proteins to NMDARs, although some, such as Ras GTPase, are also independently targeted to membrane microdomains, and some, such as CaMKII-α, are able to bind the NMDAR directly. One GEF, RasGRF1, binds the NR2B subunit directly. Thus, subunit-specific MAGUK binding may not entirely account for the differential effects of NR2B versus NR2A activation. Existing data on subunit-specific protein interactions and developmental expression do suggest that entirely different NMDAR signaling complexes exist, depending on the subunit composition of the receptor and the proteins that are available to bind. Such distinctions are exceptionally difficult to establish with biochemistry and molecular biology alone because each synapse can contain a heterogeneous array of receptor-signaling complexes, but they may hold the key to the mechanisms of NMDAR function through development.

Summary

Data from the visual and somatosensory systems support the concept of the NMDAR as a causal effector of changes in synaptic strength and structural rearrangements during the activity-dependent development of sensory pathways. Mechanisms similar to LTP and LTD appear to be involved, where NMDARs detect the coincident activation of pre- and postsynaptic neurons and strengthen or weaken synapses based on this information. Early in development when immature subunits permit synaptic depression, NMDARs eliminate synapses. In many cases of NMDAR function during development, synaptic potentiation or depression precedes the structural reinforcement or withdrawal of connections. The subunit-specific interaction of NMDARs and scaffold proteins may direct changes in synaptic strength that underlie developmental refinement of neuronal connectivity and structure; however, this has not yet been directly examined. The heterogeneity of the receptor subunits and associated proteins permits a modifiable NMDAR complex that is exquisitely sensitive to changing patterns of activity in the developing brain.

See also: Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; NMDA Receptors and Disease + C464; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation.

Further Reading

Carmignoto G and Vicini S (1992) Activity-dependent decrease in NMDA receptor responses during development of the visual cortex. *Science* 258: 1007–1011.

Cline HT, Debski E, and Constantine-Paton M (1989) NMDA receptor antagonist desegregates eye-specific stripes. *Proceedings of the National Academy of Sciences of the United States of America* 84: 4342–4345.

Cull-Candy SG and Leszkiewicz DN (2004) Role of distinct NMDA receptor subtypes at central synapses. *Science Signal Transduction Knowledge Environment* 2004(255): re16 [online].

Hestrin S (1992) Developmental regulation of NMDA receptor-mediated synaptic currents at a central synapse. *Nature* 357: 686–689.

Husi H, Ward MA, Choudhary JS, Blackstock WP, and Grant SG (2000) Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. *Nature Neuroscience* 3: 661–669.

Kennedy MB, Beale HC, Carlisle HJ, and Washburn LR (2005) Integration of biochemical signaling in spines. *Nature Reviews Neuroscience* 6: 423–434.

Kim E and Sheng M (2004) PDZ domain proteins of synapses. *Nature Reviews Neuroscience* 5: 771–781.

Komuro H and Rakic P (1998) Orchestration of neuronal migration by activity of ion channels, neurotransmitter receptors, and intracellular Ca^{2+} fluctuations. *Journal of Neurobiology* 37: 110–130.

Law MI and Constantine-Paton M (1981) Anatomy and physiology of experimentally produced striped tecta. *Journal of Neuroscience* 1: 741–759.

Monaghan DT and Buller AL (1994) Anatomical, pharmacological, and molecular diversity of native NMDA receptor subtypes. In: Collingridge GL and Watkins JC (eds.) *The NMDA Receptor,* 2nd edn., pp. 158–176. Oxford: Oxford University Press.

Mu Y and Poo MM (2006) Spike timing-dependent LTP/LTD mediates visual experience-dependent plasticity in a developing retinotectal system. *Neuron* 50: 115–125.

O'Leary DD and McLaughlin T (2005) Mechanisms of retinotopic map development: Ephs, ephrins, and spontaneous correlated retinal activity. *Progress in Brain Research* 147: 43–65.

Rumbaugh G and Vicini S (1999) Distinct synaptic and extrasynaptic NMDA receptors in developing cerebellar granule neurons. *Journal of Neuroscience* 19: 10603–10610.

VanZundert B, Yoshii A, and Constantine-Paton M (2004) Receptor compartmentalization and trafficking at glutamate synapses: A developmental proposal. *Trends in Neuroscience* 27: 428–437.

West AE, Chen WG, and Dalva MB (2001) Calcium regulation of neuronal gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11024–11031.

NMDA Receptors and Disease+C464

T P Wong, J G Howland, and Y T Wang, University of British Columbia, Vancouver, BC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

N-Methyl-D-aspartate (NMDA) receptors are a major subtype of ionotropic excitatory glutamate receptors that mediate synaptic transmission at the vast majority of excitatory synapses in the central nervous system. Functional NMDA receptors are heterotetromeric complexes of NR1, NR2A–NR2D, and NR3A and NR3B subunits. The subunit compositions of the receptors affect both their functional properties and subcellular distributions. Like other types of ionotropic glutamate receptors, NMDA receptors are ion channels that allow major monovalent cations such as sodium and potassium to cross the plasma membrane, either into or out of the neuron, depending on the chemical and electrical gradients of these ions at a given time. However, NMDA receptors possess several unique properties that distinguish them from other ionotropic glutamate receptors. In addition to monovalent cations, NMDA receptors are also highly permeable to the divalent cation Ca^{2+}, which has numerous important intracellular functions. (Note: the only other type of glutamate receptor that exhibits similar Ca^{2+} permeability is the GluR2 subunit-lacking α-amino-3-hydroxy-5-methyl-isoxazole-4-propionic acid (AMPA) receptor, which represents a minor proportion of AMPA receptors in the brain.) The permeability of NMDA receptors to Ca^{2+} places them in a unique position to control the initiation of a lengthy list of Ca^{2+}-dependent cellular events. NMDA receptors also have two unique requirements that control their activation. Unlike other glutamate receptors, NMDA receptors require co-agonists such as glycine to bind to specific sites at the same time as glutamate binds, in order to be activated. Additionally, under resting conditions, magnesium ions block the channel pore of NMDA receptors. In order for the magnesium block to be removed, the membrane must be depolarized, which enables NMDA receptors to function as a detector for the coincidental presynaptic release of glutamate and postsynaptic depolarization, a process critical for computation of neuronal signals and for the establishment of certain forms of activity-dependent synaptic plasticity. These unique properties place NMDA receptors in a position to regulate various neuronal functions, ranging from synaptic plasticity at a single synapse to complex cognitive functions such as learning and memory. However, abnormal activation of NMDA receptors may also result in cellular dysfunction and contribute to the symptoms of many disorders of the nervous system. Therefore, the contribution of NMDA receptor abnormalities to the pathogenesis of several disorders will be discussed in the following sections, with a particular focus on progress made in the last decade.

Potential Mechanisms Underlying the Roles of NMDA Receptors in Brain Dysfunction

Excitotoxicity: A Common Mechanism by which NMDA Receptors Contribute to the Pathogenesis of Neurodegenerative Disorders

Excitotoxicity refers to the overactivation of glutamate receptors as a result of increased release and/or decreased uptake of excitatory amino acid transmitters, primarily glutamate. During pathological conditions such as those experienced during ischemia, excessive release of glutamate overactivates glutamate receptors, resulting in a massive influx of sodium ions, chloride ions, and water, which cause neurons to swell and may ultimately lead to rapid neuronal loss from necrosis. Because of the high calcium permeability of NMDA receptors, their overactivation is also accompanied by an accumulation of excessive amounts of intracellular calcium, which initiates other intracellular cascades, leading to slow (delayed) apoptotic neuronal loss. The mechanisms through which excessive calcium influx leads to apoptosis are complex and not fully understood. Accumulating evidence has implicated several intracellular molecules and signaling pathways, including mitochondrial dysfunction, reactive oxygen species production, activation of catabolic enzymes which degrade proteins, nucleic acids, and other cellular components. While the detailed mechanisms through which overactivation of NMDA receptors lead to excitotoxicity may differ between disorders, a large body of evidence implicates NMDA receptor-mediated excitotoxicity as a common mechanism mediating the pathogenesis of many neurodegenerative disorders, ranging from acute events such as stroke and brain trauma to chronic neurodegenerative disorders such as Alzheimer's disease, Parkinson's disease, and Huntington's disease. Evidence for the involvement of NMDA receptor-mediated excitotoxicity in the pathogenesis of central nervous neurodegenerative disorders will be discussed in later sections using stroke and Alzheimer's disease as examples of acute and chronic neuronal degeneration, respectively.

Altered NMDA Receptor-Mediated Synaptic Plasticity May Contribute to the Pathophysiology of Brain Disorders with Abnormal Synaptic Function

Synaptic plasticity refers to the ability of synapses to change their strength. Such plastic changes in synaptic efficacy are fundamental for both neuronal development and normal brain functions. Because of the unique features of high calcium permeability and the voltage-dependent blockade by extracellular magnesium, NMDA receptors are essential for mediating many forms of synaptic plasticity in the brain throughout development. The best characterized forms of NMDA receptor-mediated synaptic plasticity are long-term potentiation (LTP) and long-term depression (LTD) of AMPA receptor-mediated synaptic transmission at CA3–CA1 synapses in the hippocampus. LTP and LTD are prominent cellular models of certain forms of learning and memory. Not surprisingly, alterations in LTP and LTD are correlated with cognitive impairments in rodent models of brain disorders such as Alzheimer's disease and fragile X syndrome. In addition, accumulating evidence supports the hypothesis that abnormal NMDA receptor function and synaptic plasticity are involved in the pathogenesis of brain dysfunctions in a number of neurological and neuropsychiatric disorders not characterized by evidence of cell death or gross structural damage to the brain. In these disorders, it is hypothesized that relatively subtle functional and structural alterations in synaptic connections between neurons in affected brain regions may underlie some of the symptoms. Examples of such brain disorders include pain, addiction, and schizophrenia. In the following sections there is a discussion of research on these disorders that has demonstrated the pathological effects of abnormal synaptic plasticity and has provided insight into a number of therapeutic strategies for reversing these abnormalities.

Examples of Disorders Hypothesized to Occur as a Result of NMDA Receptor Dysfunction

Stroke

Stroke is a sudden loss of blood supply to part of the brain that occurs when a blood vessel either bursts or is blocked. Neurons in the affected area are severely injured or die as a result of the interrupted supply of oxygen and nutrients from the blood. The specific consequences of a stroke depend on a number of factors, including the duration of blood loss and the specific brain area affected. Different cellular processes underlie neuronal injury following stroke, but particularly strong evidence suggests that excitotoxicity is the primary cause of neuronal death following stroke. For example, in both *in vitro* and *in vivo* experimental models, levels of extracellular glutamate are rapidly elevated after stroke to concentrations that produce massive increases in intracellular calcium. The levels of intracellular calcium reach concentrations sufficient to cause excitotoxic necrotic and apoptotic neuronal loss. Most importantly, administration of NMDA receptor antagonists before or immediately after stroke onset is effective in reducing the neuronal damage in various preclinical animal models.

Significant evidence thus supports the critical involvement of NMDA receptor-mediated excitotoxicity in mediating stroke-related neuronal loss and NMDA receptors as an important target for stroke therapy. As a result, numerous pharmaceutical companies have spent considerable resources in an attempt to design effective NMDA receptor antagonists for treatment of stroke. However, the effectiveness of NMDA receptor antagonists in clinical trials for stroke and traumatic brain injuries has been low. A number of factors, including undesirable side effects at therapeutically effective doses, an inability to administer the drugs within their neuroprotective windows, and heterogeneity in the patient population, likely underlie these disappointing results.

Importantly, recent advancements in understanding the detailed molecular mechanisms of NMDA receptor-mediated excitotoxicity may aid in developing novel NMDA receptor-based neuroprotective therapeutics that would reduce neurotoxicity without causing significant side effects. For example, one potential challenge in using NMDA receptor antagonists for stroke therapy is to avoid inhibiting the normal functions of NMDA receptor signaling, which are essential for neuronal survival. One promising approach to avoid completely block NMDA receptors is to use low-affinity NMDA receptor antagonists such as memantine to treat stroke. Memantine is a low-affinity, open-channel blocker of NMDA receptors. Due to its relatively low affinity, memantine only blocks significant numbers of NMDA receptors when they are highly overactivated, such as during stroke. Therefore, memantine has minimal effects on NMDA receptors in areas unaffected by the stroke, thereby reducing the side effects typical of potent NMDA receptor antagonists. Another promising therapeutic approach for stroke is to specifically target the cell death signaling pathways downstream of NMDA receptors without blocking normal functioning of the receptors. For example, one of these pathways is initiated by the association of NR2B subunit-containing NMDA receptors with the scaffolding protein PSD95. Recent experiments using preclinical models of stroke have shown that administration of a synthetic peptide that disrupts the association of the C-terminal tail

of NR2B subunits with PSD95 substantially reduces ischemia-induced damage without noticeable side effects. Given their promise in preclinical models of stroke, both low-affinity NMDA receptor antagonists and interference peptides are currently being tested in clinical trials for use in human patients suffering from stroke.

Alzheimer's Disease

Alzheimer's disease (AD) is the main cause of dementia in the elderly. Pathologically, AD is characterized by cortical atrophy, the presence of amyloid plaques and neurofibrillary tangles, and the granulovacuolar degeneration of neurons in vulnerable brain areas such as the neocortex and the hippocampus. While the mechanisms underlying AD neuropathology are not well understood, emerging evidence suggests that altered NMDA receptor function and enhanced NMDA receptor-mediated excitotoxicity may contribute to both pathological and functional abnormalities of AD. For example, amyloid plaques are primarily composed of the amyloid beta peptide (Aβ peptide). Recent studies have shown that the Aβ peptide enhances NMDA-receptor-mediated neurotoxicity by either potentiating NMDA receptor function or increasing available glutamate by inhibiting glutamate transport.

The Aβ peptide also has significant effects on NMDA receptor-mediated synaptic plasticity in the hippocampus and may be a cellular substrate for dementia in AD. Application of the Aβ peptide inhibits hippocampal LTP in both *in vitro* and *in vivo* preparations. LTP is also greatly reduced in transgenic AD animal models expressing high levels of Aβ peptide. Interestingly, the Aβ peptide may not affect or even potentiate LTD in the hippocampus. Since LTP and LTD may be mediated by distinct subpopulations of NMDA receptors, the differential effects of the Aβ peptide on LTP and LTD may result from subunit-specific actions on NMDA receptors. Future studies of the effects of Aβ peptide on NR2A- and NR2B-containing NMDA receptors are required to test this hypothesis. Whether the effects of the Aβ peptide on LTP and LTD contribute to the neurodegeneration in AD remains unclear. However, a recent study provides compelling evidence that the enhanced LTD caused by the Aβ peptide may mediate the spine and neuronal loss typically observed in AD. The importance of altered NMDA receptor signaling and synaptic plasticity for the neuropathology and cognitive dysfunction in AD is further supported by the beneficial effect of treatment of AD patients with memantine at various stages of dementia. These results also provide strong support for using NMDA receptor modulators/antagonists as part of the treatment of AD.

Pain

Pain is an unpleasant sensation associated with actual or potential tissue damage in the body. While normal pain responses are adaptive, pathological pain disorders are common and debilitating for many people. Pain disorders are caused by a number of conditions, including chronic inflammation or aberrant activity in peripheral nerves and nervous tissue (neuropathic pain). Neuropathic pain may be caused by either increased pain responses toward noxious stimuli (hyperalgesia) or decreased pain threshold to the point that normally nonnoxious stimuli such as touch or gentle pressure produce pain (allodynia). These pathological alterations are produced by hyperresponsiveness of nociceptive afferents sending sensory information to the dorsal horn of the spinal cord. After synapsing in the dorsal horn, this information is transmitted to the brain. Importantly, these pain pathways can be experimentally sensitized by repetitive stimulation of pain afferents, an effect that is mediated primarily by glutamate. Sensitization of glutamatergic transmission after noxious stimuli has also been reported in the cingulate cortex, an important forebrain region for nociception. Given the importance of the NMDA receptor in glutamatergic transmission and plasticity, it has become a research target for understanding the mechanism of central pain sensitization.

Using preclinical models, the involvement of NMDA receptors in the induction and expression of pain-induced sensitization in the spinal cord is supported by a number of findings. For example, direct administration of NMDA receptor antagonists on dorsal horn neurons abolishes the postsynaptic hyperexcitability caused by repetitive stimulation of primary pain afferents. Additionally, the intrathecal application of either NMDA receptor agonists or allosteric modulators such as D-serine induces hyperalgesia. NMDA receptor antagonists are also clinically effective in reducing hyperalgesia in patients with chronic pain.

Although NMDA receptors are important for central sensitization to pain, little is known about the mechanisms of how NMDA receptor activation contributes to causing hyperexcitability of dorsal horn neurons. However, a number of possibilities exist. For example, central sensitization may alter glutamate receptor function. Consistent with this possibility, the expression levels and kinetics of NMDA and AMPA receptors in the dorsal horn are altered in an animal model of inflammatory pain. Intriguingly, levels of AMPA receptor expression are increased in neuropathic pain, and this increase parallels the time course of hyperalgesia, suggesting an important role of increased AMPA receptor expression in the hyperexcitability of dorsal horn neurons. Given that potentiated AMPA

receptor function is also the underlying mechanism of hippocampal LTP, these enhancements of AMPA receptor function in sensitized dorsal horn neurons may be the result of an NMDA receptor-mediated mechanism similar to LTP. It is noteworthy that NMDA receptor-mediated LTP is readily induced in dorsal horn neurons by high-frequency stimulation of the primary nociceptive afferents. The recent discovery of 'silent synapses' in dorsal horn neurons also supports a potential role of LTP-like synaptic plasticity in the formation of central pain sensitization. Silent synapses are glutamatergic synapses that contain NMDA but not AMPA receptors; therefore, they are not activated during low levels of synaptic activity. Evidence suggests that during LTP, silent synapses are converted into functional synapses by recruiting AMPA receptors into the postsynaptic membrane. Recent experiments have shown that silent synapses can be activated in dorsal horn neurons, suggesting that recruitment of AMPA receptors could be a potential mechanism underlying central pain sensitization.

Available evidence also suggests that NR2B subunit-containing NMDA receptors may mediate the central sensitization of pain perception. The expression of NR2B subunits is enriched in dorsal horn neurons compared to other NR2 subunits. Additionally, specific antagonists of NR2B-containing NMDA receptors are effective analgesic agents for neuropathic pain. Finally, inflammation causes an increase in NR2B-containing NMDA receptors in the anterior cingulate cortex, while overexpression of NR2B subunits in the anterior cingulate cortex facilitates hyperalgesia caused by inflammation. Taken together, these findings highlight the importance of NR2B-containing NMDA receptors as therapeutic targets for chronic pain.

Drug Addiction

Drug addiction is a brain disorder that may be caused by maladaptive alterations in synaptic networks involving the mesocorticolimbic dopamine system. The mesocorticolimbic system consists of dopaminergic neurons which project from the ventral tegmental area (VTA) in the midbrain to forebrain areas such as the nucleus accumbens (NAc), amygdala, hippocampus, and prefrontal cortex. Alterations of dopaminergic inputs from the VTA to NAc and other forebrain regions have been investigated extensively in an effort to determine their potential role in addictive behavior. Both the VTA and NAc also receive glutamatergic inputs that undergo long-lasting modifications in response to commonly abused drugs such as cocaine and amphetamine. Accumulating evidence supports the hypothesis that the formation of addictive behavior involves plastic changes of glutamatergic inputs to various areas of the mesocorticolimbic system. Not surprisingly, there is strong evidence for the role of NMDA receptors in these plastic changes. For example, a single exposure to cocaine readily induces NMDA receptor-mediated LTP of glutamatergic inputs to the VTA (VTA-LTP). Biochemical experiments suggest that this VTA-LTP results from an increase in AMPA/NMDA receptor ratio on the postsynaptic membranes of VTA neurons. In addition, repetitive drug exposure reduces the induction threshold of VTA-LTP. However, a causal role for altered VTA synaptic plasticity in drug addiction remains to be established, partially due to a lack of specific inhibitors that interfere with the expression of LTP without affecting the normal function of NMDA receptors and downstream signaling pathways.

In addition to the plastic changes observed in the VTA, repeated administration of psychostimulants to rats enables LTD of glutamatergic inputs to the NAc (NAc-LTD), an effect which requires activation of NMDA receptors. Similar to hippocampal CA1 LTD, the expression of NAc-LTD is also mediated by the removal of synaptic AMPA receptors through GluR2 subunit-dependent clathrin-mediated endocytosis. Importantly, when a short synthetic peptide derived from the carboxyl tail region of GluR2 that blocks clathrin-mediated endocytosis was delivered into the NAc neurons, it blocked the formation of NAc LTD and the expression of amphetamine-induced behavioral sensitization. (Note: behavioral sensitization is a common behavioral model used to model some aspects of drug addiction.) Given that the peptide did not affect normal AMPA or NMDA receptor function, basal synaptic transmission, or LTP, these results provide strong evidence that NAc-LTD underlies the expression of amphetamine sensitization. These findings from the VTA and NAc provide strong evidence that abnormal NMDA receptor function and glutamatergic synaptic plasticity are involved in the pathogenesis of certain drug addition-related behaviors. Targeting these processes may provide a new avenue for developing novel drug addiction treatments.

Schizophrenia

Schizophrenia is a severe psychiatric disorder that affects approximately 1% of the population. The neurobiological causes of schizophrenia remain largely unknown but are likely related to relatively subtle structural abnormalities in various limbic areas and an imbalance of various neurotransmitters in the brain. The positive symptoms of schizophrenia, which include hallucinations and paranoia, are at least partially caused by dopamine hyperactivity, as they are reduced by dopamine antagonists (typical antipsychotics) and exaggerated by dopamine agonists. However, the cognitive and negative symptoms of schizophrenia, which are particularly debilitating, are generally resistant

to treatment with typical antipsychotic drugs. This suggests that other neurotransmitter systems may be abnormal in schizophrenia. Indeed, strong evidence suggests that alterations in the glutamatergic system exist in schizophrenia. For example, a number of psychotomimetic drugs, such as the NMDA receptor antagonists phencyclidine and ketamine, induce schizophrenic-like psychosis in normal people as well as exaggerate some of the symptoms of patients with schizophrenia. Interestingly, the effects of NMDA receptor antagonists more closely resemble those of schizophrenia than the primarily positive symptoms induced by dopamine agonists such as amphetamine. These findings suggest that hypofunction of NMDA receptors contribute to a variety of schizophrenia symptoms.

Several lines of emerging evidence support the NMDA receptor hypofunction hypothesis of schizophrenia: (1) biochemical analyses of the brains of patients with schizophrenia demonstrate abnormal distribution and expression levels of NMDA receptors in some studies, (2) various genetic approaches to disrupting NMDA receptor function are sufficient to model some of the symptoms of schizophrenia in mice, and (3) augmenting NMDA receptor function with positive modulators of the glycine binding site has displayed subtle but promising antipsychotic effects.

The complexity of schizophrenia makes determining the exact role of NMDA hypofunction in the disorder difficult. Since NMDA receptors are important for mediating the basal activity of inhibitory interneurons in cortical and limbic regions known to be involved in schizophrenia, NMDA receptor hypofunction may disrupt normal processing in these areas by decreasing inhibitory synaptic drive. Additionally, NMDA receptors have significant neurotrophic roles during early development. Current theories suggest that the etiology of schizophrenia has a strong developmental component that is likely caused by a combination of genetic and environmental effects. Therefore, early alterations in NMDA receptors could interact with other genetic or environmental effects and contribute to the development of the disorder.

Concluding Remarks

Evidence from both preclinical and clinical studies demonstrates the importance of NMDA receptor dysfunction in numerous acute and chronic neurological and psychiatric disorders. Although the detailed cellular signaling pathways likely vary between the disorders, excitotoxicity and abnormal modification of synaptic transmission are two primary mechanisms by which NMDA receptor dysfunction is involved in these disorders. Unfortunately, the results of clinical trials testing generalized NMDA receptor-based modulators and antagonists for the treatment of these disorders have achieved limited success at the present time. Ongoing research aimed at delineating the downstream signaling pathways mediated by specific subpopulations of NMDA receptors provides a rich avenue for the development of new NMDA receptor-based treatments with greater therapeutic benefits and fewer side effects. The use of low-affinity NMDA antagonists like memantine for reducing NMDA receptor-mediated excitotoxicity highlights the importance of developing NMDA receptor-based therapies with different pharmacological profiles for treating brain disorders. The success of this strategy could be further enhanced by better understanding the subunit-specific mechanisms involved in individual disorders. In addition to directly targeting the NMDA receptor, a second promising strategy is to develop novel drugs that specifically target signaling molecules downstream of NMDA receptor activation, as exemplified by the recent success in ameliorating ischemic damage by disrupting the interaction between NR2B subunits of the NMDA receptor and PSD95. Given the rapid pace of research in the area of NMDA receptor function, significant advancements in treating these disorders are likely in the near future.

See also: Alzheimer's Disease: An Overview; Excitotoxicity in Neurodegenerative Disease; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; NMDA Receptors and Development; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation.

Further Reading

Aarts M, Liu Y, Liu L, et al. (2002) Treatment of ischemic brain damage by perturbing NMDA receptor–PSD-95 protein interactions. *Science* 298: 846–850.

Bliss TV and Collingridge GL (1993) A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature* 361: 31–39.

Brebner K, Wong TP, Liu L, et al. (2005) Nucleus accumbens long-term depression and the expression of behavioral sensitization. *Science* 310: 1340–1343.

Coyle JT and Tsai G (2004) NMDA receptor function, neuroplasticity, and the pathophysiology of schizophrenia. *International Review of Neurobiology* 59: 491–515.

Cull-Candy S, Brickley S, and Farrant M (2001) NMDA receptor subunits: Diversity, development and disease. *Current Opinion in Neurobiology* 11: 327–335.

Gaunitz C, Schuttler A, Gillen C, et al. (2002) Formalin-induced changes of NMDA receptor subunit expression in the spinal cord of the rat. *Amino Acids* 23: 177–182.

Greene R (2001) Circuit analysis of NMDAR hypofunction in the hippocampus, *in vitro*, and psychosis of schizophrenia. *Hippocampus* 11: 569–577.

Harris JA, Corsi M, Quartaroli M, et al. (1996) Upregulation of spinal glutamate receptors in chronic pain. *Neuroscience* 74: 7–12.

Hsieh H, Boehm J, Sato C, et al. (2006) AMPAR removal underlies Abeta-induced synaptic depression and dendritic spine loss. *Neuron* 52: 831–843.

Li P and Zhuo M (1998) Silent glutamatergic synapses and nociception in mammalian spinal cord. *Nature* 393: 695–698.

Lipton SA (2004) Failures and successes of NMDA receptor antagonists: Molecular basis for the use of open-channel blockers like memantine in the treatment of acute and chronic neurologic insults. *NeuroRx* 1: 101–110.

Liu QS, Pu L, and Poo MM (2005) Repeated cocaine exposure *in vivo* facilitates LTP induction in midbrain dopamine neurons. *Nature* 437: 1027–1031.

Millan MJ (2005) *N*-methyl-D-aspartate receptors as a target for improved antipsychotic agents: Novel insights and clinical perspectives. *Psychopharmacology (Berlin)* 179: 30–53.

Mohn AR, Gainetdinov RR, Caron MG, et al. (1999) Mice with reduced NMDA receptor expression display behaviors related to schizophrenia. *Cell* 98: 427–436.

Olney JW and Farber NB (1995) Glutamate receptor dysfunction and schizophrenia. *Archives of General Psychiatry* 52: 998–1007.

Randic M, Jiang MC, and Cerne R (1993) Long-term potentiation and long-term depression of primary afferent neurotransmission in the rat spinal cord. *Journal of Neuroscience* 13: 5228–5241.

Ross CA, Margolis RL, Reading SA, et al. (2006) Neurobiology of schizophrenia. *Neuron* 52: 139–153.

Thomas MJ, Beurrier C, Bonci A, et al. (2001) Long-term depression in the nucleus accumbens: A neural correlate of behavioral sensitization to cocaine. *Nature Neuroscience* 4: 1217–1223.

Ungless MA, Whistler JL, Malenka RC, et al. (2001) Single cocaine exposure *in vivo* induces long-term potentiation in dopamine neurons. *Nature* 411: 583–587.

Walsh DM, Klyubin I, Fadeeva JV, et al. (2002) Naturally secreted oligomers of amyloid beta protein potently inhibit hippocampal long-term potentiation *in vivo*. *Nature* 416: 535–539.

Waxman EA and Lynch DR (2005) *N*-Methyl-D-aspartate receptor subtypes: Multiple roles in excitotoxicity and neurological disease. *Neuroscientist* 11: 37–49.

Wolf ME (1998) The role of excitatory amino acids in behavioral sensitization to psychomotor stimulants. *Progress in Neurobiology* 54: 679–720.

Zhuo M (2005) Targeting central plasticity: A new direction of finding painkillers. *Current Pharmaceutical Design* 11: 2797–2807.

NMDA Receptors, Cell Biology and Trafficking

R J Wenthold and R S Petralia, National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

The *N*-methyl-D-aspartate (NMDA) receptor is present at the postsynaptic membrane of nearly all glutamatergic synapses, where it is found with other glutamate receptors, particularly α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors. It is a key player in synaptic plasticity, and its malfunction has been implicated in a wide range of neurological and psychiatric disorders ranging from ischemia to schizophrenia. It is unique among neurotransmitter receptors since it requires activation by two agonists, glutamate and glycine, which interact with the NR2 and NR1 subunits, respectively. The ion channel of the NMDA receptor is normally blocked by magnesium, and the channel opens only after the neuron is depolarized by activation of AMPA receptors, thus allowing the NMDA receptor to serve as a coincidence detector. Finally, the ion channel of the NMDA receptor is permeable to calcium, a universal intracellular signaling molecule. These properties have made the NMDA receptor one of the most intensely studied molecules in the brain. Since neurons are highly compartmentalized, forming synaptic connections with thousands of other neurons, and the NMDA receptor has a restricted distribution within a neuron, trafficking of this receptor is central to its functional regulation. Here we define trafficking as the nonrandom movement of a protein within a neuron to, or from, a site where it is required for function. This process undoubtedly depends on interactions with other molecules, particularly other proteins, lipids, sugars, and small molecules.

Structure and Subunits

Three subunits of the NMDA receptor complex have been identified. A single NR1 subunit has eight splice variants. There are four NR2 subunits (NR2A–D) and two NR3 subunits. NR2A and NR2B are the most abundant NR2 subunits, with NR2B being predominant early in development and NR2A appearing later, along with a general decrease in NR2B expression. All subunits have the same topology, with three transmembrane domains, a reentrant loop, an extracellular N-terminus, and a cytoplasmic C-terminus, which is unusually long in all NR2 subunits. Supporting several previous functional and biochemical reports, the recent determination of the crystal structure of the extracellular domains of the NMDA receptor indicates a tetrameric protein made up of two NR1 and two NR2 subunits. In a tetramer, the NR2 subunits can be the same, such as two NR2B subunits, or different, such as one NR2B and one NR2A, in this complex. While NR3 can also assemble into complexes with NR1 and NR2, the function of NR3 remains unclear, and it has been reported that NR3 can assemble with NR1 alone to produce a functional glycine receptor. The NMDA receptor subunits interact with a large number of other proteins, which modify their function and influence their trafficking (Table 1).

The subunit composition determines the functional properties of the NMDA receptor. This is dependent mostly on the NR2 subunits forming receptors, with kinetics ranging from the slowest (containing NR2D) to the fastest (NR2A). Thus, channels with faster kinetics allow less calcium to enter the neuron, and would probably be less effective in developing plasticity. When NR3 is assembled into a complex with NR1 and NR2, it decreases channel function. NR3A knockout mice are normal, but show enhanced NMDA responses early in development.

NMDA Receptors in the Endoplasmic Reticulum

All NR2 subunits and some NR1 subunits, depending on the splice variant, are retained in the endoplasmic reticulum (ER) in their unassembled form and are released after assembly (Figure 1). This is a standard mechanism of quality control to assure that unassembled or misfolded proteins do not reach the cell surface. For the NR1 subunit, ER retention is controlled by an RXR motif (X indicates any nonacidic amino acid) in the C-terminal domain of the major splice variant, NR1-1. NR1 splice variants (NR1-2 and NR1-4) that lack this retention motif are not retained, and others (NR1-3) that contain this motif plus the C2′ cassette are also not retained. Thus, only the NR1-1 splice variant is retained in the ER; this is the major splice variant, accounting for about 60% of the total NR1. The C2′ cassette causes the subunit to terminate in the amino acid sequence – STVV, and this motif is required for overriding the RXR retention motif. This is a postsynaptic density/disc/zonula occludens-1 (PDZ) binding motif and also a site of interaction for components of the coat protein complex II (COPII),

Table 1 Proteins that bind directly to NR1 and/or NR2

NR1[a]	*NR2*
PSD-95 (to NR1-3, NR1-4)	PSD-95
SAP102 (to NR1-3, NR1-4)	SAP102
PSD-93 (to NR1-3, NR1-4)	PSD-93
SAP97 (to NR1-3, NR1-4)	SAP97
α-Actinin	α-Actinin
Tubulin	Tubulin
Spectrin	Spectrin
Myosin regulatory light chain	Myosin regulatory light chain
Dopamine D1 receptor	Dopamine D1 receptor
CaMKII	CaMKII
NADH dehydrogenase subunit 2 (ND2)[b]	NADH dehydrogenase subunit 2 (ND2)[b]
Calmodulin	S-SCAM (= MAGI-2)
Neurofilament-L	CIPP
Yotiao	mLin-7 (= Veli = Mals)
EphB receptors	Phospholipase C-γ
Apolipoprotein E receptor 2	Rack1
SALM1	μ subunit of adaptor protein complexes (AP1–4)
Sec23/24 of COPII	α1-Chimerin
	RasGRF-1
	Cyclin-dependent kinase-5 (cdk5)[b]

[a] NR1-3 and NR1-4 are splice variants.
[b] cdk5 may bind directly to NR2A, and ND2 appears to bind Src to NMDA receptors, but the exact nature of these interactions has not been determined.

which is involved in ER-to-Golgi trafficking. Thus, the PDZ or COPII interaction, or both, may control the ER export of these NR1 splice variants and override the retention that is mediated by the RXR motif. It remains unclear why some NR1 splice variants can exit the ER unassembled, or if they actually do in neurons, since there is no evidence that NR1 homomeric complexes exist in neurons or serve any function. NR1 subunits lacking retention motifs may be more readily exported from the ER assembled with NR2.

The mechanism by which NR2 subunits are retained in the ER is more complex than that of NR1 and not well characterized. There appears to be an ER retention signal(s) in the C-terminus although the specific site has not been identified. There is also an additional mechanism for ER retention since deletion of the C-terminus does not allow the remainder of the molecule to reach the cell surface unassembled, as is the case for NR1; however, this C-terminus-deleted construct can assemble with NR1 to produce functional receptors. Equally poorly understood is the mechanism by which the retention signals on NR1 and NR2 are negated to allow the assembled complex to exit the ER. Mutual masking of the retention motifs in the C-terminus, as has been demonstrated for some other proteins, does not

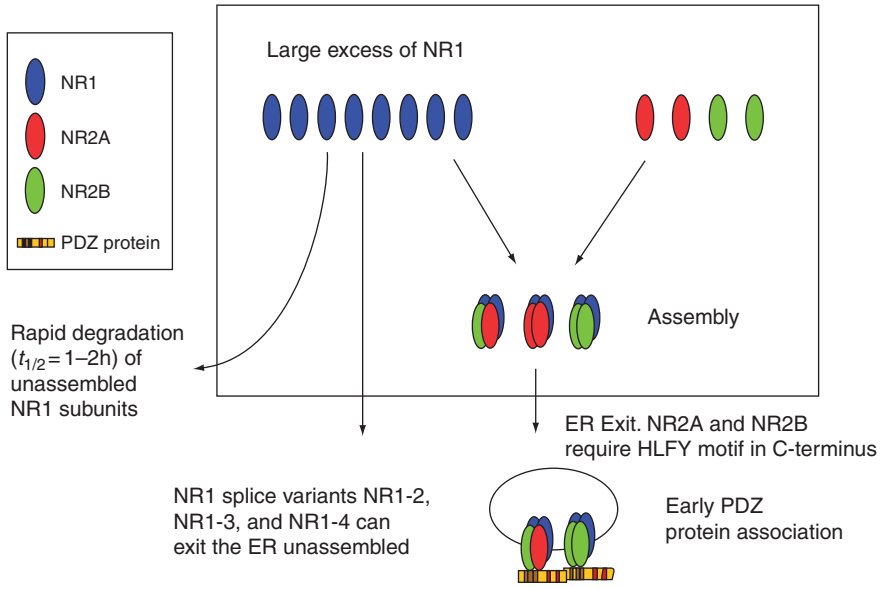

Figure 1 Assembly of the NMDA receptor in the ER. All NR2 subunits and the NR1-1 splice variant are retained in the endoplasmic reticulum (ER). Assembly negates retention and allows exit from the ER. NR1 subunits are synthesized in excess of NR2, and most are retained in the ER and rapidly degraded. In heterologous cells, some splice variants that lack the ER retention signal present in the C1 cassette, or contain the C2′ cassette, can exit the ER unassembled, probably as homodimers. It is not known if this is also the case in neurons or if it is of functional significance. Assembly can lead to receptor complexes with two identical or two different NR2 subunits. Receptors can assemble and be functional in the ER based on MK801 binding, but export requires a four-amino-acid segment (HLFY in NR2A and NR2B) immediately following transmembrane domain 4.

appear to play a role in the NMDA receptor, since subunits with a deletion of either the NR1 or NR2 C-terminal tails alone can form functional receptors. An interesting finding that further complicates our understanding of the ER retention of NR2 is that the NR2 subunit does not need to be intact to form functional receptors. The NR2A subunit can be divided into two polypeptides in the extracellular loop between transmembrane domains 3 and 4 (TM3 and TM4). Co-expression of these two fragments with NR1 produces functional receptors on the cell surface. A single neuron can contain a variety of NMDA receptor complexes (such as NR1/NR2B, NR1/NR2A, and NR1/NR2A/NR2B), and the question arises whether or not the formation of these complexes simply reflects the availability of subunits, or is regulated in some way. There is a large excess of NR1 subunits synthesized, which would ensure that most NR2 subunits would find an NR1 binding partner and form a functional receptor. This fits with the fact that the particular NR2 subunit controls the properties of the receptor.

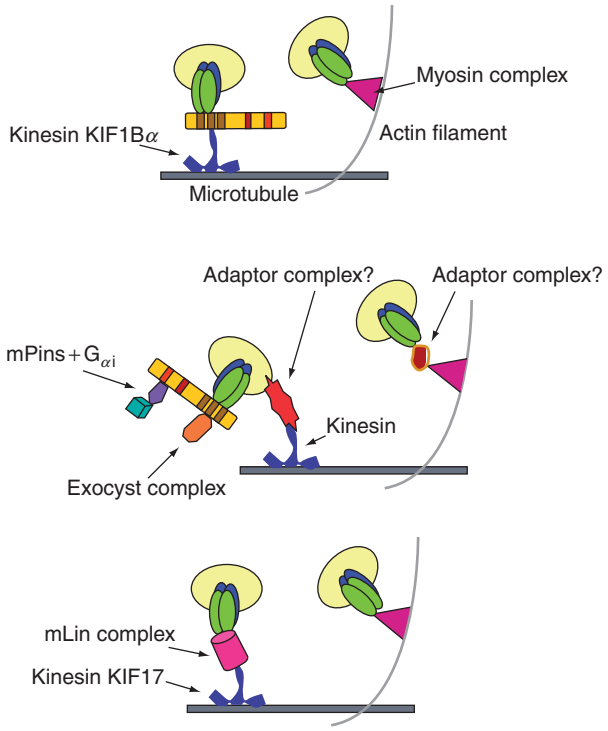

Figure 2 Proposed complexes involved in the transport of NMDA receptors. Two kinesins that indirectly interact with NMDA receptors are KIF17 and KIF1Bα, and additional kinesins may also play a role. Delivery of receptors to spines is likely to involve myosins, which remain to be identified. Both Sec8, probably in a complex with other exocyst proteins, and mPins bind to SAP102 and other related PDZ proteins (MAGUKS) and may target the NMDA receptor/PDZ protein complexes to the plasma membrane. Also, G-protein signaling via the interaction of mPins and $G_{\alpha i}$ may regulate the effect of mPins on NMDA receptor trafficking and postsynaptic spine structure.

Trafficking of NMDA Receptors in the Dendrite

NMDA receptors are concentrated at postsynaptic sites on dendrites while synthesis occurs predominantly in the cell body, requiring a mechanism to transport receptors from the cell body to synapses. This could be done in two ways. Receptors could be delivered to the plasma membrane at the cell body, and the receptors could reach the dendrite by lateral diffusion or through some form of regulated transport while they are in the plasma membrane. Alternatively, receptors can be delivered to dendritic sites through association with intracellular transport vesicles. The latter is assumed to be the major mechanism of delivery, but lateral movement to the synapse is probably also necessary for movement into the synapse after delivery to an extrasynaptic location (Figure 2). NMDA receptors can associate early in the biosynthetic pathway with PDZ proteins, including members of the PSD-95 family of membrane-associated guanylate kinases (MAGUKs), which are transported in the dendrite as a complex with the NMDA receptors. The MAGUK allows the indirect association of the NMDA receptors with multiple other proteins that can influence the trafficking of the NMDA receptor. The exocyst, a complex of eight proteins, associates with the receptor through a PDZ interaction with the PSD-95-family MAGUK, SAP102 (synapse-associated protein), and perhaps with other PDZ proteins; disruption of this interaction will dramatically reduce the surface expression of receptor. Another trafficking molecule, mPins/LGN (mammalian homolog of *Drosophila* partner of inscuteable) interacts with the receptor complex through the Src homology 3/guanylate kinase (SH3/GK) motifs of the MAGUK and disruption of this interaction also reduces NMDA receptor surface delivery. mPins/LGN interacts with a number of additional proteins, including $G_{\alpha i}$, suggesting a possible link between G-protein signaling and NMDA receptor trafficking.

Transport vesicles carrying NMDA receptors appear to be largely distinct from those carrying AMPA receptors, suggesting that these two receptor populations are packaged into distinct populations of transport vesicles as they exit the *trans*-Golgi network (TGN). In young neurons transport in dendrites involves multiple rounds of endo- and exocytosis with the plasma membrane and association of the transport vesicles with SAP102. Two motor systems

have been identified in the transport of NMDA receptors and MAGUKs. The kinesin KIF17 associates indirectly with NMDA receptors through a complex with mLin-7, mLin-2, and mLin-10 and may mediate the transport of NMDA receptors along microtubules. MAGUKs can associate directly with the kinesin KIF1Bα, making it a candidate for transporting NMDA receptors associated with SAP102. Myosin motors, which can mediate transport of proteins along actin filaments, are also likely to play a role. NMDA receptors are regulated by myosin light chain kinase and show a direct interaction with myosin regulatory light chain. The NMDA receptor, therefore, is transported as part of a macromolecular complex. However, unlike the presynaptic active zone, the postsynaptic density, a structure with which the NMDA receptor is associated at synapses, is not transported as a completely preassembled complex. The NMDA receptor complex may remain intact as it is delivered to the synapse, but a more likely scenario is that the addition and removal of proteins are a normal parts of the trafficking process to change parameters as the local environment changes.

Trafficking of NMDA Receptors at the Synapse

Synaptic Delivery and PDZ Proteins

The steps involved in delivering NMDA receptors present in transport vesicles in the dendrite to the synapse are largely unknown. This process may involve an intermediate stop in an endosomal compartment, where they may mix with recycling receptors, or direct delivery to the plasma membrane. The site of delivery is also unknown but is likely adjacent to the synapse rather than directly at the synapse. This will require movement within the plasma membrane, and these receptors may represent the extrasynaptic pool of NMDA receptors (discussed later). The number and composition of NMDA receptors at the synapse are relatively stable compared to those of AMPA receptors. However, the mechanism regulating synaptic NMDA receptors is not well understood. The number of receptors is not based on availability, since overexpression of NR2 subunits does not increase the synaptic number of receptors, but does increase the extrasynaptic number, indicating a local mechanism of control for synaptic receptors. The NR2/PSD-95 interaction was the first PDZ interaction identified (subsequent studies have shown that the other three members of the PSD-95 family of MAGUKs, SAP102, SAP97, and PSD-93, also can interact with the NR2 subunits), and it generally has been assumed that synaptic NMDA receptors are tethered to the PSD through this PDZ interaction. A number of studies have supported this by showing that the PDZ binding domain is required for an NR2 subunit to be admitted to the synapse. However, recent studies have shown that the NMDA receptor/PSD-95 interaction is not straightforward. Knockout of PSD-95 or its overexpression does not change synaptic NMDA receptors, but overexpression does increase AMPA receptors. Studies have also shown that an NR2A subunit lacking its PDZ binding motif can enter the synapse, although NR2B subunits without the PDZ binding motif cannot.

A variety of adhesion proteins are found associated with NMDA receptors at synapses and these proteins may help regulate NMDA receptor distribution and function. Postsynaptic EphB2 receptors are activated by their presynaptic ligand, ephrin-B, and bind directly to the extracellular domain of NR1, affecting NMDA receptor function directly and subsequent dendrite arborization indirectly. SALM1 can bind to the extracellular domain of NR1 and this and some other SALMs also can associate with NMDA receptors via a mutual PDZ-mediated binding to MAGUKs of the PSD-95 family. SALM1 can enhance surface expression of transfected NR2A. A number of other adhesion factors do not bind directly to NMDA receptors but instead form associations with them at the synapse and may affect NMDA receptor trafficking and activation and subsequent neuronal structure and function; examples are neuroligin, cadherin/catenin, and a number of members of the L1/NrCAM and NCAM (neural cell adhesion molecule) families of adhesion proteins. Neuroligin binds to neurexin in the presynaptic terminal and to PSD-95 and other NMDA receptor-associated MAGUKs in the postsynaptic density. Neuroligin localization at synapses may help determine if developing synapses become excitatory or inhibitory. In fact, a number of adhesion proteins such as neuroligin, cadherin, NCAM, and EphB receptors may mediate the earliest stages of synaptogenesis that are required for subsequent trafficking of MAGUKs and NMDA receptors to the nascent postsynaptic membrane.

Clathrin-Mediated Internalization and Regulation of NMDA Receptors

The number of molecules of any protein on the cell surface can be regulated by internalization, and several studies have now shown that a number of factors can increase the internalization of NMDA receptors. These include synaptic activity, activation of type 1 metabotropic receptors, and exposure to the agonist glycine. In addition, run down of extrasynaptic receptors appears to depend on internalization.

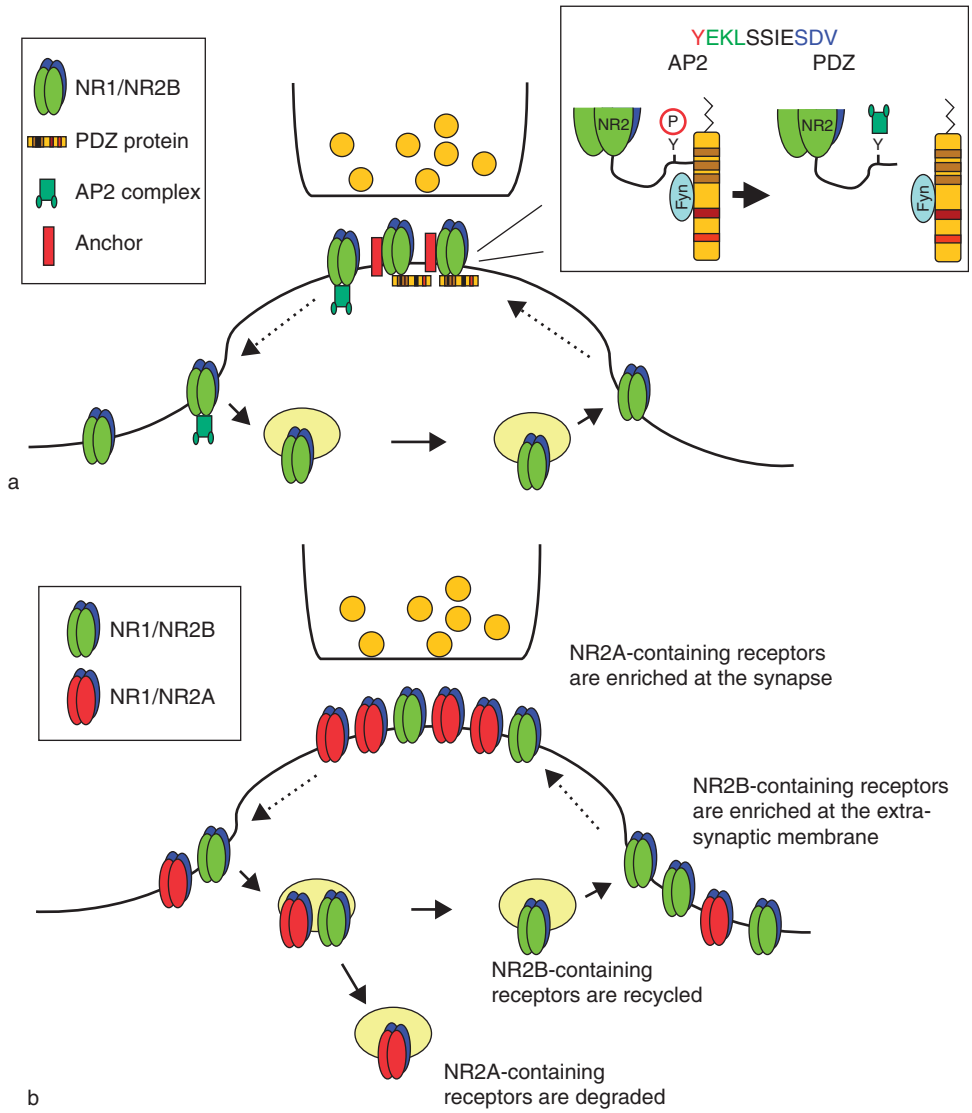

Figure 3 (a) Model for regulation of NR2B-containing receptors at the synapse. Internalization is mediated by the interaction of AP2 with the YEKL motif in the distal C-terminus of NR2B. At the synapse, Y1472 in this motif is phosphorylated by Fyn kinase and is unable to interact with AP2. Fyn is recruited by PSD-95, which associates with NR2B through its PDZ binding domain. Loss of phosphorylation of Y1472 facilitates the interaction with AP2 and leads to removal from the synapse and internalization. Retention of NMDA receptors at the synapse may involve additional anchors. (b) NR2A- and NR2B-containing receptors are regulated differently at the synapse. Several studies have suggested that NR2A-containing receptors are more abundant at the synapse while NR2B-containing receptors are more abundant at extrasynaptic sites. Several factors could contribute to this differential localization. Receptors can enter the synapse by lateral diffusion in the plasma membrane. A filter may selectively exclude NR2B-containing receptors from the synapse. Alternatively, NR2B-containing receptors may be more likely to be removed from the synapse through internalization (as described in panel (a)). Finally, internalized NR2A-containing receptors are more susceptible to degradation while NR2B-containing receptors are recycled. If they enter the extrasynaptic pool, the recycled receptors would contribute to the enrichment of NR2B-containing receptors in this pool.

NMDA receptors can be removed from the plasma membrane through a clathrin-dependent route involving the adaptor protein AP2 (Figure 3). Clathrin-dependent internalization requires the recognition of a particular motif in its cytoplasmic domain by the adaptor protein. The NR2B subunit has such a motif, YEKL, near its C-terminus, that has been shown to play a role in the surface stability of the receptor. Interestingly, a similar motif in the NR2A subunit is not involved in the internalization of this subunit, but rather an upstream dileucine motif (also a substrate for the AP2 adaptor) was shown to play such a role. The differences in the trafficking of these two subunits are highlighted in the different pathways followed by the internalized receptors. While both initially enter early endosomes, NR2B is trafficked to recycling endosomes while NR2A is trafficked to late endosomes. Therefore, NR2B-containing receptors are more likely

to be recycled and NR2A-containing receptors are more likely to be degraded after internalization.

Does internalization play a role in regulating the number of synaptic NMDA receptors? It is unlikely that internalization *per se* regulates the number of receptors, but the availability of endocytic motifs such as the YEKL motif on NR2B may. This motif contains a tyrosine, Y1472, which is a major substrate for Fyn kinase. In the striatum, synaptic membranes have a high level of phosphorylated Y1472 while intracellular membranes have a low level of phosphorylated Y1472. In cultures of cerebellar granule cells and hippocampal neurons, using transfection of NR2B subunits with mutations in this motif, it was shown that mutations that block the interaction of AP2 increase the number of synaptic receptors. Furthermore, while the PDZ binding domain of NR2B is required for synaptic localization, NR2B subunits that lack this domain and have mutations of the YEKL motif still localize to the synapse. One explanation is that the PDZ interaction may not be critical for anchoring the receptor complex to the postsynaptic protein complex, but rather plays a role in maintaining the phosphorylation of Y1472. PSD-95 has been shown to interact with Fyn kinase, and, at the synapse, PSD-95 may recruit Fyn kinase, thus maintaining the phosphorylation of Y1472 and preventing its interaction with AP2 and its subsequent internalization. The fact that the internalization of NR2A is not affected by mutations in a homologous motif, YKKM, indicates that different factors control the surface stability of NR2A-containing receptors.

The YEKL and dileucine motifs near the distal end of the C-terminus of NR2B and NR2A, respectively, are not the only endocytic motifs found on the NMDA receptor. Another tyrosine-containing motif is present in the proximal part of the C-terminus, near the last transmembrane domain. This motif was functionally characterized in receptors with the NR2A subunit, but it is present in all NR2 subunits and the NR1 subunit. Since mutation of neither this motif nor the distal motif can block all internalization, there are likely other motifs present in the NR1 and NR2 C-termini that mediate clathrin-dependent internalization. Nonclathrin mechanisms may also exist.

Extrasynaptic Receptors

NMDA receptors are found not only at the postsynaptic membrane but have also been reported on glia, presynaptic terminals, the perisynaptic membrane, and the extrasynaptic membrane. Most extensively studied are perisynaptic and extrasynaptic receptors, which may be functionally important and activated when glutamate that is released at the synapse is not sufficiently sequestered and is able to diffuse outside the synapse. While most neurons contain both synaptic and extrasynaptic receptors, in a few specific cases such as the retinal ganglion cell, NMDA receptors are found only outside the synaptic cleft. In other neurons, NR2A-containing receptors are preferred at the synapse and NR2B-containing receptors are preferred at the extrasynaptic membrane, although the separation is not complete. This subunit preference may reflect functional differences between these two pools of receptors. This, however, is difficult to reconcile with studies that show diffusion of receptors from the extrasynaptic plasma membrane into the synapse. In one study, synaptic receptors were blocked with the use-dependent antagonist, MK801, and recovery of function was monitored. The recovery of function was surprisingly rapid, and controls ruled out sources other than the extrasynaptic membrane. The preferential localization of NR2A- and NR2B-containing receptors would be consistent with these findings if there was a filter that favors the entry of NR2A into the synapse or the removal of NR2B from the synapse. NMDA receptors do recycle, although not as fast as AMPA receptors, and the enrichment of NR2B in the extrasynaptic membrane could reflect this. As noted earlier, internalized NR2B receptors are more likely to be recycled and NR2A receptors are more likely to be degraded. If receptors are recycled to the extrasynaptic membrane, the difference in intracellular trafficking could favor NR2B at the extrasynaptic membrane. Such a mechanism would also assume that there is an exchange between the synaptic and extrasynaptic pools of receptors. The functional consequences resulting from activation of these receptors is also an area of intense study. Activation of extrasynaptic receptors promotes cell death. A series of controversial studies has indicated that NR2A-containing receptors mediate long-term potentiation (LTP) and that NR2B-containing receptors mediate long-term depression (LTD). This requires more work, and it will be interesting to determine if position of the receptor – for example, synaptic versus perisynaptic – plays a role.

Developmental Changes in NMDA Receptors

During development, most neurons begin expressing predominantly NR2B, which then gradually decreases as NR2A expression increases. The functional implication of this switch is that NMDA receptors in young animals have slower kinetics (due to a preponderance of NR2B), and receptors in older animals are faster (due to a preponderance of NR2A). It has been proposed that this change, in which 'adult' NMDA receptors are faster and admit less calcium, may be related

to the decrease in synaptic plasticity in older animals. While most neurons continue to express NR2B in older animals, and therefore have a mixture of NR2A- and NR2B-containing receptors and perhaps those containing both subunits, cerebellar granule cells completely stop synthesizing NR2B. During development, NR2A and NR2C expression increases in these neurons. Trafficking plays a role in the preferential localization of subunits during development. A series of elegant studies has shown that, within hours, visual experience can cause synaptic NMDA receptors in the visual cortex to become more NR2A-like. Thus, either NR2A-containing receptors are preferentially added or NR2B-containing receptors are preferentially lost, or both. While this could depend somewhat on new protein synthesis, it is unlikely that such a change could be effected in the time period, and a more likely explanation is that NR2A-containing receptors are recruited from extrasynaptic or intracellular sites. Similar transitions from NR2B- to NR2A-containing receptors may be occurring at other synapses during development, or even at mature synapses. Thus, the mechanism underlying this switch is fundamentally important to NMDA receptor trafficking.

Phosphorylation and Trafficking of NMDA Receptors

As discussed in the preceding section, Fyn kinase phosphorylation has a major impact on trafficking of the NR2B subunits. The NMDA receptor is a substrate for several other kinases, including other members of the Src family of protein tyrosine kinases, calcium/calmodulin-dependent protein kinase II (CaMKII), cyclin-dependent kinase-5 (cdk5), cyclic AMP-dependent protein kinase A (PKA), protein kinase C (PKC), casein kinase II, and others. The NMDA receptor directly interacts with CaMKII at the postsynaptic density. This interaction is important for the translocation of CaMKII to the synapse. CaMKII has been shown to interact with all three of the common subunits, NR1, NR2A, and NR2B, through their C-termini, although the affinities of the interactions and the functional implications are very different. A recent study in which the interacting domains on NR2A and NR2B were switched showed that the CaMKII interaction with the NR2B subunit is required for LTP.

Phosphorylation of NMDA receptor subunits affects the trafficking of the individual subunits. The subcellular distribution of the NR1 subunit expressed alone in heterologous cells can be regulated by PKC phosphorylation of the serine residues in the C1 domain. These amino acids are near the ER retention signal, and mutations that mimic phosphorylation greatly increased surface expression of NR1/tac chimeric constructs.

Phosphorylation by several other kinases, particularly casein kinase II and PKC, affects the trafficking of the assembled receptor. The NR2B subunit terminates in ESDV, which is a PDZ binding domain. Serine-1480, present in this domain, is a substrate for casein kinase II, and phosphorylation of Ser1480 blocks interaction with PDZ proteins. Therefore, this may be an important mechanism for regulating receptor/PDZ interactions. PKC activation increases the delivery of NMDA receptors to the cell surface.

Degradation of NMDA Receptors

While trafficking can regulate local amounts and distributions of a protein effectively, altering either synthesis or degradation can effect global changes in proteins. Like other proteins, the NMDA receptor is continually degraded and, at steady state, replenished by synthesis of new receptors. In cultured cerebellar granule cells, the half-life of assembled NMDA receptor is about 1 day. Unassembled NR1 subunits, which are synthesized in excess compared to NR2, are retained in the ER and degraded rapidly, with a half-life of 1–2 h. As presented previously, there is a developmental change in the relative amounts of NR2A and NR2B, which could be achieved by altering synthesis or degradation. Since mRNA expression often matches protein levels, it suggests that synthesis is the main controlling factor. The degradation of NMDA receptors may not be entirely random, but may depend on the subunit composition and activity of the neuron; degradation, therefore, could change the subunit balance or the total number of receptors at a particular synapse. As noted above, receptors containing NR2A or NR2B follow different pathways after internalization. Recent studies have shown that ubiquitination plays a major role in the removal of postsynaptic proteins, including NMDA receptors. Covalent attachment of ubiquitin can cause internalization of membrane proteins and degradation in the proteasome. Several proteins of the PSD are ubiquinated, including PSD-95. NR1 can be ubiquitinated on its extracellular domain. In this case, activity-dependent NMDA receptor degradation may be mediated by retrotranslocation followed by ubiquitination, although the mechanism is not well understood.

See also: AMPA Receptors: Disease; Clathrin and Clathrin-Adaptors; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA

Receptor Role; NMDA Receptors and Development; NMDA Receptors and Disease + C464; NMDA Receptor Function and Physiological Modulation; Postsynaptic Density/Architecture at Excitatory Synapses.

Further Reading

Ali DW and Salter MW (2001) NMDA receptor regulation by Src kinase signaling in excitatory synaptic transmission and plasticity. *Current Opinions in Neurobiology* 11: 336–342.

Barria A and Malinow R (2002) Subunit-specific NMDA receptor trafficking to synapses. *Neuron* 35: 345–353.

Barria A and Malinow R (2005) NMDA receptor subunit composition controls synaptic plasticity by regulating binding to CaMKII. *Neuron* 48: 289–301.

Dunah AW and Standaert DG (2001) Dopamine D1 receptor-dependent trafficking of striatal NMDA glutamate receptors to the postsynaptic membrane. *Journal of Neuroscience* 21: 5546–5558.

Ehlers MD, Tingley WG, and Huganir RL (1995) Regulated subcellular distribution of the NR1 subunit of the NMDA receptor. *Science* 269: 1734–1737.

Kornau HC, Schenker LT, Kennedy MB, et al. (1995) Domain interaction between NMDA receptor subunits and the postsynaptic density protein PSD-95. *Science* 269: 1737–1740.

Lavezzari G, McCallum J, Dewey CM, et al. (2004) Subunit-specific regulation of NMDA receptor endocytosis. *Journal of Neuroscience* 24: 6383–6391.

Li B, Chen N, Luo T, et al. (2002) Differential regulation of synaptic and extra-synaptic NMDA receptors. *Nature Neuroscience* 5: 833–834.

McIlhinney RAJ, LeBourdelles B, Molnar E, et al. (1998) Assembly intracellular targeting and cell surface expression of the human *N*-methyl-D-aspartate receptor subunits NR1a and NR2A in transfected cells. *Neuropharmacology* 37: 1355–1367.

Mu Y, Otsuka T, Horton AC, et al. (2003) Activity-dependent mRNA splicing controls ER export and synaptic delivery of NMDA receptors. *Neuron* 40: 581–594.

Nong Y, Huang YQ, Ju W, et al. (2003) Glycine binding primes NMDA receptor internalization. *Nature* 422: 302–307.

Perez-Otano I and Ehlers MD (2005) Homeostatic plasticity and NMDA receptor trafficking. *Trends in Neuroscience* 28: 229–238.

Prybylowski K, Chang K, Sans N, et al. (2005) The synaptic localization of NR2B-containing NMDA receptors is controlled by interactions with PDZ proteins and AP-2. *Neuron* 47: 845–857.

Quinlan EM, Philpot BD, Huganir RL, et al. (1999) Rapid, experience-dependent expression of synaptic NMDA receptors in visual cortex *in vivo*. *Nature Neuroscience* 2: 352–357.

Rao A and Craig AM (1997) Activity regulates the synaptic localization of the NMDA receptor in hippocampal neurons. *Neuron* 19: 801–812.

Roche KW, Standley S, McCallum J, et al. (2001) Molecular determinants of NMDA receptor internalization. *Nature Neuroscience* 4: 794–802.

Rumbaugh G and Vicini S (1999) Distinct synaptic and extrasynaptic NMDA receptors in developing cerebellar granule neurons. *Journal of Neuroscience* 19: 10603–10610.

Sans N, Prybylowski K, Petralia RS, et al. (2003) NMDA receptor trafficking through an interaction between PDZ proteins and the exocyst complex. *Nature Cell Biology* 5: 520–530.

Sans N, Wang PY, Du Q, et al. (2005) mPins, the mammalian homologue of *Drosophila* Pins, modulates PSD-95 and SAP102 trafficking and influences NMDA receptor surface expression. *Nature Cell Biology* 7: 1179–1190.

Setou M, Nakagawa T, Seog DH, et al. (2000) Kinesin superfamily motor protein KIF17 and mLin-10 in NMDA receptor-containing vesicle transport. *Science* 288: 1796–1802.

Sprengel R, Suchanek B, Amico C, et al. (1998) Importance of the intracellular domain of NR2 subunits for NMDA receptor function *in vivo*. *Cell* 92: 279–289.

Standley S, Roche KW, McCallum J, et al. (2000) PDZ-domain suppression of an ER retention signal in NMDA receptor NR1 splice variants. *Neuron* 28: 887–898.

Tovar KR and Westbrook GL (2002) Mobile NMDA receptors at hippocampal synapses. *Neuron* 34: 255–264.

Vissel B, Krupp JJ, Heinemann SF, et al. (2001) A use-dependent tyrosine dephosphorylation of NMDA receptors is independent of ion flux. *Nature Neuroscience* 4: 587–596.

Washbourne P, Bennett JE, and McAllister AK (2002) Rapid recruitment of NMDA receptor transport packets to nascent synapses. *Nature Neuroscience* 5: 751–759.

Wenthold RJ, Prybylowski K, Standley S, et al. (2003) Trafficking of NMDA receptors. *Annual Reviews of Pharmacology and Toxicology* 43: 335–358.

Nociceptor Responses

C Belmonte and F Viana, Universidad Miguel Hernández-Consejo Superior de Investigaciones Científicas, San Juan de Alicante, Spain

© 2009 Elsevier Ltd. All rights reserved.

Nociceptors are peripheral terminals of a subpopulation of primary sensory neurons located in the dorsal root and cephalic ganglia that are functionally specialized in the detection of potentially or overtly injurious stimuli acting on bodily tissues (mechanical, chemical, and thermal). They can be considered damage-sensing terminals that encode the intensity and duration of noxious stimuli into a discharge of nerve impulses that is transmitted to upper levels of the central nervous system, ultimately evoking sensations of discomfort and pain. The name 'nociceptor' was coined by Sherrington a century ago to emphasize the noxious nature of their adequate stimulus, in contrast with the low-threshold mechano- and thermal receptors that signal innocuous forces and serve to detect non-noxious changes in the environment.

Somatic nociceptors are formed by naked (free) nerve endings of thinly myelinated (Aδ) or unmyelinated (C) sensory afferents that branch extensively within the skin, mucosae, muscles, and joint tissues. In addition, other population of Aδ and C nociceptor axons innervate the viscerae, preferentially hollow viscerae from the cardiovascular, respiratory, genitourinary, and digestive tract. The cell bodies of neurons responding to noxious stimuli of somatic tissues are exclusively located in dorsal root and cephalic ganglia, whereas those innervating the viscerae are also found in the nodose ganglion.

Nociceptors innervating the body surface have received comparatively more experimental attention than those innervating deep tissues. Cutaneous nociceptors are distinct from low-threshold receptors subserving innocuous skin sensations not only because of the high intensity of the physical force required for their stimulation – several orders of magnitude higher than in the other receptor classes – but also by their unique response to repeated stimuli. In low-threshold receptors, sustained or repetitive stimulation tends to reduce the magnitude of the evoked impulse firing (i.e., they develop adaptation and fatigue). Some nociceptors also fatigue immediately upon repeated stimulation but a few minutes afterwards they augment their response to new noxious stimuli. This process, called 'sensitization,' is characterized by a reduction of the impulse firing threshold such that propagated impulse discharges are now evoked by intensities of the stimulus within the innocuous range. Moreover, noxious stimuli elicit a stronger and more sustained impulse discharge in sensitized nociceptors (Figure 1). Finally, sensitized nociceptors often develop an irregular, low-frequency ongoing activity in the absence of stimulation. This spontaneous activity, coupled with the enhanced responsiveness to stimulating forces, is the peripheral substrate of the skin tenderness to mechanical or thermal stimulation and of the presence of pain sensation in the absence of stimulus, exhibited by injured peripheral tissues ('primary hyperalgesia').

Functionally, nociceptor neurons constitute a heterogeneous population. They differ in the presence or absence of myelin at their peripheral axons and also in the variable expression of integral membrane proteins such as certain transducer ion channels (TRPs, ASICs, P2X, and 2-pore domain K^+ channels) and receptor proteins for endogenous mediators. The content of certain enzymes, such as fluoride-resistant acid phosphatase, neurofilament 200, and certain neuropeptides (substance P, calcitonin gene-related peptide (CGRP), and somatostatin), also varies among nociceptive neurons. The structural and molecular diversity of nociceptors is reflected in differences in their preferred responses to certain forms of stimulating energy and the characteristics of their firing patterns. Such differences in responsiveness have been used to distinguish different functional subclasses of nociceptors (Figure 1).

Most information concerning the firing pattern characteristics of the different classes of nociceptor fibers has been obtained from electrophysiological recordings in anesthetized animals. However, the development of recording techniques for single fibers in sensory nerves innervating leg or arm skin in human subjects enabled researchers to correlate impulse activity with conscious sensations, confirming the fundamental analogy between nociceptors of humans and those of other mammalian species.

Functional Types of Nociceptors

Somatic Nociceptors

Nociceptors innervating superficial and deep somatic tissues have been classified by their responsiveness to mechanical forces, heat, and exogenous irritant chemicals into three main groups:

1. Mechano-nociceptors that respond initially only to mechanical forces ranging from moderate pressure with a blunt object to overtly tissue-damaging stimuli with a short discharge of impulses whose frequency increases with the intensity of the stimulus

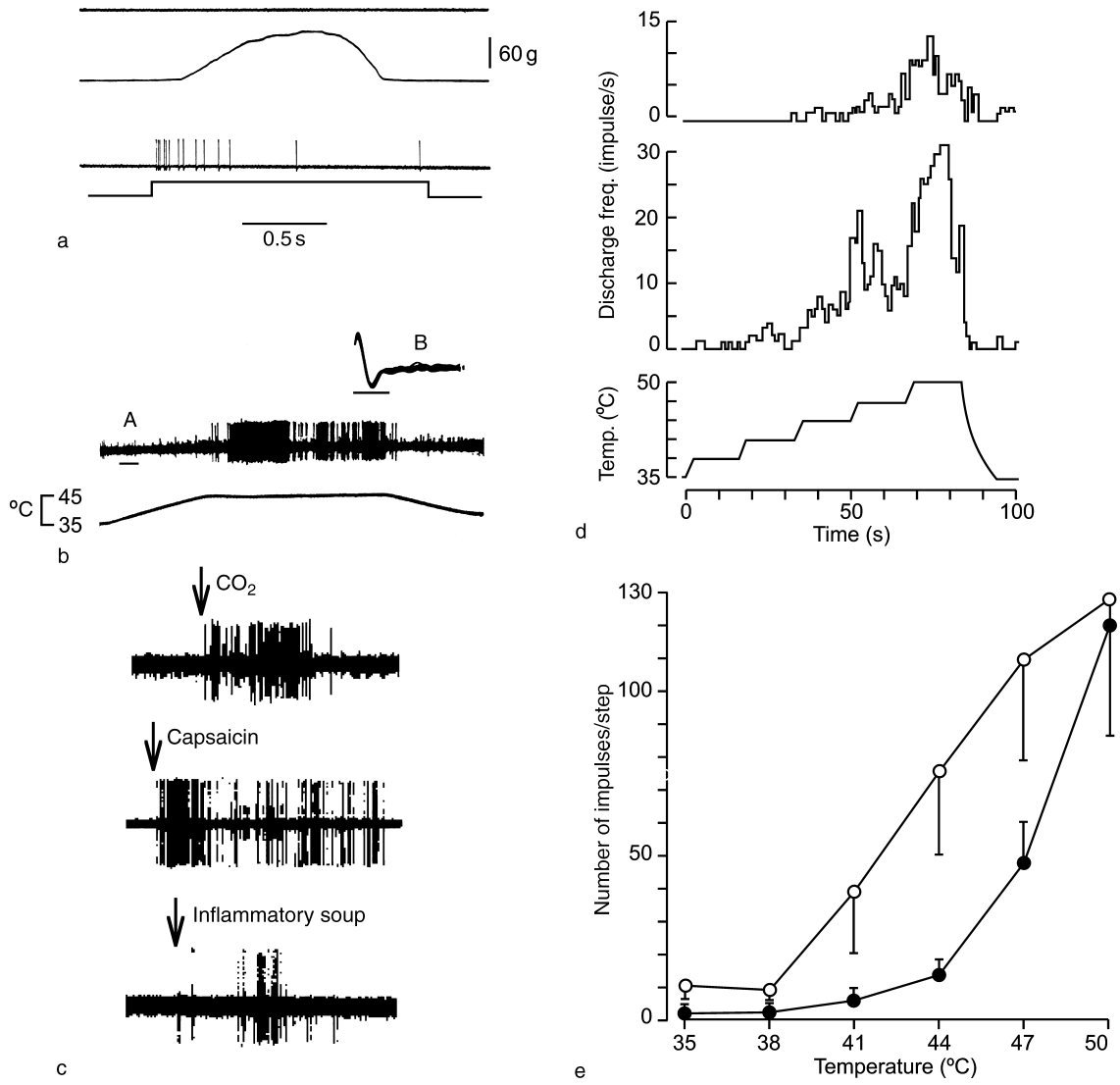

Figure 1 (a) Extracellular recording of impulse activity of a myelinated cutaneous mechano-nociceptor fiber. The top trace shows the lack of response to squeezing of the skin with blunt forceps. The middle trace shows the calibrated force record. The bottom trace shows the response to a pinch applied with serrated forces. (b) Activity of an Aδ polymodal nociceptor of the cornea evoked by heating of the corneal surface to 45 °C. (c) Response of polymodal nociceptor fibers of the cornea to acid generated by application of a CO_2 gas pulse (top trace), the irritant chemical capsaicin (middle trace), or a mixture of inflammatory mediators (bottom trace). (d) Sensitization by repeated heating of a polymodal nociceptor of the cornea; histograms show the first (top) and second (bottom) impulse responses to two identical stepwise heatings separated by 3 min. (e) Mean stimulus–response relation of eight corneal polymodal nociceptive fibers in response to first (solid circles) and second (open circles) stepwise heatings. (a) Reproduced from Burgess PR and Perl ER (1967) Myelinated afferent fibres responding to noxious stimulation of the skin. *Journal of Physiology* 190: 541–562, with permission from Blackwell. (b,c) Reproduced from Belmonte C and Galler J (1996) Corneal nociceptors. In: Belmonte C and Cervero F (eds.) *Neurobiology of Pain*, pp. 146–183. New York: Oxford University Press; and Belmonte C and Galler J (2000) The primary nociceptive neuron: A nerve cell with many functions. In: Rowe MJ and Iwamura Y (eds.) *Somatosensory Processing: From Single Neuron to Brain Imaging*, pp. 27–48. Amsterdam: Harwood, with permission. (d,e) Reproduced from Belmonte C and Giraldez F (1981) Responses of cat corneal sensory receptors to mechanical and thermal stimulation. *Journal of Physiology* 321: 355–368, with permission from Blackwell.

(Figure 1(a)). They belong generally to the Aδ fiber type, although mechano-nociceptors with C fibers have also been described.

2. Polymodal nociceptors that are activated by intense mechanical deformation producing an irregular tonic discharge roughly proportional to stimulus intensity that often outlasts the stimulus; they also respond with a sustained impulse activity to diluted acid or other irritant chemical stimuli, to heating of the skin over 40 °C, and, occasionally, to noxious cold, and they sensitize to repeated stimuli (Figures 1(b) and 1(c)). Polymodal nociceptors belong mainly to the class of C fibers, although some of them are thin myelinated, particularly in the trigeminal system.

3. Silent nociceptors, which are Aδ or C fibers characterized by their complete insensitivity or very high threshold to mechanical stimulation in the intact skin, responding only when the supporting tissue (skin and joint) is stimulated with sustained pressure or becomes inflamed (Figure 2).

Experimental studies have used mainly mechanical and thermal stimuli for nociceptor identification; for this reason, mechano- and polymodal nociceptors have also been classified as mechanically sensitive afferents (MSAs), in contrast with silent nociceptors, which form the group of mechanically insensitive afferents (MIAs) that apparently represent more than 30% of nociceptors in primates.

Within each functional type of nociceptors, those with Aδ or C fibers exhibit a number of differences. In general, nociceptors with Aδ axons respond faster and more vigorously to stimuli. Accordingly, Aδ MSA nociceptors (mechano- and polymodal) appear to be responsible for the initial sharp pricking pain experienced when a punctuate, noxious mechanical stimulus is applied to the skin, whereas long-lasting activity of MSA C polymodal nociceptor fibers generates the ensuing, persisting pain. When active, silent C nociceptors (MIAs) appear to be gradually recruited by maintained pressure and also contribute to the building pain generated by the stimulus. Likewise, the initial sharp pain sensation experienced with noxious heat is caused by nerve impulses in a subtype of fast Aδ polymodal nociceptors (type 1), followed by a transient silence and the building impulse discharge of C fibers that leads to burning pain. However, persistent heating causes adaptation of C fibers while it increases gradually in a second type of Aδ polymodal fibers (type 2), suggesting that these nociceptors maintain pain caused by continuous heat.

Visceral Nociceptors

Receptors innervating tissues of the respiratory, cardiovascular, genitourinary, and gastrointestinal systems also belong to the Aδ and C groups. Most are part of the peptide-containing class of nociceptors. Compared to those of somatic origin, the proportion of visceral afferent nerve fibers is low, which explains the diffuse nature of visceral painful sensations. Many visceral nociceptors appear to be polymodal. In hollow viscerae, nociceptors terminate in muscle, mucosa, and serosal layers. Mechanical forces, such as distension or direct prodding, and thermal and chemical stimuli have been used to characterize the

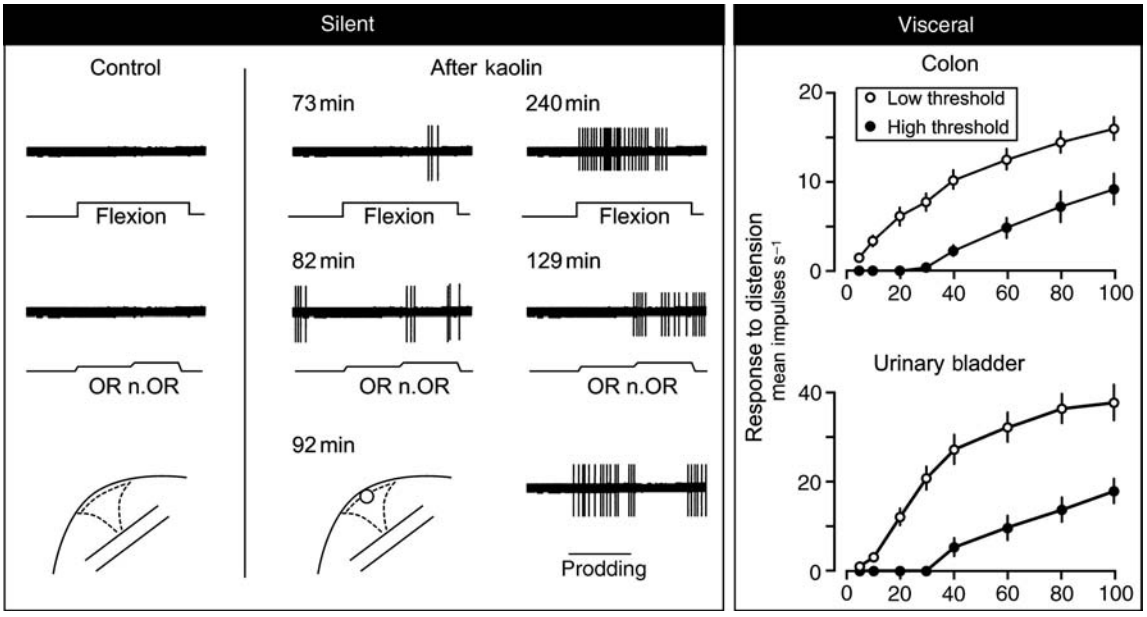

Figure 2 (Left) Response of a silent nociceptor fiber innervating the cat knee joint after experimental induction of inflammation by intra-articular injection of kaolin–carrageenan. Prior to inflammation, this fiber was recruited only with electrical stimulation of its receptive field but was unresponsive to mechanical stimulation performed by controlled flexion, nonnoxious (OR), and noxious (n.OR) outside rotation or to direct mechanical stimulation of the receptive field. (Right) Mean impulse response of low- and high-threshold pelvic visceral receptor fibers innervating the colon (top) and the urinary bladder (bottom) to graded distension of both organs, expressed as the mean increase in impulses/second over resting activity. (Left) Reproduced from Schaible H-G and Schmidt RF (1998) Time course of mechanosensitivity changes in articular afferents during a developing experimental arthritis. *Journal of Neurophysiology* 60: 2180–2195, used with permission from the American Physiological Society. (Right) Reproduced from Sengupta JN and Gebhart GF (1995) Mechanosensitive afferent fibres in the gastrointestinal and lower urinary tracts In: Gebhart GF (ed.) *Progress in Pain Research and Management: 5. Visceral Pain*. Seattle: IASP Press, with permission from International Association for the study of pain.

response properties of these receptors. Some of them, predominantly located in muscular layers, respond to low distending pressures but increase the firing frequency of their tonic discharge when noxious pressure levels are attained. They have been classified as low-threshold or intensity-encoding mechanoreceptors. High-threshold mechanoreceptors, responding only to distension that reaches the noxious range, are less abundant (approximately 20% of the afferents exhibiting mechanosensitivity). It has been suggested that low-threshold visceral receptors evoke sensations that evolve from nonpainful to painful when the intensity of the stimulus increases gradually, such as when the bladder becomes increasingly filled, whereas high-threshold visceral receptors activated only by overtly noxious stimuli mediate distinct pain sensations (Figure 2).

Transduction Mechanisms in Nociceptors

The detection of stimuli by nociceptor neurons is based on membrane signaling molecules which convert the stimulus energy into a conformational change, leading to an alteration in ionic permeability and depolarization of the nerve terminal. In nociceptors, most transduction molecules are ion channels that are directly gated by the stimulus or by intracellular messenger systems activated by a variety of chemical substances. Several classes of ion channels have been associated with the transduction of the various forms of energy and the production of generator potentials at nociceptor nerve terminals (Figure 3).

The nature of the molecular entities involved in the transduction of mechanical forces by nociceptors is uncertain. Extracellular matrix attachments have been proposed as the mechanism involved in the transmission of external mechanical forces to the neuronal surface and, subsequently, to ion channels activated or inactivated by stretch. In a second class of mechanosensitive channels, lateral forces in the membrane bilayer can directly gate the channel. Several classes of ion channels have been implicated in nociceptor mechanotransduction, but evidence from *in vivo* experiments is lacking in all cases. These include various members of the TRP superfamily of ion channels (Figure 3), such as TRPV2 and TRPV4, TRPC1, and TRPA1. Studies in mutant mice suggest that ASIC1–3 sodium channel subunits of the DEG/ENaCs superfamily contribute to mechanosensation, but the phenotypes are mild and restricted to select populations of afferents. The ionotropic ATP receptor P2X3 is located in sensory terminals and epithelial cells of hollow organs, such as the bladder and colon. Mechanical distension of these structures gives rise to ATP release and afferent activation. The role of the lipid-sensitive, mechanogated K^+ channels TREK-1, TREK-2, and TRAAK of the 2-pore domain K^+ channel family in nociceptor mechanotransduction is open to discussion.

TRPV1 is a nonselective cation channel with pronounced permeability for Ca^{2+} ions. It is the main molecular substrate for the ability of polymodal nociceptors to respond to acid, heat, and certain pungent chemicals. *In vivo*, this channel serves as a molecular integrator for multiple types of stimuli. These include exogenous chemicals, such as capsaicin and other vanilloids, and many endogenous substances that are released or locally produced upon injury, such as protons, membrane-derived lipids, cytokines, bradykinin, ATP, adenosine, and protons accumulated during tissue injury or ischemia. Other TRP channels (TRPV2, TRPV3, and TRPV4) are also gated by temperature, with thresholds ranging from warm to very hot ($>52\,^{\circ}C$) temperatures.

TRPA1 is a TRP channel initially proposed as a specific transduction molecule for noxious cold (threshold $\sim 18\,^{\circ}C$) in peripheral nociceptive neurons. *In vivo* findings have not lent full support to this view. In contrast, it is well established that TRPA1 channels are also activated by various natural pungent compounds such as cinnamaldehyde, allylisothiocyanate, and allycin (which are active ingredients of cinnamon oil, mustard oil, and garlic, respectively), by environmental irritants such as unsaturated aldehydes, and by cannabinoids, evidencing the polymodal nature of this TRP channel. TRPV1 and TRPA1 channels colocalize on many nociceptors. Furthermore, they are regulated through signaling pathways that are activated by metabotropic membrane receptors coupled to G-proteins. These include receptors for inflammatory mediators such as bradykinin, ATP, and nerve growth factor (NGF) that are coupled to phospholipase C activation. Hence, it is possible that such pathways are used by algesic substances to activate both TRPV1 and TRPA1, depolarizing nociceptor nerve terminals and giving rise to nociceptor sensitization.

In vivo data also support a role for the TRPM8 channel, which is gated by low temperatures and cooling agents such as menthol and icilin, in innocuous and noxious cold sensing (Figure 3).

Transduction channels in nociceptors are integral membrane proteins forming selective ion conductance pathways with variable permeability to monovalent (Na^+ and K^+) and divalent (Ca^{2+}) ions. Their change in activity by sensory stimuli leads to the flow of ionic currents that collapse membrane potential in the area where channels are active, acting as a sink for generator currents from neighbor, unstimulated axonal regions, thereby generating propagated action potentials. The detailed membrane biophysical

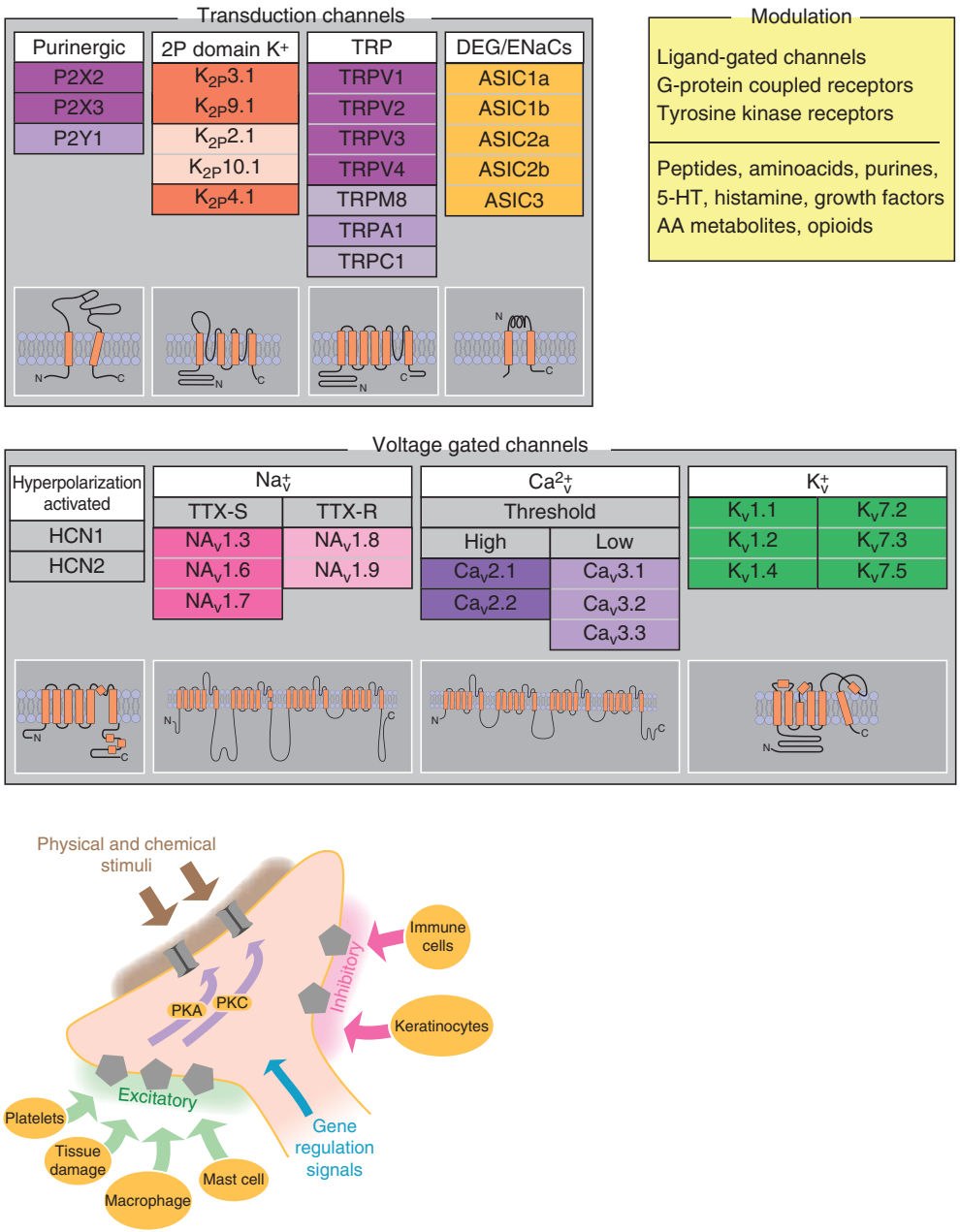

Figure 3 Transduction channels and voltage-gated channel types present in nociceptive neurons. Modulation of neuronal excitability is mediated by the activation of different classes of membrane receptor proteins by extracellular chemical agents. (Bottom) A nociceptor ending represented schematically. Noxious physical and chemical stimuli activate transduction channels. Endogenous chemical mediators acting through intracellular messenger pathways exert excitatory or inhibitory effects on channel gating. Long-term changes in membrane protein expression and channel function are determined by gene regulation triggered by injury and inflammation.

properties of the various functional types of nociceptor fibers at their terminal branches, branching points, and parent axons are not known. There exist marked differences in postspike excitability recovery and the ability to sustain repetitive discharges among nociceptor types, which suggests a variable density and distribution of voltage-dependent ion channels along the distal axonal tree of the different subclasses of nociceptive neurons.

Nociceptor Activity Following Tissue Injury

Noxious mechanical forces, extreme hot or cold, and a large number of exogenous irritants and endogenous substances released by injured tissues acting on the transduction molecules discussed previously depolarize nociceptor axons to their firing threshold, thereby evoking an immediate impulse discharge

whose duration and frequency depend on the intensity and duration of the stimulus as well as on the adaptation properties of the activated fiber. In addition, tissue injury resulting from the aforementioned stimuli acting on the tissues where nociceptors are located changes their local chemical environment and, by activation of modulatory pathways, modifies their responsiveness and firing pattern for longer time periods.

Injury causes the rapid accumulation of a large variety of inflammatory products at the lesioned area, such as ions and small chemicals contained within destroyed cells and mediators locally released by resident and blood-derived cells (mast cells, fibroblasts, leukocytes, and platelets) or formed from plasma proteins and membrane phospholipids. Molecules found in damaged tissues include H^+ and K^+ ions, adenosine and ATP, serotonin, histamine, platelet-activating factor, bradykinin, prostaglandins, leukotrienes, tromboxanes, interleukins, tumor-necrosis factor, and NGF. These mediators are responsible for local inflammatory reactions (vasodilatation, plasma extravasation, and cell migration) that act on membrane channels and receptor proteins at nociceptor nerve endings, provoking short-term changes in their gating properties. The altered ion flow and increased membrane excitability create the augmented impulse activity and reduced threshold that characterize the process of sensitization of injured nerve terminals. As discussed previously, the type of mediators released at the injured site varies with the tissue and the type of injury. In addition, the various classes of nociceptor fibers possess distinct second messenger pathways to modulate their equally characteristic population of nociceptor channels. Hence, alterations of the firing behavior of nociceptors vary according to the type of lesion and nociceptor fibers affected. For instance, sensitization to heat appears to be more prominent in type I Aδ nociceptors in the glabrous skin of humans and primates, whereas in the hairy skin, C polymodal nociceptors become sensitized by heat. Inflammatory mediators use different signaling pathways to exert their effects on channel activity. For instance, prostaglandin E_2, ATP, and adenosine or 5-HT activate the PKA signaling pathway, whereas other mediators such as bradykinin act on protein kinase C to produce sensitization (**Figure 3**). Nociceptor endings also possess receptor molecules whose activation reduces their excitability, thus causing inhibition. These include opioid, muscarinic, and GABA receptors.

In addition to the short-term, direct and indirect effects on channel behavior at the peripheral nerve endings, some of the inflammatory products released at the injured site have longer-lasting effects on nociceptive neurons through transcription-dependent changes in gene expression. This leads to the synthesis of new channel and receptor proteins, thereby producing more durable modifications of nociceptor excitability and, consequently, of the discharge pattern of propagated nerve impulses that constitutes their sensory message to consciousness.

For instance, NGF is a trophic factor that plays a critical role in the development of nociceptors in early life so that mutation of the gene encoding TrkA, the high-affinity receptor for NGF, causes the specific loss of nociceptors and congenital insensitivity to pain in humans. In adult animals, NGF also regulates channel and receptor protein expression in intact peripheral nociceptive neurons. Moreover, NGF released in augmented concentrations in injured tissue acts on nociceptive ganglion neurons containing TrkA receptors (approximately 40%, mostly CGRP and substance P peptidergic neurons), modifying their sensitivity. This effect occurs directly through interaction with TrkA receptors at the nociceptor membrane, which causes posttranslational phosphorylation of ion channels (particularly TRPV1 and voltage-sensitive Na^+ channels) through various intracellular cascades and also indirectly by releasing other inflammatory agents from local immune cells. Finally, an increased concentration of NGF in inflamed tissues affects gene expression in TrkA-expressing neurons and modifies the expression of ion channels involved in sensory transduction and impulse coding at nociceptor terminals (TRPs, ASICs, and TTX-resistant sodium channels). Upregulation of other receptor proteins and growth factors such as brain-derived neurotrophic factor, neuropeptides, and central neurotransmitters is an additional mechanism that explains the phenotypic changes experienced by nociceptive neurons during persistent inflammation. This is the basis of enhanced sensitivity and impulse firing in response to non-noxious and noxious stimuli exhibited by nociceptor terminals that defines primary hyperalgesia of the injured area.

Abnormal Firing in Injured Nociceptors

In acute pain resulting from sudden stimulation of nociceptors by tissue injury and also in inflammatory pain produced by activation of sensitized peripheral nociceptors, neural pathways involved in the detection and transmission of nerve impulses from the periphery to higher centers remain intact. Differences in the characteristics of the sensory message produced at primary sensory neurons are attributable to the previously described short- and long-term changes in neuronal excitability affecting peripheral nerve terminals, secondary to peripheral inflammation.

On the other hand, a number of pathological processes may directly affect nociceptive neural pathways, causing its defective functioning. Traumatic injuries, bacterial and viral infections, metabolic disorders, toxic effects, ischemia, hemorrhage, autoimmune diseases, and inflammation are the main causes of damage to peripheral as well as higher-order nociceptive pathways. This altered function of peripheral or central nociceptive pathways is the origin of an abnormal type of pain called 'neuropathic pain.' Neuropathic pain due to injury of the peripheral axon, cell body, or spinal branches of the primary nociceptive neurons located at dorsal root or cephalic ganglia is called 'peripheral neuropathic pain' and is far more common than central neuropathic pain.

Injured peripheral nerves form distal endbulbs entrapped in a tangled mass of fibers called 'neuroma.' Neuroma endings produce growing sprouts that may allow some of the damaged axons to regenerate into the distal nerve while many cut nerve fibers remain trapped at the injured area or grow in the retrograde direction. Recordings of electrical activity in the proximal portion of injured peripheral axons show that some of them, particularly the Aδ type but also unmylelinated C axons, develop an irregular, often bursting or paroxistic spontaneous activity (**Figure 4(a) and 4(b)**) that may have its origin in neuroma endbulbs or in more proximal regions, including demyelinated portions of the axon and the soma of the primary sensory neuron, where subthreshold oscillations of membrane potential and spontaneous firing become apparent after peripheral axotomy representing up to 75% of the total ectopic activity. In addition to the development of spontaneous ectopic activity, peripherally injured primary sensory neurons also become abnormally responsive to external stimuli of different quality. Response to mechanical stimulation with an impulse discharge of variable duration is a prominent feature of peripheral neuromas (**Figures 4(c) and 4(d)**) and may also develop along the trajectory of the injured nerve or in the dorsal root ganglion when these are directly affected by pathological process, as occurs during compression by herniated intervertebral disks. Abnormal, ectopic impulse activity is also induced by other stimuli, such as hypoxia, temperature oscillations, metabolic and ionic changes, as well as a large variety of endogenous inflammatory mediators and various neurotransmitters (particularly catecholamines), giving rise to neuropathic pain. Finally, 'cross-talk' also develops between injured and intact neurons at the peripheral axon and the cell body due to electrical coupling between membranes of neighbor cells at points where, due to the pathological process, they become in direct apposition, causing an 'ephaptic' excitation of fibers or somas not directly affected by trauma or viral infection. Chemically mediated cross-excitation due to local release of ions and neurotransmitters within the injured nerve or ganglion has also been proposed to explain hyperexcitability of intact neuron. The spread of simultaneous, abnormal repetitive activity among an increasing number of neurons caused by cross-talk has been proposed as the background mechanism for the paroxystic pain experienced by patients with trigeminal neuralgia and other peripheral neuropathies.

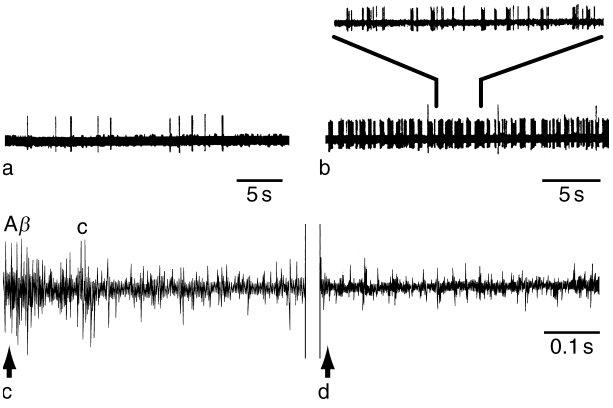

Figure 4 (a and b) Examples of ectopic, spontaneous nerve impulse activity in neuroma fibers of the cut saphenous nerve of the rat recorded *in vitro*; a segment of the recording in b has been expanded six times. (c and d) Intraneural microelectrode recordings of median nerve fibers at the wrist of a patient with an amputated hand suffering from phantom pain syndrome. Tapping the neuroma (arrows), which accentuated the phantom limb pain, induced afferent discharges in A and C nerve fibers (c). Blocking the neuroma with lidocaine (d) completely abolished the tap-induced afferent discharges and the tap-induced accentuation of the phantom pain). (a,b) Reproduced from Rivera L, Galler J, Pozo MA, and Belmonte C (2000) Responses of regenerating nerve fibres of the rat saphenous nerve neuroma to mechanical and chemical stimulation. An *in vitro* study: *Journal of Physiology* 527: 305–313. (c,d) Reproduced from Nystrom B and Hagbarth KE (1981) Microelectrode recordings from transected nerves in amputees with phantom limb pain. *Neuroscience Letters* 27: 211–216.

It is well established that when sensory nerves are cut, the ganglion neurons from which they originate also experience a number of molecular changes. After peripheral damage, the absence of the chemical signals that under normal conditions are transported along the axon centripetally to the soma, such as NGF and glial cell line-derived neurotrophic factor, modifies the expression of genes that encode ion channel and receptor proteins and regulate their trafficking to and from the periphery. In the injured neuron, the molecules transported distally by axonal flow accumulate at the neuroma endings. They include various classes of Na^+ and K^+ channels. Somatic and visceral nociceptor neurons normally express various subtypes of sodium channels. Among

these, $Na_v1.7$, $Na_v1.8$, and $Na_v1.9$ are largely responsible for the different currents that sustain the initiation and propagation of nerve impulses at their axon and cell soma. Axonal damage induces a significant upregulation of the previously silent sodium channel gene $Na_v1.3$ and downregulation of the $Na_v1.8$ and $Na_v1.9$ genes. Expression of other voltage-gated ion channels, such as K^+ and Ca^{2+}, and nonselective cation channels, such as HNC channels, which regulate the excitability and firing pattern of primary sensory neurons, is also variably altered by axotomy. This determines the modification of ionic currents present in neuroma endings and favors the development of ectopic discharges and the abnormal responsiveness to mechanical and chemical stimuli.

See also: Dorsal Root Ganglion Neurons; Neuropathic Pain; Pain Pathways: Descending Modulation; Pain and Plasticity; Pain: Central; Pain: Neuroimaging; Somatosensory Pathways (Ascending): Functional Architecture; Spinal Cord Pain Systems; Trigeminal Motor System.

Further Reading

Belmonte C and Gallar J (1996) Corneal nociceptors. In: Belmonte C and Cervero F (eds.) *Neurobiology of Pain*, pp. 146–183. New York: Oxford University Press.

Belmonte C and Cervero F (eds.) (1996) *Neurobiology of Nociceptors*. New York: Oxford University Press.

Belmonte C and Gallar J (2000) The primary nociceptive neuron: A nerve cell with many functions. In: Rowe MJ and Iwamura Y (eds.) *Somatosensory Processing: From Single Neuron to Brain Imaging*, pp. 27–48. Amsterdam: Harwood.

Belmonte C and Giraldez F (1981) Responses of cat corneal sensory receptors to mechanical and thermal stimulation. *Journal of Physiology* 321: 355–368.

Belmonte C and Viana F (2007) Transduction and encoding of noxious stimuli. In: Schmidt RF and Willis WD (eds.) *The Encyclopedia of Pain*, pp. 2515–2528. New York: Springer.

Bonica JJ (1990) *The Management of Pain*, 2nd edn. Philadelphia: Lea & Febiger.

Burgess PR and Perl ER (1967) Myelinated afferent fibres responding to noxious stimulation of the skin. *Journal of Physiology* 190: 541–562.

Burgess PR and Perl ER (1973) Cutaneous mechanoreceptors and nociceptors. In: Iggo A (ed.) *Handbook of Sensory Physiology*, vol. 2, pp. 29–78. New York: Springer.

Campbell JN and Meyer RA (2006) Mechanisms of neuropathic pain. *Neuron* 52: 77–92.

Cervero F and Laird JMA (2004) Understanding the signaling and transmission of visceral nociceptive events. *Journal of Neurobiology* 61: 45–54.

Clapham DE (2003) TRP channels as cellular sensors. *Nature* 426: 517–524.

Devor M (2006) Sodium channels and mechanisms of neuropathic pain. *Journal of Pain* 7: S3–S12.

Gebhart GF (ed.) (1995) *Progress in Pain Research and Management: 5. Visceral Pain*. Seattle: IASP Press.

Julius D and Basbaum AI (2001) Molecular mechanisms of nociception. *Nature* 413: 203–210.

Kumazawa T (ed.) (1996) *Progress in Brain Research: The Polymodal Receptor-A Gateway to Pathological Pain, vol.* 113. Amsterdam: Elsevier.

McMahon SB and Koltzenburg M (2005) *Wall and Melzack's Textbook of Pain*. London: Elsevier.

Nystrom B and Hagbarth KE (1981) Microelectrode recordings from transected nerves in amputees with phantom limb pain. *Neuroscience Letters* 27: 211–216.

Rivera L, Galler J, Pozo MA, and Belmonte C (2000) Responses of regenerating nerve fibres of the rat saphenous nerve neuroma to mechanical and chemical stimulation. An *in vitro* study: *Journal of Physiology* 527: 305–313.

Rush AM, Cummins TR, and Waxman SG (2007) Multiple sodium channels and their roles in electrogenesis within dorsal root ganglion neurons. *Journal of Physiology* 579: 1–14.

Schaible H-G and Schmidt RF (1998) Time course of mechanosensitivity changes in articular afferents during a developing experimental arthritis. *Journal of Neurophysiology* 60: 2180–2195.

Schmidt R, Schmeltz M, Foster C, Ringkamp M, Torebjörk HE, and Handwerker HO (1995) Novel classes of responsive and unresponsive C nociceptors in human skin. *Journal of Neuroscience* 15: 333–341.

Sengupta JN and Gebhart GF (1995) Mechanosensitive afferent fibres in the gastrointestinal and lower urinary tracts In: Gebhart GF (ed.) *Progress in Pain Research and Management: 5. Visceral Pain*, pp. 75–98. Seattle: IASP Press.

Nogo-A: Its Role in Axon Regeneration

M E Schwab and A Ghosh, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Injuries of the adult central nervous system (CNS) had been considered incurable over many centuries. CNS injuries such as spinal cord injury (SCI) impair basic functions of the body and lead to disability. The major cause of disability after injury is a disconnected nervous system rendering sections of the spinal cord denervated. For the past 20 years, great progress has been made toward understanding why the injury-induced disconnection in the adult CNS is persistent. One reason for the absence of spontaneous regrowth of CNS nerve fibers after injury is the presence of growth-inhibitory factors, especially in CNS myelin as well as in the scar that forms at the injury site. A key factor is a protein of the mammalian CNS now known as Nogo-A. In this article we summarize the knowledge gained over two decades that led to the conceptualization of Nogo-A as an inhibitor of growth and regeneration, and how the application of this concept has given a new perspective of making CNS injury curable.

Identification of Nogo-A and Its Expression

The Rationale behind Nogo-A and Its Identification

It had been long observed that CNS injuries at very immature ages often result in significant regeneration and sufficient plasticity to restore lost functions, contrary to the situation in adults. Moreover, the peripheral nervous system (PNS) in contrast to the CNS is able to regenerate injured axons even over long distances (Figure 1(a)). This difference in regeneration capacity does not lie in the neurons but rather in their environments: when injured neurons of the CNS are exposed to PNS tissue, they regenerate and elongate in the PNS tissue, demonstrating that adult neurons of CNS indeed have the capacity to regenerate in a favorable environment (Figure 1(b)). Experiments were then carried out to dissect the constituents of the CNS that are particularly inhibitory to neuronal growth resulting in an inhibitory environment. It was observed that neurons seeded on dissociated CNS cells particularly avoid the oligodendrocytes, the myelinating cells of the CNS (Figure 1(c)). In case of young CNS, myelination is still absent or incomplete allowing a more favorable environment for outgrowth compared to the adult. An inhibitory effect of CNS myelin was further demonstrated by induction of growth cone collapse by oligodendrocytes or myelin.

Identification of a growth-inhibitory cell type in the CNS led to the crucial hunt for the particular inhibitory constituents of the cell. Inhibition through myelin was abolished by proteases, pinpointing the role of one or several proteins. Circumstantial evidence, such as contact-based inhibition of neurite outgrowth by oligodendrocytes, pointed toward a membrane protein as a potential inhibitory candidate. Several membrane proteins were extracted from myelin and the one most inhibitory to neurite outgrowth was named NI-220/250. Antibodies were raised against NI-220/250, and their application on myelin substrates showed neutralization of myelin's inhibitory effect on neurite outgrowth. IN-1 (neutralizing antibody against NI-220/250) paved the way for regeneration after SCI, as discussed later.

The identification of peptide sequences of NI-220 led to the discovery of the gene, which was called *nogo*. It gives rise to three different protein products based on alternative splicing. The protein products are Nogo-A, Nogo-B, and Nogo C, which are 1163, 360, and 199 amino acids long rat sequences, respectively. Out of these splice variants, Nogo-A corresponds to the myelin inhibitor NI-220/250 and is recognized and neutralized by IN-1. It shares a common sequence of 188 amino acids at the carboxyl-terminus with Nogo-B and Nogo-C (Figure 1(d)). This common sequence is homologous to three other members of the reticulon (RTN) protein family. The similarity of Nogo-A to other RTN proteins is restricted to this carboxy-terminal end. RTNs are present in all eukaryotes; however, only its new member Nogo-A has known neurite inhibitory capacities. The middle Nogo-A-specific fragment of Nogo-A, together with a 66-amino-acid sequence at the carboyxl-terminus ('Nogo-66'; Figure 1(d)) gives it most of its outgrowth-inhibitory capabilities in mammalian CNS. The functions of Nogo-B and -C remain illusive.

Expression of Nogo-A

In the adult CNS, Nogo-A is prominently present in the myelin-forming oligodendrocytes. It is largely localized in the compartments of the endoplasmic reticulum, a significant proportion is expressed on the cell membrane, with the two inhibitory sites facing the extracellular space. In the developing nervous

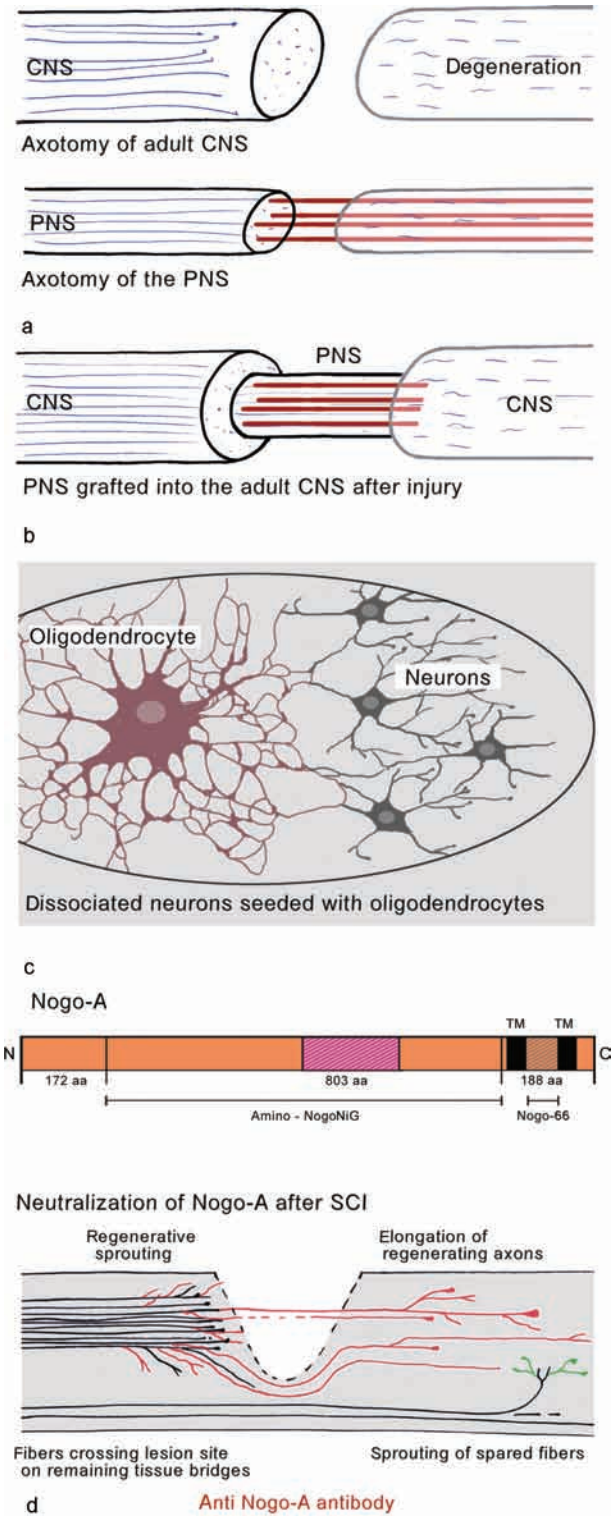

Figure 1 Key observations leading to regeneration in the adult CNS. (a) Axons of the adult CNS degenerate upon axotomy, whereas in the PNS axotomized neurons can regenerate over very long distances. (b) The inability of the CNS to regenerate is not due to the axons themselves. Experiments done by Aguayo (repeated from Tello in 1911) showed that the axons of CNS neurons can regenerate in the more permissive PNS environment. (c) CNS tissue and oligodendrocytes were identified to be inhibitory to neurite growth *in vitro*. Oligodendrocytes are myelinating cells present in the CNS and not in the PNS. (d) Nogo-A: proteins were extracted from myelin membranes and the most potent inhibitory protein was identified as Nogo-A. Neutralization of Nogo-A after SCI: neutralization of Nogo-A led to significant alternation of CNS inhibition *in vivo* resulting in enhanced sprouting and long-distance regeneration of axons after adult CNS injury.

system, Nogo-A is also transiently expressed by many neurons, especially projection neurons with long axons. Its role as a repulsive axonal guidance molecule is currently under investigation.

Spinal Cord Injuries

Majority of studies on CNS regeneration, including those on the effect of neutralization of Nogo-A after SCI, have been done in adult rats. In order to further discuss the aspects of injury on which Nogo-A neutralization acts to cause regeneration, a brief introduction to SCI and the animal models used is essential.

Brief Overview of Relevant Changes Induced by SCIs

SCI leads to a complex cascade of events, many of which are still poorly understood. Ischemia and inflammation enlarge the lesion (secondary tissue loss), and activated microglia together with astrocytes laying the foundations of the scar at the site of injury. There is a temporary opening of the blood–brain barrier. Axons that are injured but still connected to the cell body often sprout, but then retract. Axons detached from their cell bodies undergo Wallerian degeneration. For injured axons to re-innervate their targets, the axons would need to either grow through or bypass the lesion site.

Often SCIs in humans are anatomically incomplete, where some tissue is spared even if it appears insufficient to mediate functions. This spared tissue, as discussed later, can act as a bridge for regenerating fibers to reach their sites of innervations. Interestingly, CNS injury also results in a neuronal growth response. This transitory growth response manifests itself as sprouting; near the site of injury, these sprouts could be a source for regenerating fibers. In adult CNS tissue, these sprouts have often been observed to grow only over short distances and to be unable to pass through the lesion scar or across the spared tissue bridges. The occurrence of sprouting after CNS injury has been known for a long time but its regulation and possible consequences were not well studied. It was shown only recently that spontaneous sprouting after SCI in adult rats could lead to the formation of specific new intraspinal circuits that correlated with

spontaneous behavioral recovery. However, in cases of complete or near-complete SCIs, the extent of such spontaneous recovery is not large enough to lead to a recovery making the search for interventional therapy necessary.

Animal Models Used to Study SCIs

Contusion by controlled impact (weight drop or electromechanical 'impactors') or clipping, or partial or full transection of the spinal cord in rats or mice are the most frequently used models at present. The approach is by a dorsal laminectomy. Human injuries, however, are mostly ventral closed contusions. To study axonal regeneration unequivocally, a full lesion of a well-defined tract that can be traced neuroanatomically is required. Only the corticospinal tract (CST), to some extent the rubrospinal tract (RST), and the ascending dorsal column sensory tracts fulfill these requirements. Unilateral hemi or overhemisections, dorsal bilateral lesions, or the so-called T lesions (including the CST, which runs mainly in the dorsal part of the rodent spinal cord, as well as its ventral and medial components) are therefore most frequently used. These models of SCI mimic incomplete injuries, with moderate to severe functional loss. The spared tissue also makes a bridge possible, which plays a critical role in the ability of the axotomized fiber tracts to regenerate; full transactions produce a scar and often a several-millimeter-long gap that is impermeable for the growing axons.

Many studies have also used injury models that are less complicated than a large SCI. Such injuries involve transection of only certain chosen fiber tracts: for example, the dorsal column or, without any injury at the level of the spinal cord itself, the CST at the pyramidal decussation. This lesion model results in a spinal cord deprived of CST projections.

Axonal Regeneration after Neutralization of Nogo-A

Regeneration in Adult Rodents

In vitro evidence clearly showed a strong growth-inhibitory effect of Nogo-A. Immediately after conceptualization of the inhibitory effect of Nogo-A, several experiments were carried out in young and adult spinal cord-injured rats with complete CST lesions to examine the role of Nogo-A in CNS regeneration. In the initial experiments, hybridoma cells, secreting IN-1 antibody to neutralize Nogo-A, were grafted into the cortex of immune-suppressed rats (Figure 2(b)). The CST, visualized using the anterograde tracer horseradish peroxidase (HRP), was seen to regenerate up to 11 mm in the presence of IN-1 antibody in adult animals within 2–3 weeks. Control antibody-treated rats showed only the well-known, short (0.1–1 mm) sprouting response. This inducible regeneration, then seen for the first time in the absence of Nogo-A's inhibitory activity, occurred at a maximal rate of 0.5–1 mm per day. Regenerating fibers used the spared tissue as a bridge to bypass and get below the site of injury (Figure 1(d)). To date, no artificial bridges in combination with Nogo-A neutralization have been successful.

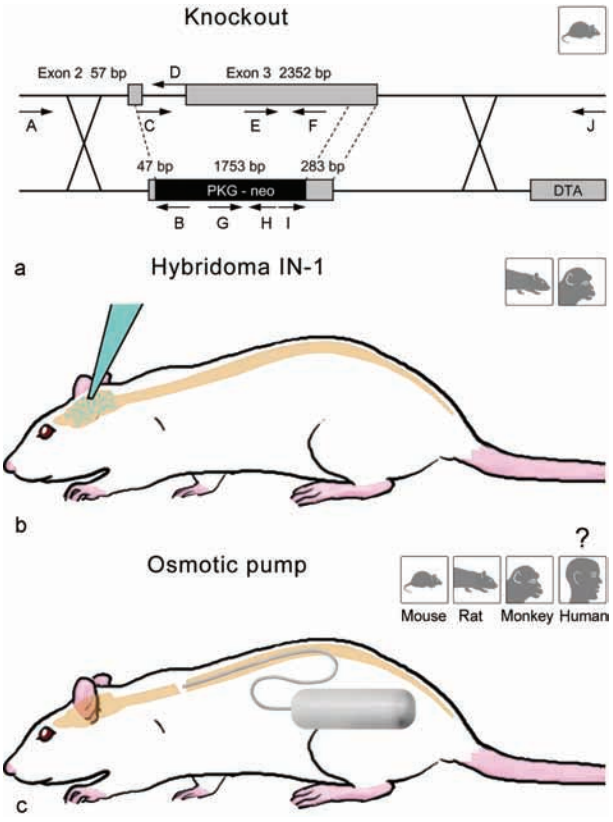

Figure 2 Strategies for neutralizing Nogo-A activity. (a) Cloning of the gene *nogo* led to the possibility of knocking out the gene and thus producing mice without Nogo-A. Such mice show enhanced regeneration compared to wild-type. However, such an approach is limited to mice and is difficult to interpret due to compensation during development by other inhibitory proteins. An inserted PKG-neo gene disrupted the production of Nogo-A and the inserted Diptheria toxin A (DTA) gene was used to select against random insertions of the gene. Transfected ES cells were screened using primes A, B, I, and J out of primers A–J. (b) Hybridoma cells were designed to secrete the antibody IN-1, which neutralizes Nogo-A. Cells are transplanted into the brain at the time of injury. This results in neutralized Nogo-A and regeneration and behavioral recovery in spinal cord- and brain-injured rats and monkeys. (c) Osmotic pumps deliver neutralizing anti-Nogo-A antibody into the cerebrospinal fluid bathing the spinal cord and brain. This gives more temporal control regarding delivery, and the dosage can be determined precisely. Nogo-A-neutralizing antibodies or peptide reagents induced axonal regeneration and behavioral recovery in mice, rats, and monkeys.

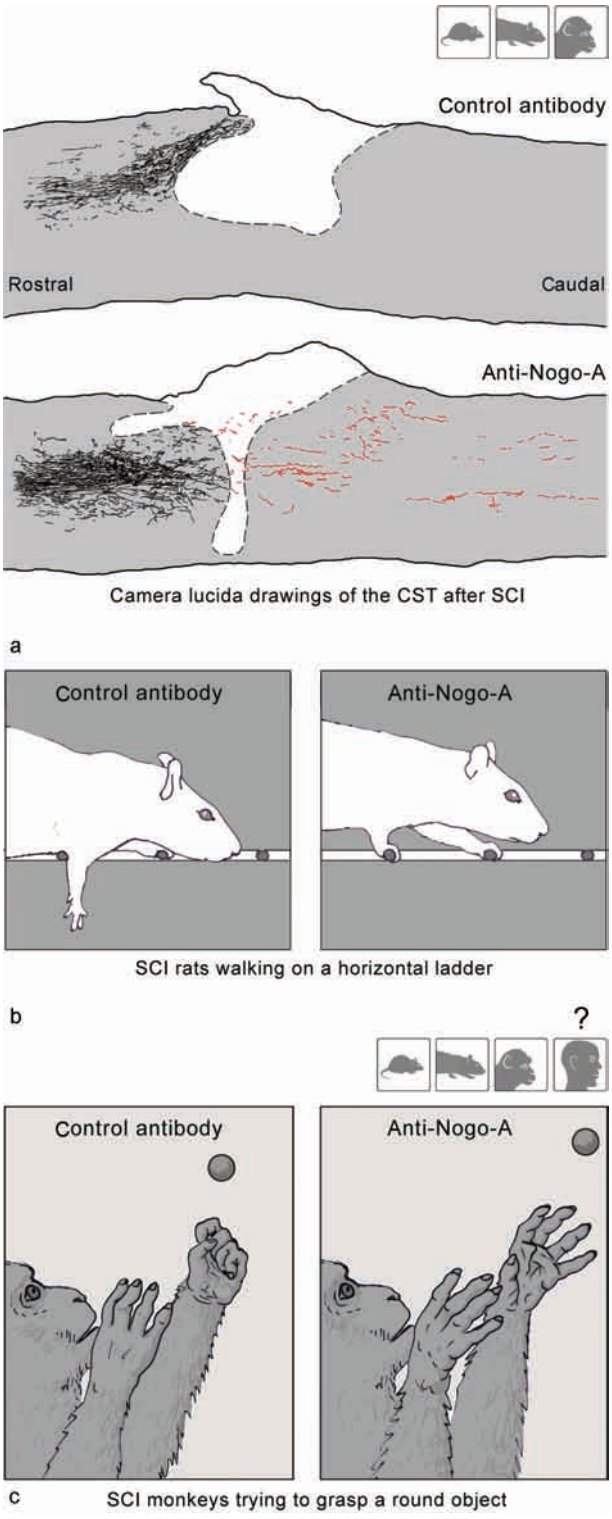

Figure 3 Regeneration and behavioral recovery after Nogo-A neutralization. (a) Nogo-A neutralization leads to enhanced regeneration after adult SCI. Regeneration is seen within a few weeks after treatment. Corticospinal neurons, thought to be difficult to regenerate, bypass the lesion site using spared tissue to re-innervate the denervated spinal cord in mice, rats, and monkeys. (b) Behavioral recovery is observed after Nogo-A neutralization in the adult spinal cord-injured animals. Spinal cord-injured rats show better coordination and body weight support and a more precise grip after treatment when crossing a horizontal ladder. (c) Macaque monkeys with high cervical SCIs resulting in loss of fine finger movements recover hand control after Nogo-A neutralization, as seen in a food-catching test.

Upon the observation of successful regeneration, challenging experiments immediately followed to examine its functional effects. Animals were tested on batteries of behavioral read-outs after SCI in the presence of Nogo-A antibody. One of the first recoveries noted was in a sensorimotor reflex, known as the contact placing response, which is strictly dependent on the CST. Rats place their limbs on a surface for support in response to a slight stimulation of the dorsal shin of the foot. This CST-dependent reflex showed a significant recovery in the presence of IN-1 antibody, but not in the controls. In addition, the spinal cord-injured rats showed improvements in a variety of locomotion tests in the presence of IN-1.

The results seen with IN-1 antibody treatment in rats led to experiments with Nogo-A inhibition in a clinically closer setting. Two new monoclonal monospecific Nogo-A antibodies (11C7 and 7B12) were generated against inhibitory fragments of Nogo-A and administered intrathecally to spinal cord-injured rats (T lesion) with osmotic minipumps at the time of injury (Figure 2(c)). Just as after IN-1 treatment secreted from hybridoma cells, the application of the new neutralizing antibodies resulted in long-distance regeneration of the CST (Figure 3(a)). They also resulted in significant overall improvement in coordination during locomotion in an open field and during swimming. Exposing animals to tests such as walking on a narrow beam or a horizontal ladder resulted in more precise hindlimb foot placements and a close to normal pattern of body weight support during walking (Figure 3(b)). These findings encouraged the possible usage of Nogo-A neutralization in a clinical setting in humans. One of the major concerns in spinal cord-injured patients is pain and allodynia. Animals treated with Nogo-A antibodies show no significant differences to control antibody-treated injured animals in tests that detect pain or allodynia.

Enhanced regeneration of injured CST axons was also observed in mice after genetic deletion of Nogo-A (Figure 2(a)). The extent of regeneration was, however, also influenced by background genes of specific mouse strains, that is, more extensive in SV 129 than in C57Bl/6.

Regeneration in Adult Primates

The rodent studies revealed Nogo-A neutralization as a potential treatment for human SCI. As a prerequisite to enter clinical trials it was essential to show that Nogo-A neutralization was effective in primates.

Initial experiments focused on using IN-1 antibody in adult marmosets (*Callitrix jacchus*). Following injury of the CST in the spinal cord, the treated marmosets showed an increase in sprouting, and regenerating fibers in the spinal cord caudal to the lesion. The much larger Macaque monkeys were studied then using anti-Nogo antibodies delivered intrathecally from subcutaneous pumps. The animals were subjected to a cervical unilateral hemisection including the CST, which rendered the corresponding hand paretic, but only minimal effects on hindlimb locomotion or vegetative functions. Upon application of the anti-Nogo-A antibodies for 4 weeks, the Macaque monkeys showed regeneration of CST axons around the lesion scar as seen in marmosets and rats. Moreover, there was a dramatic improvement in the dexterity of the hand in response to the antibody as seen in tasks like catching food (Figure 3 (c)), opening drawers, and taking raisins from wells using the precision grip.

Regeneration in Injuries Other Than That of the Spinal Cord

Regeneration due to Nogo-A neutralization is widely studied in the spinal cord; however, it is not limited to the cord. When Nogo-A was neutralized in the presence of trophic factors, the crushed optic nerve was able to regenerate. The extent of regeneration was higher than that seen due to application of the trophic factor alone. Application of anti-Nogo-A antibodies after auditory nerve lesion (the auditory nerve is myelinated mainly by oligodendrocytes) led to similar results. Functional assessments also showed some recovery of sound perception in these animals.

Compensatory Fiber Growth after Neutralization of Nogo-A

While regeneration is thought to be critical to achieve complete recovery, several other forms of anatomical changes can take place in response to a CNS injury that can play a role in functional recovery. The spontaneous growth response, known as sprouting, is one such anatomical change. A well-studied example is the sprouting of the intact CST system after unilateral CNS lesions, such as stroke or pyramidotomy. These lesions result in a permanent loss of fine movements of the forelimb, in particular the hand and fingers. Some spontaneous sprouting occurs from the intact uncrossed ventral CST fibers in the denervated half of the spinal cord. Neutralization of Nogo-A, however, leads to sprouting of the main CST from the intact side of the spinal cord; these sprouts cross the midline to innervate the denervated half of the spinal cord, leading to full recovery of fine limb movements such as those used in grasping.

After bilateral pyramidotomy, where the spinal cord has lost all its input from the cortex, the RST shows no detectable sprouting. Anti-Nogo-A antibodies resulted in pronounced sprouting of the rubrospinal fibers into areas previously innervated by the CST. Again, a very high degree of functional recovery was observed. The lack of increase in pain or allodynia after Nogo-A neutralization indicates a high degree for specificity of newly formed connections. Little is known about the underlying mechanisms, especially for the role of activity in stabilizing and fine-tuning the new connections and circuits.

Possible Mechanism of Action of Nogo-A and Its Neutralization

Neurite Growth Inhibition by Nogo-A

Nogo-A induces growth cone collapse correlated with calcium surge in the growth cone and with a pronounced activation of the small GTPase RhoA. Purified Nogo-A acts at extremely low concentrations. These observations indicated the presence of a receptor on the neuronal membrane surface coupled to an amplifying signaling pathway. As discussed above, Nogo-A has two main inhibitory regions. For one of them, the Nogo-66 sequence, an interaction partner was found, the so-called Nogo receptor, NgR. NgR binds not only to Nogo-66, but also to myelin-associated glycoprotein (MAG), oligodendrocyte myelin glycoprotein (OMgp), and probably additional ligands. Interestingly, MAG and OMgp show *in vitro* inhibitory neurite growth activity along with Nogo-A. N_gR alone does not account for the potent inhibition by Nogo-A *in vitro*. Knockout mice lacking NgR show only little regeneration. This result confirms the importance of a second, still unknown receptor (subunit) for the Nogo-A specific active site ('amino-Nogo' in Figure 1(d)) of the molecule.

Nogo-66 and Nogo-A activate the Rho family of proteins, in particular RhoA and the Rho effector ROCK. Blockade of Rho or ROCK attenuates the inhibitory effect of Nogo *in vitro*, and enhances regeneration after spinal cord lesions *in vivo*. The complete signaling cascade leading to the destabilization of the growth cone cytoskeleton as well as the precise role of receptor-associated proteins like p75, Troy, or Lingo remain to be defined.

Induction of Growth Due to the Absence of Active Nogo-A

When neutralizing antibodies against Nogo-A were applied to the CNS of intact adult rats, an increase in growth-associated protein GAP-43, neuronal cytoskeletal proteins, growth factors, transcription factors, and axon guidance molecules was observed. We speculate that Nogo-A, present in the myelin

that surrounds the mature axons, sends a retrograde signal to the cell bodies, acting as a negative regulator of growth in the adult CNS. Thus, Nogo-A neutralization enhances regeneration and sprouting not only by making the CNS tissue more permissive for growing fibers but also by the enhancement of the neuronal growth mechanisms.

Potential of Nogo-A Inactivation for Other CNS Injuries

Neutralization of Nogo-A induces regeneration in the adult CNS but it also increases structural plasticity and is potentially neuroprotective. In stroke studies, adult rats recover the lost fine motor movements to a large extent after Nogo-A neutralization. Increased plasticity in the form of sprouting of the spared cortifugal fibers to innervate the denervated regions in the spinal cord and brain is thought to be a key factor causing such recovery. Delayed treatment with anti-Nogo-A antibodies or treatment in aged rats is also effective. These findings open an exciting outlook for a possible use of Nogo-A neutralization in the clinic, not only for SCI but also for brain trauma and stroke.

Recently, Nogo-A has been shown to be involved in autoimmune demyelination as seen in the experimental autoimmune encephalomyelitis (EAE), an animal model of multiple sclerosis (MS). Mice vaccinated against Nogo-A or Nogo-A knockout mice showed reduced inflammation, demyelination, and axon loss in this model. While the exact mechanisms that led to such effects remain unknown, clearly Nogo-A is a key contributor to the growth-inhibitory nature of the adult mammalian CNS and its neutralization may alleviate the loss of neuronal connectivity in various forms of degenerations.

The identification and neutralization of Nogo-A has led to not merely an inhibitory protein, but a concept that has instead resulted in the identification of several other growth inhibitors in the CNS. Identification of growth inhibitors in CNS myelin such as MAG, OMgp, ephrinB3, and several semaphorins (4d, 5A) are part of a potentially long list of proteins being understood as negative growth regulators. The utility of the concept of Nogo-A is clearly demonstrated in the swift transformation to clinically relevant studies raising the hope of meaningful regeneration and repair of CNS injuries in humans. Clearly, regeneration in the adult CNS is not a simple phenomenon. The identification of Nogo-A has emphasized the illusive complexity of the biological system that results in an inhibited and balanced adult nervous system. Nogo-A has given us many answers; however, it has also raised many challenging questions that remain to be solved in the coming years.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Injury: Neuronal Responses; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants.

Further Reading

Filbin MT (2003) Myelin-associated inhibitors of axonal regeneration in the adult mammalian CNS. *Nature Reviews Neuroscience* 4: 703–713.

Liebscher T, Schnell L, Schnell D, et al. (2005) Nogo-A antibody improves regeneration and locomotion of spinal cord-injured rats. *Annals of Neurology* 58: 706–719.

McGee AW and Strittmatter SM (2003) The Nogo-66 receptor: Focusing myelin inhibition of axon regeneration. *Trends in Neuroscience* 26: 193–198.

Schwab ME (2002) Repairing the injured spinal cord. *Science* 295: 1029–1031.

Schwab ME (2004) Nogo and axon regeneration. *Current Opinion in Neurobiology* 14: 118–124.

Schwab ME and Bartholdi D (1996) Degeneration and regeneration of axons in the lesioned spinal cord. *Physiological Reviews* 76: 319–370.

Raineteau O and Schwab ME (2001) Plasticity of motor systems after incomplete spinal cord injury. *Nature Reviews Neuroscience* 2: 263–273.

Non-Photoreceptor Photoreception

K Y Wong, Brown University, Providence, RI, USA

© 2009 Elsevier Ltd. All rights reserved.

The mammalian visual system processes photic information from the external world in two ways. Image-forming vision entails the analysis of spatial details in the environment and requires retinal rod and cone photoreceptors. Non-image-forming vision, on the other hand, is concerned with the measurement of the overall brightness of the surroundings and leads to various reflexive responses. Examples of these non-image-forming visual responses include constriction of the pupil, synchronization (i.e., entrainment) of the master circadian clock in the suprachiasmatic nucleus (SCN) and consequently circadian behaviors to the solar cycle, inhibition of melatonin release from the pineal gland, and acute suppression of locomotor activity (i.e., negative masking) of nocturnal animals.

Non-Image-Forming Vision Does Not Require Classical Photoreceptors

Contrary to image-forming vision, non-image-forming photoreception does not require rod or cone photoreceptors. In 1927, Clyde Keeler made the initial observation that *retinal degenerate* (*rd*/*rd*) mice which lacked rods and possessed few functional cones still had the ability to constrict their pupils in response to light. Subsequently, researchers found that the circadian behaviors of *rd*/*rd* mice could be entrained to the external light–dark cycle, and that the level of plasma melatonin in some blind patients could be reduced by exposure to bright light. The possibility that these non-image-forming visual responses were mediated by a few surviving rod/cone photoreceptors was ruled out by two studies performed in the 1990s. First, Takashi Yoshimura and Shizufumi Ebihara demonstrated that the most effective wavelength (λ_{max}) for phase-shifting circadian activity rhythms in *rd*/*rd* mice is around 480 nm, significantly different from the λ_{max} for the classical photoreceptors in mice (ultraviolet cone, 359 nm; rod, 498 nm; green cone, 510 nm), implicating a previously unidentified photoreceptor. Second, Russell Foster and colleagues developed a transgenic mouse line lacking all functional rods and cones (*rd*/*rd cl*) and found that these animals were still able to exhibit light-evoked phase shifts of circadian locomotor activity, a pupillary light reflex, negative masking, and light-induced inhibition of pineal melatonin release. In the retinas of *rd*/*rd cl* mice, the entire photoreceptor layer is missing, suggesting that the photoreceptors responsible for driving the residual non-image-forming visual behaviors reside in the inner retina.

The Novel Photoreceptor Is a Ganglion Cell That Expresses Melanopsin

The non-rod, non-cone photoreceptor that can mediate non-image-forming visual functions in the absence of rods and cones was discovered in 2002 by David Berson to be a rare subtype of retinal ganglion cells (RGCs). These cells are capable of generating robust responses to light stimulation even when pharmacologically or mechanically isolated from synaptic inputs, and they are thus intrinsically photosensitive and function as photoreceptors.

Light Response Properties

Consistent with Yoshimura and Ebihara's behavioral measurement mentioned previously, the endogenous light response of these intrinsically photosensitive retinal ganglion cells (ipRGCs) has a λ_{max} of approximately 480 nm. In addition, the intrinsic ipRGC photoresponse has the following characteristics that are strikingly different from the light responses of rod and cone photoreceptors:

1. The ipRGC light response is generated by the opening of cation-selective channels in the plasma membrane, resulting in depolarization from a relatively negative resting potential (typically approximately-65 mV), similar to the case for invertebrate photoreceptors. The amplitude of the ipRGC photoresponse increases as the intensity of the light stimulus increases. In contrast, illumination closes cation channels in rods and cones, hyperpolarizing them from relatively positive resting potentials (typically approximately −40 mV).
2. The ipRGCs can generate sodium-based action potentials, which are required for these cells to propagate electrical signals along their long axons to higher brain targets. The frequency of action potentials increases as the intensity of the light stimulus increases. By contrast, rod and cone photoreceptors encode irradiance information through graded potentials only.
3. The light response of ipRGCs is extremely sluggish (Figure 1). Responses to light steps of near-threshold intensities rise slowly and can take up to several minutes to reach the peak amplitude. The time to peak decreases as light intensity increases, and the

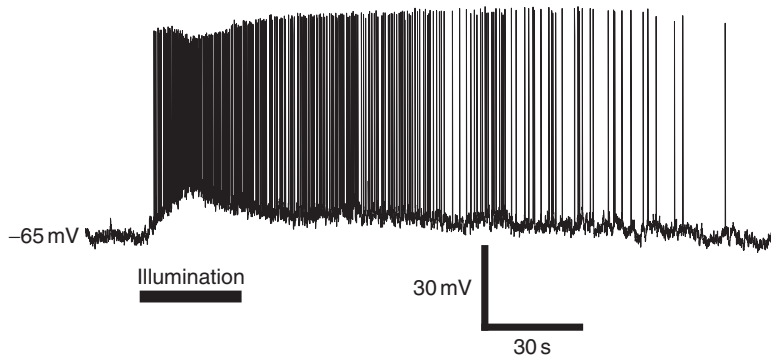

Figure 1 A melanopsin-driven light response recorded from a rat ipRGC. This response was recorded in the presence of pharmacological agents that blocked synaptic input from rod/cone-driven circuits, thus isolating the melanopsin-based photoresponse. The intensity of the light stimulus was approximately 1.5 $\log_{10}$ units above the threshold.

brightest stimuli can evoke responses that peak within several seconds. The ipRGCs also recover from light stimulation very slowly, requiring up to tens of minutes for their membrane potentials to return to baseline levels. By contrast, the photoresponses of rods and cones are always much faster, peaking within approximately 500 ms and usually returning to resting levels within a few seconds after the cessation of light stimulation.

4. Although ipRGCs can adapt to light stimulation by reducing their sensitivity, such adaptation is relatively limited and the ipRGC response to prolonged light illumination can be sustained for many hours. By contrast, cone photoreceptors adapt to light far more rapidly and completely, resulting in much more transient light responses. Thus, ipRGCs can integrate light information over much longer timescales than cones.

5. The photoresponse of ipRGCs has a threshold of approximately 10^{11} photons cm^{-2} s^{-1} at 480 nm, which is comparable to dawn light and is on the order of 5 $\log_{10}$ units less sensitive than rods and 2 $\log_{10}$ units less sensitive than cones at their corresponding peak sensitivity wavelengths. All three types of photoreceptors have comparable dynamic ranges, about 3 $\log_{10}$ units.

These properties of the endogenous ipRGC light response are advantageous to the various non-image-forming visual behaviors driven by these cells. The tonic nature of these responses enables the pupil to be constricted and nocturnal locomotor activity to be suppressed for many hours continuously. The relatively low sensitivity of these light responses and their requirement for prolonged illumination cause the ipRGCs to respond poorly to random, brief environmental photic fluctuations such as lightning, thus minimizing the impact of these events on the central circadian clock. In addition, the low sensitivity enables ipRGCs to complement the dynamic range of the more sensitive classical photoreceptors, which also contribute to non-image-forming vision (see the section titled 'Conventional photoreceptors also contribute to non-image-forming photoreception').

Some RGCs in rodents already express the photopigment melanopsin before birth (see the section titled 'Photopigment and transduction cascade') and have been shown by c-fos immunoreactivity, calcium imaging, and extracellular recording techniques to be light responsive from postnatal day 0 (P0), approximately 12 days before eye opening. By contrast, rods and cones do not become light responsive until approximately P10. The light responses of ipRGCs are more transient and approximately 1 $\log_{10}$ unit less sensitive prior to P4 than at P6 and older ages. The ipRGCs already make functional connections with the SCN and can drive light-induced expression of c-fos in the SCN from P0. The early emergence of ipRGC photosensitivity has been speculated to fine-tune the development of the non-image-forming visual pathways.

Up to three subtypes of ipRGC light responses have been demonstrated in the mouse retina by calcium imaging and extracellular spike recording, and they differ in terms of response onset latency, absolute sensitivity, and response offset kinetics. The significance of this diversity of ipRGC responses is unknown. One possibility is that these subtypes of ipRGCs project to different brain regions.

Besides the generation of an intrinsic light response, ipRGCs also display synaptically driven, rod/cone-mediated responses to light (see the section titled 'Synaptic input to ipRGCs').

Photopigment and Transduction Cascade

The photopigment in ipRGCs is melanopsin (gene symbol, opn4), which was originally discovered in

the photosensitive dermal melanophores of the frog by Ignacio Provencio and colleagues in 1998. RGCs expressing mammalian orthologs of melanopsin have been found by Provencio and other researchers in many mammalian species, including human, mouse, rat, monkey, cat, hamster, and blind mole rat. All ipRGCs express melanopsin, and in most species the ipRGCs are the only retinal neurons that express this protein. There is evidence that certain human cone photoreceptors also express melanopsin, although its functional roles in these cells are unknown. Genetically knocking out melanopsin in mice completely abolishes the endogenous photosensitivity of ipRGCs, consistent with melanopsin being the only photopigment in these cells.

Interestingly, the amino acid sequence of mammalian melanopsin shares much higher homology with invertebrate opsins than with vertebrate rod and cone opsins. Based on its amino acid sequence, melanopsin is predicted to have seven transmembrane domains, like all G-protein-coupled receptors including other opsins. Melanopsin has a long cytoplasmic tail with more than a dozen potential phosphorylation sites, suggesting that the function of melanopsin may be finely modulated by kinases.

Similar to the rod and cone photopigments, melanopsin consists of an apoprotein covalently linked to the chromophore 11-*cis*-retinaldehyde. When illuminated, the chromophore is isomerized into all-*trans*-retinaldehyde, inducing a conformational change in the melanopsin apoprotein which in turn triggers an intracellular transduction cascade that ultimately results in an increase in cation influx across the cell's plasma membrane. However, unlike the all-*trans*-retinaldehyde in rods which gets exported to the adjacent retinal pigment epithelium for the regeneration of 11-*cis*-retinaldehyde, melanopsin retains the all-*trans*-retinaldehyde and reisomerizes it to 11-*cis*-retinaldehyde upon exposure to light with appropriate wavelengths. In other words, melanopsin is bistable and functions as both photopigment and photoisomerase, just like many invertebrate photopigments. This property is critical to the maintenance of melanopsin photoresponses because the ipRGCs are physically far removed from the retinal pigment epithelium. Indeed, ipRGCs remain light responsive even when enzymes within the retinal pigment epithelium that are necessary for regenerating 11-*cis*-retinaldehyde for the rods are genetically knocked out or pharmacologically inhibited.

Considering the structural and functional similarities between melanopsin and invertebrate photopigments, it is perhaps not surprising that the phototransduction cascade in ipRGCs also appears quite similar to that in *Drosophila* photoreceptors. Melanopsin has several invertebrate-like sequence motifs in regions thought to be involved in G-protein binding, and it has indeed been demonstrated that, similar to phototransduction in *Drosophila* photoreceptors, ipRGC phototransduction requires a $G_{q/11}$ class G-protein. This G-protein appears, in turn, to signal through phospholipase C (PLC) in these cells, although the second messengers downstream from PLC are unknown. Finally, there is pharmacological evidence that in ipRGCs, the light-activated cation channel at the end point of the transduction cascade is a transient receptor potential (TRP) channel that belongs to the TRPC family.

Morphology and Distribution

The morphology and distribution of melanopsin-expressing RGCs (mRGCs) have been studied most extensively in rodent, primate, and cat retinas.

Rodents In the rat retina, melanopsin is expressed in approximately 2500 RGCs or 2.5% of all RGCs, whereas in the mouse retina approximately 750 or 1% of all RGCs express melanopsin. The density of mRGCs is slightly higher in the superior and temporal quadrants than in other parts of the retina. Approximately 95% of mRGCs are located in the ganglion cell layer, and the rest are displaced to the inner nuclear layer. The soma diameters average approximately 15 μm, which is intermediate among rodent RGCs. Each cell has two to six primary dendrites which branch sparsely, and the dendritic fields have an average diameter of approximately 500 μm (Figure 2, right). The dendrites are monostratified, usually terminating in the 'off' sublamina of the inner plexiform layer near the border with the inner nuclear layer, although the dendrites of some mRGCs appear to terminate in the proximal, 'on' sublamina instead. These two morphological types of mRGCs have been termed M1 and M2, respectively. Melanopsin is expressed in the soma, dendrites, and proximal axonal segments of each cell, and at least in the soma, melanopsin is targeted mainly to the plasma membrane. Both the soma and the dendrites of the M1 cells have been shown to be directly photosensitive, although it remains to be determined whether the M2 cells are also intrinsically photosensitive. Dendrites from adjacent mRGCs overlap extensively, forming an expansive photoreceptive net that covers the entire retina (Figure 2, left).

Primates In human and macaque retinas, melanopsin is expressed in approximately 3000 or 0.2% of all RGCs. The density of mRGC somata is the highest in

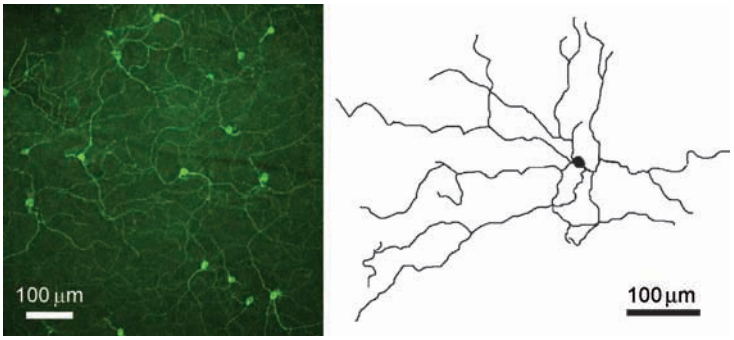

Figure 2 (Left) A photoreceptive net of ipRGCs in a flat-mounted rat retina as revealed by immunostaining against melanopsin. The image was generated by stacking images taken at several focal planes. (Right) Camera-lucida drawing of a rat ipRGC.

the parafoveal region and decreases toward the peripheral retina. Few mRGCs are found within the fovea. Approximately 60% of all mRGCs are in the ganglion cell layer, whereas the rest are displaced to the inner nuclear layer. The average soma diameter is approximately 20 μm, which is relatively large among primate RGCs. The dendrites branch sparsely and dendritic field diameters range from approximately 300 μm in the parafovea to approximately 1200 μm in the peripheral retina, much larger than those of any other known primate RGC. Each cell's dendrites are monostratified and terminate in either the on or the off sublamina in the inner plexiform layer, and both morphological types are directly photosensitive. Similar to mRGCs in rodents, the dendritic trees of adjacent primate mRGCs overlap extensively.

Cat Melanopsin is expressed in approximately 1700 RGCs or 1% of all RGCs in the cat. The density of mRGC cell bodies is fairly uniform across the entire retina. All mRGCs are in the ganglion cell layer. The average soma diameter is approximately 14 μm, which is relatively small among cat RGCs.

Axonal Projections

The axonal projections of the mRGCs have been best characterized in mice, rats, and hamsters. At least 70% of all mRGCs innervate the SCN, and at least 70% of the RGCs that project to the SCN express melanopsin. As mentioned previously, these retinal inputs synchronize the master circadian pacemaker in the SCN to the external light–dark cycle. Three other brain regions that receive most of their retinal inputs through mRGCs are the ventral subparaventricular zone (vSPZ), which is involved in photic and circadian modulation of sleep and locomotor activity, including negative masking; the intergeniculate leaflet, which integrates photic and nonphotic cues for the regulation of circadian rhythms; and the lateral habenula, which integrates signals from the limbic, motor, and circadian systems. The olivary pretectal nucleus, which controls the pupillary light reflex, receives substantial input from mRGCs, although it is also heavily innervated by conventional RGCs.

Melanopsin-expressing RGCs also make weaker axonal projections to several other parts of the brain, which derive most of their retinal inputs through non-mRGCs. These regions include the ventrolateral preoptic nucleus, which is involved in sleep–wake regulation; the peri-supraoptic nucleus, which modulates the release of hormones from the supraoptic nucleus proper; the medial amygdala, which integrates signals from the retina and the olfactory system to modulate reproductive behaviors; the superior colliculus, which controls eye movements; the ventral lateral geniculate nucleus, which is implicated in visuomotor functions such as optokinetic nystagmus; and the dorsal lateral geniculate nucleus (dLGN), which is the thalamic relay of retinal information to the primary visual cortex in the image-forming visual pathway. Innervation of the dLGN by ipRGCs has also been reported in primates, raising the possibility that ipRGCs might directly contribute to conscious visual perception.

The mRGC projections to the SCN and the vSPZ are bilaterally symmetrical, whereas all other mRGC targets receive predominantly contralateral mRGC inputs. This presumably indicates that most uncrossed mRGC axons terminate at the SCN and the vSPZ and do not continue to higher brain centers.

Despite ipRGCs' role in light-induced nocturnal suppression of melatonin release from the pineal gland, these cells do not directly project to this brain region. Instead, they signal through a polysynaptic circuit that begins at the SCN, followed by the paraventricular nucleus of the hypothalamus, the intermediolateral

nucleus of the spinal cord, the superior cervical ganglion, and, finally, the pineal gland.

Conventional Photoreceptors Also Contribute to Non-Image-Forming Photoreception

Although many non-image-forming visual functions do not require rods or cones, the classical photoreceptors nevertheless contribute to their generation in wild-type animals. Together, rod/cone photoreceptors and ipRGCs account for all types of non-image-forming photoreception in the mammalian retina.

Differential Roles of Melanopsin and Classical Photoreceptors

As discussed previously, melanopsin-driven ipRGC photoresponses differ from rod/cone light responses in various respects, and these photosensory systems indeed make largely nonredundant contributions to non-image-forming vision:

1. The rod/cone photoreceptors are more sensitive than melanopsin and have been shown to increase the absolute sensitivities of the pupillary light reflex and circadian photoentrainment. On the other hand, melanopsin is required for maximal constriction of the pupil in response to bright light stimulation. Thus, the dynamic ranges of classical photoreceptors and melanopsin complement each other.
2. Rods and cones generate faster light responses than the melanopsin-based ipRGC response, and they have been shown to shorten the latency of the pupillary light reflex. Classical photoreceptors are also important for the circadian phase-shifting response of mice to short-duration (e.g., 1 min) light pulses that are relatively ineffective in triggering melanopsin responses.
3. Melanopsin photoresponses are far more tonic than the light responses of rods and cones. In mice, melanopsin is required for prolonged light pulses to continuously suppress nocturnal locomotion, and the rod/cone photoreceptors alone can sustain negative masking for no more than 2 h. Similarly, in humans, blue light near the λ_{max} for melanopsin can continuously inhibit the release of pineal melatonin for at least 6 h, whereas longer wavelength light selective for the classical photoreceptors can maintain such inhibition for much shorter durations.

Synaptic Input to ipRGCs

The various brain regions that mediate non-image-forming visual functions receive retinal inputs through conventional RGCs as well as ipRGCs. Thus, these conventional RGCs presumably provide a pathway through which rod/cone signals can reach these brain targets. However, besides functioning as photoreceptors, the ipRGCs also receive synaptic input from rod/cone circuits, as do all other RGCs. Preliminary data have shown that at least for photoentrainment, pupillary constriction, and negative masking, genetically ablating the ipRGCs severely impaired these responses, suggesting that rod/cone signals regulating these behaviors are conveyed primarily through the ipRGCs.

As mentioned previously, ipRGCs extend their dendrites into the inner plexiform layer, the synaptic layer in which bipolar and amacrine cells transmit rod and cone signals to the RGCs. Electron microscopic data have shown that in the mouse retina, melanopsin-immunopositive dendrites are postsynaptic to both amacrine and bipolar cell terminals. The functional roles of these synaptic inputs have been best characterized in the primate and rat ipRGCs.

In primate ipRGCs, light stimulation evokes both a slower melanopsin-based response and a faster synaptically driven response that is mediated mainly by rods in dim light and by cones under light-adapted conditions. The rod-driven responses are depolarizing and fairly sustained, reflecting the relatively limited capacity of the rod photoreceptors to light adapt. These depolarizing rod-driven light responses are observed not only in ipRGCs with dendrites terminating in the 'on' sublamina of the inner plexiform layer but also in those with dendrites stratifying in the 'off' sublamina, suggesting that ipRGCs may receive unconventional 'on' bipolar cell input in the 'off' sublamina. By contrast, the cone-driven responses adapt rapidly, and the polarity of these responses is wavelength dependent: Stimuli selective for the short-wavelength (i.e., 'blue') cones elicit hyperpolarization, whereas longer wavelengths evoke depolarization. As the ipRGCs in primates project to the dorsal lateral geniculate nucleus, such color-opponent responses might lead to conscious color perception, although their relevance to non-image-forming vision remains to be determined.

The pharmacology of the synaptic inputs to ipRGCs has been studied in greater detail in rats. Rat ipRGCs receive both inhibitory GABAergic and excitatory glutamatergic inputs, presumably from amacrine and bipolar cells, respectively. These inputs are spontaneously active and can also be triggered by light. In ipRGCs, both amacrine cells and ON bipolar cells mediate light-evoked responses at light onset, although at the resting membrane potential, the ON bipolar cell input dominates, resulting in depolarizing

responses. In most ipRGCs, OFF bipolar cells also trigger a depolarization at light offset, although this OFF response is always smaller than the ON bipolar cell-driven ON response, suggesting that rat ipRGCs receive stronger synaptic input from the ON bipolar cells than from the OFF bipolar cells. This is surprising because the dendrites of rat ipRGCs terminate mainly in the 'off' sublamina of the inner plexiform layer. When most excitatory and inhibitory inputs are blocked pharmacologically, the intrinsic light response of ipRGCs increases in amplitude, suggesting that these synaptic inputs serve in part to modulate the gain of the intrinsic light response.

Besides generating light-evoked responses in ipRGCs and acutely modulating the melanopsin photoresponse, synaptic inputs have also been shown to induce longer term biochemical changes in rat mRGCs. Dopaminergic amacrine cells have been found to be presynaptic to some mRGCs, and pharmacological activation of the D2 class dopamine receptors upregulates the mRNA levels of both melanopsin and pituitary adenylate cyclase-activating peptide, one of the transmitters released by ipRGCs onto postsynaptic cells in the SCN. Since dopamine is released from the dopaminergic amacrine cells upon light exposure, its effect on melanopsin expression might allow ambient light to modulate the sensitivity of the melanopsin phototransduction cascade.

Evidence for Intraretinal Output from ipRGCs

There is preliminary evidence that ipRGCs can exert synaptic influences on rod/cone circuits within the retina. The first line of evidence has come from measurements of the electroretinogram b-wave, a retinal field potential that reflects ON bipolar cell light responses. The latency of the b-wave in the human electroretinogram is longer at night than during the day. Exposure to prolonged light stimulation at night can shorten the nocturnal b-wave latency through an intraretinal mechanism that appears to be mediated by ipRGCs because it is relatively insensitive and has a λ_{max} of approximately 480 nm. Furthermore, in mice, the amplitude and latency of the b-wave exhibit diurnal rhythms that are much more pronounced in wild-type mice than in melanopsin knockout animals, suggesting the involvement of ipRGCs in the generation of these rhythms. The second line of evidence has been obtained by analyzing light induction of c-fos expression in rodless–coneless mouse retinas. Light induces c-fos expression in neurons in both the inner nuclear layer and the ganglion cell layer, and the total number of these cells is several times greater than the number of mRGCs, suggesting that ipRGCs can signal to other retinal neurons. Finally, prolonged illumination has been shown to induce the release of dopamine in rat retinas in which nearly all of the rod/cone photoreceptors have degenerated, suggesting that the dopaminergic amacrine cells can receive light signals from ipRGCs.

The precise mechanism(s) through which ipRGCs can influence other retinal cells remains unknown. In general, RGCs are considered the output neurons of the retina and have not been shown to be presynaptic in mammalian retinas. It has been proposed that ipRGCs form gap junctions with other cells in the ganglion cell layer, and that some of these coupled cells are amacrine cells capable of modulating retinal functions through the release of neuromodulators.

Nonphotoreceptor Photoreception in Nonmammalian Retinas

Much less is known about non-rod/cone photoreception in nonmammalian vertebrate retinas. Melanopsin is expressed in some horizontal, bipolar, and ganglion cells in chicken, a few horizontal and amacrine and/or ganglion cells in cod and roach, and in some horizontal cells in frog. In addition, two other non-rod/non-cone opsins, the vertebrate ancient (VA) and the VA-long opsins, are expressed in a subset of horizontal cells in roach and zebrafish, respectively. The functional roles of these opsins are unknown. Some chicken RGCs appear to be intrinsically photosensitive because light has been shown to inhibit melatonin synthesis in some isolated RGCs, although it remains to be determined whether such an effect is mediated by melanopsin. Melanopsin and/or VA opsin has been proposed to generate a novel, delayed depolarizing light response in approximately 0.5% of the roach horizontal cells. However, this response is abolished when synaptic transmission is blocked pharmacologically, inconsistent with it being initiated within these cells.

See also: Photoreceptor Adaptation; Photoreceptors and Circadian Clocks; Photoreceptors: Physiology; Phototransduction; Retina: An Overview; Retinal Ganglion Cells: Anatomy; Retinal Development: An Overview; Vision: Light and Dark Adaptation; Visual System Development: Invertebrates; Visual System: Invertebrates.

Further Reading

Berson DM, Dunn FA, and Takao M (2002) Phototransduction by retinal ganglion cells that set the circadian clock. *Science* 295: 1070–1073.

Dacey DM, Liao HW, Peterson BB, et al. (2005) Melanopsin-expressing ganglion cells in primate retina signal colour and irradiance and project to the LGN. *Nature* 433: 749–754.

Dkhissi-Benyahya O, Gronfier C, De Vanssay W, Flamant F, and Cooper HM (2007) Modeling the role of mid-wavelength cones in circadian responses to light. *Neuron* 53: 677–687.

Freedman MS, Lucas RJ, Soni B, et al. (1999) Regulation of mammalian circadian behavior by non-rod, non-cone, ocular photoreceptors. *Science* 284: 502–504.

Graham DG, Wong KY, Pattabiraman K, and Berson DM (2007) *A Bit of Fly in the Mammalian Eye,* Abstract No. 2850. Ft. Lauderdale, FL: Association for Research in Vision and Ophthalmology.

Hankins MW and Lucas RJ (2002) The primary visual pathway in humans is regulated according to long-term light exposure through the action of a nonclassical photopigment. *Current Biology* 12: 191–198.

Hattar S, Kumar M, Park A, et al. (2006) Central projections of melanopsin-expressing retinal ganglion cells in the mouse. *Journal of Comparative Neurology* 497: 326–349.

Hattar S, Liao HW, Takao M, Berson DM, and Yau KW (2002) Melanopsin-containing retinal ganglion cells: Architecture, projections, and intrinsic photosensitivity. *Science* 295: 1065–1070.

Lucas RJ, Freedman MS, Munoz M, Garcia-Fernandez JM, and Foster RG (1999) Regulation of the mammalian pineal by non-rod, non-cone, ocular photoreceptors. *Science* 284: 505–507.

Lucas RJ, Hattar S, Takao M, et al. (2003) Diminished pupillary light reflex at high irradiances in melanopsin-knockout mice. *Science* 299: 245–247.

Panda S, Nayak SK, Campo B, et al. (2005) Illumination of the melanopsin signaling pathway. *Science* 307: 600–604.

Provencio I, Jiang G, De Grip WJ, Hayes WP, and Rollag MD (1998) Melanopsin: An opsin in melanophores, brain, and eye. *Proceedings of the National Academy of Sciences of the United States of America* 95: 340–345.

Tu DC, Zhang D, Demas J, et al. (2005) Physiologic diversity and development of intrinsically photosensitive retinal ganglion cells. *Neuron* 48: 987–999.

Wong KY, Dunn FA, Graham DM, and Berson DM (2007) Synaptic influences on rat ganglion-cell photoreceptors. *Journal of Phsiology* 582: 279–296.

Yoshimura T and Ebihara S (1996) Spectral sensitivity of photoreceptors mediating phase-shifts of circadian rhythms in retinally degenerate CBA/J (rd/rd) and normal CBA/N (+/+) mice. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 178: 797–802.

Non-Primate Models of Normal Brain Aging

C T Siwak-Tapp and P D Tapp, University of California at Irvine, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

An important goal of aging research is the preservation of cognitive function with age, which is becoming more critical as the proportion of aged people continues to rise. Despite successful modeling of numerous specific features of aging, no single animal model fully replicates all aspects of human aging. It is thus necessary to continue developing new models to explore the unique aspects of each known model and combine information from convergent studies. The most commonly used animal models of human brain aging have employed rats, transgenic mice, and nonhuman primates. Companion animals, such as the dog and cat, are now living longer due to advances in animal health care. Consequently, age-related cognitive and behavioral changes comparable to those in humans are now being observed in dogs and cats. The work described herein assesses normal brain aging in other mammalian species, with our primary focus on the dog and cat.

Brain Aging in the Dog

Veterinarians have long been aware of geriatric behavior problems in pet dogs. The field of veterinary behavioral medicine has introduced the term 'canine cognitive dysfunction' to describe age-related behavioral changes not attributable to a medical condition in dogs. Canine cognitive dysfunction is a progressive neurodegenerative disorder of senior dogs characterized by a gradual decline in a variety of behaviors over a long period of time. Five general categories of behavioral changes have been defined: (1) signs of disorientation, that is, becoming trapped in corners or behind furniture; (2) signs of decreased social interactions with family members or changes in recognition; (3) disturbances in the sleep–wake cycle, such as increased daytime sleeping and less nighttime sleeping; (4) loss of house-training; and (5) decreased activity. These behaviors involve learning and memory processes, and changes in such behaviors are used as indicators of cognitive dysfunction.

Evaluation of these behaviors is usually achieved by owners completing a checklist of noticeable changes in their dogs' behavior. These behavioral signs are observed in dogs that were normal when younger; over 60% of owners with dogs aged 11 years and older report one or more behavioral changes consistent with cognitive dysfunction, after other medical conditions have been eliminated. The prevalence of canine cognitive dysfunction increases from 32% of 11-year-old dogs to 100% of 16-year-old dogs. No sex differences have been reported.

Cognition and Behavior in the Aging Dog

The measurement of cognitive functions in the aging beagle dog using neuropsychological tests has been the focus of a series of laboratory studies for more than 15 years. Like humans, cognitive aging in the dog is not a uniform process and individual differences become more common with age. Some cognitive processes remain intact with age while others deteriorate, each at different rates. Additionally, learning and memory impairments can occur independently of one another. Cognitive decline is not, however, an inevitable consequence of aging, as some old dogs perform just as well as young dogs on some tasks. Others show a mild impairment or become severely impaired. This section will discuss the cognitive and behavioral changes that occur with age in dogs.

Age-related learning impairments in dogs In order to administer neuropsychological tests, the dogs must first learn how to perform the procedures. Learning is the process of acquiring knowledge or skills through study or experience and generally becomes more difficult with age. Procedural learning is the acquisition of new skills. The procedural learning task for the dog is designed to teach the animal the skills necessary to perform in the testing apparatus, a modified version of the Wisconsin General Test Apparatus (Figure 1) used for nonhuman primates, and consists of reward–approach and object–approach learning. Dogs learn that there is a food reward, that they can approach the food, and that they must move objects to obtain the food. The procedural learning performance of aged dogs varies with previous life experience. Old dogs raised as companion animals learn these procedures as quickly as do young dogs raised in the same way. Aged dogs raised in a kennel setting, however, are slower to acquire these tasks compared to young kennel-reared animals. This simple task reveals the importance of the rearing environment and previous experience in assessing cognitive ability in aged animals.

In a discrimination learning paradigm, used to assess associative learning, the animal must learn to associate a specific object or set of object characteristics with a food reward. Simple discrimination learning tasks, such as an object discrimination

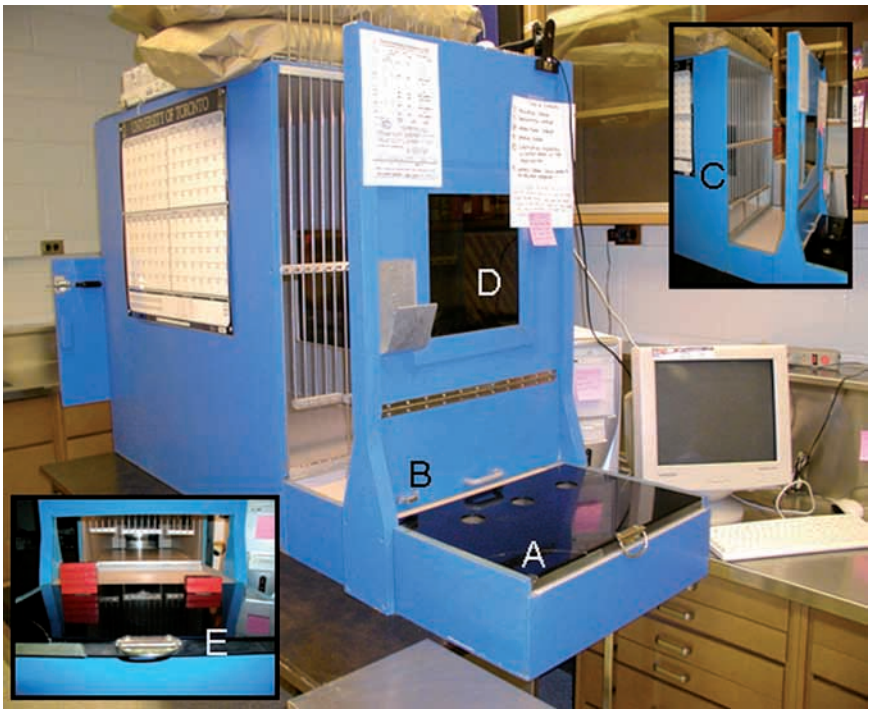

Figure 1 Toronto general test apparatus used to collect cognitive data in dogs. Objects and food rewards are placed on a sliding tray (a) and presented to the animal by lifting a door (b). Adjustable gates (c) allow animals to make responses, which are observed through a one-way mirror (c). An example of the setup for a size discrimination task is shown (e).

whereby two different objects must be compared, are easily acquired by dogs and are not sensitive to age. Increasing the complexity of discrimination problems, however, makes these tasks more sensitive to age differences. For example, the size discrimination task uses two blocks that are identical in all attributes (e.g., color, shape, texture) except size – one is taller than the other. Old dogs are slower to learn this task and make more errors relative to young dogs. A second way to increase task complexity is to increase the number of choices available. The oddity discrimination task uses three objects, two that are identical and a third that has different physical attributes. On this task, the animal must learn to choose the odd object to receive a food reward. Aged dogs are slower to learn this task compared to young dogs. Further testing using objects that appear more and more similar exacerbates the age-related impairment.

Discrimination learning tasks are often followed by reversal learning tasks, whereby the reward contingencies of the objects are switched such that the previously correct object now becomes incorrect. Reversal learning tasks are age sensitive and are often used as measures of executive functions. This will be discussed in further detail in the section titled 'Executive functions in dogs.'

Object recognition learning is more difficult than discrimination learning because correct responses are based on rule learning and not just an association between an object and reward. Object recognition learning involves identifying a previously viewed object. In the delayed-non-matching-to-sample (DNMS) task, dogs are first shown a single object covering a food reward and are allowed to respond and obtain the food. After a delay period, two objects are presented – the original object and a novel object. Each trial of the task uses random sets of objects drawn from a large assortment. The dog must learn the rule of selecting the nonmatching object in the second presentation. The object recognition learning task is very difficult for dogs to acquire. Most young dogs are able to learn the task, but only some aged dogs are successful. The aged dogs that successfully learn the DNMS task perform more poorly relative to the young dogs.

The delayed-non-matching-to-position (DNMP) task is another rule-based learning task that assesses egocentric spatial learning – the ability to locate objects in space with reference to one's own body position. The simple version of the DNMP task is the two-choice (2c) DNMP. In this task, the dog is presented with a single object over one of two lateral food wells. After a delay period, a second identical object is presented over the remaining lateral food well along with the first object still in the same location. The dog must learn to choose the object in the novel location. All young dogs and some old dogs

acquire this task, but a subset of aged dogs will show impairments in spatial learning and perform significantly more poorly. The complex version of the DNMP uses a three-choice (3c) paradigm. In the 3cDNMP, three locations are used (two lateral and one middle well). This version of the task is much more difficult for dogs to learn. Compared to the 2cDNMP, which all young dogs learn, 12% of young dogs are unable to solve the 3cDNMP problem. Similarly, only 18% of aged dogs fail the 2cDNMP but 83% are unable to learn the more complex 3cDNMP task.

The landmark task is another rule-based spatial learning task, but unlike the 2c- and 3cDNMP tasks, the landmark task measures allocentric spatial learning – the ability to locate objects in space with reference to another object or landmark. In the landmark task, dogs are rewarded for selecting the one object, of two objects, that is closest to an external landmark or referent. Thus, information about the correct response is only provided by the landmark. Like the egocentric spatial learning tasks, performance on the landmark discrimination task declines with age in dogs.

Age-related memory impairments in dogs An important observation in the canine model of cognitive aging is that learning and memory are differentially impaired with age. Some aged dogs that successfully learn a task may exhibit memory impairments. Two protocols used to examine memory function in the dog are the variable delay and maximal memory paradigms. Both protocols have been applied to the DNMP and DNMS tasks.

In the variable delay paradigm, the duration between the sample presentation and the matching phase of the task is varied between 20, 70, and 110 s. Each delay occurs an equal number of times per 12-trial session. Increasing the time between the sample and match presentation of a trial increases the demands placed on memory. On both the DNMP and the DNMS tasks, accuracy declines as the length of delay increases in both young and old dogs. The decline represents a working memory impairment since the animal successfully learned the task but has difficulty retaining the information over longer delay periods. Aged dogs perform more poorly than young dogs do at all delays.

The maximal memory paradigm uses a progressively increasing delay period until a preset number of sessions is reached. On the 2cDNMP, young dogs perform accurately at delays of over 210 s while old dogs only reach 30–50 s, indicating they have more difficulty retaining the spatial information as long as young dogs do. With the more difficult 3cDNMP, young dogs usually only reach 110 s and old dogs often fail to complete delay intervals longer than 5 s. Results from the maximal memory paradigm demonstrate that old dogs can acquire spatial information but are unable to retain information in memory for long periods of time. The age differences are more apparent on the DNMS task, in which young dogs respond accurately up to delays of 300 s.

Executive functions in dogs Executive functions describe a set of cognitive abilities that control and regulate other abilities and behaviors. These higher order cognitive processes include inhibitory control, concept abstraction, set shifting, set maintenance, and manipulation. Executive functions can be difficult to assess directly, and many of the tasks used to measure other abilities can be modified to evaluate executive functions.

Discrimination reversal tasks are often used to obtain measures of inhibitory control. Inhibitory control is the ability to inhibit interfering information or previously activated cognitive processes in order to focus attention on the relevant task requirements. On reversal tasks, animals must inhibit the previously correct response to learn a new stimulus–reward relationship using the same objects. In the dog, two important findings are reported on reversal tasks. First, aged dogs are much slower to acquire an object discrimination reversal and a size discrimination reversal compared to young dogs. Second, analysis of the types of errors made on the size discrimination reversal task indicates that the cause of reversal learning impairments changes with age. In old dogs (8–11 years old), reversal learning impairments are caused by an inability to learn the new stimulus–reward contingency. In senior dogs (over 11 years old), reversal deficits are caused by both an inability to learn new reward contingencies and an inability to inhibit a previous response. The perseverative responses observed in senior dogs suggest a deficit in inhibitory control, an executive function.

Another executive function impaired with age in the dog is the manipulation component of working memory. Maintenance involves actively holding information in working memory, while manipulation occurs when additional reorganization or updating of this information is required, an executive function. Tasks that require the manipulation of information held in working memory place greater demands on the cognitive process of working memory. The DNMP and DNMS tasks discussed earlier are tests of working memory performance. The previous protocols described varied the demands on working memory through the time required to hold the information in memory. A second way to change working memory demands is to increase the amount of information to

be retained. The spatial list learning (SLL) task assesses the manipulation component of working memory in dogs.

In the SLL task, dogs are presented with a single red disk randomly located over one of three food wells during the first phase of a trial. After a response is made, a brief delay interval occurs, followed by the second phase of the trial in which two identical red disks are presented, one in the original location and a second in one of the two remaining spatial locations. Only responses to the novel location are rewarded on the second phase, and this component forces the dog to update or manipulate the information held in working memory. After a second delay period, the third phase of the task occurs in which the dog is presented with three identical red disks, two in the familiar locations and a third in the final novel location. Only responses to the novel location are rewarded. Locations of each object are randomized across trials.

Young and aged dogs successfully learn the task but old dogs perform significantly worse than young dogs do. Old dogs take longer to acquire the task and are unable to perform at delays longer than 10 s. Even old dogs that perform well on the 3cDNMP are unable to complete the SLL task. Modified versions of the SLL which place even greater demands on working memory manipulation exacerbate age differences between young and old dogs. Together, the results suggest that age-related impairments on the SLL reflect deficits in manipulation rather than maintenance, since many of the old dogs tested on the SLL task had performed the 3cDNMP with high accuracy. Thus executive processes associated with the manipulation of information in working memory can decline even if the ability to hold information in working memory does not.

Concept learning, the ability to categorize or sort objects based on a particular dimension, is also impaired with age in the dog. In the dog, the size concept task was developed to assess concept learning. On this task, dogs are trained on a series of 2c and 3c size discrimination tasks, using various combinations of the same set of four blocks to teach the animals that size is the feature of interest. Following this training, novel stimuli differing only in size are presented to determine if the dogs learn that size could be used conceptually to solve the new tasks. In tests of size concept learning, young dogs successfully transfer the concept of size to new tests. Older dogs are unable to solve the concept tests, suggesting that dogs are capable of learning conceptual dimensions but the ability declines significantly with age.

Age-related changes in noncognitive behaviors in dogs Changes in noncognitive behaviors with age are another manifestation of the underlying brain pathology that contributes to cognitive dysfunction. Many of the noncognitive changes observed in aging dogs correlate with cognitive status. Locomotor activity, for example, decreases with age in dogs, but this decrease depends upon several factors, including the testing situation, cognitive status, and life experience. Aged dogs raised in a kennel or laboratory-type setting do not show a decrease in locomotor activity in a novel testing environment, such as an open field arena, while pet dogs show a decrease in open field activity with age. In a home area test, however, kennel/laboratory-reared dogs show declines in locomotor activity with age. This suggests that the open field and home area tests are measuring different aspects of locomotor activity. The open field test occurs in a novel setting and locomotor responses may reflect behavioral reactivity rather than spontaneous motor activity. Moreover, aged dogs that are cognitively impaired are also hyperactive in both the novel and the familiar test situation, compared to aged unimpaired dogs. Laboratory/kennel-reared dogs are also more active compared to pet raised or companion animals. Thus, environmental enrichment over the life of a dog may interact with genetic factors and contribute to the extent of locomotor decline and cognitive dysfunction.

Locomotor activity in laboratory dogs follows a circadian activity–rest pattern whereby locomotor activity is high during the day and low at night. This circadian pattern of locomotor activity also varies as a function of age, housing environment, and cognitive status. In young dogs, activity levels are higher for a longer period of time during the day compared to older dogs. In an indoor housing area, aged dogs display shorter bouts of activity during the day, more frequent rest periods, and take longer to become active once the lights are turned on compared to young dogs. These age-related differences are not observed in dogs housed in an outdoor area, suggesting that bright sunlight is effective in synchronizing and consolidating the activity rhythm. Cognitively impaired aged dogs are hyperactive and show fewer periods of rest during the day compared to old unimpaired dogs.

In addition to locomotor activity, we adapted a series of behavioral tests to assess social and exploratory behaviors in the aging dog. These tests are modified versions of the open field test whereby specific social or exploratory stimuli are placed in the arena and the response of the dog is recorded. The expression of social and exploratory behaviors in dogs varies with age and cognitive status. In response to the presence of a person sitting in the open field arena, young dogs display active social responses, including initiating physical contact, climbing, sniffing, and licking. Age-unimpaired dogs are more passive and tend to sit or lay quietly beside the person.

The overall amount of time engaged in the social behaviors does not differ with age; only the types of behaviors show an age-related difference. Cognitively impaired dogs exhibit very little interest in the person in the room and do not engage in social behaviors.

We further measured reactions to conspecifics to assess how social responses within the species may change with age. Three different stimuli were used: a cardboard silhouette of a dog, a sandcast life-size model of a dog, and a mirror. In these tests, dogs treat the cardboard silhouette like a real dog by directing their sniffing to the face and anal regions, locations that are normally investigated when live dogs interact. Young and age-unimpaired dogs habituate rapidly to the silhouette, but cognitively impaired aged dogs spend significantly more time sniffing the cardboard figure. This could reflect a deficit in habituation or stimulus recognition. To make the conspecific more realistic, we used a life-size sandcast model of a golden retriever. Young dogs spend more time sniffing and attempting to interact with the model dog compared to old dogs. Old unimpaired and impaired dogs do not differ in the amount of time spent sniffing the model dog. To make the conspecific more interactive, we also assessed the reaction of the dogs to a mirror reflection of themselves. Reactions to the reflection in the mirror depend on the dogs' ability to recognize their own species. The reactions of the dogs toward the mirror reflection are initially other-directed, treating the reflection like another dog. Responses include jumping at, barking at, and trying to play with the dog in the mirror, as well as attempts to look behind the mirror to find the other dog. Unimpaired old and young dogs habituate to the reflected image rapidly. Old impaired dogs, however, spend significantly more time reacting to the reflection in the mirror, perhaps reflecting a deficit in habituation or stimulus recognition.

Exploratory behavior, or curiosity, is another non-cognitive behavior that changes with age. To assess changes in exploratory behavior in aging dogs, we placed several novel commercially available dog toys in a large test room and measured the amount of time the dogs spent exploring the unfamiliar objects. Results from this test show that young dogs spend more time playing with and sniffing toys than does either of two groups of old dogs. Old-unimpaired dogs show reactions similar to those of young dogs but spend less time exploring objects. Old-impaired dogs generally avoid the toys and express no interest in investigating the objects.

Neuropathology

The aforementioned cognitive and behavioral measures are often combined with analyses of neuropathology to determine whether brain pathology is linked to the age-related changes observed. β-Amyloid deposition and neurofibrillary tangle formation are thought to be related to the cognitive and behavioral decline that typically occurs during aging and have received the most investigation in the dog.

β-Amyloid deposition The dog brain is very well suited for studying early stages of β-amyloid deposition in the aging brain. The earliest and most consistent distribution of β-amyloid (Aβ) plaques in the dog occurs in the prefrontal cortex starting around 8 years of age. By 14 years of age, the entorhinal and parietal cortices are also affected but the cerebellum remains relatively clear. Dogs accumulate the longer, more toxic, less soluble form of Aβ_{1-42} prior to the shorter more soluble Aβ_{1-40} type. Plaque morphology in the dog is generally diffuse and deposits are thioflavine-S negative, suggesting a lack of β-pleated sheet formation. We have found a strong correlation between β-amyloid deposition and cognitive decline in aged dogs.

Tau phosphorylation In contrast to aging humans, dogs do not develop neurofibrillary tangles in the brain but do show early stages of tangle formation characterized by tau phosphorylation and an intracellular punctuate distribution. Hyperphosphorylated tau has been observed in select neuronal populations in older dogs. The absence of neurofibrillary tangles in the dog brain could indicate that the tau protein in the dog is different from the human tau protein. The implication of this finding is that β-amyloid and not tau pathology may be sufficient to lead to cognitive and behavioral dysfunction in the aging dog.

Neuron loss and function Several studies have reported changes related to neuronal integrity with age in the dog. Age-related loss of neurons in the canine brain has been reported in the cingulate gyrus, superior colliculus, and claustrum. Selective neuron loss has also been observed in the prefrontal, temporal, and occipital cortex as well as in the hilus of the hippocampus of aged dogs.

Neurons in the prefrontal cortex of the aged dog brain show other morphological hallmarks of dysfunction, including distorted soma, loss of dendritic spines, shrinkage of dendritic branches, and tortuous apical dendrites. The presence of DNA damage and apoptosis is an additional indicator of neuron dysfunction. In dogs, the number of neurons with damaged DNA increases with age and is positively correlated with β-amyloid load in the frontal cortex.

Oxidative damage Oxidative damage to lipids and proteins in the brain is a common hallmark of aging and may compromise neuron function. In aging dogs,

malondialdehyde levels, a marker of lipid peroxidation, increase with age in the prefrontal cortex and blood serum in the dog starting around 8 years of age. Glutamine synthetase activity, an endogenous antioxidant that protects against oxidative damage, decreases with age in the prefrontal cortex. These age-related changes in oxidative damage are concomitant with increased β-amyloid deposition in the prefrontal cortex.

A role of oxidative damage in age-related cognitive dysfunction was demonstrated in a longitudinal study in dogs. Cognitive function was significantly improved and maintained in aged dogs fed a food fortified with antioxidants and mitochondrial cofactors for almost 3 years, compared to aged dogs given a control food. The fortified food contained higher amounts of vitamin E, vitamin C, *l*-carnitine, and *dl*-lipoic acid as well as a 1% inclusion of spinach flakes, tomato pomace, grape pomace, carrot granules, and citrus pulp. In addition to improving cognition and behavior, the fortified food also led to a reduction in β-amyloid deposition in some brain regions in the aged dogs. This suggests that oxidative damage is an important contributor to cognitive dysfunction and neuropathology in normal aging, and dietary interventions may significantly reduce these events.

Neuroanatomy and Hemodynamics

Noninvasive brain imaging technology permits analyses of metabolic, vascular, and structural changes associated with brain aging. Two common methods available to evaluate volumetric changes in the brain include region-of-interest (ROI) planimetry and voxel-based morphometry (VBM) using brain images acquired from magnetic resonance imaging (MRI). ROI planimetry involves manually tracing the particular brain region of interest on several consecutive images to obtain an estimate of total volume of that region. Studies using ROI planimetry reveal that global cortical atrophy and ventricular enlargement are characteristic features of the aged dog brain. Further analysis shows that the total volume of the brain, and in particular the volume of the frontal lobes and hippocampus, will decrease with age. The volume of the occipital cortex does not change with age. Moreover, these age-related decreases in frontal lobe volume correlate with increased β-amyloid deposition and measures of executive functions.

By contrast, VBM is an automated technique that performs a voxel-by-voxel comparison of local densities of gray and white matter on the brain images. In the dog brain, VBM analysis has revealed gray matter reductions bilaterally in the frontal lobes, parietal lobes, temporal lobes, thalamus, cerebellum, and brain stem. Many of the differences in structural brain aging are sexually dimorphic. In particular, aged male dogs showed a greater reduction in the frontal lobes while changes in the temporal lobes were more apparent in females. White-matter reductions with age were observed in the internal capsula and cranial nerve bundles of aged male dogs and the alveus of the hippocampus in aged females. The additional information provided by the VBM analysis suggests that brain aging in the dog varies regionally and is sex dependent.

MRI can also be used to monitor longitudinal changes in the aging dog brain using serial MRI techniques to coregister images from different time points. With this technique, we have found that dogs spontaneously develop lesions, resembling small lacunes, with age. The majority of lesions in the aged dog brain form in the frontal cortex and caudate nucleus.

A technique called dynamic susceptibility contrast (DSC)-enhanced MRI can be used to study vascular characteristics under the same conditions as anatomical MRI scans. Administration of contrast agents during DSC-MRI procedures permits kinetic modeling to obtain measures of vascular volume and permeability. In the dog, DSC-MRI revealed a decline in regional cerebral blood volume (rCBV) with age in the gray- and white-matter regions. These changes were greater in the white matter compared to the gray matter, even though gray-matter rCBV was consistently higher relative to white-matter rCBV. The decreased rCBV and increased blood–brain barrier permeability developed in tandem with cortical atrophy in this study.

Brain Aging in the Cat

Cognition and Behavior

Many of the behavioral changes typical of cognitive dysfunction in dogs also occur in elderly cats, but the age of onset may be later in cats. Information on behavioral changes with age in cats comes largely from observations in veterinary practice. Behavioral changes observed in aged cats that are consistent with feline cognitive dysfunction include signs of wandering, confusion, excessive vocalization, decreased grooming, lethargy, decreased awareness, greater time sleeping or other disruptions in the sleep/wake cycle, inappropriate elimination or loss of house-training, increased displays of affection or other alterations in social interactions, decreased recognition of owners, and increased anxiety. Estimates of prevalence suggest that 35% of aged cats, once those with possible underlying medical causes were removed, have clinical signs related to cognitive dysfunction. In cats 11–15 years old, the most common change was in social interactions, while changes in cats over 15 years of age primarily

affected activity and vocalization. Laboratory studies of cats have reported altered patterns of habituation, increased reactivity to auditory stimuli, and superior performance on spatial reversal learning tests with increasing age. No decrements in fine motor coordination were observed and locomotor activity levels did not change with age.

Neuropathology

Studies of brain pathology in cats indicate that cats develop β-amyloid and tau abnormalities. In aged cats, the β-amyloid plaques are of the diffuse variety and are localized in the prefrontal cortex, hippocampus, parahippocampal gyrus, occipital cortex, and parietal cortex. The diffuse plaques contain predominantly the $A\beta_{1-42}$ species, consistent with findings in dog and human brain tissue, but the morphology of the diffuse plaques in cat brain is different than what is usually observed in the human brain.

Hyperphosphorylated tau is also present in the aged cat brain. The pattern is consistent with that in the human brain, including area CA1 of the hippocampus, subiculum, and entorhinal/parahippocampal cortex. Aged cats with clinical signs of cognitive dysfunction are frequently positive for dystrophic neurites, but not all exhibit phosphorylated tau accumulation within neuronal cell bodies.

Brain Aging in Other Mammals

The dog and cat are more widely studied than are other mammals, likely due to the prevalence of these animals as pets. Some isolated studies exist that have examined brain aging in other animals, and are described in the following sections.

Rabbit

Eyeblink classical conditioning (EBCC) in older rabbits is used as a model system for studying the neurobiology of learning, memory, and the aging brain. Age-related impairments in EBCC occur relatively early in aging in both rabbits and humans. EBCC is also sensitive for identifying patients with Alzheimer's disease and is abnormal in Down's syndrome individuals over 35 years of age.

Investigation into the aged rabbit brain shows cell loss and gliosis in the cerebellum and hippocampus but no evidence of β-amyloid deposits or abnormal tau accumulations, even in rabbits over 7 years of age. Manipulating dietary cholesterol, however, induces neuronal accumulation of β-amyloid in New Zealand white rabbits, and this increases when copper is administered in the drinking water. Thioflavine-S-reactive features are also observed in animals supplemented with copper. Emerging evidence implicates cholesterol in the production and accumulation of β-amyloid by the 'mismetabolism' of the amyloid precursor protein. This mismetabolism causes the 42-amino-acid form of β-amyloid to form oligomers that, in turn, start a chain of events leading to the accumulation of amyloid plaques. Thus, the rabbit may be a valuable model to study initiating factors and perhaps the role of cholesterol in the deposition of β-amyloid.

Horse

Learning studies with horses are designed to provide a better understanding of trainability in this animal. Discrimination learning ability has been examined in horses using visual cues in a 3c test with food as the reward. In this study, quarter horses learned faster than thoroughbreds, and both groups learned a second discrimination task faster than the first task. In both tasks, however, there was a significant negative correlation between age and rate of learning. Separate neuropathological analysis of the brains of horses up to 27 years of age showed that all regions were devoid of hyperphosphorylated tau. No reports have been published to our knowledge about the existence of β-amyloid in the brains of horses.

β-Amyloid and Tau in Other Mammals

There are various reports of naturally occurring β-amyloid deposition or neurofibrillary tangle-like formations in several other mammals. β-Amyloid deposition has been reported in the brains of camels, coyotes, and polar bears. Tangle-like pathology has been observed in brains of sheep, goats, bears, wolverines, guanacos (similar to llamas), bison, and reindeer. To our knowledge there are no data available concerning behavioral changes with age in any of these animals.

Conclusion

Additional animal models for human brain aging have new insights to offer. A complete replication of human aging does not exist in a single animal system, but a more comprehensive picture emerges when information from several animal models is consolidated and integrated.

See also: Aging of the Brain; Amyloid: Vascular and Parenchymal; Cognition in Aging and Age-Related Disease; Executive Function and Higher-Order Cognition: Assessment in Animals; Oxidative Damage in Neurodegeneration and Injury; Strategic Control of Memory; Voxel Based Morphometry; Working Memory: Capacity Limitations.

Further Reading

Chan AD, Nippak PM, Murphey H, et al. (2002) Visuospatial impairments in aged canines (*Canis familiaris*): The role of cognitive-behavioral flexibility. *Behavioral Neuroscience* 116(3): 443–454.

Cotman CW, Head E, Muggenburg BA, et al. (2002) Brain aging in the canine: A diet enriched in antioxidants reduces cognitive dysfunction. *Neurobiology of Aging* 23(5): 809–818.

Head E, Callahan H, Cummings BJ, et al. (1997) Open field activity and human interaction as a function of age and breed in dogs. *Physiology & Behavior* 62(5): 963–971.

Head E, Milgram NW, and Cotman CW (2001) Neurobiological models of aging in the dog and other vertebrate species. In: Hof PR and Mobbs CV (eds.) *Functional Neurobiology of Aging*, pp. 457–465. San Diego: Academic Press.

Head E, Moffat K, Das P, et al. (2005) Beta-amyloid deposition and tau phosphorylation in clinically characterized aged cats. *Neurobiology of Aging* 26(5): 749–763.

Levine MS, Lloyd RL, Fisher RS, et al. (1987) Sensory, motor and cognitive alterations in aged cats. *Neurobiology of Aging* 8(3): 253–263.

Mader DR and Price EO (1980) Discrimination learning in horses: Effects of breed, age and social dominance. *Journal of Animal Science* 50(5): 962–965.

Milgram NW, Head E, Weiner E, et al. (1994) Cognitive functions and aging in the dog: Acquisition of nonspatial visual tasks. *Behavioral Neuroscience* 108(1): 57–68.

Milgram NW, Head E, Zicker SC, et al. (2005) Learning ability in aged beagle dogs is preserved by behavioral enrichment and dietary fortification: A two-year longitudinal study. *Neurobiology of Aging* 26(1): 77–90.

Ruehl WW and Hart BL (1998) Canine cognitive dysfunction. In: Dodman N and Shuster L (eds.) *Psychopharmacology of Animal Behavior Disorders*, pp. 283–304. Malden, MA: Blackwell Scientific Publications.

Siwak CT, Tapp PD, and Milgram NW (2001) Effect of age and level of cognitive function on spontaneous and exploratory behaviors in the beagle dog. *Learning & Memory* 8(6): 317–325.

Siwak CT, Tapp PD, Zicker SC, et al. (2003) Locomotor activity rhythms in dogs vary with age and cognitive status. *Behavioral Neuroscience* 117(4): 813–824.

Sparks LD (2004) Cholesterol, copper, and accumulation of thioflavine S-reactive Alzheimer's-like amyloid beta in rabbit brain. *Journal of Molecular Neuroscience* 24(1): 97–104.

Su MY, Tapp PD, Vu L, et al. (2005) A longitudinal study of brain morphometrics using serial magnetic resonance imaging analysis in a canine model of aging. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 29(3): 389–397.

Tapp PD, Siwak CT, Estrada J, et al. (2003) Size and reversal learning in the beagle dog as a measure of executive function and inhibitory control in aging. *Learning & Memory* 10(1): 64–73.

Tapp PD, Siwak CT, Gao FQ, et al. (2004) Frontal lobe volume, function, and beta-amyloid pathology in a canine model of aging. *Journal of Neuroscience* 24(38): 8205–8213.

Tapp PD, Chu Y, Araujo JA, et al. (2005) Effects of scopolamine challenge on regional cerebral blood volume. A pharmacological model to validate the use of contrast enhanced magnetic resonance imaging to assess cerebral blood volume in a canine model of aging. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 29(3): 399–406.

Tapp PD, Head K, Head E, et al. (2006) Application of an automated voxel-based morphometry technique to assess regional gray and white matter brain atrophy in a canine model of aging. *Neuroimage* 29(1): 234–244.

Tapp PD and Siwak CT (2006) The canine model of human brain aging: Cognition, behavior, and neuropathology. In: Conn PM (ed.) *Handbook of Models for Human Aging*, pp. 415–434. Amsterdam: Elsevier.

Woodruff-Pak DS and Trojanowski JQ (1996) The older rabbit as an animal model: Implications for Alzheimer's disease. *Neurobiology of Aging* 17(2): 283–290.

Noradrenaline

R D Wassall, University of Oxford, Oxford, UK
N Teramoto, Kyushu University, Fukuoka, Japan
T C Cunnane, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

From its humble beginnings, with biochemical analysis of adrenal glands and neurotransmitter release, we have come to realize that sympathetic neurotransmission is not quite as simple as first thought. We are taught that noradrenaline (norepinephrine) is the major neurotransmitter released from sympathetic nerves, and that with adrenaline (epinephrine), it is a key player in the 'fight-or-flight' response. Indeed, few people can escape the discussion of the 'adrenaline-high' associated with athletes and thrill-seekers alike, and few of us will avoid the use of drugs to alter sympathetic signaling for cardiovascular problems, anxiety or depression as we age. We might think that such an important and familiar neurotransmitter system would be well characterized, but it has not always been so. In the past, noradrenaline proved to be an elusive agent to identify when the concepts of 'neurochemical transmission' were being formed, and even today, there are a myriad of questions surrounding its co-storage, release, and postjunctional receptor mechanisms. Here, we give a brief overview of some of the facts, and unravel some of the mysteries of this celebrated molecule.

A Brief History of the Discovery of Noradrenaline

Galen (AD 130–200), arguably the greatest anatomist and physiologist of antiquity, first described the gross anatomical features of autonomic nerves. He suggested that, being hollow, they allowed the transfer of so-called 'animal spirits' between organs, producing the phenomena of 'sympathy', which was a vague term used to describe the coordination or cooperation of organs. However, the anatomical division of the parasympathetic and sympathetic nervous systems can be traced to Eustachius in the sixteenth century, who regarded the sympathetic and vagus nerves as separate, although describing incorrectly that the 'sympathetic nerve' was a continuation of the abducens (cranial nerve VI). Through the work of Winslow, Whytt, and Bichat two centuries later, an important outline of the sympathetic nervous system was laid out, with the visceral organs controlled by the ganglia, whereas voluntary action was concerned with the brain.

The first functional studies on true sympathetic nerves began with Bernard, who dilated vessels by sectioning the sympathetic nerves, and Brown-Séquard, who contracted the vessels by stimulating the cut end of the nerve. Irritation of various neuronal structures had been shown to increase the heart rate and its contractility; however, modern work on the autonomic nervous system and its functional divisions has been based largely on the work of Gaskell and Langley, who defined the essentials of the macroscopic functions of the sympathetic nervous system in the early to mid-1900s. Langley's student, Elliott, showed that adrenaline, discovered approximately 20 years earlier by Bates, had the same general effect as stimulation of the sympathetic nerves.

The chemical nature of sympathetic neurotransmission had been assumed by Elliott and Dale, who showed the sympathetic-like effects of ergot alkaloids, and when Loewi monumentally demonstrated the chemical nature of neurotransmission in a frog's heart, later found to be due to acetylcholine, it seemed that adrenaline would be an attractive candidate neurochemical in the sympathetic nervous system. The isolation of 'sympathin' from sympathetic nerves by Cannon revealed that adrenaline was similar to, but chemically distinct from, the neurotransmitter released, and its identity remained elusive until von Euler revealed that the true neurotransmitter was the non-methylated derivative of adrenaline, noradrenaline.

Structure and Biosynthesis of Noradrenaline

Noradrenaline, a catecholamine, is derived from L-tyrosine, an aromatic amino acid present in the body fluids and taken up by noradrenaline-producing cells. Through various intermediate steps (Figure 1), L-tyrosine is converted to noradrenaline and, finally, to its methylated form, adrenaline, in phenylethanolamine *N*-methyltransferase-containing cells. The first cytosolic enzyme, tyrosine hydroxylase, is inhibited by excess production of noradrenaline and is the rate-limiting step in the regulation of noradrenaline synthesis. Tyrosine hydroxylase is found only in the cytosol in catecholamine-containing cells and is used as a marker for detection of adrenergic neurons. Dopamine β-hydroxylase is also used as a selective marker for catecholamine-containing cells, but it is located in the secretory vesicles, usually membrane bound; however, a small amount is soluble and released with the vesicle contents upon exocytosis.

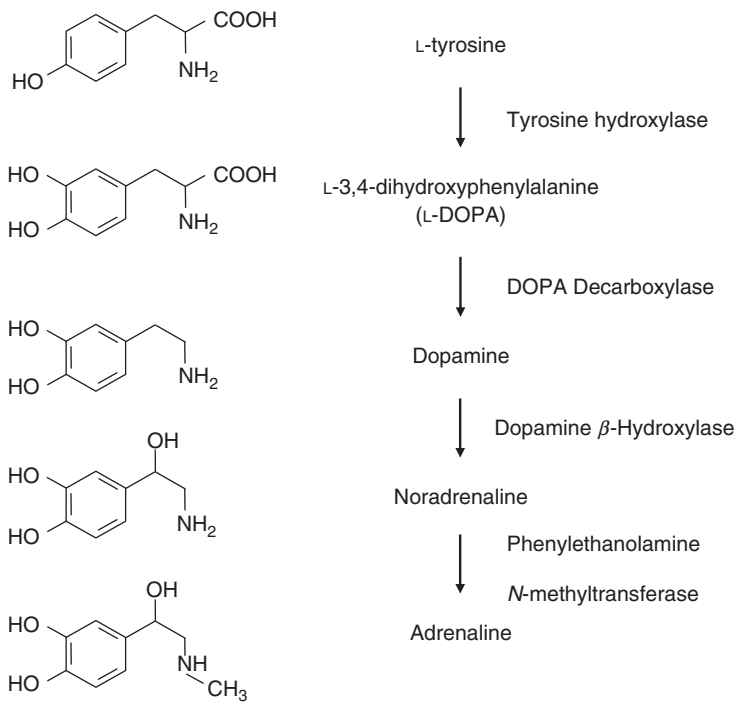

Figure 1 Biosynthetic pathway of noradrenaline and adrenaline.

Where Is Noradrenaline Found?

Noradrenaline is released with adrenaline into the blood from the medullae of the adrenal glands where it can act as a circulating hormone, but it is classically thought of as the main neurotransmitter released from postganglionic sympathetic neurons. Sympathetic fibers project to the heart and along blood vessels, controlling cardiovascular responses to maintain blood pressure, and, with the aid of circulating adrenaline, regulate bronchodilation, glycogen and fat metabolism, thermogenesis, and secretions of hormones and mucus membranes (Table 1). There are also large networks of noradrenergic and adrenergic neurons originating in or close to the locus coeruleus and reticular formation in the brain stem, extensively projecting to the cortex, hippocampus, and cerebellum, and these neurons are involved in baroreceptor and blood pressure reflexes, and also in complex behaviors such as arousal and mood. Many of these tissues are being used to study noradrenaline release and reuptake, but the vas deferens offers the ideal system in which to study sympathetic neuronal release due to its ease of isolation, dense sympathetic innervation, high noradrenaline content, and exquisite sensitivity to pharmacological manipulation.

It is perhaps useful at this stage to remind the nonspecialist reader of the anatomy of the autonomic nervous system. As shown by Langley, autonomic nerves typically have both a preganglionic nerve traveling from the central nervous system and a postganglionic nerve traveling to the effector organ. Parasympathetic fibers either arise from specific cranial nuclei and travel in the occulomotor (III), facial (VII), glossopharyngeal (IX), and vagus (X) cranial nerves or leave the spinal cord at sacral levels S2–S4 before synapsing close to the tissues in collections of cell bodies known as ganglia, and continuing by way of short postganglionic fibers to the effector cells. Sympathetic neurons, however, are generally characterized by short preganglionic fibers, exiting the spinal cord at thoracolumbar levels T1–L3, synapsing in the paravertebral chains, and continuing to the effector organs as long postganglionic fibers (possibly more than 1 m long).

The innervation of the rodent vas deferens, a well-characterized model system, is unusual in that it is supplied by short postganglionic sympathetic neurons whose cell bodies lie within the hypogastric ganglia situated close to the prostatic end. These postganglionic nerves are mainly nonmyelinated axons (0.2–2 μm in diameter) embedded in Schwann cells, together with a small number of fine myelinated fibers (1 or 2 μm in diameter), which travel through the connective tissue and divide into numerous branches as they enter the prostatic end. Traversing the connective sheath, branches pass into the smooth muscle layers, splitting into smaller bundles of two to eight axons, and become varicose. The varicosities, which are approximately 1 or 2 μm in length and 1–1.5 μm in cross section,

Table 1 Summary of major adrenoceptor signaling and some drugs that can affect it

Adrenoceptor subtype	α_1	α_2	β_1	β_2	β_3
Primary signal transduction	Stimulation of phospholipase C ($G_{q/11}$) ↑IP_3, ↑DAG, and ↑Ca^{2+}	Inhibition of adenylyl cyclase ($G_{i/0}$) ↓cAMP	Stimulation of adenylyl cyclase (G_s) ↑cAMP	Stimulation of adenylyl cyclase (G_s) ↑cAMP	Stimulation of adenylyl cyclase (G_s) ↑cAMP
Secondary signal transduction	Stimulation of phospholipase D	Stimulation of adenylyl cyclase[a]	Stimulation of guanylyl cyclase	Stimulation of guanylyl cyclase	Stimulation of guanylyl cyclase
Major physiological effects	• Vasoconstriction • Relaxation of gastrointestinal tract smooth muscle • Contraction of seminal tract and uterus • Contraction of iris radial muscle • Glycogenolysis in liver • K^+ release from salivary glands	• Vasoconstriction • Relaxation of gastrointestinal tract smooth muscle • Decreased release of acetylcholine and noradrenaline from autonomic nerve terminals • Decreased insulin release • Platelet aggregation • Inhibition of sympathetic outflow in brain stem	(Actions mainly on heart) • Positive inotropy • Positive chronotropy • Positive dromotropy • Amylase secretion from salivary glands • Renin release from juxtaglomerular cells of the kidney	• Vasodilation • Bronchodilation • Relaxation of gastrointestinal tract smooth muscle • Relaxation of seminal tract and uterus • Relaxation of ciliary muscle • Glycogenolysis in liver and skeletal muscle • Increased release of noradrenaline from nerve terminals	• Lipolysis and thermogenesis in adipose tissue • Thermogenesis in skeletal muscle
Selective agonists	Phenylephrine Cirazoline	Clonidine Medetomidine	Dobutamine Xamoterol	Salbutamol Salmeterol	BRL 37344 ZD 7114
Selective antagonists	Prazosin Doxazosin	Yohimbine Idazoxan	Atenolol Metoprolol	Butoxamine ICI 118,551	SR 59230A

[a]The role of stimulating adenylyl cyclase in addition to its inhibition by α_2-adrenoceptor activation in some systems remains unknown. For simplicity, subtypes of receptors are not shown. IP_3, inositol-1,4,5-trisphosphate; DAG, diacylglycerol; cAMP, 3′-5′-cyclic adenosine monophosphate.

are packed with vesicles and mitochondria and are separated by intervaricose regions approximately 3–5 μm in length. The diameter of the axon spanning the intervaricose regions is only 0.1 μm, and each axon gives rise to between 10 000 and 30 000 varicosities (Figure 2). The conduction velocity in these nerves is less than 1 m s^{-1}. In the vas deferens, the varicosities traverse the smooth muscle cells in complex patterns, such that at least one varicosity makes close contact (approximately 20–50 nm) with each smooth muscle cell, with the varicosities being the sites where noradrenaline is synthesized and released.

How Can Noradrenaline Release Be Measured?

Early electrophysiological investigations into sympathetic nerves relied on intracellular recording. Basing the technique on the Nobel prize-winning work of Hodgkin, Huxley, and Katz, the nerve cell bodies in ganglia were impaled using glass microelectrodes and electrical changes across the membrane were measured. As at the neuromuscular junction, electrical signals could be detected in the absence of stimulation (excitatory postsynaptic potentials in nerves and end plate potentials in the motor end plate), as could the resulting action potentials on stimulation. Although this method gave fascinating insights into the release of acetylcholine, the primary neurotransmitter at these synapses, it could not reveal anything about the terminal release of noradrenaline from the postganglionic nerve. Electrophysiological measurements from nerve terminal varicosities proved difficult due to their small size and the large microelectrode diameters; progress remained slow.

In an effort to study and localize noradrenaline, a number of methods have been employed, including high-performance liquid chromatography, overflow, radiolabeling, and voltametry techniques. These approaches all indicated a neuronal release of noradrenaline, the most abundant neurotransmitter in sympathetic neurons, and that noradrenaline was stored in the vesicles in preparation for exocytosis. However, fractional release studies showed that

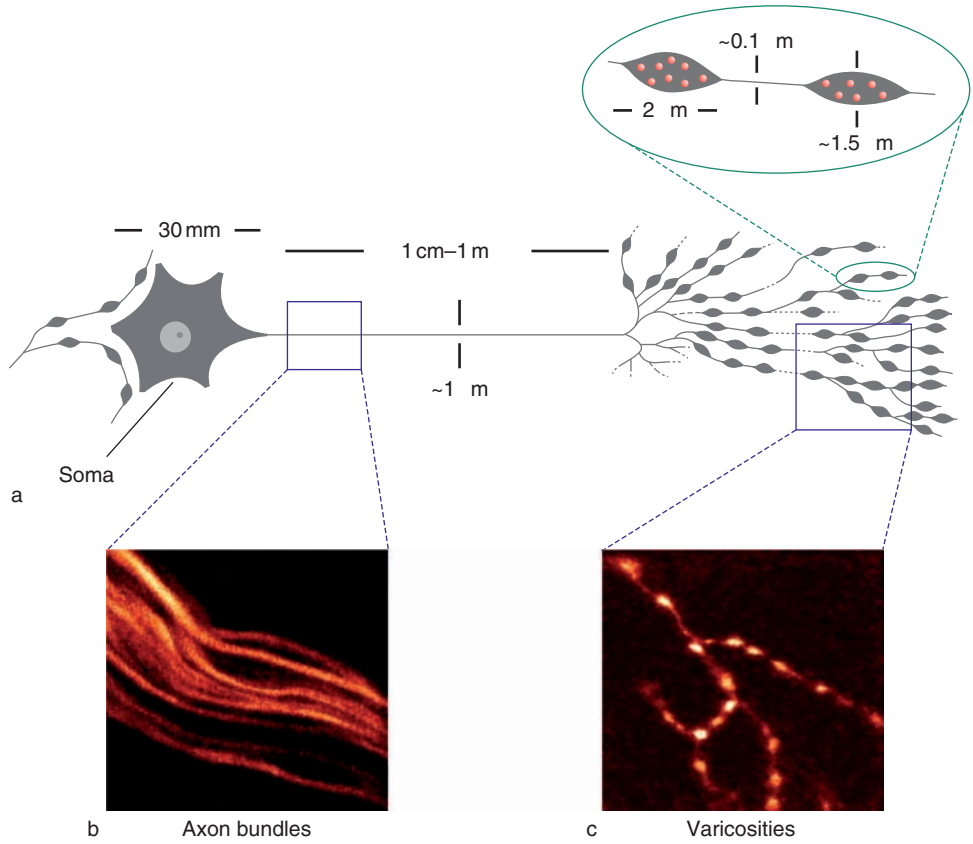

Figure 2 (a) Schematic diagram of a postganglionic sympathetic neuron showing the difference in size between cell body, axon, and varicosity. (b) A set of axon bundles loaded with the Ca^{2+} indicator Oregon Green 488 BAPTA-1 dextran. Note that only a single axon is represented in the corresponding box in (a). (c) Image of part of a sympathetic terminal loaded with the Ca^{2+} indicator Oregon Green 488 BAPTA-1 dextran. Reproduced from Jackson VM and Cunnane TC (2001) Neurotransmitter release mechanisms in sympathetic neurons: Past, present, and future perspectives. *Neurochemical Research* 26: 875–889, with permission from Springer Science and Business Media.

there was a disparity between the amount of noradrenaline released and the noradrenaline content of a vesicle.

Although these methods did not have the temporal or spatial resolution at the time to investigate the release characteristics of noradrenaline from a single varicosity, biochemical analysis of these studies provided two radically different hypotheses to explain the mechanism of noradrenaline release from the 'average' varicosity.

First, the 'fractional release' model suggested that the action potential only released a small fraction of the contents of one vesicle (1–3%) from each varicosity every time the nerve was stimulated, similar to the modern 'kiss-and-run' mechanism observed at some synapses in the central nervous system. In this scenario, on average 30–100 action potentials are required to release the equivalent of the neurotransmitter content of one vesicle from each varicosity, and this would mean that there is a constant release of neurotransmitter during neuronal activity providing an average 'biophase' concentration which bathes the smooth muscle and maintains a tonic contraction. With the assumption that the average varicosity has 1000 vesicles, this would mean that only approximately 300–400 molecules of transmitter would be released with each action potential, which should nonetheless be sufficient to reach maximal concentrations for receptor activation in a close contact neuroeffector junction.

Today, many textbooks still favor this explanation (as did the early investigators) because it appears to be a more efficient and economical way of transmitter turnover to achieve a given effector response while at the same time reusing the vesicles and allowing for efficient noradrenaline reuptake (approximately 90%). However, this would mean that sympathetic nerves would not behave like all other nerves, where quantal transmission was accepted.

An equally radical hypothesis is the 'intermittent release' model, which suggested that each action potential activates only 1–3% of all varicosities in the

tissue. When the release machinery was activated, an individual varicosity discharged a single packet of neurotransmitter equivalent to the neurotransmitter content of a single vesicle. In this case, neurotransmitter release would be highly intermittent at the level of the individual varicosity. This would cause approximately 50-fold higher noradrenaline concentrations in the junction, giving a larger 'margin of safety' for transmission, but would demand extreme efficiency of the local noradrenaline reuptake mechanisms to clear noradrenaline and terminate the response before it diffused away.

If this theory were true, it would raise more questions about why and how a varicosity would 'decide' to release neurotransmitter or not. It was understandable that the investigators favored the previous explanation. If this controversy was to be settled, then techniques with better resolution needed to be developed.

What Can We Learn About Sympathetic Neurotransmitter Release from Electrophysiology?

Burnstock and Holman, working in Melbourne, Australia, used one of the most densely innervated sympathetic tissues in the body, the vas deferens, to study neurotransmitter release. Sympathetic nerves in the body typically have low-frequency firing rates, and unlike the overflow studies, which require high frequencies to cause measurable overflow of transmitter, electrophysiology allowed the measurement of release on an impulse-to-impulse basis. Using microelectrode techniques similar to those used in ganglia, they impaled smooth muscle cells of the guinea pig vas deferens and recorded resting membrane potentials. In response to sympathetic nerve stimulation, transient depolarizations were recorded, termed excitatory junction potentials (EJPs). EJPs had a fairly slow time course (0.5–1 s), were graded with stimulus intensity, and increased in amplitude in response to the first few stimuli in a train (i.e., they exhibited facilitation). Perhaps more important, in the absence of stimulation, spontaneous EJPs (SEJPs) were recorded, which were reminiscent of miniature end plate potentials (MEPPs) at the skeletal neuromuscular junction that had famously been used to describe the phenomena surrounding quantal neurotransmission. SEJPs implied that neurotransmitter was released in multimolecular packets; however, unlike MEPPs and end plate potentials in skeletal muscle, EJPs and SEJPs had different time courses and could not be compared directly. The time course differences were probably a reflection of the tight electrical coupling between smooth muscle cells of the vas deferens, which caused rapid dissipation of charge from the point source, and also that the EJP represented the sum of electrical activity at several simultaneous release sites being activated throughout the muscle. Taken together with the overflow studies, and the difficulty in interpreting the results for a single release site, the electrical signal seemed to support the fractional release model.

High-Resolution Studies of EJPs

As the fractional release model enjoyed prime status, new methods were devised to break down EJPs into their component parts. Blakeley and Cunnane noted discontinuities between the rising phases of individual EJPs in the mouse and guinea pig vas deferens, and, on electronically differentiating the rising phases, these discontinuities were extracted as transient peaks in the rate of depolarization of the EJP, termed 'discrete events' (Figure 3). On comparing discrete events generated from SEJPs and EJPs, it was found that they matched almost perfectly, suggesting that release was intermittent and monoquantal. Various groups quickly verified the data, and so the fractional release model was moved aside for the intermittent model. Perhaps, in hindsight, it should not have come as a surprise that sympathetic nerves behave like all other nerves in the body and exhibit quantal or packeted neurotransmission.

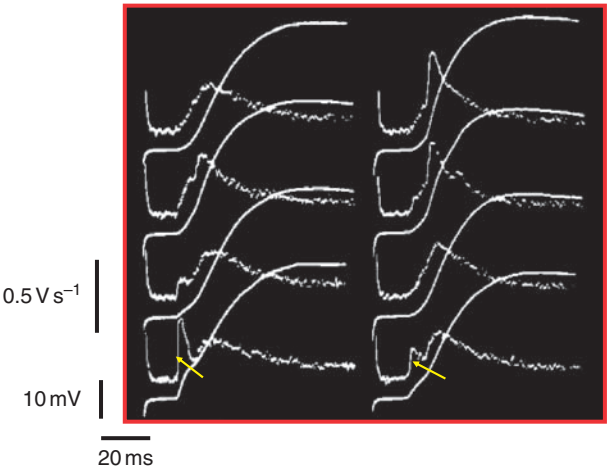

Figure 3 Intracellular recordings of the membrane potential and its time derivative in single smooth muscle cells of the guinea pig vas deferens. The preparation was stimulated at 0.91 Hz and the transient peaks in the rate of depolarization (the 'discrete event') occur intermittently, showing that action potential evoked neurotransmitter release is intermittent. Reproduced from Blakeley AG and Cunnane TC (1979) The packeted release of transmitter from the sympathetic nerves of the guinea-pig vas deferens: An electrophysiological study: *Journal of Physiology* 296: 85–96; and Jackson VM and Cunnane TC (2001) Neurotransmitter release mechanisms in sympathetic neurons: Past, present, and future perspectives. *Neurochemical Research* 26: 875–889, with permission from Springer Science and Business Media.

Mechanisms of Intermittence

Sympathetic nerves proved to be difficult to comprehend. If they only released once in every 100 stimuli or so, then how could we use them to study release? We need to understand some of the mechanisms behind intermittence before we can investigate noradrenaline release further. There were two suggestions put forward to explain intermittence:

- Intermittent failure of the action potential to propagate throughout the complex nerve terminal network of branching axons and varicosities
- A very low probability of neurotransmitter release in the invaded varicosities

To resolve this conflict, focal extracellular measurement was developed by Brock and Cunnane. By moving an electrode close to the tissue and applying gentle suction, a small area could be 'isolated' from the tissue and drugs applied either inside or outside the electrode, which was perfused with physiological solution. It allowed the simultaneous measurement of the nerve terminal impulse (NTI) and also neurotransmitter release – the excitatory junction current (EJC). In the absence of stimulation, spontaneous EJCs were recorded, and following electrical nerve stimulation it was found that the action potential (NTI) arrived each time, whereas neurotransmitter release was highly intermittent (Figure 4). Application of tetrodotoxin, an inhibitor of voltage-gated Na^+ channels, abolished the NTI and evoked neurotransmitter release, but did not abolish spontaneous EJCs. Intermittence of neurotransmitter release was concluded to be due to some failure in depolarization–secretion coupling after the sympathetic nerve terminal is invaded by the action potential rather than action potential propagation failure in the complex terminal arborization of varicosities.

One of the more interesting observations occurred when pharmacological agents were used to manipulate neurotransmission. Following blockade of α-adrenoceptors, the major type of receptors for noradrenaline in the vas deferens, there was an increase in the overflow of noradrenaline and also an increase in the electrical signals recorded with microelectrodes in the smooth muscle cells. It transpired that there were receptors located on nerve terminals, α_2-adrenoceptors, that were distinct from postjunctional adrenoceptors (mainly α_1-adrenoceptors) and limited the release of noradrenaline. This led to the novel concept of α_2-adrenoceptor-mediated inhibition of neurotransmitter release in the great majority of sympathetically innervated tissues and added a new dimension to our understanding of the control of sympathetic nerve function. Previously, the centrally determined patterns of nerve firing were primarily thought to control neurotransmitter release, but now it was shown that neurotransmitter release could be controlled locally due to feedback inhibition of the action of released noradrenaline at distinct receptors on nerve terminals producing a brake on further neurotransmitter release. There was no doubt intermittence could be regulated by this local inhibition at the level of the varicosity, and although this idea was largely born in the peripheral nervous system, it is generally accepted that presynaptic receptors exist at every synapse. Indeed, we now know that many nerves have both inhibitory and excitatory receptors located on their nerve terminals, such as the prejunctional excitatory nicotinic receptors and the prejunctional inhibitory muscarinic receptors at the neuromuscular junction, which can increase or decrease acetylcholine release, respectively.

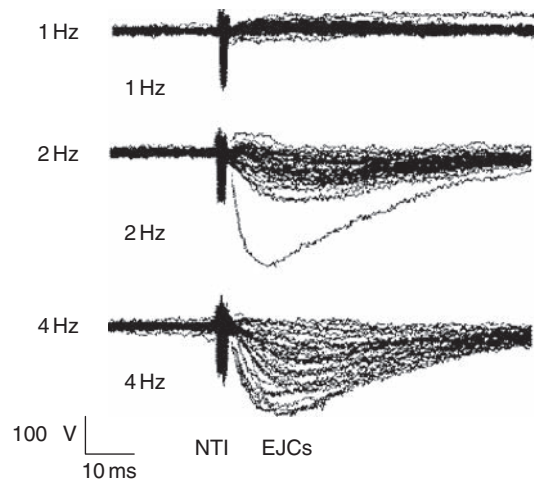

Figure 4 Frequency-dependent facilitation of neurotransmitter release in the guinea pig vas deferens. Simultaneous measurement of the nerve terminal impulse (NTI) and evoked excitatory junctional currents (EJCs) using focal extracellular recordings showed that although the impulses travel through the varicosities every time, EJCs are intermittent. Each panel shows 25 stimuli at 1, 2, and 4 Hz in the top, middle, and bottom panels, respectively. Reproduced from Jackson VM and Cunnane TC (2001) Neurotransmitter release mechanisms in sympathetic neurons: Past, present, and future perspectives. *Neurochemical Research* 26: 875–889, with permission from Springer Science and Business Media.

The importance of increases in intracellular Ca^{2+} concentrations as the trigger for exocytosis is well-known, and so perhaps intermittence was due to the failure of most varicosities to undergo appropriate changes in Ca^{2+} levels. Developments in confocal microscopy and Ca^{2+} indicators mean that we can now finally resolve Ca^{2+} dynamics in intact tissue preparations. By loading the nerve terminals through the cut end of the vas deferens, it has been shown that Ca^{2+} enters all varicosities when the action potential arrives.

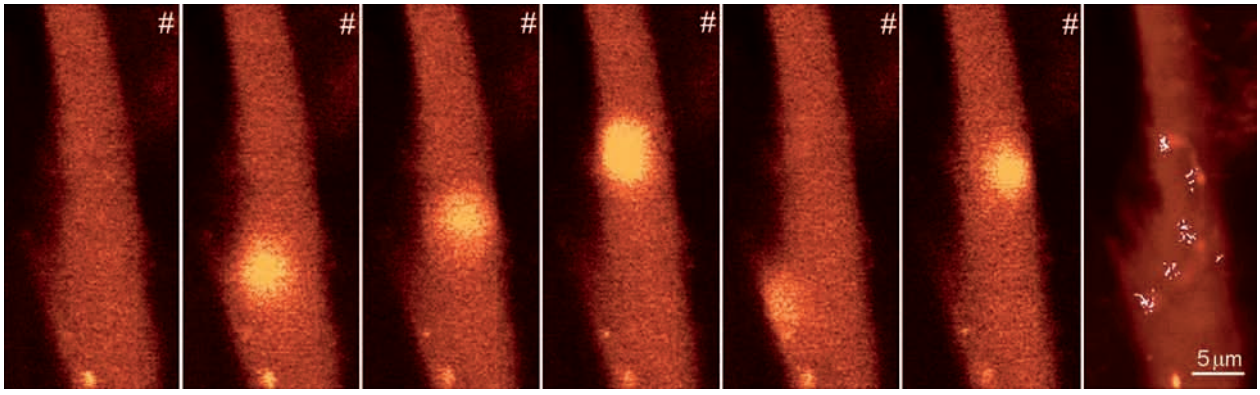

Figure 5 Discreet and intermittent smooth muscle Ca^{2+} transients following nerve stimulation. The first six frames show selected images of the same smooth muscle cell taken during 2 Hz stimulation. There is no response to most stimuli (frame 1), whereas some stimuli evoke focal Ca^{2+} transients in the smooth muscle cell (frames 2–6). The final frame was obtained from a confocal section 3 μm above that of the preceding images, and it shows an overlying nerve terminal varicosity. The white dots denote the location (epicenter) of a single smooth muscle Ca^{2+} transient occurring at some time during eight sets of recordings. Reproduced from Brain KL, Jackson VM, Trout SJ, and Cunnane TC (2002) Intermittent ATP release from nerve terminals elicits focal smooth muscle Ca^{2+} transients in mouse vas deferens. *Journal of Physiology* 541: 849–862.

Through Ca^{2+} studies on neurons, it was questioned whether it would be possible to resolve neurotransmitter release optically. Occasionally, for indeterminate reasons, some smooth muscle cells were loaded with the Ca^{2+} indicator, and distinct changes in Ca^{2+} were noted following nerve stimulation both in the varicosity and directly below in the smooth muscle cell. The focal changes in Ca^{2+} recorded in the smooth muscle cells were termed neuroeffector Ca^{2+} transients (NCTs) (Figure 5). By simultaneously monitoring both pre- and postjunctional responses evoked by low-frequency nerve stimulation with cell-permeant Ca^{2+} indicators, Brain, Jackson, and Cunnane found that these NCTs correlated well with action potential-evoked Ca^{2+} transients in varicosities and with the release probability of individual varicosities on the same axon. Surprisingly, they found that some varicosities had an apparent release probability as high as 0.09, but there was a population of varicosities that never released and remained 'silent.' Finally, there is an optical method that provides a unique approach to detect neurotransmitter release at the level of the individual sympathetic neuroeffector junction (i.e., single varicosities on the same nerve terminal branch) at an unparalleled resolution.

Currently, we still do not know the reason for intermittence. However, the ability to demonstrate neurotransmitter release at high resolution surely holds the key to unraveling the mechanisms underlying release modulation at individual varicosities on an impulse-to-impulse basis in an intact organ. Autoinhibition, variation in secretory proteins, and discrete changes in Ca^{2+} in different varicosities are among some of the current hypotheses undergoing evaluation, but we are now in a better position to study neurotransmitter release from these nerves. However, what is the neurotransmitter? Surely it is noradrenaline.

The Identity Revealed

Although there was broad agreement across all techniques that noradrenaline is released on an impulse-to-impulse basis from sympathetic nerves, there were a number of observations that did not fit the noradrenergic theory. For example, when the rodent vas deferens was depleted of noradrenaline using reserpine, which irreversibly interferes with the Mg^{2+}- and adenosine-5′-triphosphate (ATP)-dependent uptake of biogenic amines into vesicles and thereby depletes noradrenaline from nerves, electrophysiological methods showed that the EJPs were still present and often larger in amplitude. This was due to a prejunctional action rather than a postjunctional sensitization because the size of the SEJPs remained unchanged. Similarly, when yohimbine was applied to block prejunctional autoinhibitory α_2-adrenoceptors, EJPs were no longer potentiated. In other words, we have a tissue devoid of noradrenaline, but EJPs, the electrical sign of neurotransmitter release, are still present – a strange paradox.

Further investigation using selective antagonists led to the rather revolutionary concept of noradrenaline–ATP co-transmission. Imagine the controversy when Burnstock, Westfall, and others suggested that ATP, Krooll's 'universal currency of bioenergy,' could also be released from sympathetic nerves and that it was ATP, and not noradrenaline, that was responsible for the generation of the EJP. This was not so surprising in retrospect because Douglas and Poisner had shown that ATP was co-stored and co-released

with adrenaline and noradrenaline in the adrenal medulla approximately 20 years previously. Some people called purinergic transmission the 'pure imagination' hypothesis on account of its radical and unbelievable concepts, but it is now generally accepted that ATP is released from almost all nerves and acts as a separate neurotransmitter. Indeed, moving away from Dale's principle, which states that each neuron can only secrete one type of neurotransmitter, we now know that neurotransmitter release can differ at different synapses of the same cell, both in the class of neurotransmitter and also in the quantities of the cocktails of neurotransmitters released. Yet again, an idea born in the peripheral nervous system is now accepted widely throughout the central nervous system.

Work on purinergic transmission spawned the discovery of G-protein-coupled P2Y purinoceptors and of a unique class of ligand-gated ion channels, P2X receptors, which do not follow the structure of any other family of ligand-gated channels. After activation by extracellular ATP, through nonselective cation entry, P2X receptors are responsible for the EJP and SEJP because P2X receptor antagonists and desensitizers abolish both EJPs and SEJPs. NCTs are a measure of the Ca^{2+} flowing through the nonselective cation pore together with amplification from the release of Ca^{2+} from intracellular stores, and similarly, these drugs abolish NCTs in the vas deferens, blood vessels, and bladder. In the vas deferens, the purinoceptor subtype $P2X_1$ is particularly important because knockout mice that do not have $P2X_1$ receptor expression lack EJPs, SEJPs, NCTs, and spontaneous NCTs.

It would seem that the electrophysiological and confocal investigations have not been monitoring the effects of released noradrenaline but, rather, measuring how sympathetic nerves release ATP. However, if noradrenaline and ATP are co-stored in sympathetic nerve vesicles, as the biochemical analysis would suggest, then ATP release will also reflect how noradrenaline is being released assuming exocytosis of the entire neurotransmitter content of a vesicle. Indeed, research using amperometry, a direct measurement of the oxidation current of noradrenaline, has shown that noradrenaline can be released intermittently on an impulse-to-impulse basis (Figure 6). Finally, overflow studies, optical studies, and electrophysiological methods appear to be converging in order to explain sympathetic neurotransmitter release, be it ATP or noradrenaline.

The Elusive Nature of Noradrenaline – The Paradox

Noradrenaline produces its effects through G-protein-coupled receptors (Table 1) and its action is terminated by reuptake into the prejunctional neuronal (uptake 1; inhibited by cocaine, desipramine, amphetamines, and tricyclic antidepressants) and postjunctional non-neuronal (uptake 2; inhibited by corticosterone) cells. Reuptake results in either the metabolism of noradrenaline by monoamine oxidases and catechol-*O*-methyl transferases within a reaction pathway that finally produces vanillylmandelic acid or 3-methoxy,4-hydroxyphenylglycol, which are excreted in the urine, or the noradrenaline can be recycled and released again from the prejunctional nerve terminal. In the smooth muscle

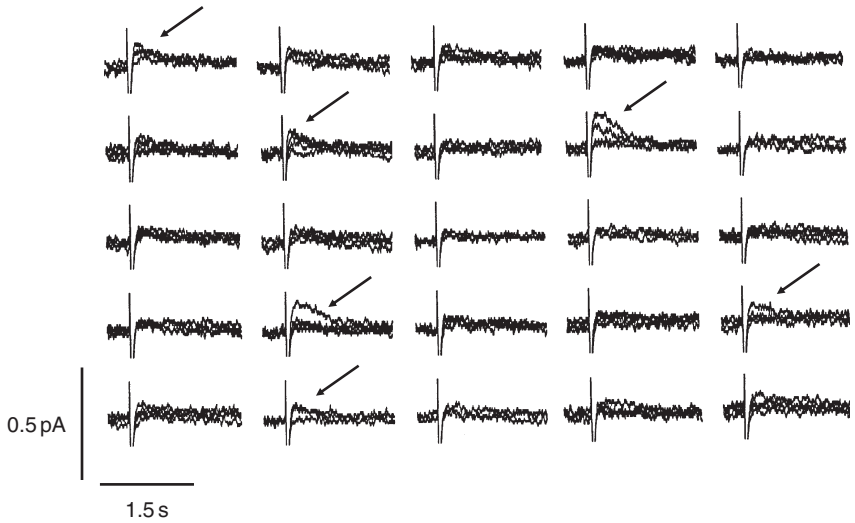

Figure 6 Intermittent release of noradrenaline from rat tail artery revealed by continuous amperometry. Five stimuli per panel at 0.1 Hz. Reproduced from Jackson VM and Cunnane TC (2001) Neurotransmitter release mechanisms in sympathetic neurons: Past, present, and future perspectives. *Neurochemical Research* 26: 875–889, with permission from Springer Science and Business Media; and Reproduced from Msghina M, Gonon F, and Stjärne L (1993) Intermittent release of noradrenaline by single pulses and release during short trains at high frequencies from sympathetic nerves in rat tail artery. *Neuroscience* 57: 887–890, with permission from Elsevier.

of the vas deferens, noradrenaline mediates most of its effects through α_1-adrenoceptors, which activate phospholipase C, and with increases in inositol-1,4,5-trisphosphate (IP_3) and diacylglycerol (DAG), respectively, lead to increases in intracellular Ca^{2+} and contraction and should be independent of an electrical change. Of course, there are exceptions to this rule, the most notable example being the rat anococcygeus, in which increases in intracellular Ca^{2+} result in an electrical depolarization by opening a Cl^- channel, but generally under low-frequency stimulations, most tissues show little or no electrical changes on α_1-adrenoceptor activation.

It would be expected that IP_3-generated Ca^{2+} waves would be detected using Ca^{2+}-imaging confocal microscopy. Indeed, slow propagating waves in small arteries following nerve stimulation can be attributed to α_1-adrenoceptor stimulation because they are blocked by prazosin, an α_1-adrenoceptor antagonist. Similarly, cholinergic stimulation in the bladder results in waves that are blocked by cyclopentolate, a muscarinic receptor antagonist, which are thought to be generated secondary to muscarinic receptor-induced IP_3 synthesis. Such waves rarely occur within the longitudinal layer of the vas deferens upon nerve stimulation, but they can be induced by applying exogenous noradrenaline. We therefore have a paradox: exogenously applied noradrenaline causes Ca^{2+} waves, whereas neuronally released noradrenaline does not readily produce waves, and yet both cause a contraction that is mediated by α_1-adrenoceptors. Once again, the remarkable sympathetic nervous system confronts us with a paradox.

The knockout mouse for $P2X_1$ purinoceptors provided the ideal opportunity to study noradrenergic transmission in isolation because $P2X_1$ receptor effects were lost, and one of the compensatory mechanisms for the purinergic 'shortfall' was the supersensitivity of the tissue to noradrenaline. Using higher frequency nerve stimulation, which is known to favor the noradrenergic component of release, even in the presence of both yohimbine, to antagonize autoinhibitory α_2-adrenoceptors to maximize noradrenaline release, and desipramine, to inhibit the reuptake of noradrenaline and maximize its concentration within the neuroeffector cleft, there were still few Ca^{2+} transients due to neuronal noradrenaline release. The preparations still responded well to exogenously added noradrenaline and appeared to have a larger contraction mediated by α_1-adrenoceptors to both neuronal and exogenous noradrenaline.

Perhaps neuronally released noradrenaline activates a different population of receptors compared to exogenously applied noradrenaline and these receptors have different pathways. Indeed, prazosin is a non-subtype selective α_1-adrenoceptor antagonist, with α_{1A}-adrenoceptors favoring the IP_3-mediated increase in intracellular Ca^{2+}, whereas α_{1B}- and α_{1D}-adrenoceptors may favor the DAG and protein kinase C pathways of contraction (which do not require such large increases in intracellular Ca^{2+}); this might explain the apparent differences due to prazosin's nonselectivity. Of course, there are crossovers between the phospholipase C (IP_3 and DAG) and D (DAG only) pathways because the receptors are promiscuous and activate many different intracellular signaling cascades, including the Rho-kinase pathway, which allows contraction without any concomitant increase in Ca^{2+} within smooth muscle cells. How these results can be explained and how the various pathways and cotransmitters interact remain under investigation. Noradrenaline and its effects remain elusive despite the current sophisticated investigative techniques.

Conclusion

We began with noradrenaline being the only neurotransmitter released from postganglionic sympathetic nerves, but investigations into the nature of neurotransmitter release have revealed major new insights into the mechanism by which sympathetic nerves function – intermittent neurotransmitter release, autoinhibition through prejunctional/presynaptic receptors, facilitation of release through prejunctional nicotinic receptors, cotransmission, and perhaps the most startling finding of all, that some varicosities do not release neurotransmitter at all. There is a rich diversity of pre- and postjunctional receptors and interactions that remain to be investigated, and we look forward to the advances that studying the sympathetic nervous system will bring. It is fair to say that in the twenty-first century we remain surprisingly ignorant of how nerves release neurotransmitters and the effects that noradrenaline and its cotransmitters have on the innervated organs.

See also: Adenosine Triphosphate (ATP); Adrenergic Receptors; Autonomic Nervous System; Brain Adrenergic Neurons; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress; Sympathetic Nervous System.

Further Reading

Blakeley AG and Cunnane TC (1979) The packeted release of transmitter from the sympathetic nerves of the guinea-pig vas deferens: An electrophysiological study. *Journal of Physiology* 296: 85–96.

Brain KL, Jackson VM, Trout SJ, and Cunnane TC (2002) Intermittent ATP release from nerve terminals elicits focal smooth

muscle Ca^{2+} transients in mouse vas deferens. *Journal of Physiology* 541: 849–862.

Brock JA and Cunnane TC (1987) Relationship between the nerve action potential and transmitter release from sympathetic postganglionic nerve terminals. *Nature* 326: 605–607.

Brock JA and Cunnane TC (1992) Electrophysiology of neuroeffector transmission in smooth muscle. In: Burnstock G and Hoyle CHV (eds.) *The Autonomic Nervous System: Autonomic Neuroeffector Mechanisms*, pp. 121–213. Chur, Switzerland: Harwood Academic.

Goodman LS, Gilman A, Hardman JG, and Limbird LE (2001) *Goodman & Gilman's the Pharmacological Basis of Therapeutics,* 10th edn., New York: McGraw-Hill.

Hague C, Chen Z, Uberti M, and Minneman KP (2003) α_1-Adrenergic receptor subtypes: Non-identical triplets with different dancing partners. *Life Science* 74: 411–418.

Jackson VM and Cunnane TC (2001) Neurotransmitter release mechanisms in sympathetic neurons: Past, present, and future perspectives. *Neurochemical Research* 26: 875–889.

Li F, De Godoy M, and Rattan S (2004) Role of adenylate and guanylate cyclases in β_1-, β_2-, and β_3-adrenoceptor-mediated relaxation of internal anal sphincter smooth muscle. *Journal of Pharmacology and Experimental Therapeutics* 308: 1111–1120.

Msghina M, Gonon F, and Stjärne L (1993) Intermittent release of noradrenaline by single pulses and release during short trains at high frequencies from sympathetic nerves in rat tail artery. *Neuroscience* 57: 887–890.

Rang HP, Dale MM, Ritter JM, and Moore PK (2003) *Pharmacology,* 5th edn., London: Churchill Livingstone.

Wenzel-Seifert K, Liu HY, and Seifert R (2002) Similarities and differences in the coupling of human β_1- and β_2-adrenoceptors to Gs_α splice variants. *Biochemical Pharmacology* 64: 9–20.

Relevant Website

http://www.iuphar.org – International Union of Pharmacology.

Norepinephrine: Adrenergic Receptors

D B Bylund, University of Nebraska Medical Center, Omaha, NE, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Norepinephrine (also called noradrenaline) is a neurotransmitter in both the peripheral and central nervous systems. Norepinephrine produces many effects in the body, the most notable being those associated with the 'fight-or-flight' response to perceived danger. The effects of norepinephrine and a related catecholamine, epinephrine (also called adrenaline), are mediated by the family of adrenergic receptors. The chemical structure of norepinephrine, as shown in Figure 1, indicates that it is a catecholamine, because it has both the catechol moiety consisting of a benzene ring (green) with two hydroxyl groups (red) and an amine (blue).

Adrenergic receptors (also called adrenoceptors) are widely distributed throughout the body. There are three major adrenergic receptor types: alpha-1, alpha-2, and beta. Each of these three receptor types is further divided into three subtypes. Adrenergic receptors are seven transmembrane receptors which consist of a single polypeptide chain with seven hydrophobic regions that form alpha-helical structures that span or transverse the membrane. Because the mechanism of action of adrenergic receptors includes the activation of guanine nucleotide regulatory binding proteins (G-proteins), they are also called G-protein-coupled receptors.

Stimulation of adrenergic receptors by endogenous catecholamines released in response to activation of the sympathetic autonomic nervous system (as well as by exogenous drugs called adrenergic agonists) results in a variety of effects such as increased heart rate, regulation of vascular tone, and bronchodilatation. In the central nervous system (CNS), adrenergic receptors are involved in many functions including memory, learning, alertness, and the response to stress.

History

In the late 1940s, Ulf von Euler in Sweden and Holtz in Germany identified norepinephrine as the neurotransmitter of the mammalian sympathetic nerves and soon thereafter also found it to be a normal constituent of mammalian brain. Norepinephrine was subsequently shown to also be a central neurotransmitter and was visualized in the brain by fluorescent and immunohistochemical techniques.

Adrenergic receptors were originally divided into two major types, alpha and beta, based on their pharmacological characteristics (i.e., rank order potency of agonists). Subsequently, the beta adrenergic receptors were subdivided into beta-1 and beta-2 subtypes, and more recently a beta-3 subtype was defined. The alpha adrenergic receptors were first subdivided into postsynaptic (alpha-1) and presynaptic (alpha-2) subtypes. After it was realized that not all alpha receptors with alpha-2 pharmacological characteristics were presynaptic, the pharmacological definition was used. The current classification scheme is based on three major types: alpha-1, alpha-2, and beta. Each of these three receptor types is further divided into three subtypes as shown in Figure 2: alpha-1A, alpha-1B, alpha-1D; alpha-2A, alpha-2B, alpha-2C; and beta-1, beta-2, beta-3.

Although adrenergic receptors are found throughout the body in tissues innervated by both the peripheral and CNSs, a few locations are of special interest. Alpha-1 adrenergic receptors are found peripherally in some blood vessels, whereas alpha-2 receptors are located on platelets and on nerve terminals in both the peripheral and CNSs. In the heart, beta-1 receptors predominate, where as in the smooth muscles of the lungs, some blood vessels, and the uterus, beta-2 receptors are relatively abundant.

Neurochemistry of Norepinephrine

Biosynthesis of Norepinephrine

Norepinephrine is synthesized in neurons starting with the amino acid tyrosine, which is obtained from the diet and can also be synthesized from phenylalanine. Tyrosine is converted to dihydroxyphenylalanine (DOPA) by the enzyme tyrosine hydroxylase, which is the rate-limiting enzyme for norepinephrine biosynthesis. DOPA, in turn, is converted to dopamine in the cytoplasm. Dopamine, which is also a neurotransmitter, is taken up into vesicles and converted to norepinephrine by the enzyme dopamine beta-hydroxylase. In the adrenal medulla and in a few brain regions, norepinephrine is converted to epinephrine by the enzyme phenylethanolamine *N*-methyltransferase.

Storage, Release, Reuptake, and Metabolism of Norepinephrine

Norepinephrine is stored in vesicles (also called storage granules) in the nerve terminals, which concentrates it and protects it from metabolism, until it is released following nerve stimulation. The major mechanism by which the effects of norepinephrine

are terminated is reuptake back into the nerve terminal by a high-affinity transporter. Norepinephrine can also be metabolized to inactive products. Inhibition of either of these processes results in an increase in the synaptic level of norepinephrine and a prolongation of its effects.

The reuptake of norepinephrine is mediated by the norepinephrine transporter (NET). Approximately 80% to 90% of the norepinephrine released into many synapses is cleared by this mechanism. NET belongs to the protein superfamily defined as Na^+/Cl^--dependent transporters and has 12 putative transmembrane domains. Some inhibitors of NET are antidepressants and are used in the treatment of clinical depression and other affective disorders. They are also sometimes used to treat anxiety disorders, obsessive–compulsive disorder, attention-deficit/hyperactivity disorder (ADHD), and chronic neuropathic pain. Several classes of antidepressant drugs inhibit NET, including the tricyclic antidepressants (e.g., imipramine, desipramine, doxepin), serotonin-norepinephrine reuptake inhibitors (SNRIs; e.g., venlafaxine, duloxetine), and the norepinephrine reuptake inhibitors (e.g., atomoxetine, reboxetine).

Norepinephrine is metabolized by the enzymes monoamine oxidase and catechol-O-methyltransferase to 3-methoxy-4-hydroxymandelic acid and 3-methoxy-4-hydroxyphenylglycol. These inactive degradation products can be quantitated in tissues, blood, and urine as a measure of norepinephrine turnover.

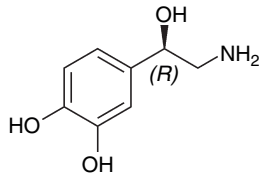

Figure 1 The structure of norepinephrine.

Ontogeny of Norepinephrine in the Brain

Cell bodies expressing tyrosine hydroxylase can be detected by 4 weeks of gestation in humans, and norepinephrine itself is detectable by 5–6 weeks. Norepinephrine levels in humans increase throughout the first trimester, especially from about 2 months of gestation forward. Following the initial increase in spinal norepinephrine, a decrease of 30–40% in concentration occurs between 6 months of gestation and early childhood. Noradrenergic neurons differentiate in rats between gestational day 10 and 13. From this point forward there is a steady differentiation and a nearly linear development of markers for noradrenergic neurons in the CNS, increasing approximately 100- to 1000-fold by adulthood. Thus, although the development of norepinephrine occurs fairly early and is reasonably rapid, it is slower than some of the other neurotransmitter systems, particularly the serotonin system.

Effects of Norepinephrine

Effects Mediated by the Autonomic Nervous System

The autonomic nervous system (also called the involuntary nervous system) is divided into two components, the sympathetic nervous system and the parasympathetic nervous system. The final nerves (postganglionic) in the sympathetic system are

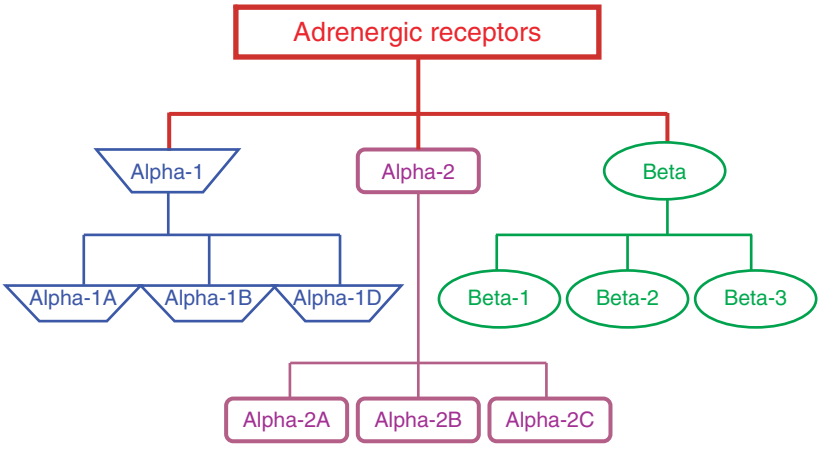

Figure 2 The classification of the adrenergic receptors.

adrenergic and thus release norepinephrine in the various tissues (end organs). The adrenal medulla, which is part of the sympathetic system, releases epinephrine into the circulation. The activation of the sympathetic system, in response to perceived danger, results in the release of large quantities of norepinephrine and epinephrine. Norepinephrine acting at alpha-1 receptors causes vasoconstriction (contraction) of cutaneous blood vessels, whereas epinephrine acting at beta-2 receptors in the blood vessels of the skeletal muscles causes vasodilatation (relaxation), resulting in blood flow being increased in the muscles. Norepinephrine and epinephrine acting at beta-1 receptors increase the force and rate of contraction of the heart, whereas epinephrine acting at beta-2receptors causes bronchodilatation in the lungs and relaxation of the smooth muscle in the uterus.

Effects Mediated by the CNS

In the CNS, the cell bodies of noradrenergic neurons are found primarily in the locus coeruleus in the brain stem. These neurons, however, project widely throughout the brain and spinal cord. The locus coeruleus, and hence norepinephrine, is an important regulator of a variety of physiologic functions, including sleep/wake cycles, attention, orientation, mood, memory, and cardiovascular as well as other autonomic and endocrine functions. Although adrenergic receptors are found throughout the brain, the alpha-2 receptors in the CNS are of particular importance because they regulate the release of norepinephrine, as well as many other neurotransmitters.

Alpha-1 Adrenergic Receptors

Three genetic and four pharmacological alpha-1 adrenergic receptor subtypes have been defined. The alpha-1A and alpha-1B subtypes were initially defined based on their differential affinity for adrenergic agents such as WB4101 and phentolamine and their differential sensitivities to the site-directed alkylating agent chloroethylclonidine. The alpha-1B subtype was subsequently cloned from the hamster, and the alpha-1A was cloned from bovine brain, although it was originally called the alpha-1c adrenergic receptor. A third subtype, the alpha-1D adrenergic receptor, was subsequently cloned from the rat cerebral cortex, although this clone was originally called the alpha-1a subtype by some investigators. A fourth pharmacological subtype, the alpha-1L, has been identified in vascular tissues from several species but may represent a conformational state of the alpha-1A receptor. The current classification scheme includes the alpha-1A, the alpha-1B, and the alpha-1D, but there is no alpha-1C (Figure 2).

Pharmacological and Molecular Characteristics of Alpha-1 Adrenergic Receptors

In addition to norepinephrine and epinephrine, alpha-1 receptors are activated by agonists as phenylephrine and methoxamine. These agonists are relatively selective for alpha-1 receptors and have lower affinity for alpha-2 and beta receptors. By contrast, they have similar affinities for the three alpha-1 subtypes and are thus non-subtype-selective agonists. Similarly, antagonists including prazosin and tamsulosin are relatively selective for alpha-1 receptors and block alpha-2 and beta receptors only at high concentrations. Several other antagonists such as phentolamine and phenoxybenzamine block both alpha-1 and alpha-2 adrenergic receptors with similar affinities. Alpha-1A-selective antagonists include 5-methylurapidil and niguldipine, whereas cirazoline appears to be a selective alpha-1A agonist.

The alpha-1 adrenergic receptors are single polypeptide chains of 446–572-amino-acid residues that span the membrane seven times, with the N-terminus being extracellular and the C-terminus intracellular. Thus, there are three intracellular loops and three extracellular loops. In contrast to the alpha-2 receptors, but similar to the beta receptors, the alpha-1 receptors have a long C-terminal tail (137–179-amino-acid residues) and a short third intracellular loop (68–73-amino-acid residues) The C-terminal tails also have multiple sites of phosphorylation which are thought to be important in the desensitization, recycling, and downregulation of the receptor.

The human alpha-1 adrenergic receptor genes consist of two exons and a single large intron of at least 20 kbp in the region corresponding to the sixth transmembrane domain. No splice variants are known for the alpha-1B and alpha-1D subtypes. By contrast, at least ten splice variants of human alpha-1A subtype have been reported, but only four produce full-length receptors.

Alpha-2 Adrenergic Receptors

Three genetic and four pharmacological alpha-2 adrenergic receptor subtypes have also been defined (Figure 2). The alpha-2A and alpha-2B subtypes were initially defined based on their differential affinity for adrenergic agents such as prazosin and oxymetazoline. These subtypes were subsequently cloned from human, rat, and mouse. A third subtype, alpha-2C, was originally identified in an opossum kidney cell line and

has also been cloned from several species. A fourth pharmacological subtype, the alpha-2D, has been identified in the rat, mouse, and cow. This pharmacological subtype is a species ortholog of the human alpha-2A subtype and thus is not considered to be a separate genetic subtype.

Pharmacological and Molecular Characteristics of Alpha-2 Adrenergic Receptors

In addition to norepinephrine and epinephrine, alpha-2 receptors are activated by clonidine and brimonidine. These agonists are relatively selective for alpha-2 receptors and have lower affinity at alpha-1 and beta receptors. Similarly, the antagonist yohimibine is relatively selective for alpha-2 receptors and blocks alpha-1 and beta receptors only at higher concentrations. Antagonists that are at least somewhat selective for one of the alpha-2 subtypes include BRL44408 for the alpha-2A, prazosin and ARC-239 for the alpha-2B (note, however, that these two agents have much higher affinities for alpha-1 receptors), and rauwolscine for the alpha-2C subtype. Ozymetazoline is a partial agonist that has a higher affinity for the alpha-2A subtype as compared to the alpha-2B and alpha-2C subtypes.

The alpha-2 adrenergic receptors are single polypeptide chains of 450–462-amino-acid residues. In contrast to the alpha-1 and beta receptors, the alpha-2 receptors tend to have long third intracellular loops (148–179-amino-acid residues) and a short C-terminal tail (20–21-amino-acid residues). The third intracellular loops have multiple sites of phosphorylation which are thought to be important in the desensitization, recycling, and downregulation of the receptor. The alpha-2 adrenergic receptor genes do not contain introns, and thus there are no splice variants.

Beta Adrenergic Receptors

Three beta adrenergic receptor subtypes have been identified. The beta-1 adrenergic receptor, the dominant receptor in heart and adipose tissue, is equally sensitive to epinephrine and norepinephrine, whereas the beta-2 adrenergic receptor, responsible for relaxation of vascular, uterine, and airway smooth muscle, is less sensitive to norepinephrine than to epinephrine. The beta-3 receptor is insensitive to the commonly used beta adrenergic receptor antagonists and was previously referred to as the 'atypical' beta adrenergic receptor. A beta-4 receptor has been postulated; however, definitive evidence of its existence is lacking, and it is now thought to be a 'state' of the beta-1 adrenergic receptor.

Pharmacological and Molecular Characteristics of Beta Adrenergic Receptors

Isoproterenol is the prototypic non-subtype-selective beta agonist which has little effect at alpha-1 and alpha-2 receptors. Epinephrine is 10- to 100-fold more potent at the beta-2 receptor as compared to the beta-1 subtype, whereas norepinephrine is more potent than epinephrine at the beta-3 subtype. Many beta-2-selective agonists, such as terbutaline and salmeterol, have been developed for the treatment of asthma. Due to their subtype selectivity they have a lower incidence of side effects mediated by the beta-1 receptor. Propranolol is the prototypic non-subtype-selective beta antagonist which has equal affinities at the beta-1 and beta-2 subtypes. Other nonselective beta adrenergic antagonists include timolol, pindolol (which is actually a weak partial agonist), and carvedilol, which is also an alpha-1 antagonist. Several beta-1-selective antagonists have been developed, such as metoprolol and esmolol.

The beta adrenergic receptors are single polypeptide chains of 408–477-amino-acid residues. In contrast to the alpha-2 receptors, but similar to the alpha-1 receptors, the beta receptors tend to have longer C-terminal tails (61–97-amino-acid residues) and shorter third intracellular loops (54–80-amino-acid residues). The C-terminal tails have multiple sites of phosphorylation which are thought to be important in the desensitization, recycling, and downregulation of the receptor.

The beta-1 and beta-2 adrenergic receptor genes do not contain introns, and thus they have no splice variants. By contrast, the beta-3 receptor has one intron, resulting in two splice variants. However, no functional differences have been found between the two splice variants.

Regulation of Adrenergic Receptors

The processes involved in desensitization and downregulation have been extensively investigated for the beta-2 adrenergic receptor. The other adrenergic receptors, as well as many other G-protein-coupled receptors, appear to behave in a similar, although not identical, manner. Initial uncoupling of the beta-2 receptor from the G-protein after agonist binding is mediated by phosphorylation of specific residues in the carboxyl tail of the receptor. The phosphorylated beta-2 receptor serves as a substrate for the binding of β-arrestin, which not only uncouples the receptor from the signal transduction process but also serves as an adapter protein that mediates the binding of additional signaling proteins and entry into the internalization pathway. The mechanisms of beta-2 adrenergic receptor downregulation appear to involve both

an increase in the rate of degradation of the receptor and a decrease in the levels of beta receptor mRNA.

Adrenergic Receptor Signal Transduction Pathways

The alpha-1 adrenergic receptors activate the $G_{q/11}$ family of G-proteins, leading to the dissociation of the α and $\beta\gamma$ subunits and the subsequent stimulation of the enzyme phospholipase C. This enzyme hydrolyzes phosphatidylinositol in the membrane, producing inositol trisphosphate (IP_3) and diacylglycerol. These molecules act as second messengers mediating intracellular Ca^{2+} release via the IP_3 receptor and activating protein kinase C. Other signaling pathways that have also been shown to be activated by alpha-1 receptors include Ca^{2+} influx via voltage-dependent and independent calcium channels, arachidonic acid release, and activation of phospholipase A_2, phospholipase D activation, and mitogen-activated protein kinase.

The alpha-2 adrenergic receptors activate the $G_{i/o}$ family of G-proteins and alter (classically inhibit) the activity of the enzyme adenylate cyclase, which in turn, decreases the concentration of the second messenger cyclic AMP. In addition, the stimulation of alpha-2-receptors can regulate several other effector systems including the activation of K^+ channels, inhibition or activation of Ca^{2+} channels, and activation of phospholipase A_2, phospholipase C, and Na^+/H^+ exchange.

The beta adrenergic receptors activate the G_s family of G-proteins and activate adenylate cyclase, thus increasing in cyclic adenosine monophosphate (AMP) concentrations. Beta adrenergic receptors interact with many other signaling proteins, including the phosphoprotein EBP50 (ezrinradixin-moesin-binding phosphoprotein-50), the Na^+/H^+ exchanger regulatory factor, and cyclic nucleotide ras guanine-nucleotide exchange factor (CNrasGEF).

Traditionally it has been assumed that intrinsic efficacy of a ligand for the adrenergic receptors is a property of the ligand. However, recent evidence has shown that certain functionally selective ligands may act as either agonists or antagonists for different functions mediated by the same receptor. This functional selectivity likely involves differences in ligand-induced intermediate conformational states, the diversity of G-proteins, scaffolding and signaling partners, and receptor oligomers.

Adrenergic Receptor Polymorphisms

Polymorphisms have been identified in some of the alpha-2 and beta adrenergic receptor subtypes, which may have important clinical implications. A common polymorphism has been identified in the third intracellular loop of the alpha-2B receptor, which consists of a deletion of three glutamate residues (301–303) and is a risk factor for acute coronary events, but not hypertension. This deletion results in a loss of short-term agonist-induced desensitization. The most common polymorphism associated with the human alpha-2C receptor is the alpha-2C-del (residues 322 to 325 are missing), which occurs in 40% of Blacks and 4% of those from other ethnic backgrounds. As the alpha-2C-del receptor appears to be less efficiently coupled to G-proteins, it is less able to mediate a response to agonists. The consequence of this in humans is an impaired feedback inhibition of catecholamine release that is associated with elevated blood pressure and an increased risk of heart failure.

The gene encoding the human beta-1 adrenergic receptor is quite polymorphic, with 18 single nucleotide polymorphisms (SNPs), seven of which cause amino acid substitutions. A total of 13 polymorphisms in the beta-2 adrenergic receptor gene and its transcriptional regulator upstream peptide have been identified. Three closely linked polymorphisms, two coding regions at amino acid positions 16 and 27 and one in the upstream peptide, are common in the general Caucasian population. The glycine-16 receptor exhibits enhanced downregulation *in vitro* after agonist exposure. In contrast, arginine-16 receptors are more resistant to downregulation. Some studies have suggested a relationship among these polymorphisms, airway responsiveness (e.g., asthma), and the responsiveness to beta adrenergic agonists.

A tryptophan-64 to arginine polymorphism has been identified in the beta-3 adrenergic receptor. The allele frequency is approximately 30% in the Japanese population, higher in Pima Indians, and lower in Caucasians. Type 2 diabetic patients with this mutation showed a significantly younger age of onset of diabetes and an increased tendency to obesity, hyperinsulinemia, and hypertension.

Drugs Which Mimic or Block the Effects of Norepinephrine

Norepinephrine itself is rarely used as a drug due to its rapid metabolism and many sites of action. Drugs which evoke responses similar to sympathetic nerve stimulation are called sympathomimetic drugs. They produce their effects either directly by stimulating adrenergic receptors (adrenergic receptor agonists) or indirectly by promoting the release of norepinephrine or by blocking its reuptake. Table 1 gives examples of various sympathomimetic drugs with their mechanism of action, the effects they produce, and their therapeutic indications.

Table 1 Drugs which mimic the effects of norepinephrine

Receptor	*Examples of drugs*	*Effect*	*Therapeutic indication*
Alpha-1	Phenylephrine	Vasoconstriction	Hypotension, nasal congestion
Alpha-2	Clonidine, brimonidine	Lowers fluid pressure	Hypertension, glaucoma
Beta-1	Dobutamine	Increases heart rate	Cardiovascular shock
Beta-2	Albuterol, terbutaline, ritodrine	Relaxes smooth muscle in lung and uterus	Asthma, premature labor
(Indirect acting)	Amphetamine	CNS stimulant	Narcolepsy, hyperactivity
(Indirect acting)	Desipramine, tomoxetine	Blocks norepinephrine reuptake	Depression
(Indirect acting)	Phenelzine	Inhibits monoamine oxidase inhibitor (MAO)	Depression

Table 2 Drugs which block the effects of norepinephrine

Receptor	*Examples of drugs*	*Effect*	*Therapeutic indication*
Alpha-1	Prazosin, terazosin	Vasodilatation	Hypertension, benign prostatic hypertrophy
Alpha-2	Mirtazapine	Antidepressant	Depression
Beta	Propranolol, timolol	Decreases heart rate	Hypertension, angina, glaucoma
Beta-1	Atenolol, metoprolol	Decreases heart rate	Hypertension

Drugs which block the responses to sympathetic nerve stimulation and thus block the effects of norepinephrine are called adrenergic receptor antagonists. They have no direct effects of their own, but act by blocking the effects of either released norepinephrine or an administered adrenergic receptor agonist. Table 2 gives examples of adrenergic receptor antagonists (also called adrenergic blockers) with their receptor selectivity, the effects they produce, and their therapeutic indication.

See also: Brain Adrenergic Neurons; Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Monoamines: Release Studies; Norepinephrine: CNS Pathways and Neurophysiology; Sleep–Wake State Regulation by Noradrenaline and Serotonin; Sympathomimetic Drugs and Adrenergic Receptor Antagonists.

Further Reading

Bylund DB (1988) Subtypes of alpha-2 adrenoceptors: Pharmacological and molecular biological evidence converge. *Trends in Pharmacological Science* 9: 356–361.

Bylund DB, Eikenberg DC, Hieble JP, et al. (1994) International Union of Pharmacology nomenclature of adrenoceptors. *Pharmacology Review* 46: 121–136.

Cooper JR, Bloom FE, and Roth RH (2003) Norepinephrine and epinephrine. In: *The Biochemical Basis of Neuropharmacology*, 8th edn., pp. 181–223. New York: Oxford University Press.

Eisenhofer G, Kopin IJ, and Goldstein DS (2004) Catecholamine metabolism: A contemporary view with implications for physiology and medicine. *Pharmacology Reviews* 5: 331–349.

Ferguson SSG (2001) Evolving concepts in G protein-coupled receptor endocytosis: The role in receptor desensitization and signaling. *Pharmacology Reviews* 53: 1–24.

Frazer A (2000) Norepinephrine involvement in antidepressant action. *Journal of Clinical Psychiatry* 61(supplement 10): 25–30.

Gordon M Autonomic (ANS) pharmacology: Introduction. In: *Medical Pharmacology and Disease-Based Integrated, Instruction*, Ch. 4. http://www.pharmacology2000.com/Autonomics/Introduction/Introobj1.htm(accessed April 2007).

Hieble JP, Bylund DB, Clarke DE, et al. (1995) International Union of Pharmacology. X: Recommendation for nomenclature of alpha-1 adrenoceptors: Consensus update. *Pharmacology Reviews* 47: 267–270.

Hokfelt T, Johansson O, and Goldstein M (1984) Central catecholamine neurons as revealed by immunohistochemistry with special reference to adrenaline neurons. In: Björklund A and Hokfelt T (eds.) *Handbook of Chemical Neuroanatomy, Vol. 2: Part 1: Classical Transmitters in the CNS*. Amsterdam: Elsevier.

Murrin LC, Sanders JD, and Bylund DB (2007) Comparison of the maturation of the adrenergic and serotonergic neurotransmitter systems in the brain: Implications for differential drug effects on juveniles and adults. *Biochemical Pharmacology* 73: 1225–1236.

Neumeister A, Charney DS, Belfer I, et al. (2005) Sympathoneural and adrenomedullary functional effects of alpha-2C adrenoreceptor gene polymorphism in healthy humans. *Pharmacogenetics and Genomics* 15: 143–149.

Rockman HA, Koch WJ, and Lefkowitz RJ (2002) Seven-transmembrane-spanning receptors and heart function. *Nature* 415: 206–212.

Small KM, McGraw DW, and Liggett SB (2003) Pharmacology and physiology of human adrenergic receptor polymorphisms. *Annual Review of Pharmacology and Toxicology* 43: 381–411.

Urban JD, Clarke WP, von Zastrow M, et al. (2007) Functional selectivity and classical concepts of quantitative pharmacology. *Journal of Pharmacology and Experimental Therapeutics* 320: 1–13.

Relevant Websites

http://www.adrenoceptor.com – Adrenoceptor Online.

http://pubchem.ncbi.nlm.nih.gov – PubChem Compound Summary: PubChem Compound 5814.

Norepinephrine: CNS Pathways and Neurophysiology

G Aston-Jones, Medical University of South Carolina, Charleston, SC, USA
C A Meijas-Aponte, National Institutes of Health, Baltimore, MD, USA
B Waterhouse, Drexel University College of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Norepinephrine Cell Groups and Pathways

Locus Coeruleus

The nucleus locus coeruleus (LC) is a dense cluster of noradrenergic neurons in the dorsorostral pons. This nucleus gained prominence in the 1960s when new anatomical approaches revealed it to be the major source of norepinephrine (NE) in brain with projections throughout most central nervous system (CNS) regions. These findings stimulated a great deal of research into this system, resulting in a wealth of knowledge at the cellular, systems, and behavioral levels. Considerable evidence indicates that these neurons are important in a variety of cognitive, affective, arousal, and other behavioral functions, as well as associated clinical dysfunctions.

Locus coeruleus efferent projections As summarized in Table 1, LC NE neurons project throughout the neuraxis, including the cerebral cortex, thalamus, cerebellum, midbrain, and spinal cord. In this regard, it is different from other 'typical' brain systems that have very limited projection targets (e.g., thalamocortical projections). Despite this high degree of divergence, LC neurons exhibit substantial regional and laminar specificity in their efferent projections. Notably, brain areas in primates that are associated with attentional processing (e.g., parietal cortex, pulvinar nucleus, and superior colliculus) have particularly dense LC projections and NE receptors. Notable areas that receive little or no input from the LC include the caudate putamen and most of the hypothalamus. Ultrastructural evidence indicates that these neurons often make conventional synaptic appositions with postsynaptic targets, although the frequency of such synaptic contacts is controversial. Extrasynaptic (paracrine) release of NE from LC terminals also appears to be possible.

Locus coeruleus afferents A variety of evidence indicates that the LC receives a wealth of different neurotransmitter inputs. This nucleus is densely innervated by fibers that contain opiates, glutamate, γ-aminobutyric acid (GABA), serotonin, epinephrine, histamine, and the peptide orexin/hypocretin. The sources of these various inputs have not been fully elucidated, though some major inputs have been identified. The nucleus paragigantocellularis lateralis (PGi) in the ventrolateral rostral medulla, a major input that strongly excites LC neurons, is a source for glutamate, GABA, enkephalin, corticotropin-releasing hormone (CRH), and epinephrine. A strongly inhibitory GABA and enkephalin input originates from the dorsomedial rostral medulla. Excitatory orexin and histamine inputs originate in the hypothalamus.

The shell of dendrites surrounding the LC nucleus offers additional extensive targets for afferent termination, and indeed it appears that several areas target these extranuclear dendrites more so than the LC nucleus proper. Thus, projections from the periaqueductal gray matter, parabrachial region, preoptic area, amygdala, and medial prefrontal cortex, among other sites, project primarily to the peri-LC region, with relatively little input to the LC proper. Although some of these projections have been shown to contact LC dendrites, additional ultrastructural studies are needed to examine others.

A recent study has extended our understanding of LC afferents to the circuit level. Using transynaptic tracing, lesions, and electrophysiology, Aston-Jones and colleagues determined that the suprachiasmatic nucleus (SCN) indirectly innervates the LC, using the dorsomedial hypothalamus as a relay. Based upon the role of the SCN as the brain's major circadian pacemaker and the role of the LC in arousal and performance (see later), these findings reveal a possible circuit for circadian regulation of arousal and performance.

Recent work by Iba, Aston-Jones, and colleagues reveal that in monkey strong projections to LC originate in cingulate and orbital cortices (Figure 1). Given the roles of these prefrontal areas in conflict monitoring, decision making, and other cognitive functions, these projections may be responsible at least in part for the different modes of activity that occur in LC neurons during cognitive performance (described later). Interestingly, these prefrontal projections to LC appear to be stronger in primates than in rats, in concert with the more highly evolved prefrontal cortex and more prominent role of this structure in regulating behavior in primates.

Taken together, these observations indicate that LC is not a relay nucleus for primary sensory or motor information and that it receives highly processed information concerning internal and external sensory stimuli as well as behavioral and affective states. This suggests a highly integrative function for this system.

Table 1 Organization of the adrenergic system

	Fiber density	*Locus coeruleus A4 + A6*	*Ventral pons and medulla cluster*				*Dorsal medulla cluster*		
			A1	*C1*	*A5*	*A7*	*A2*	*C2*	*C3*
Forebrain									
Cortex	2+	✔							
Hippocampal formation	2–3+	✔							
Amygdala									
Medial	1+		✔				✔		
Central	3+	✔	✔	✔	✔		✔	✔	✔
Basal	3+	✔							
Basal forebrain/septum									
Medial septum	2+	✔	✔	✔			✔	✔	
Medial preoptic area		✔	✔	✔			✔	✔	
Substatia innominata		✔	✔	✔			✔	✔	
Lateral septum	3+	✔	✔	✔			✔		
Bed nucleus stria terminalis	5+	✔	✔	✔	✔		✔	✔	
Ventral pallidum			✔		✔				
Zona incerta			✔	✔	✔				
Subthalamic nucleus		✔			✔				
Nucleus accumbens	1–3+						✔		
Thalamus									
Sensory									
Lateral geniculate nu	3–4+	✔							
Medial geniculate nu	1–2+	✔							
Ventroposterior medial nu		✔							
Ventroposterior lateral mu		✔							
Ventrobasal nu	3–4+	✔							
Ventromedial nu	3+	✔							
Reticular nu	3–4+	✔							
Motor									
Ventrolateral nu	3+	✔							
Ventralanterior nu	3+	✔							
Viscero-limbic									
Anteriormedial nu	4+	✔							
Anteriodorsal nu	3+	✔							
Anteriovental	2+	✔							
Mediodorsal nu	1–2+	✔							
Central medial nu	1+	✔							
Central lateral nu	1–2+	✔							
Reuniens nu	1+	✔							
Parafascicular nu	1+2	✔							
Paraventrivular nu	5+			✔	✔				
Rhomboid nu	2–3+	*							
Epithalamus									
Medial habenulla	5+	*							
Lateral habenulla	4–5+	*							
Hypothalamus									
Periventricular nucleus	5+	✔							
Suprachiamatic nucleus	1–2+	*							
Retrochiasmatic	4+		✔						
Paraventricular nuclues	3–5+	✔	✔	✔	✔		✔		
Supraoptic nuclues	4+		✔				✔		
Dorsomedial	3–5+	✔	✔	✔					
Arcute nuclues	3+	✔	✔	✔			✔		
Medial eminence	2+	✔		✔				✔	
Perifornical area	4+		✔				✔		
Lateral area	3+		✔	✔					
Brain stem									
Sensory									
Main sensory trigeminal nu	2+	✔							

Continued

Table 1 Continued

	Fiber density	*Locus coeruleus A4 + A6*	*Ventral pons and medulla cluster*				*Dorsal medulla cluster*		
			A1	*C1*	*A5*	*A7*	*A2*	*C2*	*C3*
Spinal trigeminal nu	2+	✔							
Mesencephalic trigeminal nu	4+	✔		✔					
Superior olivary complex	1+	✔							
Cochlear nu	2+	✔	✔			✔			
Vestibular nu	0–1+	✔		✔					
Motor									
Oculomotor and trochlear nu	0–1+								
Motor trigeminal nu	4+				✔	✔			
Abducens nu	0–1+								
Facial nu	3–4+				✔	✔			
Hypoglossal nu	2–3+				✔	✔			
Nucleus ambiguus	5+	✔		✔					
Inferior olive	5+	✔	✔		✔				
Red nucleus	1+	✔							
Cerebellum		✔			✔	✔			
Sensorimotor									
Superior colliculus	2+				✔				
Superficial layers		✔							
Deep layers					✔	✔			
Inferior colliculus	2+	✔							
Visceral									
Nucleus of the solitary tract including A2 and C2 areas	5+		✔	✔	✔				
Dorsal motor nucleus of vagus	5+			✔	✔				
Caudal ventrolateral medulla including A1 area	3+			✔	✔				
Rostral ventrolateral medulla including C1 area	3+		✔		✔				
Periaqueductal gray	2+5		✔	✔	✔				
Parabrachial nucleus	4+		✔	✔	✔		✔		
Area postrema	1+		✔	✔			✔		
Modulatory systems									
Substantia nigra, pars compacta	0–1+								
Ventral tegmental area	1+	✔	✔	✔					
Retrorubral field	1+	✔	✔	✔					
Raphe			✔						
Central linear	2+								
Dorsal	2+		✔						
Magnun	3+		✔	✔					
Pallidus	5+		✔	✔					
Obscurus	1+	✔		✔					
Locus coeruleus	2+			✔					✔
A5 area			✔						
Spinal cord									
Intermediolateral cell column				✔	✔	✔			
Dorsal horn		✔							
Ventral horn		**			✔	✔			

This table is based on the organization of LC versus non-LC systems as originally reviewed by Moore and Card in 1984 and published in the *Handbook of Chemical Neuroanatomy*, volume 2. The descriptions of fiber density were obtained from Moore and Card, whereas the information regarding efferent projections was updated based upon a comprehensive survey of research reports available in the literature. The information presented is meant to reflect major efferents of the adrenergic cell groups. Single asterisks indicate structures known to not receive LC afferents but for which the source of adrenergic innervation is unknown. The double asterisk indicates ventral horn innervation by LC, which is dependent upon rat strain.

Norepinephrine release Microdialysis methods have recently provided *in vivo* confirmation that certain drugs and behavioral conditions increase or decrease the release of norepinephrine from locus coeruleus axons. For example, reuptake blockers and stress both enhance extracellular NE levels in areas (such as the neocortex) that receive their only NE innervation from the LC. In addition, the

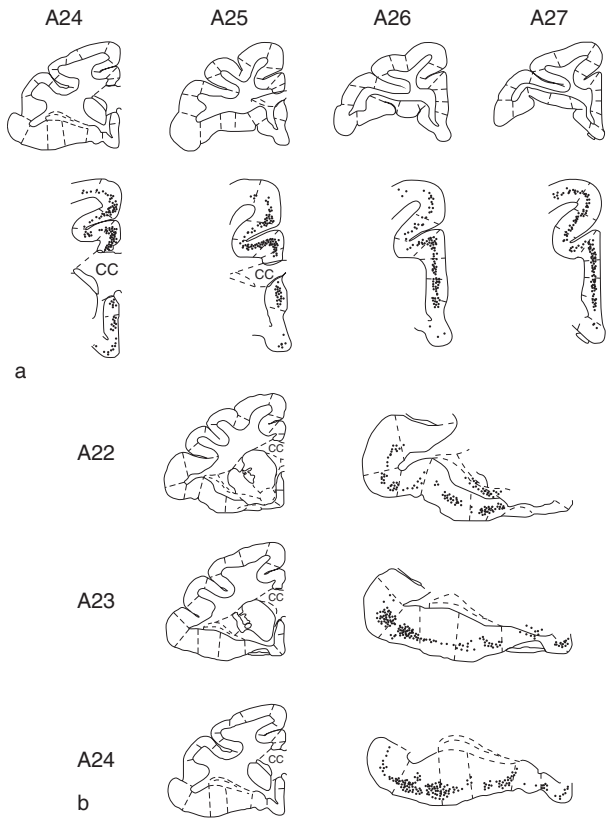

Figure 1 Plots of retrogradely labeled neurons in the nucleus accumbens and orbitofrontal cortex after injections of cholera toxin B into monkey locus coeruleus. (a) Accumbens neurons labeled from the monkey locus ceruleus. Lower sections are high-power views containing plotted cells; upper sections are low-power views to give orientation. (b) Orbitofrontal cortex neurons labeled from the monkey locus coeruleus. Sections on the right are high-power views containing plotted cells; sections on the left are low-power views to give orientation. For both panels, labels A22–A27 refer to distances (in millimeters) from the interaural line. CC, corpus callosum. From Aston-Jones G and Cohen JD (2005) Adaptive gain and the role of the locus coeruleus–norepinephrine system in optimal performance. *Journal of Comparative Neurology* 493: 99–110.

amount of NE release in cortex increases in proportion to activity within LC, even at low frequencies (2–4 Hz). Notably, burst stimulation elicits more NE release per impulse than evenly do spaced stimuli. This indicates that phasic LC responses (see later) may be particularly effective in releasing NE in target areas.

Locus coeruleus impulse activity *Sleep and waking*. Spontaneous LC activity in rats, cats, or monkeys varies with the stage of the sleep–wake cycle, with firing being most rapid during wakefulness, slower during slow-wave sleep, and virtually absent during paradoxical sleep. LC neurons are phasically activated by conspicuous unconditioned stimuli. Notably, stimuli that elicit large LC responses in either rats or monkeys also typically disrupt sleep or ongoing behavior, and evoke waking or a behavioral-orienting response. The same stimuli do not disrupt behavior if they elicit small LC responses.

Stress. Other studies have revealed strong activation of LC neurons by stressors. Stimuli such as sciatic nerve activation or other painful events strongly activate LC cells. Other stressors, such as a puff of air in the awake monkey or a variety of environmental or physiological stressors, also activate LC neurons. In addition, LC neurons are activated by corticotropin-releasing hormone, which mediates the response of LC cells to certain physiological stressors such as hypotension.

Locus coeruleus activity during cognitive performance. The preceding results suggest that, in addition to a role in sleep and waking, LC neurons may be involved in behavioral responses to sensory stimuli. Aston-Jones and colleagues tested this by recording impulse activity of LC neurons in monkeys performing a target detection task. These studies indicated that the LC exhibits two modes of impulse activity: a tonic mode in which tonic activity is high but phasic responses to targets are low or absent, and a phasic mode in which LC neurons fire tonically more slowly but exhibit consistent activation shortly following target stimuli. Behaviorally, the tonic mode corresponds to poor task performance with many false alarm errors, whereas the phasic mode is associated with nearly error-free performance. These results indicate that these different modes of LC activity may facilitate either selective, focused attention (phasic mode) or flexible behavior and scanning (tonic mode). Each of these LC and behavioral modes is adaptive, depending on the context.

LC activation at decision completion. Recent studies recording LC neural activity in monkeys performing a forced-choice task revealed that the aforementioned phasic responses described are neither purely sensory nor motor/premotor in nature. As shown in **Figure 2**, these LC responses are more closely linked to behavioral responses (lever releases) than to sensory stimuli that trigger the behavioral responses, but do not occur for lever releases elicited outside of the task context. Finally, the onset of these LC responses occurs at about 200 ms prior to the lever releases, at a time that the decision concerning lever release is being made. Aston-Jones and colleagues concluded from these results that these phasic LC responses reflected completion of a decision process, and that the modulatory gain-enhancing effects of the associated NE release (see later) facilitate brain processes and behaviors called upon by the decision just

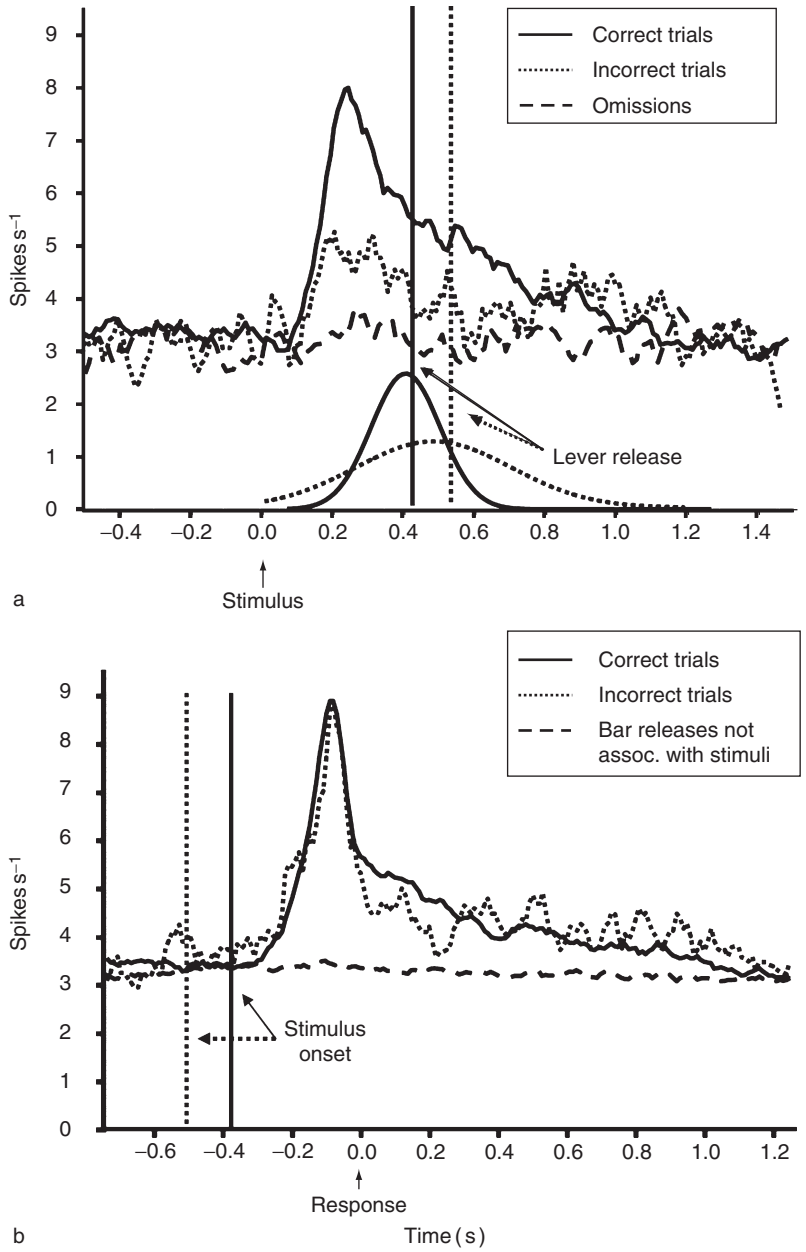

Figure 2 Stimulus- and response-locked population perievent time histograms (PETHs) showing LC responses to cues for trials yielding correct and incorrect behavioral responses. (a) Stimulus-locked population PETHs showing LC response to cues (presented at time 0) for trials yielding correct or incorrect behavioral responses. Note that the LC response peaks sooner and is less prolonged on correct, compared to incorrect, trials in this analysis (17 533 and 1362 trials, respectively). No fluctuation in LC activity was detected on omission trials (dashed line, 1128 trials). Vertical lines indicate the mean behavioral RTs. Curves below PETHs represent the normalized RT distributions for correct and incorrect trials. (b) The difference in the phasic LC response between correct and incorrect trials is no longer evident on response-locked population PETHs. In addition, no LC response occurred prior to or following lever releases not associated with stimulus presentation (dashed line, 3381 trials). Vertical lines indicate the mean stimulus onset times. From Clayton EC, Rajkowski J, Cohen JD, et al. (2004) Phasic activation of monkey locus coeruleus neurons by simple decisions in a forced choice task. *Journal of Neuroscience* 24: 9914–9920.

made. In this way, the LC phasic response supports the current task/goal and facilitates focused performance in that task. The projections to LC from orbital and cingulate cortices (described previously) may mediate these phasic LC responses associated with decision processes.

Non-LC Norepinephrine and Epinephrine Cell Groups

Neurons in other brain nuclei aside from those in the LC constitute 55% of all CNS NE and epinephrine (E) cells. These neurons are clustered in two

rostrocaudally aligned columns within the brain stem. One column is located primarily in the ventral aspect of the pons and medulla and contains 44% of all NE/E neurons, whereas the second is situated in the dorsomedial medulla. The individual clusters of non-LC NE/E neurons within these cell columns are further subdivided on the basis of their specific anatomical location and the ability to synthesize E (Figure 3).

Non-LC NE/E projections An analysis of projection patterns of LC versus non-LC NE/E neurons reveals similarities and differences (Table 1). Non-LC NE neurons project exclusively subcortically and make a preponderance of connections onto motor nuclei in the medulla and spinal cord, and onto structures associated with visceral function. Within the brain stem, targets of non-LC NE/E neurons include cranial motor nuclei such as the dorsal motor vagus, motor trigeminal, motor facial, and the hypoglossal nucleus, as well as visceral–sensory nuclei such as the nucleus of the solitary tract, parabrachial complex, and periaqueductal gray. In the forebrain, non-LC cell clusters provide dense projections to visceral-related structures such as the bed nucleus of the stria terminalis (BNST), central nucleus of the amygdala, and paraventricular hypothalamus. Thalamic and epithalamic targets include the paraventricular nucleus, rhomboid nucleus, and habenula.

Within the different non-LC adrenergic nuclei, there is a preferential innervation of motor systems by A5 and A7 groups, whereas visceral–sensory areas are primarily innervated by medullary A1 and A2 nuclei. The E-containing fibers from C1 neurons

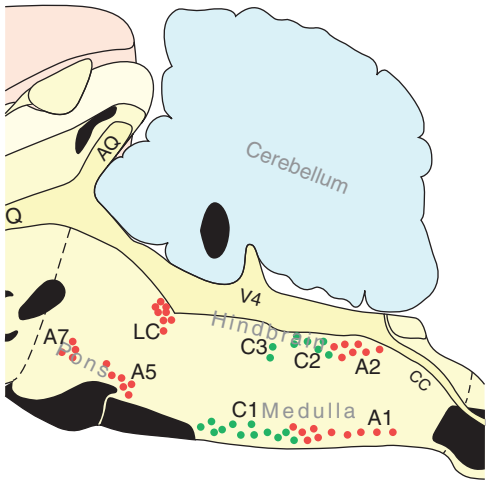

Figure 3 Localization of norepinephrine (red) and epinephrine (green) nuclei in the brain stem. The cell column in the ventral pons and medulla includes A1, C1, A5, and A7 cell groups, whereas the dorsal medulla column includes A2, C2, and C3 cell groups.

innervate both motor and visceral–sensory structures. C1 neurons send collaterals that innervate both spinal cord and brain stem targets. However, most C1 neurons that project to hypothalamus do not project to spinal cord or brain stem structures. In clear contrast to this pattern of organization, LC projections to the brain stem and spinal cord are directed almost exclusively to relay sites along ascending sensory pathways, including those associated with somatosensory, auditory, gustatory, and visual systems. This sensory-predominant bias of the LC efferent projection continues to the level of the thalamus and also includes the olfactory system. An overlap between LC and non-LC NE efferent projections to motor structures becomes evident in the cerebellar system, and the LC is the sole source of NE fibers to thalamic motor relay nuclei and motor (as well as sensory) processing regions of the cerebral cortex.

Non-LC impulse activity LC and non-LC neurons have similar electrophysiological signatures but exhibit differing responses to afferents and neurotransmitter systems. They are slow-firing neurons (<5 Hz) with wide action potentials (lasting more than 2 ms). Pharmacologically, both LC and non-LC neurons are inhibited by $\alpha 2$ adrenoceptor agonists, indicating the presence of a functional $\alpha 2$ autoreceptor. Conversely, sciatic nerve stimulation activates LC neurons, and also to a lesser extent A5 neurons, but not A2 neurons, indicating selectivity in the systems that activate the different adrenergic groups. Moreover, morphine strongly inhibits LC neurons but is reported to excite A5 neurons.

The organization of the non-LC adrenergic system and its connection to visceral-related areas prompted many studies of the non-LC adrenergic neurons; the findings indicate a diversity of roles in modulating cardiovascular, respiratory, renal, gastrointestinal, and endocrine functions. For example, the A1 area is involved in hemodynamic regulation by controlling the release of vasopressin, whereas C1 neurons are sensitive to blood pressure and are involved in regulating sodium and water balance. A5 neurons influence the respiratory rhythm generator of the rostral ventrolateral medulla, and A2 cells have a role in the regulation of food intake.

Stress Physiological (cardiovascular, osmotic, and immune challenges) as well as psychological (restraint and intermittent footshock) stressors activate both LC and non-LC NE/E neurons. These stressors also activate major targets of the non-LC system, such as the medial and central nuclei of the amygdala, BNST, and paraventricular hypothalamus, suggesting that non-LC adrenergic neurons are part of a functional stress

circuit. A significant difference between LC and non-LC cell groups may be habituation to stressors. Under chronic footshock regimens, LC neurons habituate whereas non-LC cells in A1, C1, A2, and C2 exhibit persistent stress responses. Thus, NE/E transmission in response to acute and persistent stressful conditions is differentially distributed across brain structures according to the unique efferent topographies of the LC and non-LC systems.

Drugs of abuse have profound effects on the activity of NE neurons throughout the brain. The aversiveness of opiate withdrawal is blocked by antagonizing β adrenoceptors in BNST; this NE input largely originates in A1/A2 neurons. A1/A2 neurons also appear to play a particularly prominent role in relapse to drug-seeking induced by stress. Several studies have shown that selective lesions of ascending A1/A2 NE projections, or antagonism of NE receptors in amygdala or BNST, attenuate relapse to drug-seeking induced by footshock stress.

Summary for non-LC systems In summary, several principles of organization emerge from examination of the efferent connections of the central NE/E cell groups. First, the distribution of adrenergic fibers is particularly dense in terminal fields innervated by non-LC adrenergic neurons (Table 1). Second, within the brain stem, the LC preferentially innervates primary sensory pathways while non-LC adrenergic nuclei innervate motor and visceral–sensory systems. Third, non-LC forebrain projections target visceral–sensory and autonomic circuits. Fourth, areas that subserve motor integrative functions such as the cerebellar circuit are mainly innervated by LC, with the exception of the inferior olive, which is mainly innervated by the A5 and A7 areas. Fifth, LC is the only source of NE innervation to the cortical mantle. Sixth, although all NE/E neurons have similar electrophysiological properties, LC and non-LC cells may respond differently to the same afferent inputs and neurotransmitter systems. This analysis leads to an appreciation of the complex anatomical relationships and physiological/behavioral properties of the different brain NE/E systems.

Postsynaptic Actions of Norepinephrine

Investigations in the 1960s and 1970s viewed NE as a putative conventional neurotransmitter and set out to determine its ability to increase or decrease the firing rate of target neurons. For the most part, local application of NE by microiontophoresis, or stimulation of LC, suppressed the spontaneous discharge of neurons in cerebellum, cerebral cortex, and elsewhere in the brain. These depressant responses were mimicked and blocked by adrenoceptor agonists and antagonists, respectively. Taken together, these findings support the idea that NE might function primarily as an inhibitory transmitter at central synapses.

Modulatory effects of NE on single neurons In the mid-1970s pioneering studies by Foote et al. in monkey auditory cortex demonstrated a differential depressant effect of microiontophoretically applied NE on single neurons, such that the spontaneous firing rate of recorded cells was suppressed to a greater extent than were stimulus-evoked discharges, thus yielding a net increase in the signal-to-noise ratio. This initial study prompted many other laboratories to investigate the actions of NE in a variety of sensory circuits within the mammalian brain, including olfactory bulb, sensory thalamus, visual, auditory, and somatosensory cortices. In many cases local application of NE was found to enhance extracellular responses of individual sensory neurons to synaptic stimuli; however, mixed effects were observed in other studies, with suppression of stimulus-evoked discharge predominating.

Later work by Waterhouse and colleagues examined NE actions at levels that had minimal or no effect on spontaneous discharge of target neurons. Under these conditions iontophoretic NE frequently produced absolute increases in stimulus-evoked discharge with minimal or no change in baseline firing. In other cases local administration of NE revealed robust cellular responses to otherwise subthreshold synaptic stimuli. The demonstration of NE-induced facilitation of cellular responses to direct iontophoretic application of GABA or glutamate under both normal and low-calcium/high-magnesium conditions argued that such effects were mediated postsynaptically. These findings in cerebral cortex and cerebellum of anesthetized and awake, behaving animals supported the general hypothesis that a prominent physiological function of central noradrenergic pathways is to enhance the efficacy of both excitatory and inhibitory synaptic transmission within target neuronal circuits, rather than to directly suppress cell firing.

Later work by Waterhouse et al. in cat and rat primary visual cortex showed that NE could alter specific receptive field properties (e.g., receptive field size, direction selectivity, velocity tuning) of visually responsive cells (Figure 4). Other studies showed that NE was capable of modifying the receptive field structure of cells in somatosensory and auditory cortices and cochlear nuclei. As such, these results go beyond the demonstration of simple monoamine-induced changes in the magnitude of synaptically evoked responses and begin to show that such actions can lead to selective alteration of the feature extraction properties

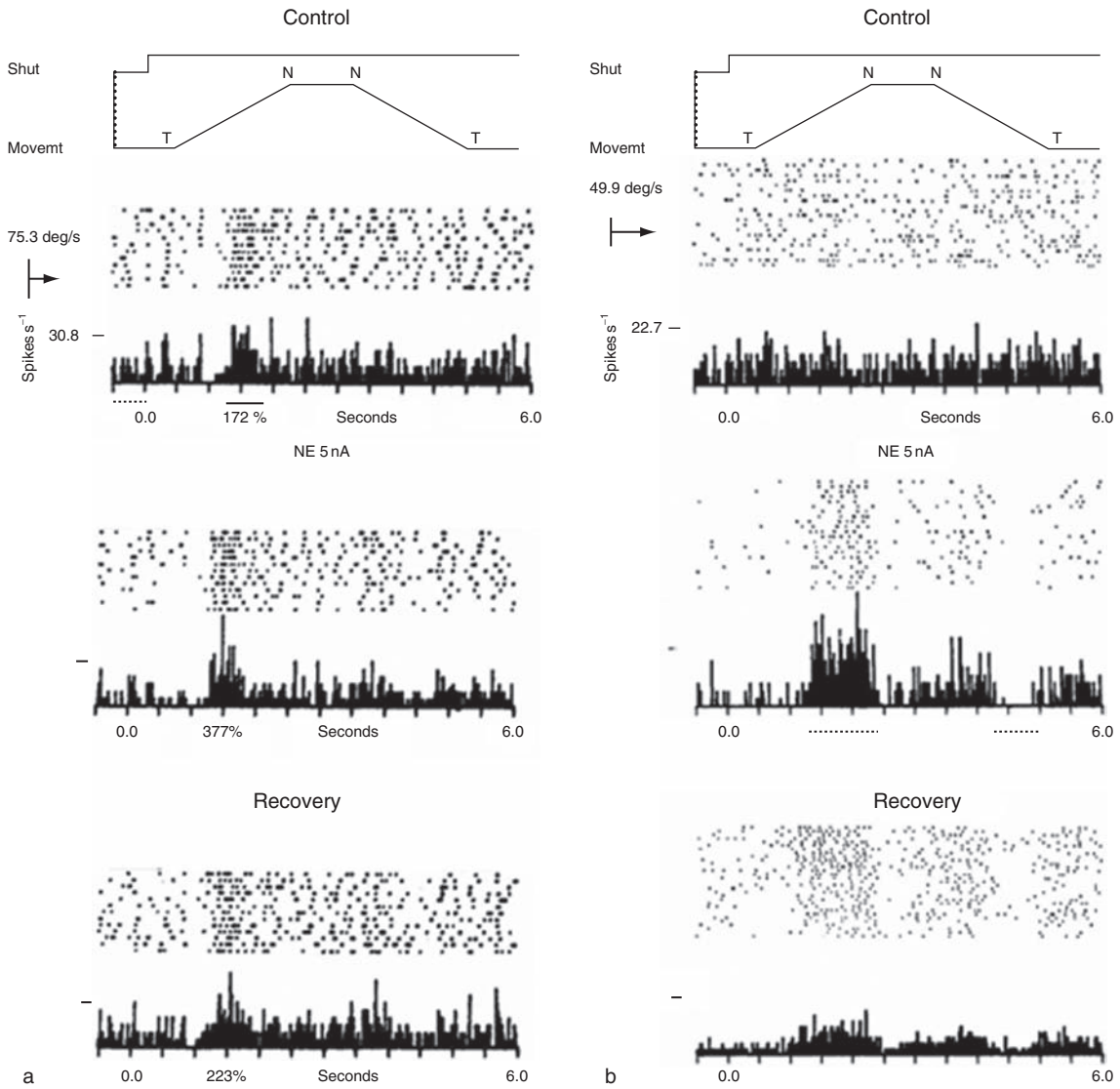

Figure 4 Effects of norepinephrine (NE) on responses of rat visual cortical neurons to visual field stimulation. (a, b) The raster and corresponding perievent histogram records show the extracellularly recorded responses of two single neurons in the rat visual cortex to movements of a vertically oriented (shown on the left) bar of light across the contralateral hemivisual field before, during, and after periods of microiontophoretic administration of NE. For each set of records, the uppermost line depicts the opening (step-up) and closing (step-down) of the shutter controlling the stimulus display. The trapezoid waveform indicates the position of the stimulus in visual space as it moves in the temporal (T) to nasal (N) direction and then back along the same path in the reverse direction. Time is indicated along the horizontal axis of the histogram, and discharge frequency of the cell is indicated along the vertical axis. Stimulus-bound excitatory or inhibitory components of the response were expressed as a percentage increase or decrease, respectively, of background firing rate (dotted line beneath histogram). (a) In the first cell the control response (top) consisted of an excitatory burst (solid bar) preceded by an inhibitory trough as the stimulus moved in the T to N direction, but no significant change in activity as the stimulus moved along the same path in the opposite direction (N to T). The visually evoked excitatory peak represented a 172% increase in firing over the basal level of discharge. Iontophoretic application of NE (5 nA) suppressed basal firing rate but produced an absolute increase in the visually evoked response (middle). Recovery to the control level of response was observed following cessation of NE administration (bottom). Each histogram sums unit activity during 13 trials; bin width = 15 ms. (b) In the second cell, there was no discernible response to the moving bar of light during the control period (top); however, during NE (5 nA) administration (middle) an obvious stimulus-evoked excitation (broken bar beneath histogram) and a stimulus-bound inhibition (dotted line beneath histogram) became evident. After cessation of NE iontophoresis (bottom), the cell gradually lost its sensitivity to the visual stimulus. Control and NE histograms, $n = 22$ trials; recovery histogram, $n = 31$ trials. Bin width = 15 ms.

of individual sensory neurons. More recent work by Arnsten, and colleagues indicates that NE and NE agonist drugs can modify the working-memory-related discharge of cells in prefrontal cortex of behaving monkeys. These latter findings suggest that the noradrenergic actions observed in sensory circuits may be generalized across brain regions that serve many different functions.

Additional studies in cortex and elsewhere have determined the pharmacologic specificity and dose dependency of NE-mediated modulatory effects. For example, the magnitude of cellular response to direct glutamate application can be maximally enhanced at a specific level of NE administration, and only moderately or minimally be affected by concentrations above and below this level (Figure 5(a)), thus revealing an inverted-U dose–response relationship for this effect of NE. Furthermore, noradrenergic enhancement of neuronal responses to excitatory synaptic stimuli are mimicked and blocked by α1-receptor agonists (Figure 5(b)) and antagonists, respectively, whereas the augmentation of inhibitory responses to GABA application (Figure 6) or afferent pathway stimulation involves activation of β receptors. Other studies, particularly those in hippocampus, provide data suggesting the opposite (i.e., NE enhancement of excitatory synaptic transmission is dependent upon β-receptor mechanisms and augmentation of synaptic inhibition is mediated by α1-receptor mechanisms). These disparate results have not been reconciled but may reflect unique receptor distributions or second-messenger linkages among neurons within different brain regions. Overall, these observations indicate that within a specified brain circuit

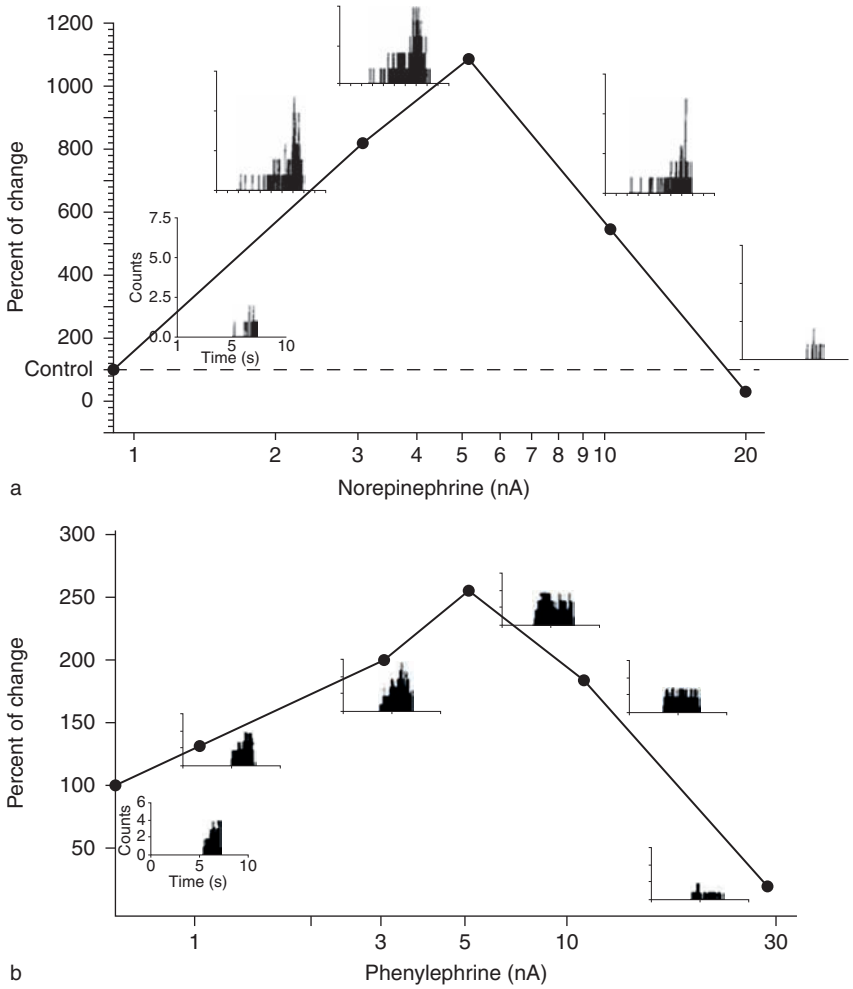

Figure 5 Modulatory actions of ionotphoretically applied norepinephrine and α1 agonist (phenylephrine) on glutamate-evoked discharges in rat barrel field cortical neurons. Responses of individual layer V neurons to iontophoretic pulses of suprathreshold levels of glutamate were quantified by analysis of perievent histograms (shown as insets) that were collected before, during, and after continuous microiontophoretic application of norepinephrine. Responses were then expressed as a percentage change from control (i.e., response to glutamate application alone = 100%) for different iontophoretic doses (nA) of norepinephrine. Each histogram and corresponding point on the graph represents the cell's response to six consecutive glutamate applications. (a) Example of a layer V cortical neuron: its excitatory response to glutamate (28 nA) was progressively increased with increasing levels of norepinephrine (peak at 5 nA) and then progressively suppressed from this optimal value with further increases in norepinephrine ejection current (6–20 nA). (b) Example of a second layer V cortical neuron: its excitatory response to glutamate (40 nA) was likewise progressively increased with increasing levels of phenylephrine (peak at 5 nA) and then progressively suppressed from this peak value with further increases in phenylephrine ejection current (10–30 nA).

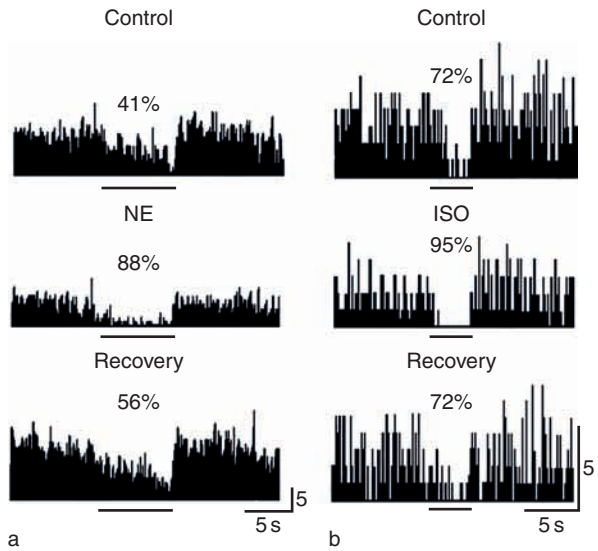

Figure 6 Actions of norepinephrine (NE) and the β agonist isoproterenol (ISO) on γ-aminobutyric acid-induced inhibition of cortical neuron spontaneous discharge. Histogram records show responses of two somatosensory cortical cells (a and b) to regularly spaced, uniform iontophoretic pulses of γ-aminobutyric acid applied before, during, and after NE (a) or ISO (b) administration. Numbers above each histogram indicate the percentage of inhibition of spontaneous firing rate produced by γ-aminobutyric acid application. Each histogram sums unit activity for an equal number of γ-aminobutyric acid pulses, $n = 5$. Both NE (10 nA) and ISO (15 nA) produced a marked enhancement of the γ-aminobutyric acid response relative to a suppression of background firing. Recovery to the control level of response was observed in each case after termination of adrenergic drug application. Calibration: vertical, 5 counts per address; horizontal, 5 s.

an array of noradrenergic modulatory actions is possible, depending upon cellular expression of receptors and levels of NE available for interaction with those receptors.

Cellular mechanisms of NE action Intracellular experiments have demonstrated multiple actions of NE on properties of neuronal membranes. For example, NE-induced blockade of a calcium-activated potassium conductance leads to accommodation block and prolongation of evoked excitatory discharges. It is widely accepted that this effect, in conjunction with a membrane hyperpolarizing action of NE, can account for noradrenergic potentiation or signal-to-noise ratio enhancement of threshold level excitatory synaptic responses observed *in vivo*. McCormick and colleagues have found that NE's depolarizing action influences a complex interaction between intrinsic membrane currents in thalamic relay neurons. This interplay moves thalamic neurons back and forth between a non-signal-transmitting oscillatory mode during sleep and a single-spike transmission, or 'relay,' mode during waking.

Available evidence implicates the β-receptor-linked cyclic AMP system in NE-mediated augmentation of GABA-induced inhibition. For example, *in vivo*, local application of agents that elevate intracellular levels of cyclic AMP mimic or augment NE-mediated enhancement of GABA-induced inhibition of cerebellar Purkinje neurons. Subsequent studies in acutely dissociated cerebellar Purkinje neurons demonstrated augmentation of GABA membrane currents via intracellular diffusion of 8-bromo-cyclic AMP or the catalytic subunit of protein kinase A. And finally, additional investigations found that subunits of the GABA-A receptor can be phosphorylated by protein kinase A. The working hypothesis is that, upon release, NE initiates a cascade of events via postsynaptic β receptors that results in activation of protein kinase A and phosphorylation of GABA-A receptor subunits. In Purkinje neurons this phosphorylation of the receptor leads to enhanced GABA binding capacity or chloride conductance, thus providing for an overall increase in GABA response.

The intracellular events associated with NE-induced enhancement of stimulus-evoked excitatory discharges have not been identified with as much certainty, although work by several laboratories identifies protein kinase A and protein kinase C as candidates for mediating transmitter/modulator effects on synaptically evoked excitation. Elucidating the intracellular events and accompanying physiological actions that underlie NE-induced enhancement of responsiveness to synaptic inputs are major issues for future studies.

Effects of locus coeruleus output on neuronal and neural circuit response properties Studies in anesthetized animals show that stimulation of the LC efferent path elicits NE-like modulatory effects on target neurons in sensory circuits. These findings support the general hypothesis that a major function of the central LC–noradrenergic efferent system is to facilitate the transfer of information through sensory circuits following events that phasically activate the LC nucleus. More recent work in waking animals has shown that phasic and tonic activation of the LC efferent path can enhance the probability of stimulus-evoked spiking in both cortical and thalamic neurons according to an inverted-U function. Thus, as described for iontophoretically applied NE, neuronal responses to afferent stimulation are increased to a maximum at intermediate levels of LC stimulation, and to submaximal levels with LC activation levels above and below this level. As reported for iontophoretically applied NE, activation of the LC efferent pathway can also reveal or 'gate' cellular responses to otherwise subthreshold synaptic inputs. LC stimulation also decreases mean response latency and reduces

trial-to-trial variability of response latency for both sensory thalamic and cortical neuron responses to afferent input. Thus, increased output from the LC is capable of altering the magnitude and timing of cellular responses to sensory-driven synaptic input as well as recruiting (gating) additional cells into a response network. These results are important in that they extend observations of LC- and NE-mediated modulatory actions into the waking condition. Further, they show that LC output can simultaneously facilitate stimulus-evoked discharge patterns at multiple levels of an ascending sensory pathway.

The observed effects of LC stimulation on response threshold, response magnitude, and response latency in individual cells suggests that under certain behavioral conditions synaptically released NE can optimize information processing within sensory circuits and provide for a network-wide enhancement of sensory signal transmission. This idea is further supported by recent multineuron recording studies in which LC activation and NE release have been shown to modulate the representation of sensory information across ensembles of functionally related thalamic neurons.

Functional and Clinical Implications

The widespread NE system appears to be functionally multifaceted. On a 24 h timescale it modulates sleep and waking, permitting sleep with inactivity but promoting arousal by becoming active. Recent results for a circuit from the SCN to the LC indicate that this system may contribute to the circadian regulation of arousal and sleep–waking. However, on a much shorter timescale of seconds, this system appears to modulate attention and performance during the waking state. In this domain the LC system regulates behavioral flexibility by promoting either stable selective attention and focused behavioral performance or a scanning type of attention and more flexible behavior.

These findings have implications for the involvement of the LC system in clinical disorders. The role proposed for the LC in attention suggests that dysfunction of this system could contribute to the cognitive and sensory signal-processing deficits that are prominent in schizophrenia, attention-deficit/hyperactivity disorder, and dementia. In regard to dementia, LC neurons have been found to degenerate in both Alzheimer's and Parkinson's diseases.

Many clinical studies have also implicated this system in depression and anxiety. This is consistent with findings that there are substantial projections of NE neurons throughout the limbic system and that the LC system is engaged by novel and alarming stimuli. The high responsiveness of LC neurons to acute and chronic stressors is also consistent with a role for this system in depression. Moreover, many of the drugs used successfully to treat anxiety and depression target the noradrenergic system. There is, as well, a growing body of clinical evidence suggesting a role for the LC–NE system in posttraumatic stress disorder.

There have also been various proposals for a role of the LC in the etiology of sleep disorders. This is consistent with both the close association of activity of these cells with different stages of the sleep–wake cycle and the disrupted sleep patterns that are a hallmark of depression. Recent results indicate that the role of the LC extends to circadian disorders in sleep and performance as well.

See also: Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Monoamines: Release Studies; Norepinephrine: Adrenergic Receptors; Sleep–Wake State Regulation by Noradrenaline and Serotonin.

Further Reading

Aston-Jones G, Chen S, Zhu Y, et al. (2001) A neural circuit for circadian regulation of arousal. *Nature Neuroscience* 4: 732–738.

Aston-Jones G and Cohen JD (2005) Adaptive gain and the role of the locus coeruleus–norepinephrine system in optimal performance. *Journal of Comparative Neurology* 493: 99–110.

Aston-Jones G and Cohen JD (2005) An integrative theory of locus coeruleus–norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience* 437: 556–559.

Berridge CW and Foote SL (1991) Effects of locus coeruleus activation on electroencephalographic activity in the neocortex and hippocampus. *Journal of Neuroscience* 11: 3135–3145.

Berridge CW and Waterhouse BD (2003) The locus coeruleus–noradrenergic system: Modulation of behavioral state and state-dependent cognitive processes. *Brain Research Reviews* 42: 33–84.

Clayton EC, Rajkowski J, Cohen JD, et al. (2004) Phasic activation of monkey locus coeruleus neurons by simple decisions in a forced choice task. *Journal of Neuroscience* 24: 9914–9920.

Florin-Lechner S, Druhan J, Aston-Jones G, et al. (1996) Enhanced norepinephrine release in prefrontal cortex with burst stimulation of the locus coeruleus. *Brain Research* 742: 89–97.

Foote SL, Bloom FE, and Aston-Jones G (1983) Nucleus locus coeruleus: New evidence of anatomical and physiological specificity. *Physiological Reviews* 63: 844–914.

Guyenet PG (1991) Central noradrenergic neurons: The autonomic connection. *Progress in Brain Research* 88: 365–380.

Hokfelt T, Johansson O, and Goldstein M (1984) Central catecholamine neurons are revealed by immunohistochemistry with special reference to adrenergic neurons. In: Bjorklund A and Hokfelt T (eds.) *Classical Transmitters in the CNS. Part I*, pp. 157–276. Amsterdam: Elsevier.

Moore RY and Card JP (1984) Noradrenaline-containing neuron systems. In: Bjorklund A and Hokfelt T (eds.) *Classical Transmitters in the CNS. Part I*, pp. 123–156. Amsterdam: Elsevier.

Ramos BP and Arnsten AF (2007) Adrenergic pharmacology and cognition: Focus on the prefrontal cortex. *Pharmacology & Therapeutics* 113: 523–536.

Sawchenko PE, Li HY, and Ericsson A (2000) Circuits and mechanisms governing hypothalamic responses to stress: A tale of two paradigms. *Progress in Brain Research* 122: 61–78.

Shaham Y, Shalev U, Lu L, et al. (2003) The reinstatement model of drug relapse: History, methodology and major findings. *Psychopharmacology (Berlin)* 168: 3–20.

Valentino RJ, Foote SL, and Page ME (1993) The locus coeruleus as a site for integrating corticotropin-releasing factor and noradrenergic mediation of stress responses. *Annals of the New York Academy of Sciences* 697: 173–188.

Woodward D, Moises H, Waterhouse B, et al. (1979) Modulatory actions of norepinephrine in the central nervous system. *Federation Proceedings* 38: 2109–2116.

Notch Pathway: Lateral Inhibition

A Chitnis, National Institute of Child Health and Human Development, Bethesda, MD, USA

© 2009 Published by Elsevier Ltd.

The Notch Signaling Pathway

Briefly, at the core of the Notch signaling pathway is the Notch receptor, a single-pass transmembrane protein (Figure 1). The mature form of the receptor expressed at the cell surface is a heterodimer of an extracellular fragment ($Notch^{EC}$) and a membrane-spanning fragment ($Notch^{TM}$), generated during synthesis in the Golgi by proteolytic cleavage of the Notch protein at the S1 site by a furin convertase (Figure 1). Interaction of Notch with the Delta, Serrate, Lag-2 (DSL) family of ligands expressed on the surface of neighboring cells helps complete a sequence of three proteolytic cleavages that result in the release of the intracellular Notch domain ($Notch^{IC}$) from the plasma membrane (Figure 1). Activation of the Notch receptor requires that the ligands undergo endocytosis, an event that involves ubiquitination by RING E3 ligases such as Mindbomb and Neuralized (Figure 2). Internalization of the DSL ligand–$Notch^{EC}$ complex by the endocytic machinery is likely to provide a force that separates the $Notch^{EC}$ and $Notch^{TM}$ fragments, facilitating subsequent cleavage of Notch that releases the transcriptionally active $Notch^{IC}$ fragment into the cell (Figure 2). While interaction of Notch and a DSL ligand in a neighboring cell (in *trans*) facilitates activation of the Notch receptor, interaction of Notch with Delta in the same cell (in *cis*) prevents effective function of both Delta and Notch (Figure 1). The $Notch^{IC}$ fragment translocates to the nucleus, where, together with a member of the DNA-binding CBF1–RBP-Jk, Suppressor of Hairless, Lag-1 (CSL) family of transcription factors and cofactor Mastermind, it drives expression of target genes with regulatory DNA sequence recognized by the CSL protein DNA-binding domain (Figure 2). In the absence of $Notch^{IC}$, the CSL protein is part of a repressor complex that maintains basal repression of target genes recognized by this DNA-binding factor (Figure 2).

Lateral Inhibition and Selection of a Cell within a Proneural Cluster

Notch signaling is best known for its role in mediating interactions that ensure adjacent cells acquire distinct fates (Figure 2 and Figure 3). During lateral inhibition it ensures that only a single cell of a larger group of cells with the potential to adopt the same fate is selected to adopt a particular fate. This role is best illustrated by the function of Notch in neurogenesis.

During *Drosophila* neurogenesis, early patterning mechanisms initiate expression of 'proneural' genes in a small group of cells called a proneural cluster (PNC). These genes, including *achaete*, *scute*, *lethal* of *scute*, and *asense* in the Achaete–Scute complex or another family of genes related to *atonal*, encode basic helix–loop–helix (bHLH) transcription factors that give the cells in the PNC the potential to adopt a neural fate. The proneural factors can drive their own expression, and this positive autoregulation can drive progressively higher levels of proneural gene expression. If sufficiently high expression is achieved, it can become self-sustaining, independent of initiating factors, and the proneural factors are able to successfully initiate expression of additional genes that facilitate subsequent adoption of a neural fate. However, positive autoregulation can be regulated by additional factors that make it more or less effective, eventually determining whether critical levels of proneural factor expression can be established in a particular cell and hence whether that cell will actually adopt a particular neural fate.

The proneural factors drive expression or facilitate the function of the Notch ligand Delta (Figure 2). So as cells express higher levels of proneural factor, they become progressively better at activating Notch in neighboring cells. When Delta activates Notch in a neighboring cell within the PNC, it initiates the expression of genes in the Enhancer of Split (E(spl)) complex. These genes interfere with proneural factor expression, making it harder for the neighbor to acquire high levels of proneural expression. Like the proneural factors encoded by genes in the Achaete–Scute complex, some genes in the E(spl) complex encode bHLH transcriptional factors; however, they are part of a larger evolutionarily conserved family of Hairy E(spl)-related factors that are characterized by a C-terminal WRPW domain, which allows them to recruit corepressor Groucho and function as repressors to prevent expression of proneural genes. By promoting function of the Notch ligand Delta and activating Notch-mediated E(spl) transcription in neighboring cells, proneural factors indirectly inhibit proneural gene expression in neighboring cells, reducing their ability to adopt a neural fate. As a consequence of this process of lateral inhibition, only a single cell eventually acquires high enough levels of proneural gene expression to adopt a neural fate within a PNC. Adjacent cells in the PNC, characterized by relatively high levels of Notch activity, are prevented from adopting a neural fate, and they

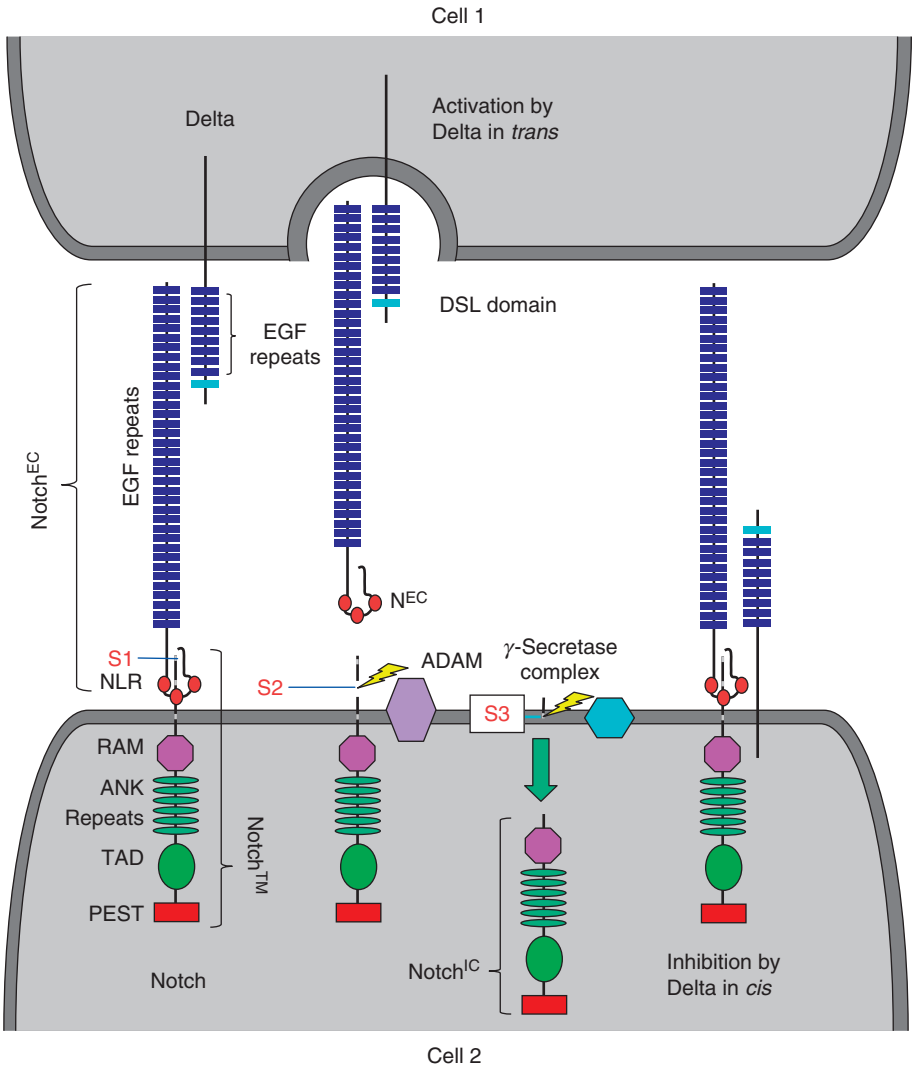

Figure 1 Schematic of Notch and Delta showing functional domains and proteolytic cleavage. The mature Notch receptor is a heterodimer containing a Notch^{EC} fragment and a Notch^{TM} fragment made by furin-based cleavage of the full length Notch protein at the S1 site. The Notch–Lin12 Repeats (NLR) in the Notch^{EC} fragment prevent access to the S2 site in the Notch^{TM} fragment. Endocytosis of Delta following interaction with Notch^{EC} in an adjacent cell (in *trans)* facilitates removal of the Notch^{EC} fragment and allows an ADAM protease access to the S2 cleavage site. S2 cleavage is followed by γ secretase-dependent cleavage at an S3 site, which releases the Notch^{IC} fragment into the cell. The Notch^{IC} fragment contains a RAM domain, ankyrin (ANK) repeats, and a transactivation domain (TAD), which allows it to function in transcriptional activation complex with Mastermind and a CSL protein like Su(H) in the nucleus. The PEST sequence allows rapid degradation of the Notch^{IC} fragment in the nucleus. Interaction with Delta in the same cell (*cis*) interferes with function of Notch. Notch^{EC}, extracellular fragment of Notch receptor; Notch^{IC}, intracellular fragment; Notch^{TM}, membrane-spanning fragment.

acquire the ability to adopt an alternate fate that is dependent on relatively high levels of Notch signaling. Failure of lateral inhibition mediated by Notch signaling results in a 'neurogenic' fate as too many cells within the PNC adopt a neural fate. This is accompanied by a deficit in the number of cells with the alternate fate that is dependent on relatively high levels of Notch signaling. Based on the neurogenic phenotype produced by loss of Notch signaling, many of the genes originally identified in the Notch signaling pathway were referred to as 'neurogenic' genes.

The genetic regulatory network described above, first discovered in *Drosophila* and subsequently in many other metazoan animals, describes the manner in which Delta–Notch signaling mediates lateral inhibition in a PNC to select a single cell that will adopt a neural fate. Though the details differ, this is how cells are selected for differentiation with distinct fates during neurogenesis in a wide range of circumstances in the central nervous system (CNS) and peripheral nervous system (PNS). In *Drosophila* during early neurogenesis, cells from PNCs in the neuroectoderm are

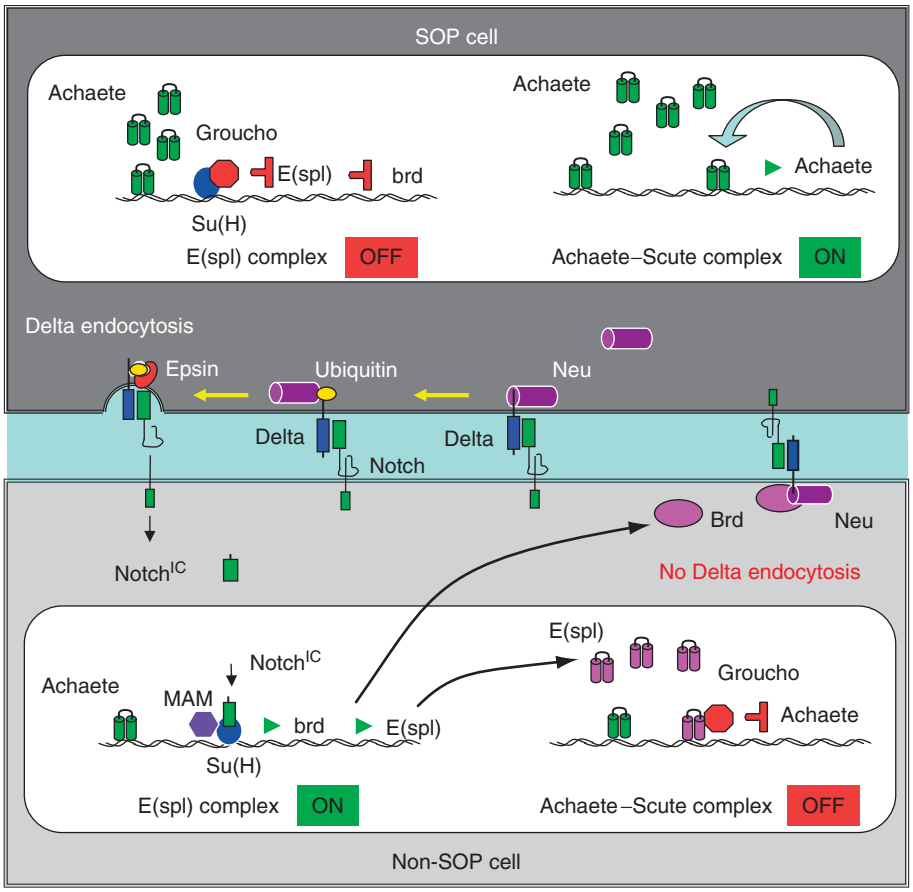

Figure 2 The dual role of Su(H) in activation and repression ensures complementary expression of genes in the Achaete–Scute and Enhancer of Split (E(spl)) complexes in the sensory organ precursor (SOP) and non-SOP cells. Relatively low levels of Notch activation in the prospective SOP cell permits effective self-activation by genes in the Achaete–Scute complex, which results in high levels of *achaete* expression. Despite presence of achaete-binding sites in the E(spl)-C enhancers, high levels of achaete proteins do not drive *E(spl)* expression in the SOP cell because, in the absence of Notch^{IC}, Su(H) maintains basal repression of genes in E(spl)-C. Suppression of E (spl) bHLH genes prevents repression of genes in the Achaete–Scute complex, and repression of *Bearded* (*brd*) genes prevents inhibition of Neuralized function in the prospective SOP. Hence the prospective SOP effectively activates Notch in the surrounding non-SOP cells, where Notch^{IC} makes Su(H) function as part of an activator complex and drives expression of genes in the E(spl)-C. In these non-SOP cells, bHLH E(spl) proteins repress *achaete* expression, while *brd* proteins prevent effective Neuralized function. E(spl), Enhancer of Split; Neu, Neuralized; Notch^{IC}, intracellular fragment of Notch receptor.

selected to become neuroblasts during formation of the CNS. Cells selected to become neuroblasts delaminate deep to the neuroectoderm and aggregate to form the neural anlage, while cells that are not selected to become neuroblasts remain superficial and eventually contribute to the formation of the epidermis. In the PNS, sensory organ precursors (SOPs) are selected from PNCs in imaginal discs. Cells in the PNC that are not selected to become an SOP also contribute to the epidermis, while the SOP cell undergoes a sequence of stereotypical divisions to produce cells of the sensory organ. Lateral inhibition also selects the R8 photoreceptors within PNCs at the advancing edge of the morphogenetic wave in the developing eye. In this context the selected cell does not undergo additional divisions; however, specification of R8 cells is critical because the R8 cell initiates the specification of the surrounding photoreceptors, and they in turn help specify additional cells of the ommatidium.

Cis-Regulatory Logic of E(Spl) Expression in a PNC

E(spl) and proneural genes are eventually expressed in complementary patterns in the PNC, with proneural expression restricted to the SOP, and E(spl) gene expression restricted to surrounding non-SOP cells. Analysis of enhancer elements that direct expression of E(spl) genes in PNCs of the wing imaginal disc reveal how the dual role of the CSL protein Su (H) in either transcriptional repression or Notch-mediated transcriptional activation establishes complementary expression of E(spl) and proneural genes in the PNC (Figure 2). The enhancer that determines Notch-mediated expression of E(spl) genes in the

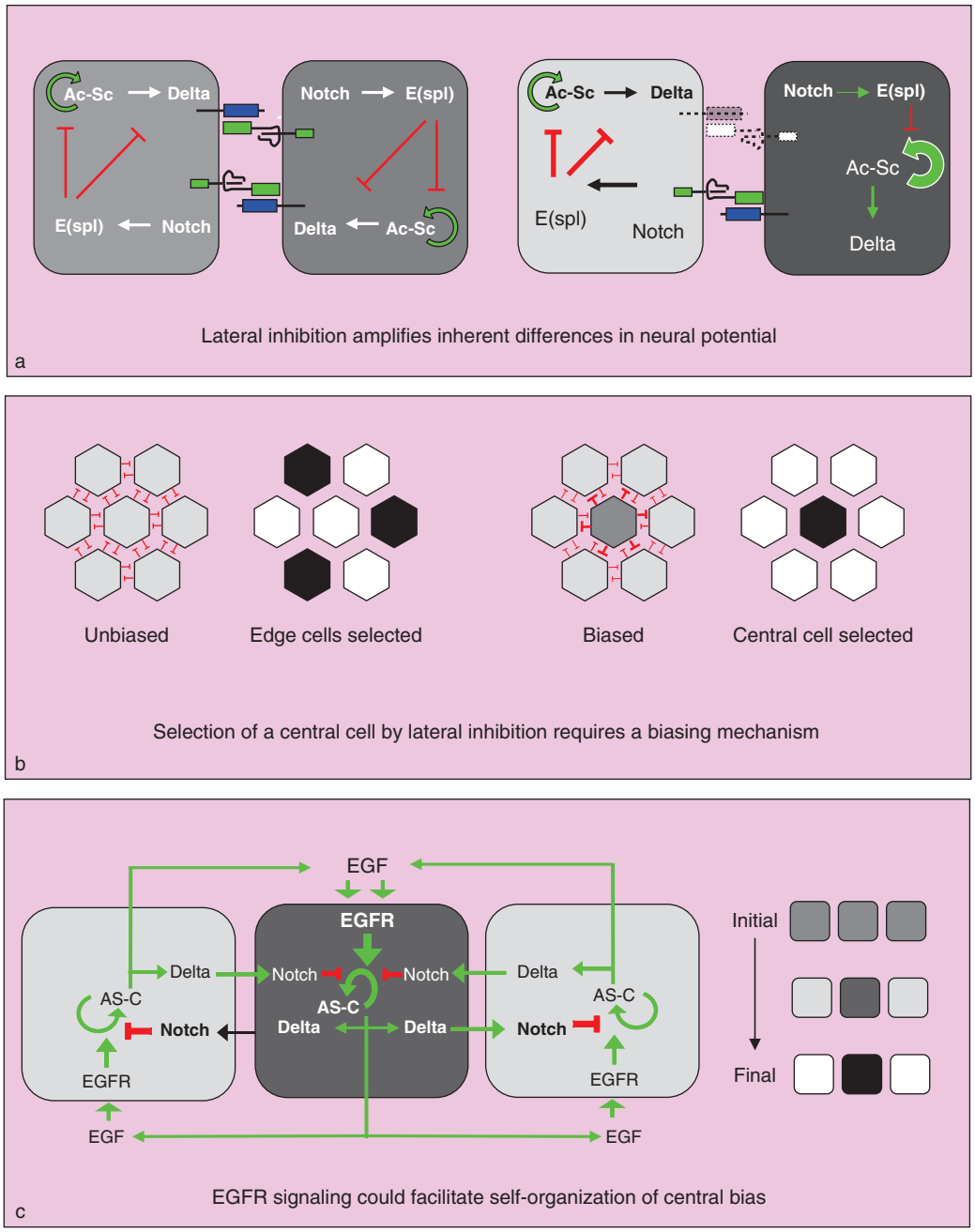

Figure 3 Lateral Inhibition selects only central cells when coupled with a central biasing mechanism. (a) In lateral inhibition, Notch signaling is part of negative feedback loop that amplifies differences in neural potential in adjacent cells. (b) If proneural genes were equally expressed in a proneural cluster (PNC), lateral inhibition would favor selection of cells at the edge of the cluster because edge cells do not receive inhibition from cells outside the PNC; in contrast, central cells receive inhibition from all surrounding neighbors. However, lateral inhibition could successfully select a central cell when coupled to a mechanism that ensures a central bias. Red bars represent lateral inhibition delivered to neighboring cells. (c) EGFR signaling promotes positive autoregulation by proneural genes. If each cell in the PNC were to be a source of a diffusible EGFR ligand, central cells would be exposed to the highest levels of the ligand and would have most efficient autoregulation of proneural genes. This defines a potential mechanism that facilitates self-organization of a central bias in proneural function that can then be amplified by Notch signaling leading to selection of a central cell. Ac-Sc, Achaete–Scute complex; EGF, epidermal growth factor; EGFR, epidermal growth factor receptor; E(spl), Enhancer of Split.

wing imaginal disc PNCs includes high-affinity sites for Su(H) binding, which allows E(spl) expression promoted by Notch^{IC}, Mastermind, and Su(H). However, the E(spl) enhancer also includes high-affinity sites for binding of proneural proteins Achaete and Scute. This seems paradoxical because the domain within the PNC where E(spl) genes are expressed is exactly where proneural expression is eventually

excluded; however, the combination of Su(H) and proneural binding sites in the enhancer ensures that Notch activation directs expression of E(spl) only within the PNC, where proneural expression has already been initiated. Though the E(spl) enhancer includes high-affinity proneural binding sites, high levels of proneural protein expression in the prospective SOP cannot independently drive E(spl) gene expression because in the absence of Notch^{IC}, Su(H) functions as part of a repressor complex that ensures genes in the E(spl) complex are not expressed. In surrounding non-SOP cells of the PNC, where Notch is activated by the SOP cell, Notch^{IC} makes Su(H) function as an activator to drive E(spl) gene expression.

The Complementary Role of Bearded Family Genes in the E(Spl) Complex

The E(spl) complex also includes non-bHLH genes of the *Bearded* (*Brd*) family. Their expression is also regulated by Notch activation; however, unlike the E(spl) bHLH genes, they do not inhibit proneural gene expression. Instead, they interfere with Delta function by inhibiting Neuralized mediated endocytosis of Delta, a step that is essential for effective activation of Notch in the neighboring cell (Figure 2). By inhibiting Neuralized function, *Brd* genes play a complementary role to bHLH E(spl) genes in making cells with Notch activation less likely to be selected as SOPs. Brd protein makes cells less effective at delivering Delta-mediated lateral inhibition to their neighbors and hence unlikely to win the competition to become an SOP. In the absence of Notch activation, Su(H) maintains repression of *Brd* genes in the prospective SOP cell, facilitating Neuralized function and making it more effective at delivering lateral inhibition to neighboring non-SOP cells.

Notch Signaling in the Progeny of Cells Selected by Lateral Inhibition

Though lateral inhibition typically refers to interactions between cells in the PNC that select a single cell, Notch signaling also has a critical role in determining distinct fates for the progeny of cells initially selected by lateral inhibition. For example, when a neuroblast divides, the daughter with less Notch activation becomes a ganglion mother cell (GMC), and the sibling with higher Notch activation becomes another neuroblast. When the GMC divides, the daughter with more Notch activity becomes a glial cell, and its sibling becomes a neuron. This lineage illustrates how the cell with less Notch activity is consistently permitted to adopt a fate that brings the cell closer to differentiating as a neuron, while the cell with greater Notch activity either remains as an undifferentiated progenitor or adopts a 'nonneural' fate. A similar pattern, discussed in more detail below, is seen in the lineage of the SOP.

Selecting a Central Cell within a PNC

The *Drosophila* epidermis develops more than 1000 bristles and other types sensory organs during development. Some of these bristles, called macrochaetes, are conspicuously large and located in stereotyped positions on the head and notum. Each macrochaete is derived from an SOP, previously also called a sense organ mother cell. In the wing imaginal disc, each SOP is, on average, selected from a cluster of 20–30 cells, called a PNC. All the cells in the PNC initially acquire the potential to become an SOP by their expression of proneural genes like *achaete* and *scute* that are part of the Achaete–Scute complex. However, competitive interactions mediated by Delta–Notch interactions eventually restrict high levels of proneural expression and SOP fate to a single cell or sometimes two cells within the PNC. High levels of proneural expression are first restricted to a smaller subset of cells within the PNC, called a proneural field. Then, within this field, the cell that attains the highest level of *achaete–scute* expression begins expressing *asense*, another proneural gene in the Achaete–Scute complex, and it acquires an SOP fate. The initial expression of proneural genes in stereotypical clusters is determined by the combinatorial function of heterogeneous 'prepattern' genes; however, the sustained high expression of proneural genes in a subset of cells depends on positive autoregulation by the proneural genes.

Biasing the Outcome of Notch Signaling

What determines which cell is eventually selected in a cluster and how is it reliably selected in a particular location within a large PNC? In the absence of additional biasing mechanisms, a central cell is least likely to be selected by lateral inhibition, and cells at the edge of a PNC are expected to have a competitive advantage (Figure 3). While cells at the edge do not receive lateral inhibition from neighbors outside the PNC, neighbors delivering lateral inhibition surround central cells.

Despite the inherent disadvantage of central cells with respect to lateral inhibition, it is often the case that they are the ones selected by lateral inhibition. Analysis of enhancer elements of the Achaete–Scute complex and epidermal growth factor receptor (EGFR) signaling suggests that interactions between EGFR and proneural factors might contribute to establishment of a central biasing mechanism during

selection of the SOP in the PNC for specific macrochaetes. The analysis of Achaete–Scute complex enhancers shows that one set of regulatory elements in a 'PNC enhancer' are responsible for initiating Achaete–Scute complex gene expression within a PNC in response to regulation by prepatterning genes, while another set of regulatory elements in a distinct 'SOP enhancer' mediate positive autoregulation by the proneural genes. The SOP autoregulatory enhancer is responsible for high levels of Achaete–Scute complex proneural expression in the cell that is eventually selected to become the SOP within the PNC. Proneural expression driven by proneural self-stimulation in the SOP enhancer is facilitated by EGFR signaling, and EGFR signaling is itself promoted by the proneural factors. As all the cells within the PNC secrete diffusible EGFR signals, central cells within a PNC are exposed to the highest levels of EGFR signaling. This results in particularly effective positive autoregulation of Achaete–Scute complex genes within central cells, biasing them for selection by lateral inhibition. Since diffusible EGFR ligands promote proneural expression by acting via the autoregulatory enhancer, their effect is limited to cells in the original PNC, where proneural expression was initiated by early patterning mechanisms, so EGFR signaling is unlikely to initiate ectopic proneural expression in cells outside the PNC.

Variations of the scheme described above play a role in biasing the outcome of lateral inhibition in various contexts. Selection of the progenitor for the *Drosophila* chordotonal organ, for example, is determined by the expression of another proneural gene, *atonal*, and it also involves an antagonistic relationship between EGFR and Notch signaling; however, in this context the initial progenitor selected by lateral inhibition continues to be the source of inductive EGFR signals, and it eventually recruits many more progenitors so that, unlike the SOP for external sensory bristles, the chordotonal organs are not isolated structures but organized to form small clusters.

Biasing Binary Cell Fate Decisions in the SOP Lineage

Competitive interactions mediated by Notch signaling continue to determine distinct cell fates in the progeny of cells initially selected by lateral inhibition. Here, the outcome of the competitive interactions is biased by the asymmetric inheritance of factors that can influence the efficacy of Notch signaling by progeny of the dividing cells. SOPs or the pI cells on the dorsal surface of the fly thorax (notum) undergo a series of stereotyped asymmetric divisions to generate cells that make cells of adult mechanosensory organs (Figure 4). The first of the asymmetric divisions is along the mediolateral axis and generates an anterior pIIb cell and a posterior pIIa cell. The pIIa cell divides once more to form a socket and a shaft cell, which are the external cells of mechanosensory organ. The pIIb cell divides twice to produce the internal cells of the sensory organ: the first division produces a glial cell and the pIIIb cell, and the pIIIb cell divides to form a sensory neuron and sheath cell. pIIa fate is dependent on high Notch activation, while pIIb fate is dependent on low Notch activation. Each sibling expresses both Delta and Notch and, in principle, could compete to acquire a dominant role in either delivering or receiving a Notch signal. However, asymmetric distribution of factors like Numb, Neuralized, and Sanpodo, determined by planar polarity signaling mechanisms, ensure that the anterior pIIb cell becomes more effective at delivering the Delta signal and the posterior pIIa cell more effective at having its Notch receptors activated. At pro-metaphase, planar polarity mechanisms segregate Partner of Inscrutable (Pins) to the anterior cortex of the pI cell, opposite to components of Par complex, including Bazooka (D-Par3), DaPKC, and DmPar6, which localize on the posterior cortex. Pins restricts localization of Baz to the posterior cortex of the dividing pI cell, and segregation of Baz along with DaPKC and DmPar6 at the posterior cortex restricts another factor, Lethal Giant Larvae, to the opposite anterior cortex. Lethal Giant Larvae in turn recruits Partner of Numb and Neuralized to the anterior cortex. As a consequence, following mitosis, Numb and Neuralized are segregated to the anterior pIIb cell. Numb prevents membrane localization of Sanpodo, a requirement for Notch function, and Numb also inhibits Notch in the pIIb cell by promoting its degradation. In addition, segregation of Neuralized to the pIIb cell allows this cell to internalize Delta in a manner that makes it effective in activating Notch in its neighbor. In contrast, absence of Numb and plasma membrane localization of Sanpodo make Notch effective in the pIIa cell, while absence of Neuralized prevents Delta in this cell from effectively activating Notch in its adjacent pIIb sibling. In this manner planar polarity mechanisms ensure that for the pII progeny of all the SOPs on the developing notum, the anterior sibling acquires a pIIb fate, and the posterior sibling acquires a pIIa fate. The coordinated fate of individual SOP progeny eventually ensures the coordinated orientation of all the sensory bristles in the notum.

A mechanism acting in parallel establishes an asymmetry in Delta recycling in the pIIa and pIIb cells (Figure 4). Shortly after division of the SOP, a Rab11-dependent recycling center is established around the centrosome in the pIIb cell. Establishment of this center is dependent on accumulation

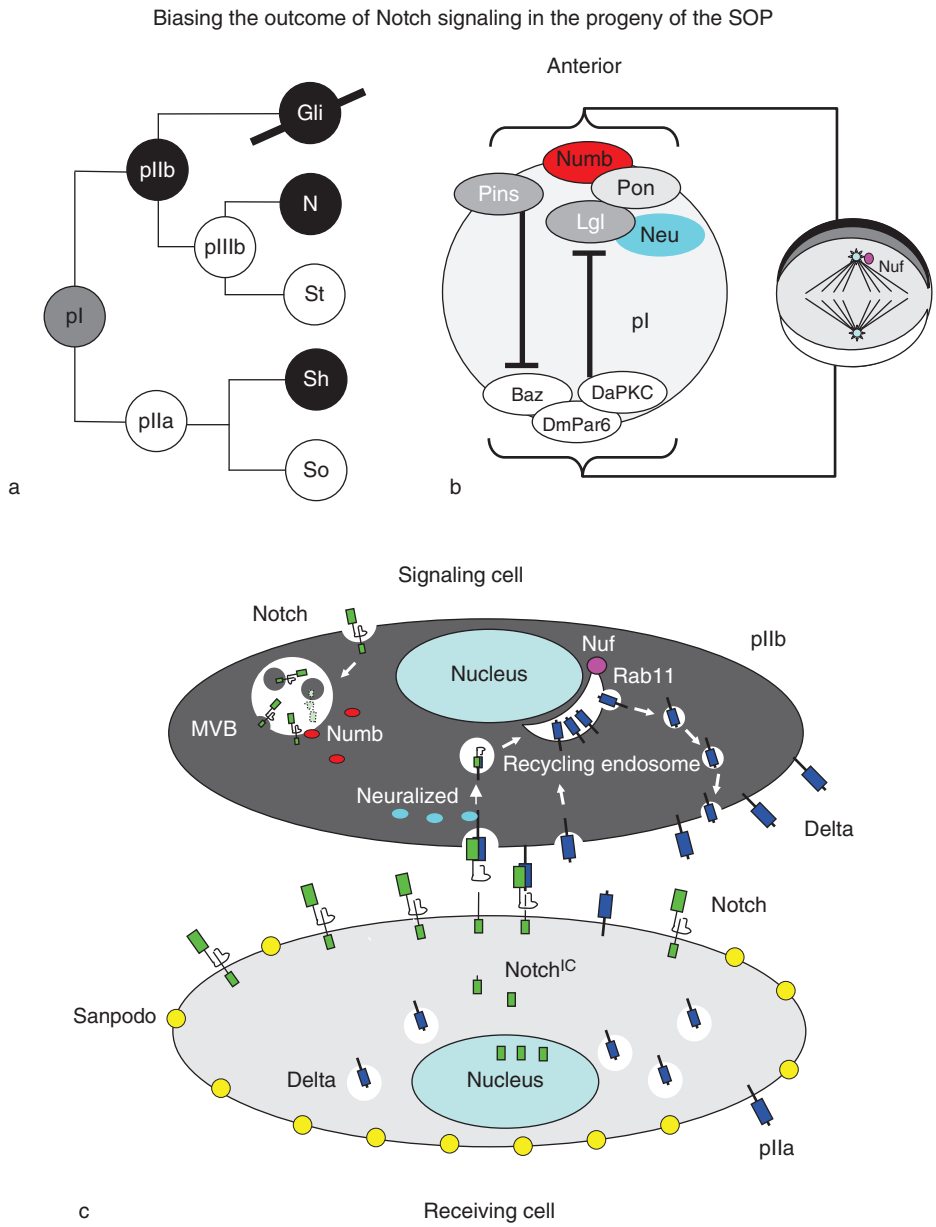

Figure 4 Mechanisms linked to planar polarity bias the outcome of Notch signaling in the progeny of the sensory organ precursor (SOP). (a) The SOP (pI) divides to form an anterior pIIb and posterior pIIa cell. The pIIa cell divides to form the Shaft (Sh) and Socket (So) cells. The pIIb cell divides to form a Glial (Gli) cell and a pIIIb cell. The pIIIb cell divides to form a Neuron (N) and a Sheath (St) cell. Cells with relatively low Notch signaling adopt the fate represented by the dark cells in the lineage. (b) At pro-metaphase, planar polarity mechanisms segregate Pins (Partner of Inscrutable) to the anterior cortex of the pI cell. Pins restricts localization of Baz to the posterior cortex of the dividing pI cell, and segregation of Baz along with DaPKC and DmPar6 at the posterior cortex restricts another factor, Lethal Giant Larvae (Lgl), to the anterior cortex. Lgl in turn recruits Partner of Numb (Pon) and Neuralized (Neu) to the anterior cortex. (c) Numb and Neuralized are segregated to the anterior pIIb cell. In addition, the Rab11 binding protein, Nuclear Fallout (Nuf), is permitted to associate with the centrosome in the pIIb cell. Numb prevents membrane localization of Sanpodo, a requirement for Notch function, and Numb also inhibits Notch in the pIIb cell by promoting its degradation in the MVB. Nuf facilitates organization of a Rab11-dependent recycling center in the pIIb cell, which makes Delta available at the cell surface. Neuralized allows the pIIb cell to internalize Delta and activate Notch in its neighbor. In contrast, absence of Numb and plasma membrane localization of Sanpodo makes Notch effective in the pIIa cell, while inhibition of Nuf association with the centrosome and absence of Neuralized makes Delta less effective at delivering a signal to the adjacent pIIb cell. MVB, multivesicular body; $Notch^{IC}$, intracellular fragment of Notch receptor.

of the Rab11-binding protein Nuclear Fallout, the *Drosophila* homolog of vertebrate Arfophilin, on the centrosome. Association of Nuclear Fallout with the centrosome is inhibited in the pIIa but allowed in the pIIb cell. As a consequence, soon after division, the pIIb cell begins recycling Delta internalized before or during mitosis, quickly establishing itself as the dominant signaling cell.

The fate of siblings following asymmetric division of the neuroblasts in the *Drosophila* CNS is also biased by asymmetric segregation of cell fate determinants like Numb. In this context, however, the asymmetric localization of cell fate determinants is not determined by planar polarity mechanisms but by asymmetries associated with apical–basal polarity in an epithelial cell. Division of the neuroblast, perpendicular to the apical–basal axis, creates a smaller basal sibling that inherits Numb. Numb inhibits Notch function and permits this cell to adopt a GMC fate. In contrast, higher Notch activation in the larger apical sibling allows that cell to retain a neuroblast fate.

Notch Signaling in Vertebrate Neurogenesis

Homologs of genes involved in Notch signaling have now been identified in almost all metazoans, and analysis of their function in vertebrate model systems including chick, zebra fish, mouse, and rat has demonstrated that Notch has a conserved role in vertebrate neurogenesis. Similarities have been shown in many contexts; two examples, from *Xenopus* and zebra fish, illustrate some of the similarities.

Early differentiating neurons, also called primary neurons in *Xenopus* and zebra fish, are distributed in three bilateral longitudinal columns in the neuroepithelium of the caudal neural plate. Within each longitudinal column, only a subset of cells become early neurons. Exaggerated Notch signaling reduces the number of neurons. In contrast, loss of Notch signaling in this context results in too many early differentiating neurons, a loss of many of the late differentiating 'secondary' neurons, and loss of the progenitor population that would normally permit neurogenesis to continue during development. In the case of primary sensory neurons, called Rohon Beard cells, Notch signaling helps neuronal progenitors choose between a neuronal fate and a neural crest fate. Loss of Notch signaling allows too many cells to become sensory neurons and not enough progenitors adopt a neural crest fate. In homozygous zebra fish Mindbomb^{ta52b} alleles, severe loss of neural crest cells results in a loss of pigment cells in the tail, which is why this allele was originally called 'white tail.' These observations in *Xenopus* and zebra fish illustrate how Delta–Notch-mediated lateral inhibition drives selection of a subset of cells that become neurons. Like neuroblasts in *Drosophila* that delaminate from the neuroectoderm, cells that are selected to become neurons release their apical or ventricular attachments, delaminate from the neuroepithelium, and eventually differentiate under the pial surface of the neural tube. Subsequent studies have shown that expression of an *atonal*-related gene, *neurogenin* (*ngn1*), helps define the proneuronal domains and gives cells the potential to become neurons. Notch signaling drives expression of Hairy–E(spl)-related genes (called ESR genes in *Xenopus*, HER genes in zebra fish, and HES genes in most other vertebrates), and Ngn1 expression is prevented by Notch signaling. If cells express high enough levels of Ngn1, they can initiate the expression of additional genes that facilitate stable acquisition of neuronal fate. For example, Ngn1 drives expression of *neuroD*, a bHLH proneuronal gene whose function is not as easily inhibited by Notch signaling. Ngn1 also initiates MyT1 expression, a factor that makes selected cells less sensitive to Notch signaling. In addition, P27^{xic1} stabilizes Ngn1 and facilitates exit from the cell cycle. Despite the obvious similarities, there are important differences in the early roles described for Delta–Notch signaling in neurogenesis: In *Drosophila*, Notch signaling selects neuroblasts that delaminate and come together to form the neural anlage, whereas in *Xenopus* and zebra fish, Notch signaling plays a role in the selection of cells that become the earliest differentiating neurons within the neural anlage (neural plate).

There are four Delta homologs in zebra fish (*deltaA*, *deltaB*, *deltaC*, and *deltaD*). Of these, *deltaA* and *deltaD* are expressed in proneuronal domains during early neurogenesis. As cells are selected to become neurons, they express higher levels of *deltaA*, and expression becomes less in surrounding cells, reflecting dynamic regulation that is consistent with the process of lateral inhibition. Though Delta–Notch signaling is likely to play a role in lateral inhibition and cells are selected in a salt-and-pepper pattern from longitudinal columns of *ngn1*-expressing cells, it is not clear whether the cells form small contiguous *ngn1*-expressing clusters from which individual cells are selected or whether early patterning mechanisms simply define a long proneuronal domain from which cells are selected in salt-and-pepper pattern by lateral inhibition. Lineage analysis shows that Notch signaling clearly regulates binary cell fate decisions and that Notch signaling has a recurring role in regulating both early and late neurogenesis. However, the precise role of individual Delta and Notch homologs in lateral inhibition and binary cell fate decisions is only beginning to be defined, and it remains unclear what mechanisms bias selection by lateral inhibition.

In the zebra fish, Delta–Notch signaling also helps select sensory hair cells in the ear and in neuromasts. Neuromasts, initially deposited by the caudally migrating posterior lateral line promordium (pllp), contain a central hair cell that is surrounded by

supporting cells. Analysis of Notch–Delta function shows that it helps select a central hair cell within a proneural cluster defined by *atonal homolog1* expression. The migrating pllp contains three or four nascent neuromasts at various stages of maturation. The least mature are at the leading edge of the pllp, where *notch3* is diffusely expressed within a cluster of cells representing the future neuromast. As the neuromast matures, *atonal homolog1*and *deltaA* expression becomes prominent in a central cell, and *notch3* expression gets restricted to surrounding cells. Following deposition, neuromasts mature to produce a cluster of hair cells, following a pattern of development reminiscent of *Drosophila* chordotonal organs. Further analysis of neuromast development is likely to provide details about how the proneural cluster is defined, how biasing mechanisms lead to selection of a central cell in this context, and how additional hair cells are added to the cluster. It will be interesting to compare how a mechanism that operates in neuromasts compares with those defined for PNCs where SOPs are selected.

Additional Roles for Notch Signaling

Examples from *Drosophila* development illustrate how Notch signaling feeds into genetic networks that determine cell fate and how the location of a cell and interactions with additional signaling mechanisms can influence the outcome of lateral inhibition. These examples provide a useful framework for understanding how similar cell fate decisions are made during neurogenesis in other metazoan systems. Mechanisms that regulate protein trafficking and cell biology have a critical role in Notch signaling.

While the role of Notch signaling in lateral inhibition during neurogenesis has been emphasized, it should be kept in mind that Notch has diverse roles in all tissues. In addition to its role in lateral inhibition and in binary cell fate decisions, where it mediates competitive interactions and prevents adjacent cells from adopting the same fate, Notch signaling is also involved in inductive interactions where expression of Notch ligand in one cell is required to activate Notch and induce a specific fate in an adjacent cell. In ddition, instead of participating in a negative feedback loop, Notch activation can be part of a positive feedback loop in which mutual activation of Notch in adjacent cells helps maintain high Notch signaling and determines similar fate in adjacent cells. At the wing margin in *Drosophila* and at rhombomere boundaries in the zebra fish hindbrain, sustained Notch signaling is linked to expression of Wingless/Wnt and helps establish signaling centers at tissue compartment boundaries. During somitogenesis, delay between transcriptional regulation of gene expression and feedback inhibition by the encoded proteins results in oscillating gene expression; furthermore, since cell–cell communication contributes to the mechanism, the oscillations are synchronized between cells, and the synchronized oscillations are used to establish the periodic pattern of somites. Finally, while Notch can influence tissue patterning through its influence on cell fate at a transcriptional level, it may simultaneously influence the organization of factors that determine cell morphology. Understanding the diverse yet conserved roles of Notch in cell fate and tissue morphogenesis remains an important and exciting challenge for the future.

See also: Helix–Loop–Helix (bHLH) Proteins: Proneural; Helix–Loop–Helix (bHLH) Proteins: Hes Family; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Notch Signal Transduction: Molecular and Cellular Mechanisms.

Further Reading

Bardin AJ and Schweisguth F (2006)) Bearded family members inhibit Neuralized-mediated endocytosis and signaling activity of Delta in Drosophila. *Developmental Cell* 10(2): 245–255.

Bray SJ (2006) Notch signalling: A simple pathway becomes complex. *Nature Reviews Molecular Cell Biology* 7(9): 678–689.

Castro B, Barolo S, Bailey AM, and Posakony JW (2005) Lateral inhibition in proneural clusters: Cis-regulatory logic and default repression by Suppressor of Hairless. *Development* 132(15): 3333–3344.

Culi J, Martin-Blanco E, and Modolell J (2001) The EGF receptor and N signalling pathways act antagonistically in Drosophila mesothorax bristle patterning. *Development* 128(2): 299–308.

Emery G, Hutterer A, Berdnik D, et al. (2005) Asymmetric Rab 11 endosomes regulate delta recycling and specify cell fate in the Drosophila nervous system. *Cell* 122(5): 763–773.

Jafar-Nejad H, Andrews HK, Acar M, et al. (2005) Sec15, a component of the exocyst, promotes notch signaling during the asymmetric division of Drosophila sensory organ precursors. *Developmental Cell* 9(3): 351–363.

Langevin J, LeBorgne R, Rosenfeld F, Gho M, Schweisguth F, and Bellaiche Y (2005) Lethal giant larvae controls the localization of notch-signaling regulators Numb, Neuralized, and Sanpodo in Drosophila sensory-organ precursor cells. *Current Biology* 15(10): 955–962.

Le Borgne R, Bardin A, and Schweisguth F (2005) The roles of receptor and ligand endocytosis in regulating Notch signaling. *Development* 132(8): 1751–1762.

zur Lage P and Jarman AP (1999) Antagonism of EGFR and notch signalling in the reiterative recruitment of Drosophila adult chordotonal sense organ precursors. *Development* 126(14): 3149–3157.

Notch Signal Transduction: Molecular and Cellular Mechanisms

A Miyamoto and G Weinmaster, David Geffen School of Medicine at the University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Pathway Overview

The Notch signaling pathway is one of a handful of signaling systems known to regulate the myriad of developmental decisions required for correct embryonic patterning and morphogenesis in metazoans. Consistent with this, Notch signaling affects the determination and differentiation of many different neural cell types in addition to regulating cell survival, adhesion, proliferation, and apoptosis. The activation of Notch by membrane-bound DSL (Delta, Serrate, LAG-2) ligands serves to limit signaling to cells in direct contact. Ligand binding leads to proteolytic release of the Notch intracellular domain (NICD) from the membrane, facilitating its role as a downstream signal transducer. The majority of signaling induced by Notch involves trafficking of NICD to the nucleus, where it interacts with a DNA-binding protein CSL (CBF1 in mammals, Su(H) in *Drosophila*, and LAG-1 in *Caenorhabditis elegans*) to recruit coactivators and turn on expression of target genes such as Hes-1 (hairy and enhancer of split-1) (Figure 1).

This mode of activation necessarily means that only one signal can be sent from each receptor molecule, and significantly, once a receptor is activated it cannot be reactivated. Given the direct nature of the signaling pathway, the core components are refreshingly few in number: ligand, receptor, and a DNA-binding downstream effector. However, this simplicity is balanced by a large, and still growing, number of requisite modulators that regulate this pathway.

Notch Receptors and DSL Ligands

The prototypic Notch receptor, *Drosophila* Notch (dNotch), is a type 1 transmembrane, cell surface protein with an ectodomain composed mainly of tandemly arrayed motifs related to epidermal growth factor (EGF-like repeats) (Figure 2(a)). Structure–function analyses indicate that the extracellular domain functions in ligand binding and repression of the signaling activity that is intrinsic to the intracellular domain. The Notch-related *Caenorhabditis elegans* proteins, GLP-1 and LIN-12, and the four mammalian Notch proteins, Notch1–4, all exhibit the same overall structure. Notch-related genes have been identified from a number of metazoans from sea urchin to human, but they are not present in plant genomes.

The DSL ligands that bind and activate the Notch/LIN-12/GLP-1 receptors comprise a family of proteins that are either Delta-like or Serrate-like (also known as Jagged) based on the structure of the two *Drosophila* ligands, Delta and Serrate (Figure 2(b)). In mammals, three Delta-like (Dll) genes have been identified (Dll-1, -3, and -4). Dll-2 has only been identified in *Xenopus* and four Dll genes have been reported for zebra fish, DeltaA–D, of which DeltaD is most similar to Dll-1. Only two Serrate-like subtypes (Serrate1/Jagged-1 and Serrate2/Jagged-2) have been isolated from humans, rats, mice, chickens, and frogs. There are three zebra fish Jagged genes (Jagged-1a, Jagged-1b, and Jagged-2).

Like the Notch receptors, the DSL ligands are single-pass cell surface proteins containing multiple EGF-like repeats; however, these ligands contain a signature motif dubbed DSL, which together with N-terminal (NT) sequences constitutes the ligand-binding domain. Although membrane attachment is thought to be important for activation of Notch, soluble DSL ligands have been found in worms, flies, and mammals. In fact, a search of the complete worm genome identified 10 DSL-containing ligands, with only five sequences predicting a transmembrane domain. In *Drosophila*, no soluble splice variants of Delta or Serrate have been reported; however, soluble forms of Delta, produced through cell surface proteolytic shedding, have been identified from embryos and cultured cells. The function of cleaved soluble DSL ligands in *Drosophila* is unclear, although when engineered soluble forms of Delta or Serrate are expressed in *Drosophila*, they produce phenotypes indicative of a loss in Notch signaling, suggesting that soluble ligands inhibit rather than activate Notch. Both Delta-like (Delta, Dll-1, and Dll-4) and Serrate-like (Jagged-1 and Jagged-2) ligands can be proteolytically processed by metalloproteases to effect extracellular domain shedding. The ability of soluble DSL ligands to either activate or inhibit signaling depends on their multimeric state; therefore, it is possible that the activity of soluble ligands may be regulated through interactions with the extracellular matrix as found for soluble growth factors. Alternatively, ligand shedding could represent a mechanism for downregulating activating ligand.

DSL ligands can also function outside of their role as activators of Notch signaling through interactions with the cytoskeleton via binding to PDZ-containing scaffolding proteins. PDZ-interaction motifs located in the C-terminus of DSL ligands can mediate interactions with MAGUK (membrane-associated guanylate

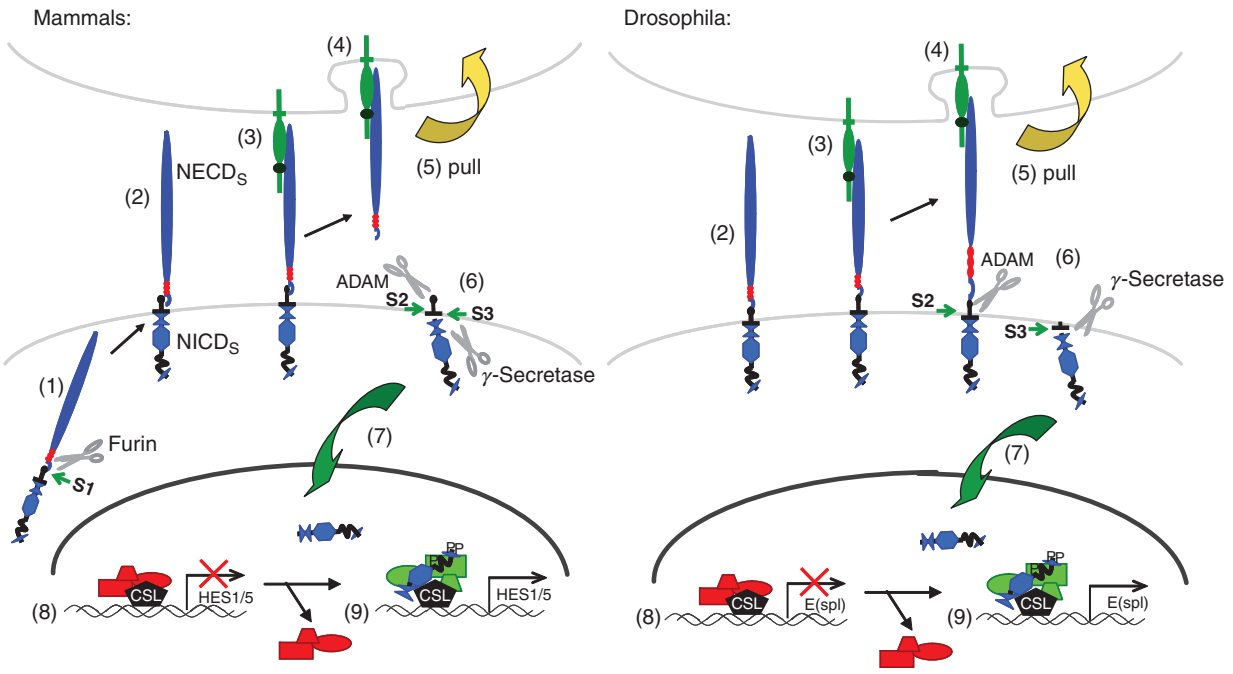

Figure 1 Notch signaling. In mammals (left), during trafficking to the cell surface mNotch is proteolytically processed by furin at the S1 site (1) and exists at the cell surface in a heterodimeric form (2). After ligand binding (3), DSL endocytosis of bound mNotch (4) is proposed to generate a 'pull' (arrow) that results in heterodimer dissociation (5) allowing removal of inhibitory sequences, including the LNR (red circles), that prevent ADAM cleavage. Following uptake of NECD$_s$ by the DSL cell, the remaining NICD$_s$ would then be exposed to activating proteolysis (6), first by ADAM to generate the S2 fragment that is subsequently cleaved by γ-secretase to produce the soluble NICD S3 fragment that translocates to the nucleus (7). Note that although mNotch S2 and S3 cleavages are shown here at the cell surface, it is also possible that either or both of these cleavages occur in an endosomal compartment. In the absence of Notch signaling, CSL recruits corepressors to inhibit gene transcription (8). When NICD is present, it displaces the corepressors and recruits coactivators to promote transcription of target genes such as Hes-1 or Hes-5 (9). In flies (right), there is evidence that cell surface dNotch is uncleaved (2). After ligand binding (3), the pull induced by DSL endocytosis (4 and 5) induces a conformational change that is thought to unmask the S2 site and allow for ADAM cleavage (6). This suggests that, unlike mammals, ligand-induced NECD release may be enzymatic, and not mechanical, in nature. Subsequent activating cleavage by γ-secretase (6) produces NICD that acts similarly to mammalian NICD in the nucleus to activate target genes such as Enhancer of split (E(spl)) (7–9).

kinase) family proteins. Mutation or deletion of the PDZ-binding domain in DeltaD, Dll-1, or Jagged-1 does not perturb Notch signaling, but rather produces phenotypes consistent with cytoskeletal involvement, such as aberrant cell migration or cell shape changes.

Regulation of Ligand-Induced Notch Signaling

Regulation of Ligand Expression

Cells take on distinct fates because Notch signaling is consistently activated in only one of the two interacting cells, highlighting the importance of establishing and maintaining signaling polarity. Studies in flies and worms have identified positive and negative transcriptional feedback mechanisms that amplify small differences in Notch and DSL ligand expression that bias which cells send or receive signals. Notch signaling can regulate gene expression of both the receptor and the ligand, and signaling from certain growth factor receptors (transforming growth factor-β (TGF-β) and vascular endothelial growth factor) also enhances DSL ligand expression. Other signaling pathways shown to potentiate DSL ligand expression include EGFR/MAP kinase, hormones, and NF-κB activation. Underscoring the importance of Notch and DSL ligand expression levels is the sensitivity to gene dosage reported for Notch-dependent processes, where both losses and gains in Notch activity produce mutant phenotypes. In humans, haploinsufficiency of either Jagged1 or Notch2 is associated with Alagille syndrome, whereas Notch1 haploinsufficiency is implicated in a subtype of inherited aortic disease.

Regulation of DSL Ligand Activity

The ability of cells to signal to Notch on adjacent cells is also regulated by endocytosis and membrane trafficking, consistent with the long-appreciated genetic interactions between Notch and shibire, which is the *Drosophila* homolog of dynamin, a key regulator of

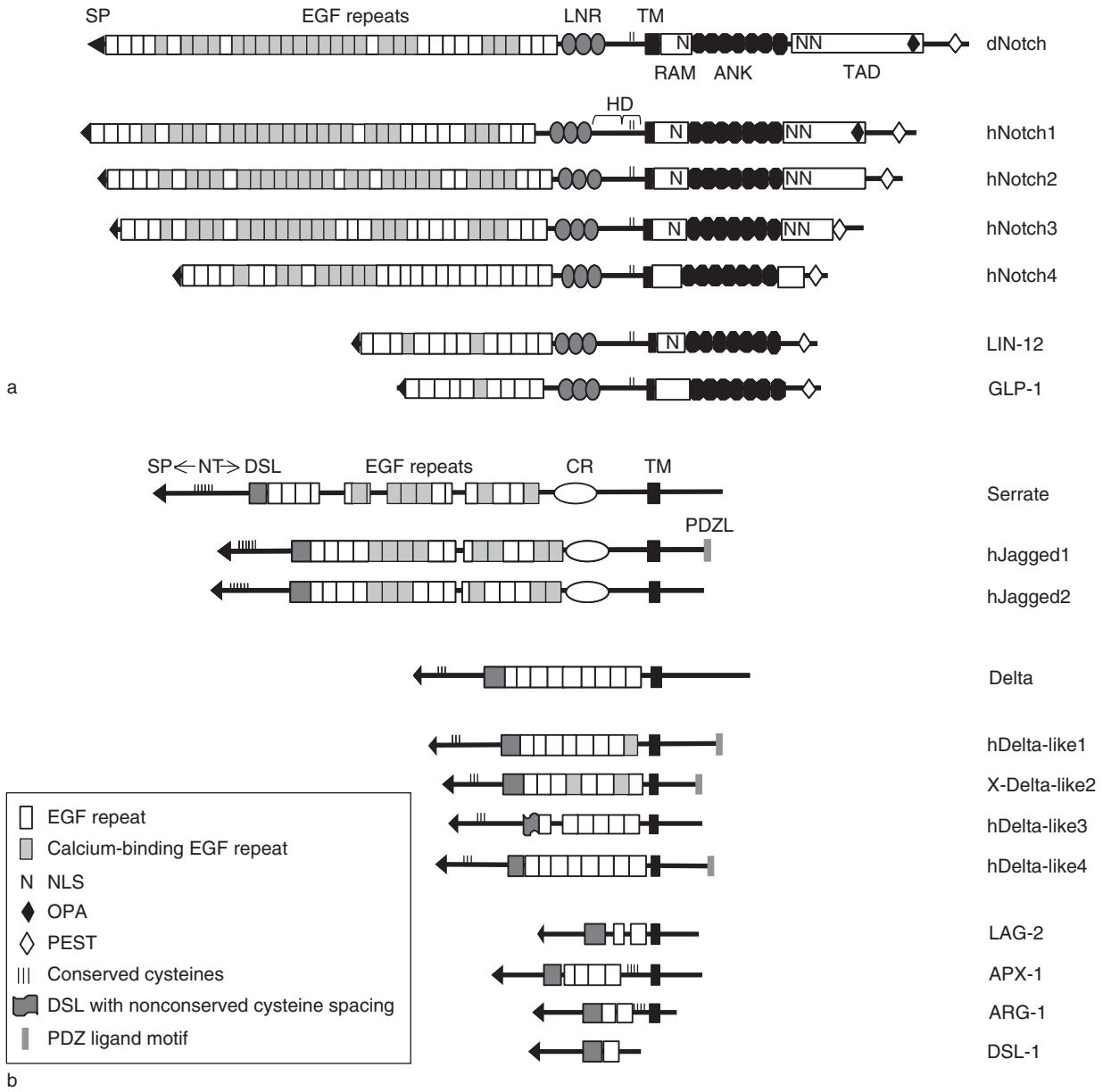

Figure 2 Notch receptors and DSL ligands. (a) The schematic is based on the following protein sequences from GenBank: *Drosophila* Notch, P07207; human Notch1, P46531; human Notch2, Q04721; human Notch3, Q9UM47; human Notch4, Q99466; *C. elegans* LIN-12, P14585; *C. elegans* GLP-1, P13508. (b) Structures are based on the following protein sequences from GenBank: *Drosophila* Serrate, P18168; human Jagged1, XP056118; human Jagged2, Q9Y219; *Drosophila* Delta, P10041; human Delta1, O00548; *Xenopus* X-Delta-2, AAB37131; human Delta3, Q9NYJ7; human Delta4, Q9NR61; *C. elegans* LAG-2, P45442; *C. elegans* APX-1, P41990; *C. elegans* ARG-1, T16213; *C. elegans* DSL-1, AAC04450. The drawings are approximately to scale. ANK, ankyrin-like repeats; CR, cysteine-rich region; DSL, Delta/Serrate/LAG-2 domain; EGF, epidermal growth factor-like; HD, heterodimerization domain; LNR, LIN-12/Notch related repeat; NLS, nuclear localization sequence; NT, N-terminal domain; OPA, glutamine-rich region; PDZL; PDZ ligand motif; PEST, motif rich in proline, glutamine, serine, and threonine residues that is important in protein turnover; RAM, CSL interaction domain; SP, signal peptide; TAD, transactivation domain; TM, transmembrane domain.

endocytosis. Dynamin is required in both signal-sending and signal-receiving cells; however, the exact functions for endocytosis in productive DSL-induced Notch signaling are unclear. Identification of additional endocytosis and membrane trafficking proteins that function in Notch signaling (Table 1) has provided some insight.

DSL endocytosis is required for Notch activation Interestingly, DSL ligands defective in endocytosis accumulate at the cell surface yet are unable to activate Notch, suggesting that cell surface ligand alone is not sufficient to initiate signaling. This is consistent with the dynamin requirement in signal-sending cells, and other endocytic proteins are required for DSL

Table 1 Endocytosis and trafficking regulation of Notch signaling

Drosophila	*C. elegans*	*Mammals*	*Role in Notch signaling*
In the signal-sending cell			
E3 ubiquitin ligases			
Neuralized	F10D7.5[a]	NEURL, NEURL2	RING-type E3 ligase that binds and ubiquitinates DSL ligands and promotes endocytosis
Mind bomb (dMib)		MIB1, 2	RING-type E3 ligase that binds and ubiquitinates DSL ligands and promotes endocytosis
Clathrin-mediated endocytosis			
Liquid facets (lqf)	Epn-1	Epsin 1, 2	Clathrin-associated sorting protein required for Delta signaling activity
Shibire	dyn-1	Dynamin1, 2	Endocytosis of ligands from PM; required for NECD *trans*-endocytosis and possible recycling
Clathrin heavy	Chc-1	CLTC	Coat protein of endocytic pits and vesicles; enhances Lqf phenotype
Auxilin	Dnj-25	Auxilin,GAK	J-domain containing protein proposed to function in CCV formation and/or uncoating; enhances Lqf
Eps-15	Ehs-1	Eps15	Required for *trans*-endocytosis and Notch activation
RalA	Ral-1	RalA	GTPase functioning in endocytosis, recycling, and actin organization; enhances Lqf phenotype
Endosomal recycling			
Rab 11	Rab-11.1	Rab 11A	Asymmetrically distributed in signal-sending cell that may promote DSL recycling
Sec 15	C28G1.3[a]	Sec15	Component of the exocyst that may function in Delta recycling
In the signal-receiving cell			
E3 ubiquitin ligases			
Su(dx)	WWP1	Itch (Mm)/AIP4 (Hs), WWP 1, 2	Nedd-type E3 ligase that binds and ubiquitinates Notch; endosomal sorting and degradation of Notch, Deltex
Dnedd4	Y92H12A.2[a]	NEDD4-1,-2 (Nedd4l)	Nedd-type E3 ligase that binds and ubiquitinates Notch; endosomal sorting and degradation of Notch
D-Smurf	Ce33003[a]	hSmurf1,2	Nedd-type E3 ligase with partial genetic redundancy with Su(dx) in *Drosophila*
Cbl	sli-1	Cbl	E3 ligase that interacts with Notch; Notch lysosomal degradation
Deltex		DTX1-4	RING-type E3 ligase involved in endosomal sorting of Notch; ubiquitination(?) and proteosomal degradation of Notch involving Kurtz; effector of ligand- and CSL-independent signaling(?); positive modulator of Notch in *Drosophila*
Endocytosis/early endosome sorting			
Shibire	dyn-1	dynamin 1, 2	Endocytosis of Notch from the PM; required for S3 cleavage of Notch to produce NICD
Merlin/Expanded		NF2	FERM protein that clears Notch from the PM; negative modulator of Notch signaling
Avalanche		syntaxin7, 12	Early endosomal SNARE protein involved in Notch degradation and endocytosis; no genetic interaction with Notch
Rab5	rab-5	rab5A,C	Early endosome trafficking protein involved in Notch degradation and endocytosis; no genetic interaction with Notch
Hrs	hgrs-1	Hrs (HGS)	Directs ubiquitnylated cargo to ESCRT machinery; no genetic interaction with Notch except in ligand-independent Notch activity induced by LOF Lethal giant discs
Multivesicular body formation			
Erupted	C09G12.9[a]	tsg 101	ESCRT-I protein that restricts ligand-independent Notch activation; MVB sorting of Notch for degradation
Vps25	WO2A11.2[a]	Vps25	ESCRT-II protein that restricts ligand-independent Notch activation; MVB sorting of Notch for degradation
Lethal giant discs	Y37H9A.3[a]	CC2D1A,B	Restricts ligand-independent Notch activation; functions in endosomal sorting after Hrs
Other proteins with roles in endosomal sorting			
Kurtz	F53H8.2[a]	ARRB1, 2	Nonvisual β-arrestin that binds Deltex; proteosomal degradation of Notch that involves Deltex

[a]Homology by protein sequence only.

Where no gene is listed, no ortholog has been identified. CCV, clathrin-coated vesicle; CSL, CBF1/Suppressor of Hairless/LAG-1 Notch downstream effector; Eps, epidermal growth factor receptor pathway substrate; ESCRT, endosomal sorting complexes required for transport; FERM, 4.1 protein/ezrin/radixin/moesin; hrs, hepatocyte growth factor receptor tryosine kinase substrate; Hs, human; LOF, loss of function; Mm, mouse; MVB, multivesicular body; NECD, Notch extracellular domain; Nedd, neural precursor cell expressed, developmentally downregulated; NICD, Notch intracellular domain; PM, plasma membrane; RING, really interesting new gene; Su, Suppressor; tsg, tumor susceptibility gene; Vps, vascuolar protein sorting.

activity (Table 1). For example, the E3 ubiquitin ligases Neuralized and Mind bomb ubiquitinate the intracellular domains of invertebrate and vertebrate DSL ligands and promote their endocytosis. DSL ligands must be ubiquitinated to signal, and only ubiquitinated DSL ligands in the presence of Liquid facets, the sole *Drosophila* epsin-related clathrin-accessory protein, are competent to activate Notch on adjacent cells. Signal-sending cells also require additional proteins that function in clathrin-mediated endocytosis, such as clathrin, dynamin, auxilin, and Eps15, for DSL ligands to signal effectively. Together, these findings highlight the importance of DSL endocytosis in Notch activation, but the exact function is unclear and remains controversial.

DSL endocytosis functions in recycling The requirement for epsin-dependent endocytosis in the signal-sending cell has been suggested to reflect a role for epsin in promoting DSL recycling to produce an 'active' ligand through clustering, posttranslational modification, or maintaining high levels at the cell surface; however, direct demonstration that DSL ligands actually recycle is lacking.

DSL ligand recycling could produce strong signal-sending cells through asymmetric localization of Neuralized that potentiates ligand activity by promoting ubiquitin–epsin-dependent endocytosis and trafficking of DSL ligand. In addition, asymmetric partitioning of the recycling endosome has also been proposed to regulate Delta levels or activity by recycling Delta to the cell surface where it can activate Notch. The differential recycling of Delta has suggested a mechanism by which ligand could be concentrated in signal-sending cells. Consistent with a role for DSL recycling in Notch activation, losses in Sec15, which functions with Rab11 in trafficking proteins from the recycling endosome to the plasma membrane, also produce cell fate transformations indicative of defects in Notch signaling. Interestingly, the mechanism responsible for the apparent polarized distribution of recycling endosomes is different from that required for asymmetric segregation of Neuralized, indicating that there are multiple mechanisms operating to establish and maintain signal-sending potential.

That only some Notch-dependent processes require Sec15 activity is inconsistent with DSL recycling producing an 'active' ligand because all Notch signaling should require this. Nonetheless, Dll1 recycling defects in bone marrow stromal cells suppress Notch-dependent phenotypic changes in lymphoid progenitors, suggesting that DSL recycling is necessary to produce an active ligand. However, the reported accumulation of Dll1 to recycling endosomes could decrease Dll1 at the cell surface, thereby accounting for the losses in Notch signaling. In this case, the need for Dll1 to pass through the recycling endosome could reflect a mechanism to replenish cell surface ligand, especially given that Dll1 density determines the lymphoid fate generated from hematopoietic stem cells.

DSL endocytosis functions in proteolytic activation of Notch DSL endocytosis could function beyond presentation of cell surface ligand for Notch engagement, such as proteolytic activation of Notch to generate the NICD that functions as the downstream signal transducer (Figure 1). DSL binding induces a series of proteolytic cleavages in Notch, first at the extracellular ADAM site followed by γ-secretase cleavage within the membrane to release the NICD required for signaling. Although DSL is required to produce NICD from full-length Notch, it is unclear how ligand binding promotes ADAM cleavage, a necessary step in activating γ-secretase proteolysis. Studies of flies have suggested that endocytosis of DSL bound to Notch on adjacent cells (designated transendocytosis) might produce a molecular strain in Notch leading to conformational changes that facilitate ADAM cleavage. Since efficient γ-secretase cleavage depends on removal of the Notch extracellular domain (NECD), DSL endocytosis could also promote Notch signaling by clearing the ADAM-shed NECD through uptake by ligand cells.

DSL endocytosis promotes physical dissociation of Notch Findings with mammalian cells support a role for NECD transendocytosis by DSL ligand cells in Notch signaling but indicate that NECD is not proteolytically released. Rather, NECD transendocytosis by DSL ligand cells in the presence of ADAM inhibitors suggests that NECD is physically removed from Notch present on the surface of interacting cells. To understand this result, it is important to note that most of the mammalian Notch (mNotch) at the plasma membrane is an intramolecular heterodimer, formed through furin-proteolytic processing during transport to the cell surface (Figure 1). The NECD and membrane-bound NICD heterodimeric subunits remain associated following furin cleavage through noncovalent interactions that keep Notch intact and inactive. Although mNotch activation relies on heterodimeric formation, the significance of this processing event has remained controversial given that *Drosophila* Notch (dNotch) does not appear to require furin processing to signal. However, an unprocessed mNotch receptor is neither dissociated nor activated by DSL cells. Since DSL binding in the absence of endocytosis does not promote heterodimer dissociation, forces created during ligand endocytosis are proposed to remove

the NECD subunit ($NECD_s$) and expose the NICD subunit ($NICD_s$) to activating proteolysis (Figure 1). In support of this idea, a majority of activating Notch mutations in T cell acute lymphocytic leukemia are missense mutations that alter heterodimer stability and effect dissociation independent of ligand.

Interestingly, amino acid insertions near the ADAM cleavage site do not alter heterodimer stability but are nonetheless constitutively active. The inserted residues are proposed to expose the ADAM cleavage site within the intact heterodimer, which may be similar to how DSL binding activates dNotch. Specifically, there may be a requirement to remove or distance the inhibitory LNR domain of Notch, located just upstream of the HD (Figure 2), for S2 cleavage to occur. In fact, the LNR domain has been shown to be necessary for inhibiting S2 cleavage but not for maintaining the heterodimeric structure of Notch1. Together, these data suggest that DSL binding followed by endocytosis activates mNotch through heterodimer dissociation, whereas conformational changes that facilitate ADAM proteolysis of unprocessed dNotch may underlie its activation.

DSL epsin-dependent endocytosis generates a pulling force Since only ligands internalized by epsin appear competent to signal, what role could epsin play in Notch activation? It is possible that Notch binding to DSL could induce ligand clustering, which would amass ubiquitin binding sites for epsin via its ubiquitin interacting motif (UIM) repeats. By assembling multiple low-affinity UIM–mono-ubiquitin interactions, strong epsin–ubiquitinated DSL interactions could be generated, which may be necessary to overcome any resistance to DSL internalization when it is bound to Notch. In fact, replacement of the Delta intracellular domain with a single ubiquitin motif that can be further ubiquitinated allows internalization and signaling activity in zebra fish. However, a nonextendable ubiquitin only weakly signals in flies even though it promotes endocytosis, supporting the idea that multiple ubiquitin interaction sites are required for DSL to activate Notch.

Epsin is a multidomain protein that in addition to binding ubiquitinated cargo interacts with membrane phospholipids, clathrin, and other clathrin adaptors as well as Cdc42 GTPase-activating proteins that regulate actin dynamics. Together with reports that the actin cytoskeleton and dynamin produce membrane constriction and tension during the process of endocytosis, it is tempting to speculate that epsin is required for ligand activity to create an endocytic vesicle endowed with sufficient force to induce conformational changes that either physically pull the mNotch heterodimer apart or promote ADAM cleavage of intact dNotch. This may explain why ubiquitinated ligands internalized in cells lacking epsin are unable to signal and why only DSL ligands internalized in an epsin-dependent manner are competent to signal.

Mechanotransduction in DSL-Induced Notch Signaling

That endocytosis, rather than proteolysis, drives mNotch dissociation suggests a mechanism of activation that involves mechanotransduction to allow activating proteolysis to occur, rather than relying completely on proteolytic cleavage as proposed for dNotch. However, if an intact dNotch can be activated, why is heterodimeric formation required for activation of mNotch? Activation of a heterodimeric receptor through physical dissociation might allow for additional mechanisms of Notch activation. In fact, noncanonical ligands have been reported, and in particular the extracellular matrix proteins MAGP1 and-2 (microfibril-associated glycoprotein) dissociate mNotch independent of ADAM cleavage and DSL binding that leads to proteolytic activation of Notch. Since MAGP binds and dissociates Notch cells autonomously, it might allow Notch to function as a mechanosensor to detect changes in shear forces produced by blood flow, which, like Notch signaling, is implicated in arterial–venous fate determinations.

How do soluble DSL ligands activate Notch? If DSL endocytosis of Notch triggers proteolytic activation, how can soluble ligands activate signaling? Although soluble DSL ligands activate Notch signaling, they are less active than cell-associated ligands and require clustering or attachment to surfaces to efficiently activate signaling and induce biological responses. In fact, prefixed Delta-expressing cells that are presumably endocytosis defective activate Notch target genes. Perhaps soluble ligands attached to the extracellular matrix or cell surface bind to Notch, generating pulling forces through cell detachment and/or Notch endocytosis. This mechanism may account for the signaling activity of naturally occurring soluble DSL ligands identified for *C. elegans*, in which the Notch-related GLP-1 is also heterodimeric.

Endocytosis and Trafficking Regulate Notch Levels and Activity

Signal-receiving cells regulate basal levels and activity of cell surface Notch through ubiquitin-dependent endocytosis, sorting, and trafficking. Removal and replenishment of cell surface Notch independent of ligand might influence signaling through availability

of receptor for activation. However, the requirement for dynamin in signal-receiving cells suggests a role for Notch endocytosis, in concert with DSL endocytosis, to effect conformational changes and/or receptor dissociation through pulling forces exerted on ligand–receptor complexes. In addition, endocytosis and trafficking of Notch could promote interactions with activating proteases since both ADAMs and γ-secretase are integral membrane proteins. Alternatively, endocytosis could function to release NICD from the membrane following γ-secretase cleavage at the cell surface.

Where Does DSL-Induced Proteolytic Activation of Notch Occur in Cells?

Whereas S2 and S3 cleavages of Notch are usually depicted to occur at the cell surface, ADAMs are found at the plasma membrane, early endosomes, and lysosome, so theoretically, cleavage could occur at any of these sites. Although evidence for intracellular cleavage of mNotch by ADAMs is lacking, γ-secretase proteolysis may occur inside cells since both monoubiquitination and clathrin-mediated endocytosis of Notch are required for signaling. Ubiquitination could allow endocytosis and/or trafficking that directs the ADAM/S2 cleavage fragment to an intracellular γ-secretase-rich compartment for S3 cleavage. However, the cellular site for γ-secretase cleavage of Notch is controversial and data support cleavage at the cell surface as well as within the cell. Nonetheless, γ-secretase activity is present in lysosomes and is enhanced by low pH. Given reports that endocytosis is required for γ-secretase cleavage of proteins other than Notch, the endosome is an attractive site for Notch proteolytic activation.

DSL-Independent Endocytosis and Trafficking of Notch

In the absence of ligand, Notch is also removed from the cell surface and may be either recycled or targeted for degradation. Since ubiquitin signals endocytosis, sorting and trafficking of cell surface proteins, it is not surprising that a number of E3 ubiquitin ligases are associated with Notch (Table 1).

In mammalian cells, the E3 ubiquitin ligases Cbl and Itch/AIP4 have been reported to interact with or modify Notch, but the biological relevance of these findings is unclear. Genetic studies of flies have provided evidence that ubiquitination and endosomal trafficking of Notch are important regulators of Notch activity. The Nedd E3 ubiquitin ligases bind and ubiquitinate Notch, and they may regulate Notch cell surface expression by promoting endocytosis and targeting for lysosomal degradation. Dominant-negative Nedd proteins increase Notch localization to recycling endosomes and at the cell surface, suggesting that losses in Notch ubiquitination promote its return to the cell surface, perhaps for increased exposure to ligands. However, coexpression of dominant-negative Nedd proteins and Notch in *Drosophila* activates signaling independent of ligand. In this case, suppressing Nedd function could protect a positive Notch effector from degradation. In fact, both *Drosophila* and mammalian Nedd family members ubiquitinate Deltex, a positive effector of Notch signaling, marking it for lysosomal degradation. Deltex is a RING-finger E3 ubiquitin ligase that, when overexpressed in flies, results in aberrant accumulation of dNotch in late endosomes/lysosomes and ligand-independent signaling. However, Deltex has also been shown to ubiquitinate dNotch directly for proteosomal degradation that would limit signaling. This negative effect of Deltex on dNotch requires a third protein, Kurtz, the single *Drosophila* ortholog of β-arrestin. Interactions between β-arrestin and dNotch may mediate integration of Notch with other signaling pathways given that in mammalian cells β-arrestins act as scaffolds/adaptors for signaling effectors.

Membrane trafficking defects result in aberrant accumulation of Notch within endosomes and in some cases unliganded Notch is activated (Table 1). In the early endosome, Notch is localized by the Hrs protein to the ESCRT machinery that promotes trafficking for lysosomal degradation. Defects in either ESCRT or Hrs proteins promote endosomal accumulation of Notch, but ectopic signaling is only activated with loss of ESCRT components (vps23 (erupted)/tsg101 and vps25) or a novel cytoplasmic protein, Lethal giant discs (Lgd), which is epistatic to Hrs. This ligand-independent Notch activation requires γ-secretase and further indicates that Notch can be activated intracellularly.

The ability of some, but not all, endosomal trafficking defects to activate Notch inappropriately provides insight into the intracellular locale where Notch is activated. Differences in endosomal acidity could limit γ-secretase activity to a specific compartment where Notch accumulates and the pH is low enough to promote conformational changes and/or receptor dissociation for ligand-independent activation. Proteolysis is possible since Notch activating proteases are also transmembrane proteins that could be trapped with Notch in the same compartment when membrane trafficking is blocked. During normal membrane trafficking, Notch and its associated proteases would transit quickly though the endosome, thereby avoiding aberrant activation.

Regulation of Notch Target Gene Expression

Transcriptional Complexes

Structure–function analyses have identified several regions of the intracellular domain necessary for transcriptional activation (Figure 2(a)) in addition to two nuclear localization sites that function in transport of NICD to the nucleus. The RAM domain, juxtaposed to the cytoplasmic face of the membrane-spanning region, is responsible for strong interactions with CSL, the major effector of Notch signaling. In addition to the RAM domain, seven ankyrin (ANK) repeats also interact with CSL proteins and are both necessary and sufficient for Notch transcriptional activity. Downstream of the ANK repeats is a region of high homology required for transcriptional activation (TAD), and at the C-terminus PEST sequences regulate turnover of NICD protein in the nucleus. Although all four mammalian Notch proteins are proteolytically processed to generate active NICD-like fragments, the individual TADs vary in potency, which may reflect differences in target gene activation induced by the different Notch receptors.

Genetic studies indicate that the CSL genes Su(H) and LAG-1 are required for normal Notch/LIN-12/GLP-1 function but act downstream of the receptors to positively regulate signaling (Figure 1). Typically, the CSL proteins function as transcriptional repressors; however, activation of Notch signaling in cells converts these DNA-binding proteins into transcriptional activators. Complexes containing NICD and CSL physically associate with DNAs containing CSL binding sites, and these interactions correlate with transcriptional activation of Notch target genes such as Enhancer of split (E(spl)) and the homologous vertebrate genes of the Hes and HERP family (Figure 3). In fact, induced expression of E(spl) and its homologs has been a convenient marker to monitor and identify cells undergoing Notch signal transduction, although some interpretation is needed because these genes can be regulated independently of Notch.

A crucial component of the NICD/CSL activation complex is mastermind (MAM). There are three mastermind-like proteins in mammals, MAML1–3, that *in vitro* have differing abilities to potentiate the activity of NICD/CSL complexes. Structure studies indicate that MAM binds to a composite site that consists of both the ANK repeats of Notch and a portion of CSL (Figure 3). Only the NT 80 amino acids of MAM are necessary for binding to the CSL/NICD interface and on their own function as a dominant-negative pan-Notch inhibitor, presumably by forming inactive complexes with activated NICD. This implies that the bulk of MAM binds and recruits other proteins required for activated transcription. In fact, in response to ligand, NICD, MAM, p300, CBP, and SKIP (Ski-interacting protein) are found associated with the Hes1 promoter. Since CSL binds to its cognate binding sites in the absence of Notch protein, it is constantly present on the Hes1 promoter (Figure 3).

In the absence of activated Notch, CSL transcriptional repression has been linked to a number of corepressor proteins and their associated histone deacetylases (HDACs) (Figure 3). Mammalian CSL proteins physically interact with two different corepressor complexes, one containing SMRT (silencing mediator of retinoid and thyroid hormone receptors), HDAC-1, and SHARP/MINT (msx-interacting protein), and the other consisting of corepressors CIR (CBF1-interacting corepressor), SAP30, and HDAC-2. Notch can displace the corepressors bound to CSL to activate transcription, whereas SMRT or CIR can compete with NICD for binding to CSL that results in a suppression of NICD-mediated CSL activation. As might be expected for a protein that requires cofactors to actively repress transcription, 'de-repression' resulting in transcriptional activation can also occur either in the absence of corepressors or due to removal of CSL from its DNA-binding sites. KyoT2, a LIM-only protein, may act in this manner because it can block CSL binding to DNAT and compete with activated Notch1 for binding to CSL.

Notch Target Genes Regulated by CSL

The best genetically and biochemically characterized direct targets are the E(spl)/Hes/HERP genes. These genes encode basic helix–loop–helix (bHLH) transcription factors that in response to Notch signaling repress both the expression and the function of genes that encode activating bHLH transcription factors, such as the proneural achaete–scute complex (AS-C) in *Drosophila* and the vertebrate homologs of the AS-C genes, MASH, neuroD, neurogenin, and MATH.

The vertebrate homologs of E(spl), the Hes genes, also serve as Notch effectors downstream of NICD/CSL activation. Three of the Hes genes (Hes1, -5, and -7) and all three Hes-related genes (HERP1–3) have been shown to be direct transcriptional targets of Notch. As in the fly, genetic loss of Hes genes often, but not always, leads to Notch-like phenotypes. Although the Hes/HERP family of bHLH repressors represents a major subset of Notch target genes that lead to repression of downstream targets, it has become evident that CSL binding sites are present in a number of promoters and that NICD can lead to transcriptional

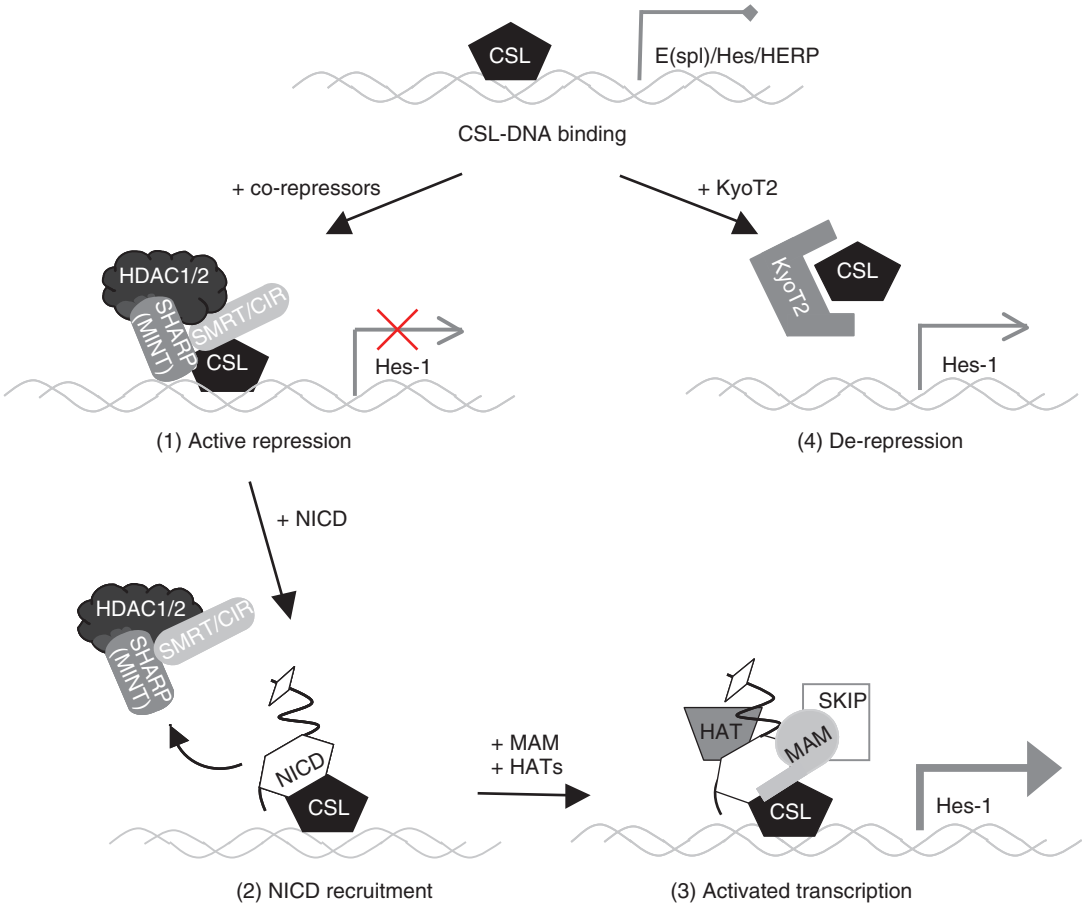

Figure 3 CSL transcriptional regulation. The DNA-binding CSL protein is common to both repression and activation complexes. In the absence of activated Notch, CSL interacts with known corepressor proteins (SMRT and SHARP(MINT)) that recruit histone deacetylases (HDAC1 or -2) to repress transcription (1). Upon Notch activation, NICD enters the nucleus and displaces the corepressor complex (2). NICD is thought to interact with CSL in a manner analogous to SMRT, although the Notch ICD and SMRT contain no identifiable sequence similarities. NICD binding to CSL creates a novel interaction site for the coactivator mastermind (MAM) that further recruits histone acetyltransferases (HAT), kinases, and elongation factors (not shown) to activate transcription (3). Loss of CSL DNA binding activity, either through interaction with KyoT2 or mutation of CSL, leads to 'de-repression' (4). In the absence of repression (and activation) complexes, some transcriptional activity is detected.

upregulation of these genes as well. The growing list includes Erbb2, glial fibrillary acidic protein, brain lipid-binding protein, GATA2, and IκBα.

Elaborations on Notch Signaling, Both Canonical and Noncanonical

The seemingly simple and direct mode of Notch signal transduction is at odds with the myriad of functions attributed to Notch, especially given the relatively small number of receptors and ligands identified for this signaling pathway. However, diversity in the output of canonical Notch signaling is most likely achieved by cross-interactions with other signaling pathways, and further diversity is achieved using noncanonical forms of Notch signaling. For example, in vascular development, the canonical Notch pathway involving NICD/CSL can cooperate with the hypoxia-inducible factors to promote blood vessel formation, perhaps by collaborating to drive expression of the Hey/HERP/Esr genes. The canonical Notch pathway can also interact with the TGF-β superfamily receptor signaling system through direct interactions between their downstream effectors. Smads activated following TGF-β binding to its cognate receptors can directly bind to NICD, leading to enhanced NICD–CSL activation of the target genes Hes/HERP. Finally, interactions have also been reported between canonical Notch signaling and the Wnt signaling pathway, but in this case the outcome is antagonistic. Cross-talk between Notch signaling and many of the major signaling pathways is likely to emerge in the future as more is learned about the global changes that occur in cells when they integrate multiple signals.

Diversity in Notch signal transduction is also achieved using pathways that diverge away from the

core one using CSL. For example, noncanonical ligands that lack DSL domains, such as DNER and F3/contactin/NB-3, have been reported to activate Notch and drive glial-specific genes such as myelin-associated glycoprotein through interactions with the downstream effector Deltex. Binding of these atypical Notch ligands leads to proteolytic activation of Notch and treatment with γ-secretase inhibitors blocks signaling. Although DNER and F3/contactin/NB-3 activation of Notch induces the translocation of NICD to the nucleus, transcriptional activation of glial-specific genes occurs independent of CSL and does not appear to involve Hes1. Diversity in Notch signaling may also be achieved using pathways that are activated independently of the transcriptional effects of NICD. In one report, for instance, DSL ligand-mediated Notch proteolysis has been proposed to induce rapid cytoplasmic responses, which involve a series of phosphorylation events on PI(3)K, Akt, and STAT3 that culminate with the translocation of phosphorylated STAT3 to the nucleus, where it stimulates Hes3 transcription. By an unknown mechanism, Hes3 leads to the production of sonic hedgehog to promote survival of neural stem cells, all independent of CSL transcriptional activation. Finally, diversity in Notch signaling may involve a pathway in which DSL ligand activates a nonheterodimeric form of Notch. Importantly, biochemical studies have documented that uncleaved, full-length Notch is endogenously expressed on the surface of a number of different mammalian cell types, and this form appears to be the major species found in *Drosophila* cells. This type of signaling has been reported to block myogenic differentiation in the absence of CSL. Notably, a full-length, furin-resistant, uncleaved Notch receptor can respond to ligand to suppress myogenesis but not activate a CSL reporter. These results indicate that an alternative form of the Notch receptor can mediate signaling which is both CSL independent and proteolysis independent.

Summary

In the past few years, there has been rapid growth in the understanding of Notch signaling mechanisms and functions. Genetically amenable model organisms that require Notch signaling have led to the identification and characterization of this essential signaling pathway. The growing number of interactions identified for Notch with other signaling pathways, as well as the isolation of additional Notch activators, likely account for the diverse and extensive nature of this highly conserved and ubiquitous signaling system. As additional components of endocytosis and membrane trafficking are linked to Notch signaling, requirement for endocytosis in Notch signaling should become clearer. What is becoming clear is that the more we investigate Notch signaling mechanisms, the more we realize just how pervasive its use is in both developmental and disease settings.

See also: Auditory System: Central Pathways; Notch Pathway: Lateral Inhibition; Retinal Color Mechanisms; Retinal Development: An Overview.

Further Reading

Androutsellis-Theotokis A, Leker RR, Soldner F, et al. (2006) Notch signalling regulates stem cell numbers *in vitro* and *in vivo*. *Nature* 442: 823–826.

Barrick D and Kopan R (2006) The Notch transcription activation complex makes its move. *Cell* 124: 883–885.

Bray SJ (2006) Notch signalling: A simple pathway becomes complex. *Nature Reviews Molecular Cell Biology* 7: 678–689.

Bush G, diSibio G, Miyamoto A, et al. (2001) Ligand-induced signaling in the absence of furin processing of Notch1. *Developmental Biology* 229: 494–502.

Eiraku M, Tohgo A, Ono K, et al. (2005) DNER acts as a neuron-specific Notch ligand during Bergmann glial development. *Nature Neuroscience* 8: 873–880.

Emery G and Knoblich J (2006) Endosome dynamics during development. *Current Opinion in Cell Biology* 18: 407–415.

Gridley T (2007) Vascular biology: Vessel guidance. *Nature* 445: 722–723.

Gustafsson M, Zheng X, Pereira T, et al. (2005) Hypoxia requires Notch signaling to maintain the undifferentiated cell state. *Developmental Cell* 9: 617–628.

Iso T, Kedes L, and Hamamori Y (2003) HES and HERP families: Multiple effectors of the Notch signaling pathway. *Journal of Cell Physiology* 194: 237–255.

Kluppel M and Wrana JL (2005) Turning it up a Notch: Cross-talk between TGFβ and Notch signaling. *BioEssays* 27: 115–118.

Le Borgne R, Bardin A, and Schweisguth F (2005) The roles of receptor and ligand endocytosis in regulating Notch signaling. *Development* 132: 1751–1762.

Louvi A and Artavanis-Tsakonas S (2006) Notch signalling in vertebrate neural development. *Nature Reviews Neuroscience* 7: 93–102.

Martinez Arias A, Zecchini V, and Brennan K (2002) CSL-independent Notch signalling: A checkpoint in cell fate decisions during development? *Current Opinion in Genetics and Development* 12: 524–533.

Nichols JT, Miyamoto A, Olsen SL, et al. (2007) DSL ligand endocytosis physically dissociates Notch1 heterodimers before activating proteolysis can occur. *Journal of Cell Biology* 176: 445–458.

Pear WS and Aster JC (2004) T cell acute lymphoblastic leukemia/lymphoma: A human cancer commonly associated with aberrant Notch1 signaling. *Current Opinion in Hematology* 11: 426–433.

NSF and SNAPs

A Morgan and R D Burgoyne, University of Liverpool, Liverpool, UK

© 2009 Elsevier Ltd. All rights reserved.

Background – NSF, SNAPs, and SNAREs

Classic work by George Palade and others in the 1960s established that secreted proteins follow a predetermined route from the endoplasmic reticulum (ER) through the various Golgi compartments to the plasma membrane. This route is known as the secretory pathway and comprises a series of transport steps, each requiring vesicle budding from the donor membrane followed by vesicle docking and fusion with the appropriate acceptor membrane. However, the molecular mechanisms underlying these budding and fusion processes were entirely unknown. In the late 1970s and early 1980s, genetic and biochemical approaches were used to shed light on this issue by studying the intracellular requirements for membrane trafficking in the secretory pathway. In the genetic approach, pioneered by Randy Schekman, yeast cells were mutagenized and then screened for mutants able to secrete at 25 °C (the permissive temperature) but not at 37 °C (the restrictive temperature). Further study revealed that these secretory (*sec*) mutants could be subdivided into 23 complementation groups, termed *sec1–23*, each representing a mutant allele of a gene required for vesicle formation or consumption at defined steps in the secretory pathway. In the biochemical approach, pioneered by James Rothman, membrane traffic through the mammalian Golgi was reconstituted *in vitro* using isolated Golgi complexes and cytosol. Subsequently, protein purification allowed the identification of several soluble proteins required for efficient intra-Golgi transport.

The first protein identified using the biochemical approach was *N*-ethylmaleimide-sensitive fusion protein (NSF), which was purified in 1988 based on its ability to reconstitute intra-Golgi transport after blockade by *N*-ethylmaleimide (NEM). NSF was soon found to be required for *in vitro* ER-to-Golgi transport and endocytic vesicle fusion, suggesting that NSF was required for multiple steps in the constitutive secretory and endocytic pathways. When the mammalian NSF gene was cloned in 1989, its coding sequence was found to be 48% identical to that of the yeast *sec18* gene. Like NSF, Sec18 is required for multiple steps in the secretory and endocytic pathways. Sec18 protein-containing yeast cytosol can substitute for NSF in mammalian *in vitro* assays of endosome fusion and intra-Golgi transport, thus emphasizing the functional conservation and near ubiquitous requirement for NSF/Sec18p in membrane traffic.

NSF exists in both cytosolic and membrane-associated forms, and the binding of NSF to membranes is mediated by soluble NSF attachment proteins (SNAPs; three isoforms – α, β, and γ). SNAPs are functional homologs of the yeast *SEC17* gene product required for ER-to-Golgi transport. NSF and SNAPs co-assemble into 20 *S* particles in the presence of detergent-solubilized membranes, indicating the presence of membrane-bound SNAP receptors (SNAREs). Using an immunoprecipitation approach with brain membranes, the Rothman lab identified the neuronal SNAREs as synaptobrevin (VAMP), syntaxin, and SNAP-25. These three proteins are proteolytic substrates for the botulinum neurotoxins, potent inhibitors of neurotransmission, thus implicating NSF and SNAPs in regulated exocytosis at the synapse. Furthermore, the neuronal SNARE proteins displayed sequence homology to various yeast proteins required for membrane traffic; for example, SNAP-25 is homologous to yeast Sec9. This discovery therefore had a tremendous impact on the field because it indicated that all membrane fusion events were mediated by NSF, SNAPs, and SNAREs.

The discovery of a complex of NSF/SNAP/SNARE proteins suggested that this might represent a single particle mediating vesicle docking and fusion. The SNARE hypothesis postulated that vesicle targeting/docking to the correct acceptor membrane was mediated by the selective interaction of specific SNAREs located on the vesicle (v-SNAREs) with appropriate SNAREs on the target membrane (t-SNAREs). Subsequent binding of SNAPs and NSF to this complex, followed by ATP hydrolysis by NSF, would disassemble the complex, leading to membrane fusion. Although appealing, this model was inconsistent with a variety of observations, not least that membrane fusion can be elicited in the absence of ATP during regulated exocytosis. Throughout the years, it has become accepted that the SNAREs execute the membrane fusion reaction, whereas NSF and SNAPs are not directly involved in fusion. Instead, NSF/SNAPs bind to dead-end SNARE complexes resident in the same membrane (*cis*-SNARE complexes) and use the energy of NSF ATPase activity to disassemble these complexes, thus freeing individual SNAREs to engage in future membrane fusion events.

SNAREs

Although originally defined biochemically as SNAP receptors, SNAREs are now defined structurally as

proteins containing a characteristic 60–70 residue domain known as the SNARE motif. Most SNAREs are integral membrane proteins with a single transmembrane domain at their C-terminus, although some lack transmembrane domains (e.g., SNAP-25 is targeted to membranes via palmitoylation). The classification of SNAREs as v- or t-, based on their localization to vesicle or target membranes, remains extremely useful conceptually. However, a rival classification system is also widely used based on the presence of a conserved glutamine (Q) or arginine (R) at the core of the SNARE motif. Most SNAREs contain a single motif, but some – notably the SNAP-25 homologs involved in exocytosis – contain two SNARE motifs. The interaction of individual SNAREs to form a complex is mediated by the formation of a helical bundle composed of four SNARE motifs (**Figure 1(a)**). In the case of the synaptic SNARE complex, one SNARE motif is provided by syntaxin (t-/Q-SNARE), two by SNAP-25 (t-/Q-SNARE), and one by VAMP (v-/R-SNARE). The formation of the three-helical t-/Q-SNARE heterodimer is the initial, rate-limiting step; upon formation, v-/R-SNARE binding can occur relatively quickly. The zippering together of the SNARE motifs to form the four-helical bundle occurs from the membrane distal N-termini toward the membrane proximal C-termini and is very energetically favorable. It is widely believed that this zippering process actively pulls the opposing membranes together to directly drive membrane fusion. This attractive idea is supported by experiments in which recombinant yeast and mammalian SNARE proteins incorporated into synthetic proteoliposomes have been shown to reconstitute membrane fusion *in vitro* in the absence of other proteins, albeit inefficiently. SNAREs, therefore, are the minimal machinery for membrane fusion.

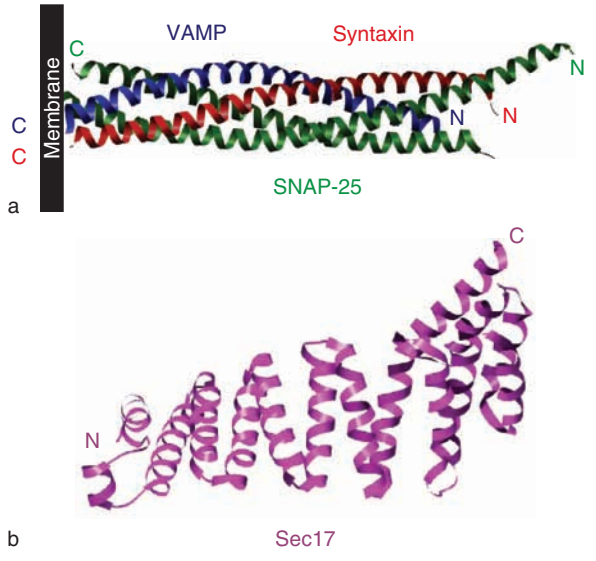

Figure 1 Structure of the SNARE complex (a) and Sec17 (b). (a) The helices are colored as follows: syntxain, red; VAMP, blue; SNAP-25, green. The relative positions of the N- and C-termini are indicated.

SNAPs

NSF is a soluble protein whose function requires attachment to membranes. Biochemical approaches led to the purification of SNAPs as the adaptor proteins that linked NSF to membranes and also provided an essential function in the intra-Golgi transport assay. Three SNAP isoforms were purified from brain: α, β, and γ, all of which are 30–40 kDa monomeric soluble proteins. α-SNAP is ubiquitously expressed, whereas β-SNAP is a brain-specific isoform that has high sequence homology to α-SNAP. Intuitively, it seems that β-SNAP must serve some brain-specific function; however, in most assays, α- and β-SNAP are interchangeable and have similar activities. γ-SNAP displays much lower sequence homology, and despite its biochemical ability to bind NSF, there are relatively few examples in which a functional role for this isoform in membrane traffic has been demonstrated. The physiological role of γ-SNAP therefore remains unclear. In contrast, abundant functional evidence supports an evolutionarily conserved role for α-SNAP in membrane traffic. Genetic evidence comes from the *in vivo* block in membrane fusion caused by mutations in the yeast homolog of α-SNAP, Sec17. Furthermore, Sec17 protein can provide SNAP activity in mammalian assays of both NSF binding and *in vitro* reconstitution of membrane traffic. α-SNAP has been shown to be involved in various membrane traffic steps, including regulated exocytosis in neuronal and endocrine cells. A mouse mutant, hydrocephalus with hop gait (*hyh*), has been found to be due to a single missense mutation in α-SNAP. These mutants are characterized by postnatal death induced by aberrant ventricular fluid regulation, suggesting an important physiological role for SNAPs in normal brain function.

The defining characteristic of SNAPs is their ability to bind NSF. However, this does not occur in solution: NSF will only bind to SNAPs that are prebound to SNAREs (or to a plastic surface, which suffices *in vitro*). Therefore, the binding site on SNAPs for NSF must normally be occluded in the cytosol, only becoming exposed upon SNARE binding. It appears that three SNAP molecules are needed to link each SNARE complex to one NSF molecule. The only available crystal structure for a SNAP is that of the yeast Sec17 protein. This reveals a highly α-helical structure comprising an N-terminal domain composed of a

series of helical hairpins capped by a C-terminal helical bundle (Figure 1(b)). Because this structure is of the protein in isolation, the residues that contact NSF and the SNAREs cannot be deduced. Nevertheless, based on biochemical analysis using mutant proteins, it seems likely that the helical hairpin domain lies along the length of the SNARE complex, with its N-terminus at the membrane-proximal region and its C-terminus in contact with NSF at the membrane-distal end of the SNARE complex. It should be noted that α-SNAP is capable of binding to syntaxin alone and also to the syntaxin/SNAP-25 t-/Q-SNARE heterodimer. In both cases, NSF can be recruited to these complexes and mediate their ATP-dependent disassembly, although whether such subcomplexes are formed *in vivo* and their functional relevance remain unknown.

Thus far, we have only considered SNAPs as adaptor proteins, linking NSF to membrane-bound *cis*-SNARE complexes. However, SNAPs also have an essential catalytic function in stimulating the ATPase activity of NSF. These two actions of SNAPs can be uncoupled by mutations in the extreme C-terminus of α-SNAP, which do not affect binding to SNARE complexes or recruitment of NSF to these complexes but abolish the SNAP-mediated stimulation of NSF ATPase activity. This results in the formation of stable NSF-SNAP-SNARE complexes that do not disassemble in the presence of hydrolyzable ATP. This demonstrates that the weak intrinsic ATPase activity of NSF is insufficient to drive SNARE complex disassembly and that stimulation of ATPase activity by SNAPs is essential for this critical function of NSF. Indeed, mutation of the highly conserved penultimate leucine 294 residue to alanine in α-SNAP produces a dominant-negative mutant that inhibits membrane traffic in both the exocytic and the endocytic pathway.

NSF

As discussed previously, abundant functional evidence from yeast and mammalian systems supports an essential physiological role for NSF in all constitutive membrane traffic steps. The general requirement for NSF in membrane trafficking extends to regulated exocytosis, as illustrated by the *Drosophila comatose* mutants. These flies appear normal at 23 °C but exhibit paralysis at 38 °C due to a block in synaptic transmission. The *D. comatose* phenotype is due to missense mutations in the *dNSF1* gene, which encodes the majority of NSF protein in adult flies. Another gene, *dNSF2*, is expressed earlier in development; mutations in this gene lead to developmental defects and overgrowth of the neuromuscular junction.

NSF is a member of the AAA+ family of ATPases associated with a variety of cellular activities. Structural and biochemical studies have shown NSF to be a homohexamer of 76 kDa subunits arranged in a ring, forming a barrel-like structure containing a central cavity. Each individual subunit comprises three domains: an N-terminal domain and two AAA+ ATPase domains, D1 and D2 (Figure 2(a)). The approximately 200-amino acid N domain is composed of two subdomains, N_A and N_B, joined by a linker (Figure 2(b)). The N domain mediates binding to the SNAP/SNARE complex, as point mutations in this region of both NSF and Sec18 inhibit SNAP/Sec17 binding. The N domain also appears to be involved in transducing the SNAP stimulation of ATPase activity, as antibodies directed at this domain result in increased ATPase activity. It is the D1 domain that is the source of this ATPase activity, although in the absence of SNAPs the intrinsic ATPase

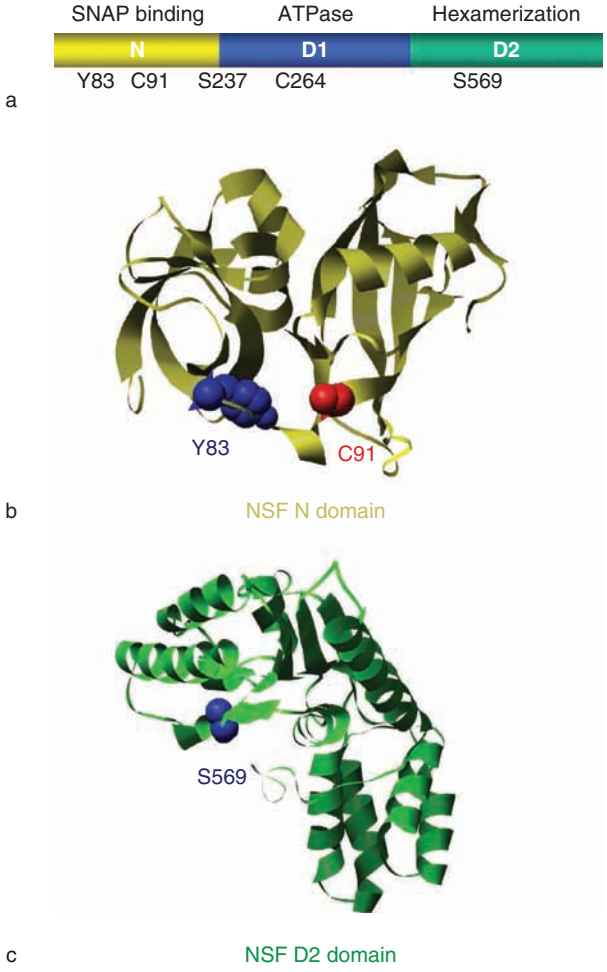

Figure 2 Domain structure and posttranslational modification of NSF. (a) Schematic diagram showing the domain structure of the NSF and the position of the residues modified by phosphorylation (Y83, S237, and S569), nitrosylation (C91), and oxidation (C264). (b) Crystal structure of the N domain, with Y83 highlighted in blue and C91 in red. (c) Crystal structure of the D2 domain, with S569 highlighted in blue.

activity of NSF is extremely low. SNAP binding to NSF stimulates D1 domain ATP hydrolysis severalfold and is essential for NSF to disassemble SNARE complexes. ATP binding and hydrolysis is coordinated by Walker A and B box motifs present in the D1 domain. Mutants in the D1 Walker A motif are unable to bind ATP and are also unable to interact with SNAP/SNARE complexes. In contrast, mutants in the D1 Walker B motif can bind ATP and interact with SNAP/SNARE complexes normally but are unable to hydrolyze the bound ATP and hence cannot disassemble these complexes. Thus, NSF binding to SNAP/SNARE complexes is dependent on the D1 ATP-bound conformation; hydrolysis of this ATP in response to SNAP stimulation both disassembles the SNAREs and releases NSF, as the ADP-bound conformation is unable to bind to SNAPs. Indeed, electron microscopy studies have revealed distinct conformations of NSF depending on its nucleotide-bound state. In the ATP-bound state, NSF's N domains extend out from the central ring, but these are withdrawn in the ADP-bound state. Mutation of residues in the D1 Walker B motif creates dominant negative mutants of NSF and Sec18, presumably because this results in the accumulation of dead-end SNAP/SNARE/NSF complexes that cannot be disassembled (just like the dominant negative SNAP L294A mutant). Although ATPase defective mutants of NSF and Sec18 are generally nonfunctional in membrane traffic, these are apparently functional in the NSF-dependent reconstitution of Golgi reassembly *in vitro*. The reason for this exception is not clear. The D2 domain is homologous to the D1 domain and also contains Walker A and B boxes (**Figure 2(c)**). However, mutations in these domains have minimal effect on NSF ATPase activity, either intrinsic or SNAP stimulated, indicating that this AAA+ ATPase domain is catalytically inactive. Instead, the D2 domain has an essential structural role in forming the hexameric structure that is required for NSF function. Indeed, truncation mutants lacking the D2 domain are not functional in membrane traffic.

Classical Functions of SNAPs and NSF

Overwhelming evidence suggests that the major, evolutionarily conserved function of SNAPs and NSF is their cooperative action in disassembling *cis*-SNARE complexes. NSF and SNAP can be thought of as chaperone and co-chaperone proteins whose job is to mediate the unfolding of such complexes wherever and whenever they form in the cell. This conclusion is supported by analysis of the temperature-sensitive yeast *sec18-1* and *D. comatose st17* and *st53* NSF mutants, which accumulate SNARE complexes upon shift to the restrictive temperature. These mutant proteins exhibit defects in SNAP binding, intrinsic ATPase activity, and SNAP-stimulated ATPase activity, respectively. It therefore seems that the *in vivo* function of SNAPs/NSF involves several sequential events: (1) SNAPs bind to *cis*-SNARE complexes; (2) NSF is recruited to form the SNARE/SNAP/NSF complex; and (3) SNAPs stimulate the intrinsic ATPase activity of NSF, triggering disassembly of the complex.

Although the requirement for SNAPs/NSF in SNARE complex disassembly is not in doubt, the precise stage at which this acts in the membrane traffic process has been controversial. Although, by definition, t-SNAREs should not reside on vesicles, it has been shown that this is in fact the case for a variety of regulated secretory vesicles. Furthermore, *cis*-SNARE complexes are present on synaptic vesicles and these can be disassembled by NSF/SNAPs. Studies of *in vitro* reconstitution of yeast vacuolar fusion indicated a predocking, priming role for Sec17/Sec18 because the function of these proteins was completed even before donor and acceptor vacuoles were mixed. One caveat of this conclusion, however, is that in such homotypic fusion systems, it is difficult to distinguish predocking from postfusion because all vacuoles presumably contain *cis*-SNARE complexes leftover from previous fusion events. Analysis of heterotypic membrane fusion in regulated exocytosis does not suffer from this problem; however, results from these studies have yielded opposite conclusions. For example, recombinant α-SNAP causes a large increase in calcium-triggered exocytosis of dense-cored granules in chromaffin cells; 99% of these granules are undocked, and there is no contribution from postfusion recycling in this assay, strongly suggesting a predocking role. Similarly, analysis of sperm exocytosis, in which the fusion of a single giant undocked vesicle, the acrosome, is triggered by calcium, reveals a stimulatory action of α-SNAP and NSF. It is difficult to see how the action of SNAPs/NSF could be anything other than predocking in these systems. In contrast, analysis of synaptic exocytosis in *D. comatose* mutants has resulted in the opposite conclusion, namely that NSF/SNAPs execute a strictly postfusion role. This is based on the observation that the entire synaptic vesicle pool at the neuromuscular junction can be exocytosed despite the inactivation of NSF, and that SNARE complexes accumulate only as a result of membrane fusion. It may be that the special requirements of the synapse for rapid exocytosis under high-frequency stimulation require a significant reserve capacity of free SNAREs. Indeed, there are approximately 70 copies of VAMP/synaptobrevin on a single synaptic vesicle. Perhaps in other regulated exocytotic events,

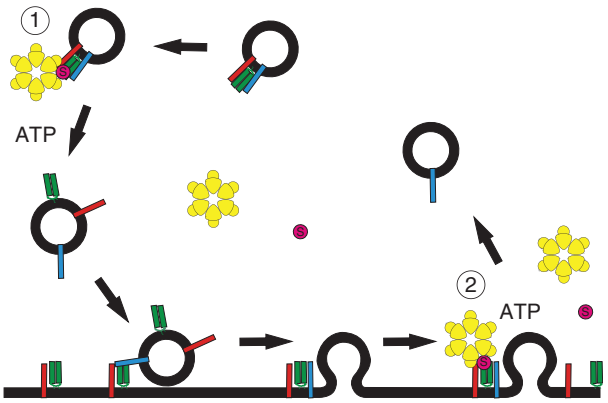

Figure 3 Role of NSF and SNAPs in membrane traffic. NSF/SNAPs associate with *cis*-SNARE complexes either on the vesicle (stage 1, predocking) or on target membranes (stage 2, postfusion). SNAP-stimulated NSF ATPase activity then disassembles the complex, releasing SNAREs to engage in further rounds of membrane fusion.

a large reserve capacity is not available, resulting in the observed acute requirement for NSF/SNAPs. It may be relevant in this regard that the initial rapid burst of exocytosis in chromaffin cells is not stimulated by SNAPs and is unaffected by NEM. Thus, it may be that in these cells, a small pool of preprimed, rapidly released vesicles can fuse without the need for acute NSF/SNAP action, whereas the bulk of release that occurs later actively requires SNAP/NSF-mediated disassembly of *cis*-SNARE complexes on the vesicle (Figure 3, stage 1) and/or plasma membranes (Figure 3, stage 2) to free up the SNAREs for subsequent fusion events.

Novel, SNARE-Independent Functions of SNAPs and NSF

In addition to its well-characterized interaction with the SNAP/SNARE complex, NSF has been reported to bind to a large number of other proteins. In most cases, this has been reported using yeast two-hybrid or coimmunoprecipitation/pull-down approaches. Note that NSF requires ATP or ADP for stability, so putative interactions observed in nucleotide-free solutions, which cause NSF aggregation, should be viewed with caution. Nevertheless, it is interesting to note that the various NSF binding proteins mainly fit into two categories. The first comprises proteins that function in various membrane trafficking steps, such as Rab protein complexes and the small, ubiquitin-like LMA1/GATE-16 proteins. These interactions may contribute to NSF's classical function in membrane traffic, albeit in a SNARE-independent manner. However, the second category of NSF binding proteins comprises plasma membrane receptors and appears unconnected with the classical function of NSF. This includes the GluR2 subunit of AMPA-type glutamate receptors, $GABA_A$ and $GABA_B$ receptor subunits, dopamine D1 and D2 receptors, adrenomedullin receptors, and several muscarinic receptor subtypes. Interestingly, the interactions generally occur on the C-terminal cytoplasmic tails of these receptors. By far the best characterized interaction is that of the GluR2 C-terminus with NSF. Originally identified by yeast two-hybrid screens, this interaction is ATP dependent but does not require SNAPs. Abrogation of the NSF–GluR2 interaction inhibits AMPA receptor-mediated synaptic transmission and affects synaptic plasticity, demonstrating the physiological importance of this interaction. The GluR2 C-terminus is a hotspot for protein interactions, with PICK1 also binding to this site. Binding to PICK1 facilitates endocytosis, and hence functional removal, of AMPA receptors. NSF can bind to this GluR2–PICK1 complex in the absence of SNAPs, but it can only disassemble the complex upon addition of SNAPs. This bears a striking resemblance to the classical SNARE disassembly function of NSF, but in this case the GluR2/PICK1/SNAP complex is the substrate for disassembly and SNAPs play a purely catalytic role in stimulating NSF ATPase activity. It is thought that disassembly of PICK1 both antagonizes endocytosis and frees up the GluR2 C-terminus to interact with other synaptic proteins such as ABP/GRIP, thus stabilizing AMPA receptor surface expression. Interestingly, overexpression of α- and β-SNAPs has distinct effects on AMPA receptor endocytosis, providing a rare example of a functional difference between these two isoforms and suggesting that the brain-specific nature of β-SNAP may be related to such nonclassical, SNARE-independent functions. It is tempting to speculate that the interaction of NSF with other receptors may also be concerned with disassembling multiprotein complexes containing these receptors. In this way, the classical role of NSF as a chaperone acting to unwind SNARE protein complexes would simply have been adapted for the unfolding of other structurally similar protein complexes.

Regulation of NSF by Posttranslational Modifications

NSF could be thought of as a housekeeping protein whose job is to continually recycle protein complexes throughout the cell. However, cellular NSF activity is not uniform and is in fact regulated by posttranslational modifications. Secretion of Weibel–Palade bodies from endothelial cells, which mediates vascular

thrombosis and inflammation, is inhibited by nitric oxide (NO) and hydrogen peroxide (H_2O_2) due to inactivation of the exocytosis machinery. Treatment of NSF with NO or H_2O_2 was found to block its ability to disassemble SNARE complexes. Cysteine 91 in the N domain and cysteine 264 in the D1 domain are the likely sites of S-nitrosylation and oxidation, respectively (**Figures 2(a) and 2(b)**). Treatment of NSF with H_2O_2 inhibits intrinsic ATPase activity, explaining the block in SNARE complex disassembly induced by this agent. This is consistent with the position of cysteine 264 in the D1 domain Walker A ATP-binding motif. Furthermore, the inhibition of exocytosis by H_2O_2 in permeabilized endothelial cells is prevented by expression of a C264T mutant NSF. In contrast, the mechanism by which NO prevents NSF-mediated SNARE complex disassembly remains unclear because NO has no effect on SNAP binding or on NSF's intrinsic ATPase activity. It may be that NO blocks ATPase activation by SNAPs, although this speculation remains to be tested experimentally. Interestingly, nitrosylation of NSF not only affects its classical function in SNARE complex disassembly in endothelial cell exocytosis but also impacts on its novel function in regulating AMPA receptors. NO treatment of NSF in this case increased binding to the GluR2 subunit, an effect that was mimicked by a C91A mutant NSF. Activation of NMDA receptors leads to both NSF nitrosylation and an NO-dependent increase in GluR2 incorporation into the plasma membrane. This suggests a model in which activation of postsynaptic NMDA receptors leads to generation of NO, which in turns nitrosylates NSF, facilitating binding to GluR2 to promote surface expression of the receptor. Such a model would be consistent with the involvement of both NO and NSF in synaptic plasticity. Whether postsynaptically liberated NO acts as a retrograde messenger to modulate NSF's classical function in SNARE complex disassembly in the presynaptic terminal is unknown.

NSF has also been shown to be regulated by phosphorylation. Phosphorylation of NSF *in vitro* on tyrosine 83 by the tyrosine kinase Fes results in two effects: an increase in intrinsic ATPase activity and a decrease in affinity for α-SNAP. Tyr83 is located at the junction between NSF's two N-terminal subdomains, and phosphorylation of this residue would likely disrupt the interface between the subdomains and hence affect SNAP binding (**Figures 2(a) and 2(b)**). Expression of a nonphosphorylatable NSF mutant, Y83F, induces the formation of large vesicles, the same phenotype seen upon expression of the tyrosine phosphatase PTP-MEG2. These results suggest that homotypic fusion of secretory vesicles is regulated by phosphorylation of NSF on Tyr83. Presumably, inhibition of NSF by tyrosine kinases causes an accumulation of dead-end *cis*-SNARE complexes and thereby inhibits homotypic fusion, whereas local dephosphorylation via vesicular PTP-MEG2 antagonizes this effect. In addition, it appears that the same phosphorylation event also affects heterotypic fusion in exocytosis because expression of phosphomimetic NSF-Y83E in Jurkat cells partially inhibited interleukin-2 secretion. Interestingly, earlier work demonstrated depolarization-induced NSF phosphorylation in brain synaptosomes, suggesting that such regulation of NSF may also occur in neurons. In this case, phosphorylation of Ser237 by protein kinase C inhibited NSF binding to the SNAP/SNARE complex – the same biochemical effect as Tyr83 phosphorylation. NSF has also been shown to be phosphorylated by the serine/threonine kinase Pctaire1. In this case, phosphorylation occurs on Ser569 in the D2 domain (**Figures 2(a) and 2(c)**) and results in destabilization of the hexamer. Phosphorylation on this residue would therefore inhibit NSF function, and indeed a nonphosphorylatable S569A mutant NSF enhances exocytosis in PC12 cells.

Conclusions

Much has been learned during the past 20 years since the discovery of NSF. We now have considerable structural, biochemical, and functional insight into NSF and SNAPs. The picture that emerges is of NSF as a molecular chaperone that acts to disassemble protein complexes in concert with its cochaperones, SNAPs (**Figure 3**). The fact that NSF/SNAPs can disassemble all the different SNARE complexes formed as a result of membrane fusion indicates that this chaperone machine recognizes a general tertiary structure rather than a specific sequence motif. This flexibility probably facilitated the adoption of different, non-SNARE complexes as substrates for disassembly later in evolution, such as the GluR2–PICK1 complex. Much remains unknown, however. For example, does this general principle hold for the other proteins that NSF has been found to bind, and what is the physiological relevance of these interactions? The physiological functions of β- and γ-SNAPs also remain unclear. It seems likely that despite being the first proteins identified in membrane fusion, the future will hold surprising discoveries regarding these key molecules.

See also: Adenosine Triphosphate (ATP); AMPA Receptors: Molecular Biology and Pharmacology; AMPA Receptor Cell Biology/Trafficking; Calcium Channels and SNARE Proteins; GABA-A Receptors: Developmental Roles; GABA-A Receptor Synaptic Functions; GABA-A

Receptors: Molecular Biology, Cell Biology and Pharmacology; GABA-B Receptors: Molecular Biology and Pharmacology; GABA-B Receptor Function; Rab3; SNAREs.

Further Reading

Barnard RJO, Morgan A, and Burgoyne RD (1997) Stimulation of NSF ATPase activity by α-SNAP is required for SNARE complex disassembly and exocytosis. *Journal of Cell Biology* 139: 875–883.

Block M, Glick BS, Wilcox CA, Wieland FT, and Rothman JE (1988) Purification of an *N*-ethylmaleimide sensitive protein catalyzing vesicular transport. *Proceedings of the National Academy of Sciences of the United States of America* 85: 7852–7856.

Clary DO, Griff IC, and Rothman JE (1990) SNAPs, a family of NSF attachment proteins involved in intracellular membrane fusion in animals and yeast. *Cell* 61: 709–721.

Hanley JG, Khatri L, Hanson PI, and Ziff EB (2002) NSF ATPase and alpha-/beta-SNAPs disassemble the AMPA receptor–PICK1 complex. *Neuron* 34: 53–67.

Matsushita K, Morrell CN, Cambien B, et al. (2003) Nitric oxide regulates exocytosis by S-nitrosylation of *N*-ethylmaleimide-sensitive factor. *Cell* 115: 139–150.

Morgan A and Burgoyne RD (1995) Is NSF a fusion protein? *Trends in Cell Biology* 5: 335–339.

Morgan A, Dimaline R, and Burgoyne RD (1994) The ATPase activity of *N*-ethylmaleimide-sensitive fusion protein (NSF) is regulated by soluble NSF attachment proteins. *Journal of Biological Chemistry* 269: 29347–29350.

Novick P, Field C, and Schekman RW (1980) Identification of 23 complementation groups required for post-translational events in the yeast secretory pathway. *Cell* 21: 205–215.

Rothman JE (1994) Mechanisms of intracellular protein transport. *Nature* 372: 55–62.

Sollner T, Whiteheart SW, Brunner M, et al. (1993) SNAP receptors implicated in vesicle targeting and fusion. *Nature* 362: 318–324.

Whiteheart SW, Griff IC, Brunner M, et al. (1993) SNAP family of NSF attachment proteins includes a brain-specific isoform. *Nature* 362: 353–355.

Whiteheart SW, Rossnagel K, Buhrow SA, Brunner M, Jaenicke R, and Rothman JE (1994) *N*-ethylmalemide-sensitive fusion protein: A trimeric ATPase whose hydrolysis of ATP is required for membrane fusion. *Journal of Cell Biology* 126: 945–954.

Whiteheart SW, Schraw T, and Matveeva EA (2001) *N*-ethylmaleimide sensitive factor (NSF) structure and function. *International Review of Cytology* 207: 71–118.

Wilson DW, Wilcox CA, Flynn GC, et al. (1989) A fusion protein required for vesicle-mediated transport in both mammalian cells and yeast. *Nature* 339: 355–359.

Zhao C, Slevin JT, and Whiteheart SW (2007) Cellular functions of NSF: Not just SNAPs and SNAREs. *FEBS Letters* 581: 2140–2149.

Nuclear Movements in Neurons

Y-H Youn and A Wynshaw-Boris, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Types of Nuclear Movements during Neuronal Development

Two basic types of nuclear migration occur during early brain development: (1) interkinetic nuclear migration of mitotically dividing neural stem cells and radial glial progenitor cells and (2) nuclear translocation during migration of postmitotic neurons. During the development of the mammalian brain, the self-renewal of neural progenitor cells can generate two daughter progenitor cells with the same cell fate by symmetric (proliferative) cell divisions, or one daughter progenitor cell and a second, different cell type (neuron or glial cell) by asymmetric (neurogenic or gliogenic) cell divisions. Prior to the onset of neurogenesis, neuroepithelial stem cells (NSCs) in the neuroepithelium, which consist of a pseudostratified single-cell layer of NSCs, display interkinetic nuclear migration whereby nuclei migrate up and down the entire apical–basal axis during the cell cycle. At the onset of neurogenesis, neuroepithelial stem cells give rise to a distinct cell type, radial glial cells. Radial glial cells represent more fate-restricted progenitors, and most neurons in the central nervous system are derived from radial glial cells. In these progenitors, interkinetic nuclear migration also occurs but it is limited to the ventricular, or proliferative, zone and is excluded from the developing intermediate zone and cortical plate. Although the exact mechanism of interkinetic nuclear migration in neuroepithelial and radial glial cells is unknown, it is known that the apical direction of nuclear migration (G2) is faster than that occurring in the basal direction (G1).

A second type of nuclear migration occurs in postmitotic neurons. The leading process of migrating neurons is in front of the soma, which contains the nucleus, and is directed toward the pial surface. The trailing process is behind the nucleus. The position of the nucleus advances into the leading process as the neuron migrates toward the pial surface during radial migration. This nuclear translocation is mediated by a microtubule-based nuclear cage that is linked to the centrosome (Figure 1), and the entire process of nuclear migration is controlled by cytoplasmic dynein.

Apical–Basal Polarity and Interkinetic Nuclear Migration during Neurogenesis

The neuroepithelium in the proliferating neural tube and the radial glial progenitors during neurogenesis display characteristic apical–basal polarity. This is manifested by strict apical localization of proteins and cellular components such as the centrosomes, prominin-1, adherens junctions, apical protein kinase C (PKC), Par3, and Par6. Apical–basal polarity is crucial to allow symmetric and/or asymmetric distribution of cellular components during mitosis and to orient the plane of cell division. Additionally, apical–basal polarity of these components results in the exposure of the migrating nucleus to different concentrations of these components during progression through the cell cycle. For example, the nucleus of proliferating neuroepithelial cells moves from the apical to the basal side of the cell during G1 phase of the cell cycle, where DNA synthesis occurs during S phase, before the nucleus moves back to the apical side during G2 to finally divide at the apical side. Although apical–basal polarity is somewhat relaxed once neurogenesis commences, it is still present in radial glial cells. These cells also display interkinetic nuclear migration that only spans the ventricular zone rather than the entire length of the cell.

Cytoskeletal–Centrosome Coordination during Neuronal Migration

The migration of progenitor neurons to their final destination within the brain requires major cytoarchitectural changes of the actin and microtubule (MT) cytoskeletons. MTs are essential to maintain the extended processes of forward-moving neurons, and actin filaments play a key role in axonal outgrowth and cell motility. Nucleokinesis mainly depends on the microtubule network in migrating neurons. The site of MT emanation just in front of the nucleus has been recognized as the centrosome – the microtubule organizing center (MTOC), a structure found in all plant and animal cells from which microtubules radiate – of the cell. A microtubule-based nuclear cage complex couples the centrosome and nucleus (Figure 1). Disturbance of this coupling complex results in migration defects. Following the advance of the centrosome into the leading process of the migrating neuron, the nucleus translocates forward in a saltatory fashion, with myosin II-based contraction at the rear of the neuron. Myosin II seems to accumulate periodically behind the nucleus, forming a short tail at the rear of the cell, suggesting that myosin II has an important role in maintaining the nucleus on the same axis as the MTOC and the extending growth cone.

In mammalian cell polarity, postsynaptic density/disc/zonula occludens-1 (PDZ)-containing proteins Par3/Par6 are required to regulate the orientation of the mitotic spindle during cell division. Cell polarity is

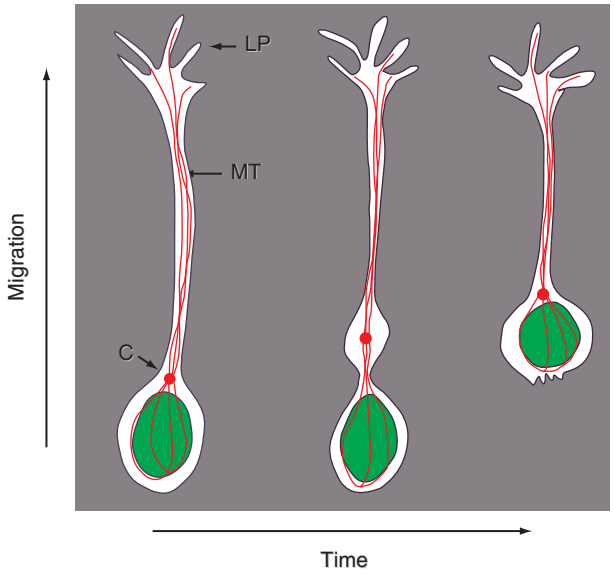

Figure 1 A model of nuclear migration in neuronal cells. A centrosome (C) is controlled by the cell polarity and guides the direction of movement during the migration. The centrosome localizes in front of the nucleus (N) and toward the leading process (LP). At the stage depicted in the middle figure, a cytoplasmic dilation forms ahead of the nucleus, and the forces generated from the centrosome and microtubule (MT) pull on the microtubule cage (red lines) surrounding the nucleus. The nucleus thus moves forward toward the centrosome, and the process can then repeat, enabling migration.

determined by spatially localized Par3/Par6 via phosphatidylinositol 3-kinase (PI3K) phosphorylation of glycogen synthase kinase-3 (GSK-3). This pathway is crucial for cell polarity, centrosome reorientation, and wound healing (cell migration) in a cultured astrocytes/wound model system. The integrin signaling pathway is also involved in cell polarity and activation of the cytoplasmic mammalian Par6 (mPar6) complex associated with PKC-ζ. During neuronal migration, the role of Par genes has been reported. Par6α is localized to the neuronal centrosome along with PKC-ζ. Deficiency of Par6α results in the scattered distribution of PKC-ζ and mislocalized perinuclear microtubule cytoskeleton.

Besides microtubules, actin filaments are also involved in the interkinetic nuclear migration of neuroepithelial cells. Cytochalasin B interferes with actin polymerization and induces defects of nuclear migration in neuroepithelial cells through the disruption of nonmuscle myosin heavy chain II-B. However, this drug does not stop the cell cycle phase, which leads to mitosis when the nucleus is localized anywhere in the ventricular zone, rather than the apical zone. On the other hand, hydroxyurea inhibits S phase progression but does not interfere with interkinetic nuclear migration (see later). These studies indicate that the two different cellular events, interkinetic nuclear migration and the cell cycle of neuroepithelial cells, are coordinated with each other during development but have different underlying regulatory mechanisms.

Dynein and Lis1/Ndel1 in the Regulation of Nucleokinesis

Type I lissencephaly is a human neuronal migration disorder that results from a disorganization of both pyramidal neurons (radial migration) and interneurons (tangential migration). It is characterized by generalized agyria/pachygyria (reduction or complete lack of gyri and sulci), four abnormal cortical layers, enlarged ventricles, and generalized neuronal heterotopias. Miller–Dieker syndrome (MDS) consists of severe classical lissencephaly, characteristic facial anomalies, and occasionally other malformations. Children with MDS are severely retarded and suffer from epilepsy. Craniofacial dysmorphisms (characteristic of abnormal development) include microcephaly with bitemporal narrowing, a high forehead, a small nose with upturned nares, thin vermilion border, and micrognathia (small lower jaw). The disorder is invariably fatal in early childhood. Isolated lissencephaly sequence (ILS) is a heterogeneous disorder consisting of variably severe lissencephaly with no other major malformations, such as the craniofacial dysmorphism seen in MDS. Children with ILS are also severely retarded and typically have epilepsy. MDS (100%) and some cases of ILS (40%) are the result of haploinsufficiency at human chromosome 17p13.3, with visible or submicroscopic deletions detectable by fluorescent *in situ* hybridization (FISH). The *LIS1* (*PAFAH1B1*) gene was cloned from this region in several MDS patients. Point mutations and an intragenic deletion in *LIS1* were identified in ILS patients; no gross structural chromosomal rearrangements were found, confirming that mutations or deletions of the *LIS1* gene are responsible for classical lissencephaly in ILS and MDS patients. The more severe neuronal migration phenotype and facial dysmorphisms displayed by MDS patients suggest that genes other than *LIS1*, such as *14-3-3e* (see later), are responsible for these phenotypes.

The Lis1 protein was first reported to be a stoichiometric component of purified platelet-activating factor acetylhydrolase 1b (PAFAH-1b). However, it was also found that Lis1 is a component of a highly conserved pathway that regulates nuclear migration in fungi. The Lis1 homolog in the slime mold *Aspergillus nidulans,* NudF, is part of a signaling pathway that regulates nuclear migration along microtubules via the regulation of dynein motor function. In this slime mold, nuclei migrate toward the growing tip of

hyphae to establish regular spacing, a process termed nucleokinesis. Several nuclear distribution mutants (*nud* mutants) have been generated in this organism. One of these, *nudF*, is the fungal homolog of *LIS1*. The identification of other *nud* mutants directly implicates cytoplasmic dynein and dynactin (a regulator of the microtubule motor protein dynein) in this process. The *nudA* and *nudG* genes code for the heavy chain and the 8-kDa light chain of cytoplasmic dynein, respectively. A mutant allele of *nudF*, *nudF6*, is suppressed by mutations in *nudA*, the cytoplasmic dynein heavy chain, providing genetic evidence that NudA and NudF are in the same genetic pathway, and suggesting that Lis1 might directly interact with cytoplasmic dynein. *NudK* codes for Arp1, which is part of the dynactin complex, and dynactin is required for dynein-mediated vesicle transport. A screen for multicopy suppressors of the *nudF* phenotype identified *nudE*, demonstrating that overexpression of NudE can rescue the *nudF* (*LIS1*) phenotype. NudE is a homolog of the nuclear distribution protein RO11 in *Neurospora crassa,* the mitotic protein MP43 in *Xenopus laevis*, and two murine homologs of NudE named NUDEL (for NudE-like) and NudE (see later). Thus, several genes responsible for nuclear movement have been isolated from *Aspergillus*. The Nud pathway is remarkably conserved in eukaryotes.

The mouse gene, *Lis1* (*Pafah1b1*), has been disrupted to examine the consequences of dosage reduction and complete deficiency of *Lis1 in vivo*. Mice with decreased levels of *Lis1* exhibit dose-dependent disorganized cortical layers, hippocampus, cerebellum, and olfactory bulb due to cell-autonomous neuronal migration defects, and are a good model for the human disorder. Complete loss of Lis1 results in peri-implantation lethality, demonstrating that *Lis1* is an essential gene. Several lines of evidence support the conclusion that there are *in vivo* migrational defects in mice with reduction of Lis1 dosage. These include histological analysis during development, bromodeoxyuridine (BrdU) birthdating experiments (a method to examine whether there are abnormalities in cortical layering by labeling postmitotic neurons with a pulse of BrdU; those neurons born just after BrdU injection are intensely labeled with BrdU, since they do not dilute the label as would occur if they divide further) and *in vitro* migration of granule cells from cerebellar cell reaggregates. For example, Lis1 is required for nuclear movement during neuronal migration by coupling the nucleus to the centrosome. Cerebellar granule cells from $Lis1^{+/-}$ mice display slowed migration and a lengthening of the distance between the centrosome and the nucleus. Lis1 is localized to the centrosome with some extension toward the nucleus. Consistent with an important role in dynein function, inhibition of dynein by overexpression of dynamitin results in a similar phenotype. Finally, studies show that interference RNA (RNAi) knockdown of *Ndel1*, *Lis1*, or *dynein* by *in utero* electroporation into the developing neocortex results in impaired neuronal positioning and causes the uncoupling of the centrosome and nucleus. These studies suggest that nuclear movement during neuronal migration is mediated by a Lis1 complex regulating cytoplasmic dynein function, a function that is conserved through evolution from *Aspergillus* to mammals.

Lis1 directly interacts with three distinct sites in the cytoplasmic dynein heavy chain, in regions that contain the basal stem or tail portion of the heavy chain molecule responsible for cargo binding and the motor domain. Lis1 also interacts with the dynamitin subunit of dynactin, which is a dynein supplemental complex that binds to dynein and regulates its binding to several cytoplasmic structures. Lis1 is colocalized with cytoplasmic dynein at the mitotic kinetochores, indicating a role of Lis1 in chromosomal behavior. For example, microinjection of anti-Lis1 antibody results in a delay of mitotic progression accompanied with chromosomal defects at the metaphase plate. Lis1 is also expressed at the cortical (marginal) area of proliferating epithelial cells, and its overexpression results in disruption of the mitotic spindle organization as a result of reductions in cortical distribution of dynein. Thus, the dynein/Lis1 complex plays an important role in the regulation of spindle orientation. These findings suggest that the mitotic spindle cleavage plane plays an important role during neurogenesis of neuroepithelial cells in the ventricular zone.

Mammalian Lis1 is in a complex associated with mNudC and with two NudE orthologs (discussed earlier), mNudE (Nde1) and its isoform NudEL (Ndel1). These three Nud homologs bind to dynein motor proteins, and these proteins also play important roles in neuronal migration via nucleokinesis. Nde1-deleted mice are viable and have been reported to have migrational defects but normal cortical lamination, though the mutant cortex has fewer neuronal cells and significantly thinner superficial cortical layer compared to wild type. Ndel1 is required for neuronal migration via interaction with Lis1. Complete loss of Ndel1 results in embryonic lethality at the peri-implantation stage and a deficiency of cell proliferation. Disruption of Ndel1 results in defective nuclear/centrosomal coupling and wider average distances between centrosome and nucleus. Ndel1 may also play an important role in MT organization, since it also interacts with katanin p60 for remodeling of MTOC. Thus, Ndel1 and Lis1, probably as well as

Nde1, have common roles for dynein regulation, the MTOC, and neuronal migration.

Ndel1 is phosphorylated and regulated by Cdk5 (Cdc2 in fibroblasts) and aurora A kinase. Cdk5 is a brain-specific kinase central for axonal guidance/growth and neuronal migration in the developing cerebral cortex. It is activated by interaction with p35 or p39, required co-activators. Ndel1 has five proline-dependent serine/threonine phosphorylation motifs and is phosphorylated by Cdk5/p35. Phospho-Ndel1 (P-Ndel1) can bind to the 14-3-3 protein epsilon subunit directly, which protects P-Ndel1 from phosphatase attack. Loss of 14-3-3e in mice results in neuronal migration defects, disruption of dynein motor function, and decreased steady-state levels of Cdk5-phosphorylated P-Ndel1. In addition, 14-3-3e is deleted in human patients with MDS, and may be responsible for some of the more severe defects displayed by MDS patients. Thus, the activity of Ndel1 and the Lis1/dynein complex appears to be tightly regulated by a typical signal transduction pathway.

DCX in the Regulation of Nucleokinesis

The major cause of X-linked lissencephaly is mutation of the *doublecortin* (*DCX*) gene. Mutations in the *DCX* gene cause gross neocortical disorganization and lissencephaly in hemizygous males, while heterozygous females show a mosaic phenotype with a normal cortex as well as a second band of misplaced (heterotopic) neurons beneath the cortex ('double cortex syndrome'). Little is known of the nature of involvement of the protein doublecortin in neuronal migration. This is partly because doublecortin appears to be present only in mammals, and so it has not been possible to study its function in fungi or *Drosophila*. However, both Lis1 and doublecortin are known to associate with microtubules in mammalian cells. Lis1 appears to increase microtubule polymerization and to decrease microtubule depolymerization rates *in vitro* and doublecortin appears to bind to and polymerize microtubules both *in vitro* and in heterologous cells. It seems then that Lis1 and doublecortin are both working in concert to maintain microtubule polymerization, and might function together to increase the pool of polymerized microtubules to promote neuronal migration, although how microtubule polymerization might facilitate successful neuronal migration remains to be elucidated.

As the nuclei move into the leading process, the network of perinuclear MTs continuously sustains the tight centrosome–nucleus coupling complex. Doublecortin plays an important role in control of the formation of the MT network and centrosome–nucleus coupling. In growing axons and dendrites, doublecortin is particularly expressed, indicating that this protein is responsible for the development of the leading process. Doublecortin may be involved in the stabilization of MTs and may be localized to the MTs bridging the nucleus and centrosome in moving cells. Overexpression of doublecortin in neurons promotes locomotion and increases the angle of nuclear–centrosome coupling, while *DCX* small interfering RNA (siRNA) leads to defects of radial migration and cortical development. However, the exact role of doublecortin in MT stabilization of moving neurons is unknown.

Concluding Remarks

Several core mechanisms regulating interkinetic nuclear migration, including cell polarity and signaling, have been shown over the past several years to reorganize cytoskeletal architecture in migrating and dividing cells. However, many important questions remain. The roles of extracellular cues in nucleokinesis need further study. Recent reports of vertebrate neurogenesis provide hints to the importance of asymmetric cell division in cell fate and migration, but the signaling molecules and cascades that regulate nucleokinesis and link it to the centrosomal positioning in the leading edge have yet to be described. Discovery of pathways that determine the architectural remodeling of cytoskeletal dynamics in the leading and trailing edges of migrating cells will undoubtedly provide a more advanced framework for understanding nucleokinesis.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Contextual Interactions in Visual Perception; Cytoskeletal Interactions in the Neuron; Cytoskeleton in Plasticity; Microtubule Associated Proteins in Neurons; Microtubules: Organization and Function in Neurons; Neural Stem Cells: Adult Neurogenesis; Neuronal Motility and Structure: Cdk5 Pathways; Neuronal Motility and Structure: MARK and GSK Pathways; Stem Cells and CNS Repair.

Further Reading

Anderson JS, Wilkinson CJ, Mayor T, et al. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 426: 570–574.

Bloch-Gallego E, Causeret F, Ezan F, et al. (2005) Development of precerebellar nuclei: Instructive factors and intracellular mediators in neuronal migration, survival and axon pathfinding. *Brain Research Reviews* 49: 253–266.

Bock HH and Herz J (2003) Reelin activates Src family tyrosine kinases in neurons. *Current Biology* 13: 18–26.

Boquet P (1999) Bacterial toxins inhibiting or activating small GTP-binding proteins. *Annals of the New York Academy of Sciences* 886: 83–90.

Cruz JC and Tsai LH (2004) A Jekyll and Hyde kinase: Roles for Cdk5 in brain development and disease. *Current Opinion in Neurobiology* 14: 390–394.

Gadde S and Heald R (2004) Mechanisms and molecules of the mitotic spindle. *Current Biology* 14: R797–R805.

Goshima G, Nedelec F, and Valle RD (2005) Mechanisms for focusing mitotic spindle pole by minus end-directed motor proteins. *Journal of Cell Biology* 171: 229–240.

Gotz M and Huttner WB (2005) The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology* 6: 777–788.

Hatten ME (2005) LIS-less neurons don't even make it to the starting gate. *Journal of Cell Biology* 170: 867–871.

Haubensak W, Attardo A, Denk W, et al. (2004) Neurons arise in the basal neuroepithelium of the early mammalian telencephalon: A major site of neurogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 101: 3196–3201.

Huttner WB and Brand M (1997) Asymmetric division and polarity of neuroepithelial cells. *Current Opinion in Neurobiology* 7: 29–39.

Kappeler C, Saillour Y, Baudoin JP, et al. (2006) Branching and nucleokinesis defects in migrating interneurons derived from doublecortin knockout mice. *Human Molecular Genetics* 15: 1387–1400.

Lechler T and Fuchs E (2005) Asymmetric cell divisions promote stratification and differentiation of mammalian skin. *Nature* 437: 275–280.

Marin O and Rubenstein JL (2001) A long, remarkable journey: Tangential migration in the telencephalon. *Nature Reviews Neuroscience* 2: 780–790.

Miyata T, Kawaguchi A, Saito K, et al. (2004) Asymmetric production of surface-dividing and non-surface-dividing cortical progenitor cells. *Development* 131: 3133–3145.

Morales-Mulia S and Scholey JM (2005) Spindle pole organization in *Drosophila* S2 cells by dynein, abnormal spindle protein (Asp), and KLP10A. *Molecular Biology of the Cell* 16: 3176–3186.

Noctor SC, Martinez-Cerdeno V, Ivic L, et al. (2004) Cortical neurons arise in symmetric and asymmetric division zones and migrates through specific phases. *Nature Neuroscience* 7: 136–144.

O'Rourke NA, Dailey ME, Smith SJ, et al. (1992) Diverse migratory pathways in the developing cerebral cortex. *Science* 258: 299–302.

Roegiers F and Jan YN (2004) Asymmetric cell division. *Current Opinion in Cell Biology* 16: 195–205.

Samuels BA and Tsai LH (2004) Nucleokinesis illuminated. *Nature Neuroscience* 7: 1169–1170.

Schaar BT, Kinoshita K, and McConnell SK (2004) Doublecortin microtubule affinity is regulated by a balance of kinase and phosphatase activity at the leading edge of migrating neurons. *Neuron* 41: 203–213.

Schaar BT and McConnell SK (2005) Cytoskeletal coordinating during neuronal migration. *Proceedings of the National Academy of Sciences of the United States of America* 102: 13652–13657.

Shu T, Tseng HC, Sapir T, et al. (2006) Doublecortin-like kinase controls neurogenesis by regulating mitotic spindles and M phase progression. *Neuron* 49: 25–39.

Solecki DJ, Model L, Gaetz TM, et al. (2004) Par6alpha signaling controls glial-guided neuronal migration. *Nature Neuroscience* 7: 1195–1203.

Tanaka T, Serneo FF, Higgins C, et al. (2004) Lis1 and doublecortin function with dynein to mediate coupling of the nucleus to the centrosome in neuronal migration. *Journal of Cell Biology* 165: 709–721.

Tanaka T, Serneo FF, Tseng HC, et al. (2004) Cdk5 phosphorylation of doublecortin ser297 regulates its effect on neuronal migration. *Neuron* 41: 215–227.

Toyo-Oka K, Sasaki S, Yano Y, et al. (2005) Recruitment of katanin p60 by phosphorylated NDEL1, an LIS1 interacting protein, is essential for mitotic cell division and neuronal migration. *Human Molecular Genetics* 14: 3113–3128.

Toyo-Oka K, Shionoya A, Gambello MJ, et al. (2003) 14-3-3epsilon is important for neuronal migration by binding to NUDEL: A molecular explanation for Miller–Dieker syndrome. *Nature Genetics* 34: 274–285.

Tsai LH and Gleeson JG (2005) Nucleokinesis in neuronal migration. *Neuron* 46: 383–388.

Wodarz A and Huttner WB (2003) Asymmetric division during neurogenesis in *Drosophila* and vertebrates. *Mechanism of Development* 120: 1297–1309.

Yingling J, Toyo-Oka K, and Wynshaw-Boris A (2003) Miller–Dicker syndrome: Analysis of a human contiguous gene syndrome in the mouse. *American Journal of Human Genetics* 73: 475–488.

Nucleic Acid Introduction into Primary Neurons and Glia

J-Y Sul*, A Sharma*, J Eberwine, and P G Haydon, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Published by Elsevier Ltd.

By modifying gene expression within the nervous system and following its impact on brain function, neuroscientists have gained insight into brain functionality in a variety of different contexts. Exogenous expression in a system lacking a gene of interest has been utilized to recapitulate a lost cellular function or reverse a behavioral phenotype implicating the gene's involvement in a certain pathway or function. Overexpression of genes has allowed for improved detection of protein-binding partners and for the unmasking of putative cellular functions. Expression of mutated forms of genes has helped establish disease models in both cell lines and organisms. The expression of exogenous nucleic acids with affinity tags and fluorescent markers has allowed for improved detection and noninvasive tracking of gene expression. The ultimate goal of gene transfection is to understand brain function from the level of single cells to that of entire organisms. Within the past couple of decades, several different methodologies have been developed for gene delivery and modified for efficient gene transfer into postmitotic neurons and mitotic glial cells. Exogenous genes can now be expressed in single cells, large population of cultured cells, cultured slices, and organisms. In this article, we discuss these methodologies and also highlight novel methods of gene delivery.

Gene delivery methods can be categorized as chemical, viral, or physical based on the mechanism of entry of the gene into the cytosol. Chemical methods are the most commonly utilized approach in which genes are delivered through lipid–nucleic acid complexes, cationic polymer–nucleic acid complexes, and calcium phosphate–nucleic acid coprecipitates. Viral gene delivery involves the engineering of viral genomes to infect target cells with a specific gene(s) of interest. Physical methods include the direct transfer of naked nucleic acids into cells. These methods include electroporation, biolistic methodologies (gene gun), microinjection, and the recently developed technique phototransfection.

Chemical and viral mechanisms of gene delivery are similar in that they first require that the nucleic acid be packaged into proper delivery vectors. For chemical methods, nucleic acids are packaged into chemical vectors that have the ability to penetrate the plasma membrane. This packaging is directed by both electrostatic interactions and the hydrophobic effect. For example, lipid chemical vectors are macromolecules containing both lengthy, hydrophobic tails and hydrophilic cationic head groups. The amphipathic nature of these molecules allows them to interact with both the nucleic acid and the lipid membrane. The hydrophobic tails, in the presence of an aqueous environment and at the proper concentration, condense into energetically favorable, spherical vesicles called liposomes that can interact with the plasma membrane. The cationic head group forms complexes with the negatively charged nucleic acids largely through electrostatic interactions. Thus, the nucleic acid is packaged on the surface of the liposome in a manner that neutralizes the negative charge and that can protect it from intracellular nuclease activity. A similar process guides the formation of polymer–nucleic acid complexes as well as calcium–phosphate–nucleic acid complexes. Viral methods of gene delivery involve directly engineering a foreign gene(s) of interest into a viral genome. This modified viral genome is then incorporated into the genome of packaging cells through a molecular recombination event. These packaging cells have the ability to assemble intracellular viral particles containing the viral genome and the gene(s) of interest.

Following nucleic acid packaging, the genes have to be transferred and expressed in target cells. For chemical and viral transfection, target cell entry is believed to occur following vector fusion with the cell membrane and internalization via an endocytosis-like process. Viral particles or chemical complexes containing nucleic acids are trapped within small membranous vesicles derived from the plasma membrane called endosomes. Viral particles have the ability to dissociate from cytosolic endosomes and can be transported to the nucleus. Some viruses require active mitosis to enter the nucleus, while others have direct nuclear membrane entry mechanisms. Following nuclear entry, viral particles dissociate in the nucleus, exposing the viral genome containing the gene(s) of interest. The nucleic acid can be expressed directly or following incorporation into the target cell genome. Chemical complexes can also be found in endosomes but do not have efficient endosomal dissociation mechanisms. Consequently, following endosome fusion with lysosomes, the nucleic acid can be degraded through the action of lysosomal enzymes. However, the inclusion of endosomal and lysosomal escape agents such as chloroquine and the modification of lipid vectors with pH buffering head groups have protected against lysosomal enzyme degradation.

*Both authors contributed equally to this work.

In addition, the use of nuclear targeting peptides has assisted in nuclear entry via chemical methods. These developments have improved transfection efficiency through chemical transfection methodologies.

Chemical transfection methods have lower transfection efficiencies (10–25%) and are generally not utilized for *in vivo* gene expression. However, they have a number of benefits and are heavily utilized in the field of neuroscience. These benefits include their ease of use, lower levels of cytotoxicity, capability for large insert expression, and minimal lag time for gene expression. A variety of parameters should be considered when optimizing these methods to maximize expression efficiency and to minimize cellular toxicity. For lipid-mediated transfection as well as other chemical transfection methods, important parameters include the ratio of nucleic acid concentration to vector concentration, the ratio of nucleic acid concentration to cell number, the vector–nucleic acid association time, the cell culture conditions, and the gene expression time.

Viral transfection systems provide significantly higher transfection efficiencies for both nonmitotic neurons and mitotic astrocytes. Viruses can have up to 95% infection efficiencies with high levels of expression in target cells whether in dissociated culture, in brain slices, or *in vivo*. However, there are also major drawbacks to consider with this method of gene delivery, including the limitations on insert size, toxicity to cells, and the time and effort needed to construct recombinant viral vectors. Furthermore, a number of different viral vectors are utilized to transfect cells of the nervous system. Hence, it is important to choose the appropriate one based on the properties of the virus and the goals of the experiment. Adenovirus (AdV) has been the most commonly used viral vector and has the ability to transfect neurons and glia both *in vivo* and *ex vivo*. Expression of transgenes can occur days after infection but lasts for weeks to months. There are limitations on insert size that should be noted as well as the finding of cytotoxicity at high titers. Lentivirus systems also can target expression to both neurons and glia but have the added advantage of stable expression for months to years. Adeno-associated virus (AAV) can provide robust gene expression selectively for neurons *in vivo* and *ex vivo* and is much less toxic than other viral vectors. However, its use is limited to inserts that are smaller than 5 kb with expression delayed 2 weeks following infection. Other viral vectors used for gene delivery include herpes simplex virus (HSV), vaccinia virus, Sindbis virus, and Semliki Forest virus (SFV) (Table 1).

Physical transfection methods employ completely different mechanisms to penetrate the cellular membrane with nucleic acid. Electroporation systems utilize cell type optimized electrical currents to make transient pores in the membranes of cells allowing for surrounding nucleic acid to enter the cell. Because the electrical current can severely reduce cell viability, millions of neurons are required for a single transfection. However, despite this drawback, the transfection rate of surviving cells can be greater than 50%. In addition, methods have been developed for single cell electroporation of dissociated neurons in culture as well as for transfection in organisms. Biolistic systems use gas pressure as a means to penetrate cells in culture or in slices with genes prepackaged onto metal bids. Careful optimization of the number of bids and gas pressure is required to minimize cell damage. Microinjection involves utilizing a sharp, micron-tipped micropipette to directly puncture and inject nucleic acid into a single cell under a microscope. With mammalian cells, this procedure is labor intensive and requires a great deal of skill compared to other transfection methods.

Because of demands for noninvasive, spatiotemporal control of gene expression in the nervous system, many of the preceding methods have been investigated for their ability to transfect certain cell

Table 1 Comparison of different viral vector systems

	Adeno-associated virus	*Adenovirus*	*Herpes simplex virus*	*Lentivirus*	*Sindbis/Semliki Forest virus*
Insert size	<4.5 kb	<7.5 kb	<30–50 kb	<7–8 kb	<6.5 kb
Onset of gene expression	Days–weeks	Days	Hours	Days	Hours
Duration of expression	Stable	Transient	Stable	Stable	Transient
Cell type	Neuron and astrocyte	Neuron and astrocyte	Neuron	Neuron and astrocyte	Neuron
Cell toxicity	Low	High	Low	Low	High

types in a controlled manner. Viral transfection methodologies have especially been developed to better meet these demands. Currently, the natural tropism of viruses largely determines which cells are targeted for expression with the gene(s) of interest. However, the recent discovery and characterization of neuron- and glial-specific promoters have helped improve upon the spatial control of gene expression, especially *in vivo*. Inducible expression systems have complemented these approaches to allow for the temporal control of transgene expression. In addition to viral methods, chemical methods have also been investigated for targeted gene expression. The *in vivo* delivery of liposomes that are targeted to certain cell types may become an option for the spatial control of gene expression. Here, antibody-coupled liposomes have been shown to penetrate the blood–brain barrier and enter the central nervous system. This may allow for targeted gene expression based upon the unique expression of surface markers by different cells types within the nervous system.

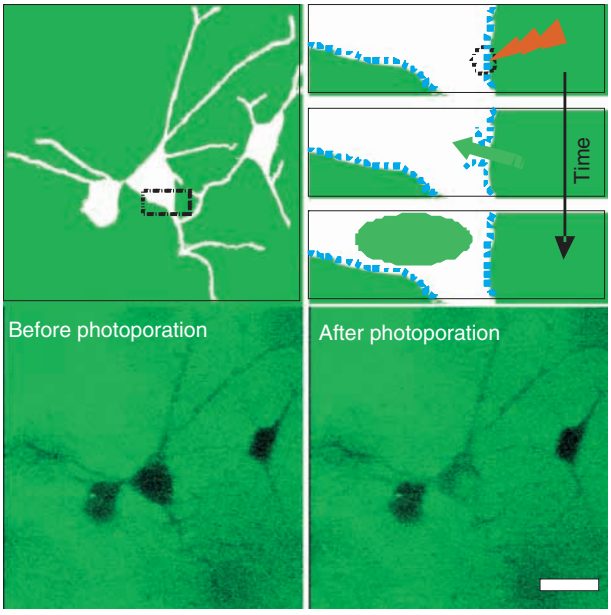

Figure 1 Two-photon laser poration successfully introduces an exogenous dye into neurons. Live hippocampal neurons are incubated with membrane-impermeable Lucifer Yellow and imaged prior to photoporation (lower left panel). A 2 μm × 2 μm area from part of the cell body of the central neuron is selected for photoporation (box, schematic upper left panel). A two-photon laser is briefly targeted to this area (red arrow, schematic upper right panel) allowing the Lucifer Yellow to enter the cell at the site of poration (schematic upper right panel). After poration, the neuron has taken up the Lucifer Yellow and become fluorescent near the site of photoporation (lower right panel). Scale bar = 20 μm.

Recently, a few research groups have developed a novel, single-cell transfection methodology called phototransfection or photoporation. This physical transfection methodology utilizes a microscope equipped with a femtosecond, pulsatile laser to form small poration sites within the plasma membrane of target cells (Figure 1). This allows bath-applied DNA or RNA to transiently enter the poration sites without detrimental affects on cell viability. By controlling the concentration of nucleic acid in the bath, the amount of nucleic acid that enters the cell can be approximated and modulated. Using this approach, astrocytes and neurons derived from diverse regions of the nervous system can be targeted. Different cellular regions, including the nucleus, cytoplasm, axon, and dendrites, can be uniquely selected for the localized expression of nucleic acid. Furthermore, subcellular regions including proximal and distal dendrites can be also be targeted. Multiple genes can be simultaneously expressed within a single cell or can be differentially expressed within different regions of the cell. In addition to the spatial control of gene expression, genes can be expressed at different time points within the same cell, either at the same region or at a different region

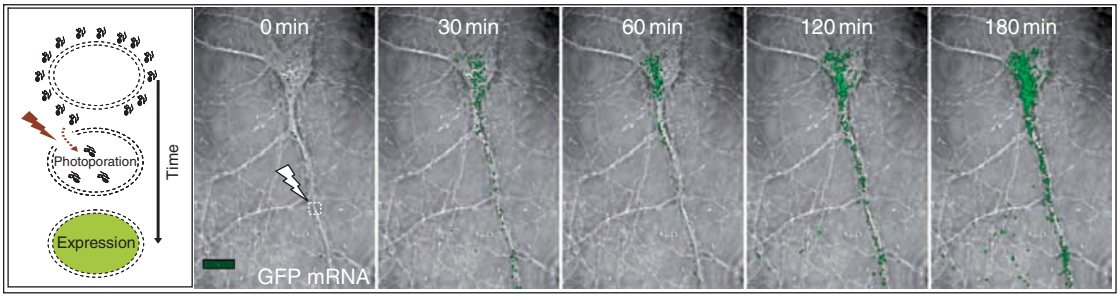

Figure 2 Phototransfection allows for the introduction of exogenous nucleic acid into neurons. Neurons are phototransfected in the presence of bath-applied capped green fluorescent protein (GFP) mRNA. The arrow represents the site of phototransfection. Differential interference contrast (DIC) images show the neuron prior to phototransfection and at indicated times after phototransfection. These images are overlaid with the GFP fluorescent signal. Scale bar = 20 μm.

of the cell. Currently, this method of gene transfection has been utilized to target gene expression within cultured primary cells of the nervous system. Efforts are under way to utilize this approach for the transfection of brain slices and for the transfection of single cells within organisms (Figure 2).

Many different gene delivery methodologies have been successfully used to study the nervous system. Currently, there is no single method that can allow for efficient, specific expression in all experimental preparations. Consequently, investigators must be aware of the advantages and drawbacks of each method and choose a method based on the experimental question and system being utilized. As the field of neuronal gene delivery moves forward, methodologies will be developed to restrict exogenous gene expression to specific cell types or pathways. This will better allow for the study of the nervous system.

See also: Engineering Viruses for CNS studies; Gene Therapy: Direct Viral Delivery; Glial Responses to Virus Infection; Viral Vectors in the CNS.

Further Reading

Barrett LE, Sul JY, Takano H, Van Bockstaele EJ, Haydon PG, and Eberwine JH (2006) Region-directed phototransfection reveals the functional significance of a dendritically synthesized transcription factor. *Nature Methods* 3(6): 455–460.

Dalby B, Cates S, Harris A, et al. (2004) Advanced transfection with lipofectamine 2000 reagent: Primary neurons, siRNA, and high-throughput applications. *Methods* 33(2): 95–103.

Deglon N and Hantraye P (2005) Viral vectors as tools to model and treat neurodegenerative disorders. *Journal of Gene Medicine* 7(5): 530–539.

Glover C, Bienemann A, Hopton M, Harding T, Kew J, and Uney J (2003) Long-term transgene expression can be mediated in the brain by adenoviral vectors when powerful neuron-specific promoters are used. *Journal of Gene Medicine* 5(7): 554–559.

Goetze B, Grunewald B, Baldassa S, and Kiebler M (2004) Chemically controlled formation of a DNA/calcium phosphate coprecipitate: Application for transfection of mature hippocampal neurons. *Journal of Neurobiology* 60(4): 517–525.

Haas K, Sin WC, Javaherian A, Li Z, and Cline HT (2001) Single-cell electroporation for gene transfer *in vivo*. *Neuron* 29(3): 583–591.

Hagemann C, Meyer C, Stojic J, et al. (2006) High efficiency transfection of glioma cell lines and primary cells for overexpression and RNAi experiments. *Journal of Neuroscience Methods* 156(1–2): 194–202.

Kariko K, Keller JM, Harris VA, Langer DJ, and Welsh FA (2001) *In vivo* protein expression from mRNA delivered into adult rat brain. *Journal of Neuroscience Methods* 105(1): 77–86.

Klimaschewski L, Nindl W, Pimpl M, Waltinger P, and Pfaller K (2002) Biolistic transfection and morphological analysis of cultured sympathetic neurons. *Journal of Neuroscience Methods* 113(1): 63–71.

Paterson L, Agate B, Comrie M, et al. (2005) Photoporation and cell transfection using a violet diode laser. *Optics Express* 13: 595–600.

Ohki EC, Tilkins ML, Ciccarone VC, and Price PJ (2001) Improving the transfection efficiency of post-mitotic neurons. *Journal of Neuroscience Methods* 112(2): 95–99.

Tinsley RB, Vesey MJ, Barati S, Rush RA, and Ferguson IA (2004) Improved non-viral transfection of glial and adult neural stem cell lines and of primary astrocytes by combining agents with complementary modes of action. *Journal of Gene Medicine* 6 (9): 1023–1032.

Washbourne P and McAllister AK (2002) Techniques for gene transfer into neurons. *Current Opinion in Neurobiology* 12(5): 566–573.

Watanabe SY, Albsoul-Younes AM, Kawano T, et al. (1999) Calcium phosphate-mediated transfection of primary cultured brain neurons using GFP expression as a marker: Application for single neuron electrophysiology. *Neuroscience Research* 33(1): 71–78.

Wellmann H, Kaltschmidt B, and Kaltschmidt C (1999) Optimized protocol for biolistic transfection of brain slices and dissociated cultured neurons with a hand-held gene gun. *Journal of Neuroscience Methods* 92(1–2): 55–64.

Numerical Intelligence: Neural Substrates

A Nieder, University of Tübingen, Tübingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Numbers can be used most flexibly to quantify, rank, and identify virtually everything that is imaginable. Although true counting and mathematics are cultural achievements that are bound to language, it has become evident during the past few decades that basic numerical skills are independent from language. Human adults, infants, and animals are able to nonverbally and approximately grasp the numerical properties of objects or events. Such nonverbal numerical systems are the precursors on which verbal numerical representations build up, and their neural foundation can be studied in animal models.

This article reviews the progress that has been made in our understanding of the neural underpinnings of numerical competence in human and nonhuman primates. It is structured according to the basic concepts numerical quantity, which refers to the empirical property cardinality (the size of a set, also termed 'numerosity'), and numerical rank, which refers to the property serial order.

Numerical Quantity (Cardinality)

Behavior

Animals can discriminate the cardinality of sets. During the past six decades, roughly half a dozen bird and mammal species have successfully been trained to perform several numerical tasks in different modalities. Importantly, animals can transfer their numerical knowledge to numerosities they have not been trained on, which indicates a true understanding of the concept of cardinalities and their sequential arrangement. Moreover, numerical intelligence is not induced by extensive training in the laboratory but can be demonstrated spontaneously in the wild. The data collected for different species indicate that quantity information is indeed exploited by animals in their natural habitat to draw informed choices, for example, in foraging situations or social interactions.

Also, preverbal human infants of several months of age have the capacity to represent cardinality, both in the visual and in the auditory domain. Infants can also engage in rudimentary arithmetic, which was first shown in experiments in which 5-month-old infants were shown basic addition and subtraction operations on small sets of objects.

Human adults can reliably compare the cardinality of sets under conditions that prevent or discourage verbal counting. In contrast to precise verbal counting, however, nonverbal discrimination performance is only inaccurate, or approximate. Some indigenous human cultures that lack number words or have a very restricted concept of verbal counting rely completely on nonverbal cardinality assessment (so-called 'one–two–many' system of 'counting'). Thus, humans without a linguistic number concept can only estimate the number of items by means of a nonverbal quantification system. Together, these studies provide evidence for an evolutionary ancient quantification system which operates independent from language and can thus be studied in animal models.

Numerosity-Selective Neurons

Recordings in monkeys demonstrated the capacity of single neurons to encode numerical quantity. In monkeys performing a visual delayed match-to-numerosity task (Figure 1(a)), the highest proportion of numerosity-selective neurons (Figure 1(c)) was found in the lateral prefrontal cortex (PFC), irrespective of co-varying nonnumerical parameters. In the posterior parietal cortex (PPC), numerosity-selective neurons were most abundant in the fundus of the intraparietal sulcus (IPS); they were few in other PPC areas or the anterior inferior temporal cortex (aITC) (Figure 1(b)). Neurons in a somatosensory-responsive region of the superior parietal lobule (part of area 5) have been reported to keep track of the number of movements, but in a movement-type-dependent manner (i.e., the neurons responded differently whether the monkey's movement was 'push' or 'turn'). Area 5 neurons were not encoding numerosity in visual displays.

Numerosity-selective neurons in the PFC and IPS were tuned for the number of items on a visual display; that is, they showed maximum activity to one of the five presented quantities – a neuron's 'preferred numerosity' (Figure 1(c)). All numerosity-selective neurons together formed a bank of overlapping numerosity filters (Figure 1(e)), thus mirroring the animals' behavioral performance (Figure 1(d)). Interestingly, the neurons' sequentially arranged overlapping tuning curves preserved an inherent order of cardinalities. This is important because numerosities are not isolated categories but, rather, exist in relation to one another (e.g., '3' is greater than '2' and less than '4'); they need to be sequentially ordered to allow meaningful quantity assignments.

The fact that IPS neurons require shorter latencies to become numerosity selective than PFC neurons suggests that the IPS might be the first cortical stage that extracts visual numerical information. Because PPC and PFC are functionally interconnected, that

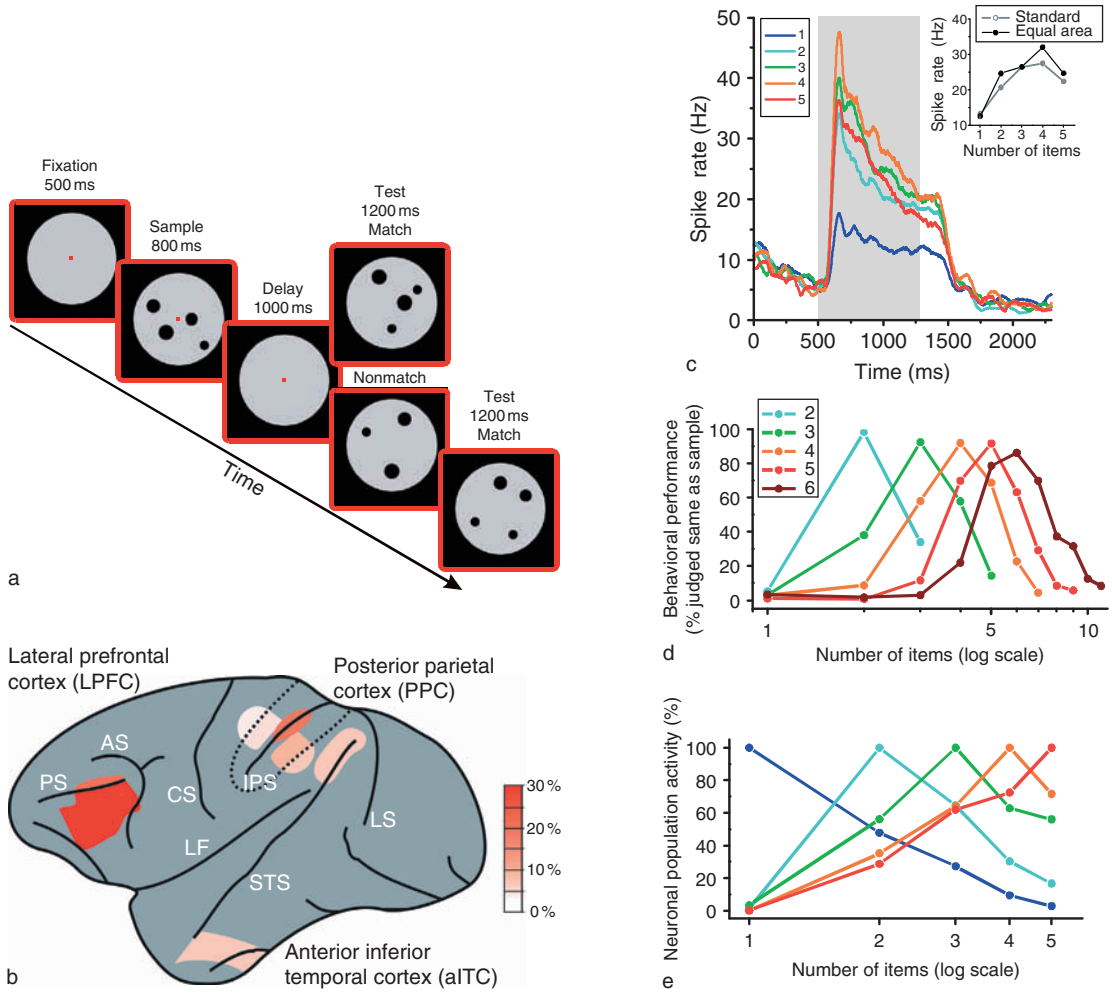

Figure 1 Representation of visual cardinality in rhesus monkeys. (a) Behavioral task. Monkeys performed a delayed match-to-numerosity task. They were required to extract the numerosity in a visual array, memorize it briefly, and match it to the one of two alternative test displays that showed the same numerosity (but with a different visual pattern). (b) Lateral view of the brain of a monkey shows the recording sites in lateral prefrontal cortex (LPFC), posterior parietal cortex (PPC), and anterior inferior temporal cortex (aITC). The proportions of numerosity-selective neurons in each area are color coded according to the color scale. As, arcuate sulcus; Cs, central sulcus; IPS, intraparietal sulcus; LF, lateral fissure; LS, lunate sulcus; PS, principal sulcus; STS, superior temporal sulcus. (c) Responses of a single neuron recorded from the PFC. The neuron showed graded discharge during sample presentation (interval shaded in gray, 500–1300 ms) as a function of numerosities 1 to 5 (color coded average discharge functions). The inset shows the tuning of the neuron and its response to different control stimuli. The preferred numerosity was '4' for this PFC neuron. (d) Behavioral numerosity discrimination functions of two monkeys. The curves indicate whether they judged the first test stimulus (after the delay) as containing the same number of items as the sample display. The function peaks (and the color legend) indicate the sample numerosity at which each curve was derived. Behavioral filter functions are skewed on a linear scale (not shown) but symmetric on a logarithmic scale. (e) The averaged single-cell numerosity tuning functions (from PFC) are also asymmetric on a linear scale (not shown) but symmetric after logarithmic transformation.

information might be conveyed directly or indirectly to the PFC, where it is amplified and maintained to gain control over behavior.

The response properties of numerosity-selective cortical cells can explain basic psychophysical phenomena in monkeys, such as the numerical distance and size effect (Figure 1(d)). The numerical distance effect states that it is easier to discriminate quantities that are numerically remote from each other (e.g., 2 vs. 6 is easier than 5 vs. 6), whereas the numerical size effect captures the finding that pairs of numerosities of a constant numerical distance are easier to discriminate if the quantities are small (e.g., 2 vs. 3 is easier than 5 vs. 6). The numerical distance effect results from the fact that the neural filter functions that are engaged in the discrimination of adjacent numerosities heavily overlap. As a result, the signal-to-noise ratio of the neural signal detection process is low, and the monkeys

would make many errors. On the other hand, the filter functions of neurons that are tuned to remote numerosities barely overlap, which results in a high signal-to-noise ratio and, therefore, good performance in cases in which the animal has to discriminate sets by a larger numerical distance.

The numerical size effect is based on the finding that neuronal tuning obeys Weber's law: the widths of the tuning curves (or neuronal numerical representations) increase linearly with preferred numerosities (i.e., on average, neurons become less precisely tuned as the preferred quantity increases). Hence, selective neural filters that do not overlap much are engaged if a monkey has to discriminate small numerosities (e.g., 1 and 2), which results in high signal-to-noise ratios and few errors for the discrimination. Conversely, if a monkey has to discriminate large numerosities (e.g., 4 and 5), the filter functions would overlap considerably. Therefore, the discrimination would show a low signal-to-noise ratio, which leads to poor performance.

The Scaling of Numerical Representations

As mentioned previously, the neurons' overlapping tuning curves are orderly arranged along a 'number line' (Figure 1(e)). However, what is the scaling scheme of such a number line and are neuronal numerical representations best described on a linear, or a nonlinear, possibly logarithmically compressed scale? The latter would be predicted if Fechner's law holds. Fechner's law states that the perceived magnitude (S) is a logarithmic function of stimulus intensity (I) multiplied by a modality- and dimension-specific constant (k). Since the behavioral discrimination and single-unit tuning functions might be regarded as the monkeys' behavioral and neural numerical representations (Figures 1(d) and 1(e)), the crucial question then concerns which scaling scheme would provide symmetric (i.e., Gaussian) probability density distributions. Both the performance and the single-unit data for numerosity judgments are better described by a compressed scale (Figures 1(d) and 1(e)) as opposed to a linear scale. Therefore, single-neuron representations of numerical quantity in monkeys obey Fechner's law.

Functional Imaging in Humans

In humans, the PFC and the parietal lobe, in particular the IPS, have long been regarded as the prime source for numerical competence. This view, however, was almost exclusively based on verbal and symbolic number tasks, which are only vaguely comparable to numerical magnitude estimation. If anatomical and functional similarities between the brains of monkeys and humans do exist, nonverbal cardinality should be processed in the human brain in equivalent areas as seen in monkeys. Indeed, such a corresponding blood oxygenation level-dependent (BOLD) activation in the IPS of humans has been found using functional magnetic resonance imaging (fMRI) adaptation with numerosities (Figure 2). Subjects were repeatedly presented with several visual displays of a fixed numerosity (e.g., 16 dots), without the requirement to discriminate them. The rationale of this protocol is the following: if any region of the brain contains a population of numerosity-selective neurons that are tuned to a specific number of dots and automatically detect numerical information, such a population of detectors should habituate (i.e., decrease its discharge) to numerosities that are repeatedly presented, whereas neurons tuned to other numerosities should not be affected. Such a habituation effect was then 'read out' by recording the event-related fMRI activation to a single deviating numerosity (e.g., 32 dots) that was presented at the end of a display sequence. In fact, the only region in the brain that was significantly habituated to numerosity was the horizontal segment of the IPS (Figure 2(a)). With this fMRI adaptation protocol, the average numerosity tuning curve of the underlying neural population could be traced indirectly in humans (Figure 2(b)). Similar to single neuron responses in monkeys, fMRI tuning curves in humans seem to be well described on a logarithmic scale, which suggests that the nonverbal number line in all primates is nonlinearly compressed.

Language allows humans to use symbolic representations, and numbers are the symbols that are employed when dealing with numerical information. Interestingly, numerical values that are cued by number symbols and set size have been shown to activate corresponding structures in the brain in humans. The IPS, which has been suggested to play a central role in basic quantity representations, is sometimes the only area specifically activated in simple number detection or comparison tasks. Eger et al. performed fMRI measurements on subjects who were asked to merely detect numerals, letters, or colors which were presented visually (written words and symbols) or acoustically (spoken words). To avoid confounds by response selection and associated cognitive states (e.g., attention), the authors analyzed the presentation of nontarget numerals (numerals that were not required to be detected) and compared it to nontarget letters or colors. The IPS was the only region that exhibited higher activation for numerals, both visually and acoustically (Figure 3(a)). Therefore, numerical activation in the IPS seems to be automatic (task independent), supramodal (visual and auditory), and notation independent (irrespective of whether numerals are spoken or written, presented in Arabic notation or spelled-out form). The IPS, however, is

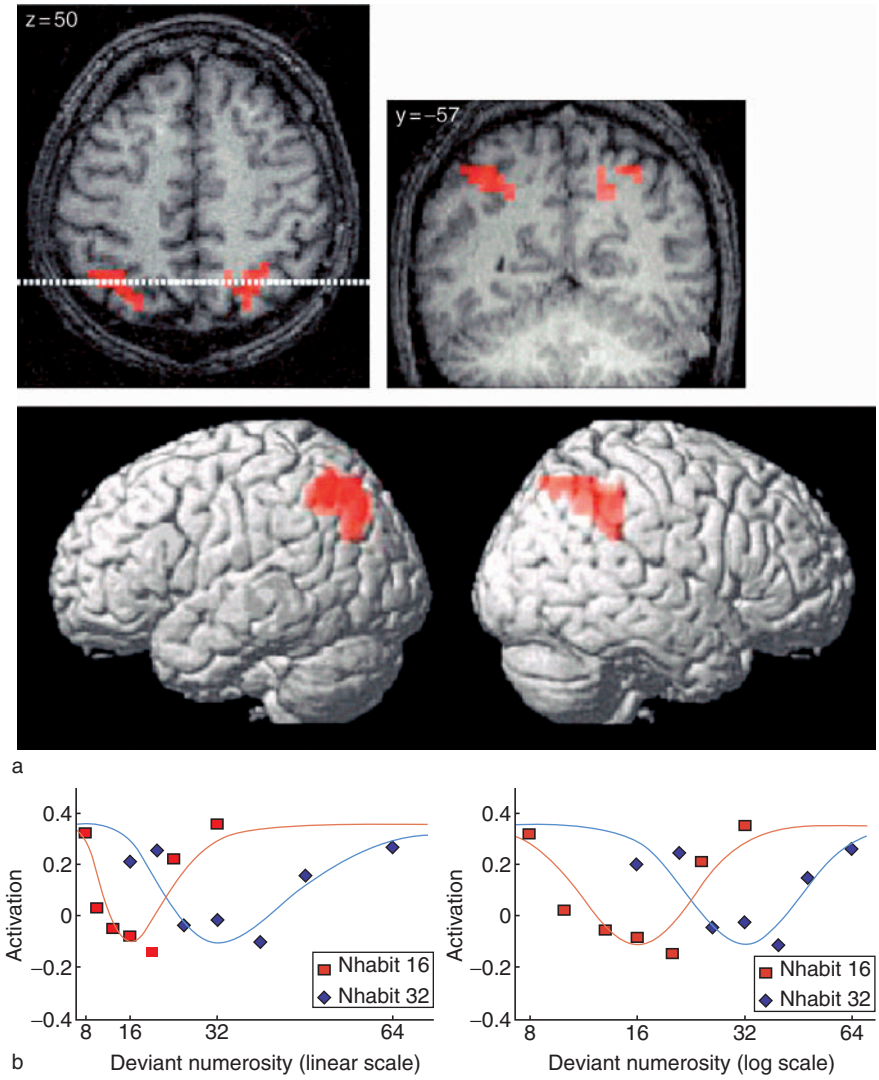

Figure 2 Functional MRI adaptation with numerosities in humans. (a) Regions in the human brain that responded to numerosity changes (i.e., exhibited numerosity selectivity). Colored areas in the axial (top left) and coronal (top right) sections and on the surface image indicate the IPS. (b) Activation difference between the habituation and deviant numerosities resulted in (inverted) numerosity tuning curves that were more symmetrically fitted on a logarithmic (right) than on a linear (left) scale. Adapted from Piazza M, Izard V, Pinel P, Le Bihan D, and Dehaene S (2004) Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron* 44: 547–555.

also engaged in more general magnitude judgments. Pinel et al. scanned subjects with fMRI while they compared Arabic numerals for luminance, font size, and numerical value. The authors observed strong overlap in the neural substrates in the three tasks. Number and size, but not luminance, activated a common parietal region, indicating toward an intimate relationship between spatial and numerical representations. This might mean that number-coding neurons are intermingled with other magnitude-coding neurons along the IPS.

Beyond mere encoding of quantity information *per se*, verbal numerical competence requires additional cognitive components. Dehaene and co-workers suggested that verbal counting and calculation engages two more parietal regions: a posterior dorsal parietal area, which is activated by shifts in spatial attention whenever subjects count, and a left angular gyrus area, which is related to linguistic processing. Moreover, simple calculation tasks (e.g., subtraction) typically activate a distributed network that involves parietal, prefrontal, and premotor cortices (Figure 3(b)).

Developmental studies confirm that impairments of arithmetical abilities correlate with abnormalities in the organization of parietofrontal networks, and the IPS in particular. Using voxel-based morphometry, Isaacs et al. compared the density of gray matter in adolescents who were born at equally severe grades of prematurity. Half of these subjects (with otherwise normal IQ) suffered from dyscalculia, and the

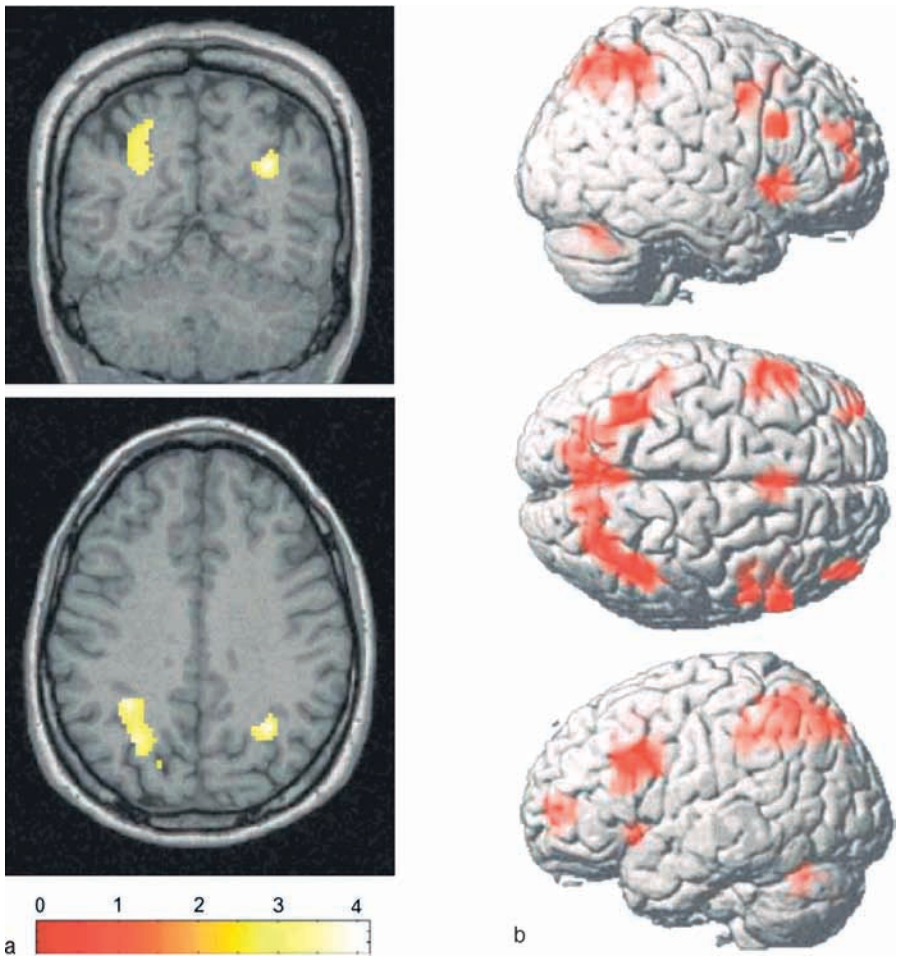

Figure 3 BOLD activation in humans to number symbols and arithmetics. (a) Cortical localization of supramodal responses to numbers (e.g., to both Arabic numerals and spoken number words). Compared to letters and colors, significant bilateral activation was only found deep in the horizontal IPS. (b) Activation pattern observed during subtraction. A covert subtraction task was contrasted with covert naming of the next number in the number line. It resulted in a distributed response pattern involving parietal, premotor, and prefrontal cortices. Adapted from Eger E, Sterzer P, Russ MO, Giraud AL, and Kleinschmidt A (2003) A supramodal number representation in human intraparietal cortex. *Neuron* 37: 719–725.

only region in the brain that showed reduced gray matter associated with arithmetical deficits was the left IPS. Therefore, dyscalculia in children might be the result of specific disabilities in basic numerical processing rather than the consequence of deficits in other cognitive abilities. Arithmetic deficits are also found in certain genetic conditions, such as the Turner syndrome (X monosomy), fragile X syndrome, and velocardiofacial syndrome. In these conditions, fMRI hypoactivation (decreased BOLD activation) was found in the IPS and wider parietofrontal networks.

Neuropsychological Studies

Brain damages in humans can cause relatively selective impairments in dealing with numbers, called acalculia. The most frequent lesions yielding acalculia involve the left inferior parietal area or the left parieto-occipito-temporal junction. In cases of left inferior parietal lesion, acalculia is frequently associated with agraphia, finger agnosia, and left–right confusion in a tetrad of deficits called Gerstmann's syndrome. In addition, calculation deficits have also been observed following left medial frontal, left and right frontal, and temporo-occipital lesions. A major issue in neuropsychological studies is the dissociation between deficits in language and numerical functions, particularly when the left hemisphere is affected. One study has indicated that syntactical mathematical operations (e.g., calculation rules such as solving equations according to brackets) are largely independently represented from linguistic syntax, at least in an adult brain.

Numerical Rank (Serial Order)

Behavior

Numerical rank is the second major numerical concept that shows biological precursors. List learning – the ability to encode and then retrieve an arbitrary list of items in their correct order – opened a window to study how the ordinal rank of objects is learned and stored by animals. Rhesus monkeys learned the serial order of items in several lists with up to seven items. Controlled manipulations of such learned lists showed that the monkeys acquired knowledge of each item's ordinal position.

Single-Unit Studies of Serial Order

Ordinal categorization of visual items requires both information about the rank of an item (e.g., based on temporal order) and its identity. The single-neuron correlate of temporal rank order information in visual lists has been studied in monkeys trained to observe and remember the order in which three visual objects appeared so that the animals could plan a subsequent triple-reaching movement in the same order (Figure 4(a)). Neurons in the ventrolateral PFC were selective for visual object properties, whereas neurons in the dorsolateral PFC were selectively tuned to the rank order of the objects irrespective of the visual properties of objects; for example, a rank-order selective neuron would be active whenever the second item of the shuffled lists appears (Figure 4(b)). A third class of neurons, found in the ventrolateral PFC, showed the most complex response that was characterized by integrating the objects' sensory and order information; such a neuron would only discharge whenever a certain object appeared at a given position in the sequence.

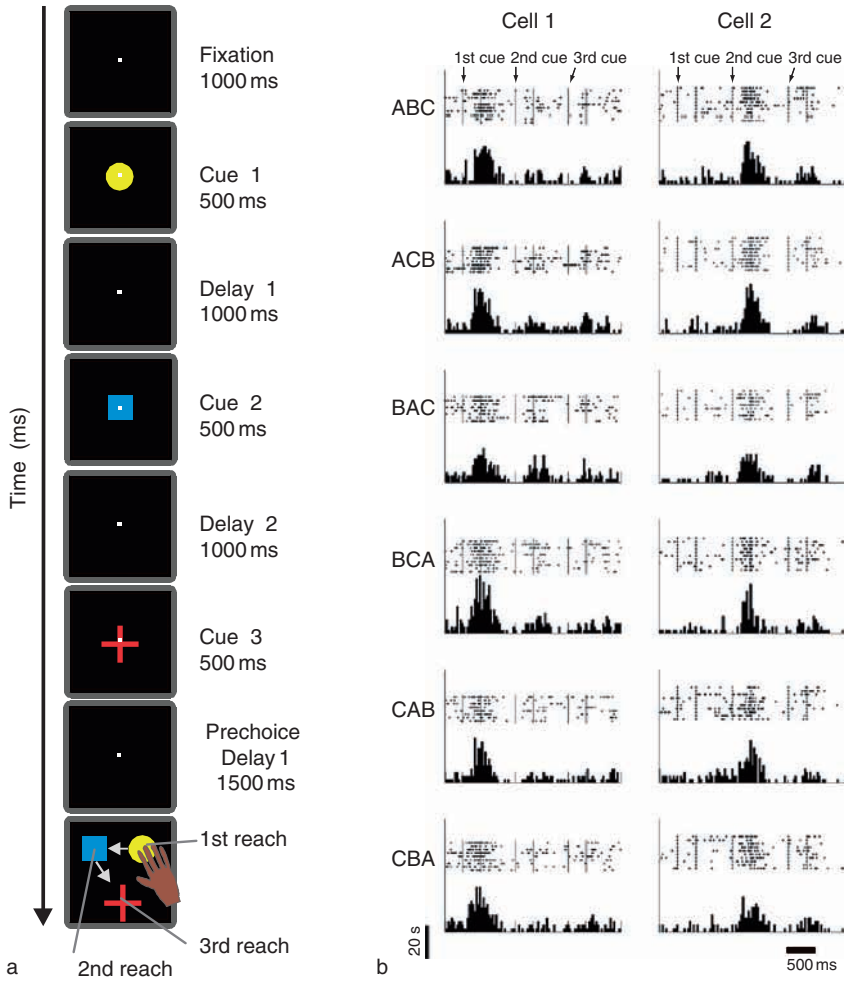

Figure 4 Temporal ordering task and single cell responses from the PFC. (a) Monkeys were required to observe and remember the order in which three visual objects appeared so that the animals could plan a subsequent triple-reaching movement in the same order. (b) Two single neurons encoding the first (cell 1) and the second rank (cell 2), irrespective of the order in which the three items (symbolized by letters ABC) appeared. Neural responses are shown in a dot-raster histogram (top; each dot represents an action potential) and averaged as peri-stimulus time histograms (bottom). Adapted from Ninokura Y, Mushiake H, and Tanji J (2004) Integration of temporal order and object information in the monkey lateral prefrontal cortex. *Journal of Neurophysiology* 91: 555–560.

The lateral PFC is an ideal region in the brain to encode both sensory object properties and rank-order information because it receives massive sensory input from the temporal and parietal lobes, and it projects to premotor and motor areas of the frontal lobe. As a result, neurons that encode the ordinal position of task-related hand or eye movements have been found frequently in prefrontal and a variety of motor-related cortical areas in trained monkeys, such as the frontal eye field (FEF), the presupplementary motor areas (pre-SMAs), the supplementary motor area (SMA), the caudate nucleus, the anterior cingulate cortex (CGa), and even the primary motor cortex (M1). Motor-related areas such as M1, SMA, pre-SMA, and FEF may receive numerical information that has been computed on earlier stages of the cortical hierarchy to perform appropriate serial-order actions.

Human Studies

Based on patient studies, the lateral PFC has been implicated in maintaining temporal order information, which is an integral aspect of episodic memory. It is well-known that damage to the human frontal cortex causes impairment in tasks that require recall of the temporal order of stimuli. A similar ordering impairment has resulted from lesioning the dorsolateral frontal cortex in monkeys, which supports the view that the dorsolateral PFC in primates is important in maintaining information about the order of events. Functional MRI studies of humans show that prefrontal and parietal cortices are more strongly activated for order information (e.g., the order of words in a list) than for item information (the presence of words in a list). In addition, the lateral frontoparietal areas, the basal ganglia, and the cerebellum are preferentially involved in ordinal control of hand movements.

See also: Animal Intelligence: The Search for Animal Intelligence; Cognition: An Overview of Neuroimaging Techniques; Executive Function and Higher-Order Cognition: Neuroimaging; Neuroimaging; Parietal Cortex and Spatial Attention; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Reasoning and Problem Solving: Models.

Further Reading

Brannon EM and Terrace HS (1998) Ordering of the numerosities 1 to 9 by monkeys. *Science* 282: 746–749.

Dehaene S (1997) *The Number Sense*. Oxford: Oxford University Press.

Dehaene S, Molko N, Cohen L, and Wilson AJ (2004) Arithmetic and the brain. *Current Opinion in Neurobiology* 14: 218–224.

Eger E, Sterzer P, Russ MO, Giraud AL, and Kleinschmidt A (2003) A supramodal number representation in human intraparietal cortex. *Neuron* 37: 719–725.

Gordon P (2004) Numerical cognition without words: Evidence from Amazonia. *Science* 306: 496–499.

Hauser MD, MacNeilage P, and Ware M (1996) Numerical representations in primates. *Proceedings of the National Academy of Sciences of the United States of America* 93: 1514–1517.

Nieder A, Diester I, and Tudusciuc O (2006) Temporal and spatial enumeration processes in the primate parietal cortex. *Science* 313: 1431–1435.

Nieder A, Freedman DJ, and Miller EK (2002) Representation of the quantity of visual items in the primate prefrontal cortex. *Science* 297: 1708–1711.

Nieder A and Miller EK (2004) A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences of the United States of America* 101: 7457–7462.

Ninokura Y, Mushiake H, and Tanji J (2004) Integration of temporal order and object information in the monkey lateral prefrontal cortex. *Journal of Neurophysiology* 91: 555–560.

Piazza M, Izard V, Pinel P, Le Bihan D, and Dehaene S (2004) Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron* 44: 547–555.

Pica P, Lemer C, Izard V, and Dehaene S (2004) Exact and approximate arithmetic in an Amazonian indigene group. *Science* 306: 499–503.

Wiese H (2003) *Numbers, Language, and the Human Mind*. Cambridge, UK: Cambridge University Press.

Wynn K (1992) Addition and subtraction by human infants. *Nature* 358: 749–750.

Nutrient Sensing: Carbohydrates

B E Levin, Veterans Affairs Medical Center, East Orange, NJ, USA

Published by Elsevier Ltd.

Introduction

Carbohydrates in the body exist largely either as glucose in the fluid compartments, including the blood, cerebrospinal fluid, and interstitial fluid, or as glucose complexed in the form of glycogen. The regulation of these circulating levels and stores is an integral part of the overall regulation of energy homeostasis in the body – that is, the balance between intake, expenditure, and storage of calories. Where intake exceeds expenditure, excess calories are stored as glycogen in liver, muscle, and brain and as fat in adipose depots. In addition to these basic functions, ingested nutrients must transit the gastrointestinal tract where they are digested, absorbed, and then assimilated into various tissues. Many of these functions are mediated by the autonomic nervous system, the function of which is modulated by an integrated network of specialized glucose sensors in the brain and periphery. The term 'glucose sensing' is used here to define the way in which these sensors utilize glucose as a signaling molecule to control their activity and output as monitors of the status of glucose in the body. In fact, glucose-sensing elements throughout the body contribute to virtually all aspects of energy homeostasis. Very low plasma and brain glucose levels stimulate, whereas high levels inhibit, energy intake. Glucose sensing is also involved in various aspects of assimilation and storage of energy, as well as its expenditure as heat (thermogenesis). In the brain, most neurons utilize glucose primarily to fuel metabolic requirements associated with changes in their activity. Glucose-sensing neurons represent a specialized class of neurons which have evolved to use glucose as a signaling molecule to control their firing rate and transmitter release. However, glucose rarely acts alone on glucose-sensing neurons. Many glucose-sensing neurons also integrate the inputs of a variety of hormones such as leptin and insulin, and of metabolic substrates such as lactate, ketone bodies, and fatty acids; they also integrate the neural inputs from peripheral sensors to control their membrane potential, neuronal activity, and gene transcription (Figure 1). Thus, although glucose sensing is the primary subject of this article, since these neurons respond to a host of other inputs from the periphery it is virtually impossible to consider their ability to sense glucose in isolation.

Glucose and the Components of Energy Homeostasis

Energy Intake

In 1953, Jean Mayer proposed the 'glucostatic hypothesis' whereby arteriovenous differences and/or glucose oxidation were monitored by specialized hypothalamic neurons as a means of controlling food intake. He believed that the hypothalamus was the site of such sensors because prior studies had shown that damage to the ventromedial hypothalamus caused both humans and animals to eat voraciously, while damage to the lateral hypothalamus caused animals to cease eating. To date, it is still unclear whether changes in blood glucose levels within the physiological range actually play a primary role in the regulation of ingestion. On the other hand, there is no question that rapid lowering of blood glucose levels (hypoglycemia) or administration of drugs which block intracellular glucose utilization (glucopenia) can and do produce feelings of hunger which drive the individual into a behavioral sequence required to seek and ingest food. However, acute drops in blood glucose rarely occur in nature and are mainly caused by inadvertent overdoses of insulin in diabetic individuals. A more physiological role for reduced glucose levels in the control of ingestion is suggested by studies showing that small (~10%) dips in blood glucose levels precede the onset of some individual meals. But, it is uncertain whether such dips are a physiological stimulant of intake or simply an epiphenomenon of underlying circadian rhythms of hormones and neural functions. On the other hand, infusions of glucose into the brain or portal vein of the liver have variably been reported to reduce food intake. Yet it is unclear whether the levels of glucose required to produce satiety are either physiologically relevant or simply act by providing a sufficient caloric load to produce satiety. Ingestion is a complicated process which requires behavioral arousal, the development of hunger, increased motor activity to seek and procure food, behavioral feedback relating to the rewarding properties of food, and finally satiety. Regardless of whether glucose can act alone to alter these various aspects of ingestive behavior, changes in ambient glucose concentrations within the physiological range clearly can alter the activity of brain glucose-sensing neurons that are known to be involved in all of these aspects of feeding (Figure 2).

Energy Assimilation

Ingested macronutrients must be absorbed and assimilated to generate energy to fuel the body's

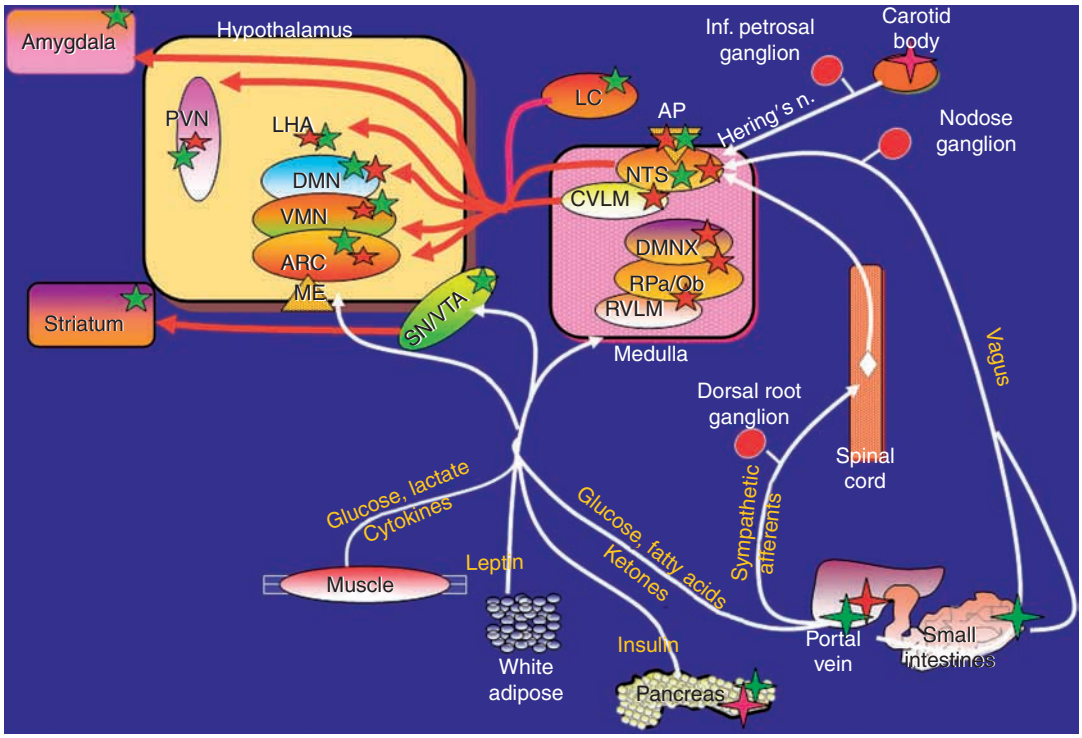

Figure 1 Location of peripheral and central glucose sensors and their connections. Vagal, glossopharyngeal, and sympathetic afferents from metabolic sensors (four-pointed stars) in the portal vein, carotid body (note Hering's nerve), and intestines converge on nucleus tractus solitarius (NTS) glucose-sensing neurons (five-pointed stars) in the medulla. These afferent inputs are integrated by NTS neurons along with metabolic and hormonal signals, which are transported across the blood–brain barrier or cross through fenestrated capillaries in the area postrema (AP) adjacent to the NTS. In addition to the NTS, glucose-sensing neurons in the dorsal motor nucleus of the vagus (DMNX) and the caudal and rostral ventrolateral medulla (CVLM, RVLM) also integrate these incoming signals and project to various hypothalamic nuclei – the arcuate (ARC), ventromedial (VMN), dorsomedial (DMN), paraventricular (PVN) – and to the lateral hypothalamic area (LHA) and the amygdala. Metabolic-sensing serotonin neurons in the raphe pallidus and obscurus (RPa/Ob) project to sympathetic preganglionic neurons in the intermediolateral cell column of the spinal cord. Metabolites and hormones also interact with glucose-sensing neurons in the hypothalamus and with dopamine neurons in the substantia nigra (SN) and ventral tegmental area (VTA). Neurons in the ARC project to additional sensing neurons in the PVN. SN neurons project to the striatum for control of motor activity and VTA neurons project to limbic areas involved in reward and motivation. ME, median eminence. Green stars are glucose-excited glucose sensors and red stars are glucose-inhibited glucose sensors.

metabolic needs. Glucose is a major source of such energy but requires insulin for uptake and incorporation into peripheral tissues. Glucose alters its own uptake and utilization by directly stimulating the release of insulin from the pancreatic β cell, the prototypic glucose-sensing cell in the body. Insulin release occurs when adenosine triphosphate (ATP) is generated by the sequential glycolysis and oxidation of glucose. An increased ratio of ATP/ADP leads to the inactivation of an ATP-sensitive K^+ (K_{ATP}) channel which depolarizes the cell membrane, leading to calcium influx and insulin release. This process is enhanced by vagal, and inhibited by sympathetic, inputs to the β cell. Both of these neural inputs are regulated by glucose acting on central glucose-sensing neurons in brain areas such as the lateral hypothalamic area (LHA), the arcuate (ARC), paraventricular (PVN), dorsomedial (DMN), and ventromedial (VMN) nuclei of the hypothalamus and the nucleus tractus solitarius (NTS), rostral ventrolateral medulla (RVLM), raphe pallidus (RPa) and raphe obscurus (ROb), and dorsal motor nucleus of the vagus (DMNX) in the medulla (Figures 1 and 2). Insulin is an anabolic hormone in the periphery in that it promotes the uptake of glucose into muscle and other organs. Insulin also inhibits breakdown of glycogen (glycogenolysis) in the liver and the release of fatty acids (lipolysis) from adipose depots. Thus, insulin is the intermediary through which glucose controls its own peripheral metabolism.

Energy Expenditure

Assimilation of nutrients also leads to the generation of heat, a process known as diet-induced thermogenesis. In humans, this occurs primarily in muscle. Rodents

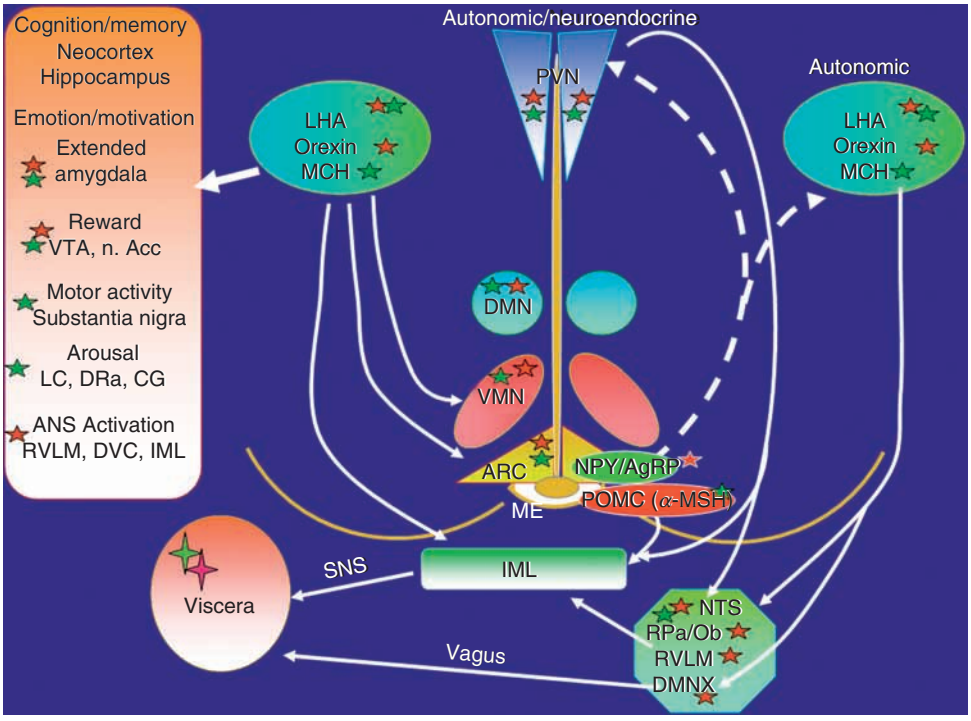

Figure 2 Efferent connections of glucose-sensing neurons involved in the regulation of energy homeostasis. Glucose-sensing neurons (green stars, glucose-excited neurons; red stars, glucose-inhibited neurons) are localized throughout the forebrain, hypothalamus, and brain stem. Arcuate (ARC) neuropeptide Y (NPY) neurons and proopiomelanocortin (POMC) neurons (which produce α-melanocyte-stimulating hormone (α-MSH)) are prototypic glucose-sensing neurons having the primary function of regulation of energy homeostasis. They project to the paraventricular nuclei (PVN) and to metabolic-sensing neurons (heavy dotted lines) orexin and melanin-concentrating hormone (MCH) in the lateral hypothalamic area (LHA). Some ARC POMC neurons also project to intermediolateral (IML) cell column preganglionic sympathetic nervous system (SNS) neurons in the spinal cord. PVN neurons project to the median eminence (ME) and pituitary to modulate neuroendocrine outputs and to sympathetic and parasympathetic output neurons in the medulla for control of visceral functions involved in energy homeostasis. In contrast to the relatively limited targets of ARC NPY and POMC neurons, the orexin and MCH LHA neurons project widely to brain areas involved in a large variety of behavioral and physiological functions, some of which are involved in the regulation of various aspects of energy homeostasis (VMN, ventromedial nucleus; NTS, nucleus tractus solitarius; RPa/Ob, raphe pallidus and obscurus; RVLM, rostral ventrolateral medulla; DMNX, dorsal motor nucleus of the vagus). Some of these targets also contain glucose-sensing neurons (VTA, ventral tegmentum area; n. Acc, nucleus accumbens; LC, locus coeruleus; DRa, dorsal raphe nucleus; CG, central gray; DVC, dorsal vagal complex).

also use brown adipose tissue, a specialized form of adipose tissue in which mitochondrial production of ATP is uncoupled from the oxidation of fatty acids leading to the generation of heat. Exercise is an additional source of thermogenesis. Glucose acts as a major stimulant of diet-induced thermogenesis and as a substrate to fuel exercise-induced heat production. This glucose-induced thermogenesis is mediated in part by direct effects of glucose on glucose-sensing neurons in the PVN, ARC, RPa/ROb, and RVLM which lead to activation of sympathetic efferents to thermogenic organs (Figure 2).

Peripheral Glucose Sensors

Aside from the pancreatic β cell, peripheral glucose sensors reside in the portal vein, the small intestines, and the glomus cells of the carotid body. While we know a great deal about the mechanisms by which β cells and glucose-sensing neurons monitor glucose levels, less is known about the mechanisms utilized by other peripheral sensors. It is known that glucose-activated signals are transmitted via vagal, glossopharyngeal, and sympathetic afferents which terminate on glucose-sensing NTS neurons, which integrate these neural signals with the effects of glucose acting directly on these same neurons (Figure 1). Portal vein sensors appear to transmit their signals via different sets of afferents, depending upon the levels of glucose. Raising portal vein glucose levels leads to a decrease in vagal afferent discharges that impinge upon NTS neurons, and these same NTS neurons are directly inhibited by glucose. On the other hand, lowering portal vein glucose levels by producing systemic hypoglycemia activates sympathetic (but not vagal) afferents. These afferents are required for the initiation of the

counterregulatory release of epinephrine from the adrenal medulla, norepinephrine from sympathetic nerves, and glucagon from the pancreatic α cells. Collectively these act in concert to raise blood glucose levels by stimulating glycogenolysis and promoting hepatic glucose production. The glomus cells of the carotid body are activated by both low glucose and low oxygen levels, and they communicate with the brain via glossopharyngeal parasympathetic afferents in Hering's nerve. Much like pancreatic β cells, intestinal glucose-sensing neurons and cells appear to utilize the K_{ATP} channel to sense glucose and they presumably communicate with the brain via vagal afferents (Figure 1). The L cells of the gut are an example of glucose-sensing cells which release peptides such as glucagon-like peptide-1 (GLP-1) in response to rising glucose levels. Such peptides may act on local gut vagal afferent receptors directly or may be released into the bloodstream to alter glucose homeostasis by acting on organs such as the pancreas, or by stimulating receptors on hepatic portal sensors.

How Do Neurons Sense Glucose?

Glucose Enters the Brain by Several Routes

Glucose enters most peripheral organs by diffusion through fenestrations between vascular endothelial cells. In the brain, endothelial cells in microvessels have tight junctions which prevent such diffusion. Thus, glucose enters the brain by facilitated transport across this blood–brain barrier via the GLUT1 glucose transporter (Figure 3). As a result, a gradient is maintained whereby brain interstitial glucose levels are held at ~20–30% of blood levels during normal or euglycemic conditions. This ratio falls to as low as 10–15% during systemic hypoglycemia. Brain glucose levels take 2–5 min to equilibrate after changes in blood glucose levels. Glucose is also transported across the choroid plexus into the cerebrospinal fluid with a delay of about 15 min. From there it can diffuse across the ependymal lining of the ventricles into the adjacent interstitial space of the brain. Glucose may also diffuse directly from the blood into the brain

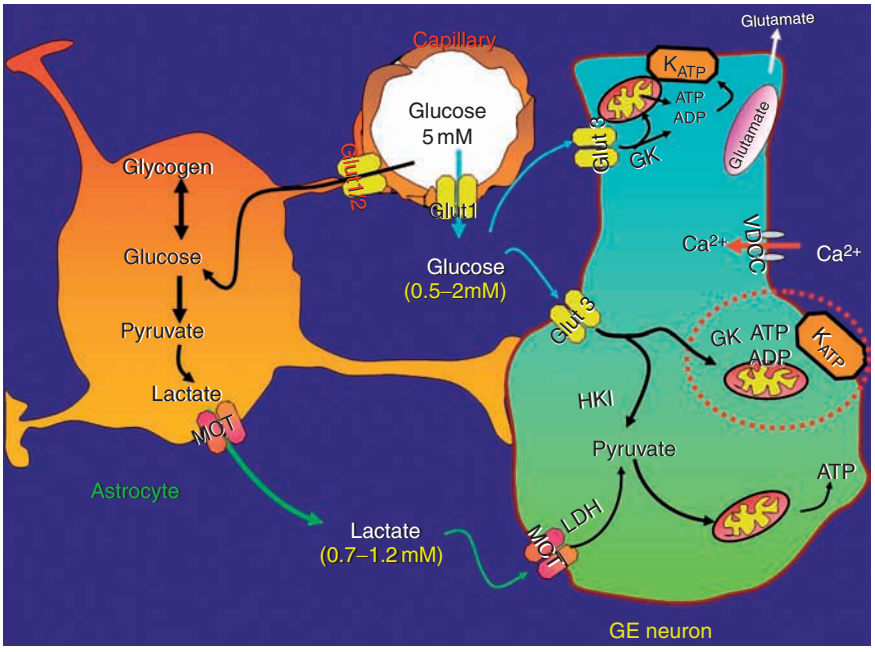

Figure 3 Hypothetical model of the mechanism of neuronal glucose sensing in a glucose-excited (GE) neuron. Glucose is transported across the endothelial cells of brain microvessels by GLUT1 transporters into the interstitial space, which maintains glucose levels at ~20–30% of blood levels. All neurons transport glucose by GLUT3 transporters. Hexokinase I (HKI) initiates glycolysis and glucose oxidation, which generates ATP for the ongoing metabolic demands of the cell. Glucose-sensing neurons also express glucokinase (GK), which is co-localized with mitochondria adjacent to K_{ATP} channels in the cell membrane. The GK-mediated generation of ATP occurs by oxidation of glucose in a concentration-dependent manner, which raises the ATP:ADP ratio and binding of ATP to the K_{ATP} channel. This binding inactivates (closes) the channel, leading to membrane depolarization and entry of extracellular calcium through a voltage-dependent calcium channel (VDCC) and propagation of an action potential. Glucose uptake and GK-mediated production of ATP and K_{ATP} channel inactivation can occur independently at axon terminals, leading to release of transmitters such as glutamate and γ-aminobutyric acid (GABA). Astrocytes have their foot processes adjacent to microvessels and they transport glucose via GLUT1 or GLUT2 transporters from the blood into the astrocytes where it is stored as glycogen. Glycogenolysis produces lactate, which is exported via monocarboxylate transporters (MCT) to the extracellular space, where it can be taken up by neurons through MCTs and converted to pyruvate by lactate dehydrogenase (LDH). Pyruvate can then be oxidized to produce an alternate source of ATP to fuel the metabolic needs of the cell.

at circumventricular organs such as the area postrema, which contains fenestrated capillaries (Figures 1 and 2). Diffusion from the area postrema can influence glucose-sensing neurons in the adjacent NTS. On the other hand, while the median eminence also contains fenestrated capillaries, glucose probably does not diffuse to the adjacent ARC because of barriers imposed by specialized glia and tanycytes, which functionally separate the two structures. Thus, depending upon the proximity of neurons to the ventricles or circumventricular organs and the rate of local glucose transport (which varies in different brain regions), glucose-sensing neurons can be exposed to glucose from up to three separate sources, each with its own temporal- and concentration-dependent patterns. Since such neurons also receive indirect neural inputs from peripheral glucose sensors, they act as comparators, integrating the arriving glucose-related signals having differing strengths and temporal patterns. The resulting effect on their membrane potential determines their firing rate and level of transmitter and neuropeptide release.

Direct Effects of Glucose on Glucose-Sensing Neurons

Glucose is transported into most neurons by the high-affinity, high-capacity GLUT3 transporter. Because of these kinetic properties, glucose transport into the neuron remains relatively constant across the physiological range of glucose levels experienced by the brain. Intraneuronal glucose is phosphorylated by hexokinase I, a high-affinity hexokinase which is saturated under physiological conditions and is also inhibited by its end product, glucose-6-phosphate. Thus, because of their kinetics, neither GLUT3 nor hexokinase I is rate limiting in the control of glucose metabolism or glucose sensing. For this purpose, glucose-sensing neurons have evolved an alternate pathway by which glucose acts in a concentration-dependent manner to generate ATP and regulate membrane potential and firing rate. There are two broad categories of glucose-sensing neurons; glucose-excited (GE) neurons increase, while glucose-inhibited (GI) neurons decrease, their activity as ambient glucose levels rise. They reverse this pattern as glucose levels fall. GE and GI neurons are located throughout the neuraxis and different areas contain differing percentages of each type of glucose sensor (Figures 1 and 2).

As with all other neurons, GE and GI neurons utilize GLUT3 and hexokinase I to provide for their ongoing metabolic needs. However, many glucose-sensing neurons also utilize glucokinase (GK; hexokinase IV) as the rate-limiting step in glucose sensing. The affinity of GK for glucose in the brain, as in the pancreas, is approximately 6–11 mM. Also, GK is not subject to end-product inhibition as is hexokinase I. At least 65% of GE and 45% of GI neurons express GK mRNA. In those GK-expressing neurons, pharmacological inhibition of GK activity decreases activity in GE neurons and increases activity in GI neurons held at glucose concentrations comparable to those seen in the brain (2.5 mM) under slightly hyperglycemic conditions in the periphery (~8 mM). On the other hand, pharmacological activation of GK inhibits the activity of GI neurons and enhances the activity of GE neurons held at glucose concentrations comparable to those seen in the brain (0.5 mM) under systemic hypoglycemic conditions (~3.5 mM). Interestingly, although GK mRNA cannot be demonstrated in all glucose-sensing neurons, an 80% reduction in GK mRNA in primary cultures of ventromedial hypothalamic neurons almost completely eliminates the ability of both GE and GI neurons to sense and respond to glucose. These data strongly suggest that GK is a critical regulator of neuronal activity in many GE and GI neurons.

Glucose-excited neurons Because of their similarity to pancreatic β cells, we know a great deal about the way in which GE neurons sense glucose (Figure 3). The K_{ATP} channel is the final common pathway for glucose sensing in both. It is an octomeric protein composed of four Kir6.2 pore-forming units for potassium and four sulfonylurea receptor subunits. Oxidation of glucose increases the ratio of ATP/ADP, causing ATP to bind to the channel. Binding of either ATP or sulfonylureas inactivates (closes) the channel and depolarizes the cell membrane. Depolarization is followed by influx of extracellular calcium through a voltage-dependent calcium channel, which in turn is often (but not always) followed by an increase in action potential frequency. Importantly, glucose-induced closure of the K_{ATP} channel at nerve terminals on some GE neurons can also release neurotransmitters independently of action potentials propagated from the cell body (Figure 3). While GK is probably the primary regulatory step in GE (and GI) neuronal glucose sensing, it is much less abundant than hexokinase I and is unlikely to contribute significantly to the overall, well-buffered pool of intracellular ATP which regulates K_{ATP} channel activity. Thus, GK may be compartmentalized in a microenvironment with mitochondria close to the K_{ATP} channels in the plasma membrane where ATP generated through the GK pathway can control channel activity (Figure 3).

Glucose-inhibited neurons While the K_{ATP} channel appears to be the final common pathway involved in GE neuron glucose sensing, much less is known about the way in which GI neurons sense glucose. In many ways, GI neurons are similar to pancreatic glucagon-secreting α cells. Both express GK and both are also inhibited as glucose levels rise and are activated as glucose levels fall. A Cl^- channel, the Na^+–K^+ ATP pump, and an ATP-responsive K^+ channel have all been proposed as possible final common pathways, but the actual effector of GI glucose sensing is unknown.

Alternate mechanisms of neuronal glucose sensing It is likely that all neurons do not depend upon GK as the regulator of glucose sensing. As with the rat pancreatic β cell, GLUT2 has been proposed as a rate-limiting step in neuronal glucose sensing. However, GLUT2 is in very low abundance in the brain and less than 30% of GE or GI neurons express GLUT2 mRNA. On the other hand, almost 70% of VMN glucose-sensing neurons co-express insulin receptor and GLUT4 mRNA. In peripheral tissues, occupation of the insulin receptor causes translocation of GLUT4 to the cell membrane, where it binds and transports glucose into the cell. This mechanism is required for glucose utilization in peripheral tissues. On the other hand, there is little convincing evidence that the brain requires insulin for glucose uptake and metabolism. However, the co-expression of insulin receptor and GLUT4 on glucose-sensing neurons suggests that insulin can act to enhance their glucose uptake under certain circumstances. For example, insulin may facilitate glucose transport during systemic hypoglycemia when brain glucose levels are lowered. Thus, insulin has little effect on the activity of ARC GE neurons at physiological glucose concentrations (2.5 mM) but causes increased neuronal activity when ambient glucose levels are lowered to those seen in the brain (0.5 mM) during hypoglycemia. In addition to this mechanism, adenosine monophosphate (AMP)-activated kinase (AMPK) is a well-described energy sensor in peripheral organs and is also localized in hypothalamic areas containing glucose-sensing neurons. Pharmacological activation of AMPK during systemic hypoglycemia enhances the systemic counterregulatory responses which normally mobilize glucose from peripheral tissues. Conversely, hypothalamic AMPK activity is regulated by ambient glucose levels. Thus, AMPK may function as an alternate emergency glucose sensor in some neurons. Finally, the sodium–glucose cotransporter may function as an alternate regulator of glucose sensing (as it appears to do in some intestinal glucose-sensing L cells); drugs which inhibit its activity decrease the responsiveness of GE neurons to glucose. However, the mRNA for this transporter is expressed in only a very few of such neurons. Thus, while the majority of glucose-sensing neurons utilize GK as their primary regulator, there are likely to be alternative mechanisms which function in these and other non-GK-expressing neurons, to allow neurons to sense and respond to glucose, particularly when glucose supply is limited.

Indirect Effects of Glucose on Glucose-Sensing Neurons

Glucose-sensing neurons do not function in isolation. They are surrounded by and provided with metabolic support by astrocytes, which readily take up and store glucose as glycogen (Figure 3). Glycogen is hydrolyzed to release lactate into the extracellular space from which it is taken up by neurons via monocarboxylate transporters (MCTs), converted to pyruvate by lactate dehydrogenase, and then oxidized to provide ATP. By this mechanism, lactate can reverse the inhibition of GE neuronal activity which occurs at low ambient glucose levels. To date, it remains unresolved what role astrocyte-derived lactate production plays in the support of non-glucose-sensing neurons and signaling in glucose-sensing neurons under physiological conditions.

Glucose-Sensing Neurons as Metabolic Sensors

GE and GI neurons have differing properties with regard to their sensitivity to glucose and to other hormones and metabolic substrates. GE neurons are more sensitive to glucose than are GI neurons, but the range over which both respond to glucose is also a function of their anatomic location. For example, GE neurons in the VMN are maximally activated over a range of 0.1–2.5 mM glucose, while ARC GE neurons continue to increase their activity up to 5 mM glucose. These differences in sensitivity may reflect inputs from surrounding astrocytes or neurons but might also be due to the fact that the K_{ATP} channel has different proportions of sulfonylurea-binding subunits, which could alter their sensitivity to glucose. Also, because the K_{ATP} channel is subject to modulatory effects of G-protein-coupled receptors, phosphoinositol 3-kinase (as modulated by both leptin and insulin), and levels of intracellular long-chain CoA molecules derived from fatty acids, the milieu in which a given neuron finds itself combines with its intrinsic metabolic and physiological characteristics to determine how it will respond to glucose. Thus, glucose-sensing neurons are really metabolic sensors in which the combined input of metabolites, peptides, transmitters, and hormones can affect both their short-term and long-term ability to sense and respond to glucose.

Role of Glucose-Sensing Neurons in the Regulation of Energy Homeostasis

The processes of ingestion, assimilation, expenditure, and storage of calories involved in the control of energy homeostasis are all ultimately controlled by the dialogue between the brain and the periphery. As with most neural functions, control of energy homeostasis resides within a distributed network of neural centers distributed throughout the brain (Figures 1 and 2). Because early studies showed a huge impact of hypothalamic lesions on the regulation of energy homeostasis, a great deal of emphasis and study has focused on the hypothalamus as a primary site of such regulation. In areas such as the ARC and VMN, 10–30% of the neurons are glucose sensing. These include neurons expressing the anabolic ARC neuropeptide Y (NPY)/agouti-related peptide (AgRP) and LHA orexin and melanin-concentrating hormone (MCH), and the catabolic ARC proopiomelanocortin (POMC) neurons. NPY/AgRP and orexin neurons are GI in type, and POMC and MCH neurons are GE in type. The ARC NPY/AgRP and POMC neurons have overlapping projections primarily to the PVN, LHA, and sympathetic preganglionic spinal cord neurons, and their major function appears to be the regulation of energy homeostasis. On the other hand, small populations of MCH and orexin neurons localized in the LHA project widely throughout the neuraxis to areas involved in cognition, emotion, motivation, reward, motor activity, arousal, autonomic, and neuroendocrine activation (Figure 2). Thus, the ARC neurons are primarily dedicated to the control of energy homeostasis while MCH and orexin neurons are effectors through which the ARC neurons modulate the complex sets of behaviors and physiological functions, of which the regulation of energy homeostasis is only a small part. Many of the ARC and LHA neurons are both glucose sensors and metabolic sensors. They also receive direct or indirect afferents from the NTS and other brain stem nuclei which are also integrators of peripheral neural, metabolic, and hormonal signals (Figure 1). Thus, no one site or set of neurons functions alone. Instead the combined input of this distributed network of metabolic-sensing neurons acts collectively to regulate energy homeostasis.

Glucose-Sensing Neurons and Obesity

Since glucose-sensing neurons are so important in the control of energy homeostasis, it is not surprising that alterations in their function might affect the long-term regulation of body weight and adiposity. In fact, rats selectively bred to develop diet-induced obesity when the caloric density and fat content of their diet are increased manifest a number of abnormalities of both glucose sensing and hypothalamic neuronal responsiveness to leptin and insulin. These abnormalities include a reduced number of hypothalamic GE neurons, reduced sensitivity of the K_{ATP} channel to both ATP and sulfonylureas in GE neurons, reduced binding of sulfonylureas to the sulfonylurea binding site on the K_{ATP} channel, aberrant activation of the sympathetic nervous system to glucose infused either systemically or directly into the forebrain via the carotid artery, and increased expression of GK mRNA in the ventromedial hypothalamus associated with a blunted response to insulin-induced hypoglycemia. All of these abnormalities antedate the development of obesity, as does a reduced sensitivity to the anorectic effects of both leptin and centrally administered insulin. Such data suggest that there may be a genetic predisposition toward reduced sensitivity of glucose-sensing neurons to catabolic feedback signals which would ordinarily limit the amount of food ingested. Such a raised threshold would provide a competitive advantage where cycles of feast and famine were common, since it would allow such individuals to eat beyond their metabolic needs, to store as many calories as fat as possible when food was readily available. This would serve as a buffer during times of food shortage. However, in our modern society, where food is abundant and can be obtained with little expenditure of energy, such a trait would predispose such individuals to become obese. Thus, a reduced sensitivity of glucose-sensing neurons to metabolic and hormonal feedback cues may be one of the factors which underlie the current worldwide epidemic of obesity.

Summary and Conclusions

Glucose is a critical substrate for neuronal function. Mammals have evolved specialized peripheral glucose sensors and central neurons which utilize glucose as a signaling molecule, as a means of sensing and regulating glucose and carbohydrate metabolism. The peripheral sensors communicate with the central glucose-sensing neurons via autonomic visceral afferents, whereas glucose is monitored directly by brain glucose-sensing neurons, many of which use GK as a rate-limiting step in the signaling pathway. These neurons also respond to a variety of hormones such as leptin and insulin and metabolic substrates such as fatty acids, lactate, and ketone bodies, which provide direct or indirect indices of the status of energy storage depots and the metabolic state of the body. All of these signals are integrated in a time- and concentration-dependent manner at the level of the cell's membrane potential, to regulate neuronal activity,

metabolism, and transmitter release. These metabolic-sensing neurons are spread throughout the brain in a distributed network which controls the regulation of energy homeostasis. Some obesity-prone individuals inherit a raised threshold for sensing metabolic and hormonal signals from the periphery. While this raised threshold would allow them to survive periods of scarce food availability, it may predispose them to become obese when food is readily available at low energetic cost.

See also: Autonomic Nervous System: Central Thermoregulatory Control; Energy Homeostasis: Thermoregulation; Energy Homeostasis: Hypothalamic Development; Gastrointestinal Signals: Stimulation; Gastrointestinal Signals: Satiety; Nutrient Sensing: Cellular Metabolism; Nutrition; Thermoregulation: Autonomic, Age-Related Changes.

Further Reading

Ashford MLJ, Boden PR, and Treherne JM (1990) Glucose-induced excitation of hypothalamic neurones is mediated by ATP-sensitive K^+ channels. *Pflugers Archiv European Journal of Physiology* 415: 479–483.

Berthoud HR and Seeley RJ (1999) *Neural and Metabolic Control of Macronutrient Intake*. New York: CRC Press.

Bouzier-Sore AK, Merle M, Magistretti PJ, and Pellerin L (2002) Feeding active neurons: (Re)emergence of a nursing role for astrocytes. *Journal of Physiology – Paris* 96: 273–282.

Kang L, Dunn-Meynell AA, Routh VH, et al. (2006) Glucokinase is a critical regulator of ventromedial hypothalamic neuronal glucosensing. *Diabetes* 55: 412–420.

Levin BE, Dunn-Meynell AA, and Routh VH (1999) Brain glucose sensing and body energy homeostasis: Role in obesity and diabetes. *American Journal of Physiology* 276: R1223–R1231.

Levin BE, Dunn-Meynell AA, and Routh VH (2002) CNS glucosensing and regulation of peripheral glucose levels. In: Dwyer DS (ed.) *Glucose Metabolism in the Brain*, 51st edn., pp. 219–258. San Diego: Academic Press.

Levin BE, Routh VH, Kang L, et al. (2004) Neuronal glucosensing: What do we know after 50 years? *Diabetes* 53: 2521–2528.

Mayer J (1953) Glucostatic mechanism of regulation of food intake. *New England Journal of Medicine* 249: 13–16.

Mobbs CV, Kow LM, and Yang XJ (2001) Brain glucose-sensing mechanisms: Ubiquitous silencing by aglycemia vs. hypothalamic neuroendocrine responses. *American Journal of Physiology – Endocrinology and Metabolism* 281: E649–E654.

Oomura Y (1983) Glucose as a regulator of neuronal activity. In: Szabo AJ (ed.) *CNS Regulation of Carbohydrate Metabolism*, pp. 31–65. New York: Academic Press.

Silver IA and Erecinska M (1998) Glucose-induced intracellular ion changes in sugar-sensitive hypothalamic neurons. *Journal of Neurophysiology* 79: 1733–1745.

Swanson RA and Magistretti PJ (1997) Glia and brain energy metabolism. In: Ransom BR and Kettenmann H (eds.) *Glia*, pp. 154.

Nutrient Sensing: Cellular Metabolism

R J Seeley, University of Cincinnati, Cincinnati, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The regulation of energy balance is the result of an accurate match between energy intake and energy expenditure. This is critical to the survival of any organism, and consequently humans have evolved a number of redundant mechanisms to ensure that the energy balance equation is maintained. To accomplish this regulation of energy balance, the central nervous system (CNS) must be able to monitor two separate variables. The first of these is the amount of stored fuel in the body. Most of this stored fuel is in the form of adipose tissue deposited in depots throughout the body (Figure 1). A wide range of evidence indicates that adipose tissue is a major endocrine organ, secreting factors that control a number of key metabolic processes. At least some of these factors appear to provide afferent information back to the CNS about the status of these stored calories. In particular, a large body of evidence indicates that the hormone leptin circulates in proportion to the amount of calories in adipose tissue. Leptin appears to cross the blood–brain barrier, where it acts upon a number of key receptor populations. One key receptor population is within the arcuate nucleus of the hypothalamus.

The second key variable is the amount of energy currently available to power necessary biological functions. Monitoring this available fuel is a key function for metabolic circuits within the CNS, since a failure to keep adequate fuel available is rapidly catastrophic. The need to monitor cellular fuel availability is not unique to cells in the CNS. Rather, this is a necessary component of every cell to ensure that key cellular processes are scaled back when the required energy is not available. Individual cells accomplish this by the complex networks of biochemical reactions that ensure that the proper amount and type of fuel is available for cellular processes. To regulate this network of biochemical reactions, cells have developed signaling pathways that act as 'fuel sensors.' These fuel sensors serve to read out the availability of key aspects of fuel to regulate broad aspects of cellular activity.

Fuel Sensors

One key energy form that cells must carefully monitor is adenosine triphosphate (ATP), since it provides the basic source of energy for most cellular processes. Because of ATP's central importance, cells have enzymes with activities that are regulated by the ratio of ATP to adenosine monophosphate (AMP). Two examples of such fuel sensors are AMP kinase (AMPK) and the mammalian Target of Rapamycin (mTOR) protein. AMPK is a highly conserved serine/threonine kinase. AMPK is activated when cellular fuel availability is low, and its activation leads to a number of cellar changes that conserve energy. This regulation is achieved because the tumor suppressor serine/threonine kinase LKB1 (now named STK11), which is itself activated by AMP, and the Ca^{2+}/calmodulin-dependent protein kinase kinase-β (CaMKKβ) regulate phosphorylation and activation of AMPK. Moreover, AMP simultaneously inhibits AMPK dephosphorylation by inhibiting protein phosphatase-2C. These effects are antagonized in the presence of high concentrations of ATP, allowing AMPK activity to precisely monitor changes in the cellular ratio of ATP:AMP.

While AMPK is activated in low-fuel situations, mTOR is activated when cellular energy is in abundance. mTOR (also known as FRAP, RAFT, RAPT, or SEP) is a ~280 kDa protein member of the phosphatidylinositol kinase-related protein kinase family that regulates critical aspects of cell growth, including transcription, translation initiation, and elongation, and cell-cycle progression. mTOR exists in two distinct multiprotein complexes: one containing a protein called raptor (mTORC1) and another containing a protein called rictor (mTORC2). The mTORC1 is the best studied and understood of the two complexes, since it appears to be the major mediator of rapamycin's potent actions. During nutrient deprivation, mTORC1 is stabilized in a manner that inhibits mTOR kinase activity, whereas it changes conformation and serves as a positive regulator of its activity in the presence of nutrients. Raptor also serves as an adaptor protein that recruits mTOR substrates, such as the 40S ribosomal proteins S6 kinase-1 and S6 kinase-2 (S6K1 and S6K2), and is necessary for their phosphorylation by mTOR. Given that the mTOR pathway is crucial for regulating rates of new protein synthesis, mTOR responds not to just fuel availability but also to those hormones and amino acids which regulate growth.

CNS Fuel Sensing

While individual cells must monitor their fuel levels to regulate specific intracellular functions, the CNS has a different problem. Individual neurons may monitor fuel levels but they must also coordinate the fuel availability they detect locally across the different

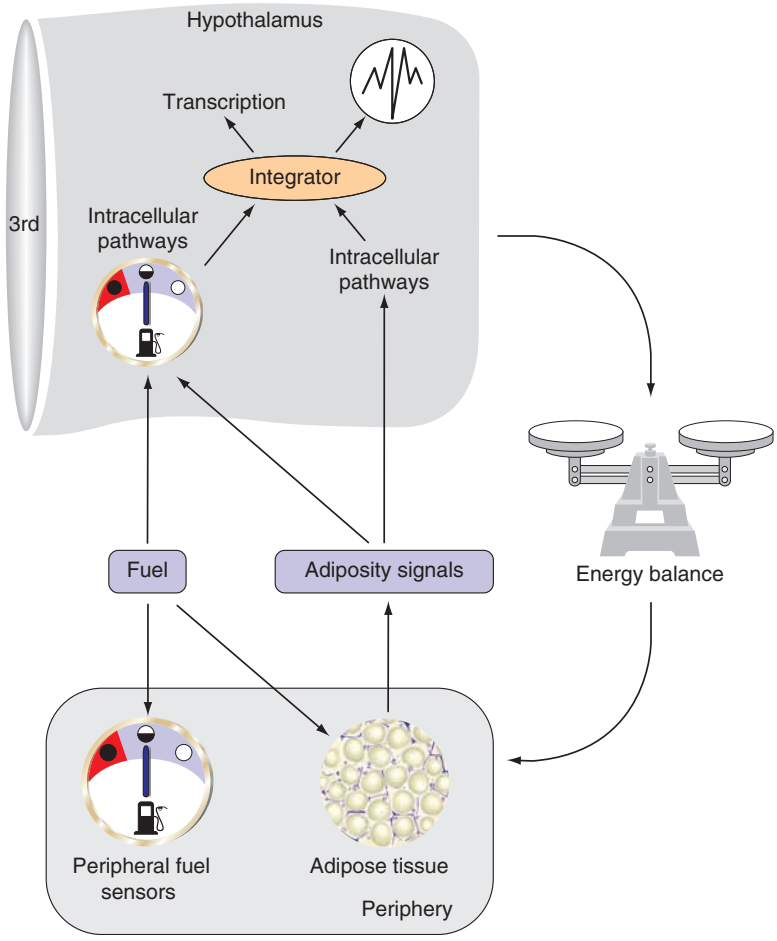

Figure 1 Regulation of energy balance.

tissues of the organism, and then direct behavior to act on critical needs. In many ways, the CNS is not an ideal organ for this function. Most neurons are heavily protected from the changing levels of fuels that peripheral cells must navigate. This protection is largely the result of glial cells, which are capable of providing needed fuel to neurons even when such fuel is in short supply. Second, neurons typically use glucose or related substitutes, such as lactate, as their primary source of fuel. While good evidence links glucose-sensing to the regulation of metabolism, many peripheral cell types, including the major energy-burning cells in muscle, can use either fatty acids or glucose to provide energy. Thus, for the CNS to regulate the organism's overall metabolic milieu it would seem necessary for neurons to be able to monitor lipid availability.

Despite some of these obvious drawbacks, growing evidence implicates specialized neuronal populations as being capable of sensing not just glucose availability but also energy more generally. The first piece of this new appreciation came from work using inhibitors of fatty acid synthase (FAS). It has long been established that FAS and fatty acid metabolism are regulated by the nutritional status of animals via lipogenic tissues, such as adipose and liver. Because the brain is not lipogenic, there was little reason to think that the *de novo* fatty acid synthesis pathway played an important role in the CNS. However, lipogenic enzymes are found in hypothalamic nuclei known to be critical for regulating both food intake and other metabolic parameters. Moreover, evidence from studies using FAS inhibitors, and using other pharmacological or genetic tools to modulate different enzymes of the FAS pathway in the CNS, suggests that this system senses and integrates signals from nutrients and nutrient-derived hormones. Administration of the natural FAS inhibitor cerulenin or a synthetic analog (C75) causes profound, dose-dependent anorexia and weight loss in rats and mice. These

effects are also observed with intracerebroventricular (icv) administration of much lower doses of C75, suggesting that the brain is the key site of action. In addition to the effects of C75, direct administration of oleic acid into the CNS results in reduced food intake and reduced hepatic glucose output. Collectively, these data point to an important ability of the CNS to monitor fuel in the form of fatty acids, rather than just the traditional fuel of neurons in the form of glucose.

CNS Fuel Sensors in the Regulation of Energy Balance

Given that at least some neuronal populations within the CNS appear to be able to respond to both lipids and glucose, one reasonable hypothesis is that these specialized populations of neurons have adapted to use the activity of cellular fuel sensors to assess the availability of fuel in the organism and to direct behavior and physiology. Solid evidence now links both AMPK and mTOR activity in the hypothalamus to the regulation of energy balance.

A number of manipulations of available energy appear to regulate AMPK activity within discrete nuclei of the hypothalamus. For instance, fasting, glucose deprivation, and administration of the melanocortin receptor antagonist/inverse agonist agouti-related protein (AgRP) increase AMPK activity in the same hypothalamic regions where refeeding, glucose, and a melanocortin receptor agonist reduce AMPK activity. Furthermore, the potent orexigen ghrelin increases hypothalamic AMPK activity, while anorexigenic treatments such as α-lipoic acid decrease hypothalamic AMPK levels. Because AMPK is activated in conditions of negative energy balance, it would be predicted to activate efferent pathways that promote the entry of fuels into the organism, in order to reestablish energy homeostasis. Increased activation of AMPK in the ventromedial hypothalamus has been found to stimulate feeding behavior and weight gain in rodents and to reduce the anorexigenic and weightreducing effects of α-lipoic acid. These data collectively support the hypothesis that AMPK activity detects changes in cellular energy to influence subsequent food ingestion.

Recent evidence also supports an important role for mTOR. By immunohistochemistry, mTOR activity is not widespread within the CNS but rather is localized to key circuits that regulate food intake and metabolism. mTOR activity is found in the arcuate nucleus of the hypothalamus and can be localized to both AgRP and proopiomelanocortin (POMC)-producing cells. The amount of mTOR activity in the arcuate also appears to be regulated by energy status, with fasting reducing mTOR activity. mTOR activity can be potently increased by specific branched-chain amino acids such as leucine. This is also true in the CNS, where leucine, but not valine, increases mTOR activity and can also reduce food intake. Collectively, these data point to an important role for mTOR as a cellular fuel sensor that is used by key populations of neurons to regulate food intake.

One important aspect of both AMPK and mTOR is that they are regulated by hormonal signals in addition to their fuel-sensing capacity. As a consequence, they are ideally suited to not only be important for regulating the CNS response to fluctuations in available fuel, but also could be a key part of the response to stored fuel. Leptin is one such key signal and it appears to regulate activity of both AMPK and mTOR in the hypothalamus. Leptin administration decreases AMPK activity in the hypothalamus, and increased constitutive AMPK activity blunts the anorexic effects of leptin. Similarly, mTOR is an important target of the phosphatidylinositol 3-kinase (PI3K) and Akt signaling cascade, and leptin has been shown to activate PI3K in the hypothalamus. CNS administration of leptin results in increased mTOR activity in the hypothalamus, and when mTOR is inhibited by CNS administration of rapamycin, it reduces the ability of leptin to reduce food intake. Thus, another role for these fuel sensors in the hypothalamus is to integrate signals concerning available metabolic fuels and signals of stored fuel, to guide ingestion and metabolism.

Conclusions

Our understanding of how CNS circuits regulate food intake and body weight has expanded rapidly in the past 15 years. An important part of that new understanding is the result of applying advances in our understanding of cellular metabolic pathways back to the CNS. Evolving multicellular organisms developed specialized metabolic organs for storing and exchanging different types of metabolic fuels. While such organs give an organism an enormous advantage, the need to coordinate across these organs is critical for survival. The CNS clearly plays a large role in this process and has developed a number of ways both to sense metabolic perturbations and to respond to them. Given this central role, however, it also means that the CNS may be an important part of the etiology of diseases of metabolic disruption, such as obesity and type 2 diabetes. A wide range of evidence points to the same circuits that regulate food intake as being crucial to the regulation of blood

glucose. While the prevailing view of how obesity leads to greatly increased incidence of type 2 diabetes is that increased fat stores lead to insulin resistance in muscle and greater strain on the β-cell, the new data suggest a more radical hypothesis. Given that such insulin resistance and β-cell failure is linked to alterations in key fuel-sensing metabolic pathways, it may be that the primary lesion is alterations in the fuelsensing capabilities in the CNS. Such a disruption would simultaneously lead to the inability to accurately match caloric intake and caloric expenditure, while limiting the ability to coordinate metabolic activity across different organs. Thus, it may be that the CNS will provide a single underlying etiology that links obesity to type 2 diabetes. Future work will need to focus on understanding how environmental and genetic factors can influence the function of these fuel-sensing pathways to increase or decrease the incidence of metabolic disease.

See also: Energy Homeostasis: Adiposity Signals; Food and Water Intake: Regulation; Gastrointestinal Signals: Stimulation; Gastrointestinal Signals: Satiety; Hormonal Signaling to the Brain for the Control of Feeding/Energy Balance; Nutrient Sensing: Carbohydrates; Nutrition.

Further Reading

Cota D, Proulx K, Blake Smith KA, et al. (2006) Hypothalamic mTOR signaling regulates food intake. *Science* 312(5775): 927–930.

Dennis PB, Jaeschke A, Saitoh M, et al. (2001) Mammalian TOR: A homeostatic ATP sensor. *Science* 294(5544): 1102–1105.

Kahn BB, Alquier T, Carling D, et al. (2005) AMP-activated protein kinase: Ancient energy gauge provides clues to modern understanding of metabolism. *Cell Metabolism* 1(1): 15–25.

Landree LE, Hanlon AL, Strong DW, et al. (2004) C75, a fatty acid synthase inhibitor, modulates AMP-activated protein kinase to alter neuronal energy metabolism. *Journal of Biological Chemistry* 279(5): 3817–3827.

Loftus TM, Jaworsky DE, Gojeb L, et al. (2000) Reduced food intake and body weight in mice treated with fatty acid synthase inhibitors. *Science* 288: 2299–2300.

Minokoshi Y, Alquier T, Furukawa N, et al. (2004) AMP-kinase regulates food intake by responding to hormonal and nutrient signals in the hypothalamus. *Nature* 428(6982): 569–574.

Obici S and Rossetti L (2003) Minireview: Nutrient sensing and the regulation of insulin action and energy balance. *Endocrinology* 144(12): 5172–5178.

Ruderman NB, Saha AK, Vawas D, et al. (1999) Malonyl-CoA, fuel sensing, and insulin resistance. *American Journal of Physiology* 276 (1, pt. 1): E1–E18.

Schwartz MW and Porte D Jr. (2005) Diabetes, obesity, and the brain. *Science* 307(5708): 375–379.

Nutrition

M P Mattson and B Martin, National Institute on Aging Intramural Research Program, Baltimore, MD, USA

Published by Elsevier Ltd.

Synaptic Plasticity and Neurogenesis

Energy restriction has been shown to increase the average and maximum life span of rats and mice by up to 50% and retard age-related deficits in motor function and learning and memory (Table 1). For example, lifelong caloric restriction (CR; 30–40% reduction below *ad libitum* levels) and intermittent fasting (IF; every-other-day fasting), starting after puberty, improved the performance of mice on tests of motor coordination and complex learning. Conversely, long-term overfeeding, particularly with diets high in fats and carbohydrates, can impair learning and memory and motor function in rodents. Electrophysiological recordings of synaptic transmission in hippocampi from aged rats that had been maintained on control and CR diets showed that long-term potentiation (LTP), believed to be a cellular correlate of learning and memory, is enhanced by low-energy diets (Figure 1). In addition, data from studies of obese and untreated diabetic humans suggest a detrimental effect of high-energy diets on cognitive function which may result from chronic elevations of circulating glucose and lipids.

The adult brain contains populations of neural stem cells that divide and are capable of differentiating into new neurons; these cells are located primarily in the dentate gyrus of the hippocampus and the subventricular zone of the cerebral cortex. However, many newly generated neurons undergo a form of programmed cell death called apoptosis and therefore never become functional. Hippocampal neurogenesis is increased in rats and mice maintained on an IF energy restriction regimen, apparently as the result of increased survival of newly generated neurons (Figure 1). On the other hand, controlled CR results in increased survival of newly generated glial cells without affecting neurogenesis. Interestingly, increased energy expenditure (regular exercise) and ischemia also enhance hippocampal neurogenesis, demonstrating a high level of sensitivity of neural stem cells to changes in energy metabolism. Excessive energy intake and diabetes suppress neurogenesis in rodent models. Because recent findings suggest that neurogenesis plays an important role in the adaptive responses of the brain to changing environments, energy intake is likely to influence those responses.

Neuronal Vulnerability to Injury and Disease

Although it is well established that moderate energy restriction can reduce the incidence of type 2 diabetes, cardiovascular disease, and many types of cancer, only recently has the influence of energy intake on neurological disorders been elucidated. Epidemiological studies in human populations have revealed associations between overeating, type 2 diabetes, and increased risk of Alzheimer's disease (AD), Parkinson's disease (PD), and stroke. Individuals with low caloric intakes are at reduced risk of these disorders. Experiments in both rodent and monkey models have directly demonstrated that CR and IF diets suppress or delay brain disease processes and are neuroprotective. IF protects hippocampal neurons of rats and mice against seizure-induced damage and also reduces brain damage and improves functional outcome in a rat model of focal ischemic stroke. An IF regime also has been shown to protect dopaminergic neurons in a mouse model of PD and protect striatal neurons in a mouse model of Huntington's disease (HD). Additionally, CR has been shown to suppress the amyloid pathology and improve cognitive function in mouse models of AD, and monkeys maintained on a CR diet exhibited improved motor function and higher striatal dopamine levels compared to monkeys on a control diet in a PD model. CR can also protect the retina against ischemic injury and auditory neural cells against age-related degeneration.

While the amount of calories consumed can affect the vulnerability of neurons to injury and disease, the specific source of the energy supply may also be important. Dietary fats that promote cardiovascular disease, including saturated fats and trans fatty acids, may also render neurons vulnerable to neurodegenerative disease. On the other hand, omega-3 fatty acids such as docosahexanoic acid may protect neural cells against disease. The ways in which the body and its cells metabolize energy may be altered in neurodegenerative disorders. For example, both HD patients and some strains of huntingtin mutant mice develop a hypermetabolic state and lose body weight progressively beginning early in the disease process. In addition, deficits in glucose utilization may occur prior to neuronal degeneration in brain regions most severely affected in AD and PD. Moreover, mitochondrial toxins can cause PD-like (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridire (MPTP) and rotenone) or HD-like (3-nitropropionic acid) pathologies and symptoms in rodents, monkeys, and humans. The ability of

Table 1 Summary of the effects of energy intake (low-energy: CR and IF; high-energy: high-fat/high-glucose diet) on brain plasticity and neuronal vulnerability

Dietary manipulation	*Plasticity*	*Neuronal vulnerability*
Caloric restriction (CR)	Enhance	Decrease
Intermittent fasting (IF)	Enhance	Decrease
High-fat/high-glucose	Decrease	Enhance

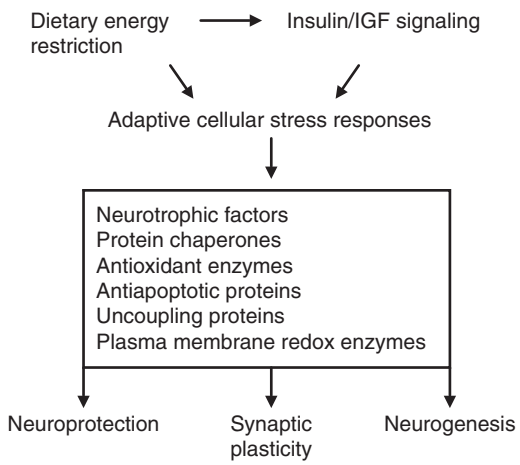

Figure 1 Hormesis model of the mechanism by which dietary energy restriction enhances neural plasticity and disease resistance. Energy restriction imposes a mild stress on neural cells, resulting in the activation of signaling pathways that upregulate the expression of proteins that promote neuronal survival, synaptic plasticity, and neurogenesis. These proteins include neurotrophic factors such as BDNF, protein chaperones, antioxidants, mitochondrial uncoupling proteins, and plasma membrane redox enzymes. Insulin and IGF-1 signaling pathways are responsive to energy restriction and may contribute to the adaptive stress responses of neural cells.

CR and IF diets to protect neurons against the latter toxins suggests an important role for energy intake in modifying key mechanisms involved in the pathogenesis of neurodegenerative disorders. However, reduced energy diets are not uniformly beneficial. Epidemiological findings and recent data from studies of mouse models of amyotrophic lateral sclerosis (ALS) suggest that motor neurons may not respond adaptively to energy restriction. Instead, high-fat diets may benefit motor neurons and protect them against ALS.

Neuroendocrine Responses to Variable Energy Availability

Highly complex integrated neuroendocrine systems control the balance between energy intake and expenditure in the context of food availability. The hypothalamus is a critical regulator of energy balance; it responds to hormones involved in energy intake (leptin and ghrelin, for example) and to higher cognitive processes in brain regions such as the hippocampus. Importantly, the hypothalamus also controls behavioral and metabolic responses to stress. Accordingly, the hypothalamic–pituitary–adrenal (HPA) axis is activated in response to energy restriction, resulting in increases of circulating glucocorticoid levels (cortisol in humans and other primates, and corticosterone in rodents). Accompanying the latter changes are increased arousal and physical activity. The state of mild stress imposed by CR and IF is part of an adaptive response that motivates the animal or human to actively seek food. Conversely, constant access to food results in satiety and reduced cognitive and physical activity. Numerous changes in neuroendocrine signaling systems occur in response to moderate energy restriction, including decreased levels of insulin and leptin (associated with increased sensitivity of insulin and leptin target cells), growth hormone, and prolactin. Such neuroendocrine changes are adaptive, orchestrating the reallocation of energy reserves among organ systems (i.e., reducing body fat and shifting reserves from reproductive processes to those necessary for survival). Severe energy restriction results in drastic changes in neuroendocrine systems that can be considered emergency measures. In females these include complete shutdown of the reproductive system and an incessant increase in physical and cognitive activities.

Fundamental aspects of cardiovascular physiology are sensitive to energy intake, including blood pressure, heart rate, and heart rate variability. Obesity and a sedentary lifestyle are associated with increased blood pressure and decreased heart rate variability in humans. When rats are maintained on CR or IF diets, their blood pressure (both systolic and diastolic) and resting heart rate decrease. In addition, cardiovascular stress adaptation is enhanced as indicated by a more rapid recovery of blood pressure and heart rate following exposure to physiological stressors. CR and IF diets result in an increase in heart rate variability, suggesting a relative increase in parasympathetic (acetylcholine-mediated) regulation of the heart. The latter changes, as well as increased insulin sensitivity, occur despite hyperactivation of the HPA axis. The hyperactivation of the HPA axis in response to energy restriction is of considerable interest because other types of stress that activate the HPA, including chronic uncontrollable psychosocial stress, are detrimental. Apparently, there are pivotal differences in how cells and organ systems respond to different types of stress; they respond adaptively to energy restriction and exercise, but maladaptively to excessive energy intake and unrelieved stress.

Cellular and Molecular Mechanisms: Hormesis and Insulin Signaling

The available data support two general mechanisms by which energy restriction extends life span in many organisms and increases disease resistance. One potential mechanism involves a decrease in oxidative stress as the result of lower amounts of glucose availability to cells and, accordingly, decreased production of mitochondrial reactive oxygen species (ROS). Accordingly, overeating enhances cellular ROS production, which can damage proteins and lipids in neurons, resulting in impaired synaptic plasticity and even cell death. ROS that are increased in response to energy intake include superoxide anion radical, hydrogen peroxide, hydroxyl radical, and nitric oxide. Such ROS damage cell membranes, DNA, and proteins, which can lead to cell dysfunction and death. There is considerable evidence that ROS contribute to the demise of neurons in essentially all neurodegenerative conditions including stroke, AD, PD, HD, and ALS. Prolonged energy restriction that occurs during IF may also protect neurons by shifting energy substrates from glucose to ketone bodies such as β-hydroxybutyrate, resulting in less ROS production.

A second mechanism by which energy restriction could be neuroprotective is by hormesis, a process in which cells subjected to a mild stress become resistant to more severe stress (Figure 1). Studies of rats and mice maintained on energy restriction diets have documented increases in levels of several stress resistance proteins, including heat-shock proteins and other protein chaperones, mitochondrial uncoupling proteins, plasma membrane redox enzymes, and sirtuins (histone deacetylases believed to play important roles in energy metabolism and aging). Interestingly, manipulations of dietary energy intake and exercise can also affect the production of several neurotrophic factors in certain brain regions such as brain-derived neurotrophic factor (BDNF) and glial cell line-derived neurotrophic factor (GDNF), which may contribute to the effects of changes in energy intake and expenditure on neuronal plasticity and vulnerability to disease. Although moderate energy restriction may protect neurons against injury and disease through decreased ROS and activation of hormesis mechanisms, the effects of severe energy restriction on these pathways are currently poorly understood.

Some of the effects of dietary energy restriction and overeating on peripheral organ systems may be mediated by neurotransmitter and neurotrophic factor signaling in the brain. Energy restriction has been reported to affect the levels of several neurotransmitters in the brains of rodents, including dopamine, norepinephrine, serotonin, and glutamate. Type 2 diabetes is associated with reductions in levels of serotonin and norepinephrine in different brain regions, and patients with clinical depression are prone to diabetes, further supporting relationships between neurotransmitter signaling in the brain and peripheral energy metabolism. Other studies have documented effects of pharmacological manipulations of such neurotransmitter signaling pathways on peripheral glucose metabolism. Energy restriction enhances insulin sensitivity peripherally, and findings from studies of BDNF-deficient mice and mice in which BDNF is infused into the brain suggest a role for BDNF signaling in the antidiabetic effects of energy restriction. Although the specific neural circuits and signaling pathways by which the nervous system mediates effects of energy intake on other organ systems remain to be established, both dietary energy restriction and excess affect blood pressure and heart rate variability by mechanisms which involve changes in the autonomic nervous system.

Finally, in light of the important roles of insulin and insulin-like growth factor (IGF) signaling pathways in the regulation of aging, the involvement of these factors in the effects of energy intake on neural plasticity is likely. Studies of the nematode *Caenorhabditis elegans* and the fly *Drosophila melanogaster* have clearly established a major role for insulin signaling in aging and life span determination. In *C. elegans,* activation of the insulin receptor homolog Daf-2 stimulates a PI_3 kinase (Age-1) which phosphorylates and activates Akt kinase; Akt phosphorylates and inhibits the forkhead transcription factor Daf16. Mutations in the insulin signaling pathway extend life span in worms, in part by derepressing daf-16, resulting in the expression of 'longevity genes' including those involved in cellular stress resistance. In mammals, insulin and IGF-1 activate receptors coupled to PI_3 kinase and Akt (Figure 2). Mammalian homologs of Daf-16 that are targets of Akt comprise the FOXO transcription factor family. FOXOs are sensitive to metabolic and oxidative stress and control the expression of genes involved in glucose metabolism, antioxidant defenses, and DNA repair. Insulin and IGF-1 have been shown to protect neurons against oxidative and metabolic insults and may also play important roles in neurogenesis and synaptic plasticity.

Dietary energy restriction increases the sensitivity of muscle and liver cells to insulin, resulting in improved glucose tolerance and resistance to diabetes and cardiovascular disease, two major causes of death in humans. Circulating levels of both insulin and IGF-1 decrease in response to energy restriction; however, levels of IGF-1 in the brain may increase. IGF-1 has been shown to act directly on neurons

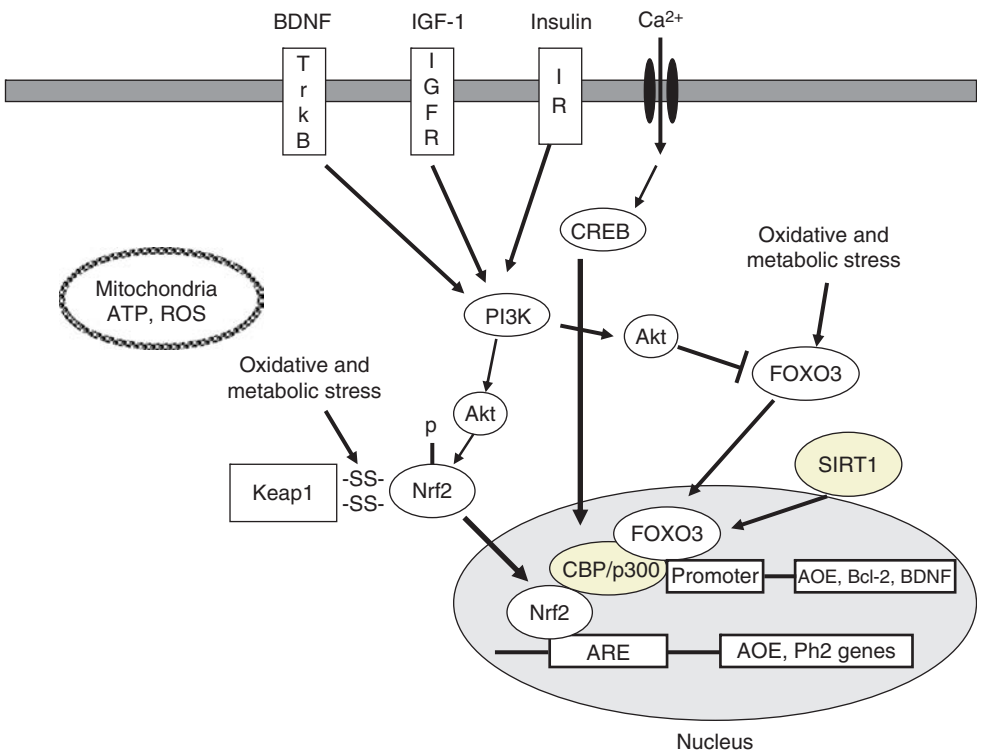

Figure 2 Insulin, IGF-1, and BDNF each activate receptors (IR, IGFR, and TrkB, respectively) coupled to the PI3 kinase–Akt pathway. Akt phosphorylates and inhibits FOXO transcription factors such as FOXO3. FOXO3 activity may be modified by oxidative and metabolic stress and by the histone deacetylase SIRT1. Akt and cellular stress also activate the transcription factor Nrf2 by causing the dissociation of Keap from Nrf2. Nrf2 translocates to the nucleus and binds to the antioxidant response element (ARE) in genes that encode cytoprotective proteins such as antioxidant enzymes (AOE) and phase 2 proteins. Calcium influx through ionotropic glutamate receptors or voltage-dependent Ca^{2+} channels activates the transcription factor CREB (cyclic AMP response element binding protein) which binds the CBP (CREB binding protein) to regulate transcription of genes encoding proteins involved in neuronal plasticity and survival, including brain-derived neurotrophic factor (BDNF) and Bcl-2.

and neural stem cells to promote plasticity and resistance to degeneration, emphasizing the importance of insulin-like signaling in brain plasticity and disease susceptibility.

Although not fully understood, there is evidence that the activities of FOXOs and sirtuins are modified by insulin/IGF-1 and in response to energy restriction, and these changes may contribute to the life span-extending and neuroprotective effects of energy restriction.

Implications for Improving Brain Health Span

As with other major organ systems, the brain is adversely affected by excessive dietary energy intake and so benefits from spartan diets. Studies of the effects of CR and regular exercise on life span in rodents suggest that energy restriction has a major impact on longevity (both the mean and maximum life span can be increased 30–40%), whereas lifelong exercise regimens are less effective (5–10% increase in mean life span). Cutting back on energy intake is therefore likely to extend the disease-free period of life for most individuals. For the many who are unable to curtail their food intake, other approaches are being developed to either promote a reduction in food intake or modulate the cellular and molecular pathways that mediate beneficial effects of energy restriction on the nervous system. To treat obesity, agents that suppress appetite are being developed, including those acting on melanocortin and cannabinoid signaling pathways; such agents would be expected to promote healthy brain aging. Attempts to develop dietary supplements that mimic the effects of CR are in progress, but their effectiveness to date has been limited by unanticipated adverse effects on some organ systems. Finally, it may be possible to activate neuronal stress resistance pathways; examples would include agents that stimulate neurotrophic factor signaling, induce protein chaperone expression, and/or enhance antioxidant systems.

See also: Alzheimer's Disease: Molecular Genetics; Amyotrophic Lateral Sclerosis (ALS); Animal Models of Parkinson's Disease; Autonomic Nervous System;

Cortical Processing of the Reward Value of Food; Food and Water Intake: Regulation; Hippocampus; Neurogenesis in the Intact Adult Brain; Nutrient Sensing: Carbohydrates; Nutrient Sensing: Cellular Metabolism.

Further Reading

Davidson TL, Kanoski SE, Walls EK, and Jarrard LE (2005) Memory inhibition and energy regulation. *Physiology Behavior* 86: 731–746.

Duan W, Guo Z, Jiang H, Ware M, Li XJ, and Mattson MP (2003) Dietary restriction normalizes glucose metabolism and BDNF levels, slows disease progression, and increases survival in huntingtin mutant mice. *Proceedings of the National Academy of Sciences of the United States of America* 100: 2911–2916.

Dupuis L, Oudart H, Rene F, Gonzalez de Aguilar JL, and Loeffler JP (2004) Evidence for defective energy homeostasis in amyotrophic lateral sclerosis: Benefit of a high-energy diet in a transgenic mouse model. *Proceedings of the National Academy of Sciences of the United States of America* 101: 11159–11164.

Greenwood CE and Winocur G (2005) High-fat diets, insulin resistance and declining cognitive function. *Neurobiology of Aging* 26: S42–S45.

Hyun DH, Emerson SS, Jo DG, Mattson MP, and de Cabo R (2006) Calorie restriction up-regulates the plasma membrane redox system in brain cells and suppresses oxidative stress during aging. *Proceedings of the National Academy of Sciences of the United States of America* 103: 19908–19912.

Ingram DK, Weindruch R, Spangler EL, Freeman JR, and Walford RL (1987) Dietary restriction benefits learning and motor performance of aged mice. *Journal of Gerontology* 42: 78–81.

Kim KY, Ju WK, and Neufeld AH (2004) Neuronal susceptibility to damage: Comparison of the retinas of young, old and old/caloric restricted rats before and after transient ischemia. *Neurobiology of Aging* 25: 491–500.

Kolta MG, Holson R, Duffy P, and Hart RW (1989) Effect of long-term caloric restriction on brain monoamines in aging male and female Fischer 344 rats. *Mechanics of Ageing and Development* 48: 191–198.

Lee J, Duan W, and Mattson MP (2002) Evidence that brain-derived neurotrophic factor is required for basal neurogenesis and mediates, in part, the enhancement of neurogenesis by dietary restriction in the hippocampus of adult mice. *Journal of Neurochemistry* 82: 1367–1375.

Mager DE, Wan R, Brown M, Cheng A, Wareski P, Abernethy DR, and Mattson MP (2006) Caloric restriction and intermittent fasting alter spectral measures of heart rate and blood pressure variability in rats. *FASEB Journal* 20: 631–637.

Maswood N, Young J, Tilmont E, et al. (2004) Caloric restriction increases neurotrophic factor levels and attenuates neurochemical and behavioral deficits in a primate model of Parkinson's disease. *Proceedings of the National Academy of Sciences of the United States of America* 101: 18171–18176.

Mattson MP, Cutler RG, and Camandola S (2007) Energy intake and amyotrophic lateral sclerosis. *Neuromolecular Medicine* 9: 17–20.

Mattson MP, Maudsley S, and Martin B (2004) BDNF and 5-HT: A dynamic duo in age-related neuronal plasticity and neurodegenerative disorders. *Trends in Neuroscience* 27: 589–594.

Nakagawa T, Ono-Kishino M, Sugara E, Yamanaka M, Taiji M, and Noguchi H (2002) Brain-derived neurotrophic factor (BDNF) regulates glucose and energy metabolism in diabetic mice. *Diabetes Metabolism Research and Reviews* 18: 185–191.

Rowland NE and Bellush LL (1989) Diabetes mellitus: Stress, neurochemistry and behavior. *Neuroscience and Biobehavioral Reviews* 13: 199–206.

Tatar M, Bartke A, and Antebi A (2003) The endocrine regulation of aging by insulin-like signals. *Science* 299: 1346–1351.

Wan R, Camandola S, and Mattson MP (2003) Intermittent food deprivation improves cardiovascular and neuroendocrine responses to stress in rats. *Journal of Nutrition* 133: 1921–1929.

Wolkow CA (2002) Life span: Getting the signal from the nervous system. *Trends in Neuroscience* 25: 212–216.

Yaffe K, Blackwell T, Kanaya AM, Davidowitz N, Barrett-Connor E, and Krueger K (2004) Diabetes, impaired fasting glucose, and development of cognitive impairment in older women. *Neurology* 63: 658–663.

Nystagmus

M J Shelhamer, The Johns Hopkins University School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Vestibular Nystagmus

An oscillatory pattern of eye movements characterized by alternating fast and slow phases. Nystagmus can be either the normal compensatory response to head/world motion, or a pathologic response to peripheral vestibular or central neural injury.

Example

When an animal rotates about a vertical axis, an eye movement is generated that attempts to keep the eyes stable in space. If the motion is prolonged for more than a few seconds and then abruptly halted, the same eye movement is generated in the opposite direction in an attempt to compensate for illusory motion in the opposite direction. This reflexive compensatory eye movement is known as nystagmus, vestibular nystagmus in this case since it is generated by the vestibular system and is manifest as the vestibulo-ocular reflex (VOR). The associated perception of spinning is known as vertigo.

Definition

The word nystagmus is derived from 'nustagmos,' Greek for 'drowsiness,' which refers to the nodding motion made by the head of a sleepy individual: slow leaning forward punctuated by rapid adjustments to the upright. Nystagmus was recognized hundreds of years ago, while systematic studies began in the late 1790s and early 1800s by Wells, Mach, Purkinje, and Crum Brown. Early controversy as to whether or not the eyes moved during the vertigo associated with body rotation was resolved in part by observing a retinal after-image during the nystagmus generated after self-rotation; its apparent motion against a stationary background demonstrated that the eyes indeed move and in this case generate postrotatory vestibular nystagmus.

Description

The most prominent feature of nystagmus is the characteristic alternating fast and slow phases (Figure 1). (An exception is pendular nystagmus, where the waveform is more sinusoidal.) The slow phases are intended to stabilize vision by matching the velocity of the visual field relative to the head, thereby keeping retinal image motion (retinal slip) at a low level in order to enhance visual acuity. The fast phases reset the eyes to select and begin tracking a new point. Nystagmus is most often labeled in terms of the direction of the fast phases, since they are so easily seen upon visual examination of a person's eyes; hence 'right-beating nystagmus' has slow phases to the left, and would be compensatory for rightward head movements. During sustained nystagmus, the eyes are on average deviated in the fast-phase direction; this is the so-called beating field.

This description alludes to the function of nystagmus as helping to maintain visual acuity during head motion. A separate but closely related eye movement pattern is optokinetic nystagmus (OKN), which is evoked by the homogeneous motion of a wide-field visual stimulus about a stationary subject.

Although nystagmus is almost always involuntary (reflexive), some normal subjects can produce a form of voluntary nystagmus (saccadic oscillations) at will.

Main Characteristics

Nystagmus can usually be suppressed by fixating on a visual target. High-performance aircraft pilots must suppress the nystagmus that would result from rapid maneuvers in order to maintain gaze on cockpit instruments. Repeated exposures to head/body rotation will habituate the response, so that it decreases in intensity. This is one way in which ice skaters can perform fast multiple spins without adverse effects. During slower turns by dancers, for example, spotting is a helpful strategy to avoid nystagmus and vertigo: the head and eyes maintain gaze on a fixed position while the body turns, then the head and eyes make a rapid movement in the direction of body rotation in order to fixate a new position. This eliminates the prolonged vestibular stimulation (head rotation) which would lead to perrotatory nystagmus.

The apparent motion of the visual world during nystagmus slow phases (the perception of the world spinning) indicates that there is no efference copy (internal monitoring) of the slow phases. It is such an efference copy that allows, for example, smooth pursuit eye movements to be made without the perception that the world is moving in the opposite direction.

Nystagmus fast phases have the same kinematic properties as saccades, and are generated by the same neural structures. The slow phases of vestibular nystagmus are a manifestation of the VOR, in which head rotation transduced by the semicircular canals of the vestibular system drives the eyes in the opposite direction so as to stabilize gaze on a fixed point in space.

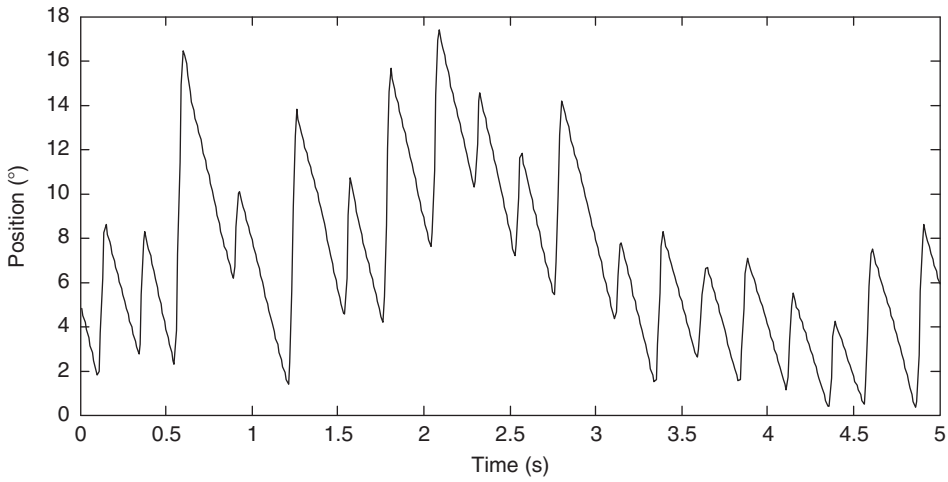

Figure 1 Example of optokinetic nystagmus (OKN), stimulated with a wide-field visual stimulus moving at 60° s^{-1}.

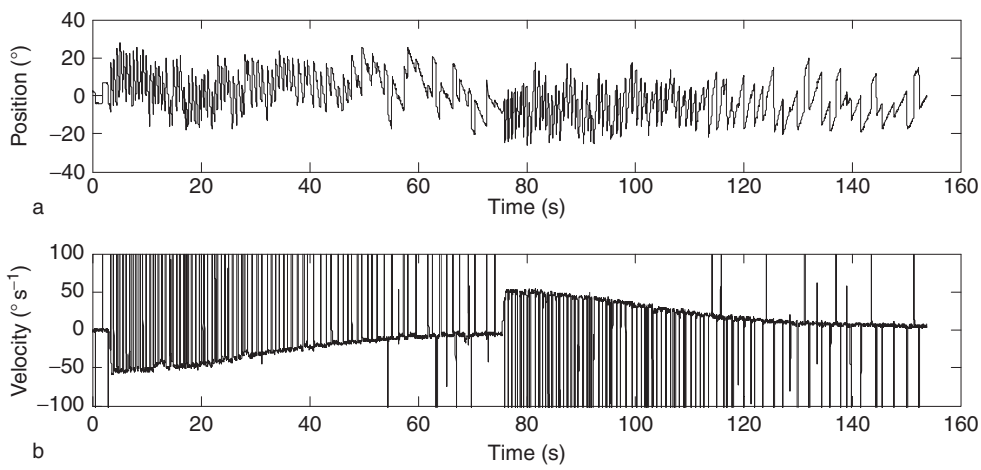

Figure 2 Examples of vestibular nystagmus. Eye position (top) and velocity (bottom) waveforms are shown. Fast phases appear as large-amplitude spikes in the velocity signal; plot scale has been set to show characteristic exponential decay of the slow-phase velocity. Perrotatory nystagmus begins at start of head rotation, near beginning of each graph. Postrotatory nystagmus begins when rotation stops, near center of each graph. Note initial rise in slow-phase velocity, followed by exponential decay, in each case. Courtesy of Mark Walker.

Vestibular nystagmus is often induced in the lab and clinic by prolonged rotation at a constant velocity, followed by a rapid halt; this results in perrotatory nystagmus and postrotatory nystagmus. In each case, the nystagmus slow-phase velocity starts at a value approximately equal to that of the rotational stimulus (a gain near 1.0). The velocity of successive slow phases decreases approximately as a decaying exponential, characterized by a dominant time constant. The time constant (highly variable but typically about 10–15 s) is much longer than the mechanical time constant of the semicircular canals; the central perseveration of activity is known as velocity storage (see Figure 2).

OKN (Figure 1) has an almost immediate rise to a near-constant velocity slow phase. It can be generated by moving a large OKN drum with a patterned visual field around the stationary subject. OKN is also generated more naturally during prolonged head turns (more than a second or so): when the VOR slow phases begin to decay in velocity they are augmented by an optokinetic response. Thus, there is a strong synergy between vestibular and optokinetic responses and the associated forms of nystagmus.

Although horizontal nystagmus is most often studied, at least in the laboratory, the appropriate natural stimuli, and some pathological conditions, can produce nystagmus with vertical or torsional

components. Rotation of the head toward one shoulder, for example, will produce torsional nystagmus.

Neural Basis

As is true for a great many types of eye movements, the neural pathways for nystagmus include two neural pools that implement the mathematical operation of integration with respect to time. The first of these is the velocity-storage integrator, which as noted above acts to perseverate vestibular activity from the semicircular canals. The other integrator converts neural eye-velocity commands to eye-position commands, both of which are needed to drive the extraocular muscles correctly. In particular, this second integrator confers the ability to maintain eccentric gaze. Deficits in the velocity-storage integrator affect the long time constant of the VOR. Deficits in the velocity-position integrator can lead to changes in the shape of the nystagmus slow phase. Normally the slow phase is straight (constant velocity), in accord with its role in stabilization of the line of sight during a prolonged head movement. If the integrator is abnormally leaky, it cannot maintain its output level, and each slow phase will decrease in velocity, leading to a scalloped shape. If the integrator is unstable, its output increases out of proportion to the input, and each slow phase will exhibit an exponential increase in velocity.

Vestibular Pathology

Nystagmus can be a reliable indicator of vestibular pathology. This discussion will be presented in terms of horizontal nystagmus, although similar reasoning applies to vertical and torsional movements, with some modifications.

If an acute peripheral lesion upsets the normal balance of tonic activity from the two sets of vestibular apparatus on either side of midline, the resulting imbalance is interpreted centrally as a prolonged rotation away from the lesioned side. This results in a strong and persistent spontaneous nystagmus, with fast phases directed away from the affected canal. Subsequent reduction in spontaneous nystagmus is one indicator of central compensation for this pathology.

The three semicircular canals are approximately aligned with the movement directions of the extraocular muscles. Thus, nystagmus that follows a specific canal plane often results from a peripheral lesion of that canal, while nystagmus in other directions is of central origin. Pure vertical or torsional nystagmus, for example, will not be of peripheral vestibular origin, but rather has a central cause. A peripheral cause also leaves fixation suppression largely intact, since central fixation mechanisms are spared. Cross-coupling (e.g., vertical nystagmus after horizontal head motions) also indicates a central cause.

Types of Nystagmus

Many specific types of nystagmus have been identified. Some are named for the particular eye motion that is involved, while others draw on the mode of stimulation for their naming.

Various types of physiologic nystagmus can be considered as the normal response to actual or perceived motion. These include perrotatory and postrotatory vestibular nystagmus, which manifest the VOR. Caloric nystagmus is generated by warm or cold stimulation of the horizontal semicircular canal, through water or air introduced into the external auditory meatus. Ideally the head is tilted back to place the horizontal canals in a vertical orientation. The resulting temperature gradient induces convection currents in the canal endolymph, mimicking a head turn. (There is also a small direct thermal effect on the nerve.)

Since nystagmus very often expresses an intent to compensate for real or illusory head motion, stimuli that mimic the sensory or motor circumstances that occur during actual head or body motion can also produce a ‘compensatory’ nystagmus. Such stimuli include movement of the outstretched arm with the hand following a surrounding drum that rotates in darkness (arthrokinetic nystagmus), and the proprioceptive input that results from apparent walking in a tight circle while actually stationary with the floor moving. Rotating sound fields can also induce nystagmus in a similar fashion.

Many specific types of pathologic nystagmus have been cataloged; in some cases the underlying neurophysiology is known and so these eye movements can serve as useful diagnostic tools. Many types of pathologic nystagmus are acquired through trauma or disease, but there also exists congenital nystagmus, which is often associated with visual deficits during development. Congenital nystagmus is most often characterized by a foveation period in each cycle, when the eyes are still or nearly so, allowing for, in some cases, normal visual acuity.

Periodic alternating nystagmus (PAN) is a spontaneous horizontal nystagmus that reverses direction approximately every two minutes; it is due to cerebellar deficit that renders the time constant of velocity storage unusually long and creates interactions with mechanisms that attempt to readjust the time constant.

Acute alcohol intoxication can change the density of the vestibular endolymph relative to that of the

cupula, and so render the semicircular canals sensitive to gravity, resulting in positional alcoholic nystagmus. Heavy water has the same effect in the opposite direction. Benign paroxysmal positional vertigo (BPPV), where debris in the canal renders it sensitive to gravity, is another cause of positional nystagmus.

Loud sounds can induce nystagmus through pathologic stimulation of the peripheral vestibular system (Tullio phenomenon, semicircular canal dehiscence syndrome).

Otolithic stimulation through linear acceleration (including changing orientation with respect to gravity) can also produce a form of vestibular nystagmus, a manifestation of the translational VOR (TVOR). Since TVOR slow-phase velocity depends strongly on viewing distance, such otolith-dependent nystagmus typically becomes stronger upon convergence. Normal subjects sometimes exhibit vertical nystagmus, probably of otolithic origin, when supine in darkness.

Gaze-evoked nystagmus occurs when a person attempts to hold the eyes in an eccentric position, such as far to the right. Here, the velocity-position integrator is leaky and does not generate a steady position signal to hold the eyes in the desired position. The eyes then drift back toward center, under the influence of elastic forces in the orbit, until corrected by a saccade back to the desired eccentric position. After extended eccentric gaze holding, a rebound nystagmus may be present upon straight-ahead gaze. Less intense forms of gaze-evoked and rebound nystagmus are not unusual in normal subjects.

Dynamics and Models of Nystagmus

It is apparent that in nystagmus, the slow phases do not take the eyes as far as they can go in one direction before being interrupted by a fast phase, and the fast phases do not occur at a regular rate. The rules – whether deterministic or random – that determine these properties are not completely understood.

One model for vestibular nystagmus is based on a 'center of interest,' a point to which the fast phases take the eyes in anticipation of seeing something interesting on which to then maintain gaze. Fast phases are generated when an internal neural copy of eye position, minus the center of interest, reaches a (variable) threshold. Another suggestion is that fast phases extend the linear range of the VOR, by triggering a fast phase when the eyes move far enough that neural responses on either side of the midline start to approach inhibitory cut-off.

Given a set of consecutive values, it is possible to predict, to a limited extent, the future values of fast-phase beginning and ending positions. However, randomized versions of these sequences, which retain linear autocorrelations, are as predictable. This indicates that each of these sequences is a temporally correlated random process (a random walk). On the other hand, fast-phase intervals are essentially unpredictable, as is characteristic of an uncorrelated random process. Although this has been demonstrated strictly for OKN, the interpretation in terms of a visual 'center of interest' as in vestibular nystagmus is evident.

See also: Basal Ganglia and Oculomotor Control; Brainstem Control of Eye Movements; Canal–Otolith Interactions; Cerebellum and Oculomotor Control; Cortical Control of Eye Movements; Eye and Head Movements; Eye Movement Disorders; Optokinetic Eye Movements; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vestibular System; Vestibulo-Ocular Reflex; Visual–Vestibular Interactions.

Further Reading

Bles W and Kapteyn TS (1977) Circular vection and human posture. I: Does the proprioceptive system play a role? *Agressologie* 18: 325–328.

Brandt T, Buchele W, and Arnold F (1977) Arthrokinetic nystagmus and ego-motion sensation. *Experimental Brain Research* 30: 331–338.

Bridgeman B (1995) A review of the role of efference copy in sensory and oculomotor control systems. *Annals of Biomedical Engineering* 23: 409–422.

Cheng M and Outerbridge JS (1974) Inter-saccadic interval analysis of vestibular nystagmus. *Acta Otolaryngologica (Stockholm)* 77: 348–353.

Chun KS and Robinson DA (1978) A model of quick phase generation in the vestibuloocular reflex. *Biological Cybernetics* 28: 209–221.

Clement RA, Abadi RV, Broomhead DS, and Whittle JP (2002) A new framework for investigating both normal and abnormal eye movements. *Progress in Brain Research* 140: 499–505.

Clement RA, Whittle JP, Muldoon MR, Abadi RV, Broomhead DS, and Akman O (2002) Characterisation of congenital nystagmus waveforms in terms of periodic orbits. *Vision Research* 42: 2123–2130.

Galiana HL (1991) A nystagmus strategy to linearize the vestibulo-ocular reflex. *IEEE Transactions on Biomedical Engineering* 38: 532–543.

Lackner JR (1977) Induction of illusory self-rotation and nystagmus by a rotating sound-field. *Aviation Space and Environmental Medicine* 48: 129–131.

Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn. Oxford: Oxford University Press.

Money KE and Myles WS (1974) Heavy water nystagmus and effects of alcohol. *Nature* 247: 404–405.

Pyykkö I and Dahlen A-I (1985) Intrabeat relationship of post-rotatory nystagmus in normal subjects. *Acta Otolaryngologica (Stockholm)* 99: 74–82.

Pyykkö I and Dahlen A-I (1985) Intrabeat relationship of post-rotatory nystagmus in patients with neurological disorders. *Acta Otolaryngologica (Stockholm)* 99: 113–121.

Raphan T, Matsuo V, and Cohen B (1979) Velocity storage in the vestibulo-ocular reflex arc (VOR). *Experimental Brain Research* 35: 229–248.

Trillenberg P, Gross C, and Shelhamer M (2001) Random walks, random sequences, and nonlinear dynamics in human optokinetic nystagmus. *Journal of Applied Physiology* 91: 1750–1759.

Trillenberg P, Zee DS, and Shelhamer M (2002) On the distribution of fast-phase intervals in optokinetic and vestibular nystagmus. *Biological Cybernetics* 87: 68–78.

Wade N and Tatler B (2005) *The Moving Tablet of the Eye: The Origins of Modern Eye Movement Research*. New York: Oxford University Press.

Obsessive–Compulsive Disorder

D Mataix-Cols, King's College London, London, UK
O A van den Heuvel, VU University Medical Center, Amsterdam, The Netherlands
D S van Grootheest, Vrije Universiteit, Amsterdam, The Netherlands
I Heyman, Maudsley Hospital, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Obsessive–compulsive disorder (OCD) is a complex and heterogeneous psychiatric disorder characterized by obsessions and compulsions (also known as 'rituals'). Obsessions are unwanted ideas, images, or impulses which repeatedly enter an individual's mind. Although recognized to be self-generated, they are experienced as 'egodystonic' (out of character and distressing). Compulsions are repetitive behaviors or mental acts which are often intended to neutralize anxiety provoked by the obsessions (Figure 1). These rituals are often driven by rules that must be applied rigidly. In order to qualify for the diagnosis, the symptoms must be disabling.

Although the core features of obsessions and compulsions appear to be remarkably consistent throughout the life span, in both sexes and in different cultures and races, the detailed content of these symptoms is varied. Two patients with OCD may present with completely nonoverlapping symptom profiles. However, more commonly, patients experience multiple types of obsessions and compulsions. The most common types of obsessions and compulsions are listed in Table 1. These symptoms tend to group in predictable ways. Indeed, multiple factor and cluster-analytical studies have identified at least four relatively independent and temporally stable symptom dimensions: (1) contamination obsessions and washing/cleaning compulsions; (2) aggressive, sexual and religious obsessions and related compulsions (often checking); (3) obsessions concerning a need for symmetry or exactness, ordering/arranging, repeating, and counting compulsions; and (4) hoarding and collecting obsessions and compulsions (Figure 2).

Aggressive and sexual obsessions in OCD must be differentiated from violent thoughts occurring in other disorders, such as urges to hurt people in psychopathy or to abuse children in pedophilia. People with OCD fear that they might commit an offense but do not carry out the feared act, and spend an excessive amount of time and energy resisting and controlling their behavior to avoid the risk of harm.

OCD often occurs together with other complicating conditions, such as depression. Screening for and treating these comorbidities is an important part of the management of the disorder (see Table 2). A majority of people with OCD of all ages understand the senseless nature of their repetitive, unwanted behaviors and intrusive, recurrent thoughts. This may lead to shame, reluctance to seek help, and poor recognition by health professionals. People with OCD have long delays in accessing effective treatments. A recent study found that the mean duration from onset of symptoms to receiving treatment was 14 years, although younger patients had access to treatment sooner.

Individuals with OCD frequently present to nonpsychiatrists for treatment (Table 3) and their symptoms often go undetected. There is a need for greater awareness of OCD in a range of nonpsychiatric health care settings, and clinicians need to be confident about recognizing it. OCD can occur throughout the life span, with children as young as 6 or 7 years of age presenting with the characteristic impairing symptoms. At the other end of the age range, cases may present for the first time in old age. A majority of adults report the onset in childhood or adolescence. OCD can result in significant disability, and the World Health Organization rates it as one of the top 20 most disabling diseases. If untreated, OCD generally persists, yet there are effective, evidence-based psychological and pharmacological treatments.

Recent epidemiological studies report prevalence rates of 0.8% in adults and in 0.25% of 5- to 15-year-old children, although earlier studies have suggested rates as high as 1–3% in adults and 1–2% in children and adolescents. The causes of OCD are unknown but, like in most complex psychiatric disorders, are likely to stem from a combination of genetic, neurobiological, cognitive-behavioral, and environmental factors.

Genetics of OCD

The influence of genetic factors in OCD has been suggested since the earliest descriptions of the disorder, and a number of study designs have been employed to determine to what extent OCD is heritable. First, family studies have convincingly shown that OCD runs in families – that is, first-degree relatives of OCD patients have an elevated risk of having OCD or subclinical obsessive–compulsive symptoms. These studies, however, do not allow disentangling whether this familiarity is due to genetic or environmental factors. For this purpose adoption or twin

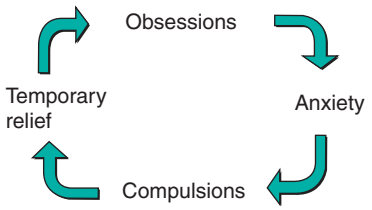

Figure 1 The obsessive–compulsive disorder cycle. Obsessions are intrusive thoughts (ideas, images, or impulses) which repeatedly enter an individual's mind against his/her will. These generate significant levels of anxiety and are difficult to dismiss. Compulsions, or rituals, are repetitive acts that are performed in an attempt to reduce the anxiety caused by the obsessions, but the relief is only temporary. Later in the course of obsessive–compulsive disorder, rituals can become more automatic and can increase, rather than reduce, the anxiety. Adapted from Heyman I, Mataix-Cols D, and Fineberg NA (2006) Obsessive–compulsive disorder. *British Medical Journal* 333: 424–429.

Table 1 Most common types of obsessions and compulsions in a large sample of patients with OCD (n = 354)

Obsession	*N*	*Percent*	*Compulsion*	*N*	*Percent*
Contamination	208	59	Cleaning	212	60
Aggressive	246	69	Checking	253	71
Sexual	68	19			
Religious	95	27			
Symmetry	159	45	Ordering	121	34
			Repeating	176	50
			Counting	125	35
Hoarding	77	22	Hoarding	73	21
Somatic	103	30			

Source: Mataix-Cols D, Baer L, Rauch SL, et al. (1999) Use of factor-analyzed symptom dimensions to predict outcome with serotonin reuptake inhibitors and placebo in the treatment of obsessive-compulsive disorder. *American Journal of Psychiatry* 156: 1409–1416.

studies are needed. Adoption studies are generally rare, and to our knowledge no such studies have been published concerning OCD. Twin studies, which compare the concordance of OCD in monozygotic versus dizygotic twins, have been conducted using either a categorical approach (presence or absence of OCD), or a dimensional approach (using the whole distribution of OC symptoms in a population). Published twin studies using a categorical approach usually suffer from several methodological problems. However, more statistically sound twin studies with a dimensional approach have clearly demonstrated the importance of genetic factors for OC symptoms, with heritabilities ranging from 45% to 65% in children and 27% to 47% in adults.

Genetic linkage analysis provides a powerful approach to elucidate the underlying genetic factors in inherited disorders. Linkage studies are designed to determine whether a behavioral phenotype, such as psychiatric disorder or a dimensional trait, is physically linked to a genetic marker, a segment of DNA with a known physical location on a chromosome. Linkage studies on OCD are rare and have been performed in relatively small clinical samples. Two independent genome scans identified a region of chromosome 9p24 that had a suggestive linkage to early-onset OCD. Recently, a large affected sib-pair sample from 219 families from the OCD Collaborative Genetics Study showed suggestive evidence for linkage in several regions, including 1q, 3q, 6q, 7p, and 15q. The same group also found significant linkage to compulsive hoarding on chromosome 14. These findings strengthen the notion that OCD is a complex and heterogeneous disorder.

Association studies offer an alternative strategy for studying genetic factors, and the aim is to demonstrate

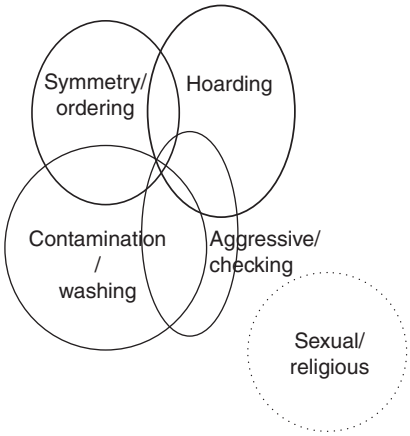

Figure 2 Schematic representation of the major symptom dimensions of obsessive–compulsive disorder. Most studies consistently identify four symptom dimensions (continuous lines); some others identify a fifth dimension consisting of sexual and religious obsessions (dashed line), but more research is needed to determine the validity of this dimension. Note the overlap between the dimensions: monosymptomatic patients are very rare.

a significantly different distribution of candidate gene variants in affected (OCD) and unaffected (non-OCD) individuals. A large number of genetic association studies on OCD have been conducted in various populations, mainly investigating the possible contribution of genes to both the serotonergic system and the dopaminergic system. Unfortunately, the results of most association studies in OCD are inconsistent and inconclusive, possibly due to a number of reasons, including small sample sizes, the use of mixed populations, failure to account for multiple testing, and the clinical heterogeneity of OCD. Recently, more statistically sound association studies have been conducted focusing on glutamatergic transmission.

Table 2 Conditions commonly occurring with OCD

Condition	*Frequency (%)*
Depression	50–60
Specific phobia	22
Social phobia	18
Eating disorder	17
Alcohol dependence	14
Panic disorder	12
Tourette's disorder	7
Schizophrenia	14

Source: Piggott TA, L'Hereux F, Dubbert B, et al. (1994) Obsessive–compulsive disorder: Comorbid conditions. *Journal of Clinical Psychiatry* 55: 15–27.

Table 3 Nonpsychiatrists likely to see patients with OCD

Professional	*Reason for consultation*
General practitioner	Depression, anxiety
Dermatologist	Chapped hands, eczema, compulsive hair pulling
Cosmetic surgeon	Concerns about appearance
Oncologist	Fear of cancer
Genito-urinary specialist	Fear of HIV/AIDS
Neurologist	OCD associated with Tourette's disorder
Obstetrician	OCD during pregnancy or puerperium
Gynecologist	Fear of contamination, vaginal discomfort from douching

Adapted from Heyman I, Mataix-Cols D, and Fineberg NA (2006) Obsessive–compulsive disorder. *British Medical Journal* 333: 424–429.

Two independent studies found an association of glutamate transporter gene *SLC1A1*, particularly in males. The encoding region for *SLC1A1* lies on chromosome 9p24, a region where suggestive linkage was reported earlier. A meta-analysis of all published case-control data on catechol-*O*-methyltranferase (COMT) found evidence of an association between the Met158 allele of COMT and OCD. The association was present in men but not in women. COMT has a role in cortical dopamine signaling and information processing. Another study reported an association between oligodendrocyte lineage transcription factor 2 (OLIG2) and OCD, especially when Tourette's disorder was absent. OLIG2 has an essential regulation role in the development of cells that produce white matter (myelin). The *OLIG2* gene is also highly expressed in the brain regions implicated in OCD.

In sum, it is clear from family and twin studies that genetic factors play a role in the etiology of OCD. As statistically sound linkage and association studies have just started to be published, it is too early to predict which specific genes are clearly implicated in OCD etiology. Furthermore, it seems essential for genetic research on OCD to take into account the heterogeneity of the disorder. For example, the use of OCD dimensions as quantitative phenotypes could be a promising approach for future genetic studies.

Functional Neuroanatomy of OCD

Current neurobiological theories of OCD suggest that specific frontal–subcortical circuits are involved in the symptoms and cognitive deficits associated with the disorder. These theories arose from various sources of evidence: the presence of OCD symptoms in some neurological conditions (Tourette's syndrome, Huntington's disease, Sydenham's chorea, and other basal ganglia disorders), the emergence of OCD-like behaviors in patients with focal brain injury, and the fact that surgical interventions that interrupt these frontal–subcortical circuits improve both mood and OCD symptoms. However, the strongest support for these models derives from the advent of modern neuroimaging techniques, which have provided a direct window into the OCD brain *in vivo*. Currently, the most widely accepted neuroanatomical model of OCD proposes the involvement of a direct and an indirect cortico–striato–thalamic pathway. In the direct pathway, an excitatory glutamatergic signal projects to the striatum, sending an inhibitory γ-aminobutyric acid (GABA)ergic signal to the internal part of the globus pallidus. This results in a decreased inhibition (disinhibition) of the thalamus and thus an increased excitatory effect on the prefrontal cortex (Figure 3). In the indirect pathway, the striatum projects an inhibitory signal to the external part of the globus pallidus and the subthalamic nucleus, sending an excitatory signal to the internal part of the globus pallidus. The net effect is an increased inhibition of the thalamus and decreased excitation on the prefrontal cortex. It is hypothesized that the direct pathway functions as a self-reinforcing positive feedback loop and contributes to the initiation and continuation of behaviors, whereas the indirect pathway provides a mechanism of negative feedback which is important in the inhibition of behaviors and in switching between behaviors. Thus, an imbalance between these frontal–striatal circuits might mediate the symptoms of OCD: an excess tone in the direct, relative to the indirect, frontal–striatal circuit is hypothesized to result in enhanced activation of the orbitofrontal cortex, ventral striatum, and medial dorsal thalamus. Based on the positive therapeutic effects of selective serotonergic reuptake inhibitors on OCD symptomatology and the inhibitory effect of serotonin on dopamine, it is suggested that failure

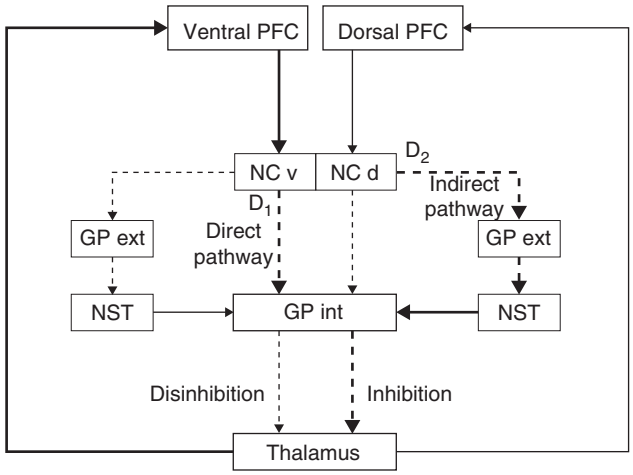

Figure 3 A widely accepted frontal–striatal model of obsessive–compulsive disorder. PFC, prefrontal cortex; NC d, dorsal part of caudate nucleus; NC v, ventral part of caudate nucleus; GP ext, external part of globus pallidus; GP int, internal part of globus pallidus; NST, subthalamic nucleus; D_1, dopamine type 1; D_2, dopamine type 2; broken lines indicate inhibitory γ-aminobutyric acid (GABA)ergic projections; solid lines indicate excitatory glutamatergic projections. From Mataix-Cols D and van den Heuvel O (2006) Common and distinct neural correlates of obsessive–compulsive and related disorders. *Psychiatric Clinics of North America* 29(2): 391–410, with permission.

of the serotonergic system results in decreased compensation of the dopaminergic influence on the frontal–striatal circuits. Dopamine (D) has a dual role on the balance between the direct and indirect frontal–striatal pathways. In the human brain, D_1 receptor expression is prominent in the ventromedial (relative to dorsolateral) prefrontal cortex and ventral (relative to dorsal) striatum. Functionally, this dopaminergic differentiation implies a stronger D_1 influence on the direct pathway of the ventromedial frontal–striatal circuit and a stronger D_2 influence on the *indirect* pathway of the *dorsolateral* frontal–striatal circuit, resulting in a hyperactivated ventral and an inhibited dorsal frontal–striatal system (see Figures 3 and 4). This corresponds with the results of functional neuroimaging studies in OCD, showing increased activation of limbic and ventral frontal–striatal regions at rest and in response to disease-relevant information and decreased responsiveness of dorsal frontal–striatal regions during executive performance.

Appealing as it may be, this model is likely to be overly simplistic, as it is based on several assumptions about the OCD brain and OCD:

- These brain structures mediate all types of OCD symptoms (i.e., OCD is a unitary nosological entity):

Psychiatric classification systems – *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV-TR) and *International Classification of Diseases* (ICD-10) – view OCD as a unitary nosological entity, but researchers are currently disputing this view. OCD may be a compendium of multiple potentially overlapping syndromes rather than a unitary entity. Each of the major symptom dimensions of OCD (i.e., contamination/washing, harm/checking, symmetry/order, and hoarding) is associated with specific patterns of comorbidity, genetic transmission, neuropsychological function, and treatment response. Several preliminary neuroimaging studies from different laboratories strongly suggest that discrete neural systems may mediate the expression of different symptom dimensions in OCD. Therefore, any revised neurobiological model of OCD will need to take into account this heterogeneity.

- These circuits represent the unique neural signature of OCD, i.e., they are not implicated in other disorders:

The picture is further complicated by the fact that OCD (currently classified as an anxiety disorder) often co-occurs with other anxiety disorders and depression. OCD has also been associated with other nonanxiety disorders that share phenomenological features, family aggregation, and (some of them) treatment response with OCD, the so-called OCD spectrum disorders. This elevated comorbidity is also reflected in a substantial degree of overlap in their neural substrates.

- There are qualitative differences between the OCD and the healthy brain:

Another assumption that remains untested is that there are qualitative differences between the OCD and healthy brains. However, the few symptom provocation studies that included healthy control groups

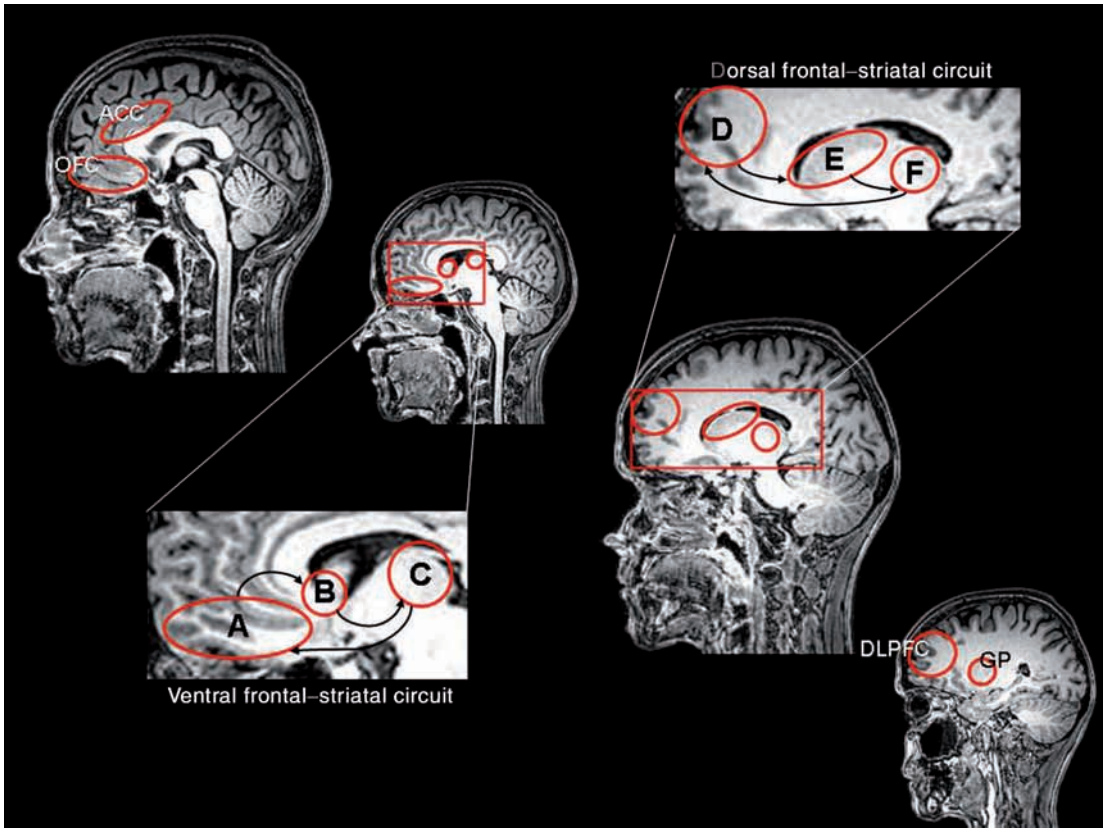

Figure 4 Schematic representation of interconnected brain regions which are part of the ventral prefrontal–striatal circuit (left) and the dorsal prefrontal–striatal circuit (right). The images are sagittal brain slices of a T_1-weighted magnetic resonance image scan, arranged from a medial view (leftmost image) to a lateral view (rightmost image). OFC, orbitofrontal cortex; ACC, anterior cingulate cortex; DLPFC, dorsolateral prefrontal cortex; GP, globus pallidus; A, orbitofrontal cortex; B, nucleus accumbens (ventral striatum); C, thalamus (medial part of mediodorsal nucleus); D, dorsolateral prefrontal cortex; E, nucleus caudatus (dorsal striatum); F, thalamus (lateral part of mediodorsal nucleus).

seem to suggest that the differences could be more quantitative than qualitative. Indeed, these studies revealed similar patterns of brain activation in both groups, although the degree of activation was, for the most part, significantly higher in the patient group. Clearly more research on this topic is needed but, if confirmed, the neuroimaging findings could reflect exaggerations of normal emotional responses to biologically relevant stimuli, rather than fundamentally abnormal neuronal responses. Furthermore, several studies have shown that the brain abnormalities seen in OCD tend to normalize after successful treatment, suggesting that these may be state dependent.

- Dysfunction in these brain regions causes OCD:

Finally, it must be emphasized that differences in brain function between patients and controls do not necessarily indicate that dysfunction in these regions causes OCD. Neuroimaging tools should be regarded as correlational techniques that allow researchers to understand the neurophysiological bases of the behaviors under study, but not necessarily their cause. At least three possible interpretations of the neuroimaging data are possible: (1) dysfunction in certain brain regions causes OCD, (2) dysfunction in certain brain regions is a consequence of OCD, and (3) another spurious variable causes both phenomena.

Immunology of OCD

In the past decade there is been renewed interest in the relationship between infections, psychiatric symptoms, and movement disorders. In particular, it was noted that some children develop OCD and/or tics in the weeks after *Streptococcus* infection. These children were categorized as suffering from pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections (PANDAS) and have been carefully investigated in a range of research studies. The most widely accepted mechanistic explanation is that antibodies, generated as part of the immune response to the streptococcal infection, cross the blood-brain barrier and interact with neuronal antigens, resulting in neurological and psychiatric symptoms. This remains a controversial area of

research, with many questions yet to be answered, such as the specificity of the psychiatric symptoms in relation to streptococcal infection, the characterization of the autoantibodies and their binding sites in human brain, and the degree of relevance to clinical practice.

Evidence for the proposed neuroimmunological mechanism includes the finding of identical antibasal ganglia antibodies in Sydenham's chorea, an autoimmune poststreptococcal illness which is characterized by a subcortical movement disorder, but also frequently incorporates psychiatric features, including obsessive–compulsive symptoms. Immunomodulatory treatments used in Sydenham's chorea have led to a trial of immunological therapy for children with PANDAS. Although this has only been attempted in a single clinical trial, one study of therapeutic plasma exchange and intravenous immunglobulin demonstrated improvement in both tic and OCD symptoms in these children. This suggested a direct pathogenic relationship between the circulating autoantibodies and the central nervous system symptoms.

There have been further attempts to characterize the autoantibodies, and it has been shown in a range of poststreptococcal illness that these antibodies recognize basal ganglia proteins of specific molecular weights. The autoantigens have been identified in some studies to be glycolytic enzymes, which also act as neuronal cell surface receptors. Other studies have proposed a mammalian lysoganglioside GM1 to be the human epitope. More research is needed to further establish the molecular mechanisms of this proposed autoimmune risk for OCD; however, all of the implicated brain binding proteins have protein homologs in streptococci, thus making them potential 'molecular mimics.'

It is unknown what proportion of childhood OCD or tics might be PANDAS; one study reported that 11% of children with tic disorders experienced worsening of tics in the 6 months following streptococcal infection. Another study has shown that patients with Tourette's syndrome or OCD were more likely than control children to have had a streptococcal infection in the 3 months before onset. The children with PANDAS had a somewhat atypical course to their OCD, with an 'on–off' trajectory to their symptoms, relapse being associated with presumed reinfection with streptococcus. Consideration has been given to whether more usually presenting OCD in children might be associated with earlier streptococcal infection. One study showed that in 50 consecutive child or adolescent patients seen in an OCD clinic, 42% had circulating antibodies to streptococcal antigens (antistreptolysin O; ASO), in contrast to 20% of healthy school children. Furthermore, 48% of these children had antibasal ganglia antibodies (ABGAs), which are only found at a rate of 3% in normal populations. This raises tantalizing questions about whether exposure to streptococcal infection could be one of the environmental factors that predispose a subgroup of children to OCD, and whether there is a specific subtype of OCD has an autoimmune cause.

There are many unanswered questions. For example, infections with streptococcus are common, yet relatively few people develop autoimmune complications. It has been proposed that genetic factors predispose vulnerable individuals to a risk of developing central nervous system complications mediated by autoreactive antibodies, following streptococcal infection. Recent work has revealed that obsessive–compulsive spectrum disorders are more common in patients with rheumatic fever, another autoimmune streptococcal illness, compared to the general population. The risk of developing OCD or a related disorder is also comparatively greater in first-degree relatives of those with rheumatic fever, demonstrating a familial relationship between these disorders, which may be genetic, although further work is required to establish this.

Currently, the clinical implications, in terms of assessment and treatment of this emerging relationship between streptococcus and OCD, remain unclear. If children present with sudden-onset OCD or tics, it is good practice to inquire about pharyngitis or other sites of infection, and if current infection is established, to treat with antibiotics as usual. Studies in which PANDAS patients have been treated prophylactically with antibiotics in the hope of preventing relapses of symptoms have given equivocal results. Similarly, immunomodulatory treatments remain experimental, and should be considered only in the context of research trials in specialist centers. Best clinical practice for children and young people with OCD, even if a streptococcal origin is suspected, is to offer the current evidence-based treatments (see later). It remains to be seen whether a subgroup of OCD with an immunological trigger has characteristics that differ from other forms of OCD, or differing treatment responsiveness.

Management of OCD

Multiple randomized controlled trials (RCTs) have established the efficacy of both a form of psychotherapy called cognitive behavior therapy (CBT) and certain medications; the medications, which inhibit the synaptic reuptake of serotonin, are the tricyclic antidepressant clomipramine and the more highly selective serotonin reuptake inhibitors (SSRIs). There is no evidence to support the efficacy of psychodynamic psychotherapy in OCD and its use is therefore not recommended.

In both adults and children, the specific CBT technique most strongly associated with good outcome in CBT studies is exposure and response prevention (E/RP). Briefly, the patient is encouraged to progressively face his/her feared situations (exposure), while refraining from performing his/her rituals (response prevention). Engaging the patients with OCD by helping them design a graded program of E/RP, and working collaboratively on easiest challenges first, are essential. Careful education about mechanisms of anxiety, understanding that repeated exposure leads to reduced anxiety, as well as reduction in obsessions, is important for success. Practice is needed, as patients with OCD will have been reinforcing their behaviors by avoiding feared situations or carrying out rituals to deal with their fears for some time. It is often needed to involve the patient's family members in this process, as many participate in (and thus help maintain) the patient's rituals.

Because many patients do not have access to specialist CBT clinics, and waiting lists for treatment are often long, there is currently much interest in developing effective tele-care and self-help programs for OCD. These have the potential to save much clinician time and to reach many more patients that would otherwise remain untreated. Indeed, recent RCTs have shown that E/RP can be successfully delivered by a therapist over the telephone and even by an interactive computer program with minimal therapist contact and without impairing efficacy.

All the SSRIs have been subject to large-scale clinical trials and have been found to be superior to placebo in both adults and children with OCD. Dose-finding studies have only been carried out in adults. Higher doses of SSRIs are needed to effectively treat OCD, compared with depression. SSRIs have largely superseded clomipramine for treating OCD because of their lesser toxicity in overdose and more favorable side-effect profile. This is especially important for children, for whom cardiac toxicity may be a risk. Head-to-head studies show equivalent efficacy and better tolerability for SSRIs relative to clomipramine. Clomipramine remains a useful option, but is usually reserved for cases in which trials of SSRIs have been ineffective.

There is some evidence supporting various medication strategies in resistant cases, including increasing the dose of the SSRI to the maximum tolerated, and switching to an alternative, as there may be idiosyncrasy in response. OCD does not respond to antipsychotics given as monotherapy. There is evidence from children and adults that adding low doses of first- and second-generation antipsychotics to SSRIs may benefit cases of resistant OCD and OCD with comorbid tics. Several studies have shown that people with OCD continue to benefit from long-term medication but that many patients relapse if medication is discontinued. For at least some cases, therefore, treatment may need to be continued indefinitely. This is not the case with E/RP, and therapeutic gains are usually maintained long-term, although some boosting sessions may be required.

Despite the demonstrated efficacy of both CBT and SSRIs, up to 40% of cases fail to respond adequately to these treatments. Recent studies suggest that patients with hoarding symptoms may be particularly treatment resistant, and new treatment approaches are being developed and tested for these patients. Some centers offer neurosurgery (cingulotomy or anterior capsulotomy) for severe, treatment-resistant OCD but, for obvious reasons, these treatments have not been evaluated in controlled trials and remain controversial. More recently a reversible form of neurosurgery consisting of deep-brain stimulation of the anterior capsule/ventral striatum has been developed and tested in OCD with promising results, but more large-scale controlled studies are needed.

See also: Anxiety: Drug Therapy; Barrel Cortex Circuits; Cognitive Neuroscience: An Overview; Cognitive Dysfunction in Psychiatric Disorders; Corticothalamic Connections: Ultrastructure; Mood Stabilizers; Psychiatric Disorders: Functional Genomics; Psychotherapeutic Approaches to Psychiatric Disorders; Tourette's Disorder.

Further Reading

Greenberg BD, Malone DA, Friehs GM, et al. (2006) Three-year outcomes in deep brain stimulation for highly resistant obsessive–compulsive disorder. *Neuropsychopharmacology* 31 (11): 2384–2393.

Hemmings SM and Stein DJ (2006) The current status of association studies in obsessive–compulsive disorder. *Psychiatric Clinics of North America* 29(2): 411–444.

Heyman I, Mataix-Cols D, and Fineberg NA (2006) Obsessive–compulsive disorder. *British Medical Journal* 333: 424–429.

Marks IM, Cavanagh K, and Gega L (2007) *Hands-on Help: Computer-Aided Psychotherapy.* London: Routledge.

Mataix-Cols D, Rosario-Campos MC, and Leckman JF (2005) A multidimensional model of obsessive–compulsive disorder. *American Journal of Psychiatry* 162: 228–238.

Mataix-Cols D, Baer L, Rauch SL, et al. (1999) Use of factor-analyzed symptom dimensions to predict outcome with serotonin reuptake inhibitors and placebo in the treatment of obsessive-compulsive disorder. *American Journal of Psychiatry* 156: 1409–1416.

Mataix-Cols D and van den Heuvel O (2006) Common and distinct neural correlates of obsessive–compulsive and related disorders. *Psychiatric Clinics of North America* 29(2): 391–410.

O'Kearney RT, Anstey KJ, and von Sanden C (2006) Behavioural and cognitive behavioural therapy for obsessive compulsive disorder in children and adolescents. *Cochrane Database of*

Systematic Reviews issue 4. Art. no. CD004856. DOI: 10.1002/14651858.CD004856.pub2.

Piggott TA, L'Hereux F, Dubbert B, et al. (1994) Obsessive–compulsive disorder: Comorbid conditions. *Journal of Clinical Psychiatry* 55: 15–27.

Steketee G and Frost R (2003) Compulsive hoarding: Current status of the research. *Clinical Psychology Review* 23(7): 905–927.

Swedo SE and Grant PJ (2005) PANDAS: A model for human autoimmune disease. *Journal of Child Psychology and Psychiatry* 46(3): 227–234.

van Grootheest DS, Cath DC, Beekman AT, et al. (2005) Twin studies of obsessive–compulsive disorder: A review. *Twin Research and Human Genetics* 8(5): 450–458.

Relevant Websites

http://www.ocfoundation.org – Obsessive–Compulsive Foundation.
http://www.ocdaction.org.uk/ocdaction/index.asp – OCD Action.
http://www.ocduk.org – OCD UK.

Octopamine and Other Monoamines in Invertebrates

T D Sotnikova and R R Gainetdinov, Duke University Medical Center, Durham, NC, USA

Crown Copyright © 2009 Published by Elsevier Ltd. All rights reserved.

Introduction

In both vertebrates and invertebrates, biogenic monoamines are important neuroactive molecules involved in neuronal communications. In invertebrates, octopamine and other monoamines can act physiologically as neurotransmitters, neuromodulators, or neurohormones. Biogenic monoamines such as octopamine, tyramine, dopamine, serotonin, and histamine regulate numerous vital functions, including cardiovascular control, movement, endocrine regulation, circadian rhythms, emotional behaviors, learning, and memory. Although the adrenergic transmitters norepinephrine and epinephrine are found at low concentrations in some invertebrate species, it is believed that their physiological roles in invertebrates are limited. Monoamines exert a large part of their effects by binding to specific receptors that primarily belong to the superfamily of G-protein-coupled receptors (GPCRs). Considerable progress has been achieved during the past few decades toward the understanding of the functional role of these monoamines in invertebrate physiology. Here, we summarize current knowledge on the biochemistry of monoaminergic pathways, regulatory mechanisms of monoaminergic transmission, receptors and signaling pathways involved in their actions, and physiological functions mediated by monoaminergic mediators in invertebrates.

Octopamine

Since its discovery more than 50 years ago in the salivary glands of *Octopus*, octopamine has been the subject of intense investigation. Although octopamine-containing neurons present only in a small subset of neurons (40–100 in insects) within the nervous system, the role of octapamine in invertebrate physiology cannot be underestimated. It is believed that octopamine and, to a lesser degree, tyramine play a pleiotropic role in invertebrate physiology that is mediated in vertebrates by adrenaline and noradrenaline. Octopamine is synthesized from the amino acid tyrosine in two steps. Decarboxylation of tyrosine by tyrosine decarboxylase (TDC) results in the production of tyramine (Figure 1). Two forms of this enzyme are known, one of which is expressed in nonneuronal cells (TDC1) and the other one (TDC2) is expressed in neurons. The second step involves the enzyme tyramine-β-hydroxylase (TβH), which converts tyramine to octopamine by the addition of a hydroxyl group in the β-position. Enzymatic inactivation of octopamine is believed to occur predominantly by N-acetylation or N-methylation. In contrast to vertebrates, monoamine oxidase (MAO) plays only a minor role in the inactivation of monoamines in invertebrates. Additional mechanisms of octopamine inactivation may involve sulfate and γ-glutamyl conjugation.

Several critical presynaptic processes governing monoaminergic transmission in mammals play a similar role in invertebrates. They include processes involved in packaging of monoamines into synaptic vesicles for subsequent storage and release, release-related events, and reuptake of released monoamines via plasma membrane transporters. Released monoamines elicit their transmitter functions by activating corresponding postsynaptic receptors, which in most cases belong to the GPCR family. All the functional elements necessary for these processes have been identified for octopaminergic transmission.

In *Drosophila*, newly synthesized octopamine is packaged into synaptic vesicles by the *Drosophila* homolog of the vesicular monoamine transporter (dVMAT). Two mRNA splice variants (dVMAT-A and-B) that differ at their C-terminus have been identified. dVMAT-A recognizes octopamine, tyramine, as well as other monoamines, including dopamine, serotonin, and histamine, as preferential substrates. Importantly, this transporter is similar to mammalian VMAT homologs with regard to pharmacological properties and can be potently inhibited by reserpine and some environmental toxins. Octopamine can be released in response to various stimuli in a calcium-dependent manner. After release and generation of corresponding signaling events at the postsynaptic cells via activation of receptors, octopamine undergoes rapid reuptake into presynaptic terminals by specific plasma membrane transporters. Reuptake is the major route of inactivation for octopamine but, at the same time, it provides recycled transmitter for subsequent release. Generally, invertebrate plasma membrane transporters share structural and pharmacological features with transporters found in vertebrates. In several insect species, such as the cabbage looper and cockroach, but not *Drosophila*, a specific phenolamine transporter that can transport octopamine with high affinity has been identified. Interestingly, since cocaine was found to interact with octopamine reuptake, the ability of cocaine to induce hyperactivity and leaf walk-off in the insects has

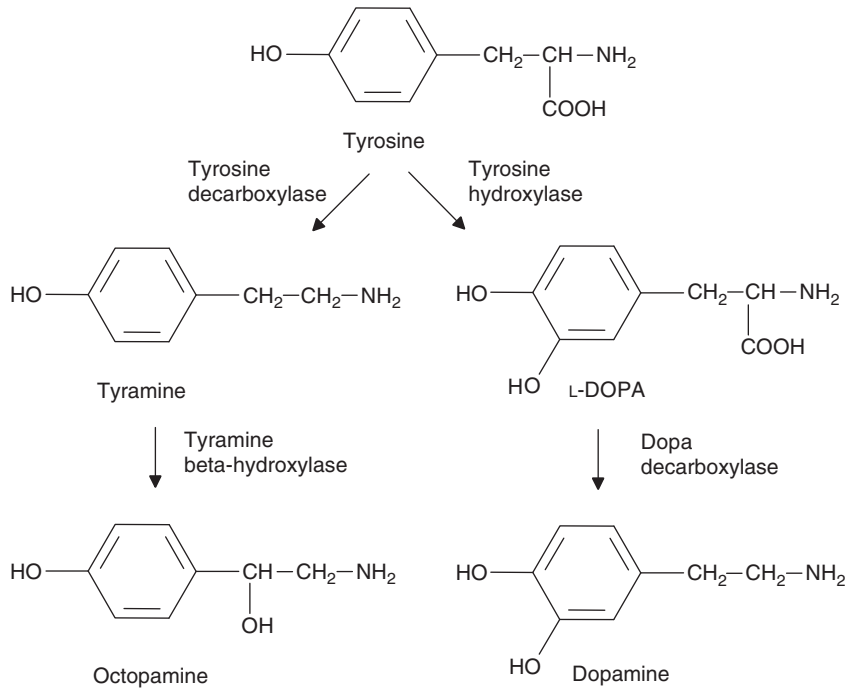

Figure 1 Synthesis of tyramine, octopamine, and dopamine in invertebrates. In the first step, tyrosine is decarboxylated by tyrosine decarboxylase to tyramine. Tyramine, in turn, is hydroxylated by enzyme tyramine β-hydroxylase to octopamine. Alternatively, as in vertebrates, tyrosine can be hydroxylated by tyrosine hydroxylase to L-DOPA. Decarboxylation of L-DOPA by DOPA decarboxylase (also known as L-aromatic amino acid decarboxylase) results in production of dopamine.

been related to octopaminergic transmission. It is believed that alkaloid of the coca plant, cocaine, has an insecticide function that prevents insects from eating the leaves via modulation of octopaminergic neurotransmission.

When octopamine is released in the synapse, it binds to corresponding GPCRs and elicits a physiological response. Octopamine receptors (OARs) share significant structural features with the corresponding adrenergic receptors. OARs have been detected in almost every peripheral organ in a large number of different insect species. Based on similarities in structure and signaling properties with vertebrate adrenergic receptors, OARs were classified as alpha-adrenergic-like octopamine receptors, beta-adrenergic-like octopamine receptors, and octopamine/tyramine or tyraminergic receptors. The most prominent subtype, alpha-adrenergic-like octopamine receptors, which is homologous to the *Drosophila* OAMB (OAR primarily expressed in the mushroom bodies), is ubiquitously expressed throughout the body and is believed to be responsible for most functions mediated by octopamine. Three splice variants of *Drosophila* OAMB have been found, and these variants may have different pharmacological properties.

Many high-affinity agonists and very few antagonists of OARs are known. They have been identified mostly in an attempt to develop insecticides that would not be effective in mammals. In fact, some OAR agonists were used as effective insecticides. Among high-affinity OAR agonists, synephrine and chlordimeform are best known. Some phenyliminoimidazolidine and aminooxazoline derivatives also display extremely high activity at OARs with K_D values in the subnanomolar range. Octopaminergic agonists that induce insecticide action typically at low doses induce hyperactivity, which is followed by the leaf walk-off syndrome, with tremor and lethality at higher concentrations. OAR antagonists do not exert insecticide actions and, in fact, very few of these compounds can inhibit octopamine action in various physiological preparations. Some OAR antagonists, such as phentolamine and mianserin, also show potent antagonistic activity at other monoaminergic receptors in invertebrates and vertebrates. Only one antagonist, epinastine, is considered to be relatively selective and is used as a tool for the dissection of the role of octopaminergic neurotransmission *in vivo* in invertebrates. Second-messenger coupling of OARs is believed to be primarily mediated by the activation of cAMP synthesis and the elevation of Ca^{2+} concentrations via involvement of G_s and G_q.

Octopamine plays a major role in essentially every organ and every function of invertebrates. Historically, the investigation of the functional roles of octopamine began with the demonstration of its action on

peripheral functions, such as muscle tension and relaxation rate. It should be noted, however, that the modulatory effects of octopamine in peripheral tissues are generally quite minor. Octopamine enhances muscle performance of the locust flight muscle, which is critical for fight-or-flight response. Stress-related enhancement of motor activity requires the change of primary energy source between rest and the onset of fight or long-term flight. Elevated octopamine levels in the hemolymph during stress induce the release of adipokinetic hormone from the corpora cardiaca. At the same time, octopamine has a direct effect on body fat, leading to the release of fatty acids into the hemolymph, that enhances the adipokinetic effect of the adipokinetic hormone. This complex adaptation to high-energy demands also involves the concomitant modulation of other organs, especially those required for transportation of metabolites and oxygen, namely the heart and the trachea/air sacs, which are in part modulated by octopamine. As a result, octopamine is involved in modulation of multiple organs to orchestrate the fight-or-flight response in times of high-energy demands.

One of the best characterized peripheral targets of octopamine action is the firefly light organ, in which octopamine functions as a true transmitter in a very straightforward paradigm. It is well established that octopamine is released from unpaired neurons located in the abdominal ganglion that innervate the light organ of the firefly. Electrical stimulation of the corresponding nerve, as well as octopamine application, induces light emission from the organ, thereby providing a simple physiological output that is easily quantifiable. Importantly, OARs are the only monoaminergic receptors expressed in this tissue, making it an excellent model in the assessment of OAR pharmacology.

Octopaminergic transmission is also critical for essentially all sensory inputs. Acoustic, visual, olfactory, gustatory, tactile, and proprioceptive functions can be modulated by octopamine. Involvement of octopamine in the modulation of the olfactory pathway in the brain may contribute to its role in learning and memory. In addition, octopamine modulation can be involved in the pheromone sensing pathway.

Octopamine has the highest concentration and plays a prominent role within the nervous system. It is well-known that octopamine injected into different parts of the insect nervous system can elicit different behaviors. Generally, octopamine controls the initiation and maintenance of various rhythmic behaviors in the central nervous system of insects. Among other important functions, octopamine is involved in the control of sensitization toward novel stimuli. The most impressive demonstration of such action was performed using the kicking reflex in locust that normally undergoes habituation and then can be completely restored by octopamine application. The role of octopamine in learning and memory can be demonstrated in the honeybee, in which injection of octopaminergic agonists can facilitate both memory formation and retrieval of memory. Octopamine also plays an important role in social behaviors. Octopamine enhances kin recognition in bees and modulates the behavior toward foreign bees. Furthermore, octopamine appears to be involved in the control of division of labor in a honeybee colony. Another related behavior that may also be controlled by octopamine is a hygienic behavior. In arthropods, octopamine is also involved in regulation of aggressive behaviors. Whereas serotonin induces aggressive behavior, octopamine seems to be involved in a submissive behavior in crustaceans. Thus, the role of this compound in the fight-or-flight reaction in insects might be related not only to modulation of peripheral targets, such as motor output or energy supply, but also to the direct influence of octopamine on the central nervous system-mediated functions.

Tyramine

The first step in the biosynthesis of octopamine, tyrosine decarboxylation by TDC, results in the formation of tyramine (Figure 1). In addition to conversion to octopamine, N-acetylation, N-methylation, potential MAO-mediated inactivation, and β-alanine conjugation seem to be important pathways for tyramine degradation. Since tyramine serves as a precursor of octopamine, demonstration of the role of tyramine as a neurotransmitter has been difficult. For example, TDC deletion mutants would result in elimination of both octopaminergic and tyraminergic neurotransmission, thereby complicating delineation of tyramine functions. However, accumulating evidence strongly indicates a neurotransmitter role of tyramine. Although it is not clearly established if both tyramine and octopamine act as transmitters in the same neurons and may be co-released, immunohistochemical studies have identified a set of larval neurons that contain tyramine but not octopamine, suggesting the existence of specific tyraminergic neurons capable of exclusive tyramine release. Tyramine can be accumulated in vesicles via vesicular monoamine transporters, such as dVMAT-A in *Drosophila*. Rapid reuptake of released tyramine by plasma membrane transporters has also been described. It has been shown that both the cocaine-sensitive *Drosophila* dopamine transporter (dDAT) and the specific phenolamine transporter discovered in the cabbage looper can transport tyramine with high affinity.

Actions of tyramine are mediated via GPCR tyraminergic receptors (TARs), which may also be involved, in part, in octopamine transmission. These receptors share structural features with the α-adrenergic receptors but have little similarity with mammalian trace amine receptors that can also be activated by tyramine. Only one TAR has been found in *Drosophila*, and thus all tyraminergic neurotransmission seems to be mediated exclusively through this receptor. In the *honoka* mutant (a *Drosophila* strain) that has a significantly lower level of TAR expression, a slight hyperactivity, impairment in some olfactory tasks, and reduced tyramine-induced muscle contraction were observed. In addition, an impairment of responses to repellents such as ethyl acetate was found, and the inhibitory effects of tyramine on the excitatory junction potential of the larval body wall muscle were significantly affected.

TAR pharmacology is quite limited, and no selective agonists (other than tyramine) are currently available. Nevertheless, some compounds among 5-phenyloxazoles show relatively high activity at TARs. Furthermore, a high-affinity antagonist of the TAR has been identified. Yohimbine, an indol alkaloid of the yohimbe tree, is mostly known for its antagonistic action on vertebrate α_2-adrenergic receptors. In invertebrates, yohimbine is a potent antagonist of TAR, and in *Drosophila* it can induce physiological effects similar to those observed in *honoka* mutants. It is believed that invertebrate TARs are coupled to the G_i protein, thereby inducing inhibition of the adenylate cyclase. Thus, regarding adenylate cyclase activity, tyramine signaling is inhibitory, whereas octopamine signaling is mostly stimulatory.

In contrast to octopamine, which induces an increase in insect activity, tyramine induces a decrease in activity. It is believed, in fact, that the balance between tyramine and octopamine activity is critical for normal locomotion in *Drosophila* and similar regulation may be important also for other functions and behaviors. Although few specific functions for tyramine have been identified, like OARs, TARs are ubiquitously distributed within the nervous system. Interestingly, tyramine has been shown to contribute to cocaine sensitization and the regulation of locomotor response to ethanol in *Drosophila*. Tyramine may also be involved in gustatory and visual sensitivity. Finally, tyramine is found in nonneuronal cells. In fact, it may act as a paracrine signal in the *Drosophila* renal system.

Dopamine

In invertebrates, dopamine synthesis processes are similar to those in mammalian neurons (**Figure 1**). It begins with the conversion of the amino acid tyrosine to L-dihydroxyphenylalanine (L-DOPA), the immediate dopamine precursor, by rate limiting in dopamine synthesis enzyme tyrosine hydroxylase (TH). This process requires the coenzyme tetrahydrobiopterin, which in turn is dependent on GTP cyclohydrolase activity. Subsequently, L-DOPA is decarboxylated to dopamine by the enzyme DOPA decarboxylase (DDC), which is also known as aromatic amino acid decarboxylase (AADC). Major metabolites of dopamine, such as dihydroxyphenylacetic acid, 3-methoxytyramine, and homovanillic acid, are found at low levels in invertebrates. Although several MAO and catechol-*O*-methyltransferase homologues exist in the *Caenorhabditis elegans* genome, these enzymes that are critical for the catabolism of dopamine in vertebrates have not been identified yet in invertebrates. In *Drosophila*, a large part of dopamine is inactivated by the enzyme *β*-alanyl-dopamine-synthetase to generate *β*-alanyl-dopamine. Degradation of dopamine also occurs via *N*-acetyltransferase-mediated conversion of dopamine to *N*-acetyl-dopamine.

Most of the studies aimed at the identification and analysis of molecules orchestrating dopamine signaling in invertebrates have been performed in the nematode *C. elegans*. Dopamine signaling in *C. elegans*, as in vertebrates, arises from a small number of cells that exert modulatory control over a larger network involved in various physiological functions. The experimental model was instrumental in the isolation and characterization of molecules important for dopamine synthesis, packaging, reuptake, and signaling that are essential for dopamine-supported behaviors. Among 302 neurons found in *C. elegans*, 8 dopamine neurons in the hermaphrodite with an additional set of 6 dopamine neurons specific to males were identified. Dopamine neurons make extensive connections that are annotated in WormBase.

A forward genetic approach involving screening for mutations that induce a loss of dopaminergic phenotype revealed that essentially all critical proteins required for dopaminergic transmission in mammals are present in invertebrates. Synthesized dopamine is rapidly packaged into synaptic vesicles by a VMAT, in which dopamine is stored available for release following neuronal depolarization. The first dopamine-related mutant identified (*cat-1*) displayed a loss of dopamine accumulation in vesicles. CAT-1 protein shows relatively high homology to human VMAT1 and VMAT2 and can support uptake of both dopamine and serotonin, which is sensitive to reserpine. Thus, it is believed that CAT-1 plays the same role as mammalian VMATs by packaging dopamine and other monoamines into synaptic vesicles. *Cat-2* has been identified as a TH homolog based on structural

similarity. Importantly, *cat-2* mutants display strikingly reduced dopamine levels, recapitulating the effect of TH deletion in mice. CAT-4 is believed to be a GTP cyclohydrolase enzyme responsible for production of the TH cofactor tetrahydrobiopterin. Similarly, AADC (*bas-1*) and the DAT (*dat-1*) homologues have been identified in *C. elegans*. It should be noted that homologues of some of these regulatory proteins were also identified in other invertebrate species. For example, *Drosophila* homologues of VMAT (dVMAT) and DAT (dDAT) are known.

Actions of dopamine in *C. elegans* are mediated by four dopamine receptors that are classified as dop-1 through dop-4. Several splice variants of these receptors have been identified (four for dop-1, two for dop-2, and four for dop-3). Although pharmacological properties of many dopamine agonists and antagonists are different at these receptors from their mammalian counterparts, dop-1 and dop-4 might be considered as analogous to mammalian D1 dopamine receptors with the major signaling pathway involving stimulation of adenylate cyclase, with dop-2 and dop-3 displaying signaling properties of mammalian D2 class dopamine receptors that induce inhibition of adenylate cyclase. By using knock-out animals, specific functions mediated by dopamine via these receptors have been delineated. Dop-1 was implicated in habituation to nose tap response, suggesting that this receptor is important for maintenance of the avoidance response. Whereas no clear function has been identified for dop-2, dop-3 was implicated in dopamine-induced paralysis and basal slowing mechanisms. Dop-4 receptor may be involved in the regulation of pharyngeal pumping, chemosensation, defecation, oxygen sensing, and male mating behavior. In general, dopamine-mediated functions in *C. elegans* include egg-laying behaviors, defecation, basal motor activity, sensation/response to food sources, and habitutation to touch.

Serotonin

Identical biochemical pathways exist in both vertebrates and invertebrates to synthesize the indolamine 5-hydroxytryptamine (5-HT; serotonin) (Figure 2). The first and rate-limiting reaction involves hydroxylation of tryptophan at the indole ring in the 5′ position by tryptophan hydroxylase. Then, 5-hydroxytryptophan is decarboxylated by DDC to serotonin. Since DDC plays a similar role in the decarboxylation of L-DOPA to dopamine (Figure 1), a defect of this enzyme simultaneously affects synthesis of both dopamine and serotonin. Once synthesized, serotonin is transported into synaptic vesicles by the vesicular transporters, and it has been shown that CAT-1 can effectively uptake serotonin into vesicles in *C. elegans*,

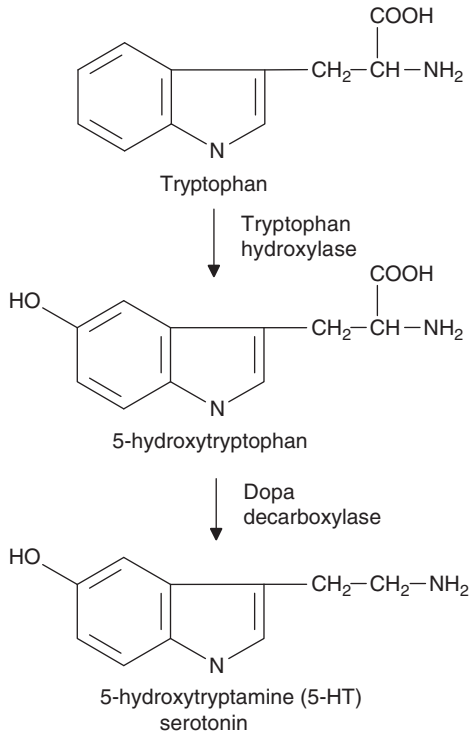

Figure 2 Synthesis of serotonin (5-hydroxytryptamine (5-HT)) in invertebrates. L-Tryptophan is hydroxylated by tryptophan hydroxylase to 5-hydroxytryptophan, which is in turn decarboxylated by DOPA decarboxylase to generate serotonin.

whereas dVMAT may have a similar function in *Drosophila*. Selective reuptake of released serotonin occurs via specific serotonin transporters, and in *Drosophila* the serotonin transporter (dSERT) has been cloned.

Most serotonergic neurons in insects are interneurons in the brain and the ventral cord with multiple projections to most parts of the nervous system. Both in mammals and in invertebrates, serotonin acts through multiple receptor subtypes and mediates a variety of functions. Altogether, 12 serotonin GPCR receptor genes have been cloned from *Drosophila*, mollusks (*Lymnaea* and *Aplysia*), and nematodes (*C. elegans* and *Ascaris*). Initial attempts to classify these receptors by using pharmacological tools developed for vertebrate receptors were largely unsuccessful. Gene structure and amino acid sequence seem to provide a more definitive way to classify invertebrate 5-HT receptors. Serotonin-mediated functions in invertebrates include aggression in lobsters, feeding and learning in snails and nematodes, and locomotion in nematodes and *Drosophila*. Serotonin is also critically involved in basic developmental processes, such as cell proliferation, migration, differentiation, and morphogenesis, both in invertebrates and in mammals.

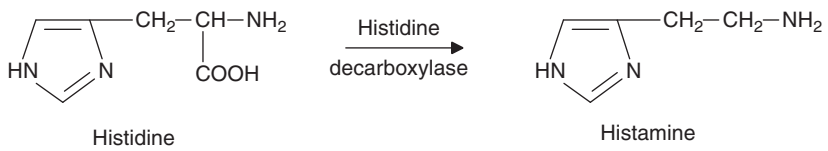

Figure 3 Synthesis of histamine in invertebrates. One-step synthesis involves decarboxylation of L-histidine by histidine decarboxylase to histamine.

Histamine

As in vertebrates, a single enzymatic step converts L-histidine to histamine in invertebrates (Figure 3). The decarboxylation reaction is mediated by histidine decarboxylase (HDC). In invertebrates, histamine has been shown to be a major neurotransmitter released from photoreceptors in response to illumination. Histamine is synthesized only in a small population of neurons in the insect brain. For example, the brain of the honeybee contains only approximately 150 histaminergic neurons. Histaminergic neurons project to many cerebral areas and may play a role in hormonal secretion, thermoregulation, cardiovascular control, and memory functions. Like other monoamines, histamine is packaged in synaptic vesicles by vesicular monoamine transporters, such as dVMAT in *Drosophila*, and can be released in response to various stimuli. Whereas four GPCRs mediate histamine functions in vertebrates, no histaminergic GPCRs have been identified in invertebrates. Two histamine-gated chloride channels were identified as putative histamine receptors. One of them is localized in the lamina of the eye, where histamine functions as a neurotransmitter released from photoreceptors. Interestingly, mutation in the gene encoding HDC results in elimination of histamine but does not cause lethality, suggesting a limited role of histamine during development or in the regulation of vital functions in adults. At the same time, HDC mutants are blind, indicating disrupted synaptic transmission between photoreceptor terminals and postsynaptic monopolar cells in the lamina or medulla. In addition, these mutants demonstrate deficits in mechanosensory transduction. In crickets, histamine plays an important role in the inhibition of an auditory interneuron within the prothoracic ganglion, indicating that histamine is involved not only in sensory transmission but also in the inhibitory modulation of interneurons in insects.

Summary

Biogenic amines are important mediators of neuronal signaling in the central and peripheral nervous systems in many invertebrate species. In general, the organization of the monoaminergic system in invertebrates shows remarkable similarity to that in vertebrates, with the most notable difference reflecting a limited role of adrenergic transmission. It is believed that the role of norepinephrine and epinephrine in invertebrates is generally played by octopamine and tyramine. Although a relatively small number of neurons contain monoaminergic transmitters in invertebrates, they nevertheless supply most of the nervous system with a fine network of projections. This organization is similar to that of mammalian adrenergic transmission, in which a relatively small number of noradrenergic neurons are able to innervate all major parts of the brain. Essentially all major proteins involved in synthesis, vesicular storage, and release, reuptake, and metabolism of monoamines have been identified in invertebrates, showing that the universal nature of molecular mechanisms of monoaminergic transmission is preserved from invertebrates to mammals. As in vertebrates, multiple receptor subtypes (mostly GPCRs) mediate the physiological actions of biogenic amines in invertebrates. It is believed that further characterization of monoaminergic transmission in invertebrates could provide important novel clues for understanding the nature of monoaminergic transmission in more complex organisms such as mammals.

See also: Dopamine – CNS Pathways and Neurophysiology; Histamine Receptors and their Ligands: Mechanisms and Applications; Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Monoamines; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Trace Monoamines and Receptors in Mammalian CNS.

Further Reading

Axelrod J and Saavedra JM (1977) Octopamine. *Nature* 265: 501–504.

Buznikov GA, Lambert HW, and Lauder JM (2001) Serotonin and serotonin-like substances as regulators of early embryogenesis and morphogenesis. *Cell and Tissue Research* 305: 177–186.

Evans PD and Maqueira B (2005) Insect octopamine receptors: A new classification scheme based on studies of cloned *Drosophila* G-protein coupled receptors. *Invertebrate Neuroscience* 5: 111–118.

Horvitz HR, Chalfie M, Trent C, and Evans PD (1982) Serotonin and octopamine in the nematode *Caenorhabditis elegans*. *Science* 216: 1012–1014.

McDonald PW, Jessen T, Field JR, and Blakely RD (2006) Dopamine signaling architecture in *Caenorhabditis elegans*. *Cellular and Molecular Neurobiology* 26: 593–618.

Nathanson JA and Greengard PA (1973) Octopamine-sensitive adenylate cyclase: Evidence for a biological role of octopamine in nervous tissue. *Science* 180: 308–310.

Nathanson JA, Hunnicutt E, Kantham L, and Scavone C (1993) Cocaine as a naturally occurring pesticide. *Proceedings of the National Academy of Sciences of the United States of America* 90: 9645–9648.

Osborne RH (1996) Insect neurotransmission: Neurotransmitters and their receptors. *Pharmacology and Therapeutics* 69: 117–142.

Roeder T (2005) Tyramine and octopamine: Ruling behavior and metabolism. *Annual Review of Entomology* 50: 447–477.

Saraswati S, Fox LE, Soll DR, and Wu CF (2004) Tyramine and octopamine have opposite effects on the locomotion of *Drosophila* larvae. *Journal of Neurobiology* 58: 425–441.

Relevant Websites

http://flybase.bio.indiana.edu – FlyBase: A Database of *Drosophila* Genes & Genomes.

http://www.wormbase.org – WormBase.

Oculomotor Control: Anatomical Pathways

J C Lynch, University of Mississippi Medical Center, Jackson, MS, USA

© 2009 Elsevier Ltd. All rights reserved.

In humans and other primates, the central retina includes a small region, the fovea, in which the receptors and related neural circuitry are highly specialized to provide a level of visual analysis many times more detailed than is possible in the more peripheral portions of the retina. This foveal region of high-acuity vision is smaller than one might imagine, subtending only a few degrees of visual angle, about the size of a 3 cm diameter coin held at arm's length. It is only within this foveal region that our vision is sharp enough for us to recognize a friend's face from a distance of 15 ft or to read normal size type. Consequently, humans and other primates make frequent eye movements as they visually scan their surroundings. These eye movements bring the retinal images of objects of interest into the region of sharpest vision, the fovea, and maintain such images on the fovea if either the visual object or the observer should move.

Most eye movements fall into one of five categories: (1) saccadic eye movements, which quickly shift the center of gaze (i.e., the fovea) from one visual object to another; (2) pursuit, or tracking, movements, which maintain the image of a moving object on the fovea so it can be seen clearly while it is followed with the eyes; (3) vergence movements, which maintain binocular fusion when a visual target moves closer to or farther away from the observer; (4) vestibulo-ocular movements, which compensate for movements of the head when the gaze is fixed on a target of interest; and (5) optokinetic movements, which compensate for slow movements of the entire visual field (as when watching the scenery from a moving car). In addition, saccades can be made voluntarily in total darkness, and numerous neurological conditions can result in abnormal spontaneous eye movements. Some eye movements are voluntary in nature; others are purely reflexive. The goal of the present article is to describe the major neuroanatomical structures and pathways that are involved in the control of two classes of voluntary eye movements: voluntary saccadic eye movements and smooth pursuit eye movements.

Cerebral Cortex and the Control of Eye Movements

Both saccadic and smooth pursuit eye movements are executed primarily with respect to visual stimuli in an individual's environment (although sometimes auditory or somatosensory stimuli can also serve to guide saccadic eye movements). It is reasonable, therefore, that the signals that initiate and control these eye movements originate in regions of the cerebral cortex that have dense connections with the visual system and, in particular, with visual association areas. Current evidence indicates that control signals from the cortex to the brain stem oculomotor centers originate not in a single cortical region but in a group of cortical regions that act together, in concert, to guide voluntary eye movements. These regions, termed cortical eye fields, are illustrated in **Figure 1(a)** and include the frontal eye field (FEF), the supplementary eye field (SEF), the prefrontal eye field (PFEF), the parietal eye field (PEF), the medial superior temporal region (MST), and the precuneus region (designated area 7m in monkeys). In each of these regions: (1) electrical stimulation evokes eye movements; (2) neural activity increases prior to the initiation of an eye movement; and (3) cortical damage or inactivation produces measurable impairment of eye movements. In addition, (4) each of these regions has direct projections to neural centers in the brain stem that are known to be concerned with eye movement control (**Figure 1(a)**); and (5) each of these regions is connected reciprocally with other cortical regions that are concerned with eye movements (**Figure 1(b)**). Finally, (6) functional magnetic resonance imaging (fMRI) studies in humans have confirmed that each of these cortical regions becomes more active during the performance of oculomotor tasks. On the basis of this evidence, these regions are referred to as eye fields. Anatomical, neurophysiological, and fMRI studies provide evidence that each eye field is composed of two distinct subregions, one devoted to the control of saccadic eye movements and the other to pursuit eye movements (**Figure 2**). Several other cortical areas, including regions in the cingulate cortex and in the temporal polysensory area, play some role in the control of voluntary eye movement, but have not yet been studied in sufficient depth to determine whether they are also eye fields in the same sense as the areas already described.

Most evidence available today supports the proposal that the several cortical eye fields comprise nodes in a distributed cortical network, with each node making distinct contributions to the whole but with some overlap in functional capability among the various nodes. For example, the several eye fields are densely and preferentially interconnected with one another and share many functional characteristics. Furthermore, although the disturbances of eye movements that are

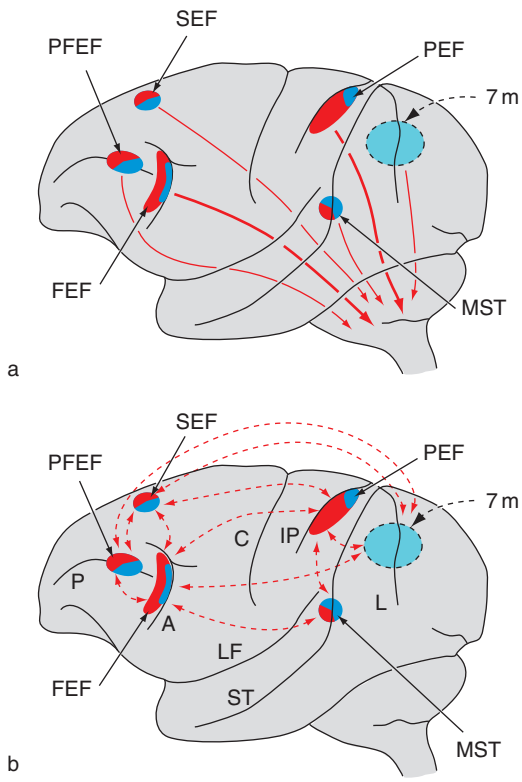

Figure 1 Cortical eye fields in the macaque monkey: (a) each eye field has direct projections to the brain stem oculomotor system; (b) each eye field is connected reciprocally with most other eye fields. Most eye fields (with exception of 7 m) contain discrete subregions that participate in the control of either saccadic (red) or smooth pursuit (blue) movements. FEF, frontal eye field; MST, medial superior temporal area; PEF, parietal eye field; PFEF, prefrontal eye field; SEF, suplementary eye field. 7 m (on the medial wall of the hemisphere) is equivalent to the precuneus in humans. Sulci: A, arcuate; C, central; IP, intraparietal; L, lunate; LF, lateral fissure; P, principal. Adapted from Lynch JC and Tian JR (2006) Cortico-cortical networks and cortico-subcortical loops for the higher control of eye movements. *Progress in Brain Research* 151: 461–501, with permission from Elsevier.

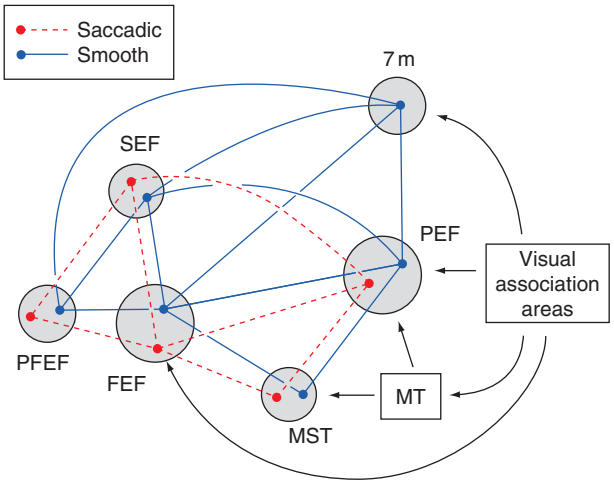

Figure 2 Parallel cortico-cortical networks for the control of saccadic eye movements and smooth pursuit eye movements. Each eye field is divided into a saccadic subregion (red) and a pursuit subregion (blue). Saccadic subregions are connected selectively with other saccadic subregions and pursuit subregions with other pursuit subregions. FEF, frontal eye field; MST, medial superior temporal area; MT, middle temporal area; PEF, parietal eye field; PFEF, prefrontal eye field; SEF, suplementary eye field. 7 m is equivalent to the precuneus in humans. Adapted from Tian JR and Lynch JC (1996) Corticocortical input to the smooth and saccadic eye movement subregions of the frontal eye field in *Cebus* monkeys. *Journal of Neurophysiology* 76: 2754–2771 used with permission.

produced by damage or inactivation in a single cortical eye field may be significant for a few hours or days, the disturbances recover very quickly. Eye movements are typically back to near normal within a few days after damage to a single cortical eye field. In contrast, damage to two cortical eye fields together produces a much more profound and long-lasting deficit of oculomotor control, suggesting that, when a single eye field is damaged, the other eye fields are able to somehow substitute for the missing functions.

Saccadic Eye Movement Pathways

The neural pathways that control saccadic eye movements are, with a few exceptions, distinct from the pathways that control pursuit eye movements. Two of the most important saccade-related pathways from the cerebral cortex to the brain stem oculomotor system are illustrated in a simplified diagram in Figure 3. The first pathway involves the superior colliculus (SC). Most of the eye fields already described, including the FEF, PEF, SEF, and PFEF, send substantial projections to the intermediate and deep layers of the SC. Neurons in these layers become active before saccadic eye movements, and electrical stimulation in these layers evokes eye movements. In contrast, neurons in the more superficial layers of the SC have visual sensory receptive fields. A major feature of the SC is that it contains a precise map of visual space, both in the superficial sensory layers and in the deeper motor layers.

The visual receptive fields of neurons in the superficial layers are arranged so that (1) neurons in the anterior part of the SC have receptive fields in the central portion of the contralateral visual hemifield, (2) neurons in the posterior SC have receptive fields in the periphery of the contralateral visual field, (3) neurons in the lateral part of the colliculus have receptive fields in the lower half of the contralateral hemifield, and (4) neurons in the medial part of the SC have visual fields in the upper half of the contralateral hemifield.

In the intermediate and deep layers, there is a motor map that is in register with the visual map in the superficial layers. For example, in the anterior part

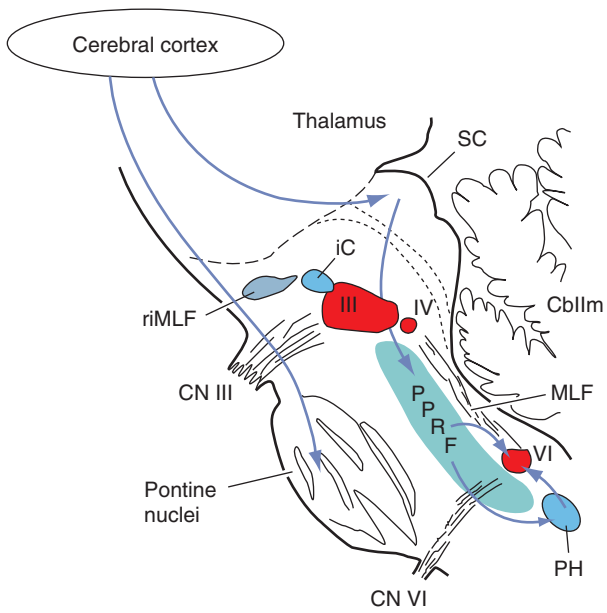

Figure 3 Brain stem structures and pathways involved in the control of saccadic eye movements. CbIlm, cerebellum; CN, caudate nucleus; iC, interstitial nucleus of Cajal; MLF, medial longitudinal fasciculus; PPRF, paramedian pontine reticular formation; riMLF, rostral interstitial nucleus of the medial longitudinal fasciculus; SC, superior colliculus. Nerves and motor nuclei: III, oculomotor nucleus; IV, trochlear nucleus; VI, abducens nucleus.

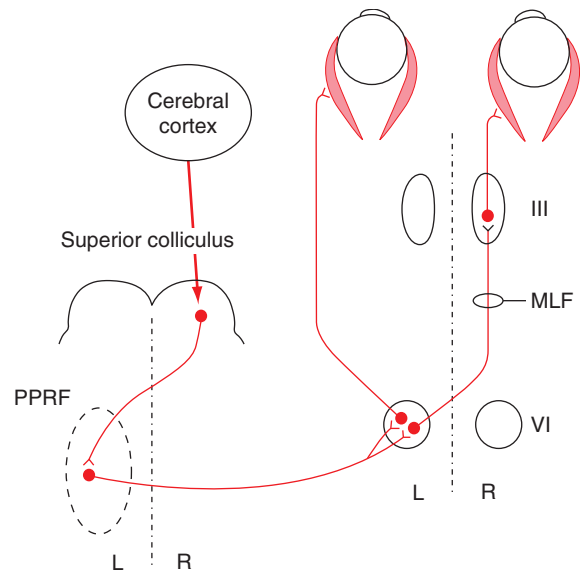

Figure 4 Simplified diagram of the brain stem neural circuitry that controls horizontal saccadic eye movements. III, oculomotor nucleus; L, left; MLF, medial longitudinal fasciculus; PPRF, paramedian pontine reticular formation; R, right; VI, abducens nucleus.

of the intermediate and deep layers of the SC, neurons are active before saccades that bring the center of gaze to the central portion of the contralateral hemifield (as in 1 above). Similarly, neurons in the intermediate and deep layers of the lateral portion of the posterior SC become active prior to saccades that bring the center of gaze into the lower half of the peripheral portion of the contralateral visual hemifield (as in 3 above).

Neurons in the intermediate layers of the SC project to the paramedian pontine reticular formation (PPRF), a region concerned with horizontal saccades, and to the rostral interstitial nucleus of the medial longitudinal fasciculus (riMLF), a structure that is concerned with vertical saccades. Neurons within the PPRF generate a step increase in activity that initiates a saccadic movement. Some of these neurons project to the nucleus prepositus hypoglossi (PH), where the step activity is integrated to produce the tonic increase in activity that is necessary to maintain the eyes in the new, postsaccade position. This tonic activity is then relayed to the abducens nucleus, where it is combined with the step activity from excitatory burst neurons in the PPRF to yield the burst-tonic activity necessary to produce an accurate saccade (Figure 3).

The axons of the excitatory burst neurons (PPRF) and tonic neurons (PH) terminate on both motor neurons and interneurons in the abducens nucleus, as diagramed in Figure 4. In this illustration, the motor neurons in the left abducens nucleus innervate the lateral rectus muscle on the left side. The axons of the interneurons in the left abducens nucleus cross the midline, ascend in the medial longitudinal fasciculus, and terminate on motor neurons in the right oculomotor nucleus. These neurons, in turn, innervate the medial rectus muscle on right side. Thus, the identical signal reaches the lateral rectus muscle on the left side and the medial rectus muscle on the right side, and a conjugate saccadic eye movement to the left is produced.

Vertical saccades are controlled by brain stem circuitry that is generally analogous to the neural circuitry previously described for horizontal saccades. This circuitry includes connections from the SC to the riMLF and the interstitial nucleus of Cajal (iC), nuclei which lie within the central mesencephalic reticular formation. The role of the riMLF in vertical saccades is analogous to that of the PPRF for horizontal saccades, and the function of the iC is analogous to that of the PH.

Paradoxically, in spite of the obvious central location of the SC in the brain stem oculomotor system, the destruction of the SC is quickly compensated by other systems and saccadic eye movements return to almost normal within a very short time. One candidate system that may mediate this rapid recovery is a projection from the FEF, PEF, SEF, and PFEF to the pontine nuclei and nucleus reticularis tegmenti pontis

(NRTP). The FEF, SEF, PFEF, and 7 m project predominantly to the NRTP and to the dorsomedial pontine nuclei (DMPN). Neurons in the NRTP and DMPN project to the oculomotor vermis (primarily lobule VII). From the vermis, signals leave the cerebellum via the oculomotor region of the fastigial nucleus (FOR) and thence to the premotor areas of the brain stem for the control of horizontal and vertical saccades, the PPRF and riMLF, respectively.

Some neurons in the NRTP and DMPN also project to the paraflocculus. This portion of the cerebellar hemisphere is particularly implicated in the control of pursuit eye movements, although there is evidence that it may also participate in the control of saccades. The pathway from the paraflocculus exits the cerebellum via the dentate nucleus and returns to the cerebral cortex by way of the ventral lateral nucleus of the thalamus.

The posterior eye field (PEF) also projects to the pontine nuclei, but, in contrast to the FEF, it projects primarily to the dorsolateral pontine nuclei (DLPN) and the lateral pontine nuclei (LPN). Neurons in the DLPN have been particularly implicated in pursuit eye movements, but activity related to saccadic eye movements has also been observed there. The DLPN projects to the oculomotor vermis and the paraflocculus, as do the DMPN and NRTP. From the vermis, signals leave the cerebellum via the fastigial nucleus, which projects to the PPRF and riMLF, the premotor areas for the control of horizontal and vertical saccades, respectively.

An additional neural circuit important for saccadic eye movements originates in the FEF, SEF, and PFEF and projects to the caudate nucleus. From there, the pathway goes to the substantia nigra, pars reticularis, and then to the intermediate layers of the SC. This pathway exerts an inhibitory influence on neurons in the SC. This inhibition reduces the probability that neural activity in the SC will cause an unintended saccade during the fixation of a stationary target. The inhibition is suspended just before an intended saccadic eye movement is initiated.

The major neural pathways involved in the control of saccadic eye movements are diagramed in Figure 5. This diagram shows only cortigofugal pathways. The cortical–subcortical–cortical feedback loops from the basal ganglia, SC, and cerebellum are addressed later in the article.

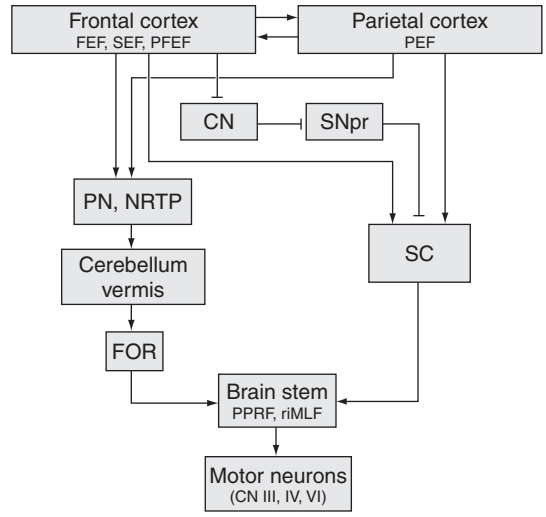

Figure 5 The primary pathways for the control of saccadic eye movements. The pathways from CN to SNpr and from SNpr to SC are inhibitory. CN, caudate nucleus; FEF, frontal eye field; FOR, fastigial oculomotor region (caudal fastigial nucleus); III, oculomotor nucleus; IV, trochlear nucleus; NRTP, nucleus reticularis tegmenti pontis; PEF, parietal eye field; PFEF, prefrontal eye field; PN, pontine nuclei; PPRF, paramedian pontine reticular formation; riMLF, rostral interstitial nucleus of the medial longitudinal fasciculus; SC, superior colliculus; SEF, suplementary eye field; SNpr, substantia nigra pars reticulata; VI, abducens nucleus. Adapted from Büttner U and Büttner-Ennever JA (2006) Present concepts in oculomotor organization. *Progress in Brain Research* 151: 1–42, with permission from Elsevier.

Pursuit Eye Movement Pathways

The neural pathways that control visual pursuit eye movements are not understood as completely as the saccadic eye movement pathways. The pursuit pathways involve the cerebellum and vestibular nuclei more directly than do the pathways for saccadic eye movements, and they avoid the SC to a large extent. The major neural pathways involved in the control of pursuit eye movements are diagramed in Figure 6. From the cortex of the posterior hemisphere, pursuit-related areas, including the middle temporal area (MT), MST, lateral intraparietal area (LIP), and 7 m, project to the DLPN. The DLPN contains neurons with activity directly related to pursuit eye movements, and lesion or inactivation of the DLPN produces disruption of pursuit movements. The pursuit subregion of the FEF (FEFsem) and the SEF project to regions of the NRTP and the DMPN that are known to participate in the control of pursuit eye movements. The DLPN projects to both the dorsal paraflocculus and to the oculomotor vermis of the cerebellum (lobules VI and VII). The projection to the dorsal paraflocculus (solid line in Figure 6) is stronger than the one to the vermis (dashed line).

The NRTP and DMPN also project to the oculomotor vermis and the dorsal paraflocculus, but in contrast to the DLPN projection, the connections of the NRTP and DMPN to the vermis (solid line) are stronger than those to the paraflocculus (dashed line). The paraflocculus then projects to the vestibular

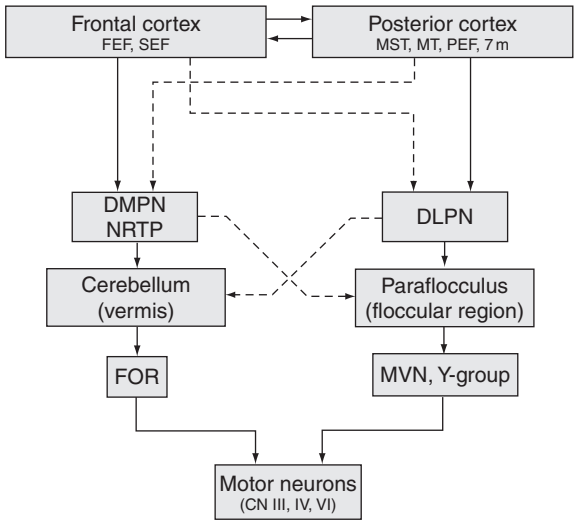

Figure 6 The primary pathways for the control of pursuit eye movements. Dashed lines indicate weaker projections than solid lines. CN, caudate nucleus; DLPN, dorsolateral pontine nuclei; DMPN, dorsomedial pontine nuclei; FEF, frontal eye field; FOR, fastigial oculomotor region (caudal fastigial nucleus); III, oculomotor nucleus; IV, trochlear nucleus; MST, medial superior temporal area; MT, middle temporal area; MVN, medial vestibular nucleus; NRTP, nucleus reticularis tegmenti pontis; PEF, parietal eye field; SEF, suplementary eye field; VI, abducens nucleus; Y-group, an accessory vestibular nucleus involved in vertical smooth pursuit. 7m (on the medial wall of the hemisphere) is equivalent to the precuneus in humans. Adapted from Büttner U and Büttner-Ennever JA (2006) Present concepts in oculomotor organization. *Progress in Brain Research* 151: 1–42, with permission from Elsevier.

nuclei (medial vestibular nucleus (MVN) for horizontal movements; Y-group cells for vertical movements). Finally, the vestibular nuclei send the pursuit-related signals directly to the motor nuclei of the oculomotor system. In contrast to the connections of the paraflocculus (a component of the cerebellar hemisphere, a phylogenetically newer structure than the vermis), the vermis projects primarily to the FOR, which, in turn, projects to the oculomotor nuclei.

The Thalamus and Eye Movement Control

In addition to the direct control exerted by the network of cortical eye fields on the oculomotor system in the brain stem, there are a number of nuclei in the thalamus that also contribute to the control of voluntary saccadic and pursuit eye movements. There are two ways in which the thalamus influences eye movements. The first is via connections between the anterior intralaminar nuclei, particularly the paracentral and central lateral nuclei, with other oculomotor structures. Some neurons in these nuclei (or immediately adjacent to them) have activity that begins prior to saccadic eye movements; some are active during the fixation of visual targets; and some neurons carry signals related to the location of saccadic targets, even in the absence of a visible target stimulus. Yet other neurons in these nuclei have activity that is related to the parameters of visual pursuit movements. The neural connections of the anterior intralaminar nuclei include the saccade subregion of FEF (FEFsac), the FEFsem, and the PEF, in addition to the basal ganglia, cerebellum, and some oculomotor structures in the brain stem. Currently available evidence therefore indicates that the paracentral and central lateral nuclei of the thalamus play a relatively direct role in the initiation and control of saccadic and pursuit eye movements.

The second mode in which thalamic nuclei participate in oculomotor control is that of relay (and processing) structures in the cortical–subcortical–cortical feedback loops. Numerous such feedback loops have been described in motor and higher function circuitry. The most important of the oculomotor loops involve feedback to the cortex from the SC, basal ganglia, and cerebellum. Two representative loops are diagramed in Figure 7. The first (A) is depicted by thick blue dashed lines. This circuit originates in the FEFsem and projects to the caudate nucleus (and, to a lesser extent, to the putamen); it goes from there to the globus pallidus. The globus pallidus, in turn, projects to the VLcr and VApc nuclei of the thalamus, which project back to the FEFsem. The function of this circuit is not yet understood, but probably is concerned with the high degree of feedback that is necessary to maintain an image of a smoothly moving visual target on the fovea.

A second feedback circuit (B in Figure 7) is more completely understood. This circuit originates in the FEFsac and projects to the oculomotor layers of the SC. Neurons in those layers project to the dorsomedial nucleus of the thalamus, where they synapse on neurons that project back to the FEFsac. This circuit serves to carry a corollary discharge signal from the SC back to the FEFsac that is proportional to an actual saccadic eye movement that has just been made. The FEFsac can then compare the magnitude and direction of the intended saccade with the saccade that actually occurred. If this circuit is inactivated at the dorsomedial nucleus, the accuracy of successive saccades is dramatically reduced because the animal is not aware of the eye position at the end of the first saccade when the second saccade is initiated. Similar feedback loops through the cerebellum have also been described, but are not illustrated here.

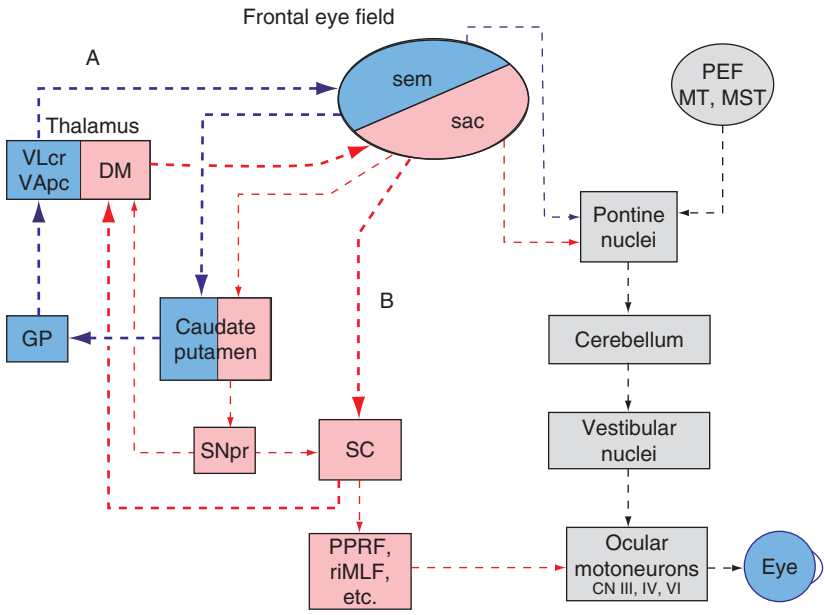

Figure 7 Diagram of representative cortical–subcortical–cortical feedback circuits. Heavy blue dashed lines indicate a feedback loop (A) related to the control of pursuit eye movements that runs from FEFsem to the striatum, globus pallidus, and thalamus and then returns to the FEFsem. Heavy red dashed lines indicate a feedback loop (B) that runs from the FEFsac to the SC, then back to the thalamus, and finally back to the FEFsac. The gray boxes on the right illustrate the primary smooth pursuit pathway and a secondary saccade pathway through the cerebellum. Feedback loops similar to A and B also exist in the cerebellar pathways, but they are not illustrated here. CN, caudate nucleus; FEF, frontal eye field; GP, globus pallidus; III, oculomotor nucleus; IV, trochlear nucleus; MST, medial superior temporal area; MT, middle temporal area; PEF, parietal eye field; PPRF, paramedian pontine reticular formation; riMLF, rostral interstitial nucleus of the medial longitudinal fasciculus; sac, saccades subregion of FEF; SC, superior colliculus; sem, pursuit subregion of FEF; SNpr, substantia nigra, pars reticulata; VI, abducens nucleus; Thalamic nuclei: DM, dorsomedial nucleus; VLcr, rostral portion of ventral lateral nucleus, pars caudalis; VApc, parvocellular portion of ventral anterior nucleus. Adapted from Cui DM, Yan YJ, and Lynch, JC (2003) Pursuit subregion of the frontal eye field projects to the caudate nucleus in monkeys. *Journal of Neurophysiology* 89: 2678–2684 used with permission.

Interactions between the Saccadic and Pursuit Pathways

Saccadic and pursuit eye movements are fundamentally different in their modes of interaction with the visual environment. Saccades are essentially preprogrammed movements that can be made at will, even in the absence of visual targets. In contrast, a smooth pursuit movement can be made only in the presence of a moving target and, by its nature, must be continuously modified in speed and direction to match the velocity of its target. These differences in the basic characteristics of saccadic and pursuit eye movements dictate that the neural circuitry controlling the two will be largely distinct from one another. Nevertheless, the saccadic system and the pursuit system must frequently interact with one another, for example, when making a saccadic eye movement to a moving target in order to initiate a pursuit movement. There is an increasing body of anatomical and physiological evidence that the two relatively independent basic pathways that have been described here come together at a number of sites in the brain stem and are, in fact, not as completely independent of one another as was formerly believed. Among the structures in the oculomotor system in which there seems to be some degree of interaction between the saccadic and pursuit systems are the DLPN and DMPN; the cerebellar vermis; the SC; the riMLF and PPRF; and the anterior intralaminar nuclei of the thalamus.

Summary and Conclusions

In summary, voluntary eye movements are used to shift the center of gaze from one point of interest to another in an individual's surroundings (saccadic eye movements) or to maintain the image of a moving object on the fovea so it can be tracked accurately with the eyes (pursuit eye movements). These two types of eye movements are controlled by largely separate neural circuits. Both classes of eye movements are controlled at the highest level by cortico-cortical networks consisting of at least six cortical eye fields acting in concert. Neural pathways concerned with saccadic eye movements originate in distinct

subregions of most of these eye fields and project to the SC, the basal ganglia, and (via brain stem relay centers) the cerebellum. Signals from the SC, in turn, engage oculomotor centers in the pontine and mesencephalic reticular formation.

Neural pathways concerned with pursuit eye movements also originate in the cortical eye fields, in subregions that are distinct from the saccadic subregions. In contrast to the saccadic pathways, the pursuit pathways project predominantly to the cerebellum (via brain stem relay centers). From the cerebellum, pursuit signals go to the vestibular nuclei and then to the ocular motor neurons. Both the pursuit and saccadic subregions of the cortical eye fields project to the basal ganglia and the cerebellum, and it is likely that some interaction between the two eye movement systems occurs in these structures. Both the basal ganglia and the cerebellum participate in cortico-subcortico-thalamo-cortical feedback loops that help to fine-tune the eye movements.

See also: Brainstem Control of Eye Movements; Cortical Control of Eye Movements; Frontal Eye Fields; Pursuit Eye Movements; Saccade–Pursuit Interactions; Saccadic Eye Movements; Superior Colliculus; Supplementary Eye Fields; Target Selection for Pursuit and Saccades; Thalamus and Oculomotor Control.

Further Reading

Alexander GE, DeLong MR, and Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience* 9: 357–381.

Büttner U and Büttner-Ennever J (2006) Present concepts of oculomotor organization. *Progress in Brain Research* 151: 1–42.

Büttner-Ennever J (ed.) (2006) *Neuroanatomy of the Oculomotor System*. New York: Elsevier.

Cui DM, Yan YJ, and Lynch JC (2003) Pursuit subregion of the frontal eye field projects to the caudate nucleus in monkeys. *Journal of Neurophysiology* 89: 2678–2684.

Dicke PW, Barash S, Ilg UJ, and Thier P (2004) Single-neuron evidence for a contribution of the dorsal pontine nuclei to both types of target-directed eye movements, saccades and smooth-pursuit. *European Journal of Neuroscience* 19: 609–624.

Haines DE (2006) *Fundamental Neuroscience for Basic and Clinical Applications*. Philadelphia: Elsevier.

Hikosaka O, Takikawa Y, and Kawagoe R (2000) Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological Reviews* 80: 953–978.

Horn AK (2006) The reticular formation. *Progress in Brain Research* 151: 127–155.

Keller EL and Missal M (2003) Shared brainstem pathways for saccades and smooth-pursuit eye movements. *Annals of the New York Academy of Science* 1004: 29–39.

Kelly RM and Strick PL (2003) Cerebellar loops with motor cortex and prefrontal cortex of a nonhuman primate. *Journal of Neuroscience* 23: 8432–8444.

Krauzlis R (2004) Recasting the smooth pursuit eye movement system. *Journal of Neurophysiology* 91: 591–603.

Leigh JR and Zee DS (2006) *The Neurology of Eye Movements*. New York: Oxford.

Lynch JC and Tian JR (2006) Cortico-cortical networks and cortico-subcortical loops for the higher control of eye movements. *Progress in Brain Research* 151: 461–501.

May PJ (2006) The mammalian superior colliculus: Laminar structure and connections. *Progress in Brain Research* 151: 321–378.

Müri RM (2006) MRI and fMRI analysis of oculomotor function. *Progress in Brain Research* 151: 503–526.

Petit L and Haxby JV (1999) Functional anatomy of pursuit eye movements in humans as revealed by fMRI. *Journal of Neurophysiology* 82: 463–471.

Sommer MA and Wurtz RH (2004) What the brain stem tells the frontal cortex. II: Role of the SC-MD-FEF pathway in corollary discharge. *Journal of Neurophysiology* 91: 1403–1423.

Sommer MA and Wurtz RH (2006) Influence of the thalamus on spatial visual processing in frontal cortex. *Nature* 444: 374–377.

Tanaka M (2005) Involvement of the central thalamus in the control of smooth pursuit eye movements. *Journal of Neuroscience* 25: 5866–5876.

Thier P and Möck M (2006) The oculomotor role of the pontine nuclei and the nucleus reticularis tegmenti pontis. *Progress in Brain Research* 151: 293–320.

Tian JR and Lynch JC (1996) Corticocortical input to the smooth and saccadic eye movement subregions of the frontal eye field in *Cebus* monkeys. *Journal of Neurophysiology* 76: 2754–2771.

Tian JR and Lynch JC (1997) Subcortical input to the smooth and saccadic eye movement subregions of the frontal eye field in Cebus monkey. *Journal of Neuroscience* 17: 9233–9247.

Voogd J and Barmack NH (2006) Oculomotor cerebellum. *Progress in Brain Research* 151: 231–268.

Wyder M, Massoglia D, and Stanford T (2004) Contextual modulation of central thalamic delay-period activity: Representation of visual and saccadic goals. *Journal of Neurophysiology* 91: 2628–2648.

Yan YJ, Cui DM, and Lynch JC (2001) Overlap of saccadic and pursuit eye movement systems in the brain stem reticular formation. *Journal of Neurophysiology* 86: 3056–3060.

Oculomotor System: Models

L M Optican, Laboratory of Sensorimotor Research, NEI, NIH, DHHS, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

The oculomotor system comprises the contents of the orbits (the globe; six extraocular eye muscles; and many fatty, fibrous, and elastic tissues) and parts of the brain. The globe undergoes a very limited amount of translation during movements, about 2 mm along the anterior–posterior axis and 0.5 mm in the frontal plane. Thus, the globe can be considered to be a spherical joint with its center fixed in the orbit. There are three main properties of the orbit that the brain must deal with. The first is the orbital mechanics created by the three-dimensional (3-D) geometry of the eye muscles and the surrounding orbital tissues. These determine the pulling action of each muscle, that is, how a contraction of a muscle is converted to a torque applied to the globe. The second is the dynamics of the orbital tissues. If the time course of the innervation sent to the muscles is not appropriate for the orbital dynamics, the eye will drift when the movement is over. The third is that 3-D rotations are not commutative. If the axis of the applied torque is not correct for the orientation of the eye, plant dynamics will cause unintended drifts or blips around other axes.

Oculomotor Plant

The oculomotor plant is very complicated, nonlinear, and still only poorly understood. Thus, many models of the plant are in use today, with linear models being used when the details can be safely lumped together to give average dynamics.

Lumped Plant Models

Early models of the plant were simplified by considering only rotations of the eye around a single (e.g., vertical) axis. Control theory models of the plant date from the suggestion by Westheimer in 1954 that rapid eye movements (called saccades) were the step response of the plant. This idea suggests a very simple saccadic system because the visual input for a saccade (a jump of the target of interest) evokes a step change in the site of innervation on the retina and, hence, in other brain areas containing maps of retinal activity. If these inputs were added together with appropriate weights (stronger for more distant targets), they would create a steplike change in the firing rate of the neurons, which could be sent directly to the plant. Westheimer proposed that the plant was a critically damped linear system, which would foveate the target as quickly as possible without overshooting it.

In 1964, Robinson measured the mechanical properties of the plant and found that its dynamics were dominated by the viscosity of the muscles and, thus, was not critically damped but instead overdamped. The dominant time constant of the plant was ~285 ms. Thus, if the innervation to the plant were changed in a steplike fashion, the eyes would drift exponentially toward the target and would only arrive after about three time constants. Because we make 3–4 saccades s^{-1}, the innervation must be able to compensate for the viscosity in the plant or our eyes would never stop drifting.

Robinson derived a fourth-order linear model of the plant's mechanics and showed that the innervation needed to drive the plant to make a saccade consisted of at least three parts: a pulse, a slide, and a step. The step determines the final orientation of the eye, the pulse generates extra force to overcome the viscous drag of the orbital tissues, and the slide compensates for some of the viscosity and the relaxation in slow orbital tissues after the rapid part of the movement is over.

The ratio of the output (eye position) to the input (innervation) in Robinson's model of the plant was represented as a transfer function (a linear input–output relationship expressed as the ratio of two complex polynomials). Robinson's plant consisted of one real zero (the value at which the numerator polynomial becomes zero), two real poles (values at which the denominator polynomial of the transfer function becomes zero), and a complex-pole pair. The complex-pole pair contributes an oscillation to the output of a movement that is very small because of the large viscosity. Thus, the plant can be usefully approximated with just two real poles and one zero. Figure 1(a) shows the corresponding mechanical model with three viscoelastic (or Voigt) elements, with the transfer function:

$$E/R = c \times ([sT_z + 1]/[sT_1 + 1][sT_2 + 1])$$

where s is the complex Laplace transform variable, T_1 and T_2 are the viscoelastic time constants, or plant poles, and T_z is the plant's zero (Figure 1(b); Table 1). The scale factor, c, is set by the net compliance of

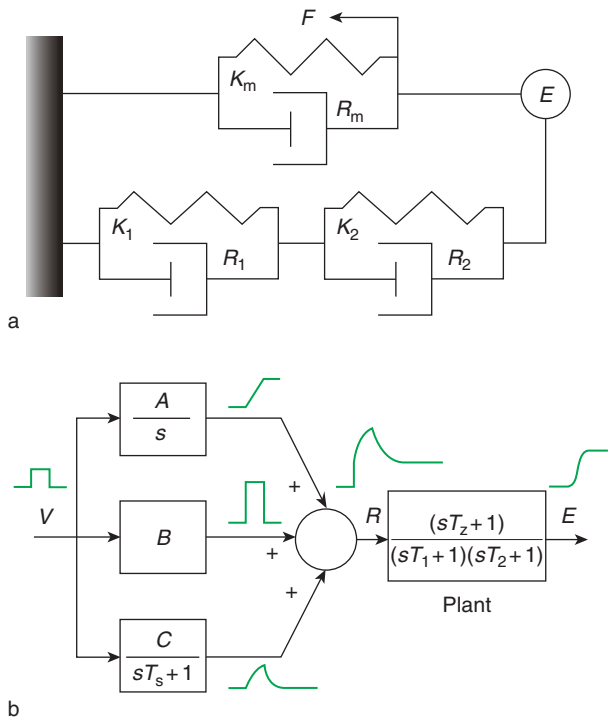

Figure 1 Second-order linear approximation to the oculomotor plant and its neural compensation: (a) mechanical model of the active and passive tissues in the plant, with the viscosity and elasticity represented by Voigt elements (parallel spring, *K*, and dashpot, *R*); (b) simplified transfer function of the model in (a) (Plant) and the circuit needed to compensate its dynamics. In (a), the force of the muscle's contractile elements is represented by *F*; the eyeball is represented by E, and the time constant of each Voigt element is *R*/*K*. In (b), the innervation (*R*) for an eye position (*E*) requires a pulse, slide, and step. For one-dimensional rotations, the slide and step can be obtained by integrating the pulse. Green traces show velocity command (*V*), step, pulse and slide of innervation, combined pulse-slide-step (*R*), and eye position (*E*). The plant has two poles (T_1 and T_2) and one zero (T_z). *A*, *B*, and *C* are weights and T_s is a time constant needed to build the slide of innervation (see **Table 1**). The transfer functions are shown in Laplace transform notation, so $1/s$ represents a perfect integrator, and $1/(sT+1)$ represents a leaky integrator (or low-pass filter).

Table 1 Lumped two-pole, one-zero plant model and its compensator

Model transfer function		
T_1	Dominant pole's time constant	0.136 s
T_2	Second pole's time constant	0.726 s
T_z	Plant's zero time constant	0.615 s
τ	Excitation–contraction coupling delay	0.008 s
Compensator (final common path)		
A	Step gain	1
B	Pulse gain	0.1605
C	Slide gain	0.0865
T_s	Slide time constant	0.615 s

the springs in the model $(K_1+K_2)/(K_1K_2)$ and the innervation–contraction coupling ratio, but it is conventionally set to 1 so that the gain of the plant is 1 for very low frequency movements. In terms of the viscosity and elasticity, the time constants are $T_i = R_i/K_i$, and the zero is given by $T_z = (R_1+R_2)/(K_1+K_2)$.

Numerically, the stiffnesses of the two passive elements are approximately 0.79 and 1.22 g/°, with viscosities of 0.0158 and 1.22 gs/°. (Note: Ophthalmologists measure muscle force during surgery with a spring scale calibrated in grams; 1g force is 9.8×10^{-3} N.) Thus, the time constants for the poles in the passive tissues are 0.020 and 1.000 s, and the zero is at 0.615 s. The viscoelasticity and active-force generation of the muscle is represented by another Voigt element (*F*, K_m, R_m). Its stiffness is about 1.58 g/° and the viscosity is about 0.316 gs/°, which gives a lumped time constant for both muscles of approximately 0.200 s. Combining the passive and active tissues, we can approximate the total plant model with two poles ($T_1 = 0.136$ s, $T_2 = 0.726$ s) and one zero ($T_z = 0.615$ s). The zero approximately cancels the second pole, making the dominant time constant about 0.150 s. Thus, in models in which the details of rapid eye movements are not important, the plant model can be reduced to just a single pole: $1/(0.150s+1)$. When time delays are taken into account, a delay of about 8 ms is added to the plant by putting the factor $e^{-0.008s}$ in the numerator of the transfer function.

Note that these values are all approximations of a more complex, nonlinear system. As such, their values can change in different species or under different experimental conditions. For example, the dominant time constant in monkeys ranges from 0.090 to 0.200 s and in humans from approximately 0.100 to 0.200 s.

One Dimension

In one dimension, the compensatory innervation (*R*) needed to drive this plant can be easily determined if the pulse command (*V*) is known (**Figure 1(b)**). Integrating the pulse with a gain (*A*) gives the step, and low-pass filtering the pulse (with gain *C* and time constant T_s) gives the slide. With perfect plant compensation, the ratio of eye movement to innervation would simply be $1/s$ (a pure integrator). Hence, for normal saccades, the gain of the step should be 1, which fixes $A=1$. The gain of the pulse must then be $B = T_1T_2/T_s$, which is about 0.1605. To compensate for the zero in the plant, $T_s = T_z$, or 0.615. The gain of the slide must be $C = T_1 + T_2 - T_s - (T_1T_2/T_s)$, which is 0.0865. **Figure 2** shows an implementation of the plant and its compensator and the output from a simulation run (cf. **Table 1**).

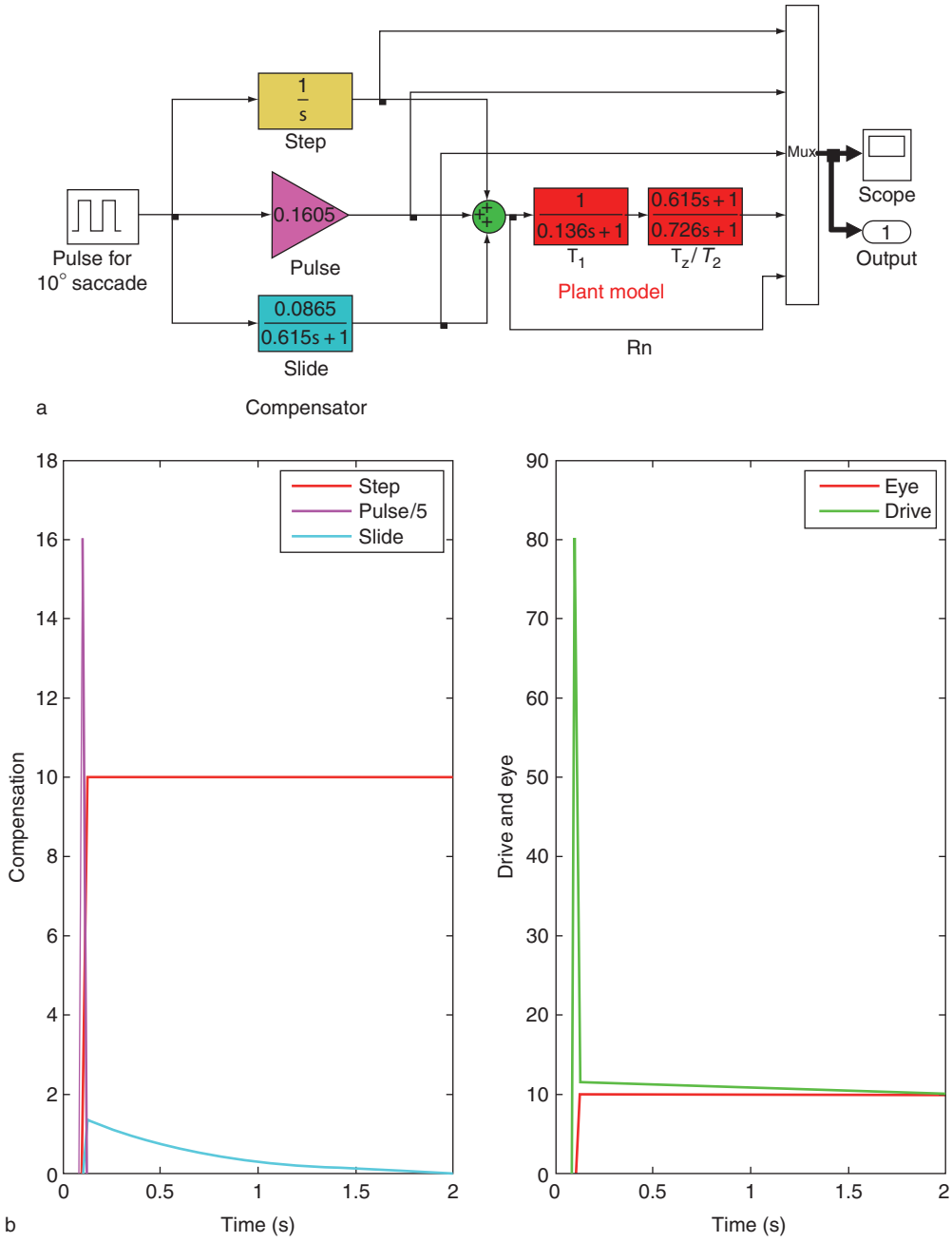

Figure 2 Simulation of the simple plant model and its neural compensator from **Figure 1**: (a) implementation of the model in MATLAB/SIMULINK visual programming environment; (b) output of a simulation run. In (a) The generator on the left puts out a pulse 500°/s high for 0.020 s (which should make an eye movement of 10°). The compensator consists of three parallel operators: an integrator (yellow), an amplifier (purple), and a low-pass filter (blue). The plant model (red) is made by combining a low-pass filter block (for T_1) with a lead–lag block (for T_z and T_2). Other blocks (white background) allow Simulink to plot the simulation output in real time (scope) and send it to the MATLAB workspace (output). In (b), the left panel shows the step, pulse (divided by 5 to scale it to fit with the others), and slide; the right panel shows the pulse–slide–step drive (green) and the output of the plant (eye). Note that the pulse part of the drive signal is about eight times bigger than the step part. Most of the energy in this pulse is eaten up by the muscles to overcome their own viscosity, and only a small part is passed to the tendon. Mux, multiplexer.

A major simplification for determining the innervation occurs in one-dimensional (1-D) models because in 1-D orientation is the integral of angular velocity, just as position is the integral of velocity. Thus, in 1-D the torque vector determined by the pulse of innervation and the torque determined by the slide and step are collinear (necessarily, because there is only one axis). Thus, if the pulse of innervation were already computed, obtaining the step of innervation would require simply the mathematical integration of

that pulse. Similarly, the slide could be obtained as the response of a leaky integrator to the pulse.

Three Dimensions

Three-dimensional rotations are more complicated than 1-D rotations because 3-D orientation is not the integral of angular velocity (i.e., rotations are not commutative). However, three dimensional muscle torques are commutative vectors, and the final eye orientation is thus determined by the balance of torques between active and passive tissues. However, during a 3-D movement, the eye deviates from a straight trajectory (e.g., with torsional blips) because of the lack of commutativity, unless the torque is modified by the current eye orientation. Torque could be modified by changing the innervation or changing the mechanics of the orbit.

Muscle Pulleys

Originally it was assumed that the muscle always went straight from its origin to its insertion, even as the eye looked into eccentric orientations. Thus, the pulling direction of the muscles would not change much as the eye moved (Figure 3(a)), and the brain would have to compute a new torque axis for each ocular orientation. However, in 1989 Miller used magnetic resonance imaging (MRI) of human orbits to show that the belly of the muscle did not move as the eye turned, thus creating an inflection point in the eye muscle path. Further research by Demer and colleagues found that this point corresponded to fibromuscular pulleys, which encircle the belly of each muscle (except the superior oblique, which goes through the bony trochlear pulley). Quaia and Optican showed in 1998 that, if the pulleys are in the right place (approximately halfway between the equator and the posterior pole of the eye), the amount by which the axis of action tips for the pulse will be just enough to compensate for the noncommutativity of 3-D rotations (Figure 3(b)), without affecting the final orientation set by the step. This makes the 3-D angular velocity vector approximately equal to the derivative of the 3-D orientation vector. That means the brain could calculate innervation in 3-D as if it were in 1-D: the pulse can simply be the derivative of the desired change in eye orientation, and the slide and step can be obtained from the pulse through (leaky) integration.

As the eye turns, the position of the pulleys must also change, so as to keep the geometrical relationship between the pulley and muscle insertion constant. If the pulley did not move, as the eye rotated the muscle insertion would run into the pulley. Fortunately, all that this requires is that the pulleys move back as much as the globe rotates in its direction. This may be facilitated by the fact that the orbital layers of the extraocular muscles insert on the pulley itself. Thus, as the muscle contracts, it moves both the eye and the pulley.

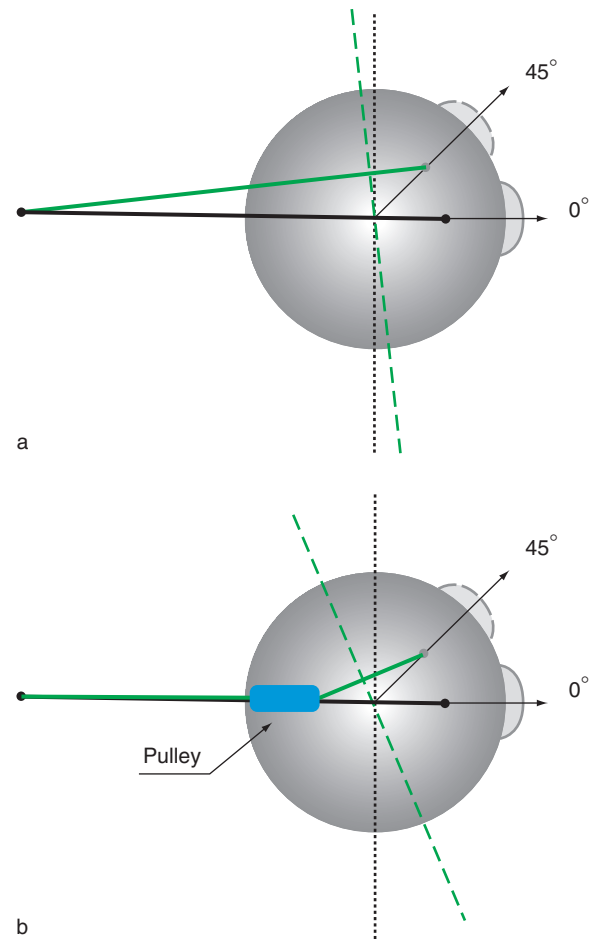

Figure 3 Effect of eye orientation on the axis of action of extraocular muscle: (a) without pulleys; (b) with pulleys. In (a), if there were no impediments the muscle would slip as the eye turned, so that the muscle always took the shortest path to the insertion. In this case, even for a 45° elevation, the axis of action (green dotted line) tips very little (~6°). In (b), a specialized structure in the orbital tissues, called a pulley (blue), holds the muscle in place as the eye turns. In this case, for a 45° elevation, the pulley redirects the axis of action of the muscle (green dotted line) by tilting it in the same direction as the globe has moved, but by only half as much (22.5°). Adapted from Quaia C and Optican LM (2003) Three-dimensional rotations of the eye. In: Kaufman PL and Alm A (eds.) *Adler's Physiology of the Eye: Clinical Application*, pp. 818–829. New York: Mosby.

Eye Movement Models

Eye movements can be divided up into subtypes by their speed, latency, and whether they are voluntary

or reflexive. These include vestibular, optokinetic, smooth pursuit, saccadic, vergence, and short-latency ocular following movements. Many parts of the brain respond to several eye movement types, so the brain does not necessarily process these types independently. Saccades are the most demanding movements to model from the point of view of dynamics, and they are used as an oculomotor model prototype next.

Final Common Integrator

The task of computing oculomotor innervations comes down to generating an appropriate pulse of innervation and then integrating that to obtain a slide and a step for plant compensation. Skavenski and Robinson showed that the phase lag introduced by the brain from the head-velocity signal on the vestibular afferents to the oculomotor neurons was ~90°. This indicated that the brain must be performing an integration of desired eye velocity to get the eye position signal needed by the motor neurons. This observation gave rise to the concept of a final common path, which includes the motor neurons and the neural integrator. Any eye movement command (e.g., vestibular, pursuit, or saccade) need only generate an eye velocity command and feed it into the final common path. All the other mechanisms from there to the orbit are shared by all systems. This sharing makes the neural integrator a key element of all oculomotor movements. Lesion and single-unit recording studies have localized the neural integrator to the vestibular nuclei, midbrain interstitial nucleus of Cajal (NIC; for vertical and torsional components), the brain stem nucleus prepositus hypoglossi (NPH; for horizontal components), and corresponding parts of the cerebellum.

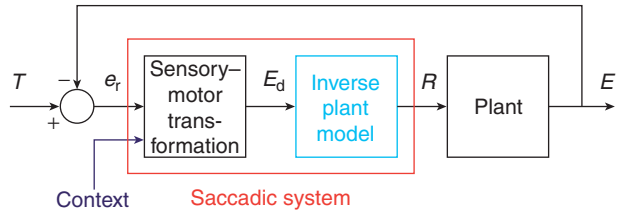

Figure 4 The saccadic system must consist of two functional subdivisions: a context-dependent sensory-to-motor transformation that changes the input signals into a desired eye orientation signal, E_d, and an inverse of the plant model that converts desired eye orientation into the innervation needed to drive the eye until $E = E_d$.

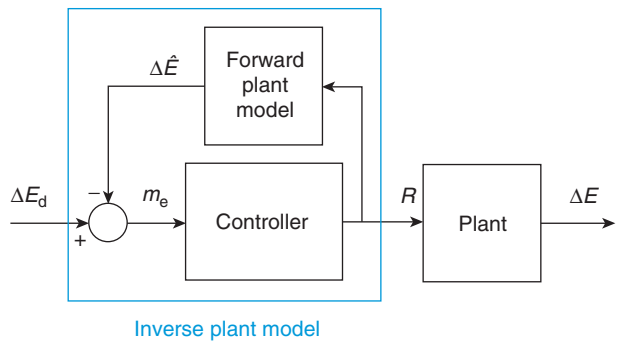

Figure 5 Local feedback loop, proposed by Robinson, that corresponds to the inverse plant model in Figure 4. The feedback loop contains a forward model of the plant that takes innervation as an input and puts out eye orientation. (The forward plant model is represented by a resettable leaky integrator.) The forward plant model's output, $\Delta\hat{E}$, is called the efference copy of the change in eye orientation, ΔE. The comparator subtracts efference copy from desired change in orientation to calculate dynamic (i.e., changing during the movement) motor error, m_e.

Saccadic System Functions

The saccadic system (Figure 4) takes target location in retinotopic coordinates and additional information about the movement's context (such as the orientation of the eyes and motion of the target) and produces the innervation required to get the eye (E) on target (T). In essence, the saccadic system must consist of two functional subdivisions, a sensory-to-motor transformation that changes the input signals into a desired eye orientation signal, E_d, and an inverse model of the plant that can convert an eye orientation into the corresponding innervation needed to drive the eye until $E = E_d$.

Robinson realized that an inverse plant model can be made simply by placing a forward model of the plant in a feedback loop around a high-gain amplifier (Figure 5). This loop defines a new signal, the instantaneous, or dynamic, motor error (m_e). It is the difference between the desired eye orientation change and the current estimate of the eye orientation change ($\Delta E_d - \Delta\hat{E}$) obtained by feeding a copy of the innervation that goes to the plant into the forward model (the forward model can be well approximated by an integrator that resets to zero before every saccade). The summing junction that subtracts $\Delta\hat{E}$ from ΔE_d is called a comparator. If the gain of the controller is high enough, the negative feedback loop drives the output of the comparator (motor error) to zero and the eye's orientation becomes equal to the desired orientation. The sensory-to-motor transformation is much more complicated because it must take into account the current orientation of the eyes, whether the target is moving or stationary, what behavior is desired, and so on. How this transformation is carried out is not well understood, but by moving from a lumped to a distributed model various mechanisms can be proposed.

Brain Areas and the Lumped Model

The brain can calculate the needed innervation if it knows the change in eye orientation that is desired. There are two obvious areas in the brain that could provide this signal: the superior colliculus (SC) and the cerebellum (CB).

Superior colliculus It is not easy to incorporate the SC into a lumped model of the saccadic system as the source of the desired eye movement for three reasons. First, the SC contains a spatially coded map. However, the feedback signal that comes back to the comparator (Figure 5) is temporally coded. Thus, the SC output must be converted into a temporal code before it can feed into the comparator. The second problem is that the locus of activity on the SC encodes the desired target in retinotopic coordinates, not the movement necessary to acquire it. The third problem is that the SC cannot be the only input to the saccadic system because, after SC lesions, saccades are accurate, although they start with a longer latency and have a lower peak speed than normal. This suggests that another area must cooperate with the SC to make a movement.

Oculomotor cerebellum The CB is the area most concerned with making accurate eye movements. The oculomotor vermis (OV) and the caudal part of the fastigial nucleus (cFN) are involved in controlling saccades.

Just as SC-receptive fields form a map of visual space, the cells in the CB form a map of movement space. This map is not as well defined as the SC map, and many cells participate in almost every movement. To aid in interpretation more than to be descriptive, the midline CB can be represented as two maps (Figure 6). The vermis is the only cortical structure in the brain that is continuous across the midline. When unfolded, it is actually much longer than it is wide, but this conceptual map simply represents small saccades toward the center of the vermis and larger saccades toward the edges. The only output from the cerebellar cortex is the inhibitory axon of the Purkinje cell. It projects directly to the ipsilateral cFN.

The cFN in primates is a very small structure. Although it is near the midline, the left and right nuclei are not interconnected. There are inhibitory cells that project out of the cFN, but they go only to the inferior olive and are not discussed here. The outputs from the excitatory cells in the cFN cross over the midline and go to the brain stem. In particular, they project to the contralateral medium lead burst neurons (inhibitory burst neurons (IBNs) and excitatory burst neurons (EBNs)) in the brain stem.

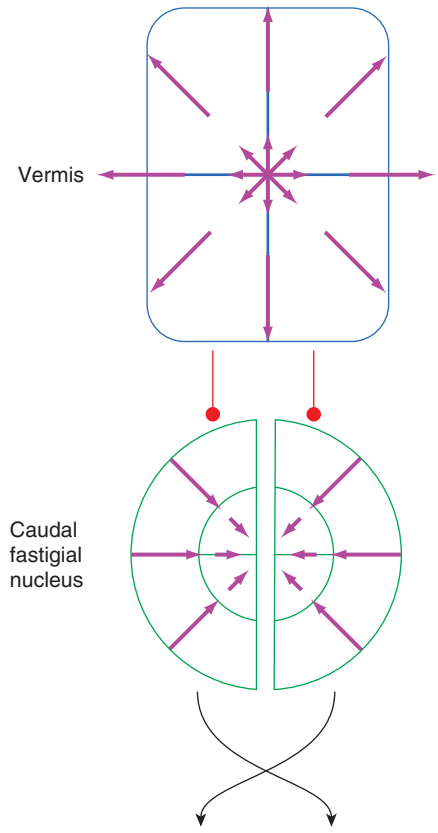

Figure 6 Conceptual maps in the cerebellum. For modeling, the cerebellum can be regarded as having two areas: the cortex and the deep cerebellar nuclei. The areas related to horizontal movements are the midline structures (posterior vermis and caudal fastigial nucleus, cFN). The vermis has a coarse map of saccade size and direction (purple arrows). The cFN is quite small, and the cells receive convergent input from many Purkinje cells. Thus, it is unlikely that they contain any kind of topographical map. Nonetheless, it is convenient to think of the cFN as repeating the map in the vermis but with the sign reversed. The output of the fastigial nucleus crosses over and goes to the brain stem on the other side. Thus, activity in the left nucleus corresponds to rightward saccades. The output of the vermis goes to the ipsilateral fastigial nucleus, but the output comes only from the Purkinje cells, which are inhibitory. Thus, a rightward saccade requires a pause in the left vermis. Correspondingly, an increase in activity in the right vermis inhibits the right fastigial nucleus, unbalancing the fastigial output to the brain stem. This results in a rightward movement.

The direction associated with activation of Purkinje cells on the left is thus a movement to the left because it inhibits the left cFN which projects to rightward premotor burst neurons (MLBNs), thus tipping the balance toward a leftward movement. In contrast, activation of the left cFN causes a rightward eye movement. It thus makes sense to extend our conceptual map of the CB to include vertical movements by placing up and down cells at the top and bottom of the vermis and down and up cells at the top and bottom of the cFN, respectively (Figure 6).

As more data on the role of cells in the midline CB, and related areas such as the interpositus deep cerebellar nucleus, become available, these conceptual maps will need to be updated to maintain a closer isomorphism with the brain. However, the concept of a map, at least in the cerebellar cortex, will probably still be useful.

Neuromimetic Model of the Saccadic System

A lumped model simplifies the description of a system by using simple connections of physically meaningful signals (e.g., target position and eye velocity) among complex elements (e.g., a resettable integrator). Unfortunately, the signals flowing through the brain often do not correspond to physical signals, and individual neurons are not very complex information processors. Instead, the complexity of the brain comes from the massive interconnection of many simple neurons. Thus, to build a model of the saccadic system that incorporates some of this complexity while maintaining its physiological isomorphisms with the brain requires a distributed model. If the distributed model contains elements with anatomical correspondences with the brain and if the elements carry signals observed in neurophysiological recordings, the model can be called neuromimetic.

Now that the conceptual visual (SC) and motor (CB) maps have been presented, they can be combined into a distributed neuromimetic model of the saccadic system (Figure 7). The cerebral cortex is the source of visual information, cognitive choices about what the eyes should look at and the context of the movement. The drive for the eye comes from two parallel pathways, the SC and the CB. These project down to the brain stem areas for controlling the eye muscles. The cerebrum and the SC cooperate to select the target. No specific signal to start the movement is needed. Once the SC determines the target, it stops sending a veto signal to the omnipause neurons (OPNs), which inhibit the saccade generator. The SC also sends a fixed-direction drive signal (in retinotopic coordinates) to the brain stem. The information about the selected target and the movement's context (e.g., target is moving) are sent to the CB. The CB sends a variable-direction drive (or pilot) signal to the brain stem. The sum of the drive from the CB and the SC form the velocity command that moves the eye. This command is also fed back to the CB. The velocity signal can be used to update the current state of the CB. If necessary, the CB can then change the direction of the pilot signal to bring the eyes to the target. When the CB decides that the movement is over, it sends a choke signal to the brain stem (presumably to contralateral IBNs) that cuts off the input to the motor neurons, stopping the saccade.

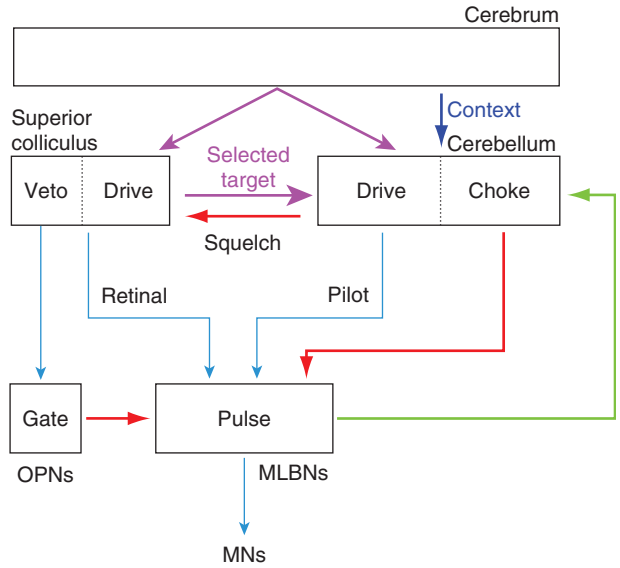

Figure 7 Distributed model of the saccadic system. The cerebral cortex and the superior colliculus (SC) determine what target to look at. The SC stops its input signal to the omnipause neurons (OPNs), shutting them off and disinhibiting the saccade generator. The SC also sends a drive command to the brain stem premotor burst neurons (MLBNs) to get the eye moving. The cerebellum sends a variable-direction drive (pilot) signal to the brain stem. The sum of the drive from the cerebellum and the SC form the velocity command that moves the eye. If the cerebellum decides that the movement is over, it sends a choke signal to the brain stem (via the contralateral inhibitory burst neurons) that cuts off the input to the motor neurons (MNs), stopping the saccade. The output of the SC is directionally fixed, and so it is called a directional, or retinal, drive. The cerebellum receives a feedback signal (efference copy of eye position and/or velocity) during the movement and, thus, can adjust its output to steer the saccade, so its output is called a pilot drive. When we compare this feedback loop with the one in **Figure 5**, it clear that the key element of Robinson's model, the local feedback loop, goes from the MLBN to the cerebellum (green arrow). Thus, the role of the forward plant model in **Figure 5** is played here by the cerebellum.

The neurons in the left SC fire for rightward targets, and the neurons in the right SC fire for leftward targets. However, cFN neurons fire for both ipsi- and contralateral saccades. How can the same neuron fire for both directions of movement and accomplish anything? One possible interpretation is suggested by the observation that cFN neurons fire with a shorter latency for contra- than for ipsilateral saccades. The contralateral burst is time-locked to the start of the movement, but the ipsilateral burst occurs with a high degree of variability, usually near the end of the movement. This timing difference of cFN activity can be exploited. In the brain stem, there is some evidence that IBNs are stronger than EBNs. This breaks the anatomical symmetry and allows the early contralateral output to be a drive signal, whereas the later ipsilateral output acts as a choke.

Thus, for a rightward saccade, the left OV must pause and the left cFN must fire, which drives the right IBNs and EBNs. The right EBNs drive the right lateral rectus and left medial rectus. The right IBNs cross over and inhibit the EBNs, IBNs, and motor neurons on the left side. Near the end of the saccade, the right vermis pauses and the right cFN fires. The output from the right cFN crosses over and drives the left IBNs and EBNs. At this time, the left IBNs and EBNs are being inhibited by the right IBNs. However, if we assume that the excitation from the right cFN is more powerful than the inhibition from the right IBNs on the left IBNs, it will begin firing. The left IBNs cross over to the right side and inhibit the right IBNs, right EBNs, and right motor neurons (probably presynaptically). This crossed inhibition chokes off the drive to the motor neurons and stops the movement.

During the movement, the CB is also sending a signal to the SC that inhibits the SC more and more toward the end of the movement (squelch). Thus, the SC is inside a feedback loop, but it is not a classic error control feedback loop. It simply causes the SC output to decay somewhat during the movement, which may make it easier to start a new saccade after the current one is completed.

A diagram of how the system works is shown in Figure 8(a). On the left side, the bullet-shape represents the retinotopic map on the SC. On the right side, the upper rectangle represents the vermis (continuous across the midline), and the two squares represent the left and right cFN oculomotor regions (FOR). The appearance of the target causes a locus of activity somewhere in the SC (blue disk). This is communicated to the CB, but the projection diverges widely and, thus. cells in most of the vermis become active, which shuts down the FOR (red arrows). The innervation for the movement is the sum of the outputs from the SC and the CB (with the sign reversed for the ipsilateral FOR, making it a choke signal). This sum drives the brain stem burst neurons, which in turn provide feedback to the vermis. During the saccade, the feedback must cause some kind of updating in the CB, which causes different parts of it to pause as the movement progresses. This can be thought of as a spreading wave across the CB cortex. However, it is important to keep in mind that the mechanism underlying this apparent spread is unknown. Other mechanisms (such as a population of neurons with different thresholds) could also cause the activity in the ipsilateral vermis to pause at the right time. Nonetheless, it is important to realize that the spread of the pause in the vermis cannot be generated simply by some kind of diffusion process within the CB or by transport delay (along the parallel fibers that cross the midline). If that were the case, the movement would be unable to compensate for motor noise in the system. The only way the CB can steer the movement to the target is by processing instantaneous feedback information.

An example of how movement context is incorporated into the model is shown in Figure 8(b). Suppose that the target appears on the horizontal meridian, at the same eccentricity as in Figure 8(a). Then, the locus of activity in the SC would be same in both cases (blue disk). However, the CB also receives information that the target is moving. Based on its previous experience, the CB has learned that to acquire this target the initial locus of activity must be moved a bit down and to

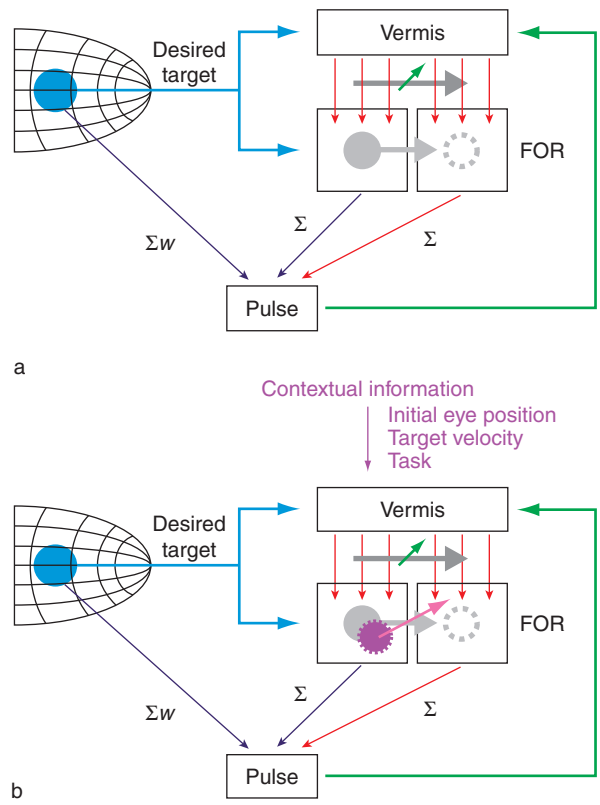

Figure 8 Interaction of drive signals from superior colliculus (SC) and cerebellum determine the saccade (two-dimensional maps of the SC and cerebellum show where activity must be for the desired eye movement): (a) under normal circumstances; (b) when contextual information (e.g., target velocity, behavioral commands, or initial eye position) indicate that a different movement is called for. In (a), the output of the SC and the corresponding output of the cerebellum are simply summed (with the sign of the choke reversed, red arrow). Feedback (green arrows) causes the activity at different places in the vermis to pause at the right time (as if a wave of inhibition were spreading across the vermis, gray arrow). In (b), the initial locus of the pause in the vermis is displaced, creating a new drive signal from the fastigial oculomotor region (FOR), as shown by the magenta disk. During the eye movement, feedback drives the spread of the Purkinje cell (PC) pause along the magenta arrow. Thus, the total output (sum of SC and cerebellar drives) gets the eye on target. $\sum_w$, weighted sum.

the right (magenta disk). The sum of the SC and CB outputs now causes the eye to move up and to the right. Feedback causes the pause in the vermis (or burst in the FOR) to move up and to the right (magenta arrow). The eye is thus successfully steered to it rendezvous with the moving target.

The Brain Does Not Have a Desired Eye Movement Signal

The division of the drive signal across different areas in the distributed model has profound consequences. The most important question is: where in the brain is the desired eye movement represented? Let us consider in more detail a saccade to a moving target. Figure 9(a) shows the target configurations and the corresponding eye movements. Blue symbols and traces show a saccade to a stationary target at 20° eccentricity. Green symbols and traces show a saccade to a stationary target at 15°. Orange symbols and traces show a saccade to a target (*) appearing at 20° and moving toward the fixation point (+) at 50° s^{-1}. If the delay between target appearance and the saccade is 100 ms, then the target will be 5° closer to the fixation point at the time of the saccade than when it appeared. Thus, to rendezvous with the moving target the eye needs to make only a 15° saccade (and then pursue the target).

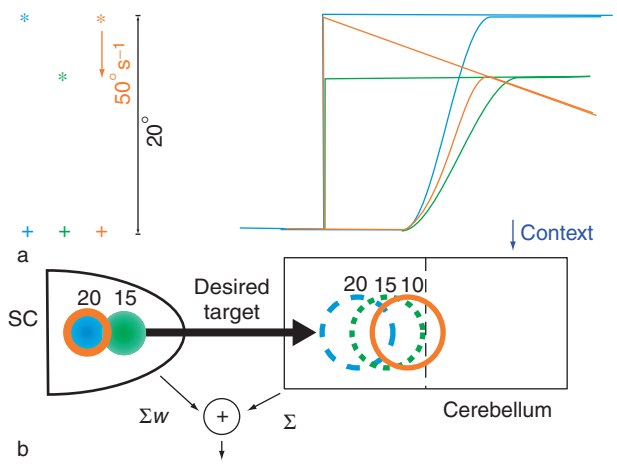

Figure 9 The brain's representation of the desired target is distributed: (a) the target configurations and the corresponding eye movements; (b) diagrams of the superior colliculus (SC; retinotopic) and cerebellum (CB; motorotopic). In (a), suppose the saccade is made from the fixation point (blue plus) to a stationary target at 20° (blue asterisk). The representation of the target in the (SC) is at the 20° locus (blue disk). The representation of the target in the CB is also at the 20° locus (dashed blue circle). The resultant saccade goes to the target (blue trace). If the saccade is made to a stationary target at 15° (green), the locus of activity in the SC is more rostral and the locus of activity in the CB is more medial (green disk and dotted circle). However, when the saccade is made to a moving target that appears at 20° and moves toward the fixation point at 50° s^{-1}, the loci of activity in the SC and CB differ (b). If the saccade occurs in 100 ms, the target will have moved closer, so only a 15° saccade needs to be made. In this case, the locus of activity in the SC is still at the 20° site (orange ring) because the SC encodes the location of the appearance of the target. The CB receives the location of the original appearance, but it also receives information about the velocity of the target (Context). Based on previous experience, the CB has learned that it needs to make only a 10° contribution for this movement. Thus, activity starts at that site (orange ring). The total outputs of the SC and the CB are now correct to make a 15° movement. However, neither area alone encoded the actual size of the movement. The most fundamental signal of the lumped model, the desired eye movement, thus does not exist in the distributed model. $\sum_w$, weighted sum.

Figure 9(b) shows a diagram of the SC (retinotopic) and CB (motorotopic) maps. From the results just described, we know where the activity in the SC and CB must be localized during saccades to stationary targets. For the 20° stationary target, the locus of activity in the SC is at the appropriate caudal site (blue disk), and in the CB the pause is at the 20° site (blue dashed circle). If the effective drive is the average of the retinal and pilot signals, then a 20° saccade results. The same argument reveals that the loci in the SC and the CB must be at the 15° sites for the shorter saccade to a closer target (green disk and green dotted circle).

Now, consider the orange target. The target appears at 20° eccentricity. Thus, the locus of activity in the SC, which knows only which target was selected in retinotopic coordinates, must be at the 20° site (orange circle). Where is the locus of the pause in the CB? If it were also at the 20° site (blue dashed circle), a 20° saccade would result, which would be too large. If the locus in the CB were at the 15° site (green dotted circle), which is the calculated size of the desired movement based on the target's motion, then the saccade would still be too big because the SC is still putting out a command for a 20° saccade. Instead, the pause in the CB must be at the 10° site (orange circle) because, if the SC puts out too much drive, the CB must put out a smaller drive and/or choke sooner. The average of those two drives gets the eye on target. Thus, neither the SC nor the CB encodes the size of the movement. Desired movement size is distributed across several brain structures and is an emergent property of the neuronal circuit.

The distributed model does not contain a signal encoding the desired eye movement. Why is this so important? Because all classical models based on control theory or optimization theory work by driving a motor error signal to zero. However, the existence of an instantaneous motor error signal depends on a comparator to calculate the difference between desired and accomplished eye movements (Figure 10(b)). Since Robinson's original proposal of the local feedback loop in 1975, no neurophysiological evidence

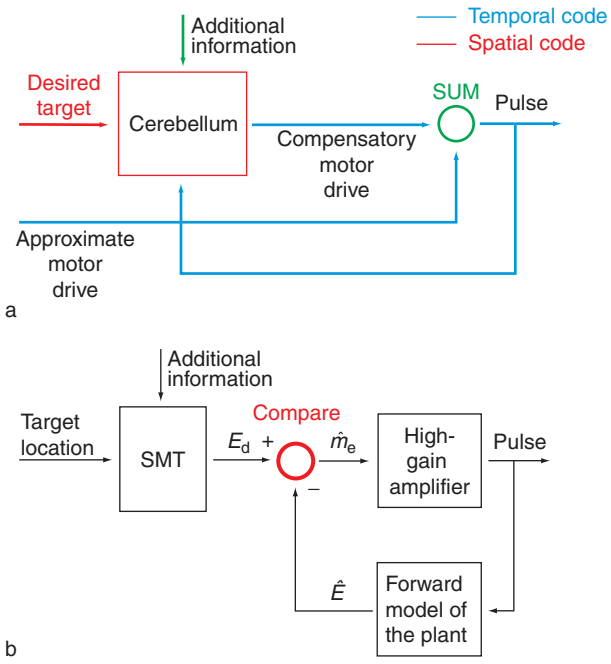

Figure 10 Comparison of models of the saccadic system: (a) distributed; (b) lumped. In (b), the lumped model, the signals all encode physical values. The input to the pulse generator is the desired eye orientation (E_d). It is compared to the efference copy of eye orientation, producing motor error (m_e). However, the distributed model of the saccadic system (a) has no signal corresponding to the desired eye movement. Thus, no comparator is present in the distributed model and there is no motor error signal. In the distributed model, the information processed in the cerebellum is spatially coded (red), but in the downstream areas information is temporally coded (blue). Note that the two types of signals never mix (i.e., there is no comparator). Instead, the temporal activity of the feedback signal causes an updating of the spatial locus of activity in the cerebellum. $\hat{E}$, estimate of eye orientation; STM, sensory-to-motor transform.

of a neuronal population encoding the desired eye movement, let alone dynamic motor error, has ever been found. The neuromimetic, distributed model does not use a desired eye movement signal, and thus it cannot compute a motor error; hence, it does not need a comparator. Instead, the model uses a desired target signal in retinotopic coordinates and knowledge of the current context of the movement, and it works because, after repeated saccades, the CB learns to make the correct contribution, on average, in that situation. Instantaneous feedback of an efference copy signal allows the CB to compensate for motor noise on an individual saccade by updating the locus of activity in the CB. Thus, the brain does not appear to have, and does not need, either a motor error or a desired movement signal.

See also: Brainstem Control of Eye Movements; Cerebellum and Oculomotor Control; Cerebellum: Models; Cortical Control of Eye Movements; Neural Integrator Models; Oculomotor Control: Anatomical Pathways; Pursuit Eye Movements; Saccade–Pursuit Interactions; Saccades and Visual Search; Saccadic Eye Movements; Superior Colliculus; Target Selection for Pursuit and Saccades.

Further Reading

Büttner-Ennever JA (ed.) (2005) *Neuroanatomy of the Oculomotor System*, Progress in Brain Research, vol. 151. Amsterdam: Elsevier.

Carpenter RHS (1977) *Movements of the Eyes*. London: Psion.

Demer JL, Miller JM, Poukens V, Vinters HV, and Glasgow BJ (1995) Evidence for fibromuscular pulleys of the recti extraocular muscles. *Investigative Ophthalmology & Visual Science* 36(6): 1125–1136.

Jürgens R, Becker W, and Kornhuber HH (1981) Natural and drug-induced variations of velocity and duration of human saccadic eye movements: Evidence for a control of the neural pulse generator by local feedback. *Biological Cybernetics* 39(2): 87–96.

Lefèvre P, Quaia C, and Optican LM (1998) Distributed model of control of saccades by superior colliculus and cerebellum. *Neural Networks* 11(7–8): 1175–1190.

Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*. New York: Oxford.

Miller JM (1989) Functional anatomy of normal human rectus muscles. *Vision Research* 29(2): 223–240.

Quaia C, Lefèvre P, and Optican LM (1999) Model of the control of saccades by superior colliculus and cerebellum. *Journal of Neurophysiology* 82(2): 999–1018.

Quaia C and Optican LM (2003) Three-dimensional rotations of the eye. In: Kaufman PL and Alm A (eds.) *Adler's Physiology of the Eye: Clinical Application*, pp. 818–829. New York: Mosby.

Ramat S, Leigh RJ, Zee DS, and Optican LM (2005) Ocular oscillations generated by coupling of brainstem excitatory and inhibitory saccadic burst neurons. *Experimental Brain Research* 160(1): 89–106.

Robinson DA (1981) Models of the mechanics of eye movements. In: Zuber BL (ed.) *Models of Oculomotor Behavior and Control*, pp. 21–41. Boca Raton, FL: CRC Press.

Skavenski AA and Robinson DA (1973) Role of abducens neurons in vestibuloocular reflex. *Journal of Neurophysiology* 36(4): 724–738.

Tweed D and Vilis T (1987) Implications of rotational kinematics for the oculomotor system in three dimensions. *Journal of Neurophysiology* 58(4): 832–849.

Zee DS, Optican LM, Cooke JD, Robinson DA, and Engel WK (1976) Slow saccades in spinocerebellar degeneration. *Archives of Neurology* 33(4): 243–251.

Olfaction in Invertebrates: *Drosophila*

R R H Anholt, North Carolina State University, Raleigh, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Drosophila as a Genetic Model

Ever since Thomas Hunt Morgan's groundbreaking work in the 1920s, *Drosophila* has been recognized as a powerful genetic model system. The genome of *Drosophila* consists of three major chromosomes (the X chromosome and second and third chromosomes) and a small, mostly heterochromatic fourth chromosome, which contains few genes. The genome encoded by the major chromosomes is divided into 100 cytological bands, of which 1–20 are located on the X chromosome, 21–40 and 41–60 are on the left and right arms of the second chromosome, respectively, and 61–80 and 81–100 are on the left and right arms of the third chromosome, respectively. Each cytological band is subdivided into six regions, indicated by the letters A–F. Recombination occurs only in females. Mutations can be readily generated and stably propagated through the use of balancer chromosomes, which contain large inversions and suppress recombination. Gene disruption by transposon mobilization has become the preferred way to introduce mutations. Transposons can also be used as vectors to drive transgenes in specific cells. Such transgenes can either be used to rescue mutant phenotypes or to drive fluorescent reporter genes as neuroanatomical markers; for toxic or apoptotic proteins that can ablate cells or render neurons dysfunctional; and for RNAi constructs that can suppress expression of specific genes. Large numbers of mutant stocks in controlled genetic backgrounds are available to *Drosophila* researchers from national stock centers. In addition, fly stocks with overlapping deficiencies, in which parts of chromosomes are missing, that cover about 70% of the genome are also available and can be used for high-resolution gene mapping.

The genome of *Drosophila melanogaster* was sequenced in 2000 and has been extensively annotated. In addition, genomes of several other *Drosophila* species have also been sequenced, enabling comparative and evolutionary analyses. In addition to its advantages as a genetic model, *Drosophila* is also readily amenable to electrophysiological, behavioral, and neuroanatomical experimentation. These experimental advantages and the relative simplicity of its olfactory system have made *Drosophila* an attractive model for studies on chemosensation.

Anatomy of the *Drosophila* Olfactory System

The olfactory system of *Drosophila* shows remarkable similarities in its functional organization to the mammalian olfactory system and, because of its reduced numerical simplicity, has become one of the best characterized chemosensory systems. It consists of ~1200 olfactory neurons compartmentalized in three morphologically distinct sensilla – the basiconic, coeloconic, and trichoid sensilla, in the third antennal segments – and ~120 olfactory neurons in basiconic sensilla in the maxillary palps. Basiconic sensilla are further subdivided into large and small basiconic sensilla. The sensilla are fluid-filled porous structures that comprise dendrites from one to four olfactory neurons. The axons of olfactory sensory neurons project via the antennal nerve to the antennal lobes, where multiple axons from primary olfactory neurons converge to form synapses with a single or few projection neurons in complex spherical structures, 'named glomeruli,' of the neuropil. The glomeruli also contain synaptic contacts with local interneurons which are intrinsic to the antennal lobe and are primarily inhibitory. Each antennal lobe contains ~43 individually identifiable glomeruli. Axons from olfactory sensory neurons project ipsilaterally, but bifurcate in the vicinity of the antennal lobe to form an additional projection to the contralateral antennal lobe via a commissure that connects the two antennal lobes. Output neurons from the antennal lobes project ipsilaterally to Kenyon cells of the mushroom bodies and the lateral horn of the proto cerebrum in the central brain. Cell bodies of output neurons are organized as three distinct clusters in the antennal lobes – the dorsal, ventral, and lateral clusters. The primary olfactory projection in larvae is similar to that in the adult brain in that olfactory sensory neurons project to glomeruli in the larval antennal lobe, but the larval olfactory system is even simpler, consisting of only 21 olfactory sensory neurons, contained bilaterally in two symmetrically located dorsal organs.

Development of the Olfactory System in *Drosophila*

Early development of the olfactory system in *Drosophila* is regulated by basic helix–loop–helix transcription factors, including atonal, achaete-scute, and amos, and by the transcriptional regulator lozenge. The pattern of atonal expression in the antennal imaginal disc is influenced in turn by hedgehog, wingless, and epidermal growth factor patterning genes. Atonal specifies the development of peripheral glial cells which wrap around growing olfactory sensory axons early during

development as they segregate into three fascicles that exit the antennae toward the antennal lobes. There are at least two subclasses of glia, designated Mz317 and GH146 glial cells. The latter arise in the brain and migrate to the antennae along the developing olfactory axons. Overexpression or misexpression of atonal in all sensory precursor neurons in the imaginal antennal disc results in the conversion of all sensillar classes into coeloconic sensilla, indicating that atonal specifies the formation of this particular sensillar type. Axon collaterals that project to the contralateral antennal lobes are guided by slit-dependent robo2 signaling.

The first adult olfactory sensory neurons arise from the eye-antennal disc at around 16 h after puparium formation, and well-defined glomeruli are observed at around 48 h after puparium formation. The projection sites of axons of olfactory sensory neurons appear to be predetermined. Since the first odorant receptors are expressed 60 h after puparium formation, well after the formation of glomerular connections, odorant receptors appear not to be involved with axonal convergence and targeting during development of the *Drosophila* olfactory system, in contrast to the vertebrate olfactory system. The mosaic analysis using a repressible cell marker (MARCM) method, a genetic technique that uses mitotic recombination to drive expression of a reporter gene in progeny of neuronal precursor cells, has shown that antennal lobe projection neurons form dendritic arborizations at predetermined glomerular locations prior to the arrival of axons of olfactory sensory neurons. The maintenance of the projection neuron map, either at the dendritic sites in the glomeruli or at the axonal projection sites in the protocerebrum, is independent of sensory input. Thus, the *Drosophila* olfactory system is hardwired, and development of the functional adult olfactory pathway at the level of glomeruli requires the precise alignment and integration of the olfactory sensory neuron map with the projection neuron map.

The p21-activated serine–threonine kinase (PAK) and the SH2/SH3 adapter protein Dock have been implicated in precise targeting of axons to their appropriate glomerular destinations. In addition, multiple forms of the cell surface protein Dscam (Down syndrome cell adhesion molecule) are expressed in axons of developing olfactory sensory neurons, and the expression of Dscam is essential for correct wiring between axons and their glomerular targets. Thus, specific adhesion molecules that interact with the extracellular environment and regulate intracellular signaling pathways form a framework within which the axons of olfactory sensory neurons that will express a particular odorant receptor are sorted and guided toward their appropriate glomerular targets.

Odorant-Binding Proteins

When odorants diffuse into the porous sensilla they partition into the aqueous fluid that surrounds the chemosensory dendrites of olfactory neurons (the perilymph) and contains odorant-binding proteins (Obps). Obps are secreted by support cells and comprise a family of about 50 proteins of 18–20 kDa. Although the sequences of these rapidly evolving proteins are divergent, Obps are characterized by six conserved cysteine residues that are thought to be important for maintaining their three-dimensional structures.

Genes encoding Obps are located on all three major *Drosophila* chromosomes and occur both as single genes and in clusters. The recommended nomenclature for these genes is the prefix *Obp* followed by the cytological band location and a lower case identifying letter; for example, *Obp56d* is located at cytological location 56, where it is the fourth member of this family at this chromosomal location (indicated by the letter d).

The functions of Obps remain poorly defined. It has been postulated that they bind hydrophobic odorants to facilitate their partitioning in the aqueous sensillar perilymph and their transport to odorant receptors at the chemosensory membrane. A role for protecting odorants from degradation or a role in the removal of odorants has also been postulated. However, a different protein, named Pinocchio (also designated smell impaired 21F, or smi21F), which is widely and abundantly expressed in the third antennal segment and which is only distantly related to the Obp family, has been implicated in the removal of xenobiotics from the perilymph.

There is good evidence that Obps are important for transporting odorants in the perilymph to their receptors. First, functional and structural studies on the pheromone-binding protein of the silk moth, *Bombyx mori*, reveal a pH-dependent mechanism by which this protein binds and releases bombykol. Second, flies in which one Obp gene had been deleted were attracted to high concentrations of ethanol and short-chain alcohols that were repellant to wild-type flies.The gene encoding this Obp was, therefore, called *lush*, and is synonymous with *Obp76*a. Subsequent electrophysiological and behavioral experiments showed that *lush* mutants did not respond to the pheromone 11-*cis*-vaccenyl acetate, suggesting that Lush may function both in general odorant recognition of short-chain alcohols and pheromone recognition. Normal electrophysiological and behavioral responses could be restored either by delivering recombinant Lush protein along with the pheromone via a recording pipette to trichoid sensilla, where *Obp76a* is normally expressed,

or by introducing an *Obp76a* transgene into the mutant background. The crystal structure of the alcohol-binding site of this Obp has also been reported.

Despite the absolute requirement of Lush for responses to 11-*cis*-vaccenyl acetate, the importance of other Obps in olfaction remains to be determined. Heterologous expression of the *Drosophila* Or43a receptor in *Xenopus* oocytes in the absence of Obps resulted in odorant-induced current responses. Similarly, response profiles of odorant receptors expressed under an *Or22a* promoter in an *Or22a* deletion mutant resembled the response profiles of the same receptors in their native sensilla. It should be noted, however, that high concentrations of odorants were applied to obtain electrophysiological responses in the heterologously expressed Or43a receptor, and it is possible that response profiles of odorant receptors would shift to lower odorant concentrations in the presence of the appropriate Obps.

Some Obps show sexually dimorphic expression. For example, *Obp99a* and *Obp99b* show profound sex-biased expression in females and males, respectively, suggesting that these genes have evolved through gene duplication of an ancestral gene with subsequent sex-specific functional diversification. Altered regulation of expression of some Obps has been observed following mating, during exposure to starvation stress, during the development of alcohol tolerance after exposure to alcohol, and as a correlated response to artificial selection for divergent levels of copulation latency and aggression. These observations suggest that this multigene family of proteins serves important functions that broadly affect the organism's physiology and behavior.

Odorant Receptors

Drosophila olfactory neurons express a unique odorant receptor from a repertoire of 56 odorant receptor (*Or*) genes together with a common odorant receptor, Or83b. These receptors belong to the superfamily of seven-transmembrane domain G-protein-coupled receptors. Their sequences are divergent and they are unrelated to vertebrate odorant receptors. The common Or83b receptor forms a heterodimer with the unique odorant receptor, and this dimerization is essential for transport and insertion of the specific odorant receptors in the chemosensory dendritic membrane. Evidence suggests that *Drosophila* odorant receptors have a unique membrane topology among G-protein-coupled receptors, with their N-termini located on the cytoplasmic side of the cell and the C-termini extracellularly.

Neurons that express the same receptor project bilaterally to one, or rarely two, of about 43 individually identifiable symmetrically located glomeruli in the antennal lobes, where olfactory information is encoded in a spatial and temporal pattern of glomerular activation. Four cases have been identified in which olfactory sensory neurons co-express two receptors together with Or83b, and one case has been observed in which a glomerulus receives input from two populations of olfactory sensory neurons expressing different receptors. Sexually dimorphic glomeruli, which may be involved in courtship and mating behavior, receive input from three groups of olfactory sensory neurons expressing unique odorant receptors.

Organization and Response Properties of Odorant Receptors

Elegant extracellular recordings from sensilla in the maxillary palps and third antennal segments of *D. melanogaster* have characterized the response properties of olfactory sensory neurons and shown that sensilla can be classified as specific types that contain invariant combinations of neurons with defined molecular receptive fields. Each maxillary palp contains only 120 olfactory neurons housed in 60 basiconic sensilla with two neurons per sensillum. Based on their odorant response profiles, olfactory sensory neurons in the maxillary palp can be classified into six functional classes, which follow a strict pairing rule such that two neurons with particular response profiles always occur together. Thus, sensilla in the maxillary palps can be classified into three groups, designated pb1, pb2, and pb3 (pb indicates palp basiconic), and responses from individual neurons within these sensilla can be distinguished based on their spike amplitudes. A similar organization is found in the antennae, where two olfactory sensory neurons are housed together within the basiconic sensilla, with one sensillar type containing four olfactory sensory neurons. Again, neurons with defined odorant response profiles are housed within these sensilla according to strict pairing rules. Based on a diagnostic panel of 12 odorants, there are 16 classes of olfactory sensory neurons housed in seven different types of sensilla in the antennae, designated ab1–ab3 for the large basiconic sensilla and ab4–ab7 for the small basiconic sensilla (ab indicates 'antenna basiconic'; see Figure 1). The distribution of the different sensilla varies across the antennae. In addition, the response kinetics to different odorants will vary among olfactory sensory neurons, and both excitatory and inhibitory responses to odorants have been observed.

A deletion mutant lacking the *Or22a* and *Or22b* gene enables ectopic expression of receptors under

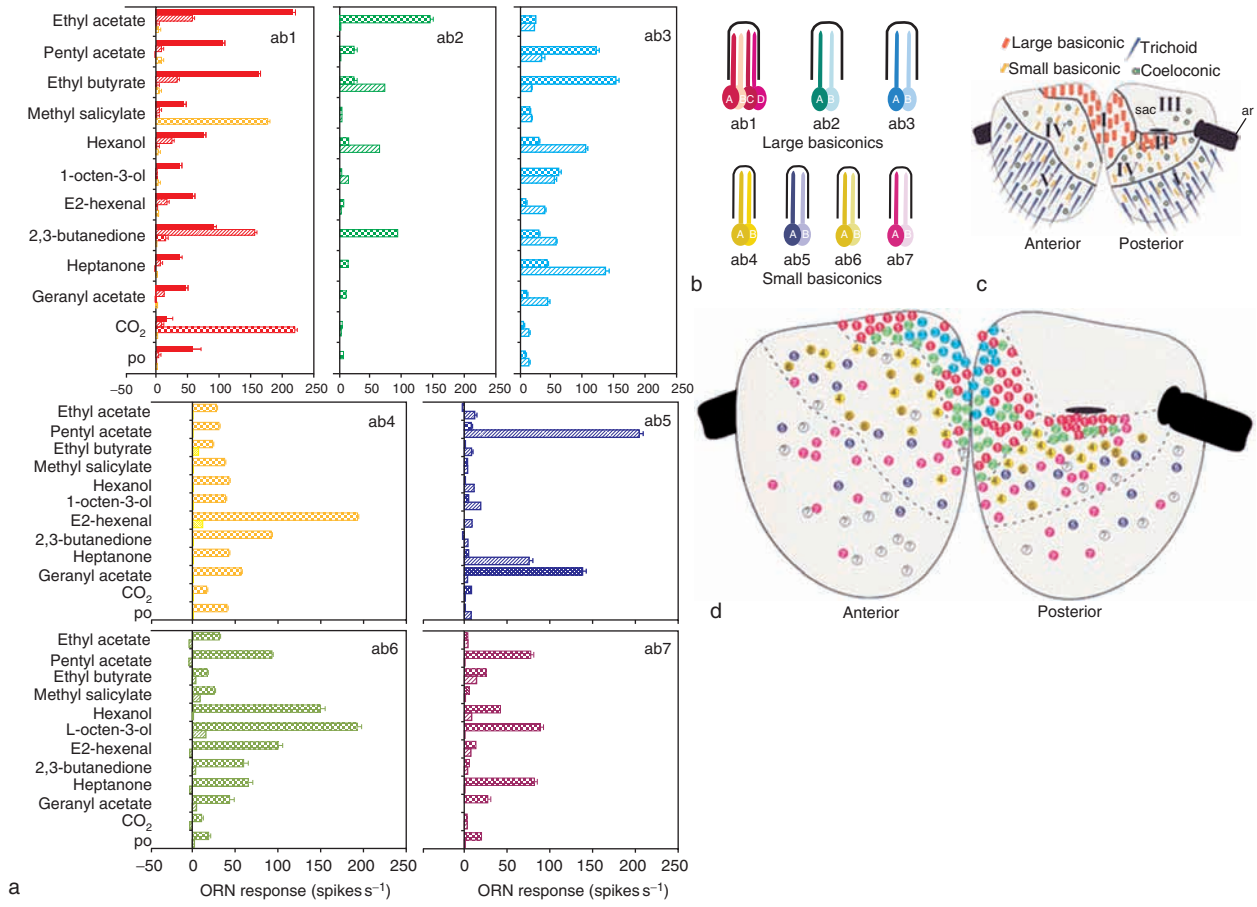

Figure 1 Functional characterization of *Drosophila* olfactory receptor neurons. (a) Characterization of olfactory receptor neuron (ORN) response profiles for three types of large basiconic sensilla (ab1, ab2, and ab3) and four types of small basiconic sensilla (ab4, ab5, ab6, and ab7) with 11 diagnostic stimuli and the solvent control, paraffin oil (po). Responses are measured as the increase (or decrease) in spikes per second compared to the spontaneous firing frequency. Sixteen classes of olfactory sensory neurons occur in characteristic combinations in the seven basiconic sensilla types. (b) Schematic representation of the seven functionally characterized basiconic sensilla types in the antenna. (c) Distribution of morphological subtypes of sensilla across the third antennal segment. The arista (ar) is a mechanosensory structure that extends from the antenna. (d) The distribution of functional types of basiconic sensilla described in (a and b). Each circle represents a recording from one sensillum (question marks represent recordings that could not be classified).The dotted lines represent boundaries between morphological categories of sensilla. Dorsal is at the top; medial is along the center of the diagram. Note the distribution of functionally characterized basiconic sensilla in this panel and the location of morphologically characterized small and large basiconic sensilla in (c). Adapted from De Bruyne M, Foster K, and Carlson J (2001) Odor coding in the *Drosophila* antenna. *Neuron* 30: 537–552, with permission from Elsevier.

the *Or22a* promoter in otherwise silent mutant neurons and establishes an elegant *in vivo* assay system for a comprehensive characterization of odorant receptor specificities, including larval odorant receptors, odorant receptors from other species (such as the female-specific AgOr1 receptor from the malaria vector, *Anopheles gambiae*, that recognizes the human host odorant 4-methylphenol), and odorant receptors expressed in the maxillary palps and antennae of adult flies. This analysis was subsequently extended to a comprehensive characterization of the molecular receptive fields and response characteristics of 24 odorant receptors with a panel of more than 100 odorants.

Central Processing of Olfactory Information

The glomerular map is decoded in central brain structures – the lateral horn of the protocerebrum and the mushroom bodies. Axons from projection neurons with dendritic arbors in specific glomeruli in the antennal lobes course through the antennal cerebral tract, form branches in the mushroom body, and terminal axonal arborizations in the lateral horn of the protocerebrum. The axonal arborization patterns in the lateral horn are highly stereotyped and invariant among individuals (Figure 2). Furthermore, projection neurons with similar terminal axon fields receive input

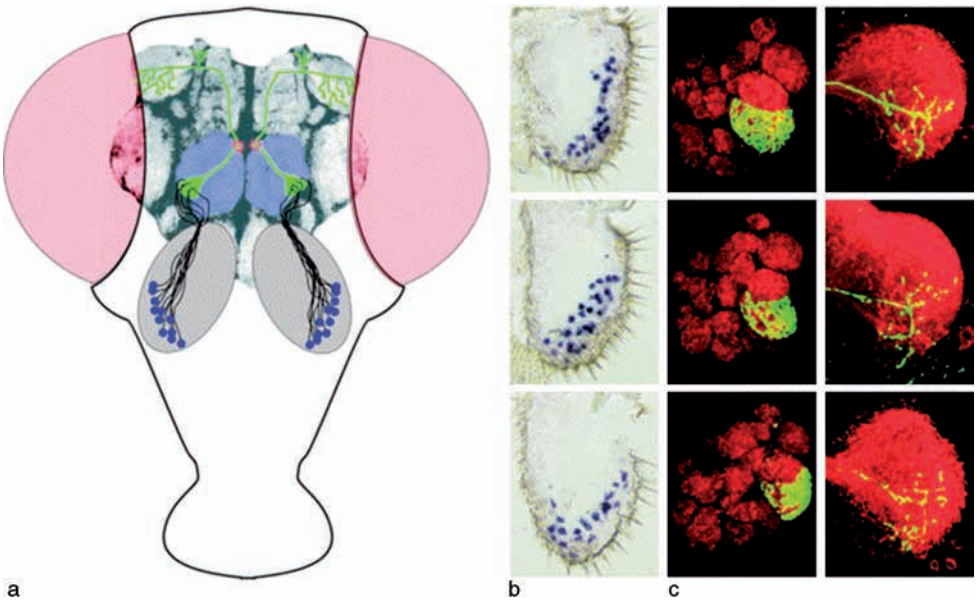

Figure 2 Neural projections in the *Drosophila* olfactory system, illustrating the stereotypic labeled line projection pathway of neurons that express the Or47b receptor. (a) Schematic representation of the *Drosophila* olfactory pathway. The diagram shows a frontal view of the *Drosophila* head, with the compound eyes shaded in red, the antennae in gray, and the antennal lobes in blue. Olfactory sensory neurons that express the Or47b receptor are shown as blue dots, and their axons converge on the VA1l/m glomerulus (green). Axons of output neurons (green) project to the calices of the mushroom bodies and the lateral horn of the protocerebrum. (b) *In situ* hybridization shows stereotypical expression of Or47b in olfactory sensory neurons located in a restricted lateral dorsal region of the third antennal segment. (c) The left-hand panels show that axons from olfactory sensory neurons, in which the Or47b promoter drives expression of green fluorescent protein, converge upon the VA1l/m glomerulus in the antennal lobes. The right-hand panels show stereotypical axonal arborizations of a single projection neuron from the VA1l/m glomerulus in the ventral region of the lateral horn of the protocerebrum. Reproduced from Keller A and Vosshall LB (2003) Decoding olfaction in *Drosophila*. *Current Opinion in Neurobiology* 13: 103–110, with permission.

from adjacent glomeruli. In this aspect the organization of the insect olfactory system is reminiscent of the vertebrate system, where adjacent glomeruli in the olfactory bulb tend to respond to structurally related odorants. The stereotypical projection pattern from the antennal lobe to the lateral horn suggests positional preservation of the glomerular chemotopic map onto the axonal projection map in the lateral horn. The projection to the lateral horn, however, is more diffuse than is the primary olfactory sensory projection to the antennal lobe glomeruli, with substantial interdigitation among axonal innervation fields. The innervation patterns of axons of projection neurons onto the dendrites of Kenyon cells in the calyx of the mushroom body are more variable than is the strict stereotypy found in the lateral horn, suggesting greater synaptic plasticity in the mushroom bodies, which have been implicated with learning and memory. It has been suggested that odor recognition may be a primary function of the lateral horn, whereas the mushroom bodies may be the principal sites for odorant-associated memory formation.

High-resolution calcium imaging with calcium-sensitive fluorescent reporter transgenes enables real-time monitoring of odorant-induced glomerular activation. Such experiments have shown that glomeruli are functional units that are either activated or not, but never show partial activation. Different odorants elicit distinct glomerular activation patterns that are invariant among different individuals and are faithfully transmitted to projection neurons in the activated glomeruli. At low odorant concentrations few glomeruli are activated, but as odorant concentrations increase they elicit activity in a larger segment of the glomerular field, which most likely reflects recruitment of odorant receptors with lower affinity. To what extent temporal characteristics of the glomerular activation map are essential for odorant recognition and olfactory discrimination in *Drosophila* still remains an unresolved issue.

Genetic Architecture of Olfactory Behavior

Odorant Recognition and Behavior

Electrophysiological and behavioral studies have shown that, as in mammalian olfaction, odorant recognition in

Drosophila is combinatorial. A given odorant can interact with multiple receptors and, although receptors differ in the scope of their response profiles, odorant receptors can interact with multiple ligands. In the larval olfactory system systematic ablation of any one of the 21 olfactory sensory neurons results in disruption of chemotactic responses to small subsets of odorants, but when only a single neuron remains intact (i.e., functional redundancy in odorant recognition is removed), larvae can still exhibit chemotaxis. Nevertheless, olfactory sensory neurons acting in concert in this system facilitate odorant recognition. Behavioral assays that can quantify both attractant and avoidance responses of adult flies to odorants have been developed, including Y-maze assays, olfactory trap assays, computer-assisted video-monitoring systems, and simple dipstick assays, in which repellent responses are monitored away from a cotton-wool swab dipped in an odorant solution. Most larval responses to odorants are attractant, whereas both attractant and repellant responses are common in adult flies. However, the direction of a behavioral response to an odorant is often concentration dependent, with attractant responses at low and intermediate concentrations giving way to repellent responses at high odorant concentrations. The response to carbon dioxide has been especially well characterized. Carbon dioxide is a potent repellent, which elicits avoidance behavior by activating olfactory sensory neurons that project to a single glomerulus in the antennal lobe, the V glomerulus.

Gene Networks That Mediate Olfactory Behavior

The behavioral response to external chemical signals requires not only odorant recognition, but also signal transduction, signal propagation, signal processing and interpretation, and generating an appropriate behavioral output. A vast proportion of the genome is required to enable behavioral responses to chemical stimuli. Genes that are essential for olfactory behavior can be readily identified by *P*-element insertional mutagenesis, whereby a transposable element, a genetically engineered DNA construct, inserts in or near a gene that contributes to the behavior, thereby disrupting its function and generating a mutant phenotype. Quantitative genetic studies and whole-genome transcriptional profiling of *P*-element insertional olfactory mutants in controlled genetic backgrounds have shown that the genetic architecture of olfactory behavior consists of interactive networks of multifunctional (pleiotropic) genes characterized by nonadditive (epistatic) effects on the phenotype. These networks are plastic and adjust dynamically to environmental changes. They encompass, among others, genes implicated in development of the nervous system. Hypomorphic mutations in these genes, which show no overt or only subtle neuroanatomical effects, can have profound effects on adult olfactory behavior. A complete understanding of the neurogenetic underpinnings of olfactory behavior will ultimately require linking genetic networks to neural circuitry, a central challenge for future investigations.

See also: Olfaction in Invertebrates: Manduca; Olfaction in Invertebrates: Honeybee; Olfactory Receptors; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Cortex Physiology; Olfactory System Theory; Olfactory System: Circuit Dynamics and Neural Coding in the Locust.

Further Reading

Anholt RRH (2004) Genetic modules and networks for behavior: Lessons from *Drosophila*. *BioEssays* 26: 1299–1306.

Anholt RRH, Dilda CL, Chang S, et al. (2003) The genetic architecture of odor-guided behavior in *Drosophila*: Epistasis and the transcriptome. *Nature Genetics* 35: 180–184.

Clyne P, Warr C, Freeman M, et al. (1999) A novel family of divergent seven-transmembrane proteins: Candidate odorant receptors in *Drosophila*. *Neuron* 22: 327–338.

De Bruyne M, Clyne P, and Carlson J (1999) Odor coding in a model olfactory organ: The *Drosophila* maxillary palp. *Journal of Neuroscience* 19: 4520–4532.

De Bruyne M, Foster K, and Carlson J (2001) Odor coding in the *Drosophila* antenna. *Neuron* 30: 537–552.

Fishilevich E, Domingos AI, Asahina K, et al. (2005) Chemotaxis behavior mediated by single larval olfactory neurons in *Drosophila*. *Current Biology* 15: 2086–2096.

Hallem E, Ho M, and Carlson J (2004) The molecular basis of odor coding in the *Drosophila* antenna. *Cell* 117: 965–979.

Jhaveri D and Rodrigues V (2002) Sensory neurons of the Atonal lineage pioneer the formation of glomeruli within the adult *Drosophila* olfactory lobe. *Development* 129: 1251–1260.

Keller A and Vosshall LB (2003) Decoding olfaction in *Drosophila*. *Current Opinion in Neurobiology* 13: 103–110.

Laissue P, Reiter C, Hiesinger P, et al. (1999) Three-dimensional reconstruction of the antennal lobe in *Drosophila melanogaster*. *Journal of Comparative Neurology* 405: 543–552.

Marin EC, Jefferis GS, Komiyama T, et al. (2002) Representation of the glomerular olfactory map in the *Drosophila* brain. *Cell* 109: 243–255.

Shandbhag S, Muller B, and Steinbrecht A (1999) Atlas of olfactory organs of *Drosophila melanogaster*. 1. Types, external organization, innervation and distribution of olfactory sensilla. *International Journal of Insect Morphology and Embryology* 28: 377–397.

Vosshall L, Amrein H, Morozov P, et al. (1999) A spatial map of olfactory receptor expression in the *Drosophila* antenna. *Cell* 96: 725–736.

Vosshall L, Wong A, and Axel R (2000) An olfactory sensory map in the fly brain. *Cell* 102: 147–159.

Wang J, Wong A, Flores J, et al. (2003) Two-photon calcium imaging reveals an odor-evoked map of activity in the fly brain. *Cell* 112: 271–282.

Wong AM, Wang JW, and Axel R (2002) Spatial representation of the glomerular map in the *Drosophila* protocerebrum. *Cell* 109: 229–241.

Xu P, Atkinson R, Jones DN, et al. (2005) *Drosophila* OBP LUSH is required for activity of pheromone-sensitive neurons. *Neuron* 45: 193–200.

Relevant Websites

http://www.fruitfly.org – Berkeley Drosophila Genome Project (National Human Genome Research Institute, National Cancer Institute, and Department of Energy).

http://flybase.bio.indiana.edu – Database of *Drosophila* genes and genomes (National Human Genome Research Institute, National Institutes of Health).

Olfaction in Invertebrates: Honeybee

R Menzel, Freie Universität Berlin, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Olfactory Receptor Neurons and Glomeruli

The honeybee's strong points as a model system for olfaction research are the facts that odor processing can be studied through learning tests, with odor as an appetitive stimulus, and that odor-induced neural excitation can be monitored both at the single-neuron level and at the neural network level. A total of 60 000 olfactory receptor neurons (ORNs) on each antenna project their axons to the antennal lobes (ALs), which are subdivided into approximately 160 identified glomeruli. Most of the ORNs are in groups of 25–20 in pore plates (sensilla placodea). The number and distribution of receptor molecules are unknown, but in analogy with the findings in *Drosophila* and the mouse – for which a correspondence between the number of olfactory receptor genes and the number of glomeruli has been found – it is assumed that approximately 160 different ORNs may exist on the bee antennae.

Extracellular recordings from ORNs in pore plates showed that the respective chemoprofiles are broad and overlapping. Furthermore, the temporal response characteristic of single ORNs can be complex, including both excitatory and inhibitory components to particular odors, indicating antagonistic processing either at the level of intracellular cellular pathways initiated by more than one receptor protein in each ORN, or by neural processing between ORNs within one pore plate.

Each ORN innervates only one glomerulus, and axons from ORNs within one pore plate reach different glomeruli. The glomeruli are interconnected by about 4000 local interneurons, and about 800 projection neurons (PNs) lead from the glomeruli to higher order brain centers, such as the mushroom bodies and the lateral protocerebrum. Using calcium imaging, it is possible to measure odor-evoked glomerular activity patterns in about 40 of the 160 glomeruli. Using a computerized morphological atlas of AL glomeruli, it was possible to map the identity of glomerular units onto the physiological recordings done with calcium imaging. With this procedure, it was possible to compare the response profiles of individual glomeruli among specimens (Figure 1).

Glomerular Activity Pattern and Odor Identity

Two main issues were addressed with the imaging technique: first, whether odor representation is conserved within the species, and second, whether the activity pattern elicited by an odor is sufficient to predict the odor stimulus. The combinatorial pattern of glomerular activity is indeed highly conserved between individuals, indicating a tight developmental control of ORN input processing in the AL. Furthermore, statistical analysis carried out on 18 of the 160 glomeruli of the bee (11%) using a discriminant analysis showed that odor representations of each given odor form a coherent 'cloud' in the multidimensional space where axes are defined by each of the identified glomeruli. In 86% of these cases the odor could be identified from the glomerular activity.

In a more extensive analysis the generalization profiles of 16 × 16 pairs of odors were determined using the odor conditioning paradigm and compared with the results from optophysiological recordings. The perceptual distances correlate well with physiological distances. It was concluded that the functional groups of the primary and secondary aliphatic alcohols, aldehydes, and ketones, and carbon-chain lengths, are inner dimensions of the honeybee olfactory space and that neural activity in the AL reflects the perceptual quality of odors. In both of these studies the perceptual similarity measure was the generalization profile, the response probability to an odor that was not learned after the animal had been trained to a particular odor.

Another measure of stimulus similarity can be response time. In most examples studied so far, response time and accuracy are inversely related to each other and this appears to apply also for color discrimination of honeybees. Under given conditions of reward and motivation, the brain appears to accumulate the evidence against or in favor of a certain choice until a determined threshold is reached. Free-flying bees can be tested regarding how long they sniff an odor probe, and whether sniffing time differs for the discrimination of more similar or less similar odors. In this study, a behavioral assay was used in which bees had to discriminate odors consisting of mixtures of two components, thus odors that varied greatly between very similar and very different odors. Bees learned to discriminate all of the mixtures, including very similar ones, and even different concentrations of the same odor. Even though discriminating two very similar

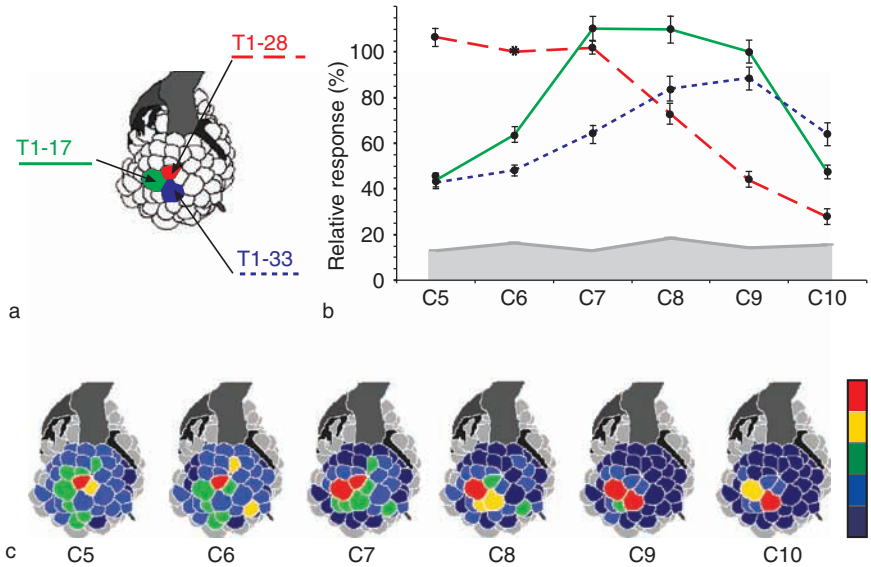

Figure 1 Representation of aliphatic alcohols. (a) Schematic view of the honeybee antennal lobe (AL), with the three glomeruli T1-28, T1-17, and T1-33 as indicated. These three glomeruli are direct neighbors. (b) Responses of the T1-28, T1-17, and T1-33 glomeruli to a series of alcohols, varying in carbon-chain length from C5 (1-pentanol) to C10 (1-decanol). Note that T1-28 responds most strongly to short-chain alcohols, T1-17 to intermediate-chain lengths, and T1-33 to longer chain lengths. Each point represents the average of 14-21 individuals (error bars shown). Responses are shown relative to the response to hexanol in glomerulus T1-28 (asterisk). The shaded area indicates noise levels. (c) Spatial activity patterns elicited by the same six alcohols as those in (b). Red indicates strong activity and dark blue indicates low activity. Glomeruli that could not be measured are shown in gray. Note the continuous shift in activity between T1-28, T1-17, and T1-33 as the carbon-chain length increases. Also note that the response patterns are not limited to these neighboring glomeruli. Adapted from Sachse S, Rappert A, and Galizia CG (1999) The spatial representation of chemical structures in the AL of honeybees: Steps towards the olfactory code. *European Journal of Neuroscience* 11: 3970–3982.

odors appears to be a more difficult task than discriminating two very distinct substances, it was found that the time needed to make a choice (690 ms) for or against an odor was independent of odor similarity.

These data suggest that, irrespective of the nature of the olfactory code, the bee olfactory system evaluates odor quality after a constant interval. This may ensure that odors are only assessed after the olfactory network has optimized its representation, an aspect which will be addressed in the following sections.

Neural Processing in the AL

Glomeruli are connected by about 4000 local interneurons, many of which are immunoreactive to γ-aminobutyric acid (GABA). Comparing the activity patterns of glomeruli as measured predominantly from the input and the output sites of the glomeruli indicated a sharpening of the odor-induced patterns, decorrelating the activity patterns for different odors. Frequently inhibitory responses and 'off' responses were found in the output neurons of the glomeruli, the PNs. Interglomerular inhibition is most likely to occur between glomeruli with similar response profiles in order to sharpen their somewhat fuzzy response profiles.

Glomeruli with similar responses were often found to be direct neighbors, and such neighbors may mutually inhibit each other demonstrating lateral inhibitory mechanisms. However, interglomerular inhibition is not limited to direct neighbors, for the following reasons. First, all glomeruli are approximately equidistant from the center of the antennal lobe: the AL is spherical, with all glomeruli covering the outside of the sphere. Local interneurons innervate several glomeruli; however, their neurites do not travel from one glomerulus to a neighbor, but rather from one glomerulus to the central neuropil and from there to other glomeruli. Second, not all glomeruli with similar response profiles are direct neighbors. To investigate the underlying mechanisms, the GABA-receptor antagonist picrotoxin (PTX) was applied and two separate inhibitory networks were found: one that is GABAergic and modulates overall AL activity, another that is PTX-insensitive and glomerulus-specific.

The net result of the two inhibitory networks together is a globally modulated, contrast-enhanced, and predictable representation of odors in the olfactory output neurons.

Connecting the ALs with the Mushroom Bodies

The second-order neuropils of olfactory processing in the insect brain are the mushroom bodies and the lateral horns (LHs) (Figure 2 (a)). Four different morphological types of PNs leave the AL, two of which project to the lip region of the mushroom body (MB) and to the lateral protocerebrum, the LH. PNs in these two pathways have been analyzed more closely by intracellular recordings. PNs in the median antennocerebral tract (m-ACT) code odors by latency differences or specific inhibitory phases in combination with excitatory phases, have more specific activity profiles for different odors, and thus convey the information with odor-specific delay. The PNs of the lateral antennocerebral tract (l-ACT) code odors by spike rate differences, have broader activity profiles for different odors, and convey the information more quickly. Thus, less precise information about the olfactory stimulus appears to reach the MB and the LH via neurons of the l-ACT, and odor information subsequently becomes more specified by activities of neurons of the m-ACT.

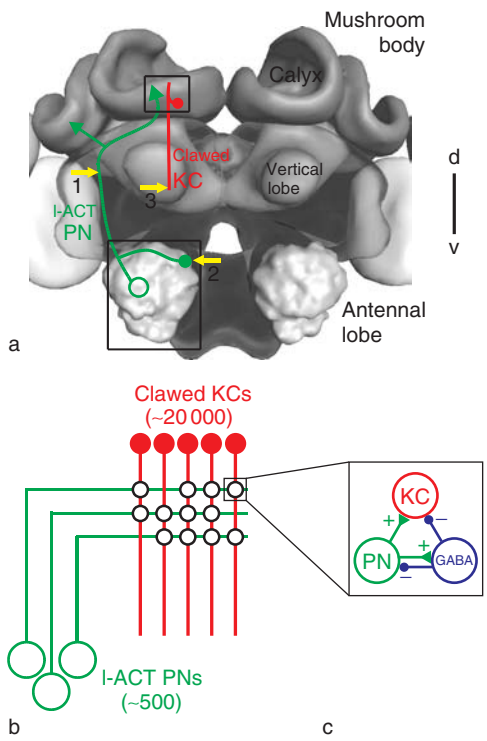

Figure 2 The olfactory pathway in the bee brain and olfactory processing in the MB. (a) Scheme of the bee brain. Lateral antennocerebral tract (l-ACT) projection neurons (PN, green) were optically recorded at their dendrites in the antennal lobe (AL) and at their presynaptic boutons in the lip of the MB calyx. Dendrites and somata of clawed Kenyon cells (clawed KC, red) were recorded in the MB calyx. Yellow arrows indicate sites of dye injection: 1 for recording PN dendrites in the AL, 2 for recording PN boutons in the MB calyces, and 3 for recording KCs in the MB calyces. Squares represent the two imaged areas, AL (for measuring PN dendrites) and MB calyx (for measuring PN boutons and/or clawed KCs). (b) The wiring diagram illustrates the divergent and convergent connectivity between PNs and clawed KCs. About 400 cholinergic l-ACT PNs synapse onto roughly 20 000 clawed KCs. The dendrites of clawed KCs are small, arranged in columns, and feature few clawlike synaptic specializations (black circles). (c) Within the lip region, PNs synapse onto GABAergic neurons which, in turn, make inhibitory synapses with PNs and KCs. Moreover, GABAergic feedback neurons receive input in the MB lobes and send their axons to calyx lip region. Since they leave out the ventral vertical lobe, they presumably do not receive input from clawed KCs. Thus, GABAergic neurons may provide local, PN-driven inhibitory microcircuits within the MB calyx lip, and a more global inhibitory feedback loop between the MB output and input region. Adapted from Ganeshina O and Menzel R (2001) GABA-immunoreactive neurons in the mushroom bodies of the honeybee: An electron microscopic study. *Journal of Comparative Neurology* 437: 335–349.

It was concluded that the separation into two neural pathways from the primary to the secondary centers of olfactory processing is not related to the distinction between different odors, but rather relates to a dual coding of the same odor by two different neuronal strategies in the time domain. Part of the olfactory information may also lie in the timing of action potentials. For example, when olfactory processing is disturbed by applying the GABA-antagonist PTX (which also affects the temporal pattern of AL neuron firing, but may also affect the spatial representation of odors), similar odors are no longer distinguished by honeybees.

Many more experiments are needed to understand the relationship between temporal and spatial odor representation in both the insect and the vertebrate brain. One difficulty lies in the paucity of electrophysiologically recorded cells whose innervated glomeruli have been identified, and which can therefore be used to correlate the observed spatial and temporal activity patterns. The atlas of the AL in the bee brain is an excellent tool to establish, for each glomerulus, an olfactory response profile and temporal response patterns, both for the ORN input and the output via the PNs.

Learning-Related Plasticity in the AL

The plasticity of the neural network of the AL has been studied from two perspectives, that of neural correlates of sensory memory and that of associative memory after olfactory reward conditioning.

Sensory memory is a short-lived persistence of a sensory stimulus in the nervous system, such as iconic memory in the visual system. The effect of an odor stimulus on the postsynaptic responses in PNs was measured within the glomeruli. A single-odor presentation changed the timing of spontaneous activity across glomeruli, enhancing the probability

of coactivity of glomeruli that had been active during odor stimulation shortly before. Moreover, during the first few minutes after odor presentation, correlations between the spontaneous activity fluctuations suffice to reconstruct the stimulus. These results were interpreted to reflect modifiable fluctuations as substrates for Hebbian reverberations and sensory memory, a mechanism that might well be generalized to other neural systems.

Imaging glomerular activity during appetitive learning of an odor leads to an increased response in those glomeruli that are activated by the learned odor, but not in those that respond to a specifically untrained odor. The odor-specific activity pattern was not changed qualitatively, but only quantitatively. Since glomerular activity can also be enhanced by higher concentration of the odor, it is not yet clear how the learning-induced enhancement can be separated from the odor-concentration effect. Since these experiments were carried out under conditions in which the animal was trained to an odor, they not only documented a learning-specific plasticity in the AL, but also proved that the staining with the Ca-fluorescence dye and the imaging procedure did not interfere with the neural processing in the bee brain.

Recently similar associative conditioning experiments combined with Ca imaging were performed with specific staining of a subset of the postsynaptic elements in the glomeruli, the uniglomerular PNs of the lateral antennoglomerular tract. It was found that their responses to odors were remarkably resistant to plasticity following a variety of appetitive olfactory learning paradigms. There was no significant difference in the changes of odor-evoked activity between single- and multiple-trial forward or backward conditioning, differential conditioning, or unrewarded successive odor stimulation. PNs of the lateral antennoglomerular tract may thus be more involved in reliable odor coding rather than modified by learning. The role in olfactory learning and memory retrieval of PNs other than those of the imaged lateral antennoglomerular tract remains to be investigated.

Odor Processing in the MB

PNs convey the output of glomeruli to the MB and the LH (**Figure 2 (a)**). Whereas nothing is known so far about odor processing in the LH, the MBs have been studied recently with Ca-imaging techniques. The MBs are multisensory integration centers that play a dominant role in odor learning. They are densely packed with 170 000 Kenyon cells (KCs), which receive second-order sensory input in the MB calyces, with different modalities innervating spatially distinct areas. Olfactory input is confined to the lip region. KC axons target output neurons in the vertical and medial lobes of the MB. In these areas, GABAergic feedback neurons receive input and project back to the MB lip region, forming an inhibitory loop (**Figure 2 (b)**).

Electrophysiological recordings in locusts and imaging experiments in *Drosophila* indicated that odor representations do indeed differ remarkably in the AL and the MB. Unlike PNs, KCs respond to odors in a sparse way. However, it is unclear whether this transformation of odor representation is a result of the KCs' integration properties (as suggested for locust), or pre- and postsynaptic processing within the MB lip region. Since the bouton-like PN terminals in the MB lip are involved in reciprocal and feedback and forward microcircuits between GABAergic neurons and PNs, information flow from PNs onto KCs may not only be shaped by feed-forward processes, but also may include interactions between PNs and GABAergic neurons within the MB microcircuits.

It was, therefore, asked whether output activity of PN boutons is modified by presynaptic processing within the MB, by comparing odor-evoked responses in PN dendrites within the AL glomeruli and their boutons at the KC synapse. To reveal transformations taking place in the postsynaptic KCs, odor-evoked responses in KCs were recorded and compared to their presynaptic input from PNs (**Figure 3**). At all three processing stages (AL glomeruli, presynaptic boutons of PNs in the MB lip, and postsynaptic spine activity of KCs) odors reliably evoke combinatorial activity patterns. However, in contrast to PNs, KCs code odors in a sparse way, and generate only brief responses at stimulus onset. The KCs' high degree of odor specificity originates at two steps: first, in a presynaptic sharpening of PN synaptic output, and second, in a diminishing KC response. Interestingly, the temporal sharpening of KCs' responses is established only at the postsynaptic side. These results also show that PN activity generated within the first 200 ms determines whether a KC will respond. Thus, the complex temporal response patterns of PNs are transformed into brief phasic responses in KCs.

Thus, two types of transformations occur within the MB: diminishment of a combinatorial code, mediated by pre- and postsynaptic processing within the MB microcircuits, and temporal sharpening of postsynaptic KC responses, probably involving a broader loop of inhibitory recurrent neurons.

Sniffing and Odor Discrimination

As mentioned previously, bees need a constant sniffing time to discriminate similar and different odors. A task-independent choice time implies that quality is assessed after a predetermined time span, and not as soon as the neural codes for distinct odors differ to a sufficient degree in order to justify a decision. As

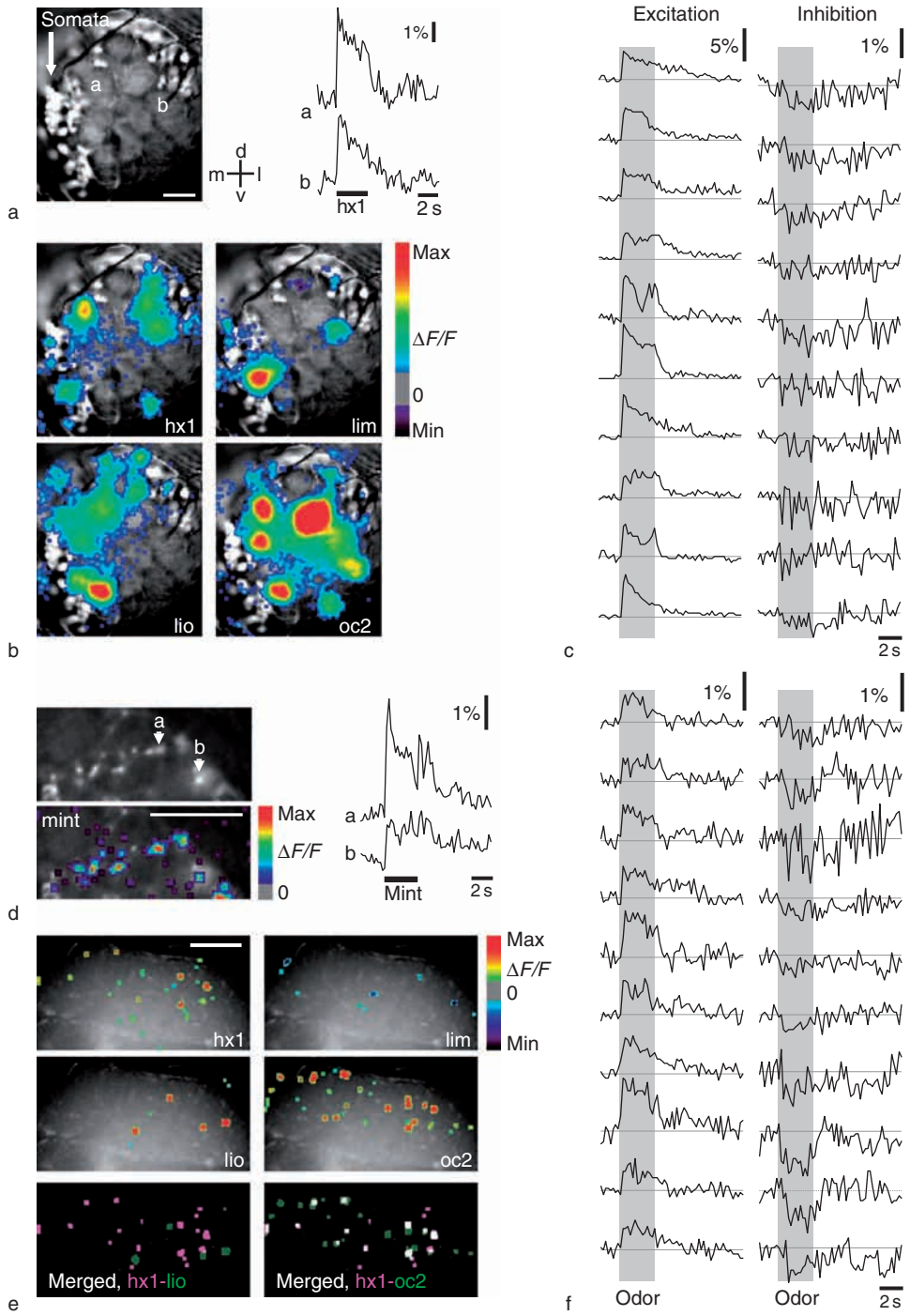

Figure 3 Response properties of l-ACT PNs. (a) Dendritic responses to odor stimulation in uniglomerular l-ACT PNs imaged in the frontal AL glomeruli. Traces represent the time courses of Ca^{2+} transients evoked by 1-hexanol in two glomeruli, (a, b). d, dorsal; m, medial; l, lateral; v, ventral. (b) Color-coded Ca^{2+} signals superimposed on the raw fluorescence image show that different odors activated overlapping combinatorial sets of PN dendrites. (c) Traces show representative excitatory and inhibitory dendritic odor responses in the AL ($n = 10$ bees). Ca^{2+} transients typically followed complex temporal dynamics. Odor stimulus is highlighted. (d) Odor responses in axon terminals (boutons) of PNs in the MB calyx. Boutons are visible in the raw fluorescence image. Stimulation with peppermint induced Ca^{2+} increase in many of the boutons (single measurement). (e) Different odors evoked excitatory and inhibitory responses in distinct boutons. The merged images of excitatory 1-hexanol (hx1, red) and linalool (lio, green) responses visualize distinct sets of activated boutons. In contrast, 1-hexanol (red) and 2-octanol (oc2, green) activated overlapping sets of boutons. All traces represent single measurements; images represent the averages of three stimulations. (f) Odor-evoked Ca^{2+} transients in PN boutons ($n = 8$ bees) show complex dynamics, as is the case in the dendrites. In the entire figure, traces $= \Delta F/F$; scale bars $= 50\,\mu m$. Adapted from Szyszka P, Ditzen M, Galkin A, et al. (2005) Sparsening and temporal sharpening of olfactory representations in the honeybee mushroom bodies. *Journal of Neurophysiology.* 94: 3303–3313.

pointed out by Ditzen *et al*. in 2003, this finding may be important in the context of olfactory coding, since it implies that information about the odor is only available at a predefined time point after stimulus onset. In the case of a 'combinatorial code,' one would have to postulate that the across-glomeruli activity patterns are only read out after a given delay. An internal 'clock' may provide such a signal in the form of odor-evoked oscillations. If these oscillations have a frequency of about 30 Hz in the bee brain – as in the locust brain – a time span of 690 ms gives a high-end estimate of 21 cycles, from which the time not involved in odor detection, such as further processing in the brain, and motor commands, will have to be subtracted.

Both the 'synchrony code' (as suggested for the locust) and a combinatorial code that develops its optimal spatial coding over time (as shown for the honeybee AL) rely on the sequence of activity patterns. The sequences of two very different odors may differ earlier than the sequences of two very similar odors. Under these conditions, one would expect that the brain would continuously monitor the evidence in favor of each alternative, and make a choice as soon as the evidence suffices for one, as has been shown in other experimental paradigms. However, the results reported in the foregoing text suggest that odor evaluation is not incremental, but rather occurs in single information units, consisting of the entire sequence of neural activity following an odor sample. It has been suggested by Ditzen et al. that an interim storage of the sequence would be needed in order to explain the time constancy. Obviously, in the odor coding models provided so far, there is a missing component: a mechanism that would allow for a constant readout time of olfactory quality. This constant time requirement for odor choice may guarantee that odor quality is assessed when its coding is optimal.

See also: Olfaction in Invertebrates: Manduca; Olfaction in Invertebrates: Drosophila; Olfactory Receptors; Olfactory Cortex Physiology; Olfactory System Theory; Olfactory System: Circuit Dynamics and Neural Coding in the Locust.

Further Reading

Chittka L, Dyer AG, Bock F, et al. (2003) Psychophysics: Bees trade off foraging speed for accuracy. *Nature* 424: 388.

Ditzen M, Evers JF, and Galizia CG (2003) Odor similarity does not influence the time needed for odor processing. *Chemical Senses* 28: 781–789.

Faber T, Joerges J, and Menzel R (1999) Associative learning modifies neural representations of odors in the insect brain. *Nature Neuroscience* 2: 74–78.

Galán RF, Weidert M, Menzel R, et al. (2006) Sensory memory for odors is encoded in spontaneous correlated activity between olfactory glomeruli. *Neural Computation* 18: 10–25.

Galizia CG, Küttner A, Joerges J, et al. (2000) Odour representation in honeybee olfactory glomeruli shows slow temporal dynamics: An optical recording study using a voltage-sensitive dye. *Journal of Insect Physiology* 46: 877–886.

Galizia CG and Menzel R (2000) Odour perception in honeybees: Coding information in glomerular patterns. *Current Opinion in Neurobiology* 10: 504–510.

Galizia CG, Sachse S, Rappert A, et al. (1999) The glomerular code for odor representation is species specific in the honeybee *Apis mellifera*. *Nature Neuroscience* 2: 473–478.

Ganeshina O and Menzel R (2001) GABA-immunoreactive neurons in the mushroom bodies of the honeybee: An electron microscopic study. *Journal of Comparative Neurology* 437: 335–349.

Getz WM and Akers RP (1994) Honeybee olfactory sensilla behave as integrated processing units. *Behavioral and Neural Biology* 61: 191–195.

Guerrieri F, Schubert M, Sandoz JC, et al. (2005) Perceptual and neural olfactory similarity in honeybees. *PLoS Biology* 3: e60.

Joerges J, Küttner A, Galizia CG, et al. (1997) Representation of odours and odour mixtures visualized in the honeybee brain. *Nature* 387: 285–288.

Laurent GJ (2003) Olfactory network dynamics and the coding of multidimensional signals (review). *Nature Reviews Neuroscience* 3: 884–895.

Müller D, Abel R, Brandt R, et al. (2002) Differential parallel processing of olfactory information in the honeybee, *Apis mellifera* L. *Journal of Comparative Physiology A* 188: 359–370.

Peele P, Ditzen M, Menzel R, et al. (2006) Appetitive odor learning does not change olfactory coding in a subpopulation of honeybee antennal lobe neurons. *Journal of Comparative Physiology A* 192: 1083–1103.

Sachse S and Galizia CG (2002) The role of inhibition for temporal and spatial odor representation in olfactory output neurons: A calcium imaging study. *Journal of Neurophysiology* 87: 1106–1117.

Sachse S and Galizia CG (2003) The coding of odour-intensity in the honeybee antennal lobe: Local computation optimizes odour representation. *European Journal Neuroscience*. 18: 2119–2132.

Stopfer M and Laurent GJ (1999) Short-term memory in olfactory network dynamics. *Nature*. 402: 664–668.

Stopfer M, Bhagavan S, Smith BH, et al. (1997) Impaired odour discrimination on desynchronization of odour-encoding neural assemblies. *Nature*. 390: 70–74.

Stopfer M, Jayaraman V, and Laurent G (2003) Intensity versus identity coding in an olfactory system. *Neuron*. 39: 991–1004.

Szyszka P, Ditzen M, Galkin A, et al. (2005) Sparsening and temporal sharpening of olfactory representations in the honeybee mushroom bodies. *Journal of Neurophysiology*. 94: 3303–3313.

Olfaction in Invertebrates: *Manduca*

A M Dacks, P G Guerenstein, C E Reisenman, J P Martin, H Lei, and J G Hildebrand, University of Arizona, Tucson, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Invertebrate models have been instrumental to the discovery of many of the fundamental principles of neuroscience. The use of these models in comparative studies provides information about basic principles of neural processing and highlights the constraints under which nervous systems have evolved. In addition to their inherent value to basic neuroscience, invertebrates merit study because of their economic and medical importance. In the past several decades, the hawkmoth *Manduca sexta* (hereinafter referred to as *Manduca*) has become established as an excellent model for the study of the neurobiology of olfaction. The relatively large and experimentally tractable nervous system of *Manduca*, combined with experimentation guided by the natural history of this insect, has placed *Manduca* at the forefront of olfactory research.

Natural History

Much of the research on the neural mechanisms of olfactory processing in *Manduca* has been guided by knowledge of the olfactory environment encountered by moths in nature. The nocturnally active adults must find flowers on which to feed, males must find mates, and females must find the proper plants on which to lay their eggs. For each task, *Manduca* must detect, identify, and discriminate volatile chemicals, both individually and in characteristic blends that signal host flowers (Figure 1(a)), conspecific mates (Figure 1(b)), and healthy leaves on which to lay eggs (Figure 1(c)). *Manduca* prefers the characteristic floral blends produced by certain plants but can learn to identify others as sources of nectar. The characteristic blend of sex pheromone components released by female moths may be present at very low concentrations and is encountered as discontinuous filaments blown downwind from their source. In addition, females avoid laying eggs on plants that are releasing elevated quantities of certain volatile compounds in response to damage by herbivory. A complete description of the neural mechanisms of olfaction in *Manduca* must explain how moths process information in the context of their specific olfactory environment.

Anatomy of the Olfactory System of *Manduca*

Olfactory Receptor Cells

The distalmost segment of the antenna – the flagellum – is the main olfactory organ of moths (Figure 2(a)). It carries thousands of hairlike, multiporous olfactory organules called sensilla, each containing one to a few sensory cells that function as odor detectors. Female antennas possess only short to medium-length sensilla, whereas the male antennas carry additional, male-specific, long sensilla. The antennal flagellum is divided into subsegments (flagellomeres or annuli) each bearing several types of sensilla. Different morphological types of sensilla are categorized based on their length, shape, surface features, pore density, and dendritic branching of their sensory cells. The best-studied types of sensilla in *Manduca* are the trichoid and basiconic sensilla. In addition to the antennal flagellum, there is also an olfactory organ, called the labial palp pit organ (LPO), in the third (distalmost) segment of each of the paired labial palps.

Trichoid and basiconic sensilla typically house two or three olfactory receptor cells (ORCs) that use acetylcholine as their neurotransmitter (Figure 2(b)). The dendrites of ORCs reside within the shaft of the sensillum, whereas the somata of ORCs and supporting sheath cells are located in the epithelium at the base of the sensillum. The axons of ORCs leave the sensillum via the antennal nerve (AN) and project to and terminate in the primary olfactory center in the brain, the antennal lobe (AL), which possesses an array of condensed neuropil structures called glomeruli. The ORCs in the sensilla of the LPO project axons via the labial palp nerve to the subesophageal ganglion and then through it to terminate in the AL. ORCs express specific olfactory receptor proteins (ORs), and axons of ORCs expressing the same OR(s) converge on a single glomerulus. Thus, activation of a particular type of OR can be observed in the activation of its corresponding glomerulus. ORCs expressing different ORs are intermingled throughout the antennas and the AN, but during pupal development the ORC axons, as they approach the AL through the AN, reorganize themselves into fascicles destined for specific glomeruli (Figure 2(c)). This occurs when the growth cones of the developing axons encounter a region rich in glial cells called the 'sorting zone' near the entrance to

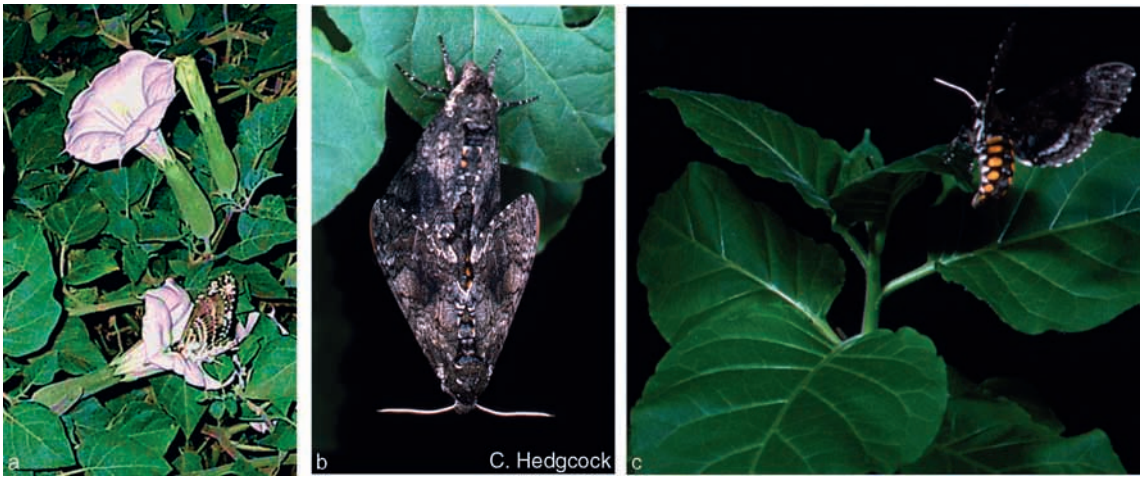

Figure 1 Examples of natural olfactory-guided behaviors of *Manduca*. (a) An adult moth feeding from a flower of one of its preferred host plants, *Datura wrightii*. (b) Two moths mate on a leaf after the male has located the female by following her sex pheromone plume. (c) A female moth oviposits (lays an egg) on the underside of a *D. wrightii* leaf. (b, c) Courtesy of C Hedgcock.

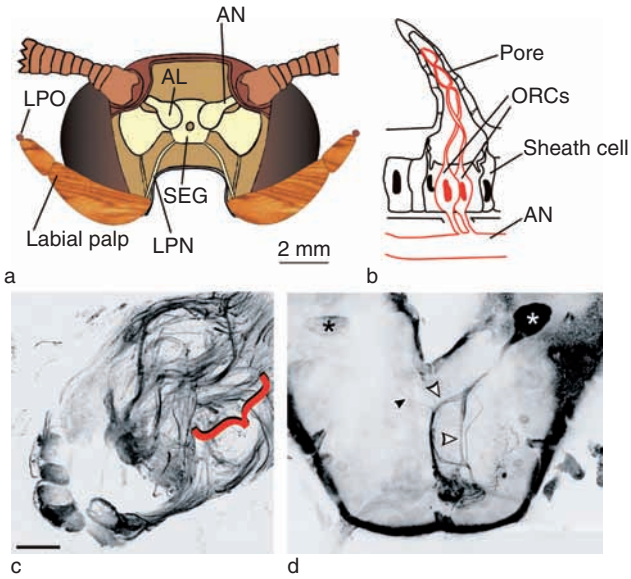

Figure 2 Peripheral sensory structures of the olfactory system of *Manduca*. (a) Schematic frontal view of the moth's head. Each antennal lobe (AL) receives input from olfactory receptor cells in the antenna via axons projecting through the antennal nerve (AN). The AL also receives input from receptor cells located in the labial palp pit organs (LPOs) in the distalmost segment of each of the paired labial palps. The axons of these cells project from each LPO through its labial palp nerve (LPN) to the subesophageal ganglion (SEG) and through it to the AL. (b) Illustration of a multiporous sensillum which contains olfactory receptor cells (ORCs) and sheath cells. (c) Anterogradely stained ORC axons from one antenna of *Manduca*. As the axons of ORCs pass through the sorting zone (bracket), they reorganize into distinct bundles bound for specific glomeruli. (a) From Guerenstein PG, Christensen TA, and Hildebrand JG (2004) Sensory processing of ambient-CO_2 information in the brain of the moth *Manduca Sexta*. *Journal of Comparative Physiology A* 190: 707–725; (b) From Payne T (1979) Pheromone and host odor perception in bark beetles. In: Narahashi T (ed.) *Neurotoxicology of Insecticidal and Pheromones*, pp. 27–57. New York: Plenum; (c) Courtesy of M Higgins and LP Tolbert. (d) Anterogradely stained projections (arrowheads) from an LPO. LPO ORCs project bilaterally to the LPO glomerulus of each AL (white arrowheads and asterisk, ipsilateral projection; black arrowhead and asterisk, contralateral projection).

the AL. At the developmental stage when the ORC axons grow through the sorting zone, the AL has a relatively undifferentiated, coarse neuropil surrounded by a rind of AL glia. The ORC axons project through that glial sheath to form a 'protoglomerular layer.' Soon after, the developing processes of AL intrinsic neurons innervate the protoglomerular layer, and AL glia surround and stabilize each protoglomerulus. Thus, the ORCs confer both a physiological and an anatomical identity on each glomerulus.

The Antennal Lobes

Glomeruli are uniquely identifiable, functional units of the AL and the first synaptic sites within the olfactory system. Glomeruli are composed of densely packed neurites of, and synapses among, ORCs, AL neurons, and centrifugal neurons and are surrounded by a sheath of glial cells. In *Manduca*, chemical synapses in the AL are confined to the glomeruli, which surround a region of coarse neuropil comprising neurites of AL neurons and centrifugal neurons projecting into the AL from elsewhere in the central nervous system. The ALs of both males (Figure 3(b)) and females (Figure 3(a)) have 63 ± 1 glomeruli, all of which have been characterized on the basis of their characteristic shapes, sizes, and relative positions. Some of these glomeruli also have been characterized physiologically. The AL of *Manduca* includes two olfactory subsystems, one of which is composed of approximately 60 sexually isomorphic glomeruli that process information about plant volatiles and possibly other odors. The second subsystem comprises three sexually dimorphic glomeruli situated near the entrance of the AN into the AL. In males, these specialized glomeruli (called cumulus, toroid 1, and toroid 2) belong to the macroglomerular complex (MGC) (Figure 3(b)) and are devoted to processing information about the sex pheromone released by conspecific female moths. The male-specific MGC glomeruli receive input from the long male-specific trichoid sensilla on the antenna. The female-specific glomeruli – a pair of 'large female glomeruli' (LFG) and a smaller glomerulus – process information about specific plant volatiles (Figure 3(a)). With the exception of the identified glomerulus that receives input from the ORCs of the LPO (the labial palp pit organ glomerulus (LPOG)) (Figures 2(d), 3(a), and 3(b)), AL glomeruli receive input from ORCs located in the antenna. Two classes of AL neurons have been identified: local interneurons (LNs; $n \approx 360$) (Figure 3(c)) and projection neurons (PNs; $n \approx 800$) (Figure 3(d)). The dendrites of uniglomerular PNs are restricted to a single glomerulus; multiglomerular PNs have arborizations in several glomeruli. The axons of PNs project from the AL to certain higher brain centers (Figures 3(e) and 3(f)). The LNs, which are all or mostly GABAergic inhibitory neurons, are restricted to the AL and vary greatly in morphology, arborizing in variable numbers of glomeruli. Thus, LNs interconnect glomeruli within the AL in a complex inhibitory network. In addition to the intrinsic AL neurons, there are also centrifugal neurons (including neurons that are believed to secrete serotonin and octopamine) that apparently provide feedback from other regions of the brain into the ALs.

Beyond the Antennal Lobe

The AL is connected to the protocerebrum (PC) by five well-defined tracts: the inner, middle, outer, dorsal, and dorsomedial antennocerebral tracts (IACT, MACT, OACT, DACT, and DMACT, respectively) (Figure 3(f)). The major targets of PN axons (via the IACT, MACT, or OACT) are the lateral horn of the PC, inferior lateral PC (ILPC), and calyces of the ipsilateral mushroom body (Figures 3(e) and 3(f)). The IACT is the most prominent connection between the AL and the PC and mainly carries the axons of uniglomerular PNs. In general, IACT fibers extend from the AL posteriorly, then turn laterally and send collateral branches into the calyces before terminating in the lateral PC. The axons of the MGC PNs in the IACT innervate the ILPC. The PNs that follow the MACT appear to be multiglomerular and terminate in the lateral horn, although the MGC PNs in the MACT project to the ILPC. The OACT carries axons primarily of multiglomerular PNs that project to the lateral horn or the ILPC, although some project to the contralateral lateral horn. Some PN axons in the IACT and OACT project first to the lateral PC and then to the calyces of the ipsilateral MB. Neurons in the two dorsal tracts terminate in the contralateral ILPC and lateral horn. The heterogeneity of the projection patterns from the AL to the PC suggests a complicated process of information segregation and integration at the level of the PC. In addition to PNs projecting from the AL to the PC, there are many neurons that project via the DACT or DMACT from the AL to the PC with somata in either the PC or the SEG, respectively.

Physiology and Behavior

The olfactory system of *Manduca* has evolved not only to identify odorants and discriminate them from one another but also to assess the concentration of odorants, identify natural odor mixtures in which components are presented at particular ratios, and follow the rapid changes that occur in natural odor plumes. Studies of olfaction in *Manduca* benefit from the facts that this sensory modality is decisive for the control of behaviors (e.g., finding mates, food, and oviposition sites), much is known about the chemical composition of natural olfactory stimuli (pheromonal and plant-derived volatiles), and glomeruli are identifiable and accessible for electrophysiological recording (Figure 4).

Odor Identification and Discrimination

Sexually receptive female *Manduca* release a sex pheromone comprising at least eight compounds

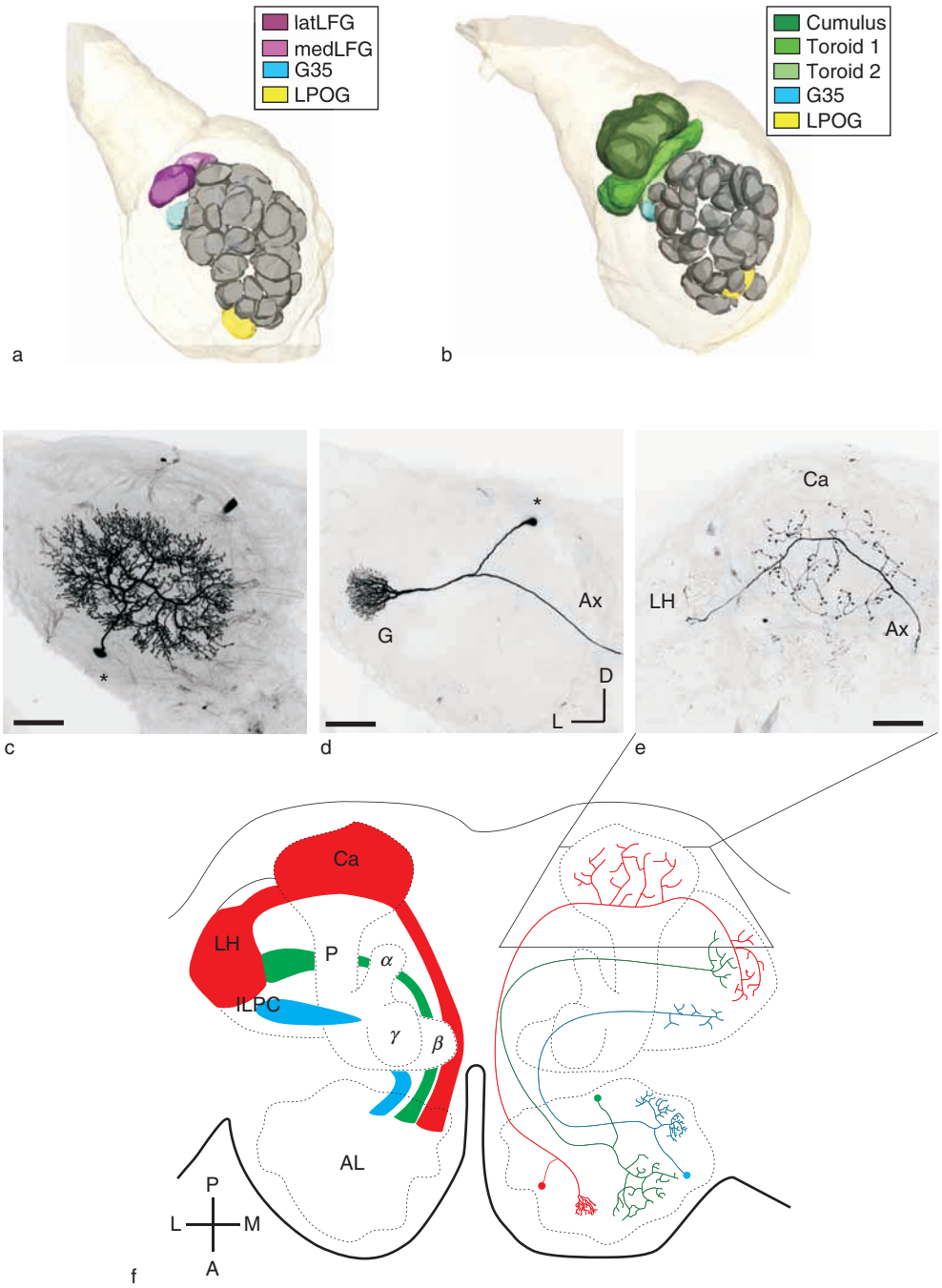

Figure 3 Neuroanatomy of the olfactory system of *Manduca*. (a) Three-dimensional reconstruction of a female antennal lobe (AL) highlighting the lateral and medial large female glomeruli (latLFG and medLFG, respectively), the labial palp pit organ glomerulus (LPOG), and the G35 glomerulus. (b) Three-dimensional reconstruction of a male AL highlighting the three glomeruli (cumulus, toroid 1, and toroid 2) of the macroglomerular complex as well as the LPOG and glomerulus G35. (c) Confocal microscopic scan of a local neuron showing the cell body (asterisk) in the lateral cell group of the AL and arborizations in all of the glomeruli in the AL. (d) Confocal microscopic scan of a uniglomerular projection neuron with cell body (asterisk) in the medial cell group of the AL, dense arborizations confined to one glomerulus (G), and an axon (Ax) that projects out of the AL and, shown in (e), to the protocerebrum where it terminates in the calyces (Ca) of ipsilateral mushroom body and the lateral horn (LH). (f) Projection neurons send axons to the protocerebrum through three major output tracts (left; representative neurons are shown on the right). The inner antennocerebral tract (red) projects to the Ca and LH and carries the axons of a majority of uniglomerular projection neurons. Neurons with axons in the outer antennocerebral tract (blue) and the middle antennocerebral tract (green) are largely multiglomerular and terminate in the inferior lateral protocerebrum (ILPC) and the LH. Note that the dorsal tracts (DACT and DMACT) are not shown in this image. (d–f) P, peduncle; α, α lobe; β, β lobe; γ, γ lobe; A, anterior; D, dorsal; L, lateral; M, medial; P, posterior. Scale bars = 100 μm (c–e).

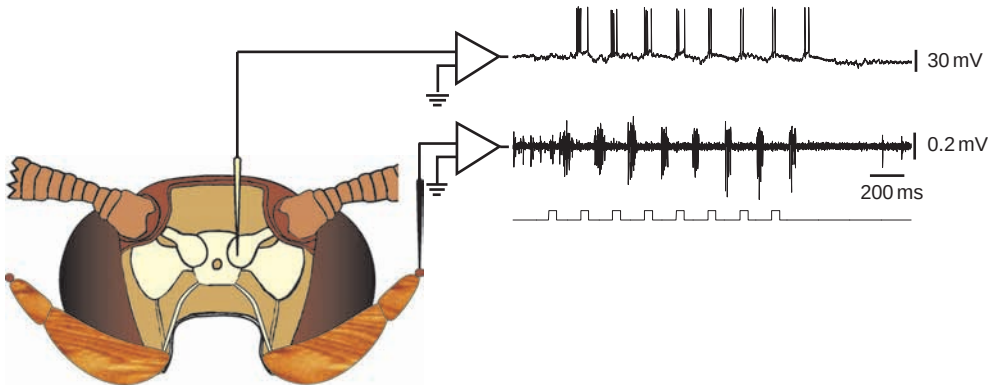

Figure 4 Electrophysiological recordings from cells in the labial palp pit organ (LPO) and the antennal lobe responding to 50-ms pulses of air bearing an increased level of CO_2, delivered at a frequency of five per second. The top trace shows an intracellular recording from a projection neuron in the LPO glomerulus, and the middle trace shows a simultaneous extracellular recording from olfactory receptor cells in the LPO. Stimulus concentration was 600 ppm CO_2 increased from a baseline level of 280 ppm. Stimulus pulses are indicated in the third (bottom) trace. From Guerenstein PG, Christensen TA, and Hildebrand JG (2004) Sensory processing of ambient-CO_2 information in the brain of the moth *Manduca Sexta*. *Journal of Comparative Physiology A* 190: 707–725.

that have been shown to affect the behavior of recipient male moths, but a blend of just two major components ((*E*,*Z*)-10,12-hexadecadienal (bombykal) and (*E*,*E*,*Z*)-10,12,14-hexadecatrienal) is sufficient to attract conspecific males. The individual components alone are ineffective. There are at least four male-specific types of antennal ORCs and one type of sexually isomorphic ORC, all of which are highly sensitive and specific to single pheromone components, including bombykal and (*E*,*E*,*Z*)-10,12,14-hexadecatrienal. ORCs responding to bombykal project exclusively to toroid 1, whereas ORCs responding to (*E*,*E*,*Z*)-10,12,14-hexadecatrienal project to the cumulus. This chemotopic arrangement is preserved in the glomerular output: PNs with dendritic arborizations restricted to toroid 1 are excited selectively by bombykal, and PNs confined to the cumulus are excited selectively by (*E*,*E*,*Z*)-10,12,14-hexadecatrienal (Figures 5(a) and 5(b)). There are some multiglomerular PNs with dendrites in both the cumulus and toroid 1 that are excited by both components. Some neurons in the PC exhibit long-lasting excitation only in response to the proper mixture of the components of the blend, showing response selectivity for a blend at this higher level of processing.

Although the pheromone-specific subsystem in *Manduca* is a paradigm of high sensitivity and specificity, efforts have addressed the widespread notion that the ORCs and PNs associated with other, sexually isomorphic, glomeruli are broadly tuned. The degree of specificity of sexually isomorphic antennal ORCs to host plant volatiles is variable, from specialized ORCs that detect terpenoids to ORCs of intermediate specificity that detect aromatics and generalist ORCs that respond to compounds of different chemical classes at relatively high concentrations. Non-MGC PNs have also been shown to have a range of different specificities. For example, PNs arborizing in the lateral LFG (latLFG) in the ALs of females show preferential responses to the (+) enantiomer of the host plant monoterpenoid, linalool. PNs arborizing in an adjacent glomerulus respond preferentially to the (−) enantiomer, and PNs in other neighboring glomeruli respond equally well to both enantiomers. Thus, it is possible that the AL evaluates the naturally occurring ratios of the enantiomers. A high degree of specificity is also found in the ORCs of the LPO, which detect carbon dioxide (CO_2). CO_2 is emitted in high quantities by newly opened, nectar-producing flowers, and moths preferentially visit flowers releasing high levels of CO_2, suggesting that the moths use CO_2 levels to assess the quality of a nectar resource. Indeed, ORCs in the LPO of both sexes respond exclusively to CO_2, at levels comparable to those emitted by the flowers of host plants. CO_2-responsive ORCs project to the LPOG, which houses arborizations of PNs that respond exclusively to CO_2 (Figures 5(c) and 5(d)). PNs with arborizations limited to another sexually isomorphic glomerulus (G35) show preferential responses to the host plant volatile *Z*-3-hexenyl acetate and also respond to structurally similar odorants presented at higher concentrations (Figures 5(e) and 5(f)). These PNs were shown to be morphologically and physiologically equivalent in both sexes.

Although a PN arborizing in an individual glomerulus typically responds preferentially to a certain odorant (Figure 5), a single odor compound may activate a number of glomeruli (albeit with different intensities). Both multielectrode and optical imaging techniques have demonstrated that glomeruli in different regions of the AL respond preferentially to odorants of different chemical classes. Two structurally similar

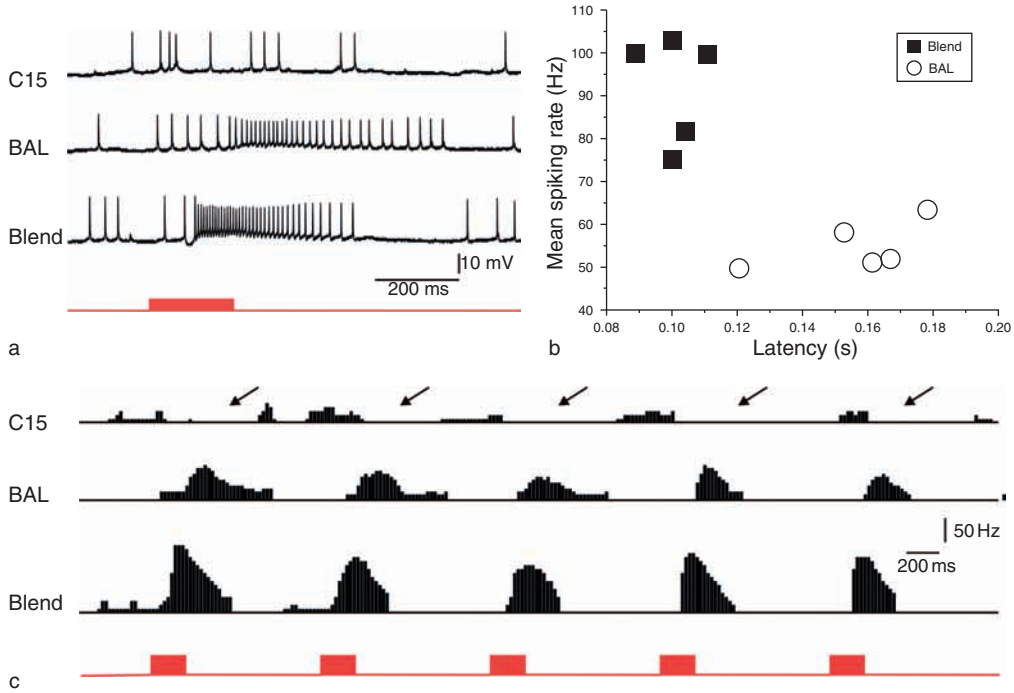

Figure 5 Synergistic effect of stimulation with a blend of the two key pheromone components on pheromone-responsive macroglomerular complex neurons. (a) Intracellular traces showing the responses of a toroid 1 projection neuron (PN) evoked by a chemical mimic ('C15') of the natural (*E,E,Z*)-10,12,14-hexadecatrienal component of the sex pheromone, by bombykal (BAL), and by a blend (mixture of C15 and BAL in a 1:1 ratio). The C15 stimulus decreased the spiking rate, whereas BAL and the blend markedly increased the spiking rate. Note that the blend-evoked excitatory response had a shorter delay and higher spiking rate than the BAL-evoked responses. The stimulus marker is indicated below the traces. (b) The same PN was tested in five trials using the blend and BAL as antennal stimuli. The response latency and mean spiking rate were quantified from all five trials, and the two variables are plotted against each other. The differential responses to these two stimuli resulted in apparent clustering in the plot, indicating a consistent synergistic effect (shorter delay and higher spiking rate) induced by the blend stimuli across trials. (c) Rate histograms showing the time course of the PN's responses to five repeated stimuli. Again, the C15 stimuli suppressed the spiking activity (arrows) of this PN, and the blend stimuli evoked stronger excitatory responses than did the BAL stimuli.

odorants may elicit spatially similar excitatory responses from an ensemble of AL neurons, but the odor-activated correlations elicited by the two odorants among the neurons in the ensemble can differ significantly. Thus, the identity of an odor stimulus is encoded by an ensemble of neurons through both its spatial and its temporal activity patterns. Moreover, experiments using multielectrode recording and behavioral assays have shown that there is a direct correlation between neural (in the AL) and behavioral discrimination of two odorants. In addition, the pattern of odor activation in the AL can be altered with learning. As a moth learns to associate a particular odorant with a food reward, additional AL neurons become responsive to that odor stimulus, and the activity of other neurons changes.

Sensitivity/Concentration Coding

Natural odors are released by different sources at variable rates. Furthermore, the amount and relative composition of components in a blend change as the animal approaches the source. In order to mediate appropriate odor-oriented behavior, the olfactory system must encode information about both the identity and the quantity of an odor. Both ORCs and intrinsic AL neurons solve this problem by increasing their spiking activity as stimulus concentration increases. In contrast to generalist ORCs, specialist ORCs and PNs have low response thresholds and dose–response curves that span several orders of magnitude of odor intensity. Importantly, their highest specificity occurs at relatively low stimulus concentrations, such as those encountered in nature.

Pheromone-responsive ORCs typically exhibit higher sensitivities than ORCs that respond to plant volatiles. CO_2-responsive ORCs have a response range from less than one order of magnitude below the ambient level (~370 ppm) to at least two orders of magnitude above it. The population dynamic range is expanded, moreover, due to the fact that individual neurons have different but overlapping response ranges. As a consequence of the high degree of convergence of ORCs on a single glomerulus, the response threshold of PNs is generally lower than that of

ORCs. The response thresholds of sexually isomorphic PNs in the female are lower than those of the corresponding male PNs, which is likely due to the higher number of plant-responsive ORCs in female antennas. Furthermore, the sensitivity of the AL circuitry is subject to circadian modulation, at least in part, by serotonin, such that AL neurons are most sensitive in the evening when the moths are most active.

Temporal Resolution/Inhibition

Odors in a plume downwind from its source exist in packets or filaments of odor-bearing air interspersed with clean air. Flying moths orienting toward an odor source thus are not confronted with a concentration gradient but, rather, with discrete odor pulses of 10–50 ms duration at a frequency of 1–20 s^{-1}. The olfactory system of moths (and most other flying insects) has evolved to recognize this temporal variation as a characteristic feature of the stimulus. Within a pheromone plume, male moths fly in a zigzagging pattern that aids in the localization of the pheromone source. This characteristic pattern of behavior is elicited only when the olfactory system is stimulated intermittently. Neurons throughout the olfactory pathway encode the natural spatiotemporal dynamics of the odor stimulus. Pheromone-specific ORCs respond better to short pulses of pheromone components (up to a frequency of at least 3/s) rather than to prolonged exposure, due to adaptation and desensitization, and many pheromone-responsive PNs track odor pulses up to 10–20 s^{-1} (Figure 4). In addition, both CO_2-responsive ORCs and PNs can track pulses accurately up to at least 10 s^{-1} and are excited by increases in CO_2 level and inhibited by decreases. The capability of following high-frequency CO_2 fluctuations might help moths to find nearby CO_2 sources (e.g., flowers). PNs that respond to plant volatiles can also follow short odor pulses with brief bursts of action potentials. Thus, the olfactory system of *Manduca* encodes an important characteristic of the natural odor signal – its intermittency.

Extensive studies on the MGC have revealed neuronal mechanisms that underlie the ability of the AL to encode the temporal qualities of an olfactory stimulus. It was first observed that the responses of some pheromone-specific PNs are temporally complex. For instance, the PNs with arborizations confined to the cumulus receive excitatory input from ORCs responsive to the pheromone component (*E,E,Z*)-10,12,14-hexadecatrienal. These PNs also receive inhibitory input from LNs that have arborizations in numerous glomeruli, including the cumulus and toroid 1, and are activated by bombykal. The converse is true for PNs with arborizations limited to toroid 1. The presentation of a blend of these two key components of the sex pheromone at their natural blend ratio elicits a triphasic response that consists of a brief hyperpolarization followed by a depolarization with a burst of action potentials and, finally, a delayed, slow, and long-lasting hyperpolarization (Figure 5(b)). The ability of these PNs to follow the frequency and duration of pulses is improved when they are stimulated with the correct blend (i.e., with the correct balance of excitation and inhibition). However, unnaturally prolonged stimulation is not faithfully encoded by MGC PNs. It should be noted, however, that a triphasic response is also produced in PNs of sexually isomorphic glomeruli (Figures 5(d) and 5(f)). In addition to improving their pulse-tracking capability, the excitation/inhibition balance is also believed to contribute to the synergistic responses of some MGC PNs when stimulated with the blend. For instance, some MGC PNs produce a response with a faster onset (i.e., shorter latency) and higher rate of spiking upon stimulation of the antenna with the pheromone blend than with the excitatory component alone (Figure 6).

Although pheromone-responsive neurons in the PC often exhibit prolonged depolarization (long-lasting excitation) in response to the pheromone blend, a subset of these neurons exhibit a state-dependent response to stimulation that may preserve some of the temporal structure of the stimulus. When stimulated while in a state of low, background firing, these neurons respond to the pheromone blend with excitation that can last up to 30 s, regardless of the duration of the stimulus. When stimulated during this excited state, the neurons reduce their firing rate for several seconds after the onset of the stimulus before gradually increasing it to the initial, higher level. Combined with neurons that produce only brief excitation in response to an odor, it seems that the PC neurons can encode the pulsatile nature of the pheromone plume, which is necessary to produce the characteristic upwind flight response of males to female sex pheromone.

Inhibitory interactions (mediated by GABA) appear to have additional roles in olfactory coding. For example, in the MGC, slow, inhibitory, interglomerular interactions promote synchronization of firing across glomerular outputs, which may enhance the representation of a complex olfactory stimulus (e.g., the pheromone blend) in downstream, higher-order brain centers. Spiking activity in pairs of PNs innervating the same glomerulus is more coherent when the neurons are activated by antennal stimulation with the pheromonal blend rather than with the single excitatory component. In this way, when two functionally related glomeruli (e.g., toroid 1 and cumulus) receive their odor input simultaneously (the sex

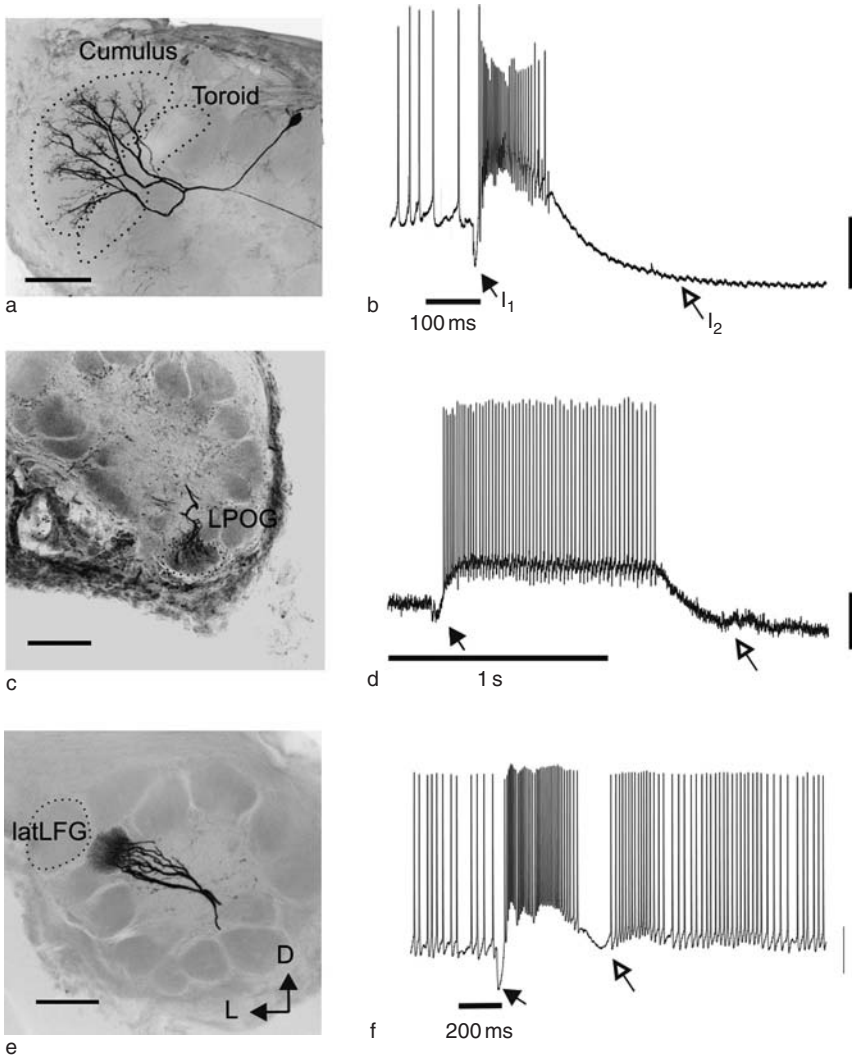

Figure 6 Examples of projection neurons (PNs) arborizing in three reference glomeruli with similar levels of specificity and temporal response dynamics. (a) A PN limited to the cumulus (dotted outline) of the macroglomerular complex (note that toroid 1 is also outlined in this image). (b) Intracellular recording from the PN in a responding to stimulation with the sex pheromone blend. Note the triphasic response consisting of an initial short hyperpolarization (synaptic inhibition, I_1), followed by a burst of action potentials and then a longer lasting hyperpolarization (I_2). (c) A PN from the sexually isomorphic labial palp pit organ glomerulus (LPOG) (dotted border). (d) Intracellular recording from a CO_2-responsive PN in the LPOG, demonstrating the triphasic response to stimulation with increased CO_2. (e) A PN from the sexually isomorphic glomerulus G35, which lies ventromedial to the lateral large female glomeruli (latLFG; dotted border). (f) Intracellular recording demonstrating the triphasic response of the PN in (e) to stimulation with *Z*-3-hexenyl acetate. D, dorsal; L, lateral. Scale bars = 100 μm. (d) From Guerenstein PG, Christensen TA, and Hildebrand JG (2004) Sensory processing of ambient-CO_2 information in the brain of the moth *Manduca Sexta*. *Journal of Comparative Physiology A* 190: 707–725.

pheromone blend), the temporal tuning of output from each glomerulus is enhanced by reciprocal inhibitory interactions. Interglomerular inhibitory interactions also occur among other glomeruli in the AL. In females, stimulation with the compound that excites latLFG PNs hyperpolarizes PNs in the adjacent glomerulus G35, but apparently not vice versa. This suggests that glomeruli that encode the components of a behaviorally relevant blend are organized in functionally related, synaptically interconnected clusters. In addition, the LNs in an activated glomerulus suppress the spontaneous activity in other glomeruli, thus providing contrast enhancement of the spatial pattern of activation elicited by an olfactory stimulus. In summary, inhibition serves multiple tasks, including shaping the temporal response characteristics of PNs, contrast enhancement, and enhancing the representation of behaviorally relevant blends.

Conclusions

Naturalistic experimental approaches are particularly useful in sensory neurobiology. The preferred stimulus for a neuron or a neural circuit can be more easily

selected from myriad possible stimuli on the basis of understanding what the animal is likely to experience in its natural environment. Every olfactory system has evolved under selective pressures that have been imposed by the characteristics of their olfactory environment. The olfactory systems of animals that specialize in a small number of food sources have been under different selective constraints than have the olfactory systems of generalist feeders. Specialists may have to be more sensitive to certain odorants, whereas generalists may have to be able to detect and discriminate a larger range of odors. Likewise, animals in a more dynamic olfactory environment, such as that encountered during flight, have to be able to encode the spatiotemporal patterns of odors such as those in odor plumes. Adult *Manduca* must detect discrete, intermittent packets of sex pheromone in order to locate potential mates, and they must evaluate complex blends of low-concentration odorants that vary with increasing distance from their source in order to locate food and appropriate oviposition sites. Neurobiological and neuroethological experimentation within the framework of the natural history of the animal fosters understanding of how the different features of an odor are encoded by the olfactory system within the context of the evolutionary constraints exerted upon particular species under study. This approach highlights the basic principles and mechanisms by which olfactory systems encode information about the identity, intensity, and intermittency of natural odors and is essential for understanding the functional basis by which odors are sensed and perceived.

See also: Olfaction in Invertebrates: *Drosophila*; Olfaction in Invertebrates: Honeybee; Olfactory Receptors; Olfactory Cortex Physiology; Olfactory System Theory; Olfactory System: Circuit Dynamics and Neural Coding in the Locust.

Further Reading

Christensen TA and Hildebrand JG (2002) Pheromonal and host-odor processing in the insect antennal lobe: How different? *Current Opinion in Neurobiology* 12: 393–399.

Christensen TA, Waldrop BR, and Hildebrand JG (1998) Multitasking in the olfactory system: Context-dependent responses to odors reveal dual GABA-regulated coding mechanisms in single olfactory projection neurons. *Journal of Neuroscience* 18: 5999–6008.

Hansson BS and Christensen TA (1999) Functional characteristics of the antennal lobe. In: Hansson BS (ed.) *Insect Olfaction*, pp. 126–164. Berlin: Springer.

Hildebrand JG (1995) Analysis of chemical signals by nervous systems. *Proceedings of the National Academy of Sciences of the United States of America* 92: 67–74.

Hildebrand JG (1996) King Solomon Lecture – Olfactory control of behavior in moths: Central processing of odor information and the functional significance of olfactory glomeruli. *Journal of Comparative Physiology A* 178: 5–19.

Hildebrand JG and Shepherd GM (1997) Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. *Annual Review of Neuroscience* 20: 595–631.

Homberg U, Montague RA, and Hildebrand JG (1988) Anatomy of antenno-cerebral pathways in the brain of the sphinx moth *Manduca sexta*. *Cell and Tissue Research* 254: 255–281.

Oland LA and Tolbert LP (1996) Multiple factors shape development of olfactory glomeruli: Insights from an insect model system. *Journal of Neurobiology* 30: 92–109.

Strausfeld NJ and Hildebrand JG (1999) Olfactory systems: Common design, uncommon origins? *Current Opinion in Neurobiology* 9: 634–639.

Olfactory Bulb Anatomy

T Kosaka and K Kosaka, Kyushu University, Fukuoka, Japan

© 2009 Elsevier Ltd. All rights reserved.

Laminar Organization of the Main Olfactory Bulb

The MOB of most mammals has a roughly concentric arrangement of seven layers, which are, from the surface (Figures 1–3), (1) the olfactory nerve layer (ONL), (2) glomerular layer (GL), (3) external plexiform layer (EPL), (4) mitral cell layer (MCL), (5) internal plexiform layer (IPL), (6) granule cell layer (GCL), and (7) subependymal–ependymal layer (SEL). In some species, the lateral ventricle extends into the olfactory bulb.

Fascicles of unmyelinated olfactory axons from the sensory epithelium penetrate the cribriform plate as fila olfactoria, reach the MOB from the ventromedial aspect, and comprise the outermost layer of the MOB, the ONL. In the ONL, bundles of olfactory nerves (ONs) are enclosed by ensheathing glia and penetrate into glomeruli in the GL. Glomeruli are one of the most distinctive structures in the brain, analogous to barrels or columns in the cerebral cortex. In the mammalian MOB, they are compactly packed almost two-dimensionally near the MOB surface, forming the GL. Each glomerulus is composed of neuropil-rich spheroidal structures surrounded by a distinctive shell of small to medium-sized neurons and glial cells (Figures 1–4). Within each glomerulus, ON axon terminals make contact with dendritic tufts of various types of neurons, which include both projection neurons (mitral and tufted cells) and intrinsic local circuit neurons (Figures 2 and 3).

Neurons in the periglomerular region (the region around and subjacent to glomeruli) are apparently heterogeneous but occasionally collectively referred to as 'juxtaglomerular neurons,' which are morphologically classified into three different types of neurons: periglomerular cells, superficial short-axon cells, and external (or superficial) tufted cells. Among these neurons, periglomerular cells and external tufted cells send dendrites into glomeruli, whereas superficial short-axon cells are believed to send dendrites into the periglomerular region but not into glomeruli. Somata of tufted cells and some types of so-called short-axon cells are located in the EPL. Secondary dendrites of mitral cells as well as tufted cells make dendrodendritic synapses in the EPL, including reciprocal synapses with granule cell dendrites and short-axon cell dendrites (Figures 5 and 6). The MCL is the thin layer where large somata of mitral cells and smaller somata of some other local circuit neurons are located (Figure 7). The IPL is a relatively cell-free thin layer. In the GCL, many clusters or rows of granule cell somata are compactly packed, among which a few larger somata of deep short-axon cells are scattered. The core region of the MOB is the SEL, which is a part of the pathway of the rostral migratory stream from the subventricular zone of the lateral ventricle.

Neuron Types and Synaptic Organization, Chemical Synapses, and Gap Junctions

As in other brain regions, there are two types of neuron groups in the MOB – the principal neurons and local circuit neurons. Both are heterogeneous in various aspects and summarized in Table 1 and Figures 2 and 3. It is important to note that recent studies, mainly immunocytochemical studies, have revealed various types of chemically defined neurons in MOBs of various mammals, many of which appear not to correspond well to any of the neuron types described previously in classical Golgi studies and do not yet have specific names but are collectively called short-axon cells (Figure 2(c)). Thus, future studies will classify these short-axon cells into several subtypes with characteristic structural and physiological features.

In considering the neuronal organization of the MOB, it is important to note one of the characteristic structural features of olfactory bulb neurons; that is, except for some short-axon cells, somata and dendrites of most neurons are not only postsynaptic but also presynaptic, making dendrodendritic synapses including reciprocal pairs or reciprocal synapses (Figures 5 and 6).

Principal Neurons

In the MOB, the principal neurons are mitral cells and tufted cells, most of which are considered to project to extrabulbar brain regions, although some tufted cells are considered to be intrabulbar association neurons rather than projection neurons. Mitral cells and tufted cells generally send their single primary dendrites into single glomeruli making characteristic tufts and several secondary dendrites in the EPL, although some external tufted cells lack secondary dendrites (Figures 2 and 3). Mitral and tufted cells show marked differences in the dendritic morphologies which tend to be correlated with the laminar

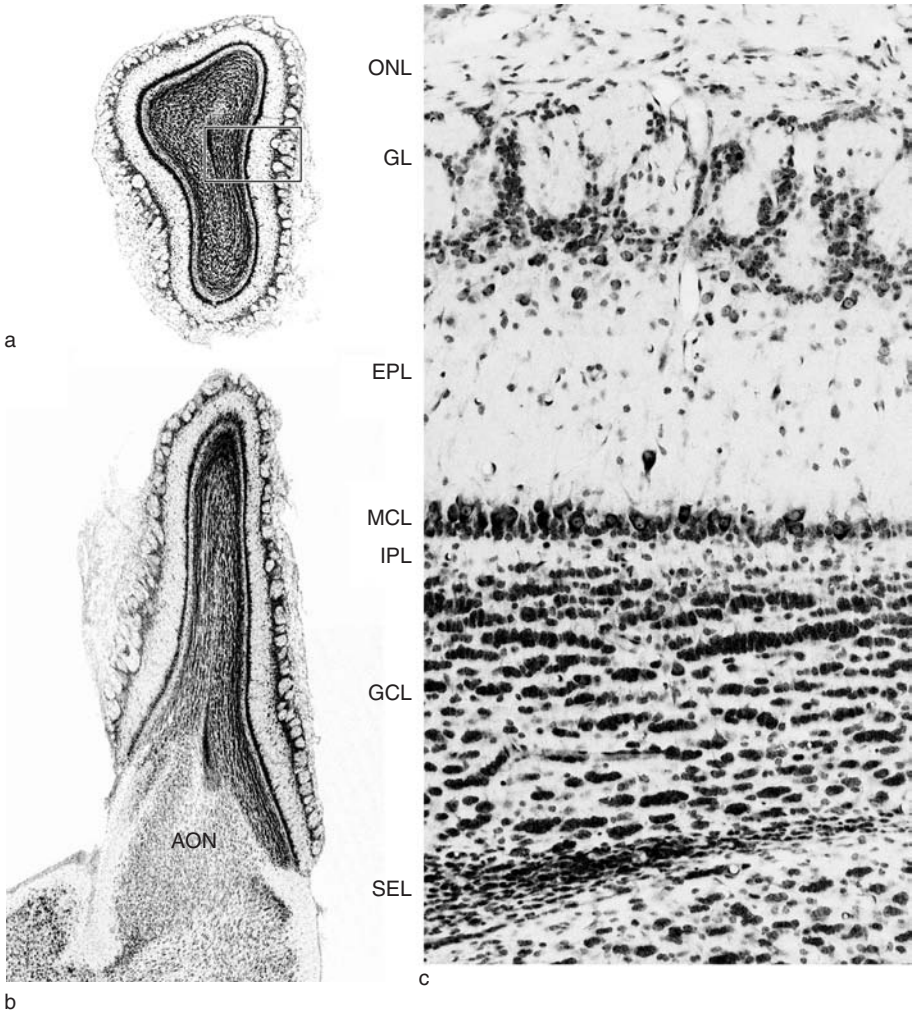

Figure 1 Confocal laser scanning microscope images of propidium iodide-stained 50 μm-thick sections of mouse olfactory bulbs. Medial side is right in (a) and (b). (a) Coronal section; projection image. (b) Horizontal section; projection image. (c) High magnification part is outlined with rectangle; single optical image. For clearer visualization, black and white are inverted. AON, anterior olfactory nucleus; EPL, external plexiform layer; GCL, granule cell layer; GL, glomerular layer; IPL, internal plexiform layer; MCL, mitral cell layer; ONL, olfactory nerve layer; SEL, subependymal–ependymal layer.

position of their somata. In addition, there are also differences in their interaction with different types of granule cells as well as in their axonal projections. Thus, as shown in Figure 3 and Table 1, mitral cells are classified into two subtypes, type 1 (M1) and type 2 (M2), based on the laminar distribution of their secondary dendrites in the EPL. M1 cells extend their secondary dendrites mainly in the inner sublamina of the EPL, whereas M2 cells extend their secondary dendrites mainly in the intermediate to outer sublamina. Tufted cells are classified into three subtypes – deep, middle, and external tufted cells – based on the location of their somata. In addition, mitral cells and tufted cells have also been shown to be chemically and physiologically different. Thus, the principal neurons of the MOB, mitral and tufted cells, are apparently heterogeneous in their structural, chemical, and physiological features. Furthermore, external tufted cells have been revealed to be heterogeneous in their axonal projections and are classified into three types: conventional projection-type external tufted cells, external tufted cells with the associational pathway between the opposing region of the medial and lateral sides of the MOB, and the local, interneuron-like (superficial) tufted cells. Thus, even the classical neuron type 'external tufted cell' is never regarded as a single type of neuron but consists of apparently diverse neurons. Among these external tufted cells, those with the intrabulbar associational system may be interesting because they send a topographically organized projection to the IPL of the opposite side of the same bulb, making a

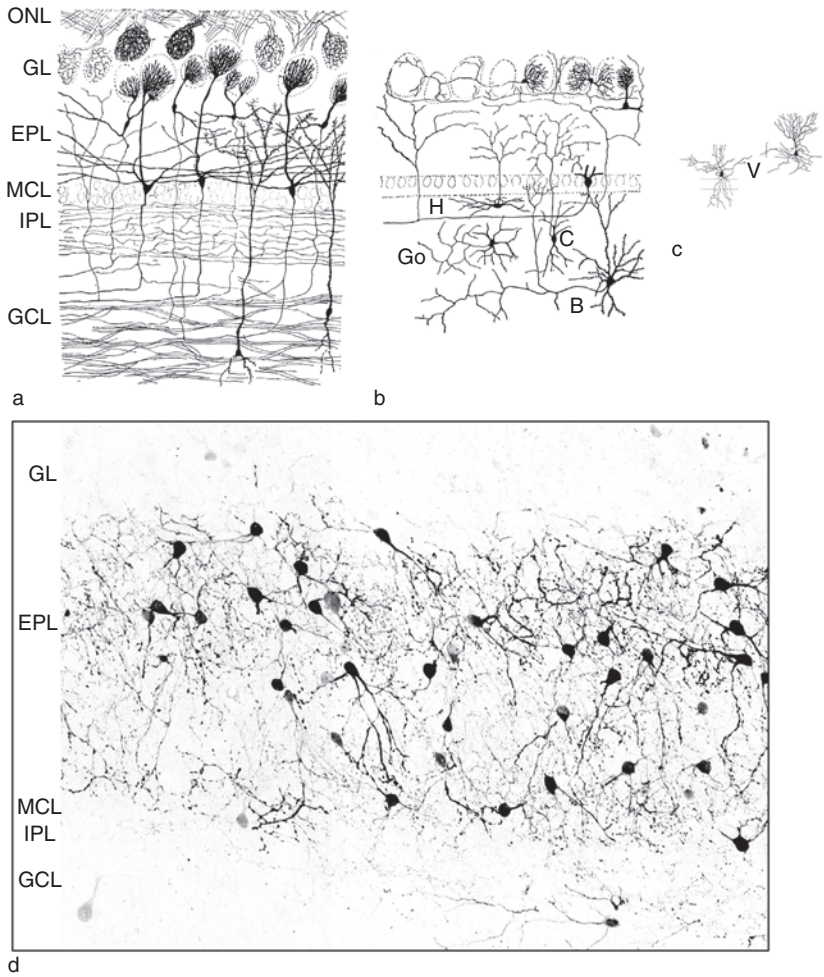

Figure 2 (a–c) Drawing of Golgi impregnated neurons. (a) Olfactory nerves, mitral cells, tufted cells, and granule cells. Modified from figure 412 of Ramón y Cajal S (1911) *Histologie du Système Nerveux de l'homme et des Vertébrés. Tome II* (translated by L Azoulay from Spanish to French). Paris, Maloine (republished 1952 by Consejo Superior de Investigaciones Cientificas, Madrid). (b) Neurons in the periglomerular region and deep short-axon cells. Modified from figure 415 of Ramón y Cajal S (1911) *Histologie du Système Nerveux de l'homme et des Vertébrés. Tome II* (translated by L Azoulay from Spanish to French). Paris, Maloine (republished 1952 by Consejo Superior de Investigaciones Cientificas, Madrid). B, Blanes cell; C, Cajal cell; Go, Golgi cell; H, horizontal cell; (c) Van Gehuchten cells (V). Cells 26 and 27 from Van Gehuchten A and Martin I (1891) Le bulbe olfactif chez quelques mammiféres. *La Cellule* 7: 204–237. (d) parvalbumin immunoreactive cells in the mouse main olfactory bulb. Various types of short-axon cells are seen. EPL, external plexiform layer; GCL, granule cell layer; GL, glomerular layer; IPL, internal plexiform layer; MCL, mitral cell layer; ONL, olfactory nerve layer.

point-to-point reciprocal connection between isofunctional glomeruli.

Local Circuit Neurons

In contrast with most principal neurons, local circuit neurons are considered to extend their axons within the olfactory bulb, or some are axonless. There are at least three types of local circuit neurons in the MOB – granule cells, periglomerular cells, and short-axon cells – although each type is apparently heterogeneous and further consists of some subtypes. In addition, there appear to be some uncharacterized or unidentified types of presumed local circuit neurons.

Periglomerular cells and granule cells are considered to be inhibitory neurons and make synapses with mitral/tufted cell dendritic tufts in glomeruli and secondary dendrites in the EPL, respectively (intraglomerular dendritic processes of mitral and tufted cells and those in the EPL are considered to be indistinguishable by electron microscope and thus are usually termed collectively as mitral/tufted cell dendrites). Some short-axon cells are also considered to be inhibitory neurons, but most of them remain to be analyzed.

Granule cells Granule cells are axonless and send peripheral dendrites into the EPL, where they bear

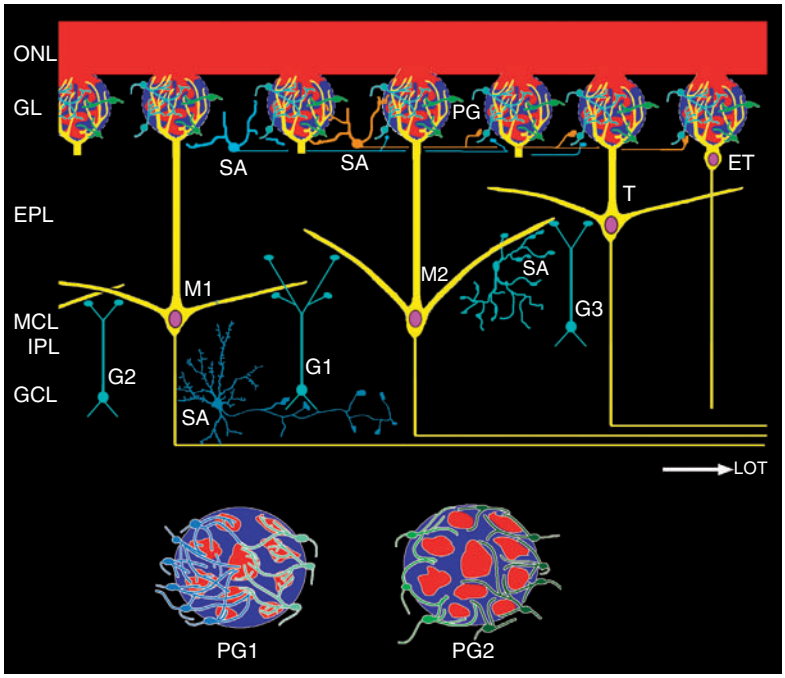

Figure 3 Schemes of the neuronal organization of rodent main olfactory bulb showing neuron types and their interaction. G1–G3, types 1–3 granule cells; M1 and M2, type 1 and type 2 mitral cells; PG1 and PG2, type 1 and type 2 periglomerular cells; EPL, external plexiform layer; GCL, granule cell layer; GL, glomerular layer; IPL, internal plexiform layer; MCL, mitral cell layer; ONL, olfactory nerve layer; ET, external tufted cell; T, tufted cell; SA, short-axon cell.

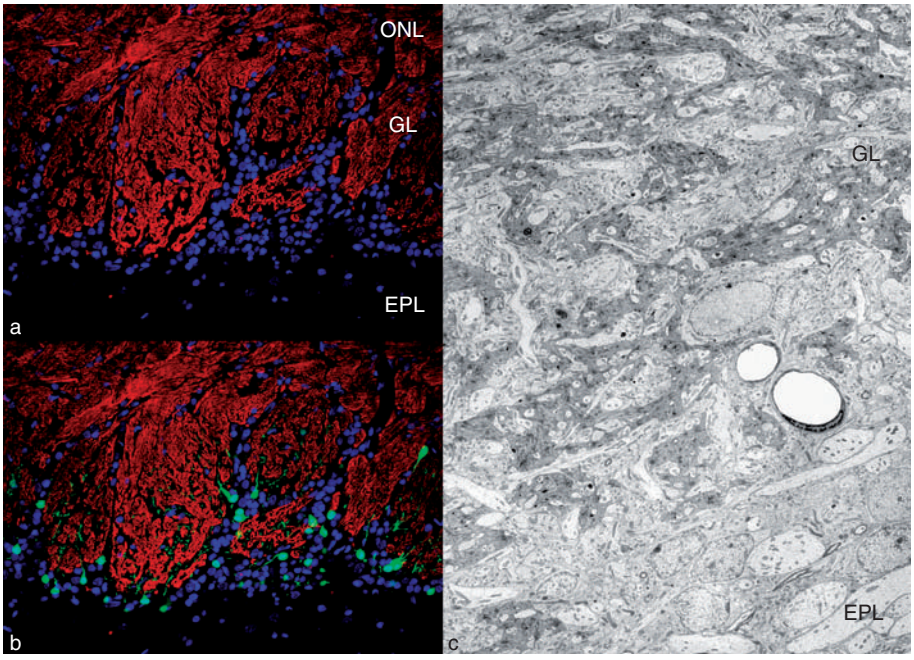

Figure 4 (a, b) Pseudo-colored confocal laser scanning microscope images of glomeruli in multiple immunostained 50 μm-thick section of rat main olfactory bulb. (a) The compartments of glomeruli, olfactory marker protein (OMP)-positive ON zone and OMP-negative non-ON zone. (b) Calbindin D28K-positive type 2 periglomerular cells extending dendrites in the non-ON zone. (c) Low-magnification electron micrograph of the glomerulus and the superficial part of the external plexiform layer (EPL). GL, glomerular layer; ONL, olfactory nerve layer.

large spine-like protrusions, gemmules, and central processes with spines into the GCL. Gemmules contact mainly secondary dendrites of mitral/tufted cells and make reciprocal pairs of synapses (Figures 5 and 6). Based on the location of their somata and the distribution of their peripheral dendrites in the EPL, granule cells are proposed to be classified into three subtypes (Figure 3): type 1 (G1) cells extend their

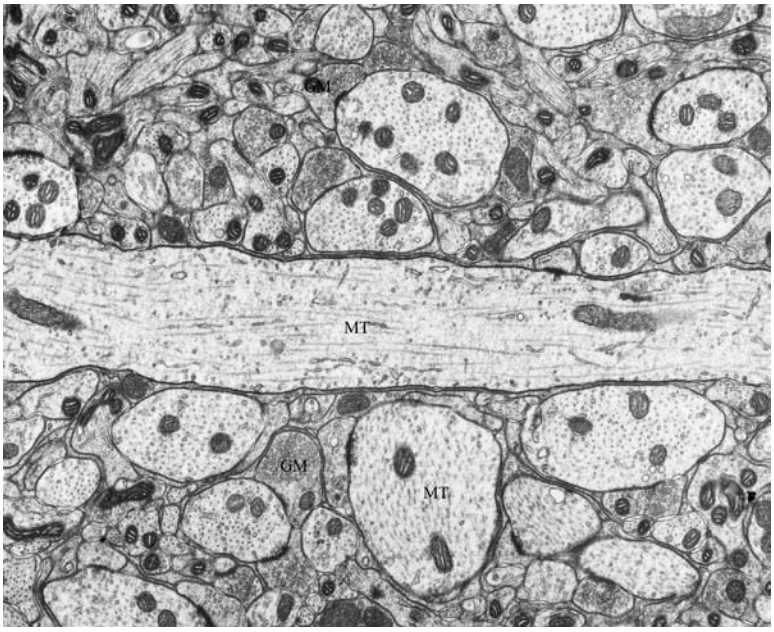

Figure 5 Electron micrograph of the external plexiform layer of the rat main olfactory bulb. GM, granule cell gemmule; MT, mitral/tufted cell.

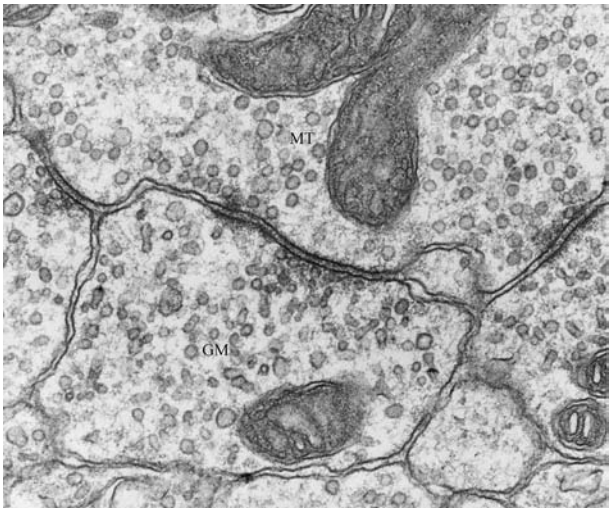

Figure 6 Electron micrograph of a reciprocal synapse between a mitral/tufted cell dendrite (MT) and granule cell gemmule (GM) in the external plexiform layer of the mouse main olfactory bulb.

dendrites throughout the EPL, type 2 (G2) in the deep half of the EPL, and type 3 (G3) in the superficial half of the EPL. Somata of type 2 and type 3 granule cells are mainly located in the deep and superficial parts of the GCL and are also termed deep and superficial granule cells, respectively. Most granule cells are believed to be GABAergic. In the rat MOB, some granule cells (approximately 10% of GABAergic granule cells) are enkephalin-like immunoreactive, located mainly in the deeper half of the GCL; consequently, the deeper half of the EPL is more intensely enkephalin-like immunoreactive than the superficial half of the EPL. Some other granule cells are calretinin immunoreactive, although calretinin is also contained in various neuronal elements in the MOB, such as ONs, subpopulations of mitral cells, periglomerular cells, and short-axon cells.

Periglomerular cells Periglomerular cells are located in the periglomerular region and send their dendrites into glomeruli and extend axons laterally. In glomeruli, their dendrites intermingle with mitral/tufted cell dendritic tufts (Figure 8(g)) and ONs. They have been considered to receive synapses from the ONs and make synapses, including reciprocal synapses, with mitral/tufted cell dendrites (Figure 8). However, some periglomerular cells have been reported to receive no or few synapses from the ONs. Thus, there appear to be at least two different types of periglomerular cells, which are closely related to the compartmental organization of the glomerulus (Figures 3, 4, and 8). Within the glomerulus, ONs do not spread diffusely but, rather, they cluster to make many irregular tortuous strands or islands with many small holes (Figure 4). At the electron microscopic (EM) level, clusters of electron-dense ONs are intermingled with electron-lucent dendrites. Some electron-lucent dendrites appear to penetrate into clusters of ONs, corresponding to

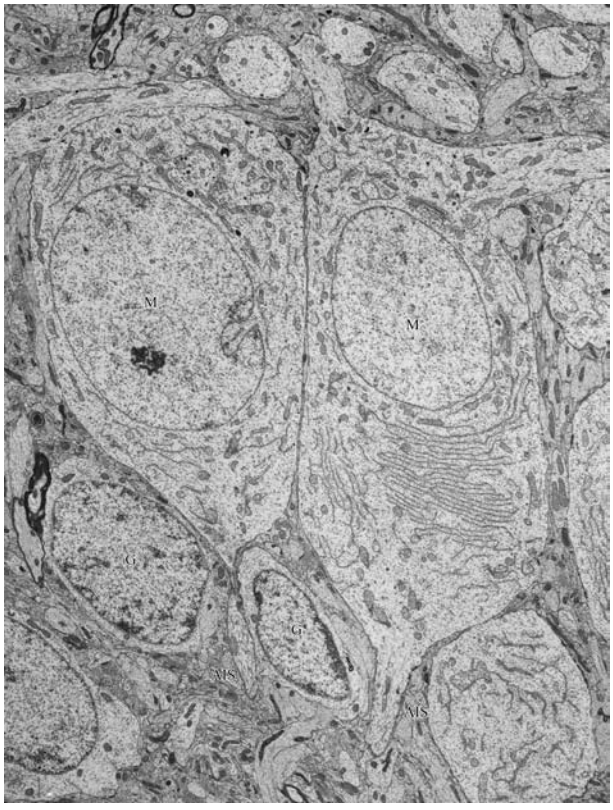

Figure 7 Electron micrograph of the mitral cell layer of the rat main olfactory bulb. AIS, axon initial segment; G, granule cell; M, mitral cell.

small holes in the ON strands (Figures 4 and 8). Thus, it is now considered that each glomerulus consists of two compartments or zones, the ON zone (or axonal or sensory subcompartment) and the non-ON zone (or dendritic or central subcompartment). The ON zone is mainly occupied by ON preterminals and terminals as well as fine dendritic processes of intrinsic neurons, indicating that the ON zone is the site where olfactory receptor cells make synapses on their targets. The non-ON zone is occupied mainly by dendritic processes of intrinsic neurons, where dendrodendritic interactions between intrinsic neurons occur. However, as described previously, many ON negative patches in the ON zone are continuous with the area of the non-ON zone (Figures 4 and 8). Thus, it is difficult to strictly define the border of the ON and non-ON zones.

Immunocytochemical studies at the light microscopic and EM levels have revealed that one group of periglomerular cells extend their dendrites in both the ON and non-ON zones and receive synapses from the ONs, whereas another group of cells restrict their dendrites in the non-ON zone as if to avoid the ONs and receive no or few synapses from the ON terminals (Figures 3 and 4). The former group of periglomerular cells correspond to the traditionally characterized periglomerular cells which are called type 1 periglomerular cells (PG1), whereas the latter group of periglomerular cells are newly recognized and called type 2 periglomerular cells (PG2). The differences in dendritic projection of the two groups of PG cells, especially their prominent differences in the synaptic inputs from the ON terminals, indicate their possible functional differences in the local circuits, which remain to be identified in future physiological studies. These two structurally different groups are also different in their chemical properties. In the rat MOB, PG1 cells include neurons immunoreactive for GABA, its synthesizing enzyme glutamic acid decarboxylase (GAD), and dopamine-synthesizing enzyme tyrosine hydroxylase (TH). PG2 cells include those immunoreactive for calbindin and calretinin. It is interesting to note that anatomically revealed direct contacts between ON terminals and PG1 cell dendrites correspond well with the results obtained by physiological studies, although the direction of the structurally identified synapses is reverse in relation to the function; dopamine and GABA presumably spilled over from PG1 cell dendrites contacting ON terminals, regulating the transmitter release from the ON terminals via D2 and GABA-B receptors on the ON terminals. However, we must be careful not to oversimplify the classification of neurons. For example, even in the rat MOB, we cannot regard all GABA-positive cells as PG1 cells *a priori*. We must consider the possibility that GABA-positive cells are heterogeneous and some are PG1 cells but others are PG2 cells. Furthermore, because there are many uncharacterized cells in the periglomerular region, there must be GABA-negative PG1 cells and calbindin negative/calretinin-negative PG2 cells. It is also noteworthy that in some other animals, such as laboratory musk shrews, almost all calbindin-positive cells are also GABA positive, but they are apparently PG2 cells. Thus, there are both type 1 and type 2 GABA-LIR and/or GAD-IR periglomerular cells in the MOBs of some mammals.

Short-axon cells In the MOB, mitral/tufted cells, granule cells, and periglomerular cells are major groups of neurons, but there are other types of neurons which are collectively named short-axon cells. Short-axon cells are distributed throughout the layers – that is, in the GL, EPL, ML, IPL, and GCL – and include diverse types of neurons, only some of which have been identified by classical Golgi studies and reported as distinct types of local circuit neurons (Figures 2(b) and 2(c); Table 1). Immunocytochemical studies have identified several chemical markers that characterize diverse types of short-axon cells, although most of these chemical markers are not

Table 1 Neuron types in the main olfactory bulb

	Neuron types	*Subtypes*	*Sub-subtypes*	*Neuroactive substances*	*Other chemical markers*
Principal neurons	Mitral cell (M)	Type 1 (M1)		Glutamate, CRF	Calretinin
		Type 2 (M2)			
	Tufted cell (T)	Deep tufted cell (DT)		Glutamate, CRF	
		Middle tufted cell (MT)		Glutamate, CRF	
		External tufted cell (ET)	Projection	Glutamate	
			Association	Glutamate, CCK	
			Local circuit	Glutamate?	
Local circuit neurons (interneurons)	Periglomerular cell (PG)	Type 1 (PG1)		GABA, GABA + DA, somatostatin, CCK, NO (nitric oxide synthase, NADPH diaphorase)	
		Type 2 (PG2)		Enkephalin? (GABA in some species)	Calbindin, calretinin
	Granule cell (GR)	Type 1 (G1)		GABA, enkephalin	Calretinin
		Type 2 (G2)			
		Type 3 (G3)			
	Short-axon cell (SA)	Deep short-axon cell (DSA)	Blanes cell (B)	GABA	
			Golgi cell (Go)	?	Parvalbumin
			Cajal cell (C)	?	
			Horizontal cell (H)	?	
			Unclassified	GABA, VIP, NPY, NO	Parvalbumin, calbindin, calretinin
		Superficial short-axon cell (SSA)	Van Gehuchten cell (V)	GABA, VIP	Parvalbumin
			Unclassified	GABA? glutamate? NO	Parvalbumin calbindin, calretinin
Afferent nerves	Olfactory nerves (ON)			Glutamate	Olfactory marker protein, calretinin, carnosine
	Centrifugal afferents	Higher olfactory cortices		Glutamate?	
		Nucleus of the diagonal band		ACh, GABA	
		Locus coeruleus		Noradrenaline	
		Raphe nuclei		Serotonin	

ACh, acetylcholine; CCK, creatine Kinase; CRF, corticotropin-releasing factor; DA, dopamine; NPY, neuropeptide Y; VIP, vasoactive intestinal polypeptide.

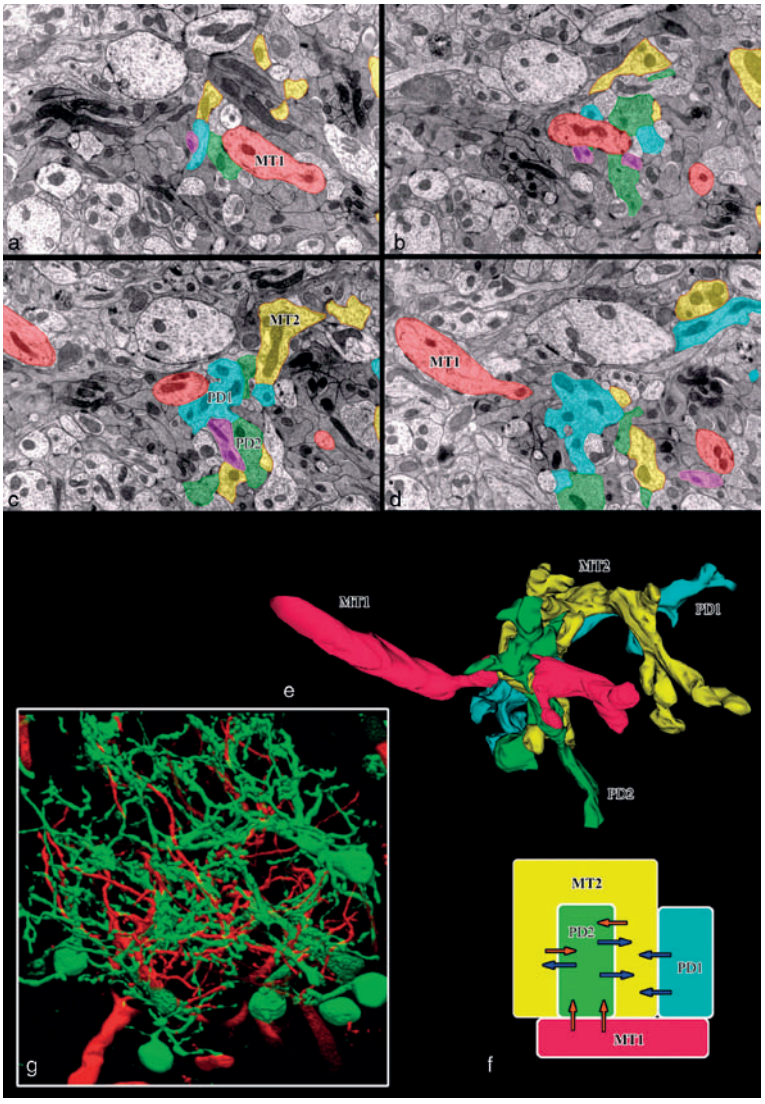

Figure 8 (a–d) Electron micrographs of part of a glomerulus from serial sections. Color-labeled processes, two mitral/tufted cell dendrites (MT1 and 2), and two type 1 periglomerular cell dendrites (PD1 and 2) are reconstructed in (e). Synaptic connections among these four processes are shown in (f), in which synapses from ONs are not shown. (g) Volume-rendered three-dimensional confocal laser scanning microscope image showing the tracer (BDA)-labeled mitral cell dendrite (red) and CB-IR type 2 periglomerular cells (green), which are intermingled with each other.

necessarily specific to short-axon cells but, rather, are contained in other types of neurons; these chemical markers include calcium-binding proteins such as parvalbumin (Figure 2(d)), calretinin, calbindin, and neurocalcin and neuroactive substances such as vasoactive intestinal polypeptide (VIP), neuropeptide Y, somatostatin, and nitric oxide (nitric oxide synthase and NADPH diaphorase). Among these, some parvalbumin-positive short-axon cells in the EPL are characterized at the EM level and have been confirmed to make reciprocal synapses with mitral/tufted cell dendrites, suggesting that they are inhibitory local circuit neurons. On the other hand, some VIP-positive deep short-axon cells are also reported to be GABAergic and make symmetrical synapses on other local circuit neurons. Furthermore, a type of deep short-axon cell, Blanes cell, is also a GABAergic inhibitory neuron mediating persistent feed-forward inhibition onto granule cells. As described previously, short-axon cells are usually inhibitory local circuit neurons. However, 'superficial short-axon cells' have been reported to send interglomerular axons over long distances to form excitatory synapses with inhibitory periglomerular cells up to 20–30 glomeruli away, resulting in an interglomerular center–surround inhibitory network. These excitatory superficial short-axon cells are reported to have poorly branched dendrites but appear to

display intraglomerular dendrites or small tufts and thus may correspond to local, interneuron-like (superficial) tufted cells described previously. Whether these two groups of neurons correspond to each other remains to be determined. Most short-axon cells remain to be analyzed in detail. Furthermore, it remains to be determined whether so-called short-axon cells are really nonprojecting local circuit neurons or whether at least some of them project out of the OB. Compared to the neurons in the retina, in which diverse subtypes of each neuron type are characterized in detail, neuronal subtypes in the MOB remain to be identified, characterized in detail, and classified.

Chemical Synapses and Gap Junctions

Neuronal interactions are mediated in diverse ways. Chemical synapses and gap junctions are two major pathways through which neurons send signals to and receive signals from other neurons (Figures 9 and 10). A characteristic feature of chemical synapses in the MOB is the dominance of dendrodendritic reciprocal pairs of asymmetrical and symmetrical synapses between mitral/tufted cells and periglomerular/granule cells in the glomeruli and the EPL, respectively. Gap junctions are frequently observed in the MOB, especially in glomeruli, in which gap junctions occur between mitral/tufted cell dendrites and between mitral/tufted cell dendrites and dendrites of periglomerular cells and some uncharacterized local circuit neurons. In extraglomerular regions, the mitral/tufted cell somata, dendrites, and axon hillocks make gap junctions and mixed synapses with various types of interneuronal processes, including apparent granule cell dendritic gemmules and some types of sheetlike or calyx-like processes contacting the somata or large dendrites of mitral/tufted cells. The major chemical synaptic connections as well as gap junctions among neuronal elements in the rodent MOB are summarized in Figure 10.

Figure 9 Electron micrograph of the glomerulus of the mouse main olfactory bulb. A presumed mitral/tufted cell dendrite receives an asymmetrical synapse from an ON terminal and makes a gap junction (arrow; inset) with a presumed periglomerular cell dendrite containing vesicles.

Centrifugal Afferents

MOBs receive dense centrifugal inputs from diverse sources which are considered to be important for integration and modulation of the information. The centrifugal afferents are classified into two groups: afferents from olfactory-related structures and modulatory afferents from nonolfactory subcortical structures. Olfactory-related structures sending afferents to the MOB usually receive inputs from the MOB, making the reciprocal connections. On the other hand, there are three major sources of modulatory afferents from nonolfactory subcortical structures: the nucleus of the diagonal band, raphe nuclei, and locus coeruleus.

Afferents from the nucleus of the diagonal band contain both cholinergic and GABAergic nerves. Cholinergic nerves originate mainly from the Ch3 cell groups and distribute in the GCL, IPL, the deeper one-half to one-third of the EPL, and the GL of the MOB. Retrograde labeling studies have identified GABAergic neurons in the nucleus of the diagonal band projecting to the MOB, but their distribution in the MOB remains to be determined due to difficulty in discriminating the extrinsic GABAergic elements from the numerous intrinsic ones.

Raphe nuclei and locus coeruleus send serotonergic and noradrenergic fibers to the MOB, respectively. Serotonergic fibers distribute in the glomeruli, IPL, and GCL, whereas noradrenergic fibers distribute mainly in the IPL and GCL.

Structure of Nonrodent Mammalian Olfactory Bulbs, with Reference to the Organization of the Insectivore Olfactory Bulb

The major cell types and the clear lamination of the MOB are generally considered constant throughout mammals, and thus it is assumed that there is a considerable amount of similarity in the MOBs of mammals. However, the MOBs of laboratory musk shrews and moles, two insectivores, have a neuronal organization that is different from the well-known general organization of the rodent MOBs (Figure 11).

The compartmental organization of glomeruli and the existence of two types of periglomerular cells

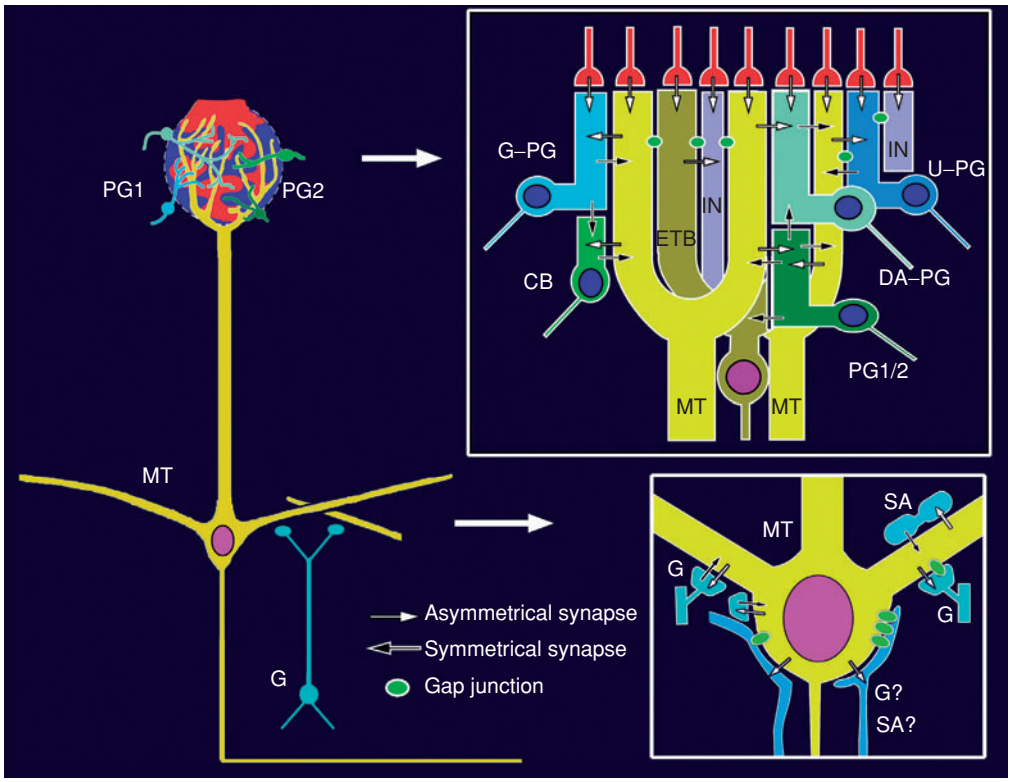

Figure 10 Schemes of the neuronal organization in the rat main olfactory bulb. Summary of neuronal connections via chemical synapses and gap junctions in the glomerulus and in the extraglomerular region. In the glomerulus, synaptic connections among olfactory nerves (ONs) and dendrites of various types of neurons are shown. In the glomerulus, dendritic tufts of mitral/tufted (MT) cells and type 1 periglomerular cells (PG1) receive synapses from olfactory nerves (ONs), whereas type 2 periglomerular cells (PG2) do not receive synapses from ONs. GABA/GAD-positive cell dendrites occasionally make reciprocal pairs of synapses with MT cell dendrites. TH-positive cells also receive asymmetrical synapses from MT cells and make symmetrical synapses to MT cells. However, no TH-positive dendrites making reciprocal synapses have been encountered. CB-positive type 2 PG cells receive no or rare synapses from ONs but make many synapses, including reciprocal pairs, with MT cell dendrites. On the other hand, GABA/GAD-positive, TH-positive, and CB-positive PG cell dendrites receive symmetrical synapses from other PG cell dendrites (PG1 and-2), at least some of which are GABA/GAD positive. Gap junctions are seen between MT cell (including type B external tufted cell) dendrites, between MT cell dendrites and type 1 PG cell dendrites with as yet uncharacterized properties (U-PG), and between MT cell dendrites and dendrites of some interneurons (IN) different from PG cells. Type B external tufted cells are a type of external tufted cells which show cytological features slightly different from those of typical MT cells and show no reciprocal synapses with PG cells, even though they make asymmetrical synapses to PG cells and other interneurons and receive symmetrical synapses from PG cells as well as asymmetrical synapses from ON terminals. In the extraglomerular region MT cell somata and dendrites make chemical synapses including reciprocal pairs of synapses with granule cell gemmules and processes of other types of interneurons, such as Van Gehuchten cells. Gap junctions and mixed synapses are also seen between MT cell somata and dendrites and various processes of presumed interneuron origins. CB, calbindin or calretinin-positive type 2 periglomerular cell; DA-PG, tyrosine hydroxylase and GABA-positive type 1 periglomerular cell; ETB, type B external tufted cell, which shows cytological and synaptic features somewhat different from usually reported mitral/tufted cells; G-PG, GABA and/or glutamic acid decarboxylase-positive type 1 periglomerular cell; MT, mitral/tufted cell.

are confirmed even in the MOBs of these two insectivores. However, their MOBs contain a unique spheroidal synaptic region, the nidus, obviously different from the overlying glomerulus and never observed in the MOBs of rodents. In addition, there are at least two types of neuron specifically associated with the nidus, namely tasseled cells and perinidal cells. Tasseled cells are a type of projection neuron with small tuftlike dendritic branches, whereas perinidal cells are mostly a type of GABAergic interneuron which appears to be related to PG2 cells. In the laboratory musk shrew and mole MOBs, there are at least two separate projection neuron groups: the mitral/tufted cell group receives numerous direct inputs from ONs, and the tasseled cell group receives no or rare direct inputs from them (Figure 11). These two projection systems are believed to interact with each other via several types of GABAergic (and also presumably GABA-negative) local circuit eurons at multiple sites, including the nidus and the EPL. Thus, tasseled cells and mitral/tufted cells should be regarded as interacting at several different sites via different types of GABAergic interneurons (and presumably via other interneurons, such as GABA-negative,

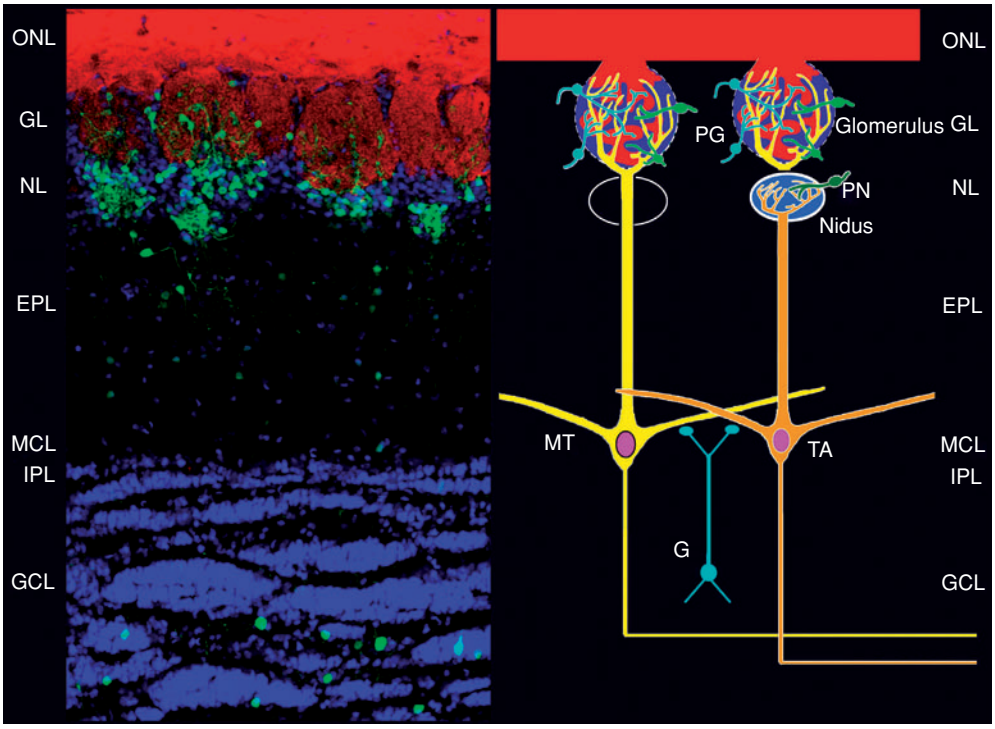

Figure 11 Pseudo-colored confocal laser scanning microscope image of the shrew main olfactory bulb (MOB) (right) and scheme showing its organization (left). Double-immunostained for olfactory marker protein (red) and calbindin D28K (green) combined with the DAPI nuclear staining (blue). In the shrew MOB, there are at least two separate projection systems – the mitral/tufted (MT) cell system and the tasseled (TA) cell system. The MT cell system in the shrew MOB resembles that of the rat MOB, consisting of MT cells, type 1 and 2 periglomerular cells (PG), and granule cells (G). On the other hand, the TA cell system in the shrew MOB consists of TA cells and perinidal cells (PN). In contrast with MT cells receiving numerous direct inputs from the olfactory nerves (ONs), TA cells extend their dendritic tufts in the nidus instead of glomerulus and receive no or rare direct inputs from the ONs. EPL, external plexiform layer; GCL, granule cell layer; GL, glomerular layer; IPL, internal plexiform layer; MCL, mitral cell layer; NL, nidal layer; ONL, olfactory nerve layer.

calretinin-immunoreactive perinidal cells) and thus assumed to mutually inhibit each other, intensifying or focusing the centrally transmitted information.

See also: Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Olfactory Receptors; Olfactory Epithelium; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory Higher Centers Anatomy; Olfactory Glomeruli: Structure and Circuitry; Vomeronasal Accessory System.

Further Reading

Halász N (1990) *The Vertebrate Olfactory System: Chemical Neuroanatomy, Function and Development*. Budapest: Akadémiai Kiadó.

Kosaka T and Kosaka K (2005) Synaptic organization of the glomerulus in the main olfactory bulb: Compartments of the glomerulus and heterogeneity of the periglomerular cells. *Anatomical Science International* 80: 80–90.

Macrides F, Schoenfeld TA, Marchand JE, and Clancy AN (1985) Evidence for morphologically and functionally heterogeneous classes of mitral and tufted cells in the olfactory bulb. *Chemical Senses* 10: 175–202.

Mori K (1987) Membrane and synaptic properties of identified neurons in the olfactory bulb. *Progress in Neurobiology* 29: 275–320.

Pinching AJ and Powell TPS (1971) The neuron types of the glomerular layer of the olfactory bulb. *Journal of Cell Science* 9: 305–345.

Pinching AJ and Powell TPS (1971) The neuropil of the glomeruli of the olfactory bulb. *Journal of Cell Science* 9: 347–377.

Pinching AJ and Powell TPS (1971) The neuropil of the periglomerular region of the olfactory bulb. *Journal of Cell Science* 9: 379–409.

Price JL and Powell TPS (1970) The morphology of the granule cells of the olfactory bulb. *Journal of Cell Science* 7: 91–123.

Price JL and Powell TPS (1970) The synaptology of the granule cells of the olfactory bulb. *Journal of Cell Science* 7: 125–155.

Ramón y Cajal S (1911) *Histologie du Système Nerveux de l'homme et des Vertébrés. Tome II* (translated by L Azoulay from Spanish to French). Paris, Maloine (republished 1952 by Consejo Superior de Investigaciones Cientificas, Madrid).

Shepherd GM, Chen WR, and Greer CA (2004) Olfactory bulb. In: Shepherd GM (ed.) *The Synaptic Organization of the Brain*, 5th edn., pp. 165–216. New York: Oxford University Press.

Shipley MT, Ennis M, and Puche AC (2004) Olfactory system. In: Paxinos G (ed.) *The Rat Nervous System*, 3rd edn., pp. 923–964. San Diego: Elsevier.

VanGehuchten A and Martin I (1891) Le bulbe olfactif chez quelques mammiféres. *La Cellule* 7: 204–237.

Olfactory Bulb Mapping

K Mori, University of Tokyo, Tokyo Japan

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The olfactory bulb is the first relay station of the central olfactory system in the vertebrate brain and contains in its superficial layer a few thousand glomeruli, spherical neuropils with sharp borders (Figure 1). In mammals the diameter of glomeruli ranges from 30 to 200 μm. Individual olfactory sensory neurons in the epithelium of nose express only a single odorant receptor gene among a repertoire of numerous odorant receptor genes (e.g., nearly 1000 genes in mice). This is the one cell–one receptor rule in olfactory sensory neurons.

Each olfactory sensory neuron projects a single axon to a single glomerulus in the olfactory bulb. Within the glomerulus, olfactory axons form excitatory synaptic connections on the dendritic tufts of olfactory bulb neurons, mitral cells, and tufted cells. The olfactory axon projection to the glomeruli is precisely organized in such a way that olfactory sensory neurons expressing a given odorant receptor converge their axons to a few topographically fixed glomeruli (Figure 1). All the axons converging onto a single glomerulus originate from sensory neurons expressing the same odorant receptor (the glomerular convergence rule).

Because of the one cell–one receptor rule and the glomerular convergence rule, individual glomeruli represent a single odorant receptor. Therefore, the glomerular sheet of the olfactory bulb forms odorant receptor maps. This article summarizes an emerging view of the spatial organization of the odorant receptor maps in the olfactory bulb.

Each Olfactory Bulb Contains Two Symmetric Maps

In mice and rats, each olfactory bulb has two mirror-symmetric maps of odorant receptors, one map in the rostro-dorso-lateral part (lateral map) and the other in the caudo-ventro-medial part (medial map) of the olfactory bulb. A majority of odorant receptors are thus represented by two sites in the glomerular sheet of the olfactory bulb, one site in the lateral map and the other site in the medial map. In contrast, a small group of odorant receptors shows a single glomerulus representation and is represented by the glomeruli at the ventral midline of the olfactory bulb. The functional significance of the dual maps in each olfactory bulb is not well understood yet, although the two maps connect each other topographically via axon collaterals of tufted cells.

Zonal Organization

For a majority of odorant receptors, individual receptors are expressed by sensory neurons that are distributed widely in one of the four zones (zones 1, 2, 3, and 4 from the dorsomedial to the ventrolateral direction) in the olfactory epithelium (Figure 1). Glomeruli in both the lateral and medial maps of the olfactory bulb are arranged in a similar zonal fashion, and axons of olfactory sensory neurons expressing the zone-specific odorant receptors show a zone-to-zone projection pattern. Olfactory sensory neurons in zone 1 (and thus expressing zone 1 odorant receptors) project their axons selectively to zone 1 of the olfactory bulb. Sensory neurons in zones 2–4 project to zones 2–4 of the olfactory bulb.

Several studies related the molecular structure of odorants to the zonal organization of odorant receptor expression in the olfactory epithelium and to the zonal organization of odorant receptor representation in the olfactory bulb. For example, odorants that activate individual glomeruli in zone 1 have one or more polar functional groups in common, suggesting that the polar functional groups play a key role in activating zone 1 odorant receptors and thus zone 1 glomeruli. In contrast, odorants that activate individual glomeruli in zones 2–4 tend to have a similar hydrocarbon profile or relatively nonpolar molecular profiles, suggesting that the overall molecular shape is critical in activating zone 2–4 odorant receptors and zone 2–4 glomeruli.

Glomeruli as Molecular Feature-Detecting Units

When a large panel of odorants with systematic variation of their molecular structure is used to stimulate the olfactory epithelium in the mammalian nose, it is possible to estimate the range of odorants that activate an individual glomerulus. This is called the molecular receptive range (MRR) of the glomerulus. It is also possible to compare the molecular features of odorants that activate an individual glomerulus.

Recent optical imaging studies show that individual glomeruli typically respond to a range of odorants that share a specific combination of molecular features. Different glomeruli show different MRR properties, suggesting that each glomerulus in either the lateral map or the medial map may be unique in its

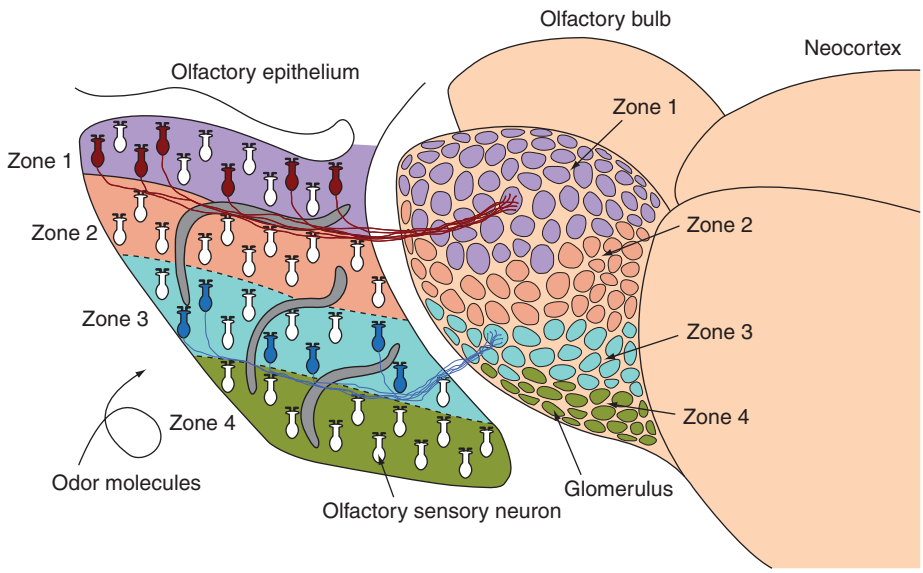

Figure 1 Axonal projection pattern of olfactory sensory neurons to the glomeruli of the rodent olfactory bulb. The olfactory epithelium in rats and mice is divided into four zones (zones 1–4). A given odorant receptor is expressed by sensory neurons located within one zone of the epithelium. Individual olfactory sensory neurons express a single odorant receptor. Olfactory sensory neurons expressing a given odorant receptor are distributed widely in the epithelial zone and converge their axons onto a few topographically fixed glomeruli that are located within a corresponding zone of the olfactory bulb. Thus, each glomerulus represents a single odorant receptor. The glomerular sheet of the olfactory bulb forms odorant receptor maps. Adapted from Mori K, Takahashi YK, Igarashi KM, and Yamaguchi M (2006) Maps of odorant molecular features in the mammalian olfactory bulb. *Physiology Review* 86: 409–433.

MRR property. The results support the idea that each glomerulus functions as a molecular feature-detecting unit.

Extracellular recording of spike activity from single neurons also shows that individual mitral and tufted cells in the mammalian olfactory bulb respond to a range of odorants that share a specific combination of molecular features. Because individual mitral and tufted cells in the mammalian olfactory bulb project a single primary dendrite to a single glomerulus and receive the olfactory axon inputs exclusively within the glomerulus, the MRR property of individual mitral and tufted cells may strongly reflect that of the affiliated glomerulus.

Thus, individual glomeruli are thought to be molecular feature-detecting modular units, just as individual odorant receptors are thought to be molecular feature-detecting molecular units. How are the molecular feature-detecting units spatially arranged in the glomerular sheet of the olfactory bulb?

Molecular Features of Odorants

Olfaction is mediated by a vast variety of odorants that are inhaled into the nose. To understand the molecular feature maps in the olfactory bulb, it is important to summarize the knowledge of molecular features of odorants. Odorants are small volatile compounds with molecular mass less than 300 Da. It has been estimated that more than 400 000 different compounds are odorous to the human nose. Organic and fragrance/flavor chemists and psychophysicists have documented a variety of relationships between the molecular structure of odorants and their subjectively perceived odor. This is called the structure–odor relationship.

The chemical and psychophysical studies have noted the importance of two categories of odorant molecular features in the structure–odor relationship. One category is a polar functional group that contains an oxygen, a nitrogen, or a sulfur atom. Polar functional groups that have a strong influence on the perceived odor include the carboxyl group (–COOH), aldehyde group (–CHO), hydroxyl group (–OH), ketone group (>CO), ester group (–COO–), amino group ($-NH_2$), and thiol group (–SH), among others. The other category is the molecular profile that is formed mainly by the overall molecular shape.

Combinatorial Coding

In order to spatially map the odorant-induced glomerular activity in the olfactory bulb, a variety of methods have been used. The mapping methods include 2-deoxyglucose uptake, optical imaging with voltage-sensitive dye, electrophysiological single-cell recordings from individual mitral/tufted cells, expression of immediate early genes such as c-fos, c-jun,

zif-268, and activity-regulated cytoskeleton-associated (ARC) genes, optical imaging of intrinsic signals, expression of phosphorylated extracellular signal-regulated kinase (ERK), imaging using pHluorin, and functional magnetic resonance imaging (fMRI).

The results obtained by these methods clearly show that a given odorant activates a specific combination of glomeruli that are located and distributed at stereotypical positions in the olfactory bulb. Different odorants are coded by different combinations of activated glomeruli. For each odorant, activated glomeruli are spatially arranged in a roughly symmetrical fashion across medial and lateral maps.

In summary, each odorant is coded by a specific combination of odorant receptors (molecular feature-detecting receptors), by a specific combination of olfactory sensory neurons (molecular feature-detecting cells), and by a specific combination of glomeruli (molecular feature-detecting modules). In other words, information carried by individual odorants is fractionated according to the odorant receptors that they bind and handled by a specific combination of molecular feature-detecting glomerular modules in the olfactory bulb. The principle of combinatorial coding at the odorant receptor level, olfactory sensory neuron level, and glomerular level is found in the olfactory system of a wide variety of vertebrates and invertebrates.

Molecular Feature Clusters of Glomeruli in the Olfactory Bulb

Comparison of the molecular features of odorants that activate individual glomeruli shows that glomeruli with a similar MRR property are located in proximity and form molecular feature clusters. The molecular feature clusters (also called molecular feature modules or domains) are located at stereotypical positions in the olfactory bulb. Figure 2 shows the spatial arrangement of nine molecular feature clusters (clusters A–I) in the lateral map of the rat olfactory bulb. These clusters were defined by the method of optical imaging of intrinsic signals.

Although different methods define the molecular feature clusters in different ways, the results of these studies are consistent with the idea that glomeruli that detect a similar combination of molecular features gather together in proximity and form molecular feature clusters. The molecular feature clusters are the basic spatial organization and appear to cover virtually all parts of the glomerular maps in the rodent olfactory bulb.

The finding of molecular feature clusters of glomeruli suggests that glomeruli in the same cluster represent odorant receptors having similar MRR properties and thus have similar odorant receptive site structures. This idea is consistent with the evidence that olfactory sensory neurons expressing homologous odorant receptor genes project their axons to neighboring glomeruli. These results suggest cluster-specific grouping and representation of subsets of odorant receptors.

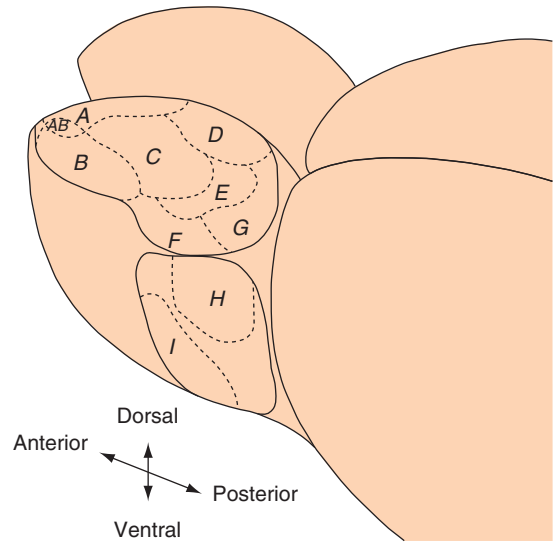

Figure 2 Molecular feature clusters of glomeruli in the dorsal and postero-lateral surfaces of the olfactory bulb. Molecular feature clusters A–D are located in zone 1 of the glomerular map, while clusters F–I are in zones 2–4. Adapted from Mori K, Takahashi YK, Igarashi KM, and Yamaguchi M (2006) Maps of odorant molecular features in the mammalian olfactory bulb. *Physiology Review* 86: 409–433.

The clustering of glomeruli that detect similar molecular features is reported not only in the mammalian olfactory bulb but also in the fish olfactory bulb and even in the *Drosophila* antennal lobe. However, it remains to be determined whether the human olfactory bulb has similar molecular feature clusters of glomeruli.

Molecular Feature Maps

Because each glomerulus appears to be unique in its MRR property, an individual glomerulus can be characterized by determining the specific combination of molecular features that are shared by odorants effective in activating the glomerulus. This is called the characteristic molecular feature of a glomerulus. By superimposing the three-dimensional structure of odorants that are effective in activating a glomerulus, it is possible to deduce the characteristic molecular features of the glomerulus.

Although the detailed analysis is limited to only a small part in the glomerular maps, comparison of the characteristic molecular features of individual glomeruli along the axes of the glomerular maps

suggests that the glomerular sheet represents the characteristic molecular features in a systematic, gradual, and multidimensional fashion.

Representation of Basic Odor Quality

Odorants having a similar combination of molecular features tend to have a similar odor quality to the human nose. Because glomeruli with a similar MRR property form a molecular feature cluster in the mammalian olfactory bulb, molecular feature clusters might also be present in the human olfactory bulb and are part of the neural representation of basic odor quality.

Glomeruli in cluster C (Figure 2) in the rat olfactory bulb, for example, respond to phenol family odorants and phenyl ethers. These odorants share the molecular features composed of a benzene ring with a hydroxyl group or an alkoxyl group. To the human nose, these odorants contain phenolic and medicinal odor in common. This suggests that cluster C glomeruli, if present in the human olfactory bulb, may participate in the representation of phenolic and medicinal odor quality.

The idea that the molecular feature clusters in the olfactory bulb may participate in the neuronal representation of basic odor quality raises the issue of whether parosmia, smelling something other than the presented odor, might in part originate from the distorted maps in the human olfactory bulb or abnormal activity in it.

Functional Roles of Molecular Feature Maps

In addition to the possible participation in the perception of basic odor quality, the spatial arrangement of the molecular feature maps is critical in determining the manner of interactions among olfactory bulb neurons that are affiliated with different glomeruli. In the mammalian olfactory bulb, an individual glomerulus is innervated by several thousands of converging olfactory axons, ~10–20 mitral cells, and ~50–70 tufted cells.

Mitral/tufted cells affiliated with a particular glomerulus can interact with those affiliated with different glomeruli via local neuronal circuits in the olfactory bulb. One of the major local neuronal circuits involves dendrodendritic reciprocal synaptic connections between the long secondary dendrites of mitral/tufted cells and apical dendrites of granule cells. Another major local circuit involves juxtaglomerular interneurons.

First, mitral/tufted cells associated with a glomerulus in a particular molecular feature cluster can interact with those associated with neighboring glomeruli in the same cluster. In the cluster A of the rabbit olfactory bulb, for example, individual mitral cells are activated by a range of fatty acids with different carbon chain lengths. In addition, individual mitral cells are typically inhibited by fatty acids with carbon chain lengths slightly longer or slightly shorter than the effective odorants. The tuning of individual mitral cells to odorant molecular features is thus enhanced by the lateral inhibition from mitral cells associated with neighboring glomeruli.

Interaction among mitral cells associated with different molecular feature clusters is also observed in the rat olfactory bulb. Alkylamines that are generated by spoiled foods have an unpleasant fatty fishy odor and activate glomeruli in cluster A. Herb spices such as fennel and clove activate glomeruli in the neighboring molecular feature clusters and temporally inhibit the alkylamine-induced response of mitral cells in cluster A. Because fennel and clove are known to mask the fatty fishy smell as well as to add their own flavor, the masking of fatty fishy odor may be mediated at least in part by the lateral inhibition among mitral/tufted cells associated with different molecular feature clusters in the odor maps.

The local neuronal circuits are known to mediate synchronized oscillatory activity among mitral cells associated with different glomeruli. Because the synchronized oscillatory activity may be crucial in activating olfactory cortex neurons to which mitral cell axons project, the molecular feature maps provide the substrate for the local neuronal circuits to select specific subsets of mitral cells for the synchronization.

Olfactory Cortex Reads the Odor Maps in the Olfactory Bulb

Mitral cells project their axons via the lateral olfactory tract to various regions of the olfactory cortex. Recent studies show that mitral cells receiving input from a given odorant receptor (and thus associated with a few specific glomeruli) project their axons to specific clusters of olfactory cortex neurons that are arranged at stereotypical positions. Individual olfactory cortex neurons appear to receive inputs from multiple different odorant receptors and thus from multiple different glomeruli. However, it is not well understood how individual olfactory cortex neurons integrate signals derived from multiple different odorant receptors and thus from multiple different glomeruli. The knowledge of the molecular feature maps in the olfactory bulb may provide a basis for understanding how the olfactory cortex reads the odor maps in the olfactory bulb.

See also: Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Olfactory Receptors; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Higher Centers Anatomy; Olfactory Glomeruli: Structure and Circuitry; Vomeronasal Accessory System.

Further Reading

Axel R (1995) The molecular logic of smell. *Scientific American* 273: 154–159.

Buck LB (1996) Information coding in the vertebrate olfactory system. *Annual Review of Neuroscience* 19: 517–544.

Leon M and Jhonson BA (2003) Olfactory coding in the mammalian olfactory bulb. *Brain Research* 42: 23–32.

Mori K, Nagao H, and Yoshihara Y (1999) The olfactory bulb: Coding and processing of odor molecule information. *Science* 286: 711–715.

Mori K, Takahashi YK, Igarashi KM, and Yamaguchi M (2006) Maps of odorant molecular features in the mammalian olfactory bulb. *Physiology Review* 86: 409–433.

Rubin BD and Katz LC (1999) Optical imaging of odorant representation in the mammalian olfactory bulb. *Neuron* 23: 499–511.

Shepherd GM, Chen WR, and Greer CA (2004) Olfactory bulb. In: Shepherd GM (ed.) *The Synaptic Organization of Brain*, pp. 165–216. New York: Oxford University Press..

Uchida N, Takahashi YK, Tanifuji M, and Mori K (2000) Odor maps in the mammalian olfactory bulb: domain organization and odorant structural features. *Nature Neuroscience* 3: 1035–1043.

Xu F, Greer CA, and Shepherd GM (2000) Odor maps in the olfactory bulb. *Journal of Comparative Neurology* 422: 489–495.

Zou Z, Horowitz LF, Montmayeur JP, Snapper S, and Buck LB (2001) Genetic tracing reveals a stereotyped sensory map in the olfactory cortex. *Nature* 414: 173–179.

Relevant Websites

http://leonlab.bio.uci.edu – Glomerular Activity Response Archive.

http://www.uchsc.edu – Restrepo Laboratory, University of Colorado at Denver and Health Sciences Center.

http://senselab.med.yale.edu – Senselab; see the Olfactory Bulb Odor map DataBase (OdorMapDB).

Olfactory Bulb Physiology

A V Masurkar and W R Chen, Yale University, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The olfactory bulb is the first processing station in the central olfactory pathway. It receives olfactory nerve input from the sensory epithelium in the nose and sends mitral/tufted cell axonal output to the olfactory cortex. One critical determinant of olfactory bulb function is the nose-to-bulb projection. Olfactory sensory neurons expressing a given odor receptor gene project their axons into two or very few specific spherical structures called olfactory glomeruli. Such a unique sensory projection indicates that the glomerulus is a basic coding unit for olfactory signals. A single odorant activates one set of glomeruli that are distributed widely across the olfactory bulb surface. Different odorants activate distinct but overlapping sets of glomeruli. Thus, a central function of the olfactory bulb is to process odor coding patterns and relay processed odor codes to the higher olfactory brain centers.

Odor processing starts at the level of the olfactory glomerulus. Besides olfactory nerve terminals, the glomerulus is surrounded by hundreds of juxtaglomerular cells. It also receives dendritic projections from tens of mitral/tufted cells located more deeply. As mitral/tufted cells (and most juxtaglomerular cells) send their dendrites into only a single glomerulus for receiving olfactory nerve input, a glomerulus and its associated mitral/tufted and juxtaglomerular cells form a functional module. Thousands of glomerular modules serve as parallel channels for relaying odor signals throughout the olfactory bulb to the piriform cortex. Between these vertical or radial modules are two levels of horizontal circuits mediating lateral intermodular interactions. One is via short-axon cells in the glomerular layer, and the other is through mitral/tufted cell basal secondary dendrites in the external plexiform layer.

Radial Glomerular Modules

In a glomerular module, odor signals are transmitted from olfactory nerve terminals to distal dendritic tufts of mitral/tufted cells in the glomerulus and then from these glomerular tufts to the mitral/tufted cell soma–axon region via apical primary dendrites.

Olfactory Nerve Synapses

Olfactory nerve terminals make excitatory synapses with mitral/tufted cells and certain types of juxtaglomerular cells. These synapses are glutamatergic, and postsynaptic effects involve α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and *N*-methyl-D-aspartate (NMDA) receptors as well as metabotropic glutamate receptors. Transmitter release is highly reliable at these terminals, with a release probability of ≥ 0.8 and a linear dependence on extracellular calcium concentration. The synapses show strong paired pulse depression due to presynaptic activation of γ-aminobutyric acid (GABA) class-B ($GABA_B$) receptors and dopamine D_2 receptors by GABA and dopamine released from juxtaglomerular cells. Therefore, monosynaptic and disynaptic activation of GABAergic and dopaminergic juxtaglomerular cells can provide feedback inhibition onto olfactory nerve terminals (**Figure 1(a)**, upper-left inset).

The mitral cell response to olfactory nerve electric stimulation usually exhibits a long-lasting excitatory postsynaptic potential that is mediated by AMPA, NMDA, and type-1 metabotropic glutamate receptors. This process is initially triggered by release of glutamate from olfactory axon terminals and then prolonged by secondary and regenerative glutamate release from excited mitral/tufted-cell dendrodendritic synapses in the glomerulus. Such recurrent excitation further involves glutamate spillover within a glomerulus (**Figure 1(a)**, lower-left and upper-right insets). One likely function for these mechanisms is to amplify weak olfactory input. They are also involved in coordinating intraglomerular neuronal synchrony, as described further below in the section titled 'Intraglomerular microcircuits.'

Mitral Cell Apical Primary Dendrite

In the olfactory bulb, mitral/tufted cells are projection neurons sending axonal output to the olfactory cortex. Each mitral cell extends one apical primary dendrite superficially into a single glomerulus to form a tuft of branches containing both post- and presynaptic sites. Odor signal transmission from the glomerulus to the mitral cell soma–axon region is mediated by both passive spread of synaptic potentials and active propagation of action potentials along this dendrite. Dual-patch recordings made simultaneously from the mitral cell soma and distal primary dendrite have revealed that mitral cells possess some unique dendritic properties compared with other types of central neurons.

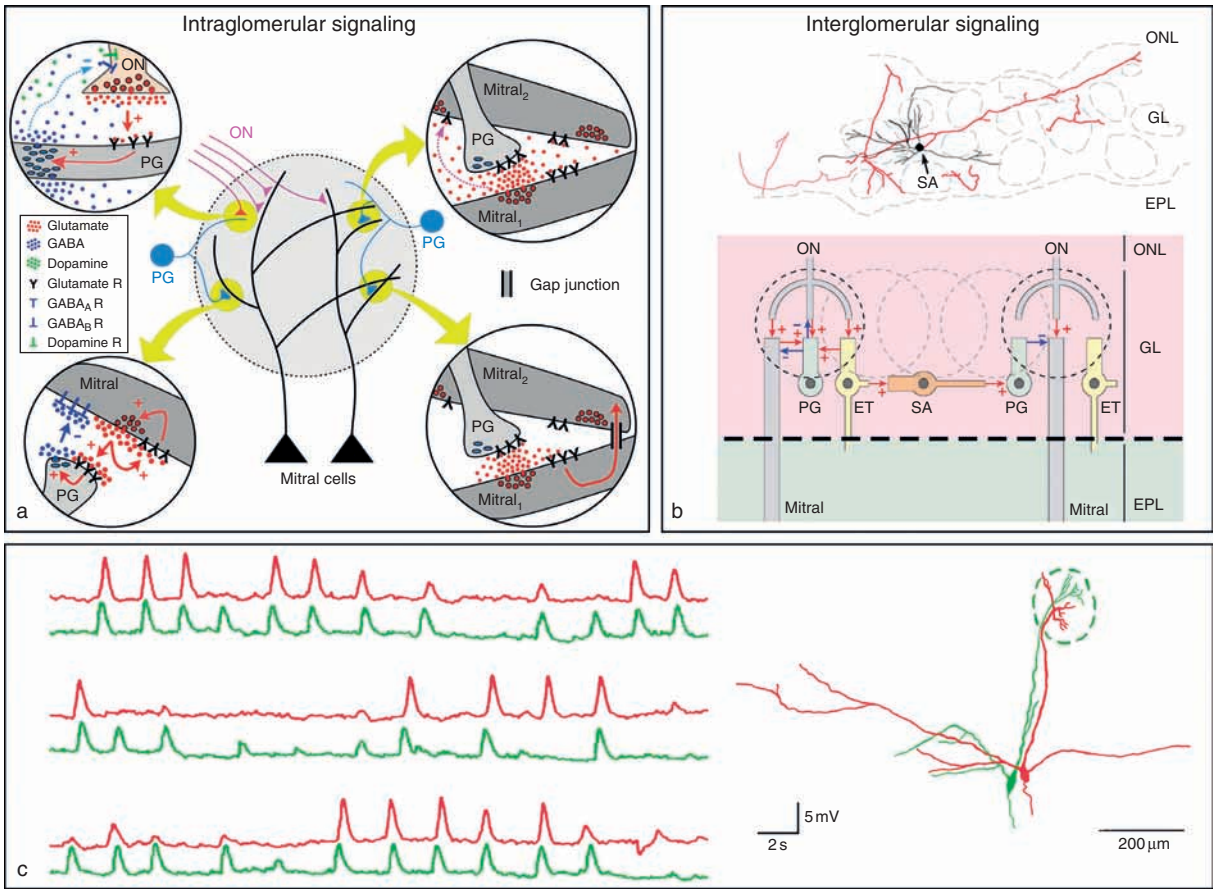

Figure 1 (a) Intraglomerular circuits. Upper-left circular inset, γ-aminobutyric acid (GABA) or dopamine released from periglomerular cells can act presynaptically via GABA$_B$ and dopamine D$_2$ receptors on olfactory nerve terminals to inhibit olfactory sensory input. Lower-left inset, a mitral cell glomerular dendritic branch releases glutamate to excite an inhibitory periglomerular cell dendrite, which in turn releases GABA back onto the mitral cell dendrite. Glutamate released from the mitral cell tuft branch can also act on autoreceptors, causing further regenerative release of glutamate. (Upper-right) Glutamate released from one mitral cell dendritic branch spills over to another mitral cell dendrite, causing lateral excitation. (Lower-right) Glutamate released from one mitral cell dendrite acts on autoreceptors, causing self-excitation which spreads into another mitral cell dendrite via electrical coupling to mediate lateral excitation. (b) Interglomerular circuits. Mitral cells, external tufted cells, and subpopulations of periglomerular cells receive monosynaptic inputs from olfactory nerve terminals. Excitation of GABAergic and dopaminergic periglomerular cells can inhibit olfactory nerve input by a presynaptic mechanism. Both mitral and external tufted cells can excite periglomerular cells, which can provide recurrent inhibition. The external tufted cell excites the short-axon cell, which in turn activates the GABAergic periglomerular cell in another glomerulus, causing interglomerular lateral inhibition. EPL, external plexiform layer; ET, external tufted cell; GL, glomerular layer; ON, olfactory nerve; ONL, olfactory nerve layer; PG, periglomerular cell; SA, short-axon cell. Adapted from Aungst, JL, Hayward, PM, Puche, AC, et al. (2003) Centre-surround inhibition among olfactory bulb glomeruli. *Nature* 426: 623–629. (c) Slow rhythmic activity synchrony between two mitral cells projecting to a common glomerulus. Adapted from Carlson GC, Shipley MT, and Keller A (2000) Long-lasting depolarizations in mitral cells of the rat olfactory bulb. *Journal of Neuroscience* 20: 2011–2021.

There are four different modes of action potential initiation and propagation along the mitral cell soma–primary dendrite axis (Figure 2).

With weak to moderate excitation of the mitral cell glomerular dendrites by olfactory nerve input, odor signals are first transmitted along the primary dendrite in the form of passive spread of postsynaptic potentials. Once this input drives the membrane potential reaching the firing threshold, an action potential is initiated in the axon initial segment and propagates forward down the axon to the olfactory cortex and backward into both the primary and secondary dendrites (Figure 2(a)). Increasing olfactory nerve synaptic excitation shifts the spike initiation site from the axon initial segment to the glomerular tuft branches, and in this case signal transmission takes form of forward spike propagation along the primary dendrite to the mitral cell soma and axon (Figure 2(b)). When strong olfactory nerve input is paired with strong synaptic inhibition near the cell body, an action potential is initiated in the distal primary dendrite but fails to propagate forward to the soma. This transforms the mitral cell from a projection neuron to a local interneuron producing dendrodendritic synaptic

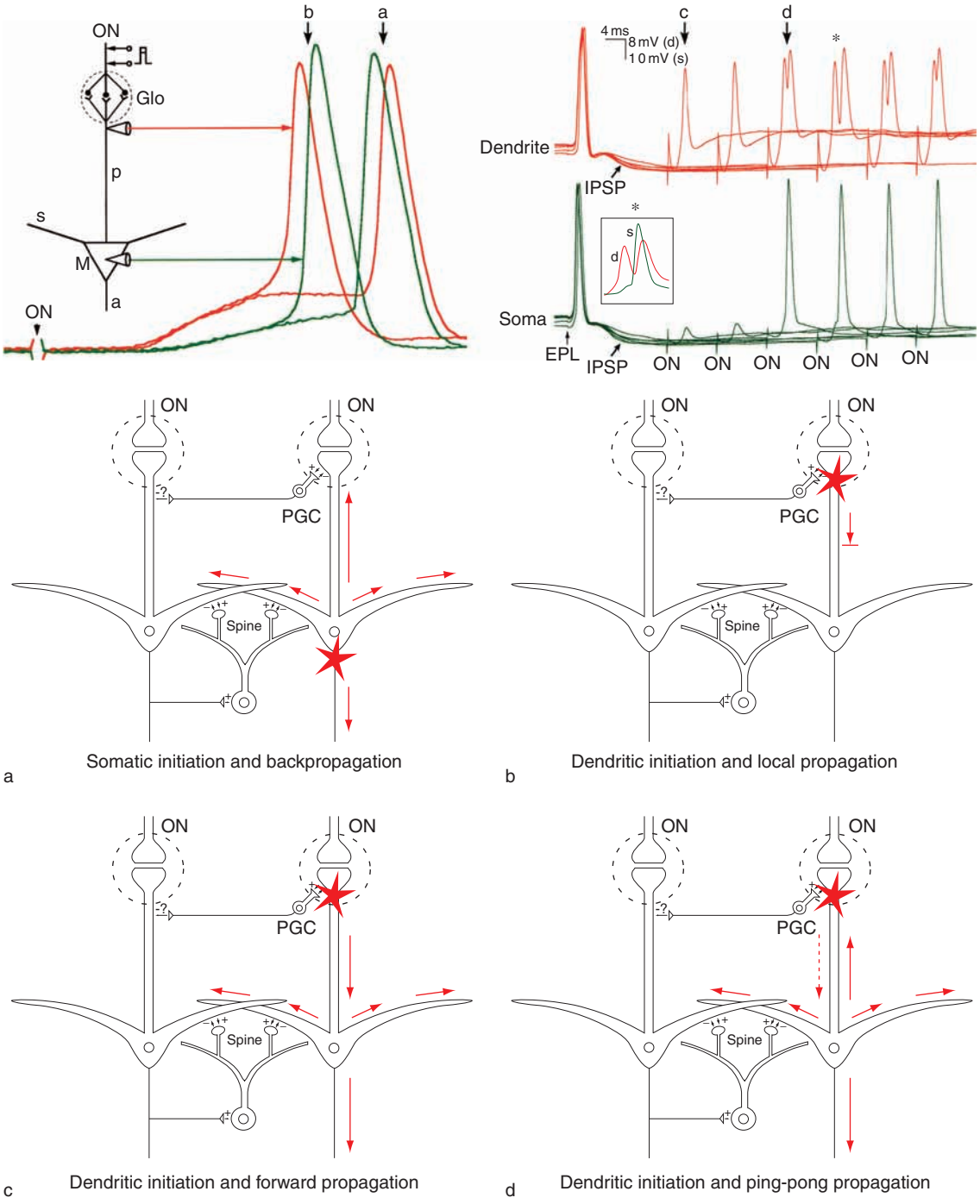

Figure 2 Four different modes of action potential initiation and propagation along the mitral cell apical primary dendrite. (Top row) Dual-patch recordings from the mitral cell soma and distal primary dendrite in response to increasing olfactory nerve input from (a) to (b) (left) and to pairing a strong olfactory nerve input with granule cell inhibition evoked at different time intervals (right). Panels (a) and (b) are illustrations showing four different modes of spike initiation and propagation in the mitral cell primary dendrite, as evidenced by the corresponding membrane potential traces shown in top row. See the text for a more detailed description. a, axons; d, double spikes; EPI, external plexiform layer; Glo, glomerulus; IPSP, inhibitory postsynaptic potential/space; M, mitral cell; ON, olfactory nerve; P, primary dendrite; PGC, periglomerular cell; S, secondary dendrite; S (inset), single spike. Adapted from Chen WR, Shen GY, Shepherd GM, Hines ML, and Midtgaard J (2002) Multiple modes of action potential initiation and propagation in mitral cell primary dendrite. *Journal of Neurophysiology* 88: 2755–2764.

output only within an olfactory glomerulus (Figure 2 (c)). With strong distal excitation paired with intermediate proximal inhibition, an action potential is initiated in the glomerular tuft. In the course of forward propagation, the spike jumps to the axon initial segment to generate a mitral cell axonal output and meanwhile to induce a spike back-propagation (Figure 2(d)).

These four different modes of spike initiation and propagation indicate that odor signal transmission within a vertical glomerular module is subject to modulation by lateral inhibitory synaptic circuits and that an odor map is not always faithfully transferred from the glomerular plane down to the mitral cell soma layer. Significant coding pattern transformation can occur within the olfactory bulb.

Intraglomerular Microcircuits

In each glomerular module, olfactory signals are relayed to about 20–30 mitral/tufted cells. The activity of these output neurons is coordinated by several intraglomerular mechanisms. First, surrounding a glomerulus are hundreds of juxtaglomerular neurons, which are heterogeneous and include external tufted, periglomerular, and short-axon cells. External tufted cells receive monosynaptic input from olfactory nerve terminals. They have unique intrinsic membrane properties that generate long-lasting depolarization plateau potentials and spike bursts. These neurons provide strong excitatory dendritic output to other neurons within the glomerulus and appear to play a key role in synchronizing the entire glomerular circuit. Second, mitral cells sharing a common glomerulus are connected by electrical coupling via gap junctions located in their glomerular dendritic branches (Figure 1(a), lower-right inset). This mechanism ensures that all principal neurons of a glomerular module can respond to sensory inputs synergistically as a single unit. Third, the glomerular dendritic tufts of mitral/tufted cells carry glutamate autoreceptors in their presynaptic membrane, which can mediate regenerative release of glutamate from these dendrites and eventually lead to glutamate spillover and intraglomerular neuronal synchrony (Figure 1(c)). As the glomerulus is a chemical compartment isolated by surrounding glial cells, glutamate spillover-mediated synchrony is specific to each glomerulus, which futher consolidates individual glomeruli as functional units.

In vivo olfactory bulb physiology reveals that mitral/tufted cells usually show two different timescales of synchronized activity. One is related to slow oscillations that are phase locked with animal respiratory cycles, and the other involves spike synchrony with a temporal precision less than 10 ms. Both forms of synchrony can happen among the cells either sharing a common glomerulus or belonging to different glomeruli. The mechanisms for interglomerular synchrony are described in the next section.

Lateral Interglomerular Circuits

As olfactory signals generated by a single odorant are usually distributed in multiple parallel glomerular modules, lateral interglomerular interactions are crucial for both synthesizing a unified odor percept and discriminating between different odorants. Such lateral interaction occurs in two distinct layers, the glomerular layer and external plexiform layer, each of which may represent a different processing stage.

Short-Axon Cells in the Glomerular Layer

Although neuronal circuits in the glomerular layer are mostly oriented to individual glomeruli, at least one type of neuron – the short-axon cell – presents an exception. In contrast to their name, these neurons give rise to a rich and extensive axonal plexus, which projects to multiple glomeruli located up to several hundred microns away (Figure 1(b), upper panel). The short-axon cells do not receive monosynaptic input from olfactory nerve terminals but are instead excited disynaptically via external tufted cells. The short-axon cells are considered to be glutamatergic in some studies and provide an excitatory synaptic input to periglomerular and external tufted cells in other glomeruli (Figure 1(b), lower panel).

The extensive innervation of individual short-axon cells gives rise to a long-range lateral circuit within the glomerular layer. The signal flow of this circuit is from olfactory nerve terminals to external tufted cells to short-axon cells in one glomerulus, and then through their axonal network to periglomerular cells and further to mitral cells in another distally located glomerulus. Many periglomerular cells are inhibitory, so strong activation of one glomerulus can lead to inhibition of mitral-cell tufts in other weakly activated glomeruli via this lateral circuit, yielding lateral inhibition in the glomerular layer.

Dendrodendritic Reciprocal Synapses

Just deep to the glomerular layer lies the external plexiform layer. Each mitral cell extends several lateral secondary dendrites in this layer over a long distance (>1000 μm). One unique feature of olfactory bulb circuits is reciprocal synapses between these secondary dendrites and dendritic spines of axonless granule cell interneurons. The mitral cell dendrite releases glutamate, activating both AMPA and NMDA receptors in the granule cell spine, which in turn release GABA back onto the mitral cell dendrite, causing $GABA_A$ receptor-mediated feedback or recurrent inhibition

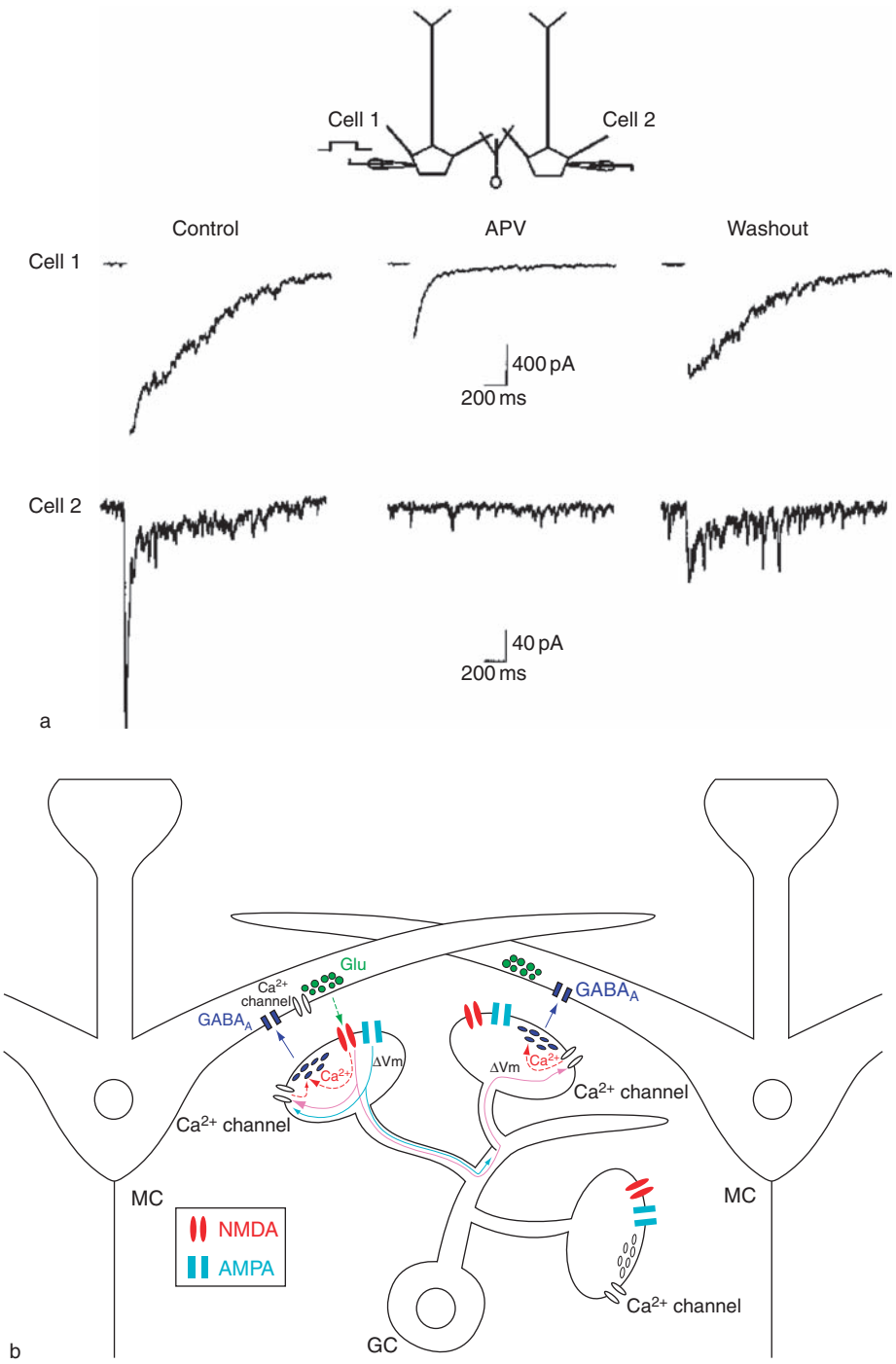

Figure 3 Olfactory bulb dendrodendritic reciprocal synapses mediate mitral cell recurrent and lateral inhibition. (a) Pair recordings from two mitral cells. A depolarization pulse in Cell 1 induced a feedback inhibitory current in Cell 1 and also a lateral inhibitory current in Cell 2, both of which are reversibly blocked by an *N*-methyl-D-aspartate (NMDA) receptor antagonist (APV). (b) A diagram illustrating signal flow underlying recurrent and lateral inhibition. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; GC, granule cell; Glu, glutamate, $GABA_A$, γ-aminobutyric acid class-A receptor; MC, mitral cell. Dashed-line arrow, Ca^{2+} diffusion; solid-line arrow, spread-out of postsynaptic depolarization potential. (a) Reproduced from Isaacson JS and Strowbridge BW (1998) Olfactory reciprocal synapses: Dendritic signaling in the CNS. *Neuron* 20: 749–761, with permission from Elsevier.

(Figure 3(a)). The feedback GABA release depends critically on the activation of NMDA rather than AMPA receptors, indicating that calcium influx through NMDA receptors can provide a direct route for triggering presynaptic transmitter release (Figure 3(b)). As different spines of a granule cell usually contact dendrites from different mitral cells, spread of excitatory postsynaptic potentials among these spines can lead to lateral inhibition between mitral cell pairs (Figure 3). Besides mediating feedback and lateral inhibition, the mitral–granule cell reciprocal synapses have been considered as the key loci for storage of odor memory. The dominant role of NMDA receptors at these synapses may confer on them an extraordinary capacity for plasticity.

Mitral Cell Basal Secondary Dendrites

The long extension of mitral cell secondary dendrites presents an anatomical basis for one glomerular module to interact laterally with hundreds of other modules in the olfactory bulb. This interaction is mediated by spike propagation along these dendrites. Unlike most other central neurons, mitral cells possess membrane mechanisms that can support active propagation of fast sodium spikes from the cell body all the way to the distal ends of secondary dendrites (Figure 4(a)). During such propagation, action potentials trigger dendritic release of glutamate to activate the mitral-to-granule cell dendrodendritic synapses. On the other hand, inhibitory granule-to-mitral cell synapses, activated either by other mitral cells or via centrifugal fibers, can block spike propagation dynamically at variable distances from the soma (Figures 4(b) and 4(c)). Such an interaction between spike propagation and dendrodendritic synaptic activation leads to the emergence of functional domains in the mitral cell secondary dendrites, which can expand and contract dynamically to provide a circuit mechanism for coupling and decoupling different sets of glomerular modules for the processing of different odorants.

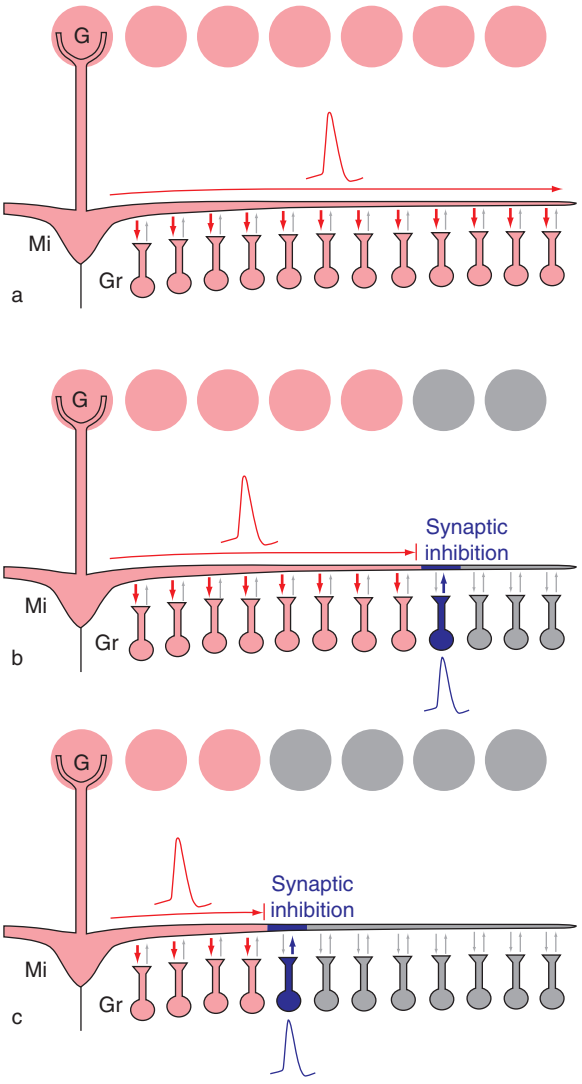

Figure 4 (a) An action potential can propagate actively throughout the entire length of a mitral cell (Mi) secondary dendrite. (b, c) Granule cell (Gr) inhibitory synaptic inputs along the secondary dendrite can dynamically gate spike propagation and thus determine the spatial activation range of mitral-to-granule cell dendrodendritic synpases. G, glomerulus; red thick arrow, activated excitatory synapse; blue thick arrow, activated inhibitory synapse; gray thin arrow, nonactivated synapse. Adapted from Xiong W and Chen WR (2002) Dynamic gating of spike propagation in the mitral cell lateral dendrites. *Neuron* 34: 115–126.

Granule Cell Dendritic Signal Integration

As axonless interneurons in the olfactory bulb, granule cells extend their spiny dendrites perpendicularly with respect to mitral cell secondary dendrites, establishing the dendrodendritic reciprocal synapses at the site of contact. Although the horizontal span of dendritic branches is only about the dimension of one or two glomeruli, these cells receive excitatory synaptic inputs via mitral cell lateral dendrites from a wide range of glomerular modules and display a much broader odor tuning property than mitral cells do. For each granule cell, there are approximately a couple hundred spines, on which the reciprocal synapses rest. Both experimental and computational analyses have revealed that at least three different levels of processing can take place within a granule cell.

The first level of processing is confined to individual spines and mediates short-circuit recurrent inhibition of a connected mitral cell partner (Figure 3(b)). The second level is mediated by local spread of excitatory postsynaptic potentials out of originally activated spines into neighboring dendrites and spines. This

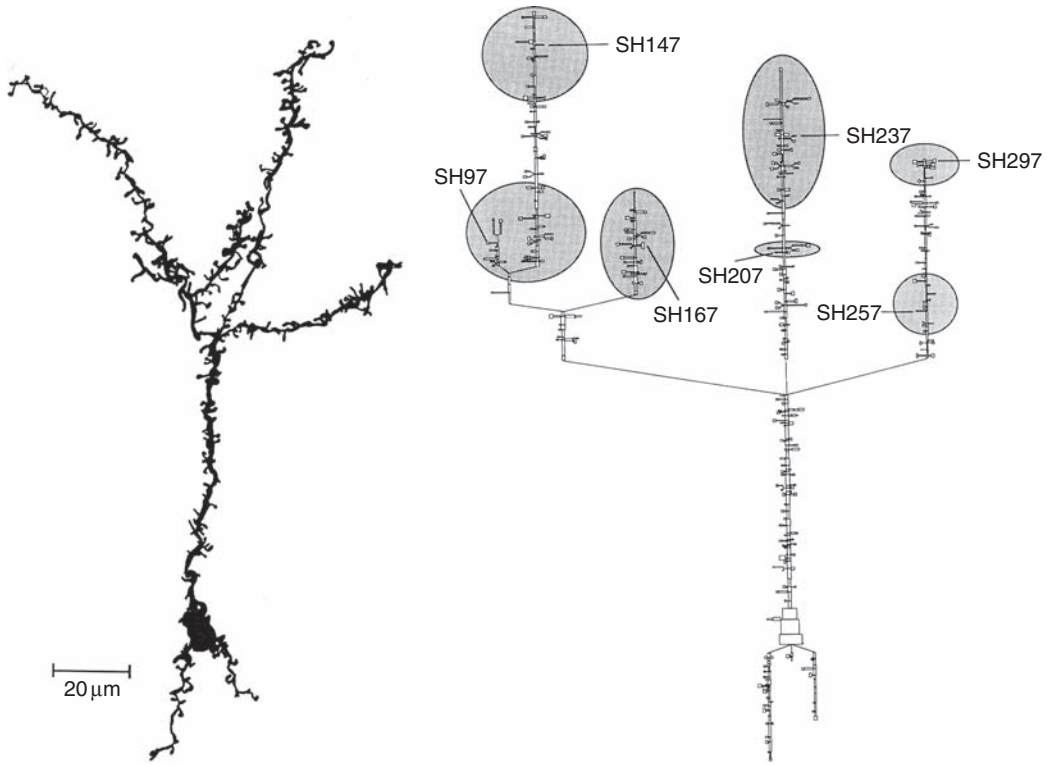

Figure 5 (Left) A granule cell morphology revealed by Golgi silver staining. (Right) A computational model based on morphometric data obtained from the cell shown on the left. A fast conductance simulating excitatory synaptic input is activated in each one of the seven dendritic spines, as denoted by 'SHxxx.' Each shaded area indicates a dendritic domain in which the steady state voltage differs by no more than 5% from the input spine, and all spines within such a domain are considered a functional cluster. Adapted from Woolf TB, Shepherd GM, and Greer CA (1990) Local information processing in dendritic trees: Subsets of spines in granule cells of the mammalian olfactory bulb. *Journal of Neuroscience* 11: 1837–1854.

spread of depolarizing potentials sums up enough depolarization to trigger GABA release synergistically from a cluster of spines (Figure 5) without raising the granule cell soma above its firing threshould. This local dendritic mechanism can mediate lateral inhibition between different mitral cells connected to the spine cluster. Due to the cable filtering effect, the NMDA receptor-mediated slow synaptic potential is more effective than the fast AMPA component in dendritic transmission and summation (Figure 3(b)). This renders the lateral inhibition also dependent critically on NMDA receptor activation. The third level of processing requires the summation of excitatory synaptic inputs from a wide range of spines for spike initiation. The action potential can subsequently propagate throughout the entire granule cell dendrites to generate a global inhibitory output to all connected mitral cells.

Such multilevel processing can be further extended by an inactivating A-type potassium current and a low-threshold calcium current. The A current in granule cell dendrites can specifically suppress AMPA receptor-mediated fast synaptic potential and thus operate as a temporal filter that delays spike initiation and enables NMDA receptors to set a slow pace for GABA release. The low-threshold calcium spike, on the other hand, can mediate in granule cells a new scale of inhibition that operates in a membrane potential range below the threshold for sodium spike generation. In addition, these different modes of processing in granule cells can be dynamically modulated by GABAergic inputs from Blanes cells distributed in the granule cell layer.

Functions of Lateral Inhibition

As in the visual and other sensory systems, lateral inhibition in the olfactory bulb can provide a mechanism to enhance the contrast of spatial coding patterns of similar odorants. In response to a homologous series of aliphatic aldehydes, the mitral cell odor-tuning curve is usually sharpened by lateral inhibition into a configuration of center excitation and surround inhibition (Figure 6).

Another possible function for olfactory lateral circuits is to mediate oscillatory firing synchrony among mitral cells connected either to a common glomerulus or to a group of coactivated glomeruli. One

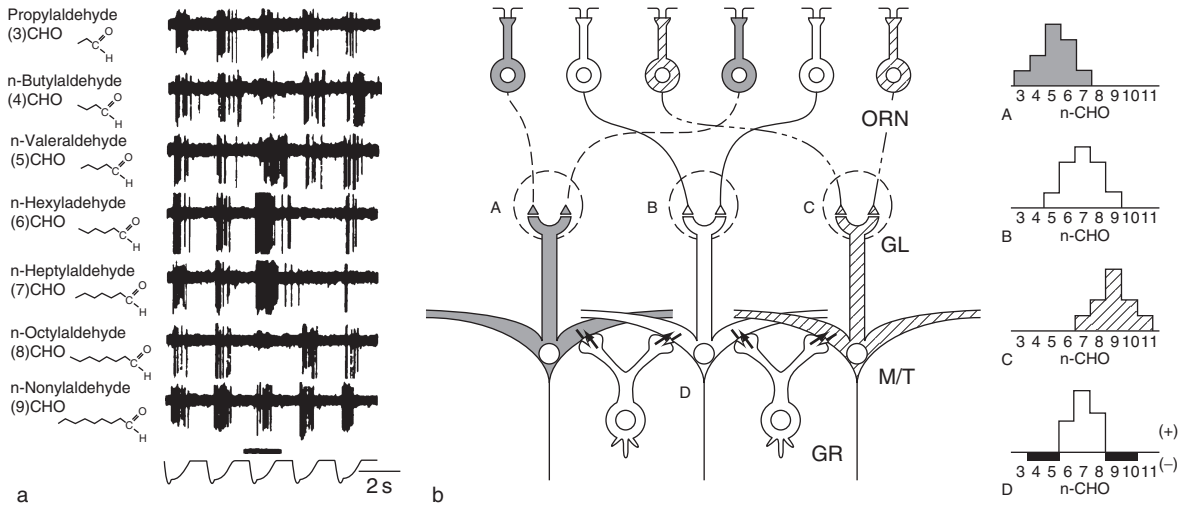

Figure 6 (a) Single-unit recordings of a mitral cell in response to a homologous series of aliphatic aldehydes. The lowest trace records animal respiratory cycles, and the horizontal bar indicates the timing of odor delivery. This mitral cell fired spontaneously in each respiratory cycle. Aldehydes of five- to seven-carbon-chain length significantly increased its firing frequency, while four- or eight-carbon-chain aldehydes inhibited this cell from spiking. (b) A diagram illustrating possible circuit mechanisms that contribute to the sharpening of mitral cell odor-tuning specificity. On the right, different tuning curves of three glomeruli (A–C), reflecting different odor molecular receptive ranges of the receptor cell populations that project to them. The granule cell-mediated lateral inhibition sharpens the mitral cell tuning properties (D). GL, glomerulus; GR, granule cell; M/T, mitral/tufted cell; ORN, olfactory receptor neuron. Adapted from Yokoi M, Mori K, and Nakanishi S (1995) Refinement of odor molecule tuning by dendrodendritic synaptic inhibition in the olfactory bulb. *Proceedings of the National Academy of Sciences of the United States of America* 92: 3371–3375.

prominent feature of the olfactory system is the use of a combinatorial coding strategy. A single odorant is represented by multiple widely distributed glomeruli, and different odorants activate distinct but overlapping sets of glomeruli. Neuronal synchrony, orchestrated by lateral inhibition, can function as an association signal to 'bind' such a distributed combinatorial code for synthesizing a unitary odor percept.

Centrifugal Modulation

In addition to odor processing via vertical and horizontal circuits, the olfactory bulb receives centrifugal inputs from various brain regions. These centrifugal projections are charterized by several different transmitter systems, including noradrenergic fibers from locus coeruleus, cholinergic and GABAergic inputs from the horizontal limb of the diagonal band of Broca, serotoninergic axons from raphe nuclei, and glutamatergic projections from the olfactory cortex. This indicates that odor processing in the bulb is subject to modulation by signals from higher-order olfactory centers as well as nonolfactory subcortical modulatory systems.

Centrifugal fibers from the olfactory cortex terminate mostly in the granule cell layer. They form excitatory synapses on granule cells, and may thus promote inhibition of mitral cell activity via dendrodendritic synapses as described above. The function of this long-loop feedback remains to be established. One of these projection systems may mediate bilateral interbulbar interactions. The tufted cells from one side of the olfactory bulb project to the ipsilateral anterior olfactory nucleus, which sends pyramidal cell axons topographically to the other bulb via the anterior commissure.

Nonolfactory subcortical systems project differentially to the olfactory bulb. Noradrenaline-containing fibers innervate mainly the granule cell and external plexiform layers, whereas the glomerular layer is nearly devoid of adrenergic input. The action of noradrenaline in the olfactory bulb is complicated. It can directly excite mitral cells via α_1 receptors. Meanwhile, it acts on α_2 receptors to suppress GABA release from granule cell spines and thus indirectly enhances mitral cell excitability. When the locus coeruleus is activated by novel or unanticipated events, noradrenaline leads to a transient increase in mitral cell sensitivity to weak odor stimuli. Both serotonin- and acetylcholine-containing fibers have a dense innervation within the glomerular layer. They also exhibit complicated modulatory actions on the olfactory bulb circuits. All these three modulators are critically involved in several forms of olfactory learning and memory, such as odor preference memory and maternal recognition of newborn pups.

Parallel Olfactory Pathways

The vertebrate olfactory bulb is not a homogeneous structure but is instead composed of several subsystems. The main olfactory bulb receives regular sensory input from the main olfactory epithelium, while the accessory olfactory bulb receives pheromone signals from the vomeronasal organ. In addition, the main olfactory bulb can be further divided at least into the regular glomerular system and necklace glomerular system on the basis of different olfactory signal transduction mechanisms employed in their respective sensory neurons. These different subsystems constitute parallel pathways for processing different types of odor signals.

In addition to these parallel pathways defined by separate sensory inputs, the mitral and tufted cell populations represent another set of parallel pathways based on their different intrabulbar circuits. Compared with the mitral cells, the tufted cell bodies are located more superficially, their lateral dendrites extending for a shorter distance, and their axonal projections more confined to the anterior part of the olfactory cortex. Within the main olfactory bulb, there appear to be at least three different populations of granule cells. Those with their soma located superficially in the granule cell layer tend to make reciprocal synapses with the tufted cells, while those deeply located establish synaptic contacts primarily with the mitral cells. The intermediate granule cells interact with both mitral and tufted cells. These two parallel pathways, even sharing inputs from a common glomerulus, may process different aspects of odor signals. The tufted cells, because of their relatively smaller size and higher input resistance, may be more sensitive to weak odor stimuli. On the other hand, their lateral circuits are less extensive than those of mitral cells. Accordingly, the tufted cell pathway may have high odor sensitivity but low discriminability for odorants with similar molecular structures.

A third set of parallel pathways comprises mirror glomeruli in the olfactory bulb. Sensory neurons expressing a given olfactory receptor usually project their axons to two glomeruli located on the medial and lateral sides, and thus each bulb possesses two identical sets of glomeruli. The functional significance of such duplication is not clear. It is interesting that these two sets of glomerular modules have point-to-point mutual inhibitory connections. The superficial tufted cells of one glomerulus project their axon collaterals to the other side, terminating in the internal plexiform layer right below the corresponding mirror glomerulus. These axon terminals make excitatory synapses with inhibitory granule cells and presumably mediate inhibition between the mirror glomeruli.

Circuit Rewiring via Adult Neurogenesis

Throughout the entire life span, sensory neurons in the nasal epithelium and interneurons in the olfactory bulb are continually replaced by newborn neurons every 3–6 months. One major function of adult neurogenesis in the olfactory bulb is to allow plasticity in already mature and relatively fixed circuits. For example, sensory deprivation can dramatically decrease the number, dendritic length, and spine density of newborn granule cells whereas preexisting granule cells show few morphological changes. Bulbar neurogenesis may also represent an adaptive mechanism to ever-changing olfactory peripheral input, which is due to the noxious environment exposed to sensory epithelium. Given such a continuous rewiring of olfactory neuronal circuits, it remains to be explored how odor perceptual constancy is maintained.

See also: Chemical Senses: Overview; Dendritic Signal Integration; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Olfactory Coding; Olfactory Receptors; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Olfactory Higher Centers Anatomy; Olfactory System Theory; Olfactory Glomeruli: Structure and Circuitry; Olfactory System: Circuit Dynamics and Neural Coding in the Locust; Vomeronasal Accessory System.

Further Reading

Aungst JL, Hayward PM, Puche AC, et al. (2003) Centre-surround inhibition among olfactory bulb glomeruli. *Nature* 426: 623–629.

Carlson GC, Shipley MT, and Keller A (2000) Long-lasting depolarizations in mitral cells of the rat olfactory bulb. *Journal of Neuroscience* 20: 2011–2021.

Chen WR, Shen GY, Shepherd GM, Hines ML, and Midtgaard J (2002) Multiple modes of action potential initiation and propagation in mitral cell primary dendrite. *Journal of Neurophysiology* 88: 2755–2764.

Chen WR, Xiong W, and Shepherd GM (2000) Analysis of relations between NMDA receptors and GABA release at olfactory bulb reciprocal synaspes. *Neuron* 25: 625–633.

Christie JM, Bark C, Hormuzdi SG, Helbig I, Monyer H, and Westbrook GL (2005) Connexin36 mediates spike synchrony in olfactory bulb glomeruli. *Neuron* 46: 761–772.

Isaacson JS and Strowbridge BW (1998) Olfactory reciprocal synapses: Dendritic signaling in the CNS. *Neuron* 20: 749–761.

Jahr CE and Nicoll RA (1982) An intracellular analysis of dendrodendritic inhibition in the turtle in vitro olfactory bulb. *Journal of Physiology* 326: 213–234.

Kashiwadani H, Sasaki YF, Uchida N, and Mori K (1999) Synchronized oscillatory discharges of mitral/tufted cells with different molecular receptive ranges in the rabbit olfactory bulb. *Journal of Neurophysiology* 82: 1786–1792.

Lledo PM and Saghatelyan A (2005) Integrating new neurons into the adult olfactory bulb: Joining the network, life-death decisions, and the effects of sensory experience. *Trends in Neurosciences* 28: 248–254.

Mori K (1987) Membrane and synaptic properties of identified neurons in the olfactory bulb. *Progress in Neurobiology* 29: 275–320.

Schoppa NE (2006) Synchronization of olfactory bulb mitral cells by precisely timed inhibitory inputs. *Neuron* 49: 271–283.

Schoppa NE and Westbrook GL (1999) Regulation of synaptic timing in the olfactory bulb by an A-type potassium current. *Nature Neruroscience* 2: 1106–1113.

Shepherd GM, Chen WR, and Greer CA (2004) Olfactory bulb. In: Shepherd GM (ed.) *The Synaptic Organization of Brain*, pp. 165–216. New York: Oxford University Press.

Wilson RI and Mainen ZF (2006) Early events in olfactory processing. *Annual Reivew of Neuroscience* 29: 163–201.

Woolf TB, Shepherd GM, and Greer CA (1990) Local information processing in dendritic trees: Subsets of spines in granule cells of the mammalian olfactory bulb. *Journal of Neuroscience* 11: 1837–1854.

Xiong W and Chen WR (2002) Dynamic gating of spike propagation in the mitral cell lateral dendrites. *Neuron* 34: 115–126.

Yokoi M, Mori K, and Nakanishi S (1995) Refinement of odor molecule tuning by dendrodendritic synaptic inhibition in the olfactory bulb. *Proceedings of the National Academy of Sciences of the United States of America* 92: 3371–3375.

Olfactory Coding

M Bazhenov, The Salk Institute for Biological Studies, La Jolla, CA, USA
M Stopfer, National Institutes of Health, Bethesda, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Olfactory circuitry provides unique advantages for the analysis of information processing by neurons. The task of olfaction is challenging; it demands the ability to rapidly detect, identify, categorize, and prepare for memory storage of myriad odorants that vary in molecular structure and concentration. Yet, olfaction is achieved by relatively few layers of neurons, with anatomical structures and physiological mechanisms that appear repeatedly in widely divergent species. Thus, a study of olfaction offers the promise of insight into a successful and perhaps optimal biological algorithm for processing complex information. Theoretical and computational approaches have contributed substantially toward understanding the functions of olfactory circuitry.

Much current thinking about olfactory processing has been shaped by the compelling observation that the projections of olfactory receptor neurons delineate spatial maps within the olfactory bulb: histological and *in situ* hybridization studies demonstrate convincingly that receptor neurons expressing the same receptor gene converge on a small and specific subset of glomeruli. Based on these observations and on electrophysiological evidence, a model of olfactory information processing has emerged that proposes that individual structural features of odorants are recognized by different receptor neurons, each expressing a specific receptor gene, and that each receptor cell responds to many odorants sharing the same structural features. According to one version of this model, each odorant elicits responses from, and is thus represented by, a more or less static map – a unique spatial ensemble of activated glomeruli.

In addition to such results, a variety of experimental approaches have revealed that olfactory neurons (and those of other sensory systems as well) respond to a static stimulus with oscillatory and temporally structured responses. An example of temporal structure in responses of the olfactory system is the relatively slow, complex sequences of excitation and inhibition in olfactory neurons superimposed on stimulus-evoked oscillations. These patterns have been observed in neurons of the olfactory bulb of amphibians and mammals and within the antennal lobe of insects. In the locust olfactory system, these slow temporal firing patterns generally vary with the odor, and they are reproducible over repeated trials. Intracellular recordings made *in vivo* from locust antennal lobe projection neurons (PNs) have also revealed that individual PNs transiently phase-lock with population oscillations at times that vary with the stimulus. Thus, there is a fine temporal structure to the timing of PN action potentials within the population response that is stable over trials and is different for different PNs. This temporal structure appears to arise mainly from circuit interactions within the antennal lobe. Evidence for odor-evoked spatiotemporal patterns of neuronal activity is not unique to the insect; in salamander olfactory bulb complex spatiotemporal patterns of activity were revealed by voltage-sensitive dye imaging. Recordings from the turtle olfactory bulb have revealed that different assemblies of cells are activated during different cycles of field oscillations, supporting the hypothesis that odors are encoded by spatiotemporal patterns. Thus, the antennal lobe and olfactory lobe appear to process, or reformat, the olfactory information from a diverse population of receptors into a spatiotemporal code.

In vertebrates, output fibers from olfactory bulb neurons project onward to another level of processing in the olfactory (piriform) cortex. In insects, odor information processed in the antennal lobe is transferred to the mushroom body, a structure analogous to the olfactory cortex, and one important for memory formation. Normal cyclic AMP signaling specifically in mushroom body neurons is required for olfactory learning. In honeybees, mecamylamine injection into the mushroom body calyces strongly impaired olfactory learning, suggesting that the cholinergic network in the mushroom body is involved in memory formation. Genetic studies indicate that synaptic transmission from mushroom body neurons is required during memory retrieval but not during memory acquisition or storage. This indicates that acquisition and storage of olfactory memory may both occur upstream of synaptic transmission from the mushroom body, and it suggests that processes contributing to associative learning may reside within mushroom body neurons.

Whatever strategies the antennal lobe or olfactory bulb employs for encoding odors, its follower mushroom body, olfactory cortex, or other circuitry should be able to decode and categorize the information it receives. Therefore, it is of great interest to explore how odor decoding strategies at the downstream levels (mushroom body or olfactory cortex) match the encoding achieved at the first olfactory relay

(antennal lobe and olfactory bulb), providing efficient and optimal odor information processing. In this article, we discuss these coding strategies and specifically emphasize the use of computer models to test several interesting hypotheses of odor encoding and processing.

Encoding Olfactory Information at the First Olfactory Relay

The relative simplicity of the invertebrate olfactory system makes it an attractive system for theoretical and computational studies of olfactory coding. In the locust, each antenna contains approximately 50 000 cholinergic olfactory receptor neurons (numbers range from 2000 to 350 000 in different insect species) that send their axons to the ipsilateral antennal lobe. There, each axon terminates in a few glomeruli, where it forms synapses with the dendrites of GABAergic local neurons (LNs; ~300 in locust) and cholinergic projection neurons (PNs; ~830 in locust). PNs provide the only pathway for olfactory information from the antennal lobe. The number of glomeruli varies among insect species (e.g., from ~50 in *Drosophila melanogaster* to ~1000 in locust). Within the antennal lobe, PN projections can be mainly uniglomerular (e.g., *D. melanogaster*, bee, and cockroach) or multiglomerular (e.g., locust and wasp). Each PN makes connections with a large fraction of LNs, and each LN connects with many PNs as well as with other LNs.

PNs of the antennal lobe provide olfactory input to the mushroom body. In the locust, each mushroom body contains approximately 50 000 Kenyon cells (KCs). In addition to the direct excitatory input from PNs, the KCs of the locust mushroom body also receive competitive, inhibitory input, provided by lateral horn interneurons, which are themselves driven by PNs.

To a remarkable extent, the antennal lobe shares many features with the vertebrate olfactory bulb (OB). Similar to LN–PN circuitry of the antennal lobe, the OB includes interacting excitatory mitral cells and inhibitory granular cells. Lateral inhibition has been reported between granule/periglomerular cells of OB. Axons of the mitral cells project to the olfactory cortex, where they make connections with pyramidal cells and inhibitory interneurons that provide feed-forward inhibition to the pyramidal neurons. The anatomy and functions of OB circuitry are under intensive study.

In this article, we focus on computational approaches to olfactory coding in the insect olfactory system, emphasizing, when possible, similarities to the vertebrate system.

Oscillations and Fine Temporal Structure of Spiking in Olfactory Neurons

The presence of odor-elicited oscillations in the olfactory systems of many animals including insects suggests that oscillatory neural synchrony may play a fundamental role in odor encoding. A number of experimental results provide strong support for this idea. In the locust, pharmacologically blocking the inhibitory $GABA_A$ receptors in the antennal lobe by locally injecting picrotoxin eliminates the 20 Hz oscillations and leads to a breakdown in the specificity of odor responses in beta lobe interneurons, which are followers of KCs. Furthermore, behavioral experiments with honeybees demonstrate that oscillatory synchronization of the antennal lobe neurons is needed for fine odor discrimination; picrotoxin-induced antennal lobe desynchronization causes a loss of precise odor discrimination. Similarly, suppressing synchronous oscillations in the procerebral lobe of the terrestrial slug *Limax* by blocking nitric oxide with L-NAME impairs the animal's ability to discriminate between similar odorants.

In insects, physiological studies in which $GABA_A$ receptors were blocked with picrotoxin indicate that fast feedback inhibition within the antennal lobe is essential for the generation of network oscillations during odor stimulation. Also, similar roles for feedback inhibition mediated by local interneurons have been established for many well-investigated brain models, including those for cortical and hippocampal gamma oscillations. In these models of synchronized network oscillations, spikes in excitatory cells trigger spikes in postsynaptic inhibitory neurons, whose spikes in turn produce inhibitory postsynaptic potentials in excitatory cells. This inhibition systematically delays and synchronizes the generation of the next excitatory response, thereby creating widespread network oscillations. The frequency of the oscillations depends on the time constant of inhibition and synaptic conductance strength. **Figure 1** presents an example of periodic oscillations recorded *in vivo* from the locust antennal lobe (**Figure 1(a)**) and comparable oscillations simulated by a computational network model of the antennal lobe (**Figure 1(c)**) that includes populations of randomly connected PNs and LNs stimulated by an external input (**Figure 1(b)**). The oscillations in the model were triggered by a constant (except for low-amplitude additive noise) depolarizing input to an odor-specific subset of antennal lobe neurons (**Figure 1(b)**) and were mediated by excitatory–inhibitory connections between PNs and LNs.

Oscillation-generating networks in a variety of species have been studied with computational models. Typically, the emergence of synchronized oscillations has been shown to depend on fast feedback inhibition

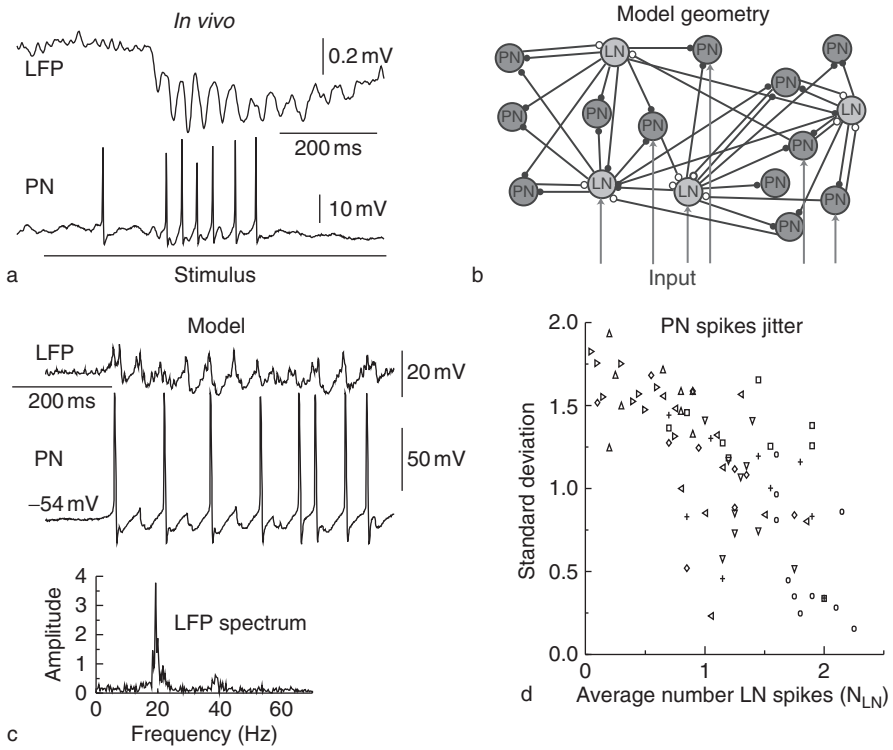

Figure 1 Synchronization of PN responses in the antennal lobe. (a) Typical odor responses recorded *in vivo*. Odor stimulation evoked oscillations in the local field potential (top, recorded in the mushroom bodies) and action potentials in a PN (bottom). (b) Antennal lobe model. (c) Averaged PN activity (LFP) in the model shows periodic oscillations at approximately 20 Hz during stimulation. (d) For each PN, standard deviation of PN spike phases across trials plotted against the total number of LN spikes in all presynaptic LNs. Adapted from Bazhenov M, Stopfer M, Rabinovich M, et al. (2001) Model of transient synchronization in the locust antennal lobe. *Neuron* 30: 553–567, with permission from Elsevier.

within a circuit. Ermentrout and colleagues described stimulus-evoked synchronization of olfactory neurons in a model of the *Limax* olfactory lobe that requires input from inhibitory interneurons. A model of spike synchronization in the honeybee antennal lobe, also dependent on local inhibition, has been proposed by Linster and Cleland to explain the experimentally observed impairment of sensory discrimination when inhibition is blocked. Also, a role for inhibition in memory formation in the *Limax* olfactory lobe has been proposed by Ermentrout and colleagues. In this model, global inhibitory connections between oscillating cells were proposed to suppress activity in all but a small population of neurons, thus forming a single and unique 'memory band.' In this model, odor localization was transient and became possible only after the network activity was synchronized.

Interestingly, antennal lobe oscillations are generally not evoked by the first presentation of an odorant. As shown by Stopfer and Laurent, oscillations emerge gradually during the course of multiple encounters with a given odorant, provided the encounters occur close together in time. Electrophysiological and computational evidence suggests that the gradual increase in oscillatory synchronization results from activity-dependent neural plasticity within the antennal lobe. Models of this mechanism proposed by Bazhenov and colleagues indicate that this plasticity, when engaged by repeating odor stimuli (as often occurs in nature), achieves robust stability against noise. Such stability is essential in a neural system that works to amplify small differences in input patterns.

Odor-induced oscillatory activity in the gamma (30–60 Hz) frequency range has been described in many mammals, including primates. Despite intensive studies, the primary mechanism for generating these oscillations is unclear. Olfactory bulb circuitry includes feedback inhibition mediated by granule cells. Interactions between granule and mitral cells in the OB appear to resemble those between the PNs and LNs of the insect antennal lobe. A number of computational models support the idea that excitatory–inhibitory interactions between mitral cells and granule cells could underlie gamma oscillations in the OB. This hypothesis is also supported by recordings from OB neurons under conditions that mimic natural odor stimuli. Electrical coupling between mitral cells whose primary dendrites form

synapses within the same glomeruli appears to enhance the synchrony of mitral cell oscillations. However, important questions remain. The relatively slow decay time of inhibition from granule cells may not be consistent with the synchronization underlying fast oscillations observed in mitral cells. A different type of model suggests that aperiodic but correlated input from widely branching granule cells could provide for the synchronization of mitral cells. Galan and colleagues showed that the slow inhibition from granule cells can lead to the synchronization of mitral cells in the gamma frequency range when partially correlated aperiodic fluctuations are fast and mitral cells fire at a rate that is roughly constant.

Computational modeling work has led to several alternative proposals for mechanisms for the synchronization of olfactory neurons in the OB. One mechanism that does not rely on synaptic interactions among neurons within the bulb but, rather, on common oscillatory drive from receptor cells has been suggested by Hopfield and Brody. According to this model, mitral cells synchronize when they receive similar inputs during odor stimulation. If this input represents the sum of common drive from receptors, and a random bias current which is different for different mitral cells is applied, then only a fraction of the mitral cell population would become synchronized during a given odor stimulation. This and number of other models suggest, therefore, that similar to antennal lobe dynamics, the patterns of synchronization across mitral cells are critical for odor representation in the OB.

A model of the OB based on chaotic attractors has been proposed by Skarda and Freeman. In this model, the resting state of the OB before odor recognition is chaotic; this state keeps the bulb far from equilibrium and allows a more rapid conversion to limit cycles (corresponding to previously reinforced odorants) during odor recognition.

In the locust, $GABA_A$-dependent synchronization of a given pair of PNs is generally transient. Usually after only a few hundred milliseconds of firing spikes that are phase-locked to the oscillations, the synchrony between action potentials in a given PN and the field potential oscillations is lost; then, other subsets of PNs become synchronized. However, the times during repeated trials at which a given PN will phase lock are reliable and odor specific. A number of hypotheses have been proposed to explain mechanisms behind such odor-specific, transient synchronization of the olfactory network. Computer models of the locust antennal lobe have suggested that the transient nature of PN synchronization could be explained by variations in inhibitory drive from inhibitory LNs over the duration of the odor-elicited response. In a model by Bazhenov and colleagues, as *in vivo*, a portion of the PNs become phase locked to the oscillatory local field potential at different times when driven by an odor; the identities of the PNs and the times at which they phase-lock change when a different stimulus is presented. When the activity of sets of LNs that are presynaptic to each PN in the network, and for each cycle of oscillation, was analyzed, it became clear that the total number of Ca^{2+} spikes in presynaptic LNs and the degree of coordinated firing in their PNs were highly correlated (**Figure 1(d)**). As feedback inhibition increased in strength, the timing precision of PN spikes across trials also increased: Jitter in PN spike timing decreased (**Figure 1(d)**). Note that in this model, inhibition only controls the timing of PN spikes when PN firing is induced by excitatory input from receptors. In other studies, disinhibition complimented by intrinsic conductances could trigger firing in PNs. Factors such as lateral LN–LN inhibition, spike adaptation in LNs, and the identities of activated neurons contribute to the variations in LN activity in this model, consistent with other theoretical studies.

A number of computational models have also investigated a potential role for lateral inhibitory projections in contrast enhancement – a process hypothesized to amplify differences between sensory representations, improving the signal-to-noise ratio. In these models, stronger inputs typically suppress nearby, weaker inputs. Some models of the OB implement contrast enhancement either at the level of lateral mitral cells dendrites (by granule cells) or at the glomerular layer (by periglomerular cells). It was shown by Linster and Hasselmo that the latter mechanism could help keep the number of active mitral cells stable, independent of the intensity of the olfactory input. Earlier models explored a potential role for lateral inhibition in promoting competition among olfactory glomeruli, and a model of olfactory memory based on altering lateral inhibition in the honeybee's antennal lobe was proposed by Linster and Masson. Computational studies by Linster and colleagues suggest that the contrast enhancement between odorants in the honeybee antennal lobe is best achieved when inhibition between each pair of glomeruli is organized based on glomerular odor response profiles rather than on anatomical neighborhood relations. Further experimental and modeling work is needed to clarify the possible role of lateral inhibition in contrast enhancement in the olfactory system.

Slow Temporal Patterning of Olfactory Neuron Responses

Experiments performed *in vivo* have established that only the synchronization of PNs in the insects'

antennal lobe is eliminated by picrotoxin; the slower firing pattern structures (alternating periods of depolarization and hyperpolarization) remain intact. Such temporal structure has been described in the projection neurons of locusts and *Drosophila*, as well as in the mitral cells of vertebrates. To explain the origins of slow temporal patterns in the insect antennal lobe, two alternative hypotheses have been proposed and evaluated using computer models. According to a model by Bazhenov and colleagues, the slow inhibitory receptors between LNs and PNs, which are not sensitive to picrotoxin and which operate with a time constant of a few hundred milliseconds, could account for the stimulus-specific slow temporal patterning of PN responses. When slow inhibition was provided in this model, odor stimuli induced characteristic slow temporal patterns in the responses of individual cells. The pattern consisted of alternating depolarizing modes, during which Na^+ spikes were generated, and hyperpolarizing modes, during which only subthreshold oscillations were evident in PNs. Each mode usually lasted a few hundred milliseconds, in agreement with recordings made *in vivo*. The sequences in which these modes appeared differed among PNs and varied reliably with the stimulus. Blocking slow inhibitory receptors in the model, a test not currently possible in the locust, eliminated the slow temporal patterns; groups of PNs then showed similar responses that did not vary when a new stimulus was applied. Also, similar to findings obtained *in vivo*, the slow temporal patterns in individual PN responses remained essentially intact after the model's $GABA_A$ inhibition was blocked.

Work on *Drosophila* by Olsen et al. suggests that excitatory interactions within the antennal lobe, likely mediated by interneurons, also contribute to the spatiotemporal structure of olfactory responses.

In a model proposed by Sivan and Kopell, relatively slow calcium and calcium-dependent potassium channels of PNs, rather than slow inhibitory input from LNs, were used to provide for picrotoxin-resistant slow temporal patterning in projection neurons. This model, as well as the one by Bazhenov et al., displayed little change in the average firing rate of PNs after the blockade of fast inhibition, in agreement with experimental data from locust. Both approaches highlight the power of computational modeling because they lead to predictions that can be tested *in vivo*.

A possible role for inhibitory input from granule cells in controlling mitral cell firing rate has been explored in a simplified model of the OB by Linster and Gervais. This model demonstrated that modulation of granule cell inhibition can alter the firing frequencies of mitral cells and can also decrease overlap between output patterns. Reminiscent of models of the insect antennal lobe, this study suggests that alternations of granule cell activity over the duration of an odor response can mediate odor-specific slow temporal patterning in mitral cell activity.

In *Drosophila*, Wilson et al. have shown that the extent of activity and the complexity of temporal responses in groups of PNs exceed those of the sensory neurons that provide direct input. What is the functional significance of the temporal structure in olfactory neuron responses? As shown by Friedrich and Laurent, in zebra fish, the slow temporal patterning in mitral cells appears to play a major role in the decorrelation of odor representations. The responses of a population of mitral cells to chemically similar odorants were highly correlated immediately after the onset of the odor stimulation, but they became progressively less correlated over the duration of the odor response. This type of gradually evolving decorrelation was also tested by Bazhenov and colleagues with a computational model of the locust antennal lobe. When similar odors were presented, the response patterns of PNs at first greatly overlapped. Over the course of a few hundred milliseconds, overlaps between the distributed firing patterns representing similar odorants decreased, and the initial clusters of activated PNs disappeared; cross-correlations between PN activity patterns (calculated across all PNs) was reduced by 30–50%. These works suggest that the intrinsic dynamics of antennal lobe circuitry depend on interactions between PNs and LNs, and such dynamics can amplify small differences between similar odors, thus improving odor discriminability over time. Therefore, the evolving temporal structures appear to be important for the resolution of odor identity. However, behavioral evidence obtained from rodents suggests that some olfactory tasks can be achieved within the time of a single sniff, less than 150 ms. In another study, it has been proposed that spatial representation of odor in the OB evolves across a sniff; therefore, difficult discriminations of similar odors may require the olfactory system to 'wait' for later activated components.

Decoding Olfactory Information at Downstream Levels of Processing

In insects and vertebrates, output from relatively small populations of olfactory neurons in the antennal lobe or the OB fans out broadly to large interneuronal ensembles in the mushroom body or the olfactory cortex. Work on locust by Jortner et al., for example, demonstrates that half of all PNs synapse directly upon each KC, a connectivity scheme that maximizes the coding space available for odors.

Electrophysiological recordings from insects indicate that KCs of the mushroom body respond with high specificity to odors: A given odor induces a reliable response in only a very small subpopulation of KCs. Several mechanisms operating in parallel appear to contribute to this response specificity, including the intrinsic nonlinear membrane properties of the KCs. Intracellular recordings from KCs suggest that activation of certain intrinsic conductances can amplify and sharpen excitatory postsynaptic potential-in'duced depolarizations, thus generating narrow time windows during which KCs can integrate inputs from multiple PNs. Another mechanism, drawing on inhibitory input from lateral horn interneurons (LHIs), regulates the integration window as well: Feed-forward inhibition from LHIs, which are driven by PNs, faithfully follows, with a delay, the periodic excitation of KCs by PNs, thus creating independent and brief temporal windows over which KCs can summate their input from PNs. The existence of these relatively short integration windows suggests that KCs are extremely sensitive to synchronized inputs from the antennal lobe; KCs may respond only when sufficient numbers of input spikes arrive together.

A realistic computer model by Bazhenov and colleagues of an olfactory network endowed with coincidence detection properties was used to test the extent to which KCs can detect synchrony in the input spike trains from a population of PNs (**Figure 2(a)**). Each stimulus consisted of many spike trains, which were delivered to a single cell through a large set of distributed cholinergic synapses. Individual spike trains had temporal structures designed to match experimentally observed PN firing patterns and included transient synchrony among synaptic inputs. For each stimulus, only a very small fraction of KCs responded reliably with a Na^+ spike on each trial despite small variations in the inputs from one trial to the next (input spike number and timing were slightly varied) (**Figure 2(b)**). As *in vivo*, different KCs spiked at different times, and their spike timing depended on the timing of transient correlations between inputs. In most cases, however, KCs were silent or produced only a few random spikes when input spikes happened to coincide, exceeding the KC's threshold. Results from this computational model suggest that a network of KCs with a sufficiently small integration window can respond with a unique spatiotemporal pattern for different inputs, decoding the correlation structure of the input spike train.

Modeling work by Assisi et al. indicates that the integration window of KCs can vary with odor concentration. Higher concentrations of odors, which elicit greater oscillatory synchrony, also elicit stronger feedback inhibition from LHIs. The resulting shifting balance of coordinated excitation and inhibition maintains the sparseness of spiking in KCs over wide ranges of stimulus intensity.

Physiological and modeling results from Cassenaer and Laurent indicate that the precise timing established by antennal lobe circuitry is maintained not only within KCs but also by their neural followers in the beta lobe. Interestingly, odor-specific timing precision is maintained in the beta lobe neurons, in part, by spike timing-dependent plasticity.

Modeling studies of odor decoding in the mushroom body further suggest that a strategy of coincidence detection offers a number of advantages for encoding sensory information compared to a classical integrator model. The sensitivity of a coincidence detector to timing on a very fine scale provides an additional coding dimension that could allow the olfactory system to discriminate between similar odors, whose neural representations may differ only in the fine temporal structure of the antennal lobe input. In contrast, a decoding system integrating over longer timescales would have no such ability. Furthermore, because the coincidence detector model focuses on specific time windows within an oscillation cycle, in agreement with theoretical studies in other systems, it is thus less sensitive to uncorrelated noise in its input. When the KC model operating as a coincidence detector was explicitly compared to an integrator model possessing a large integrating window, the integrator model showed much greater sensitivity to the presence of random noise in the input (**Figure 2(c)**). In the locust, these coincidence detection mechanisms provide a powerful strategy with which to solve complex problems that every olfactory system encounters.

In the olfactory cortex, local inhibitory interneurons receive afferent input and provide feed-forward inhibition to the pyramidal cells, thus creating a circuit that can operate as a precise coincidence detector. This circuitry is similar to that found in insect mushroom body. It is unknown, however, whether the vertebrate olfactory system utilizes a similar coding strategy. Many types of studies in this rapidly developing field are under way. Studies in mice have shown that afferent inputs from mitral cells connected to specific OB glomeruli terminate in broad patches which overlap with the terminal patches of other glomeruli. As a result, individual odorants are represented by subsets of sparsely distributed cortical neurons. Individual pyramidal cell of the olfactory cortex can make a few thousand connections to other pyramidal cells, further facilitating the redistribution of afferent activity patterns. Furthermore, molecular analysis strategies applied to mouse olfactory cortex have revealed

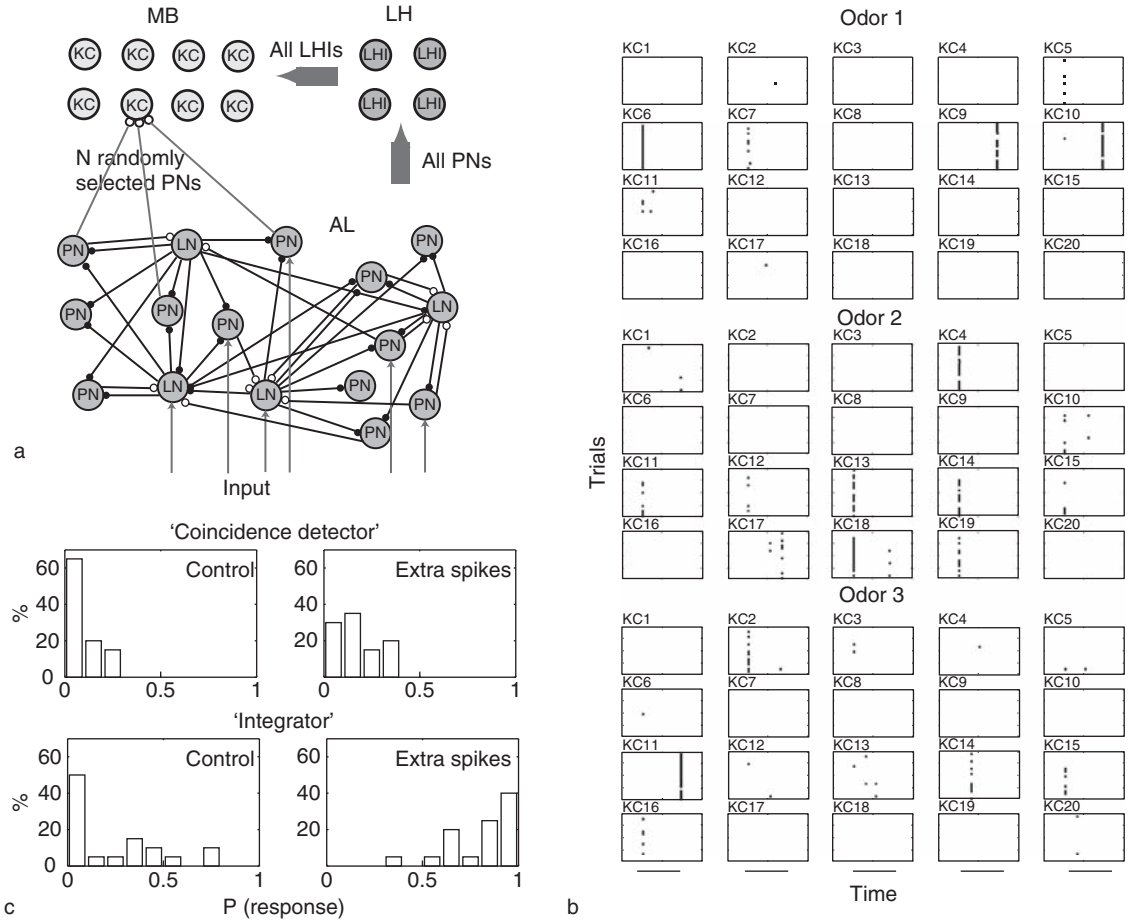

Figure 2 Decoding olfactory information in the mushroom body (MB). (a) Olfactory model. (b) KC responses for odor stimuli. Each box represents one KC operating as a coincidence detector; 20 cells are shown. Each stimulus (1 s duration; bars) was presented 20 times (raster) and included small variations between trials. Two to four KCs fired reliably in response for each stimulus. Identities of these active KCs changed between odors. (c) Frequency distribution of KC response probabilities before and after noise (extra spikes) was added to the input. The coincidence detector model shows only a slight increase in response probability, whereas the integrator model exhibits a drastic reduction in response specificity. Adapted from Perez-Orive J, Bazhenov M, and Laurent G (2004) Intrinsic and circuit properties favor coincidence detection for decoding oscillatory input. *Journal of Neuroscience* 24(26): 6037–6047, with permission from the Society for Neuroscience.

that binary odorant mixtures activate many cortical neurons that do not respond to the individual components of the mixture. This finding supports a coincidence detection function for the olfactory neurons. However, electrophysiological work in slice preparations has revealed that relatively small numbers of mitral cells are sometimes able to activate individual pyramidal cells in rat olfactory cortex. Based on this finding, it has been proposed that complex odors are represented in olfactory cortex in a broadly distributed manner. It is unknown to what extent the temporal structure found in activity patterns of the OB neurons is reflected in olfactory cortex. Multichannel and single-unit electrophysiological recordings made in the rat olfactory cortex suggest that individual odorants evoke activity in a spatially scattered ensemble of olfactory cortex neurons, and the ensemble activity includes a rich temporal structure.

The literature regarding modeling of the information processing in the olfactory cortex is growing but is currently rather limited. Aforementioned work by Hopfield and Brody incorporates models of the cortical cell population with individual units operating as coincidence detectors on the input from the OB. Results from a study by Wilson and Bower, using a large-scale computer model of the olfactory cortex, suggest that during each 40 Hz cycle of EEG activity, there is a convergence of afferent information from the OB and from caudal cortex. This model reproduced patterns of activity recorded physiologically in response to electrical shocks to the afferent input. A series of models simulating cortical EEG dynamics were proposed by Walter Freeman.

Conclusion

Experimental and modeling studies of odor coding have developed, extended, and tested a wide range of ideas about information processing by networks of neurons. Computational models of the insect olfactory system have explored the proposal that antennal lobe circuitry reformats information about odors, distributing it into the identities of PNs responding to the odor and into the relative timing of spikes in those PNs. Although temporal patterning may not be required for relatively easy tasks, such as the discrimination of ensembles of PNs that do not significantly overlap with one another (activated, perhaps, by odorants with chemically distinct structures), temporal features appear to be critical when discrimination tasks are more challenging – when it is necessary to discriminate activated PN ensembles that substantially overlap. Thus, the antennal lobe actively sharpens information about the olfactory stimulus by employing time and transient synchrony as coding dimensions. The massive convergence of afferents from the olfactory epithelium to the antennal lobe contributes to the creation of reliable and reproducible odor-specific patterns of activity in the antennal lobe. The intrinsic dynamic properties of the antennal lobe optimize these spatiotemporal structures, helping to reduce overlap between representations of similar odors. Contained within the spatiotemporal patterns of antennal lobe activation is information not only about odor identity but also about odor concentration. KCs of the mushroom body, postsynaptic targets of the PN ensemble, decode the stimulus-specific spatiotemporal patterns of PN activity. Operating as coincidence detectors, and thus employing the useful coding dimension of precise timing, KCs transform spatiotemporal patterns of antennal lobe activity into new and very sparse distributed firing sequences. The very large number of KCs enables the insect olfactory system to minimize the overlap between response patterns elicited by even very similar odors. Thus, temporal coding implemented by the antennal lobe can magnify differences in odor representations while minimizing the influence of noise, dramatically reducing odor code overlap in the mushroom body compared to the initial coding by olfactory receptor cells. Understanding and testing hypotheses about these complex biological processes has been greatly facilitated by the use of computational models.

See also: Neural Synchrony and Feature Binding; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory System: Circuit Dynamics and Neural Coding in the Locust; Population Coding.

Further Reading

Buck LB (1996) Information coding in the vertebrate olfactory system. *Annual Review of Neuroscience* 19: 517–544.

Cleland TA and Linster C (2005) Computation in the olfactory system. *Chemical Senses* 30(9): 801–813.

Kay LM and Stopfer M (2006) Information processing in the olfactory systems of insects and vertebrates. *Seminars in Cell & Developmental Biology* 17: 433–442.

Laurent G (1999) A systems perspective on early olfactory coding. *Science* 286(5440): 723–728.

Laurent G (2002) Olfactory network dynamics and the coding of multidimensional signals. *Nature Reviews Neuroscience* 3: 884–895.

Shepherd GM, Chen WR, and Greer CA (2004) Olfactory bulb. In: Shepherd GM (ed.) *Synaptic Organization of the Brain*, 5th edn., pp. 165–216. Oxford: Oxford University Press.

Relevant Website

http://senselab.med.yale.edu – Senselab.

Olfactory Cortex Physiology

D A Wilson, University of Oklahoma, Norman, OK, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The olfactory cortex is defined as that cortical region of the mammalian brain that receives direct afferent input from the olfactory bulb. The olfactory cortex can be divided into multiple subregions based on anatomical characteristics, the largest of which is the piriform cortex (PCx, also known as the pyriform or prepyriform cortex). The PCx has received the greatest attention of any olfactory cortical region, and thus PCx physiology is the focus of this description.

The PCx plays a critical role in odor discrimination and perception, synthetic processing of complex odorant mixtures, experience- and state-dependent sensory gating, short-term odor habituation, and odor memory. Afferent and intrinsic glutamatergic synapses demonstrate both a short- and long-term plasticity, contributing to these processes. The specific neural architecture of this paleocortical structure (not described here) also contributes significantly to its function.

Circuit Structure

An overview of PCx circuitry is necessary to provide a reference for physiological function. In brief, the primary afferent to the PCx is via mitral and tufted cell axons, which form the lateral olfactory tract. This input terminates on the distal half (layer Ia) of the apical dendrites of the principal cortical neurons, the pyramidal cells, the somata of which lie in layers II and III. The inner half of these dendrites (layer Ib) is dominated by intrinsic and commissural fibers. The lateral olfactory tract extends approximately to the midpoint of the rostral–caudal extent of the PCx, and its termination point is an anatomical marker for division of the PCx into anterior and posterior regions. The posterior PCx receives afferent input from distributed mitral/tufted cells axons, though the termination zone for afferent input is reduced to a thinner layer in posterior, compared to anterior, regions. In addition to afferent input from the olfactory bulb, the PCx also receives input from the entorhinal and orbitofrontal cortices, and input from several neuromodulatory systems, including the noradrenergic nucleus locus coeruleus and the cholinergic nucleus of the horizontal limb of the diagonal band of Broca. Finally, in addition to the principal pyramidal cells, the PCx contains several classes of inhibitory interneurons, including feedforward inhibitory interneurons in layer I, which receive direct input from the lateral olfactory tract, and feedback and lateral inhibitory interneurons in layer III, which receive input from PCx pyramidal cells. Pyramidal cells receive γ-aminobutyric acid (GABA)ergic inhibitory inputs on both their somata and apical dendrites (Figure 1).

Synaptic Physiology

Afferent input to the PCx evokes a sequence of synaptic actions on pyramidal neurons (Figure 2). The glutamatergic afferent input, though limited to distal apical synapses, is sufficiently strong to allow spike generation evoked by simultaneous input from just a few afferent fibers. This could provide an early mechanism for odorant feature synthesis. Spiking evoked by the afferent excitatory postsynaptic potential (EPSP), however, is controlled by an early-onset GABAergic, Cl^--mediated, somatic inhibitory postsynaptic potential (IPSP) that is depolarizing at resting potential in these cells and serves to shunt excitatory currents. In addition, the EPSP is followed by a longer (several hundred ms) K^+-mediated dendritic IPSP. The myelinated lateral olfactory tract provides nearly simultaneous activation of pyramidal neurons within anterior PCx. However, due to the fact that posterior PCx is innervated by long axon collaterals posterior to the termination of the lateral olfactory tract, the activation comes as a rostral-to-caudal wave.

If pyramidal cell spikes are evoked in response to afferent input, a second wave of excitation is generated via the pyramidal cell-mediated intrinsic fiber system. Individual pyramidal neurons are synaptically connected to several thousand other neurons throughout the PCx (and neighboring regions, including the olfactory bulb), thus providing a mechanism for rapid and widespread distribution of afferent input. The intrinsic fiber system provides a second substrate for convergence of information regarding multiple odorant features.

There is some evidence that the anatomical segregation of afferent and intrinsic fiber synapses on distal and proximal apical dendrites, respectively, provides a substrate for differential modulation of these two excitatory pathways. Thus, for example, both norepinephrine and acetylcholine appear to preferentially,

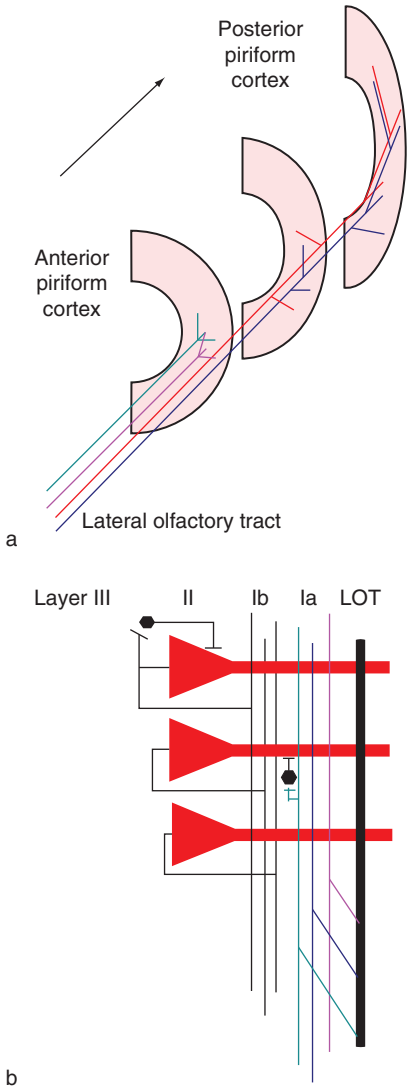

Figure 1 Basic piriform cortex circuitry. (a) Mitral/tufted cell axons conveying information from specific olfactory receptor types terminate in overlapping patches in anterior piriform cortex, and in broader, more overlapping regions in posterior piriform cortex. (b) Mitral/tufted cell axons form the lateral olfactory tract (LOT), which runs along the superficial layer of the anterior piriform cortex. Afferent axon branches (blue line) leave the lateral olfactory tract and terminate on the distal apical dendrites of piriform cortex pyramidal neurons (red) and also on a small population of feedforward inhibitory interneurons. Pyramidal cell axons form a dense, broadly projecting network of intrinsic association fibers (green line), which terminate on the proximal half of the pyramidal dendritic tree and on basal dendrites (not shown).

though not exclusively, modulate intrinsic fiber synapses compared to afferent synapses within PCx.

The combination of an afferent input-generated EPSP–IPSP sequence and the concomitant activation of intrinsic fiber-mediated spread of EPSPs can lead to circuit oscillations. Activity within the PCx expresses a theta band (2–10 Hz) frequency oscillation phasic with the respiratory cycle, as well as beta (10–40 Hz) and gamma (40–100 Hz) band oscillations. The higher frequency oscillations can occur as waves of activity riding the slower respiration-entrained oscillations (Figure 2). Beta frequency oscillations may originate within the PCx, and can then be transmitted to the olfactory bulb. Isolation of the bulb from the PCx eliminates olfactory bulb beta band oscillations. Gamma frequency oscillations have been further functionally divided into higher and lower subbands. These high-frequency beta and gamma oscillations are influenced by descending input from limbic and other cortical structures, as well as neuromodulatory inputs such as acetylcholine. High-frequency PCx oscillations may be indicative of, and correlate with, behavioral state, expectation, and stimulus familiarity/novelty. Their specific role in encoding stimulus identity is controversial.

The oscillatory nature of stimulus-evoked PCx activity may also contribute to neural plasticity in this system by enhancing synchrony between afferent and intrinsic fiber activity on sequential oscillatory cycles. Both afferent and intrinsic fiber synapses express use-dependent plasticity. Associative long-term synaptic potentiation (LTP) has been described at both afferent and intrinsic synapses, though LTP at intrinsic synapses appears most robust. LTP in the PCx is *N*-methyl-D-aspartate (NMDA) receptor dependent and can be modulated in a state-dependent manner through cholinergic regulation of GABAergic inhibition. Activation of cholinergic muscarinic receptors reduces GABAergic inhibition and enhances NMDA receptor-dependent LTP in PCx.

In addition to LTP, afferent synapses display an activity-dependent short-term depression, mediated by presynaptic group III metabotropic glutamate receptors. As described in the following section, this short-term depression contributes to cortical odor adaptation and short-term habituation.

Sensory Physiology

The PCx is generally considered the primary sensory cortex for olfaction, though it may also serve roles comparable to those of the higher order cortex of thalamocortical sensory systems. Cutting the lateral olfactory tract (LOT) input to PCx disrupts odor discrimination, and the PCx projects to limbic, hypothalamic, and neocortical regions that may underlie behavioral and autonomic responses to odor stimulation.

Odor stimulation single-unit and local field potential responses have been recorded in the PCx of awake and anesthetized animals. Odor-evoked activity can also be observed in the human PCx using functional magnetic resonance imaging (fMRI) techniques.

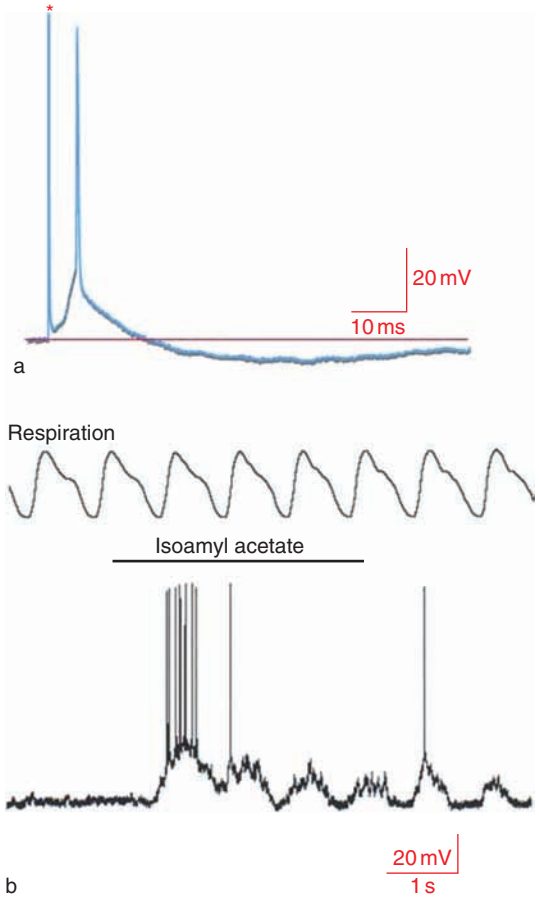

Figure 2 Lateral olfactory tract-evoked and odor-evoked responses in rat piriform cortex. (a) Electrical stimulation of the lateral olfactory tract (red asterisk) evokes a monosynaptic excitatory postsynaptic potential in layer II/III pyramidal neurons. This excitatory postsynaptic potential is followed by a K^+-mediated inhibitory postsynaptic potential. A short-latency Cl^--mediated inhibitory postsynaptic potential (IPSP) precedes the slower inhibitory postsynaptic potential (not visible in this record at resting membrane potential, −72 mV). (b) Odor-evoked responses recorded intracellularly in an anesthetized rat show strong depolarizations in phase with the respiratory cycle. As in extracellular field potential recordings, these intracellular depolarizations may show high-frequency oscillatory components.

While olfactory receptor input to the olfactory bulbs is exclusively ipsilateral, individual neurons within the PCx can respond to odor stimulation from either the ipsilateral or contralateral receptor sheet, with different neurons expressing different relative lateralized responsiveness. PCx neurons in rodents also respond to a variety of nonolfactory stimuli, such as during approach to the odor sampling port, during consummation of water reward, or in response to foot shock, though the specific sensory nature of these responses is not clear. For example, given the modulatory inputs to the PCx and to the olfactory bulb, unit activity in PCx may change due to changes in arousal or behavioral state, rather than as a direct consequence of a specific sensory input. Nonetheless, activity changes within the PCx (and olfactory bulb) do not solely reflect primary odor input.

Single units in the PCx can display either excitatory or suppressive responses to odorants (Figure 2). Responses may be phasic with the respiratory cycle or more prolonged, extending across multiple cycles. Excitatory responses are associated with large depolarizing postsynaptic potentials giving rise to evoked spiking. Instantaneous evoked firing rates can attain 300 Hz during single inhalations. Combined with the theta frequency of the respiratory cycle, this pattern of activation (high-frequency bursts at theta frequency) is ideal for inducing LTP. Individual cells respond to multiple odorants, and based on both single-unit recordings and c-fos mapping studies, neurons responsive to a particular odorant are widely distributed throughout PCx. Neighboring cells may respond to different odorants. The distributed nature of odorant-evoked responses is consistent with patchy mitral cell termination zones and the broadly dispersed intrinsic fiber system, and suggests that odorant identity is encoded by a broad ensemble of PCx neurons.

The overlapping patches of input from mitral cells conveying information from specific olfactory receptors, and the amplification of this overlap and convergence via intrinsic fibers, provide an anatomical basis for synthesis of multiple odorant feature input within the PCx. This is particularly relevant given that most naturally occurring odors are mixtures of many different molecular species. Using immunohistochemical techniques, evidence has been presented suggesting that some single neurons may respond only to combinations of specific odorants, and not to the individual components of those mixtures. However, these combination-specific neurons have not been observed with electrophysiological techniques to date. In anesthetized rats, there are significant differences in single-unit activity between the anterior and posterior PCx. Posterior PCx neurons have lower average spontaneous activity and less intense odor-evoked activity compared to anterior PCx neurons. The anterior PCx can be further subdivided into dorsal and ventral regions based on local circuit anatomy. Immunohistochemical techniques suggest enhanced odor responsiveness in the dorsal anterior PCx compared to ventral anterior PCx.

In addition to their role in odor quality coding, PCx single units may encode odorant intensity. This topic has received sparse attention, but there may be a rostral-to-caudal gradient in PCx unit sensitivity to odorants, with more rostral regions showing greatest sensitivity. Based on functional imaging techniques and single-unit recording, increases in odorant

intensity activate increasingly more caudal regions of the anterior PCx. This gradient of sensitivity could reflect the gradient of mitral and tufted cell projections to the PCx, wherein tufted cells tend to terminate in more rostral regions while mitral cells terminate throughout. Tufted cells, especially superficial tufted cells, have a lower threshold for olfactory nerve input, and thus could produce an enhanced sensitivity of rostral regions of the anterior PCx.

Single-unit and local field potential responses to odorants are strongly influenced by behavioral state and past experience. Behavioral state (e.g., hunger/satiety or sleep/wakefulness) may modulate PCx responses either directly and/or via upstream modulation of olfactory bulb activity. Regulation of cortical responses by behavioral state is at least partially mediated by cholinergic input from the horizontal limb of the diagonal band of Broca. For example, pharmacological blockade of muscarinic receptors can eliminate the hunger/satiety modulation of PCx responses to food odor.

A shift from fast-wave to slow-wave neocortical activity in anesthetized rats, nominally similar to a shift from wakefulness to deep sleep, reduces PCx responsiveness to odorants. This state-dependent sensory gating in PCx is similar to that seen in the thalamus in thalamocortical sensory systems. PCx neurons during slow-wave state are hyperpolarized, with large membrane potential fluctuations. In fast-wave state, membrane potential is relatively depolarized and stable, allowing afferent input to more reliably evoke activity. The state-dependent change in membrane potential and PCx odor responsiveness is presumably due to fluctuations in modulatory inputs such as ACh, though this has not yet been determined.

Odorant-evoked PCx activity also reflects recent odor experience. As noted earlier, afferent inputs express a use-dependent, metabotropic glutamate receptor-mediated presynaptic depression. Prolonged or rapidly repeated odorant stimulation induces profound adaptation of cortical responses, despite relatively maintained olfactory bulb activity. Blockade of these presynaptic receptors can prevent cortical odorant adaptation and habituation of odorant-evoked autonomic reflexes. This presynaptic depression is also regulated by noradrenergic β-receptors, which could contribute to state- or attention-mediated blockade of adaptation to biologically important stimuli. PCx odor adaptation is highly odorant specific, and has been demonstrated to serve as a mechanism for odorant figure–ground separation when an odor is presented against a relatively stable background. The homosynaptic afferent depression allows the PCx to selectively respond to those dynamic components of the inhaled stimulus, while filtering the more stable, potentially background components.

Finally, PCx physiology and odorant-evoked responses are affected by odorant familiarity and associative conditioning, and some aspects of this plasticity may vary between the anterior and posterior PCx. Associative conditioning modifies the strength of odor-evoked local field potentials and the coherence of oscillations between different subregions of the piriform cortex and olfactory bulb. Using electrical stimulation of the olfactory bulb or lateral olfactory tract as artificial odors in a conditioning task enhances the PCx response evoked by that electrical stimulation, but only if the animal successfully learns the discrimination. Together, these two findings suggest that, as animals learn about odors and their meanings, synaptic strengths of afferent and intrinsic connections are modified in the PCx, leaving a long-term trace of the learned odor. The success of olfactory learning can be impaired by NMDA antagonists, suggesting that NMDA-dependent LTP is involved in the learned synaptic plasticity. Furthermore, neuromodulators such as acetylcholine and norepinephrine are also required for cortical changes.

Learned changes in PCx function not only store information about the learned odor, but also affect the discriminability of the familiar odor. Perceptual learning is an increase in discriminability with sensory experience and occurs in both human and nonhuman olfactory systems. Familiar odorants are more easily discriminated from molecularly similar odorants than from novel odorants. Perceptual learning is associated with enhanced PCx single-unit odorant discrimination. In cross-adaptation tests of discrimination, familiar odorants are more easily discriminated by anterior PCx single units, and single-unit populations, than are novel odors. Importantly, experience with odorant mixtures produces enhanced discrimination of those mixtures from their components by anterior PCx neurons. This synthesis of mixtures into stimuli distinct from their components is consistent with the synthetic or configural perception of odor mixtures and the impaired behavioral ability to correctly identify components of complex mixtures. Thus, an experience-dependent merging of multiple odorants and/or odorant features into unique perceptual odor objects maybe a critical function of the PCx. Perceptual learning at both the behavioral and cortical level is impaired by the cholinergic antagonist scopolamine.

Thus, as odors are learned, not only does the PCx store a long-term trace of that odor for subsequent recognition, but also anterior PCx encoding of the odorant is modified to enhance stimulus discriminability. In addition to these two changes following

odor learning, a series of short-term changes occur in PCx that increase the ability to learn new odors. Olfactory learning displays a form of rule learning, or learning-to-learn phenomenon, wherein learning one or more odor discriminations enhances the speed at which subsequent discriminations are learned. Odor rule learning is associated with increased excitability of PCx pyramidal cells, increased pyramidal cell apical dendrite spine density, and enhanced presynaptic transmitter release from intrinsic fibers – a series of effects that facilitate induction of long-term synaptic plasticity. Importantly, these changes are transient, lasting for at most a few days after the odor learning experience. Thus, a relatively rapid series of learning experiences increases the ability of the PCx to learn new odors, an advantage that fades after several days. As with the other forms of learning described earlier, olfactory rule learning is acetylcholine dependent.

Functional Implications

Based on cortical architecture, theoretical and computational models suggest that the PCx functions

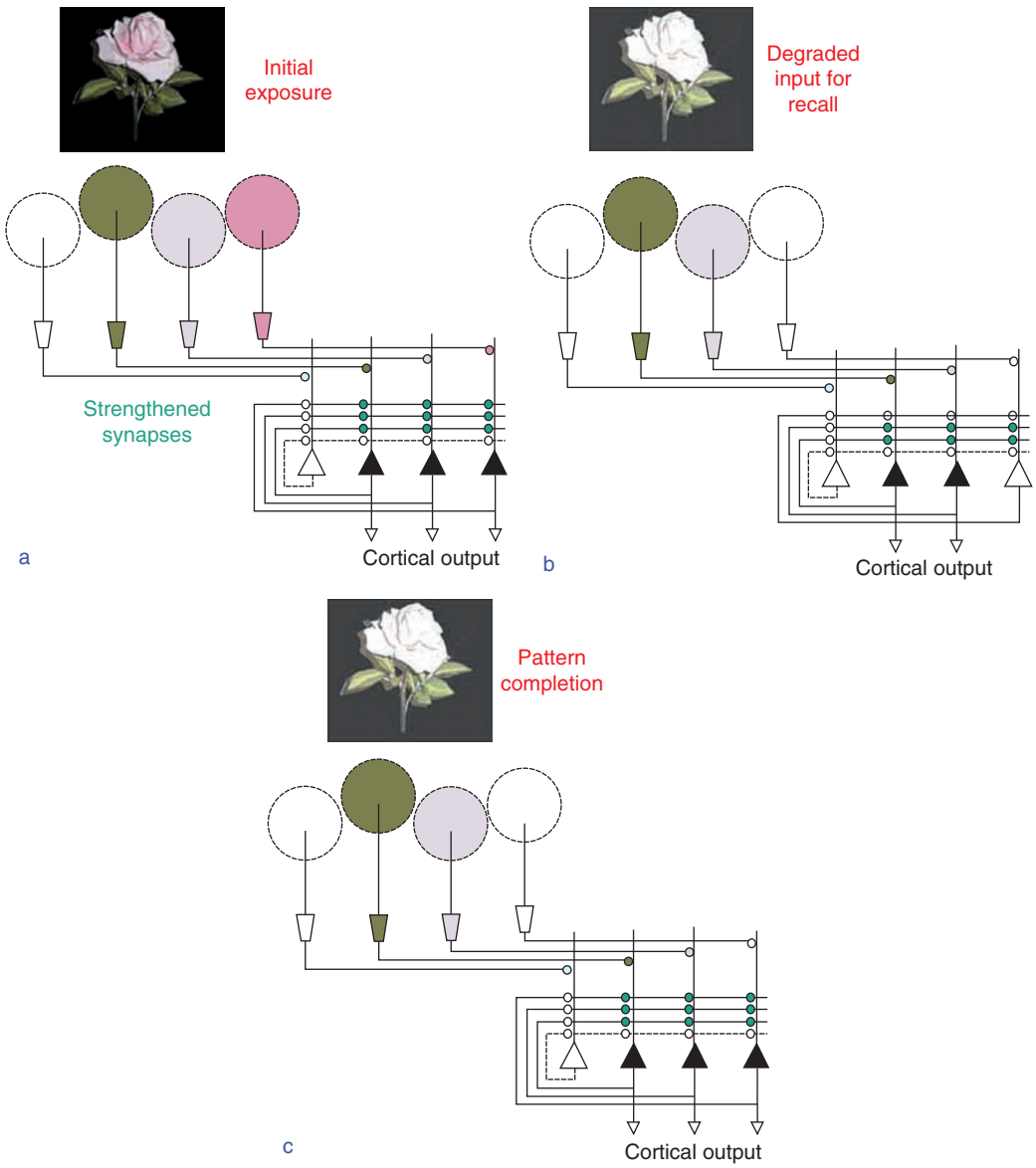

Figure 3 Autoassociative network. (a) Initial exposure to an olfactory stimulus activates specific olfactory bulb glomerular and mitral/tufted cell patterns (shaded circular glomeruli and trapezoidal mitral cells), which in turn, via patchy terminations in the piriform cortex, activate a pattern of pyramidal cells. The coincident excitation of specific afferent and intrinsic association fiber synapses (blue dots) induces long-term potentiation, effectively stored as a template of the experienced pattern. (b) On reexposure to the same or partially degraded stimulus, the potentiated intrinsic association fiber synapses allow pattern completion (c) and a cortical output that matches the original, signifying stimulus recognition.

as an autoassociative, content-addressable memory system. Patterns of afferent input are stored through changes in weight of intrinsic association fiber synapses. Information is encoded by an ensemble of distributed, interconnected neurons, and familiar input patterns are matched with existing, stored templates through the enhanced retrieval of patterns with strengthened synapses. Increasingly, physiological evidence supports this model.

Autoassociative, content-addressable memory systems are capable of generating stable output patterns, given complex and variable inputs. For example, noisy or reduced inputs can reliably be reconstructed into completed output patterns, due to the activation of strengthened synapses within the stored template (Figure 3). Assuming a major function of the PCx is to discriminate and recognize odors, the autoassociative system is especially useful for at least two reasons. First, most naturally occurring odors are complex mixtures of different molecules, and individual olfactory receptor neurons and olfactory bulb mitral cells respond to multiple odorant molecules. This leads to a complex input pattern to the PCx, where information concerning odor identity or quality is encoded in the combinatorial pattern, rather than in the activity of a specified individual neuron or neuron class. Thus, complex pattern recognition by PCx circuits becomes critical for odor recognition and discrimination – a feat performed well by the autoassociative network. Second, given the complexity of odorant mixtures and the frequent presence of background odorants, a particular odor may vary in its precise combination or relative strengths of individual components, yet needs to be consistently identified. Again, the autoassociative network functions well in this circumstance to reconstruct the complete odor from degraded input patterns.

If the autoassociative, content-addressable memory model is applicable to PCx function then several predictions about PCx physiology can be made. First, PCx circuits must be capable of expressing use-dependent associative plasticity necessary for storing previous input patterns. As described earlier, several forms of synaptic plasticity have been described at PCx afferent and intrinsic fiber synapses, and this plasticity has been linked to behavioral odor memory. Second, discriminability of odors should improve with familiarity as those input patterns become stored. In support of this, perceptual learning has been described both behaviorally and in the ability of single PCx neurons to discriminate similar odorants. Pharmacological manipulations (e.g., cholinergic blockade) that disrupt normal PCx synaptic plasticity impair both behavioral and cortical experience-dependent improvement in odor discrimination. Third, the stored and reconstructed patterns may include more than simple olfactory-based afferent input. Given that the PCx receives extensive descending input from entorhinal and orbitofrontal cortex, stored patterns may include multimodal, contextual, or state-dependent information. This could allow context and expectation to influence the PCx output (and consequent perception) resulting from a given pattern of olfactory receptor activation. Behavioral evidence supports this effect (e.g., differentially coloring liquids can influence the odors assigned to them), and more recent local field potential recordings suggest that expectation and experience strongly influence PCx activity.

See also: Chemical Senses: Overview; Olfactory Cortex: Comparative Anatomy; Olfactory Bulb Mapping; Olfactory Higher Centers Anatomy; Olfactory System Theory; Sensory Aging: Chemical Senses.

Further Reading

Barkai E (2005) Dynamics of learning-induced cellular modifications in the cortex. *Biological Cybernetics* 92: 360–366.

Fontanini A and Bower JM (2006) Slow-waves in the olfactory system: An olfactory perspective on cortical rhythms. *Trends in Neurosciences* 29: 429–437.

Granger R and Lynch G (1991) Higher olfactory processes: Perceptual learning and memory. *Current Opinion in Neurobiology* 1: 209–214.

Haberly LB (2001) Parallel-distributed processing in olfactory cortex: New insights from morphological and physiological analysis of neuronal circuitry. *Chemical Senses* 26: 551–576.

Hasselmo ME, Wilson MA, Anderson BP, et al. (1990) Associative memory function in piriform (olfactory) cortex: Computational modeling and neuropharmacology. *Cold Spring Harbor Symposium on Quantitative Biology* 55: 599–610.

Neville KR and Haberly LB (2004) Olfactory cortex. In: Shepherd GM (ed.) *The Synaptic Organization of the Brain*, pp. 415–454. New York: Oxford University Press.

Wilson DA, Best AR, and Sullivan RM (2004) Plasticity in the olfactory system: Lessons for the neurobiology of memory. *Neuroscientist* 10: 513–524.

Wilson DA and Stevenson RJ (2006) *Learning to Smell: Olfactory Perception from Neurobiology to Behavior.* Baltimore: Johns Hopkins University Press.

Olfactory Cortex: Comparative Anatomy

K R Illig, University of Virginia, Charlottesville, VA, USA
D A Wilson, University of Oklahoma, Norman, OK, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The vertebrate olfactory system faces an enormous challenge. When one or a mixture of a seemingly limitless set of odorous molecules is encountered, the olfactory system must identify it, discriminate it from similar odors, recall information that has already been learned about the odor's significance, and incorporate any new information that might accompany the current presentation. Understanding the transfer of chemical information and the circuitry involved at the initial stages of processing yields some clues about how the olfactory system performs such tasks, but it is at the level of the olfactory cortex – complex, often multimodal sensory and association areas characterized by a high degree of convergence and interconnectedness – that odor information is integrated with other sensory cues, learned associations, and internal motivational states to generate a meaningful response to an odor.

This comparative review of the vertebrate olfactory cortical areas first examines what is known about the olfactory cortical areas in rodents, which have been the most well studied and widely used vertebrate experimental subjects in olfactory research. The structure and organization of the two major olfactory cortical areas, the piriform cortex and the anterior olfactory nucleus, are examined in detail, and the afferent, reciprocal, and efferent connections that these two areas have with other cortical structures are considered. Next, the structure and organization of the rodent olfactory cortex are discussed in relation to other mammalian and nonmammalian vertebrate species, and the article concludes with remarks about the potential function of the cortical areas in the vertebrate olfactory system.

What Is Olfactory Cortex?

In the broadest sense, olfactory cortex is defined as any cortical region that receives information from the olfactory system. The first structure that could be designated in such a way is the main olfactory bulb, which is the target for incoming information from the olfactory sensory neurons in the olfactory epithelium. Although the olfactory bulb is not usually regarded as a cortical structure, it has the features considered by neuroanatomists to characterize cortex: it is divisible into more than two tangential layers, the most superficial layer is a plexiform layer of incoming afferent fibers; it is connected by a network of neuronal processes that give rise to columnar functional units, and these functional units can interact with each other both within and across laminae. Regardless of whether it can be termed olfactory 'cortex,' the olfactory bulb is the first stage of olfactory information processing in the vertebrate central nervous system, and a great deal of synthesis and refinement of the olfactory signal occurs at this stage.

The two classes of principal cells in the olfactory bulb, the mitral and the tufted cells, project to a number of cortical structures, and these third-order structures are traditionally designated olfactory cortex (Figure 1). The largest of these is the piriform cortex, which is often referred to as 'primary olfactory cortex,' although this designation has more to do with its size than its functional similarities to other primary sensory cortices. The piriform cortex is also designated the pyriform or prepiriform cortex. The second-largest cortical recipient of mitral and tufted cell fibers is the anterior olfactory nucleus (AON), a large structure (cortical in nature, despite its name) situated within the olfactory peduncle between the olfactory bulb and the piriform cortex. Although the AON has received relatively little experimental attention, it likely plays a central role in processing olfactory information. These two cortical structures will be the focus of this review.

There are other cortical areas that receive direct input from the olfactory bulb, including the olfactory tubercle, lateral entorhinal cortex, prefrontal cortex, and cortical areas associated with the amygdala, although the direct olfactory bulb input to these areas often is nominal. Many of these areas also have reciprocal projections with the piriform cortex and the AON and can thus be designated, along with perirhinal, perihippocampal, and orbitofrontal cortices, as members of higher-order olfactory or multimodal cortical circuits.

Rodent Piriform Cortex

Organization

The piriform cortex is the largest cortical recipient of direct olfactory bulb projections, and it is also a very prominent part of the rodent brain, accounting for up to 10% of cortical volume in some species. Although it is generally agreed that the area is involved in higher-order processing of olfactory information, its function appears to be complex, and is not altogether well understood. Unlike 'primary' cortical structures

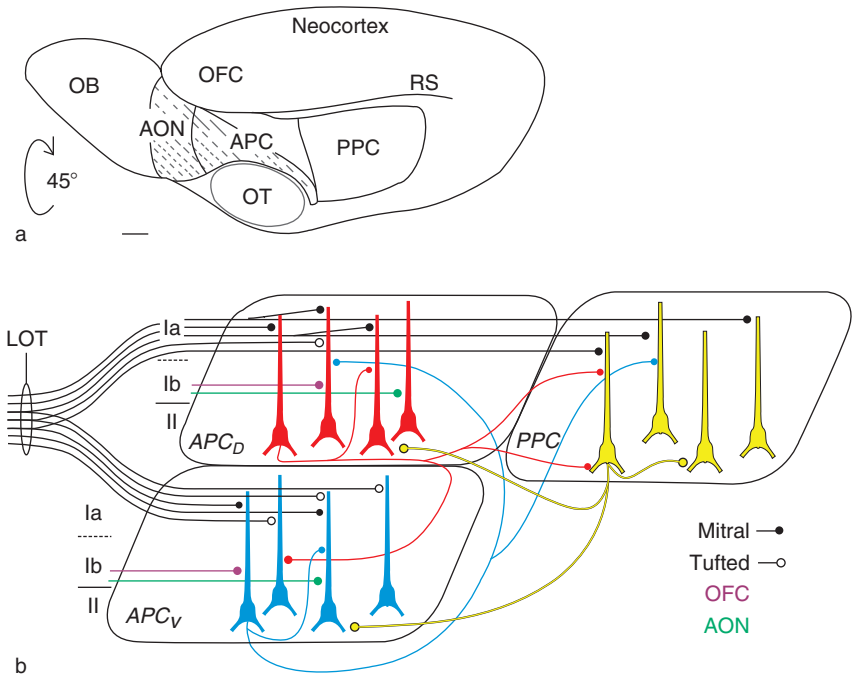

Figure 1 Schematic diagram showing the relative location and connectivity of olfactory cortical areas in the rodent. (a) Ventrolateral view of the rat brain, with the dorsal surface of the brain rotated 45° medially, showing the locations of the olfactory bulb (OB), anterior olfactory nucleus (AON), and the anterior (APC) and posterior (PPC) piriform cortices. The lateral olfactory tract (LOT) overlies the AON and a portion of the APC and is indicated by cross-hatching. Also shown are the orbitofrontal cortex (OFC) and olfactory tubercle (OT). The dorsal portion of the APC forms the ventromedial wall of the rhinal sulcus (RS). Scale bar = 1 mm. (b) Circuitry of the piriform cortex. Axon collaterals from the mitral and tufted cells exit the LOT to synapse on the distalmost portion of pyramidal cell apical dendrites in layer Ia of the piriform cortex. Although a light input from tufted cells is shown in dorsal APC (APC_D), it has been suggested that ventral APC (APC_V) may be the only region of piriform cortex that receives input from tufted cells. Tufted cell axons do not project to PPC. Axons from the OFC and from the AON project proximal to the input from the OB in layer Ib of APC. The remainder of input to pyramidal cell dendrites in layer Ib is from associative projections from piriform cortex, often originating in the same subregion as the target cell (so-called 'autoassociative' input). The projection from PPC onto other piriform pyramidal cells is primarily to basal dendrites in layer III. Other inputs to layer III, for example from the amygdala and OFC, are not shown.

in other sensory systems that are devoted to processing a single type of sensory information, the piriform cortex has a highly autoassociative architecture, with an extensive network of interconnections within and among its various subregions. Moreover, direct afferent fibers from the olfactory bulb make up only a small fraction of the input that neurons in piriform cortex receive on their dendritic tree; all other input to piriform cortex is from other neurons within piriform cortex, from higher-order areas such as orbitofrontal cortex or from structures that are most often associated with other systems or sensory modalities, such as entorhinal cortex and the amygdala.

A second feature of the rodent piriform cortex is its discrete laminar organization perpendicular to the cortical surface (Figures 1 and 2). Piriform cortex is three-layered paleocortex. The most superficial layer (layer I) is a highly ordered plexiform layer consisting of two parts: layer Ia, a superficial sublayer that includes axons from the mitral and tufted cells in the olfactory bulb, and a deeper layer, Ib, which consists of associative axons from other neurons in piriform cortex, from other ipsilateral cortical structures such as the AON, and from commissural axons from contralateral olfactory cortical neurons (Figure 1). The next deepest layer (layer II) is a compact cell body layer that houses the majority of the principal cells within piriform cortex. layer IIa is a superficial sublamina that houses a specialized population of pyramidal-type cells that lack basal dendrites, termed semilunar cells. The deeper layer, IIb, contains the cell bodies of superficial pyramidal cells and of small multipolar cells (Figure 1). Deep to layer II is layer III, which contains cell bodies of deep pyramidal cells, two populations of multipolar cells, and a large number of associational fibers, both from within piriform cortex and from other brain structures.

Deep to layer III is the endopiriform nucleus, which has also been termed layer IV of piriform cortex. This structure is dominated by multipolar neurons, and although it is interconnected with overlying piriform cortex, it does not receive direct input from the

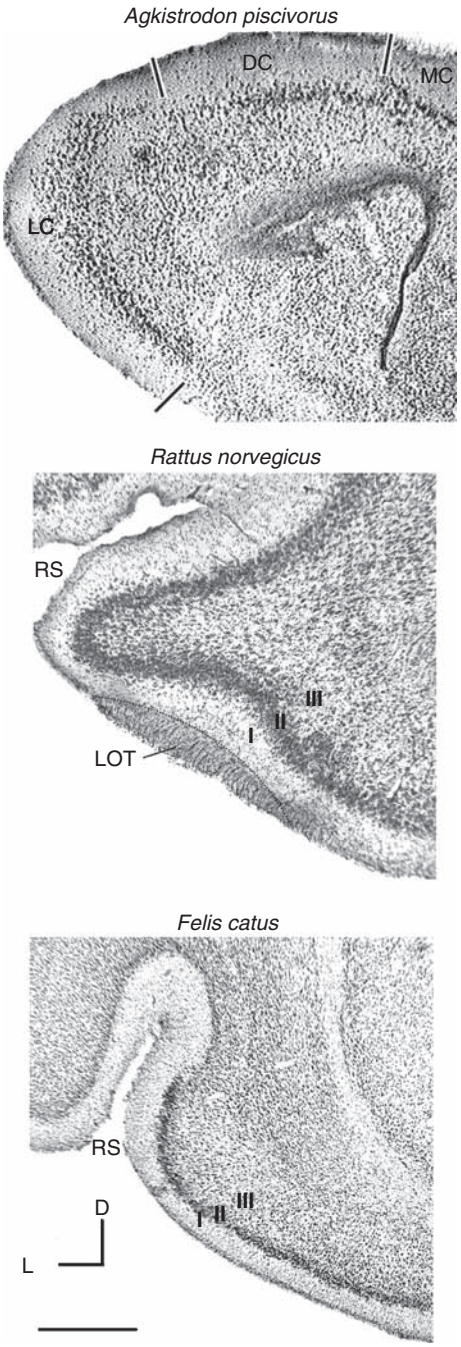

Figure 2 The appearance of olfactory cortex in mammalian and nonmammalian vertebrates. Nissl-stained coronal sections through one hemisphere of one nonmammalian species, the cottonmouth snake (*Agkistrodon piscivorus*) and of two mammalian species, the Norway rat (*Rattus norvegicus*), and the domestic cat (*Felis catus*). In *A. piscivorous*, the medial (MC), dorsal (DC), and lateral (LC) cortices are labeled. The anterior region of LC is the primary olfactory cortex, while more posterior regions of lateral cortex receive convergent olfactory and vomeronasal input. In the mammals, the piriform cortex is shown ventromedial to the rhinal sulcus (RS), with its characteristic three-layered appearance visible in these sections (I, II, and III). LOT, lateral olfactory tract. Scale bar = 500 μm.

olfactory bulb, and its role in olfactory information processing has not been investigated directly.

As in most other cortical structures, the principal neurons are glutamatergic pyramidal cells, and γ-aminobutyric acid (GABA)ergic interneurons are distributed throughout the cortex. Further, the piriform cortex is the target of a variety of neuromodulators, including acetylcholine from the horizontal limb of the diagonal band of Broca, norepinephrine from the locus coeruleus, and serotonin from the raphe nucleus. These neuromodulatory inputs heavily terminate in layers Ib, II, and III, targeting both dendritic and axonal elements.

Piriform cortex is often referred to as a single structure, but it can be divided into five or more subregions on the basis of chemo- and cytoarchitectural features. This review will focus on three main subdivisions: anterior piriform cortex (APC), ventral APC (APC_V), and posterior piriform cortex (PPC; see Figure 1). These areas appear similar to one another, but there is good evidence that they receive different inputs, have distinct intrinsic architecture, and provide unique patterns of output. For example, APC_V may be the only portion of piriform cortex that receives input from tufted cells, and it receives a smaller complement of associative fibers originating in piriform cortex than do other subregions. In the rat, this region of the cortex also lacks characteristic GABAergic 'cartridges,' inhibitory features found elsewhere in APC that dampen neuronal activity and control recurrent excitation levels, which may account for the observation that this subregion is among the most epileptogenic sites within the rodent brain. The output from APC_V is almost exclusively to the dorsal APC and to orbitofrontal cortex, which sends a projection back to APC_V.

The PPC receives a much lighter input from the olfactory bulb, and this input is exclusively from mitral cells; tufted cell axons do not reach into the region of the PPC. Layer Ia is distinctly smaller in PPC than in APC, and PPC has a much thicker layer Ib than APC does, reflecting a much heavier associative input from the more anterior subdivisions of piriform cortex.

There is good evidence in the literature that the divisions made according to anatomical features reflect functional differences among subregions of piriform cortex. For example, responses to odors recorded from awake, behaving animals seem to be more stimulus specific in APC whereas responses in PPC appear to incorporate more associative information, perhaps reflecting the relative differences in afferent and associative input seen in each area.

Output

Piriform cortex projects to a widespread and diverse set of forebrain structures. These structures include the olfactory bulb, AON, lateral amygdala nucleus, entorhinal cortex, orbitofrontal cortex, and mediodorsal nucleus of the thalamus. The projections to the olfactory bulb terminate on GABAergic granule cell interneurons and thus allow a cortical feedback control of olfactory bulb excitability. It is not known how these back projections are organized within the olfactory bulb relative to the patterns of olfactory receptor afferent input. The projection to the amygdala derives primarily from the PPC. Both the APC and the PPC receive input from the basolateral amygdala nuclei, though this projection is most pronounced to the PPC. Thus, the PPC has a unique reciprocal relationship with the amygdala, which may be important in the incorporation of hedonic or associative meaning with the representation of odors. The relationship between the piriform cortex and the orbitofrontal cortex is similarly complex, with both direct and indirect (via the dorsomedial thalamic nucleus) projections.

The Piriform Cortex in Other Mammalian Species

In general structure, the piriform cortex in mammalian species outside the class Rodentia is recognizable by its characteristic lamination (Figure 2). Experimental study has been made of the piriform cortex of the opossum, cat, dog, and primates. Its relative size within the brain diminishes greatly with the enlargement of the frontal lobes seen in these mammals. In rodents and other small nonprimates such as possums and cats, the piriform cortex is located along the ventral or ventrolateral edge of the forebrain, with the myelinated lateral olfactory tract (LOT) overlying the anterior regions. In primates, the piriform cortex has moved more medially as neocortex has expanded and resides along the dorsal medial edge of the temporal lobe. Both gross anatomy and, as far as is known, local circuitry are generally conserved across these species.

Rodent AON

The AON is a large structure that lies within the olfactory peduncle, caudal to the olfactory bulb and rostral to the piriform cortex, with which it merges at its caudal end. As noted above, the AON is a cortical structure that was named by early investigators who recognized the outside-in pattern of development of the AON as seen in other brain nuclei, as opposed to the inside-out pattern observed in cortex. Nevertheless, the concept of the AON as a cortical structure has been argued for at least 80 years. The argument is based on structural features – a rigid laminar structure, the presence of pyramidal-shaped principal cells, and its continuity with the piriform cortex caudally – and on its functional role in the olfactory system. Its continued designation as the AON persists as a matter of historical consistency, although it has been proposed that functional groupings within the AON, when identified, should be renamed in a manner consistent with its cortical nature.

Organization

The AON exhibits a laminar structure similar to that of piriform cortex; it receives a direct projection from the mitral and tufted cells within the superficial plexiform layer (layer I), and these synapse on the apical dendritic tree of glutamatergic pyramidal cells that make up the principal cell population of the AON. Pyramidal cell bodies lie in layer II, and deep to layer II is an amorphous, poorly differentiated layer of neuromodulatory and corticofugal inputs to the AON. GABAergic cells and terminals are found distributed throughout the AON, suggesting internal processing of olfactory information within the AON, but the proportion of GABAergic cells that are intrinsic to the AON relative to those which project elsewhere is currently unknown.

One question that has surrounded the internal organization of the AON is whether it can be subdivided into functionally or anatomically discrete units. Traditionally, the AON has been divided into five subregions, termed *pars* (Latin for 'part'), that have been designated according to their physical location: pars medialis, pars dorsalis, pars lateralis, pars ventralis (or ventroposterioralis), and pars externa. Unfortunately, with the exception of pars medialis and pars externa, these designations have been made without a cytoarchitectural basis, and there are no anatomically or functionally defined features to discriminate among the remaining subregions. Nevertheless, ongoing research suggests that the AON is not homogeneous and is likely a collection of functionally discrete subregions that perform separate roles in processing olfactory information.

Output

Three structures are the primary targets for pyramidal cells in the AON: the olfactory bulb, the contralateral AON, and the piriform cortex. The output to the olfactory bulb is widespread, with terminations found within each layer, suggesting that the AON plays an important role in the initial stages of odor processing. It is interesting that there are substantial regional differences in the patterns of projections to

the olfactory bulb; for example, projections from pars medialis terminate in the deep granule cell layer of the ipsilateral bulb, while projections from pars externa predominately innervate the inner plexiform layer in the contralateral bulb. The remaining subregions have heterogeneous, bilateral projections to broader regions of the bulb.

The projection to the contralateral AON is substantial and widespread, and the AON is the first structure in the olfactory pathways to exhibit such interhemispheric connections. Because the subregions of the AON that receive input from the contralateral hemisphere also project back to the ipsilateral olfactory bulb, the AON may be an important component in a network that allows interbulbar feedback regulation of activity. Further, because the AON is the first structure to receive input from both olfactory bulbs, it is in a position to compare the incoming signals, perhaps to facilitate localizing olfactory cues in the environment.

The output from the AON to the piriform cortex is heaviest to the central and ventromedial portions of APC, deep to the LOT and extending from the border with the olfactory tubercle laterally to just beyond the border of the LOT. This projection arises predominantly from pars lateralis, with lesser involvement from adjacent regions of pars dorsalis and pars ventroposterior. The projection from the AON to the piriform cortex is bilateral, but evidence from tract-tracing experiments suggests that pars dorsalis may project more heavily to the ipsilateral piriform cortex, while the pars ventroposterior projects more heavily to the contralateral hemisphere.

Perhaps the most striking feature of the projection from the AON to the piriform cortex is that these fibers terminate on pyramidal cells at the most proximal portion of the apical dendritic tree in deep layer Ib. This suggests that the AON is an important source of input for driving activity in piriform cortex, and it indicates that higher-order cortical processing of olfactory information involves complex reciprocal connections between these two cortical structures.

AON: Other Mammalian Species

The AON in mammalian species outside the class Rodentia is recognizable by its location and laminar appearance but has remained poorly studied. In most primates, the AON loses its circular structure as the forebrain expands, and it appears instead as extended groupings or islands of cells at the end of the olfactory stalk. In the macaque monkey, the AON may be only 12 pyramidal cells wide in some coronal sections. This diminished size, and the consequent difficulty in reaching the area with experimental techniques, may contribute the current lack of scientific study of the AON in other mammals.

Comparative Anatomy of Olfactory Cortex: Nonmammalian Vertebrates

Nonmammalian terrestrial vertebrates lack neocortex, but many have three-layered paleocortex that receives olfactory projections (see Figure 2). Output from the olfactory bulb in many of these species can be divided into at least two main projections, a medial olfactory tract and a LOT, often with the presence of a third, intermediate tract. Mammals have medial projections from the olfactory bulb, for example to the tenia tecta, but do not have a fully separate medial olfactory tract distinct from the LOT. In reptiles, the LOT projects bilaterally to the lateral cortex and to portions of the amygdala. The medial (and intermediate) olfactory tract projects to more medial structures, again bilaterally, including structures along the medial wall of the forebrain.

Similarly, the medial and lateral olfactory pathways in birds project to three (or more) distinct cortical areas; medially, olfactory bulb afferents innervate the regio retrobulbaris and the 'piriform' or 'prepiriform' area, while lateral projections innervate the regio periamygdalaris. Each of these areas receives heavy input from the olfactory bulb, and each exhibits a characteristic three-layered structure. While it has been postulated that the regio retrobulbaris in birds corresponds to the mammalian AON, and the piriform area corresponds to the mammalian piriform cortex, the degree to which any of these areas is similar to mammalian olfactory cortex is not known.

In amphibians and fish, with no true cortex, the olfactory bulb projects to the lateral pallium via the LOT and to the medial pallium via the medial olfactory tract.

Does the olfactory cortex in nonmammalian animals serve a functional role similar to that postulated for mammals? Some evidence suggests that it does. In birds, a portion of the lateral cortex (a region that is analogous to mammalian piriform cortex) is critical for the use of olfactory information. Lesioning this region in homing pigeons does not affect homing behavior when birds are released from familiar locations, but it drastically impairs performance on homing from unfamiliar locations, where olfactory cues are necessary. Further, recent physiological recordings in catfish suggest that the projection to the olfactory forebrain may be spatially organized such that food odors (e.g., amino acids) may be processed in regions distinct from social odors (e.g., bile salts), mirroring a similar functional segregation in the fish olfactory

bulb. These findings suggest that the olfactory forebrain in fish and the lateral (olfactory) cortex in birds may play a role in associating chemosensory cues with behavioral outcomes, a role that is analogous to that attributed to piriform cortex in the mammalian brain.

Summary

The last decade has seen an expansion of cellular and molecular biological techniques applied to sensory and systems neuroscience. Arguably, there is no area of study that has benefited more from this influence than has the study of olfaction, where this work has led directly to an understanding of the structure and function of olfactory receptor neurons and the organization of their projections to the olfactory bulb.

Nevertheless, critical questions regarding the processing of olfactory information remain unanswered, and the answers are likely to come from a study of the olfactory cortical regions. For instance, rodents have little difficulty making discriminations between complex odor mixtures that share components, even though these mixtures evoke activity in overlapping populations of olfactory receptors and mitral and tufted cells. How are such discriminations carried out? How is the pattern of cells activated in the olfactory bulb deciphered by the cortex? Do cortical areas play an active role in shaping the responses of cells in the bulb, as might be suggested by the heavy cortico–bulbar projections?

This review has outlined the structure of the two most prominent olfactory cortical areas in the vertebrate, the AON and the piriform cortex. By virtue of their connections and their position in the olfactory pathways, these structures are likely to play significant roles in olfactory information processing. Although some aspects of organization and circuitry are species specific, each of these cortical structures is characterized by heavy reciprocal relationships with the olfactory bulb, and with higher-order cortical areas, in all vertebrate species that have been studied. Consequently, a complete understanding of olfaction will be achieved only after the roles of these cortical structures have been identified.

See also: Flavor Physiology; Olfactory Coding; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Olfactory System Theory; Retronasal Olfaction.

Further Reading

Brunjes PC, Illig KR, and Meyer EA (2005) A field guide to the anterior olfactory nucleus (cortex). *Brain Research Reviews* 50: 305–335.

Carmichael ST, Clugnet MC, and Price JL (1994) Central olfactory connections in the macaque monkey. *Journal of Comparative Neurology* 346: 403–434.

Haberly LB (2001) Parallel-distributed processing in olfactory cortex: New insights from physiological and morphological analysis of neuronal circuitry. *Chemical Senses* 26: 551–576.

Lanuza E and Halpern M (1998) Efferents and centrifugal afferents of the main and accessory olfactory bulbs in the snake Thamnophis sirtalis. *Brain, Behavior and Evolution* 51: 1–22.

Neville KR and Haberly LB (2004) Olfactory cortex. In: Shepherd GM (ed.) *The Synaptic Organization of the Brain*, pp. 415–454. New York: Oxford University Press.

Shipley MT and Ennis M (1996) Functional organization of olfactory system. *Journal of Neurobiology* 30: 123–176.

Olfactory Ensheathing Cells: Regeneration Following Spinal Cord Injury

M Nieto-Sampedro, Consejo Superior de Investigaciones Científicas, Madrid, Spain; and National Hospital of Paraplegics, Toledo, Spain

© 2009 Elsevier Ltd. All rights reserved.

Central Nervous System Regeneration Failure

The difficulty of central nervous system (CNS) lesion repair has been known since the pioneering work of Tello and Ramón y Cajal at the beginning of twentieth century. The failure of injured adult mammalian CNS to regenerate is due to (1) the formation in the injury area of the glial scar, a cellular–molecular structure inhibitory of axonal regrowth; (2) neuronal death secondary to the primary injury and delayed with respect to it; and (3) the postnatal loss of intrinsic CNS axonal growth capacity (Figure 1). Oligodendrocytes, reactive astrocytes, and reactive microglia are responsible for the unfavorable growth environment found in the CNS. Because Schwann cells in the peripheral nervous system favor axon regeneration by producing growth-promoting extracellular and soluble molecules, dorsal rhizotomy has been used extensively to investigate CNS regeneration. The gliotic tissue formed at the dorsal horn/dorsal root entry zone (DREZ) or in other CNS loci after injury has inhibitory properties that prevent penetration in the CNS tissue of the abundant sprouts formed after damage of the sensory axon central branch.

Transplantation near the DREZ of olfactory bulb ensheathing cells (OBECs) modifies the spinal environment, causing astrocyte reactivity to diminish/disappear and promoting regeneration and spinal ingrowth of damaged sensory fibers. However, due to innervation overlap by adjacent sensory roots, the functional deficit caused by damage to a single dorsal root is not clear enough to permit assessment of the recovery achieved with the help of OBEC transplants. Accordingly, multiple rhizotomy was performed at the cervical, lumbar, and lumbosacral levels, and OBEC transplants were used to repair the damage. Transplant site and technique influenced the extent of sensory function recovery and its anatomical and electrophysiological correlates. More recently, Nieto-Sampedro's proposal of using OBEC transplantation as a general strategy for CNS injury repair has been tested by several groups and partial recovery after complete spinal cord transection has been achieved in rodents.

Multiple Rhizotomy: Repair and Partial Recovery of Function

Variations of the OBEC transplantation technique (Figure 2) have been used to repair multiple dorsal rhizotomy at the cervical, lumbar, and lumbosacral levels, with partial recovery of the functions mediated by the repaired afferents. These studies showed that OBEC transplantation promoted axonal regrowth and reinnervation of the dorsal horn, activation of postsynaptic dorsal horn neurons, and restitution of polysynaptic reflex activity. Reinnervation led to recovery of anatomical and electrophysiological parameters, correlating with cutaneous, proprioceptive, and autonomic sensory function, as well as restoring withdrawal reflexes and organized suprasspinal nocifensive responses to forepaw stimulation.

Unilateral C4-to-T2 dorsal rhizotomy in the rat led to severe forepaw sensory deficits. Transplantation of OBECs medial to the DREZ of cut dorsal roots C7/C8 promoted significant recovery of sensory response to noxious stimuli as judged by C-*fos* protein expression. The increased expression of c-*fos* protein in intrinsic dorsal horn neurons of the superficial dorsal horn laminae, induced by immersion of the ipsilateral forepaw in water at 52 °C, was observed in nonoperated animals or in sham-operated controls, abolished in rhizotomized animals, and partially restored after ipsilateral OBEC transplantation in the dorsal horn. Therefore, OBEC transplants helped re-establish functional contacts between injured sensory axons and intrinsic dorsal horn neurons.

Direct evidence of functional restitution of cutaneous reflex circuits was obtained by recording biceps reflex activity in response to stimulation of the median nerve, ipsilateral to C4–T2 rhizotomy and C7/C8 OBEC transplantation. Stimulus–response curves were obtained for nonoperated animals, rhizotomized rats transplanted with cell culture medium, and rats transplanted with ensheathing cells. Median nerve activation with short (0.1 ms) or long (0.5 ms) pulse duration stimuli indicated that only small-diameter afferents regenerated in transplanted animals. Biceps reflexes recovered after OBEC transplantation showed a higher 'gain' than either normal or

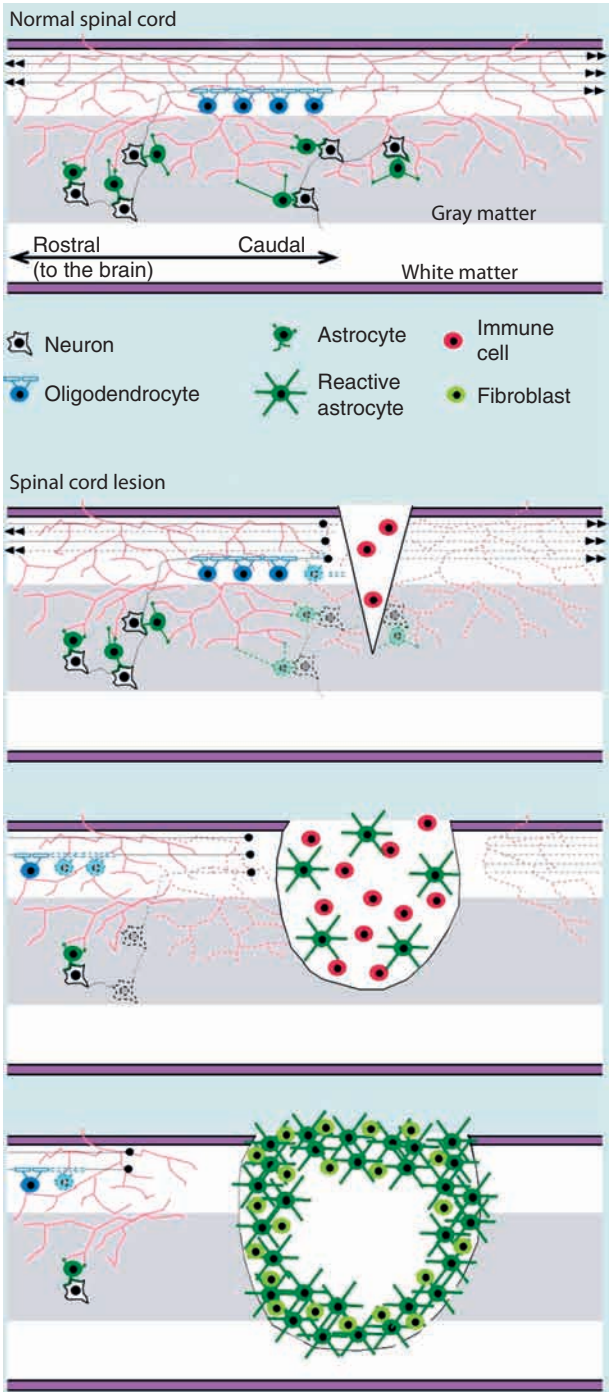

Figure 1 The spinal cord before and after a traumatic lesion.

nontransplanted rhizotomized rats. The synchronized electromyographic discharge evoked by median nerve stimulation in the biceps in normal adult rats was lost after rhizotomy or rhizotomy plus transplant of DMEM alone. However, similar stimulation of animals rhizotomized and transplanted OBECs 6 months after injury evoked an unsynchronized biceps discharge, reminiscent of immature activity.

The repair of sensory input observed electrophysiologically occurred in parallel with restoration of sensory reflexes in the awake animal. Withdrawal reflex and supraspinal nocifensive responses to both radiant heat stimuli and mechanical stimuli applied to the ipsilateral forepaw plantar pad were assessed behaviorally before and up to 2 months after cervical dorsal rhizotomy. Mechanical stimulation of the forepaw after brachial rhizotomy did not evoke reflex activity in the nontransplanted group or in the OBEC-transplanted group (data not shown). In contrast, reflex responses to thermal stimuli showed significant recovery after OBEC transplantation. During the first month following rhizotomy, ipsilateral withdrawal reflexes were abolished in control groups, and they were partially restored in the OBEC-transplanted group during the second month after injury (10–15 s; Figure 3(a)), significantly better than in nontransplanted controls (20–22 s; Figure 3(b)). The same behavioral test was also used to measure the recovery of organized and unlearned supraspinal responses in transplanted animals. A defined nonparametric scale was developed to grade orientation of the head and paw-licking behavior in response to thermal stimulation of the forepaw. This nocifensive score (SNS) declined from a maximal value of 3 in normal animals to 0 after rhizotomy. The score remained at 0 in nontransplanted animals (indicating lack of awareness of thermal stimulation; Figure 3(c)). In contrast, the group of OBEC-transplanted rats recovered an SNS score of 1 during the second month following rhizotomy, indicating orientation toward the stimulated forepaw (Figure 3(d)). The recovery of supraspinal nocifensive responses after OBEC implantation correlated significantly with the recovery of withdrawal flexor activity, indicating that both reflexive and supraspinal behavioral responses measured activity evoked by regenerating afferents.

Repair of Transected Rat Spinal Cord

Spinal cord transection leads to permanent loss of central control of motor, sensory, and autonomic functions below the injured segment. The seemingly irreversible deficits are caused by neuronal death, together with the failure of axons to regenerate across the injury site. Upregulation of proteoglycans and myelin-associated inhibitors, lack of adequate supply of neurotrophic factors, and formation of a glial scar are the major obstacles of axonal regeneration. However, given the appropriate cellular environment, central axons can sprout and regrow long distances. We proposed transplantation of olfactory bulb ensheathing cells as a general environmental modification strategy for CNS injury repair. Such intervention

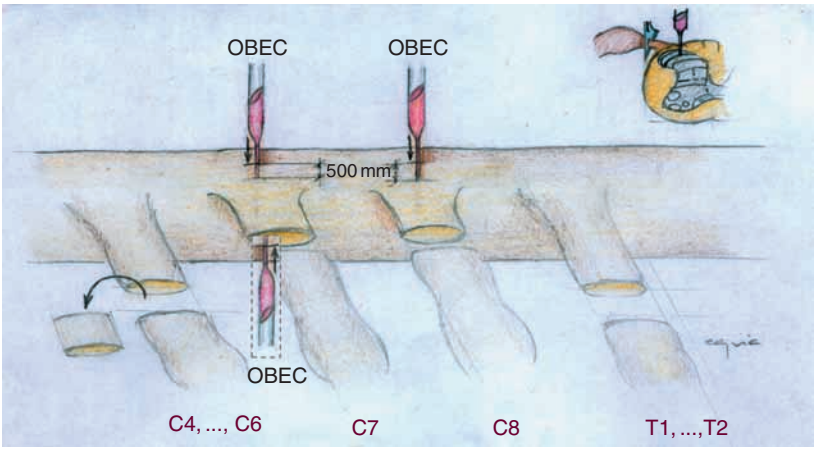

Figure 2 To optimize functional recovery mediated by afferents regrowing into the spinal cord after dorsal rhizotomy, various surgical techniques were used depending on the cord segment. In cervicothoracic segments, the OBEC suspension was injected in the dorsal horn, near the DREZ, and tens of micrometers below the Lissauer band. At the lumbar level, the OBEC suspension was injected in the central stump of the cut root. In the case of lumbosacral deafferentation, the transplant was carefully injected in the cord, near the sacral parasympathetic nucleus. In all cases, OBEC migration, both toward the central canal and toward the DREZ, was observed. Adapted from Taylor JS, Muñetón-Gómez VC, Equía-Recuero R, and Nieto-Sampedro M (2001) Transplants of olfactory bulb ensheathing cells promote functional repair of multiple dorsol rhizotomy. *Progress in Brain Research* 132: 651–664, with permission from Elsevier.

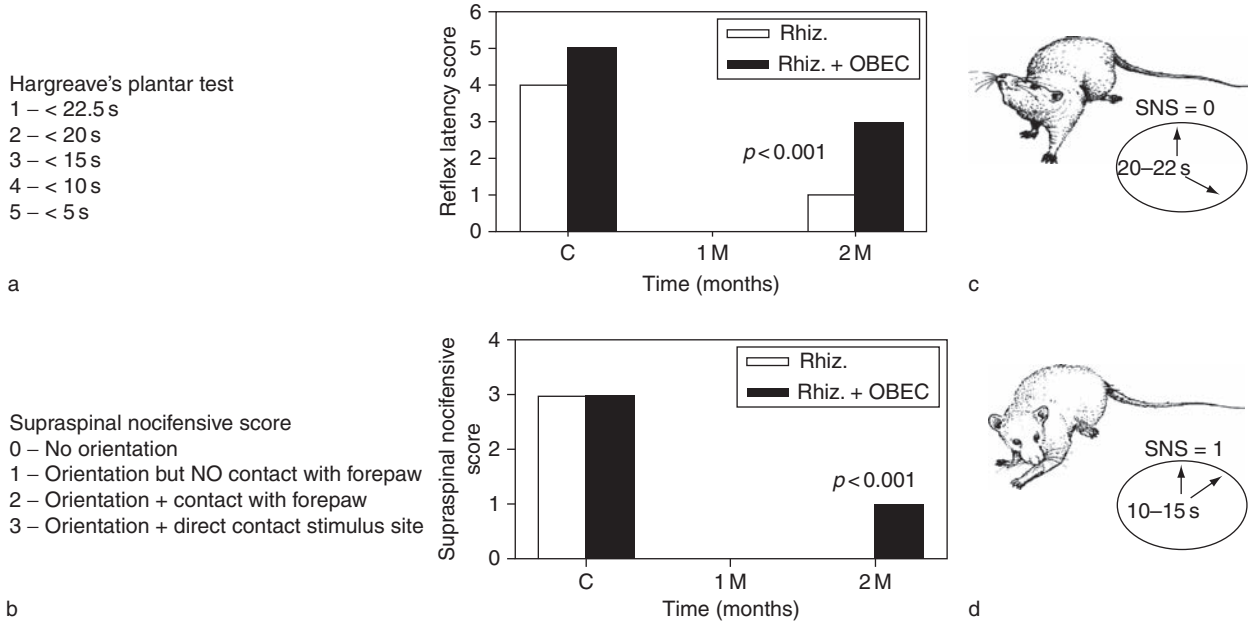

Figure 3 Supraspinal nocifensive response. (a) Withdrawal reflex latency was measured before and up to 2 months after complete ipsilateral brachial plexus (C4–T2) dorsal rhizotomy in response to radiant heat applied to the forepaw of rats. Rhizotomized animals received dorsal horn injections of DMEM only (open bars) or a suspension of OBECs in DMEM (solid bars). (b) Supraspinal nocifensive response to radiant heat as in a, before and up to 2 month after rhizotomy (format as in (a)). (c, d) Schematic of the behavioral results observed during the second month after rhizotomy indicating the reflex latency and supraspinal nocifensive score in nontransplanted (c) and OBEC-transplanted rats (d). Adapted from Taylor JS, Muñetón-Gómez VC, Equía-Recuero R, and Nieto-Sampedro M (2001) Transplants of olfactory bulb ensheathing cells promote functional repair of multiple dorsol rhizotomy. *Progress in Brain Research* 132: 651–664, with permission from Elsevier.

appears limited by the migration preferences of OBEC in the adult CNS and by the survival of transplanted cells in the free radical-rich injured tissue. Previous work in our laboratory showed that OBEC transplanted into mechanically or photochemically hemisected spinal cord had neuroprotective and pro-regenerative effects, improving functional outcome. Although the mechanisms underlying the neuroprotective/regenerative properties of OBEC are not completely understood, the publication of the

rhizotomy repair results encouraged the use of OBEC transplants to repair corticospinal tract lesions, demyelination, and complete spinal transections. The conditions for corticospinal tract repair of Li et al. were optimal in that experimental lesions (selective electrolytic lesions) were very small (approximately 500 μm) and neither bleeding nor fibroblast entry in the cord was allowed. Modulation of astrocyte reactivity and proteoglycan expression, promotion of neoangiogenesis, and release of neurotrophic factors are considered the main beneficial actions of the transplanted cells. In addition, OBECs promote regenerative and collateral axon sprouting and remyelination. Experimental studies by Navarro's group in Barcelona, Spain, analyzed the effects of OBEC grafts on axonal regeneration and target reinnervation after complete thoracic spinal cord transection in the adult rats, acutely or 1 week postlesion. Electrophysiological, immunohistochemical, and behavioral techniques were used to investigate the modulation by the transplants of the expression of glial fibrillary acidic protein (GFAP) and NG2. The CNS repair potential of OBEC transplants was manifested by their use for repair of optic or facial nerves.

However, the neuroprotective and pro-regenerative effects of OBEC grafts after spinal cord injury or dorsal rhizotomy have been questioned by other groups. Differences in the results may be due to different cell sources (olfactory bulb, olfactory nerve layer, or olfactory mucosa) and differences in cell culture, purification, and preservation protocols applied to different trauma models. Despite controversy, several clinical trials are ongoing that use OBEC grafts to treat human spinal cord injury, without concluding results. The National Institute of Neurological and Communication Disorders and Stroke and International Campaign for Cures of Spinal Cord Injury Paralysis panel guidelines to assess functional repair after spinal cord injury require histological and electrophysiological evidence of nerve disconnection after injury, of axonal regeneration across the lesion site, of postsynaptic responses to stimulation of regenerated fibers, and, finally, of functional recovery mediated by the new connections. The dangers of extrapolating the results from cell culture and rodent experimental models to the clinic, as well as the importance and potential pitfalls of the various models, have been carefully discussed by Dobkin.

Available and Foreseeable Central Nervous System Injury Repair Tools

Although it appears likely that OBEC transplantation will be the way to repair spinal cord injuries in the immediate future, several important problems remain to be solved before reliable clinical trials can be initiated. One problem concerns the amount of cells required to repair large injuries and the possibility of growing sufficient cells in primary culture or using immortalized cell lines. Another problem concerns the immunological compatibility of transplant donor and receptor, and a third problem regards the migration preferences of OBECs in the CNS. Various laboratories throughout the world are addressing these problems.

To search for OBEC-like glia with other migration preferences, we screened the adult rat brain for glial cells that, like OBECs, would be GFAP-, vimentin-, and p75-positive and could grow in culture from adult tissue dissociates. OBEC-like macroglial cells fulfilling these criteria were found in several brain loci, including the hypothalamus and the hypophysis. The macroglial types in these loci, tanycytes and pituicytes, have functions that involve rapid and reversible axon ensheathment. Furthermore, tanycytes and pituicytes have regeneration-promoting properties similar to both OBECs and Schwann cells. Tanycytes, pituicytes, olfactory bulb ensheathing cells, and pineal glia, in common with nonmyelinating Schwann cells, survive well when cultured from adult rodent brain tissue; exhibit morphologies similar to OBECs; and express a common distinctive set of immunological markers, namely p75 neurotrophin receptor, O4 antigen, estrogen receptor-α type, and insulin-like growth factor-1 (IGF-1) (Table 1).

Tanycytes extended long, fine, p75 immunoreactive processes, parallel to each other, from the ventricular wall to the ventral brain surface. All p75-positive tanycytes in both median eminence and infundibulum co-expressed GFAP, vimentin, estrogen receptor-α, and antigen O4. Immunoreactivity for estrogen receptor of the α type had previously been found exclusively in neuronal nuclei and in neuronal cytoplasmic processes. However, the well-characterized polyclonal antibody used by us (National Hormone and Pituitary Program, Rockville, MD), prepared against a synthetic peptide from residues 270–284 of domain D of estrogen receptor-α, immunostained in the adult rat brain the nuclei of hypothalamic neurons as well as the plasma membranes of olfactory bulb ensheathing cells and tanycytes. Preincubation with the peptide used as immunogen abolished both neuronal and glial immunostaining. Confocal microscopy of arcuate nucleus, median eminence, and pituitary stalk sections, doubly labeled with antibodies to the estrogen receptor D domain and to p75, showed co-localization of both receptors in the nonneuronal cells characteristic of these loci. These cells, when co-cultured with neurons, associated with

Table 1 Comparative immunological markers of conventional central macroglia, aldynoglia, and Schwann cells[a]

Cell type	*O4*	*p75*	*S100*[b]	*ER*	*Vim*	*GFAP*
Ast†	−	−	+	−	−	++
Olig†	++	−	±	−	−	−
SC*	+	++	++	++	++	++
Tan†	+	+	++	++	++	++
Tan*	+	++	++	+	++	+
Pit†	+	+	++	++	++	++
Pit*	+	++	++	+	++	+
OBECs†	+	+	++	+	++	++
OBECs*	+	++	++	+	++	+

[a]Cell immunoreactivity was very intense (++), clearly positive (+), or negative (−). Reaction with the immunological markers indicated was examined in adult tissue *in situ* (†) or in cultures prepared from adult tissue (*). Astrocytes (Ast), oligodendrocytes (Olig), and cultured nonmyelinating Schwann cells (SC) were used as standards of immunostaining intensity under the same experimental conditions as the rest of the cell types.
[b]Immunostaining of tissue sections with polyclonal anti-S100 was not selective – that is, weak staining of every type of glial cell was superimposed over intense OBEC or tanycyte staining.
ER, estrogen receptor-α; GFAP, glial fibrillary acidic protein (polyclonal IgG); OBECs, olfactory ensheathing cells; O4, seminolipid, sulfatide antigen; p75, low-affinity NGF receptor; Pit, pituicytes; S100, Ca^{2+} binding protein S100; Tan, tanycytes; Vim, vimentin.
Data from Gudiño-Cabrera G and Nieto-Sampedro M (2000) Schwann-like macroglia in adult rat brain. *Glia* 30: 49–63.

the neurites in a manner suggestive of enveloping and expressed peripheral myelinating phenotype.

The Schwann-like properties shown by these cells define a CNS macroglial type that would not fit in any of the conventional central glia classes (Table 1). We have called this distinct group of growth-promoting CNS macroglia 'aldynoglia' (Greek αλδαινο meaning 'make grow'). The proliferative and growth-promoting properties of aldynoglia seem to be retained during the entire lifetime of an organism in those CNS loci in which normal function depends on continuous axon renewal. The concomitant expression of estrogen receptor and low-affinity neurotrophin receptor may confer Schwann-like plasticity to these glial cells. Other glial cells, such as Müller cells, Bergmann glia, and cortical and spinal cord radial glia, show growth properties and immunological markers reminiscent of aldynoglia. Aldynoglia appears as an intermediate in the ontogeny of both neurons and macroglia. Its plasticity seems totally or partially lost with age, where and when it is no longer essential, as in the case of adult astroglia.

The problem of growing a sufficient number of cells in culture, as required for injury repair, has been approached in our laboratory as well as other laboratories, attempting to differentiate aldynoglia from neural stem cells and to transdifferentiate hematopoietic stem cells into neural stem cells and then into aldynoglia. These two projects differ in both their potential importance and the present state of the consensus. Neurospheres may be generated in abundance from embryonic rat striatum and differentiated to aldynoglia-like cells when plated onto an adhesive substrate in olfactory bulb ensheathing cell conditioned medium, supplemented with epidermal growth factor and basic fibroblast growth factor. The immunophenotype, messenger RNA expression (microarray analysis), and functional properties of the cells differentiated from neurospheres were similar to those of aldynoglia, although remarkably, they did not express the MHC II complex, an important result that needs further confirmation. Aldynoglia-like cells expressed, with respect to neurosphere cells, unchanged *nestin* levels and high GFAP, suggesting a glial phenotype with a possible role as precursor of neuronal and glial cells. Selective differentiation of these precursor cells may control the proportion of the various cell types in neural tissue during development and following injury.

Transdifferentiation, on the other hand, is proposed to explain how tissue-specific stem cells could generate cells typical of other organs, supporting the idea of enhanced adult stem cell plasticity. Although spontaneous cell fusion, rather than transdifferentiation, may explain some cell fate changes *in vivo*, such a mechanism does not explain possible transdifferentiation events *in vitro*, including the generation of neural cell types from cultured bone marrow-derived stem cells. Croft and Przyborski showed that bone marrow-derived stem cells expressed neural proteins in culture and formed neural-like structures under defined growth conditions. These changes in cell structure and protein expression were not typical of neural development, and the ability of bone marrow-derived stem cells to adopt a neural phenotype *in vitro* possibly occurred as a result of cellular stress and in response to removing cells from their niche and making them grow in different environmental conditions. Whatever the explanation for the growth behavior of bone marrow-derived cells, the most important question from a clinical standpoint is whether these cells will or will not perform as neural stem cells do when autotransplanted in CNS injury sites.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Injury: Neuronal Responses; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Glial Responses to Injury; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Microglial Response to Injury; Nogo-A: Its Role in Axon Regeneration; Spinal Cord Regeneration after

Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Regeneration and Functional Recovery: Strategies.

Further Reading

Dobkin BH (2007) Curiosity and cure: Translational research strategies for neural repair-mediated rehabilitation. *Developmental Neurobiology* 67: 1133–1147.

Fawcett JW, Curt A, Steeves JD, et al. (2007) Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: Spontaneous recovery after spinal cord injury and statistical power needed for therapeutic clinical trials. *Spinal Cord* 45: 190–205.

Fraher JP (2000) The transitional zone and CNS regeneration. *Journal of Anatomy* 196: 137–158.

Gudiño-Cabrera G and Nieto-Sampedro M (1996) Ensheathing cells: Large scale purification from adult olfactory bulb, freeze-preservation and migration of transplanted cells in adult brain. *Restorative Neurology and Neuroscience* 10: 25–34.

Gudiño-Cabrera G and Nieto-Sampedro M (2000) Schwann-like macroglia in adult rat brain. *Glia* 30: 49–63.

Gudiño-Cabrera G, Pastor AM, de la Cruz RR, Delgado-García JM, and Nieto-Sampedro M (2000) Limits to the capacity of olfactory ensheathing glia to promote axonal regrowth in the CNS. *NeuroReport* 11: 467–471.

Imaizumi T, Lankford KL, Waxman SG, Greer CA, and Kocsis JD (1998) Transplanted olfactory ensheathing cells remyelinate and enhance axonal conduction in the demyelinated dorsal columns of the rat spinal cord. *Journal of Neuroscience* 18: 6176–6191.

Keyvan-Fouladi N, Raisman G, and Li Y (2003) Functional repair of the corticospinal tract by delayed transplantation of olfactory ensheathing cells in adult rats. *Journal of Neuroscience* 23: 9428–9434.

Li Y, Field PM, and Raisman G (1998) Regeneration of adult rat corticospinal axons induced by transplanted olfactory ensheathing cells. *Journal of Neuroscience* 18: 10514–10524.

Li Y, Sauve Y, Li D, Lund RD, and Raisman G (2003) Transplanted olfactory ensheathing cells promote regeneration of cut adult rat optic nerve axons. *Journal of Neuroscience* 23: 7783–7788.

Lopez-Vales R, Forès J, Verdú E, and Navarro X (2007) Chronic transplantation of olfactory ensheathing cells promotes partial recovery after complete spinal cord transection in the rat. *Glia* 55: 303–311.

Mingorance A, Sole M, Munetón V, et al. (2006) Regeneration of lesioned entorhino-hippocampal axons *in vitro* by combined degradation of inhibitory proteoglycans and blockade of Nogo-66/NgR signaling. *FASEB Journal* 20: 491–503.

Navarro X, Valero A, Gudiño-Cabrera G, et al. (1999) Ensheathing glia transplants promote dorsal root regeneration and spinal reflex restitution after multiple lumbar rhizotomy. *Annals of Neurology* 45: 207–215.

Nieto-Sampedro M (1998) Neuron–glia ensembles andmammalian CNS lesion repair. In: Castellano B, González de Mingo B, and Nieto-Sampedro M (eds.) *Understanding Glial Cells*, pp. 255–270. Norwell, MA: Kluwer.

Pascual JI, Gudiño-Cabrera G, Insausti R, and Nieto-Sampedro M (2002) Spinal implants of olfactory ensheathing cells promote axon regeneration and bladder activity after bilateral lumbosacral dorsal rhizotomy in the adult rat. *Journal of Urology* 167: 1522–1526.

Polentes J, Stamegna JC, Nieto-Sampedro M, and Gauthier P (2004) Phrenic rehabilitation and diaphragm recovery after cervical injury and transplantation of olfactory ensheathing cells. *Neurobiology of Disease* 16: 638–653.

Raisman G and Li Y (2007) Repair of neural pathways by olfactory ensheathing cells. *Nature Reviews Neuroscience* 8(4): 312–329.

Ramón-Cueto A and Nieto-Sampedro M (1994) Regeneration into the spinal cord of transected dorsal root axons is promoted by ensheathing glia transplants. *Experimental Neurology* 127: 1–13.

Taylor JS, Muñetón-Gómez VC, Equía-Recuero R, and Nieto-Sampedro M (2001) Transplants of olfactory bulb ensheathing cells promote functional repair of multiple dorsal rhizotomy. *Progress in Brain Research* 132: 651–664.

Olfactory Epithelium

T Leinders-Zufall, University of Saarland School of Medicine, Homburg, Germany
M Ma, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Functional Organization of the Mammalian Nose

The olfactory system recognizes a vast range of molecules that represent vital information about an animal's environment, including the location of prey and predators; the sexual, hormonal, and reproductive state of mating partners; and the level of aggression in rivals. How does the nose organize this complex mixture of olfactory information? Evidence from rodents has shown that there are several spatially segregated subpopulations of sensory neurons within the mammalian nose (Figure 1), each projecting to different areas of the brain and each likely serving as a somewhat distinct channel for different types of chemosensory information. Some of these subpopulations of sensory neurons use different signaling cascades for the transduction of information from chemosensory molecules into electrical membrane signals. The use of genetically altered animals in which each of these chemosensory information routes (Figure 1) is rendered nonfunctional in a selective manner will eventually lead to a better understanding of the precise roles that each subsystem plays in odor-dependent behaviors.

This article focuses on the mechanisms of odor transduction in the sensory neurons of the mammalian olfactory system, excluding two types of neurons from the vomeronasal organ (Figure 1) which are discussed separately. The mouse has rapidly emerged as one of the best model systems for investigating the cellular and molecular basis of odor transduction as well as the relationship between mechanisms underlying transduction and the significance of these signaling steps for odor-dependent behaviors. Knowledge gleaned from this work should also provide a solid foundation for understanding the molecular basis of odor transduction in other vertebrates, including humans, despite the fact that there are specific adaptations in each species regarding the overall organizational and structural features underlying olfactory sensing. Olfactory transduction in invertebrates is discussed elsewhere in this encyclopedia.

Olfactory Sensory Neurons of the Main Olfactory Epithelium

Signal Transduction: Olfactory Receptors, cAMP-Gated Channels, and Ca^{2+}

Odor transduction is best understood in ciliated olfactory sensory neurons (OSNs) of the main olfactory epithelium (MOE), which function as general sensors of the external chemical environment. Approximately 6–10 million OSNs form a sensory neuroepithelium that lines the nasal cavity in mammals (Figures 2(a) and 2(b)). The OSNs are bipolar neurons with a single dendrite that projects to the apical surface of the epithelium ending in a knoblike swelling from which emanate 10–30 very fine cilia (200–300 nm in diameter; Figure 2(c)). The cilia, which are covered by a thin layer of mucus, house the olfactory signal transduction machinery and serve as the primary interface between chemostimuli in the external environment and nervous system signaling. On the basal pole of each OSN, a thin unmyelinated axon projects to the main olfactory bulb (MOB). OSN axons form synapses with second-order neurons (mitral cells) and interneurons in dense regions of neuropil in the MOB that are collectively known as olfactory glomeruli.

The main olfactory system, like the immune system, is built on the condition that it is not possible to predict what molecules it might encounter. Therefore, it is necessary to maintain an indeterminate yet nonetheless precise sensory array. The olfactory system manages this complicated detection and discrimination task by expressing a surprisingly large family of G-protein-coupled receptors (GPCRs). This family of olfactory receptors, first identified by Buck and Axel, consists of as many as 1400 genes in mammals, making it the largest family of GPCRs and the largest gene family in the entire genome. Efforts in sequencing the entire human and mouse genomes have provided large databases of candidate olfactory receptors that can be accessed via the World Wide Web (Genbank and the Olfactory Receptor Database (ORDB)). However, specific ligands have only been identified for some receptors. Thus, many members of the olfactory receptor family are still considered orphan receptors.

How does the organization of the MOE allow for encoding an almost unlimited number of structurally diverse odor molecules? Although this critical question is still under investigation, several general principles have emerged. The first principle is that each OSN expresses only a single olfactory receptor (with

some exceptions) and that axons from OSNs that express the same receptor converge onto small sets of olfactory glomeruli at reproducible locations in each olfactory bulb. This allows for the creation of precise maps of receptor activation in the MOB. A second principle is that these receptors are relatively broadly tuned to odors. In other words, each OSN is sensitive to a different but overlapping set of odor ligands, and these tuning curves become less specific as odor concentration is increased. Thus, OSNs with the lowest threshold for a given odor molecule are activated first, and the less sensitive ones are recruited at higher concentrations, giving rise to a combinatorial coding scheme (Figure 3). The combinatorial coding scheme dictates that each odor ligand activates a unique subset of OSNs (and hence olfactory receptors), and this activation pattern depends on both the molecular identity of the stimulus and its concentration, allowing for a vast number of combinations.

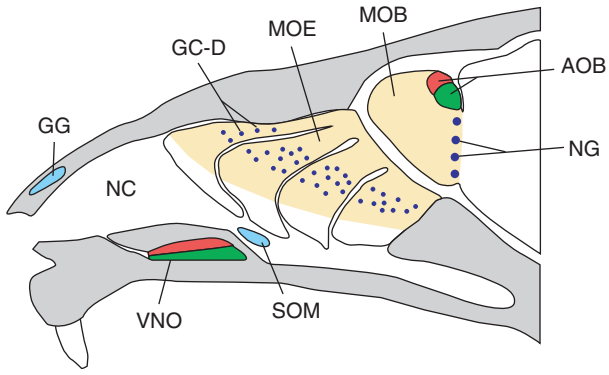

Figure 1 Anatomical organization of chemosensory systems in the nose. Midsagittal view of the rodent nasal cavity (NC) and forebrain. Olfactory sensory neurons of the main olfactory epithelium (MOE) project to glomeruli in the main olfactory bulb (MOB; yellow). GC-D (guanylyl cyclase D-expressing) neurons (blue) are also located in the MOE but project to a limited number of necklace glomeruli (NG) in the caudal MOB. The septal organ of Masera (SOM) is an isolated island of olfactory sensory neurons that project axons to glomeruli in the ventromedial olfactory bulb. The Grueneberg ganglion (GG) projects axons to specific glomeruli in the caudal region of the MOB. Sensory neurons in the vomeronasal organ (VNO) project to the anterior (red) or posterior (green) accessory olfactory bulb (AOB).

Once the olfactory receptors have bound an odor ligand, a cascade of events is initiated, resulting in the generation of an electrical membrane current – a process called primary odor transduction. The odor-induced sensory current underlies the graded membrane depolarization known as receptor potential (Figures 4(a) and 4(b)). The basic transduction apparatus in the OSN cilia is composed of a G-protein-coupled cyclic adenosine 3′,5′-monophosphate (cAMP)-mediated second messenger cascade located in the ciliary plasma membrane (Figure 4(c)). The ligand-bound receptors activate a stimulatory GTP-binding protein ($G\alpha_{olf}$), which in turn activates type III adenylyl cyclase (ACIII), leading to the rapid production of the second messenger cAMP. Mutant mice with targeted deletions in either $G\alpha_{olf}$ or ACIII are unable to smell most odor molecules (anosmia). The main target for odor-induced cAMP is a cyclic nucleotide-gated (CNG) channel, which is closed in the absence of an odor stimulus but directly opened (gated) through the binding of cAMP to a specific site on the intracellular face of the channel. This CNG channel is a heteromultimeric protein composed of three distinct

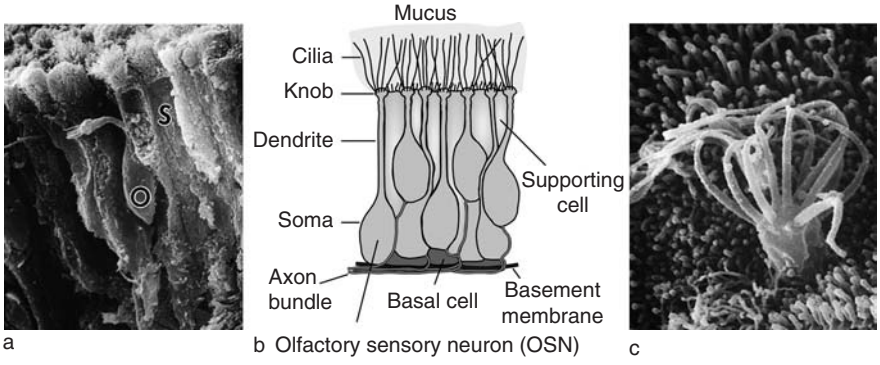

Figure 2 Cytoarchitecture of the main olfactory epithelium. (a) Three-dimensional scanning view of a human olfactory epithelium. An olfactory sensory neuron (O) with its dendrite, knob, and cilia can be seen among supporting cells (S). Magnification, ×1241. (b) The main olfactory neuroepithelium (MOE) is a relatively simple tissue consisting of three major cell types: (1) the olfactory sensory neurons (OSNs; the only neuronal cell type), (2) supporting or sustentacular cells (S; a kind of glial cell whose function is not well understood), and (3) a stem cell population, known as basal cells, from which new OSNs are generated. (c) Three-dimensional view of a dendritic knob of a human olfactory neuron giving rise to cilia that gently taper as they extend over the epithelial surface. The cilia house the olfactory signal transduction machinery and serve as the primary interface between chemostimuli in the external environment and nervous system signaling. Magnification, ×18 525. (a, c) Reprinted from Morrison EE and Costanzo RM (1990) Morphology of the human olfactory epithelium. *Journal of Comparative Neurology* 297: 1–13.

Figure 3 Combinatorial receptor code for odorant detection. In this model, the receptors shown in dark red are those that recognize the odor ligand shown on the left. The identities of different odor ligands are encoded by different combinations of odor receptors (ORs). However, each OR can serve as one component of the combinatorial receptor code for many odor ligands. Given the immense number of possible combinations of ORs, this scheme could allow for discrimination of an almost unlimited number and variety of different odor ligands. Reprinted from Malnic B, Hirono J, Sato T, and Buck LB (1999) Combinatorial receptor codes for odors. *Cell* 96: 713–723, with permission from Elsevier.

subunits: CNGA2, CNGA4, and CNGB1b. CNGA2 is the principal subunit; mutant mice lacking this subunit are largely anosmic. CNGA4 and CNGB1b are modulatory subunits that play specific roles in odor adaptation. The CNG channel is permeable to monovalent cations such as Na^+ and K^+, but it also has a substantial permeability to the divalent cation Ca^{2+}. At rest, OSNs maintain a membrane potential near −65 mV. Odor stimulation leads to an influx of Na^+ and Ca^{2+} through the CNG channels and thus causes the inside of the cell to become approximately 30 mV less negative, generating a membrane depolarization. This, in turn, elicits action potentials which then propagate along the OSN axon into the forebrain, where they initiate nerve impulses in second-order neurons (mitral cells) in the olfactory bulb.

The odor-induced rise in intracellular Ca^{2+} concentration (Figures 4(c)–4(h)) is of particular significance for the transduction process because it acts as a negative feedback in olfactory signaling and allows rapid adaptation in this sensory system. As intracellular Ca^{2+} increases during the odor response, it acts on the CNG channel to decrease its sensitivity to cAMP (Figure 4(c)), thus diminishing the probability of opening the channel. This effect of Ca^{2+} requires the ubiquitous Ca^{2+}-binding protein calmodulin. Also essential for this feedback to function correctly are the modulatory CNG channel subunits CNGA4 and CNGB1b, which accelerate the Ca^{2+}/calmodulin-dependent channel desensitization. Mutant mice with a deletion in the CNGA4 gene are defective in odor adaptation. Thus, the Ca^{2+}-mediated negative feedback regulation of the CNG channel is essential for animals (and probably for humans as well) in continually assessing changes in odor environment and intensity in the presence of adapting background stimuli. Furthermore, the increase in Ca^{2+} has also been shown to mediate the Ca^{2+}/calmodulin kinase II-dependent attenuation of adenylyl cyclase activity.

Besides triggering adaptation, another important role of Ca^{2+} entering the olfactory cilia through the CNG channels is to mediate a special amplification mechanism based on the activation of a Ca^{2+}-gated Cl^- channel (Figure 4(c)). Normally, activation of Cl^- channels leads to membrane hyperpolarization. OSNs, however, maintain an unusually high intracellular Cl^- concentration of approximately 80 mM (with a corresponding Cl^- equilibrium potential of approximately −15 mV). Consequently, an intracellular increase in Ca^{2+} triggers Cl^- efflux through the activation of Cl^- channels, thereby further depolarizing the cells. The coexistence of ion fluxes through CNG channels and Ca^{2+}-gated Cl^- channels in the odor-induced sensory current may serve to reduce variations of the odor response resulting from changes in external ionic concentrations in the mucus surrounding the olfactory cilia. Identification of genes encoding this Cl^- channel and the targeted disruption of these genes should provide a more complete picture of the biological role of this amplification mechanism.

Not much attention has been directed toward understanding the downstream events responsible for the extrusion of Ca^{2+}. The Na^+/Ca^{2+} exchanger (NCX) is thought to be the primary Ca^{2+} extrusion mechanism in OSNs following odor stimulation. The Na^+ influx resulting from NCX activity is suggested to be recycled by an Na^+/K^+-ATPase. An additional pathway for Ca^{2+} extrusion could be provided by a plasma membrane calcium pump which has been described in OSNs of some species.

The Chemical Nature of Odorants

The main olfactory system has generally been regarded as a nonselective molecular analyzer of volatile odorants. Various lines of evidence have challenged this classical view, and a far greater complexity of chemosensory detection of olfactory cues has been suggested.

Terrestrial vertebrates aim their noses at regions of interest and sniff in order to acquire olfactory cues.

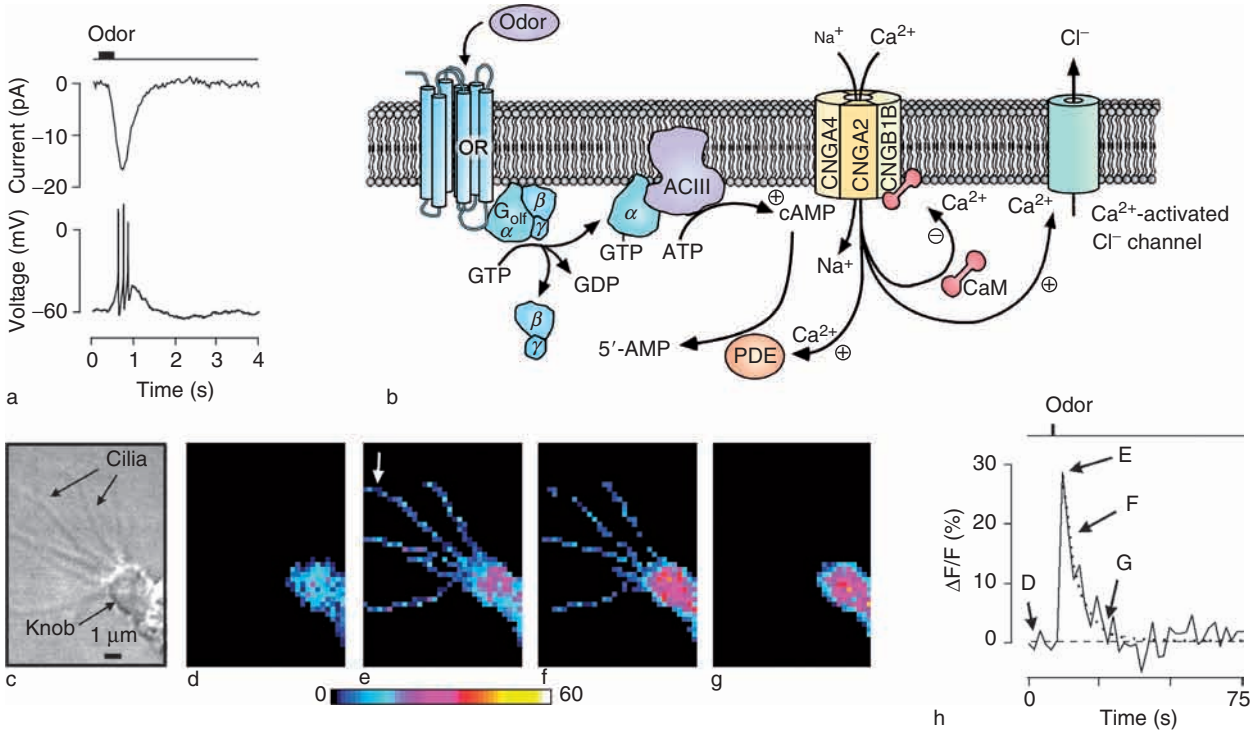

Figure 4 The cAMP-mediated chemoelectrical signal transduction cascade of odor-induced responses in the olfactory sensory neurons (OSNs). (a) Odor response measured in an isolated OSN using the perforated patch clamp technique. In the same OSN, the sensory current is measured using the voltage clamp technique at −60 mV and the receptor potential using the current clamp technique. Note the generation of a train of action potentials due to the receptor potential depolarization. (b) Schematic representation of the cAMP second messenger cascade present in the olfactory cilia of OSNs (as described in detail in the text). OR, olfactory receptor; G_{olf}, α subunit of heterotrimeric GTP-binding protein; ACIII, type III adenylyl cyclase; CNGA2/CNGA4/CNGB1b, subunits that form a heteromultimeric cyclic nucleotide-gated channel; PDE, phosphodiesterase; CAM, calmodulin; cAMP, adenosine 3′,5′-cyclic monophosphate. (c–h) Transient Ca^{2+} elevation in individual olfactory cilia following the application of a single pulse of an odor ligand using optical imaging techniques. (c) Phase contrast image of an OSN showing several cilia emanating from an olfactory knob. (d–g) Fluorescence images (pseudocolor scale) taken at rest in the absence of odor ligand (d), near peak fluorescence (e), and 4 s (f) and 16 s (g) after stimulation. At rest, the knob and part of the dendrite are visible, but resting Ca^{2+} fluorescence in the cilia is below the resolution limit. (h) Time course of an odor-induced ciliary Ca^{2+} transient (taken from the cilium labeled by white arrow). Labeling indicates the time points for which individual frames were printed. (c–h) Reprinted from Leinders-Zufall T, Greer CA, Shepherd GM, and Zufall F (1998) Imaging odor-induced calcium transients in single olfactory cilia: Specificity of activation and role in transduction. *Journal of Neuroscience* 18: 5630–5639, with permission from the Society for Neuroscience.

Schlieren imaging, an optical airflow visualization method, revealed that animals use specific exhale and inhale sniffing paradigms which enable them to sample olfactory cues of different sizes and volatility. Other fluid dynamic methods have indicated that fine charcoal particles (0.5–5 μm) can be found as far into the nasal cavity as a dog's olfactory ethmoturbinates. Bodily secretions such as urine contain a plethora of chemosignals, from small volatile organic molecules to short nonvolatile peptides and larger proteins. Some of the nonvolatile constituents present in urine and other bodily secretions have been shown not only to gain access to the MOE following direct physical contact with the source but also to be detected by its sensory neurons. It also was surprising that some of the chemosignals established earlier as sensory stimuli of vomeronasal organ (VNO) neurons (often incorrectly portrayed as pheromones because they activate these types of neurons) are recognized by OSNs. Thus, the olfactory system can be stimulated by ligands (common odors and pheromones) independent of their chemical volatility.

Guanylyl Cyclase D-Expressing Neurons of the Main Olfactory Epithelium

Not all of the chemosensory transduction in the mammalian nose is mediated by cAMP. Ciliated OSNs of the main olfactory epithelium comprise a heterogeneous population, with groups of cells displaying different immunochemical and histochemical markers or unique patterns of gene expression. One of the most intriguing subpopulations represents less than 1% of all OSNs and expresses a receptor guanylyl cyclase, GC-D (Figure 1). The chemosensory stimuli detected by these neurons have not been identified, but it is clear that these cells exhibit both unique anatomical and functional characteristics. Unlike most OSNs,

which project their axons to hundreds of glomeruli on all aspects of the MOB, GC-D neurons project to a small group of approximately nine glomeruli, the necklace glomeruli, which circumscribe the caudal MOB and have been implicated in neonatal behaviors such as suckling. In contrast to most OSNs, GC-D neurons continue to provide active input to their glomerular targets in CNGA2 null mutant mice. GC-D neurons express none of the transduction components characteristic of most OSNs, including ACIII, $G\alpha_{olf}$, CNGA2, the cAMP-specific phosphodiesterase PDE4A, or the Ca^{2+}/calmodulin-sensitive phosphodiesterase PDE1C2. However, GC-D neurons do express two proteins reminiscent of a cGMP-mediated transduction pathway: a CNG channel similar or identical to the cGMP-specific channel subunit expressed in cone photoreceptors (CNGA3) and a cGMP-stimulated PDE2. Thus, it seems highly likely that these chemosensory neurons use cGMP, but not cAMP, as a second messenger for signal transduction. The targeted disruption of genes, such as that encoding GC-D, should allow for specific testing of this hypothesis and help identify the chemosensory cues that are sensed by this unique subset of OSNs.

The Septal Organ of Masera

In addition to the main olfactory epithelium, the septal organ of Masera (SOM) is another epithelium containing chemosensory neurons. It is located bilaterally on the nasal septum near the entrance of the nasopharynx and is completely encircled by respiratory epithelium (Figure 1). This island of olfactory sensory neurons projects axons to the MOB and targets a subset of glomeruli located in the posterior ventromedial bulb. Neurotracing studies revealed that a few of these glomeruli receive inputs mainly (if not exclusively) from the SOM, whereas others receive mixed inputs from both the SOM and the MOE. The SOM has been described in various vertebrate species, including rat and mouse, but seems to be absent in some mammals, such as cat and ferret. It resembles the MOE in its cytoarchitecture containing chemosensory neurons, supporting and basal cells, as well as Bowman's glands. The morphology of the chemosensory septal organ neurons shows some unique features, including flattened somata, shorter dendrites, and larger dendritic knobs.

Identical mechanisms of olfactory transduction are observed in these neurons compared to the MOE. The majority of septal organ neurons express $G\alpha_{olf}$ and ACIII. Adenylyl cyclase inhibitors suppress its odor-induced responses, suggesting a critical role for the cAMP signaling pathway in these neurons. Again, similar to the MOE, there is a small subset of cells expressing a cGMP-stimulated phosphodiesterase (PDE2), a marker for GC-D neurons. These results support the notion that the SOM might have evolved by segregating from the ventral part of the MOE during prenatal development.

A relatively complete expression profile of the olfactory receptors in the SOM has been achieved by combining cDNA cloning, Affymetrix GeneChips covering all mouse olfactory receptor (OR) genes, and *in situ* hybridization. Although cDNA cloning using degenerate primers allows identification of more than 120 candidate receptors from a repertoire of more than 1000 ORs, the majority of the septal organ neurons express only a few receptors in varying patterns. All the major septal ORs are expressed in the most ventral, lateral area of the MOE. Their expression patterns, however, do not match those in the SOM, in which some receptors are present in unusually high densities. The most prevalent receptor (MOR256-3) is expressed in nearly 50% of the cells, and the eight most prevalent receptors are expressed in more than 90% of the cells. This finding made it possible to test whether a single SOM neuron expresses only one receptor. Double-labeling studies using a systematic combination of all the major OR genes expressed in the SOM confirmed this hypothesis. The function of the septal organ remains obscure. Physiological and behavioral studies are needed to elucidate the function of this enigmatic chemosensory organ. In terms of chemoreceptor composition, the SOM is the best described region in the mammalian nose.

The Grueneberg Ganglion

The Grueneberg ganglion (GG) is a compact cluster of neurons located bilaterally at the anterior end of the nasal septum (Figure 1). This ganglion was previously thought to be nonsensory and part of the terminal nerve system. However, studies have indicated that these neurons express olfactory marker protein, a marker for mature chemosensory neurons. The axons of these neurons fasciculate into several nerve bundles that terminate in a subset of atypical glomeruli, the same region as the glomerular target of the GC-D neurons. It is not clear if these two different chemosensory neurons share the same glomeruli or if these glomeruli have different functions depending on their innervation by GC-D or GG neurons. However, it appears that the GG is part of the olfactory system, although its function remains to be determined.

Conclusion

The vertebrate olfactory system is classically divided into two anatomically separate, and morphologically distinct, sensory systems: the accessory and the main olfactory system. The accessory olfactory system,

discussed elsewhere in this encyclopedia, consists of the vomeronasal organ (enclosed in the vomer bone) and targets the accessory olfactory bulb (**Figure 1**). The main olfactory system consists of the OSNs in the nasal cavity that project to the MOB. However, it has become evident that the nasal cavity does not contain only the main olfactory system with its OSNs utilizing the cAMP-mediated signal transduction cascade. As discussed previously, there are additional olfactory neuroepithelia present, including the SOM and the GG. These chemosensory subsystems contain more than one type of sensory neuron, which can be distinguished based on morphology, signal transduction cascade, or its target in the brain. Some of these cells are microvilli cells that have been detected in the MOE. A population of nonneuronal chemosensory cells is scattered within the rostral nasal cavity. These seem to be innervated by trigeminal afferent nerve fibers. The use of transgenic animals has resulted in the discovery of new subpopulations. One subtype of these newly recognized chemosensory neurons can be identified due to the presence of a specific ion channel, the transient receptor potential channel subunit TRPM5 (a key element in taste transduction). Thus, the mammalian olfactory system is composed of several distinct subsystems. Some of them, such as OSNs in the MOE and the sensory neurons in the VNO, have been shown to detect partially overlapping sets of social signals, giving rise to a model that involves parallel processing of the same molecules in these systems. This proposal implies that information about the same olfactory stimuli derived from each system is integrated in higher brain centers to regulate mammalian social behavior. Current interest is focused on the role that each of the olfactory subsystems plays in detecting molecular information and regulating chemosensory-dependent behaviors.

See also: Olfactory Neuron Patterning and Specification; Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Olfactory Higher Centers Anatomy; Electronic Nose; Olfactory System Theory; Pheromones in Mammals.

Further Reading

Buck L and Axel R (1991) A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* 65: 175–187.

Frings S, Reuter D, and Kleene SJ (2000) Neuronal Ca^{2+}-activated Cl^{-} channels: Homing in on an elusive channel species. *Progress in Neurobiology* 60: 247–289.

Fuss SH, Omura M, and Mombaerts P (2005) The Grueneberg ganglion of the mouse projects axons to glomeruli in the olfactory bulb. *European Journal of Neuroscience* 22: 2649–2654.

Giannetti N, Saucier D, and Astic L (1995) Analysis of the possible altering function of the septal organ in rats: A lesional and behavioral study. *Physiology & Behavior* 58: 837–845.

Kaluza JF, Gussing F, Bohm S, Breer H, and Strotmann J (2004) Olfactory receptors in the mouse septal organ. *Journal of Neuroscience Research* 76: 442–452.

Leinders-Zufall T, Greer CA, Shepherd GM, and Zufall F (1998) Imaging odor-induced calcium transients in single olfactory cilia: Specificity of activation and role in transduction. *Journal of Neuroscience* 18: 5630–5639.

Levai O and Strotmann J (2003) Projection pattern of nerve fibers from the septal organ: DiI-tracing studies with transgenic OMP mice. *Histochemistry and Cell Biology* 120: 483–492.

Ma M, Grosmaitre X, Iwema CL, Baker H, Greer CA, and Shepherd GM (2003) Olfactory signal transduction in the mouse septal organ. *Journal of Neuroscience* 23: 317–324.

Malnic B, Hirono J, Sato T, and Buck LB (1999) Combinatorial receptor codes for odors. *Cell* 96: 713–723.

Meredith M (1991) Sensory processing in the main and accessory olfactory systems: Comparisons and contrasts. *Journal of Steroid Biochemistry and Molecular Biology* 39: 601–614.

Mombaerts P (2004) Genes and ligands for odorant, vomeronasal and taste receptors. *Nature Neuroscience* 5: 263–278.

Morrison EE and Costanzo RM (1990) Morphology of the human olfactory epithelium. *Journal of Comparative Neurology* 297: 1–13.

Settles GS (2005) Sniffers: Fluid-dynamic sampling for olfactory trace detection in nature and homeland security – The 2004 Freeman Scholar Lecture. *Journal of Fluids Engineering* 127: 189–218.

Shepherd GM (2006) Smells, brains and hormones. *Nature* 439: 149–151.

Storan MJ and Key B (2006) Septal organ of Grueneberg is part of the olfactory system. *Journal of Comparative Neurology* 494: 834–844.

Tian H and Ma M (2004) Molecular organization of the olfactory septal organ. *Journal of Neuroscience* 24: 8383–8390.

Xu F, Schaefer M, Kida I, et al. (2005) Simultaneous activation of mouse main and accessory olfactory bulbs by odors or pheromones. *Journal of Comparative Neurology* 489: 491–500.

Zufall F, Firestein S, and Shepherd GM (1994) Cyclic nucleotide-gated channels and sensory transduction in olfactory receptor neurons. *Annual Review of Biophysics and Biomolecular Structure* 23: 577–607.

Zufall F and Leinders-Zufall T (2000) The cellular and molecular basis of odor adaptation. *Chemical Senses* 25: 473–481.

Relevant Websites

http://www.achems.org – Association for Chemoreception Sciences.

http://bip.weizmann.ac.il – Bioinformatics & Biological Computing; see Human Olfactory Receptor Data Exploratorium (HORDE).

http://www.leffingwell.com – Leffingwell & Associates: Services and Software for the Perfume, Flavor, Food and Beverage Industries.

http://md1.csa.com – ProQuert; see The Chemoreception Web.

http://senselab.med.yale.edu – SenseLab; see the Olfactory Receptor Database (ORDB).

Olfactory Glomeruli: Structure and Circuitry

M T Shipley and A C Puche, University of Maryland School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Glomeruli as Anatomical and Functional Units

The projection of receptor neurons to olfactory glomeruli is massively convergent, with thousands of olfactory receptor neurons (ORNs) converging onto each glomerulus of the mammalian olfactory bulb. Each of the several thousand ORNs expressing the same seven-transmembrane odorant receptor (OR) converges onto one, two, or a few glomeruli in the olfactory bulb (OB). The result of this organization is such that, at the level of sensory input to the OB, OR identity is mapped across the surface of the OB, with glomeruli serving as the anatomical unit of this map. Several lines of evidence indicate that the glomerulus also serves as a functional unit in representing odor information. First, ORNs expressing the same OR show similar odorant response specificities *in vitro*, implying that ORN input to a glomerulus is functionally homogeneous. Second, glomeruli can be discretely activated in an odorant-specific manner as seen using a variety of methods including 2-deoxyglucose mapping, functional magnetic resonance imaging, intrinsic optical imaging, imaging with fluorescent activity probes such as calcium-sensitive dyes and voltage-sensitive dyes, genetically encoded activity probes, and electrophysiology recordings. Third, odors activate multiple glomeruli with overlapping, but different, response specificities. Thus, information about odorant identity is carried in the combination and relative level of activation of different glomeruli. This feature is not surprising given that the ORs which are represented by glomeruli also show distinct but overlapping odorant specificities. Together, these data suggest that, at the level of ORN input, the glomerulus is the smallest detectable unit contributing to the functional representation of odor information.

Glomerular activity maps show a loosely organized chemotopy across the OB. In mammals, structurally similar odorants preferentially activate glomeruli in similar OB regions. These 'modules' or 'domains' encompass between 10 and 50 glomeruli. As with the glomeruli, each module is duplicated on the medial and lateral halves of the bulb and is roughly conserved across individuals. Glomerular modules have only been defined functionally (i.e., based on their odorant response properties); it is unclear whether they correspond to particular OR subclasses or even how sharply delineated are the boundaries between different modules.

Intrinsic and Synaptic Properties of Juxtaglomerular Neurons

The spatial and temporal characteristics of glomerular activity maps discussed to this point are the results of properties inherent in ORNs and the 'rules' governing their glomerular targeting. These characteristics neither involve nor require interactions among ORNs. Upon synapsing in the glomerulus, ORNs release glutamate and engage postsynaptic circuits that initiate the neural processing of olfactory input. From this point on, neural processing involves not only the intrinsic properties of neurons but also the synaptic relationships that organize them into neural circuits.

Glomeruli are composed of a 'shell' of juxtaglomerular (JG) neurons and astrocyte cell bodies surrounding a globular assembly of neuropil. The neuropil contains the synaptic contacts between ORN axons, the apical dendrites of mitral and tufted cells, the dendrites of JG neurons, and synaptic linkages among these neurons. Although classical neuroanatomists recognized at least three types of JG neurons, there has been a tendency to view the glomerulus as composed of 'periglomerular' cells, which function as local GABAergic interneurons that negatively modulate ORN to mitral/tufted cell synaptic throughput. Research indicates that this view is too simplistic and that each glomerulus contains complex micronetworks that perform multiple, distinct operations on ORN input.

Early Golgi and electron microscopy (EM) studies identified three broad types of JG neurons: external tufted (ET), periglomerular (PG), and short axon (SA) cells. These cells can be distinguished not only by their morphologies but also by equally distinctive physiological characteristics, including intrinsic and synaptic properties (Figure 1). The advent of modern chemoanatomical methods revealed an additional level of complexity by showing that different JG cell subpopulations express a variety of neurotransmitters, receptors, and calcium-binding proteins. Cellular physiology analyses of JG neurons led to the recognition that the synaptic circuits are also are far more complex and play more roles in olfactory processing than previously appreciated.

ET cells are distinguished by their elaborate dendritic arbor, which ramifies extensively throughout a single glomerulus. The ET cell dendrite is actually

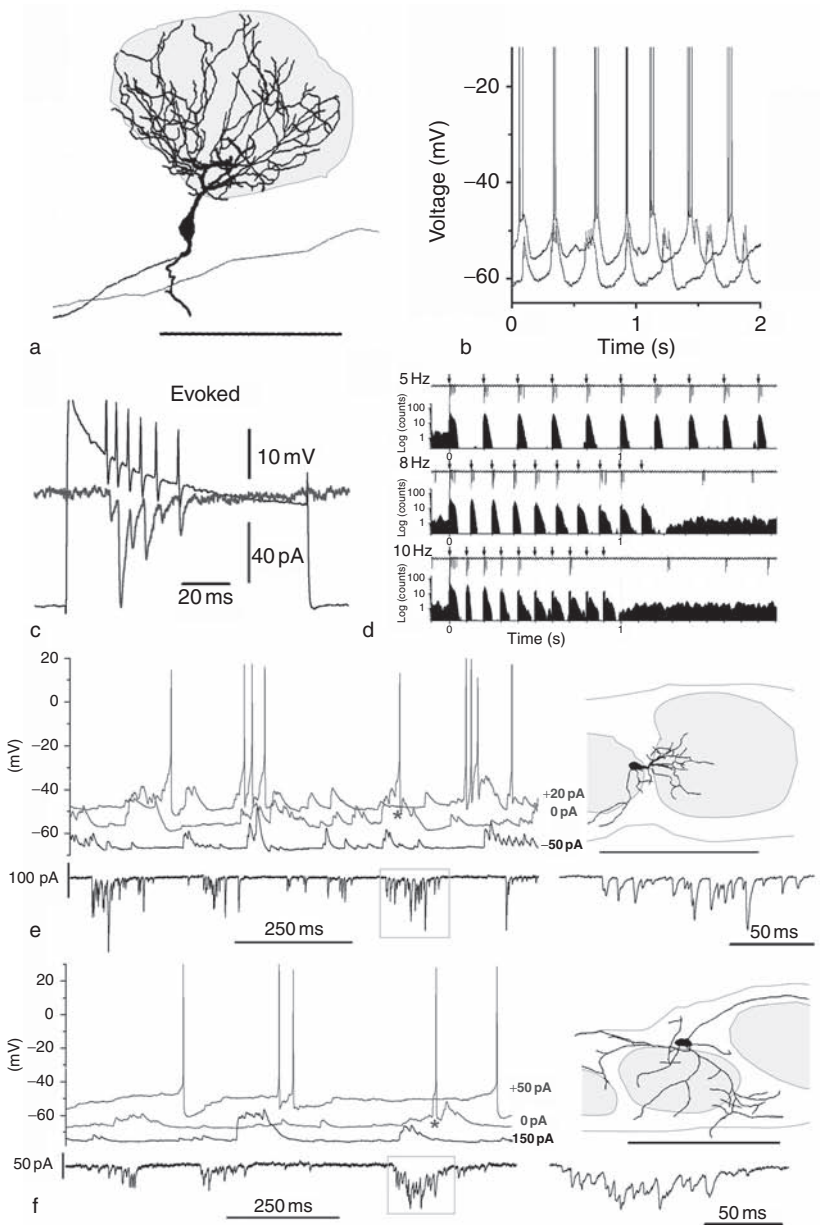

Figure 1 Physiology and morphology of ET, PG, and SA cells. ET cells: (a) reconstruction of a biocytin-filled ET cell. Note the highly branched tufted dendrites that ramify throughout only a single glomerulus. (b) The same ET cell in cell-attached mode and whole cell configuration. (c) ET cells monosynaptically excite PG and SA cells. Dual recordings from an ET cell in cell-attached (bottom traces) and a PG cell in whole cell voltage clamp (top traces). Example of a burst of EPSCs in a PG cell (bottom trace) evoked by a burst of spikes (top trace) elicited by extracellular application of a positive current pulse in a paired ET cell. Note the short, invariant latency between evoked spikes in the ET cell and EPSCs in the PG cell. (d) ET cells are entrained by patterned ON input across a frequency spectrum similar to breathing/sniffing frequencies. Peristimulus time histograms showing the activity of an ET cell in response to a train of 10 ON stimuli (arrows) delivered at 5 Hz (top), 8 Hz (middle), and 10 Hz (bottom). PG cells: (e) (top traces) current-clamp recordings show effects of steady hyperpolarizing or depolarizing current injection in a typical PG cell. Depolarizing or hyperpolarizing current injections do not generate spike bursts. The cell receives spontaneous long-duration bursts of EPSCs and generates spikes infrequently, notably during bursts of EPSPs (asterisk). (Bottom traces) Voltage-clamp recording (HP = −60 mV) shows bursts of spontaneous EPSCs in the same PG cell; the region enclosed in the box is shown at a faster timescale at right. (Right) Reconstruction of the recorded cell shows typical morphology of PG cells, including relatively small soma and one to three relatively thick primary dendritic shafts that give rise to thinner branches ramifying within a subregion of a single glomerulus. SA cells: (f) the resting activity of SA cells and their responses to current injection are similar to those of PG cells. SA cells receive bursts of spontaneous EPSCs. Reconstruction of the recorded cell at left shows typical morphology of SA cells, including several poorly branched dendrites extended into or between two to four glomeruli. (a, b, d–f) Reproduced from Hayer A, Karnup S, Shipley MT, and Ennis M (2004) Olfactory bulb glomeruli: External tufted cells intrinsically burst at theta frequency and are entrained by patterned olfactory input. *Journal of Neuroscience* 24:1190–1199, with permission from Society for Neuroscience. (c) Reproduced from Hayer A, Karnup S, Ennis M, and Shipley MT (2004) External tufted cells: A major excitatory element that coordinates glomerular activity. *Journal of Neuroscience* 24: 6676–6685, with permission from Society for Neuroscience.

more extensive on average than that of the mitral cell. A second distinctive morphological feature is that ET cells rarely extend lateral dendrites and, when present, do not enter into the external plexiform layer (EPL). Thus, unlike mitral and classical (superficial, middle, and deep) tufted cells, ET cells are not regulated by inhibitory inputs from granule cells. ET cells have a single axon that exits the glomerular layer and traverses the EPL without branching. The target(s) of ET cell axons is unclear, but they do appear to project beyond the bulb at least as far as the anterior olfactory nucleus.

The most distinctive characteristic of ET cells, however, is physiological: ET cells spontaneously generate rhythmical bursts of action potentials. Their ability to generate rhythmic bursts is due to intrinsic membrane properties and not to rhythmic synaptic input. Indeed, when synaptic transmission is pharmacologically blocked, the bursting frequency of each ET cell becomes more regular. ET cell bursting depends on the interplay of several intrinsic conductances, including a persistent sodium current that appears to be necessary to initiate the burst, an L/T-type calcium current that sustains the burst and gates a Ca^{2+}- and voltage-dependent K^{+} current, which terminates the burst. A second fundamental property of ET cells is that they all receive monosynaptic input from the olfactory nerve. This is in sharp contrast to the majority of PG cells and all SA cells, which do not receive monosynaptic olfactory nerve (ON) input.

PG cells have a less elaborate dendritic arbor that occupies a much smaller glomerular region than that of ET cells. Their physiological characteristics are also very different from ET cells. PG cells do not generate rhythmic burst of action potentials and in general have infrequent spontaneous action potentials. The intrinsic membrane properties of PG cells are more heterogeneous than those of ET cells, perhaps corresponding to their more heterogeneous morphologies and heterogeneous expression of transmitters, receptors, and calcium-binding proteins. PG cells can be divided into two groups based on synaptic input. Approximately 70–80% of all PG cells receive distinctive rhythmic bursts of spontaneous synaptic input and do not receive direct ON input; the remaining 20–30% receive direct olfactory nerve input but do not exhibit bursts of spontaneous synaptic events. Thus, although classic EM studies suggested that all PG cells receive monosynaptic ON input, only approximately 20–30% actually do. As discussed later, the vast majority of PG cells receive olfactory input via ET cells.

SA cells are the third and historically most obscure of the classical JG cell types. Studies have disclosed that SA cells play a key role in a newly discovered interglomerular lateral inhibitory network (Figure 2). Consistent with classical descriptions, SA cell dendrites – unlike those of ET and PG cells – ramify among several neighboring glomeruli. However, in contrast to their name, 'short axon' cells actually give rise to a rich and extensive interglomerular axonal network, which projects to glomeruli as far as 1 mm distant. The contradiction between their classical name and their actual morphology probably reflects the failure of the Golgi method to adequately impregnate their fine axons. Physiologically, the spontaneous activity of SA cells is very similar to that of the majority of PG cells; they exhibit rhythmic bursts of spontaneous synaptic events and do not receive monosynaptic ON input. Perhaps due to their extensive axonal arbor, SA cells have larger cell bodies and lower input resistances but otherwise are difficult to distinguish from PG cells based on spontaneous and ON-evoked activity.

Glomerular Circuits

The intrinsic neurons that comprise the glomerular network – ET, PG, and SA cells – are more complex and physiologically diverse than previously suspected. Based on new information on the major synaptic relationships among ET, PG, and SA neurons, at least four principal glomerular circuits can be postulated: the ON→M/T circuit, the ON→PG circuit, the ON→ET→PG circuit, and the ON→ET→SA circuit (Figure 3).

The ON→M/T→PG Circuit

A fundamental canon of olfactory circuitry is that ON terminals synapse directly on the apical intraglomerular dendrites of M/T cells (i.e., an ON→M/T circuit). Based on ultrastructural features, classical EM studies distinguished two types of intraglomerular dendrites postsynaptic to ON terminals: M/T cells and PG cells. These studies also posited that M/T cell intraglomerular dendrites are presynaptic to PG cells, leading to the idea that ON-evoked excitation of M/T cells causes M/T excitation of PG cells, thus providing feedback inhibition of M/T cells. This view needs to be reconsidered in light of new information about ET and PG cells. ET dendrites ramify at least as extensively within the glomerulus as M/T dendrites, and all ET cells receive monosynaptic ON input. The morphological criteria used to identify M/T cells in EM studies would certainly have included ET cell dendrites. Therefore, it is likely that some (perhaps appreciable) fraction of the ON→M/T synapses identified by classical EM studies were actually ON→ET synapses.

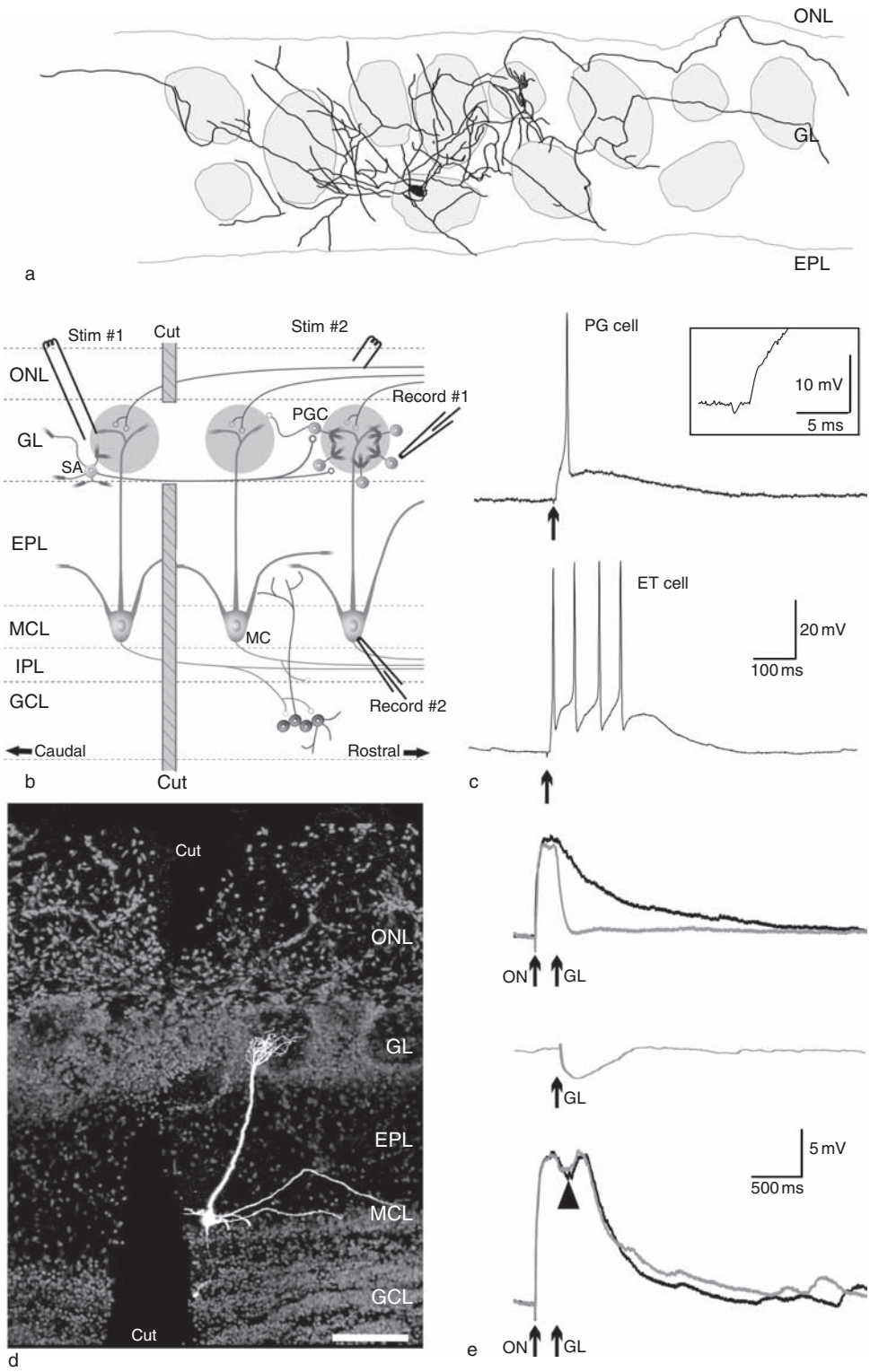

Figure 2 Center surround inhibitory network among olfactory bulb glomeruli. (a) Reconstruction of a biocytin-filled SA cell showing extensive branched axonal projections across multiple glomeruli and a less complicated dendritic projection (dendrites, gray; axon, black). (b) Diagram showing center surround inhibition via excitatory SA cells. (c) PG and ET cells responded with EPSPs to distant glomerular layer stimulation. The inset shows a 2 ms response latency, consistent with a monosynaptic event. (d) Biocytin-filled mitral cell (white) and nuclear stain (gray). Cuts through the ONL and deep bulb layers isolating the interglomerular circuit are clearly visible. (e) (Top trace) Mitral cells respond to ON stimulation with an LLD (black trace) that is abruptly terminated by stimulation of a distant glomerulus (arrow, gray trace). (Middle trace) Stimulation of interglomerular connections evokes an IPSP in mitral cells. (Lower trace) In the presence of gabazine, olfactory nerve evoked LLDs are no longer suppressed by glomerular layer stimulation (black trace, ON stimulation alone; gray trace, ON plus interglomerular stimulation).

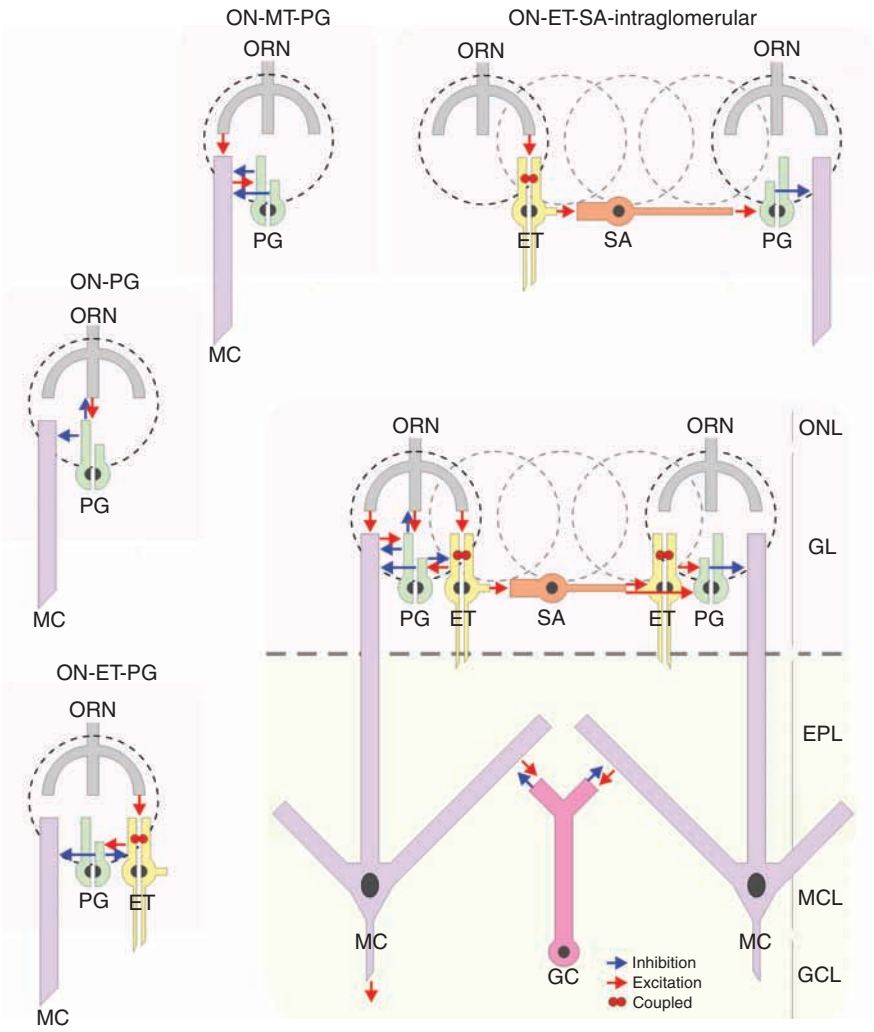

Figure 3 Principal glomerular circuits. The small panels show four principal glomerular circuits: ON→M/T→PG, ON→PG, ON→ET→PG, and ON→ET→SA→PG→M/T interglomerular circuit. The central panel shows all four glomerular circuits together with the mitral–granule–mitral (M→G→M) circuit.

What about the physiological evidence for monosynaptic ON inputs to M/T cells? Mitral cells respond to ON stimulation with a short latency, low-variance excitatory postsynaptic current (EPSC). Thus, there is physiological support for the conclusion that M/T cells receive monosynaptic ON input. However, this initial response is followed by a series of longer, more variable latency EPSCs that sum to produce large, long-lasting events termed 'long-lasting depolarizations' (LLDs). The events contributing to the LLD are too long latency and recur over too long a time frame to be due to monosynaptic ON input. LLDs have been attributed to recurrent input via 'glutamate spillover' from the apical dendrites neighboring mitral and/or tufted cells. The finding that ET cells generate prolonged, repetitive bursts of action potentials in response to monosynaptic ON input suggests the possibility that ON-evoked ET cell bursts also play a role in LLDs by amplifying and prolonging ON activation of M/T cells. Additional investigation is needed to determine if the ON→M/T excitatory circuit is supplemented by ON→ET→M/T excitation.

The ON→PG Cell Circuit

Classical EM studies also showed that PG cells receive ON synaptic contacts, leading to the inference that ON input directly excites PG cells, which in turn provide feed-forward inhibitory input to MT cells. Therefore, the second principal glomerular circuit is the ON→PG circuit. In contrast to the classical anatomical picture, electrophysiological studies show that the olfactory nerve provides monosynaptic excitatory input to only approximately 20–30% of all

PG cells. Thus, PG cells are divided into two broad groups on the basis of synaptic input – the approximately 20–30% of 'mono' PG cells which receive direct ON input and the 70–80% of 'poly' PG cells which receive indirect polysynaptic ON input. However, although less is known about its influence on the ON→PG cell circuit, it should not be dismissed as 'minor' simply because it comprises only 20–30% of the PG cell population. PG cells comprise approximately 70% of the entire JG cell population; therefore, the 'minor' monoPG cells are as numerous as ET or SA cells. The ON→PG cell circuit may correspond to the classical ON→PG→M/T cell circuit postulated by early EM studies. This circuit might operate in parallel with the ON→ET→PG→M/T circuit (discussed later), both providing delayed feed-forward inhibition of M/T cells.

Another established role for the ON→PG circuit is presynaptic inhibition of ON terminals via an ON→PG→ON circuit. Feedback from PG cells onto ORN axon terminals suppresses transmitter release (i.e., presynaptic inhibition). In mammalian glomeruli, both GABA and dopamine are involved in this presynaptic inhibition. ORNs express $GABA_B$ and dopamine D_2 receptors on their axon terminals, and both $GABA_B$ and D_2 agonists suppress postsynaptic activity evoked by ON shock. Direct depolarization of GABAergic periglomerular interneurons (PG cells) also suppresses ON input, indicating that the circuit mediating feedback presynaptic inhibition is intrinsic to the glomerular layer. Presynaptic inhibition of ORNs is predominately mediated by a suppression of calcium influx through N-type voltage-sensitive calcium channels.

The ON→ET→PG Circuit

ON provides monosynaptic input to all ET cells, which in turn monosynaptically excite PG cells. ON-evoked monosynaptic EPSCs in ET cells are followed by a longer latency, compound inhibitory postsynaptic current, which is due to GABAergic input from PG cells. Thus, one extension of the ON→ET→PG circuit is an ON→ET→PG→ET feedback circuit. The temporal sequence of excitation followed by feedback inhibition generated by this circuit may function to shape postsynaptic responses to sensory input possibly by creating a temporal 'window of opportunity' for sensory signals to achieve glomerular throughput and influence downstream olfactory circuits. Since only 20% of PG cells receive monosynaptic ON input, some of the intraglomerular inhibition of M/T cell dendrites, classically attributed to the ON→M/T→PG→M/T circuit, may actually be due to the operation of an ON→ET→PG→MT cell circuit.

The ON→ET→SA Interglomerular Circuit

SA cells lack monosynaptic ON input but do receive excitatory monosynaptic input from ET cells. Thus, the fourth principal glomerular circuit is the ON→ET→SA cell circuit. SA cells give rise to a rich axonal plexus that contacts multiple glomeruli up to a millimeter distant. The existence of this interglomerular network suggested that SA cells mediate interactions among glomeruli. Experiments using voltage-sensitive dyes showed these interglomerular connections to be excitatory. SA cells make glutamatergic synapses with neurons in distant glomeruli and stimulation of interglomerular axons produce EPSCs in PG and ET cells of target glomeruli. This raised the possibility that the interglomerular circuit could inhibit M/T cells via an SA→PG→M/T cell circuit. This was confirmed by showing that stimulation of interglomerular axons produced a compound inhibitory postsynaptic potential (IPSP) in mitral cells; the IPSP was blocked by the $GABA_A$ receptor antagonist gabazine. The inhibition produced by activating the interglomerular circuit was sufficient to terminate ON-evoked LLDs in mitral cells. Thus, there is a network composed, minimally, of an ON→ET→SA→(distal glomerulus)→PG→M circuit, which provides interglomerular lateral inhibition. Strong ON activation of a given glomerulus activates the ON→ET→SA→(distal glomerulus)→PG→M circuit, which inhibits mitral cells – and thus glomerular throughput – in weakly activated glomeruli. Such lateral inhibition could function to enhance contrast among differentially activated glomeruli.

The length of SA axons does not necessarily limit the spatial propagation of interglomerular activity. Since stimulation of the SA interglomerular axons excites ET cells of distant glomeruli, it is possible that ET cells excite additional local SA cells which further inhibit more distant glomeruli, via an ON→ET→SA→(distal glomerulus) ET→SA→(distal glomerulus) multisynaptic interglomerular circuit. Mathematical modeling of this propagating network suggests that the interglomerular circuit may provide input to all glomeruli that represents the average of glomerular activity over much of the bulb. A consequence of this is that an overall increase in glomerular activity, such as during high-concentration odor exposure, would result in a higher level of activity in the interglomerular network and a higher level of inhibition onto mitral cells. This putative averaging function could 'normalize' postsynaptic activity to changes in odor concentration.

The External Tufted Cell: A Glomerular Gatekeeper?

The ET cell is a key player in the glomerular circuitry. ET cells are the primary JG cell targets of monosynaptic ON input. Only 20–30% of PG cells and no SA cells are synaptically contacted by ON terminals. Thus, ET cells are ideally positioned to mediate sensory input to other PG and SA cells. Simultaneous dual whole cell recordings from ET and PG or SA cells show that spontaneous and ON-evoked bursts of action potentials in ET cells are highly correlated with subsequent bursts of EPSCs in simultaneously voltage-clamped PG/SA cells, suggesting that the ET cell population provides excitatory input to PG and SA cells. In many paired recordings, spike-triggered averaging revealed activity correlations indicative of monosynaptic ET to PG/SA cell input. In such cases, direct stimulation of the ET cell generated short, invariant latency EPSCs in the PG/SA cell. These events were blocked by an AMPA/kainate glutamate receptor antagonist, demonstrating that ET cells provide monosynaptic, glutamatergic input to PG and SA cells. Thus, the ET cell directly mediates olfactory nerve input to the majority of other intrinsic glomerular neurons. The additional fact that ET cells respond to suprathreshold ON input with a burst of action potentials suggests that ET cells amplify and/or increase the synaptic security of sensory input to other postsynaptic glomerular targets. Thus, the ET cell is a key excitatory element in the glomerular network.

Two properties of ET cells suggest that glomerular circuits are specialized to deal with the dynamics that sniffing confers upon odorant sampling. One property results from ET cell bursting and the second from the observation that ET cells of the same glomerulus have coupled (i.e., quasi-synchronous) activity which may function, in part, to coordinate and possibly synchronize intraglomerular activity.

Synchrony

ET cells of the same glomerulus have correlated spontaneous activity, which is expressed at the level of action potentials as well as oscillations of membrane potential. ET cells associated with different glomeruli do not have such correlated activity. Thus, the mechanisms that couple ET cells are of intraglomerular origin. Several intraglomerular synaptic and electrical coupling mechanisms have been identified, indicating that both synaptic transmission and gap junction coupling contribute to this correlated spontaneous bursting of ET cells. The coupled activity of ET cells of the same glomerulus, in addition to the fact that two of the four principal glomerular circuits – ON→ET→PG and ON→ET→SA – are primarily driven by the ET cell, suggests that glomerular circuits are specialized for highly coordinated activity. This coordinate activity may function to synchronize and temporally gate or window glomerular output.

Several lines of evidence indicate that glomerular output is synchronized by intraglomerular mechanisms. ON stimulation produces a short-latency EPSC in mitral cells. This short-latency monosynaptic EPSC is then augmented by recurrent EPSCs, leading to a long-lasting (400–800 ms) depolarization. Mitral cells whose dendrites ramify in the same glomerulus have synchronous LLDs, whereas LLDs in mitral cells associated with different glomeruli are not correlated. Thus, LLDs are initiated within the glomerulus and intraglomerular mechanisms synchronize at least some aspects of mitral cell activity. Stronger ON stimulation also triggers slow (2 Hz) oscillations in mitral cells. These events, like LLDs, and conventional EPSPs are also highly synchronous in mitral cells associated with the same glomerulus. Depolarization of mitral cell pairs generates action potentials that are correlated on a timescale of $\leq$10 ms for cells associated with the same glomerulus. This correlated spiking appears to depend on both intraglomerular electrical (gap junction) and chemical (AMPA receptors) coupling. Thus, at several temporal levels, the activity of mitral cells of the same glomerulus is synchronous and the mechanisms regulating this synchronous activity are of intraglomerular origin. ET cells may also play a role in this synchronization. A single shock to the olfactory nerve generates a brief monosynaptic excitatory response in ET cells. This brief input is converted into a prolonged output by the burst of action potentials. ET cell bursts could contribute to the synchrony of mitral/tufted activity cells via glutamate spillover and gap junctions in a manner similar to mitral cells whose dendrites are associated with the same glomerulus.

ET cells monosynaptically excite PG cells, which in turn provide intraglomerular feedback and inhibition of ET cells. This arrangement is similar to the mitral→granule→mitral cell (M→G→M) circuit in which mitral cells excite granule cells, which in turn provide feedback to the same and lateral inhibition to other mitral cells. The M→G→M circuit provides contrast enhancement of strongly versus weakly activated glomeruli (i.e., interglomerular spatial contrast enhancement). It is not obvious what functional role would be served by intraglomerular spatial contrast enhancement via the ET→PG→ET circuit. However, the M→G→M circuit also provides a temporal

windowing function via feedback inhibition of mitral cells. After strong excitation, M→G→M feedback inhibition temporarily reduces mitral cell excitability, thus reducing sensitivity to sensory input. The ET→PG→ET circuit is well suited to provide a similar intraglomerular temporal windowing function. Thus, the ET→PG→ET intraglomerular circuit may function first to amplify olfactory nerve input and second to generate a temporal window, within which sensory inputs are integrated to generate synchronous glomerular output.

ET Bursting and Sniffing

The ability of ET cells to generate rhythmical bursts of action potentials is due entirely to ET cell intrinsic membrane properties. When synaptic inputs are blocked with receptor antagonists, not only do ET cells continue to rhythmically burst but also the bursting frequency becomes significantly more regular. Under such conditions, each ET cell expresses a bursting frequency characteristic for that cell. This 'native' or 'steady-state' bursting frequency is different for each cell. The distribution of bursting frequencies across the entire population of ET cells comprises a relatively narrow spectrum ranging from approximately 0.2 to 10 Hz. This is nearly identical to the spectrum of normal breathing and sniffing frequencies exhibited by rats and mice. This correspondence suggested that the ability of ET cells to generate intrinsic bursting might be a physiological specialization that allows them to respond optimally to rhythmical (i.e., sniff-related odorant input). Indeed, ET cells follow repetitive ON stimulation at all frequencies in the range from approximately 0.2 to 10 Hz and a given ET cell follows (i.e., is entrained) by frequencies at or above the cell's native frequency.

This finding has important functional implications. As the frequency of sniffing increases, more ET cells will be entrained as each cell follows all input frequencies above its native bursting frequency. Since ET cells provide the excitatory input to the majority of PG and to all SA cells, as the number of entrained ET cells increases with higher sniffing frequency, the excitatory drive on and temporal synchrony of PG and SA cells will increase. The increased excitatory drive on PG cells will increase the inhibitory feedback onto ET cells via the ON→ET→PG→ET circuit and onto M/T cells via the ON→ET→PG→M/T circuit. The combination of ON→ET entrainment and PG→ET feedback inhibition should generate an increasingly narrow temporal window for the integration of sensory input as sniff frequency increases. This may impose critical timing constraints on glomerular throughput. Sniff frequency should also impact the dynamics of the ON→ET→SA→interglomerular circuit. As sniffing frequency increases, more ET cells are entrained and the excitatory drive on SA cells will increase and become more synchronous. This should increase SA excitatory drive on PG cells in distant glomeruli, which in turn will increase GABAergic inhibition of M/T cells in those glomeruli. Thus, at higher sniffing frequencies, interglomerular inhibition should increase, with the result that more strongly (and synchronously?) activated glomeruli should more effectively inhibit weakly activated glomeruli – that is, interglomerular contrast enhancement should increase in proportion to sniffing frequency.

Role of Glomerular Processing in Shaping Odor Representations

How might the presynaptic and postsynaptic glomerular network function together to shape spatial and temporal patterns of activity as olfactory input is transmitted from ORNs to MT cells? We speculate that this network is optimally organized to impose brief temporal windows over which MT cells can be excited by rhythmic ORN inputs while at the same time maintaining responsiveness across a wide range of input strengths. We further speculate that spatial patterns of MT activity are sharpened via interglomerular inhibition. The interglomerular circuit may also normalize relative levels of glomerular activity in the face of changing concentration. A key feature of the circuitry underlying these functions is that the strength and dynamics of both intra- and interglomerular inhibition are dependent on the frequency with which odorants are sampled via sniffing. At low sniff frequencies characteristic of anesthetized and resting awake animals (1–3 Hz), ORN input to glomeruli occurs as brief bursts which are repeated with each inspiration and which drive transient activity in postsynaptic neurons. Intraglomerular excitatory mechanisms synchronize MT firing with the transient ORN inputs, whereas intraglomerular inhibitory pathways (mediated by the ET→PG cell circuit) mediate fast $GABA_A$-mediated inhibition of MT cells and slow $GABA_B$-mediated presynaptic inhibition of ORN input. The dynamics of the slower presynaptic inhibition are ideally suited to suppress ORN input in a manner that is graded with sniff frequency without affecting the dynamics of input evoked by an individual sniff. Thus, we hypothesize that as sniff frequency increases and more ET cells are entrained to fire synchronously, ORN inputs become more strongly suppressed. At the same time, the strength of fast inhibitory input to MT cells

increases as more ET cells are recruited, preventing excessive autoexcitation and further sharpening the temporal window over which excitatory inputs are integrated.

How could this circuitry affect spatial patterns of MT cell activity? First, given that the latency and rise time of sniff-driven ORN input to a glomerulus varies across glomeruli and across odorants, the temporal window over which inputs are able to excite MT cells is critical in determining how spatial maps of input to glomeruli are transformed into MT cell response patterns. Because this temporal window narrows at higher sniff frequencies, one possibility is that only MT cells innervating glomeruli receiving the fastest onset input are activated. At the same time, increased ET cell activation at higher sniff frequencies would be expected to increase the strength of interglomerular inhibition via the ET→SA→PG cell network, additionally sharpening MT response patterns such that MT cells innervating glomeruli receiving weaker and longer latency input are not excited. These processes are conjectural, but one implication of this organization is that the degree to which spatial patterns of glomerular activity are sharpened by glomerular processing can change depending on the frequency of sniffing. This hypothesis, which is testable, implies that odor representations depend not only on the identity of the olfactory stimulus but also on the way in which it is sampled.

See also: Olfactory Coding; Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping.

Further Reading

Aungst JL, Heyward PM, Puche AC, et al. (2003) Centre-surround inhibition among olfactory bulb glomeruli. *Nature* 426: 623–629.

Bozza T, Feinstein P, Zheng C, and Mombaerts P (2002) Odorant receptor expression defines functional units in the mouse olfactory system. *Journal of Neuroscience* 22: 3033–3043.

Carlson GC, Shipley MT, and Keller A (2000) Long-lasting depolarizations in mitral cells of the rat olfactory bulb. *Journal of Neuroscience* 20: 2011–2021.

Cleland TA, Johnson BA, Leon M, and Linster C (2007) Relational representation in the olfactory system. *Proceedings of the National Academy of Sciences of the United States of America* 104: 1953–1958.

Hayar A, Karnup S, Ennis M, and Shipley MT (2004) External tufted cells: A major excitatory element that coordinates glomerular activity. *Journal of Neuroscience* 24: 6676–6685.

Hayar A, Karnup S, Shipley MT, and Ennis M (2004) Olfactory bulb glomeruli: External tufted cells intrinsically burst at theta frequency and are entrained by patterned olfactory input. *Journal of Neuroscience* 24: 1190–1199.

Hayar A, Shipley MT, and Ennis M (2005) Olfactory bulb external tufted cells are synchronized by multiple intraglomerular mechanisms. *Journal of Neuroscience* 25: 8197–8208.

Hildebrand JG and Shepherd GM (1997) Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. *Annual Review of Neuroscience* 20: 595–631.

Johnson BA, Woo CC, and Leon M (1998) Spatial coding of odorant features in the glomerular layer of the rat olfactory bulb. *Journal of Comparative Neurology* 393: 457–471.

McGann JP, Pirez N, Gainey MA, Muratore C, Elias AS, and Wachowiak M (2005) Odorant representations are modulated by intra but not interglomerular presynaptic inhibition of olfactory sensory neurons. *Neuron* 48: 1039–1053.

Murphy GJ, Darcy DP, and Isaacson JS (2005) Intraglomerular inhibition: Signaling mechanisms of an olfactory microcircuit. *Nature Neuroscience* 8: 354–364.

Parrish-Aungst S, Shipley MT, Erdelyi F, Szabo G, and Puche AC (2007) Quantitative analysis of neuronal diversity in the mouse olfactory bulb. *Journal of Comparative Neurology* 501: 825–836.

Pinching AJ and Powell TP (1971) The neuron types of the glomerular layer of the olfactory bulb. *Journal of Cell Science* 9: 305–345.

Schoppa NE and Westbrook GL (2001) Glomerulus-specific synchronization of mitral cells in the olfactory bulb. *Neuron* 31: 639–651.

Schoppa NE and Westbrook GL (2002) AMPA autoreceptors drive correlated spiking in olfactory bulb glomeruli. *Nature Neuroscience* 5: 1194–1202.

Toida K, Kosaka K, Aika Y, and Kosaka T (2000) Chemically defined neuron groups and their subpopulations in the glomerular layer of the rat main olfactory bulb. IV: Intraglomerular synapses of tyrosine hydroxylase-immunoreactive neurons. *Neuroscience* 101: 11–17.

Wachowiak M and Cohen LB (2001) Representation of odorants by receptor neuron input to the mouse olfactory bulb. *Neuron* 32: 723–735.

Olfactory Higher Centers Anatomy

J L Price, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Published by Elsevier Ltd.

Introduction

As with other sensory modalities, olfactory information must be transmitted from peripheral olfactory structures (the olfactory epithelium) to more-central structures (the primary olfactory cortex), integrated to detect and discriminate specific stimuli, and then transferred to other parts of the brain in order to reach sensory consciousness and affect behavior. The central olfactory system, however, and particularly the olfactory cortex, is different from the other sensory systems in two key ways.

First, olfactory sensory activity is transferred directly from the olfactory bulb to the primary olfactory cortex, and from there to the orbital neocortex, without a thalamic relay. Although there is a projection from the olfactory cortex to the mediodorsal thalamic nucleus, which is reciprocally connected with the posterior orbital/agranular insular cortex, this transthalamic pathway involves fewer neurons than the monosynaptic projection from the primary olfactory cortex to the same orbital/insular areas. The transthalamic projection is therefore not required for relay of olfactory sensory information to the neocortex.

The second difference is that the analysis of olfactory stimuli may not involve a topographic organization beyond the olfactory bulb. Olfactory stimuli are not distributed along an obvious spatial or frequency map, and it has been difficult to determine how the stimuli are organized. An organization has been found in the projection from the olfactory epithelium to the olfactory bulb, but there is little solid indication of a topographic organization from the bulb to the olfactory cortex. Instead, axons of single mitral or tufted cells have multiple axon collaterals that make synapses throughout a large part of the olfactory cortical areas. A more complex organization may eventually be discovered, but on the present evidence, it is also likely that specific odors are discriminated by differences in the spatiotemporal pattern of activity across large regions of the olfactory cortex.

Axons from the olfactory bulb extend through the olfactory tract to a varied set of cortical regions on the surface of the ventral forebrain that can generally be referred to as the olfactory cortical areas or primary olfactory cortex (Figures 1 and 2). The olfactory tract is often called the lateral olfactory tract, but it is now recognized that there is no medial olfactory tract. Although the primary olfactory cortex should not be considered equivalent to the primary visual, somatic sensory, or auditory areas in the cerebral neocortex, it is a major site for integration of sensory stimuli and appears to be involved in olfactory detection, discrimination, and even memory.

Throughout the olfactory cortex, fibers from the olfactory bulb end in the superficial lamina of layer I (layer Ia). There is also a complex set of associational connections that originate in many of the olfactory cortical areas and terminate in the deeper lamina of layer I (layer Ib) and layer III. Projections from the several olfactory areas convey information to many other parts of the brain, including the hypothalamus, amygdala, hippocampal formation, frontal and insular cortex, mediodorsal thalamus, and ventral striatum. These not only support olfactory sensation but also allow olfactory stimuli to influence both conscious behavior and subconscious visceral function.

Projection of Olfactory Bulb and Accessory Olfactory Bulb

The relay cells in the olfactory bulb and accessory bulb (the mitral and tufted cells) send their axons into the lateral olfactory tract, which runs caudally through the lateral part of the olfactory peduncle and the rostral part of the olfactory cortex. Axon collaterals branch off the larger axons in the tract and extend medially, laterally, caudally, and deep to the olfactory tract itself, to synapse in all the olfactory cortical areas. The thickness of layer Ia and the density of termination are greatest in the regions nearest to the lateral olfactory tract (for example in the medial edge of the piriform cortex immediately deep to the tract), and become progressively less in areas of cortex more distant from the tract.

The smaller tufted cells have less extensive axonal ramifications and reach only the anterior half to two-thirds of the olfactory cortex, whereas the larger mitral cells send axons to all parts of the cortex. Each parent axon in the lateral olfactory tract appears to give off a number of collateral branches, each of which makes numerous synapses as it runs through layer Ia of the cortex. Therefore, each mitral or tufted cell forms synapses in a wide region of the olfactory cortex.

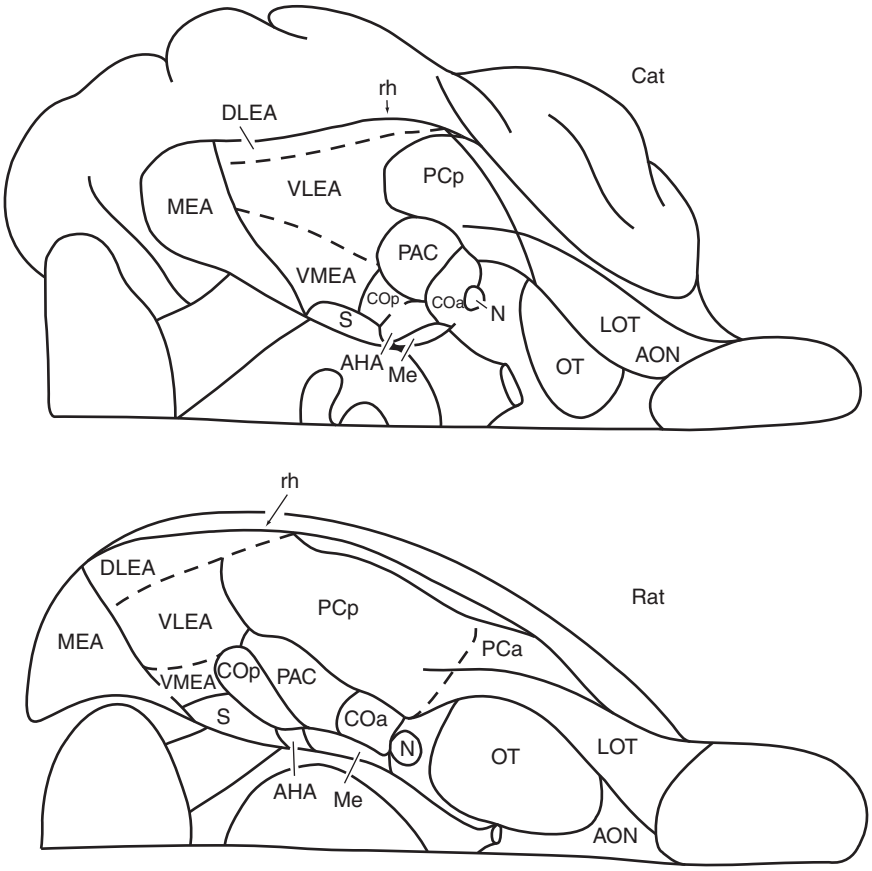

Figure 1 Drawings of the ventral surface of rat and cat brains, showing the several olfactory cortical areas that make up the primary olfactory cortex. AHA, amygdalo-hippocampal area; AON, anterior olfactory nucleus; COa, anterior cortical amygdaloid nucleus; COp, posterior cortical amygdaloid nucleus; DLEA, VLEA, and VMEA, dorsolateral, ventrolateral, and ventromedial parts of lateral entorhinal cortex; LOT, lateral olfactory tract; Me, medial amygdaloid nucleus; MEA, medial entorhinal cortex; N, nucleus of lateral olfactory tract; OT, olfactory tubercle; PAC, periamygdaloid cortex; PCa and PCp, anterior and posterior parts of piriform cortex; rh, rhinal sulcus; S, subiculum. From Krettek JE and Price JL (1978) *Journal of Comparative Neurology* 172: 723–752.

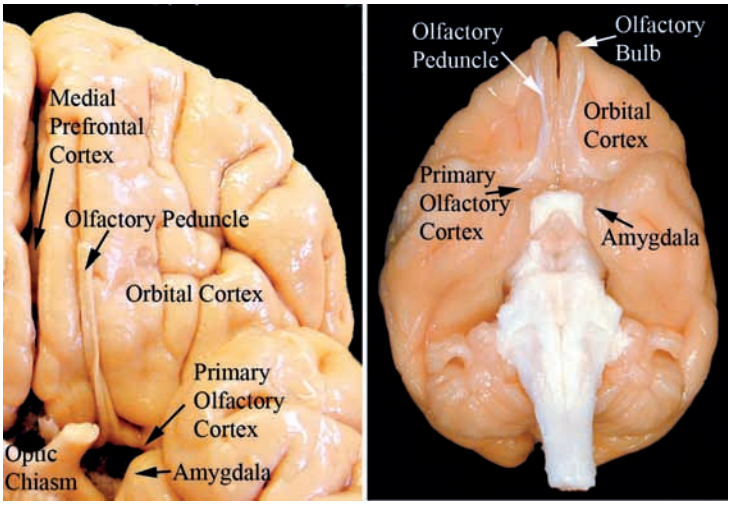

Figure 2 Photographs of the ventral surface of monkey (*M. fascicularis*; right) and human brains, showing the position of the primary olfactory and orbital cortex, as well as the olfactory bulb and peduncle and the amygdala.

Primary Olfactory Cortical Areas: Structure

At least seven distinct areas can be recognized that receive a monosynaptic input from the olfactory bulb and can therefore be considered olfactory cortical areas. All these areas are on the surface of the brain and have a generally cortical structure, in that they are composed primarily of pyramidal-type cells arranged in layers, with dendrites arrayed in a more-or-less radial manner. The cortical lamination is considerably simpler than in the neocortex, such that most of the areas have only two or three layers. There is considerable variation among different areas, however, and some have a more elaborate 'cortical' structure than others.

Anterior Olfactory Nucleus

The anterior olfactory nucleus (AON) is the most rostral of the olfactory cortical areas. In rats and most other small mammals, it consists of a ring of cells situated around the olfactory peduncle, surrounding the anterior limb of the anterior commissure. It is usually subdivided into dorsal, lateral, ventral, medial, and external parts. In primates, including humans, the olfactory bulb is pulled rostrally, and the olfactory peduncle is a long, slender stalk. A 'bulbar' part of the AON is located in the deep part of the olfactory bulb, while medial and lateral portions of the AON are near the point where the olfactory tract fuses to the frontal lobe. As in other olfactory cortical areas, axons of the mitral and tufted cells synapse on the distal segments of dendrites that extend into the plexiform layer around the neuron clusters. More-caudal subdivisions of the AO are situated on either side of the lateral olfactory tract where it joins the orbital cortex, at the posterior end of the olfactory peduncle.

Piriform Cortex

Caudal to the AON, the lateral olfactory tract runs along the junction of the olfactory tubercle, medially, and the piriform cortex, dorsal and laterally (Figure 1). The piriform cortex is the largest and most distinctive olfactory cortical area (Figure 3). Functional and behavioral studies suggest that it plays a central role in the detection and discrimination of olfactory sensory stimuli. In rodents and other small mammals, the piriform cortex occupies a large part of the ventral surface of the cerebral hemisphere, ventromedial to the rhinal sulcus. It forms a much smaller part of the cortical surface in monkeys and humans, on the caudolateral part of the ventral frontal lobe and the rostral part of the medial temporal lobe. The piriform cortex is characterized by a densely packed layer II composed of moderately large pyramidal cell bodies, and a less dense layer III of slightly larger pyramidal cells and other neurons. Layer III is well developed only in the caudal part of the cortex. Deep to the cortex, a ventral continuation of the claustrum forms the endopiriform nucleus. There is little data on this nucleus in primates, but in rodents it appears to represent a deep component of the overlying piriform cortex.

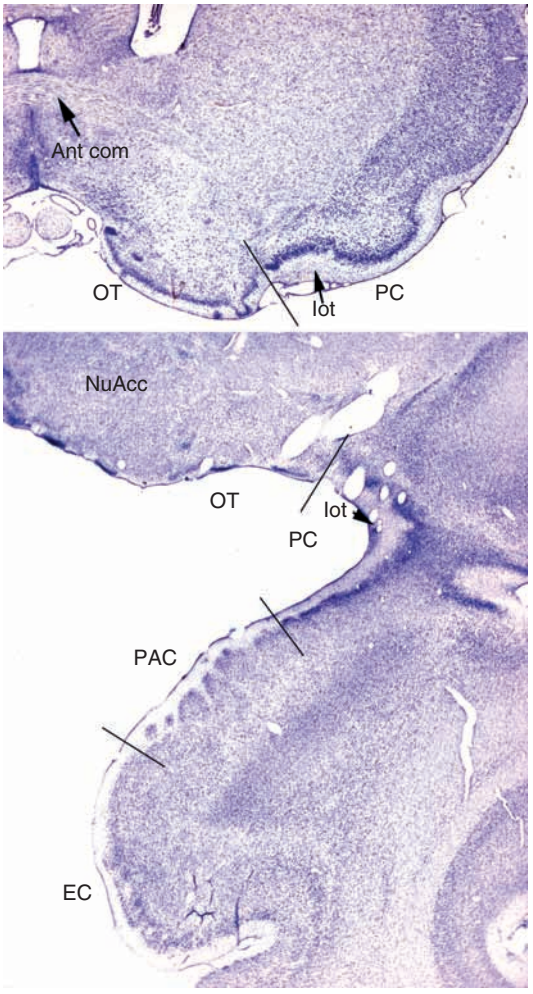

Figure 3 Photomicrographs of sections of rat (above) and monkey (below) brains, taken through the piriform cortex (PC), olfactory tubercle (OT), and (in monkey) the periamygdaloid (PAC) and entorhinal cortex (EC). Ant Com, anterior commissure; NuAcc, nucleus accumbens; lot, lateral olfactory tract.

As discussed in the section titled 'Intracortical associational fiber system,' the piriform cortex gives rise to a substantial associational fiber system that extends throughout all parts of the primary olfactory cortex. The piriform cortex also provides substantial projections to the ventral striatum and the agranular insula/posterior orbital cortex. There are further outputs through the deep cells of the endopirifom nucleus to the mediodorsal thalamic nucleus and the lateral hypothalamus.

Olfactory Tubercle

The olfactory tubercle is a prominent structure in rodents and carnivores, with a well-developed laminar structure similar to that in other olfactory cortical areas (Figure 3). It is much less distinct in primates but still has a laminar arrangement of cell bodies and afferent fibers. In all mammals, collaterals of axons in the lateral olfactory tract leave the medial side of the tract and run through the superficial part of layer I of the olfactory tubercle to contact apical dendrites of cells with somata in layers II and III. Although this laminar structure is clearly cortical, other features of the olfactory tubercle resemble the underlying corpus striatum, and it has often been included with the nucleus accumbens in the ventral striatum. For example, cells of the tubercle resemble the medium spiny cells of the striatum, and it has a high concentration of acetylcholinesterase, as well as a prominent dopaminergic input from the ventral midbrain. Its major output is to the ventral pallidum, which is similar to and dorsally continuous with the globus pallidus.

Clusters of small neurons called the islands of Calleja are inserted into layers II and III of the olfactory tubercle. Ventral pallidal elements are distributed somewhat loosely within the deep part of the tubercle rostrally, with more distinct components of the ventral pallidum found farther caudally and dorsally. The nucleus accumbens extends to the ventral surface in both monkeys and humans and occupies much of the area referred to as the olfactory tubercle in gross anatomical terms.

Anterior Cortical Nucleus of the Amygdala and Periamygdaloid Cortex

In the caudal part of the olfactory cortex, the axons from the olfactory bulb extend into several small areas on the medial surface of the amygdala. The anterior cortical amygdaloid nucleus is characterized by a relatively loosely packed layer II and even more diffuse layer III. The periamygdaloid cortex is a larger area, which is heterogeneous in monkeys and humans. There is a relatively thin zone of termination of olfactory bulb axons (layer Ia) in the periamygdaloid cortex, and the most caudal part does not receive any olfactory bulb input in primates.

Entorhinal Cortex

The entorhinal cortex has a more elaborate laminar structure than other parts of the olfactory cortex have, with at least five layers. It is notable because it provides the major input to the hippocampus. In rodents and carnivores, it is common to divide the entorhinal cortex into a relatively large lateral part, which receives fibers from the olfactory bulb, and a medial part that does not. In contrast, in primates the olfactory bulb input is limited to a small rostral olfactory zone that makes up only 10% or less of the entorhinal cortex as a whole. Even in the rostral part of the monkey entorhinal cortex, the olfactory input is relatively slight.

Accessory Olfactory Cortical Areas

In addition to the olfactory bulb, the accessory olfactory bulb provides a chemosensory input to specific areas of the amygdala, the medial and posterior cortical amygdaloid nuclei. This projection is almost entirely complementary to the olfactory bulb projection. The medial nucleus is situated immediately adjacent to the anterior cortical amygdaloid nucleus, while the posterior cortical nucleus is farther caudal. Both of these have important projections to the anterior and medial hypothalamus. In rodents, these connections have been shown to be involved in mating behavior and a variety of related functions.

Intracortical Associational Fiber System

There is an extensive system of intracortical connections within the primary olfactory cortex, which seems to interact with the input from the olfactory bulb to create patterns of activity across the cortex. It is possible that the olfactory cortex functions as a correlative system, possibly comparable to higher-order association cortex in other sensory systems.

The piriform cortex provides the greatest number of association fibers, but the fibers originate in all the olfactory cortical areas except the olfactory tubercle. The fibers are distributed throughout the olfactory cortex, and individual neurons in the piriform cortex have axons that extend into all or most of these areas. The association fibers terminate primarily in the deep part of layer I (layer Ib) and in layer III, complementary to the pattern of termination of the fibers from the olfactory bulb in layer Ia. In the tangential dimension of the cortex, there is a broad and relatively complex organization within the association fibers, which has been best defined in rats. Cortical regions near the lateral olfactory tract project to other parts of the olfactory cortex near the tract, whereas regions distant to the tract tend to project to regions at the edges of the olfactory cortex.

The AON and the anterior part of the piriform cortex also project through the anterior commissure to the contralateral olfactory cortex. The laminar and areal pattern of termination in the contralateral cortex are approximately the same as on the ipsilateral side.

Presumably, the association system interacts with the sensory activity from the olfactory bulb to support olfactory discrimination. In contrast to the olfactory bulb, neither of these fiber systems has the sort of detailed topographic organization that would be expected if odors were represented in a spatial code. It is possible that cells related to different odors are dispersed in many parts of the cortex, and appropriate bulbar and association fibers synapse on them in a selective manner. It may be more likely, however, that olfactory discrimination depends on a system of spatiotemporal patterning across the cortex, in which adjustments in synaptic strength are built up from sensory experience.

Olfactory Projections beyond the Primary Olfactory Cortex

Olfactory information is transmitted from the primary olfactory cortex to several other parts of the forebrain, including the orbital cortex, amygdala, hippocampus, ventral striatum, hypothalamus, and mediodorsal thalamus (Figure 4). The connections have been best studied in rodents, but there is also some experimental data from monkeys, and functional imaging studies have recently provided data on humans.

Projections to the Amygdala and Hippocampus

Both the amygdala and hippocampus have often been considered to be closely associated with olfaction, especially in lower animals. These inputs arise primarily in the periamygdaloid cortex and the olfactory part of the entorhinal cortex, which project both to deep amygdaloid nuclei and to several parts of the hippocampus. In rats, there is an especially strong olfactory input to the hippocampus, because more than half of the entorhinal cortex receives olfactory inputs. In primates, amygdala and hippocampus have become dominated by other sensory inputs, especially vision, but they still have direct olfactory connections.

Projections to the Hypothalamus

In rats, both electrophysiological recordings and axonal tracer experiments indicate that there are olfactory inputs to several parts of the hypothalamus. The most direct projection arises from cells in the deepest layer of the piriform cortex and other olfactory cortical areas. Although the fibers run through the full rostrocaudal extent of the hypothalamus, in the medial forebrain bundle, they terminate predominantly in the caudal half of the lateral hypothalamic area. Axons from the anterior cortical nucleus and medial nucleus of the amygdala end in more rostral and medial parts of the hypothalamus.

The olfactory inputs to the hypothalamus have not been as well defined in monkeys. Experiments with retrograde tracer injections in the lateral hypothalamus, however, label neurons in most of the olfactory cortical areas, as well as in related orbital/insular cortical areas. Electrophysiological responses to olfactory stimuli have also been recorded in the lateral hypothalamic area in monkeys.

Olfactory Projections to the Striatum, Pallidum, and Thalamus

Along with the amygdala, the olfactory cortex projects to the ventral part of the striatum, including

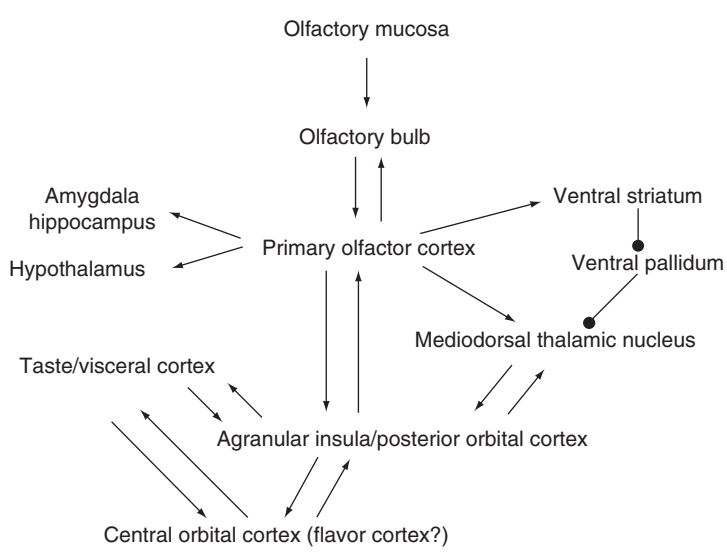

Figure 4 Diagram of the connections of the central olfactory system.

the accumbens nucleus and the olfactory tubercle. The olfactory projections are largely restricted to the olfactory tubercle and the caudal, ventrolateral part of the nucleus accumbens. These areas project out to the ventral pallidum, a continuation of the globus pallidus ventral to the anterior commissure. The major output of the ventral pallidum, in turn, is to the mediodorsal thalamic nucleus. As with the dorsal striatopallidothalamic system, the inputs to the ventral striatum are glutamatergic, but the striatopallidal and pallidothalamic projections are γ-aminobutyricacidergic (GABAergic).

As in other mammals, both electrophysiological recording and axonal tracing in monkeys also indicate excitatory olfactory input to the mediodorsal thalamic nucleus (MD) from the olfactory cortex. The neurons that project to MD are primarily located in the deep layers of the piriform cortex and other olfactory cortical areas. There are relatively small numbers of these neurons, and it appears that they may relay convergent activity from a relatively large portion of the olfactory cortex. Within MD in monkeys, fibers from small areas of the olfactory cortex or the amygdala end in small 'patches,' suggesting that inputs from restricted regions converge onto a few thalamic neurons. Similarly, olfactory-responsive units are restricted to a relatively small region of the medial, magnocellular part of MD. Compared with the hypothalamus, the olfactory-related units in MD appear to be relatively broadly tuned and respond to many different odors.

In addition to MD, anatomical and electrophysiological studies in both rats and monkeys indicate that there is an olfactory projection to the anteroventral portion of the submedial thalamic nucleus. This small nucleus is separated from MD only by the internal medullary lamina, and it may represent a portion of the same nucleus functionally. In rats, the projection to the submedial nucleus arises in a relatively restricted zone at the junction of the piriform cortex and olfactory tubercle.

The portion of medial MD that receives olfactory input is reciprocally connected with several areas in the posterior orbital cortex and rostral agranular insular cortex. The cortical connections of SM are to a more restricted area near the junction of the olfactory peduncle and the frontal lobe (especially area 13a). It might be presumed, therefore, that MD would relay olfactory information to the orbital cortex. As discussed in the section titled 'Olfactory projections to neocortical areas,' however, these same cortical areas receive more-numerous, monosynaptic projections directly from the primary olfactory cortex. These corticocortical projections appear to be better suited to relay detailed sensory information. It is likely, therefore, that the transthalamic projection does not represent a sensory relay as such.

Olfactory Projections to Neocortical Areas

In the 1970s and 1980s, Takagi and his colleagues reported that odorant stimuli or electrical stimulation of the olfactory bulb could evoke neuronal responses in the orbital cortex. Two areas, in the lateral and central parts of the posterior orbital cortex, were identified. In the lateral area, the units were relatively specific, responding to only one or two odorants, whereas in the more central area, units responded to several odorants. These studies suggested that the sensory information reached the orbital cortex from the primary olfactory cortex through the hypothalamus and thalamus. There is now clear anatomical evidence in monkeys and other animals, however, that the principal pathway is directly from the primary olfactory cortex to the orbital cortex.

Injections of anterograde axonal tracers in the piriform cortex of monkeys label axons in several areas of the agranular insula/posterior orbital cortex (areas Iam, Iapm, Iai, Ial, 13a, 13m, and 14c; **Figure 5**). All these cortical areas are agranular or dysgranular, and the areas nearest to the primary olfactory cortex have sometimes been referred to as periallocortex. The projection is different from more usual sensory inputs to cortex that are relayed through the thalamus, because the label is heaviest in layer I. Retrograde axonal tracer injections into the same agranular insula/posterior orbital areas label neurons in many parts of the primary olfactory cortex, including the AON, piriform cortex, anterior cortical amygdaloid nucleus, periamygdaloid cortex, and olfactory part of the entorhinal cortex. The labeled cells are located in both layer II and layer III. Electrical stimulation of the olfactory bulb also evokes unit and field potential responses in most of the areas where there is anatomical evidence of olfactory input.

The projections from the primary olfactory cortex are largely reciprocated by fibers from agranular insula/posterior orbital areas. Anterograde axonal tracer injections into several of the agranular insular areas label axons in the rostral parts of the primary olfactory cortex. These include the olfactory tubercle, which is the only olfactory cortical area that does not project to the orbital cortex.

The agranular insula/orbital cortical areas that receive olfactory input interact with other areas on the orbital surface through corticocortical connections to integrate olfactory sensory information with

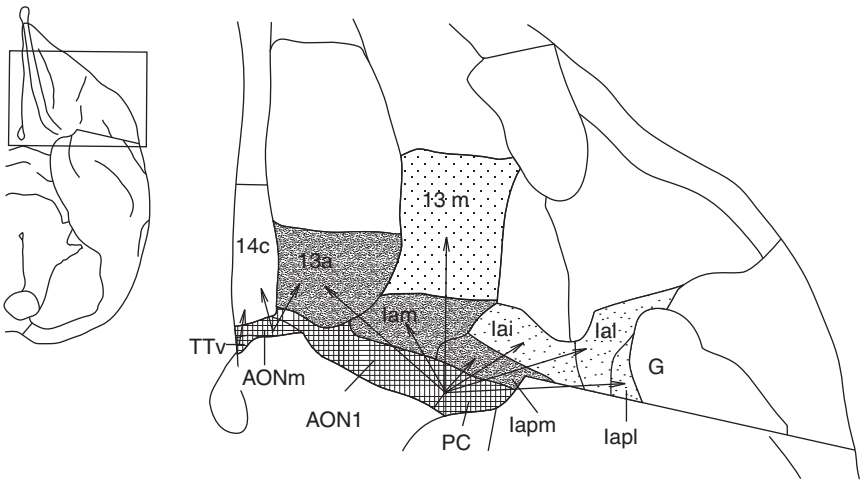

Figure 5 Diagrammatic summary of the projections from the primary olfactory cortical areas to the neocortical areas in the posterior orbital cortex in monkey. ai, agranular insula; AON, anterior olfactory nucleus.

other sensory modalities. Taste information reaches the orbital cortex through the thalamic gustatory relay and the primary gustatory cortex. The orbital cortex is the first place where olfactory information and taste information converge, so it presumably underlies the sensation of flavor, which depends on both of the primary modalities. In addition, there are corticocortical somatosensory and visual inputs to the orbital cortex from the parietal and inferior temporal cortex. The somatosensory inputs appear to be related to the hand and mouth. The corticocortical connections between orbital areas form an orbital network that appears to function in the integration and analysis of food-related sensory information.

Recordings in the orbital cortex in monkeys show neuronal responses to food and food-related stimuli, including visual stimuli that are associated with food. It is important to note that the responses reflect the affective or reward significance of the stimuli as well as their sensory properties. For example, the neuronal response to a particular visual stimulus (e.g., a triangle) will change markedly if the association of the stimulus with a food reward changes. In addition, if the animal is fed to satiety with a food stimulus, the neuronal response to that food will decrease.

Imaging of Olfactory Sensory Activity in Humans

Several studies have used functional imaging methods to identify the human cortical areas activated by olfactory stimuli. Although odorant-induced responses have been obtained in the region of primary cortex, other studies have failed to find substantial activation of primary olfactory areas. This discrepancy is probably due to technical issues that make it difficult to image activity in the primary olfactory cortex, including the rapidly adapting nature of olfactory sensation.

Olfactory-related activity has more consistently been detected in the central orbital cortex. Activation in the lateral orbital/anterior insular cortex has also been found following stimulation with aversive odors. The amygdala was also activated to the same odors.

The central orbital cortical region that responds to odorants also is activated by taste stimuli. This agrees well with the anatomical evidence in monkeys that corticocortical interconnections within the orbital cortex correlate olfactory and gustatory sensory information. Other sensory modalities, especially related to visual and somatic-sensory aspects of food stimuli, are also represented within the orbital cortex. The sensory responses in the orbital cortex also depend on hedonic properties of the stimuli. For example, functional magnetic resonance imaging responses to food-related olfactory stimuli are decreased after animals are fed to satiety with that food. Taken together, the evidence indicates that the orbital cortex is part of a system for analysis and assessment of food and food-related stimuli. Possibly because food is a primary reward, the system also may provide more-general analysis of reward. It should not be considered as simply secondary olfaction cortex or secondary taste cortex.

See also: Olfactory Cortex: Comparative Anatomy; Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Olfactory System Theory.

Further Reading

Carmichael ST, Clugnet M-F, and Price JL (1994) Central olfactory connections in the macaque monkey. *Journal of Comparative Neurology* 346: 403–434.

Haberly LB (2001) Parallel-distributed processing in olfactory cortex: New insights from morphological and physiological analysis of neuronal circuitry. *Chemical Senses* 26: 551–576.

Illig KR and Haberly LB (2003) Odor-evoked activity is spatially distributed in piriform cortex. *Journal of Comparative Neurology* 457: 361–373.

Price JL (2003) The olfactory system. In: Paxinos G (ed.) *The Human Nervous System,* 2nd edn., pp. 1198–1212. San Diego, CA: Elsevier Academic Press.

Rolls ET (2001) The rules of formation of the olfactory representations found in the orbitofrontal cortex olfactory areas in primates. *Chemical Senses* 26: 595–604.

Rolls ET (2005) Taste, olfactory, and food texture processing in the brain, and the control of food intake. *Physiology & Behavior* 85: 45–56.

Olfactory Insights from Transcriptional Profiling

T S McClintock and N Sammeta, University of Kentucky, Lexington, KY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The peripheral olfactory system of mammals has two unusual properties – the ability to detect external chemical signals (odors), of course, and also the capacity for adult neurogenesis. The odor signals animals detect are often linked to biological imperatives such as reproduction, other social interactions, and assessing the quality of the local environment (food, shelter, and predators). These are powerful selective forces driving the detection of external chemicals at high sensitivity, a problem solved by exposing a set of neurons to the external environment. The seemingly unavoidable consequence of damage to these exposed neurons appears to be why the olfactory epithelium retains substantial neurogenic capacity throughout adulthood, in contrast to better protected regions of the nervous system. Much effort investigating the olfactory system goes toward understanding these two properties, and a growing component of our advancing knowledge of them stems from work using expression profiling approaches.

The history of gene discovery approaches in the study of the olfactory system stretches back more than 25 years to the identification of olfactory marker protein (OMP), an enigmatic protein that has proved useful as a specific marker of mature olfactory sensory neurons. Since that time, molecular identification and analyses of individual genes have paid dividends for our understanding of the olfactory system. Prime examples are the discovery of the odorant receptors and consequent elucidation of their role in determining the odor quality map of the olfactory bulb glomeruli that must be the basis for odor discrimination by the CNS. Such studies of single genes and proteins are the foundation for recent functional genomics studies that continue to expand our knowledge of genes important to the olfactory system, and promise to build a system-level understanding of the molecular functions that give the olfactory system its capacity to detect odors and replace damaged neurons. The most compelling advances in olfactory genomics have arisen from studies that creatively combined expression profiling techniques with other approaches. After all, expression profiling techniques measure just one parameter, mRNA abundance – albeit with the advantage of simultaneously measuring (nearly) all mRNAs. Interpretation of expression profiling data has therefore depended critically on providing additional context to these data.

Functional Genomics of the Olfactory Epithelium

Olfactory Sensory Neurons

The task of detecting and encoding odor signals falls to the olfactory sensory neurons (OSNs) of the main olfactory epithelium and to the vomeronasal sensory neurons (VSNs) of the vomeronasal organ. These two neural epithelia are located in the nasal cavity and their neurons project axons ipsilaterally to the main and accessory olfactory bulbs of the CNS, respectively (Figure 1). While these epithelia contain several cell types (Figure 2), the vast majority of studies have understandably concentrated on the neurons, and therefore on the genes they express. As a group, OSNs and VSNs express an unusually large number of genes that have highly restricted expression patterns. This category of gene is dominated by the odorant receptors (ORs) and vomeronasal receptors (V1Rs and V2Rs), but several other genes also appear to be expressed solely or primarily in these sensory neurons. By definition, such genes encode proteins that are often directly involved in the detection and encoding of odor signals, but sometimes playing supporting roles in chemosensory neurons instead. Although the genes that encode the chemosensory receptors and their associated transduction mechanisms certainly differ between OSNs and VSNs, and while no quantitative measures have yet been made, OSNs and VSNs appear to overlap extensively in the remainder of their gene expression patterns.

The abundance of olfactory-enriched gene products has contributed to the successful use of single-cell polymerase chain reaction (PCR) as an expression profiling method to identify olfactory-enriched sequences. For example, the V1R family of vomeronasal receptors was discovered in this way. Similarly, single-cell PCR methods led to the discovery of a putative OR chaperone/trafficking complex, components (e.g., Rtp1 and Rtp2) of which are examples of olfactory-specific gene products that play supporting roles in the odor detection process. Furthermore, cells of the olfactory epithelium were the substrate for pioneering work that combined single-cell PCR amplification and microarray analysis. The advantage of knowing the cell expressing each mRNA meant that each mRNA could

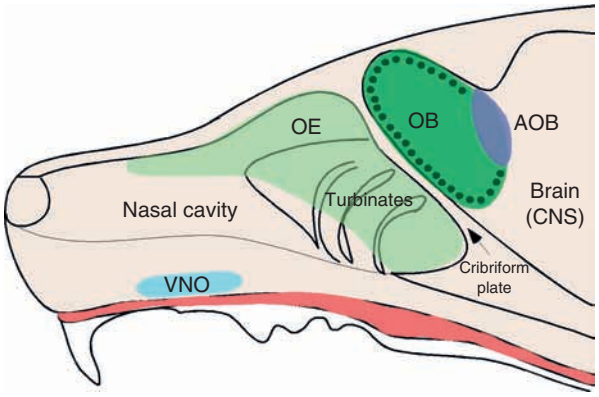

Figure 1 The location of the main components of the first two levels of the olfactory system, depicted in sagittal section of a rodent head. The main olfactory epithelium (OE; light green) resides in the nasal cavity, where it covers the turbinates laterally and septum medially. The olfactory sensory neurons in the olfactory epithelium project axons directly through the cribriform plate to the main olfactory bulb (OB; dark green) of the forebrain. Vomeronasal sensory neurons, which reside in the vomeronasal organ (VNO; light blue) at the base of the septum, send their axons to the accessory olfactory bulb (AOB; dark blue).

be interpreted in the context of adding one more item to the list of molecular functions present in that cell type.

Genes expressed in OSNs have also been identified by detecting mRNAs that decrease when mature OSNs are forced to apoptose, accomplished by severing OSN axons during removal of the olfactory bulbs (bulbectomy). This approach tends to detect only mRNAs enriched in mature OSNs, because most mRNAs shared with neighboring cells, which survive damage to the OSNs, show relatively small changes in abundance. More inclusive data that identified what appears to be the majority of mRNAs present in OSNs required the purification of a large pool of OSNs and the development of microarrays that assess nearly the entire genome. By comparing ratios of mRNA abundance in OSNs versus the pool of neighboring cells with cell type identification from *in situ* hybridization data, these analyses gave a probability of expression in OSNs for each mRNA and predicted that OSNs express more than 10 000 genes. Applying bioinformatics analyses to this list of genes predicts not only individual molecular functions present in OSNs, but also the biological processes that require them (Figure 2).

One such process is ciliogenesis, which happens as immature OSNs complete differentiation and become mature. The previously known molecular link between OSNs and sperm, revealed by evidence that a few OR genes mediate sperm chemotaxis, pales in comparison to the extensive molecular similarity between OSN cilia and sperm flagella. These similarities include components of the axoneme, the structural foundation of cilia and flagella, and the proteins of cilia support structures, such as basal bodies, ciliary rootlets, and trafficking compartments.

Another intriguing process enriched in OSNs is chromatin remodeling/assembly. Components of all of the major chromatin remodeling complexes are expressed in OSNs or OSN progenitor cells, including SWI/SNF complexes, Polycomb complexes, and NuRD complexes. The modification of chromatin is a mechanism of transcriptional control that acts at the level of regions of DNA as well as at the single-gene level. As such, it is critical to development and differentiation, consistent with evidence that many chromatin remodeling mRNAs are most abundant in OSN progenitors and immature OSNs. However, several mRNAs associated with chromatin remodeling (*Cbx4*, *Cbx8*, *H2afy*, *Tlk1*, and *Setd7*) are most abundant in mature OSNs, suggesting that epigenetic mechanisms also have a role in the fully differentiated phenotype.

Other surprises relate to the cell signaling capacities of OSNs. For example, the wealth of nonchemosensory receptor mRNAs found in OSNs included 16 growth factor receptors and at least 30 G-protein-coupled receptors, most not previously detected in OSNs. These include receptors for glutamate, acetylcholine, somatostatin, serotonin, Wnts, latrophilins, prostaglandins, epidermal growth factor, insulin-like growth factors, interleukins, and members of the transforming growth factor-β family of signals (especially bone morphogenetic proteins). OSNs also express 14 growth factor mRNAs and at least 80 mRNAs encoding proteins involved in cell adhesion. OSNs appear to pay attention to, and to send, many more signals than previously suspected.

The number of transcription factors expressed by OSNs is impressive. The data predict that OSNs express at least 268 mRNAs that are known or implicated by sequence similarity to be proximal regulators of transcription. Some illustrative examples are listed in Figure 2. The functions in OSNs of very few of these transcription factors are known, but in other tissues some are known to be involved in processes consistent with known OSN functions: stress responses, differentiation, ciliogenesis, and responses to Ca^{2+} signals or neural activity, for example. A likely harbinger of a trend was expression profiling of olfactory epithelia of mice lacking the transcription factor, Klf7. Klf7-deficient mice have reduced amounts of several synaptic mRNAs and the OSN-specific mRNAs, including OMP and adenylyl cyclase type 3, implicating Klf7 in the terminal events of OSN differentiation.

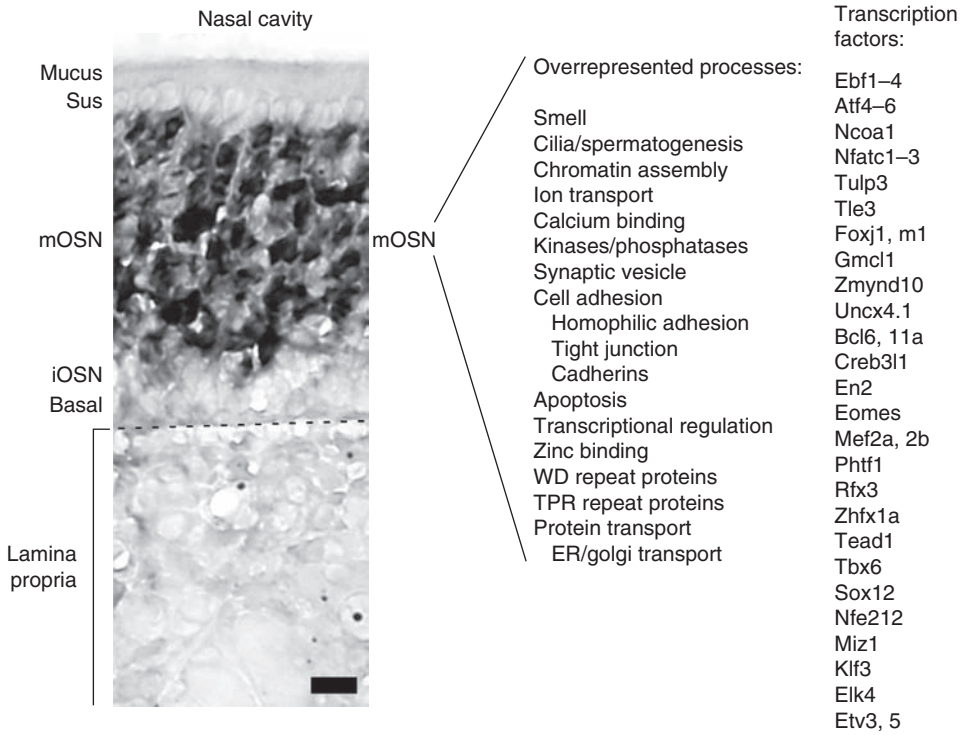

Figure 2 Mature olfactory sensory neurons (mOSNs) have been a major focus of expression profiling studies in the olfactory system. Their cell bodies occupy the middle of the pseudostratified olfactory epithelium, as shown in this *in situ* hybridization for olfactory marker protein, the mRNA for which is specific to mOSNs. Biological processes overly abundant in expression profiling data from OSNs are listed (overrepresented processes). Also shown is a partial list of the more than 250 transcription factors identified among the >10 000 genes expressed by OSNs. Dashed line, location of the basal lamina; Sus, sustentacular cell (a glial-like supporting cell); iOSN, immature OSN layer. Progenitor cells with the capacity to proliferate new OSNs reside in the basal (globose and horizontal) cell layer. Scale bar = 10 μm.

Odorant Receptors

At more than 1000 members, odorant receptors are encoded by the largest family of genes in mammalian genomes. ORs are also very diverse, with sequence similarities ranging down to about 20%. To cope with such size and diversity, a GeneChip microarray devoted to mouse ORs and V1Rs was developed. As expected, use of this GeneChip determined that at least 800 of the ORs are expressed in the olfactory epithelium. While a few dozen OR mRNAs were detected in other tissues (cerebellum, heart, liver, muscle, and testis), very few were more abundant in these tissues than in the olfactory epithelium, and many were several-fold more abundant in the olfactory epithelium. OR gene classification, though based largely on protein length, a few small conserved motifs, and a predicted seven-transmembrane domain topology, rather than on sequence similarity, therefore appears to be largely correct.

The OR GeneChip has also provided a broader view of the zonality of OR gene expression. Data from just a few dozen ORs indicate that ORs are expressed in restricted regions along the dorsoventral axis of the olfactory epithelium. Expression profiling of dorsal versus ventral olfactory epithelium supports and extends this conclusion to more than 600 ORs. Class I ORs, which have homology to fish ORs, were detected only in dorsal region samples, while the more recently evolved class II ORs were found in both dorsal and ventral regions. Given that ORs differ in their odorant agonists, these findings are consistent with long-standing evidence that regions of the olfactory epithelium are differentially sensitive to odorants. They are also consistent with suppositions about the evolutionary origins and expansion of the OR family from fish, with a single OR expression domain, to amphibians, with separate air and water OR compartments, to mammals, in which the entire olfactory epithelium is exposed to air but the ancestral expression pattern of two OR expression domains is at least partially retained and even divided into smaller zones in the case of the ventral domain.

Other Cell Types in the Olfactory Epithelium

While other cell types in the olfactory epithelium and in central olfactory pathways have not yet been investigated in detail by expression profiling methods, one of the by-products of the existing expression

profiling data has been the discovery of new tools that help discriminate cell types. These tools include *Reg3g* for respiratory epithelium, *Cbr2* and *Epas1* for sustentacular cells, *Ptgds* and *Igf2* for olfactory bulb leptomeninges, and *Ntf3* for a subset of olfactory bulb mitral cells. Another by-product comes from the comparison of the expression profile of the pool of immature OSNs, sustentacular cells, basal cells, respiratory epithelium, and lamina propria cells against the profile of purified mature OSNs. While more difficult to interpret because the identities of the source cells are ambiguous, more than 13 000 mRNAs were detected in this sample. Expression profiling data on adult olfactory neurogenesis provide insight into the function of some of these gene products in the olfactory epithelium (see later).

Novel Genes

Not surprisingly, a majority of the more than 14 000 mRNAs detected in the olfactory epithelium encode proteins with no known function. The genes are difficult to investigate, but they also represent opportunities to make truly novel contributions. The available evidence does provide some guidance for many of these mRNAs. Some, like *Tmem16b*, are at most weakly expressed in other tissues, thereby predicting olfactory-specific functions. Others, like *Susd4* and *Tspyl4*, are abundant only in neural tissues, predicting a role important to neurons in general. Still others, like *1110012D08Rik*, appear to be ubiquitous and therefore part of a function required by all cells.

Functional Genomics of Adult Olfactory Neurogenesis

Olfactory Sensory Neuron Response to Axotomy

The ability of the olfactory epithelium to replace damaged OSNs even in adulthood and yet not lose the capacity to discriminate odors is a remarkable example of the plasticity of the nervous system. It is also an example of the potential of neural replacement therapies for the treatment of neurological disorders. The cellular events that support adult neurogenesis in the olfactory epithelium are well understood (Figure 3). Damaged OSNs apoptose, then infiltrating immune cells (predominantly macrophages) clear the resulting cellular debris and also secrete signals that help stimulate basal progenitor cell proliferation, leading to the production of new immature OSNs that differentiate into mature OSNs. As a model for adult neurogenesis, OSN replacement has the advantage that it can be massively and synchronously induced by selectively damaging the OSNs in any of several ways, but most easily by bulbectomy, the removal of the olfactory bulbs and consequent severing of the OSN axons.

While a few important molecular markers of this process were known, recent expression profiling results have identified hundreds of additional mRNAs which change in abundance during the replacement process. The most striking finding is the degree of agreement between molecular and cellular events. The vast majority of affected mRNAs changed abundance along temporal profiles that matched the profiles of the critical cellular events of OSN apoptosis, immune cell infiltration, basal cell proliferation, and OSN differentiation. The identities of the mRNAs in each of these temporal profiles corresponded to functions necessary for the matching cellular events.

Within the OSNs, the expression profiling data reveal a common apoptotic theme. OSNs shift into a proapoptotic state reflected by a decrease in antiapoptotic mRNAs and an increase in proapoptotic mRNAs. The major antiapoptotic genes affected are *Bcl2* and *Birc6*. The major proapoptotic genes affected are *Bcl2l11* (*Bim*), *Bax*, *Bid*, *Casp3*, *Casp6*, *Casp8*, *Gadd45a*, and *Gadd45b*. Exactly how these mRNA abundance changes are triggered by signals, such as activated caspases, traveling back from the severed axons of the OSNs is not yet clear.

Immune Cell Infiltration

Also anticipated among the events evident during OSN replacement was the infiltration of immune cells, especially macrophages, that follows the apoptosis of OSNs. Increases in chemokines such as chemokine (C-C motif) ligand 2 (CCL2) and interleukin-1β (IL-1β) are consistent with the documented recruitment of macrophages. Molecular evidence that dendritic cells may also be recruited has not yet been substantiated. Later increases in mRNAs associated with macrophage maturation, such as *Cd1d1*, *Cd83*, *Lsp1*, *Pparg*, and *Csf1r*, and antigen presentation mRNAs suggest that infiltrating macrophages reduce their rate of migration and mature while in the olfactory epithelium. These events are important not only because macrophages phagocytose debris from apoptotic cells, but also because they secrete signals that enhance the proliferation of basal cells. A clever approach to assessing this involved profiling mRNA abundance in the olfactory epithelium of mice lacking scavenger receptor A, important for macrophage responses to apoptosing cells. The data showed evidence of molecular changes after bulbectomy that are consistent with cellular defects in these knockout mice – fewer infiltrating macrophages, reduced basal cell proliferation, and accumulation of apoptotic debris.

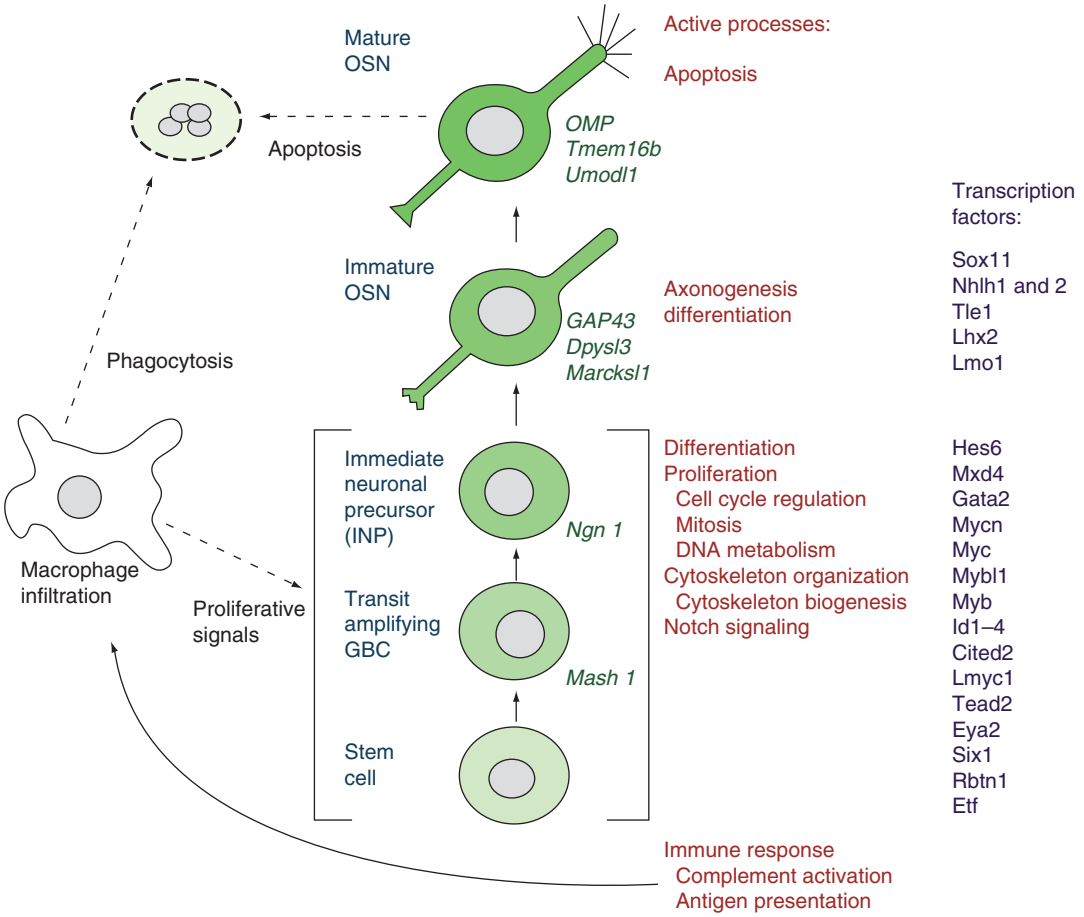

Figure 3 Expression profiling has had a significant impact on our understanding of the control of adult neurogenesis in the olfactory epithelium. The olfactory sensory neuron (OSN) cell lineage begins with an as yet unidentified stem cell (hypothesized to be a subtype of basal cell) and progresses through at least two types of proliferative globose basal cells (GBCs) – a transit-amplifying GBC and an immediate neuronal precursor GBC – before differentiation generates immature OSNs and subsequently mature OSNs. With the exception of the stem cell, known differentially expressed genes (green italics) allow identification of the cell types. The other important cellular event is immune cell (primarily macrophages) infiltration; the infiltrating immune cells aid neurogenesis by phagocytosing debris and secreting proliferative signals. Expression profiling studies have identified processes that are active during adult neurogenesis (active processes; red text). These are depicted next to their presumed locations in the cell lineage. Similarly, expression profiling predicts transcription factors responsible for the molecular changes that control adult neurogenesis. A subset of these transcription factors is listed (purple text).

Proliferation and Differentiation

The synchronization of OSN replacement by bulbectomy leads to a robust increase in cell proliferation that peaks about 6 days later in mice. Expression profiling data from this time period reveal increases in many positive regulators of the cell cycle and other markers of cell proliferation. Of paramount interest are the proteins that control this proliferative process. Proteins known to stimulate proliferation or inhibit lineage commitment detected as increased mRNAs include transcription factors Gata2, Mycn (Nmyc1), Myc (c-Myc), Mybl1, Myb, Id1, and Id3 (Figure 3). These transcription factors are therefore candidates to control the proliferative events that underlie OSN replacement. Also important are transcription factors that promote exit from the cell cycle, determine cell fate, and contribute to the subsequent differentiation of OSNs. Expression profiling identified Nhlh1, Neurog1 (Ngn1), Lmo1, Sox11, Hes6, and Mxd4 as candidates to control these events. Associated with differentiation of neurons should be the extension of neurites, and increases in numerous mRNAs encoding proteins involved in axon outgrowth were observed: *Gap43*, *Basp1*, *Ablim1*, *Marcksl1*, *Stmn1*, *Chl1*, *Dbn1*, *Dpysl3*, *Syn1*, *p116Rip*, *Celsr1*, *Fyn*, *Dlx5*, *Myh10*, *Crym*, *Rasa3*, and *Ppp2cb*. These mRNAs are very likely to be enriched in immature OSNs and relatively less abundant in mature OSNs. *In situ* hybridization data have confirmed this prediction for several of these mRNAs.

Adult and Embryonic Neurogenesis Are Transcriptionally Similar

More broadly, expression profiling of induced OSN replacement speaks to the question of whether adult neurogenesis is substantially different from embryonic neurogenesis. When compared with expression profiling of mouse brain embryonic development, >300 mRNAs are shared – more than half of the mRNAs that increased during the period of basal cell proliferation and OSN differentiation. Comparison of embryonic and adult neurogenesis within the olfactory epithelium would probably show even more similarity. While some differences between adult and embryonic neurogenesis probably exist, the available evidence suggests that many molecular elements are shared.

Functional Genomics of the Olfactory Bulb and Behavior

Expression profiling methods have as yet had little impact on our understanding of central olfactory pathways. Several interesting opportunities exist, however. For example, the olfactory central nervous system (CNS) has functional and cellular plasticity that should be intriguing as a substrate for expression profiling. Associative learning involving odors is an example of robust functional plasticity. Cellular plasticity happens continuously in the olfactory bulb, where interneurons are replaced from progenitors that arise in the subventricular zone and migrate to the bulb in the rostral migratory stream. The one question that has been probed by microarray analysis is the regionalization of the olfactory bulb. The olfactory bulb has an overall regional organization that may help organize the odor quality map of OR-specific glomeruli, a map that is presumably critical for odor discrimination. At least nine mRNAs are differentially distributed across the olfactory bulb. Several of them (*Jag1*, *Cadm1* (*RA175C*), *Ntf3*, *Vtn*, *Col5a2*, and *Cdh11*) have functional links to synapse formation and axon connectivity, and therefore might contribute to the synaptic organization of the bulb or its cortical targets.

Expression profiling should also be of value in linking genes and behaviors mediated by olfaction. Heritable mutations that alter olfactory-guided behaviors are most likely of two types: changes in proteins that are central to the detection of odors, such as ORs, or changes in proteins that cause pleiotropic effects that are not so obviously related to olfaction. Pleiotropic effects resulting from single gene mutations are likely to involve changes in the abundance of gene products and are therefore suitable targets for expression profiling. One such situation has been investigated by profiling five co-isogenic smell-blind mutant *Drosophila* lines known to have epistatic relationships. Significant overlap was detected in mRNA abundance changes among these epistatic fly lines, with more than 150 altered mRNAs shared by at least two lines. The inverse of this, mRNAs specific to each line, is also informative. As should happen with genomic level analyses, the identities of the affected mRNAs suggest hypotheses about affected processes. For example, mutant line *smi21F* was associated with chemical recognition mRNAs, *smi28F* was associated with ligand carrier protein and biotransformation enzyme mRNAs, and *smi26D* was associated with synaptic function mRNAs. Ways in which altered levels of these gene products could alter odor-stimulated behaviors are easy to imagine.

Expectations for the Future

Expression profiling has much more to reveal about the olfactory system. Most of what has been learned to date is descriptive, predictive, and correlative, but the breadth and depth of these data make them valuable resources. We can expect the trend to characterize the capacities of individual cell types to expand to additional cell types in the olfactory epithelium and probably even to the neurons of the olfactory bulb and higher olfactory centers. We should expect to gain insight into physiological and developmental details by more expression profiling of gene knockout mice or mice with more sophisticated mutations that generate proteins with altered function. In addition, we can expect expression profiling to shed light on the importance of alternative splicing within the olfactory system, a topic as yet ignored. We can also expect to see more work on model systems in which gene expression patterns can be linked to olfactory behaviors.

See also: Olfactory Receptor Genes: Human Loss During Evolution; Olfactory Neuron Patterning and Specification; Olfactory Coding; Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Plasticity and Activity-Dependent Regulation of Gene Expression; Transcriptional Networks and the Spinal Cord.

Further Reading

Anholt RR, Dilda CL, Chang S, et al. (2003) The genetic architecture of odor-guided behavior in *Drosophila*: Epistasis and the transcriptome. *Nature Genetics* 35: 180–184.

Dulac C and Axel R (1995) A novel family of genes encoding putative pheromone receptors in mammals. *Cell* 83: 195–206.

Getchell ML, Li H, Vaishnav RA, et al. (2006) Temporal gene expression profiles of target-ablated olfactory epithelium in mice with disrupted expression of scavenger receptor A: Impact on macrophages. *Physiological Genomics* 27: 245–263.

Getchell TV, Liu H, Vaishnav RA, et al. (2005) Temporal profiling of gene expression during neurogenesis and remodeling in the olfactory epithelium at short intervals after target ablation. *Journal of Neuroscience Research* 4: 74–82.

Goldstein BJ, Kulaga HM, and Reed RR (2005) Cloning and characterization of SLP3: A novel member of the stomatin family expressed by olfactory receptor neurons. *Journal of the Association for Research in Otolaryngology* 80: 309–329.

Kajimura D, Dragomir C, Ramirez F, et al. (2007) Identification of genes regulated by transcription factor KLF7 in differentiating olfactory sensory neurons. *Gene* 388: 34–42.

Lin DM, Yang YH, Scolnick JA, et al. (2004) Spatial patterns of gene expression in the olfactory bulb. *Proceedings of the National Academy of Sciences of the United States of America* 101: 12718–12723.

Saito H, Kubota M, Roberts RW, et al. (2004) RTP family members induce functional expression of mammalian odorant receptors. *Cell* 119: 679–691.

Sammeta N, Yu TT, Bose SC, et al. (2007) Mouse olfactory sensory neurons express 10,000 genes. *Journal of Comparative Neurology* 502: 1138–1156.

Shetty RS, Bose SC, Nickell MD, et al. (2005) Transcriptional changes during neuronal death and replacement in the olfactory epithelium. *Molecular and Cellular Neuroscience* 30: 90–107.

Tietjen I, Rihel JM, Cao Y, et al. (2003) Single-cell transcriptional analysis of neuronal progenitors. *Neuron* 38: 161–175.

Yu TT, McIntyre JC, Bose SC, et al. (2005) Differentially expressed transcripts from phenotypically identified olfactory sensory neurons. *Journal of Comparative Neurology* 483: 251–262.

Zhang X, Rogers M, Tian H, et al. (2004) High-throughput microarray detection of olfactory receptor gene expression in the mouse. *Proceedings of the National Academy of Sciences of the United States of America* 101: 14168–14173.

Relevant Websites

http://firestein.bio.columbia.edu – Stuart Firestein (Columbia University, New York, NY) neurobiology lab odorant receptor database.

http://app.mc.uky.edu/ – University of Kentucky College of Medicine: McClintock lab data downloads (genes expressed in the olfactory epithelium).

http://senselab.med.yale.edu – Yale University Olfactory Receptor Database.

Olfactory Neuron Patterning and Specification

C L Beites, S Kawauchi, and A L Calof, University of California at Irvine, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview of Olfactory Epithelium and Vomeronasal Organ Structure and Cell Types

In terrestrial vertebrates, the primary relay neurons of the olfactory system are contained within two specialized sensory epithelia in the head – the main olfactory epithelium (OE), which lines the nasal cavity and is involved in detection of odors, and the vomeronasal sensory epithelium, which lines the vomeronasal organ (VNO; a bean-shaped structure located in the anterior–ventral region of the nasal septum) and mediates detection of nonvolatile pheromones (Figure 1(a)). Cell bodies of olfactory receptor neurons (ORNs) lie within the epithelia of the OE and VNO. ORNs are bipolar primary sensory neurons: they have apical dendrites, with cilia- or microvilli-covered surfaces that extend into the nasal or VNO cavities, and they extend axons that synapse upon neurons within the central nervous system. In the case of the OE, the axons of ORNs form the first cranial nerve and synapse upon cells of the main olfactory bulb (MOB) of the forebrain. The sensory neurons of the VNO extend axons to form the vomeronasal nerve, which synapses upon the accessory olfactory bulb (AOB), a prominent structure in rodents that lies immediately adjacent to the MOB.

Because rodents have a highly developed sense of smell and have been used for many years as model organisms in the study of olfaction, information provided in this article concentrates primarily on studies of OE and VNO in rodents, particularly rats and mice. Studies of primary olfactory pathway development in mouse mutants have shown that the process of neurogenesis in both the OE and the VNO is regulated to a large extent by the same genes, which are expressed by cells at distinct stages in the neuronal lineages of the two epithelia (Figure 1(b)). The OE and VNO are also similar in that, once their structures are established during the final third of gestation, the neural cell types within them are arranged in roughly comparable laminar organizations (Figure 1(c)). Thus, in both epithelia, neural stem cells and committed neuronal progenitors are located in the basal compartment (near the basement membrane, which forms the boundary between sensory epithelium and underlying connective tissue stroma – i.e., the lamina propria). Committed neuronal progenitors give rise to terminally differentiated sensory neurons (ORNs), which comprise the majority of cells within both epithelia. Another major cell type within the two sensory epithelia is the supporting, or sustentacular, cell, best characterized in the main OE. The cell bodies of supporting cells lie in a distinct layer in the apical compartment of the OE and their processes extend to the basal lamina; they appear to be analogous to glial cells of the brain, and studies of mouse mutants suggest that ORNs and supporting cells are both derived from a common neural stem cell in the OE.

Another cell type found associated with the OE is the olfactory ensheathing cell (OEC). OECs, which encircle bundles of ORN axons in the stroma underlying OE, possess characteristics of both Schwann cells and astrocytes. Interestingly, transplanted OECs appear to promote recovery in a variety of nerve lesion models, and it has been proposed that their presence is responsible for the ability of lesioned ORN axons to regenerate and reinnervate the central nervous system. Some experimental embryology studies suggest that OECs may originate from the olfactory placode, and tissue culture studies have also reported that cells resembling OECs arise in cultures made from OE proper. However, no definitive lineage-tracing study has been performed to demonstrate unequivocally that OECs arise from the olfactory placode or OE proper, and at present their origin remains uncertain. Thus, although the provocative possibility exists that OECs are, like ORNs, products of a multipotential OE stem cell, this hypothesis remains to be proved, and OECs are not discussed further in this article.

Neuronal Cell Types of the OE and VNO

Newly generated neurons in the OE and VNO are products of differentiation pathways known to contain at least three distinct proliferating cell types. These cell types have been most extensively characterized in the main OE (Figure 1(b)): (1) neural stem cells divide to give rise to a population of (2) committed neuronal progenitor cells that express the proneural gene *mammalian achaete scute homolog 1* (*Mash1*; also known as *Ascl1*), and *Mash1*-expressing progenitor cells are committed to a neuronal fate and undergo amplifying divisions to rise to (3) immediate neuronal precursors (INPs), which express a different proneural gene, *neurogenin1* (*Ngn1*; also known as *Nurog1*). Studies using retroviral lineage tracing and

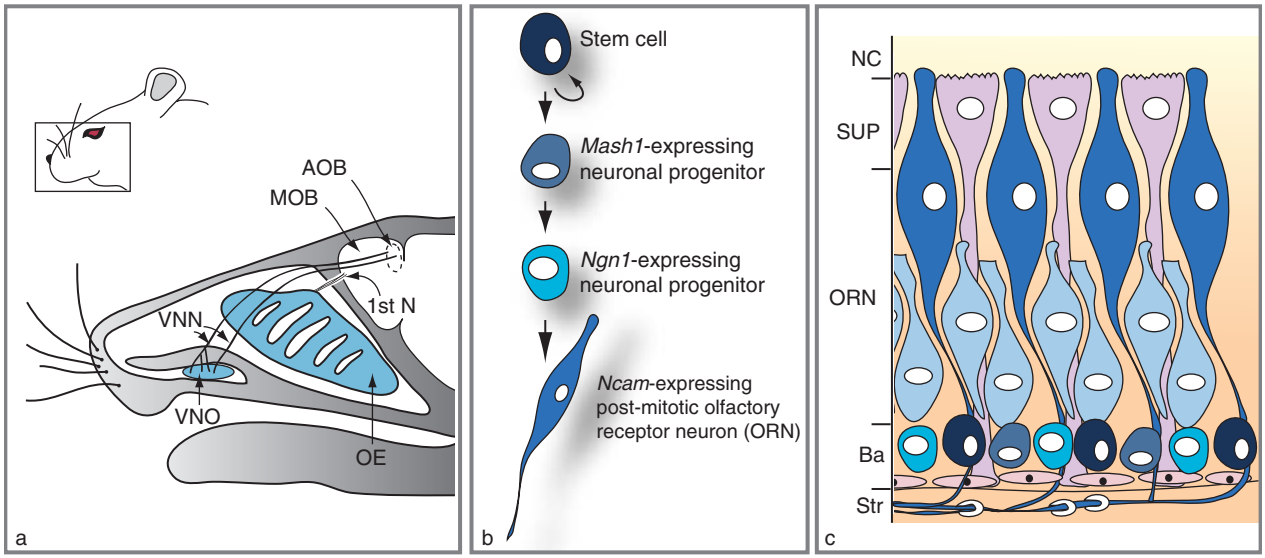

Figure 1 Location of main olfactory epithelium (OE) and vomeronasal organ (VNO) in mouse head, and lineage and distribution of different cell types within the OE. (a) Sagittal section through an adult head reveals the position of the VNO relative to the OE. While the vomeronasal nerves (VNN) target the accessory olfactory bulb (AOB), axons of olfactory receptor neurons (first cranial nerve; 1st N) target the main olfactory bulb (MOB). (b) Scheme of the neuronal differentiation pathway in the OE. (c) Histological arrangement of cells in mature OE. Neuronal stem cells (black) located at the basal layer (Ba) adjacent to the stroma (Str) give rise to transit-amplifying progenitors expressing the *Mash1* gene (round gray-blue cells) followed by *Ngn1*-expressing precursors (turquoise cells). The immature ORNs (light blue elongated cells) arise from the *Ngn1*-positive cells and mature into *Ncam*-expressing neurons (dark blue elongated cells). Supporting cells (SUP) lie in a single layer on the apical surface of the OE, at the nasal cavity (NC), while olfactory ensheathing cells wrap around nerve bundles in the underlying stroma. The VNO (not shown) has a similar histological arrangement, with basal cells at the boundary between epithelium and mesenchyme and neurons populating the bulk of the epithelium. Adapted from Kawauchi S, Beites CL, Crocker CE, et al. (2004) Molecular signals regulating proliferation of stem and progenitor cells in mouse olfactory epithelium. *Developmental Neuroscience* 26: 166–180, with permission. Copyright 2004 by S. Karger AG, Basel, Switzerland.

molecular markers for different cell types indicate that, in the established OE (see later), INPs, *Mash1*-expressing neuronal progenitors, and neural stem cells are all found in the basal compartment of the epithelium. INPs are, like *Mash1*-expressing progenitors, transit-amplifying cells, but INPs are committed ultimately to generating daughter cells that undergo terminal differentiation into ORNs. ORNs express many genes generally characteristic of terminally differentiated neurons, as well as genes specific to their sensory function (odorant receptors, cell-type-specific channels, etc.). Two well-known markers for ORNs are the neuronal cell adhesion molecule (NCAM1) and the olfactory marker protein (OMP). Interestingly, in the main OE, the differentiation and maturation of ORNs from stem and progenitor cells observed during embryonic development appear to be maintained during regenerative neurogenesis, which occurs when ORNs in the main OE have been lost due to disease or environmental insult.

ORNs of the main OE and VNO express different classes of odorant receptors (the catalytic receptors that transduce odor stimuli into cytoplasmic signals), reflecting their different functions: detection of 'common' odors in the main OE, and pheromone detection in the VNO. Odorant receptors in both sensory epithelia are seven-transmembrane, G-protein-coupled receptors. In the main OE, each ORN appears to express only one odorant receptor gene (of over 1000 estimated receptor genes), and those ORNs expressing the same odorant receptor converge their axons homotypically onto selected glomeruli of the MOB. Thus, a specific set of ORNs, with receptors that have been activated by a specific odorant, will stimulate postsynaptic neurons (mitral and tufted cells) within a specific set of MOB glomeruli, which then transduce this information to higher-order olfactory structures in the brain. In the VNO, two families of receptors, the V1Rs and V2Rs, serve as pheromone receptors, and expression of different receptors from these two families appears to divide VNO sensory neurons into two distinct populations: neurons within the apical zone of the VNO neuroepithelium express V1Rs, and their axons project to the anterior portion of the AOB, while neurons in the basal zone of the VNO neuroepithelium express V2Rs and project to the posterior portion of the AOB.

Two Phases of Olfactory Neurogenesis

Both anatomical studies and, more recently, studies of mouse mutants indicate that olfactory neurogenesis can be divided into two phases. The first is an early, morphogenetic phase, in which the neuronal lineage is established and the basic structures of the primary olfactory pathway (OE and VNO) are set up (primary neurogenesis); the second phase is established neurogenesis, during which the characteristic morphology of the nasal cavity and mature pattern of distribution of different cell types within the epithelia emerge, and the signaling systems that control expansion and regenerative neurogenesis come to predominate in the regulation of neuronal cell number. These phases of neurogenesis are depicted in Figure 2.

During primary neurogenesis, the OE forms from the olfactory placodes, thickenings of embryonic ectoderm that first appear as two oval epithelial patches in the anterolateral region of the head around day 9 of gestation (e9) in the mouse. As development proceeds, the olfactory placodes invaginate to form the nasal pits, which by e10.5 are already lined by neuroepithelium that contains cells at every stage of the neurogenic pathway. By e11.5, the nasal pits continue to deepen and fold, and it is during this period that the VNO becomes recognizable as a thickening in the epithelium of the medial wall of the nasal pit. At e12.5, the end of primary neurogenesis, developing nasal turbinates – the elaborate foldings of which will permit a greatly expanded OE surface area – begin to be recognizable in the main olfactory cavity. It is during this period that VNO morphogenesis takes place, as the ventral medial fold of developing neuroepithelium invaginates and pinches off from the main nasal cavity, forming a tube that is closed posteriorly, and anteriorly communicates via its own duct with the developing oral cavity, nasal cavity, or sometimes both.

In the OE that lines the developing nasal pit as early as e10, the three mitotic neuronal cell types (neural stem cells, *Mash1*-expressing progenitors and *Ngn1*-expressing INPs), as well as *Ncam*-expressing ORNs, are all present. These cell types are initially localized in concentric patterns, with the least differentiated cells lying closest to the rim and the most differentiated cells (ORNs) at the center (Figure 3(a)). Expression of *Sox2*, which encodes a marker for OE neural stem cells during established neurogenesis, defines the entire neuroepithelial domain of the developing nasal pit during primary neurogenesis. However, a small group of *Sox2*-expressing cells, which lie closest to the rim of the pit, also express *fibroblast growth factor 8* (*Fgf8*). *Fgf8* encodes a signaling molecule known to be required for OE neurogenesis, nasal cavity formation, and development of the VNO (see later); the cells in the rim of the olfactory pit that co-express *Sox2* and *Fgf8* are thought to be primordial neural stem cells responsible for initiating primary olfactory neurogenesis (by analogy to the primordial germ cells that ultimately give rise to gametes) (Figure 3(b)).

Around e14.5 in the mouse, the overall number of mitotic figures in the OE starts to decrease in the

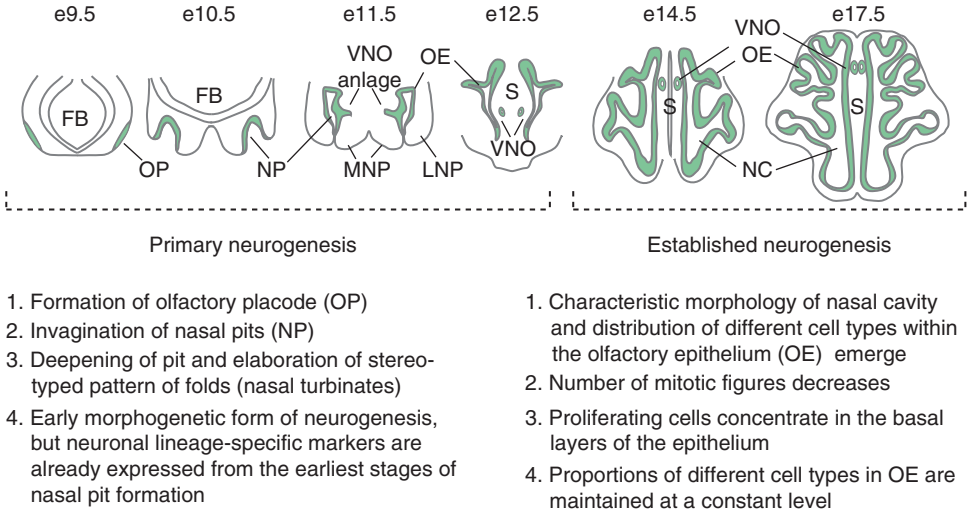

Figure 2 Schematic model and description of primary and established neurogenesis during mouse olfactory epithelium development. Coronal sections through heads at days 9.5–12.5 of gestation (e9.5–e12.5) are depicted for the phase of primary neurogenesis, while horizontal sections at e14.5 and e17.5 (immediately before birth) are portrayed for the phase of established neurogenesis. FB, forebrain; MNP, medial nasal process; VNO, vomeronasal organ; LNP, lateral nasal process; S, nasal septum; NC, nasal cavity. Adapted from Kawauchi S, Shou J, Santos R, et al. (2005) Fgf8 expression defines a morphogenetic center required for olfactory neurogenesis and nasal cavity development in the mouse. *Development* 132(23): 5211–5223.

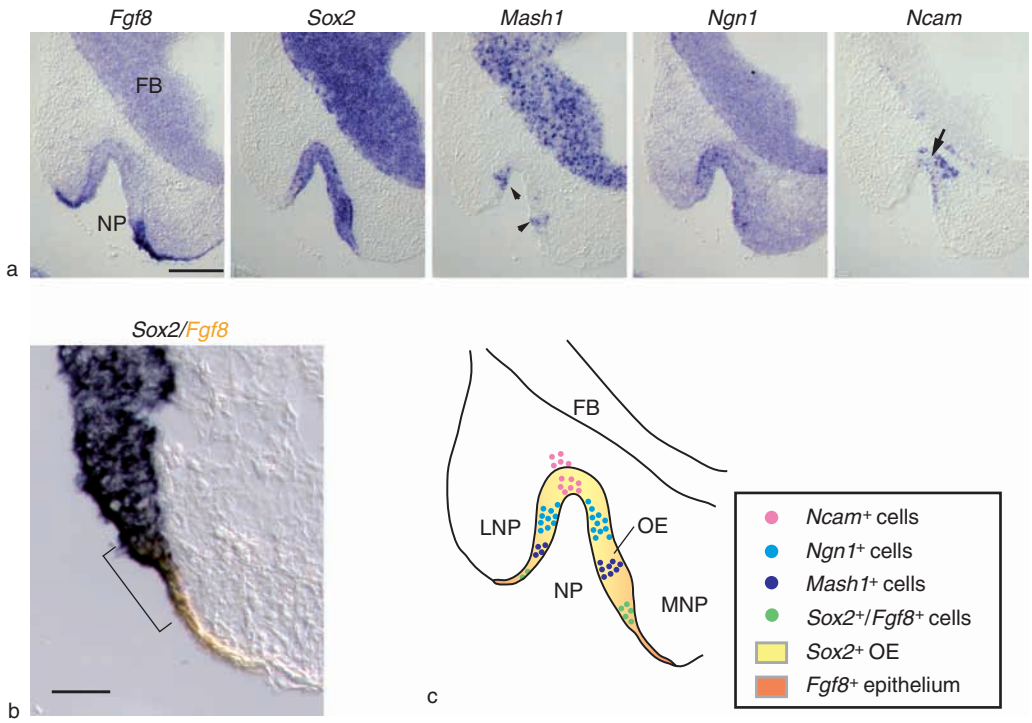

Figure 3 Expression of *Fgf8* and neuronal cell markers in developing olfactory epithelium (OE). (a) Five successive images show *in situ* hybridization for *Fgf8* and OE neuronal lineage markers in invaginating nasal pit (NP) at day 10.5 of gestation (FB, forebrain). While *Sox2* is expressed throughout the neuroepithelium, *Fgf8* is localized to the borders of the invaginating pit. *Mash1* (arrowheads) expression is located next to the *Fgf8*-expressing cells at the inner rim of the nasal pit, while *Ncam*-expressing neurons (arrow) are located at the center of the pit. (b) Double-label *in situ* hybridization for *Fgf8* (orange) and *Sox2* (blue) demonstrates overlap of the two markers in a small rim of surface ectoderm and adjacent invaginating neuroepithelium (bracket). (c) Model of peripheral-to-central process of neuronal differentiation in developing OE and origin of *Sox2*-expressing neural stem cells from *Fgf8*-expressing ectoderm (LNP, lateral nasal process; MNP, medial nasal process). Scale bar = 200 μm (a), 50 μm (b). Adapted from Kawauchi S, Shou J, Santos R, et al. (2005) Fgf8 expression defines a morphogenetic center required for olfactory neurogenesis and nasal cavity development in the mouse. *Development* 132(23): 5211–5223.

epithelium, and different cell types of the OE begin to take on the laminar positions that they will ultimately maintain throughout life. This is the onset of the established phase of neurogenesis, when the mature structure of the nasal cavity also emerges: turbinates and definitive septum form, the sensory epithelium becomes localized to the posterior and dorsal nasal cavity, and the more anterior nasal cavity comes to be lined with respiratory epithelium. Within the OE proper, supporting (sustentacular) cells emerge as a distinct cell population at this time, and their cell bodies begin to form the apical cell layer of the neuroepithelium. *Ncam*-expressing ORNs now lie in the middle compartment of the epithelium, and *Mash1*- and *Ngn1*-expressing neuronal progenitors predominate in the basal compartment of both OE and VNO (Figure 4). With the emergence of this organized pattern begins a new phase of neurogenesis Scale bar = 200 μm (a), 50 μm (b) that is dedicated to maintaining and regulating ORN number. From this time on, the proportions of different cell types are maintained at a fairly constant level within the OE, and the locations of populations expressing different cell-type-specific genes within the epithelium remain essentially the same throughout life.

During the established phase of neurogenesis, an asymmetry develops in the VNO epithelial lining, such that its lateral wall is lined with nonsensory epithelium, whereas the medial wall is lined with a thicker sensory neuroepithelium. This is analogous to the asymmetry that develops in the main nasal cavity, wherein nonsensory, respiratory epithelium predominates in the lining of the anterior and ventral regions of the cavity, whereas the thicker sensory neuroepithelium containing ORNs tends to be localized to more posterior and dorsal regions, covering the turbinates and the posterior two-thirds of the nasal septum.

Intrinsic Factors: Transcription Factors Regulating Mitotic Cell Populations

Neurogenesis and nerve cell renewal take place throughout life in both the OE and the VNO. This

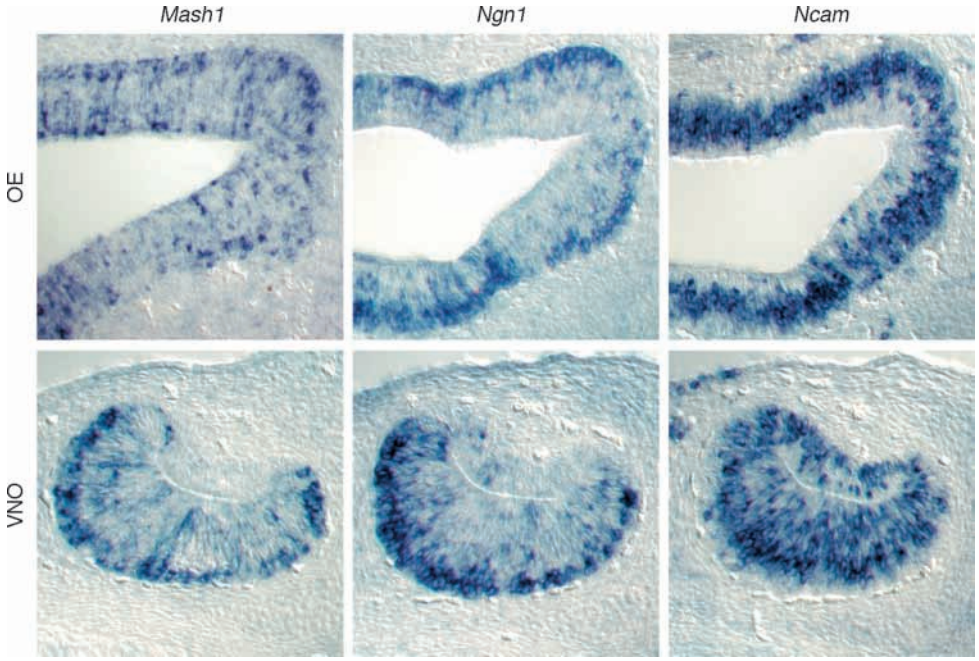

Figure 4 Cell-type-specific markers for committed neuronal progenitors and differentiated olfactory receptor neurons in the established phase of neurogenesis in main olfactory epithelium (OE) and vomeronasal organ (VNO). *In situ* hybridization is shown for mRNAs encoding *Mash1*, *Ngn1*, and *Ncam* in the OE and VNO of mice at day 14.5 of gestation, when the pattern of neurogenesis becomes established in the two sensory epithelia. Even at this stage in prenatal development, the characteristic laminar patterns of cells are apparent, and the relative proportions of cells at different stages of differentiation are similar to what is seen in the postnatal epithelia. (Top panel) *Mash1* and *Ngn1* messages are all expressed in basal areas of the OE, while some apical cells express *Mash1* but not *Ngn1*. *Ncam*, an olfactory receptor neuron marker, is expressed throughout the OE but is absent from the supporting cell layer that lines the nasal cavity. (Bottom panel) *Mash1* and *Ngn1* are similarly located in the VNO, at the boundary between basal epithelium and underlying connective tissue. *Ncam*-positive cells populate the majority of the sensory epithelium, from the concave side of the lumen to the basal layer. Adapted from Murray RC, Navi D, Fesenko J, et al. (2003) Widespread defects in the primary olfactory pathway caused by loss of Mash1 function. *Journal of Neuroscience* 23(5): 1769–1780, with permission. Copyright 2003 by the Society for Neuroscience.

capacity for ongoing neurogenesis is coupled with the ability to regenerate the sensory neuron compartment quickly, at least in the main OE. When ORN number is reduced dramatically by surgical or chemical ablation, progenitor cells in the basal compartment of the OE rapidly upregulate proliferation and produce new differentiated ORNs to replenish the damaged epithelium. These observations imply that stem and neuronal progenitor cells, as well as the microenvironment in which they reside, produce signals that stimulate proliferation and differentiation. Moreover, it must also be the case that OE neuronal stem and progenitor cells express intrinsic factors that endow them with the capacity to respond to these signals.

Work from a number of groups has demonstrated that as cells progress through the OE neuronal lineage, they successively express transcription factors that are characteristic of, and required for, differentiation of stem cells into committed neuronal progenitors and, ultimately, ORNs. Thus, expression of these cell-intrinsic factors both provides unique molecular 'signatures' for neuronal progenitor cells at specific developmental stages and determines the ultimate fates of these cells. This feature makes the OE an ideal system for studies of the regulatory roles that such transcription factors play in neurogenesis, and also provides the molecular markers needed to decode effects of extrinsic factors, such as signaling molecules, on the regulation of neurogenesis. Indeed, studies of mice with mutations in genes encoding regulatory transcription factors have proved to be extremely informative in understanding the roles that such factors play in regulating ORN and VNO development.

Although no definitive stem cell marker has been found for OE and VNO stem cells, a likely candidate is the transcription factor, Sox2. Sox2, a transcription factor of the SoxB1-type SRY transcription factor family, is thought to be a general neuronal stem cell marker: It is expressed throughout the neural primordium in rodents, is an important regulator of embryonic development, and has been shown to direct neural progenitor identity. Moreover, in a number of

neural tissues, *Sox2*-expressing cells have been shown to be capable of both self-renewal and differentiation, suggesting that *Sox2* gene expression is a trait shared by stem cells in many neural systems. During primary neurogenesis in the OE, expression of *Sox2* mRNA defines the neuroepithelial domain of the invaginating olfactory pit, and co-expression of *Sox2* and *Fgf8* has recently been reported to define a population of primordial neural stem cells that will give rise to all subsequent neural stem and progenitor cell types of the OE (**Figure 3**). Indeed, in mice in which the *Fgf8* gene is inactivated in anterior neural structures, this *Sox2–Fgf8* co-expressing cell population undergoes apoptosis, leading to a failure in subsequent OE neurogenesis, nasal cavity formation, and morphogenesis of the VNO (see later). Thus, although no stem cell marker has been identified for the VNO, the observation that the VNO fails to develop at all in *Fgf8* conditional mutant mice strongly suggests that the *Sox2–Fgf8* co-expressing primordial neural stem cells, observed to play a critical role in the early stages of primary neurogenesis in the OE, give rise to the neurogenic population of the VNO as well.

During primary neurogenesis at e10–e11, cells expressing the proneural gene, *Mash1*, are found in close apposition to *Sox2–Fgf8* co-expressing primordial neural stem cells near the rim of the invaginating olfactory pit (**Figure 3**). A day or so later in development (e12.5), *Mash1* mRNA can be detected in cells found in the apical, middle, and basal compartments of the OE, coincident with the location of mitotic figures at this age. As the OE matures and enters the phase of established neurogenesis, *Mash1*-expressing cells come to be located primarily in the basal compartment of the OE, suggesting that the action of *Mash1* is required early in the ORN lineage. *Mash1* has been shown in genetic studies to be required for ORN development, and in studies *in vitro*, in OE cultures, and *in vivo*, in surgical models of induced neurogenesis, *Mash1* has been shown to be expressed by early-stage transit-amplifying progenitors of the ORN lineage (**Figure 1(b)**). Indeed, in mice with targeted inactivation of the *Mash1* gene, *Ngn1*-expressing INPs, as well as ORNs, fail to develop, indicating that *Mash1* acts upstream of *Ngn1* to direct neuronal differentiation in the OE.

Recent studies indicate that, in the absence of *Mash1* function, the OE reverts to a state in which it maintains high levels of both proliferation and apoptosis. Proliferating cells express *Sox2*, the *Mash1* 3′-untranslated region (3′-UTR; which is still present in the targeted mutant), and *Steel*, a marker of supporting cells. This has led to the hypothesis that these proliferating cells in the OE of *Mash1*$^{-/-}$ mice are 'frozen' at an early stage of differentiation, and the fact that these cells co-express markers of both supporting cells (*Steel*) and early neuronal progenitors (*Mash1* 3′-UTR) suggests that they would be capable, if they did not undergo apoptosis (the precipitating cause of neurogenic failure in the *Mash1* mutant), of giving rise to both ORNs and supporting cells. These observations provide indirect evidence that the neural stem cell of the OE is a bipotential stem cell, capable of giving rise to both glial and neuronal cell types. Interestingly, neurogenesis also fails in the VNO of *Mash1*$^{-/-}$ mice, and in the same manner as in main OE: *Ngn1* expression fails to occur and neurons fail to form, while the *Mash1* 3′-UTR is expressed in abundant proliferating cells that undergo high levels of apoptosis. Thus, the developmental hierarchy of gene expression in both the main OE and the VNO appear to be fundamentally similar.

As they progress through the ORN lineage, *Mash1*-expressing progenitors lose expression of *Mash1* and upregulate expression of a different proneural gene, *Ngn1*. *Ngn1* expression defines the immediate neuronal precursor, which has been shown by tissue culture and genetic studies to be committed to ORN differentiation after one to two rounds of division. Evidence for the function of *Ngn1* as a neural determination gene first came from studies in *Xenopus*. In *Xenopus*, misexpression of an *Ngn1* homolog can convert nonneurogenic ectodermal cells to neurons. In *Ngn1* mutant OE, most ORNs fail to develop and differentiate (at least by the end of primary neurogenesis at e12.5), suggesting that mammalian *Ngn1* plays a role similar to that of its *Xenopus* counterpart. Moreover, although *Ngn1* expression is severely reduced in the OE of *Mash1* mutant mice, *Mash1* expression is not significantly affected in the OE of *Ngn1*-null mice, indicating that *Mash1* and *Ngn1* expression are essential at different stages of differentiation and that *Mash1* acts upstream of *Ngn1* in the ORN lineage.

A number of other transcription factor genes play roles in regulating neuronal differentiation in the OE, including *RunX1* and *NeuroD*. *Runx1* encodes a member of the Runt/Runx family of transcription factors, whereas *NeuroD*, like *Mash1* and *Ngn1*, encodes a basic helix–loop–helix transcription factor. In the OE, expression of *Runx1* and *NeuroD* is restricted primarily to cells in the basal half of the epithelium. Evidence from developmental genetic studies suggests that *NeuroD* is expressed at the stage when late, *Ngn1*-expressing neuronal progenitors are just differentiating into ORNs. Gene expression studies indicate that only a few cells co-express *Runx1* and *Mash1*, whereas virtually all cells expressing *NeuroD* also express *Runx1*. Thus, expression data suggest that both *Runx1* and *NeuroD* act at the time when

late-stage neuronal progenitors (INPs) are undergoing terminal differentiation into ORNs. OE development in mice with targeted inactivation of the *Runx1* gene has only been examined up to the end of the primary phase of neurogenesis, since homozygous nulls die at e12.5. Interestingly, in the OE of e12.5 *Runx1*$^{-/-}$ embryos, the total number of cells, and the number of *Mash1*-expressing cells, appear to be unchanged; however, there is a decrease in the number of *NeuroD*-expressing cells and an increase in cells expressing the early neuronal marker β-III tubulin. Since Runx1 is also known to repress expression of cyclin-dependent kinase inhibitors (which act as 'brakes' on mitotic cells in the G1/S transition; see later), these observations have been interpreted as showing a role for Runx1 in regulating *NeuroD* expression and terminal differentiation of OE neuronal progenitors into postmitotic ORNs. Since *NeuroD* is expressed in the VNO in a pattern analogous to its expression in OE, by extension it seems possible that Runx1 may function in this tissue as well to regulate sensory neuron differentiation, although this has not yet been investigated.

Control of Neurogenesis by Extrinsic Factors

Developmental transitions in expression of transcriptional regulators, cell proliferation, cell differentiation, and intraepithelial cell location in the OE are directed by the actions of extrinsic signaling molecules. Different signaling molecules appear to predominate during the primary and established phases of neurogenesis, and these factors are produced both within the OE itself and by its underlying mesenchymal stroma (also known as the lamina propria of the mature epithelium). The actions of these secreted signaling molecules in regulating programs of neuronal cell proliferation and differentiation in the OE are classified into two categories: (1) proneurogenic effects, which are positive effects on OE neurogenesis and include stimulation of progenitor cell proliferation and cell survival, and (2) antineurogenic effects, which include suppression of cell proliferation and, directly or indirectly, increases in cell death (apoptosis). Both categories of action are important for OE and VNO morphogenesis and acquisition and maintenance of proper sensory neuron number, and therefore for OE function during development and postnatal life. Over the past several years, studies have revealed that OE neurogenesis is critically dependent on signaling molecules from two different polypeptide growth factor superfamilies, fibroblast growth factors (FGFs) and transforming growth factor-βs (TGF-βs). Each of these superfamilies of signaling molecules includes many members, all of which have specific patterns of expression and differing functions during development and tissue homeostasis. During OE neurogenesis it has been found that different factors from these two superfamilies interact in at least two ways: first, opposing signals converge on neuronal stem and progenitor cells at specific developmental stages in the ORN lineage, regulating proliferation and the stepwise progression of neuronal differentiation. Second, TGF-βs and their secreted antagonists play key roles in feedback loops that regulate the size of progenitor cell pools and the number of ORNs that differentiate from these cell pools. Since these developmental pathways have been worked out almost exclusively in the main OE of rodents, particularly the mouse, the following discussion deals primarily with mouse main OE.

Fibroblast Growth Factors

During OE development, several different FGFs are known to be expressed, and since expression of these FGFs is developmentally regulated, it is thought that the functions of different FGFs predominate at different developmental stages. It has recently been established that FGF8 plays a crucial role in OE development during the early stages of primary neurogenesis. Tissue-specific inactivation of the *Fgf8* gene has shown that *Fgf8* expression, which is highest in the ectoderm and neuroepithelium that outlines the rim of the invaginating pit at e10–e11 in the mouse, defines a morphogenetic center of function that is crucial for OE neurogenesis, nasal cavity morphogenesis, and development of the VNO. As mentioned previously, a subpopulation of *Sox2*-expressing neural stem cells, located in the outermost rim of the invaginating OE neuroepithelium, also express *Fgf8* (**Figure 3**); these cells have been termed 'primordial' neural stem cells of the OE. In mice with tissue-specific inactivation of *Fgf8*, these *Sox2–Fgf8* co-expressing cells undergo apoptosis (**Figure 5(a)**). As a result, OE neurogenesis fails, as does nasal cavity formation; indeed, by e14.5, only a small vestige of remaining OE, with just a few neural cells, can be observed – and this is seen only in the least severely affected mutants (**Figures 5(b) and 5(c)**). These animals also lack any VNO, indicating that FGF8 is a crucial signal for morphogenesis and neurogenesis in the VNO as well as in the main OE. The effects of the neurogenic deficits resulting from absence of FGF8 are long lasting: although animals survive until birth, no OE or VNO ever forms, and only a few, scattered neural cells are observed in the malformed oral–nasal cavity that does develop. These observations indicate

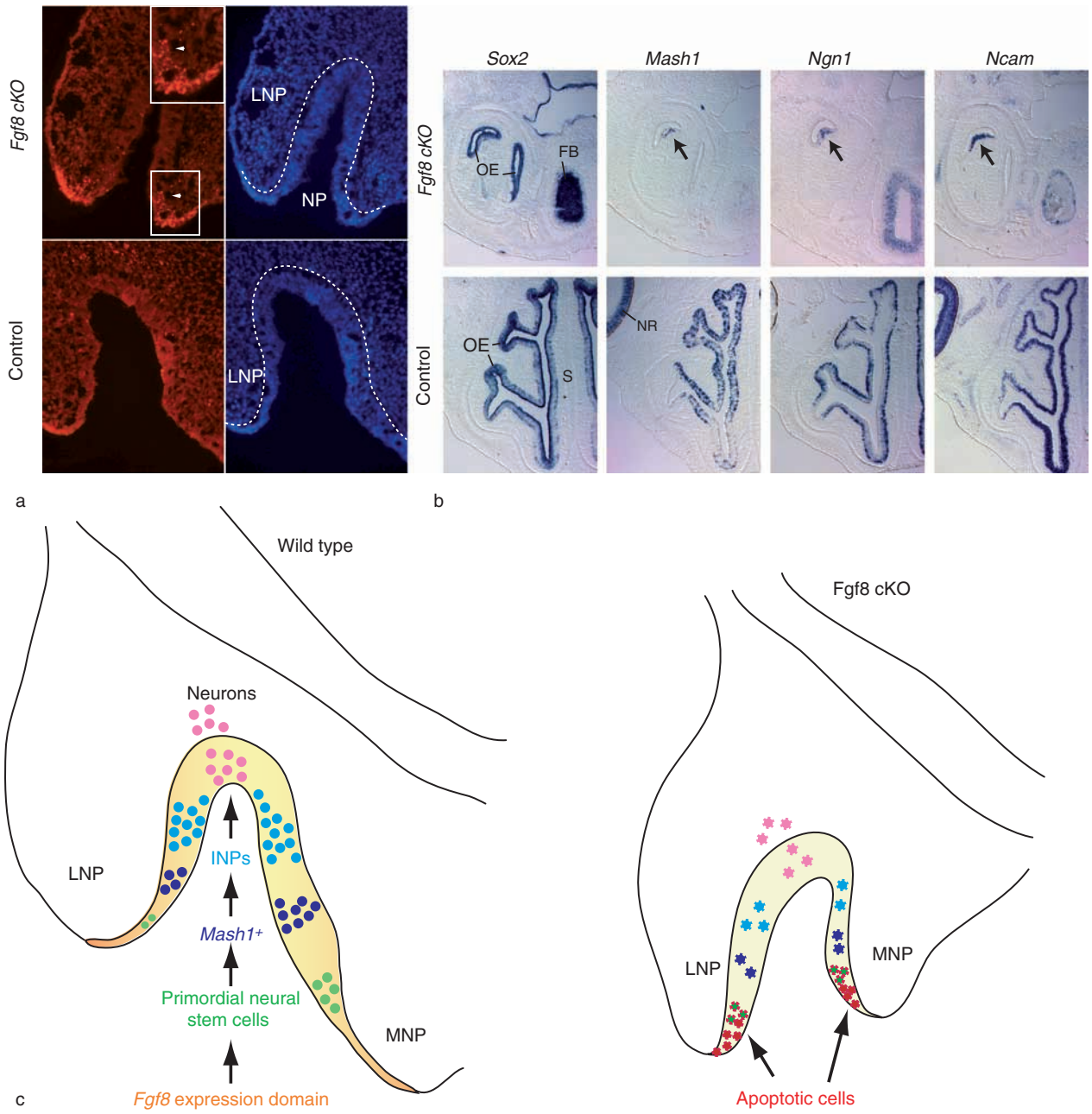

Figure 5 *Fgf8* is required for cell survival in the neurogenic domain. (a) Terminal dUTP nick end labeling (TUNEL) of olfactory epithelium of mutant (cKO; conditional knockout) and control littermates at day 10.5 of gestation shows a high number of apoptotic cells in mutants in ectoderm and olfactory epithelium (white arrowhead; magnified in inset) of invaginating nasal pit (NP). Hoescht panel (blue, right) shows extent of invaginating NP. Broken white line indicates the boundary of the neuroepithelium lining the NP and lateral nasal process (LNP). (b) *In situ* hybridization on horizontal sections of mutant and control littermates (day 14.5 of gestation) shows the near absence of OE in the mutants and the lack of neuronal markers in the remaining OE (arrow) (FB, forebrain; S, nasal septum; NR, neural retina). (c) Role of *Fgf8* in olfactory neurogenesis. Schematic of primary neurogenesis at day 10.5 of gestation in wild-type OE and *Fgf8* mutant OE, illustrating the relative positions (MNP, medial nasal process) of the *Fgf8* expression domain and different neuronal cell types: *Fgf8* expression domain, orange; *Sox2*-expressing neuroepithelium, yellow; *Sox2*- and *Fgf8*-expressing primordial stem cells, green; *Mash1*-expressing progenitors, blue; immediate neuronal precursors (INPs),turquoise; *Ncam*-expressing olfactory receptor neurons, pink. In the mutant invaginating pit, cells undergoing apoptosis due to *Fgf8* inactivation are shown in red and apoptotic primordial neural stem cells are green with red jagged borders. Diminished populations of other neuronal cell types are shown in their corresponding colors, but with jagged edges. Adapted from Kawauchi S, Shou J, Santos R, et al. (2005) Fgf8 expression defines a morphogenetic center required for olfactory neurogenesis and nasal cavity development in the mouse. *Development* 132(23): 5211–5223.

that FGF8 is required not only for the survival and expansion of the neural stem cell pool during primary neurogenesis (Figure 5(c)), but also that it is required for maintenance of primary neurogenesis and, as a result, the initiation of the established phase of neurogenesis. Interestingly, *Fgf8* expression declines by the end of primary neurogenesis (e12.5), when another closely related FGF, encoded by *Fgf18*, begins to be expressed within the neuroepithelium of the OE. It has been hypothesized that FGF18 assumes FGF8's role in maintenance of the stem cell population during the embryonic stages of the established phase of neurogenesis, although this idea has not been tested directly in *Fgf18* mutant mice.

Unlike *Fgf8*, *Fgf2* does not appear to be expressed at significant levels in prenatal mouse OE, but in adult mice, it is expressed throughout the neuronal cell layers of the main OE. By 3 weeks of age, FGF2 expression can also be detected in axon bundles that converge below the basal lamina, en route to the olfactory bulb. Like FGF8, FGF2 can also function as a proneurogenic factor: In cultures of OE, addition of FGF2 increases the number of proliferating INPs, by promoting multiple rounds of INP divisions before INP daughter cells undergo terminal differentiation into ORNs; FGF2 has also been shown to be capable of promoting survival and proliferation of putative stem cells in OE cultures. Taken together with the pattern of *Fgf2* expression, these observations suggest that FGF2 is likely to be involved in maintaining neuronal stem and progenitor cells, as well as in serving as a mitogen for these cells, in order to regulate ORN number in postnatal OE.

TGF-β Superfamily

Bone morphogenetic proteins Many studies indicate that the bone morphogenetic proteins (BMPs), a subfamily of TGF-βs structurally related to the *Drosophila* signaling molecules Dpp and 60A, have important roles in regulating neuronal fate determination and neurogenesis during vertebrate development. For example, endogenous BMP4 promotes acquisition of an epidermal fate, at the expense of neural tissue, in developing ectoderm, and BMP2 and BMP4 have been shown to inhibit proliferation and/or induce apoptosis of neural progenitor cells in several systems. Studies of the OE have shown that BMPs can exert both pro- and antineurogenic effects, depending on the concentration and identity of the BMP in question and the identity of the target cell that is acted upon.

Bmp2, *Bmp4*, and *Bmp7* are all expressed in and/or near OE proper during embryonic development in the mouse, and all three of the secreted proteins (BMP4, BMP2, and BMP7) can inhibit OE neurogenesis *in vitro* by acting on *Mash1*-expressing neuronal progenitors. Exposure to any of the three BMPs causes these progenitors to target preexisting *Mash1*-encoded protein for proteasome-mediated degradation, resulting in apoptosis and termination of the ORN developmental pathway at the *Mash1*-expressing stage. Interestingly, however, low concentrations of BMP4, but not BMP7, stimulate OE neurogenesis: In this case, the action of BMP4 is to promote survival of newly generated ORNs. Since BMP4 and BMP7 selectively activate different subsets of serine–threonine kinase type I BMP receptors, which are known to be differentially expressed within the OE, it is likely that differential responsiveness of specific OE neuronal stem/progenitor cells to different concentrations of specific BMPs is dictated by the identity of the cell surface receptors that are activated in the various target cell types. The OE is thus proving to be a useful model system for dissecting out the cellular basis of concentration-dependent responses to BMPs, an area of broad interest to both developmental biologists and cancer biologists.

Growth and differentiation factor 11 and feedback inhibition of neurogenesis Studies *in vitro* and *in vivo* have shown that generation of new ORNs by their progenitors is inhibited by a signal produced by OE neuronal cells themselves; this process has been termed feedback inhibition of neurogenesis. Recently, growth and differentiation factor 11 (GDF11), a member of a small subgroup of activin-like TGF-β s that includes the muscle regulatory factor GDF8 (also called myostatin), has been shown to mediate feedback inhibition of INP proliferation during the established phase of OE neurogenesis. *Gdf11* is expressed by both committed neuronal progenitors and newly differentiated ORNs in the OE, as are its serine–threonine kinase transmembrane receptors (ALK5 and ActRIIb). Both *Gdf11* and the gene that encodes its secreted antagonist, follistatin (FST), are first expressed at significant levels at the onset of the established phase of neurogenesis, and their expression continues through adult life (*Gdf11* is expressed only within OE neuroepithelium proper; *Fst* is expressed in OE stroma and, at lower levels, in OE proper). Recombinant GDF11 inhibits OE neurogenesis *in vitro*, by reversibly arresting INP divisions; this action is accompanied by induction of the cyclin-dependent kinase inhibitor $p27^{Kip1}$ in these neuronal progenitors. The antiproliferative effect of GDF11 on INPs is abrogated by addition of FST to OE cultures, but, interestingly, GDF11 actions

predominate over the proliferative effect of FGFs on INPs. *Gdf11*-null mice show increased cell proliferation in the OE, and the number of *Ngn1*-expressing INPs is increased, as is overall OE thickness, relative to that of wild-type control littermates. These supernumerary INPs go on to give rise to ORNs, as demonstrated by an increase in *Ncam*-expressing cells in *Gdf11*-null OE (Figure 6(a)). FST, a secreted protein that prevents effective binding and signaling by GDF11, is able to completely antagonize the antineurogenic effect of GDF11 on INPs in OE cultures. Importantly, the number of INPs and ORNs is dramatically decreased in $Fst^{-/-}$ OE, consistent with what would be expected with removal of a crucial brake on the antineurogenic actions of endogenous GDF11 in the OE neuroepithelium (Figure 6(b)). These observations have led to the proposal of a model in which GDF11 and FST interact to maintain neuron number in the OE by tightly regulating INP divisions and ORN production. Since GDF11 arrests INP proliferation but does not kill INPs, its role in feedback inhibition of OE neurogenesis is likely to be one of maintaining INPs in stasis, such that they are capable of responding rapidly to a decrease in ambient GDF11 levels (as when OE neuronal cells are lost through injury or disease) by rapidly reentering the cell cycle and giving rise to ORNs (Figure 6(c)).

Summary and Conclusions

In addition to revealing major steps in morphogenesis and neurogenesis of the OE and VNO, studies of the development of these two sensory neuroepithelia have revealed important aspects of the molecular regulation of neurogenesis. Studies of main OE development in particular have led to the concept that neurogenesis in olfactory sensory epithelia proceeds in a linear manner, with expression of specific transcription factors marking the progressive restriction in fate observed as neuronal stem and committed progenitor cells differentiate through the ORN developmental pathway. Both these intrinsic signals, which operate in a cell-autonomous manner, and extrinsic (secreted) signals present in the local microenvironment of developing cells (the stem/progenitor cell niche) play important roles in regulating this process. The factors that predominate during primary olfactory neurogenesis play an expansionary, morphogenetic role. For example, FGF8, which is crucial for increasing the size of the stem cell pool, is of vital importance at this stage in order to establish the ORN lineage in the main OE and to permit formation of the VNO (Figure 7). Once the OE neurogenesis enters its established phase, both FGFs and TGF-βs participate in regulating neuronal cell number. Studies of the

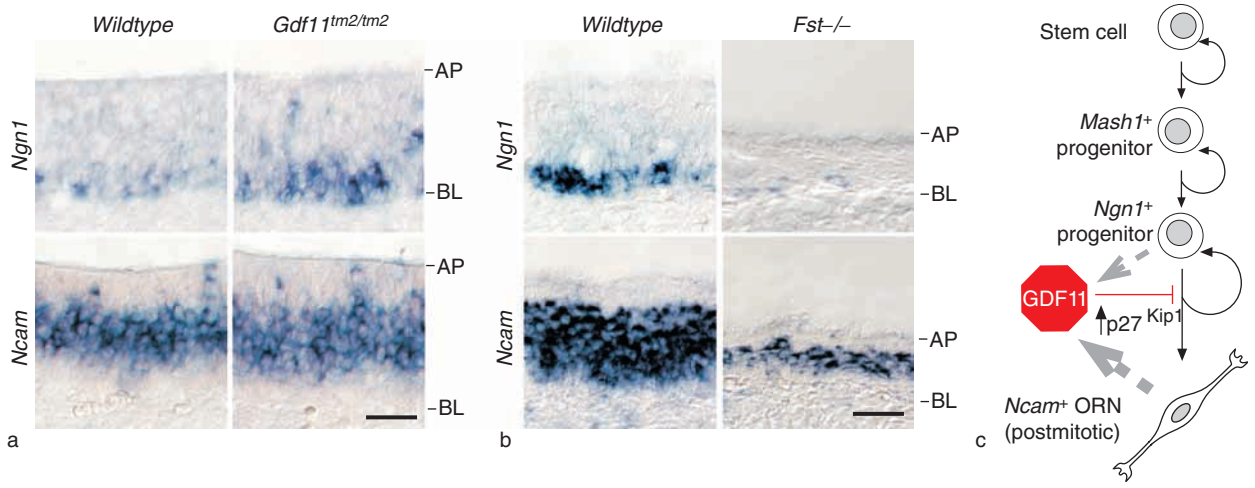

Figure 6 Disruption of neurogenesis in mice with loss or gain of function of *Gdf11*. (a) *In situ* hybridization on horizontal sections (AP, apical layer; BL, basal layer) of olfactory epithelium of $Gdf11^{tm2/tm2}$ mice reveals an increase in *Ngn1*-expressing cells and a corresponding increase in *Ncam*-expressing cells: the *Ncam*-expressing cell layer is thicker by 20% (9 μm), about the diameter of one olfactory receptor neuron (ORN), in $Gdf11^{tm2/tm2}$ olfactory epithelium, compared to in wild type. (b) Mice lacking a functional *follistatin* (*Fst*) gene show decreased olfactory epithelium neurogenesis. *In situ* hybridization for *Ngn1* and *Ncam* shows large decreases in expression of both markers as well as aberrantly thin olfactory epithelium in the *Fst* mutant. (c) Schematic model of growth and differentiation factor 11 (GDF11) action in ORN neurogenesis. GDF11 is produced by both *Ngn1*+ progenitors and *Ncam*+ ORNs (gray broken-line arrows). GDF11 reversibly arrests *Ngn1*+ progenitors through induction of the cycle-dependent kinase inhibitor $p27^{Kip1}$, thus preventing ORN generation. Reproduced from Wu HH, Ivkovic S, Murray RC, et al. (2003) Autoregulation of neurogenesis by GDF11. *Neuron* 37: 197–207, with permission from Elsevier.

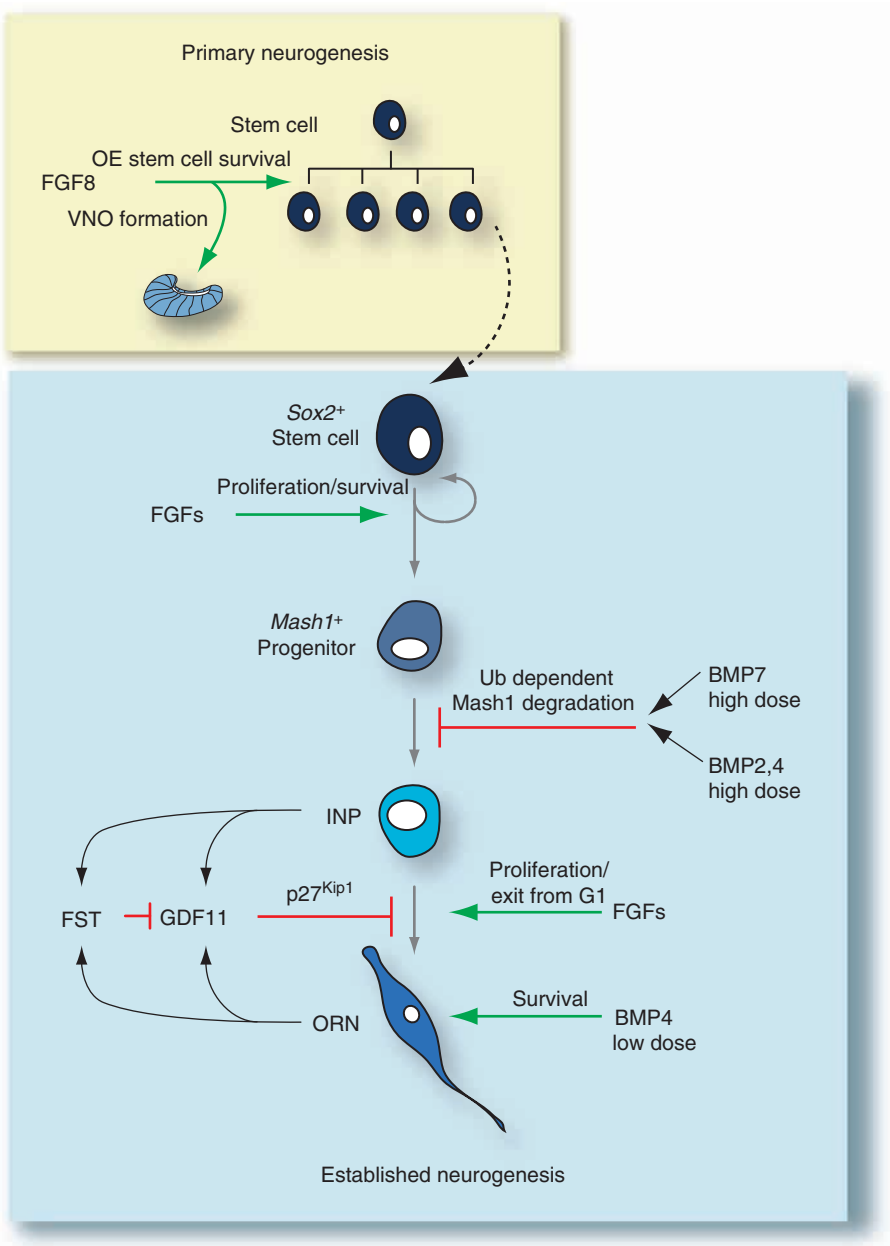

Figure 7 Summary of molecular signals involved in both primary and established olfactory epithelium (OE) neurogenesis. During primary neurogenesis (yellow box), fibroblast growth factor 8 (FGF8) and other factors are involved in producing the stem cell pool and initially establishing the different progenitors in the olfactory receptor neuron lineage. FGF8 is also necessary for initial vomeronasal organ (VNO) formation. After day 12.5 of gestation, once the lineage is established (blue box), FGFs and transforming growth factor-βs (BMP, bone morphogenetic protein; FST, follistatin; GDF, growth and differentiation factor) converge on different cell types in the OE lineage to achieve and maintain proper neuron number. Green arrows indicate stimulatory interactions; red bars indicate inhibitory interactions. Adapted from Kawauchi S, Beites CL, Crocker CE, et al. (2004) Molecular signals regulating proliferation of stem and progenitor cells in mouse olfactory epithelium. *Developmental Neuroscience* 26: 166–180, with permission. Copyright 2004 by S. Karger AG, Basel, Switzerland.

actions of TGF-βs and their secreted antagonists in regulating OE neurogenesis have led to the concept of feedback inhibition of neurogenesis, in which feedback circuits consisting of endogenous secreted antineurogenic factors regulate the size of the stem/progenitor cell pool and total neuron number. Ultimately, determining whether similarities in these processes exist between the OE and other regions of the nervous system should provide important knowledge for understanding how neuronal stem/progenitor cell pools are controlled in all neural tissues, and for overcoming roadblocks to recovery in regions of the

adult nervous system where – unlike the OE – neuronal regeneration is limited or absent.

See also: Drosophila Apterous Neurons: from Stem Cell to Unique Neuron; Olfactory Ensheathing Cells: Regeneration Following Spinal Cord Injury; Olfactory Receptor Genes: Human Loss During Evolution; Olfactory Receptors: Microarray Analysis; Olfactory Epithelium; Vomeronasal System Evolution; Vomeronasal Accessory System.

Further Reading

Beites CL, Kawauchi S, Crocker CE, et al. (2005) Identification and molecular regulation of neural stem cells in the olfactory epithelium. *Experimental Cell Research* 306: 309–316.

Bhattacharyya S and Bronner-Fraser M (2004) Hierarchy of regulatory events in sensory placode development. *Current Opinion in Genetics & Development* 14: 520–526.

Buck L and Axel R (1991) A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* 65: 175–187.

Caggiano M, Kauer JS, and Hunter DD (1994) Globose basal cells are neuronal progenitors in the olfactory epithelium: A lineage analysis using a replication-incompetent retrovirus. *Neuron* 13: 339–352.

Cau E, Gradwohl G, Fode C, et al. (1997) Mash1 activates a cascade of bHLH regulators in olfactory neuron progenitors. *Development* 124: 1611–1621.

Chang H, Brown CW, and Matzuk MM (2002) Genetic analysis of the mammalian transforming growth factor-beta superfamily. *Endocrine Reviews* 23: 787–823.

Dulac C and Torello AT (2003) Molecular detection of pheromone signals in mammals: From genes to behaviour. *Nature Reviews Neuroscience* 4: 551–562.

Farbman A (1992) *Cell Biology of Olfaction*. Cambridge, UK: Cambridge University Press.

Guillemot F, Lo C-C, Johnson JE, et al. (1993) Mammalian achaete-scute homolog 1 is required for the early development of olfactory and autonomic neurons. *Cell* 75: 463–476.

Halpern M and Martinez-Marcos A (2003) Structure and function of the vomeronasal system: An update. *Progress in Neurobiology* 70: 245–318.

Itoh N and Ornitz DM (2004) Evolution of the Fgf and Fgfr gene families. *Trends in Genetics* 20: 563–569.

Kawauchi S, Beites CL, Crocker CE, et al. (2004) Molecular signals regulating proliferation of stem and progenitor cells in mouse olfactory epithelium. *Developmental Neuroscience* 26: 166–180.

Kawauchi S, Shou J, Santos R, et al. (2005) Fgf8 expression defines a morphogenetic center required for olfactory neurogenesis and nasal cavity development in the mouse. *Development* 132(23): 5211–5223.

Mombaerts P (2001) How smell develops. *Nature Neuroscience* 4(supplement): 1192–1198.

Murray RC and Calof AL (1999) Neuronal regeneration: Lessons from the olfactory system. *Seminars in Cell and Developmental Biology* 10: 421–431.

Murray RC, Navi D, Fesenko J, et al. (2003) Widespread defects in the primary olfactory pathway caused by loss of Mash1 function. *Journal of Neuroscience* 23: 1769–1780.

Nicolay DJ, Doucette JR, and Nazarali AJ (2006) Transcriptional regulation of neurogenesis in the olfactory epithelium. *Cellular and Molecular Neurobiology* 26: 801–819.

Pevny L and Placzek M (2005) SOX genes and neural progenitor identity. *Current Opinion in Neurobiology* 15: 7–13.

Vincent AJ, West AK, and Chuah MI (2005) Morphological and functional plasticity of olfactory ensheathing cells. *Journal of Neurocytology* 34: 65–80.

Wu HH, Ivkovic S, Murray RC, et al. (2003) Autoregulation of neurogenesis by GDF11. *Neuron* 37: 197–207.

Olfactory Receptor Genes: Human Loss During Evolution

Y Gilad, University of Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Buck and Axel identified the odorant receptor gene family more than a decade ago, based primarily on the observation that the candidate genes were exclusively expressed in olfactory epithelium in rats, but not detected in a number of other nonolfactory tissues. They originally hypothesized that there were at least 100–200 olfactory receptor (OR) genes in the rat genome. Nearly all mammalian OR genes are encoded by only one coding exon of roughly 1 kb. They are members of the G-protein-coupled receptor (GPCR) gene superfamily and based on their protein sequence are assumed to have the hallmark seven transmembrane domain structure. These features were used to recognize additional coding regions of OR genes in mammalian genomic sequence. Specifically, the search relied on their similarity to the first set of confirmed OR genes and by the presence of certain protein motifs.

Mammalian OR genes are currently divided into two classes, 17 families and ~250 subfamilies, based on protein sequence similarity. Class I OR genes are closely related to OR genes that are found in fish and hence are referred to as 'fishlike,' while class II OR genes are specific to tetrapods. OR genes from the same family and subfamily are defined as sharing 40% and 60% or more of their amino acid sequence, respectively. It is thought that OR genes from the same subfamily bind the same type of odorants, while differing in their affinity to close structural variants.

Recently, the complete genomic sequence of a number of mammalian species became available, permitting interspecies comparisons of complete OR gene repertoires. Such interspecies comparisons increase our understanding of the evolution of the OR gene repertoires and, in particular, of the molecular basis for inter- and intraspecies differences in the sense of smell. The analyses of published mammalian genomes suggested that the original estimate of the size of the mammalian OR gene repertoire was a severe underestimate. Indeed, it is now thought that mammalian genomes carry 800–1400 OR genes, which are typically organized in gene clusters and are found on many chromosomes. With roughly 3% of all genes coding for odorant receptors, OR genes are by far the largest gene family in mammalian genomes.

In humans, OR genes have a tendency to cluster in subtelomeric regions, and they are found on all chromosomes except for chromosomes 20 and Y. Class I OR genes are all found in a single cluster on chromosome 11. These genes were originally thought to be an evolutionary relic, the original function of which was to bind water-soluble odorants (in fish). However, many class I OR genes remained intact, putatively functional, in human as well as in other mammalian species. This observation raises the possibility that class I OR genes now have a function in terrestrial mammals.

The largest cluster of class II human OR genes is found on chromosome 11, and is thought to be the original tetrapod-specific OR gene cluster from which all other class II OR genes have originated. The OR gene family is thought to have evolved in mammals through a process of gene birth and death, with an evolutionary pressure for functional diversification of the repertoire (through mutations and/or gene conversion events).

Human Loss of OR Genes

The completion of the human genome project enabled the identification of the entire repertoire of human OR genes. Humans have roughly 900 OR genes, but more than 50% of them carry one or more coding region disruptions and therefore are annotated as pseudogenes. However, not all OR pseudogenes represent gene loss. For example, roughly 80 of the human OR pseudogenes belong to subfamily 7E, a recent expansion of a cluster of three OR pseudogenes that is thought to have occurred only in apes.

Early anecdotal observations of a human-specific OR coding region disruption raised the hypothesis that the rapid accumulation of OR pseudogenes occurred in parallel to a reduction in the sense of smell in primates and, specifically, in humans. This hypothesis found some measure of support when the mouse genome, and consequently its complete repertoire of OR genes, became available. The mouse OR gene repertoire is ~30% larger than that of man, but contains only 20% pseudogenes. Thus, the mouse putative functional OR gene repertoire is estimated to be more than three times larger than that of humans. Similarly low proportions of OR pseudogenes were later found in dogs and rats.

Among apes, the relative rates of accumulating OR coding region disruptions were first estimated by direct sequencing of 50 randomly chosen human OR genes and orthologs from chimpanzee, gorilla,

orangutan, and one Old World monkey (rhesus macaque). Using this data set, OR coding region disruptions could be mapped to specific lineages (Figure 1) and their lineage-specific fixation rates could be estimated. This analysis suggested that while apes have been accumulating OR pseudogenes faster than rodents, humans have been accumulating OR pseudogenes much faster than any of the other apes.

Recently, the availability of the chimpanzee genomic sequence enabled a direct comparison of the entire OR gene repertoires of humans with that of our closest evolutionary relative. In agreement with previous estimates based on small samples of OR genes, it was found that humans have a significantly smaller number of intact OR genes than chimpanzees (Figure 2). Moreover, the hypothesis, that humans and chimpanzees accumulated OR pseudogenes at a constant rate since the divergence of human and chimpanzee could be rejected, again suggested that the rate of OR gene disruption was higher in the human lineage.

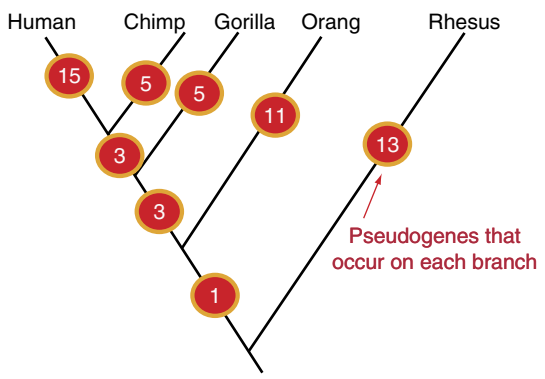

Figure 1 Gilad et al. resequenced 50 orthologous OR genes from five primate species. Of these 50 OR genes, the number of OR genes that were lost in each branch is indicated. The branch lengths represent the species divergence time. As can be seen from this sample, humans have lost OR genes much more rapidly than any of the other species.

Thus, the picture that emerges was that of a decrease in the size of the putatively functional OR gene repertoire in apes relative to other mammals, with a further deterioration specifically in the human evolutionary lineage. This relaxation of evolutionary constraint on human OR genes is thought to be the result of differences between humans and other apes in several aspects of the lifestyle that have been shown to be influenced by the olfactory system, including differences in social behavior, shifts in nutrition, and the factors that determine mating choice.

Although the causes are unclear, it seemed reasonable to speculate that the high fraction of OR pseudogenes in apes compared to rodents reflects a decreased reliance on the sense of smell in species for whom auditory and visual cues may be more important. To examine this, a random set of 100 OR genes were sequenced from 19 primate species, including apes, Old World monkeys, and New World monkeys. These OR loci were not sets of orthologous genes, so that the lineage-specific rates of accumulating OR coding region disruptions could not be estimated. However, under the assumption that the OR gene repertoire is of similar size in all the studied species (a reasonable assumption given that the size of the OR gene repertoire in human, chimpanzee, mouse, rat, and dog was found to be similar), the number of putatively functional OR genes in each species can be estimated using the proportion of pseudogenes in the random samples that were sequenced. Using this approach, the intact OR gene repertoire of most New World monkeys was estimated to be similar to that of mouse, dog, and rat. In contrast, significantly higher proportion of OR pseudogenes were found in two distinct evolutionary lineages: in the ancestors of Old World monkeys and apes; and in only one New World monkey, the howler monkey. Interestingly, the howler monkey is also the only New World monkey to possess full trichromatic color vision, along with all Old World monkeys and apes. Hence, these

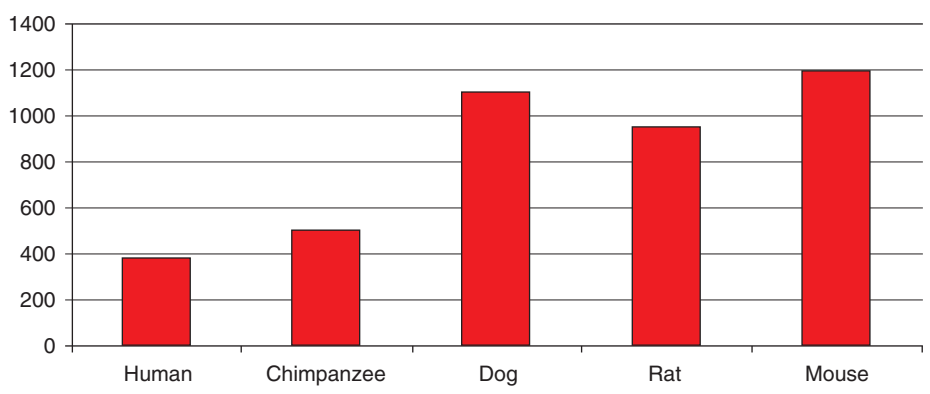

Figure 2 An estimate of the number of intact, putatively functional OR genes in mammalian species with available genomic sequence.

findings suggest that the deterioration of the functional OR gene repertoire in primates may have been associated with the acquisition of full trichromatic color vision.

Segregating Pseudogenes in Humans

The observation that OR pseudogenes have accumulated in the human lineage at a rapid rate suggested that this process is ongoing. In particular, approximately 70 of the hundreds of human OR pseudogenes have only one disruption in their coding region. These loci are likely to be the youngest human OR pseudogenes, as additional coding region disruptions have not had the time to accumulate. This suggests that the disruptive mutations in these loci may not yet have reached fixation in the entire human population; in other words, that both the intact and the pseudogene alleles are still expected to segregate in human populations.

To test this hypothesis, the single disruptive mutation in 50 human OR pseudogenes was genotyped in a sample of 192 individuals, half of whom were Europeans and half African Americans. Of the 50 loci, 26 were found to segregate both the intact and pseudogene alleles. Because these segregating mutations disrupt the OR coding regions, they represent promising candidates for functional polymorphisms that affect the sense of smell and may underlie, at least in part, the observed interindividual variation in olfactory sensitivity. For example, individuals that carry two OR pseudogene alleles at a given locus (i.e., homozygotes for the null allele) probably experience a loss of olfactory function, possibly specific anosmia. If this is the case, then when all 26 segregating pseudogenes are considered, each of the 192 individuals that were genotyped is inferred to have a different functional OR gene repertoire (Figure 3).

In that respect, it is interesting to note that human populations may vary in the mean size of the functional olfactory receptor gene repertoire. As an illustration, in the 26 segregating pseudogenes, the intact OR gene alleles tended to be at significantly higher frequency in African Americans compared to Europeans. Thus, on average, European individuals in the sample carry two null alleles more often than do the African American individuals. While intriguing, there currently exist no data with which to test the functional implications of this observation.

Positive Selection on Human OR Genes

The difference in the proportion of OR pseudogenes between different mammals suggests that the selective pressures on the OR gene repertoire vary among species. One extreme example are dolphins, which no longer require the ability to detect volatile odorants, and consequently lost all of their class II OR genes. The high proportion of OR pseudogenes in humans raises two related questions: (1) Does the rapid accumulation of OR pseudogenes in the human lineage suggests that humans do not need a keen sense of smell? (2) Are we in the process of losing all of our OR genes?

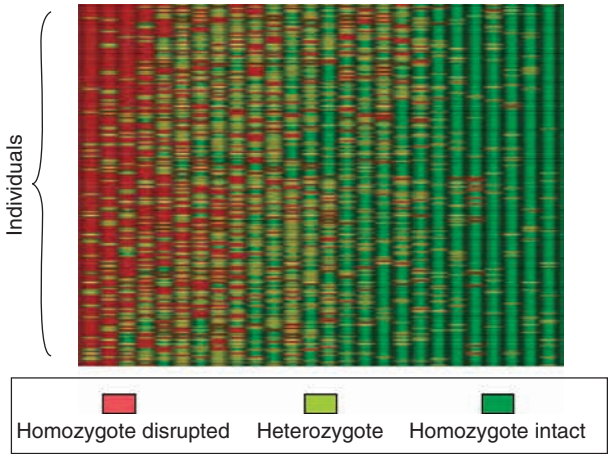

Figure 3 Menashe et al. genotyped the coding region disruption in 26 segregating pseudogenes from 192 human individuals. This is a representation of the genotype of the 192 individuals across the 26 OR loci. No two individuals have the same genotype. If loss of function is assumed for individuals who are homozygotes for a pseudogene allele, even by considering only these 26 OR loci, no two individuals have the same functional OR gene repertoire.

By examining patterns of within- and between-species variation at the DNA level, it is possible to infer the selective forces that shaped different genomic regions. The most popular approach is to estimate the ratio Dn/Ds, the ratio of the rate of nucleotide substitutions that changed amino acids (Dn) to the rate of silent nucleotide substitutions that did not change the amino acid (Ds). When both classes of substitutions are neutral, that is, are not affected by selection, then a Dn/Ds ratio of 1 is expected. In contrast, if mutations that replace the amino acid are selected against and only rarely are allowed to become fixed, while synonymous mutations are effectively neutral, then a lower Dn/Ds ratio is expected. Finally, a Dn/Ds larger than 1 reflects selection in favor of mutations that alter the amino acid, leading these to become preferentially fixed.

Using this approach, OR pseudogenes were inferred to have evolved under no selective constraint in both human and chimpanzee. This result is expected, as pseudogenes are assumed not to have a function and hence mutations in them are not expected to affect phenotype and be subject to natural

selection. In contrast, in chimpanzees, but not in humans, intact OR genes showed patterns of variability that were consistent with purifying selection. This finding is consistent with the observation that chimpanzee maintained a larger number of intact OR genes compared to humans. Indeed, based on the observation of a rapid accumulation of human OR pseudogenes, we expect human OR genes to evolve under less evolutionary constraint than OR genes in other apes.

Interestingly, several studies also reported that in humans, patterns of genetic variation at a subset of intact OR genes were consistent with the action of positive selection. This may seem surprising, given the relaxation of selection that seems to have occurred on the human lineage. However, these apparently contradictory observations can be reconciled if humans do not require most of their OR genes, but there exists a class of OR genes that have adapted in response to hominid-specific needs. Again, this could be caused by the pronounced differences in life style between humans and nonhuman apes. In addition to a greater reliance on vision and hearing which rendered the human sense of smell overall less important, such differences may have led to certain novel human olfactory needs, not shared with other primates. For example, humans are the only primates who consume cooked food, with potentially widespread effects on nutrition, ecology, and social relationships. This shift may have had a strong impact on the OR gene repertoire, since the sense of taste is largely a function of olfaction.

What Exactly Did We Lose?

The human decrease in the number of intact OR genes is assumed to result in reduced sense of smell compared to other mammals. However, when only the intact (putatively functional) OR gene repertoires of human and mouse are compared, it emerges that more than 160 (~65%) of the different OR gene subfamilies are shared by both species. If, as is commonly assumed, OR genes from the same subfamily bind the same type of odorants, this observation might suggest that, in spite of the much smaller functional repertoire of human relative to mouse, the repertoires of odors that can be sensed by the two species are similar.

Very little functional psychophysical data exist with which to test this hypothesis. In one of the only olfactory functional studies that involved primates, Laska et al. compared threshold sensitivities between several species for nine odorants, from four chemical groups. They showed that humans perform as well or better than the squirrel monkey (a New World monkey) and that, with respect to the odorants tested, humans also perform as well or better than dog and/or rat. Based on these observations, it was suggested that the human sense of smell may be as keen as that of macrosmatic mammals such as mice, dogs, and rats, typically thought to be highly dependent on their sense of smell. If true, there would appear to be a contradiction between the interspecies psychophysical data on one hand and the relative size of the OR gene repertoires on the other. One possibility is that interspecies differences in nonreceptor aspects of the olfactory mechanism (such as the size of the olfactory bulb, the size of the brain, and the nasal cavity) are responsible for differences in the sense of smell.

To account for the discrepancy between the relative sizes of OR gene repertoires and the psychophysical data as well as the observation of multiple modes of selection acting on human OR genes, we need a better understanding of the relationship between olfactory receptors and the binding of odorants. Unfortunately, almost all of the OR genes are orphan receptors, as very few ligand (odorant)–receptor interactions have been demonstrated for OR genes. Similarly, there exist few interspecies odorant detection comparisons. Future high-throughput approaches to study the binding capacity of OR genes may help addressing this issue.

Conclusions

It is unclear whether the size of the functional OR gene repertoire is the main determinant of how good a species' sense of smell is. Nevertheless, the reduction in the number of functional OR genes was found to be more rapid in mammalian species that rely less on their sense of smell. Humans have fewer intact OR genes than any studied mammalian species, and the process of OR gene loss in humans is still ongoing. Given this observation, it is reasonable to assume that humans reduced reliance on the sense of smell rendered many human OR genes unnecessary and therefore they are free to become pseudogenes. However, humans did not forfeit the sense of smell entirely, as there is evidence that a subset of OR genes evolved under positive selection in humans, presumably to respond to human-specific olfactory requirements.

See also: Olfactory Cortex: Comparative Anatomy; Olfactory Coding; Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Cortex Physiology.

Further Reading

Buck L and Axel R (1991) A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* 65: 175–187.

Firestein S (2001) How the olfactory system makes sense of scents. *Nature* 413: 211–218.

Gilad Y, Bustamante CD, Lancet D, and Paabo S (2003) Natural selection on the olfactory receptor gene family in humans and chimpanzees. *American Journal of Human Genetics* 73: 489–501.

Gilad Y, Man O, and Glusman G (2005) A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Research* 15: 224–230.

Gilad Y, Man O, Paabo S, and Lancet D (2003) Human specific loss of olfactory receptor genes. *Proceedings of the National Academy of Sciences of the United States of America* 100: 3324–3327.

Gilad Y, Wiebe V, Przeworski M, Lancet D, and Paabo S (2004) Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biology* 2: E5.

Laska M, Seibt A, and Weber A (2000) 'Microsmatic' primates revisited: Olfactory sensitivity in the squirrel monkey. *Chemical Senses* 25: 47–53.

Menashe I, Man O, Lancet D, and Gilad Y (2003) Different noses for different people. *Nature Genetics* 34: 143–144.

Mombaerts P (1999) Molecular biology of odorant receptors in vertebrates. *Reviews in the Neurosciences* 22: 487–509.

Shepherd GM (2004) The human sense of smell: Are we better than we think? *PLoS Biology* 2: E146.

Young JM and Trask BJ (2002) The sense of smell: Genomics of vertebrate odorant receptors. *Human Molecular Genetics* 11: 1153–1160.

Zhang X and Firestein S (2002) The olfactory receptor gene superfamily of the mouse. *Nature Neuroscience* 5: 124–133.

Olfactory Receptors

K Touhara, The University of Tokyo, Chiba, Japan

© 2009 Elsevier Ltd. All rights reserved.

Background

Based on evidence for the involvement of G-protein-mediated channel activation in olfactory signal transduction, Buck and Axel hypothesized that receptor proteins for odorants must be expressed in the olfactory epithelium and must belong to a family of seven-transmembrane G-protein-coupled receptors (GPCR). In the early 1990s state-of-the-art molecular biological techniques allowed them to identify a multigene family that appears to encode candidate receptors for odorants. The genes were named olfactory receptor or odorant receptor (OR) genes: olfactory receptor because the proteins are expressed in olfactory neurons and play a crucial role in olfaction, and odorant receptor because the proteins bind odorant molecules.

Structure

OR proteins are classified as part of the GPCR superfamily because of the presence of structural features common to GPCRs and because of their function in coupling and activating heterotrimeric G-proteins. ORs possess seven hydrophobic transmembrane domains, a disulfide bond between the conserved cysteines in the extracellular loops, a conserved glycosylation site in the N-terminal region, and several amino acid sequence patterns that are conserved in the OR family (Figure 1). These conserved motifs probably contribute to proper folding of ORs in the plasma membrane so that the ORs can function in binding odorants and in coupling to the appropriate G-proteins. In contrast, the transmembrane regions play a role in forming odorant-binding pockets; for this reason, the sequences are relatively variable, so that ORs can cover a wide variety of odorant molecules in the ligand recognition spectrum.

The size of the OR family was estimated to be at least several hundred when it was originally found in rat. In the past several years, whole-genome sequencing projects have allowed for comprehensive analysis of the OR gene family, revealing the genomic structure and distribution of OR genes in various organisms. The most recent data for vertebrate ORs are shown in Table 1. In mammals, the OR repertoire comprises 800∼1500 members, while fish have a relatively small OR family of ∼100. Thus, OR gene expansion occurred when animals shifted their living space from an aqueous to a terrestrial environment.

A significant portion of the OR gene family has been pseudogenized in vertebrates; hominoids have a high pseudogene content (∼50%), while the proportion of pseudogenes is ∼20% in mouse and dog and 25–35% in primates. The OR family was found to comprise the most changeable set of genes in the lineage of each animal species, undergoing rapid molecular evolution. The fraction of pseudogenes increased during evolution in the order rodents, monkeys, and humans, suggesting that reduced function of olfaction correlates with a loss of functional OR genes. Indeed, in whale and dolphin, in which the auditory system is dominant, 70–80% of OR genes appear to be pseudogenes.

The OR genes constitute an exon that contains an intronless coding region of ∼1 kb, and upstream exons whose number varies from one to four. The OR genes form genomic clusters and are distributed widely in chromosomes. The distribution, however, is not even; for example, almost half of mouse OR genes are located on chromosomes 2 and 7, which have orthologous relationships with chromosome 11 in humans. Further, OR genes are highly polymorphic, as suggested for human leukocyte antigen-linked OR genes and dog OR genes, possibly accounting for individual differences in odor perception ability and sometimes causing specific anosmia.

Expression Pattern

In rodents, OR genes start to be expressed at embryonic stages. The onset of expression in mouse was determined to be during mouse embryonic days E12–E14 using *in situ* probes for several ORs. Microarray analysis of the developmental course of OR gene expression revealed that the number of transcribed OR genes remains at a relatively low level until birth (100∼200 ORs), and a large number of OR genes appear to be detected only after birth; within 2–3 weeks postnatally, the number of detected OR genes reaches a peak (∼500 ORs). Recently, specific antibodies against ORs have become available. The OR proteins can be visualized in dendritic knobs of olfactory sensory neurons by specific antibodies as early as E12, before the initiation of ciliogenesis. As the olfactory cilia are formed at E13, the OR proteins appear to migrate into the developing cilia, and strong labeling by the antibody is observed in expanded tips of cilia. Up to E18, the number of cilia per neuron continues to increase, and OR immunostaining is clearly visible during this time.

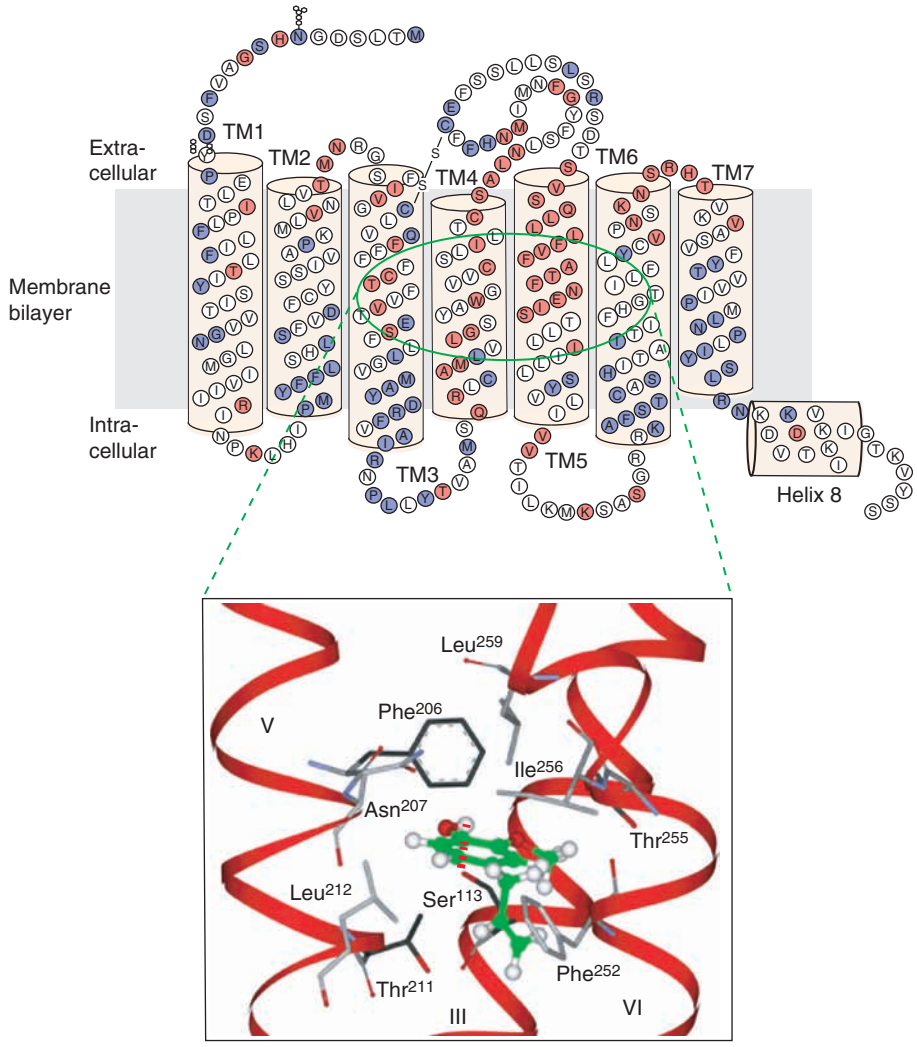

Figure 1 Structure and topology of an olfactory receptor protein, and a model for the odorant-binding mode. (Top) Schematic depicting components in the olfactory receptor family: red, variable amino acids; blue, conserved amino acids; TM, transmembrane domain. (Bottom) Schematic depicting (in green) the ligand, eugenol, bound to the space formed by transmembrane regions of a mouse olfactory receptor, mOR-EG; TMIII, TMV, and TMVI are shown as red ribbons.

Table 1 Number of functional olfactory receptor genes and proportion of pseudogenes in various vertebrate species

Organism	*Total number of genes*	*Number of intact genes*	*Pseudogene %*
Human	802	388	52
Mouse	1391	1037	25
Rat	1493	1202	20
Dog	1094	872	20
Frog	888	410	54
Zebra fish	133	98	26
Puffer fish	94	40	57

In situ hybridization experiments using different OR probes have suggested that each OR gene is expressed only in a certain region of the olfactory epithelium (Figure 2). Thus, the epithelium can be divided into at least four spatial zones in which different sets of OR genes are expressed. The distinct spatial expression pattern appears to be conserved among the vertebrate species, although expression of fish OR genes is not as clearly segregated as is that of mammalian OR genes. Initially, the spatial zones observed in rodents were thought to be sharply segregated, but careful *in situ* hybridization studies have revealed that each OR gene is expressed in a distinct zone in a continuous and overlapping manner. Therefore, multiple overlapping bands are formed along the dorsomedial/ventrolateral axis, although the boundary of the most dorsal region, corresponding to that of the olfactory cell adhesion molecule (OCAM)-positive neuronal layer, is maintained.

There are some exceptions to the zonal expression rule. First, some OR genes are expressed in the septal

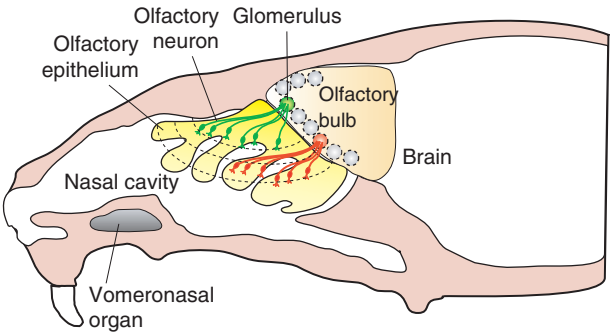

Figure 2 Anatomical structure of the mouse nasal cavity. The olfactory receptor genes are expressed in a unique and distinct zone in a continuous and overlapping manner, and olfactory sensory neurons expressing a given olfactory receptor project their axons to defined glomeruli in the olfactory bulb (shown by red and green neurons expressing distinct olfactory receptors).

organ, an isolated island of olfactory sensory neurons located on the nasal septum near the entrance of the nasopharynx. Second, the OR37 subfamily shows patch-like distribution in clustered populations of neurons in the central turbinates. Although there are some promoter motifs that may control the expression domain in the nasal cavity, the regulatory factors that determine positional information remain to be identified. Together, ORs appear to play a critical role in the development of olfactory sensory neurons for proper spatial and temporal neuronal network formation.

Single Receptor Choice

Each olfactory sensory neuron expresses just one OR from thousands in the OR repertoire. The one-neuron/one-receptor rule has been confirmed by a variety of techniques: *in situ* hybridization, single-cell reverse transcriptase polymerase chain reaction (RT-PCR) analysis, and transgenic experiments. Further, the selected OR appears to be transcribed from only one of two alleles, either the maternal or paternal allele. The mutually exclusive expression pattern is preserved even between transgenes and endogenous copies, supporting the stochastic model. Indeed, a single OR choice is observed from a minigene comprising a few kilobases upstream of the transcription start site, implicating *cis*-regulatory elements. In contrast, the expression of one cluster of OR genes, including MOR28 in transgenic mice, requires a long upstream region of ~100 kb containing an enhancer element, called the H-region. The expression levels differ up to 300-fold among ORs due to inequalities in the numbers of expressing cells and in cell transcript levels. In addition to positive regulation of OR gene expression by *cis*-regulatory elements, negative-feedback regulation must also be present to suppress the expression of additional receptors. It appears that functional OR proteins, although not mRNA, play an inhibitory role by preventing further activation of other OR genes, although the molecular mechanism remains to be elucidated.

Odorant Response Assays

To assess OR function, olfactory neurons were first targeted as an OR expression system, because they should possess the appropriate cellular machinery for OR expression and transmission of odorant signals. Adenovirus-mediated gene transfer was used for overexpression of ectopic ORs in the olfactory epithelium and the ligand responses were measured either by electro-olfactography (EOG) or by Ca^{2+} imaging of infected neurons. In addition, a gene-targeting approach was used to tag defined OR-expressing olfactory neurons using green fluorescent protein, and to record the responses of the fluorescent cells to cognate odorants. However, both virus-mediated gene transfer and the gene-targeting approach are tedious and involve much time and labor; therefore, the establishment of heterologous expression systems is still awaited.

It is not easy to express functional ORs in heterologous cells because ORs appear not to be efficiently translocated to the plasma membrane. In some cases, addition of an N-terminal leader sequence from another GPCR has resulted in limited expression of functional ORs in the plasma membrane, and in successful odorant-response recording in heterologous systems such as HEK293 cells. Thus, in HEK293 cells co-expressing tagged ORs and promiscuous G-protein, ($G\alpha15$), Ca^{2+} responses are observed when the cells are stimulated with their ligands. Without co-expression of $G\alpha15$, ORs activate endogenous $G\alpha s$ upon ligand stimulation, resulting in an increase in cAMP level in various mammalian cell lines (e.g., HEK293, COS-7, and CHO-K1 cells). A luciferase-reporter assay system using the zif268 promoter allows luminescent detection of cAMP increases upon stimulation with an odorant. *Xenopus laevis* oocyte is another good heterologous expression system for ORs.

None of these expression systems, however, is applicable to all ORs. To overcome this problem, a chaperone molecule which helps with membrane translocation of ORs has been investigated. A one-transmembrane receptor transporter protein, RTP1, was found to enhance cell surface expression of ORs, and some ORs have been deorphanized by co-expressing RTP1. In addition, Ric-8B, a putative guanine nucleotide exchange factor for $G\alpha_{olf}$, promotes efficient signal transduction of ORs. Thus, it

is thought that the problem can be solved if we can introduce all of the factors required for OR expression, which are present in olfactory neurons but lacking in the heterologous expression system. In addition, a methodology to allow stable retention of ORs in the cell surface should be developed, because ORs have been shown to undergo continuous internalization and recycling in heterologous cells. Thus, the level of surface expression still appears to be critical for efficient OR assay.

Pharmacology

The best-characterized OR is a mouse OR, mOR-EG (MOR174-9), which was originally isolated from a single eugenol (EG)-responsive neuron by Ca^{2+}-imaging and single-cell RT-PCR techniques. The mOR-EG recognizes 22 odorants that share certain molecular determinants, with EC_{50} values ranging from a few micromolars to several hundred micromolars (Figure 3(a)). Structure–activity relationship studies for mOR-EG suggest that ORs have a broad but selective molecular receptive range, and that selectivity is determined by the shape, size, and length of a ligand. This pharmacological concept is common to many ORs, as has been shown by studies on rat I7 and other deorphanized ORs. However, there seem to be some ORs whose ligand spectra are relatively narrowly tuned. Ligand specificity appears to be determined by the environment of the binding site in each OR. Considering that there are a thousand ORs, each of which has a unique ligand spectrum, and that each odorant is recognized by multiple ORs, it is reasonable to conclude that odorant information is coded by a combination of ORs activated by the odorant.

As is the case for other GPCRs, odorants function as agonists for ORs, but also as antagonists for other ORs, suggesting that the interaction between ORs and odorants in the mixture is sometimes more complex than expected. For example, the eugenol response of mOR-EG is potently blocked by certain structurally related odorants, such as methyl isoeugenol and isosafrol (Figure 3(b)). The antagonism is observed in mOR-EG-expressing olfactory neurons using intact olfactory epithelium slice preparation. Further, rat I7 has been shown to be antagonized by citral, and undecanal has been identified as an inhibitor for hOR17-4. Thus, OR antagonism between odorants may contribute significantly to formation of the receptor code for an odorant mixture in the olfactory system. However, the possibility of mixture interaction in real life may be low, because the chance of being an antagonist is not sufficiently high. Indeed, optical recording of the olfactory bulb shows that the receptor code for natural odors such as coffee and cumin is more or less the sum of the responses to its individual components. Quantitative recording using a Ca^{2+} imaging technique, however, demonstrates that partial inhibition of the response by antagonists is observed in the OR-defined glomerulus, which may contribute to a change in odor perception.

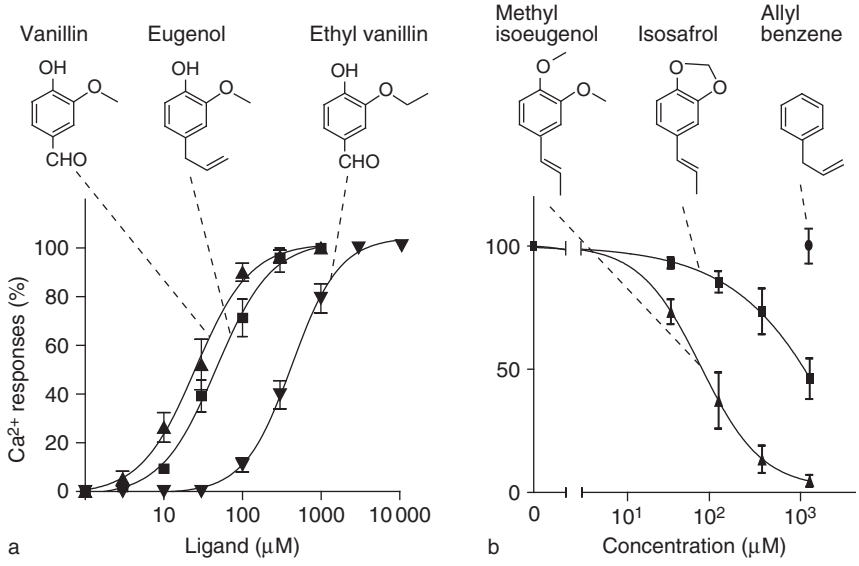

Figure 3 Pharmacology of an olfactory receptor expressed in HEK293 cells with Gα15. (a) Dose-dependent Ca^{2+} response curves of a mouse olfactory receptor, mOR-EG, responding to odorant agonists. (b) Dose-dependent inhibition curves of Ca^{2+} responses of a mouse olfactory receptor, mOR-EG, to eugenol by odorant antagonists.

Odorant-Binding Site

Multiple alignment analysis of the OR superfamily has allowed the determination of highly conserved and variable regions that are likely to be involved in structural organization and ligand recognition, respectively. Some of the earlier attempts, such as correlated mutation analysis and Fourier analysis of multiple OR sequences, supported the hypothesis that variable residues are responsible for odorant binding. An alternative method for identification of odorant-sensing residues is prediction of the binding site using a combined approach of computer modeling and docking simulation when cognate ligands for ORs are identified. Based on the coordinates of bacteriorhodopsin, a non-GPCR seven-transmembrane (TM) helix protein, and, more recently, on the X-ray structure and computationally determined three-dimensional structure of rhodopsin, using only the primary structure, computer modeling and calculation of lowest ligand-binding energy for several ORs has led to hypothesis that 10–20 residues in TM3, TM4, TM5, and TM6 are involved in ligand recognition and are responsible for odorant specificity.

A systematic experimental approach to decipher the odorant-binding site was taken for mOR-EG. Thus, functional analysis was carried out on a series of site-directed mutants, and ligand docking simulation studies were performed in parallel to define the odorant-binding environment. It was found that most of the critical residues involved in odorant recognition, and therefore sensitive to mutation, were hydrophobic, and that the binding pocket was in the space formed by TM3, TM5, and TM6 (Figure 1). The spatial location of the binding pocket is similar to that for other biogenic GPCRs; however, other GPCRs tend to form multiple electrostatic interactions with ligands, while hydrophobic amino acids appear to play critical roles in OR odorant recognition. Further, the accuracy of the binding model has been validated by the fact that single amino acid changes result in predictable changes in agonist and antagonist specificity. Thus, it can be concluded that the stereospecific receptor theory is proved by functional identification of odorant-binding site, and ligand information appears to be transduced by the three-dimensional configuration of the binding pocket and its specific odorant ligand.

Other Functions

Olfactory sensory neurons expressing a given OR project their axons to defined glomeruli in the olfactory bulb (Figure 2). Genetic manipulation that results either in deletion or in receptor swap suggests that the OR protein is instructive in axon guidance of olfactory neurons. Indeed, the OR proteins are present in olfactory axonal terminals as well as in cells of the cribriform mesenchyme involved in fasciculation and sorting of outgrowing axons. A point mutation in the OR protein results in formation of novel glomeruli, indicating that the OR amino acid sequence encodes information on axonal identity. The OR, however, is not a sole player in the axonal convergence; membrane receptors such as ephrin receptor, and cell adhesion molecules, are coordinatively involved in determination of the sensory map in the olfactory bulb. In addition, the instructive role in axon guidance is not restricted to ORs; a β_2-adrenergic receptor can substitute for an OR in glomerular formation, suggesting that GPCRs may have a common function. It should be noted, however, that both ORs and β_2-adrenergic receptors couple to the stimulatory G-proteins, Gαs. It is consistently found that OR-mediated Gα signaling is a positional determinant. Thus, the OR proteins expressed on the olfactory neuronal cilia and dendrites recognize odorants from the external environment, whereas the same OR proteins localized on axon termini appear to function as a critical factor in transmitting information for proper axon guidance.

Analogous to axon guidance, wherein olfactory neurons are attracted and navigated to a defined glomerulus by the ORs, testicular ORs are probably involved in proper navigation of sperm toward egg in mammals. Human OR17-4, expressed in sperm, recognizes bourgeonal, a floral aldehyde odorant, resulting in induction of chemotaxis and chemokinesis. MOR23, expressed in mouse sperm, recognizes lyral, again a floral aldehyde odorant, resulting in regulation of sperm motility. Since both bourgeonal and lyral are synthetic fragrances, identification of endogenous ligands for these ORs is necessary in order to elucidate the biological function of sperm ORs. Multiple testicular ORs have been shown to be expressed by various types of spermatids and spermatocytes in different developmental stages in mice, suggesting that some ORs are involved in spermatogenesis rather than in sperm physiology. In any event, ORs may function as a chemosensor in nonolfactory tissues.

Signal Transduction

The odorant signal received by ORs is converted to an electrical signal in olfactory sensory neurons (Figure 4). The first step in signal transduction is the activation of olfactory G-proteins, $G\alpha_{olf}$, by odorant-bound activated ORs. Some amino acids in

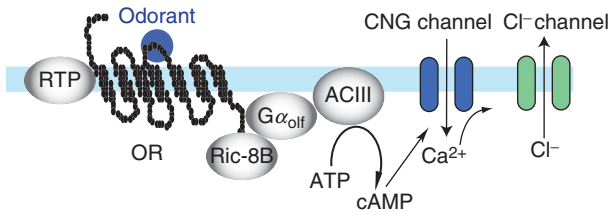

Figure 4 Signal transduction in olfactory sensory neurons. An olfactory receptor (OR), which is properly translocated to the plasma membrane by virtue of a chaperone receptor transporter protein (RTP), binds an odorant and couples to stimulatory G-protein ($G\alpha_{olf}$), which in turn activates type III adenylyl cyclase (ACIII), resulting in a cyclic adenosine monophosphate (cAMP) increase followed by an opening of cyclic nucleotide-gated (CNG) channel and Cl^- channel. Ric-8B, a guanine nucleotide exchange factor, enhances the G-protein-coupling efficiency.

the C-terminal domain and the third intracellular loop of ORs appear to be involved in coupling and activation of $G\alpha_{olf}$. Unlike in the case of rhodopsin, signal amplification does not occur in the OR–G-protein activation cycle. Instead, Ric-8B, a guanine nucleotide exchange factor, enhances accumulation of $G\alpha_{olf}$ at the cell membrane, thus improving the efficiency of OR coupling to $G\alpha_{olf}$. The activated $G\alpha_{olf}$ in turn stimulates adenylyl cyclase III, resulting in cAMP increase. The elevated cAMP interacts with and opens up the cyclic nucleotide-gated (CNG) channel, leading to cation influx and depolarization of the receptor neuron. Further, Ca^{2+} activates a Cl^- channel, leading to an efflux of Cl^-, which contributes to amplification of sensory depolarization. Bestrophin-2 is a candidate for a molecular component of the olfactory Ca^{2+}-activated Cl^- channel.

Gene knockout mice lacking either of the three molecular components (i.e., $G\alpha_{olf}$, adenylyl cyclase III, or CNG channel) failed to respond to odorants, suggesting that the cAMP cascade was dominant in transmitting the odorant signal in olfactory neurons. However, it was later demonstrated that CNGA2 knockout mice could detect a subset of odorants including 2-heptanone and 2,5-dimethylpyrazine, which are constituents of urine. Two additional olfactory pathways have been proposed: one involves membrane-bound guanylyl cyclase and phosphodiesterase type 2, and another involves phospholipase C and a transient receptor potential (TRP) channel. The multiple pathways in olfactory neurons may allow mice to respond readily to complex biological signals for social and sexual communication.

The activated olfactory neurons must return to the steady state, which is called desensitization. As is the case for other GPCRs, G-protein-coupled ORs could be phosphorylated upon odorant binding by protein kinases such as protein kinase A (PKA) and G-protein-coupled receptor kinase (GRK), resulting in desensitization. Indeed, knockout mice lacking GRK show slower recovery from cAMP increases, implying involvement of GRK in OR desensitization. In addition, it has been shown that PKA-mediated phosphorylation and subsequent internalization of ORs along with β-arrestin lead to reduction in the amount of ORs in the membrane. Although direct evidence for OR phosphorylation has not been obtained, desensitization mechanisms are likely to exist at the OR level. In addition, the influx of Ca^{2+} ions via the CNG channel downregulates itself, leading to closing of the CNG channel. Ca^{2+} also negatively regulates adenylyl cyclase activity. These Ca^{2+}-mediated negative feedback mechanisms allow activated olfactory neurons to revert to the steady state and prepare for the next stimulus.

See also: Electronic Nose; Olfactory Neuron Patterning and Specification; Olfactory Receptors: Microarray Analysis; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Olfactory Higher Centers Anatomy; Olfactory System Theory; Retronasal Olfaction; Sensory Aging: Chemical Senses.

Further Reading

Buck L and Axel R (1991) A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* 65: 175–187.

Firestein S (2004) A code in the nose. *Science's STKE* 227: pe15 [DOI: 10.1126/stke.2272004pe15].

Katada S, Hirokawa T, Oka Y, et al. (2005) Structural basis for a broad but selective ligand spectrum of a mouse olfactory receptor: Mapping the odorant-binding site. *Journal of Neuroscience* 25: 1806–1815.

Krautwurst D, Yau KW, and Reed RR (1998) Identification of ligands for olfactory receptors by functional expression of a receptor library. *Cell* 95: 917–926.

Malnic B, Hirono J, Sato T, et al. (1999) Combinatorial receptor codes for odors. *Cell* 96: 713–723.

Mombaerts P (2004) Genes and ligands for odorant, vomeronasal and taste receptors. *Nature Reviews Neuroscience* 5: 263–278.

Niimura Y and Nei M (2005) Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proceedings of the National Academy of Sciences of the United States of America* 102: 6039–6044.

Oka Y, Katada S, Omura M, et al. (2006) Functional odorant–receptor map in the mouse olfactory bulb: *In vivo* sensitivity and specificity. *Neuron* 52: 857–869.

Saito H, Kubota M, Roberts RW, et al. (2004) RTP family members induce functional expression of mammalian odorant receptors. *Cell* 119: 679–691.

Serizawa S, Miyamichi K, and Sakano H (2004) One neuron-one receptor rule in the mouse olfactory system. *Trends in Genetics* 20: 648–653.

Shepherd GM (1994) Discrimination of molecular signals by the olfactory receptor neuron. *Neuron* 13: 771–790.

Spehr M, Schwane K, Riffell JA, et al. (2006) Odorant receptors and olfactory-like signaling mechanisms in mammalian sperm. *Molecular and Cellular Endocrinology* 250: 128–136.

Touhara K, Sengoku S, Inaki K, et al. (1999) Functional identification and reconstitution of an odorant receptor in single olfactory neurons. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4040–4045.

Touhara K, Katada S, Nakagawa T, et al. (2006) Ligand screening of olfactory receptors. In: Haga T and Takeda S (eds.) *G Protein-Coupled Receptors: Structure, Function, and Ligand Screening*, pp. 85–109. Boca Raton, FL: CRC Press.

Touhara K (2002) Odor discrimination by G protein-coupled olfactory receptors. *Microscopy Research and Technique* 58: 135–141.

Zhao H, Ivic L, Otaki JM, et al. (1998) Functional expression of a mammalian odorant receptor. *Science* 279: 237–242.

Olfactory Receptors: Microarray Analysis

X Zhang and S Firestein, Columbia University, New York, NY, USA
X Zhang, NimbleGen Systems Inc., Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction to Olfactory Receptors

Organisms use their olfactory systems to detect and discriminate among thousands of low-mass molecules that reveal food sources, signal the presence of predators, and influence the organisms' social and sexual behaviors. Two olfactory systems have been developed in most animals. The main olfactory system senses environmental signals, which is the primary sense used by animals to find food, detect predators and prey, and recognize environmental cues. To enhance the ability of species and individuals to defeat their competitors, the olfactory system has evolved to have great discriminatory power and breadth, accomplished by a large family of olfactory receptors (ORs) residing in the nasal sensory epithelium. A second, or accessory, system developed to specialize in some species for finding a receptive mate. The sensory organ of this system is known as the vomeronasal organ (VNO). It is the sensor for species-specific olfactory signals, produced by one gender and perceived by the other, which play important roles in sexual behaviors and other social behaviors, such as territoriality, aggression, and suckling. These behaviors are initially executed by two types of receptors, known as vomeronasal receptors 1 and 2 (V1Rs and V2Rs), residing in the accessory olfactory system and binding with pheromones and related molecules. This review focuses on the ORs.

ORs were first discovered in 1991 by Linda Buck and Richard Axel, who were awarded the Nobel Prize in 2004 for their landmark work. As expected, OR genes turn out to be G-protein-coupled receptors (GPCRs), just like the receptors used in neurotransmission, photoreception, and many other cellular processes. Unexpectedly, the OR gene family size was found to be as large as 1000 in mammals, making it the largest gene family of GPCRs. The number of OR genes discovered was relatively small at the beginning because labor-intensive reverse transcription-polymerase chain reaction (RT-PCR) and rapid amplification of complementary DNA ends (RACE) studies were used to identify new OR genes. When the genome was sequenced, however, it became possible to computationally identify OR genes based on genomic sequences. From comprehensive data mining research in different genomes, it was found that rodents, such as mouse and rat, have in the range of 1000–1500 functional ORs, and that primates, such as human and chimp, have only around 350 functional ORs but a high percentage (~60%) of pseudogenes. Fish and amphibians are less well endowed, with about 100 ORs, and fly and other insects have about 60 ORs (Table 1).

Expression of ORs was first verified through traditional molecular methods such as RT-PCR. Subsequently, *in situ* experiments were carried out to examine the expression pattern and choice of ORs in olfactory sensory neurons. It was confirmed with single-cell RT-PCR that only one receptor gene is expressed in each sensory neuron, hence the so-called one neuron, one receptor rule. An advance was made through molecular-genetic studies, which showed that the H region in the 5′ upstream region of an OR cluster on chromosome 14 is involved with OR choice. OR expression also reveals spatial patterning; each receptor gene is expressed only in a subset of neurons that are randomly distributed in one of the several zones of olfactory epithelium (OE). In addition, it was reported that ORs are expressed in non-olfactory tissues, notably in testis. It is particularly interesting that at least one OR-like gene was found to function as a chemotaxis sensor in germ cells, directing their movement to the ovum.

Necessity of Microarray

Critical to the development of olfactory studies are genome-wide approaches. A family of more than 1000 genes could not be handled easily with standard molecular techniques. However, they are ideally suited to computational and high-throughput analysis. The protein coding regions of ORs, in common with many other GPCRs, are encoded by a single exon, which indicates that OR diversity could not be realized through gene rearrangement or alternative splicing, as occurs in the immune system. The fact that no exon–intron prediction is needed facilitates the data mining of receptor genes on the basis of similarity models. As the genome sequences become available, the OR repertoire has been examined computationally. However, it is not possible for molecular biologists to perform more than 1000 *in situ* experiments to verify the expression of each OR with regard to time and efficiency. Microarray technology is a good technique for studying OR expression as arrays allow the detection of many transcripts at the same time. Because of extensive similarity among paralogs,

Table 1 Total number of olfactory receptors in different organisms

Organism	*Number of putative functional genes*
Worm	1100[a]
Fruit fly	62
Fugu fish	47
Zebra fish	99
Chicken	347
Mouse	1191
Rat	1284
Canine	757
Chimp	411
Human	356

[a]This is actually the total number of sensory receptor genes for both smell and taste.

selecting unique probes for *in situ* study of ORs is an issue. One advantage of microarray design is that the probes on the chip do not need to be continuous in the original gene sequence; therefore, the regions with most cross-hybridization potential can be avoided in array design.

Standard commercial chips usually contain genes with known messenger RNA (mRNA) sequences and annotations from biological experiments, but few OR genes have mRNA sequences available. As a result, most OR genes are not represented on commercial chips. Furthermore, traditional gene expression design relies on 3′ ends of the mRNA sequences, which are largely unknown for OR genes. The limit of biological experiments and lack of OR genes on the commercial chips prompted the design of a custom chip specialized for ORs.

Design of OR Custom Array

As already discussed, an OR custom array is an excellent way to explore the expression of the whole OR repertoire. The Affymetrix Expression Array is based on reverse transcription (RT) of mRNAs with oligo (dT) primers. The RT reaction is 3′ biased, so only probes close to the 3′ end of the mRNA will detect signals in the array. Designing an array for genes with unknown mRNA sequences is challenging because the locations of the 3′ ends of the mRNAs are not known. The 3′ untranslated region (3′ UTR) downstream of the coding region for each OR was predicted by looking for polyadenylic acid (polyA) signals. Sequence analysis with the coding regions and the corresponding 3′ flanking sequences revealed that the latter are much more variable, which is an additional advantage to facilitate this unique probe selection. For each OR gene, polyA sites were predicted up to 5 kbp downstream of the coding region, and probe sets were designed closely upstream of the predicted polyA sites. To compensate for any inaccuracy of the prediction program, multiple probe sets were used per gene. Each probe set consisted of 11 probe pairs; each pair contained one perfect match probe and one mismatch probe. The mismatch probe, designed as the interior negative control, had almost the same nucleotide composition as the perfect match probe except for one mutation in the middle position of the 25-mer oligo. Besides 3′ UTRs, the coding regions of OR genes were also represented in the array. For OR genes with long 3′ UTRs, only the UTR probe sets would pick up signal, while the coding probe would have no signal. However, random-priming RT of mRNAs is an alternative to oligo(dT) primer-based RT. Samples from random-priming RT will have full-length coverage of the mRNA, and the coding probe sets will become useful.

Validation and Data Analysis of OR Custom Array

Custom high-density oligonucleotide arrays containing probes for all the putative ORs and VRs (V1R and V2R) were designed and manufactured in consultation with Affymetrix. Two generations of OR arrays were designed as the OR repertoire was updated, and more 3′ UTR sequences are available with the improved genome assembly. The improvement in the 3′ UTR sequences is important for detecting the expression of OR genes because the oligo(dT) priming process in the mRNA labeling requires the probes designed to be near the 3′ end of the transcripts. False negative results would be obtained for those ORs with long 3′ UTRs if probes for them were selected from the coding region. The custom arrays were produced with the same technology used extensively by Affymetrix in biological and medical applications, and subject to the same quality control.

Extensive tests were carried out to validate the custom arrays and optimize the experimental conditions and data analysis. Because each neuron expresses only one of more than 1000 ORs, each gene is present on average in about 0.1% of neurons, resulting in a very low expression level for each individual OR. To optimize the overall detection capability, the arrays were tested with different amounts of labeled cRNAs to find the appropriate point at which the number of detected genes reaches the maximum but noise level remains low. For data analysis of the GeneChip data, there are also multiple options. The dChip was chosen because of its flexibility in normalization matches with the property of the custom array. Other software may be better in probe-level

analysis, but distribution-based normalization may result in high false discovery rates. The perfect match–mismatch (PM-MM) difference model was used to calculate the expression level for each probe set, which was subject to differential analysis using Significance Analysis of Microarrays (SAM) software. SAM is based on adjusted *t*-statistics for small sample size data, producing an estimated false discovery rate through multiple permutations.

Expression of ORs in OE

Because the majority of sequences of ORs were obtained from computational prediction, it remains possible that some of them do not function as ORs or even not as a functional gene at all. Their specific expression in the sensory organ will confirm whether these genes are chemosensory receptors or not. The expression of genes was first compared between OE and the VNO, which are thought to mediate different types of environmental stimuli. It was found that most OR genes are highly enriched in the OE, compared with the VNO and other tissues, such as lung, liver, etc. From the first-generation mouse OR array, a total of 817 ORs were classified as enriched in OE, while about 1000 ORs were classified to be enriched in the OE with the second-generation mouse OR array. More ORs were detected in the second-generation array because multiple probes were designed for each gene when the correct polyA sites were unknown. Similarly, V1R and V2R genes are mostly specifically expressed in VNO, consistent with their roles as pheromone receptors.

The expression level varies dramatically among ORs. Since hybridization intensity could be affected by the nucleotide composition, the expression level used was calculated by GCRMA, in which the effect of GC content is compromised to some extent, for comparison among different ORs. ORs with extremely high identity in their protein sequences were found to have manifold differences in their mRNA expression level; similarly, ORs located next to each other in the genome can also have quite different expression levels. These data indicate that gene expression levels of ORs are regulated individually. The different levels of expression observed for very similar ORs suggest that the array approach was able to distinguish these genes, presumably by selecting unique probes in the 3′ UTRs. It should be noted that the data from microarrays are proportional to the multiplicative product of cell numbers for a certain OR and the mRNA transcripts per cell. The expression variance among ORs could result from the diverse number of sensory neurons for each receptor or different transcript copies or both.

OR Expression in Nonolfactory Tissues

OR gene expression has been reported in various tissues, particularly in testis, and also in brain, heart, and other tissues. These reports are generally isolated examples without systematic discovery. OR custom array is ideally suited to examine the level of OR expression in these tissues. Besides OE, OR mRNAs were also detected in the olfactory bulb, testis, kidney, heart, lung, and other tissues. The detected message in olfactory bulb could be from the transported mRNAs along axons. In total, about 270 ORs were found to be expressed in non-OE tissues, and about 100 of them are not enriched in OE. These genes may not be real ORs in terms of their functions, but just OR-like genes because of sequence similarity. So far their functions are rarely known, except that by analogy they may be receptors recognizing low molecular weight molecules in other tissues. Combing all the tissues studied, the expression of about 1100 ORs, which constituted 92% of the explored putatively functional ORs, was detected.

Temporal Pattern of OR Expression

Before the application of OR custom arrays, the developmental courses of most ORs have not been documented. Mouse OE tissues were collected at various ages, and the expression level of ORs was compared. There was observation of OR expression as early as embryonic day 12 (E12); however, there are only a few OR-positive cells, making the detection by microarray marginal. It was thought that ORs begin to be expressed in relatively large numbers of cells between E15 and E16, which was also the stage when significant OR expression was first observed with the array. From the custom array, it was found that the number of detected ORs and their expression level jumped after birth. In the 10-day-old mouse, the majority of ORs are already detected, while their expression levels continue to increase as the mouse matures. At 3–4 months, the average expression level reaches a maximum, then descends as the mice age. At 18 months, about 15% of ORs are no longer detected. At 24 months, about 40% of ORs are lost. The birth and loss of ORs are most likely attributable to the turnover of sensory neurons, though it is not known whether the transcript copies per cell also fluctuate temporally (**Figure 1**).

Zonal Expression of ORs

One of the earliest findings about ORs was their expression in one of the four parallel zones that run rostral to caudal across the turbinates within the OE.

To obtain zonal expression data of ORs, the OE in the dorsal recess above the turbinates, a region that only includes zone 1, was dissected; the ventral proportions of the turbinates were then dissected such that only zone 4 and parts of zones 2 and 3 were included (shown in Figure 2). The dissection left out the dorsal portion of the turbinates to ensure a clean separation between zone 1 and other zones. From the microarray data, ~350 ORs were significantly enriched in zone 1 and ~400 ORs were enriched in other zones. For the nonclassified ORs, it is likely because part of the epithelium was excluded during the microdissection. Among the zone 1 expressed ORs, 110 of them are class I ORs, and the rest of them are class II ORs. However, among the ORs expressed in zones 2–4, none of them belong to class I. Since class I ORs are segregated from class II ORs on the chromosomes, this finding indicates that ORs segregated in the genome are also segregated spatially in the OE.

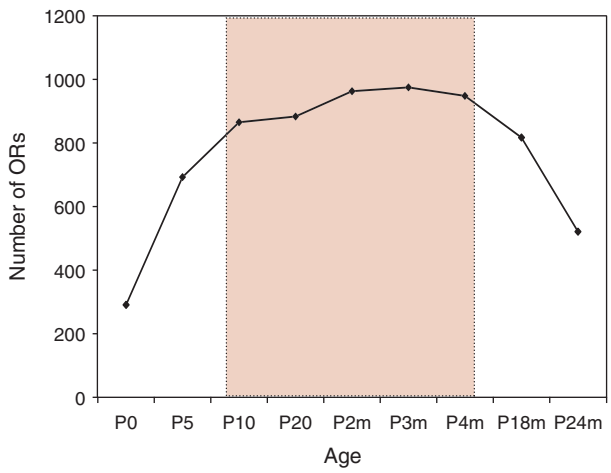

Figure 1 The number of expressed olfactory receptors (ORs) at different developmental stages. Four samples per time point were used for differential analysis compared to the vomeronasal organ. From postnatal day (P) 10 to postnatal month 4 (P4m) (shaded pink), the total number of expressed ORs is quite flat, whereas OR birth and loss happen at an early postnatal age and at old age, respectively.

3′ UTR

In this strategy, probe sets showing a specific expression signal should demonstrate the correct position of polyA sites, by which the 3′ UTR sequences of ORs were inferred. Multiple probe sets for different poly-A site predictions were put in the design, and the experimental data showed which of the predictions were correct. From the analysis, it was found that the length of 3′ UTRs varies among ORs. In fact, 40% of them have 3′ UTR shorter than 500 bp, while a few are longer than 2 kbp. The sequences of 3′ UTRs are much more diverse than the corresponding coding regions; therefore, they are perfect resources for unique probes selected for *in situ*.

The 3′ UTRs have been shown to be important regions for gene regulation; they are involved in mRNA stability, silencing, or activation mediated by microRNA or RNA-binding proteins, and so forth.

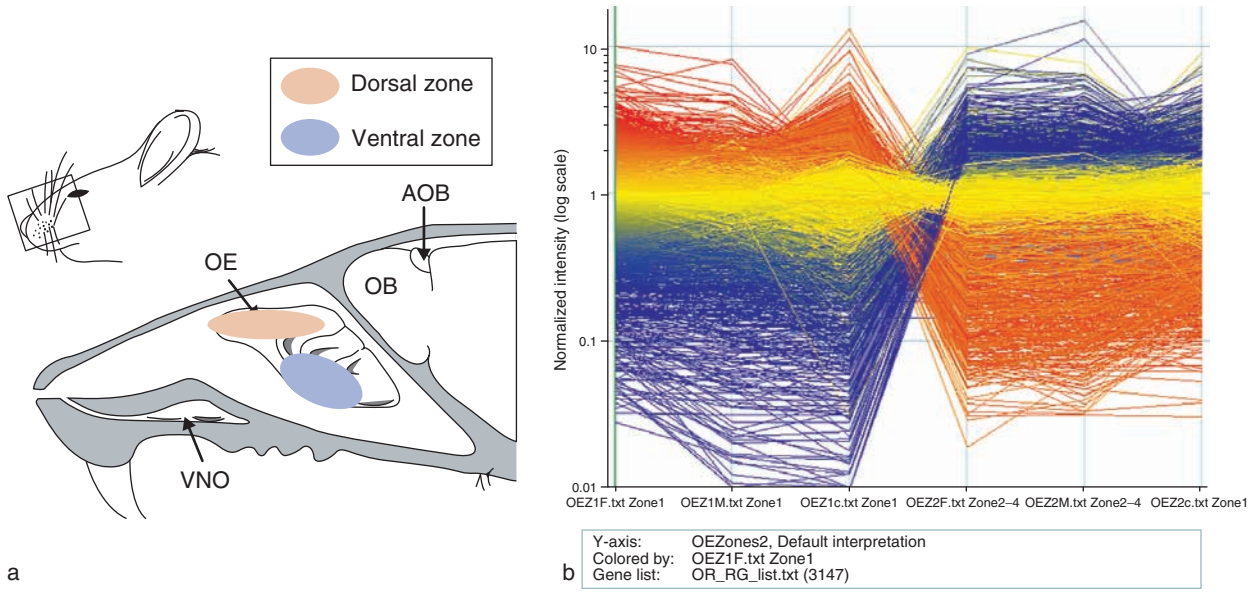

Figure 2 Olfactory receptor (OR) genes are differentially expressed in different zones of olfactory epithelium (OE). (a) Diagram showing the microdissection of the OE. Zone 1 (dorsal) is shaded in orange and zones 2–4 (ventral) in light blue. AOB, accessory olfactory bulb; OB, olfactory bulb; VNO, vomeronasal organ. (b) Differential expression of OR genes in zone 1 and zones 2–4. Each line represents the expression level of one OR. Orange lines represent ORs enriched in zone 1, and blue lines represent ORs enriched in zones 2–4.

So far very little is known about the function of 3′ UTRs of OR genes. The array data characterize the 3′ UTRs of a large number of OR genes, and this is a useful resource for future research on posttranscriptional regulation.

Future Perspectives for OR Microarrays

Microarrays have been used to study mRNA expression of ORs in different tissues, developmental time courses, and zones of OE, and to infer the 3′ UTR sequences for the ORs. However, the exploration of 5′ noncoding exons of OR genes has not been done yet. The transcription start site of OR genes may shed light on the regulation mechanisms. Current molecular techniques, such as RNA ligase-mediated-RACE, will be too time consuming to explore the 5′ exons for all OR genes. Tiling arrays would be a good choice to investigate this question. In this array design, probes would be tiled densely across the complete OR gene loci. Random primer-based reverse transcription would be used as the first step to amplify OR mRNAs to get complete coverage of the transcripts. Such experiments would generate the complete gene structure for OR genes. One could also use tiling arrays to explore the transcription factor binding sites by doing chromatin immunoprecipitation coupled with microarray and therefore generate data to decipher how OR genes are transcriptionally regulated. In addition, from computational analysis, many single-nucleotide polymorphisms (SNPs) were found in ORs and vomeronasal receptors between different strains. SNP arrays or allele-specific arrays could be used to associate genetic variance with variable odor perception.

Summary

Although the complete family of OR genes has been identified by genomic analysis, subsequent functional study of all these genes remains a challenge. OR microarrays have enabled high-throughput gene expression study for this large gene family, both temporally and spatially. The expression profiling results have provided important insights about the development and organization of the olfactory system. As cutting-edge technology, microarray-based tools will certainly play more important roles in the future study of neuroscience.

See also: Chemical Senses: Overview; Olfactory Receptors; Olfactory Epithelium; Olfactory System Theory; Vomeronasal System Evolution; Vomeronasal Accessory System.

Further Reading

Buck L and Axel R (1991) A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* 65: 175–187.

Firestein S (2001) How the olfactory system makes sense of scents. *Nature* 413: 211–218.

Firestein S (2004) A code in the nose. *Science's STKE: Signal Transduction Knowledge Environment:* pe15.

Malnic B, Hirono J, Sato T, and Buck LB (1999) Combinatorial receptor codes for odors. *Cell* 96: 713–723.

Mombaerts P (1999) Molecular biology of odorant receptors in vertebrates. *Annual Review of Neuroscience* 22: 487–509.

Parmentier M, Libert F, Schurmans S, et al. (1992) Expression of members of the putative olfactory receptor gene family in mammalian germ cells. *Nature* 355: 453–455.

Ressler KJ, Sullivan SL, and Buck LB (1993) A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell* 73: 597–609.

Serizawa S, Miyamichi K, Nakatani H, et al. (2003) Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse. *Science* 302: 2088–2094.

Tusher VG, Tibshirani R, and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 98: 5116–5121.

Vosshall LB, Amrein H, Morozov PS, Rzhetsky A, and Axel R (1999) A spatial map of olfactory receptor expression in the Drosophila antenna. *Cell* 96: 725–736.

Zhang X and Firestein S (2002) The olfactory receptor gene superfamily of the mouse. *Nature Neuroscience* 5: 124–133.

Zhang X, Rogers M, Tian H, et al. (2004) High-throughput microarray detection of olfactory receptor gene expression in the mouse. *Proceedings of the National Academy of Sciences of the United States of America* 101: 14168–14173.

Olfactory System Theory

T A Cleland and C Linster, Cornell University, Ithaca, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The most difficult challenges in understanding sensory systems arise in the context of natural scenes. The problems inherent in the sampling, representation, and processing of sensory information within natural environments are substantial; some would even be considered crippling were it not already clear that functioning sensory systems do exist. Hence, understanding these processes as complex systems not only will yield insight into neural processing *per se*, but also offers the hope of fundamental new insights into stimulus processing for engineering and machine sensation, particularly in areas such as source separation, object construction and identification, and the utility of prior expectations. Given recent insights into the molecular and biochemical events underlying fundamental sensory transduction mechanisms, the complex neural algorithms underlying postacquisition stimulus processing are increasingly the limiting factors in understanding sensory system function. Studying sensory processing mechanisms as complex systems is a critical component of modern neuroscience.

The olfactory modality is particularly amenable to study at the systems level. While many of the difficulties imposed by olfactory natural scenes are analogous to those encountered in other sensory modalities, some of the resulting computational problems are considerably more tractable. For example, there is no clear olfactory analog to the problem in vision that a given visual object does not reliably generate a replicable primary representation – that is, the pattern of activation of retinal photoreceptors in response to a given visual stimulus object differs arbitrarily based upon both eye position and the spatial orientation of the object. In contrast, while the breadths of olfactory primary representations are affected by changes in stimulus intensity and other environmental factors, odor-evoked activation patterns among primary olfactory sensory neurons (OSNs) are fundamentally specific to odor quality and hence are reliable as channel-based primary representations, potentially rendering the olfactory binding problem a considerably more tractable goal than is its visual counterpart. In contrast, other elements of olfactory stimulus processing are more complex than are their analogs in other modalities. The regulation of receptive field breadth, for example, is classically mediated by lateral inhibitory interactions in modalities such as visual retinotopy or auditory frequency tuning, in which stimulus similarities can be mapped onto physical neighborhood relationships in the brain. In olfaction, however, the similarities among stimuli are mapped high dimensionally, rendering lateral interactions in functionally two-dimensional cortical networks incapable of mediating this basic sensory filtering process. Novel neural mechanisms are required to perform similarity-dependent transformations on high-dimensional sensory information.

While a comprehensive, systems-level analysis of olfactory processing that engages the problems posed by natural scenes becomes considerably more complex than are analyses limited to interpreting the results of highly constrained experimental studies, it also provides vital constraints on functional hypotheses that otherwise would go unrecognized. A robust neural system engages in numerous homeostatic tasks unrelated to stimulus coding *per se* – for example, pH regulation, energy conservation, prevention of excitotoxicity, maintenance of the appropriate dynamic ranges of individual neurons and networks, and compensatory responses to durable changes in synaptic weights or the incorporation of newly generated neurons that may be necessary for network stability. Such processes may correlate with sensory inputs and hence can easily be interpreted as contributing to 'stimulus coding' without casting much light on the dynamics of sensory system operation. Comprehensive, systems-level study of neural circuit operations in complex environments is required in order to come to understand the broader biological networks that underlie robust, functioning sensory systems *in vivo*.

Systems Analysis of the Olfactory Modality

In natural environments, airborne chemical stimuli are distributed unpredictably in time and space, and odorants from innumerable sources intermix freely. Furthermore, the mixtures of volatile molecules that define meaningful odors in an animal's environment are inherently variable both in quality and in intensity, depending on numerous external factors, including time of day, season, temperature, weather conditions, and the proximity of the animal to various odorant sources. An effective olfactory system must be able to detect potential signals of interest within these chemically noisy environments, correctly extract these signals from a complex and changing odor background

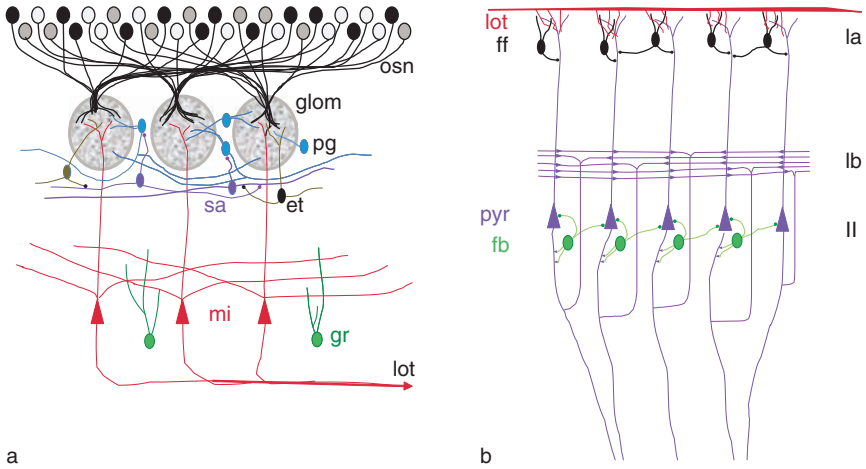

Figure 1 Schematic illustration of olfactory circuitry. (a) Olfactory sensory neurons and olfactory bulb circuitry. Olfactory sensory neurons (osn) expressing the same olfactory receptor project to a given glomerulus (glom) in the olfactory bulb, within which they make synaptic contacts with mitral cells (mi), periglomerular cells (pg), and external tufted cells (et). Periglomerular and mitral cells make reciprocal dendrodendritic synaptic contacts with each other within the glomerular neuropil: periglomerular cells inhibit mitral cells while mitral cells excite periglomerular cells. External tufted cells, together with short axon cells (sa), form a densely connected lateral excitatory network that spans the glomerular layer and they thus deliver excitatory inputs onto periglomerular cells. Deeper within the bulb, within the external plexiform layer, mitral cell secondary dendrites extend broadly and laterally, forming reciprocal dendrodendritic synapses with inhibitory granule cells (gr). Spike trains from mitral cells, the principal neurons of the olfactory bulb, are delivered to multiple secondary olfactory structures via the lateral olfactory tract (lot). (b) Circuitry of the piriform (olfactory) cortex. Within cortical layer Ia, mitral cell axons of the lateral olfactory tract (lot) excite both the primary apical dendrites of pyramidal cells (pyr) and the dendrites of feed-forward inhibitory interneurons (ff), which in turn locally inhibit the pyramidal cell dendrites. In layer Ib, the pyramidal cell dendrites also receive excitatory inputs from other pyramidal cells. These dense excitatory feedback connections are called association fibers and are known to undergo long-term potentiation. Pyramidal cells also excite a second class of inhibitory local interneurons called feedback interneurons (fb), which in turn inhibit a small number of nearby pyramidal cells. The cell bodies of pyramidal cells and feedback interneurons comprise layer II of piriform cortex.

to form stimulus representations, identify these constructed representations with respect to previously experienced odors, differentiate relevant from irrelevant stimuli, and cue an appropriate response. Current theoretical and experimental approaches have tended to break this process down into sets of roughly sequential problems, such as the following sequence: (1) detect and transduce airborne chemical signals, (2) improve the signal-to-noise ratio, (3) modulate selectivity (contrast), (4) translate the representation from a form optimized for stimulus sampling to one compatible with common cortical processing mechanisms (generally considered to be sparse and spike timing sensitive), filter out multiple confusing sources of variance, including (5) differences in intensity (normalization) and (6) unrelated background odorants, and (7) identify relevant stimuli with reference to previous experience. The neural tissues involved in odor detection and recognition processes contain specialized circuitry to perform these essential tasks. Primary OSNs transduce chemical stimuli into electrical signals with high sensitivities and broad dynamic ranges that initially appear to defy mass action law. The axons of all OSNs expressing the same olfactory receptor converge upon one or two common neuropilar regions in the olfactory bulb input layer, termed glomeruli; the resulting high convergence ratios enable substantial improvement in the signal-to-noise ratios of olfactory sensory responses. Local inhibitory circuits within the olfactory bulb filter this primary representation, regulating stimulus selectivity and segregating odor quality information from response differences owing to different stimulus intensities. Other olfactory circuits are thought to utilize short-term synaptic plasticity for odor–background segmentation, whereas long-term plasticity and associative memory networks within the piriform cortex may contribute to the memorization and identification of stimuli sampled from noisy backgrounds. Extensive feedback interactions between the olfactory bulb and deeper olfactory cortices synchronize the activity of related circuits, providing a timebase for the representation and processing of olfactory information on short timescales (loosely termed 'temporal coding'). Here, we outline a functional description of olfactory signal processing emphasizing the functional roles of different neural circuits within olfactory structures (see Figure 1).

Stimulus Transduction and Signal-to-Noise Enhancement

Primary OSNs number in the millions in rodents. Specialized cellular properties within individual OSNs have been hypothesized to improve their sensitivities to

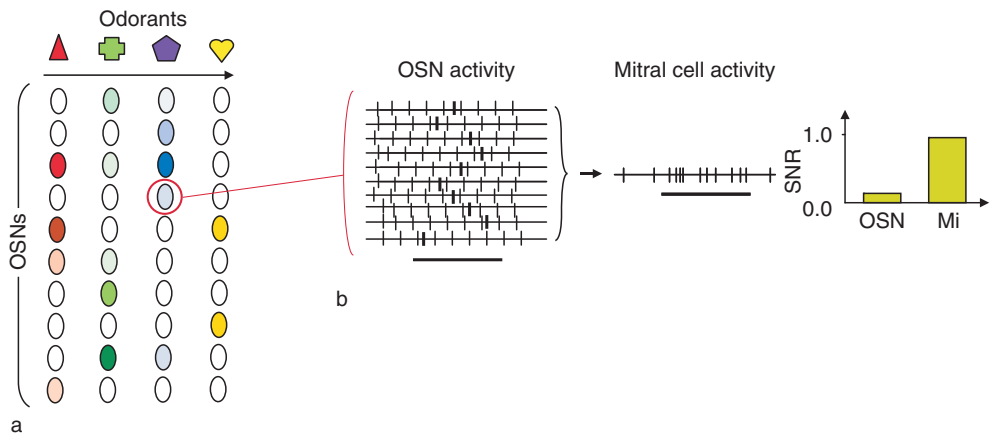

Figure 2 Combinatorial stimulus representation and signal-to-noise improvement. (a) Every odorant chemical binds to several types of olfactory receptors with different characteristic affinities, hence activating several different classes of olfactory sensory neurons (OSNs) to differing degrees (depicted by color saturation). Furthermore, the same OSN may be activated by many different odorants. Consequently, each chemical activates a unique combination of receptors with a characteristic pattern of activation levels. With only 10 OSNs, assuming only four discrete activation levels for each OSN (as depicted), 4^{10}, or 1 048 576, odors can be uniquely represented by the ensemble. Furthermore, combinatorial representations degrade gracefully; substantial information about odorant quality is preserved even if subpopulations of OSNs are destroyed or inactive. (b) OSN convergence improves the olfactory signal-to-noise ratio (SNR). In the present context, the signal-to-noise ratio can be defined as the number (or density) of spikes evoked by the signal of interest (odor presentation) divided by the total number (density) of spikes produced during the time of observation. Spikes specifically evoked in OSNs by odorant presentation are depicted by thicker vertical lines; thinner lines denote background spiking activity. Hundreds or thousands of OSNs with similar selectivity for odorants converge upon each mitral cell in the olfactory bulb. Even if the signal-to-noise ratio is extremely low in individual OSNs (i.e., if the stimulus-evoked increase in activity is nearly undetectable with respect to the OSN's baseline activity, as depicted), the convergence of many such neurons will evoke a substantial increase in the density of incoming spikes impacting a single mitral cell (Mi), reliably affecting its level of activation. This corresponds to a higher SNR in mitral cells when compared with any of its cognate OSNs.

odor ligands well beyond their ligand–receptor dissociation constants, while their large, redundant populations and correspondingly high convergence ratios onto the glomeruli of the olfactory bulb have been proposed to yield advantages such as improved signal-to-noise ratios and broader dynamic ranges. The molecular receptive ranges, or chemical receptive fields, of different odorant receptors overlap substantially, such that the identity of odorants is not associated with the activation of a specific receptor, but rather is represented by a distributed, combinatorial representation. From a theoretical point of view, this type of representation has several advantages: first, the number of unique odor representations is not limited to the number of different receptor types, but can be estimated as m^n, where n denotes the number of receptor types and m denotes the number of recognizable states that each sensor can assume, ultimately limited by the signal-to-noise ratio of the system. Second, the fact that structurally and perceptually similar odorant molecules activate correspondingly overlapping sets of olfactory receptors establishes a basis for the recognition of stimulus similarity in the olfactory system. A reliable metric for stimulus similarity facilitates a tolerance for variance among repeated stimulus samples and also is prerequisite for basic postsensory cognitive processes such as generalization. Combinatorial representations not only exhibit greater coding capacities than do labeled-line systems, but also are more robust to degradation, noise, and natural variation in stimulus parameters. Figure 2 illustrates the characteristics and advantages of combinatorial stimulus representation in the olfactory system.

Contrast Modulation and Normalization

Distributed patterns of activity in response to chemical stimuli are transmitted to the olfactory bulb via OSN axons that terminate in the glomeruli of its input layer. The olfactory bulb is believed to filter and transform these incoming sensory data, performing normalization, contrast enhancement, and similar operations before conveying the processed olfactory information to several different secondary olfactory structures via mitral cell axon collaterals. Because of the patchy and nontopographically organized representation of olfactory stimulus similarity across the olfactory bulb, contrast enhancement operations in this system cannot rely on physical proximity-based mechanisms such as center-surround lateral inhibition. (Indeed, olfactory stimulus similarity is not a static property; maps of olfactory similarity space depend on the statistics of the odor environment as well as on the structures of olfactory receptors expressed in the nose.) Instead,

the bulb appears to mediate an analogous operation relying solely on local feed-forward inhibition coupled with global, nonselective feedback inhibition. In addition to enabling contrast enhancement in a high-dimensional modality such as olfaction, this nontopographical mechanism does not require a built-in foreknowledge of the similarities in molecular receptive ranges expressed by different olfactory bulb glomeruli in order to distribute inhibition correctly, and is entirely independent of the physical location of glomeruli within the olfactory bulb. Briefly, OSN axonal arbors in each glomerulus deliver glutamatergic excitation upon mitral cell dendrites as well as the dendritic spines of local inhibitory neurons known as periglomerular cells. These periglomerular cell spines directly deliver γ-aminobutyric acid $(GABA)_A$ergic shunt inhibition onto mitral cell dendrites in parallel to the excitatory inputs that the latter receive from OSNs (Figure 3(a)). Theoretical models indicate that this configuration enables nontopographical contrast enhancement in a similarity space naturally inherited from the current chemosensory environment; mitral cells associated with a given odorant receptor are activated only by the highest affinity odor ligands, responding to lesser degrees of cognate OSN activation with a net inhibitory response (Figure 3(b)). Odor-evoked activity patterns across mitral cell ensembles (secondary representations) are consequently sparser and less overlapping than are the corresponding primary representations observed among OSNs, as visualized in imaging studies of glomerular activation patterns (Figure 3(c)). Furthermore, the stringency of

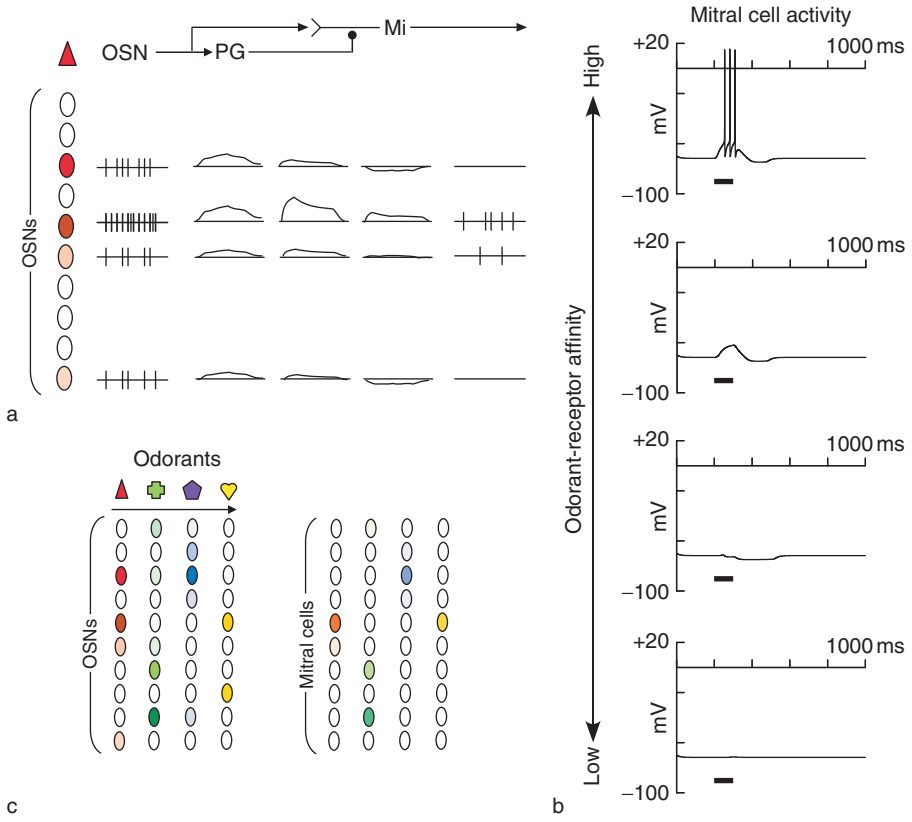

Figure 3 Local computation of contrast enhancement. (a) Odor presentation evokes a distributed pattern of activation among olfactory sensory neurons (OSNs), communicated to the olfactory bulb via convergent axonal projections. OSN activation excites both mitral (Mi) and periglomerular (PG) cells in the olfactory bulb input layer, while periglomerular cell dendritic spines concurrently inhibit mitral cell primary dendrites, resulting in a final activation pattern among mitral cells that is narrower and more precise than that originally evoked among OSNs. Corresponding to the schematic at the top are the five columns depicting neuronal activity (from left to right): spike trains in sensitive OSNs, synaptic activation of periglomerular cells by OSNs, synaptic activation of mitral cells by OSNs, net synaptic activation of mitral cells after factoring in periglomerular shunt inhibition, and evoked spiking activity in mitral cells. (b) Computational modeling of glomerular circuitry, demonstrating nontopographical contrast enhancement. Odorants with low affinity for the depicted mitral cell's cognate odorant receptor evoke no response in the mitral cell depicted, while odorants with moderate affinity activate OSNs but evoke a net inhibitory response in the corresponding mitral cell. Only the highest affinity odorants evoke action potentials in mitral cells. (c) These local computations result in distributed, odor-specific activity patterns at the level of mitral cells that are considerably sparser than those conveyed by the OSNs, thereby establishing greater contrast (less overlap) between the representations of structurally similar odorants. (b) Adapted from Cleland TA and Sethupathy P (2006) Non-topographical contrast enhancement in the olfactory bulb. *BMC Neuroscience* 7: 7, with permission.

this transformation appears to be regulated by centrifugal neuromodulatory inputs, specifically including inputs mediated by nicotinic cholinergic receptors in the glomerular layer.

In general, higher stimulus intensities both increase the activity of sensitive primary sensory neurons and broaden the neural response by recruiting additional, more weakly tuned primary sensory neurons into the activated ensemble. In the olfactory system, this corresponds to increased probabilities of ligand–receptor binding as ligand concentration rises, eventually binding significantly even to receptors for which the ligand in question has very low affinities. Hence, as odor concentrations rise, an increasingly broad range of OSNs become activated, substantially broadening the primary representation. In contrast, mitral cell activity in response to increasing odorant concentrations is variable; individual mitral cells may increase or reduce their response levels (to greater or lesser degrees), may exhibit shorter spike latencies, or may transition from a net excitation to inhibition, or vice versa. They do not, however, exhibit reliably monotonic increases in activation levels comparable to those observed in individual OSNs or populations of convergent OSNs (the latter observed via glomerular activity imaging). That is, some form of normalization process is clearly active between the primary representation across the OSN population and the secondary representation composed of mitral cell spiking activity patterns.

Normalization processes in sensory systems are essential for segregating quality from concentration effects and for constructing intensity-independent representations of stimulus quality. Indeed, identifying methods of forming intensity-independent representations is a major research concern in many sensory modalities. In principle, intensity normalization processes require global feedback inhibition, in which uniform inhibition is delivered to all units in proportion to the mean activity level of all sensory inputs. In the olfactory system, this normalization has been proposed to rely upon a lateral excitatory network formed by external tufted and short axon cells in the glomerular layer of the olfactory bulb. Briefly, this widespread, densely connected lateral network is excited by direct OSN excitation of external tufted cells within glomeruli. The lateral excitatory network integrates these heterogeneous activation levels across the bulbar input layer and delivers a uniform level of excitation onto a subclass of periglomerular cells, which in turn inhibit mitral cells. The result of this computation, manifested in the pattern of afferent input to mitral cells, is a preservation of relative activity levels among glomeruli independent of absolute stimulus intensity (Figure 4).

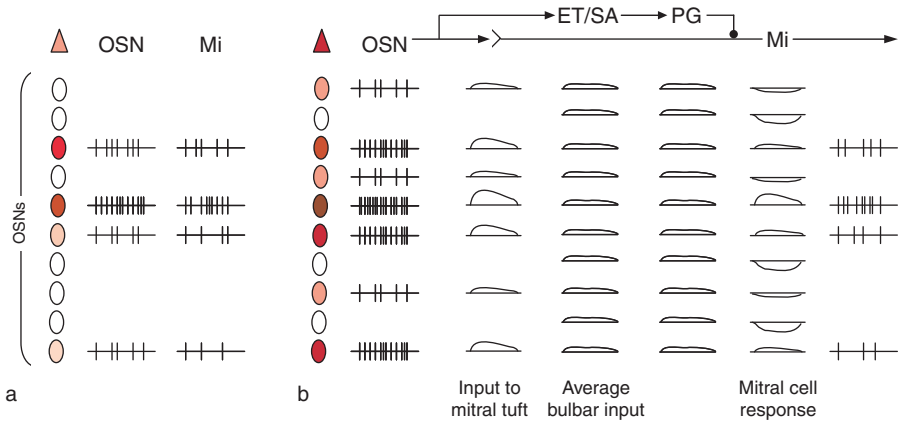

Figure 4 Normalization of odorant concentrations by glomerular-layer processing. (a) At the lowest odor concentrations, for sake of argument, the spatial pattern of activity among mitral cells (Mi) may largely reflect the pattern of activated glomeruli in turn reflecting presynaptic olfactory sensory neuron (OSN) activity (activation profiles of bulbar interneurons not shown). (b) When higher concentrations of the same odor are presented, increased ligand–receptor binding evokes both stronger activation of sensitive OSNs and the recruitment of additional OSNs expressing lower affinity receptors into the activated ensemble, broadening the primary representation (column 1, OSN). This intensified and broadened OSN activity evokes correspondingly broader and more intense excitatory synaptic input to mitral cell dendritic tufts (column 2) as well as to periglomerular cells (not depicted here; see Figure 3) and external tufted cells (ET). External tufted cells, along with short-axon cells (SA), are interconnected in a dense lateral excitatory feedback cellular network that is capable of integrating the input activity across all bulbar glomeruli and computing an average level of bulbar activation (column 3, ET/SA). This summed activity is transmitted to periglomerular (PG) cells (column 4), which oppose the direct excitation of mitral cells with shunt inhibition. The outcome of this computation is to subtract the mean bulbar activity from the activation at each glomerulus, resulting in a mitral cell representation that is normalized with respect to the mean (column 5, Mi). This normalization generates relatively concentration-independent representations of odor quality at the level of mitral cell output, presumably facilitating the recognition of odors irrespective of intensity (column 6; compare with a).

Synchronization and Binding

After the excitatory and inhibitory inputs converging on mitral cell primary dendrites have been integrated, olfactory stimuli are represented in broadly distributed levels of mitral cell activation across the olfactory bulb that largely determine whether any given mitral cell will evoke action potentials in response to a given odorant presentation. The timing of mitral cell spikes, however, is influenced by the intrinsic dynamics of the olfactory bulb as well as by the dynamics of feedback from other brain areas. Both intrinsic mitral cell membrane properties and network dynamics between mitral and granule cells are believed to influence mitral cell spike timing; indeed, synchronization of the spikes produced by those mitral cells most responsive to a given stimulus can further refine the representation of this stimulus and alter its capacity to drive activity in postsynaptic neurons (Figure 5(a)). Furthermore, due to activity-dependent synaptic plasticity between inhibitory granule cells and mitral cell lateral dendrites, the degree of synchronization as well as the identities of the neurons most tightly synchronized with one other can be regulated dynamically (Figure 5(b)). Indeed, mitral cell synchronization properties depend strongly on descending cortical and neuromodulatory inputs, suggesting additional mechanisms by which odor representations can be modified by centrifugal factors.

Odor–Background Segmentation and Mixture Interactions

Olfactory segmentation problems arise when two or more odors from different sources are present at the same time. The basis of the problem is that there is no *a priori* reason why the multitude of odorous molecules that together comprise one natural odor, such as the scent of oranges, should be processed as a single complex stimulus, while other odorous molecules, present at the same time, do not contribute to the perceived scent of the stimulus. The outcome is that an orange can be identified by scent largely irrespective of whether it is being smelled while standing on a beach,

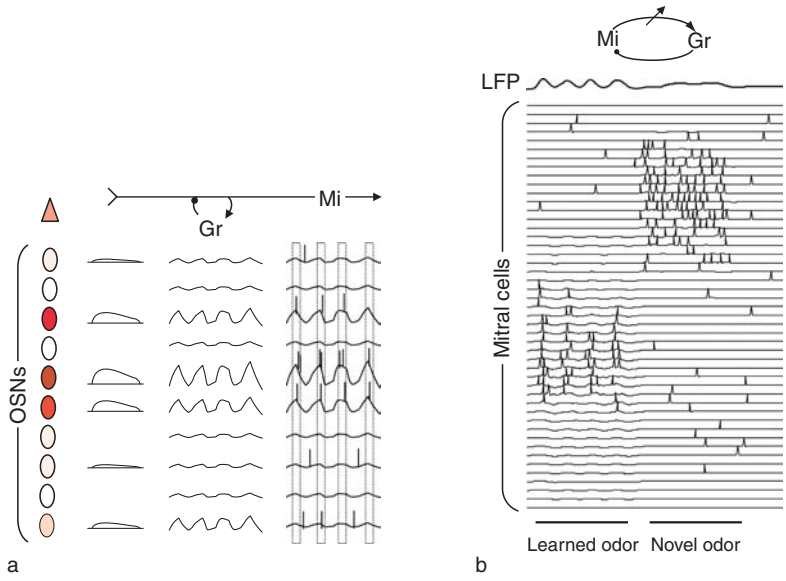

Figure 5 Synchronization of mitral cell firing. (a) The timing of mitral cell (Mi) action potentials is shaped by glomerular computations. Intrinsic subthreshold oscillations within mitral cells as well as recurrent synaptic interactions among mitral and granule cells (Gr) can regulate spike timing so as to regulate the synchronization of spiking activity among mitral cells. Mitral cells receive odor-specific patterns of afferent input excitation from olfactory sensory neurons (OSNs; column 1); activated mitral cells in turn activate granule cells. Due to the excitatory–inhibitory feedback interactions among mitral and granule cells, as well as between the olfactory bulb and higher olfactory cortices, rapid field potential oscillations in the olfactory bulb (column 2) are evoked by olfactory input as well as by the process of active sampling. The regular, synchronized delivery of inhibitory synaptic inputs onto mitral cell lateral dendrites interacts with the differing levels of input activity among mitral cells to shape the timing of evoked action potentials in mitral cells (column 3). Groups of activated mitral cells thereby form synchronized assemblies representing the odor stimulus presented. (b) Due to activity-dependent synaptic plasticity between mitral and granule cells, bulbar oscillatory dynamics in response to learned or behaviorally relevant odorants can be enhanced with respect to those evoked by novel odorants. Here, computational modeling output showing mitral cell response profiles to two different odorants (activating different subsets of mitral cells) is depicted. Owing to the pattern of altered synaptic weights between mitral and granule cells induced by prior learning, the odorant previously learned by the model olfactory system exhibits coherent oscillations in mitral cell membrane potentials and in the local field potential (LFP), as well as in more sharply synchronized mitral cell spikes. The novel odorant evokes comparable levels of mitral cell spiking, but this activity is less able to induce coherent oscillations and synchronize spike timing among mitral cells.

in a perfumer's shop, or on a factory floor. The core mechanistic problem is how the broad profile of activated receptors can be correctly deconstructed into groups or patterns associated with given stimulus sources, even when the arrays of odorant receptors sensitive to each stimulus source overlap substantially.

Odor segmentation problems can be classified into two categories: odor–background segmentation, in which a novel odor is detected and identified in the presence of a preexisting odorous background, and mixture segmentation, in which multiple odor stimuli are detected at roughly the same time. Recent electrophysiological studies have shown that odor–background segmentation can be achieved at the level of second-order sensory synapses. Synaptic interactions between olfactory bulb mitral cells and pyramidal cells in the piriform cortex undergo a form of short-term plasticity which results in the suppression of information transmission between these two structures after approximately 50 s when an odorant is continuously present. Consequently, piriform pyramidal cells rapidly suppress their responses to background odorants, but remain excitable in response to novel odorants (Figure 6(a)). Interestingly, this adaptation of pyramidal cell responses is specific to configural odors, rather than to odor elements, such that responses to the individual elements of a fully adapted odor mixture, if presented separately, are not suppressed to the same degree as is the response to the mixture (Figure 6(b)). This clearly suggests a regulation of synaptic suppression that is more complex than simple feed-forward desensitization, and that is deployed with reference to a preestablished configural odor representation, or odor template. This, of course, begs the question of how such configural odor representations are constructed and retained at this level.

Simultaneous-onset mixture interactions pose a substantially more complex computational problem. While mixture processing has been studied at the perceptual level using intramodal blocking, overshadowing, and comparable behavioral paradigms, to date there is no satisfactory model of how simultaneously presented odor mixtures can be segmented and identified. Computational efforts toward addressing this problem have relied upon built-in cues specifying the odor source to which a given response element belongs, avoiding the heart of the problem. Indeed, current theories suggest that this problem may not be solvable by feed-forward properties alone, but rather may rely critically on learned, preexisting odor templates. If correct, this suggests an essential role for associative memory and network

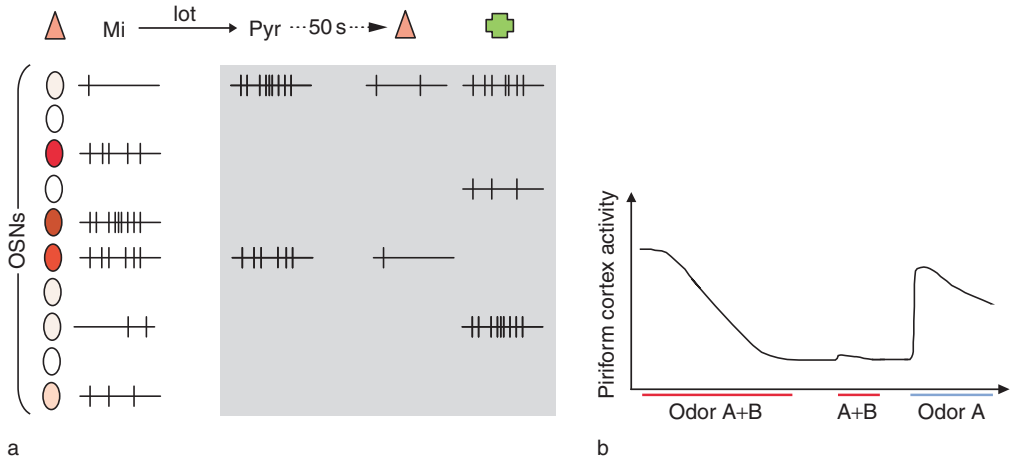

Figure 6 Odor–background segmentation in the piriform cortex. (a) Mitral cell (Mi) axons project from the olfactory bulb via the lateral olfactory tract (lot) to synapse with pyramidal cells (Pyr) in the piriform cortex. Piriform pyramidal cells and olfactory bulb mitral cells have receptive fields for odorants that are comparably broad. The excitatory synapse between mitral cells and piriform pyramidal cells undergoes short-term synaptic plasticity when an odorant is continuously presented for roughly 50 s, suppressing the pyramidal cell activity previously evoked by that odorant. Due to internal computations within the piriform cortex, this adaptation to prolonged odor stimulation is relatively specific to the broad, complex activity patterns evoked by configural odors rather than to each elemental chemical feature (odotope). Hence, even pyramidal cells that are fully adapted in response to a background odor still remain responsive to novel odors. No topological specificity is retained between mitral cells and piriform pyramidal cells; the alignments of pyramidal cell spike trains in shaded regions do not correspond to the rows illustrating connectivity between olfactory sensory neurons (OSNs) and their cognate mitral cells. (b) Prolonged presentation of a binary odorant mixture (A+B) leads to adaptation of evoked activity in piriform cortex due to suppression of mitral cell synaptic inputs. Subsequent presentations of that mixture elicit minimal responses. However, presentation of a single component (A) of this binary mixture again evokes increased activity, demonstrating that pyramidal cell adaptation is specific to complex, configural odors.

attractor dynamics, not only in olfactory learning as traditionally construed, but also in the very process of odor identification.

Odor Binding and Associative Memory

Binding among the multiple features of any given odor stimulus has been proposed to occur in the piriform cortex. Briefly, the ensemble of activated mitral cells that fire quasi-synchronously in response to a given odor stimulus activates a corresponding population of pyramidal cells. Whereas, at the level of the olfactory bulb, individual mitral cell responses predominantly reflect the elemental chemical features (odotopes) of odorant molecules, the combined representations of these features are integrated at the level of piriform cortex and are durably associated with one another via long-term synaptic plasticity across the pyramidal cell network. The feedback excitatory 'associative fibers' that mediate these patterns of association are prominent features of piriform cortex, along with feed-forward and feedback inhibitory pathways. Indeed, the piriform cortex has long been modeled as an autoassociative memory network capable of learning odorrepresentations and later reconstructing these learned representations from partial or degraded inputs, while also excluding unexpected or spurious additional inputs, a property known as pattern completion

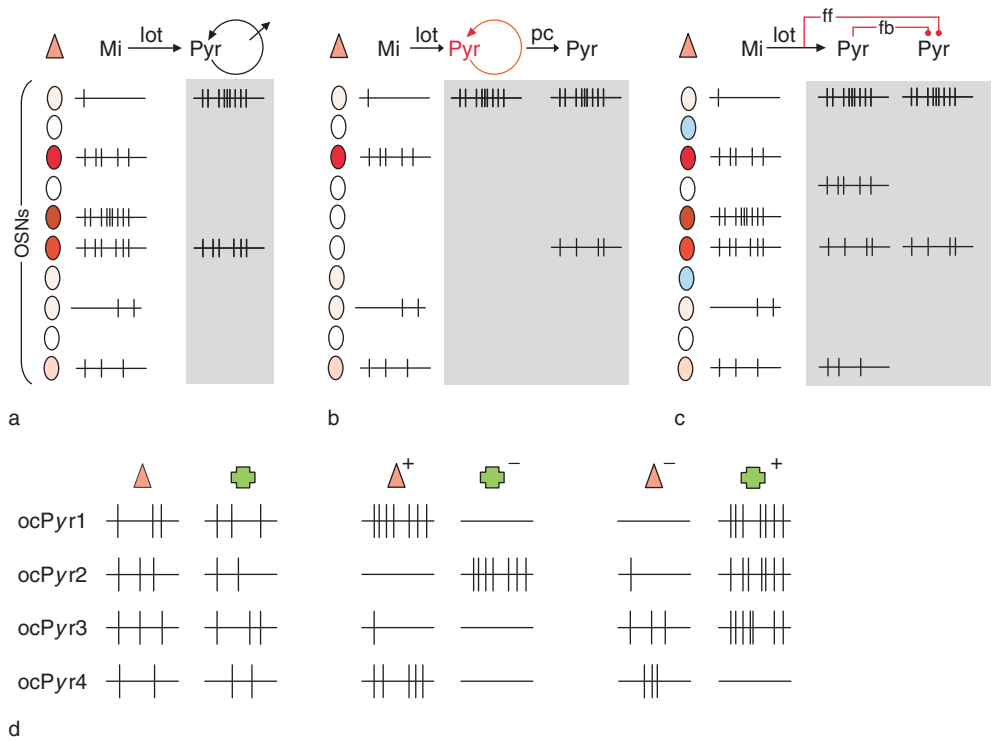

Figure 7 Associative memory and reward associations. (a) The highly interconnected association fiber network of the piriform cortex is thought to underlie learned associations between the elements of odor stimuli, generating and imprinting the patterned representations of complex odors. During the presentation of a rewarded odor stimulus, excitatory associative synapses from mitral cells (Mi), arriving via the lateral olfactory tract (lot), are strengthened among those piriform pyramidal cells (Pyr) responsive to that stimulus. No topological specificity is retained between mitral cells and piriform pyramidal cells; the alignments of pyramidal cell spike trains in shaded regions do not correspond to the rows illustrating connectivity between olfactory sensory neurons (OSNs) and their cognate mitral cells. (b) Upon subsequent presentation of the same odor, even if a subset of the normally activated pyramidal cells is not activated by the afferent inputs, associative circuitry within piriform cortex can perform pattern completion (pc), recreating the complete odor-specific activity pattern in pyramidal cells. (c) Other piriform circuitry incorporating feedback (fb) and feed-forward (ff) inhibitory interneurons can perform the complementary operation, reducing noise in the representation by suppressing pyramidal cell activity extraneous to the learned activity pattern. Both pattern completion and noise reduction are important for the robust comparison of noisy and variable stimulus representations with the memorized patterns of previously encountered stimuli. (d) Pyramidal neurons in orbitofrontal cortex (ocPyr) can be responsive to olfactory stimuli, but their responses are primarily dependent on learned associations and context. Most orbitofrontal pyramidal neurons change their responses to an odor stimulus depending on its predictive properties (e.g., for a reward, or even for the absence of a reward in a context where reward is associated with other odor cues). Weak responses are depicted in response to two novel odors (left-hand pair of columns). After both odorants are used in rewarded odor discrimination training, with one predicting reward (+) and the other predicting no reward (−), the responses of orbitofrontal pyramidal cells to either odorant are durably altered (middle pair of columns). After reversal training, in which each odorant is associated with the opposite contingency, orbitofrontal responses to these odors are again modified (right-hand pair of columns).

(Figures 7(a)–7(c)). While considerable work remains to be done in this area, it is likely that the circuitry of piriform cortex is essential to the construction of potentially meaningful, configural odor representations.

Forming discrete, identified sensory objects is traditionally considered necessary before such objects can be associated with a contingency such as reward or punishment. In the orbitofrontal cortex (a multimodal sensory cortex which receives direct input from the piriform cortex as well as from structures associated with other sensory modalities), neurons can be activated by odor presentation, but their response properties depend substantially less on physical stimulus attributes and more on the learned, expected contingency of the odor (Figure 7(d)). Hence, orbitofrontal odor responses may be better described as representing expectation, or the acquired meaning of an odor stimulus, rather than as coding for odor identity in the physicochemical sense. Reciprocal orbitofrontal projections to anterior piriform cortex further suggest that these learned, configural representations may be used dynamically to shape cortical responses to odors.

Learning, Experience, and Neuromodulation

The neural activity elicited by odorant stimulation depends increasingly on experience, behavioral relevance, and the motivational state of the animal as the representation progresses centrally from the primary sensory epithelium, factors that are thought to be mediated by centrifugal projections into these olfactory structures. Indeed, most olfactory areas interact with one another and with deeper, multimodal cortices via feed-forward and feedback connections, and also receive neuromodulatory inputs from the horizontal limb of the diagonal band of Broca (cholinergic), the locus coeruleus (noradrenergic), and the raphe nucleus (serotonergic). These cortical and neuromodulatory inputs dramatically influence neural activity patterns and the dynamics of olfactory processing within the olfactory cascade, and presumably mediate these experience- and state-dependent contributions to sensory representations.

Conclusion

The promise of the olfactory system as a model for understanding sensory processing as a complex system – from physical stimulus properties to cognitive object – is becoming increasingly clear. The peripheral processing of odorant stimuli in the olfactory epithelium and olfactory bulb is predominantly oriented toward adequately sampling the volatile chemical environment. Olfactory neuronal properties and bulbar circuit architecture are adapted to the physical and statistical properties of chemical stimuli and serve to mitigate the ambiguities and limitations inherent in the sensory transduction of stimuli in this modality. Impressively, orbitofrontal cortical representations of the acquired meaning of odors are only three synapses removed from the primary sensory representation, suggesting a singularly compact and tractable cascade of neural representations that may provide unique opportunities to understand how afferent sensory inputs are integrated with the neural representations of memories and behavioral state in order to create meaningful percepts. While the problem must not be oversimplified – for example, mitral cells also project to the cortical amygdala, the subcortical olfactory tubercle, and a series of medial cortices, and piriform and entorhinal cortical circuitry is considerably more intricate than discussed herein – the olfactory system remains a particularly attractive target for research in neural systems.

See also: Olfactory Receptors; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Orbitofrontal Cortex: Visual Functions.

Further Reading

Cleland TA and Linster C (2005) Computation in the olfactory system. *Chemical Senses* 30(9): 801–813.

Cleland TA and Sethupathy P (2006) Non-topographical contrast enhancement in the olfactory bulb. *BMC Nueroscience 7: 7*.

Doty RL (ed.) (2003) *Handbook of Olfaction and Gustation,* 2nd edn., New York: Dekker.

Linster C and Cleland TA (2002) Cholinergic modulation of sensory representations in the olfactory bulb. *Neural Networks* 15(4–6): 709–717.

Rolls ET (2001) The rules of formation of the olfactory representations found in the orbitofrontal cortex olfactory areas in primates. *Chemical Senses* 26(5): 595–604.

Rouby C, Schaal B, Dubois D, et al. (eds.) (2002) *Olfaction, Taste, and Cognition.* Cambridge, UK: Cambridge University Press.

Schoenfeld TA (ed.) (2006) What's in a sniff?: The contributions of odorant sampling to olfaction. *Chemical Senses* 31(2): special issue.

Shepherd GM (2003) *The Synaptic Organization of the Brain,* 5th edn., New York: Oxford University Press.

Wilson DA and Stevenson RJ (2006) *Learning to Smell: Olfactory Perception from Neurobiology to Behavior.* Baltimore: Johns Hopkins University Press.

Olfactory System: Circuit Dynamics and Neural Coding in the Locust

V Jayaraman and G Laurent, California Institute of Technology, Pasadena, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Sensory information needs to be converted from its 'raw' format (e.g., photons, molecules, and pressure) into a format that neurons in the brain can use for efficient internal computations. This format – which usually builds on the basic unit of neural communication, the action potential, or spike – is an important determinant of how the brain solves such problems as identifying and discriminating between stimuli, tracking time-varying stimuli in noisy environments, and remembering past encounters with a particular stimulus. However, such communication protocols or encoding strategies are difficult to study because they usually involve many neurons that together generate and use a so-called 'population code' or 'ensemble code.' These codes may depend, in ways not known *a priori*, on particular correlative aspects of activity in those neurons. In addition, the problem is usually underconstrained; that is, many different encoding and decoding strategies can be consistent with the limited sample of neural data collected by an experimenter. One approach to such issues is to record from as many neurons as possible and to decipher the encoding and decoding strategies used by the system itself. This is often difficult and far easier to achieve in smaller, simpler systems, such as the insect brain. Principles of circuit function extracted from simpler systems can then be tested on or applied to larger and more-complex brains as well.

Anatomy of the Locust Olfactory System

Insects are the most successful animal (metazoan) inhabitants of our planet and account for the majority of recorded species. A large part of their adaptive success must be correlated with remarkable sensory systems; indeed, insects occupy most ecological niches. Of all sensory systems, olfaction is considered the oldest – even bacteria have basic chemosensory capabilities. There are notable similarities between olfactory circuits across phyla, and insect olfactory circuits share structural and functional analogies with mammalian ones. This makes studying olfaction in insects all the more relevant to understanding the rules of olfactory coding in general.

Anatomy: Olfactory Receptor Neurons

As with other olfactory systems, information about an odor is first encoded across combinations of olfactory receptors (ORs). Different odorants interact with different subsets of ORs, and the activity of individual OR neurons (ORNs) is determined by the OR they express. Thus, a combinatorial activation of subsets of ORNs is the first odor-related electrical input to the brain. In both mammals and insects, the axons of ORNs expressing the same ORs converge to the same glomeruli in the olfactory bulb or antennal lobe (AL). There are ~90 000 ORNs in each locust antenna. Each ORN terminates in four to eight glomeruli organized in a semicircle. In the AL, ORNs contact two populations of neurons, the excitatory projection neurons (PNs) and the local neurons (LNs).

Anatomy: AL

Locust PNs, which, as in other insects, are the sole output of the AL, each extend dendrites to multiple glomeruli (each PN projects to about ten of about 1000 glomeruli in the AL). PN axons leave the AL in a common medial tract and project ipsilaterally to the mushroom body (MB) and to the lateral protocerebrum, or lateral horn (LH; see Figure 1(a)), where they make excitatory outputs mediated by acetylcholine. In the MB, the PN axon collaterals branch out profusely in a region known as the calyx, for its cuplike shape. The LNs are axonless, and each arborizes throughout the AL, influencing the spread of information across PN channels. In locust, LNs do not fire sodium-based action potentials; instead, they respond to odors with graded potentials and calcium spikelets. Transmitter release can occur both phasically, during a spikelet, and tonically, during maintained depolarization. LNs are γ-aminobutyric-acidergic (GABAergic) and inhibitory, although a significant fraction of them (up to 25%) may be cholinergic (as demonstrated in *Drosophila*).

Although all somata in the AL probably belong to PNs and LNs, the AL also contains the projections of a few extrinsic neurons. Most of these appear to be aminergic (i.e., they release biogenic amines, such as serotonin or octopamine) and are thought to be modulatory. One of these neurons is known as VUMmx1 in honeybees; it is octopaminergic and has a counterpart in locust and probably most other insect species. The AL also appears to be rich in the enzyme nitric oxide synthase (expressed by LNs). The gas neurotransmitter NO is therefore also a possible mediator of information processing in the AL.

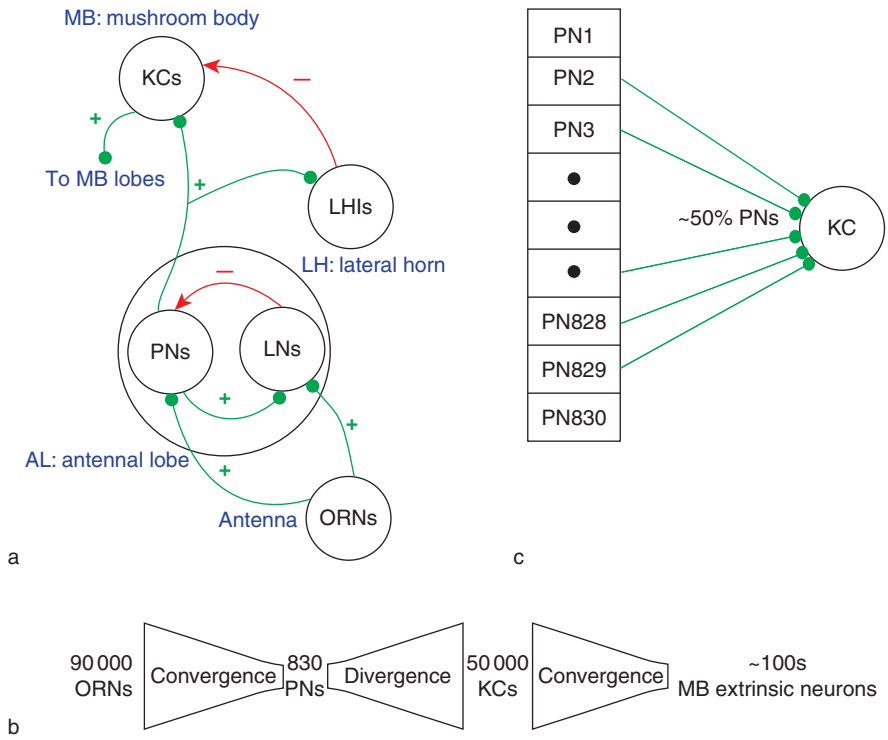

Figure 1 Architecture of the early olfactory system of the locust. (a) Odor information is carried into the antennal lobe (AL) by olfactory receptor neurons (ORNs). After processing by projection neurons (PNs) and local neurons (LNs) in the AL, information travels to the lateral horn (LH) and mushroom body (MB). MB Kenyon cells (KCs) then send information on to the MB output lobes (alpha and beta). (b) Information converges from a large number of ORNs onto a much smaller set of PNs. By contrast, PN output diverges to the KC layer, and KC output converges again onto a small number of MB extrinsic neurons. (c) PN–KC connectivity is dense, with each KC receiving input from close to 50% of the PNs in an AL. LHI, lateral horn interneuron.

Anatomy: MB

Several decades of behavioral and genetic experiments in fruit flies and honeybees have established that the MB is required for associative learning and recall. The intrinsic MB neurons are known as Kenyon cells (KCs). Their small cell bodies are packed tightly above the calyx (as grapes in a wine glass); the dendrites of each KC occupy a small fraction of the volume of the calyx. Whereas the MB has been primarily associated with olfactory learning, KCs may receive inputs from other sensory modalities as well. There is good anatomical evidence for this in the honeybee.

There are many more KCs than PNs: ~50 000 versus ~830 in locusts. Thus, the early olfactory circuit features numerical convergence (ORNs onto PNs) followed by divergence (PNs to KCs), as depicted in Figure 1(b). ORN-to-PN convergence is highly ordered; in *Drosophila*, the ORNs that converge onto one PN generally originate from ORNs of the same molecular type. This indicates that ORN space and PN space are of similar dimensionalities and suggests that ORN convergence could be useful for noise reduction by signal averaging. The connectivity between PN and KC populations is only beginning to be studied. It is extensive in locust: approximately 50% of the PNs in an AL appear to connect to any given KC (Figure 1(c)). Whether this convergence is ordered or predetermined (genetically or developmentally) is not known.

The next stages of the KC pathway are in the MB output lobes. KC axons leave the calyx in a dense tract called a peduncle. There is morphological evidence for reciprocal chemical synaptic connections between adjacent KCs in this axon bundle. After they process down this tract, KC axons bifurcate and enter output regions called the α and β lobes, where they synapse onto MB-extrinsic neurons; KC output is excitatory. The MB-extrinsic neurons are odor-responsive, and their morphology has been studied to some extent as well. Their precise connectivity to KCs and their role in olfactory circuits is only beginning to be understood. Among the most interesting recent findings about KC input to MB-extrinsic neurons is the existence, in locust, of Hebbian spike timing-dependent plasticity, with consequences for the regulation of spike time precision and output synchronization and obvious potential roles for learning and memory.

Anatomy: LH

Not much is known about this region of the insect brain. It is generally thought that the LH plays a role in the processing of innate rather than learned responses to odor (a role assigned to the MB). In locust, among the LH neurons is a subpopulation of ~60 GABAergic neurons, called LH interneurons (LHIs), with convergent input from the PN population. Although these neurons respond to odors, their responses are not odor-specific. LHIs in turn send axonal projections to the MB, thus acting as a feedforward inhibitory circuit onto KCs.

Circuit Dynamics and Ensemble Coding in the Locust Olfactory System

Oscillations are a prominent feature of brain dynamics across phyla and brain structures. They are typically detected in field potentials, generated by the summed activity (electrical and synaptic) of populations of neurons. Oscillatory field potentials were first reported in human electroencephalogram recordings by Hans Berger in 1929 and in local field potential (LFP) recordings in the olfactory system of hedgehogs by Edgar Adrian in 1942. Oscillations have been hypothesized to be involved in numerous phenomena, such as attention, perceptual binding, memory, information transfer between brain regions, signal amplification, and gating of sensory input during sleep. They have also been proposed, by others, to be meaningless artifacts and epiphenomena. Oscillations are common in olfactory systems (from mollusks to mammals), and their possible function in olfactory processing/perception has received considerable attention.

Some studies have successfully correlated the presence or absence of oscillations with behavior (discrimination between sensory inputs), but it is increasingly clear that demonstrating causal links between such disparate levels of description is very difficult. The approach in this article is to present the potential role of oscillations by exploiting the relative simplicity of the locust olfactory system.

Fast and Slow Dynamics in the AL

One of the most conspicuous circuit-level features of the AL is the recurrent connectivity between excitatory (cholinergic) PNs and inhibitory (GABAergic) LNs. Indeed, PNs make local outputs within the AL (probably from glomerular neurites and, so far as is known, only onto LNs) in addition to axonal output to the MB and LH. The interplay between the PNs and LNs, together with fast and slow kinetics of $GABA_A$ and $GABA_B$ receptors, gives rise to AL dynamics on two timescales. First, PNs respond to odors with slow temporal spike patterns that last more than a second and outlast the odor stimulus. The patterns include both excitatory and inhibitory epochs (Figure 2(a)). Second, active PNs transiently synchronize with each other during an odor presentation, due to fast $GABA_A$-mediated negative feedback via LNs. This transient pairwise, distributed synchronization gives rise to odor-evoked, collective ~20 Hz oscillations, detected in LFPs recorded in the MB – a consequence of summed PN synaptic output (Figure 2(b)). These odor-evoked oscillations are also visible in the subthreshold odor-evoked activity of PNs, LNs, and the recipients of PN output in the MB, the KCs. Mechanistically, $GABA_A$ conductances in PNs and LNs underlie the synchronous periodicity of PN output and thus the oscillations. Synchronization can be abolished relatively selectively by the application of picrotoxin, a Cl^- channel blocker, in the AL. This pharmacological manipulation has been used to test the potential functions of synchronization, using both behavior and electrophysiology as functional assays. Oscillation frequency is independent of the stimulus and appears to be determined in great part by inhibitory postsynaptic potential kinetics.

Ensemble Coding and Population Dynamics in the AL

The combination of slow temporal patterning and synchronization by phasic inhibitory feedback during odor responses (Figure 2(c)) has important consequences for the format of PN output sent to the next layer of olfactory processing. PN population output can be thought of as sequences of high-dimensional state vectors the width of an oscillation cycle (Figure 2(d)). Each dimension (that is, each row of the PN ensemble vectors shown in Figure 2(d)) captures the state of one PN and tracks the number of spikes fired by this PN in each oscillation cycle. These ~830-dimensional vectors (assuming one dimension for each PN in the AL, although the number of independent dimensions could be smaller if, for example, there are multiple PNs of identical type) are updated at each oscillation cycle. The dimensions define the coordinates of PN phase space within which the PN ensemble vectors exist. An individual odor response produces a progression of ensemble states and therefore a trajectory in PN phase space (Figure 2(e) shows the trajectory for PNs 1 and 2 from the red box in Figure 2(d)). The evolution of a particular odor trajectory depends on the sequence of specific combinations of transiently synchronous PNs; the rate of evolution of the

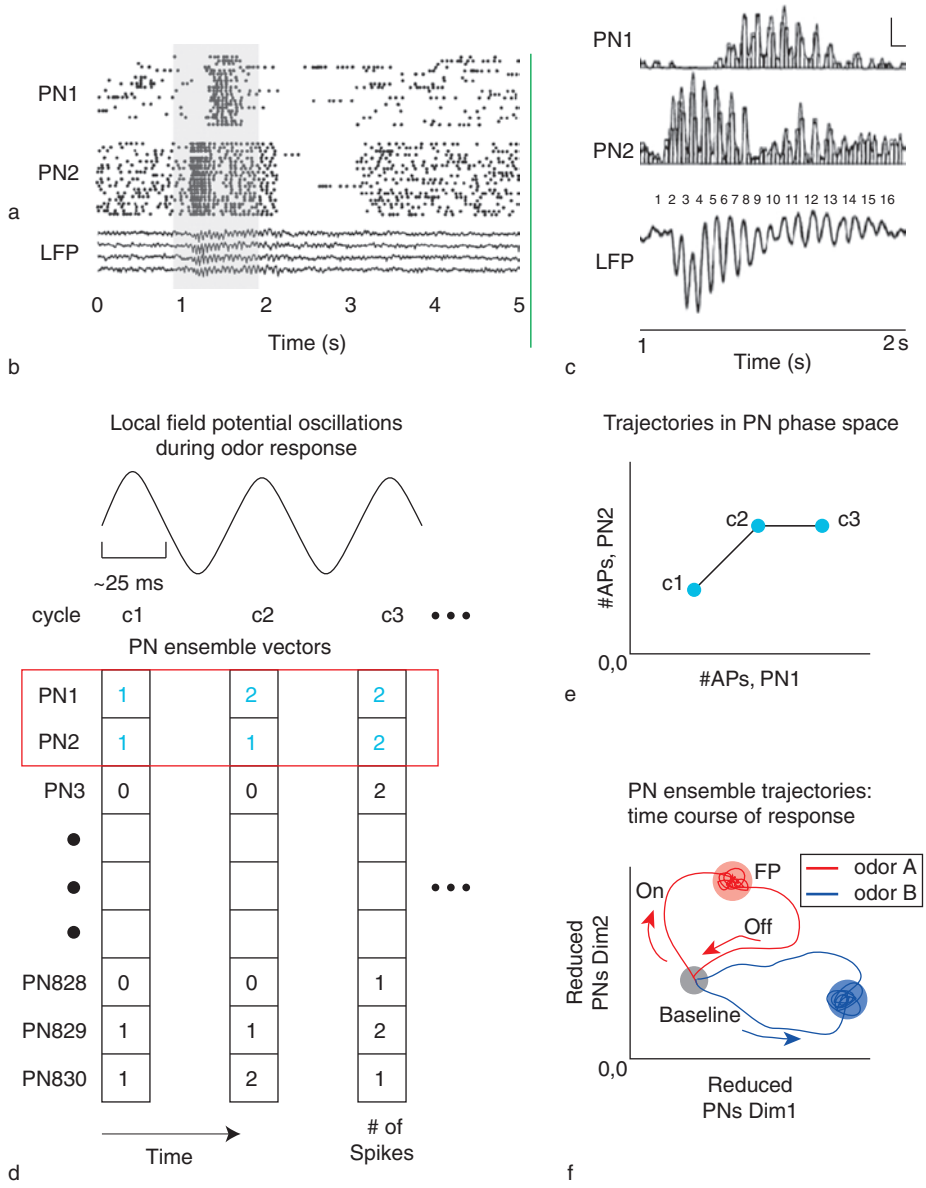

Figure 2 Dynamics of projection neuron (PN) ensemble odor responses. (a) Odor responses of two simultaneously recorded PNs. Each row of dots for PN1 represents a spike train recorded at the same time as the corresponding row for PN2. Shown are 21 successive trials with the same odor. Response patterns are highly reliable across trials. (b) Local field potential (LFP) recorded in the mushroom body (MB) shows ~20 Hz oscillations during the period of odor presentation. (c) PNs transiently synchronize their firing (thanks to fast local inhibitory feedback through γ-aminobutyricacidergic (GABAergic) local neurons (LNs) in the antennal lobe (AL)) in response to odors, producing ~20 Hz LFP oscillation that can be recorded in the MB. (d) Oscillations define a clock for this olfactory network. PN spike trains can be thought of as being formatted into bins the width of an oscillation cycle for the next layer in the network. The state of the PN ensemble is represented during each cycle by the 830-dimensional vector of spikes across PNs during that cycle. (e) The state of the PN ensemble (in this illustration, including only PNs in the red box in (d), PNs 1 and 2) can be represented on a phase plot in which each point represents the number of spikes produced by each neuron during one cycle. The entire response is then represented by the trajectory connecting the sequence of such points. (f) If more than two PNs are included in the ensemble, dimensionality reduction techniques (such as principal components analysis) are needed to visualize an approximation of the high-dimensional trajectories. Shown here is a schematic of PN ensemble trajectories to two different odors. The dynamics of the response are restricted to periods near odor onset and offset, with longer pulses producing periods of stationarity in the PN ensemble response (as represented by the fixed point, FP). AP, action potential.

trajectory is thus affected by both the slow and fast dynamics of individual PN responses. If the animal is exposed to a sustained odor, the slow temporal patterning usually ceases after ~2 s, at which time the PN ensemble reaches a fixed point: individual PNs are either stably active or silent until the odor is switched off. At that time, there is further patterning before a return to near-baseline levels of activity (Figure 2(f)).

The off-transient so defined is always different from the on-transient but is also odor-specific. The correlation width (the time on either side of a given cycle during which PN ensemble vectors are still significantly correlated) is on the order of 100 ms early in the response and can be many hundreds of milliseconds in the later stages of the response. It is interesting that the greatest pairwise distances between the trajectories defined by different odors and the lowest noise levels (correspondingly, the highest success in odor classification by observers) are reached early (within about 200 ms) in the on- and off-transient phases of the response rather than at the fixed points. The fixed points are thus not optimal for odor identification. Said differently, the optimal epochs for stimulus identification from PN activity vectors are when the PN population is in motion.

During a typical odor response (for a 1 s-long stimulus), more than half the PNs are active at some point at least during the response period. In a given oscillation cycle, however, only about 100–150 action potentials are produced. Of these, 55% are produced by a reliable PN (representing about 10% of the entire PN population). The remaining spikes originate from a pool of less reliable PNs (representing about 30% of the PN population). The remaining 60% of the PNs are reliably silent. Thus, the ensemble vector that represents the state of the PN population includes casts of reliably and unreliably firing PNs, whose identities change from oscillation cycle to oscillation cycle. This becomes relevant for how the PN vectors are decoded, which is discussed in the next section.

Decoding PN Ensemble Activity: Sparse Coding in the KCs

MB KC firing patterns are very different from those of their input PNs (cf. Figures 3(a) and 3(b)). KCs are almost silent at baseline and respond to ~1–10% of odors with only approximately two spikes per response. The timing of those spikes is cell- and odor-specific. Thus, in contrast to the broad PN tuning (and resulting dense odor representations in the AL), odor representations by KCs are very sparse.

Sparse coding refers to an encoding and representational scheme that uses only a small number of narrowly tuned cells for any one item to be encoded. This coding strategy is found in other systems (e.g., song production circuits in zebra finches, rodent auditory cortex, and the medial temporal lobes of humans). Neurons whose responses are extremely rare and contain only few action potentials (two on average with KCs) can be very hard to detect with extracellular recordings; such neurons and encoding strategies may thus be vastly underrepresented in the literature. There are important reasons to expect that neurons that are involved in associative memory, as KCs are thought to be, might be sparse encoders. Sparseness helps reduce the probability of overlap and potential interference between memories; that is, forming a new association is less likely to affect existing ones. Similarly, the number of synapses that need to be modified to form an association is small. Sparse strategies are thus efficient, particularly when implemented by almost binary spiking neurons. Finally, sparse codes lead to synthetic representations (ones that lack detail of the constituent parts). This is in fact how odors appear to be perceived: multicomponent mixtures, such as ground coffee or flower fragrances that excite a wide array of receptors, are generally perceived as singular. Sparse coding is limiting in terms of capacity, but this disadvantage can be offset by a large number of neurons; KCs are indeed the most numerous central neurons in the insect olfactory system (with about a third of a million KCs in cockroaches and up to 10 million KCs in the MB of a related arthropod, the horseshoe crab *Limulus*).

Mechanisms of Sparsening

How do sparse responses arise, or, stated differently, how is high response specificity generated? In locust KCs, experimental results indicate a combination of contributory mechanisms:

1. Each KC appears to be connected, with a narrow range of synaptic weights, to ~50% of its input population (PNs in the AL). This high number and the enormous space of PN combinations that it allows ($^{830}C_{400} \cong 10^{240}$) imply that KCs could, in principle, encode a very large number of input patterns or stimuli. It also implies that if only a minute fraction of this large number of possible combinations is realized (there are only 50 000 KCs per MB and thus only a maximum of 50 000 possible realizations), the corresponding input patterns are likely to be very different from one another. This could begin to explain why KC responses are so specific. Such high convergence (many PNs to one KC) and divergence (one PN to many KCs) should excite many KCs, however, unless KC firing thresholds are set appropriately high. This appears to be the case, with a requirement for >60 coincident PN spikes per oscillation cycle.

2. Intrinsic properties of KCs contribute to their action as coincidence detectors, allowing them to fire even with high voltage thresholds. Voltage-dependent conductances serve to sharpen excitatory postsynaptic potentials (EPSPs) when the KCs are depolarized. Thus, a PN spike that arrives synchronously with others can contribute disproportionately toward bringing a KC to spike threshold.

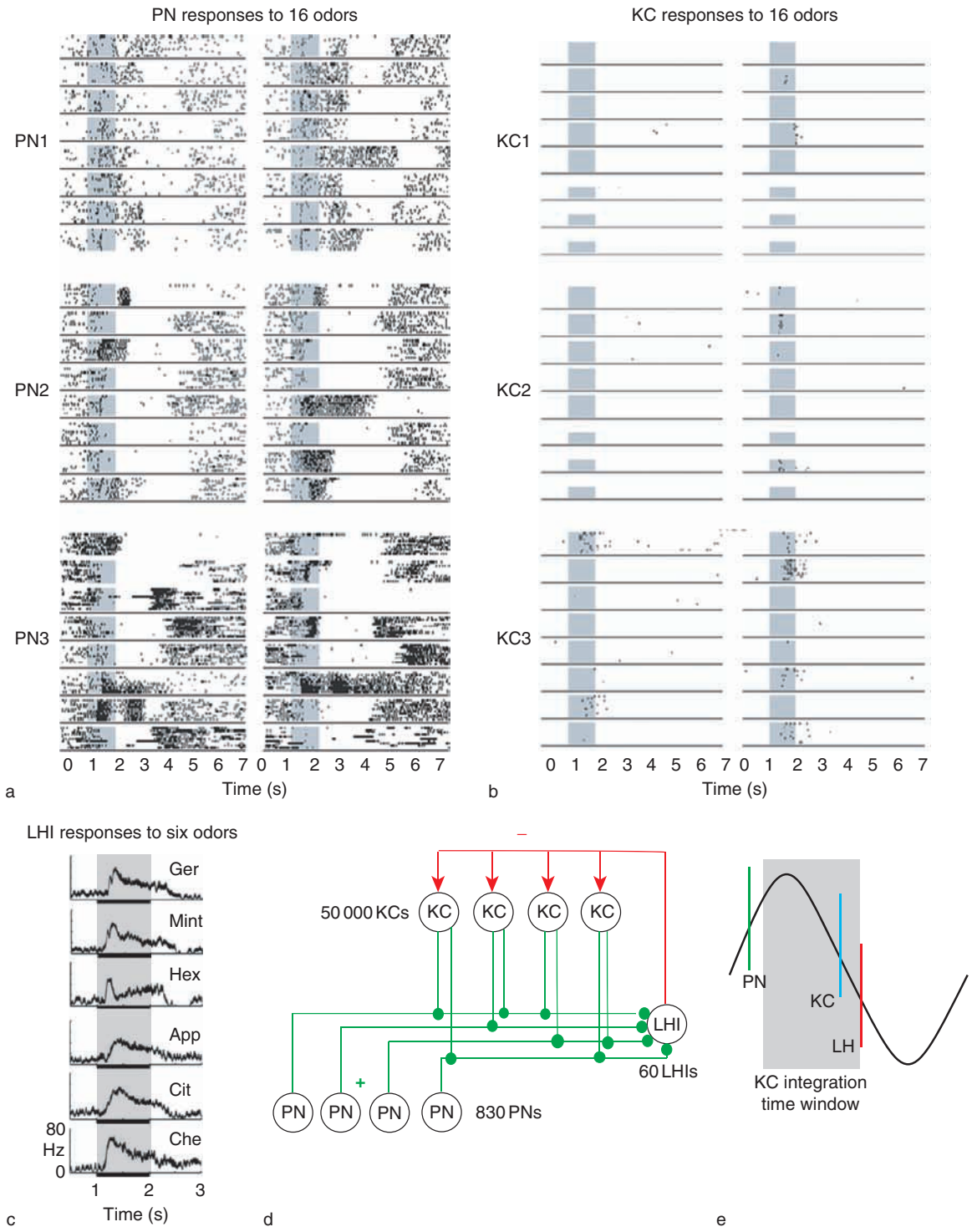

Figure 3 Transformation of odor representations from projection neurons (PNs) to Kenyon cells (KCs). (a) Responses of three simultaneously recorded PNs to 16 different odors. Most PNs respond to a large percentage of odors with a variety of spike patterns. These dense representations of odors are transformed to sparse representations in the mushroom body (MB). (b) Each KC responds to only a small number of odors, and KCs typically respond with very few spikes. (c) Lateral horn interneurons (LHIs) respond to all odors with similar response patterns. (d) Fan-out from PN to KC populations is ~50%; each KC thus samples about half of all PNs. KCs also receive delayed feed-forward inhibitory input from LHIs. (e) Oscillations constrain the time available to each KC to integrate the odor-specific excitatory input it receives from the PNs connected to it. PNs (excitation) tend to fire during the rising phase of the oscillation cycle; LHIs fire about half a cycle later, that is, between the middle and the end of the falling phase of the local field potential. This phase-lagged, nonspecific inhibitory input from LHIs curtails KC integration time to a half oscillation cycle, after which the KC membrane potential is effectively reset by an LHI inhibitory postsynaptic potential. Cit, citral; ger, geraniol.

3. KCs receive odor-evoked feed-forward inhibition from the LH neurons. These GABAergic neurons, which respond nonspecifically to odors (Figure 3(c)), have extensive axonal arborizations in the MB.

4. Excitatory (from PNs) and inhibitory (from LHIs) inputs to KCs arrive at specific times, clocked by the odor-evoked oscillations. PNs tend to fire preferentially during the rising phase of each cycle. LHIs, which receive their input also from the PNs, tend to fire with a delay, during the falling phase of each cycle. Thus, in each cycle during an odor response, each KC receives excitatory input from a subset of PNs (those among its presynaptic set that are coactive). Immediately after this, each KC receives nonspecific inhibitory input from the LHIs; this inhibitory input resets KC membrane potential, limiting or suppressing integration of EPSPs across oscillation cycles (Figures 3(d) and 3(e)). This inhibitory reset means that each KC has the opportunity to fire only during a small fraction of each oscillation cycle, and then only if it gets the appropriately synchronous input from a large enough fraction of the PNs it is connected to. KC firing phases reflect the strong effect of this oscillatory input. KCs are strongly phase-locked with the LFP and tend to fire in the period between the arrival of PN and LHI input (gray bar in Figure 3(e)). The importance of this phase-lagged inhibitory input has been tested pharmacologically. Blocking inhibition in the MB leads to a very significant increase in the width of KC tuning (and therefore, to a decrease in the sparseness of their responses) and to the loss of KC phase locking to the LFP.

In summary, dynamic PN population output, oscillatory PN output, delayed feed-forward inhibition of KCs, PN–KC connectivity, and nonlinear intrinsic properties of KCs all contribute to making the KCs sparse encoders. During an odor stimulus, each KC can be thought of as matching its PN connectivity matrix (the PNs it is connected to) with the entire ensemble PN state vector at every oscillation cycle. If there is significant overlap (i.e., a large-enough number of its presynaptic PNs fire synchronously near the top of a given cycle), the KC reaches threshold and fires. If this does not happen, the KC's membrane potential is reset by delayed inhibition from LHIs. The KC must wait until the next cycle to have a new opportunity to fire. However, because only ~55% of spikes (during a given odor-evoked oscillation cycle) are reliable, the threshold must be reached with the help of spikes from unreliable PNs. That is, although these spikes are not individually reliable, they contribute reliably as a collective during each cycle. Whether this design facilitates robustness to noise or confers other advantages is an open question. It has been observed, however, that intertrial noise in PN responses is smallest during the transient phases of the PN population output.

Processing of Complex Stimuli

The olfactory system has presumably evolved to process not only simple pulses of odors but also (and probably more likely) complex and dynamic stimuli such as caused by turbulent plumes. Realistic stimuli must therefore often be intermittent, unpredictable, and variable in duration, concentration, and chemical composition.

Intensity Invariance

The concentration of an odor can vary across many orders of magnitude depending on the strength of the source, distance from the source, and wind conditions. Animals can typically recognize odors across many concentrations. However, responses of ORNs typically increase with increasing concentration, and individual PNs are strongly affected by the concentration of the odor. Indeed, PN responses to different concentrations of an odor are often qualitatively similar to responses of a PN to different odors. How then does intensity coding differ from identity coding? The olfactory system resolves this seeming confusion by using gain control and ensemble coding. Total firing rate across the PN population does not increase significantly with concentration. Yet at the ensemble level, PN responses are more similar across concentrations of the same odor than they are across different odors. In PN phase space, this can be seen in the fact that PN trajectories for different concentrations of an odor lie within odor-specific surfaces (Figure 4(a)). Thus, KCs that receive inputs from PN ensembles that respond more similarly across concentrations could have similar responses to a wide range of odor intensities, presumably allowing the animal to recognize odors in an intensity-invariant manner. Concentration-invariant KCs do exist (see KC1 and KC2 in Figure 4(b); KC3 is an example of an odor- and concentration-specific KC), but whether invariance results from such connectivity is as yet unknown.

Effect of History and Initial Conditions

A sensory system's ability to process a particular stimulus correctly is often affected by the lasting effects of prior exposure to other stimuli. Such history dependence can limit or help the performance of a system under dynamic conditions, such as when multiple stimuli arrive in rapid succession. When presented with sequences of odors, for example, the AL and MB can respond differently to a given odor, depending on the

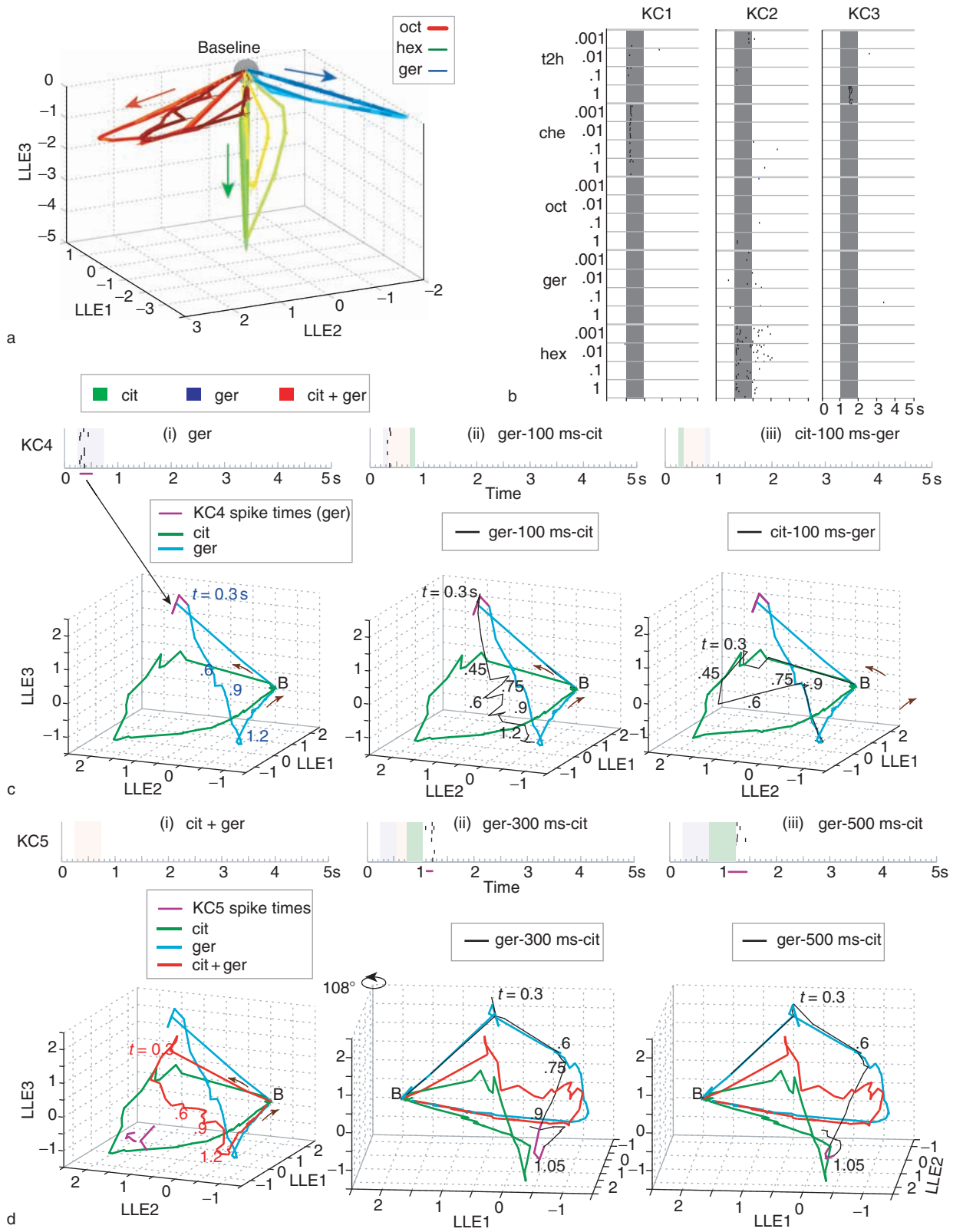

Figure 4 Projection neuron (PN) ensemble encoding and Kenyon cell (KC) decoding: intensity invariance and the effect of stimulus history. (a) Responses of 110 PNs to four different concentrations of three different odors, represented as trajectories after dimensionality reduction using locally linear embedding (LLE). PN ensemble responses to different concentrations of the same odor are more similar to each other than to responses to different odors. PN response trajectories for different concentrations of an odor lie on the same

timing and identity of the preceding stimulus. This effect is illustrated in Figures 4(c) and 4(d).

In general, PN ensembles are able to successfully track an individual odor even when it is preceded by another odor. This is seen in the successful switching of the black trajectory (representing the PN ensemble response during an overlapping odor sequence) from that corresponding to one odor to another in Figure 4(c)(iii). However, the trajectory that the ensemble of PNs follows under such conditions is different from the one it would follow if the odor was presented in isolation (shown in blue and green in the same figures). In some cases, the system is unable to make the switch (Figure 4(c)(ii)), and the second odor is effectively 'masked.' This has consequences for the downstream KCs as well. Each KC is, by virtue of its connectivity to a specific PN, tuned to a specific volume of PN phase space; the size and shape of this volume likely depend on factors such as the synaptic weights associated with its input PN vector, the range of input PN instantaneous firing rates, and the KC's own spike threshold and active properties. If the ensemble PN response trajectory passes through the KC's region of interest, that KC will respond, but not otherwise. In the case illustrated in Figure 4(c), the KC tuned to a particular early phase of the geraniol PN trajectory responds fully (Figure 4(c)(i)), less well (Figure 4(c)(ii)), or not at all (Figure 4(c)(iii)), depending on the path taken by the PN ensemble trajectory (and the overlap between the two odor pulses). In Figure 4(d), the converse is illustrated: the recorded KC responds neither to the two odors presented alone nor to their binary mixture. It responds only for particular conditions of odor overlap that cause the PN ensemble response trajectory to transition through a specific and distinct region of PN phase space. Thus, KCs can be highly tuned to the coactivation of specific PN combinations.

Conclusions and Future Directions

Summary

Understanding olfaction requires understanding how chemical signals, once transduced by ORN arrays, are processed by the brain for a variety of purposes, such as discrimination, tracking, orientation, association, learning/memorization, and recall. Studies in the locust olfactory system provide a window into the sorts of transformations that must occur as information makes its way from the periphery of a sensory system to higher brain centers. In this system, information about the odor environment is represented in dynamically evolving ensembles of PN firing patterns. Each odor condition causes the PN ensemble to travel through a specific sequence of states. These PN ensemble states are relevant for their targets in the MB, the KCs. KCs encode odors as sparse activity vectors, in a format suitable for associations with other sensory modalities and valence. Their responses are consistent with an oscillation-cycle-by-cycle detection of synchronous firing across a large fraction of their PN ensemble input. The design of the system seems suitable both for encoding a large number of odor conditions and for responding rapidly to changes in odor conditions.

Downstream Targets of KCs

The MB's involvement in associative memory in flies and honeybees suggests that downstream MB neurons may be critical for odor learning. It is known that beta lobe neurons respond to odors with phase-locked spikes in stimulus-specific patterns, and recent results show that their inputs from KCs are modulated by Hebbian spike timing-dependent plasticity rules, with a time window of plasticity that corresponds precisely to the oscillation period.

odor-specific surface. (b) KC responses can show varying degrees of intensity invariance, consistent with their receiving input from a large percentage of PNs. Consistent with their decoding of PN ensemble response patterns, KC responses are often odor- and time-specific but not intensity-specific. (c) PN ensemble and KC responses to overlapping presentations of two odors. The responses of individual KCs occur during particular epochs, specific to each odor–KC pair, and thus correspond to particular segments of the PN ensemble response trajectories. Both responses are sensitive to stimulus history. KC4 tends to fire specifically during the first 200 ms after the presentation of geraniol (ger). Its response is masked if citral (cit) is presented within 100 ms of the presentation of ger. If cit precedes ger, KC4 does not fire at all. (i) The period during which KC4 responds to ger (purple bar) is highlighted on the matching epoch of the PN ensemble response (purple segment on blue trajectory). Any PN response trajectory that passes through this part of PN phase space has a high probability of evoking a response in this KC. If the response trajectory (black, corresponding to ger- and cit-pulse onsets delayed between one another by 100 ms) passes through only a small part of this volume (ii), KC4 fires fewer spikes. In (iii), the PN ensemble response to cit preceding ger by 100 ms does not intersect the PN ensemble response to pure ger until much later than the period of KC4's response. Correspondingly, KC4 does not spike. (d) KC5 fires specifically during a transition from ger to cit for particular overlapping odor sequences. Consistent with the PN–KC decoding model, the regions of PN phase space that these transitory periods occupy (shown with purple bar under KC responses and on PN phase plots) are not traversed by any PN ensemble responses to pure odor conditions.

Large Odor Sets, Other Stimulus Paradigms, and Underlying Dynamics

How is the universe of odors represented in the locust brain? Although we are beginning to understand how the AL and MB process odors, the topology of PN and KC phase spaces remains unknown. Discovering structure in these spaces, if it exists, will require very large data sets, with many tens of stimuli. Finally, investigators need to understand the behavior of these networks as dynamical systems: clearly, the underlying motions are not described by classical attractor dynamics (fixed points or limit cycles). Alternative models, such as the recently described 'winnerless competition' must be explored and tested. Successful theoretical frameworks should ultimately explain the stability and advantages of transients for sensory encoding.

Comparison to Other Systems

Several computational features of the locust olfactory system find analogs in other systems. Sparse coding is found across sensory and motor systems and across phyla, with examples ranging from the song production circuits of zebra finches to rodent auditory cortex and the medial temporal lobe of humans. The circuit motif of feed-forward excitation and delayed feed-forward inhibition is seen across brain circuits, and its function in promoting the narrowing of integration time windows for target neurons has also been noted in thalamocortical networks in the rodent barrel cortex. Encoding with dynamic neural assemblies is also a feature of many nervous systems, from leech motor decision-making circuits to mammalian cortical and hippocampal networks. Oscillations are a prominent feature of numerous brain circuits although their precise functional role in other systems is not as clear. Thus, insights gained from a careful analysis of the locust olfactory system may have applications across animals, brains, and functional networks.

See also: Information Coding; Olfactory Coding; Olfactory Receptors; Population Codes: Theoretic Aspects; Population Coding.

Further Reading

Buzsaki G and Draguhn A (2004) Neuronal oscillations in cortical networks. *Science* 304(5679): 1926–1929.

Cassenaer S and Laurent G (2007) Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information. *Nature* 448(7154): 709–713.

Jortner R, Farivar S, and Laurent G (2007) A simple connectivity scheme for sparse coding in an olfactory system. *Journal of Neuroscience* 27(7): 1659–1669.

Kanerva P (1988) *Sparse Distributed Memory.* Cambridge, MA: MIT Press.

Laurent G (2002) Olfactory network dynamics and the coding of multidimensional signals. *Nature Reviews Neuroscience* 3(11): 884–895.

Leonardo A (2005) Degenerate coding in neural systems. *Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology* 191(11): 995–1010.

Mazor O and Laurent G (2005) Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron* 48(4): 661–673.

Perez-Orive J, Mazor O, Turner G, et al. (2002) Oscillations and sparsening of odor representations in the mushroom body. *Science* 297(5580): 359–365.

Stopfer M, Jayaraman V, and Laurent G (2003) Intensity versus identity coding in an olfactory system. *Neuron* 39(6): 991–1004.

Strausfeld NJ and Hildebrand JG (1999) Olfactory systems: Common design, uncommon origins? *Current Opinion in Neurobiology* 9(5): 634–639.

Oligodendrocyte and Schwann Cell Identification Methods

J Trotter and K Karram, Johannes Gutenberg University of Mainz, Mainz, Germany

© 2009 Elsevier Ltd. All rights reserved.

Definition of an Oligodendrocyte and a Schwann Cell by Morphology Using Light and Electron Microscopy

Oligodendrocytes (OLs) and Schwann cells (SCs) are the myelinating cells of the central nervous system (CNS) and peripheral nervous system (PNS), respectively, ensheathing large-diameter axons with multilamellar wrappings of myelin, a specialized subdomain of their plasma membrane. Myelination facilitates saltatory conduction; in addition, signaling between myelinating glia and axons is essential for maintenance of axonal health and integrity during development and in the adult. An oligodendrocyte is able to myelinate multiple axonal internodes; in contrast, an SC myelinates only one internode in the PNS. OLs and SCs were initially identified morphologically using light and electron microscopy (EM). The term oligodendrocyte was first used in 1921 by Rio Hortega and described cells with few processes visible by light microscopy after metallic impregnation techniques on brain sections. Subclasses of oligodendrocytes were defined by differences in the density of their cytoplasm visible by EM. In general, oligodendrocytes have a dense cytoplasm and nucleus, lack fibrils and glycogen in the cytoplasm, and have large numbers of microtubules in their processes. Later EM studies from the group of Richard Bunge showed definitively the connection of the oligodendrocyte to the myelinated axon and proved that one oligodendrocyte can myelinate several axonal internodes. Experiments in which myelinating oligodendrocytes in CNS sections are filled with a fluorescent dye such as Lucifer yellow allow the visualization of oligodendrocytes and their processes. OLs in white (myelinated) matter often appear as interfascicular rows of cells lying between the axon bundles. These can be most clearly seen in well-ordered areas such as the corpus callosum. Both OLs and SCs include nonmyelinating cells; in the adult PNS these nonmyelinating SCs lie interspersed between axon and also enwrap but do not myelinate groups of small-diameter axons. It is thought that these cells can be induced to differentiate into myelinating SCs after damage and demyelination of the PNS. In the CNS, the situation is more complicated: Nonmyelinating OL progenitors are present in gray matter (so-called satellite oligodendrocytes) and the identification and developmental potential of this cell population is a matter of debate.

Identification of Oligodendrocytes and Schwann Cells Using Immunostaining for the Expression of Myelin Proteins

A restricted number of proteins are expressed in CNS and PNS myelin, synthesized by the myelinating glia. Their expression can be determined by staining with specific antibodies *in vitro* and *in vivo*, permitting their use as markers to identify these cell types (Figure 1). In the PNS, SCs upregulate myelin proteins upon axonal contact. The expression of the proteins P0, PMP22, P2, and myelin-associated glycoprotein (MAG) characterize myelinating SCs. In the CNS, OLs can express all the main myelin proteins *in vitro* in the absence of axons, although the amount made and their subcellular location appear to be regulated by neuronal presence. The myelin proteins proteolipid protein, myelin basic protein, MAG, 2′,3′-cyclic nucleotide 3′-phosphohydrolase (CNP; Figure 2), and myelin oligodendrocyte glycoprotein are expressed by OLs. OLs also express minor proteins, such as the GPI-linked adhesion molecules NCAM 120 and F3. In contrast to OLs, SCs express the cell adhesion molecule L1, which is also expressed by neurons. Expression of L1 is downregulated on myelinating SCs. During myelination, the complex structure of the myelin sheath has to be elaborated and several minor myelin proteins are expressed by OLs and SCs which then are directed to specialized locations such as the paranodal areas of uncompacted myelin. These include members of the connexin family (e.g., connexin 29, which is in OLs and SCs), members of the claudin family (e.g., claudin 11 in OLs), and specific cytoskeletal-associated proteins such as ermin in OLs and periaxin in SCs. The myelin lipids sulfatide and galactocerebroside are present in both PNS and CNS myelin and thus antibodies against these lipids characterize OLs and SCs *in vivo* and *in vitro*. Immunostaining can also yield information about the site of expression of a gene product within the cell as well as heterogeneity within a population: Thus, myelin basic protein is expressed strongly in the myelin compartment of actively myelinating OLs, whereas the cell body is less strongly stained.

Delineation of OL and SC lineage and the protein and lipid antigens expressed by OLs and SCs during

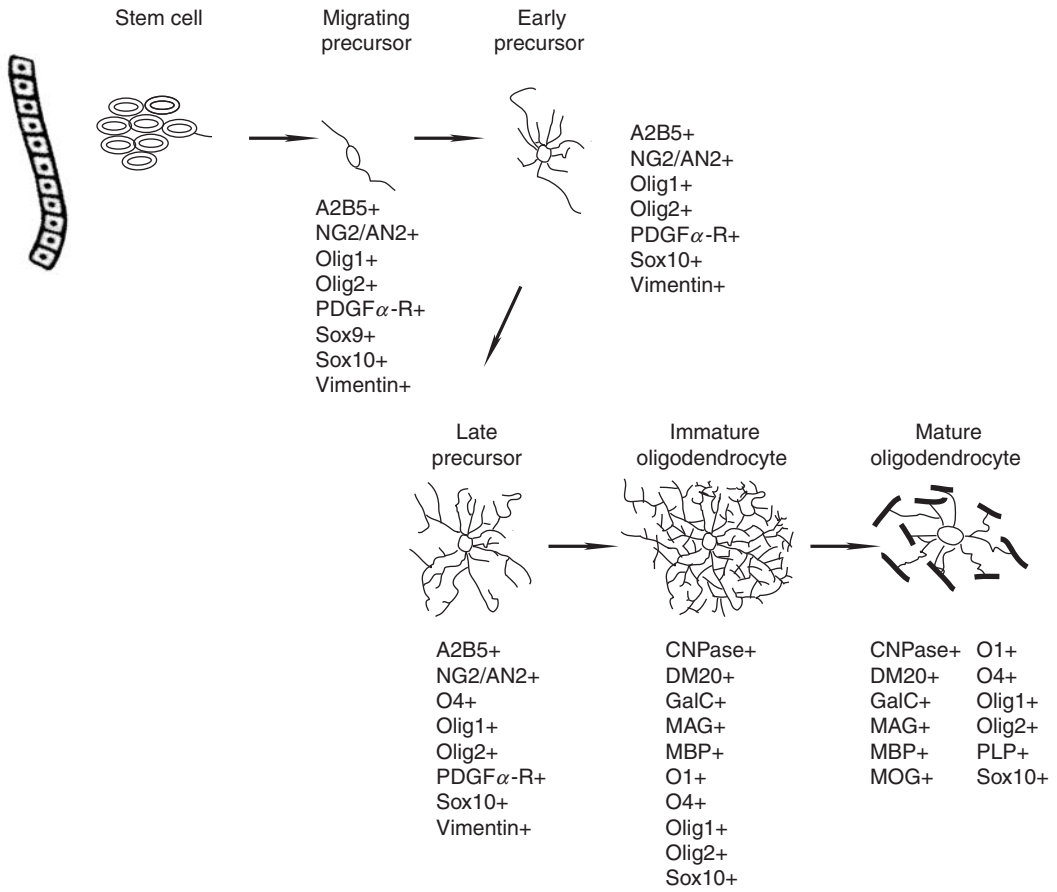

Figure 1 Schematic summary of oligodendrocyte development. Antigenic profile of oligodendrocyte development in the central nervous system.

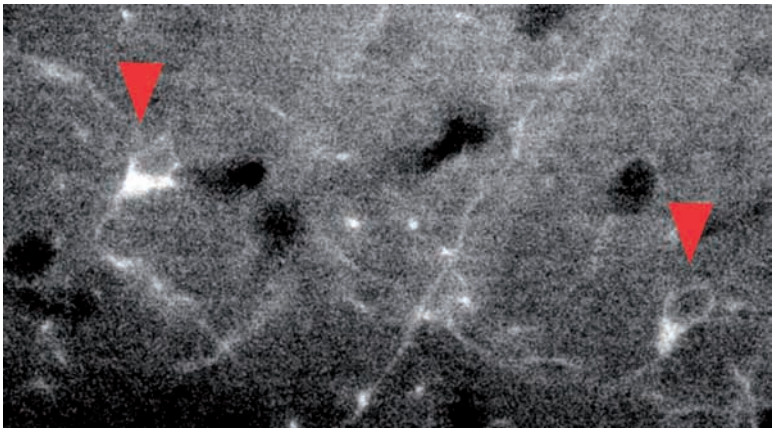

Figure 2 Microscope image of the cortex of a 75-day-old mouse stained with an antibody that recognizes CNP (arrows point to oligodendrocytes).

development has been aided by studies in cell culture. In the PNS, co-culture of SCs and, for example, dorsal root ganglion neurons has allowed the characterization of the expression and role of SC-expressed proteins in myelination. In the CNS, co-cultures of OL and DRG neurons will myelinate within 2 weeks and the use of explants of CNS tissue is common (Figure 3). These *in vitro* systems aid in the characterization of different populations of OLs and their precursor cells and allow a good comparison with

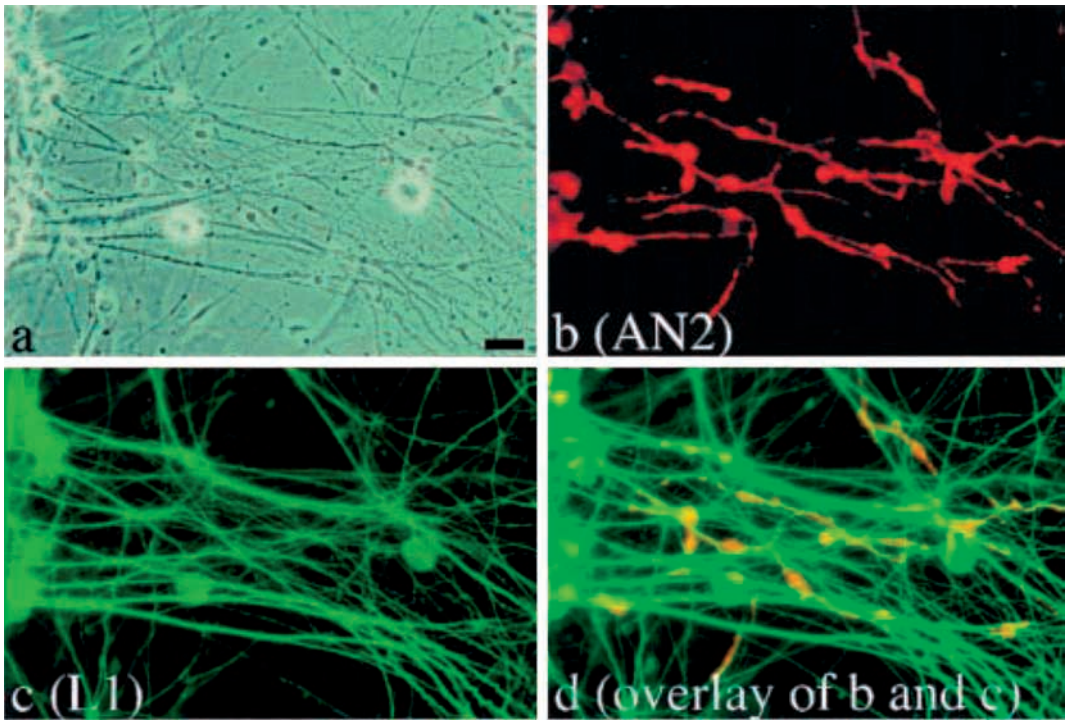

Figure 3 Immature murine Schwann cells in co-culture with DRG neurons express the L1 glycoprotein and the NG2 proteoglycan. The co-cultures were stained with AN2 polyclonal antibodies recognizing NG2 (b) and L1 monoclonal antibodies (c). (d) shows the overlay of (b) and (c), and (a) shows the corresponding phase-contrast picture. Scale bar = 25 μm. Reproduced from Schneider S, Bosse F, D'Urso D, et al. (2001) The AN2 protein is a novel marker for the Schwann cell lineage expressed by immature and nonmyelinating Schwann cells. *Journal of Neuroscience* 21(3): 920–933, with permission. Copyright 2001 by the Society for Neuroscience.

in vivo models. Since cell cultures can be manipulated, functional studies of new gene products can be carried out in dissociated and explant culture, and complement studies performed using genetically manipulated mice in which the gene product has been mutated or eliminated.

In Situ Hybridization to Detect the Presence of Myelin Proteins

Multiple studies have analyzed the *in situ* expression of myelin proteins by localizing the expression of mRNA for the protein. This has additionally facilitated the identification and *in situ* localization of OLs and SCs in sections of CNS and PNS.

Expression of Transcription Factors Specific for Oligodendrocytes and Schwann Cells

Expression of a range of transcription factors characterizes myelinating glia. Thus, Olig1, Olig2, and Sox10 are expressed in developing and mature OLs, and Sox10 is expressed by SCs. However, the expression of some of these transcription factors is not exclusive to OLs; for example, Olig2 is important for OL development and expressed by mature OLs but is also essential for the generation of groups of motor neurons. Sox9 is a transcription factor for glial determination and appears early in the OL precursor, but it is downregulated in late OL precursors, when they start expressing Sox10 and are destined to be myelinating OLs. In the PNS, Sox10 is a key regulator of differentiation of SCs, and Sox10 knockout mice lack SCs.

Electrophysiological Characterization of Oligodendrocytes and Schwann Cells

Myelinating glial cells exhibit characteristic electrophysiological properties due to the expression of a range of ion channels and neurotransmitters which were initially thought to be exclusively expressed by neuronal cells. These were first demonstrated in cell culture and subsequently in sections. In both cases, whole cell patch clamp is extensively used in combination with a range of antagonists and agonists of the specific receptors and channels. For example, OL lineage cells express receptors for glutamate: These include the AMPA/kainate and NMDA classes of

receptors, although the latter appear to be expressed mainly on the cell processes, making detection by patch clamp techniques difficult. The AMPA receptors are expressed predominantly by precursor cells, as are receptors for GABA (the A subtype). SCs also express neurotransmitter receptors and channels which differ from those expressed by OLs.

Use of Genetically Manipulated Mice to Identify and Study Oligodendrocytes and Schwann Cells *In Situ* and Follow Oligodendrocyte and Schwann Cell Development

Transgenic approaches have been utilized to aid in the identification of OLs and SCs *in situ*. These approaches use the promoter for a myelin protein gene to drive the expression of a fluorescent reporter such as enhanced green fluorescent protein (EGFP). In many instances, the constructs are inserted as transgenes randomly in the mouse genome after injections into oocytes. A more elegant approach consists of placing the fluorescent indicator under the expression of the endogenous promoter of the gene (so-called knockin technology), usually by homologous recombination in exon 1 of the gene in embryonic stem cells followed by blastocyst injection. This latter approach has the advantage that the expression of the reporter mimics the endogenous gene expression, and there is less risk of interfering with the expression of other genes due to random insertion of the transgene in the mouse genome. However, there is a risk that the endogenous promoter activity may not be sufficient to generate enough reporter protein to visualize the cells. Marking the cells with fluorescent reporters allows the visualization of cells in living brain slices and facilitates their analysis by electrophysiological and imaging techniques such as two-photon microscopy. In addition, technical advances in which cells can be visualized in living, anesthetized mice through the skull clearly increase the amount of biological information that can be obtained from such animals using relatively noninvasive methods. Furthermore, use of such reporters allows direct isolation of the labeled cells via fluorescence-activated cell sorting, without prior antibody binding. This is particularly advantageous when the antigen of interest is readily degradable by the proteases required to dissociate postnatal neural tissue. A further advantage of the knockin approach over the transgenic insertion of genes is the generation of a knockout mouse for the gene of interest when the knockin mice are bred to homozygosity. In this instance, the cells in which both alleles of the gene have been deleted are labeled, thus facilitating detailed comparison with the heterozygous mice which still carry one copy of the normal allele.

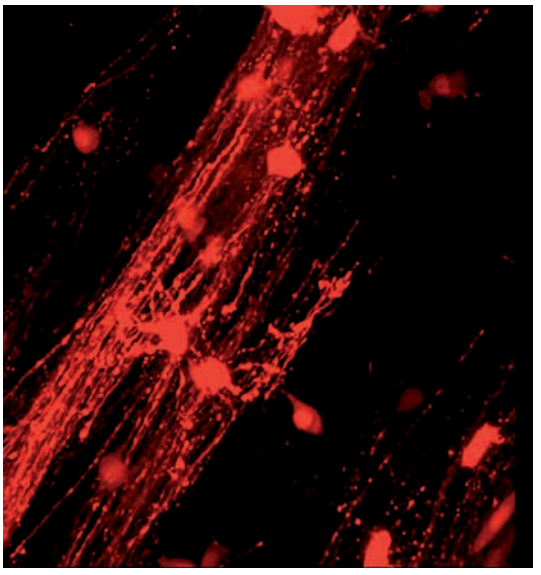

Figure 4 Confocal section of the brain of an adult mouse expressing the PLP-DsRed transgene. The transgene is expressed in oligodendrocytes and myelin. Courtesy of Dr. Frank Kirchhoff.

The bacterial Cre recombinase can be inserted after the endogenous OL or SC promoter; such mice can then be crossed to mice in which a reporter gene (e.g., EGFP) has been preceded by floxed stop sites. In cells in which the specific promoter is active, Cre recombinase is synthesized and deletes the stop sites, yielding labeled cells at the first incidence of promoter activity. In this way, all progeny resulting from an initial promoter activity are labeled. In contrast, in mice in which the reporter gene is inserted directly after the promoter of interest, the cells are labeled only during the period of promoter activity. More differentiated progeny in which the promoter is no longer active lose the label. In addition, animals expressing Cre under the influence of an OL-specific promoter can be used to delete genes (and thus proteins) specifically in OLs by crossing the Cre mice with mice in which the genes to be deleted are floxed.

PLP-EGFP transgenic mice, PLP-DsRed transgenic mice (Figure 4), and CNP-EGFP transgenic mice have been reported. The CNP-Cre knockin mouse has proven extremely useful to specifically target genes in the OL lineage.

Problems of Precursor Cells

Definition of the direct precursors to myelinating SCs and OLs is important. The SC lineage is relatively

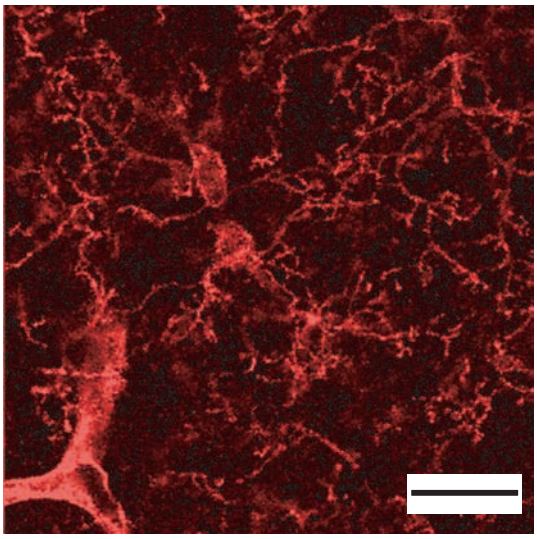

Figure 5 Confocal image scan of the cortex of a 5-day-old mouse stained with the monoclonal AN2 antibody recognizing mouse NG2. NG2-expressing progenitor cells and NG2-expressing pericytes around a capillary are visible. Scale bar = 20 μm.

well-defined, but there is considerable confusion about oligodendrocyte progenitors. This has implications for human diseases, most notably multiple sclerosis, in which resynthesis of myelin does not occur in chronic lesions despite the presence of a population of cells bearing markers of OL precursors. OL precursors express the polysialylated form of N-CAM, receptors for platelet-derived growth factor AA, as well as antigens recognized by the O4 antibody. Expression of the NG2 proteoglycan was originally thought to define oligodendrocyte precursor cells, but there is now considerable speculation regarding whether all NG2+ cells which are found in white and gray matter in developing and adult brain are capable of developing into OLs *in vivo* (Figure 5). Furthermore, the question as to whether some of these cells can also generate neurons is a hotly debated area of research.

Additional Animal Models to Study Myelination

Zebra fish have been explored as a tool for the study of myelination. It was shown that the genes for the main myelin proteins are largely conserved in zebra fish. The advantages of zebra fish are the rapid embryonic development and the transparent embryos, permitting the visualization of fluorescently labeled cells during development in the living animal. It is hoped that the ability to rapidly mutagenize genes in the fish by adding substances to the aquarium water and the relative ease of screening many fish embryos expressing fluorescent constructs will permit the rapid isolation of mutants affecting OL and SC development and permit the delineation of subpopulations of these cells and their functions.

See also: Fluorescence Microscopy in the Neurosciences; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission; Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells; Oligodendrocyte Morphology; Oligodendrocyte Specification; Schwann Cells and Axon Relationship; Schwann Cell Development; Schwann Cells and Plasticity of the Neuromuscular Junction; Schwann Cell Morphology.

Further Reading

Bunge MB, Bunge RP, and Pappas GD (1962) Electron microscopic demonstration of connections between glia and myelin sheaths in the developing mammalian central nervous system. *Journal of Cell Biology* 12: 448–453.

Butt AM and Nishiyama A (2002) Novel NG2–glia. *Journal of Neurocytology* 31: 419–565.

Butt AM and Ransom BR (1993) Morphology of astrocytes and oligodendrocytes during development in the intact rat optic nerve. *Journal of Comparative Neurology* 338: 141–158.

Girault JA and Peles E (2002) Development of nodes of Ranvier. *Current Opinion in Neurobiology* 12: 476–485.

Hirrlinger PG, Scheller A, Braun C, et al. (2005) Expression of reef coral fluorescent proteins in the central nervous system of transgenic mice. *Molecular and Cellular Neuroscience* 30: 291–303.

Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

Lappe-Siefke C, Goebbels S, Gravel M, et al. (2003) Disruption of Cnp1 uncouples oligodendroglial functions in axonal support and myelination. *Nature Genetics* 33: 366–374.

Lazzarini RA, Griffin JW, Lassmann H, Nave K, and Trapp BD (2004) *Myelin Biology and Disorders*. New York: Elsevier.

Ligon KL, Fancy SP, Franklin RJ, and Rowitch DH (2006) Olig gene function in CNS development and disease. *Glia* 54: 1–10.

Peters A, Palay SL, and Webster HD (1976) *The Neurons and Supporting Cells*. Washington, DC: Saunders.

Reynolds R and Hardy R (1997) Oligodendroglial progenitors labeled with the O4 antibody persist in the adult rat cerebral cortex *in vivo*. *Journal of Neuroscience Research* 47: 455–470.

Salzer JL (2003) The AN2 protein is a novel marker for the Schwann cell lineage expressed by immature and nonmyelinating Schwann cells. *Journal of Neuroscience* 3: 920–933.

Schneider S, Bosse F, D'Urso D, et al. (2001) The AN2 protein is a novel marker for the Schwann cell lineage expressed by immature and nonmyelinating Schwann cells. *Journal of Neuroscience* 21(3): 920–933

Sommer I and Schachner M (1981) Monoclonal antibodies (O1 to O4) to oligodendrocyte cell surfaces: An immunocytological study in the central nervous system. *Developmental Biology* 83: 311–327.

Sternberger NH, Itoyama Y, Kies MW, and Webster HD (1978) Myelin basic protein demonstrated immunocytochemically in oligodendroglia prior to myelin sheath formation. *Proceedings of the National Academy of Sciences of the United States of America* 75: 2521–2524.

Trapp BD, Andrews SB, Cootauco C, and Quarles R (1989) The myelin-associated glycoprotein is enriched in multivesicular bodies and periaxonal membranes of actively myelinating oligodendrocytes. *Journal of Cell Biology* 109: 2417–2426.

Verity AN and Campagnoni AT (1988) Regional expression of myelin protein genes in the developing mouse brain: *In situ* hybridization studies. *Journal of Neuroscience Research* 21: 238–248.

Verkhratsky A and Steinhauser C (2000) Ion channels in glial cells. *Brain Research: Brain Research Reviews* 32: 380–412.

Yoshida M and Macklin WB (2005) Oligodendrocyte development and myelination in GFP-transgenic zebrafish. *Journal of Neuroscience Research* 81: 1–8.

Yuan X, Chittajallu R, Belachew S, et al. (2002) Expression of the green fluorescent protein in the oligodendrocyte lineage: A transgenic mouse for developmental and physiological studies. *Journal of Neuroscience Research* 70: 529–545.

Oligodendrocyte Morphology

A M Butt, University of Portsmouth, Portsmouth, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Oligodendrocytes are defined as the cells that produce the myelin sheaths which insulate central nervous system (CNS) axons. Oligodendrocytes are one of the two major kinds of macroglial cell, the other being astrocytes. Embryologically, oligodendrocytes and astrocytes are of neural origin and are distinct from the third class of glial cell – microglia – which are nonneural in origin. Myelinating oligodendrocytes are most numerous within CNS white matter, and it is their fatty myelin sheaths that give white matter its translucent appearance when the CNS is labeled with histochemical dyes. The cellular morphology of myelinating oligodendrocytes was first described in 1928 by Del Rio Hortega, who used classical metal-impregnation staining techniques. He subdivided myelinating oligodendrocytes into four subtypes, types I–IV, characterized by the number of their processes and the size of the axons they myelinated. The cellular connection between oligodendrocytes and myelin sheaths was not confirmed until the 1960s by electron microscopy (EM). In recent decades, the use of intracellular dye injection has enabled us to visualize in their entirety individual myelinating oligodendrocyte units, defined as the cell body, processes, and myelin sheaths (Figure 1). Immunocytochemical and *in situ* hybridization techniques have provided further insight into, for example, the functions of myelin proteins, receptors, and transcription factors in myelination.

In addition to myelinating oligodendrocytes, there are also nonmyelinating or satellite oligodendrocytes found within the CNS gray matter, either closely apposed to neuronal cell bodies (perineuronal) or to blood vessels (perivascular). These cells were recognized by Del Rio Hortega, but they are only minor populations and their functions remain unresolved. The mature CNS also contains a significant population of adult oligodendrocyte precursors (OPCs), which were identified in the 1980s in cell cultures and *in vivo* by imunolabeling with antibodies for the NG2 chondroitin sulfate proteoglycan. These NG2-expressing glia are numerous in the gray and white matter, and it is not clear that all these cells are OPCs. Within the last few years, it has been proposed that NG2-expressing glia represent a fourth kind of morphologically distinct mature glial cell, termed synantocytes; their function is unresolved, although some may be capable of regenerating myelinating oligodendrocytes, which is of particular importance following demyelination.

Morphological Subtypes of Myelinating Oligodendrocytes

The rodent optic nerve and anterior medullary velum have been particularly useful model systems in which to examine oligodendrocyte morphological heterogeneity, using both intracellular dye injection and immunolabeling with the Rip antibody. Dye injected into the soma diffuses throughout the cytoplasm of the cell, delineating the cytoplasmic continuity between the soma, processes, terminal paranodal loops, and inner and outer tongue processes of myelinating oligodendrocytes (Figure 1). Similarly, Rip is almost unique in that it labels oligodendroglial cell somata, processes, and myelin sheaths (Figure 2); the majority of oligodendroglia-specific antibodies tend to label either the myelin sheaths (e.g., myelin basic protein) or cell bodies (e.g., carbonic anhydrase).

Characteristics of Oligodendrocyte Types I–IV

Type I and II oligodendrocytes have small somata from which extend four or more very fine primary processes, which then branch to contact and myelinate 10–30 small-diameter axons, generally less than 2 μm in diameter, with relatively short internodal lengths of 50–150 μm (Figures 1(a) and 2(a)). Type III oligodendrocytes have much larger cell bodies than type I and II oligodendrocytes, and they have thick primary processes which rarely branch and myelinate a small number (usually fewer than five) of large-diameter axons, greater than 4 μm, with long internodal lengths of 200–500 μm (Figures 1(b) and 2(b)). Type IV oligodendrocytes are directly applied to a single large-diameter axon (similar to the one-to-one relationship between a myelinating Schwann cell and an axon) and have internodal lengths as great as 1000 μm. The myelin sheaths of type III and IV oligodendrocytes have a network of cytoplasmic reticulations, called the Schmidt–Lantermann incisures, which are cytoplasmic intrusions into the compacted myelin sheaths.

Distributions of Oligodendrocyte Types I–IV

A range of transitional morphologies exist between types I and II and between types III and IV, but the range is less between types I–II, on the one hand, and types III–IV, on the other. It appears, therefore, that type I–II and type III–IV oligodendrocytes are two

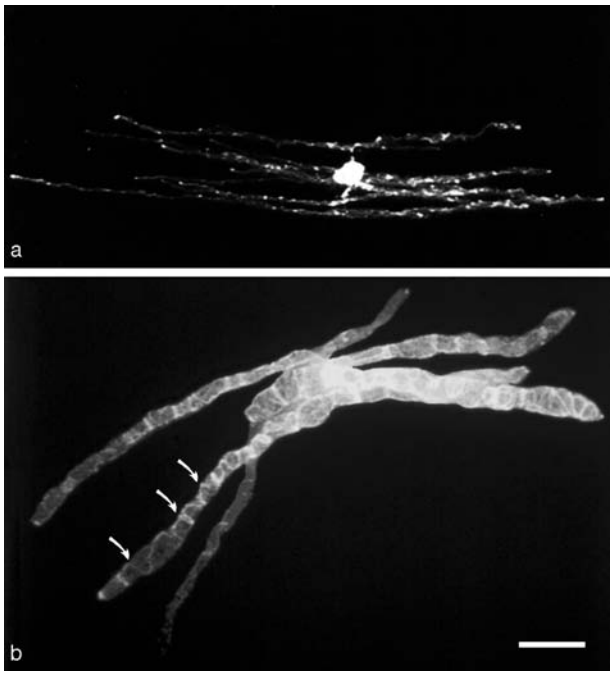

Figure 1 Dye-filled oligodendrocytes: (a) type II oligodendrocyte from the rodent optic nerve, which myelinates multiple small-diameter axons; (b) type III oligodendrocyte from the rodent anterior medullary velum, myelianting two large-diameter axons. The injection of dye into the cell body distributes throughout the cell to fill the cell body, processes, and the cytoplasmic parts of the internodal myelin sheaths, which appear as parallel arrays of dye-filled processes. Dye injection reveals the morphological heterogeneity of oligodendrocytes. In (b), the dye-filled cytoplasmic ridges are seen spiraling around the axon (arrows). Scale bar = 25 μm. (a) Reproduced from Butt AM, Colquhoun C, and Berry M (1994) Confocal imaging of glial cells in the intact rat optic nerve. *Glia* 10: 315–322, with permission from John Wiley & Sons. (b) Reproduced from Berry M, Ibrahim M, Carlile J, Ruge F, Duncan A, and Butt AM (1995) Axon–glial relationships in the anterior medullary velum of the adult rat. *Journal of Neurocytology* 24: 965–983, with permission from Springer-Verlag Berlin.

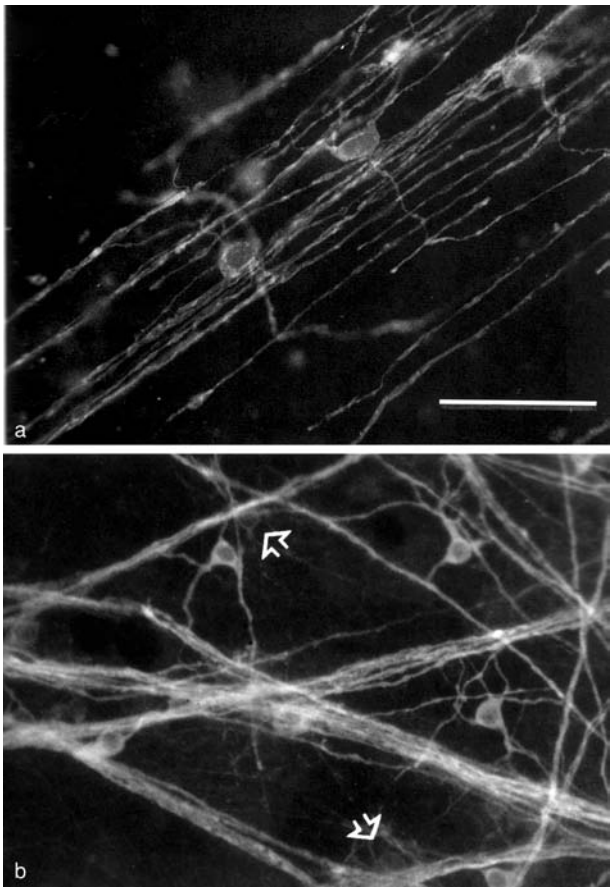

Figure 2 Immunolabeling with the Rip antibody of oligodendrocytes in the postnatal rat anterior medullary velum: (a) distribution of type I–II oligodendrocytes, myelinating small-diameter axons; (b) distribution of type III–IV oligodendrocytes, myelinating large-diameter axons. Arrows indicate immature oligodendrocytes. Scale bar = 50 μm. (a) Reproduced from Butt AM, Ibrahim M, Ruge FM, and Berry M (1995) Biochemical subtypes of oligodendrocyte in the anterior medullary velum of the rat as revealed by the monoclonal antibody Rip. *Glia* 14: 185–197, with permission from John Wiley & Sons. (b) Reproduced from Butt AM, Ibrahim M, and Berry M (1997) The relationship between developing oligodendrocyte units and maturing axons during myelinogenesis in the anterior medullary velum of neonatal rats. *Journal of Neurocytology* 26: 327–338, with permission from Springer-Verlag Berlin.

discrete populations which myelinate small- and large-diameter axons, respectively. Hence, type I–II oligodendrocytes are found in areas with the smallest-diameter axons – type I in the forebrain, cerebellum, and spinal cord and type II in the white matter, such as the corpus callosum; optic nerve; and cerebellum. Type III oligodendrocytes are found in areas with large-diameter axons, mainly the cerebral and cerebellar peduncles, the medulla oblongata, and the spinal cord. Type IV oligodendrocytes are restricted to tracts containing the largest-diameter fibers and occur near the entrance of nerve roots into the CNS. Thus, although oligodendrocytes are morphologically heterogeneous overall, they are generally quite homogeneous within a given tract determined by the diameter of axons below and above a critical diameter of 2–4 μm.

The Myelin Sheath

Consecutive myelin sheaths along axons are separated by nodes of Ranvier, and the myelin sheath between nodes is the internode. Each myelin sheath comprises a large sheet which is wrapped around the axon to form concentric layers or lamellae. If unwrapped, the myelin sheath would appear as an extraordinarily large trapezoid extension of the oligodendroglial cell membrane, with most of the cytoplasm extruded to leave a cytoplasmic ridge which surrounds the compacted myelin. The cytoplasmic

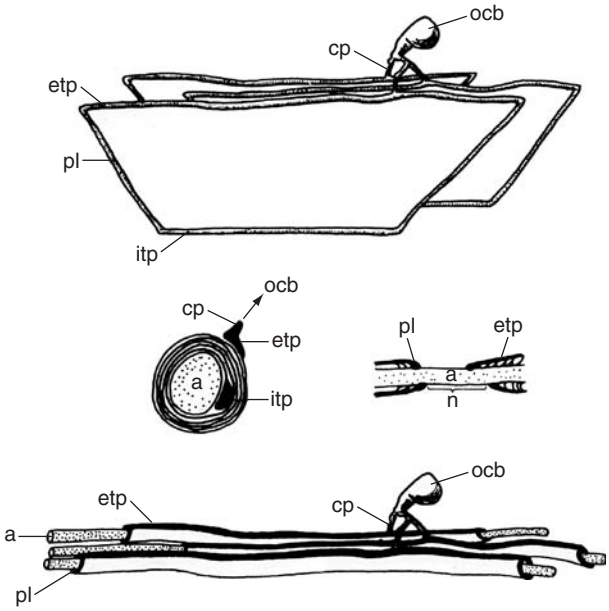

Figure 3 Dimensions of the myelin sheath. If they could be unwrapped, myelin sheaths would appear as large sheets of compacted cell membrane with ridges of cytoplasm, which correspond to the external tongue process (etp), paranodal loops (pl), and internal tongue process (itp) observed in electron microscopic transverse and longitudinal sections. Dye injected into the cell body (ocp) passes down the connecting processes (cp) to fill the cytoplasmic ridges, tracing the pathway by which myelin components are transported from the cell body to the myelin sheets. Reproduced from Butt AM and Ransom BR (1989) Visualization of oligodendrocytes and astrocytes in the intact rat optic nerve by intracellular injection of Lucifer yellow and horseradish peroxidase. *Glia* 2: 470–475, with permission from John Wiley & Sons.

ridge is continuous with the cytoplasm of the oligodendroglial processes and cell body (Figure 3). Thus, each myelin sheath is formed by the fusion of apposed cytoplasmic faces of the plasma membrane (to form the major dense line), which is wrapped around the axon by fusion of the outer faces of the plasma membrane (to form the intraperiod line). When wrapped around the axon, the inner and outer cytoplasmic ridges form the mesaxons and the lateral cytoplasmic ridges build up at the termini of the internodal sheath to form the paranodal loops which are apposed to the axolemma by complex axoglial junctions (Figure 3). The cytoplasmic ridges are the intracellular pathway along which all of the myelin products are transported from the cell body and targeted to the myelin sheath, over hundreds and potentially thousands of micra. This is achieved by intracellular transport involving microtubules and other components of the cytoskeleton. One of the most interesting aspects of protein targeting in oligodendrocytes is the myelin basic proteins. These proteins are not translated in the cell body because they are strongly charged and this would impede their transport to the myelin sheath; instead, the mRNA for myelin basic protein is translocated to the cytoplasmic ridges, where the protein is translated and inserted into the myelin sheath.

Dimensions of the Myelin Sheath

There are strict relationships between the dimensions of the myelin sheath and axon diameter, and these determine the conduction properties of the axon. The thickness of the myelin sheath (i.e., the number of lamellae, N) and the internodal length (L) are directly and positively correlated with axon diameter (D); that is, the larger the axon, the thicker and longer its myelin sheath is. The number of lamellae determines the insulating properties of the myelin sheath, and the internodal length determines the speed of conduction, both of which are greater for large-diameter axons. Thus, axons in type III–IV axon–oligodendrocyte units conduct faster than those in type I–II axon–oligodendrocyte units. The dimensions of each myelin sheath can be calculated as πDNL, and to determine the total amount of myelin supported by a single oligodendrocyte, this must then be multiplied by the number of sheaths within the unit. The ratio of axons to oligodendrocytes is less than 5:1 in type III–IV units and greater than 15:1 in type I–II units and is negatively correlated to the diameter of axons in the unit. There is a sharp demarcation between type I–II oligodendrocytes, which support myelin volumes of approximately 500 μm^3, and type III–IV oligodendrocytes, which support myelin volumes of approximately 30 000 μm^3, and those that myelinate the largest-diameter spinal cord axons as much as 150 000 μm^3. Hence, a single oligodendrocyte would be visible to the naked eye if the entire expanse of its myelin were laid out. The production of these large membranes places a considerable metabolic load on the cell, and it is perhaps not surprising that oligodendrocytes direct their entire energy to the exclusive function of myelin production. The large cell bodies and thick radial processes of type III–IV units and their prominent Schmidt–Lantermann incisures facilitate the distribution of myelin gene products.

Composition of the Myelin Sheath

Compacted myelin comprises multiple fused phospholipid bilayers, and it is these lipids that provide the insulating properties of the sheath. Galactocerebroside (GalC) is a specific oligodendroglial lipid and is a useful immunohistochemical tool for studying oligodendrocytes. The other main constituent of myelin are proteins (30% of its dry weight), the major ones being myelin basic protein (MBP) and

proteolipid protein (PLP), which are specific to myelin and are used immunohistochemically to identify and study myelinating oligodendrocytes. The main functions of MBP and PLP are in the fusion of the cytoplasmic and extracellular interfaces of the myelin lamellae, respectively. In addition, there are a number of proteins that in the CNS are specific to oligodendrocytes and which are important in myelin formation, including 2′,3′-cyclic nucleotide-3′-phosphodiesterase (CNP), myelin/oligodendrocyte specific protein (MOSP), myelin-associated/oligodendrocyte basic protein (MOBP), transferrin, carbonic anhydrase, and members of the tetraspan-protein family. Still other myelin-specific proteins are important in cell–cell interactions, for example, myelin-associated glycoprotein (MAG), myelin oligodendrocyte glycoprotein (MOG), and oligodendrocyte-myelin glycoprotein (OMgp). Myelin is one of the most complex structures in the body and this is underscored by the large number of specialized proteins required for its formation and maintenance.

Oligodendrocyte Development

During development, myelinating oligodendrocytes develop from OPCs that originate in ventricular zones in the brain and spinal cord, and migrate to their final sites in the brain and spinal cord, where they undergo local proliferation and differentiation, or apoptosis, in response to multiple cellular and extracellular signals. Oligodendrocytes pass through a series of distinct developmental phases that can be identified by the expression of stage-specific antigens (Figure 4): OPCs express platelet-derived growth factor α receptors (PDGFαR) and NG2; as they differentiate into immature oligodendrocytes they first gain expression of the sulfatides recognized by the O4 antibody and then GalC, as they lose PDGFαR and NG2 but not O4; finally, oligodendrocytes then begin to express myelin-related gene products, such as MBP and Rip.

Morphological Changes during Myelination

A progression of morphological alterations during myelination has been identified using intracellular dye-filling and immunohistochemistry (Figure 4). Immature oligodendrocytes express myelin-related proteins prior to axon ensheathment and extend a large number of fine processes in all directions for 3–20 μm. Processes that engage axons bifurcate and extend initiator processes that begin to spiral longitudinally along axons. These initial ensheathments are uncompacted myelin sheaths, and their growth along the axons is remarkably uniform and symmetrical

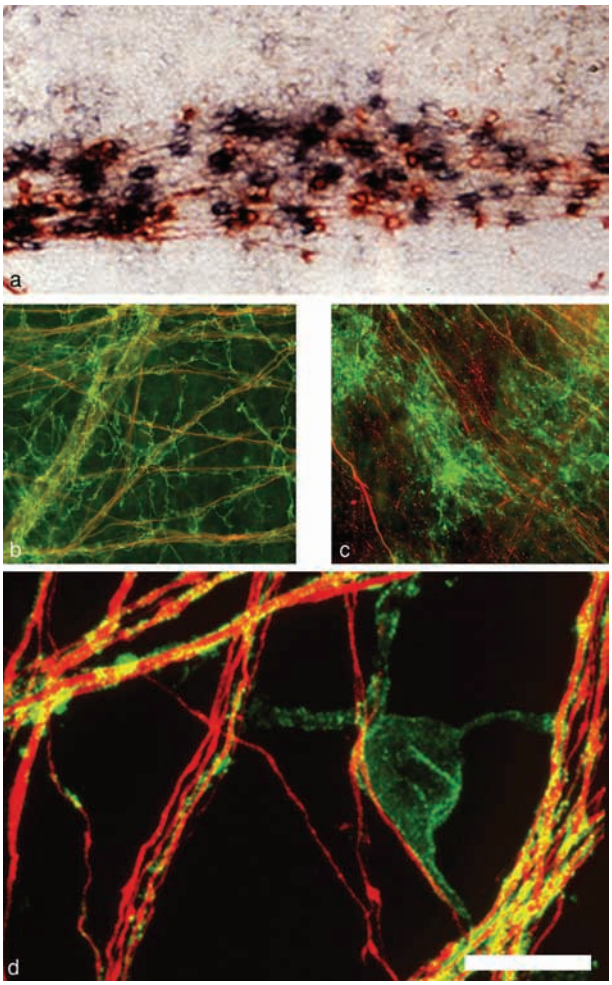

Figure 4 Oligodendrocyte development of the postnatal anterior medullary velum: (a) double *in situ* hybridization labeling; (b–d) immunofluorescence labeling. Oligodendrocyte progenitor cells are identified by the expression of PDGFαR (in panel (a) blue) and NG2 (in panel (b), green). As oligodendrocyte progenitors differentiate into oligodendrocytes, they begin to express myelin proteins, such as MBP (in panel (a), red) and Rip (in panel (c), green). Progenitors extend numerous radial processes which form multiple contacts with unmyelinated axons (in panel (a), red). Premyelinating oligodendrocytes have a more complex morphology and extend innumerable fine branching processes that contact axons via fine filipodia and extend short initiator processes along axons (in panel (b), red). In (d), oligodendrocytes go through a remodeling phase during which they lose non-ensheathing processes and the number of sheaths per unit is established. Scale bar = 100 μm (a), 75 μm (b,c), and 20 μm (d). (a) Reprinted from Butt AM, Hornby MF, Ibrahim M, Kirvell S, Graham A, and Berry M (1997) PDGF-α receptor and myelin basic protein mRNAs are not coexpressed by oligodendrocytes *in vivo*: A double *in situ* hybridization study in the anterior medullary velum of the neonatal rat. *Molecular and Cellular Neuroscience* 8: 311–322, with permission from Elsevier. (d) Reproduced from Butt AM, Ibrahim M and Berry M (1997) The relationship between developing oligodendrocyte units and maturing axons during myelinogenesis in the anterior medullary velum of neonatal rats. *Journal of Neurocytology* 26: 327–338, with permission from Springer-Verlag Berlin.

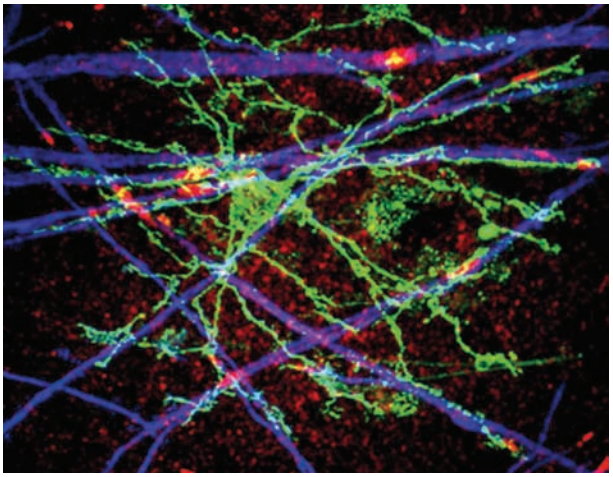

Figure 5 NG2-expressing synantocyte. Confocal micrograph of triple immunolabeled anterior medullary velum. A synantocyte (labeled with antibodies to NG2, green) extends process along myelin sheaths (labeled with antibodies to myelin basic protein, blue) to form contacts multiple nodes of Ranvier (labeled with antibodies to ankyrin-3G, red). Reproduced from Butt AM, Hamilton N, Hubbard P, Pugh M, and Ibrahim M (2005) Synantocytes: The fifth element. *Journal of Anatomy* 207: 695–706, with permission from Blackwell Publishing.

within a single unit. Oligodendrocytes then go through a remodeling phase in which nonensheathing processes are lost and the number of axons per oligodendrocyte unit is established. The final ratio is greater for type I–II oligodendrocytes than for type III–IV oligodendrocytes, but the factors regulating the phenotypic divergence are unknown. Definitive internodal lengths are established by longitudinal growth during the transition to compact myelin, accompanied by the development of paranodal axoglial junctions and the establishment of nodes of Ranvier. The growth of the myelin sheath (both radially, by increasing the number of lamellae, and longitudinally, by growth of the internode) and the radial growth of axons are interrelated, but the factors controlling these relationships are unresolved.

Nonmyelinating Oligodendrocytes

OPCs that do not form myelin persist in the adult as NG2-expressing synantocytes, which are equally numerous in CNS white and gray matter. In addition, there are satellite oligodendrocytes in the gray matter, although their developmental relations to myelinating oligodendrocytes are not clear. Synantocytes and satellite oligodendrocytes are distinguished from astrocytes by their lack of cytoplasmic intermediate filaments and from myelinating oligodendrocytes by the lack of expression of myelin proteins.

Synantocytes (Adult Oligodendrocyte Progenitor Cells)

During development, OPCs give rise to both myelinating oligodendrocytes and synantocytes, which retain the antigenic phenotype of OPCs. Synantocytes, or at least a subpopulation of them, can regenerate myelinating oligodendrocytes following demyelination. The distribution and process outgrowth of synantocytes is directly related to neurons, and they form exquisite contacts with neurons at synapses and nodes of Ranvier (Figure 5). Synantocytes respond to neurotransmitters released during neuronal activity, and they may be specialized to monitor and respond rapidly to changes in neuronal activity. The physiological functions of synantocytes are unknown, but it is known that synantocytes respond rapidly to any insult to the CNS by a stereotypic reactive gliosis, and this may be a primary function of these cells in the adult CNS.

See also: Demyelinating Diseases; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells; Oligodendrocyte Specification; Oligodendrocyte and Schwann Cell Identification Methods.

Further Reading

Barres BA and Raff MC (1994) Control of oligodendrocyte number in the developing rat optic nerve. *Neuron* 12: 935–942.

Bergles DE, Roberts JDB, Somogyi P, and Jahr CE (2000) Glutamatergic synapses on oligodendrocyte precursor cells in the hippocampus. *Nature* 405: 187–191.

Berry M, Ibrahim M, Carlile J, Ruge F, Duncan A, and Butt AM (1995) Axon–glial relationships in the anterior medullary Velum of the adult rat. *Journal of Neurocytology* 24: 965–983.

Bunge RP (1968) Glial cells and the central myelin sheath. *Physiological Reviews* 48: 273–284.

Butt AM, Colquhoun, and Berry M (1994) Confocal imaging of the glial cells in the intact rat optic nerve. *Glia* 10: 315–322.

Butt AM, Hamilton N, Hubbard P, Pugh M, and Ibrahim M (2005) Synantocytes: The fifth element. *Journal of Anatomy* 207: 695–706.

Butt AM, Ibrahim M, Ruge FM, and Berry M (1995) Biochemical subtypes of oligodendrocyte in the anterior medullary velum of the rat as revealed by the monoclonal antibody Rip. *Glia* 14: 185–197.

Butt AM, Ibrahim M, and Berry M (1997) The relationship between developing oligodendrocyte units and maturing axons during myelinogenesis in the anterior medullary velum of neonatal rats. *Journal of Neurocytology* 26: 327–338.

Butt AM and Ransom BR (1989) Visualization of oligodendrocytes and astrocytes in the intact rat optic nerve by intracellular injection of Lucifer yellow and horseradish peroxidase. *Glia* 2: 470–475.

Butt AM, Hornby MF, Ibrahim M, Kirvell S, Graham A, and Berry M (1997) PDGF-α receptor and myeling basic protein mRNAs are

not coexpressed by oligodendrocytes *in vivo*: A double *in situ* hybridization study in the anterior medullary velum of the neonatal rat. *Molecular and Cellular Neuroscience* 8: 311–322.

Dawson RL, Levine JM, and Reynolds R (2000) NG2-expressing cells in the central nervous system: Are they oligodendrocyte progenitors? *Journal of Neuroscience Research* 61: 471–479.

Del Rio-Hortega P (1928) Tercera apotación al conocimiento morfologica e interpretacion funcional de la oligodendroglia. *Memorias de la Real Socieda Española de Historia Natatural Madrid* 14: 5–122.

ffrench-Constant C and Raff MC (1986) Proliferating bipotential glial progenitor cells in adult rat optic nerve. *Nature* 319: 499–502.

Hildebrand C, Remahl S, Persson H, and Bjartmar C (1993) Myelinated fibres in the CNS. *Progress in Neurobiology* 40: 319–384.

Hirano A (1968) A confirmation of oligodendroglial origin of myelin in the adult rat. *Journal of Cell Biology* 38: 637–640.

Kaplan MR, Meyer-Franke A, Lambert S, et al. (1997) Induction of sodium channel clustering by oligodendrocytes. *Nature* 386: 724–728.

Kettenmann H and Ransom BR (eds.) (2005) *Neuroglia,* 2nd edn. Oxford: Oxford University Press.

Levine JM and Card JP (1987) Light and electron microscopic localization of a cell surface antigen (NG2) in the rat cerebellum: Association with smooth protoplasmic astrocytes. *Journal of Neuroscience* 7: 2711–2720.

Mori S and Leblond CP (1970) Electron microscopic identification of three classes of oligodendrocytes and a preliminary study of their proliferatice activity in the corpus callosum of young rats. *Journal of Comparative Neurology* 139: 1–30.

Nishiyama A, Lin X-L, Giese N, Heldin C-H, and Stallcup WB (1996) Co-localization of NG2 proteoglycan and PDGF α-receptor on O2A progenitor cells in the developing rat brain. *Journal of Neuroscience Research* 43: 299–314.

Penfield W (1932) Neuroglia: Normal and pathological. In: Penfield W (ed.) *Cytology and Cellular Pathology of the Nervous System*, vol. 2, pp. 423–479. New York: Hoeber.

Peters A (1964) Observations on the connexions between myelin sheaths and glial cells in the optic nerves of young rats. *Journal of Anatomy* 98: 125–134.

Oligodendrocyte Specification

W D Richardson, University College London, London, UK

© 2009 Published by Elsevier Ltd.

Specification of Oligodendrocyte Precursor Cells in the Embryonic Neural Tube

All the neurons and glial cells of the central nervous system (CNS) are derived from the neuroepithelial progenitor cells that form the walls of the embryonic neural tube. The neuroepithelial cell layer (ventricular zone (VZ)) of the spinal cord and brain stem can be subdivided along the dorsal–ventral axis into 11 distinct microdomains, defined by expression of different combinations of transcription factors (Figure 1). These progenitor domains generate different types of neurons followed by glial cells (oligodendrocytes or astrocytes). Fully differentiated glia are not formed directly from the neuroepithelium but via intermediate glial precursors. During or after neuronogenesis, the neuroepithelial progenitors give rise to so-called radial glial cells, which express characteristic gene products such as the glutamate transporter GLAST and the radial cell antigen RC2. Radial glia subsequently generate dedicated astrocyte or oligodendrocyte precursors. In a given domain of the VZ, radial glia produce either mainly astrocytes or mainly oligodendrocytes. For example, the pMN domain of the ventral spinal cord (Figure 1) generates motor neurons followed by oligodendrocyte precursors (OPCs) but few astrocytes. Overall, approximately 80% of spinal cord oligodendrocytes are derived from ventrally derived OPCs, mainly pMN, and the remaining 20% from OPCs that originate in dorsal domains dP3–5 (Figure 1).

Visualizing Oligodendrocytes and Their Precursors *In Situ*

From the earliest stages of their development, migratory OPCs can be visualized by immunohistochemistry or *in situ* hybridization against a characteristic set of markers, including the platelet-derived growth factor receptor (α subunit) (PDGFRA), the proteoglycan NG2, the transcription factor SOX10, and, in chicken only, surface antigens recognized by monoclonal antibody O4. In rodents, O4 antigen does not appear until a later, premyelinating stage (called pro-oligodendrocytes) when the precursors have stopped migrating and are beginning to differentiate into oligodendrocytes. In both birds and rodents, actively myelinating oligodendrocytes can be easily recognized by the expression of proteins (or corresponding mRNAs) that are abundant structural components of myelin, such as myelin basic protein and myelin proteolipid protein.

Using these and other molecular markers, it has been shown that OPCs are first generated in the pMN progenitor domain of the ventral spinal cord at 12.5 days postfertilization (E12.5) in mice, E14 in rats, and E6 or E7 in chick. In human spinal cord, OPCs appear on or before 45 days postfertilization. The following discussion, unless otherwise stated, refers to mouse. After the OPCs first appear at the ventricular surface, they proliferate and migrate throughout the cord, becoming evenly distributed through the gray matter and developing white matter before birth. Starting at approximately E18 (birth is approximately E19 in mice), oligodendrocyte differentiation starts in the ventral white matter and continues for approximately 6 weeks, peaking at approximately postnatal days 10–14. In human spinal cord, myelination of corticospinal tracts continues into the teenage years.

Spatial Control of OPC Generation by Signals from the Ventral and Dorsal Midline

In the ventral spinal cord, specification of both motor neurons and OPCs depends on signaling molecules that are secreted from the notochord and floor plate at the ventral midline, notably the secreted protein Sonic hedgehog (SHH). If chicken embryos are treated *in ovo* with cyclopamine, an inhibitor of SHH signaling, then the appearance of OPCs in the ventral spinal cord is blocked. Conversely, supplying an ectopic source of SHH (e.g., by implanting SHH-expressing cells or fragments of notochord adjacent to the spinal cord) induces extra OPCs in misplaced positions. Members of the bone morphogenetic protein (BMP) family of signaling molecules are also involved in limiting the spatial extent of OPC generation in the ventral cord. BMPs (notably BMP2 and BMP4) are expressed in the dorsal spinal cord and exert long-range influences along the dorsal–ventral axis. This has been demonstrated by surgically removing the BMP-expressing domain *in ovo* or by transplanting cells that express the competitive BMP inhibitor Noggin next to the chick spinal cord. Both these manipulations caused the OPC-producing pMN domain to expand dorsally. The extra OPCs were generated at the expense of astrocytes, which are normally produced in the p2 progenitor domain, immediately dorsal to pMN. Therefore, it seems that the

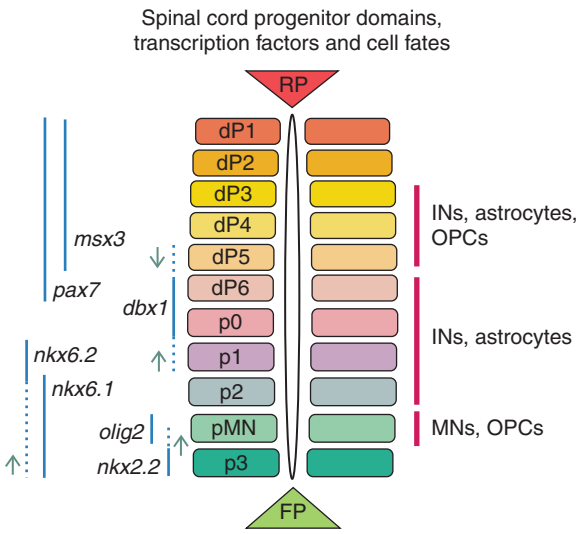

Figure 1 Diagram of progenitor domains in the VZ of the embryonic (~E11) mouse spinal cord. Domains are known as p3, pMN, etc. in the ventral half of the cord and dP6, dP5, etc. in the dorsal half. To the left of the diagram are shown the expression limits of transcription factors, some of which are mentioned in the text. Dotted lines with small arrows indicate that the expression domain expands or shrinks after it is first established. On the right are shown the cell fates of progenitors in the indicated domains (the origins of astrocytes are still tentative). FP, floor plate; INs, interneurons; MNs, motor neurons; OPCs, oligodendrocyte precursor cells; RP, roof plate.

spatial extent of OPC and astrocyte production in the ventral cord is limited by mutually antagonistic actions of SHH and BMPs.

The majority (approximately 80%) of OPCs in the spinal cord are generated in the ventral VZ. The remainder are generated in other progenitor domains, including dorsal domains dP3–5. The dorsally derived OPCs appear later in development than the pMN-derived OPCs (approximately E15 vs. E12.5) and they migrate less widely than their ventrally derived counterparts. It seems unlikely that these dorsal progenitor domains are under the influence of SHH from the floor plate, suggesting that there might be a SHH-independent route to OPC specification. Indeed, OPCs can arise in cultures derived from SHH null spinal cord or in the presence of cyclopamine, if fibroblast growth factor-2 (FGF-2) is also present in the culture medium. OPC generation in the dorsal spinal cord might therefore depend on FGF signaling, possibly combined with a decline in BMP expression in the dorsal cord during late embryogenesis. There might even be a biochemical overlap between SHH and FGF signaling because it has been shown that both pathways depend on MAP kinase activity.

It is not known whether or not ventrally and dorsally derived OPCs in the spinal cord are functionally specialized. However, there is evidence that ventrally and dorsally derived OPCs in the forebrain can substitute for one another, implying that they are functionally equivalent.

OPC Generation in the Forebrain

The adult forebrain, including the cerebral cortex, develops from an embryonic structure called the telencephalon. Like the spinal cord, this starts as a simple neuroepithelial tube, although it becomes progressively more convoluted during development. There is no notochord underlying the telencephalon and no floor plate; however, the ventral neuroepithelial cells express SHH and its receptors, Patched (PTC) and Smoothened (SMO). At approximately E13.5, some cells in the ventral neuroepithelium (medial ganglionic eminence (MGE)) start to express OPC markers such as PDGFRA and SOX10 and migrate away from their origin through the ventral telencephalon. Generation of these early OPCs depends on SHH because they do not appear in embryos that lack SHH expression in the ventral telencephalon (*Nkx2.1* null mice). Later, other parts of the VZ generate OPCs in a temporal wave of production from ventral to dorsal. OPCs from the MGE and lateral ganglionic eminence (LGE) initially populate the developing forebrain including the cerebral cortex before birth, to be joined after birth by OPCs that originate within the cortex. The latter OPCs remain in the cortex and do not migrate ventrally.

The question of whether all these OPCs and the oligodendrocytes that they give rise to are functionally specialized, or equivalent, was addressed by killing specific subpopulations of OPCs at source. This was achieved by targeting diphtheria toxin A-chain expression to subpopulations of OPCs that originate in different domains of the telencephalic neuroepithelium (MGE, LGE, and cortex) by combinatorial use of an oligodendrocyte lineage-specific gene promoter (*Sox10*) and one of several promoters that are active in different parts of the VZ (*Nkx2.1*, *Gsh2*, or *Emx1*). These experiments showed that when OPCs originating within the cortical VZ (normally approximately 50% of all OPCs in the cortex) were ablated, the remaining ventrally derived populations expanded to make up the loss and the animals survived and lived a normal life span. Even when all telencephalic OPCs from MGE, LGE, and cortex were ablated simultaneously, they were replaced by OPCs that migrated forward from the diencephalon. A normal OPC cell density was restored – with a significant but ultimately harmless delay – indicating that OPCs from different parts of the embryonic neuroepithelium are not intrinsically different from one another despite the very different signaling environments in which they arise.

Of course, there might be subtle differences that would not be detected without detailed behavioral analysis of the ablated mice, but this remains to be investigated.

Role of Transcription Factors in OPC Development: The OLIG Genes

A major step forward in understanding the molecular control of oligodendrocyte lineage development resulted from the discovery of transcription factors that orchestrate OPC specification and differentiation. Prime among these are the oligodendrocyte lineage (OLIG) transcription factors, OLIG1 and OLIG2. These are members of the large family of basic helix–loop–helix factors that also includes the pro-neural proteins NGN1/2 and MASH1 and the cell lineage regulators MYOD and NEUROD. In the developing spinal cord, SHH induces expression of OLIG2 in pMN, prior to and during motor neuron (MN) production. OLIG2 is downregulated rapidly in postmitotic MNs but remains on in OPCs as they proliferate and migrate away through the parenchyma. OLIG2 is required for both MN and OPC specification because both cell types are lost in *Olig2* null spinal cords. Pockets of OPCs persist in the brains of *Olig2* null mice, but no OPCs are found anywhere in the CNS of *Olig1/2* compound nulls. Therefore, OLIG1 might partly compensate for loss of OLIG2 in the brain. There is no effect on OPC generation in either the brain or the spinal cord of *Olig1* null mice, although there can be severe myelination defects later. OLIG1 therefore seems to be mainly involved in the later stages of oligodendrocyte differentiation and myelination.

Note that different laboratories have reported different experiences with *Olig1* knockout mice, depending on whether or not the *PGK-Neo* cassette used to inactivate the locus was removed prior to establishing mouse lines. When the *PGK-Neo* cassette was left in place, there was no obvious dysmyelinating phenotype during development, possibly because of *cis*-acting effects of the *PGK-Neo* transcription unit on the nearby *Olig2* gene. However, even in the latter mice there was a striking effect on remyelination of adult mice that were subjected to gliotoxin-induced experimental demyelination; such demyelinated lesions are rapidly repaired in wild-type mice but remyelination was blocked in the *Olig1* knockouts.

The underlying mechanism by which OLIG2 governs the sequential production of MNs and OPCs ('neuron–glial switch') is not known. Presumably, OLIG2 switches binding partners (transcriptional cofactors) during the transition from MN to OPC production. During the period of MN production, OLIG2 and the homeodomain transcription factor NKX2.2 are mutually repressive so that the OLIG2 and NKX2.2 expression domains (pMN and p3, respectively) are sharply demarcated (Figure 2). Later, after MN production has ceased, the cross-repression seems to relax because NKX2.2 expression creeps dorsally into pMN and an overlap region develops. Initially, it was thought that OPCs might be generated specifically within this overlap region under the cooperative activities of OLIG2 and NKX2.2. In chick, it does appear that PDGFRA^{+} OPCs are formed exclusively within the overlap region, but in mice OPCs arise in all parts of the OLIG2-expressing pMN domain, not just the region of overlap with NKX2.2 (Figure 2). Moreover, OPC specification is not affected in *Nkx2.2* null mice,

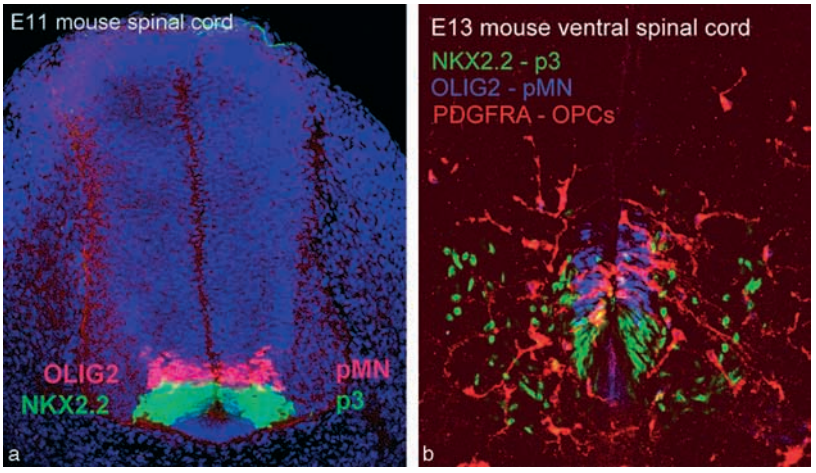

Figure 2 (a) Immunolabeling of ventral progenitor domains p3 and pMN in the mouse spinal cord with antibodies against transcription factors NKX2.2 and OLIG2, respectively. Motor neurons are still being generated at E11. (b) Three-color immunolabeling of NKX2.2 (p3, green), OLIG2 (pMN, blue), and PDGFRA (newly forming OPCs, red) in the mouse ventral spinal cord at E13. The first OPCs are formed mainly within the OLIG2-expressing pMN domain in mouse, outside the NKX2.2-expressing p3 domain. After they are formed at the ventricular surface, the OPCs migrate away rapidly in all directions to populate the spinal cord.

although the later stages of oligodendrocyte differentiation and myelination are delayed. A similar delay in myelination is observed in mice that lack the high mobility group transcription factor SOX10, which is known to interact with OLIG1 and OLIG2. OLIG2 has also been shown to interact physically with NKX2.2. Therefore, a picture is emerging in which OLIG1/2, NKX2.2, and SOX10 cooperate in controlling myelin gene expression and axon ensheathment. OLIG2 and other members of the SOX family (including SOX8 and SOX9) are required earlier for OPC and/or MN specification, but so far there is no compelling explanation for the neuron–glial switch.

Notch–Delta signaling is clearly involved in the MN/OPC fate choice as in many other binary decisions during development. It has been shown that, in zebra fish, abrogation of Notch signaling results in excess production of MNs at the expense of OPCs. Conversely, constitutive activation of Notch signaling results in excess OPCs at the expense of MNs. A well-established general function of Notch–Delta signaling is to maintain precursor cells in an immature state and to inhibit premature cell cycle exit and differentiation, thus preserving part of the precursor pool for generation of later-born cell types. Since MNs are formed before OPCs, the gain- and loss-of-function experiments in zebra fish spinal cord can be interpreted in those terms – implying that Notch signaling might play a permissive rather than an instructive role in the MN/OPC fate choice.

Control of OPC Proliferation

After they are specified in the embryonic VZ, OPCs proliferate and migrate away from their sites of origin, distributing widely and uniformly throughout the CNS before associating with axons and differentiating into myelin-forming oligodendrocytes (Figure 3). Several mitogenic polypeptides have been shown to influence OPC proliferation in culture and/or *in vivo*, including PDGF, FGF, and cytokines. These probably act in concert with one another, possibly in different combinations in different regions of the CNS and/or at different stages of OPC development.

OPCs express receptors for PDGF (PDGFRA subtype), and many neurons and astrocytes synthesize PDGF-A. *PDGF-A* null mice have reduced OPC numbers (approximately 10% normal numbers in the spinal cord and approximately 50% in the cerebral cortex) and correspondingly reduced amounts of myelin. Therefore, homodimeric PDGF-AA is an essential mitogen for OPCs *in vivo*. PDGF-AA is also strongly mitogenic for OPCs in mixed neural cell cultures, such as mixed glial cell cultures derived from perinatal rat optic nerve. However, PDGF-AA is not very mitogenic for pure, immunoselected populations of OPCs cultured on their own in the absence of other cell types. There are other essential mitogens or mitogenic cofactors released into the extracellular medium of mixed cell cultures that are required to cooperate with PDGF-AA. One of these is the CXC cytokine GRO-alpha. Astrocytes are a major source of GRO-alpha in optic nerve cultures and probably also *in vivo*. Cell adhesion molecules, notably integrin family members, also synergize with PDGF and possibly other polypeptide mitogens to drive OPC proliferation at limiting mitogen concentrations.

FGF-2 also enhances the mitogenic effect of PDGF-AA for OPC glial cell cultures. OPCs cultured with both PDGF-AA and FGF-2 can proliferate seemingly indefinitely without differentiating, whereas PDGF-AA in the absence of FGF allows limited OPC proliferation followed by oligodendrocyte differentation. FGF-2 on its own in the absence of PDGF is mitogenic for a late stage of OPC development, when OPCs stop migrating and express the O4 antigen prior to

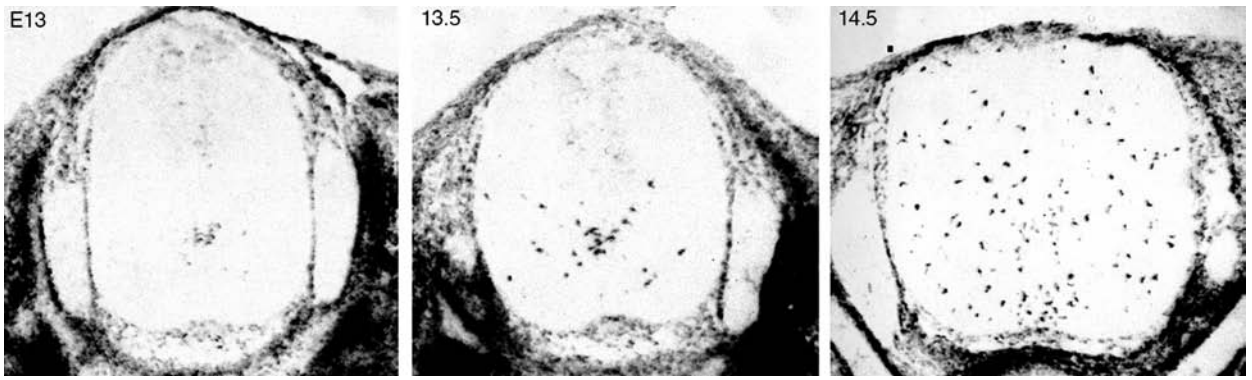

Figure 3 OPCs, visualized by *in situ* hybridization for *Pdgfra* mRNA, in the developing mouse spinal cord. The first OPCs appear in pMN at approximately E12.5, and then they proliferate in response to PDGF-AA and other mitogenic cofactors and migrate throughout the cord, becoming widely distributed by approximately E15. They first start to differentiate into oligodendrocytes in the ventral and dorsal axon tracts soon before birth (E19 in mice).

terminal oligodendrocyte differentiation – so-called 'pro-oligodendrocytes.' Studies on the role of FGF *in vivo* are complicated by the large number of potential ligands (>20 FGF family members) and the fact that OPCs express different FGF receptor subtypes (FGFR1–3) at different stages of their development. Nevertheless, retrovirus-mediated inhibition of FGF-2 activity *in vivo* has shown that FGF-2, acting through FGFR1, negatively regulates OPC terminal differentiation into oligodendrocytes during normal development and after experimental demyelination in adult mice.

Other signaling systems that have been implicated in the control of OPC proliferation include neurotrophin-3, neuregulin-1/glial growth factor (NRG1/GGF), and insulin-like growth factor-1. There is also a growing awareness of the importance of ion channels and neurotransmitters in OPC proliferation and differentiation control.

Control of OPC Differentiation

After OPCs have disseminated throughout the CNS and into the future gray and white matter, something triggers them to differentiate into oligodendrocytes. One key signal seems to be the thyroid hormone triiodothyronine (T3), a constituent of the defined culture medium commonly used for primary nerve cells. When T3 is omitted from the culture medium, OPCs divide for an extended period in response to PDGF-AA, well beyond their normal limit in the presence of T3. It appears that T3 is required for withdrawal from the cell cycle prior to terminal differentiation. This effect of T3 can be mimicked by glucocorticoids or retinoic acid.

There is an established literature on the role of thyroid hormone (TH) in brain development and myelination. The start of myelination is delayed in hypothyroid rats and accelerated in hyperthyroid rats or rats that receive postnatal injections of T3. Moreover, TH normally only becomes available in the rat when the thyroid gland becomes active after birth, which is approximately the time when myelin first starts to appear. Oligodendrocytes and OPCs possess receptors for T3, so it is likely that T3 acts directly on OPCs to control their differentiation *in vivo*.

Notch signaling is also implicated in the control of oligodendrocyte differentiation. OPCs express Notch-1, and adding the Notch ligand Jagged to optic nerve cell cultures inhibits OPC differentiation into oligodendrocytes *in vitro*. Jagged is expressed on optic nerve axons *in vivo*, suggesting that differentiation of OPCs in the optic nerve might be triggered in part by downregulation of Jagged in the retinal ganglion cells that project their axons through the nerve. In support of this general idea, conditional deletion of Notch-1 in OPCs in transgenic mice leads to premature expression of oligodendrocyte differentiation markers such as the myelin-associated glycoprotein.

Control of Myelination by Neuregulin-1/Glial Growth Factor

Not all axons are myelinated: Axons must reach a certain minimum size (diameter) threshold before myelination is triggered. Below this threshold the axon remains naked. This is true in both the CNS and the peripheral nervous system (PNS). In addition, there is a direct relationship between the diameter of a myelinated axon and the thickness of the myelin sheath that develops around that axon (i.e., larger diameter axons have more myelin wraps). The myelin-modulating signal resides on the axonal surface and is interpreted by the myelinating glia (oligodendrocytes or Schwann cells) which adjust their synthesis of myelin membrane accordingly.

Neuregulins are a family of signaling proteins encoded in three loci, *Nrg1–3*, each of which generates alternatively spliced products that can be either secreted or membrane bound. They bind to receptors of the ErbB family (ErbB1–4), which includes the epidermal growth factor receptor ErbB1. NRG1 type III, also known as GGF, is now known to be the main axon-bound regulator of myelination in the PNS, acting through ErbB receptors on Schwann cells. The expectation that one or more NRG isoforms might also regulate myelinogenesis in the CNS is supported by the finding that expressing a dominant negative ErbB transgene in oligodendrocytes results in thinner myelin sheaths and associated physiological deficits, including reduced conduction velocity and, unexpectedly, increased dopaminergic activity of neurons. The precise NRG isoform responsible is not known. These initial results are exciting because of the implied links between myelin deficiency, neuronal activity, and, potentially, psychiatric illness.

OPCs in the Adult CNS

Myelination continues for at least 6 weeks postnatally in mice and rats, peaking 2 or 3 weeks after birth. However, cells with the antigenic phenotype and morphology of OPCs persist in large numbers in the adult CNS (approximately 4% of all cells). Like their perinatal counterparts, they continue to express both PDGFRA and the NG2 proteoglycan and retain the capacity to generate oligodendrocytes in culture or following experimentally induced demyelination. However, because there has been no compelling reason to think that significant numbers of new

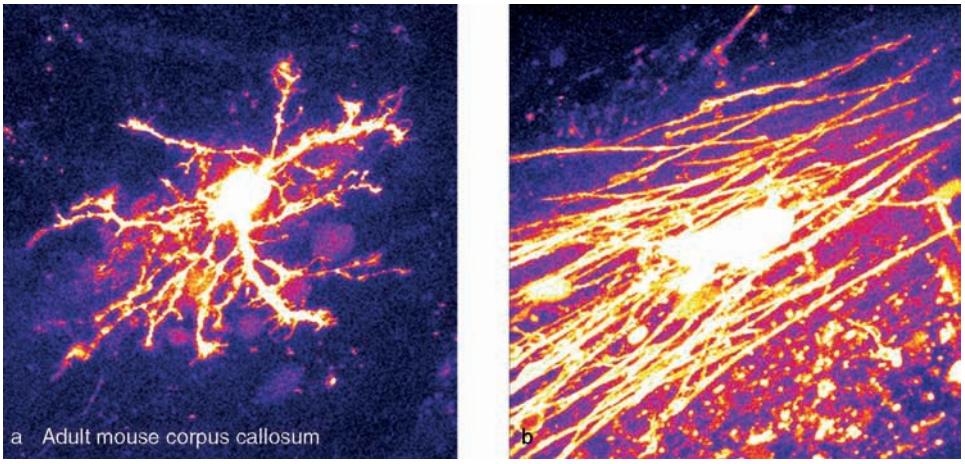

Figure 4 Individual oligodendrocyte lineage cells labeled *in situ* (in 300-μm live sections of postnatal mouse corpus callosum) by microinjection of fluorescent Alexa dye. The myelinating oligodendrocyte (b) was formed between 6 and 8 weeks after birth from a PDGFRA-expressing adult OPC/NG2 cell (a). The long rod-like processes in b are myelin sheaths ('internodes') around nerve axons. Micrographs courtesy of M Rizzi.

oligodendrocytes are needed during normal healthy life, why large numbers of OPCs should survive in the adult has been puzzling.

It has been recognized that OPCs in the adult make contact with nodes of Ranvier (the gaps between adjacent myelin sheaths on an axon), they receive synaptic input from neurons, and they can even fire spontaneous action potentials. This has led to the notion that OPCs participate in the normal physiology of the adult CNS and that their primary role is not necessarily as glial precursors. They have been called 'a fourth glial cell type' (after astrocytes, oligodendrocytes, and microglia), 'synantocytes,' 'polydendrocytes,' or, more commonly, 'NG2 cells.' The physiological role of these cells is under intense scrutiny.

We should not reject the idea that NG2 cells might be needed to generate new or replacement oligodendrocytes throughout life. Cre-lox fate mapping in adult mice (using tamoxifen-inducible *Pdgfra-CreER*^{T2}) shows that they continue to divide and generate significant numbers of new myelinating oligodendrocytes in the corpus callosum of adult animals (Figure 4). Whether this ongoing oligodendrogenesis occurs to replace cells that die through natural turnover or to add extra oligodendrocytes to the existing pool is not known. It is now accepted that new neurons are generated in some parts of the adult brain throughout life; for example, new olfactory interneurons are continuously generated from stem cells that reside in the adult subventricular zones of the forebrain. It is conceivable that some adult-born neurons might need to be myelinated, and this could be one important function of the adult NG2 cells. Alternatively, or in addition, the axonal diameters of some previously unmyelinated neurons might increase during life, taking them over the threshold for *de novo* myelination.

Finally, NG2 cells might have more general cell fate potential. It is known that OPCs in perinatal optic nerve cell cultures can give rise to neurons, astrocytes, and oligodendrocytes if cultured in an appropriate manner. It is therefore possible, though not yet demonstrated, that OPCs in the adult CNS might be able to generate new neurons or astrocytes as well as oligodendrocytes. If so, adult OPCs/NG2 cells would qualify as a type of adult neural stem cells.

See also: Drosophila Apterous Neurons: from Stem Cell to Unique Neuron; Motor Neuron Specification in Vertebrates; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells; Oligodendrocyte Morphology; Oligodendrocyte and Schwann Cell Identification Methods.

Further Reading

Barres BA and Raff MC (1994) Control of oligodendrocyte number in the developing rat optic nerve. *Neuron* 12: 935–942.

Coman I, Barbin G, Charles P, Zalc B, and Lubetzki C (2005) Axonal signals in central nervous system myelination, demyelination and remyelination. *Journal of Neurological Science* 233: 93–97.

Karadottir R and Attwell D (2007) Neurotransmitter receptors in the life and death of oligodendrocytes. *Neuroscience* 145: 1426–1438.

Miller RH (2002) Regulation of oligodendrocyte development in the vertebrate CNS. *Progress in Neurobiology* 67: 451–467.

Nishiyama A (2007) Polydendrocytes: NG2 cells with many roles in development and repair of the CNS. *Neuroscientist* 13: 62–76.

Polito A and Reynolds R (2005) NG2-expressing cells as oligodendrocyte progenitors in the normal and demyelinated adult central nervous system. *Journal of Anatomy* 207: 707–716.

Raff M (2006) The mystery of intracellular developmental programmes and timers. *Biochemical Society Transactions* 34: 663–670.

Richardson WD (2001) Oligodendrocyte development. In: Jessen KR, and Richardson WD (eds.) *Glial Cell Development*, 2nd edn., pp. 21–54. Oxford: Oxford University Press.

Richardson WD, Kessaris N, and Pringle NP (2006) Oligodendrocyte wars. *Nature Reviews Neuroscience* 7: 11–18.

Rowitch DH (2004) Glial cell specification in the vertebrate neural tube. *Nature Reviews Neuroscience* 5: 409–419.

Wegner M and Stolt CC (2005) From stem cells to neurons and glia: A Soxist's view of neural development. *Trends in Neurosciences* 28: 583–588.

Olivocerebellar System

R R Llinas and K D Walton, New York University Medical School, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Morphology

The olivocerebellar system comprises the bilateral inferior olivary nuclei (IO), located in the ventral aspect of the brain stem, and the overlying cerebellum. These two structures are linked through the cerebellar peduncles. Axons from the IO neurons (Figure 1(b)) course up the contralateral peduncle, through the cerebellar white matter, into the cerebellar cortex, where they innervate the Purkinje cells (Figure 1(a), PC). On reaching the Purkinje cell layer, the IO axons establish synaptic contact with single Purkinje cells, forming a one-to-one junction at their soma and extend over the length of their primary and secondary dendrites. These IO afferents were thus named 'climbing fibers' by Ramon y Cajal (Figure 1(a), CF). The climbing fiber–Purkinje cell synapse is the largest synaptic junction in the vertebrate central nervous system. Activation of a climbing fiber elicits an all-or-nothing burst of spikes in the PC, and this stereotypic response is called a 'complex spike' (see Figure 2(c)). There are about ten times more PCs than IO neurons and, on average, each IO neuron generates ten climbing fibers.

The PC axons, the only output from the cerebellar cortex, terminate in the cerebellar nuclei, where they form inhibitory synapses (Figure 3(f), blue). In turn, the axons of the cerebellar nuclei are the only output of the cerebellum. The olivocerebellar axons also contact, via collaterals, the cerebellar nuclei neurons. The axons of half of all cerebellar nuclear neurons return to the contralateral olive as an inhibitory system and terminate in the IO onto structures called 'glomeruli' (Figures 1(a) and 1(d)). Each olivary glomerulus contains approximately five to eight spines (Figure 1(d), ST) from distal dendrites of different IO neurons (Figure 1(d), IOD). The glomeruli contain both excitatory and inhibitory synapses of extrinsic origin. As in other central nervous system structures, gap junctions provide the main communication pathway between the IO neurons (Figures 1(c) and 1(d)). Different types of γ-aminobutyric acid-A receptor subtypes have been identified on the soma and dendrites of IO neurons (Figure 3(f), green). The other half of the cerebellar nuclear system projects as excitatory afferents to the brain stem nuclei, spinal cord, and thalamus, the last as part of the cerebrocerebellar system.

The second major input to the cerebellum is provided by the mossy fibers that innervate PCs through the mossy fiber–parallel fiber system (Figure 1(a)). Activation of this pathway elicits simple spikes in PCs.

Electrophysiology

The function of the olivocerebellar system in motor execution seems to be closely related to so-called physiological tremor. This can be easily observed by measuring the velocity of human finger movements (Figure 2(a)). Vallbo and Wessberg reported that oscillations at 8–10 Hz represent movement steps where their rather stereotyped frequency is close to fixed and is independent of movement speed.

Electrophysiologically, inferior olive neurons are characterized by their ability to generate 'low-threshold' and high-threshold calcium spikes (see Figure 2(d)) as well as spontaneous subthreshold membrane potential oscillations near 10 Hz (Figure 2(b)). These oscillations modulate spike initiation, and so 'spontaneous' action potentials are generated and fire at 1–10 Hz (Figure 2(d)).

Very significantly, the phase of subthreshold oscillations may be influenced by activity, as shown in Figure 2(b) from an *in vitro* experiment. Indeed, extracellular stimulation of the nearby white matter may generate full action potential followed by membrane hyperpolarization (Figure 2(b), boxed area). If it is strong enough, such a stimulus will reset the oscillatory phase, but not the amplitude, of subthreshold oscillations in a large proportion of neurons, and being electrically coupled, they tend toward oscillatory entrainment (Figure 2(c)). Dynamically constrained groups of IO neurons oscillating in phase can synchronously activate a population of PCs and thereby control patterns of synchronous activity in the cerebellar output during motor coordination.

This ability of the olivocerebellar system to generate rhythmic activity has been attributed to the intrinsic oscillatory properties of the IO neurons themselves, and their electrotonic coupling contributes to the synchronicity of activation. Several types of voltage-gated calcium and potassium channels, in addition to those underlying fast action potential generation, enable IO cells to oscillate and fire rhythmically. These include a high-threshold Ca^{2+} channel ($Ca_V2.1$), a somatic low-threshold Ca^{2+} channel ($Ca_V3.1$), a Ca^{2+}-activated K^+ channel ($K_Vx.x$), and a hyperpolarization-activated cationic channel ($K_Vx.x$).

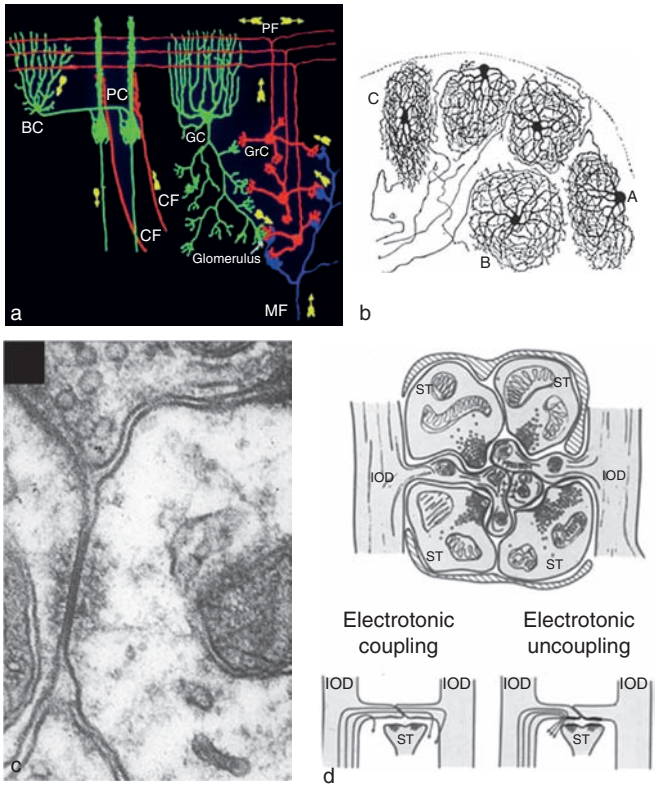

Figure 1 (a) Cerebellar cortex. BC, basket cell; PC, Purkinje cell; CF, climbing fiber; PF, parallel fiber; GrC, granule cell; MF, mossy fiber. (b) Inferior olivary neurons. Note their spherical dendritic tree. (c) Electronmicrograph showing gap junction between spines of IO dendrites within IO glomerulus. (d) Drawing of IO glomerulus. The center shows spines from IO dendrites (IODs) coupled by gap junctions. Bottom left, coupling current path from one IO dendrite to another. Bottom right, current flow is shunted when γ-aminobutyricacidergic synapse is active at the gap junction. (a, b) From Ramon y Cajal S (1888) Sobre las fibras nerviosas de la capa molecular del cerebelo. *Revista Trimestral de Histología Normal y Patológica* 2: 33–41. (c) Adapted from Sotelo C, Llinas R, and Baker R (1974) Structural study of inferior olivary nucleus of the cat: Morphological correlates of electrotonic coupling. *Journal of Neurophysiology* 37: 541–559. (d) Adapted from Llinas R (1974) Eighteenth Bowditch lecture. Motor aspects of cerebellar control. *Physiologist* 17: 19–46.

When IO cells fire, the resulting PC complex spikes (Figure 2(e)) display a close to 10 Hz rhythmicity. In fact, *in vivo* simultaneous recordings from many PCs have demonstrated that groups of PCs fire synchronous complex spikes (see Figure 4). Such synchronous activation of multiple PCs by the IO nucleus suggests that clusters of IO neurons fire rhythmically and in unison as expected from the fact that they are electrically coupled to each other.

In fact, the electrotonic coupling has been proposed to play a crucial role both in synchronizing IO oscillations and in generating groups of concurrently oscillating neurons. Modification of this coupling by γ-aminobutyricacidergic terminals is thought to influence the spatial grouping of the PC complex spike activity over the cerebellar cortex. This dynamic modulation of the geometrical distribution of Purkinje cell activation via the climbing fibers was found to be implemented by the inhibitory cerebellar nuclear inputs to the IO that shunt electrotonic coupling at the IO glomeruli. Such inhibitory input can control the degree and distribution of synchronous oscillatory activity in the IO nucleus and cerebellum by uncoupling gap junctions (Figure 1(c)). Cerebellar nuclear cells in turn are also known to have low-threshold calcium-dependent activity that can be modulated at close to 10 Hz.

Voltage-Dependent Dye Imaging

The spatial profile of synchronized IO neuron oscillations has been studied with voltage-sensitive dye imaging (Figure 5(a)), the method of choice for studying the geometrical distribution of activity in large neuronal ensembles (34). The imaging response comprises an initial change in background fluorescence (induced by direct electrical stimulation) followed by several oscillatory waves that spread throughout the IO nucleus (Figures 5(b) and 5(c)). These ensemble IO oscillations were calcium derived and were also observed in concurrently obtained intracellular recordings from IO neurons.

These optically recorded IO oscillations comprised clusters of coherent neuronal activity in which each

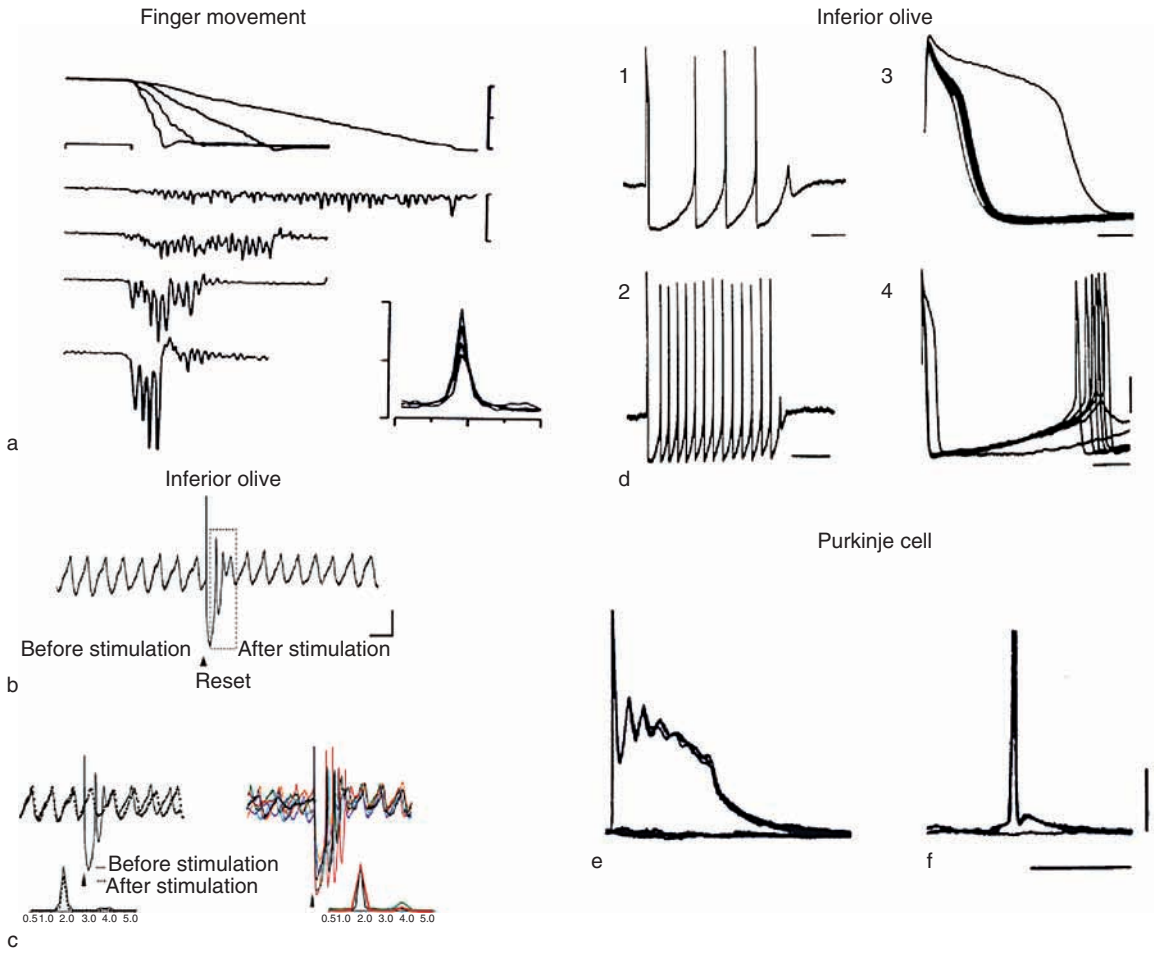

Figure 2 (a) Traces showing different degrees of finger movement. Lower traces show electromyograpic recording of muscle showing gradation of activity with finger movement. Plot shows that the frequency of activity is independent of movement amplitude. (b) Intracellular recording of spontaneous oscillations at 2 Hz interrupted by an extracellular stimulus. After extracellular stimulation (marked with an arrowhead), oscillations disappeared for 750 ms (boxed area) and then resumed. The membrane potential was at −60 mV. (c) Left, intracellular recordings of spontaneous (dashed black line) and stimulus-evoked (solid black line) oscillations from the same cell are superimposed. Their corresponding power spectra are shown below. Note that extracellular stimulation modified only the phase of spontaneous oscillations without affecting its amplitude and frequency. Right, six individual intracellular traces of stimulus-evoked oscillations from the same cell are superimposed. Each trace is shown in a different color. Their corresponding power spectra are displayed below. In every recording, the frequency of stimulation-evoked oscillation was the same (2.0 Hz). Note that in each trace, the stimulation-induced shift in the cell's oscillatory rhythm is remarkably similar. Oscillations are clearly seen after the stimulus-induced reset, but they can barely be detected before stimulation. Calibration bar: 1 mV; (b) 500 ms. Calibration bar: 1 mV; B, 1 s; (c), 500 ms. (d) Intracellular recordings of spontaneous activity in an IO neuron. 1, Four spikes are followed by a subthreshold calcium spike; 2, longer burst of spikes; 3, superposition of spikes in two different times, showing that the spike duration progressively decreased until spike generation failed; 4, same as in 2, showing somatic calcium spikes rise from afterhyperpolarization. (e) Superimposed spontaneous complex spike in Purkinje cell. (f) Superimposed Purkinje cell simple spikes showing synaptic potential. (a) Adapted from Vallbo A and Wessberg J (1993) Organization of motor output in slow finger movements in man. *Journal of Physiology* 469: 673–691. (b, c) Adapted from Leznik E, Makarenko V, and Llinas R (2002) Electrotonically mediated oscillatory patterns in neuronal ensembles: An *in vitro* voltage-dependent dye-imaging study in the inferior olive. *Journal of Neuroscience* 22: 2804–2815. (d) Adapted from Llinas R and Yarom Y (1986) Oscillatory properties of guinea-pig inferior olivary neurones and their pharmacological modulation: An *in vitro* study. *Journal of Physiology* 376: 163–182. (e, f) From Llinas R and Muhlethaler M (1988) An electrophysiological study of the *in vitro*, perfused brain stem-cerebellum of adult guinea-pig. *Journal of Physiology* 404: 215–240.

cluster was a localized functional event (Figure 5(c), second and third image panels). These oscillatory groupings of cells began in several independent regions of the slice, but they had the same oscillation frequency and were phase coherent. Such phase coherence is most clearly observed after the phase reset triggered by an electrical stimulus. These optically recorded oscillations were highly repeatable in

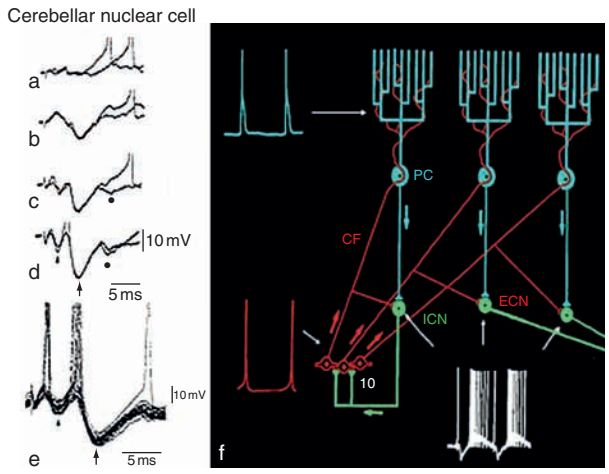

Figure 3 (a–e) Intracellular recordings from cerebellar nuclear neuron in isolated brain preparation showing change in inhibitory postsynaptic potential (IPSP) elicited by white matter stimulation. (a–d) Increasing stimulus strength shows progressive appearance of early excitatory postsynaptic potential (EPSP)–IPSP (arrowhead in (d)) and late EPSP–IPSP (arrow in (d), dots in (c) and (d)). (e) Superimposed responses showing early (arrowhead) and late (arrow) IPSPs. (f) Drawing of olivocerebellar network showing the production of rhythmic activity in the presence of harmaline in Purkinje cells (PC; blue), inferior olivary (IO) cells (red), and cerebellar nuclear cells (green). Rhythmic activity in IO cells is synchronized by gap junctions. This elicits rhythmic and synchronous climbing fiber (CF) activation of PC and collateral activation of inhibitor cerebellar nuclear cells (ICN) and excitatory cerebellar neurons (ECN). PC activation elicits IPSPs in nuclear neurons that are followed by rebound activation. Activation of ECN is transmitted to the rest of the central nervous system through the cerebellar peduncles. Activation of ICN modulates the excitability of IO cells through the γ-aminobutyricacidergic nucleo-olivary pathway. Adapted from Llinas R and Muhlethaler M (1988) Electrophysiology of guinea-pig cerebellar nuclear cells in the *in vitro* brain stem-cerebellar preparation. *Journal of Physiology* 404: 241–258.

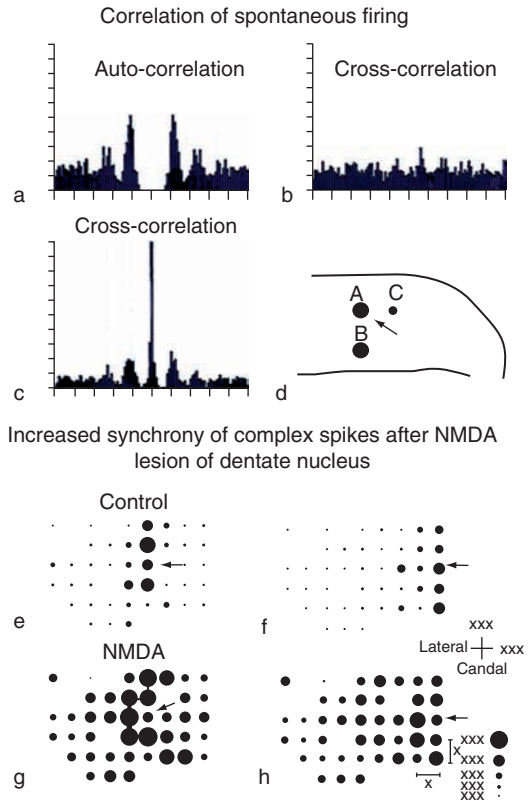

Figure 4 (a–d) Activity in three spontaneously active Purkinje cells recorded *in vivo*. (a) Auto-correlogram showing that a Purkinje cell (cell 'A' in panel (d)) fires at a regular interval close to 10 Hz. (b) Cross-correlogram for activity in cell 'B' with respect to cell 'A.' (c) Cross-correlogram of activity in cell 'C' with respect to cell 'A'. (d) Cell 'A' is the master cell (arrow), cell 'B' is 500 μm caudal, and cell 'C' is 500 μm medial to this cell. (e–h) Increase in climbing fiber synchronicity following lesion of the dentate nucleus in anesthetized rat. Simultaneous recordings from 42 PCs. (e, f) Control recording from left and right vermis, 31.3 min duration. (g, h) A 17.1 min recording after injection of *N*-methyl-D-aspartate (NMDA). (e, g), right Crus 2A, (f, h), left Crus 2a. The area of the dots in (d)–(h) reflects the degree of cross-correlation with the master cell (arrow). (a–d) Adapted from Sasaki K, Bower J, and Llinas R (1989) Multiple Purkinje cell recording in rodent cerebellar cortex. *European Journal of Neuroscience* 1: 572–586. (e–h) Adapted from Lang EJ, Sugihara I, and Llinas R (1996) GABAergic modulation of complex spike activity by the cerebellar nucleoolivary pathway in rat. *Journal of Neurophysiology* 76: 255–275.

time and space; the average of three oscillatory cycles had the same spatiotemporal profile as that of the individual oscillations (compare the second and third image panels in Figure 5(c)).

The frequency of optically recorded clusters, 1.6–6.5 Hz, was within the expected range for intracellularly recorded subthreshold oscillations. The optically recorded oscillatory clusters had a dynamic spatial organization. They changed their size depending on the oscillation phase, so that they embraced the largest area during the upward phase of the oscillations (Figure 5(c), second image panel, third image). Hence, each cluster comprised a core region and the adjoining area. The core region demonstrated a close-to-constant size, but the extent of the adjoining area was phase dependent.

The optical data, therefore, indicate that at the network level, the IO nucleus is organized into functionally coupled activity clusters. Each cluster contains several hundred cells, which may act in unison to simultaneously activate groups of cerebellar PCs, as has been previously observed with multiple electrode recordings. Further, considering the distribution of complex spike activity in the cerebellum cortex, these clusters are very likely responsible for the synchronized activation of the PCs observed in multielectrode experiments (Figure 4). Further, clustering is a direct consequence of the combined electrotonic and intrinsic properties of coupled IO neurons as supported by computational modeling of IO neuronal ensembles.

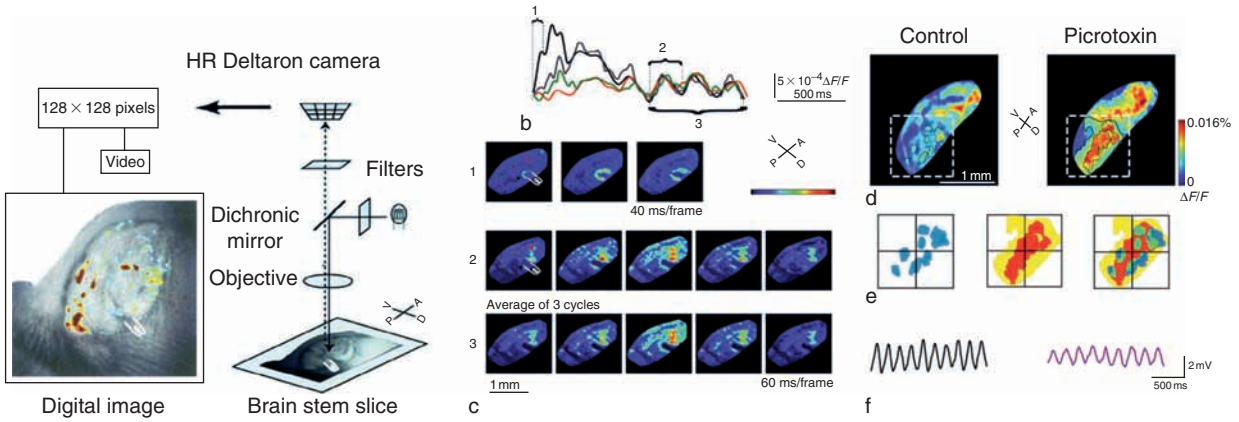

Figure 5 (a) Drawing showing method for voltage-sensitive dye imaging. (b) A complete time course of optical responses for four pixels. Locations of the pixels are marked with circles of different color in (c). The optical response comprised an initial response to a train of extracellular stimuli (two stimuli at 10 Hz) followed by three oscillatory cycles. The frequency of the optically recorded oscillations was 4 Hz. (c) Several multiframe displays for each of the regions marked in (b). Only the area of the inferior olive (IO) is shown. Position of the stimulating electrode is indicated. The color bar gives the color scale, with dark purple corresponding to lowest fluorescence and red corresponding to the highest fluorescence from 0 to 2.7×10^{-3} $\Delta F/F$. 1, Response of the IO nucleus to the first stimulus. 2, Location and spatial spread of one cycle of the ensemble oscillations. The oscillations emanate from several clusters of closely spaced cells throughout the IO (see the change in amplitude with time). These clusters are highly repeatable from cycle to cycle over an oscillation sequence. 3, An average of three oscillatory cycles further underlines the stability of fluorescent clusters. (d) Frames of imaged ensemble neuronal oscillating clusters in control conditions (left) and after the addition of 20 μm of picrotoxin (right) are shown. Several representative clusters in the lower right part of the IO are outlined in black. Clusters were defined as areas with pixel values above the selected threshold level. After picrotoxin, an additional threshold level of higher value was chosen to delineate the highest areas of activity. (e) Clusters' contours shown within the boxed areas (1 × 1 mm) are enlarged for the control condition (right panel) and after the addition of picrotoxin (middle panel). (Left) Areas shown in right and middle panels are superimposed. Control clusters are shown in blue, picrotoxin clusters of lower threshold are shown in yellow, and those of higher threshold are shaded in red. Overlapping regions of control clusters and picrotoxin clusters of lower threshold are shown in green. Note that picrotoxin significantly increased the size of clusters by activating additional regions and merging several small areas into larger activation areas. (f) Intracellular recording from an IO neuron showing spontaneous subthreshold oscillations before (black) and after (purple) the addition of 20 μmol l^{-1} picrotoxin. Adapted from Leznik E, Makarenko V, and Llinas R (2002) Electrotonically mediated oscillatory patterns in neuronal ensembles: An in vitro voltage-dependent dye-imaging study in the inferior olive. *Journal of Neuroscience* 22: 2804–2815.

Previous *in vivo* multielectrode experiments have shown that synchronized complex spikes are primarily observed among PCs located within the same parasagittal bands. Groups of IO neurons that define such parasagittal bands comprise up to several hundred neurons. Because this number is on the same order of magnitude as the calculated number of cells in our fluorescent clusters, it has been proposed that such optical measurements might reflect the ensemble physiological properties formerly observed with multiple recordings at the PC level. Hence, each IO cluster can control activity of a single parasagittal band in the cerebellum. Simultaneous activation of several IO clusters would result in coherent complex spike activity between different parasagittal bands, as observed during coordinated movements.

In conclusion, dimensions of clusters are probably determined by the IO electrical coupling coefficient and therefore by the magnitude and distribution of the return inhibition from the cerebellar nuclear feedback, as has been demonstrated in previous *in vivo* experiments and supported by mathematical modeling.

Modeling Inferior Olive Clusters

As mentioned above, the IO nucleus appears to play a significant role in the temporal binding of organized movement. The anatomical and physiological properties of these neurons are sufficiently well established to warrant the development of quantitative computer modeling of both single cell and network properties. Nonlinear mathematical analysis of experimental data from single neurons (Figure 6) and from coupled pairs of IO neurons indicates that despite being almost perfectly periodic, IO oscillations have a weakly chaotic nature. This property markedly increases the sensitivity of the dynamics independent of input origin. It is surprising that in this neuronal system, the chaotic degree of the dynamics does not increase (i.e., the Lyapunov spectrum remains practically the same) regardless of the number of elements in the system. Indeed, the synchronization itself was found to belong to a very specific type, known as chaotic phase synchronization and observed in a notoriously robust dynamic system.

That the IO nucleus plays a role in movement coordination is shown by its ability to control functional

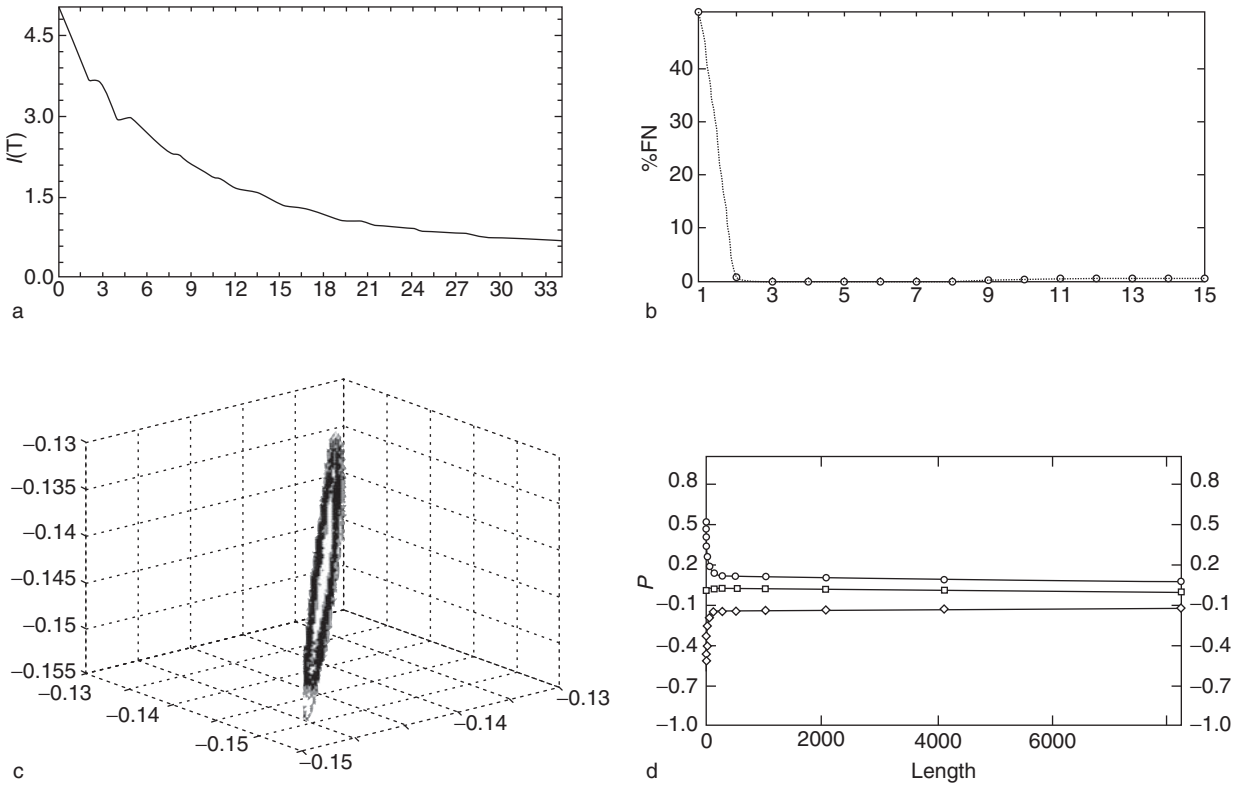

Figure 6 Reconstruction of attractor from a time series of membrane potential for a single neuron. The attractor has a structure close to a limit cycle with a regular periodic trajectory. (a) Average mutual information and (b) false nearest neighbors methods were used to reconstruct (c) the attractor and (d) Lyapunov exponents. Exponent values (small positive, 0, and negative) indicate that the system displays low-dimensional chaotic dynamics. Adapted from Makarenko V and Llinas R (1998) Experimentally determined chaotic phase synchronization in a neuronal system. *Proceedings of the National Academy of Sciences of the United States of America* 95: 15747–15752.

parameters in time, locality, and number of degrees of freedom. It has been proposed that to meet these requirements, the IO uses its intrinsic oscillatory properties for temporal organization, that is, its ability to organize, support, and change cluster size and location to control spatial organization. In addition, its weakly chaotic properties exhibit the necessary fluidity to control the number of degrees of freedom by rapidly modulating the phase relation between clusters.

Results from computer simulation of single neurons and a network of a considerable number of neurons support this view of IO functional dynamics. The simulated network of locally coupled neurons demonstrated that clustering had features similar to those observed in the *in vitro* IO studied in brain stem slices with voltage-sensitive dye imaging. The main conclusion from this group of studies is that the IO is a truly integrated system in which the neuronal network and the macroscopic dynamic features (cluster permutation agility) are so well matched that their control properties almost match cellular time constants. This allows the system to effectively control innumerable movements without overloading brain function.

Implications Concerning Olivocerebellar Motor Control

The synchronous, rhythmic firing of the olivocerebellar system has led to the hypothesis that climbing fibers may perform a timing function in motor coordination. Thus, IO activity would provide a timing signal for the motor system by generating synchronous and rhythmic activation of the cerebellar nuclei. The synchronized inhibitory barrage originated by climbing fiber activation of PC complex spikes thus controls the timing of the output of the cerebellar nuclei. Evidence to support the motor timing proposal includes the following: synchronous firing of a population of PCs in anesthetized and awake animals, recordings of rhythmic inhibitory potential in the deep cerebellar nuclei, and temporal correlation between firing of the olivocerebellar system and execution of movements.

To this effect the IO seems to be dynamically organized in clusters of synchronously oscillating neurons that can generate rhythmic and coherent activity of the cerebellar output. Ultimately, as further insight is gained into the oscillatory property of neurons and their relation to macroscopic motor execution and control, a general message becomes apparent: The dynamic properties of single neurons and the channel kinetics that generate them seem to hone the 'impedance matching' from the molecular to the macroscopic levels to the point that our motor timing is nothing other than the echo of our IO oscillation.

See also: Cerebellar Microcircuitry; Cerebellar Deep Nuclei; Cerebellum: Models; Sporadic Degenerative Ataxias and the Dominantly Inherited Spinocerebellar Ataxias.

Further Reading

Armstrong D and Schild R (1970) A quantitative study of the Purkinje cells in the cerebellum of the albino rat. *Journal of Comparative Neurology* 139: 449–456.

Bal T and McCormick DA (1997) Synchronized oscillations in the inferior olive are controlled by the hyperpolarization-activated cation current *I*(h). *Journal of Neurophysiology* 77: 3145–3156.

Bell C and Kawasaki T (1972) Relations among climbing fiber responses of nearby Purkinje cells. *Journal of Neurophysiology* 35: 155–169.

Benardo L and Foster R (1986) Oscillatory behavior in inferior olive neurons: Mechanism, modulation, cell aggregates. *Brain Research Bulletin* 17: 773–784.

Bennett M (2000) Electrical synapses, a personal perspective (or history). *Brain Research. Brain Research Reviews* 32: 16–28.

Bleasel AF and Pettigrew AG (1992) Development and properties of spontaneous oscillations of the membrane potential in inferior olivary neurons in the rat. *Developmental Brain Research* 65: 43–50.

Crill W (1970) Unitary multiple-spiked responses in cat inferior olive nucleus. *Journal of Neurophysiology* 33: 199–209.

Devor A, Fritschy J, and Yarom Y (2001) Spatial distribution and subunit composition of GABA(A) receptors in the inferior olivary nucleus. *Journal of Neurophysiology* 85: 1686–1696.

De Zeeuw C, Holstege J, Ruigrok T, and Voogd J (1990) Mesodiencephalic and cerebellar terminals terminate upon the same dendritic spines in the glomeruli of the cat and rat inferior olive: An ultrastructural study using a combination of [^{3}H]leucine and wheat germ agglutinin coupled horseradish peroxidase anterograde tracing. *Neuroscience* 34: 645–655.

De Zeeuw C, Simpson J, Hoogenraad C, Galjart N, Koekkoek S, and Ruigrok T (1998) Microcircuitry and function of the inferior olive. *Trends in Neuroscience* 21: 391–400.

De Zeeuw CI, Lang EJ, Sugihara I, et al. (1996) Morphological correlates of bilateral synchrony in the rat cerebellar cortex. *Journal of Neuroscience* 16: 3412–3426.

Eccles JC, Llinas R, and Sasaki K (1966) The excitatory synaptic action of climbing fibres on the Purkinje cells of the cerebellum. *Journal of Physiology* 182: 268–296.

Fredette BJ and Mugnaini E (1991) The GABAergic cerebello-olivary projection in the rat. *Anatomy & Embryology* 184: 225–243.

Fukuda M, Yamamoto T, and Llinas R (2001) The isochronic band hypothesis and climbing fibre regulation of motricity: An experimental study. *European Journal of Neuroscience* 13: 315–326.

Hutcheon B and Yarom Y (2000) Resonance, oscillation and the intrinsic frequency preferences of neurons. *Trends in Neurosciences* 23: 216–222.

Lampl I and Yarom Y (1997) Subthreshold oscillations and resonant behavior: Two manifestations of the same mechanism. *Neuroscience* 78: 325–341.

Lang E (2001) Organization of olivocerebellar activity in the absence of excitatory glutamatergic input. *Journal of Neuroscience* 21: 1663–1675.

Lang E (2002) GABAergic and glutamatergic modulation of spontaneous and motor-cortex-evoked complex spike activity. *Journal of Neurophysiology* 87: 1993–2008.

Lang EJ, Sugihara I, and Llinas R (1996) GABAergic modulation of complex spike activity by the cerebellar nucleoolivary pathway in rat. *Journal of Neurophysiology* 76: 255–275.

Lang EJ, Sugihara I, Welsh JP, and Llinas R (1999) Patterns of spontaneous Purkinje cell complex spike activity in the awake rat. *Journal of Neuroscience* 19: 2728–2739.

Leznik E, Makarenko V, and Llinas R (2002) Electrotonically mediated oscillatory patterns in neuronal ensembles: An *in vitro* voltage-dependent dye-imaging study in the inferior olive. *Journal of Neuroscience* 22: 2804–2815.

Llinas R (1974) Eighteenth Bowditch lecture. Motor aspects of cerebellar control. *Physiologist* 17: 19–46.

Llinas R, Baker R, and Sotelo C (1974) Electrotonic coupling between neurons in cat inferior olive. *Journal of Neurophysiology* 37: 560–571.

Llinas R, Lang E, Welsh J, and Makarenko V (1997) A new approach to the analysis of multidimenshional neuronal activity: Markov Random Fields. *Neural Networks* 10: 785–789.

Llinas R and Muhlethaler M (1988) An electrophysiological study of the *in vitro*, perfused brain stem-cerebellum of adult guinea-pig. *Journal of Physiology* 404: 215–240.

Llinas R and Muhlethaler M (1988) Electrophysiology of guinea-pig cerebellar nuclear cells in the *in vitro* brain stem-cerebellar preparation. *Journal of Physiology* 404: 241–258.

Llinas R and Sasaki K (1989) The functional organization of the olivo-cerebellar system as examined by multiple Purkinje cell recordings. *European Journal of Neuroscience* 1: 587–602.

Llinas R and Yarom Y (1981) Electrophysiology of mammalian inferior olivary neurones *in vitro*: Different types of voltage-dependent ionic conductances. *Journal of Physiology* 315: 549–567.

Llinas R and Yarom Y (1981) Properties and distribution of ionic conductances generating electroresponsiveness of mammalian inferior olivary neurones *in vitro*. *Journal of Physiology* 315: 569–584.

Llinas R and Yarom Y (1986) Oscillatory properties of guinea-pig inferior olivary neurones and their pharmacological modulation: An *in vitro* study. *Journal of Physiology* 376: 163–182.

Makarenko V and Llinas R (1998) Experimentally determined chaotic phase synchronization in a neuronal system. *Proceedings of the National Academy of Sciences of the United States of America* 95: 15747–15752.

Manor Y, Rinzel J, Segev I, and Yarom Y (1997) Low-amplitude oscillations in the inferior olive: A model based on electrical coupling of neurons with heterogeneous channel densities. *Journal of Neurophysiology* 77: 2736–2752.

Ramon Y and Cajal S (1888) Sobre las fibras nerviosas de la capa molecular del cerebelo. *Revista Trimestral de Histología Normal y Patológica* 2: 33–41.

Ruigrok T and Voogd J (1995) Cerebellar influence on olivary excitability in the cat. *European Journal of Neuroscience* 7: 679–693.

Sasaki K, Bower J, and Llinas R (1989) Multiple Purkinje cell recording in rodent cerebellar cortex. *European Journal of Neuroscience* 1: 572–586.

Sotelo C, Llinas R, and Baker R (1974) Structural study of inferior olivary nucleus of the cat: Morphological correlates of electrotonic coupling. *Journal of Neurophysiology* 37: 541–559.

Szentagothai J and Rajkovits K (1959) Ueber den Ursprung der Kletterfasern des kleinhirns. *Zeitschrift für Anatomie und Entwicklungsgeschichte* 121: 130–141.

Thach W (1968) Discharge of Purkinje and cerebellar nuclear neurons during rapidly alternating arm movements in the monkey. *Journal of Neurophysiology* 31: 785–797.

Vallbo A and Wessberg J (1993) Organization of motor output in slow finger movements in man. *Journal of Physiology* 469: 673–691.

Velarde MG, Nekorkin VI, Kazantsev VB, Makarenko VI, and Llinas R (2002) Modeling inferior olive neuron dynamics. *Neural Networks* 15: 5–10.

Welsh JP, Lang EJ, Suglhara I, and Llinas R (1995) Dynamic organization of motor control within the olivocerebellar system [see comment]. *Nature* 374: 453–457.

Welsh JP and Llinas R (1997) Some organizing principles for the control of movement based on olivocerebellar physiology. *Progress in Brain Research* 114: 449–461.

Yamamoto T, Fukuda M, and Llinas R (2001) Bilaterally synchronous complex spike Purkinje cell activity in the mammalian cerebellum. *European Journal of Neuroscience* 13: 327–339.

Operant Conditioning of Reflexes

J R Wolpaw and X Y Chen, New York State Department of Health and State University of New York, Albany, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Spinal cord reflexes are simple behaviors produced by central nervous system (CNS) pathways that lie entirely within the spinal cord. The sensory afferent fibers that evoke these reflexes enter the spinal cord and activate spinal motor neurons directly or through a chain of one or more spinal interneurons. Although these pathways are entirely spinal, they are affected by descending pathways from the brain, either directly or through other spinal interneurons. Through these descending pathways, the brain exerts both short-term and long-term influence over spinal cord reflex function. In the short term, the brain rapidly adjusts spinal reflexes to suit the needs of different tasks (e.g., standing vs. walking vs. running). In the long term, it gradually shapes spinal reflexes during development, during skill acquisition later in life, and in response to CNS trauma and disease. The long-term changes that the brain induces in spinal cord reflexes involve activity-dependent plasticity in the spinal cord itself.

The spinal cord is the simplest and most technically accessible part of the mammalian CNS. Thus, spinal cord reflexes, the brain's influence over them, and the spinal cord plasticity this influence produces provide the basis of a powerful experimental protocol for studying the mechanisms and substrates of learning. By rewarding a subject (whether monkey, rat, mouse, or human) for a larger or smaller reflex, this protocol operantly conditions the subject to maintain descending influence appropriate to the reward contingency (i.e., descending influence that increases or decreases reflex size). In a laboratory analog of the brain's long-term shaping of spinal reflexes during normal life, the continued maintenance of this appropriate descending influence produces plasticity in the spinal pathway of the reflex and thereby changes reflex size appropriately. According to a standard definition of a skill as an adaptive behavior acquired through practice, the larger or smaller reflex is a simple motor skill, and the plasticity directly responsible for it is located in the spinal cord, where it is accessible to detailed physiological and anatomical analyses.

A lengthy series of studies using this protocol has provided important new insight into the remarkably complex spinal and supraspinal plasticity underlying the acquisition and maintenance of a simple motor skill. Those studies also provide the basis for a promising new approach to restoring effective motor function after spinal cord injury or other trauma or disease. This work has focused on operant conditioning of the simplest spinal cord reflex, the spinal stretch reflex (SSR), and its electrical analog, the Hoffmann (H)-reflex. These reflexes are produced largely by a two-neuron monosynaptic pathway consisting of the primary afferent fiber from the muscle spindle, its synapse on the alpha motor neuron, and the alpha motor neuron. This pathway is shown in Figure 1. In the SSR, the primary afferent is excited by muscle stretch; in the H-reflex, it is excited directly by electrical stimulation of the nerve.

The Conditioning Protocol and Its Results

Figure 2(a) shows the standard conditioning protocol as it is implemented in the rat. The monkey, human, and mouse H-reflex and SSR conditioning protocols are similar. The rat is chronically implanted with fine-wire electromyographic (EMG) recording electrodes in the soleus muscle and a nerve cuff on the posterior tibial nerve (which innervates the soleus). The wires connect through a head-mounted tether and a commutator to an amplifier and stimulator. The tether allows the rat to move freely about the cage and remains connected continuously 24 h day^{-1}. Whenever the ongoing background EMG in soleus remains in a defined range for several seconds, a stimulus through the nerve cuff at M response (direct muscle response) threshold elicits the H-reflex. H-reflex size is measured as EMG in the H-reflex interval (typically 6–9 ms after the stimulus). For the first 10–20 days, data are collected in the control mode to determine the initial size of the H-reflex. Then, the rat is switched to the up-conditioning (HRup)-mode or the down-conditioning (HRdown)-mode, in which a food-pellet reward is given when H-reflex size is above (HRup-mode) or below (HRdown-mode) a criterion value.

This protocol has two crucial features. First, soleus background EMG is stable throughout, so that the background, or prestimulus, level of excitation of the soleus motor neuron pool does not change. Second, the H-reflex is elicited at unpredictable times. Thus, because the H-reflex is the earliest possible CNS response to the nerve cuff stimulus, the rat can change H-reflex size only by being prepared ahead of time, that is, by maintaining mode-appropriate descending influence over the wholly spinal pathway of the reflex. It is this continued descending influence that gradually

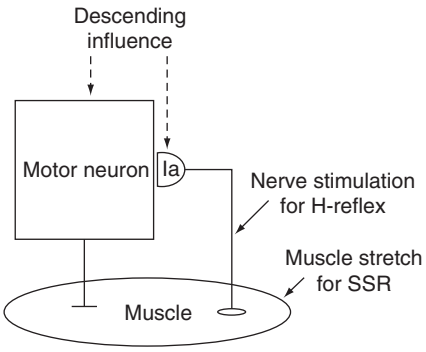

Figure 1 Monosynaptic pathway of the spinal stretch reflex (SSR) and its electrical analog, the H-reflex. The pathway comprises the Ia afferent neuron from the muscle spindle, its synapse on the alpha motor neuron, and the motor neuron itself. When the Ia afferent is excited, it excites spinal motor neurons innervating the same muscle and its synergists. If the afferent is excited in the normal fashion, that is, by muscle stretch, the muscle's response is the SSR. If it is excited by direct electrical stimulation, the response is the H-reflex. The SSR and the H-reflex are measured by electromyograph or by their mechanical effects. While their pathway is wholly spinal, both reflexes are affected by descending influences on the Ia terminal (exerted presynaptically) and on the motor neuron, and the SSR is also affected by descending control of muscle spindle sensitivity. As a result of this descending influence and the plasticity it produces in the spinal cord, the brain can gradually modify these spinal reflexes so as to increase reward. From Wolpaw JR (1997) The complex structure of a simple memory. *Trends in Neurosciences* 20: 588–594.

produces the activity-dependent plasticity in the spinal cord that underlies H-reflex change.

The primary observation is that imposition of the HRup- or HRdown-mode changes reflex size appropriately over days and weeks. Figures 2(b) and 2(c) show the progressive SSR or H-reflex increases and decreases as they occur in rats, monkeys, humans, and mice. Reflex change appears to occur in two phases, a small, rapid phase-1 change in the first few hours or days and a much slower phase-2 change that continues for weeks. In Figure 2, this two-phase course is seen most clearly in the monkey SSR data. Phase 1 probably reflects rapid mode-appropriate change in descending influence over the spinal arc of the reflex. This mode-appropriate change in descending influence is operantly conditioned by the reward contingency. In contrast, phase 2 appears to reflect the gradual spinal cord plasticity produced by the long-term continuation of the descending input responsible for phase 1. This spinal cord plasticity survives for some days after removal of all descending influence. This indicates that the conditioning changes the spinal cord.

The Complex Spinal Cord Plasticity Accompanying Reflex Conditioning

H-reflex conditioning is associated with plasticity at multiple sites in the spinal cord. Downconditioning produces a positive shift in motor neuron firing threshold and a reduction in axonal conduction velocity. These two changes together suggest that a positive shift in sodium channel activation voltage occurs throughout the motor neuron soma and axon. As illustrated in Figure 3(a), the change in threshold can account in large part for the smaller reflex. Although activity-dependent synaptic plasticity is often assumed to be the basis of learning, the shift in motor neuron threshold produced by down-conditioning seems to be an example of neuronally based (as opposed to synaptically based) learning. The threshold change could also account for the decrease in motor neuron axonal conduction velocity that is associated with H-reflex down-conditioning (Figure 3(b)). In addition to this neuronal plasticity, H-reflex conditioning modifies Ia afferent motor neuron synapses and several other populations of synaptic terminals on the motor neuron. The effects on GABAergic terminals are particularly striking (Figure 3(c)). Conditioning probably changes spinal cord interneurons as well, either those in di- or trisynaptic pathways that can contribute to the H-reflex and/or those that convey descending influence to the motor neuron. At present, it is most likely that up-conditioning is due to change in oligosynaptic primary afferent input to the motor neuron. Thus, down-conditioning and up-conditioning are probably not mirror images of each other but rather have different mechanisms. In addition to this complex plasticity in the spinal cord ipsilateral to the conditioned reflex, H-reflex conditioning also causes plasticity in the contralateral spinal cord. The precise location and nature of this contralateral plasticity are as yet unknown.

Complex Plasticity Is Necessary and Inevitable

Some of the complex plasticity accompanying reflex conditioning appears to account for the behavioral change, that is, the larger or smaller H-reflex, produced by the conditioning protocol. For example, the positive shift in motor neuron threshold after down-conditioning explains why the H-reflex is smaller. Other changes, however, such as that on the contralateral side of the spinal cord, seem to be unrelated to the behavior, a smaller H-reflex that is the goal and most obvious product of the down-conditioning protocol. Nevertheless, these apparently inexplicable changes are probably both necessary and inevitable.

The spinal cord neurons and synapses affected by H-reflex conditioning support many other motor behaviors, such as locomotion. The plasticity underlying H-reflex conditioning (such as a positive shift in motor neuron firing threshold) is likely to disrupt

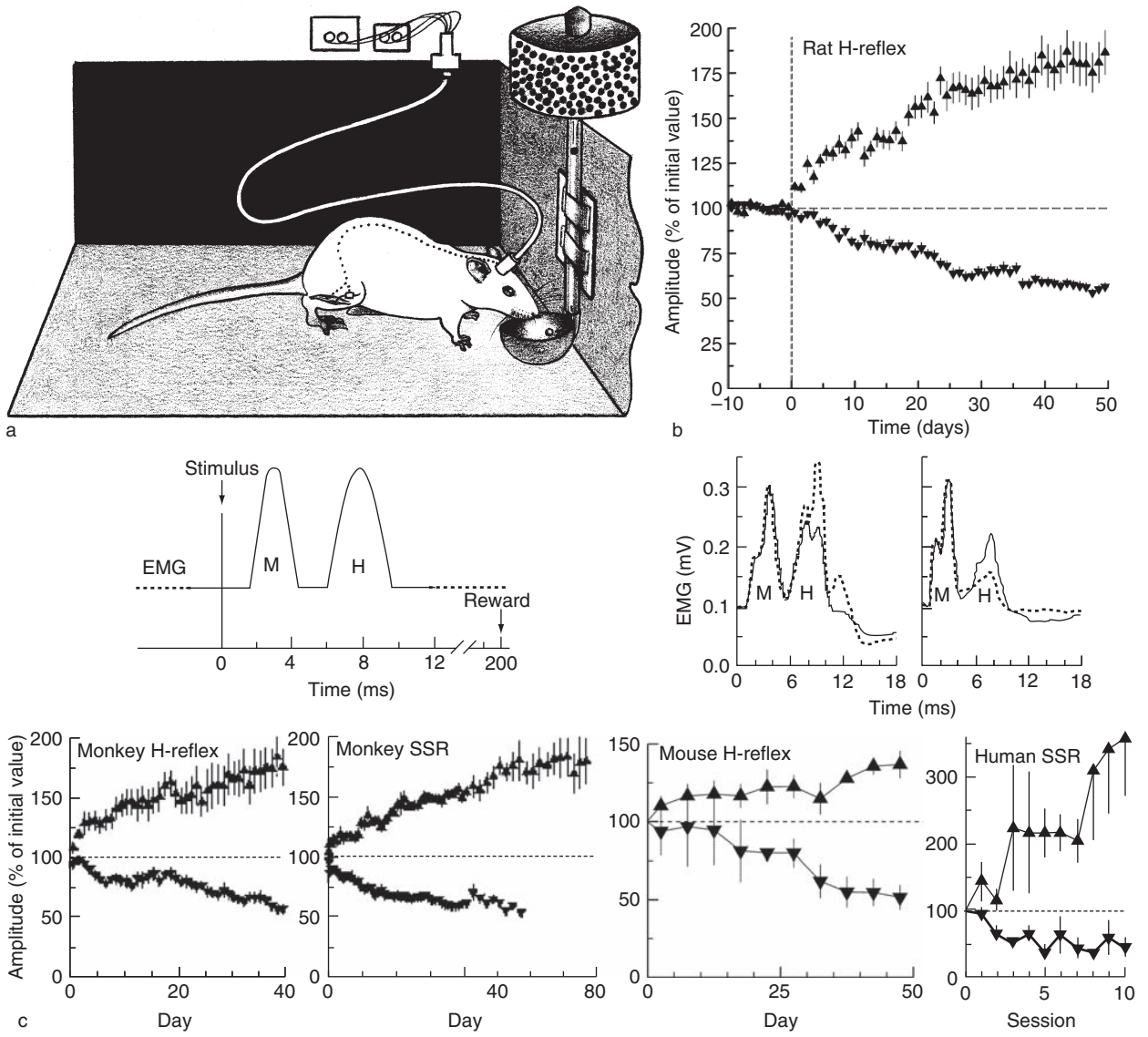

Figure 2 The reflex operant conditioning protocol and its results. (a) Soleus electromyograph (EMG) is monitored continuously (24 h per day) in a rat with chronically implanted EMG electrodes and a tibial nerve cuff. The implant wires pass subcutaneously to a head-mounted connector and then through a flexible cable and a commutator to amplifiers and stimulator. The rat is able to move freely about the cage. Whenever the absolute value of its soleus EMG remains within a specified range for a randomly varying 2.3–2.7 s period, a nerve cuff stimulus elicits a threshold direct muscle (M) response and an H-reflex. For the first 10 days, the rat is exposed to the control mode, in which no reward occurs and the H-reflex is simply measured to determine its initial size. For the next 50 days, the rat is exposed to the up-conditioning or down-conditioning mode, in which a food-pellet reward occurs if the H-reflex is above (up-mode) or below (down-mode) a criterion value. Background EMG and M response remain constant throughout. Successful conditioning, defined as a change of at least 20% in the correct direction, occurs in 75–80% of the rats (the remainder change less than 20%). (b) The top graph shows average daily H-reflexes (± standard error) from successful up-conditioned (▲) and downconditioned (▼) rats for the control-mode exposure (i.e., days 10 to 0) and for the subsequent up-mode or down-mode exposure (i.e., days 0 to 50). Under the up-mode, the H-reflex rises gradually to about 175% of its initial value, while under the HRdown mode, it falls gradually to about 55%. The bottom graphs show average poststimulus EMG for representative days from an up rat (left) and a down rat (right) under the control mode (solid) and near the end of up-mode or down-mode exposure (dashed). The H-reflex is much larger after up-conditioning and much smaller after down-conditioning, while background EMG (indicated here by EMG at zero time) and M responses are unchanged. If a rat is switched from the up-mode to the down-mode or vice versa, the H-reflex change reverses in the same gradual fashion. (c) Average results for up-conditioning (▲) and down-conditioning (▼) of (left to right) monkey triceps surae H-reflex, monkey biceps brachii SSR, mouse triceps surae H-reflex, and human biceps brachii SSR. The courses of change are similar to those found in the rat. Error bars = standard error. Modified from Wolpaw JR (1997) The complex structure of a simple memory. *Trends in Neurosciences* 20: 588–594 and Carp JS, Tennissen AM, Chen XY, and Wolpaw JR (2006) H-reflex operant conditioning in mice. *Journal of Neurophysiology* 96: 1718–1727.

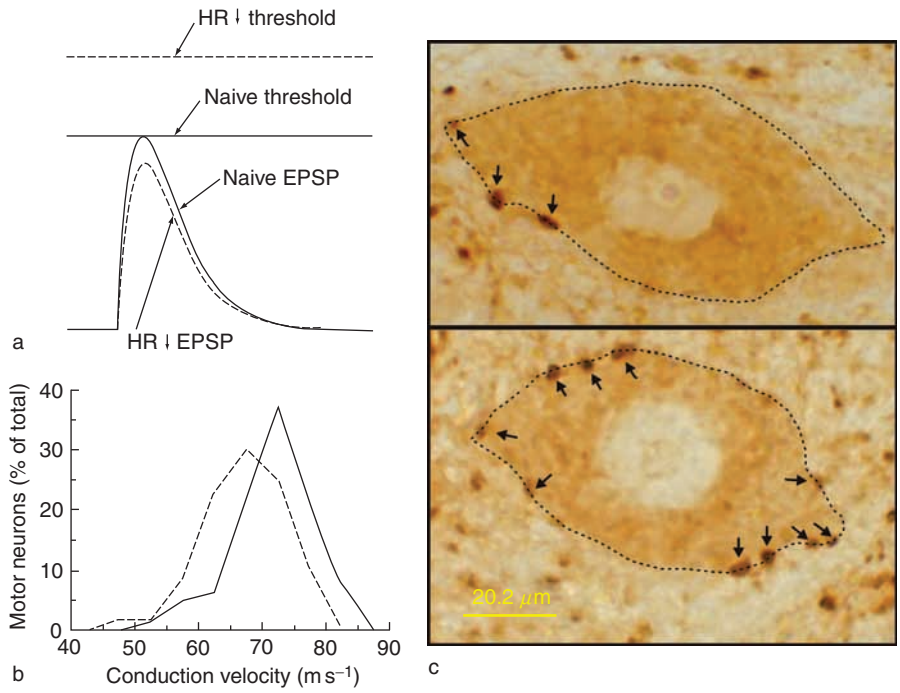

Figure 3 Examples of the multisite spinal cord plasticity underlying H-reflex conditioning. (a) Motor neurons on the conditioned side of H-reflex (HR) down-conditioned animals have more positive firing thresholds and slightly smaller Ia excitatory postsynaptic potentials (EPSPs). Together, these two findings can explain why the H-reflex becomes smaller. (b) Distributions of soleus motor neuron axonal conduction velocities in unconditioned (solid) and down-conditioned (dashed) rats. Down-conditioning has a comparable effect in monkeys. The more positive motor neuron firing threshold can explain this drop in conduction velocity. (c) Soleus motor neurons (dashed lines) from an unconditioned rat (top) and a down-conditioned rat (bottom), with arrows pointing to γ-aminobutyricacidergic (GABAergic) terminals on the somatic membrane. GABAergic terminals are detected by glutamic acid decarboxylase (GAD_{67}) immunoreactivity. After down-conditioning, soleus motor neurons have more GABAergic terminals, and these terminals are more densely labeled and occupy more of the somatic membrane. (a) Modified from Carp JS, Chen XY, Sheikh H, and Wolpaw JR (2001) Operant conditioning of rat H-reflex affects motoneuron axonal conduction velocity. *Experimental Brain Research* 136: 269–273. (b) Modified from Carp JS, Chen XY, Sheikh H, and Wolpaw JR (2001) Operant conditioning of rat H-reflex affects motoneuron axonal conduction velocity. *Experimental Brain Research* 136: 269–273.

these other behaviors. Thus, additional compensatory plasticity may be necessary to preserve these previously acquired behaviors in spite of the disruptive effects of the plasticity that underlies the conditioned reflex change. For example, the primary afferent input that produces the soleus H-reflex is a major contributor to the burst of soleus activity that helps support the stance phase of locomotion. As a result, when H-reflex up-conditioning or down-conditioning changes the size of the H-reflex, there are comparable effects on the soleus locomotor bursts: the burst becomes larger after H-reflex up-conditioning and smaller after H-reflex down-conditioning. Figure 4 shows this effect on the locomotor burst. It is likely to disrupt locomotion and thereby induce compensatory plasticity elsewhere in the spinal cord or brain that is necessary to preserve normal locomotion.

Furthermore, the fact that activity-dependent plasticity can occur at many sites throughout the CNS implies that H-reflex conditioning is likely to be associated with additional reactive (or downstream) plasticity. Such reactive plasticity is an inevitable result of the changes in activity caused by the plasticity directly underlying the H-reflex change and/or by compensatory plasticity. In sum, even a skill as simple as a larger or smaller H-reflex is necessarily and inevitably accompanied by a complex pattern of activity-dependent plasticity that goes well beyond the changes accounting for the modified H-reflex.

Supraspinal Control of the Spinal Cord Plasticity

The descending influence that is operantly conditioned by the protocol and changes the spinal cord is conveyed in the corticospinal tract (CST). Although CST transection alone does not have a long-term effect on H-reflex size, CST transection (or ablation of contralateral sensorimotor cortex, which is the origin of the CST) prevents both down-conditioning and up-conditioning.

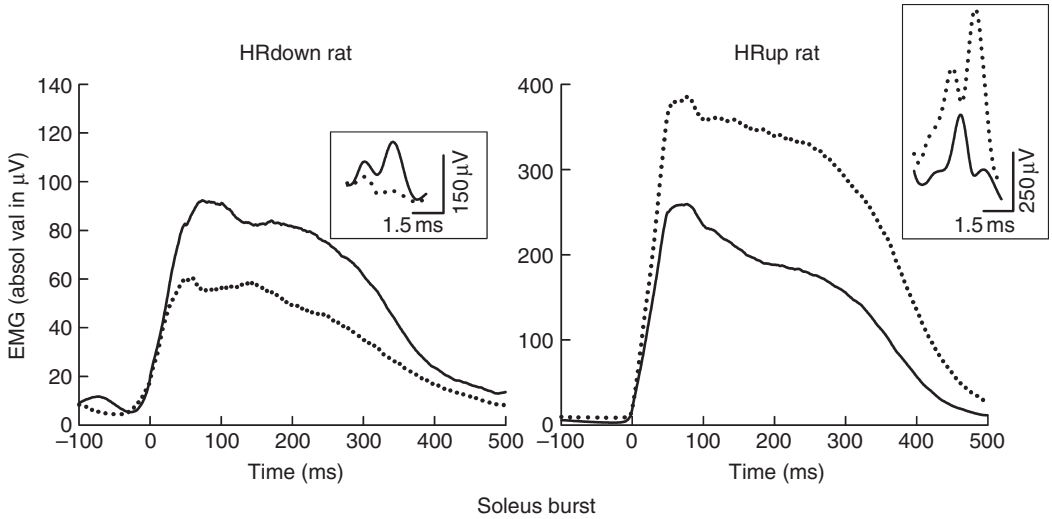

Figure 4 Soleus H-reflex conditioning affects soleus activity during locomotion. Average right soleus locomotor bursts before (solid) and after (dotted) H-reflex conditioning from an up-conditioned (HRup) and a down-conditioned (HRdown) rat. The insets show the average right soleus H-reflexes elicited during the stance phase of locomotion before (solid) and after (dashed) conditioning. After conditioning, both the stance H-reflex and the soleus burst are smaller in the down-conditioned rat and larger in the upconditioned rat. Primary afferent excitation is a major contributor to the locomotor burst. Thus, the changes in the strength of this excitation that result from H-reflex conditioning affect the size of the burst. EMG, electromyograph; absol val, absolute value. Modified from Chen Y, Chen XY, Jakeman LB, et al. (2005) The interaction of a new motor skill and an old one: H-reflex conditioning and locomotion in rats. *Journal of Neuroscience* 25: 6898–6906.

Transection of other major descending tracts or of the dorsal column ascending tract does not impair conditioning. Furthermore, if the CST is transected in rats in which the soleus H-reflex has already been reduced by down-conditioning, the reflex decrease disappears within 5–10 days. Figure 5 illustrates this effect. It implies that the spinal cord plasticity underlying the smaller H-reflex can survive only a few days on its own.

Down-conditioning also requires the cerebellum. It does not occur in rats in which the cerebellar output nuclei dentate and interpositus have been ablated. The cerebellar contribution is mediated through cerebellar output to cortex rather than to spinal cord, since transection of the entire lateral column (which contains the rubrospinal tract) does not impair conditioning. The cerebellum is also needed for the long-term maintenance of conditioning. If the cerebellar output nuclei are ablated in rats in which the soleus H-reflex has already been reduced by down-conditioning, the reflex decrease persists for about 40 days and then disappears. This remarkable finding implies that the CST activity that maintains the spinal cord plasticity that directly underlies the H-reflex decrease depends on supraspinal plasticity that can survive for 40 days after loss of cerebellar influence, but not longer.

A Hierarchy of Plasticity

Figure 6 summarizes current knowledge and hypotheses concerning the complex multisite plasticity that is produced by H-reflex conditioning. Changes occurring both in the spinal cord and in the brain appear to constitute a hierarchy of plasticity: The reward contingency induces and maintains plasticity in the brain that produces the CST activity that induces and maintains the spinal cord plasticity that directly underlies the acquired skill, that is, a smaller or larger H-reflex. The existence of this hierarchy is revealed by the effects of specific lesions. The brain plasticity responsible for the CST activity that changes the spinal cord can survive 40 days after loss of cerebellar influence, and the spinal cord plasticity that directly underlies the smaller H-reflex can survive only 5–10 days after loss of the CST activity. This simple skill is acquired and maintained through a hierarchy of plasticity in the brain and spinal cord.

Comparable Plasticity throughout Life

The complex spinal cord plasticity associated with H-reflex conditioning is not simply a laboratory phenomenon. Gradual activity-dependent plasticity, driven by descending and peripheral inputs, shapes

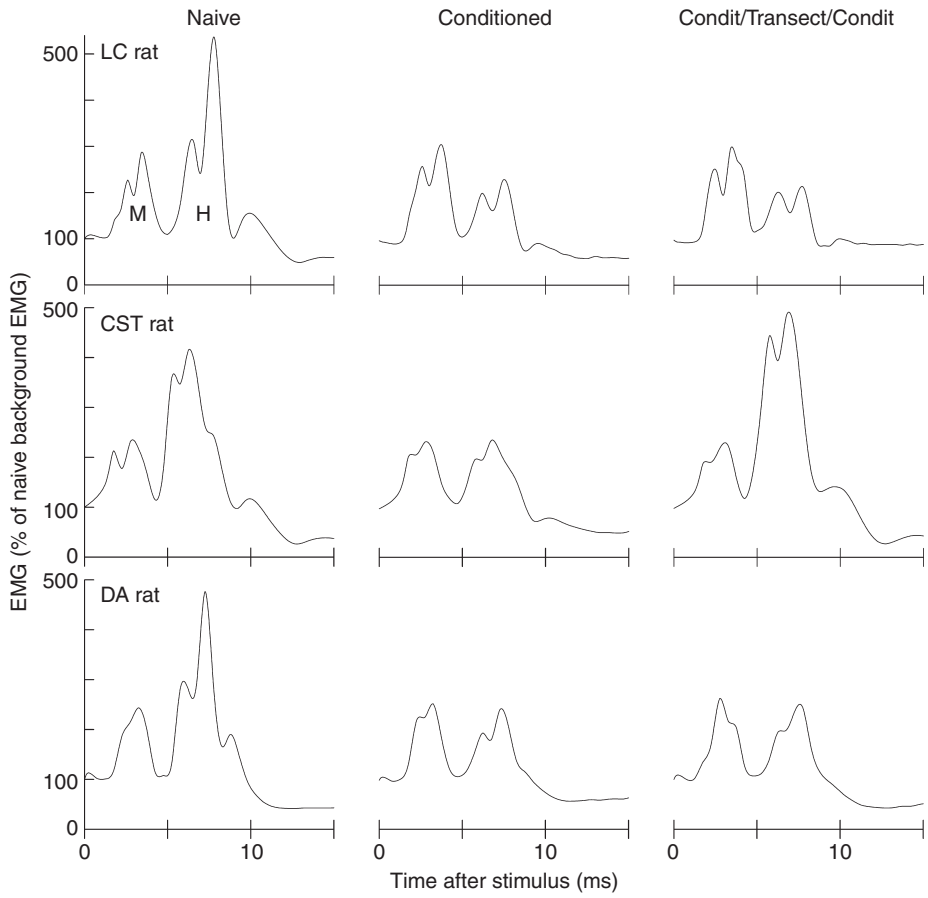

Figure 5 The essential role of the corticospinal tract (CST) in maintenance of H-reflex down-conditioning. Average poststimulus electromyograph (EMG) from three rats for representative days: prior to down-conditioning (left); at the end of down-conditioning and prior to transection of the right column (LC rat), the corticospinal tract (CST rat), or the dorsal ascending tract (DA rat; center); and after transection and continued down-conditioning (Condit/Transect/Condit; right). All three rats decreased the H-reflex (H) in response to down-conditioning. LC or DA transection did not affect the H-reflex decrease. In contrast, CST transection led eventually to loss of the H-reflex decrease and to an H-reflex larger than the original, control-mode H-reflex. Background EMG (shown by the EMG at time zero) and muscle response (M) did not change over the course of study. From Chen XY and Wolpaw JR (2002) Probable corticospinal tract control of spinal cord plasticity in rats. *Journal of Neurophysiology* 87: 645–652.

spinal cord function during development and continues to adjust it throughout life. This plasticity establishes and maintains spinal cord pathways in a state that supports effective motor behaviors.

Early in life, descending input to the spinal cord guides the plasticity that leads to an adult spinal cord that has normal reflexes, supports standard motor skills like locomotion, and can also support specialized skills like dancing or playing the piano. Descending inputs and associated peripheral inputs establish normal patterns of spinal proprioceptive reflexes, flexion withdrawal reflexes, and autonomic reflexes. If perinatal supraspinal damage (e.g., cerebral palsy) disrupts descending input, infantile reflex patterns may persist into adulthood and contribute to disability. Figure 7(a) illustrates the effect of neonatal loss of descending input on flexion withdrawal reflexes of adult rats.

Mastery of motor skills later on in life is associated with changes in spinal cord reflexes comparable to those produced in the laboratory. Spinal reflexes are affected by the nature, intensity, and duration of past physical activity and by specific training regimens. SSRs and H-reflexes differ in athletes and nonathletes and among different kinds of athletes. Such differences may contribute to the skills possessed by these different groups. For example, both H-reflexes and disynaptic reciprocal inhibition in the leg are very small in professional ballet dancers. From the perspective of performance, the lessened peripheral influence on motor neurons indicated by the smaller reflexes might increase cortical control and thereby improve the precision of movement.

The impact of specific training regimens furnishes additional evidence of activity-dependent plasticity in spinal cord reflex function. For example, Figure 7(b)

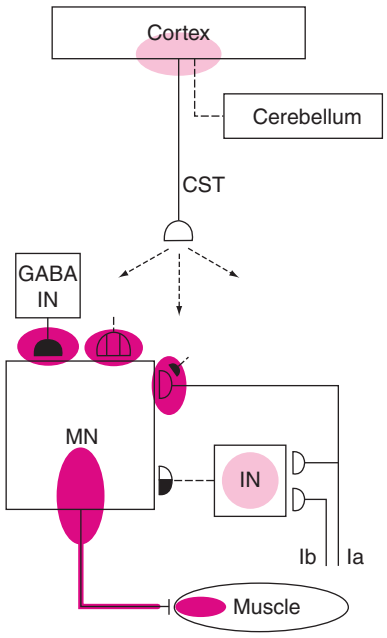

Figure 6 The hierarchy of spinal and supraspinal plasticity underlying spinal reflex conditioning. The shaded ovals indicate the spinal and supraspinal sites of the plasticity associated with operant conditioning of the spinal stretch reflex (SSR) or its electrical analog, the H-reflex. MN is the motor neuron, CST is the main corticospinal tract, IN is a spinal interneuron, and GABA IN is a γ-aminobutyricacidergic (GABAergic) interneuron. Open synaptic terminals are excitatory, solid ones are inhibitory, half-open one could be either, and the subdivided one is a cluster of C-terminals. Dashed pathways imply the possibility of intervening spinal interneurons. The monosynaptic and probably oligosynaptic SSR/H-reflex pathway from Ia and Ib afferents to the motor neuron is shown. Definite (red shading) or probable (pink shading) sites of plasticity include the motor neuron membrane (i.e., firing threshold and axonal conduction velocity), motor unit properties, GABAergic inhibitory terminals, and C-terminals on the motor neuron, the Ia afferent synaptic connection, terminals conveying disynaptic group I inhibition or excitation to the motor neuron, and sensorimotor cortex. The essential roles of the CST (originating in sensorimotor cortex) and of cerebellar output to cortex are indicated. The spinal plasticity that is directly responsible for H-reflex conditioning appears to be induced and maintained by cortical plasticity that itself depends for its long-term survival on the cerebellum. Modified and updated from Wolpaw JR and Tennissen AM (2001) Activity-dependent spinal cord plasticity in health and disease. *Annual Review of Neuroscience* 24: 807–843.

shows that practice of a backward-walking task 15 min daily leads to a gradual change in the relationship between soleus H-reflex size and the time in the step cycle when the reflex is evoked. On the first day, the reflex is large well before the beginning of the soleus burst in the stance phase of the cycle. By the tenth day, the H-reflex is not detectable until the burst begins. the earlier reflex increase on the first day may compensate for uncertainty as to when the stance phase, with its need for greater soleus activity, will begin: the increased sensitivity of the reflex arc promotes rapid excitation by foot contact. The time of contact becomes more predictable as training progresses, so that this compensation is not necessary.

The reflex changes associated with aging are further evidence of adaptive spinal cord plasticity in adult life and in response to specific demands. Age-related alterations in reflex size and task-dependent reflex modulation (e.g., Figure 7(c)) are likely to reflect both the direct and indirect effects of aging; that is, the direct effects of aging on the reflex circuitry itself and the indirect effects of aging due to its effects elsewhere in the CNS or its effects on the peripheral sensory and motor apparatus involved in movement.

Therapeutic Uses of Spinal Reflex Conditioning

Activity-dependent changes in spinal cord reflexes contribute not only to normal motor behaviors but also to the complex motor disabilities that accompany spinal cord injuries and other chronic neurological disorders. Recognition of this additional importance has coincided with and has been encouraged by the energy and optimism now focused on efforts to restore spinal cord structure and function after injury. As described above, an effective adult spinal cord is the result of appropriate activity-dependent plasticity during development and throughout later life. In contrast, a spinal cord that is newly regenerated after injury is likely to display diffuse infantile reflexes and other disordered and ineffective behaviors. As a result, new methods for properly reeducating the regenerated spinal cord to restore more appropriate reflex patterns are likely to be required.

Operant conditioning of the H-reflex can be used to modify the performance of other behaviors that depend on the same pathways. For example, as Figure 4 illustrates, up-conditioning of the right soleus H-reflex increases the amplitude of the soleus locomotor burst during the stance phase of walking, and H-reflex down-conditioning decreases the burst. These effects are consistent with the established importance of primary afferent input in producing the locomotor burst.

Furthermore, H-reflex conditioning can improve locomotion after partial spinal cord injury. Midthoracic transection of the right lateral column (which leaves the CST intact) produces a lasting asymmetry in treadmill locomotion. The right stance phase appears to be abnormally short. As Figure 8 shows, H-reflex up-conditioning, which strengthens the right soleus burst, eliminates this asymmetry. These results suggest that reflex conditioning protocols might improve motor function in people with partial spinal cord injuries. Such protocols could be especially useful

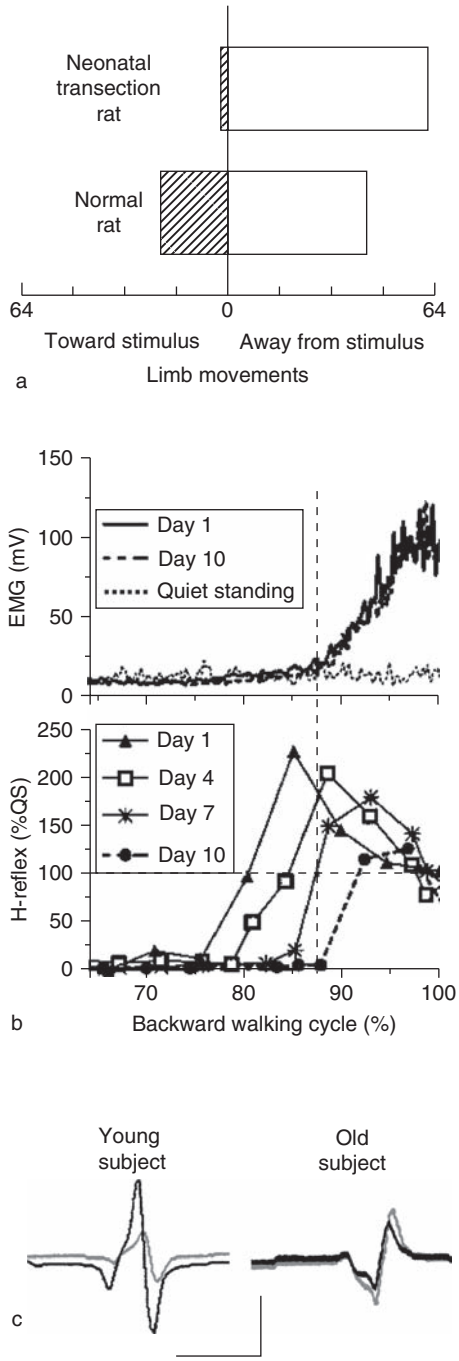

Figure 7 Spinal cord reflexes change throughout life. (a) The direction of limb movements produced by flexion withdrawal responses to painful stimuli in normal adult rats and in adult rats that had undergone spinal cord transection just after birth. The direction of limb movement is almost always appropriate, i.e., away from the stimulus, in normal adults but is often inappropriate in transected adults. Neonatal spinal cord transection prevents descending input from gradually shaping normal (i.e., appropriate) flexion withdrawal responses. Modified from Levinsson A, Luo XL, Holmberg H, and Schouenborg J (1999) Developmental tuning in a spinal nociceptive system: Effects of neonatal spinalization. *Journal of Neuroscience* 19: 10397–10403. (b) Change in soleus H-reflex size as a function of time in the backward-walking step cycle as a human masters backward walking over 10 days. Top: Soleus electromyographic (EMG) activity on day 1 (solid) and day 10 (dashed) just before and after onset of the soleus burst associated with the stance phase of the step cycle. The dotted line shows soleus EMG during quiet standing. Bottom: H-reflex size (as % of size during quiet standing; QS) versus time in the backward-walking step cycle for days 1, 4, 7, and 10 of training. Soleus EMG does not change with training. In contrast, the marked increase in H-reflex size prior to the soleus burst that is seen on day 1 disappears by day 10. Modified from Schneider S and Capaday C (2003) Progressive adaptation of the soleus H-reflex with daily training at walking backward. *Journal of Neurophysiology* 89: 648–656. (c) Soleus H-reflexes in prone (black) and standing (gray) positions from a young person and an old person. In old participants, the H-reflex tends to be smaller and less affected by body position. Modified from Koceja DM, Markus CA, and Trimble MH (1995) Scale: Vertical scale is 2.4 mV for the young subject and 1.2 mV for the old subject; horizontal scale is 16 ms. Postural modulation of the soleus H reflex in young and old subjects. *Electroencephalography & Clinical Neurophysiology* 97: 387–393.

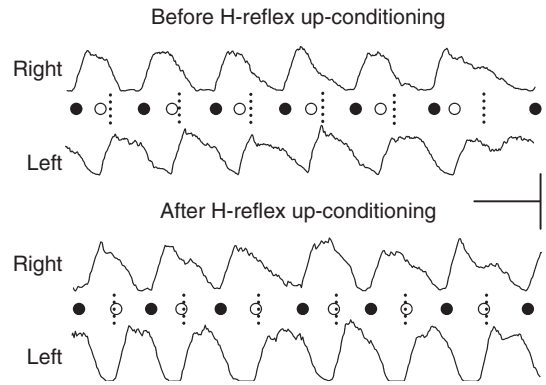

Figure 8 H-reflex conditioning may improve walking after spinal cord injury. The effects of H-reflex conditioning on locomotion in a rat with midthoracic transection of the right lateral column. The traces show the bursts of electromyographic (EMG) activity from right and left soleus muscles during treadmill locomotion before (top) and after (bottom) H-reflex up-conditioning has increased the size of the right soleus H-reflex. The presumed onsets of the right and left stance phases of locomotion are indicated by the solid and open circles, respectively. The short vertical dashed lines mark the midpoints between right burst onsets, which is where the left burst onsets should occur. Before H-reflex up-conditioning, the left burst onset occurs too early; after up-conditioning has strengthened the right soleus burst, the left burst onset occurs on time. Horizontal scale bar: 0.5 s; vertical scale bar: 100 and 150 μV for the right and left bursts, respectively. From Chen Y, Chen XY, Jakeman LB, *et al.* (2006) Operant conditioning of H-reflex can correct a locomotor abnormality after spinal cord injury in rats. *Journal of Neuroscience* 26(48): 12537–12543.

when regeneration becomes practical and precise methods for reeducating the regenerated spinal cord become essential in order to restore effective function. The recent demonstration that reciprocal inhibition can also be operantly conditioned extends the options for using reflex conditioning to improve motor function.

See also: Corticospinal Development; Primate Interneurons; Reflex Circuits; Sensorimotor Plasticity and Control of Movement Following Spinal Cord Injury +D18; Sensorimotor Integration: Models; Spinal Cord Regeneration and Functional Recovery: Strategies; Spinal Cord Injuries; Spinal Motor Neurons: Properties; Synaptic Mechanisms of Learning.

Further Reading

Brown WF (1984) *The Physiological and Technical Basis of Electromyography*. Boston: Butterworths.

Carp JS and Wolpaw JR (1994) Motoneuron plasticity underlying operantly conditioned decrease in primate H-reflex. *Journal of Neurophysiology* 72: 431–442.

Chen Y, Chen XY, Jakeman LB, et al. (2005) The interaction of a new motor skill and an old one: H-reflex conditioning and locomotion in rats. *Journal of Neuroscience* 25: 6898–6906.

de Groat WC (2002) Plasticity of bladder reflex pathways during postnatal development. *Physiology & Behavior* 77: 689–692.

Dobkin BH and Havton LA (2004) Basic advances and new avenues in therapy of spinal cord injury. *Annual Review of Medicine* 55: 255–282.

Henneman E and Mendell LM (1981) Functional organization of motoneuron pool and inputs. In: Brooks VB (ed.) *Handbook of Physiology, Section 1, The Nervous System, Vol. 2: Motor Control*, pp. 423–507. Bethesda, MD: American Physiological Society.

Matthews PBC (1972) *Mammalian Muscle Receptors and Their Central Actions*. pp. 319–409. Baltimore: Williams & Wilkins.

Nielsen J, Crone C, and Hultborn H (1993) H-reflexes are smaller in dancers from the Royal Danish Ballet than in well-trained athletes. *European Journal of Applied Physiology* 66: 116–121.

O'Sullivan MC, Miller S, Ramesh V, et al. (1998) Abnormal development of biceps brachii phasic stretch reflex and persistence of short latency heteronymous reflexes from biceps to triceps brachii in spastic cerebral palsy. *Brain* 121: 2381–2395.

Schneider S and Capaday C (2003) Progressive adaptation of the soleus H-reflex with daily training at walking backward. *Journal of Neurophysiology* 89: 648–656.

Stein RB (1995) Presynaptic inhibition in humans. *Progress in Neurobiology* 47: 533–544.

Wang Y, Pillai S, Wolpaw JR, and Chen XY (2006) Motor learning changes GABAergic terminals on spinal motoneurons in normal rats. *European Journal of Neuroscience* 23: 141–150.

Wolpaw JR (1997) The complex structure of a simple memory. *Trends in Neurosciences* 20: 588–594.

Wolpaw JR and Chen XY (2006) The cerebellum in maintenance of a motor skill: A hierarchy of brain and spinal cord plasticity underlies H-reflex conditioning. *Learning & Memory* 13: 208–215.

Wolpaw JR and Tennissen AM (2001) Activity-dependent spinal cord plasticity in health and disease. *Annual Review of Neuroscience* 24: 807–843.

Zehr EP (2002) Considerations for use of the Hoffman reflex in exercise studies. *European Journal of Applied Physiology* 86: 455–468.

Opioid Peptides and Receptors

B L Kieffer, IGBMC,CNRS/INSERM/ULP, Illkirch, France

© 2009 Elsevier Ltd. All rights reserved.

The Opioid System: A Short History

Discovery of the opioid system stems from the use and abuse of opium in ancient history. Opium, extracted from poppy seeds (*Papaver somniferum*; see Figure 1(a)), has powerful pain-relieving properties and produces euphoria. This substance has been used both medicinally and recreationally for several millennia. Morphine, named after the god Morpheus, is the most active ingredient of opium. The compound was isolated in 1805 and rapidly became the clinical treatment of choice to alleviate severe pain. Heroin was synthesized chemically by morphine diacetylation in the late 1800s and was commercialized as the first nonaddictive opiate to treat cough and asthma. The strong addictive properties of heroin were soon acknowledged, and both heroin and opium were prohibited in 1910. Today morphine remains the best pain-killer in contemporary medicine, despite a wide array of adverse side effects (respiratory depression, constipation, tolerance, and dependence). Heroin is a main illicit drug of abuse, and heroin addiction represents a major public health issue. Because of their extraordinarily potent analgesic and addictive properties, opiates have prompted scientists to seek to understand their mode of action in the brain.

In 1973, three independent teams showed that opiates bind to membrane receptors in the brain, and these receptors were named μ, δ, and κ a few years later. In 1975, two pentapeptides, Met- and Leu-enkephalin were isolated as the first endogenous ligands for these receptors. Many peptides followed, forming the opioid peptide family. Three distinct genes encoding the peptides were isolated in the late 1970s and early 1980s; three other genes encoding the receptors were cloned a decade later. In the mid 1990s the entire endogenous opioid system was characterized at the molecular level (Figure 1(b)).

Molecular Components of the Opioid System

The Peptides

Opioid peptide genes encode large precursor proteins, which are further cleaved into shorter peptides. All opioid peptides share a common N-terminal Tyr-Gly-Gly-Phe signature sequence, which interacts with opioid receptors. The preproenkephalin gene encodes several copies of the pentapeptide Met-enkephalin and one copy of Leu-enkephalin. This gene also encodes longer versions of Met-enkephalin. Notably, one of them, the opioid peptide bam22, activates opioid receptors via its N-terminal sequence while activating another receptor family known as sensory neuron-specific (SNS) receptors via its C-terminal part. The preprodynorphin gene also encodes multiple opioid peptides, including dynorphin A, dynorphin B, and neo-endorphin. The preprodynorphin-derived peptides all contain the Leu-enkephalin N-terminal sequence and basic amino acid residues in their C-terminal end. Finally the preproopioimelanocortin gene encodes a single opioid peptide, β-endorphin, among other peptides with nonopioid biological activity. β-endorphin is the longest opioid peptide (31 amino acids) and shows highly potent and long-lasting analgesic properties.

The Receptors

Opioid receptors are membrane receptors with a seven-transmembrane topology that belong to the large G-protein-coupled receptor (GPCR) superfamily. This family comprises several hundred members within the mammalian genome. μ-, δ-, and κ-opioid receptor genes constitute a GPCR subfamily, together with a fourth member encoding the orphanin FQ/nociceptin receptor. The last receptor and its endogenous ligand share high structural homology with the opioid receptors and peptides. However, the receptor does not bind opioids and the peptide orphanin FQ/nociceptin shows no affinity toward μ, δ, or κ receptors. All four receptor genes display similar genomic organization, with conserved intron–exon junctions, suggesting a common ancestor gene. These genes have been characterized throughout vertebrates, but cannot be identified earlier in evolution. The four genes are referred as to Oprm1 (μ), Oprd1 (δ), Oprk1 (κ), and Oprl1 (orphanin FQ/nociceptin) (genomic information is available on the NCBI website). The encoded receptors are classically named MOR, DOR, KOR, and ORL-1, respectively, throughout the literature. A novel nomenclature MOP (μ), DOP (δ), KOP (κ), and NOP (orphaninFQ/nociceptin) has recently been proposed by the International Union of Basic and Clinical Pharmacology (IUPHAR) to unify receptor abbreviations. Many single-nucleotide polymorphisms have been identified in both coding and regulatory sequences of human opioid receptor genes, and large genetic investigations are underway to associate haplotypes with neurological or psychiatric diseases. At present, coding

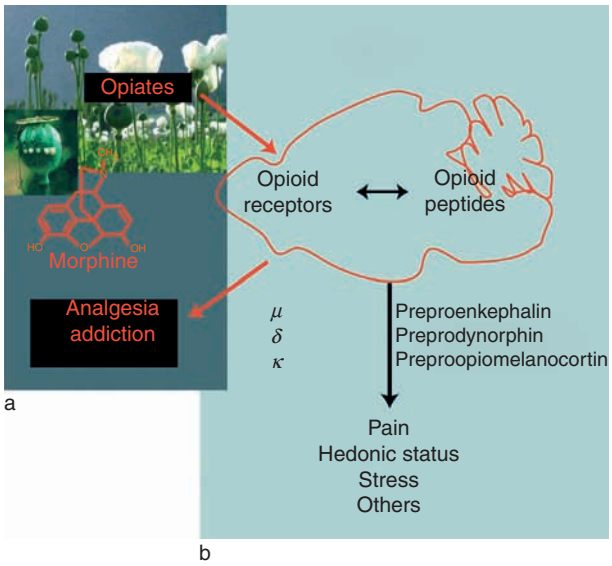

Figure 1 The opioid system: (a) exogenous opiates activate opioid receptors; (b) the endogenous opioid system. In (a), morphine is the main active ingredient of opium and has strong analgesic and addictive properties, which result from a direct interaction of the drug with opioid receptors. In (b), the endogenous opioid system is a neuromodulatory system. The three opioid receptors naturally interact with a family of opioid peptides, produced by proteolytic cleavage of three large precursor proteins. Receptors and peptides are broadly expressed throughout the nervous system and control nociceptive pathways (pain), mood, well-being (hedonic status), and responses to stress. The system also controls respiration, gastroinstestinal motility, and endocrine and immune functions.

polymorphisms as well as silent or noncoding polymorphisms that may influence receptor expression levels in the brain are under study.

Physiology of the Opioid System *In Vivo*

Opioid peptides and receptors are strongly expressed in the limbic system, as well as in nociceptive pathways and along the hypothalamic–pituitary–adrenal axis. In these neural circuits, the opioid system regulates addictive and emotional behaviors, and controls responses to pain or stress.

Pharmacology

Three decades of pharmacology have demonstrated an analgesic activity for all three μ, δ, and/or κ agonists. Generally, opioid analgesics used in the clinic are μ agonists. These compounds show strongest efficacy in the treatment of severe acute pain, but their activity remains variable in situations of chronic pain. κ agonists are potential candidates in the treatment of visceral pain. Significantly, studies of μ and κ compounds have highlighted opposing activities of μ and κ agonists on mood. μ agonists show potent rewarding properties and abuse liability, whereas κ agonists are strongly dysphoric. The latter activity has hindered the development of centrally acting κ agonists for pain control in the clinic. The pharmacology of δ agonists has shown slower progress. The notion that δ agonists may represent useful analgesics with low abuse liability has been proposed for many years and is currently being explored.

Genetics Approach

For a decade, mice lacking μ, δ, or κ receptors, as well as preproenkephalin, preprodynorphin, or β-endorphin, have been created by gene-targeting (knockout mice). Single mutant mice, or even the triple receptor knockout mice, are viable and fertile and show no obvious developmental deficit, indicating that the opioid system is not essential for survival. These mutant mice have been extensively analyzed either for spontaneous behaviors or in response to opioid and nonopioid drugs. Parallel to the pharmacology, this genetic approach has clarified the specific contribution of each molecular component of the opioid system in opioid-controlled physiology and behaviors *in vivo*. In particular, the comparative analysis of μ, δ, and κ knockout mice has clearly shown very distinct activity patterns for each receptor *in vivo* (Figure 2). Similar comparisons are currently underway for peptide knockout mice. Main conclusions from gene knockout mice are the following.

μ receptors represent the primary molecular target for morphine *in vivo* and mediate both beneficial and adverse effects of the most broadly used opiate. μ receptors also mediate the rewarding properties of nonopioid drugs of abuse including cannabinoids, alcohol, and nicotine or even natural reinforcers such as social interactions. μ receptors therefore represent a key molecular trigger for reward, and most likely contribute to the initiation of addictive behaviors. κ receptors, as predicted by the pharmacology, mediate dysphoric activities of both κ opioids and cannabinoids and oppose μ receptors in regulating the hedonic tone. δ receptors are less directly involved in hedonic control. Very distinct from μ and κ receptors, δ receptors regulate emotional responses and show anxiolytic and antidepressant activity. This specific function of δ receptors is now confirmed both by gene knockout and pharmacological studies using SNC80, the only commercially available highly selective δ compound. Further analysis of δ knockout mice and the development of more selective compounds will probably reveal other activities of delta receptors.

μ, δ, and κ receptor-deficient mice all exhibit enhanced pain sensitivity. This indicates that the three receptors, activated by endogenous opioid peptides, tonically inhibit nociceptive responses. Noticeably,

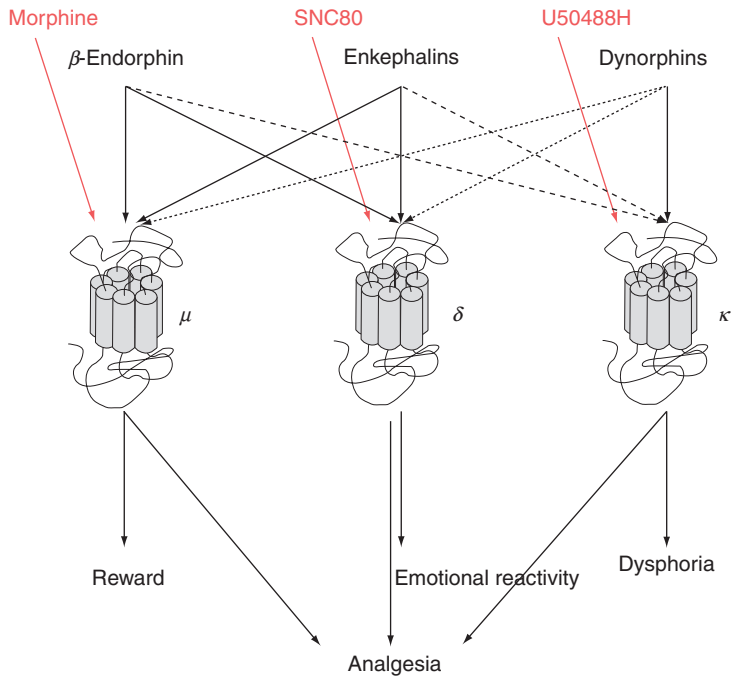

Figure 2 Physiological role of opioid receptors and peptides *in vivo*. Opioid peptides show some, although weak, selectivity toward μ, δ, and κ receptors. A highly selective exogenous (morphine) or synthetic (SNC80 and U50488H) agonist is indicated for each receptor. The combination of pharmacological studies with gene-knockout approaches (targeted gene inactivation in mice) showed that each receptor controls specific neural responses. μ and κ receptors produce euphoria or dysphoria, respectively, whereas δ receptors control levels of anxiety and depressive-like behaviors (emotional reactivity). The activation of all three receptors inhibits pain (analgesia), but each receptor controls a distinct panel of pain modalities.

the phenotypes of mutant mice differ across pain assays. In models of physiological or acute pain, μ receptors modulate mechanical, chemical, and supraspinally controlled thermal nociception, whereas κ receptors modulate spinally mediated thermal nociception and visceral pain. Again, δ receptors differ from μ and κ receptors in that there is no obvious regulation of acute pain. In contrast, there is evidence for a role of δ receptors in reducing hyperalgesia in situations of inflammatory and neuropathic pain. These data, combined with many pharmacological studies, clearly demonstrate a specific role for each receptor in regulating the broad diversity of pain modalities.

Finally, due to the short half-life and structural similarity of the peptides, identifying a specific role for each endogenous opioid *in vivo* has been difficult. Data from opioid peptide knockout mice do not necessarily follow data from receptor knockouts, consistent with the notion that multiple peptides probably act at each receptor. Although affinity studies and pharmacological approaches suggest a relationship between β-endorphin and μ receptors, enkephalin and δ receptors, and dynorphin and κ receptors (Figure 2), knockout data provide a more complex picture. It is likely that the specific anatomical location of receptors and peptide release largely drive opioidergic synapse functioning.

Receptor Structure

Receptor Binding

Overall μ, δ, and κ receptors show a 60% amino acid sequence identity. Closest homology occurs within the seven-transmembrane helical core, which contains the opioid-binding pocket (Figure 3(b)). Extracellular domains, including three extracellular loops and the N-terminal domain, determine μ, δ, and κ selectivity. These domains differ strongly among receptors and probably act as a gate filtering μ, δ, and κ agonists or antagonists entering the binding pocket. Note that synthetic compounds have been developed with high μ, δ, and κ selectivity, whereas endogenous peptides show limited receptor preference (Figure 2).

Receptor Signaling

As in all GPCRs, opioid receptors convey extracellular signals within the cell by activating heterotrimeric G-proteins, which interact with cytoplasmic domains of the receptor. Agonist binding modifies helical packing of the receptor, and a rearrangement in the positioning of transmembrane domains 3, 6, and 7 has been proposed to drive the transition between inactive and active conformations of the receptor.

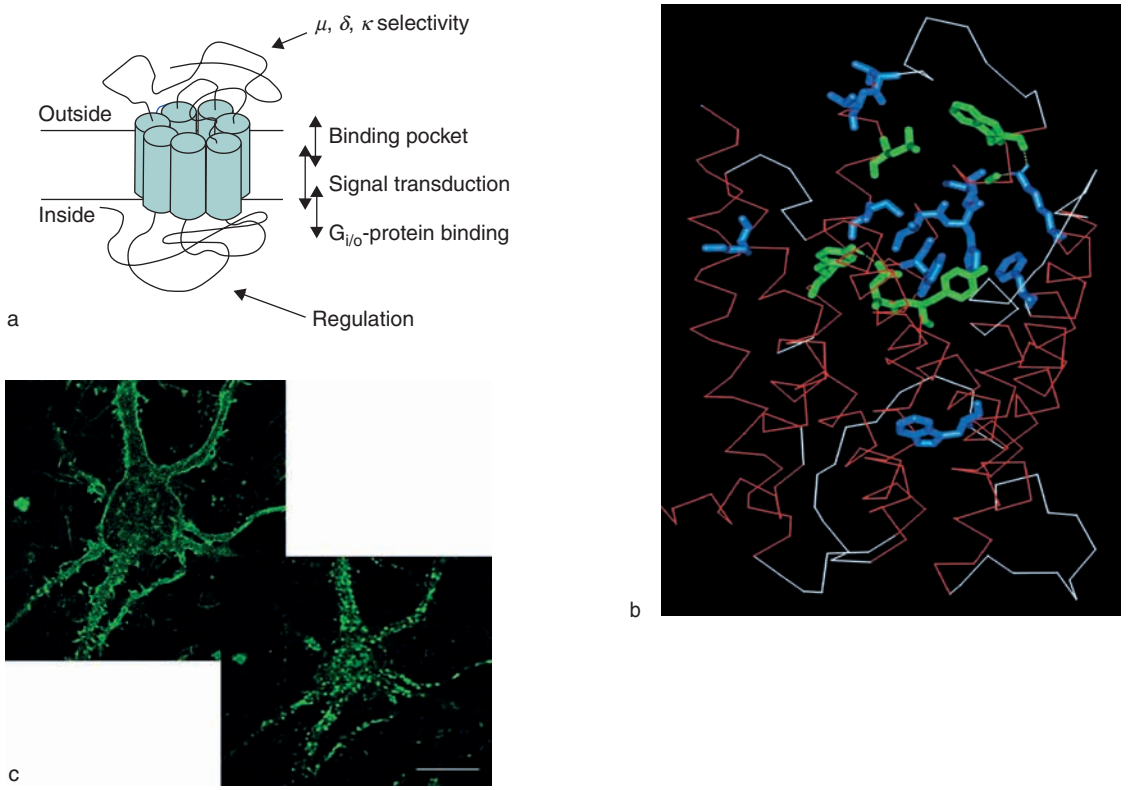

Figure 3 Opioid receptors: (a) receptors belong to the G-protein-coupled receptor superfamily with a seven-transmembrane topology; (b) three-dimensional computer model of the human δ-opioid receptor; (c) direct visualization of δ receptors in native neurons. As shown in (a), μ, δ, and κ receptors are highly homologous. Each receptor domain contributes to receptor function, as indicated. Regulation occurs immediately after the receptor is activated by an opioid agonist and includes receptor phosphorylation, uncoupling from the G-protein, and internalization. In (b), the seven-transmembrane helices are shown in red, and portions of extrahelical loops are in white. The binding pocket penetrates halfway within the helical bundle, and side chains of amino acid residues forming the main binding site are shown in blue and green. As shown in (c), hippocampal neurons extracted from a genetically engineered mouse endogenously produce fluorescent δ-opioid receptors. The receptor is observed at the surface of both cell bodies and processes of the neuron (see continuous green fluorescence on the plasma membrane, top). Exposure to the agonist produces receptor internalization, visible under the form of highly fluorescent dots within the neuron. Note the concomitant fading of surface fluorescence (bottom).

This helical movement modifies the receptor's intracellular structure, hence the receptor–G-protein interaction. Intracellular loops of the receptor form a large part of the receptor–G-protein interface. These intracellular receptor domains are almost identical across μ, δ, and κ receptors, consistent with the fact that all three receptors interact with inhibitory G-proteins of the G_o/G_i type. G-protein subunits dissociate from the activated receptor and, in turn, modulate intracellular effectors and pathways.

Several opioid-evoked signaling events have been identified both in transfected cells and native tissues. Opioids inhibit voltage-dependent Ca^{2+} channels or activate inwardly rectifying potassium channels, thereby diminishing neuronal excitability. Opioids also inhibit the cyclic adenosine monophosphate (cAMP) pathway and activate mitogen-activated protein kinase (MAPK) cascades, both of which affect cytoplasmic events and transcriptional activity of the cell. Overall, opioids inhibit neurons by decreasing either neuronal firing or neurotransmitter release, depending on the post- or presynaptic localization of the receptors. Finally, opioid receptors are expressed on both excitatory and inhibitory neurons and can therefore exert inhibition or disinhibition within neural circuits.

Regulation of Receptor Signaling

The C-terminal domain differs largely across opioid receptors. Together with intracellular loops, this receptor domain contributes to receptor–G-protein coupling. Significantly, both the C-terminus and intracellular loops also interact with other cellular effectors following receptor activation. Effectors include receptor-specific (G-protein-coupled receptor kinase (GRK)) or-nonspecific (protein kinase A (PKA) and protein kinase C (PKC)) kinases, arrestins that act as adaptors between the receptor and the endocytic machinery, and other recently identified proteins.

These interacting proteins drive the desensitization of receptor signaling, which typically occurs as an adaptive response after receptor stimulation. Desensitization mechanisms include receptor phosphorylation and uncoupling from the G-protein, as well as receptor endocytosis (Figure 3(c)) and redistribution between the cell surface and intracellular compartments. Largely determined by C-terminal epitopes, these dynamic phenomena differ across μ, δ, and κ receptors. Many studies have addressed the mechanisms of opioid receptor trafficking and desensitization in cellular models and, more recently, *in vivo*. At present, studies on native neurons suggest that agonist-stimulated δ receptors are mainly targeted toward degradation pathways after endocytosis, whereas internalized μ receptors are resensitized in early endosomes and rapidly recycle back to the cell surface. Understanding the dynamic aspects of receptor trafficking to and from the cell surface has become a major field in GPCR research.

Chronic Opiates

Regulatory events that occur at the receptor level, as already described, represent adaptive mechanisms that classically contribute to cellular homeostasis under acute agonist stimulation. Chronic exposure to opiates, however, has long-term consequences, which lead to strong and perhaps irreversible alterations of brain functioning at the cellular, synaptic, and network levels. Behaviorally, these adaptations are manifested by tolerance, defined as a reduced sensitivity to the drug effects, and dependence revealed by drug craving and withdrawal symptoms. Many efforts have been made for several decades to understand the molecular basis of these elaborate responses to chronic opiates. Modifications of opioid receptor-associated ion channels and second-messenger pathways have been demonstrated in several brain areas. Also several protein kinases, including PKA, PKC, and calcium/calmodulin-dependent protein kinase (CaMK)II; glutamate receptors; cytoskeleton proteins or neurotrophic factors are involved. Finally opposing neurotransmitter systems – called anti-opioid systems – are recruited. Integrating these observations into a coherent explanation for tolerance and dependence *in vivo* remains a challenge.

Opioid Receptor Heterogeneity

The Existence of Opioid Receptor Pharmacological Subtypes

Opioid receptor pharmacology is complex and the existence of multiple μ, δ, and κ receptor types has been proposed since the early 1970s. Gene cloning led to the characterization of only three receptor genes, and the molecular basis for pharmacological diversity has long remained a matter of debate. Alternative splicing has been reported, but it has been extremely difficult to establish the biological relevance of these alternative transcripts *in vivo* and to correlate their existence with the multiple opioid receptor subtypes that were described earlier by the pharmacology. Today, it is admitted that the three main Oprm1-, Oprd1-, and Oprk1-encoded receptors are highly dynamic proteins, which may indeed account for the wide diversity of pharmacological subtypes.

Three Receptor Proteins, Many More Cellular Responses

There are several ways to explain pharmacological heterogeneity of opioid receptors. First, increasing evidence supports the notion that μ, δ, and κ receptors may adopt multiple active conformations. Mutagenesis data suggest the existence of multiple binding modes for opioids within the binding pocket. Further, signaling studies in cellular models show that receptor activation and subsequent regulations are strongly drug-dependent. Hence the ligand–receptor complex, rather than the receptor itself, determines the ultimate physiological cellular response. A second source of heterogeneity is the direct cellular environment of the receptor. Heterotrimeric G-proteins differ across cell types, and the number of possible G-protein-associated signaling pathways has expanded dramatically. The variable combinations of G-protein subunits and the nature of associated signaling networks necessarily generate neuron-specific, or even neuron compartment-specific, responses. Also, many regulatory proteins directly interact with the receptor C-terminal tail (as already discussed) and potentially influence opioid-receptor pharmacology, as was demonstrated for several other GPCRs. A third potential receptor modulator is another receptor molecule. The possibility that GPCRs exist as dimeric or oligomeric complexes has gained evidence in the recent years. Co-expression data suggest that the physical association of opioid receptors either as homodimers or heterodimers, or even with other GPCRs, creates novel receptor entities with unique pharmacological properties, which increases opioid receptor heterogeneity. Whether receptor dimerization truly modulates opioid pharmacology *in vivo* remains an important question.

In conclusion, molecular approaches have provided a novel view of opioid receptors. The receptor is now considered a dynamic multicomponent unit rather than a single protein entity. It is likely that the complexity of opioid responses will extend beyond the previously

reported pharmacological subtypes as *in vivo* molecular pharmacology develops. Rational design of opioid compounds that activate a specific subset of μ, δ, or κ receptor-associated signaling pathways is a possible strategy for developing novel drugs of high therapeutic value and low adverse activities, but this still remains a far distant goal.

Conclusion

The last decade in opioid research has led to a better understanding of the genetic basis of opioid-controlled behaviors and physiology. It has also opened avenues toward the development of novel therapeutic opioids in the field of chronic pain, emotional disorders, and addictive diseases. Recombinant technologies have allowed high-throughput screening for novel opiate compounds, the detailed molecular analysis of receptor structure and function, and the identification of associated proteins regulating receptor signaling.

Ongoing studies have returned to analyzing opioid receptor and peptides operating in their physiological environment, using molecular and genetic tools as well as high-resolution imaging techniques. There is a need to identify neural sites where opioid peptide release and subsequent opioid receptor activation controls nociceptive, emotional, and motivational circuits. Also, it is important to characterize signaling pathways that are relevant to specific behavioral or physiological responses to opioids *in vivo*. In addition, novel insights into brain function will arise from understanding the interactions of the opioid system with other neurotransmitter systems. For example, whether the opioid system interacts with the cannabinoid system or anti-opioid systems at the molecular, cellular, or circuit level remains an open question. Last, the elucidation of molecular and network adaptations to chronic opiates will undoubtedly shed light on the general mechanisms of brain plasticity.

See also: Addiction: Neurobiological Mechanism; Drug Addiction: Behavioral Neurophysiology; Invertebrate Neurohormone GPCRs; Neuropeptides: Pain; Single Photon Emission Computed Tomography (SPECT): Technique.

Further Reading

Akil H, Watson SJ, Young E, et al. (1984) Endogenous opioids: Biology and function. *Annual Review of Neuroscience* 7: 223–255.

Bailey CP and Connor M (2005) Opioids: cellular mechanisms of tolerance and physical dependence. *Current Opinion in Pharmacology* 5: 60–68.

Brownstein MJ (1993) A brief history of opiates, opioid peptides and opioid receptors. *Proceedings of the National Academy of Sciences of the United States of America* 90: 5391–5393.

Chang KJ, Porreca F, and Woods JH (eds.) (2004) *The Delta Receptor.* New York: Dekker.

Contet C, Kieffer BL, and Befort K (2004) Mu opioid receptor: A gateway to drug addiction. *Current Opinion in Neurobiology* 14: 370–378.

Devi LA (2001) Heterodimerization of G-protein-coupled receptors: Pharmacology, signaling and trafficking. *Trends in Pharmacological Sciences* 22: 532–537.

Dickenson TH and Kieffer BL (2005) Opiates – basic. In: Koltzenburg M, and McMahon SB (eds.) *Wall & Melzack's Textbook of Pain,* 5th edn., ch. 27. London: Elsevier.

Evans CJ (2004) Secrets of the opium poppy revealed. *Neuropharmacology* 47(supplement 1): 293–299.

Ikeda K, Ide S, Han W, et al. (2005) How individual sensitivity to opiates can be predicted by gene analyses. *Trends in Pharmacological Sciences* 26: 311–317.

Kieffer BL (1997) Molecular aspects of opioid receptors. In: Dickenson A and Besson J-M (eds.) *Handbook of Experimental Pharmacology, Vol. 130: The Pharmacology of Pain*, pp. 281–303. Berlin: Springer.

Kieffer BL and Gaveriaux-Ruff C (2002) Exploring the opioid system by gene knockout. *Progress in Neurobiology* 66: 285–306.

LaForge KS, Yuferov V, and Kreek MJ (2000) Opioid receptor and peptide gene polymorphisms: Potential implications for addictions. *European Journal of Pharmacology* 410: 249–268.

Mansour A, Fox CA, Akil H, et al. (1995) Opioid-receptor mRNA expression in the rat CNS: Anatomical and functional implications. *Trends in Neuroscience* 18: 22–29.

Nestler EJ (1996) Under siege: The brain on opiates. *Neuron* 16: 897–900.

Pierce KL, Premont RT, and Lefkowitz RJ (2002) Seven-transmembrane receptors. *Nature Reviews Molecular Cell Biology* 3: 639–650.

Williams JT, Christie MJ, and Manzoni O (2001) Cellular and synaptic adaptations mediating opioid dependence. *Physiological Reviews* 81: 299–343.

Relevant Website

http://www.ncbi.nlm.nih.gov – NCBI.

Opsoclonus/Myoclonus and Neuroblastoma

M R Pranzatelli, Southern Illinois University School of Medicine, Springfield, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

What Is Opsoclonus–Myoclonus Syndrome?

Opsoclonus–myoclonus syndrome (OMS) is a rare, unique, and important disorder, affecting children and adults. The following observations are based on the experience and research of the National Pediatric Myoclonus Center, a world center for the evaluation and treatment of children with OMS, and on published literature.

Opsoclonus was described in 1913 and later in 1927 by Orzechowski. The association of opsoclonus and myoclonus as a pediatric syndrome was made in 1962 by Kinsbourne, who also identified adrenocorticotropic hormone (ACTH; corticotropin) as an effective therapy. Although Foster-Kennedy had described a child with probable opsoclonus and neuroblastoma in the early 1900s, it was not until 1968 that Solomon and Chutorian alerted clinicians to the fact that OMS heralds occult neuroblastoma. Since then, over 250 cases have been reported in the literature.

In the typical pediatric case, toddlers acutely develop a constellation of neurological abnormalities, which rapidly progress over days to weeks. Most striking is the gait ataxia or inability to walk. Parents report staggering or a drunken gait, with wide spacing of the legs and falling. The diagnosis is straightforward once opsoclonus, or randomly darting eyes (saccadomania), appears, although this may be one of the later features of the syndrome. This results in a misdiagnosis of acute cerebellar ataxia in the majority of cases. Orzechowski said "the ocular globes are in a state of continuous agitation, being shaken and increasingly displaced by very rapid and unequal movements, which generally take place in the horizontal plane." Opsoclonus may occur in spells or bursts, with eyelids open or closed in sleep. It is increased by saccadic movements or fixation and seldom decreased by fixation. The action myoclonus gives the appearance of tremulousness or quivering, which has been called polyminimyoclonus. The distribution of myoclonus in OMS may include the face, head, and neck; trunk; and limbs, specifically fingers and hands. It is evoked by action and other stimuli, such as noise, light, or visual threat, and is exacerbated by crying, excitement, or stress. Opsoclonus, myoclonus, and ataxia are usually all present in typical OMS, but there may be variability in the proportions of each and the sequence of appearance. The difference in severity between children is remarkable. In atypical cases, either opsoclonus or myoclonus may be absent or so trivial as to be missed. A minority of children, perhaps 5%, have convulsions.

Children with ataxia in combination with either opsoclonus or myoclonus should never be dismissed as having acute cerebellar ataxia. The child's extreme irritability and characteristic age vulnerability should prompt a search of occult neuroblastoma. Besides opsoclonus, myoclonus, and ataxia, the children exhibit rage, fragmented nonrestorative sleep, expressive language impairment (dysarthria to anarthria), and hyperacussis. The keenly observant clinician should be looking for hints of these features in a toddler with acute ataxia. Later sequellae often include obsessive–compulsive disorder, attention-deficit hyperactivity disorder, oppositional defiant disorder, and cognitive impairment, even mental retardation.

In contrast to the neurological disaster, the prognosis for surviving the tumor is excellent in OMS compared to neuroblastoma without OMS. The tumors are neural-crest derived and appear along the sympathetic chains. They may arise in the neck, chest, abdomen, or pelvis in paraspinal, retroperitoneal, or adrenal locations. Thin-cut high-resolution computed tomography (CT) scans probably have the highest diagnostic yield. A nuclear medicine metaiodobenzylguanidine (MIBG) scan is sensitive for neuroblastoma, but has some false negatives and positives. Urine catecholamine screening is only helpful if the levels are increased; the false negative rate is high. Serum neuron-specific enolase or ferritin may be elevated.

Other diagnostic studies are unhelpful. There may or may not be initial cerebrospinal fluid (CSF) pleocytosis, and protein and glucose are normal. Oligoclonal bands are usually absent. Head magnetic resonance imaging (MRI) studies reveal no evidence of any major cerebellar edema or mass lesion, and in most cases are normal. Electroencephalograms are usually normal, but may show nonspecific slowing.

The belief that OMS can be caused either by a tumor or a virus is based on series that show approximately 50% yield in being able to diagnose a neuroblastoma despite using modern imaging techniques. However, recent evidence challenges the idea that there are two different etiologies. First, on the grounds of logic, it is hard to rationalize that an identical syndrome could be produced by such disparate etiologies. Even an experienced clinician would not be able to tell on neurological grounds whether a

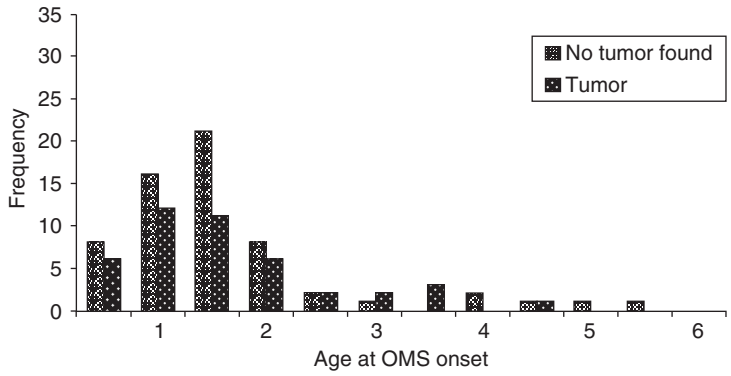

Figure 1 Frequency plot of age at OMS onset in all OMS cases according to presumed etiology. The no-tumor and tumor groups were not significantly different.

given child did or did not have a tumor. It is true that some viruses, DNA or RNA, have been documented in children with OMS. However, are they the primary etiologic factors or rather secondary triggers? A viral-like prodrome of vomiting or upper respiratory symptoms usually occurs before OMS, whether or not a tumor is found. Neuroblastoma is notorious for spontaneous involution; no other tumor has such a propensity. The possibility exists, therefore, that children may have already had neuroblastoma by the time their clinical symptoms prompt a look for the tumor. The argument that the tumor may have been eradicated by the immune system is based on the high degree of lymphocytic infiltration in neuroblastomas, especially those that are associated with OMS. Also, the distribution of onset age does not differ between groups (Figure 1). Whether or not a tumor is found does not appear to have a major effect on outcome.

There is also evidence that OMS is not a uniform disease. Subgroups can be defined on clinical and immunological levels. Children with mild presentation and a monophasic course often undergo complete remission with conventional treatment. The more common scenario is the child with one or more relapses, prompted either by infection or attempts to taper steroids or corticotropin. Whether these children have a multiphasic versus a monophasic illness that never was successfully eradicated is unclear. A group of children has relapsing-progressive disease. Mental retardation and significant motor and behavioral/cognitive disability is characteristic of this group.

What Causes the Autoimmune Response in OMS?

In building a hypothesis of etiopathophysiology, one must begin with the tumor – neuroblastoma is the only commonly occurring tumor in pediatric OMS – and work toward the neural targets. The assumption, which is backed by some data, is that an onconeural antigen, found both on the tumor and on embryologically related nerve cells, provokes an autoimmune disorder – a case of 'friendly fire' on the brain.

There may be more than one immunological defect. The tumor escaped immune surveillance; it was allowed to grow. In some children it grows so large that it is inoperable. In other children, however, the tumor may be only a few millimeters in size. Some genetic vulnerability may explain this, but there is no simple Mendelian inheritance. Family members with OMS, including twins, have never been reported. However, complex immunoregulatory gene defects would be more difficult to track down.

A second possible defect is that once the immune system does find the neuroblastoma, there seems to be an overreaction on some level. Fortunately, this is uncommon, occurring in only 3–4% of all neuroblastomas. But in those individuals, immune system dysregulation or a state of autoimmunity occurs. Secondary factors may further promote the development of an actual autoimmune disorder.

The currently favored immunologic hypothesis of OMS is that it is humorally mediated. This is due in part to various putative autoantibodies that have been described, but studies on T cells and other lymphocytes are lacking. Anti-Ri antibodies are associated with breast cancer in women with OMS, but have not been reported in pediatric OMS. Anti-Hu antibodies, found in neuroblastoma with or without OMS, probably do not cause OMS. Both IgM and IgG autoantibodies to high-molecular-weight neurofilament and other smaller antigens have been reported in OMS, but can also be found in some controls. Commercial screening for 'paraneoplastic' autoantibodies by Western blots in children with OMS has not been productive.

When, instead, the powerful new method SEREX (serological identification of antigen by recombinant

cDNA expression library) was employed, various other autoantigens were detected, but no one autoantigen was found in each of the cases of OMS studied. This method utilizes a cDNA library of a brain area to obtain the specific antigens from it. Another new tool, a newly introduced competitive binding assay, demonstrated surface binding of antibodies from OMS sera directed against the same autoantigen in different children with OMS. Autoantibodies bound to the surface of isolated rat cerebellar granular neurons. No study has demonstrated a relation between autoantibody titers and neurological severity. As a result, the intriguing autoantibody hypothesis of OMS is in question.

Although paraneoplastic autoantibodies have been injected into rodents and cavies, no signs of OMS developed – hence, no animal model of OMS. Also, because most children with OMS do not die from their condition, there have been limited postmortem studies. In some cases, there was evidence of brain inflammation; no abnormalities were found in others.

It had been hypothesized that B and T cells enter the brain, once activated by peripherally induced disease induction, and compromise brain function through production of autoantibodies or inflammatory cytokines, or by direct cell-to-cell toxicity. Thereby, normal immune cell trafficking can progress to lymphocyte recruitment. The National Pediatric Myoclonus Center validated this hypothesis in OMS by performing immunophenotyping of CSF lymphocytes through fluorescence-activated cell sorting (FACS), focusing on the same cell types known to infiltrate neuroblastomas and defend against tumors (natural killer cells and $\gamma\delta$ T cells, conventional T cells, and B cells). This was a cross-sectional study, so individuals at any stage in their disease were included. Each of the patients was videotaped, and the videotape was scored by a trained, blinded observer using the OMS Evaluation Scale, a 12-item scale rated from 0 to 3, with a maximum abnormality of 36 (Table 1). This scale is meant to provide an objective documentation of the motor abnormalities of OMS.

CSF exhibited several abnormalities. The frequency of CD19+ B cells was increased significantly. The frequency of CD4+ helper/inducer cells was decreased but that of CD8+ cytotoxic/suppressor T cells was increased. This resulted in a reduction in the CD4/CD8 T cell ratio. In some individuals, this was profound, resulting in a ratio <1:1 instead of the normal 3:1. Using the OMS evaluation scale to assess mild, moderate, and severe levels of abnormality, comparisons were made between the three OMS groups. CD4+ T cells decreased significantly with motor severity while the frequency of CSF B-cells (Figure 2) and $\gamma\delta$ T cells (unconventional T cells) increased with severity. The degree of cellular abnormality was variable. Exploring the B cell hypothesis further, B-cell expansion was more prevalent in acute cases (Figure 3) and tended to decrease over time. In chronic, symptomatic cases, however, the percentage did not return to normal. As to T cell activation in maturation markers, the only abnormality was an increase in HLA-DR+ activated T cells. No other cell abnormalities correlated with severity.

Table 1 OMS Evaluation Scale – motor performance

Item no.	*Description*
1	Walking: side-to-side imbalance
2	Walking: front-to-back imbalance
3	Walking: wide base
4	Instability while standing (feet apart)
5	Difficulty achieving standing position
6	Truncal instability while sitting
7	Targeting difficulty
8	Difficulty grasping with one hand
9	Difficulty with pincer grasp
10	Abnormal eye movements while tracking (fixation)
11	Abnormal eye movements while resting
12	Speech abnormality (dysarthria)

Videotapes were made to include a segment of each item needed for scoring. Each item is scored from 0 to 3 as follows: 0, normal; 1, mild; 2, moderate; 3, severe.

Analysis of different treatment groups yielded novel observations. Conventional immunotherapy, such as with steroids or intravenous immunoglobulins (IVIg), did not result in a statistically significant difference between groups. Children who had been treated previously with cyclophosphamide or multiagent chemotherapy for their tumors did not have a different CSF immunophenotype compared to those who were not treated. These data indicate that while conventional treatments and standard cancer chemotherapy may work by different mechanisms, they have not prevented the persistence of neurological problems. The blood mirrored some but not all of the immunologic abnormalities found in CSF. It reflected the changes in conventional CD4 and CD8+ T cells, but not in B cells. Whether these T cell abnormalities could have some role as treatment markers remains to be demonstrated by longitudinal studies.

Neopterin, a known indicator of cellular immune activation in a variety of disorders, was increased in OMS CSF, but tests are not sensitive enough to allow neopterin to serve as a biomarker of disease activity. Neopterin was more elevated in acute and untreated patients, and the magnitude of the increase was similar to that reported for central nervous system viral infections. No differences were found between the children who did or did not have a demonstrable

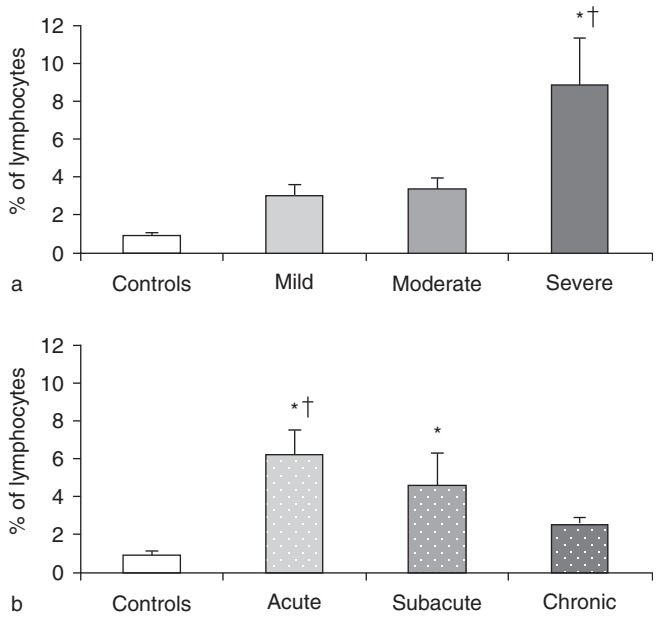

Figure 2 (a) Relation of neurological severity to percentage of B cells in cerebrospinal fluid. As an aid to clinical interpretation, neurological severity was defined as mild if the videotape total score was $\leq$12, moderate if 13–24, and severe if 25–36. (b) Relation of syndrome duration to the percentage of cerebrospinal fluid B cells. The syndrome duration category was defined as acute if $\leq$3 months, subacute if 3–12 months, and chronic if $>$1 year. Controls are shown for comparison. Asterisks signify significant difference compared with controls. Crosses indicate significant difference among duration categories, $P < 0.05$ by Duncan test. From left to right, $n = 18$, 8, 10, and 18. As a point of comparison, neurological severity (total score) also varied with the duration of the illness ($F_{2,51} = 4.6$; $P = 0.014$, ANOVA). Scores in the chronic phase (11.8 ± 8.8) were about one-half of those in the acute and subacute phases ($P < 0.05$, Duncan test).

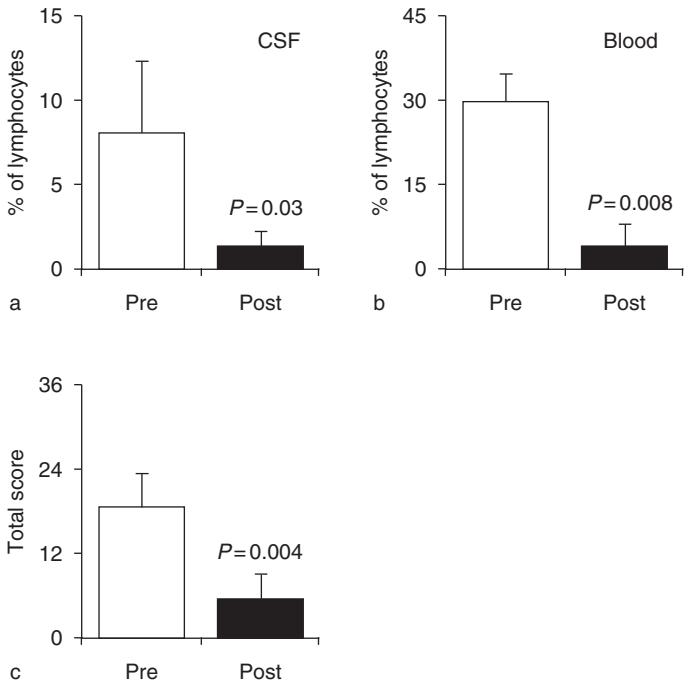

Figure 3 Effects of rituximab on cerebrospinal fluid (CSF) B cells and severity of OMS. Six months after completion of rituximab infusions, CSF (a) and blood B cells (b) were depleted. The total score (c) decreased from the upper limit of moderate to the upper limit of mild categories. Clinical severity was classified as mild if the total score was 0–12, moderate if the score was 13–24, and severe if the score was 25–36.

tumor. Previous studies on CSF neopterin revealed that it rises after allogenic T cell activation, but monocytes and macrophages also can be induced by cytokines to produce neopterin. Elevated neopterin supports the autoimmune hypothesis of OMS.

Neural Targets in OMS

The cerebellum, which contains half of all brain neurons, is a primary target in OMS, but whether it explains all features is less clear. It plays a role in language development and motor learning, and some behavioral or affective symptoms may be cerebellar in origin. An adult with infantile-onset OMS (41 years earlier) exhibited cerebellar vermis atrophy. The patient had gait ataxia as well as truncal and appendicular ataxia. Cerebellar atrophy also was reported in some children with OMS and neuroblastoma, but is not typical. Using conventional proton MR spectroscopy, however, no abnormal cerebellar spectra have been identified.

As to opsoclonus, the consensus is that the mechanisms of saccadic eye movements reside in the brain stem, but the contrary view is cerebellar. Mesencephalic and pontine ocular gaze centers contain burst cells and omnipause cells that control saccadic eye movements. Burst neurons are silent until just before or during a saccade, when they drive ocular motor neurons to create saccades. Omnipause neurons cease firing when burst cells are firing and inhibit burst cells during fixation. Opsoclonus could be due to excessive activity of burst neurons or lack of inhibition.

Myoclonus can originate from cortical and subcortical structures. Limited neurophysiological studies in OMS identified the myoclonus as subcortical in origin, consisting of < 100 ms synchronous discharges without back-average cortical correlate or exaggerated responses to evoked-potential testing or long-latency electroencephalographic/electromyographic responses. In the brain stem, the nucleus gigantocellularis reticularis of the medulla is a critical nucleus for myoclonus. Myoclonus is often associated with ataxia, but most cerebellar disorders are not associated with myoclonus. The special vulnerability of the infant and toddler may relate to parallel, progressive phylogenetic development of the neocerebellum and frontal cerebral cortex, two of the most slowly maturing brain areas. Frontal connections of the cerebellum are crucial for language development. Lesions of the cerebellar vermis in children are associated with mutism (anarthria), irritability, aggressive behavior, poor socialization, language disorder, and possible autistic-like features. Mental retardation is a feature of many pediatric cerebellar ataxias, particularly when vermis foliation is hypoplastic.

It has been proposed that the cerebellum is the seat of motor learning through the mechanism of climbing fiber synapses on Purkinje cells. However, lesions of the inferior olives, from which climbing fibers emanate, prevent motor learning, suggesting that the inferior olives are principally involved. The cerebellum contributes to oculomotor plasticity. There is some evidence that the cerebellum also has a role in cognitive behaviors besides motor learning, for which the repository site of learning may be the cerebellar target areas rather than the cerebellum itself. These behaviors include spatial learning (connections between cerebellum and frontoparietal association cortex, limbic system, and superior colliculus), discrimination learning (projections to cerebellum via pontine nuclei), and emotions such as fear (reticular activating system, limbic system, and hypothalamus). The cerebellum also has a role in language development, as early cerebellar lesions may result in mutism. An active role of the cerebellum in early language development, not just speech articulation, is supported by studies in humans and experimental animals.

How do these structures and their systems interlinkages generate the physiological phenomena that characterize the syndrome? In the absence of functional studies, the answer is unknown. One could speculate that interruptions of cerebellar Purkinje cell firing (whether due to direct immunological assaults on Purkinje cells or other cerebellar neurons, such as granule cells) give rise both to the dyskinesias and ataxia. Concurrent brain stem involvement through cerebellar circuits would explain the opsoclonus. Cognitive impairment, attention deficit disorder, and seizures imply broader cerebral involvement, either as a result of indirect cerebellar-frontal connections, or due to direct effects of the immune system on these more remote sites. Integrating the various phenomena probably requires dysfunction of a diffuse neural network originating in the cerebellum with brain stem and basal ganglia-frontal connections.

In summary, it would seem that a neuroblastoma can result in at least a cerebellar injury, especially involving the vermis, but network failure involving remote neural sites is also likely. Proper functioning of Purkinje cells, the fastest firing brain neurons and main outflow to deep cerebellar nuclei, is requisite to preventing movement disorders.

How Should the Tumor Be Handled in OMS?

Should the neuroblastoma be removed or left alone? Until recently, it has been removed without question. The argument for tumor resection, as opposed to

biopsy, includes reduction of the antigen load that presumably precipitated OMS, and preventing tumor progression. Certain tumors encroaching on spinal foramina or otherwise causing damage present urgent problems and are not at issue. Some parents really want the tumor to be taken out for peace of mind. But children can come to harm as the result of the resection, developing a postsurgical Horner's syndrome (miosis, ptosis, anhydrosis), in the case of chest or neck resections, or plexopathy, neuropathy, or an unstable spine, causing scoliosis.

The argument for leaving the tumor alone is to prevent iatrogenic injury, reliance on scanning to verify gradual tumor involution or absence of progression, and reluctance to perturb the relation of the immune system and the tumor. Also, neuroblastomas in OMS rarely increase in stage. In the International Neuroblastoma Staging System (INSS), they are usually stage 1, sometimes stage 2, and seldom stages 3 or 4. Histologically, they become more mature (ganglioneuroblastoma or ganglioneuroma) over time.

One newer alternative to surgery is the use of radioactive iodine-labeled MIBG to shrink the tumor when it has wrapped around vital structures and is inoperable. Multiagent cancer chemotherapy is the more standard approach. The use of chemotherapy in OMS for neuroblastoma *per se* is also controversial. One view is that the neuroblastoma tends to be more indolent and low-stage tumors do not require chemotherapy.

An important management debate is which should take priority: treating the tumor or the OMS? Because of the devastation caused by delaying immunotherapy for OMS, the National Pediatric Myoclonus Center advocates treatment as soon as the diagnosis is made. For all the reasons previously discussed, the tumor has lower priority. This view can be difficult for oncologists to accept, especially because neuroblastoma without OMS is so aggressive and lethal.

Current Immune Therapies for OMS

Conventional Therapy

The choice of immunologic agents and overall strategy as a treatment remains one of the greatest controversies in OMS. Corticotropin, intravenous immunoglobulins, and corticosteroids are the typical conventional agents. In the experience of the National Pediatric Myoclonus Center, high-dose ACTH is superior to corticosteroids in inducing a neurological remission, but there are no controlled trials. The use of pulse dosing with dexamethasone or methylprednisolone has gained in popularity among steroid prescribers. Having seen the pitfalls of using single or even dual conventional therapies in children with moderately severe to severe disease, the Center no longer accepts them as adequate therapy. Others have had similar negative experiences. This is not to say that corticosteroids, IVIg, or ACTH could not be used alone in the mildest of cases, but moderately severe and severe cases usually require some form of chemotherapy or alternative treatment.

The wait-and-watch approach is popular but dangerous. Sometimes no treatment is started until months after tumor resection, based on the belief that resection alone will stop OMS. Actually, two-thirds of such cases do not improve or get worse after surgery. Then a single agent is initiated; if the child does not do well, perhaps a second conventional agent is added. Some clinicians will consider the use of chemotherapy, but many will not. Children who do not have a tumor are at the highest risk of not being offered chemotherapy for their neurological syndrome. This is unfortunate. Probably, there is a window of opportunity for successfully eradicating the underlying autoimmune disease without allowing permanent brain injury to occur. Once mental retardation develops, it is not reversible.

Newer Approaches

Given the data on various putative autoantibodies and CSF B cell recruitment in OMS, the National Pediatric Myoclonus Center tested the theory of B cell autoaggression in OMS. The first step was to determine whether an anti-B-cell immunotherapy eradicated CSF B cells, and what the effect of such treatment would be on the clinical course. The chimeric anti-CD20 monoclonal antibody rituximab, approved in the United States for the treatment of B cell lymphoma, not only eradicated B cells in blood, as had been known previously, but it significantly lowered the frequency of B cells in CSF. Many patients after treatment had undetectable B cells. A few children did not respond or need retreatment a year later, but for most, rituximab seemed to confer lasting beneficial effect on the frequency of CSF B cells. The relation between clinical and immunological effects was a reduction in total score on the OMS Evaluation Scale (Table 1). Before treatment, the average total score was in the moderate severity range, whereas after therapy, it was in the mild severity category. This will be apparent to the practitioner.

Results of CSF lymphocyte immunophenotyping can be very useful in determining what type of chemotherapy should be given. FACS is available at most major medical centers, therefore the procedure could be adopted widely. If flow cytometry is not available, the clinical rule of thumb is that if a child does not completely return to normal on ACTH therapy, there

is always active underlying autoimmune disease. Because conventional agents do not alter the frequency of pathological lymphocytes in CSF, and presumably brain, the child is at risk for ongoing brain injury and relapse when the conventional agents are tapered or removed. Because no one knows exactly how long the therapeutic window of opportunity is open in OMS, the National Pediatric Myoclonus Center is a strong proponent of front-end multimodality therapy, combining ACTH, IVIg, and rituximab, for example. In the most difficult cases, other agents or approaches may be necessary as well, depending on the initial response.

A protocol that combines cyclophosphamide with oral corticosteroids and IVIg also is being evaluated. However, in a pilot FACS study, cyclophosphamide as add-on therapy did have anti-B-cell properties in the CSF, but did not reduce B cells to normal levels at the doses used in this protocol. This suggests that the combination will be more useful in milder cases. Dose-response studies are needed. Children with the severest disease require the use of multiple types of immunotherapy over a few to several years to put their disease into remission. Mycophenolate mofetil, which has replaced azathioprine in inflammatory diseases, does not reduce the frequency of CSF B cells in OMS. Other agents that have been tried are 6-mercaptopurine, methotrexate, and cyclosporine.

Given the humoral theory of OMS, plasma exchange (PEX) would seem logical. It has been tried empirically; case reports are positive but sparse. A newer strategy is to couple PEX with immunoadsorption, a technique that utilizes a column of staphylococcal A protein, to which IgG binds avidly. Toddlers are not ideal candidates for this procedure because of their lower blood volume – they may need to be transfused – but it can be done safely in older children through a large-bore central catheter in six to eight sessions over several weeks. How PEX works is not well understood, as it does not seem to remove enough Ig to rid the body of autoantigens. It can modify the immunophenotype, as well as render antigen–antibody complexes more vulnerable to immune-mediated destruction.

Future Therapies

Once the responsible onconeural antigen(s) is found, it will be possible to devise specific therapies for OMS, such as a T cell vaccine. Meanwhile, advances in biological response modifiers will probably be used to advantage in OMS. These could include blockers of type 1 T helper cell (Th1) cytokines or treatment with Th2 cytokines. Newer anti-B-cell monoclonal antibodies, with greater efficacy at sanctuary sites, also will become available.

Summary

Many converging lines of evidence support the autoimmune theory of OMS. Cancer-induced immunity cross-reacts with select neural tissues to engender a paraneoplastic syndrome. The multiple immunological abnormalities of OMS involve B and T cells, sometimes persisting for years, and show significant individual variability. CSF B and T cell recruitment is directly linked to neurological dysfunction and disease duration, possibly accounting for relapses and disease progression. As a result, B and T cells are useful biomarkers of disease activity in OMS. The observation that the tumor and no-tumor-found groups were not immunologically different based on the phenotype of CSF lymphocytes suggests a uniform paraneoplastic model of primary disease pathogenesis, with possible secondary provocative roles for viral infections and immunizations. Because OMS is paraneoplastic until proved otherwise, the neurological syndrome should not be treated differently based on whether one does or does not find an underlying tumor. The data also clearly indicate that tumor therapy, including both tumor resection and conventional cancer chemotherapy, usually is insufficient as treatment for this paraneoplastic disorder. As a result, there is a need for more selective and potent immunological agents, which will be the future of OMS therapy.

See also: Autoimmune Autonomic Neuropathy; Dystonia: Myoclonus–Dystonia; Neural Crest; Tumors of the Brain and the Spinal Cord.

Further Reading

Bataller L, Rosenfeld MR, Graus F, et al. (2003) Autoantigen diversity in the opsoclonus-myoclonus syndrome. *Annals of Neurology* 53: 347–353.

Blaes F, Fuhlhuber V, Korfei M, et al. (2005) Surface-binding autoantibodies to cerebellar neurons in opsoclonus syndrome. *Annals of Neurology* 58: 313–317.

Connolly AM, Pestronk A, Mehta S, et al. (1997) Serum autoantibodies in childhood opsoclonus-myoclonus syndrome: An analysis of antigenic targets in neural tissues. *Journal of Pediatrics* 130: 878–884.

Cooper R, Khakoo Y, Matthay KK, et al. (2001) Opsoclonus–myoclonus-ataxia syndrome in neuroblastoma: Histopathologic features – a report from the children's cancer group. *Medical and Pediatric Oncology* 36: 623–629.

Kinsbourne M (1962) Myoclonic encephalopathy of infants. *Journal of Neurology, Neurosurgery & Psychiatry* 25: 221–276.

Mitchell WG, Davalos-Gonzalez Y, Brumm VL, et al. (2002) Opsoclonus-ataxia caused by childhood neuroblastoma: Developmental and neurologic sequelae. *Pediatrics* 109: 86–98.

Nishihira H, Toyoda Y, Tanaka Y, et al. (2000) Natural course of neuroblastoma detected by mass screening: A 5-year prospective study at a single institution. *Journal of Clinical Oncology* 18: 3012–3017.

Pranzatelli MR (1992) The neurobiology of opsoclonus-myoclonus. *Clinical Neuropharmacology* 15: 186–228.

Pranzatelli MR (1996) The immunopharmacology of the opsoclonus-myoclonus syndrome. *Clinical Neuropharmacology* 19: 1–47.

Pranzatelli MR (2000) Friendly fire. *Discover* April: 35–36.

Pranzatelli MR, Tate ED, Dukart WS, et al. (2005) Sleep disturbance and rage attacks in opsoclonus-myoclonus syndrome: Response to trazodone. *Journal of Pediatrics* 147: 372–378.

Pranzatelli MR, Tate ED, Kinsbourne M, et al. (2002) Forty-one year follow-up of childhood-onset opsoclonus-myoclonus: Cerebellar atrophy, multiphasic relapses, response to IVIG. *Movement Disorders* 17: 1387–1390.

Pranzatelli MR, Tate ED, Travelstead AL, et al. (2005) Immunologic and clinical response to rituximab in a child with opsoclonus-myoclonus syndrome. *Pediatrics* 115: e115–e119.

Pranzatelli MR, Travelstead A, Tate ED, et al. (2004) B- and T-cell markers in pediatric opsoclonus-myoclonus syndrome: Immunophenotyping of CSF lymphocytes. *Neurology* 62: 1526–1532.

Pranzatelli MR, Travelstead AL, Tate ED, et al. (2004) CSF B-cell expansion in opsoclonus-myoclonus syndrome: A biomarker of disease activity. *Movement Disorders* 19: 770–777.

Riva D and Giorgi C (2000) The cerebellum contributes to higher functions during development. *Brain* 123: 1051–1061.

Rudnick E, Khakoo Y, Antunes NL, et al. (2001) Opsoclonus–myoclonus-ataxia syndrome in neuroblastoma: Clinical outcome and antineuronal antibodies – a report from the Children's Cancer Group Study. *Medical and Pediatric Oncology* 36: 612–622.

Solomon GE and Chutorian AM (1968) Opsoclonus and occult neuroblastoma. *New England Journal of Medicine* 279: 475–477.

Tate ED, Allison T, Pranzatelli MR, et al. (2005) Neuroepidemiologic trends in 105 US cases of pediatric opsoclonus-myoclonus syndrome. *Journal of Pediatric Oncology Nursing* 22: 8–19.

Relevant Website

http://www.omsusa.org – Opsoclonus–Myoclonus U.S.A. and International.

Optic Nerve, Optic Chiasm, and Optic Tracts

R A Rachel, National Cancer Institute – Frederick, Frederick, MD, USA
B E Reese, University of California at Santa Barbara, Santa Barbara, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

The projection neurons of the vertebrate retina are the ganglion cells, which extend axons across the inner surface of the retina in the optic fiber layer and exit the eye via the optic nerve head. These optic axons subsequently course along the optic nerve to the base of the brain at the optic chiasm, where, with optic axons from the opposite nerve, they intermingle and sort to form the optic tracts. The final destination of these axons is the target visual nuclei in the brain (Figure 1). A small proportion of optic axons course in an accessory optic tract posterior to the main optic tract (and innervate a distinct group of accessory optic nuclei), and there exist a small number of centrifugal fibers innervating the retina as well as commissural fibers of the supraoptic tracts. The contributions of the accessory optic, centrifugal, and supraoptic axons vary depending on the species and developmental age (there is also a retino-retinal projection that is entirely eliminated during development), and are not be considered further in this article.

Organization of the Optic Pathway

The retinofugal pathway establishes synaptic connections in both the diencephalon and the mesencephalon, where axons or collateral branches leave the pathway and innervate target visual nuclei. Different optic axon classes, arising from morphologically and physiologically distinct retinal ganglion cell classes and often discriminable by their size, depart the pathway at different locations to innervate distinct targets: for example, the retinal projection to the hypothalamus departs the pathway at the optic chiasm, while the projection to the midbrain extends through a continuation of the optic tract to the superior colliculus (optic tectum). This optic innervation forms a 'retinotopic map' in some of these target nuclei, by which the array of ganglion cells across the retina is faithfully replicated by the order of their synaptic connections within the target. Retinotopic precision in primates is greatest in the projection to the dorsal lateral geniculate nucleus, the thalamic relay nucleus that projects upon the primary visual cortex.

In mammals, many optic axons from the temporal retina do not cross the midline of the optic chiasm but rather turn to project into the optic tract ipsilateral to their origin. The area of the temporal retina giving rise to this nondecussating, or uncrossed, projection varies across species, dependent on the extent of binocular overlap within the visual field and constrained by the position of the eyes in the head. This 'partial decussation' at the optic chiasm enables the binocularly congruent parts of the two retinas to project directly to the same side of the brain, where processing of the two retinal images in the visual cortex mediates stereoscopic depth perception. Mammals with forward-facing eyes tend to have greater binocular overlap though a narrower field of vision relative to those with laterally placed eyes (Figures 2(a) and 2(b)). The balance between these parameters in any species reflects the behavioral ecology of the organism. For example, rodents and rabbits have laterally positioned eyes, enabling a visual field adapted to viewing a broad expanse while monitoring for predators from above and behind (Figure 3). Additionally, in such species, the proportion of cells in the temporal retina giving rise to uncrossed optic axons is as low as 10%. This proportion rises in species with forward-facing eyes, such that virtually every ganglion cell in the temporal retina of primates projects ipsilaterally. The resulting narrower field of vision in primates, on account of having eyes in the front of the head, is partially offset by the ability to turn the head in many directions. The greater proportion of retinal ganglion cells with uncrossed optic axons in the temporal retina in primates, coupled with their smaller receptive fields, renders the population of binocularly driven cortical neurons sensitive to slight retinal disparities, thereby providing the substrate for fine stereoscopic depth perception.

The partial decussation of optic fibers at the chiasm is a feature unique to mammals. In most nonmammalian vertebrates, such as the favorite experimental models of fish and chick, the eyes are positioned laterally, maximizing the breadth of the visual field and minimizing or eliminating regions of binocular overlap. This arrangement of the eyes with respect to the brain results in an optic chiasm with no decussating fibers and a completely crossed retinal projection. Some nonmammalian species with frontally placed eyes, such as owls and frogs, make use of other neural pathways that cross the midline at locations central to the retinofugal targets in order to achieve binocularity.

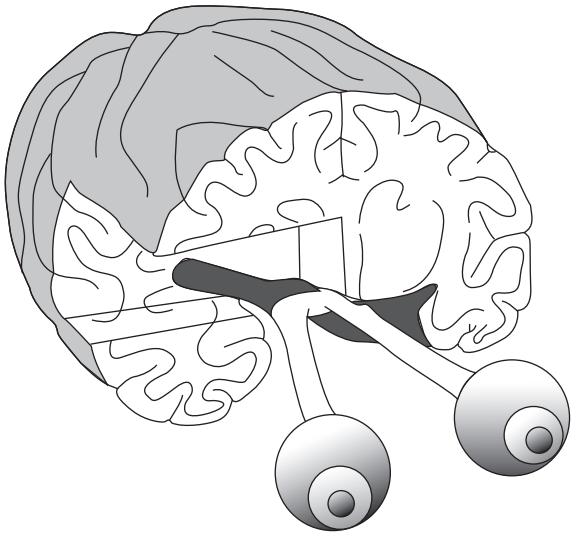

Figure 1 The optic nerves, chiasm, and tracts are defined by the course of the optic axons as they project from eye to brain. The optic pathway becomes fused with the base of the brain in the region of the optic chiasm (dark stipple, posterior chiasm and tract). The optic nerve (light stipple, anterior chiasm and nerve) comprises the second cranial nerve. Unlike the other cranial nerves, the optic nerve is a part of the central nervous system (as is the retina), being a developmental outgrowth of the ventral diencephalon. Modified from Hoyt WF and Lewis O (1963) The primate chiasm. *Archives in Ophthalmology* 69: 69–85.

Fiber Organization in the Optic Nerve, Chiasm, and Tract

Traditional accounts of the primary visual pathway assumed that optic axons coursed in a retinotopically ordered fashion along the length of the optic nerve, and that axons arising from binocularly conjugate retinal loci were brought together in the optic chiasm to yield a fiber order within the optic tract that accurately anticipated the binocular and retinotopic features of the visual targets. Clinical observations following focal damage to the optic pathway were, however, often incompatible with those traditional accounts: for example, selective impairments of distinct visual functions, affecting binocularly disparate retinal loci, were often a consequence of partial damage to the optic tract, rather than precise, binocularly conjugate, blind fields ('scotomas'), as should be expected from standard textbook accounts of the visual pathways. Subsequently, neuroanatomical investigations in mammals revealed that optic axons are only very coarsely organized into a retinotopic fiber order at the optic nerve head, and that this order deteriorates along the course of the nerve. In particular, the fibers arising from the nasal and temporal half-retinas are intermingled extensively in many mammalian species,

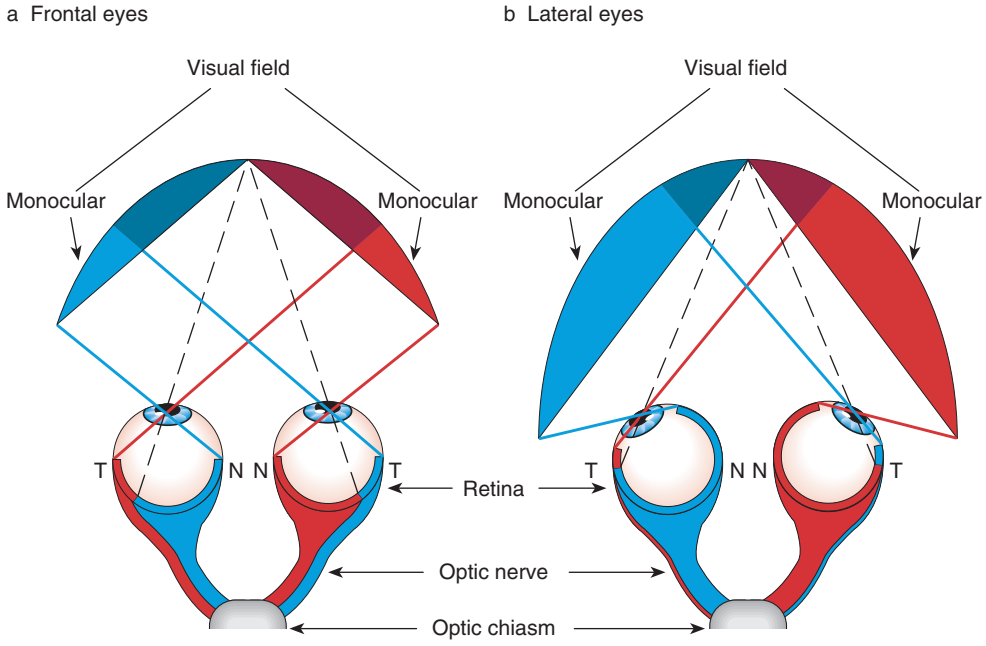

Figure 2 The visual field of mammalian species with frontally facing eyes (a) displays more binocular overlap but a narrower overall field of vision compared to mammals with laterally placed eyes (b) who tend to have a smaller binocular overlap but a wider overall field of vision. Red represents information from the right visual field and blue represents information from the left visual field. Note that both eyes transmit visual information from both visual hemifields; in animals with frontal eyes, this information comes more equally from both eyes, while in mammals with lateral eyes, one eye carries most of the information from a single hemifield. N, nasal; T, temporal. Modified from Jeffery G and Erskine L (2005) Variations in the architecture and development of the vertebrate optic chiasm. *Progress in Retinal and Eye Research* 24: 721–753.

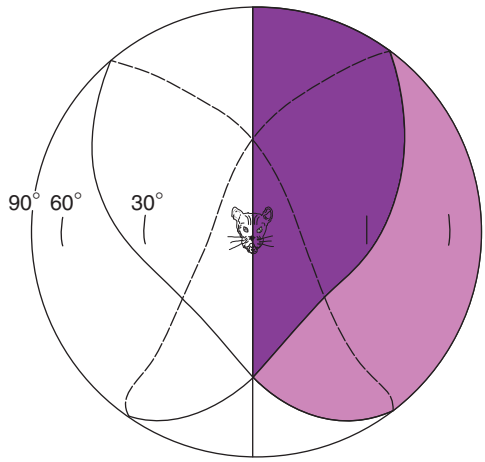

Figure 3 One visual hemifield is shown for the rat, with the binocular portion (dark purple) viewed by the right eye's temporal retina (situated ventrotemporally as a crescent upon the retina) and the left eye's corresponding nasal retina (situated between that eye's temporal retina and the optic nerve head). The remaining portion of the left eye's retina nasal to the optic nerve head views the peripheral (monocular) visual field (light purple). Despite the relatively lateral position of the eyes in the head, the rat's binocular visual field at its widest extent approaches 100°s of visual angle. Modified from Hughes A (1979) A schematic eye for the rat. *Vision Research* 19: 569–588, and from Reese BE (1988) 'Hidden lamination' in the dorsal lateral geniculate nucleus: The functional organization of this thalamic region in the rat. *Brain Research Reviews* 13: 119–137.

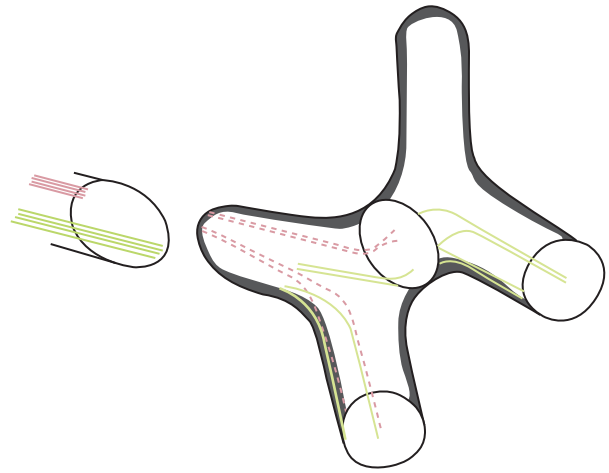

Figure 4 Optic axons become segregated by functional class (two of which are indicated here in different colors) as they pass through the chiasmatic region. This is evidenced in all adult mammalian species as a segregation of axons according to their size within the optic tract. Modified from Reese BE and Ho K-Y (1988) Axon diameter distributions across the monkey's optic nerve. *Neuroscience* 27: 205–214.

so that fiber position in the nerve cannot be a basis for the partial decussation at the optic chiasm. Further, an entirely different organizing principle is found central to the optic chiasm: while functionally distinct optic axons are intermingled within the optic nerve, they are partially segregated within the tract (Figure 4). In addition, those distinct axon classes regain some degree of retinotopic fiber order upon entering the tract. Rather than being a passive conduit for bestowing a retinal map upon the visual targets by maintaining precise neighbor relationships, the optic pathway displays conspicuous fiber rearrangements along its length. These observations on the mature visual pathway indicate that optic axons must change their ordering as they navigate a course from eye to brain during development.

Development of the Optic Pathway

Axonal Outgrowth and Optic Nerve Formation

The embryonic retina develops as an out-pouching of the diencephalon, the optic vesicle. This vesicle gradually invaginates to produce the optic cup, the inner surface of which becomes the germinal neuroepithelium producing all of the neurons of the mature retina; the outer surface becomes the nonneuronal pigmented epithelium. The neck of the optic vesicle, the optic stalk, differentiates into the optic nerve as the first postmitotic cells of the retinal neuroepithelium (the ganglion cells) extend axons, which grow among the cells of the stalk within its ventral wall. Those stalk cells, in turn, differentiate into the glia of the developing optic nerve, yielding a mature nerve composed of fascicles of axons and interfascicular astrocytes and oligodendrocytes. Electron microscopic and dye-labeling studies have confirmed that growing optic axons frequently change position between the fascicles, so that neighbor relationships are not maintained along the developing nerve. By the time the growing optic axons arrive at the base of the brain, those from different quadrants of the retina are conspicuously intermingled in the developing rodent and carnivore optic nerve.

Decussation Decisions at the Optic Chiasm

The microenvironment of the developing optic pathway changes dramatically at the base of the brain: axons no longer course in fascicles amid glial cell bodies, but rather course subpially among radial glial processes which extend from somata at the ventricular surface. As the axons from the two nerves meet, those from the nasal retina of each eye cross the midline; those from the temporal retina, by contrast, adopt an uncrossed course. Distinct populations of glial processes and neurons inhabit this 'decision region' prior to and during axonal invasion, and within this region the growth dynamics of newly arriving optic axons change, evidenced by the changing morphology and advance rate of their growth

cones. *In vitro* assays implicate a membrane-bound molecule to which only the temporal retinal growth cones are sensitive, acting as a contact-mediated inhibitory substrate, probably via a receptor–ligand interaction. One such ligand–receptor pair mediating divergence at the optic chiasm has recently been identified as belonging to the Eph/ephrin family, which is known to mediate a number of neuronal and glial interactions in the developing nervous system.

Curiously, a portion of optic axons from the temporal retina of hypopigmented eyes are misdirected at the chiasmatic midline, crossing the midline to project aberrantly into the opposite optic tract, suggesting a reduced sensitivity to the midline signal. Such a reduced sensitivity of axons from temporal retina may also occur during later development in normally pigmented nonprimate mammals, since the later-generated cohorts of retinal ganglion cell in these species extend axons that all decussate, irrespective of retinal locus. The molecular signature of ganglion cells from temporal retina during early development is different from that of their decussating neighbors forming at later developmental stages and contributes to the ability of those later-extending temporal fibers to cross the midline. The transient expression of Zic2, a zinc finger transcription factor, in retinal ganglion cells giving rise to uncrossed axons is necessary to prevent these fibers from crossing the midline at the optic chiasm. Thus, like other neuronal populations in the developing brain and spinal cord, retinal axons are molecularly specified to respond to signals at midline choice points such as the optic chiasm.

With respect to development of the optic chiasm, a unique but experimentally informative species is the frog, which initially has an entirely crossed visual projection. Beginning at metamorphosis, as the eyes shift from the side of the head to the front, the transforming tadpole develops a small ipsilateral component arising from the temporal retinal periphery. This projection is thyroxine dependent and corresponds with the appearance of a thyroxine-responsive stem cell population in the peripheral retina.

Chronotopic Fiber Reordering

The axons of progressively later-generated cohorts of retinal ganglion cells (destined to differentiate into distinct functional classes) grow amid the older optic axons down the nerve, but as they enter the optic chiasm and tract, they become segregated according to their time of arrival, creating a 'chronotopic' map. This is evidenced in electron micrographs by the accumulation of growth cones in the subpial parts of the optic tract, and by the selective labeling of older axons in the deeper parts of the tract. The fiber reordering has been related to the changing glial environment mentioned earlier, and to an associated changing molecular milieu in which proteoglycans create an unfavorable environment for newly arriving axons. This chronotopic reordering of optic axons ultimately yields a mature optic tract in which functionally distinct optic axons, having differentiated into their discriminable size classes, are segregated (Figure 4).

Optic Tract Formation: Retinotopic Reordering and Positioning

As the optic axons enter the tract, they regain a degree of retinotopic order that is not present in the prechiasmatic optic nerve: axons from dorsal and ventral retina are intermingled within the nerve, yet they become segregated across the width of the optic tract, establishing a coarse mapping of this retinal axis. The cause of this retinotopic reordering is unclear, but since the change is relatively sudden, it is likely to be triggered by some changing feature of the immature diencephalon. As this new fiber order, created afresh in the pathway at the base of the brain, presages the orderly termination of axonal arbors in the target nuclei, a role in the formation of retinotopic maps within the targets is suggested.

Independent of signals that reorder the optic axons relative to one another, other cellular and molecular features of the perichiasmatic region may be critical for determining the locus of pathway formation. For example, embryonic neurons in the chiasmatic region may provide a template for the future optic pathway. Ablation of these neurons prevents pathway formation, leaving optic axons stalled within the prechiasmatic nerves. Mutations producing a completely uncrossed optic pathway (shown in both humans and in dogs, in which the optic nerves fail to fuse) may reflect an anomalous specification of these cells or their surface properties that normally defines the characteristic crossing of the retinal axons. Other local features at the surface of the diencephalic neuroepithelium may provide a template for the developing optic tract, defined by both growth-promoting and unfavorable substrates for axonal elongation. These appear to be independent of other fiber tracts in the vicinity (at least in amphibians), since manipulations that allow optic axons to invade a less mature diencephalon, prior to the formation of these other pathways, do not prevent the formation of the optic tract.

Target Innervation

As optic axons invade their targets, their growth rate slows down and the morphology of their growth cones increases in complexity as they begin to form terminal arborizations. The interactions leading

to target recognition, axonal invasion, and synapse formation are still relatively undefined, but are likely to be different in midbrain and thalamus, suggested by *in vitro* studies, transplantation studies, and studies of their normal and abnormal patterns of innervation. Innervation of the lateral geniculate nucleus occurs via collateral sprouting after optic axons have extended within the optic tract beyond the level of the target, suggestive of a precise target-recognition mechanism. Yet, deprived of normal visual targets, optic axons will elaborate axonal arbors and form retinotopically organized synaptic connections in nearby nonretinal targets in the thalamus, provided that those nuclei have been suitably deafferented. Innervation of the optic tectum, by contrast, may be mediated by a longer range diffusible attraction, since transplantation studies have shown that this target can guide optic axons to it via novel (nonoptic) pathways. As occurs at the optic chiasm and in other regions of the central nervous system, receptor–ligand interactions are important for setting up the initial retinotopic map in target regions of the diencephalon. Molecules of the Eph/ephrin family, along with transcription factors, extracellular matrix proteins, neurotrophic factors, and cell death-regulating factors, including gradients of guidance molecules in the target regions responding to corresponding gradients in the retina, all contribute to the process of topographic mapping.

The foregoing developmental considerations are of interest not only because they begin to clarify the mechanisms governing axonal navigation and pathway formation: they also make comprehensible the mature organization of the optic nerve, chiasm, and tract. This in turn provides a fuller understanding of the disorders of vision that result following focal damage to the pathway.

Genetic Mutations in Pigment-Related Genes Reduce Binocular Vision by Reallocating Retinal Fibers at the Optic Chiasm

As mentioned previously, in mammals with reduced pigmentation, such as found in albino species, a portion of the uncrossed retinal projection fails to be inhibited by the midline at the optic chiasm and incorrectly crosses into the contralateral optic tract with the normally decussating fibers (Figure 5). Decades of study have revealed the precise wiring defects that result from the aberrant crossed projection within the lateral geniculate nucleus, as well as the downstream consequences within the primary visual cortex. These studies have determined that the fiber abnormality in hypopigmentation mutants arises from a lack of melanin in the eye rather than being due to any abnormality within the optic nerve or chiasm. Indeed, a study of mouse lines having mutations in genes that decrease pigment specifically in either the retinal pigment epithelium or in the outer choroidal layer of the eye have confirmed that only those with reduced or absent pigment in the retinal pigment epithelium display a smaller uncrossed retinal projection indicative of a fiber misrouting at the optic chiasm.

A number of inherited human diseases affect both pigmentation and vision, on account of absent or abnormally produced melanin in the retinal pigment epithelium. The visual system defects in these individuals include hypersensitivity to light, nystagmus, reduced visual acuity, and reduced binocular vision and depth perception, the last because of the retinal axon misrouting defect, thus leading to severe visual impairment in affected individuals. Similar abnormalities result from mutations in a variety of different

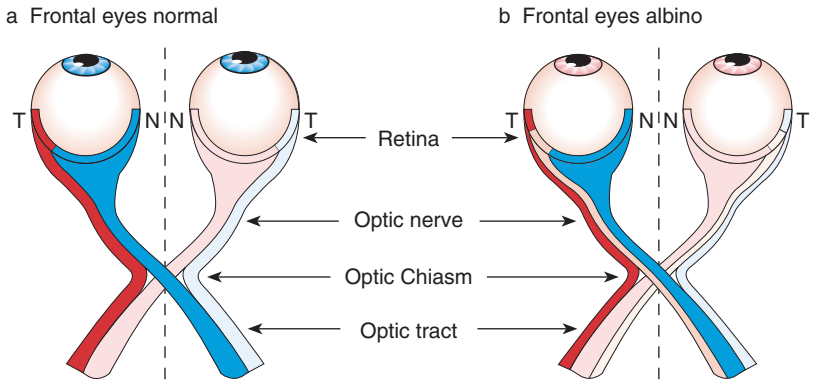

Figure 5 As in Figure 2, information from the right visual field is represented in red while that from the left visual field is depicted in blue. An albino individual from a mammalian species with frontal eyes (b) has reduced binocular vision due to the reallocation of fibers and a misrouting of some normally uncrossed fibers at the optic chiasm (represented by lighter coloring of fibers in the optic nerve, chiasm, and tract). N, nasal; T, temporal.

genes, the common feature of each phenotype being reduced or abnormal melanin in the retinal pigment epithelium. These genetic disorders include the tyrosinase-negative albino, in which no melanin is formed in the skin, hair, or eyes. The albino patient represents the most severe form of oculocutaneous albinism. Mutations in other genes of the melanin synthesis and melanosome biogenesis pathways likewise result in the visual disturbances found in albinism. Hermansky–Pudlak syndrome (HPS) results from mutations in a cluster of genes involved in various aspects of melanosome biogenesis. Ocular albinism (OA1) is yet another melanosome biogenesis disorder that affects vision. While treatment for these disorders is not yet possible, we do have mouse models of each genetic mutation and further research is under way to determine the causes and potential cures.

See also: Animal Models of Inherited Retinal Degenerations; Retina: An Overview; Retinal Development: An Overview; Retinal Glia; Retinal Models; Retinitis Pigmentosa; Visual Development.

Further Reading

Guillery RW, Mason CA, and Taylor JSH (1995) Developmental determinants at the mammalian optic chiasm. *Journal of Neuroscience* 15: 4727–4737.

Herrera E, Brown L, Aruga J, et al. (2003) Zic2 patterns binocular vision by specifying the uncrossed retinal projection. *Cell* 114: 545–557.

Hoskins SG and Grobstein P (1984) Induction of the ipsilateral retinothalamic projection in *Xenopus laevis* by thyroxine. *Nature* 307: 730–733.

Hoyt WF and Lewis O (1963) The primate chiasm. *Archives in Ophthalmology* 69: 69–85.

Hughes A (1979) A schematic eye for the rat. *Vision Research* 19: 569–588.

Jeffery G and Erskine L (2005) Variations in the architecture and development of the vertebrate optic chiasm. *Progress in Retinal and Eye Research* 24: 721–753.

Mann F, Harris WA, and Holt CE (2004) New views on retinal axon development: a navigation guide. *International Journal of Developmental Biology* 48: 957–964.

Oster SF and Sretavan DW (2003) Connecting the eye to the brain: The molecular basis of ganglion cell axon guidance. *British Journal of Ophthalmology* 87: 639–645.

Rachel RA, Dolen G, Hayes NL, et al. (2002) Spatiotemporal features of early neuronogenesis differ in wild-type and albino mouse retina. *Journal of Neuroscience* 22(11): 4249–4263.

Rasband K, Hardy M, and Chien CB (2003) Generating X: Formation of the optic chiasm. *Neuron* 39: 885–888.

Reese BE (1988) 'Hidden lamination' in the dorsal lateral geniculate nucleus: The functional organization of this thalamic region in the rat. *Brain Research Reviews* 13: 119–137.

Reese BE (1993) Clinical implications of the fibre order in the optic pathway of primates. *Neurological Research* 15: 83–86.

Reese BE (1996) The chronotopic reordering of optic axons. *Perspectives on Developmental Neurobiology* 3: 233–242.

Reese BE and Ho K-Y (1988) Axon diameter distributions across the monkey's optic nerve. *Neuroscience* 27: 205–214.

Williams RW, Hogan D, and Garraghty PE (1994) Target recognition and visual maps in the thalamus of achiasmatic dogs. *Nature* 367: 637–639.

Williams SE, Mason CA, and Herrera E (2004) The optic chiasm as a midline choice point. *Current Opinion in Neurobiology* 14: 51–60.

Optic Tectum: Development and Plasticity

S C Sharma and K A Patil, New York Medical College, Valhalla, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term 'optic tectum' is generally used in lower vertebrates to describe the homolog of the mammalian superior colliculus. In fish, amphibians, reptiles, and birds, it is represented by a single pair of dorsally situated dilatations in the mesencephalon, which are their principal visual structure in the brain. The optic tectum contains a visual map that is activated by its principal input, retinal axons. It responds through descending efferent pathways to brain stem and spinal motor areas giving a movement reaction to visual stimuli. At progressive levels of vertebrate phylogenetical transformation, the optic tectum becomes more structurally compartmentalized, giving rise in eutherian mammals to both the inferior and the superior colliculi, the latter maintaining more of the visual functions ascribed to the original optic tectum.

Nevertheless, as the cerebral hemispheres begin to play a more important role in the processing of neural information, the optic tectum is reduced in size relative to the rest of the brain. During the evolutionary process of 'telencephalization,' the cerebral hemispheres are thought to take over several of the visual tasks performed by the primitive optic tectum, and greatly extend the analyses and associative discrimination of visual stimuli. Although simplified, the human superior colliculus still preserves a laminar organization, receiving in its most superficial layers input from the retina and visual cortex, and in its deeper layers input from the auditory and somatic sensory systems. These different types of sensory information are arranged in three separate spatial maps, parallel to the surface of the colliculus, and integrated to direct eyes and head movements toward an external stimulus in a coordinated response. Functions of the superior colliculus also include aspects related to attention and pattern discrimination, working in closer interdependence with the geniculo-cortical system. Thus, in correlation with its physiological importance in the nonmammalian brain, the primal optic tectum shows a higher degree of anatomic complexity than its equivalent, the superior colliculus in the primates. Because of its behavioral relevance, and its convenience in the study of laminar neuronal organization, the optic tecta of fish and amphibians have been popular models in the elucidation of (1) axonal growth and targeting during development and regeneration of the central nervous system; (2) neuronal plasticity following injury; and (3) the neuroanatomic basis for visual-related behavior.

Cytoarchitecture

The optic tectum in lower vertebrates is composed of two prominent structures, which are symmetrically situated along the midline of the brain, and between the olfactory lobes and the diencephalons anteriorly, and the cerebellum posteriorly. It displays a lamination pattern made by a series of alternating layers of neural fibers and cells. Some variations exist in the histological organization of the optic tectum in different species. Here we will exemplify using the optic tectum of teleost fish, although most tecta follow a similar general plan. A commonly accepted nomenclature divides the optic tectum, from surface to depth, into the following: stratum marginale, stratum opticum, stratum fibrosum et griseum superficiale (SFGS), stratum griseum centrale (further subdivided into the inner plexiform layer and the inner gray layer), stratum album centrale, and stratum periventriculare. Each of these layers extends parallel to the surface throughout most of the optic tectum. Information from different regions in the visual field is integrated in different locations along the optic tectum's homogeneous preservation of laminar organization (Figure 1).

The stratum marginale is characteristic of fish and contains three elements: (1) the dendritic trees of pyramidal neurons (their somas are situated in the SFGS), which make synaptic boutons; (2) the marginal fibers originating in the torus longitudinalis, running parallel and laterally among glial elements; and (3) the stratum marginale ascending dendrites, which originate in neurons in the stratum griseum centrale.

The stratum opticum (SO; Figure 1), directly underlying the stratum marginale (SM; Figure 1), contains myelinated axons constituting part of the retinal afferents (contralateral), and large axons of tectal origin. These fibers make synaptic contacts with dendritic arbors from neurons whose somata are located in the stratum griseum centrale, the stratum album centrale, and the stratum periventriculare. They also synapse with dendrites of horizontal neurons whose cell bodies are located within the stratum opticum itself. Horizontal neurons show a spheroidal soma and dendrites that extend horizontally. The next deeper lamina is the stratum fibrosum et griseum superficiale, which is conformed by abundant myelinated

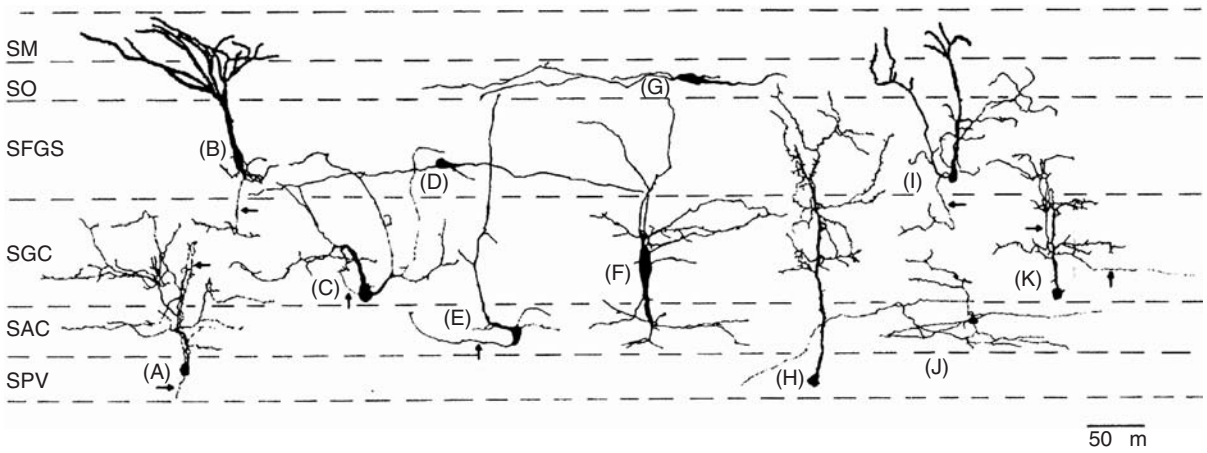

Figure 1 Composite diagram illustrating the principal neuronal types in different layers of the goldfish optic tectum. A, H, and K are pyriform neurons; B and I are pyramidal neurons; C, E, and J are nonoriented neurons; D and G are horizontal neurons; F is a fusiform neuron. SM, stratum marginale; SO, stratum opticum; SFGS, stratum fibrosum et griseum superficiale; SGC, stratum griseum centrale; SAC, stratum album centrale; SPV, stratum periventriculare.

axons and conspicuous cell bodies. Most of these axons are retinotectal afferents, whose terminal arbors synapse with dendrites extended by neurons located in deeper layers, and also with dendrites of pyramidal and bipolar neurons, located within this stratum. The pyramidal neurons are conspicuous in this layer. Horizontal neurons are present in this layer as well. The stratum griseum centrale (SGC; Figure 1) is divided into the outer plexiform layer and the inner gray layer. Most abundant in this lamina are the vertical dendritic arbors of neurons located deeper in the tectum (large piriform and other cells). Also present are the descending axons of pyramidal neurons. Several neuronal types have been found in this stratum: large fusiform neurons, extending dendrites into the SFGS, and an axon which leaves the tectum through the stratum album centrale; small pyriform neurons, giving rise to the stratum marginale ascending dendrites; small multipolar neurons, some of which extend axons that leave the tectum; and horizontal neurons with dendrites arborizing locally. The stratum album centrale (SAC; Figure 1) contains mostly bundles of myelinated axons, although unmyelinated fibers can also be recognized. The majority of these axons are of tectal origin, but some are originated in the contralateral retina, and the ipsilateral and contralateral telencephalon. Neurons here are the ganglionic and the large pyriform types, having large ascending dendritic arbors. Branches of the latter have been shown to be contacted by retinotectal terminals. Nonoriented neurons can be found here as well, representing a less defined type. The deepest laminae are the stratum periventriculare (SPV; Figure 1). It contains the 'periventricular neurons,' which extend dendritic processes into the overlying strata. However, the most characteristic cell type in this layer is the ependymoglial cell, which gives rise to an ascending process that reaches the uppermost aspect of the tectum while extending profuse branches along the way.

Afferents into the tectum are axonal terminals originating from (1) the retina, (2) the telencephalon, (3) the nucleus isthmi, (4) the dorsolateral and dorsomedial thalamic nuclei, (5) several pretectal nuclei, (6) the nucleus preglomerulosus, (7) the torus longitudinalis and the torus semicircularis, (8) the contralateral tectum and the dorsolateral tegmentum, and (9) diverse brain stem cell groups. Fibers originating in the contralateral eye course through the optic nerve, reaching the tectum in a retinotopic distribution. They terminate in the stratum opticum and the SGFS, and scarcely between the stratum album centrale and the stratum periventriculare. They synapse with horizontal neurons, pyramidal neurons, small bipolar neurons, fusiform neurons, and large fusiform neurons. In addition, optic tectum also receives a strong input from the auditory system and contains map of auditory space especially in owl and pigeons.

Tectal efferents originate from (1) the small bipolar neurons of the SFGS, (2) the fusiform neuron in the stratum griseum centrale, (3) the large multipolar neurons of the stratum griseum centrale and the stratum album centrale, and (4) large pyriform neurons of the stratum griseum centrale and the stratum album centrale. Tectal efferents form three efferent fiber groups: (1) the ascending group, running mostly toward the ipsilateral pretectum and thalamus, but some toward contralateral structures (including tecto-ocular fibers); (2) the medial group, coursing toward the contralateral tectum (forming the tectal

commissure, and terminating mostly in the stratum griseum centrale), but a few reaching the torus longitudinalis; and (3) the descending group, which extends into the ipsilateral and contralateral tegmentum and lower brain stem, including the nucleus isthmus (the nonmammalian homolog of the parabigeminal nucleus).

The optic tectum displays visually elicited electrical activity. Discrete areas of the retina receiving light from punctiform stimuli in the visual field project to discrete regions of the contralateral tectum (although species having binocularity project also ipsilaterally, mainly through an isthmotectal relay).The visual projections are arranged in a topographic manner, such that: nasal retina is represented in the caudal region of the tectum, and temporal retina in the rostral region; the dorsal retina projects to the lateral region of the tectum, and the ventral to the medial. This retinotopic organization extends across all laminae of the optic tectum. The topographic organization of axonal terminations is conserved also in the isthmotectal and tectoisthmal projections. There is orderly point to point correspondence between the retina and the tectum. The nasotemporal axis is represented along the caudorostral axis and the dorsoventral retinal axis is represented on the lateromedial axis of the tectum (Figure 2).

The mechanism of retinotectal map formation has been studied in mature fish as their optic nerves have a capability of regeneration. Seminal work was done by Roger Sperry who proposed that during development cells of the retina and tectum acquire cytochemical labels denoting their position within the retina or tectum. The growing retinal axons carrying the unique positional specificity on them are guided toward the tectal cells having corresponding labels; thereby, a selective and appropriate contact is made. By creating retinotectal mismatch or size disparity, Attardi and Sperry showed the regenerating optic axons reached their appropriate terminal sites by passing the inappropriate sites (Figure 3) Further research led to many aspects of neuronal specificity providing evidences of neuronal plasticity in which retinal fibers either 'compress' or 'expand' their projection depending on the available sites. In all such situations one element that remained constant was

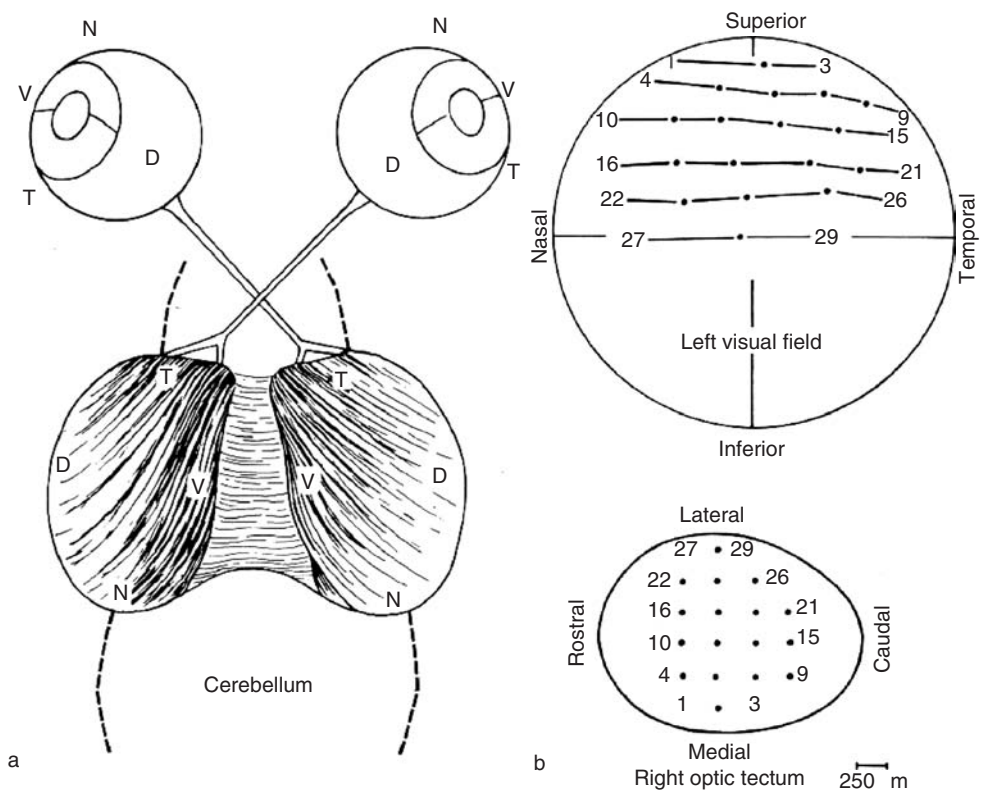

Figure 2 (a) Schematic representation of the projection of the retina onto the optic tectum in teleosts, showing the complete crossing of the optic nerves, and the division of the optic tracts into medial and lateral bundles before reaching the tectum. The optic nerve fibers spread in a caudolateral direction across the surface of the tectum from their entrance near rostromedial pole. Nasal retinal fibers (N) terminate in caudal tectum, temporal fibers (T) in rostral tectum, dorasal fibers (D) in lateral and ventral tectum, and ventral fibers (V) in medial tectum. (b) Visuotopic map from a normal left eye to the right optic tectum. Each number in the visual field shows the position of a visual stimulus that gave a localized response at the electrode positions indicated by the same number on the tectum.

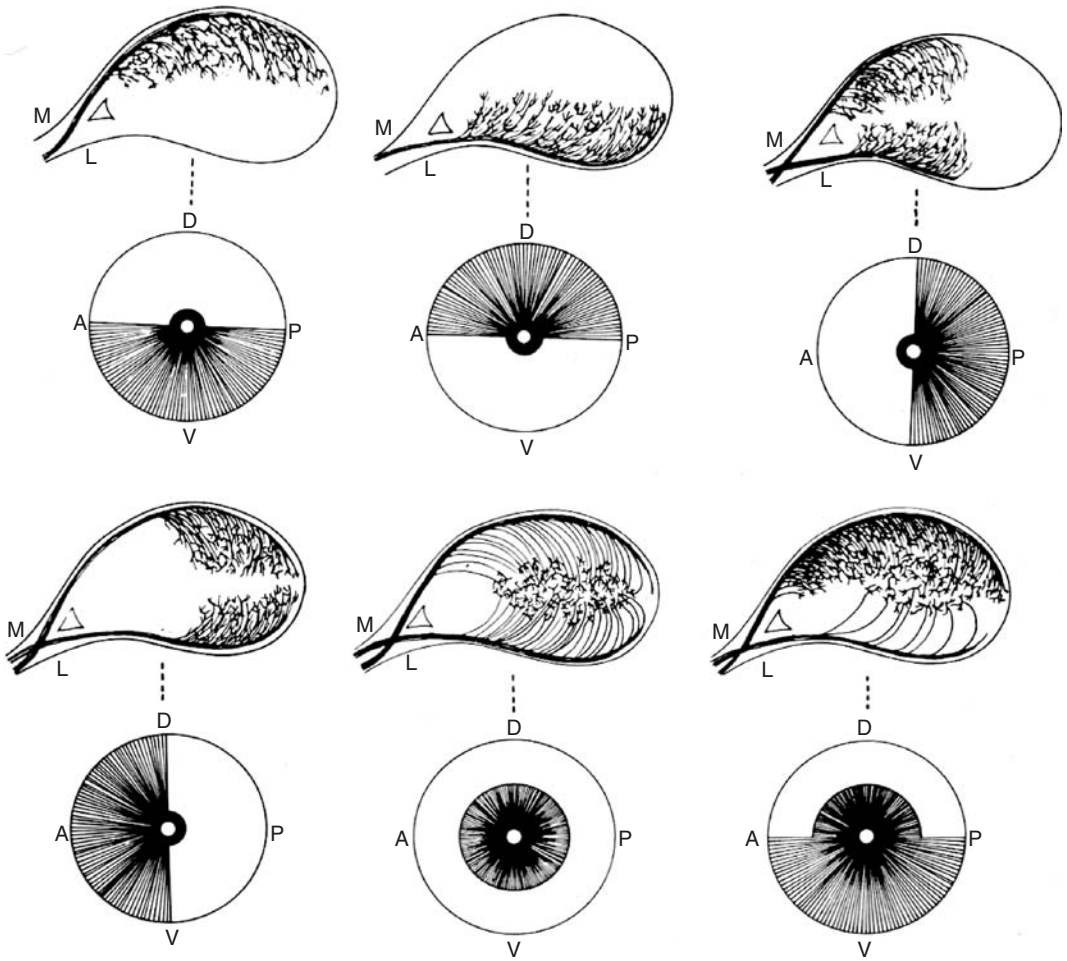

Figure 3 Schematic drawings of the regenerated optic fiber patterns observed in the tectum using Bodian protagrol method after optic nerve section and various partial retinal ablations. In each retina diagram, the white area denotes the ablated region. On the retinal diagrams: A, anterior; P, posterior; D, dorsal; V, ventral. On the tectum diagrams: M and L, medial and lateral branches of the optic tract. Redrawn with permission from Dr. RW Sperry.

the order of topography. These and other relevant aspects have been summarized by Sharma and Fawcett.

Retinotectal terminal arborizations show electrophysiological segregation among different tectal layers. While no light-evoked responses have been detected in the stratum marginale, unitary responses are predominantly 'on–off' and transient in the stratum opticum and the superficial-SFGS. Units recorded in the deep-SFGS give sustained responses and are either on-center or off-center type. Electrically active tectal dendritic elements with their highly organized spatial arrangement are an important component of the visual processing apparatus, rather than just relays of action potentials. Similar to the mammalian visual cortex, neuronal elements in the optic tectum display a high degree of specialization for different components in which visual information is separated. Studies of tectal field potentials show a directional selectivity to moving spots of light. Tectal neurons also show velocity sensitivities, since their firing rate depends on the stimulus velocity. There are also specific tectal neurons in fish, which respond only to certain orientations of polarized light, as received by specialized photoreceptors.

There are several neuronal types in the optic tectum, which respond directly to photic stimuli because of direct synapses with retinal ganglion cell axons. Pyramidal cells located in the SFGS, which send dendrites within the SFGS, to the stratum opticum, and to the stratum marginale, have a short latency of response. These cells have strong spontaneous activity, display slow adapting responses with movement directionality, and are perhaps involved in the orientation circuit. Pyriform neurons located in the stratum periventriculare, which send dendritic trees in the upper tectal strata, respond in a delayed form to retinal input, and they are thought to receive direct connections from retinal ganglion cells, via slow conducting fibers reaching the deep tectum. These cells

have a wide range of electrical activity, although smaller pyriform neurons tend to have stronger spontaneous activity and photic responses. Horizontal neurons represent a heterogenous population of tectal neurons arranged in several strata. Their electrophysiological responses have a wide range of variation. Many of them display strong spontaneous activity and short latencies to retinal stimulation, demonstrating their direct connections with retinal afferents.

Multiple types of molecules have been identified in recent years as playing a role in interneuronal communication in the optic tectum of nonmammalian species. Their variety serves as an indication of the complexities of tectal circuitry. Neurotransmitters and neuromodulators that have been shown in tectal neurons include: glutamate, γ-aminobutyric acid (GABA), acetylcholine, serotonin, somatostatin, substance P, neurotensin, neuropeptide Y, gonadotropin-releasing hormones, corticotrophin-releasing factor, and arginine vasotocin. Glutamate-positive retinal fibers terminate in the stratum opticum and the SFGS. Glutamate receptors of different types are present on tectal neuron dendrites. Glutamate is not only necessary for visual information relay, but in addition, *N*-methyl-D-aspartate (NMDA) receptor has been shown to be important in the electrical-activity-driven refinement of the retinotectal terminal arborizations. This process occurs during development and regeneration, and leads to sharpening of the retinotopic map. Glutamate mediates transmission into the tectum by marginal fibers projecting from the torus longitudinalis, and has been found in optic tectum neurons as well, including pyramidal, multipolar, fusiform, large pyriform, and periventricular. GABA, present in neurons and fibers scattered across the frog optic tectum, enhances the excitatory synaptic transmission of optic nerve terminals. Acetylcholine has an important role in the isthmotectal projections and in the intrinsic circuitry of the optic tectum. Cholinergic input modulates the transmission of these fibers via presynaptic nicotinic receptors. Two molecules of interest in the sexual dimorphism that the lower vertebrate brain displays are arginine vasotocin and galanin. They are present in higher levels in the optic tectum (among several brain regions) of male fish, frogs, and newts than in that of females, or sexually unresponsive males.

Development of Tectum and Axonal Guidance

The optic tectum develops simultaneously with the eye. It differentiates from the alar plate of the mesencephalon. Four phases can be considered: cell proliferation; migration; fiber growth and synaptogenesis; and a phase of regressive changes including cell death and neurite pruning. The development of the topographic connection of the retina with the tectum has been a focus of research for many years. In the prospective mesencephalon, the local anterior–posterior polarity is generated through the activity of a long-range signal from the posterior boundary with the rhombencephalon, the isthmus. Many of the successive genetic interactions that determine tectal evolution in the midbrain have been identified at the molecular level.

The development of the ectodermal midbrain/hindbrain organizer is induced by mesoderm. The spatially correct expression of the homeobox genes, Otx-2 (anterior) and Gbx-2 (posterior), is important for the development of the isthmic-signaling region. Otx-2 is expressed in the anterior neuronal tube down to the mid-hindbrain boundary. Gbx-2 is expressed in the rhombencephalon and borders with the expression of Otx-2. Then a diffusible substance, such as fibroblast growth factor 8 (FGF-8), emanating from the organizer determines midbrain polarity. Several genes (including En-1, Otx-2, Pax-2, and Wnt-1) regulate FGF-8 itself transcriptionally. Pax proteins are transcription factors that contain paired box and homeobox DNA-binding domains. Pax-2 covers the whole of the mesencephalon, whereas Pax-5 is expressed in high levels in the isthmic region. Positive inductions by FGF-8, in conjunction with inhibitory influences from the mesencephalic–diencephalic junction, produce a decreasing caudal–rostral gradient of engrailed (En-1 and En-2) expression along the posterior–anterior axis of the midbrain. Eventually, the engrailed distribution determines the positional specificity of tectal neurons. Ephrins and the Eph receptors are among the cell-surface molecules directly responsible for neural guidance of afferent retinal fibers toward specific tectal positions. Ephrin-A5 and Ephrin-A2 are the two molecules shown to be engrailed dependent *in vivo*. They distribute in gradients, increasing from anterior to the posterior poles, with Ephrin-A5 restricted to the caudal half of the tectum. Eph receptors are expressed in gradients across the early retina and on the growth cones of ganglion cells. Eph-A3 is the preferred receptor for Ephrin-A5 and Ephrin-A2. It is present in a retinal gradient, increasing from nasal to temporal. The tectal distribution seems to be a repellent guidance system that stops growing retinal axons at different anterior–posterior positions. Grg-4 is expressed early in the mesencephalon and causes repression of molecules, including FGF-8, Pax5, En-2, and Ephrin-A2. Thus, Grg-4 might contribute to setting the anterior limit of the tectum and its rostrocaudal polarity formation (Figure 3).

The dorsoventral patterning of the mesencephalon is determined by the Shh molecule, which is expressed in the floor plate of the neural tube up to the diencephalon. In the tectum, if Shh is expressed ectopically in the dorsal region, the dorsal tectum receives dorsal retinal fibers, which normally terminate at the ventral tectum. Furthermore, Shh represses the expression of the tectum-related genes (En-1, En-2, Pax-2, Pax-5, and FGF-8) and induces ventral markers, such as HNF3b and ptc. Ephrin-B ligands and their receptors appear to be the guidance cues for spatial segregation along the dorsal–ventral axis. Wnt3 is expressed in a medial–lateral decreasing gradient in chick optic tectum and mouse superior colliculus (Schmitt et al.). Retinal ganglion cell (RGC) axons from different dorsal–ventral positions showed graded and biphasic response to Wnt3 in a concentration-dependent manner. Wnt3 repulsion is mediated by Ryk, expressed in a ventral-to-dorsal decreasing gradient, whereas attraction of dorsal axons at lower Wnt3 concentrations is mediated by Frizzled(s). A classical morphogen, Wnt3, acting as an axon guidance molecule, plays a role in retinotectal mapping along the medial–lateral axis, counterbalancing the medial-directed EphrinB1-EphB activity.

Several molecules known to affect neurite behavior are concentrated in the retinorecipient laminae making them plausible candidates to participate in the regulation of laminar specificity. These include the adhesion molecule SC1/DM-GRASP/BEN, semaphoring/collapsing receptors and ephrin-B2; in addition, N-cadherin is concentrated in the synaptic cleft at retinotectal synapses. N-cadherin and glycoconjugates recognized by the plant lectin Vicia villosa agglutinin-B4 (VVA) guides retinal axons. Versican is an important layer-specific cue for presynaptic maturation of retinal ganglion cells.

Reelin (RELN) is an extracellular matrix protein largely related with laminar organization in several brain areas. In the optic tectum (OT) of trout, as in amniotes, RELN immunoreactivity increases within specific cell layers as lamination proceeds, and decreases when it is complete, except in the stratum opticum (SO), where RELN-immunoreactive cells are observed throughout life. Time-course expression of RELN in the OT suggests a role in the early modeling of synaptic contacts and the accommodation of new retinal arriving axons throughout life.

In summary, the spatial identity of the neurons within the optic tectum seems to be tightly regulated by the graded induction of a series of asymmetric genes, transcriptional factors, and receptors for signal transduction during early development. The exact sequences of different gene activation leading to the formation of adult tectum are not yet clear.

Behavioral Correlates

The optic tectum is involved in (1) perception of form (related to food catching and escape behavior); (2) responses that can be conditioned to visual stimuli; (3) ocular convergence movements; and (4) 'food searching' forward body movements. However, as evidenced in fish, there are visual responses in which the tectum is not directly involved. The optokinetic and dorsal light reflex appear to be mediated by nontectal targets (e.g., diencephalic and pretectal groups), since they persist after tectal ablation. The optic tectum might also be involved in visual memory, as long-term potentiation has been elicited in it. Nevertheless, visuo-spatial conditioning memory involves other brain structures and persists even after abolishing the tectum.

In both fish and frogs, electrical microstimulation of the optic tectum can elicit complex stereotypic body movements. It has been demonstrated that tectoreticular circuit helps in the generation of saccadic eye movements in goldfish. The optic tectum of goldfish, as in other vertebrates, plays a major role in the generation of orienting movements, including eye saccades. To perform these movements, the optic tectum sends a motor command through the mesencephalic and rhombencephalic reticular formation, to the extraocular motorneurons. These motor responses have been studied in the two primary behaviors of vertebrates, feeding and avoidance. Distance of the prey from the eye determines whether frogs will merely fixate on it, or snap at the stimulus. This difference seems to be related to the size of the retinal area stimulated by moving objects. As would be expected, tectotectal connections are involved in behavior involving conjugated eye movements, such as turning toward prey, as shown in studies of the effects of severance of these fibers. In contrast, hemisection of the tectal descending fibers eliminates the frog's ability to snap forward, while preserving that of turning. On the opposite end of the behavioral spectrum, the optic tectum of nonmammalian vertebrates is involved in the avoidance to possibly threatening objects. Both frogs and fish react by jumping or flipping away when they visually spot large dark objects approaching. This behavior is completely eliminated by tectal ablation. Smaller lesions to areas of the tectum abolish the above-mentioned responses to stimuli within specific corresponding regions of their visual field. Stimulation of the rostral tectum induces turns of the head and body toward the frontal field; activation of caudal tectum elicits body orientation toward the rear field.

The primitive optic tectum, like the superior colliculus of primates, integrates other sensory modalities

in addition to vision. In weakly electric fish, electrosensory stimuli (used in the detection of objects in their surroundings) are processed in the tectum. This type of input is directed to areas in the optic tectum in such a way that it overlaps with visual information coming from the corresponding space region. Processed information closes the stimulus-response circuit as the tectum projects to the diencephalic nucleus, which has efferents to the mesencephalic prepacemaker, connecting in turn with the medullary pacemaker nucleus, which finally controls the electric organ discharge (used in interactions with other fish, such as in courtship). In snakes, another sensorial type of information is important. Temperature changes in its surroundings are detected by the heat-focusing pit organ, which connects with the optic tectum to process these stimuli.

In spite of the enormous progress made during the past five decades, our knowledge of the functioning of the optic tectum leaves further opportunities for investigation. Among others, studies are needed to elucidate in further detail the electrochemical features of the tectal circuitries, and to advance our understanding of the integration mechanisms in which the optic tectum accomplishes simultaneous processing of different modalities of sensory information, in order to provide coherent responses to the environment.

See also: Optic Nerve, Optic Chiasm and Optic Tracts; Photoreceptor Adaptation; Retinal Pharmacology: Inner Retinal Layers; Retinal Development: An Overview.

Further Reading

Angeles LM, Pilar Perez-Perez M, Herrero L, and Torres B (2005) Involvement of the optic tectum and mesencephalic reticular formation in the generation of saccadic eye movements in goldfish. *Brain Research. Brain Research Reviews* 49(2): 388–397.

Attardi DG and Sperry RW (1963) Preferential selection of central pathways by regenerating optic fibers. *Experimental Neurology* 7: 46.

Candal EM, Caruncho HJ, Sueiro C, et al. (2005) Reelin expression in the retina and optic tectum of developing common brown trout. *Brain Research. Developmental Brain Research* 154(2): 187–197.

Ewert JP, Capranica RR, and Ingle DJ (eds.) (1983) *Advances in Vertebrate Neuroethology.* New York: Plenum Press.

Guthrie DM and Sharma SC (1991) Visual responses of morphologically identified tectal cells in the goldfish. *Vision Research* 31(3): 507–524.

Heiligenberg W (1990) Electrosensory systems in fish. *Synapse* 6(2): 196–206.

Katahira T, Sato T, Sugiyam S, et al. (2000) Interaction between Otx2 and Gbx2 defines the organizing center for the optic tectum. *Mechanisms of Development* 91(1–2): 43–52.

Matsunaga E, Araki I, and Nakamura H (2000) Pax6 defines the di-mesencephalic boundary by repressing En1 and Pax2. *Development* 127(11): 2357–2365.

Nakamura H (2001) Regionalization of the optic tectum: Combinations of gene expression that define the tectum. *Trends in Neuroscience* 24(1): 32–39.

Nemec P, Altmann J, Marhold S, et al. (2001) Neuroanatomy of magnetoreception: The superior colliculus involved in magnetic orientation in a mammal. *Science* 294(5541): 366–368.

Nomura T and Fujisawa H (2000) Alteration of the retinotectal projection map by the graft of mesencephalic floor plate or sonic hedgehog. *Development* 127(9): 1899–1910.

Romeskie M and Sharma SC (1979) The goldfish optic tectum: A Golgi study. *Neuroscience* 4(5): 625–642.

Schmitt AM, Shi J, Wolf AM, and Lu CC (2006) Wnt-Ryk signalling mediates medial–lateral retinotectal topographic mapping. *Nature* 439(7072): 31–37.

Sharma SC and Fawcett JW (eds.) (1993) *Formation and Regeneration of Nerve Connection.* Boston, MA: Birkhauser.

Vanegas H (ed.) (1984) *Comparative Neurology of the Optic Tectum.* New York: Plenum Press.

Wang SR (2003) The nucleus isthmi and dual modulation of the receptive field of tectal neurons in non-mammals. *Brain Research. Brain Research Reviews* 41(1): 13–25.

Yamagata M and Sanes JR (1995) Target-independent diversification and target-specific projection of chemically defined retinal ganglion cell subsets. *Development* 121(11): 3763–3776.

Yamagata M and Sanes JR (1999) Formation of lamina-specific synaptic connections. *Current Opinion in Neurobiology* 9(1): 79–87.

Yamagata M and Sanes JR (2005) Versican in the developing brain: Lamina-specific expression in interneuronal subsets and role in presynaptic maturation. *Journal of Neuroscience* 25(37): 8457–8467.

Optic Tectum: Sensorimotor Integration

H Luksch, Technische Universität München, Freising-Weihenstephan, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The vertebrate midbrain consists of three major regions: the caudal isthmus, the ventral tegmentum, and the dorsal tectum (Latin for roof). The tectum can again be differentiated into an auditory center, the torus semicircularis, and a visually dominated center, the optic tectum. This nomenclature is consistent throughout the vertebrates, with the exceptions of birds, in which the auditory center is termed nucleus mesencephalicus lateralis pars dorsalis (Mld), and mammals, in which the optic tectum corresponds roughly to the outer layers of the superior colliculus and the auditory midbrain is called inferior colliculus.

The optic tectum is characterized by several features. First, it is a multisensory center, with a primary visual input and additional input from sense organs that yield a topographic representation of the external world (including auditory, somatosensory, and lateral line, infrared, and electrosensory where present). Second, the sensory input is always organized in the form of maps that are in register with one another, thus forming a coherent spatial representation. And third, the optic tectum is a highly laminated structure, with alternating layers of cell bodies and neuropils, characteristic cell types in the different layers, and a strict separation of input and output structures. Together, these features have rendered the optic tectum one of the best-studied model systems for developmental map formation, the alignment of multimodal maps, plasticity after injury and sensory misalignment, and the cellular processing of sensory input.

The functions of the optic tectum can generally be summarized as being related to spatial orientation. The tectum integrates all available sensory modalities and generates an unambiguous premotor signal which is then relayed to premotor and motor nuclei via descending efferent pathways. This premotor signal aims at aligning the sensory organs to a specific location in space and has led to the notion of the tectum being a visual (or multimodal) reflex center. In the lower vertebrates (i.e., nonmammalian species), the size of the optic tectum suggests that it is the major sensory processing center, a finding that is in line with the idea of mostly reflexive behavior in these animals (Figure 1). In contrast, the relatively small size of the mammalian superior colliculus is often taken as an indication that the telencephalon has taken over most of the tectal functions, thus rendering the superior colliculus a subordinate region under complete control of forebrain efferents.

However, this classic view of the optic tectum can no longer be upheld. The size of the optic tectum is quite variable within a vertebrate class and appears to be dependent on other factors, such as the binocularity of a given species. Also, the optic tectum gives rise to many different projection systems, the largest being the tectofugal ascending pathway that relays information toward the forebrain. Many tectal projections are involved in attentional and integrative networks, thus participating in the general cognitive control of behavior. In addition, a remarkable conservation of cell types and their projections has been found in comparative studies of the vertebrate optic tectum. And finally, physiological as well as lesion studies have shown that throughout the vertebrates the role of the optic tectum is far from being a simple reflex circuit for sensory orientation. Due to the extensive interconnections with many areas throughout the entire brain, the optic tectum is involved in many different behaviors, and likewise exerts its functions only in concert with all other visual and sensory processing areas. This unity of the visual system has to be taken into account when attempting to isolate and analyze functions of any given area in the vertebrate brain.

Cytoarchitecture and General Connectivity

In most vertebrates, the optic tecta can easily be identified as two symmetrical prominent lobes. When sectioned in the transversal plane, these structures display a pronounced pattern of lamination which may be absent only in secondarily simplified brains (e.g., salamanders). These laminae extend parallel to the curved surface of the tectum and may change in their relative thickness, depending on the part of the visual field they are representing. Even though often identical anatomical terms are used, the nomenclature of the alternating cell and fiber layers varies among vertebrate classes and cannot easily be homologized (Figure 2). In general, the optic tecta of fish and amphibia, on the one hand, and reptiles and birds, on the other hand, can be distinguished by the greater differentiation of the laminae in the most superficial layers that are collectively termed the stratum griseum et fibrosum superficiale (SGFS).

For simplicity, the following cytoarchitectonic description focuses on the avian optic tectum (Figure 2(d)). In birds, two nomenclatures for the tectal laminae have been developed. Ramón y Cajal differentiated 15 laminae, starting from the superficial layer. Later, a different

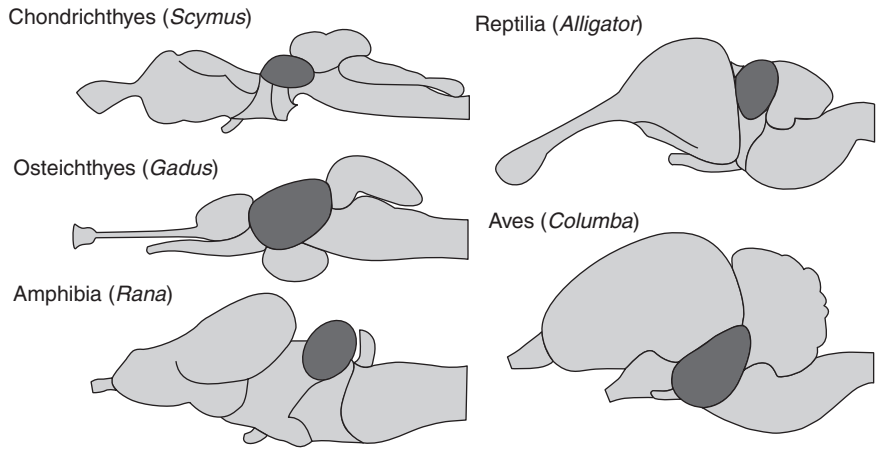

Figure 1 Position and size of the optic tectum (dark gray area) in vertebrates.

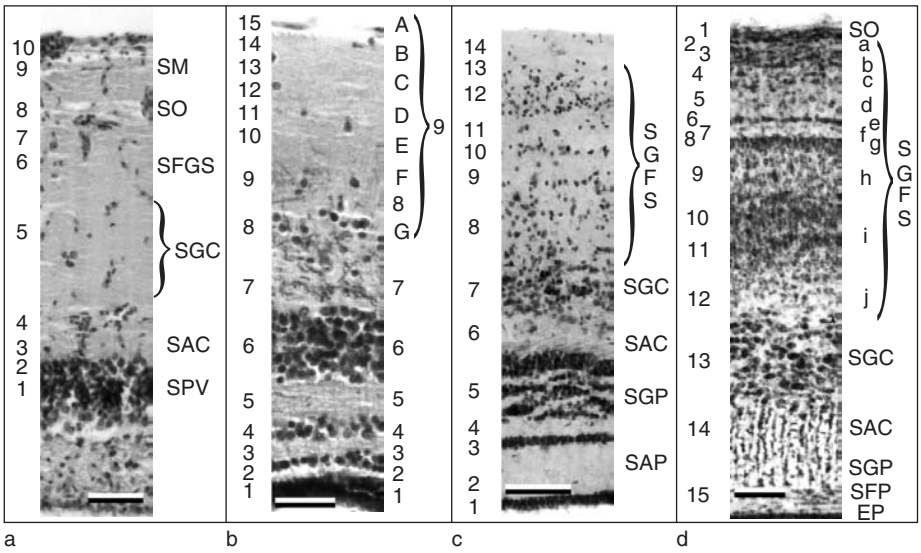

Figure 2 Cytoarchitecture of the optic tectum in various vertebrate classes in histological sections: (a) fish, *Carassius auratus*; (b) frog, *Discoglossus pictus*; (c) reptile, *Gecko gecko*; (d) bird, *Gallus gallus*. The layer numbers to the left refer to the original nomenclature suggested by Ramón y Cajal and Pedro Ramón, the acronyms to the right were introduced by various authors. Both nomenclatures can be found in the current literature. EP, ependym; SAC, stratum album centrale; SAP, stratum album periventriculare; SFGS, stratum fibrosum et griseum superficiale; SFP, stratum fibrosum profundum; SGC, stratum griseum centrale; SGFS, stratum griseum et fibrosum superficiale; SGP, stratum griseum profundum; SM, stratum marginale; SO, stratum opticum; SPV, stratum periventriculare. Scale bar = 50 μm (a, b) and 100 μm (c, d). Panel (a) provided by DTT Plachta; panel (c) provided by Y Tang and CE Carr.

nomenclature was introduced that divided the tectum into six layers, five of which are located in the deep tectum and consist of only one cellular or plexiform lamina whereas the intermediate and superficial laminae are subsumed in the SGFS (Figure 2(d)). In the current literature, both nomenclatures are used and often mixed, and some authors subdivide the layers even further. In addition, for developmental studies, a third nomenclature has been established to accurately describe the complex sequence of layer formation.

The different laminae of the optic tectum contain specific cell types. Although earlier studies assumed that each lamina has a characteristic cell morphology and a specific connectivity, it was later found that layers may include cells with comparable morphology but varying connectivity. As an example, the projection neurons of the stratum griseum centrale (SGC) have large somata, large dendritic trees, and distal dendritic specializations, but they receive input from different retinal ganglion cells and project to various target areas. Thus, a complete description of the cell types in the avian optic tectum is not feasible. This article, rather, gives an overview of tectal connectivity pertaining to the laminar organization

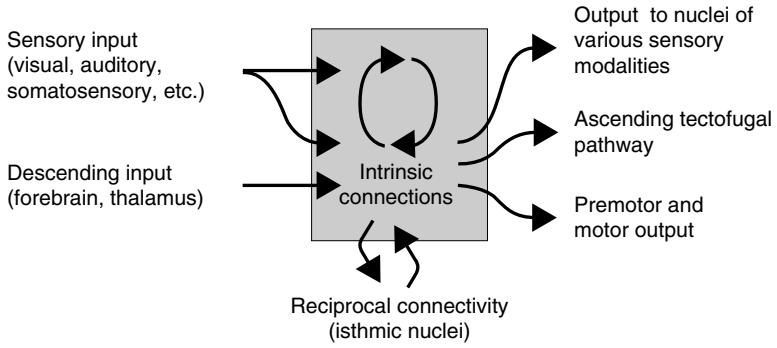

Figure 3 Schematic of the connectivity of the optic tectum.

and gives a detailed description only of some well-characterized neurons.

The major sensory input to the tectum originates in the retina. Generally, a large proportion of the retinal ganglion cells project contralaterally into the optic tectum, even in those animals that have a well-developed thalamofugal pathway (the pathway from the retina via the dorsolateral thalamus to the forebrain). Throughout the vertebrates, the degree of ipsilateral projection varies and seems to correlate at least partially with the degree of binocularity in the species. In birds, however, the projection is exclusively contralateral, even in highly binocular species such as owls and even though a transient ipsilateral projection is found during embryogenesis. The topographical projection of the retina produces a map of visual space. Its development during embryogenesis has been studied in great detail as a model system of molecular developmental biology.

In all nonmammalian vertebrates, the optic fibers course over the surface of the tectum, turn sharply inward at their appropriate position, and terminate in the upper layers (2–5 and 7). The retinal afferents from various retinal ganglion cell classes form small terminal fields in individual layers, as determined by immunocytochemical and tracing studies. Additional sensory input is fed into the intermediate and deep layers. The nonsensory input to the tectum is generally not restricted to a single lamina, but can span most of the tectal layers, often in a columnar fashion; however, data on termination zones of specific projections are scarce. It is noteworthy that even in the retinorecipient upper layers, synapses of nonretinal input outnumber the retinal input, emphasizing the point that the optic tectum is not only a visual structure.

The output of the optic tectum is derived from the intermediate and deep layers. The major output is constituted by layer 13, where the ascending tectofugal pathway to the thalamic nucleus rotundus and descending pathways to premotor and motor targets originate. The intermediate layers constitute many projections to other nuclei of the sensory pathways, as well as a prominent reciprocal connection with the isthmic nuclei. In addition to these projections, both Golgi and intracellular studies suggest that many tectal cells have termination zones within the optic tectum, thus constituting a multitude of intrinsic connections which boost the computational capacity of the system. A simplified scheme of tectal connectivity is depicted in Figure 3.

Although it is not the focus of this article, a comment on the comparison between the nonmammalian tectum and the mammalian superior colliculus might be appropriate here. Even though the superior colliculus can also be differentiated into laminae, a direct comparison of laminar cell types and functions is not feasible. As a first approximation, the entire optic tectum of nonmammalian species can be considered homologous to the superficial layers of mammals (including the stratum zonale, stratum griseum superficiale, and stratum opticum), whereas the intermediate and deep layers of the superior colliculus may correlate to subtectal and tegmental areas in nonmammals. However, this rule of thumb breaks down for many aspects; for example, the motor output pathways originate from the deep layers in mammals and from the optic tectum proper in nonmammals. In general, homologies and comparisons between the mammalian and the nonmammalian visual midbrain should be made with great caution to avoid confusion, especially when it comes to functional considerations.

Identified Cell Types

A variety of cell types have been analyzed in such detail that they can be considered an identified cell type, indicating that their morphology, input, projection, and biochemical features are sufficiently documented. This kind of work is fundamental for a deeper understanding of tectal function, a challenge often hampered by the fact that anatomical studies rarely take physiological data into account and that physiological studies

often disregard efforts to ascribe physiological findings to specific cellular circuitry. Therefore, the wealth of data on the optic tectum obtained in anatomical, physiological, and also molecular studies is often difficult to reconcile; however, for some tectal cell types, the integration of the data is possible. Three types of neurons are briefly described here to demonstrate the level of analysis obtainable in this system.

Horizontal Cells of Layer 5

Even in the earliest Golgi studies of the avian optic tectum, horizontal cells were described in various superficial layers. A detailed intracellular study of layer 5 horizontal cells found multipolar cells with rather thick dendrites and arborizations confined to the layer where the soma was located (Figure 4(a)). Dendrites could extend more than 1000 μm in each direction. Because horizontal cells in layer 5 were found to have no axon, their dendrites act as both post- and presynaptic sites. Electron microscopy has shown that these cells receive input from retinal afferents and make connections with dendrites of tectal projection neurons. These local microcircuits are often surrounded by a glial sheet. Intracellular recordings have demonstrated an unusual cell physiology, with an oscillatory response to depolarization. The horizontal cells are immunopositive for calbindin and are likely to form a GABAergic output. On the basis of these features, models have been put forward to explain the contribution of horizontal cells to tectal information processing, an effort also motivated by the finding that horizontal cells with comparable characteristics are also found in the mammalian superior colliculus.

Isthmo-Optic Projection Neurons

The isthmo-optic nucleus (ION) receives a direct input from the optic tectum and constitutes an output to the retina that influences retinal ganglion cell firing rate, thus forming a short feedback circuit to the retina. The tectal neurons that project to the ION have never been observed in Golgi preparations, but have been described in detail with tracing techniques (Figure 4(b)). The somata of these neurons are located in tectal layer 9, close to layer 10, and have a thick axon that projects to the ION via the tecto-isthmal tract. The dendritic arborization has a columnar appearance, with apical dendrites restricted to layer 8 and only occasionally reaching layer 7, and basal dendrites reaching into layers 9–12 where the highest density of dendritic arborization is observed. Because retinal input does not reach deeper than layer 7, direct retinal input is probably not the major input to these cells. The transmitters of the ION-projecting neurons are glutamate and nitric oxide (NO), and the cells have been found positive for the synthesizing enzyme of NO. However, even though the circuit from the retina via the optic tectum and the ION back to the retina is anatomically rather simple, the function of this feedback has not been clearly established.

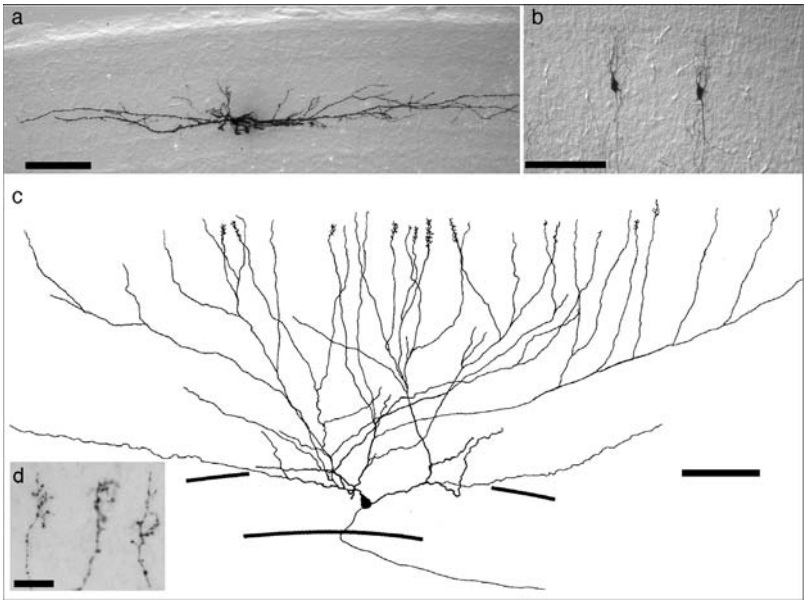

Figure 4 Identified cell types of the avian optic tectum: (a) horizontal cells in layer 5 of the chick, intracellular fill; (b) neurons of layer 9 that project to the ION in the chick after tracing; (c) SGC neuron in the pigeon, reconstruction of an intracellular fill; (d) detail of dendritic endings. Scale bar = 100 μm (a, b); 200 μm (c); and 20 μm (d).

Projection Neurons of Layer 13

Due to the different nomenclatures, the neurons of layer 13 are often called SGC neurons. These cells constitute the major output of the optic tectum and project bilaterally to the thalamic nucleus rotundus and, in addition, to various nuclei of the tectothalamic tract on the ipsilateral side (nucleus subpretectalis, nucleus interstitio-pretecto-subpretectalis, and nucleus posteroventralis thalami). These projections have been shown to be lateralized (i.e., specific projections differ in strength depending on the side of the brain), with the laterality being introduced by the differential illumination of the eyes during embryonic development in the egg. The output of the SGC is glutamatergic via α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors.

The morphology of SGC neurons is unique in the tectum. All SGC cells are multipolar and have several primary dendrites that course obliquely through the tectum toward the superficial layers, thus constituting very large dendritic fields with spans of several millimeters (covering up to 40° of visual angle in the retinotopic map). At their distal tips, the dendrites show specializations called bottlebrush endings, which consist of many small filiform arborizations (Figure 4(c)). Intracellular work has shown that each SGC neuron has a multitude of these endings, which are located exclusively in a specific tectal layer but are randomly distributed in the lateral domain. On the basis of the position of the bottlebrush endings and the position of the soma, several types of SGC neurons have been differentiated so far that also project to different subdivisions of the nucleus rotundus. In these thalamic subdivisions, various aspects of visual stimuli are represented (e.g., luminance, colour, and motion). Taken together, the tectal SGC neurons form computational channels, sampling information from specific retinal input layers (and thus specific retinal ganglion cell populations) and feeding this information into the functional subdivisions of the thalamic target. In this process, the simple retinotopic organization of the optic tectum is discarded in favor of a complex interdigitating topographic arrangement, in which each point in the nucleus rotundus receives input from a population of cells sparsely distributed throughout the entire tectum. SGC cell types also possess specific physiological characteristics that allow a differentiation of the cells even with extracellular recording techniques. Most remarkable in respect to cellular physiology is the phasic signal transfer from retinal afferent synapses on to the dendritic endings of SGC neurons. Together with the spatial arrangement of the dendritic endings in the visual map of space, simulations have shown that a complex response characteristic is generated, which, for one specific SGC-subtype, results in a differential response to stationary versus moving objects (including second-order motion stimuli).

In addition to these anatomical and physiological characteristics, SGC cell types express specific cell adhesion molecules during development that mediate pathfinding of the axonal projection and, through combinatorial expression in target nuclei, lead to the correct wiring of the output channels. The expression strength of a specific adhesion molecule can be experimentally manipulated, opening the door to manipulations of specific tectal projection systems and thus to the study of, for example, behavioral effects of identified circuitry.

Cells with the characteristics of SGC neurons are found in all avian species examined so far and also in reptiles and mammals, suggesting that basic functions of the optic tectum are performed by these neurons and that these functions appear to have been partially conserved during evolution.

Multisensory Integration in the Optic Tectum

In most vertebrates, the optic tectum is the essential multisensory integration center that enables the animal to orient and interact with its spatial environment. The organization of the visual sensory afferents is essentially topographic, with distortions in the representation of retinal coordinates when retinal specializations (e.g., the area centralis) exist. The retinal topography is transferred to the optic tectum through molecular gradients, which in essence endow each axon with an information about the appropriate position in the tectum it has to project to. In contrast, the integration of additional sensory input to the tectum is much more difficult because spatial information is not necessarily contained in the sensory epithelia. For example, spatial relations in the auditory system have to be computed from binaural cues through a complex circuitry in the brain stem, and space-specific neurons are first found in a subnucleus of the auditory midbrain. From this subnucleus, a topographical projection conveys the information to the intermediate and deep layers of the tectum. The alignment of the two differentially obtained spatial maps in the optic tectum is therefore of critical importance and has to be dynamically regulated because changes in the interaural distance during development, for example, affect binaural parameters and thus the map computation.

The mechanisms of these alignments have been worked out in the midbrain of barn owls in which the

sensory input had been manipulated during development (in the original literature, the auditory midbrain in owls is termed inferior colliculus). In a seminal series of experiments, young owls were raised with prismatic spectacles that displaced the visual field horizontally. Under these conditions, both the visual and the auditory space maps in the optic tectum were found to be shifted relative to the normal situation, indicating that the visual system generates the master coordinate system in the tectum and that the auditory afferents have to find their correct position in this map. Through a combination of anatomical and physiological work, it was shown that the plastic changes occur within the auditory midbrain prior to the subnucleus that projects on the optic tectum. The changes are achieved through axonal outgrowth, the formation of new glutamatergic synapses, and the inhibition of old synaptic connectivity through GABAergic mechanisms. The capacity for huge plastic changes in the auditory midbrain circuitry is most pronounced in young animals, whereas in juvenile individuals the step size for rewiring is drastically reduced. However, there is evidence that in juvenile animals an additional site of adaptive plasticity exists in the optic tectum itself. In experiments with prism attachment to juvenile owls, the auditory map in the superficial layers was found to be shifted further than the auditory map in the deep layers. Likewise, the topography of intrinsic connections between the superficial and the deep layers was found to be systematically modified. Unfortunately, it is difficult to ascribe these changes to specific cell types of the optic tectum because the cellular elements whose responses constitute the bimodal map in the tectum have not been identified. Even though auditory responses can be readily recorded throughout the tectal layers (with the shortest latencies in the deep layers), the dendritic spread of tectal neurons impedes the determination of the cells. The size of the visual receptive fields suggests that a cell type with large dendritic fields (e.g., an SGC neuron) is involved, but the evidence is only circumstantial. This is also a problem for the identification of the neurons that, on mismatch of visual and auditory maps in the tectum, generate the error signal to the auditory pathway that initiates plastic changes. Under normal conditions, this error signal is under GABAergic inhibition and thus not recordable in the auditory midbrain. After the blocking of inhibition in the tectum, a topographic visual signal can be recorded in the auditory midbrain; the cellular origin of this signal in the tectum is, however, still under debate.

The integration of other sensory modalities into the multimodal maps in the optic tectum appears to follow the pattern laid out by the auditory system. In fish and aquatic amphibia, lateral line inputs are computed into a map of hydrodynamic topography, and cells responsive to visual and lateral line inputs are found in the deep layers of the tectum. Likewise, in weakly electric fish the electrosensory afferents are fed into the deep tectal layers where bimodal cells have been recorded. Also, somatosensory and even infrared input is topographically represented in the deep layers. With respect to magnetic input, the literature is ambiguous; although in older studies avian neurons in the superficial and intermediate layers were reported to respond to directional changes in the ambient magnetic field, these data have been questioned and remain an unresolved issue.

Behavioral Aspects

As for any given brain area, the function of the optic tectum can only be completely understood when the integration of this structure into the entire processing circuits of the brain is taken into account. However, much of the research on the tectum has been motivated by the wish to understand a specific behavior on a mechanistical level, and many researchers have thus focused on specific visual functions. Because the prey-catching behavior in anuran amphibia is a famous example for this approach and is often referred to as the defining function of the tectum, a short overview of the research in this field is given next.

The complete sequence of feeding in a toad can be divided into prey recognition, prey localization and distance estimation, prey approach behavior, and prey catching. Because a partial lesioning of the optic tectum resulted in a neglect of prey objects in the corresponding visual field, it was hypothesized that much of the necessary computations take place in the optic tectum. In a series of studies, the response characteristics of many neurons in the retina and the optic tectum were analyzed and classified according to their stimulus preferences. On the basis of these cell classes, a network was proposed that included the various cell types in a hierarchical network. A prey object was thought to excite this network and to eventually lead to the activation of a specific cell type (the famous T 5.2 neuron) that was considered to function as the main tectal output neuron toward premotor and motor areas. Upon activation of T 5.2 neurons in combination with other tectal cell types, various prey-approach behaviors were elicited, until finally the snapping response was activated. Although this concept was very influential, it was criticized in several respects, one being the unnecessary revival of the detector neuron or command neuron concept. More significantly, it was shown that the tectum was not the place where a complete transformation of

visual information into motor commands was performed. Rather, other nuclei in the forebrain, thalamus, and tegmentum contribute to the network activity through interconnections and projections to the premotor areas. Thus, although the optic tectum seems to be the essential area to identify and locate a prey object, the final decision on the behavioral response appears to be made on the level of the premotor and motor nuclei, not on the level of the optic tectum.

The example of the neural control of prey catching in toads again emphasizes the point that brains are highly integrated networks with a massively parallel architecture. At the same time, it highlights the level of analysis obtainable by a thorough analysis of the circuitry and the physiology of a vertebrate optic tectum. This is especially promising because the investigation of the optic tectum's role in more complex levels of brain function is under way. As an example, the tectal contribution to bottom-up attentional mechanisms is being studied in a variety of species. Because the tectofugal pathway conveys information on many aspects of a visual stimulus toward the forebrain and attentional mechanisms are most likely implemented early on to allow for fast analysis and fast responses, tectal circuitry might well turn out to be of fundamental importance. From this circuit analysis point of view, the tectum remains a system of choice, even after more than 100 years of research into the cellular elements and their functions.

See also: Activity in Visual Development; Midbrain Patterning; Multisensory Convergence and Integration; Retinal Models; Sensorimotor Integration: Models; Sound Localization: Neural Mechanisms; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Visual Cortex in Humans; Visual Attention; Visual Cortical Models of Orientation Tuning.

Further Reading

Butler AB and Hodos W (1996) *Comparative Vertebrate Neuroanatomy.* New York: Wiley-Liss.

Ewert JP, Capranica RR, and Ingle DJ (eds.) (1983) *Advances in Vertebrate Neuroethology.* New York: Plenum Press.

Hartline PH, Kass L, and Loop MS (1978) Merging of modalities in the optic tectum: Infrared and visual integration in rattlesnakes. *Science* 199: 1225–1229.

Heiligenberg W (1990) Electrosensory systems in fish. *Synapse* 6(2): 196–206.

Hofmann MH, Wojtenek W, and Wilkens LA (2002) Central organization of the electrosensory system in the paddlefish (*Polyodon spathula*). *Journal of Comparative Neurology* 446: 25–36.

Karten HJ (1979) Visual lemniscal pathways in birds. In: Granda AM and Maxwell JH (eds.) *Neural Mechanisms of Behavior in the Pigeon*, pp. 409–430. New York: Plenum Press.

Knudsen EI (1982) Auditory and visual maps of space in the optic tectum of the owl. *Journal of Neuroscience* 2: 1177–1194.

Luksch H (2003) Cytoarchitecture of the avian optic tectum: Neuronal substrate for cellular computation. *Reviews in the Neurosciences* 14: 85–106.

Luksch H, Gauger B, and Wagner H (2000) A candidate pathway for a visual instructional signal to the barn owl's auditory system. *Journal of Neuroscience* 20: RC70.

Mey J and Thanos S (2000) Development of the visual system of the chick. I: Cell differentiation and histogenesis. *Brain Research Reviews* 32: 343–379.

Roth G (1987) *Studies of Brain Function, Vol. 14: Visual Behavior in Salamanders.* Berlin: Springer.

Thanos S and Mey J (2001) Development of the visual system of the chick. II: Mechanisms of axonal guidance. *Brain Research Reviews* 35: 205–245.

Vanegas H (ed.) (1984) *Comparative Neurology of the Optic Tectum.* New York: Plenum Press.

Witten IB and Knudsen EI (2005) Why seeing is believing: Merging auditory and visual worlds. *Neuron* 48: 489–496.

Optical Imaging of Intrinsic Signals

N Pouratian, University of Virginia, Charlottesville, VA, USA
S Sheth, Massachusetts General Hospital, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Optical imaging of intrinsic signals maps the brain by measuring intrinsic activity-related changes in tissue light reflectance and absorbance. Functional physiological phenomena, such as changes in local blood volume, hemoglobin oxygenation, and light scattering, that are coupled to neuronal activity, result in intrinsic tissue light reflectance changes that are exploited to map functional brain activity. This approach offers a distinct advantage over extrinsic signal imaging, such as fluorescent dye imaging, which may cause phototoxicity in *in vivo* preparations and may therefore alter normal physiology. By not requiring any tissue contact, optical imaging of intrinsic signals is also ideally suited for studying chronic preparations that an investigator may wish to image over days or months or for intraoperative mapping of the human cortex during neurosurgery.

Compared to other neuroimaging modalities, optical imaging of intrinsic signals offers a superior combination of spatial sampling, spatial resolution, and temporal resolution. The spatial resolution of optical imaging is unparalleled among *in vivo* imaging techniques (on the order of micrometers), making it ideal for studying fine functional organization at the level of cortical columns and their supplying microvasculature. Although the temporal resolution of optical imaging is inferior to electrophysiological techniques, imaging is commonly performed at video frame rates (30 Hz), which is more than sufficient for imaging perfusion-related responses, which evolve over seconds.

Intrinsic optical changes were first used to map cortical activity *in vivo* in 1986 by Amiram Grinvald and colleagues. Since this initial report, intrinsic optical changes have been reported in rodents, cats, monkeys, and humans. Because of the advantages it offers, the number of studies using optical imaging of intrinsic signals has grown rapidly.

Sources of Intrinsic Signals and Wavelength Dependency

Rather than directly measuring neuronal activity, optical imaging of intrinsic signals detects hemodynamic events triggered by neuronal activity. An increase in local neuronal activity results in the release of various metabolites and neurotransmitters, including nitric oxide, lactate, adenosine, ATP, calcium, and potassium. Their release in turn triggers a cascade of metabolic and vascular events, including increased oxygen consumption and vasodilation of regional feeding arterioles, resulting in local changes in blood flow, volume, and hemoglobin oxygenation. These events occur over hundreds of milliseconds to seconds following neuronal activation and extend in space over hundreds of micrometers to millimeters from the locus of activity. Functional maps based on hemodynamics are the result of a convolution of neuronal activity with this vascular and metabolic response, and they are therefore spatially and temporally broader than electrophysiologic events. Since optical signals are based on these hemodynamic responses, they must be interpreted appropriately and not accepted blindly as maps of neuronal activity *per se*.

The various components of the hemodynamic response can be selectively imaged by choosing the appropriate illumination wavelength during data acquisition (Figure 1). Each component has its own characteristics, advantages, and disadvantages. The most important contributors to optical signals are changes in blood volume, oxygenation, and light scattering.

Blood Volume Component

Hemoglobin is the most important absorber of light in the visible spectrum. The most prevalent and relevant isoforms are oxyhemoglobin and deoxyhemoglobin. At wavelengths at which the absorption spectra of oxy- and deoxyhemoglobin intersect ('isosbestic wavelengths,' including 550, 569, and 586 nm), light is equally absorbed by these two moieties. Therefore, light absorption is a function of the total hemoglobin content, which is proportional to cerebral blood volume (CBV; assuming constant hematocrit). Optical maps at these wavelengths therefore provide maps of CBV.

Following neuronal activation, relaxation of smooth muscle tone in feeding arterioles produces an increase in vessel diameter and regional CBV. The spatial density of the vascular tone control mechanisms determines the spatial specificity of CBV changes: the higher the spatial density, the greater the specificity. In regions such as primary sensorimotor cortices, regional blood flow is finely regulated at the level of precapillary arterioles resulting in CBV maps with high spatial resolution and the ability to distinguish submillimeter functional columns. The magnitude of CBV signals scales linearly with changes in underlying neuronal activity, enabling accurate quantification of changes in underlying neuronal activity.

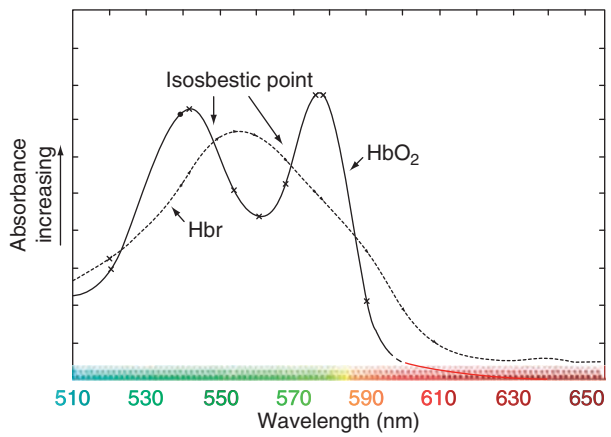

Figure 1 Hemoglobin spectra and selection of wavelengths for optical imaging. Different physiological responses can be emphasized by imaging at different wavelengths. The absorption curves for oxy- and deoxyhemoglobin are displayed to highlight how imaging at different wavelengths may emphasize different phenomena.

Oxygenation Component

In the red region of the visible spectrum (>600 nm), deoxyhemoglobin absorbs light several times more than oxyhemoglobin. Reflectance changes in this wavelength band therefore provide estimates of changes in hemoglobin oxygenation. If other components are constant, a relative deoxygenation results in decreased reflectance, and a relative hyperoxygenation results in increased reflectance. This signal is closely related to the blood oxygenation level-dependent (BOLD) signal commonly used in functional magnetic resonance imaging (fMRI).

Within a few hundred milliseconds following neuronal stimulation, the cerebral metabolic rate of oxygen ($CMRO_2$) increases. Because this increase in $CMRO_2$ precedes changes in blood flow or volume, oxygen extraction necessarily increases, resulting in local deoxygenation. This can often be seen as a decrease in the optical signal in the red region of the visible spectrum. This phenomenon is known as the 'initial dip' in the fMRI literature. The subsequent influx of oxygenated blood washes out this effect and tissue oxygenation overshoots the baseline, producing a relative hyperoxygenation and increase in the optical signal that peaks between 3 and 5 s.

Because the initial dip only occurs in the stimulated region, the spatial specificity of the response is high. Its magnitude is small, however, making it difficult to detect and highly susceptible to degradation by vasomotion or other effects of the subject's baseline physiological state. The spatial distribution of the later hyperoxygenation response is biased toward draining veins, which show the largest magnitude hyperoxygenation change. These findings are consistent with BOLD fMRI. The correlation between changes in the magnitude of these oxygenation responses and underlying neuronal activity is less than that of CBV signals.

Light Scattering Component

At long visible or near infrared wavelengths (>670 nm), hemoglobin absorption decreases significantly. In this regime, light scattering becomes an important contributor to the optical signal. Light scattering changes arise from ion and water flux; morphological changes (i.e., expansion or contraction) in vascular, cellular, and extracellular conformations (which may be related to ion and water movement); and blood volume changes. Increased scatter results in decreased reflectance and thus a decrease in the optical signal, typically peaking at 4–6 s after stimulation onset.

Resolving the Components of Optical Imaging Signals

Single-Wavelength Imaging

As alluded to in the previous section, the different components of the optical signal are weighted differentially at different wavelengths. The different hemodynamic signals can therefore be emphasized by imaging at specific wavelengths of light that either accentuate the differences between the chromophores (e.g., at 610 nm, deoxyhemoglobin absorbs significantly more than oxyhemoglobin) or minimize the difference between the chromophores (e.g., 569 nm is an isosbestic point at which oxy- and deoxyhemoglobin absorb equally).

The time course and spatial involvement of optical responses at different wavelengths are unique. For single-wavelength imaging, there are two major patterns of intrinsic signal time courses to physiological cortical activation: monophasic (decreased reflectance) and biphasic (initially decreased reflectance followed by increased reflectance). In the visible spectrum up to ~590 nm, functional intrinsic signal changes monophasically decrease reflectance. Between 600 and 760 nm, a biphasic pattern is observed. Finally, at wavelength greater than 760 nm, a monophasic pattern reminiscent of wavelength less than 590 nm is observed. For monophasic responses, responses generally appear approximately 1 s after stimulus onset, peak 3 or 4 s after stimulus onset, and return to baseline by approximately 8 s. For the biphasic responses, the initial response to stimulation appears within 500 ms, peaks 1 or 2 s after stimulus onset, reverses phases at approximately 2 or 3 s, peaks in the opposite polarity approximately 4 or 5 s after stimulus onset, and returns to baseline by approximately 10 s.

Imaging reflectance changes at a single wavelength provides high-resolution, two-dimensional functional maps. However, single-wavelength imaging suffers from an important shortcoming: ambiguous etiology. Reflectance changes at a particular wavelength emphasize different aspects of the hemodynamic response, but the distinction is not complete. Each component (i.e., blood volume, oxygenation, scattering, and others) contributes to the optical signal at every wavelength to some degree. Cross-talk between the hemodynamic responses can make single-wavelength imaging difficult to interpret.

Optical Spectroscopy

Optical intrinsic signal spectroscopy is a variant of optical imaging designed to better isolate the physiological processes that give rise to the optical signals. Before capturing reflected light on the imaging camera, the light passes through a spectrographic slit. The one-dimensional column of light is then incident upon a diffraction grating that disperses the light into its constituent wavelengths along a second orthogonal axis. This two-dimensional 'spatiospectral' image is then refocused on a second image plane and captured by the camera. The x-dimension of the image represents the wavelength of light at a particular point along the slit, and the y-dimension represents vertical position. Spatiospectral images are taken over time to capture the hemodynamic response, as in other techniques.

This approach essentially sacrifices one dimension of spatial information for an extra dimension of spectral information. The advantage gained is the ability to fit the spectra acquired over time to a model containing physiological parameters that are known to change during the hemodynamic response and affect light reflectance. Optical spectroscopy therefore provides changes in physiologically relevant variables rather than reflectance changes.

Optical Spectroscopic Imaging

Optical imaging offers high spatial and temporal resolution while optical spectroscopy proves more accurate physiologic measures of hemoglobin oxygenation and concentration. Studies combining the advantages of these two techniques are now beginning to emerge. For spectroscopic imaging, full field images are acquired at multiple wavelengths, either in an interleaved manner with a fast filter changer or simultaneously using a beam splitter. The multiple wavelengths at each pixel are then treated as a discrete spectrum, and these pixel-by-pixel 'spectra' are fit to the spectroscopic model to extract value for the physiological parameters at each pixel. This approach produces high-resolution functional maps of hemodynamic responses, permitting the investigation of their evolution within parenchymal, arteriolar, and venous compartments with excellent detail and specificity.

Technique

Selection of Species

Optical imaging of intrinsic signals is done in a variety of species, including rodent, cats, nonhuman primates, and humans. Each offers distinct advantage and drawbacks. The rodent model offers a well-characterized lissencephalic somatosensory cortex with little intersubject variability, the potential to perform population-based studies, and the availability of genetically modified strains. The major drawback of the rodent model is the question of extensibility of results to human subjects. Cats have been used extensively for imaging of intrinsic optical signals, furthering our understanding of both neurophysiology and the organization of the visual cortex. The larger size of the brain relative to the rodent and its greater complexity are both significant advantages of using this subject over rodents. The higher complexity of the cat brain makes it more likely that results can be extrapolated to humans. Although results in primates are most extensible to human subjects, this model is limited by cost and the inability to perform population studies. Furthermore, because of the thickness and density of the cranium and dura, imaging in primates requires craniotomy and removal of the dura. Intraoperative mapping in humans provides an unparalleled opportunity to examine the basic physiology and organization of the functioning human brain. At the same time, intraoperative mapping poses unique challenges for data acquisition, analysis, and interpretation. These include the unique demands of the operating room environment with its physical constraints of sterility and mobility, time constraints, cooperation of awake patients, and other considerations.

Cortical Immobilization

To take advantage of the superior spatial resolution of optical imaging, it is critical that images acquired before and during activation be in identical locations since the former will be subtracted from the latter during analysis. The movement of the brain due to respiration and heartbeat presents a major obstacle. Five different strategies have been devised to overcome this challenge: the cranial window, the thin-skull preparation, stabilization using glass plates, synchronization with respiration and heartbeat, and postacquisition image registration. These approaches are not mutually exclusive, and in several circumstances more than one strategy can be used in a single experiment.

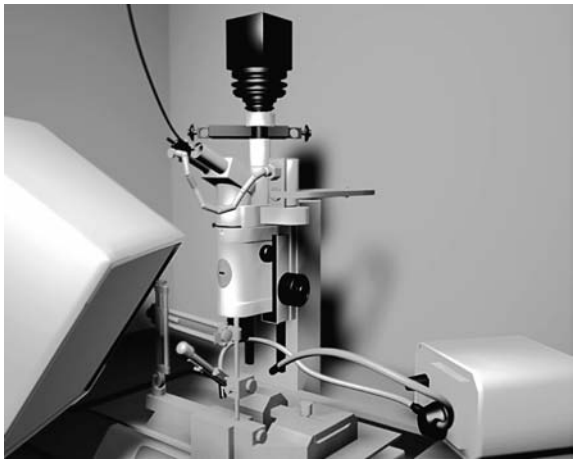

Figure 2 Optical imaging/spectroscopy apparatus. The experimental setup centers on the stereotax, in which the subject's head is immobilized. The cortex is illuminated by a light source, and reflected light is captured by the microscope optics, filtered for wavelength selection, and directed to both an imaging camera and, if desired, a spectrophotometer via fiber optic. Concurrent electrophysiology may be performed by mounting a recording electrode on the stereotax. Reproduced from Sheth SA, Nemoto M, Guiou M, et al. (2004) Linear and nonlinear relationships between neuronal activity, oxygen metabolism, and hemodynamic responses. *Neuron* 42: 347–355, with permission from Elsevier.

Data Acquisition

The imaging apparatus consists of a camera (a photodiode array, video camera, or a charge-coupled device) and an illumination source (voltage stabilized with appropriate filters for wavelength selection). The imaging camera is mounted on a microscope or lens assembly focused on the subject. For spectroscopy, a spectrophotometer and camera are also necessary. These may be directly attached to the microscope, or light may be guided to the spectrophotometer via fiber-optic cables. The latter option is particularly useful when simultaneously combining imaging and spectroscopy (Figure 2).

Once the cortical area of interest has been exposed, the imaging device is positioned appropriately, and images of the cortex are taken at rest and during activation. Analysis of optical images often requires at least one baseline image (i.e., prior to stimulus onset) in order to determine the change in signal from baseline. Most of the components that contribute to intrinsic signals return to baseline 8–10 s after stimulus offset.

Optical Imaging Data Analysis

Analysis of Single-Wavelength Data

For ratio analysis, a reference image is subtracted and divided from all images in a time series on a pixel-by-pixel basis. These ratio images represent fractional change from baseline at each spatial location. The reference image may be a prestimulus baseline image or a so-called cocktail image, which is a sum of multiple individual cortical images under different stimulus conditions, intended to represent the uniformly activated cortex. Using a cocktail blank assumes *a priori* knowledge about cortical functional architecture and neuronal interactions. Subsequent analyses can then be performed to study the spatial and temporal properties of the response, depending on the study purpose.

Analysis of Spectroscopic Data

Optical spectroscopy models are usually based on the Beer–Lambert law, which describes light attenuation in the presence of absorbers:

$$\mathrm{Abs} = \sum_i \varepsilon_i c_i l$$

where ε_i is the extinction coefficient of the ith absorber, c is the absorber concentration, and l is the pathlength through the tissue. In living tissue under normal physiological circumstances, deoxy-(Hbr) and oxyhemoglobin (HbO_2) are the most important absorbers in visible wavelengths. Calculation of attenuation due to absorption requires knowledge of the optical pathlength in the tissue, which is typically much greater than the tissue thickness due to scattering and is wavelength dependent. The pathlength may be calculated using Monte Carlo simulation or measured using a phantom. Attenuation is thus the exponential decrease in light intensity along the nonlinear path taken by the scattered photons:

$$\log\frac{I_0^\lambda}{I(t)} = \left(\varepsilon_{\mathrm{Hbr}}^\lambda \cdot \Delta[\mathrm{Hbr}](t) + \varepsilon_{\mathrm{HbO_2}}^\lambda \cdot \Delta[\mathrm{HbO_2}](t)\right) \cdot l^\lambda + \Delta S$$

where I_0 is the measured prestimulus intensity, $I(t)$ is the intensity time course, $\Delta[\mathrm{Hbr}](t)$ and $\Delta[\mathrm{HbO_2}](t)$ are the time courses of change in absorbers, S is scattering, and superscript λ indicates wavelength dependence. Spectral data are fit to the model using a least-squares analysis to extract the time courses of the model parameters (i.e., Hbr and HbO_2).

Combining Optical Imaging with Other Techniques

Intermodality comparisons are useful because each imaging technique offers different information about the cortical organization, anatomy, and physiology of the cortex of interest. Because each technique offers slightly different information about perfusion responses and holds different assumptions, optical imaging can be combined with electrophysiologic recordings or

other perfusion-related brain mapping modalities to characterize the etiology of perfusion-related neuroimaging signals.

Combining Optical Imaging with Electrical Recordings

Evoked field potential measurement or single-unit recordings can also be performed simultaneously with intrinsic signal imaging since the presence of electrodes does not interfere with optical responses. Such studies have been instrumental for investigating characteristics of neurovascular coupling and corroborating hemodynamic responses with actual measurements of neuronal activity. Optical signals correlate better with field potentials, which measure aggregate input activity in the region, than with spiking activity, which measures neuronal output.

Combining Optical Imaging and fMRI

BOLD fMRI is one of the mostly commonly used perfusion-based brain mapping modalities. Studies comparing BOLD with optical intrinsic signals are particularly advantageous for characterizing BOLD signals because the etiology of optical signals has been well scrutinized. Although many cite optical imaging studies to support the etiology of the BOLD signal, only a few intrasubject comparisons of optical intrinsic signals and BOLD fMRI exist.

A spatial/temporal mismatch has been noted between the early negative 610 nm response (which is attributed to increases in deoxyhemoglobin) and the BOLD signal in response to brief (2 s) somatosensory stimulation. Whereas the BOLD signal is centered on sulci, the negative 610 nm response is centered on gyri. In contrast, the positive 610 nm response and the BOLD signal demonstrate much greater spatial correlation (than the negative 610 nm response and the BOLD signals). Despite the greater spatial correlation, the positive 610 nm response still demonstrates greater gyral activations than BOLD signals.

Comparing fMRI signals, which are tomographic, with optical signals, which are surface signals, presents a challenge. In order to compare mapping signals from multiple modalities within a single subject, the different modalities must be warped into a common space. The common space used is the three-dimensional cortical model of each subject's brain (generated from high-resolution T1-weighted MR scans).

Combining Optical Imaging and Laser Doppler Flowmetry

Laser Doppler flowmetry (LDF) measures cerebral blood flow (CBF) at a single point on the cortical surface using a confocal laser and detector pair. This method can easily be incorporated into the experimental setup to provide information on CBF, a hemodynamic variable not directly assessable by optical imaging. Simultaneous imaging of changes in CBF and hemoglobin oxygenation permits the calculation of $CMRO_2$. Two-dimensional images of CBF can also be obtained by scanning the laser in a raster pattern across the surface or using laser speckle imaging, a related modality that generates two-dimensional CBF maps. When combining LDF with optical imaging, an appropriate filter should be placed in the optical imaging light path to prevent interference between the two. Most LDF devices use a helium–neon laser that operates at 632.8 nm; thus, this wavelength must be excluded both from incidence on the cortex by the optical imaging illuminator and from detection, or at least analysis, by the imaging camera.

Applications

Characterization of Visual and Somatosensory Cortices

The high spatial resolution of optical imaging has made it an ideal modality for characterizing the fine functional architecture of sensory (both visual and somatosensory) cortices on a submillimeter scale. By imaging visual cortex responses to gratings of different orientations, Bonhoeffer and Grinvald demonstrated that neurons responsive to gratings of the same orientation are grouped together and are organized in pinwheel-shaped patterns. This underlying organizational element is consistent in visual cortex orientation domains from cats to primates.

Similarly, optical imaging of intrinsic signals has been used to demonstrate ocular dominance domains in primary and secondary visual cortex of macaque monkeys. Studies in conjunction with single-unit electrophysiology and cytochrome oxidase (CO) histology have better characterized the functional organization within the CO stripes of visual area V2 of primates, detailing previously unrecognized subcompartments within individual CO stripes that are specific for color, orientation, and retinal disparity.

Optical imaging has similarly been applied in human subjects to characterize language and somatosensory (hand and face) cortices. Pouratian and colleagues have studied the organization of language cortices both within and across languages and demonstrated unique spatial and temporal activation patterns of language areas in a task- and language-dependent manner.

Characterization of Functional Perfusion and Neurovascular Physiology

Since perfusion-related imaging techniques (e.g., fMRI and optical imaging) measure hemodynamic

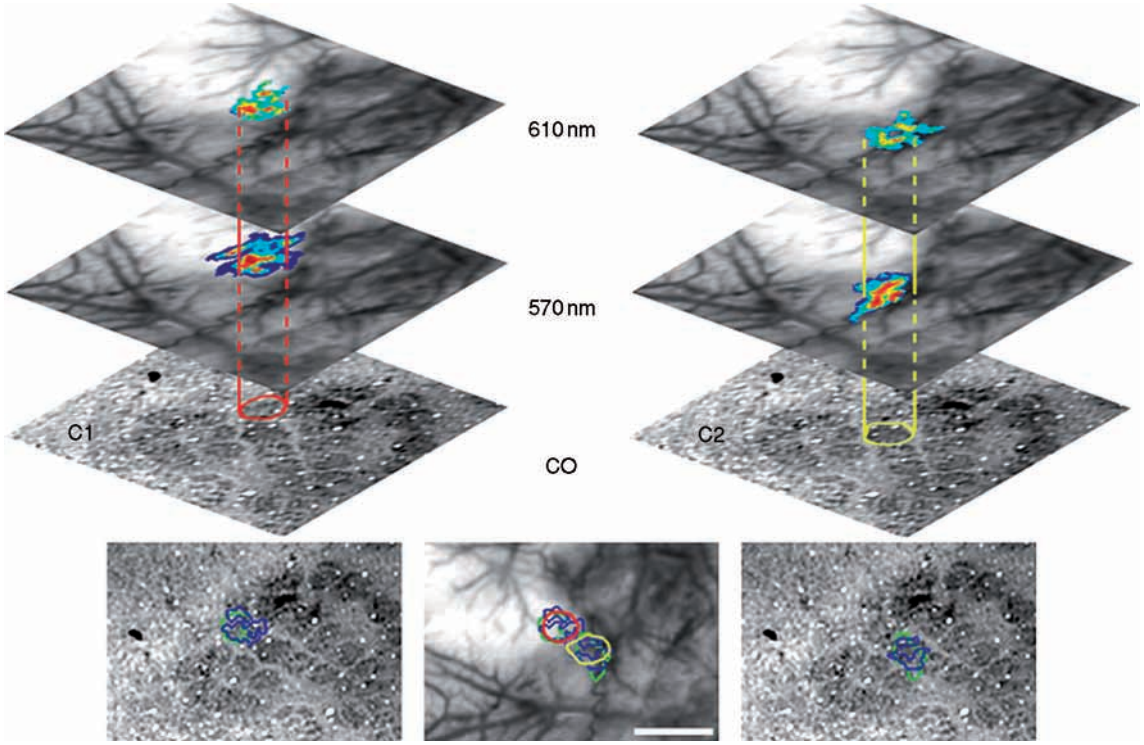

Figure 3 Columnar specificity of optical maps. Columnar specificity of the hemodynamic responses was studied in the rat barrel cortex. Vertical and flat projections of neighboring columns identified by cytochrome oxidase (CO) staining are shown in red and yellow. These projections pass through peak CBV and early deoxygenation responses overlaid on raw images, demonstrating localization of both responses with underlying anatomy. Contours enclosing the peak response area for CBV (green) and early oxygenation (blue) responses are overlaid on CO images. Contours for both responses demonstrate the spatially distinct representations of both hemodynamic responses. Reproduced from Sheth SA, Nemoto M, Guiou M, et al. (2004) Columnar specificity of microvascular oxygenation and volume responses: Implications for functional brain mapping. *Journal of Neuroscience* 24: 634–641, with permission. Copyright 2004 by the Society for Neuroscience.

responses instead of neuronal activity directly, it is imperative to understand the relationship between the two. Several optical imaging studies have investigated neurovascular coupling with simultaneous optical imaging and electrophysiology. Sheth and colleagues showed that the CBV response co-localizes to active cortical columns more precisely than the late oxygenation response, reflecting the fact that blood volume is controlled on a fine spatial scale at the level of precapillary arterioles, whereas hyperoxygenation changes are greatest in downstream draining veins that can be hundreds of micrometers to millimeters from active neurons (Figure 3). In addition, changes in the magnitude of CBV correlate linearly with input synaptic activity, as measured by evoked field potentials. On the other hand, hyperoxygenation changes are nonlinearly related, making their interpretation more challenging.

The degree to which neurons rely on oxidative metabolism has been a topic of controversy for years. Measuring cerebral oxygen metabolism with optical spectroscopy and LDF, Mayhew and colleagues demonstrated that oxygen utilization increases within 1 s of neuronal activity. Sheth et al. additionally found that oxygen metabolism increases linearly with neuronal activity beyond a certain threshold, suggesting the presence of an oxygen buffer or the reliance on nonoxidative respiration for low-level activity.

Imaging Development and Plasticity

Several groups have taken advantage of the ability to use optical imaging for chronic imaging to characterize both developmental and plasticity-related changes in functional architecture and perfusion-related signals. The development of orientation domains has been characterized in ferret visual cortex, from the time that they are first observed postnatally to the time at which they are fully developed. Pouratian and collaborators characterized developmental changes in intrinsic signals in the rodent barrel cortex, showing that whereas some aspects of hemodynamic signals change throughout the developmental period, others remain relatively stable and unchanged throughout development.

Optical imaging has also been used to image neocortical plasticity. Whereas Prakash and collaborators demonstrated rapid representational changes in barrel cortex response after addition of either brain-derived neurotrophic factor or nerve growth factor, others have used optical imaging to demonstrate more long-term changes in functional representations secondary to physical manipulations of the subject. Polley and colleagues showed that long-term whisker removal and differing degrees of environmental stimulation alter neuronal responses in the barrel cortex. While a single whisker's functional representation underwent expansion following innocuous removal of all neighboring whiskers, a contraction could also be induced, using the same manipulation, if the animal was given a brief opportunity to use its whiskers for active exploration of a different environment. Grinvald et al. investigated the stability of ocular dominance columns in macaque visual cortex with chronic optical imaging and found that they remained stable over a period of more than 1 year.

Imaging Disease

Haglund and colleagues were the first to observe optical signals in humans, reporting activity-related changes in cortical light reflectance during cortical stimulation, epileptiform afterdischarges, and cognitive tasks. Large optical signals were found in the sensory cortex during tongue movement and in Broca's and Wernicke's language areas during naming exercises.

Pouratian et al. investigated spatial and temporal properties of hemodynamic responses intraoperatively in sensory, motor, and language cortices. They demonstrated functional segregation of languages in a bilingual patient in expressive language areas

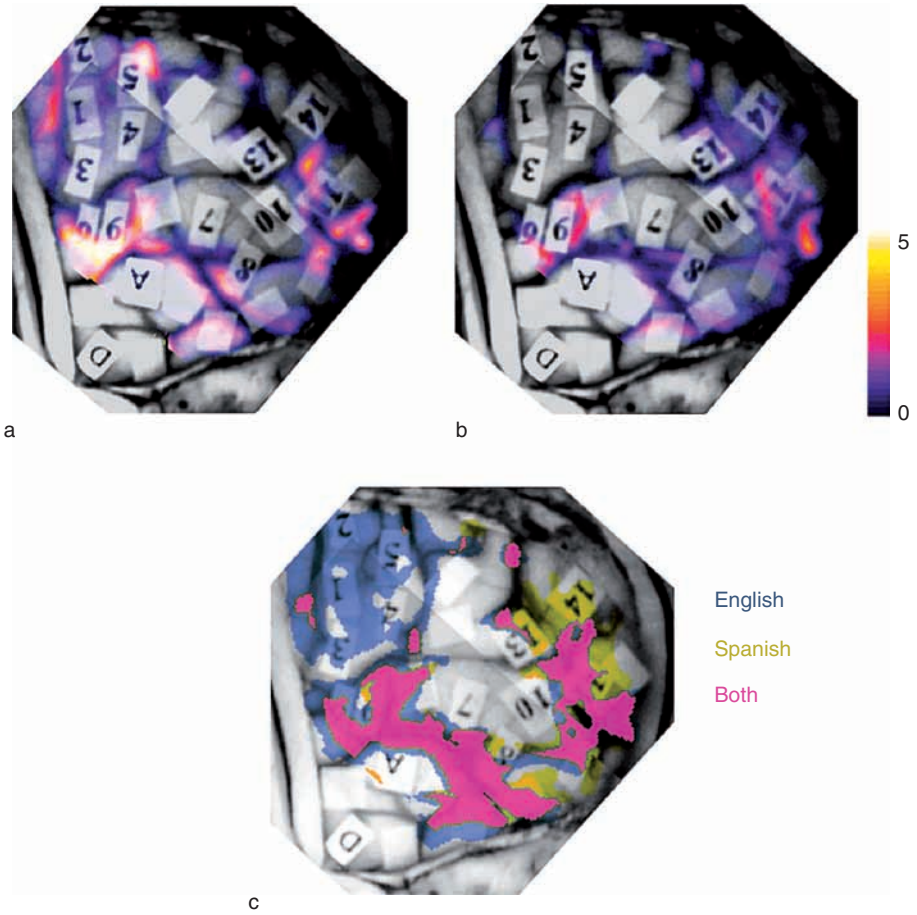

Figure 4 Optical imaging of language representations in a bilingual patient. This Spanish–English bilingual patient was undergoing brain surgery for tumor resection. She was awakened during surgery to map her language function. (a) Map of optical activity observed during English visual object naming. (b) Map of optical activity during Spanish visual object naming. Color bar represent percent optical signal change from base line. (c) Comparison of English and Spanish visual object naming. Optical maps from images shown in (a) and (b) are thresholded and displayed in blue and yellow, respectively. The pink color indicates areas of activation overlap between the two languages tested. Adapted from Pouratian N, Bookheimer SY, O'Farrell AM, et al. (2000) Optical imaging of bilingual cortical representations. *Journal of Neurosurgery* 93: 676–681, with permission from AANS.

(Figure 4). Of clinical interest, these studies evidence a developing role for optical imaging in intraoperative guidance and surgical planning, helping the neurosurgeon delineate eloquent areas that must be spared during the operation.

Conclusions

Optical imaging based on intrinsic signals is unique among functional neuroimaging techniques in that it offers both high spatial (on the order of micrometers) and temporal (on the order of milliseconds) resolution simultaneously with high spatial sampling. The resolution of optical imaging makes it ideal for a variety of experimental investigations, including studying the functional organization of and functional relationships within the brain, characterizing neurovascular physiology, and examining pathophysiological changes due to disease.

See also: Cognition: An Overview of Neuroimaging Techniques; Evoked Potentials: Recording Methods; fMRI: BOLD Contrast; Neuroimaging; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging.

Further Reading

Bonhoeffer T and Grinvald A (1991) Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* 353: 429–431.

Frostig RD, Lieke EE, Ts'o DY, and Grinvald A (1990) Cortical functional architecture and local coupling between neuronal activity and the microcirculation revealed by *in vivo* high-resolution optical imaging of intrinsic signals. *Proceedings of the National Academy of Sciences of the United States of America* 87: 6082–6086.

Grinvald A, Lieke E, Frostig RD, Gilbert CD, and Wiesel TN (1986) Functional architecture of cortex revealed by optical imaging of intrinsic signals. *Nature* 324: 361–364.

Haglund MM, Ojemann GA, and Hochman DW (1992) Optical imaging of epileptiform and functional activity in human cerebral cortex. *Nature* 358: 668–671.

Jones M, Berwick J, Johnston D, and Mayhew J (2001) Concurrent optical imaging spectroscopy and laser-Doppler flowmetry: The relationship between blood flow, oxygenation, and volume in rodent barrel cortex. *NeuroImage* 13: 1002–1015.

Martindale J, Mayhew J, Berwick J, et al. (2003) The hemodynamic impulse response to a single neural event. *Journal of Cerebral Blood Flow & Metabolism* 23: 546–555.

Mayhew J, Johnston D, Berwick J, et al. (2000) Spectroscopic analysis of neural activity in brain: Increased oxygen consumption following activation of barrel cortex. *NeuroImage* 12: 664–675.

Pouratian N, Bookheimer SY, O'Farrell AM, et al. (2000) Optical imaging of bilingual cortical representations. *Journal of Neurosurgery* 93: 676–681.

Pouratian N, Sheth SA, Martin NA, and Toga AW (2003) Shedding light on brain mapping: Advances in human optical imaging. *Trends in Neuroscience* 26: 277–282.

Pouratian N, Sicotte N, Rex D, et al. (2002) Spatial/temporal correlation of BOLD and optical intrinsic signals in humans. *Magnetic Resonance in Medicine* 47: 766–776.

Sheth SA, Nemoto M, Guiou M, et al. (2004) Columnar specificity of microvascular oxygenation and volume responses: Implications for functional brain mapping. *Journal of Neuroscience* 24: 634–641.

Sheth SA, Nemoto M, Guiou M, et al. (2004) Linear and nonlinear relationships between neuronal activity, oxygen metabolism, and hemodynamic responses. *Neuron* 42: 347–355.

Shtoyerman E, Arieli A, Slovin H, Vanzetta I, and Grinvald A (2000) Long-term optical imaging and spectroscopy reveal mechanisms underlying the intrinsic signal and stability of cortical maps in v1 of behaving monkeys. *Journal of Neuroscience* 20: 8111–8121.

Toga AW, Cannestra AF, and Black KL (1995) The temporal/spatial evolution of optical signals in human cortex. *Cerebral Cortex* 5: 561–565.

Vanzetta I and Grinvald A (1999) Increased cortical oxidative metabolism due to sensory stimulation: Implications for functional brain imaging. *Science* 286: 1555–1558.

Vanzetta I, Hildesheim R, and Grinvald A (2005) Compartment-resolved imaging of activity-dependent dynamics of cortical blood volume and oximetry. *Journal of Neuroscience* 25: 2233–2244.

Optical Monitoring of Exo- and Endocytosis

E T Kavalali, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Fluorescent Visualization of Synaptic Vesicle Fusion and Recycling

Despite their versatility, electrophysiological methods typically report neurotransmitter release from a population of synapses or release sites. In most cases, unitary quantal transmission can only be detected using recordings of spontaneous neurotransmission or under settings that activate a very small number of synapses. Moreover, these methods usually provide little or no spatial information about the origin of release. Optical experiments using fluorescent probes, on the other hand, report release at the level of individual synapses. These methods either monitor Ca^{2+} influx through postsynaptic NMDA receptors or directly monitor presynaptic properties using styryl dye imaging and green fluorescent protein (GFP)-tagged synaptic vesicle proteins.

NMDA receptor-mediated Ca^{2+} influx provides a valuable reporter of postsynaptic receptor activation in response to presynaptic glutamate release. NMDA receptors have a very high affinity for glutamate ($\sim$1 μM); thus, they can detect even subtle fusion events that may not release sufficient glutamate to activate AMPA receptors. In addition, in response to a single quantal fusion event, postsynaptic NMDA receptors are not saturated, which enables measurements of release with a large dynamic range despite their high affinity. Therefore, this approach has provided evidence of multivesicular release in central synapses. Ca^{2+} indicator dyes with adequate linear response range and affinity can be used to detect Ca^{2+} signal in the single dendritic spine evoked by quantal release events. This method, by its very nature, is limited to glutamatergic synapses or synapses with reliable postsynaptic Ca^{2+} selective channels (e.g., some subtypes of nicotinic acetylcholine receptors or purinergic receptors, such as P2X receptors). Moreover, the properties of neurotransmitter release can only be inferred postsynaptically; therefore, the same caveats that apply to electrophysiological recordings in principle apply to this method as well. However, the ability to monitor neurotransmission at the level of individual synapses and the availability of a wide variety of Ca^{2+} indicators make this a method of choice for the investigation of synaptic activation in intact tissues such as brain slices using two-photon imaging.

In addition to postsynaptic Ca^{2+} imaging, several fluorescence imaging modalities have been introduced which primarily report synaptic vesicle fusion and endocytosis. These methods either take advantage of styryl dyes that can be loaded into recycling vesicles or they utilize exogenous expression of synaptic vesicle proteins tagged with a pH-sensitive variant of GFP at regions facing the vesicle lumen. Styryl dyes, such as FM1-43 and FM2-10, are amphipathic compounds that label recycling synaptic vesicles in an activity-dependent manner. This optical approach is also useful for monitoring the kinetics of replenishment for the release ready pool of vesicles in a synapse. Furthermore, uptake and release of these dyes provides an accurate indication of synaptic vesicle reuse during activity. GFP-tagged synaptic vesicle proteins such as synaptobrevin and synaptophysin, on the other hand, provide a molecularly specific way to examine surface exposure and reinternalization of these proteins during synaptic vesicle exo- and endocytosis.

Functional Analysis of Exocytosis and Vesicle Recycling Using Styryl Dyes

In the past decade, significant advances in our understanding of the dynamics of the synaptic vesicle cycle were made possible by the availability of new optical probes. The first set of probes were the styryl dye FM1-43 and its analogs. FM1-43 is a partially lipophilic dye, which increases fluorescence 100-fold when it partitions into membranes and thus it is virtually invisible in aqueous solution. Because of its chemical structure FM1-43 cannot cross lipid membranes, thus limiting the fluorescence labeling to outer membranes and recycling vesicles (Figure 1). Using this methodology, one can detect activity-dependent vesicular dye uptake and release in synaptic connections in multiple preparations, including central neuronal cultures. Studies using FM dyes have found that activity-induced dye uptake and release operates with high fidelity. Specifically, the extent of fluorescent dye endocytosed during a particular stimulation matches the magnitude of fluorescence loss due to exocytosis in response to the same stimulation. This observation suggests that synaptic vesicles recycle without intermixing with endosomal compartments and thus the fluorescence probes do not get diluted.

The properties of styryl dyes make them useful tools for obtaining quantitative information on vesicular recycling within a presynaptic terminal. These dyes can be inserted into membranes within milliseconds; however, with increasing lengths of their hydrocarbon chains, they have decreasing rates of dissociation from

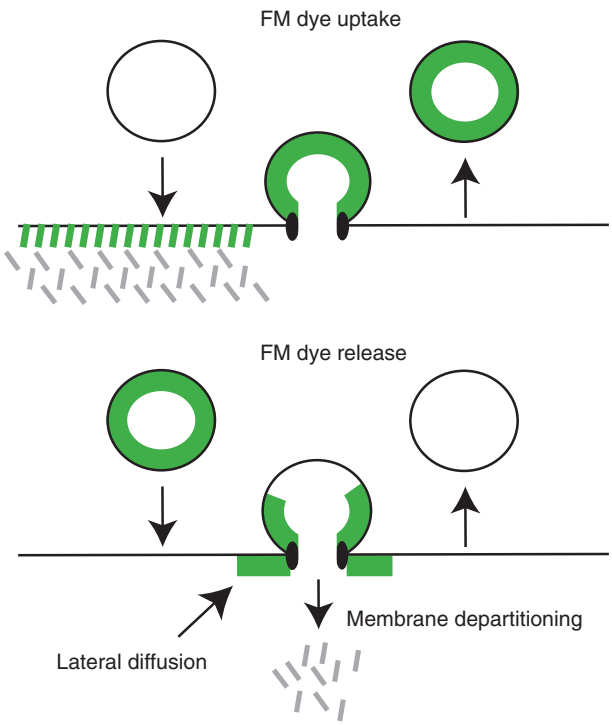

Figure 1 Optical monitoring of vesicle recycling using FM dyes. FM1-43 and its analogs are partially lipophilic dyes which fluorescence increases almost 100-fold when they partition into membranes. Therefore, they are virtually invisible in aqueous solution. Because of its chemical structure FM1-43 cannot cross lipid membranes, thus limiting the fluorescence labeling to outer membranes and recycling vesicles. During a typical experiment, stimulation in the presence of extracellular dye leads to uptake of dye molecules into vesicles that have undergone exocytosis followed by endocytosis. Subsequent washout of extracellular dye uncovers fluorescent staining specific to actively recycling vesicles during the stimulation paradigm. Consequent stimulation results in fusion of dye-loaded vesicles and loss of fluorescence trapped in synapses. This fluorescence loss is due to departitioning of the dye into aqueous solution or lateral diffusion of dye in neuronal membrane. During kiss-and-run endocytosis, a protein barrier that makes up the fusion pore restricts FM dye lateral diffusion. If vesicle retrieval is faster than the dye departitioning rate, then endocytosed synaptic vesicle may still contain residual dye.

membranes. Previous measurements have shown that FM1-43 departitions from the lipid to aqueous environment with a time constant of 2.5 s, whereas its analogs, FM2-10 and FM1-84, departition with time constants of 0.6 and 4.8 s, respectively. These differential departitioning properties of FM dyes can be exploited to estimate the kinetics of endocytosis in hippocampal synapses. Because of FM2-10's membrane dissociation rate, destaining experiments with the dye provide the closest estimates of vesicle mobilization and exocytosis. In contrast, the same experiments performed with FM1-43 may be vulnerable to dye trapping due to fast endocytosis. Since the introduction of FM dyes, several studies have reported partial release of FM1-43 upon synaptic vesicle fusion during multiple forms of stimulation. In some studies, destaining rates of synaptic boutons stained with FM2-10, FM1-43, and FM1-84 were slower in proportion to the increasing time constants of departitioning. These initial observations were later extended by experiments examining FM dye loss from single vesicle fusion events. The common interpretation of partial FM dye release is the presence of a fast endocytotic mechanism, which preferentially traps the slowly departitioning dyes, FM1-43 and FM1-84, in synaptic vesicles. Partial dye release is only possible if a barrier (presumably a protein-based fusion pore) between the vesicle membrane and the plasma membrane delimits lateral diffusion of FM dye following the formation of the vesicle–plasma membrane interface. However, in other systems such as goldfish bipolar cells, all FM dyes have been reported to laterally diffuse out of fused vesicles and thus lose fluorescence irrespective of their departitioning rates in aqueous solution.

A significant proportion of endocytosis occurs following the cessation of stimulation, albeit on a slower timescale. This slow endocytosis may require formation of clathrin coats, followed by a recycling pathway in which vesicles go through an endosomal intermediate structure and recover all of the necessary molecular components. Alternatively, recycling vesicles can be formed from membrane invaginations called 'cisternae' that develop adjacent to the active zone. Reformation of synaptic vesicles from these endosomal structures requires clathrin and clathrin adaptor proteins that bind to specific synaptic vesicle membrane proteins. Studies of the frog neuromuscular junction have shown that a slow endocytosis pathway, characterized by large membrane invaginations formed in response to prior activity, operates at rest for at least 10 min or more after cessation of exocytosis and selectively traps FM1-43 but not FM2-10. The slowly endocytosed membrane can form synaptic vesicles, which in turn can mix with the reserve pool and become reavailable for release with a delay. FM2-10 uptake, in contrast, occurs in rapidly endocytosing vesicles, which preferentially populate the readily releasable pool (RRP) and give rise to the fast component of destaining.

In a typical experiment to estimate slow endocytosis, one can stimulate synaptic terminals with 10–30 Hz field stimulation for 30–60 s and rapidly perfuse FM1-43 onto the preparation (for 60 s up to 10 min) at gradually increasing delays following the cessation of stimulation. The time course of decrease in the amount of fluorescence trapped within synapses provides an estimate of the time constant of slow endocytosis. This time constant has previously been

shown to slow with increasing extent of depletion of vesicles in the neuromuscular junction. However, in these studies, parallel control experiments are essential to account for the potential effect of changes in the level of exocytosis. Evidence indicates that the number of fused vesicles can regulate the rate of endocytosis. Under the same conditions more rapid endocytosis can be quantified when dye is applied concomitantly during stimulation and washed out rapidly afterwards with no delay. Fast endocytosis is typically tightly coupled to exocytosis, and it is thought to be critical for maintenance of sustained neurotransmission.

The measurement of the size of functionally distinct vesicle pools is critical in linking the structure and function of presynaptic terminals. Styryl dyes are also quite useful in the determination of vesicle pool sizes at the level of individual synapses. For these measurements, it is important to distinguish changes in the Ca^{2+} dependence of the release machinery from changes in numbers and maturational (release competent) states of synaptic vesicles. This distinction can be achieved by using saturating levels of Ca^{2+}-dependent stimulation or Ca^{2+}-independent stimulation, such as by hypertonic sucrose application (typically +500 mOsm), to mobilize vesicles. The recycling kinetics of RRP vesicles can be determined using uptake and release of the styryl dye FM2-10 in response to hypertonic stimulation. To determine the total recycling pool, uptake and release of styryl dyes can be induced by sustained depolarization achieved via application of elevated potassium solution. This strong stimulation can induce mobilization of virtually all vesicles capable of recycling within individual synapses. Comparison of the two fluorescence measurements (RRP vs. the total recycling pool) can reveal critical information regarding the functional allocation of vesicles to a particular pool in given condition.

Detecting the release of FM dyes after their uptake into the synapse is a commonly used method to quantify synaptic vesicle reuse in synapses. In these experiments, FM dye is applied during and/or immediately after stimulation to label endocytosing vesicles. The dye-containing solution is then swiftly washed out to prevent further dye uptake. This dye application stains a number of vesicles depending on the stimulation strength as well as the duration of dye presence. Application of a second stimulation at gradually delayed time points (typically 0–120 s) causes dye loss due to the reavailability of labeled vesicles. However, the proximity of stimulations inducing dye uptake and release makes it difficult to image actual dye loss during the second stimulation because of the high fluorescence background originating from residual dye. This shortcoming is usually circumvented by monitoring the remaining detectable staining at a later time point (typically ~5 min) after complete dye washout. The releasable fluorescence trapped in the synapse gradually decreases following stimulation applied at longer intervals or for longer durations after dye washout. The time course of this decrease in fluorescence provides an estimate of the time constant of vesicular reuse, which is the time it takes for an endocytosed vesicle to become release ready again. This is a very sensitive method for detecting vesicle reuse. As stated previously, the major caveat is the background fluorescence: although usually not sufficient to restain vesicles during stimulation, it makes fast imaging nearly impossible. Because of this problem, the vesicle reuse kinetics is typically reconstructed from a population of synapses rather than from repeated measurements at individual presynaptic terminals.

Detection of Synaptic Vesicle Exocytosis and Endocytosis Using pHluorin Tagged Synaptic Vesicle Proteins

Synaptophluorin (spH) is another fluorescent probe that is commonly used to monitor exocytosis and endocytosis of synaptic vesicles. spH was originally engineered by Gero Miesenbock and colleagues as a fusion construct of the synaptic vesicle protein synaptobrevin (also called vesicle-associated membrane protein) with a pH-sensitive enhanced green fluorescence protein (EGFP) at its C-terminal (located in the synaptic vesicle lumen). Synaptic vesicle lumen normally has an acidic pH of approximately 5.5, at which spH fluorescence is quenched. When vesicles fuse, lumenal EGFP is exposed to the extracellular pH, which results in a marked increase in its fluorescence. Studies using spH have provided valuable information on the tight coupling between exocytosis and endocytosis. One advantage of the spH probe is the molecular specificity of the fluorescence signal since the signal originates from a synaptic vesicle protein as opposed to nonspecific lipid-based fluorescence of FM dyes. Proton diffusion from a synaptic vesicle is not expected to be limited, and the time course of vesicle reacidification is thought to be fast without imposing a significant delay in the fluorescence signal during spH endocytosis. Although most studies have used synaptobrevin as a carrier for the pHluorin tag, recent studies have taken advantage of other synaptic vesicle proteins as pHluorin attachments. Surprisingly, in these experiments, different synaptic vesicle proteins tagged with pHluorin gave different results with respect to their background fluorescence (i.e., surface exposure) and retrieval. For instance, both synaptophysin (a four transmembrane domain synaptic vesicle protein) and the vesicular glutamate transporter reveal very little surface exposure (thus fluorescence) in the absence

of stimulation. Once stimulated, they show a swift increase in fluorescence without significant lateral diffusion upon fusion, which was a salient feature of spH-based measurements. The dependence of the properties of pHluorin-based measurements on the identity of the tagged protein complicates a straightforward interpretation of these experiments in terms of synaptic vesicle recycling. Despite this caveat, these measurements reveal critical information with regard to copy numbers and specific cell biological attributes of synaptic vesicle proteins (Figure 2).

Inhibition of vesicle reacidification by blockers of the vacuolar ATPase is useful for dissecting the kinetics of exocytosis and endocytosis in pHluorin-based measurements. In response to brief field stimulation, synaptophysin-pHluorin (or spH) fluorescence shows a rapid increase and a gradual slow return to baseline, presumably due to slow compensatory endocytosis. However, when the same experiment is repeated in the presence of a vacuolar ATPase blocker (e.g., folimycin), synaptophysin-pHluorin fluorescence shows a larger increase detectable within the first 500 ms. The difference between these two signals can be taken as an indication of endocytosis with the assumption that reacidification of vesicles upon endocytosis proceeds rapidly and requenches internalized pHluorin. However, current estimates of synaptic vesicle reacidification differ widely. According to some estimates, reacidification may take place within 500 ms upon endocytosis. In contrast, findings in hippocampal cultures suggest a relatively slow time course (~5 s). These slower estimates derive from measurements of synaptic vesicle reacidification during delayed endocytosis following stimulation and therefore may not take into consideration a fast component of vesicle reacidification. The vesicles that are endocytosed with a delay after stimulation may end up being slowly reacidified due to slower reassembly of their protein complements, whereas vesicles that are endocytosed rapidly during stimulation may be reacidified rapidly because their protein components (including the vacuolar ATPase) are more likely to be intact.

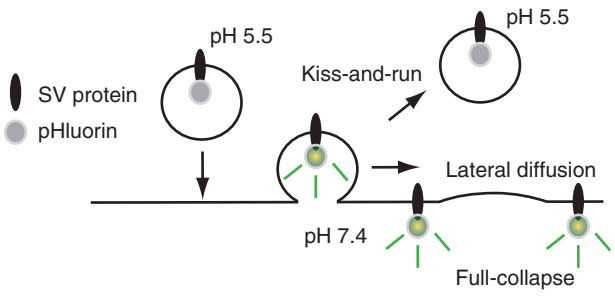

Figure 2 Synaptophluorin (spH) is a fusion construct of the synaptic vesicle protein synaptobrevin with a pH-sensitive EGFP at its C-terminal (located in the synaptic vesicle lumen). The synaptic vesicle lumen normally has an acidic pH of approximately 5.5, at which spH fluorescence is quenched. When vesicles fuse, lumenal EGFP is exposed to the extracellular pH, which results in a marked increase in its fluorescence. During endocytosis, pHluorin fluorescence is requenched as vesicle lumen becomes acidic. Following fusion, the absence of a diffusion barrier between the vesicle membrane and the plasma membrane may result in lateral diffusion of spH.

Experiments have shown that once the surface fluorescence is quenched, it is possible to monitor single vesicle fusion events using spH as a probe. Using this setting, Gandhi and Stevens detected single vesicle fusion events with square pulse-like rapid alkalinization and reacidification behavior akin to single ion channel openings. These findings suggest that the number of spH molecules per vesicle is preserved during the exocytosis and endocytosis cycle, in a manner consistent with the kiss-and-run hypothesis. However, the same results were not reproducible with a pHluorin fusion construct of synaptophysin, suggesting that rapid reversal of pHluorin signal may be due to lateral diffusion of synaptobrevin because the synaptophysin-pHluorin does not laterally diffuse upon vesicle fusion.

Comparison of Styryl Dye Imaging and pHluorin-Based Visualization of Exo- and Endocytosis

Although FM dyes and spH are commonly used to measure endocytosis and exocytosis, there are advantages and disadvantages to both. FM dye measurements have been instrumental in tracing vesicles during their cycle in the synapse. Using uptake and release of FM dyes, one can obtain results regarding vesicle mixing with endosomal compartments or repeated use of the same vesicles during activity which would be difficult to attain using probes such as spH. FM dyes report lipid cycling; spH reports the fate of synaptic vesicle protein synaptobrevin (or others) upon fusion and endocytosis. Sph is a genetically expressed probe which facilitates more reproducible repeated measurements on the same set of synapses (because it does not require dye uptake). Furthermore, spH can also be transgenically expressed in mice, which makes *in vivo* measurements of synaptic vesicle trafficking possible. However, measurement of endocytosis with spH is confounded by two factors. First, lateral diffusion of vesicular spH following fusion is commonly observed. It is unclear whether reuptake of spH truly reflects endocytosis of vesicles or, rather, reclustering and reinternationalization of synaptobrevin. It is also unknown whether endocytosis of lipid and protein components of synaptic vesicles represents

distinct processes or whether they are coupled. Second, the dependence of endocytosis measurements on vesicle reacidification may delay the fluorescence signal after spH internalization. There are also concerns with regard to using FM dyes. Typically, FM dye measurements suffer from nonspecificity. FM dyes stain the surface membrane indiscriminately; thus, they are less specific to synaptic vesicles. Moreover, changes in FM dye fluorescence typically occur slowly, on the order of seconds, making it difficult to detect events in the millisecond timescale, which more closely matches the time course of neurotransmission. In addition, after FM dye uptake, preparations need to be washed with dye-free solution for a significant period to reduce background fluorescence, which limits the rapid detection of synaptic vesicle reavailability after endocytosis.

In summary, spH provides a molecularly specific probe to monitor the coupling between exocytosis and endocytosis, whereas FM dyes are better at tagging vesicles during their cycling within the synapses. Thus, these two optical probes complement each other in monitoring distinct aspects of the synaptic vesicle cycle.

See also: Endocytosis and Presynaptic Scaffolds; Endocytosis: Kiss and Run; Fusion Pore; Synaptic Depression; Synaptic Vesicles; Vesicle Pools.

Further Reading

Aravanis AM, Pyle JL, and Tsien RW (2003) Single synaptic vesicles fusing transiently and successively without loss of identity. *Nature* 423: 643–647.

Betz WJ, Mao F, and Smith CB (1996) Imaging exocytosis and endocytosis. *Current Opinion in Neurobiology* 6: 365–371.

Gandhi SP and Stevens CF (2003) Three modes of synaptic vesicular recycling revealed by single-vesicle imaging. *Nature* 423: 607–613.

Granseth B, Odermatt B, Royle SJ, and Lagnado L (2006) Clathrin-mediated endocytosis is the dominant mechanism of vesicle retrieval at hippocampal synapses. *Neuron* 51: 773–786.

Kavalali ET, Klingauf J, and Tsien RW (1999) Properties of fast endocytosis at hippocampal synapses. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 354: 337–346.

Li Z and Murthy VN (2001) Visualizing postendocytic traffic of synaptic vesicles at hippocampal synapses. *Neuron* 31: 593–605.

Miesenbock G, De Angelis DA, and Rothman JE (1998) Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins. *Nature* 394: 192–195.

Murthy VN and Stevens CF (1998) Synaptic vesicles retain their identity through the endocytic cycle. *Nature* 392: 497–501.

Oertner TG, Sabatini BL, Nimchinsky EA, and Svoboda K (2002) Facilitation at single synapses probed with optical quantal analysis. *Nature Neuroscience* 5: 657–664.

Pyle JL, Kavalali ET, Piedras-Renteria ES, and Tsien RW (2000) Rapid reuse of readily releasable pool vesicles at hippocampal synapses. *Neuron* 28: 221–231.

Richards DA, Guatimosim C, and Betz WJ (2000) Two endocytic recycling routes selectively fill two vesicle pools in frog motor nerve terminals. *Neuron* 27: 551–559.

Ryan TA and Smith SJ (1995) Vesicle pool mobilization during action potential firing at hippocampal synapses. *Neuron* 14: 983–989.

Sankaranarayanan S and Ryan TA (2000) Real-time measurements of vesicle-SNARE recycling in synapses of the central nervous system. *Nature Cell Biology* 2: 197–204.

Voglmaier SM, Kam K, Yang H, et al. (2006) Distinct endocytic pathways control the rate and extent of synaptic vesicle protein recycling. *Neuron* 51: 71–84.

Zenisek D, Steyer JA, Feldman ME, and Almers W (2002) A membrane marker leaves synaptic vesicles in milliseconds after exocytosis in retinal bipolar cells. *Neuron* 35: 1085–1097.

Optokinetic Eye Movements

M J Mustari and S Ono, Emory University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

General Introduction: Optokinesis and Visual–Vestibular Interaction

Much of vertebrate behavior is dependent on optimal processing of visual information. The image of a stationary or moving object must be maintained relatively stable on the retina if the object is to undergo detailed examination. This is achieved by different motion compensation and tracking systems, including vestibular ocular, optokinetic, smooth pursuit, saccadic, and vergence eye movement systems. Previous studies in humans have shown that retinal image motion speeds greater than a few degrees per second cause degradation in optimal visual performance. Stable gaze during locomotion is necessary to maintain the visual image of an object of interest from slipping on the retina.

The main gaze stabilizing reflex is the vestibulo-ocular reflex (VOR). The VOR generates compensatory eye movements during head movements, allowing viewers to maintain their line of sight, or gaze, on a stationary object. Linear and rotational accelerations of the head are sensed by otolith-related and semicircular canal-related neurons, respectively. Vestibular afferents carry a head velocity signal to central vestibular neurons, which in turn drive oculomotor neurons to produce the rotational and translational vestibular–ocular reflexes (rVORs and tVORs). This simple three-neuron VOR arc can be adjusted by visual inputs traveling in relatively direct and indirect pathways.

The optokinetic system works synergistically with the vestibular ocular system to extend its dynamic range. For example, VOR gain (eye velocity/head velocity) is typically less than 1.0 when measured in darkness. However, in the presence of a visual target, the visually enhanced VOR has gain of 1.0. Visual enhancement of the VOR could be due at least in part to summation of the VOR and visually mediated tracking, including reflexlike optokinetic and volitional smooth pursuit-generated responses to retinal image motion. Such visual following responses support the VOR to maintain clear vision across a broad range of head accelerations. In addition, visual error signals provided by direction- and speed-sensitive neurons of the optokinetic system are essential for adjusting VOR gain in response to physiological or pathological challenges to vestibular–ocular function. Finally, because the adequate stimulus for the VOR is head acceleration, compensation during constant-velocity head movements would fail without the support of the optokinetic system, which produces compensatory eye movements during continuing visual motion. In this article, the focus is on the organization of horizontal optokinetic eye movements and their development.

Optokinetic Eye Movements: Afoveate Mammals

The optokinetic system plays an essential role in stabilizing the visual image on the retina by producing compensatory eye movements in the direction of visual motion. Direction-selective (DS) retinal ganglion cells provide initial visual motion information to nuclei of the accessory optic system (AOS, see later) appropriate for driving slow optokinetic eye movements of afoveate mammals.

The preferred directions for DS retinal ganglion cells of rabbits are aligned close to the activation/inactivation planes of the semicircular canals, facilitating visual–vestibular interaction.

In this article, the focus is on the horizontal optokinetic eye movements. In animals capable of making significant eye movements, large-field visual motion produces reflex following responses of the eyes. Optokinetic eye movements can follow relatively high-frequency oscillation of the visual world, as might occur during head movements associated with locomotion (Figure 1).

During unidirectional large-field visual motion over an extended time, a pattern of eye movements called optokinetic nystagmus (OKN) occurs. OKN is composed of slow, compensatory, following movements in the direction of visual motion and quick resets (Figure 2). An optokinetic response (OKR) has been observed in all vertebrates tested, from fish to humans. Here, the focus is on the OKR of mammals. Mammals lacking a fovea or significant central retinal specialization (e.g., fovea or area centralis), such as mice, rats, guinea pigs, and rabbits, typically have laterally placed eyes with minimal binocular overlap in the visual field representation from each eye. In afoveate mammals the is strongest during low-speed ($<10^\circ\ s^{-1}$) visual motion, directed toward the nose (nasalward).

The major pathway supporting optokinetic eye movements in afoveate mammals involves direct retinal inputs to brain stem nuclei of the AOS. There is little evidence for cerebral cortical control of optokinetic eye movements in afoveate mammals. The adequate stimulus for eliciting an OKR in lower

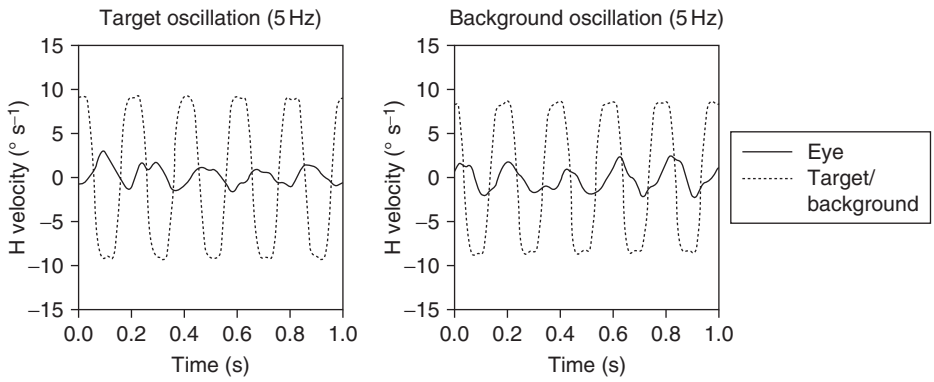

Figure 1 Examples of optokinetic eye movements elicited during high-frequency (5 Hz), low-amplitude (± 1°) oscillation of either small-diameter (0.1°; left panel) or large-field (± 75°; right panel) visual stimuli. Stimuli were rear projected on a tangent screen 57 cm distant from the eyes. Visual stimuli were moved by driving a galvanometer with a triangle-wave produced by a function generator or computer D/A. Five successive cycles of oscillation are shown. The gain of these eye movements is low (0.1–0.2) for such high-frequency, periodic motion. Traces as indicated in the key. Eye movements were recorded from a mature monkey (*Macaca mulatta*) using an electromagnetic method (CNC Engineering, Seattle, WA) employing scleral search coils.

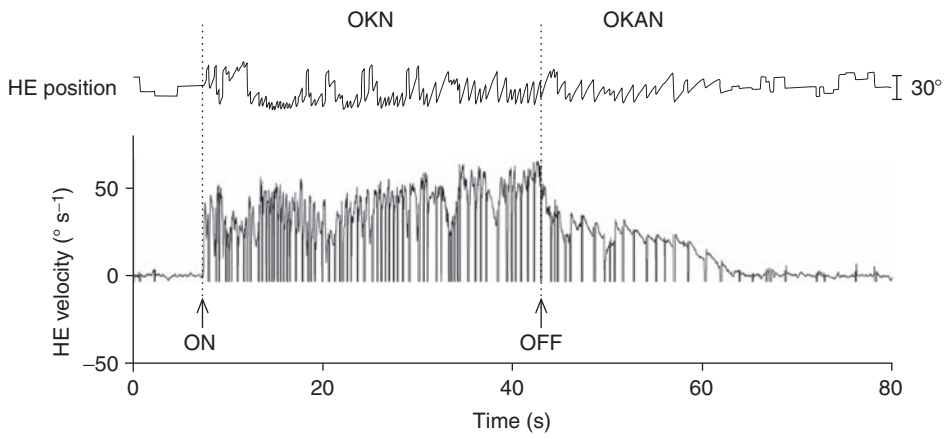

Figure 2 Example of OKN and after-nystagmus (OKAN) recorded from a mature monkey (*Macaca mulatta*). Rightward continuous visual motion (50° s^{-1}) was projected on a cylindrical screen using a planetarium projector for 30 s. Prior to stimulus onset, the monkey sat in complete darkness with the head held stable. Visual motion commenced after a variable amount of time (arrow, ON) producing a rapid rise in eye velocity followed by a more gradual buildup to visual stimulus velocity. When the visual stimulus was turned off, leaving the monkey in complete darkness, eye movements continue as OKAN. Traces from the top: horizontal eye (HE) position and horizontal eye velocity. Occasional rightward quick-phases have been removed and leftward quick-phases have been clipped for clarity. Vertical dashed lines indicate visual stimulus onset and offset times. Eye movements were recorded with the scleral search coil method.

mammals is large-field visual motion; however, small-field stimuli can also produce optokinetic following movements in frontal-eyed, foveate mammals (Figure 1). The gain of the OKR (eye velocity/visual velocity) approaches 1 for low-speed (e.g., $<10° s^{-1}$) continuous motion in afoveate mammals but can reach much higher speeds ($>100° s^{-1}$) in cats and primates.

Optokinetic Eye Movements: Primates

Primate optokinetic eye movements have similarities to those observed in afoveate mammals, with additional features related to adaptations for frontally placed eyes and foveal vision. For example, to maintain the image of a moving object on the fovea, it may be necessary to suppress the OKR or the VOR. This would occur when an object of interest moves in the direction of head movement. Suppression of OKN or VOR is made possible by extensive descending cerebral cortical projections to specific brain stem nuclei (see later).

During continuous visual motion as might occur during slow head turns, compensation would cease when the eyes reached the limit of rotation. This problem is minimized by insertion of resetting eye movements called quick-phases (Figure 2) that bring the eyes back into a usable portion of the oculomotor

range. Therefore, during unidirectional, constant-speed visual motion a characteristic pattern of slow, conjugate, compensatory eye movements with quick resets occurs; this is called OKN (Figure 2). Quick-phases have amplitude–velocity relationships that fall on the main sequence for saccadic eye movements. The quick-phases are high-velocity (100–800° s^{-1}) resetting movements that play no role in compensation *per se*, other then getting the eye back into a usable portion of oculomotor range. In contrast, the slow, compensatory eye movements of OKN can match visual stimulus velocity over a wide range (e.g., 1–180° s^{-1}) of speeds in macaques. Compensatory eye movements associated with OKN are best illustrated in eye velocity plots where saccades are removed from the records (Figure 3).

An effective method for observing all components of the OKR is to expose an alert viewer to sudden onset of large-field unidirectional visual motion. OKN is induced when the viewer is observing a moving pattern either projected on a screen or placed on the surface or a cylinder rotating around the viewer.

Figure 2 illustrates OKN of a macaque monkey during unidirectional, horizontal visual motion presented on a cylindrical screen 57 cm distant. The gain (measured at steady state) of the OKR is modulated by extent of field stimulated, stimulus contrast, and spatial frequency. To obtain an optimal response, the entire visual surround should contain moving contours and no fixation points, because fixation can stop OKN. In primates, strong optokinetic eye movements can be elicited even when only central 10° of the retina is stimulated, reflecting the foveal/ parafoveal importance in primate vision. To examine the in isolation, the head is held stable. At the start of the experiment, the animal is sitting in complete darkness, when at a random time unidirectional visual motion is provided by a rotating striped cylinder or a pattern projected with an appropriate continuous motion system (e.g., planetarium; Figures 2 and 3) is suddenly illuminated. This produces a rapid rise in eye velocity at 100 to 150 ms latency, followed by a gradual buildup in eye velocity over 4–7 s, until eye velocity matches visual stimulus velocity during steady-state OKN. These two components of OKN are thought to reflect activity in different neural circuits (see later). Compensatory eye speed during OKN saturates at about 180° s^{-1} in macaques and OKN gain is less than 1.0 for greater speeds. The initial rapid jump in eye speed can reach 40–80% of the maximum compensatory OKR and probably reflects contributions from reflex following and smooth pursuit. Eye velocity continues to increase following this initial jump in eye speed, to match visual stimulus velocity. The gradual buildup in eye velocity takes 10–15 s and is due to charging a so-called velocity storage mechanism that is shared between the optokinetic and vestibular ocular systems. This velocity storage system is essential for extending the effective time constant of the vestibular ocular response due to head acceleration or during constant-velocity head movement in the light.

The gain of OKN is affected by the state of the viewer, as demonstrated in testing human study participants. If viewers passively stare at the optokinetic stimulus, gain is less (0.85–0.9) than if they actively track individual stimulus contours, where gain approaches 1.0. It is thought the higher gain during active tracking reflects the addition of volitional smooth pursuit to optokinetically driven eye movements. Such modulation of OKN in frontal-eyed, foveate animals depends on cortical–brain stem pathways (see later).

Optokinetic Eye Movements and Short-Latency Ocular Following Response

As reviewed previously, the latency for eye motion onset with respect to visual motion onset during OKN testing is rather slow (100–150 ms). However, this rather long latency may not capture the actual functional capability of the optokinetic system. A short-latency (50 ms), ocular following response (OFR) can be produced by sudden movement of a large-field visual stimulus presented on a tangent screen in front of the viewer. The OFR stimulus simulates visual motion that occurs during head translation and the tVOR. OFR gain is modulated by viewing distance and retinal image disparity, appropriately supporting the tVOR in gaze stabilization.

The OFR response occurs with maximum gain and shortest latency (50 ms) when the step of visual motion is presented 10–50 ms following a centering saccade. The saccadic target is extinguished before saccade termination, preventing fixation. A major difference between OFR and OKN testing can be found in the method used for delivering the visual stimulus. Typically, OKN is produced by suddenly illuminating a rotating and striped cylinder already moving continuously around the viewer. This sudden onset of the visual surround activates the transient response mechanisms of middle temporal/medial superior temporal (MT/MST) neurons. Subsequently, specific DS mechanisms take over to produce an optokinetic eye movement at longer latency than observed during OFR.

The stereotyped nature of the OFR has proved to be helpful in studying the neural substrate for transformation of visual motion signals into commands for

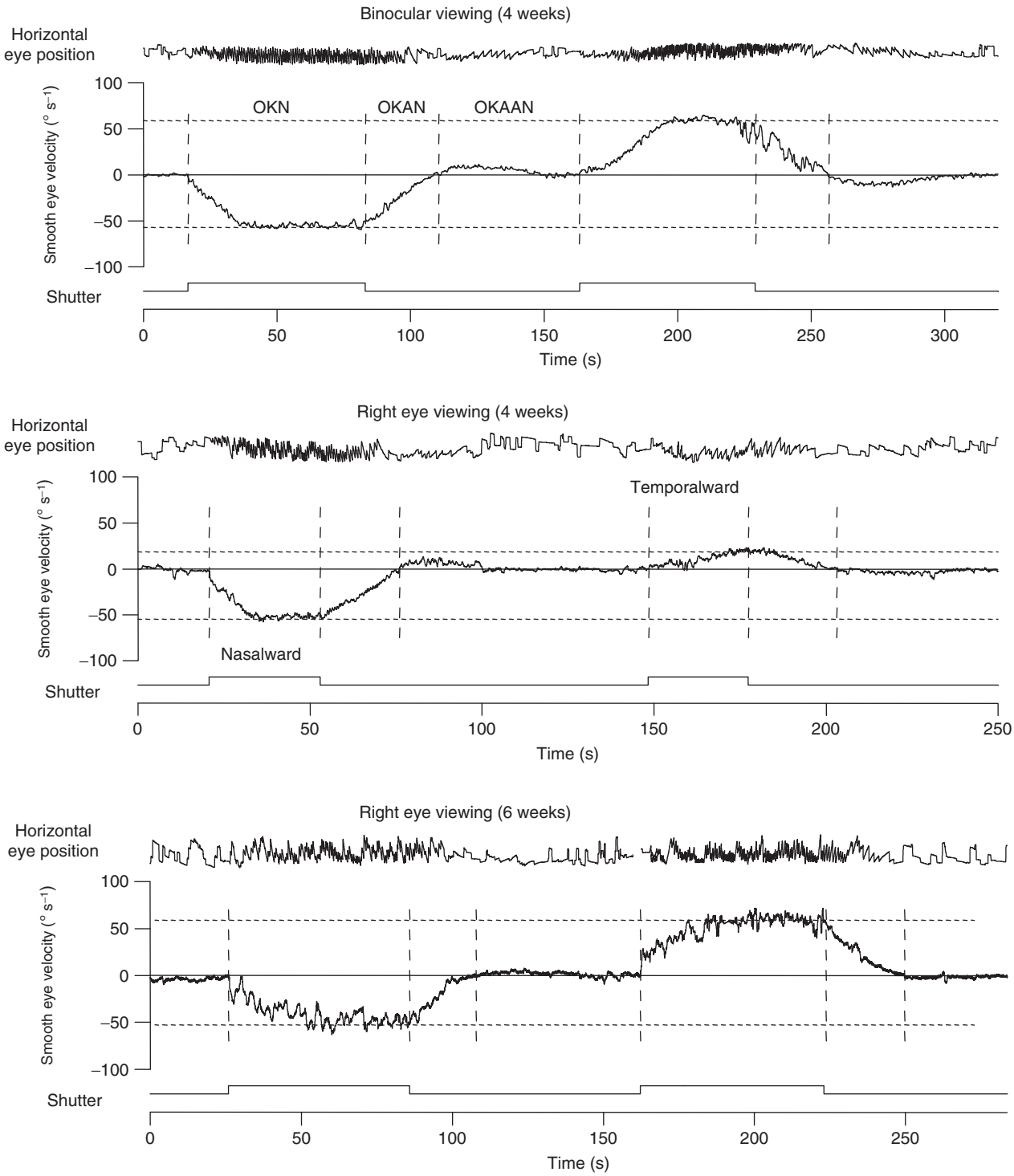

Figure 3 Immature pattern of horizontal OKN recorded from an infant monkey (*Macaca mulatta*) using the scleral search coil technique. Rightward and leftward visual motion ($55^\circ\,s^{-1}$) was presented sequentially on a cylindrical screen (57 cm distant), using a servo-controlled, planetarium projector. Upward deflection of the shutter trace indicates when the visual stimulus was on (OKN). Following visual stimulation, eye velocity continued as after-nystagmus (OKAN) and reversed direction as OKAAN. When OKAAN eye velocity approached $0^\circ\,s^{-1}$, visual motion was presented in the opposite direction. During binocular viewing (top panel) a 4-week-old monkey produces symmetrical OKN that gradually builds (no rapid rise) to match visual stimulus velocity. During monocular viewing, the 4-week-old monkey shows an immature pattern of OKN with strong nasalward and weak temporalward responses (middle panel). At 6 weeks of age (bottom panel), monocular OKN approaches a mature pattern, with nearly symmetric nasalward and temporalward responses. A rapid rise in eye velocity at the start of OKN is now evident. Eye traces from the top: horizontal eye position and desaccaded horizontal velocity. Horizontal dashed lines indicate the steady-state eye velocity reached during $55^\circ\,s^{-1}$ visual motion.

eye movements. Multivariate statistical and inverse dynamics modeling studies of neuronal response related to retinal image or eye motion (position, velocity, and acceleration) have demonstrated that MST, dorsolateral pontine nucleus (DLPN), and pretectal nucleus of the optic tract (NOT) neurons are best modeled using visual motion parameters. In contrast, ventral paraflocculus (vPF) neuronal response appears to code eye motion *per se*. Therefore, visual motion signals appear to be converted into motor commands for OFR in the vPF of the cerebellum.

Signals carried in the MT–DLPN–vPF pathway (Figure 4) play a role in OFR and smooth pursuit. However, volitional smooth pursuit certainly involves additional pathways, including the frontal eye fields, nucleus reticularis tegmenti pontis (NRTP), and cerebellar vermis. The MST–DLPN–vPF pathway has been considered to be part of the direct or early pathway for producing OFR and a rapid rise in eye velocity during OKN. Pathways traveling from the pretectal NOT and terminal nuclei of the AOS, including the lateral and medial terminal nuclei,

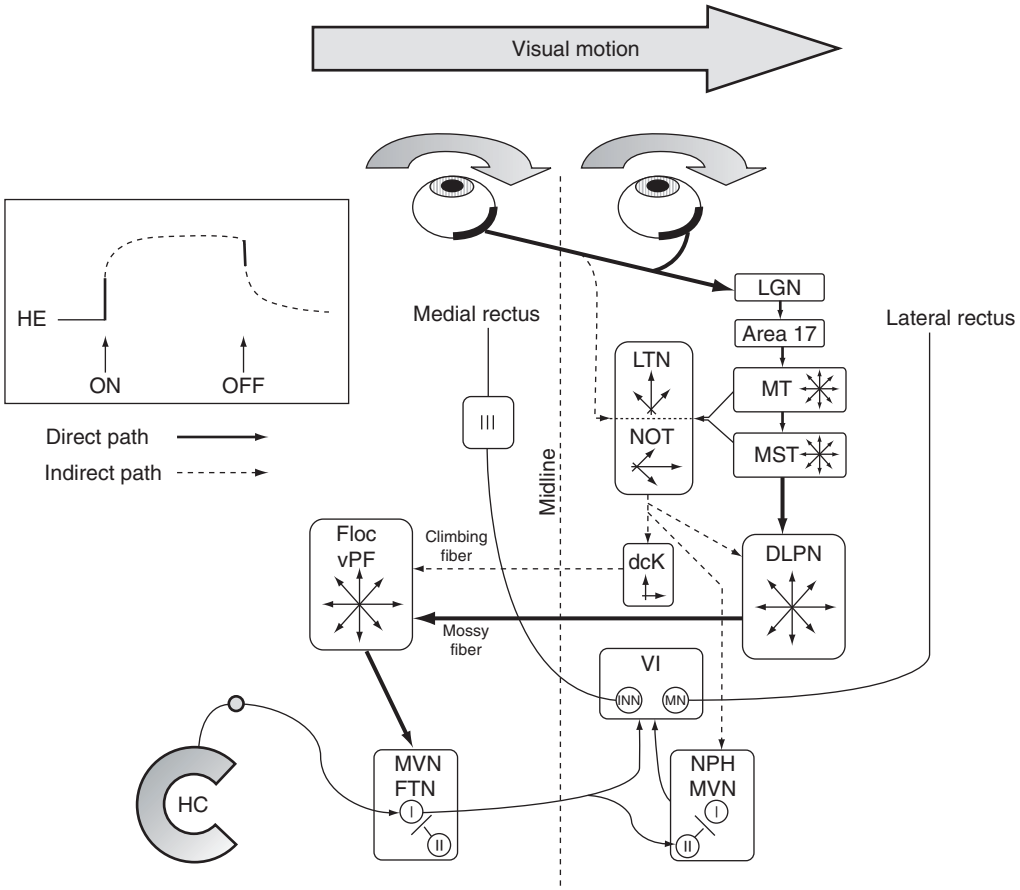

Figure 4 Simplified diagram indicating major sources for visual motion signals that play a role in optokinetic, ocular following, and smooth pursuit eye movements of primates. For example, rightward small- or large-field visual motion leads to rightward optokinetic following movements. Visual motion signal travel in the retinal AOS (LTN/NOT, lateral terminal nucleus/pretectal nucleus of the optic tract) emphasizes the contralateral eye (nasal retina). Visual motion signals also travel in retino-geniculo-striate-extrastriate pathways essential for visual motion processing. The extrastriate visual area (MT, middle temporal) provides binocular, high-speed visual motion information to the AOS/NOT. The distribution of DS visual motion signals is indicated for neurons in the MT, medial superior temporal visual area (MST), dorsolateral pontine nucleus (DLPN), LTN, NOT, and ventral paraflocculus (vPF). The right side of the diagram shows components of OKN, thought to be driven mostly by separate so-called direct and indirect pathways. Direct pathways (solid lines) may serve smooth pursuit and ocular following preferentially, while indirect pathways (dashed lines) have access to the velocity storage mechanism. Production of rightward horizontal optokinetic eye (HE) movements depends on activation of appropriate populations of horizontal (rightward) neurons and inactivation of neurons with other direction preferences. Such signals are found in the NOT/dorsal terminal nuclei-driven and in MT/MST-driven pathways. Signals in the NOT/AOS–dorsal cap of Kooy-inferior olive (dcK) pathway provide visual climbing fiber information that plays a role in visual modification of the VOR Signals in NOT–medial vestibular nucleus (MVN)/nucleus prepositus hypoglossi (NPH) pathways play a role in velocity-storage. Visual motion information in cortico-ponto-cerebellar pathways play a role in all components of the OKR, ocular following, and aspects of smooth pursuit. Floc, flocculus; FTN, flocculus target neurons; INN, internuclear neurons; MN, motor neurons; VI, abducens motor neurons; III, oculomotor motor neurons; LGN, lateral geniculate nucleus; HC, horizontal semicircular canal; type I and type II, neurons in the MVN.

provide major visual afferents for the indirect pathway related to velocity storage. The direct (early) and indirect (delayed) terminology has been helpful in highlighting differences in sources of visual motion information and efferent connections essential for production of reflexlike (OFR, OKR, and OKN) and foveal smooth pursuit (Figure 4).

Development of Optokinetic Eye Movements

The OKR is immature at birth in cats, monkeys, and humans and has some properties that resemble the afoveate pattern of lower mammals. In the adult monkey or human, steady-state OKN gain is symmetric during monocular viewing of a large-field stimulus, moving in either temporalward or nasalward directions. Furthermore, mature OKN in rhesus monkeys includes an initial rapid rise in eye velocity during OKN. OKN immaturity early in life includes a decreased steady-state OKN gain, with increasing visual stimulus velocity or spatial frequency. During the first few weeks of life the monocular OKR is asymmetric, with much higher gain during nasalward compared to temporalward visual motion. In fact, there is little to temporalward response in monkey OKN during the first 2–3 weeks of life.

Examination of Figure 3 shows that immature OKN lacks a rapid rise in eye velocity during OKN, but does show a strong velocity storage component. This is indicated by the gradual buildup in eye velocity during OKN and the presence of after-nystagmus (OKAN). This slow buildup in eye velocity is supported by a velocity storage mechanism, supported in part by activity in the medial vestibular nucleus/nucleus prepositus hypoglossi (MVN/NPH). A more mature pattern of OKN is obtainable by 6 weeks of age in the monkey (6 months in humans). Figure 3 shows the immature OKN response of a 4-week-old rhesus monkey. During monocular viewing of continuous visual motion, there is a strong nasalward response but the temporalward response is just beginning to develop. Another important aspect of immature OKN in macaques is the lack of an early rapid rise in eye velocity during OKN until about 6 weeks of age, when a symmetric monocular horizontal OKN appears.

NOT Binocular Sensitivity and OKN Development

The maturation of the early asymmetric OKR to the mature symmetric pattern is thought to depend on the development of cortical and NOT binocular sensitivity. The best supported hypothesis is that the NOT receives direct retinal input from the contralateral nasal retina at birth. This projection is likely to arise, at least in part, from DS ganglion cells in the gamma class. Postnatal visual experience is necessary to allow development of cortical–NOT pathways representing the ipsilateral eye.

DS retinal ganglion cells could play an important role in cortical–NOT/AOS pathway formation by reinforcing development of cortical neurons with appropriate direction selectivity. For example, even though all directions of visual motion are represented in the MT, only those with ipsiversive direction preference project to the NOT of mature primates (Figure 5). Furthermore, this projection is absent in the first weeks of life, when monocular OKN is still asymmetric.

If coordinated binocular vision is impaired during early life – as occurs in strabismus, anisometropia, or a congenital cataract in one eye (monocular deprivation) – cortical–NOT connections representing the ipsilateral eye may fail to develop, leading to OKN asymmetry. The common underlying deficit is most likely loss of binocularly sensitive neurons in cortex, failure to develop cortical–NOT connections that provide representation for the ipsilateral eye, leading to loss of NOT binocular sensitivity. In addition, cortical–NOT pathways are essential for providing NOT neurons with high-velocity sensitivity, which supports OKN to speeds well over $100^{\circ}\,s^{-1}$. Thus, any disruption of early binocular visual experience might result in the development of asymmetric OKN. This points out the importance of a sensitive period during the first 6 weeks of life in monkeys (first 6 months in humans), when the visual–oculomotor system is most vulnerable to visual experience. The critical period for other visual–oculomotor behaviors may have different time courses.

NOT Direction Selectivity

Decreased binocular interactions early in life disrupt the maturation of cortical–NOT circuits, leaving the NOT driven either exclusively or mostly by the contralateral eye. Such contralateral ocular dominance could help to explain asymmetric, monocular OKN.

Directional mechanisms are certainly important in the development of symmetric monocular OKN and smooth pursuit. Visual motion in the nasalward direction, which produces strong OKN at birth, may initially rely on direct retinal–NOT projections. Such projections have been demonstrated to emphasize the nasal retina, which develops earlier than the temporal retina in primates. Ipsilateral projecting retinal afferents may require longer postnatal development times

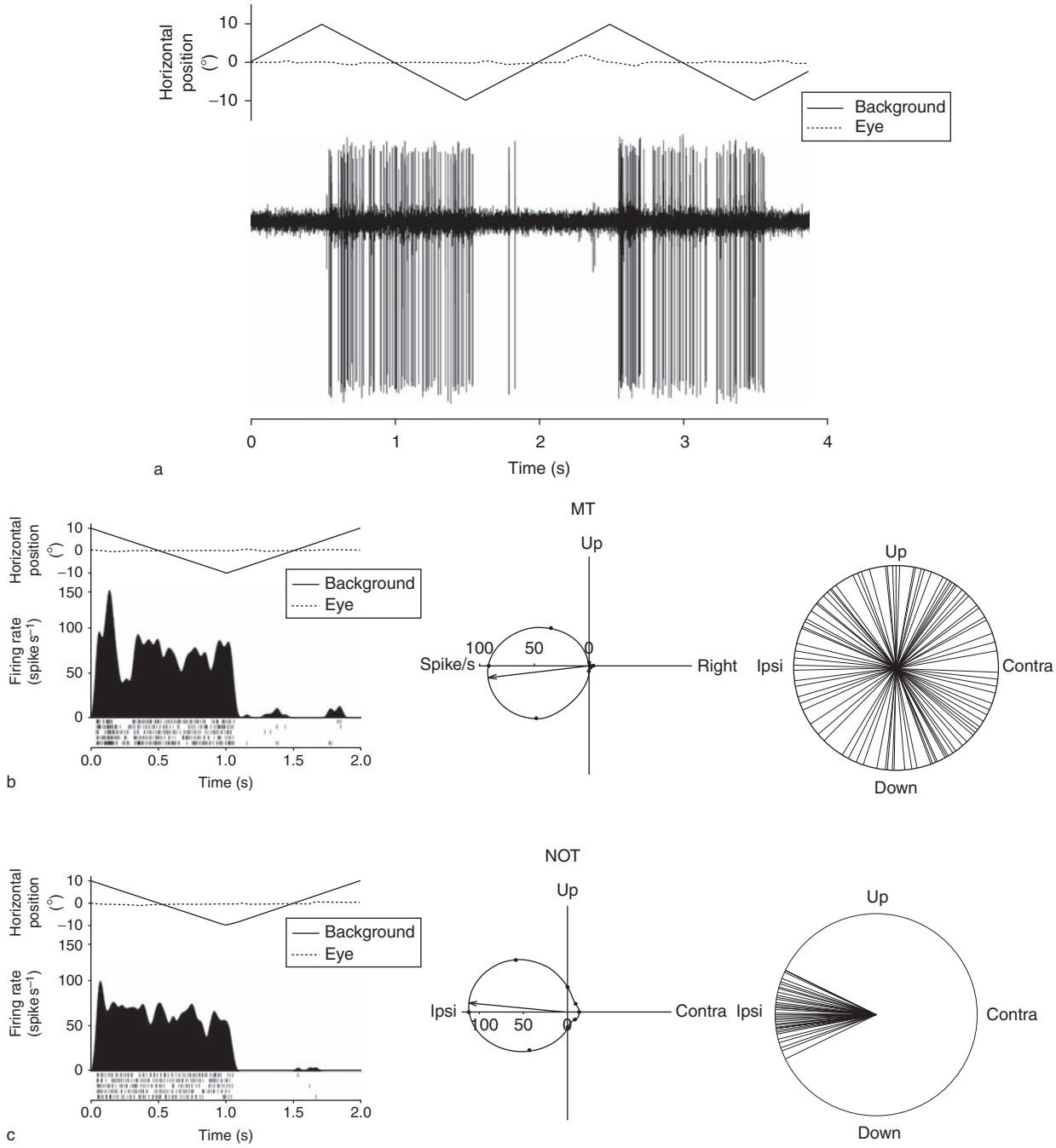

Figure 5 Response properties of a representative primate middle temporal (MT) neuron during visual motion testing. (a) The neuronal response during horizontal large-field visual motion presented on a tangent screen while the monkey (*Macaca mulatta*) fixated a stationary target spot. A strong DS response occurs at short latency (50 ms) after the start of leftward visual motion. (b) Left: rasters for successive trials aligned on the start of leftward visual motion and the average firing rate represented as a spike-density function ($\sigma = 10$ ms) for five cycles of horizontal visual motion. Middle: directional tuning data for a representative MT neuron. Large-field visual motion was presented in different directions while the monkeys fixated a stationary target. The average firing rate is indicated at 45° intervals (filled circles), with a curve fit to estimate the direction preference for the unit. The point where the fit curve reached a maximum was taken as the preferred direction for the neuron (arrow). Right: calculated preferred directions for each of 100 MT neurons. MT neurons with ipsiversive direction preference provide visual motion information to the nucleus of the optic tract (NOT). (c) The ipsiversive (leftward) direction response for a NOT unit (left), the direction tuning curve for this neuron (middle), and the calculated preferred directions for a representative population of 50 NOT neurons (right).

than do their counterparts. For example, studies in the kitten reveal that until 3 weeks of age, stimulation delivered to the contralateral eye is effective in driving NOT units, but stimulation of the ipsilateral eye is ineffective, even until about 7 weeks of age, when ipsilateral eye stimulation becomes more effective. There is evidence that sensitivity to the ipsilateral eye is cortically derived, rather than being derived from direct retinal inputs from the ipsilateral eye, because sensitivity to the ipsilateral eye appears at the same time that NOT cells can be orthodromically activated by electrical stimulation of area 17 and area 18 in the cat. It seems likely that a similar pattern of development could be at work in the primates.

Neural Substrate for Optokinetic Eye Movements: AOS

Converging anatomical, single-unit, electrical stimulation and lesion evidence points to the AOS and pretectal NOT as essential sources of DS visual information for optokinetic eye movements and visual vestibular interaction.

The AOS includes DS retinal ganglion cells and their targets the lateral, medial, and dorsal terminal nuclei (LTN, MTN, and DTN, respectively), distributed along the mesencephalic brain stem. All of these nuclei receive a direct input, mostly from the contralateral retina, and indirect inputs from ipsilateral striate and extrastriate visual cortex (Figure 4). Early studies in rabbits identified DS retinal ganglion cells that project to the terminal nuclei; DS ganglion cells are also found in cats and monkeys. Neurons in the LTN and MTN respond preferentially to vertical visual motion while those in the DTN respond optimally for visual motion toward the side of the specific nucleus (ipsiversive).

In lower mammals, the terminal nuclei are easily separated from surrounding structures. In the primate, the LTN is the only consistently demonstrable terminal nucleus, although all divisions of the AOS are probably present. Primate LTN neurons are sensitive to upward visual motion. The DTN is sensitive to horizontal large-field visual motion.

The terminal nuclei and the NOT (see later) provide the only known source of visual motion information to the dorsal cap of Kooy-inferior olive (dcK), thus providing the afferent signals for generation of visual complex spikes in cerebellar Purkinje cells. This activity is thought to play a role in visually guided motor learning in the VOR, especially during high-frequency visual–vestibular interactions. The terminal nuclei and the NOT are reciprocally connected by γ-aminobutyric acid (GABA)ergic neurons, perhaps to improve the efficacy of directional drive for the OKR. Neurons in the terminal nuclei are most sensitive to low visual motion speed ($<5^\circ\,s^{-1}$), especially in lower mammals.

The NOT is often included as a member of the AOS because it has similar anatomical connections and some of its units behave like those of the DTN. However, neurons in the NOT of cats and monkeys often have much wider speed-tuning ranges than do those in the terminal nuclei. In the primate NOT, neurons are sensitive to either large-field or foveal/parafoveal visual motion. These neurons may reflect properties of MT neurons with ipsiversive direction selectivity, which have been shown to project to the NOT. Small-field visual neurons have not been reported in the NOT of lower mammals.

The NOT and AOS can influence smooth eye movements of foveal, optokinetic, or vestibular origin by at least three possible pathways (Figure 4). First, both the NOT and terminal nuclei of the AOS project to the dcK, which in turn provides the sole source of visual climbing fibers to eye movement-associated regions of the cerebellum, such as flocculus, vPF, and vermis. This pathway could play a role in making gradual plastic changes in slow eye movements (e.g., VOR and smooth pursuit) in response to development, aging, and disease. However, it seems unlikely that the NOT projection to the dcK is well suited to support the contribution to the initiation of OKN, ocular following, or smooth pursuit. The low firing rates (<10 spikes s^{-1}) of dcK neurons are consistent with a role related to modulation of ongoing simple spike activity.

Purkinje cells of the oculomotor cerebellum provide essential commands for all components of OKN, OFR, and smooth pursuit. The contraversive visual motion sensitivity of Purkinje cell complex spikes could be derived from DTN and NOT inputs to the dcK (Figure 4). Second, the NOT projects to the NRTP and to the DLPN, which provides visual mossy fiber inputs to the oculomotor cerebellum. These inputs could effect moment-by-moment changes in smooth eye movements of OKN, OFR, or smooth pursuit. Neurons in the NRTP and DLPN of primates discharge during either large- or small-field visual motion over a broad range of speeds and directions. Finally, neurons in both the NOT (and AOS nuclei) project to the NPH and MVN. The NPH is known to contain neurons that participate in all types of horizontal eye movements. In addition, the NPH and MVN are essential for velocity-storage during OKN. Lesions along either direct or indirect pathways produce impairments in different aspects of optokinetic eye movements.

Conclusions

The optokinetic system plays an essential role in gaze stabilization responses during head rotation and

translation. The optimal stimulus for production of an OKR is large-field visual motion, although small-field visual motion may contribute to foveally based tracking systems such as smooth pursuit.

Optokinetic eye movements play a major role in visual–vestibular interactions for gaze stabilization during locomotion and for adjusting the gain of the VOR in response to developmental, age-related, or pathological challenges. The short-latency OFR shares much of the same circuitry as the OKR and may be particularly important in supporting the tVOR.

The OKR of frontal-eyed species (e.g., cats and primates) is strongly influenced by descending visual cortical inputs. These inputs extend the operating range of optokinetic eye movements to include compensation for high-speed visual motion.

The optokinetic system is immature at birth in frontal-eyed animals (e.g., cats and primates) and presents with an OKN pattern similar to that observed in lateral-eyed, afoveate mammals, where temporal–nasal directions of visual motion are most effective. Maturation of the OKR is associated with development of cortical binocularity and motion sensitivity. Appropriate postnatal visual experience is required for this maturation and development of descending cortical connections with the AOS/NOT. Development of cortical binocularity takes place along with development of symmetrical, monocular OKR and improved sensitivity to higher visual motion speeds.

The major visual afferent pathways supporting the OKR comprise both direct retinal and indirect cortical (V1, V2, and MT) inputs to the AOS/NOT. Interconnections between nuclei of the AOS/NOT are mostly GABAergic and could serve to improve directional OKRs. Efferent connections of the AOS/NOT to the dorsal cap of the inferior olive and basilar pontine nuclei (DLPN and NRTP) provide visual climbing and mossy fiber inputs, respectively, to the oculomotor and vestibular cerebellum. Direct AOS/NOT projections to the MVN and NPH may play a role in velocity storage. The optokinetic system plays an essential role in gaze stabilization, visual adjustments to posture, volitional smooth eye movements, and visually guided motor learning.

See also: Cerebellum and Oculomotor Control; Eye and Head Movements; Pursuit Eye Movements; Vestibular System; Vestibulo-Ocular Reflex; Visual Motion Models; Visual–Vestibular Interactions.

Further Reading

Andersen RA, Snowden RJ, Treue S, et al. (1990) Hierarchical processing of motion in visual cortex. *Cold Spring Harbor Symposium* 55: 741–748.

Bremmer F, Duhamel JR, Ben Hamed S, et al. (2000) Stages of self-motion processing in primate posterior parietal cortex. *International Review of Neurobiology* 44: 173–198.

Büttner U and Büttner-Ennever JA (2006) Present concepts of oculomotor organization. *Progress in Brain Research* 151: 1–42.

Distler C, Hoffmann K-P, and Mustari MJ (2002) Origins of cortical projections to the NOT-DTN and DLPN in macaques. *Journal of Comparative Neurology* 444: 144–158.

du Lac S, Raymond JL, Sejnowski TJ, et al. (1995) Learning and memory in the vestibulo-ocular reflex. *Annual Review of Neuroscience* 18: 409–441.

Fuchs AF and Mustari MJ (1993) The optokinetic response in primates and its possible neuronal substrate. In: Miles FA and Wallman J (eds.) *Reviews of Oculomotor Research, Volume 5. Visual Motion and Its Role in the Stabilization of Gaze*, pp. 343–369. Amsterdam: Elsevier.

Hoffmann K-P, Distler C, and Ilg U (1992) Callosal and superior temporal sulcus contributions to receptive field properties in the macaques monkey's NOT and DTN. *Journal of Comparative Neurology* 321: 150–162.

Inoue Y, Takemura A, Kawano K, et al. (2000) Role of the pretectal nucleus of the optic tract in short-latency ocular following responses in monkeys. *Experimental Brain Research* 131: 269–281.

Kawano K, Shidara M, Watanabe Y, et al. (1994) Neural activity in cortical area MST of alert monkey during ocular following responses. *Journal of Neurophysiology* 71: 2305–2324.

Kawato M and Gomi H (1992) The cerebellum and VOR/OKR learning models. *Trends in Neurosciences* 15: 445–453.

Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn., New York: Oxford University Press.

Miles FA, Kawano K, and Optican LM (1986) Short-latency ocular following responses of monkey. I: Dependence on temporospatial properties of visual input. *Journal of Neurophysiology* 56: 1321–1354.

Mustari MJ and Fuchs AF (1990) Discharge patterns of neurons in the pretectal nucleus of the optic tract (NOT) in the behaving primate. *Journal of Neurophysiology* 64: 77–79.

Mustari MJ, Fuchs AF, Kaneko CRS, et al. (1994) Anatomical connections of the primate pretectal nucleus of the optic tract. *Journal of Comparative Neurology* 349: 111–128.

Raymond J and Lisberger SG (1996) Behavioral analysis of signals that guide learned changes in the amplitude and dynamics of the VOR. *Journal of Neuroscience* 16: 791–7802.

Wallman J (1993) Subcortical optokinetic mechanisms. In: Miles FA and Wallman J (eds.) *Reviews of Oculomotor Research, Volume 5. Visual Motion and Its Role in the Stabilization of Gaze*, pp. 321–342. Amsterdam: Elsevier.

Orbitofrontal Cortex: Visual Functions

E T Rolls, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Visual Inputs to the Orbitofrontal Cortex, and Visual Stimulus–Reinforcer Association Learning and Reversal

Rolls and colleagues have shown that there is a major visual input to many neurons in the orbitofrontal cortex, and that what is represented by these neurons is in many cases the reinforcer (reward or punisher) association of visual stimuli. Many of these neurons reflect the relative preference or reward value of different visual stimuli, in that their responses decrease to zero to the sight of one food on which the monkey is being fed to satiety, but remain unchanged to the sight of other food stimuli. In this sense the visual reinforcement-related neurons predict the reward value that is available from the primary reinforcer, the taste. The visual input is from the ventral, temporal lobe, visual stream concerned with 'what' object is being seen, in that orbitofrontal visual neurons frequently respond differentially to objects or images (but depending on their reward association). The primary reinforcer that has been used is taste.

The fact that these neurons represent the reinforcer associations of visual stimuli has been shown to be the case in formal investigations of the activity of orbitofrontal cortex visual neurons, which in many cases reverse their responses to visual stimuli when the taste with which the visual stimulus is associated is reversed by the experimenter. An example of the responses of an orbitofrontal cortex neuron that reversed the visual stimulus to which it responded during reward reversal is shown in Figure 1. This reversal by orbitofrontal visual neurons can be very fast, in as little as one trial – that is, a few seconds.

These neurons thus reflect the information about which stimulus is currently associated with reward during reversals of visual discrimination tasks – they are reward-predicting neurons. If a reversal occurs, then taste cells in the orbitofrontal cortex provide the information that an unexpected taste reinforcer has been obtained, another group of cells shows a vigorous discharge that reflects the error between the expected reward and the reward actually obtained (see later), and the visual cells with reinforcer association-related responses reverse the stimulus to which they are responsive. These neurophysiological changes take place rapidly, in as little as 5 s, and are presumed to be part of the neuronal learning mechanism that enables primates to alter their knowledge of the reinforcer association of visual stimuli so rapidly. This capacity is important whenever behavior must be corrected when expected reinforcers are not obtained, in, for example, feeding, emotional, and social situations. In that these neurons reflect whether a visual stimulus is associated with a reward or a punisher, they reflect the relative preference for different stimuli. Consistent with this evidence that the responses of some orbitofrontal cortex neurons reflect the learned predictive reward value of visual stimuli, Thorpe and colleagues found that orbitofrontal cortex neurons learned to respond differently to new stimuli that did or did not predict reward. Different neurons in the orbitofrontal cortex are tuned to different learned or conditioned reinforcers, with, for example, approximately 5% responding to visual stimuli associated with taste reward, and 3% to visual stimuli associated with taste punishment.

In the visual discrimination reversal task, a second class of neuron was found that codes for particular stimuli only if they are associated with reward, and not if they are associated with punishment. Such a neuron might respond to a green stimulus associated with reward, after reversal not respond to the green stimulus when it was associated with punishment, and not respond to a blue stimulus irrespective of whether it was associated with reward or punishment. An example is shown in Figure 2. These neurons may be described as conditional visual stimulus-to-taste reward neurons, and may be important in a process by which the rule for which stimulus is treated as being currently rewarded can be reversed in one trial.

This reversal learning found in orbitofrontal cortex neurons probably is implemented in the orbitofrontal cortex, for it does not occur one synapse earlier in the visual inferior temporal cortex, and it is in the orbitofrontal cortex that there is convergence of visual and taste pathways onto the same neurons. The way in which the ventral visual stream builds position, size, and even view invariant representations of objects, as described in *Computational Neuroscience of Vision* by Rolls and Deco, is very important, for it enables brain regions (such as the orbitofrontal cortex) which learn about the associations of objects to generalize to different positions, sizes, and views of the same object when they are seen later.

A possible mechanism for visual-reinforcer learning in the orbitofrontal cortex is associative (Hebbian) modification of synapses conveying visual input onto taste-responsive neurons, implementing a pattern-association network. In this model the unconditioned stimulus forcing the output neurons to respond is the

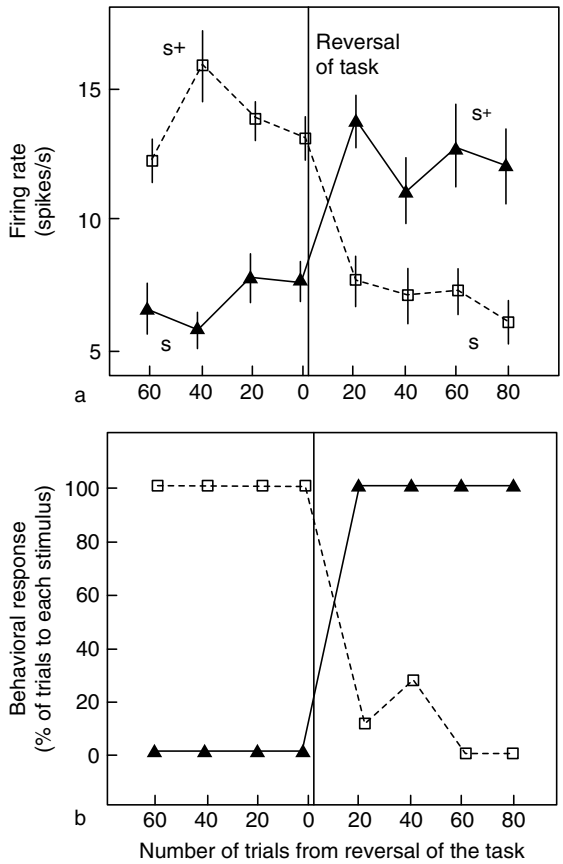

Figure 1 Orbitofrontal cortex: visual discrimination reversal. The activity of an orbitofrontal visual neuron during performance of a visual discrimination task and its reversal. The stimuli were a triangle (▲) and a square (□) presented on a video monitor. (a) Each point represents the mean poststimulus activity in a 500-ms period of the neuron based on approximately 10 trials of the different visual stimuli. The standard errors of these responses are shown. After 60 trials of the task the reward associations of the visual stimuli were reversed. (+ indicates that a lick response to that visual stimulus produces fruit juice reward; - - - indicates that a lick response to that visual stimulus results in a small drop of aversive-tasting saline). This neuron reversed its responses to the visual stimuli following the task reversal. (b) The behavioral response of the monkey to the task. It is shown that the monkey performs well, in that he rapidly learns to lick only in response to the visual stimulus associated with fruit juice reward. Adapted Rolls ET, Critchley HD, Mason R, et al. (1996) Orbitofrontal cortex neurons: Role in olfactory and visual association learning. *Journal of Neurophysiology* 75: 1970–1981.

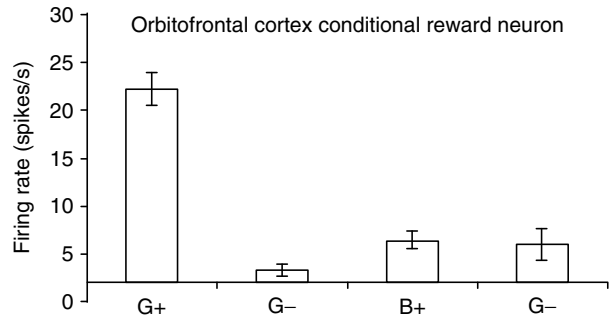

Figure 2 A conditional reward neuron recorded in the orbitofrontal cortex which responded only to the green stimulus when it was associated with reward (G+), and not to the blue stimulus when it was associated with reward (B+), or to either stimuli when they were associated with a punisher, the taste of salt (G− and B−). The mean firing rate +/− the SEM is shown. Data from Thorpe SJ, Rolls ET, and Maddison S (1983) Neuronal activity in the orbitofrontal cortex of the behaving monkey. *Experimental Brain Research* 49: 93–115.

(taste) primary reinforcer, and the visual conditioned stimulus becomes associated with this by associatively modifiable synapses. Such a pattern association network could in principle unlearn the association by using associative synapses that incorporate long-term depression (LTD). Although reversal might be implemented by having long-term synaptic depression for synapses that represented the reward-associated stimulus before the reversal, and long-term potentiation (LTP) of the synapses activated by the new stimulus that after reversal is associated with reward, this would require one-trial LTP and one-trial heterosynaptic LTD to account for one-trial stimulus–reward reversal. To implement the reversal learning very rapidly, in as little as one trial after a number of reversals when a reversal learning set has been acquired, a special rule-switching network in the orbitofrontal cortex may be required, and a model of this has been described (e.g., in *Emotion Explained*, by ET Rolls).

Studies in macaques provide evidence on the details of the neuronal representations in the primate orbitofrontal cortex which are essential for building a computational understanding for exactly what information is represented in it and how it is represented, how these representations differ from preceding and succeeding stages, and thus how the orbitofrontal cortex operates computationally. However, the types of visual–reward associations that have been studied in primates (and confirmed as applying in humans) include objects associated with taste rewards or punishers. It has therefore been useful not only to confirm that these concepts do indeed apply to humans, but also to extend the types of visual conditioned reinforcers to quite abstract reinforcers such as monetary reward. In a functional magnetic resonance imaging (fMRI) study, O'Doherty and colleagues used a visual discrimination task in which one stimulus was associated with monetary reward, and a different visual stimulus with monetary loss (punishment). The actual amounts of money won on reward trials and lost on punishment trials were probabilistic. This part of the design, and the fact that unexpected visual discrimination reversals occurred so that there were trials on

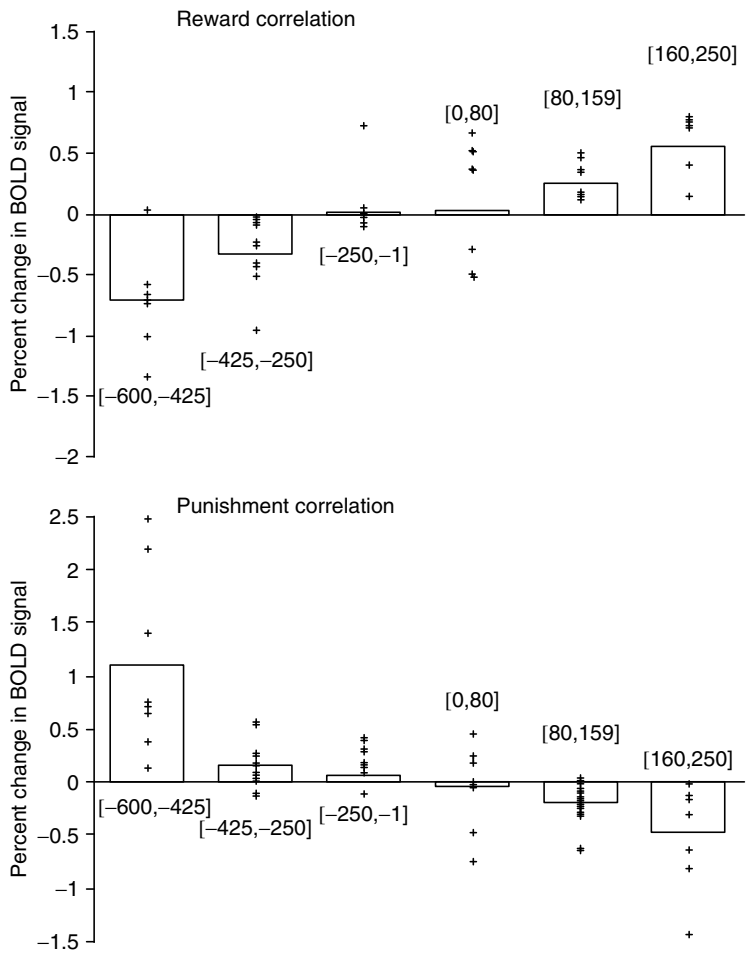

Figure 3 Correlation of brain activations with the amount of money won or lost in a visual discrimination reversal task with probabilistic monetary reward and loss, showing the mean percent change in blood oxygen-level-dependent (BOLD) signal from baseline across individuals for six different category ranges of reward and punishment. The signal was averaged across a category range within each individual and then the average signal change from each category was averaged across individuals. This is plotted on the right for voxels in the medial orbitofrontal cortex that significantly correlated with reward, and on the left for voxels in the lateral orbitofrontal cortex that significantly correlated with punishment. The ranges of monetary reward and punishment in each category are shown on the chart and were determined by their relative frequencies, which follow from the experimental design. (Large positive values are large wins; large negative values are large losses.) Adapted from O'Doherty J, Kringelbach ML, Rolls ET, et al. (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience* 4: 95–102.

which money was lost, showed that the magnitude of the activation of the medial orbitofrontal cortex was correlated with the amount of money won on each trial, and the magnitude of the activation of the lateral orbitofrontal cortex was correlated with the amount of money lost on each trial, as shown in Figure 3.

Another way in which it has been shown that the visual neurons in the orbitofrontal cortex reflect the expected reward value predicted by visual stimuli is by reducing the reward value by feeding to satiety. With this sensory-specific satiety (or reward devaluation) paradigm, it has been shown that the visual (as well as the olfactory and taste) responses of orbitofrontal cortex neurons in the macaque decrease to zero as the monkey is fed to satiety with one food, but remain unchanged in response to another food not eaten in the meal. In that these neurons parallel the changing preference of the monkey for the food being eaten to satiety versus the food not being eaten to satiety, they reflect the relative preference for different visual stimuli.

Error Neurons in the Orbitofrontal Cortex, and Visual Stimulus–Reinforcer Association Learning and Reversal

In addition to the neurons that encode the reward association of visual stimuli, other neurons (3.5%)

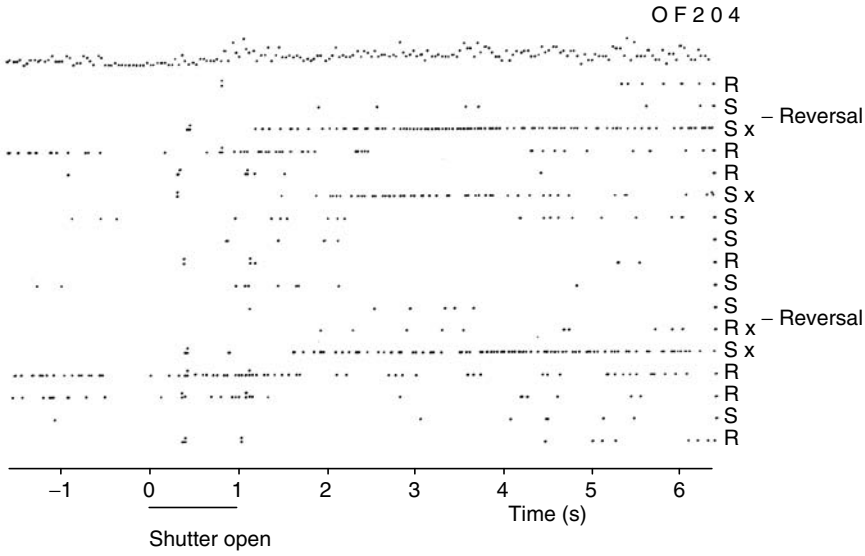

Figure 4 Error neuron. Responses of an orbitofrontal cortex neuron that responded only when the monkey licked in response to a visual stimulus during reversal, expecting to obtain fruit juice reward, but actually obtaining the taste of aversive saline, because it was the first trial of reversal. Each single dot represents an action potential; each vertically arranged double dot represents a lick response. The visual stimulus was shown at time 0 for 1 s. The neuron did not respond on most reward (R) or saline (S) trials, but did respond on the trials marked x, which were the first trials after a reversal of the visual discrimination on which the monkey licked to obtain reward, but actually obtained saline, because the task had been reversed. It is notable that after an expected reward was not obtained due to a reversal contingency being applied, on the very next trial the macaque selected the previously nonrewarded stimulus. This shows that rapid reversal can be performed by a nonassociative process, and must be rule based. A model for this has been described by G Deco and ET Rolls. Adapted Thorpe SJ, Rolls ET, and Maddison S (1983) Neuronal activity in the orbitofrontal cortex of the behaving monkey. *Experimental Brain Research* 49: 93–115.

in the orbitofrontal cortex detect different types of nonreward. For example, some neurons respond in extinction, immediately after a lick has been made in response to a visual stimulus that has previously been associated with fruit juice reward, and other neurons respond in a reversal task, immediately after the monkey has responded to the previously rewarded visual stimulus, but has obtained the punisher of salt taste rather than reward (see example in Figure 4).

Different populations of such neurons respond to other types of nonreward, including the removal of a formerly approaching taste reward, and the termination of a taste reward. The fact that different nonreward neurons respond to different types of nonreward (e.g., some to the noise of a switch that indicated that extinction of free licking for fruit juice had occurred, and others to the first presentation of a visual stimulus that was not followed by reward in a visual discrimination task) potentially enables context-specific extinction or reversal to occur. (For example, the fact that fruit juice is no longer available from a lick tube does not necessarily mean that a visual discrimination task will be operating without reward.) Thus the error neurons can be specific to different tasks, and this could provide a mechanism for reversal in one task to be implemented, while at the same time not reversing behavior in another task.

The presence of these neurons is fully consistent with the hypothesis that they are part of the mechanism by which the orbitofrontal cortex enables very rapid reversal of behavior by stimulus–reinforcer association relearning when the association of stimuli with reinforcers is altered or reversed. This information appears to be necessary for primates to rapidly alter behavioral responses when reinforcement contingencies are changed, as shown by the effects of damage to the orbitofrontal cortex. To the extent that the firing of some dopamine neurons may reflect error signals, one might ask where the error information comes from, given that the dopamine neurons themselves may not receive information about expected rewards (e.g., a visual stimulus associated with the sight of food) and obtained rewards (e.g., taste), and would have to compute an error from these signals. On the other hand, the orbitofrontal cortex does have all three types of neurons and the required neuroanatomically defined inputs, and this is an important site in the brain for computing error signals.

In that most neurons in the macaque orbitofrontal cortex respond to reinforcers and punishers, or to

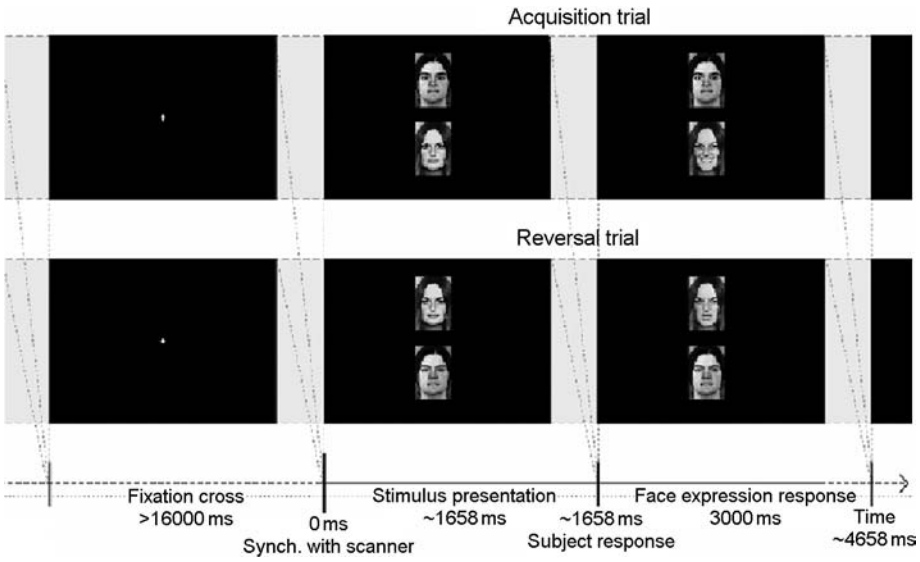

Figure 5 Social reversal task. The trial starts synchronized with the scanner, and two people with neutral face expressions are presented to the individual. The individual has to select one of the people by pressing the corresponding button, and the person will then either smile or show an angry face expression for 3000 ms, depending on the current mood of the person. The task for the individual is to keep track of the mood of each person and choose the 'happy' person as much as possible (upper row). Over time (after between 4 and 8 correct trials) this will change so that the 'happy' person becomes 'angry' and vice versa, and the individual has to learn to adapt her choices accordingly (bottom row). Randomly intermixed trials with either two men, or two women, were used to control for possible gender and identification effects, and a fixation cross was presented between trials for at least 16 000 ms. Adapted Kringelbach ML and Rolls ET (2003) Neural correlates of rapid reversal learning in a simple model of human social interaction. *NeuroImage* 20: 1371–1383.

stimuli associated with rewards and punishers, and do not respond in relation to responses, the orbitofrontal cortex is closely related to stimulus processing, including the stimuli that give rise to affective states. When it computes errors, it computes mismatches between stimuli that are expected, and stimuli that are obtained, and in this sense the errors are closely related to those required to correct affective states. This type of error representation may thus be different from that represented in the cingulate cortex, in which behavioral responses are represented, where the errors may be more closely related to errors that arise when action–outcome expectations are not met, and where action–outcome rather than stimulus–reinforcer representations need to be corrected.

We have also been able to obtain evidence that nonreward used as a signal to reverse behavioral choice to visual stimuli is represented in the human orbitofrontal cortex. Kringelbach and Rolls used the faces of two different people, and if one face was selected then that face smiled, and if the other was selected, the face showed an angry expression. After good performance was acquired, there were repeated reversals of the visual discrimination task (see Figure 5). It was found that activation of a lateral part of the orbitofrontal cortex in the fMRI study was produced on the error trials – that is, when the human chose a face, and did not obtain the expected reward (Figure 6). Control tasks showed that the response was related to the error, and the mismatch between what was expected and what was obtained, in that just showing an angry face expression did not selectively activate this part of the lateral orbitofrontal cortex. An interesting aspect of this study that makes it relevant to human social behavior is that the conditioned stimuli were faces of particular individuals, and the unconditioned stimuli were face expressions. Moreover, the study reveals that the human orbitofrontal cortex is very sensitive to social feedback when it must be used to change behavior.

A Visual Representation of Faces in the Orbitofrontal Cortex

Another type of information represented in the orbitofrontal cortex is information about faces. There is a population of orbitofrontal cortex face-selective neurons that respond in many ways similarly to those in the temporal cortical visual areas. The orbitofrontal face-responsive neurons, first observed by Thorpe and colleagues in 2003, then by Rolls et al. in 2006, tend to respond with longer latencies than do temporal lobe neurons (140–200 ms typically, compared with

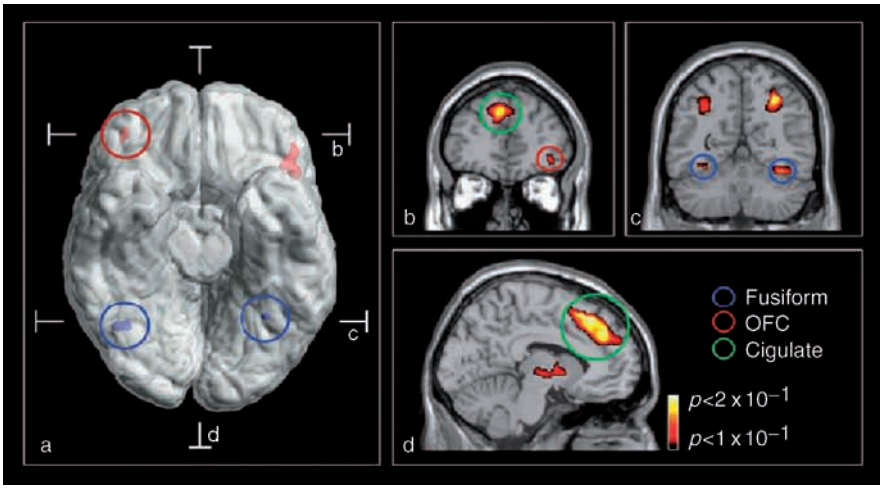

Figure 6 Social reversal. Composite figure showing that changing behavior based on face expression is correlated with increased brain activity in the human orbitofrontal cortex. (a) The figure is based on two different group statistical contrasts from the neuroimaging data, which are superimposed on a ventral view of the human brain with the cerebellum removed, and with indication of the location of the two coronal slices (b, c) and the transverse slice (d). The red activations in the orbitofrontal cortex (OFC, red circle; maximal activation: $z = 4.94$: 42, 42, −8; and $z = 5.51$; X, Y, $Z = -46$, 30, −8) shown on the rendered brain arise from a comparison of reversal events with stable acquisition events, while the blue activations in the fusiform gyrus (blue circles; maximal activation: $z > 8$; 36, −60, −20 and $z = 7.80$; −30, −56, −16) arise from the main effects of face expression. (b) The coronal slice through the frontal part of the brain shows the cluster in the right orbitofrontal cortex across all nine individuals when comparing reversal events with stable acquisition events. Significant activity was also seen in an extended area of the anterior cingulate/paracingulate cortex (green circle; maximal activation: $z = 6.88$; −8, 22, 52). (c) The coronal slice through the posterior part of the brain shows the brain response to the main effects of face expression, with significant activation in the fusiform gyrus and the cortex in the intraparietal sulcus (maximal activation: $z > 8$; 32, −60, 46; and $z > 8$: −32, −60, 44). (d) The transverse slice shows the extent of the activation in the anterior cingulate/paracingulate cortex when comparing reversal events with stable acquisition events. Group statistical results are superimposed on a ventral view of the human brain with the cerebellum removed, and on coronal and transverse slices of the same template brain (activations are thresholded at $p = 0.0001$ for purposes of illustration to show their extent). Adapted Kringelbach ML and Rolls ET (2003) Neural correlates of rapid reversal learning in a simple model of human social interaction. *NeuroImage* 20: 1371–1383.

80–100 ms); they also convey information about which face is being seen, by having different responses to different faces (see Figure 7), and are typically rather harder to activate strongly compared to temporal cortical face-selective neurons, in that many of them respond much better to real faces than to two-dimensional images of faces on a video monitor. Some of the orbitofrontal cortex face-selective neurons are responsive to face gesture or movement. The findings are consistent with the likelihood that these neurons are activated via the inputs from the temporal cortical visual areas in which face-selective neurons are found. The significance of the neurons is likely to be related to the fact that faces convey information that is important in social reinforcement, both by conveying face expression, which can indicate reinforcement, and by encoding information about which individual is present, also important in evaluating and utilizing reinforcing inputs in social situations.

Consistent with these findings in macaques and, as previously described, in humans, activation of the lateral orbitofrontal cortex occurs when a rewarding smile expression is expected, but an angry face expression is obtained, in a visual discrimination reversal task. This is an example of the operation of a social reinforcer, and, consistent with these results, activation of the orbitofrontal cortex is found when humans are making social judgments. In addition, activation of the medial orbitofrontal cortex is correlated with face attractiveness. In relation to possible functions in memory of the orbitofrontal cortex, it is of interest that a population of orbitofrontal cortex neurons responds to novel visual stimuli.

Effects of Damage to the Orbitofrontal Cortex on Visual Processing

Consistent with the neurophysiological and neuroimaging results described in this article, lesions in humans and macaques of the orbitofrontal cortex impair visual discrimination reversal. In humans, the ability to identify face expressions is also impaired, though face recognition is not.

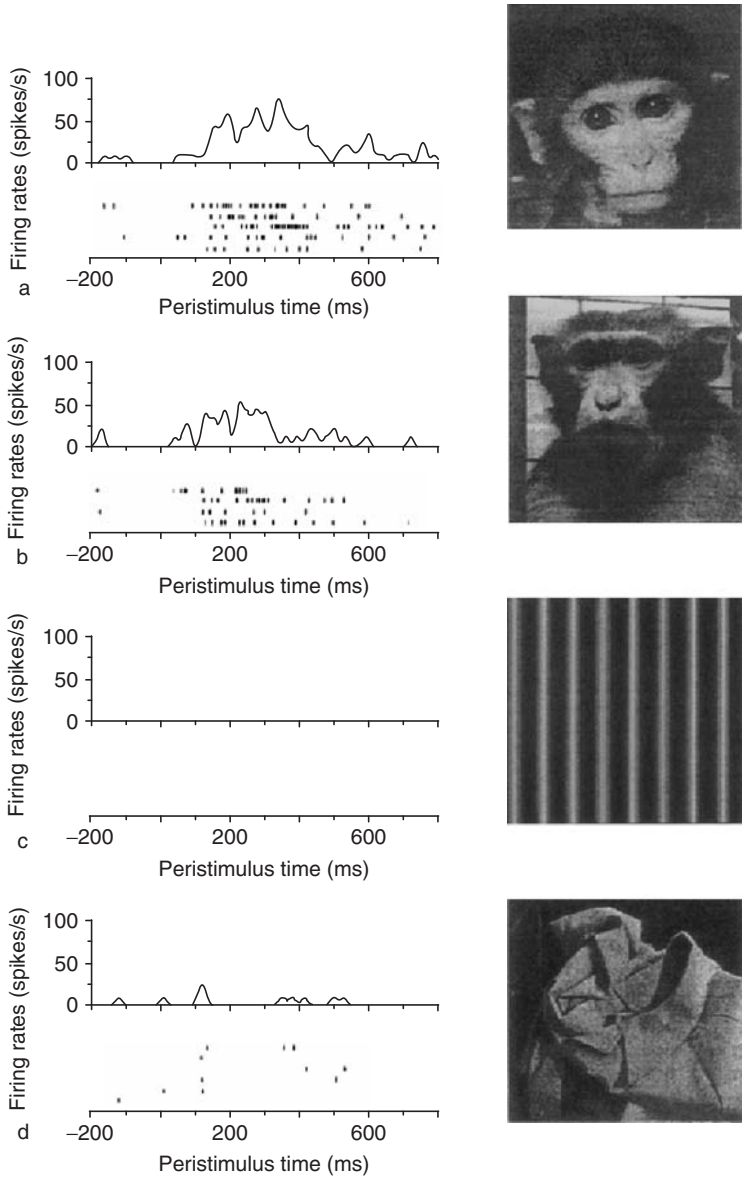

Figure 7 Orbitofrontal cortex face-selective neuron as found in macaques. Peristimulus rastergrams and time histograms are shown. Each trial is a row in the rastergram; several trials for each stimulus are shown (the ordinate is in spikes per second). The neuron responded best to face a; it also responded, though less, to face b, had different responses to other faces (not shown), and did not respond to nonface stimuli (c and d). The stimulus appeared at time 0 on a video monitor. From Rolls ET, Critchley HD, Browning AS, et al. (2006) Face-selective and auditory neurons in the primate orbitofrontal cortex. *Experimental Brain Research* 170: 74–87.

See also: Learning, Action, Inference and Neuromodulation; Reward and Learning; Reward Neurophysiology and Orbitofrontal Cortex; Visual System: Functional Architecture of Area V2; Visual Cortical Models of Orientation Tuning.

Further Reading

Berlin H, Rolls ET, and Kischka U (2004) Impulsivity, time perception, emotion, and reinforcement sensitivity in patients with orbitofrontal cortex lesions. *Brain* 127: 1108–1126.

Deco G and Rolls ET (2005) Attention, short-term memory, and action selection: A unifying theory. *Progress in Neurobiology* 76: 236–256.

Deco G and Rolls ET (2005) Synaptic and spiking dynamics underlying reward reversal in orbitofrontal cortex. *Cerebral Cortex* 15: 15–30.

Hornak J, Bramham J, Rolls ET, et al. (2003) Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain* 126: 1691–1712.

Hornak J, O'Doherty J, Bramham J, et al. (2004) Reward-related reversal learning after surgical excisions in orbitofrontal and dorsolateral prefrontal cortex in humans. *Journal of Cognitive Neuroscience* 16: 463–478.

Hornak J, Rolls ET, and Wade D (1996) Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia* 34: 247–261.

Kringelbach ML and Rolls ET (2003) Neural correlates of rapid reversal learning in a simple model of human social interaction. *NeuroImage* 20: 1371–1383.

Kringelbach ML and Rolls ET (2004) The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology* 72: 341–372.

O'Doherty J, Kringelbach ML, Rolls ET, et al. (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience* 4: 95–102.

Rolls ET (2004) The functions of the orbitofrontal cortex. *Brain and Cognition* 55: 11–29.

Rolls ET (2005) *Emotion Explained*. Oxford: Oxford University Press.

Rolls ET, Browning AS, Inoue K, et al. (2005) Novel visual stimuli activate a population of neurons in the primate orbitofrontal cortex. *Neurobiology of Learning and Memory* 84: 111–123.

Rolls ET, Critchley HD, Browning AS, et al. (2006) Face-selective and auditory neurons in the primate orbitofrontal cortex. *Experimental Brain Research* 170: 74–87.

Rolls ET, Critchley HD, Mason R, et al. (1996) Orbitofrontal cortex neurons: Role in olfactory and visual association learning. *Journal of Neurophysiology* 75: 1970–1981.

Rolls ET and Deco G (2002) *Computational Neuroscience of Vision*. Oxford: Oxford University Press.

Rolls ET, Hornak J, Wade D, et al. (1994) Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery and Psychiatry* 57: 1518–1524.

Thorpe SJ, Rolls ET, and Maddison S (1983) Neuronal activity in the orbitofrontal cortex of the behaving monkey. *Experimental Brain Research* 49: 93–115.

Waelti P, Dickinson A, and Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412: 43–48.

Orofacial Motor Control

B J Sessle, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

This article focuses on the brain stem and higher brain center mechanisms involved in the reflex regulation and sensorimotor coordination of the musculature in the orofacial region. Other works may be consulted for the structural and functional features of this musculature and its motor units and muscle fibers and sensory innervation (see the section titled 'Further reading').

Most of the literature on motor control focuses on limb motor control, and reviews of the topic often completely neglect the motor control mechanisms required for orofacial movements, many of which necessitate complex processing to provide for the exquisite motor control that is required for the muscles in the orofacial region. Yet these muscles serve important and diverse functions, some of which are vital for sustaining the life of the organism. In addition, there are unique features of the orofacial motor system that distinguish it from the spinal motor systems. Furthermore, many orofacial movements involve muscles innervated by several different cranial nerves. Given this and the large number of muscles involved, intricate bilateral motor control is essential to ensure that coordinated and appropriate motor patterning is carried out. Some of the muscles are multipennate, and thus the different compartments of these muscles may contribute differentially to the various motor functions carried out by each of these muscles.

Each muscle may be guided in these different functions by sensory inputs to its motor neurons that derive from a large array of different types of receptors. Many of these receptors have features that are similar to those in the spinal motor system, but some are associated with tissues unique to the orofacial region (e.g., tooth, cornea, taste receptors). The orofacial receptors are associated with primary afferents that project to the brain stem. The major sensory nuclei include the trigeminal (V) brain stem sensory nuclear complex (VBSNC), which receives most of the general somatosensory afferent input from the orofacial tissues, and the solitary tract nucleus (NTS), which receives visceral afferents (e.g., those supplying lingual taste buds, laryngeal and pharyngeal taste buds, and mechanoreceptors). The main cranial nerve motor nuclei include the Vth motor nucleus (Motor V), which provides the motor innervation of most jaw muscles; the VIIth motor nucleus, supplying the muscles of facial expression; the XIIth motor nucleus, supplying the intrinsic and extrinsic tongue muscles; and the nucleus ambiguus, which mainly supplies muscles of the palate, larynx, and pharynx.

Types of Motor Functions

The types of movements and other motor functions involving this musculature are also extensive, and some are unique to the orofacial region. They include voluntary movements that are driven by centers in the central nervous system (CNS) and that generally can be refined through sensory inputs to the CNS and reflex circuits. These voluntary movements can range from 'simple' movements involving a limited number of muscles (e.g., jaw opening or tongue protrusion), to the diversity of tongue positions and shapes that some of us can achieve, to the different facial expressions that we can manifest to express our feelings to others, to the complex and sophisticated movements that characterize speech. Reflexes are other types of movements whereby an afferent input to the CNS evokes a rather stereotyped motor response. Again, some of these reflex movements involve relatively simple reflex circuits and are numerous in the orofacial region (e.g., jaw-opening reflex and blink reflex). Other orofacial reflexes are more complicated and involve several motor neuron pools bilaterally (e.g., cough and gag), and others are even more complex and require the activity of numerous muscle groups and highly integrated CNS circuits (e.g., swallowing). Some of these complex motor events may also include a voluntary component as well as a reflex component (e.g., the oral vs. buccopharyngeal phases of a swallow).

Some of the orofacial motor functions may also be cyclic in nature. Chewing is a good example: the cyclic events (e.g., jaw opening and closing) are driven by a CNS program. Nonetheless, these cyclic jaw movements and other associated muscle activities (e.g., of tongue and facial muscles) that characterize chewing are sensitive to afferent inputs acting through brain stem reflex circuits or higher brain centers. Since chewing involves the production of a wide range of bite forces between opposing upper and lower teeth, highly coordinated processes are necessary, on the one hand, for example, to allow for light forces to explore and ascertain the texture of a foodstuff and on the other hand, to allow for the development of large forces that are necessary for biting through tough foodstuffs while avoiding self-injury.

Some of these sophisticated motor functions are laid down at birth, and others develop as the infant matures. Whereas swallowing develops *in utero*, and most mammals are born as suckling and swallowing animals, chewing is a learned motor function. As the infant matures, the rhythmic jaw movements come increasingly under central control and more sophisticated, engaging to a greater degree other muscle groups (e.g., tongue, facial) that allow for more refinement and management of chewing; the postnatal eruption of the teeth in humans provides important sensory inputs from periodontal receptors that also assist in the development of this control. Other highly developed oral functions also appear as the animal matures, the most striking being speech in humans, which involves an elaborate integration of sensory and motor pathways and central programming of muscles of the jaw, face, tongue, palate, pharynx, larynx, and other parts of the respiratory tract. Thus, there is the need for sophisticated neural circuits that allow for integration bilaterally of these diverse and complex motor functions and for coordination with other motor functions, such as respiration. Virtually all these functions of course are not purely 'motor' but indeed are sensorimotor since they depend on or utilize sensory inputs or feedback to initiate or guide them. Furthermore, there are several features that distinguish these functions and their underlying mechanisms from spinal sensorimotor processes and movements. In addition to the large number of muscles that may be involved, often requiring bilateral muscle activities, these features include the arrangement of the various sensory nuclei and motor nuclei into distinct neuronal pools in the brain stem, and unique aspects of the peripheral mechanisms.

Peripheral Mechanisms

Receptors and afferent inputs involved in orofacial motor control will not be considered in detail in this article. Nonetheless, it needs to be pointed out that the numerous orofacial afferent inputs that access the brain stem and are relayed to thalamus and cerebral cortex are involved not only in perceptual processes but may also be involved at each of these levels in sensorimotor integration and control. Furthermore, the distinctive features of some of the receptor systems in the orofacial region need to be kept in mind. These include the unique receptor systems noted above, the limited number of muscles that have muscle spindles and Golgi tendon organs that are so much a feature of limb motor control, and the presence of the primary afferent cell bodies of those jaw muscles that do have spindles (i.e., jaw-closing muscles) within the CNS, in the V mesencephalic nucleus.

Brain Stem Reflex Processes

As noted above, other important sensory and cranial nerve motor nuclei include the NTS, and the VIIth and XIIth motor nuclei, and the nucleus ambiguus. Furthermore, in addition to the VBSNC and NTS, important interneuronal sites include regions immediately adjacent to these sites (e.g., intertrigeminal nucleus, supratrigeminal nucleus), as well as components of the medial and especially lateral reticular formation. These sites receive and integrate afferent inputs from various orofacial tissues and from other brain stem or suprabulbar areas and thereby provide a neural substrate allowing for the initiation or modulation of the brain stem reflexes (e.g., by afferent inputs, descending controls, and conscious state). In addition, they provide some of the neural circuitry and processes that constitute the central programs (pattern generators) that are so crucial in the initiation and control of complex functions such as chewing and swallowing. In addition to the brain stem-based reflex responses evoked by these stimuli, afferent signals are relayed from the brain stem to higher brain centers involved in sensorimotor control of the muscles.

Brain Stem Reflexes

The motor neurons supplying the orofacial muscles receive reflex afferent inputs from the free nerve endings as well as specialized receptors in the orofacial tissues. Thus, the orofacial neuromuscular system can be reflexly influenced by the sensory inputs into the brain stem from receptors that signal pain, touch, joint position, muscle stretch or tension, and so forth. Via their connections with the interneuronal circuits noted above, these afferent inputs can activate or inhibit the cranial nerve motor neurons that supply the orofacial musculature. Brain stem circuits also underlie the reflex changes in heart rate, blood pressure, breathing, salivation, and more complex behaviors that can be evoked by nonnoxious or noxious stimulation of orofacial tissues. On the basis of these various types of responses, several human behavioral paradigms have been developed in order to study the effects of orofacial stimuli in humans; these include changes in autonomic function (e.g., heart rate and salivation), muscle reflexes, and facial expression.

Given the large number of muscles in this region and the diversity of receptors and afferent inputs, the number of reflexes is also immense, and only representative examples are noted below; more extensive reviews should be consulted for more details.

For example, several jaw reflex responses involving brain stem circuits and modulation by afferent and descending influences can be elicited by mechanical stimulation of the jaw-closing muscles. Stretch of these muscles evokes myotatic stretch reflexes through activation of jaw muscle spindle afferents. The primary afferent cell bodies of these spindle afferents are located in V mesencephalic nucleus, from which impulses in their central axons can monosynaptically activate jaw-closing motor neurons in Motor V. Synapses occur on these primary afferent cell bodies, and they and their central axons have intriguing electrophysiological and neurochemical features.

Other well-studied jaw reflexes include the jaw-opening reflex and the reflex effects of stimulation of periodontal receptors around the root of the tooth. A brief excitatory reflex may be elicited under certain conditions, but an inhibitory reflex involving one or more 'silent' periods in the jaw-closing muscles has received particular attention. A protective or regulatory function (e.g., of bite force) during jaw-closing movements is usually viewed as the role of this inhibitory reflex, and even its usefulness has been proposed as a diagnostic tool in certain orofacial pain or motor disorders. It is, however, subject to variation because of several interacting factors, and furthermore, inhibitory periods in the jaw-closing muscles can be induced by stimulation of different types of receptors in various orofacial tissues other than just periodontal tissues. The importance of other inputs is underscored by findings that feed-forward facilitation of masseter muscle activity produced by placing objects of varying hardness between the teeth is dependent on muscle afferent inputs as well as periodontal afferent influences.

Reflex excitatory or inhibitory effects can also be elicited in tongue, facial, and other muscles in the orofacial region. For example, stimulation of receptors in periodontal and other orofacial tissues can elicit reciprocal excitatory and inhibitory reflexes in tongue retrusive and protrusive muscles, and well-studied reflex responses in facial muscles in humans and laboratory animals are the corneal reflex and the blink reflex. It should also be noted that stimulation of many orofacial tissues, including laryngeal and pharyngeal mucosa, can evoke a number of respiratory-related reflexes that facilitate or prevent air intake into the airway (e.g., sniffing vs. coughing). Analogous 'supportive' or 'protective' reflexes exist for the alimentary tract (e.g., swallowing vs. gagging).

In the case of the reflex effects of noxious stimulation of orofacial tissues, brief high-intensity stimulation can elicit a transient jaw-opening reflex in the anterior digastric muscle and reflex activation in other muscles (e.g., tongue, facial); some stimuli can also, or instead, evoke transient inhibitory effects in these muscles. Algesic chemicals (such as hypertonic saline, capsaicin, glutamate, or mustard oil) applied to the temporomandibular joint, muscle, or other orofacial tissues of anesthetized animals can elicit more sustained afferent input into the brain stem that has been shown to activate neurons in subnucleus caudalis or in other components (e.g., subnucleus oralis) of the VBSNC. The activation of these neurons and their connection with the brain stem reflex centers, such as Motor V, can result in prolonged increases in electromyographic activity of both the jaw-opening and jaw-closing muscles, and several chemical mediators and receptor mechanisms (e.g., *N*-methyl-D-aspartate, opioids, γ-aminobutyric acid) in Motor V and interneuronal sites have been shown to be involved. It has been suggested that the cocontraction of these muscles may provide a 'splinting' effect that limits jaw movements in pathophysiological conditions affecting deep tissues such as the temporomandibular joint and muscle. There is, however, no consensus on whether the electromyographic activity of these muscles decreases, increases, or remains unchanged during experimentally induced or clinical orofacial pain in humans, and a number of factors may account for the disparity in experimental and clinical pain data. Influences from brain stem and higher brain centers involved in stress, emotion, alertness, sleep, and wakefulness are examples of other possible contributing factors, and the next sections briefly note some of these circuits and processes. Thus, the issue of how pain interacts with the neuromuscular system is not entirely clear, and theories have been proposed to account for the motor effects of pain that are seen clinically. The vicious cycle theory, for example, proposes that pain can lead to muscle hyperactivity and vice versa, whereas the pain adaptation model proposes changes in agonist and antagonist jaw muscles that allow the masticatory system to adjust to or adapt to the painful condition; most of the limited scientific evidence available favors more the latter concept.

Descending Influences

Many neurons in the VBSNC or NTS that relay to thalamus or other brain regions are subject to segmental (afferent) modulation and to descending modulation from various cortical and subcortical structures. Because many of these neurons also contribute as interneurons to reflex and other behavioral responses evoked by stimulation of orofacial tissues, these responses can also be regulated by these modulatory influences on the interneurons or, in some cases, on the motor neurons themselves that are part of the reflex circuits. These descending modulatory influences include those from the amygdala and other parts of

the limbic system, the lateral hypothalamus, the lateral habenular nucleus, the basal ganglia, the anterior pretectal nucleus, the red nucleus, the cerebellum, the sensorimotor cerebral cortex, the cortical premotor and supplementary motor areas, and the cortical masticatory and swallowing areas. It is through these descending excitatory or inhibitory influences that the higher brain centers can exert control over the brain stem processes and activities of motor neurons supplying the orofacial musculature and thereby initiate, guide, or regulate orofacial motor functions. The next section examines in more detail these higher brain center influences and processes.

Higher Brain Center Processes

Subcortical Processes

As noted above, several areas in the CNS project directly or indirectly to the cranial nerve motor neuron pools and can exert modulatory influences on motor behavior. There are also numerous connections between these areas, and so the neural circuitry involved in CNS control of motor function is extensive and complex. Because of space limitations, it is not possible to review these circuits in detail here. Also, only limited study has been done on these pathways as they apply to orofacial motor control compared with limb motor control, and most of the few studies have focused on the cerebral cortex (see the section titled 'Cortical processes') as well as basal ganglia, amygdala, hypothalamus, and brain stem regions.

In the case of the brain stem, the descending inputs as well as the afferent inputs from peripheral receptors access the reflex interneurons that project to and modulate motor neurons in the cranial nerve motor nuclei and that were noted above. Several of these regions also act in concert to form the neural circuitry of the central pattern generators (CPGs) for chewing, swallowing, and other analogous complex motor behaviors. Most research attention has focused on the CPGs underlying swallowing and especially mastication. The CPG for swallowing (the 'swallow center') appears to involve neurons within and adjacent to the NTS. Collectively, their output provides the time-locked, patterned drive to the different motor neuron pools supplying the muscles that participate in swallowing. While the swallow CPG neurons are triggered into action by sensory inputs, the CPG is nonetheless relatively insensitive to sensory feedback or descending controls once the swallow has started. The masticatory CPG (the 'chewing center'), in contrast, can generate chewing-like movements independent of orofacial sensory inputs. Nonetheless, studies in humans and animals indicate that it can utilize these inputs, especially those from periodontal mechanoreceptors and jaw muscle spindles, in concert with other brain regions accessing it, to provide for modification and guidance of masticatory movements. Thereby, the CPG-dependent stereotyped movements typical of chewing can be varied, and can function in an integrated manner with movements of the cheeks and tongue, to allow for repositioning of the food bolus and for alterations in masticatory force, velocity, and jaw displacement as the food is crushed and manipulated. These features explain how several factors (e.g., number of teeth, food composition and hardness, bite force) can influence the masticatory process and provide for the appropriate reduction of the food to a particular size suitable for swallowing. As part of this process, the CPG can also modulate sensory inputs such that central control can on the one hand mask out undesirable perturbances and reflexes that might disrupt the ongoing masticatory process and yet on the other hand allow nociceptive reflex inputs to access the masticatory motor neurons and thereby provide protection of the masticatory apparatus. One common example of this process in operation is the interruption of chewing by the jaw-opening reflex elicited by a noxious stimulus such as a fish bone piercing the oral mucosa while a person is eating a fish meal.

There are several higher brain centers that access the motor neurons directly or indirectly via the various brain stem interneuronal groups, including the CPGs. A particularly important series of inputs derives from various components of the basal ganglia and substantia nigra, and many cerebral cortical output pathways project indirectly to the brain stem via the basal ganglia (as well as directly to the brain stem). Orofacial motor deficits can arise after lesions of some of these components, and basal ganglia neurons have been recorded (e.g., in putamen and globus pallidus) that show activity during orofacial movements such as chewing, and many of these neurons may receive orofacial sensory inputs. Pharmacological studies have also underscored the relevance of these structures to orofacial motor control. For example, an imbalance in dopamine (e.g., by drugs that alter dopamine actions, such as amphetamines and cocaine) may change basal ganglia functional activities and perhaps be a factor leading to abnormal motor functions such as those seen in oral dyskinesia and sleep bruxism. As in the subcortical processes involved in limb motor control, several other neurochemicals and subcortical structures and interconnecting circuits are involved in orofacial motor control. They include acetylcholine, γ-aminobutyric acid, glutamate, serotonin, vasopressin, catecholamines, and opioids, and the structures include the hypothalamus, amygdala, subthalamic nucleus, red

nucleus, anterior pretectal nucleus, superior colliculus, periaqueductal grey, and cerebellum.

Cortical Processes

Several cortical areas exert descending influences on brain stem and other subcortical regions. Studies using transcranial magnetic stimulation or imaging (e.g., functional magnetic resonance imaging, positron-emission tomography) in humans have revealed that orofacial movements may be associated with activation of the face primary motor cortex (M1), primary somatosensory area (S1), premotor cortex, supplementary motor area (SMA), cortical masticatory area (CMA), insula, and anterior cingulate gyrus, to name just a few. The area(s) activated may vary depending on the type of movement (e.g., swallowing vs. chewing vs. clenching or tapping the teeth together), probably reflecting in part whether the movement involves evoked sensory inputs that project to the cortical area(s). This is consistent with intracortical microstimulation (ICMS) studies in monkeys and subprimates (e.g., rats) revealing evoked movements from only some of these sites and indicating that the pattern of some evoked movements (e.g., mastication) may vary between sites. In addition, the ICMS studies have documented that each orofacial muscle or movement is represented multiple times within M1, indicating that each output zone of M1 control one of the many contextual functions in which a muscle participates. This multiple representation includes ipsilateral and contralateral elemental movements (e.g., jaw-opening, facial twitch, tongue protrusion), as well as complex activities such as mastication and swallowing. For example, mastication or swallowing can be evoked by ICMS, not only from the deep and principal parts of the classical CMA lateral to face M1, but also from within face M1 and even within the face S1 in the monkey, and there is considerable overlap of face M1, S1, and CMA in subprimates. Furthermore, selective cold block or ablation of each of these regions can disrupt chewing and swallowing to varying degrees, indicating that each may be involved differentially in the production and patterning of chewing and swallowing, including the coordination of the tongue, facial, and jaw movements that is necessary for the proper ingestion and transport of food or liquid. It is to be noted also that ICMS-evoked orofacial movements have also been reported for SMA, an area which, like premotor cortex, has been implicated in the preparation for movement. However, little information is available on the role of these regions in orofacial motor control except for recent studies indicating that premotor cortex, as well as other cortical areas (e.g., parietal cortex), may be involved in the preparation and planning for ingestion, visuomotor control, and perhaps even in cognitive functions related to the understanding and communication of ingestive motor actions, facial recognition, and other complex behaviors involving the orofacial region.

Single neuron recordings in monkeys or subprimates have revealed that many face M1 and S1 neurons discharge in relation to chewing or swallowing, but most also discharge in association with more elemental movements, with some active in relation to jaw movements (especially jaw opening) and many others active in relation to tongue movements. These findings are consistent with the above ICMS findings as well as with evidence that ablation or cold block of the monkey's face M1 or S1 disrupts the animal's ability to perform a learned tongue task but causes much less disruption of a biting task. The single neuron recordings in face M1 have also revealed that some neurons are active during preparation for movement, although many more neurons in premotor areas appear to show this feature, at least during visually guided jaw movements, consistent with cortical findings related to limb motor control that have placed more emphasis on cortical (and subcortical) regions other than just M1 in mechanisms involved in motor planning and preparation.

The neuronal and ablation findings noted above also underscore the importance of the somatosensory cortex as well as the motor cortex in the fine motor control of orofacial movements. The face S1 can influence orofacial movements by its projections to face M1 and other cortical areas and to subcortical regions such as the VBSNC, NTS, reticular formation, and even directly into the cranial nerve motor nuclei in primates. The face S1 has a somatotopically arranged array of sensory inputs which are predominantly from facial skin and intraoral structures but may also include some inputs from deep tissues (e.g., muscle). The face M1 and CMA also receive inputs from afferents supplying deep and especially superficial (e.g., skin, mucosa, teeth) tissues of the face, mouth, and jaws. The substantial number of face M1 and CMA neurons that receive bilateral inputs from the orofacial tissues is probably related to the need for bilateral sensorimotor coordination in orofacial muscle function. These inputs may be used by CMA to help guide masticatory-related movements. Face M1 may also use its orofacial afferent inputs for generating and regulating orofacial movements in order to refine ongoing cortical motor activity and shape the appropriate motor response, for example, in the control of voluntary orofacial movements such as the manipulation of the food bolus after it is placed in the mouth, since many face M1 neurons are active

during the food preparatory phase. Sensory inputs from the orofacial regions presumably are utilized by these M1 neurons for this purpose.

Thus, face M1 and S1 may use orofacial afferent inputs to guide, correct, and control movement by the use of sensory cues prior to movement and by using sensory information generated during movement. These processes may involve intracortical processing, cortical gating, and transfer of somatosensory information, as well as corticofugal projections to subcortical sites that modulate and select somatosensory information ascending through subcortical relay neurons in the brain stem (e.g., VBSNC and NTS) and thalamus. These inputs also play critical roles in motor learning and in the motor adjustments or adaptations that take place after a change in the sensory environment. Therefore, this article now turns to a consideration of neuroplasticity, especially as it applies to the cortical mechanisms of orofacial motor control.

Both S1 and M1 have been shown in numerous studies to undergo neuroplastic changes following manipulations of afferent inputs or in association with learning of novel motor skills. While nearly all these studies have focused on limb S1 and M1, as well as some subcortical regions (e.g., basal ganglia), there is some recent evidence that comparable neuroplastic changes may occur in face S1 and M1. For example, dental extraction can lead to the loss of dental representations and an enhanced representation of adjacent orofacial structures in face S1 of the mole rat, and alterations of the dental occlusion can be associated with modifications of the motor representation of jaw and tongue muscles in face M1. These findings suggest that physiological changes occur in face S1 and M1 when the oral environment is altered, and they raise the possibility that such cortical changes reflect adaptive mechanisms that may be crucial in determining how well or poorly a person adapts to such alterations (e.g., dentures, implants, pain, and nerve damage).

A role for face M1 in orofacial motor skill acquisition is also suggested by recent novel findings in monkeys and humans of face M1 neuroplasticity associated with the learning of a novel tongue-protrusion task. No comparable changes were seen in CMA in the monkey study. These studies have underscored that face M1 plays a crucial role in sensorimotor integration and control and that it has a remarkable capacity for neuroplasticity that may reflect dynamic, adaptive cortical mechanisms mediating complex motor and cognitive phenomena related to motor learning and memory. They also suggest that there may be cortical templates for a variety of familiar, learned orofacial motor activities. Speech is one such activity. While its consideration is beyond the scope of this article, cortical mechanisms are clearly crucial to the fine control and coordination of the various muscles participating in voice production and articulation of speech sounds, both when a child is learning to speak and once speech has been learned.

See also: Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Corticomotoneuronal System; Motor Control of Feeding and Drinking; Neural Circuitry in the Somatosensory System; Posterior Parietal Cortex and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape; Sensation from the Face; Sensorimotor Integration: Models; Speech Production: Development; Trigeminal Motor System.

Further Reading

Dubner R, Sessle BJ, and Storey AT (1978) *The Neural Basis of Oral and Facial Function*. New York: Plenum.

Hannam AG and McMillan AS (1994) Internal organization in the human jaw muscles. *Critical Reviews in Oral Biology and Medicine* 5: 55–89.

Hiiemae KM and Palmer JB (2003) Tongue movements in feeding and speech. *Critical Reviews in Oral Biology and Medicine* 14: 413–429.

Jean A (2001) Brain stem control of swallowing: Neuronal network and cellular mechanisms. *Physiological Reviews* 81: 929–969.

Lavigne GJ, Kato T, Kolta A, and Sessle BJ (2003) Neurobiological mechanisms involved in sleep bruxism. *Critical Reviews in Oral Biology and Medicine* 14: 30–46.

Lund JP and Kolta A (2006) Generation of the central masticatory pattern and its modification by sensory feedback. *Dysphagia* 21 (3): 167–174.

Miles TS, Nauntofte B, and Svensson P (2004) *Clinical Oral Physiology*. Copenhagen: Quintessence.

Sessle BJ, Adachi K, Avivi-Arber L, et al. (2006) Neuroplasticity of face primary motor cortex control of orofacial movements. *Archives of Oral Biology* 52(4): 334–337.

van der Bilt A, Englen L, Pereira LJ, van der Glas HW, and Abbink JH (2006) Oral physiology and mastication. *Physiology & Behavior* 89(1): 22–27.

Yamada Y, Yamamura K, and Inoue M (2005) Coordination of cranial motoneurons during mastication. *Respiratory Physiology & Neurobiology* 147: 177–189.

Osmoregulation

A K Johnson, University of Iowa, Iowa City, IA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Life began and initially evolved in the sea. The oceans have always been composed of a watery broth of mineral salts, molecules, and dissolved gases. Water and these constituents provide the necessary ingredients for the generation of more-complex molecules and for the reproduction, growth, maintenance, and repair of cells. About 150 years ago, biologists such as the noted French physiologist Claude Bernard first recognized that complex animals had, in essence, internalized the primeval sea in which life evolved. This bodily ocean bathes cells and provides them with an immediacy of ingredients necessary for life. Maintaining a high degree of consistency of the fluids outside of cells (extracellular fluid) is vital and gives advanced organisms the opportunity for life independent of the ancient seas.

In the early twentieth century, Walter Cannon coined the term 'homeostasis' to describe the concerted actions of physiological systems to maintain many physical (e.g., blood pressure) and chemical (e.g., blood pH and extracellular glucose concentration) bodily parameters within relatively narrow limits. Because the viability of cells greatly depends on the constituents of bodily fluids, processes that determine total body water and its distribution are of pivotal significance in the maintenance of virtually all homeostatic processes. Given the importance of body fluid homeostasis, it should not be surprising that there are multiple systems directed at assuring that fluid resources are optimally maintained. Many of these complex physiological mechanisms are related to the process of osmoregulation.

For many neuroscientists and physiologists who study osmoregulation, the concept entails the mechanisms involved in sensing the concentration of extracellular osmoles and the corrective actions taken if the extracellular condition is different from the ideal (i.e., a deviation from a set point). Regulatory responses usually are initiated within a few seconds of detection of an altered osmotic state. In the case of hyperosmolality, provided water is accessible, the compensatory and restorative responses run their course within minutes to at most a couple of hours.

The following review provides a discussion of the nature of osmosensing and a description of some of the cellular and central neural mechanisms found to activate the effector systems responding to a hyperosmotic state. However, it begins with a broader treatment, including important control systems that contribute to more-global aspects of body fluid homeostasis. As an introduction to this topic, it is helpful to gain an appreciation of the nature of body fluids and of the physical determinants of solute and solvent distribution between the key fluid compartments of the body.

Body Fluid Compartments and the Distribution of Water between Them

Approximately 60% of body weight is water. Under normal conditions the totality of water contained in all the cells of the body (i.e., in the intracellular fluid compartment) is about twice the amount of all the water that is outside the cells (i.e., in the extracellular fluid compartment). The 2:1 ratio of intra- to extracellular fluid volume is mainly the result of concentrations of various impermeable ions and molecules located on each side of the cell membrane. The movement of water in and out of cells is largely unimpeded as it can pass through the cell membrane by simple diffusion and is augmented by movement through water channels formed by proteins called aquaporins. Experimentally, it is easy to dehydrate the intracellular compartment by increasing the concentration of impermeable solute in the extracellular space. This generates an osmotic force that initiates the movement of water (osmosis) out of cells, dilution of extracellular solute, and shrinkage of the intracellular space. Using the same principle, it is possible to increase the volume of the intracellular compartment by decreasing the concentration of extracellular impermeable solute, for example by administering into the extracellular compartment water or a hypotonic fluid (i.e., a solution containing a lower concentration of impermeable ions or molecules than found in the extracellular fluid). In this case, water moves down its concentration gradient and enters cells until the extracellular fluid becomes iso-osmotic (isotonic) with the intracellular fluid. Although there are many different types of molecules and ions that do not readily cross cell membranes, ionic sodium (Na^+) is the solute in the extracellular compartment that is the major factor in determining the dynamic apportioning of water between the intra- and extracellular compartments.

The extracellular compartment includes both the interstitial fluid that is located in the space between cells and the plasma of the cardiovascular system. To all intents and purposes, circulating plasma (excluding blood cells and large proteins) is in ready communication with the interstitial fluid, so that a

contraction (e.g., hemorrhage) or an expansion (e.g., plasma transfusion) of blood volume is rapidly followed by a proportional reduction or increase of interstitial fluid. Manipulating blood volume by removing or adding plasma that is iso-osmotic with the intracellular fluid will not immediately change the volume of the intracellular compartment. For example, an acute hemorrhage, which produces an absolute deficit in total body fluid, does not produce a rapid decrease in total intracellular fluid volume.

Responses to Dehydration of the Intra- and Extracellular Fluid Compartments

Under physiological (e.g., exercise, exposure in a warm environment, and normal renal loss) and pathophysiological (e.g., emesis and diarrhea) conditions, animals lose water and Na^+ to the environment. Depending on the cause or type of dehydration accrued, the relative proportion of Na^+ to water loss from the body will vary between a net hypotonic and a net hypertonic fluid loss (Figure 1; note hypertonic fluid loss can occur via the kidneys), and the immediate consequence will be a disproportionate fluid loss from the intra- or the extracellular compartments. Under specifiable experimental conditions, it is possible to selectively dehydrate the intracellular compartment without reducing extracellular fluid volume, and vice versa. Fluid loss from either the intra- or extracellular fluid compartment produces an activation of both reflexive (i.e., autonomic nervous and endocrine systems) and behavioral (i.e., water drinking) responses that serve first to minimize the consequences of dehydration and second to correct such homeostatic disruptions.

Dehydration causes increased activity in the sympathetic nerves innervating the cardiovascular system and the kidneys. The effect of increased sympathetic cardiovascular tone is intended to maintain cardiac output and the perfusion of critical vascular beds (e.g., brain, heart). Sympathetic action on the renal tubules promotes Na^+ and water reabsorption, thereby reducing the rates of the progressing loss of one or both of these substances. The predominant endocrine response to dehydration is the release of the antidiuretic hormone, vasopressin (VP), from the posterior pituitary. This hormone acts on the kidney to promote the reabsorption of water. The initial behavioral response of a dehydrated animal is to seek out and drink water. Such drinking behavior is the result of central nervous system (CNS) activation producing a drive for water (i.e., the motivational state of thirst). Under conditions when there is a loss of Na^+ from the body, such as after the loss of isotonic plasma, a drive to obtain and consume NaCl is activated in conjunction with thirst. This

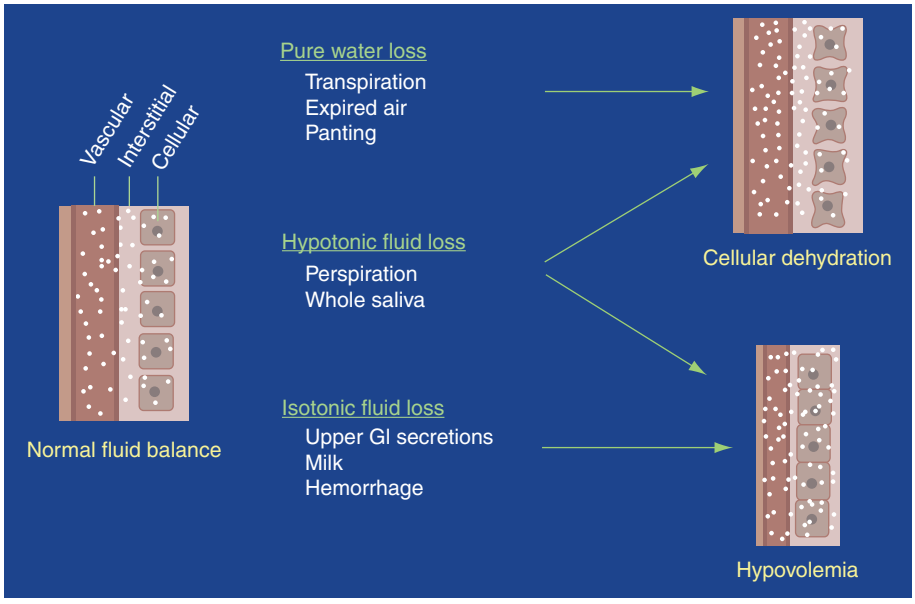

Figure 1 Physiological challenges and nonrenal loss of water and sodium from the body result in differential dehydration of the intra- and extracellular fluid compartments. The effects of different types of dehydration varying between the loss of pure water to the loss of isotonic body fluid on the intra- and extracellular fluid compartments. The extracellular fluid compartments consist of the fluids in blood vessels and in the interstitial space. The fluids in the vasculature and interstitial space are in ready communication with one another, and the induction of hypovolemia shrinks proportionally the vascular and interstitial volumes. Given sufficient time, renal mechanisms can cause the redistribution of fluid between the extra- and intracellular fluid compartments. GI, gastrointestinal.

motivational state is defined as salt (or sodium) appetite.

Coordinated endocrine and autonomic reflexes and motivated behaviors to cope with dehydration require an exquisite set of integrated physiological responses. The release of VP and sympathetic activation act to reduce the rate of renal fluid loss and to maintain vascular perfusion in the face of moderately reduced fluid resources. These reflexes permit an animal to remain behaviorally competent. Evidence indicates that the threshold for release of VP in response to dehydration is slightly lower than that for activation of thirst-related behaviors. There is probably a similar difference in dehydration thresholds for activating a reflex increase in sympathetic tone versus the mobilization of thirst. The earlier activation of sympathetic and endocrine reflexes allows an animal to continue its normal pursuits (e.g., seeking food or mates, parental care, and sleeping) before thirst and the compulsion to find and consume water are aroused. It should be appreciated that although hormonal and autonomic countermeasures play an important temporizing role in the dehydrated state, it is only through the consumption and absorption of water that homeostasis normally is restored.

It is also worth recognizing that a degree of independence in the control of the rates of renal Na^+ loss versus renal water loss makes it possible, in principle, to partially exchange a state of cellular dehydration for a state of extracellular dehydration or vice versa. For example, if for some reason an animal experiences a pure cellular dehydration (e.g., dogs panting in the heat) but has no access to water, it will over an extended period of time apportion the water deficit between the cellular and extracellular compartments by excreting hyperosmotic urine.

In contrast to dehydration, an excess of Na^+ or water or acute hypertension will activate inhibitory mechanisms. For example, in a hyperosmotic state or under a condition in which blood volume is expanded, NaCl appetite is inhibited, and the release of hormones (e.g., oxytocin in some species and atrial natriuretic peptide) that increase renal Na^+ excretion will be facilitated.

In order to appropriately activate reflexes to minimize the consequences and progression of altered hydration and to mobilize or inhibit homeostatic behaviors, the brain must be informed of the disposition of body fluids. Several types of sensors are involved in detecting the status of total body fluid and the regional distribution of Na^+ and water.

Both mechano- and chemoreceptors have been implicated in the detection of extracellular volume. The most thoroughly studied receptor systems implicated in body fluid and cardiovascular homeostasis are mechanoreceptors, known collectively as vascular baroreceptors (i.e., pressure receptors). There are two subclasses of baroreceptors – one located on the low-pressure side of the circulation in the large veins (e.g., superior vena cava) and cardiac atria, and a second set on the high-pressure side of the circulation in the carotid sinuses and aortic arch. Baroreceptor nerve endings are insinuated into blood vessels, and their membrane potentials change in proportion to the degree of distention of the regions they innervate. Increased deformation of an innervated blood vessel results in an increased frequency of action potentials in these sensory (afferent) nerves. On the venous side of the circulation, with its high degree of compliance, graded increases in blood volume are reflected as increased pressure and stretch in the innervated low-pressure vessels and atria. Because of the low compliance of blood vessels on the high-pressure side of the circulation, arterial baroreceptors are not tuned to sensing small changes in blood volume. However, they do respond when extracellular fluid loss becomes so large that arterial blood pressure falls, such as in the case of hypotensive hemorrhage.

Information from baroreceptors reaches the brain by the IXth and Xth cranial nerves, which synapse in the nuclei of the solitary tract (NTS). When blood volume or arterial pressure falls below normal levels, afferent baroreceptor nerve activity declines and signals the reflex release of VP and sympathetic nervous system activation. With sustained hypotension or hypovolemia, behaviors leading to Na^+ and water intake will be initiated. Alternatively, when blood pressure or blood volume acutely increases above its normal set point, most regulatory endocrine and autonomic reflexes and behaviors are inhibited.

Paralleling neural afferent input reflecting blood volume and arterial pressure are humoral factors that reinforce the neural information entering the CNS. The most thoroughly studied are components of the renin-angiotensin-aldosterone system. In states of hypotension, hypovolemia, or hyponatremia, the enzyme renin is released from the kidneys to trigger a metabolic cascade to ultimately generate the octapeptide angiotensin II (ANG II). ANG II is an opportunistic hormone that acts at innumerable sites and through multiple mechanisms to produce an expansion of blood volume and an increase in blood pressure. Beyond its potent smooth muscle vasoconstrictor actions, ANG II acts to enhance sympathetic nervous system activation by blunting baroreceptor depressor reflexes. This effector peptide also enhances VP release, is a potent dipsogenic and natriorexigenic agent, and a major controller of aldosterone release. The hormone aldosterone acts on the kidney to conserve Na^+, and in many species this steroid also targets the brain to induce salt appetite.

ANG II has a molecular weight of about 1000 Da and does not readily gain access to the brain. Circulating ANG II exerts CNS-mediated effects on autonomic reflexes, VP release, and behavior (i.e., thirst, salt appetite) by acting on a set of small, brain-related structures referred to as sensory circumventricular organs (CVOs). The parenchyma of these regions lacks a blood–brain barrier, and CVOs have been identified as putative receptive sites for various circulating factors. The three sensory CVOs are midline (or near midline) periventricular structures. Two of them, the subfornical organ (SFO) and the organum vasculosum of the lamina terminalis (OVLT), lie, respectively, along the dorsal and ventral aspects of the rostral wall of the third cerebral ventricle. The area postrema (AP), the third sensory CVO, is located on the floor of the fourth ventricle in the caudal medulla. Various lines of evidence indicate that each of the sensory CVOs is responsive to ANG II. The SFO and OVLT are implicated in ANG II sympathetic activation, VP release, thirst, and salt appetite. Results suggest that ANG II action on the AP blunts baroreceptor-mediated reflex depressor responses.

Information about both intra- and extracellular fluid status uses both afferent nerve activity and humoral stimuli as body–brain signaling mechanisms. Input from these two types of sensory modalities enters into an extensive neural network, the visceral neuraxis. This interconnected neural matrix processes information to coordinate and activate regulatory reflexes and behaviors to restore homeostasis. Certain nuclei and brain regions have been implicated as especially critical for maintaining body fluid and cardiovascular homeostasis. Among these are the AP, NTS, caudal and rostral ventrolateral medulla, parabrachial nuclei, raphe nuclei, locus coeruleus, paraventricular and supraoptic hypothalamic nuclei, several nuclei of the extended amygdala, the perifornical region of the lateral hypothalamus, and structures lying along the lamina terminalis.

Counterposed to systems which enhance retention or acquisition of water or Na^+ are systems that inhibit body fluid expansion in the case of overhydration. In rats, in addition to released systemic oxytocin acting to induce natriuresis, oxytocinergic mechanisms also act in the brain to inhibit solute (e.g., Na^+) intake. Other centrally acting factors limiting excess intake of water or Na^+ include brain natriuretic peptide, serotonin, cholecystokinin, corticotropin releasing hormone, and tachykinins. One or the other of these agents has been implicated as acting in periventricular tissues, the parabrachial nuclei, the dorsal vagal complex, or the AP to inhibit sympathetic and fluid retaining hormonal reflexes and prevent excess water or Na^+ intake.

Extracellular Osmoles and the Concept of the Osmoreceptor

An insight into the nature of the stimuli involved in signaling the status of the composition of the extracellular compartment was gained from experiments conducted on thirst in dogs by Alfred Gilman in the mid-1930s. Although it had long been proposed that increased extracellular fluid solute was an adequate stimulus for thirst, Gilman demonstrated that there was an essential property of the solute that was necessary in order to trigger drinking in the dog. Specifically, he showed that for a hyperosmotic solution to be an effective stimulus for thirst, its solute must not readily cross cell membranes. Gilman demonstrated that intravenously infused impermeable solutes, such as NaCl, that dehydrate cells act as excellent dipsogenic stimuli. In contrast, equiosmotic solutions (as measured by either freezing point depression or a vapor pressure osmometer) comprised of solutes (e.g., urea) that pass through membranes will not effectively dehydrate cells, and they do not induce significant thirst.

The nature of the receptors hypothesized to mediate the release of antidiuretic hormone in response to hypertonic solutions was proposed shortly after Gilman by EB Verney. By administering various hyperosmotic agents, Verney came to conclude that antidiuretic hormone (VP) release was mediated by a receptor that was specifically sensitive to stimuli that caused the receptive cells to shrink. He hypothesized that these cells were sensitive to their own volume and coined the term osmoreceptor. It was only a short time until the concept of the osmoreceptor was appropriated to describe the hypothetical sensors that mediate hyperosmotically induced thirst. Within the same period, Bengt Andersson serendipitously discovered that injections of hypertonic saline into the hypothalamus produced avid drinking in goats. Consequently, the brain became the prime focus as the site housing osmoreceptors.

Advances in neuroscience in the 1950s and 1960s and experiments from his own laboratory led Bengt Andersson to propose that the brain osmoreceptor may actually be a Na^+ receptor. He reasoned that in prior experiments in which various hyperosmotic solutes were delivered systemically to induce thirst or VP release, the presence of the blood–brain barrier caused the brain to be dehydrated. In other words, the brain becomes dehydrated and hypertonic because water moves out of the brain more rapidly than does Na^+. The consequence of this unequal movement of water and Na^+ is that brain extracellular Na^+ concentration is increased. Andersson went on to speculate that brain Na^+ receptors are responsible for the activation of regulatory actions to systemic extracellular

hyperosmolality. In response, Michael McKinley and colleagues pointed out that in experiments generating evidence for his hypothesis, Andersson and co-workers had made their hypertonic solutions for central administration by using only sugars (e.g., sucrose) in water. It is possible such infusates produce a dilution of extracellular ions, particularly Na^+, and this may disrupt normal neural function. McKinley's laboratory found that intracerebroventricular infusions are quite effective in stimulating drinking when solutions were made hypertonic by adding impermeable sugars to an artificial cerebrospinal fluid (aCSF; aCSF contains Na^+, among other ions) as a vehicle. One caveat noted by McKinley et al., however, was that solutions made hypertonic by adding NaCl to aCSF are more effective as stimuli than are equiosmolar, sugar-plus-aCSF-constituted hyperosmotic solutions. These findings imply that both osmo- and Na^+ receptors in the brain are likely to be involved in osmoregulation.

The osmoreceptor as a sensor of extracellular osmolality was initially proposed as a hypothetical construct to account for a set of experimentally observed phenomena. However, in 60 years, no one has unequivocally demonstrated a specific cell type, neuron or otherwise, that functions to sense osmolality and trigger increased neural activity in systems controlling thirst, VP release, or sympathetic outflow. However, the concept has engendered productive thought and investigation on issues such as receptor localization and what actually is being sensed (osmole concentration or Na^+ concentration) under physiological conditions.

Although the brain is firmly implicated in the process of maintaining consistency of extracellular osmolality or Na^+, there are also osmoregulatory responses triggered by peripheral mechanisms that almost immediately sense the nature of a given hydrational load. These so-called presystemic sensing mechanisms act to trigger compensatory regulatory actions before there are detectable changes in the osmolality of the systemic circulation or of the brain extracellular fluid. For example, the oral consumption of fluid immediately inhibits VP release and reduces the amount of subsequent drinking in dogs, humans, and other species. Experimental evidence indicates that presystemically signaled regulatory responses to prevent overhydration may derive from different sources and through various mechanisms. These include oropharyngeal monitoring of fluid type and volume, metering of swallowing, detection of gastric volume and/or osmolality, and the sensing of osmolality or Na^+ concentration in the hepatic portal circulation. Undoubtedly, different types of chemo- and mechanoreceptors are involved in these diverse sensory systems. Among the presystemic receptors and sensory pathways most thoroughly studied are the osmo-/sodium receptors in the splanchnic circulation and the afferents from these receptors. Manipulating the osmolality of solutions infused into the hepatic portal vein alters afferent nerve activity, and water is more effective in producing diuresis when it is infused into the portal vein than when infused systemically. Thirst and salt appetite have also been shown to be appropriately facilitated or inhibited by varying the concentration of NaCl in hepatic portal vein infusates. Input from the visceral receptors sensing extracellular osmolality or Na^+ concentration travels in vagal nerve afferents and enters the CNS via the NTS. Although all mammals studied to date have presystemic signaling that contributes to osmoregulation, there appear to be differences in the specific mechanisms a given species employs.

It is likely that there are many sites and cell types in the brain that are osmo- or Na^+ receptive. In the 1950s, Verney and his co-worker, Jewell, speculated that the hypothalamic cells that synthesize VP might themselves be osmoreceptors. Use of *in vitro* intracellular recording methods allowed investigators more recently to demonstrate that increases in extracellular osmolality will depolarize hypothalamic vasopressinergic magnocellular neurons. However, when the hyperosmotic challenge does not exceed physiological levels, the degree of depolarization is not sufficient to generate action potentials. Many investigators believe that magnocellular neurons are likely to receive input from other osmo- or Na^+ receptive regions in the brain or body and that these act to stimulate VP release.

The periventricular tissue surrounding the anteroventral third cerebral ventricle (AV3V) has received significant attention in recent years as a likely site containing osmo- or Na^+ receptors influencing not only VP release but also sympathetic tone and thirst. Awareness was drawn to this region in the mid-1970s by studies conducted in rats that destroyed a small volume of AV3V tissue and resulted in chronic deficits in drinking, VP release, and pressor responses to hypertonic challenges. AV3V lesions that produce these global alterations in behavioral and reflex responses to hypertonic stimuli include the OVLT, the ventral median preoptic nucleus (MnPO; i.e., the MnPO below the anterior commissure), and the preoptic/anterior hypothalamic periventricular nuclei. Functional studies conducted in larger species (sheep and dogs) have suggested that osmoreceptors, as defined by Verney, are likely to reside in one or more of the sensory CVOs.

The OVLT is regarded as a major candidate among the sensory CVOs as a site housing central osmo- and/or Na^+ receptors. However, lesions restricted to the OVLT reduce but do not abolish responses to systemic hypertonic challenges, whereas OVLT lesions that invade the ventral MnPO virtually abolish all

osmotically induced regulatory responses. It has been suggested that the ventral MnPO is osmo- or Na^+ sensitive and that it is a key site for integrating input from osmo- or Na^+ receptors from various regions, including input from the SFO and OVLT, and ascending visceral input from the periphery (e.g., hepatic portal circulation). Recently, lamina terminalis tissues have been the focus of studies using molecular techniques to identify ion channels as candidates for osmo- and Na^+ sensing. Two classes or types of channels, (1) the transient receptor potential ion channel, vanilloid subfamily (TRPV) and (2) the Na_x Na^+ channel, are of particular interest.

The TRPV type 4 (TRPV4) was among the first implicated in osmosensing. TRPV4 is a vertebrate homolog of an osmotically gated ion channel in the worm *C. elegans* and was identified through a candidate gene approach. *In situ* hybridization studies indicate that abundant TRPV4 messenger RNA is found in the OVLT, SFO, and MnPO, as well as in the choroid plexus. When expressed as a homomeric channel in a heterologous system, TRPV4 channels have been shown to be activated by cell volume expansion but not by cell shrinkage. Hence, this type of channel may be ideally suited to activate brain systems that inhibit excess fluid intake or retention. Studies indicate that *trpv4* null (knockout) mice have small alterations in mechanisms both facilitating and inhibiting thirst in hyper- and hypo-osmotic states. Another TRPV channel suggested to be a variant of the TRPV1 channel has recently been implicated in mechanotransduction in supraoptic magnocellular and OVLT neurons. TRPV1 knockout mice are hyperosmotic under basal conditions. Thirst and VP release to cellular dehydration have been reported to be attenuated in *trpv1* null animals.

The Na_x channel is present in the OVLT, SFO, and MnPO (in rat) and has been implicated in the control of salt appetite. Na_x null mice drink large amounts of hypertonic saline under experimental conditions in which salt appetite is normally inhibited in wild-type mice. This response is corrected in rescued mice produced by transducing the SFO with Na_x complementary DNA. The Na_x Na^+ channel involved in salt appetite control has been suggested to be associated with glia.

Recent discoveries that specific ion channels are likely to play a role as osmo- and Na^+ receptors offers great promise for achieving a more thorough understanding of the central mechanisms by which changes in body and brain extracellular fluid osmolality and/or Na^+ concentration are sensed and transduced. Given the likely involvement of many different types of channels and cells, including glia, the classical concept of an osmoreceptor is likely to change.

Summary and Conclusions

Organisms use multiple strategies to maintain consistency of their internal fluid environment. One of the factors carefully regulated by neural and hormonal reflexes and by behavior is extracellular osmolality. Na^+ is the most important extracellular solute, and Na^+ concentration or a suitable proxy (e.g., osmolality or total impermeable extracellular molecules and ions per volume of fluid) is probably sensed by Na^+ receptors or osmoreceptors. Other receptor systems are also linked in important ways to broader aspects of osmoregulation. These include systemic high- and low-pressure baroreceptors and receptors for circulating humoral factors that are released in response to changes in water and sodium balance or distribution and blood pressure changes (**Figure 2**).

Fluid-related information derived from systemic receptor systems converges on the brain through visceral afferent nerves and by accessing parts of the CNS more directly. Body–brain signaling peptides such as ANG II activate central pathways by first acting on brain-associated structures lacking a blood–brain barrier, the sensory CVOs. Afferent signaling molecules that are less impeded by the blood–brain barrier, such as the steroid aldosterone, gain access directly to central neurons to affect osmoregulatory mechanisms by generating a salt appetite. The input arriving in the brain is conveyed into the visceral neuraxis, which is a neural network involved in neural integration of homeostatic information. The result of this central processing of information is the orchestration of a pattern of reflex and behavioral responses appropriate for the animal to deal with the prevailing internal and external conditions and thereby to maximize survival.

Different components of the visceral neuraxis have unique roles in maintaining body fluid homeostasis. Of particular importance are the neural structures lying along the lamina terminalis. The sensory CVOs of the lamina terminalis have been extensively implicated in sensing humoral factors, in particular extracellular osmolality or Na^+ concentration and ANG II. The MnPO processes information from the SFO and OVLT, along with other fluid-related inputs (**Figure 2**). Output from the lamina terminalis structures projects to key premotor nuclei influencing reflexes and behaviors controlling body fluid and cardiovascular homeostasis (**Figure 2**).

Recently, stretch- and ion-gated channels have been identified in cells of lamina terminalis structures, and emerging functional evidence has implicated these in various processes involved in osmoregulation. The development of new methods used to pursue the molecular and cellular nature of osmo- and Na^+ receptors holds great promise for advancing the

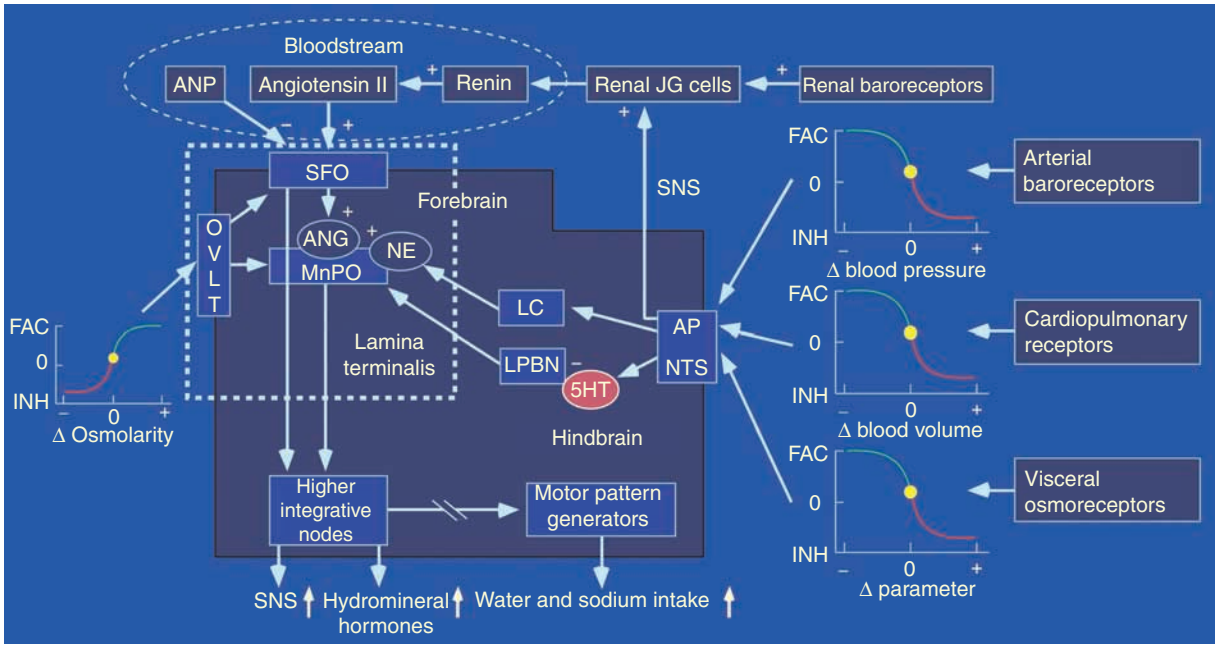

Figure 2 A schematic depicting the nature of neural and hormonal inputs into the brain and the central neural pathways that mediate sensory integration of signals for sympathetic and hormonal reflexes and behaviors (drinking (thirst) and sodium ingestion (salt appetite)) which reduce the consequences of dehydration and sodium loss and restore body fluid homeostasis. Both inhibitory and excitatory inputs from the periphery derive from arterial and cardiopulmonary baroreceptors and probably other visceral receptors (e.g., gastric, hepatic, portal, or renal). Information carried in afferent nerves projects mainly to the nuclei of the solitary tract (NTS). Circulating angiotensin II and other fluid-related hormones act on the subfornical organ (SFO) and the other sensory circumventricular organs – the organum vasculosum of the lamina terminalis (OVLT) and the area postrema (AP). Ascending information to the forebrain is carried in from noradrenergic cell groups (e.g., from the locus coeruleus (LC)) in the hindbrain which are activated by arterial and cardiopulmonary receptor input under conditions of hypotension and/or hypovolemia. Centrally generated angiotensin functioning as a neurotransmitter in certain central neural pathways and noradrenergic (NE) inputs act synergistically in the forebrain nuclei (e.g., those nuclei along the lamina terminalis). A hindbrain inhibitory pathway originating in the AP and medial NTS ascends to the lateral parabrachial nucleus (LPBN). This projection uses serotonin (5HT) to guard against excessive sodium and water intake, thereby limiting inappropriate expansion of extracellular fluid volume. Inhibitory input also ascends the neuraxis to act in the forebrain structures. ANG, angiotensin; ANP, atrial natriuretic peptide; FAC, facilitatory; INH, inhibitory; MnPO, median preoptic nucleus; SNS, sympathetic nervous system; JG, juxtaglomerular cells (source of renal renin). Reproduced from figure 11 in Johnson AK and Thunhorst RL (1997) The neuroendocrinology of thirst and salt appetite: Visceral sensory signals and mechanisms of central integration. *Frontiers in Neuroendocrinology* 18: 292–353, with permission from Elsevier.

understanding of the sensory and effector mechanisms that lead to the global process of osmoregulation.

See also: Blood Pressure: Baroreceptors; Circumventricular Organs; Fluid and Electrolyte Homeostasis: Clinical Disease; Food and Water Intake: Regulation; Neuromodulation of Sodium Channels; Salt Appetite; Sodium Channels; Thirst; Transient Receptor Potential (TRP) Channels.

Further Reading

Andersson B (1971) Thirst and brain control of water balance. *American Scientist* 59: 408–415.

Bourque CW (1998) Osmoregulation of vasopressin neurons: A synergy of intrinsic and synaptic processes. *Progress in Brain Research* 119: 59–76.

Brody MJ and Johnson AK (1980) Role of the anteroventral third ventricle region in fluid and electrolyte balance, arterial pressure regulation, and hypertension. In: Martini L and Ganong WF (eds.) *Frontiers in Neuroendocrinology,* vol. 6, pp. 249–292. New York: Raven Press.

Denton D (1982) *The Hunger for Salt*. New York: Springer.

Fitzsimons JT (1961) Drinking by rats depleted of body fluid without increase in osmotic pressure. *Journal of Physiology* 159: 297–309.

Fitzsimons JT (1998) Angiotensin, thirst, and sodium appetite. *Physiological Reviews* 78: 583–686.

Fluharty SJ (2002) Neuroendocrinology of body fluid homeostasis. In: Pfaff DW, Arnold A, Etgen A, Fahrbach S, and Rubin R (eds.) *Hormones, Brain, and Behavior*, pp. 525–570. Boston, MA: Academic Press.

Gilman A (1937) The relation between blood osmotic pressure, fluid distribution, and voluntary water intake. *American Journal of Physiology* 120: 323–328.

Honda K, Negoro H, Dyball RE, Higuchi T, and Takano S (1990) The osmoreceptor complex in the rat: Evidence for interactions between the supraoptic and other diencephalic nuclei. *Journal of Physiology* 431: 225–241.

Johnson AK, de Olmos J, Pastuskovas CV, Zardetto-Smith AM, and Vivas L (1999) The extended amygdala and salt appetite. *Annals of the New York Academy of Sciences* 877: 258–280.

Johnson AK and Gross PM (1993) Sensory circumventricular organs and brain homeostatic pathways. *FASEB Journal* 7: 678–686.

Johnson AK and Thunhorst RL (1997) The neuroendocrinology of thirst and salt appetite: Visceral sensory signals and mechanisms of central integration. *Frontiers in Neuroendocrinology* 18: 292–353.

Johnson AK and Thunhorst RL (2007) The neuroendocrinology, neurochemistry and molecular biology of thirst and salt appetite. In: Lajtha A and Blaustein JD (eds.) *Handbook of Neurochemistry and Molecular Neurobiology: Behavioral Neurochemistry, Neuroendocrinology and Molecular Neurobiology*, pp. 641–688. Berlin: Springer.

King MS and Baertschi AJ (1991) Central neural pathway mediating splanchnic osmosensation. *Brain Research* 550: 268–278.

Kraly FS, Kim YM, Dunham LM, and Tribuzio RA (1995) Drinking after intragastric NaCl without increase in systemic plasma osmolality in rats. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology* 269: R1085–R1092.

Liedtke W and Friedman JM (2003) Abnormal osmotic regulation in TRPV4-/-mice. *Proceedings of the National Academy of Sciences of the United States of America* 100: 13698–13703.

McKinley MJ, Denton DA, and Weisinger RS (1978) Sensors for antidiuresis and thirst: Osmoreceptors or CSF sodium detectors? *Brain Research* 141: 89–103.

Naeini RS, Witty MF, Seguela P, and Bourque CW (2006) An N-terminal variant of TRPV1 channel is required for osmosensory transduction. *Nature Neuroscience* 9: 93–98.

Noda M (2007) Hydromineral neuroendocrinology: Mechanism of sensing sodium levels in the mammalian brain. *Experimental Physiology* 92: 513–522.

Stricker EM and Hoffmann ML (2007) Presystemic signals in the control of thirst, salt appetite, and vasopressin secretion. *Physiology & Behavior* 91: 404–412.

Stricker EM, Hosutt JA, and Verbalis JG (1987) Neurohypophyseal secretion in hypovolemic rats: Inverse relation to sodium appetite. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology* 252: R889–R896.

Thrasher TN and Keil LC (1987) Regulation of drinking and vasopressin secretion: Role of organum vasculosum laminae terminalis. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology* 253: R108–R120.

Thunhorst RL, Lewis SJ, and Johnson AK (1994) Effects of sinoaortic baroreceptor denervation on depletion-induced salt appetite. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology* 267: R1043–R1049.

Toney GM, Chen QH, Cato MJ, and Stocker SD (2003) Central osmotic regulation of sympathetic nerve activity. *Acta Physiologica Scandinavica* 177: 43–55.

Verbalis JG, Mangione MP, and Stricker EM (1991) Oxytocin produces natriuresis in rats at physiological plasma concentrations. *Endocrinology* 128: 1317–1322.

Verney EB (1947) The antidiuretic hormone and the factor which determines its release. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 135: 27–106.

Otoacoustic Emissions

D T Kemp, Centre for Auditory Research, UCL EAR Institute, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Otoacoustic emissions (OAEs) are low-level sounds generated by the ear as a natural by-product of the hearing process. The sounds are recordable by means of an ear piece or probe containing a microphone, usually also a means of delivering sound stimulation to the ear. This probe is inserted into the ear canal so as to seal it as in Figure 1. The microphone signal is processed electronically to identify and extract sounds of cochlear origin usually during or following sound stimulation. Figure 2 shows an example of the cascade of sound pressure vibrations typically emanating from a human ear in response to a click stimulus together with its frequency spectra. The emission of a wide range of sound frequencies in response to appropriate stimuli indicates healthy functioning of a set of sensory cells in the cochlea known as the outer hair cells. OAEs also indicate functional integrity of the middle ear since it is the middle ear that conducts the stimulus into the cochlea and then conducts the secondary vibrations from the cochlea to the eardrum to create OAEs in the ear canal. The physical vibration of the eardrum needed to create typical OAEs is less than 0.1 nm, that is, smaller than atomic dimensions.

The outer hair cells of the cochlea (Figures 3 and 4) are the source of OAEs. These cells are uniquely capable of generating sound frequency vibrations. Excitation by mechanical deflection of the stereocilia that protrude from the apex of these cylindrical cells changes their receptor potential and in response the protein prestin in the cell's lateral membrane undergoes a configurational change. This results in a simultaneous length change in the isolated cell while *in vivo* it results in a force being applied to the stiff structure of the organ of Corti in which these cells are embedded (Figure 4). The force is naturally synchronized to the applied mechanical stimulation and this plays an essential role in optimizing the vibratory properties of the cochlea as described below. Normal operation of the mechano-electromotile action of outer hair cells depends critically on the composition and electrochemical status of the endolymph which fills the scala media and bathes the apical faces of both inner and outer hair cells. Inner hair cells also depend on the endolymph to perform their essential role of transducing the sound vibration reaching them and activating the auditory nerve, to initiate hearing. It is the leakage back to the middle ear of some of the vibratory energy generated by the outer hair cells which results in OAEs.

OAE therefore requires healthy status of the middle ear, outer hair cells, and cochlear endolymph. This makes them a valuable tool for the objective detection of a variety of peripheral auditory pathologies. The method is in common use to detect hearing impairment in neonates and for clinical diagnosis. In laboratory applications, emissions are used to monitor cochlear status and to record subtle changes.

Background

The OAE phenomenon was first reported in 1978. Initial observations were on human ears, but the phenomenon of OAE as it is commonly known was subsequently confirmed in a variety of animals. The first reports triggered a controversy which eventually merged with a major reevaluation of peripheral auditory function. At the time of the discovery of OAE, the dominant model of cochlear function was based on the work of the 1961 Nobel prize winner George von Bekesy. His pioneering observations in the 1940s on how sound stimulation vibrated cochlear structures led to the understanding that the spiral elastic basilar membrane of the cochlea (Figure 1(d)), loaded by the fluid filling the cochlea, performed an initial analysis of sound by frequency. This is achieved in a purely mechanical way by the natural formation of a frequency-dispersive, apically traveling, transverse wave. In Bekesy's model when sound energy is introduced into the cochlea by vibration of the stapes in the oval window, the traveling wave transports vibrational energy to specific regions along the tapered basilar membrane determined by the frequencies present in the sound. In this way, high notes become concentrated or focused on sensory cells near to the base, whereas low notes become focused on places near to the apex of the spiral organ as illustrated in Figure 5.

The focusing or concentration of incident sound energy of each frequency on small group of sensory cells serves to concentrate excitation and increase hearing sensitivity. The separation of frequency components across an extended sensory array prior to transduction is valuable for hearing because it permits the detection and subsequent simultaneous neural transmission of multiple frequency-specific channels of sensory information without mutual interference and degradation. This is exploited to maximum advantage in mammals and allows them to detect and perceive the information carried by sounds at frequencies far in excess of the maximum firing rate of auditory neurons. However, the efficiency of this mechanism is severely limited by the natural viscous damping occurring in the small fluid-filled spaces of

the cochlea. von Bekesy's and all subsequent observations of the ear's physical interaction with sound prior to 1978 appeared to show that in practice the process worked too inefficiently to yield the degree of frequency separation (resolution) required to explain the frequency-resolving properties of normal hearing (see Figure 6). An additional unknown mechanism was therefore presumed to act at or after the transduction stage so as to enhance frequency selectivity to the degree that was already known to exist from single-unit auditory nerve fiber studies.

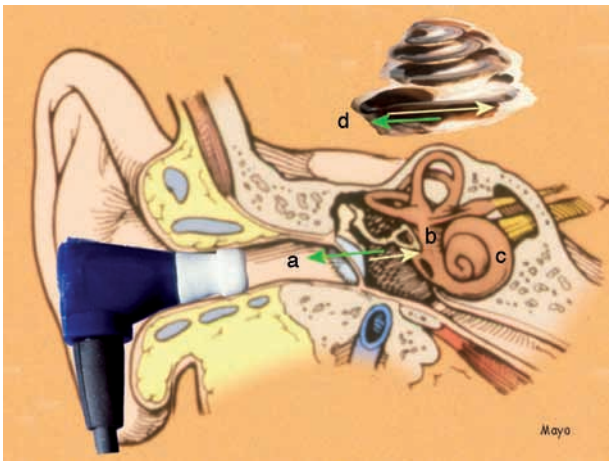

Figure 1 Otoacoustic emissions are recorded by a probe inserted into the ear canal (a) so as to seal it. Stimulation (yellow arrows) is conducted through the middle ear (b) and enters the cochlea (c). It progresses up the spiral basilar membrane of the cochlea (d) propagating in a traveling wave (see Figure 5). This wave mode directs energy at different frequencies to different places on the cochlea's sensory epithelium. The frequency resolution achieved in the process depends critically on the level of viscous damping. Damping is actively reduced by the electromotile action of outer hair cells (see Figure 3) and a by-product of this is the return of sound energy back to the eardrum after a delay. With thanks to M. Robinette and A. Pye for image material.

The discovery in 1978 of sound emissions by the intact healthy ear severely challenged Bekesy's model. The emitted sounds had nonlinear properties indicative of a biological origin. Exploiting this nonlinearity in suppression studies allowed the mechanical frequency tuning of their source to be determined and this was found to be similar to that of auditory nerves. But contemporary cochlear models viewed the cochlea as a passive absorber of sound energy, not an emitter, and they predicted poor frequency selectivity and linear behavior. OAEs on the other hand suggested that the cochlea performed frequency analysis much more efficiently than had been seen in the laboratory and that it did not simply absorb sound but also emitted it. The implications of these challenging observations took some years to be digested. They challenged most previous work on cochlear biophysics. The problem was that no physical mechanism was then known which could account for a reversal of Bekesy's apically traveling wave to cause sound emission into the ear canal. Furthermore, in some ears, sound emission was observed to occur spontaneously at precise frequencies without any stimulation being present at all. No biological mechanism then known had such a capability. This all led to an initial rejection of OAEs as being an epiphenomenon of no relevance to cochlear physiology.

Interestingly, Bekesy's model of the cochlea had been challenged as early as 1948 by Thomas Gold. He reasoned from physical principles that a biological vibration amplifier mechanism was necessary for the

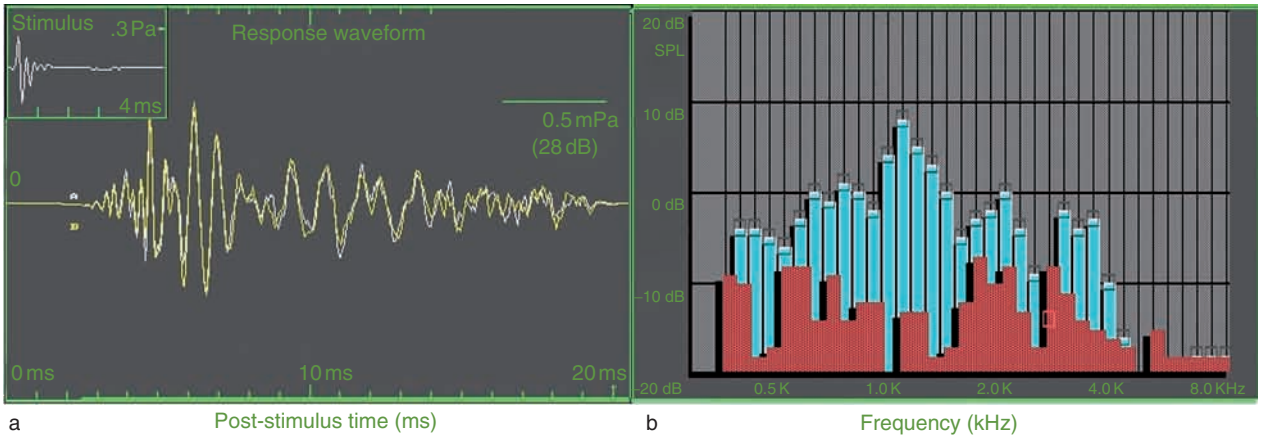

Figure 2 Transient-evoked otoacoustic emission (TEOAE). (a) Left, the sound pressure waveform recorded in a sealed human adult ear canal following a brief click stimulus (top left). The oscillations in sound pressure are the result of vibrations of the eardrum caused by a mechanical response of outer hair cells within the cochlea. The initial 3 ms is blanked out because the recording is overloaded by the click stimulus itself. An eighth octave frequency analysis of this waveform is shown on the right. The upper blue columns indicate the sound pressure level of otoacoustic emission at each frequency for this ear. Both the waveform shape and the spectral intensity structure are highly individual but normally very stable. Sensory hearing loss at any frequency usually abolishes the emission of that sound frequency.

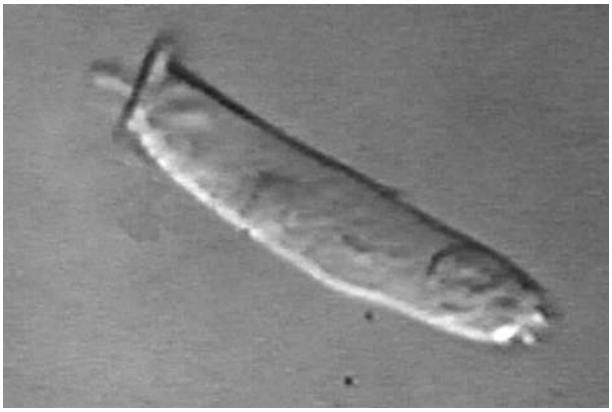

Figure 3 The source of otoacoustic emissions is vibrations from the outer hair cells of the cochlea. In the isolated cell shown, the protruding bundle of stereocilia is just visible on the top left. *In vivo* displacement of the stereocilia by fluid vibration in the excitatory direction opens ion channels in their tips causing a current flow which depolarizes the cell. This causes the cell to shorten. The change is very rapid and generates audio frequency vibration at the stimulating frequency and at distortion frequencies. The electromotility can be demonstrated *in vitro* by applying alternating potential to the cell. Prepared by David Brass.

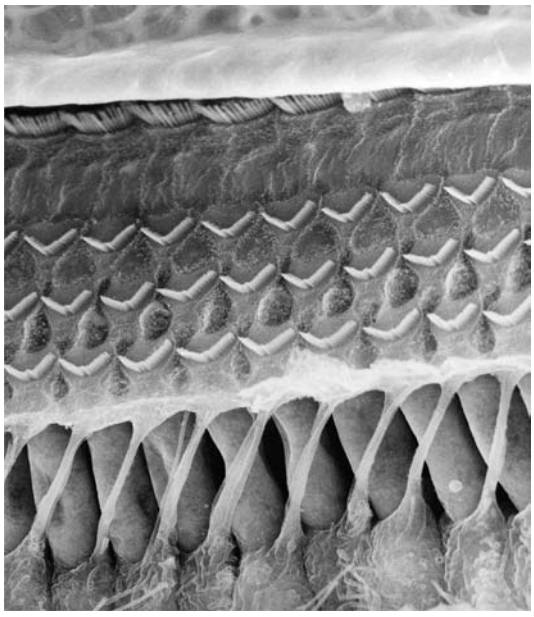

Figure 4 The sensory hair cells of the cochlea. In the foreground of this longitudinal section of the organ of Corti, the outer hair cell bodies are seen embedded inside the structure. The V-shaped arrays of stereocilia belonging to three rows of outer hair cells are seen above protruding from the apical surface of the organ of Corti. The hairs of the inner hair cells are just visible beneath the tectorial membrane at the top. Hearing occurs only via the inner hair cells which directly synapse onto auditory nerve fibers. Electromotility of the three rows of outer hair cells act as a biological hearing aid, amplifying weak stimulation. Electron micrograph courtesy of A. Forge.

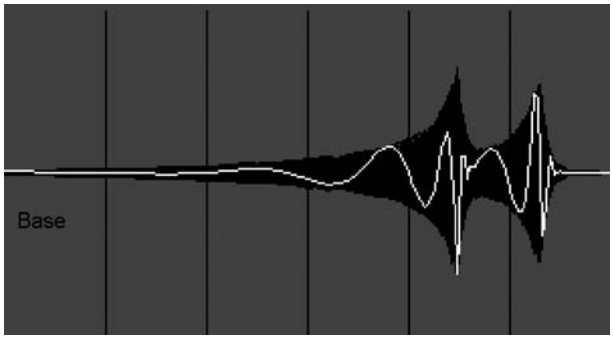

Figure 5 Schematic illustration of the cochlear traveling wave in response to two pure tone frequencies. The dark shaded region indicates the envelope of the wave over time, while the white line indicates (greatly exaggerated) the instantaneous displacement imposed by the wave on the basilar membrane at one arbitrary instant in time. The wave moves from left (base) to right. In this 'low damping' model, there is a sharp peak in the overall envelope for each of the two different frequencies where excitation is maximal for each frequency.

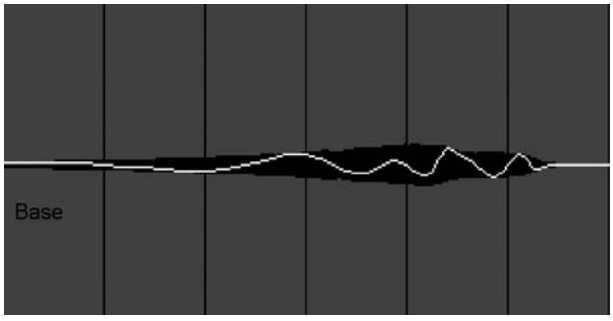

Figure 6 From the same computer model as in Figure 5 and with the same input stimulation, this figure illustrates how the frequency peaks of traveling wave intensity decrease and merge together when substantial viscous damping is present. The traveling waves in this figure represent those in a dead cochlea or one in which outer hair cell motile activity is missing. This corresponds to an ear with sensory transmissive hearing loss.

ear to achieve its remarkable sensitivity and frequency selectivity but that such an active mechanism would obviously not have been operating in the damaged or dead specimens studied in the laboratory. Gold's 1948 model was that biologically derived electrical energy was converted into vibratory energy by some unknown piezoelectric mechanism in synchrony with the incoming sound vibration and was then coupled back into basilar membrane motion so as to replace the energy lost by the viscous drag of the cochlear fluid. With this 'positive feedback,' the cochlea would be able to perform its mechanical frequency separation and focusing role much more efficiently – turning the dull response in Figure 6 into the sharp response of Figure 5.

Gold's hypothesis was generally discounted but the discovery of OAE triggered fresh uncertainty about

the degree to which frequencies where physically separated by Bekesy's traveling wave. This uncertainty was ended in 1982 by new and refined measurements of vibration in undamaged animal cochleae made by Sellick, Patuzzi, and Johnstone. The sharp mechanical frequency tuning they observed in the carefully prepared cochlea matched that seen in auditory nerve fibers. This meant that the healthy undamaged cochlea was somehow able to avoid the previously observed effects of viscous damping, just as Gold had postulated. The discovery in 1983 of outer hair cell electromotility by Brownell provided a suitable mechanism to implement Gold's positive feedback and to generate OAEs. The stimulus-activated electromotility of the outer hair cells is now believed to more than replace the energy lost through viscosity – not exactly by positive feedback as Gold envisaged but by a more subtle interaction with basilar membrane mechanics. This new aspect of cochlea's biophysical function became known as the cochlear amplifier. The cochlear traveling wave amplifier provides up to a 1000 times (60 dB) increase in the level of excitation reaching the sensory inner hair cells over and above that demonstrable in the dead cochlea. OAEs are the result of some of this added energy escaping back to the middle ear, that is, a by-product of the cochlear amplifier.

The output of the cochlear amplification saturates with increasing stimulus intensity so that the gain it provides reduces. This has the effect of reducing the dynamic range of stimulation from over 100 dB (i.e., more than 100 000:1) at the outer ear to maybe 50 dB (~300:1) at the inner hair cells and is valuable in two ways. Weak stimulations receive most amplification to raise them above the relatively high threshold of inner hair cells whereas strong stimuli need no amplification to do that. Also, auditory nerves have a dynamic range in terms of firing rate of only around 40 dB (100:1), and so the compression provided by the nonlinearity of the cochlear amplifier helps the auditory nerve accommodate the full range of environmental sound intensities. The level of stimulus frequency OAE (SFOAE), which is the dominant part of transient-evoked OAE (TEOAE) and SOAE (spontaneous OAE, see below), exhibits a compressive nonlinearity with stimulation level following an approximately 1/3 power law.

OAEs and Hearing Loss

Hearing impairment can take many forms, most of which include an elevation of hearing threshold, that is, a decrease in sensitivity to sound. Sensitivity can be lost due to poor transmission of sound energy though the middle ear to the cochlea (conductive loss), poor transmission and frequency analysis of sound energy through the structures of the cochlea onto the sensory inner hair cells (sensory transmissive loss), poor sensory transduction (sensory transductive loss), or poor transmission of transduced information along the auditory pathways (neural transmissive loss). There are important dimensions to hearing impairment beyond that of simply losing sensitivity to quiet sounds. These arise from degradation of the information extracted from the sound stimulation. Whereas a finite loss of sensitivity can always be completely corrected for by applying external amplification, a loss of information caused, for example, by inadequate frequency separation in the cochlea, by defective transduction, or by impaired neural transmission cannot be corrected for externally. It is therefore of import to ascertain the nature of the loss before prescribing amplification.

Determining whether OAEs are present or absent helps define the nature of a hearing loss. They are absent in the case of sensory transmissive hearing loss due to the failure of the cochlear amplifier through a loss of the electromotile action of outer hair cells. When this happens, the distribution of stimulus sound energy within the cochlea reverts to that seen by Bekesy in dead ears – notably a damping of the traveling wave leading to poor frequency-specific focusing of stimulation onto the sensory hair cell as in Figure 6. There is a loss of sharpness and consequently of the intensity delivered to these cells (compare Figures 6 and 5), and also there is an increase in dynamic range. Loss of outer hair cell function has been shown experimentally to cause up to a 60 dB (50%) loss in hearing. Decreased dynamic range contributes to the clinical symptom of recruitment, and the loss of frequency resolution contributes to poor hearing in noise.

OAEs are absent in sensory transmissive hearing loss because no vibrational energy is being added to enhance the excitatory traveling wave. The prevalence of OAE loss in the entire hearing impaired population with normal middle ear function is very high, leading to the conclusion that loss of outer hair cell function is the major cause of, or is a necessary concomitant of, nearly all nonconductive hearing losses. This empirical fact justifies the widespread use of the technique for newborn hearing screening in those infants without risk of neurological damage and identifies outer hair cells as the most vulnerable component of the auditory system – a fact first learned histologically.

Methods Used to Observe OAE

OAEs are generated when a stimulus of any kind enters the cochlea, but for the purposes of measurement specific types of acoustic stimuli are employed.

Different stimuli emphasize different aspects of outer hair cell function and they also require different signal processing to identify and separate the emission from the incident stimulus.

In the TEOAE technique, a short click sound is delivered to the ear and the probe microphone records ear canal sounds for up to 20 ms afterward (Figure 2). During the first 10 ms in humans and a shorter time in most animals, the stimulation will have propagated as a traveling wave along the entire length of the cochlea. In doing so, the wave will have activated outer hair cells and been strengthened by their electromotile reaction – enhancing the traveling wave and ensuring sharp focusing of each frequency component onto its own frequency place. Natural anatomical or functional irregularities in the system lead to the enhanced traveling wave being partially reflected back to the middle ear and ear canal and so creating an OAE.

The click intensity used to record TEOAEs is typically 50 dB above threshold with a peak intensity of 80–85 dB sound pressure level (SPL). Repeated click stimulation and signal averaging is used to overcome background noise and recover the emission which in human ears ranges from 0 to 30 dB SPL. To avoid including the stimulus or the acoustic response of the middle ear in the measurement, the recording is delayed until several milliseconds after the stimulus. Even so, there is a risk of acoustic oscillations in the ear canal being confused with the cochlear emission. To overcome this, a differential technique is employed in which responses to different intensities of stimulation are subtracted in proportion so as to cancel any linearly proportionate acoustic responses. This leaves only nonlinear signals which are most likely to be of biological origin. TEOAE responses are then analyzed by frequency in order to confirm outer hair cell function along the length of the cochlea. Responses have been obtained between 500 Hz and 8 kHz but the TEOAE technique is most effective over the 1–4 kHz range.

The complex waveform of the response seen in Figure 2 occurs because the various frequency components making up the click take different times to enter the cochlea, to travel to their natural place, and then after reflection to travel back to the middle ear. The delay is frequency dependent and this results in a time dispersion of response components. High frequencies emerge first and low frequencies last. The intensity of each component is irregularly dependent on frequency. Both the delay and intensity profiles are individual and tend to be very stable with time. The TEOAE waveform can be exactly reproduced weeks or even years later, provided there is no damage or deterioration in the cochlea or middle ear. TEOAE is therefore a very sensitive indicator of physiological change in the peripheral auditory system. The technique has been successfully used on other primates and with somewhat greater difficulty on a wide range of other animals including frogs. It is often the method of choice for newborn hearing screening and for research with human subjects, but is rarely used in laboratory research.

A second commonly used method of OAE observation involves measuring distortions produced during stimulation with two pure tones approximately one-third an octave apart. The significance of this separation is that the individual frequency places in the cochlea for such a stimulus just overlap and so compete for the responses of some sensory cells. Due to their nonlinearity, outer hair cells in places which provide amplification to both stimulus frequencies are unable to do so without generating distortion also. This is because the transfer function from mechanical stimulation to amplified mechanical output is nonlinear and saturates at quite low levels. The mechanical force these cells impose on the basilar membrane does not exactly mimic the applied stimulus waveform. Vibrations at new frequency components are generated. For stimulus frequencies f_1 and f_2, nonlinearity of the hair cell's response introduces harmonics distortions at $2f_1$, $3f_1$, $4f_1$, etc., and $2f_2$, $3f_2$, $4f_2$, etc., and also intermodulation distortion products at frequencies of $2f_1 - f_2$, $3f_1 - 2f_2$, $4f_1 - 3f_2$, etc., and $2f_2 - f_1$, $3f_2 - 2f_1$, $4f_2 - 3f_1$, etc. Despite their algebraic complexity, intermodulation distortion actually comprises a simple array of tones, each separated by the stimulus difference frequency $f_2 - f_1$ (f_1 less than f_2) which surround the applied stimuli. It needs to be emphasized that although all these distortion products might be generated at the cell only a few of them escape to form OAEs.

Those distortions that reach the ear canal form distortion product OAEs or DPOAES. DPOAE can be extracted by subjecting the ear canal probe microphone signal to narrow-band frequency analysis and this is illustrated in Figure 7. Because the frequencies of the emitted sounds are absolutely defined by (but are different from) the applied frequencies, any particular distortion product can be readily identified and its intensity measured. DPOAE levels range from −10 up to 30 dB SPL. The measurement of DPOAE is practically limited by the level of background noise present in the duration of the recording. There is also a risk of noise and artifactual distortion generated by the sound delivery and recording systems being mistaken for OAEs. This risk is minimized by making control recordings in an artificial ear canal and by statistical assessment of the signal-to-noise ratio. DPOAEs characteristically emerge after

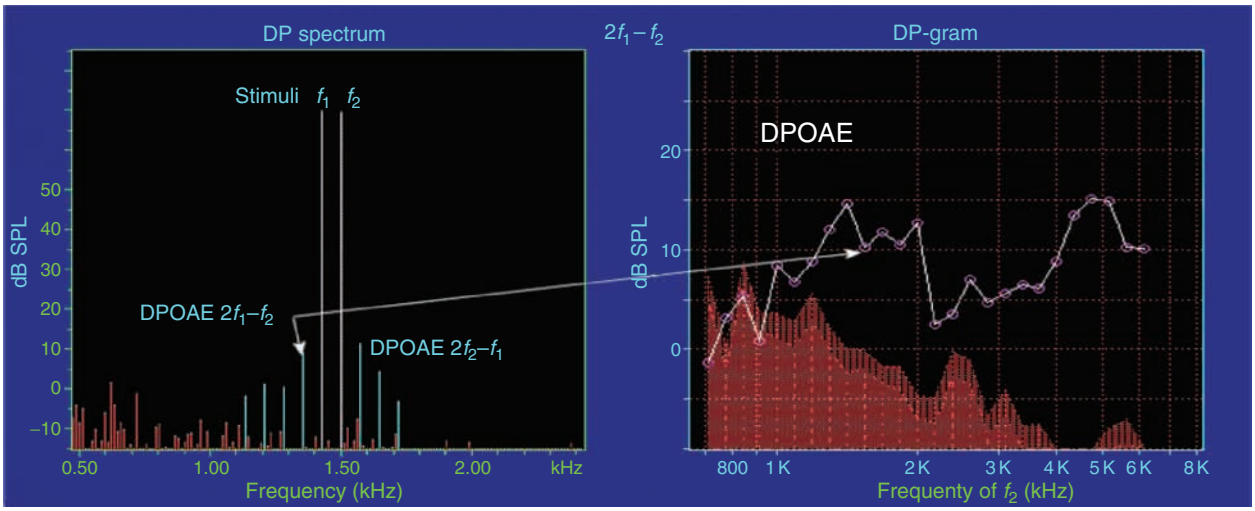

a b

Figure 7 An example of distortion product otoacoustic emissions. (a) Shows the spectrum of sounds in a human ear canal during stimulation by two pure tones $f_1 = 1420$ Hz and $f_2 = 1500$ Hz, each at 70 dB SPL. The additional blue lines are intermodulation distortion products generated by the nonlinear activity of outer hair cells. The most commonly measured component is at the frequency given by the formula $2f_1 - f_2 = 1360$ Hz in this case. (b) A DP-gram formed by plotting the level of $2f_1 - f_2$ for a series of f_1, f_2 frequency steps across the audio frequency range. The irregularity is quite typical and does not imply a similar fluctuation in hearing. The area shaded red shows the 95 percentile limits of noise recorded during the measurement. In this example, the ratio f_2/f_1 is 1.05, which is well below the optimum value for $2f_1 - f_2$ generation. It has been chosen because this close ratio allows the maximum number of distortion products to escape. This reveals the systematic arrangement of distortion components.

a delay of several milliseconds, and this feature can also be used to discriminate between cochlear and artifactual signals.

For a particular distortion product to emerge from the cochlea to form a DPOAE, there has to be mechanism for its basal propagation. Those with distortion frequencies significantly above the originating place frequency, for example, $2f_1$ and $4f_2 - 3f_1$, cannot propagate at all and also cannot be perceived as their intensity is so much lower than that of the causative stimuli. Those distortions with frequencies lower or slightly above the natural frequency of their place of origin, for example, $2f_1 - f_2$ ($f_1 < f_2$), naturally propagate apically, and these can be perceived as aural combination tones. However, lower-frequency intermodulation products can escape back to the middle ear in at least two ways. First, for values of f_2/f_1 near to 1.2, the spatial distribution and phasing of the $2f_1 - f_2$ distortion produced by each hair cell naturally favors its propagation back to the middle ear which then results in DPOAEs. The ratio 1.2 is the optimum ratio for the creation of DPOAE $2f_1 - f_2$. Different f_2/f_1 ratios favor other lower-frequency distortion emissions. Second, any lower-frequency distortions propagating apically from their site of origin to there own frequency places can be reflected back to the middle ear by the same irregularities which are responsible for the creation of TEOAE and SOAE. Figure 8 illustrates how two stimulus tones introduce a variety of different waves in the cochlea and how they escape to form OAEs.

DPOAEs are a complex phenomenon originating in the mechanical nonlinearity of outer hair cells spread over a significant length of the basilar membrane and emerge by a dual reverse transmission path, which can cause spurious fluctuations of DPOAE intensity with frequency through interference. The intensity of the emission also depends greatly on the stimulus level and on the intensity and frequency ratios of the two stimulus components. Despite this, the technique is valuable in the laboratory and clinic as an indication of outer hair cell function. Typically, measurements of the $2f_1 - f_2$ DPOAE are made over a sequence of f_1, f_2 frequencies and combined to form a DP-gram as shown in Figure 7(b). The method works well up to high frequencies. Stimulation by two equal-intensity stimuli f_2 and f_1 at levels above 70 dB SPL yields DPOAEs related to outer hair cell integrity, providing a psuedo-cochleogram. DPOAEs obtained using lower intensity and unequal stimuli, for example, f_2 55 dB SPL and f_1 65 dB SPL, are more dependent on the functional quality of these hair cells and indicate normal hearing threshold.

A third method of observing OAEs which is mainly employed in specialized research applications involves the application of just one pure tone to the ear. OAEs are created only at the applied frequency due to wave reflection at cochlear irregularities.

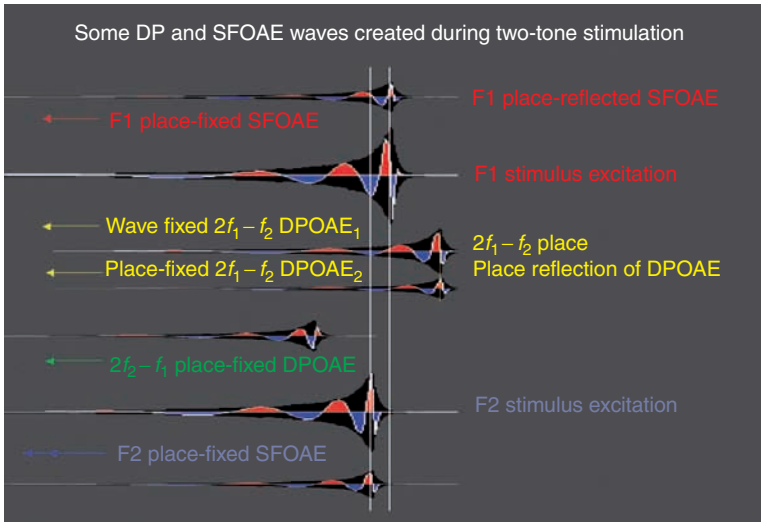

Figure 8 Schematic illustration of the cochlear waves that result in the major otoacoustic emissions in response to a two-tone stimulus comprised of frequencies f_1 and f_2. The base (input) and higher-frequency region of the cochlear is left, the apex and lower frequency is right. Stimulation waves move left to right. OAE-generating waves move right to left. Emission waves caused by reflection at a place are called place-fixed emissions. Emissions arising from wave interactions are called wave-fixed or distortion emissions. The most commonly used OAE is the $2f_1 - f_2$ DPOAE which can be seen to have two sources. TEOAEs are composed mostly of SFOAE with some DPOAE.

Separation of the SFOAE from the applied stimulation requires careful processing. As with TEOAEs, a differential technique can be employed whereby responses to high- and low-level stimuli are scaled and subtracted so as to leave only the nonlinear part of the ear canal sound. Alternatively, a second stronger tone is introduced at a nearby frequency to divert outer hair cell activity and so suppress the emission in response to the first tone. Again, a subtraction method serves to remove the unaffected stimulus signal. SFOAEs can be used to monitor cochlear status with time at one selected frequency or they can be measured over a range of frequencies by sweeping the stimulus frequency. The method has the advantage of a greater simplicity of data interpretation but for the most part gives no more information than the TEOAE and DPOAE techniques.

Other techniques need to be mentioned briefly. Observation of SOAEs requires only the narrowband frequency analysis of the ear canal sound. They consist of one or more continuous oscillations at very narrowly defined frequencies – typically of less than 1 Hz bandwidth (see Figure 10). SOAEs are due to the repeated self-stimulation of the cochlea by closed-loop reflections of vibration between the middle ear and the same internal reflecting irregularities that give rise to TEOAEs and SFOAEs.

Each of the stimulated emission techniques reviewed above has variants. TEOAEs, for example, can be derived from stimuli other than clicks and both tone bursts and pseudo-random noise (maximum length sequences) have been successfully employed. DPOAEs can be obtained simultaneously from multiple tone pair stimuli, and also from modulated stimuli. Both techniques have been used for research purposes. DPOAEs have also been obtained in the presence of an additional suppressing tone intended to reduce the reflection component so as to simplify interpretation.

Interpretation of OAEs

As indicated above, the ear emits sound because the healthy cochlea contains an active wave amplifier. Some of the amplifier's energy escapes due to the creation of distortions by nonlinearities in the amplification mechanism. This escape occurs from locations where the spatial phase configurations of the distortion naturally propagate waves back to the middle ear. Other energy escapes because the forward wave propagation characteristics of the basilar membrane and/or the amplification that is present along its length exhibits irregularities (in primates at least), which cause the reflection of both stimulus and distortion waves back to the middle ear. Large numbers of studies in human ears have confirmed that ears with healthy middle ear function but no OAE have a sensory hearing loss typically of 30 dB or more.

OAEs are broadly frequency specific, because failure to exhibit an emission over at least a half-octave frequency range (given suitable stimulation) correlates with hearing loss over a related frequency range. With TEOAEs and SFOAEs, the frequency of hearing loss is the frequency at which OAEs are absent. For $2f_1 - f_2$ DPOAEs, it is the frequency of

f_2 which determines the place in the cochlea being examined, whereas the emission response may emerges about one-half octave below that frequency if the optimum stimulus frequency ratio is used.

However, a direct correspondence of the frequencies of emission loss and hearing loss cannot be totally relied upon. A fine structure of OAE intensity with frequency can exist due to interference between multiple exit routes (see **Figure 8**). Peaks and dips of intensity on a scale smaller than one-half octave are not generally functionally significant. Furthermore, although it is tempting to assume a one-to-one correspondence between frequency and place in the cochlea since, as this is the basis of its function, it is not the case that all emissions originate from the appropriate frequency place. Wave and cell interactions occur over a substantial length of the basilar membrane even for single stimulus frequencies and these contribute to the OAE. Emissions should be regarded as a somewhat spurious by-product of the cochlear amplification and filtering process.

It is also tempting to assume that the intensity of an emission at a given stimulation level, or the lowest stimulus level at which an emission can be detected, will correspond in some way to the sensitivity of hearing. This is partly true. OAE intensity reduces systematically as hearing loss progresses. Many population studies have shown about a 60% correlation between measures of emission strength and hearing threshold sensitivity. This is sufficient to demonstrate the underlying relationship but is insufficient to allow a meaningful prediction of an individual's hearing threshold from OAE measurements. This is not surprising since the cochlear amplifier mechanism feeds but does not involve the transduction process. It is the transduction process that determines the threshold for excitation of the auditory nerve. Furthermore, the transmission of waves back to the middle ear to form OAEs involves the fortuitous escape of energy via mechanisms not related to the hearing process. Even the recording process affects the intensity of emissions recorded. Vibrations of the eardrum driven by waves in the cochlea create measurable emission sound pressures only when the ear canal is sealed by the measurement probe. The volume of air between the probe and eardrum then plays a significant role in determining the sound pressure developed. These cautions raise the following question: what value are OAEs beyond their use for the detection of a loss of outer hair cell function?

The Wider Significance and Importance of OAE

Outer hair cells are known to be the most vulnerable components of the auditory system and their malfunction is implicated in the vast majority of sensory hearing loss pathologies. The stability of hearing sensitivity depends on the mechanisms ensuring the functional homeostasis of the cochlea. The nature of outer hair cells vulnerability and the mechanisms that maintain their optimal function are still inadequately understood. OAEs provide data specific to the functional status of outer hair cells in the intact ear and facilitate direct comparisons between human and animal cochleae. So although quantitative interpretations can be difficult, OAEs have a uniquely important role in auditory research at this time.

The functional mechano-electromotile status of outer hair cells is key to hearing. It can be characterized *in vitro* by their input–output function and operating point. This saturating nonlinear characteristic can be used to predict the pattern of distortion product generated locally but the complexity of the distortion product emission mechanism (see **Figure 9**) prevents the unambiguous determination of the average hair cell input–output characteristics at one frequency region from static DPOAE measurements. On the other hand, changes in OAE level can be readily measured and related to functional change. It is relatively easy to demonstrate the effects of noise and ototoxic trauma and potentially the effects of otoprotective drugs using DPOAE and TEOAE. Long exposures to intense noise can cause permanent hearing loss, abolition of OAE, and even hair cell loss. Lesser exposures cause a temporary threshold elevation and a correlated depression of OAEs. Very short duration exposures (~1 min) can trigger a biphasic alteration in both threshold and OAE which incorporates an enhancement phase. Challenge by even mild overexcitation reveals the characteristics of cochlear control mechanisms (see **Figure 10**).

Changes in OAE intensity can be interpreted as either due to a change in outer hair cell input–output functions or in their operating point. The latter can be experimentally altered by the application of a more intense low-frequency tone, which by delivering a modulated mechanical bias to each hair cell modulates their operating point and consequently modulates OAEs. From this modulation, hair cell characteristics can be deduced. Some pathologies – specifically cochlear hydrops – cause hearing loss by fluid pressure imbalance within the cochlea, and the offset of hair cell operating point can be observed using OAE.

There is extensive evidence for some central control of cochlear status although its function is not well understood. In particular, fibers from the medial olivary complex in the auditory brain stem synapse directly onto outer hair cells and so can alter their operating point and maybe their dynamic function. Excitation of the medial cochlear efferent system has

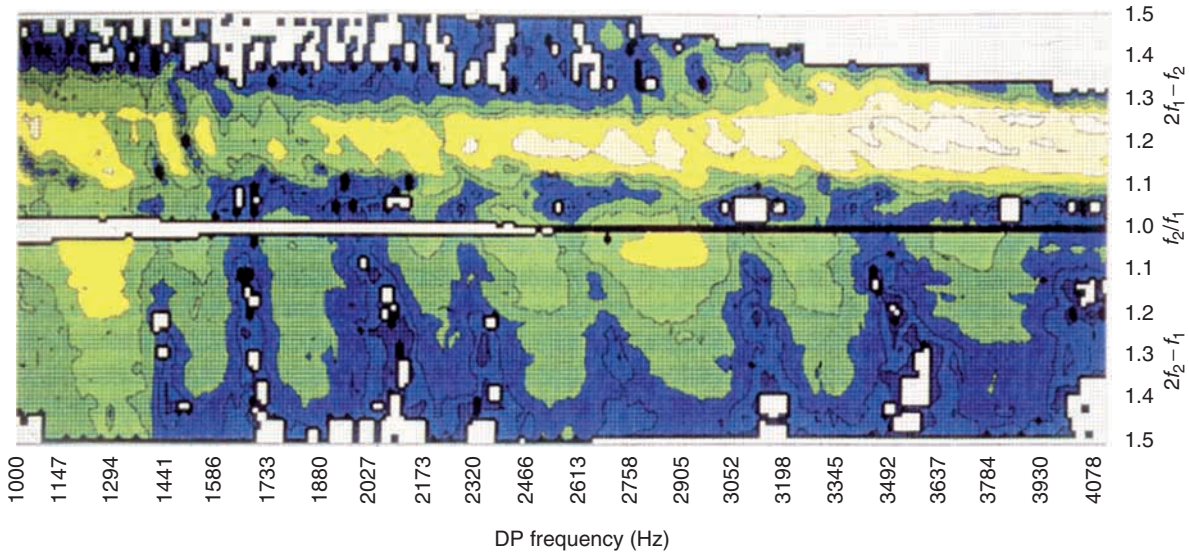

Figure 9 A DPOAE map of a human cochlea showing a detailed individual structure. This ear has normal hearing. The structure is the result of the complexity of the emission process and interference phenomena. Its broader features probably are related to functionally minor irregularities in cochlear anatomy and physiology which may or may not be predictive of the ear's robustness and genetic factors. The map shows DPOAE intensity against its frequency and stimulus pair frequency ratio obtained by recording DPOAEs at one stimulus pair level for all stimulus frequencies and frequency ratios. Upper and lower panels represent different OAEs.

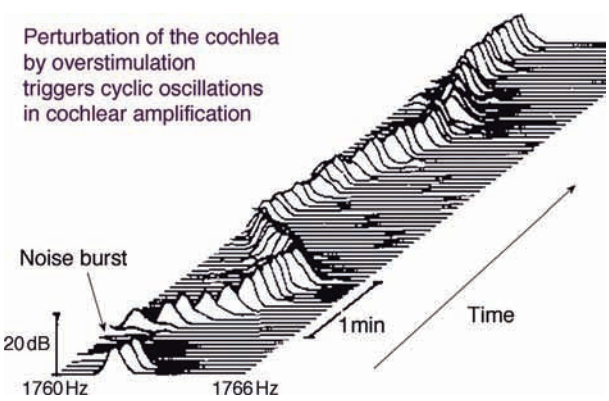

Figure 10 An example of the cyclic change in cochlear amplification following brief overstimulation as reflected in the intensity and frequency of a spontaneous otoacoustic emission (SOAE).

a small suppressive effect on hearing and on OAEs. As a large proportion of these fibers are activated by input from the contralateral ear, this suppression of activity (called the cochlea reflex) can be most easily demonstrated by applying noise or other fluctuating sound stimulation contralaterally and observing the depression of OAE. This occurs over a timescale of the order of hundreds of milliseconds. There is experimental evidence that a strong cochlear reflex correlates with low susceptibility to overstimulation.

OAEs are a powerful tool for researching the mechanisms that maintain functional cochlear homeostasis. Abnormal malfunction of these mechanisms might be early indicators of pathology or be predictive of individual hypersusceptibility. The absolute intensity of OAEs has not proved to be a valuable measure in itself but the magnitude and time course of changes in OAE intensity seen in response to experimental perturbations of the auditory system are uniquely valuable.

See also: Cochlear Mechanics; Cochlear Development; Deafness; Hair Cells: Sensory Transduction; Hair Cell Differentiation.

Further Reading

Brownell WE, Bader CR, Bertrand D, and de Ribaupierre Y (1985) Evoked mechanical responses of isolated cochlear outer hair cells. *Science* 227: 194–196.

Gold T (1989) Historical background to the proposal 40 years ago, of an active model for cochlear frequency analysis. In: Wilson JP and Kemp DT (eds.) *Cochlear Mechanisms: Structure, Function and Models*, pp. 299–306. London: Plenum.

Hall JW, III (1999) *Handbook of Otoacoustic Emissions, Thomson Learning, USA. A Singular Audiology Text*. San Diego, CA: Singular Press.

Kemp DT (1978) Stimulated acoustic emissions from within the human auditory system. *Journal of the Acoustical Society of America* 64: 1386–1391.

Knight RD and Kemp DT (2001) Wave and place fixed DPOAE maps of the human ear. *Journal of the Acoustical Society of America* 109: 1513–1525.

Manley G and Popper A (2007) Active processes and otoacoustic emissions. Springer Handbook of Auditory Research. New York: Springer.

Moore BCJ, Ashmore JF, and Haggard M (eds.) (2002) New developments in hearing and balance. *British Medical Bulletin 63*. New York: Oxford University Press.

Robinette M and Glattke TJ (2007) *Otoacoustic Emisions: Clinical Applications,* 3rd edn. New York: Theime.

Sellick PM, Patuzzi R, and Johnstone BM (1982) Measurement of basilar membrane motion in the guinea pig using the Mossbauer technique. *Journal of the Acoustical Society of America* 72(1): 131–141.

Shera CA and Guinnan JJ (1999) Evoked otoacoustic emissions arise by two fundamentally different mechanisms. *Journal of the Acoustical Society of America* 105: 782–798.

Oxidative Damage in Neurodegeneration and Injury

I G Gazaryan and R R Ratan, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Published by Elsevier Ltd.

Introduction

The etiology of dysfunction and cell death in neurodegenerative diseases remains enigmatic. Most neurodegenerative conditions are insidious in onset and run a gradually progressive and painfully inexorable course for those afflicted and their families. Prototypic neurodegenerative conditions include Alzheimer's disease (AD), Huntington's disease (HD), Parkinson's disease (PD), and amyotrophic lateral sclerosis (ALS). By contrast, acute neurodegenerations are monophasic and characterized by stroke and spinal cord injury. Many neurodegenerative conditions are linked to inherited mutations, resulting in accumulation of the damaged proteins or their wrongly processed variants. Sporadic neurodegenerative conditions are also associated with accumulations of misfolded proteins and their aggregates. Downstream of these events, metal-catalyzed oxidation leading to oxidative stress has been implicated as a common final pathway of injury (Figure 1).

Aging and Neurodegeneration

Mitochondria as a Major Source of Superoxide under Normal Conditions

Neurodegeneration is directly modified by cell aging. Some theories invoke the mitochondria as the major sites of generation of deleterious free radicals that promote aging. The chemistry of reactive species that could result in oxidative damage under normal physiological conditions is the same for any cell. The major species in cells is the superoxide radical, produced as a by-product of the mitochondrial respiration, in particular by complex I and complex III of the respiratory chain. Complex III breaks a two-electron reducing equivalent into two single-electron reducing equivalents inside the membrane (in what is known as the Q cycle), and the resulting quinone radical reacts with oxygen, giving rise to the superoxide radical. This side reaction is known as 'mitochondrial leakage' and comprises up to 3–4% of the total oxygen consumption. If in a lifetime we consume nearly 60 000 l of oxygen per kilogram of body weight, it means that we produce approximately 2000 l of superoxide per kilogram of body weight. What happens to the released superoxide and how does a cell handle the consequent 'oxidative' load?

Antioxidant/Antiaging Mechanisms: Superoxide Scavenging

In both mitochondria and the cytosol, superoxide is scavenged and converted into hydrogen peroxide and oxygen by superoxide dismutase (SOD), although the nature of SOD in mitochondria and the cytosol is different (i.e., MnSOD and CuZnSOD, respectively). The formed hydrogen peroxide is further decomposed into water with the help of catalase or is reduced by the glutathione peroxidase/glutathione system. Very recent studies show that the thioredoxin reductase/thioredoxin system is also capable of reducing hydrogen peroxide to water.

There are a number of theoretical reasons why the nervous system in general and neurons in particular are under the biggest load of oxidative stress under basal conditions. Neurons have a very high metabolic rate in order to meet the demands of electrical signaling. Specifically, our brain uses 20% of the total oxygen consumed, although the brain is only 2% of the body weight. An expected corollary of high oxygen utilization is a tenfold higher level of radical production. Paradoxically, despite higher radical production, catalase is absent in brain mitochondria. This is not the case for heart mitochondria.

Given the higher level of radical production and the absence of catalase in brain mitochondria, what are the mechanisms at play inside neuronal mitochondria to neutralize the formed hydrogen peroxide and keep oxidative damage to tolerable levels? In addition to glutathione peroxidase, the only other known system that neutralizes hydrogen peroxide is that of thioredoxin and peroxiredoxin. Reduced thioredoxin provides peroxiredoxin with reducing equivalents to reduce hydrogen peroxide to water. Oxidized thioredoxin is reduced back by a mitochondrial thioredoxin reductase, selenocysteine flavodithiol oxidoreductase. Why mitochondrial catalase is absent in neurons is not well understood, but could be related to neuronal signaling functions of peroxide that are not necessary in other tissues. The absence of catalase from neuronal mitochondria places the burden of neutralizing hydrogen peroxide generated in mitochondria squarely on the shoulders of the thiol-based detoxification systems, particularly glutathione, thioredoxin, and peroxiredoxin. A causal role for mitochondrial peroxide in aging has been supported by the generation of transgenic mice that overexpress catalase in the nucleus, the peroxisome, and the mitochondria. Significant extension of life span was only seen in the animals in which catalase was overexpressed in the mitochondria. The extension of life span was associated with improvement

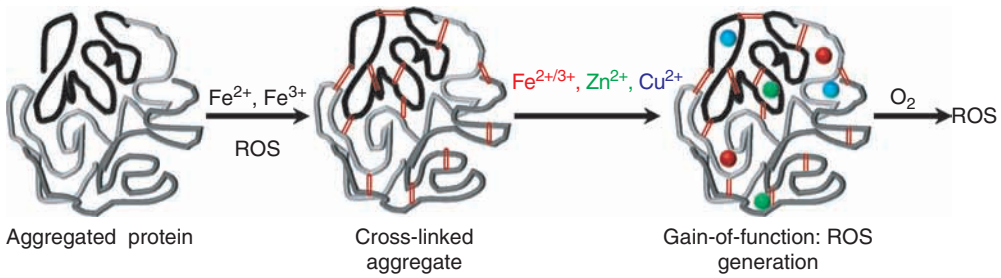

Figure 1 Widely accepted hypothetic gain-of-function transformation of cross-linked aggregates of misfolded proteins. Confirmed for AD, PD, and ALS, and supposed for HD. ROS, Reactive oxygen species.

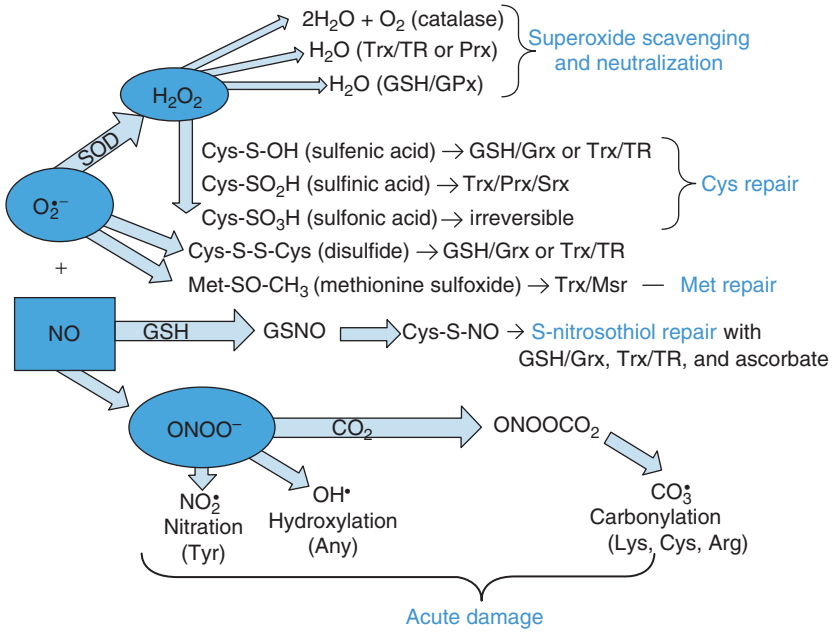

Figure 2 Reactive oxygen/nitrogen species generation and neutralization reactions and systems: Upper level: Superoxide scavenging reactions catalyzed by superoxide dismutase (SOD), catalase, thioredoxin reductase (TR), peroxiredoxin (Prx), and glutathione peroxidase (GPx) and mediated by thioredoxin (Trx) and glutathione (GSH). Center: Oxidation states of cysteine and methionine and their repair; glutathione repairs most of the oxidative modifications by glutathionylation, followed by glutathione removal with the glutaredoxin (Grx) system; methionine is repaired with the methionine sulfoxide reductase (Msr)/Trx system; sulfonic acid modification is irreversible, while sulfinic acid can be repaired with the use of a recently characterized ATP- and Mg^{2+}-dependent enzyme, sulfiredoxin (Srx), which is capable of reducing the sulfinic form of peroxiredoxin. Lower level: Peroxynitrite-generated radicals cause largely nonrepairable damage to proteins, DNA, and lipids.

in cardiac function, decreased development of cataracts, and diminished mitochondrial DNA mutations. It is unclear to what extent the increased life span reflects improvements in brain function.

Antioxidant/Antiaging Mechanisms: Repair of Oxidized Residues

Unscavenged radicals are thermodynamically driven to oxidize cellular constituents, including protein cysteines and methionines. These critical amino acid residues can be subsequently reduced only with the help of specialized enzyme systems (Figure 2). Cysteines are reduced with a glutaredoxin/glutathione system, and methinones are reduced with a methionine sulfoxide reductase (Msr) system. The Msr system includes two enzymes, only one of which belongs to the same enzyme class as glutathione reductase and thioredoxin reductase.

The glutaredoxin/glutathione enzyme system has beneficial effects on functional activity of a number of proteins, including the thiol-containing mitochondrial complex I, the inhibition of which is caused by administration of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP), a neurotoxin that produces PD-like symptoms in primates, including humans. Interestingly, glutathione depletion is an early feature of

Table 1 Changes in key proteins regulating the cellular redox status in disease, injury, and aging

Protein/enzyme	*AD (brain)*	*PD (substantia nigra)*	*ALS (spinal cord)*	*Injury, excitotoxicity (brain)*	*Aging (brain)*
Trx1	Protein ↓	n.a.	n.a.	n.a.	n.a.
TR	Activity ↑	n.a.	n.a.	n.a.	n.a.
MsrA	Activity ↓	n.a.	n.a.	n.a.	Activity ↓, protein ↓, mRNA ↓
Prx1	Protein unchanged	Protein ↓	n.a.	n.a.	n.a.
Prx2	Protein ↑	Protein ↓	Protein ↓	Protein ↑	Protein ↓
Prx3	Protein ↓	Protein ↓		Protein ↓	Protein ↓
GPx1	mRNA ↑	mRNA ↓	Protein ↓	n.a.	Activity ↑
GR	mRNA ↑	n.a.	n.a.	n.a.	Activity ↓
GCL	n.a.	Activity unchanged	n.a.	n.a.	Activity ↓, protein ↓
Grx1	Protein ↑	n.a.	n.a.	Protein ↑	n.a., ↓ in interfibrillary mitochondria
SOD1	Protein ↑	n.a.	Activity ↓	Protein ↑	n.a.
NADH:UQ reductase	Protein ↓	Protein ↓	n.a.	n.a.	n.a.

Trx1, thioredoxin; TR, thioredoxin reductase; MsrA, methionine sulfoxide reductase A; Prx1–3, peroxiredoxins; GPx1, glutathione peroxidase 1; GR, glutathione reductase; GCL, glutamate cysteine ligase; Grx1, glutaredoxin 1; SOD1, superoxide dismutase; NADH: UQ reductase, NAD(P)H:ubiquinone oxidoreductase (antioxidant response element-dependent, two-electron reduction of quinones); n.a., not available.

PD and appears to be disease specific. The depletion of glutathione may explain defects in complex I function found in sporadic forms of PD.

Msr, another important enzyme in repair of proteins, is proposed to play a central role in neurodegeneration and aging. Genetically engineered organisms overexpressing the enzyme live longer. On the contrary, the reduced or suppressed activity of the enzyme results in shortened life span, hippocampal degeneration, and increased sensitivity to oxidative stress. The enzyme can repair oxidized methionines and restore the function of many important proteins – for example, calmodulin. Msrs act on oxidized calmodulin and repair all eight methionine sulfoxide residues initially present in the inactive protein.

As mentioned, neuronal mitochondria protect themselves against hydrogen peroxide formed inside the matrix solely with the thiol-dependent enzyme systems. Moreover, repair systems also largely depend on reduced thiols (see earlier). In many cases, the catalytically important enzyme thiols are shown to be protected against oxidative damage by glutathionylation. Glutathionylation is a posttranslational, reversible redox modification of proteins by thiol/disulfide exchange. Indeed, glutathione used to be a dominant theme in brain neurodegeneration, and only recently the significance of thioredoxin-based systems has been fully appreciated. Measurements of reduced nicotinamide adenine dinucleotide phosphate (NADPH)-dependent thiol reducing activity in brain mitochondria show that glutathione reductase and thioredoxin reductase are equally important. Thiol-dependent systems of hydrogen peroxide neutralization and subsequent repair are also targets for oxidative modification (Table 1), and this obviously creates a threshold of oxidative damage that can be handled by neuronal mitochondria. Once reached, it results in mitochondrial failure and subsequent cell death by either apoptotic or necrotic pathways.

The delicate balance existing under normal conditions between oxidant production and antioxidant defensive mechanisms (Figure 3) can either be slowly shifted with aging or disturbed at once upon acute injury to produce oxidative stress. The adaptive response to tissue damage includes the activation of transcription factors such as nuclear factor-kappa B (NF-κB), among others, to switch on the inflammation response to limit damage and promote repair.

Injury and Neurodegeneration

The most common form of neurodegeneration occurs after an acute injury. The causes of neuronal injury are many, and include stroke, epilepsy, trauma, DNA damage from radiation, chemotherapeutic agents, and exposure to environmental neurotoxins. Acute neuronal damage involves a complex combination of processes, including excitotoxicity, inflammation, necrosis, and apoptosis.

Excitotoxicity

Excitotoxicity is a particularly important event that initiates acute neurodegeneration. Accumulation of glutamate within the synaptic cleft leads to Ca^{2+}

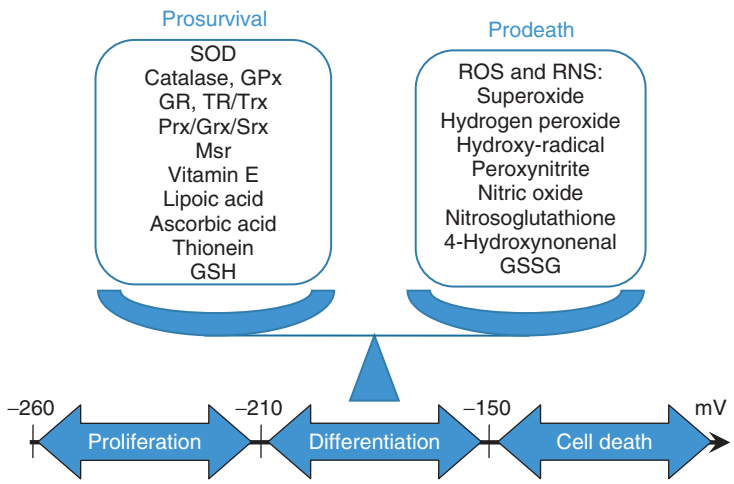

Figure 3 Delicate balance between prodeath and prosurvival factors determines the cell fate *in vivo*. For cultured nerve cells, the depletion of intracellular redox potential, shown as that for the glutathione/oxidized glutathione (GSSG/GSH) couple, below a certain level determines whether the cells will differentiate or die. SOD, superoxide dismutase; ROS, reactive oxygen species; RNS, reactive nitrogen species; GPx, glutathione peroxidase; GR, glutathione reductase; TR/Trx, thioredoxin reductase/thioredoxin; Prx/Grx/Srx, peroxiredoxin/glutaredoxin/sulfiredoxin; Msr, methionine sulfoxide reductase.

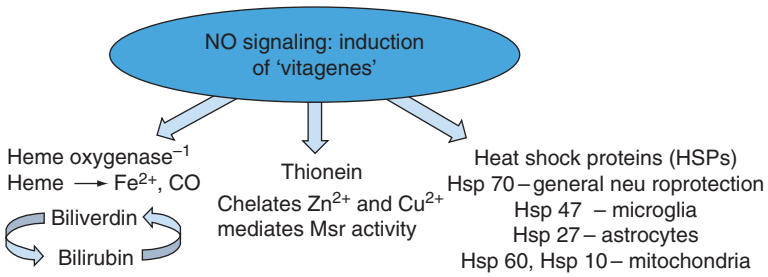

Figure 4 Prosurvival role of nitric oxide (NO): induction of synthesis of heme oxygenase, which provides cells with an antioxidant bilirubin/biliverdin/biliverdin reductase system; synthesis of thioneins, which chelate metal ions and in addition can mediate/enhance methionine sulfoxide reduction by methionine sulfoxide reductase (Msr) and thioredoxin (Trx); and synthesis of heat shock proteins.

influx via hyperactivation of *N*-methyl-D-aspartate receptors, voltage-gated Ca^{2+} channels, and nonspecific cation conductances. The latter nonspecific cation channels are represented by the recently discovered transient receptor potential (melastatin) (TRPM) ion channels, among which TRPM7 and TRPM2, shown to be permeable for Ca^{2+} and inhibited by gadolinium, appear to play a critical role in anoxic cell death.

Excess Ca^{2+} via ionotropic glutamate receptors results in the initiation of neurotoxic signaling cascade by activating calmodulin and neuronal nitric oxide synthase (nNOS). Current opinion holds that the intracellular redox state is the critical factor determining whether in brain cells NO is toxic or protective (Figure 4). NO is known to signal the induction of heme oxygenase-1, which is considered as a prosurvival enzyme. It degrades heme, yielding ferrous iron, carbon monoxide (CO), and biliverdin, which cycles between the oxidized and reduced form, bilirubin. The latter exhibits strong antioxidant properties. Biliverdin reductase is present in brain in large functional excess, suggesting that such redox cycling amplifies antioxidant effects of heme oxygenase expression. On the other hand, NO also reacts with glutathione, generating nitrosoglutathione, which is able to modify and thus inactivate protein thiols. However, the most dangerous species – peroxynitrite – is generated via the direct interaction of superoxide radical and NO, a comparatively stable and harmless molecule. Peroxynitrite is extremely reactive and capable of nitrating tyrosine residues in proteins (Figure 2). A downstream effect of peroxynitrite production is the activation of TRPM channels. Molecular deletion of TRPM2 or TRPM7 channels renders neurons resistant to hypoxia/aglycemia. Whether peroxynitrite induces activation of TRPM channels via nitration of its critical components or some upstream modulatory protein is unclear. Protein nitration has been shown to

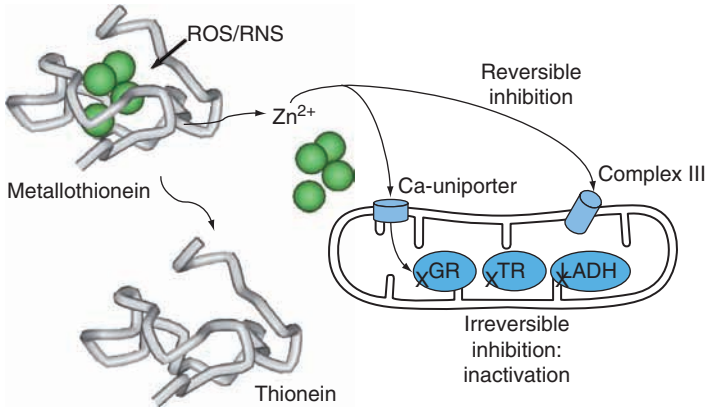

Figure 5 Zn-induced damage to mitochondrial enzymes of energy production and antioxidant defense. GR, glutathione reductase; TR, thioredoxin; LADH, lipoamide dehydrogenase.

take place in brain injuries and in some but not all neurodegenerative diseases.

Superoxide- and especially peroxynitrite-induced modifications of the metallothionein thiols result in the release of zinc. Neurodegenerative diseases are characterized by a mobilization of intracellular zinc. The latter has been shown to mediate NO-induced neuronal death by directly affecting mitochondrial respiration through inhibiting complex III. In addition, it has recently been shown that zinc is capable of entering mitochondria and directly inactivating lipoamide dehydrogenase, the terminal enzyme of major multienzyme energy-producing complexes, as well as glutathione reductase and thioredoxin reductase. The inactivation is irreversible, and only the newly synthesized proteins transported into mitochondria may compensate for the damage (Figure 5). Thus, acute oxidative damage, resulting in massive release of intracellular zinc, will cause mitochondrial failure and cell death.

Inflammation

Oxidative stress can directly or indirectly initiate an inflammatory cascade. Collective evidence from many recent studies suggests that increased phospholipase A2 (PLA2) activity plays a central role in acute inflammatory responses in the brain as well as in oxidative damage associated with AD, PD, and multiple sclerosis (MS). PLA2 contributes to pathogenesis of neuroinflammation by attacking neural membrane phospholipids, to yield arachidonic acid and lysophospholipids. These are subsequently metabolized to a variety of proinflammatory lipid mediators such as prostaglandins, leukotrienes, thromboxanes, and platelet-activating factor. Arachidonic acid metabolism is also one of the major sources of oxidative damage (Figure 6). Arachidonic acid undergoes catalytic oxidation by cyclooxygenases 1 and 2, and, in addition, can react with superoxide or peroxynitrite to generate one of the most potent modifying agents, 4-hydroxy-2-nonenal (HNE). The latter can modify lysine, cysteine, and histidine residues in proteins and can bind to free amino acids and deoxyguanosine. Immunostaining for HNE-modified proteins shows that such modification is characteristic of acute oxidative stress, occurring during brain and spinal cord injury.

Inflammation is characterized with the increased deposition of iron. Iron, especially in the ferrous or unbound form, is able to catalyze the formation of free radicals and could be a cause of neuronal injury. Depletion of antioxidants in the brain, rise in iron-dependent oxidative stress, and monoamine oxidase B (MAO B) activity are key characteristics of aging that contribute to the onset of neurodegenerative disorders. MAO B is a mitochondrial flavin-dependent enzyme that catalyzes oxidative deamination of neurotransmitters and exogenous arylalkylamines. MPTP, an impurity in synthetic heroin, being activated by MAO B, affords a widely used chemically induced model of PD. MPP^+, the toxic product of MPTP conversion, increases superoxide formation by suppressing activity of NADH dehyrogenase (complex I) and increasing leak of electrons to oxygen. Superoxide is then available to attack Fe/S cluster proteins such as aconitase. Destabilization of cytosolic as well as mitochondrial aconitase results in an increase in iron regulatory protein-1 (IRP-1), an RNA-binding protein that signals cellular iron deficiency. This leads to the paradoxical and maladaptive increase in iron in the cell. The mechanisms underlying iron cellular toxicity are only beginning to emerge. A representative trend in neuroprotective drug development is to combine an iron chelator and a monoamine oxidase inhibitor in the same compound (see later). Iron chelators have recently been shown to

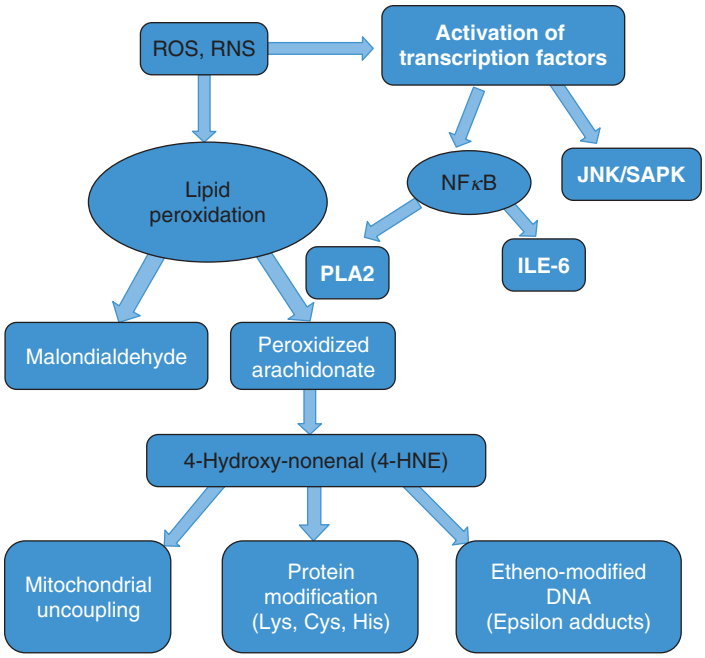

Figure 6 4-Hydroxynonenal, the most damaging product of lipid peroxidation. Phospholipase A2 (PLA2) inhibitors were shown to be neuroprotective. ROS, reactive oxygen species; RNS, reactive nitrogen species; NFκB, nuclear factor-kappa B; JNK/SAPK, c-Jun N-terminal kinase/stress-activated protein kinase; ILE-6, interleukin-6.

have the ability to induce adaptive gene expression via the stabilization of factors such as hypoxia-inducible factor-1 (HIF-1). HIF is a heterodimeric transcriptional complex that mediates the induction of more than 70 genes involved in hypoxic compensation, including those for vascular endothelial growth factor (VEGF) and erythropoietin (Epo).

Activated neutrophils and macrophages generate widespread secondary damage at the traumatic site by releasing cytokines and free radicals. They produce inducible NO synthase and NAD(P)H oxidase (Phox), generating nonmitochondrial superoxide; both enzymes cooperate to generate peroxynitrite and thus expose cells to further oxidative damage. In addition to peroxynitrite, leukocytes possess myeloperoxidase, which generates hypochlorite from hydrogen peroxide and chloride anion. Hypochlorite is a strong oxidizing agent capable of chlorinating protein residues and oxidizing membrane lipids. In all cases of rodent neurodegeneration, medications reducing inflammatory responses were shown to exhibit beneficial effects on the disease progression.

Apoptosis

The discovery that oxidative stress can trigger a program of neuronal cell death with features of apoptosis was significant in several aspects. First, it showed that oxidative stress does not always result in random and disordered cell damage. Second, it demonstrated the possibility of a free radical triggering an endogenous program of cell suicide.

Oxidative stress has been shown to activate a host of downstream signaling pathways leading to apoptosis. In some schemes, the c-Jun N-terminal kinase (JNK) signaling pathway is activated, resulting in Bax translocation, cytochrome *c* release, and apoptosis. In other schemes, Erk activation or poly(ADP-ribose) polymerase (PARP) activation leads to translocation of an apoptosis inducing factor and a caspase-independent cell death. Oxidative stress has been implicated in activation of a cell cycle resulting in cell death, although two studies in which oxidative stress has been induced by downregulating antioxidant defenses failed to demonstrate a protective effect of cell cycle inhibitors.

The superoxide released by mitochondria is directed both 'in' to and 'out' of the mitochondrial inner membrane and, therefore, is capable of damaging mitochondrial DNA, matrix proteins, inner membrane proteins and lipids, and cytosolic proteins and nuclear DNA. With respect to DNA damage, more than 1000 DNA-damaging events occur in each mammalian cell every day, due to replication errors and cellular metabolism. To cope with the deleterious consequences of DNA lesions, cells are equipped with efficient defense mechanisms to remove DNA damage by DNA repair pathways, control cell cycle progression,and eliminate

damaged cells via apoptosis. The complicated network of DNA repair mechanisms includes base excision repair, transcription-coupled repair, global genome repair, mismatch repair, homologous recombination, and nonhomologous end-joining damage. Evolution has overlaid the core cell cycle machinery with a series of surveillance pathways termed 'cell cycle checkpoints.' Checkpoints in proliferating cells tightly control progress through the cell cycle; cells may be arrested at any of the checkpoints, and either DNA will be repaired or cells will die by apoptosis. It appears that apoptosis induced by oxidative damage may be promoted by multiple pathways, both cell cycle dependent and independent. The determining factor for the pathways induced is an area of active exploration (Figure 7).

Nature of Aggregated Deposits in Neurodegenerative Diseases

Alzheimer's Disease

First described by Alois Alzheimer in 1906, AD has two characteristic 'deposits,' tangles and plaques. The latter are found in brain parenchyma and in blood vessels and are formed from amyloid mixed with inflammatory cells. Amyloid presents cross-linked Aβ-peptides formed via wrong processing of amyloid precursor protein (APP). Genetic AD is associated with mutations in APP or in presenilins (1 and 2), the proteins responsible for APP processing. Neurofibrillary tangles can be formed from the oxidatively cross-linked intraneuronal tau protein and occur mainly in large pyramidal neurons. The oxidative stress results in more and more dense plaques and tangles. Once formed, tangles and plaques directly and indirectly enhance the oxidative stress. The direct pathway represents the oxidative activity of metal-bound (Zn^{2+}, Fe^{2+}, or Cu^{2+}) plaques and tangles. The indirect pathway includes the activation of microglial cells, which in turn generate inflammatory messengers, putting the brain on permanent 'red alert.' This indirect pathway had led some to call AD 'arthritis of brain.' Thus oxidative stress may precipitate neuropathological hallmarks of AD or may be a consequence of these neuropathological changes. Disappointing therapeutic results with antioxidants in this disease could reflect an inability to adequately neutralize damaging radicals in all of the cell types and subcellular compartments afflicted in AD.

Parkinson's Disease

PD is a common neurodegenerative disorder affecting 1% of the population over the age of 65 years. The disease is characterized by the loss of dopaminergic neurons in the substantia nigra pars compacta, as well as by the presence of Lewy body inclusions in these cells. Five genes have been identified that are associated with relatively rare familial PD, by either autosomal dominant or recessive transmission. These include genes encoding α-synuclein and dardarin/leucine-rich kinase 2 (LRRK2), which are associated with dominantly inherited PD, and genes for parkin, DJ-1, and pten-induced kinase 1 (PINK1), which cause recessively inherited PD. α-Synuclein readily aggregates and is a major fibrillar component of Lewy bodies.

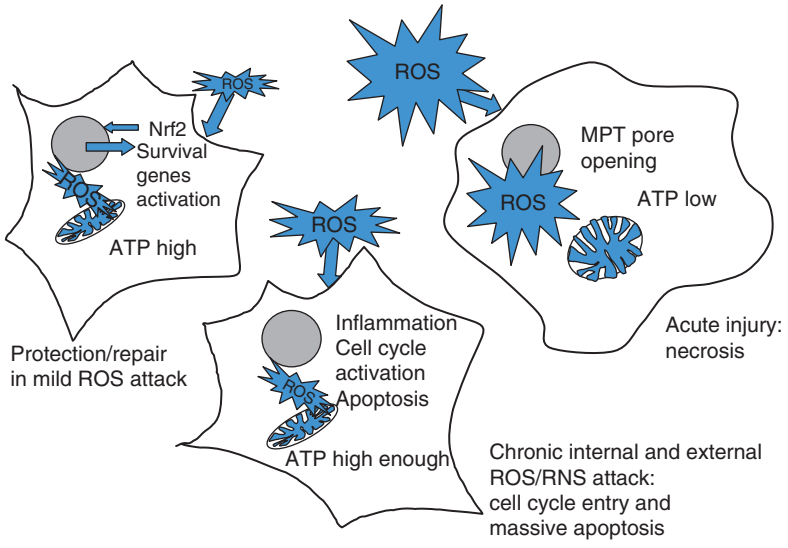

Figure 7 Cell fate depends on the level of oxidative stress. ROS, reactive oxygen species; MPT, mitochondrial permeability transition; Nrf2, nuclear factor E2-related factor 2; RNS, reactive nitrogen species.

Aggregation of α-synuclein is enhanced with tissue transglutaminase: Lewy bodies in PD patients are positively immunostained with antibodies recognizing isodipeptide bonds, a marker of tissue transglutaminase cross-linking. Phosphorylation also promotes α-synuclein aggregation. Cultured cells overexpressing α-synuclein generate reactive oxygen species (ROS). One mechanism by which aggregated synuclein leads to ROS generation in PD is via trapping trace metals into the aggregates. These metals can then easily accept electrons from reducing substances such as superoxide or glutathione and transfer them to oxygen to form superoxide radical. DJ-1, encoded by a gene that when mutated is associated with autosomal recessive parkinsonism, has antioxidant properties which are lost if the mutation destroys its dimeric structure. The monomer redistributes from cytosol to the mitochondria and nucleus.

Polyglutamine Diseases

Polyglutamine diseases (the CAG trinucleotide repeat/polyglutamine diseases) are characterized by the occurrence of protein aggregates within neurons. The most well characterized among ten known diseases of this origin are HD, dentatorubral and pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy, and multiple forms of spinocerebellar ataxia. Each disease is caused by a distinct gene product with expanded polyglutamine repeats.

Huntington's Disease

The hallmarks of the genetic disorder known as HD are a progressive chorea combined with dementia. Historically, the lurching madness was mistaken for possession by witchcraft (some of the Salem witches burnt in 1693 may have actually had HD). HD is caused by a single, dominant gene (i.e., only one copy is sufficient to cause the disease), unlike most genetic disorders, which are recessive (i.e., two 'bad' copies are needed to cause the disease). The average onset is from 35 to 40 years of age. Huntingtin is a large (350 kDa) protein of unknown function; once the polyglutamine tail crosses the threshold of 38 residues, the mutant begins to form aggregates. The latter precipitate in cytosol and also form nuclear inclusions. The mechanisms by which huntingtin aggregates launch the disease are still disputable. It is supposed to be a gain-of-function event. It is documented that mutant huntingtin can interfere with gene expression that is associated with adaptation to oxidative stress or mitochondrial dysfunction. Thus mutant huntingtin may be directly toxic to mitochondria and this toxicity may be sustained by mutant huntingtin's suppression of compensatory gene expression. While defects in energy metabolism are widely documented in human HD and associated animal models, the only evidence for oxidative stress is oxidative DNA damage.

Amyotrophic Lateral Sclerosis

ALS (Lou Gehrig's disease) is a chronic, progressive disease marked by gradual degeneration of motor neurons. Most incidences are sporadic, with ~10% of cases inherited in an autosomal dominant manner. The cause of ALS is unknown; however, in 2% of all cases (20% of familial disease) the primary cause is a mutation of the CuZnSOD (*SOD1*) gene. Approximately 117 mutations have been identified; they occur on the protein surface and all yield dominantly inherited disease. As in AD and PD, intracellular and extracellular aggregates are present in ALS, and the aggregates are intensely immunoreactive with antibodies to SOD1, supporting the theory that mutant SOD1 is toxic because of aggregation and coaggregation with other essential proteins. Peripherin, a neuronal intermediate filament protein, is associated with inclusion bodies in motor neurons of patients with ALS. Proteomics study has shown the upregulation of peripherin in ALS patients, and the downregulation of mitochondrial membrane transport and respiratory chain proteins. Study of transgenic mice that express the SOD1 mutation has shown that the mutant protein aggregates onto mitochondria only in cells within the spinal cord. This finding explains the selectivity of SOD1 mutations to affect motor neurons. Theories to explain ALS include oxidative and nitrosative stress, aggregation, mitochondrial damage, and glutamate and microglial cell-mediated damage.

Neuroprotection Strategies: Problems and Perspectives

By the time a patient is diagnosed as having a neurological illness, extensive neuronal damage has usually already occurred. Consequently, there is a great need for the discovery of biomarkers that allow early diagnosis and intervention. Common features among neurodegenerative diseases – genetic mutations (SOD1, parkin, huntingtin), protein misfolding and aggregation (Lewy bodies, amyloid plaques, tangles), and mitochondrial dysfunction and apoptosis – have implications for disease prevention and for development of effective therapies. Despite common final pathways, it is unlikely that a single drug or targeting a single mechanism will be sufficient to halt neurodegenerative processes.

Chronic neurodegeneration is age related, and thus delay in biological aging will decrease the occurrence of age-related diseases, with resulting prolongation of a healthy life span. Available evidence suggests that to delay aging one has to maintain healthy mitochondria and reduce oxidative stress. One of the proposed approaches is to maintain or recover the activity of the so-called vitagenes. In this regard, the positive effect of heme oxygenase is linked to the production of bilirubin and biliverdin (see earlier), the administration of which, after the first few weeks of life, in doses slightly above normal levels, results in cytoprotective effects. Also, positive effects of CO administration in doses comparable to that of heavy smoker have been shown in suppression of ischemia–reperfusion injury.

Another approach to activate vitagenes is administration of nutritional antioxidants. Curcumin, the most prevalent nutritional and medicinal compound used by the Indian population, has the potential to inhibit lipid peroxidation and efficiently neutralize reactive oxygen and nitrogen species. It has been recently shown that curcumin inhibits NF-κB activation and induces heme oxygenase-1. Administration of curcumin for 6 months resulted in a suppression of indices of inflammation and oxidative damage in the brain of an AD transgenic mouse model (APPSw). Caffeic acid phenethyl ester, an active component of propolis, has been shown to induce heme oxygenase-1 in astroglial cells. Gene induction in both cases occurs through the antioxidant response element (ARE), and this led to the conclusion that the increased expression of genes regulated by the ARE may provide the CNS with protection against oxidative stress. Indeed, numerous studies have supported a role for ARE activators in the prophylaxis against acute and chronic neurodegenerative conditions.

Anti-inflammatory and anti-apoptotic treatments have also shown benefit for many forms of neurodegeneration, although it is also becoming clear that parts of the inflammatory response must be maintained to facilitate repair. A combinatorial approach that highlights the promise of this strategy is the synergistic neuroprotective effects of minocycline (MC) and glatiramer acetate (GA) in a MS model. MC showed remarkable benefit in animal models of MS, PD, HD, ALS, stroke, and spinal cord injury, and in preliminary clinical data for 36-month human trials of MS. MC has a role in interdicting a host of deleterious pathways, including inhibiting matrix metalloproteases (MMPs), chelating Ca^{2+} divalent metals, and inhibiting p38 mitogen-activated protein kinase (MAPK), reducing microglial activation and inhibiting cytochrome *c* release from mitochondria. GA polarizes T cells toward T helper type 2 (Th2) status. GA-reactive Th2 cells cross the blood–brain barrier, and upon reactivation by myelin antigens, release Th2 cytokines. In the CNS, Th2 cells confer neuroprotection, as demonstrated in animal models of PD, ALS, optic nerve crush, ocular glutamate excitotoxicity, etc.

A recent development in chelators involves the design and synthesis of multifunctional drugs that have the ability to bind iron, inhibit a particular enzyme, and exhibit antioxidant properties (free radical scavengers). Metal (in particular, iron) chelation therapy has been proposed as a way of reducing the level of redox-active metals in neurodegenerative diseases. The green tea catechin epigallocatechin gallate (EGCG; known for its iron-chelating and antioxidant properties), the antibiotic iron chelator clioquinol, and intracerebroventricularly injected desferal (DFO) are potent neuroprotective agents. Obviously, iron chelation targets not only unbound iron and that in the aggregated deposits, but, more significantly, also iron dioxygenases such as HIF prolyl hydroxylase. The latter is an emerging target for neuroprotection, although HIF may not be the only substrate of this enzyme. More than 70 genes of putative nonheme iron oxygenases have been identified in the human genome, but only a number have the physiological functions ascribed. An assumption on the uniqueness of an inhibitory action of a particular drug selected among others may be wrong, if only one enzyme candidate has been tested. The best example is probably EGCG, which targets HIF prolyl hydroxylase, MAO B, and MICAL ('molecule interacting with Cas ligand'), a flavin monooxygenase implicated in axonal guidance and highly homologous to MAO B. The recently discovered M30 as an antioxidant/chelator/MAO inhibitor has also been shown to stimulate neurite outgrowth and thus may actually target MICAL as well.

Conclusions

Therapies targeted at reducing oxidative damage in the nervous system must achieve several goals in order to be effective. First, they must interdict pathological oxidant interactions without affecting physiological signaling by radicals. Over the past two decades, peroxide and nitric oxide have been shown to play the role of messengers in the nervous system. Antioxidants that inadvertently abrogate these signaling functions would not be desirable. Second, antioxidants or, alternatively, effective repair strategies must be augmented in distinct cell types and subcellular compartments. Oxidative and nitrosative stress are not an undifferentiated whole and are mediated by distinct species produced in distinct cellular

compartments and distinct cell types. The great challenge has been to divine a multimodal strategy that inhibits a cassette of targets without the expected toxicity that arises as the specificity of the therapy decreases. We propose that understanding endogenous homeostatic pathways for protecting against oxidative and nitrosative stress is the way forward. These homeostatic pathways involve the activation of preexisting proteins as well as *de novo* gene expression. Small molecules that activate homeostatic responses to oxidant stress are expected to reap large therapeutic benefits. Evidence that such an approach is effective and safe continues to emerge from preclinical studies. The ultimate proof will be the demonstration of neuroprotection in a human clinical trial. Such success will also provide long overdue evidence supporting a role for oxidative damage in human neurological disease.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Apoptosis in Nervous System Injury; Apoptosis in Neurodegenerative Disease; Axonal Transport and Neurodegenerative Diseases; Excitotoxicity in Neurodegenerative Disease; Huntington's Disease: Neurodegeneration; Inflammation in Neurodegenerative Disease and Injury; Mitochondrial Dysfunction in Nervous System Injury; Neuroprotection: Endogenous Mechanisms; Neuroprotection: Pharmacological Approaches; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Aarts M, Iihara K, Wei WL, et al. (2003) A key role for TRPM7 channels in anoxic neuronal death. *Cell* 115: 863–877.

Becker EB and Bonni A (2004) Cell cycle regulation of neuronal apoptosis in development and disease. *Progress in Neurobiology* 72: 1–25.

Binda C, Hubalek F, Li M, et al. (2006) Structure of the human mitochondrial monoamine oxidase B: New chemical implications for neuroprotectant drug design. *Neurology* 67: S5–S7.

Calabrese V, Butterfield DA, Scapagnini G, et al. (2006) Redox regulation of heat shock protein expression by signaling involving nitric oxide and carbon monoxide: Relevance to brain aging, neurodegenerative disorders, and longevity. *Antioxidants & Redox Signaling* 8: 444–477.

Dhib-Jalbut S, Arnold DL, Cleveland DW, et al. (2006) Neurodegeneration and neuroprotection in multiple sclerosis and other neurodegenerative diseases. *Journal of Neuroimmunology* 176: 198–215.

Farooqui AA, Ong WY, and Horrocks LA (2006) Inhibitors of brain phospholipase A2 activity: Their neuropharmacological effects and therapeutic importance for the treatment of neurologic disorders. *Pharmacological Reviews* 58: 591–620.

Fratelli M, Demol H, Puype M, et al. (2002) Identification by redox proteomics of glutathionylated proteins in oxidatively stressed human T lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America* 99: 3505–3510.

Ghezzi P and Bonetto V (2003) Redox proteomics: Identification of oxidatively modified proteins. *Proteomics* 3: 1145–1153.

Jenner P, Dexter DT, Sian J, et al. (1992) Oxidative stress as a cause of nigral cell death in Parkinson's disease and incidental Lewy body disease. *Annals of Neurology* 32: S82–S87.

Johnson MD, Yu LR, Conrads TP, et al. (2005) The proteomics of neurodegeneration. *American Journal of Pharmacogenomics* 5: 259–270.

Langley B and Ratan RR (2004) Oxidative stress-induced death in the nervous system: Cell cycle dependent or independent? *Journal of Neuroscience Research* 77: 621–629.

Lehtinen MK, Yuan Z, Boag PR, et al. (2006) A conserved MST-FOXO signaling pathway mediates oxidative-stress responses and extends life span. *Cell* 125: 987–1001.

Maher P (2006) Redox control of neural function: Background, mechanisms, and significance. *Antioxidants & Redox Signaling* 8: 1941–1970.

Ong WY and Farooqui AA (2005) Iron, neuroinflammation, and Alzheimer's disease. *Journal of Alzheimer's Disease* 8: 183–200; discussion 209–215.

Poole LB, Karplus PA, and Claiborne A (2004) Protein sulfenic acids in redox signaling. *Annual Review of Pharmacology and Toxicology* 44: 325–347.

Schwartz EI, Smilenov LB, Price MA, et al. (2007) Cell cycle activation in postmitotic neurons is essential for DNA repair. *Cell Cycle* 6: 318–329.

Siddiq A, Aminova LR, and Ratan RR (2007) Hypoxia inducible factor prolyl 4-hydroxylase enzymes: Center stage in the battle against hypoxia, metabolic compromise and oxidative stress. *Neurochemical Research* 32: 931–946.

Singh IN, Sullivan PG, Deng Y, et al. (2006) Time course of post-traumatic mitochondrial oxidative damage and dysfunction in a mouse model of focal traumatic brain injury: Implications for neuroprotective therapy. *Journal of Cerebral Blood Flow & Metabolism* 26: 1407–1418.

Wang JY, Wen LL, Huang YN, et al. (2006) Dual effects of antioxidants in neurodegeneration: Direct neuroprotection against oxidative stress and indirect protection via suppression of glia-mediated inflammation. *Current Pharmaceutical Deign* 12: 3521–3533.

Zheng H, Gal S, Weiner LM, et al. (2005) Novel multifunctional neuroprotective iron chelator-monoamine oxidase inhibitor drugs for neurodegenerative diseases: *In vitro* studies on antioxidant activity, prevention of lipid peroxide formation and monoamine oxidase inhibition. *Journal of Neurochemistry* 95: 68–78.

Oxytocin (Peripheral/Central Actions and their Regulation)

J A Russell and P J Brunton, University of Edinburgh, Edinburgh, UK

© 2009 Elsevier Ltd. All rights reserved.

Oxytocin

Functions

Oxytocin is a neuropeptide hormone produced by neurons in the hypothalamus. Oxytocin secreted from the mother's posterior pituitary gland into the general circulation during birth promotes powerful uterine contractions and hence fetal ejection. It is essential for stimulating milk ejection during suckling in lactation and hence for survival of the young of mammals.

Oxytocin has actions in the brain that complement its peripheral actions: it promotes social interaction and mating behaviors, and it facilitates the rapid expression of maternal behavior after birth. It can also have anxiolytic, stress-reducing, and antinociceptive actions. Both males and females express the oxytocin peptide and receptor genes. In males, oxytocin is involved in the neural network stimulating penile erection.

Oxytocin Peptide and Gene

Oxytocin was the first peptide reproductive hormone to be discovered, in 1906, from its activity in a posterior pituitary extract in stimulating uterine contractions and milk release. In 1953, oxytocin was characterized as a nonapeptide, Cys-Tyr-Ile-Gln-Asn-Cys-Pro-Leu-GlyNH_2, comprising a ring formed by a disulfide bridge between the two cysteine groups, with a three-amino-acid tail, and differing from vasopressin by substitutions at positions 3 and 8. These modifications, from mutations after duplication of a common ancestor gene, result in dramatically different actions of the two peptides.

Essentially, only mammals have an oxytocin gene, which in humans is located on chromosome 20p13 and contiguous except for 3–11 kb with the vasopressin gene, but in opposite orientation. The oxytocin gene codes for a large precursor of 106 amino acids, which includes neurophysin, a larger peptide than oxytocin but without signaling function. The gene has three exons – the first codes for the signal peptide and oxytocin, and the second and third code for neurophysin – and two introns. The oxytocin gene is expressed in distinct groups of neurons in the hypothalamus.

Marsupials may express an oxytocin gene or a similar gene (mesotocin, which is identical to oxytocin except that isoleucine replaces leucine at position 8). Various invertebrates (e.g., earthworm and octopus) have oxytocin-like peptides and receptors in their nervous systems, and these peptides (annetocin and cephalotocin, respectively) promote egg laying.

Regulation of Oxytocin Gene

Regulatory sequences comprise a composite hormone response element (−164 to −145 and −110 to −79 upstream) through which nuclear receptors such as estrogen receptor-β, thyroid hormone, and retinoic acid receptors stimulate, and orphan receptors stimulate (SF-1; TR (testicular orphan receptor)4) or competitively repress (COUP-TF), expression. Control of hypothalamus-specific expression is exerted via the intergenic region between the oxytocin and vasopressin genes.

Progesterone, or its neurosteroid metabolite allopregnanolone, restrains oxytocin gene expression in pregnancy, and a transient postpartum increase in expression follows progesterone withdrawal and stimulation by estrogen. Progesterone action is indirect since oxytocin neurons lack progesterone receptors, but allopregnanolone acts to amplify inhibitory GABA actions via allosteric modulation of $GABA_A$ receptors on oxytocin neurons. Some rostral inputs to oxytocin neurons (in the medial preoptic area and lamina terminalis) contain estrogen receptor-α, indicating an indirect regulation by estrogen.

During lactation, oxytocin gene expression, as reflected in cell body content of oxytocin mRNA, is increased only after 1 week of repeated stimulation by the suckling stimulus, when the posterior pituitary oxytocin store is depleted. Evidently, the level of oxytocin gene expression is set to provide ample oxytocin to meet sustained demand with a wide safety margin for normal reproductive purposes.

Copeptides

Oxytocin neurons also express other neuropeptide genes. Under basal conditions, vasopressin mRNA is expressed in only 1% of magnocellular oxytocin neurons, increasing 17-fold with prolonged stimulation, as in lactation. Proenkephalin-A mRNA expression leads to translation to enkephalins, which have autoinhibitory actions. Other peptides that can be produced include cholecystokinin, corticotropin-releasing factor, and cocaine and amphetamine-regulated transcript.

Neurohypophysial System

Oxytocin Secretion

In parturition and lactation, oxytocin is secreted into the circulation, in pulses a few minutes apart, from

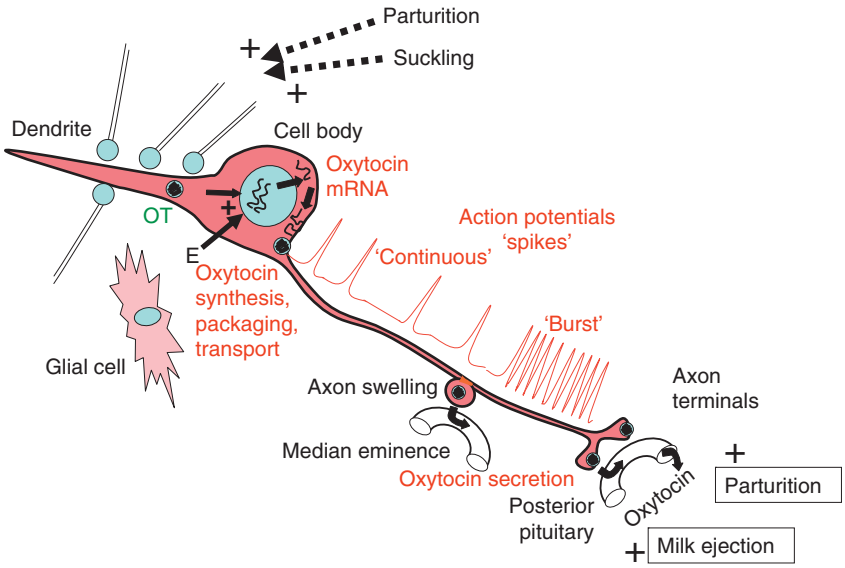

Figure 1 Summary of features of a magnocellular oxytocin neuron. Oxytocin is synthesized in the cell body and transported to the axon terminals in the posterior pituitary; it is secreted by exocytosis when action potentials arrive from the cell body. Some oxytocin is secreted in the median eminence into the hypothalamo-hypophysial portal blood. Oxytocin (OT) is also released from the dendrites in the hypothalamus; this is dependent on Ca^{2+} but not depolarization. Neural stimulation by parturition or suckling causes intermittent high-frequency synchronous discharge ('bursts') of the oxytocin neurons, resulting in secretion of pulses of oxytocin; slow 'continuous' firing is seen in other conditions. Sustained stimulation, as in lactation, increases oxytocin gene expression; estradiol (E) acting via estrogen receptor-β can have a similar effect. +, stimulatory; −, inhibitory. Adapted from Burbach JPH, Young LJ, and Russell JA (2005) Oxytocin: Synthesis, secretion, and reproductive functions. In: Neill JD (ed.) *Knobil and Neill's Physiology of Reproduction*, 3rd edn., pp. 3055–3128. San Diego: Elsevier, with permission from Elsevier.

the millions of axon terminals in the posterior pituitary (neurohypophysis, neural lobe) of the approximately 7000 simultaneously activated neurosecretory oxytocin neurons in the hypothalamus. The cell bodies of these neurons are in the paired supraoptic and paraventricular nuclei in the hypothalamus. These oxytocin neurons are large (20–40 μm in diameter) and designated magnocellular neurons, reflecting their abundant protein synthetic and processing machinery necessary to produce sufficient oxytocin to repetitively attain circulating concentrations high enough to strongly stimulate uterine contractions or milk ejections (Figure 1).

Development

Magnocellular neuron progenitor cells arise at midgestation (embryonic day 10.5 and 12.5 in mouse and rat, respectively) in the neuroepithelium of the ventricular region. Within 4 days, they migrate to form the paraventricular and supraoptic nuclei and then send their axons into the posterior pituitary and begin to express the oxytocin gene. Physiological responsiveness of the oxytocin system develops after birth in the rat. The axons are attracted and guided into the region where the posterior pituitary forms, and they have a trophic effect on the pituicytes, which are modified glial cells. A hierarchy of four transcription factor genes (*Sim1*, *Arnt2*, *Otp*, and *Pou3/2*) is essential for normal embryonic development of oxytocin neurons.

Oxytocin Synthesis and Storage

After mRNA translation on ribosomes in the cell bodies, the oxytocin–neurophysin prohormone is transported through the endoplasmic reticulum, into the Golgi apparatus, and enzymatically cleaved by a prohormone convertase, and the oxytocin now bound covalently to neurophysin is guided into the regulated secretory pathway. The vesicles, or neurosecretory granules, emerging from the Golgi apparatus contain oxytocin and the coproduced peptides, and are continually and slowly transported from the cell bodies within the axons at approximately 8 mm h^{-1}, maintaining a large store of oxytocin in the terminals (Figure 1). A small fraction of the total stored (approximately 4 μg or 2 U, the international unit based on biological activity) is secreted with stimulation (in the rat, approximately 1 mU per milk ejection). During pregnancy, this store is increased 30–50% by decreased release rather than increased gene expression; much of this extra store is secreted during parturition.

Some of the oxytocin–neurophysin-containing secretory vesicles are transported into the dendrites, where

they can release oxytocin by exocytosis. This provides a local feedback mechanism, with possible distant effects of oxytocin diffusing in the brain.

Centrally Projecting System

Oxytocin neurons that project their axons centrally are predominantly in the paraventricular nucleus, with some in the medial preoptic area and bed nucleus of the stria terminalis. These are smaller (cell bodies approximately 10–15 μm in diameter), designated parvocellular neurons, producing and storing less oxytocin to achieve their effects than magnocellular neurons.

Some of these neurons project from the paraventricular nucleus to the brain stem (nucleus tractus solitarii and locus coeruleus) or spinal cord (intermediolateral cell column and dorsal horns), whereas others project to the limbic brain (amygdala and hippocampus), nucleus accumbens, or ventromedial hypothalamic nucleus. In these regions, there is a range of action in appetite control, autonomic regulation, antinociception, anxiety, and sex behavior.

Oxytocin Receptor and Gene

The specific actions of oxytocin depend on its relatively high affinity for the oxytocin receptor and low affinity for the three vasopressin receptors (V1a, V1b, and V2, although the V1b receptor has some affinity for oxytocin), the tissue-selective expression of the receptor, and the pattern of oxytocin secretion.

The oxytocin receptor is a seven transmembrane G-protein-coupled receptor composed of 389 amino acids; in marsupials, the mesotocin receptor is very similar. Only a single oxytocin receptor gene has been identified. The human oxytocin receptor gene is located at chromosomal position 3p25, spans 19 kb, and comprises four exons. The oxytocin receptor gene has evolved from duplication and mutation of an ancestral gene.

Regulation of Oxytocin Receptor Gene

The receptor gene has multiple potential proximal regulatory elements. In general, estrogen upregulates expression and progesterone is inhibitory. The sequence of estrogen and progesterone treatment over several days followed by progesterone withdrawal, mimicking pregnancy levels of these sex steroids, upregulates oxytocin receptor expression in brain regions critical for maternal behavior, which is rapidly induced after birth. In the rat hypothalamic ventromedial nucleus, in which oxytocin promotes female lordosis behavior, estrogen increases oxytocin receptor mRNA expression and acts via protein kinase C to upregulate oxytocin receptor expression. Oxytocin acting in the central amygdala is anxiolytic, and here dopamine regulates oxytocin receptor expression via protein kinase A and an upstream cAMP response element in rats (Figure 2). In contrast, regulation in mice is estrogen dependent, acting via estrogen receptor-α. The oxytocin receptor gene also has interleukin response elements, which may mediate inhibition of expression.

Oxytocin Receptor Distribution in the Brain

The oxytocin receptor is widely distributed, expressed in neurons, notably in olfactory processing regions (olfactory nuclei and piriform cortex), limbic brain structures (including amygdaloid nuclei, septal nuclei, and medial preoptic nucleus), hippocampus, hypothalamus (ventromedial, dorsomedial, premammillary nuclei, and magnocellular nuclei), and brain stem (vestibular nuclei, motor and sensory cranial nerve nuclei regulating face and mouth, and autonomic centers (nucleus tractus solitarii)).

The distribution of oxytocin receptor among brain regions is highly species specific, even between closely related species. In the prairie and montane voles, the different distribution relates to striking differences in social affiliative behavior (Figure 2).

Central Actions of Oxytocin: Receptor Activation

Occupancy of the extracellular binding pocket of the receptor by oxytocin (or agonist derivatives or a nonpeptide synthetic agonist) changes the conformation of the receptor, allowing binding of G-protein ($G_{\alpha q/11}$) to the intracellular domain (Figure 3). This activates phospholipase C, producing inositol trisphosphate (IP_3) and diacylglycerol. IP_3 liberates Ca^{2+}, stimulating exocytotic neuropeptide or transmitter release.

Another pathway leads to prostaglandin and endocannabinoid generation, which mediates autofeedback actions of oxytocin on magnocellular oxytocin neurons. Prostaglandin depolarizes the neurons via extracellular actions on cationic conductances; firing rate is increased by presynaptic inhibition of GABA release. Endocannabinoids are inhibitory via presynaptic inhibition of glutamate and promotion of GABA release (Figure 3).

Oxytocin can excite or inhibit electrical activity of other neurons. Neurons excited by the action of oxytocin on membrane ion channels include those in the bed nucleus of the stria terminalis, the ventromedial hypothalamic nucleus, the dorsal motor nucleus of vagus (parasympathetic preganglionic motor neurons), and spinal cord preganglionic sympathetic neurons. In the spinal cord dorsal horns, oxytocin inhibits pain transmission.

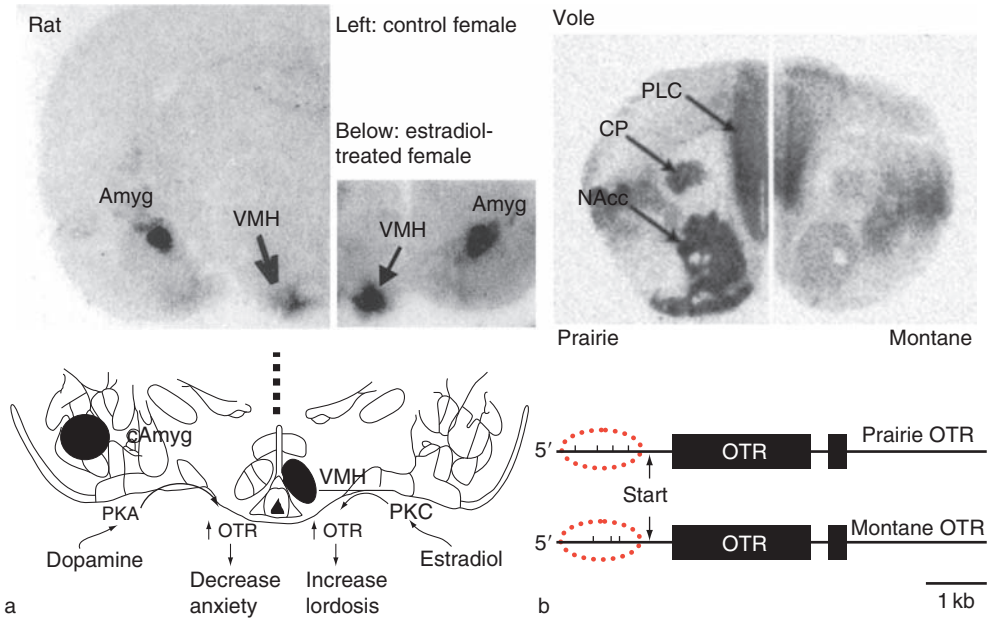

Figure 2 Differential oxytocin receptor expression in the brain. Autoradiographs of coronal hypothalamus sections labeled with oxytocin receptor (OTR) radioligand. (a) Estrogen regulation. Estrogen increases OTR expression in the rat ventromedial hypothalamus (VMH) but not in the amygdala (Amyg). (Bottom) Estrogen regulates OTR expression via protein kinase C (PKC) in the VMH; in the amygdala, dopamine regulates OTR expression via protein kinase A (PKA). In the rat, estrogen is important for oxytocin action on lordosis in the VMH but not for anxiolysis in the amygdala. Adapted from Bale TL, Davis AM, Auger AP, Dorsa DM, and McCarthy MM (2001) CNS region-specific oxytocin receptor expression: Importance in regulation of anxiety and sex behavior. *Journal of Neuroscience* 21: 2546–2552, with permission from the Society for Neuroscience. (b) Species differences. Compared to the female montane vole, in the female prairie vole OTR is strongly expressed in the nucleus accumbens (NAcc) and caudate putamen (CP). Expression is similar between species in the prelimbic cortex (PLC). OTR expression in the CP and NAcc in prairie voles is important in pair bonding and monogamy in this species; montane voles do not form pair bonds. (Bottom) These species differences are due to differences in OTR gene expression. The OTR gene differs only at a few sites (red dotted ovals) in the 5′ flanking region between vole species. Adapted from Lim MM and Young LJ (2006) Neuropeptidergic regulation of affiliative behavior and social bonding in animals. *Hormones and Behavior* 50: 506–517, with permission from Elsevier.

Once released, oxytocin is subject to degradation by placental leucine aminopeptidase (cysteine aminopeptidase–oxytocinase), which cleaves the N-terminal ring structure of oxytocin and is expressed in oxytocin target regions.

Control of Oxytocin Neurons: Electrophysiological Properties

Magnocellular oxytocin neurons have rich excitatory and inhibitory synaptic inputs; there are 37×10^6 synapses/mm^3 in the supraoptic nucleus. Oxytocin neurons, like other neurons, produce action potentials when summation of inhibitory and excitatory postsynaptic potentials depolarizes the cell body above a critical level for action potential generation at the axon hillock (Figure 1). In rats, in which oxytocin is natriuretic, magnocellular oxytocin neurons are directly osmosensitive via stretch-inactivated cationic channels. Such depolarization interacts with excitatory postsynaptic potentials to increase the probability of action potential initiation. Their patterns of firing determine the oxytocin release patterns that are most effective and efficient for the functions of oxytocin. For fetal ejection and milk ejection, brief high-frequency bursts, synchronized among the magnocellular neurons, cause secretion of pulses of oxytocin (Figure 1). Their firing and secretory activity increases slowly and progressively in response to other stimuli, such as systemic cholecystokinin (CCK) administration.

Excitatory Inputs

Noradrenaline The nucleus tractus solitarii (NTS) in the medulla oblongata relays autonomic afferent neural information from the uterus and birth canal during parturition. The NTS contains noradrenergic neurons (the A2 group), which project directly to oxytocin neurons, exciting them via α_1 receptors (Figure 4). Coproduced peptides in these neurons, including neuropeptide Y and prolactin-releasing peptide, supplement excitation by noradrenaline. The NTS also relays information in vagus nerve afferents detecting intra-abdominal CCK and cytokines

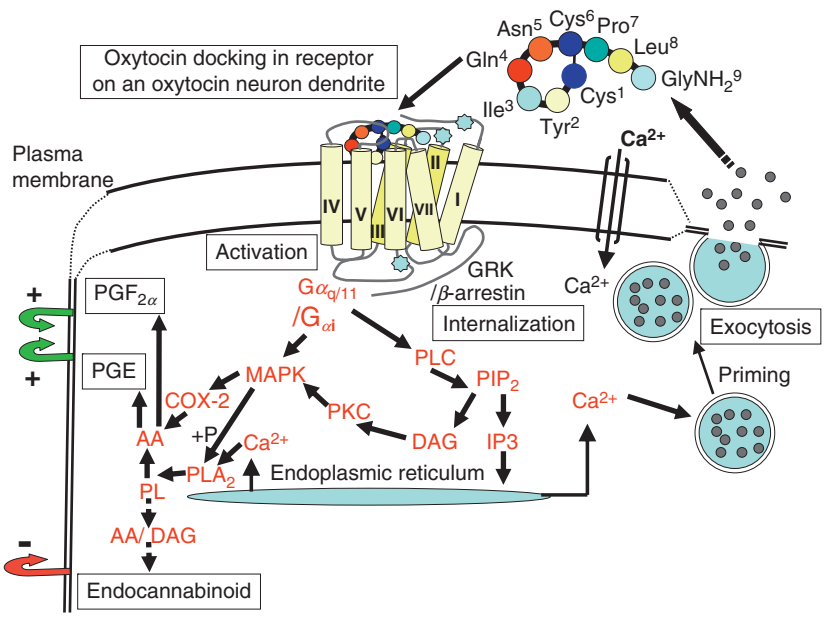

Figure 3 Oxytocin actions on neurons: postreceptor signaling. Schematic based mainly on studies of autoregulation of oxytocin neurons by oxytocin (enzymes are italicized). Oxytocin stimulates dendritic oxytocin secretion by exocytosis and the production of prostanoids and cannabinoids, which modulate excitatory and inhibitory amino acid synapses. The oxytocin receptor is a seven transmembrane domain G-protein-coupled receptor. Oxytocin, a nonapeptide (color-coded residues 3 and 6 are different in vasopressin), docks in an extracellular pocket: stars show key residues in domains 1 and 2. Receptor occupancy leads to coupling to $G\alpha_{q11}$ and $G\alpha_i$; this requires the starred residue on the third intracellular loop. Consequent stimulation of phospholipase C (PLC) triggers several pathways. First, PLC catalyzes phosphatidylinositide 4,5-bisphosphate (PIP_2) breakdown to diacylglycerol (DAG) and inositol 1,4,5-trisphosphate (IP_3). IP_3 triggers Ca^{2+} release from intracellular stores and entry from outside via Ca^{2+} channels. Increased $[Ca^{2+}]_i$ primes oxytocin secretory granules and triggers exocytosis. Second, Ca^{2+} causes phospholipase A_2 (PLA_2) binding to the endoplasmic reticulum; mitogen-activated protein kinase (MAPK) phosphorylates PLA_2, and membrane glycerophospholipid (PL) is broken down to yield arachidonic acid (AA) and prostanoids. Cyclooxygenase 2 (COX-2; induced by MAPK) action forms PGH_2; synthases form PGE_2 or $PGF_{2\alpha}$. Third, oxytocin autostimulates production of endocannibinoids from glycerophospholipid; this is Ca^{2+} dependent via N-acyltransferase/aPLD and PLC/diacylglycerol lipases. Prostaglandins are autoexcitatory by direct effects and by presynaptically inhibiting GABA release. Endocannabinoid can presynaptically inhibit both glutamate and GABA release, with a net inhibitory action. The occupied receptor interacts with G-protein-coupled receptor kinase (GRK), allowing binding to β-arrestin and receptor internalization. Adapted from Burbach JPH, Young LJ, and Russell JA (2005) Oxytocin: Synthesis, secretion, and reproductive functions. In: Neill JD (ed.) *Knobil and Neill's Physiology of Reproduction*, 3rd edn., pp. 3055–3128. San Diego: Elsevier, with permission from Elsevier.

(in rats, signaling satiety or sickness and provoking stress-induced oxytocin secretion); these peptides also activate NTS neurons from the circulation, through the adjacent area postrema.

Suckling stimulus This also acts via brain stem noradrenaline neurons, but the rostral suckling pathway is not fully characterized. The pathway begins with stimulation by the suckled young of somatic sensory nerve endings in and around the nipples. Then, the afferents synapse in the dorsal horn and ascend ipsilaterally, via the dorsolateral funiculi in the spinocervicothalamic tract, into the lateral cervical nucleus and decussate in the lower brain stem. The primary projection from the brain stem to oxytocin neurons is likely direct via the medial forebrain bundle, from noradrenergic neurons (other than A2 neurons in the nucleus tractus solitarii), which also project to regions that regulate nursing behavior.

Glutamate A direct glutamatergic projection from the olfactory bulbs is activated during parturition and may be involved in stimulating oxytocin release into the brain and hence in the initiation of maternal behavior.

A functionally distinct afferent pathway comes from the lamina terminalis, specifically the subfornical organ–nucleus medianus–organum vasculosum lamina terminalis complex. These neurons use glutamate or GABA as neurotransmitter. This input mediates stimulatory actions of circulating relaxin (a pregnancy peptide hormone from the corpus luteum), angiotensin II, and hyperosmolarity.

Inhibitory Inputs

GABA Local GABA neurons around the magnocellular nuclei inhibit oxytocin neurons via $GABA_A$ receptors (Figure 4).

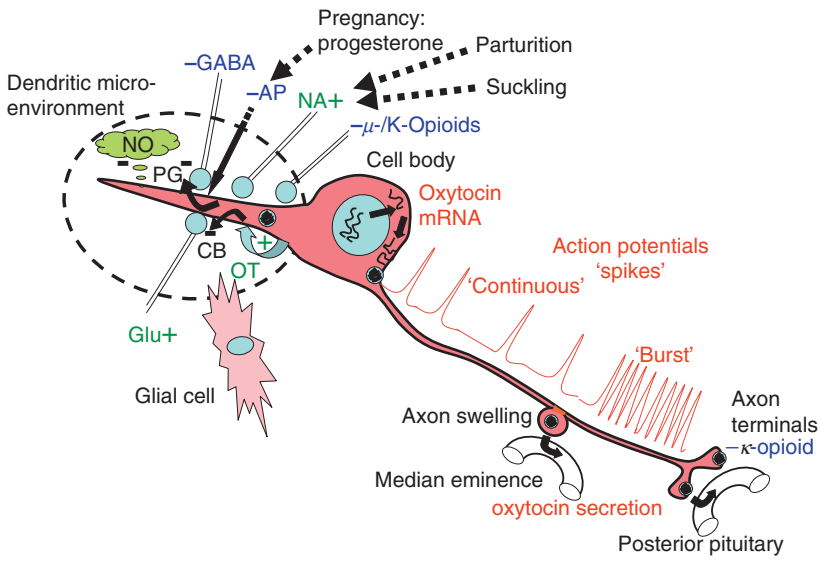

Figure 4 Summary of excitatory and inhibitory inputs to oxytocin neurons and signaling from dendrites that regulate burst-firing. Noradrenaline (NA) is an important transmitter for both inputs from parturition and suckling, and it interacts with local glutamate (Glu) and GABA circuitry to generate bursts. Dendritically released oxytocin (OT) has an essential role in burst-firing: it stimulates its own release and the production of prostaglandins (PG) and endocannabinoid (CB), which presynaptically modulate the glutamate and GABA inputs, thus permitting synchronized burst-firing (see **Figure 3**). Before the end of pregnancy, the activity of oxytocin neurons is restrained by nitric oxide (NO), endogenous opioids (acting centrally and in the posterior pituitary), and allopregnanolone (AP; potentiating GABA action via $GABA_A$ receptors) formed from progesterone. Adapted from Burbach JPH, Young LJ, and Russell JA (2005) Oxytocin: Synthesis, secretion, and reproductive functions. In: Neill JD (ed.) *Knobil and Neill's Physiology of Reproduction*, 3rd edn., pp. 3055–3128. San Diego: Elsevier, with permission from Elsevier.

The neuroactive steroid allopregnanolone, a progesterone metabolite which is abundant until the end of pregnancy, is an allosteric modifier of the $GABA_A$ receptor amplifying the action of GABA by extending the opening time of the Cl^- channel in the $GABA_A$ receptor, prolonging inhibitory postsynaptic currents. This is a mechanism restraining oxytocin neurons from premature excitation near the end of pregnancy.

Opioids κ and μ opioid agonists potently inhibit the firing of oxytocin neurons in rats, even when strongly stimulated. For example, during parturition, morphine, a μ agonist, inhibits oxytocin secretion and stops births. Opioids can act centrally, on inputs to oxytocin neurons or on the neurons themselves, and on axon terminals in the posterior pituitary. Central endogenous opioid inhibition emerges in late pregnancy, as shown by excitatory actions of naloxone. The opioid responsible for this inhibition may be from the brain stem or arcuate nucleus projections to oxytocin neurons. The interruption of parturition by stress involves activation of opioid inhibition of reflex stimulation of oxytocin neurons.

Oxytocin neurons are the only peptidergic neurons that develop opiate (morphine) dependence and tolerance, providing a robust model to study the cellular mechanisms of opiate addiction.

Histamine Oxytocin neurons receive an inhibitory input from histaminergic neurons in the premammillary nuclei. This input may be important during parturition and suckling.

Axon Terminals

Oxytocin and coproduced peptides are released when action potentials arrive at the axon terminals in the posterior pituitary from the cell bodies, traveling at approximately 5 m/s along the unmyelinated oxytocin-containing axons in the pituitary stalk. The arrival of action potentials triggers Ca^{2+} entry into the nerve terminals, leading to exocytosis of oxytocin stored in secretory vesicles, and entry into the systemic circulation via fenestrae in the walls of the adjacent capillaries. Action potentials that are clustered together release more oxytocin. This frequency facilitation results from maintained excitation between spikes due to extracellular K^+ accumulation and depletion of extracellular Ca^{2+} via Ca^{2+} channel activity, both enhancing Na^+ entry.

ATP released by excited terminals initially augments oxytocin release via enhanced Ca^{2+} entry, but the purine adenosine, formed by ATP hydrolysis, then inhibits stimulated oxytocin release by blocking N-type Ca^{2+} channels. These local negative feedback

mechanisms limit release and may contribute to shaping pulses of oxytocin.

Oxytocin release at the axon terminals is modulated by co-released neuropeptides, notably opioids derived from proenkephalin-A, which act on κ or μ receptors on the terminals and inhibit release (Figure 4). This mechanism is downregulated in rats during late pregnancy, enabling more effective control by action potentials from the cell bodies.

Dendrites

The dendrites of supraoptic nucleus oxytocin neurons are broad and bundled together in the ventral glial lamina of the nucleus. Dendrites of magnocellular paraventricular nucleus oxytocin neurons project toward the third ventricle. There are many synaptic boutons on these dendrites. However, the dendrites of oxytocin neurons not only receive signals but also transmit them (Figures 3 and 4). Oxytocin released from the dendrites of magnocellular neurons is an essential component of the mechanism that causes synchronized burst-firing of the neurons during parturition and suckling in lactation, and dendritic oxytocin release is a mechanism of more widespread volume transmission in the brain. Dendritic release of oxytocin can occur independently of initiation of spikes in the cell body, and it is the result of activation of mechanisms that increase intracellular Ca^{2+}; both oxytocin and α-melanocyte-stimulating hormone (from the arcuate nucleus projection) act in this way.

Firing Patterns

Continuous activity Under basal conditions, oxytocin neurons fire action potentials in a continuous irregular manner (at approximately 3 spikes per second); this pattern distinguishes oxytocin neurons from vasopressin neurons, which fire phasically (bursts of activity lasting from a few seconds to minutes with similar silences between bursts). Without parturition or suckling, oxytocin neurons more gradually increase their firing rate, and hence oxytocin secretion, when stimulated.

In this mode of activity, nitric oxide is generated by oxytocin neurons, which express neuronal nitric oxide synthase, and nitric oxide attenuates their responses to stimulation (Figure 4).

Synchronized bursts Magnocellular oxytocin neurons are stimulated during parturition and suckling to fire in brief, intermittent, high-frequency bursts, synchronized among the supraoptic and paraventricular nuclei. This results in secretion of pulses of oxytocin a few minutes apart. This stimulation involves positive feedback mechanisms, such that passage of the fetus through the birth canal (Ferguson reflex) or sucking at the nipples stimulates pulsatile oxytocin secretion, which drives gradual fetal ejection or causes milk ejection and further oxytocin secretion until the feedback loop is broken by birth of the fetus or satiation of the suckled infant (Figure 4).

The mechanism of this synchronized burst-firing of oxytocin neurons has been intensively investigated but is incompletely understood. Burst-firing is a property of the oxytocin neurons but requires synaptic input; it also requires a local auto- or paracrine action of oxytocin released from magnocellular neuron dendrites. This acts back on oxytocin receptors on the oxytocin neuron dendrites, stimulates further oxytocin release, and activates endocannabinoid production, which after release presynaptically inhibits glutamate and GABA synapses (Figures 3 and 4). These presynaptic actions seem to be important in preparing the oxytocin neurons for synchronized burst-firing. Bursts involve drive by a local glutamatergic network but can be evoked by local noradrenaline action. Bilateral projections to oxytocin neurons from neurons in the medulla oblongata may explain the synchronized bursting among the paired supraoptic and paraventricular nuclei during suckling.

Several brain regions facilitate the milk ejection bursts of oxytocin neurons when acted on by oxytocin, including the ventral tegmentum, dorsomedial hypothalamus, and bed nucleus of stria terminalis.

Glia

Oxytocin neuron dendrites in the supraoptic nucleus project into the ventral glial lamina. Here, astrocytic processes are intimately close to the dendrites but withdraw when the neurons are chronically and strongly stimulated, as in lactation. This permits synaptic remodeling, stimulated by 17β-estradiol and oxytocin, with rapidly increased numbers of axon terminal boutons in contact with the dendrites. However, this reversible plasticity is not essential for the burst-firing capacity of the neurons during lactation.

Central Oxytocin Systems

Neurosecretory material was identified in neurons that project within the mammalian brain in 1954, and some of these projections were later shown with immunocytochemistry to contain oxytocin. Further studies since the mid-1970s have shown oxytocin to be an important neurotransmitter or modulator in the brain, coordinating reproductive behavior with the peripheral actions of oxytocin secreted from the posterior pituitary. The central oxytocin system is

separate from the posterior pituitary system so that it cannot be monitored by measuring oxytocin in the circulating blood. Conversely, the blood–brain barrier effectively prevents circulating oxytocin from entering the brain, and oxytocin receptors are not expressed in the circumventricular organs. However, intranasal application of a large dose of oxytocin has effects on the brain, perhaps via the olfactory bulbs or by entering the brain extracellular fluid; this is useful in human experiments (Figure 5). Circulating oxytocin indirectly influences brain function via its peripheral actions in the fetus ejection and milk ejection reflexes.

Oxytocin and Affiliative Behaviors

Oxytocin action in the brain is important in the processes leading to social, mating, or maternal behaviors – that is, motivation, recognition of others, and some degree of bonding.

Motivation to relate Oxytocin promotes male and female mating behavior, social affiliation, and maternal behaviors. It also switches off eating and aggressive behavior and reduces anxiety to favor reproductive and affiliative behaviors.

Female sexual behavior Lordosis (receptive) behavior in rats is promoted by oxytocin acting in the medial preoptic area and the ventromedial nucleus on oxytocin receptors induced by estrogen (Figure 2). In female mice, oxytocin is not necessary for receptive behavior, and in some species (vole and sheep) oxytocin is inhibitory.

Male sexual behavior In rats, erectile function is promoted by oxytocin actions in the paraventricular nuclei on descending projections. However, in male mice with engineered deletion (knockout) of the oxytocin or oxytocin receptor gene, reproductive capability seems normal. In contrast with inhibitory actions of nitric oxide on magnocellular oxytocin neurons, the activity of parvocellular oxytocin neurons involved in penile erection is stimulated by nitric oxide.

Oxytocin secretion from the posterior pituitary is increased in humans with sexual arousal and at ejaculation and female orgasm; this may help sperm transfer.

Memory and recognition The first studies in 1978 of the central actions of oxytocin on memory showed amnesic actions in an avoidance paradigm. It is now clear that oxytocin has a key role in establishing an olfactory memory for a conspecific in rodents and sheep. Unlike wild-type mice, male oxytocin knockout mice fail to remember recent olfactory contact with the same male. Female oxytocin knockout mice have a similar deficit. The olfactory bulb-to-pyriform cortex processing pathway responds normally, and

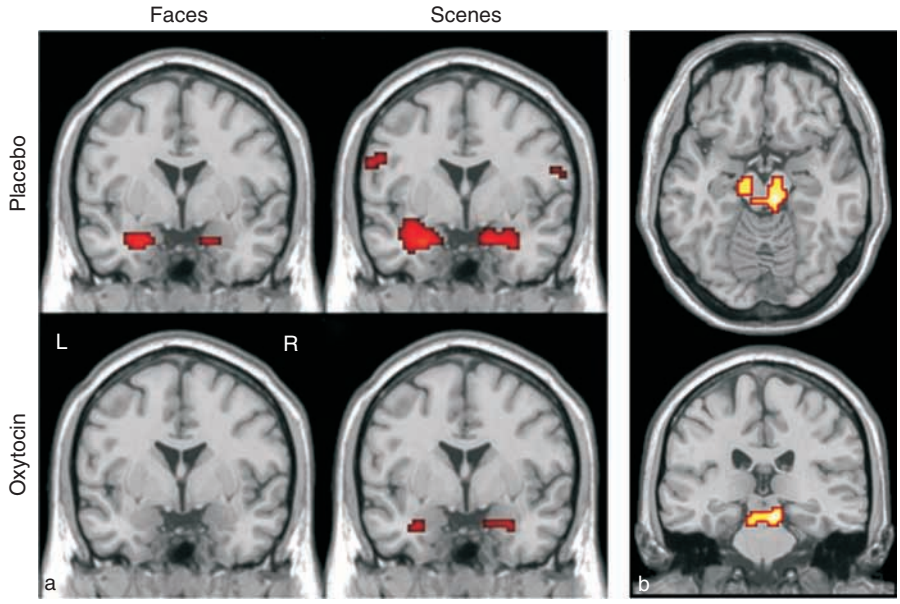

Figure 5 Oxytocin reduces activation of fear circuitry in humans. Blood oxygen level-dependent functional magnetic resonance imaging of human brain coronal sections in subjects given intranasal placebo or oxytocin. (a) Subject exposed to images of frightened or angry faces or frightening scenes. Activation of the amygdala (red) by viewing the images is greatly reduced by oxytocin. L, left; R, right. (b) Oxytocin reduces coupling of amygdala and brain stem; red/yellow (periaqueductal gray and reticular formation) represent relatively greater brain stem activation after placebo versus oxytocin. Adapted from Kirsch P, Esslinger C, Chen Q, et al. (2005) Oxytocin modulates neural circuitry for social cognition and fear in humans. *Journal of Neuroscience* 25: 11489–11493, with permission from the Society for Neuroscience.

the memory defect is due to failure to activate the medial amygdala; oxytocin injection into the medial amygdala rescues the social memory defect. The medial amygdala is a gateway to a circuit involving the bed nucleus of stria terminalis and medial preoptic area, where oxytocin also promotes social memory. In the mouse (but not rat) amygdala, estrogen receptor-α regulates oxytocin receptor expression, and oxytocin peptide gene expression in the paraventricular nucleus is regulated by estrogen receptor-β. These interacting genes form a four-gene micronet that is important in social recognition; inactivation of any of them impairs social recognition through olfactory cues.

Stress Control and Anxiolysis

Oxytocin is a stress hormone in the rat that is secreted from the posterior pituitary as a result of activation of the magnocellular neuron cell bodies. Oxytocin released by axons of these neurons close to the hypothalamopituitary portal vessel system in the median eminence is transported to the anterior pituitary, where in rats it acts on the V1b receptor on corticotrophs to enhance corticotropin-releasing factor (CRF) action in stimulating ACTH secretion. In humans, oxytocin secretion has the opposite action and may indicate anxiety about social attachment (relational distress).

Stress increases oxytocin release in brain regions that regulate neuroendocrine and emotional stress responses, and it may assist stress coping. In the paraventricular nucleus, oxytocin reduces CRF neuron responses to stress, and in the amygdala it is anxiolytic. In humans, in a controlled stress test intranasal oxytocin interacted with social support to reduce anxiety and cortisol secretion.

Appetite

Oxytocin acts in the brain to reduce salt appetite. Oxytocin knockout mice have an uninhibited salt appetite; they also have an increased appetite for sweet drinks. Hence, oxytocin restricts hedonic appetite. Here, oxytocin action in the nucleus tractus solitarii is important; through this region, oxytocin also inhibits gastric motility. Centrally projecting oxytocin neurons are activated by leptin and cholecystokinin, which reduce appetite and meal size. Hence, oxytocin receptors in appetite-regulating brain regions are potential anti-obesity drug targets.

Bonding

Social Bonding

In prairie voles, which are monogamous and form heterosexual pair bonds after mating, central release of oxytocin during copulation in females leads to bonding (vasopressin has this role in males). Oxytocin receptor expression is essential in the network critical for this behavior (lateral septum, medial preoptic area, bed nucleus of stria terminalis, and nucleus accumbens). Hence, the montane vole, an asocial species that does not bond, is not responsive to central oxytocin and lacks oxytocin receptors in these regions (Figure 2). This different receptor distribution is ascribed to differences in the promoter sequence in the 5′ flanking region in the oxytocin receptor gene. Oxytocin action in promoting pair bonding in prairie voles involves dopamine action in the nucleus accumbens in the mesolimbic reward circuit.

In the rat, activation of oxytocin receptors on GABA neurons in the central amygdala (lateral and capsular divisions) inhibits neurons in the medial central amygdala that project to the brain stem to produce fear responses when excited by vasopressin. Thus, oxytocin opposes fear and anxiety-generating actions of vasopressin in the central amygdala, and it functionally uncouples the central amygdala from the brain stem. This has been shown in human functional magnetic resonance neuroimaging studies after intranasal oxytocin administration, which reduces amygdala activation in response to fear-inducing images (Figure 5). In rats and humans, oxytocin appears to promote social interaction by reducing fear or increasing trust in others.

Maternal Bonding

Female rats and sheep are not spontaneously maternal. In sheep, maternal bonding is a prerequisite for maternal behavior, and bonding depends on rapid formation of an olfactory memory by the ewe selectively for her own newborn lamb, which requires oxytocin action in the olfactory bulbs to engage the memory-forming circuitry. Passage of the lamb through the birth canal reflexly activates brain stem noradrenergic neurons, which excite paraventricular nucleus oxytocin neurons projecting to the olfactory bulbs. Here, prior estrogen exposure upregulates oxytocin receptor expression in the granule layer. Oxytocin establishes olfactory memory by simultaneously promoting noradrenaline release in the olfactory bulb, inhibiting GABA neurons and exciting glutamate glomerular cells, which are coincidentally activated by the odor of the newborn lamb. The subsequent strengthening of the activated synaptic connections, via local nitric oxide generation, represents learning of the lamb odor.

In humans, oxytocin secretion stimulated by vaginal delivery or suckling soon after birth correlates with ease of mother–infant bonding. Visual recognition is important in human mothers; looking at pictures of their children activates brain regions

that express oxytocin receptor and that, animal studies show, are important in maternal behavior.

Maternal Behavior

Initiation Bonding between the mother and individual offspring is not seen in rats or mice, but the initiation of maternal behavior is accelerated by central oxytocin release. This action of oxytocin was first demonstrated in 1979 in virgin rats pretreated with estrogen and progesterone, and it is most important when environmental disturbance impairs maternal behavior. Key sites of oxytocin action in rodents are the medial preoptic area and ventral tegmental area since oxytocin antagonist infused into these areas inhibits the onset of maternal behavior.

Even virgin mice express maternal behavior readily, as do postpartum oxytocin knockout mice. Hence, the central oxytocin system is not essential for maternal behavior in mice; however, detailed studies of oxytocin receptor knockout mice show striking defects in natural maternal behavior toward their newborn. Moreover, oxytocin knockout mice are more aggressive and infanticidal than wild-type mice.

Oxytocin-containing nerve fibers, the release of oxytocin, and oxytocin receptors are all found in the many brain regions that control the different elements of maternal behavior. These regions include the paraventricular nuclei (source of oxytocin); hippocampus (recognition of young); limbic structures – medial preoptic area, bed nucleus of stria terminalis, lateral septum, and medial amygdala (reduced aggression and aversion to the young); mesolimbic reward circuitry – ventral tegmental area and nucleus accumbens (hedonic aspects of maternal behavior); and the medial preoptic area and substantia nigra (control of nursing posture). At these sites, oxytocin modulates the release of GABA, noradrenaline, and dopamine to elicit the components of maternal behavior. Coordination of activity in this network to display the combined elements of maternal behavior is likely achieved via widespread axonal projections of parvocellular paraventricular nucleus oxytocin neurons. However, volume transmission by oxytocin release from dendrites of supraoptic and paraventricular magnocellular neurons, as occurs during parturition and suckling, may be important.

Maintenance Release of oxytocin in the brain has a role in maintaining maternal behavior as well as in its initiation. The quantity of oxytocin receptor expression in regions important for the continued expression of maternal behavior, such as the medial preoptic area, central nucleus of amygdala, bed nucleus of the stria terminalis, and nucleus accumbens, correlates with the quality of maternal behavior. Also, brain regions that are important for maternal behavior are activated by suckling and by oxytocin, suggesting that nursing reinforces maternal behavior via central oxytocin (Figure 6).

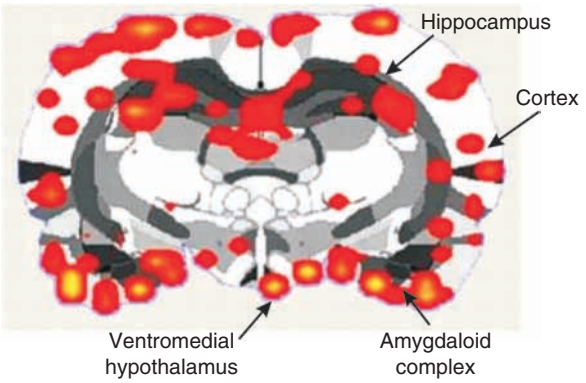

Figure 6 Imaging *in vivo* stimulation of brain activity by oxytocin. Blood oxygen level-dependent magnetic resonance image of a coronal brain slice of a lactating rat given oxytocin into a lateral cerebral ventricle. Red/yellow indicate activation versus a control lactating rat given vehicle. Oxytocin activates widespread regions known to express the oxytocin receptor and that are also activated by suckling, evidently by stimulating central oxytocin release. Adapted from Febo M, Numan M, and Ferris CF (2005) Functional magnetic resonance imaging shows oxytocin activates brain regions associated with mother–pup bonding during suckling. *Journal of Neuroscience* 25: 11637–11644, with permission from the Society for Neuroscience.

Disease and Oxytocin System Malfunction

No defined natural disorders are clearly attributable to mutations of the oxytocin or oxytocin receptor genes. Dominant disabling mutations would not be transmissible through the female line, except in humans, since oxytocin is indispensable for milk transfer to the young. There is evidence from animal studies of a potential major impact that reduced receptor expression or impaired function in the brain could have on behaviors in humans. Abnormalities might include impaired affiliative behavior, affecting quality of mothering and sociability, and increased aggressiveness.

The extensive use of oxytocin in obstetric practice to promote birth has led to questioning whether exposure of newborns to oxytocin has long-lasting adverse effects on the brain. Massive doses of oxytocin have been used in published studies on neonatal rats to show enduring changes in the central oxytocin system and related behaviors. The issue assumes more importance with the development of nonpeptide agonists and antagonists which may enter the brain more readily than oxytocin.

Autism

Reduced oxytocin system activity has been implicated in autism, a highly heritable condition. Findings include loss of oxytocin neurons, low plasma oxytocin, and increased precursor level, suggesting abnormal proconvertase expression. Large-scale molecular genetic studies show linkage to the chromosomal locus containing the oxytocin receptor gene (3p25.3), and polymorphisms have been found in noncoding regions and exons in the oxytocin receptor gene in autistic individuals. In autistic adults, intravenous oxytocin infusion may reduce repetitive behaviors and improve processing and retention of social information. However, abnormalities in more than 20 genes or their proteins may underlie different forms of autism.

Prader–Willi Syndrome

This condition (due to deletion of paternally inherited genes on chromosome 15q11–13) involves childhood obesity and insatiable appetite, perhaps due to reduced central satiety action of oxytocin since the number of oxytocin neurons in the paraventricular nucleus is greatly reduced.

The Future

Research on oxytocin systems in the brain is expected to result in targeted treatments in selected cases of autism, obesity due to overeating, aggressive behavior, male impotence, and anxiety.

See also: Magnocellular Neurosecretory System: Organization, Plasticity, Model Peptidergic Neurons; Neuroendocrine Control: Maternal Behavior; Neuroendocrinology; Neuroendocrinology of Social/Affiliative Behavior; Pituitary Gland (Cell Types, Mediators, Development); Vasopressin/Oxytocin and Receptors.

Further Reading

Argiolas A and Melis MR (2005) Central control of penile erection: Role of the paraventricular nucleus of the hypothalamus. *Progress in Neurobiology* 76: 1–21.

Ashworth DM, Batt AR, Baxter AJ, et al. (2006) Nonpeptide oxytocin agonists. *Drugs of the Future* 31: 345–353.

Born J, Lange T, Kern W, McGregor GP, Bickel U, and Fehm HL (2002) Sniffing neuropeptides: A transnasal approach to the human brain. *Nature Neuroscience* 5: 514–516.

Brown CH and Russell JA (2004) Cellular mechanisms underlying neuronal excitability during morphine withdrawal in physical dependence. Lessons from the magnocellular oxytocin system. *Stress* 7: 97–107.

Burbach JPH, Young LJ, and Russell JA (2005) Oxytocin: Synthesis, secretion, and reproductive functions. In: Neill JD (ed.) *Knobil and Neill's Physiology of Reproduction,* 3rd edn., pp. 3055–3128. San Diego: Elsevier.

Catheline G, Touquet B, Lombard M-C, Poulain DA, and Theodosis DT (2006) A study of the role of neuro-glial remodeling in the oxytocin system at lactation. *Neuroscience* 137: 309–316.

Choleris E, Kavaliers M, and Pfaff DW (2004) Functional genomics of social recognition. *Journal of Neuroendocrinology* 16: 383–389.

Febo M, Numan M, and Ferris CF (2005) Functional magnetic resonance imaging shows oxytocin activates brain regions associated with mother–pup bonding during suckling. *Journal of Neuroscience* 25: 11637–11644.

Israel J-M, Le Masson G, Theodosis DT, and Poulain DA (2003) Glutamatergic input governs periodicity and synchronization of bursting activity in oxytocin neurons in hypothalamic organotypic cultures. *European Journal of Neuroscience* 17: 2619–2629.

Kendrick KM (2000) Oxytocin, motherhood and bonding. *Experimental Physiology* 85(supplement): 111S–124S.

Kirsch P, Esslinger C, Chen Q, et al. (2005) Oxytocin modulates neural circuitry for social cognition and fear in humans. *Journal of Neuroscience* 25: 11489–11493.

Landgraf R and Neumann ID (2004) Vasopressin and oxytocin release within the brain: A dynamic concept of multiple and variable modes of neuropeptide communication. *Frontiers in Neuroendocrinology* 25: 150–176.

Lim MM and Young LJ (2006) Neuropeptidergic regulation of affiliative behavior and social bonding in animals. *Hormones and Behavior* 50: 506–517.

Ludwig M and Leng G (2006) Dendritic peptide release and peptide-dependent behaviours. *Nature Reviews Neuroscience* 7: 126–136.

Marrero HG and Lemos JR (2005) Frequency-dependent potentiation of voltage-activated responses only in the intact neurohypophysis of the rat. *Pflugers Archiv – European Journal of Physiology* 450: 96–110.

Raggenbass M (2001) Vasopressin- and oxytocin-induced activity in the central nervous system: Electrophysiological studies using *in vitro* systems. *Progress in Neurobiology* 64: 307–326.

Russell JA and Brunton PJ (2006) Neuroactive steroids attenuate oxytocin stress responses in late pregnancy. *Neuroscience* 138: 879–889.

Russell JA, Leng G, and Douglas AJ (2003) The magnocellular oxytocin system, the fount of maternity: Adaptations in pregnancy. *Frontiers in Neuroendocrinology* 24: 27–61.

Takayanagi Y, Yoshida M, Bielsky IF, et al. (2005) Pervasive social deficits, but normal parturition, in oxytocin receptor-deficient mice. *Proceedings of the National Academy of Sciences of the United States of America* 102: 16096–16101.

Wang Y-F and Hatton GI (2006) Mechanisms underlying oxytocin-induced excitation of supraoptic neurons: Prostaglandin mediation of actin polymerization. *Journal of Neurophysiology* 95: 3933–3947.

P2X Receptors

Z Li, S Harris, and T M Egan, Saint Louis University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

P2X receptors are ligand-gated ion channels that open in the presence of extracellular adenosine triphosphate (ATP). They are found in both simple parasitic organisms such as *Schistosoma mansoni* and more complex animals, such as humans. Activation usually results in a rapid depolarization of the target cell. This occurs as ATP binds to an extracellular site and opens an integral transmembrane pore that allows small monovalent and divalent cations to flow down their electrochemical gradients. At physiological membrane potentials, the overriding effect of activation is an influx of sodium and calcium (Figure 1(a)). The resulting rise in intracellular sodium depolarizes the membrane and renders the cell hyperexcitable. The influx of calcium activates a number of intracellular calcium-dependent processes such as muscle contraction and synaptic vesicle fusion. Some P2X receptors, particularly those expressed in muscle, also have a demonstrable permeability to chloride, although the concurrent influx of sodium still causes depolarization. At times, P2X receptors initiate the movement of larger molecules across the membrane. For example, the release of proinflammatory interleukin-1β from macrophages is stimulated by activation of native $P2X_7$ receptors, and activation of some recombinant receptors (e.g., rat $P2X_2$ $P2X_4$, and $P2X_7$ receptors and human $P2X_5$ receptors) expressed in model cells leads to the opening of membrane pores permeable to the large nucleic acid stains, ethidium and YO-PRO-1 (Figure 1(b)). The original hypothesis that dye uptake reflects a time-dependent dilation of the intrinsic ion channel (also known as the 'pore-dilation' hypothesis) has come under attack. Specifically, recent data suggest that separate pathways mediate the ATP-gated movement of small cations and large dyes; the evidence that supports this claim is discussed in more detail below.

Genes and Gene Products

The first two P2X receptor genes ($P2X_1$ and $P2X_2$) were identified in 1996 by expression cloning from tissues derived from rat. These two then served as templates for the subsequent discovery of five additional rat genes ($P2X_3$–$P2X_7$) by polymerase chain reaction. Orthologous sequences are found in a number of organisms, including trematodes, fish, frogs, chickens, rodents, cows, and humans. Paralogs and splice-variants have also been described. The seven full-length rat genes encode proteins that vary in length from 384 to 595 amino acids and are 40–50% identical in sequence. Functional receptors are oligomeric complexes of three subunits that selectively combine in both homomeric and heteromeric fashions. Each subtype has a phenotype that is defined by (1) the amplitude of the ATP-gated current; (2) the relative permeability of cations and anions; (3) the percentage of the current carried by calcium; (4) the ability to initiate uptake of nucleic acid stains; (5) the rate and extent of desensitization and resensitization; (6) the ability to form heteromeric complexes with other family members; and (7) sensitivity to a range or agonists, antagonists, and allosteric modulators. Six of the seven genes ($P2X_{1-5}$, $P2X_7$) are capable of encoding functional homomeric receptors that mediate ATP-gated transmembrane currents in heterologous expression systems (Figure 1(c) and (d)). The $P2X_6$ receptor is the exception. Biochemical analyses suggest that monomers of this subtype are partially processed and retained by the endoplasmic reticulum. Despite this fact, there are occasional reports of ATP-gated ionic currents generated in cells expressing only the $P2X_6$ gene. These findings are somewhat controversial and may reflect the unintended formation of heteromeric receptors containing recombinant $P2X_6$ and native (but apparently silent) non-$P2X_6$ subunits. This hypothesis is plausible because several laboratories have reported that the $P2X_6$ subtype forms functional heteromeric complexes in a subunit-specific manner when co-expressed with other family members. Indeed, most P2X receptors are capable of forming functional heteromers that display unique phenotypes when expressed in heterologous systems. Although there are numerous examples of functional heteromeric receptors made of two different subtypes, there is no report of a P2X receptor containing three different family members, even though the existence of such a receptor is possible.

Structure and Function

Because the crystal structure of the receptor is unsolved, a range of less direct biochemical and pharmacological approaches was used to develop a preliminary map of the protein as it sits in the membrane. All functional P2X receptors have the following topology: intracellular N-terminus/first transmembrane segment (TM1)/extracellular segment/second transmembrane segment (TM2)/intracellular C-terminus (Figure 1(e)).

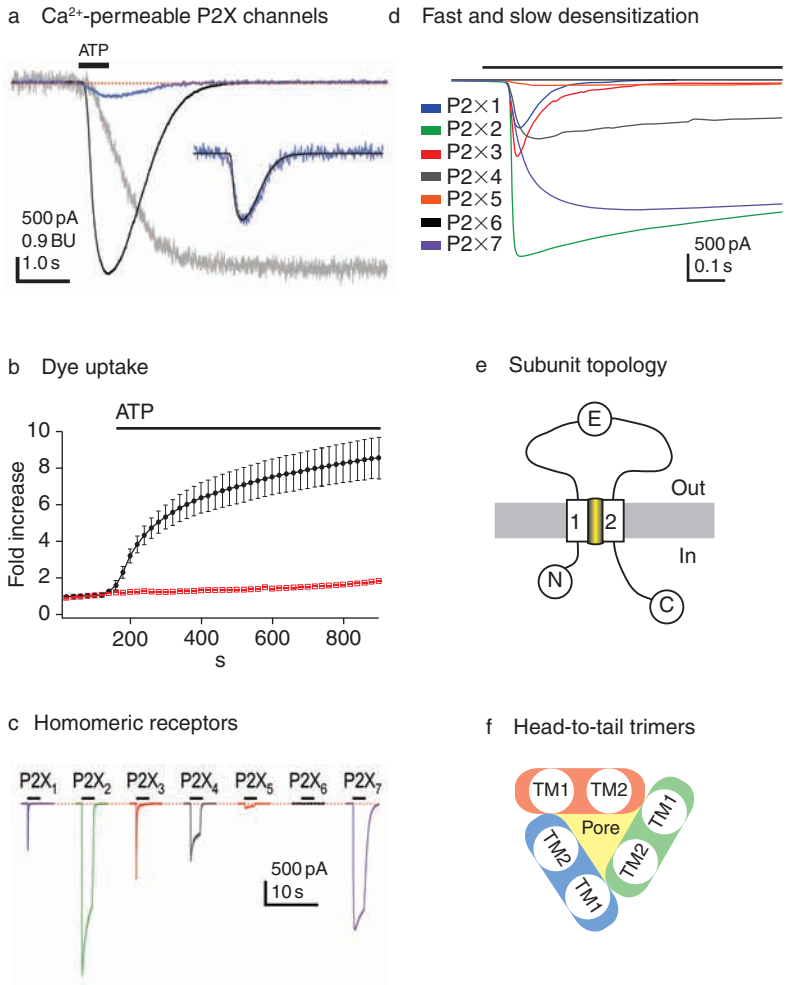

Figure 1 (a) P2X receptors are cation channels permeable to sodium and calcium. The traces show total membrane current (black traces) and fura-2 fluorescence recorded from a HEK-293 cell expressing homomeric rat $P2X_2$ receptors. In this experiment, a decrease in fluorescence (excitation 380 nm; emission 510 nm) indicates an increase in intracellular free calcium. An application of 10 μmol l^{-1} ATP (solid black bar) causes an inward current and a decrease in fura-2 fluorescence. The fluorescence was calibrated, filtered, and differentiated to give the ATP-gated calcium current (blue trace). About 7% of the total current is carried by calcium (compare the amplitudes of the blue and black traces). In the inset, the peak currents of the total current and calcium current are normalized to show that their time courses are identical, as expected if all the rise in intracellular calcium comes from the inward flow of calcium current through the P2X channel. (b)YO-PRO-1 uptake. The traces show the averaged data from 45 test (filled black circles) and 12 control (empty red squares) experiments. In the test experiments, ATP (600 μmol l^{-1}) was applied to six to ten HEK-293 cells expressing the $P2X_2$ receptor and a reporter protein (AsRed). The drug was applied 2 min after the start of the recording. YO-PRO-1 fluorescence (491/509 nm) was collected by a photomultiplier. Transfected cells showed little background fluorescence in the absence of ATP but fluoresced brightly during agonist application (solid bar). Control cells expressed only the reporter protein. (c) The traces show ATP-gated currents recorded from seven different HEK-293 cells transiently expressing one subtype of P2X receptor. All the cells except those expressing the $P2X_6$ receptor show inward current in response to ATP. In some cases ($P2X_1$ and $P2X_3$ receptors), the receptors show rapid desensitization during the ATP application. In another case ($P2X_5$ receptor), the size of the current is relatively small compared with the other subtypes. The concentrations of ATP are as follows: $P2X_1$ and $P2X_3$, 1 μmol l^{-1}; $P2X_2$ and $P2X_4$, 30 μmol l^{-1}; $P2X_5$, 100 μmol l^{-1}; and $P2X_7$, 1 mmol l^{-1}. (d) The current traces of (c) are shown at an expanded timescale to show the rate of onset and desensitization of the six functional homomeric receptors. (e) A representation of a P2X subunit. Functional receptors are formed of three subunits; the diagram shows one subunit sitting in the membrane (gray shaded area). Both the N- and C-termini are intracellular. Two transmembrane domains (1,2) span the lipid bilayer, and these are connected by a large extracellular loop (E) that makes up the bulk of the receptor protein. (f) Oligomeric receptors comprise three subunits that form a cation-permeable pore. Although the pore contains parts of both transmembrane domains, TM2 seems to contribute more than TM1 does to the permeation pathway.

Individual subunits are arranged in a head-to-tail fashion so that the TM1 of each subunit abuts the TM2 of its neighbor (Figure 1(f)). The presence of only two transmembrane segments distinguishes P2X receptors from other ligand-gated cation channels, which typically possess three or more. The N-tail is relatively short (about 25 amino acids) and contains a consensus site for protein kinase C-mediated phosphorylation;

removal of this site leads to an accelerated fade in the amplitude of the ATP-gated current of the $P2X_2$ receptor. The C-tail is longer (about 25–250 amino acids) and more variable and contains multiple sites that mediate subtype-specific effects. These effects include receptor desensitization; membrane trafficking; altered permeation; and interactions with proteins, phospholipids, and lipid polysaccharides. $P2X_7$ receptors with their C-tails altered by truncation or site-directed mutagenesis have normal sodium and calcium currents but lose the ability to show a time-dependent increase in the permeability of larger cations like *N*-methyl-D-glucamine ($NMDG^+$). It is interesting that the time-dependent increase in permeability to ethidium and YO-PRO-1 is unaffected by a truncation that obliterates $NMDG^+$ permeability, suggesting that separate pathways underlie pore dilation (i.e., the gradual increase in permeability to $NMDG^+$) and dye uptake (i.e., the gradual increase in permeability to nucleic acid stains).

Both transmembrane segments line the pore and may contain sites affecting channel function. An explicit role for the TM1 has not been forthcoming, although data suggest that it is involved in transducing agonist binding into channel gating. The functional domains of the TM2 are better described. Within the TM2 are specific amino acids that affect the two properties that P2X receptors share with all ion channels, permeation and gating. Further, the TM2 helps direct the assembly of subunits into oligomeric complexes and is partly responsible for determining the rate and amount of desensitization experienced by different family members.

The two transmembrane segments are connected by a long string of amino acids that form the extracellular extent of the receptor. This string contains ten conserved cysteines, some of which participate in disulfide bonds that help define the tertiary structure of the protein. In contrast; disulfide bonds do not form between adjoining subunits; so they are not involved in stabilizing the quaternary structure. The extracellular loop also contains consensus sequences for the N-linked glycosylation involved in targeting the receptor to the cell surface membrane, for agonist and antagonist binding sites, and for intersubunit binding sites responsible for allosteric modulation of the ATP response by hydrogen ions and zinc.

The ATP Binding Site

The ATP binding site resides at the interface of neighboring subunits of the multimeric receptor complex. The negative charge of the phosphate chain of ATP is attracted to the formal positive charge of conserved lysines and arginines located near the TM1 and TM2, and the adenine ring is coordinated by aromatic amino acids. Different family members respond to different agonists. For example, $\alpha\beta$-methylene ATP is a potent agonist at $P2X_1$ and $P2X_3$ receptors but has little or no effect on $P2X_2$, $P2X_4$, $P2X_5$, and $P2X_7$ receptors. Further, the median effective concentration (EC_{50}) of ATP on the $P2X_7$ is much higher than that of all other family members. The differences in agonist (and antagonist) profiles suggest that the structure of the agonist and antagonist binding sites varies in a subunit-specific manner.

Ion Permeation and Gating

P2X receptors are generally classified as ligand-gated cation channels because most of their physiological effects result from the inward flow of cations across the cell surface membrane. However, some subtypes initiate the movement of both cations and anions across cell surface membranes, making the cation channel classification somewhat inaccurate.

Cation Permeability

All P2X receptors are freely permeable to the three most important physiological cations: sodium, potassium, and calcium. Sodium and potassium permeate equally well. Ding and Sachs have suggested that cations interact with an anionic binding site of low to moderate field strength located at about 20% of the electrical distance from the extracellular surface of the channel. If P2X receptors have only two transmembrane segments, then this binding site must be made of parts of TM1 and/or TM2. The only conserved formal negative charge in these domains is an aspartate in TM2 (e.g., Asp^{349} of the $P2X_2$ receptor). Neutralizing Asp^{349} has no effect on cation permeability, making it unlikely that it contributes to regulation of current flow through the channel. Rather, the site may be formed of polar amino acids that reside in a narrow constriction of the pore, although such a model is not applicable to all P2X subtypes (see below). An alternative hypothesis is that other parts of the protein form a reentrant pore loop that forms a selectivity filter in the pore, analogous to the P loop of the potassium-selective KcsA channel or the M2 loop of the glutamate receptor. However, no such reentrant loop has been identified for any P2X receptor to date.

The relative permeability to calcium is higher than that to sodium or potassium and varies in a subunit-specific manner. For example, about 6% of the current through the $P2X_2$ receptor is carried by calcium, and this is large enough to facilitate excitatory neurotransmission onto stratum radiatum interneurons of the

hippocampus. By contrast, the calcium currents of $P2X_1$ and $P2X_4$ receptors are twice as large as those of the $P2X_2$ receptor. The molecular mechanism(s) responsible for the subunit-specific variability are unknown. Two complementary methods have been proposed. First, site-directed mutagenesis of hydroxyl-bearing amino acids within the TM2 of the $P2X_2$ receptor decreases calcium flux, perhaps by interfering with a binding site involved in partially dehydrating permeating ions as they pass through a narrow constriction in the pore. However, the polar residues are missing in some of the subtypes, and therefore hydroxyls are not fully required for calcium flux throughout the family. Second, the COO^- side chains of acidic amino acids found at the extracellular ends of the transmembrane domains may provide an electrostatic force that interacts with calcium in the mouth of the channel pore. Indeed, subtypes that possess these formal charges (i.e., $P2X_1$ and $P2X_4$ receptors) transduce significantly higher calcium fluxes than those that do not (i.e., $P2X_2$ receptors).

Anion Permeability

Hume and Thomas were the first to describe an anion-permeable, ATP-gated ion channel. They measured a shift in the reversal potential of the ATP-gated current of chick skeletal muscle when extracellular chloride was replaced by larger anions. Such a shift suggests that the outward movement of chloride contributes to the ATP-evoked depolarization. The responsible gene was cloned by Soto and others and found to encode a chick $P2X_5$ receptor capable of transducing significant chloride flux in heterologous expression systems. Orthologs have been described in zebra fish, bullfrog, mouse, rat, and humans. Most of these are poorly characterized because their ATP-gated currents are small (e.g., less than 100 pA even at the extremes of physiological membrane potentials), and therefore their anion permeability is unknown. North et al. described a full-length human $P2X_5$ receptor that is different from other $P2X_5$ receptors in that it transduces large (nA) currents in response to ATP. Like the smaller currents of the chick $P2X_5$ receptor, the human ortholog displays a significant permeability to chloride.

Pore Dilation versus Dye Uptake

Sodium, potassium, calcium, and chloride carry most of the ATP-gated membrane current during the first few seconds of application of ATP. Longer applications evoke an additional conductance as the membrane permeability to large organic cations (like $NMDG^+$) gradually increases. The increase in permeability suggests that the narrowest portion of the pore dilates by at least 3 Å from an initial diameter of 9–11 Å. At the same time that the permeability to $NMDG^+$ begins to increase, cells expressing $P2X_2$, $P2X_4$, $P2X_5$, or $P2X_7$ receptors also gain the ability to take up large nucleic acid dyes (i.e., ethidium bromide or YO-PRO-1). A simple hypothesis that could explain both effects is that the P2X pore moves through three states. The first is the nonconducting, closed state seen in the absence of ATP. The second is open in the presence of ATP and permeable to small cations with geometric diameters of $<\sim 11$ Å. The third opens in the continued presence of ATP and with a delay; it occurs when the initial open state widens to a new diameter that is big enough to allow large molecules to pass through the channel. The three-state model is sufficient to explain much of the biophysical data currently available. However, at least in the case of the $P2X_7$ receptor, recent data suggest that the uptake of dye does not occur through the pore of the $P2X_7$ channel *per se*. Surprenant, North, and others have shown that deleting a short stretch of 18 amino acids just intracellular to the end of the TM2 (called the Cys-rich region because it contains many cysteines) prevents the change in $NMDG^+$ permeability but has no effect on the uptake of ethidium. Such an effect is difficult to explain if both molecules use the same pathway to enter the cell. Further, both the Surprenant and Dahl laboratories have shown that RNA interference directed against recently discovered pannexin-1 hemi-channels significantly reduces dye uptake without effecting cation flux through the pore of the $P2X_7$ channel. Thus, it seems that the $P2X_7$ channel uses two different methods to move large ions across the cell surface membrane. The first is a gradual dilation of the integral P2X pore that ultimately results in an increase in permeability to $NMDG^+$. The second is an indirect activation of ethidium-permeable pannexin-1 channels that may result from a protein–protein interaction with the $P2X_7$ receptor. These findings are exciting and novel, but additional experiments are needed to prove the hypothesis they seem to support.

Pharmacology

Although the field has advanced in the past few years, it remains difficult to discriminate among P2X receptors on the basis of pharmacology alone. Gever et al. have provided an excellent review of the recent advances in purinergic pharmacology.

Not only do many of the currently available reagents discriminate poorly among the P2X receptors themselves; these same drugs also act on ATP-gated

metabotropic P2Y receptors. However, a few pharmacological trends do hold true. P2X receptors can be divided into three general classes on the basis of the phenotype of the ATP-gated current and the pharmacology of the receptor. The first class includes the $P2X_1$ and $P2X_3$ receptors. These receptors display agonist-gated currents that desensitize rapidly (<1 s), have a high affinity for both ATP and its analog α,β-methylene ATP (the EC_{50} for both agonists is about 1 $\mu mol\,l^{-1}$), and are blocked by nanomolar concentrations of the purinergic antagonist TNP-ATP. The second class includes $P2X_2$, $P2X_4$, and $P2X_5$. These receptors show a much slower rate of desensitization (tens of seconds), are less sensitive to ATP (EC_{50} about 6–30 $\mu mol\,l^{-1}$), are relatively insensitive to α,β-methylene ATP, and are blocked by micromolar concentrations of TNP-ATP. Heteromeric receptors that contain members of both the first and the second classes generally adopt the agonist and antagonist sensitivities of the first class; that is, they show robust responses to low concentrations of both ATP and α,β-methylene ATP and are readily antagonized by TNP-ATP. The third class is made of the $P2X_7$ receptor alone. This receptor shows the lowest ATP potency ($EC_{50} >$ 100 $\mu mol\,l^{-1}$) and is the only member of the family in which prolonged activation is reported to result in cell vesiculation and death.

Distribution and Physiology

P2X receptors are expressed throughout the body. They are found in a range of tissues, including nerve, muscle, glia, epithelium, bone, and hemopoietic. Most tissues express more than one subtype, and current dogma suggests that the majority of native receptors are heteromeric complexes of two or more subtypes. The $P2X_1$ receptor is predominately expressed in smooth muscle. Genetically altered mice that lack this receptor are infertile because the ATP-mediated contraction of the vas deferens is missing. Further, $P2X_1$ receptors are also widely expressed in the cardiovascular system, where they are found in heart, blood vessels, and platelets. However, no obvious effect on cardiovascular function is found in $P2X_1$-knockout mice, and the role of ATP in these tissues is still unknown. $P2X_2$ and $P2X_4$ receptors are found in neurons and glia throughout the central and peripheral nervous systems, where they facilitate neurotransmission by both presynaptic and postsynaptic actions. In contrast, the distribution of the $P2X_3$ receptor is more restricted. It is predominately found on the sensory afferent neurons of the peripheral nervous system, where it mediates gustation; mechanosensory regulation of urinary bladder function; and transmission of persistent, chronic, neuropathic and inflammatory pain. $P2X_5$ receptors are found in skeletal muscle and may play a role in regulating proliferation and differentiation of tumors. $P2X_7$ receptors are found in mast cells, macrophages, and lymphocytes, where they participate in the immune response.

Presynaptic and Postsynaptic Effects

In general, two separate but related functions are described for these receptors in neurons. First, P2X receptors allow cations to enter the cell, leading to membrane depolarization. Direct somatic depolarization by release of endogenous ATP has been difficult to demonstrate, perhaps because of the rapid degradation of the nucleotide by endonucleotidases. However, good examples exist including depolarization of smooth muscle in the vas deferens and neurons of the celiac ganglia, spinal cord, myenteric plexus, medial habenula, hippocampus, and somatosensory cortex. The outcome of depolarization is the propagation of excitation from one cell to the next. The second effect of ATP is presynaptic. P2X receptors increase intracellular calcium in nerve terminals, both directly, by conducting calcium, and indirectly, by activating voltage-gated calcium channels. In both cases, the rise in intraterminal calcium can lead to an increased probability of neurotransmitter release.

Conclusion

P2X receptors are some of the newest members of the ligand-gated ion channel superfamily. Their unique topology, ubiquitous distribution, high calcium flux, and ability to induce the movement of very large (up to about 900 kDa) molecules across the cell-surface membrane suggest that they play important roles in regulating the homeostasis of a range of organisms. In upcoming years, a high-resolution map of a P2X receptor structure may well be forthcoming; this map, in combination with functional studies, will greatly improve our understanding of the molecular physiology and pharmacology of the receptor. Further, recent studies that demonstrate direct protein–protein interactions of P2X receptors and hemichannels, other ligand-gated channels, and Fe65 suggest that the influence of ATP in the physiology and pathophysiology of humans is underappreciated.

See also: Adenosine; Adenosine Triphosphate (ATP); Adenosine Triphosphate (ATP) as a Neurotransmitter; Adenosine Receptor Mediated Functions; Calcium Waves: Purinergic Regulation; Pharmacology of Sleep: Adenosine; Purinergic Receptors; Purines and Purinoceptors: Molecular Biology Overview.

Further Reading

Gever JR, Cockayne DA, Dillon MP, Burnstock G, and Ford AP (2006) Pharmacology of P2X channels. *Pflügers Archiv: European Journal of Physiology* 452: 513–537.

Khakh BS and North RA (2006) P2X receptors as cell-surface ATP sensors in health and disease. *Nature* 442: 527–532.

North RA (2002) Molecular physiology of P2X receptors. *Physiological Reviews* 82: 1013–1067.

Roberts JA, Vial C, Digby HR, et al. (2006) Molecular properties of P2X receptors. *Pflügers Archiv: European Journal of Physiology* 452: 486–500.

Samways DS, Li Z, and Egan TM (2006) Binding, gating, and conduction of ATP-gated ion channels. In: Arias H (ed.) *Biophysical Aspects of Ligand-Gated Ion Channel Receptor Superfamilies*, pp. 419–443. New York: Research Signpost.

Pain and Genes

M Devor, Hebrew University of Jerusalem, Jerusalem, Israel

© 2009 Elsevier Ltd. All rights reserved.

Pain Variability

Chronic pain is an unsolved health problem of major proportions worldwide. Although pain *per se* is rarely life threatening, it takes an enormous toll in quality of life and accounts for huge financial outlays for treatment and compensation. Large surveys indicate that at any given time nearly 20% of adults report suffering from chronic pain of an intensity and duration that embitters their lives, the most common conditions being back pain, arthritis, and headache, including migraine. One of the greatest puzzles concerning pain, both acute pain in response to noxious stimuli and more prolonged pain following injury or disease, is its great variability from individual to individual, even when the provoking stimulus or disorder is essentially identical. A dramatic example is stump and phantom limb pain in amputees. Some amputees report no pain whereas in others severely disabling pain lasts a lifetime, even when the level of the amputation is the same. The fact that re-amputation usually fails to relieve phantom limb pain points to factors associated with the patient rather than the surgery as the underlying cause. Variability is also the rule in the day to day. A slap in the face that provokes sobs in one person will be laughed off by another. We each have a different pain threshold.

Variability in pain expression has traditionally been attributed to psychosocial and cultural factors – personality and upbringing. It is also irrevocably entangled with the question of how much the outward expression of pain actually reflects the pain felt by the individual, inside. In some societies, for example, people do not hesitate to express their pain, whereas in others children are taught to be stoical. Even for a given individual and a stimulus of fixed intensity, pain expression may vary wildly with the circumstances. Did that slap in the face indicate a challenge or an insult, or did the tears come just because it hurt? In most cultures, there is a dramatic sex difference in the degree to which pain expression is sanctioned. Boys are taught not to be cry babies, but girls are expected to show tears. Does the resulting difference in outward pain behavior only reflect social norms, or is there also a real difference in the amount of pain felt among boys and girls in response to identical painful stimuli? A large body of evidence indicates that women are much more likely than men to suffer from a variety of chronic pain conditions and that women are less tolerant of noxious stimuli when tested in laboratory experiments. To what extent does this reflect pain expression as opposed to pain experience? Is it possible that the socialization of boys and girls actually alters the amount of pain experienced later in life and not just how pain is expressed?

Pain is inherently a private sensory and emotional experience. It cannot be shared. It is felt in the first person (I) and can only be inferred by outward behavior and language in the second and third persons (you and he or she). This fact of life is likely to remain with us for the foreseeable future. There is no method on the horizon that will permit me to crawl into your brain and sample your pain experience. However, advances in noninvasive brain imaging may soon make it possible to know whether someone is faking his or her tears because the amount of neural activity in certain parts of the brain tends to correlate with reported pain experience. Moreover, these brain activations increase and decrease when the amount of pain felt in response to a fixed stimulus is altered by factors such as anticipation, hypnosis, and the presence of analgesic drugs.

Heritability of Pain

In addition to socialization and other psychosocial and environmental factors, variability in pain response is also due to differences in the wiring of our individual brains as determined by our genes. The first solid indications of this comes from rare familial disorders inherited by certain family members, but not by others, in which pain is either absent (e.g., congenital insensitivity to pain with anhydrosis (CIPA)) or excessive (e.g., Fabry's disease or familial hemiplegic migraine (FHM)). Such conditions can usually be traced to disease-causing mutations in specific genes. Evidence for heritability is also available from the study of twins. Pain as a normal trait (e.g., in response to noxious heat and cold) and the presence of common painful conditions such as backache and sciatica are more concordant in identical than fraternal twins. The idea here is that identical twins share all their gene variants whereas fraternal twins share only half. Recently, investigators have begun to undertake studies using modern genetic methodologies aimed at actually identifying the specific genes that control normal pain sensitivity and the likelihood of suffering pain in response to injury or disease.

Basically, all biological traits are determined by genes, environment, and interactions between genes

and environment. Why should we be surprised to learn that pain susceptibility is affected by nature as well as by nurture? The short answer is that we should not be. Nonetheless, because of the overwhelming influence of environment on both the clinical manifestations of pain response in patients and in our day-to-day experience of pain, the emergence of an important genetic component took many people by surprise.

Beyond the simple knowledge that pain response is in part heritable, the new awareness of pain genes has at least two important implications. First, contemporary biology has powerful new ways of exploiting this knowledge in the interests of improving our understanding of pain mechanisms and ultimately improving diagnosis, prognosis, and treatment. Second, knowledge that there are pain genes has immediate implications for pain patients here and now. People who report having more pain than is expected from easily observable signs of injury and disease are often stigmatized as complainers and malingerers. Sometimes they are even suspected of being outright liars, trying to cheat the system and obtain undeserved benefits from insurance companies, employers, or caring family members. The knowledge that pain is variable and that one person may have much more pain than another for genetic reasons, through no fault of his or her own, may provide some comfort. Of course, the fact that pain genes exist does not mean that malingerers and cheats do not.

Pain Genes

What are pain genes? Genes are parts of DNA molecules that encode for proteins. Genes do not encode for sensory and emotional experiences such as pain. Only a conscious brain can experience pain. But this does not mean that all of the 25 000-odd genes that are required to make a human body are pain genes. A pain gene is a gene with a polymorphism (i.e., a variation in the sequence of DNA base pairs) in a region that encodes for a protein that affects pain susceptibility. Sequence variations can also affect pain if they occur in a region that regulates the expression of pain genes. For most of our genes and regulatory regions (99.9%), the base pair sequence is identical from one person to the next and the proteins that genes yield are no different in the way that they work. For some genes, however, there are rare variants (mutations) or common variants (polymorphisms) that do affect the functioning of the proteins encoded. Sequence variants of a given gene are called 'alleles' of the gene. For most genes with functional variations among people, there are only a few common alleles. What makes us genetically different from one another is the mix-and-match combination of alleles that we carry.

At present, we have no idea how many genes there are in which allelic variation affects pain response. The question "How many pain genes are there?" is not even well defined. Some gene variants may have major effects on pain, whereas others may have subtle effects. Polymorphisms in some genes affect pain only if accompanied by polymorphisms in others, such that pain sensitivity is determined by a combination of alleles, not by just a particular one. Some pain genes may have no effect on pain at all except under unusual environmental circumstances (e.g., at high altitudes, or if your sister-in-law made you particularly exasperated). This is a gene–environment interaction.

An essential distinction needs to be made between alleles that cause (or increase the predisposition to) diseases that may be painful and alleles that predispose to pain in the presence of identical injury or disease. An example of a disease-susceptibility gene is a gene that increases the likelihood of developing diabetes, a metabolic condition that often leads to pain-provoking nerve injury (diabetic neuropathy). Indeed, quite a large number of genetic mutations have been identified that cause various types of inherited peripheral neuropathies, and some of these conditions are associated with pain. A prominent example is the $Thr^{124}Met$ base pair substitution that causes painful Charcot–Marie–Tooth (CMT-2) neuropathy. In each of these neuropathies, there is visible damage to nerve fibers in affected patients, although this in itself does not explain why some of these conditions are painful and others are not. A far less obvious example of genetic variation that predisposes a person to experiencing pain is the alleles associated with risk-taking behavior. The more risks you take, the more likely you are to suffer from a painful injury. Like pain, there are no genes that encode for risk-taking behavior. But there may be a chain of causality from gene to protein to behavior – for example, an allele that encodes for muscle proteins may make a young man tend to be more attractive and athletic, more outgoing, more likely to engage in risky sports, more likely to be hurt playing football, and hence more likely to suffer pain.

Disease-susceptibility genes that affect pain directly are of obvious medical importance, and even the ones that affect pain indirectly, like risk-taking genes, may have more than simple entertainment value. But if our aim is to better understand the neural mechanisms that determine pain intensity and to facilitate the development of better analgesic drugs, pain-susceptibility genes are of more immediate interest than disease-susceptibility genes. For example, it is possible that pain-susceptibility genes determine which amputee will have severe phantom limb pain and which, with the same amputation level (the same disease), will have none. Identification of the gene variants that make the

difference might point directly to the neurobiological processes that are responsible for phantom limb pain in the first place. Among the various possibilities is that an amputation pain-susceptibility gene might determine to what extent sensory neurons that have been damaged as a result of the amputation generate spontaneous abnormal electrical impulse discharges. Alternatively, the critical genetic polymorphism may have nothing to do with primary sensory neurons but, rather, affect amplification mechanisms within the spinal cord. Either type of effect, and many others, could account for the difference in pain between the suffering and the pain-free amputee. And knowing which could lead scientists to new and better pain relievers.

Finding Pain Genes

How do we find pain-susceptibility genes? The search for pain genes, like the search for genes that control any other inherited trait, begins by comparing the genetic makeup of individuals or groups with contrasting pain phenotypes (i.e., pain responses). For example, which alleles in which genes are identical in amputees with phantom pain and different from amputees without pain? It is relatively straightforward to identify a Mendelian gene, a gene that affects a pain phenotype so powerfully that its alleles alone determine whether you will have pain or not. The approach, linkage analysis, involves studying families in which some members carry the mutation (and have the painful condition) and others do not. Genetic differences are identified which correlate with the pain phenotype of the individual family members. Nearly all medically important disease-susceptibility genes identified to date were identified by tracing linkage in large families. This is also true of conditions that are painful. A striking example is FHM, for which specific mutations have been found in four separate genes (in different FHM families), all of which cause the painful migraine with hemiplegia characteristic of this disease. Another example is erythermalgia, an inherited condition associated with red warm hands and/or feet and severe burning pain. The cause in all families studied to date is a rare mutation in a single gene. Interestingly, in both conditions, the mutated genes encode for proteins whose function is known and makes sense in terms of current pain physiology. There is also the anticipation that this line of work might uncover genes never before linked to pain. Such information might open up new chapters in pain physiology.

An important limitation to linkage analysis in the search for pain genes in families is that it only works for traits that arise spontaneously, such as migraine headache and erythermalgia. Many chronic pain conditions, phantom limb pain in amputees, for example, result from acquired injuries or disease. It is exceedingly unlikely that we will ever find large families in which all the members have had a below-the-knee amputation or suffer from a uniform degree of diabetic neuropathy. In addition, susceptibility to chronic pain is likely to be affected by several genes and to be further modified by environmental factors. Linkage analysis in families works efficiently only for genes of major effect. The alternative is to compare genetic makeup in matched groups of patients, with and without pain, who are not related. This is called an association study. In a typical association study, we compare the genomes of two unrelated groups of individuals that differ in the trait of interest but that are as similar as possible with regards to other parameters, such as environment and ethnicity. Because of the much greater variability in genetic background among people who are unrelated, the number of people who must be tested is much larger. The specific number required depends on how strong an effect the gene(s) in question has in determining the trait. For a Mendelian gene, a few dozen samples may be sufficient. However, in the more common situation in which any given gene has only a partial effect in determining the trait, hundreds or even thousands of individuals need to be included in each group. Due to practical (often budgetary) constraints, association studies are frequently conducted with smaller sample sizes than necessary. The reliability of results obtained from such studies, as assessed by successful replication, is often disappointing. An example is an early report that the $Val^{158}Met$ polymorphism, a simple base pair substitution in *COMT*, the gene that encodes for an enzyme (catechol-*O*-methyl transferase) involved in noradrenaline synthesis, greatly increases pain susceptibility. A number of attempts to replicate this finding have obtained a much weaker signal or none at all. Single nucleotide polymorphisms (SNPs) of this sort have also been reported to affect pain in osteoarthritis, low back pain, postherpetic neuralgia, and other common medical conditions.

A potential alternative to association analysis in humans, actually a complementary approach, is to begin by carrying out linkage analysis using an animal model of a human pain trait. Such studies are almost always done in mice because of the superior experimental resources available for genetics studies in this species (**Figure 1**). The hope is that, after identifying a pain-susceptibility gene in mice or a modest number of gene candidates, we might succeed at finding important polymorphisms in the homologous human gene(s) using patient cohorts of manageable size. The mouse study tells us in which part of the human genome to look. An interesting example is the 2004 work by Seltzer and colleagues, who used

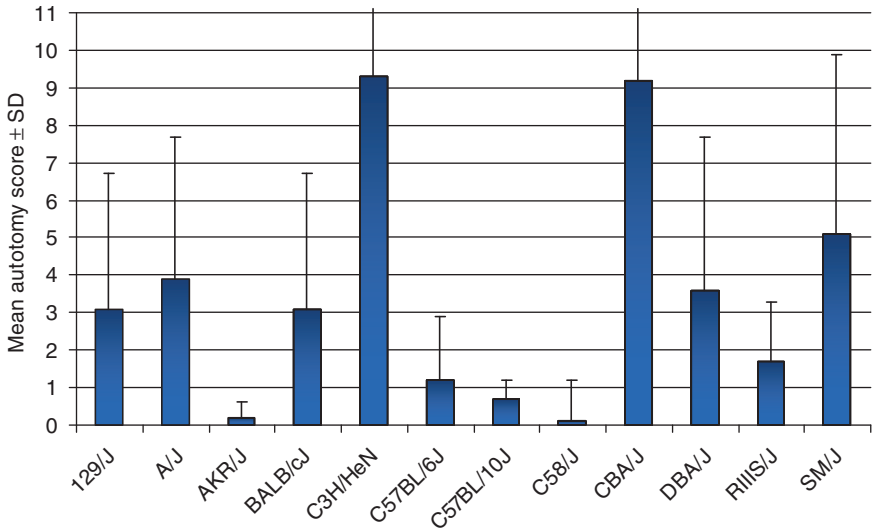

Figure 1 Graph of pain sensitivity in the neuroma model of neuropathic pain in 12 different inbred mouse strains (all males). The most sensitive strains are C3H/HeN and CBA/J, and the least sensitive strains are C58/J and AKR. Individual mice within an inbred strain are genetically identical to one another. Therefore, trait variability within each strain (column error bars) gives an indication of the effect that environmental factors had on the animals' response, whereas variability among strain means gives an indication of the degree to which the pain trait is heritable. SD, standard deviation. Data from Mogil JS, Wilson SG, Bon K, et al. (1999) Heritability of nociception, I: Responses of 11 inbred mouse strains on 12 measures of nociception. *Pain* 80: 67–82.

the recombinant inbred approach (a variant of family linkage analysis) in the mouse neuroma model of neuropathic amputation pain. This analysis identified a region in the mid-part of mouse chromosome 15 that apparently harbors a gene of major effect for this pain trait. Seltzer was then able to focus his attention on the corresponding region of the human genome, chromosome 22, in DNA samples from a cohort of limb amputees and women who had undergone mastectomy, some of whom had chronic pain and some of whom did not. This strategy has yielded a promising pain gene candidate, but verification is still needed.

Pharmacogenetics and Other Applications

Another example of the mouse–human comparison approach illustrates the newly emerging field of pharmacogenetics. Mogil and colleagues (2003) found that contrasting sensitivity to κ-opioid-type pain relievers in females (but not males) of two inbred mouse strains is due to an allelic difference in a genetic region close to *Mc1r,* the mouse melanocortin-1 receptor gene. Allelic variation in the human homolog of *Mc1r* was previously known to occur in light-skinned redheads. Putting one and one together, they found that redheaded women are much more sensitive to κ-opioid analgesics than are men or women who are not redheads. Common base pair substitutions in *OPRM1,* the gene encoding for the human morphine receptor, and in *CYP2D6,* the gene that encodes for a liver enzyme that metabolizes codeine, have an important influence on the efficacy of morphine and codeine as pain relievers in people who carry different alleles of these genes. One of the first clinical applications of pain genetics is likely to be arrays (gene chips) that will indicate which drugs, including pain relievers, might be particularly effective, or toxic, for a given patient.

Genetic linkage analysis and association studies may point to a region of DNA that harbors many genes, with more than one being an attractive candidate for the gene responsible for the trait in question. We still need to chose which is the gene that really counts. And even when a polymorphism in a particular gene is indicated, it is wise to verify that we have indeed found the target gene using an independent methodology. A variety of methods are available to do this, mostly using animal models. Pharmacological reagents, agonists or antagonists, are sometimes helpful. Drugs, however, often interact with molecular targets in addition to the intended one. A more specific, although resource-intensive approach, is to knock out the gene in question using the method of homologous recombination, so that its protein product is not expressed in any cell in the body (global knockout) or in a particular class of cells (conditional knockout). The key question is whether this knockout will duplicate the phenotype attributed to the candidate gene being tested. For example, will the knockout mouse have more pain sensitivity than normal littermates? Particular genes can likewise be expressed in excess, leading to the prediction that this will cure the

pain phenotype. Another approach is to use antisense oligonucleotides or small inhibitory RNA (siRNA) molecules to block the transcription of the target gene into messenger RNA and hence to prevent the expression of the gene's protein. Specific antibodies can also be used to neutralize the protein product itself. Such ablative procedures can, in principle, form the basis of new pain therapies. Therapeutic antibodies for pain control in certain conditions have recently become available, and there is optimism that siRNA technology will also prove its practicability. Further down the road are proposals to develop gene therapy for pain control. For example, a patient's cells could be altered genetically so that they express endogenous morphine-like protein pain-relieving molecules specifically in body locations where they are required. More speculatively, it might be possible in the future to actually repair mutated genes in patients or insert altered, or even completely novel, genes using viruses as carriers (vectors) of genetic information. Genetic engineering is fast becoming a medical reality.

Finally, there are a variety of ways in which genetic thinking can be applied to the discovery of pain mechanisms that do not involve actually identifying pain genes. For example, in 1990 Devor and Raber applied selective breeding to a pain phenotype in outbred rats. Hindlimb nerves were injured to induce experimental neuropathic pain, and individual male and female rats that showed high pain sensitivity were then interbred, as were males and females that showed low pain sensitivity. First-generation offspring did not show much of a change in pain phenotype. But after selection pressure had been applied for five successive generations, nearly all offspring of pain-sensitive animals were pain sensitive and nearly all offspring of pain-resistant animals were pain resistant. These high and low pain-prone rats could then be used to see what other traits co-selected with pain. Because the animals are alike genetically except for genes that affect the pain response, it is likely that any behavioral or neurophysiological characteristic that contrasts between them is determined by the same genetic variants that are responsible for the difference in pain sensitivity. A related approach, correlational analysis, does not require genetic selection lines. In this approach, 10–15 strains of inbred mice are graded on a particular trait (**Figure 1**), say sensitivity to heat stimuli applied to the paw. If another trait, say neuropathic pain, shows the same rank order of sensitivity across strains, it can be concluded that these two traits are controlled by similar genes and hence probably share a common underlying physiological mechanisms. Unlike gene knockout, gene-based pharmacological approaches, and genetic engineering, information on how the pain system works can be gleaned from linkage analysis, association studies, genetic selection, and correlational analysis without prior knowledge, assumptions, and hypotheses.

The genetic revolution has only just started to impact on the field of pain science and medicine. There is well-founded optimism about the potential power of linkage analysis and association studies, particularly when undertaken using the mouse–human comparison approach, to generate new knowledge. The payoff in terms of new therapeutic options is likely to be not too far down the road.

See also: Neuropathic Pain; Pain and the Sympathetic Nervous System; Pain Pathways: Descending Modulation; Pain Management; Pain: Central; Pain: Neuroimaging; Spinal Cord Pain Systems.

Further Reading

Battie MC, Videman T, and Parent E (2004) Lumbar disc degeneration: Epidemiology and genetic influences. *Spine* 29: 2679–2690.

Breivik H, Collett B, Ventafridda V, Cohen R, and Gallacher D (2006) Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. *European Journal of Pain* 10: 287–333.

Craft RM, Mogil JS, and Aloisi AM (2004) Sex differences in pain and analgesia: The role of gonadal hormones. *European Journal of Pain* 8: 397–411.

Devor M and Raber P (1990) Heritability of symptoms in an experimental model of neuropathic pain. *Pain* 42: 51–67.

Mogil JS (ed.) (2004) *The Genetics of Pain: Progress in Pain Research and Management,* vol. 28. Seattle, WA: IASP Press.

Mogil JS, Wilson SG, Bon K, et al. (1999) Heritability of nociception, I: Responses of 11 inbred mouse strains on 12 measures of nociception. *Pain* 80: 67–82.

Mogil JS, Wilson SG, Chesler EJ, et al. (2003) The melanocortin-1 receptor gene mediates female-specific mechanisms of analgesia in mice and humans. *Proceedings of the National Academy of Sciences of the United States of America* 100: 4867–4872.

Rainville P, Duncan GH, Price DD, Carrier B, and Bushnell MC (1997) Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science* 277: 968–971.

Seltzer Z, Wu T, Max MB, and Diehl SR (2001) Mapping a gene for neuropathic pain-related behavior following peripheral neurectomy in the mouse. *Pain* 93: 101–106.

Zubieta JK, Heitzeg MM, Smith YR, et al. (2003) COMT val158-met genotype affects mu-opioid neurotransmitter responses to a pain stressor. *Science* 299: 1240–1243.

Pain and Plasticity

R-R Ji and Y Kawasaki, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Nociceptor and Nociceptive Pain

During evolution, living organisms develop a specialized apparatus to sense their immediate environment in order to protect themselves from potentially hazardous situations. This special apparatus is called nociceptor, which can detect noxious stimuli from the environment. Intense stimulation of high-threshold nociceptors will elicit a pain sensation, through a pathway that is initiated from the action potential generated in the peripheral nociceptor terminal, conducted via thin unmyelinated C fibers and myelinated Aδ fibers to the spinal cord dorsal horn in the central nervous system (CNS), finally projecting to the cortex via a relay in the thalamus. This acute 'ouch' pain is called nociceptive pain or physiological pain, and it has a protective role that warns of potential tissue damage. Many pain transduction molecules have been identified, such as the thermal receptor transient receptor potential (TRP) ion channel family. Whereas TRPV1 and TRPV2 detect heat stimuli, TRPA1 and TRPM8 may sense cold stimuli. Activation of TRPV1 after a heat stimulus (>42 °C) generates inward currents in the nociceptor peripheral terminal and results in action potentials in the nociceptor axon, leading to the sensation of pain.

Nociceptive Sensitization and Clinical Pain

A further evolution of the early pain system was the development of the capacity to produce increases in sensitivity after injury, termed nociceptive sensitization. This sensitization results in hyperalgesia, an increased responsiveness to noxious stimuli, and allodynia, a painful response to normally innocuous stimuli. Nociceptive sensitization was found even in simple organisms such as *Aplysia*, in which an intense noxious stimulus can produce a long-lasting sensitization of the gill withdrawal reflex due to increased synaptic strength. The persistence of a heightened responsiveness of the nervous system after a brief noxious stimulus has clear parallels with memory, where information needs to be stored and retrieved. Although the gill withdrawal reflex in *Aplysia* could be used as a model for nociceptive sensitization, it has been accepted as a model for learning and memory since many synaptic mechanisms that work in this model also apply to learning and memory behaviors in rodents.

Whereas nociceptive pain is typically transient, clinical pain or pathological pain is characterized by persistent pain hypersensitivity. Clinical pain is often associated with either peripheral tissue damage/inflammation (inflammatory pain) or lesions to the nervous system (neuropathic pain), and it is typically characterized by hyperalgesia and allodynia, as well as by spontaneous pain – pain experienced in the absence of any obvious peripheral stimulus. Animal models of inflammatory pain are usually produced by injection of irritative substances such as formalin, carrageenan, and complete Freund's adjuvant (CFA) into a hindpaw or joint of rats and mice, and this pain lasts from hours to days or weeks. In animal models, neuropathic pain is often induced by lesion to a peripheral nerve (e.g., partial lesion of the sciatic nerve and its branches in rats and mice), and this pain can last for many weeks to several months. In addition, animal models are also developed for cancer pain and postoperative pain (e.g., pain after incision), which share some features with inflammatory and neuropathic pain but also have distinct characteristics. Long-term exposure of opioids such as morphine, a traditional analgesic, may also produce clinical pain symptoms. Pain hypersensitivity is not only produced in the injured tissue or territory (innervated by the injured nerve) but also spread to the adjacent noninjured regions or the extraterritory. Current drugs have had only limited success in treating some clinical pain, such as neuropathic pain, in part due to our incomplete knowledge about how nociceptive sensitization is produced and maintained. It is generally believed that persistent pain or clinical pain is an expression of neural plasticity, which can occur in both the peripheral and the central nervous system.

Neural Plasticity in the Peripheral Nervous System: Peripheral Sensitization

Tissue injury such as that caused by surgical incision produces inflammation and heightens the pain experience. This arises from the production and release of inflammatory mediators such as prostaglandin E_2 (PGE_2), bradykinin, ATP, protons, and nerve growth factor (NGF) from nonneuronal cells (e.g., fibroblasts, mast cells, neutrophils, and platelets) and from primary sensory terminals. After nerve damage, proinflammatory cytokines such as tumor necrosis factor-α (TNF-α) and interleukin-1β (IL-1β) are released from enclosing Schwann cells and damaged axons. The soma and axons of primary sensory

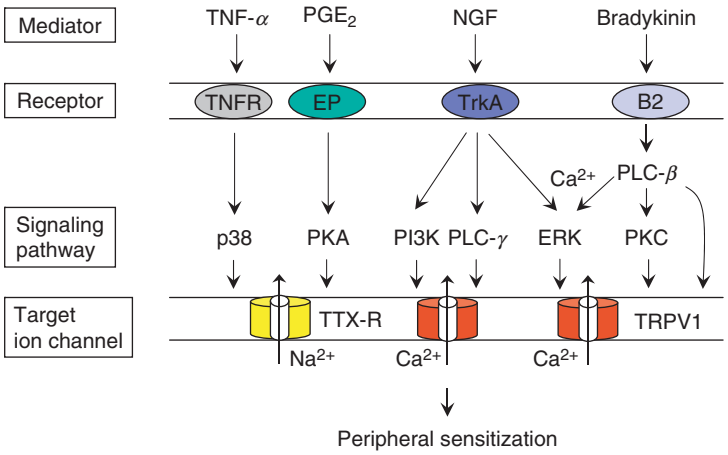

Figure 1 Induction of peripheral sensitization in nociceptor neurons. Peripheral tissue injury or nerve damage leads to the production of various inflammatory mediators, such as TNF-α, PGE_2, bradykinin, and NGF, which are released and stimulate the appropriate receptors on terminals, axons, or the cell body of primary sensory neurons. TNF-α is known to activate p38 MAPK in DRG neurons. Stimulation of G-protein-coupled receptors (GPCR) by PGE_2 activates adenylate cyclase through G_s and subsequently stimulates PKA. Other GPCR such as bradykinin B2 receptor activate phospholipase C-β (PLC-β) through G_q, thus increasing intracellular Ca^{2+} and leading to the activation of ERK and PKC. Stimulation of the tyrosine receptor kinase A (TrkA) receptor by NGF results in an activation of ERK, PI3K, and PLC-γ signaling pathways. Activation of PLC-β or PLC-γ results in the hydrolysis of plasma membrane lipids and activation of PKC, both proposed to enhance TRPV1 function. In particular, NGF can rapidly increase membrane expression of TRPV1 through PI3K. PI3K activates Src that binds to and phosphorylates TRPV1, leading to membrane insertion of TRPV1. As a consequence of these regulations, the threshold for the activation of TRPV1 is reduced so that lower temperature (e.g., body temperature 37 °C) can activate the receptor following tissue injury. In addition, activation of PKA, PKC, p38, and ERK results in potentiation of TTX-resistant sodium channel (TTX-R, including $Na_v1.8$ and $Na_v1.9$). Hyperactivity of TRPV1 and TTX-R causes increased sensitivity of primary sensory neurons at the peripheral terminals, axons, and cell bodies, termed 'peripheral sensitization.'

neurons contain receptors for these inflammatory mediators, and activation of the receptors leads to the activation of multiple intracellular signaling pathways, increasing the sensitivity and excitability of nociceptors (Figure 1). This sensitization in the peripheral nervous system (PNS) is called peripheral sensitization. Several ion channels are essential for the generation of peripheral sensitization. Posttranslational regulation of these channels can occur within minutes by phosphorylation, which is responsible for the induction of peripheral sensitization. However, transcriptional regulation of these channels takes hours to manifest, leading to persistent peripheral sensitization. In addition, regulation through transcription-independent translation or axonal transport can also change the sensitivity of nociceptors.

Peripheral sensitization produces increases in pain sensitivity restricted to the site of tissue injury. An example is heat pain after sunburn; a normally warm shower now feels burning hot.

Onset of Peripheral Sensitization

Various types of ion channels, such as TRPV1, TRPV2, TRPM8, P2X3, acid-sensitive channel ASIC, and tetrodo toxin (TTX)-resistant sodium channels (TTX_R, including $Na_v1.8$ and 1.9), are expressed in primary sensory neurons of the dorsal root ganglion (DRG) or trigeminal ganglion. The sensitivity of these channels can be regulated by inflammatory mediators. Current studies focus on two molecules – TRPV1, a transduction molecule that can convert extracellular stimuli into electrical activity on the membrane, and TTX_R, a conduction molecule that is important for the conduction of action potential to the spinal cord. TRPV1 is specifically expressed in C-fiber nociceptive neurons and is required for the generation of heat hyperalgesia. Inflammatory heat hypersensitivity following bradykinin, NGF, CFA, and carrageenan is almost abolished in TRPV1 null mice. TTX_R sodium channels are also expressed in nociceptors and crucial for the generation of hyperexcitability of sensory neurons. Knockdown of $Na_v1.8$ with antisense oligodeoxynucleotides results in decreased inflammatory pain and neuropathic pain. It is generally believed that hypersensitivity of TRPV1 and TTX_R underlies the induction of peripheral sensitization. Several protein kinases, such as phospholipase C, PKC, phosphatidyl inositide-3 kinase (PI3K) and extracellular signal-regulated kinase (ERK), are implicated in TRPV1 sensitization (Figure 1). In particular, the PI3K pathway may contribute to peripheral sensitization by inducing rapid trafficking and membrane insertion of TRPV1. Activation of PKA, PKC, and p38 mitogen-activated protein kinase

(MAPK) is also involved in regulating the sensitivity of TTX_R (Figure 1). A membrane translocation of PKC-ε appears to be important for its activation of TTX_R.

Persistence of Peripheral Sensitization

Tissue and nerve injuries induce gene transcription and protein synthesis in primary sensory neurons contributing to persistent pain hypersensitivity. Although NGF is required for the survival of sensory neurons during fetal development, it is not necessary for the survival of mature sensory neurons. Rather, it maintains the phenotype of nociceptive sensory neurons in the adult and is thus required for the continued expression of many genes. C-fiber nociceptors in the DRG are roughly divided into two subsets, NGF-responsive and glia-derived growth factor (GDNF)-responsive, despite the fact that there is a significant overlap between these two subsets of nociceptors. NGF is absolutely critical for gene activation in tyrosine receptor kinase A (TrkA) expressing C-fiber nociceptors. Many nociceptive genes, such as those encoding for substance P, calcitonin gene-related peptide (CGRP), brain-derived neurotrophic factor (BDNF), TRPV1, and $Na_v1.8$, are induced by NGF in DRG neurons. Therefore, an increase in NGF levels in DRG neurons after peripheral inflammation leads to an upregulation of these nociceptive genes. In contrast, a decrease in NGF levels in DRG neurons results in a downregulation of these genes in axon-injured neurons, which can further be rescued by NGF. It has been shown that GDNF family member artemin is also produced in inflamed tissue, which can induce TRPV1 expression in DRG neurons through the GFRα3 receptor. The GFRα3 receptor is coexpressed with TrkA, thus enhancing the function of NGF.

Target-derived NGF is retrogradely transported to the cell body in the DRG, where it activates p38 MAPK, leading to an upregulation of TRPV1 and other nociceptive genes (e.g., BDNF, substance P, and TRPA1, a receptor required for cold allodynia). Therefore, the active form of p38 (phosphorylated p38) is coexpressed with TRPV1 and inhibition of this kinase suppresses heat hyperalgesia (Figure 2). It remains to be investigated whether artemin also activates this kinase to induce TRPV1 expression. Apart from p38, ERK/MAPK is also activated by NGF in DRG neurons and involved in NGF induction of nociceptive genes. Activation of the transcription factor cAMP response element binding protein (CREB) could be downstream to MAPK activation. The CREB binding sites (CRE) are found in nociceptive genes that encode BDNF, substance P, and CGRP in sensory neurons.

In addition to transcriptional regulation, MAPK also mediates gene expression via translational regulation. Peripheral inflammation and NGF induce an increase in TRPV1 protein levels but not in TRPV1 mRNA levels in the DRG. This nontranscriptional regulation of TRPV1 is mediated by p38 in a NGF-dependent way. Finally, increased axonal transport of nociceptive proteins also contributes to persistent pain hypersensitivity. For

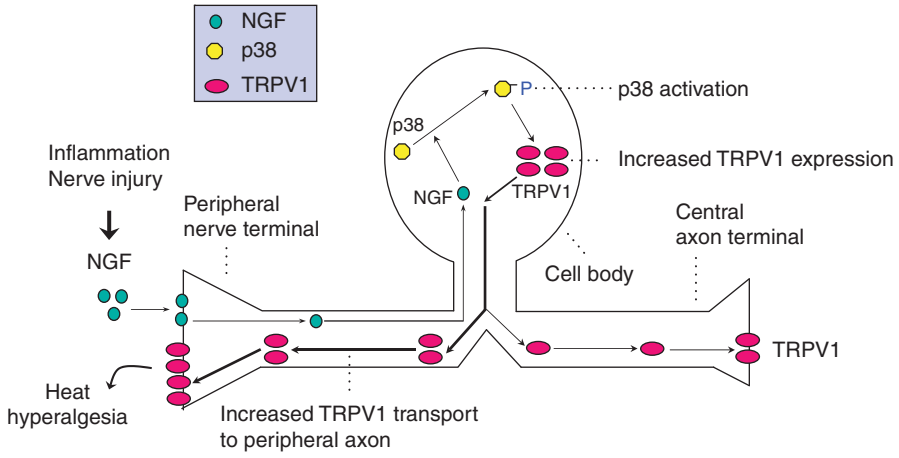

Figure 2 Regulation of TRPV1 expression and persistent heat hyperalgesia by p38 MAPK activation in primary sensory neurons. In normal conditions, TRPV1 is synthesized in the cell body of primary sensory neurons in the dorsal root ganglion (DRG) and transported to both peripheral terminals in the skin/muscle and central terminals in the spinal cord. After peripheral tissue inflammation (e.g., arthritis), NGF is produced in inflamed tissue. NGF is also produced by Schwann cells after peripheral nerve damage. Upon release, NGF is taken up by intact peripheral nerve terminals and axons and retrogradely transported to the cell body in the DRG. NGF in the cell body activates p38, which in turn increases TRPV1 translation. Furthermore, increased TRPV1 is anterogradely transported to peripheral terminals, producing persistent peripheral sensitization (heat hyperalgesia). Adapted from Ji RR, Samad TA, Jin SX, Schmoll R, and Woolf CJ (2002) MAPK activation by NGF in primary sensory neurons after inflammation increases TRPV1 levels and maintains heat hyperalgesia. *Neuron* 36: 57–68, with permission from Elsevier.

example, inflammation increases TRPV1 axonal transport from the DRG cell body to peripheral nerve terminals after inflammation (Figure 2).

Neural Plasticity in the CNS: Central Sensitization

Tissue and nerve injuries produce plastic changes in the pain pathway not only at the peripheral sites but also at the central sites. An important advance in pain research is to establish that activation of peripheral nociceptors by high-threshold and persistent noxious stimuli (heat, mechanical, and electrical) or by inflammatory mediators (chemical stimuli) also results in an activity- or use-dependent neuronal plasticity in the CNS. This plasticity modifies the performance of the nociceptive pathway by enhancing and prolonging the responses to subsequent peripheral stimuli. These changes in the spinal cord, as well as in the brain, are defined as central sensitization. Although central sensitization is often but not always induced by peripheral sensitization, it is responsible for the pain, after injury, by normally innocuous low-threshold afferent inputs (mechanical allodynia) and the spread of hypersensitivity to regions beyond injured tissue. The behavioral consequences of central sensitization can be readily detected in human psychophysical experiments. For example, intradermal injection of capsaicin, the pungent ingredient in chili peppers that activates the TRPV1 receptor, produces an intense but transient burning pain. This is followed by heightened sensitivity outside of the primary region of the capsaicin injection to high-threshold pinprick (secondary mechanical hyperalgesia) and to low-threshold brush (secondary mechanical allodynia), both of which are induced by central sensitization. Clinically, central sensitization contributes to pain hypersensitivity in the skin, muscle, joints, and viscera, such as in patients with fibromyalgia. Neuropathic pain after spinal cord injury is called 'central pain,' in which peripheral sensitization may not be robust and central sensitization can be self-maintained.

Current studies on central sensitization focus on dorsal horn neurons, the first-order relay neurons in the pain pathway. Small nociceptive fibers from primary sensory neurons mainly project to the superficial dorsal horn (laminae I–II), as well as lamina V, whereas large myelinated fibers terminate in the deep dorsal horn (laminae III and IV). Primary afferent inputs to dorsal horn neurons are excitatory, which is primarily contributed by glutamate. The dominant type of inhibitory synaptic transmission in the dorsal horn is mediated by γ-aminobutyric acid (GABA) and glycine receptors, which are ligand-gated Cl^- channels. In normal conditions, the inhibitory control on the pain system is very powerful, serving as a 'gate control.' Disturbance of the balance between excitatory and inhibitory synaptic inputs will lead to abnormal pain sensitivity. For example, there is a reduction of inhibitory synaptic transmission but an enhancement of excitatory synaptic transmission in the dorsal horn following peripheral nerve injury.

Pain and memory are often associated in daily life. Persistence of a heightened responsiveness of the nervous system after noxious stimuli or injuries has clear parallels with memory. Studies on neural plasticity in the hippocampus reveal that long-term potentiation (LTP) – a brief tetanic electrical stimulation that induces long-term enhancement of synaptic transmission in hippocampal neurons – underlies the generation of memory. LTP is also induced in dorsal horn neurons. It is not only induced by nerve stimulation at high frequencies (100 Hz) but also may be induced by nerve injury or even by low-level activity in nociceptive nerve fibers after inflammation. However, LTP is more difficult to induce in dorsal horn neurons than in hippocampal neurons. Therefore, central sensitization is a broader phenomenon, including LTP (homo- or heterosynaptic), wind up (homosynaptic, a transient synaptic plasticity during a noxious stimulus), and activity-dependent classic central sensitization (heterosynaptic, a synaptic plasticity induced by low-threshold Aβ stimulation, which outlasts initial C-fiber stimulation) and other forms of hypersensitivity of CNS neurons. Whereas long-term memory and the late phase of hippocampal LTP require gene transcription, short-term memory and early phase LTP require only posttranslational modifications. This same dichotomy appears to apply to central sensitization and pain hypersensitivity: the induction of central sensitization and pain hypersensitivity appears to require only posttranslational processes, whereas their maintenance seems to depend on changes in gene transcription. Although neural circuits in the spinal dorsal horn and hippocampus are quite different, glutamate is the predominant neurotransmitter in both circuits and activation of NMDA receptors is essential for the plasticity of both dorsal horn and hippocampal neurons. The activity of NMDA and AMPA receptors can be regulated by several protein kinases through posttranslational regulation. Moreover, the ERK-CREB pathway might be critical for the transcriptional regulation of genes in both the dorsal horn and the hippocampus that maintain central sensitization and LTP.

Onset of Central Sensitization in the Spinal Cord

Mild noxious stimulation generates fast excitatory postsynaptic potentials (EPSPs) lasting milliseconds.

This fast synaptic transmission is mediated by ionotropic glutamate receptors, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), and kainate receptors. Another type of ionotropic glutamate receptor, the *N*-methyl-D-aspartate (NMDA) receptor, contributes little to the EPSPs or action potentials under normal conditions due to Mg^{2+} blockade of the channel. However, this receptor contributes importantly to the onset of different forms of central sensitization after removal of Mg^{2+} blockade via depolarization. In addition to glutamatergic synaptic transmission, intense noxious stimulation also releases neuropeptides, such as substance P, producing slow postsynaptic responses via G-protein-coupled natural killer (NK)-1 receptors. Also, C-fiber stimulation can release the neurotrophin BDNF. Like substance P, BDNF is located in large, dense-cored vesicles. Inflammation not only increases the synthesis of substance P and BDNF in primary sensory neurons but also increases the expression of their receptors (NK-1 and TrkB) in postsynaptic spinal neurons. Both increases would enhance the responses of the receptors because the spontaneous activity that is generated in axons and cell bodies of primary sensory neurons after injuries will cause more release of neurotransmitters.

Activation of protein kinases and intracellular signaling cascades is required for the development of central sensitization. Protein kinases can phosphorylate NMDA and AMPA receptors via posttranslational regulation, enhancing synaptic transmission and producing central sensitization (**Figure 3a**). For example, a PKC-mediated phosphorylation of the NMDA receptor in dorsal horn neurons removes its voltage-dependent Mg^{2+} block, enabling glutamate to generate a greater inward current through the NMDA ion channel at resting membrane potentials. Tyrosine kinase Src enhances NMDA current via phosphorylation of the NR2A/B subunit of the NMDA receptor, increasing channel open time and kinetics. Noxious simulation, inflammation, as well as stimulation of TrkB receptor induce the phosphorylation of NMDA receptor subunits in the dorsal horn neurons.

NMDA receptor-mediated activation of CaMKII is critical for the induction of hippocampal LTP. CaMKII not only phosphorylates GluR1-containing AMPA receptor subunits on Ser831, enhancing the single channel conductance of synaptic AMPA receptors, but also induces rapid trafficking and incorporation of GluR1 into hippocampal synapses as well as docking of NMDA receptors to postsynaptic sites in dorsal horn neurons. Painful stimulation was also shown to induce CaMKII phosphorylation and membrane insertion of GluR1 receptor subunits in spinal cord neurons.

In addition to enhanced excitatory synaptic transmission, suppression of inhibitory synaptic transmission also plays an important role in the generation of central sensitization. PGE_2, an inflammatory mediator strongly implicated in peripheral sensitization, also contributes to central sensitization by suppressing glycine currents in dorsal horn neurons via activation of PKA (**Figure 3(a)**). BDNF has been implicated in disinhibition via modulation of chloride influx and alternation in anion reversal potentials.

In particular, ERK is specifically activated in superficial dorsal horn neurons by nociceptive inputs (**Figure 3(b)**). Activation of NMDA and TrkB receptors, as well as metabotropic glutamate receptors, results in ERK activation, which is required for behavioral manifestation of central sensitization. ERK can suppress the activity of $K_v4.2$ potassium channels, increasing the excitability of dorsal horn neurons. Also, ERK may phosphorylate NMDA receptor in dorsal horn neurons and mediate membrane insertion of AMPA receptors during synaptic plasticity.

Persistence of Central Sensitization in the Spinal Cord

Pain and memory also share mechanisms for long-term changes in synaptic plasticity. In particular, the transcription factor CREB plays an essential role in persistent neuronal plasticity in both hippocampal and dorsal horn neurons and is required for both long-term memory and persistent pain (e.g., neuropathic pain). CREB is known to regulate the expression of numerous genes that are important for plasticity, cell survival, and synaptic formation. Noxious stimulation induces a rapid phosphorylation of CREB in dorsal horn neurons. Persistent CREB phosphorylation was also shown in the spinal cord after inflammation and nerve injury. CREB may maintain central sensitization by inducing the transcription of immediate early genes (e.g., c-*fos*, *Zif268*, and *Cox-2*) and later response genes (e.g., *NK-1*, *TrkB*, and *prodynorphin*). Importantly, these genes are induced in the dorsal horn following intense noxious stimulation, or inflammation and nerve injury, and they all contain CREB binding sites (CRE) in their promoter regions. Although multiple signaling cascades are implicated in CREB activation, the ERK pathway appears to be critical for this activation. Phosphorylated ERK (pERK) is colocalized with pCREB in dorsal horn neurons following noxious stimuli. Upon activation of ERK, pERK is translocated to the nucleus, activating the CREB kinase RSK2 (**Figure 3(c)**). Since multiple signaling cascades, such as PKA, PKC, PI3K, and Src, can cause ERK activation, they should also converge on CREB phosphorylation. Central sensitization is maintained not

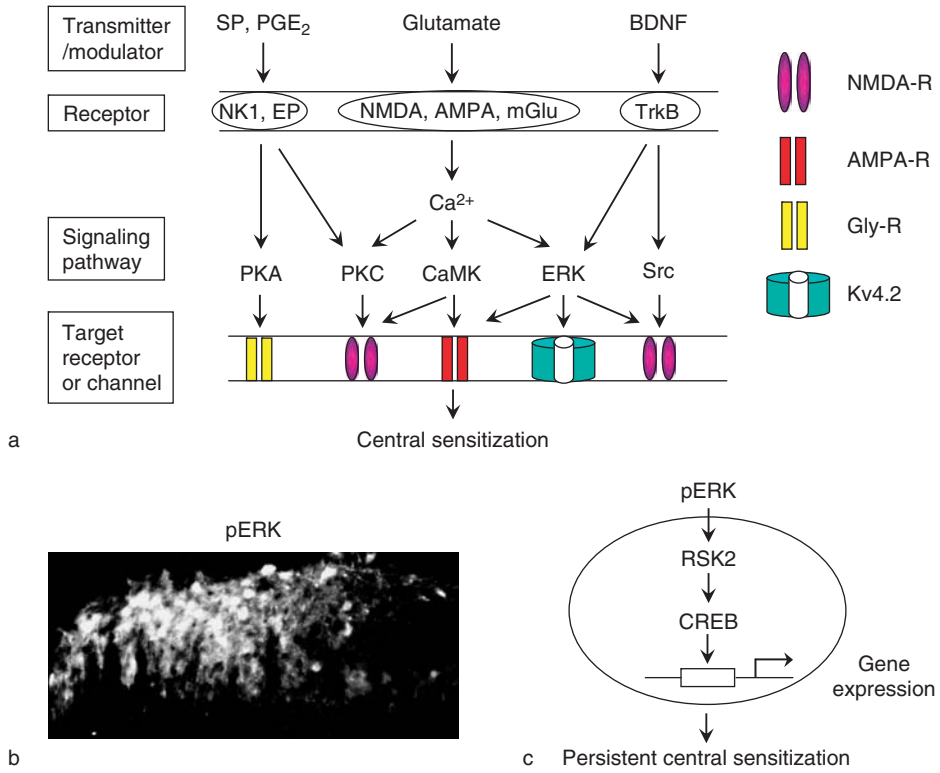

Figure 3 Mechanisms of central sensitization induction and maintenance in spinal cord dorsal horn neurons. (a) Induction of central sensitization. Following tissue or nerve injury, there is increased release of the neurotransmitters glutamate, substance P, and BDNF from the central terminals of primary afferents in the spinal cord. Activation of postsynaptic NMDA receptors and Ca^{2+}-permeable AMPA receptors in the spinal cord leads to Ca^{2+} influx. Stimulation of mGluRs (e.g., group I) and NK-1 receptors activates PLC and stimulates Ca^{2+} release from intracellular stores. The increase in intracellular Ca^{2+} activates CaMK, PKC, and ERK. Activation of TrkB leads to the activation of the Ras-Raf-MEK-ERK pathway as well as Src, which is also activated by PKC. Activation of these protein kinases increases the function of NMDA and AMPA receptors. In addition, ERK activation suppresses the activity of the $K_V4.2$ potassium channels, increasing the excitability of dorsal horn neurons. PGE_2 is also released in the spinal cord under pain conditions; activation of its EP receptor reduces the function of glycine receptors via PKA. Collectively, these posttranslational steps result in enhanced excitatory synaptic transmission and reduced inhibitory synaptic transmission, contributing to an overall increase in dorsal horn neuron sensitivity that generates central sensitization. (b) Activation of ERK in superficial dorsal horn neurons following a noxious stimulation. Injection of a C-fiber activator capsaicin into a hindpaw of rats produces intense burning pain. It also rapidly (within minutes) phosphorylates ERK, as shown by pERK immunofluorescence, in the medial superficial spinal cord, where C-fiber primary afferents from a hindpaw terminate. However, pERK-labeled neurons are not found in the nonstimulated spinal cord (e.g., contralateral side). (c) Maintenance of central sensitization. Activation of ERK in dorsal horn neurons after injury leads to a nuclear translocation of pERK, activating the transcription factor CREB via RSK2. CREB activation results in increased transcription of many dorsal horn genes such as *NK-1*, *Cox-2*, c-*fos*, *TrkB*, and *prodynorphin*. Persistent activation of these genes maintains heightened excitability and synaptic responsiveness and thus supports central sensitization and persistent pain. Scale bar = 50 μm (b).

only by an upregulation of gene expression but also by a downregulation of gene expression. For example, a reduction of spinal glutamate transporters (e.g., EAAC1 and GLAST1) following nerve injury will decrease the uptake of glutamate, leading to sustained glutamate activation in synapses and enhanced excitatory synaptic transmission.

In addition to persistent change in excitatory transmission, peripheral nerve injury also produces long-term changes in inhibitory transmission. For example, nerve lesion induces a rapid downregulation of the potassium chloride cotransporter KCC2 in dorsal horn neurons, leading to an accumulation of intracellular chloride that can convert inhibitory inputs of GABA and glycinergic interneurons into excitatory ones. Furthermore, there might be a delayed but permanent loss of inhibitory neurons in the dorsal horn after nerve injury.

Glial Regulation of Central Sensitization in the Spinal Cord

Studies on pathological pain primarily focus on the responses of the neurons and neural circuits. There is accumulating evidence indicating a role of spinal glial cells in persistent pain. In particular, microglia in the spinal cord are strongly implicated in the genesis of neuropathic pain. Microglial markers such as CD11b are rapidly induced in spinal cells after nerve injury.

Nerve injury also produces robust morphological changes of microglia, from resting status (with ramified processes) in the normal conditions to hypertrophy status (with thick and short process) after lesion. Upon activation, microglia become a major source of inflammatory mediators, producing proinflammatory cytokines (TNF-α, IL-1β, and IL-6), PGE_2 (synthesized by the enzymes COX-1 and COX-2), and NO (synthesized by the enzyme iNOS). Nerve lesion elevates the levels of these cytokines and enzymes in the spinal cord. BDNF is also produced in cultured microglia, although it is barely detected in spinal microglia *in vivo* under normal or injury conditions. Upon release, all these substances are neuroactive. Although PGE_2 and BDNF have been implicated in central sensitization, little is known about the central effects of proinflammatory cytokines. IL-1β appears to enhance excitatory postsynaptic transmission in dorsal horn neurons, in addition to its transcriptional regulation of COX-2. The involvement of microglia and inflammatory mediators in central sensitization is also supported by behavioral studies: spinal injection of glial toxins (e.g., fluorocitrate), microglial inhibitor (minocycline), and inhibitors of proinflammatory cytokine pathways and COX has been shown to reduce neuropathic pain. In addition, several receptors, such as ATP receptor P2X4, chemokine receptors CCR2 and CX3CR1, and probably Toll-like receptor (TLR-4), are induced in spinal microglia after nerve injury; blocking these receptors results in reduced neuropathic pain.

Signal transduction in spinal microglia begins to be revealed. Nerve injury induces a very robust activation of p38 MAPK exclusively in spinal microglia. Furthermore, intrathecal injection of p38 inhibitors attenuates nerve injury-induced neuropathic pain. p38 activation in microglia is known to induce the synthesis and release of multiple inflammatory mediators in the spinal cord, enhancing and prolonging spinal sensitization (**Figure 4**). It is generally believed that spontaneous electrical activity produced in the axons and cell bodies after nerve injury is essential for the genesis of neuropathic pain. This activity also appears to be important for microglial activation of p38. Therefore, nerve conduction blockade proximal to nerve injury prevents spinal p38 activation. TNF-α is also implicated in p38 activation. In particular, the chemokine fractalkine (CX3CL1) plays an important

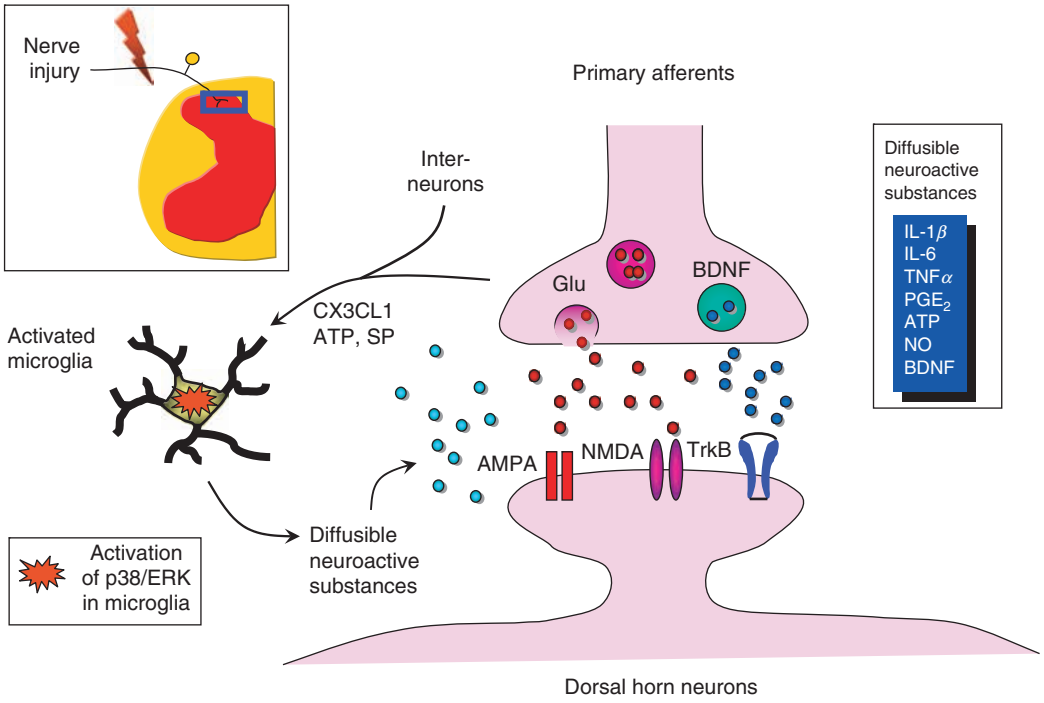

Figure 4 Glial regulation of spinal sensitization following peripheral nerve injury. After peripheral nerve injury, hyperactive primary sensory neurons or spinal neurons can release neuromodulators such as substance P, ATP, and fractalkine (CX3CL1), leading to the activation of the MAPK (p38 or/and ERK) signaling pathway in spinal microglia. The activated microglia synthesize and release multiple inflammatory mediators, such as IL-1β, TNF-α, IL-6, nitric oxide (NO), and PGE_2, as well as the neurotrophin BDNF. These mediators serve as diffusible neuroactive substances, acting on postsynaptic dorsal horn neurons, enhancing the function of NMDA and AMPA receptors or reducing the function of GABA and glycine receptors. These mediators can also act on presynaptic sites, leading to increased release of neurotransmitters (e.g., glutamate and BDNF). Moreover, these mediators will act on surrounding glia (microglia or astrocytes), causing further activation of glia. Taken together, a positive feedback loop between neurons and glia in the spinal cord enhances and maintains central sensitization and neuropathic pain.

role in mediating neural/glial interaction for several reasons. First, fractalkine is produced in neurons (including both primary sensory and spinal neurons). Second, CX3CR1 is the sole receptor for fractalkine and this receptor is only expressed in spinal microglia. Finally, nerve injury induces a marked upregulation of CX3CR1 in spinal microglia. Activation of CX3CR1 also results in p38 activation, and CX3CR1/p38 cascade in spinal microglia is essential for the development of neuropathic pain. In addition to p38, ERK is also activated in spinal microglia by nerve lesion, which is essential for neuropathic pain sensitization.

Apart from a rapid activation of spinal microglia, nerve injury induces a delayed but more persistent activation of astrocytes. Interestingly, c-Jun N-terminal kinase (JNK)/MAPK is persistently activated in spinal astrocytes and contributes to late maintenance of neuropathic pain. Since astrocytes have more close contact with synapses, they should regulate spinal neuron sensitivity via different mechanisms than microglia.

Central Sensitization in the Brain

Activity-dependent central sensitization is extremely robust and has been reported in rodent, cat, and primate dorsal horn neurons, including spinothalamic neurons. A central sensitization-like phenomenon can also be generated in neurons in other parts of the pain system, such as the rostroventral medulla (RVM), anterior cingulate cortex (ACC), and amygdala, following tissue injury. Several symptomatic features of headache reflect central sensitization produced in trigeminal nuclei within the brain stem. For example, inflammation induces phosphorylation of AMPA receptor (GluR1) in the RVM, which may regulate descending facilitation. Calcium-dependent adenylate cyclase AC1 and AC8 are highly expressed in the ACC, and inflammatory pain is attenuated in mice lacking AC1 and AC8. The action of AC1 and AC8 appears to be mediated by the cAMP-PKA-CREB pathway. Neuronal activity in the ACC could influence spinal nociceptive transmission by activating the descending facilitatory system.

Pain consists of different components. In addition to a sensory-discriminative component that describes the quality, intensity, and spatiotemporal characteristics of the sensation, pain experience also includes an emotional-affective component, referring to the unpleasantness or aversion of sensation. Emotional distress could be the most disruptive and undesirable feature of painful experiences. ACC also plays an important role in pain affect. Noxious stimulation (e.g., formalin)-induced conditioned place avoidance (noxious conditioning) has been established as an animal model to examine affective pain. Lesions of the ACC produce severe deficits in avoidance conditioning. Furthermore, the PKA/ERK pathway in the ACC is also involved in affective pain. Therefore, neural plasticity also plays an important role in the generation of affective pain.

Summary

Under normal conditions, pain is acute and protective, manifested as responsiveness to noxious stimuli. However, after tissue or nerve injury, pain often becomes pathological and chronic, with increased responsiveness to noxious stimuli and even becoming responsive to innocuous stimuli. Pain hypersensitivity following intense nociceptive input or injuries results from neural plasticity that occurs both in the peripheral nervous system (e.g., primary sensory neurons), termed peripheral sensitization, and in the central nervous system (e.g., dorsal horn neurons), termed central sensitization. Activation of several intracellular signal pathways in nociceptive neurons leads to both induction and maintenance of both types of sensitization. Furthermore, activation of spinal glial cells following injuries can enhance and prolong central sensitization by producing inflammatory mediators centrally. Current treatment for clinical pain is not very effective, in part due to a focus on nociception. Development of drugs that suppress signal transduction pathways in neurons and glia may offer new therapies for chronic pain by normalizing neural plastic changes in the pain pathway.

See also: Inflammation in Neurodegenerative Disease and Injury; Neuropathic Pain; Nociceptor Responses; Pain Pathways: Descending Modulation; Pain and Genes; Pain Management; Pain: Central; Pain: Neuroimaging.

Further Reading

Caterina MJ, Leffler A, Malmberg AB, et al. (2000) Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science* 288: 306–313.

Chen L and Huang LY (1992) Protein kinase C reduces Mg^{2+} block of NMDA-receptor channels as a mechanism of modulation. *Nature* 356: 521–523.

Coderre TJ, Katz J, Vaccarino AL, and Melzack R (1993) Contribution of central neuroplasticity to pathological pain: Review of clinical and experimental evidence. *Pain* 52: 259–285.

Dougherty PM and Willis WD (1992) Enhanced responses of spinothalamic tract neurons to excitatory amino acids accompany capsaicin-induced sensitization in the monkey. *Journal of Neuroscience* 12: 883–894.

Dubner R and Ruda MA (1992) Activity-dependent neuronal plasticity following tissue injury and inflammation. *Trends in Neuroscience* 15: 96–103.

Gold MS, Levine JD, and Correa AM (1998) Modulation of TTX-R INa by PKC and PKA and their role in PGE2-induced

sensitization of rat sensory neurons *in vitro*. *Journal of Neuroscience* 18: 10345–10355.

Hokfelt T, Zhang X, and Wiesenfeld-Hallin Z (1994) Messenger plasticity in primary sensory neurons following axotomy and its functional implications. *Trends in Neuroscience* 17: 22–30.

Hunt SP and Mantyh PW (2001) The molecular dynamics of pain control. *Nature Reviews Neuroscience* 2: 83–91.

Ikeda H, Stark J, Fischer H, et al. (2006) Synaptic amplifier of inflammatory pain in the spinal dorsal horn. *Science* 312: 1659–1662.

Ji RR, Kohno T, Moore KA, and Woolf CJ (2003) Central sensitization and LTP: Do pain and memory share similar mechanisms? *Trends in Neuroscience* 26: 696–705.

Ji RR, Samad TA, Jin SX, Schmoll R, and Woolf CJ (2002) MAPK activation by NGF in primary sensory neurons after inflammation increases TRPV1 levels and maintains heat hyperalgesia. *Neuron* 36: 57–68.

Ji RR and Strichartz G (2004) Cell signaling and the genesis of neuropathic pain. *Science STKE* 252: reE14 [online].

Julius D and Basbaum AI (2001) Molecular mechanisms of nociception. *Nature* 413: 203–210.

Pezet S and McMahon SB (2006) Neurotrophins: Mediators and modulators of pain. *Annual Review of Neuroscience* 29: 507–538.

Porreca F, Ossipov MH, and Gebhart GF (2002) Chronic pain and medullary descending facilitation. *Trends in Neuroscience* 25: 319–325.

Treede RD, Meyer RA, Raja SN, and Campbell JN (1992) Peripheral and central mechanisms of cutaneous hyperalgesia. *Progress in Neurobiology* 38: 397–421.

Tsuda M, Inoue K, and Salter MW (2005) Neuropathic pain and spinal microglia: A big problem from molecules in "small" glia. *Trends in Neuroscience* 28: 101–107.

Watkins LR, Milligan ED, and Maier SF (2001) Glial activation: A driving force for pathological pain. *Trends in Neuroscience* 24: 450–455.

Woolf CJ (1983) Evidence for a central component of post-injury pain hypersensitivity. *Nature* 306: 686–688.

Woolf CJ and Salter MW (2000) Neuronal plasticity: Increasing the gain in pain. *Science* 288: 1765–1769.

Pain and the Sympathetic Nervous System

W Jänig, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The autonomic nervous system is by definition an efferent system that is involved in the regulation of a variety of target cells, such as vascular and nonvascular smooth muscle cells, cardiac muscle cells, and secretory epithelia. Based primarily on anatomical criteria, we distinguish parasympathetic, sympathetic, and enteric nervous systems. Primary afferent neurons with small-diameter myelinated (Aδ) or unmyelinated (C) nerve fibers that also innervate the autonomic target cells (e.g., blood vessels in somatic tissues or viscera) and have their cell bodies in the dorsal root ganglia, trigeminal ganglia, or ganglia associated with vagal afferent neurons (inferior, or nodose, and superior, or jugular, ganglia of the vagus; petrosal ganglion) are not called autonomic (sympathetic or parasympathetic) afferent neurons. We have no anatomical, histochemical, functional, or other criteria (e.g., distinct central synaptic connectivity of these afferent neurons) to classify visceral afferent neurons as autonomic. Afferent neurons of the enteric nervous system which have their cell bodies in enteric ganglia are called intrinsic primary afferent neurons. This conceptual distinction between efferent sympathetic or parasympathetic neurons and afferent neurons that innervate autonomic target cells does not collide with the notion that these afferent neurons are functionally closely associated with autonomic regulation in the sense that they monitor the inner state of the body.

Any noxious (tissue-damaging or potentially tissue-damaging) event within the body or impinging on the body affects the sympathetic nervous system and neuroendocrine systems and leads to integrative responses that further healing and recuperation. These responses are orchestrated by the brain, the level of integration occurring in spinal cord, brain stem, and forebrain (hypothalamus and cerebral hemispheres) (Figure 1). The integrated responses during pain and nociception involving the sympathetic nervous system have to be distinguished from the role of the sympathetic nervous system in the generation or enhancement of pain. This aspect appears to be, at first sight, rather unconventional, since it assigns to an efferent system, which is normally involved in regulation of autonomic target cells, functions that may lead to activation or sensitization of the nociceptive system and therefore to pain. This article is restricted to this point, arguing that the sympathetic nervous system may be involved in the generation of pain.

Conceptual Background

Table 1 lists the different functions of the sympathetic nervous system related to the generation of pain. Some of these functions may be relevant under biological conditions and others only under pathobiological conditions. They involve primarily peripheral mechanisms. However, the functional relevance of these peripheral mechanisms will only be revealed when the central mechanisms are taken into account (Figure 2).

Whether explicitly formulated in the literature or not, the centerpiece of the models behind all descriptions of the putative role of the sympathetic nervous system in the generation of pain is a positive feedback pathway from the spinal cord to the periphery of the body via the sympathoneural or sympathoadrenal system. Figure 2 illustrates in which context this efferent loop has to be seen. Peripheral inflammation leads to nociceptor sensitization, followed by central sensitization of second-order neurons (e.g., in the spinal or trigeminal dorsal horn), with an amplification of the transmission of nociceptive impulses. Peripheral nerve lesion is followed by generation of ectopic impulses in nociceptive and nonnociceptive primary afferent neurons leading to central hyperexcitability. Both chronic pain states are accompanied, in the periphery and in the central nervous system, by plastic changes of the nociceptive system (and other neural systems). These plastic changes are expressed in biochemical, morphological, and functional changes of the neurons and neuron circuits. Depending on the time after trauma, changes associated with peripheral inflammation are probably reversible, and those being associated with nerve lesions not. Both categories of pain states (i.e., inflammatory pain and neuropathic pain) are expressed in spontaneous and evoked pain as well as possibly other somatosensory abnormalities and autonomic, somatomotor, and possibly endocrine responses. An important but still not well-defined component in this concerted action of the brain is the role of the sympathetic nervous system in initiation and maintenance of the pain states via the efferent feedback to the peripheral tissues (red in Figure 2). This function of the sympathetic nervous system may be relevant under pathophysiological conditions but less so under physiological conditions. The idea about the involvement of the sympathetic

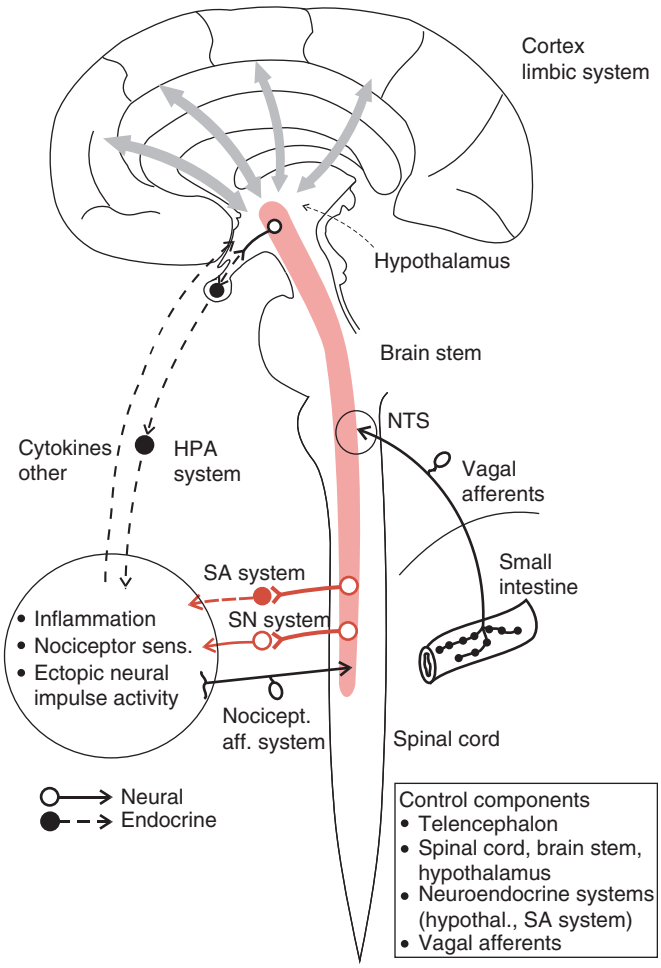

Figure 1 Spinal cord, brain stem, and hypothalamus contain neuronal circuits (red shaded area) that modulate nociceptor sensitivity, inflammation, and ectopic activity in lesioned afferent neurons in the periphery of the body via the sympathoadrenal (SA) system, the sympathoneural (SN) system, and the hypothalamic–pituitary–adrenal (HPA) system. Feedback information from the peripheral inflammatory process occurs via nociceptive primary afferent neurons and cytokines. Abdominal vagal afferents signal events from the inner defense line of the body (gut-associated lymphoid tissue) to the lower brain stem (NTS, nucleus tractus solitarii). The telencepalon controls inflammation and sensitivity of nociceptors via the circuits in the spinal cord, brain stem, and hypothalamus (shaded arrows).

nervous system in the generation of pain is based on various clinical observations as well as on animal experimentation.

In the following sections, the focus is on the different modes of sympathetic–afferent coupling in the periphery of the body (as outlined in the top half of Table 1); these are believed to be the basis of sympathetically maintained pain. The roles of the sympathetic nervous system in inflammation and hyperalgesia, briefly outlined in the lower half of Table 1, will not be further discussed.

Sympathetically Maintained Pain: Clinical Background

Pain that is dependent on activity in sympathetic neurons is called 'sympathetically maintained pain' (SMP). SMP includes generically spontaneous pain and pain evoked by mechanical or thermal stimuli. It may be present in the complex regional pain syndrome (CRPS) type I and type II and in other pain syndromes. The concept of involvement of the (efferent) sympathetic nervous system in the generation of pain is based on various clinical observations that have been documented in the literature for decades. Representative of these multiple observations on patients with pain in which the sympathetic nervous system is causally involved are the following groups of experimental and clinical findings involving patients who had SMP:

- Blockade of impulse transmission in sympathetic paravertebral ganglia may relieve spontaneous pain as well as mechanical and cold allodynia in patients with CRPS.

Table 1 Sympathetic nervous system and the generation of pain

Sympathetic–afferent coupling in the periphery[a]
• Coupling after nerve lesion (noradrenaline, α-adrenoceptors)
• Coupling via the micromilieu of the nociceptor and the vascular bed
• Sensitization of nociceptors mediated by sympathetic terminals, independent of excitation and release of noradrenaline
• Sensitization of nociceptors initiated by cytokines or nerve growth factor and mediated by sympathetic terminals
• Sympathoadrenal system and nociceptor sensitization
Role of the sympathetic nervous system in inflammation and hyperalgesia[b]
• Complex regional pain syndromes and sympathetic nervous system
• Immune system and sympathetic nervous system
• Rheumatic diseases and sympathetic nervous system
• Systemic chronic pain and sympathetic nervous system (e.g., fibromyalgia)

[a]Based on data from Jänig W and Baron R (2001) The role of the sympathetic nervous system in neuropathic pain: Clinical observations and animal models. In: Hansson PT, Fields HL, Hill RG, et al. (eds.) *Neuropathic Pain: Pathophysiology and Treatment*, pp. 125–149. Seattle: IASP Press; Jänig W and Häbler H-J (2000) Sympathetic nervous system: Contribution to chronic pain. *Progress in Brain Research* 129: 453–470; and Jänig W and Levine JD (2006) Autonomic-endocrine-immune responses in acute and chronic pain. In: McMahon SB and Koltzenburg M (eds.) *Wall and Melzack's Textbook of Pain*, 5th edn., pp. 205–218. Edinburgh: Churchill Livingstone. [b]Based on data from Jänig W and Baron R (2003) Complex regional pain syndrome: Mystery explained? *Lancet Neurology* 2: 687–697; Jänig W and Häbler H-J (2000) Sympathetic nervous system: Contribution to chronic pain. *Progress in Brain Research* 129: 453–470; Jänig W, Khasar SG, Levine JD, et al. (2000) The role of vagal visceral afferents in the control of nociception. *Progress in Brain Research* 122: 273–287; Jänig W and Levine JD (2006) Autonomic-endocrine-immune responses in acute and chronic pain. In: McMahon SB and Koltzenburg M (eds.) *Wall and Melzack's Textbook of Pain*, 5th edn., pp. 205–218. Edinburgh: Churchill Livingstone; Straub RH, Baerwald CG, Wahle M, et al. (2005) Autonomic dysfunction in rheumatic diseases. *Rheumatic Disease Clinics of North America* 31: 61–75; and Vierck CJ Jr. (2006) Mechanisms underlying development of spatially distributed chronic pain (fibromyalgia). *Pain* 124: 242–263.

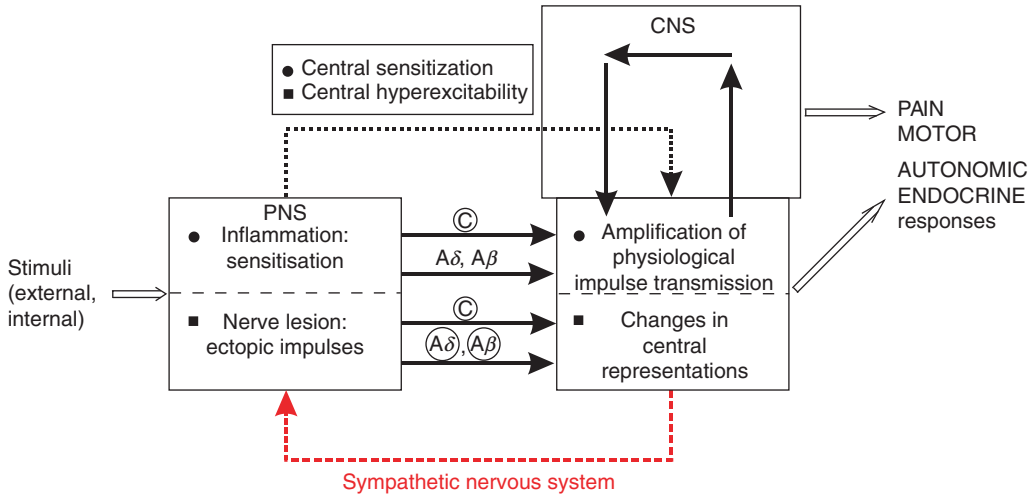

Figure 2 Concept of generation of peripheral and central hyperexcitability during inflammatory pain or neuropathic pain involving the sympathetic nervous system (CNS, central nervous system; PNS, peripheral nervous system). The pain is always associated with motor, autonomic, and endocrine responses. The upper dashed arrow indicates that the central changes are generated (and possibly maintained) by persistent activation of nociceptors with unmyelinated (C) fibers (e.g., during chronic inflammation), called here 'central sensitization,' or after trauma with nerve lesion by ectopic activity in myelinated (Aδ, Aβ) and C-afferents and other changes in the lesioned afferent neurons, called here 'central hyperexcitability.' The lower dashed arrow, in red, indicates the efferent feedback via the sympathetic nervous system (including the sympathoadrenal system). The transmission of nociceptive impulses is under multiple control of the brain (clear arrows, upper right). Primary afferent nociceptive neurons (in particular those with C-fibers) are sensitized during inflammation. After nerve lesion, all lesioned primary afferent neurons (unmyelinated as well as myelinated ones) undergo biochemical, physiological, and morphological changes, which become irreversible with time. These peripheral changes entail changes of the central representations (of the somatosensory system), which become irreversible if no regeneration of primary afferent neurons to their target tissue occurs. The central changes, induced by persistent activity in afferent nociceptive neurons or after nerve lesion, are also reflected in the sympathetic efferent feedback system that may establish positive feedback loops to the primary afferent neurons.

- Long after limb amputation, injection of adrenaline or physiological doses of noradrenaline around a stump neuroma may be intensely painful, compared to saline injection.
- Intraoperative electrical stimulation of the sympathetic chain may increase pain in patients with CRPS type II but not in patients with hyperhydrosis.

- In patients with CRPS or posttraumatic neuralgias, intracutaneous application of noradrenaline in a symptomatic area rekindles spontaneous pain and dynamic mechanical and cold allodynia that has been relieved by sympathetic blockade. This supports the idea that human nociceptors, after partial nerve lesion, develop noradrenergic sensitivity which can be maintained over several years.
- Some patients with postherpetic neuralgia experience increased spontaneous pain and mechanical hyperalgesia/allodynia after intracutaneous injection of adrenaline or phenylephrine (an α_1-adrenoceptor agonist).
- Patients with CRPS type I and SMP experience more spontaneous pain and mechanical allodynia when sympathetic cutaneous vasoconstrictor neurons are activated by cold stress.

These findings on patients with SMP are interpreted as follows (Figure 2): primary afferent nociceptive neurons are excited or possibly sensitized by noradrenaline released by sympathetic postganglionic fibers; these nociceptors may have expressed α-adrenoceptors, or their increased excitability may be generated indirectly – for example, by decrease of blood flow and subsequent ischemia or by other components modulated by sympathetic fibers. However, the mechanisms underlying SMP in the patients with nerve lesion could be rather different. In these patients, following nerve lesion (whether induced mechanically, metabolically, or by a virus), the lesioned primary afferent neurons and corresponding central representations (e.g., in the dorsal horn, thalamus, and cortex) may undergo dramatic biochemical, morphological, and physiological changes. Sympathetic coupling to afferent neurons with large-diameter axons may now be important, since activity in these neurons could generate central hyperexcitability and feed into central circuits that have changed as a consequence of the peripheral nerve lesion.

SMP following Nerve Lesion Simulated in Behavioral Animal Models

In animals (rats), signs of pain are measured by quantifiable behavioral components; examples of such behaviors include paw withdrawal latency in reaction to thermal noxious stimulation (heat, cold), frequency and threshold of paw withdrawal in reaction to mechanical stimulation, degree of self-mutilation (autotomy), or spontaneous lifting and licking of paws. These reactions are interpreted as equivalent to mechanical or thermal hyperalgesia or allodynia or spontaneous pain in humans. Behavioral models allow to test whether the sympathetic nervous system contributes to these hyperalgesic/allodynic behaviors or to spontaneous pain behavior. Animals are given clinically relevant nerve lesions, and behavioral experiments are performed before and after interventions aimed at targeting the sympathetic supply of the affected extremity. The strengths of these models include the range of possible experimental manipulations performed at the sympathetic nervous system and the ability to use different strains of animals or transgenic animals. Quantitative data obtained in these behavioral experiments lead to the formulation of hypotheses that can be tested in reduced experiments *in vivo* and *in vitro* using various methodical approaches.

Several rat models of neuropathic pain involving controlled nerve lesions have been used to determine the contribution of the sympathetic nervous system to neuropathic pain (Table 2). The interventions included surgical or chemical sympathectomy; systemic or local application of adrenoceptor blockers, including phentolamine (α_1, α_2), prazosin (α_1), or yohimbin (α_2), or of adrenoceptor agonists, including phenylephrine (α_1) or clonidine (α_2); and intraperitoneal application of guanethidine, which is taken up by the noradrenergic terminals and depletes noradrenaline.

Overall, the results obtained using these rat behavioral models of nerve injury were somewhat disappointing and disconcerting, since they did not lead to straightforward answers. The controversial results from these models provide conflicting evidence on the role of the sympathetic nervous system in generating pain behavior, indicating that behavioral animal models must be designed with utmost care. Small changes in the experimental procedures may create major behavioral changes. Behavioral testing should be performed in a blinded fashion if possible, and the models must be designed in close association with the clinical situation to be certain to test what is intended to be tested. The results of systemic pharmacological interventions (e.g., application of adrenoceptor agonists or antagonists, guanethidine, or chemical sympathectomy with 5-hydroxydopamine) do not necessarily allow functional interpretations, partly because of the widespread systemic effects of these agents. Finally, the controversial results obtained from the animal behavior models may argue that the models do not reflect SMP as defined clinically.

This controversial situation does not mean that behavioral animal models are useless and cannot simulate the clinical situation. On the contrary, there are good reasons to assume that we are dependent on the design of these animal models if we want to unravel the mechanisms behind the different types of pathological pain in which the sympathetic nervous system is involved. However, a single model

Table 2 Nerve lesion animal (rat) models to study sympathetically maintained pain[a]

Model	*Nerve lesion*	*Behavioral measurements*	*Intervention at the sympathetic supply*	*Results*
Autotomy model[b]	Ligation and transection of sciatic and saphenous nerve	Autotomy of affected paw (self-mutilation)	Intraperitoneal guanethidine	Autotomy ↓
			Subcutaneous noradrenaline	Autotomy ↑
Partial lesion of the sciatic nerve[c]	Ligation of 33–50% of sciatic nerve	Heat HA	Intraperitoneal guanethidine	Heat HA ↓, mech HA ∅
		Mech HA	Noradrenaline local	Heat HA ↑, mech HA ↑
Chronic constriction, injury of the sciatic nerve[d]	Loose ligation of sciatic nerve	Heat HA	Surgical sympathectomy	Heat HA ↓, mech HA (↓), ∅
		Mech HA	Intraperitoneal guanethidine	Heat HA↓, mech HA (↓), ∅
Spinal nerve ligation model[e]	Ligation and transection of spinal nerve L5, L5/L6	Mech HA/allodynia	Surgical sympathectomy	Mech HA ↓ or ∅[g]
		Heat HA	Surgical sympathectomy	Heat HA ↓
Cryolysis of the sciatic nerve[f]	Cryolysis of sciatic nerve	Mech allodynia	Surgical sympathectomy	Mech allodynia ∅

[a]Key to measurements and results: HA, hyperalgesia; mech, mechanical; ↓, decrease; ↑, increase; ∅, no effect.
[b]Based on data from Wall PD, Scadding JW, and Tomkiewicz MM (1979) *Pain* 6: 175–182; Wall PD, Devor M, Inbal R, et al. (1979) *Pain* 7: 103–113; Coderre TJ, Abbott FV, and Melzack R (1984) *Pain* 18: 13–23; Coderre T, Grimes RW, and Melzak RM (1986), *Pain* 26: 61–84; and Coderre T and Melzack RM (1986) *Experimental Neurology* 92: 713–722.
[c]Based on data from Seltzer Z, Dubner R, and Shir Y (1990) *Pain* 43: 205–280; Shir Y and Seltzer Z (1991) *Pain* 45: 309–320; and Tracey DJ, Cunningham JE, and Romm MA (1995) *Pain* 60: 317–327.
[d]Based on data from Bennett G and Xie YK (1988) *Pain* 33: 87–107; Neil A, Attal N, and Guilbaud G (1991) *Brain Research* 565: 237–246; Desmeules JA, Kayser V, Weil-Fuggaza J, et al. (1995) *Neuroscience* 67: 941–951; and Kim SH, Yoon YW, and Chung JM (1997) *Experimental Brain Research* 113: 200–206.
[e]Based on data from Kim SH and Chung JM (1991) *Neuroscience Letters* 134: 131–134; Kim SH, Na, HS, Sheen K, et al. (1993) *Pain* 55: 85–92; Kim SH, Yoon YW, and Chung JM (1997) *Experimental Brain Research* 113: 200–206; Kinnman E and Levine JD (1995) *Neuroscience* 64: 751–767; and Ringkamp M, Eschenfelder S, Grethel EJ, et al. (1999) *Pain* 79: 143–153.
[f]Based on data from Willenbring S, Beauprie IG, and DeLeo JA (1995) *Anesthesia & Analgesia* 81: 544–548.
[g] The negative effects of surgical sympathectomy reported by Ringkamp et al. (1999; *Pain* 79: 143–153) were obtained independently by two different groups.

cannot be expected to represent SMP in different groups of patients. Furthermore, the animal models so far tested are related to nerve lesions. Rat models simulating SMP in patients with pain states developing after trauma without nerve lesion (see Table 1) practically do not exist (e.g., for patients with CRPS type I).

Effects of Sympathetic Postganglionic Neurons on Afferent Neurons

Physiologically sympathetic postganglionic neurons do not directly excite primary afferent neurons innervating skin or deep somatic tissues. Some small effects may occur indirectly (e.g., via the vascular bed and subsequent changes of blood flow). The situation may change under pathophysiological conditions: excitation of sympathetic neurons may activate or sensitize afferent neurons. Furthermore, sympathetic postganglionic terminals may mediate sensitization of nociceptive afferents for mechanical stimulation during inflammation. Finally, the sympathoadrenal system may be involved in the sensitization of nociceptors for mechanical stimulation. The ways of sympathetic–afferent coupling discussed in the following sections are schematically illustrated in Figure 3.

Coupling between Lesioned Postganglionic and Afferent Nerve Terminals

Subsequent to nerve section or ligation, a few afferent myelinated and unmyelinated nerve fibers ending in a neuroma can be excited following electrical stimulation of the sympathetic supply or by noradrenaline or adrenaline injected systemically. The coupling occurs in young neuromas but less so in old ones, weeks and months after nerve lesion. The frequency of electrical stimulation, applied to the sympathetic trunk, necessary to excite the lesioned afferent nerve fibers is high. Physiologically sympathetic neurons do not discharge at these frequencies. Clinically, neuroma

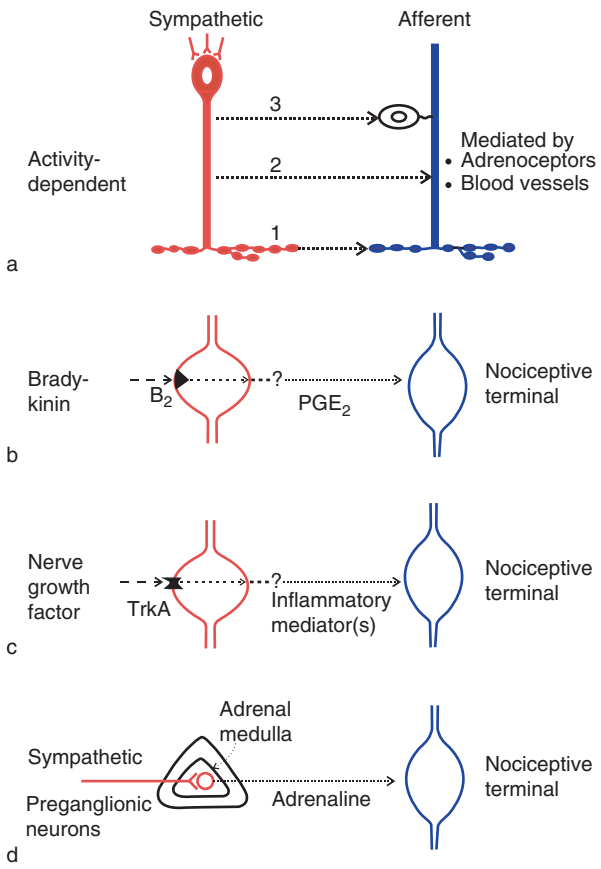

Figure 3 Ways hypothesized to couple sympathetic and primary afferent neurons following peripheral nerve lesion (a) or during inflammation (b–d). (a) These types of coupling depend on the activity in the sympathetic neurons and on functional adrenoceptors expressed by the afferent neurons, or are mediated indirectly via the blood vessels (blood flow). Coupling can occur in the periphery (1), in the dorsal root ganglion (3), or possibly also in the lesioned nerve (2). (b) The inflammatory mediator bradykinin acts at B_2 receptors in the membrane of the sympathetic varicosities or in cells upstream of these varicosities, inducing release of prostaglandin E_2 (PGE_2) and sensitization of nociceptors. This way of coupling is probably not dependent on activity in the sympathetic neurons. (c) Nerve growth factor released during an experimental inflammation reacts with the high-affinity receptor TrkA, in the membrane of the sympathetic varicosities, inducing release of an inflammatory mediator (or mediators) and sensitization of nociceptors. This effect is probably not dependent on activity in the sympathetic neurons. (d) Activation of the adrenal medulla by sympathetic preganglionic neurons leads to release of adrenaline, which generates sensitization of nociceptors. The question mark in (b) and (c) indicates that PGE_2 or other inflammatory mediators may be released by cells other than the sympathetic varicosities. Adapted from Jänig W and Häbler H-J (2000) Sympathetic nervous system: Contribution to chronic pain. *Progress in Brain Research* 129: 453–470.

pain is usually not dependent on sympathetic activity. Histological observations show that catecholamine-containing axon profiles are rare within the neuroma and for several centimeters proximal to it. The coupling is chemical and occurs via noradrenaline acting on α-adrenoceptors, although other mediator substances may also be involved. Ephaptic coupling between sympathetic fibers and afferent fibers has so far not been observed in a neuroma.

The situation is different when afferent and sympathetic fibers are allowed to regenerate to the target tissue. This has been shown experimentally for the chronic situation more than a year after cross-union of nerves (e.g., adaptation of the proximal stump of the sural or superficial peroneal nerve to the distal stump of the tibial nerve, in the cat) and after reinnervation of appropriate and inappropriate target tissues. In these cases, unmyelinated afferent fibers may be vigorously excited by electrical low-frequency stimulation of the sympathetic supply. This excitation is adrenoceptor mediated (Figure 4). It can be mimicked by adrenaline or noradrenaline and blocked by α-adrenoceptor blockers (e.g., phentolamine).

Coupling between Unlesioned Postganglionic and Afferent Nerve Terminals following Partial Nerve Lesion

Intact C-fiber polymodal nociceptors in skin may develop sensitivity to catecholamines following partial nerve injury. Sympathetic nerve stimulation and noradrenaline may excite the polymodal nociceptors or sensitize them for heat stimuli. The activation and sensitization are already seen 4–10 days after partial nerve lesion and are maintained for at least 150 days. This sympathetic–afferent coupling apparently involves the nonlesioned polymodal nociceptive afferent axons that project through the lesion site, and nonlesioned postganglionic axons. It is suggested that the unlesioned unmyelinated afferents develop hyperreactivity to catecholamines following degeneration of the sympathetic postganglionic axons. Expression of adrenoceptors in afferent fibers may be triggered by collateral sprouting of both afferent fibers and postganglionic fibers in the target tissue. However, activation or sensitization of nociceptors via sympathetic–afferent coupling requires sympathetic stimulation at high non-physiological frequencies.

Coupling in the Dorsal Root Ganglion and Collateral Sprouting following Peripheral Nerve Lesion

Chemical sympathetic–afferent coupling following peripheral nerve lesion (e.g., of a spinal nerve, the sciatic nerve, or another limb nerve) may occur proximally to the injury site in the nerve or in the dorsal root ganglion (DRG; Figure 3(a), site 3). Sympathetic postganglionic fibers reach the spinal nerves via gray rami. Most of these fibers project distally to

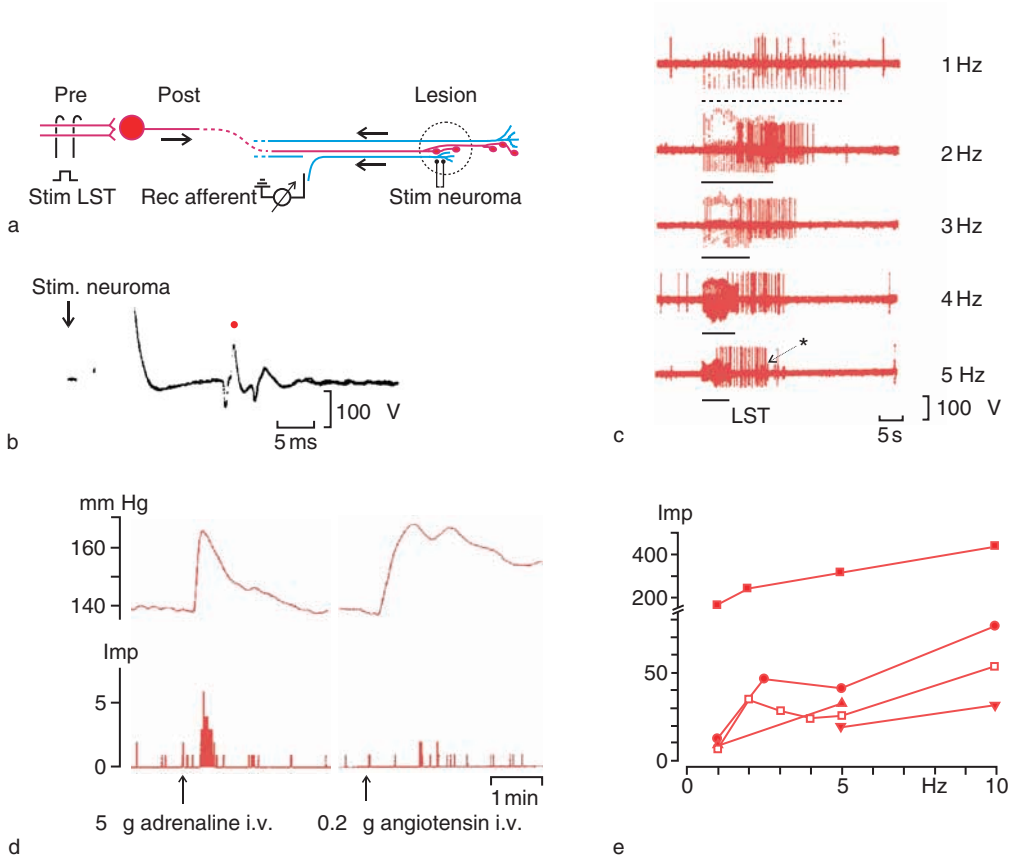

Figure 4 Sympathetic–afferent coupling after nerve lesion, with subsequent fiber regeneration. Excitation of unmyelinated afferent fibers by electrical stimulation of sympathetic fibers following nerve injury. Unmyelinated primary afferents were recorded in cats 11–20 months after the central cut stump of the sural or superficial peroneal nerve (both skin nerves) had been adapted to the distal stump of the transected tibial nerve (mixed nerve). This preparation was designed to mimic the consequences of lesion of a mixed nerve. A 'neuroma-in-continuity' at the lesion site was present, and cutaneous nerve fibers had regenerated into skin and deep somatic tissue supplied by the mixed nerve. (a) Experimental setup. Post, postganglionic; pre, preganglionic; stim, stimulation; rec, recording; LST, lumbar sympathetic trunk. (b) The afferent fibers were identified as unmyelinated by electrical stimulation of the neuroma with single pulses. The signal indicated by the red dot was recorded from the same afferent fiber as in (c) and (d). (c) Record from a single unmyelinated afferent fiber. Supramaximal stimulation of the LST with trains of 30 pulses at 1–5 Hz (trains and stimulation artifacts indicated by bars). Note that the afferent fiber had some low rate of ongoing activity (impulses before the trains at 1 and 4 Hz) and a second fiber was activated at 5 Hz stimulation (marked by an asterisk). (d) Adrenaline (5 μg, injected intravenously) activated the fiber. Angiotensin (0.2 μg, injected intravenously) generated a large increase of mean arterial blood pressure but did not activate the afferent fiber. (e) Stimulus response curves for the single afferent fiber (□) and four filaments containing two or three (●,▲/▼) and more than five (■) afferent fibers. Ordinate scale is the total number of impulses (imp) exceeding ongoing activity in response to variable stimulation frequency of the LST. Adapted from Häbler H-J, Jänig W, and Koltzenburg M (1987) Activation of unmyelinated afferents in chronically lesioned nerves by adrenaline and excitation of sympathetic efferents in the cat. *Neuroscience Letters* 82: 35–40.

peripheral target cells; some project proximally (e.g., to the DRG) and are normally found along blood vessels. After an experimental nerve lesion (e.g., transection and ligation of the sciatic nerve in rats), many perivascular catecholamine-containing axons of the unlesioned proximally projecting neurons start to sprout in the DRGs which contain somata with lesioned axons. The extent of this collateral sprouting increases with time after nerve lesion. For more than 70 days after nerve lesion, some somata are partially or almost completely surrounded by varicose catecholaminergic terminals, and the frequency of these catecholamine-fluorescent structures increases. The noradrenergic axons sprout preferentially to axotomized large-diameter afferent cell bodies.

Afferent neurons projecting in the lesioned nerve can be excited or depressed in their activity via the DRG by electrical stimulation of sympathetic neurons and by catecholamines:

1. In the first 2–3 weeks after nerve lesion, most A-fiber neurons with spontaneous activity and only very few silent A-fiber neurons are excited, yet the spontaneous activity is not maintained by

activity in sympathetic neurons. The sympathetically evoked activity in afferent neurons probably is partly mediated indirectly by ischemia in the DRG. When the catecholaminergic sprouting in the DRG is more prominent at later time periods, most spontaneously active A-fiber neurons are inhibited during stimulation of the sympathetic supply, but are rarely excited.
2. Only very few afferent neurons with unmyelinated axons respond to electrical stimulation of the sympathetic trunk; most of them are inhibited in their activity.
3. The coupling from sympathetic postganglionic fibers to afferent cell bodies in the DRG occurs only to spontaneously active muscle afferent neurons with Aδ-fibers.
4. Activation of the afferent neurons, via the DRG, requires high-frequency stimulation of the sympathetic neurons at 5–20 Hz, discharge rates that occur rarely or not at all physiologically.
5. Noradrenaline and preferentially α_1-adrenoceptors are involved in sympathetic–afferent coupling in the DRG.
6. Mechanical allodynic/hyperalgesic behavior shown by rats with L5 spinal nerve injury (transection and ligation) is not dependent on the innervation of the DRG by sympathetic neurons.

The mechanism leading to the collateral sprouting of postganglionic noradrenergic fibers in the DRG is possibly related to neurotrophic signals (and their receptors), generated by afferent cells with lesioned axons in the DRG and the surrounding satellite glia cells. Whether these aberrant pathological connections can account for spontaneous pain and allodynia in some patients after peripheral nerve lesions awaits further investigations.

Indirect Involvement of the Sympathetic Nervous System in the Generation of Pain

Modulation of Excitability of the Nociceptive Afferent by Its Micromilieu

Nociceptive afferents are embedded in a complex micromilieu (Figure 5). The state of this micromilieu surrounding the receptive terminals depends on mediator substances, released during inflammatory processes following trauma, from nonneural cells such as mast cells, polymorphonuclear leucocytes, macrophages, fibroblasts, endothelial cells, or other cells. The microcirculation is under neural control of sympathetic vasoconstrictor neurons. Activation of subgroups of nociceptive primary afferents not only causes orthodromic impulse traffic but also arteriolar (precapillary) vasodilation and (in most tissues, but not in human skin) venular plasma extravasation by release of neuropeptides from the receptive terminals (e.g., substance P and calcitonin-gene-related peptide). Both afferent-induced changes are called collectively 'neurogenic inflammation.' This shows that there are possibilities for indirect coupling between sympathetic and afferent nerve terminals:

1. Vascular perfusion of the micromilieu surrounding the nociceptors after nerve trauma may change as consequences of denervation and reinnervation by postganglionic vasoconstrictor neurons and afferent nociceptive neurons.
2. Hyperreactivity of blood vessels to activity in sympathetic neurons or circulating catecholamines may develop.
3. Blood vessels exhibit changes in vasodilation in response to antidromic activation of reinnervated peptidergic unmyelinated afferents. Thus, after a nerve lesion, reinnervated blood vessels may be under stronger than normal vasoconstrictor influence, which can no longer be counteracted by an afferent-mediated vasodilation. *In vitro* investigation of the rat tail artery has shown that the functional recovery of neurovascular transmission is permanently compromised after nerve lesion.
4. Nociceptors may be sensitized by inflammatory mediators such as prostaglandins and interleukins, the release of which from nonneuronal cells of the micromilieu may be mediated by noradrenergic nerve terminals.

Sensitization of Nociceptors Mediated by Sympathetic Terminals Independent of Excitation and Release of Noradrenaline

Following a single intradermal injection of the inflammatory mediator bradykinin into the dorsal skin of the rat paw, the withdrawal threshold to mechanical stimulation of the rat hindpaw decreases dose dependently for more than 1 h. This type of mechanical hyperalgesic behavior is mediated by the B_2 bradykinin receptor. In normal rats, bradykinin-induced hyperalgesic behavior is blocked by the cyclooxygenase inhibitor indomethacin and therefore is mediated by a prostaglandin (probably prostaglandin E_2) that sensitizes nociceptors for mechanical stimulation.

The decrease in paw withdrawal threshold in response to mechanical stimulation generated by intradermal injection of bradykinin is significantly reduced after surgical sympathectomy. Decentralization of the paravertebral sympathetic ganglia (denervating the postganglionic neurons by cutting the preganglionic sympathetic axons) does not abolish the bradykinin-induced mechanical hyperalgesic behavior (Figure 6). This indicates that the sensitizing effect

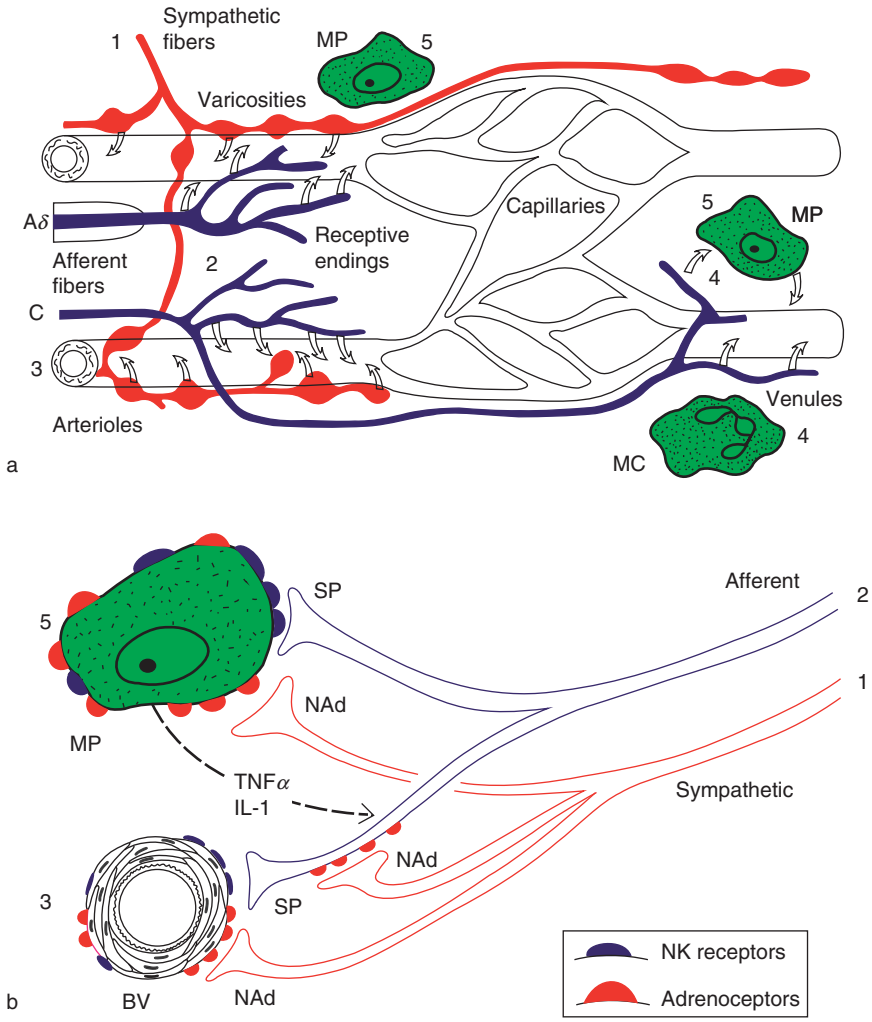

Figure 5 (a, b) The micromilieu of nociceptors. The microenvironment of primary afferents is thought to affect the functional properties of the receptive endings of myelinated (Aδ) and unmyelinated (C) afferent fibers with nociceptive function. This has been particularly documented for inflammatory processes. The vascular bed consists of arterioles (directly innervated by sympathetic and afferent (Aδ and C) fibers), capillaries (not innervated and not influenced by nerve fibers), and venules (not directly innervated, but influenced by nerve fibers). The micromilieu depends on several interacting components. Neural activity in postganglionic noradrenergic fibers (1) supplying blood vessels (3) causes release of noradrenaline (NAd) and possibly other substances causing vasoconstriction. Excitation of primary afferents (Aδ- and C-fibers; 2) causes vasodilation in precapillary arterioles and plasma extravasation in postcapillary venules (C-fibers only) by the release of substance P (SP) and other vasoactive compounds (e.g., calcitonin gene-related peptide). Some of these effects may be mediated by nonneuronal cells, such as mast cells (MC; 4) and macrophages (MP; 5). Other factors that affect the control of the microcirculation are the myogenic properties of arterioles (3) and more global environmental influences, such as a change of the temperature and the metabolic state of the tissue. Hypothetical relation (shown in b) between sympathetic noradrenergic nerve fibers (1), peptidergic afferent nerve fibers (2), blood vessels (BV; 3), and macrophages (4). The activated and sensitized afferent nerve fibers activate macrophages (via substance P release). The immune cells start to release cytokines, such as tumor necrosis factor-α (TNF-α) and interleukin-1 (IL-1), which further activate afferent fibers by enhancing sodium influx into the cells. Substance P (also calcitonin gene-related peptide; not shown) released from the afferent nerve fibers reacts with neurokinin-1 (NK) receptors (calcitonin gene-related peptide receptors) in the blood vessels (arteriolar vasodilation, venular plasma extravasation; neurogenic inflammation). The sympathetic nerve fibers interact with this system potentially on three levels: via adrenoceptors (mainly α) on the blood vessels (vasoconstriction), via adrenoceptors (mainly β) on macrophages (further release of cytokines), and via adrenoceptors (mainly α) on afferents (further sensitization of these fibers). Adapted from Jänig W and Baron R (2003) Complex regional pain syndrome: Mystery explained? *Lancet Neurology* 2: 687–697.

of bradykinin is not dependent on activity in the sympathetic neurons innervating skin, and therefore also not on release of noradrenaline, and that bradykinin stimulates the release of prostaglandin from the sympathetic terminals or from other cells in association with the sympathetic terminals in the skin (Figure 3(b)). Mechanical hyperalgesic behavior generated by intracutaneous injection of bradykinin, with its dependence

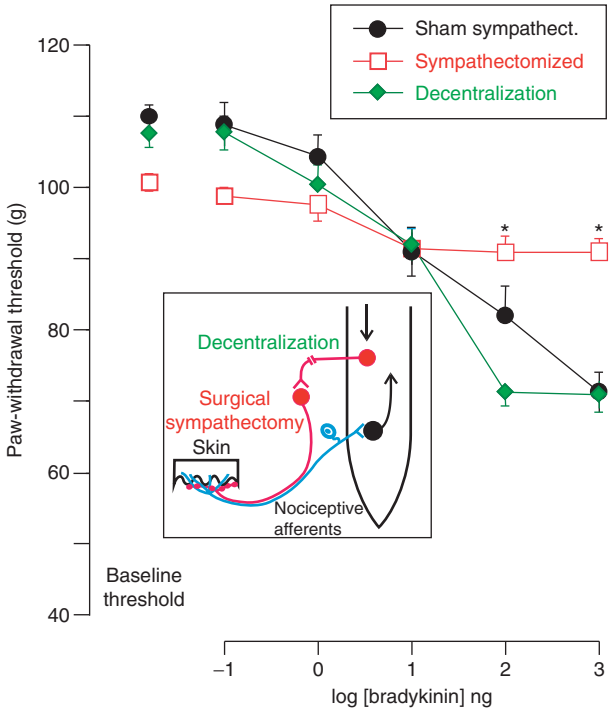

Figure 6 Mechanical hyperalgesic behavior and sympathetic innervation. Cutaneous nociceptors on the dorsum of the paw are stimulated by a linearly increasing mechanical force. Paw withdrawal threshold is defined as the mean (+S.E.M.) minimum force (in grams, ordinate scale) at which the rat withdraws its paw. Abscissa scale: log dose of bradykinin (in nanograms) injected in a volume of 2.5 μl of saline into the dermis of the skin. Decrease of paw withdrawal threshold in response to mechanical stimulation induced by intradermal injection of bradykinin in sham-sympathectomized (control) rats (●, six hindpaws), in surgically sympathectomized rats (lumbar paravertebral ganglia L2–L4 removed 8 days before; □, 12 hindpaws), and in rats with decentralized lumbar sympathetic chains (section of white rami with preganglionic axons to paravertebral ganglia L2 and L3 and chain rostral to ganglion L2 8 days before; ◆, ten paws). Sham sympathectomy and sympathetic decentralization were both significantly different from sympathectomy groups at doses of bradykinin of 10^2 and 10^3 ng (Δ, $P < 0.01$). Adapted from Khasar SG, Miao FJ-P, Jänig W, et al. (1998) Modulation of bradykinin-induced mechanical hyperalgesia in the rat skin by activity in the abdominal vagal afferents. *European Journal of Neuroscience* 10: 435–444.

on the sympathetic innervation of the skin, is an interesting phenomenon. However, the idea that it is due to prostanoids sensitizing nociceptors (since indomethacin prevents it) appears to be too simple. The reasons for this are that (1) sympathetic fibers mediate this hyperalgesia, independent of neural activity and release of noradrenaline; (2) indomethacin does not block this behavior under certain conditions (e.g., when the adrenal medullae are activated after vagotomy or after sound stress); (3) sensitization by bradykinin of nociceptive afferents innervating deep somatic tissues, in response to mechanical stimulation, has been demonstrated in neurophysiological experiments in the cat, yet not in humans, for cutaneous pain; and (4) in skin, there is a spatial dissociation between afferent nociceptive fibers and sympathetic fibers, many of the former ending in the epidermis and subepidermal plexus, whereas most of the latter terminate in deeper dermal layers, raising the question as to the nature of the interaction between both groups of fibers.

Sensitization of Nociceptors Initiated by Nerve Growth Factor or Cytokines, Possibly Mediated by Sympathetic Terminals

Based on behavioral and pharmacological experiments it has been deduced that nociceptive afferents can be sensitized by nerve growth factor (NGF) or proinflammatory cytokines involving noradrenergic sympathetic postganglionic fibers. The underlying mechanisms of such interactions are largely unknown and have to be worked out by recording the activity of nociceptor neurons with electrophysiological techniques.

NGF Systemic injection of NGF is followed by a transient thermal and mechanical hyperalgesia in rats and humans. During experimental inflammation (evoked by Freund's adjuvant in the rat hindpaw), the concentration of NGF increases in the inflamed tissue, paralleled by the development of thermal and mechanical hyperalgesia. Both are prevented by anti-NGF antibodies. The mechanisms responsible are sensitization of nociceptors via high-affinity NGF receptors (TrkA receptors) and an increased synthesis of calcitonin-gene-related peptide (CGRP) and substance P in the afferent cell bodies; the latter is induced by NGF being taken up by the afferent terminals and transported to the cell bodies. The NGF-induced sensitization of nociceptors, or part of it, also seems to be mediated indirectly by the sympathetic postganglionic terminals. In rats, heat and mechanical hyperalgesic behavior generated by local injection of NGF into the skin are significantly reduced after chemical or surgical sympathectomy. These experiments suggest that NGF released during inflammation by inflammatory cells acts on the sympathetic terminals via high-affinity TrkA receptors, inducing the release of inflammatory mediators (the source of which is unknown) and subsequently the sensitization of nociceptors for mechanical and heat stimuli (Figure 3(c)). It is unclear whether this sensitization of nociceptors mediated by terminal sympathetic nerve fibers is dependent on or independent of neural activity in the sympathetic neurons, release of noradrenaline, and adrenoceptors expressed in the nociceptive afferent neurons.

Proinflammatory cytokines Tissue injury, injection of the bacterial cell wall endotoxin lipopolysaccharide,

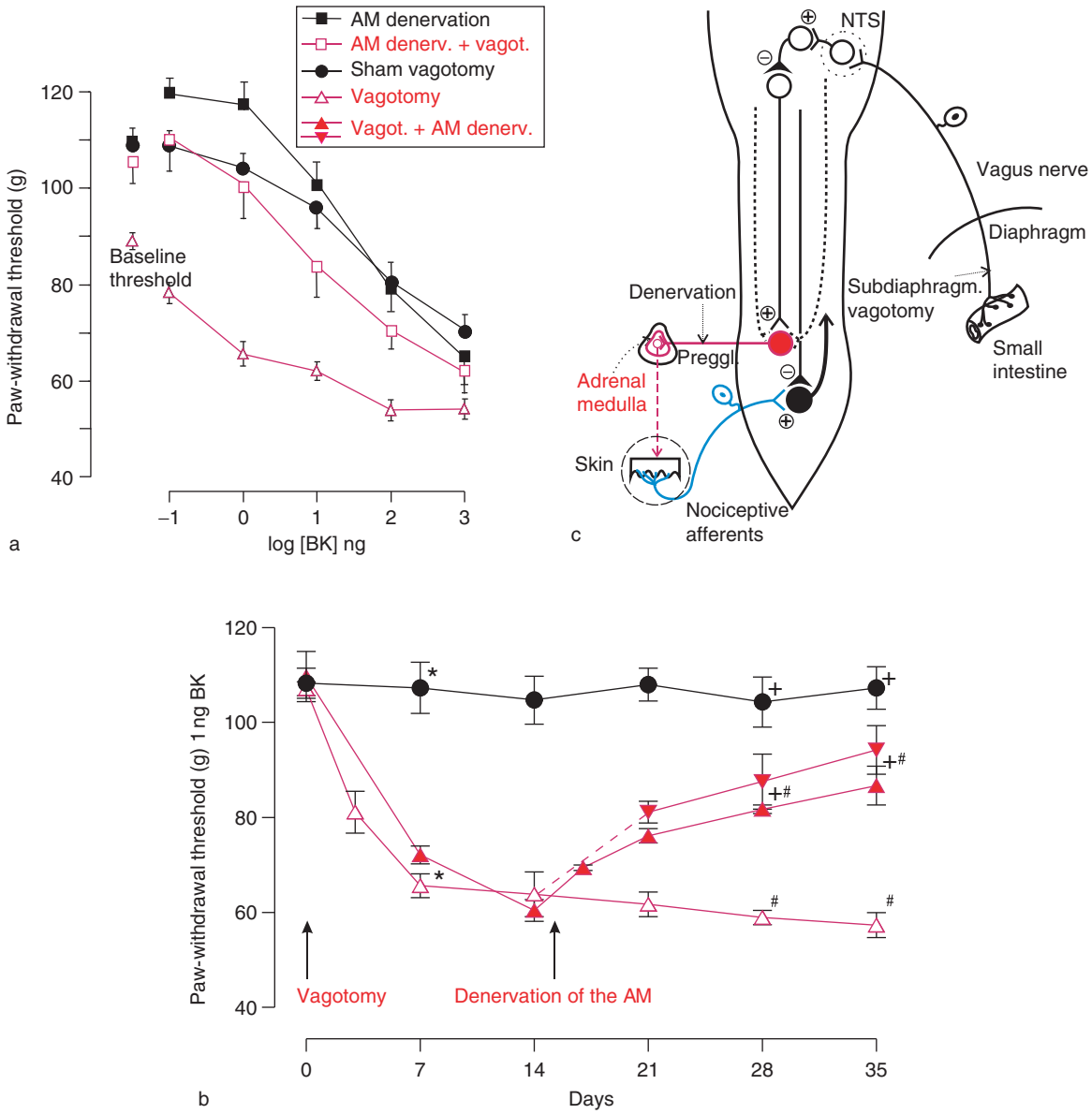

Figure 7 Long-term enhancement of bradykinin(BK)-induced mechanical hyperalgesia after activation of the adrenal medullae (by vagotomy) and its disappearance after denervation of the adrenal medullae (AM). (a) Decrease of paw withdrawal threshold in response to mechanical stimulation of the dorsum of the rat hindpaw induced by BK injected intradermally (BK-induced behavioral mechanical hyperalgesia), in sham-vagotomized rats (●, $n = 18$), in vagotomized rats (Δ, $n = 16$; 7 days after subdiaphragmatic vagotomy), in rats with denervated adrenal medullae (■, $n = 6$), and in vagotomized rats with denervated AM (□, $n = 6$) (for details, see **Figure 6** caption). (b) Total change of paw withdrawal threshold (baseline plus BK-induced) in response to intradermal injection of 1 ng of BK in rats before and 7–35 days after sham vagotomy (●, $n = 8$), before and 7–35 days after vagotomy (Δ, $n = 6$), and in rats whose AM are denervated 14 days after vagotomy and for measurements taken up to 35 days after initial surgery (▲, rats tested after vagotomy and denervation of the AM, $n = 6$; ▼, rats only tested after additional denervation of the AM, $n = 4$). Note that 1 ng of BK injected intradermally does not significantly change the paw withdrawal threshold in rats with intact vagus nerves (see a); *, $p < 0.01$, ▲/Δ vs. ● on day 7; +, $p > 0.05$, ▲/▼ vs. ● on days 28 and 35; #, $p < 0.01$, ▲/▼ vs. Δ on days 28 and 35. (c) Schematic diagram showing the proposed spinal cord and brain stem neural circuits which modulate nociceptor sensitivity via the sympathoadrenal system (adrenal medulla). Sensitivity of cutaneous nociceptors for mechanical stimulation is modulated by adrenaline released by the AM. Activation of the AM increases the sensitivity of the nociceptors. Activity in preganglionic neurons innervating the AM depends on activity in vagal afferents from the small intestine; this activity has an inhibitory influence on the central pathways to these preganglionic (preggl) neurons. Thus, interruption of the vagal afferents leads to activation of the AM. It is hypothesized that these neuronal (reflex) circuits in the brain stem are under the control of upper brain stem, hypothalamus, and telencephalon. Dotted lines indicate axons of sympathetic premotor neurons in the brain stem which project through the dorsolateral funiculi of the spinal cord to the preganglionic neurons of the AM; NTS, nucleus tractus solitarii; +, excitation; − , inhibition. Adapted from Khasar SG, Miao FJ-P, Jänig W, et al. (1998) Vagotomy-induced enhancement of mechanical hyperalgesia in the rat is sympathoadrenal-mediated. *Journal of Neuroscience* 18: 3043–3049.

or injection of carrageenan (a plant polysaccharide) stimulate tissue inflammation and lead to a sensitization of nociceptors that is reflected in mechanical or heat hyperalgesia. Systematic pharmacological interventions using blockers or inhibitors of the various mediators demonstrate that the proinflammatory cytokines tumor necrosis factor-α (TNF-α) and interleukins (IL-1, IL-6, and IL-8) may be involved in this process of nociceptor sensitization. Pathogenic stimuli activate resident cells, with release of inflammatory mediators, including bradykinin. Activation of macrophages, monocytes, and other immune-related cells leads to release of TNF-α. Nociceptors are believed to be sensitized by two possible pathways: (1) TNF-α induces production of IL-6 and IL-1β by immune cells. These interleukins stimulate cyclooxygenase-2 and the production of prostaglandins E_2 and I_2 (PGE_2, PGI_2), which in turn react with the nociceptive terminal via E-type prostaglandin receptors. (2) TNF-α induces the release of IL-8 from endothelial cells and macrophages. IL-8 reacts with the sympathetic terminals that are supposed to mediate sensitization of nociceptive afferent terminals by release of noradrenaline, to act via β_2-adrenoceptors. These two peripheral pathways by which nociceptive afferents are hypothesized to be sensitized involving cytokines are under inhibitory control of anti-inflammatory interleukins (e.g., IL-4 and IL-10) and circulating glucocorticoids.

Sympathoadrenal System and Nociceptor Sensitization

Activity of the sympathoadrenal system (adrenal medullae) is regulated by central circuits that are different from those regulating other sympathetic systems. Activation of this system (Figure 7(c)), but not of the sympathoneural system, generates mechanical hyperalgesia and enhances bradykinin-induced mechanical hyperalgesia (decrease of paw withdrawal threshold in response to mechanical stimulation of the skin; compare • and Δ in Figure 7(a)). Both develop slowly over 7–14 days during continuous activation of the adrenal medullae (AM; e.g., induced by subdiaphragmatic vagotomy) and are maintained over the 35-day test period (Δ in (Figure 7(b)). In rats with denervated AM (adrenal nerves with preganglionic axons innervating the cells of the AM sectioned; (Figure 7(c)), subdiaphragmatic vagotomy has only a small effect (□ and ■ in Figure 7(a)), generated by removal of inhibition of nociceptive impulse transmission in the spinal dorsal horn (Figure 7(c)). Denervation of the AM, 14 days after vagotomy, when the decrease of paw withdrawal threshold is maximal, is followed by a slow increase of paw withdrawal threshold, taking more than 3 weeks (▲ and ▼ in Figure 7(b)). Thus, the decrease of paw withdrawal threshold following vagotomy is reversed after denervation of the AM. Application of adrenaline through a minipump for 3–14 days slowly enhances the development of mechanical hyperalgesia induced by bradykinin injected intradermally. Continuous application of a β_2-adrenoceptor blocker over 7–14 days by way of a minipump prevents the enhancement of this hyperalgesic behavior generated by activation of the AM.

The results of this experiment imply that adrenaline released by persistent activation of the AM sensitizes cutaneous nociceptors for mechanical stimuli. This sensitization of nociceptors and its reversal are slow and take days to develop. The slow time course implies that the nociceptor sensitization cannot be acutely blocked by an adrenoceptor antagonist given intracutaneously or systemically. Adrenaline probably does not act directly on the cutaneous nociceptors but instead acts on cells (e.g., mast cells, macrophages, or keratinocytes) in the microenvironment of the nociceptors. These cells then release substances that induce slow changes, generating nociceptor sensitization. Change in the sensitivity of a population of nociceptors, generated by adrenaline released by the AM, regulated by the brain, would be a novel mechanism of sensitization, different from mechanisms that lead to activation and/or sensitization of nociceptors by sympathetic–afferent coupling, as discussed earlier (Figure 3(d)).

This peripheral mechanism of pain and hyperalgesia involving adrenaline released by the sympathoadrenal system and acting on nociceptors may operate in ill-defined pain syndromes such as irritable bowel syndrome, functional dyspepsia, fibromyalgia, and chronic fatigue syndrome. Conclusions about long-term sensitization of nociceptors for mechanical stimulation by adrenaline fully rest on behavioral experiments. Neurophysiological experiments *in vivo* on primary afferent nociceptive neurons are required to test directly whether all or only a subpopulation of nociceptors are sensitized.

See also: Adrenergic Receptors; Autonomic Nervous System; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neuropathic Pain; Neuropeptides: Pain; Nociceptor Responses; Norepinephrine: Adrenergic Receptors; Pain Pathways: Descending Modulation; Pain and Plasticity; Pain Management; Pain: Central; Spinal Cord Pain Systems; Sympathetic Nervous System; Sympathoadrenal System: Neural Arm of the Stress Response; Visceral Pain.

Further Reading

Ali Z, Raja SN, Wesselmann U, et al. (2000) Intradermal injection of norepinephrine evokes pain in patients with sympathetically maintained pain. *Pain* 88: 161–168.

Baron R, Schattschneider J, Binder A, et al. (2002) Relation between sympathetic vasoconstrictor activity and pain and hyperalgesia in complex regional pain syndromes: A case-control study. *Lancet* 359: 1655–1660.

Bennett G and Xie YK (1988) *Pain* 33: 87–107.

Burnstock G (2001) Purine-mediated signaling in pain and visceral perception. *Trends in Pharmacological Sciences* 22: 182–188.

Coderre T and Melzack RM (1986) *Experimental Neurology* 92: 713–722.

Coderre T, Grimes RW, and Melzak RM (1986) *Pain* 26: 61–84.

Coderre TJ, Abbott FV, and Melzack R (1984) *Pain* 18: 13–23.

Desmeules JA, Kayser V, Weil-Fuggaza J, et al. (1995) *Neuroscience* 67: 941–951.

Häbler H-J, Jänig W, and Koltzenburg M (1987) Activation of unmyelinated afferents in chronically lesioned nerves by adrenaline and excitation of sympathetic efferents in the cat. *Neuroscience Letters* 82: 35–40.

Jänig W (2006) *The Integrative Action of the Autonomic Nervous System. Neurobiology of Homeostasis*. Cambridge, NY: Cambridge University Press.

Jänig W and Baron R (2001) The role of the sympathetic nervous system in neuropathic pain: Clinical observations and animal models. In: Hansson PT, Fields HL, Hill RG, et al. (eds.) *Neuropathic Pain: Pathophysiology and Treatment*, pp. 125–149. Seattle: IASP Press.

Jänig W and Häbler H-J (2000) Sympathetic nervous system: Contribution to chronic pain. *Progress in Brain Research* 129: 453–470.

Jänig W and Baron R (2003) Complex regional pain syndrome: Mystery explained? *Lancet Neurology* 2: 687–697.

Jänig W and Levine JD (2006) Autonomic-endocrine-immune responses in acute and chronic pain. In: McMahon SB and koltzenburg M (eds.) *Wall and Melzack's Textbook of Pain*, 5th edn., pp. 205–218. Edinburgh: Churchill Livingstone.

Jänig W and Baron R (2007) Sympathetic nervous system and pain. In: Schmidt RF and Willis WD (eds.) *Encyclopedia of Pain*, pp. 2356–2366. Berlin: Springer.

Jänig W, Khasar SG, Levine JD, et al. (2000) The role of vagal visceral afferents in the control of nociception. *Progress in Brain Research* 122: 273–287.

Khasar SG, Miao FJ-P, Jänig W, et al. (1998) Modulation of brady kinin-induced mechanical hyperalgesia in the rat skin by activity in the abdominal vagal afferents. *European Journal of Neuroscience* 10: 435–444.

Khasar SG, Miao FJ-P, Jänig W, et al. (1998) Vagotomy-induced enhancement of mechanical hyperalgesia in the rat is sympatoadrenal-mediated. *Journal of Neuroscience* 18: 3043–3049.

Kim SH and Chung JM (1991) *Neuroscience Letters* 134: 131–134.

Kim SH, Na HS, Sheen K, et al. (1993) *Pain* 55: 85–92.

Kim SH, Yoon YW, and Chung JM (1997) *Experimental Brain Research* 113: 200–206.

Kinnman E and Levine JD (1995) *Neuroscience* 64: 751–767.

Neil A, Attal N, and Guilbaud G (1991) *Brain Research* 565: 237–246.

Poole S, Cunha FQ, and Ferreira SH (1999) Hyperalgesia from subcutaneous cytokines. In: Watkins LR and Maier SF (eds.) *Cytokines and Pain*, pp. 59–87. Basel: Birkhäuser Verlag AG.

Price DD, Long S, Wilsey B, et al. (1998) Analysis of peak magnitude and duration of analgesia produced by local anesthetics injected into sympathetic ganglia of complex regional pain syndrome patients. *Clinical Journal of Pain* 14: 216–226.

Ringkamp M, Eschenfelder S, Grethel EJ, et al. (1999) *Pain* 79: 143–153.

Seltzer Z, Dubner R, and Shir Y (1990) *Pain* 43: 205–280.

Shir Y and Seltzer Z (1991) *Pain* 45: 309–320.

Straub RH, Baerwald CG, Wahle M, et al. (2005) Autonomic dysfunction in rheumatic diseases. *Rheumatic Disease Clinics of North America* 31: 61–75.

Torebjörk HE, Wahren LK, Wallin BG, et al. (1995) Noradrenaline-evoked pain in neuralgia. *Pain* 63: 11–20.

Tracey DJ, Cunningham JE, and Romm MA (1995) *Pain* 60: 317–327.

Vierck CJ Jr. (2006) Mechanisms underlying development of spatially distributed chronic pain (fibromyalgia). *Pain* 12: 242–263.

Wall PD, Devor M, Inbal R, et al. (1979) *Pain* 7: 103–113.

Wall PD, Scadding JW, and Tomkiewicz MM (1979) *Pain* 6: 175–182.

Willenbring S, Beauprie IG, and DeLeo JA (1995) *Anesthesia & Analgesia* 81: 544–548.

Pain Management

A R Molloy and M J Cousins, University of Sydney, Sydney, NSW, Australia

Crown Copyright © 2009 Published by Elsevier Ltd. All rights reserved.

Introduction

Advances in neuroscience and psychology over the past 30 years have revolutionized our understanding of the reasons for the variability and complexity of pain and have provided new evidence-based treatment options. Pain is now viewed as a complex perception associated with an array of interactions at the level of the peripheral and central nervous systems, involving sensory, emotional, cognitive, and environmental factors. Instead of treating pain report in isolation, there is a range of treatment options available to the multidisciplinary pain management team, targeting relevant outcomes such as function, mood, quality of life, acceptance of pain, and optimal adjustment to persisting pain. A number of key events in the evolution of pain management as a specialty should be noted. The establishment of the first multidisciplinary pain clinic in Seattle at The University of Washington in 1960 by Professor John Bonica led to widespread recognition of this approach. The publication of the 'gate control' theory of pain in 1965 by Melzack and Wall was significant for two reasons: it recognized the role of the spinal cord in modulation of signals from both nociceptors in the periphery and descending pathways and highlighted the potential role of psychological factors in activating descending pain-inhibitory pathways. Since the establishment of the International Association for the Study of Pain in 1974, this organization has attracted over 6000 members from a range of disciplines in more than 80 countries. Chronic pain has been described as the 'hidden epidemic' affecting one in five adults, up to 80% of patients with advanced cancer, and in AIDS the estimated prevalence is 30–80%. Based on the wealth of scientific research available, it has been proposed that persistent pain is a disease entity and warrants the same health focus as other major diseases such as diabetes and cancer. Linked to this is the call to recognize evidence-based treatment of pain as a basic human right, as studies consistently show that pain is often undertreated even when resources are available.

Acute Pain

'Everyday' acute pain is associated with stimulation of nociceptors, which are peripheral sensory receptors on Aδ and C-fibers. Other Aδ and C-fibers, called 'silent nociceptors,' are only activated by noxious stimuli in the presence of tissue damage. Nociceptors convert a range of noxious environmental stimuli (thermal, mechanical, and chemical) into action potentials that reflect the intensity and duration of the insult. The effect on nociceptors is mediated through a wide range of inflammatory mediators, including ATP, bradykinin, prostinoids, serotonin, histamine, and hydrogen ions (tissue acidosis). Nociception is not pain, as pain can only be experienced by the conscious brain. Diagnosis is usually straightforward in acute pain, as there is usually broad concordance between symptoms, physical signs, and pathology; however, there is no direct relationship between the amount of tissue injury and pain. Acute pain is typically dull and aching and anatomically related to the area of pathology. Acute pain usually resolves in a predictable period once the noxious stimulus is removed and normal healing progresses. Multimodal analgesic and interventional techniques which target different mechanisms contributing to the pain are recommended for moderate to severe pain to reduce side effects associated with individual drugs, in particular opioids. Options include drugs such as nonsteroidal anti-inflammatory drugs (NSAIDs), acetaminophen, ketamine, gabapentin, wound infiltration, regional local anesthetic blocks and spinal opioid administration. Uncontrolled acute pain should rarely be a clinical problem except in special cases, such as the patient with previous trauma to the tissues (e.g., surgery or injury), incomplete healing, persistent inflammation, or nerve injury (either peripheral or central), and those on large doses of opioid medication, or if there is a history of substance abuse. In these cases factors such as peripheral and central sensitization and tolerance may either singly or in tandem result in the heightened experience of pain, and nonstandard analgesic regimes such as a subcutaneous or systemic infusion of ketamine and lignocaine (singly or in combination) may be required.

Chronic or Persistent Pain

Chronic or persistent pain refers to the pain that a significant proportion of patients will continue to experience despite appropriate medical or surgical assessment and intervention. For epidemiological purposes a period of 3 months is taken, but in practical terms chronic pain may be present as early as 2 weeks. Epidemiological studies in different countries are remarkably consistent. Approximately 20% of the population will experience persisting pain, with about half of those patients distressed and disabled.

There are three main categories of factors contributing to chronic pain: nociceptive, neuropathic, and psychological/environmental. Nociceptive pain refers to the pain associated with ongoing tissue damage that can last long after the initial noxious stimulus. Components of this type of pain are involved in the pain of osteoarthritis, rheumatoid arthritis, unhealed fractures, and chronic infections.

A different clinical presentation is associated with neuropathic pain, which is due to damage, disease, or section of the peripheral or central nervous system. Pain descriptors include a range of dysesthetic sensations such as burning, shooting, or electric shocks. The topographical distribution of neuropathic pain tends to be regional rather than dermatomal. For example, in a variant of neuropathic pain, complex regional pain syndrome (CRPS), the dysesthetic pain may occupy the whole of the upper limb following a trivial injury to the hand. Common associated phenomena are allodynia (pain experienced after a nonpainful stimulus – e.g., touching sunburned skin), hyperalgesia (severe pain experienced after a mildly painful stimulus), abnormal motor responses, and disturbance of sympathetic activity such as sweating, swelling, and color change. In the case of persistent low back pain, there may be neuropathic, nociceptive, and psychosocial components contributing to the pain experience. In many cases there will be no evidence of nerve or tissue damage – for example, in CRPS type 1, where ongoing pain is considered to be due to dysfunction of part of the nervous system.

Neuropathic components of pain may be classified according to the etiology of the pain, the anatomical site, or a mechanism-based classification. Such systems have been of variable utility clinically, as it is likely that multiple pain mechanisms are operating simultaneously. Also, as plasticity is an inherent feature of animal and human sensory pathways, neuronal systems involved in the experience of pain may alter over time.

The biomedical model is generally inadequate in cases of persisting pain. The biopsychosocial model proposed by Engel has been widely accepted, as it provides a better analysis of all of the factors contributing to persistent pain. Figure 1 shows the typical factors that inadvertently cause an increase in suffering.

The biopsychosocial model requires that the bio- (medical), psycho- (psychological), and social (environmental) factors are assessed so that all treatable

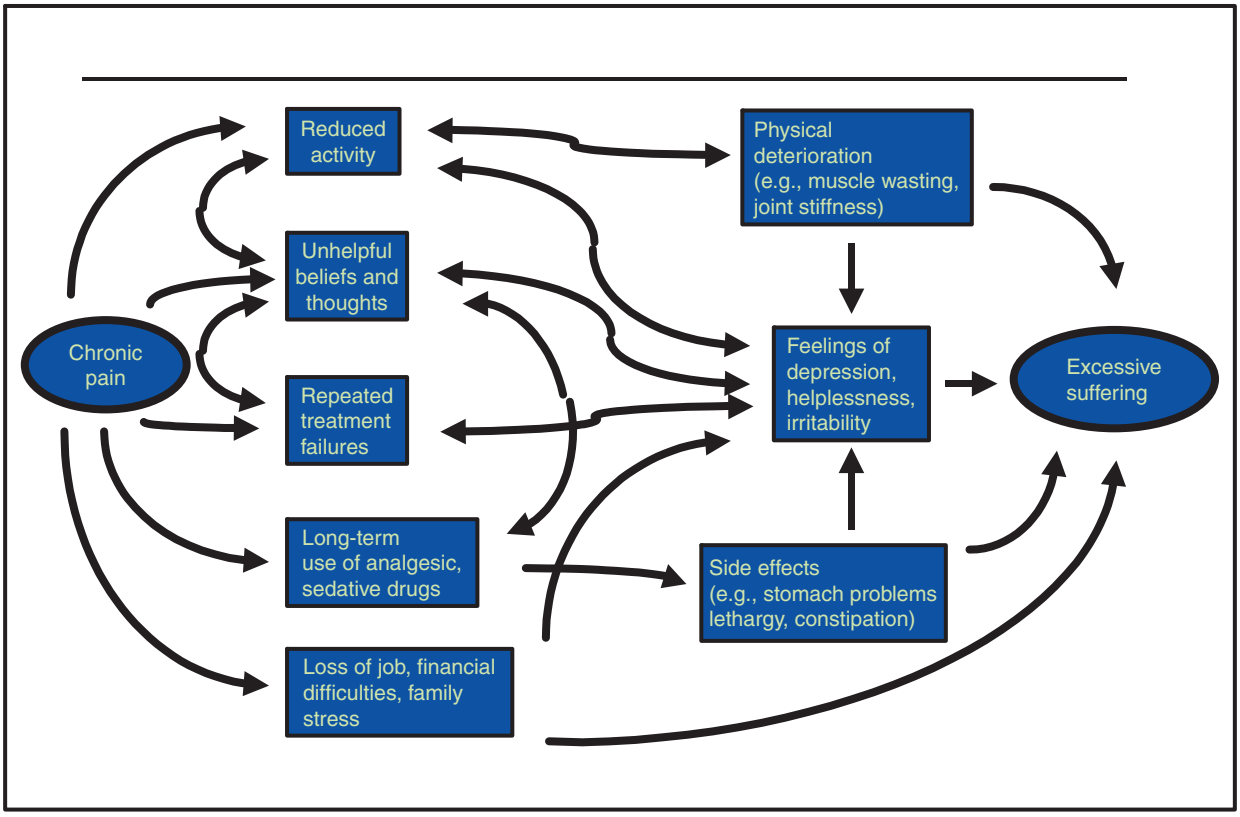

Figure 1 When chronic pain becomes a problem. Reproduced with permission from MK Nicholas, Pain Management and Research Centre, Royal North Shore Hospital, St. Leonards, NSW, Australia.

factors are identified. A diagnostic formulation and treatment plan are then arrived at after a 'round table' interdisciplinary meeting. A diagnosis may result from the medical assessment, but a formulation looks at the impact of the patient's pain and possible modulating factors. The importance of a formulation lies in the fact that a patient may have the same diagnosis, but the impact in terms of mood, function, and quality of life may be very different.

An important concept is that persistent, physical pain changes the nervous system, and adverse psychological/environmental adaptations become self-sustaining; this process represents a disease entity. Patients with persisting pain should be referred to a pain center as early as possible. Figure 2 is a suggested algorithm.

Cancer Pain

Pain is experienced by approximately 60–80% of patients with advanced cancer. Nociceptive and neuropathic components of pain occur either singly or in combination, as in chronic pain. However, the underlying disease process is usually progressive, resulting in the need for regular reviews and adjustments of the analgesic regime. Analgesic titration is further compounded by drug interactions with other therapeutic agents and by symptoms such as fatigue, constipation, dyspnea, and impaired cognitions. By tailoring multimodal analgesic drug therapy to each patient's needs, at least 85% of patients should experience effective pain relief. Other options include chemotherapy and radiotherapy, which are also associated with pain relief by reducing tumor bulk. Biphosphonates produce morphological change and apoptosis of osteoclasts which are responsible for osteoclast-induced skeletal metastases. Pinning or surgical treatment of fractures may be required, as up to 30% of patients with bony metastases develop long bone fractures.

Breakthrough pain is a heterogeneous group of complaints related to spontaneous neuropathic components, incident pain associated with movement, and pain secondary to fluctuations in analgesic levels. Distinction between these causes is important, as treatment will be different. For incident pains such as associated with dressing changes, a preemptive dose of an opioid, such as immediate-relief morphine, should be taken.

Approximately 10% of patients do not experience adequate pain relief. Therapeutic options must take account of prognosis, location of pain, type of pain, coexistent disease, and psychosocial factors. Traditionally neurolytic blocks were favored but increasingly spinal drug administration is used, which may vary from a percutaneous intrathecal or epidural catheter to an implantable programmable pump. Outcome studies have shown that intrathecal morphine administration is associated with improved pain control, reduced side effects, less use of adjuvant pain medications, and improved survival in a cancer pain population, as compared to standard medical management.

Pharmacological Treatments

The application of evidence-based medicine, the introduction of novel analgesic agents, and the publication of a range of guidelines (e.g., Oxford League Tables) have resulted in a more rational approach to prescribing. A commonly used tool to describe analgesic efficacy is numbers needed to treat (NNT). This is defined as the number of patients who need to receive an analgesic, for one patient to report a predetermined level of pain relief (usually 50%) more than the placebo group. While being a useful tool, few studies provide a cohort of the more than 500 patients needed to calculate a credible NNT. An analgesic with high efficacy has an NNT of 2–3. The flip side is that these measures are not only drug, dose, and context specific but also do not reflect minor and major adverse events (expressed as numbers needed to harm (NNH)).

Acetaminophen

Acetaminophen is one of the most commonly used analgesics worldwide. This drug is available as an oral, rectal, and intravenous preparation. Its mode of action remains undefined. Hypothesis suggest an action on a number of systems, including inhibition of central cyclooxygenases COX-2 and COX-3, modulation of descending serotoninergic pathways, action on the *N*-methyl-D-aspartate (NMDA) receptor, or an action related to nitric oxide. The evidence for the clinical efficacy of acetaminophen is good and it is usually recommended as a first-line analgesic in mild to moderate acute and cancer pain, and as part of a multimodal approach in severe pain. It is regularly combined with other analgesics, such as codeine. It is well tolerated and side effects are rare, except for liver failure associated with an overdose.

NSAIDs

The introduction of a number of selective COX-2 inhibitors (coxibs), in addition to aspirin and the traditional nonselective (COX-1 and COX-2) NSAIDs, has provided a range of new treatment options but has also highlighted a number of concerns over this drug class. All NSAIDs have analgesic, anti-inflammatory,

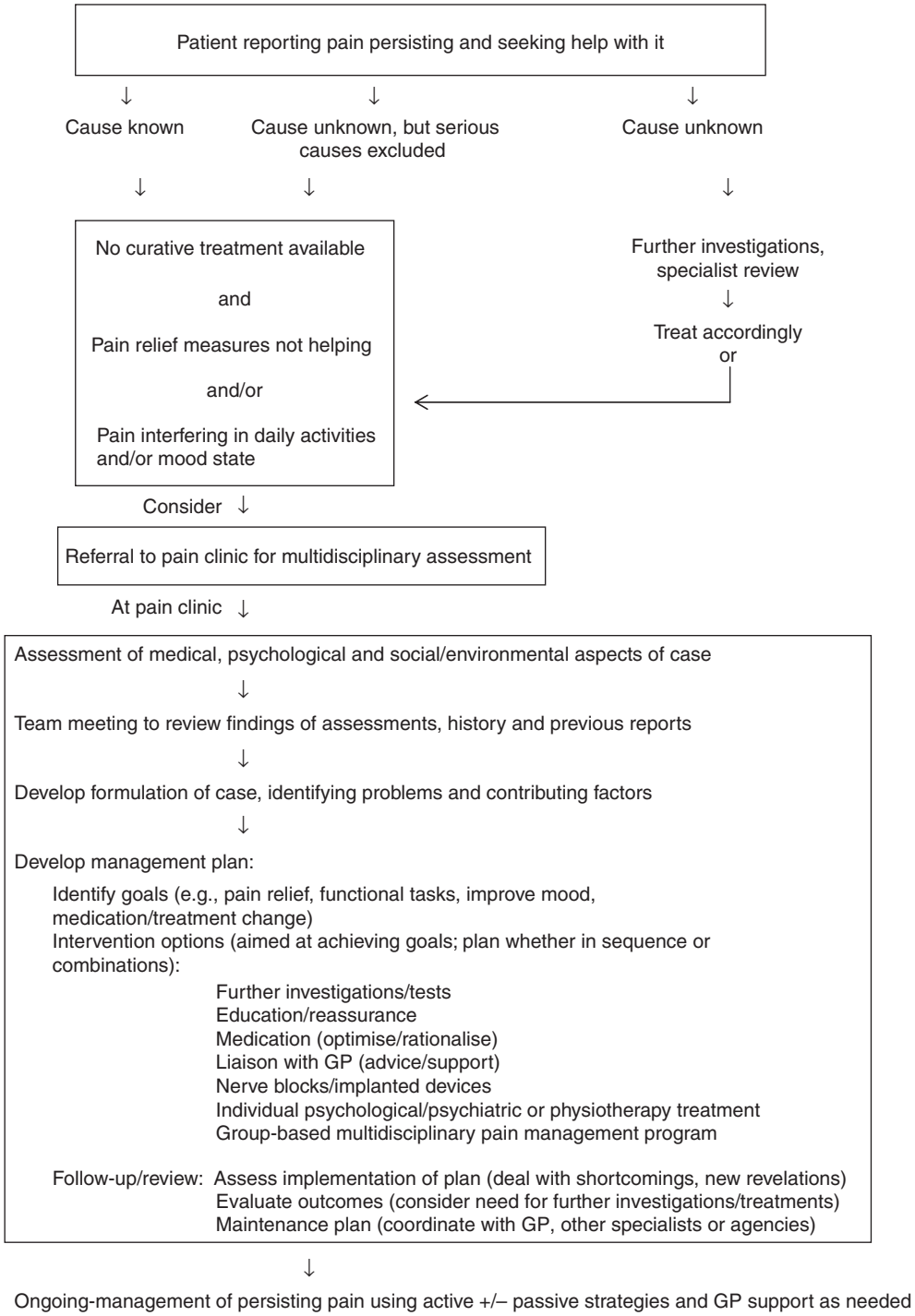

Figure 2 Algorithm: who and when to refer to a pain clinic. Adapted from Nicholas MK (2004) When to refer to a pain clinic. *Best Practice & Research Clinical Rheumatology* 18(4): 613–629.

and antipyretic actions which result from their action in reducing prostaglandin concentrations both peripherally and centrally. The mechanism of action also explains the range of adverse events witnessed clinically. The principal COX-1 side effects are gastrointestinal (GI) ulceration and bleeding (COX-1 is involved in the production of prostaglandin E_2 and prostacyclin, which have a protective action on the gastric mucosa) and a platelet-inhibiting effect (COX-1 has a role in the production of thromboxane A). The latter

'side effect' is used therapeutically to reduce platelet aggregation and the risk of stroke and myocardial infarction.

Coxibs have the same clinical efficacy as do traditional NSAIDs but are associated with significantly reduced risk of GI complications, although the same renal precautions apply. While they do not interfere with the action of thromboxane A in platelets, the inhibition of prostacyclin in the endothelium may explain the high incidence of thrombogenic adverse events observed with rofecoxib that led to the withdrawal of this drug worldwide. This raised the question as to whether this was a class effect of coxibs. Rofecoxib (methyl sulfone) has a chemical structure different from that of the other coxibs (arysulfonamides), which may explain the clinical differences. Coxibs are still represented in the market by celecoxib (oral preparation) and parecoxib (intravenous administration).

Antidepressants, Systemic Local Anesthetics, and Anticonvulsants

The increasing evidence base showing improved pain control with tolerable side effects has resulted in the antidepressant, systemic local anesthetic, and anticonvulsant groups of drugs being accepted as specific analgesics for neuropathic pain. Most studies have been in painful diabetic neuropathy (PDN) and postherpetic neuralgia (PHN), as these are common neuropathic pain conditions. Use of these drug classes for other neuropathic pain syndromes is on the presumption that similar neurobiological processes are involved, although it is almost certain that a number of mechanisms underlie the majority of neuropathic pain syndromes and that more than one drug may be required. At present there are no widely accepted drug administration algorithms for neuropathic pain and few studies comparing one drug with another or the outcomes associated with multiple drug use.

While almost all studies with tricyclic antidepressants (TCAs) have shown that their use is associated with pain relief, their clinical use is limited by the frequency of associated side effects, which includes dry mouth, constipation, blurred vision, urinary hesitancy, and postural hypotension. This has led to studies on the newer antidepressants, which have a more favorable side effect profile. These include the selective serotonin reuptake inhibitors (SSRIs) and those that inhibit the uptake of both serotonin and noradrenaline (SNRIs). To date the results have not been clear-cut but more evidence is accumulating to support their use.

Anticonvulsants and systemic local anesthetics are drugs which have the ability to suppress ectopic neuronal activity. Initially their use was an extension of their established role in epilepsy but now they are considered as specific analgesics for chronic pain; also more recently a role in acute pain has been identified. Their main action is to differentially suppress spontaneous ectopic neuronal activity at drug concentrations that do not block normal neuronal conduction (approximately 3 times less). In broad terms this effect may be achieved either by drugs such as carbamazepine, lamotrigine, mexiletine, and lignocaine, which block sodium channels, or gabapentin and pregabalin, which also bind to the $\alpha_2\delta$ subunit of the voltage-gated calcium channel. A variety of side effects have been reported, ranging from lightheadedness and dizziness to ataxia, severe skin rashes (lamotrigine), and arrhythmias (mexiletine).

Opioids

Opioids are the mainstay of management of moderate to severe acute pain. In most cases the opioid acts through the μ receptor, which is similar among species, permitting extrapolation from animal studies. Attempts to clinically target the δ receptor, κ receptor, and opioid-like receptor (ORL1) have been largely unsuccessful.

Although morphine has been the traditional first-line drug, there is now a range of opioids available with differing pharmacokinetics and pharmacodynamics. These include fentanyl, codeine, meperidine, oxycodone, hydromorphone, buprenorphine, and methadone. Fentanyl is a synthetic opioid which is up to 100 times more potent than morphine and has rapid onset and offset due to high lipid solubility. Codeine may be administered orally or intramuscularly or in tablet form combined with paracetamol or an NSAID. Meperidine is a synthetic opioid that has a short half-life and is being used less due to problems with addiction and the risk of convulsions with normeperidine, the major metabolite. Oxycodone and hydromorphone are useful alternatives to morphine. Hydromorphone has approximately seven times the potency when given spinally. Methadone is a synthetic opioid with a long and variable half-life. It has been used increasingly in chronic pain management due to an action on the NMDA receptor.

The usual routes of administration are oral or parenteral (subcutaneous, intramuscular, or intravenous). As there is a wide variation in opioid requirements, patient-controlled analgesia (PCA) devices are often used postoperatively to allow patients to titrate opioid doses to their needs. Increasingly the benefits of epidural and intrathecal opioid administration are being recognized.

Side effects of opioids include respiratory depression, nausea, vomiting, endocrine abnormalities, and bowel dysfunction, in particular constipation.

Ziconotide

This peptide is a selective for the N-type calcium channel and is approved by the US Food and Drug Administration (FDA) for intractable pain in those for whom intrathecal infusion is appropriate and who have failed to respond to more conservative measures and intrathecal morphine. This drug has a narrow therapeutic index with side effects such as dizziness, nystagmus, confusion, gait disturbance, and urinary retention.

Ketamine

Glutamate acting through the NMDA receptor in the dorsal horn is associated with amplification and prolongation of nociceptive input, termed 'wind-up.' Central hypersensitivity that results is implicated in persistent pain states. NMDA receptor antagonists, including ketamine, dextrometorphan, and memantine, have been successful in treating clinical pain conditions with significant components of hyperalgesia and allodynia. Ketamine is the most widely used drug in this group and has been used orally, subcutaneously, and intrathecally. Because of actions at other NMDA receptors, particularly in the CNS, side effects are common and include somnolence and dysphoria. Although it has been used intrathecally, concerns have been raised about neurotoxicity.

α_2 Agonists

Clinical efficacy has been demonstrated for α_2 agonists given spinally in nociceptive and neuropathic pain. Although used orally and topically, clonidine is most commonly administered intrathecally, usually in combination with an opioid. Side effects include sedation and hypotension.

Methods of Drug Delivery

Studies have shown that the delivery of both established and new drugs by modern drug delivery devices and novel routes of administration are often associated with improved analgesic outcomes. In a particular patient, contemporary multimodal acute pain management may involve many routes, such as oral, intravenous, and intraspinal drug administration. Although in persistent pain management there is a greater focus on contributing psychosocial aspects, it is important to minimize side effects associated with analgesic therapy so that function can be promoted. Oral administration is recommended if bowel function is normal, or parenteral (subcutaneous, intramuscular, or intravenous) if it is impaired, such as postoperatively. Other routes are generally indicated to shorten the onset of action (e.g., transmucosal or nasal fentanyl), to address concurrent medical problems such as fentanyl patches in short bowel syndrome, to target specific sites (e.g., topical lignocaine patches in postherpetic neuralgia), or to access specific analgesic receptors, such as peripheral or spinal opioid receptors. Low-dose buprenorphine patches, changed every 7 days, provide long-duration, well-tolerated analgesia, particularly in the older age group.

Invasive Treatments

Peripheral Nerve Stimulation

The technique of peripheral nerve stimulation, which is thought to activate large, low-threshold afferent fibers, may be considered when persistent pain is well localized and lead placement in the region of the pain is technically feasible. Drawbacks are that the leads may displace or break. An increasingly common indication is headache due to occipital neuralgia, postherniorraphy nerve pain, and other types of postoperative neuropathic pain. This technique should be avoided where there is a significant central component to the pain.

Spinal Cord Stimulation

Highly sophisticated systems are now available to electrically stimulate parts of the spinal cord and brain. The proposed mechanism of action of spinal cord stimulation (SCS) is a combined activation of descending inhibitory pathways and release of a cascade of neuroactive substances in the region of stimulation, particularly γ-aminobutyric acid (GABA). Nociceptive pain is poorly responsive to SCS, but conditions such as neuropathic pain, ischemic limb, cardiac pain, and CRPS are accepted indications. There is a significant incidence of complications, in particular infection, lead breakage, and electrode movement.

Deep Brain and Motor Cortex Stimulation

Deep brain stimulation and motor cortex stimulation have been used for pain that has been unresponsive to all other therapeutic modalities. Deep brain stimulation usually involves stimulation of the sensory thalamic nuclei, which target neuropathic pain, and the periacqueductal gray, which is effective for certain types of nociceptive pain. Motor cortex stimulation shows promise for severe 'central pain.'

Intraspinal Drug Delivery

Modification of spinal function by epidural and intrathecal drug administration depends on administration of either local anesthetics which block neuronal sodium channels in spinal nerve roots or drugs which

act at a spinal cord dorsal horn level on other receptor systems (e.g., opioid receptors). Local anesthetics have a rapid onset, as they block nerve roots which pass through the epidural or intrathecal space. Drugs targeting spinal cord μ or GABA sites have a slower onset, as they need to target receptors on primary afferent endings or at supraspinal sites. Epidurals are widely used for the management of labor pain and postoperative pain management, particularly in patients with potential respiratory compromise. Larger volumes of local anesthetic are required to achieve sensory blockade as compared to intrathecal analgesia.

Spinal drug administration also has an established role in both chronic noncancer and cancer pain. Up to 10% of patients experience severe cancer-related pain despite optimal medical management. Many patients with nonmalignant pain also fail to report benefit with oral or parenteral medication or experience unacceptable side effects. Intraspinal drug delivery should be considered in these patients, as a range of spinal and supraspinal receptors can be targeted which may be associated with improved analgesia and a reduction in side effects associated with systemic administration. Delivery may be epidural, intrathecal, or intracerebroventricular through either an externalized system or a fully implanted system. Currently a range of battery-powered or gas-driven implantable infusion systems are available and also a complementary hand-held device with the facility to deliver bolus doses as needed. Equivalent analgesia may be obtained with epidural administration at one-tenth of the parenteral morphine dose and intrathecally at one-tenth of the epidural dose. Intraventricular, compared to intrathecal, administration requires smaller doses. Although currently at least 16 drugs have been used intrathecally, only morphine sulfate, ziconitide, and baclofen have been approved by the US FDA for intrathecal use. There is concern over the long-term safety of many of the other compounds and drug mixtures used. Spinal drug administration may be associated with a different spectrum of side effects. For example, intrathecal morphine may be associated with the growth of catheter tip inflammatory masses. Outcome studies have shown that intrathecal morphine administration is associated with improved pain control, reduced side effects, less use of adjuvant pain medications, and improved survival in a cancer pain population as compared to standard medical management. This approach can also be of benefit in carefully selected patients with severe and disabling chronic nonmalignant pain. Combination of these approaches with an intensive cognitive behavioral pain program may produce a better outcome than either treatment alone.

Neurodestructive

Although neurodestructive techniques are used in certain chronic noncancer conditions such as trigeminal neuralgia and spinal medial branch rhizotomy, they are to be avoided in most other cases due to the high incidence of deafferentation pain. Neurolytic procedures using phenol or alcohol, which permanently interfere with the function of nerves, are usually reserved for well-localized cancer-related pain, in patients with a short life expectancy. Common clinical examples include celiac plexus block for pancreatic cancer, and subarachnoid neurolytic blocks for pelvic pain and for unilateral chest wall pain associated with mesothelioma.

Radiofrequency

Radiofrequency techniques have been used for over 25 years to interrupt nerve transmission at various anatomical points. Most commonly the medial branch of the posterior primary division supplying the affected spinal segment (sacral, lumbar, thoracic, or cervical) is targeted after selection on the basis of concordant relief, with injection of one or more different local anesthetics or saline onto the medial branch or into the facet joint. Up to 80% of patients experience initial relief, with a variable period of sustained benefit from weeks to over 1 year. The dorsal root ganglion, the cervical spinothalamic tract in the treatment of cancer pain (known as a cordotomy), and the lumbar sympathetic chain are also potential targets for this technique. The role for these techniques is not clear, but they are relatively simple and well tolerated.

Treatment of Psychological and Environmental Factors

Psychosocial Assessment and Management of Pain

Psychosocial assessment, which includes assessment of psychological and environmental factors, is widely accepted as an essential part of the evaluation of persisting pain conditions. The assessment aims to identify the main presenting problems and the contributing somatic, psychosocial, and demographic factors. A formulation of relevant domains is then presented at a multidisciplinary conference and targets are identified for intervention. Typically a series of validated self-report questionnaires covering the relevant domains such as mood, function, coping strategies, pain behaviors, self-efficacy, and tendency to catastrophize are also completed by the patient. These can also be completed after treatment and should give a reliable indication of change and act as a check on patients' verbal reports.

Interventions can be delivered either serially (over weeks or months), as part of an intensive pain program, or as a component of a coordinated management plan (e.g., spinal cord stimulation and pain program). As these treatments are aimed at helping the patient to live with their pain, it is recommended that all medical and surgical treatments targeting pain relief are completed before commencing the program. Most evidence is for multidisciplinary treatments that are delivered according to cognitive behavioral principles, whereby patients learn a set of active pain management skills, which replace other unhelpful strategies. Treatments have to be tailored to the individuals needs and also integrated with the workplace if return to work is a goal. In general terms there is a dose-related response such that the more disabled patients do better with more intensive treatment. Significant improvements in pain, mood, function, and quality of life have been shown. Inevitably some patients do not benefit from this approach. Possible reasons are inadvertent reinforcement of passivity, lack of acceptance of persisting pain, poor patient compliance, and conflicting treatment advice. Such patients may seek further treatments targeting pain reduction after a pain program, and the risk is that further passive approaches will be introduced, reinforcing disability. In such cases individual sessions with a clinical psychologist may be appropriate.

See also: Deep Brain Stimulation; Neuropathic Pain; Neuropeptides: Pain; Nociceptor Responses; Pain Pathways: Descending Modulation; Pain and Genes; Pain: Central; Pain: Neuroimaging; Spinal Cord Pain Systems; Synaptic Transmission: Models.

Further Reading

Australian and New Zealand College of Anaesthetists and Faculty of Pain, Medicine (2005) *Acute Pain Management: Scientific Evidence,* 2nd edn. Melbourne: ANZCA.

Cousins MJ, Brennan F, and Carr DB (2004) Pain relief: A universal human right. *Pain* 112(1–2): 1–4.

Cousins MJ and Bridenbaugh PO (eds.) (1998) *Neural Blockade in Clinical Anesthesia and Management of Pain,* 3rd edn. Philadelphia: Lippincott-Raven.

Dolin SJ, Cashman JN, and Bland JM (2002) Effectiveness of acute postoperative pain management. Evidence from published data. *British Journal of Anaesthesia* 89: 409–423.

Flor H, Kalso E, and Dostrovsky J (eds.) (2006) *Proceedings of the 11th World Congress on Pain.* Seattle: IASP Press.

Hassenbusch SJ, Portenoy RK, Cousins M, et al. (2004) Polyanalgesic Consensus Conference 2003: An update on the management of pain by intraspinal drug delivery – Report of an expert panel. *Journal of Pain Symptom Management* 27(6): 540–563.

Loeser JD (ed.) (2001) *Bonica's Management of Pain,* 3rd edn. Philadelphia: Lippincott, Williams & Wilkins.

McMahon S and Koltzenberg M (eds.) (2005) *Wall and Melzack's Textbook of Pain,* 5th edn. London: Elsevier Science.

Moore A, Edwards J, Barden J, et al. (eds.) (2003) *Bandolier's Little Book of Pain.* Oxford: Oxford University Press.

Morley S, Eccleston C, and Williams A (1999) Systematic review and meta-analysis of randomised controlled trials of cognitive behaviour therapy and behaviour therapy for chronic pain in adults, excluding headache. *Pain* 80: 1–13.

Nicholas M, Molloy A, Tonkin L, et al. (2001) *Manage Your Pain.* Sydney: ABC Books.

Nicholas MK (2004) When to refer to a pain clinic. *Best Practice & Research Clinical Rheumatology* 18(4): 613–629.

Siddall PJ and Cousins MJ (2004) Persistent pain as a disease entity: Implications for clinical management. *Anesthesia & Analgesia* 99(2): 510–520.

Relevant Websites

http://www.anzca.edu.au – Australian and New Zealand College of Anaestheticists: Acute Pain Management: Scientific Evidence.

http://www.iasp-pain.org – International Association for the Study of Pain.

http://www.kovacs.org – Kovacs Foundation, European guidelines for prevention of low back pain.

http://www.kovacs.org – Kovacs Foundation, European guidelines for the management of chronic nonspecific low back pain.

http://www.mskcc.org – Memorial Sloan-Kettering Cancer Center.

http://www.pmri.med.usyd.edu.au – Pain Management Research Institute, Royal North Shore Hospital, University of Sydney.

Pain Pathways: Descending Modulation

M H Ossipov, University of Arizona College of Medicine, Tucson, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In spite of the tremendous progress made over the course of human history in the understanding of the sensorium, the concept of pain as a unique and separate sensation has only been appreciated within the past half century. For the most part, it was generally thought that pain did not arise from specific nociceptors but that the temporal patterns of sensory inputs from sensory fibers led to the interpretation of the inputs as being painful. Although a specific pattern had been suggested, its acceptance was hampered by the fact that the correlation between stimulus intensity and pain perception is quite variable. Moreover, individuals with an injury or inflammation report pain arising from normally nonnoxious stimuli (allodynia) or as being much more intense than is normally experienced by a given noxious stimulus (hyperalgesia). A veritable explosion of studies aimed at understanding the mechanisms of pain over the past four decades has led to our understanding of not only specific pain receptors (nociceptors) and pain pathways but of pain regulatory systems in the central nervous system (CNS) that serve to both inhibit and augment pain perception. A key moment in the history of pain research was the 1965 publication of the 'gate control theory' by Melzack and Wall. This theory is considered to be the first attempt to provide a coherent mechanism describing an endogenous mechanism through which the CNS can modulate pain. Further studies clearly demonstrated the existence of brain regions that can be activated to produce a complete blockade of responses to noxious inputs, providing compelling evidence of an endogenous pain-control system. This concept was bolstered by the discovery of selective stereospecific receptors for opioids and by the discovery of endogenous peptidic ligands for these receptors. These discoveries spurred an intense and growing field of neuroscience dedicated to the investigation of pain-control mechanisms. We now understand not only that the brain receives nociceptive inputs from the periphery, which are interpreted at cortical sites as pain but that there also exist complex interactions among several regions of the brain that result in activation of descending modulatory systems that both inhibit and facilitate pain. The balance among these opposing modulatory systems can produce either analgesia or an enhanced pain state.

Ascending Noxious Inputs to the Brain Yield Pain

It has been well appreciated over the past 100 years that noxious inputs from the periphery enter the CNS through synaptic contacts between the peripheral nerves and second-order neurons of several ascending projections. One of the most prominently considered tracts is the spinothalamic tract (Figure 1), which transmits sensory inputs from the spinal cord to the distinct thalamic nuclei, including the posterior part of the ventromedial nucleus, ventroposterior nuclei, and ventrolateral nucleus. The spinothalamic tract consists predominantly of projections from lamina I of the spinal dorsal horn; it also includes projections from laminae IV to V and from laminae VII to VIII. The parabrachial nucleus receives inputs from laminae I and laminae V through the parabrachial tract and transmits these inputs to more rostral sites, including the amygdala, hypothalamus, and thalamic nuclei. Nociceptive inputs from the spinal dorsal horn are also transmitted to the noradrenergic pontine nuclei, which include the locus coeruleus (A6), the subcoeruleus and Kölliker–Füse nuclei (A7). These regions modulate cardiovascular, respiratory, and autonomic functions in addition to nociception and may participate in the integration of pain with contextual and situational cues. The ascending pain pathways also project to the reticular formation, including the raphe nuclei, and to the ventrolateral periaqueductal gray (PAG). Ultimately, the ascending nociceptive sensory inputs are integrated with nonnociceptive sensory inputs from other regions of the CNS and are projected to cortical regions, most notably the somatosensory cortex and the anterior cingulate cortex (ACC), where nociception is interpreted as pain, including its emotional context. These regions give rise to caudally projecting pathways to medullary and spinal sites, through which further noxious inputs may be either inhibited or facilitated, and thus serve as a complex endogenous pain modulatory system.

Descending Inhibition of Pain

The Periaqueductal Gray

The PAG is one of the first brain structures that was recognized as playing an important role in pain and analgesia. Tsou and Jang, working in Shanghai in 1964, reasoned that because morphine blocks pain at doses that do not affect other sensory modalities, it probably worked through a site specific for pain control. They either applied morphine directly onto the surface of the brain, injected it into the cerebral

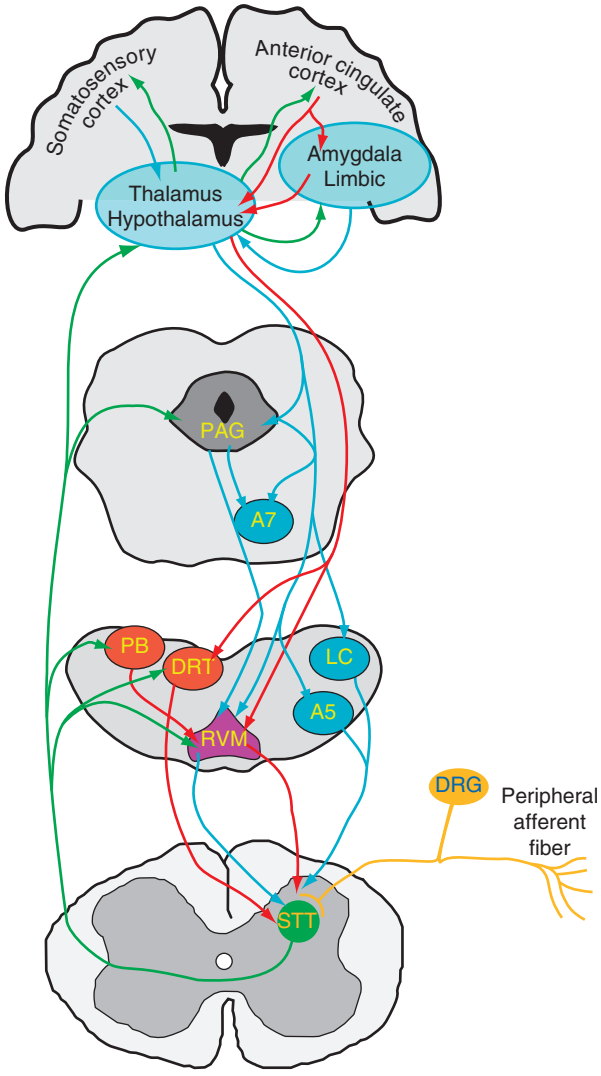

Figure 1 Schematic representation of pain modulatory circuitry. Ascending sensory pathways are shown in green, descending pain inhibitory pathways are shown in blue, and descending pain facilitatory pathways are shown in red. Noxious stimuli applied to the free nerve endings (nociceptors) of peripheral afferent fibers are transmitted to the central nervous system through synaptic transmission with second-order neurons in the dorsal horn of the spinal cord. The ascending fibers of the spinothalamic tract (STT) neurons transmit sensory inputs to the thalamus, which relays the sensory signals to nuclei of limbic structures, including the amygdala and to cortical sites, where the pain experience is interpreted in the context of environmental and other cues. Collateral projections from the ascending tracts innervate the parabrachial region (PB), the mesencephalic reticular formation, and the periaqueductal gray (PAG) as well. The ascending nociceptive inputs activate a cascade of descending pain inhibitory (blue) and facilitatory (red) circuits that modulate further noxious inputs. The somatosensory cortex, thalamus, and limbic regions communicate with one another, and with the PAG, which is a predominant pain-inhibitory locus with reciprocal communications with the rostroventromedial medulla (RVM). Inhibitory signals are also elicited from noradrenergic nuclei, including the A7, A5, and locus coeruleus (LC), which may project directly to the spinal cord or through the RVM. Descending pain facilitatory signals may arise from the anterior cingulate cortex (ACC), which communicates with limbic structures as well as the RVM. The parabrachial region (PB) and the dorsal reticular nucleus (DRT) also activate pain-facilitatory signals, either directly to the spinal dorsal horn or via synaptic transmission in the RVM. The RVM acts as a final relay for the integration of descending pain modulatory systems. It receives inputs from rostral structures and possesses both pronociceptive facilitatory and antinociceptive inhibitory projections to the spinal dorsal horn. The predominant net effect either inhibits further noxious inputs, thus serving as an endogenous antinociceptive system, or facilitates nociceptive inputs, thus acting as a pronociceptive modulatory system that promotes states of enhanced pain. DRG, dorsal root ganglion.

ventricles, or microinjected it into discrete brain regions and discovered that it produced a powerful antinociceptive effect when microinjected into the PAG. This discovery led to a rapid succession of mapping experiments in which opioids were microinjected or electrical stimulation was applied to discrete loci about the PAG, identifying the ventrolateral aspect of this region as the pain modulatory site. In one study, electrical stimulation of the PAG produced antinociception of such intensity that it was possible to perform a laporotomy on a conscious rat without producing overt signs of behavioral distress. Electrical stimulation of the PAG was rapidly adapted to relieve intractable cancer pain in humans, and it is commonly cited as one of the most rapid applications of an experimental finding to clinical application. Although PAG stimulation has been largely discontinued because of other side effects, such as anxiety and distress, deep brain stimulation aimed at other regions remains an important means of controlling otherwise intractable pain. Significantly, antinociception produced by electrical stimulation of the PAG or microinjection of morphine produced antinociception that lasted much longer than the stimulus application, suggesting that an endogenous analgesic system was being activated. Analgesia produced by electrical stimulation of the PAG is blocked by opioid antagonists, showing that an endogenous opioid system mediates this antinociceptive effect. Further mapping experiments demonstrated that there is considerable overlap between PAG sites that produce antinociception in response to electrical stimulation and those that are responsive to morphine microinjection. It is now known that several neurotransmitter systems are capable of activating pain modulatory circuitry from the PAG. Activation of the CB1 cannabinoid receptors in the PAG is believed to underlie opioid-independent stress-induced analgesia. Other studies demonstrated that agonists acting at metabotropic glutaminergic receptors; *N*-methyl-D-aspartate (NMDA) receptors; and receptors for neurotensin, substance P, calcitonin gene-related peptide

(CGRP), or neuropeptide Y also produce varying degrees of antinociception when microinjected into the PAG. The PAG receives inputs from central sites that receive nociceptive projections from the periphery, including the thalamus, hypothalamus, parabrachial regions, cortical regions, and the limbic system (Figure 1), and it may be a site where nociceptive signals are integrated with autonomic and stress-mediated responses. In spite of its prominent role in the modulation of pain, the PAG has few direct projections to the spinal cord. Rather, the PAG possesses reciprocal connections with medullary sites that project to the spinal cord. It is suggested that these caudal projections from the PAG are diffuse, rather than discrete, because a single midline injection of local anesthetic into the rostral medulla failed to abolish antinociception from electrical stimulation of the PAG, whereas stimulation-produced antinociception was blocked by local anesthetic microinjected into the midline and lateral aspects of the rostral medulla. It is now recognized that the PAG mediates its antinociceptive effect through medullary relays with spinal projections that inhibit further nociceptive inputs at the level of the spinal cord.

Noradrenergic Inhibition of Pain

Early investigations found that antinociception mediated from stimulation of the PAG, either electrically or by microinjection of morphine, could be attenuated by the spinal injection of α-adrenergic blockers. Other studies demonstrated that the spinal administration of α-2 adrenergic agonists acts synergistically with opioids to enhance antinociception. These studies indicated that the endogenous pain-control mechanisms depend, at least in part, on a noradrenergic system. Consistent with this hypothesis, it was found that several noradrenergic sites, the A5 (locus coeruleus), A6, and A7 (Kölliker–Füse) nuclei, possess communications with the PAG and contribute to its antinociceptive actions. Critically, the locus coeruleus is the major source of noradrenergic projections to the spinal cord, and the A7 is also an important source of noradrenergic projections to the superficial laminae of the spinal dorsal horn. These noradrenergic spinopetal projections may inhibit the response of second-order neurons of the spinal dorsal horn. In addition, the A7 projections may also attenuate the activity of primary afferent nociceptors through activation of the α-2 adrenergic receptors residing on the nerve terminals. Several studies showed that electrical or chemical stimulation of these noradrenergic regions produced antinociception sensitive to blockade by spinally administered adrenergic antagonists. Electrical or neurotoxin lesions of the noradrenergic nuclei attenuate the antinociceptive effect of opioids and enhance the sensitivity of dorsal horn neurons to noxious stimuli. The noradrenergic regions also receive inputs from structures related to the pain-processing pathways, including the parabrachial region, and send projections to other sites, including the nuclei of the rostral ventromedial medulla (RVM). This region is generally regarded as the final relay in the modulation of pain.

Descending Inhibition from the Rostral Ventromedial Medulla

The RVM is defined as the medulla ventral to the nucleus gigantocellularis and extending between the caudal facial nucleus and the inferior olivary complex, and it includes the serotonergic nucleus raphe magnus (NRM) and the nucleus gigantocellularis pars α and the surrounding reticular formation. The RVM exhibits reciprocal connections with the PAG as well as inputs from the A5 and A7 noradrenergic nuclei, parabracial region, and limbic and hypothalamic sites (Figure 1). Early studies aimed at determining the antinociceptive pathways linked to the PAG found that electrical stimulation or morphine microinjection into the RVM produced behavioral signs of analgesia and inhibited dorsal horn neuronal responses to noxious stimuli. Local anesthetic block of the RVM or electrolytic or chemical lesions of the RVM abolished antinociception arising from stimulation of the PAG. The RVM is a major source of bulbospinal projections that terminate in laminae I, II, and V of the spinal dorsal horn, where they can form synapses with second-order neurons and terminals of primary afferent nociceptors, thus inhibiting further nociceptive inputs.

Extensive studies have been performed over the past three decades to characterize the neurons of the RVM. Based on the response characteristics to noxious heating of the tail and measured by electrophysiological methods, there are three classes of neurons within the RVM. The on cells were so named because of a marked acceleration of firing rate that occurs just prior to the nociceptive tail-flick reflex. The off cells cease firing immediately prior to the nociceptive tail-flick reflex. The changes in firing rate were related to the nociceptive reflex and not to the intensity of the stimulus, indicating a physiological connection between the neuronal responses of these cells and behavioral signs of nociception. The third class of cells, the neutral cells, do not demonstrate any consistent changes with noxious stimuli and might be related to physiological processes other than modulation of nociception. Doses of morphine that produce behavioral signs of analgesia or inhibit dorsal horn responses to noxious stimuli given

either systemically or microinjected into either the PAG or RVM result in increased off cell activity and the suppression of on cell activity. Further studies established that the suppression of the off cell pause is necessary for the antinociceptive effect of opioids, whereas the inhibition of on cell activity is not necessary for antinociception to occur. Doses of opioid antagonists that abolish the antinociceptive effect of opioids also restore the cessation of the off cell activity. It is generally accepted that the activation of the descending projections of the off cells block nociceptive transmission at the level of the spinal cord and thus moderate the input of pain signals to higher centers of the brain. Descending inhibition from the RVM off cells is a key component of endogenous pain-control systems.

Pain Facilitation from Supraspinal Sites

Until the early 1990s, the main focus of studies of the role of the RVM was related to its inhibition of spinal nociceptive inputs. However, as studies of this region progressed, it became clear that a dual function may be evoked from the RVM. That is, stimulation of the RVM was found to either produce antinociception or to actually facilitate responses to noxious stimulus, indicating a pronociceptive effect. Electrical stimulation applied to the RVM at low intensities (e.g., 5–25 μA) facilitated the nociceptive tail-flick reflex in anesthetized rats, whereas higher intensities of stimulation (50–100 μA) inhibited the response. Likewise, the electrophysiological responses of dorsal horn neurons to noxious stimuli were facilitated by low-intensity stimuli applied to the RVM and inhibited by higher-intensity stimulus. Chemical stimulation of the RVM by the microinjection of excitatory substances such as glutamate or neurotensin produced a similar biphasic effect. Lower doses facilitated behavioral and electrophysiological responses to noxious stimuli, whereas higher doses inhibited the same responses. The microinjection of a neurotensin antagonist into the RVM also elicited a biphasic effect related to dose, leading to the conclusion that endogenous neurotensin in the RVM mediates the facilitation of nociception and, to a lesser extent, inhibition, which is masked under normal conditions.

Just as the off cells mediate descending inhibition, there is sufficient evidence to support the concept that the on cells mediate a descending facilitation of pain. The electrophysiological characteristics of the on cells are consistent with a pronociceptive function because they become active after a noxious stimulus is applied. Moreover, manipulations that facilitate responses to nociceptive stimuli also increase on cell activity For example, prolonged delivery of a noxious thermal stimulus facilitated the tail-flick reflex to noxious heat and also caused an increase in the firing rate of the on cells. This facilitated response was abolished by the microinjection of the local anesthetic lidocaine into the RVM. Withdrawal from opioids and the subdermal injection of formalin into a hindpaw of a rat both produced exaggerated behavioral responses to noxious stimuli, and both provoked increased responses of on cells of the RVM.

Whereas the concept of endogenous systems that inhibit pain was readily accepted and well studied, the concept that endogenous mechanisms promote nociception may at first appear to be counterintuitive. However, enhanced sensitivity to pain (hyperalgesia) may serve as a protective mechanism. The enhanced pain state may be beneficial to healing after an injury. For example, the inflammatory response and hyperalgesia accompanying a broken limb discourage its use during the healing process. Considerable evidence has recently been generated to point to the RVM as a principal contributor to the maintenance of enhanced pain states, and it may be responsible for the maintenance of chronic pain states. More recent studies have postulated the existence of a spinal-bulbo-spinal loop that may be important to the development and maintenance of exaggerated pain behaviors associated with chronic pain states.

Descending Facilitation from the Rostral Ventromedial Medulla Mediates Chronic Pain States

Considerable evidence shows that descending facilitation arising from the RVM maintains enhanced pain states. Hyperalgesia elicited by formalin injection into the hindpaw or by electrical stimulation of the RVM were both blocked by the spinal administration of methysergide, which blocks the facilitatory serotonergic projections from the RVM. Neurogenic inflammation caused by the application of mustard oil produced facilitation of dorsal horn neurons in response to noxious peripheral stimuli that was abolished by microinjection of lidocaine into the RVM. Neurotoxic lesions of the RVM also abolished behavioral signs of pain facilitation in animals with inflammation of a hindlimb. The microinjection of lidocaine into the RVM of rats with experimental nerve injury abolished hyperesthesias to sensory stimuli. Strong circumstantial evidence suggests that the population of RVM neurons that express the μ-opioid receptor corresponds to the RVM neurons that drive descending facilitation and correlate to the function of on cells. By using the μ-opioid receptor as a portal of entry into these pain facilitatory cells, it was

possible to selectively destroy a large part of this neuronal population. The ablation of this population of RVM neurons, presumed to correspond to on cells, abolished behavioral signs of enhanced pain in animal models of neuropathic pain.

Enhanced abnormal pain associated with peripheral nerve injury has been linked to the enhanced release of excitatory neurotransmitters from primary afferent neurons, which appears to be secondary to neuroplastic changes that include enhanced descending facilitation from the RVM. The stimulated release of neuropeptides from spinal cord preparations or cultured dorsal root ganglion (DRG) neurons has emerged as a useful tool for assessing the function of primary afferent nociceptors. Chronic pain states, both inflammatory or neuropathic, are associated with enhanced capsaicin-evoked release of substance P or CGRP. The disruption of descending facilitation from the RVM, through either physical ablation of the descending projections from the RVM in the dorsolateral funiculus (DLF) or by selective ablation of the pain facilitatory neurons of the RVM, abolished behavioral signs of neuropathic pain as well as the enhanced capsaicin-evoked release of peptidic transmitters from primary afferent nerve terminals. It is now hypothesized that persistent or repetitive input of noxious stimuli elicits supraspinal neuroplastic changes that results in the activation of a descending pain facilitation from the RVM. This spinopetal pain facilitatory process serves to promote additional release of excitatory neurotransmitters from primary afferent terminals, which in turn promotes the maintenance of the tonic descending facilitatory system, thus perpetuating a chronic pain state.

Descending Pain Facilitation Can Be Elicited by Neuroplastic Changes in the Rostral Ventromedial Medulla

Although the role of descending pain modulatory systems has been extensively studied, the role of the ascending inputs from the periphery to the brain in triggering this process is less clearly understood. In the normal state, acute nociceptive inputs are relayed to supraspinal sites, in part through the spinothalamic neurons of the superficial lamina of the spinal dorsal horn that express the neurokinin (NK-)1 receptor. These neurons are the primary targets of peptidergic primary afferent nociceptors and account for more than 77% of the nociceptive-responsive cells of lamina I. These lamina I projection neurons send terminations to thalamic nuclei, which integrate sensory inputs and serves as a relay to the somatosensory cortex, insular cortex, and amygdala. There is also a significant projection from lamina I to the parabrachial region, which receives sensory and nociceptive inputs from mechanical, thermal, and cold stimuli and which may act as a site of integration and discrimination of noxious inputs that are then passed on to the hypothalamus and amygdala. These projections lead to the activation of descending processes from the cortical and limbic regions to stimulate the PAG and ultimately the RVM, the final relay in processing of descending control of nociceptive inputs The selective ablation of the spinal neurons expressing the NK-1 receptor abolished behavioral signs of enhanced abnormal pain in rats with peripheral nerve injury or inflammation of the hindpaw, but it did not alter responses to acute nociceptive stimuli. Likewise, the augmented electrophysiological and neurochemical responses of dynamic wide-range neurons to peripheral stimuli were abolished by ablation of the NK-1-expressing neurons. It is now hypothesized that the activation of this ascending limb relays persistent nociceptive messages to the parabracial and thalamic regions, which ultimately communicate with the RVM. Persistent noxious inputs result in neuroplastic adaptations within the RVM to induce an activation of descending pain facilitatory systems from this region.

The RVM and surrounding region include the nuclei of prominent spinopetal serotonergic projections that constitute both inhibitory and facilitatory projections from this region. It has also recently become evident that spinal serotonin (5-HT)1A receptors mediate the inhibition of pain transmission, whereas the 5-HT3 receptors facilitate pain transmission in the spinal cord. Persistent noxious stimulation evoked the expression of the protooncogene product Fos, which is a neurochemical marker for neuronal activation, in the nucleus raphe magnus (NRM). Furthermore, double-labeling studies revealed that a subpopulation of these neurons also expressed serotonin, indicating that nociceptive inputs indirectly activate descending serotonergic fibers. Accordingly, the spinal administration of the 5-HT3 antagonist ondansetron mimicked the behavioral and electrophysiological consequences of ablation of the spinal neurons that express the NK-1 receptor in animal models of inflammation and of nerve injury. Genetically altered mice that do not express the 5-HT3R-A subunit demonstrate normal responses to acute nociceptive stimuli, but demonstrate significantly reduced behavioral and electrophysiological responses to inflammatory pain. It has been suggested that descending serotonergic facilitation of nociception mediated through the 5-HT3 receptor becomes active during conditions of chronic pain. Taken together, these studies provide growing evidence that chronic pain states are maintained by a spinal-bulbo-spinal excitatory system, in which incoming

pain causes supraspinal neuroplastic changes that, in turn, serve to facilitate further nociceptive inputs, thus maintaining a condition of enhanced pain.

Other Pain Facilitatory Sites

Anterior Cingulate Cortex

The anterior cingulate cortex (ACC) is recognized as an important cortical center of integrations of pain with emotional and situational cues. The ACC has reciprocal connections with the medial and ventrolateral thalamic nuclei, and it receives input from the hippocampus, amygdala, and hypothalamus. The ACC, together with these limbic regions, encodes emotional and autonomic responses to pain. The ACC sends projections to the PAG and the RVM, regions known to participate in the modulation of pain.

Electrophysiological recordings have shown that neurons within the ACC are activated in response to noxious stimulation. Recordings performed in monkeys suggest that the ACC may encode the anticipation of pain because they may be activated by conditioned aversive stimuli. Stimulation of the ACC enhanced the activity of spinal dorsal horn neurons that are responsive to noxious stimuli, and lesions of the ACC or lidocaine microinjection into this region resulted in decreased responsiveness of these neurons. Electrolytic lesions or microinjection of lidocaine into the ACC reduced behavioral responses to acute noxious stimuli as well as measures of enhanced pain in animal models of inflammation. Behavioral studies suggest that ACC lesions may block the aversive component of pain rather than modulating the response to pain intensity. Avoidance of noxious stimuli was attenuated by lesions of the ACC. In humans, brain imaging studies show that the ACC is activated by acute painful stimuli as well as by pain conditions such as headache and neuropathic pain. Functional magnetic resonance imaging (fMRI) studies also indicate that the ACC responds in anticipation of pain. Patients with lesions of the ACC demonstrate a reduced affective response to chronic pain states. Based on such observations, it is believed that the ACC is integral in the integration of the motivational and affective components of pain. This region integrates inputs from the limbic, sensory, and motor regions to perform a global assessment of situational cues along with the noxious inputs and to form the interpretation of the quality of pain.

The ACC may facilitate pain in part though the activation of the pain facilitatory neurons of the RVM because microinjection of lidocaine or an α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)/kainate receptor antagonist into the RVM blocked the facilitation of nociceptive reflexes produced by electrical or chemical stimulation of the ACC. Other studies suggest that pain facilitation from the ACC is processed through the medullary dorsal reticular nucleus (DRT).

Dorsal Reticular Nucleus

The DRT is found in the dorsal caudal – most portion of the medulla. This region has reciprocal connections with the spinal dorsal horn, and its spinopetal projections are consistent with the laminar distribution of nociceptive-encoding neurons. The DRT also communicates with the PAG, the nuclei of the RVM, the locus coeruleus, and the A5 and A7 noradrenergic nuclei as well as with thalamic and limbic structures. These connections implicate a role in the pain-processing pathways. Electrophysiological studies have demonstrated that the DRT is a pain facilitatory site that forms a spinal-bulbo-spinal feedback amplification loop that enhances nociceptive input. The neurons of the DRT appear to be excited almost exclusively by noxious stimuli. Stimulation of the DRT facilitates behavioral responses to noxious stimuli and enhances electrophysiological and neurochemical responses of dorsal horn neurons to nociceptive stimuli. In contrast, lesions of the DRT increase pain thresholds and result in reduced responses to nociceptive stimuli. Inflammation is associated with increased activity of the DRT, and lesions of the DRT attenuate augmented responses to noxious stimuli in animal models of inflammation. Lesions in or lidocaine injection into the DRT blocks enhanced responses to nociception produced by stimulation of the ACC. This nucleus provides an additional mechanism through which the enhancement of nociception and production of hyperalgesia in chronic pain states may be maintained.

Conclusion

Tremendous progress in our understanding of the mechanisms that modulate pain have been made over the preceding half century since pain research came into its own. Many exciting discoveries were made that established the existence of endogenous mechanisms that inhibit the expression of pain. The characterization of the pain-inhibitory pathways, along with the discovery of opioid receptors and the endogenous opioid peptides, not only enhanced our understanding of the mechanisms through which opioid analgesics may produce their therapeutic effects but also revealed the existence of dedicated pain-control circuitries. The system that is best characterized at the present time is the PAG-RVM-spinal pathway. The PAG appears to be a central collection and processing site for nociceptive inputs from caudal and rostral sites, where a pain-modulatory message

might be transmitted to the spinal cord through intermediate relays. Cortical regions integrate sensory, affective, and motivational inputs and may activate both inhibitory and facilitatory systems; this may partly explain the observations that environmental, psychological, and emotional cues alter pain responses. The RVM is capable of relaying both inhibitory and facilitatory signals to the spinal cord. The relative strength of these opposing signals influences the intensity of further incoming pain signals at the level of the spinal cord. Activation of pain-facilitatory systems from the RVM and the DRT also provides a mechanism through which chronic pain states may be maintained. Increasing evidence suggests that pain facilitation may serve a protective role in wound healing but that it may also become maladaptive, permitting long-lasting, abnormal pain that persists as neuropathic or other chronic pain states.

See also: Neuropathic Pain; Neuropeptides: Pain; Nociceptor Responses; Pain and the Sympathetic Nervous System; Pain Management; Pain: Central; Pain: Neuroimaging; Spinal Cord Pain Systems; Visceral Pain.

Further Reading

Gebhart GF (2004) Descending modulation of pain. *Neuroscience and Biobehavioral Reviews* 27: 729–737.

Ji RR, Kohno T, Moore KA, and Woolf CJ (2003) Central sensitization and LTP: Do pain and memory share similar mechanisms? *Trends in Neuroscience* 26: 696–705.

Lima D and Almeida A (2002) The medullary dorsal reticular nucleus as a pronociceptive centre of the pain control system. *Progress in Neurobiology* 66: 81–108.

Ossipov MH and Porreca F (2006) Role of descending facilitation in neuropathic pain states. In: Campbell JN, Basbaum AI, Dubner R, Dray A, Dworkin RH, and Sang CN (eds.) *Emerging Strategies for the Treatment of Neuropathic Pain*. Seattle, WA: IASP Press.

Pertovaara A and Almeida A (2006) Endogenous pain modulation: Descending inhibitory systems. In: Cervero F and Jensen TS (eds.) *Handbook of Clinical Neurology*, pp. 179–192. New York: Elsevier.

Porreca F, Ossipov MH, and Gebhart GF (2002) Chronic pain and medullary descending facilitation. *Trends in Neuroscience* 25: 319–325.

Romanelli P and Esposito V (2004) The functional anatomy of neuropathic pain. *Neurosurgery Clinics of North America* 15: 257–268.

Suzuki R and Dickenson A (2005) Spinal and supraspinal contributions to central sensitization in peripheral neuropathy. *Neurosignals* 14: 175–181.

Suzuki R, Morcuende S, Webber M, Hunt SP, and Dickenson AH (2002) Superficial NK1-expressing neurons control spinal excitability through activation of descending pathways. *Nature Neuroscience* 5: 1319–1326.

Suzuki R, Rygh LJ, and Dickenson AH (2004) Bad news from the brain: Descending 5-HT pathways that control spinal pain processing. *Trends in Pharmacological Sciences* 25: 613–617.

Vanderah TW, Ossipov MH, Lai J, Malan TP Jr., and Porreca F (2001) Mechanisms of opioid-induced pain and antinociceptive tolerance: Descending facilitation and spinal dynorphin. *Pain* 92: 5–9.

Vanegas H and Schaible HG (2004) Descending control of persistent pain: Inhibitory or facilitatory? *Brain Research. Brain Research Reviews* 46: 295–309.

Pain: Central

J Boivie, University Hospital, Linkoping, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Definitions

The International Association for the Study of Pain (IASP) has defined central pain as pain caused by a lesion or dysfunction in the central nervous system (CNS); nowadays central pain is often referred to as 'central neuropathic pain.' Note that, according to the IASP definition, the cause shall be a primary process in the CNS. Thus peripherally induced pain with central mechanisms is not central pain, even if the central mechanisms are prominent.

The term 'thalamic pain' is often used in a general sense for all central pain, and the expression 'pseudothalamic pain' is sometimes used for central pain caused by extrathalamic lesions. The term 'dysesthetic pain' is also sometimes used for central pain in general. Dysesthetic pain can have either central or peripheral causes. It is recommended that the general term, central pain, be used in most instances, and that only central pain caused by lesions in the thalamus should be labeled thalamic pain.

The term 'anesthesia dolorosa' has chiefly been used for head and face pain, and in particular for the neuropathic pain after destructive surgery or nerve blocks to treat neuralgias. The term 'deafferentation pain' is used for similar conditions, but it is more commonly used in patients with lesions of spinal nerves.

Historical Perspective

Central pain is puzzling for patients and their friends and families, but also for personnel in the health care system – doctors, nurses, and others who provide care for patients. It seems strange to the patient that that there is apparently nothing wrong in a leg or arm, but that it hurts so badly and feels so peculiar that it is difficult to find words to describe how it feels. It is also puzzling for the research worker, because it is difficult to find a plausible explanation for many of the features of central pain.

Knowledge about central pain in the late 1930s was summarized in an impressive series of publications in 1938 by Riddoch, who added observations of his own. Many of Riddoch's ideas about central pain showed considerable foresight in the light of today's view of the clinical features of patients with central pain. For instance, he discussed the difference between unpleasant and painful sensations, the presence of both spontaneous and evoked pain, the diversity in the quality of central pain, the overreaction to somatic stimuli, the many kinds and locations of CNS lesions that can cause central pain, the uncertainty of the role of autonomic dysfunction, and the role played by the capacity to cope with the disability central pain causes. Regarding the pathophysiology of central pain, he supported Head and Holmes' original idea that the crucial lesion is one of the dorsal column–medial lemniscal pathways, which we now know to be incorrect. A historical review of the literature was provided by Boivie in 2005.

With modern techniques and methods much new knowledge has been acquired, but there are still aspects of central pain that remain unknown, the most important one being our lack of knowledge about the underlying cellular mechanisms. One reason for this is the lack of experimental models for central pain caused by brain lesions. Experience from techniques to produce experimental lesions in the spinal cord gives hope that the use of this model will lead to new insights into central pain.

Historically, central pain appears to have first been knowingly described as early as 1883 by Greiff in a patient who developed lasting pain following cerebrovascular lesions. The lesion included the thalamus. Due to the description from 1906 by Dejerine and Roussy of the thalamic syndrome, including thalamic pain, this pain has remained the best known form of central pain, but only a minority of central pains are related to thalamic lesions.

In the oldest literature many fascinating detailed descriptions of the symptoms and signs are found, but no systematic studies were done contemporaneously. Many of the published descriptions show that the character of this pain can vary considerably from patient to patient and that central pain can be excruciating. They vividly illustrate how individuals interpret this kind of pain: as a crushed feeling, as a scalding sensation, as if boiling water was being poured down the arm, as if the leg was bursting, as if something was crawling under the skin, as pain pumping up and down the side, as if the painful region was covered with ulcers, as if pulling a dressing from a wound, as if a log of wood was hanging down from the shoulder, like a wheel running over the arm, and as a cold stinging feeling. As a result of descriptions such as these, central pain has commonly been thought of as being excruciating pain with bizarre character, and covering large areas of the body. This is only part of the truth, however, because central pain can appear in many guises. It can have a trivial

character and be restricted to a relatively small area, such as distal pain in one arm, or in the face. Central pain is mostly severe in the sense that it causes the patient much suffering, even though its intensity may be relatively low, because it is commonly irritating and constant.

Epidemiology

There are large differences in the prevalence of central pain among the disorders that may lead to such pain. The highest prevalences are found with traumatic spinal cord injury (SCI), multiple sclerosis (MS), stroke, and syringomyelia (Table 1). The latter is a rare disease with a very high incidence of central pain.

Two prospective epidemiological studies have been published, one of central poststroke pain and the other on pain following SCI. In the stroke study an incidence of 8.4% was found during the first 12 months, and in the spinal cord study 41% developed at-level pain (at least to some extent central pain) and 34% developed below-level pain (certain central pain) during the 5 years the patients were followed. Among the stroke patients with somatosensory deficit (42% of all stroke patients) the incidence of central pain was 18%. The corresponding figure in a mainly retrospective study of central pain in 63 patients with brain stem infarct was 44%, with an overall incidence for central poststroke pain (CPSP) of 25%.

A recent study of the prevalence of central pain in 364 patients with MS has found that 28% of patients experience, or have experienced, central pain. The prevalence of central pain in MS has not been specifically investigated in the previous studies of pain in MS, but the information given in some of the studies support the conclusion that the prevalence is in the order of 25–30%. Trigeminal neuralgia occurs in 5% of MS patients.

The highest prevalence of central pain is found in syringomyelia. It appears to be in the order of 67–75% of all patients. It has been suggested that this high incidence depends on the fact that the syrinx directly injures the dorsal horn of the spinal cord, where processes that are of fundamental importance in pain mechanisms reside.

Table 1 Estimated prevalence of major disorders with central pain and calculated number of patients per 10 million inhabitants

Disease	*Patients with CP (no.)*	*Patients with CP (%)*
Spinal cord injury	3200	35
Multiple sclerosis	2800	28
Stroke	6800	8.4

Lesions That Cause Central Pain

Many etiologically different lesions in the brain and spinal cord can cause central pain (Table 2). These lesions and dysfunctions are due to many different disease processes. The macrostructure of the lesion is probably less important than its location as regards the probability that it will induce central pain. This does not mean that the structure of the lesion has no significance, because it is conceivable that the microstructure of the lesion in some instances is critical, but there is to date no information from research on this matter. Lesions include both rapidly and slowly developing ones. There appears to be no correlation between the tendency to cause central pain and the speed with which they develop.

The role of thalamic lesions is one of the recurring questions in discussions regarding the location of the lesions that cause central pain. In their prospective study, Andersen et al. had 25% thalamic involvement among the 191 CPSP patients (as assessed by computed tomography scans), but in a study of 70 patients using the more sensitive method of magnetic resonance imaging (MRI) it was found that about 60% had lesions involving the thalamus.

The importance of the location of the lesion within the thalamus was elucidated in a study of thalamic infarct. The results showed that only patients with lesions including the ventroposterior thalamic region developed central pain. Many studies have shown that the first requisite for a disease process to produce central pain is that it affects structures involved in somatic sensibility, which is not surprising, since pain is part of somesthesia. Abnormalities in somatic sensibility are actually the only symptoms and signs besides pain that are present in all patients with central pain.

The dominating features of the sensory abnormalities are abnormal sensibility to temperature and

Table 2 Causes of central pain

Vascular lesions in the brain and spinal cord
Infarct
Hemorrhage
Vascular malformation
Multiple sclerosis
Traumatic spinal cord injury
Cordotomy
Traumatic brain injury
Syringomyelia and syringobulbia
Tumor
Abscess
Inflammatory diseases other than multiple sclerosis
Myelitis caused by viruses, syphilis
Epilepsy
Parkinson's disease (?)

pain, and hyperesthesia. The results from studies on patients with central pain following stroke, MS, SCI, and syringomyelia, employing quantitative methods to assess the sensory abnormalities, form the basis for the hypothesis that central pain only occurs after lesions that affect the spinothalamic pathways (i.e., the pathways that are most important for temperature and pain sensibility). If this hypothesis turns out to be correct it means that lesions of the dorsal column–medial lemniscal pathways are not necessary for the occurrence of central pain, but many patients also have such lesions and they certainly affect sensibility. Likewise many patients have severe deficits in temperature and pain sensibility and yet they do not develop central pain.

There is a large variation in the spectrum of sensory abnormalities among patients with central pain. It ranges from a slightly raised threshold for one of the submodalities, to complete loss of all somatic sensibility in the painful region. Hyperesthesia to touch, moderate cold, and heat; allodynia to touch and cold; and hyperalgesia to cold, heat, or pinprick are common in many central pain conditions, and often coexist. For instance, some stroke patients with severely decreased tactile sensibility, even with total loss, have tactile allodynia. Head and Holmes, as early as 1911, claimed that overreaction to somatic stimuli was the most typical sign of central pain.

Neurophysiological Examinations

Because of the fact that the sensory disturbances in central pain indicate that the lesions affect the spinothalamic pathways, it is of interest to study somatosensory evoked potentials (SEPs) evoked by peripheral stimulation of afferents that activate the spinothalamic pathways. This can be done by using lasers to stimulate cutaneous heat receptors. Studies using this technique on patients with CPSP have shown that abnormalities in the laser-evoked cortical potentials, which have a long latency, correlate well with abnormalities in the sensibility to temperature and pain, but not to touch and vibration.

Pain Characteristics

Pain Location

In stroke about 70% of patients have hemipain, which in a minority affects the face (see Table 3 and Figure 1). In MS central pain dominates in the lower extremities (87%), and occurs over the upper extremities (31%). About 5% have trigeminal neuralgia. Central pain is experienced as superficial or deep pain, or with both superficial and deep components, but the high incidence of cutaneous hyperesthesia contributes to the impression that superficial pain dominates, although deep pain is also common.

Table 3 Common locations of central pain

Stroke
All of one side
All of one side except the face
Arm and/or leg on one side
Face on one side, extremities on the other side
Face
Multiple sclerosis
Lower half of the body
One or both legs
Arm and leg on one side
Trigeminal neuralgia
Spinal cord injury
Whole body below the neck
Lower half of the body
One leg
At the level of injury

Quality of Pain

No pain quality is pathognomonic for central pain. Central pain is thus not always burning or 'dysesthetic.' In fact, central pain can have any quality, and the variation among patients is great, although some qualities are more common than others (Table 4). Another basic feature is the presence of more than one pain quality in most patients. Different pains can coexist in a body region, or may be present in different parts of the body. One would expect the location of the lesion to be a deciding factor regarding the quality of pain. This appears to be partly true, but it is also apparent that similar lesions can lead to different pain qualities.

Intensity of Pain

The intensity of central pain ranges from low to extremely high. However, even if the pain is of low or moderate intensity, the patients assess the pain as severe because it causes much suffering, due to its irritating character and constant presence.

Central pain usually has a relatively constant intensity, but in some patients the intensity varies. These variations seem to occur spontaneously or under the influence of external somatic stimuli or psychological stimuli, or due to internal events, such as cutaneous stimuli, body movements, visceral stimuli, emotions, and changes in mood. It is also common that patients with central pain experience an immediate increase in pain after sudden fear, joy, loud noise, or bright light. Experience from clinical practice indicates that central pain is as affected by psychological factors, as other pain conditions (i.e., anxiety and depression) aggravate central pain.

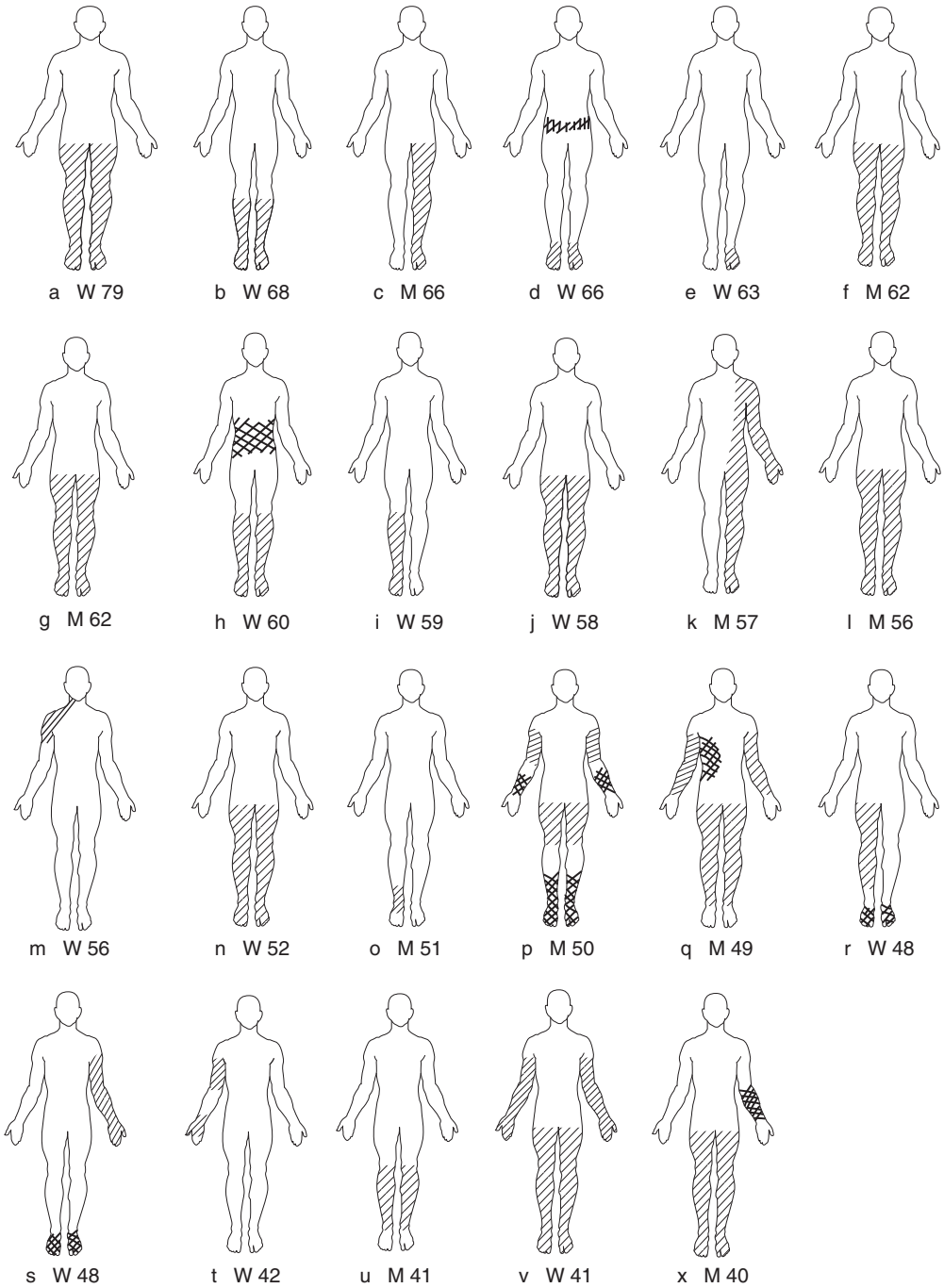

Figure 1 Pain drawings of 23 patients with central pain. Sex (W, female; M, male) and age in years are shown for each patient. From Österberg A and Boivie J, unpublished results.

Temporal Pattern of Pain

Central pain may start almost immediately after occurrence of the lesion, or it may be delayed for up to several years. By its nature it is often not possible to define this in disseminated or slowly progressive disease such as MS or syringomyelia. Delays are well known in stroke. This delay may be as long as 2–3 years, but in most patients the pain starts within a few of weeks of the stroke. Most spontaneous central pain is present constantly, with no pain-free intervals. Unfortunately, central pain is usually permanent, but it may remit completely. CPSP may successively decrease and cease completely, but usually continues throughout life. A few cases have been reported in which a new supratentorial stroke abolished the pain.

Table 4 Qualities of pain reported by patients with central pain

Burning[a]	Shooting	Stabbing
Aching[a]	Squeezing	Cramping
Lancinating[a]	Throbbing	Smarting
Pricking[a]	Cutting	Pulling
Lacerating[a]	Crushing	Sore
Pressing[a]	Splitting	Icy feeling
	Stinging	

[a]The most commonly reported qualities.

Pathophysiology

The pathophysiology of central pain can be summarized as follows:

1. The disease process that causes central pain involves the spinothalamic pathways, including the indirect spino-reticulo-thalamic and spino-mesencephalic projections or their trigeminal equivalents, as indicated by abnormalities in the sensibility to pain and temperature. Thalamocortical lesions are also able to precipitate central pain.
2. The lesion probably does not have to involve the dorsal column–medial lemniscal pathways to elicit central pain.
3. The lesion can be located at any level of the neuraxis, from the dorsal horn to the cerebral cortex.
4. Many etiologically diverse processes may cause central pain, but the probability of central pain occurring varies greatly between these diseases, from being rare to occurring in the majority of patients.
5. As yet no single region has been shown to be crucial in the processes underlying central pain, but three thalamic regions have been focused upon – namely, the ventroposterior, reticular, and medial/intralaminar regions. The role of the cerebral cortex in central pain is unclear.
6. The pain and hypersensitivity experienced by central pain patients are believed to result from increased neuronal activity and increased neuronal reactivity along the somatosensory pathways, and decreased inhibitory mechanisms.
7. The cellular processes underlying central pain are still unknown, but processes involving excitatory amino acids, and in particular *N*-methyl-D-aspartate (NMDA) receptors, have been implicated.

Hypotheses Concerning the Mechanisms Involved

Disinhibition by lesions in the medial lemniscal pathways The notion that central pain is caused by lesions in the dorsal column–medial lemniscal pathways was one of the most favored hypotheses in the first half of the past century. The crucial physiological consequence of the lesions was thought to be a disinhibition of neurons in the pain-signaling system.

Lesion in the spinothalamic pathway In the past two decades most investigators have found evidence suggesting that the spinothalamic system is affected in the majority of central pain patients (see earlier). This is the basis for the currently favored hypothesis that central pain only occurs after lesions affecting the spinothalamic system. In many patients the lemniscal pathways are also affected, but this does not appear to be necessary for the occurrence of central pain. It even appears possible that involvement of the lemniscal pathways does not affect the character of the pain. Based on results from research on patients with central pain following spinal cord injuries, it has been concluded that spinothalamic tract lesions are necessary for the development of central pain, but that some other kind of perturbations may also be necessary for its development.

Disinhibition by removal of cold-activated spinothalamic projections The most recent hypothesis is that of Craig. He proposed that "central pain is due to the disruption of thermosensory integration and the loss of cold inhibition of burning pain." This disruption is, according to the hypothesis, caused by a lesion somewhere along the spinothalamic projections to the thalamus.

Treatment

General Aspects

Treating central pain is no easy task, because there is no universally effective treatment. This means that one often must try various treatment modalities to get the best results (Table 5). Treatment usually reduces the pain, rather than giving complete relief, but relatively small decreases in pain intensity are often highly valued by the patients.

One of the similarities between central pain and peripheral neuropathic pain is type of treatment. In both pain categories antidepressants and antiepileptic drugs are the most frequently used drugs, and have the best documented effects and are the only drug types tested in well-conducted clinical trials. They are the first-line treatments, together with transcutaneous electrical nerve stimulation (TENS). There is a body of literature containing thorough discussion of the treatments of neuropathic pain; here only some brief comments are made on this subject.

It is conceivable that treatment affects some aspects of central pain but not others. Therefore it would be desirable to assess the effect of treatment on each pain

Table 5 Treatment modalities used for central pain

Pharmacologic
Tricyclic antidepressant drugs
Antiepileptic drugs
Analgesics
Antiarrhythmic drugs, local anesthetic agents
Other drugs
Sensory stimulation
Transcutaneous electrical stimulation
Spinal cord stimulation
Deep brain stimulation
Motor cortex stimulation
Neurosurgery
Dorsal root entry zone lesions

modality separately, in accordance with ideas about mechanism-based treatments of pain. Another important but still largely unanswered question is whether the different central pain conditions respond differently to a particular treatment. This has not been systematically studied, but such differences appear to exist. From the literature and from clinical experience, one gets the impression that central poststroke pain, as compared to central pain in spinal cord injury (SCI) and MS, responds better to antidepressants. Conversely, paroxysmal pain in MS seems to respond much better to antiepileptic drugs than do other kinds of central pain. It is still uncertain to what extent patients with central pain benefit from the use of strong analgesics. The conclusion is that some, but only a minority, of the patients experience sufficient relief to continue long-term treatment with analgesics. Many of the other treatments listed in Table 5 are more of an experimental character, although some of them are used quite frequently.

Electrical stimulation of the spinal cord or the brain is an exclusive mode of treatment that should be reserved for particularly severe and treatment-resistant pain conditions. The exquisite pain suffered by many central pain patients fulfills such a criterion. From a review of the literature, and based on experience with his own patients, Tasker concluded that spinal cord stimulation is not effective enough in central pain to be recommended, a view shared by Gybels and Sweet, Nashold, and Pagni. Instead Tasker favors deep brain stimulation (DBS) or motor cortex stimulation (MCS). In recent years the focus has been on surface stimulation of the motor cortex, with several groups reporting good effects, particularly on CPSP.

Many different surgical lesions have been tried to provide relief of central pain, but no particular lesion has been found that reliably results in successful outcome. Dorsal root entry zone (DREZ) lesions have gained interest over recent years for the treatment of central SCI pain, but so far results have not been consistent.

Antidepressants Table 5 presents the first-line drugs for central pain. Controlled trials have been done with antidepressants (ADs) only on CPSP (15 patients) and the central pain in SCI, with conflicting results. In the CPSP study there was a statistically significant reduction in pain as compared to placebo. The number-needed-to-treat (NNT) value for amitriptyline was 1.7 (1.1–3.0). These results contrast with those from a controlled study of amitriptyline on central pain in 44 patients with SCI in which no significant effects were found, compared to active placebo, but the doses were relatively low, judging by the plasma concentrations.

Antiepileptic drugs An effect on central pain has only been demonstrated for the antiepileptic drug (AED) lamotrigine in central poststroke pain. Gabapentin has been recommended for the treatment of neuropathic pain. Several case reports describing successful treatment of central pain have been published, but so far no good controlled studies on central pain have been published. Recently pregabalin, closely related to gabapentin, was shown to give good relief of central pain in a trial on 137 patients with SCI. In trigeminal neuralgia it is clear that the pain-relieving effect of carbamazepine is strongly correlated with the plasma concentration, as in epilepsy. It is not known weather this is also the case with AEDs in central pain. No such correlation was found in the study on CPSP.

Like the antidepressants, the AEDs have a tendency to cause troublesome side effects, and must therefore be managed with caution. It is possible that some groups of patients are more amenable to the side effects than are other patients with neurological diseases, because many neurologists have the impression that carbamazepine, and possibly also other AEDs, cause more problems in MS patients than in patients with idiopathic trigeminal neuralgia or epilepsy. This could be explained by the fact that carbamazepine has effects on cerebellar centers for the coordination of movements, which are frequently affected by the disease process in MS.

Many AEDs might cause allergic reactions, mainly in the skin. The most serious reactions occur in rare instances with the use of lamotrigine. The best way to prevent this is to increase the dose slowly.

Analgesics, cannabinoids The question of whether central pain responds to analgesics is still controversial. The results from acute, single-blind tests of opioids, as well as a controlled clinical trial with the potent μ-opioid agonist levorphanol on patients with CPSP, indicate low sensitivity to opioids. This is similar to the experience of many patients with central pain who undergo operations and receive opioids

postoperatively – namely, that they have a good effect on the pain related to the operation, but no effect on the central pain.

In a study on five patients with stroke and ten with spinal cord injury, intravenous morphine did not give significant pain relief, but there was a strong tendency to obtain better pain suppression by the morphine, compared to placebo. However, in the open posttrial oral medication period, only three patients (= 17%) continued with morphine for at least 12 weeks. The others stopped before that, because of side effects and/or poor relief. The conclusion was that opioids may be useful for a minority of patients with central pain, and that the effect is modest in most of these. This is in accord with common clinical experience.

For many years the use of cannabinoids in neuropathic pain has been discussed. Two controlled clinical trials on central pain in MS have now been reported. A weak but significant pain-relieving effect by oral dronabinol was found in 24 patients during 3 weeks of treatment. In the second study on 64 patients (34 took a whole-plant based spray for 5 weeks) a clear significant pain relief was recorded.

See also: Neuropathic Pain; Pain Management; Spinal Cord Injuries; Spinal Cord Pain Systems; Stroke; Stroke: Injury Mechanisms; Syringomyelia.

Further Reading

Andersen G, Vestergaard K, Ingeman-Nielsen M, et al. (1995) Incidence of central post-stroke pain. *Pain* 61: 187–193.

Attal N and Bouhassira D (2006) Pain in syringomyelia/bulbia. In: Cervero F and Jensen TS (eds.) *Handbook of Clinical Neurology,* vol. 81, pp. 705–713. Oxford: Elsevier.

Boivie J (2005) Central pain. In: McMahon SB and Koltzenburg M (eds.) *Textbook of Pain,* 5th edn., pp. 1057–1074. Edinburgh: Churchill Livingstone.

Boivie J, Leijon G, and Johansson I (1989) Central post-stroke pain – a study of the mechanisms through analyses of the sensory abnormalities. *Pain* 37: 173–185.

Bowsher D (1996) Central pain: Clinical and physiological characteristics. *Journal of Neurology, Neurosurgery & Psychiatry* 61: 62–69.

Bowsher D, Leijon G, and Thuomas K-Å (1998) Central post-stroke pain: Correlation of MRI with clinical pain characteristics and sensory abnormalities. *Neurology* 51: 1352–1358.

Ducreux D, Attal N, Parker F, et al. (2006) Mechanisms of central neuropathic pain: A combined psychophysical and fMRI study in syringomyelia. *Brain* 129: 963–976.

Leijon G, Boivie J, and Johansson I (1989) Central post-stroke pain – neurological symptoms and pain characteristics. *Pain* 36: 13–25.

MacGowan DJL, Janal MN, Clark WC, et al. (1997) Central post-stroke pain and Wallenberg's lateral medullary infarction: Frequency, character, and determinants in 63 patients. *Neurology* 49: 120–125.

Österberg A, Boivie J, and Thuomas K-Å (2005) Central pain in multiple sclerosis – prevalences, clinical characteristics and mechanisms. *European Journal of Pain* 9: 531–542.

Rog DJ, Nurmikko TJ, Friede T, et al. (2005) Randomized, controlled trial of cannabis-based medicine in central pain in multiple sclerosis. *Neurology* 65: 812–819.

Siddall PJ, Cousins MJ, Otte A, et al. (2006) Pregabalin in central neuropathic pain: A placebo-controlled trial. *Neurology* 67(10): 1792–1800.

Siddall PJ, McClelland JM, Rutkowski SB, et al. (2003) A longitudinal study of the prevalence and characteristics of pain in the first 5 years following spinal cord injury. *Pain* 103: 249–257.

Tasker RR (2001) Central pain states. In: Loeser J (ed.) *Bonica's Management of Pain,* 2nd edn., pp. 433–457. Philadelphia: Lippincott, Williams and Wilkins.

Vestergaard K, Nielsen J, Andersen G, et al. (1995) Sensory abnormalities in consecutive, unselected patients with central post-stroke pain. *Pain* 61: 177–186.

Pain: Neuroimaging

R C Coghill, Wake Forest University School of Medicine, Winston-Salem, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

The identification and characterization of brain mechanisms involved in the processing of pain is essential for a better understanding of the complex phenomenon that is the pain experience. Although many advances have been made in the understanding of central nervous system mechanisms involved in pain processing, delineating the roles of the supraspinal structures in pain has been a difficult task, in part because of the complexity and multiplicity of mechanisms supporting this essential sensory experience. For example, lesions of the cerebral cortex rarely produce a complete loss of the capacity to experience pain. Furthermore, pain is only rarely reported in response to direct electrical stimulation of the cerebral cortex in unanesthetized patients undergoing neurosurgery. These two classic findings have led to the long-held view that the cerebral cortex is not substantially involved in the processing of pain. Yet modern research employing single-neuron recording techniques in animals has identified several cerebral cortical regions that appear to be involved in pain processing. How can this apparent paradox be resolved to better understand the contribution of the cerebral cortex to the pain experience?

The answer lies in the possibility that simultaneous activation of multiple regions distributed throughout the brain may be necessary for many aspects of pain perception. Testing this hypothesis, however, is not feasible with the traditional single-neuron recording techniques that permit investigation of only small numbers of neurons in discrete brain regions. Other methodological approaches which can evaluate the responses of many brain regions simultaneously are needed to assess the extent and distribution of activation that may be evoked during the pain experience. To these ends, functional imaging techniques have been successfully employed to map the responses of multiple cerebral cortical and subcortical areas to pain.

To date, the most effective imaging techniques used to study pain have been positron emissions tomography (PET) and functional magnetic resonance imaging (fMRI). Both techniques typically examine changes in cerebral blood flow (CBF). CBF has long been known to be correlated with neuronal activity and, in many studies of cognitive tasks, has proven a reliable index of activity. Furthermore, this approach has the particularly powerful advantage that human subjects can be studied without anesthesia and can undergo multiple scans without significant radiation exposure. Autoradiographic techniques using similar indices of activity are employed effectively in animal studies as well.

Since 1991, a number of functional imaging studies have examined the processing of pain by the human brain and, despite the use of different painful stimuli and different experimental conditions, have produced a remarkably consistent map of brain regions involved in pain processing. Many investigations have now identified pain-induced activation of a number of brain regions which, together, appear sufficient to support the complex sensory, affective, autonomic, and motor components of the pain experience. These regions and their possible roles in the pain experience are detailed next.

Overview of Brain Regions Active during Pain

Thalamus

A substantial portion of nociceptive information entering the brain from the spinal cord is received by a number of nuclei in the thalamus. These include lateral thalamic nuclei, such as the ventroposteriolateral and the ventroposteriomedial nuclei, and medial thalamic areas, such as the intralaminar nuclei and mediodorsal nucleus. Neurons responsive to painful stimuli have been identified in both the lateral and medial thalamic nuclei – those in the lateral thalamic nuclei generally have unilateral receptive fields representing only a portion of the body, whereas neurons in the medial thalamic nuclei generally have large bilateral receptive fields. Neurons in both regions have been demonstrated to encode pain intensity with an accuracy sufficient to support behavioral discrimination. Accordingly, most functional imaging studies of acute pain in normal subjects have detected significant activation in the thalamus. This activation is typically located contralateral to the stimulated area, but may become bilateral if stimuli are sufficiently intense (**Figures 1(c) and 1(d)**). Determining exactly which thalamic nuclei are activated remains a difficult issue because neither PET nor fMRI offers sufficient spatial resolution to easily distinguish among the small, tightly packed nuclei of the thalamus.

Primary Somatosensory Cortex

Neurons in the lateral thalamic nuclei constitute the bulk of subcortical input to the primary somatosensory cortex (SI), a functional region consisting of Brodmann's areas 3, 1, and 2 of the postcentral gyrus.

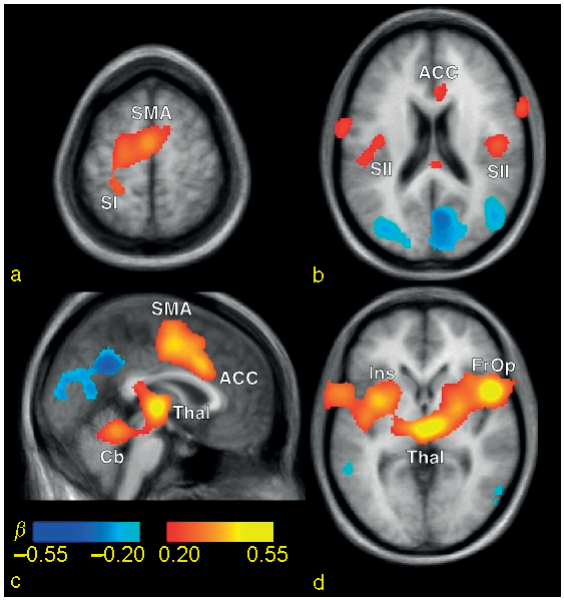

Figure 1 Distributed processing of pain intensity by the human brain: (a) activation of primary somatosensory cortex (SI); (b) activation of secondary somatosensory cortex (SII); (c) activation of thalamus (Thal) and cingulate cortex; (d) activation of Thal and insular cortex. Multiple regression analysis of positron emission tomography (PET) data revealed statistically reliable relationships between perceived pain intensity and activation of a functionally diverse group of brain regions, including those important in sensation, motor control, affect, and attention. Pain intensity-related activation occurred bilaterally in the cerebellum (Cb), Thal, insula (Ins), anterior cingulate cortex (ACC), and SII; contralaterally in the SI and supplementary motor area (SMA); and ipsilaterally in the frontal operculum (FrOp). Statistical maps derived from functional PET data (color) are superimposed on structural magnetic resonance images of brain structure (gray). The color scale denotes the regression coefficient of brain activation, which exhibited a statistically reliable relationship with subjective ratings of pain intensity. Modified from Coghill RC, Sang CN, Maisog JM, and Iadarola MJ (1999) Pain intensity processing within the human brain: A bilateral, distributed mechanism. *Journal of Neurophysiology* 82: 1934–1943, used with permission from The American Physiological Society.

Although SI has long been known to play a key role in the processing of innocuous stimuli, its role in the processing of pain has until recently been controversial. Classical studies of patients with brain lesions identified few disruptions of pain when the postcentral gyrus was damaged. Electrical stimulation of this area rarely led to the subjective experience of pain in conscious patients during neurosurgery. However, neurons responsive to painful stimuli have been identified in this area and exhibit a somatotopy similar to neurons that process exclusively innocuous stimuli. Nociceptive neurons in SI are notoriously difficult to identify and generally represent only a small portion of neurons encountered during recording sessions. Functional imaging studies by different groups have now demonstrated significant activation of SI during pain (Figure 1(a)). This activation is almost always contralateral to the stimulated region and occurs in areas consistent with the known somatotopy of SI. This region is activated in a fashion that is dependent on the perceived intensity of pain. The somatotopic focus of pain-induced activation in SI, combined with limited observations that lesions of SI disrupt the ability to extract some of the sensory features of painful stimuli, suggests that this region makes an important contribution to some of the sensory-discriminative aspects of the pain experience. Given the equivocal stimulation and lesion results in humans, it appears that although this region may contribute to pain perception, other regions must necessarily be activated to appreciate the full experience of pain.

Secondary Somatosensory Cortex and Posterior Insular Cortex

Other cerebral cortical areas probably involved in processing some of the sensory-discriminative aspects of pain include the secondary somatosensory cortex (SII) and related somatosensory fields of the posterior insula. These areas receive input from the SI, but also, in primates, receive direct input from thalamic regions likely to be involved in pain processing. Nociceptive neurons with large, typically bilateral receptive fields have been identified in these regions. Unilateral lesions of this area elevate pain thresholds and degrade tactile discrimination abilities, but do not produce a complete elimination of pain sensibility. Functional imaging studies frequently identify bilateral activation within the SII and adjacent areas (Figure 1(b)). Like the SI, the magnitude of this activation is also significantly related to the perceived intensity of pain. Functional imaging studies have also identified multiple and complex somatopically organized body representations in multiple subregions of the SII.

Cingulate Cortex

The cingulate cortex (Figures 1(c) and 2) is one of the most consistently activated regions during functional imaging studies of pain. Classic studies of chronic pain patients who have undergone transection of the cingulum bundle, a white matter pathway through which much input and output of the cingulate cortex travels, have led to the idea that the cingulate cortex plays a role in the processing of pain-related affect. Multiple functional imaging studies have shown that this area is involved with processes related to general affect and to pain-related affect. In humans, selective hypnotic amplification of the unpleasantness, but not the intensity, of pain has been shown to increase activation within this region, indicating that

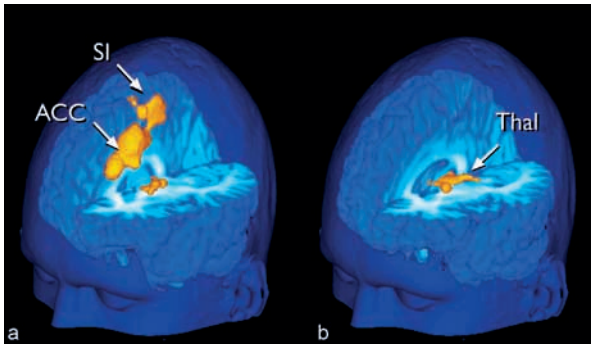

Figure 2 Pain-induced activation of the anterior cingulate cortex (ACC; upper orange blob) and thalamus (Thal; lower orange blob) displayed on a volume rendering of a human brain: (a) high-sensitivity individuals; (b) low-sensitivity individuals. SI, primary somatosensory cortex.

it is involved the processing of the negative emotional aspects of pain. However, this region also responds in a graded fashion to progressive increases in nociceptive stimulus intensity, indicating that it has access to and can use intensity-related information.

On the basis of both anatomic and functional criteria, this complex structure has been subdivided into the anterior cingulate cortex, midcingulate cortex, posterior cingulate cortex, and retrosplenial cortex. In studies of pain, regions of the anterior and midcingulate cortex in Brodmann's area 24 (typically referred to as the anterior cingulate cortex) are most frequently activated. Portions of this structure receive direct nociceptive input from the medial dorsal nucleus and intralaminar nuclei of the thalamus and have extensive connections with the amygdala, prefrontal cortex, and anterior insular cortex. Although nociceptive neurons have been identified in the cingulate cortex, this region is by no means exclusively involved with pain processing. It appears to be important in analgesia. The cingulate cortex has high concentrations of opioid receptors and is also active during placebo analgesia. Consistent with a role in analgesia, but in contrast to classic lesion studies in chronic pain patients, lesions of the anterior/midcingulate cortex and cingulum bundle increase responses to acute pain in nonpain patients.

The anterior cingulate cortex is involved with autonomic regulation and processing of emotional memories. This portion of the cingulate cortex is also frequently activated in functional imaging studies with high attentional demands and appears to be involved in avoidance learning. The midcingulate cortex is involved in reward-based decision making as well as response selection. Anterior portions of the midcingulate cortex have cortico-spinal projections and may play a role in some of the motoric responses to pain and other somatosensory stimuli. All these diverse functions are important for the processing of pain, and many careful investigations will probably be needed to fully characterize the multiple roles of the cingulate cortex in pain.

Anterior Insular Cortex

Similar to the cingulate cortex, the anterior insular cortex probably plays multiple complex roles in the processing of pain (Figure 1(d)). This region is frequently activated during functional imaging studies of pain and exhibits monotonic increases in activity as perceived pain intensity increases. The anterior insular cortex is reciprocally connected with anterior portions of the cingulate cortex and similarly receives direct nociceptive input from the thalamus. Like the cingulate cortex, it is connected with the amygdala and prefrontal cortex. However, it is highly connected with the posterior insular cortex, which, in turn, is connected with the SII. As such, it has been proposed to serve as a cortico-limbic pathway for somatosensory information. Limited stimulation and lesion studies support anatomic evidence for this region's serving to link somatosensory and limbic processes. Lesion patients have been reported to have altered affective responses to noxious stimuli, whereas nonpain patients undergoing stimulation of this area report feelings of fear. Functional imaging studies indicate that this region is not exclusively involved in unimodal somatosensory processing but responds in a multisensory fashion to visual, auditory, and tactile stimuli. This area, thus, may be critical for integrating ongoing afferent information with memory, allowing a full appreciation and evaluation of the meaningfulness of multidimensional aspects of the stimulus in light of previous experience.

Prefrontal Cortex

The prefrontal cortex is heavily connected with both the anterior cingulate cortex and the anterior insular cortex. This region has been thought to be involved in aspects of pain since the prefrontal lobotomy studies in the late 1940s. Patients with prefrontal lobotomies still experience pain and, in some cases, have been reported to have decreased thresholds for acute pain. However, during chronic pain, these patients exhibit disrupted pain affect and appear to have difficulty understanding the meaning and future implications of their pain. The prefrontal cortex is frequently seen to be activated in functional imaging studies of pain. In some instances, activation of this region is related to perceived pain intensity, particularly when analyses are weighted toward individual differences in subjective reports of pain. In other studies, prefrontal activation appears to be related to the cognitive difficulty of generating a rating of the stimulus. Such roles are not mutually exclusive because the

prefrontal cortex is extremely important in attention, affect, working memory, future planning, and decision making. It is clearly activated during the evaluation of stimuli of multiple sensory modalities. The prefrontal cortex is heavily connected with the amygdala and hippocampus and, as such, is well positioned to integrate information from the anterior cingulate and anterior insular cortices with both historical autobiographical information and information from the environmental milieu. All these processes are probably important in the evaluation of afferent sensory information as well as the application of a meaning and future significance to this information.

The prefrontal cortex also appears to be important in the modulation of pain. Studies of expectation and anticipation indicate that prefrontal activation prior to the stimulus can influence the subjective experience of the stimulus. Such roles of expectation are crucial elements of the placebo response. Consistent with this functionality, the prefrontal cortex is significantly involved in placebo analgesia.

Supplementary Motor Cortex and Motor Processing

A number of other brain regions are activated during the pain experience. Such areas include regions thought to be important in motor processing, such as the cerebellum, putamen, and supplementary motor cortex (Figures 1(a) and 1(c)). Escaping from and/or orienting to a painful stimulus is a critical response necessary to minimize possible injury and identify the cause of pain and is a key component of the pain experience. Given that subjects do not generate actual movements during the scanning period, activation in these motor-related areas has been postulated to reflect the neural infrastructure that processes movement preparation or suppression. The supplementary motor cortex is thought to be important in internally directed motor planning, timing, sequencing, and action retrieval. All these processes are probably engaged when a conscious human subject actively suppresses withdrawal responses while simultaneously considering how to escape from a noxious stimulus if the pain becomes intolerable.

Basal Ganglia

Typically, the functions of the basal ganglia are discussed in terms of motor processing. However, the basal ganglia are involved in numerous other processes and are frequently activated during pain. Of the basal ganglia, the putamen is most frequently activated during pain. Nociceptive neurons have been identified in this region as well as the caudate nucleus. Both the caudate and putamen represent important structures for cortical input into the basal ganglia. Much output of the basal ganglia is sent back to the same cortical region (via the thalamus) as the input. This loop architecture may provide a substrate for sequential evaluation of sensory information. In addition, the basal ganglia are known to play important roles in reward and reinforcement and to respond robustly to novel and behaviorally salient stimuli. All these processes are highly relevant to pain.

Cerebellum

Much like the basal ganglia, the cerebellum is typically associated with motor processing; its role in pain remains poorly explored and poorly understood. Both the cerebellar vermis and the cerebellar hemispheres are activated frequently in functional imaging studies of pain (Figure 1(c)). Much like the basal ganglia, the cerebellum is connected with the cerebral cortex in a loop architecture. Projections from the cortex synapse on pontine neurons, which in turn project to the cerebellum. The cerebellum then projects back to the same cortical region via the thalamus. These loops involve the motor and premotor cortices, as well as the prefrontal cortex. The cerebellum is frequently activated in cognitive studies that do not involve movement, and lesions of the cerebellum have been reported to disrupt affect. Thus, at least some of the cerebellar activation observed during pain may be related to cognitive and affective processes involving the prefrontal cortex.

Brain Regions Deactivated during Pain

As already discussed, numerous brain regions exhibit increased activity while subjects undergo noxious stimulation. However, several brain areas appear to be less active during pain than during rest or during nonpainful stimulation. These regions include the posterior cingulate cortex and the adjacent precuneus (Figure 1(c)), as well as the ventromedial prefrontal cortex. Similar deactivations are seen during a wide variety of experimental tasks that do not involve pain. Interpretation of these consistently deactivated areas has remained limited due to the difficulty of designing appropriate control conditions and the lack of fully quantitative assessments of brain activation. However, emerging evidence suggests that these areas are highly active during the rest condition when conscious subjects are frequently allowed to think in an unconstrained fashion. Then, when subjects are engaged in a task, these random thoughts are terminated and produce an apparent deactivation.

Brain Activation during Chronic Pain

When pain is evoked by normally nonpainful stimuli in patients with chronic pain, brain activation patterns are generally similar to those produced by acute pain in healthy subjects. However, during some forms of

chronic neuropathic pain, patterns of cerebral blood flow are markedly different than those seen in healthy subjects. During resting pain, patients with complex regional pain syndrome exhibit paradoxically low cerebral blood flow in the thalamus that is contralateral to their affected limb. The physiological significance of this unusual pattern of cerebral blood flow is unclear. It may reflect alterations in metabolic activity that result in perturbed relationships between neural activity and blood flow. Alternatively, it may represent a loss of inhibition, such that afferent nociceptive information is minimally modified before being distributed to the cerebral cortex.

Brain Mechanisms Supporting Individual Differences in the Subjective Experience of Pain

Pain is a profoundly individualized experience. Some individuals are highly sensitive to pain, whereas others are surprisingly insensitive. Prior to the advent of functional imaging, developing an understanding of these individual differences in the subjective experience of pain remained difficult due to the challenges of separating interindividual experiential differences from interindividual differences in the reporting of pain to a third-person observer. However, functional imaging studies of individuals reporting vastly different pain from a standardized heat stimulus indicate that, consistent with their subjective reports, highly sensitive individuals exhibit significantly more frequent and significantly greater activation of the primary somatosensory cortex, anterior cingulate cortex, and prefrontal cortex (Figure 2). In contrast, sensitive and insensitive individuals exhibit roughly similar levels of thalamic activation, suggesting that experiential differences emerge once afferent information has traveled beyond the suprathalamic levels.

Interindividual differences in the experience of pain may arise from a convolution of an individual's past history, present environment, and perceived future implications of a stimulus with incoming afferent information. The delineation of mechanisms by which these factors influence afferent processing has been accomplished, in part, by examining the effects of learned expectations on pain. Expectations are future predictions of the nature of an experience that are derived from previous experience with a stimulus as well as the present context in which the stimulus is presented. Brain regions including the prefrontal cortex, anterior cingulate cortex, and anterior insular cortex are activated in a manner that is proportional to expected pain, indicating that they are involved in the construction of a mental representation of an impending stimulus. All these regions are also activated during the actual experience of pain and, as such, can provide a substrate for the integration of expected information and afferent information. This convergence appears to be critical in the elaboration of the pain experience because reduced expectations of pain result in reduced reports of pain as well as reduced pain-related brain activation. Accordingly, expectations may provide a sensory set that substantially influences how afferent information is processed.

Distributed versus Parallel Processing of Pain

The multiple brain areas activated during functional imaging studies of pain reflect the complex mix of sensory, emotional, attentional, and motoric components of the experience that we know as pain. Although each region mentioned may be preferentially involved with some specific processing task, the number of different brain regions activated by painful stimulation and the high degree of connectivity among these areas strongly suggest that pain is processed by complex distributed cortical and subcortical networks. Furthermore, many of these networks are not only involved exclusively with the processing pain but are also probably involved in the processing of innocuous stimuli and in processes underlying cognitive evaluation of the situation within our environment.

A distributed system of pain processing explains the difficulty in eliciting painful sensations by cortical stimulation because simultaneous activation of several brain regions could be necessary for pain. Similarly, a distributed pain system explains why discrete cortical lesions seldom lead to a complete reduction in pain but may alter certain aspects of the pain experience, and why chronic pain may initially be alleviated by lesions at different levels of the neuraxis but usually returns after a period of months. This resiliency of chronic pain may involve a plasticity of networks involved in the processing of pain, whereby functions usually performed by one region are taken over by another. Such resiliency is of obvious evolutionary value because the ability to detect, understand, and escape pain is essential for survival. Future studies in this rapidly advancing field of functional imaging will probably provide a better characterization of the complex processes that underlie the pain experience.

See also: Cerebral Cortex; Cingulate Cortex; fMRI: BOLD Contrast; Nociceptor Responses; Pain Pathways: Descending Modulation; Pain: Central; Positron Emission Tomography (PET); Somatosensory Cortex: Functional Architecture; Somatosensory Cortex; Thalamus: Evolution in Vertebrates.

Further Reading

Coghill RC Brain Mechanisms of pain. http://www1.wfubmc.edu/Nba/Faculty/Labs/Coghill/Overview.html (accessed Jul. 2008).

Coghill RC, McHaffie JG, and Yen YF (2003) Neural correlates of interindividual differences in the subjective experience of pain. *Proceedings of the National Academy of Sciences of the United States of America* 100: 8538–8542.

Coghill RC, Sang CN, Maisog JM, and Iadarola MJ (1999) Pain intensity processing within the human brain: A bilateral, distributed mechanism. *Journal of Neurophysiology* 82: 1934–1943.

Koyama T, McHaffie JG, Laurienti PJ, and Coghill RC (2005) The subjective experience of pain: Where expectations become reality. *Proceedings of the National Academy of Sciences of the United States of America* 102: 12950–12955.

Price DD (2000) Psychological and neural mechanisms of the affective dimension of pain. *Science* 288: 1769–1772.

Rainville P, Duncan GH, Price DD, Carrier B, and Bushnell MC (1997) Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science* 277: 968–971.

Vogt BA (2005) Pain and emotion interactions in subregions of the cingulate gyrus. *Nature Reviews Neuroscience* 6: 533–544.

PANDAS (Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal Infections)

P J Hoekstra, R B Minderaa, and C G M Kallenberg, University of Groningen, Groningen, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The acronym PANDAS stands for 'pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections.' The notion that tics may be associated with infections was raised by the German psychiatrist Erwin Straus in 1927, and 2 years later by the American neurologist Laurence Selling, but subsequently faded into oblivion. In the mid-1990s, however, several case studies entered the literature, describing children who suddenly demonstrated severe tics and obsessive–compulsive symptoms that appeared to be temporally linked to streptococcal infections. Confronted with a failure of conventional treatment approaches, the authors of those studies described the successful application of immune-based treatments in some of these cases, consisting of therapeutic plasma exchange, intravenously administered immunoglobulins, or simple use of antibiotics to treat the infectious process. In some cases, streptococcal reinfections were associated with the reinduction of symptoms.

There also was the notion that patients with Sydenham's chorea, which has a well-established link to group A streptococcal infections, often show behavioral symptoms, including obsessions, compulsions, marked irritability, and emotional lability, and a substantial number of children with obsessive–compulsive disorder (OCD) were reported to show choreiform movements or tics. Furthermore, a study had suggested the presence of autoantibodies reacting with brain tissue in the serum of patients with tics and/or OCD. These observations led researchers at the National Institute of Mental Health to formulate clinical criteria for a putative subgroup of children with OCD and/or a tic disorder, designated PANDAS; in these children, symptom exacerbations were abrupt, dramatic, and temporally related to group A beta-hemolytic streptococcal (GABHS) infections. The concept of PANDAS strongly centers on Sydenham's chorea as a putative disease model, including the suggested pathogenic role of autoantibodies that cross-react with basal ganglia antigens as a hypothesized consequence of structural homology with streptococcal antigens. Research criteria were outlined for the PANDAS subgroup (see Table 1). What percentage of children with a tic disorder and/or OCD meet criteria for PANDAS is unknown. Although PANDAS criteria require prepubertal symptom onset, cases with adult-onset symptoms that otherwise meet PANDAS criteria have been described.

PANDAS is not an established clinical entity nor enlisted in formal disease classifications. Indeed, the very idea that tics or OCD symptoms might be related to autoimmunity remains, as yet, controversial, even when and if there is an established association with infections. Moreover, some of the criteria for PANDAS have been criticized for being vague and insufficiently operationalized. Especially, it is hard to say when precisely the criterion of a temporal association between GABHS infections and symptom onset and exacerbations has been met. Interestingly, in Sydenham's chorea, a latent period as long as 6 months, between onset of symptoms and the preceding GABHS infection, is not uncommon. GABHS infections are fairly common in children in general, as are remissions and exacerbations in children with tic disorders. Another criterion for PANDAS, the presence of an episodic course of symptom severity with explosive exacerbations and remissions, is not easy to apply either. An episodic course of symptoms is characteristic of pediatric tic and OCD symptoms in general, but it is hard to say when exacerbations can exactly be called explosive.

A more fundamental concern regarding the concept of PANDAS is the risk of circularity. From a scientific perspective, research that attempts to define the extent of immune involvement in neuropsychiatric disorders should better preclude the *a priori* selection of cases that, by definition, demonstrate a temporal relationship with streptococcal infections. Indeed, much of the evidence supporting a possible role for autoimmunity in tic disorders and OCD has been based on unselected individuals. Findings in unselected cases and in cases fulfilling criteria for PANDAS have been similar in magnitude. Still, introduction of the concept of PANDAS has certainly done much to stimulate research and clinical interest in the potential relevance of immune factors in tic and related disorders. In the following sections, after outlining the main clinical features of PANDAS, we present an overview of findings supporting, as well as challenging, the immune hypothesis of tic and related disorders.

Clinical Features

The first PANDAS criterion requires the presence of a tic disorder and/or OCD, two clinical entities that are generally regarded distinct in phenomenology and

Table 1 Criteria for diagnosis of PANDAS

Presence of obsessive–compulsive disorder and/or a tic disorder
Pediatric onset
Episodic course of symptom severity
Association with group A beta-hemolytic streptococcal infection
Association with neurological abnormalities

presumed etiology. Involuntary, rapid, recurrent, nonrhythmic motor movements and/or vocalizations are the defining features of tic disorders. Common motor tics include eye blinking, neck jerking, shoulder shrugging, and facial grimacing, whereas typical examples of vocal tics are throat clearing, barking, sniffing, and hissing. Tic intensity can vary substantially, ranging from barely visible or audible tics to extremely forceful or loud expressions. Quite often, tics are mild, in which case they may hardly attract attention from others and may not interfere with everyday life at all. However, powerful and frequent tics may severely interfere with everyday activities, including speech, driving, and walking. In exceptional cases, tics may lead to physical injury, including joint dislocation and other tissue damage. Tics usually start around the age of 6 years, are often transient, and tend to improve after adolescence, but may persist into adult life. Frequently, tic disorders are associated with a wide variety of emotional disturbances, especially obsessional phenomena, sometimes justifying a comorbid diagnosis of OCD. OCD is characterized by recurrent and persistent thoughts (obsessions) and/or repetitive behaviors (compulsions) that the person feels driven to perform in response to an obsession, or according to rules that must be applied rigidly. Only a subgroup of pediatric OCDs may be etiologically related to tic disorders, characterized by a family history of tics, a more familial subtype in general, male preponderance, association with disruptive behavior and developmental disorders, and a less striking association with mood disorders. The predominant phenomenology of obsessions/compulsions also differs between tic-related and non-tic-related OCDs. Presence of hoarding/symmetry and sexual and aggressive symptoms predominates in the former, whereas in non-tic-related OCD symptoms of contamination/checking prevail. In tic-related OCD, compulsions do not generally appear to be anxiety driven, but rather could be phenomenologically regarded as complex tic behaviors. It is well known that so-called mental tics form a part of tic symptomatology, and in some cases certain obsessions and compulsions could be better conceptualized as mental and complex tics, respectively. Thus, it may well be that the presence of tics – or, more generally, a movement disorder, in contrast to anxiety-driven obsessions and compulsions – forms the distinctive feature of those neuropsychiatric disorders which may be related to autoimmunity.

The PANDAS criterion of an episodic course of symptom severity refers to a clinical course characterized by the abrupt onset of symptoms or by dramatic symptom exacerbations. Typically, the onset of a specific symptom exacerbation can be assigned to a particular day or week, at which time the symptoms seemed to 'explode' in severity. Symptoms usually decrease significantly between episodes and occasionally resolve completely between exacerbations. Symptom exacerbations must be temporally related to GABHS infection (i.e., associated with positive throat culture and/or elevated anti-GABHS antibody titers). The temporal relationship between the GABHS infection and the symptom exacerbation may vary over the course of the illness, with recrudescences following the GABHS infections at much shorter intervals, often with a time lag of only several days to a few weeks, than are seen in first episodes. Because fever and other stressors of illness are known to increase symptom severity, the exacerbations should not occur exclusively during the period of acute illness. Finally, common examples of the required association with neurological abnormalities include motoric hyperactivity and adventitious movements (including choreiform movements or tics).

Laboratory Abnormalities

The diagnosis of PANDAS is made on clinical criteria alone. There are no laboratory tests that can diagnose PANDAS. Still, a number of the laboratory findings that have been reported in tic disorders and/or OCD may support the involvement of immune dysregulation. These include the presence of serum autoantibodies reacting with basal ganglia, increased streptococcal antibody titers, alterations in T and B cell subsets, and changes in levels of cytokines and adhesion molecules. It should be noted that these laboratory measures have virtually no place in the clinical management of patients with a tic disorder, OCD, or suspected PANDAS, perhaps with the exception of streptococcal antibody titers.

Antineuronal Antibodies

The presence of autoantibodies against basal ganglia in the peripheral blood of patients with tics and/or OCD is potentially a strong line of evidence in favor of the autoimmune hypothesis of these disorders, and has been repeatedly reported by several research groups. Results, however, have been partially conflicting, and relatively high levels of antineuronal staining have been found in sera from healthy controls. It is unclear how to interpret this overlap between patients and controls. It

raises suspicions of methodological flaws, suggesting rather high levels of nonspecific binding, albeit not necessarily, given that autoantibodies in healthy individuals have repeatedly been reported. Unfortunately, as yet, no accepted standardized methodology of assessing anti-basal-ganglia antibodies is available. Rather, different research centers have all used different methods, which makes results difficult to compare. Varying substrates of brain tissue have been utilized across studies, ranging from immortalized neuronal cell lines to animal and human brains, and have, moreover, been prepared in different ways. Also, methods of assessing antibody binding to brain tissue have differed from center to center. Early studies used immunofluorescence microscopy, which implies that assessment of presence of antibody binding depends on subjective human judgment, whereas more recent studies have made use of enzyme-linked immunosorbent assays (ELISAs), which is a more objective and quantitative method.

Assessed with an indirect immunofluorescence technique, autoantibody binding to human caudate tissue in patients with tics and/or OCD has been reported in 44–50% of tic disorder patients versus 21–24% of healthy controls. Subsequent studies using ELISA techniques against either an immortalized neuronal cell line, human basal ganglia, or rat brain in general have confirmed the increased levels of serum antineuronal antibodies in patients with tic disorders.

An important consideration when interpreting studies regarding anti-basal-ganglia status is whether observed antibody–antigen bindings are truly specific antineuronal bindings. Brain tissue contains a host of antigens which are not exclusively neuronal. The technique of Western blot, which separates proteins according to molecular weight, facilitates the identification and subsequent sequencing (molecular characterization) of specific candidate autoantigens. Common binding to basal ganglia antigens of 40, 45, and 60 kDa has been observed; these autoantigens have been identified as neuronal glycolytic enzymes (i.e., aldolase C, enolase, and pyruvate kinase M1, respectively). Neuronal glycolytic enzymes are multifunctional proteins that are expressed intracellularly as well as on the neuronal cell surface. On the neuronal plasma membrane, these enzymes are involved in energy metabolism, cell signaling, and synaptic neurotransmission. Interestingly, antibodies to pyruvate kinase have been found to also react strongly with surface antigens of infectious strains of *Streptococcus*, and conversely, antibodies to streptococcal M proteins react with pyruvate kinase.

The mere presence of autoantibodies does not necessarily imply a causal involvement of autoimmunity. Autoantibodies against a range of tissue antigens are normally present in healthy individuals, probably as part of the immune system's ongoing process to discriminate between self and non-self. Whether these autoantibodies cause damage and, thus, may have pathological significance is what counts, not the presence *per se*. Autoantibodies are also found when tissue damage is caused by trauma or infection. In other words, autoantibodies can result from, rather than be the cause of, tissue damage. Currently, the pathogenetic meaning of antineuronal antibodies in tic disorders and OCD is unclear. Some support for the direct involvement of antineuronal autoantibodies in the disease process stems from two different research findings. First, intravenously administered immunoglobulins and therapeutic plasma exchange were reported to be effective treatments in cases fulfilling PANDAS criteria. These treatment modalities are thought to block or remove the antineuronal autoantibodies, respectively. A second pillar possibly supporting a pathogenic role of antineuronal antibodies is formed by induction of stereotypic movements in rodents through passive transfer of antibodies by means of striatal microinfusions. Rats microinfused with sera from patients with a tic disorder subsequently show a significant increase in ticlike behaviors (e.g., licks and forepaw shakes) and episodic utterances, which are not observed when sera from healthy controls are microinfused. These findings have not uniformly been replicated, however.

In conclusion, the assessment of serum autoantibodies has appeared to be a fruitful research field in tic disorders. However, this area is very much a work in progress, especially with regard to standardization of methods and elucidating the pathogenic relevance of the antibody–antigen interactions in animal models.

Association with Antistreptococcal Antibody Titers

It is remarkable that studies that did not preselect tic disorders according to PANDAS criteria provide the best available evidence for an association of tic disorders with streptococcal infections, even when the exact nature of this possible association remains unclear. Such studies have repeatedly found markedly increased serum levels of antistreptococcal antibodies in a single-time-point design. Interestingly, it has been reported that levels of streptococcal antibody titers correlate with tic severity and that elevated streptococcal antibody titers are associated with the presence of anti-basal-ganglia antibodies. The cross-sectional nature hampers interpretation of these findings, especially since antibody levels are typically not assessed at the time of the first appearance of the tics, nor at a time of symptom exacerbation, but rather at an arbitrary point in time.

The cross-sectional data of increased antistreptococcal titers fit well with a longitudinal study that

found that the average annual rate of new streptococcal infections was 0.42 infections per patient per year among unselected pediatric patients with Tourette's disorder or OCD, compared with 0.28 infections per person per year for controls, even when no association was found between new streptococcal throat infections and symptom exacerbations. In another longitudinal study, however, positive correlations were detected between streptococcal titers and obsessive–compulsive severity rating changes in a subgroup of patients with large symptom fluctuations over time. These subjects were also more likely to have elevated streptococcal titers in comparison with patients without such dramatic fluctuations. Furthermore, there are data indicating that patients with tic disorder or OCD are more likely than controls to have experienced streptococcal infection in the 3 months prior to disease onset date, as well as multiple streptococcal infections within the preceding 12 months. Finally, in a longitudinal study involving a large number of healthy children, a strong relationship has been demonstrated between positive streptococcal throat swab cultures and both hyperactive behavior and choreiform movements. All in all, these studies indicate an intriguing association between streptococcal infections and tic and OCD symptoms. Large-scale prospective studies are needed to fully understand the relevance of streptococcal infections.

Alterations in Other Immune Parameters

Studies of lymphocyte subsets in tic disorder/OCD patients have provided some interesting preliminary findings. A D8/17-specific mouse monoclonal immunoglobulin M (IgM) antibody has been reported to bind to an increased percentage of B lymphocytes in rheumatic fever patients and patients with tic disorder, OCD, or PANDAS, in comparison with healthy controls. These studies seemed to suggest that increased D8/17 B cell expression could not only serve as a peripheral susceptibility marker for rheumatic fever, as had previously been investigated, but also as a marker for movement disorders and/or OCD associated with preceding streptococcal infections. Antibodies against D8/17 were originally prepared from fusions of spleen cells from mice that had repeatedly been immunized with isolated human B lymphocytes obtained from rheumatic fever or rheumatic heart disease patients. Unfortunately, several problems considerably reduce the enthusiasm for D8/17. More recent studies that used more sophisticated methods (flow cytometry vs. immunofluorescence microscopy; proper isotype control to rule out nonspecific binding) have failed to replicate earlier findings, perhaps due to insufficient stability of the antibody. A major problem is the lack of molecular characterization of the antibody. Therefore, the precise meaning of elevated D8/17 expression remains elusive.

New data suggest that patients with tic disorders may have a decreased capacity to inhibit autoreactive lymphocytes through a deficit in T regulatory cells. A significant decrease in T regulatory cells is observed in patients with moderate to severe tic symptoms compared with healthy age-matched control children. A decrease in T regulatory cell number is also noted during symptom exacerbations. The T regulatory cells play a major role in preventing autoimmunity, and the observation of decreased numbers offers an interesting potential mechanism for the induction of autoimmunity in tic disorders.

Only a limited number of studies have investigated cytokine profiles in patients with tic disorders and/or OCD. Interleukin-12 and tumor necrosis factor-α concentrations are found to be elevated in pediatric patients, compared with healthy controls. Both of these markers appear to be further increased during periods of symptom exacerbation. These findings suggest that symptom exacerbations are associated with an inflammatory process propagated by systemic and local cytokine synthesis that might involve the central nervous system. This is compatible with an autoimmune hypothesis of tic disorders and pediatric-onset OCD. In contrast, in adult patients with OCD, tumor necrosis factor-α has generally been reported to be lowered.

Children fulfilling PANDAS criteria have also been demonstrated to have increased levels of soluble intercellular adhesion molecule-1, but not vascular cell adhesion molecule-1 and E-selectin. This also points to the involvement of systemic inflammation and supports the hypothesis of an ongoing immune-mediated process in PANDAS. Finally, in a peripheral leukocyte-based gene expression study, tic disorders were found to be characterized by several gene expression clusters, including one cluster in which six genes were overexpressed, all of which were associated with immune cell function. This had been assessed by microarrays, which provide the opportunity to analyze thousands of genes at once.

To summarize, alterations in a number of immune parameters, including B and T cell subsets, cytokine profiles, and adhesion molecules, in patients with a tic disorder and/or OCD could be in line with inflammatory processes and/or autoimmunity. Still, as yet, these have no place in the clinical assessment of patients with (suspected) PANDAS.

Treatment

As described in case studies, a number of immunomodulatory interventions have been applied with some

success in children fulfilling PANDAS criteria. These include immunosuppression with corticosteroids, therapeutic plasmapheresis, or intravenous immunoglobulins. In some of these case studies, after successful plasmapheresis, attempts have been made to prevent new GABHS infections with penicillin prophylaxis.

There has been one placebo-controlled study on the efficacy of immunomodulatory therapy. In that study, 30 children meeting PANDAS criteria were randomly assigned to either plasmapheresis, intravenous immunoglobulins, or a placebo condition (saline solution, given in the same manner as intravenous immunoglobulins), and subsequently each patient was given a regimen of penicillin prophylaxis. After 1 month, the intravenous immunoglobulins and plasma-exchange groups showed striking improvements in severity of OCD symptoms, anxiety, and overall functioning. Tic symptoms were significantly improved by plasma exchange only. The children in the placebo condition did not show any amelioration. Interestingly, these improvements were maintained at 1 year after treatment for both plasmapheresis and intravenous immunoglobulins.

Another treatment option could be to treat or prevent streptococcal infections with antibiotics. One-year treatment with antibiotics, either azithromycin or penicillin, appeared to decrease the number of neuropsychiatric symptoms in a group of children with PANDAS when compared to the year preceding antibiotic treatment. As a final possible treatment option, a case report of two brothers has been described, one with OCD and the other with a tic disorder; both improved significantly after undergoing adenotonsillectomy for treatment of their recurrent tonsillitis. Currently, however, immune-based treatments should not be routinely given. Antibiotic treatment may be associated with development of resistance. Also, plasma exchange and intravenous immunoglobulins are invasive treatments that should only be considered in controlled trials. The best clinical approach for patients fulfilling PANDAS criteria is to consider treating troublesome tics with antipsychotic agents such as risperidone, or with behavioral therapy, and to treat OCD symptoms with behavioral therapy.

Conclusion

Currently, only circumstantial evidence points to a possible role for autoimmunity in tic disorders and pediatric OCD. To classify a disease as autoimmune, transmissibility of symptoms to animals through passive transfer of antibodies or T lymphocytes should be demonstrated. Thus, demonstration of symptom induction through anti-basal-ganglia antibodies in animal models remains a high research priority. At present, results have been conflicting in this area. In a fascinating recent study, however, a subset of mice that were immunized and boosted with a streptococcal homogenate in Freund's adjuvant was demonstrated to exhibit motoric and behavioral disturbances in association with the presence of serum antibodies that were immunoreactive to several brain regions, including globus pallidus and thalamus. This was not the case with mice that were boosted with Freund's adjuvant only. These results are consistent with the hypothesis that immune response to streptococci can result in motor and behavior alterations through antistreptococcal antibodies cross-reactive with brain components.

Large-scale longitudinal data are needed to more fully investigate the role of infections with regard to symptom fluctuation. All in all, tic disorders and pediatric OCD are model neuropsychiatric disorders in which to investigate the complex relationship between brain functioning, autoimmunity, and human behavior, in an interplay between genetic and environmental factors. Ultimately, it is hoped that unraveling the genetic background of tic and related disorders will lead to a better understanding of environmental factors and more targeted treatment.

See also: Antipsychotic Drugs; Autoimmune Autonomic Neuropathy; Autonomic Nervous System: Respiratory Control; Executive Functions: Eye Movements and Neuropsychiatric Disorders; Obsessive–Compulsive Disorder; Tourette Syndrome as a Neurotransmitter Disorder; Tourette's Disorder.

Further Reading

Garvey MA, Giedd J, and Swedo SE (1998) PANDAS: The search for environmental triggers of pediatric neuropsychiatric disorders. Lessons from rheumatic fever. *Journal of Child Neurology* 13: 413–423.

Giovannoni G (2006) PANDAS: Overview of the hypothesis. *Advances in Neurology* 99: 159–165.

Harris K and Singer HS (2006) Tic disorders: Neural circuits, neurochemistry, and neuroimmunology. *Journal of Child Neurology* 21: 678–689.

Hoekstra PJ, Kallenberg CG, Korf J, et al. (2002) Is Tourette's syndrome an autoimmune disease? *Molecular Psychiatry* 7: 437–445.

Hoekstra PJ and Minderaa RB (2005) Tic disorders and obsessive–compulsive disorder: Is autoimmunity involved? *International Review of Psychiatry* 17: 497–502.

Murphy TK, Goodman WK, Ayoub EM, et al. (2000) On defining Sydenham's chorea: Where do we draw the line? *Biological Psychiatry* 47: 851–857.

Murphy TK, Husted DS, and Edge PJ (2006) Preclinical/clinical evidence of central nervous system infectious etiology in PANDAS. *Advances in Neurology* 99: 148–158.

Murphy TK, Petitto JM, Voeller KK, et al. (2001) Obsessive compulsive disorder: Is there an association with childhood streptococcal infections and altered immune function? *Seminars in Clinical Neuropsychiatry* 4: 266–276.

Murphy TK, Sajid MW, and Goodman WK (2006) Immunology of obsessive–compulsive disorder. *Psychiatry Clinics of North America* 29: 445–469.

Swedo SE, Leonard HL, Garvey M, et al. (1998) Pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections: Clinical description of the first 50 cases. *American Journal of Psychiatry* 155: 264–271.

Snider LA and Swedo SE (2003) Post-streptococcal autoimmune disorders of the central nervous system. *Current Opinion in Neurology* 16: 359–365.

Snider LA and Swedo SE (2004) PANDAS: Current status and directions for research. *Molecular Psychiatry* 9: 900–907.

Trifiletti RR and Packard AM (1999) Immune mechanisms in pediatric neuropsychiatric disorders. Tourette's syndrome, OCD, and PANDAS. *Child and Adolescent Psychiatry Clinics of North America* 8: 767–775.

Relevant Website

http://intramural.nimh.nih.gov – Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal Infections, National Institute of Mental Health.

Panic Disorder

L E Heuer, S J Mathew, and D S Charney, Mount Sinai School of Medicine, New York, NY, USA

Published by Elsevier Ltd.

Introduction

Panic disorder (PD) is a relatively common anxiety disorder (1–3% prevalence in epidemiological surveys) marked by recurrent, often unexpected panic attacks, which are surges of rapid-developing anxiety in a crescendo pattern. The nature of attacks in PD are heterogeneous, but are generally marked by palpitations, chest pain or pressure, dyspnea, and cognitive, neurological, or gastrointestinal symptoms. Besides the discrete attacks, marked anticipatory anxiety about having future attacks, behavioral avoidance of situations which might provoke attacks (such as bridges, elevators, enclosed spaces), and psychological concerns about 'going crazy' or 'losing one's mind' are commonly observed. Lifetime prevalence estimates are 22.7% for panic attacks, 3.7% for PD without agoraphobia, and 1.1% for PD with agoraphobia. Herein, we review the clinical features of PD, the neurobiology of PD, and therapies presently available for PD.

Clinical Presentation

According to the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (DSM-IV), a patient with PD has recurrent and unexpected panic attacks (Table 1). A minimum of one of the attacks must have been followed by 1 month (or more) of one (or more) of the following symptoms: persistent concern about having additional attacks, worry about the implications of the attack or its consequences, and/or a significant change in behavior related to the attacks. The panic attacks must not be due to effects of a substance of abuse (e.g., stimulant) or general medical condition (e.g., hyperthyroidism), and the panic attacks are not better accounted for by another psychiatric disorder, such as posttraumatic stress disorder, social phobia, and specific phobia.

It has been suggested that the current DSM-IV categorical system fails to capture the full spectrum of subclinical and atypical manifestations associated with PD and implies that if an individual fails to meet the full criteria for a 'disorder' he or she has no disorder. Recent research has supported the view that subthreshold disorders might be as significant as those meeting full DSM criteria and can lead to a decrease in quality of life.

Assessment Instruments

There are several validated instruments for the assessment of panic and related symptoms. The Panic and Agoraphobia Scale (PAS) is a widely used instrument that contains five subscales – panic attacks, avoidance, anticipatory anxiety, disability, and health worries – found to reduce quality of life in PD patients. The application of the PAS in a double-blind-placebo-controlled PD trial demonstrated its sensitivity to detect differences between panic treatments. Another measurement tool, the Panic Disorder Severity Scale (PDSS), is a seven-item instrument assessing each core symptom of DSM-IV symptoms of PD, with or without agoraphobia, as well as work and social impairment. Items include frequency of panic attacks and limited-symptom episodes, distress caused by panic and limited-symptom episodes, anticipatory anxiety, agoraphobic fear/avoidance, and work and social impairment.

In addition to the clinician-administered PDSS, a self-report version of the PDSS was found to have good test–retest reliability, good sensitivity to change in treatment, and a reliable format, thus providing another useful tool for clinical and research settings. Other tools which can be used in busy outpatient primary care offices include the Hospital Anxiety and Depression Scale, the Patient Health Questionnaire, and a simple screening question ("In the last four weeks, have you had an anxiety attack – suddenly feeling fear or panic?"). Finally, the Anxiety Sensitivity Index (ASI) is a 16-item scale developed to measure 'fear of fear,' that is, the degree to which a person believes that physical symptoms of anxiety have negative consequences. Anxiety sensitivity is a construct that has been specifically linked to PD. Persons with high ASI scores tend to attribute physical signs of arousal as representing a serious illness (e.g., palpitations signaling imminent cardiac arrest) rather than a more benign cause (palpitations due to drinking coffee).

Panic-Agoraphobic Spectrum

The Structured Clinical Interview for Panic-Agoraphobic Spectrum (SCI-PAS), a 114-item tool assessing eight primary domains, was developed in accordance with the view that PD patients have heterogeneous domains of symptomatology. The first domain is that of separation sensitivity, which is highly correlated

Table 1 Diagnostic criteria for PD with/without agoraphobia

A. Both 1 and 2:
  1. Recurrent unexpected panic attacks
  2. At least one of the attacks has been followed by 1 month (or more) of one (or more) of the following:
     (a) Persistent concern about having additional attacks, (b) worry about the implications of the attack or its consequences (e.g., losing control, having a heart attack, 'going crazy'), (c) a significant change in behavior related to the attacks
B. Presence/absence of agoraphobia
C. The panic attacks are not due to the direct physiological effects of a substance (e.g., a drug of abuse, a medication) or a general medical condition (e.g., hyperthyroidism)
D. The panic attacks are not better accounted for by another mental disorder, such as social phobia (e.g., occurring on exposure to feared social situations), specific phobia (e.g., on exposure to a specific phobic situation), obsessive–compulsive disorder (e.g., on exposure to dirt in someone with an obsession about contamination), posttraumatic stress disorder (e.g., in response to stimuli associated with a severe stressor), or separation anxiety disorder (e.g., in response to being away from home or close relatives)

Adapted from First MB (2000) *Diagnostic and Statistical Manual – Text Revision (DSM-IV-TR)*. Arlington, VA: American Psychiatric Publishing Group.

with separation anxiety in childhood. Other features include the patients' uneasiness with sleeping alone or away from home, reluctance to leave their home for vacation or work, fearing harm to their family members, and developing highly dependent, unusually intense, close relationships. The second domain is paniclike symptoms. With this domain, patients report panic attacks without fear, only experiencing physical symptoms, such as rapid heart rate, chest tightness, and sweating. In addition, the patient may change their emotional life and behavior in ways that can cause the development of an anxious, dependent, avoidant personality. Stress sensitivity is the third domain, in which a patient will experience an onset of panic symptoms after a period of high tension or after exiting a stressful situation, rather than during the stressor. This may result in the person avoiding stressors, negative news, and resisting change in their daily routine. The fourth domain is medication and substance abuse sensitivity, which may manifest in the patient's resistance to taking psychoactive drugs. In addition, these patients seem to have increased sensitivity to ingested substances (over-the-counter medications, street drugs, caffeine) and increased sensitivity to withdrawal (benzodiazepines, alcohol). The person may develop a persistent state of vigilance or alertness even in the absence of frequent panic attacks, known as anxious expectation, which is the fifth domain. Agoraphobia is the sixth domain and may be marked by a fear of being home alone, in a public square or mall, traveling through a tunnel or over a bridge, or being in crowded places. Additionally, less obvious, 'atypical' symptoms of agoraphobia should be accounted for. These may include perceived obstacles to breathing (ties, turtlenecks), fear of natural phenomena (thunderstorms, high winds, flooding, earthquakes), and fear of humiliation if panic symptoms should arise, all of which can interfere with social or work relationships. Domain seven is illness phobia and hypochondriasis, usually manifesting as a fear of bodily sensations. The last domain, or domain eight, is reassurance orientation. Patients that fit into this domain will experience immediate relief from anxiety when reassured by 'protective' figures.

Course of Illness

PD, much like major depressive disorder, is generally episodic in nature, although approximately 20–25% percent of patients have a more pernicious and unremitting course. In one longitudinal study over a course of 11 years, it was found that many patients recover from panic attacks and disabilities, and some will achieve full remission. Specifically, 66.7% reported no panic attacks during the year before follow-up, 87.5% reported no panic attacks the month before follow-up, 54% reported no or only mild phobic avoidance, and 33% of the patients were completely remitted according to a composite remission criterion. Another study examined the long-term follow-up (2–14 years, median 8 years) of 132 patients with PD with agoraphobia successfully treated by exposure therapy alone. Two hundred patients satisfying the DSM-IV criteria for PD with agoraphobia were originally treated with a standard therapy protocol consisting of 12 sessions of psychotherapy and exposure homework. One hundred and thirty-six patients became panic free following this treatment and 132 of these patients were available for followup. The study found a high probability of remaining in remission 2–14 years following treatment. Specifically, three out of four patients remained well after 7 years. Positive associations with remission included a younger patient and the absence of any personality disturbances or residual agoraphobia. Higher initial levels of depressed mood and use of antidepressant drugs were related to a poor outcome.

Using observational/longitudinal data from the Harvard/Brown Anxiety Disorders Research Program,

recurrence of panic symptoms after remission in women were compared to men. Although the rates of remission at 5 years for patients with PD with or without agoraphobia were equivalent in both men and women (39%), 25% of these women and 15% of these men reexperienced symptoms in the following 6 months. At the 8-year follow-up, cumulative remission rates were equivalent among men and women. However, relapse events for uncomplicated panic were threefold higher in women than in men by 8 years.

Neurochemistry and Neuroimaging of PD

Varied neurotransmitter and neuropeptide systems have been associated with PD, most notably the norepinephrine (NE), serotonin (5HT), and γ-aminobutyric acid (GABA) systems. Pharmacological probes of the α-adrenergic receptor have served a useful role in developing neurobiological models of PD for more than three decades. Most of these studies have used adrenergic antagonist (yohimbine) or agonist (clonidine) medication, and have compared biochemical, physiologic, and behavioral responses between patient groups and healthy volunteers. Administration of yohimbine, an α-2-adrenergic presynaptic antagonist which increases the release of NE, generally results in increased plasma NE metabolites, blood pressure, heart rate, and anxiety responses in PD patients versus healthy volunteers. The increased sensitivity has been theorized to be attributed to the α-2-adrenergic autoreceptor being less sensitive to the inhibitory feedback of NE in the synapse. Several groups have studied platelet membranes as models for α-2-adrenoreceptor activity. In one study, the α 2-adrenoreceptor density in platelets, as reflected by binding of [^{3}H]yohimbine, was not different between PD and healthy subjects. The study did find, however, a decreased platelet α-2-adrenoreceptor affinity in the PD patients, a finding which has been replicated in subsequent trials. These findings may be due to either increased postsynaptic adrenergic receptor density, enhanced coupling to G-protein, or abnormal postreceptor signal transduction mechanisms.

Multiple studies have demonstrated a blunted growth hormone (GH) response in patients with PD to clonidine, an α-2-adrenoreceptor agonist, possibly reflecting the subsensitivity (increased postsynaptic adrenergic receptor density, enhanced coupling to G-protein, or abnormal postreceptor signal transduction mechanisms) of the postsynaptic α-2-adrenergic receptor. Additionally, blunted GH response in patients with PD was also found with GH releasing factor. A recent investigation examined the question of whether an overall hypoactive GH system, rather than a specific α-2-adrenergic abnormality, was responsible for the blunted GH response to clonidine. The results did not support a reduced basal secretory GH activity, thus increasing the likelihood that a reduced sensitivity of α-2-adrenoreceptors is a mechanism for these findings.

Regarding GABA, the major inhibitory amino acid neurotransmitter, PD patients were found to have regional and/or global reductions in GABA. Preclinical pharmacological models of anxiety have observed that deficits in GABA neuronal function have been associated with anxiogenesis, whereas enhancement of GABA function tends to be anxiolytic. In a study using proton magnetic resonance spectroscopy (MRS), patients with PD had a 22% reduction in total occipital cortex GABA concentrations compared with controls. Central benzodiazepine (BZD) receptor binding has also been assessed in PD. Using single-photon computed tomography (SPECT), a reduced uptake of the selective BZD receptor ligand iomazenil was demonstrated in patients with PD relative to the control group. Additionally, using flumazanil and positron emission tomography (PET) scanning, a global reduction in BZD binding sites in PD patients relative to healthy controls was found.

The 5HT-1A receptor has been investigated in PD, building on previous work demonstrating lower 5HT-1A receptor density in patients with major depressive disorder. One PET study evaluated the volume of distribution of a radioligand for this receptor in unmedicated patients with PD and healthy controls. It was found that patients with PD as compared to controls had a significantly lower volume of distribution in the anterior cingulate, posterior cingulate, and raphe portions of the brain.

Few morphometric studies have been conducted in PD. Magnetic resonance imaging (MRI) studies have found morphometric deficits in temporal lobe, with normal hippocampal volume. A recent study using voxel-based morphometry techniques found increased gray matter volume in the midbrain and rostral pons of brain stem, potentially supporting models of PD that hypothesize that the panic attack stems from brain stem loci.

Genetics of PD

PD has a complex genetic contribution, with an estimated heritability of 30–46%. First-degree relatives of patients with PD have an eightfold increased risk of PD, and there is an emerging literature regarding specific genes implicated in PD. The adenosine 2A receptor gene (*ADORA2A*) is one of two genes to be

implicated in multiple studies in PD as susceptibility loci, with the other being catechol-*O*-methlytransferase gene (*COMT*). Regarding adenosine, caffeine is an antagonist to the adenosine 2A receptor, and has long been known to be anxiogenic in PD. A linkage study analyzed five DNA sequence variations in *ADORA2A*, in a large sample of PD pedigrees (70 pedigrees and 83 child–parent trios). Linkage analysis revealed elevated logarithm of odds (lod) scores for the silent substitution (1976C/T) in the second coding exon, which was consistent with an earlier association study showing an association between the 1976T/T genotype and increased vulnerability to PD. Recent studies have found an association between the V158M *COMT* polymorphism and PD, with a single nucleotide polymorphism in the *COMT* gene leading to a three-to fourfold difference in enzymatic activity. Findings have supported evidence for a susceptibility locus for PD either within the *COMT* gene or in a nearby region of chromosome 22.

A genome scan was performed on 587 individuals in 60 pedigrees segregating PD and bladder/kidney problems and found that there were genes on chromosome 13 (and possibly chromosome 22) that support the existence of a genetic syndrome involving PD and several associated conditions. Common medical conditions observed in family members of PD probands include interstitial cystitis, migraine headaches, thyroid conditions, and mitral valve prolapse. This association may be explained using the model of dysregulation of the noradrenergic system found in PD and the fact that the function of the bladder is regulated through innervation from autonomic nuclei. A genome scan in 120 multiplex PD pedigrees of 1591 individuals was performed, with evidence for linkage in one region on chromosome 15q (near $GABA_A$ receptor subunit genes *GABRB3* and *GABRA5*) and 2q. Several strategies are currently being pursued, including fine mapping of promising regions, investigation of endophenotypes (including gene expression patterns and candidate gene studies), and development of methods to investigate gene–gene interactions.

Therapies for PD

Both medications and psychotherapies, such as cognitive behavioral therapy (CBT), have demonstrated efficacy for PD. Selective serotonin reuptake inhibitors (SSRIs) are considered to be the first-line treatment for PD, although serotonin and norepinephrine reuptake inhibitors (SNRIs) such as venlafaxine XR have recently gained US Food and Drug Administration (FDA) approval.

Three classes of medications have received extensive evaluation in PD, namely SSRIs, tricyclic antidepressants, and benzodiazepines. Each have been found equally effective in treating panic and associated anxiety symptoms, as well as demonstrating a clinically significant response in decreasing the frequency of panic attacks in 50–80% of patients. Safety and side effect profile have therefore become a major determinant when deciding upon treatment for an individual patient. Due to their safety and overall tolerability, SSRIs are recommended as first-line therapy in the treatment of PD. Starting these drugs at low doses with a slow titration helps to minimize initial side effects, which can include nausea, anorexia, tremors, anxiety, sexual dysfunction, jitteriness, and insomnia. If a patient does not respond to an initial trial of SSRI therapy, clinical experience suggests that trying an alternative SSRI may provide the desired result. If there is still no adequate response, switching to another class of drugs, such as tricyclic antidepressants, benzodiazepines, or SNRIs, is recommended (venlafaxine XR). With regard to long-term administration of benzodiazepines, there is concern with respect to dependence, medication abuse, and the rapid reemergence of symptoms after discontinuation. Therefore, benzodiazepines are used most appropriately concurrently with SSRIs at the beginning of treatment to hasten the antipanic response, and for as-needed use for breakthrough anxiety.

CBT is a form of psychotherapy in which a patient receives a limited number of sessions (usually 12–16) over a period of approximately 3–4 months. It focuses on recreating the feared symptoms and then modifying the patient's usual response to these symptoms. One large multicenter trial compared medication and CBT and their combination in the treatment of PD with or without agoraphobia. Three hundred and twelve patients meeting DSM-III-R criteria for PD were randomly assigned to receive imipramine, CBT, placebo, CBT plus imipramine, or CBT plus placebo. Although the follow-up period was limited to 6 months after treatment discontinuation, the intent to treat results demonstrated a 41% response rate for CBT plus placebo, 31.9% for CBT alone, 19.7% for imipramine alone, 13% for placebo, and 26.3% for CBT combined with imipramine. There was no significant difference found between CBT alone and imipramine alone, combined treatment was found superior to CBT and CBT plus placebo, and responders to imipramine, with or without CBT, fared significantly worse in the no-treatment follow-up period than did those who received either CBT alone or CBT plus placebo. These results support the durability of CBT over a 6-month follow-up

period, and suggest high rates of relapse following cessation of medication use in PD.

See also: Anxiety Disorders; Aversive Emotions: Genetic Mechanisms of Serotonin; Gamma-Aminobutyric Acid (GABA); Genetics of Human Anxiety and Its Disorders; Panic Disorder as an Emotional Disorder; Psychotherapeutic Approaches to Psychiatric Disorders; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology.

Further Reading

Caspi A and Moffitt TE (2006) Gene-environment interactions in psychiatry: Joining forces with neuroscience. *Neuroscience* 7: 583–590.

Charney DS (1985) Increased anxiogenic effects of caffeine in panic disorders. *Archives of General Psychiatry* 42: 233–243.

Charney DS and Heninger GR (1986) Abnormal regulation of noradrenergic function in panic disorders. Effects of clonidine in healthy subjects and patients with agoraphobia and panic disorder. *Archives of General Psychiatry* 43: 1042–1054.

Charney DS, Woods SW, Krystal SW, et al. (1992) Noradrenergic neuronal dysregulation in panic disorder: The effects of intravenous yohimbine and clonidine in panic disorder patients. *Acta Psychiatrica Scandinavica* 86: 273–282.

Coplan JD, Liebowitz MR, Gorman JM, et al. (1992) Noradrenergic function in panic disorder. Effects of intravenous clonidine pretreatment on lactate induced panic. *Biological Psychiatry* 31: 135–146.

First MB (2000) *Diagnostic and Statistical Manual – Text Revision (DSM-IV-TR)*. Arlington, VA: American Psychiatric Publishing Group.

Fyer AJ, Hamilton SP, Durner M, et al. (2006) A third-pass genome scan in panic disorder: Evidence for multiple susceptibility loci. *Biological Psychiatry* 60: 388–401.

Goddard AW, Mason GF, Appel M, et al. (2004) Impaired GABA neuronal response to acute benzodiazepine administration in panic disorder. *American Journal of Psychiatry* 161: 2186–2193.

Hamilton SP, Fyer AJ, Durner M, et al. (2003) Further genetic evidence for a panic disorder syndrome mapping to chromosome 13q. *Proceedings of the National Academy of Sciences of the United States of America* 100: 2550–2555.

Hamilton SP, Slager SL, DeLeon AB, et al. (2004) Evidence for genetic linkage between a polymorphism in the adenosine 2A receptor and panic disorder. *Neuropsychopharmacology* 29: 558–565.

Katon WJ (2006) Panic disorder. *New England Journal of Medicine* 354: 2360–2367.

Klein DF (1993) False suffocation alarms, spontaneous panics, and related conditions. An integrative hypothesis. *Archives of General Psychiatry* 50: 306–317.

Neumeister A, Bain E, Nugent AC, et al. (2004) Reduced serotonin type 1A receptor binding in panic disorder. *Journal of Neuroscience* 24: 589–591.

Shear MK, Cassano GB, Frank E, et al. (2002) The panic-agoraphobic spectrum: Development, description, and clinical significance. *Psychiatric Clinics of North America* 25: 739–756.

Weissman MM, Fyer AJ, Haghighi F, et al. (2000) Potential panic disorder syndrome: Clinical and genetic linkage evidence. *American Journal of Medical Genetics* 96: 24–35.

Panic Disorder as an Emotional Disorder

G M Sullivan, Columbia University College of Physicians and Surgeons, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Panic disorder (PD) is a disabling psychiatric disorder characterized by paroxysmal unexpected panic attacks and ongoing anxiety regarding the attacks. A panic attack is a discrete episode of abruptly developing fear, terror, or extreme discomfort accompanied by a varied constellation of intense physical symptoms, distortions of perception, and emotional distress. This may include palpitations or heart racing, sweating, trembling, dyspnea, gastrointestinal distress, flushing, dizziness, paresthesias, derealization, depersonalization, and the fear one is losing control, one is going crazy, or that death is imminent. Respiratory symptoms and hyperventilation are prominent in some patients, as can be a powerful urge to flee from where the attack is occurring. By definition, the panic symptoms crescendo within 10 min. They often abate or resolve in 20–30 min, although they can last significantly longer. Panic attacks occur in other psychiatric conditions such as posttraumatic stress disorder; and approximately 10% of individuals with no psychiatric diagnosis experience at least one panic attack in their lifetime. Requisite for the diagnosis of PD is that at least one of the panic attacks is followed by at least 1 month of persistent concern about future attacks, irrefutable catastrophic worry about what the attacks portend (e.g., cardiac disease or insanity), and/or significant changes in behavior related to the attacks. In approximately 25% of patients with PD, there is also the development of agoraphobia, which is fear or distress about situations in which escape might be difficult or embarrassing or help may be unavailable should a panic attack develop. Typically, agoraphobic fears and the resulting avoidant behaviors center on being outside the home alone; being in crowds, in a line, on a bridge; or traveling within a vehicle.

The age of onset of PD is usually between late adolescence and the mid thirties. Women are over two times more likely to develop PD than men. Lifetime prevalence is 2–5%, and 1 year prevalence is 1–3%. Comorbidity with other anxiety, mood, and substance use disorders is common. Family and twin studies suggest that genetic factors play a role in the pathophysiology of PD, and heritability estimates approach 50%.

The term panic is derived from the name of the Greek god Pan, who was said to induce terror in travelers of the forests of ancient Greece by repeatedly rustling bush along their path while remaining hidden from view. Panic attacks have been described in the medical literature since at least the American Civil War, when Jacob Mendez DaCosta coined the term irritable heart for a constellation of severe cardiac-like symptoms in the absence of cardiac pathology. Later, terms such as neurasthenia and anxiety neurosis were used for PD. With the discovery of the panic-ablating effects of the tricyclic antidepressant imipramine in the 1960s, PD was proposed to be a separate anxiety condition. It was not until 1980 that PD became established as an officially recognized psychiatric diagnosis, ushering in a boom of research aimed at understanding the neurobiology underlying PD. This article summarizes the current understanding of the neurobiology of PD as derived from clinical studies and basic research efforts.

Challenge Studies

Panic attacks are particularly amenable to study in the laboratory setting, in large part because they are discrete, time-limited events and can be artificially induced by the administration of particular chemical agents. The term challenge study refers to the research procedure in which a panic-eliciting agent, or panicogen, is administered to PD patients and comparison subjects so that the behavioral, autonomic, and endocrine responses may be measured. An array of remarkably diverse chemical panicogens, which precipitate panic attacks in most PD patients but only a small subset of controls, has been identified. Intravenous sodium lactate was the first panicogen to be identified, followed by list that includes inhaled CO_2; noradrenergic (NA) agents, including isoproterenol, yohimbine, and epinephrine; serotonergic agents, including fenfluramine and *m*-chlorophenylpiperazine (mCPP); caffeine, an adenosine 2A antagonist; doxapram, a respiratory stimulant; cholecystokinin (CCK)-derived peptides CCK-4 and pentagastrin; and the γ-aminobutyric acid ($GABA_A$) antagonist flumazenil. One of the realizations that came from the variety of challenge studies is the remarkable pharmacodynamic diversity among the panicogens. Yet preclinical work with these panicogens points to cohesive elements, in terms of common activation of central pathways, potentially explaining the panicogenic effects (see the section titled 'Anxiety sensitivity theory of panic disorder').

Heuristic Models for Understanding Panic Disorder Neurobiology

Modern Learning Theory of Panic Disorder

One model of PD is based on modern learning theory and posits the involvement of neural circuits that mediate conditioned fear and avoidance. Panic attacks are considered normal but misfired fight-or-flight responses, activated in the absence of a proximate threat. PD develops as a result of conditioning. In learning theory terminology, when the adverse unconditioned stimulus (US), the panic attack, is paired with previously innocuous exteroceptive (environmental) and/or interoceptive (visceral) conditioned stimuli (CS), an association is learned such that the CS alone can trigger anxiety and panic when reencountered. Conditioning to exteroceptive stimuli therefore can result in agoraphobic behaviors. Conditioning to interoceptive cues can result in seemingly spontaneous or out-of-the-blue panic attacks or heightened anxiety on future encounters with similar visceral stimuli. The elicitation of this conditioned anxiety may potentiate further panic attacks in a vicious cycle, leading to PD.

The neural correlates of conditioned fear are well described. For example, the pathway mediating auditory fear conditioning in the rat involves pathways by which somatosensory information about the US (pain) and auditory information about the CS (tone) converge on the temporal lobe structure known as the amygdala. Relayed by thalamic and cortical processing regions, neural signals representing the US and CS pathways meet on single cells in the lateral nucleus of amygdala (LA). The LA connects directly and indirectly with the central nucleus of amygdala (CE), considered the output station of the amygdala. The outputs from the CE project to diverse brain regions that control components of the fear response. Those to the periaqueductal gray (PAG) mediate freezing and analgesia; those to the lateral hypothalamus (LH), parabrachial nucleus (PBN), and locus coeruleus (LC) mediate autonomic changes (cardiovascular and respiratory); and those to the paraventricular nucleus of hypothalamus (PVN) mediate hypothalamic–pituitary–adrenal (HPA) axis activation. Distinct from its role in explicit (conscious declarative) memory, the hippocampus provides CS information to the amygdala about the context in which the adversity is occurring. Extinction is the process whereby the learned conditioned response (CR) is weakened or dissipated, such as by repeated CS presentations nonreinforced by the US. Extinction appears to involve interactions between the prefrontal cortex (PFC) and subcortical brain areas, including the amygdala. Through extinction, the ability of the CS to control CRs via the LA-to-CE communication becomes regulated by the ventromedial PFC. Activity between these regions may also be modulated by contextual information provided by the hippocampus. The modern learning theory hypothesis of PD is consistent with a notion of a failure to limit the amygdala-mediated fight-or-flight responses by the PFC and/or hippocampus. Figure 1 illustrates how the fear circuit may be central to PD pathophysiology.

False Suffocation Alarm Hypothesis of Panic Disorder

The false suffocation alarm hypothesis focuses on the panic attack as the aberrant response in PD. The hypothesis assumes the existence of an early evolved brain monitoring and alarm system for usable air. When tripped by poor air, emergency response escape behaviors are activated. The panic attack is the alarm response tripped in absence of poor air. The distinction between panic and conditioned fear is emphasized, noting that HPA axis activation is an element of the fear response, whereas in panic attacks it is not requisite. Acute anxiety is viewed as a more recently evolved response involving anticipation of danger and deriving from fear. Potential neural correlates are suggested by Ondine's curse, a disorder involving deficient activation of the brain stem metabolic respiratory system. This condition, which has the opposite defect proposed in PD, is associated with a deficiency in cholinergic neurotransmission in the arcuate nucleus of hypothalamus. In this light, it is of interest that pathologically enhanced muscarinic cholinergic neurotransmission has been proposed as underlying the hypersensitivity to suffocation stimuli and hypercapnia in PD.

Anxiety Sensitivity Theory of Panic Disorder

This model posits that particular individuals have a cognitive predisposition to interpret the arousal symptoms of anxiety as harmful beyond the inherent discomfort or displeasure experienced. Anxiety symptoms are interpreted as threatening to their physical, psychological, and/or social well-being, and the harm they cause may accumulate over time. PD patients score high on measures of anxiety sensitivity, such as the widely employed Anxiety Sensitivity Index (ASI). Notably, an ASI score is a good predictor of panic to a panicogen. In this model, the enhanced response of PD patients to challenge with a panicogen stems from catastrophic misinterpretations about the induced arousal symptoms rather than direct stimulation of a biochemical defect. Yet the anxiety sensitivity model is not inconsistent with a neurobiological explanation involving an enhanced sensitivity to afferent viscerosensory input transmitted via the glossopharyngeal and vagus nerves. These viscerosensory

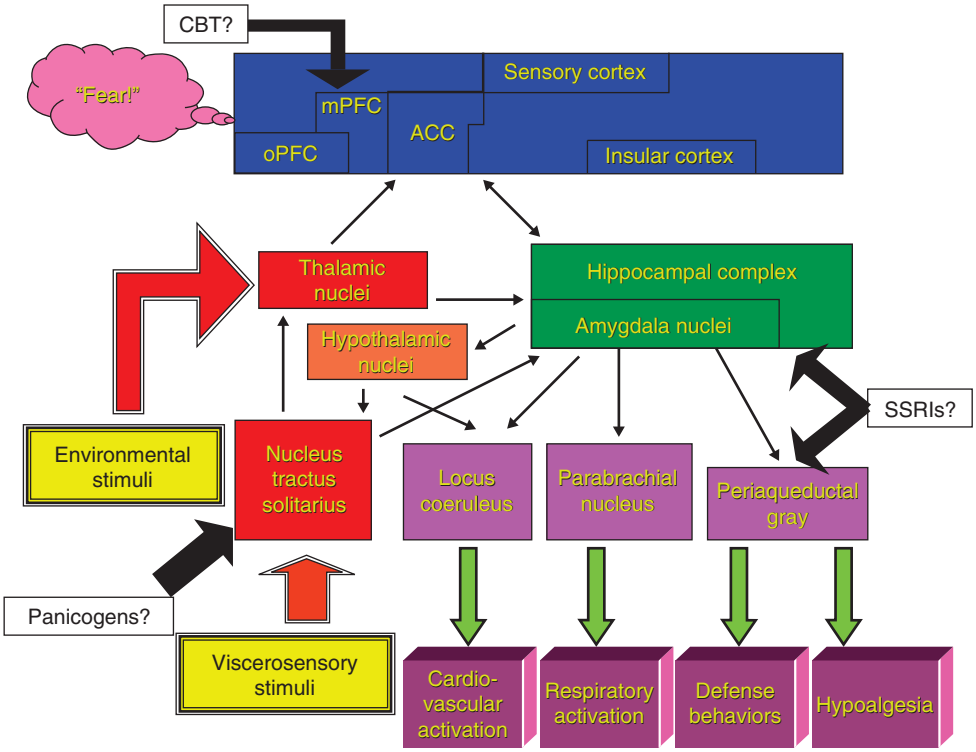

Figure 1 Hypothetical neurocircuitry of panic disorder. The proposed circuit for panic disorder centers on amygdala-mediated processing of environmental and viscerosensory stimuli (yellow). The stimuli are transmitted to amygdala directly by thalamus and nucleus tractus solitarius (red) and indirectly via cortical processing areas including the sensory, insular, and anterior cingulate cortices (blue). Contextual information is relayed from the cortex via hippocampus. The amygdala also influences attention, perception, and memory during heightened anxiety by connections with cortical areas such as the orbital prefrontal cortex (oPFC) and anterior cingulate cortex (ACC; blue). A panic attack results when the amygdala activates, in parallel, hypothalamic (brown) and brain stem nuclei (purple) that mediate particular components of the fight-or-flight response (dark purple). Panicogens may cause panic attacks through a common mechanism of enhancing activity in viscerosensory pathways mediated through nucleus tractus solitarius. The selective serotonin reuptake inhibitors (SSRIs) may act by enhancing serotonergic neurotransmission to the amygdala, hippocampus, and periaqueductal gray. Cognitive behavioral therapy (CBT) may have its effects by enhancing the inhibitory influence of the medial prefrontal cortex (mPFC) on the amygdala.

pathways synapse in the nucleus tractus solitarius (NTS), and the signals are relayed to the amygdala and hypothalamus, as well as the PBN and LC, potentially lowering the threshold for a fight-or-flight response. In this light, it is notable that an action common to most the panicogens is the activation of the viscerosensory areas of NTS (see Figure 1). Here too, inadequate top-down control of the amygdala, limiting the impact of cortically processed information provided to it through the hippocampus and/or PFC, could further lower the threshold for an inappropriately triggered fight-or-flight response.

Neurotransmitter Systems and Panic Disorder

Neurotransmitter systems are the mediators and modulators of neurotransmission within all neural circuits. Therefore, specific pathology in a neurotransmitter system could lead to dysfunction in circuitry involved in PD. Yet it is also possible that identified dysfunction in a neurotransmitter system is simply a reflection of changes in circuit functioning caused by a pathology of a completely different etiology.

Noradrenergic System

Both rodent models of stress and fear and human challenge studies point to a role for a dysfunction in the NA system in PD. The LC is the primary brain stem nucleus for NA neurons, and these neurons project to key cortical, limbic, and other brain stem regions involved in visceral sensitivity and fearful responding. There are reciprocal connections involving corticotropin-releasing hormone (CRH)-secreting neurons between the LC and amygdala. The presence of threat increases the firing rate of the LC, resulting in cortical norepinephrine (NE) release and neural activity reflected in electroencephalograms. Laboratory manipulations of the early developmental environment in the rodent, such as in handling and maternal deprivation paradigms, lead to long-term alterations in CRH receptor and α_2-adrenergic autoreceptor

densities in the LC. This suggests a significant environmental sculpting of the stress response system.

In PD patients, clonidine, an α_2-agonist that decreases LC firing by feedback inhibition at α_2-auto receptors, decreases vulnerability to panic on challenge with lactate. Yohimbine, an α_2-antagonist, produces greater paniclike anxiety in PD patients accompanied by a decrease in frontal cortical cerebral blood flow (CBF). This contrasts with the response to yohimbine reported in healthy volunteers in which there is CBF increase in the medial PFC.

Recent application of cardiac imaging and microneurography in PD suggests impaired cardiac sympathetic function due to a central mechanism. Anxiety sensitivity is predictive of lower cardiac NE reuptake, implicating NA system dysfunction in vulnerability to panic. In this light, animal studies showing that chronic imipramine treatment decreases α_2-autoreceptor density in the LC and the A2 region of the NTS suggest that this treatment may work through correction of the effects of adverse early environmental factors on the stress response system. It is also important to emphasize the interdependence of the NA and serotonergic systems. As an example of the tight cross-regulation of these systems, stress-induced sympathetic activity and fight-or-flight defensive behaviors are inhibited by serotonin release in the PAG and rostral ventrolateral medulla (RVLM), in part via actions on serotonin 5-HT_{1A} receptors.

Serotonergic System

Evidence for serotonin system dysfunction in PD is also extensive. Perhaps most compelling is the efficacy of the serotonin selective reuptake inhibitors (SSRIs) for symptoms of PD. It is also notable that a wealth of preclinical experiments have demonstrated a modulating role of serotonin neurotransmission on various fearful and avoidant behaviors. Yet in some anxiety paradigms serotonin is anxiogenic and in others it is anxiolytic, often with site-specific and paradigm-specific effects. A provocative hypothesis for PD reconciles the seemingly contradictory animal data by suggesting opposite roles for serotonin in panic versus anxious anticipation. Raphe projections to the PAG and medial hypothalamus normally inhibit fight-or-flight behaviors, so deficient serotonin neurotransmission is permissive to the panic response. In contrast, raphe projections to the amygdala and hippocampus are proposed to enhance anticipatory anxiety and resilience to chronic stress, respectively, but, here too, deficiency in serotonin signaling is proposed to unbridle a normal restraining influence of this upstream fear circuitry on a PAG-mediated panic response (see Figure 1). Recent studies testing an SSRI on conditioned fear memory in the rodent suggest a similar pattern of effect of SSRIs on conditioned fear as they have on PD symptoms. Acute SSRI administration before training increased conditioned fear behavior, whereas several weeks of SSRI reduced the behavior. In PD patients, SSRIs often increase panic and anxiety when first started, the panic-ablating effect only beginning after several weeks.

Challenge of PD patients with serotonergic agents such as mCPP and fenfluramine produces heightened anxiety and can provoke panic, similar to the acute effects of an SSRI. Abnormal serotonin metabolism in PD is suggested by a sevenfold higher level of the main serotonin metabolite in jugular venous samples from nonmedicated, nonpanicking PD patients. Acute depletion of tryptophan, the amino acid precursor of serotonin, results in SSRI-treated PD patients becoming newly vulnerable to flumazenil-induced panic, suggesting enhanced serotonin neurotransmission is necessary for the protective effects of SSRI treatment. Interestingly, female PD patients homozygous for a functional polymorphism in the transcriptional control region of the serotonin reuptake protein (i.e., the serotonin transporter) have a poor response in terms of ablation of panic attacks to treatment with an SSRI.

A particular receptor for serotonin, the 5-HT_{1A} receptor, is located at presynaptic sites on serotonin neurons of the brain stem raphe nuclei, where it acts as an inhibitory feedback autoreceptor; it is located on postsynaptic neurons in brain regions that receive serotonin terminals projecting from raphe. Experiments with transgenic mice involving site-specific rescue of the 5-HT_{1A} receptor have indicated the importance of the postsynaptic receptors in anxiolysis and the necessity of this receptor during early development for normal serotonergic mediation of anxiety. Of note, when context-conditioned 5-HT_{1A} knockout mice are tested in the same context but with novel cues added, they do not have the normal decrement in fear behavior of the wild-type mice. This suggests the 5-HT_{1A} knockout is biased toward threatening cues in ambiguous environments, analogous to the memory bias of PD patients to panic-related cues in an otherwise neutral environment.

Patients with PD and depression have been shown to have attenuated hypothermic and neuroendocrine responses to challenge agents acting on the 5-HT_{1A} receptor, implying a subsensitivity of the receptor. In the gene coding for 5-HT_{1A}, there is a functional promoter polymorphism, -1019C/G, believed to de-repress 5-HT_{1A} autoreceptor expression, thereby enhancing negative feedback and reducing overall serotonergic neurotransmission. PD patients homozygous for the G-allele have been reported to have less activation of the right PFC regions in response to fearful stimuli. This suggests a dysfunction of 5-HT_{1A}

autoreceptor activity in PD may contribute to deficient top-down control. In other words, the response to threat may be conceded to subcortical fear structures that produce a stereotyped fear response rather than a more appropriate response influenced by complex, cortically processed information provided by the PFC. It must be noted that the G-allele has been associated with PD with agoraphobia but not with PD overall. It would be of interest to know whether the G-allele is associated with enhanced fear conditioning, as might be predicted by the learning theory model of PD. It is notable that in clinical depression the G-allele has been associated with higher cortical postsynaptic 5-HT_{1A} binding by positron emission tomography (PET) and poorer response to antidepressants.

Recent work in healthy volunteers using both PET and functional magnetic resonance imaging (fMRI) demonstrates an inverse relationship between raphe 5-HT_{1A} autoreceptor binding and amygdala reactivity to angry and fearful faces. This suggests another direct link between 5-HT_{1A} autoreceptor functioning and subcortical fear system hyperreactivity. PET radioligand studies assessing 5-HT_{1A} binding in PD indicate lower 5-HT_{1A} binding both in the cortex and brain stem raphe. Somatic and hypochondriacal anxiety, which predominate in PD, inversely correlate with cortical postsynaptic 5-HT_{1A} receptor binding, whereas psychic anxiety, more germane to depression and generalized anxiety disorder, positively correlates with it. PET and fMRI studies in this area are only beginning to be reported and promise to clarify the links among the G-allele, 5-HT_{1A} receptor binding, and fear system reactivity.

GABA System

GABA is the predominant inhibitory neurotransmitter throughout the brain. The two main receptor subtypes for GABA are the $GABA_A$ and $GABA_B$ receptors. $GABA_A$ is a multisubunit receptor with a central ligand-gated chloride ion channel. It can be allosterically modulated due to multiple binding sites for benzodiazepines, neurosteroids, barbiturates, and other compounds. The $GABA_A$ receptor complex comprises five protein subunits, and there are a variety of combinations of subunit types, termed α, β, γ, δ, and ρ, that make up the $GABA_A$ receptor. Murine knockouts of individual subunits have provided clues to the various physiological effects, and particular subunits appear more involved in anxiolysis than others. Interestingly, 5-HT_{1A} knockout mice have been shown to have downregulation in $GABA_A$ α-subunits, reduced $GABA_A$ binding, and benzodiazepine-resistant anxiety, indicating the impact of serotonergic system dysfunction on the GABA system. There is evidence for GABA system dysfunction in PD, although it is not clear whether this is indicative of primary pathology, compensatory changes, or, in some cases, the effects of long-term treatment with benzodiazepines. PET and single photon emission computed tomography (SPECT) studies of benzodiazepine binding in PD indicated decreased global binding as well as specifically decreased binding in the hippocampus and frontal cortical regions, although there are conflicting reports. The higher-potency benzodiazepines, such as alprazolam, clonazepam, and lorazepam, have played a large role in the routine pharmacological treatment of PD, although the potential for abuse of this class of medications has led to a more judicious selection of patients appropriate for this treatment approach. The benzodiazepines also have a role in the first several weeks of treatment with SSRIs when the SSRIs are initially ineffective or paradoxically anxiogenic for some PD patients. GABA reuptake inhibitors, inhibitors of GABA metabolism, and partial agonists at the $GABA_A$ receptor are now being explored as potential anxiolytics in PD and other anxiety conditions.

Glutamate System

Glutamate is the major excitatory neurotransmitter throughout the central nervous system, playing a role in functioning of up to 40% of synapses. The two major classes of glutamate receptors are the metabotropic (second-messenger) glutamate receptors and the ionotropic (ion channel) glutamate receptors, including *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methylisoxazole-4-propionate (AMPA), and kainate receptors. The metabotropic receptors appear to modulate activity of the ionotropic receptors via actions at pre- and postsynaptic sites as well as presence on glial cells and GABAergic neurons. The ionotropic receptors are generally postsynaptic and mediate fast synaptic transmission through sodium- and calcium-permeable ligand-gated ion channels. These receptors are also involved in synaptic plasticity in which activity-induced structural changes in synapses occur through processes dependent on gene transcription and new protein synthesis. Under development as anxiolytics are agents that modulate glutamatergic activity through actions on either metabotropic or ionotropic receptors. There are also several existing anticonvulsant medications that act through glutamatergic mechanisms, and some, such as valproic acid, appear to have some efficacy in the treatment of PD. D-Cycloserine, a partial agonist at NMDA receptors, facilitates extinction of fear memory in the rodent model. Agents such as D-cycloserine have been suggested as useful adjuncts to psychotherapies, such as cognitive-behavioral therapy (CBT), that employ techniques aimed at extinguishing anxious

responses to conditioned interceptive and exteroceptive stimuli. Preliminary trials to date have shown promising results.

Neuropeptide Systems

CRH is a 41-amino-acid peptide that plays a key role in the coordinated behavioral, autonomic, neuroendocrine, and neuroimmune responses to stress. First described for its endocrine role in the HPA axis response, CRH was later found also to act as a neurotransmitter or neuromodulator in widespread brainregions involved in fear and anxiety including the cortical areas, amygdala, bed nucleus of the stria terminalis (BNST), hippocampus, nucleus accumbens, raphe nuclei, and LC. CRH has actions via two receptor subtypes, CRH_1 and CRH_2; and CRH_2 has three functional splice variants, $CRH_{2A–C}$. Blocking activity at CRH_1 produces anxiolytic-like responses, and CRH_1 knockout mice have reduced anxiety-like behaviors. There is also a role for CRH_2 in anxious responding, although it has been shown to be anxiogenic, anxiolytic, and anxio-neutral in different animal models. Therefore, CRH receptors have been prime targets for anxiolytic drug development.

CCK, a peptide found in the gastrointestinal tract and brain, is co-localized with other neurotransmitters, particularly dopamine and GABA. CCK stimulates both cortical and hippocampal neurons via two described receptors, CCK_1 and CCK_2, effects that are blocked by benzodiazepines. In PD, the endogenous CCK level has been reported to be lower in the cerebrospinal fluid (CSF). CCK-tetrapeptide (CCK-4) is cleaved from the N-terminus of CCK. It has selective activity at the CCK_2 receptor and has been shown to bind throughout the fear system regions. CCK-4 is also a potent panicogen. PD patients have reduced CBF in the medial PFC following challenge with the peptide, similar to the effect shown with yohimbine. And, like most panicogens, CCK-4 administered systemically induces cFos-like immunoreactivity in the NTS and several hindbrain regions. Thus, the central CCK receptor system is also a target for therapeutics, and CCK-4 challenge is being used to test potential panic-ablating medications.

Neuropeptide Y (NPY) and galanin are two peptides widely expressed in key stress and fear circuits. Both are co-localized with NE, and galanin is also co-localized with serotonin. Both peptides have anxiolytic-like activity, apparently through modulation of the monoaminergic systems. Similarly, substance P is a peptide found in brain and sensory afferents and is co-localized with glutamate, serotonin, and acetylcholine. Substance P is released under aversive conditions. It binds to several neurokinin (NK) receptors, including NK-1. Central administration of substance P is anxiogenic in animal models. All three of these neuropeptide systems are targets for therapeutic drug development in PD and other anxiety conditions.

Animal Models Relevant to Panic Disorder

Several rodent models have particular relevance to the behavioral and physiological responses of PD. One focuses on the dorsomedial hypothalamus, in which local disruption of GABA activity results in vulnerability to lactate-induced paniclike behavioral and physiological responses. Similarly, the repeated local delivery of exogenous CRH to the basolateral nucleus (BL) results in vulnerability to lactate, suggesting both sites are involved in sensitivity to the panicogen. Even in the absence of pharmacological manipulation of these regions, several panicogens, including lactate, CO_2, and doxapram, have been shown to enhance anxiety-related behaviors in diverse paradigms of fear, conflict, and avoidance. Systemic administration of doxapram, which activates carotid body chemoreceptors, is reported to increase endogenous CRH expression in the CE. CCK-8, a CCK cleavage product that activates gut CCK_1 receptors, illustrates how viscerosensory activation of fear circuitry may occur. The visceral signal induced by CCK-8 is transmitted via the vagus nerve to the caudal NTS, where it is transferred to NA neurons projecting to the hypothalamus, PBN, and amygdala. If such NA neurons co-express CRH, a similar rise in CRH could be expected in the CE. And, like doxapram and epinephrine, CCK-8 enhances the retention of conditioned fear memories. Therefore, sensitivity of the fear circuitry to viscerosensory pathway activation in PD is implied by both preclinical models, human challenge studies, and the anxiety sensitivity theory. Animal models specific to PD are useful for dissecting the neural pathways key to pathological anxiety and panic, and advances in brain imaging are allowing the direct testing of the implications for PD pathophysiology.

Functional Imaging in Panic Disorder

Although there have been several reports of regional brain-volume differences in PD by MRI, no consistent pattern has yet emerged. In contrast, functional imaging in PD has added considerably to our understanding of the pathophysiological mechanisms. For example, magnetic resonance spectroscopy (MRS) studies in PD showed higher brain lactate levels in response to controlled hyperventilation in

PD, suggesting there may be an exaggerated global cerebrovascular response in PD to hyperventilation-induced hypocarbia. A tilt table challenge of PD patients demonstrated an exaggerated decrease in middle cerebral artery flow by transcranial Doppler in both remitted and symptomatic patients, indicating such cerebrovascular dysautonomia may be a trait of PD. PET and SPECT studies of regional CBF and glucose metabolism have been aimed at establishing both baseline and panic state differences in PD from controls. Several studies point to baseline differences in left/right ratios of hippocampal and parahippocampal activity in PD as well as decreased left hippocampal activity. Anxiety and panic states in PD have been associated with a decrease or a smaller increase in global CBF compared with controls. Regional differences from controls have included relative decreases in the PFC regions and increases in the insula and striatum. In particular, the panic state induced by CCK-4 was associated with decreased medial PFC activity. A recent ^{18}F-fluorodeoxyglucose PET study of PD patients with high prescan state anxiety showed elevated activity in the bilateral amygdala, hippocampus, thalamus, brain stem, and cerebellum. Moreover, successful treatment with CBT for PD was associated with reduced activity in the hippocampus, anterior cingulate, cerebellum, and pons and increased activity in the medial PFC (see Figure 1). In summary, functional imaging in PD supports abnormal global cerebrovascular response, deficiency in PFC-mediated breaking on subcortical activity, hyperactivity in the subcortical fear network, and lateralized abnormalities in hippocampal activity.

Comments

Although the pathophysiology of PD has yet to be delineated, the past 25 years of research have focused attention on the most likely brain regions, circuits, and neurotransmitter systems involved. Building on the neurobiological foundations established by preclinical neuroscience research, plausible heuristic models explaining PD symptomatology have been proposed. Converging evidence from challenge studies, genotyping for functional polymorphisms, and radioligand receptor imaging studies implicates substantial roles for noradrenergic and serotonergic system dysfunction in PD. A better appreciation of key neuropeptide systems may soon allow more specific targeting of dysfunctional fear and anxiety circuits, with the promise of more efficacious, more tolerable, and safer therapies. Functional imaging has indicated exaggerated cerebrovascular responses in PD and some data support the notion of deficient top-down control of hyperresponsive subcortical fear circuits by regions of the PFC. The rapid advances in our understanding of PD have been propelled by an extraordinary array of scientific disciplines and research modalities. Such efforts should make PD one of the first psychiatric disorders to yield to an accurate identification of the neuropathology.

See also: Amygdala: Contributions to Fear; Anxiety Disorders; Anxiety: Drug Therapy; Aversive Emotions: Genetic Mechanisms of Serotonin; Panic Disorder.

Further Reading

Battaglia M and Ogliari A (2005) Anxiety and panic: From human studies to animal research and back. *Neuroscience and Biobehavioral Review* 29: 169–179.

Bouton ME, Mineka S, and Barlow DH (2001) A modern learning theory perspective on the etiology of panic disorder. *Psychological Review* 108: 4–32.

Gorman JM, Kent JM, Sullivan GM, and Coplan JD (2000) Neuroanatomical hypothesis of panic disorder, revised. *American Journal of Psychiatry* 157: 493–505.

Keck ME and Strohle A (2005) Challenge studies in anxiety disorders. In: Holsboer F and Ströhle A (eds.) *Handbook of Experimental Pharmacology, Vol. 169: Anxiety and Anxiolytic Drugs*, pp. 449–468. New York: Springer.

Kent JM, Mathew SJ, and Gorman JM (2002) Molecular targets in the treatment of anxiety. *Biological Psychiatry* 52: 1008–1030.

Klein DF (1993) False suffocation alarms, spontaneous panics, and related conditions: An integrative hypothesis. *Archives of General Psychiatry* 50: 306–317.

Maron E and Shlik J (2006) Serotonin function in panic disorder: Important, but why? *Neuropsychopharmacology* 31: 1–11.

Pare D, Quirk GJ, and Ledoux JE (2004) New vistas on amygdala networks in conditioned fear. *Journal of Neurophysiology* 92: 1–9.

Pollack MH (2005) The pharmacotherapy of panic disorder. *Journal of Clinical Psychiatry* 66(supplement 4): 23–27.

Sullivan GM, Coplan JD, Kent JM, and Gorman JM (1999) The noradrenergic system in pathological anxiety: A focus on panic with relevance to generalized anxiety and phobias. *Biological Psychiatry* 46: 1205–1218.

Parasitic Diseases, Nervous System Effects

K Leder, Monash University, Melbourne, VIC, Australia
P F Weller, Beth Israel Deaconess Medical Center, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Of the large number of protozoan and metazoan parasites that may infect humans, many have the potential to affect the nervous system. Protozoan parasites are unicellular and have the ability to replicate and proliferate within the human host. In contrast, metazoan parasites are multicellular and include nematodes (roundworms) and two groups of flatworms, trematodes (flukes) and cestodes (tapeworms). In general, almost all metazoan parasites lack the ability to replicate within the human body and hence augmentation of intensity of infection requires additional exogenous reinfection. Metazoan, but not protozoan, parasites are characteristically associated with an immunological response that includes increased numbers of eosinophilic leukocytes in the blood and tissues.

Some parasites have a characteristic predilection for involvement of the nervous system, whereas other infections are only rarely associated with neurological manifestations. This article does not exhaustively list all parasites reported to have possible neurological effects, but instead discusses those infections that are most likely to present with symptoms related to the nervous system.

Protozoan Parasites

Malaria, widely prevalent in tropical and subtropical areas of the world, is caused by mosquito-transmitted infections with one of four species of *Plasmodium*. Of the four species, the most serious infection is caused by *Plasmodium falciparum*, which accounts for almost all deaths due to malaria. In falciparum malaria, after an initial hepatic phase of development, parasites invade cells and engage in an erythrocytic cycle of development that eventuates in red cell lysis and liberation of parasitic forms that perpetuate the infection. Fever is typical and may, but need not, occur in daily or every-other-day paroxysms with chills. One of the most serious complications of falciparum malaria is the development of cerebral malaria, which may develop rapidly over many hours and is associated with high mortality. Neurological manifestations include disturbances of consciousness (ranging from somnolence to frank coma), acute organic brain syndromes with behavioral or intellectual changes or hallucinations, major motor seizures, meningismus, or rarely focal neurological alterations. There are a number of potential pathogenic mechanisms underlying cerebral symptoms in malaria, and the neurological complications are often part of a multisystem dysfunction. Adherence of falciparum-infected erythrocytes to vascular endothelial cells within the brain leads to obstruction of blood flow within microvasculature and secondary cerebral ischemia. In addition, cytokine release and metabolic derangements such as hypoglycemia and lactic acidosis can also contribute to cerebral dysfunction. Untreated, cerebral malaria is uniformly fatal, so prompt diagnosis and therapy are mandatory. Diagnosis is made by finding malarial parasites in red cells on stained blood smears or via antigen-capture test kits. In most geographic regions, there is increasing resistance of falciparum malarial parasites to chloroquine and sulfadoxine-pyramethamine drugs, so combination therapy with alternative agents such as mefloquine, quinine, or artemisinin derivatives is required. Even with therapy, cerebral malaria is associated with up to 30% mortality, particularly in children, and permanent neurologic sequelae occur in up to 10% of survivors.

Toxoplasmosis, due to the intracellular parasite *Toxoplasma gondii*, is a globally distributed zoonotic infection. Infection, which is acquired either congenitally or by ingestion of oocysts or tissue cysts present in meat, can result in one of four clinical syndromes. Acute toxoplasmosis in normal hosts usually results in asymptomatic lymphadenopathy or lymphadenitis; only very rarely does encephalitis occur. In contrast, acute toxoplasmosis in an immunodeficient host is frequently marked by central nervous system (CNS) involvement, including encephalitis, meningoencephalitis, or mass lesions. The second form of toxoplasmosis arises from reactivation of latent parasites and occurs in immunosuppressed individuals, including those with AIDS. CNS mass lesions and meningoencephalitis are frequent manifestations in immunodeficient hosts. Patients present with headache, altered levels of consciousness, seizures, and/or focal neurological findings. Ocular toxoplasmosis manifests as chorioretinitis and may develop from acute acquired or congenital infection. In the fourth form, congenital toxoplasmosis may be associated with sequelae of varying severity due to CNS involvement. Diagnosis is made by finding appropriate morphological forms of the parasites in tissues or with the assistance of serological tests. Findings on computed tomography (CT) scans or magnetic resonance imaging (MRI) vary considerably with the different forms of infection, but

typically consist of intracerebral calcifications or hypodense lesions involving the thalamus and basal ganglia that show ring enhancement following contrast. Treatment is not usually required for primary infection in immunocompetent hosts, but pregnant women, immunocompromised individuals, and children with congenital infection should be treated. Unless contraindicated, therapy typically consists of a combination of oral pyrimethamine, sulfadiazine, and folinic acid.

Infections with amebic protozoa can involve the nervous system. *Entamoeba histolytica* causes amebiasis, which is principally a colonic infection. Hematogenous spread from the bowel can lead to the development of an amebic abscess, most commonly in the liver, but on occasion involving the brain. In addition, other free-living amebas of *Naegleria*, *Acanthamoeba*, and *Balamuthia* species, which are found in soil and water, can cause neurological infections. *Naegleria*, presumably entering through the cribriform plate, can produce a fulminant primary amebic meningoencephalitis resembling bacterial meningitis. Victims tend to be young and healthy, typically with a recent history of water-sport activities. Infections with *Acanthamoeba* are more indolent and often occur in immunocompromised hosts. Hematogenous spread to the CNS from primary pulmonary or skin lesions are thought to occur, resulting in a subacute granulomatous encephalitis manifest by headache, seizures, altered mental status, and/or focal neurological deficits. Infections due to *Balamuthia mandrillaris* also cause chronic granulomatous meningoencephalitis, occurring in both immunocompetent and immunocompromised individuals. Diagnosis of free-living amebic infections generally relies on identification of the parasites in cerebrospinal fluid (CSF) or brain tissue. All species are difficult to treat and are commonly fatal.

Microsporidia are obligate intracellular parasites. A number of different species can infect humans, but symptomatic infections are seen predominantly in immunosuppressed hosts. Microsporidia belonging to the *Encephalitozoon* species are most common and can result in serious disseminated infections, particularly in patients with human immunodeficiency virus (HIV) infection. Transmission occurs mainly via ingestion of contaminated food or water. Primary infection usually begins in the intestines or lungs, but widespread dissemination resulting in sinusitis, keratoconjunctivitis, and CNS involvement can follow. CNS symptoms vary according to the location of the lesion, but seizures are the most common manifestation. The diagnosis is made by detection of organisms in body fluids or tissue biopsy specimens via microscopy or polymerase chain reaction (PCR). In CNS infections, parasite spores are often present in peripheral blood and occasionally may also be present in CSF. Imaging of the brain usually shows multifocal ring-enhancing lesions. Treatment is with albendazole.

The causative organisms of African trypanosomiasis, or sleeping sickness, have a special predilection for invading the CNS. The disease, caused by hemoflagellates of the genus *Trypanosoma*, is prevalent in equatorial Africa. Two forms infect humans. West African trypanosomiasis is present in tropical forests of western and central Africa and is caused by *Trypanosoma brucei gambiense*, a parasite not found in major animal reservoirs. East African trypanosomiasis is due to *T. brucei rhodesiense*, which exists in wild animal reservoirs and is prevalent in the savanna and woodlands of eastern tropical Africa. Both forms of infection are transmitted by tsetse flies. East African trypanosomiasis is an acute febrile illness, and death may occur from pancarditis prior to CNS involvement. In contrast, West African trypanosomiasis is a chronic illness, and progressive CNS involvement produces meningoencephalitis and meningomyelitis which develop slowly over years. Increasing somnolence and lassitude progress to coma and death from inanition or intercurrent infection. Definitive diagnosis depends on finding the parasite in blood, CSF, lymph node aspirate, or bone marrow specimens. CSF analysis generally shows pleocytosis when the CNS is involved. Treatment relies on the chemotherapeutic agents suramin, pentamidine, arsenical derivatives, or eflornithine, all of which have significant toxicities and at best moderate efficacy, particularly in advanced stages of disease.

There are three stages of disease associated with American trypanosomiasis (Chagas disease) due to *Trypanosoma cruzi* infections: acute, intermediate, and chronic. In the acute stage, symptomatic infection typically manifests with fever, facial edema, lymphadenopathy, myocarditis, and occasionally meningoencephalitis. The encephalitis is due to direct destructive and inflammatory changes by the parasite in the CNS, which can be life-threatening, although spontaneous resolution is usual. The intermediate stage is not associated with clinical symptoms or with significant nervous system involvement. The chronic stage is characterized by immune mediated destruction of many nerve components, especially the autonomic innervation of the heart and gastrointestinal tract. Cardiac involvement causes arrhythmias, congestive heart failure, and thromboembolic events, and parasympathetic denervation of the gastrointestinal tract leads to dilatation of the esophagus or colon. In addition, immunosuppression due to malignancy, HIV, or following transplantation has recently been shown to be associated with reactivation of infection, often

resulting in meningoencephalitis and brain abscesses. Available therapeutic agents include nifurtimox and benznidazole, but neither has significant efficacy once complications from chronic infection have developed.

Metazoan Infections

Nematode Infections

Eosinophilic meningitis may develop when humans are infected with the rat lungworm, *Angiostrongylus cantonensis*. This parasite is present principally in the Far East and the islands of the Pacific basin, including Hawaii, and has been found in Cuba. Third-stage larvae of the parasite are infective. Humans are accidental, dead-end hosts and acquire the infection by ingesting infected land snails or slugs or by consuming inadequately cooked carrier hosts (freshwater shrimp, aquatic and amphibious crabs, and certain marine fish) that have ingested infected mollusks. After the larvae are ingested, they migrate to the brain, spinal cord, and nerve roots, where they induce an eosinophilic inflammatory reaction. Common symptoms include severe headache, nuchal rigidity, vomiting, weakness, and paresthesias. Fever is not prominent. Seizures and extraocular muscle palsies occur uncommonly. CSF pleocytosis develops, with greater than 20% of the cells being eosinophils. The illness is usually self-limited and resolves over several weeks. Anthelminthic agents should not be given as they can result in an enhanced inflammatory reaction to dying parasite larvae within the CNS, with a consequent clinical deterioration.

Infective larvae of species of *Gnathostoma* are found in tissues of freshwater fish, in eels, frogs, and snakes, and in poultry or pigs fed infected fish. Ingestion of these infected foods by humans introduces larvae which can migrate within human tissues, usually within the skin. Infection has been noted principally in the Far East. Less commonly, larvae migrate along a nerve track and enter the CNS, producing radiculomyelitis and an eosinophilic meningoencephalitis. Characteristic symptoms include headache, severe radicular pain, migratory parasthesias, or paralysis. Subarachnoid hemorrhage and coma can also ensue. Examination of the CSF will show an eosinophilic pleocytosis which is often xanthochromic or bloody, and cerebral imaging typically shows areas of hemorrhage.

Strongyloidiasis arises from infection with *Strongyloides stercoralis*, which is unique because it can internally reinfect humans. This internal autoinfection cycle allows the parasite to perpetuate its infection for decades. The autoinfection is modulated by the human immune system; if immunosuppression develops, such as by the administration of corticosteroids, the autoinfection cycle can become unbridled, leading to the hyperinfection syndrome. Large numbers of infective larvae are produced and penetrate the colon to disseminate hematogenously. Any organ, including the brain, may be invaded by larvae. In addition, secondary bacterial infections, including polymicrobial pneumonia or meningitis, are not infrequent complications of disseminated strongyloidiasis. Treatment in these cases will consist of antibiotic therapy plus ivermectin.

Ingestion of larvae of *Trichinella spiralis*, encysted in the flesh, especially pork, of carnivorous mammals, leads to the development of trichinosis. The intestinal phase of trichinosis develops first as adult parasites generate new larvae that enter the bloodstream. Thereafter, the muscle phase of the illness ensues, with the common development of periorbital and facial edema, subconjunctival and retinal hemorrhages, and subungual splinter hemorrhages. About 3 weeks after infection, larvae begin to encyst in striated muscle, and signs and symptoms of myositis, fever, and eosinophilia occur. Although larvae encyst only in striated muscle, inflammatory lesions may develop due to larval migration in the heart, lungs, and CNS. Neurologic involvement can result from vascular obstruction, eosinophil infiltration, or toxic effects of parasite antigens. Clinical manifestations include meningeal irritation resembling meningitis, intracerebral hemorrhages, and occasionally seizures, paresis, delerium, or coma. Diagnosis relies on serology or direct demonstration of organisms in biopsy specimens. Treatment is not always necessary, but in severe cases prednisolone and/or specific anthelminthic therapy with albendazole or mebendazole can be given.

Visceral larva migrans develops when the infectious eggs of dog or cat ascarids, *Toxocara* species, are ingested. Since these eggs are found in soil contaminated with dog or cat feces, the illness is seen in those who ingest soil, particularly children with geophagous pica. Larvae hatch from the eggs and are carried from the intestinal tract in the bloodstream. The larvae then migrate to various organs and induce an eosinophil-rich inflammatory response before they die. Although many with this visceral form of the infection are minimally symptomatic, malaise, weight loss, fever, hepatomegaly, and rashes may occur. With heavy infection neurological involvement may occur. The spectrum of neurological manifestations includes muscular weakness, sensory abnormalities, seizures, behavioral disorders, and coma. Peripheral eosinophilia will often be present. A second form of the illness, the ocular form, is also recognized. This form presents solely with ocular findings due to a retinal granuloma formed around a larva; less commonly,

endophthalmitis or uveitis may occur. Therapy for toxocaral infections is not always necessary and is primarily symptomatic. Another potential cause of the larva migrans syndrome is infection with the raccoon-derived ascarid parasite, *Bayliascaris*. Larvae can enter the brain, causing an often-fatal eosinophilic meningoencephalitis.

Trematode Infections

Schistosomiasis is a major, snail-transmitted infection due principally to one of three species of *Schistosoma*: *Schistosoma japonicum* in the Orient, *Schistosoma mansoni* in Africa and the Americas, and *Schistosoma hematobium* in Africa and Arabia. Adult parasites localize in the venous system draining the colon (for *S. mansoni* and *S. japonicum*) or the urinary tract (*S. hematobium*). Much of the morbidity of schistosomiasis is due to granulomatous responses developing around eggs deposited in the colon or urinary tract or, for the former two species, carried hematogenously into the liver. Neurological complications usually occur in chronically infected individuals but can also occur in the acute stage. These complications arise either as a result of embolization of eggs with deposition in nervous tissue or due to the anomalous migration of adult worms. Cerebral schistosomiasis occurs predominantly with *S. japonicum* and presents as focal or generalized seizures. Spinal schistosomiasis, due principally to *S. mansoni*, results in the development of transverse myelitis. Diagnosis is usually presumptive since confirmatory biopsy tissue is not available, but positive serology and/or visualization of eggs in urine or feces may be helpful. Praziquantel is highly efficacious in the killing of adult worms. Adjunctive corticosteroids are sometimes given to reverse the inflammatory pathology associated with egg deposition.

Paragonimiasis is due to infection with flukes of *Paragonimus westermani* in the Orient and other *Paragonimus* species in Africa and the Americas. Infection is acquired by consumption of freshwater crabs or crayfish. Young flukes migrate principally into the lungs, and pulmonary manifestations are usual. However, flukes may occasionally migrate into the CNS. An acute meningoencephalitis or progressive neurological signs including paraplegia, visual disturbances, seizures, or coma develop secondary to a space-occupying mass lesion in the brain or spinal cord. Neuroimaging findings and serological testing of serum or CSF can aid in the diagnosis. The active stage of infection shows a characteristic 'grapelike' cluster of multiple, ring-enhancing lesions, whereas old, chronic lesions are often calcified and are typically described as showing 'soap-bubble' calcification. Treatment is usually with an extended course of praziquantel.

Fasciola hepatica, the liver fluke, which is widely distributed in Europe, Asia, Africa, and the Americas, may cause human infection if parasitic cysts attached to aquatic plants, especially watercress, are ingested. Although fascioliasis involves primarily the liver, aberrant migration of flukes to the CNS can occur. Imaging, serology, or visualization of eggs on fecal microscopy can help make the diagnosis. Triclabendazole is now the treatment of choice.

Cestode Infections

The fish tapeworm, *Diphyllobothrium latum*, does not leave the intestinal tract. It does, however, compete with the host for dietary vitamin B12, and megaloblastic anemia and neuropathy due to vitamin B12 deficiency can develop in some infected patients.

The pork tapeworm, *Taenia solium*, may cause cysticercosis. Neurocysticercosis is the most common parasitic infection of the brain, affecting over 50 million people worldwide. Pigs serve as the intermediate hosts for this parasite; humans who eat undercooked pork containing cysts develop intestinal tapeworm infection rather than cysticercosis. Instead, cysticercosis is contracted by ingestion of *T. solium* eggs, which usually occurs via contamination of food such as fruits or vegetables. CNS lesions can occur in the brain parenchyma, in the subarachnoid space, in the ventricular system, or in the spinal cord, and multiple cysts are not uncommon. Symptoms will vary depending on the number, size, viability, and location of the lesions. Patients commonly present with seizures; other potential complications include headache, hydrocephalus, or focal neurological signs. CT scans or MRIs often suggest the diagnosis, and serological tests are also available. Optimal treatment varies, ranging from observation, symptomatic therapy only, or chemotherapy with albendazole $\pm$ praziquantel, to surgical intervention, depending on the individual characteristics of the lesion(s).

CNS lesions, similar to those of cysticercosis, also occur with coenurosis, which develops from infection with dog tapeworms of the genus *Multiceps*. Sparganosis is another tapeworm infection due to larvae of the species *Spirometra*. The disease is most common in East Asia. Humans can be accidentally infected by drinking water that is contaminated by larval stages of the parasite, by ingesting infected amphibians or mammals that serve as intermediate hosts, or by applying poultices of raw infected frog flesh to a wound or eye. Subcutaneous tissue, subconjunctival tissue, and other organs can be affected in addition to the CNS. Brain abscesses and spinal cord involvement can occur, causing seizures, paresis, paresthesias, and other symptoms. MRI or CT scanning can help establish the diagnosis, but definitive diagnosis and treatment

usually relies on surgical excision, which is indicated particularly if the worm is still viable.

Echinococcosis in humans results predominantly from infection with either *Echinococcus granulosus* or *Echinococcus multilocularis*. Cystic hydatid disease is more common and is due to *E. granulosus*, whereas *E. multilocularis* causes alveolar echinococcosis. The adult parasites live in the intestines of dogs and wild canids such as foxes and wolves. Infective eggs, passed in the feces, are ingested by intermediate hosts, including sheep, cattle, and humans. The life cycle is maintained when dogs or wolves ingest the carcasses of infected intermediate hosts. Hence, cystic hydatid disease is most prevalent in sheep- and cattle-raising countries. In infected humans, unilocular cysts most commonly develop in the liver, but other organs, including the brain, may be involved either from primary inoculation or via secondary spread of a ruptured cyst. Variable clinical findings develop referable to the location and size of space-occupying cystic lesions. The spinal cord can also be involved, with secondary cord compression. Neurological hydatid disease is associated with concomitant hepatic cysts in up to 80% of cases. The diagnosis may be suspected on CT scan or MRI, and serological tests can help establish the etiology. Definitive therapy of lesions usually involves surgical excision, but treatment with albendazole can also be considered. Further information regarding echinococcal infection can be found at Echinonet Online Version of Echinonews.

See also: Gene Expression Dysregulation in CNS Pathophysiology; Neural Stem Cells and CNS Diseases; Stem Cells and CNS Repair.

Further Reading

Bia FJ and Barry M (1986) Parasitic infections of the central nervous system. *Neurologic Clinics* 4: 171–206.

Chang KH and Han MH (1998) MRI of CNS parasitic diseases. *Journal of Magnetic Resonance Imaging* 8: 297–307.

Golden RF and Hill DR (1993) Serodiagnosis of parasitic infection of the central nervous system. *Seminars in Neurology* 13: 219–233.

Guerrant R, Walker DH, and Weller PF (2006) *Tropical Infectious Diseases: Principles, Pathogens and Practice*, 2nd edn. Philadelphia: Churchill Livingstone.

Lowichik A and Ruff AJ (1995) Parasitic infections of the central nervous system in children. I: Congenital infections and meningoencephalitis. *Journal of Child Neurology* 10: 4–17.

Lowichik A and Ruff AJ (1995) Parasitic infections of the central nervous system in children. II: Disseminated infections. *Journal of Child Neurology* 10: 77–87.

Lowichik A and Siegel JD (1995) Parasitic infections of the central nervous system in children. III: Space-occupying lesions. *Journal of Child Neurology* 10: 177–190.

Moffitt MC (1994) Parasitic infections of the central nervous system. *Pediatric Annals* 23: 424–425, 429–433.

Nishimura K and Hung T (1997) Current views on geographic distribution and modes of infection of neurohelminthic diseases. *Journal of the Neurological Sciences* 145: 5–14.

Strickland GT and Magill AJ (2000) *Hunter's Tropical Medicine*, 8th edn. Philadelphia, PA: WB Saunders.

Walker MD, Kublin JG, and Zunt JR (2006) Parasitic central nervous system infections in immunocompromised hosts: Malaria, microsporidiosis, leishmaniasis, and African trypanosomiasis. *Clinical Infectious Diseases* 42: 115–125.

Walker MD and Zunt JR (2005) Neuroparasitic infections: Cestodes, trematodes, and protozoans. *Seminars in Neurology* 25: 262–277.

Walker MD and Zunt JR (2005) Neuroparasitic infections: Nematodes. *Seminars in Neurology* 25: 252–261.

Relevant Websites

http://www.cdfound.to.it – Carlo Denegri Foundation Atlas of Medical Parasitology, Central Nervous System (CNS) Parasites.

http://www.cdc.gov/health/diseases.htm – Centers for Disease Control and Prevention (A–Z index for health information).

http://www.medicalweb.it – Echinonet Online Version of Echinoneus.

http://www.medletter.com – The Medical Letter (on drugs, therapeutics, and treatment guidelines and for parasitic infections).

http://www.who.int – World Health Organization (A–Z index of health topics; see for trypanosomiasis, Chagas disease, malaria, and schistosomiasis).

Parasomnias

M W Mahowald, M A Cramer Bornemann, and C H Schenck, Hennepin County Medical Center, Minneapolis, MN, USA; and University of Minnesota Medical School, Minneapolis, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

There are many different parasomnias. They may be divided into two broad categories: primary, which represent abnormalities of the sleep states *per se*, and secondary due to activity of other organ systems, which occur primarily or predominately during sleep. The primary parasomnias can be divided with regard to the parent state: nonrapid eye movement (NREM) sleep, rapid eye movement (REM) sleep, or a miscellaneous category not respecting sleep states. The secondary parasomnias can be categorized by the organ system involved. The most common primary parasomnias are the disorders of arousal, REM sleep behavior disorder, and rhythmic movement disorder. Common secondary parasomnias are nocturnal seizures and, less common, sleep-related expiratory groaning.

Primary Sleep Parasomnias

The concept that sleep and wakefulness are not invariably mutually exclusive states, and that the various state-determining variables of wakefulness, NREM sleep, and REM sleep may occur simultaneously or oscillate rapidly, is key to the understanding of the primary sleep parasomnias. The admixture of wakefulness and NREM sleep would explain confusional arousals (sleep drunkenness), automatic behavior, or microsleeps. The tonic and phasic components of REM sleep may become dissociated, intruding or persisting into wakefulness, thus explaining cataplexy, wakeful dreaming, lucid dreaming, and the persistence of motor activity during REM sleep (REM sleep behavior disorder).

Parasomnias are most apt to appear as the brain becomes reorganized across states; therefore, they often occur during the transition periods from one state to another. In view of the large number of neural networks, neurotransmitters, and other state-determining substances which must be recruited synchronously for full state declaration and the frequent transitions among states during the wake–sleep cycle, it is surprising that errors in state declaration do not occur more frequently.

In addition to the phenomenon of state dissociation, in which two states of being overlap or occur simultaneously, there are likely additional underlying physiologic phenomena which contribute to the appearance of complex motor behaviors during sleep. These include activation of locomotor centers during sleep, sleep inertia upon arousal, and sleep state instability.

NREM Parasomnias

Disorders of Arousal

The disorders of arousal are the most impressive and most common of the NREM sleep parasomnias. These share common features: They tend to arise from the deepest stages of NREM sleep (slow-wave sleep (stages 3 and 4 of NREM sleep)), thus usually occurring in the first third of the sleep cycle (and rarely during naps); and they are common in childhood, usually decreasing in frequency with increasing age. Disorders of arousal may be primed by febrile illness, alcohol, prior sleep deprivation, physical activity, emotional stress, or medications. Such factors should be considered as priming events in susceptible individuals, and not causal. Numerous other sleep disorders which result in arousals (obstructive sleep apnea, nocturnal seizures, or periodic limb movements) may also trigger disorders of arousal.

Persistence of these behaviors beyond childhood or their development in adulthood have erroneously been taken as an indication of underlying psychopathology. Numerous studies have dispelled this myth, indicating that significant psychopathology is usually not present in adults with disorders of arousal.

Disorders of arousal occur on a broad spectrum ranging from confusional arousals, through somnambulism (sleepwalking), to sleep terrors. Some take the form of 'specialized' behaviors, such as sleep-related eating and sleep-related sexual activity– without conscious awareness.

Confusional Arousals

These are often seen in children and are characterized by movements in bed, occasionally thrashing about, or inconsolable crying. 'Sleep drunkenness' is probably a variation on this theme. The prevalence of confusional arousals in adults is approximately 4%.

Sleepwalking

Sleepwalking is prevalent in childhood (1–17%), peaking at 11 or 12 years of age, and is far more common in adults (nearly 4%) than generally acknowledged. Sleepwalking may be either calm or agitated, with varying degrees of complexity and duration.

Sleep Terrors

The sleep terror is the most dramatic disorder of arousal. It is frequently initiated by a loud, bloodcurdling scream associated with extreme panic, followed by prominent motor activity such as hitting the wall and running around or out of the bedroom, or even out of the house, resulting in bodily injury or property damage. A universal feature is inconsolability, and attempts at consolation are fruitless and may serve only to prolong or even intensify the confusional state. Complete amnesia for the activity is typical but may be incomplete. Sleep terrors are much more prevalent in adults than generally acknowledged (4% or 5%). Although usually benign, these behaviors may be violent, resulting in considerable injury to the victim or others or damage to the environment, occasionally with forensic implications.

Treatment of disorders of arousal is often not necessary. Reassurance of their typically benign nature, lack of psychological significance, and the tendency to diminish over time is often sufficient. The tricyclic antidepressants and benzodiazepines may be effective, and they should be administered if the behaviors are dangerous to person or property or extremely disruptive to family members. Nonpharmacologic treatment such as hypnosis is recommended for long-term management. Anticipatory awakening has been reported to be effective in treating sleepwalking in children. The avoidance of priming factors such as drugs, alcohol, and sleep deprivation is also important.

REM Sleep Parasomnias

REM Sleep Behavior Disorder

The most common REM sleep parasomnia is the REM sleep behavior disorder (RBD). One of the defining features of REM sleep is active paralysis of all somatic musculature (with the exception of the diaphragm to permit respiration during REM sleep). In experiments reported in 1965, bilateral lesions of pontine regions adjacent to the locus coeruleus in cats caused the absence of the expected atonia associated with REM sleep, allowing the cats to demonstrate prominent motor behaviors during REM sleep (oneiric behaviors). Patients with RBD lack the paralysis of REM sleep, permitting the acting out of dream mentation, often with violent or injurious results. The prevalence of RBD is approximately 0.5%.

The presenting complaint is that of vigorous sleep behaviors usually accompanying vivid striking dreams. These behaviors may result in repeated injury, including ecchymoses, lacerations, and fractures. The potential for injury to self or bed partner raises interesting and difficult forensic medicine issues. Two striking demographic characteristics of RBD are that it predominately affects males (approximately 90%) and usually begins after the age of 50 years. RBD in humans occurs in both an acute and a chronic form.

Until recently, most reported cases of acute transient RBD fell in the toxic/metabolic category, with the best-studied conditions being the withdrawal states – most commonly involving ethanol. A growing cause of acute RBD appears to be iatrogenic, particularly associated with tricyclic antidepressants and serotonin-specific and noradrenergic-specific reuptake inhibitors – both widely prescribed categories of antidepressants.

The chronic form is most often either idiopathic or associated with neurological disorders. Each basic category of neurological disease (vascular, neoplastic, toxic/metabolic, infectious, degenerative, traumatic, congenital, and idiopathic) could be expected to manifest this disorder. Careful longitudinal evaluation has demonstrated that the majority of patients with what initially appears to be 'idiopathic' RBD will eventually develop neurodegenerative disorders, most notably the synucleinopathies (Parkinson's disease, multiple system atrophy, including olivopontocerebellar degeneration and the Shy–Drager syndrome, or dementia with Lewy body disease). RBD may be the first manifestation of these conditions, and it may precede any other manifestation of the underlying neurodegenerative process by more than 10 years. With such a high percentage of patients with RBD eventually developing other manifestations of neurodegenerative diseases, it has been proposed that the term idiopathic RBD be replaced by 'cryptogenic' RBD – the implication being that if followed long enough, the underlying neurologic condition will declare itself.

Systematic longitudinal study of patients with such neurological syndromes indicates that RBD and REM sleep without atonia may be far more prevalent than previously suspected. Although the exact prevalence of RBD in extrapyramidal disease is unknown, subjective reports indicate that 25% of patients with Parkinson's disease have behaviors suggestive of RBD or sleep-related injurious behaviors, and polysomnographic studies found RBD in up to 47% of patients with Parkinson's disease with

sleep complaints. In one large series of patients with multiple system atrophy, 90% were found to have REM sleep without atonia and 69% had clinical RBD.

There is a higher incidence of RBD in patients with narcolepsy. This association could be expected, because both narcolepsy and RBD are characterized by 'state boundary control' abnormalities. Furthermore, tricyclic antidepressants, SSRIs, and monoamine oxidase inhibitors, often prescribed to treat cataplexy associated with narcolepsy, can trigger or exacerbate RBD in this population.

Neuroimaging studies indicate dopaminergic abnormalities in RBD. Single-photon emission computed tomography (SPECT) studies have found reduced striatal dopamine transporters, and decreased striatal dopaminergic innervation has been reported. Decreased blood flow in the upper portion of the frontal lobe and pons has been reported, as has functional impairment of brain stem neurons. Positron emission tomography and SPECT studies have revealed decreased nigrostriatal dopaminergic projections in patients with multiple system atrophy and RBD. Decreased blood flow in the upper portion of the frontal lobe and pons has been found in one magnetic resonance imaging and SPECT study. Whole brain perfusion changes in patients with RBD are similar to those seen in Parkinson's disease. Impaired cortical activation as determined by electroencephalographic spectral analysis in patients with idiopathic RBD supports the relationship between RBD and neurodegenerative disorders.

There are a number of fascinating variations on RBD (beyond the scope of this article), all likely representing abnormalities of state boundary control. These include status dissociatus, the parasomnia overlap syndrome, and agrypnia excitata.

Clonazepam 0.5–1.0 mg at bedtime is effective in approximately 90% of cases, with little evidence of tolerance or abuse. Melatonin, often at doses of 3–12 mg at night, may also be effective. Pramipexole or levodopa may be effective. The other essential therapeutic intervention concerns environmental safety.

Rhythmic Movement Disorder

Rhythmic movement disorder (RMD), formerly termed jactatio capitis nocturna, refers to a group of behaviors characterized by stereotyped movements (rhythmic oscillation of the head or limbs and head banging or body rocking during sleep) seen most frequently in childhood. Its persistence into adulthood is not uncommon. It may be familial, and it is not the manifestation of any underlying psychiatric or psychological condition. RMD may arise from either REM or NREM sleep and may occur in the transition from wake to sleep. Significant injury from repetitive pounding may result. The etiology of RMD is unknown, and no systematic studies of pharmacologic or behavioral treatment have been reported.

Numerous other conditions may perfectly mimic the primary sleep parasomnias. These include obstructive sleep apnea, nocturnal seizures, psychogenic dissociative disorders, or malingering.

Isolated, often bizarre, sleep-related events may be experienced by perfectly normal people, and most do not warrant further extensive or expensive evaluation. The initial approach to the complaint of unusual sleep-related behavior is to determine whether further evaluation is necessary. The patient should be queried regarding the exact nature of the events. Because many of these episodes may be associated with partial or complete amnesia, additional descriptive information from a bed partner or other observer may prove invaluable. Home videotapes of the clinical event may be quite helpful. In general, indications for formal evaluation of parasomnias include behaviors that are potentially violent or injurious, are extremely disruptive to other household members, result in the complaint of excessive daytime sleepiness, or are associated with medical, psychiatric, or neurological symptoms or findings.

Secondary Sleep Parasomnias

Nocturnal Seizures

Sleep and epilepsy are common bedfellows. In most cases, epilepsy is highly state dependent: NREM sleep promotes seizures, whereas REM sleep is a relatively antiepileptic state. This state-dependent vulnerability to seizures reflects the dramatic reorganization of the entire central nervous system as it moves across the three states of being: wakefulness, NREM sleep, and REM sleep. There is extensive overlap and potential for confusion between other parasomnias and nocturnal seizures, particularly nocturnal frontal lobe epilepsy.

Sleep-Related Expiratory Groaning (Catathrenia)

Groaning during sleep has been termed 'catathrenia.' The behaviors occur intermittently during either REM or NREM sleep and are characterized by prolonged, often very loud, socially disruptive groaning sounds during expiration. It is poorly understood and awaits further definition and therapeutic studies. There is no known effective treatment.

Conclusion

Evaluation of problematic behaviors arising from the sleep period is best performed by experienced,

multidisciplinary teams of sleep medicine professionals in centers with extensive polysomnographic monitoring capability.

See also: Cataplexy; Narcolepsy; Sleep Architecture; Sleep Oscillations and PGO Waves; Sleep and Sleep States: PET Activation Patterns; Sleep Apnea; Sleeping Sickness.

Further Reading

Albin RL, Koeppe RA, Chervin RD, et al. (2000) Decreased striatal dopaminergic innervation in REM sleep behavior disorder. *Neurology* 55: 1410–1412.

Cramer Bornemann MA, Mahowald MW, and Schenck CH (2006) Parasomnias. Clinical features and forensic implications. *Chest* 130: 605–610.

Derry CP, Davey M, Johns M, et al. (2006) Distinguishing sleep disorders from seizures. Diagnosing bumps in the night. *Archives of Neurology* 63: 705–709.

Gagnon J-F, Postuma RB, Mazza S, et al. (2006) Rapid-eye-movement sleep behaviour disorder and neurodegenerative diseases. *Lancet Neurology* 5: 424–432.

Gagnon J-F, Postuma RB, and Montplaisir J (2006) Update on the pharmacology of REM sleep behavior disorder. *Neurology* 67: 742–747.

Hurwitz TD, Mahowald MW, Schenck CH, et al. (1991) A retrospective outcome study and review of hypnosis as treatment of adults with sleepwalking and sleep terror. *Journal of Nervous and Mental Disease* 179: 228–233.

Jouvet M and Delorme F (1965) Locus coeruleus et sommeil paradoxal. *Comptes Rendus Des Seances de la Societe de Biologie et de ses Filiales* 159: 895–899.

Kohyama J, Masukura F, Kimura K, and Tachibana N (2002) Rhythmic movement disorder: Polysomnographic study and summary of reported cases. *Brain and Development* 24: 33–38.

Maaza S, Soucy JP, Gravel P, et al. (2006) Assessing whole brain perfusion changes in patients with REM sleep behavior disorder. *Neurology* 67: 1618–1622.

Mahowald MW (2004) Parasomnias. *Medical Clinics of North America* 88: 669–678.

Mahowald MW and Schenck CH (2000) Parasomnias: Sleepwalking and the law. *Sleep Medicine Reviews* 4: 321–339.

Mahowald MW and Schenck CH (2001) Evolving concepts of human state dissociation. *Archives Italiennes de Biologie* 139: 269–300.

Mahowald MW and Schenck CH (2005) NREM sleep parasomnias. *Neurologic Clinics* 23: 1077–1106.

Nightingale S, Orgill JC, Ebrahim IO, et al. (2005) The association between narcolepsy and REM behavior disorder (RBD). *Sleep Medicine* 6: 253–258.

Ohayon M, Guilleminault C, and Priest RG (1999) Night terrors, sleepwalking, and confusional arousal in the general population: Their frequency and relationship to other sleep and mental disorders. *Journal of Clinical Psychiatry* 60: 268–276.

Ohayon MM, Caulet M, and Priest RG (1997) Violent behavior during sleep. *Journal of Clinical Psychiatry* 58: 369–376.

Pressman MR (2007) Factors that predispose, prime, and precipitate NREM parasomnias in adults: Clinical and forensic implications. *Sleep Medicine Reviews* 11: 5–30.

Schenck CH and Mahowald MW (1990) Polysomnographic, neurologic, psychiatric, and clinical outcome report on 70 consecutive cases with REM sleep behavior disorder (RBD): Sustained clonazepam efficacy in 89.5% of 57 treated patients. *Cleveland Clinic Journal of Medicine* 57(supplement): S9–S23.

Schenck CH and Mahowald MW (1994) Review of nocturnal sleep-related eating disorders. *International Journal of Eating Disorders* 15: 343–356.

Schenck CH and Mahowald MW (2005) REM sleep parasomnias. *Neurologic Clinics* 23: 1107–1126.

Shapiro CM, Trajanovic NN, and Fedoroff JP (2003) Sexsomnia – A new parasomnia? *Canadian Journal of Psychiatry* 48: 311–317.

Shouse MN, da Silva AM, and Sammaritano M (1996) Circadian rhythm, sleep, and epilepsy. *Journal of Clinical Neurophysiology* 13: 32–50.

Tobin JD Jr. (1993) Treatment of somnambulism with anticipatory awakening. *Journal of Pediatrics* 122(3): 426–427.

Vetrugno R, Provini F, Plazzi G, et al. (2001) Catathrenia (nocturnal groaning): A new type of parasomnia. *Neurology* 56: 681–683.

Parasympathetic Nervous System

J B Furness, University of Melbourne, Parkville, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

The parasympathetic nervous system is a division of the autonomic nervous system (ANS). The parasympathetic nervous system is not truly a system; it does not act as a coordinated whole, except to the extent that it is part of the ANS. The reasons for classifying a part of the ANS as the parasympathetic nervous system are buried in the history of the investigation of the anatomy and physiology of the ANS. The major reasons are (1) that the autonomic outflows from the central nervous system are divided anatomically into cranial, thoraco-lumbar, and sacral outflows (between these are sections of the neuraxis without autonomic outflows) and (2) that some pathwys in the cranial and sacral parts provide vasodilator commands, whereas the thoraco-lumbar outflows include vasoconstrictor pathways.

The organs and functions that are controlled through the parasympathetic division of the ANS are summarized in Table 1. An inspection of Table 1 emphasizes that the parasympathetic nervous system is not a system in a functional sense. For example, there is no obvious physiological link between micturition and gastric acid secretion or between lacrimation and penile erection. In another instance, for control of the airway caliber, both a pathway for bronchoconstriction and a pathway for bronchodilation travel in the parasympathetically defined vagus nerve.

Nevertheless, a number of features do link the components of the parasympathetic nervous system. The preganglionic parasympathetic neurons form synapses in the ganglia close to the organs that they influence, in some cases within the walls of the organs. This is generally not the case for the sympathetic pathways, in which most peripheral ganglia are remote from the innervated organs. Exceptions include the spleen and kidney, which are close to the sympathetic ganglia that supply them. There are even reports of sympathetic

Table 1 Parasympathetic pathways and their functions

Function/organ	*Pathway from CNS*[a]	*Ganglia of supply*	*Final transmitters to effector*[b]
Heart (negative chronotropic action)	Vagus	Cardiac	ACh
Heart (negative ionotropic action)	Vagus	Cardiac	ACh
Vasodilation, cranial blood vessels	Facial nerve	Sphenopalatine, otic and cranial microganglia	NO (VIP, ACh)
Vasodilation, genital organs (penile erection, clitoral and labial engorgement), and uterus	Pelvic nerves	Pelvic (hypogastric)	NO (VIP)
Pupillary constriction	Oculomotor nerve	Ciliary	ACh
Accommodation of lens for near vision	Oculomotor nerve	Ciliary	ACh
Lacrimation	Facial nerve	Sphenopalatine	ACh
Parotid gland	Glossopharyngeal nerve	Otic	ACh (VIP)
Submaxillary gland	Facial nerve	Submaxillary	ACh
Gastric glands (acid secretion)	Vagus	Via enteric ganglia	ACh
Gastric relaxation (primarily fundus)	Vagus	Via enteric ganglia	NO (ATP, VIP)
Gastric contraction (primarily antral)	Vagus	Via enteric ganglia	ACh, TK
Exocrine pancreas	Vagus	Via enteric (pancreatic) ganglia	ACh
Pulmonary glands	Vagus	Tracheal ganglia	ACh
Bronchoconstriction	Vagus	Tracheal ganglia	ACh
Bronchodilation	Vagus	Tracheal ganglia	NO (VIP)
Glands of genital organs	Pelvic nerves	Pelvic ganglia	ACh
Gall bladder contraction	Vagus	Intramural ganglia of gall bladder	ACh
Contraction of urinary bladder (micturition reflex)	Pelvic nerves	Pelvic ganglia and intramural ganglia of bladder	ACh, ATP
Initiation of defecation reflex	Pelvic nerves	Via enteric ganglia	ACh, TK

[a]The nerves that carry vagal pathways are cranial nerves III (oculomotor), VII (facial), IX (glossopharyngeal), and X (vagus) and the sacral outflows S2–S4 whose autonomic components continue as the pelvic nerves. There are slight differences in which sacral roots participate, between individuals and between species.
[b]Primary transmitters are indicated, with co-transmitters in brackets.
ACh, acetylcholine; ATP, adenosine triphosphate; CNS, central nervous system; NO, nitric oxide; TK, tachykinin; VIP, vasoactive intestinal peptide family of peptides.

neurons in the capsule of the kidney. Another exception is the pelvic ganglia, which contain the cell bodies both of parasympathetic and of sympathetic postganglionic neurons.

A second feature of the parasympathetic nervous system is that the primary transmitters of the postganglionic neurons are acetylcholine (ACh) and nitric oxide (NO). So far as is known, NO is not a primary transmitter of postganglionic sympathetic neurons. Conversely, noradrenaline is a sympathetic postganglionic transmitter, but it is not used by parasympathetic postganglionic neurons. ACh is a transmitter of a majority of parasympathetic postganglionic neurons, but it is a transmitter only of those sympathetic neurons that innervate the sweat glands. Adenosine triphosphate (ATP) is a transmitter of some postganglionic neurons in each autonomic division.

It is often written that sympathetic pathways become active in emergency functions (flight or fight) and that parasympathetics are active in resting functions, such as digestion. Like other generalizations, this in only partly true. The maintenance of blood pressure at rest is a function of the sympathetic vasoconstrictor neurons. Penile and clitoral engorgement is ordinarily not regarded as a resting function, nor is micturition, although these are all parasympathetic functions. In addition, it is hard to imagine an emergency as profound as drowning, or the threat of drowning, a situation in which there is strong discharge of parasympathetic pathways to the heart.

See also: Acetylcholine Neurotransmission in CNS; Autonomic Nervous System; Nitric Oxide; Stress and Parasympathetic Control; Sympathetic Nervous System.

Further Reading

Furness JB (2006) The organisation of the autonomic nervous system: Peripheral connections. *Autonomic Neuroscience* 130(1–2): 1–5.

Jänig W (2006) *The Integrative Action of the Autonomic Nervous System*. Cambridge, UK: Cambridge University Press.

Robertson D (2004) *Primer on the Autonomic Nervous System*. Amsterdam: Elsevier.

Parietal Cortex and Spatial Attention

M E Goldberg, Columbia University, New York, NY, USA

© 2009 Published by Elsevier Ltd.

Introduction

The brain has a problem: sights, sounds, smells, and touches bombard our sensory apparatus constantly, and the primate brain cannot possibly deal with all of them simultaneously. Instead, it chooses among the myriad objects in our intense sensory world those most relevant to its behavior for further processing. This act of selection is called attention. In 1890, William James described attention as "the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought... It implies withdrawal from some things in order to deal effectively with others." James then described two different kinds of attention:

> It [attention] is either passive, reflex, non-voluntary, effortless or active and voluntary. Voluntary attention is always derived; we never make an effort to attend to an object except for the sake of some remote interest which the effort will serve. In passive immediate sensorial attention the stimulus is a sense-impression, either very intense, voluminous, or sudden...big things, bright things, moving things...blood. (William James (1890), *The Principles of Psychology*)

The attention you are now paying to this article is voluntary. Your momentary distraction by a thunderclap or a flash of lightning would be involuntary. Recently, these two kinds of attention have been described as endogenous, or 'top-down' attention, and exogenous, or 'bottom-up' attention, respectively.

Studies of attention have relied upon a number of different methods to establish the locus of attention. The simplest, but least precise, is *post hoc*: if a person has responded to something, then at some point he/she attended to it. There are three other precisely measurable effects of attention. (1) Attention lowers reaction time: if a cue flashes in the visual field, for the next 300 ms an individual has a faster reaction time to respond to a second stimulus that appears at the position of the cue, and a longer reaction time to respond to a second stimulus that appears away from the cue. (2) Attention improves perception: when for some reason an individual attends to a position in space (e.g., if a cue flashes there), the contrast sensitivity – how different the stimulus must be from the background – improves at the attended location relative to another part of the visual field. (3) Attention limits the effect of distractors: it is easier to pick out a searched-for object among cluttering distractors at the locus of attention.

The posterior parietal lobe, Brodmann's area 7, has long been thought to be important in the neural mechanisms underlying spatial attention. This concept arose from the study of patients with lesions of the parietal lobe, who exhibit two different sorts of deficits. The most severe deficit is a neglect of the contralateral world: patients with right parietal lesions do not recognize that the paralyzed contralateral limb is theirs, nor do they recognize that they are ill. Although they cannot see the left visual field, or the left side of their egocentric space, they are unaware of the deficit. As this extreme deficit recovers they begin to perceive their own contralateral space and limbs, but can have three types of milder deficits. One is a residual neglect: these patients have difficulty constructing the contralateral side of objects. When they are asked to draw a clock they cram the numbers into the right half of the clock (Figure 1). When they are asked to bisect a line they place the midline far to the right of the true midpoint. When they are asked to strike out lines on a page they begin at the far right and often leave out a few lines at the far left. The second deficit is extinction: these patients have an inability to see an object in the affected visual field when a similar object appears in the unaffected visual field, although they can see a single object in the affected field perfectly well. Extinction can be tactile as well – a patient who can appreciate a light tap on his left hand may not perceive the same tap when his right hand is tapped simultaneously. These patients behave as if objects in their unaffected field grab their attention, so they neglect objects in the affected field. The third deficit is in making accurate visually guided movements: patients with parietal lesions often misreach. All of these clinical observations suggest that the parietal lobe is important for three interrelated functions – attention, spatial perception, and the generation of accurate movements to visual stimuli.

Although introspectively we tend to think of perceptual space as being unitary, a homogeneous sphere of which we are at the center, neuropsychological studies have questioned this idea. Thus, certain patients who can bisect a line accurately when they shine a laser pointer at a distant light cannot use a pencil to bisect the same line when it is within arm's reach. Thus, neural mechanisms of spatial attention must distinguish more than the azimuth and elevation of the attended object, but also its distance. One way to explain this multiplicity of attentional spaces is to view them as

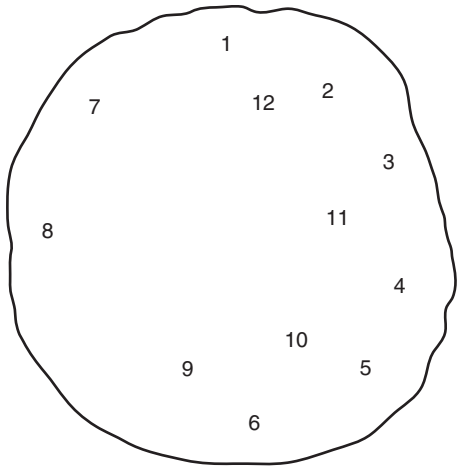

Figure 1 A clock drawn by a parietal patient. The examiner draws the circle and asks the patient with a right parietal lesion to write in the numerals of the clock phase in their proper position. The patient places 1 through 6 in their proper position, but fails to place the remaining numbers properly, putting them on the right side of the clock rather than the left. Adapted from Critchley M (1953) *The Parietal Lobes*. London: Edward Arnold.

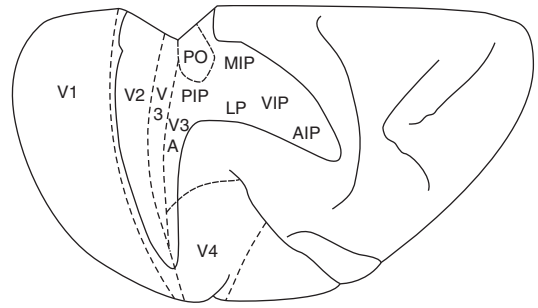

Figure 2 Location of visually responsive areas in rhesus monkey cortex. In this dorsal view of the right hemisphere, the lunate and intraparietal sulci are opened up to show the locations of several extrastriate areas and visually responsive areas within the intraparietal sulcus. Visual areas V2, V3, and V3A are shown in the lunate sulcus. Other areas: PO, the parieto-occipital area; PIP, the posterior intraparietal area; MIP, the medial intraparietal area; LIP, the lateral intraparietal area; VIP, the ventral intraparietal area; AIP, the anterior intraparietal area. Area 7a lies on the surface of the prelunate gyrus. Adapted from Colby CL, Gattass R, Olson CR, et al. (1988) Topographic organization of cortical afferents to extrastriate visual area PO in the macaque: A dual tracer study. *Journal of Comparative Neurology* 269: 392–413.

related to motor workspaces. Thus, attention for far space ranges through the part of space uniquely explored by eye movements. Attention for near space ranges through the workspace for reaching movements.

A second major function of the posterior parietal lobe is to describe the world for action. When you reach out to grab a thick object such as an apple, you open your hand so that when your fingers reach the apple they are opened wider than the apple, and you easily close your hand around the apple. When you reach out to grasp a pencil, you open your fingers much less widely than you did when you grasped the apple. Patients with parietal lesions cannot do this: they cannot use visual information to set their grasp, even though they can tell you whether the apple or the pencil is wider.

Physiological studies of parietal function have emphasized both of these major functions of the parietal lobe: generating spatial attention and analyzing the visual world for action. Most of our understanding of the neurophysiology of the posterior parietal lobe comes from studies of the parietal lobe of the rhesus monkey. Brodmann described the posterior parietal cortex as a single area, area 7, on the basis of its cellular morphology. Recent studies have shown that area 7 is divided into a number of different regions, with differing anatomical projections and physiology. Many of these subdivisions lie within the intraparietal sulcus (Figure 2). Each of the areas in the intraparietal sulcus has neurons which respond to the appearance of visual stimuli. These neurons do not respond to the appearance of stimuli everywhere in the visual field; instead, they respond only to stimuli which appear in a limited part of the visual field. The part of the visual field which can excite the neuron is called its receptive field, and the receptive fields of parietal neurons usually lie in the visual field opposite the hemisphere in which the neuron lies – the contralateral visual field. The properties of each individual area make them suitable for representing objects which can be the target for a specific set of movements – a 'motor workspace'.

Each area in the intraparietal sulcus sends projections to premotor areas which define the actual movements in that workspace. Thus, the ventral intraparietal (VIP) area has neurons which represent the near space around the mouth, and the tactile areas on the face. The neurons have congruent responses: neurons which represent the upper visual field also represent the upper face (Figure 3). Many VIP neurons respond to visual stimuli close to the face, but not to stimuli with the identical visual angle further away. The VIP area projects to area F4 in premotor cortex, which represents movements of the face and head. The medial intraparietal (MIP) area has neurons responding to tactile stimulation of the arm and discharges when the monkey reaches to targets. The MIP area projects to F2 in premotor cortex, which represents movements of the arm. The anterior intraparietal (AIP) area has neurons which respond to visual stimuli which can be grasped, and which discharge when a monkey makes a grasping movement in the dark. The AIP area projects to area F5 in premotor cortex, which is important in the generation of grasping movements.

The most extensively studied area in the intraparietal sulcus is the lateral intraparietal (LIP) area. This area has reciprocal connections with two areas important in

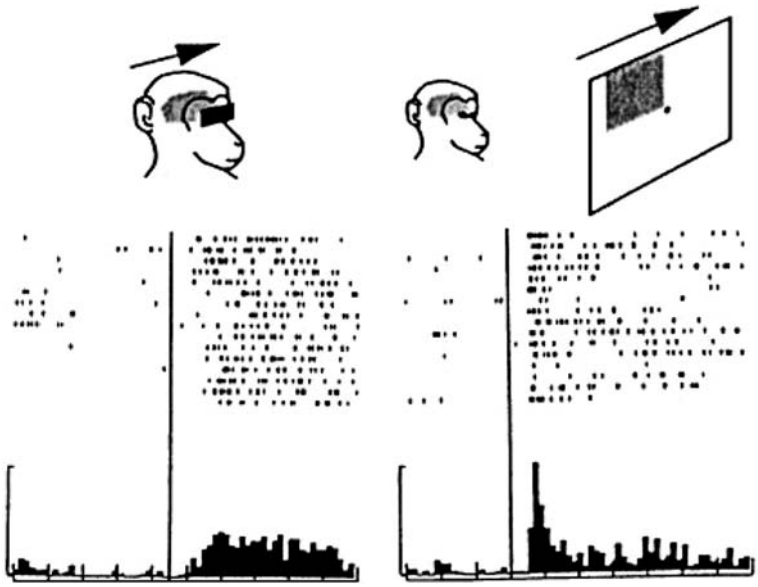

Figure 3 Correspondence of visual and tactile receptive fields in the ventral intraparietal area of the monkey. The top of each column has a drawing depicting the experiment, and the raster diagram of the activity of a single neuron in the ventral intraparietal area under different circumstances is below, showing the individual neuronal response. Each dot is an action potential; each line is a single trial. Successive lines are synchronized on a single event – on the left when the experimenter began to stroke the animal, and on the right when the visual stimulus appeared. The vertical lines show when the synchronizing event began. Histograms beneath the rasters sum the activity of the rasters. Each tick in the raster is 200 ms. The raster in the left column shows the tactile responses and that in the right column shows the visual responses of a neuron in ventral intraparietal area. The receptive fields correspond if one considers the mouth to be analogous to the fovea. Each depicted neuron was directionally selective: Visual and tactile stimuli moving in the opposite direction had little or no effect on the neuron (not shown here). Reproduced from Duhamel J-R, Colby CL, and Goldberg ME (1998) Ventral intraparietal area of the macaque: Congruent visual and somatic response properties. *Journal of Neurophysiology* 79: 126–136, with permission.

the generation of rapid eye movements (saccades), the frontal eye field (FEF) and the superior colliculus (SC). The projection from the SC to the LIP area is disynaptic via the thalamus; the projections to and from the FEF are monosynaptic. The LIP area also has reciprocal connections with areas important for the analysis of patterns in the visual world: prestriate area V4, and TEO and TE in the inferior temporal cortex. Activity in the LIP area is consistent with providing information both for generation of eye movements and for pinning attention in the visual field.

LIP neurons do not respond to every object in the visual field. Most of the objects in the visual field, especially those that are stable and irrelevant to an individual's behavior, are not represented by activity in the LIP area. Instead, only those objects which can be the object of voluntary or involuntary attention are represented. The same object can evoke little or no response when it appears in the receptive field of neuron, or a very strong response, depending upon the behavioral significance of the object. LIP neurons are closely related to the neural mechanisms for the control of saccadic eye movements. When a monkey plans a memory-guided saccade to the spatial location of a recently vanished stimulus in a LIP neuron's receptive field, the neuron discharges in response to the appearance of the stimulus, during the delay period, and before the saccade itself. This activity is consistent with saccade planning, and also voluntary attention, because attention, as measured by an improvement in perceptual threshold, lies at the goal of the memory-guided saccade. However, LIP neurons also respond to the abrupt onset of a distractor which appears in the receptive field when the animal is planning a saccade elsewhere. Attention lies at the goal of the planned saccade unless a distractor appears elsewhere in the visual field, at which point the mechanisms for involuntary attention overwhelm the voluntary attention induced by the eye movement, and attention shifts away from the goal of the saccade to the location of the distractor. The activity of the LIP area in this case parallels the location of the monkey's attention. The time at which attention returns from the distractor to the saccade goal is predicted by the time at which the activity evoked by the distractor falls below that evoked by the saccade plan (Figure 4). Under the circumstances of this experiment, although LIP reflects a saccade plan, its activity does not necessarily correlate with when, where, or if the monkey will make a saccade – the distractor has no effect on the monkey's eye movements. Instead, its activity correlates with the

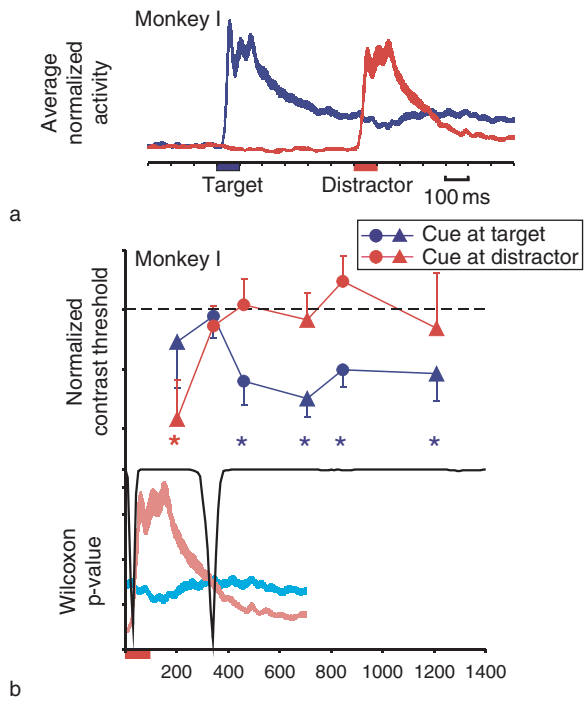

Figure 4 Top-down and bottom-up attentional activity in the lateral intraparietal area. (a) Average activity of 23 lateral intraparietal neurons during the delayed saccade task. Blue trace: the saccade target appears briefly in the receptive field (blue bar). The normalized neuronal spike density is calculated with a sigma of 10 ms (ordinate) plotted against time (abscissa). The thickness of the line is the standard error of the mean of the activity, Note the onset transient in response to the target appearance, followed by delay period activity well above the background activity, despite the fact that the target is no longer present. The delay activity continues until the monkey makes the saccade (not shown). Red trace: distractor appears in the receptive field (red bar) while the monkey is planning a saccade away from the receptive field. Note the brisk onset transient and decay of activity beneath the level of the saccade delay period activity. (b) Relationship of the monkey's attention to neuronal activity in the lateral intraparietal area. The top trace shows the monkey's perceptual performance during the task. The normalized contrast threshold for a stimulus flashed for one video frame (abscissa) is plotted against time from the appearance of the distractor (ordinate). A significant decrement in contrast threshold denotes that the monkey's attention lies at the spatial location of the stimulus. Before the distractor appeared, the perceptual threshold was lower at the saccade goal (not shown). At 200 ms after the distractor appeared, the perceptual threshold was significantly lower at the distractor site but not at the saccade goal. The bottom trace shows the neuronal activity during this epoch. Note that when the neuronal activity was greater at the distractor site (red trace) than it was at the saccade goal (blue trace), the distractor site was the locus of attention. When there was no difference between the activity at the distractor site and the saccade goal (the black line shows the Wilcoxon *p* value of the difference between the two points), there was no attentional advantage at either site. When the neuronal advantage became significant again at the saccade goal, the attentional advantage returned to the saccade goal and remained there throughout the delay period.

location of the monkey's attention on a millisecond-by-millisecond basis, including the attention that is pinned to a location in space by the plan to make a saccade there. In contrast, when a monkey is exploring the visual field in a search task, activity in the LIP area predicts the goal and reaction time of the saccade. Thus activity in the LIP area is best understood as representing a map of the salient locations in the visual field in the far space explored by eye movements. The visual system presumably uses this representation to pin attention at the peaks in the salience map; the oculomotor system can drive the eyes to the peak of the salience map when a saccade is appropriate.

Neurons in the LIP area have retinotopic receptive fields. The visual, memory, and saccade-related signals carried by these neurons describe stimuli in terms of the distance and direction of the stimulus or saccade location relative to the center of gaze. This retinotopic organization presents a paradox: how can accurate spatial information be derived from retinotopic input? Every time we move our eyes, each object in our surroundings activates a new set of retinal neurons. Despite these changes, we experience the world as stable and move accurately in it. This perceptual stability has long been understood to reflect the fact that what we see is not a direct impression of the external world but an internal image of it. The brain must construct a representation that can compensate for changes in eye position. Neurons in the LIP area compensate for eye movements: if a stimulus flashes outside the receptive field of a neuron, but an eye movement brings the spatial location of the vanished stimulus into the receptive field, neurons in the LIP area will respond, after the eye movement, to a stimulus which never appeared in their receptive field, but would do so if it were still illuminated after the eye movement (Figure 5). Some neurons respond to stimuli which will be brought into their receptive fields by an impending saccade even before the saccade. Thus, LIP neurons update their representation of the visual world, compensating for eye movements, so that there is a continuously spatially accurate representation of the visual world in the LIP area despite a moving eye. Neurons in the MIP area, which describe the workspace for reaching movements, also describe the world in retinal coordinates and compensate for saccades in a similar way. In keeping with this, patients with right parietal lesions cannot make this compensation: they cannot make a saccade to a stimulus which appears and disappears before an intervening saccade. This suggests that without an intact parietal lobe they cannot compensate for

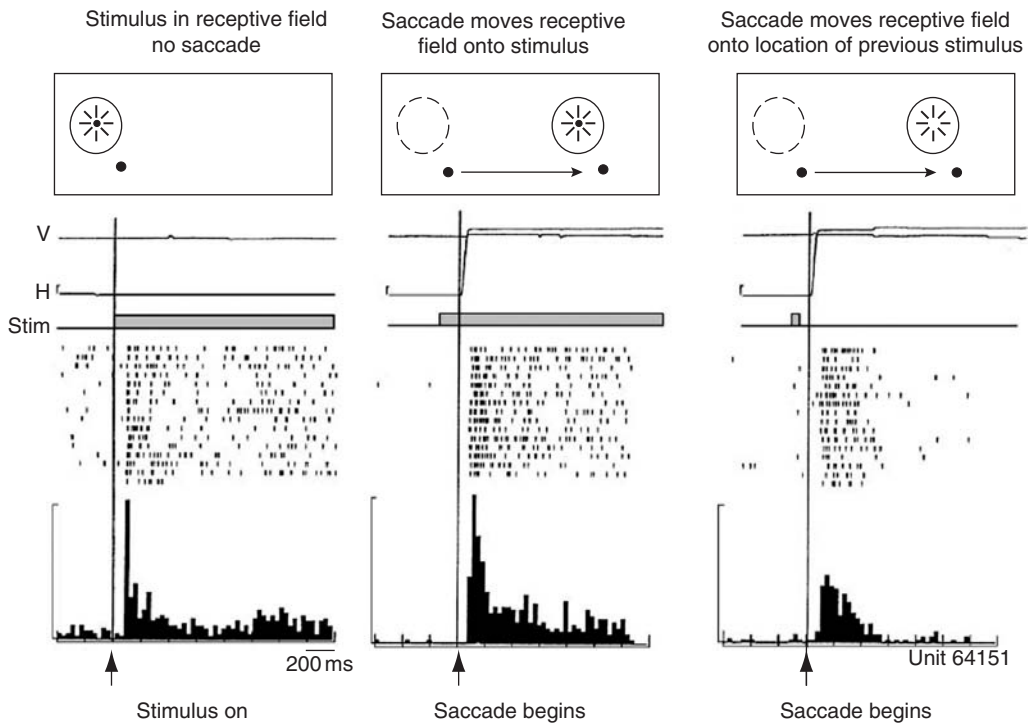

Figure 5 Remapping of visual activity in the lateral intraparietal area. Diagrams show the arrangement of the fixation point, receptive field, stimulus location, and saccade. (Left) In the fixation task, the neuron responds to the onset of a stimulus in the receptive field. (Middle) The same neuron responds when a saccade moves the receptive field onto the location of a recent stimulus. (Right) Response after a saccade moves the receptive field onto a previously stimulated location. The stimulus is flashed on for only 50 ms and is extinguished before the saccade begins. The neuron is responding to the remapped memory trace of the previous stimulus. Control experiments (not shown) indicate that neither the stimulus alone nor the saccade alone can drive the neuron. Adapted from Duhamel J-R, Colby CL, Goldberg ME, et al. (1992) The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* 255: 90–92.

their own eye movements. Thus, the neural activity in the LIP area reflects the three functions described by lesions in patients: attention, spatial perception, and describing the visual world for movement. Whether the other areas in the intraparietal sulcus perform the same functions for the motor workspaces they describe will be the subject for future research.

See also: Attentional Networks in the Parietal Cortex; Neglect Syndrome and the Spatial Attention Network; Neural Coding of Spatial Representations; Spatial Memory: Assessment in Animals; Spatial Cognition and Executive Function; Spatial Cognition; Spatial Cognitive Maps; Vision for Action and Perception; Visual–Vestibular Interactions.

Further Reading

Colby CL, Gattass R, Olson CR, et al. (1988) Topographic organization of cortical afferents to extrastriate visual area PO in the macaque: A dual tracer study. *Journal of Comparative Neurology* 269: 392–413.

Colby CL and Goldberg ME (1999) Space and attention in parietal cortex. *Annual Review of Neuroscience* 23: 319–349.

Critchley M (1953) *The Parietal Lobes*. London: Edward Arnold.

Duhamel J-R, Colby CL, Goldberg ME, et al. (1992) The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* 255: 90–92.

Duhamel J-R, Colby CL, and Goldberg ME (1998) Ventral intraparietal area of the macaque: Congruent visual and somatic response properties. *Journal of Neurophysiology* 79: 126–136.

James W (1890) *The Principles of Psychology*. New York: Holt.

Parkinsonian Syndromes

C Schwarz, New York Presbyterian Hospital – Weill Cornell Medical Center, New York, NY, USA
C Henchcliffe, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Clinical Features

Parkinsonian syndromes occur as part of multiple underlying pathologies that involve the striatum (Table 1). This article focuses on Parkinson's disease (PD), which is the most common cause of parkinsonism, accounting for approximately 80% of cases. PD is the second most common neurodegenerative disease after Alzheimer's disease. Most cases of PD occur between ages 40 and 70 years, with the peak of onset in the sixth decade. In the United States, the prevalence is 1% by age 65 years and 3.5% by age 85 years. PD is suggested to arise due to both environmental and genetic causes. In approximately 90% of cases, no apparent genetic defect can be detected, and it is referred to as idiopathic. The disease is progressive, and treatment is mainly symptomatic. Clinically, PD is characterized by cardinal motor features that are caused by a loss of dopaminergic neurons in the substantia nigra, which project to the striatum. They include tremor at rest with a frequency of 4–7 Hz, lead-pipe rigidity of the extremities on examination, bradykinesia and akinesia (slowness and paucity or absence of movement), and postural instability later in the course of disease. Patients suffer from hypophonia (decreased voice volume), dysphagia, drooling, hypomimia (loss of facial expression or 'masked facies'), micrographia (decreased size of handwriting), flexed posture, shuffling gait, and freezing (motor block, often affecting gait). The onset is gradual and usually asymmetric. Dystonia may be associated with PD early in a minority of patients or, more often, later in the course of the disease. PD patients also suffer from a variety of nonmotor symptoms, including cognitive and behavioral changes, depression and anxiety, olfactory loss, sleep disorders, and autonomic symptoms. These are related to neurodegeneration in brain regions other than the substantia nigra and may precede motor symptoms. The longitudinal course of PD is followed clinically using the Unified Parkinson Disease Rating Scale, which consists of ratings for mentation, behavior and mood, the ability to conduct activities of daily living, and motor performance.

The previously mentioned features of PD are found in varying combinations in other parkinsonian syndromes (Table 1), raising a diagnostic challenge in some individuals. The presence of atypical features, however, may raise suspicion. Progressive supranuclear palsy (PSP) typically manifests with symmetric parkinsonism, impaired oculomotor function (particularly vertical gaze), prominent axial rigidity, and early falls. The cortical involvement of corticobasal-ganglionic degeneration may lead to apraxia, alien limb, cortical sensory impairment, and myoclonus, in the setting of asymmetric parkinsonism. Multiple system atrophy (MSA) may comprise prominent parkinsonism (striatonigral degeneration or MSA-P); cerebellar signs and symptoms (olivopontocerebellar atrophy or MSA-C); or autonomic dysfunction, including orthostatic hypotension and syncope, urinary symptoms, constipation, and sexual dysfunction (Shy–Drager syndrome or MSA-A). These Parkinson's-plus syndromes typically display more rapid progression and little, if any, response to dopamine replacement. The presence of dementia early in disease should raise suspicion of dementia with Lewy bodies (DLB), typically associated with hallucinations and fluctuating clinical symptoms. However, parkinsonism is also a feature of Alzheimer's disease and frontotemporal dementia in some patients, and tau mutations on chromosome 17 lead to autosomal dominant inherited frontotemporal dementia with parkinsonism (FTDP-17).

Neuroimaging

In general, patients with parkinsonian syndromes undergo conventional magnetic resonance imaging (MRI) to rule out any structural abnormalities that may lead to parkinsonism, such as tumor, stroke or infection of the basal ganglia, or hydrocephalus. The neurodegenerative diseases under discussion in this article typically have subtle or no abnormalities revealed by MRI. However, in some with MSA, T2-weighted images demonstrate hyperintense signals along the lateral aspect of the putamina, the 'hot cross bun' sign due to degeneration of neural pathways in the pons, and/or cerebellar atrophy. Positron emission tomography (PET) and single photon emission computerized tomography (SPECT) allow imaging of basal ganglia and nigrostriatal tract integrity using appropriate radioligands. Fluorodopa PET has revealed progressive presynaptic dopaminergic nigrostriatal damage in PD patients: while not a satisfactory measure of progression, it may help to assess nigrostriatal tract integrity in atypical cases. Fluorodeoxyglucose PET shows normal or increased metabolism in the

striatum in PD, whereas in Parkinson's-plus syndromes there may be decreased striatal metabolism. Currently, however, PET and SPECT are little used in clinical management of PD and other parkinsonian syndromes.

Table 1 Differential diagnosis for parkinsonian syndromes

Parkinson's disease (sporadic, familial)
Secondary parkinsonism
Toxic: MPTP, manganese, copper, cyanide
Trauma
Tumor
Vascular
Postencephalitic
Hypoxia
Carbon monoxide
Normal pressure hydrocephalus
Drug-induced: dopamine antagonists and depletors
Hemiatrophy–hemiparkinsonism
Multisystem degeneration (Parkinson's-plus syndromes)
Multiple asystem atrophy (MSA)
MSA-P (striatonigral degeneration)
MSA-A (Shy–Drager)
MSA-C (sporadic olivopontocerebellar atrophy)
Progressive supranuclear palsy
Corticobasal ganglionic degeneration
Progressive pallidal atrophy
Lytico–Bodig (Parkinson–dementia–ALS complex of Guam)
Dementia associated: dementia with Lewy bodies, frontotemporal dementia, Alzheimer's disease
Hereditary neurodegenerative diseases
Lubag
Dopa-responsive dystonia
Olivopontocerebellar and spinocerebellar atrophy
Huntington's disease
Mitochondrial myopathies
Neuroacanthocytosis
Wilson's disease
Neurodegeneration with brain iron accumulation type1

Pathology

The disease is characterized by loss of dopaminergic neurons in the substantia nigra pars compacta (SNpc), which affects nigrostriatal dopaminergic pathways leading to striatal dopamine (DA) deficiency. Depigmentation of the SNpc in autopsy tissue is caused by loss of these neurons, which contain neuromelanin. Findings indicate a more widespread pathology with neurodegeneration also involving the olfactory bulb, the dorsal motor nucleus of the vagus nerve, the nucleus basalis of Meynert, the superior cervical ganglia, the mesenteric plexus, and the locus coeruleus (Table 2). Neurodegeneration in these areas likely precedes neuronal death in the SNpc and accounts for some of the nonmotor symptoms in PD. Clinical symptoms of PD emerge when approximately 80% of striatal dopamine and 50% of nigral dopaminergic neurons are lost. Thus, early diagnosis remains a challenge.

A pathological hallmark of PD is the formation of intraneuronal cytoplasmic inclusion bodies, termed Lewy bodies (Figure 1). The major components of Lewy bodies are α-synuclein and ubiquitin. Lewy bodies occur not only in the SN but also in other areas involved in neurodegeneration mentioned previously. The relationship between Lewy bodies and cell death remains to be determined, and of note, they are not found in PD arising from parkin mutations nor in all cases of PD associated with LRRK2 mutations.

Etiology and Pathogenesis

Multiple lines of evidence point to a combination of environmental factors and genetic mutations or susceptibility factors as the cause of PD, and their relative contributions vary between individuals.

Table 2 Proposed pathologic stages of Parkinson's disease (Braak classification)

Stage	*Location*	*Structures affected*
1	Medulla oblongata	Dorsal nucleus of vagal and glossopharyngeal nerves and/or intermediate reticular zone; olfactory bulb and anterior olfactory nucleus
2	Medulla oblongata and pontine tegmentum	Pathology of stage 1 plus caudal raphe nuclei, gigantocellular reticular nucleus, and coeruleus–subcoeruleus complex
3	Midbrain	Pathology of stages 1 and 2 plus midbrain, especially in the pars compacta of the substantia nigra
4	Basal prosencephalon and mesocortex	Pathology of stages 1–3 plus prosencephalic and cortical involvement confined to the temporal mesocortex (transentorhinal region) and allocortex (CA2 plexus)
5	Neocortex	Pathology of stages 1–4 plus higher order sensory association areas of neocortex and prefrontal neocortex
6	Neocortex	Pathology of stages 1–5 plus sensory association areas of the neocortex and premotor areas; occasionally mild changes in primary sensory areas and the primary motor field

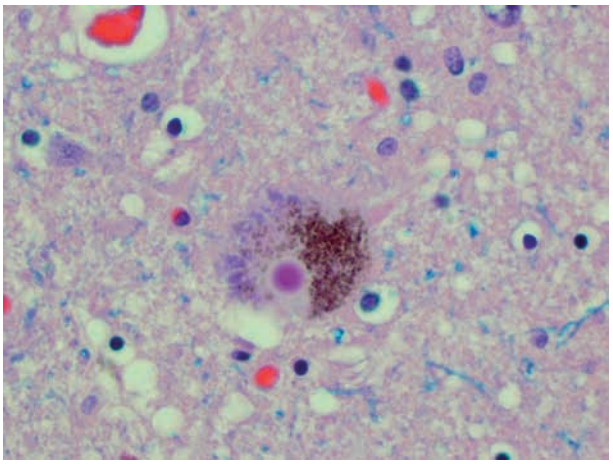

Figure 1 Photomicrograph of a hematoxylin–eosin-stained section taken from the substantia nigra pars compacta at autopsy of a PD patient. A surviving dopaminergic neuron in the center of the field contains a cytoplasmic Lewy body, with a central eosinophilic dense core surrounded by a halo. Also in the cytoplasm are numerous neuromelanin granules (brown pigment). Cortical Lewy bodies are less well-defined in structure. Courtesy of Ehud Lavi, MD, Department of Pathology, Weill Medical College of Cornell University.

Several studies have examined the role of environmental factors in the pathogenesis of PD. These have identified a number of risk factors, including rural living, pesticide exposure, well water consumption, as well as mining and welding. Exposure to manganese, a component of herbicides, pesticides, and welding fumes, produces a parkinsonian syndrome and leads to damage of the globus pallidus with sparing of the substantia nigra. Exposure to the pesticide rotenone has been shown to reproduce neuropathological features of PD with cytoplasmic inclusions that contain ubiquitin and α-synuclein in rats. Ingestion of annonacin in the Caribbean is associated with parkinsonism in humans and induces nigral and striatal neurodegeneration in rats. Exposure of humans to 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) results in an acute parkinsonian syndrome. These findings have been used to create animal models of PD by administering a variety of the previously mentioned toxins to different species. In contrast, other environmental factors, such as cigarette smoking, alcohol, and caffeine, are associated with a lower risk for PD, although the reason for this negative association remains to be elucidated.

PD Genetics: Implications for PD Pathogenesis

In the past few years, various gene mutations have been identified in familial PD (Table 3), beginning with the discovery of α-synuclein mutations in the large Contursi kindred in 1997. Even though only approximately 10% of PD cases are clearly hereditary, the study of genetic forms of the disease offers valuable insights into the pathogenesis of PD (Table 3). Aided by the discovery of specific genes mutated in familial PD, genetic animal models in different species have been generated to address the previous questions and to test pharmacological interventions that could be applied to nonfamilial forms of PD as well. The discovery of mutations in several PD genes involved in the function of the ubiquitin–proteasome system provides evidence that dysfunction of this pathway leads to abnormal protein aggregation, Lewy body accumulation, and neuronal death in PD. Study of normal and mutant function of several PD genes also strongly implicates a role for mitochondrial dysfunction and oxidative stress in PD pathogenesis. In addition, oxidative stress or environmental toxins most likely contribute to proteasomal dysfunction.

α-Synuclein (PARK1) was the first gene to be identified in hereditary PD, but mutations in this gene affect only a very small number of families. α-Synuclein is a 140-amino acid presynaptic protein that is involved in vesicle trafficking and neurotransmitter release, and it is a major component of Lewy bodies in PD and DLB. Three missense mutations and several genomic multiplications of the α-synuclein gene have been identified in hereditary PD. However, promoter polymorphisms may also be associated with sporadic PD. Mutant or overexpressed α-synuclein is misfolded and aggregates within the cell, thus impairing degradation by the ubiquitin–proteasome system. Transgenic mouse lines have been generated, including α-synuclein knockouts, and lines overexpressing human α-synuclein under the control of various promoters. The behavioral phenotype, such as decreased locomotor activity and increased righting time on an inverted screen, correlates with the level of overexpression of the protein. However, only a few of the lines have dopaminergic neurodegeneration in the striatum, which limits their usefulness as PD models. In contrast, injection of viral vectors containing human α-synuclein into adult rat SN leads to formation of α-synuclein-positive, Lewy body-like inclusions and dopaminergic neuronal cell death. A *Drosophila* model of PD overexpressing wild-type or mutant α-synuclein has been developed. Dopamine neuron loss in these flies was associated with Lewy body-like inclusions and behavioral changes that were corrected by administration of dopamine agonists and levodopa.

Mutations in the parkin gene (PARK2) cause juvenile autosomal recessive parkinsonism. They are found at high frequency in early onset PD (18% sporadic early

Table 3 Parkinson's disease genetics

Gene	*Inheritance*	*Function*	*Observations*
α-Synuclein (PARK1, -4)	Dominant	?	Aggregation into oligomers or fibrils Substrate for ubiquitin–proteasomal system Knockout mutant: ↑ resistance to MPTP (mice) Mutation: A53T overexpression; abnormal mitochondria (mice)
Parkin (PARK2)	Recessive	Ubiquitin E3 ligase	Ubiquitin–proteasomal system impairment Mitochondrial outer membrane association (partial) Mutation: abnormal mitochondria, ↑ sensitivity to oxidative stress (*Drosophila*)
UCH-L1 (PARK5)	Recessive	Ubiquitin C-terminal esterase L1	Ubiquitin–proteasomal system impairment Requires confirmation as PD gene in further families
PINK1 (PARK6)	Recessive	Mitochondrial serine–threonine kinase	Mitochondrial membrane localization siRNA: ↑ sensitivity to MPP+, rotenone Mutation: abnormal mitochondria, ↑ sensitivity to oxidative stress (*Drosophila*)
DJ-1 (PARK7)	Recessive	Oxidative stress sensor, chaperone	Oxidative stress causes re-localization to mitochondria (matrix/intermembrane space) Protects against oxidative stress siRNA: ↑ sensitivity to oxidative stress (*Drosophila*) Mutation: ↑ sensitivity to rotenone, paraquat, hydrogen peroxide (*Drosophila*); ↑ sensitivity to oxidative stress (mice)
LRRK2 (PARK8)	Dominant	Serine–threonine kinase	GTPase or MAPKKK dysregulation
Lysosomal ATPase (PARK9)	Recessive	?	Responsible for Kufor–Rakeb syndrome of parkinsonism with dementia

onset and 49% familial autosomal recessive early onset). Heterozygous mutations in the parkin gene might constitute a risk factor in sporadic PD, and polymorphisms in the parkin gene appear to confer an increased risk toward the development of idiopathic PD. Parkin is a ubiquitin E3 ligase that is involved in the ubiquitination of proteins prior to degradation by the proteasome. Mutations in parkin therefore interfere with normal clearing of proteins by the cell and lead to accumulation of parkin substrates, such as α-synuclein, Pael-R, synphilin 1, and CDCrel-1. A role for parkin in mitochondrial function has also been suggested. Parkin knockout mice, however, do not display accumulation of parkin substrates and there is no loss of dopaminergic neurons. These mice do show altered dopaminergic neurotransmission as well as behavioral changes but are not regarded as a good model for parkinsonism. Ubiquitin–proteasomal system dysfunction is further implicated by the finding of ubiquitin C-terminal hydrolase L1 (UCH-L1) mutation in a family with autosomal recessive PD, but its role in PD remains to be confirmed with additional families.

PINK1 (PTEN-induced kinase1) is another cause of autosomal recessive PD, although far less commonly than parkin. It localizes to the mitochondria and possibly exerts a protective effect on the cell.

Deletions or point mutations in the DJ-1 gene (PARK7) cause rare autosomal recessive early onset PD. DJ-1 is a multifunctional oncogene product, a regulatory subunit of RNA binding protein, and has been implicated in the cellular response to oxidative stress. DJ-1 knockout mice show an increased vulnerability to oxidative stress and decreased dopamine release from striatal neurons but no loss of dopaminergic neurons. These mice have decreased locomotor activity and increased sensitivity to the neurotoxin MPTP.

Leucine-rich repeat kinase 2 (LRRK2) is increasingly recognized as an important cause of autosomal dominant PD in particular populations. Its function remains to be elucidated, but its structure comprises a Ras-GTPase domain and a MAPKKK domain, thus implicating dysfunction of multiple cell signaling pathways.

Several important questions remain in the pathogenesis of genetic PD, such as the basis for selective neuronal vulnerability and the exact mechanism of cell death. It is unclear why there is a delayed clinical presentation in adulthood in the genetic forms of PD, even though the gene is expressed at or before birth. It is very likely that there are molecular changes in the affected cells before expression of the phenotype, which cannot yet be studied in patients.

Neurotoxin-Based Animal Models of PD

Nongenetic, neurotoxin-based animal models of PD have played a critical role in understanding the pathways that lead to selective degeneration of dopaminergic neurons in sporadic PD. Toxin-induced animal

models are based on the assumption that dopaminergic neurons follow a common pathway of cell death regardless of the initial stimulus. The toxins used are all thought to provoke the formation of reactive oxygen species. Elucidating the cascade that leads to cell death in these models might contribute to a better understanding, identify novel molecular pathways that might be targeted to prevent degeneration, and might lead to the discovery of novel treatment strategies. Moreover, they may prove valuable for preclinical testing of novel treatments that modulate the ubiquitin–proteasome pathway or mitochondrial function.

Exposure of humans and nonhuman primates to MPTP, a by-product of chemical synthesis of a synthetic heroin analog, results in an acute parkinsonian syndrome that recapitulates many clinical features of PD. Its metabolite, 1-methyl-4-pyridinium (MPP+), binds to the dopamine transporter and thus accumulates in dopaminergic neurons. It inhibits NADH dehydrogenase and NADH CoQ reductase of complex I of the mitochondrial electron transport chain (ETC), resulting in cell death of dopaminergic neurons of the SNpc.

Local stereotactic injection of the neurotoxin 6-hydroxydopamine (6-OHDA) into the SN or striatum of animals induces selective degeneration of the nigrostriatal dopamine system, resulting from preferential uptake by dopamine and noradrenergic transporters. Animals with 6-OHDA-induced lesions undergo extensive loss of DA neurons in the basal ganglia, with corresponding neurological impairments including akinesia, dyskinesias in response to dopaminergic treatment, and improved motor function after the administration of dopaminergic agents. However, 6-OHDA injections do not produce Lewy body-like inclusions.

The herbicide, paraquat, and the pesticide, rotenone, are also used to produce toxic animal models of PD. Intravenous infusion of rotenone, which binds to and inhibits complex I of the mitochondrial respiratory chain, produces dopaminergic degeneration in the SN and induces α-synuclein-positive Lewy body-like inclusions in rats. Animals treated with rotenone develop bradykinesia and abnormal posturing. Paraquat shows structural similarity to MPP+. When infused into mice, it leads to dopaminergic degeneration in the SN and formation of α-synuclein-positive inclusions.

Oxidative Stress and Mitochondrial Dysfunction in PD Pathogenesis

There is mounting evidence supporting mitochondrial dysfunction, either primarily or secondarily, in the PD pathogenetic process, and several PD genes are suggested to be involved in mitochondrial function and oxidative stress (Table 3). The mitochondrial ETC is a major source of highly reactive free radicals, particularly when impaired, that may oxidatively damage neighboring molecules. Moreover, impaired mitochondrial function increases susceptibility to oxidative stress.

Complex I is the largest of the ETC complexes, and it is the major site of superoxide production in the ETC. Its activity in PD patients is decreased in the SN and other brain regions, and cell cybrids (fusion of platelets isolated from PD patients with cell lines deficient in mitochondria) can have a complex I deficiency. Moreover, higher levels of acquired mitochondrial DNA deletions have been demonstrated in autopsy brain tissue in PD compared with non-PD controls. Finally, several studies have suggested that specific clusters of mitochondrial DNA haplogroups may decrease the risk of PD. Postmortem analysis of SN tissue has provided direct evidence of oxidative stress in PD, with increased levels of malondialdehyde and cholesterol lipid hydroperoxides and others. Moreover, oxidative damage impairs the ubiquitination and degradation of proteins mediated by the ubiquitin–proteasomal system, suggesting a link between oxidative damage and characteristic abnormal protein aggregates. In addition to mitochondrial ETC dysfunction, impaired antioxidant mechanisms and increased exposure to environmental or endogenous sources of oxidative stress may all contribute to increased oxidative stress in PD. Glutathione, an important antioxidant, is decreased in the SN by approximately 50% in PD patients. Iron can act as a catalyst in the formation of hydroxyl radicals from hydrogen peroxide. Neuromelanin, present in some dopamine neurons, chelates iron but is present at decreased levels in PD substantia nigra. Several studies have demonstrated increased iron levels in the brain of PD patients, as well as decreased amounts of ferritin, which inhibits iron's catalytic properties. Therefore, the potential use of iron chelators, combined with monoamine oxidase (MAO) inhibition, is being pursued.

Current Treatments for PD

Pharmacologic treatment is symptomatic. The mainstay of PD treatment is the DA precursor L-DOPA (L-3,4-dihydroxyphenylalanine). L-DOPA provides adequate control of akinesia and tremor for several years, but over time treatment is complicated by the development of motor fluctuations and dyskinesias, requiring more advanced therapeutic strategies. Many clinicians therefore prefer to use alternative therapies, such as dopamine agonists or MAO-B inhibitors, when possible, early in the disease. These may also be used as adjunctive therapies in later disease, along with

catechol-*O*-methyltransferase inhibitors such as entacapone, which inhibit levodopa metabolism and help decrease motor fluctuations. Other agents in use are amantadine, with multiple mechanisms of action including NMDA receptor blockade, and anticholinergic agents, which are helpful for tremor. Side effect profiles of these, however, are often limiting. Cholinesterase inhibitors seem to have mild benefit for cognitive dysfunction, and various antidepressants are under investigation for PD-associated depression.

Deep brain stimulation (DBS) is a surgical approach to treating advanced PD in which electrodes are implanted into specific brain regions. Continuous high-frequency stimulation is then delivered via an externally programmable pulse generator. Stimulation of either the subthalamic nucleus (STN) or the internal segment of the globus pallidus (GPi) relieves bradykinesia, rigidity and tremor, as well as drug-induced dyskinesias in PD patients. DBS of the GPi and STN can reduce the need for dopaminergic medications and alleviate motor fluctuations. Stimulation of the ventral intermediate nucleus of the thalamus can reduce tremor associated with PD, but since other symptoms usually become more prominent over time, this target is seldom used. Although the exact mechanism of action remains to be determined, it is thought that DBS modulates a motor circuit formed by the basal ganglia, thalamus, and cortex that is dysfunctional in PD. Dopamine depletion in the basal ganglia induces changes in electrical activity in this circuit, resulting in increased output from the GPi and STN. DBS appears to act in restoring electrical imbalance in the affected pathways. Indications for DBS in PD are intractable tremor, drug-related motor fluctuations, and dyskinesias, with a good previous clinical response to medical treatment with L-DOPA being the best prognostic indicator for success of DBS. Unfortunately, 'midline symptoms,' such as postural instability, and nonmotor symptoms, such as cognitive decline, may continue to worsen after DBS.

New Directions in PD Therapy

A great deal of research is dedicated toward identifying neuroprotective agents to prevent or treat PD. Several trials have investigated the neuroprotective properties of selegiline, a selective, irreversible inhibitor of MAO-B, which inhibits dopamine metabolism. Its metabolite, desmethylselegiline, has been shown to protect animals against MPTP toxicity by inducing transcriptional changes leading to upregulation of antiapoptotic and downregulation of proapoptotic genes. In PD patients, selegiline significantly delayed the need for levodopa compared to placebo, but this was most likely due to symptomatic effect, and any potential neuroprotective activity remains highly controversial. The Food and Drug Administration has approved rasagiline (*N*-propargyl-1R-aminoindan), a novel, highly potent irreversible MAO-B inhibitor for the treatment of PD. Studies have shown a neuroprotective effect in cell culture via activation of the antiapoptotic protein Bcl-2 and downregulation of the proapoptotic Bax family of proteins by its propargylamine moiety. A delayed-start design phase III clinical trial demonstrated greater benefit in subjects taking rasagiline early, suggesting a possible neuroprotective effect that remains to be clarified and is currently under investigation.

Based on increasing understanding of underlying pathogenetic mechanisms of PD, avenues to retard disease progression that include modulation of mitochondrial function, oxidative stress, and ubiquitin–proteasomal system function are being evaluated. One such promising compound is coenzyme Q10 (CoQ10), an electron acceptor for mitochondrial complex I, shown to be decreased in mitochondria of PD patients. Treatment of PD patients with CoQ10 leads to an increase of platelet mitochondrial complex I and III activity. In a phase II clinical trial, high-dose CoQ10 was well tolerated and it may slow clinical progression in PD patients. High-dose CoQ10 is under further evaluation in a large phase III clinical trial.

Restorative therapy is a future goal for treatment, potentially allowing arrest or reversal of disease progression. Transplantation of fetal and stem cells has attracted considerable attention since it might lead to a more physiologic replacement of dopamine, potentially avoiding the motor complications associated with long-term medical treatment. Based on evidence from animal models of PD that transplantation of fetal dopaminergic neurons can relieve PD symptoms, PD patients have received bilateral intrastriatal transplants of fetal mesencephalic cells, with demonstrated dopaminergic cell survival. However, development of 'runaway dyskinesias' – that is, dyskinesias persisting after overnight withdrawal of dopaminergic medication – indicates the need for more sophisticated understanding and modulation of transplanted cell function. Embryonic stem cells are pluripotent cells that can be differentiated into dopaminergic neurons. Transplantation of these cells into the striatum of MPTP-lesioned monkeys and rats after 6-OHDA administration leads to behavioral improvement. No clinical trials have been conducted in PD patients.

Neuroregeneration by infusion of growth factors in a site-specific manner would potentially avoid side effects resulting from activation of other neuronal circuits. Neurotrophic factors, which undergo retrograde axonal transport over intact nigrostriatal axons, have been infused into the striatum of PD patients.

A phase I trial continuously infused glial cell line-derived neurotrophic factor (GDNF) into the putamen of PD patients, which resulted in improvement of certain aspects of motor function with reduction of medication-induced dyskinesias. However, a subsequent randomized controlled study found no detectable clinical benefit from infused intraputamenal GDNF.

Finally, considerable interest has focused on the possibility of using viral vectors to deliver genes to the central nervous system. Lentiviral delivery of GDNF by injecting lenti-GDNF into striatum and substantia nigra of the MPTP primate model of PD reversed functional deficits. Clinical trials of gene therapy in PD patients are under way using adeno-associated virus to deliver glutamic acid decarboxylase to the STN and also aromatic L-amino acid decarboxylase and the trophic factor neurturin to the striatum.

See also: Animal Models of Parkinson's Disease; Cell Replacement Therapy: Parkinson's Disease; Deep Brain Stimulation and Parkinson's Disease; Deep Brain Stimulation and Movement Disorder Treatment; Dopamine – CNS Pathways and Neurophysiology; Dopaminergic Agonists and L-DOPA; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Transgenic Models of Neurodegenerative Disease; Ubiquitin–Proteasome System and Plasticity.

Further Reading

Au WL, Adams JR, Troiano A, et al. (2006) Neuroimaging in Parkinson's disease. *Journal of Neural Transmission Supplement* 70: 241–248.

Beal MF (2005) Mitochondria take center stage in aging and neurodegeneration. *Annals of Neurology* 58: 495–505.

Betchen SA and Kaplitt M (2003) Future and current surgical therapies in Parkinson's disease. *Current Opinion in Neurology* 16: 487–493.

Braak H, Del Tredici K, Rub U, et al. (2003) Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of Aging* 24: 197–211.

Chaudhuri KR, Healy DG, and Shapira AH (2006) Non-motor symptoms of Parkinson's disease: Diagnosis and management. *Lancet Neurology* 5: 235–245.

de Lau LM and Breteler MM (2006) Epidemiology of Parkinson's disease. *Lancet Neurology* 5: 525–535.

Fahn S (2003) Description of Parkinson's disease as a clinical syndrome. *Annals of the New York Academy of Science* 991: 1–14.

Farrer MJ (2006) Genetics of Parkinson disease: Paradigm shifts and future prospects. *Nature Reviews Genetics* 7: 306–318.

Hardy J, Cai H, Cookson MR, et al. (2006) Genetics of Parkinson's disease and parkinsonism. *Annals of Neurology* 60: 389–398.

Lim KL, Dawson VL, and Dawson TM (2003) The cast of molecular characters in Parkinson's disease: Felons, conspirators, and suspects. *Annals of the New York Academy of Sciences* 991: 80–92.

Waters CH (2006) *Diagnosis and Management of Parkinson's Disease*. Caddo, OK: Professional Communications.

Relevant Websites

http://www.MichaelJFox.org – Michael J. Fox Foundation for Parkinson's Research.

http://www.ninds.nih.gov – National Institute of Neurological Disorders and Stroke.

http://www.parkinson.org – National Parkinson Foundation.

http://www.pdf.org – Parkinson's Disease Foundation.

http://www.psp.org – Society for Progressive Supranuclear Palsy.

http://www.nlm.nih.gov – U.S. National Library of Medicine.

http://www.wemove.org – Worldwide Education and Awareness for Movement Disorders.

Parkinson's Disease: Alpha-Synuclein and Neurodegeneration

M Goedert, Medical Research Council Laboratory of Molecular Biology, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Parkinson's disease (PD) is the most common movement disorder. It affects on the order of 1–2% of the general population older than 65. Clinical manifestations include bradykinesia, increased muscle tone, resting tremor, and postural imbalance. Neuropathologially, PD is defined by the loss of dopaminergic nerve cells in the substantia nigra and the presence there of abnormal proteinaceous inclusions known as Lewy bodies. Similar inclusions are also seen in nerve cell processes, where they are called Lewy neurites. The clinical symptoms of PD become manifest when, as a result of dopaminergic nerve cell loss in the substantia nigra, striatal dopamine levels are reduced by more than 70%. In addition, nerve cell loss and Lewy body pathology are found in a number of other brain regions, such as the dorsal motor nucleus of the vagus nerve, the nucleus basalis of Meynert, and some autonomic ganglia. Lewy bodies and Lewy neurites also constitute the defining neuropathological characteristics of Lewy body dementias (encompassing dementia with Lewy bodies and Parkinson's disease with dementia), common late-life dementias that exist in a pure form or overlap with the neuropathological characteristics of Alzheimer's disease (AD). Unlike PD, Lewy body dementias are characterized by the presence of a large number of Lewy bodies and Lewy neurites in cortical brain areas. Ultrastructurally, Lewy bodies and Lewy neurites consist of abnormal filamentous material. Even though Friedrich Lewy described the Lewy body as long ago as 1912, its molecular composition remained unknown until 1997.

A new direction to research on the etiology and pathogenesis of PD and Lewy body dementias was imparted by the twin discoveries that a missense mutation in the *α-synuclein* gene is a rare genetic cause of PD and that α-synuclein is the main component of Lewy bodies and Lewy neurites in PD, dementia with Lewy bodies, and several other diseases. Subsequently, the filamentous glial and neuronal inclusions of multiple system atrophy (MSA), also known as Papp–Lantos bodies, were found to be made of α-synuclein, revealing an unexpected molecular link with Lewy body diseases. These findings have placed dysfunction of α-synuclein at the center of PD and several atypical parkinsonian disorders (Table 1).

The Synuclein Family

Synucleins are abundant brain proteins whose physiological functions are only poorly understood. The human synuclein family consists of three members – α-synuclein, β-synuclein, and γ-synuclein – which range from 127 to 140 amino acids in length and are 55–62% identical in sequence, with a similar domain organization. The N-terminal half of each protein is taken up by imperfect 11-amino acid repeats that bear the consensus sequence KTKEGV. Individual repeats are separated by an interrepeat region of five to eight amino acids. The repeats are followed by a hydrophobic intermediate region and a negatively charged C-terminal domain. By immunohistochemistry, α- and β-synuclein are concentrated in nerve terminals, with little staining of somata and dendrites. Ultrastructurally, they are found in proximity to synaptic vesicles. In contrast, γ-synuclein is present throughout nerve cells in many brain regions. In rat, α-synuclein is most abundant throughout telencephalon and diencephalon, with lower levels in more caudal regions. Beta-synuclein is distributed fairly evenly throughout the central nervous system, whereas γ-synuclein is most abundant in midbrain, pons, and spinal cord, with much lower levels in forebrain areas.

Synucleins are natively unfolded proteins with little secondary structure and have been identified only in vertebrates. Mammals and birds have three synuclein genes; a fourth gene (called γ2-synuclein) is present in teleost fish. Experimental studies have shown that α-synuclein can bind to lipid membranes through its N-terminal repeats, indicating that it may be a lipid-binding protein. It adopts structures rich in α-helical character on binding to synthetic lipid membranes containing acidic phospholipids. In presence of the lipid mimetic sodium dodecyl sulphate, this conformation is taken up by amino acids 1–98 and consists of two α-helical regions (residues 1–42 and 45–98) that are interrupted by a break of two amino acids (residues 43 and 44). Residues 99–140 are unstructured. In cell lines and primary neurons treated with high fatty acid concentrations, α-synuclein was found to accumulate on phospholipid monolayers surrounding triglyceride-rich droplets. Beta-synuclein binds in a similar way, but γ-synuclein fails to bind to lipid droplets and remains cytosolic. Accordingly,

Table 1 α-Synucleinopathies

Idiopathic Parkinson's disease
Dementia with Lewy bodies
Pure autonomic failure
Lewy body dysphagia
Inherited Lewy body diseases
Incidental Lewy body disease
Rapid eye movement sleep behavioral disorder
Multiple system atrophy

α-synuclein binds to fatty acids *in vitro*, albeit with a much lower affinity than physiological fatty acid-binding proteins. Both α- and β-synuclein have been shown to inhibit phospholipase D2. This isoform of phospholipase D localizes to the plasma membrane, where it may be involved in signal-induced cytoskeletal regulation and endocytosis. A fraction of α-synuclein in normal brain is phosphorylated at S129. Casein kinase-1 and casein kinase-2 phosphorylate S129 of α-synuclein *in vitro*, as do several G-protein-coupled receptor kinases.

Inactivation of the *α-synuclein* gene by homologous recombination does not lead to a neurological phenotype, with the mice being largely normal. A loss of function of α-synuclein is therefore unlikely to account for its role in neurodegeneration. Analysis of mice lacking β- or γ-synuclein has similarly failed to reveal any gross abnormalities. The same is true of mice double knockout for α- and β-synuclein. The most compelling evidence for a physiological function comes from experiments showing that increased expression of α-synuclein rescues mice lacking cysteine string protein α (CSPα), a presynaptic molecular chaperone that aids in the folding and refolding of synaptic proteins critical for neurotransmitter release and vesicle recycling. CSPα-deficient mice are normal at birth but develop progressive neurodegeneration that causes defects in synaptic transmission after 2–3 weeks of age and is lethal after 1–4 months. Deletion of CSPα inhibits soluble NSF (N-ethylmaleimide-sensitive factor) attachment protein receptor (SNARE) protein assembly, and transgenic α-synuclein ameliorates this inhibition in a phospholipid binding-dependent manner. Alpha-synuclein may therefore play a physiological role in maintaining synaptic integrity.

Parkinson's Disease and Other Lewy Body Diseases

Alpha-Synuclein Mutations Cause Familial PD

In 1990, an autosomal dominantly inherited form of PD was described in an Italian-American kindred (the Contursi kindred). It was the first familial form of disease in which Lewy bodies were shown to be present. In 1997, a missense mutation (A53T) in the *α-synuclein* gene was identified as the cause of disease in the Contursi kindred (Figure 1). Subsequently, a second mutation (A30P) was identified in a German family with PD. More recently, a third mutation (E46K) was described in a Spanish family with PD and Lewy body dementia. All three missense mutations are located in the repeat region of α-synuclein.

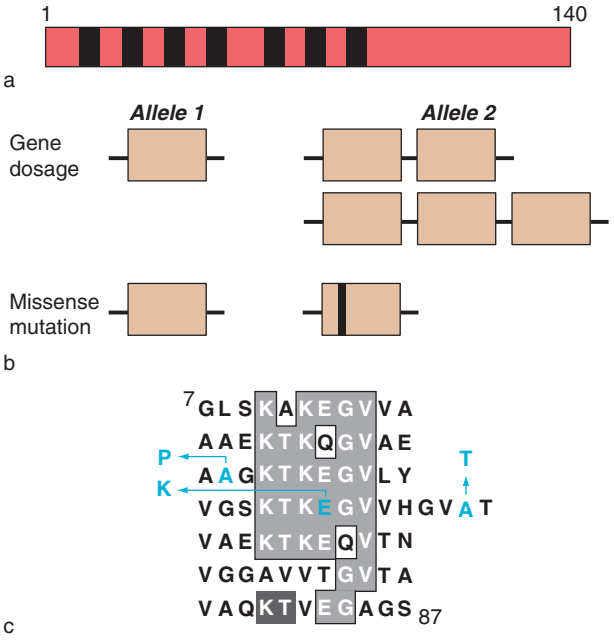

Figure 1 α-Synuclein. (a) Schematic representation of the 140-amino-acid protein, with the N-terminal, lipid-binding repeats shown as black bars. (b) Duplications and triplications of the *α-synuclein* gene and missense mutations (black box) in the *α-synuclein* gene cause inherited forms of Parkinson's disease/Lewy body dementia. (c) Three missense mutations in *α-synuclein* are shown (in blue).

In addition to changes in primary structure, overexpression of wild-type α-synuclein has also been identified as a cause of inherited PD and Lewy body dementia. Thus, duplications and triplications of the *α-synuclein* gene are responsible for autosomal dominantly inherited forms of disease with abundant and widespread Lewy body pathology. Triplications lead to early onset, rapidly progressive parkinsonism dementia with autonomic dysfunction. Duplication mutations result in disease that is more closely related to PD, with later onset parkinsonism, slower progression, and without autonomic dysfunction. In addition, polymorphic variations in the 5′ noncoding region of the *α-synuclein* gene have been found to be associated with idiopathic PD in some studies. They probably influence the level of α-synuclein expression. Variation in the expression levels of α-synuclein may be an important risk factor for the development of Lewy body diseases.

Lewy Body Filaments Are Made of α-Synuclein

Shortly after the identification of the genetic defect responsible for PD in the Contursi kindred, Lewy bodies and Lewy neurites in the substantia nigra from patients with sporadic PD were shown to be strongly immunoreactive for α-synuclein (Figure 2). Subsequently, Lewy body filaments were found to be decorated by antibodies directed against α-synuclein. The same was true of the Lewy body pathology of dementia with Lewy bodies.

Filaments associated with PD and dementia with Lewy bodies are unbranched, with a length of 200–600 nm and a width of 5–10 nm. Full-length α-synuclein is present, with its N- and C-termini being exposed on the filament surface (Figure 3). The core of the filament extends over a stretch of about 70 amino acids that overlaps almost entirely with the lipid-binding region of α-synuclein. Of the three human synucleins, only α-synuclein is associated with the filamentous inclusions of Lewy body diseases. Prior to this work, ubiquitin staining was the most sensitive marker of Lewy body pathology. By the use of double-labeling, the number of α-synuclein-positive structures was found to be greater than that stained for ubiquitin, suggesting that ubiquitination of α-synuclein occurs after assembly. Phosphorylation and nitration are two additional posttranslational modifications of filamentous α-synuclein. Phosphorylation at S129 has been documented in Lewy bodies and Lewy neurites, and it has been suggested that it may trigger filament assembly. However, it remains to be determined whether phosphorylation at S129 occurs before or after filament assembly in the human brain. Nitration of filamentous α-synuclein was found using antibodies specific for nitrated tyrosine residues. As for ubiquitin, staining for nitrotyrosine was less extensive than staining for α-synuclein, suggesting that nitration of α-synuclein occurs after its assembly into filaments.

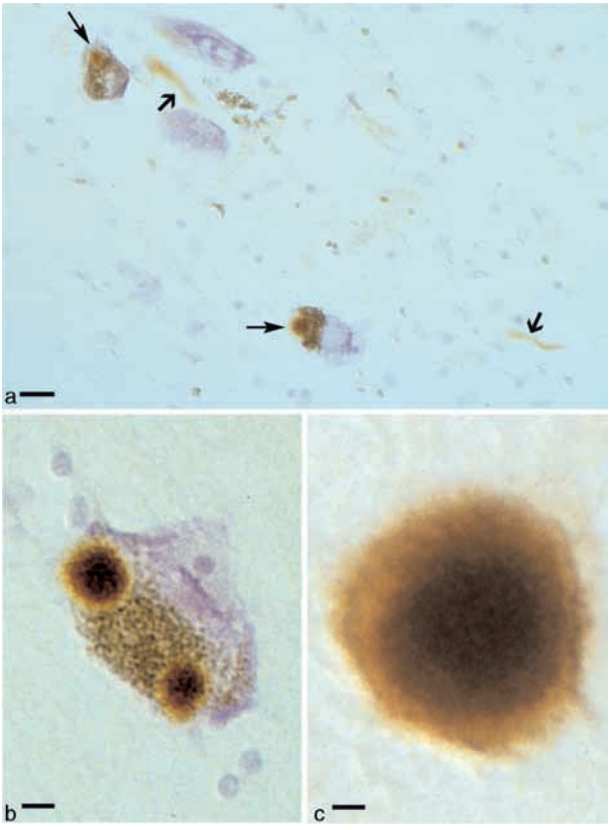

Figure 2 Substantia nigra from patients with Parkinson's disease immunostained for α-synuclein. (a) Two pigmented nerve cells (large arrows), each containing an α-synuclein-positive Lewy body. Lewy neurites (small arrows) are also immunopositive. (b) Pigmented nerve cell with two α-synuclein-positive Lewy bodies. (c) α-Synuclein-positive extracellular Lewy body. Scale bar = 20 μm (a), 8 μm (b), 4 μm (c).

Lewy body pathology is also the defining feature of several other, rarer diseases, such as Lewy body dysphagia and pure autonomic failure. In these diseases, Lewy bodies and neurites are largely limited to the enteric and peripheral nervous systems. In PD, Lewy body pathology is also present in the enteric and autonomic nervous systems.

Incidental Lewy body disease describes the presence of small numbers of Lewy bodies and Lewy neurites at autopsy. It is observed in 5–10% of the general population older than 60 and is believed to represent a preclinical form of Lewy body disease. In cases with incidental Lewy body disease, the first α-synuclein-positive structures in the brain form in the dorsal motor nucleus of the vagus nerve, the intermediate reticular zone, the olfactory bulb, and the anterior olfactory nucleus. Rapid eye movement sleep behavioral disorder and olfactory dysfunction can be the clinical correlates of these abnormalities. The pathology ascends from vulnerable regions in the medulla oblongata to the pontine tegmentum, midbrain, basal forebrain, and cerebral cortex. Alpha-synuclein deposits may form even earlier in the enteric nervous system, the peripheral nervous system, and layer I of the dorsal horn of the spinal cord, suggesting that Lewy body diseases may begin outside the brain. Autonomic dysfunction and pain can be the clinical correlates of these abnormalities. The first deposits develop in the form of Lewy neurites, suggesting that the filamentous assembly of α-synuclein in axons may precede assembly in cell bodies and dendrites. Incidental Lewy body disease may be at one end of the spectrum of Lewy body diseases, with Lewy body dementias at the other end, and with Lewy body dysphagia, pure autonomic failure, and PD in between. These findings underscore the fact that there is more to PD than the mere degeneration of dopaminergic cells in the substantia nigra.

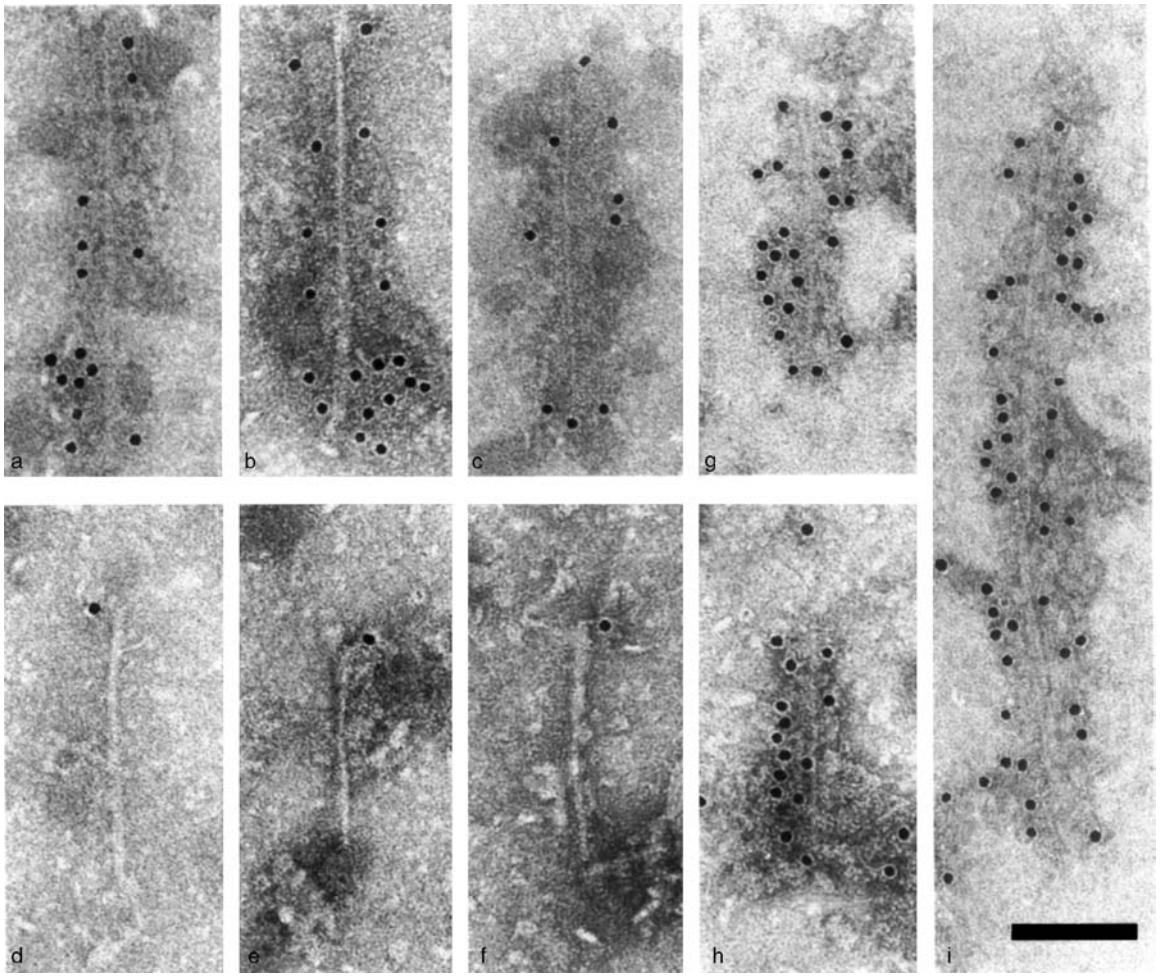

Figure 3 Decoration of filaments from substantia nigra of Parkinson's disease brain with α-synuclein antibodies. (a–c) Antibody recognizing the C-terminus of α-synuclein. (d–f) Antibody recognizing the N-terminal region of α-synuclein. (g–i) Antibody recognizing the N-terminus of α-synuclein. The gold particles indicating positive staining appear as black dots. Scale bar = 100 nm.

Besides amyloid-β and tau deposits, Lewy body pathology is also present in some cases of familial AD, indicating that mutations in the amyloid precursor protein and the presenilin genes can lead to dysfunction of α-synuclein. Similarly, Lewy bodies and Lewy neurites are present in some cases of sporadic AD. Conversely, plaques and tangles are often found in Lewy body dementias. The amygdala is most consistently affected, suggesting that α-synuclein pathology may start there. These findings in AD define a type of α-synucleinopathy that is distinct from the one described above.

MSA

Filamentous α-synuclein deposits characterize this atypical parkinsonian disorder, but Lewy body pathology is not generally observed. Instead, glial cytoplasmic inclusions (Papp–Lantos bodies), which consist of filamentous aggregates, are the defining neuropathological feature of MSA. They are found mostly in the cytoplasm and, to a lesser extent, the nucleus of oligodendrocytes. Inclusions are also observed in the cytoplasm and nucleus of some nerve cells and in neuropil threads. The principal brain regions affected are the substantia nigra, striatum, locus coeruleus, pontine nuclei, inferior olives, cerebellum, and spinal cord. Typically, nerve cell loss and gliosis are observed. The formation of glial cytoplasmic inclusions may be the primary lesion that will eventually compromise nerve cell function and viability.

Glial cytoplasmic inclusions are strongly immunoreactive for α-synuclein, and filaments isolated from the brains of patients with MSA are labeled by α-synuclein antibodies. Assembled α-synuclein is nitrated and phosphorylated at S129, and the number of α-synuclein-positive structures exceeds that stained by antiubiquitin antibodies, confirming that the

accumulation of α-synuclein precedes ubiquitination. Filament morphologies and their staining characteristics were found to be similar but not identical to those of filaments extracted from the brains of patients with PD and dementia with Lewy bodies. This work revealed an unexpected molecular link between MSA and Lewy body diseases. The main difference is that in MSA, most of the α-synuclein pathology is found in glial cells, whereas in Lewy body diseases, most of the pathology is in nerve cells.

Synthetic α-Synuclein Filaments

Recombinant α-synuclein assembles into filaments that share many of the morphological characteristics of the disease filaments. Assembly is a nucleation-dependent process that occurs through the N-terminal repeats of α-synuclein. The C-terminal region, in contrast, is inhibitory. Assembly is accompanied by the transition from random coil to a β-pleated sheet. By electron diffraction and solid-state nuclear magnetic resonance spectroscopy, filaments assembled from recombinant wild-type human α-synuclein show a conformation characteristic of amyloid fibers, with evidence for the existence of filaments with different morphologies. Assembly of monomeric α-synuclein into filaments probably passes through an oligomeric stage that after prolonged incubation results in the formation of elongated protofibrils that disappear on filament formation. Oligomeric α-synuclein may be toxic to nerve cells. Under the conditions of these experiments, β- and γ-synucleins failed to assemble, consistent with their absence from the filamentous lesions of the human diseases. Both β- and γ-synucleins have been shown to inhibit the assembly of α-synuclein *in vitro*. Inhibition of α-synuclein assembly by a number of small organic molecules has also been described.

Mutations E46K and A53T increase the rate of assembly of α-synuclein, indicating that this may be their primary effect. The mechanism of action of the A30P mutation is less clear. One study reported that it causes the accumulation of oligomeric, nonfibrillar α-synuclein, implying that this may be a toxic species. Mutation A30P reduces the binding of α-synuclein to rat brain vesicles, whereas mutation E46K increases lipid binding. Mutation A53T has no significant effect on lipid binding. Many factors have been shown to influence the rate and/or extent of α-synuclein assembly *in vitro*. They include nitration, oxidation, dopamine, cytochrome C, polyamines, some pesticides, several di- and trivalent cations, and sulphated glycosaminoglycans. However, the relevance, if any, of these factors for the transition from soluble to filamentous α-synuclein in nerve cells in the brain remains to be established.

Animal Models of Human α-Synucleinopathies

Rodents and Primates

Mice expressing wild-type and mutant human α-synuclein in nerve cells or glial cells have been shown to develop numerous α-synuclein-immunoreactive cell bodies and processes. However, filament formation and nerve cell loss were a less consistent feature. One study has documented the presence of α-synuclein filaments in brain and spinal cord of mice transgenic for A53T α-synuclein. The formation of inclusions correlated with the appearance of a severe movement disorder, suggesting a possible cause and effect relationship. A minority of inclusions was ubiquitin immunoreactive. Signs of Wallerian degeneration were in evidence in ventral roots, but motor neuron numbers were not reduced. A major difference with PD was the absence of significant pathology in dopaminergic cells of the substantia nigra. However, a more recent study of transgenic mice expressing a C-terminally truncated form of α-synuclein that fibrillizes more readily than the full-length protein *in vitro* has shown some evidence of filament formation and neurodegeneration in substantia nigra. Expression of high levels of human α-synuclein in oligodendroglia recapitulates features of MSA, including the accumulation of filamentous human α-synuclein aggregates in oligodendrocytes linked to their degeneration. In addition, mouse α-synuclein also accumulated in axons in association with neuronal cell loss and motor impairments.

Viral vector-mediated gene transfer differs from standard transgenic approaches by being targeted to a defined region of the central nervous system and by being inducible at any point during the life of the animal. Adeno-associated and lentiviral vectors have been used to express human wild-type and mutant α-synuclein in rodent and primate substantia nigra. Lewy body-like inclusions formed, and a significant proportion of nerve cells degenerated. In the rat, about a quarter of animals were impaired in spontaneous and drug-induced behavior. In the marmoset monkey, expression of human A53T α-synuclein in the substantia nigra resulted in the formation of inclusions and nerve cell death in most (but not all) animals. Fewer inclusions and less nerve cell death were observed on expression of wild-type human α-synuclein.

A neurotoxin model of α-synuclein pathology has been developed in the rat through the chronic administration of the pesticide rotenone, a high-affinity inhibitor of complex I. The rats developed a progressive degeneration of nigrostriatal neurons and Lewy body-like inclusions that were immunoreactive for α-synuclein and ubiquitin. Behaviorally, they showed bradykinesia, postural instability, and resting

tremor. Under these conditions, the inhibition of complex I was only partial, indicating that a bioenergetic defect with adenosine triphosphate (ATP) depletion was probably not involved. Instead, oxidative damage might have played a role since partial inhibition of complex I by rotenone stimulates the production of reactive oxygen species. It would therefore seem that oxidative stress can lead to the assembly of α-synuclein into filaments.

Flies, Worms, and Yeast

One of the first reports to describe the overexpression of human α-synuclein made use of *Drosophila melanogaster*, an organism without synucleins. Expression in nerve cells resulted in the formation of filamentous Lewy body-like inclusions and an age-dependent loss of some dopaminergic nerve cells. The inclusions were α-synuclein- and ubiquitin-positive. The ensuing motor defect was reversed by L-dopa and dopamine agonists. Toxicity of α-synuclein in this system depended on aggregation and required phosphorylation of S129. Chaperones modulated these effects. Thus, co-expression of heat shock proteins reduced the toxicity of α-synuclein, whereas a reduction in endogenous chaperones exacerbated nerve cell loss. Overexpression of wild-type and mutant human α-synuclein in *Caenorhabditis elegans* resulted in dopaminergic nerve cell loss and motor deficits, but in the apparent absence of α-synuclein fibrils.

When expressed in the yeast *Saccharomyces cerevisiae*, human wild-type and A53T α-synucleins were found to localize at the plasma membrane, whereas the A30P mutant had a cytosolic distribution, consistent with lipid binding *in vitro*. Unlike the A30P mutant, wild-type and A53T α-synucleins were toxic to yeast cells. A genome-wide screen to identify genetic enhancers of α-synuclein toxicity identified genes involved in lipid metabolism and vesicle-mediated transport as the predominant functional class. It has been suggested that toxicity of overexpressed α-synuclein in yeast may not require filament formation but may result from the coating of the plasma membrane and internal membranes via amphipathic α-helix formation, resulting in the disruption of normal membrane processes, including vesicle trafficking. Expression of α-synuclein has been shown to block selectively endoplasmic reticulum to Golgi transport, causing endoplasmic reticulum stress. The ensuing toxicity could be suppressed by overexpression of the small GTPase Ypt1/Rab1. It is interesting that overexpression of the latter also rescued nerve cell loss in invertebrate models of α-synuclein-induced neurodegeneration.

Outlook

The work of the past 10 years has shown conclusively that a neurodegenerative pathway leading from soluble to insoluble, filamentous α-synuclein is central to Lewy body diseases and MSA. The development of experimental models of α-synucleinopathies has opened the way to the identification of the detailed mechanisms by which the formation of inclusions can cause disease. These model systems have also made it possible to identify disease modifiers that may lead to the development of mechanism-based therapies for these diseases. At a conceptual level, it will be important to understand whether α-synuclein has a role to play in inherited parkinsonian disorders caused by mutations in other genes. They include autosomal-recessive juvenile forms of parkinsonism caused by mutations in the *Parkin*, *DJ-1*, and *PINK-1* genes and the dominantly inherited diseases caused by mutations in the lysosomal type 5 adenosine triphosphatase gene (*ATP13A2*) and the leucine-rich repeat kinase-2 gene (*LRRK2*). The latter is particularly interesting since activating missense mutations in the kinase domain of LRRK2 are a relatively common cause of Lewy body PD, demonstrating the existence of a pathway leading from activation of LRRK2 to aggregation of α-synuclein. The molecular dissection of this pathway is likely to shed new light on the understanding of Lewy body diseases.

See also: Animal Models of Parkinson's Disease; Basal Ganglia: Functional Models of Normal and Disease States; Basal Ganglia: Evolution; Cell Replacement Therapy: Parkinson's Disease; Deep Brain Stimulation and Parkinson's Disease; Oxidative Damage in Neurodegeneration and Injury; Parkinsonian Syndromes; Synucleins.

Further Reading

Baba M, Nakajo S, Tu PS, et al. (1998) Aggregation of α-synuclein in Lewy bodies of sporadic Parkinson's disease and dementia with Lewy bodies. *American Journal of Pathology* 152: 879–884.

Betarbet R, Sherer TB, MacKenzie G, et al. (2000) Chronic systemic pesticide exposure reproduces features of Parkinson's disease. *Nature Neuroscience* 3: 1301–1306.

Braak H, Del Tredici K, Rüb U, et al. (2003) Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of Aging* 24: 197–211.

Chandra S, Gallardo G, Fernández-Chacón R, Schlüter OM, and Südhof TC (2005) α-Synuclein cooperates with CSPα in preventing neurodegeneration. *Cell* 123: 383–396.

Chartier-Harlin MC, Kachergus J, Roumier C, et al. (2004) Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* 364: 1167–1169.

Feany MB and Bender WW (2000) A *Drosophila* model of Parkinson's disease. *Nature* 404: 394–398.

Fujiwara H, Hasegawa M, Dohmae N, et al. (2002) α-Synuclein is phosphorylated in synucleinopathy lesions. *Nature Cell Biology* 4: 160–164.

Goedert M (2001) Alpha-synuclein and neurodegenerative diseases. *Nature Reviews Neuroscience* 2: 492–501.

Ibanez P, Bonnet AM, Debarges B, et al. (2004) Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. *Lancet* 364: 1169–1171.

Krüger R, Kuhn W, Müller T, et al. (1998) Ala30Pro mutation in the gene encoding α-synuclein in Parkinson's disease. *Nature Genetics* 2: 106–108.

Lee VMY and Trojanowski JQ (2006) Mechanisms of Parkinson's disease linked to pathological α-synuclein: New targets for drug discovery. *Neuron* 52: 33–38.

Lewy F (1912) Paralysis agitans. In: Lewandowski M and Abelsdorff G (eds.) *Handbuch der Neurologie, vol. 3*, pp. 920–933. Berlin: Springer.

Paisán-Ruiz C, Jain S, Whitney Evans E, et al. (2004) Cloning of the gene containing mutations that cause *PARK8*-linked Parkinson's disease. *Neuron* 44: 595–600.

Papp MI, Kahn JE, and Lantos PL (1989) Glial cytoplasmic inclusions in the CNS of patients with multiple system atrophy. *Journal of Neurological Sciences* 94: 79–100.

Polymeropoulos MH, Lavedan C, Leroy E, et al. (1997) Mutation in the α-synuclein gene identified in families with Parkinson's disease. *Science* 276: 2045–2047.

Singleton AB, Farrer M, Johnson J, et al. (2003) α-Synuclein locus triplication causes Parkinson's disease. *Science* 302: 841.

Spillantini MG, Crowther RA, Jakes R, Hasegawa M, and Goedert M (1998) α-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. *Proceedings of the National Academy of Sciences of the United States of America* 95: 6469–6473.

Spillantini MG, Schmidt ML, Lee VMY, et al. (1997) α-Synuclein in Lewy bodies. *Nature* 388: 839–840.

Zarranz JJ, Alegre J, Gomez-Esteban JC, et al. (2004) The new mutation, E46K, of α-synuclein causes Parkinson and Lewy body dementia. *Annals of Neurology* 55: 164–173.

Zimprich A, Biskup S, Leitner P, et al. (2004) Mutations in *LRRK2* cause autosomal-dominant parkinsonism with pleomorphic morphology. *Neuron* 44: 601–607.

Paroxysmal Paralysis

M Salajegheh and A A Amato, Brigham and Women's Hospital, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Periodic paralyses are a group of muscle disorders characterized by episodic attacks of muscle weakness, secondary to abnormal sarcolemmal excitability, which may progress into paralysis. They are often induced by certain triggers, most often (but not always) are associated with changes in serum potassium level and are occasionally accompanied by myotonia (the inability to relax muscle after contraction). The familial cases are usually inherited in an autosomal dominant pattern, but can also occur sporadically. Classically, based on serum potassium levels or response to potassium administration, they have been classified as hyper-, hypo- or normo-kalemic periodic paralysis. Since their pathogenesis is related to disorders of muscle membrane sodium, calcium, and potassium ion channels (channelopathies), we prefer to classify them based on the various channels involved (Table 1). All forms of familial periodic paralyses share the common final pathway of aberrant depolarization, which results in the inactivation of sodium channels and electrical inexcitability of muscle fibre. Differences in attack frequency, severity, precipitating factors, and electrodiagnostic findings suggest the channel gene that is affected, and the specific alterations in mutant channel function. The episodic nature of these channelopathies suggests that while channel mutations are necessary to cause the disease, they are not sufficient. Understanding potential triggers that could induce the attacks are imperative to the management of the various channelopathies. Secondary cases also exist and are related to alterations in serum potassium levels and thyrotoxicosis, which need to be excluded in patients with periodic paralysis and no family history.

Sodium Channelopathies

Sodium channelopathies include potassium-sensitive (hyperkalemic) periodic paralysis (HyperPP), paramyotonia congenita (PMC), potassium aggravated myotonias (PAMs) (e.g., myotonia fluctuans, myotonia permanens, and acetazolamide responsive myotonia), and familial hypokalemic periodic paralysis type 2 (HypoPP2) which are autosomal dominant myopathies with certain clinical similarities as well as differences. These disorders have been demonstrated to be allelic, and caused by missense mutations in the pore-forming α-subunit of the voltage-gated skeletal-muscle sodium channel NaV1.4 (encoded by the gene *SCN4A*). In general, each missense mutation of *SCN4A* is consistently associated with one of the four allelic sodium channel disorders, suggesting the presence of separate classes of functional defects. However, some variability exists and the distinction is often blurred between PMC and HyperPP, even for affected members of the same family.

The skeletal-muscle voltage-gated sodium channel is a heterodimer consisting of α- and β-subunits that are encoded on chromosomes 17q23-25 and 19q13.1, respectively. Point mutations in the α-subunit gene, *SCN4A*, are responsible for the mentioned disorders. The α-subunit of the sodium channel is composed of four homologous domains (1–4) each containing six hydrophobic segments (S1–S6) that traverse the plasma membrane. Studies have suggested that the S4 segment is critical for inactivation of the open channel, while the S3 segment plays a role in the recovery of inactivated channels. Thus, the S3 and S4 segments interact during the transition to and from inactivation states. Over 30 different point mutations in *SCN4A* have been described and most are located in regions of the α-subunit critical for fast inactivation. Almost all tested missense mutations produce gain-of-function defects, either disrupted inactivation or enhanced activation. The notable exception is the mutations associated with HypoPP2, which are all clustered in the voltage sensor region of the second repeat (D2S4) and diminish activity by enhancing inactivation The β-subunit is noncovalently linked to the α-subunit and functions to increase the amplitude of the current and speed of inactivation.

Normally, upon depolarization to the threshold potential, the voltage-gated sodium channels open (activate) quickly and then close to a fast-inactivated state within milliseconds, from which further openings are rare (Figure 1). This is accompanied by an inward sodium current that peaks within milliseconds, followed by a drop back to zero as the channel inactivates. In order for the sodium channels to recover from the inactivated state, and become available for opening up with subsequent depolarizations, the muscle membrane must become hyperpolarized by the influx of potassium. Thus, fast inactivation of channels limits the duration of the action potential, results in a refractory period that limits the firing rate of muscle fibers, and initiates repolarization of the muscle membrane. In doing so, it regulates the number of sodium channels available for activation. Gain-of-function mutations in PMC

Table 1 Hereditary channelopathies with clinical features of periodic paralysis

Channelopathy type	*Inheritance*	*Gene/ location*	*Onset*	*Triggers*	*Duration*
Sodium channel					
Potassium-sensitive periodic paralysis (HyperPP)	Autosomal dominant	*SCN4A*/ 17q23-25	First decade	Rest following exercise, fasting, potassium, stress, fatigue, cold[b]	Hours
Paramyotonia congenita[a] (PMC)	Autosomal dominant	*SCN4A*/ 17q23-25	First decade	Exercise, cold, potassium	Hours
Hypokalemic periodic paralysis type 2 (HypoPP2)	Autosomal dominant	*SCN4A*/ 17q23-25	First two decades	Rest following exercise, carbohydrates, stress	Hours to days
Calcium channel					
Hypokalemic periodic paralysis type 1 (HypoPP1)	Autosomal dominant	*CACNA1S*/ 1q31-32	First two decades	Rest following exercise, carbohydrates, alcohol, cold, stress and lack of sleep, viral illness, menstruation and specific medications[c]	Hours to days
Potassium channel[d]					
Andersen–Tawil syndrome (ATS)	Autosomal dominant	*KCNJ2*/ 17q23-25	First two decades	Prolonged rest following exercise	Hours to days

[a]As opposed to myotonia, muscle stiffness is induced or worsened by exercise in paramyotonia.
[b]Cooling more commonly provokes weakness in patients with paramyotonia.
[c]β-Agonists, corticosteroids, and insulin.
[d]In some reports, KCNE3 gene (11q13-q14) mutation has been suggested in HypoPP and thyrotoxic periodic paralysis, although not fully verified.

Amato A and Dumitru D (2002) Hereditary myopathies. In: Dumitru D, Amato A, and Zwarts M (eds.) *Electrodiagnostic Medicine*, 2nd edn., pp. 1265–1370. Philadelphia: Hanley & Belfus, Inc.
Miller TM, Dias da Silva MR, Miller HA, et al. (2004) Correlating phenotype and genotype in the periodic paralyses. *Neurology* 63(9): 1647–1655.

and PAM often slow the rate of inactivation by three- to fivefold, resulting in increased duration of the action potential, and the availability of Na^+ channels (i.e., fraction not inactivated) at the end of each action potential, in turn augmenting the membrane excitability and causing myotonia. In addition to slow inactivation, many PMC mutations also disrupt the final extent of inactivation. The steady inward Na current generated through the small fraction of mutant channels that have failed to inactivate depolarizes the membrane to a new stable resting potential of approximately -50 mV. This in turn results in inactivation of the wild-type and most of the mutant Na channels, resulting in a system that is refractory from generating an action potential, leading to a paralytic attack. In these patients, increased extracellular potassium results in further depolarization of the muscle membrane, thus leading to muscle fiber inexcitability.

Primary Hyperkalemic or Potassium-Sensitive Periodic Paralysis (Adynamia Episodica Hereditaria)

We generally prefer to use the term potassium-sensitive periodic paralysis, since while patients are sensitive to and become weak when given potassium, not all attacks are associated with hyperkalemia.

Clinical features Potassium-sensitive periodic paralysis is transmitted in an autosomal dominant fashion, with a high degree of penetrance in both females and males. In the majority of patients, symptoms manifest within the first decade of life. While attacks of weakness usually develop in the morning, they can occur at any time, and are precipitated by rest following exercise or fasting. The attacks can be mild or severe, with the latter more commonly occurring after strenuous physical activity. These may also be triggered by intake of potassium-rich food, stress, and fatigue. Paresthesia and a sense of heaviness in the legs or small of the back may herald the onset of an attack in some individuals. The weakness then manifests in the thighs and calves and progresses to involve multiple other muscle groups. The weakness can also be focal, and unlike hypokalemic periodic paralysis, it rarely develops into generalized flaccid paralysis. While as with all periodic paralyses the sphincter muscles are well preserved during attacks, the degree of weakness may be severe enough to affect bulbar and respiratory muscles in some patients. During attacks, the reflexes are diminished or absent, while sensation remains normal. Sustained mild exercise after a period of strenuous activity may postpone or prevent weakness from developing in the exercising muscles while resting muscle groups become weak.

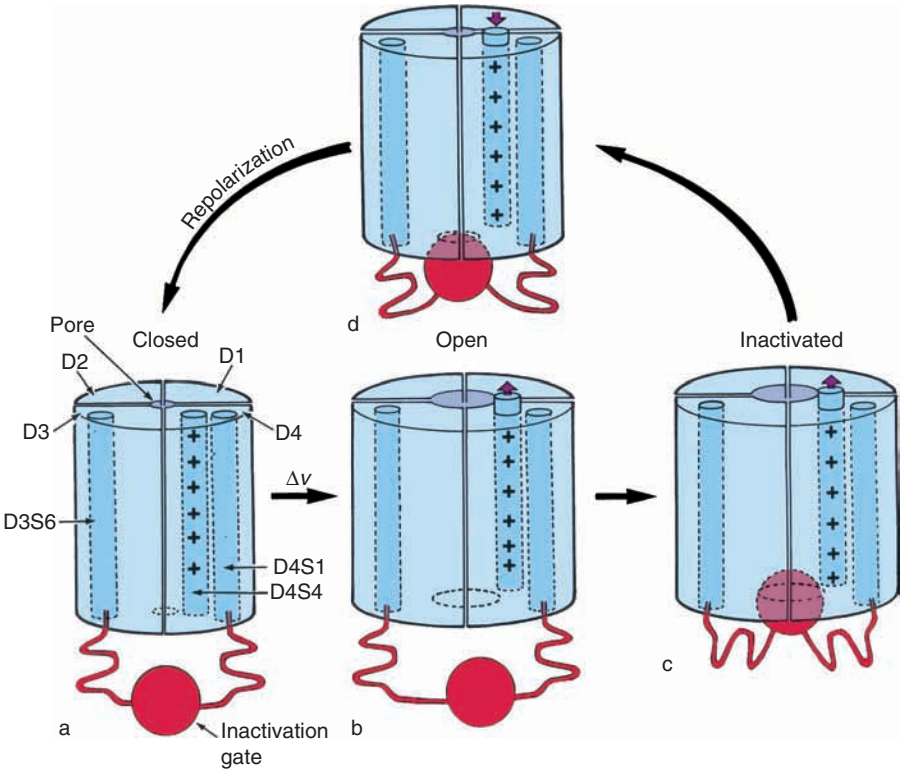

Figure 1 Cation channel model showing the four domains (D1–D4) of the cation channel arranged around the ion pore in the membrane. The S4 segments of cation channels contain a repeating motif of positively charged (+) amino acids (arginine or lysine) at every third position separated by two neutral amino acids. In the closed state (a), the pore is closed and the inactivation gate is open. In response to depolarization, the S4 segments move slightly under the influence of electrostatic forces. Depolarization leads to a conformational shift, which results in pore opening (b). Channel closing is thought to result from a 'ball-valve' mechanism where a cytoplasmic loop between domains D3 and D4 falls into the ion pore, thus blocking it (c). In this model, repolarization would then result in the protein assuming its closed conformation (d) and release of the inactivation gate. The channel is now in the closed state (a) and is ready to open in response to the next depolarization. Reproduced from Ptacek L (1998) The familial periodic paralyses and nondystrophic myotonias. *American Journal of Medicine* 105(1): 58–70, with permission from Elsevier.

Following a bout of weakness, it is not uncommon for pain to be experienced in the affected muscles up to several days. The duration of weakness is usually less than 2 h, although mild weakness can persist for a few days. The frequency of attacks is highly variable, and while it may occur several times a day in some patients others experience difficulty once a year. In addition, there is great variation of attack severity and frequency within and between families, and the frequency of paretic attacks often decreases with age. Between the attacks, sensation and deep tendon reflexes are normal and lid lag or eyelid myotonia may be the only clinical signs present. Over time, some patients develop fixed weakness, unrelated to attack frequency, usually involving the more proximal muscles.

Potassium-sensitive periodic paralysis can occur in three forms: (1) without myotonia, (2) with clinical or electrical myotonia, or (3) associated with PMC. The course of the attacks of weakness is the same in all three forms, except that cooling more commonly provokes weakness in patients with PMC. Clinical myotonia is usually mild, and can be elicited in the face/eyelids, tongue, finger extensors, and the thenar eminence with percussion or activity. Patients with PMC exhibit paradoxical myotonia in which muscle stiffness is induced or worsened by exercise or cold temperature.

While provocative testing can be performed when the diagnosis is unclear, the availability of commercial genetic testing usually makes it unnecessary. Provided the patient has no significant underlying cardiac or renal disease, a potassium load can be administered to increase the serum potassium levels, and to see whether an attack of weakness can be provoked. The patient's strength, electrolytes, and cardiac function (electrocardiogram (EKG)) need to be closely monitored during the study.

Laboratory features Although the attacks of weakness are usually associated with serum potassium levels of up to 5 to 6 $mEq\,l^{-1}$, serum levels may also

remain within normal limits. During this time, urinary excretion of potassium increases and serum sodium levels decrease. At the end of an attack, diuresis and transient hypokalemia commonly occur. In between the attacks, serum potassium levels are within normal limits and serum creatine kinase (CK) levels can be normal or mildly elevated. Increased amplitudes of the precordial T waves are occasionally noted on the EKG and are secondary to hyperkalemia (Table 2).

Histopathology Muscle biopsy in patients with potassium-sensitive periodic paralysis is frequently associated with vacuoles in the myofibers, and patients with fixed weakness may have nonspecific myopathic changes (Table 2).

Electrophysiologic findings Both motor and sensory nerve conduction studies are normal between and during attacks, while the compound muscle action potential (CMAP) amplitude is reduced during an attack of weakness. Abnormalities are usually apparent with the short and long exercise tests. The short-exercise test is performed by having the patient isometrically exercise a muscle (e.g., abductor digiti minimi) for 10 s, followed by measurement of CMAP 2 s immediately after exercise and every 10 s thereafter up to 50 s. This test is usually repeated twice more with a rest period of 60 s between trials. The long exercise test is performed by having the patient isometrically exercise a muscle (e.g., abductor digiti minimi) for 5 min (with 3–4 s of rest every 30–45 s), while CMAPs are being recorded every minute during the exercise period, 2 s immediately after cessation of exercise, then every minute for 5 min and finally every 5 min for 40–45 min. In normals, CMAP amplitudes mildly increase during exercise, and slightly decrease immediately after exercise completion along with increased duration, resulting in an increase of the total CMAP area. Recovery of preexercise values occurs after 30–60 s and remains unchanged within the following 40–50 min (Table 2).

After the short exercise test in patients with potassium-sensitive periodic paralysis, most show an increase in CMAP amplitudes that are significantly higher and last longer than that observed in unaffected individuals, and repetition of the short exercise amplifies the increase in CMAP amplitude. With the long exercise test, during the exercise period and immediately afterwards, there is an initial increase in CMAP amplitude. This is followed by a progressive decline in amplitude over the subsequent 40–45 min. A short exercise trial during the paretic phase may induce a marked CMAP increment, partially correcting the loss of excitability.

Needle electromyographic examination findings depend on whether myotonia is present or absent. Electrical myotonia is found in 50–75% of affected individuals, while it is clinically apparent in less than 20%. In patients with myotonia, examination of the muscle between attacks of weakness reveals an increase in insertional activity, in the form of fibrillation potentials and positive sharp waves, along with mild to moderate amounts of spontaneous runs of myotonic potentials. These are reflective of the underlying hyperexcitability or instability of the muscle membrane and are not due to denervation. A reduction in limb temperature may exacerbate the runs of myotonic potentials. Analysis of motor unit action potential (MUAP) parameters may reveal either normal potentials or a slight increase in the percentage of small duration, polyphasic potentials. If an attack of weakness is provoked during the examination, there may be a slight initial increase in spontaneous activity followed by a reduction in all electrical activity, both spontaneous and voluntary; the number of MUAPs detected decreases and the recruitment pattern may be reduced to a few motor units of normal or short duration that are firing rapidly.

In patients with potassium-sensitive periodic paralysis without myotonia, the insertional activity is normal and needle movement or voluntary contraction does not provoke the appearance of fibrillation potentials, positive sharp waves, or myotonic potentials. During an attack of weakness, there is a progressive reduction in the electrical activity, without any transitory increase in membrane excitability. The MUAPs not only decrease in number but also in duration and amplitude. In patients with profound weakness, complete electrical silence may be observed in some muscles with respect to voluntary MUAPs. In some patients, there may be a mild increase in the number of short-duration polyphasic MUAPs suggestive of a mild and persistent myopathy.

Molecular genetics Potassium-sensitive periodic paralysis is caused by mutations in the α-subunit of the voltage-gated sodium channel (*SCN4A*).

Treatment Preventive therapy with a low-potassium, high-carbohydrate diet and avoidance of fasting, strenuous activity, and cold is recommended. The acute attacks of weakness are often mild in severity, short lasting, and do not require treatment. Ingestion of simple carbohydrates (e.g., fruit juices, glucose-containing candies) results in insulin excretion and a decrease in serum potassium levels, which can lead to an improvement in strength. Beta-adrenergic agonists (e.g., metaproterenol, albuterol, and salbutamol)

Table 2 Diagnostic features of some channelopathies presenting with periodic paralysis

Diseases	*Short exercise*	*Long exercise*	*Needle EMG*	*Musce biopsy*	*CK*	*K^+*
Potassium-sensitive periodic paralysis	Absence of PEMPs, CMAP increase postexercise amplified by exercise repetition	CMAP increase immediately postexercise, with gradual and prolonged decrement in rest phase	Myotonia[a] and increased spontaneous activity,[b] normal MUAP or slight increase in small duration, polyphasic potentials in patients with fixed myopathy	Fiber size variation, occasional vacuolated muscle fibers	↑ or normal	↑ or normal during attack
Paramyotonia congenita	Postexercise PEMPs and decrement in CMAP, worsening with exercise repetition[c]	CMAP decrement during and following exercise, occasionally lasting for hours	Myotonia, worsened by cold,[c] increased jitter, and occasional blocking on SFEMG	Fiber size variation, occasional vacuolated muscle fibers, myofibrillar disarray, and tubular aggregates	↑ or normal	↑ or normal during attack
Hypokalemic periodic paralysis type 1	Absence of PEMPs or CMAP changes postexercise	CMAP unchanged immediately postexercise, but delayed decrement in rest phase	Absence of myotonia, normal MUAPs or slight increase in small duration, polyphasic potentials in patients with fixed myopathy	Fiber size variation, vacuolated muscle fibers, sarcoplasmic reticulum dilation, tubular aggregates	↑ or normal	↓ during attack
Andersen–Tawil syndrome		CMAP may show delayed decrement in rest phase	Absence of myotonia	Fiber size variation, tubular aggregates	↑ or normal	↑, normal, or mostly ↓ during attack

[a]Myotonia is seen in a subgroup of patients with potassium-sensitive periodic paralysis.
[b]Fibrillation potentials and positive sharp waves are seen in the subgroup of patients with myotonia, and are reflective of the underlying hyperexcitability or instability of the muscle membrane.
[c]In patients with PMC local cooling results in decrement in CMAPs, dense fibrillation-like potentials on EMG with gradual reduction in MUAP activity.
CK, serum creatine kinase levels; CMAP, compound muscle action potential; EMG, electromyography; K^+, serum potassium levels; SFEMG, single fiber electromyography; PEMP, postexercise myotonic potential.

Amato A and Dumitru D (2002) Hereditary myopathies. In: Dumitru D, Amato A, and Zwarts M (eds.) *Electrodiagnostic Medicine*, 2nd edn., pp. 1265–1370. Philadelphia: Hanley & Belfus, Inc.
Fournier E, Arzel M, Sternberg DL, et al. (2004) Electromyography guides toward subgroups of mutations in muscle channelopathies. *Annals of Neurology* 56(5): 650–661.
Venance SL, Cannon SC, Fialho D, et al. (2006) The primary periodic paralyses: Diagnosis, pathogenesis and treatment. *Brain* 129(Pt. 1): 8–17.

may be effective in improving strength and can be used safely provided there are no associated cardiac arrhythmias. Only rarely, and with severe attacks, is treatment with intravenous glucose, insulin, or calcium carbonate indicated. Prophylactic treatment with acetazolamide (125–1000 mg daily), chlorothiazide (250–1000 mg daily), and dichlorphenamide (50–150 mg daily) may be beneficial in reducing the frequency of attacks and may also help with myotonia. Mexiletine also diminishes the myotonia.

Normokalemic Periodic Paralysis

Some patients with potassium-sensitive periodic paralysis develop weakness while serum potassium levels remain within normal limits. Most patients in this group develop weakness when given potassium, and have clinical and electrophysiological features which are similar to those with weakness and elevated serum potassium levels. Furthermore, mutations in the muscle sodium channel gene have been demonstrated in families with normokalemic periodic paralysis.

PMC (Eulenburg Disease)

The prefix 'para' derives from the paradoxical reaction these patients show toward exercise, where in contrast to the warm-up phenomenon observed in other myotonic syndromes, repeated exercise worsens their muscle stiffness. Cold temperature can also precipitate myotonia and occasionally weakness in individuals with PMC.

Clinical features PMC is an autosomal dominant disorder with high penetrance and is allelic to potassium-sensitive periodic paralysis, with some kinships having clinical features of both (paralysis periodica paramyotonica). The attacks of myotonia are usually induced by exercise or cold, and some patients develop either focal or a more generalized weakness occasionally progressing to flaccid paralysis. The facial, hand, pharyngeal, and tongue muscles are the most sensitive to these phenomena. Patients usually display evidence of the disease within the first decade of life. Infants may be noted to have the disease during a crying spell, when they are unable to open their eyes secondary to exercise-induced myotonia of the orbicularis oculi muscles. Alternatively, it may be noticed that washing the infant's face with a cold washcloth may result in a cold-induced myotonic spasm of the facial muscles. Patients who live in cold climates may note an increase in exercise-induced muscle contraction leading to weakness during the shoveling of snow, and there may be an occasional patient who complains of hand weakness when shopping in the frozen food section of the supermarket. A cold-induced attack can last for several hours even after return to a warm environment. Weakness can also be induced in some cases by potassium intake. Paramyotonia, particularly of the eyelids, is typically evident in most patients. While percussion myotonia may be demonstrated, it is usually not prominent, and myalgias are not a common complaint. Muscle strength is normal in between attacks of paralysis, in those with pure PMC, as are the deep tendon reflexes and sensory examination. These symptoms are usually nonprogressive and the patients either remain stable or improve slightly with aging.

Laboratory features Serum CK levels may be mildly to moderately elevated in some patients. Serum potassium levels may be elevated during an attack of paralysis (Table 2).

Histopathology Minimal changes are noted on muscle biopsy, although there might be mild fiber size variation with a mixture of normal, atrophic, and hypertrophic fibers. While fiber type differentiation is usually normal, few reports have suggested abnormalities of differentiation. A few patients may display intracytoplasmic vacuoles, particularly those with superimposed periodic paralysis. Electron microscopy may reveal myofibrillar disarray and tubular aggregates (Table 2).

Electrophysiologic findings Routine sensory and motor nerve conduction studies are normal between attacks, while abnormalities are usually apparent during the short-exercise test. There is repetitive activity immediately after exercise, where the first component of the CMAP response, evoked by a single supramaximal stimulus, is followed by several signals of decreasing amplitude that occur every 7–10 ms (post-exercise myotonic potentials or PEMPs). There is also evidence for a decrement in the CMAP in some patients. Repetition of the short exercise, even in those who do not show any CMAP decline after the first trial, displays a marked reduction of CMAP amplitudes after the third trial. In some patients, there is also a gradual decrease in PEMPs.

With a long-exercise test, both during and after exercise, there is often marked reduction of CMAPs from baseline which can remain so for prolonged periods, sometimes exceeding an hour. Furthermore, if the muscle is cooled and the patient asked to exercise the muscle, a pronounced (50% or more)

decrement in the CMAP can be documented in most patients with paramyotonia.

While needle electromyography reveals normal MUAPs, this may be quite difficult to observe secondary to the overt induction of profuse trains of myotonic potentials. The myotonia is usually worse in the distal compared to proximal aspects of the limbs. In patients with PMC and periodic paralysis, local cooling of the muscle results in dense fibrillation-like potentials and the gradual reduction in MUAP activity. As the muscle becomes flaccid, electrical myotonia vanishes and complete electrical silence is observed. In contrast in patients with pure PMC, while cooling results in increased myotonic discharges, the muscle strength, MUAP morphology, and recruitment remain normal and the patients continue to generate maximal interference pattern (Table 2).

Molecular genetics PMC with and without episodes of periodic paralysis are caused by mutations in the α-subunit of the voltage-gated sodium channel (*SCN4A*).

Treatment While tocainide and mexiletine have both shown to prevent the muscle stiffness and weakness associated with cold, mexiletine is the preferred medication due to the side effects associated with tocainide. Mexiletine has been demonstrated to correct the cold-induced decrements of baseline CMAPs following exercise or repetitive stimulation, in correlation with the observed clinical improvement. Hydrochlorothiazide has also been shown to be effective in relieving the myotonia.

Hypokalemic Periodic Paralysis Type 2

While approximately 70% of familial hypokalemic periodic paralysis cases are caused by mutations in the L-type Ca channel gene *CACNA1S* (HypoPP1), mutations in *SCN4A* (HypoPP2) account for approximately 10% of genetically confirmed families. Missense mutations in the highly conserved arginine residues (R699 and R672) of the voltage-sensing segment 4 (S4) of domain 2 (D2) are associated with HypoPP2. These mutations enhance inactivation to produce a net loss-of-function defect, as opposed to the gain-of-function observed in potassium-sensitive periodic paralysis, although the mechanism linking this defect to depolarization and hypokaliemia remains obscure.

Clinical Features Whereas no definite clinical features truly identify HypoPP1 from HypoPP2, a large retrospective series, including patients characterized with molecular lesions for HypoPP1 (42) and HypoPP2 (7), showed that age of onset was earlier (average 10 years) and the duration of episodes longer (average 20 h) in HypoPP1 versus HypoPP2 (16 years and 1 h, respectively), and more than 70% of HypoPP1 patients had fixed proximal weakness compared with none of the HypoPP2 patients.

Histopathology Unlike HypoPP1, in which muscle biopsies demonstrate myofibers that contain vacuoles, the myofibers in patients with hypoPP2 show mostly tubular aggregates.

Electrophysiologic findings With the long exercise test, 10–20 min into the rest phase and after cessation of exercise, there is a decrease of CMAP amplitude which is significantly milder than that observed in HypoPP1 patients. The needle electromyographic examination between attacks of muscle paralysis does not show any detectable myotonia or other abnormalities in the majority of patients.

Treatment It is important to note that patients with hypoPP2 may worsen with acetazolamide treatment, unlike patients with hypoPP1 or the potassium-sensitive periodic paralysis caused by *SCN4A* mutations.

Calcium Channelopathies

Primary hypokalemic periodic paralysis type 1 (HypoPP1) is caused by mutations in the muscle L-type voltage-gated calcium channel gene, specifically in the α1-subunit (*CACNA1S*), which is encoded on chromosome 1q31-32. This calcium channel (CaV1.1) is composed of five subunits (α1, α2, β, δ, and γ). The α1-subunit is composed of four domains, each containing six transmembrane segments (S1–S6), and contains the dihydropyridine (DHP) receptor. This receptor functions not only as a pore, conducting calcium ions into the T-tubule of skeletal muscle, but also as a voltage sensor for excitation–contraction coupling. The other subunits of the calcium channel serve to regulate the function of the α1-subunit. The S4 segment of the α1-subunit is important in conferring voltage-sensing properties to the channel, and is the site for most of the mutations identified in patients with hypokalemic periodic paralysis. The mechanism for the induced attacks of weakness in HypoPP is not understood. As with other forms of periodic paralysis, muscle fibers are depolarized and inexcitable during an attack. *In vitro*, as opposed to the hyperpolarization seen in wild-type fibers in low

K solutions, muscle fibers from patients with HypoPP1 paradoxically depolarize. However, the source of the depolarizing current has remained elusive. The available data suggest a loss-of-function defect with reduced ionic current density. A study on fibers biopsied from a patient with an R528H calcium channel mutation detected a reduction in ATP-sensitive K current, more easily tying it to the depolarization seen with hypokalemia and suggesting a secondary channelopathy resulting from altered calcium homeostasis. It has been also suggested that reduced calcium influx through the T-tubule may be the cause of impaired excitation–contraction coupling.

Primary Hypokalemic Periodic Paralysis Type 1

Clinical features Hypokalemic periodic paralysis type 1 is the most frequent form of periodic paralysis with an estimated prevalence of 1/100 000. It is an autosomal dominant disorder, with reduced penetrance in women (a male to female ratio of 3 or 4 to 1). Symptom onset is usually in the first two decades when patients note the development of episodic weakness. The bouts of weakness can occur at any time of day, although early morning hours appear to have a slightly higher propensity for being associated with weakness. Most patients note that some form of strenuous exercise or exertion followed by rest or sleep usually precipitates an attack. Other aggravating factors include heavy meals rich in carbohydrates and sodium, alcohol consumption, exposure to cold, emotional stress, intercurrent viral illness, lack of sleep, menstruation and specific medications (e.g., beta agonists, corticosteroids, and insulin). The episodes of weakness are often preceded by a vague sensation of heaviness in the lumbosacral region, thighs, and calves which spreads to involve other muscle groups, primarily those in the proximal upper limbs. If the patient notes the above prodrome, mild exercise may stave off the full blown attack of weakness; however, this is not always successful. Severity of an attack can range from mild weakness of an isolated muscle group to severe generalized paralysis. The facial and respiratory muscles as well as the sphincter muscles are typically spared or only minimally involved during particularly severe attacks; however, during peak weakness, reflexes are absent and the muscle is electrically unexcitable. Myotonia or paramyotonia are not seen in hypokalemic periodic paralysis. Death from respiratory muscle involvement and cardiac arrhythmia secondary to hypokalemia have been reported. Profound weakness may last for several hours to more than a day in some instances. Some patients may note a residual weakness for several days following an attack. Characteristically, those muscles affected last are the first to return to normal strength. The frequency of these attacks of weakness is highly variable and can occur from several times per week to a few lifetime episodes. It generally decreases after the age of 30 years and some patients become attack free in their 40s or 50s. Many patients eventually develop fixed permanent weakness after multiple attacks of paralysis, independent of their frequency or severity, with the proximal muscles being the most prone, especially in the legs.

While provocative testing can be performed when the diagnosis is unclear, the availability of commercial genetic testing usually makes it unnecessary. The goal of provocative testing is to reduce serum potassium levels to see whether typical attack of weakness could be induced. This could be accomplished by an oral or intravenous load of glucose, although insulin is also sometimes required to sufficiently lower the potassium level. The patient's strength, cardiac function (EKG), and electrolytes need to be carefully monitored during the study.

Laboratory features Between bouts of weakness, the serum potassium is normal, while attacks of weakness are associated with decreased serum potassium levels, usually below 3.0 mEq l^{-1}. During severe attacks, there is oliguria with urinary retention of sodium, potassium, chloride, and water. Obstipation and diaphoresis may also occur. The EKG may demonstrate bradycardia, flattened T waves, prolonged PR and QT intervals, and notably U waves, secondary to the hypokalemia. Serum CK levels are normal or only mildly elevated between attacks and increased during an attack of weakness (Table 2).

Histopathology The muscle biopsy may demonstrate multiple or single intracellular vacuoles, tubular aggregates, and dilation of the sarcoplasmic reticulum (Table 2). Muscle fiber size variation, split fibers, hypertrophic and atrophic fibers may also be present, and rarely necrotic and degenerating muscle fibers are noted.

Electrophysiologic findings Sensory and motor nerve conduction studies are normal between attacks of weakness. During paralytic attacks, sensory studies remain normal and motor responses continue to demonstrate a normal conduction velocity for as long as a response can be obtained. During the attack, the CMAP amplitudes decline precipitously secondary to muscle membrane inexcitability. Repetitive stimulation during this time can maintain

the CMAP amplitudes to some degree, supporting the clinical impression that mild exercise can stave off an attack.

In between the attacks, contrary to patients with potassium-sensitive periodic paralysis, there are minimal changes in CMAP immediately after the exercise phase of the short- or long-exercise tests, or with repeated trials of short exercise. The long-exercise test shows a delayed decline in CMAP amplitudes 10–20 min after the end of exercise.

The needle electromyographic examination between attacks of muscle paralysis does not show any detectable abnormality in the majority of patients. However, in the patients who develop a myopathy characterized by persistent muscle weakness, MUAPs with decreased amplitudes and durations and an increase in the number of polyphasic potentials is noted. Rarely, sparse fibrillation potentials may be observed, and there are indications of early recruitment pattern during voluntary activity. Early in an attack of weakness, there is a slight increase in spontaneous potentials (fibrillation potentials and positive sharp waves) and insertional activity, which are a reflection of the hyperirritable muscle membranes and not indicative of denervation. As the paralytic attack progresses, there is a decrease in the amplitude and duration of voluntary MUAPs as well as an overall decrease in the number contributing to the interference pattern. When the paralytic attack is maximal there is marked reduction or complete absence of insertional activity, and there are minimal, if any, voluntary MUAPs. A careful search usually demonstrates end plate activity, although this is also reduced in amplitude. Recovery from muscle paralysis essentially displays the above-noted needle electromyographic changes in reverse.

Between attacks, single fiber electromyography shows normal jitter along with a slight increase in fiber density. The latter is likely the result of a mild myopathic process which is also noted on histopathology. During an attack, the reduction in muscle membrane excitability results in a dropout of single muscle fibers and a decrease in fiber density, compared to both normal and interattack values for the patients. This is accompanied by a slight increase in jitter with occasional potential blocking (Table 2).

Molecular genetics Primary hypokalemic periodic paralysis is caused by mutations in the α-subunits of skeletal muscle L-type calcium channel gene (*CACNA1S*) in up to 70% of cases (HypoPP1). In 10% of affected individuals (HypoPP2), the mutations are present in the α-subunits of skeletal muscle sodium channel gene, *SCN4A*, and no mutation has been identified for the remaining 20% of patients. For both, the mutations occur in highly conserved arginine residues in the voltage-sensing segments. A small number of definitely affected kindreds do not have a mutation in either the calcium or sodium genes and are not linked to these loci. The possibility that they are caused by mutations in the potassium channel gene (*KCNE3*) has been suggested, but not verified. HypoPP1 may be allelic with malignant hyperthermia; malignant hyperthermia susceptibility without PP has been described with mutation in *CACN1AS*. Hence, all HypoPP patients undergoing surgery should be monitored for malignant hyperthermia-like reactions (rigidity and marked elevation in CK). However, postoperatively, paralysis in this group of patients is more likely related to the stress of surgery.

Treatment Preventive measures include avoidance of provocative factors (e.g., ingestion of high-carbohydrate meals and strenuous exercise). Acetazolamide (125–1000 mg daily) and potassium salts (0.25–0.5 mEq kg^{-1}) can be prophylactically administered to prevent hypokalemia and attacks of weakness. Dichlorphenamide (50–150 mg daily) appears to be effective in reducing attack frequency and severity. Triamterene (25–100 mg daily), and spironolactone (25–100 mg daily) may also be used to prevent attacks and improve interattack weakness, when acetazolamide and dichlorphenamide are not effective. Patients may take oral potassium (15–30 mEq) at the beginning of an attack to alleviate or shorten the episode. We treat acute attacks with oral potassium salts (0.25 mEq kg^{-1} body wt.) every 30 min until strength improves. If the patient's condition precludes oral potassium, intravenous potassium (KCL bolus 0.05–0.1 mEq kg^{-1} body wt. or 20–40 mEq l^{-1} of KCL in 5% mannitol) may be administered. Cardiac monitoring is essential throughout treatment.

Potassium Channelopathies

Andersen–Tawil Syndrome (Andersen Syndrome, or Klein–Lisak–Andersen Syndrome)

Clinical features Andersen–Tawil syndrome (ATS) is a rare ion channel disorder characterized by the clinical triad of periodic paralysis, ventricular arrhythmias associated with long QT and skeletal developmental anomalies. A diagnosis of ATS can be made when an individual exhibits two of these three cardinal features. Sixty percent of affected individuals manifest the complete triad while ~80% express two of the three cardinal features. Inheritance is autosomal dominant although *de novo* mutations

are frequent, and phenotypic expression is extremely variable. Prevalence is unknown but estimated at 1/10 that of HypoPP, and symptom onset is typically within the first or second decade. ATS must be excluded in all individuals presenting with episodic weakness and/or a long QT syndrome.

Neuromuscular manifestations. The neuromuscular manifestations of ATS consist of intermittent weakness that may arise spontaneously or be triggered by rest following exertion and prolonged rest. The attacks of paralysis usually begin in the first or second decade of life and may be associated with elevated, normal, or most commonly reduced ictal serum K^+ levels. Patients with ATS do not show evidence of myotonia or paramyotonia. The attacks vary in duration (hours to days), severity, and frequency between and within affected individuals, and might range from a single lifetime event to daily bouts of weakness. Permanent proximal weakness often develops over time.

Cardiac manifestations. Ventricular arrhythmias occur as a primary manifestation of ATS, and a number of potentially life-threatening arrhythmias, including bidirectional ventricular tachycardia (VT), polymorphic VT, and torsades de pointes may occur. The cardiac arrhythmias may be asymptomatic or manifest as palpitations, syncope, or even cardiac arrest necessitating defibrillator implantation. No specific triggers have been associated with the ventricular arrhythmias of ATS. One of the most common ECG findings is a long QT interval, which is recognized as an integral feature of ATS and may serve as a trigger for fatal ventricular arrhythmias.

Developmental anomalies. Various developmental abnormalities are associated with ATS, most commonly including clinodactyly, hypertelorism, low set ears, mandibular hypoplasia, syndactyly, and scoliosis. Other features less commonly associated are short stature, a broad nose and forehead, cleft or high-arched palate, short digits, vaginal atresia, brachydactyly, cardiac valve abnormalities, and hypoplastic kidney. A recent study on ten subjects with the *KCNJ2* mutation has suggested a distinct neurocognitive phenotype in this group of patients, characterized by deficits in executive function and abstract reasoning.

While provocative testing with the induction of hypo- or hyperkalemia may be performed when the diagnosis is unclear, there is an increased risk of precipitating arrhythmias. Genetic testing is also available to aid in the diagnosis.

Laboratory features Serum CK is normal or only mildly elevated (less than 5 times normal). Serum potassium levels can be normal, elevated, or decreased during attacks of weakness. EKG manifestations include prolongation of the corrected QT interval (present in 80% of patients), prominent U waves, premature ventricular contractions, ventricular bigeminy, and polymorphic VT. A subset of patients manifests a unique form of VT, bidirectional VT, which ischaracterized by beat-to-beat alternating QRS axis polarity (Table 2).

Histopathology Muscle biopsy frequently reveals tubular aggregates, and patients with fixed weakness may have nonspecific myopathic changes (Table 2).

Electrophysiologic findings Motor and sensory nerve conduction studies are normal, and there are no myotonic discharges. Electromyography is normal between the attacks. The long-exercise protocol may detect a delayed decrement in ATS patients, similar to what is seen in HypoPP1 (Table 2).

Molecular genetics and pathophysiology The molecular basis for the majority (approximately two-thirds) of ATS cases is missense mutations or small deletions in the *KCNJ2* gene (ATS1), which encodes an inwardly rectifying potassium channel (Kir2.1) and is predominantly expressed in the heart, skeletal muscle, and brain. While this mutation is known to cause symptomatic impairment of muscle membrane depolarization and repolarization, its role in the associated skeletal developmental anomalies and possible CNS manifestations is not fully clear. Kir2.1 channels stabilize the resting membrane potential in skeletal and cardiac muscle. The majority of *KCNJ2* mutations result in the failure of appropriate conduction, and many alter the binding of phosphatidylinositol 4,5 bisphosphate, an important regulator of Kir2.1 channel function. It is postulated that reduced Kir2.1 channel function in skeletal muscle may result in sustained membrane depolarization, failure of action potential propagation, and flaccid paralysis. It may also prolong the most terminal phase of repolarization in the heart and lead to delayed after-depolarizations. Targeted disruption of *KCJN2* in homozygous knockout mice results in neonatal lethality owing to aspiration and asphyxia from a severe cleft palate, which suggests a role for Kir2.1 in skeletal development and may be the basis for dysmorphic features in ATS. In approximately 20% of patients with an ATS phenotype, no mutation of *KCJN2* has been identified (ATS2). Possible molecular candidates include genes homologous to Kir2.1, and gene products that modulate Kir2.1 channel expression, trafficking, and function.

Treatment Early recognition of the potential cardiac conduction abnormalities is important, since it might necessitate treatment with antiarrhythmic agents or an implantable cardioverter-defibrillator. The involvement of both cardiac and skeletal muscle in ATS presents a unique treatment challenge, since specific treatments may have paradoxical effects on cardiac and skeletal muscle function. Treatment to date has been largely anecdotal and empirical, and while small doses of acetazolamide may prevent paralytic attacks in some patients, it has not been shown to be beneficial in all cases.

Other Causes of Periodic Paralysis

Secondary Hyperkalemic or Hypokalemic Paralysis

Secondary hyperkalemia and hypokalemia may cause generalized weakness, mimicking the familial forms of periodic paralyses, and must be excluded particularly in patients with no family history (Table 3). In comparison with potassium-sensitive periodic paralysis, the secondary cases do not exhibit clinical or electrical myotonia on examination. Secondary hypokalemia might also precipitate bouts of weakness as well as causing a necrotizing myopathy. In both cases, muscle weakness resolves with correction of the electrolyte imbalance.

Thyrotoxic Periodic Paralysis

Thyrotoxic periodic paralysis resembles hypokalemic periodic paralysis, except that it occurs sporadically and the associated hypokaliemia is often profound. It should be excluded in patients presenting after age 20 years or without a family history, by looking for a suppressed thyroid stimulating hormone (TSH) and elevated free T4 or free T3 levels. It is possible that the patients have an underling channel defect that clinically manifests itself when the patient becomes hyperthyroid. An initial report of an R83H mutation in the *KCNE3* gene causing thyrotoxic PP proved erroneous, since it is present in 1% of healthy controls. While hyperthyroidism is more common in women, the majority of cases of thyrotoxic periodic paralysis are male. The disorder is most prevalent in Asian men, where it may be inherited, although it has been described in all races. The patients can develop progressive muscle weakness secondary to hyperthyroidism, with the attacks of weakness resolving with treatment of the thyrotoxicosis and recurring with exacerbation of the hyperthyroid state. Unlike the familial forms of hypokalemic and hyperkalemic periodic paralysis, acetazolamide does not appear to have any significant benefit; however, β-blockers may be effective in reducing the frequency and severity of the paralytic attacks.

See also: Calcium Channels; Epilepsy: Channelopathies; Hemiplegic Migraine and Channel Disorders; Inwardly Rectifying Potassium Channels; Sodium Channels; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Table 3 Selected causes of hyperkalemic and hypokalemic paralysis

Hyperkalemic paralysis
Addison's disease
Chronic renal failure
Hypoaldosteronism
Iatrogenic (spironolactone, triamterene, and potassium supplementation)
Rhabdomyolysis
Hypokalemic paralysis
Alcoholism
Barium poisoning
Bartter's syndrome
Renal tubular acidosis
Hyperaldosteronism
Hyperthyroidism
Iatrogenic (diuretics, corticosteroids, Amphotericin B toxicity, and licorice)

Adapted from Amato A and Dumitru D (2002) Hereditary myopathies. In: Dumitru D, Amato A, and Zwarts M (eds.) *Electrodiagnostic Medicine*, 2nd edn., pp. 1265–1370. Philadelphia: Hanley & Belfus, Inc.

Further Reading

Amato A and Dumitru D (2002) Hereditary myopathies. In: Dumitru D, Amato A, and Zwarts M (eds.) *Electrodiagnostic Medicine,* 2nd edn., pp. 1265–1370. Philadelphia: Hanley & Belfus, Inc.

Cannon SC (2006) Pathomechanisms in channelopathies of skeletal muscle and brain. *Annual Review of Neuroscience* 29: 387–415.

Davies NP and Hanna MG (2003) The skeletal muscle channelopathies: Distinct entities and overlapping syndromes. *Current Opinion in Neurology* 16(5): 559–568.

Fournier E, Arzel M, Sternberg D, et al. (2004) Electromyography guides toward subgroups of mutations in muscle channelopathies. *Annals of Neurology* 56(5): 650–661.

Katz JS, Wolfe GI, Iannaccone S, et al. (1999) The exercise test in Andersen syndrome. *Archives of Neurology* 56(3): 352–356.

Kung AW (2006) Clinical review: Thyrotoxic periodic paralysis: A diagnostic challenge. *Journal of Clinical Endocrinology & Metabolism* 91(7): 2490–2495.

Miller TM, Dias da Silva MR, Miller HA, et al. (2004) Correlating phenotype and genotype in the periodic paralyses. *Neurology* 63(9): 1647–1655.

Ptacek L (1998) The familial periodic paralyses and nondystrophic myotonias. *American Journal of Medicine* 105(1): 58–70.

Venance SL, Cannon SC, Fialho D, et al. (2006) The primary periodic paralyses: Diagnosis, pathogenesis and treatment. *Brain* 129(Pt. 1): 8–17.

Yoon G, Quitania L, Kramer JH, et al. (2006) Andersen–Tawil syndrome: Definition of a neurocognitive phenotype. *Neurology* 66(11): 1703–1710.

Relevant Websites

http://rarediseasesnetwork.epi.usf.edu – Consortium for Clinical Investigations of Neurological Channelopathies, Rare Diseases Clinical Research Network.

http://www.clinicaltrials.gov – NIH Clinical Trials Website.

http://www.periodicparalysis.org – Periodical Paralysis Research Center, Periodical Paralysis Association.

Paroxysmal Sympathetic Storms

E E Benarroch, Mayo Clinic, Rochester, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Clinical Manifestations and Pathophysiology

Clinical Manifestations

Massive sympathoexcitation, as occurs in acute neurologic catastrophes, results in activation of both sympathoneural (peripheral sympathetic ganglia) and sympathoadrenal (adrenal medulla) discharge, as reflected by increased muscle and skin sympathetic nerve activities and circulating levels of norepinephrine and epinephrine. Paroxysmal sympathetic hyperactivity produces hypertension, tachycardia, pallor, and excessive sweating (hyperhidrosis), commonly associated with abnormalities in thermoregulation (hypothermia or hyperthermia). Massive sympathoexcitation may result in severe hypertension, malignant tachyarrhythmias, myocardial injury, and pulmonary edema. Sympathetic hyperactivity may occur with disorders affecting the central or peripheral nervous systems (Table 1). In general, sympathetic hyperactivity of central origin tends to begin abruptly after the acute injury, then fluctuates, and persists for several hours or days. Sympathetic hyperactivity due to peripheral disorders is often paroxysmal and recurs over longer periods.

Pathophysiology

The manifestations of paroxysmal sympathetic hyperactivity reflect excessive activation of vascular α-adrenergic receptors, cardiac β-adrenergic receptors, and cholinergic muscarinic receptors in the sweat glands. Sympathetic hyperactivity may occur with disorders affecting the cerebral cortex, diencephalon, brain stem, spinal cord, or peripheral nervous system. The mechanisms triggering sympathetic hyperactivity are incompletely understood, and vary according to the underlying cause (Figure 1).

In focal seizures, sympathetic hyperactivity results from paroxysmal discharges of neurons in the temporal or frontal lobes connected with hypothalamic and brain stem autonomic centers. In stroke, it may reflect loss of descending sympathoinhibitory output to these areas. Immediately after an intracranial catastrophe, such as traumatic or hypoxic brain injury, subarachnoid or intracerebral hemorrhage, or mass lesions affecting the brain stem, there is a massive sympathoneural and sympathoadrenal activation, resulting in increase in sympathetic discharge and catecholamine release. This dramatic response reflects direct excitation or disinhibition, or both, of sympathoexcitatory areas of the hypothalamus or the medulla. These areas include the paraventricular nucleus of the hypothalamus, which is involved in the stress response, and the rostral ventrolateral medulla, which contains sympathoexcitatory neurons that are activated by hypoxia or ischemia. The rostral ventrolateral medulla neurons are thought to be responsible for the hypertension experienced during the classical 'Cushing reflex.' This response is characterized by hypertension, bradycardia, and apnea in response to an increase in intracranial pressure, resulting in decrease in cerebral perfusion pressure. An initial tachycardia associated with hypertension occurs before the onset of bradycardia.

Another mechanism of acute sympathetic hyperactivity is massive activation of segmental spinal sympathoexcitatory reflexes in the setting of interruption of descending modulatory pathways from the hypothalamus and brain stem. This is called autonomic dysreflexia and occurs in patients with spinal cord injury above the T5 level. Episodes of sympathoexcitation may also occur in patients with failure of baroreflex inhibition of sympathetic output, as occurs in patients with Guillain–Barré syndrome or porphyria. In some disorders, the mechanisms of sympathoexcitation may be both central and peripheral. In these conditions, sympathetic hyperactivity is associated with confusional state and with motor abnormalities, such as muscle rigidity, tremor, or myoclonus. Important examples are alcohol withdrawal delirium, neuroleptic malignant syndrome, and serotonin syndrome. Stimulant drugs, such as cocaine or amphetamine, may lead to excessive release or impaired uptake of catecholamines, resulting in similar manifestations. In many conditions, sympathetic hyperactivity is associated with increased muscle tone. This suggests that somatosympathetic reflexes from muscle mechanoreceptors or chemoreceptors during the episodes of increase muscle tone may trigger or contribute to excessive sympathetic activation.

Syndromes of Sympathetic Hyperactivity

Sympathetic hyperactivity is a critical manifestation of specific syndromes that require prompt recognition for appropriate management.

Acute Sympathetic Hyperactivity in Neurologic Catastrophes

Severe brain injury An episodic hypersympathetic state occurs in patients with severe brain injury. This

Table 1 Causes of autonomic hyperactivity

Forebrain disorders
Insular stroke
Temporal or anterior cingulate seizures
Close head injury
Subarachnoid hemorrhage
Mass lesions producing increased intracranial pressure
Episodic hyperhydrosis with or without hypothermia
Familial or sporadic fatal insomnia
Limbic encephalitis
Posterior fossa disorders
Brain stem ischemia
Cerebellar or brain stem neoplasms
Syringobulbia
Tetanus
Leigh syndrome
Spinal cord disorders
Traumatic injury
Devic syndrome
Stiff-man syndrome
Peripheral nerve disorders
Guillain–Barré syndrome
Porphyria
Thallium intoxication
Morvan syndrome
Iatrogenic disorders
Neuroleptic malignant syndrome
Serotonin syndrome
Monoamine oxidase inhibitors
Drug intoxication or withdrawal
Stimulant (cocaine, amphetamines) intoxication
Alcohol withdrawal
Opioid withdrawal

episodic sympathetic hyperactivity has been variably referred to in the literature as diencephalic seizures, paroxysmal sympathetic storm, and paroxysmal autonomic instability with dystonia. It occurs most commonly in the setting of severe head trauma with diffuse axonal injury and in the postresuscitation period after severe anoxic–ischemic brain insults, but may also occur in the setting of intracerebral hemorrhage, hydrocephalus, brain stem ischemia, or mass lesions. The manifestations of this episodic hypersympathetic state include fever (often $>39\,^{\circ}C$), tachycardia, hypertension, diaphoresis, pupil dilatation, tachypnea, and shivering, associated with muscle hypertonia and extensor posturing. The paroxysms usually start 5–7 days after the injury, but they may appear earlier. The episodes follow a regular pattern, occur on average 1–3 times per day, and each lasts between 2 and 10 h. Although in most patients sympathetic outflow diminishes over time and the manifestations of sympathetic hyperactivity tend to disappear within a few days, some patients develop recurrent paroxysms of sympathetic hyperactivity during the acute hospitalization phase and later recovery phase. Over time, the episodes tend to become less frequent but more prolonged and severe, before subsiding several months after the acute event.

The differential diagnosis of this syndrome is based on the recognition of the regular pattern of hypersympathetic spells in a patient with acute intracranial lesions. This syndrome should be distinguished from other disorders that present with autonomic instability, including seizures, impending brain herniation causing the Cushing response, tetanus, malignant hyperthermia, neuroleptic malignant syndrome, and lethal catatonia. In all patients with suspected sympathetic storms, serial blood cultures and cerebrospinal fluid examination should be performed to exclude the possibility of sepsis, meningitis, or encephalitis. All patients should also have a computed tomography or magnetic resonance scan of the brain to exclude the presence of hydrocephalus or brain stem distortion. In patients with traumatic head injury, the possibility of acute spinal injury above the fifth thoracic level producing acute autonomic dysreflexia should also be excluded by imaging study. An electroencephalogram should be obtained to exclude seizures.

Drugs that have been reported helpful to treat these sympathetic paroxysms include morphine, bromocriptine, labetalol, clonidine, clonazepam, or dantrolene. Antiepileptic drugs are typically not helpful, but should be tried if there is a persistent possibility of an epileptic nature of the spells. Autonomic storms have been associated with poor functional outcome after severe traumatic brain injury.

Subarachnoid hemorrhage Subarachnoid hemorrhage is the prototype of an acute neurologic catastrophe associated with massive sympathoexcitation. One mechanism is development of acute hydrocephalus, resulting in mechanical distortion of periventricular hypothalamic areas, particularly the paraventricular nucleus, which is critical for sympathetic activation in response to stress. The acute sympathoadrenal excitation may produce electrocardiographic (ECG) changes of cardiac injury, including depression or elevation of the ST segment, inverted or biphasic T waves, QT prolongation, and Q waves. The ECG changes resembling cardiac ischemia in patients with severe subarachnoid hemorrhage are associated with increased levels of creatine kinase and troponin, and there is echocardiographic evidence of ventricular hypokinesia. In general, the ECG and enzymatic abnormalities reverse within 1–4 days. Some patients may develop pulmonary edema, which may be due to neurogenic or cardiogenic causes. Myocardial injury is not due to cardiac ischemia, as coronary angiogram is normal in these patients. Myocardial damage results from massive release of norepinephrine from nerve terminals adjacent to the myocardium, leading to

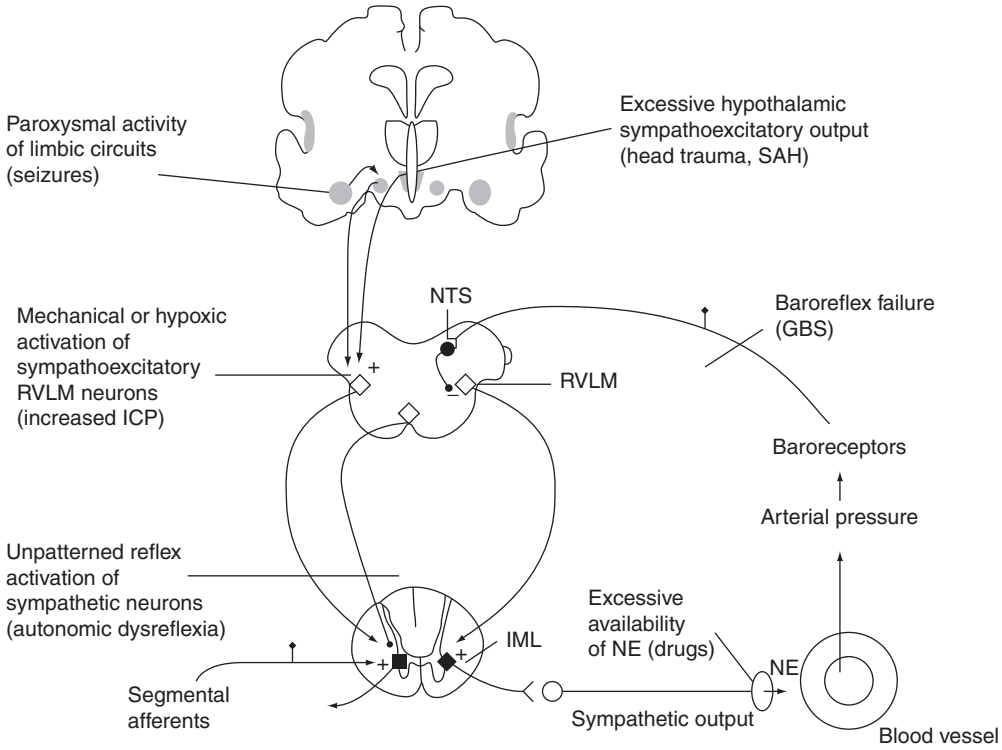

Figure 1 Pathophysiological mechanisms of paroxysmal sympathetic hyperactivity and representative examples. Sympathetic hyperactivity results from excessive sympathoexcitatory outflow from limbic circuits (as occurs with temporal lobe seizures) or the hypothalamus (as occurs in severe head injury) or subarachnoid hemorrhage (SAH). These circuits activate the preganglionic sympathetic neurons of the intermediolateral cell column (IML), either directly or via the sympathoexcitatory neurons of the rostral ventrolateral medulla (RVLM). The RVLM may be directly activated by hypoxia or mechanical distortion, as occurs in the setting of increased intracranial pressure (ICP). The sympathoexcitatory output of the RVLM is tonically inhibited by the baroreflex, via the nucleus of the solitary tract (NTS). Therefore, baroreflex failure, as occurs in the Guillain–Barré syndrome (GBS), results in paroxysmal sympathetic discharge. Interruption of descending modulatory inputs from the hypothalamus and medulla unmasks segmental sympathoexcitatory reflexes, resulting in massive sympathetic activation in response to visceral or somatic stimulation, as typically occurs in autonomic dysreflexia following high spinal cord injuries. Sympathoexcitatory effects are mediated by norepinephrine (NE). Therefore, drugs that increase NE availability, such as cocaine, amphetamines, or monoamine oxidase inhibitors, produce excessive sympathetic activity.

subendocardial contraction band necrosis ('stunned heart'). Several tachyarrhythmias may occur in the setting of subarachnoid hemorrhage. Although supraventricular are more common than ventricular arrhythmias, a prolonged QT interval is a predictor of serious arrhythmias (such as torsades de pointes) that may produce sudden death in some cases.

Chronic Recurrent Episodes of Sympathetic Activation

Seizures Seizures of temporal or extratemporal origin, and generalized tonic–clonic seizures, most commonly activate the sympathetic system, producing tachycardia, hypertension, diaphoresis, pupillary dilatation, pallor or facial flushing, and piloerection. These manifestations resemble panic attacks, or pheochromocytoma, and when they are the sole manifestation of partial seizures, the diagnosis is often delayed or missing. Sinus tachycardia can occur in more than 85% of complex partial and generalized tonic–clonic seizures. Patients with refractory epilepsy have one or more abnormalities in repolarization and/or arrhythmias during or immediately after seizures. These include atrial fibrillation, supraventricular tachycardia, premature ventricular complexes, bundle-branch block, and complete atrioventricular block. Cardiovascular and respiratory dysregulation are postulated to be a major factor in sudden unexplained death in epileptic patients. Ictal and postictal apnea and neurogenic pulmonary edema, rather than cardiovascular effects, appear to be the major cause of sudden unexplained death in epilepsy. Pulmonary edema is the result of sympathetically triggered pulmonary vasoconstriction and increased left atrial pressure from systemic hypertension; both lead to increase in intrapulmonary vascular pressure and pulmonary edema. Status epilepticus is a neurological and medical emergency. Sympathetic hyperactivity associated with prolonged seizures may lead to

death due to cardiac arrhythmias and to myocardial injury (contraction band necrosis).

Episodic hyperhidrosis with or without hypothermia This syndrome consists of spontaneous episodes of severe hyperhidrosis, in general associated with hypothermia that may recur several times a day and last several hours. During the episodes, the patients may be confused and poorly responsive. This disorder has been associated with diencephalic lesions (such as tumors leading to obstructive hydrocephalus), agenesis of the corpus callosum (Shapiro syndrome), basal forebrain malformation, hypothalamic strokes, and infections with human immunodeficiency type 1 virus, but commonly occur with no apparent cause. Clonidine, cyproheptadine, anticholinergic drugs (such as oxybutynin or glycopyrrolate), carbamazepine, chlorpromazine, and phenobarbital have been shown to be helpful in reducing hyperhidrosis or hypothermia, or both, in some individual cases.

Sympathetic Hyperactivity, Insomnia, Delirium, and Excessive Motor Activity

In several neurologic disorders, sympathetic hyperactivity is associated with confusional apathetic state, fluctuating delirium, and excessive motor activity. This syndrome is characterized by insomnia with reduced slow-wave sleep and increased incidence of rapid-eye-movement sleep without atonia. Disorders associated with this syndrome include delirium tremens (alcohol withdrawal delirium), fatal insomnia, Morvan syndrome, and limbic encephalitis.

Delirium tremens Delirium tremens is the most severe manifestation of the alcohol withdrawal syndrome and constitutes a medical emergency. It is characterized by acute or subacute confusional state with anxiety, insomnia, and tremor; vivid visual, auditory, or tactile hallucinations; and fluctuating levels of psychomotor activity. These manifestations are accompanied by sympathetic hyperactivity with tachycardia, fever, diaphoresis, flushing, and hypertension. In general, delirium appears after 2–3 days of cessation of drinking and usually lasts 48–72 h, but occasionally lasts much longer. A similar clinical picture may occur after sudden withdrawal of barbiturates or benzodiazepines. The main treatment is benzodiazepines, such as chlordiazepoxide or lorazepam. Clonidine and propranolol can reduce hypertension and tachycardia in selected patients but should only be given together with benzodiazepines.

Fatal insomnia Fatal insomnia is a prion disease that can be sporadic or familial. Patients complain of early insomnia, with inability to fall asleep at night and frequent awakening. Insomnia is associated with perspiration, salivation, tachycardia, increased arterial pressure, and pyrexia. Later in the course of the disease the patients develop increased autonomic hyperactivity and periods of hallucinatory behavior and enacted dreams; this is followed by stupor and eventual coma, preceding death. In addition to insomnia and autonomic hyperactivity, there is increased cortisolemia and suppression of circadian rhythms.

Morvan syndrome Morvan syndrome is a rare disorder characterized by severe insomnia associated with sympathetic hyperactivity manifested with profuse perspiration, tachycardia, hypertension, and increased body temperature. The patients display hallucinatory behavior and motor unit hyperexcitability (neuromyotonia), as well as cramps and myoclonus. Morvan syndrome may be associated with antibodies against voltage-gated K^+ channels, and may be a paraneoplastic manifestation of thymoma or small cell lung carcinoma.

Limbic encephalitis Limbic encephalitis is an inflammatory disorder that affects the medial portion of the temporal lobe (hippocampus and amygdala) and may also involve the insular and orbitofrontal cortex and hypothalamus. It is characterized by subacute onset of memory impairment, changes in personality, temporal lobe seizures, and autonomic instability. The common causes include herpes simplex encephalitis and autoimmune disorders. Autoimmune limbic encephalitis may be a paraneoplastic manifestation of the presence of anti-Hu autoantibodies in patients with small cell lung cancer, anti-Ma2 in patients with testicular germ cell tumors, or voltage-gated K^+ channels in patients with or without associated Morvan syndrome and thymoma. The management of this paraneoplastic syndrome includes identification and treatment of the underlying malignancy, if present, and immunosuppressive therapy with corticosteroids, intravenous immunoglobulin, or plasma exchange.

Autonomic Dysreflexia in Spinal Cord Injury

Acute lesions of the upper spinal cord, resulting from traumatic injuries or severe inflammatory conditions, such as Devic disease, commonly produce autonomic failure, resulting in orthostatic hypotension, urinary retention, and ileus during the state of flaccid paralysis (spinal shock). However, upon recovery from the acute spinal shock state, patients with lesions above the T5 level may experience episodes of severe sympathetic hyperactivity triggered by reflex stimuli below the level of the lesion. This generalized reflex

sympathetic activation, called autonomic dysreflexia, is due to interruption of descending sympathomodulatory inputs from the brain stem, which control segmental sympathetic reflexes and mediate baroreflex inhibition of the sympathetic preganglionic neurons. Autonomic dysreflexia may be triggered by a variety of visceral or somatic stimuli, most commonly bladder distension. The episodes of autonomic dysreflexia may be life threatening. The most serious manifestation is severe hypertension, which may produce hypertensive encephalopathy or intracranial, subarachnoid, or retinal hemorrhages. The episodes are often heralded by headaches and face flushing, and are accompanied by piloerection, extensive diaphoresis above the level of the lesion, and vagally mediated reflex bradycardia. Autonomic dysreflexia may be prevented by appropriate medical care of patients with spinal cord injury, including bowel and bladder programs and skin care. Both the patient and family should be educated about possible causes, prevention, presentation, and first-aid measures. In the acute setting, the patient is placed in an upright position, with removal of the precipitating stimulus. When hypertension persists, α-adrenergic blockers, such as phenoxybenzamine or prazosin, or nifedipine may be used.

Guillain–Barré Syndrome

Autonomic instability is a leading cause of mortality in Guillain–Barré syndrome, and tends to occur in severe cases, associated with respiratory failure. One important mechanism is impaired baroreflex modulation of sympathetic cardiovascular output due to demyelination of vagal afferents in this disorder. Manifestations include paroxysmal or sustained hypertension, orthostatic hypotension, vasodepressor responses, tachyarrhythmias, and vagal-induced bradycardia during intubation or tracheal suction. Severe hypertension may result in subarachnoid hemorrhage or posterior leukoencephalopathy with cortical blindness. In general, spontaneous fluctuations of blood pressure in patients with Guillain–Barré syndrome are best left untreated unless there is compromise of target organs. These patients may suffer severe hypotension or hypertension in response to labetalol or adrenergic agonists, respectively. Treatment of precipitating causes, particularly pain, is critical.

Baroreflex failure may also result from damage of afferent structures, including the arterial baroreceptors, baroreceptor afferents, or nucleus of the solitary tract where these afferents terminate. Common causes include extensive neck surgery and radiation therapy for cancers of the neck, bilateral carotid tumor resections, and repeated trauma to the neck. Baroreflex failure manifests with episodes of paroxysmal hypertension, usually accompanied of tachycardia, headache, sensation of warmth or flushing, diaphoresis, and palpitations. These episodes are commonly triggered by psychological stress, physical exercise, or pain. Patients with baroreflex failure are susceptible to dramatic changes in blood pressure after ingestion of vasoactive drugs, including over-the-counter decongestants. One important differential diagnosis is pheochromocytoma. Clonidine profoundly reduces arterial pressure and plasma norepinephrine levels in patients with baroreflex failure, whereas it has no effect in patients with pheochromocytoma.

Iatrogenic Causes of Sympathetic Hyperactivity

Neuroleptic malignant syndrome Neuroleptic malignant syndrome (NMS) is an idiopathic reaction resulting from impaired central dopaminergic transmission in the hypothalamus and basal ganglia. The two common triggers are drug-induced blockade of D_2 dopaminergic receptors and sudden withdrawal of dopaminergic agonists in patients with Parkinson disease. The drugs that commonly trigger NMS include the first-generation antipsychotic agents (e.g., haloperidol, phenothiazines), and antiemetics (metoclopramide), but NMS has also been described with 'atypical' antipsychotic drugs such as clozapine, olanzapine, risperidone, and quetiapine. NMS may also develop in patients with Parkinson's disease after withdrawal of levodopa, dopaminergic agonists, catechol-*O*-methyltransferase inhibitors, or amantadine. Patients with Lewy body disease are at particular risk to develop NMS after receiving even a low dose of antipsychotic drugs. NMS may also result from overdose of tricyclic antidepressants, citalopram, amphetamines, cocaine, or lithium.

The clinical manifestations of NMS include mental status change ranging from confusional state to coma, severe muscle rigidity with or without tremor, dystonia and mutism, fever (>38 °C), and autonomic hyperactivity, including diaphoresis, tachycardia, and hypertension. Two important laboratory features are leukocytosis and elevated creatine kinase levels, reflecting rhabdomyolysis secondary to muscle rigidity.

Typically, NMS evolves over 24–72 h, but may rarely develop slowly over a few days. In uncomplicated cases, the course lasts 7–10 days. Management of NMS includes immediate withdrawal of the precipitating drug, supportive measures (including intravenous fluids and nutrition, temperature reduction), and pharmacologic treatment with dopaminergic agonists such as bromocriptine, and muscle relaxants such as dantrolene.

The main differential diagnosis includes malignant hyperthermia, acute lethal catatonia, neuroleptic-induced heat stroke, central anticholinergic syndrome, and serotonin syndrome. Malignant hyperthermia typically occurs with the use of potent halogenated inhalation anesthetics (such as halothane or isoflurane) and depolarizing muscle agents such as succinylcholine. Management of this condition consists of immediate withdrawal of the offending agent, supportive measures, and dantrolene sodium. Acute lethal catatonia is typically preceded by a history of behavioral changes (including agitation, automatisms, negativism, and stereotypies) before the onset of motor symptoms, which include not only rigidity but also abnormal postures (catalepsy) and waxy flexibility. Lethal catatonia typically occurs in the setting of schizophrenia but may also follow meningitis, encephalitis, head trauma, and metabolic disturbances such as Wernicke encephalopathy or porphyria. Treatment includes benzodiazepines and electroconvulsive therapy. The central anticholinergic syndrome occurs following administration of drugs that block central and peripheral muscarinic receptors, particularly tricyclic antidepressants such as amitriptyline. However, unlike NMS, fever and tachycardia in anticholinergic syndrome are associated with decreased sweating, dry mouth, mydriasis, and urinary retention.

Serotonin syndrome Serotonin syndrome results from excessive stimulation of central and peripheral serotonin receptors, particularly 5-HT_{2A} receptors. Unlike NMS, serotonin syndrome is a dose-dependent phenomenon and develops rapidly following administration of single doses of, or combinations of, drugs that enhance serotonergic transmission. These include serotonin reuptake inhibitors such as fluoxetine, fluvoxamine, paroxetine, sertraline, citalopram, venlafaxine, trazodone, nefazodone, clomipramine); 5-HT_1 agonists (buspirone, triptan); monoamine oxidase inhibitors; opioids (meperidine, fentanyl, tramadol, dextromethorphan), triptans, valproate, lithium, antiemetic agents (metoclopramide, ondansetron), drugs of abuse such as lysergic acid diethylamide and methylenedioxymethamphetamine ('Ecstasy'), and dietary supplements and herbal products (tryptophan, St. John's wort). The typical manifestations of serotonin syndrome include agitation, hyperreflexia, clonus, fever, mydriasis, tachycardia, diaphoresis, and hyperactive bowel sounds or diarrhea. Severely affected patients may have agitated delirium, muscle rigidity, hyperthermia, and severe hypertension and tachycardia that may abruptly be followed by hypotension and shock in patients. The most distinguishing features of serotonin syndrome are clonus and diarrhea, which help differentiate it from NMS and central anticholinergic syndrome, respectively. Management includes removal of precipitating drugs, supporting measures, control of the hyperthermia and autonomic instability, control of agitation with benzodiazepines, and administration of 5-HT_{2A} antagonists such as cyproheptadine. Propranolol is not indicated because it can produce severe hypotension, and bromocriptine is contraindicated as it can trigger serotonin syndrome.

See also: Autonomic Disorders; Pain and the Sympathetic Nervous System; Paroxysmal Paralysis; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Sympathetic Nervous System.

Further Reading

Baguley IJ, Nicholls JL, Felmingham KL, et al. (1999) Dysautonomia after traumatic brain injury: A forgotten syndrome? *Journal of Neurology, Neurosurgery & Psychiatry* 67: 39–43.

Blackman JA, Patrick PD, Buck ML, et al. (2004) Paroxysmal autonomic instability with dystonia after brain injury. *Archives of Neurology* 61: 321–328.

Bhanushali MJ and Tuite PJ (2004) The evaluation and management of patients with neuroleptic malignant syndrome. *Neurology Clinics* 22: 389–411.

Boyer EW and Shannon M (2005) The serotonin syndrome. *New England Journal of Medicine* 352: 1112–1120.

Brouwers PJ, Wijdicks EF, Hasan D, et al. (1989) Serial electrocardiographic recording in aneurysmal subarachnoid hemorrhage. *Stroke* 20: 1162–1167.

Bullard DE (1987) Diencephalic seizures: Responsiveness to bromocriptine and morphine. *Annals of Neurology* 21: 609–611.

Devinsky O (2004) Effects of seizures on autonomic and cardiovasacular function. *Epilepsy Currents* 4: 43–46.

Di Pasquale G, Andreoli A, Lusa AM, et al. (1998) Cardiologic complications of subarachnoid hemorrhage. *Journal of Neurosurgical Sciences* 42: 33–36.

Karlsson AK (1999) Autonomic dysreflexia. *Spinal Cord* 37: 383–391.

LeWitt PA, Newman RP, Greenberg HS, et al. (1983) Episodic hyperhidrosis, hypothermia, and agenesis of corpus callosum. *Neurology* 33: 1122–1129.

Montagna P and Lugaresi E (2002) Agrypnia excitata: A generalized overactivity syndrome and a useful concept in the neurophysiology of sleep. *Journal of Neurophysiology* 113: 552–560.

Ropper AH (1997) Management of the autonomic storm. In: Low PA (ed.) *Clinical Autonomic Disorders: Evaluation and Management*, 2nd edn., pp. 791–801. Philadelphia: Lippincott-Raven.

Zochodne DW (1994) Autonomic involvement in Guillain–Barré syndrome. *Muscle Nerve* 17: 1145–1155.

Relevant Website

http://www.mc.vanderbilt.edu/ – American Autonomic Society.

Pattern Generation

S Grillner, Karolinska Insititutet, Stockholm, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Already in the end of the nineteenth century, the question was asked whether the spinal cord contains networks that coordinate different patterns of motor behavior or, as an alternative, whether they were coordinated through chain reflexes, requiring sensory feedback. The early strategy to address such questions was to compare the motor repertoire before and after a transection of all the afferents to the limb or appendage investigated. The German investigator Munk showed in 1897 that monkeys could use their arm after deafferentation provided that the intact arm was not free to move. Very extensive experiments with deafferentation and other approaches in fish and other vertebrate species have established beyond any doubt that central networks play a very important role in generating and controlling the vertebrate movement repertoire.

Although sensory feedback, in general, is not required for a motor pattern to be produced, it nevertheless plays a prominent role in many types of movements, precision movements as well as larger synergies such as locomotor movements and postural adjustments. For optimal control of most movements, central programming provides the template, but sensory signals are used before the movement starts, to initiate the appropriate motor coordination and to adjust the movement trajectory during the movements when time permits. During slow movements, there may be ample time for sensory control during the actual ongoing movement, but during faster, particularly ballistic movements, there is no time available – and the movement depends entirely on central programming.

The Capacity of the Spinal Cord to Generate Integrated Patterns of Motor Behavior

A very important question to address was to what a degree the spinal cord contained networks that could coordinate behaviorally meaningful motor patterns. One way to address this was by an analyses of what the spinal cord could produce after a complete lesion of the spinal cord, interrupting all nerve fibers from the brain to the spinal cord. In all classes of vertebrates, it is now well established that the spinal cord devoid of any supraspinal control can coordinate locomotor movements, scratching or wiping reflexes, protective reflexes, and also reflexes that control bladder function and other autonomic functions.

Cyclostomes and Fish – Central Programming and Sensory Feedback

Some fish, such as dogfish and eel, show spontaneous well-coordinated swimming activity after a spinal transection with appropriate motor coordination along the body. Species that normally swim continuously apparently have a very high excitatbility in their spinal cord networks at rest. The movement can, however, be inhibited by certain sensory input and accelerated by others. Most fish, however, do not show any coordinated movements immediately after a transection, but sensory input from, for example, the tailfin can generate brief bouts of locomotor activity. Pharmacological activation of the spinal cord with glutamate agonists, serotonin (5-HT) agonists, or noradrenergic agonists can, in general, elicit locomotor activity in the spinal cord.

The spinal cord with afferents is thus able to generate well-coordinated locomotor activity in fish. The next question is whether this motor coordination results from a chain reflex, a view held by many investigators until the 1970s, or depends on a network within the spinal cord itself. Both in the isolated spinal cord, in which the motor coordination can be recorded from the ventral roots along the spinal cord, and in the spinal fish after transection of the sensory dorsal roots, the swimming motor coordination is retained (Figure 1). These experiments clearly demonstrate, beyond any doubt, that there is a spinal neuronal network organization that coordinates the locomotor activity. These networks are often referred to as a central pattern generator networks (CPGs).

The level of activity in the spinal locomotor CPGs is normally controlled from the mesencephalic and diencephalic locomotor command regions in the brain stem. These command regions turn on the locomotor CPGs and determine the level of activity (e.g., fast or slow swimming). The locomotor command regions provide a bilaterally symmetrical activation of the glutamatergic reticulospinal nuclei in the lower brain stem.

The fact that there is a spinal locomotor CPG that can generate the motor pattern without sensory feedback in no way means that the sensory input is without importance. In all species investigated, the activity level of the CPGs can be accelerated by tonic input from the tail region and can be inhibited from other areas such as the skin over the abdomen (for fish resting at the bottom of a lake or the sea). As important is the phasic sensory input onto the CPGs. In lamprey and dogfish, it has been shown that stretch receptors located at the lateral margin spinal cord senses the progression of each locomotor cycle. This

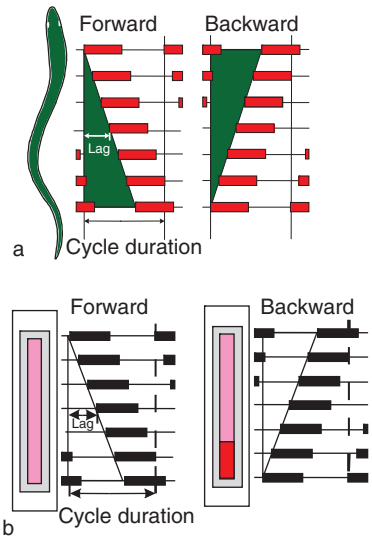

Figure 1 Intersegmental coordination through phase coupling – undulatory locomotion: (a) cyclostomes, fish, and some amphibians and reptiles; (b) in the isolated spinal cord (control). Cyclostomes, fish, and some amphibians and reptiles swim by producing a mechanical wave that is transmitted along the body. As illustrated in (a), during forward locomotion there is a lag between consecutive segments in the spinal cord. This lag is always a certain proportion of the cycle duration and is therefore referred to as a constant phase lag. During backward locomotion, the wave is instead propagated from tail to head. As shown in (b), in the isolated spinal cord a rostrocaudal phase lag can also be produced, and thus the ability to generate a constant phase lag is inherent to the spinal cord. The pattern can be reversed if extra excitation is added to the caudal spinal cord (right). This is explained by the fact that the segments of the caudal part then get a higher excitability and generate a somewhat higher burst rate that is able to entrain the segmental networks in the rostral part of the spinal cord. The rostral segments. then, have a lower inherent burst rate and will therefore be entrained but with a certain lag.

input is fed directly into the CPG network, and if the movement does not proceed according to the plan due to external perturbations, the swim cycle becomes adapted to the external event and is either prolonged or shortened. After discussing the spinal coordination in the other classes of vertebrates next, we then discuss the intrinsic function of the CPGs and the sensory input.

Amphibians and Reptiles

In amphibians and reptiles, the spinal cord can also generate the motor pattern after a spinal transection. The swimming movements in the tadpoles as well as the adult salamander are characterized by an undulatory wave transmitted along the body in a way similar to that of fish. A CPG for swimming has been demonstrated also in this case, as has sensory modulation. During swimming, the limbs are kept close to the body and do not take part in generating the propulsive thrust.

During walking, however, the limbs are responsible for the propulsive thrust, and there is a CPG for each limb that generates a complex output pattern. The wiggling movements of the trunk are part of the walking locomotor synergy in newt and reptiles, and they serve to prolong the step during the swing phase of each limb. To achieve this, the trunk coordination is modified so that it is linked to the step cycle, and instead of the traveling wave used during swimming, the pattern of coordination is that of a standing wave linked to each step cycle.

During development, the tadpoles of frogs, toads, and newts are all very similar and swim through an undulatory wave transmitted along the tail. From a very fast undulatory movement, the frequency decreases as the size of the tadpole increases. Initially there is only one spike in the motor neurons per swim cycle, and later a regular burst occurs and, at the same time, there are changes in the properties of the neurons within the CPG. Later, when the limbs start to develop, their motor activity becomes linked to the tail-muscle activity with synchronous activity in the flexors and extensors, and subsequently a progressive change takes place so that the limb CPG gradually becomes independent. Frogs and toads later lose their tail, whereas the newt maintains it. After losing the tail, frogs acquire a jumping locomotor coordination in which the two hindlimbs are active in-phase – the in-phase relation between the two hindlimbs is, in principle, similar to that occurring in a mammalian gallop. Frogs, however, are also able to generate alternating movements between the two hindlimbs as they move slowly, as do toads.

Birds

As has long been known, spinal birds can generate the walking synergy if the lateral stability is assisted in some way. The synchronous movements of the wings can also be produced by spinal networks – in principle, similar to the in phase coordination occurring during a gallop or a frog jump. Also, the spinal walking and flying networks are normally controlled from the mesencephalic locomotor command region.

The progressive development of the different patterns of motor coordination has been studied in the chick. It is a classic experimental model and very experimentally amenable because the nervous system is accessible in the egg at different developmental stages and can be manipulated in different ways. As in other vertebrates, the isolated spinal cord can be made to generate the different spinal motor patterns.

Mammals (including Humans) – Spinal Generation, Central Pattern Generator Networks, and Sensory Feedback

Although early studies by Philippson 100 years ago demonstrated walking movements in spinal dogs, these results did not impact the neuroscience community for many decades. It was only accepted when it was demonstrated with modern techniques in the 1970s and 1980s that the spinal cord contained the necessary motor programs. This required simultaneous recordings of locomotor movements performed on the treadmill (Figure 2(a)) with different speeds and the timing of the different muscles with electromyography (EMG). These experiments established that the spinal animals are able to generate the same motor pattern as intact walking animals. It is now generally accepted that the motor programs for walking is present in the spinal cord of mammals, including primates.

Locomotor activity can be generated directly after a spinal transection, if the CPG is activated by pharmacological means (e.g., by noradrenergic agonists). However, after a complete transection, the movements of a spinal adult animal take time to reappear (without any further pharmacological manipulation). For the recovery of locomotor function, it has proved to be important to regularly expose the spinal animal to limb movements similar to walking, for instance, on the treadmill. Chronic spinal mammals can, after extensive training, be made to perform well-controlled walking movements on the treadmill, and the range and vigor of the movements can be influenced by sensory stimuli that affect the CPG excitability. Pharmacologically, particularly 5-HT agonists have proven to be effective in facilitating the locomotor movements in most vertebrates. Also in humans with a complete spinal cord lesion, leaving the spinal CPGs intact, treadmill training with appropriate postural support can uncover a spinally generated locomotor pattern.

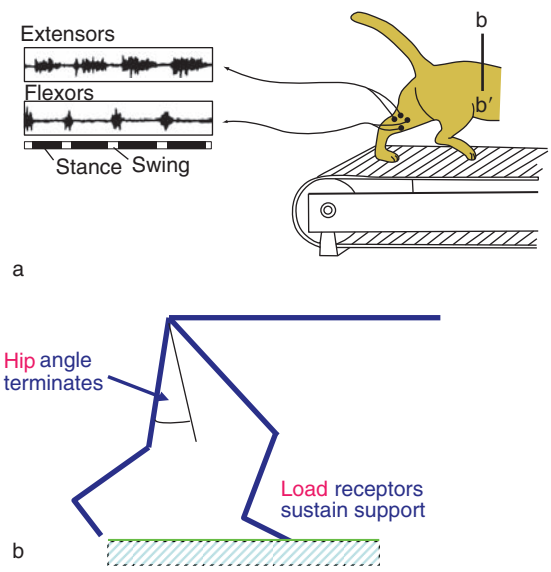

Figure 2 Locomotor movements in a spinal animal and sensory control mechanisms: (a) spinal cat (lesion at lower thoracic level) walking on a treadmill; (b) representation of the two main sensory mechanisms that help regulate the duration of the step cycle, a position-dependent (extension of the proximal joint (hip)) mechanism and a load-dependent mechanism. As shown in (a), both movements and electromyogrphical activity can be recorded simultaneously and have established a motor pattern virtually identical to that of the intact cat walking at the same speed Panel (a) modified from Pearson KG (2000) Neural adaptation in the generation of rhythmic behavior. *Annual Review of Physiology* 62: 723–753.

That a CPG is responsible for generating a complex output pattern to the limbs and also the muscles of the trunk is now well established. Graham Brown, in the early twentieth century, demonstrated that alternating muscle contractions could be produced in a pair of antagonistic muscles without the contribution of sensory afferents or in the form of a reflex. However, a much more detailed analyses and experimentation was required before these results were accepted in the research community.

From the fact that spinal animals can adapt to the treadmill speed, it follows that sensory input has access to the CPG network. In mammals, this has been studied particularly in the cat. Two major sources for the sensory regulation have been defined (Figure 2(b)). The switch from the support to swing phase is influenced by two main types of sensory signals from the limb: (1) load receptors (as long as there is a significant load on the limb, the transition from the stance to the swing phase is prevented) and (2) hip extension angle (the more the hip extension (posterior position of the leg movement trajectory), the easier it is to trigger the transition to flexion, and, conversely, if the limb is prevented from extending, the transition is also blocked). These two input sources appear functionally very meaningful; they constitute an integral part of the control of each step cycle and play a major role in adapting the movement pattern to external perturbations.

Intrinsic Function of the Central Pattern Generator Network

Lower Vertebrates

The intrinsic function of the segmental CPG of lower vertebrates has been analyzed particularly in the lamprey and in the developing amphibian tadpole; it is now well understood with regard to the main synaptic, cellular, and molecular mechanisms. These networks are arranged on the segmental level with reciprocal

inhibition between the left and right sides. Figure 3(a) depicts the core of the spinal network in the lamprey, consisting of excitatory interneurons (EIN) that excite one another within the pool and at the same time activate commissural inhibitory interneurons that inhibit the contralateral CPG. The network is turned on from the brain stem (command–drive signal). Figure 3 (c) illustrates several cellular and synaptic mechanisms that contribute throughout a cycle. The initial depolarization is thus due to, (1) the excitatory drive from the brain stem, (2) excitation from other EINs, and finally (3) the boosting of the depolarization by voltage-dependent channels such as *N*-methyl-D-aspartate (NMDA) and low-voltage-activated (LVA) Ca^{2+} channels. In turn the Ca^{2+} entry during the depolarization leads to a progressive activation of K_{Ca^+} channels during the depolarization, which leads to a progressive hyperpolarization and a closure of NMDA channels and, finally, a termination of the burst.

The ability to generate burst activity is distributed along the spinal cord, and if isolated, each segment, and even a hemisegment, can generate rhythmic burst activity. In the lamprey, the axons of EINs are asymmetric, with a descending branch extending over five to six segments and a rostral branch extending roughly three segments. There may be around 40 EINs/segment, and they are estimated to interact with only approximately 10% of the EINs within

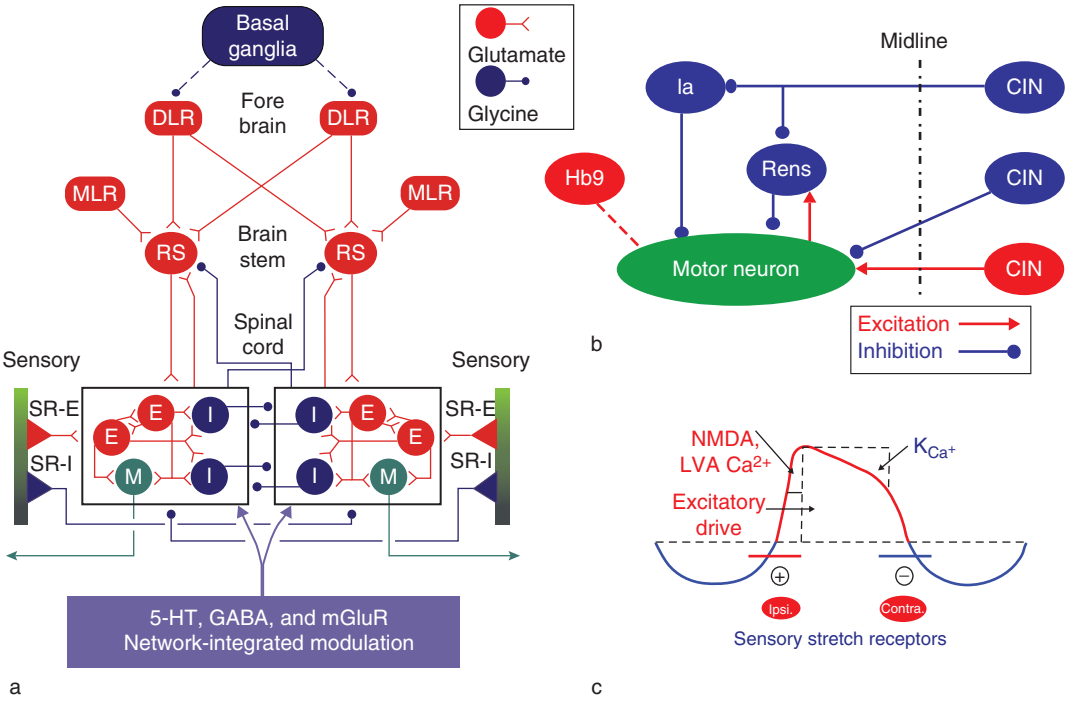

Figure 3 Neuronal elements: (a) lamprey CPG; (b) mammalian hindlimb CPG; (c) factors that contribute to the initiation of the depolarizing phase, its maintenance, and its termination (lamprey locomotor CPG). Shown in (a) is a representation of the forebrain, brain stem, and spinal components of the lamprey neural circuitry that generates rhythmic locomotor activity. All neuron symbols denote populations rather than single cells. The reticulospinal (RS) glutamatergic neurons excite all classes of spinal interneurons and motor neurons (M). The excitatory interneurons (E) excite all types of spinal neurons, that is, the inhibitory glycinergic interneurons (I) that cross the midline to inhibit all neuron types on the contralateral side and motor neurons. The stretch receptor neurons are of two types: one excitatory (SR-E), which excites ipsilateral neurons, and one inhibitory (SR-I), which crosses the midline to inhibit contralateral neurons. RS neurons receive excitatory synaptic input, the diencephalic and the mesencephalic locomotor regions (DLR and MLR), which in turn receives input from the basal ganglia as well as visual and olfactory input. In addition, metabotropic receptors are also activated during locomotion and are an integral part of the network (5-HT, GABA, and mGluR). As shown in (b), ipsilateral inhibitory premotor interneurons (1a and Renshaw (Rens)) are known to be rhythmically active during locomotion. They receive input also from contralaterally projecting interneurons. Motor neurons receive input directly from both excitatory and inhibitory contralateral interneurons. One ipsilateral glutamatergic locomotor-related interneuron (Hb9) has also been identified. In (c), in addition to conventional synaptic excitation, voltage-dependent NMDA receptors and low-voltage-activated (LVA) Ca^{2+} channels and Na^+ channels may be activated. Ca^{2+} will enter the cell through these channels, cause the activation of K_{Ca^+}, and thereby cause a progressive hyperpolarization leading to the closure of the NMDA channels. In the behaving animal, the initiation of the depolarizing phase is facilitated by the activation of ipsilateral SR-Es, whereas the termination of the depolarized phase is partially a result of activation of contralateral SR-Is. 5-HT, serotonin; CPG, central pattern generator network; E, excitatory; GABA, γ-aminobutyric acid; I, inhibitory interneuron; mGluR, metabotropic glutamate receptor; NMDA, *N*-methyl-D-aspartate.

the field of innervation. In this way, a continuous rostro-caudal network along the spinal cord is formed.

The spinal cord CPG circuits that generate the undulatory wave that results in swimming in fish, lampreys, and amphibians can, in a simplified way, be regarded as a series of segmental oscillator circuits, as already discussed. Normally, the rostral segments are activated first, and the other segments follow from head to tail-fin, with a fixed phase lag of approximately 1% of the cycle duration (lamprey), whether the swimming frequency is low or high (Figure 1). If the opposite situation is induced with a higher relative excitability in the caudal segments (by inhibition of the rostral segments or additional excitation of the most caudal segments) the direction of the wave is reversed and goes from tail to head instead, and backward swimming results, which is required behaviorally if a lamprey gets caught in a corner. In the isolated spinal cord, forward and backward coordination with a constant phase coupling can be generated experimentally in a similar fashion.

How is this type of flexible phase coupling along the spinal cord achieved? We may consider, in a simplified way, a series of unit CPGs coupled in a chain with excitatory connections between the various unit CPGs in both directions. In this case, the unit CPG with the highest inherent frequency becomes the leader, and the other unit CPGs are pulled along – the trailing one is the one with the lowest excitability.

Mammalian Locomotor Central Pattern Generator Networks

The elucidation of the mammalian locomotor CPG has been more demanding than that of the tadpole or the lamprey. In those two cases, a symmetric alternation between the left and right sides generated by pools of excitatory and crossed inhibitory interneurons suffices to generate the segmental motor pattern, whereas mammalian locomotion is dependent on a more complex pattern of muscle activity in each limb, which is coordinated by the limb CPG.

The circuits controlling locomotion resides in the ventral spinal cord because the locomotor pattern can be induced in the isolated rodent spinal cord, even when the dorsal aspect has been removed, and also in each hemicord because the motor pattern of a limb can be produced after the isolation of each side. Moreover, the CPG circuitry of each limb appears to be distributed along the spinal cord, as in the lamprey, although the more rostral segments controlling the proximal hip joint seem to have a more powerful role in entraining the circuits controlling the knee, ankle, and foot joints. In order to elicit alternating flexor and extensor activity, inhibitory mechanisms are, of course, required. After the inhibition has been blocked pharmacologically, burst activity commences, and therefore it appears that, also in this case (in accordance with the findings from lower vertebrates), there is a core excitatory burst-generating circuit that does not require inhibition to generate burst activity.

Considerable progress has come from the development of an *in vitro* preparation of the neonatal rodent brain stem–spinal cord. The details of the CPG are still not clear, but different commissural excitatory and inhibitory premotor interneurons have been identified (Figure 3(b)). Developmental findings in the mouse have led to the genetic delineation of subtypes of precursor interneurons (V0–V3) that give rise to some of the classic types of spinal interneurons. V0 and V3 give rise to neurons that project to the contralateral side, whereas V1 and V2 are ipsilateral. Among the V1 interneurons are the inhibitory 1a and Renshaw premotor interneurons which are active during locomotor activity. The V2 group contains both inhibitory and EINs. These genetic markers have opened up the possibility of expressing different fluorescent probes (e.g., green fluorescent proteins) or other molecules such as receptors in particular subtypes of interneurons. The combination of developmental, molecular, and neurophysiological techniques promises to yield rapid progress in the understanding of the neuronal bases of motor coordination in mammals.

Systems of Central Pattern Generator Networks – Modular Organization

Interaction between the Limb Central Pattern Generator Networks during Various Gaits

Each limb is controlled by a separate spinal CPG that provides a standard pattern of activation of the muscles of that limb (adaptable with regard to support phase duration). Between two limb CPGs acting at the same girdle, there are two preferred modes, alternation and in-phase (approximately). Alternation, as in walking or trotting, is in general produced by reciprocal inhibition between the two CPGs – as a rule due to glycinergic inhibition mediated by commissural interneurons. Mutual excitation between the two CPGs instead results in that their activity becomes synchronous, as during a gallop. The four limbs are, finally, coordinated in different ways during walking, pacing, trotting, and galloping. Each limb CPG is thus part of a system of interacting limb CPGs (Figure 4(a)) that can be coupled in a few stable modes.

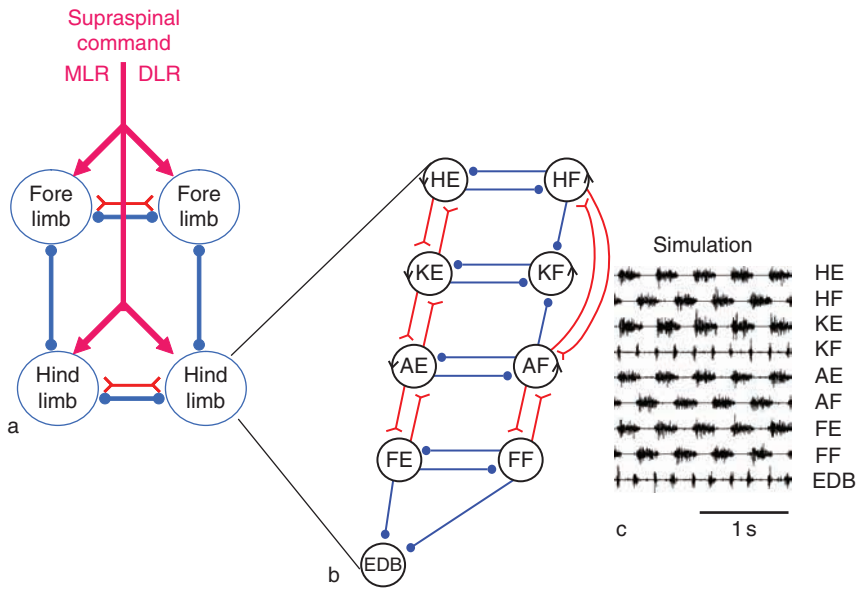

Figure 4 Systems of interacting CPGs, inter- and intralimb coordination: (a) command regions in the brain stem; (b) further subdivision in unit CPGs; (c) exploratory simulation of locomotor activity. In (a) the command regions in the brain stem (the mesencephalic locomotor region (MLR) and the diencephalic locomotor region (DLR)) control the level of activity in the CPG networks for each of the four limbs. Each limb CPG is responsible for the complex motor pattern of that limb, whereas the interlimb coordination, as in walking or galloping, results from the interaction between the different limb CPGs. The interaction takes place through inhibitory reciprocal interaction (blue inhibitory) during walking and excitatory mutual excitation during galloping (red). As shown in (b) within each limb CPG there is most likely a further subdivision in unit CPGs controlling the synergists at one joint, such as the hip extensor (HE), knee extensor (KE), ankle extensor (AE), foot extensor (FE), and ankle flexor (AF). Extensor digitorum brevis (EDB) has a particular pattern. The normal pattern of activity results from the interaction between the different unit CPGs at different joints. The advantage is that the unit CPGs could be recombined as in backward or forward walking, in the same way as the different limb CPGs, which can be recombined in the different gaits. Circles indicate inhibition, and forks/triangles indicate excitation. In (c), the network arranged so each unit CPG is designed in a way similar to the lamprey unit CPGs, consisting of 100 excitatory interacting neurons, and the interaction among the nine-unit CPGs arranged as in (b). The output of the activity captures essential features of the locomotor output. CPG, central pattern generator netowork; DLR, diencephalic locomotor region; MLR, mesencephalic locomotor region.

Is there Flexibility within Each Limb Central Pattern Generator Network? – Walking Forward and Backward

The flexibility of the locomotor behavior is considerable, we can thus walk forward, backward, sideways, and even on our knees. Do we need to have a separate limb CPG for each direction of movement? With a hard-wired system for the entire limb, this would actually be the case. Most likely the limb CPG is also composed of a system of limb CPGs. Indirect evidence suggests that each limb CPG can be further subdivided into smaller functional modules, which can be recombined to generate a varied and flexible motor output. Each limb CPG is thought to be composed of a series of unit CPGs, each controlling a group of close synergists such as hip flexors or ankle extensors. Such a system adds flexibility, achieved by changing the mode of interaction between the different unit CPGs, in much the same way as between the limb CPGs in the different gaits.

The conceptual scheme presented in Figure 4(b) proposes that there are unit CPGs for each group of close synergists at a given joint. The model assumes that each unit CPG can burst by itself but that the final pattern is due to three factors: (1) the reciprocal interaction between antagonists (flexors and extensors) at the same joint, (2) mutual excitation between unit CPGs at different joints that are active together, and (3) the possibilities for more complex patterns, with muscles being active in the interval between the main flexors and extensors. To test this conceptual scheme in an exploratory fashion, we made use of the segmental unit CPGs defined in the lamprey system and plugged them in as unit CPGs in the limb system. With no further modification, we simulated the motor pattern of the hindlimb and also the complex pattern of activity of the different muscles of the limb (Figure 4(c)). Although the limb CPG most likely is more complex, it is thought provoking that a distributed circuit of this type can reproduce the motor pattern and add the flexibility of reconfiguring the network to other patterns of coordination. To account for this type of requirement for flexibility, another model has also been considered in which a central rhythmogenic circuit is thought to pace

the rhythm and alternative interneuronal output channels to different motor nuclei provide the necessary flexibility.

Scratching and Wiping Reflexes

The scratch reflex provides another interesting example of a spinal rhythmic precision movement that is present in mammals and other vertebrates with limbs. The objective is to remove an irritant somewhere on the body surface through a limb movement targeting this particular location. It is often elicited by a hindlimb, but it can also be performed with a forelimb. It can reach out to all points on the body surface that are anatomically possible. It is mostly performed as a rhythmic motor act involving rhythmic movements centered around, particularly, the skin area in which the irritant is located.

Well-coordinated scratch reflexes toward a particular skin area (called wiping reflexes in the frog) can be elicited also after a spinal transection. Because the irritant can be located on widely different skin regions, it also means that the motor patterns must be flexible and differ radically between, for instance, an irritant at the head and the lower back. These flexible patterns have been particularly well analyzed in the turtle by Paul Stein. The skin irritant, say, in the cervical area, must provide a signal to the spinal cord which informs about the location of the irritant (proximal or distal location in a given segment). This segmental input must then, in turn, be able to activate the lumbosacral scratch CPG and provide instructions to produce a scratching movement toward the particular target area. A high degree of flexibility is thus required. The turtle scratch reflex is organized with modular units that can be recombined, as already described for the unit CPG organization in the context of the locomotor CPGs. This organization with discrete unit CPGs that can be recombined in different ways also allows for the possibility that they may be integrated into other, widely different patterns of motor behavior.

The spinal cord itself is thus able to provide positional information about the skin irritant on the body surface. If the irritant is located on the distal part of a limb, the orientation of the limb also needs to be considered (pointing forward or backward in relation to the body). The spinal cord at the stimulus segment thus has the astounding ability to direct the movement to the irritant regardless of the skin area stimulated and the orientation of the appendage in relation to the body. This clearly indicates that the spinal cord itself contains information about the spatial coordinates of the body in relation to the trunk and that this information can be used to design the appropriate motor command.

Posture, Static Limb Positions, and Protective Reflexes

Also with posture and static limb position, there is evidence to suggest a subdivision into spinal modules controlling groups of synergists in the lumbosacral spinal cord. Co-activation of different modules produces different positions of the limb.

The flexion and crossed extension reflexes, in Sherringtonian terms, was for a long period considered to be standard example of a behaviorally meaningful integrated spinal reflex. It now turns out that it, instead, represents a family of very precise reflex modules that removes the skin area stimulated by the stimulus. Different combinations of muscles are required depending on the location of the stimulus (e.g., on the lateral or medial part of the foot). A set of different interneuronal modules can account for the family of different site-specific protective reflexes in the limb.

Modular Organization Appears to Be a General Organizational Principle

This type of modular organization appears to apply to a variety of systems, both in vertebrates and invertebrates. It seems likely that these different spinal modules – these integrated circuits – could also be recruited during, for instance, voluntary limb movements or when adapting the human locomotor synergy to other gross movements such as skiing or bicycling.

In Conclusion

The spinal cord contains a number of preformed circuits that can coordinate a variety of movements. The basic design features appear to be evolutionary conserved from cyclostomes to primates, although the fine details may differ. Many aspects of our movement repertoire are, thus, produced through a recruitment of different spinal motor circuits and through the combination of different spinal networks.

See also: Central Pattern Generators: Sensory Feedback; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Motor Primitives; Reflex Circuits; Sensorimotor Plasticity and Control of Movement Following Spinal Cord Injury +D18; Spinal Motor Neurons: Properties; Vestibulospinal System and Eye–Head/Neck Movement.

Further Reading

Earhart GM and Stein PS (2000) Step, swim, and scratch motor patterns in the turtle. *Journal of Neurophysiology* 84(5): 2181–2190.

El Manira A and Wallen P (2000) Mechanisms of modulation of a neural network. *News in Physiological Science* 15: 186–191.

Grillner S (1985) Neurobiological bases of rhythmic motor acts in vertebrates. *Science* 228(4696): 143–149.

Grillner S (1996) Neural networks for vertebrate locomotion. *Scientific American* 274(1): 64–69.

Grillner S (2003) The motor infrastructure: From ion channels to neuronal networks. *Nature Reviews Neuroscience* 4(7): 573–586.

Grillner S (2006) Biological pattern generation: The cellular and computational logic of networks in motion. *Neuron* 52(5): 751–766.

Kiehn O (2006) Locomotor circuits in the mammalian spinal cord. *Annual Review of Neuroscience* 29: 279–306.

Pearson KG (2000) Neural adaptation in the generation of rhythmic behavior. *Annual Review of Physiology* 62: 723–753.

Rossignol S, Dubuc R, and Gossard JP (2006) Dynamic sensorimotor interactions in locomotion. *Physiological Review* 86(1): 89–154.

Schmidt BJ and Jordan LM (2000) The role of serotonin in reflex modulation and locomotor rhythm production in the mammalian spinal cord. *Brain Research Bulletin* 53(5): 689–710.

Sillar KT, McLean DL, Fischer H, and Merrywest SD (2002) Fast inhibitory synapses: Targets for neuromodulation and development of vertebrate motor behaviour. *Brain Research Brain Research Reviews* 40(1–3): 130–140.

Peptidergic Receptors

G Casini and D Cervia, Università della Tuscia, Viterbo, Italy

© 2009 Elsevier Ltd. All rights reserved.

Peptides and Intercellular Communication

Neuropeptides are peptides involved in nervous system function. They are synthesized in cells in large precursor proteins, and generally several biologically active peptides are contained in the same precursor molecule. In the 1970s the hunt for new peptides was intense, and an ever-growing list could be compiled. Peptidergic signaling molecules are widely distributed throughout the central and peripheral nervous systems and in the peripheral organs. They may act in a neurocrine, paracrine, autocrine, or endocrine manner; the same peptides may participate in intercellular communications through different modalities. For instance, many peptides expressed in the enteric system belong to a class of agents known as brain–gut peptides, functioning both as neuropeptides and gut hormones.

Neuropeptides may be co-stored with other neuropeptides or, alternatively, may coexist with 'classical' neurotransmitters in different cellular compartments. Different combinations of transmitters and peptides are found in ganglionic cells of the autonomic nervous system. For instance, the sympathetic postganglionic input to the heart arises in a large part from the stellate ganglion; these neurons commonly contain neuropeptide Y (NPY) in addition to norepinephrine, similar to almost all sympathetic postganglionic neurons innervating the vasculature of the gut and hindlimbs. Other sympathetic postganglionic neurons in the stellate ganglion project to the sweat glands using acetylcholine as their major transmitter (instead of the expected norepinephrine), and also contain at least two other peptides, that is, calcitonin gene-related peptide (CGRP), and vasoactive intestinal polypeptide (VIP). In addition, many postganglionic parasympathetic neurons contain acetylcholine and VIP. Thus, while autonomic neurons are commonly referred to as cholinergic or adrenergic, they may also contain permutations of other neuropeptide transmitters including angiotensin II (Ang II), enkephalin, neurotensin, substance P (SP), somatostatin, CGRP, etc. Functionally, specific permutations of transmitters may comprise a 'chemical code' for neurons subserving specific actions. Co-stored peptides and classical transmitters are released together at all terminals of a given neuron, and they act together in determining the responses of target cells. It is a general rule that when a peptide and a classical transmitter coexist, the first mediates long-lasting responses and the latter short-term synaptic events in the target cells.

Peptidergic Receptors

Independent of the modality, peptides act through specific receptors which are generally located on the plasma membrane of target cells, although there is growing evidence that certain receptors are also present on the nuclear membrane. Peptidergic receptors belong to the superfamily of heterotrimeric G-protein-coupled receptors (GPCRs), which are characterized by the presence of seven transmembrane domains. The specialized feature of these receptors is to achieve information transfer via the binding of ligands to a recognition domain and allosterically transmit the presence of that ligand to an intracellular domain that leads to G-protein activation. The human genome may code for more than 5000 GPCRs. Less than 10% are known, of which those activated by peptides represent a large proportion. In particular, mammalian peptidergic receptors are grouped into two families of GPCRs: the rhodopsin-like family (the largest family of GPCRs, also called class A), which contains the majority of the known peptidergic receptors, and the secretin-like family (class B). The ligand peptides for the receptors in the secretin-like family are relatively large polypeptides (27–141-amino-acid residues), and seven of them – glucagon, glucagon-like peptides 1 and 2, gastric inhibitory polypeptide, VIP, pituitary adenylate cyclase-activating peptide (PACAP), and growth hormone-releasing hormone – are structurally related.

It has been generally assumed that peptides exert their actions on GPCRs, but there are some exceptions. For instance, FMRFamide (Phe-Met-Arg-Phe-NH_2) and related peptides typically exert their action through GPCRs. However, two ionotropic receptors for these peptides have recently been identified which may have functional roles in the central and the peripheral nervous systems. Another example is one of the neurotensin receptors, which is a single transmembrane domain protein that belongs to a recently identified family of sorting receptors.

Although there is at least one receptor cloned for almost all of the peptides discovered so far, it still remains to be elucidated how many receptors can be

found for each of the peptides. In many cases, it does not seem unreasonable to assume that there are several receptors for one peptide. However, their number may not approach those described for the receptors of some classic transmitters, such as the 15 or so receptors for serotonin. At present, many 'orphan peptides' (for which there are no known receptors) still exist, and many others may be discovered soon. Recently, MrgX2 receptor has been identified as the cortistatin-specific receptor. In contrast, some peptides have been found via one of the many 'orphan receptors,' that is, receptors for which the endogenous ligand is unknown. For instance, orexin-A and -B were initially identified as endogenous ligands for an orphan GPCR. In addition, the endogenous ligand for the growth hormone secretagogue receptor was recently discovered and named ghrelin, while a study focusing on another orphan receptor, the glucocorticoid receptor 3, revealed that a prolactin-releasing peptide in the brain represents an endogenous ligand for this receptor. Also, many naturally occurring peptides exhibit a high degree of promiscuity across GPCRs. The physiological significance of this phenomenon and the degree to which it occurs is not well characterized, but it does not seem an exception in the nervous system. For instance, cortistatin binds to somatostatin receptors with similar high affinity to that of somatostatin itself and it may produce physiological effects similar to those of somatostatin. In addition, the orphan peptide histidine methionine-27 was recently shown to be a potent agonist at the human calcitonin receptors.

Functionally, there is a general agreement that peptidergic receptors are coupled to multiple components of transduction pathways including adenylyl cyclase/cyclic AMP, phospholipase C/diacylglycerol/inositol triphosphate, phospholipase A2/arachidonic acid, guanylyl cyclase, protein kinases and phosphatases, ionic channels and exchanger, reactive oxygen species, metalloproteases, transcription factors, and other intracellular cytoplasmic and nuclear regulators. The diversity of the transduction pathways reflects the pleiotropic actions of peptides. The prevailing mechanisms in any given cell depend on many factors, including receptor distribution and the signaling elements expressed by the cell. Often the situation is further complicated by cross-talk between different pathways. In addition, at the cellular level, many mechanisms are involved in the amplification effects, which occur in the signaling cascade. In particular, posttranslational modification of receptors, different expression of factors interacting with G-proteins, and differences in G-protein repertoire and signaling isoenzymes may affect function in a process which is generally referred to as cross-talk and receptor trafficking. On the other hand, recent results suggest the existence of multiple ligand-specific receptor conformations, which may differentially couple to the intracellular pathways within the signal transduction network. Indeed, peptidergic receptor reserve, coupling efficiency, and probably receptor trafficking may not only depend on transduction mechanisms, but also on the nature of the ligand and the receptor at play. In addition, the type and the magnitude of the response elicited by a ligand is also dictated by the level of receptor available at the plasma membrane.

Endocytosis and recycling of receptors are important for desensitization and resensitization in response to endogenous ligands and drugs. Recent studies of peptidergic receptors have provided exciting insights into specific mechanisms that control endocytosis of receptors from the plasma membrane (internalization) and regulate proteolytic degradation of receptors. For instance, the internalization of the AT_1 receptor (a specific receptor for Ang II) occurs via clathrin-coated pits, but there is evidence that, in contrast to the internalization of other GPCRs, it is independent of dynamin and beta-arrestin. In addition, AT_1 receptor internalization requires two regions in the cytoplasmic tail of the receptor, but it is independent of G-protein activation. The dependence of AT_1 receptor internalization on the presence of a serine–threonine-rich region suggests that phosphorylation of the receptor tail may regulate the internalization process.

GPCRs have been found to be dysfunctional/dysregulated in a growing number of human diseases and the property of 'constitutive activity' which is associated with GPCRs provides a persistent tonic mechanism to fine-tune synaptic transmission during both acute and chronic information transfer. In this respect, GPCRs have been a very successful target for drug discovery and development and almost a third of currently available prescription drugs function as GPCR ligands. In particular, the clinical potential of neuropeptides is well known and, needless to say, the multiplicity of peptidergic receptors offers unique and important openings for the development of specific drugs, which may be useful for treatment of various disorders. In fact, receptor-subtype-selective antagonists and agonists have been developed, and recently an SP-receptor antagonist has been shown to have clinical efficacy in the treatment of major depression and chemotherapy-induced emesis. In addition, the use of somatostatin and its analogs is well established for certain indications, such as acute esophageal variceal bleedings, gastrointestinal disorders, neuroendocrine tumors, and acromegaly as well as the diagnosis and radiotherapy of some tumors. Several other neuropeptide receptor ligands are on clinical trials for various indications.

Peptidergic Receptors in the Autonomic Nervous System and in Peripheral Organs

Multiple peptides and receptor subtypes are expressed in the autonomic nervous system and in peripheral organs, where they are involved in a variety of functions (Table 1). In this section, data from recent scientific reports and review articles have been collected to provide examples of the occurrence and functional actions of peptides and their receptors in the autonomic nervous system and peripheral organs.

Modulation of Parasympathetic Neurons

The submandibular ganglion (SMG) is a parasympathetic ganglion, which receives inputs from preganglionic cholinergic neurons and innervates the submandibular salivary gland to control saliva secretion. Neurotransmitters and neuropeptides acting via GPCRs change the functional activity of neurons in this ganglion by modulating voltage-dependent Ca^{2+} channels.

Opioid peptides Endogenous opioid peptides and their receptors are integral components of neuronal circuits modulating pain transmission. They also produce various effects, such as analgesia, inhibition of diarrhea, respiratory depression, and catalepsy. Opioid peptides act at three distinct types of opioid receptors, namely, μ-, δ-, and κ-opioid receptors. Met-enkephalin causes hyperpolarization, while μ-, δ-, and κ-opioid receptor agonists inhibit currents in voltage-dependent Ca^{2+} channels in a dose-dependent manner. Similar actions of opioid peptides have been reported using different receptor subtype-specific agonists in rat dorsal root ganglion neurons, rat parasympathetic intracardiac neurons, neurons of the rat nodose ganglion, and avian ciliary ganglion.

Tachykinin peptides Three types of tachykinin receptors, neurokinin (NK)-1, NK-2, and NK-3, have been cloned and characterized pharmacologically. They preferentially bind the endogenous tachykinins SP, neurokinin A (NKA), and neurokinin B, respectively. SP is a well-known pain-related peptide synthesized and secreted from primary afferent neurons. SP and NKA are reported as neurotransmitters of the SMG neurons. SP, acting at NK-1 receptors, causes depolarization of SMG neurons by inhibiting M-type K^+ channels. In addition, an NK-1 receptor agonist inhibits voltage-dependent Ca^{2+} channel currents. SP also facilitates ganglionic transmission in sympathetic

Table 1 Examples of the main effects mediated by some of the peptidergic receptors expressed in sympathetic neurons or in sympathoeffector cells

Neuropeptides	*Receptor subtype*	*Site of expression*	*Effect*
Calcitonin gene-related peptide	CGRP1	Postganglionic vasomotor neurons	Vasodilatation
Cholecystokinin	CCK1	Preganglionic neurons of rostral ventrolateral medulla	Inhibition of postganglionic neuron activity
	CCK2	Preganglionic vasomotor neurons	Increase of postganglionic neuron activity
Endocannabinoids	CB1	Preganglionic nerve terminals	Inhibition of transmitter release
Endothelin	ETA	Postganglionic nerve terminals of vasculature/vascular smooth muscles	Increase of vasoconstrictor transmitter release/inhibition of vasodilatator transmitters
	ETB	Postganglionic neurons/postganglionic nerve terminals of vasculature	Increase of oxidative stress/increase of vasoconstrictor transmitter release
Enkephalins	Opioid δ and κ	Postganglionic neurons	Inhibition of transmitter release
Neuropeptide Y	Y1	Vascular smooth muscles	Vasoconstriction
	Y2	Postganglionic nerve terminals of vasculature	Inhibition of vasoconstrictor transmitter release
Somatostatin	sst2 and sst3	Postganglionic nerve terminals	Inhibition of transmitter release
Tachykinins	NK1	Postganglionic neurons of superior cervical ganglion/spinal preganglionic neurons and postganglionic axons of vasculature/vascular smooth muscles	Increase of neuron excitability/ increase of vasoconstrictor transmitter release/vasodilatation
	NK2	Postganglionic neurons in gastrointestinal tract	Smooth muscle relaxation
Vasoactive intestinal polypeptide/pituitary adenylate cyclase activating peptide	PAC1	Postganglionic neurons of superior cervical ganglion	Increase of transmitter release
	VPAC1 and VPAC2	Heart and vascular smooth muscles	Vasodilatation

ganglia, while it inhibits currents in voltage-dependent Ca^{2+} channels in a variety of neuronal cells.

CGRP CGRP is generated through an alternate splicing of the calcitonin gene primary transcript. CGRP receptors are heterodimeric complexes composed of the heptahelical calcitonin receptor-like receptor and the single transmembrane receptor activity-modifying protein 1. They are widely distributed throughout the nervous system and have been studied in several species. Similar to SP, in SMG neurons, CGRP causes depolarization by inhibition of voltage-insensitive K^+ channels. In addition, CGRP facilitates voltage-dependent Ca^{2+} channels.

Ang II Ang II has various physiological effects, including stimulation of increased blood pressure, water and sodium intake, vasopressin secretion, and modulation of baroreflex function. In addition, Ang II acts as a neurotransmitter at two types of receptors (AT_1 and AT_2 receptors). In SMG neurons, Ang II causes depolarization by the combination of an increase in Na^+ conductance and a decrease of K^+ conductance. On the other hand, Ang II inhibits voltage-dependent Ca^{2+} channel currents via AT_1 with transduction mechanisms involving protein kinase C.

VIP and PACAP VIP was first isolated from porcine intestine as a peptide capable of inducing vasodilatation. PACAP was initially isolated from ovine hypothalamus for its ability to stimulate pituitary cell membrane adenylyl cyclase and was further shown to exist in two biologically active forms of 27 and 38 amino acids. At least three receptors for these peptides have been identified in mammals: $VPAC_1$, $VPAC_2$, and PAC_1. $VPAC_1$ and $VPAC_2$ respond to VIP and PACAP with comparable affinity, while PAC_1 binds PACAP with high affinity and VIP with much lower affinity. VIP coexists with acetylcholine in postganglionic SMG neurons, and following stimulation of parasympathetic nerve, it is released from the submandibular gland together with acetylcholine. In SMG neurons, VIP causes depolarization by inhibition of voltage-insensitive K^+ channels. Both VIP and PACAP inhibit voltage-dependent Ca^{2+} channel currents. The inhibition mediated by $VPAC_1$/$VPAC_2$ receptors involves both protein kinase A and protein kinase C, while the PAC_1 receptors inhibit the voltage-dependent Ca^{2+} channel via the $\beta\gamma$ subunits of a Gs-protein. VIP and PACAP may also facilitate voltage-dependent Ca^{2+} channels; however, VIP- and PACAP-induced facilitation is less frequently reported than inhibition.

Functional implications A variety of neuropeptides and their receptors act in SMG neurons. However, such a diversity of signaling substances would presumably not be necessary to only signal neuronal excitation or inhibition. The voltage-dependent Ca^{2+} channels are essential in the regulation of many neuronal processes including membrane excitability, enzyme activity, gene expression, neuronal outgrowth, neuronal regeneration, differentiation, and proliferation. The G-protein-dependent signals elicited by the action of multiple peptidergic receptors and their interactions may provide the necessary versatility to properly accomplish those neuronal functions.

Presynaptic Modulation at Sympathoeffector Junctions

The transmitter release at sympathetic varicosities is modulated by a plethora of different signaling molecules, including classical transmitters, ATP, histamine, prostanoids, and a variety of peptides. Presynaptic receptors located on the sympathetic terminals are activated by transmitters that are released either from the same nerve ending or from a different axon terminal, and the receptors involved are called auto- or heteroreceptors, respectively. The peptidergic receptors involved in this regulation include receptors for NPY as autoreceptors and for Ang II, bradykinin, cannabinoid peptides, endothelin, natriuretic peptide, opioid peptides, somatostatin, and VIP/PACAP as heteroreceptors.

NPY Y_2 receptors NPY shares a family of common receptors with peptide YY and pancreatic polypeptide, which consists of at least five members. These are named Y_1, Y_2, and Y_4 through Y_6 receptors. Selective antagonists have been developed for Y_1 and Y_2 receptors. NPY is co-stored and co-released together with ATP and norepinephrine in most sympathetic axon terminals. In the vasculature, NPY has been found to induce contractions mediated by Y_1 receptors localized on smooth muscle cells. In some cases, postsynaptic effects of NPY may also involve Y_2 receptors. In *in vitro* preparations of noradrenergic postganglionic sympathetic neurons, NPY reduces transmitter release. This inhibitory effect of NPY is typically mimicked by Y_2 receptor agonists. In addition, there is evidence that presynaptic Y_2 receptors are indeed autoreceptors. However, there are reports that a selective Y_2 receptor antagonist has no effect on excitatory junction potentials, indicating a lack of NPY-dependent feedback modulation of sympathoeffector transmission. Thus, it remains to be established under which conditions NPY may mediate an autoinhibition of sympathetic transmitter release.

Ang II AT_1 receptors In a variety of sympathetically innervated tissues obtained from a large number of different species, a facilitation of norepinephrine release by Ang II has been reported. In addition, evidence has been provided that this stimulatory effect is mediated by AT_1 receptors. However, in some instances, presynaptic angiotensin effects have been observed to be antagonized by both AT_1 and AT_2 receptor antagonists, revealing some uncertainty concerning the subtypes of presynaptic angiotensin receptors in sympathetic axonal terminals.

Bradykinin B_2 receptors Actions of bradykinin are mediated by at least two types of receptors, B_1 and B_2, and the existence of a B_3 receptor has also been suggested. B_1 receptors are mainly restricted to the vasculature, whereas B_2 receptors have been detected in most types of tissues including the central and peripheral nervous systems. There are reports that bradykinin may modulate sympathetic transmitter release in a bimodal way, causing inhibition via B_1 and facilitation via B_2 receptors, but in most cases bradykinin enhances stimulation-evoked norepinephrine release via B_2 receptors. However, in a considerable number of investigations, these apparently presynaptic effects of bradykinin have been shown to be indirect and to involve Ang II with AT_1 receptors, CGRP, or SP.

Cannabinoid CB_1 receptors At least two types of cannabinoid receptors, CB_1 and CB_2, have been identified. Aside from their psychoactive and immunomodulatory effects, cannabinoids exert pronounced cardiovascular actions such as vasodilatation, tachycardia, and changes in blood pressure, all mediated most likely by CB_1 receptors. Because central sites of action are not involved in these effects, cannabinoids are believed to elicit vasorelaxation through peripheral mechanisms, which may include inhibition of sympathoeffector transmission. Indeed, there is experimental evidence that cannabinoids may cause sympathoinhibition through presynaptic CB_1 receptors on sympathetic nerve terminals, the activation of which reduces transmitter release.

Endothelin receptors Endothelin-1 was identified originally as an endothelium-derived vasoconstricting peptide, but the family of endothelins (endothelin-1, -2, and -3) has recently been suggested to act as neuropeptides in the central and peripheral nervous systems. The actions of endothelins are mediated by either ET_A or ET_B endothelin receptors. There are data showing presynaptic facilitatory and inhibitory and postsynaptic effects of endothelins. Thus, it appears that endothelins may modulate sympathetic neurotransmission via facilitatory and inhibitory presynaptic receptors, but the receptor subtypes involved remain to be unequivocally identified.

Natriuretic peptide receptors The family of receptors for natriuretic peptides (atrial natriuretic peptide, brain natriuretic peptide, and C-type natriuretic peptide) comprises two members, the ANP_A and the ANP_B receptors. Both are membrane receptors with one transmembrane domain and intracellular guanylyl cyclase activity. The ANP_A receptor is preferentially activated by atrial and brain natriuretic peptides, whereas C-type natriuretic peptide is the endogenous agonist for ANP_B. A role of presynaptic receptors for natriuretic peptides on sympathetic axon terminals remains to be fully elucidated, and the natriuretic peptides may rather be assumed to control the sympathetic nervous system through central sites of action.

Opioid receptors Opioid receptor agonists have long been known to modulate transmitter release in the peripheral nervous system. Evidence has been obtained that each of the three known opioid receptor subtypes may be located on sympathetic axon terminals and, upon activation, reduce transmitter release.

Somatostatin receptors Somatostatin is widely distributed in the nervous system, and five different G-protein-coupled somatostatin receptors (sst_1–sst_5) have been characterized by molecular cloning. Inhibitory effects of somatostatin on transmitter release from sympathetic terminals have been documented. In addition, this action of somatostatin may be mimicked by sst_2 or sst_3 receptor agonists; however, a definite pharmacological characterization of inhibitory presynaptic somatostatin receptors on sympathetic axonal terminals is still missing.

VIP/PACAP receptors Various forms of PACAPs and VIPs are present in and released from sympathetic nerve terminals. Nevertheless, autoregulation of sympathetic transmitter release by VIP or PACAP has not been demonstrated. Therefore, the respective receptors are considered presynaptic heteroreceptors. Presynaptic receptors for VIP and/or PACAP may either facilitate or inhibit sympathetic transmitter release, but the receptor subtypes mediating these opposite effects remain to be identified.

Regulation of Gastrointestinal Motility

The peptide cholecystokinin Cholecystokinin (CCK) is a potent regulator of gastrointestinal motility. It and the related peptide gastrin belong to the class of brain–gut peptides. The actions of CCK include stimulation

of exocrine and endocrine secretion, motility and growth in the gastrointestinal tract, and regulation of satiety, anxiety, pain, and behavior in the central and peripheral nervous systems. Alterations in CCK release and tissue responses to the peptide have been implicated in the pathogenesis of irritable bowel syndrome. There are two principal sources of CCK: endocrine cells in the duodenum and peptidergic nerves, both in the enteric and central nervous systems. In the periphery, CCK-containing neurons are found in the myenteric plexus, submucosal plexus and muscle layers of the small intestine and colon, and in the celiac plexus and the vagus nerve.

CCK receptors The biological actions of CCK are mediated by CCK receptors that are distributed throughout the gut. They include two distinct types that are the products of two distinct genes and are termed CCK_1 and CCK_2 receptors. There is identity between the CCK_2 receptor and the gastrin receptor. The CCK_1 receptor has an approximately 1000-fold greater affinity for CCK than for gastrin, while the CCK_2 receptor has the same high affinity for both CCK and gastrin. The activated G-protein leads to activation of phospholipase C, but high agonist concentrations can also activate adenylyl cyclase via a different G-protein. Although the nucleotide sequences of cloned cDNAs of CCK_1 receptors from different sources are identical, their affinity states may vary. The available evidence suggests that the CCK_1 receptor exists in both high- and low-affinity states and that CCK binding results in the initiation of different intracellular events and consequent biological responses. The CCK analog JMV-180, which is an agonist at the high-affinity CCK_1 receptors and an antagonist toward the low-affinity ones, is a useful tool for functionally distinguishing the two receptor states.

CCK receptors in the gastrointestinal tract In the colon, the main target of CCK is the myenteric plexus, which has predominantly CCK_1 receptors. In addition, CCK_1 receptors are also present, at moderate-to-low density, in the longitudinal muscle, while the circular muscle is negative for both CCK_1 and CCK_2 receptors. Neither CCK_1 nor CCK_2 receptors are in the blood vessels, lymphoid tissue, mucosa, or muscularis mucosa. These data suggest that CCK affects colonic motility by two fundamentally different pathways: acting on neurons in the myenteric plexus and directly on the smooth muscle cells.

A large number of studies implicate vagal afferent mechanisms in the control of several CCK actions such as relaxation of proximal stomach, gallbladder contraction, and pancreatic enzyme secretion.

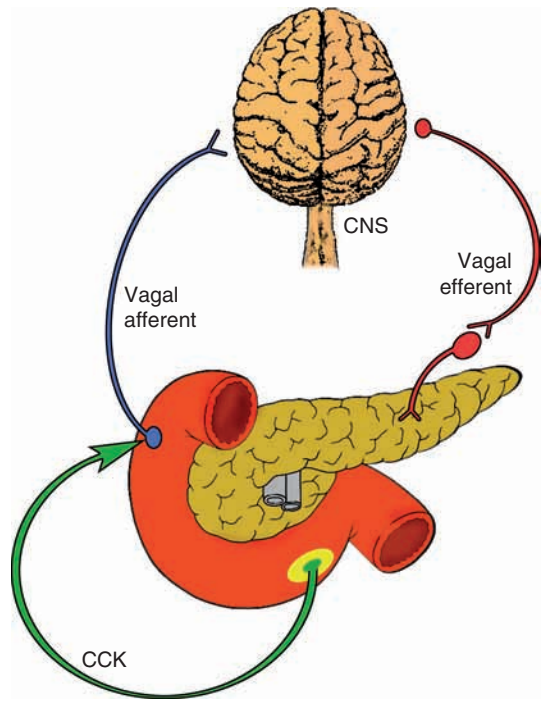

Figure 1 CCK released by endocrine cells in the duodenum acts on CCK receptors expressed by vagal afferent fibers originating in the duodenal mucosa. The response carried by vagal efferent fibers stimulates pancreatic secretion.

Expression of both CCK_1 and CCK_2 receptors has been demonstrated in vagal afferent fibers and in the nodose ganglion. Several studies indicate that CCK acts via an afferent vagal pathway originating in the duodenal mucosa to stimulate pancreatic secretion in rats as well as in humans (Figure 1). Furthermore, activation of these receptors is thought to initiate a vagovagal reflex inhibition of gastric motor function.

CCK receptor agonists/antagonists in disease treatment Because CCK is involved in sensory and motor responses to distension in the intestinal tract, it may contribute to the symptoms of constipation, bloating, and abdominal pain that are often characteristic of functional gastrointestinal disorders and irritable bowel syndrome. These dysfunctions are associated with motor abnormalities in the small intestine and colon that may depend on exaggerated and prolonged CCK release. CCK_1 receptor antagonists have been evaluated clinically to stimulate transit and treat irritable bowel syndrome. CCK also belongs to the growing list of factors modulating food consumption, and CCK molecules are actually promising targets for anti-obesity drugs. Since the initial discovery of its property as a food-intake inhibitor, CCK was demonstrated to be a short-term, meal-reducing signal in most mammalian species including

humans. The availability of potent and specific CCK agonists and antagonists has allowed the identification of the CCK_1 receptor as the receptor type involved in the satiety actions of CCK. Indeed, specific CCK_1 receptor agonists potently inhibit food intake, whereas agonists of the CCK_2 receptor do not. The CCK_1 receptors mediating CCK action on food intake are localized on the vagus nerve, which relays to the hypothalamus via brain stem areas such as the nucleus tractus solitarius and the area postrema. In the brain, CCK contributes to the sensation of satiety, and it may be involved in the mediation of the control of food intake. In addition to the CCK_1 receptors, the CCK_2 receptor may also be involved in the CCK control of food intake. Indeed, CCK_2 receptors are expressed in the vagus nerve brain stem complex that mediates the satiety effect of peripheral CCK, and the CCK_2 receptor is the predominant form found in the brain. Consistent with a possible involvement of CCK_2 receptors, centrally or peripherally administered CCK_2 receptor antagonists have been reported to increase food intake.

See also: Magnocellular Neurosecretory System: Organization, Plasticity, Model Peptidergic Neurons; Neuroendocrine Peptide Processing; Neuropeptide Signaling in Invertebrates; Neuropeptide Inactivation or Metabolism; Neuropeptide Receptors – Drug Development; Neuropeptide Release; Neuropeptides in Autonomic Neurons; Neuropeptides Internalization; Neuropeptides and Coexistence; Neuropeptides: Discovery; Neuropeptides: Enteric Nervous System.

Further Reading

Boehm S and Kubista H (2002) Fine-tuning of sympathetic transmitter release via ionotropic and metabotropic presynaptic receptors. *Physiological Reviews* 54: 43–99.

Brain SD and Grant AD (2004) Vascular actions of calcitonin gene-related peptide and adrenomedullin. *Physiological Reviews* 84: 903–934.

Dakhama A, Larsen GL, and Gelfand EW (2004) Calcitonin gene-related peptide: Role in airway homeostasis. *Current Opinion in Pharmacology* 4: 215–220.

Dufresne M, Seva C, and Fourmy D (2006) Cholecystokinin and gastrin receptors. *Physiological Reviews* 86: 805–847.

Dvoráková MC (2005) Cardioprotective role of the VIP signaling system. *Drug News and Perspectives* 18: 387–391.

Endoh T (2004) Modulation of voltage-dependent calcium channels by neurotransmitters and neuropeptides in parasympathetic submandibular ganglion neurons. *Archives of Oral Biology* 49: 539–557.

Groneberg DA, Rabe KF, and Fischer A (2006) Novel concepts of neuropeptide-based drug therapy: Vasoactive intestinal polypeptide and its receptors. *European Journal of Pharmacology* 533: 182–194.

Harmar AJ (2004) Receptors for gut peptides. *Best Practice and Research Clinical Endocrinology and Metabolism* 18: 463–475.

Huidobro-Toro JP and Donoso MV (2004) Sympathetic co-transmission: The coordinated action of ATP and noradrenaline and their modulation by neuropeptide Y in human vascular neuroeffector junctions. *European Journal of Pharmacology* 500: 27–35.

Leeb-Lundberg LMF, Marceau F, Müller-Esterl W, Pettibone DJ, and Zuraw BL (2005) International Union of Pharmacology. XLV. Classification of the kinin receptor family: From molecular mechanisms to pathophysiological consequences. *Pharmacological Reviews* 57: 27–77.

Ohki-Hamazaki H, Iwabuchi M, and Maekawa F (2005) Development and function of bombesin-like peptides and their receptors. *International Journal of Developmental Biology* 49: 293–300.

Pedrazzini T (2004) Importance of NPY Y1 receptor-mediated pathways: Assessment using NPY Y1 receptor knockouts. *Neuropeptides* 38: 267–275.

Peeters TL (2003) Central and peripheral mechanisms by which ghrelin regulates gut motility. *Journal of Physiology and Pharmacology* 54: 95–103.

Sternini C (2001) Receptors and transmission in the brain-gut axis: Potential for novel therapies. III: μ-Opioid receptors in the enteric nervous system. *American Journal of Physiology – Gastrointestinal and Liver Physiology* 81: G8–G15.

Relevant Websites

http://www.gproteins.com – G Protein Coupled Receptors (GPCR) and G Protein Research Portal.

http://www.gpcr.org – Information System for G Protein-Coupled Receptors (GPCRs) (GPCRDB).

http://www.iuphar-db.org – Receptor Database (IUPHAR).

http://www.signaling-gateway.org – UCSD Nature Signaling Gateway Home.

Perception and Eye Movements

L S Stone and B R Beutter, NASA Ames Research Center, Moffett Field, CA, USA
M P Eckstein, University of California at Santa Barbara, Santa Barbara, CA, USA
D B Liston, San Jose State University Foundation, NASA Ames Research Center, Moffett Field, CA, USA

© 2009 Published by Elsevier Ltd.

The study of eye movements has long enjoyed a special status in biomedical research. First, compared to other sensorimotor systems, quantitative visual stimuli are relatively easy to create and quantitative eye movement measurements are relatively easy to acquire. Second, prior to the development of advanced imaging technologies, eye movements were the most important noninvasive diagnostic tool to localize brain pathology. Third, the eyeball has a constant rotational inertia and does not lift anything; thus, eye movements can be made largely independent of the on-line proprioceptive feedback needed to calibrate the force–displacement relationships of other motor systems. In summary, oculomotor control was perceived to be a straightforward, simple model system for understanding human sensorimotor control.

On the motor side, the 1960s brought about a revolution in oculomotor research, driven largely by the work of David A Robinson and Robert Wurtz. These researchers pioneered the combined use of the rigorous analytical tools of linear systems theory and of the powerful new experimental capability to reliably and precisely monitor eye movements while recording for extended periods of time from single visuomotor neurons of monkeys performing oculomotor tasks. In the subsequent half century, by systematically tracing, recording from, lesioning, stimulating, and ultimately modeling the major oculomotor pathways in cortical and subcortical structures, their two laboratories and those of their scientific progeny and colleagues made amazing progress in extending our understanding of how visual signals are transformed into eye movement motor commands.

On the vision side, the 1960s brought a similar revolution with John Robson, Fergus Campbell, Christina Enroth-Cugell, and others applying linear systems theory to visual perceptual performance. The 1980s brought a second revolution in vision science driven by David Marr and others, who applied a machine vision, computational perspective to human vision. Their synthetic approach of seeking biologically inspired (and constrained) solutions to real-world engineering and robotics problems by emulating the information processing within the human brain served to highlight the full complexity of the visual processing needed for the effective control of motor actions.

During approximately the past two decades, it has become increasingly clear that the initial hope of the 1960s – that quasi-linear feedback models that transform a retinal position signal (and its derivatives) into an eye position signal (and its derivatives) might explain the critical aspects of voluntary oculomotor behavior – was overly optimistic. Indeed, recent results have highlighted that highly complex and nonlinear aspects of visual perception play a critical role in guiding both voluntary smooth pursuit and saccades. Decades of research using small-spot visual stimuli that were nearly infinitely detectable, identifiable, and localizable revealed much about the motor factors and circuits that limit human oculomotor performance; however, much less was learned about the visual mechanisms that both limit and enrich oculomotor performance. A number of laboratories have begun to measure human saccadic and pursuit performance under more realistic visual conditions in order to examine how visual processing shapes oculomotor behavior.

Action and Perception

In 1982, Ungerleider and Mishkin made a major contribution to our understanding of primate visual information processing by identifying two major ascending processing streams through extrastriate cortex: a dorsal stream through parietal cortex with a hierarchical set of areas that focus on static and dynamic spatial characteristics of the stimulus (the 'where' pathway) and a ventral stream through the inferotemporal cortex with a hierarchy of areas that focus on identifying features of the stimulus (the 'what' pathway). Later, based largely on clinical observations, Goodale and Milner pushed this dichotomy a step further by proposing that the dorsal and ventral streams represent separate, largely independent, visual pathways supporting action and perception, respectively. They argued that object shape information from the ventral stream used for perception is not available to motor systems, and that spatial information from the dorsal pathway was not available to support perception. The critical finding from Newsome's laboratory that local microstimulation of the middle temporal (MT) or medial superior temporal (MST) areas generates spatially localized and directionally specific biases in visual motion discrimination performance demonstrates that the dorsal pathway does play a major causal role in perception. However, these

findings do not rule out the possibility that ventral stream information may be inaccessible to motor systems.

A careful reexamination of early reports of dissociations of visual responses for perception and action used to support the Goodale–Milner hypothesis has shown that the purported failure to find a vulnerability to visual illusions for motor actions could be accounted for by subtle differences in the visual demands of the perceptual and motor tasks or by offsetting effects of a misperceived target location and of a distorted motor reference frame that obscure the visual impact of the illusion on motor action. When perception and motor actions are tested either simultaneously or using carefully matched conditions and tasks, both are similarly vulnerable to illusions and other properties of the visual stimulus, undermining the view that visual information for perception is not shared with those systems controlling motor actions.

This article reviews key findings in the recent literature examining the extent to which the visual signals supporting perception are also used to drive voluntary human smooth pursuit and saccadic eye movements.

Smooth Pursuit and Visual Motion Perception

The linear systems view of pursuit has been dominated by retinal image motion models (Figure 1(a)). In this view, pursuit is essentially a simple, quasi-linear,

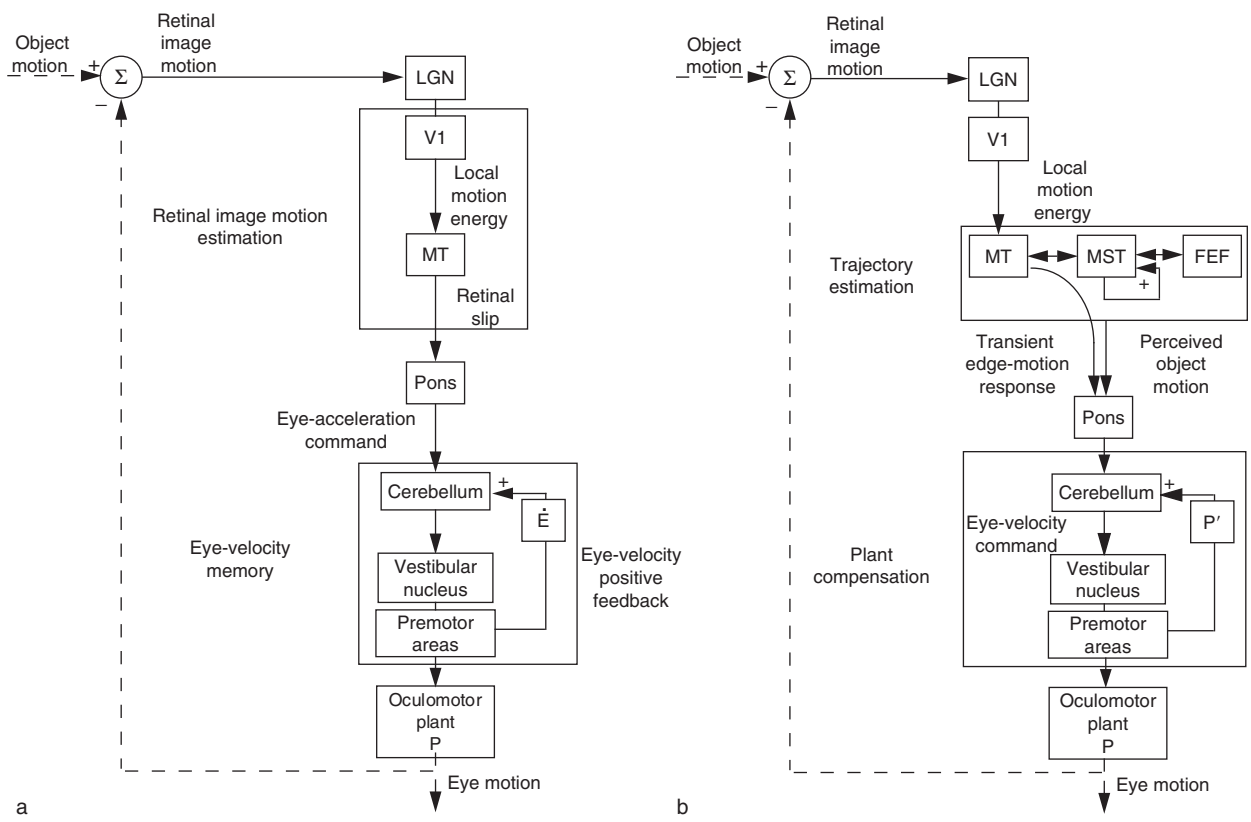

Figure 1 Pursuit models. (a) Retinal image motion models. In this framework, retinal image motion is analyzed in area MT (and associated extrastriate cortical areas in the superior temporal sulcus) to generate retinal position, velocity, and acceleration error signals. Weighted versions of these signals are summed and then sent to the brain stem–cerebellar structures as a command for pursuit eye acceleration. A positive feedback loop between the brain stem and cerebellum sustains ongoing eye velocity and integrates the eye-acceleration commands to generate changes in eye velocity. (b) Object motion models. In this framework, retinal image motion is first segmented in order to identify local edge motions which are subsequently selectively integrated to yield a measure of object motion. To attempt to resolve the inherent ambiguities in this problem, bottom-up retinal motion data, top-down cognitive assumptions and expectations, and either feed-forward or feedback signals related to ongoing eye motion are combined in a sluggish process that involves at least parietal and frontal cortex. The resulting visual object–motion signal is then sent to brain stem–cerebellar premotor circuits that prefilter the object-motion signal to compensate for the oculomotor plant before sending it on to motor neurons as the command for eye motion. (a) Adapted from Lisberger SG, Morris EJ, and Tychsen L (1987) Visual motion processing and sensory–motor integration for smooth pursuit eye movements. *Annual Review of Neuroscience* 10: 97–129. (b) Adapted from Stone LS, Beutter BR, and Lorenceau J (2000) Visual motion: Integration for perception and pursuit. *Perception* 29: 771–787.

continuous control system that attempts to minimize retinal image velocity through negative feedback. A retinal motion error signal or retinal slip (primarily a retinal velocity signal) is computed in area MT and sent to the brain stem as a command to drive eye acceleration. Meanwhile, ongoing eye velocity is maintained by a positive feedback circuit through the cerebellum which acts as an integrator of the acceleration commands. Although retinal image motion models can do a good job of explaining pursuit responses to a single white spot moving on a dark background (the only situation in which retinal image motion is identical to object motion), they cannot explain human pursuit responses to the wider range of visual conditions (and retinal stimuli) experienced when tracking a realistic object (which generates a set of disparate retinal local-edge motions) across a realistically textured environment.

Early studies revealed problems with the core idea that retinal motion error and eye velocity positive feedback form the command signal ultimately driving pursuit eye movements. The earliest pursuit studies observed zero phase-lag pursuit of sinusoidal motion, demonstrating a predictive component to the pursuit response that cannot be accounted for by eye velocity memory. In 1976, Steinbach showed that humans could pursue the horizontal motion of a rolling wagon wheel when only a few illuminated dots were visible on the rim. The perceived horizontal motion of an invisible rolling wheel is tracked, even though the retinal motions of the visible dots are cycloidal. Any observer who pursues this wagon wheel stimulus becomes immediately aware that pursuit does not simply respond to the motion in the retinal image. An additional red flag was raised in 1988 when Newsome, Wurtz, and Komatsu found signals related to ongoing eye motion in area MST, an area shown by both lesions and microstimulation studies to be causally involved in motion perception as well as generating the motor command for pursuit. Thus, the cortical signal driving smooth pursuit is not simply a retinal error signal but, rather, one more related to object motion in head- or world-centric coordinates. Moreover, that same year, Dursteler and Wurtz found that MST lesions cause direction-specific pursuit deficits, with no effort to make smooth corrections of the large residual retinal motion error during the poor steady-state tracking performance. These findings put into question any model of pursuit that proposes a negative feedback circuit that simply attempts to drive retinal motion to zero.

Numerous studies have shown clear links between pursuit and perceived motion, further distancing pursuit behavior from raw retinal velocity errors. First, perception and pursuit show the same speed and direction discrimination performance and are similarly affected by nonlinear directional anisotropies such as the oblique effect, suggesting that rate-limiting neural speed and direction signals are shared by perception and pursuit. Furthermore, analysis of the trial-by-trial variation in human perceptual and pursuit direction responses shows higher correlation than expected by chance from independent mechanisms (**Figure 2(a)**). Thus, perception and pursuit share a limiting internal direction noise source, thereby undermining the theory that pursuit and perception have separate and independent, but fortuitously equal, underlying direction signals. Second, high-level cognitive expectations affect perception and pursuit equally. For example, in 2001, Krauzlis and Adler showed that when the likely direction of stimulus motion is cued prior to each trial, humans generate similar perceptual and pursuit performance biases toward the likely direction. Third, pursuit and perception appear to be fooled in the same way by visual illusions. For example, in 1996, Ringach, Hawken, and Shapley showed that humans generate a pursuit vergence response to the illusory motion in-depth of the kinetic depth effect. Fourth, the time course of the computation of object motion is similar for perception and pursuit. In particular, a human psychophysical study by Shiffrar and Lorenceau in 1996 and a human pursuit study by Masson and Stone in 2002 showed that the temporal integration of an initial set of local edge motion signals into a bona fide object motion signal is similarly sluggish (on the order of a few hundred milliseconds) for both perception and pursuit. In 2001, Pack and Born found a similarly sluggish temporal evolution of the direction signals encoded in the firing rate of MT neurons and those driving smooth eye movements in monkeys. Fifth, when presented with ambiguous image motion, perception and pursuit choose the same arbitrary interpretation of object motion, influenced by the same static image properties. In particular, when Madelain and Krauzlis asked observers to pursue a motion stimulus whose direction is completely ambiguous such that an observer can select a leftward or rightward percept at will, each perceptual reversal was associated with a pursuit reversal. Furthermore, on average, the perceptual reversal precedes the pursuit reversal by nearly 100 ms, consistent with perception driving pursuit within an object-motion pursuit model (**Figure 3**).

Finally, in 2001, Alais and Lorenceau showed that the spatial configuration or shape of ambiguous line drawings affects their perceived motion. In this regard, humans show similar effects of object shape on pursuit (**Figure 4**). For example, the motion of a diamond occluded by vertical windows appears rigid and is both veridically perceived and accurately

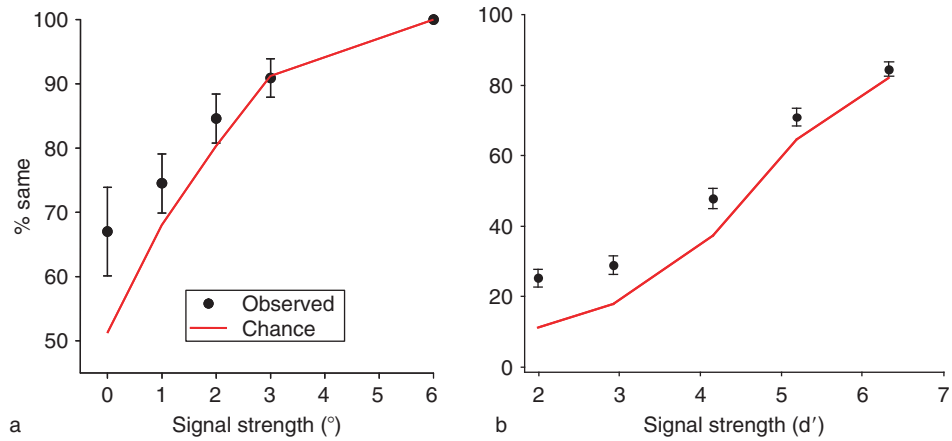

Figure 2 Covariation analysis (also called % same or choice probability). (a) Shared internal direction noise for pursuit and perception. In a simultaneous study of perceptual and pursuit direction signals, the binary psychometric and oculometric decisions were highly correlated across the entire range of signal strengths tested. The correlation expected by chance for two independent mechanisms is 50 % in the no-signal condition and increases monotonically to 100 % in the high-signal condition in which both decisions are always correct and therefore trivially the same. This figure plots the % same (the percentage of trials for which the pursuit and perceptual binary direction decisions were the same) for a single observer as a function of signal strength (direction difference in degrees from a cardinal axis). Note that the observed correlation was significantly higher than expected by chance (error bars show 95% confidence intervals). (b) Shared response to external noise for saccades and perception. In sequential studies of perceptual and saccadic localization with an identical set of noisy search images, the 10-alternative forced choice (10AFC) psychometric and oculometric decisions were correlated across the entire range of signal strengths tested (signal-to-noise levels in d′ units). The correlation expected by chance from two independent mechanisms is 10 % in the zero SNR condition and increases monotonically to 100 % for an extrapolated high SNR condition in which both decisions are always correct and therefore trivially the same. This figure plots the % same (the percentage of image samples for which the perceptual and saccadic 10AFC decisions were the same) for a single observer as a function of signal strength (ratio of peak luminance of the target to the amplitude of the pixel luminance noise). Note that the observed correlation was significantly higher than expected by chance (error bars show standard errors). (a) Adapted from Stone LS and Krauzlis RJ (2003) shared motion signals for human perceptual decisions and oculomotor actions. *Journal of Vision* 3: 725–736. (b) Figure based on data from Beutter BR, Eckstein MP, and Stone LS (2003) Saccadic and perceptual performance in visual search tasks: I. Contrast detection and discrimination. *Journal of the Optical Society of America A* 20: 1341–1355.

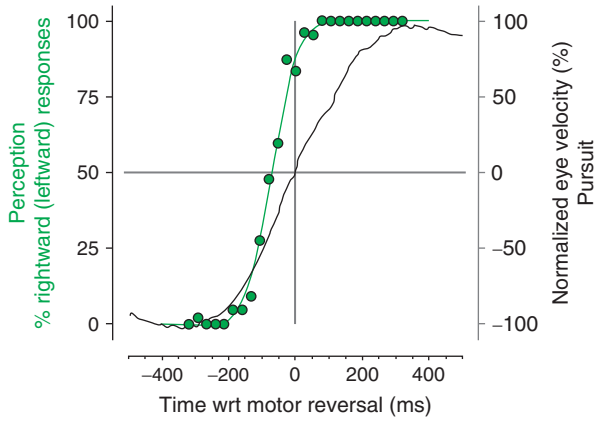

Figure 3 Correlated pursued and perceived direction reversals for completely ambiguous motion. For a stimulus that can be perceived as moving rightward or leftward at will, the data show that the perceptual reversal (plotted as percentage rightward or leftward judgments) preceded the oculomotor direction reversal (plotted as mean eye velocity) by slightly less than 100 ms, consistent with perceived motion driving pursuit. The retinal image motion is constant throughout but pursuit changes from rightward to leftward (or vice versa) synergistically with the reversals in perceived motion, which is inconsistent with all retinal image motion models of pursuit. Adapted from Madelain L and Krauzlis RJ (2003) Pursuit of the ineffable: Perceptual and motor reversals during the tracking of apparent motion. *Journal of Vision* 3: 642–653.

pursued. However, the motion of a similarly occluded cross (defined by the exact same segment image motion but merely spatially rearranged) is perceived as nonrigid; the coherent two-dimensional (2-D) object motion interpretation is presumably interfered with by the distracting vertical motion of the segments. Pursuit of a diamond moving along an isotropic trajectory shows an isotropic response (i.e., vertical and horizontal pursuit gains are equal), consistent with the motion of a rigid object. However, pursuit of the cross shows lower overall gain and a clear vertical bias, consistent with the nonrigid object motion percept and the vertical segment motion inference. This finding suggests that object shape information, presumably from the ventral pathway, is influencing perception and pursuit in parallel. Moreover, even an extended version of the retinal image motion model that attempts to minimize the sum (or vector average) of all the retinal segment motions would predict that pursuit of the cross and pursuit of the diamond would be identical.

The emerging picture is that pursuit is driven by a signal related to perceived object motion and not directly by retinal image motion. An object motion model of pursuit (Figure 1(b)) can not only explain

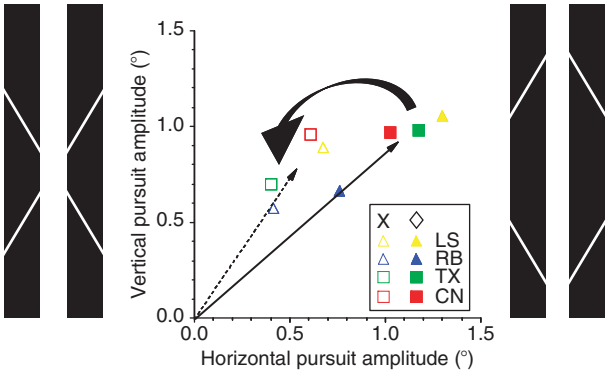

Figure 4 Different pursuit for identical retinal motion related to differences in object shape. Moving occluded line figures with identical object motion trajectories and identical segment motions (but with the segments in different relative retinal locations) can nonetheless be perceived and pursued differently. The data show the horizontal and vertical pursuit amplitudes of four observers in response to a diamond (solid) or a cross (open) configuration representing the same object motion along an isotropic 2-D Lissajous figure. The diamond is perceived to move rigidly and is pursued with an isotropic high-gain response. The cross, however, appears to move nonrigidly and is pursued with a low-gain trajectory biased in the vertical direction of the segment motion. The fact that both of these stimuli generate 2-D trajectories at all is inconsistent with retinal image motion models, which would generate nearly exclusively vertical pursuit in response to the vertical retinal segment motion. More important, the fact that identical physical object motion is pursued differently when different non-motion information generates different object motion percepts is strong evidence that pursuit is driven by an object motion signal shared with perception. Adapted from Stone LS, Beutter BR, and Lorenceau J (2000) Visual motion: Integration for perception and pursuit. *Perception* 29: 771–787.

the wider range of new phenomena described previously but also takes into consideration the wealth of physiological findings demonstrating an extensive role for both parietal (MST) and frontal (frontal eye field (FEF) pursuit area) cortex in trajectory estimation (including prediction). It also acknowledges that these areas, not just MT, provide a major component of the descending pathway to brain stem and cerebellar structures involved in the generation of the motor output signal driving pursuit. The role of cortex in this new view is not to calculate derivatives of retinal motion but, rather, to compute a best estimate of object motion from the available visual information (both dynamic and static) within a cognitive and attentional context. If an object motion signal is sent down to the brain stem (instead of a retinal error signal), then the role of the cerebellar feedback loop must also be reassessed. If object motion is the descending cortical output signal, then the only processing needed to optimize the oculomotor command to the plant is to high-pass prefilter the eye velocity command to compensate for the low-pass filtering of the oculomotor plant. Indeed, the output signals from the cerebellum supporting pursuit have been shown to be a prefiltered version of the subsequent eye velocity signal. Such prefiltering could be achieved via a positive feedback loop that mimics the plant. Another important virtue of this model, in addition to its ability to reconcile a wide range of behavioral and physiological pursuit observations, is that it is easily generalized to other motor systems, whereby the cerebellum would act as a compensator for other motor plants all driven by the same cortical object motion signal (or another relevant perceptual signal). Thus, the hand and the eye can work synergistically when tracking a moving object.

Saccades and Visual Target Selection

The linear systems theory view of the saccadic system has been dominated by variants of a retinal offset model (Figure 5(a)). In this view, the saccadic system is essentially a simple, quasi-linear, bang-bang control system that rapidly corrects retinal position offset by driving this offset to zero through negative feedback. A retinal position error signal from primary visual cortex (V1), modulated by inputs from other cortical areas, is sent down to the superior colliculus (SC) of the midbrain as the retinal target position signal. When a particular locus in the SC map reaches a threshold, that location relays a command for desired eye position to brain stem structures that generate the neural pulse and step signals needed to accelerate the eye and then hold it in its new position, respectively. Although this model includes a detailed description of brain stem structures, as with the retinal image motion pursuit model, the entire visual and decision-making process is the trivial calculation of an unambiguous retinal error signal – in this case, position instead of velocity. Again, such a retinal signal is only unambiguous in the degenerate case of a small spot on a dark background. Furthermore, it is has long been known that nonretinal signals can drive voluntary saccades, such as auditory or visual memory signals.

During approximately the past decade, a number of laboratories have turned their attention to understanding the visual processes underlying saccadic target selection in search and search-like tasks using more realistic visual stimuli with targets of variable signal strength, embedded in external image noise, among similarly salient nontarget 'distractors,' and/or with differing *a priori* location probabilities. These studies have revealed a rich complexity of signal processing, requiring broadly distributed interactions between parietal and frontal cortex and the midbrain.

Numerous studies have illustrated clear links between saccades and perception, further distancing saccadic behavior from raw retinal offset signals. First, it

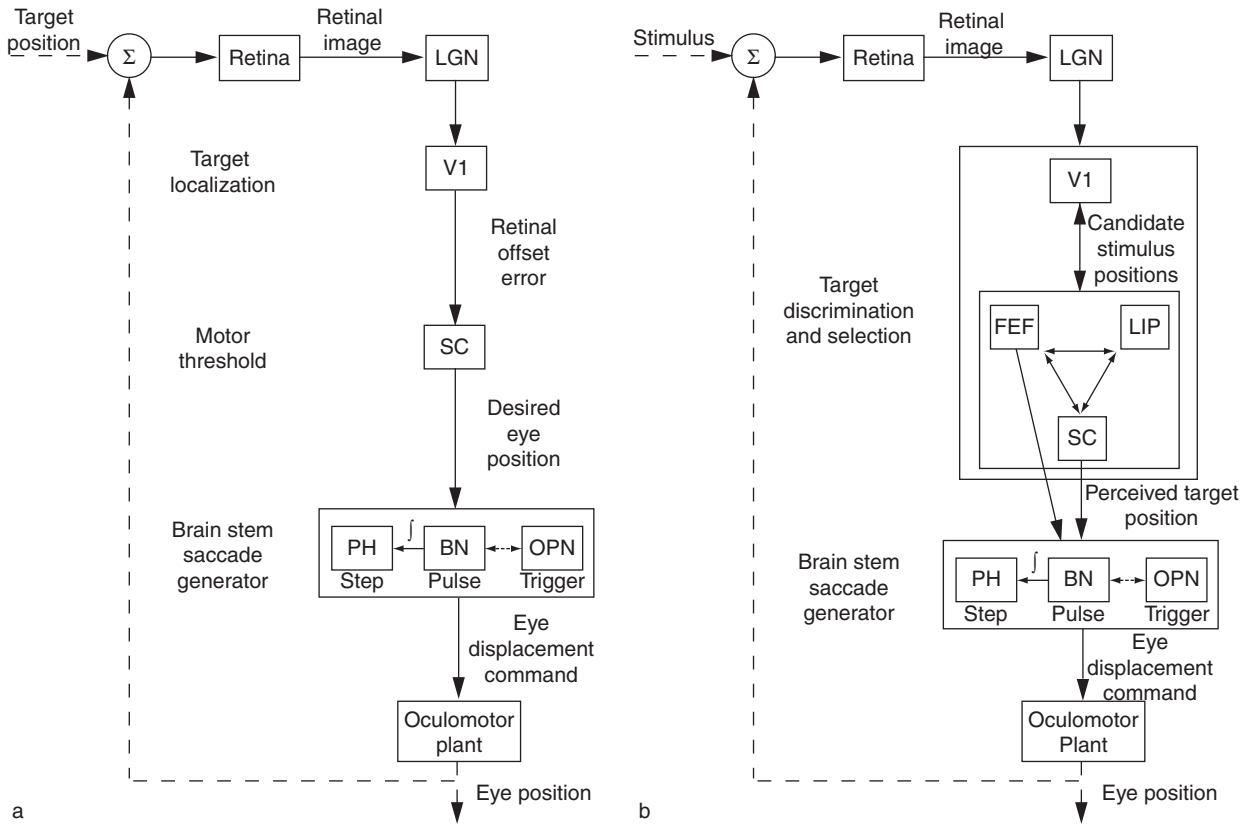

Figure 5 Saccade models. (a) Retinal offset error models. In this framework, a retinal offset signal is computed in primary visual cortex (V1) (with input from parietal and frontal cortical areas) and then send to the superior colliculus (SC) of the midbrain. When that signal reaches threshold, the SC sends a desired eye position signal to brain stem saccadic structures, which generate the command for eye displacement. (b) Target selection models. In this framework, the totality of the retinal image is processed in parallel to identify candidate target locations. A network that includes the lateral–intraparietal area (LIP), the frontal eye fields (FEF), and the SC works to select a single target using the available bottom-up sensory data modulated by top-down cognitive and attentional factors. The resulting perceived target position signal is then sent to brain stem saccadic structures, which generate the command for eye displacement.

has been shown that saccades and perception appear similarly affected by spatial illusions. For example, saccades (and hand pointing as well) show systematic spatial deviations in the same direction as the spatial misperceptions for stimuli that evoke the Muller–Lyer illusion. Although early reports of this effect on saccades were conflicting, when the saccadic task relies on a visual feature that is distorted by the illusion, the motor response reflects the perceptual illusion. Second, contrast detection and discrimination thresholds are equal for perception and saccades. More important, using noisy search stimuli, the correlation between the external image noise-driven responses of saccades and perception is higher than expected by chance from two independent mechanisms (Figure 2(b)). Thus, saccades and perception appear equally vulnerable to those random spatial features of a particular noise sample that appear target-like. Third, using classification image analysis (a form of reverse correlation), by combining the hundreds of noise samples (sans target) that generated erroneous responses, one can construct an image of the virtual target that saccades and perception were looking for (i.e., an estimate of the template or overall receptive field used in the decision-making process). Not only are the shapes of the classification images from saccadic and perceptual decisions indistinguishable but also their common shape differs significantly from that of the target (Figure 6). Indeed, both images are a high-passed version of the target, which may indeed be optimal if one is searching for the target embedded in the low-pass filtered noise commonly associated with visual neural signals.

The emerging view is that a complicated network between the parietal and frontal cortex, as well as the SC, converges to select a single target using both bottom-up sensory data and top-down cognitive knowledge, expectation, and/or anticipation of reward. Furthermore, ascending feedback from the SC to visual cortical areas appears to play an important role in this process by driving attention-like modulations of visual processing, thereby influencing

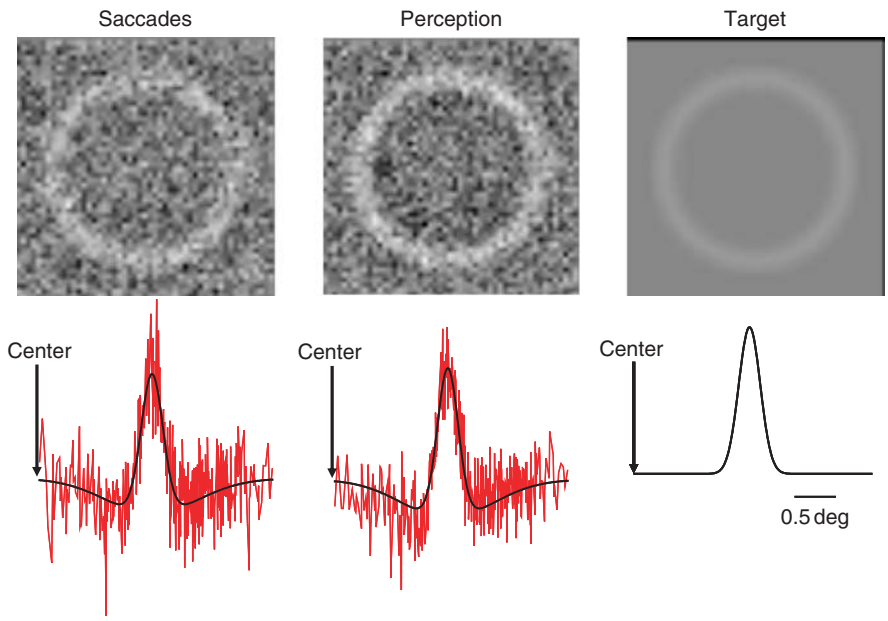

Figure 6 Classification image (CI) analysis reveals the saccadic and perceptual 'receptive field' used to search for the brightest 'O' among dimmer Os. The top row shows the raw CI for saccades and perception, respectively. The bottom row shows the normalized and polar averaged cross-section of the receptive fields. Note that the perceptual and saccades receptive fields are both high-pass filtered versions of the Gaussian stimulus cross-section and are well described by difference of Gaussian functions (solid black lines under the red data traces). Thus, not only is the receptive field shape the same for saccades and perception but also the perceptual and saccades receptive fields are more similar to each other than they are to the actual target, providing strong evidence of a shared target template used by perception and saccades during search.

perception, although the exact nature of these modulations is not well understood. The target selection model of saccades (Figure 5(b)) acknowledges the simple fact that the saccadic system is not finding a single target presented in a vacuum but, rather, is selecting the most behaviorally relevant target from among many candidate possibilities. Thus, rather than thinking of visual target localization as a trivial thresholding of a single highly salient spot, studies have reframed saccadic targeting within the context of signal detection theory, addressing the nontrivial question of how to decide where to look next when the target is inconspicuous or is similar to surrounding distractors. Saccadic and perceptual search can be described as a template-matching task in the presence of noise and spatial uncertainty; the template captures essential contrast and shape information about the searched-for target albeit not necessary perfectly matched to the target (Figure 6). Indeed, the similarity of the classification images for perception and saccades (and their shared difference with the actual target shape) strongly suggests that the saccadic system has access to the same shape information as perception, presumably from shared ventral pathways supporting both perception and action.

Although the new framework for saccadic targeting embraces a novel view of visual processing, unlike pursuit, the proposed new visual front end for saccades does not require a fundamental reinterpretation of the role of brain stem saccade-generating structures. However, it is becoming increasingly clear that the SC is not merely a feed-forward output relay to brain stem structures but, rather, plays a key role in the target selection process, along with the lateral intraparietal (LIP) area and the FEF. Consistent with this distributed responsibility for target selection, Schiller and colleagues showed that bilateral removal of either the SC or the FEF leaves the ability to make saccades intact, whereas bilateral removal of both areas destroys that ability. Furthermore, studies in both the Wurtz and Newsome laboratories have demonstrated that stimulation of the SC can modulate perceptual performance, justifying the elevation of the SC into the target-selection network, which was formerly considered to be exclusively cortical.

A New Framework for the Visuomotor Control of Pursuit and Saccades

The major paradigm shift that has occurred in the study of pursuit and saccades is the realization that the visual processing part of the sensorimotor problem is more complicated than originally thought.

When humans 'see' by taking the neural signals coming from the approximately 1 million ganglion cells in each eye to create a full-color 3-D model of the world around them, quite a complicated process has occurred. The sensory information available represents two 2-D projections of the world, one from each retina, and is therefore necessarily incomplete. The human visual system must use the available sensory data (bottom-up information) as well as *a priori* knowledge about the world (top-down information) to infer a best guess of what is actually out there. As such, all visual perception is illusion; what we call illusions are merely those rare instances when the brain's best guess is obviously wrong.

Complex interactions between frontal, parietal, inferotemporal, and occipital cortex, along with the midbrain, are used to combine bottom-up and top-down information into a time-varying best estimate of the oculomotor target. For pursuit, the reconstruction of the target object's motion in the world takes time. An early quick estimate of object motion using the vector average of local edge motions appears to drive early perceptual and pursuit responses, with a more accurate estimate emerging with a time constant on the order of 100 ms. The standard linear system control theory approach of considering only the initial 100-ms 'open-loop' pursuit response to characterize the internal visuomotor transfer function therefore misses much of this elegant, yet sluggish, computational process. Object motion estimation of realistic targets takes time because, after the initial sensory transduction, the visual system must first segregate from the background those pieces of the retinal image that represent individual local motion features. The system must then selectively integrate those pieces that properly belong together into a whole object whose global motion can then be computed. This process is inherently ambiguous because the data set is incomplete (although over time the ambiguity of the object trajectory can resolve itself). Thus, the perceptual and pursuit systems must make visual choices. The preponderance of evidence gathered during approximately the past decade argues that perception and pursuit make the same choice using shared visual motion processing mechanisms broadly distributed across parietal and frontal cortex, with likely input from inferotemporal cortex as well.

For saccades, localizing the target object is also a complex, time-consuming visual process involving both parallel and serial components: parallel because all portions of the visual field are being processed simultaneously and serial because, at low target salience, a sequence of discrete fixation steps is traced as one searches for the target. Presumably, during each fixation, a target-like template is being compared with incoming noisy sensory data in order to determine the 'best' next location to foveate, with multiple locations constantly competing for attention. The preponderance of evidence indicates that perception and saccades make the same choices using shared spatial visual processing mechanisms broadly distributed between parietal and frontal cortex, with inputs from the superior colliculus and inferotemporal cortex as well.

Another difference between the traditional oculomotor models (**Figures 1(a)** and **5(a)**) and the new frameworks (**Figures 1(b)** and **5(b)**) is the coordinate system of the visual signals driving motor actions. The traditional view of saccades and pursuit assumes that the controlled visual signals are in raw retinal coordinates: retinal position, velocity, or even acceleration error signals to be minimized by negative feedback. Not only do the aforementioned MST studies rule this out for pursuit but also a behavioral study of the direction signals driving perception during fixation and pursuit by Krukowski and colleagues in 2003 showed that the limiting direction noise and even the oblique effect are not in retinal coordinates but, rather, in head- or world-centered coordinates. This finding shows that performance in an apparently simple direction discrimination task is limited by neural signals beyond area MT (the highest area mapped in retinal coordinates) and thus in area MST (an area in which retinal motion signals are combined with eye and head motion signals) or another motion area further along in the dorsal pathway. Indeed, there is even direct physiological evidence that some MST neurons encode object motion in world coordinates.

The cortical signals driving saccades are also not simply in retinal coordinates. LIP and FEF both encode target location in 'oculocentric' coordinates, which comprise a coordinate system in which target location is constantly being updated with respect to the current eye position using anticipatory signals about future eye movements (even in the absence of visual feedback). This updating process, first described by Colby, Duhamel, and Goldberg, effectively keeps the target location stably encoded in head-centered space – yet bouncing around in the retinal map with every saccade – presumably as part of the mechanism keeping the world perceptually stable during saccades. Sommer and Wurtz characterized an oculomotor feedback pathway from SC to FEF, via the mediodorsal thalamus, and have further shown that interruption of the corollary discharge signal from the SC disrupts the proper updating of target location, thereby demonstrating that the SC plays a key role in the coordinate system transformation for target localization performed within the cortex.

Clearly, the frameworks in Figures 1(b) and 5(b) are mere cartoons that require much fleshing out. Nonetheless, they succinctly represent the concatenation of a number of conceptual advances that are guiding the next generation of oculomotor researchers and cortical physiologists. Regrettably, this short article can mention only a small number of the many important recent findings. Many researchers are actively working to understand the visual processing for saccade target selection and pursuit object motion estimation and the associated neurophysiological underpinnings. Although much progress has been made, the picture of how an extended network of cortical (and even subcortical) areas interact to shape synergistic perception and sensorimotor control remains incomplete.

Conclusion

Because the sensory 'image' of the world is inherently ambiguous, humans must generate a best-estimate solution to this complex puzzle, using noisy incomplete data and cognitive strategies to construct an internal 3-D model of visual space, in order to interact with their environment and ultimately to survive in the real world. Perhaps, our coherent multisensory perception of the world evolved from the need for the brain to provide a single, decisive interpretation of an ambiguous and incomplete set of visual (and other) signals in order to coordinate the actions of multiple motor systems. Indeed, above and beyond the parsimony argument about limited cortical processing resources, it is difficult to imagine an evolutionary advantage of each neural system developing a set of independent, potentially conflicting, private solutions to the visual problems of detection, discrimination, and target selection.

See also: Active Perception; Attention and Eye Movements; Basal Ganglia and Oculomotor Control; Brainstem Control of Eye Movements; Cerebellum and Oculomotor Control; Contextual Interactions in Visual Perception; Cortical Control of Eye Movements; Eye and Head Movements; Eye Movement Disorders; Oculomotor Control: Anatomical Pathways; Optokinetic Eye Movements; Pursuit Eye Movements; Saccade–Pursuit Interactions; Saccades and Visual Search; Saccadic Eye Movements; Target Selection for Pursuit and Saccades; Vision for Action and Perception.

Further Reading

Beutter BR, Eckstein MP, and Stone LS (2003) Saccadic and perceptual performance in visual search tasks: I. Contrast detection and discrimination. *Journal of the Optical Society of America A* 20: 1341–1355

Dassonville P and Bala JK (2004) Perception, action, and Roelofs effect: A mere illusion of dissociation. *Public Library of Science Biology* 2: e364.

DeGrave DDJ, Franz VH, and Gegenfurtner KR (2006) The influence of the Brentano illusion on eye and hand movements. *Journal of Vision* 6: 727–738.

Eckstein MP, Beutter BR, Pham BT, Shimozaki SS, and Stone LS (2007) Similar neural representations of the target for saccades and perception during search. *Journal of Neuroscience* 27: 1266–1270.

Goldberg ME, Bisley JW, Powell KD, and Gottlieb J (2006) Saccades, salience and attention: The role of the lateral intraparietal area in visual behavior. *Progress in Brain Research* 155: 157–175.

Ilg UJ (2002) Commentary: Smooth pursuit eye movements: From low-level to high-level vision. *Progress in Brain Research* 140: 279–298.

Ilg UJ, Schumann S, and Thier P (2004) Posterior parietal cortex neurons encode target motion in world-centered coordinates. *Neuron* 43: 145–151.

Krauzlis RJ (2005) The control of voluntary eye movements: New perspectives. *Neuroscientist* 11: 124–137.

Krauzlis RJ, Liston D, and Carello CD (2004) Target selection and the superior colliculus: Goals, choices, and hypotheses. *Vision Research* 44: 1445–1451.

Krauzlis RJ and Stone LS (1999) Tracking with the mind's eye. *Trends in Neuroscience* 22: 544–550.

Krukowski AE and Stone LS (2005) Expansion of direction space around the cardinal axes revealed by smooth pursuit eye movements. *Neuron* 45: 315–323.

Lisberger SG, Morris EJ, and Tychsen L (1987) Visual motion processing and sensory–motor integration for smooth pursuit eye movements. *Annual Review of Neuroscience* 10: 97–129.

Madelain L and Krauzlis RJ (2003) Pursuit of the ineffable: Perceptual and motor reversals during the tracing of apparent motion. *Journal of Vision* 3: 642–653

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. New York: Oxford University Press.

Stone LS, Beutter BR, and Lorenceau J (2000) Visual motion: Integration for perception and pursuit. *Perception* 29: 771–787.

Schiller PH and Tehovnik EJ (2005) Neural mechanisms underlying target selection with saccadic eye movements. *Progress in Brain Research* 149: 157–171.

Stone LS and Krauzlis RJ (2003) shared motion signals for human perceptual decisions and oculomotor actions. *Journal of Vision* 3: 725–736

Smeets JB and Brenner E (2006) 10 years of illusions. *Journal of Experimental Psychology: Human Perception and Performance* 32: 1501–1504.

Sparks DL (2002) The brainstem control of saccadic eye movements. *Nature Reviews Neuroscience* 3: 952–964.

Thompson KG and Bichot NP (2005) A visual salience map in the primate frontal eye field. *Progress in Brain Research* 147: 251–262.

Perception of Surfaces and Forms

D Purves, Duke University, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The fundamental problem in understanding perception in any sensory modality is the indirect relationship between our perceptions and the physical world. The challenge this fact presents is especially evident in vision and has been recognized since the beginning of the eighteenth century. The problem is that the images on the retina cannot, by their nature, uniquely specify the physical objects in a scene. For example, physical objects are three-dimensional and what we see also appears to be three-dimensional. Nevertheless, the retinal image is only two-dimensional. As a result, the actual size and arrangement of objects in space are not indicated by the retinal projection as such, and this information cannot be deduced by any logical operation on the retinal image (Figure 1(a)). The physical properties associated with the perceptual qualities of brightness and color are, for the same reasons, unspecified by the information in the retinal image (Figure 1(b)). This quandary in vision is referred to as the inverse optics problem, and the key question in visual perception is how the visual brain solves it.

Beginning with the work of Hermann von Helmholtz in the latter part of the nineteenth century, many vision scientists have supposed that the strategy for solving the inverse problem in perception is to make sensory systems depend in some degree on empirical information (i.e., information about the meaning of ambiguous stimuli based on past experience with what the same or similar stimuli meant). In this way, Helmholtz proposed that the visual system might 'disambiguate' images by an empirical boost that would help out their 'interpretation.' Only recently, however, have vision scientists considered the possibility that percepts might be generated entirely on the basis of the empirical success or failure of visual experience in the past of both the species and the individual. If that were the case, then the purpose of visual processing would essentially be to determine relative probabilities of the possible sources of inherently uncertain stimuli rather than representing the 'features' in the retinal projection.

As a result of these very different concepts of vision – seeing the 'real world' by directly detecting its features versus seeing probable stimulus sources based on statistical evidence – there is considerable debate about what visual processing is actually doing and how best to interpret the wealth of information about the anatomy and physiology of visual neurons that modern neuroscience has provided during the past 50 years (work that extends far beyond the scope of this article). One camp supposes that the way of resolving this debate in vision science is to continue obtaining more information about the receptive field properties of visual neurons, with the expectation that sooner or later it should be apparent how these properties give rise to visual perception. Another camp has tended to focus on visual percepts as such, asking whether what people actually see can tell us about the strategy of visual processing. In fact, both approaches make sense and will eventually need to come together to understand how percepts are generated. The following discussion focuses on perception and the evidence it provides about how this problem is being solved by the visual system.

The Perception of Surface Qualities

Brightness

A good place to begin is the perception of light and dark elicited by different intensities of light, which is called brightness (technically, brightness refers to the appearance of a light source, such as a light bulb, and lightness to the appearance of a surface; for present purposes, brightness is used in its general inclusive sense). Not the least of the reasons for beginning with brightness is that such percepts are arguably the most fundamental visual qualities that humans see. Vision in us or other animals cannot occur without this perceptual quality, whereas other qualities (e.g., color) are expendable (some animals have color vision and some do not). The approximate physical correlate of brightness percepts is luminance, a physical measure of light intensity made by a photometer and expressed in units such candelas per square meter. Like all percepts, brightness is not subject to direct measurement, and it can only be evaluated indirectly by asking observers to report the appearance of one object or surface relative to that of another.

Some Puzzling Discrepancies between Luminance and Brightness

Given that brightness is the subjective experience of the intensity of a light stimulus, logical assumptions would be that (1) luminance and brightness are directly proportional since increasing the luminance of a stimulus increases the number of photons captured by photoreceptors and (2) two objects in a scene that return the same amount of light to the eye should

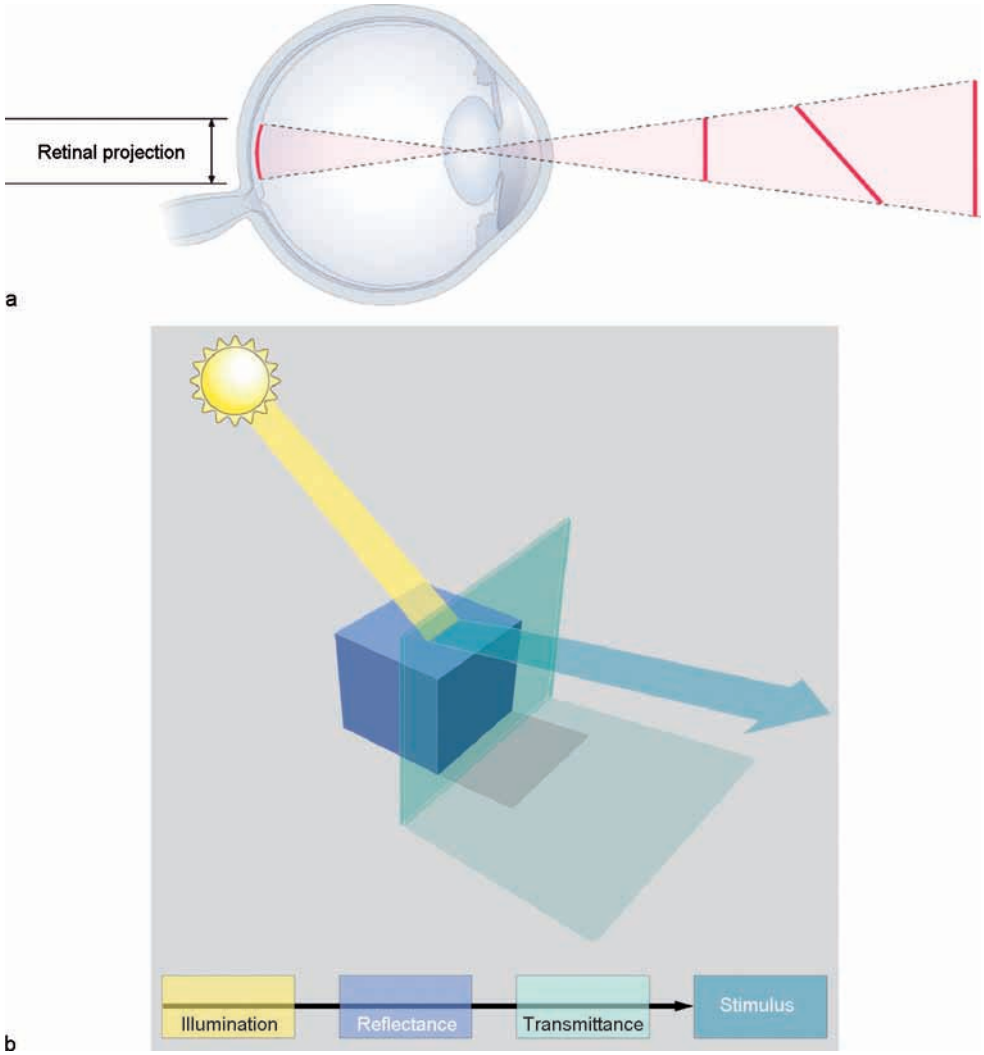

Figure 1 The inverse optics problem. (a) The inherent ambiguity of retinal stimuli, illustrated here in terms of the perception of objects in space. Note that the same linear projection on the retina can derive from an infinite number of linear objects at different distances, of different sizes, and in different spatial orientations. As a result, a retinal image cannot specify its physical source. (b) The inevitable conflation in a light stimulus of illumination, reflectance, and transmittance, the factors that an observer must parse in order to respond appropriately to the pattern of luminance values and spectral power distributions in any visual stimulus.

appear equally bright. It has long been known, however, that perceptions of brightness fail to meet these seemingly simple expectations. For example, two patches returning equal amounts of light to the eye are perceived as looking differently bright when placed on backgrounds that have different luminances (Figure 2(a)). Thus, a patch on a background of relatively low luminance appears somewhat brighter than the same patch on a background of higher luminance, a phenomenon called simultaneous brightness contrast. This effect can be made much more dramatic when the stimulus includes more information about the possible real-world conditions underlying the stimulus (Figure 2(b)), as the German psychologist Adhemar Gelb showed many years ago in a room contrived such that objects could be illuminated by hidden lights.

Until relatively recently, the explanation of this effect was based on the properties of neurons at the input level of the visual system (usually retinal ganglion cells) and the lateral interactions among them. Presumably as a means of enhancing the detection of luminance contrast boundaries (edges), the central region of the receptive fields of lower order visual neurons has a surround of opposite functional polarity (Figure 3). The number of action potentials generated per unit time by neurons whose receptive fields intersect a contrast boundary will therefore differ from the activity of neurons whose receptive fields fall entirely on one side of the boundary or the other.

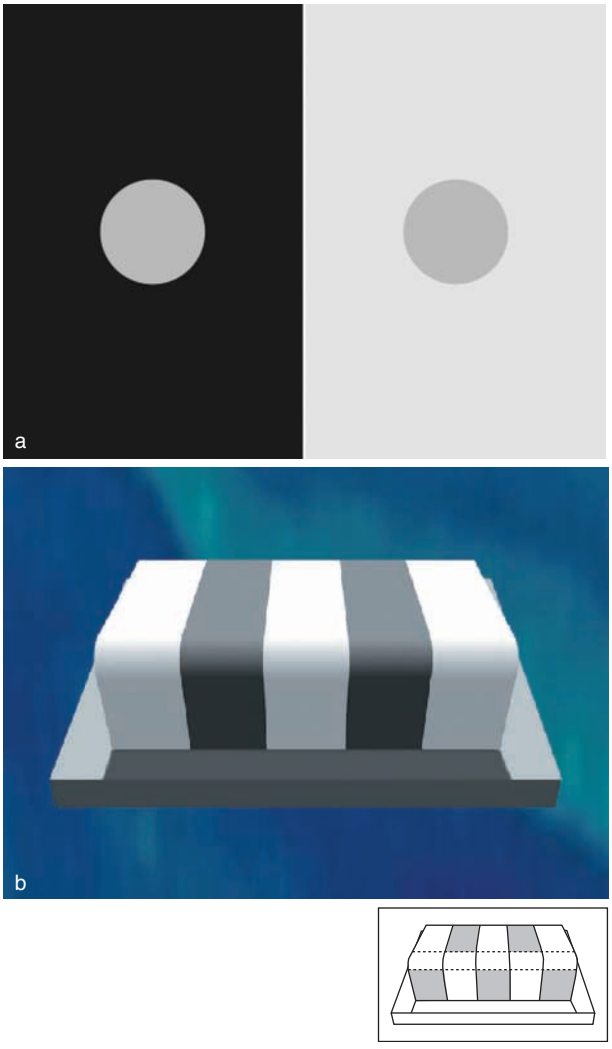

Figure 2 Simultaneous brightness contrast. (a) Standard presentation of this effect; the two circular patches have exactly the same luminance, but the one in the dark surround looks somewhat brighter. (b) Enhanced effect arising from mutually supportive contextual information. Simultaneous brightness contrast can be much greater when the scene is more realistic, as it is here; as shown in the key, the patches that look very different in brightness again have the same luminance. These stimuli show that the amount of light returned to the eye is not a good predictor of the brightness seen. (b) Adapted from Purves D and Lotto RB (2003) *Why We See What We Do: An Empirical Theory of Vision*. Sunderland, MA: Sinauer.

For example, the neurons whose receptive field centers lie just within the target on the dark background in Figure 3 will fire at a higher rate than the neurons whose receptive field centers lie just within the target on the light background (because the former are less inhibited by their oppositely disposed receptive field surrounds than the latter). As a result, many early investigators supposed that the patch on the dark background looks brighter than the patch on the light background because of this difference in retinal output.

The percepts elicited by other stimulus patterns, however, undermine the idea that simultaneous brightness contrast effects are an incidental consequence of an anomalous retinal output in response to edges. In the pattern in Figure 4, for example, the target patches on the left are surrounded by a greater area of higher luminance (lighter) territory than lower luminance and yet appear brighter than the targets on the right, which are surrounded by more lower luminance (darker) territory than higher. Although the average luminance values of the surrounds in this stimulus are effectively opposite those in standard simultaneous brightness stimulus in Figure 3, the brightness differences elicited are approximately the same in both direction and magnitude as in the standard presentation.

If the output of retinal neurons cannot account for the relative brightness values seen in response to stimuli such as the simple patterns in Figures 2 and 4, what then is the explanation? The answer seems likely to be based on the inverse problem as it applies to brightness. In natural circumstances, the luminance of any part of a retinal image is determined by three fundamental aspects of the physical world: the illumination of objects, the reflectance of object surfaces, and the transmittance of the space between the objects and the observer. As indicated in Figure 1(b), however, these factors are inevitably conflated in the retinal image. Thus, an infinite number of different combinations of illumination, reflectance, and transmittance can give rise to the same value of luminance; as a result, there is no logical or direct way in which the visual system can determine how these factors are actually combined to generate a particular retinal luminance value and, thus, what behavior in response to the stimulus is appropriate.

Since successful behavior requires responses that accord with the physical source of a stimulus, this uncertainty means that making brightness percepts directly proportional to luminance values in a stimulus would not provide a guide to successful behavior. A more effective approach to generating brightness percepts – and perhaps the only plausible one – would be to determine the percepts empirically according to the past success or failure observers had experienced interacting with the different combinations of illumination, reflectance, and transmittance generated by natural scenes. In this empirical framework, brightness should always correspond to the relative frequency with which different possible sources had been the source of the same or similar stimuli in the enormous number of visual scenes witnessed by the species during the course of evolution, as well as by

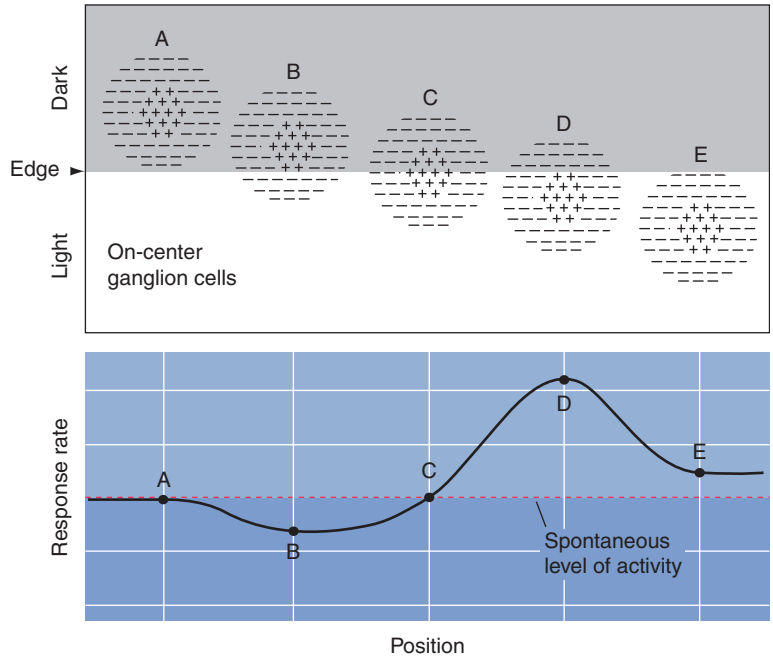

Figure 3 Lateral interactions in the retina and their effect on retinal output. Diagram of the different firing rates of retinal ganglion cells as a function of their position with respect to a light–dark contrast boundary.

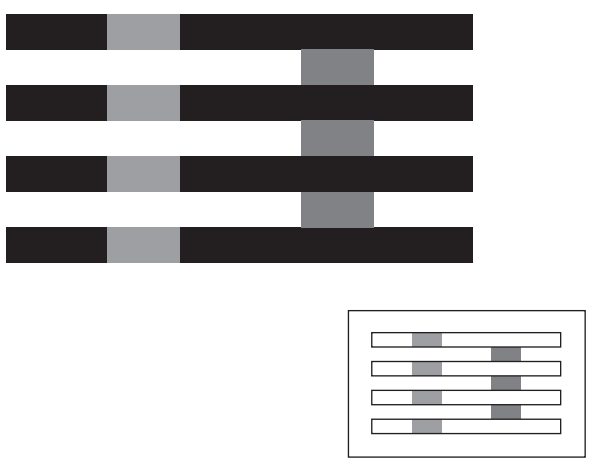

Figure 4 Stimulus pattern that elicits perceptual effects that cannot be explained in terms of lateral interactions arising from local contrast effects. The pattern and its effect are called 'White's illusion' after the psychologist who first described it. Adapted from White M (1979) A new effect of pattern on perceived lightness. *Perception* 8: 413–416.

individual observers during their lifetimes. This idea harks back to Helmholtz's suggestion in the nineteenth century that empirical information might be needed to augment what he took to be the more or less veridical (i.e., 'true to life') information supplied by sensory mechanisms. However, because retinal (and other sensory) receptors cannot provide unambiguous information about the state of the world, the more radical idea currently being examined by cognitive neuroscientists is that visual processing is fundamentally statistical. In this conception, the brightness values seen by an observer should accord with the empirical significance of the stimulus rather than with the physical intensities of light falling on the retina. The biological rationale for this way of seeing brightness is that by using the outcome of experience accumulated by trial and error during phylogenetic and ontogenetic experience (i.e., what percept worked in response to a given stimulus), percepts – and, more important, behaviors – evolve to have an increasingly better chance of successfully representing inherently ambiguous retinal images.

Color

A second major quality of human visual percepts is color. Recall that brightness is defined as the perceptual category elicited by the overall amount of light in a visual stimulus. Color is defined as the perceptual category generated by the distribution of that amount across the visible spectrum (i.e., the relative amount of energy at short, middle, and longer wavelengths across the light spectrum). In fact, the experience of color comprises three perceptual qualities: (1) hue, which is the perception of the relative redness, blueness, greenness, or yellowishness of a stimulus; (2) saturation, which is the degree to which the perception approaches a neutral gray (e.g., a highly unsaturated red is a percept that appears largely gray but

nonetheless has an appreciable reddish tinge); and (3) color brightness, which is the same perceptual category described in the previous section but applied to a stimulus that elicits a discernible hue. Together, these qualities describe a perceptual color space (Figure 5). The ability to see colors has presumably evolved in humans and many other mammals (but not all, showing that the ecological value of color is far less than brightness) because percepts that use information derived from spectral differences allow observers to distinguish surfaces in the natural world more effectively than distinctions made solely on the basis of luminance.

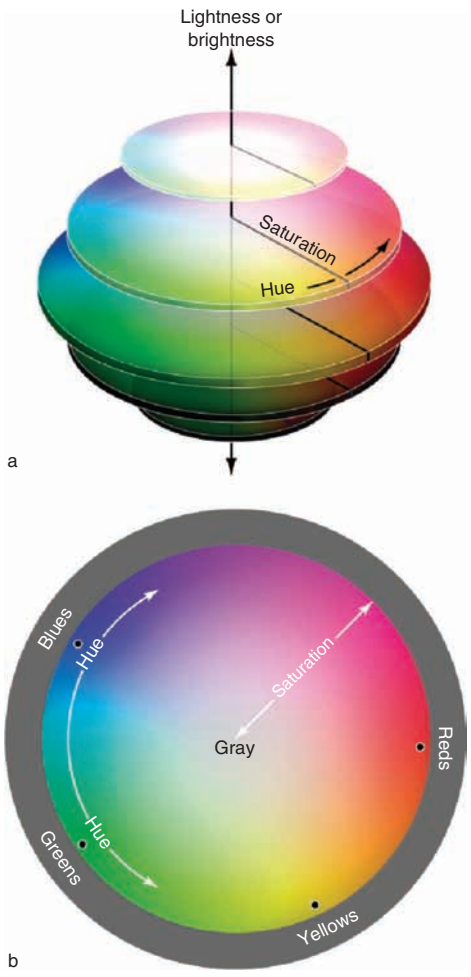

Figure 5 Diagram of perceptual 'color space' viewed from the side (a) and looking down on a single plane (b). At any particular level of light intensity (which evokes a sensation of color brightness), movements around the perimeter of the relevant plane correspond to changes in hue (i.e., changes in the apparent contribution of red, green, blue, or yellow to the percept), whereas movements along the radial axis correspond to changes in saturation (i.e., changes in the approximation of the color to the perception of a neutral gray). Each of the four primary color categories (red, green, blue, and yellow) is characterized by a unique hue (indicated by dots) that has no apparent admixture of the other three (i.e., a color experience that cannot be seen or imagined as a mixture of any other colors). These four colors are considered primary because of their perceptual uniqueness, which has nothing to do with the definition of 'primary colors' one may have learned about in an art class.

Color Contrast and Constancy

Like perceptions of brightness, the colors we see are strongly influenced by context. For example, a stimulus patch generating exactly the same distribution of power at various wavelengths can appear quite different in color depending on its surroundings, a phenomenon termed color contrast. Conversely, patches in a scene returning different spectra to the eye can appear to be much the same color, an effect called color constancy (brightness contrast and constancy are equally apparent) (Figure 6). Although these phenomena were well-known to psychologists and vision scientists in the late nineteenth and early twentieth centuries, they were not paid much attention to until Edwin Land's work in the late1950s. Land – a photochemist who among other achievements invented polarizing filters and instant photography and founded the Polaroid Corporation – used three adjustable illuminators generating short-, middle-, and long-wavelength light to illuminate a collage of colored papers. With a telemetric photometer, he showed that two patches that in white light appeared quite different colors (e.g., green and brown) continued to look their respective colors even when the three illuminators were adjusted so that the light being returned from the 'green' surfaces produced exactly the same readings on the three telephotometers that had previously come from the 'brown' surface – a striking demonstration of color constancy. A simpler demonstration of color constancy is shown in Figure 6.

Color contrast and constancy effects present the same problem for understanding color processing as the contextual brightness effects described previously do for understanding how achromatic percepts are generated; together, these phenomena have led to a debate about brightness and color percepts that now spans more than a century. Land tried to explain such effects by a series of postulated algorithms that integrated the spectral returns of different regions over the entire scene (the implication being that color processing in the visual system implemented these algorithms). It was recognized even before Land's death in 1991, however, that this so-called retinex theory did not work in all circumstances and was in any event a description rather than an explanation. Although it far exceeds the limitations of this article, there is growing evidence that, like

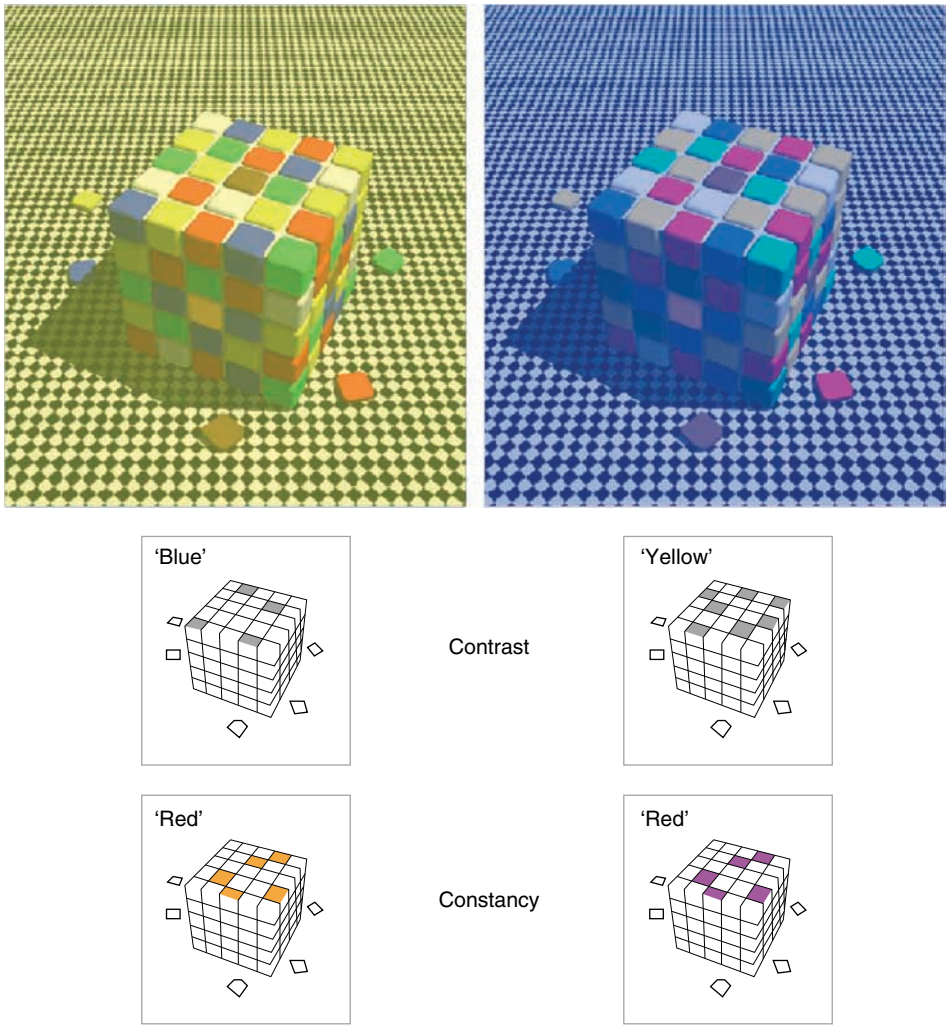

Figure 6 Demonstrations of color contrast and constancy. Color contrast: Identical gray patches on the surface of the two cubes (see upper key) can be made to appear either blue or yellow by changing the context in which they occur. Color constancy: Other patches on the surface that have very different spectra and different color appearances in a neutral setting (see lower key) can be made to look more or less the same color (red or reddish in this case), again by manipulation of contextual information. From Purves D and Lotto RB (2003) *Why We See What We Do: An Empirical Theory of Vision*. Sunderland, MA: Sinauer.

achromatic brightness, perceptions of hue, saturation, and color brightness are generated according to the empirical significance of spectral stimuli in past experience.

The Perception of Surface Form

Another basic quality of visual perception is the form of objects. In the simplest case, perceptions of form entail simple geometrical characteristics such as the length of lines, their apparent orientation, and the angles they make with other lines.

The Apparent Length of a Line

A simple starting point in exploring how the visual system generates perceptions of form is how we perceive the distance between two points in a stimulus, as occurs in the apparent length of a line or the dimensions (size) of any simple geometrical shape. It is logical to suppose that the perception of a given line (e.g., a line drawn on a piece of paper or on a computer screen) should correspond more or less directly to its projected length in the retinal image. However, as in the case of brightness and color, this expectation is not met. A well-studied example is the variation in the perceived length of a line as a function of its orientation (Figure 7(a)). As investigators have repeatedly shown during the past 150 years, a line oriented more or less vertically in the retinal image appears to be significantly longer than a horizontal line of the same length, with the maximum length being seen when the stimulus is oriented approximately 30° from vertical (Figure 7(b)). This effect is evidently a particular manifestation of a general tendency to

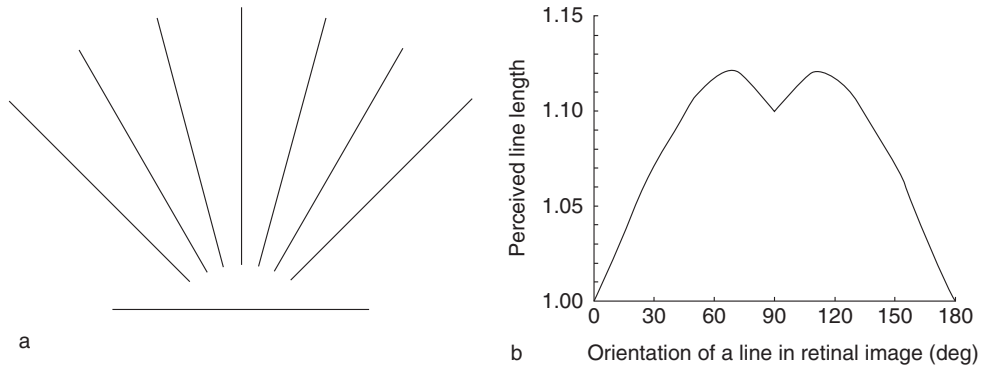

Figure 7 Variation in apparent line length as a function of orientation. (a) The horizontal line in this figure looks significantly shorter than the vertical or oblique lines, despite the fact that all the lines are identical in length. (b) Quantitative assessment of the apparent length of a line reported by subjects as a function of its orientation in the retinal image (orientation is expressed as the angle between the line and the horizontal axis). The maximum length seen by observers occurs when the line is oriented approximately 30° from vertical, at which point it appears approximately 10–15% longer than the minimum length seen when the stimulus is horizontal.

perceive the extent of any spatial interval differently as a function of its orientation in the retinal image. For instance, as the psychologist Wilhelm Wundt first showed in 1862, the apparent distance between a pair of dots varies systematically with the orientation of an imaginary line between them, and a perfect square or circle appears to be slightly elongated along its vertical axis.

The Perception of More Complex Aspects of Geometry

There is a rich literature on other perceptual distortions elicited by geometrical stimuli, showing in each case that measurements made with instruments such as rulers or protractors are often at odds with the corresponding percepts. Some of the most familiar geometrical illusions – and the ones whose etiology has been most hotly debated – are illustrated in Figure 8. The first example is attributed to Hering, who showed that two parallel lines (indicated in red in Figure 8(a)) appear bowed away from each other when presented on a background of converging lines. In the Poggendorff illusion, the continuation of a line interrupted by a bar appears to be displaced vertically, even though the two line segments are actually collinear (Figure 8(b)). In the inverted-T illusion (Figure 8(c)) the vertical line looks somewhat longer than the horizontal line (see also Figure 7). In the more complex Müller–Lyer illusion, the line terminated by arrow tails looks longer than the same line terminated by arrowheads (Figure 8(d)). In the Ponzo illusion, the upper horizontal line appears longer than the lower one, despite the fact that they are identical (Figure 8(e)). It would be a mistake to conclude that because these stimuli are simple figures that bear little resemblance to real-world objects, they are pertinent to natural scenes. In the tabletop illusion created by Stanford psychologist Roger Shepard, the point is again made that such effects are actually more dramatic when the stimulus is made more realistic (Figure 8(f)).

Studies that have measured the geometry of scenes by laser scanning suggest that these classical geometrical illusions are determined by the statistical distribution of the sources of such geometries in the natural scenes that humans have always been exposed to, consistent with the empirical explanations of brightness and color percepts described previously. The rationale for this interpretation is based on the fact that in the perception of form, as in the perception of brightness and color, the sources of the relevant stimuli are inherently uncertain and that stimulus meanings must therefore be adduced according to past experience.

Conclusions

A basic challenge in understanding vision is explaining how inherently uncertain retinal stimuli give rise to definite perceptions and generally successful visually guided behavior. In each the major visual perceptual domains (e.g., brightness, color, and form), the evidence gleaned over many years increasingly points to a fundamentally empirical solution to the inverse optics problem, which is arguably the central challenge in biological vision. The idea that what we see in response to inherently ambiguous retinal images is a statistical manifestation of accumulated past experience rather than a logical analysis of retinal image features *per se* runs counter to our intuitions about vision. Nevertheless, the phenomenology of what we actually see is difficult to explain in any other way. The biological advantage of generating vision in this manner is that percepts – and the visual circuitry that underlies them – will have progressively evolved to

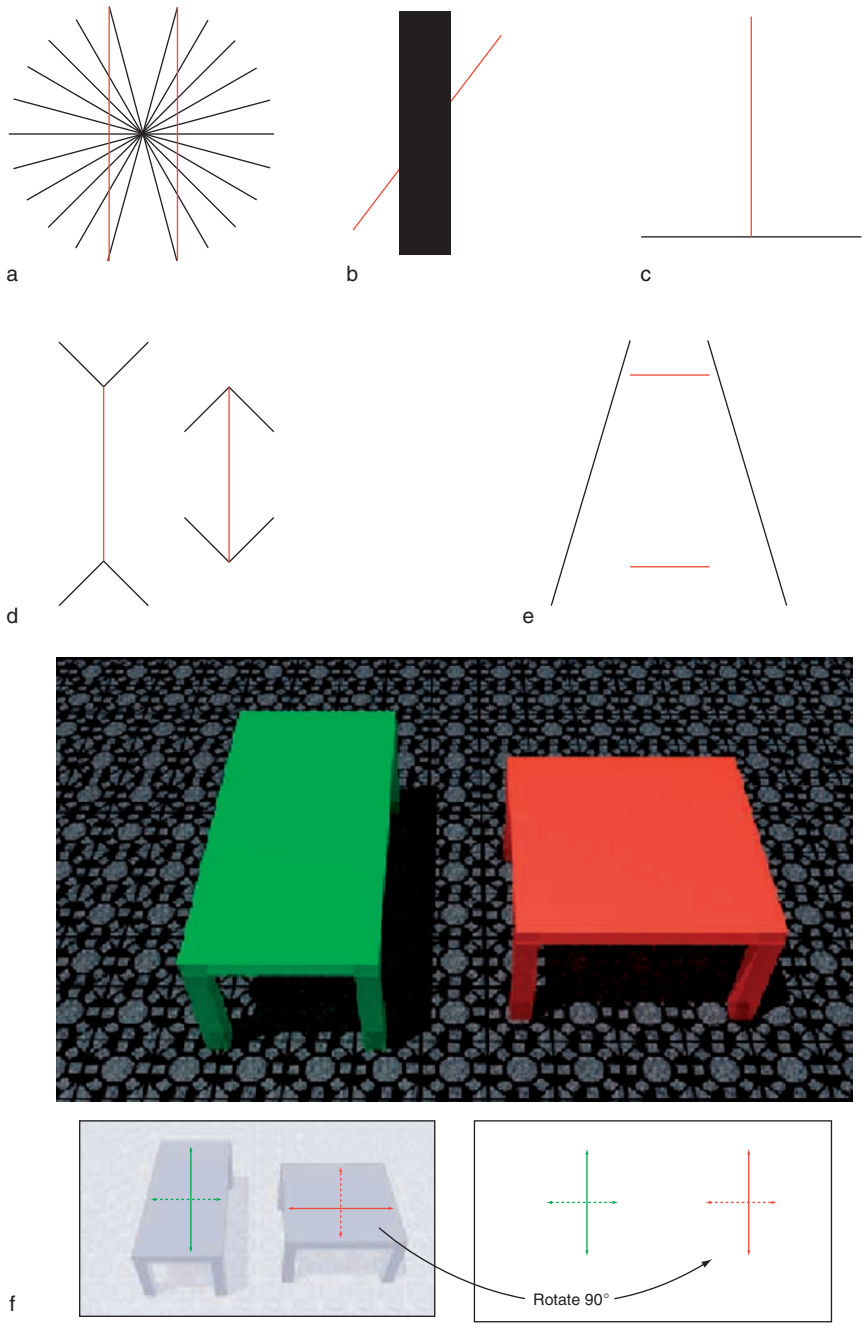

Figure 8 Examples of some much-studied geometrical illusions: (a) the Hering illusion, (b) the Poggendorff illusion, (c) the inverted-T illusion, (d) the Müller–Lyer illusion, (e) the Ponzo illusion, and (f) the tabletop illusion.

incorporate the vast amount of information derived from the experience of both the species and the individual, much to our advantage. The implication of this framework for understanding visual perception is that the main function of visual processing is instantiating and modulating probabilities that link images to their real-world sources.

See also: Active Perception; Contextual Interactions in Visual Perception; Contextual Interactions in Visual Processing; Perception and Eye Movements; Perceptual Learning and Sensory Plasticity; Retinal Color Mechanisms; Statistical Analysis of Visual Perception; Vision for Action and Perception; Vision: Surface Segmentation.

Further Reading

Adelson EH (2000) Light perception and lightness illusions. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences*, 2nd edn., pp. 339–351. Cambridge: MIT Press.

Berkeley G (1709/1976) *A New Theory of Vision*. New York: Everyman's Library.

Cornsweet TN (1970) *Visual Perception*. New York: Academic Press.

Farah MJ (2000) *The Cognitive Neuroscience of Vision*. Oxford, UK: Blackwell.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Erlbaum.

Helmholtz HLFv (1866/1924) *Helmholtz's Treatise on Physiological Optics*. Washington, DC: Optical Society of America.

Howe CQ, Lotto RB, and Purves D (2006) Empirical approaches to understanding visual perception. *Journal of Theoretical Biology* 241: 866–875.

Howe CQ and Purves D (2005) *Perceiving Geometry: Geometrical Illusions Explained by Natural Scene Statistics*. New York: Springer.

Hubel DH (1988) *Eye, Brain and Vision*. New York: Freeman.

Land EH (1986) Recent advances in retinex theory. *Vision Research* 26: 7–21.

Maloney LT (2002) Statistical decision theory and biological vision. In: Heyer D and Mausfeld R (eds.) *Perception and the Physical World: Psychological and Philosophical Issues in Perception*, pp. 145–189. New York: Wiley.

Marr D (1982) *Vision: A Computational Investigation into Human Representation and Processing of Visual Information*. San Francisco: Freeman.

Milner AD and Goodale MA (1995) *The Visual Brain in Action*. Oxford: Oxford University Press.

Mollon JD (1995) Seeing colour. In: Lamb T and Bourriau J (eds.) *Colour: Art & Science*, pp. 127–150. Cambridge, UK: Cambridge University Press.

Purves D and Lotto RB (2003) *Why We See What We Do: An Empirical Theory of Vision*. Sunderland, MA: Sinauer.

Purves D, Williams SM, Nundy S, and Lotto RB (2004) Perceiving the intensity of light. *Psychological Review* 111(1): 142–158.

Rao RPN, Olshausen BA and Lewicki MS (eds.) (2002) *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge: MIT Press.

Robinson JO (1998) *The Psychology of Visual Illusions*. New York: Dover. [Corrected republication of the 1972 edition published by Hutchinson and Co. in England].

Rock I (1984/1995) *Perception*. New York: Freeman.

Simoncelli EP and Olshausen BA (2001) Natural images statistics and neural representation. *Annual Review of Neuroscience* 24: 1193–1216.

Wheatstone C (1838) Contributions to the physiology of vision. I: On some remarkable and hitherto unobserved phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London Series B* 128: 371–394.

White M (1979) A new effect of pattern on perceived lightness. *Perception* 18: 413–416.

Yang Z and Purves D (2004) The statistical structure of natural light patterns determines perceived light intensity. *Proceedings of the National Academy of Sciences of the United States of America* 101: 8745–8750.

Relevant Website

http://www.purveslab.net – Provides links to numerous relevant sites.

Perceptual Learning and Sensory Plasticity

M Fahle, Bremen University, Bremen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Perceptual Learning and Sensory Plasticity: Definitions and Overview

Humans are able to adapt to changing environments through several mechanisms, prominent among them sensory plasticity and learning. Perceptual learning (PL) as a special case of learning is defined as a long-lasting improvement in detecting, discriminating, or categorizing sensory stimuli based on preceding experience and leading to unconscious inference, in the nomenclature of Helmholtz. PL extracts previously unused sensory information and reduces unwanted noise. Thus, it improves the cortical representation of a stimulus, leading to a better detection. It also improves the task of discriminating a stimulus from other stimuli and hence improves stimulus categorization, into categories that also may be learned. As one form of sensory plasticity, it changes both perceptual performance and brain physiology, improving the representation of the outer world in the brains of humans and animals and modifying behavior, usually toward improvement. Sensory plasticity in general coordinates and normalizes the perception of different types of sensory stimuli, for example, by adjusting the neuronal gain to coordinate the reference systems of different sense modalities. The term 'perceptual learning' is more often used in relation to the systemic level – that is, to changes in behavior – while 'sensory plasticity' often denotes changes on the neuronal level. Here, 'sensory plasticity' is used as a more general term including both behavioral and physiological changes in perceptual learning, sensomotor control, and long-term adaptation as a result of experience.

The improvement in PL tends to persist over weeks and months, distinguishing PL from short-term adaptation, sensitization, habituation, attention, and priming, which all produce more transient changes of performance. PL is a form of learning not leading to 'knowing that' and lacking memory traces that can be true or false. Hence, PL is of the procedural type, forming implicit memory traces as opposed to explicit memory, and cannot be communicated to others, while declarative forms of memory can. PL appears to directly modify the neuronal mechanisms underlying the processing of the task required, in contrast to declarative forms of learning. These declarative forms of learning usually lead to memory traces, stored in specialized brain regions such as the hippocampus that are clearly separated from the cortical structures analyzing the stimuli. In contrast to associative forms of learning, PL does not normally bind together contents that were initially separated, as does conditioning, but it improves both the detection and the discrimination of stimuli that could not be detected or discriminated before the training, and coordinates signals from different sense organs.

Through PL, we all learn to extract and discriminate the most relevant features from the multitude of signals supplied by the sense organs, and this sensory plasticity continues throughout life to adapt cortical processing on all levels to new situations. Usually, PL requires attention to the target, thus reserving learning to important sensory tasks, while some learning may also occur without attending to the specific stimulus feature due to general arousal. Hence, on a phenomenological level, PL and attention lead to similar results, both improving performance in perceptual tasks. But the effects of attention are similar to those of habituation and sensitization in that they do not permanently change behavior – as soon as attention is moved to another position or object, performance drops to the initial level. Perceptual learning, on the other hand, improves the processing of stimuli throughout their cortical processing chain in a way that leaves permanent traces. Even after months without training, a perceptual task that has been learned yields better results than another one that was never trained.

Sensory plasticity and PL take place on quite a number of different cortical levels, even on rather early ones. An important distinction between different types of PL is whether the improvement generalizes to stimuli similar to those practiced – a distinction closely related to the question of the cortical mechanisms underlying improved performance. 'Early' stages of processing are more specific for low-level features such as position and orientation of a stimulus. Therefore, a high specificity of PL with little transfer to other orientations or positions indicates a neuronal substrate at early cortical levels, while generalization over stimulus position and orientation is more compatible with a higher level of neuronal changes. The results of psychophysical, electrophysiological, and imaging studies indeed show an involvement of early sensory cortices in PL, with feedback from 'higher' to 'lower' levels of processing. PL involving rather early levels of cortical processing, levels that have been well investigated, thus allows one to probe the mechanisms underlying learning in a relatively well-known part of the brain, and in a brain area that is directly accessible to sensory stimulation.

Stimulus Specificity in 'Early' Perceptual Learning: 'Early Selection'

A perceptual task well suited to test PL is vernier acuity, a type of visual hyperacuity, in which even untrained observers attain discrimination thresholds around 10 arcsec – thresholds which are at least slightly below the spacing of foveal photoreceptors. Through training, thresholds can improve down to 2–3 arcsec. Obviously, performance in these tasks is not limited primarily by the optics of the eye nor by the photoreceptor spacing, but rather by the signal-to-noise ratio of neuronal processing. Hyperacuity, being a very sensitive measure, thus is well suited for studying learning processes in visual perception.

Performance typically improves fast for about 10–20 min of training, slowing down thereafter (see Figures 1). Improvement usually does not generalize to a stimulus rotated by 90° for low-level stimuli (Figures 1(a) and 1(b)), and rotation by as little as 4° is sufficient to decrease performance and increase thresholds (Figure 1(b)) virtually to baseline in vernier discrimination. Improvement is similarly specific for position in the visual field. Training of vernier discrimination at a position in the visual periphery improved performance, on average, by 7% (e.g., increasing from 80% to 87% correct responses). Stimuli subsequently presented at a new visual field position yielded 7% fewer correct responses. Hence improvement did not transfer at all between these visual field positions. Performance improves also monocularly in a number of low-level tasks, similar to binocular training. However, improvement does not fully transfer to the contralateral eye; learning is at least partly specific for the trained eye (Figure 1(c)). In many tasks, training improves performance up to 30% within an hour, but large interindividual differences exist even in quite homogeneous groups of individuals, such as students. The causes for these differences in the ability to improve perception through training are so far poorly understood.

PL is specific for a number of stimulus parameters, such as motion speed and motion direction, improving motion discrimination even with 'paired-dot' stimuli. This finding again indicates that improvement may occur on early levels of processing, since these paired-dot stimuli are supposed not to activate the medial temporal (MT) area, the cortical area most involved in motion detection. Improvement of observers training in a shape discrimination task based on luminance contrast transfers to the detection of a shape defined by temporal structure, but not the other way round, possibly indicating a sequential order in cortical processing. Improvement with a visually masked stimulus is specific for the mask, hence caution is required when judging the specificity of improvement based on masked stimuli.

Even detection of luminance contrasts improves through training, and – unlike vernier discrimination – generalizes across orientations. This finding is surprising, given the fact that humans should be overtrained for contrast detection through everyday life, especially under low illuminances. Obviously, training of some tasks is orientation specific, while generalization occurs in others, indicating different mechanisms underlying PL, with generalization of learning being more pronounced on later (cortical) levels. High stimulus specificity can be considered as a sign for modifications on early levels of processing, selecting the relevant information on these early levels.

Generalization of Improvement in Late Visual Perceptual Learning and 'Late Selection'

PL is not restricted to the simple types of stimuli described in the preceding section, but can also involve learning to discriminate between quite complex classes of stimuli. Examples of more complex tasks include the ability to discriminate, say, between the histology of different tissues, or between different types of wines. Individuals identify stimuli as belonging to a specific class or category, and these categories can even be formed through PL, hence these are not necessarily innate processes (see next section).

Often, the amount of generalization is higher for more complex tasks than for simpler ones (and increases greatly with noisy stimuli), indicating a selection of the relevant information only at a later stage of cortical processing. For example, playing action video games partly generalizes to other tasks, and in visual search two aspects are learned: first, specific stimulus features (not generalizing to other stimuli), and second, a global search strategy that generalizes to other stimuli. This is to say that perceptual learning improves the detection of stimuli as well as their discrimination and categorization, not only by improving their cortical representations and by better defining discriminative features, but also because PL contains a learning of the complex setting for the task at hand, including the type of decision and response strategies required. These learning processes probably take place on a rather cognitive level of processing.

Visual training through reading text seems to modify the way people perceive printed words: reading-related training improves perception of words but not of nonwords. Moreover, the size of the visual span (the range around the center of gaze in which letters

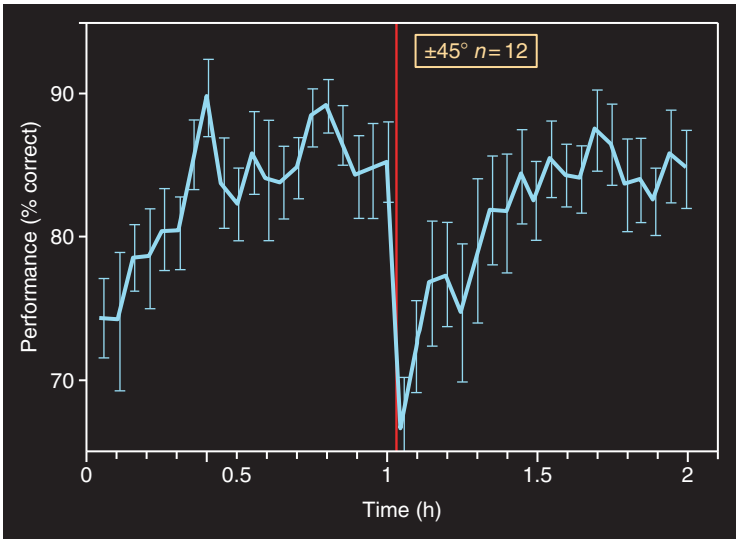

a

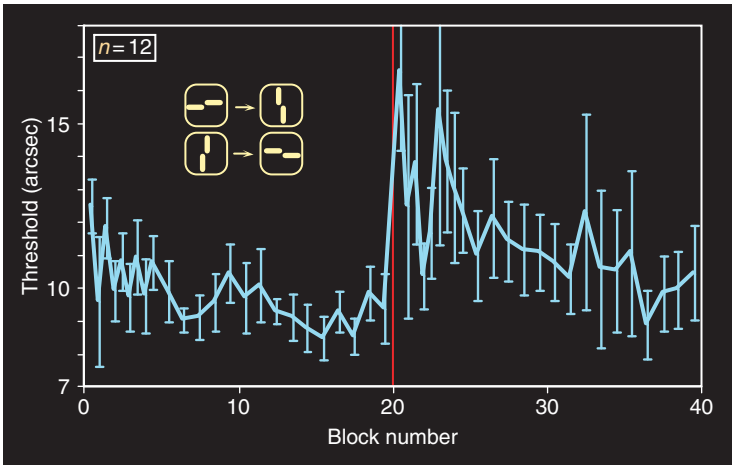

b

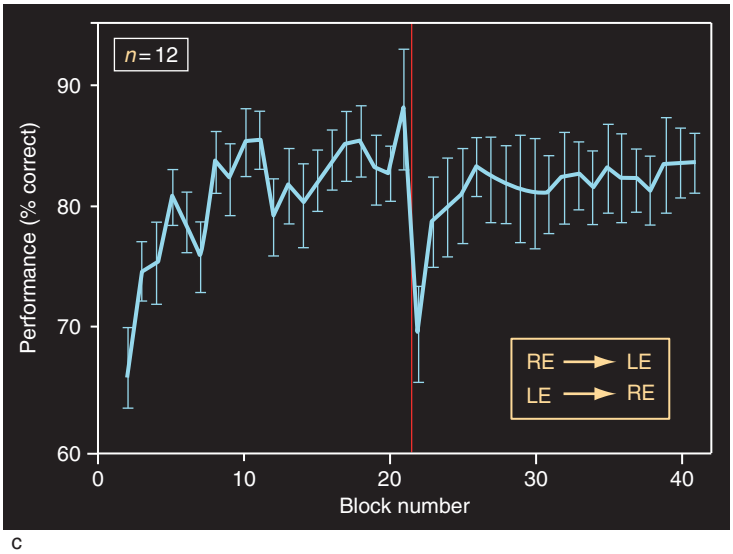

c

Figure 1 Continued

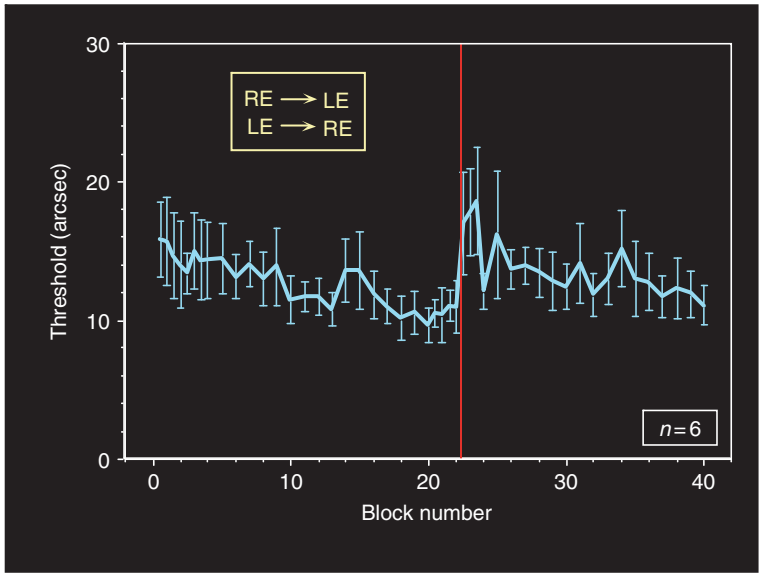

d

Figure 1 Improvement through perceptual learning in a vernier discrimination task. (a) Mean performance of 12 observers initially improved rapidly in a 1 h training session, but rotation of the stimulus by 90°, indicated by the vertical red line, significantly decreased performance – the improvement achieved through training does not transfer to the new stimulus orientation. (b) Mean thresholds of another 12 observers gradually improved over five 1 h training sessions, but massively deteriorated after stimulus rotation. (c) Performance also improved for monocular vision, for example, with the left eye (LE) patched. However, there was hardly any transfer of improvement when the patch was moved to the right eye (RE). (d) The same specificity was found for thresholds in the condition with 5 h of training for each eye in another six observers. (a) Reproduced Fahle M and Edelman S (1993) Long-term learning in hyperacuity. *Vision Research* 33: 397–412, with permission from Elsevier. (c) Reprinted from Fahle M, Edelman S, and Poggio T (1995) Fast perceptual learning in hyperacuity. *Vision Research* 35: 3003–3013, with permission from Elsevier.

are recognized) expands through training, with an associated increase in maximum reading speed. These findings encourage the use of PL in visual rehabilitation.

Formation of New Visual Categories

Categorical perception denotes the capability of observers to better discriminate between members of different categories, as compared to members of the same category that are equally different. Hence, categorical perception emphasizes both the differences between categories and the similarities within categories, and it is often cited as indicating cognitive influences on perception. Categorical learning improves discrimination in domains such as color perception and shape discrimination, for example, in compound gratings (see Figure 2). In color perception, the discrimination between green tiles and blue tiles is easier than it is between tiles of different shades of green, even if the differences between all of the tiles in physical terms (i.e., in terms of wavelength) are of equal size. In the more artificial discrimination between different gratings, observers discriminate better – after a short training phase – between stimulus pairs from different categories just learned. They detect differences between gratings belonging to different categories more often than before the training, but this is true only for the trained grating orientation, not for the perpendicular orientation. Hence, training of categorical perception improves performance only for pairs of stimuli belonging to different categories and only for the trained stimulus orientation, with a rather sharp decline even for small differences in orientation. The high specificity of categorical learning indicates involvement of 'early' mechanisms unable to generalize across different stimulus orientations. On the other hand, another form of 'category' learning – namely, pop-out in search tasks – can be learned and generalized between different types of stimuli.

Learning in Other Sensory Modalities

PL is by no means restricted to the visual domain. The fast improvement with training auditory tasks is, at least in tone frequency discrimination, mostly due to PL. Discrimination of temporal intervals is relatively specific for interval length, but generalizes to other spectral (tone) frequencies. Similarly, as in the visual system, noise is eliminated at an early level, hence loss of efferent feedback (to the cochlea) degrades auditory perception in noisy stimuli. Reorganization of primary auditory cortex and other brain regions was found after some forms of PL but not in others. A fast

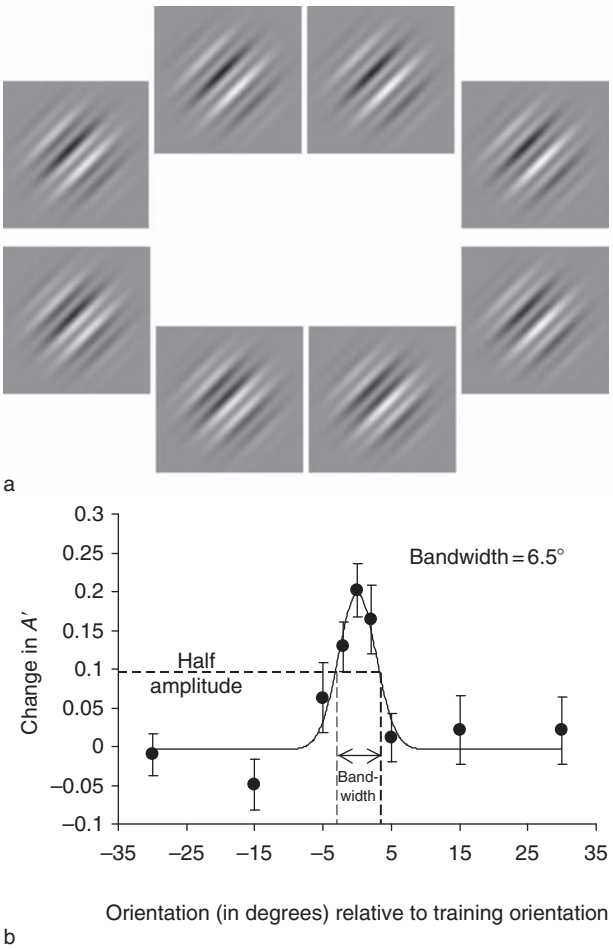

Figure 2 Perceptual learning of new categories. (a) Observers learned to discriminate gratings in the upper half of the figure from those in the lower half of the figure. Discrimination between any two 'neighbors' was equally good initially, but it was better between members of the upper versus the lower half than between members of the same group. This superiority of between-category discrimination disappeared for relatively small rotations of the gratings (b). The graph indicates the orientation bandwidth of the transfer between orientations. (b) Reproduced from Notman LA, Sowden PT, and Ozgen E (2005) The nature of learned categorical perception effects: A psychophysical approach. *Cognition* 95(2): B1–B14, with permission from Elsevier.

reorganization was coupled with high sensory thresholds, while thresholds for the 'reorganized' frequencies were close to normal after several months.

PL induces strong changes in the motor and somatosensory cortices, with significant improvement of two-point tactile resolution and enlargement of the cortical representations of those body parts used during training, possibly due to Hebbian coactivation. Even Hebbian coactivation by means of transcranial magnetic cortical stimulation coupled with tactile stimulation improves tactile discrimination ability, and tactile coactivation induces significant changes in somatosensory cortex. Olfactory learning may rely on changes in the odorant receptive fields of second- and/or third-order olfactory neurons of animals, hence again on modifications on a relatively peripheral level of processing, leading to the synthesis of experience-dependent receptive fields as a result of PL. Finally, PL improves 'taste', for example, by significantly lowering the detection threshold for salt in rats.

Psychophysical Evidence Leading to Models of Perceptual Learning

The steadily increasing amounts of data and insights in PL allow the formulation of ever-improving models of PL, incorporating both specificity and generalization of learning. Important features of these models are the implementation of recurrent (feedback) connections, incorporating internal and external noise, and assessing the change of internal templates. The models show that (1) the reported sharpening of orientation tuning curves can account for the psychophysical data, and that (2) PL seems to improve elimination of external noise, possibly by (3) retuning the internal templates.

At least two different mechanisms seem to achieve PL. According to the reverse hierarchy model, learning starts with high speed at high cortical areas (where it generalizes) and progresses backward toward lower levels if necessary (where it is specific and requires more time). A slightly different view contrasts an early selection theory of PL with a late selection, in analogy to theories on attention, and argues that selection should occur as early as possible to optimize signal-to-noise ratio. The specificity of the slow type of PL for stimulus features such as the exact stimulus orientation, position, and even the eye trained under monocular conditions suggests such an involvement of 'early' stages of cortical information processing, where neurons are specific for visual field position and are activated monocularly, as in the primary visual cortex, V1 (Figure 3). Hence, PL seems to involve changes in functional connectivity between neurons already on the level of the primary visual cortex.

In one respect, such a fine-tuning of information processing on as early a level as possible seems reasonable, removing irrelevant noise as early as possible during cortical processing. In another respect, however, the fine-tuning of peripheral levels of processing is unacceptable for a mature nervous system, since learning of a new task would modify the neuronal front end used to process all stimuli. Hence learning on an early level would interfere with performance in

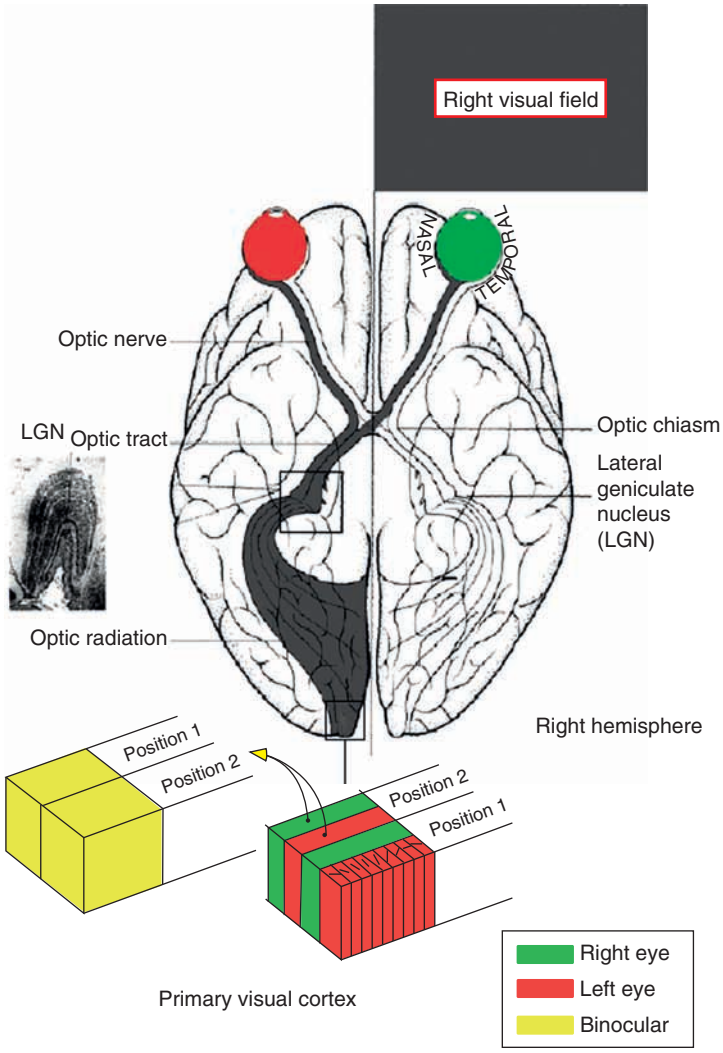

Figure 3 Hypothetical model for the cortical location of perceptual learning, based on the specificity of perceptual learning on stimulus orientation and the eye trained. Since receptive fields in the retina and thalamus are rotation symmetric, orientation-specific learning would be difficult to realize there. Since single-cell recordings in visual cortices find most neurons beyond the primary visual cortex to be binocularly activated, eye specificity of improvement is not trivial to achieve there. Adapted from Fahle M (2004) Perceptual learning: A case for early selection. *Journal of vision* 4: 879–890, with permission from ARVO.

other tasks. For example, shallow luminance gradients are best detected by large receptive fields, which in turn are unable to detect fine gratings. Training to detect fine gratings would therefore improve their detection by decreasing receptive field size, but would deteriorate detection of shallow gradients. A solution of this apparent paradox would be a task-dependent switching of early cortical signal processing under top-down influence. For example, the balance between (excitatory) receptive field centers and (inhibitory) surrounds could be changed under top-down control, or else the gain of feed-forward connections could be modified, or lateral inhibition on the level of the early sensory cortices may be switched on and off or varied in strength. According to this view, filtering and processing on early cortical stages would be chosen from a repertoire of previously learned alternatives under the influence of top-down control. Of course, some learning may take place exclusively on higher processing levels.

The fast and general mechanism, representing more 'conventional' forms of learning, may be located in more central, or 'late', parts of sensory cortices, such as the parietal and temporal lobes, while the slow and specific mechanism seems to involve early cortical areas, including the primary sensory cortices. Perceptual learning and motor learning are actually similar in several respects and may both rely on changes in neuronal tuning functions.

Generalization is at the heart of useful learning in both systems, and apparently relies on strong top-down influences. It will be an important challenge

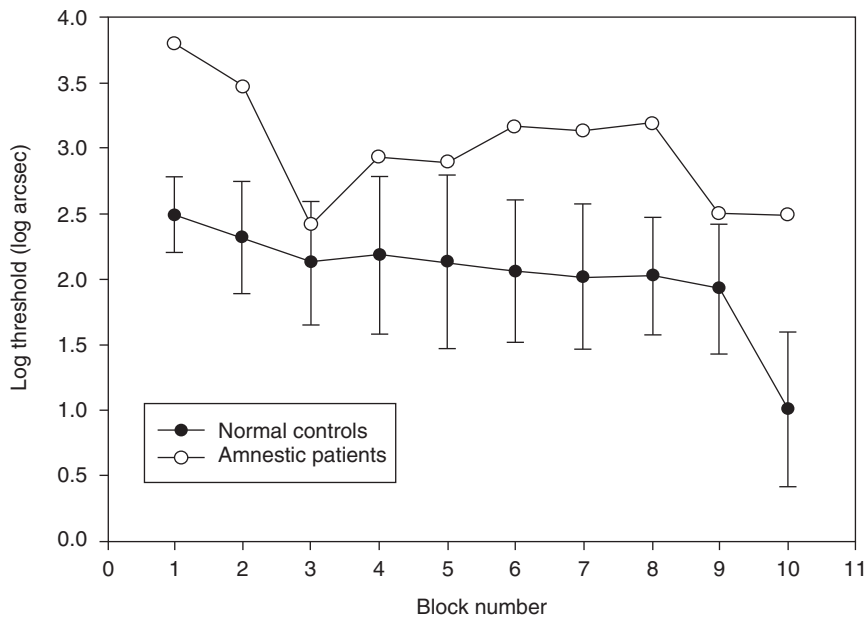

Figure 4 Mean thresholds of six amnesic patients who were hardly able to remember the task required throughout one session, but who nevertheless improved significantly over the course of ten blocks of training on a hyperacuity task; performance was much poorer than that in normal controls. Reprinted from Fahle M and Daum I (2002) Perceptual learning in amnesia. *Neuropsychologia* 40: 1167–1172, with permission from Elsevier.

for research in perceptual learning to find conditions of training, allowing for at least some generalization of the improvement attained through training.

A study of six patients suffering from amnesic syndrome supports the hypothesis of a 'low' rather than cognitive level of the underlying mechanisms. Two of the six patients tested clearly improved performance within as few as two sessions with five blocks of 80 presentations each, tested at a 1 week interval (Figure 4). Another two patients improved somewhat, while performance of the remaining two patients stayed constant – similar to the results of about 15% of the healthy student population of around 300 observers we have tested so far. While the patients had no recollection at all after the 1 week gap following the first session that they had participated in such an experiment (and did not remember the experimenter), performance of the six patients as a group improved significantly as a result of training. This finding indicates that PL does not require the neuronal circuits underlying explicit or declarative learning. Finding the reasons why some persons strongly improve through learning remains another challenge for research in perceptual learning. A first hint is that especially those persons improve fast who start from a below-average level of performance, though the correlation between speed of improvement and initial level of performance is weak.

Possible Mechanisms of Sensory Plasticity

Receptive fields of neurons in the primary visual cortex typically consist of an elongated excitatory field center that determines the orientation specificity of the neuron. The neuron is most strongly activated by light falling on the receptive field center without extending on the inhibitory surround. The receptive fields of neurons with different orientation preferences and slightly differing receptive field positions are clearly able to discriminate between a straight vernier and an offset one. The precision of discrimination depends, among other factors, on the width of the receptive field center. So a straightforward and plausible hypothesis regarding the neuronal changes underlying PL with vernier stimuli would be that learning modifies the width of receptive field centers and hence their orientation bandwidth, as indicated previously (see Figure 5). The orientation half-width of neurons in the primary visual cortex seems to vary between around 5° and 40°, while that on later cortical stages is still wider. The orientation half-width of perceptual learning, on the other hand, is around 5°, corresponding better to early than to late levels of cortical processing. So far, there are no clear indications that orientation bandwidth of neurons changes as a result of training.

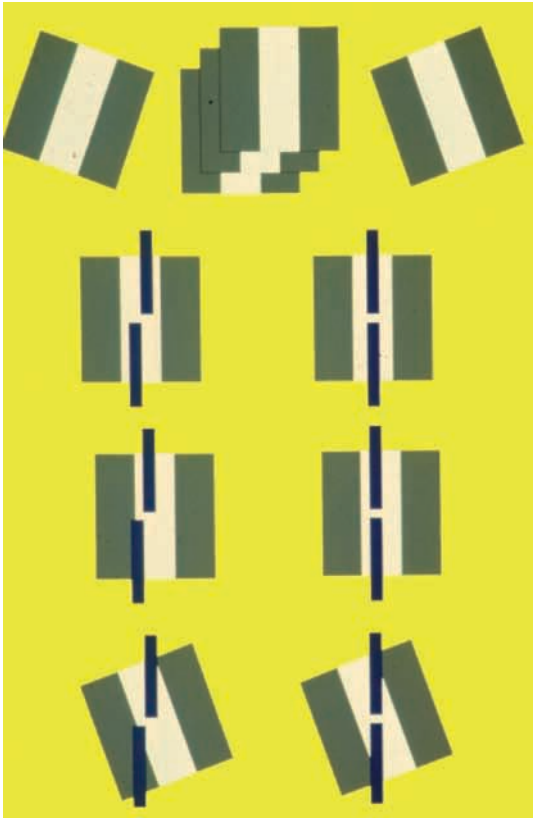

a

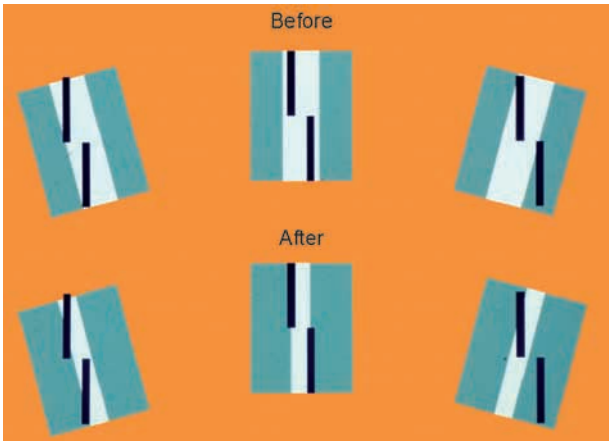

b

Figure 5 A hypothetical model of perceptual learning based on changes in the width of the excitatory receptive field center of early visual cortex neurons. (a) The offset in vernier stimuli is detected by receptive fields with different preferred orientations and different visual field positions (see text). (b) Modifying (here narrowing) the width of the receptive field center may improve detection of vernier offsets. These changes would have to be implemented in a task-dependent way under top-down influence. Adapted from Fahle M (2002) Learning to perceive below the foveal photo receptor spacing. In Fahle M and Poggio T (eds.) *Perceptual Learning,* pp. 197–218. Cambridge, MA: MIT Press, with permission from The MIT press.

Physiological Evidence for Sensory Plasticity

What are the neuronal cortical mechanisms underlying the improvement of vernier discrimination through training, leading to PL? It is notoriously difficult to indicate the level of neuronal plasticity on the basis of psychophysical results. Other methods are of help here, and, indeed, both single-cell and sum-potential recordings clearly show plasticity at early levels of sensory processing. While 20 years ago the primary visual cortex appeared as a hard-wired first stage of analysis, more recent electrophysiological evidence points to plasticity even in the adult lateral geniculate nucleus and primary visual cortex, supporting the hypothesis that PL may involve plasticity even of primary visual cortex. However, plasticity on early levels may be less pronounced in the visual system than in other sensory systems.

An early study found changes in early evoked sum potentials of humans as a result of PL, and several others corroborated the basic finding of training-induced changes in evoked potentials. Training an orientation discrimination task, surprisingly, decreases in monkeys the number of neurons in primary visual cortex representing the trained orientation, without any evident changes in receptive field properties there or in inferior temporal cortex. However, neurons in V4, with receptive fields in the trained region of the visual field, narrow their orientation tuning and increase responses as a result of training. Moreover, neurons in V1 change not only their contextual influences but also the classical receptive field properties, depending on the animal's actual task, thereby optimizing the information on the relevant stimulus feature under top-down influence. Similarly, neurons in inferotemporal cortex can show target-selective neuronal responses during visual search. Neurons in V4 respond more strongly to noisy images the monkey has been trained with than to untrained images containing the same amount of noise. Hence, training improves both perception and single-cell responses that are noise specific. In humans, training of visual texture discrimination increases activity in the corresponding quadrant of the visual field representation in V1, as demonstrated by functional magnetic resonance imaging (fMRI), compatible with the hypothesis of modification of 'early' sensory processing.

The neuronal networks modified by PL are influenced by pharmacological agents, as well as by transcranial magnetic stimulation. For example, improvement by coactivation in tactile two-point discrimination is eliminated by lorazepam, a γ-aminobutyric acid receptor

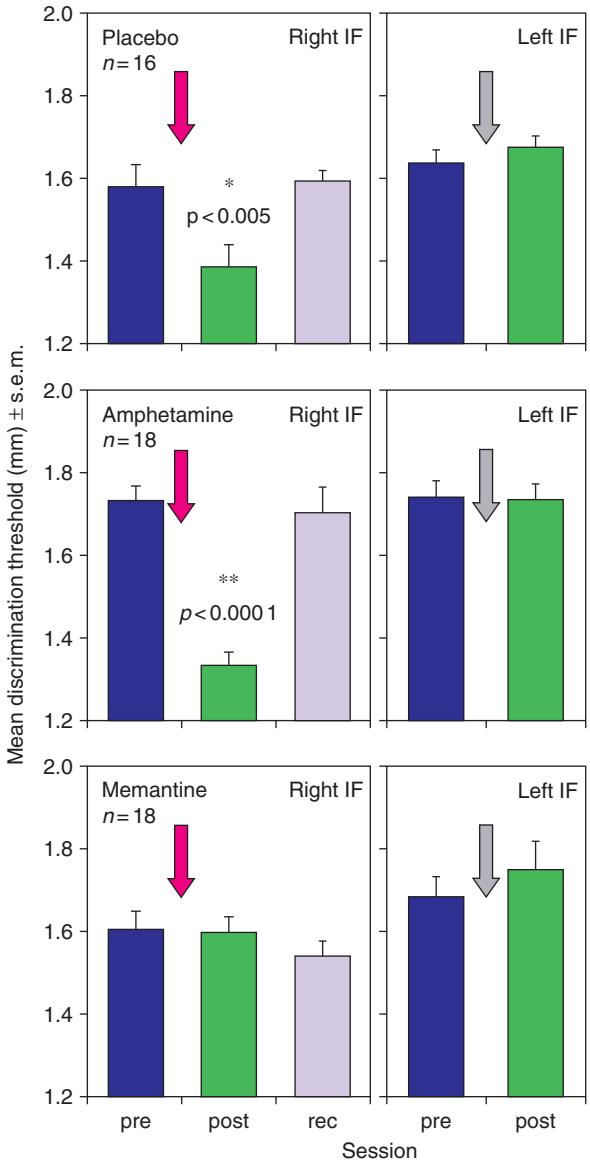

Figure 6 Influences of pharmaceuticals on perceptual learning, pre- and posttraining, using the left or right index finger (IF) for coactivation (rec, recovery). Memantine abolishes perceptual learning by blockade of *N*-methyl-D-aspartate receptor, as does the γ-aminobutyric acid receptor antagonist lorazepam (not shown). Amphetamine increases improvement in perceptual learning; under placebo administration, the learning is less pronounced than under amphetamine. Reprinted from Dinse HR, Ragert P, Pleger B, et al. (2003) Pharmacological modulation of perceptual learning and associated cortical reorganization. *Science* 301: 91–94, with permission from AAAS.

agonist, as well as by memantine, which blocks *N*-methyl-D-aspartate receptors. Amphetamine, on the other hand, increases the improvement (Figure 6).

Top-Down Influences in Perceptual Learning and Sensory Plasticity

Attention is a top-down influence affecting even primary sensory levels, and certainly plays an important role in PL. Several studies have demonstrated that a stimulus feature is better detected if it is attended to – mere presentation of a feature usually does not improve its perception over time through PL. But when a nonattended feature is always presented simultaneously with the attended stimulus, the simultaneous feature improves, too. This indicates that a general 'arousal' lasting less than a second may improve perception not only for attended features, but also for those presented simultaneously. In these rare cases, PL evolves without individuals attending to a specific stimulus feature.

In hyperacuity learning, attention also plays an important role. In one study, two vernier stimuli were presented simultaneously, resembling a cross, but only one of the verniers was attended. Then offset discrimination only for the attended vernier improved over the course of training. As mentioned previously, both PL and attention are able to improve performance, but on rather different time scales: attention immediately but without any permanent change, PL (under the guidance of attention) over the course of an extended training, but leaving long-lasting effects.

Improvement through learning in hyperacuity tasks is possible even in the absence of error feedback, but often is significantly faster in the presence of feedback. If only half of the incorrect responses are followed by an error signal (incomplete feedback), the observer's improvement is about as fast as with complete feedback, whereby each incorrect response leads to an error signal. This finding poses severe problems for all theories of (perceptual) learning based on a teacher signal, which allows the observer to classify each stimulus. With partial feedback, half of the incorrect responses would be classified as belonging to the incorrect class – and this should virtually abolish learning, while it does not. But improvement through training is indeed abolished if observers receive even manipulated, more precisely random, feedback, without any correlation between the correctness of their response and the error feedback (while observers assume that they receive correct error feedback). Improvement through learning is about as fast with block feedback when the percentage of correct responses after each block of 80 presentations is indicated as it is for complete trial-by-trial feedback. Most surprisingly, even manipulating the block feedback in a way similar to the condition of random feedback, by presenting a number uncorrelated with the actual performance of the observer, prevents improvement. Hence, feedback can strongly influence the speed and extent of visual learning, indicating that several top-down influences, not just attention, must play a major role in this type

of learning, even if it occurs partly on early levels of cortical information processing.

Conditions for Consolidation of Learning and Visual Rehabilitation of Patients

New experimental data have broadened the evidence that consolidation of improvement achieved through PL requires sleep or at least rest. Both sleep and restful waking facilitate auditory tone learning, and auditory training with a small number of words improves performance with these words, generalizing to novel words using the same phonemes but with different acoustic patterns. Sleep consolidates improvement in a spoken-language task and the automaticity in an auditory discrimination task. Training identification of vowels both improves performance and enhances amplitude of evoked potentials, decreasing their latency.

Amblyopic patients are clearly benefiting from PL. Training improves contrast sensitivity of amblyopic eyes by about a factor of 2, and increases their acuity significantly. Finally, PL is certainly required when deaf or blind patients learn to use the information supplied by an external microphone or camera via direct electrical stimulation of their cochlea, retina, or cortex. Patients wearing a cochlear implant learn to partially adjust to changes in frequency-to-electrode assignment. Unilateral lesions in macaques, on the other hand, severely disrupt learning and performance in a visual match-to-sample task.

Conclusions and Outlook

Perceptual learning can greatly improve the sensitivity of perception, with thresholds decreasing by a factor of 2 and more within a rather short period of training. This seems to rely on more than one process, with the different processes probably located on different (cortical) levels of signal processing. Improvement through modifications on an early cortical level tends to be more specific for the exact stimulus dimensions than are mechanisms relying on more 'cognitive' processes, and may require longer training. PL usually requires some form of attention (hence, no learning without attending to a stimulus) and is often highly specific for rather low-level features of stimuli, such as their orientation. To achieve optimal performance especially for this type of feature, irrelevant signals and noise have to be eliminated as early as possible during processing. To prevent interference of learning one task with performance in other tasks, the changes achieved on early levels have to be activated, in a task-dependent way, by top-down signals. Easier tasks may not have to involve modifications on the early processing levels, leading to generalization of improvement. Perceptual learning is an important mechanism, enabling humans to improve perception and to cope rapidly with ever-changing demands of sensorimotor requirements when interacting with the external world.

See also: Auditory System: Central Pathway Plasticity; BDNF in Synaptic Plasticity and Memory; Multisensory Convergence and Integration; Neural Coding of Spatial Representations; Parietal Cortex and Spatial Attention; Perceptual Learning: Neural Mechanisms; Sensorimotor Integration: Attention and the Premotor Theory; Spatial Cognition; Spatial Cognitive Maps; Supplementary Eye Fields; Transcription Factors in Synaptic Plasticity and Learning and Memory; Vision for Action and Perception.

Further Reading

Calford MB, Chino YM, Das A, et al. (2005) Neuroscience: Rewiring the adult brain. *Nature* 438: E3.

Dinse HR, Ragert P, Pleger B, et al. (2003) Pharmacological modulation of perceptual learning and associated cortical reorganization. *Science* 301: 91–94.

Fahle M (2002) Learning to perceive below the foveal photo receptor spacing. In Fahle M and Poggio T (eds.) *Perceptual Learning,* pp. 197–218. Cambridge, MA: MIT Press.

Fahle M (2004) Perceptual learning: A case for early selection. *Journal of vision* 4: 879–890.

Fahle M (2005) Perceptual learning: Specificity versus generalization. *Current Opinion in Neurobiology* 15: 154–160.

Fahle M and Daum I (2002) Perceptual learning in amnesia. *Neuropsychologia* 40: 1167–1172.

Fahle M and Edelman S (1993) Long-term learning in hyperacuity. *Vision Research* 33: 397–412.

Fahle M, Edelman S, and Poggio T (1995) Fast perceptual learning in hyperacuity. *Vision Research* 35: 3003–3013.

Fahle M and Poggio T (eds.) (2002) *Perceptual Learning.* Cambridge, MA: MIT Press.

Fiorentini A and Berardi N (1980) Perceptual learning specific for orientation and spatial frequency. *Nature* 287: 43–44.

Ghose GM (2004) Learning in mammalian sensory cortex. *Current Opinion in Neurobiology* 14: 513–518.

Gibson EJ (1969) *Principles of Perceptual Learning and Development.* New York: Appleton-Century-Crofts.

Gilbert CD, Sigman M, and Crist RE (2001) The neural basis of perceptual learning. *Neuron* 31: 681–697.

Helmholtz H von (1867) Handbuch der physiologischen Optik. In: Karsten G (ed.) *Allgemeine Enzyklopädie der Physik vol. IX.* Leipzig: Voss .

Herzog MH and Fahle M (1998) Modeling perceptual learning: Difficulties and how they can be overcome. *Biological Cybernetics* 78: 107–117.

Kaas JH, Krubitzer LA, Chino YM, et al. (1990) Reorganization of retinotopic cortical maps in adult mammals after lesions of the retina. *Science* 248: 229–231.

Karni A and Sagi D (1993) The time-course of learning a visual skill. *Nature* 365: 250–252.

Notman LA, Sowden PT, and Ozgen E (2005) The nature of learned categorical perception effects: A psychophysical approach. *Cognition* 95(2): B1–B14.

Poggio T, Fahle M, and Edelman S (1992) Fast perceptual learning in visual hyperacuity. *Science* 256: 1018–1021.

Rainer G, Lee H, and Logothetis NK (2004) The effect of learning on the function of monkey extrastriate visual cortex. *PLoS Biology* 2: 275–283.

Recanzone GH, Merzenich MM, and Dinse HR (1992) Expansion of the cortical representation of a specific skin field in primary somatosensory cortex by intracortical micro stimulation. *Cerebral Cortex* 2: 181–196.

Sagi D and Tanne D (1994) Perceptual learning: Learning to see. *Current Opinion in Neurobiology* 4: 195–199.

Perceptual Learning: Neural Mechanisms

W Li and C D Gilbert, The Rockefeller University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Learning and memory is manifested in different forms. The conscious form of memory, such as the recollection of previously seen objects and experienced events, is referred to as declarative memory or explicit memory, which can be acquired with a single exposure. The unconscious form of memory, which includes acquired habits and improved motor and perceptual skills, is called nondeclarative memory or implicit memory, which usually results from repeated experience. Perceptual learning is the unconscious acquisition of improved ability in simple discrimination and detection tasks with practice. It is an important form of implicit memory whereby the brain adapts for better and more efficient processing of recurring stimuli.

Different forms of memory are mediated by different brain structures. Studies on patients with specific damage to the brain memory system, the medial temporal lobe, have shown that despite an inability to form any long-term explicit memory, the patients are able to acquire and retain new perceptual skills, such as making a significant improvement in discriminating small misalignments between two collinear lines, a so-called vernier discrimination task in psychophysical studies (Figure 1(c)). The prevailing view is that all declarative memory shares the medial temporal lobe memory system, and that implicit memory can involve any area of the neocortex or cerebellum. In particular, studies have indicated the sensory cortices, including the primary sensory cortex, as the loci of perceptual learning.

This article focuses on typical perceptual learning that involves simple perceptual tasks and early sensory cortices. For complex visual learning, particularly for that pertaining to the recognition of complex objects and the association between objects, the learning is related to the formation of a more declarative form of memory, which is encoded in higher visual areas such as the inferior temporal cortex (IT) of monkeys and the lateral occipital complex of humans. This form of higher level visual learning is not discussed.

Characteristics of Perceptual Learning

Perceptual learning evolves with practice and time. For an untrained task such as vernier discrimination, a naïve subject's discrimination ability can be significantly improved with a few tens of training trials, indicating an early rapid phase in perceptual learning. Following the initial quick changes, slow and steady improvements are seen across daily training sessions until the subject's performance reaches a plateau. The effects of perceptual learning are virtually permanent. Subjects are able to retain the improved discriminability even if they stop performing the learned task for a couple of years; the longest duration has ever been tested. This enduring improvement suggests long-term plastic changes in the brain.

Although perceptual learning occurs in various perceptual tasks with repetitive exposure to the same stimulus and task conditions, the improved perceptual skill is usually highly specific to the conditions used for training, a phenomenon known as the specificity of perceptual learning. There is usually neither transfer nor interference between learning different tasks. The commonly reported specificity in visual perceptual learning can be divided into four types. First, perceptual learning is restricted to the visual field location where the task stimulus is presented. As the stimulus is moved away from the trained location, the observer's performance gradually regresses to the pretraining level (Figure 1(a)). Second, the learning is specific to the stimulus used for training. A change in simple stimulus attributes such as orientation (Figure 1(b)), or even a change in stimulus context (Figure 1(c)), diminishes or abolishes the learning effect. Third, learning in some visual tasks is specific to the eye being trained. If only one eye is exposed to the task stimuli and the other eye is occluded, the improvement does not transfer to the untrained eye. Finally, the learning is specific to the particular perceptual task being practiced and does not generalize to other tasks. Task dependency of perceptual learning indicates that top-down influences such as selective attention play a critical role in perceptual learning. For the same visual stimulus, a subject's ability to discriminate a subtle change in a stimulus attribute can be improved only if that attribute is attended and used in the perceptual task. The other unattended attributes are unaffected in perception even though the subjects are repeatedly exposed to the same stimulus. For example (Figure 1(d)), orientation and brightness are two attributes of a straight line. When both attributes are slightly altered from trial to trial, the discrimination ability of human subjects is improved only for the attended attribute.

Despite the converging evidence that perceptual learning requires attention to the task-relevant stimulus, studies have shown that an unattended visual stimulus can also be learned. However, in this case of task-irrelevant learning, the unattended stimulus

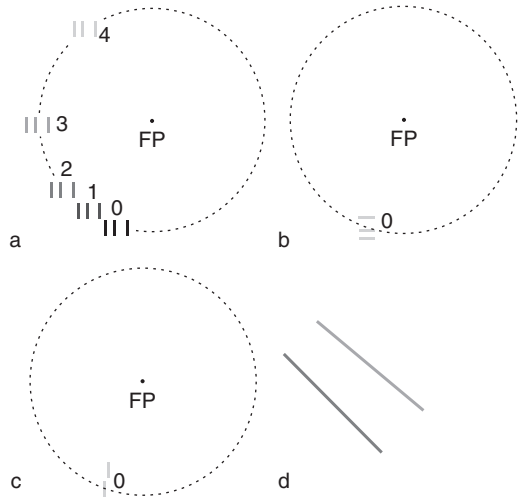

Figure 1 Illustration of the specificity of perceptual learning. (a) Specificity to the trained visual field location. In a three-line bisection discrimination task, the two outer stimulus lines are fixed as a reference, whereas the central line is slightly displaced from the middle to either side. The subject's task is to determine whether the central line appears closer to the left or to the right reference line. With a few weeks' training, the minimum displacement of the central line for reliable detection of the positional offset (i.e., the threshold for bisection discrimination) greatly decreases. The improved discrimination ability, however, is restricted to the location where the task is practiced (labeled as position '0'). When the same stimulus is moved away from the trained location (from 0 to 4) with the eccentricity kept constant, the training-induced improvement gradually disappears (denoted by a decrease in the darkness of lines). (b) Specificity to stimulus orientation. At the trained location, a 90-degree rotation of the stimulus greatly diminishes the learning effect. (c) Specificity to stimulus context. Even if the task is to discriminate the lateral displacement of the same line at the same location, the learning effect does not transfer to a condition of new stimulus context in which a reference line is fixed at one end of the target line. This discrimination task is also the vernier task mentioned in the text. (d) Specificity to the task-relevant attribute. When both orientation and brightness of a target line are changed from trial to trial relative to a reference line, the discrimination ability of human subjects is improved only for the attended attribute. (a) Adapted from Crist RE, Kapadia MK, Westheimer G, and Gilbert CD (1997) Perceptual learning of spatial localization: Specificity for orientation, position, and context. *Journal of Neurophysiology* 78: 2889–2894.

must coincide with an attended task stimulus for the learning effect to take place. One possible explanation for this subliminal learning is that attentively performing a task activates diffusive reinforcement signals, which interact not only with the task-relevant signals but also with those task-irrelevant signals that are temporally associated with the task stimulus. Therefore, even for the task-irrelevant learning it is still necessary to engage the attentional system of the brain.

Another form of top-down influence on perceptual learning involves feedback that indicates the observer's performance during training. External audio or visual feedback is not a prerequisite to perceptual learning because the learning still occurs without it. However, correct external feedback usually results in faster and more extensive improvement than without any feedback, whereas feedback that is uncorrelated with the subject's responses completely prevents the learning. To reconcile these phenomena, it has been proposed that an evaluation process, which is based either on feedback provided externally or on feedback derived internally from confident trials, is of key importance in perceptual learning.

The characteristics of perceptual learning, when considered along with the functional organization of sensory systems, provide implications for the cortical loci and mechanisms involved in perceptual learning. In the visual system, more than 30 cortical areas are associated with visual processing, and these areas are functionally specialized and hierarchically organized, forming partially segregated modules and pathways for processing different visual information. The primary visual cortex (area V1) is the first stage of visual cortical processing, where neurons are selective for simple stimulus attributes such as orientation. In addition, the receptive fields (RFs) of V1 neurons are, on average, the smallest among all visual areas, and the visual inputs from the two eyes are still partially separate in V1. In higher order visual areas, response properties of neurons get more complex, the RFs get much bigger, and the inputs from the two eyes become indistinguishable. That is, higher visual areas gain selectivity for more complex stimuli such as faces but loose spatial resolution and ocular identity. Therefore, given the specificity of learning to simple stimulus attributes, spatial location, and the trained eye, the involvement of early visual areas in visual perceptual learning can be deduced from the functional organization of the visual system. On the other hand, the dependence of perceptual learning on attention to specific stimulus attributes and tasks, together with the remarkable effects of feedback signals on learning, suggests an involvement of top-down control and an interaction between cortical areas.

Neuronal Mechanisms

In addition to the implication from psychophysical studies, a line of physiological evidence suggesting the involvement of early sensory cortices in perceptual learning comes from the accumulated findings that these cortical areas are highly plastic even in adults. The primary sensory cortices, including the primary somatosensory (area S1), auditory (area A1), and visual (area V1) areas, are well organized into topographic maps to represent the sensory surfaces, forming so-called somatotopic, tonotopic, and retinotopic

maps, respectively. After injury to the peripheral sensory surface in adult animals, the topographic maps will undergo pronounced reorganization processes. A similar plastic change in all these early sensory cortices is that the regions and neurons that used to represent the lesion surface start to represent adjacent intact surface. For instance, a laser-induced retinal scotoma removes visual input to the corresponding retinotopic region in V1 and silences cortical neurons within that region, producing a cortical scotoma (Figure 2). However, over a period of 2 months following the retinal injury, the cortical scotoma substantially shrinks: neurons within the cortical scotoma regain responsiveness by shifting their RFs outside the retinal scotoma. This reorganization has been shown to be cortically mediated rather than being a rearrangement of thalamocortical afferents. The plasticity of early sensory cortices observed during recovery from injury has been suggested to serve as the neural basis of perceptual learning under normal physiological conditions.

In agreement with the speculations from psychophysical and lesion studies, plastic changes associated with perceptual learning have been directly demonstrated in early sensory cortices. These changes can be grouped into several categories, which to some extent depend on the nature of perceptual tasks and could reflect different cortical mechanisms of perceptual learning.

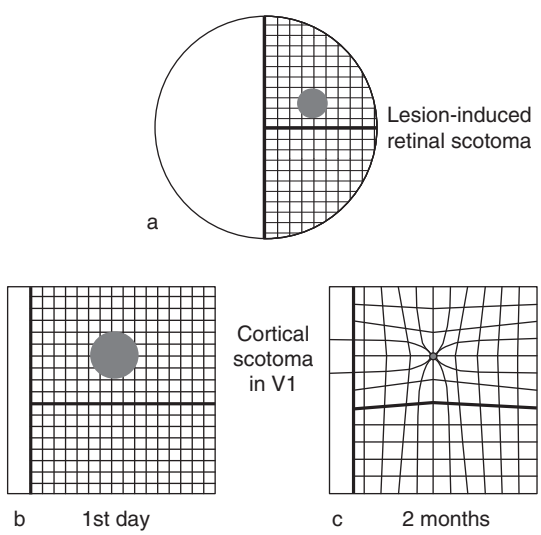

Figure 2 Plasticity in adult V1. A retinal lesion (a) immediately silences the corresponding retinotopic region in V1 (b). The retinal lesion induces plastic changes in V1 in such a way that neurons within the cortical scotoma regain responsiveness to visual input from the area surrounding the retinal scotoma (c). Adapted from Gilbert CD (1992) Horizontal integration and cortical dynamics. *Neuron* 9: 1–13.

Expansion of Cortical Representation

The topographic maps in S1 and A1 have been shown to be able to reorganize with perceptual learning. This process is known as cortical recruitment, whereby the cortical area representing the trained sensory surface expands by recruiting the adjacent cortical areas previously representing the untrained surface. For example, after training monkeys to perform a tactile frequency discrimination task using a small surface of one digit, remarkable reorganization of the somatotopic map in S1 has been observed, characterized by a significant increase in the size and complexity of the territory representing the trained skin area. Similarly, training on an acoustic frequency discrimination task results in an extensive increase in cortical area representing the trained frequencies within A1.

Intuitively, with the participation of a larger area of cortex and a proportionally greater number of neurons, better performance should be achieved in the discrimination task at the trained location due to probability summation: the probability of detecting a stimulus increases with the number of detectors (neurons), assuming that the detectors are at least partially independent and that a subset is enough to signal the presence of the stimulus. Nevertheless, it is a matter of some debate whether the learning-associated cortical recruitment is actually responsible for the improved discrimination ability. Although there has been some supporting evidence that cortical recruitment in S1 and A1 can be correlated with improved performance in discrimination tasks, results have also been reported that there is no cortical recruitment in A1 associated with training on an auditory frequency discrimination task. Moreover, an expansion of cortical representation of the trained sensory surface occurs at the expense of the representation of adjacent surface. If an increase in representation is responsible for the enhanced discrimination ability on the trained sensory surface, a decrease in representation would presumably result in impaired discrimination ability on the adjacent untrained surface. Compelling evidence for such a phenomenon is still lacking. In the visual system, the absence of transfer between very similar tasks also argues against cortical recruitment. For example, in the bisection (Figure 1(a)) and vernier (Figure 1(c)) discrimination tasks mentioned previously, the spatial offset of the same target line is discriminated in the same visual field location but within different stimulus contexts. If the improvement in one task is mediated by cortical recruitment, one would expect to see an improvement in the other task. In fact, studies have shown that no change occurs in the retinotopic map in V1 after extensive training on bisection discrimination. However,

changes do occur in V1 neuronal response properties that can account more for the context and task specificity of the learning.

Changes in Stimulus Selectivity of Neurons

Instead of getting more neurons involved, another potential mechanism of perceptual learning is to increase neuronal selectivity for the stimulus attributes that are relevant to the trained task, as supported by electrophysiological studies in monkey V1. One of the remarkable properties of V1 neurons is their sensitivity to stimulus context, a phenomenon called contextual modulation whereby neuronal responses to a stimulus in the RF are modified by other stimuli located outside the RF. Contextual influences seen in V1 indicate that V1 neurons are selective for more complex features in visual scenes in addition to basic stimulus attributes. Studies have revealed that the neuronal mechanism of contextual modulation can be retuned by perceptual learning so as to fit the requirements of the discrimination task.

In the bisection discrimination task (Figure 1(a)), the positional offset of the target line is determined relative to the flanking lines. Earlier studies on contextual modulation in V1 showed that neuronal responses to a line positioned in the RF and orientated at the cell's preferred orientation are usually slightly inhibited by a second parallel line placed on either side of the RF. However, as shown by a study in which monkeys were trained on a three-line bisection

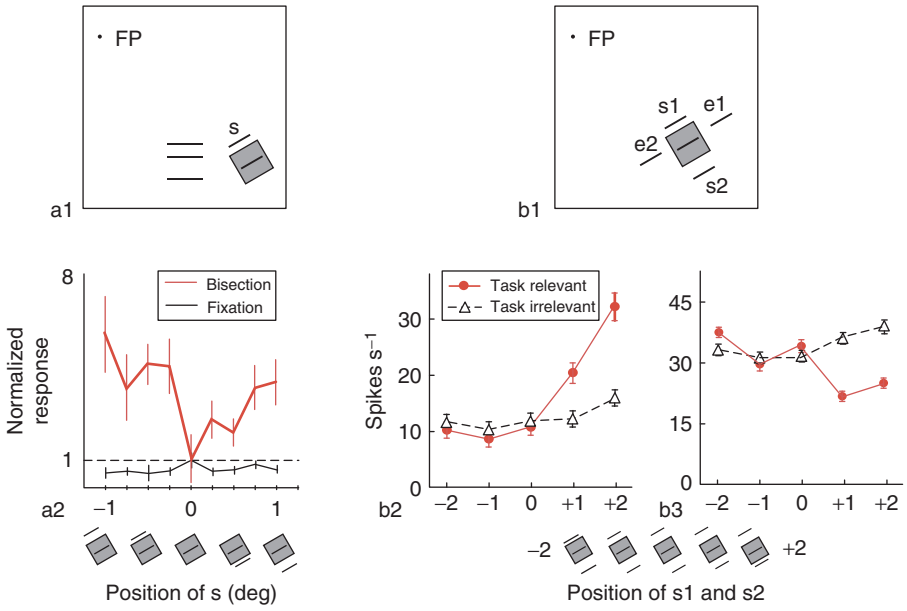

Figure 3 Perceptual learning modifies response properties of V1 neurons in a task-specific manner. (a1) After training monkeys on a bisection discrimination task using three horizontal lines, responses of single V1 neurons to another test stimulus were recorded when the animal either did the trained bisection task or simply maintained its fixation at the fixation point (FP). The test stimuli consisted of two lines, an optimally oriented line fixed in the center of the RF (denoted by the small gray square) and a second parallel line (indicated by 's') placed at different locations on either side of the RF (bottom of (a2); at position 0 degrees the two lines were superimposed). (a2) The normalized responses of a V1 cell to the test stimuli as a function of the position of line 's.' When the animal was performing the simple fixation task, placing 's' on either side of the RF slightly suppressed neuronal responses relative to the responses at position 0 degrees, where a single line was located in the RF. In contrast, when the animal was performing the bisection task, the weak contextual inhibition was changed into strong facilitation. The difference in contextual modulation between the fixation and bisection task conditions was seen only in the trained area. (b1) To further isolate the effect of perceptual task *per se*, monkeys were trained to do two different discrimination tasks with identical stimulus patterns at the same visual field location. The stimuli consisted of five simultaneously presented lines: an optimally oriented line fixed in the RF center and flanked by four additional lines surrounding the RF. In different trials, the arrangement of the two side flankers (s1 and s2) was randomly assigned from a set of five different configurations (illustrated at the bottom of (b2) and (b3); labeled from −2 to +2). Each configuration differs from the others in the separation between the three side-by-side lines (in condition '0' the three lines were equidistant; in the other conditions either s1 or s2 was closer to the central line). In the same trials, the two end flankers (e1 and e2) were also independently assigned a random configuration from a set of predefined arrangements, such that the end flankers were collinear with each other but misaligned with the central line to either side. The animal was cued to perform either a bisection task based on the three side-by-side lines or a vernier task based on the three end-to-end lines. (b2, b3) Two examples of V1 cells in response to the five-line stimuli under the two task conditions. Neuronal responses were examined as a function of the position of the two side flankers s1 and s2 when the animal either performed the bisection task, in which s1 and s2 were task relevant, or performed the vernier task, in which the same s1 and s2 were task irrelevant. (a1, a2) Adapted from Crist RE, Li W, and Gilbert CD (2001) Learning to see: Experience and attention in primary visual cortex. *Nature Neuroscience* 4: 519–525. (b1–b3) Adapted from Li W, Piech V, and Gilbert CD (2004) Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience* 7: 651–657.

discrimination task, this contextual effect in V1 can be greatly modified (Figures 3(a1) and 3(a2)). For a large proportion of cells, the modulatory effect caused by the second line is profoundly increased during the performance of the bisection task. In a subset of these cells, weak contextual inhibition is even reversed to strong facilitation (Figure 3(a2)). The changes in magnitude and, in some cases, polarity of contextual modulation are present only in the trained area while the animal performs the learned task, suggesting a task-specific modification of neuronal response properties. A more rigorous examination of task-specific contextual modulation was provided by a subsequent study. Monkeys were presented with an identical set of stimulus patterns on which they were trained to perform either a bisection or a vernier discrimination task (Figure 3(b1)). This study showed that V1 neuronal responses are strongly modulated by the stimulus attribute relevant to the immediate task, but that they are little affected by the other task-irrelevant attribute (Figures 3(b2) and 3(b3)). Information theoretic analyses revealed that, compared with the task-irrelevant condition, neuronal responses carry significantly more information about the same stimulus attribute when the attribute is used by the animal to solve the discrimination task. Taken together, these findings suggest that V1 neurons in the trained animals take on novel response properties related to the discrimination task, and that the exercise of these acquired response properties requires the same stimulus and behavioral context as that during training. This neuronal mechanism coincides with the specificity of perceptual learning.

Besides the selectivity for stimulus context, basic tuning properties of V1 neurons can also be modified by perceptual learning, leading to an improvement in neuronal selectivity for simple stimulus attributes. For example, in both area V1 and area V4, a steepening in orientation tuning curves around the trained orientation has been reported in a subset of neurons after training monkeys on orientation discrimination. The orientation tuning curves of V4 neurons are also generally narrower in the trained region.

Learning-associated cortical recruitment, as well as changes of neuronal response properties, requires attention to the stimulus attribute relevant to the task. This is consistent with the psychophysical finding that the encoding of the learned information is subject to top-down influences. Moreover, the reorganized cortical maps in A1 and S1 are seen when the trained animals are under anesthesia, suggesting that long-term plastic changes indeed take place in early sensory areas. However, neuronal response properties in V1 that emerge from learning the bisection and vernier discrimination tasks are present only when the animal is executing the learned task. This suggests that top-down influences play a triggering role in recall of these acquired properties for processing the perceptual task at hand. One possible neuronal mechanism is that the information related to a given stimulus attribute is represented at the level of subsets of inputs to a cell, which can be plastically tuned and dynamically gated by top-down signals, probably via interactions between feedback connections from higher cortical areas and intrinsic connections within V1. This flexible mechanism enables multiple learned attributes to be represented by the same neurons without cross-talk, greatly expanding information processing capability in V1. Therefore, the top-down influences are critical in both encoding and retrieving the learned information.

Enhanced Neuronal Responsiveness

An increase in neuronal activity as a result of higher responses from individual neurons or from recruiting additional neurons could account for perceptual learning due to an improvement in the signal-to-noise ratio or to probability summation. On the other hand, a sharpening of tuning can also account for improvement in discrimination, yet this would mean that fewer cells would respond to the stimulus, leading to an overall decrease in neuronal responses.

Some discrimination tasks involve making a fine distinction between a target and a reference along a given stimulus dimension such as frequency, orientation, or shape. It is not surprising, then, that training on such tasks would enhance neuronal selectivity for the task-related stimulus attributes. The changes in tuning properties are, in some cases, accompanied by changes in neuronal responsiveness. Some studies have reported increases in neuronal activity associated with discrimination learning. For example, over the course of learning a tone discrimination task in adult owl monkeys, an increase in sensitivity of A1 neurons to all frequencies is observed during the initial phase of improvement in performance. Subsequently, neuronal responses to the target stimuli are differentially intensified relative to the reference and nontarget stimuli. In monkey visual area V4, after training the animal on orientation discrimination, a narrowing of orientation tuning curves in the trained area is accompanied by an overall increase in neuronal responsiveness. In contrast, an imaging study has shown decreased orientation-specific activity in early visual areas associated with orientation discrimination training. In the case of training on bisection and vernier discrimination, both an increase (Figure 3(b2)) and a decrease (Figure 3(b3)) in neuronal responses are observed in different subsets of V1 neurons, but both changes are associated with

increased selectivity for the task-relevant attribute. The diverse observations suggest that the changes in neuronal responsiveness with discrimination training may be more closely related to the change in tuning properties, which better accounts for the improvement in discrimination ability.

Unlike the discrimination tasks mentioned previously, some perceptual tasks involve detection or identification of the presence or absence of a target, either in isolation near the detection threshold or within a noisy background. In this case, the intensity or signal-to-noise ratio of the target stimulus is more directly related to the task than any other stimulus attribute. Therefore, the responsiveness of neurons to the target is crucial for solving the detection task. Depending on the attributes of the target to be identified, activity in different visual areas is significantly increased with training. It has been reported that an improvement in monkeys' performance on detection of coherent motion of dynamic random dots is correlated with an increase in neuronal activity in areas MT and MST that are involved in motion processing. Visual area V4 is at an intermediate stage along the hierarchical pathway for object recognition. When monkeys are trained to identify natural images degraded by noise, V4 neuronal responses to those familiar and degraded pictures are specifically enhanced. In early visual areas including V1, evidence from electrophysiological and functional magnetic resonance imaging (fMRI) studies shows that training on detection of a target embedded in a background of noise or distracters significantly boosts activity in these early visual areas, and the increased activity is correlated with the improvement in target detection. Training on target detection near the contrast threshold and in the absence of any noise or distracter shows a similar effect. After training human subjects to detect a near-threshold grating patch, the fMRI signal in V1 for the trained orientation is significantly increased relative to the control orientation, and this change is correlated with the improvement in target detection. Based on these accumulated findings, an important neuronal mechanism of learning in target detection near threshold or in noisy environments involves increasing the responsiveness of sensory neurons to the target and thus increasing the signal-to-noise ratio.

Shift of Cortical Representation

A special kind of detection task is visual search, in which a target is camouflaged in a background of similar distracters. Psychophysical evidence indicates that extensive training on searching for a hidden target significantly increases the efficiency of target detection, resulting in a pop-out of the target. Functional imaging studies have shown that, in addition to increased activity in early retinotopic areas, learning to search for a particular shape among distracters results in decreased activity both in higher visual areas involved in shape processing and in prefrontal and parietal areas responsible for attentional control. This suggests that another mechanism of perceptual learning involves a shift of cortical representation of the learned stimulus from higher to lower visual areas for more efficient and less effortful processing of the task stimulus. These findings also indicate that the attentional mechanism is pliable, being modified with training.

Changes in Temporal Response Properties of Neurons

The neural codes for sensory stimuli are included not only in the firing rates of neurons but also in their temporal firing patterns, such as the synchrony of discharges across a population of neurons. It has been shown that temporal response properties of neurons can be modified by perceptual learning. For example, neuronal responses in monkey S1 become more coherent with training on tactile frequency discrimination. This change is highly predictive of the improved performance and thus correlates better with the learning than does cortical recruitment. Similarly, after training rats to use the repetition rate of acoustic noise pulses to find food in a maze, with higher repetition rates indicating shorter distances between the animal and the food, A1 neuronal responses to high-frequency pulses become stronger and more phase-locked to the stimuli. In early visual areas of cat, gamma oscillations – an indication of response synchrony among neurons – are increased for a behaviorally relevant stimulus on which the animals were trained. These findings suggest that the enhanced temporal response properties of a population of neurons can serve as another neuronal mechanism of perceptual learning.

Conclusions

Experience-dependent plasticity in sensory cortex serves as the neural basis of perceptual learning throughout the life. Depending on the attributes of the stimuli to be learned, different neuronal mechanisms are evoked to encode the learned information, which can be stored by modifying the functional properties and architecture of early sensory cortices. These changes usually take place in a task-specific fashion under top-down control. Similar top-down influences are required, at least in some visual perceptual learning, to retrieve the learned information relevant to a specific task. This cortical mechanism provides

an effective way to avoid any interference between learning and performing different perceptual tasks.

See also: Adult Cortical Plasticity; Attentional Functions in Learning and Memory; Auditory System: Central Pathway Plasticity; Contextual Interactions in Visual Perception; Contextual Interactions in Visual Processing; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Motor Skill Learning; Neural Synchrony and Feature Binding; Perception of Surfaces and Forms; Perceptual Learning and Sensory Plasticity; Visual Associative Memory; Visual Attention.

Further Reading

Bao SW, Chang EF, Woods J, and Merzenich MM (2004) Temporal plasticity in the primary auditory cortex induced by operant perceptual learning. *Nature Neuroscience* 7: 974–981.

Blake DT, Strata F, Churchland AK, and Merzenich MM (2002) Neural correlates of instrumental learning in primary auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America* 99: 10114–10119.

Brown M, Irvine DRF, and Park VN (2004) Perceptual learning on an auditory frequency discrimination task by cats: Association with changes in primary auditory cortex. *Cerebral Cortex* 14: 952–965.

Crist RE, Kapadia MK, Westheimer G, and Gilbert CD (1997) Perceptual learning of spatial localization: Specificity for orientation, position, and context. *Journal of Neurophysiology* 78: 2889–2894.

Crist RE, Li W, and Gilbert CD (2001) Learning to see: Experience and attention in primary visual cortex. *Nature Neuroscience* 4: 519–525.

Fahle M (2004) Perceptual learning: A case for early selection. *Journal of Vision* 4: 879–890.

Furmanski CS, Schluppeck D, and Engel SA (2004) Learning strengthens the response of primary visual cortex to simple patterns. *Current Biology* 14: 573–578.

Gilbert CD (1992) Horizontal integration and cortical dynamics. *Neuron* 9: 1–13.

Lee TS, Yang CF, Romero RD, and Mumford D (2002) Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. *Nature Neuroscience* 5: 589–597.

Li W, Piech V, and Gilbert CD (2004) Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience* 7: 651–657.

Rainer G, Lee H, and Logothetis NK (2004) The effects of learning on the function of monkey extrastriate visual cortex. *PLoS Biology* 2: 275–283.

Recanzone GH (2000) Cerebral cortical plasticity: Perception and skill acquisition. In: Gazzaniga MS (editor-in-chief) *The New Cognitive Neurosciences,* 2nd edn., pp. 237–247. Cambridge: MIT Press.

Salazar RF, Kayser C, and Konig P (2004) Effects of training on neuronal activity and interactions in primary and higher visual cortices in the alert cat. *Journal of Neuroscience* 24: 1627–1636.

Schiltz C, Bodart JM, Dubois S, et al. (1999) Neuronal mechanisms of perceptual learning: Changes in human brain activity with training in orientation discrimination. *Neuroimage* 9: 46–62.

Schoups A, Vogels R, Qian N, and Orban G (2001) Practising orientation identification improves orientation coding in V1 neurons. *Nature* 412: 549–553.

Seitz A and Watanabe T (2005) A unified model for perceptual learning. *Trends in Cognitive Sciences* 9: 329–334.

Shiu LP and Pashler H (1992) Improvement in line orientation discrimination is retinally local but dependent on cognitive set. *Perception & Psychophysics* 52: 582–588.

Sigman M, Pan H, Yang YH, Stern E, Silbersweig D, and Gilbert CD (2005) Top-down reorganization of activity in the visual pathway after learning a shape identification task. *Neuron* 46: 823–835.

Yang TM and Maunsell JHR (2004) The effect of perceptual learning on neuronal responses in monkey visual area V4. *Journal of Neuroscience* 24: 1617–1626.

Zohary E, Celebrini S, Britten KH, and Newsome WT (1994) Neuronal plasticity that underlies improvement in perceptual performance. *Science* 263: 1289–1292.

Perfusion MRI

T T Liu, E C Wong, and R B Buxton, University of California at San Diego, La Jolla, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

The first demonstration that the signal measured with magnetic resonance imaging (MRI) can be made sensitive to neural activity used an injected intravascular contrast agent that alters the MR signal as it passes through the brain. The investigators compared the dynamics of the agent in visual cortex during a rest condition and, with a second injection, during sustained visual stimulation. Although this contrast agent approach ignited the field of functional neuroimaging with MRI, and has now become a standard technique in clinical practice, it has had only limited applications in basic neuroscience studies of brain function. However, it demonstrated that local perfusion changes associated with neural activity can be detected with MRI and then used to map patterns of activation. Soon after this demonstration, the blood oxygenation level-dependent (BOLD) effect was discovered, and this quickly supplanted the earlier perfusion-based technique. With BOLD imaging, the signal changes associated with neural activity arise from a local increase in blood oxygenation, because the fractional blood flow increase is larger than the fractional change in the oxygen metabolic rate. No injections are needed, because the BOLD effect is an intrinsic change in the MR signal related to magnetic properties of hemoglobin, making experiments easier to perform, and also allowing measurement of dynamic changes.

Although imaging methods based upon BOLD contrast are currently the predominant method for functional magnetic resonance imaging (fMRI) of the brain, the interpretation of changes in the BOLD signal can be complicated by factors such as age, disease, and medication, because the magnitude of the BOLD response depends on the matching of changes in blood flow and oxygen metabolism. In addition, BOLD methods can only detect changes between two states, and so are insensitive to chronic conditions. A different approach to fMRI, based upon arterial spin labeling (ASL) methods, offers a useful complement to BOLD fMRI. Like BOLD, the ASL approach requires no injections, and exploits the ability of MRI techniques to selectively 'tag' water in different regions by inverting the proton magnetization, a noninvasive and easily repeatable method to label inflowing blood. It can provide quantitative measures of both resting-state levels of cerebral blood flow (CBF) and of the functional CBF response associated with evoked neural activity. It is also insensitive to slow signal drifts, which can contaminate the BOLD signal. When combined with measures of the BOLD response, measures of functional CBF responses can be used to estimate functional changes in the cerebral metabolic rate of oxygen.

Cerebral Blood Flow

Cerebral blood flow, or perfusion, is a measure of the rate of delivery of arterial blood to a capillary bed in tissue. The standard unit of measurement for CBF is milliliters of blood per 100 g^{-1} of tissue per minute, and a typical value in the human brain is 60 ml 100 $g^{-1}min^{-1}$. Assuming an average brain tissue density of 1 g ml^{-1}, the average CBF may also be written as 60 ml 100 ml^{-1} $min^{-1} = 0.01\ s^{-1}$. In a typical ASL experiment, about 1 s is allowed for the delivery of blood. This corresponds to 1 ml of blood delivered to 100 ml of tissue. As a result, the overall MR signal change due to the delivered blood is only about 1% of the total signal due to the tissue. This small percentage contributes to the low intrinsic signal-to-noise ratio (SNR) of ASL methods.

Overview of ASL

MRI is based on manipulating the magnetization of protons, primarily in water molecules, to generate a detectable MR signal, and the MR image is then a spatial map of this signal. Initially the magnetic moments of the protons partially align with the main magnetic field of the MRI system to create a local magnetization. When radiofrequency (RF) magnetic fields are applied, this magnetization can be tipped over to create a detectable MR signal, or fully inverted by a 180° RF pulse. After an RF pulse the magnetization relaxes back to alignment with the main magnetic field, typically requiring a few seconds, and the magnetization can then be tipped again by another RF pulse to generate a new MR signal. The effects of the RF pulses can be quite selective, exciting only a thin plane or a broad band. Most ASL methods measure CBF by taking the difference of two sets of images: tag images, in which the magnetization of arterial blood is inverted or saturated, and control images, in which the magnetization of arterial blood is fully relaxed. In this way the arterial blood delivered to an image voxel has

a positive magnetization in one image and a negative magnetization in the other, while the static protons contribute the same signal to both and so subtract out in the difference image. The basic steps are as follows:

1. Arterial blood flowing toward the region of interest (ROI) is tagged by magnetic inversion or saturation (see Figure 1). This tagged blood flows into the imaging slices or volume and also relaxes toward equilibrium with the longitudinal relaxation time constant of blood (see red curves in the bottom half of Figure 1).
2. After a delay TI (known as the inversion time) to allow for inflow of tagged blood, an image is acquired in the slice(s) or volume of interest. This image is referred to as the tag image.
3. As discussed previously, the delivered blood signal is only about 1% of the static tissue signal. In order to remove the contribution of the static tissue to the tag image, a control image of the same slice is acquired in which inflowing blood is not tagged. The magnetization of this control image is shown by the dashed blue lines in the bottom half of Figure 1.
4. Taking the difference of the control and tag images yields an image $\Delta M = M_{\text{control}} - M_{\text{tag}}$ that is proportional to CBF. This is shown as the difference of the blue and red lines in Figure 1.
5. Tag and control images are typically acquired in an interleaved fashion, and the running difference of the control and tag images is used to form a perfusion time series (see later, section titled 'Processing of ASL data').

ASL Methods

There are currently three classes of ASL methods, pulsed, continuous, and velocity selective, as summarized in Figure 1.

Pulsed ASL – Tagging Based on Location

In pulsed ASL (PASL), short (5–20 ms) RF pulses are used to saturate or invert a slab of magnetization that is proximal to the region of interest. These techniques have the advantages that they provide high inversion efficiency and use little RF power. Disadvantages are that they depend on the coverage and uniformity of the transmit RF field to determine the geometry of the applied tag.

Continuous ASL – Tagging Based on Location and Velocity

In continuous ASL (CASL), long (1–3 s) RF pulses are used in conjunction with a static gradient field to irradiate a narrow plane of spins with RF energy. The plane is chosen proximal to the region of interest so that inflowing arterial blood flows through in a direction that is roughly perpendicular to the plane. When the RF amplitude and gradient fields are properly adjusted, flowing blood undergoes adiabatic flow-dependent inversion when traversing the irradiated plane. Because the tag can be applied closer to the region of interest (on average), this can result in a higher overall tagging efficiency than is possible with pulsed techniques. However, this technique requires a large amount of average RF power and

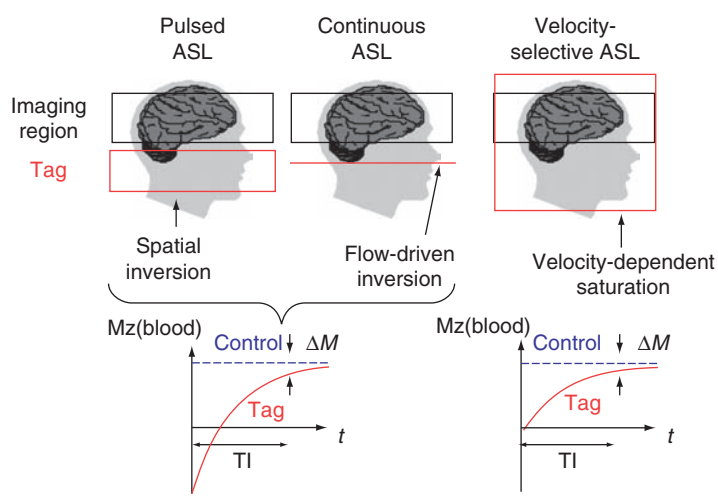

Figure 1 Pulsed, continuous, and velocity-selective arterial spin labeling (ASL) methods use various methods to either invert or saturate the magnetization of arterial blood. The difference, ΔM, between control and tag images is proportional to cerebral blood flow. Adapted from Liu TT and Brown GG (2007) Measurement of cerebral perfusion with arterial spin labeling. Part 1: Methods. *Journal of the International Neuropsychological Society* 13: 526–538, copyright © 2007 INS. Reprinted with permission.

can be limited by specific absorption rate (SAR) considerations at higher fields. A recently introduced form of CASL, dubbed pseudo-CASL, uses repeated RF pulses instead of a continuous RF wave.

Velocity-Selective ASL – Tagging Based on Velocity

In the velocity-selective ASL (VS-ASL) technique, an RF pulse train is applied that selectively saturates flowing spins with no spatial selectivity. Because it tags all blood that is flowing above a selected cutoff velocity, this technique has the advantage that the applied tag is close to all the target tissue. This results in a small and uniform transit delay for the delivery of the tagged blood to the tissues of interest (see below regarding transit delays).

Perfusion Quantification

The ASL difference signal ΔM is proportional to CBF but also has a complex dependence on a number of physiological parameters. Variations in these physiological parameters can cause systematic errors in the CBF measurements, and the goal of quantitative ASL methods is to minimize these errors. The primary sources of errors are (1) the transit delay Δt between the tagging region and the imaging slice, (2) the temporal width τ of the arterial bolus in PASL techniques, (3) the presence of tagged intravascular blood that has not yet perfused its target capillary bed, (4) the dependence of relaxation on the exchange of water between blood and tissue compartments, and (5) the clearance of water by outflow.

Transit delays arise in spatial tagging methods because of the need to place the tagging slab in PASL or the tagging plane in CASL some distance (1–3 cm) from the imaging slice in order to minimize the perturbation of spins in the imaging region. To allow all of the tagged blood to travel from the tagging region to the imaging region, the inversion time TI must satisfy the condition $\mathrm{TI} > \Delta t + \tau$, where Δt is the transit delay and τ is the temporal width of the tagged bolus. Although increasing the inversion time reduces the sensitivity to transit delays, it also reduces the magnitude of the difference signal ΔM, which is decaying exponentially with a relaxation time that is of the same order as the transit delay (the T_1 of blood is approximately 1600 ms at a magnetic field of 3 T). Typical inversion times are 1000 and 1400 ms for CASL and PASL, respectively. In contrast to CASL and PASL, transit delays are typically negligible in VS-ASL because the velocity-selective tagging process saturates arterial blood with velocities down to 1–2 cm s^{-1}, corresponding to vessels (roughly 50–100 μm arterioles) that are close to the capillary bed and within the imaging region. This insensitivity to transit delay makes VS-ASL a particularly attractive method for functional ASL studies of those individuals, such as stroke patients, who may exhibit long transit delays.

In CASL methods the temporal bolus width τ is determined by the temporal duration (typically 1–3 s) of the continuous inversion pulse. However, in PASL methods, the temporal width depends on the velocity of arterial blood in the tagging region, which typically increases with functional activation. To reduce the sensitivity to the temporal width, a spatial saturation pulse can be applied to the tagging region at a time TI_1 after the inversion pulse. If $\mathrm{TI}_1 < \tau$, then the temporal width of the bolus is TI_1, and the measured signal is proportional to the product of CBF and TI_1.

A related source of error is the presence of tagged blood in the arteries and arterioles that has entered the imaging slice but is destined to perfuse to more distal slices. If the image is acquired before this blood has time to leave the imaging slice, the intravascular signal will contribute to the ASL difference signal and lead to an overestimate of CBF. Typically, this is not a problem when using pulse sequence parameters that also control for transit delays. In addition, flow-weighting gradients can be used to attenuate the intravascular signal, at the cost of a reduction in the SNR of the desired difference signal.

Improper modeling of the exchange of water between blood and tissue components can lead to significant errors in CBF quantification, especially in tissues with low CBF, such as white matter. For these cases, a simple two-compartment model of exchange can be used to improve accuracy. Finally, errors due to the clearance of water by venous outflow are usually negligible given the inversion times used in most ASL experiments, but may need to be considered in high-flow conditions or when long inversion times are used.

Processing of ASL Data

In most ASL fMRI experiments, control and tag images are acquired in an interleaved fashion, as shown in Figure 2. A perfusion time series is then formed from the running subtraction of the control and tag images. An example of this process is shown in Figure 2, where the perfusion images are formed from the difference between each control image and the average of the two surrounding tag images or from the difference between the average of two surrounding control images and a tag image. The type of differencing shown in Figure 2 is referred to as

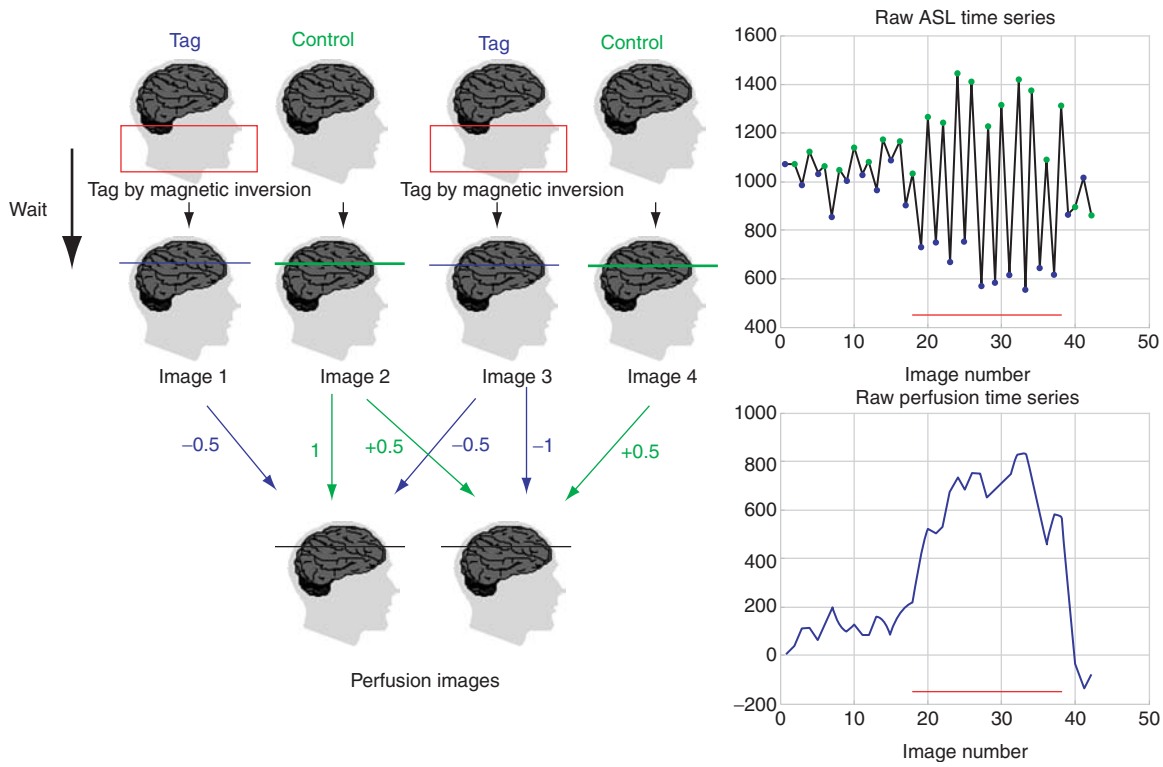

Figure 2 Formation of a perfusion time series from the 'surround subtraction' of control and tag images. An example arterial spin labeling (ASL) time series of control (green) and tag (blue) signals is shown in the upper right-hand plot. The perfusion time series created by surround subtraction is shown in the lower right-hand plot. The red bars indicates the time of stimulus. Adapted from Liu TT and Brown GG (2007) Measurement of cerebral perfusion with arterial spin labeling. Part 1: Methods. *Journal of the International Neuropsychological Society* 13: 526–538, copyright © 2007 INS. Reprinted with permission.

surround subtraction, and represents one specific approach to the filtered subtraction of control and tag images. Other common approaches are pairwise subtraction and sinc subtraction. These approaches all share the property that they can be modeled as the modulation of the control and tag images (e.g., multiply all control images by +1 and all tag images by –1) followed by convolution with a lowpass filter. For fMRI experiments with block designs (e.g., long periods of on/off), surround and sinc subtraction approaches tend to provide the best performance. In contrast, for randomized event-related experimental designs, pairwise subtraction can provide better performance (e.g., less filtering of the hemodynamic response function). However, all filtered subtraction approaches lead to inevitable broadening of the hemodynamic response function. Unfiltered approaches based upon general linear models of the ASL experiment can be used to eliminate this broadening and can also provide improvements in statistical power.

As with BOLD fMRI experiments, cardiac and respiratory fluctuations are a major source of noise in ASL experiments, especially at higher field strengths. Since the inherent SNR of ASL methods is typically much lower than that of BOLD fMRI, the need for methods to reduce physiological noise is especially pronounced. Retrospective image-based methods have been shown to significantly reduce physiological noise in ASL data and should be used whenever possible. An example is shown in Figure 3.

Simultaneous Estimates of CBF and BOLD

ASL data for fMRI are often acquired with single-shot acquisition methods such as echo-planar imaging (EPI) or spiral readouts. In addition to their sensitivity to flow introduced by the tagging process, these images can also exhibit BOLD contrast. For example, with gradient echo readouts, the images exhibit a BOLD weighting. Dual echo acquisitions are particularly useful for simultaneous perfusion and BOLD imaging. Images acquired at a short echo time (e.g., 3 ms) can be used to form a perfusion time series with relatively little BOLD weighting, while images acquired at a later echo time (e.g., 30 ms) can be used to form the BOLD time series.

A BOLD time series can be formed from the control and tag images through a running average approach that is analogous to the filtered subtraction approach used to form the perfusion time series. In general, the

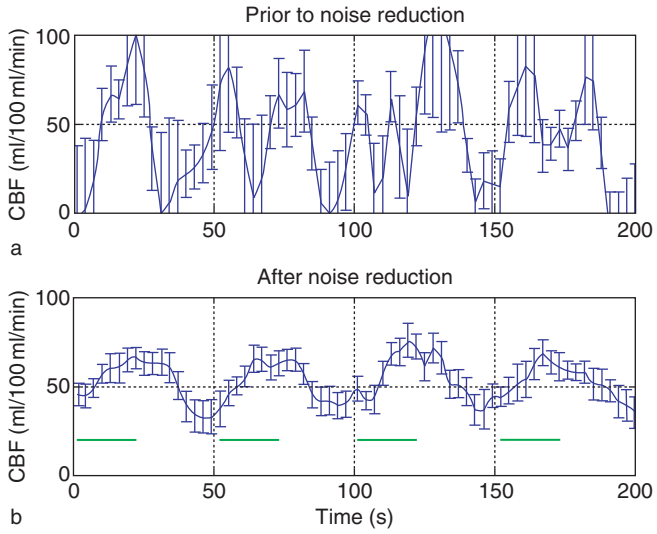

Figure 3 Example of the hippocampal cerebral blood flow (CBF) response to a memory encoding task before (top panel) and after (bottom panel) the application of physiological noise correction. Mean responses from a population of ten older individuals (mean age 74 years) are shown with standard error bars. The horizontal green bars denote the times during which novel images were shown.

BOLD time series is formed by the convolution of the control and tag images with a lowpass filter. For example, the method analogous to surround subtraction would be a surround average consisting of the average of each image with the average of its two nearest neighbors. The BOLD time series are inherently flow-weighted because of the tagging process. In PASL experiments the flow-weighting can be reduced by the application of a presaturation pulse applied in the imaging plane either immediately prior to or immediately after the application of the tagging pulse.

Temporal Resolution

The temporal resolution of perfusion fMRI is inherently poorer than BOLD fMRI because of the necessity to form tag and control images and to allow time for blood to be delivered from the tagging region to the imaging slice. In a typical PASL experiment, a repetition time (TR) of 2 s or more is used, so that one tag and control image pair is acquired every 4 s. By comparison, the temporal resolution of BOLD fMRI is typically 1–2 s and can be as low as 100 ms for specialized applications. Methods for improving temporal resolution include turbo-ASL and single-shot ASL. In turbo-ASL, a control image is acquired after the inversion pulse at a delay (typically 100 ms) that is too small for blood to have flowed into the imaging slice. A tag image is then acquired at an interval of *TR* (typically 1 s) after the control image. In single-shot ASL, background suppression is used to suppress static tissue, thus eliminating the need for a control image. While these methods can improve temporal resolution, the quantification of the resultant ASL signals is more complicated than are the more standard methods.

SNR and Contrast-to-Noise Ratio

As discussed on the first page, the SNR of ASL methods is inherently low because the inflowing blood magnetization is typically only about 1% of the tissue. The low SNR combined with the poor temporal resolution of ASL results in a low contrast-to-noise ratio (CNR) for quantitative ASL fMRI experiments, as compared to BOLD. Due to the low inherent SNR of ASL methods, noise reduction methods (such as the physiological noise reduction method discussed earlier) should be used wherever possible. Background suppression methods that attenuate the static tissue component also have the potential to reduce noise, but their application for simultaneous CBF/BOLD experiments is not straightforward. For functional mapping experiments in which quantification of CBF is not required, nonquantitative versions of ASL, such as turbo-ASL and close-tag CASL, can be used to improve the CNR. For example, the CNR of close-tag CASL at 1.5 T has been shown to approach that of the BOLD signal for a finger-tapping experiment.

In brain regions such as the orbital frontal cortex, the presence of pronounced magnetic susceptibility gradients can lead to severe signal dropouts in gradient-echo BOLD imaging, due to the long echo times employed. Because perfusion fMRI measures do not require BOLD contrast, these signal dropouts can be reduced through the use of either short echo time

readouts or a refocusing pulse sequence. The relative SNR of perfusion fMRI measures can therefore be higher than that of BOLD measures in these brain regions.

Spatial Coverage

At present, most ASL fMRI studies use two-dimensional (2-D) multislice acquisition methods in which the data are acquired on a slice-by-slice basis. For CASL and PASL methods, the slices are typically acquired in an inferior-to-superior manner so that slices that are further from the tagging region are acquired at later times. The number of slices that can be acquired depends on (1) the acquisition time for each slice, (2) the time TR – TI that is available for slice acquisition, and (3) the need to acquire slices before the difference signal ΔM has decayed away due to the longitudinal relaxation of blood. Due to these considerations, ASL studies typically acquire a smaller number of slices (e.g., 3–15) than do whole-brain BOLD studies (e.g., 30–40 slices), with thicker slices (e.g., 5–8 mm) than would be used in a BOLD study (e.g., 3–4 mm). Recently, ASL fMRI studies with single-shot 3-D acquisitions have been demonstrated. The 3-D acquisition methods with refocusing pulse trains can reduce susceptibility artifacts, are well-suited to the use of background suppression methods, and are a natural match for the nonspatial tagging scheme used in VS-ASL. However, 3-D methods are not as well suited for simultaneous measurements of CBF and BOLD.

Applications of ASL in fMRI

Quantitative Assessment of Brain Activity

Because it can provide a quantitative measure of a fundamental physiological quantity (i.e., CBF), ASL has the potential to better reflect neural activity as compared to BOLD, which is a complex function of a number of physiological variables. As compared to BOLD, ASL measures of functional CBF changes have been shown to exhibit decreased interindividual and intersession variability and a more linear relation to neural stimulus, possibly reflecting a more direct link between CBF and neural activity.

Although the majority of studies using perfusion fMRI have focused on the visual and motor cortices, there have been a growing number of studies aimed at measuring the CBF response to more complex cognitive tasks. Perfusion fMRI has been used to successfully detect CBF responses associated with language generation, visual attention, memory encoding, working memory, and sequence learning.

When combined with BOLD measures, ASL measures of CBF can be used to derive estimates of functional changes in the cerebral metabolic rate of oxygen ($CMRO_2$), which has been shown in animal studies to be correlated with functional changes in neuronal spiking frequency. In human fMRI studies, the addition of simultaneous measurements of CBF and BOLD responses to a hypercapnic stimulus (administered carbon dioxide) can provide information that is then used to calibrate the BOLD response and estimate functional changes in $CMRO_2$. This approach has been used to demonstrate a linear coupling in the visual cortex between functional changes in CBF and $CMRO_2$. The use of combined CBF and BOLD measures is especially important for fMRI studies of populations, because factors such as medication, age, and disease can complicate interpretation when only BOLD measures are acquired.

Simultaneous measures of CBF and BOLD with ASL are also useful for understanding the mechanisms underlying the BOLD response. For example, ASL measurements have been used to examine whether poststimulus undershoots in the BOLD signal are due primarily to a slow return to baseline in cerebral blood volume or a prolonged elevation of $CMRO_2$.

Insensitivity to Low-Frequency Noise

Low-frequency drifts are present in most fMRI time series data and tend to reduce the statistical power of BOLD experiments. As discussed in a previous section, perfusion time series are formed from the filtered subtraction of control and tag images. This subtraction process greatly attenuates low-frequency drifts and can make ASL more sensitive than BOLD for experimental paradigms with long periods. For example, the insensitivity of ASL to low-frequency drifts has been used to study increases in CBF that persisted for at least 8 min after the completion of a stressful task. In another experiment demonstrating the stability of ASL measurements, patterns of activation were detectable when comparing resting images on one day with activation images on the next day, a comparison that would not be possible with BOLD imaging.

Spatial Localization

Because CBF directly measures the amount of arterial blood that has been delivered to the capillary beds, it has been hypothesized that the functional CBF signal may be more localized to the sites of neuronal activation than the BOLD signal. With the quantitative ASL methods described above, the measured CBF signal is

primarily from small arterioles, capillaries, and brain tissue. There is relatively little signal in the arteries because the inversion delays are chosen to allow enough time for the blood to flow to the target tissue. In addition, there is relatively little outflow of the tagged blood into the venous circulation, because (1) the inversion delays are on the order of the T_1 of blood, which is typically less than the capillary transit time, and (2) there is rapid exchange of water between the vascular and tissue compartments, which effectively slows the outflow. By contrast, the BOLD signal derives primarily from veins and their surrounding tissues. In work examining the visual orientation columns in the cat brain, it was found that the ASL signal provided a more convincing functional map of these very high-resolution structures than did the BOLD signal.

Summary

Perfusion fMRI using ASL offers a useful complement to standard BOLD techniques. ASL methods provide a quantitative measure of CBF that has the potential to better characterize the magnitude and localize the sites of the underlying neuronal activity. Two important drawbacks of ASL techniques for fMRI are that both the temporal resolution and the sensitivity of ASL are lower than those of the BOLD contrast mechanism, so the choice of these methods should be tailored carefully to the biological questions at hand, or used together to provide additional physiological or metabolic information.

See also: Alzheimer's Disease: MRI Studies; Cognition: An Overview of Neuroimaging Techniques; fMRI: BOLD Contrast; Neuroimaging; Positron Emission Tomography (PET); Statistical Tests and Inferences.

Further Reading

Aguirre GK, Detre JA, and Wang J (2005) Perfusion fMRI for functional neuroimaging. *International Review of Neurobiology* 66: 213–236.

Buxton RB (2002) *Introduction to Functional Magnetic Resonance Imaging*. Cambridge: Cambridge University Press.

Buxton RB, Frank LR, Wong EC, et al. (1998) A general kinetic model for quantitative perfusion imaging with arterial spin labeling. *Magnetic Resonance in Medicine* 40(3): 383–396.

Calamante F, Thomas DL, Pell GS, et al. (1999) Measuring cerebral blood flow using magnetic resonance imaging techniques. *Journal of Cerebral Blood Flow and Metabolism* 19(7): 701–735.

Detre JA, Leigh JS, Williams DS, et al. (1992) Perfusion imaging. *Magnetic Resonance in Medicine* 23: 37–45.

Detre JA and Wang J (2002) Technical aspects and utility of fMRI using BOLD and ASL. *Clinical Neurophysiology* 113(5): 621–634.

Golay X, Hendrikse J, and Lim TC (2004) Perfusion imaging using arterial spin labeling. *Topics in Magnetic Resonance Imaging* 15(1): 10–27.

Liu TT and Brown GG (2007) Measurement of cerebral perfusion with arterial spin labeling: Part 1: Methods. *Journal of the International Neuropsychological Society* 13: 526–538.

Liu TT and Wong EC (2005) A signal processing model for arterial spin labeling functional MRI. *Neuroimage* 24(1): 207–215.

Parkes LM (2005) Quantification of cerebral perfusion using arterial spin labeling: Two-compartment models. *Journal of Magnetic Resonance Imaging* 22(6): 732–736.

Wong EC, Buxton RB, and Frank LR (1998) A theoretical and experimental comparison of continuous and pulsed arterial spin labeling techniques for quantitative perfusion imaging. *Magnetic Resonance in Medicine* 40(3): 348–355.

Wong EC, Cronin M, Wu WC, et al. (2006) Velocity-selective arterial spin labeling. *Magnetic Resonance in Medicine* 55(6): 1334–1341.

Peripheral Circadian Oscillators

G Tosini, A Davidson, C Fukuhara, and K N Paul, Morehouse School of Medicine, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the past few years, the core circadian clock genes that make up the 'transcriptional/translational' feedback loop that generates circadian rhythms at the molecular level have been identified. We now know that although the master pacemaker in the suprachiasmatic nuclei (SCN) regulates the timing, either directly or indirectly, of most circadian rhythms in mammals, many organs/tissues of the body can generate circadian rhythms independently of the SCN *in vitro,* and are thus called 'peripheral circadian oscillators.'

These revolutionary developments have opened up an entirely a new area of biomedical research focused on the importance of internal timing for health and disease. In this article we review the current knowledge regarding peripheral circadian oscillators and the role that these oscillators may play in the regulation of health status of an organism.

The Circadian Clock in the Retina

The alternation of light and dark is the most reliable timing cue on our planet, and therefore it is not surprising that the retina has evolved a precise timing mechanism that allows it to anticipate and then to adapt to the more than 1 million-fold change in light intensity during a 24 h period.

The retina was the first extra-SCN oscillator to be discovered in mammals. Several studies have now demonstrated that many of the physiological, cellular, and molecular rhythms that are present within the retina are under the control of a circadian clock, or more likely a series of circadian clocks that are present within this tissue (Figure 1). For example, the disk shedding that occurs in the rod photoreceptors is under circadian control. Shedding persists in animals with SCN lesions or a transected optic nerve, indicating its independence from the central circadian pacemaker. Additional studies have reported that sensitivity to light-induced photoreceptor damage is modulated by the circadian clock via a cyclic adenosine monophosphate (cAMP)-dependent pathway. Other important retinal functions, such as visual sensitivity, are also under circadian control.

Although results from these studies suggested that retinal physiology was regulated by a circadian clock, they were not sufficient to conclude that an independent circadian pacemaker was located within the retinal tissue. The definitive demonstration of the presence of an autonomous retinal clock in mammals was achieved a few years ago when it was shown that a circadian rhythm of melatonin release persisted in mammalian retinas maintained in culture. In light/dark cycles, melatonin levels were high during the night and low during the day. In constant darkness, the circadian rhythm of melatonin release free-ran, exhibiting a period close to 24 h. The circadian rhythm of melatonin release in the retina can be entrained by light *in vitro* and is temperature compensated. Such results demonstrated that the retina can be considered a *bona fide* circadian pacemaker, since it satisfies the three fundamental properties (i.e., free-running, entrainment, and temperature compensation) that describe a circadian rhythm.

The retina is a heterogeneous tissue with many different types of specialized cells, several of which have been shown to express clock genes. Although we know that the circadian pacemaker driving melatonin synthesis is localized within the neural retina, we still do not know which cell types contain the circadian clock. The most recent experimental evidence indicates that the dopaminergic neurons, but not the photoreceptors, express mRNA for several clock genes and therefore these cells are the most likely candidates for the location of the retinal circadian pacemaker. Additional investigations have indicated that mutations in 'clock genes' influence the retinal clock a manner similar to that observed in the SCN, thus suggesting that the molecular mechanism responsible for the generation of circadian rhythms in these two structures is the same.

Previous work has clearly established that the mammalian circadian system is hierarchically organized and that the SCN is the master circadian oscillator that orchestrates circadian rhythms throughout the body (see later). New evidence suggests the presence of a similar hierarchical network of circadian clocks within the mammalian retina. In contrast, circadian rhythmicity of dopamine and its metabolites is not necessary for the generation of the melatonin circadian rhythm in photoreceptors. Thus, the circadian pacemaker controlling melatonin synthesis is insensitive to destruction of the inner retina, as well as to lesions of the SCN, and therefore should be considered the master circadian clock within the mammalian retina.

Finally, it must be mentioned that the retina plays a vital role in circadian organization of the whole organism, since it is the only source of photic information to the SCN (and hence to the rest of the body).

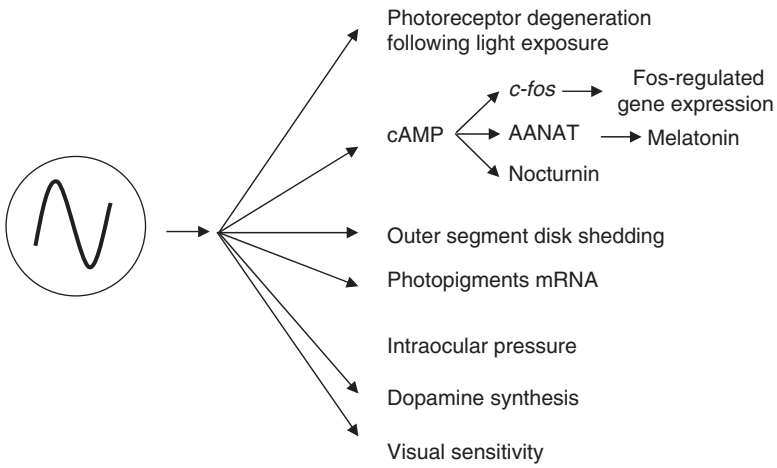

Figure 1 The circadian clock in the mammalian retina controls several rhythms that are important for the functioning of the retina and for the health of the photoreceptors. cAMP, cyclic adenosine monophosphate; AANAT, arylalkylamine *N*-acetyltransferase.

Removal of the retinas may affect the stability and the period of behavioral rhythms as well as the rhythmicity in some parts of the SCN.

Peripheral Circadian Oscillators

During the past decade our understanding of the functional organization of the mammalian circadian system has undergone significant advances. The cloning of putative clock genes has provided new tools with which to study mammalian circadian organization. It has become apparent that the expression of these genes is not restricted to the SCN and the retina, but instead they are expressed throughout the body. Furthermore, in many organs and tissues the genes are rhythmically expressed with a pattern similar to that seen in the SCN, albeit delayed by several hours. While some studies using population sampling methods have indicated that the rhythmic expression of these genes in peripheral tissues disappears in animals in which the SCN is removed, controversy remains, since these studies would not be able to detect a rhythm among animals no longer synchronized within the group.

The view that the SCN drives the rhythms in peripheral structures has been challenged by two seminal discoveries. First, Schibler's team reported that a serum shock can induce circadian oscillation in immortalized fibroblasts. Later, Menaker's group developed a transgenic rat in which the Period 1 (*PER1*) promoter was fused to the firefly luciferase gene to provide a real-time index of gene transcription in living tissue. This new powerful tool was used to demonstrate that many organs and tissues contain oscillators that are capable of generating circadian rhythms *in vitro*. Altogether, these early studies demonstrated three important points: (1) circadian oscillation in mammals can be produced by nonneuronal cells, (2) circadian oscillators are widespread within the mammalian body, and (3) only the SCN is capable of producing long-lasting circadian oscillations (Figure 2). However, more recent evidence has suggested that this third conclusion may not be true. A study using an engineered mouse that carries the luciferase enzyme fused to the Period 2 clock gene (*PER2*) has suggested that peripheral oscillators remain rhythmic in the absence of the SCN, but are no longer coherently organized.

The control and coordination of peripheral oscillators appears to differ among tissues. For example, a 6 h phase shift (delayed or advanced) of the light/dark cycle immediately resulted in phase shifts in the Period 1–luciferase activity rhythms in the SCN, while skeletal muscles and lung took 6 days to adjust to new light/dark regimens. The results obtained with the liver were even more surprising, since the delay phase shift was not completed even after 16 days in the new light cycle. The mechanisms responsible for such differences in the characteristics of circadian rhythms of the SCN and of the various peripheral tissues are still unknown. These results have been used to explain the malaise that results from rapid transmeridian travel and, more importantly, from shift-work schedules.

Aging has a variety of influences on rhythms in peripheral tissues, such as the advance of peak phase in the pineal gland and kidney or the complete loss of rhythmicity in the lung. Also, microarray studies have revealed that while between 8% and 12% of the genome is oscillating with high amplitude in each peripheral tissue, there is little overlap among these tissues with respect to which transcripts oscillate. This key finding suggests that peripheral rhythms serve local functions specific to the organs in which they are generated.

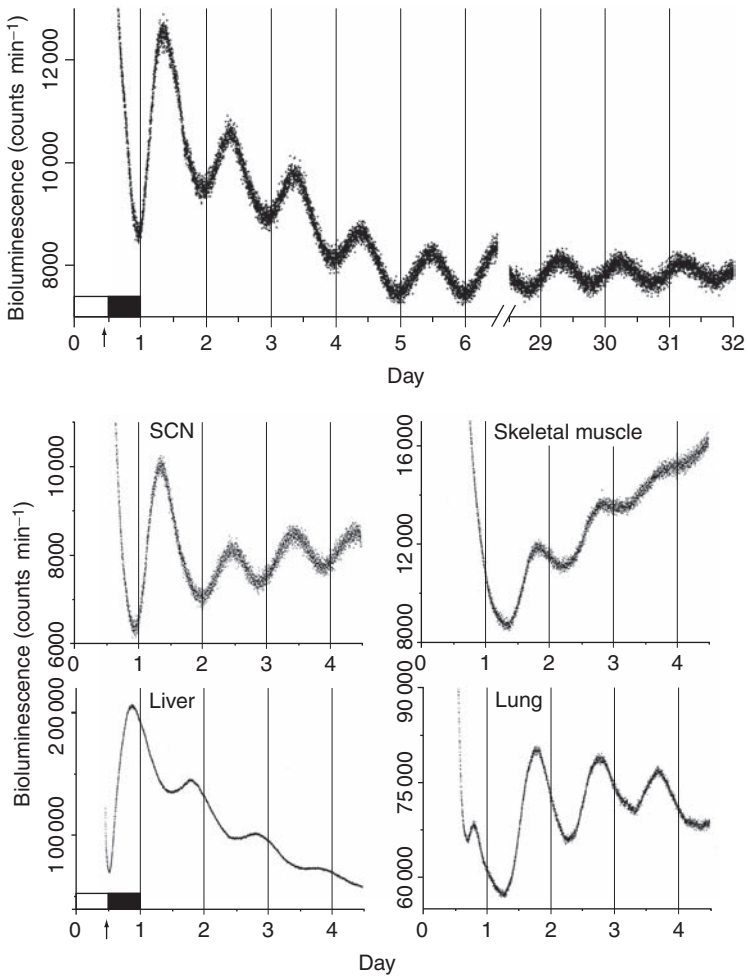

Figure 2 Representative circadian rhythms of bioluminescence from cultured suprachiasmatic nuclei (SCN) and peripheral organs explanted from a Period 1 and luciferase (*PER1–luc*) transgenic rat. Black and white bars show the animal's previous light/dark conditions. Rhythmicity in the SCN persisted for several cycles (top), but only for a few cycles in the peripheral organs. Reprinted with permission from Yamazaki S, Numano R, Abe M, et al. (2000) Resetting central and peripheral circadian oscillators in transgenic rats. *Science* 288: 682–685. copyright (2000) AAAS.

Characterization of human rhythmicity is of high therapeutic value to the diagnosis and treatment of disease. In particular, humans undergoing chronomodulated chemotherapy could be more effectively treated if clock phase and amplitude could be easily assayed for each patient. Using current approaches, the characterization of human circadian rhythmicity is complicated and expensive, since each experimental participant needs to be housed in a specially designed room within a hospital or research facility. Therefore, it would be of great value to devise new techniques that will allow for easier and more economical monitoring of circadian parameters in humans.

There is compelling evidence that the cellular and molecular mechanisms that generate circadian rhythms are similar between the SCN neurons and nonneuronal cells located in organs and tissues outside the brain. Therefore, circadian characterization of tissue biopsies has potential as a surrogate approach to studying the whole person. Indeed, recent work has demonstrated the feasibility of this approach. Human fibroblast cultures were monitored for circadian rhythmicity *in vitro*. From this initial study it emerged that the period lengths of human fibroblasts showed a greater interindividual variation than what has been observed in studies at the whole organism level. This new technology represents a major step forward in our ability to monitor human circadian rhythms and will allow us to develop new diagnostic tools to screen for genetic mutations as well to assess the effects of shift work, jet lag, and sleep deprivation on the circadian system.

Synchronization of Peripheral Oscillators

A poorly understood aspect of circadian organization in mammals is the means by which the SCN synchronizes circadian rhythms in the rest of the body (Figure 3). Experimental evidence suggests that both humoral and neural factors are involved in this process. Previous studies have shown that several humoral factors may act as synchronization factors. In the central nervous system, transforming growth factor-α (TGF-α) is one, but probably not the only, diffusible factor responsible for the transmission of circadian information from the SCN to the other brain regions. In the periphery it appears that glucocorticoids, angiotensin II, and retinoic acid may each play a role in the entrainment of peripheral oscillators. The results from studies examining glucocorticoids have shown that the analog dexamethasone can induce circadian gene expression in cultured rat-1 fibroblasts, and systemic injection in the mouse affects the phase of circadian gene expression in the liver, kidney, and heart, but not in SCN neurons.

It is worthwhile to mention a new study in which the authors, using a parabiosis approach whereby two animals share a blood supply but have independent nervous systems, have shown that nonneural (behavioral or bloodborne) signals are adequate to maintain circadian rhythms of clock gene expression in liver and kidney, but not in the heart, spleen, or skeletal muscle. Furthermore, although SCN transplants may restore the circadian rhythms in locomotor activity, this procedure fails to restore circadian rhythms in melatonin, serum corticosterone, and cortisol levels as well as gonadal and testicular regression, thus suggesting that the SCN must employ more than just humoral signals to control peripheral rhythmic events. Indeed, new neuroanatomical data using retrograde tracing techniques have demonstrated that the SCN projects to several organs, including the hypothalamic–pituitary–adrenal axis, the hypothalamic–pituitary–gonadal axis, the hypothalamic–pineal axis, and the hypothalamic–pancreatic axis. Taken together, these results indicate that the SCN orchestrates the expression of circadian rhythms in different peripheral organs by diverse pathways, and more importantly that these signals can be very specific in their target action.

Behavioral Rhythmicity Driven by Structures outside the SCN

So far we have described endogenously generated rhythms in retinal melatonin synthesis and period 1 and period 2 gene expression in peripheral tissues. These rhythms likely have local action, organizing physiology within the time domain in a specific manner for each tissue. However, circadian rhythmicity generated outside the SCN also has the capacity to influence global physiology and behavior. Food-anticipatory activity arising in animals fed a single daily meal and chronic methamphetamine-induced rhythmicity are two circadian behaviors that challenge the notion that the SCN plays a unique role at the top of the circadian hierarchy.

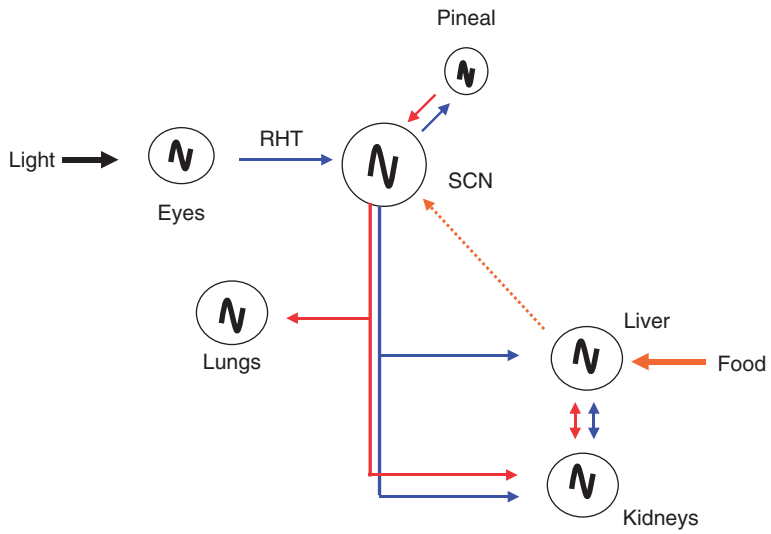

Figure 3 Schematic model representing the organization of the mammalian circadian system. The suprachiasmatic nuclei (SCN) receives the photic information from the retina via the retinohypothalamic tract (RHT), then the SCN conveys this information to other circadian oscillators via neural (blue lines) or humoral (red lines) pathways. Some peripheral oscillators can also influence the rhythmicity in the SCN via humoral (pineal) or neural (retina) pathways. Food can also acts as a potent synchronizing signal for the liver and for the SCN.

Food intake patterns are governed by endogenous and exogenous cues, and likely act to reinforce the cyclic nature of behavior and physiology. For example, nocturnal rats eat mostly at night. This food intake acts to control the circadian phase of digestive tissues such as the alimentary tract and the liver. Experimentally altering the timing of food intake has dramatic effects on physiology and circadian behavior. Restricting food access to a 4 h window during the day leads to expression of a circadian behavior termed food-anticipatory activity (FAA). This behavioral pattern is characterized by an increase in locomotor activity in the hours preceding meal access. This can be observed by measuring feeder approaches, wheel-running activity, general cage activity, and core body temperature increases. FAA in rats exhibits some of the hallmarks of a circadian rhythm, including persistence during food deprivation, resetting transients following an abrupt shift of the feeding time, and limits of entrainment in the circadian range such that a meal interval of 22–28 h will produce FAA, but an interval that is larger or smaller will not.

FAA has been observed in a wide variety of mammalian species, including several rodents, rabbits, marsupials, and primates. Among nonmammals, honeybees, birds, and fish have all been observed exhibiting FAA in some form. While the conserved nature of the ability to predict mealtime suggests its adaptive importance to organisms at many levels of the foodchain, its function is unknown. It has been speculated that for rodents at least, FAA is analogous to foraging. FAA may be an arousal signal that is activated several hours prior to the time at which food was successfully and safely found on previous days, thereby allowing lead time for the animal to leave its burrow and travel to that location.

While it is well established that the SCN is necessary for the generation of behavioral circadian rhythms entrained by environmental light, the mammalian SCN is not required for FAA. Lesions placed in the SCN, as well as lesions of many other brain structures, have failed to abolish this behavior. Studies have implicated the paraventricular nucleus, nucleus accumbens core, the parabrachial area, and, most recently, the dorsomedial hypothalamic nucleus in some behavioral measures of FAA. However, none of these regions seems to be completely necessary for the prediction of mealtime when multiple behavioral measures are considered.

While peripheral structures do quickly entrain to meal access when that meal is provided at times that are at odds with the light cycle, studies have suggested that these peripheral rhythms are independent from FAA. For example, one can observe FAA without entrainment of metabolism by providing highly palatable food on a schedule, but allowing *ad libitum* access to regular chow. Also, gastrointestinal tissues exhibit molecular rhythmicity that in some circumstances does not match the pattern of FAA.

The circadian clock responsible for FAA remains somewhat of a mystery. Another circadian behavior for which the underlying mechanism is totally unknown is chronic methamphetamine-induced circadian rhythmicity. First described by Shinichi Honma and his group in Sapporo, Japan, this behavior is typified by a long free-running rhythm in constant darkness or constant light, even in mice or rats with SCN lesions. It is induced by simply placing low-dose methamphetamine in the drinking water. If the animal is rhythmic, the rhythm will lengthen and in some cases split into two components with different periods. Presumably these two rhythms are the outputs of the SCN, and another circadian clock that has not yet been discovered but is currently identified as the methamphetamine-sensitive circadian oscillator (MASCO). If the animal is SCN lesioned or is arrhythmic from long-term exposure to constant light, a rhythm will emerge after a latency that can vary from 1 week to one or several months. The period of this rhythm is dependent on the concentration of methamphetamine in the drinking water.

It has been argued that the rhythms induced with chronic methamphetamine are the product of an hourglass timer, rather than a circadian oscillator. If a rat or mouse ingests water containing methamphetamine, activity will increase, leading to further water intake, lasting until the animal tires. Sleep is then followed by a new bout of water/drug intake and increased activity. If the period of this sleep/wake pattern was similar to 24 h, it would appear as if an endogenous rhythm was driving the behavior, even though it is actually caused by a self-reinforcing cycle of drug intake followed by rest. However, the methamphetamine-driven free-running rhythm exhibits characteristics that strongly suggest that it is driven by MASCO rather than by an hourglass, or interval-timing, mechanism. For example, an 8 h removal of methamphetamine from the water results in either no change in the rhythm or an advance of the phase. An interval mechanism would predict a phase delay. Also, there is evidence of persistence of the rhythm after sustained withdrawal of the methamphetamine, and of transients during entrainment to a methamphetamine delivery schedule. These observations are all consistent with the hypothesis that MASCO, a true circadian oscillator independent from the SCN, drives this behavior.

Since both methamphetamine-induced rhythmicity and FAA are non-SCN-based behavioral rhythms that are typified by hyperactivity, it has been speculated

that they arise as the output of the same, as-of-yet unidentified, circadian clock in the brain. If so, identification of the substrate for one will solve the mystery of the other, thereby greatly advancing our understanding of the mammalian circadian system.

Conclusions

Several studies in the past 10 years have demonstrated that circadian oscillators are widely distributed within the mammalian body and that possibly each cell within the body contains a circadian clock. However, these peripheral oscillators are organized by the central circadian pacemaker located in the SCN, indicating that the mammalian circadian system is hierarchically organized. The mechanisms by which the SCN controls peripheral oscillators, as well as the roles that these peripheral oscillators play in the maintenance of normal physiology and health, are topics of active investigation.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Organization in Non-Mammalian Vertebrates; Circadian Regulation in Invertebrates; Circadian Function and Therapeutic Potential of Melatonin in Humans; Melatonin Regulation of Circadian Rhythmicity in Vertebrates.

Further Reading

Balsalobre A, Damiola F, and Schibler U (1998) A serum shock induces circadian gene expression in mammalian tissue culture cells. *Cell* 93: 929–937.

Balsalobre A, Brown SA, Marcacci L, et al. (2000) Resetting of circadian time in peripheral tissues by glucocorticoid signaling. *Science* 289: 2344–2347.

Brown SA, Fleury-Olela F, Nagoshi E, et al. (2005) The period length of fibroblast circadian gene expression varies widely among human individuals. *PLoS Biology* 3: e338.

Davidson AJ (2006) The search for the feeding-entrainable circadian oscillator: A complex proposition. *AJP – Regulative, Integrative and Comparative Physiology* 290: R1524–R1526.

Davidson AJ, Yamazaki S, and Menaker M (2003) SCN: Ringmaster of the circadian circus or conductor of the circadian orchestra? *Novartis Foundation Symposium* 253: 110–121.

Davidson AJ, Poole A, Yamazaki S, et al. (2003) Is the food-entrainable oscillator in the digestive system? *Genes, Brain and Behavior* 2: 32–39.

Fukuhara C and Tosini G (2003) Role of peripheral oscillators in mammals. *Frontier in Bioscience* 8: 642–651.

Gooley JJ, Schomer A, and Saper CB (2006) The dorsomedial hypothalamic nucleus is critical for the expression of food-entrainable circadian rhythms. *Nature Neuroscience* 9: 398–407.

Guo H, Brewer JM, Champhekar A, et al. (2005) Differential control of peripheral circadian rhythms by suprachiasmatic-dependent neural signals. *Proceedings of the National Academy of Science of the United States of America* 102: 3111–3116.

Homma S, Kanematsu N, and Honma K (1992) Entrainment of methamphetamine-induced locomotor activity rhythm to feeding cycles in SCN-lesioned rats. *Physiology & Behavior* 52: 843–850.

Kramer A, Yang FC, Snodgrass P, et al. (2001) Regulation of daily locomotor activity and sleep by hypothalamic EGF receptor signaling. *Science* 294: 2511–2515.

Stephan FK (2002) The "other" circadian system: Food as a Zeitgeber. *Journal of Biological Rhythms* 17: 284–292.

Tataroglu Ö, Davidson AJ, Benvenuto LJ, et al. (2006) The methamphetamine-sensitive circadian oscillator (MASCO) in mice. *Journal of Biological Rhythms* 21: 185–194.

Tosini G and Fukuhara C (2002) The mammalian retina as a clock. *Cell and Tissue Research* 309: 119–126.

Yamazaki S, Numano R, Abe M, et al. (2000) Resetting central and peripheral circadian oscillators in transgenic rats. *Science* 288: 682–685.

Yoo SH, Ko CH, Lowrey PL, et al. (2005) A noncanonical E-box enhancer drives mouse Period2 circadian oscillations *in vivo*. *Proceedings of the National Academy of Science of the United States of America* 102: 2608–2613.

Peripheral Nerve Regeneration: An Overview

P M Richardson, Queen Mary, University of London, London, UK

© 2009 Published by Elsevier Ltd.

Spontaneous Recovery of Function in the Central Nervous System

Following axonal injury in the adult mammalian nervous system, some neurological function is usually regained spontaneously over a period of months, despite the absence of functional axonal regeneration. The importance of alternative pathways in such recovery of neurological function can be demonstrated by a second lesion, after initial recovery of function. For example, after lesion of the deep cerebellar nuclei, monkeys initially develop tremor and lose coordination, and then slowly recover. This recovered function is then irretrievably lost if somatosensory regions of the cerebellar cortex are ablated at a second intervention. The efficacy of compensation by spared neural elements varies among neurological pathways and among functions mediated by given pathways. For example, after division of the corticospinal tract in rats, locomotion is transiently impaired whereas fine reaching movements of the forelimb never recover.

The formation of new synapses in the adult mammalian brain was convincingly proved for the first time in an electron microscopic study in 1969. New synapses from spared afferents are formed on deafferentated postsynaptic sites in the septal nucleus, hippocampus, red nucleus, spinal cord, and many other neuronal nuclei in the central nervous system (CNS) and are electrophysiologically functional. A second pattern of sprouting is seen proximal to the site of axonal transection, analogous to growth after horticultural pruning. In the spinal cord, new connectivity can provide a polysynaptic pathway to replace a monosynaptic pathway. The capacity to form new connections through extensive axonal elongation, over micrometers rather than millimeters, accounts for the recovery of function that is anticipated after partial spinal cord injury but which is exceptional after clinically complete loss of function.

Other forms of synaptic plasticity are demonstrable. With contemporary imaging techniques, is now possible to visualize in real time retraction and extension of dendritic spines in hippocampal slice preparations. Synaptic plasticity may be molecular rather than morphological, a classic example being long-term potentiation in the hippocampus, a brain-derived neurotrophic factor-dependent phenomenon whereby postsynaptic responses are enhanced by previous electrical activity. This type of synaptic plasticity underlies learning, and presumably relearning, after neural injury. Disinhibition of synapses within minutes of injury can unmask previously silent connections. For example, after unilateral cervical spinal cord injury, activity in the ipsilateral phrenic nerve ceases only to reappear immediately after a second injury to the contralateral phrenic nerve. In the 'crossed phrenic reflex,' preexisting connections between contralateral descending fibers and phrenic motorneurons are believed to be activated by a hypoxic stimulus.

Another potential mechanism of spontaneous recovery of function is remyelination of axons demyelinated by inflammation or trauma. Demyelinated axons not only fail to conduct impulses but also are vulnerable to Wallerian degeneration because of loss of trophic support. Endogenous oligodendrocytes are capable of remyelination in toxic, inflammatory, and traumatic animal models of demyelination.

Morphological and molecular synaptic plasticity and, to a lesser extent, remyelination are major mechanisms of recovery of function after neural injury. The potential to enhance such spontaneous recovery by electrical activity is demonstrated by observations that locomotion in cats with spinal cord lesions can be improved by early assisted training on a treadmill.

Peripheral Nerve Regeneration

Cellular events underlying Wallerian degeneration and axonal regeneration have been well characterized for many years. Distal to the site of injury, axons degenerate and myelin disintegrates over several days and the debris is phagocytosed by macrophages within 2 weeks (in the rat). The environment created by Schwann cells, macrophages, fibroblasts, and other endoneurial cells supports axonal regeneration at a rate of several millimeters per day.

During Wallerian degeneration, Schwann cells increase in number and become aligned in bands of Büngner. They stop synthesizing mRNAs for myelin protein, ciliary neurotrophic factor, and other proteins and start to synthesize another group of mRNAs, such as those for p75 nerve growth factor receptor and monocyte chemoattractant protein-1. The Schwann cell phenotype during Wallerian degeneration, distinct from that of immature Schwann cells or myelinating or unmyelinating Schwann cells in an intact nerve, is determined by inflammatory cytokines and the loss of axonal signals to Schwann cells. If Schwann cells are deprived of axons for months,

they eventually atrophy and die and the endoneurium ceases to support axonal regeneration well. The prototypic, but not unique, axon-to-Schwann cell signaling pathway, comprising neuregulin and its erbB2–erbB3 receptor dimer, stimulates Schwann cell survival, mitosis, and myelin formation. Two classes of molecules in the endoneurium promote axonal regeneration. Components of the extracellular matrix, in particular laminin, guide and stimulate axons through interaction with their integrins. Diffusible neurotrophic molecules such as nerve growth factor, brain-derived neurotrophic factor, and leukemia inhibitory factor have local actions on the growth cone and retrogradely influence the nerve cell body: their synthesis is strongly increased during Wallerian degeneration. The ability of Schwann cells to redifferentiate in response to axonal degeneration is one critical component of the capacity for regeneration in the peripheral nervous system.

Macrophages in the endoneurium of injured nerves increase in number through recruitment of macrophages and proliferation of resident cells. Macrophages also accumulate around axotomized neurons in dorsal root ganglia. Leukocyte infiltration into the peripheral nervous system after injury is selective for monocytes/macrophages, with minimal involvement of lymphocytes and polymorphonuclear neutrophils: it is regulated by chemokines and cell adhesion molecules that mediate interactions between leukocytes and the vascular endothelium, in particular, monocyte chemoattractant protein-1. Macrophages phagocytose axonal and myelin debris and contribute to the induction of neurotrophic factors, in particular, nerve growth factor. If macrophage invasion into the distal nerve segment is prevented, then axonal regeneration is impaired.

Another series of morphological changes is observed in the cell bodies of axotomized neurons in dorsal root ganglia or in the spinal cord or brain stem. Neurons may die after division of their major peripheral axons, the likelihood being greater if the injury is close to the cell body or if the animal is young. 'Chromatolysis' of neurons refers to the dispersion of ribosomes and peripheral displacement of the nucleus. That inducible events in the nerve cell bodies contribute to axonal regeneration is demonstrated by a variety of 'conditioning' phenomena whereby axonal regeneration is accelerated by a previous or additional axonal injury. Regeneration of the peripheral axons of primary sensory neurons is accelerated by a more distal nerve injury 1 week before the crush injury and regeneration of the central axons of these bipolar neurons is strongly enhanced by a concomitant injury to the corresponding peripheral axons. The propensity for regeneration induced in neurons by axonal injury persists for months, but not indefinitely. Changes in gene expression in neurons following nerve injury include the appearance of messenger ribonucleic acids (mRNAs) for regeneration-associated genes and alteration in either direction of neuropeptide mRNAs. These changes are triggered by changes in retrograde axonal transport. For example, the retrograde influence of nerve growth factor is decreased at least until axons encounter a compensatory source in the distal nerve segment, and this deprivation accounts for a decrease in neurofilament mRNA and many of the fluctuations of neuropeptide mRNAs. The retrograde axonal signals that induce regeneration-associated genes are less well defined. The products of regeneration-associated genes in axotomized sensory neurons include transcription factors (c-Jun), growth factors (brain-derived neurotrophic factor and interleukin-6), growth cone proteins (growth-associated protein-43), and proteins of less well-known function, such as small proline-rich repeat 1A, and their induction is believed to underlie the propensity for regeneration in these neurons. The ability of neurons to redifferentiate is a second factor in the capacity for regeneration in the peripheral nervous system (PNS).

A major limitation to the results of nerve repair stems from the lack of guidance cues to lead axons to their former destinations. Functional recovery is much better after crush injury with preservation of endoneurial architecture than after repair by nerve suture or nerve graft when this physical guidance is lost. Although some cues specific to motor or sensory axons appear to exist, their ability to restore appropriate connections with the periphery is limited. The most important determinant of recovery following nerve repair in human beings is the nature of the injured nerve. For example, the radial nerve with relatively few sensory fibers and an uncomplicated pattern of muscle innervation recovers much better than does the ulnar nerve, a mixed nerve innervating diverse muscles.

Peripheral nerve injury evokes changes in connectivity in the spinal cord, thalamus, and cerebral cortex. Particularly after partial nerve injury, synaptic rearrangement in the dorsal horn of the spinal cord may lead to painful phenomena, including allodynia and hyperesthesia. After a neurotization procedure providing novel peripheral connections (for example, between the intercostal nerve and the biceps), plasticity in the cerebral cortex may improve functional recovery.

Neural Repair in the CNS

Cellular and molecular strategies to repair the brain or spinal cord have been devised. Cellular therapies aim to replace either neurons or glial cells, and

molecular therapies aim to stimulate the nerve cell body or counteract inhibitory influences in CNS glia.

Replacement of neurons, appropriate to the treatment of neurodegenerative disorders, was pioneered by Bjorklund and Stenevi. Following extensive animal studies, fetal mesencephalic tissue was transplanted into the striatum of patients with Parkinson's disease. The cells survived and yielded modest, long-lasting clinical improvement: possible mechanisms of action include integration of grafted neurons into host circuits, release of dopamine by the grafted cells, and beneficial humoral actions of fetal neural tissue. More limited success has been obtained through the grafting of fetal neural tissue into the brains of patients with Huntington's disease. Because of the considerable practical difficulties in obtaining human fetal nervous tissue for transplantation, attempts have been made to generate neuronal cell lines from embryonic stem cells. Also, efforts are under way to enhance the proliferation and differentiation of stem cells indigenous to the adult mammalian brain. Although proof of the principle has been obtained that neuronal replacement can improve or arrest decline of function in neurodegenerative disease, neuroprotective measures such as local administration of neurotrophic factors perhaps hold more hope for future success.

Glial cell transplantation to promote axonal growth in the CNS is based on the premise that such growth could be supported better by replacing astrocytes and oligodendrocytes with more permissive cell types. Reports a century ago that CNS axons could regenerate into a peripheral nerve graft were confirmed many years later by new axonal tracing techniques. It is now apparent that axons from many but not all classes of CNS neurons can grow into the environment provided by Schwann cells and other endoneurial cells. However, two factors have severely limited the success of this strategy in promoting functional recovery. First, the percentage of axons from any given CNS fiber tract that grow into a peripheral nerve graft is small, particularly if the injury is distant from the nerve cell body. Second, axons that enter such grafts have limited ability to exit from them and form new synapses. Olfactory ensheathing cells hold promise to overcome the second obstacle. They are a mixed population of Schwann cells and other cells which support olfactory system axons with the unique ability to grow from the PNS into the CNS in adult mammals. Whereas axons regenerating in dorsal spinal nerve roots fail to penetrate the PNS–CNS interface at the root entry zone, axons regenerating in the olfactory nerve successfully enter and form synapses in the olfactory bulb. Although the cellular and molecular differences between cells in the endoneurium of peripheral nerves and olfactory ensheathing cells have not been fully defined, the unique ability of the latter cells to shepherd axonal growth into the CNS provides a rationale for using them in repair of the CNS. Indeed, transplantation of olfactory ensheathing cells after partial spinal cord injury in rats has led to remarkable recovery of motor function. The rapidity of behavioral improvement, within 10 days, suggests that the recovery is likely due to short-distance sprouting rather than to long-distance regeneration. Other peripheral ensheathing cells – for example, from immature animals – may also support axonal growth across CNS injury sites better than do adult Schwann cells. A less daunting challenge for glial cell transplantation is to restore conductivity to nerve fibers demyelinated by inflammation or trauma. Schwann cells, olfactory ensheathing cells, and oligodendrocyte precursor cells have all shown some promise in this regard.

Two major classes of molecules that inhibit axonal growth in adult CNS glial tissue are myelin-associated proteins and chondroitin sulfate proteoglycans. Three proteins associated with CNS myelin, myelin-associated glycoprotein, Nogo, and oligodendrocyte myelin protein, each strongly inhibits neurite growth *in vitro*. All three act through a receptor complex comprising the Nogo receptor and p75 neurotrophin receptor, which activates the small GTPase Rho and its downstream target Rho-associated kinase, which, in turn, inhibits growth cone extension. Local administration of antibodies to the Nogo receptor and of Rho-associated kinase inhibitors have been found to promote regeneration and recovery of function after spinal cord injury in rodents. Curiously, mice with null deletion of the Nogo receptor do not show enhancement of regeneration of spinal axons. Likewise, delivery of chondroitinase to the injured spinal cord improves both functional recovery and axonal regeneration. For both types of treatment, the improvement of functional recovery precedes and exceeds the enhancement of axonal regeneration and may be mediated through increased synaptic plasticity.

A second impediment to axonal regeneration in the CNS is the difficulty in raising a strong regenerative response in axotomized neurons. Some success has been achieved in enhancing regenerative responses in CNS neurons through neurotrophic factors and small molecules. For example, infusion of brain-derived neurotrophic factor in the red nucleus induces regeneration-associated genes in the neurons and increases the regeneration of rubrospinal axons into a peripheral nerve graft, and injection into the globe of a viral vector delivering ciliary neurotrophic factor increases axonal regeneration from retinal ganglion cells. In general, neurotrophic factors must be

delivered locally in the vicinity of nerve cell bodies, as their side effects upon systemic administration are prohibitive. Cyclic adenosine monophosphate (cAMP) also increases the propensity of neurons to regenerate *in vitro* and *in vivo*, and the need for local delivery can be obviated by systemic administration of a phosphodiesterase inhibitor. Research in the past generation has revealed unexpected plasticity and capacity for recovery in the adult mammalian CNS and has characterized some of the underlying molecular, cellular, and physiological mechanisms.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Injury: Neuronal Responses; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Neuronal Plasticity after Cortical Damage; Nogo-A: Its Role in Axon Regeneration; Sensory Re-education; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants.

Further Reading

Bareyre FM, Kerschensteiner M, Raineteau O, et al. (2004) The injured spinal cord spontaneously forms a new intraspinal circuit in adult rats. *Nature Neuroscience* 7: 269–277.

Barnett SC and Riddell JS (2004) Olfactory ensheathing cells (OECs) and the treatment of CNS injury: Advantages and possible caveats. *Journal of Anatomy* 204: 57–67.

Bjorklund A and Stenevi U (1979) Reconstruction of the nigrostriatal dopamine pathway by intracerebral nigral transplants. *Brain Research* 177: 555–560.

Bonilla IE, Tanabe K, and Strittmatter SM (2002) Small proline-rich repeat protein 1A is expressed by axotomized neurons and promotes axonal outgrowth. *Journal of Neuroscience* 22: 1303–1315.

Bradley WG and Asbury AK (1970) Duration of synthesis phase in neuilemma cells in mouse sciatic nerve during degeneration. *Experimental Neurology* 26: 275–282.

Brushart TM, Gerber J, Kessens P, et al. (1998) Contributions of pathway and neuron to preferential motor reinnervation. *Journal of Neuroscience* 18: 8674–8681.

D'Antonio M, Michalovich D, Paterson M, et al. (2006) Gene profiling and bioinformatic analysis of Schwann cell embryonic development and myelination. *Glia* 53: 501–515.

Dahlin LB (1995) Prevention of macrophage invasion impairs regeneration in nerve grafts. *Brain Research* 679: 274–280.

Dailey AT, Avellino AM, Benthem L, et al. (1998) Complement depletion reduces macrophage infiltration and activation during Wallerian degeneration and axonal regeneration. *Journal of Neuroscience* 18: 6713–6722.

Fu SY and Gordon T (1995) Contributing factors to poor functional recovery after delayed nerve repair: Prolonged axotomy. *Journal of Neuroscience* 15: 3876–3885.

Fu SY and Gordon T (1995) Contributing factors to poor functional recovery after delayed nerve repair: Prolonged denervation. *Journal of Neuroscience* 15: 3886–3895.

Goldberger ME and Growdon JH (1973) Pattern of recovery following cerebellar deep nuclear lesions in monkeys. *Experimental Neurology* 39: 307–322.

Heumann R, Lindholm D, Bandtlow C, et al. (1987) Differential regulation of mRNA encoding nerve growth factor and its receptor in rat sciatic nerve during development, degeneration, and regeneration: Role of macrophages. *Proceedings of the National Academy of Sciences of the United States of America* 84: 8735–8739.

Kobayashi NR, Fan D-P, Giehl KM, et al. (1997) BDNF and NT-4/5 prevent atrophy of rat rubrospinal neurons after cervical axotomy, stimulate GAP-43 and Tα1-tubulin mRNA expression, and promote axonal regeneration. *Journal of Neuroscience* 17: 9583–9595.

Li Y, Decherchi P, and Raisman G (2003) Transplantation of olfactory ensheathing cells into spinal cord lesions restores breathing and climbing. *Journal of Neuroscience* 23: 727–731.

Lindvall O, Sawle G, Widner H, et al. (1994) Evidence for long-term survival and function of dopaminergic grafts in progressive Parkinson's disease. *Annals of Neurology* 35: 172–180.

Lu X and Richardson PM (1993) Responses of macrophages in rat dorsal root ganglia following peripheral nerve injury. *Journal of Neurocytology* 22: 334–341.

Malessy MJ, Bakker D, Dekker AJ, et al. (2003) Functional magnetic resonance imaging and control over the biceps muscle after intercostal-musculocutaneous nerve transfer. *Journal of Neurosurgery* 98: 261–268.

McQuarrie IG and Grafstein B (1973) Axon outgrowth enhanced by a previous nerve injury. *Archives of Neurology* 29: 53–55.

Moreno DE, Yu X-J, and Goshgarian HG (1992) Identification of the axon pathways which mediate functional recovery of a paralyzed hemidiaphragm following spinal cord hemisection in the adult rat. *Experimental Neurology* 116: 219–228.

Mueller M, Leonhard C, Wacker K, et al. (2003) Macrophage response to peripheral nerve injury: The quantitative contribution of resident and hematogenous macrophages. *Laboratory Investigation* 83: 175–185.

Nguyen QT, Sanes JR, and Lichtman JW (2002) Pre-existing pathways promote precise projection patterns. *Nature Neuroscience* 5: 861–867.

Raisman G (1969) Neuronal plasticity in the septal nuclei of the adult rat. *Brain Research* 14: 25–48.

Richardson PM and Issa VMK (1984) Peripheral injury enhances central regeneration of primary sensory neurons. *Nature* 309: 791–793.

Richardson PM, McGuinness UM, and Aguayo AJ (1980) Axons from CNS neurones regenerate into PNS grafts. *Nature* 284: 264–265.

Rossignol S, Bouyer L, Langlet C, et al. (2004) Determinants of locomotor recovery after spinal injury in the cat. *Progress in Brain Research* 143: 163–172.

Sabel BA and Schneider GE (1988) The principle of "conservation of total axonal arborizations": Massive compensatory sprouting in the hamster subcortical visual system after early tectal lesions. *Experimental Brain Research* 73: 505–518.

Subang MC and Richardson PM (2001) Influence of injury and cytokines on synthesis of monocyte chemoattractant protein-1 mRNA in peripheral nervous tissue. *European Journal of Neuroscience* 13: 521–528.

Verge VMK, Richardson PM, Wisenfeld-Hallin Z, et al. (1995) Differential influence of nerve growth factor on neuropeptide expression *in vivo*. *Journal of Neuroscience* 15: 2081–2096.

Taveggia C, Zanazzi G, Petrylak A, et al. (2005) Neuregulin-1 type III determines the ensheathment fate of axons. *Neuron* 47: 681–694.

Thomas PK (1974) Nerve injury. In: Bellairs R and Gray EG (eds.) *Essays on the Nervous System*, pp. 44–70. Oxford: Clarendon.

Woolf CJ (1983) Evidence for a central component of post-injury pain sensitivity. *Nature* 306: 686–688.

Perirhinal Cortex

G W Van Hoesen and A Solodkin, University of Iowa, Iowa City, IA, USA

© 2009 Published by Elsevier Ltd.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

The perirhinal cortex in terms of topography can be localized 'around or in the vicinity of' the rhinal sulcus in all mammals and marsupials. Almost always, it occupies the depths of the sulcus and/or its lateral bank, interposed between the entorhinal cortex medially and the neocortex along the lateral surface of the cerebral hemisphere. In nonprimates it is a narrow strip of cortex seemingly only a few cell diameters in width, but in primates and, especially, humans it is a sizable area of cortex. The location of the perirhinal cortex in the human brain can be somewhat confusing since the rhinal sulcus is shallow and truncated typically, bordering only the anteriormost parts of the entorhinal cortex. This results in the repositioning of the perirhinal cortex posteriorly to the medial bank of the collateral sulcus when the rhinal sulcus ends (Figure 1(a)). Nevertheless, it continues to separate the entorhinal cortex from the neocortex and is 'peri' to these rather than to the rhinal sulcus *per se*. Braak and Braak introduced the term 'transentorhinal cortex' to account for this arrangement in the human brain and to emphasize what they regarded as the transitional nature of this cortex.

The cytoarchitecture of the perirhinal cortex is unique, and Brodmann designated it as area 35, setting it apart from the entorhinal cortex, or Brodmann's area 28, medially and the medial part of the inferior temporal gyrus, or Brodmann's area 36, laterally. It lacks a well-defined internal granular layer or layer IV, and layer II is accentuated, having an admixture of small and slightly larger neurons with darker staining characteristics in cell stains. Layer III resembles that of the isocortex but is often thin in comparison with a notable lack of density. Layers V and VI contain very large neurons and are not well differentiated. These are classic cytoarchitectural features of what Sanides termed the proisocortex, the cortex that forms the bulk of the limbic lobe. Accordingly, this cortex characterizes dorsal parts of the limbic lobe such as the cingulate gyrus and ventral parts such as the perirhinal cortex. While it is tempting to view these as 'transitional' cortical areas interposed between the older allocortex and newer isocortex, this is in all likelihood too great a generality. They are too large in size to be viewed in this fashion, they have their own unique cell types, layers and neurotransmitters, and such a view precludes assigning unique functional attributes to the proisocortices – for which there are many. The term meso or intermediate cortex seems to contain less prejudice.

Little is known about the subcortical axonal connections of the perirhinal cortex, but a substantial amount of information exists regarding its cortical inputs and outputs. For example, the perirhinal cortex is connected to many other parts of the limbic lobe including the cingulate, insular, temporal polar, posterior parahippocampal and orbital frontal cortices. In the monkey, strong neocortical projections course to the perirhinal cortex as well, but these arise primarily from the visual association cortex on the lateral surface of the temporal lobe and the visual association cortex that occupies the junction of the temporal and occipital lobes both laterally and ventrally. These along with posterior parahippocampal projections link the perirhinal cortex powerfully with the visual cortices at least in primates. However, it is equally important to appreciate that some of the stronger connections of the perirhinal cortex are made with the entorhinal cortex that lies medial to it. Since the latter has strong reciprocal connections with the hippocampal formation, it follows that the perirhinal cortex is a critical link in reciprocal cortical neural systems that couple together the visual association and limbic cortices with the hippocampal formation. Although the details are not as well known, it would not be surprising if other cortical areas functionally linked to additional modalities access and receive projections from the hippocampal formation via the perirhinal cortex.

Recent investigations in humans and nonhuman primates add substantially to a growing neuroscience interest in the perirhinal cortex. For example, in the human brain acetylcholinesterase staining for axons partitions the perirhinal cortex into distinct columns spanning the superficial layers (Figure 1(b)). These appear to be axonal plombs with a regular periodicity. Immunostaining for the calcium-binding protein parvalbumin reveals a similar columnarity or modularity (Figure 1(c)). This acquires greater interest in view of the fact that neurofibrillary tangles in Alzheimer's disease begin first in the perirhinal cortex and Alz-50 immunostaining for the abnormally phosphorylated tau protein A68 and fluorochrome staining with Thioflavine-S for neurofibrillary tangles reveals sharp columns of affected neurons with pathology-free interspaces separating the columns (Figures 2(a) and (b)).

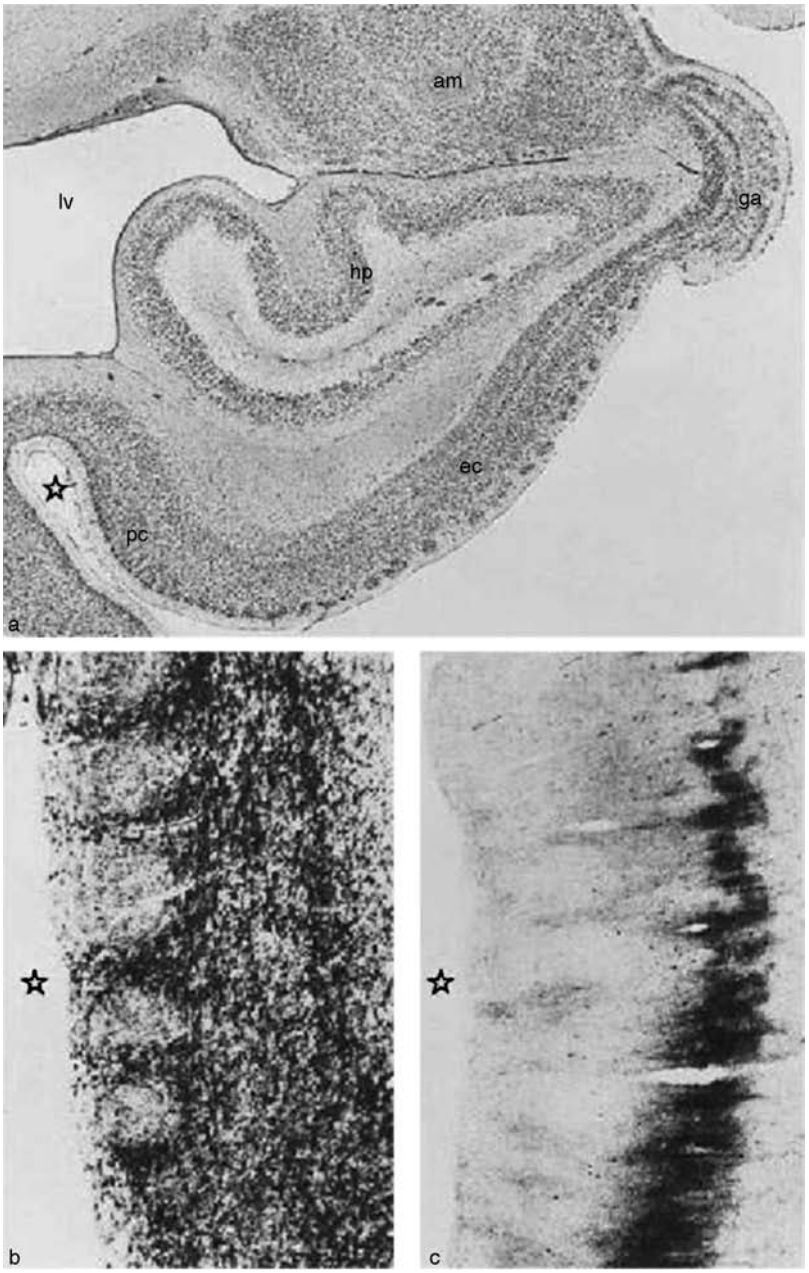

Figure 1 (a) Cross-section through the ventral medial part of the temporal lobe in the human brain stained with the Gallyas cell stain showing the location of the perirhinal cortex (pc) along the medial bank of the collateral sulcus (star). The parahippocampal gyrus at this level is dominated by the entorhinal cortex (ec) which also covers the gyrus ambiens (ga). am, amygdala; hp, hippocampus; lv, inferior horn of the lateral ventricle. (b) Distribution of acetylcholinesterase in the perirhinal cortex of the human brain on the medial bank of the collateral sulcus (star). Note its modular appearance in the superficial cortical layers and the distinct columns of staining. (c) Immunostaining for parvalbumin on the medial bank of the collateral sulcus (star) in the human perirhinal cortex. Note the modular bands of immunoreactivity in both the superficial layers (more lightly stained) and deeper layers (more darkly stained).

Thus, there is a growing body of evidence that the perirhinal cortex is distinctly modular in the human brain and that this organization is partially disassembled in Alzheimer's disease.

Last, it is noteworthy that ablations in monkeys confined to the cortex in the depths and lateral bank of the rhinal sulcus produce object recognition memory impairments as great or greater than hippocampal ablations and exacerbate the impairment when combined with entorhinal or hippocampal lesions. These findings, now observed by several groups, suggests that the perirhinal cortex not only plays an important role in memory, but may, in fact, have a unique or even pre-eminent role relative to other ventromedial temporal lobe structures.

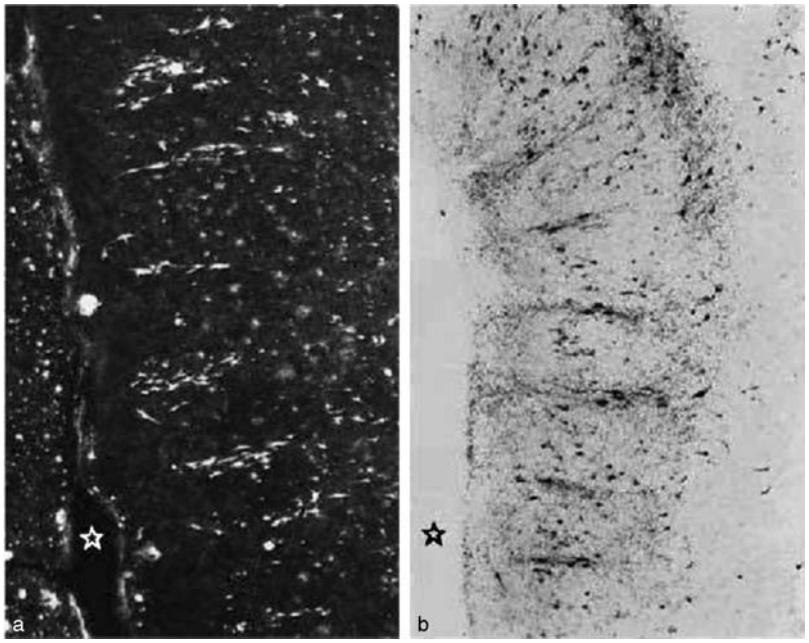

Figure 2 (a) Thioflavine-S fluorochrome staining for neurofibrillary tangles in Alzheimer's disease in the perirhinal cortex along the medial bank of the collateral sulcus (star). Note the distinct columns of diseased neurons and the pathology-free interspaces between them. (b) Alz-50 immunostaining for the abnormally phosphorylated tau or A68 protein in Alzheimer's disease in the perirhinal cortex along the medial bank of the collateral sulcus (star). Note the columns of immunopositive neurons that characterize this cortex.

In summary, the perirhinal cortex is a pivotal cortical area in the medial temporal lobe interconnecting the hippocampal formation together with other parts of the limbic lobe and with the lateral temporal and occipitotemporal association cortices. Its cytoarchitecture and cortical input and output have been known for some time in several species, but recent behavior and pathological data suggest that it plays a role in visual memory. These observations are of great interest since the perirhinal cortex is affected early in Alzheimer's disease and appears to be the first cortical area that contains neurofibrillary tangles in this disorder. Finally, in the human brain a distinct modular organization exists suggesting that this cortex is specialized functionally and is not a simple transition area separating the entorhinal and inferior temporal cortices.

See also: Animal Models of Alzheimer's Disease; Brodmann's Areas; Cerebral Cortex; Entorhinal Cortex; Memory Consolidation: Cerebral Cortex; Recognition Memory.

Further Reading

Hyman BT, Van Hoesen GW, Damasio AR, and Barnes CL (1984) Alzheimer's disease: Cell-specific pathology isolates the hippocampal formation. *Science* 225: 1168–1170.

Insausti R, Amaral DG, and Cowan WM (1987) The entorhinal cortex of the monkey: II. Cortical afferents. *Journal of Comparative Neurology* 264: 356–395.

Meguro K, Blaizot X, Kondo Y, and LeMestric C (1999) Neocortical and hippocampal glucose hypometabolism following neurotoxic lesions of the entorhinal and perirhinal cortices in the nonhuman primate as shown by PET: Implications for Alzheimer's disease. *Brain* 122: 1519–1531.

Meunier M, Bachevalier J, Mishkin M, and Murray EA (1993) Effects on visual recognition of combined and separate ablations of the entorhinal and perirhinal cortex in rhesus monkeys. *Journal of Neuroscience* 13: 5418–5432.

Suzuki WA and Amaral DG (1994) Perirhinal and parahippocampal cortices of the macaque monkey: Cortical afferents. *Journal of Comparative Neurology* 350: 497–533.

Suzuki WA and Amaral DG (1994) Topographic organization of the reciprocal connections between the monkey entorhinal cortex and the perirhinal and parahippocampal cortices. *Journal of Neuroscience* 14: 1856–1877.

Suzuki WA, Zola-Morgan S, Squire LR, and Amaral DG (1993) Lesions of the perirhinal and parahippocampal cortices in the monkey produce long-lasting memory impairment in the visual and tactual modalities. *Journal of Neuroscience* 13: 2430–2451.

Van Hoesen GW (1982) The parahippocampal gyrus: New observations regarding its cortical connections in the monkey. *Trends in Neuroscience* 5: 345–350.

Van Hoesen GW and Pandya DN (1975) Some connections of the entorhinal (area 28) and perirhinal (area 35) cortices of the rhesus monkey. I. Temporal lobe afferents. *Brain Research* 95: 1–24.

Van Hoesen GW and Solodkin A (1993) Some modular features of temporal cortex in humans as revealed by pathological changes in Alzheimer's disease. *Cerebral Cortex* 3: 465–475.

Zola-Morgan S, Squire LR, Clower RP, and Rempel NL (1993) Damage to the perirhinal cortex exacerbates memory impairment following lesions to the hippocampal formation. *Journal of Neuroscience* 13: 251–265.

Perirhinal Cortex: Neural Representations

M W Brown and M Eldridge, University of Bristol, Bristol, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

This article primarily concerns what is known of the neural representations in the perirhinal cortex that relate to the identification of stimuli and their associations, and to judgments about prior occurrence. After some background anatomical information, a summary is given of the results of ablation experiments that indicate the functions of the perirhinal cortex. These functions include roles in perception, particularly for object-related information, and in memory, notably but not exclusively recognition memory and paired-associate learning. A brief overview of recent attempts to find synaptic plastic substrates for the memory functions of the perirhinal cortex is also included. The article does not include the results of human brain imaging, except when mentioned in passing when of particular pertinence.

Perirhinal Cortex

In the last 20 years, the perirhinal cortex has received increasing scientific attention because of its important roles in perception and learning. The perirhinal cortex is located in the anterior and medial portion of the inferior aspect of the medial temporal lobe of primates (Figure 1). Although in the past there has been considerable variation in the definition of the boundaries of the perirhinal cortex, there is now broad agreement that in the monkey it extends immediately lateral to the full extent of the rhinal sulcus and includes the cortex corresponding to both areas 35 and 36 of Brodmann. In the rat, the perirhinal cortex (areas 35 and 36) is located on either side of the caudal part of the rhinal sulcus. The precise extent of the corresponding region in the human brain remains uncertain.

Overview of Connections

The perirhinal cortex is highly interconnected with many other brain regions. It receives highly processed sensory information of all modalities and provides an important route for transferring information to and from the hippocampus. Its position at the top of the sensory processing hierarchy and at the gateway to the limbic system means that it is ideally placed to play important roles in sensory perception and memory. Once a stimulus may be identified, the perirhinal cortex processes aspects of its past history and associations.

In the monkey, the most prominent input arises from the inferior temporal cortex (unimodal visual areas TE and TEO), although it also receives input from the auditory, olfactory, and somatosensory association areas of cortex (Figure 1(a)). Further substantial inputs arise from polymodal association areas, including the prefrontal and cingulate cortices, entorhinal cortex (and, thereby, indirectly from the hippocampus), parahippocampal cortex (areas TH and TF), and cortex of the dorsal bank of the superior temporal sulcus. Perirhinal outputs largely mirror their inputs with strong connections to the entorhinal cortex (again, therefore indirectly connecting to the hippocampus), parahippocampal gyrus, prefrontal cortex, and unimodal and polymodal sensory association areas. The perirhinal cortex also has reciprocal connections with the amygdala and thalamus (mediodorsal and midline nuclei) and sends outputs to the tail of the caudate nucleus and ventral putamen. In addition, it receives input from brain stem monaminergic cell groups and cholinergic input from the basal forebrain. A parallel, similarly widespread pattern of connectivity is found in the rat (Figure 1(b)).

The perirhinal cortex is juxtallocortex, cortex of a type that is transitional between the neocortex and the archicortex and paleocortex of the hippocampal and entorhinal regions. Unlike the neocortex, area 35 lacks an inner granular layer (layer 4) and this layer is very narrow in area 36. Again, in contrast to the neocortex, the perirhinal cortex does not have an obviously columnar organization.

Overview of Functions

There is a very large body of evidence indicating that the perirhinal cortex plays a central role in certain aspects of recognition memory. The evidence is clearest for familiarity discrimination for single items. There is also much evidence of a role in perceptual functions, although this role is not universally accepted. In addition, the perirhinal cortex is involved in other types of memory linked to learned associations made with items (notably reinforcement learning and paired-associate learning). Details of the neuronal responses related to such functions are given later.

Overall, ablation studies provide strong evidence for a role for the perirhinal cortex in object identification, although some findings have challenged this conclusion. Monkeys with perirhinal lesions have been shown to be impaired in a series of tasks designed to test object identification. Thus, there is

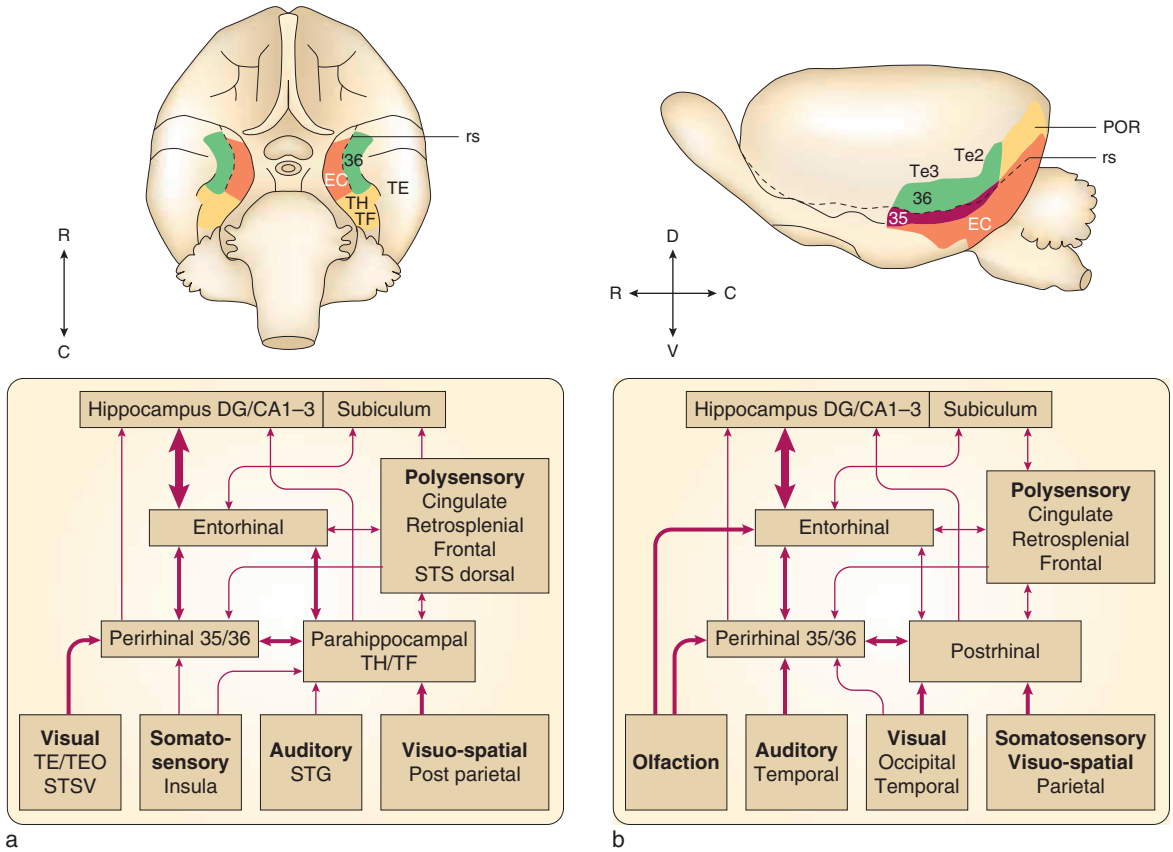

Figure 1 Position and some major connections of the perirhinal and neighboring cortices: (a) in the monkey (ventral view); (b) in the rat (lateral view). Below each brain image, routes are shown by which sensory information may reach the perirhinal cortex and the connections from the perirhinal cortex to the hippocampus. The arrow thickness indicates the relative size of the projections. 35 and 36, perirhinal cortex areas; C, caudal; D, dorsal; DG, dentate gyrus; EC, entorhinal cortex; POR, postrhinal cortex; R, rostal; rs, rhinal sulcus; STG, superior temporal gyrus; STS, superior temporal sulcus; SUBIC, subicular complex; TE, TEO, Te2, Te3, visual association cortex areas; TH and TF, parahippocampal cortex areas; STSV, superior temporal sulcus ventral. From Witter MP, Groenewegen HJ, Lopes da Silva FH, and Lohman AHM (1989) Functional organization of the extrinsic and intrinsic circuitry of the parahippocampal region. *Progress in Neurobiology* 33: 161–253.

impairment in perceptual oddity tasks, in which the odd one of a set of images must be chosen; tasks requiring object identification from different views or partial views of the same object; and tasks in which an object previously experienced only by touch must be selected using vision. In addition, monkeys with perirhinal lesions are particularly impaired on tasks requiring discrimination between visual stimuli with a high level of feature ambiguity. It has been argued that, when perceptual deficits have not been found, the tasks may have been solved using discrimination based on individual features rather than whole objects as entities in themselves. When test has been made, the evidence indicates that it is possible to dissociate the mnemonic and perceptual functions of the perirhinal cortex.

The ability to recognize the novelty or familiarity of individual items is an important aspect of recognition memory. Such recognition memory relies not only on identification (perception) but also on judgment of prior occurrence (memory). The involvement of the perirhinal cortex in the judgment of prior occurrence for individual items is widely agreed on. In one view it forms part of a unified medial temporal lobe memory system; an alternative view is that its role is differentiable from that of the hippocampus (the hippocampus is concerned with recollective, complex associational, and spatial aspects of recognition memory, whereas the perirhinal cortex is concerned with single-item familiarity discrimination and simple associations). The pattern of impairments after perirhinal cortex lesions has been doubly dissociated from neighboring areas such as the middle temporal gyrus, which corresponds to the dorsal part of visual area TE; hippocampus; and amygdala. Thus, whereas lesions of the middle temporal gyrus impair color discrimination but spare recognition memory, lesions of the perirhinal cortex impair object

recognition memory but not color discrimination. Further, in both monkeys and rats, hippocampal or fornix lesions lead to marked impairment in spatial tasks with relatively mild or no effect on recognition memory, whereas perirhinal cortex lesions result in the opposite pattern of deficits. Amygdalar lesions impair food preference learning but not delayed matching to sample with infrequently repeated stimuli, whereas perirhinal lesions result in the opposite pattern of impairment. These double dissociations demonstrate that the perirhinal cortex is necessary for visual recognition memory even when functionally isolated from structures to which it is connected both in the ventral visual processing stream and in the medial temporal lobe. In addition to these results for visual tasks, perirhinal lesions impair tactile recognition memory in monkeys and olfactory recognition memory in rats, although auditory recognition memory seems to be spared.

Ablation studies also indicate that the perirhinal cortex is necessary for making certain (although not all) associations involving objects. Such work has shown that perirhinal lesions impair tasks involving associations of objects with other objects, odors, tastes, and tactile or auditory information. Such associations may also be abstract. Monkeys with perirhinal lesions are impaired in a task in which they must respond to a sequence of discriminations to obtain a reward; the results suggest a role of the perirhinal cortex in forming associations between objects and position in a series or proximity to reward in a sequence. In addition, perirhinal lesions in rats produce impairments in conditioned avoidance tasks, notably contextual fear conditioning; the impairment relates to processing of the conditioned stimulus rather than the spatial context of the task.

Neural Responses Related to Stimulus Identification and Linked Associational Functions

Stimulus Identification

Visual information processing is largely divided between two pathways: a dorsal visual stream primarily associated with spatial perception and sensorimotor output, and a ventral visual stream primarily associated with object identification. The latter pathway runs in the posterior–anterior direction through successive processing areas from the primary visual cortex, via area V4, to the inferior temporal cortex, including area TE, which supplies strong projections to the perirhinal cortex. Electrophysiological evidence indicates that neurons in the anterior regions of the ventral visual stream code more complex visual representations than do neurons in the more caudal regions. The perirhinal cortex can be considered to be at the top of the hierarchy of the processing regions, and it has been suggested that it is necessary for the perception of objects as entities as opposed to the identification of their individual features.

Most studies of neuronal response properties concerning stimulus identification in the anterior temporal lobe have concerned visual stimuli and the inferior temporal cortex. Although it is clear that some such studies have included perirhinal neurons, the sensory response properties of perirhinal neurons have rarely been distinguished from those of other anterior inferior temporal neurons. (Past uncertainties about the boundaries of the perirhinal cortex have not assisted in making such distinctions.) Indeed, only a few distinctions have been described between the responses of neurons in anterior TE and in the perirhinal cortex, and these chiefly relate to associative or mnemonic properties (see the sections titled 'Correlates of long-term memory and learned associations of frequently repeated stimuli,' 'Neuronal responses related to paired-associate learning,' and 'Neural responses related to judgment of prior occurrence').

In conscious monkeys, typically more than half of recorded perirhinal neurons are reported as being visually responsive. When reported, more than half of these visually responsive neurons respond to the majority of complex stimuli tested, although by no means with equal vigor to every stimulus. Nevertheless, some perirhinal neurons respond only to restricted classes of stimuli, such as faces or stimuli of a particular color. Accordingly, the responses of perirhinal neurons signal information about the physical features of visual stimuli. It has been reported that a higher proportion of neurons are stimulus selective in the perirhinal cortex than in area TE. Unfortunately, in spite of the suggestion that the perirhinal cortex is critical to object identification, no recordings examining this issue have included the perirhinal cortex. As in TE, when single stimuli are viewed, perirhinal receptive fields are large and may encompass a region extending into both visual hemifields. There has been little study of perirhinal responses in relation to stimulus constancy, but again evidence indicates that it is similar to that found in TE, with similar responses being found when stimuli are shown across different positions or at different sizes. Although some tendency for clustering of similar perirhinal responses has been reported, there is no strong anatomical or physiological evidence that the perirhinal cortex has a columnar organization.

Anatomical connectivity indicates that the perirhinal cortex is a polymodal area, yet there is still

no published major study of polymodal perirhinal responses (although commonly perirhinal neuronal activity in both monkeys and rats is found to vary with all aspects of the behavioral task during which recordings are made and not merely to the presentation of sensory stimuli). Studies of neuronal responses in rats indicate that the responses of perirhinal neurons encode stimulus-specific visual and olfactory information but do not encode detailed spatial information.

Correlates of Attention and Short-Term Memory

Two types of neuronal activity that occur in many brain regions in relation to attentional and short-term memory tasks are also found in the perirhinal cortex: response suppression and delay activity. When monkeys are trained to select a target stimulus presented among nontarget stimuli, neural responses to the behaviorally irrelevant stimuli are suppressed, whereas the response to the behaviorally relevant target stimulus remains unchanged (see Figure 2). This change in responsiveness reflects a neural mechanism encoding the learned salience of a given stimulus. Delay activity is found during the interval between stimulus presentations in short-term or working memory tasks such as delayed matching to sample. In such tasks, a stimulus is presented in an acquisition phase and, after a possibly variable but relatively brief (<1 min) delay interval, there follows a choice phase in which a decision must be made on the basis of the stimulus presented in the acquisition phase. For many perirhinal neurons, the activity in the delay period carries information concerning the stimulus presented during the acquisition phase (i.e., it is stimulus specific). Such activity maintains a representation of that stimulus online and may be a correlate of the stimulus being held in mind. In certain cases, the response of the neuron to the stimulus may then be enhanced when it is repeated on match trials. Again, the activity change shows that perirhinal activity is modulated by attentive and mnemonic factors.

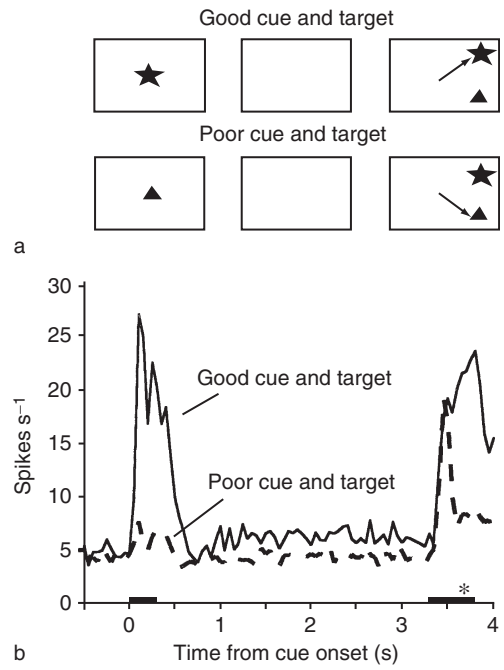

Figure 2 Differential delay activity and response suppression: (a) experiment design; (b) graph of population response for 22 monkey inferior temporal (including perirhinal) neurons. For each neuron, one stimulus that evoked a strong response and one that evoked a weak response were chosen. As shown in (a), either stimulus was presented as the cue (first bar on time scale in (b)), and after a delay, the cue stimulus had to be selected (by making a saccadic eye movement to it; second bar in (b)) from a display of both stimuli in the choice phase. Thus, whichever stimulus was the cue, the display during the choice phase was of both stimuli. The activity was averaged for each neuron for trials using the good and poor stimuli and then across neurons to produce a population response. Note that the delay activity between the end of the cue stimulus presentation and the appearance of the choice stimuli is greater for the good than for the poor stimuli. Also note that the response to the display of both stimuli during the choice phase is rapidly suppressed when the poor stimulus is the target. Thus, across the population of neurons only those that respond strongly to the target are strongly active during the delay and the choice phases. The responses of other neurons are rapidly (~200 ms) suppressed in the choice phase. * indicates the start of saccadic eye movement (note that this is later than start of suppression). From Desimone R (1996) Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences of the United States of America* 93: 13494–13499.

Correlates of Long-Term Memory and Learned Associations of Frequently Repeated Stimuli

A number of different types of long-term change have been reported in the responses of perirhinal neurons trained on different tasks in which stimuli are seen many times and associated with reward. The findings emphasize the plasticity of the perirhinal cortex in relation to different processing demands. This plasticity and adaptation to specific learning situations may explain what are, at first, apparent inconsistencies in some reported findings. Let us review information available about perirhinal neuronal responses related to examples of such long-term changes and learned associations that are concerned with frequently presented stimuli.

First, the response change that has been reported most often is found in delayed matching tasks using small stimulus sets in which the stimuli are frequently repeated. In such experiments, the response to a stimulus changes (becoming variously either stronger or weaker for different neurons) according to whether it occurs as the choice stimulus on a match or mismatch trial. Such response changes establish that perirhinal neuronal responses can be determined

by more than the physical characteristics of the presented stimulus; the responses can also reflect the behavioral context in which the stimulus is presented. In one description, such responses therefore contain information about the learned associations of the stimulus and, hence, may be said to convey information concerning the behavioral meaning of the stimulus. In another description, the responses contain information about the recent occurrence or nonoccurrence of the stimulus and, hence, are related to a temporal rather than any more abstract context and therefore to judgment of prior occurrence.

Second, sequence learning provides stronger evidence that perirhinal neurons can respond on the basis of the behavioral meaning of a stimulus (its abstract associations) rather than solely according to the physical characteristics of the stimulus. Consider findings from experiments in which a monkey has been trained in a task in which different length sequences of stimulus presentations occur before the availability of reward. The animal learns that certain stimuli signal the position in the sequence (e.g., first or last). Certain perirhinal neurons respond according to the position in the sequence leading to reward that is signaled by a stimulus rather than according to the physical attributes of the particular stimulus. This type of response to a stimulus' implication rather than its appearance was not observed for neurons in TE. Recent results indicate that the development of such perirhinal responses can be blocked by local infusion of an oligonucleotide that interrupts dopaminergic transmission.

Third, when the same stimuli are used repeatedly for tens or hundreds of trials and the animal must attend to these stimuli to gain reward, the responses to these highly familiar stimuli show a gradual increase in magnitude. This finding (along with others) indicates that perirhinal neuronal representations of stimuli are plastic even in the adult. They adapt to experience, presumably to optimize the processing of behaviorally important information.

Neuronal Responses Related to Paired-Associate Learning

Correlates of the Sequential Pairing of Stimuli

Both recording and lesion experiments in monkeys have established the importance of the perirhinal cortex for visual paired-associate learning. In this task, the animal learns across many trials to associate arbitrarily assigned pairs of complex visual stimuli, so that, after a delay following an initial (cue) stimulus presentation, the animal selects (e.g., by touching) the other stimulus of the pair (its paired associate). Two neural substrates of this form of associative learning have been identified in perirhinal cortex: (1) stimulus-specific delay activity and (2) the occurrence of strong responses to both of the stimuli that have been paired (such paired responsiveness occurs more frequently than expected by chance; see Figure 3). Thus, the learning results in single neurons that can code for both of a pair of stimuli; that is, information from both stimuli has converged on individual perirhinal neurons. Such pairs have not been found in TE; there, neurons are responsive only to individual stimuli.

Although sensory information is fed forward from TE to the perirhinal cortex, memory retrieval is fed

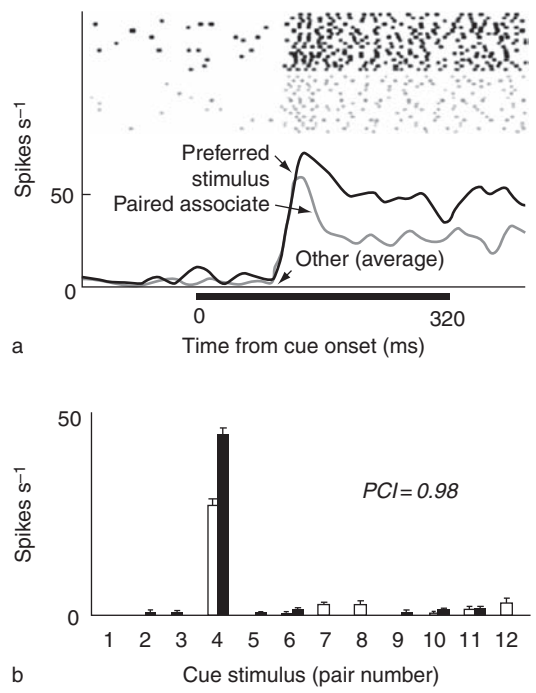

Figure 3 Example stimulus-selective neuronal response for learned paired-associate stimuli: (a) Raster display and peristimulus time histogram for the trials; (b) mean (+ SEM) discharge rates during the cue period (60–320 ms from the cue onset) for each of 12 learned pairs of stimuli. The example monkey perirhinal neuron was strongly activated by both members of a particular pair of associated stimuli. In contrast, its responses were negligible when presented with members of any of the 11 other stimulus pairs. In (a) the response to the preferred stimulus as the cue stimulus is indicated by the thick black line and the response to its paired associate is indicated by the gray line. The thin black line denotes the averaged responses in the trials in which other stimuli were used as the cue. The horizontal gray bar indicates the cue presentation period. Note the continuing (delay) activity for the preferred stimuli after the cue stimulus is turned off. Note in (b) that the responses to both members of pair 4 are much greater than the response to any member of the other 11 paired associates. *PCI*, pair-coding index. Reproduced from Naya Y, Yoshida M, and Miyashita Y (2003) Forward processing of long-term associative memory in monkey inferotemporal cortex. *Journal of Neuroscience* 23: 2861–2871, copyright © 2003, with permission from the Society for Neuroscience.

back from the perirhinal cortex to TE. Figure 4 illustrates the firing rate of a perirhinal neuron and that of a neuron from area TE in response to the stimulus to which they both respond preferentially. Memory-retrieval signals appear first in the perirhinal cortex, after which TE neurons are gradually recruited to represent the sought target. It has been established that the perirhinal changes involve and are dependent on the immediate early gene *Zif268*. Anatomical tracing indicates that paired-associate learning is associated with changes in the pattern of axonal connectivity between TE and the perirhinal cortex. Such alterations in the functional architecture of the perirhinal cortex may facilitate the reactivation of newly formed object representations and, moreover, may form a neural basis for the process of memory retrieval.

Correlates of the Simultaneous Pairing of Stimuli

In contrast to findings for paired stimuli presented sequentially, when pairs of stimuli are shown simultaneously, inferior temporal (including perirhinal) neuronal responses to the stimuli shown individually appear to be unchanged even after many such pairings. However, responses to the stimuli shown as a pair are slowly changed by learning. In such experiments, the animal learns to respond (saccade) to whichever one of the two stimuli has been previously shown as a cue. After training, when both stimuli are displayed simultaneously the neuronal response is enhanced if the animal must saccade to the stimulus to which the neuron responds better, but it is reduced when the neuron must respond to the stimulus to which it responds more weakly. Thus, the cortical representation of the pair appears to have become more differentiated. This differential response develops across trials in which the two stimuli are paired when the pairing is repeated many times. The alteration of the representation of the pair of stimuli in the cortex is presumably a result of the altered effectiveness of interconnections within the cortex. Moreover, measured across a range of stimuli, two neighboring perirhinal neurons tend to respond more similarly if the stimuli are familiar (presented many times on the previous day) than if they are relatively novel (not seen before the day of recording). The finding suggests that as a result of repeated experience perirhinal neurons that respond similarly to particular stimuli tend to become grouped.

Neural Responses Related to Judgment of Prior Occurrence

In addition to the changes detailed in previous sections, the responses of certain neurons in the perirhinal and neighboring cortices signal information concerning the previous occurrence of infrequently encountered stimuli. Most strikingly, responses of such neurons are strong for stimuli that have not been seen previously and weak once these stimuli appear again. We term such responses repetition sensitive. They have characteristics necessary for making judgments of prior occurrence, and they have been related to the recency and familiarity discrimination aspects of recognition memory. These response changes have also been called stimulus-specific adaptation and response suppression; however, these phrases carry implications concerning the underlying mechanisms or functionality of the changes that cannot

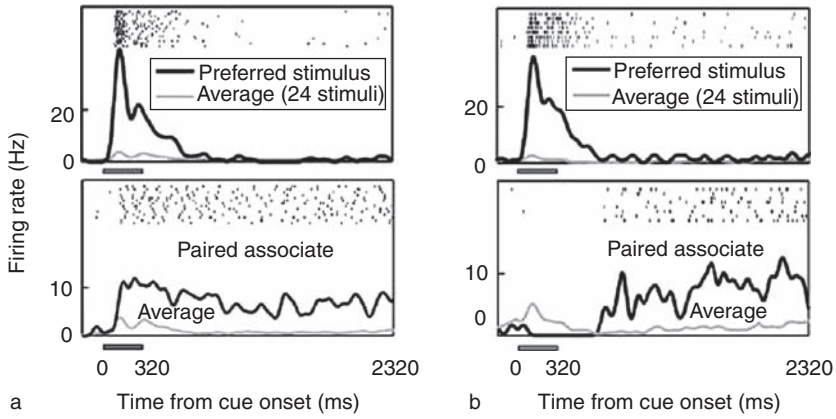

Figure 4 Neuronal activity during paired associate memory retrieval in monkey for a neuron: (a) in the perirhinal cortex (A36); (b) in area TE. Shown for each neuron are raster displays and peristimulus time histograms for trials in which the preferred stimulus (upper graphs) or its paired associate (lower graphs) were presented as the cue stimulus. The black lines indicate responses to the preferred cue stimulus (upper graphs) or its paired associate (lower graphs); the gray lines indicate the mean responses to all 24 stimuli. Note that for the paired associate of the preferred stimulus, the delay activity in the perirhinal cortex is high from the time that stimulus presentation ends, whereas there is a delay before such activity appears in TE. The delay activity in TE is dependent on a signal fed back from the perirhinal cortex. From Naya Y, Yoshida M, and Miyashita Y (2001) Backward spreading of memory retrieval signal in primate temporal cortex. *Science* 291: 661–664.

currently be justified. In particular, to date no evidence has been found that the response reductions (decrements) are dependent on an active suppressive or inhibitory mechanism. Moreover, the term response suppression is used in relation to attentional processes; the use of the same term for a different process involved in familiarity discrimination is inappropriate.

Most information about repetition-sensitive responses comes from monkeys and visual stimulation. Next, we present the characteristics of these neuronal responses, with generalizations across modalities and species (discussed later).

Response Characteristics

Up to 25% of perirhinal neurons have a reduced response to familiar (compared to novel) visual stimuli, representing approximately 60% of the visually responsive neurons (Figure 5). Note that such change divides perirhinal neurons into two classes: those that have repetition-sensitive responses signaling information concerning even a single prior occurrence of a stimulus and the remainder that do not. The overall tendency to response reduction on repetition is sufficiently large that it is readily seen in population

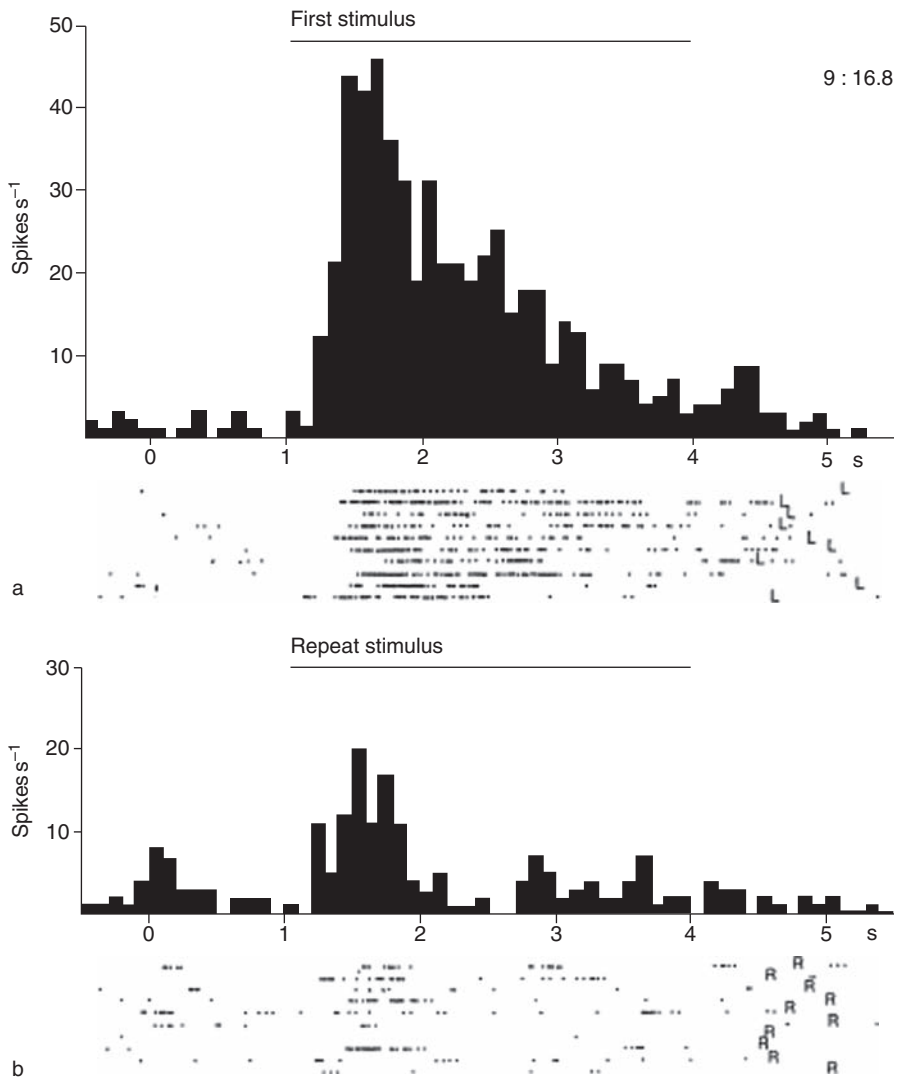

Figure 5 Example of repetition-sensitive monkey perirhinal neuronal response: (a) first appearance of ten initially novel pictures; (b) repeat appearance of the ten pictures. Responses are stronger to first than for the repeat appearances. One stimulus was shown (bars above the histograms) on each trial of a serial recognition memory task. To be rewarded, the monkey had to press left (L) for the first appearance and right (R) for a repeat appearance of the stimulus. The first and repeat trials have been grouped in this display, but were intermixed when presented. Controls (not shown) established that the difference in activity was not due to the different behavioral responses. Note also that the neuron does not respond equally strongly to the first presentations of the ten different pictures; that is, its activity carries information about stimulus features as well as prior occurrence. From Brown MW (1990) Why does the cortex have a hippocampus? In: Gabriel M and Moore J (eds.) *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pp. 233–282. Cambridge, MA: MIT Press.

measures of neuronal activity from this cortex. Except under specific circumstances (see next section), response increases with stimulus repetition occur at a level below that expected by chance. Response reductions with repetition have been described for three-dimensional objects and two-dimensional pictures of objects, faces, patterns, and scenes. Repetition-sensitive responses also encode information concerning the physical features of the stimuli, just as do responses that are not repetition sensitive. Thus, for example, some neurons respond only to novel faces and not novel objects or only to novel stimuli of a certain color. Similar response changes are found in the perirhinal cortex (areas 35 and 36), adjacent visual association cortex (anterior TE), and neighboring lateral entorhinal cortex.

In experiments in which both novel and highly familiar stimuli have been repeated, it has been possible to show that there is more than one pattern of response change on stimulus repetition. Under the conditions of the original experiments, these response types were termed novelty, recency, and familiarity because of the type of information signaled. Novelty responses were strong to first presentations of novel stimuli and weak when these stimuli were repeated or already familiar stimuli were shown; the responses gave out a novelty signal. Recency responses were strong for novel stimuli and for highly familiar stimuli that had not been seen for some time (e.g., 24 h) but weak when either type of stimulus had been seen recently; accordingly, such response changes contained information about the recency of presentation of stimuli. In contrast, familiarity responses were strong for novel stimuli even when they were seen for a second time after a brief interval (e.g., <5 min) but weak for highly familiar stimuli even if they had not been seen for a long time (e.g., 24 h); accordingly, such response changes contained information about the relative familiarity of stimuli. Significantly, these different patterns of response change indicated that information about recency of occurrence could be encoded separably from that about relative familiarity. Collectively, perirhinal neurons signaled information about both familiarity and recency. Further study of the properties of the response changes indicated that the term 'familiarity response' was potentially misleading. A better description is that they are slow-change responses because such responses are reduced when an initially novel stimulus is shown for a second time if the interval between its presentations is sufficiently long (e.g., >10 min). They are not reduced if the interval is brief. In contrast, the novelty and recency responses are reduced even when an initially novel stimulus is repeated after an interval of <1 s; correspondingly, such responses are fast change (Figure 6). This new designation does not alter the fact that more than one type of information is being signaled and that there must be more than one type of underlying synaptic change (see the section titled 'Synaptic plasticity and modeling of perirhinal neuronal response changes').

Evidence of Relationship to Recognition Memory

The findings of a large number of ablation studies in both rats and monkeys are in agreement that lesions involving the perirhinal cortex cause major impairment of recognition memory tasks that can be solved using familiarity discrimination for single items. The perirhinal cortex is thus necessary for the solution of such tasks. Repetition-sensitive perirhinal responses provide sufficient information to solve such tasks. First, they signal information concerning stimulus novelty or familiarity and recency of occurrence. Second, the responses are changed by a single occurrence of a new stimulus; they therefore evidence single-exposure learning, as does animal and human recognition memory. Third, the system appears to have a very large capacity; even after a monkey has been shown many hundreds of stimuli, the neurons still respond on the basis of whether an individual stimulus has been seen before. The responses are in this way highly stimulus selective. This property also has an important corollary, namely, that the neurons do not respond weakly to a repeated stimulus because of response fatigue; they continue to respond strongly to a subsequent new stimulus. Fourth, the response changes show access to information held in long-term memory. In particular, the response to the second appearance of a stimulus is reduced even when many other stimuli have been seen and a long time has elapsed since the stimulus was first seen. For many such perirhinal neurons ($\sim$70%, compared to $\sim$45% in anterior TE), responses are reduced even when a stimulus has not been seen for 24 h (Figure 6). Neuronal memory spans (the length of time during which reduced responses are still found) of a few days have also been reported, but longer delays have not been tested. These properties establish that the underlying mechanism differs from that of simple habituation produced by a monotonously repeated stimulus. Moreover, the change is not a simple analog of adaptation to prolonged sensory stimulation. Although the fast-change responses might be related to priming mechanisms, evidence from human lesion and psychological studies makes this unlikely. Slow-change responses do not have the characteristics necessary to explain priming. It should be noted that, for many neurons, response reductions fade with the passage of time and that this provides information within the network concerning how recently a stimulus was presented. In addition, response

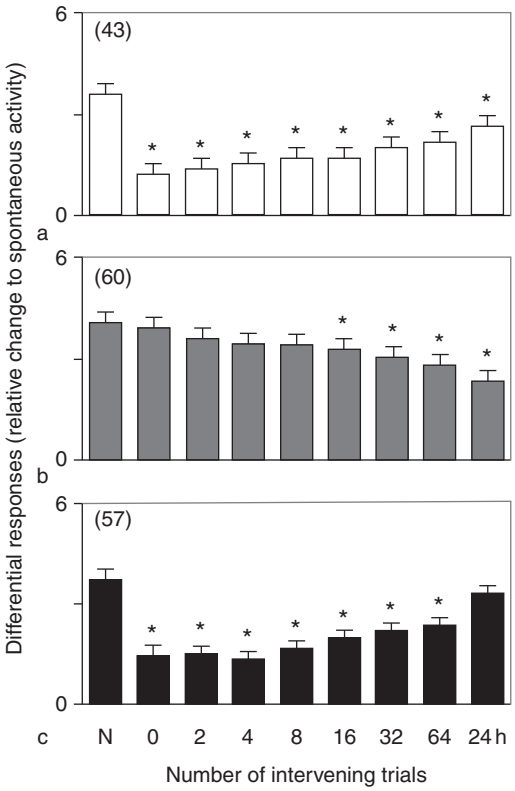

Figure 6 Mean population memory spans for neurons in the monkey anterior inferior temporal cortex: (a) recency neurons; (b) familiarity neurons; (c) novelty neurons. Mean population responses (averaged within neurons and then across neurons) are shown for first presentations (N) and second presentations after different intervals (numbers of intervening trials in a serial recognition memory task or 24 h). For these intervals, memory across the population increased with time elapsed for familiarity neurons but decreased for novelty and recency neurons. The decrement developed rapidly (<10 s) for recency and novelty neurons but took more than 5 min to become significant for the population of familiarity neurons. * indicates $P < 0.05$ difference in mean response to second presentations compared to first presentations. From Brown MW (2000) Neuronal correlates of recognition memory. In: Bolhuis JJ (ed.) *Brain, Perception and Memory*, pp.185–208. Oxford: Oxford University Press.

reductions with stimulus repetition are found both when stimuli are presented within a recognition memory task (so that the animal must use its relative familiarity to gain reward) and when they are presented without the animal being trained in familiarity discrimination. The response changes are therefore automatic and endogenous; they are not solely induced by training.

Three other types of perirhinal neuronal response change may also be observed during delayed matching to sample tasks in which judgments about the prior occurrence of a target stimulus have to be made in a choice phase following a delay. They signal information of importance to the solution of particular types of delayed matching to sample tasks, but do not have the properties necessary to explain the behavioral features of general familiarity discrimination for infrequently repeated individual visual stimuli. First, when only one stimulus needs be held in mind at a time and a decision related to its prior occurrence made within a short period (seconds rather than many minutes), increased firing (delay activity) is found between the initial stimulus presentation and the subsequent presentation of the choice (match or nonmatch) stimulus. Such delay activity is not observed when more than one stimulus must be held in mind and the interval before the decision is long, so delay activity cannot provide the substrate for general familiarity discrimination. Second, if a monkey is trained to distinguish between a repeated, unrewarded and a repeated rewarded stimulus, some perirhinal neurons show an increase in response to the repeated rewarded stimulus; however, such increases have been found only when there is a single repeated target stimulus and after training on the specific task. Thus, again, the mechanism appears to be specific to a particular type of task. Third, in delayed matching tasks using frequently repeating stimuli, many perirhinal neurons respond differently to the choice stimulus on match and nonmatch trials. However, it has been shown that for many neurons (notably those outside the perirhinal cortex in, for example, the hippocampus) that these match/nonmatch differences are the same for novel as for familiar stimuli; hence, such differences do not signal the novelty or familiarity of a stimulus. Match/nonmatch differences therefore cannot provide a substrate for general familiarity discrimination, although they may provide information important to judging the recency of occurrence of the stimulus.

One other aspect of perirhinal neuronal activity must also play an important role in familiarity discrimination – the correlated firing of action potentials of neurons within local and distal networks. There is still relatively little known about such interactions or the makeup of such networks in the perirhinal cortex. Nevertheless, it has been established that when the activity of synchronously recorded pairs of perirhinal neurons is studied, the firing of one neuron is quite commonly correlated with that of the other. So far, the evidence from such pairs indicates that synchronized firing (when the two neurons fire at approximately the same time) carries less information about the relative familiarity of stimuli than does the firing rate of the neurons.

Evidence for the Local Generation of the Response Changes

Neuronal response changes with the repetition of infrequently repeated stimuli are also found in areas

other than the perirhinal and its neighboring cortices. For example, such changes are found in parts of prefrontal cortex and the striatum, and a small number (~1%) are found in the hippocampus. (Note that neuronal response changes with frequently repeating stimuli are found in even more regions, including the many that demonstrate habituation.) Nevertheless, there is good evidence that response changes with the necessary properties to underlie the familiarity discrimination component of visual recognition memory are initially generated in the perirhinal cortex, although also possibly in the neighboring anterior inferior temporal cortex.

More posteriorly, and earlier in the visual processing pathway, repetition-sensitive response changes have memory spans that last for only a minute or so and are disrupted by one or two intervening stimuli. Such changes cannot explain those found in the perirhinal cortex. Therefore, perirhinal response changes are not merely passive reflections of responses earlier in the visual pathway. There is also evidence that the changes are not passive reflections of those occurring in subsequent processing areas. This evidence comes from the very fast latencies of the response changes found in the monkey perirhinal cortex and neighboring area TE. In the fastest cases, the change between the response to a novel stimulus and that to a familiar stimulus occurs within milliseconds (measurements are accurate to only 10–20 ms) of the response to the visual stimulus itself (70–80 ms). Accordingly, there is no time for feedback from areas further on in the processing hierarchy. Indeed, when such latencies have been measured in the hippocampus, prefrontal cortex, and striatum, they are significantly longer than in the perirhinal cortex. The conclusion is that at least the initial response change must be first generated in the perirhinal cortex (and/or anterior TE). This conclusion receives support from experiments in which drugs have been infused locally into the rat perirhinal cortex. Local infusions of different glutamate receptor antagonists block the acquisition of longer- or shorter-term recognition memory, and no other brain region is able to substitute for the perirhinal cortex for either type of delay period (see the section titled 'Synaptic plasticity and modeling of perirhinal neuronal response changes').

The recording experiments in which repetition-sensitive response changes have been found have employed many controls to establish that the changes are not explicable by behavioral or motivational changes, including changes in alertness, eye movements, and pupil diameter. The speed of the response changes further establishes that they are not artifactually generated because they occur too soon to be explicable as secondary to ocular or attentional changes. It also establishes that these neurons provide a very rapid signal indicating the presence of a novel stimulus. It has been argued that the evolutionary advantage conveyed by the capacity to attend early to novelty provides a reason for the existence of such a network.

Generalization across Modalities and Species

Similar neuronal response changes on stimulus repetition have been described in both the monkey and rat perirhinal cortex. Moreover, recent functional magnetic resonance imaging (fMRI) studies have indicated that signals in the analogous human cortical region are also weaker for familiar than for novel stimuli. Thus, there is a strong indication that the same type of repetition-sensitive response reductions are likely to occur generally across mammalian species.

Such response changes have been largely studied using visual stimuli because of the availability of such stimuli in the very large numbers necessary for such experiments. Because the connections of the perirhinal cortex indicate that it should be a multimodal area, it might be expected that similar results will be found whichever modality is used. Indeed, findings in the rat for olfactory stimuli are consistent with those for visual stimuli. Neuronal activity measurements have not been made with somatosensory stimuli, although perirhinal lesions impair tactile familiarity discrimination. However, lesion experiments suggest that the perirhinal cortex may not be necessary for auditory familiarity discrimination. When neuronal activity is imaged using Fos (which has proved to be a useful marker for imaging activity changes related to visual recognition memory), activity is higher for novel than familiar sounds in the auditory association cortex (as it is for visual stimuli in visual association cortex) adjacent to the perirhinal cortex but not in the perirhinal cortex itself. The implication is that either another region performs the role of the perirhinal cortex for auditory stimuli or that there is a difference in the nature or complexity of the auditory and visual stimuli being employed. The latter suggestion relates to the proposed role of the perirhinal cortex in the perception of objects and the question as to whether the particular auditory stimuli employed are processed as objects. If they are not, this could explain the lack of differential activation of the perirhinal cortex by the novel and familiar sounds.

Synaptic Plasticity and Modeling of Perirhinal Neuronal Response Changes

Descriptions of different types of perirhinal neuronal response changes related to learning have already been presented. It is possible to list these under the

following headings: (1) feature extraction (perceptual categorization), (2) familiarity/recency discrimination (potentially several different mechanisms), (3) paired-associate learning, and (4) reward-association learning. Yet other changes, thought to be related to attentional and/or short-term memory mechanisms, have also been described, such as delay activity and response suppression. It is already clear that there are also developmental changes. The interrelationships and independencies of these different response changes and their underlying synaptic plastic mechanisms are largely unknown, but in the list there are at least eight types of activity change, all of which are observable in the activity of perirhinal neurons under different conditions. The changes observed are dependent on the particular conditions of the experiment and on the particular perirhinal neuron under study. Not all perirhinal neurons show all types of changes, although some may show more than one; again information here is limited.

The perirhinal cortex provides a promising region for the investigation of neural substrates of memory. In recent years, increasing efforts have been made to relate perirhinal neuronal response changes to underlying synaptic plastic changes. This work has variously encompassed paired-associate learning, reward learning, developmental changes, and familiarity discrimination. It has also involved the study of synaptic plasticity mechanisms in slices of the perirhinal cortex.

Perirhinal Plasticity Studied in Brain Slices

In vitro studies have established that both long-term potentiation (LTP) and long-term depression (LTD) can be evoked in the perirhinal cortex by suitable stimulation. The induction of both LTP and LTD are dependent on *N*-methyl-D-aspartate (NMDA) glutamate receptors in the perirhinal cortex, as elsewhere. However, LTD has been shown to have an unusual dependency on activation of metabotropic glutamate receptors. Moreover, in the adult perirhinal cortex LTP is dependent on the activation of NMDA receptors containing NR2A subunits, whereas LTD (and depotentiation) is dependent on NMDA receptors containing NR2B subunits. Again, as in other regions, there are differences between the adult and immature cortices in NMDA glutamate receptor composition and LTD induction mechanisms. Indeed, recently, it has been shown that visual experience triggers such a change between immature and adult perirhinal LTD induction mechanisms.

Signaling Mechanisms Related to Paired-Associate and Reward Sequence Learning

As already mentioned, studies of paired-associate learning in the monkey have established that underlying perirhinal changes involve and are dependent on the immediate early gene *Zif268*. Similarly, reward-sequence learning in the monkey involves dopaminergic mechanisms because perirhinal neuronal response changes are blocked by local infusion of an oligonucleotide that downregulates the expression of D2 dopaminergic receptors. Interestingly, antagonism of the NMDA receptors did not prevent the development of the response changes.

Signaling Mechanisms Related to Recognition Memory

Several studies have been carried out that look at potential substrates of familiarity discrimination in the rat. This work has established that recognition memory is impaired following local infusions into the perirhinal cortex of the muscarinic receptor antagonist scopolamine or of selective antagonists of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), NMDA, kainate, and metabotropic glutamate receptors. When AMPA receptors are blocked, both acquisition and retrieval are impaired. Otherwise, all the studied drug actions involve acquisition and not retrieval. Interestingly, blockade of NMDA and metabotropic glutamate receptors produces recognition memory impairment only when the memory has to be held across a long (24 h) and not shorter (20 min) delay. In contrast and most unusually, kainate receptor antagonism impairs recognition memory after a 20 min but not after a 24 h delay. These results parallel the evidence provided by perirhinal neuronal responses that there must be more than one substrate underlying familiarity discrimination. The indication is that there must be a rapidly induced shorter-term process that is dependent on kainate receptor activation and one or more longer-term processes, more slowly induced and dependent on NMDA and/or metabotropic glutamate receptor activation. In addition, viral transduction of the perirhinal cortex has shown that phosphorylation of cyclic adenosine monophosphate (cAMP) responsive element binding protein (CREB) is necessary for both perirhinal LTP and long-term (24-h) recognition memory, a dependency that parallels findings elsewhere (e.g., in the hippocampus).

Theoretical Models and Plasticity Mechanisms

There are now several theoretical neuronal models both of networks designed to extract features and thereby achieve stimulus categorizations leading to perception and of networks designed to achieve familiarity discrimination. None of these models yet uses specific details of perirhinal circuitry (about which there is still relatively little information). Most of the models have employed synaptic enhancement

(i.e., LTP-like mechanisms) as their primary synaptic plasticity mechanism. One way of achieving the reduced neuronal responses for familiar compared to novel stimuli in such models is by increasing inhibition, but experimental evidence of such increased inhibition is lacking. Recently, it has been argued that models that use separate (but interlinked) networks for feature extraction and familiarity discrimination are more economical. The idea of separate networks is consistent with the observation that perirhinal neurons have responses that are either repetition-sensitive or not, rather than some mixture of the two. Moreover, the calculations indicate that the use of synaptic weakening as the primary mechanism is considerably more efficient than synaptic enhancement. Heuristically, this is because when features repeat across different stimuli, synaptic enhancement leads to certain synapses being successively strengthened so that the neurons with these synapses come to fire selectively strongly for these features (i.e., they signal the presence of such features). In contrast, synaptic weakening has the opposite effect; it emphasizes what is not common rather than what is common across stimuli – but this is what is required for novelty detection. Hence, neuronal modeling has provided an explanation for why perirhinal responses are weaker for familiar than novel stimuli. Under this hypothesis, the synapses that are most strongly activated by a novel stimulus are weakened; when the stimulus occurs again (i.e., when it is familiar), the weakened synapses give rise to a weaker response.

Summary

The position of the perirhinal cortex at the top of the hierarchy of sensory processing areas and at the gateway to the limbic system means that it is ideally placed to play important roles in sensory perception and memory. Ablation evidence indicates that the perirhinal cortex is indeed involved in both perceptual (stimulus categorization) and mnemonic functions. The results of recording studies established that the responses of perirhinal cortical neurons carry information about complex stimuli and relate to stimulus identity, but they have yet to provide evidence for the categorization of objects as entities (as suggested by certain lesion studies). There is much evidence from recording studies that perirhinal neurons are involved in attentive and short-term memory, as well as multiple long-term memory tasks including sequence learning for reward, paired-associate learning, and recognition memory. A variety of different neuronal response changes have been related to these different memory tasks, and studies of their underlying synaptic plastic mechanisms are underway.

See also: Adult Cortical Plasticity; Cortical Plasticity and Learning: Mechanisms and Models; Hippocampus and Neural Representations; Memory Representation; Neural Patterning: Arealization of the Cortex; Neural Coding of Spatial Representations; Neuronal Plasticity after Cortical Damage; Perceptual Learning: Neural Mechanisms; Perceptual Learning and Sensory Plasticity; Perirhinal Cortex; Recognition Memory; Shape Representation in Inferotemporal Cortex; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Mechanisms of Learning; Visual Associative Memory.

Further Reading

Barker GRI, Warburton EC, Koder T, et al. (2006) The different effects on recognition memory of perirhinal kainate and NMDA glutamate receptor antagonism: Implications for Underlying plasticity mechanisms. *Journal of Neuroscience* 26: 3561–3566.

Brown MW (1990) Why does the cortex have a hippocampus? In: Gabriel M and Moore J (eds.) *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pp. 233–282. Cambridge, MA: MIT Press

Brown MW (2000) Neuronal correlates of recognition memory. In: Bolhuis JJ (ed.) *Brain, Perception and Memory*, pp.185–208. Oxford: Oxford University Press

Brown MW and Aggleton JP (2001) Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience* 2: 51–61.

Brown MW and Bashir ZI (2002) Evidence concerning how neurons of the perirhinal cortex may effect familiarity discrimination. *Philosophical Transactions of the Royal Society of London B* 357: 1083–1095.

Brown MW and Xiang JZ (1998) Recognition memory: Neuronal substrates of the judgement of prior occurrence. *Progress in Neurobiology* 55: 149–189.

Desimone R (1996) Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences of the United States of America* 93: 13494–13499.

Eichenbaum H (2000) A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience* 1: 41–50.

Henson RN, Cansino S, Herron JE, Robb WG, and Rugg MD (2003) A familiarity signal in human anterior medial temporal cortex? *Hippocampus* 13: 301–304.

Massey PV, Johnson BE, Moult PR, et al. (2004) Differential roles of NR2a and NR2b-containing NMDA receptors in cortical long-term potentiation and long-term depression. *Journal of Neuroscience* 24: 7821–7828.

Miyashita Y (2004) Neural mechanisms of cognitive memory. *Keio Journal of Medicine* 53: 59–68.

Murray EA and Bussey TJ (1999) Perceptual-mnemonic functions of the perirhinal cortex. *Trends in Cognitive Neuroscience* 3: 142–151.

Murray EA and Richmond BJ (2001) Role of perirhinal cortex in object perception, memory, and associations. *Current Opinion in Neurobiology* 11: 188–193.

Naya Y, Yoshida M, and Miyashita Y (2001) Backward spreading of memory retrieval signal in primate temporal cortex. *Science* 291: 661–664

Naya Y, Yoshida M, and Miyashita Y (2003) Forward processing of long-term associative memory in monkey inferotemporal cortex. *Journal of Neuroscience* 23: 2861–2871

Rugg MD and Yonelinas AP (2003) Human recognition memory: A cognitive neuroscience perspective. *Trends in Cognitive Sciences* 7: 313–319.

Squire LR, Stark CE, and Clark RE (2004) The medial temporal lobe. *Annual Review of Neuroscience* 27: 279–306.

Suzuki WA and Eichenbaum H (2000) The neurophysiology of memory. *Annals of the New York Academy of Sciences* 911: 175–191.

Witter MP, Groenewegen HJ, Lopes da Silva FH, and Lohman AHM (1989) Functional organization of the extrinsic and intrinsic circuitry of the parahippocampal region *Progress in Neurobiology* 33: 161–253.

Witter MP, Naber PA, VanHaeften T, et al. (2000) Cortico-hippocampal communication by way of parallel parahippocampal-subicular pathways. *Hippocampus* 10: 398–410.

Peroxisomal Disorders and Neurological Disease

N Braverman, Johns Hopkins University School of Medicine, Baltimore, MD, USA
F Eichler, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Peroxisomes are ubiquitous, usually spherical, single-membrane-bound organelles, 0.2–1 μm in diameter and numbering several hundred per mammalian cell. First described in 1958 as microbodies in mouse kidney by Johannes Rhodin, their name derives from the peroxide generated in these organelles as a by-product of oxidative reactions. Zellweger syndrome (ZS) was originally classified as a congenital malformation syndrome, the cerebrohepatorenal syndrome, and X-linked adrenoleukodystrophy (X-ALD), a sudanophilic leukodystrophy referred to as Schilder disease. A peroxisome etiology for these disorders was not recognized until the assignment of enzyme defects to the peroxisome, thus defining a new category of human metabolic disease. More than 50 enzymes have since been localized to the peroxisome matrix. The majority are involved in lipid metabolism, including the oxidation of very long chain fatty acids (VLCFAs) and the synthesis of bile acids and ether phospholipids. We now recognize two general classes of peroxisome disorders, the single-enzyme (or protein) deficiencies (SED), including X-ALD, and the peroxisome biogenesis disorders (PBDs). Failure of the peroxisome to assemble correctly leads to multiple enzyme deficiencies in PBDs. Although the underlying gene defects in these disorders have now been identified, the functions of the encoded proteins are still being defined; it remains to relate the clinical features with the disturbances in biochemical pathways. This should lead to improved therapeutic approaches. In this article, we review the clinical phenotypes and molecular genetics of X-ALD and PBDs and discuss current investigations that relate to pathogenesis.

X-Linked Adrenoleukodystrophy

Clinical Phenotypes

X-ALD causes varying degrees of malfunction of the adrenal cortex and nervous system myelin. The estimated overall frequency is 1:17 000. There are several distinct phenotypes. Approximately 35% of patients have the childhood cerebral form. These affected boys develop normally until 4–8 years of age, then suffer dementia and a progressive neurologic deficit that leads to a vegetative state and death. More than 90% of these patients have adrenal insufficiency. In approximately 35–40% of patients, the disorder presents in young adulthood as a slowly progressive paraparesis with sphincter disturbances that primarily involve the long tracts in the spinal cord, referred to as adrenomyeloneuropathy (AMN). Recently, Van Geel et al. demonstrated that 20% of these patients develop progressive inflammatory demyelination, to which they succumb in 1–2 years. This appears to follow the same pattern as the childhood form of cerebral adrenoleukodystrophy, albeit slower in progression. Female heterozygotes of X-ALD are also known to have symptoms related to the underlying metabolic disorder. Approximately 60% have abnormalities resembling adrenomyeloneuropathy. A third of these patients have severe symptoms, another third have slight to moderately severe symptoms, and the remaining third are asymptomatic but have hyperreflexia and diminished vibration sense in the lower extremities.

Genetics

The adrenoleukodystrophy gene, located at Xq28, encodes an integral peroxisome membrane protein with the consensus structure of an ATP-binding cassette (ABC) half-transporter. The gene name has been designated *ABCD1*, for subfamily D, member 1 of a large family of proteins with similar domains, conserved through evolution. The protein is still referred to as the adrenoleukodystrophy protein (ALDP). The functional role of ALDP is currently being investigated. It has been shown to both homodimerize and heterodimerize with related peroxisomal membrane ABC half-transporters, ALDRP, PMP70, and PMP70R, presumably to form a full transporter. The expression pattern of these related proteins varies according to tissue type and time in development, and the substrates transported might depend on the assembled transporter. Since VLCFAs accumulate in X-ALD, and peroxisomal β-oxidation is reduced in spite of the presence of normal enzymatic machinery inside the peroxisome, it has been assumed that ALDP transports either VLCFAs, VLCFA-CoA, or the VLCFA acyl-CoA synthetase. However, the specific substrates for the peroxisome ABC transporters remain to be demonstrated. In the interim, evidence has been presented to suggest that the accumulation of VLCFAs in X-ALD might be secondary to enhanced chain elongation of fatty acids, rather than reduced β-oxidation. This is supported by reports that

ABCD1-deficient mice have elevated VLCFAs, but normal peroxisomal β-oxidation in tissues, and mice with disruption of very long-chain acyl-CoA synthetase do not accumulate VLCFAs, but are deficient in β-oxidation.

To date, more than 450 mutations have been identified in *ABCD1* and are described in the X-ALD database. There are no common alleles, nor a general correlation between genotype and phenotype. Linkage analysis in families has provided support for the hypothesis that at least one autosomal gene, acting in a dominant fashion, can influence phenotype determination. Candidate genes include those encoding proteins involved in the inflammatory response or those involved in the biological pathway of ALDP. However, within a family, the same mutation, whether a missense or deletion allele, can result in any of the ALD phenotypes, and even monozygotic twins with discordant phenotypes have been reported. Thus the disease phenotype that develops might be determined by the independent inheritance of modifier genes, as well as exposure to unknown environmental factors.

VLFCAs and Membrane Stability

VLCFAs may play an important role in the pathogenesis of cerebral demyelination of peroxisomal disorders. Recently it was shown that elevations of C26:0 in lysophosphatidylcholine can be reliably demonstrated in blood spots by tandem mass spectrometry, opening up the possibility of screening for these disorders. It is worth considering the biophysical properties of VLCFAs in this context. Long-chain fatty acids have backbones of 14–18 carbon atoms, designated C14–18, and comprise the majority of dietary fatty acids. VLCFAs have 20 or more carbons. Normally saturated VLCFAs occur in highest concentration in myelin lipids and red blood cell sphingomyelin (C26:0 accounts for 1–5% of total fatty acids in brain cerebrosides and sulfatides). The VLCFAs that accumulate in X-ALD are saturated and unbranched (mainly C24:0 and C26:0). The excess is most striking in the cholesterol ester and ganglioside fractions of affected brain white matter and adrenal cortex, with the phosphatidylcholine fraction containing a 39-fold excess of C26.

In X-ALD, the initial locus of demyelination occurs with reliable regularity in the splenium or genu of the corpus callosum. These regions are most notable for their tight fiber packing. Since VLCFAs diffuse quickly across membranes but are slow to desorb, the tight fiber packing may lead to 'trapping' within the phospholipid bilayer. Further, thickening of the membrane and alteration of membrane topology may be more harmful to regions of higher, rather than lower, fiber density. It is thought that the decreased mobility of VLCFAs makes them 'sitting targets' for activation through acyl-CoA, hence leading to further incorporation into membrane phospholipids.

The greater length of the aliphatic chain causes VLCFAs to be insoluble and alters their physiological properties. The most direct evidence that abnormally high VLCFA levels can alter membrane function is provided by the study of Whitcomb et al., who assessed adrenocorticotropic hormone (ACTH)-stimulated cortisol release in cultured human adrenocortical cells. The addition of C26:0 or C24:0 to the culture medium in concentrations equivalent to those in X-ALD plasma increased the microviscosity of adrenocortical cell membranes and decreased ACTH-stimulated cortisol secretion. These investigators concluded that the VLCFA excess altered membrane structure and suppressed the availability of the ACTH receptor.

Neuroimaging and Pathology

Cerebral involvement in patients with X-ALD is most often first detected in neuroimaging studies (**Figure 1**). Indeed, neuroimaging provides valuable information about prognosis and evaluation of therapies. Magnetic resonance image (MRI) abnormalities most often precede onset of neurologic symptoms, but given the rapid progressive course only precede death by 1–2 years. Eighty percent of ALD patients with cerebral involvement show white-matter lesions that are characteristic with respect to location and attenuation pattern. In these patients, the lesions are symmetric and involve periventricular white matter in the posterior parietal and occipital lobes. These regions are generally hyperintense on fluid-attenuated inversion recovery (FLAIR) sequence or T2-weighted magnetic resonance images. Use of the contrast material gadolinium diethylenetriamine pentaacetic acid (DPTA) demonstrates a garland of accumulated contrast material adjacent and anterior to the posterior hypodense lesions shown on T1-weighted MRIs. Contrast enhancement at the edges of the white-matter lesions is so characteristic as to be almost diagnostic. Proton magnetic resonance spectroscopy (MRS) can anticipate future MRI abnormalities in white matter of ALD patients while it still appears normal on a conventional MRI: MRS spectra demonstrate increased choline and decreased *N*-acetylaspartate (NAA), presumed to reflect biochemical alterations in myelin sheaths and axonal dysfunction, respectively. Even in the normal-appearing white matter of patients with only the Addisonian phenotype, NAA is lower than normal controls.

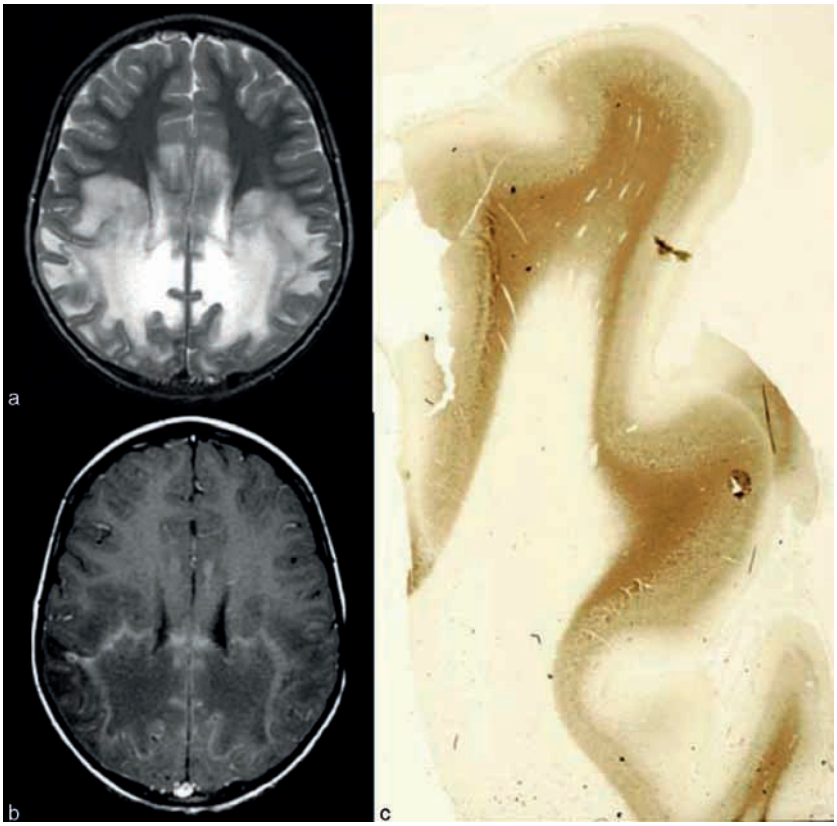

Figure 1 X-linked adrenoleukodystrophy patient with confluent white-matter lesions on T2-weighted imaging (a) and garland of contrast enhancement following gadolinium injection (b). Major basic protein staining of the occipital lobe shows sparing of the arcuate fibers (c).

Accumulation of contrast material on the MRI correlates well with the histopathological zones described by Schaumburg et al. (Figure 1). Demyelination usually affects the parieto-occipital regions most severely and earliest in the course of the disease. As confirmed by neuroimaging studies, these lesions progress, often asymmetrically, toward the frontal lobes. The edge between the lesion and 'normal' white matter can be distinct or blurred, and subcortical U fibers are usually spared. Atrophy of the corpus callosum and optic nerves is typical, as is the early and prominent involvement of the splenium and posterior limbs of the internal capsule.

The cortical and deep gray matter are unremarkable, as is the brain stem, except for some patients that exhibit foci in the basis pontis or in the pyramidal tracts within the cerebral peduncles, basis pontis, and medulla. Particularly in adulthood the spinal cord may show atrophy of the lateral corticospinal tracts and dorsal columns (adrenomyeloneuropathy). This is not seen in childhood except in the course of extensive downward progression from cerebral lesions. ALD is best known for the intensity and extent of its central nervous system (CNS) white-matter disease, which consists of confluent areas of myelin and oligodendrocytic loss that is replaced by fibrillary astrogliosis. It is important to note that neither the extent nor the intensity of these lesions has been correlated with clinical (hypocortisolism) or pathological defects in the adrenal cortex.

Inflammatory Demyelination in X-ALD

After the initial injury at the site of tightly packed fibers in the corpus callosum, a cascade of inflammatory demyelination ensues. Correlation of clinical and pathologic findings indicates that the progression in cerebral X-ALD is related to the perivascular infiltration of lymphocytes in the brain white matter. This is the most distinctive hallmark of cerebral X-ALD and sets it apart from all other leukodystrophies.

The white-matter lesion shows marked chronic inflammation consisting largely of T lymphocytes, particularly cytotoxic CD8 cells in the early stage, and macrophages. This resembles multiple sclerosis (MS), but differs from it by the paucity of

B cells/plasma cells and the localization of the lymphocytes at the lesion edge in MS and within the demyelinated region in ALD. Areas in which these lesions are present show breakdown of the blood–brain barrier and accumulation of contrast material on gadolinium MRI. Multiple immune cells and effector molecules have been identified in the lesions and on microglial, astrocytic, and endothelial cells, including major histocompatibility complex (MHC) class I and II molecules, CD4 and CD8 cells, tumor necrosis factor-α (TNF-α), interleukin-1, intercellular adhesion molecule (ICAM), nitric oxide, and CD1.

While the exact triggers for the inflammatory response remain unknown, it is speculated that the intracellular accumulation of VLCFAs and the resulting alteration in myelin metabolism poisons cells, leading to cell death and the induction of an inflammatory response. The cascade of inflammatory demyelination appears to be cytokine mediated, perhaps by stimulating TNF-α production in a manner analogous to the well-known effect of lipopolysaccharide. Ganglioside, phosphatidylcholine, proteolipid, and cholesterol ester fractions show the greatest excess of VLCFAs and may be plausible candidates as triggers of demyelination.

Animal Models

ABCD1 null mice have been generated and show tissue accumulation of VLCFAs leading to a detectable clinical phenotype comparable to AMN. Around 15 months of age the ALD-deficient mice exhibit an abnormal neurological and behavioral phenotype. This correlates with slower nerve conduction, and with myelin and axonal anomalies detectable in the spinal cord and sciatic nerve, but not in brain. Thus this model should be useful for investigating the pathogenesis of AMN.

The murine locus for *ABCD2*, encoding ALDRP, was also inactivated, and these mice develop late-onset cerebellar and sensory ataxia. This correlates with a loss of cerebellar Purkinje cells and dorsal root ganglia cell degeneration. Interestingly, none of the mouse models so far has developed the inflammatory demyelination characteristic of human ALD. Research on triggers of inflammation in the setting of elevated VLCFAs is currently lacking. An interesting feature found in both the *ABCD1* and the *ABCD2* knockout mouse models is the presence of ultrastructural alterations of mitochondria in axons and neuronal bodies. The contribution of these mitochondrial alterations to disease pathogenesis remains to be clarified.

Therapy

Table 1 summarizes the therapeutic approaches in peroxisomal disorders. All patients with impaired adrenal function should receive replacement therapy with adrenal steroids. However, this does not impact neurological progression. As neurological disease is thought to be related to lipid accumulation and ensuing inflammation, disease-modifying therapies have focused on both metabolic reversal and immunomodulation.

In asymptomatic boys whose brain MRI is normal, dietary therapy with 'Lorenzo's oil (LO)' helps prevent the onset of cerebral disease. LO is a 4:1 mixture of glyceryl trioleate and glyceryl trierucate. Oral administration of this oil, combined with moderate reduction of fat intake, normalizes or significantly lowers the levels of VLCFAs in the plasma of patients

Table 1 Management of X-linked adrenoleukodystrophy and peroxisome biogenesis disorders

Phenotype	*Follow-up tests*	*Recommended therapy*
Asymptomatic X-ALD	Monitor MRI and adrenal function; neurological and neuropsychological test; family screening	Lorenzo's oil, HSCT if MRI abnormal
Addison-only X-ALD	Monitor MRI; neurological and neuropsychological test	Lorenzo's oil, HSCT if MRI abnormal, adrenal Rx
Cerebral X-ALD	Neurological and neuropsychological test; family screening; adrenal function	HSCT if mild, adrenal Rx
AMN	Brain MRI, adrenal function; SSEP; EEG; psychiatry; family screening	Adrenal Rx, physical therapy, Lorenzo's oil (?), HSCT (?)
X-ALD heterozygotes	Adrenal function; MTS; SSEP; EEG; family screening	Genetic counseling, physical therapy, adrenal Rx, Lorenzo's oil (?), general support
PBD	VEP; BAERs; ophthalmological examination; brain MRI; renal and cardiac ultrasound; hepatic enzymes and clotting factors	Vitamin K supplementation; dietary measures (?) (DHA and bile acid supplementation, phytanic acid restriction)

AMN, adrenomyeloneuropathy; BAER, brain stem auditory evoked response; DHA, docosahexanoic acid; EEG, electroencephalogram; HSCT, hematopoietic stem cell transplantation; MRI, magnetic resonance imaging; MTS, mitochondrial targeting signal; PBD, peroxisome biogenesis disorder; SSEP, somatosensory evoked potential; VEP, visual evoked potential; X-ALD, X-linked adrenoleukodystrophy. Adapted from Moser HW, Raymond GV, Lu S-E, et al. (2005) Follow-up of 89 asymptomatic patients with adrenoleukodystrophy treated with Lorenzo's oil. *Archives of Neurology* 62(7): 1073–1080.

with X-ALD within 4 weeks. This does not significantly alter the rate of disease progression in boys who are already symptomatic when therapy is initiated, particularly those with the cerebral inflammatory phenotypes. However, a preventive effect is seen in asymptomatic boys whose brain MRI is normal. Another candidate for lowering VLCFAs may be lovastatin. Treatment with lovastatin may reduce high plasma levels of hexacosanoic acid (C26:0) by a shift to monounsaturated VLCFAs, but the clinical benefit remains to be demonstrated. Recently it was recognized that omega oxidation may be an alternative pathway for VLCFA degradation. This pathway accounts for 5–10% of total fatty acid oxidation in the liver and is catalyzed by P450 enzymes. Finally, overexpression of ALDRP can correct VLCFA levels in human cells and in the *ABCD1* mouse model. In the future, stimulation of these enzymes may be available for therapeutic intervention in X-ALD.

Cerebral inflammation occurs in less than half of patients with the known gene defect and remains the disease aspect requiring utmost vigilance and monitoring. For patients in the early stage of the cerebral inflammatory phenotype, hematopoietic stem cell transplantation (HSCT) leads to improved neurologic outcome and survival. Although the mechanisms of the benefit of HSCT are unclear, it is thought that donor-derived cells enter the nervous system and gradually replace perivascular microglia, thereby ameliorating the metabolic milieu. Alternatively, or in addition, the favorable effect may be due to the immunosuppression implemented during the HSCT preregimen. Peters et al. have reported the follow-up of 126 patients with X-ALD who had received either bone marrow or umbilical cord HSCT during the period 1981–99. Outcome was favorable in patients with the following characteristics at the time of transplant: early stage of illness, limited neurological deficits, and limited MRI involvement. In this group the 5-year survival was 92% and the neurological and functional status far superior to that in the nontransplanted group. For the group with more advanced deficits, 5-year survival for the transplant group was less than that in the nontransplant group (45% vs. 60%). Reduced-intensity preparative regimens for HSCT may become an important approach in X-ALD patients who have marked progression of neurologic symptoms to reduce regimen-related toxicities and multiorgan system complications. Additionally, cord blood transplantation may be an option. Regardless of procedure, HSCT is strongly recommended only for patients in the early stage of cerebral X-ALD.

Immunosuppressive treatment in cerebral ALD has shown limited success, though has only been tested in small numbers of patients with significant neurologic deficit and advanced disease. Interferon-β and intravenously administered immunoglobulin (IVIg) do not appear to halt the course of progressive cerebral ALD. The use of cyclophosphamide and steroids has been reported in five patients (ages 6–20 years) with rapidly progressing neurologic symptoms and abnormal brain computed tomography (CT) scans 4–7 months after their first neurologic symptom. Rate of progression for four patients was unchanged compared to controls; however, the fifth patient remained stable at 3 years of follow-up. Given the success of immunomodulation in autoimmune diseases such as multiple sclerosis, the testing of available immunotherapies in cerebral ALD remains an area of great interest.

Peroxisome Biogenesis Disorders

Clinical Phenotypes

The PBDs are a heterogeneous group of autosomal recessive diseases with an estimated incidence of 1 in 50 000. Based on clinical, biochemical, and molecular criteria, they are divided into the Zellweger syndrome spectrum (ZSS) and rhizomelic chondrosyplasia punctata (RCDP). ZSS includes ZS, neonatal adrenoleukodystrophy (NALD), and infantile Refsum disease (IRD) and represents a continuum of disease, from severe to mild, rather than distinct disorders. ZS patients are distinguished by facial dysmorphology, profound hypotonia, and minimal psychomotor development. Clinical features include cataracts, glaucoma, and optic atrophy, progressive hepatocellular disease, renal cortical cysts, and epiphyseal stippling (or chondrodysplasia punctata) of the patella. Adrenal function may be deficient on provocative testing or during episodes of stress. The incidence of cardiac and other malformations is increased over the general population. The EEG, auditory, and somatosensory evoked responses are typically abnormal. Most ZS patients die within the first year of life due to secondary infection, aspiration, or progressive liver disease. NALD patients usually survive the first year, but can develop a leukodystrophy with a degenerative course characterized by demyelination and cortical atrophy. The IRD group may survive through adolescence and manifest moderate to severe mental retardation, hypotonia, progressive retinopathy, and sensorineural deafness. Facial dysmorphology is less apparent in NALD and IRD. Around half of PBD ZSS patients have phenotypes consistent with IRD. The full spectrum of phenotypes continues to unfold and has included patients with cerebellar ataxia and cholestasis, Usher-like syndromes, peripheral neuropathy, and late-onset leukodystrophy.

PBD RCDP includes classic RCDP1 and milder variants. The incidence of RCDP is around 1/100 000. Patients rarely survive beyond late childhood and have rhizomelia, extensive chondrodysplasia punctata, cataracts, seizures, and profound growth and mental retardation. Cervical spine stenosis is reported and can be accompanied by spinal cord compression. As in PBD ZSS, there is increased incidence of congenital malformations. Death is often secondary to respiratory complications from chronic aspiration, restricted chest wall movement, and a small pulmonary cage. The recent identification of several patients with features of adult Refsum disease (ARD) and *PEX7* gene defects suggests that RCDP1 might also be viewed as a disease continuum.

Genetics

Around 15 mammalian genes required for peroxisome assembly have now been identified; they are termed PEX genes and the proteins they encode are peroxins. PBD ZSS is associated with defects in 12 PEX genes, whereas RCDP1 is associated only with *PEX7*. The frequency of gene defects and roles for the encoded peroxins in the assembly process are listed in Table 2. Despite the high frequency of a few founder alleles, it is not possible to predict the PEX gene defect of a ZSS patient based on phenotype alone. However, PEX mutations predicted to result in null alleles are associated with ZS, whereas missense mutations are more often associated with NALD or IRD. Indeed, a number of patient surveys have shown that homozygosity for the nonsense alleles, *PEX1*-Ile700TyrfsX42 and *PEX10*-Leu272-ValfsX65, predicted a ZS phenotype, whereas homozygosity for *PEX1*-Gly843Asp predicted IRD. For RCDP1, all patients homozygous for the frequent nonsense allele, *PEX7*-Leu292X, have classical disease.

Biochemical Deficiencies

Mammalian peroxisomes are a hub of vital enzymatic activities that include β-oxidation of VLCFAs, oxidation of D-amino acids and L-lysine, ether phospholipid synthesis, and glyoxylate and polyamine catabolism, as well as oxidoreductase reactions that both generate and degrade oxygen radicals. The peroxisomal β-oxidation pathway degrades straight chain and branched, saturated and unsaturated

Table 2 PEX gene defects associated with peroxisome biogenesis disorders

PEX gene[a]	*Frequency in peroxisome biogenesis disorders in Zellweger syndrome spectrum*	*Role in peroxisome* assembly[b]
PEX16	<1%	Membrane synthesis
PEX19	<1%	
PEX3	<1%	
PEX5	2%	Peroxisome targeting signal 1 (PTS1) receptor
PEX13	<1%	Docking site for matrix protein import
PEX14	<1%	
PEX2	3%	Translocation site for protein import
PEX10	3%	
PEX12	5%	
PEX26	5%	Recycling of targeting receptors
PEX6	10%	
PEX1	70%	
PEX7	100% of RCDP1, a few patients with ARD[c]	Peroxisome targeting signal 2 (PTS2) receptor

[a]The designation for the encoded protein is peroxin (e.g., *PEX16* encodes peroxin 16).
[b]Reviewed by Weller S, Gould SJ, and Valle D (2003) Peroxisome biogenesis disorders. *Annual Review of Genomics and Human Genetics* 4: 165–211.
[c]RCDP1, Rhizomelic chondrodysplasia punctata type 1; ARD, adult Refsum disease. Data from Braverman N, Chen L, Lin P, et al. (2002) Mutation analysis of PEX7 in 60 probands with rhizomelic chondrodysplasia punctata and functional correlations of genotype with phenotype. *Human Mutation* 20(4): 284–297, and from Van den Brink DM, Brites P, Haasjes J, et al. (2003) Identification of PEX7 as the second gene involved in Refsum disease. *American Journal of Human Genetics* 72(2): 471–477.
Adapted from Steinberg S, Chen L, Wei L, et al. (2004) The PEX gene screen: Molecular diagnosis of peroxisome biogenesis disorders in the Zellweger syndrome spectrum. *Molecular Genetics and Metabolism* 83: 252–263. Based on 300 patients studied at Kennedy Krieger Institute.

fatty acids with a carbon chain length >C22, as well as long-chain dicarboxylic acids. In addition, peroxisomal β-oxidation is required for the synthesis of bile acids and docosahexanoic acid (DHA), a polyunsaturated fatty acid found in myelin and photoreceptor cells, and the inactivation of certain prostaglandins, leukotrienes, and xenobiotics. The peroxisomal enzymes dihydroxyacetone phosphate acyltransferase and alkyldihydroxyacetone phosphate synthase initiate ether phospholipid synthesis. Subsequent modifications lead to the generation of plasmalogen species. Plasmalogens are important membrane constituents in a variety of tissues, including myelin. There is evidence to support a role for these molecules in membrane fusion events and signaling pathways. In addition, plasmalogens can protect cell membranes from free radical damage through their vinyl–ether linkage.

The PBDs are diagnosed by documenting deficiencies in more than one biochemical pathway of the peroxisome. Tests routinely used are measurements of plasma VLCFAs and phytanic acid (PA), and erythrocyte plasmalogens. In addition, pipecolic acid, produced from lysine, and bile acid intermediates can be measured. Once metabolite levels consistent with peroxisome enzyme deficiencies are obtained, it is recommended that the enzymatic defects be confirmed in fibroblast cultures. In general, the degree of the enzymatic defect and biochemical disturbance correlates with the severity of disease. RCDP1 patients have a subset of biochemical abnormalities restricted to defects in peroxisome targeting signal 2 (PTS2) protein import: they are deficient only in plasmalogen synthesis and phytanic acid oxidation (PAO). Although peroxisomal thiolase is a PTS2 protein, RCDP1 patients have normal β-oxidation of VLCFAs, compensated by a second peroxisomal enzyme with thiolase activity, sterol carrier protein 2 (SCP2).

Around 15–20% of patients with a clinical diagnosis of ZSS will have a SED in β-oxidation, either acyl-CoA oxidase (ACOX1) or multifunctional protein-2 (MFP-2). Similarly, around 5% of RCDP patients will have a SED in the two peroxisomal enzymes required for plasmalogen synthesis. Given the congruity in these SED and PBD clinical phenotypes, it has been suggested that the major pathology of ZSS is determined by defects in β-oxidation and that plasmalogen deficiency determines RCDP. While this seems to be the case for RCDP, it remains to define the normal functions of plasmalogens in tissues and relate this to the disease process. Narrowing the pathway in ZS is more challenging, since peroxisomal β-oxidation affects a broad range of cellular metabolites and processes.

Cellular Phenotypes

In addition to the correlation between clinical phenotype and genotype in the PBDs, there is also a correlation between these parameters and cellular phenotype in ZSS patients. Peroxisome morphology and matrix protein import can be visualized in cells using immunohistochemistry. In ZS patients, peroxisomes are reduced in number, enlarged in size, and have a limited capacity to import matrix proteins; they are aptly called 'ghosts.' Patients with milder disease tend to have more abundant peroxisomes that import more matrix proteins. In fact, residual peroxin function in cell lines from patients with milder disease can be further enhanced, and this is attractive for the design of future drug-based therapies. First, when fibroblasts from NALD and IRD patients are cultured at 30 °C, there is partial recovery of peroxisome number, matrix protein import, and enzymatic activities. The mechanism was shown to be improved protein folding at lower temperature or increased amounts of an otherwise degraded peroxin. Furthermore, when PBD cells are cultured in 4-phenylbutyrate, a compound that stimulates peroxisome proliferation, matrix protein import and enzyme activity improve. Finally, fibroblasts from NALD and IRD patients routinely cultured at 37 °C frequently exhibit a characteristic 'mosaic' pattern in which some cells import matrix proteins, but adjacent cells in the same microscopic field do not, implicating microenvironmental differences that might promote residual peroxin function.

Neuroimaging and Pathology

A number of neuropathological investigations have been reported and delineate three major types of neurological involvement in ZS: neuronal migration abnormalities, hypomyelination, and selective neuronal involvement. Gross inspection frequently shows polymicrogyria and pachygyria, hypoplasia of the corpus collosum, and deficient olfactory lobes. The inferior olive is disorganized, and hypoplasia of the cerebellar vermis can be present. These changes coincide microscopically with heterotopias of cerebral neurons and Purkinje cells, the presence of immature neurons, and reduction in neuron numbers. The pattern of medial pachygyria and lateral polymicrogyria is a distinguishing feature of ZS (Figure 2). Migration defects in the NALD brain are more variable and less severe. Polymicrogyria and neuronal cell loss can be seen, but extensive demyelination is the hallmark of this disorder, apparent in the cerebrum, midbrain, and cerebellum. Most cases have shown a widespread sudanophilic leukodystrophy, often associated with perivascular accumulation

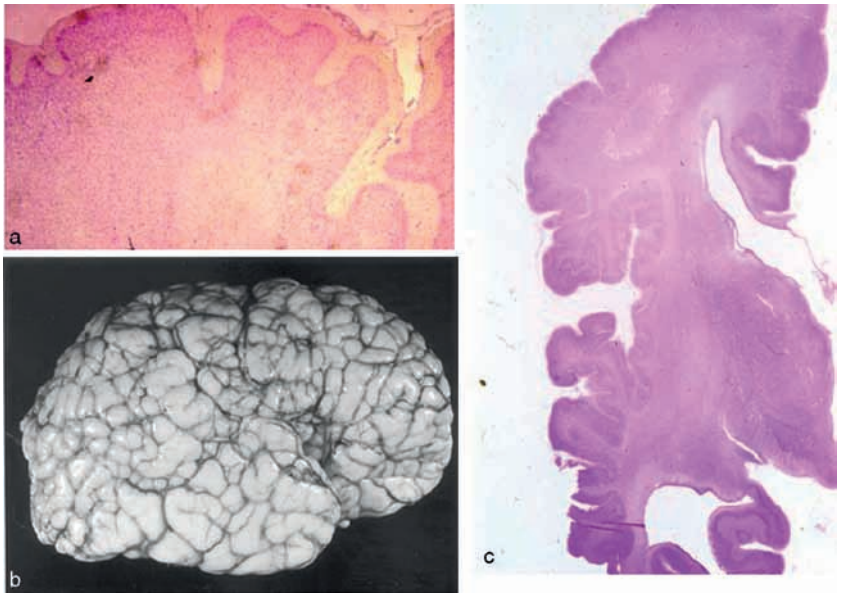

Figure 2 Zellweger patient with perisylvian polymicrogyria on hematoxylin and eosin stain (a, c) and gross brain inspection (b).

of lymphocytes as in X-ALD. The boundaries of the demyelinated areas are usually sharply demarcated and there is reactive gliosis. Studies of IRD brain, although limited, show reduced neuron number and cerebellar atrophy without active demyelination. There can be dysplastic olives and Purkinje cell heterotopia. Thus early defects in brain morphogenesis correspond to the degree of peroxisome dysfunction. Defects in β-oxidation remain a likely suspect in pathogenesis, as brains from patients with D-bifunctional protein deficiency also show a similar spectrum of abnormalities.

MRI of the brain shows abnormal cortical gyral patterns, with regions of microgyria primarily in the frontal and perisylvian cortex, together with thickened pachygyric cortex in the perirolandic and occipital regions. Ventricular dilatation and subependymal germinolytic cysts are seen. In RCDP, cranial MRI shows delayed myelination and both cerebral and cerebellar atrophy. Decreased choline:creatine ratios and increased levels of mobile lipids have been observed on MRS and might reflect plasmalogen deficiency, the major biochemical abnormality in RCDP. The milder patients have less severe skeletal involvement and psychomotor defects, and may have normal MRI studies.

VLCFAs accumulate in membrane phospholipids in ZS brain and in intracellular trilamellar inclusions composed of VLCFA cholesterol esters. The accumulation of VLCFAs might have a toxic effect on cells, inhibit neuron migrations, and disrupt myelin. Plasmalogen levels in ZS brain are also severely deficient and likely contribute to pathology. However, striking neuronal migration abnormalities are not seen in RCDP. In RCDP brain, there is a general decrease in neuron number and white matter, dysplastic olives, and progressive cerebellar degeneration. An additional consequence of plasmalogen deficiency is the accumulation of the fatty alcohol substrates, demonstrated by Rizzo et al. in plasma from RCDP patients. Fatty alcohols may contribute to the neurological features in RCDP, based on features similar to Sjögren–Larsson syndrome.

Animal Models

In order to gain further insight into the pathogenesis of ZS, mouse models with complete deficiency of *PEX2*, *PEX5*, *PEX13*, and *PEX11β* have been engineered. These mice recapitulate many of the aspects of the ZS phenotype, including the metabolic consequences of peroxisome elimination and abnormalities in brain formation, hypotonia, and early demise. Detailed investigations of these models reveal an altered distribution of cortical neurons and defects in neuronal migration, commencing early in the cytogenetic epoch and demonstrated by birthdating experiments using 5-bromodeoxyuridine (5-BrdU). Other abnormalities include a delay in neuronal differentiation and extensive neuronal apoptosis (*PEX5*, *PEX11β*). Cerebellar layer formation is delayed, and there is increased apoptosis in granular neurons. Purkinje cell morphology is abnormal, displaying stunted dendritic trees and abnormal arborization (*PEX2*). In *PEX11β* null mice, peroxisome abundance is decreased, but there are no defects in matrix protein import and only mild deficiencies of β-oxidation and ether lipid biosynthesis, challenging the idea that the

clinical features are directly related to the severity of the metabolic abnormalities. An alternative model of pathogenesis has been proposed in which the disease phenotype correlates with the abundance and distribution of peroxisomes, in addition to their metabolic activity. Further, mice with targeted disruptions of *PEX5* in neuron and glial cells, or in hepatocytes, have been engineered. Although complete peroxisome elimination is not achieved until birth, the accumulation of VLCFAs at midgestation is as severe as that seen in the generalized *PEX5* null model. Targeted elimination of peroxisomes in the brain results in moderate cortical migration defects that correct postnatally and might reflect migration delay. Local elimination of peroxisomes from the liver results in postnatal arrest of neuron migration, particularly in the cerebellum, which still develops in the postnatal period. Thus, peroxisome metabolism in extraneuronal tissues influences the development of the murine brain.

Variable abnormalities of mitochondria have been noted in liver, brain, and muscle from ZS patients, specifically pleomorphic mitochondria with distorted cristae and reduction in activity of respiratory chain complexes. This might result from increased oxidative stress secondary to peroxisome deficiency and might equally be a factor in the pathogenesis of ZS. Similar mitochondrial abnormalities are observed in tissues from PEX null mice. The mouse models have also been used to test potential therapeutic interventions. Supplementation with DHA, beginning *in utero*, normalizes DHA levels in brain but does not improve the migration disorder. Bile acid supplementation improves cerebellar abnormalities, implicating the accumulation of C27 bile acid intermediates in the cerebellar toxicity observed postnatally in these mice.

Management

There is no specific therapy for these disorders and any therapy should address the fact that features begin prenatally and therefore should be tailored to disease severity. Systematic evaluation of patients with ZSS should include visual and auditory evoked response testing, ophthalmological examination, brain MRI, renal and cardiac ultrasound, and evaluation of hepatic enzymes and clotting factors (see Table 1). ACTH and cortisol levels should be measured to determine the extent of adrenal dysfunction. Infants may require a gastrostomy tube to ensure appropriate nourishment. Water-miscible preparations of fat-soluble vitamins, including vitamin K, should be provided. Dietary interventions have included restriction of metabolites that accumulate and replacement of those that are deficient. Reports of these approaches in patients are mostly anecdotal and have not been systematically studied. For example, repletion of erythrocyte plasmalogen levels has been achieved by dietary administration of plasmalogen precursors such as batyl alcohol, but this did not improve the clinical status in the few patients in which it was tried. Oral bile acid administration improved hepatobiliary function in an infant with ZS. More recently, patients have received DHA supplementation, but the clinical benefit to this therap–y has not yet been proved. An IRD patient received an orthotopic liver transplant, resulting in improvement in biochemical parameters, but clinical improvement is not yet known. Due to the benefit of dietary phytanic acid restriction in ARD, this is often prescribed for surviving PBD patients.

Conclusion

The underlying gene defects in peroxisomal disorders have only partially revealed insight into the neuropathogenesis of X-ALD and PBDs. Where there is a strong genotype–phenotype correlation in the PBDs, there is none in X-ALD. The elevations in VLCFAs seen in these disorders have been implicated both in the process of inflammatory demyelination and in cortical migration defects. The murine models will be useful tools to sort out both AMN and the cortical migration defects in ZS. The presence of ultrastructural alterations of mitochondria in axons and neuronal bodies in both disorders suggests that the cross-talk between mitochondria and peroxisomes might contribute to the pathogenesis. As the complex functions of the peroxisome are still under investigation, any therapeutic success will depend on a clear understanding of the altered proteins and their exact function. This understanding will be expedited by the combined investigations in patients, cellular, and animal models.

See also: Demyelinating Diseases; Demyelination and Demyelinating Antibodies; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Peroxisomes: Organization and Transport in Neurons.

Further Reading

Braverman N, Chen L, Lin P, et al. (2000) PEX7 gene structure, alternative transcripts, and evidence for a founder haplotype for the frequent RCDP allele, L292ter. *Genomics* 63(2): 181–192.

Braverman N, Chen L, Lin P, et al. (2002) Mutation analysis of PEX7 in 60 proands with rhizomelic chondrodysplasia punctata and functional correlations of genotype with phenotype. *Human Mutation* 20(4): 284–297.

Eichler FS, Barker PB, Cox C, et al. (2002) Proton MR spectroscopic imaging predicts lesion progression on MRI in X-linked adrenoleukodystrophy. *Neurology* 58(6): 901–907.

Hubbard WC, Moser AB, Tortorelli S, et al. (2006) Combined liquid chromatography–tandem mass spectrometry as an analytical method for high throughput screening for X-linked adrenoleukodystrophy and other peroxisomal disorders: Preliminary findings. *Molecular Genetics and Metabolism* 89(1-2): 185–187.

Ito M, Blumberg BM, Mock DJ, et al. (2001) Potential environmental and host participants in the early white matter lesion of adreno-leukodystrophy: Morphologic evidence for CD8 cytotoxic T cells, cytolysis of oligodendrocytes, and CD1-mediated lipid antigen presentation. *Journal of Neuropathology & Experimental Neurology* 60(10): 1004–1019.

Moser AB, Rasmussen M, Naidu S, et al. (1995) Phenotype of patients with peroxisomal disorders subdivided into sixteen complementation groups. *Journal of Pediatrics* 127(1): 13–22.

Moser H, Smith KD, Watkins PA, et al. (2000) X-linked adrenoleukodystrophy. In: Scriver C, Beaudet AL, Sly WS, et al. (eds.) *The Metabolic and Molecular Bases of Inherited Disease*, pp. 3257–3301. New York: McGraw-Hill.

Moser HW, Raymond GV, Lu S-E, et al. (2005) Follow-up of 89 asymptomatic patients with adrenoleukodystrophy treated with Lorenzo's oil. *Archives of Neurology* 62(7): 1073–1080.

Peters C and Steward CG (2003) Hematopoietic cell transplantation for inherited metabolic diseases: An overview of outcomes and practice guidelines. *Bone Marrow Transplantation* 31(4): 229–239.

Powers JM and Moser HW (1998) Peroxisomal disorders: Genotype, phenotype, major neuropathologic lesions, and pathogenesis. *Brain Pathology* 8(1): 101–120.

Pujol A, Hindelgang C, Callizot N, et al. (2002) Late onset neurological phenotype of the X-ALD gene inactivation in mice: A mouse model for adrenomyeloneuropathy. *Human Molecular Genetics* 11(5): 499–505.

Theda C, Moser AB, Powers JM, et al. (1992) Phospholipids in X-linked adrenoleukodystrophy white matter: Fatty acid abnormalities before the onset of demyelination. *Journal of Neurological Science* 110(1-2): 195–204.

Van den Brink DM, Brites P, Haasjes J, et al. (2003) Identification of PEX7 as the second gene involved in Refsum desease. *American Journal of Human Genetics* 72(2): 471–477.

Van Geel BM, Bezman L, Loes DJ, et al. (2001) Evolution of phenotypes in adult male patients with X-linked adrenoleukodystrophy. *Annals of Neurology* 49(2): 186–194.

Wanders RJ and Waterham HR (2006) Biochemistry of mammalian peroxisomes revisited. *Annual Review of Biochemistry* 75: 295–332.

Weller S, Gould SJ, and Valle D (2003) Peroxisome biogenesis disorders. *Annual Review of Genomics and Human Genetics* 4: 165–211.

Steinberg S, Chen L, Wei L, et al. (2004) The PEX gene screen: Molecular diagnosis of peroxisome biogenesis disorders in the Zellweger syndrome spectrum. *Molecular Genetics and Metabolism* 83: 252–263.

Steinberg SJ, Dodt G, Raymond G, et al. (2006) Peroxisome biogenesis disorders. *Biochimica et Biophysica Acta* 1763(12): 1733–1748.

Relevant Websites

http://www.x-ald.nl – X-Linked Adrenoleukodystrophy Database.
http://www.dbpex.org – db PEX, PEX Gene Database.

Peroxisomes: Organization and Transport in Neurons

N Rayapuram and S Subramani, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Role of Peroxisomes in Neurons

The potential role of peroxisomes in neurons is not well understood, except for the metabolism of very long-chain fatty acids (VLCFA). In human patients and in mice with peroxisome biogenesis disorders, neuronal migration is clearly affected. It is unclear whether this migration defect requires the function of peroxisomes in neuronal tissue or whether peroxisomes in nonneuronal tissues can provide this function. Experiments in mice show that the tissue-specific restoration of peroxisomes in the liver of mice deficient in peroxisome biogenesis does partially correct the neuronal migration deficits in the brain, but the restoration of peroxisomes in both tissues, rather than in only one, is more effective in restoring neuronal migration. It is also clear that a β-oxidation deficiency alone cannot cause neuronal migration defects, so other metabolic functions of peroxisomes may be involved. Mice lacking Pex11β, which is involved in peroxisome division, have only mild to no metabolic defects with respect to fatty acid β-oxidation and plasmalogen synthesis, but they still exhibit neuronal migration defects. These results emphasize that we still do not understand how or why peroxisomes affect neuronal migration and whether this is a direct or indirect effect.

Peroxisomes are present in neural and astrocytic processes. The presence of peroxisomes in active growth cones with well-developed filopodia, varicosities, and neural terminal-like (presynaptic axon terminal) structures suggests their importance in the development and/or maintenance of neurites. Whereas many peroxisomes are found in long neurites with growth cones, only a small number exist in short neurites, leading to the suggestion that the short neurites are probably at rest since no growth cones have been observed. Therefore, it was postulated that the presence of peroxisomes in neurites helps their extending activity.

Peroxisome Movement in Mammalian Cells

Based on their motility, peroxisomes fall into one of two groups. The first, comprising a majority (90–95%) of the peroxisomes in African green monkey kidney (CV1), Chinese hamster ovary (CHO), HepG2 cells, and neurites, exhibits randomly oriented, low-velocity movements (average velocity of 0.024 $\mu m\,s^{-1}$) without net organelle translocation in a distinct direction. This movement is independent of actin filaments and microtubules.

The second category of peroxisomes, representing 5–10% of these organelles, exhibits fast movement (with average velocities of 0.26 $\mu m\,s^{-1}$, peak velocities of up to 0.75 $\mu m\,s^{-1}$, and sustained directional velocities up to 0.45 $\mu m\,s^{-1}$ over distances of 11.5 μm), which is saltatory (discontinuous and erratic) and bidirectional (toward and away from the cell center) and results in the displacement of the organelle over distances of 10 μm. The bidirectional movement suggests the involvement of motor proteins of the dynein and/or kinesin families. Peroxisomes belonging to these two populations can spontaneously interchange and behave like the other. The fast movement, but not the slow kind, depends on energy in the form of ATP hydrolysis. Apart from these two movements, peroxisomes can also exist in a motionless 'frozen' state upon complete depletion of metabolic energy.

Whereas the slow movement of peroxisomes is not affected by actin-depolymerizing agents, such as cytochalasins or anticytoskeletal drugs, microtubule-destabilizing agents, such as nocodazole, vinblastine, and demecolcine, completely abolish the fast, saltatory movement indicative of the involvement of microtubules in this process. Peroxisomes are often found in close proximity with microtubules when observed by triple confocal imaging of CV1 cells expressing GFP in peroxisomes, propidium iodide staining of the nucleus, and the immunolabeling of microtubules using anti-β-tubulin antibodies (Figure 1). In contrast to the microtubule-based movement of peroxisomes in mammalian cells, the movement of this organelle in fungi and plants depends on actin and myosin.

Peroxisome–Microtubule Binding *in Vitro*

Although the *in vivo* imaging of peroxisomes and microtubules has exhibited a close association between the two, formal evidence of peroxisome binding to microtubules has come from *in vitro* binding studies. A novel *in vitro* assay was developed in which purified rat liver peroxisomes associate with microtubules coated onto a microtiter plate (Figure 2). The binding was visualized using a confocal laser-scanning microscope and quantified by immunoblotting. The binding was time, temperature, and pH dependent.

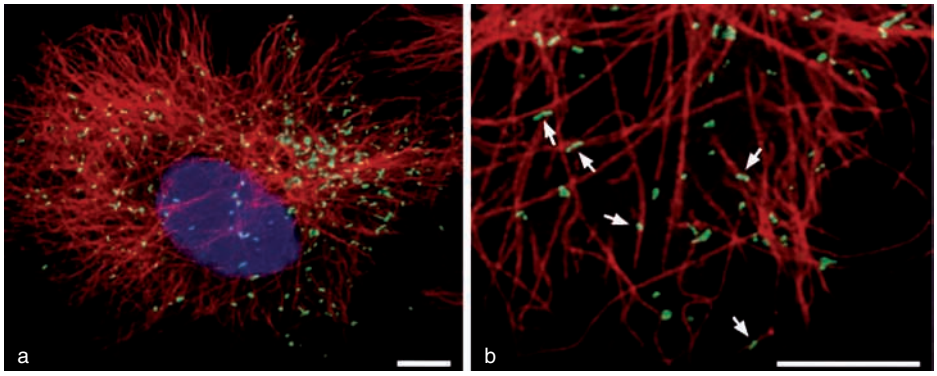

Figure 1 Association of peroxisomes with microtubules. Triple fluorescence confocal imaging of CV1 cells expressing GFP in peroxisomes (green), propidium iodide staining of the nucleus (pseudo-colored blue), and immunolabeling of microtubules using mouse anti-tubulin followed by goat anti-mouse IgG-CY5 (pseudo-colored red). (a) A single cell showing the overall distribution of peroxisomes relative to microtubules. (b) An enlarged portion of the image of the same cell showing the close association of many individual peroxisomes (arrows) with microtubules. Scale bar = 10 μm. Reprinted from Wiemer EA, Wenzel T, Deerinck TJ, Ellisman MH, and Subramani S (1997) Visualization of the peroxisomal compartment in living mammalian cells: Dynamic behavior and association with microtubules. *Journal of Cell Biology* 136: 71–80.

Further investigation into this association showed that addition of ATP increased binding, whereas ATP depletion reduced binding. Stripping the peroxisome surface of peripheral membrane proteins by treatment with KCl reduced the binding, which was restored by the cytosol. However, treatment of either the peroxisomes or the cytosol with proteases or *N*-ethylmaleimide decreased binding and this was irreversible. These experiments clearly show the involvement of a peroxisomal membrane protein, probably a peroxisomal cytoplasmic linker protein (CLIP) and some cytosolic factor(s) for binding of peroxisomes to microtubules (Figure 3). When a motor-enriched fraction from chick embryo brain was added to the peroxisomes bound to microtubules, movement of the organelle occurred in an ATP-dependent manner, suggesting the involvement of microtubule-dependent motor proteins in peroxisome movement.

Motors Involved in Peroxisome Movement

Microtubule-dependent motor proteins, namely conventional kinesin (kinesin-1) and cytoplasmic dynein, are implicated in organelle trafficking and cell division. The use of RNAi to selectively inhibit several kinesins (kinesin-1, two members of the kinesin-2 family, three members of the kinesin-3 family, and a member of kinesin-14 or the C-terminal kinesin family) and cytoplasmic dynein in *Drosophila* S2 cells showed that only kinesin-1 and dynein heavy-chain motors are responsible for the movement of peroxisomes. The plus (+)-end directed kinesins carry organelles to the cell periphery, whereas the minus (−)-end directed dyneins ferry organelles back toward the center of the cell. The movement of peroxisomes in neurites of a subline of PC12 pheochromocytoma (PC12D) cells and *Drosophila* S2 cells was studied by time-lapse high-resolution fluorescence microscopy and by fluorescence imaging with one nanometer accuracy (FIONA), respectively, and found to occur in both the anterograde (cell center to cell periphery) and retrograde (cell periphery to cell center) directions (Figure 4). Overexpression of the dynamitin subunit of the dynein activator, dynactin, inhibited peroxisome movement in monkey kidney (COS7) cells. The tau protein, a neuronal microtubule-associated protein, inhibits the kinesin-dependent movement of peroxisomes in primary cortical neurons, retinal ganglion cells, and neuroblastoma cells and makes these cells more susceptible to oxidative stress.

Given the bidirectional movement of peroxisomes, the important question regarding whether the kinesin and dynein motors act simultaneously or sequentially on peroxisomes was addressed experimentally in *Drosophila* S2 cells. The conclusion from these studies was that the two kinds of motors did not act against each other simultaneously but, rather, up to 10 or 11 molecules of one or the other type of motor work together to produce up to 10 times the *in vitro* speed of peroxisome movement. These studies explain why the movement of peroxisomes is saltatory, fast, and occurs over distances that are far greater than the average step size of a single motor molecule.

Regulation of Peroxisome Motility

Peroxisomal motility is sensitive to the extracellular environment and probably to receptor-mediated

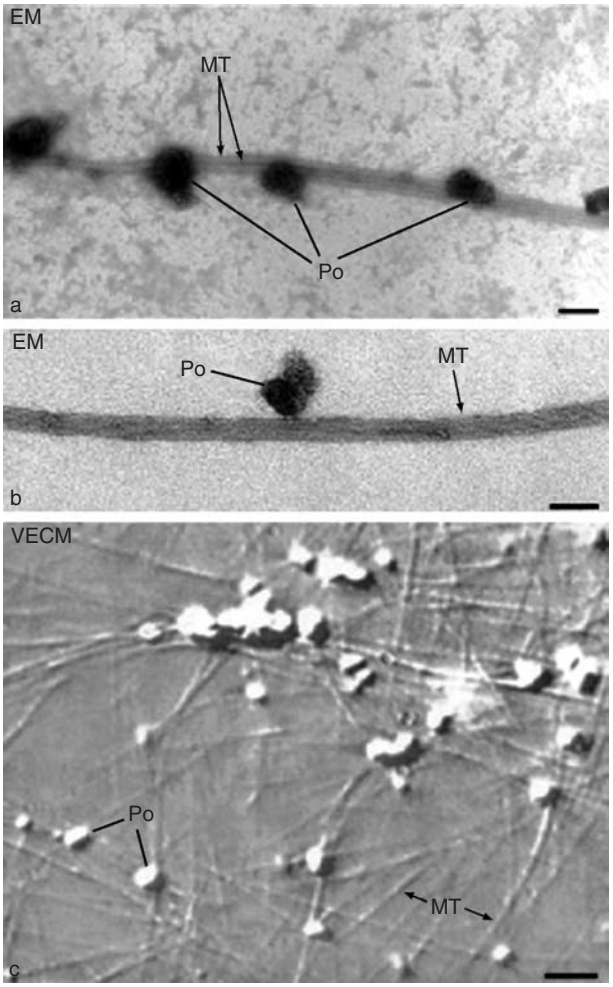

Figure 2 Electron micrographs of peroxisomes bound to microtubules *in vitro*. Highly purified rat liver peroxisomes (Po) are specifically associating with taxol-stabilized microtubules from bovine brain (MT). (a and b) Electron micrographs (EM) of negatively stained suspended peroxisome–microtubule complexes with 2% phosphotungstic acid. (c) Video-enhanced contrast microscopy (VECM) of microtubules and peroxisomes. Scale bars = 0.2 μm (a), 0.05 μm (b), and 1 μm (c). Reprinted from Schrader M, Thiemann M, and Fahimi HD (2003) Peroxisomal motility and interaction with microtubules. *Microscopy Research and Technique* 61: 171–178.

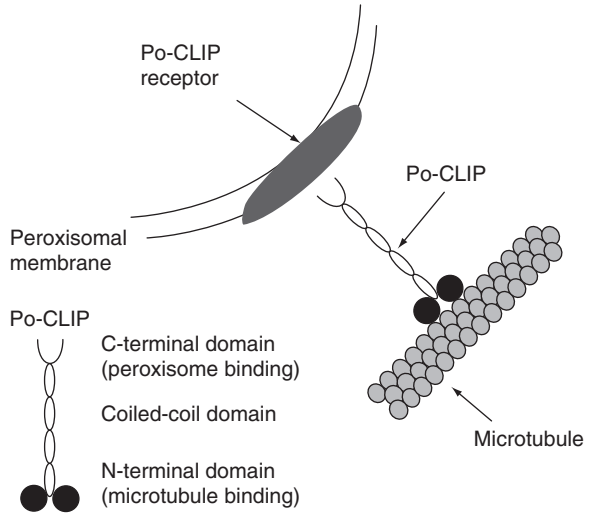

Figure 3 Schematic model for the binding of peroxisomes to microtubules. Possible structure of the putative peroxisomal cytoplasmic linker protein (Po-CLIP) according to the model proposed for CLIP-170 and the presumed association of Po-CLIP to microtubules and peroxisomes. Reprinted from Thiemann M, Schrader M, Volkl A, Baumgart E, and Fahimi HD (2000) Interaction of peroxisomes with microtubules. *In vitro* studies using a novel peroxisome microtubule binding assay. *European Journal of Biochemistry* 267: 6264–6275.

regulation in CHO and human umbilical vein endothelium cells. For example, the combination of ATPγS and lysophosphatidic acid (LPA) blocks peroxisome movement probably by action via G-protein-coupled ATP and LPA receptors.

Peroxisome motility in CHO cells is regulated by a GTP-binding protein. Upon microinjection of GTPγS, a nonhydrolyzable GTP analog, there is a complete loss of peroxisome motility, although the effect is not immediate. A similar effect is elicited by activators of heterotrimeric G-proteins, such as aluminum fluoride and mastoparan.

Pertussis toxin, which antagonizes the action of the G_i/G_o class of proteins, prevents the action of GTPγS in inhibiting peroxisome motility, suggesting that this class of G-proteins is involved in the regulation of peroxisome movement. Additionally, inhibition of phospholipase C, classical protein kinase C isoforms, mitogen-activated kinase kinase, or phospholipase A2 (PLA2) affects the ability of this signaling pathway to inhibit peroxisome motility, so these could be signaling molecules acting downstream of the G-proteins. The PLA2 signaling cascade is believed to act via arachidonic acid, whose release inhibits peroxisome movement.

Although much of this information on the regulation of peroxisome movement comes from non-neuronal cells, the principles are likely to be conserved because microtubule-based peroxisome movement is common to mammals and the signaling cascades controlling peroxisome movement are widely distributed in mammalian tissues.

Peroxisome Morphology

Although peroxisomes appear spherical in cross-section under the microscope, they are a highly dynamic subcellular compartment. Peroxisomes often exist in the form of an interconnected network or reticulum. They assume morphologically distinct

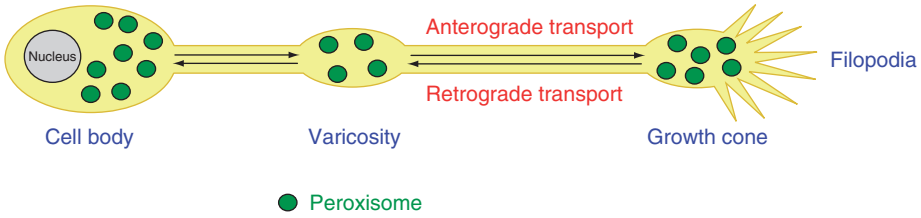

Figure 4 A model of peroxisome transport from cell body to growth cone. Peroxisomes, which are formed in the cell body, move to the growth cone via the varicosity by anterograde transport. Peroxisomes play an important role at the growth cone and then return to the cell body by retrograde transport. During transportation, peroxisomes exhibit saltatory movement and/or are oscillating repeatedly. Reprinted from Ishikawa T, Kawai C, Sano M, and Minatogawa Y (2001) Peroxisomes exist in growth cones and move anterogradely and retrogradely in neurites of PC12D cells. *Experimental Cell Research* 266: 260–269.

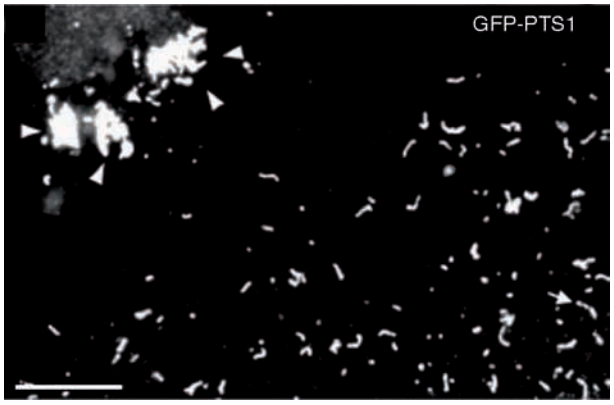

Figure 5 GFP-PTS1 is targeted to peroxisomal aggregates in COS-7 cells, which contain elongated peroxisomes forming reticular structures. In some COS-7 cells, tubular peroxisomes and/or large peroxisomal aggregates (arrowheads) can be visualized by GFP-PTS1. An elongated peroxisome with a segmented appearance is visible (arrow). Scale bar = 10 μm. Reprinted from Schrader M, King SJ, Stroh TA, and Schroer TA (2000) Real time imaging reveals a peroxisomal reticulum in living cells. *Journal of Cell Science* 113: 3663–3671.

shapes, including tubular structures that fuse to form a reticulum (Figure 5), when cultured cells are treated with defined growth factors (e.g., nerve growth factor) or with polyunsaturated fatty acids, free radicals, or ultraviolet (UV) irradiation. Even tissues engaged in the synthesis of special kinds of lipids exhibit this behavior. Both spherical and tubular peroxisomes move inside cells with similar velocities. Elongated peroxisomes could change shape while moving by shortening in length, collapsing into round structures, or elongating, and they also interacted with other peroxisomes.

Tubular peroxisomes of up to 5 μm can be induced in 40–50% of HepG2 cells by UV irradiation, and this induction occurs via a tyrosine kinase signal transduction pathway because it is sensitive to K 252b, an inhibitor of tyrosine kinase. Hydrogen peroxide also induces tubulation, but hypoxia reduces tubule formation. The overexpression of Pex11β results in tubulation and fission of elongated peroxisomes.

The exact function and the precise mechanism of formation of these morphologically distinct peroxisomal compartments are still under investigation, but tubulation and fission of peroxisomes have been suggested to contribute to peroxisome proliferation and to the response of cells to reactive oxygen species.

Failure of Peroxisome Biogenesis and/or Movement Is Observed in Disease

Whereas peroxisomes are abundant in normal hepatocytes and fibroblasts, their numbers are drastically decreased in peroxisome biogenesis disorders. These include the cerebrohepatorenal or Zellweger syndrome, rhizomelic chondrodysplasia punctata (RCDP), and neonatal adrenoleukodystrophy. Although diseased cells lack detectable peroxisomes when analyzed using antibodies to peroxisomal matrix markers, remnant membranous structures termed 'ghosts' with peroxisomal membrane proteins (PMPs) are present. The presence of these ghosts with PMPs is explained by the fact that the signals and machineries involved in peroxisomal matrix and membrane protein import are different, and most (but not all) of these biogenesis disorders affect only matrix protein import.

Peroxisomes are assumed to play a crucial role in the formation of cortex, cerebellum, and the inferior olivary nucleus because there are severe morphogenic defects in these tissues in peroxisome biogenesis disorders. Analysis of peroxisome-deficient fetuses with focus on neurodevelopment showed defects typical of Zellweger syndrome, such as a defect in the neuronal migration in the neocortex, delay in neuronal maturation, increased apoptosis in the cortical plate, and structural abnormalities in the inferior olivary nucleus.

Human patients lacking the multifunctional protein 2 (MFP2), a D-bifunctional protein with enoyl-CoA hydratase and 3-hydroxy-acyl-CoA dehydrogenase

activities, necessary for peroxisomal β-oxidation of bile acid intermediates and pristanic acid, were found to have neuronal migration defects. MFP2 is involved in the metabolism of VLCFAs. The accumulation of VLCFAs as a result of peroxisomal *β*-oxidation deficiency, depletion of plasmalogens, or reduced content of the polyunsaturated fatty acid docosahexaenoic acid (DHA) play a major role in impairment of neuronal migration. However, in mice lacking MFP2, the neuronal migration defect was not obvious.

In fibroblasts from human patients with PEX1 or D-bifunctional enzyme deficiencies, peroxisome abundance, movement, and association with microtubules are affected. This is also true in mice lacking PEX13. Perhaps these are deficient in the recruitment of a protein that links peroxisomes with microtubules.

Peroxisome Biogenesis Disorders That Affect Neuronal Migration

Peroxisome biogenesis disorders result from defects in *PEX* genes, which code for proteins involved in peroxisome biogenesis and protein import. An important pathological feature of peroxisome biogenesis disorders is defective neuronal migration.

The following are peroxisome biogenesis disorders:

1. Zellweger syndrome or cerebro-hepato-renal syndrome results in abnormal central nervous system neuronal migration due to the lack of functional peroxisomes. Zellweger syndrome phenotype is caused by mutations in any of several *PEX* genes, such as *PEX1*, *PEX2*, *PEX3*, *PEX5*, *PEX6*, *PEX10*, *PEX12*, *PEX13*, *PEX16*, *PEX19*, or *PEX26*. The most striking pathological abnormalities are apparent in the brain. As indicated previously, most of these affect peroxisomal matrix protein import, but a few, such as *PEX3*, *PEX16*, and *PEX19*, affect both PMP and matrix protein import.
2. Neonatal adrenoleukodystrophy is caused by milder mutations in the genes affected in Zellweger syndrome.
3. Infantile Refsum disease (IRD) is the mildest form of impairment among peroxisomal biogenesis disorders. It is also called infantile phytanic acid storage disorder. The biochemical abnormalities are not restricted to phytanic acid but also include accumulation of VLCFAs, di- and trihydroxycholestanoic acid, and pipecolic acid. Mutations in *PEX1*, *PEX2*, or *PEX26* can cause IRD.
4. RCDP: Biochemically, RCDP patients have subnormal levels of red cell plasmalogens and progressive accumulation of phytanic acid. Mutations in the PTS2 receptor, *PEX7*, cause RCDP.
5. Zellweger-like syndrome.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Microtubule Associated Proteins in Neurons; Microtubules: Organization and Function in Neurons; Peroxisomal Disorders and Neurological Disease.

Further Reading

Baes M, Gressens P, Huyghe S, et al. (2002) The neuronal migration defect in mice with Zellweger syndrome (Pex5 knockout) is not caused by the inactivity of peroxisomal beta-oxidation. *Journal of Neuropathology & Experimental Neurology* 61: 368–374.

Brites P, Motley AM, Gressens P, et al. (2003) Impaired neuronal migration and endochondral ossification in Pex7 knockout mice: A model for rhizomelic chondrodysplasia punctata. *Human Molecular Genetics* 12: 2255–2267.

Cole NB and Lippincott-Schwartz J (1995) Organization of organelles and membrane traffic by microtubules. *Current Opinion in Cell Biology* 7: 55–64.

Fagarasanu A, Fagarasanu M, Eitzen GA, Aitchison JD, and Rachubinski RA (2006) The peroxisomal membrane protein Inp2p is the peroxisome-specific receptor for the myosin V motor Myo2p of *Saccharomyces cerevisiae*. *Developmental Cell* 10: 587–600.

Faust PL (2003) Abnormal cerebellar histogenesis in PEX2 Zellweger mice reflects multiple neuronal defects induced by peroxisome deficiency. *Journal of Comparative Neurology* 461: 394–413.

Faust PL, Banka D, Siriratsivawong R, Ng VG, and Wikander TM (2005) Peroxisome biogenesis disorders: The role of peroxisomes and metabolic dysfunction in developing brain. *Journal of Inherited Metabolic Disease* 28: 369–383.

Gressens P (2000) Mechanisms and disturbances of neuronal migration. *Pediatric Research* 48: 725–730.

Gressens P (2006) Pathogenesis of migration disorders. *Current Opinion in Neurology* 19: 135–140.

Huber C, Saffrich R, Anton M, et al. (1997) A heterotrimeric G protein phospholipase A2 signaling cascade is involved in the regulation of peroxisomal motility in CHO cells. *Journal of Cell Science* 110: 2955–2968.

Huber CM, Saffrich R, Ansorge W, and Just WW (1999) Receptor-mediated regulation of peroxisomal motility in CHO and endothelial cells. *EMBO Journal* 18: 5476–5485.

Ilieva HS, Nagano I, Murakami T, et al. (2003) Age-related changes in peroxisomal membrane protein 70 and superoxide dismutase 1 in transgenic G93A mice. *Neurological Research* 25: 423–426.

Ishikawa T, Kawai C, Sano M, and Minatogawa Y (2001) Peroxisomes exist in growth cones and move anterogradely and retrogradely in neurites of PC12D cells. *Experimental Cell Research* 266: 260–269.

Janssen A, Gressens P, Grabenbauer M, et al. (2003) Neuronal migration depends on intact peroxisomal function in brain and in extraneuronal tissues. *Journal of Neuroscience* 23: 9732–9741.

Kural C, Kim H, Syed S, et al. (2005) Kinesin and dynein move a peroxisome *in vivo*: A tug-of-war or coordinated movement? *Science* 308: 1469–1472.

Li X, Baumgart E, Morrell JC, et al. (2002) PEX11*β* deficiency is lethal and impairs neuronal migration but does not abrogate peroxisome function. *Molecular and Cellular Biology* 22: 4358–4365.

Mathur J, Mathur N, and Hulskamp M (2002) Simultaneous visualization of peroxisomes and cytoskeletal elements reveals actin and not microtubule-based peroxisome motility in plants. *Plant Physiology* 128: 1031–1045.

Maxwell M, Bjorkman J, Nguyen T, et al. (2003) Pex13 inactivation in the mouse disrupts peroxisome biogenesis and leads to a Zellweger syndrome phenotype. *Molecular and Cellular Biology* 23: 5947–5957.

Maxwell MA, Allen T, Solly PB, et al. (2002) Novel PEX1 mutations and genotype–phenotype correlations in Australasian peroxisome biogenesis disorder patients. *Human Mutation* 20: 342–351.

Nagase T, Shimozawa N, Takemoto Y, et al. (2004) Peroxisomal localization in the developing mouse cerebellum: Implications for neuronal abnormalities related to deficiencies in peroxisomes. *Biochimica et Biophysica Acta* 1671: 26–33.

Nguyen T, Bjorkman J, Paton BC, and Crane DI (2006) Failure of microtubule mediated peroxisome division and trafficking in disorders with reduced peroxisome abundance. *Journal of Cell Science* 119: 636–645.

Rapp S, Saffrich R, Anton M, et al. (1996) Microtubule-based peroxisome movement. *Journal of Cell Science* 109: 837–849.

Schrader M (2001) Tubulo-reticular clusters of peroxisomes in living COS-7 cells: Dynamic behavior and association with lipid droplets. *Journal of Histochemistry and Cytochemistry* 49: 1421–1429.

Schrader M and Fahimi HD (2004) Mammalian peroxisomes and reactive oxygen species. *Histochemistry and Cell Biology* 122: 383–393.

Schrader M, King SJ, Stroh TA, and Schroer TA (2000) Real time imaging reveals a peroxisomal reticulum in living cells. *Journal of Cell Science* 113: 3663–3671.

Schrader M, Thiemann M, and Fahimi HD (2003) Peroxisomal motility and interaction with microtubules. *Microscopy Research and Technique* 61: 171–178.

Schrader M, Wodopia R, and Fahimi HD (1999) Induction of tubular peroxisomes by UV irradiation and reactive oxygen species in HepG2 cells. *Journal of Histochemistry and Cytochemistry* 47: 1141–1148.

Stamer K, Vogel R, Thies E, Mandelkow E, and Mandelkow EM (2002) Tau blocks traffic of organelles, neurofilaments, and APP vesicles in neurons and enhances oxidative stress. *Journal of Cell Biology* 156: 1051–1063.

Steinberg S, Chen L, Wei L, et al. (2004) The PEX GeneScreen: Molecular diagnosis of peroxisome biogenesis disorders in the Zellweger syndrome spectrum. *Molecular Genetics and Metabolism* 83: 252–263.

Subramani S (1998) Components involved in peroxisome import, biogenesis, proliferation, turnover, and movement. *Physiological Reviews* 78: 171–188.

Suzuki Y, Shimozawa N, Imamura A, et al. (2001) Clinical, biochemical and genetic aspects and neuronal migration in peroxisome biogenesis disorders. *Journal of Inherited Metabolic Disease* 24: 151–165.

Thiemann M, Schrader M, Volkl A, Baumgart E, and Fahimi HD (2000) Interaction of peroxisomes with microtubules *in vitro* studies using a novel peroxisome microtubule binding assay. *European Journal of Biochemistry* 267: 6264–6275.

Wanders RJ (2004) Metabolic and molecular basis of peroxisomal disorders: A review. *American Journal of Medical Genetics A* 126: 355–375.

Wiemer EA, Wenzel T, Deerinck TJ, Ellisman MH, and Subramani S (1997) Visualization of the peroxisomal compartment in living mammalian cells: Dynamic behavior and association with microtubules. *Journal of Cell Biology* 136: 71–80.

Pharmacology of Fear Extinction

M Davis and K M Myers, Emory University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The mechanisms of fear inhibition are attracting increasing interest. In the laboratory, fear inhibition most often is studied through a procedure in which a previously fear-conditioned animal is exposed to a fear-eliciting cue in the absence of any aversive event. This procedure results in a decline in conditioned fear responses that is attributed to a process called fear extinction. Extensive empirical work by psychologists has revealed basic behavioral characteristics of extinction, and theoretical accounts have emphasized extinction as a form of inhibitory learning as opposed to an erasure of acquired fear. Guided by this work, neuroscientists have begun to dissect the neural mechanisms involved, including the regions in which extinction-related plasticity occurs and the cellular and molecular processes that are engaged. Several careful reviews on both the behavioral aspects of extinction and the brain areas that are critical for extinction are available, but this article reviews only the pharmacology of extinction. The literature on neurotransmitter involvement in fear extinction is built on the assumption that extinction is a form of inhibitory learning which leaves acquisition-related plasticity relatively intact. Therefore, the same neurotransmitter systems involved in fear conditioning might be expected also to be involved in fear extinction, although certain differences might stem from the excitatory versus inhibitory nature of these two forms of learning.

Before proceeding, it should be noted that the term 'extinction' is commonly used in more than one sense, to refer to (1) the experimental procedure used to produce a decrement in the fear response, (2) the decremental effect of this procedure upon the fear response, which can be measured both at the time of the experimental procedure and at a later time, or (3) the hypothesized associative or cellular process responsible for that effect. For the purposes of this article, we define the experimental procedure as extinction training, the decrement in the fear response measured during extinction training as within-session extinction, and the decrement measured at some interval after extinction training as extinction retention. The term extinction is reserved for the theoretical process underlying the loss of the fear response. Occasionally we refer to short-term extinction and long-term extinction; these terms correspond to stages of extinction memory similar to those defined for other forms of learning.

γ-Aminobutyric Acid

FG-7142, a compound that functionally decreases γ-aminobutyric acid (GABA) transmission, the major inhibitory neurotransmitter in the brain, dose-dependently impaired within-session extinction and, when administered prior to test, blocked extinction retention in the context of extinction training, but had no effect on performance in a novel context. The blockade of extinction could not be explained by state dependency.

Gephryin, a protein that contributes to the regulation of GABAergic neurotransmission by clustering $GABA_A$ receptors at the synapse, is upregulated in the basolateral amygdala (BLA) following extinction training based on mRNA, protein levels, and receptor binding. GABA also has been implicated in the consolidation of extinction. Post-extinction training systemic administration of the GABA antagonist picrotoxin enhanced extinction retention, as assessed with active avoidance. Intra-BLA infusions of the GABA receptor antagonist bicuculline immediately but not 3 h after extinction training facilitated extinction of contextual freezing, and pre-extinction training systemic administration of diazepam, which enhances GABA transmission, impaired extinction retention, although state dependency was not ruled out.

Glutamate

N-Methyl-D-aspartate (NMDA) receptor antagonists, a glutamate receptor subtype, block extinction of freezing and fear-potentiated startle when administered prior to extinction training either systemically or directly into the BLA, or step-down avoidance extinction when administered into CA1 of the hippocampus. In all measures the effects were dose dependent and not due to state dependency. These effects may result from a blockade of extinction consolidation. Thus, pre-extinction training administration of the NMDA receptor antagonist (±)-3-(2-carboxypiperazin-4-yl)-propyl-1-phosphonic acid (CPP) blocked extinction retention measured 24 but not 48 h later. However, when CPP was administered prior to extinction training and again 24 h later, without further extinction training, extinction retention tested drug-free 48 h after extinction training was again

impaired. This suggests an NMDA receptor-dependent consolidation process that can be delayed when NMDA receptors are not functional within the normal consolidation time frame. However, using step-down avoidance, it was found that intrahippocampal NMDA antagonist infusion immediately following the first of four daily extinction exposures produced a lasting impairment, which persisted even with additional extinction training without infusions. A similar impairment with pre-extinction training infusions into entorhinal cortex was also seen.

NMDA-dependent extinction consolidation also is seen with the NMDA receptor partial agonist D-cycloserine (DCS). DCS was initially shown to facilitate extinction of fear-potentiated startle when administered systemically or directly into the BLA prior to extinction training. Later it was found that DCS facilitated extinction of freezing when administered prior to or up to 120 min after extinction training, suggesting modulation of consolidation. DCS-facilitated extinction is less susceptible to reinstatement and generalized to a comparably trained but nonextinguished cue, and may only work if some within-session extinction occurs. The effect of DCS appears to be mediated through mitogen-activated protein kinase (MAPK) and phosphatidylinositol 3-kinase (PI3K) in the BLA and requires both transcription and protein synthesis.

Dopamine

Although fear conditioning is generally facilitated by dopamine, extinction seems to be disrupted. Systemic administration of cocaine prior to extinction training had no effect on within-session extinction of fear-potentiated startle but impaired extinction retention, an effect not due to state dependency. A similar disruption following administration of cocaine, amphetamine, or the D1 agonist SKF 38393 was also reported. Cocaine and SKF 38393 also disrupted extinction retention when administered prior to test in animals extinguished drug-free. Systemic administration of the D2 receptor agonist quinpirole prior to extinction training impaired extinction retention, although a possible state-dependent effect was not ruled out.

In mice, facilitation of extinction retention by pre-extinction training systemic administration of sulpiride, a D2 receptor antagonist, has been reported. However, using inhibitory avoidance, it was found that D1 receptor knockout mice were impaired in extinction, although ceiling effects in acquisition make the results difficult to interpret. None of this work involved targeted manipulations, so we do not know the locus of these effects.

Acetylcholine

Systemic administration of scopolamine, a muscarinic acetylcholine receptor antagonist, prior to the seventh of eight weekly extinction exposures in a shuttle box impaired extinction retention via a central effect. This effect was dose dependent and observable as long as 9 months after initial training.

Opioids

A critical question regarding extinction is 'how is nonreinforcement detected?' That is, what is the 'error signal' that results from detection of the nonoccurrence of an expected unconditioned stimulus (US)? Most error signal theoretical accounts assume that the magnitude of the associative change produced by a US is inversely related to the degree to which the conditioned stimulus (CS) predicts the US. Hence, the discrepancy between the 'actual' and 'predicted' US drives learning: when the discrepancy is positive (i.e., the US is not fully predicted), excitatory learning occurs; when the discrepancy is negative (i.e., the US is overpredicted), inhibitory learning occurs; and when the discrepancy is zero (i.e., the US is perfectly predicted), no learning occurs. Neural theories propose that as CS learning occurs, it inhibits brain areas normally activated by the US, reducing its effective impact. If the US is withheld, CS-induced inhibition occurs in the absence of any counteractive US-induced excitation, and this somehow initiates extinction-related plasticity.

It has been hypothesized that, over the course of acquisition, the CS induces release of endogenous opioids that reduce the functional intensity of the US by producing conditioned analgesia. Opiate antagonists (e.g., naloxone) given prior to CS–US pairings facilitate fear acquisition by blocking conditioned analgesia. This analgesia-based error correction hypothesis is compelling and accounts for certain difficult-to-explain observations. However, there are indications that opioid involvement in fear conditioning is not limited to the mitigation of functional US intensity. In particular, opioids modulate learning that occurs when there is no shock involved, such as extinction. In fact, earlier work suggested that extinction of active avoidance is impaired by opiate antagonists and facilitated by certain opioids.

Consistent with these older findings, it has been found that systemic administration of naloxone prior to extinction training profoundly retards the development of extinction. This effect is mediated by central, not peripheral, opioid receptors and is not a consolidation or expression effect. Infusions of naloxone or

the μ opioid receptor antagonist D-Phe-Cys-Tyr-D-Trp-Arg-Thr-Pen-Thr-NH_2(CTAP), directly into ventrolateral periaqueductal gray (vlPAG), but not the dorsal PAG, mimic the systemic effect, suggesting that μ opioid receptors are specifically involved in extinction. Extinction also is retarded by intra-vlPAG infusions of the cyclic adenosine monophosphate (cAMP) analog 8-Br-cAMP, suggesting that the effects of naloxone or CTAP on extinction are mediated by cAMP inhibition. Finally, pre-extinction training infusion of RB101(S), an inhibitor of endogenous opioid catabolizing enzymes, facilitates extinction dose dependently and site specifically when infused into vlPAG prior to extinction training.

Cannabinoids

Cannabinoid (CB) receptor (the CB1 receptor) knockout mice showed normal fear conditioning but a profound impairment of both within- and between-session extinction. This effect is mimicked in wild-type mice by giving the CB1 antagonist SR141716A prior to, but not immediately after, extinction training. WIN 55,212-2, a CB1 agonist, had no effect on extinction retention of fear-potentiated startle, whereas AM404, an inhibitor of CB reuptake and enzymatic degradation, dose dependently facilitated extinction retention measured 1 and 24 h after extinction training and impaired reinstatement. These effects were blocked by coadministration of SR141716A.

Neuropeptides

Administration of adrenocorticotropic hormone (ACTH) or vasopressin before extinction training impaired extinction of active avoidance, passive avoidance, and conditioned heart rate. Removal of the posterior pituitary, which decreases ACTH, facilitated extinction with no effect on acquisition. ACTH produced an impairment of extinction retention when administered prior to test, similar to that seen with unsignaled footshock or a shift of context.

This might have occurred by restoring an internal state or 'context' that was present during acquisition because ACTH did not restore extinguished avoidance responses in animals administered dexamethasone, a synthetic glucocorticoid that inhibits the release of endogenous ACTH, prior to acquisition and extinction training. However, ACTH impaired extinction retention even when administered 7 days prior to test, presumably long after any internal state associated with ACTH had dissipated. Although ACTH-induced stress, sensitization, or increased fear may play a role, intracerebroventricular (i.c.v.) and local microinjection of ACTH and vasopressin produced an impairment of extinction retention without concomitant autonomic responses indicative of a stress reaction. Moreover, ACTH attenuates extinction just as effectively in appetitive paradigms as in aversive ones. ACTH and vasopressin may act at the level of the parafascicular nucleus (PFN), believed to be important for processing noxious stimulation such as footshock. Given this localization, ACTH might act to restore the value of a deflated US representation, one explanation of extinction.

Glucocorticoids

In addition to modulating fear memories, glucocorticoids may also modulate the development and consolidation of fear extinction. Systemic pre-extinction training injection of metyrapone, which blocks glucocorticoid synthesis, impaired extinction retention of freezing in mice, although there was no control for state dependency. Dexamethasone, which would mimic the effects of glucocorticoids, dose dependently facilitated extinction when administered systemically pre- or immediately post-extinction training, as did intra-BLA infusions of the glucocorticoid receptor (GR) agonist RU28362. Impaired extinction by metyrapone was blocked by coadministration of a low dose of dexamethasone that was not sufficient to facilitate extinction significantly. The GR antagonist mifepristone likewise impaired extinction dose dependently when infused into the BLA prior to extinction training and blocked the facilitation of extinction by dexamethasone.

Interestingly, posttraumatic stress disorder (PTSD), characterized by a failure to extinguish traumatic memories, has been associated in some studies with abnormally low cortisol levels that may contribute to the persistence of PTSD symptoms by impairing extinction. In fact, daily administration of low doses of hydrocortisone to PTSD patients reduced certain PTSD symptoms including reexperiencing, hyperarousal, and avoidance, perhaps in part through a facilitation of extinction.

Norepinephrine

In mice, systemic administration of yohimbine, which leads to norepinephrine (NE) release, facilitated within-session context and cue extinction and extinction retention. Under conditions that lead to 'incubation' of fear, yohimbine blunted this effect and the B-noradrenergic antagonist propranolol exaggerated it. Using a contextual fear conditioning paradigm in rats, intra-BLA infusions of propranolol immediately after extinction training did not significantly affect

extinction retention but did prevent the facilitation of extinction by intra-BLA bicuculline when the two compounds were coinfused. Intra-BLA infusions of NE after extinction training dose dependently facilitated extinction retention, an effect not blocked by coinfusion of muscimol, suggesting that extinction retention effects of GABA are upstream of noradrenergic signaling.

Brain-Derived Neurotrophic Factor

A time-dependent upregulation of brain-derived neurotrophic factor (BDNF) mRNA within the BLA in extinguished rats relative to nonextinguished controls has been found. Intra-BLA virally mediated infusion of a truncated form of the TrkB receptor, which would block BDNF signaling, prior to extinction training produced a deficit in extinction retention (measured 1 or 24 h after extinction training) but not within-session extinction, suggesting a role for BDNF in the consolidation of extinction memory.

Intracellular Signaling Pathways

Voltage-Gated Calcium Channels

Systemic administration of the L-type voltage-gated calcium channel (L-VGCC) blockers nifedipine and nimodipine prior to extinction training produced a dose-dependent impairment of both within- and between-session extinction in mice. These effects were evident in both context and tone extinction and were not state dependent. Interestingly, nifedipine did not impair extinction when extinction training occurred immediately after acquisition, but was evident when extinction training was delayed by 1 or 3 h, perhaps consistent with behavioral evidence suggesting a difference in the mechanisms of short- and long-interval extinction.

Adenylyl Cyclase/cAMP

Transgenic mice overexpressing type 1 adenylyl cyclase within the forebrain had increased adenylyl cyclase activity, long-term potentiation, and phosphorylated MAPK and cAMP response element binding protein (CREB), a transcription factor, within the hippocampus. In a Pavlovian fear conditioning paradigm, the transgenic mice acquired both tone and context fear normally, but were delayed in context fear extinction relative to wild-type mice. Despite the apparent lack of effect in acquisition, the delay of extinction may have reflected stronger fear memory formation in transgenic animals.

Rolipram, a phosphodiesterase inhibitor that increases cAMP concentrations in the brain, facilitates acquisition of fear and delays extinction. Once again, it is not clear whether rolipram impaired extinction or, alternatively, just increased fear acquisition, which would be expected to delay extinction.

Protein Kinase A

Infusion of the PKA inhibitor Rp-cAMPs into the CA1 region of hippocampus either before or immediately after the first of several extinction trials in a step-down avoidance paradigm produced a persistent impairment of extinction. However, intra-BLA infusions of the PKA activator 6-BNZ-cAMP immediately following each of four daily tone extinction training sessions had no effect on extinction of freezing.

Mitogen-Activated Protein Kinase

Intra-BLA (but not intrahippocampus) infusions prior to extinction training of the MAPK inhibitor PD98059 impaired extinction retention in the fear-potentiated startle paradigm. This was not due to state dependency, although it is not clear whether PD98059 affects the expression of extinction because experiments involving pretest infusions in animals extinguished drug-free were not conducted. A similar disruption of extinction of step-down avoidance following infusion of PD98059 into the CA1 region of hippocampus also has been seen. In the former study, infusions both before and after extinction training were effective, suggesting a role for MAPK in consolidation of extinction memory. Immediate but not delayed (2 or 4 h) post-extinction training infusions of PD98059 into medial prefrontal cortex (mPFC) impaired later extinction retention, using freezing to a tone as a measure of fear.

An upregulation of phosphorylated MAPK (pMAPK) was seen in extinguished wild-type mice relative to nonextinguished controls, and pMAPK actually decreased in CB receptor knockout (CB1−/−) mice that exhibit a profound impairment in extinction. Using a weak extinction training protocol, only a modest increase in pMAPK within the BLA was seen, but this effect was much larger when this same extinction protocol was combined with DCS, an effect blocked by intra-BLA administration of the MAPK inhibitors U0126 or PD98059.

Phosphatidylinositol 3-Kinase

Although direct measurement of changes in PI3K phosphorylation targets such as Akt has been inconsistent, intra-BLA infusions of the PI3K inhibitor wortmannin prior to extinction training blocked extinction of fear-potentiated startle, and intra-CA1 administration of the PI3K inhibitor LY 294002 immediately following daily extinction exposures to

a previously trained context impaired extinction in mice. DCS facilitation of extinction via systemic pre-extinction training administration of DCS was blocked by intra-BLA infusions of wortmannin.

Calcium/Calmodulin-Dependent Protein Kinase II

Intra-CA1 infusions of the CaMKII inhibitor KN-62 prior to, or immediately following, the first of several extinction exposures produced a persistent impairment extinction of step-down avoidance in rats. KN-62 also produced a persistent impairment in extinction when infused into the entorhinal cortex prior to the first extinction exposure. Post-extinction training infusions were not included in this latter study.

Src Family of Tyrosine Kinases

Infusions into the CA1 region of the hippocampus of the Src family inhibitor 4-amino-5-(4-chlorophenyl)-7-(*t*-butyl) pyrazolo [3,4-D] pyrimidine (PP2) immediately but not 180 min after each of four daily extinction exposures blocked extinction, suggesting a role for Src family kinases in consolidation of extinction memory. Subsequent extinction proceeded normally when extinction training was continued in the absence of infusions, indicating the infusions did not produce lasting damage to the hippocampus.

Cytoplasmic Polyadenylation Element Binding Protein

Cytoplasmic polyadenylation element binding protein (CPEB) knockout mice and wild-type controls acquired context fear normally, but the knockouts showed slower rates of between-session extinction. The knockouts also showed deficits in extinction in the Morris water maze, suggesting that CPEB may play a role in the extinction of hippocampus-dependent learning as opposed to fear extinction *per se*.

Calcineurin

The findings reviewed thus far suggest that a significant number and variety of kinases are activated in extinction. In apparent conflict with this, however, are data suggesting that extinction involves upregulation of the protein phosphatase calcineurin (also known as protein phosphatase 2B or PP2B), which would dephosphorylate (inactivate) many of these same kinases. Thus, both amygdalar depotentiation *in vitro* and fear extinction in behaving animals are associated with increased BLA calcineurin protein and enzymatic activity, and both depotentiation and extinction are blocked by pre-administration of calcineurin inhibitors. Like fear extinction, depotentiation also is blocked by NMDA receptor antagonists and L-VGCC inhibitors. Likewise, both extinction and 'quenching' of fear (defined as a loss of recently acquired fear responses following administration of depotentiation-inducing electrical stimulation to behaving animals) are associated with a reversal of fear conditioning-induced increases in BLA phosphokinases (Akt and MAPK), and in both cases the reversal is blocked by pre-administration of calcineurin inhibitors. Based on these similarities, it has been suggested that fear extinction involves a reversal of acquisition-related plasticity through upregulation of calcineurin and induction of depotentiation.

In other words, the argument is that extinction involves 'unlearning,' which, as we have seen, generally is problematic in light of data indicating that extinguished fear responses reappear, but may be an accurate description of the mechanism of extinction under some circumstances. On the other hand, there are some apparent inconsistencies within this body of work that await a satisfactory explanation. For example, if extinction is dependent on calcineurin-dependent depotentiation, why should MAPK and PI3K inhibitors impair extinction when MAPK and Akt are targets of calcineurin and are dephosphorylated following depotentiation-inducing stimulation? It is probable that some of these kinase and phosphatase cascades are modulated differentially within different BLA cell types, such as principal neurons and interneurons, which could explain the apparently conflicting observations. Hence, finer dissection of the cellular processes involved in extinction will require cell type-specific manipulations.

Transcription

Extinction has been associated with an increase in the expression of immediate-early genes in mPFC and BLA and an upregulation of gephyrin and BDNF mRNA in the BLA. A persistent impairment of extinction of step-down avoidance has been found when the transcription inhibitors a-amanitin or 5,6-dichloro-1-b-D-ribofuranosyl benzimidazole (DRB) were infused into the CA1 region of hippocampus 15 min before, but not 1 or 3 h after, the first of several extinction exposures to the apparatus. The lack of effect of post-extinction training administration is somewhat surprising given the likelihood that mRNA transcription is involved in consolidation of extinction memory, although it is possible that the onset of transcription occurs fairly rapidly (<1 h) following extinction training. Facilitation of extinction of fear-potentiated startle by pre-extinction training DCS was blocked by pre-administration of the transcription inhibitor actinomycin D into the BLA, an effect not due to state dependency.

Translation

Most studies employing pre- or post-extinction training administration of protein synthesis inhibitors have reported a profound impairment of extinction, despite differences in the organism being studied, the specifics of the training protocol, the method and timing of drug delivery, and the measure of conditioned fear. However, the issue of whether extinction is dependent on protein synthesis is complicated greatly by the procedural similarities between experiments examining extinction and those examining reconsolidation of conditioned fear.

Clinical Implications

The clinical implications of the work on fear extinction are particularly straightforward. Fear- and anxiety-related disorders such as phobias and PTSD seem to be disorders of fear dysregulation in which inhibition of fear is absent or insufficient in situations that are patently safe. Procedurally, extinction is very similar to several types of exposure-based psychotherapeutic treatments, and thus drugs found to facilitate extinction in rodents are now being tested to see if they will facilitate exposure-based psychotherapy.

DCS Facilitates Exposure Therapy

As we have already seen, DCS facilitates the functioning of the NMDA subtype of glutamate receptor and facilitates extinction of conditioned fear in animals. It has now been shown that DCS facilitates exposure-based psychotherapy in patients with fear of heights, social phobia, and obsessive-compulsive disorder, and similar studies are being conducted by several other groups for PTSD and other disorders. These studies represent outstanding examples of 'translational research' and give hope that the study of extinction in rodents will lead to more and more drugs that may improve exposure-based psychotherapy in several types of psychiatric disorders.

See also: Amygdala: Contributions to Fear; Aversive Emotions: Molecular Basis of Unconditioned Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Emotion: Neuroimaging; Extinction: Anatomy; Gamma-Aminobutyric Acid (GABA).

Further Reading

Barad M (2005) Fear extinction in rodents: Basic insight to clinical promise. *Current Opinion in Neurobiology* 15: 710–715.

Borowski TB and Kokkinidis L (1998) The effects of cocaine, amphetamine, and the dopamine D1 receptor agonist SKF 38393 on fear extinction as measured with potentiated startle: Implications for psychomotor stimulant psychosis. *Behavioral Neuroscience* 112: 952–965.

Bouton ME (2004) Context and behavioral processes in extinction. *Learning and Memory* 11: 485–494.

Chhatwal JP, Myers KM, Ressler KJ, and Davis M (2005) Regulation of gephyrin and GABAA receptor binding within the amygdala after fear acquisition and extinction. *Journal of Neuroscience* 25: 502–506.

Chhatwal JP, Stanek-Rattiner L, Davis M, and Ressler KJ (2006) Amygdala BDNF signaling is required for consolidation but not encoding of extinction. *Nature Neuroscience* 9: 870–872.

Delamater AR (2004) Experimental extinction in Pavlovian conditioning: Behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology B* 57: 97–132.

Falls WA, Miserendino MJD, and Davis M (1992) Extinction of fear-potentiated startle: Blockade by infusion of an NMDA antagonist into the amygdala. *Journal of Neuroscience* 12: 854–863.

Ledgerwood L, Richardson R, and Cranney J (2003) D-Cycloserine facilitates extinction of conditioned fear as assessed by freezing in rats. *Behavioral Neuroscience* 117: 341–349.

Lin CH, Yeh SH, Leu TH, Chang WC, Wang ST, and Gean PW (2003) Identification of calcineurin as a key signal in the extinction of fear memory. *Journal of Neuroscience* 23: 1574–1579.

Lu KT, Walker DL, and Davis M (2001) Mitogen-activated protein kinase cascade in the basolateral nucleus of amygdala is involved in extinction of fear-potentiated startle. *Journal of Neuroscience* 21: RC162.

Marsicano G, Wotjak CT, Azad SC, et al. (2002) The endogenous cannabinoid system controls extinction of aversive memories. *Nature* 418: 530–534.

Myers KM and Davis M (2007) Mechanisms of fear extinction. *Molecular Psychiatry* 12: 120–150.

Quirk GJ, Garcia R, and Gonzalez-Lima F (2006) Prefrontal mechanisms in extinction of conditioned fear. *Biological Psychiatry* 60: 337–343.

Rescorla RA (2001) Experimental extinction. In: Mowrer RR and Klein S (eds.) *Handbook of Contemporary Learning Theories*, pp. 119–154. Mahwah, NJ: Erlbaum.

Ressler KJ, Rothbaum BO, Tannenbaum L, et al. (2004) Cognitive enhancers as adjuncts to psychotherapy: Use of D-cycloserine in phobic individuals to facilitate extinction of fear. *Archives of General Psychiatry* 61: 1136–1144.

Richardson R, Riccio DC, and Devine L (1984) ACTH-induced recovery of extinguished avoidance responding. *Physiology and Psychology* 12: 184–192.

Santini E, Ge H, Ren K, Pena de Ortiz S, and Quirk GJ (2004) Consolidation of fear extinction requires protein synthesis in the medial prefrontal cortex. *Journal of Neuroscience* 24: 5704–5710.

Santini E, Muller RU, and Quirk GJ (2001) Consolidation of extinction learning involves transfer from NMDA-independent to NMDA-dependent memory. *Journal of Neuroscience* 21: 9009–9017.

Szapiro G, Vianna MR, McGaugh JL, Medina JH, and Izquierdo I (2003) The role of NMDA glutamate receptors, PKA, MAPK, and CAMKII in the hippocampus in extinction of conditioned fear. *Hippocampus* 13: 53–58.

Walker DL, Ressler KJ, Lu K-T, and Davis M (2002) Facilitation of conditioned fear extinction by systemic administration or intra-amygdala infusions of D-cycloserine as assessed with fear-potentiated startle in rats. *Journal of Neuroscience* 22: 2343–2351.

Yang YL, Chao PK, and Lu KT (2006) Systemic and intra-amygdala administration of glucocorticoid agonist and antagonist modulate extinction of conditioned fear. *Neuropsychopharmacology* 31: 912–924.

Pharmacology of Sleep: Adenosine

R Basheer, J T McKenna, and R W McCarley, Boston VA Healthcare System and Harvard Medical School, Brockton, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Approximately one-third of the human lifetime is spent in sleep. Undoubtedly, sleep is essential for the physical and mental well-being of the individual. Sleep disorders often lead to impaired daytime performance, with at times dire social and financial consequences. Approximately 40 million Americans suffer from chronic sleep disorders each year, and approximately 20 million experience occasional sleeping difficulties. Sleep pathologies such as insomnia, hypersomnia, or parasomnia occur either as a primary disorder or as a secondary disorder in conjunction with another illness. Research has focused on developing new therapeutic strategies, utilizing an increasing knowledge of the brain regions, neurotransmitters, and related molecular/cellular mechanisms that regulate sleep–wake behavior. In this article, we review the current status of pharmacologic agents and their molecular targets used in sleep therapy, followed by an extensive description of developments concerning adenosinergic mechanisms of sleep regulation. Finally, we discuss the prospects of using adenosine receptors or transporters as therapeutic targets for treating sleep disorders.

General Pharmacology of Sleep

Significant advancements have been made in understanding the brain mechanisms involved in sleep–wake regulation (Figure 1). The three behavioral states, also termed vigilance states, are waking, non-rapid eye movement (NREM) sleep, and rapid eye movement (REM) sleep (also known as paradoxical sleep or dream sleep). The transitions between these states involve different neurotransmitters, their receptor subtypes, and brain regions. Consequently, the pharmacological advances in sleep-promoting and wake or alertness-promoting drugs target a wide range of these neurotransmitter systems.

The first generation of sleep-promoting drugs composed of hypnotics or sedatives are known to enhance the action of the neurotransmitter GABA at the ionotropic $GABA_A$ receptor. Sedatives such as benzodiazepines (diazepam and valium), barbiturates, neurosteroids, and general anesthetics (isoflurane and propofol) target the $GABA_A$ receptor, potentiating the action of endogenous GABA. These drugs specifically bind to $GABA_A$ receptors that contain the γ subunit, in addition to one of the α subunits (1, 2, 3, or 5). These $GABA_A$ receptors are widely distributed in the cortex and other wake-related areas, such as the thalamus, tuberomammillary nucleus, basal forebrain, and hypothalamus. These drugs may act on targets of the sleep-promoting ventrolateral preoptic area (e.g., the tuberomammillary nucleus) by means of enhancing the inhibitory action of GABA to promote sleep.

A second generation of 'non-benzodiazepine' compounds has been developed which targets the $GABA_A$ receptor but is more selective for the α_1 subunit. Such compounds, including zolpidem, zopiclone, and zalplon, are termed the 'z' drugs. A newer class of compounds with similar actions, such as indiplon, is currently being evaluated. Also, a direct-acting $GABA_A$ agonist ($GABA_A$ receptor with α_4–δ subunits), gaboxadol, has been shown to improve sleep induction, maintenance, and sleep architecture. Interestingly, the selective $GABA_A$ receptor types that bind gaboxadol do not respond to benzodiazepines and are present in high density in the cortex, thalamus, and limbic system. At the cellular level, gaboxadol induces c-Fos expression in VLPO neurons, indicating activation of sleep-inducing neurons.

Drugs that act as antagonists of neurotransmitter receptors on wake-active neurons also exert sedative effects. For example, the histamine blockers dipenhydramine, eplivanserin, and pruvanserin are being developed as potential hypnotic agents. Since it is known that activation of the $5HT_{2A}$ receptor induces arousal, the specific receptor antagonist ketanserin has been developed.

Several recently developed compounds have been used to explore new avenues of pharmacological treatment of sleep disorders. GHB (γ-hydroxybutyrate) has been used to treat cataplexy/narcolepsy (sodium oxybate), and this compound is being considered for treatment of insomnia, although its use as a recreational drug has led to strict controls over prescriptions. Melatonin, secreted by the brain's pineal gland, has been considered a possible dietary supplemental treatment for insomnia, and it may help reduce jet lag. The M1 and M2 melatonin receptor agonist ramelteon has received much attention as a novel therapeutic agent for the treatment of insomnia.

Several drugs are targeted toward enhancing the activities of wake-active neuronal circuitry in order to treat conditions of hypersomnia. The most widely accepted drugs for enhancing wake and alertness are dopamine transport blockers, such as methylphenidale and modafinil. The histamine H_3 autoreceptor

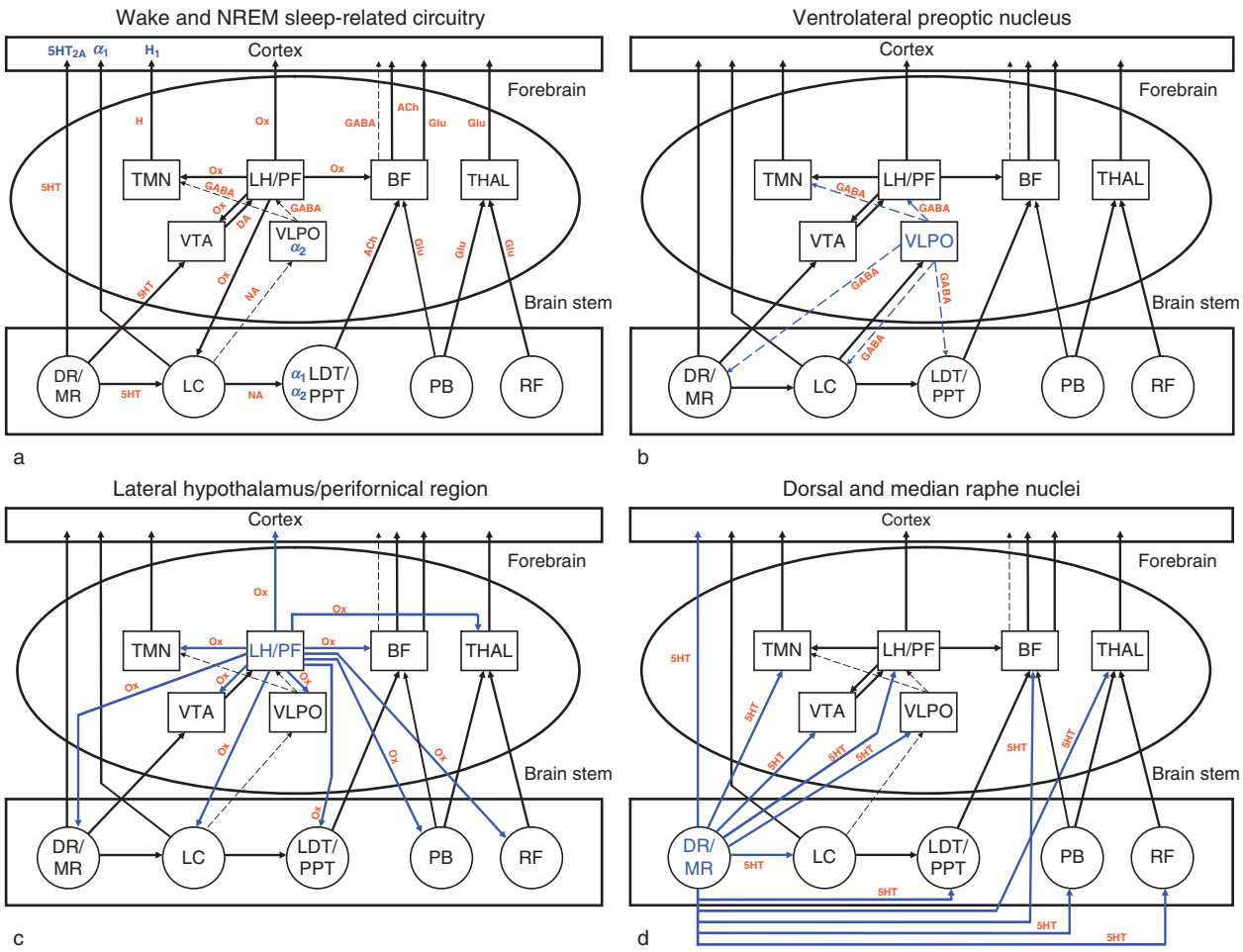

Figure 1 (a) Schematic diagram of the major brain circuitry involved in wake and NREM sleep. The arrows show the known connectivity between the brain regions. Solid lines denote excitatory neuronal outputs, and dashed lines denote inhibitory neuronal outputs. The neurotransmitters (red) shown include the following: ACh, acetylcholine; DA, dopamine; GABA, γ-aminobutyric acid; Glu, glutamate; H, histamine; 5HT, serotonin; NA, noradrenaline; Ox, orexin. Forebrain region: BF, basal forebrain. Cortex: LH/PF, lateral hypothalamus, perifornical region; THAL, thalamus; TMN, tuberomammillary nucleus; VLPO, ventrolateral preoptic nucleus; VTA, ventral tegmental area. Brain stem regions: DR/MR, dorsal and median raphe nuclei; LC, locus coeruleus; LDT/PPT, laterodorsal and pedunculopontine tegmental nuclei; PB, parabrachial nucleus; RF, reticular formation. The receptors known to mediate the effects in specific regions (blue) include the following: α_1, excitatory α_1 adrenoceptors; α_2, inhibitory α_2 adrenoceptors; 5HT2A, excitatory serotonin receptors; H1, excitatory histamine receptor. (b) Schematic diagram of proposed inhibitory projections (blue dashed lines) from the ventrolateral preoptic nucleus to proposed wake and NREM sleep-related brain regions. Note that these GABAergic projections are also galaninergic. (c) Schematic diagram of proposed orexinergic excitatory projections (blue solid lines) from the lateral hypothalamus/perifornical region to proposed wake and NREM sleep-related brain regions. (d) Schematic diagram of proposed serotonergic excitatory projections (blue solid lines) from the dorsal and median raphe nuclei to proposed wake and NREM sleep-related brain regions. Adapted from Szabadi E (2006) Drugs for sleep disorders: Mechanisms and therapeutic prospects. *British Journal of Clinical Pharmacology* 61: 761–766; and Saper CB, Scammell TE, and Lu J (2005) Hypothalamic regulation of sleep and circadian rhythms. *Nature* 437: 1257–1263.

antagonists ciproxifan and thioperamide can enhance histamine release and increase waking.

In addition to the previously mentioned neurotransmitters, much research has also focused on endogenous homeostatic factors that integrate sleep regulation with feeding and metabolism, such as the delta sleep factor, cytokines, growth factors, and peptides (orexin, cholecystokine octapeptide, neuropeptide Y, and ghrelin). For example, antagonists of the orexin receptors, such as compound GW649868, are being tested for use as a sedative. Overall, therapeutic strategies that aim to design sleep-promoting and wake-promoting drugs utilize a multitude of approaches developed through knowledge of brain circuitry and manipulation of the different known neurotransmitter systems.

Adenosine as a Potential Target of Hypnotics

The nucleoside adenosine is an endogenous neuromodulator and a homeostatic sleep factor. Developments with regard to adenosine's role in sleep regulation and the delineation of underlying adenosinergic molecular mechanisms have presented several possibilities for treatment of insomnia and hypersomnia. In the following sections, we discuss the brain regions, receptors, and transporters of adenosine that are associated with sleep regulation, and the pharmacological possibilities in designing therapeutic compounds are also discussed.

Adenosine is a ubiquitous molecule, being an integral constituent of several macromolecules such as DNA, RNA, coenzymes, and the high-energy molecule adenosine triphosphate (ATP). The first description of the hypnogenic effects of adenosine was published in 1954 on cats, followed by studies on dogs. Since then, several laboratories have shown that systemic or intracerebroventricular administration of adenosine or agonists of adenosine receptors induces sleep. In 1981, researchers showed that the often-consumed compound caffeine (a theophylline derivative) is behaviorally stimulating due to its structural similarity to adenosine and its ability to block adenosine receptors.

Investigators originally postulated that adenosine levels would increase throughout the brain during wake due to higher neuronal metabolic activity related to ATP breakdown. Thus, in the brain, in which energy metabolism is tightly regulated, adenosine acts as a 'fatigue factor' indicating the need for rest/sleep. In cats, levels of extracellular adenosine, measured by microdialysis, increased 15–20% during wake compared to NREM sleep in brain regions associated with sleep–wake regulation, such as the basal forebrain, cerebral cortex, thalamus, preoptic area of hypothalamus, dorsal raphe nucleus, and pedunculopontine tegmental nucleus. However, prolonged waking achieved by gentle handling revealed a site-specific significant increase in extracellular adenosine in only the magnocellular basal forebrain (140% of the baseline) and, to a lesser extent, in the cingulate cortex. Infusion of adenosine into the rodent or feline magnocellular basal forebrain increased sleep, thus indicating that this region is a major site for the sleep-inducing effects of adenosine. Although adenosine perfusion in other brain regions, such as the pedunculopontine tegmental nucleus and the subarachnoid space ventral to the rostral forebrain, has also been shown to induce sleep, the sleep deprivation-induced increase in extracellular adenosine was rather selective to the magnocellular basal forebrain, suggesting that this region is one of the major sites for the homeostatic regulation of sleep by adenosine.

Four subtypes of adenosine receptors – A1, A2a, A2b, and A3 – have been identified based on their specific order of potency for agonists and antagonists and their particular structure ascertained by gene cloning. Of these, the A1 receptor is widespread throughout the brain, whereas A2a receptors are far more abundant in the basal ganglia and olfactory tubercle, and they are less abundant in the cortex. A2b and A3 receptors are less abundant and their role in sleep is not clear. A1 and A2a receptors are known to regulate adenylate cyclase and cellular activity in opposite ways. The A1 receptor, coupled to the G_i subtype of the G-protein, exerts inhibitory effects, whereas the A2a receptor, coupled to the G_s subtype of the G-protein, exerts excitatory effects on cellular activity. Both A1 and A2a receptors have been implicated in mediating the sleep-inducing effects of adenosine. The effect of adenosine on sleep varies depending on the receptor and the brain region of interest, as described later (**Figure 2**).

Subtype-specific activation of receptors has varying intracellular consequences depending on the receptor-coupling proteins inside the cell membrane (**Figure 3**). The A1 receptor couples to the G_i subtype heterotrimeric G-protein and is recognized for its dual signaling capabilities. It can either inhibit adenylate cyclase activity via the $G_{\alpha i}$ subunit, resulting in reduced cAMP production, or activate phospholipase C (PLC) via the $G_{\beta\gamma}$ subunits, which can lead to inositol trisphosphate (IP_3) production and subsequent release of intracellular calcium by activating IP_3 receptors. In neurons, A1 receptors can activate pertussis toxin-sensitive K^+ channels, as well as K_{ATP} channels, and inhibit Q-, P-, and N-type Ca^{2+} channels. The selectivity of activation and intracellular signaling is also influenced by homo- or heterodimerization. For example, adenosine A1 receptor function can be influenced by the homodimer or heterodimer or by interaction with ectoadenosine deaminase.

A2a receptors are coupled to the G_s-subtype G-protein, and hence increase adenylate cyclase activity. The A2b receptors are positively coupled to adenylate cyclase and PLC. The A3 receptor, like A1, inhibits adenylate cyclase, as well as stimulates PLC and induces intracellular calcium release.

Adenosine has been shown to inhibit neuronal activity by means of A1 adenosine receptors in many of the arousal-related brain regions. In the basal forebrain, where predominantly wake-active neurons have been described, the A1 adenosine receptor has been localized to both cholinergic and non-cholinergic neurons. Perfusion of the A1 receptor

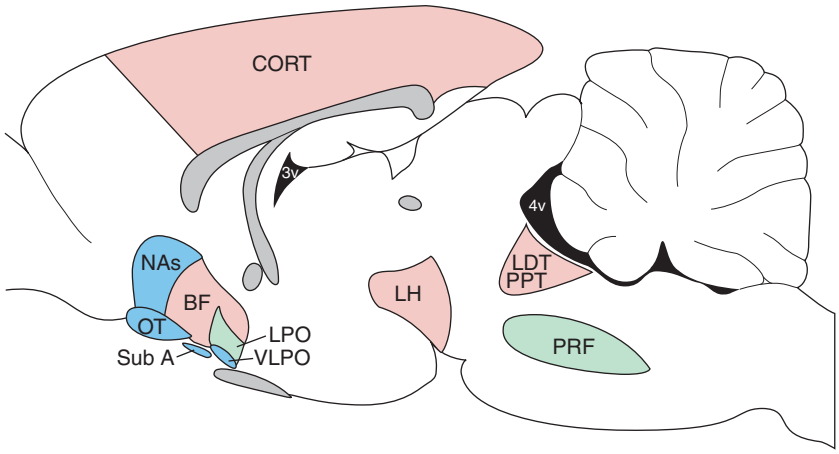

Figure 2 Sagittal representation of the rodent brain showing brain regions with documented adenosine effects on sleep–wake behavior. The brain regions shaded in pink are recognized for A1 adenosine receptor-, blue for A2a receptor-, and green for both A1 and A2a receptor-mediated effects on sleep. BF, basal forebrain; CORT, cortex; LDT/PPT, laterodorsal and pedunculopontine tegmental nuclei; LH, lateral hypothalamus, perifornical region; LPO, lateral preoptic area; NAs, shell of the nucleus accumbens; OT, olfactory tubercle; PRF, pontine reticular formation; SubA, subarachnoid space; VLPO, ventrolateral preoptic area; 3V, third ventricle; 4V, fourth ventricle.

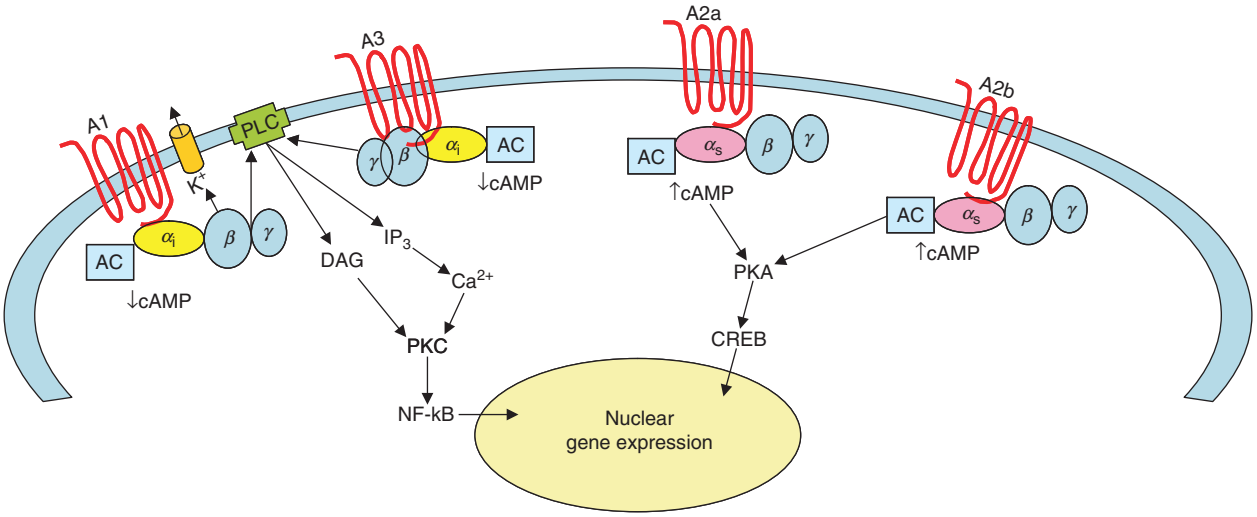

Figure 3 Intracellular signaling pathways of adenosine receptors: A1 and A3 receptors are coupled to pertussis toxin-sensitive G_i-subtype G-protein. Their activation either inhibits adenylate cyclase activity, preventing cAMP production, or activates phospholipase C, releasing diacylglycerol and inositol trisphosphate, which leads to protein kinase C activation. A2a and A2b receptors are coupled to cholera toxin-sensitive G_s-subtype G-protein. Activation of these receptors can increase cAMP production and activation of protein kinase A. The activation kinases can lead to phosphorylation of cellular proteins, including transcription factors such as NF-κB and CREB, leading to changes in gene expression. AC, adenylate cyclase; CREB, cAMP response element-binding protein; DAG, diacylglycerol; IP3, inositol trisphosphate; PKA, protein kinase A; PKC, protein kinase C; PLC, phospholipase C.

antagonist, cyclo-pentyl-1, 3-dimethylxanthine, increased waking and decreased sleep. Reduction of the A1 receptor by infusing antisense for the A1 receptor into the rodent basal forebrain resulted in a significant decrease in slow-wave delta activity during NREM sleep following 6 h of sleep deprivation.

Actions of adenosine at the A1 receptor vary according to its localization, but almost all promote sleep. In cortex, the target of wake-active cholinergic neurons, activation of presynaptic A1 receptors inhibited acetylcholine release, thereby promoting sleep. Presynaptic A1 receptor activation in the ventrolateral preoptic area, known to be sleep active, disinhibited sleep-active GABAergic neurons and led to sleep induction. Also, activation of the somatic A1 receptor on wake-active orexinergic neurons may induce sleep. The wake and REM-active cholinergic neurons of the brain stem laterodorsal and pedunculopontine tegmental nuclei are also directly inhibited by adenosine acting via the A1 receptor. Lastly, in the pontine reticular formation, both A1 and A2a receptors have been implicated in sleep regulation.

A2a receptors are located in select brain nuclei/regions involved in sleep regulation, such as the subarachnoid space, which is located below the rostral basal forebrain and the shell of the nucleus accumbens. Administration of adenosine or the A2a receptor agonist CGS21680 into the subarachnoid space below the rostral basal forebrain or the shell of the nucleus accumbens near the olfactory tubercle has been shown to promote sleep. Data suggest that prostaglandin D2 may induce the release of adenosine and, via A2a receptors, promote sleep. The sleep-inducing effects of both prostaglandin D and the A2a agonist CGS21680 can be blocked by the A2a receptor antagonist KF17937. In the lateral preoptic area, which contains numerous sleep-active neurons, a clear distinction between the behavioral effects of A1 and A2a receptor activation has been reported. A1 receptor activation of the sleep-active neurons in this area resulted in waking, possibly by inhibition, whereas A2a receptor activation induced sleep by stimulating the activity of sleep-active neurons. The olfactory tubercle contains a high density of A2a receptors, but their role in sleep is unknown.

In addition to activation or inhibition of adenosine receptors, changes in the adenosine metabolizing enzymes or adenosine transporters could potentially contribute to extracellular adenosine levels and promote sleep. For example, administration of the adenosine kinase blocker ABT-702 has been shown to increase sleep. Binding of the equilibrative transporter ENT1 was reduced in the basal forebrain following sleep deprivation in rats, suggesting that a decrease in adenosine transport into cells may contribute to the sleep deprivation-induced increase in extracellular adenosine and subsequent sleep.

Thus, research on the effect of adenosine on sleep has revealed a large number of potential targets for the development of pharmacological drugs. The agonists of adenosine receptors may have potential for treating insomnia, and the agents that antagonize adenosine receptors may be useful for treating hypersomnia and increasing alertness.

Adenosine Receptors as Therapeutic Targets

Adenosine receptor modulation has been utilized in the development of useful therapeutics for treating a variety of pathologies, including inflammatory, cardiovascular, and neurodegenerative diseases. A large number of compounds have been generated as agonists or antagonists that are essentially the derivatives of adenosine and methylxanthine molecules, respectively (Figure 4). Advances in molecular modeling have allowed definition of interactions between potential agonists/antagonists and adenosine receptors. Adenosine receptors, like all G-protein-coupled receptors, have a central core consisting of seven transmembrane helices connected with three intracellular and three extracellular loops. The selectivity of these receptors and their respective ligands is often coded by the varying lengths of the N-terminal extracellular domain and the C-terminal intracellular domain, as well as the intracellular loops. In the case of adenosine receptors' consensus sites, N-linked glycosylation has been identified in the extracellular regions, but the location of these sites within the extracellular region and the mode of posttranslational modification vary for each receptor subtype. Site-directed mutagenesis of single amino acids, critical for ligand recognition and responsible for discerning between agonists and antagonists, has been central to the medicinal chemistry of the drug–receptor interaction.

Several potent and selective adenosine receptor agonists have been utilized for treatment, and many are being tested for use in inflammation, analgesia, benign prostatic hypertrophy, and cardiovascular disease, in contrast to the relative paucity of development for use in sleep and wakefulness. Some adenosine antagonists are derivatives of xanthines, but these compounds have limited water solubility and bioavailability. Consequently, nonxanthine compounds with increased water solubility and bioavailability have been synthesized. Examples of A1, A2, and A3 selective agonists/antagonists are presented in Tables 1–3, respectively.

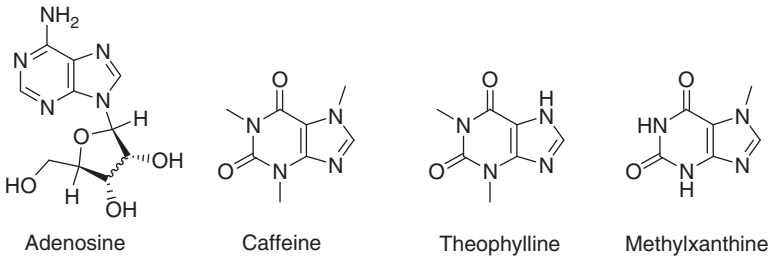

Figure 4 Chemical structures of adenosine, caffeine, theophylline, and methylxanthine.

Table 1 Select human A1 receptor agonists and antagonists

Adenosine A_1 receptor subtype	*Compound*	*K_i value for AR (nM)*			
		A_1AR^*	*$A_{2A}AR^*$*	*$A_{2B}AR^*$*	*A_3AR^**
Agonists					
	CPA	2.3	794	18 600	72
	CCPA	0.83	2270	18 800	38
	S(−)-ENBA	0.38	>10 000	>10 000	915
	ADAC	0.85	210	N.D.	13.3
	AMP579	5.1	56	N.D.	N.D.
	NNC-21–0136	10	630	N.D.	N.D.
	GR79236	3.1	1300	N.D.	N.D.
	CVT-510 (Tecadenoson)	6.5	2315	N.D.	N.D.
	SCZWAG 994	23	25 000	>10 000	N.D.
	Selodenoson	1.1	306	N.D.	N.D.
Antagonists					
	DPCPX	3.9	129	56	3980
	WRC-0571	1.7	105	N.D.	7940
	BG 9719	0.43	1051	172	3870
	BG 9928	29	4720	690	42 110
	FK453	18	1300	980	>10 000
	FR194921	2.9	10 000	N.D.	>10 000
	KW3902	1.3	380	N.D.	N.D.

N.D., not determined.
Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Drug Discovery* (Jacobson KA and Gao Z-G (2006) Adenosine receptors as therapeutic targets. *Nature Reviews Drug Discovery* 5: 247–264), copyright (2006).

Table 2 Select human A2 receptor agonists and antagonists

Adenosine A_2 receptor subtypes	*Compound*	*K_i value for AR (nM)*			
		A_1AR^*	*$A_{2A}AR^*$*	*$A_{2B}AR^*$*	*A_3AR^**
Agonists					
A_{2A}	NECA	14	20	140	25
	CGS21680	289	27	>10 000	67
	DPMA	168	153	>10 000	106
	Binodenoson	48 000	270	430 000	903
	ATL-146e	77	0.5	N.D.	45
	CV-3146	>10 000	290	>10 000	>10 000
A_{2B}	LUF5835	4.4	21	10	104
Antagonists					
A_{2A}	KW6002	2830	36	1800	>3000
	CSC	28 000	54	N.D.	N.D.
	SCH 58261	725	5	1 110	1 200
	SCH 442416	1110	0.048	>10 000	>10 000
	ZM241,385	774	1.6	75	743
	VER 6947	17	1.1	112	1 470
	VER 7835	170	1.7	141	1 931
	'Schering compound'	82	0.8	N.D.	N.D.
A_{2B}	MRS1754	403	503	2	570
	MRE 2029-F20	245	>1000	3	>1000
	OSIP-339391	N.D.	N.D.	0.5	N.D.

N.D., not determined.
Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Drug Discovery* (Jacobson KA and Gao Z-G (2006) Adenosine receptors as therapeutic targets. *Nature Reviews Drug Discovery* 5: 247–264), copyright (2006).

Nucleoside Transporters as Therapeutic Targets

Adenosine is hydrophyllic and its transport across the cell membrane is carried out by two kinds of nucleoside transporters: the equilibrative nucleoside transporters (ENTs) and the concentrative nucleoside transporters (CNTs). The ENTs allow facilitated diffusion of nucleosides, whereas CNTs are cation symporters. ENT1–4 vary in their tissue distribution pattern. In the brain, ENT1 and ENT4 are

Table 3 Select human A3 receptor agonists and antagonists

A_3 receptor subtype	*Compound*	*K_i value for AR (nM)*			
		*A_1AR**	*$A_{2A}AR$**	*$A_{2B}AR$**	*A_3AR**
Agonists					
A_3	IB-MECA	51	2900	11 000	1.8
	Cl-IB-MECA	220	5360	>100 000	1.4
	LJ568	193	223	N.D.	0.38
	CP-608039	7200	N.D.	N.D.	5.8
	MRS3558	260	2300	>10 000	0.29
	MRS1898	136	784	N.D.	1.5
Antagonists					
A_3	OT-7999	>10 000	>10 000	>10 000	0.61
	MRS1292	12 100	29 800	N.D.	29.3
	PSB-11	1640	1280	2 100	4
	MRS3777	>10 000	>10 000	>10 000	47
	MRS1334	>100 000	>100 000	N.D.	2.7
	MRE 3008-F20	1200	141	2100	0.82
	MRS 1220	305	52	N.D.	0.65
	MRS1523	15 600	2050	N.D.	18.9
	Novartis compound	197	1670	3	10

N.D., not determined.

Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Drug Discovery* (Jacobson KA and Gao Z-G (2006) Adenosine receptors as therapeutic targets. *Nature Reviews Drug Discovery* 5: 247–264), copyright (2006).

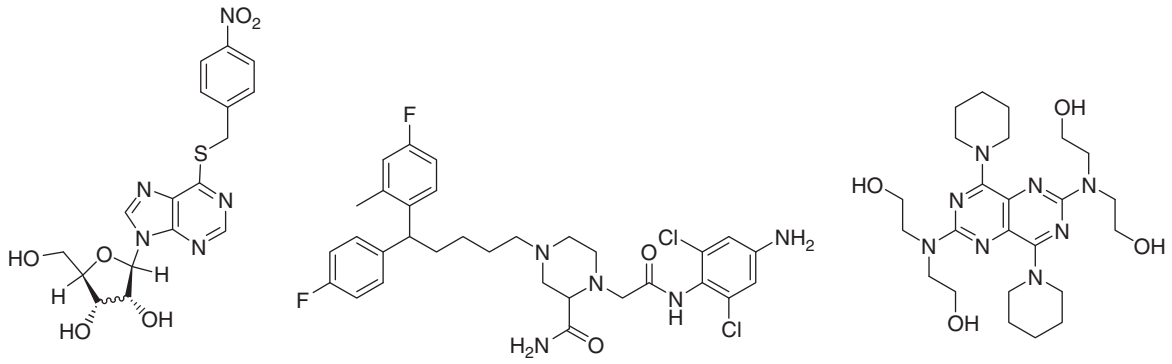

Figure 5 Chemical structures of compounds that block nucleoside transporters.

more abundant. There are three subtypes of CNTs (CNT1–3), and CNT2 is most abundant in brain. Although the affinity for adenosine is much greater for CNT2 compared to ENT1 (K_m of 6 μM for CNT2 and 40 μM for ENT1), the adenosine transport rate is much slower with CNT2. In the cortex, CNT2 mRNA levels and, in the magnocellular basal forebrain, ENT1 binding to NBMPR were attenuated after 24 and 6 h of sleep deprivation, respectively. The downregulation of these transporters was suggested to contribute to the observed increase in extracellular adenosine in these two brain regions following sleep deprivation. Consequently, inhibitors of adenosine transporters that increase the levels of extracellular adenosine may be promising compounds for treatment of such sleep disorders as insomnia. Experimental proof of this phenomenon has been obtained in the rat with infusion of the ENT1 inhibitor NBMPR into the basal forebrain, where levels of extracellular adenosine increased and sleep was induced. Several such ENT1 inhibitors, including dilazep, dipyridamole, draflazine, and lidoflazine, all with different molecular structures, have been identified for their ability to enhance the extracellular concentration of adenosine by blocking adenosine reuptake (Figure 5).

Allosteric Modulators as Therapeutic Agents

Adenosinergic receptor activity may also be influenced by compounds that bind to receptor allosteric sites that are topologically distinct from the orthosteric sites to which endogenous ligands, as well as

exogenous agonists and antagonists, bind. Allosteric ligand binding does not overlap with the orthosteric site, thus allowing simultaneous binding of compounds at both sites. These compounds can thus selectively enhance the response to adenosine in only those organs or a localized area of a given organ in which the production of adenosine is increased. At the functional level, molecules that bind to allosteric sites modulate receptor activity by causing conformational changes, leading to either decreased (positive regulators or allosteric enhancers) or increased (negative regulators) dissociation kinetics of the orthosteric ligand. Therefore, the use of allosteric modulators as therapeutic agents has become very attractive. Allosteric modulators exhibit greater subtype selectivity and fewer side effects. Moreover, the neurotransmitter receptors are less sensitive to desensitization or downregulation by allosteric enhancers compared to exogenous agonists. Typical examples are the benzodiazepines, such as diazepam, acting as allosteric enhancers of $GABA_A$ receptors with limited side effects, whereas direct agonists of $GABA_A$ receptors have widespread side effects and are therefore unsuitable for treatment. Similarly, an allosteric modulator of the nicotinic receptor, the alkaloid galanthamine, potentiates nicotinic cholinergic neurotransmission without major side effects and is used for treatment of Alzheimer's disease.

Allosteric Modulators of the A1 Adenosine Receptor Subtype

Allosteric modulators have also been described for adenosine receptors and transporters and are potential therapeutic agents (Figure 6). The benzoylthiophene derivatives, such as PD 81,723 (2-amino-4,5-dimethyl-3-thienyl-[3[(trifluromethyl)phenyl]-methanone) and related analogs, were the first compounds to be identified as A1 receptor-selective allosteric enhancers due to their ability to increase the levels of agonist binding. In the past decade, multiple derivatives of PD 81,723, with variable potency for enhancing the agonist binding to the A1 receptor, have been synthesized. Among them, PD 71,605 (T-62) proved to be a potent enhancer of agonist binding to membranes of Chinese hamster ovary cells expressing the human A1 receptor. 2-Aminothiazlium salts are a new class of allosteric enhancers of the A1 receptor that are suggested to be more stable than benzoylthiophenes. Most of the tests for the potency of this compound have involved *in vitro* systems, however, and clinical use of these compounds has not been reported. On the other hand, tests in animal studies showed that PD 81,723 exhibited some degree of tissue selectivity, modulating neuronal but not adipocyte adenosine receptors. This compound's effects were also species specific, showing greater enhancement of agonist binding in the guinea pig and dog brain compared to the rodent brain. T-62 has been shown to reduce mechanical hypersensitivity in a rat model of neuropathic pain, thus demonstrating its potential use in pain therapy.

Allosteric Modulators of the A2a and A2b Adenosine Receptor Subtypes

The allosteric modulators of the A2a receptor thus far described lack absolute specificity. For example, PD 120918 (4-methyl-7([methylamno]carbonyl)oxy)-2H-1-benzopyran-2-one) has been shown to slow the dissociation of the adenosine agonist from the rat brain A1 receptor, therefore appearing to be not very selective for the A2a receptor. Other compounds, such as amiloride analogs and SCH-202676, increased the antagonist dissociation from the A2a receptor but did not show any effect on the dissociation rate of agonist CGS21680. To our knowledge, no known allosteric compound modulates the A2b adenosine receptor subtype. It is suggested, however, that SCH-20676 can also function as an allosteric modulator of the A2b subtype receptor (Figure 7).

Allosteric Modulators of the A3 Adenosine Receptor Subtype

Two distinct compounds and their derivatives have been identified to act as allosteric modulators of the A3 adenosine receptor subtype. 3-(2-pyridinyl) isoquinoline derivatives were very selective for the

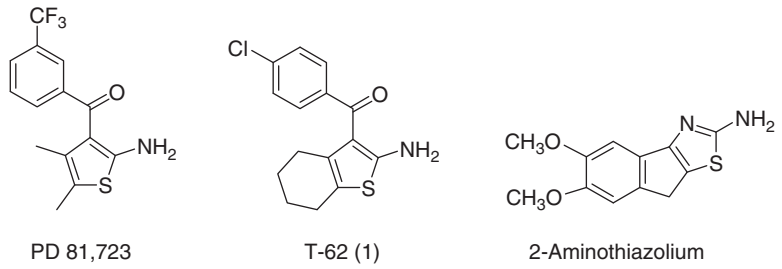

Figure 6 Chemical structures of allosteric modulators of A1 receptors.

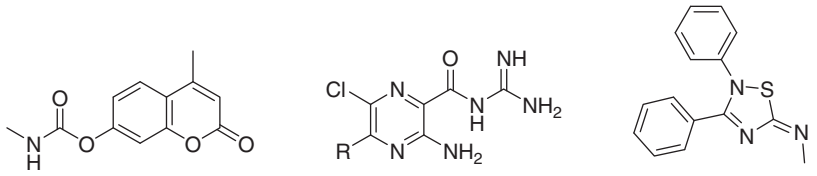

Figure 7 Chemical structures of allosteric modulators of A2a receptors.

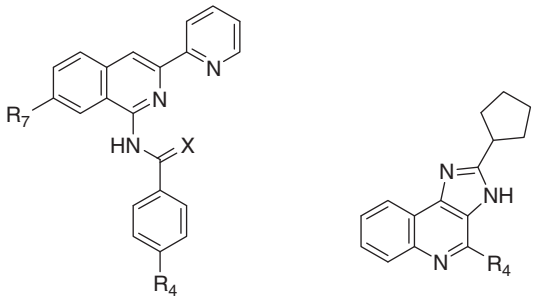

Figure 8 Chemical structures of allosteric modulators of A3 receptors.

A3 receptor, and they have been shown to enhance binding of the agonist [^{125}I]-N6(4-amino-3-iodobenzyl)-5′-*N*-methyl carboxamidoadenosine (I-AB-MECA, also known as IB-MECA) in a concentration-dependent manner. These derivatives not only enhanced agonist binding to the A3 receptor but also enhanced A3 receptor function (i.e., inhibition of forskolin-induced cAMP production). However, these compounds also demonstrated antagonist-like activity. Further studies have shown that the removal of a 7-methyl group decreased the antagonistic function without any loss in the allosteric enhancing activity, suggesting that a structural requirement for allosteric enhancement might be distinct from those for competitive antagonism (Figure 8).

Another set of compounds, 1H-imidazo-[4,5-c] quinoline and its derivatives, exhibited A3 receptor-selective allosteric enhancement of agonist binding, specific to the human A3 receptors. A few derivatives of this compound decreased the dissociation of the agonist IB-MECA without any effect on A3 receptor antagonist binding or A1/A2a receptor agonist binding. A few of these derivatives, however, tested positive for antagonistic binding to the receptor at the orthosteric site. These compounds demonstrated enhanced inhibition of forskolin-stimulated cAMP production in cells that expressed human A3 receptors.

In addition, another subset of the previously mentioned compounds for the A2a receptor, the amiloride analogs, may be more selective for the A3 receptor. Studies have shown that the former was more effective as an allosteric enhancer for the A3 receptor and acted as allosteric inhibitors of the antagonist binding to A1, A2a, and A3 receptors. Specifically, it was shown that the dissociation rates for IB-MECA binding of the human A3 receptor, as well as forskolin-stimulated cAMP production, were significantly reduced without any effect on the binding of A1 and A2a agonists to their respective receptors, suggesting a higher selectivity of amiloride compounds for the A3 receptor.

Adenosine as a Therapeutic Agent for Other Diseases

In the drug discovery arena, the development of compounds with receptor subtype-specific pharmacodynamics without undesired side effects has been a constant challenge. Several adenosinergic drugs have been described for the treatment of both central nervous system (CNS) and non-CNS-related illnesses. Some of these compounds are already in use, others are in later stages of clinical trials, and many more are being tested in animals. In patients with paroxymal supraventricular tachycardia, a condition defined as an abnormally fast heart rate due to the rapid firing of electrical impulses from the atrioventricular node, intravenous infusion of adenosine has been found to activate A1 receptors and restore a normal heart rhythm. Another A1 receptor-selective agonist, called selodenoson, tecadenoson, or CVT 2759, in clinical trials, has been developed for its capacity to slow heart rate without lowering blood pressure. However, it has been found to have the detrimental side effect of producing renal toxicity.

Activation of A1 receptors has also been recognized to provide cytoprotection during ischemia. The A1 receptor antagonist BG9719 is undergoing phase II clinical trials for improvement of renal function and treatment of congestive heart failure. Adenosine is being used in imaging for the evaluation of coronary artery disease. Another A2a receptor agonist, regadenoson, is also being used for myocardial perfusion imaging. The A2a receptor agonist GW328267 is undergoing phase II clinical trials for the treatment of asthma, a chronic obstructive pulmonary disease.

Other adenosinergic compounds have been developed to treat other pathologies. Some CNS drugs, such as cabamazepine, exhibit pain-relieving effects which are mostly mediated through A1 and, partially, through A2 receptors. Another A1 receptor-selective

agonist, GW-493838, is in clinical trials for the treatment of pain and migraine headaches. An allosteric enhancer of the A1 receptor, T-62, is in clinical trials for treating neuropathic pain. An A1 receptor antagonist, FR194921, is being investigated for use as a potential treatment of dementia and anxiety disorders. NNC-21–0136, another A1 receptor selective agonist, has been shown to be selectively neuroprotective during global ischemia, with significantly few cardiovascular effects. Lastly, the A2a receptor antagonist istradefylline (KW-6002) is in phase III clinical trials for the treatment of Parkinson's disease.

Conclusion

Research has strongly supported the role of adenosine as a potent somnogenic factor. Although no drugs targeting adenosine receptors or transporters have been utilized as compounds for treating insomnia or hypersomnia, there appears to be potential for such therapeutic strategies in the future. With improved understanding of the brain regions and the receptor subtypes mediating the effects of adenosine, it is time to test or design specific modulators of adenosine receptors or transporters for treating sleep disorders. However, it is important to note that one of the basic challenges in using adenosine receptor or transporter modulators as therapeutic agents is the ubiquitous distribution of these molecules and the possibility of undesirable side effects. Strategies that include mediation of effects specific to the brain and, better still, in a localized area of the brain will be important. For example, allosteric modulators may potentiate an adenosine receptor agonist's action in an agonist concentration-dependent manner, which could hold great promise for enhancement of the localized effects of adenosine. The concept of engineering adenosine receptors (neoceptors) in concert with uniquely modified/designed ligands (neoligands) is being explored for use in organ-targeted gene therapy. Early studies have demonstrated successful synthesis of neoceptors for the A2a and A3 adenosine receptors. This approach may hold promise for the future of targeted drug therapy. In summary, the adenosine receptors present a promising target for pharmacological intervention, one heretofore not utilized in the treatment of insomnia and hypersomnia.

See also: Adenosine; Adenosine Receptor Mediated Functions; Herbal Products and GABA Receptors; Nightmares; Reticular Activating System; Sleep Architecture; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function; Sleep in Aging; Stimulant and Wake-Promoting Substances.

Further Reading

Basheer R, Strecker RE, Thakkar MM, and McCarley RW (2004) Adenosine and sleep–wake regulation. *Progress in Neurobiology* 73: 379–396.

Boutrel B and Koob GF (2004) What keeps us awake: The neuropharmacology of stimulants and wakefulness promoting medications. *Sleep* 27: 1181–1194.

Fredholm BB, Chen J-F, Masino SA, and Vaugeois J-M (2005) Actions of adenosine at its receptors in the CNS: Insights from knockouts and drugs. *Annual Review of Pharmacology and Toxicology* 45: 385–412.

Fredholm BB, Ijzerman AP, Jacobson KA, Klotz KN, and Linden J (2001) International Union of Pharmacology: XXV. Nomenclature and classification of adenosine receptors. *Pharmacological Reviews* 53: 527–552.

Gao Z-G, Kim S-K, Ijzerman AP, and Jacobson KA (2005) Allosteric modulation of the adenosine family of receptors. *Mini Reviews in Medicinal Chemistry* 5: 545–553.

Huang Z-L, Urade Y, and Hayaishi O (2007) Prostaglandins and adenosine in the regulation of sleep and wakefulness. *Current Opinion in Pharmacology* 7: 33–38.

Jacobson KA and Gao Z-G (2006) Adenosine receptors as therapeutic targets. *Nature Reviews Drug Discovery* 5: 247–264.

King AE, Ackley MA, Cass CE, Young JD, and Baldwin SA (2006) Nucleoside transporters: From scavengers to novel therapeutic targets. *Trends in Pharmacological Sciences* 27: 416–425.

Moro S, Gao Z-G, Jacobson KA, and Spalluto G (2005) Progress in the pursuit of therapeutic adenosine receptor antagonists. *Medicinal Research Reviews* 26: 131–159.

Porkka-Heiskanen T, Strecker RE, Thakkar M, et al. (1997) Adenosine: A mediator of the sleep-inducing effects of prolonged wakefulness. *Science* 276: 1265–1268.

Saper CB, Scammell TE, and Lu J (2005) Hypothalamic regulation of sleep and circadian rhythms. *Nature* 437: 1257–1263.

Steriade M and McCarley RW (2005) *Brain Control of Wakefulness and Sleep*. New York: Kluwer/Plenum.

Strecker RE, Morairty S, Thakkar MM, et al. (2000) Adenosinergic modulation of basal forebrain and preoptic/anterior hypothalamic neuronal activity in the control of behavioral state. *Behavioural Brain Research* 115: 183–204.

Szabadi E (2006) Drugs for sleep disorders: Mechanisms and therapeutic prospects. *British Journal of Clinical Pharmacology* 61: 761–766.

Pheromones and other Chemical Communication in Animals

T Wyatt, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Chemical communication is arguably the most widely shared sensory modality, used almost universally across the animal kingdom. From the simplest bacteria to humans, almost all organisms use chemical communication. Within a species, the molecules used for chemical communication are called pheromones. The word pheromone comes from the Greek *pherein*, to carry or transfer, and *horman*, to excite or stimulate. Pheromones are often divided by function, such as sex pheromones and aggregation pheromones.

A broader term for chemicals involved in animal communication is semiochemical (from the Greek *semeion*, sign). Chemical signals broadcast to the world can be detected by individuals from other species. So, for example, pheromones released by bark beetles may attract their predators – which use the pheromones as a kairomone to locate their bark beetle prey. Chemical signals in deceit or propaganda are termed allomones (discussed in the section titled 'Eavesdropping and propaganda').

Evolution of Chemical Communication

Pheromones evolve from chemicals released to the environment that give an advantage to the receiver (and, often, the sender). The clue to such an evolutionary path is often given by the identity of the compounds used as pheromones. For example, the sex pheromones released by female goldfish appear to be very similar to the steroid hormones that are circulating in her blood while her eggs are developing. Male goldfish have evolved great sensitivity to these hormonelike compounds. Presumably, over evolutionary time, males able to detect these compounds, initially inadvertently leaking from females, would gain more fertilizations as they would be better able to find the females about to spawn. There would be selection for greater and greater sensitivity in the 'noses' of the males – to the extent that now male goldfish can detect the waterborne sex pheromones down to nanomolar and picomolar threshold concentrations. Females would gain more-reliable fertilization, so there would be selection for females that, instead of simply leaking these compounds, actually secreted them. Thus, theoretically, something which started as eavesdropping by males could evolve into a pheromone signal produced by females specifically to attract males.

Another example comes from the alarm pheromones in ants, released when threatened by a predator or invading ants from another colony. A bizarre variety of compounds are found as alarm pheromones in different taxonomic groups of ants. The pattern is largely explained when the alarm pheromones are compared with the toxic compounds that are released in fighting or defense in those species – the alarm pheromones appear, in many species, to have evolved from defense compounds.

Thus all kinds of molecules can potentially evolve to become used as pheromones in chemical communication. The size and stability of compounds can often be related to their role in communication. For example, ant alarm pheromones tend to be small, volatile molecules which diffuse quickly. Hyena territorial marking pheromones are high-molecular-weight involatile compounds, which will last a long time. A wide variety of the compounds are used by different species as pheromones, including steroids, fatty acids, and esters. Some aquatic invertebrates and vertebrates, including amphibians, use short peptides as pheromones. For example, the Asian red-bellied newt (*Cynops pyrrhogaster*) uses a decapeptide, sodefrin, as its male sex pheromone. The sea slug mollusk *Aplysia* similarly uses a short peptide.

Small molecules as well as large ones are used by animals large and small. However, terrestrial vertebrates may be different from other animals as their low-molecular-weight pheromones are often delivered loosely in the cleft of a protein, which increases the longevity of the signal. For example, mouse pheromones such as dehydro-*exo*-brevicomin are presented in association with lipocalin mouse urinary proteins (MUPs). Large quantities of MUPs are present in the urine of dominant male mice. The MUPs may also give some individuality to the signal of the male.

As it happens, a closely related compound *exo*-brevicomin is used as a pheromone by some bark beetles as well as male mice. Such a shared use is not unusual. The Asian elephant (*Elephas maximus*) shares its female sex pheromone with some 140 species of moth. The compound is a small, volatile molecule, (*Z*)-7-dodecen-1-yl acetate. This phenomenon comes about because of the common origin of life: basic enzyme pathways are common to all multicellular organisms. In addition, the number of potential different small, and thus volatile, molecules is limited by the number of permutations of combining a small number of atoms. However, moths and elephants are unlikely to be confused by each

other as female elephants produce prodigious amounts of the pheromone (and the amounts produced by a female moth will not be noticed by a male elephant) and male moths are attracted to species-specific combinations of pheromone components, of which the elephant pheromone is just one.

Most chemical communication, such as that of moths, is by mixtures of small molecules which give species specificity to messages. Female moths' pheromones tend to consist of a species-specific combination of fatty acids or alcohols. Males respond to the whole blend as a signal. Unique compounds such as the complex molecule periplanone, used as a sex pheromone by the American cockroach, are unusual, although animals using peptides as their pheromones can have species-specific sequences.

In moths we can see speciation reflected in their chemical communication. Related species may have similar pheromone blends, distinguished by different proportions of the compounds or the addition or loss of certain components. In some moth families, the synthetic pathways for the pheromone blend components are well understood. With this background, the variation in blends produced by different species can be explained by gene frequency changes and mutations, resulting in up- or downregulation of particular enzymes or changes in reaction product or substrate. There may be regional variation in pheromone blends within a moth species, and incipient speciation can sometimes be observed.

Mechanism of Olfaction, Receptor Proteins, and the Evolution of Pheromones

In animals with an olfactory system, pheromones are detected by the olfactory system rather than the gustatory system. The organization of the olfactory system in invertebrates and vertebrates is remarkably similar in functional terms. The great flexibility of the olfactory system means that almost any kind of chemical can potentially evolve into a pheromone if it provides useful information to the receiver. This is because olfaction works by stimulation of a wide range of relatively nonspecific G-protein-coupled receptor proteins (each expressed on different olfactory sensory neurons linked to glomeruli). The system can respond to the widest possible range of odors, even as entirely new compounds are created (in evolving biological systems). Although a given odor may be unlikely to fit any one receptor perfectly, it is likely to stimulate some. If detection of a molecule gives significant selective advantage, then over evolutionary time, the receptors may evolve great specificity to such molecules; so, for example, on the male moth antennae there will be large banks of sensory neurons, each with receptors very finely tuned to one of the key components of that species' female sex pheromone.

Many vertebrates have a second olfactory system, the vomeronasal organ, which projects to the accessory olfactory lobe, as well as the main olfactory system. Depending on species, pheromones are detected by either the main olfactory system or the vomeronasal organ or, as seems increasingly likely, both systems.

Releaser and Primer Pheromones

The first pheromones recognized were sex pheromones which elicited (or 'released') an almost instantaneous behavioral response. For example, in moths, female sex pheromones elicit orientation flights by males toward the female, and when he finds her, in many species, the male fans his wings as part of courtship. This formed the bioassay used to identify the pheromone (see the section titled 'Techniques for studying chemical communication').

Some pheromones produce longer-term physiological changes (primer effects) in the receiver. For example, the female goldfish sex pheromones, besides attracting males, also cause the receiving males to increase their sperm and semen production. Primer pheromones have their effect indirectly via the receiver's endocrine system. The distinctions between primer and releaser pheromones now seem to be less clear-cut than when first defined. There are many interactions between pheromones and the central and the peripheral nervous systems and endocrine systems. Some primer effects can be rapid, and the effects of some behavioral interactions mediated by pheromones can be long lasting.

Another class are modulator pheromones, which are proposed to affect mood in humans though the evidence is not yet strong for these, at least at physiologically relevant concentrations of the putative pheromones.

Odor Signatures versus 'Simple' Chemical Communication (Pheromones)

The pheromones above are defined molecules or mixtures of molecules. There are other situations, such as mate choice based on the odors associated with the immune system, where arguably there is chemical communication, but not of the traditional pheromone type. This is information in the form of an 'odor signature' which differs between individuals or family groups of the same species. This variability in odor

signature makes recognition of individuals or families possible by smell. The odor signature is very complex and is made up from all sources of variation, both genetic and environmental, including recent feeding history or infection. For odor signatures, the differences between individuals or clans are the message. (Some authors have termed these odors 'signaler pheromones.') The odor identity of the stud male mouse in the Bruce effect, in which the odor of a new male causes spontaneous abortion or reabsorption of the embryo in the female, is remembered by neuronal circuits in the accessory olfactory lobe after stimulation via the vomeronasal organ. In contrast, in the female sheep, the odor identity of her lamb is remembered in the main olfactory lobe.

An important feature of odor signatures is that, unlike species-specific pheromones, odor signatures all have to be learned – in mammals, for example, by growing up with litter mates (which are later recognized as kin – and avoided as mates when adult). In ants, the colony identity odor has to be gained from ants in the colony and learned on emergence from the pupa. The importance of learning is demonstrated by cross-fostering experiments, which show a sensitive period of imprinting in the life of many young mammals. Similar results are found with social insects.

The one theoretical exception to the need for learning is the Green Beard scenario, in which three linked genes produce (1) an external marker, (2) a genetic ability to recognize the marker, and (3) a genetically determined appropriate response to look after other individuals with this marker. However, there are only a tiny number of examples, notably in the fire ant and possibly in slime molds.

Production of Pheromones

Pheromones are secreted by a variety of specialized glands in insects and mammals. In the same way that compounds become co-opted as pheromones, glands are found in many different places in animals, usually as adapted skin glands. In mammals these occur in species-specific locations, for example near the eye, around the anus, or on the chin or flanks. Some pheromones are produced by internal organs; for example, the male sex pheromones of the sea lamprey, *Petromyzon marinus*, are bile acids produced by the liver.

In some species, the chemicals used for pheromone communication are collected; for example, male tropical euglossine bees must collect perfume compounds from orchid flowers to use as their sex pheromone. Some butterflies and moths – either as larvae or as adults – sequester plant poisons, and these may be adapted as sex pheromones in courtship (see the section titled 'Sex pheromones').

Bark beetle pheromones may include compounds ingested from the tree sap and then chemically modified by the beetle itself or by symbiotic bacteria in the beetle's gut. Pheromones (or signature odors) are produced by bacterial and fungal fermentation of fatty acids and other compounds secreted in the anal glands of dogs, badgers, and other carnivores.

Range of Roles for Chemical Communication

Pheromones are used for a wide range of functions by animals. As might be expected, they are used especially by nocturnal species. Pheromone signals may be combined with other modalities – for example, on detecting a predator, black-tailed deer signal alarm with both a pheromone and a visual tail-flick signal. Pheromones can be classified by their apparent uses or functions – though a given pheromone can have more than one role in a species, depending on context.

Sex Pheromones

The first pheromones identified were the sex pheromones of moths. It had long been known that male moths could find females at night; pheromones proved to be the means. At first each moth species was thought to have its own single unique compound as its pheromone. As more moth species were investigated, it was discovered that instead, each species had a unique blend of compounds as its pheromone. Different populations of the same species may have subtle variations of the blend, rather like a dialect. Goldfish sex pheromones have already been mentioned.

Pheromones can be the subject of sexual selection. Pheromones may give clues to the status or quality of the pheromone emitter. For example, in many rodents, only high-ranking individuals, with high testosterone titers, will produce the male pheromones.

In some moths which sequester plant poisons for defense, male pheromones are derived from their chemical store in a way which offers an honest signal of the quantity of defensive plant poison he will transfer to the female at mating.

Territorial Pheromones

A key use of chemical communication by terrestrial vertebrates is for the marking of territories. An advantage of pheromones is that they can 'shout' ownership when the owner is not in that part of the territory. Male house mice mark their territories extensively with urine streaks. In the house mouse,

only the dominant male will mark, and he defends this privilege vigorously should a subordinate male be rash enough to mark in the territory of the dominant male. Scrub-living antelopes mark their territories with orbital-gland pheromones, smearing secretions from the gland onto twigs at antelope head height. Some lizards mark their territories with leg gland secretions which have a double signal – a visual one (in the ultraviolet range, which lizards can see) to make it conspicuous from a distance and a volatile chemical signal detectable at close range.

Alarm Pheromones

Alarm pheromones offer a warning or a call for help and marshal attack. Social insects – ants, bees, wasps, and the unrelated termites, all show colony defense coordinated by alarm pheromones. Honeybees release an alarm pheromone when they sting an intruder such as a bear intent on robbing their nest of its honey. The alarm pheromone attracts other bees from the nest to join in the attack. This collective defense can be very effective.

Many freshwater fish species have an alarm pheromone released from skin glands when these are damaged by a predator's closing its jaws on the victim. Other fish detecting the alarm pheromone show escape behavior and greater vigilance.

The evolution of alarm pheromones is relatively straightforward to understand in social insects or other animals living in kin groups. It is harder to explain in fish, which appear to live in groups of unrelated animals. This remains a mystery though one suggestion may be that the alarm pheromones attract other predators, which interfere with the first predator and thus the fish in its mouth escapes.

Aggregation Pheromones

Some animals use pheromones to attract conspecifics. Bark beetles arriving on a tree use aggregation pheromones to attract the critical number of conspecifics needed to overwhelm the tree's defenses. There is an optimum population density for the beetles, below which the beetles do less well – an Allee effect. As beetle density gets greater than the optimum, the pheromone messages released change, and potential settlers are deterred and go to other trees instead.

Many marine animals such as barnacles respond to conspecific chemical cues to select the place to make their once-in-a-lifetime metamorphosis from planktonic larva to sessile adult glued to a rock. The presence of proteins left by adult barnacles indicates that the place is a good one to settle because members of that species have survived to maturity. It also ensures there will be other conspecific barnacles within reach for mating, which is important as these animals fertilize internally.

Externally fertilizing marine animals coordinate density in time as well as space by using pheromones for a chemical duet between males and females which ensures that both sexes release their eggs and sperm into the water at the same time and in the same place. Some of the most spectacular examples come from marine polychaete worms, which spawn in enormous numbers on particular nights, all coordinated by pheromones.

Eavesdropping and Propaganda

Broadcast signals can be 'overheard' or eavesdropped. This happens to pheromone communication in just the same way. Bark beetle pheromones are intended for other conspecifics, but predators have evolved great sensitivity to these compounds and can use them as kairomones to home in on calling bark beetles. Parasitic wasps can use the female moth's pheromones to track down the moths whose eggs they parasitize.

Some species of ant which take over the nests of other ant species use pheromones as allomones ('propaganda'), releasing large quantities of the alarm pheromone of their victim species. This causes the victim species to run about chaotically but leaves the invading species unaffected.

Bolas spiders and some orchids independently show some of the most spectacular uses of allomones to manipulate other species: bolas spiders synthesize the female sex pheromones used by moths and thereby lure male moths to their death. A single spider may produce the pheromones of different moth species over the course of an evening to match the flight times of those species.

Orchids similarly synthesize female bee pheromones to attract male bees – though in this case the deception is for pollination, not to catch prey. Male bees attracted by the apparently authentic pheromone, mistaking it for a female bee, attempt to mate with the orchid flower, pick up pollen, and then fly on to be fooled again, delivering the pollen to another flower of the same orchid species.

Techniques for Studying Chemical Communication

Chemical communication has proved more difficult than other modalities to study. The first reason is conceptual – our own worlds are so visual and auditory that we find it difficult to imagine the central role of chemical communication in the lives of most other organisms. The second has to do with the difficulties

of chemical analysis. Much chemical communication uses vanishingly small quantities of chemical released by the signaler and perceived, at very low concentration, by the receiver of the signal. The quantities can stretch the limits of human instruments for detection (often we can detect the signal only by using the output from the animal's own sense organs – as used in the electroantennogram, an electronic recording made from the antenna of a moth in response to chemical signals). The third challenge is the 'playback' of chemical signals. Organisms are very sensitive to precise molecular shapes of pheromones – even the chirality of molecules has to be correct. This presents real challenges to the skills of synthetic chemists. In contrast to studies of communication by sound, there is no equivalent for pheromones of simple playback of a signal from a tape recorder or a computer-generated artificial signal.

Apart from recordings from the sense organs of animals, a key requirement in the identification of molecules used for chemical communication is a bioassay. This can be behavioral, such as testing whether the candidate molecules elicit the same behavior as the putative pheromone (in testing potential pig sex pheromones, for example, the bioassay could be the lordosis (back arching and presentation behavior of the female pig) given in the presence of odors of the male pig (boar). Or the bioassay could be a physiological one, such as the quantity of semen produced by a male fish after exposure to putative female sex pheromones.

Techniques for visualizing brain activity are gaining ground as tools to investigate the processing of chemical communication. Fluorescing dyes stimulated by free calcium concentrations have been used to investigate brain activity in response to odors in a variety of animals, including honeybees, salamanders, zebra fish, and salamanders. Brain imaging using functional magnetic resonance imaging or positron-emission tomography has been used by researchers investigating chemical communication in vertebrates, including humans.

Exploiting Chemical Communication for Agriculture and Medicine

The importance of chemical communication in the lives of animals offers many opportunities for applied use in agriculture, forestry, and, potentially, medicine. These uses are now well established for insect pest control, but with a few exceptions, their use with vertebrates is only just beginning.

Mating disruption of pest moths in apple orchards by the release of synthetic female sex pheromone provides successful control of codling moth *Cydia pomonella* and other important pests of fruit. The synthetic pheromones seem to prevent males' and females' finding each other, so the females' eggs are not fertilized – and no caterpillars result. Pheromones are particularly useful where pesticide resistance has developed, as has occurred for many pests of cotton. Pheromones provide a way to get off the 'pesticide treadmill,' in which farmers find they have to use additional pesticides to control previously innocuous insects which become pests when their predators are killed by the earlier rounds of pesticides. The specificity of pheromones means that very small quantities can be used and that nontarget species are not affected. Unlike pesticides, pheromones leave predators unharmed, so these methods are compatible with biological control measures.

Pest vertebrates too might be controlled by pheromones. The sea lamprey *Petromyzon marinus* has ravaged the Great Lakes fisheries in North America. Two lamprey pheromones, the male sex pheromone and the larval pheromone, are being explored to see whether they could be used to reduce lamprey populations by selective trapping or diversion. Mice and other pest rodents use pheromones for communication, and these perhaps could be used to control them.

An understanding of priming pheromones could be useful in animal husbandry to, for example, advance puberty and reduce the postpartum period in pigs, and end-seasonal anestrus in sheep. Artificial insemination requires knowledge of when estrus is about to occur – the predictive lordosis response of the female pig is more reliable if they are sprayed with synthetic male pheromone (5α-androstenone, Boar Mate). Fish priming pheromones could be used in aquaculture to replace the individual hormone injections given to male fish to stimulate sperm production before semen collection.

Pheromones can be helpful in modifying the behavior of companion animals, too. A synthetic cat cheek gland pheromone (Feliway) is used to reduce the frequency of spraying by neutered tomcats.

Pheromones may have a role to play in the future control of insect vectors of human diseases, such as the bloodsucking reduviid insects which carry Chagas disease. Nematodes, many of which cause human disease, use pheromones for communication, which might provide a route to control.

Ever since the discovery of the Bruce effect in mice, there has been interest in possible pheromone communication in humans. Humans may have pheromones, but only one effect, on menstrual synchrony, has been well demonstrated. In the distant future, this might lead to discovery of a new kind of birth control.

Conclusions

We are at an early stage in the investigation of chemical communication. Of all the senses, olfaction, because of the chemical communication it enables, is probably going to give us the most surprises in the future.

See also: Agonistic and Affiliative Signals: Resolutions of Conflict; Communication Networks and Eavesdropping in Animals; Endocrinology of Animal Communication: Behavioral; Hormones and Behavior; Kinship Signals in Animals; Multimodal Signaling in Animals; Olfactory Receptor Genes: Human Loss During Evolution; Olfactory Receptors; Olfactory Epithelium; Pheromones in Mammals; Pheromones in Humans and Social Chemosignals; Sexual Selection and the Evolution of Animal Signals; Signal Transmission in Natural Environments.

Further Reading

Agosta WC (2002) *Thieves, Deceivers and Killers: Tales of Chemistry in Nature*. Princeton, NJ: Princeton University Press.

Baxi KN, Dorries KM, and Eisthen HL (2006) Is the vomeronasal system really specialized for detecting pheromones? *Trends in Neurosciences* 29: 1–7.

Dusenbery DB (1992) *Sensory Ecology: How Organisms Acquire and Respond to Information*. New York: Freeman.

Greenfield MD (2002) *Signalers and Receivers: Mechanisms and Evolution of Arthropod Communication*. Oxford, UK: Oxford University Press.

Hurst JL, Beynon RJ, Roberts SC, and Wyatt TD (eds.) (2008) *Chemical Signals in Vertebrates 11*. New York: Springer Press.

Müller-Schwarze D (2006) *Chemical Ecology of Vertebrates*. Cambridge, UK: Cambridge University Press.

Ngai J (2006) An extra dimension to olfaction. *Nature* 442: 637–638.

Shepherd GM (2006) Smells, brains and hormones. *Nature* 439: 149–151.

Wyatt TD (2003) *Pheromones and Animal Behaviour: Communication by Smell and Taste*. Cambridge, UK: Cambridge University Press.

Wysocki CJ and Preti G (2004) Facts, fallacies, fears, and frustrations with human pheromones. *Anatomical Record* 281A: 1201–1211.

Relevant Website

http://www.pherobase.com – The Pherobase: Database of Insect Pheromones and Semiochemicals.

Pheromones in Humans and Social Chemosignals

M K McClintock, University of Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Pheromone Definitions

In order to summarize and evaluate the research on human pheromones, it is essential to first clarify the distinction between pheromones and social odors. A pheromone is a chemical signal; it is secreted by one individual of a species and triggers a specific behavioral, neuroendocrine, or developmental response in another individual of that same species. This means that pheromones enable one individual to affect the behavior or physiology of another, and are therefore fundamentally social chemical signals. In addition, pheromones are required only in minute amounts to produce their functional effects, typically several orders of magnitude less than is required to produce the perception of an odor. The relatively minute amount of pheromones necessary to trigger a response is an essential, but often overlooked, characteristic of these chemosignals.

Pheromones were first defined in insect species, in which their behavioral effects are dramatic. For example, the female tobacco budworm moth releases a sex attractant pheromone that is carried downwind, creating a chemical plume. This sex attractant pheromone consists of a mixture of six different kinds of molecules, and only a few molecules are needed to trigger specific neural programs in the male moth, enabling him to fly upwind in a stereotyped search pattern to find the broadcasting female. As pheromones were discovered in other species, ranging from yeast to mammals, it became clear that pheromones serve diverse functions, coordinating behavior and physiology within a social group. Human pheromones are best understood as including different types that are distinguished by the myriad functional effects they produce.

Types of Pheromones

Pheromones are defined by their function, not by their chemical identity or by the receptor or neural systems that mediate their effects. They must be vagile (i.e., capable of being dispersed) either because they are volatile, airborne, or exchanged through physical contact with another individual or a communal surface. Four functional types of pheromones have been identified among all species, of which only three are found in humans.

Releaser pheromones operate quickly on the nervous system to trigger, within seconds to minutes, relatively stereotyped behavior patterns or sets of functionally related behaviors. A mixture of androstenone and androstenol (the common terms for the 16-androstene steroids, 5α-androst-16-en-3-one and 5α-androst-16-en-3-ol) is the releaser pheromone in domestic pigs, *Sus scrofa*. The boar produces androstenone in his saliva, and by shaking his head, flings his saliva onto the sow, triggering her virtually catatonic mating stance, enabling him to mount. Indeed, androstenone is sold in aerosol cans to pig breeders. Even this steroid alone is such a prepotent stimulus that, sprayed on the snout of the sow, the compound elicits the mating stance and allows for artificial insemination.

The suggestion that pheromones could create irresistible sexual attraction may be a good marketing strategy for commercial fragrances, but it is not based on any scientific evidence. On the other hand, several less extreme effects have been demonstrated in humans.

Primer pheromones produce long-term neuroendocrine changes that regulate endocrine systems or development in the recipient. That function must define and classify pheromones, rather than their chemical identities, is clearly illustrated by androstenol and androstenone, which function as both a releaser and a primer pheromone. These androstenes in the boar's saliva accelerate puberty in the immature gilt when she is exposed to the saliva over several days. As a primer pheromone, the androstenes from the boar accelerate her fertility and eventual pregnancy.

Modulator pheromones adjust ongoing behavioral responses to a rich set of social and physical stimuli. They also regulate psychological reactions to a particular context, but without triggering specific behaviors or thoughts. As a consequence, modulator pheromones do not have a prepotent effect on behavior. Rather, they function by changing sensitivity to specific types of stimuli, including their salience, emotional valence, and sensorimotor integration. In essence, they create a context that modulates more prepotent stimuli. In rats (*Rattus norvegicus*), male pheromones increase the intensity of the female's arched-back lordosis reflex elicited when the male mounts her and palpates her flanks. In contrast to the domestic pig, however, the males' modulator pheromone alone is not sufficient to trigger the female rats' mating posture.

Signaler pheromones indicate or 'signal' to other members of the same species attributes of the sender, such as identity, emotional state, reproductive status, or genetic information. The recipient may or may not act on this information and, therefore, in contrast to

releaser and modulator pheromones, signaler pheromones have the least prepotent effect on behavior. Perhaps the best example of an unequivocal modulator pheromonal effect are airborne compounds associated with the major histocompatability complex (MHC), a cell surface protein that uniquely identifies individuals. In mice, this information can be used for mate choice, in selecting another female with which to communally nurse young, and even in determining whether a male is indeed the sire of the pups carried *in utero* by a pregnant female.

Pheromones and Odors

Many pheromones are also odors at high concentrations; they stimulate the olfactory system to create a percept. Percepts have well-characterized functions different from those of a pheromone. For example, androstenol has an odor foul enough to increase depression in people living near pig farms. Nonetheless, at the minute levels at which pheromones operate within a species, none of the four types of pheromone relies on the functional properties characteristic of odors.

Odors have specific functional properties, and the absence of these particular properties can be used to distinguish pheromones from odors. One such property is the capacity to serve as a conditioned stimulus for taste aversion. When mice are made nauseous while an arbitrary odor is paired with their food, the odor becomes a conditioned stimulus for avoiding the taste of the food. In contrast, female pheromones that elicit male mating behavior are only minimally effective as conditioned stimuli for taste aversion. In contrast to odors, pheromones produce little sensory adaptation or stimulus generalization. Behavioral responses to odorants adapt relatively quickly. Pheromonal responses take much longer to adapt, if at all. For example, a sow living next to a boar adapts to his sight and sound, but not to androstenone, his mating pheromone. This behavioral finding is consistent with recent neurophysiological evidence that neurons in the mouse vomeronasal epithelium do not adapt with prolonged exposure to a specific pheromone.

Human Primer Pheromones

Ovarian Function

Women produce ovarian pheromones that regulate the timing of the preovulatory surge of luteinizing hormone (LH) in other women. When presented daily throughout two menstrual cycles, axillary compounds from the follicular phase of the menstrual cycle shorten the follicular phase and advance the preovulatory LH surge of recipient women. In contrast, daily exposure to periovulatory compounds lengthens the follicular phase and delays the preovulatory LH surge. These compounds have no effect on the duration of menstruation, nor on the life span of the corpus luteum, indicating that they affect the neuroendocrine mechanisms of follicular development and the timing of the preovulatory LH surge.

As predicted, axillary compounds from the follicular phase are reported to accelerate the pulses of LH in recipient women, while ovulatory phase compounds slow them. The compounds that serve as ovarian pheromones have yet to be identified. One candidate for the periovulatory pheromone is 3α-androstenol (5α-androst-16-en-3α-ol). It is detectable in female axillary secretions and slows the pulsatile release of LH in women.

The key demonstration of compounds from the axillae of women is the effect on the frequency of LH pulses, the length of the follicular phase, and the timing of the preovulatory surge of LH. A computer simulation of such phase-advance and phase-delay pheromones demonstrates that under specific conditions, ovarian synchrony can develop within groups of females exchanging pheromones. Indeed, menstrual synchrony has been reported among groups of women living together under specific conditions. The conditions under which social synchronization occurs include women: (1) who are living together in small groups, (2) who are close friends, (3) who share activities, (4) who are co-workers, particularly those with high job interdependency but low stress, or (5) who are mothers and daughters.

Other potential forms of the social control of ovulation in women are asynchrony, regularization of the cycle, or changes in the rate of ovulatory cycles. In studies in which menstrual synchrony has not been observed, these other manifestations of pheromonal interactions could be operating. In addition, studies in which menstrual synchrony did not occur include groups of women: (1) who are without the opportunity to synchronize (i.e., have too few concurrent cycles that are both consecutive and ovulatory), (2) who are lesbian couples with prolonged intense contact, (3) who are members of multiple groups, or (4) who are people with particular personality traits.

Although ovarian pheromones were the first human pheromones reported, they may not have been measured in the most significant context in evolutionary terms. From an evolutionary perspective, having many consecutive ovulatory menstrual cycles is a social artifact, not a species trait. Typically, because contraception was inadequate or not practiced, women in their fertile years were less likely to have menstrual cycles than they were to have birth cycles of pregnancy and lactation. Indeed, in agrarian and hunter–gather

populations, an ovulatory menstrual cycle occurs on average no more than three times in 2 years. This is also true for most mammals with spontaneous ovulatory cycles. Therefore, pheromones likely evolved to regulate ovulation in harsher environmental contexts, helping achieve or suppress an ovulation from an infertile state. Such contexts would be the initiation of ovulatory cycles after lactational anovulation, at the end of droughts, at the beginning of reproductive senescence, and at the initiation of puberty. For example, the presence of older sisters, who are undoubtedly ovulating, is associated with delayed puberty in girls.

In other animals, signals from breeding females regulate the fertility of other females in their social group. Indeed, breast-feeding pheromones from lactating women and their infants regulate follicular development in a context-specific manner, increasing the variability of the ovarian cycle in women exposed to the breast-feeding compounds. Only trace amounts of compounds are necessary to exert their effects, and they are effective without being consciously detected as an odor in a way that is associated with ovarian responses.

Primer pheromones from men also affect the timing of ovulation. Women in an all-women college, who are simply in social proximity with men more than three times a week, have shorter and more regular cycles than do women who see men less often. Indeed, exposure to an extract of men's axillary compounds accelerates the pulse frequency of LH, a precursor to the larger surge of LH that triggers ovulation.

Modulator, Not Releaser, Pheromones

Two steroids, α4,16-androstadien-3-one (common name androstadienone) and 1,3,5,(10),16-estratetraen-3-ol (common name estratetraenol), were originally claimed to function as human releaser pheromones that stimulate specific social cognitions. Androstadienone was purported to make women feel specifically more open to social interactions and at ease. Estratetraenol was purported to stimulate self-confidence in men. Subsequent behavioral and psychological data, however, do not support these strong claims for releaser pheromones on specific kinds of social cognition, but indicate instead that these steroids are likely modulator pheromones, which adjust mood responses already being driven by particular social interactions. It should be noted, nonetheless, that continued research may reveal that these steroids are vasanas, compounds with odorlike properties, but that they function below the threshold of conscious detection.

Androstadienone affects diverse aspects of psychological state, depending on the social and psychological context of the recipient. It improves women's mood, particularly in laboratory settings that typically dampen mood (moreover, this improved mood is sustained after returning to everyday life, outside the lab). These effects are obtained with exposure to less than 9 nmol of steroid passively inhaled and verified to be undetectable as an odor, and even when also masked by the strong odor of clove. Simply waving a swab soaked in a 250 μM solution briefly below the nose is sufficient to produce enduring effects. Moreover, similar psychological effects are not produced either by a structurally similar 16-androstene steroid or by muscone, which has a similar odor. Therefore, in minute amounts, androstadienone does not depend on odor properties for its functional effects on the mood of women. But mood is not the only psychological state which androstadienone modulates in women. Women also report feeling more focused and alert when answering computerized questionnaires, and even experience more physical pain in some circumstances. In other contexts, they report feeling more sensual.

The importance of context in determining how the effects of androstadienone are psychologically manifest is particularly salient when men and women interviewers follow the same experimental protocol. When women are in a testing room with a man, which itself modulates mood, androstadienone increases positive mood after the experience; with a woman interviewer, which does not alter mood, androstadienone has no effect. Likewise, when men and women are alone, there are no demonstrable effects of the steroids.

For women, estratetraenol, the 'female' steroid produced by the placenta in pregnant women, has similar, albeit weaker, positive effects on mood, indicating that these steroids do not function simply as sex attractants. In fact, women in the late follicular phase of their menstrual cycles respond more strongly and positively to both steroids than do men. Women feel less anxiety and are less 'down' than during exposure only to the carrier solution. In contrast, men actually feel worse after exposure to the steroids in the same testing environment. This is especially true in response to estratetraenol, directly refuting the hypothesis that estratetraenol might serve as a female sex attractant for men. Indeed, when men have a woman as a tester, which prevents an increase in negative mood during the experiment, estratetraenol exacerbates their negative mood, counteracting the effect of being in the presence of a woman.

When men and women actively sniff androstadienone presented in crystalline form in amounts large enough to have an odor, as well as crystalline estratetraenol, an odorless compound, both still function as modulator pheromones. When men are saddened by a tragic film, androstadienone intensifies their sadness. In women, it has the opposite effect, both ameliorating

the women's sadness and preventing the typical drop in happiness. During an erotic film, however, androstadienone enhances sexual arousal in both men and women. At high, as at low, concentrations, the effects of estratetraenol are similar to those of androstadienone, but weaker. Neither steroid modulates mood when people watch a neutral film. Thus at both low and high concentrations, these compounds consistently modulate emotional state created by other stimuli. Even though they do increase sexual arousal and sensuality in an erotic context, they are clearly not a 'male pheromone' that affects only women and a corresponding 'female pheromone' that affects only men – similar to the sex-specific compounds evolved as releaser pheromones in insects. Note that neuroimaging of hypothalamic activity in response to androstadienone in a neutral context does not follow this pattern, indicating that its effect on sexual arousal likely involves cortical as well as subcortical mechanisms.

Androstadienone and estratetraenol also increase skin conductance, a veridical marker of sympathetic nervous system tone, because the eccrine sweat glands are innervated only by the sympathetic, and not the parasympathetic, system. The increase in sympathetic tone is more pronounced in women, particularly during the fertile phase of their menstrual cycle. In men, the effect is an increase in parasympathetic tone, indicated by warmer hands in men and colder hands in women. The effect of context on these autonomic nervous system effects will help determine whether they play a direct role in modulating psychological reactions to a situation or if they are a consequence of that change in emotional state.

Both men and women produce androstadienone from a variety of sources. The ovary produces androstadienone, as does the testis. Moreover, androstadienone levels are not correlated with testosterone production in men. Both sexes synthesize the steroid in the adrenal. Indeed, it is synthesized in the axillae of some men and women by resident coryneform bacteria, rather than being excreted by apocrine glands. In contrast, estratetraenol is known only to be produced in the third trimester of pregnancy, synthesized from androstadienone in a single step. Although some research reports the effects of 'estratetraene,' this refers to a large class of steroids. Estratetraenol is the more precise common name, indicating the ketone at the third carbon position.

Human Signaler Pheromones

The MHC is a complex, highly polymorphic protein evolved to carry molecules from inside the cell and present them at the cell surface. In the development of the immune system, the MHC and the molecules it carries select out T cell populations that recognize 'self' and would otherwise cause autoimmune disease. When the T cells bind to this complex, recognizing the cell as 'non-self,' they trigger an immunological cascade, removing the foreign cell. It was first discovered in mice that this genetically determined information about individual identity, is associated with airborne compounds that males and females use to discriminate differences of a single allele encoding the structure of the MHC. Behaviorally, this genetic information is used not only for mate choice, but also for guided paternal care, nursing, and other social choices.

Humans also have the capacity to detect and utilize such information. Married couples in an isolated religious community are less likely than random to share more than 5 out of 10 human leukocyte antigen (HLA) alleles (the human MHC), suggesting that humans may choose marriage partners based on signaling pheromones. However, now that humans have been assessed in laboratory tasks, it is clear that perception of this genetic information and choices made are determined by the pleasantness of the body scent. Therefore, the compounds are potentially functioning as odors, and are not signaling pheromones.

When women are asked which odor they would choose if they had to smell it all of the time, without knowing that they were human scents, women choose, and report as more pleasant, the scent of a male with whom they share at least one MHC allele. In a more sexualized context, women report that the scents from a male with more allelic matches are more pleasant. And, ironically, in a neutral context focused only on odor quality, women are indifferent if the scents are from donors' MHC with more matches, and are more likely to find them neither pleasant nor unpleasant. In terms of human pheromones, these data indicate that even in animals, the MHC genetic information may not be a signaling pheromone at all; rather, the genetic information may simply affect the pleasantness of a consciously perceived body scent.

Routes of Action

In principle, therefore, there are several routes by which human pheromones could exert their effects. In vertebrates, pheromones are not defined by a specialized receptor organ, and exert their effects via the main olfactory epithelium, the vomeronasal system, and absorption through skin or mucosa. However, anatomical evidence for a specialized pheromone system in humans is currently ambiguous.

Main Olfactory System

There are genes for several hundred olfactory receptors; these receptors are G-protein-coupled receptors on dendritic cells in the olfactory epithelium and they weakly bind a spectrum of odorants. In larger amounts, in which androstadienone functions as a social odor, repeated sniff-training or exposure over several weeks increases sensitivity fourfold, in contrast to the stable intensity of nonsocial odors such as amyl acetate. The time course of the sensitization and an accompanying shift in the unpleasantness of the odor indicate that there are at least two receptors that are induced by repeated exposure.

Human primer and modulator pheromones may also act via the main olfactory epithelium, as do some other mammalian pheromones. In addition to genes for main olfactory receptors, there are also genes for two other chemoreceptors that are structurally distinct. In rodents, these are also expressed in the main olfactory epithelium and have the high sensitivity and narrow specificity typical of pheromone receptors and function.

The human genes *hV1RL1–hV1RL5* are similar to rodent pheromone *V1R* receptor genes. *hV1RL1* has an intact open reading frame and is expressed in human olfactory epithelium, although the full distribution of receptor cells is not yet established. *V1R* genes have become diversified in primates, suggesting a positive selection pressure for species-specific function, even as the complexity of the primate vomeronasal organ (VNO) has diminished over evolution. In humans, however, the olfactory epithelium does not express the gene for Trp2, a cation channel protein necessary for neural signal transduction in rodent V1R neurons. Thus, if functional, the transduction from chemical to neural information must take place through another mechanism.

The second family of potential pheromone or specialized social odor receptors in the human main olfactory epithelium includes trace amine-associated receptors (TAARs), also G-protein-coupled receptors. In mice, eight classes are expressed in neurons of the main olfactory epithelium. In cell culture, neurons expressing these receptors bind amine-based pheromones and social odors. Genome sequence data indicate that humans have five members in the TAAR family. It now needs to be determined whether message for these genes will be detected in humans, and whether they bind social chemosignals with primer or modulator pheromone functions.

Finally, transepithelial absorption and humoral effects on the brain are also possible. Androstenone (5α-androst-16-en-3-one), the boar primer pheromone, can act humorally, being absorbed through the mucosa and transferred from venous blood to the arterial supply of the brain. Humans, too, can absorb steroids and other molecules through the nasal mucosa and olfactory epithelium, where such molecules enter the blood supply. Additionally, some molecules are carried through retrograde transport along the olfactory nerve to the brain.

Vomeronasal System

Humans also have a VNO along each side of the nasal septum and an associated nasopalatine duct extending from the base of the septum toward the roof of the mouth. In many vertebrates, pheromones are transported through the nasopalatine duct and bind in the VNO at V1R and V2R receptors. Neurons from the VNO then project to an accessory olfactory bulb. In humans, however, there is no evidence that the nasopalatine duct is patent, except in dysmorphologies, neither V1R nor V2R receptors are expressed in the VNO and Trp2 is absent. Likewise, no myelinated neurons have been detected leaving the VNO and adults do not have a distinct VNO accessory olfactory bulb. Finally, occlusion of the vomeronasal duct does not affect the perception of androstadienone in amounts where it is a social odor.

In the human fetus, however, the vomeronasal structures are readily apparent and have cells positive for LH releasing hormone (LHRH) and neuron-specific enolase, indicating they are olfactory receptor cells. Likewise, the human fetus has an accessory olfactory bulb detectable until 26 weeks of age, but no bulb is found at the end of gestation. Thus, many have concluded that this structure only functions to guide neurogenesis of LHRH-containing neurons in the fetus.

A comparative perspective argues that the vomeronasal structures on the nasal septum may have evolved in catarrhine monkeys and apes to serve a function different from that of the VNOs of strepsirhine primates (e.g., lemurs). Although receptors and neural tissue are missing in humans and chimpanzees, in both species the septal tube is lined with ciliated, pseudostratified columnar epithelium, which may serve a glandular secretion and transport function.

Central Nervous System

Androstadienone modulates the state of the brain without being perceived consciously as an odor, changing glucose utilization in distributed systems associated with its affects on mood and attention. When women perform a simple visual–motor task during 20 min of glucose uptake, androstadienone increases glucose utilization in the primary visual

cortices (V1), anterior thalamus, and basal ganglia. Motor systems are modulated, decreasing glucose utilization (superior parietal cortex, premotor cortex, and cerebellum). Finally, even though there is no detectable effect on the primary olfactory system, there are changes in brain regions that receive projections from the olfactory system (inferior temporal lobe, hippocampus, amygdala, hypothalamus, and thalamus).

Even when androstadienone is detectable by women as an odor, and is presented briefly to them as they lie blindfolded in a positron emission tomography (PET) scanner, it does not preferentially activate the olfactory cortical regions affected by other odors. It does, however, alter blood flow in a variety of nonolfactory cortical areas, the most consistent being inferior parts of the prefrontal cortex and the superior temporal cortex. This is the opposite effect seen with the control odors, indicating that androstadienone functions differently even at higher concentrations.

Androstadienone may be processed differently than other odorants by an as yet unidentified secondary subcortical circuit, functionally similar to the secondary visual system that processes an emotionally salient social system. In amounts large enough to be detected as an odor, cortical processing of androstadienone is 20% faster than is processing of odorants that are perceptually similar in intensity and hedonic valence. Whether this is also the case for subthreshold amounts remains an open question.

One possibility is suggested by observations in *Maccaca irus,* an Old World monkey, that there is olfactory input to the lateral hypothalamus, which is more finely tuned to specific odors than is typical of the main olfactory system. This pathway does not pass through the main olfactory areas (prepyriform–entorhinal area or the olfactory tubercle), and instead passes through the septum and nucleus accumbens.

Tantalizing support for this hypothesis comes from PET studies of men and women, both heterosexual and homosexual. Crystalline androstadienone, presented in vials as an odor and without asking individuals to do anything particular other than cooperate with the PET scan, increases oxygen utilization in the medial preoptic area of the anterior hypothalamus of both heterosexual women and homosexual men. In parallel, crystalline estratetrenol affects the paraventricular nucleus of the anterior hypothalamus of heterosexual men and homosexual women. These findings have been presented as evidence for sexual pheromones, specifically linked to sexual orientation – androstadienone linked to attraction to men and estratetrenol linked to attraction to women, putatively serving as sexual pheromones. These data, however, are not consistent with the modulating effects on mood and attention when the component is not consciously perceived as an odor. An alternative hypothesis is that these crystalline steroids function as conditioning odors during sexual interactions, and are consequences, not modulators, of sexual motivation and orientation.

See also: Hormones and Behavior; Neuroendocrinology of Puberty; Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Olfactory Higher Centers Anatomy; Pheromones in Mammals; Sexual Behavior: Neuroendocrine Control; Vomeronasal System Evolution; Vomeronasal Accessory System.

Further Reading

Grus WE, Shi P, Zhang YP, et al. (2005) Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals. *Proceedings of the National Academy of Sciences of the United States of America* 102: 5767–5772.

Jacob S, Kinnunen LH, Metz J, et al. (2001) Sustained human chemosignal unconsciously alters brain function. *NeuroReport* 12: 2391–2394.

Jacob S, McClintock MK, Zelano B, et al. (2002) Paternally inherited HLA alleles are associated with women's choice of male odor. *Nature Genetics* 30: 175–179.

Johnston RE (1998) Pheromones, the vomeronasal system, and communication. From hormonal responses to individual recognition. *Annals of the New York Academy of Sciences* 30: 333–348.

Lundstrom JN, Olsson MJ, Schall B, et al. (2006) A putative social chemosignal elicits faster cortical responses than perceptually similar odorants. *NeuroImage* 30: 1340–1346.

McClintock MK (1998) Whither menstrual synchrony? *Annual Review of Sexual Research* 9: 77–95.

McClintock MK (2002) Pheromones, odors and vasanas: The neuroendocrinology of social chemosignals in human and animals. In: Pfaff D, Arnold A, Etgen A, et al. (eds.) *Hormones, Brain and Behavior,* vol. 1, pp. 797–870. San Diego, CA: Academic Press.

Savic I, Berglund H, and Lindstrom P (2005) Brain response to putative pheromones in homosexual men. *Proceedings of the National Academy of Sciences of the United States of America* 102: 7356–7361.

Smith TD, Bhatnagar KP, Shimp KL, et al. (2002) Histological definition of the vomeronasal organ in humans and chimpanzees, with a comparison to other primates. *Anatomical Record* 267: 166–176.

Sobel N, Prabhakaran V, Desmond JE, et al. (1998) Sniffing and smelling: Separate subsystems in the human olfactory cortex. *Nature* 392: 282–286.

Stern K and McClintock MK (1998) Regulation of ovulation by human pheromones. *Nature* 392: 177–179.

Witt M, Georgiewa B, Knecht M, et al. (2002) On the chemosensory nature of the vomeronasal epithelium in adult humans. *Histochemistry and Cell Biology* 117: 493–509.

Wysocki CJ and Preti G (2004) Facts, fallacies, fears, and frustrations with human pheromones. *Anatomical Record A: Discoveries in Molecular, Cellular, and Evolutionary Biology* 281: 1201–1211.

Relevant Websites

http://www.cf.ac.uk – Human Pheromones (TJC Jacob, Cardiff University, Cardiff Wales).
http://www.sfn.org – Society for Neuroscience: Pheromones.
http://bioportal.weizmann.ac.il – The Human Olfactory Receptor Data Exploratorium (Crown Human Genome Center and Weizmann Institute of Science).
http://www.pherobase.net – The Pherobase: Database of insect pheromones and semiochemicals.

Pheromones in Mammals

C J Wysocki and G Preti, Monell Chemical Senses Center and University of Pennsylvania, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Definition of Pheromone and Its Transition

The term pheromone was coined in 1959 by Peter Karlson and Martin Lüscher, who introduced the term as being

> derived from the Greek *pherin*, to transfer; *hormān*, to excite. Pheromones are defined as substances which are secreted to the outside by an individual and received by a second individual of the same species, in which they release a specific reaction, for example, a definite behavioral or a developmental process.

The introduction of this neologism was not accepted by all. HS Micklem, on reading the 1959 paper in *Nature*, responded with this letter to the editor of the journal, which published the original pheromone paper:

> Karlson and Lüscher[1] have recently proposed the term 'pheromone' for a class of substances which, while resembling hormones in some respects, cannot be included among them because they are not of an endocrine origin. It is claimed that the latter part of the proposed word is derived from the Greek *hormān* (which should properly be *hormōn*), meaning 'to excite'. This is not so. It was apparently the intention of the authors to find a term which, while being derivable from *hormān*, bore little overt resemblance to the word 'hormone'. This they achieved by the improper expedient of dropping the letter 'r', thus mutilating the root of the Greek word. It should be a principle, when a scientific term is derived from classical origins, that the term must by its structure be traceable to those origins. In the case of the word 'pheromone' that principle is not observed: there is no structural clue to its derivation from *hormān*. Indeed, the only Greek verb (apart from *pherein*) which might conceivably be involved in its ancestry means 'to swear'.

In the original definition, a pheromone was a single chemical compound or limited set of compounds that, when released by the sender and detected by the recipient (typically of the same species), would elicit a stereotypical response each and every time. Unfortunately, for most pheromonal responses, each of the components of the definition has fallen by the wayside. In insects, wherein much more is known about pheromones than for mammals, the complexity is enormous. For some responses the chemical stimulus is a bouquet, composed of a delicately balanced mixture of many compounds. In other situations, the pheromone crosses species lines. For example, the pheromone may be released by a predator which is sending out a mimic message – the sex-attractant of the female of the prey species. The male responds to the amorous message only to find itself in the jaws of its predator. In other situations, there may be considerable learning or context specificity, not at all what was in mind when the original definition was presented. Furthermore, there is considerable diversity in the type of message or response that can occur. Its origin aside, some scientists suggest that the term pheromone is redundant or too general, especially as it applies to mammals. Indeed, the word odorant, whether it has a special property, such as influencing physiological response, or providing general information, such as indicating the source of food or the appreciation of a floral fragrance, has been suggested to be fully sufficient. This argument, as usually stated, relies on the following logic: there are no unique terms in vision, audition, or somesthesia, so why use such a unique term in olfaction? Others have proposed that the definition of pheromone be limited to chemical communication wherein both sender and receiver benefit from such communication. Although logically legitimate, this restrictive definition may omit inclusion of pheromonal communication for informational purposes and may include chemical signaling that may not be pheromonal in nature. For example, mother rodents pass to offspring the safety of foods to eat by chemical signals originating in the oral cavity; the mothers eat a food, exhale into the nasal cavity of the offspring, and thereby inform them that the food is safe to eat. To this end, both sender and receiver benefit: the mother secures a safe future for her genetic investment (her genes continue to exist in the form of her offspring and, she might hope, the successors of her offspring), and the offspring learn what is safe to eat. In this context, it would be difficult to mold the odor of food into a pheromonal communication system in which both sender and receiver obtain benefits, although other aspects of this exchange appear to meet the criteria.

Among humans, pheromonal communication may exist as a language that is difficult to understand at the cognitive level. Research suggests that such information may be processed via subcortical pathways and hence may be influencing human responses much more than previously acknowledged.

Types of Pheromones and Typical Responses

The general scheme for mammalian pheromones posits odorants that have specific or selective effects on the physiology or immediate or potential outcome of behavior of recipients. In the present context, pheromone is a legitimate term. Of the four categories of pheromones in mammals, primer, releaser, signaler, and modulator, the last was described from research on humans before research on other mammals, which is an unusual outcome. Typically, research with nonhuman animals initially defines a pheromonal response, and later research extends the category to include humans.

Primer Pheromones

Primer pheromones have many effects that influence endocrine or neuroendocrine mechanisms. These include induced pregnancy failures, shifts in rates of sexual maturation, alterations in levels of circulating testosterone in males, and inhibition of or modifications to estrous and menstrual cycles in females.

Pregnancy At least in rodents (and perhaps in other mammals), primers determine the success of a pregnancy; for example, exposing a recently pregnant female mouse (within the first 5 days of 21-day gestation) to the chemical signals from a strange male mouse (a male that she hasn't yet met or mated with) results in failure of the pregnancy – the fetuses are reabsorbed and, for the female, physiology proceeds as if pregnancy did not occur. In some species of voles, this effect can be observed as late as day 18 of a 22-day gestation. Usually, pheromones from the male with whom she mated do not induce failure of pregnancy. This results from a memory trace of the male's chemical signals that is encoded within the vomeronasal system (in the accessory olfactory bulb; see below); however, the system can be manipulated. After mating, if the original male and his pheromones are removed and replaced by a novel male, the female's vomeronasal system encodes information about this second male. During the subsequent pregnancy, although he is not the impregnating male, the female's pregnancy is unaffected by exposures to this male's pheromones; however, pheromones from the impregnating male will induce pregnancy failure.

Sexual development Chemical cues from adults will influence sexual development in both males and females. In general, in a group of juvenile males in the presence of an adult (usually dominant) male, pheromones from the adult male(s) will inhibit sexual maturation of immature individuals. The inhibited males can and often do escape this inhibition once they migrate from the group to establish a territorial domain elsewhere. An analogous situation is true for females: dominant adult females inhibit the growth of reproductive potential in developing females. This has been found in New World primates – the scent-mark of a mother marmoset monkey, in the absence of her physical presence, can inhibit puberty in her daughters well past the anticipated date of its onset. It is interesting that in situations in which developing young in laboratory conditions are exposed to the chemical signals from adults of the opposite sex, puberty occurs earlier than in siblings not exposed to these chemicals. Whether this pheromonally induced early onset of puberty is a common occurrence in natural populations outside the laboratory is a topic for future research; however, there has been speculation that this phenomenon may occur in humans – young girls enter puberty earlier if they live in a household with a nonbiological father.

Hormone shifts In adulthood, pheromones are known to affect endocrine systems: exposing adult male rodents to the chemical signals from novel adult females results in a surge in luteinizing hormone (LH), as measured by changes in LH-releasing hormone (LHRH). The increase in LHRH subsequently, and quite quickly, increases the levels of circulating testosterone. Presumably, this activates sexual motivation in males and tends to facilitate reproductive behavior. In females a similar situation occurs. Pheromones from novel members of the species (including humans) can influence the pulse cycle of LHRH (as inferred by measures of LH), which may affect the onset of ovulation and/or disrupt normal cycling. Depending on the source of the chemical signals, ovulation may be advanced (by exposing the female to chemical signals from adult males or from females in the preovulatory phase of their cycle) or retarded (by exposing the female to chemical signals from other females in the postovulatory phase of their cycle). In the extreme, when females live among a group in which there is a dense population of females, estrous cycling tends to cease. Exposing a singly housed female rodent to bedding from a cage of group-housed females is sufficient to disrupt cycling. The estrous cycling can be restored by exposing the female(s) to pheromones from adult males. This apparently overrides the endocrine block induced by group housing. In some mammalian species, such as the prairie vole, females do not have an estrous cycle. To become physiologically receptive, the female must first be exposed to a male – that is, to

his pheromones – before she can mate. These chemical cues prime the reproductive axis, for instance by increasing LH, and facilitate successful mating.

Releaser Pheromones

Releaser pheromones trigger specific, often stereotypical, behavioral responses. These include, but are not limited to, selective interest in chemical signals from the opposite sex, sexual attraction, precopulatory communication between male and female (e.g., ultrasonic calling in mice), induced mating postures in females and actual mating behavior, territorial marking and sexual receptivity, intermale aggression, maternal aggression, maternal-like behavior in nulliparous females, and at least in rabbits, nipple attachment and suckling.

Investigation of sexual cues and sexual attraction In many species, pheromones will support active investigation of chemical cues from the opposite sex. For some, the mere investigation of the cues appears to provide positive reward, and a feed-forward loop is established. The reward component, at least in guinea pigs, is supported by inputs through the vomeronasal system. These sexual attractants appear to be the vanguard of myriad sexual signals, many of which may be nonpheromonal, employed to ensure successful mating; however, in some species reliance on pheromones is critical – typically, for males, the absence of pheromonal inputs precludes actual courtship and mating behaviors.

During encounters with other conspecifics, especially the initial meeting of a male and a female, much can be learned, even to the point that synthetic odorants can acquire pheromone-like properties. In one experiment, male mice were exposed to female mice to which Coty Wild Musk cologne had been applied. During this first meeting with an adult female and in all subsequent encounters with females, this group of male mice smelled the cologne in the presence of the female. In subsequent tests with the cologne, in the absence of the females, males treated the Coty Wild Musk as a natural female-associated pheromone and engaged in courtship-related behavior, viz., ultrasonic calling, in its presence. Other groups of males, not having the specific pairing of cologne with females, did not treat the cologne as a pheromone. This rapid and enduring odorant-specific associative learning becomes important when trying to understand apparently conflicting results of experiments in which mechanisms of pheromonal detection were explored.

Sexual attraction, the first pheromonal response to be described in moths, is also the most discussed among humans. A Google search with two terms, viz., 'pheromone human,' revealed over 2 million websites; most offered human pheromones for sale, primarily targeting males. These products typically claimed enhanced success at attracting females; however, caveat emptor is totally relevant (see the section titled 'Releaser pheromones' under 'Evidence for pheromones in humans').

The first mammalian pheromone to be identified chemically was from boars. Found in high concentrations in boar saliva, androstenone (5α-androst-16-en-3-one), and its alcohol derivative, androstenol (5α-androst-16-en-3-ol), when presented to sexually/physiologically receptive sows, elicits a specific postural response, that is, lordosis, from the sow. In lordosis the sow is stationary and the back is arched, allowing mating to progress. Other specific chemical compounds that are active mammalian pheromones have since been identified (see Table 1) albeit none in humans, as discussed in the section titled 'Evidence for pheromones in humans.'

Territorial marking and sexual receptivity Many animals with a home range will actively scent mark their territory, thereby providing a chemical signal of territorial occupation that transcends time. Same-species visitors to the marked territory subsequently can extract information from the chemical signals and may avoid crossing the boundary (to minimize risk of an aggressive encounter), may ignore the territorial mark altogether, or may overmark in an attempt to establish their own territory or dominance. In analogous fashion, both males and females of many species, especially those with limited mating seasons, will advertise sexual receptivity by depositing chemical signals, such as urine or glandular secretions, at discrete locations visited by the opposite sex. For example, moose bulls will dig a shallow pit with their antlers, urinate in the pit, roll in it to become immersed in their odors, and advertise their readiness to mate to receptive females. Cows also engage in such advertisement in a mating arena, urinating profusely at the perimeter of the arena and calling to elicit approach by a bull.

Aggression Both intermale aggression and maternal aggression toward intruders at the nest containing young rely on pheromonal communication. Typically, males, especially dominant males, will attack another male. Attacks often are not directed at females; however, if a female is painted with urine from a male, she often will elicit attack. Females with offspring will defend their young by engaging in maternal aggression directed at intruders. Such behavior relies on the female's recognizing the intruder, usually a male, as such. This is often accomplished through pheromones.

Table 1 Identified pheromones in mammals

Source	*Class*	*Effect*	*Chemical compound(s)*
Male mouse urine	P and R	Attracts female mice; induces estrus; stimulates intermale aggression	2-*sec*-Butyl-4,5-dihydrothiazole 2,3-dehydro-*exo*-brevicomin
Male mouse urine	P and R	Attracts female mice; induces estrus; stimulates intermale aggression	α- and β-Farnesenes
Male mouse urine	P	Accelerates female mouse onset of puberty	Isobutyl amine; isoamyl amine
Male mouse urine	P	Accelerates female mouse onset of puberty	6-Hydroxy-6-methyl-3-heptanone
Female mouse urine	P	Accelerates female mouse onset of puberty	2,5-Dimethylpyrazine
Hamster vaginal secretion	R	Attracts male conspecifics	Dimethyl sulphide
Hamster vaginal secretion	S	Signals estrus	Methyl thiobutyrate
Cat male and female urine	S	Suggested as territorial and sexual status markers	3-Mercapto-3-methyl-1-butanol, 3-mercapto-3-methylbutyl formate, 3-methyl-3-methylthio-1-butanol, and 3-methyl-3-(2 disulfanyl)-1 butanol
Milk from lactating rabbit and her ventrum	R	Seek and seize nipple	2-Methylbut-2-enal
Sebaceous glands around nipples of nursing female cat	M	Suggested as a mixture to tranquillize kittens and adult conspecifics in stress situations	Oleic acid, palmitic acid, linoleic acid, myristic acid, lauric acid, stearic acid
Sebaceous glands around nipples of nursing dog	M	Suggested to have tranquilizing effect on puppies and adult conspecifics in stress situations	Oleic acid, palmitic acid, linoleic acid, myristic acid, lauric acid, pentadenoic acid, stearic acid
Vaginal fluid of estrous female dog	R	Stimulates male sexual arousal and sexual behavior	Methyl *p*-hydroxybenzoate
Boar testis (also found in saliva and other sources)	R	Stimulates lordosis in estrous sows; stimulates oxytocin release in estrous sows; reduces agonistic behavior in piglets; induces feeding behavior in piglets	5α-Androst-16-en-3-one
Boar testis (also found in saliva and other sources)	R	Stimulates lordosis in estrous sows	5α-Androst-16-en-3α-ol; 5α-androst-16-en-3β-ol
Sow, sebaceous glands around milking sow nipples	P	Reduces agonistic behavior in piglets; induces feeding behavior in piglets	Palmitic acid, oleic acid, linoleic acid, lauric acid, myristic acid, capric acid
Ram fleece and anteorbital gland secretions	P	Stimulates secretion of LH in and emission of urine from ewes	1,2-Hexadecanediol, 1,2-octadecanediol, and (unknown) acid compound(s)
Goat, buck fleece	P	Stimulates LH secretion in female conspecifics	4-Ethyl octanoic acid derivative(s)
Asian elephant, facial temporal gland secretions, urine and breath of bulls	R	Stimulates mating-related behavior in follicular-phase female conspecifics	1,5-Dimethyl-6,8-dioxabicyclo [3.2.1] octane (frontalin)
Asian elephant, female urine	R	Stimulates sexual behavior in male conspecifics	(*Z*)-7-dodecen-1-yl-acetate
African elephant, male urine	S and P	Signals reproductive status to female conspecifics	Butanone (elevated level during premusth), acetone and 2-pentanone (elevated levels during musth), and 2-nonanone (elevated level during late-musth)
African elephant, female urine	S and P	Signals reproductive status to male conspecifics	Frontalin, *exo*-brevicomin, α-farnesene and β-farnesene

M, modulator; P, primer; R, releaser; S, signaler.
Adapted from Van den Hurk R (2007) *Intraspecific Chemical Communication in Vertebrates with Special Attention to Its Role in Reproduction*. Zeist, The Netherlands: Pheromone Information Centre.

Without the ability of the female to detect these pheromones, maternal aggression is eliminated, demonstrating the import of these chemical cues.

Other maternal behaviors and nursing At least in rats, chemical cues associated with young pups will elicit maternal-like behavior from nulliparous females. Here, too, without the ability to detect these pheromones, females do not demonstrate this maternal-like behavior.

Young pups use pheromonal cues to locate nipples on the mother. This is readily demonstrated in

rabbits. In this species, it is important for the offspring to find and secure a nipple – there is but a single daily feeding. To facilitate feeding, the ventrum of the doe and her milk contain a nipple-search pheromone. The chemical identity of this pheromone has been described (Table 1).

Signaler Pheromones

This class of pheromones is the broadest and perhaps the most controversial. Some have objected to inclusion of this class because it appears to allow virtually all forms of chemical communication. It is included here because significant information about the sender can be gleaned from chemical cues alone. Information contained in chemical cues includes, but is not limited to, identification of conspecifics; sex; age; status of health; social dominance; pregnancy; level of stress of the sender; family membership; and individual identity, as conferred by genes within the major histocompatibility complex (MHC), which is a set of >50 genes that regulates the immune system. The MHC is perhaps the most variable set of genes in mammals, having many different alleles at myriad loci. Indeed, variation is so extreme that within natural populations, no two individuals (except identical twins) are thought to have identical MHC haplotypes. This indicates that an MHC-related odor profile is unique to each individual, rendering an individual-specific odortype. Recognition of MHC type influences choice of mate; for example, mice typically will choose to mate individuals possessing an MHC type that differs from their own. Experiments have shown that subtle changes within the MHC, such as a single nucleotide polymorphism, can be detected in the respective odortypes of the donors. It is interesting that the odortypes of fetuses 'leak' into and become infused in the odortype of the pregnant female. At least under experimental control, test animals can use this information to indicate paternity of the fetus by matching the fetal signal contained in the mother's odortype to the odortype of the biological father. Perhaps this MHC-related odortype is used by females to identify their biological offspring as early as the day of birth, as illustrated by the ability of a ewe to identify her lamb by smell.

Modulator Pheromones

This class of pheromones was first introduced in 2000 by Martha McClintock on the basis of her studies of responses of humans to chemical signals from other people. It is currently unknown whether analogous pheromones exist in nonhumans; however, based on the existing literature, data suggest that similar communication exists among nonhuman mammals. The most prominent example comes from the results of experiments involving the chemical cues from animals placed in mildly stressful situations. For example, mice temporarily housed in cages whose floors can deliver mild electrical shock can alter their body odor to reflect this stressful situation, as can be measured in responses from mice exposed to the chemical signals from the stressed mice: information about the stressful situation is conveyed in chemical signals to the other animals, which will then avoid the environment associated with stress-inducing stimuli.

In addition to avoiding the location where stressed animals were situated, the recipient animals experience a rise in their cortisol levels, perhaps indicating elevated levels of stress. This perhaps represents a crossover of primer and modulator effects. Similar changes in cortisol levels have been reported in rodents after they are exposed to odors from predators. This is not to suggest that there is cross-species pheromonal communication among mammals, but rather it illustrates how multiple input pathways can affect similar responses.

Evidence for Pheromones in Humans

Primer Pheromones

Often, women who live together tend to synchronize menstrual cycles. Originally studied and described by McClintock, this phenomenon and follow-up studies describing shifts in the onset of ovulation or length of the menstrual cycle on exposure to chemical signals obtained from other women or men, or changes in LH pulsing on exposure to chemical signals from men, offer the best evidence for primer effects in humans. Experiments performed in the 1980s in the laboratory of George Preti suggested that women with irregular menstrual cycles became more regular, for example, closer to a 28-day cycle, after being exposed to chemical cues contained in collections of axillary secretions from men. Axillary secretions from other women tended to shift the menstrual cycle of recipient females in the direction of synchrony. In the McClintock lab, the ovulation-related surge in LH was advanced and the entire cycle shortened in women receiving odorant stimulation from axillary secretions derived from other women who were in the early, follicular phase of their menstrual cycle. Axillary secretions collected from the same women, but at the stage of ovulation, elicited the opposite response in recipient females. In addition to the surge in LH that is related to ovulation, LH release throughout the menstrual cycle exhibits a pulsatile characteristic, albeit at much lower levels. In follow-up work in the Preti lab, the onset of the next LH pulse occurred

more rapidly in women after they were exposed to axillary secretions from men relative to the onset of the LH pulse after exposure to a placebo.

Women who breastfeed their infants, or the infants themselves (or both the women and the infants), appear to generate yet another pheromone that influences recipient, nulliparous women. Relative to women who received placebo treatments, women with a regular sexual partner reported increased sexual desire after exposure to chemical signals collected from the breast area and axillae of lactating women; women without a sexual partner reported many more sexual fantasies.

Releaser Pheromones

As noted above, considerable emphasis has been focused on the sale of sex attractants, usually for men (although some products target women). Many of these products contain androstenone or androstenol, true pheromones, but for pigs. The first pheromone to be identified chemically was determined to be bombykol, the releaser emitted by female silkworm moths to attract males. The identity of the compound was determined by systematically reducing the number of potential candidate chemicals by repeatedly taking apart the complex mix of chemicals emitted by the female moth. By measuring odorant-induced upwind flights by male moths, the investigators were able to determine which subcomponent contained the active ingredient. In an iterative process of taking apart the subcomponents and testing the different parts in a wind tunnel, the investigators were able to identify the critical chemical. It is important to note that there is no publication in the biomedical literature that relies on such a bioassay-driven or response-driven approach to identify a human pheromone, even in its crudest form, such as axillary secretions, which contain hundreds of chemicals.

With one exception, none of the releaser effects discussed above have been implicated in human behavior. The exception is displayed by neonates. Babies will orient in the direction of a breast pad containing chemical cues from the lactating mother. Furthermore, neonates old enough to exhibit controlled movement will crawl in the direction of the breast pad, demonstrating attractant properties. The babies' respective mothers were used as a source of the pheromone, so it is not yet known whether breast odor from another lactating woman would elicit such behavior.

Signaler Pheromones

There are numerous demonstrations of these types of pheromones in humans, and some of the accounts may have overlapping underlying mechanisms. Not all purported situations can be summarized, but the well-documented examples are included.

Mothers can recognize their newborn infant by smell alone shortly after delivery (but fathers cannot). This difference may result from exposure of the mother's olfactory receptor cells to circulating odorants generated by the fetus (olfactory receptors can be activated by odorants contained in blood), which may give the mother many months of exposure during which she potentially could learn the smell of her offspring (but, as noted above, the mother's odor changes too in response to the fetus's cells and immune response).

Individuals can recognize members of their family by smell, which may be based on MHC relatedness or result from years of experience with the odor of family members. In a related discrimination, both men and women can infer underlying MHC type by sniffing clothing worn by individuals; they typically choose the donor who is least similar to the MHC type of the evaluator. It is interesting that the McClintock group members have extended these findings. They have noted that among women, preferences for MHC-related chemical cues are related to the MHC genes they inherited from their father, and they choose chemical cues from donors having some, but not all, alleles inherited from their father. These preferences also are unrelated to the MHC genotypes to which they have been exposed, via MHC-related odorants of the parents, throughout their life, but did not inherit.

Axillary odor from men and women can be discriminated, indicating that information about biological sex is contained within the axillary chemical cues. Recent findings suggest that the ability to discriminate, using the sense of smell, lies in qualitative, rather than quantitative (intensity), differences in axillary odor between men and women. Sexual orientation also influences the production of these axillary cues and how people respond to them. The chemical cues from the four sources, heterosexual men, heterosexual women, gay men, and lesbians, could be discriminated one from the other, at least by members of one or more of the four groups of evaluators. Furthermore, in two-alternative forced-choice tests, in which participants were asked to indicate which sample they preferred, the only participants to consistently choose the gay samples when they were pitted against axillary odors from the other sources were gay men; all other groups chose the alternative sample, that is, that from heterosexual males or lesbians. In the forced-choice comparison between samples from females or heterosexual males, the only group to show a preference was gay males; they uniformly rejected the samples from heterosexual males and chose samples from lesbians.

Modulator Pheromones

More data in this area exist for humans than for any other mammal. For the most part, unfortunately, only half the story has been told. If people are exposed to situations that elicit stress or happiness, body odor changes, presumably to signal the emotive component of the situation. These differences can be discriminated in olfactory-based tests administered to other people. Although few people are able to identify whether the odor originates from a stressed or from a happy person, they can tell that the samples are different. What remains to be determined is whether individuals who are exposed to the chemical signals from a target person who has an altered emotional status also show a shift in their emotional status to mirror that of the target.

In another report, the emotional status of donor males was not manipulated. Rather, chemical cues from the axillae of these donors were presented to females. The hormonal status of the females was monitored, and the participants were asked to rate their mood at selected times during the experiment. Relative to placebo exposures, the chemical cues from males induced a more relaxed and less tense status in heterosexual females, suggesting that there may be a male pheromone that modulates female mood and emotion.

Detection of Mammalian Pheromones

Within the mammalian nose reside six chemical detection subsystems (at least for non-Old World primates and with exceptions for many marine mammals). These include the following:

1. Olfaction *per se*, which receives input from millions of bipolar receptor neurons residing within the main olfactory epithelium (MOE) and projecting axons to the main olfactory bulb (MOB).
2. The vomeronasal system, which expresses bipolar receptor neurons within the tubular vomeronasal organ (VNO), projects axons to the accessory olfactory bulb, and expresses genes for molecular receptor proteins, viz., V1Rs and V2Rs. These genes differ from the large family of olfactory receptor genes (ORG) found in the olfactory epithelium.
3. The septal organ, also with bipolar receptor neurons projecting to the MOB and expressing a subset of ORGs.
4. The Grueneberg ganglion, a small collection of bipolar cells found in the nasal vestibule and projecting axons to the MOB. Many of these neurons express molecular receptors from the vomeronasal family, and some cells express molecular signatures of MOE neurons.
5. Chemesthesis, supported by afferents of the trigeminal nerve.
6. The nervus terminalis, which during fetal development gives rise to most forebrain LHRH-containing neurons that migrate from the nasal cavity to take up residence in their respective destinations.

A final possible type is indicated by the modified glomerular complex and associated necklace glomeruli located at the posterior margin of the MOB surrounding the accessory olfactory bulb. This subsystem is associated with a distinct chemical identity and appears to be activated by suckling behavior in infant rats.

With respect to detection of pheromones, there is no functional evidence to suggest involvement of the septal organ, the Grueneberg ganglion (although it is intriguing that its neurons express receptors from both the vomeronasal and olfactory systems), trigeminal afferents, or elements of the nervus terminalis; however, considerable evidence implicates both olfaction and the VNO in the detection of pheromones. It is important to note that both systems are known to respond to common, nonpheromone odorants.

Two Major Inputs

Most mammals have both olfactory and vomeronasal inputs to the central nervous system; however, the latter appears to be nonfunctional in Old World primates, including humans. This does not mean that Old World primates do not maintain pheromonal communication, as many primary publications and Internet sites suggest. The logic employed in the argument to refute pheromone communication in Old World primates relies on the false premise that pheromones are detected by the VNO and anything that activates receptors in the VNO must be a pheromone. This view held sway for nearly 2 decades, and for good reason; experiment after experiment showed that removal or destruction of the VNO eliminated responses to pheromones. Slowly, however, evidence began to emerge that the VNO was not critical for every pheromonal response. For example, sows can respond to androstenone even though their VNO is rendered nonfunctional; sheep do not need the VNO to detect their lambs; suckling rabbits do not rely on input from the VNO to respond to the doe's cues; and as noted above, sensory cells in the VNO respond to typical odorants not known to be pheromones. It is important to recall that although lacking a functional VNO, humans respond to pheromones. The accumulating evidence strongly suggests that the MOE is involved in the detection of pheromones.

Olfaction has multiple pathways as chemosensory information leaves the MOB. Some pathways may

support cognitive processing of odor whereas others, perhaps subcortical, may support pheromonal responses. Inherently, there is no a priori reason why pheromones should stimulate only neural networks that support unconscious processing of information; pheromones may have primer, releaser, signaler, or modulator effects while at the same time stimulating cognitive appreciation of the smell of the adjacent source of the pheromone. Aah, the chemistry of love.

See also: Hormones and Behavior; Neuroendocrinology of Puberty; Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory Cortex Physiology; Olfactory Higher Centers Anatomy; Pheromones in Humans and Social Chemosignals; Pheromones and other Chemical Communication in Animals; Sexual Behavior: Neuroendocrine Control; Vomeronasal System Evolution; Vomeronasal Accessory System.

Further Reading

Baxi KN, Dorries KM, and Eisthen HL (2006) Is the vomeronasal system really specialized for detecting pheromones? *Trends in Neurosciences* 29: 1–7.

Brennan PA and Binns EK (2005) Vomeronasal mechanisms of mate recognition in mice. *Chemical Senses* 30(supplement 1): i148–i149.

Fleischer J, Schwarzenbacher K, Besser S, Hass N, and Breer H (2006) Olfactory receptors and signalling elements in the Grueneberg ganglion. *Journal of Neurochemistry* 98: 543–554.

Halpern M (1987) The organization and function of the vomeronasal system. *Annual Review of Neuroscience* 10: 325–362.

Karlson P and Lüscher M (1959) 'Pheromones': A new term for a class of biologically active substances. *Nature* 183: 55–56.

Restrepo D, Lin W, Salcedo E, Yamazaki K, and Beauchamp G (2006) Odortypes and MHC peptides: Complementary chemosignals of MHC haplotype? *Trends in Neurosciences* 29: 604–609.

Shepherd G (2006) Smells, brains and hormones. *Nature* 439: 149–151.

Varendi H and Porter RH (2001) Breast odour as the only maternal stimulus elicits crawling towards the odour source. *Acta Paediatrica* 90: 372–375.

Van den Hurk R (2007) *Intraspecific Chemical Communication in Vertebrates with Special Attention to Its Role in Reproduction*. Zeist: Pheromone Information Centre.

Wyatt TD (2003) *Pheromones and Animal Behaviour: Communication by Smell and Taste*. Cambridge, UK: Cambridge University Press.

Relevant Website

http://neuro.fsu.edu – Meredith, M. The vomeronasal organ.

Phobia and Human Evolution

A Öhman, Karolinska Institutet, Stockholm, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Fear, Phobia, and Anxiety Disorder

Fears and Phobias

Fear is a normal emotion anchored in mammalian defense systems. It denotes an activated, aversive emotional state that serves to motivate attempts to cope with events that provide threats to the survival or well-being of the organism. The coping attempts are typically centered on defensive behaviors such as immobility (freezing), escape, or attack. Such defense behavior has an evolutionary origin and typically is structured according to the closeness of the threat. For example, distant threats elicit immobility (freezing), which provides camouflage and gives time for scanning the environment to evaluate the threat and potential escape routes. If the threat is closer (e.g., is interpreted as a predator preparing an attack), the likely response is flight and, if necessary, fight. These different behavioral responses are associated with different patterns of metabolically supporting autonomic and somatic activity, which may be used to index fear.

Even though basically adaptive, intense fear may sometimes interfere with normal adaptive functioning. A phobia is an excessive, involuntary, and rationally unfounded fear, which provokes maladaptive avoidance of a specific object or situation. The sufferer typically has insight into the irrationality of the fear given the objective danger involved but nonetheless is unable to voluntarily control the fear and its consequences. However, it is not the intensity or irrationality of the fear in itself that is decisive for a diagnosis of phobia. Rather it is the negative impact on the person's life by the phobic avoidance that qualifies it for the diagnosis. As long as the phobic object or situation can be successfully avoided without interfering with the goals of the individual, the fear does not provide a significant impediment to normal function. For example, an intense fear of flying may be of little consequence for an individual until he or she is offered an attractive job that requires frequent air travel. Then the fear comes in conflict with important goals, and the individual may seek professional help and be diagnosed with flight phobia. Nonetheless, there is a general, albeit not a perfect, correlation between fear intensity and avoidance. Thus, we can meaningfully talk about different thresholds for common fear, intense fear, and phobia, with many more people having strong irrational fears than those showing dysfunctional avoidance qualifying for a diagnosis of phobia.

Types of Phobias

In the fourth edition of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV), phobias are classified among the anxiety disorders, which also include panic disorder, posttraumatic stress disorder, generalized anxiety disorder, and obsessive–compulsive disorder. Phobias are classified into categories depending on the feared object or situation.

Specific phobias concern circumscribed objects or situations in nature (e.g., heights, water, lightening, thunder, darkness, loud noises, or sharp objects), animals, (e.g., snakes or other reptiles, worms, spiders, insects, birds, dogs, cats, or horses), blood-injection-injury (seeing blood or injured people, taking injection, or dental treatment), situations such as enclosed spaces, elevators, bridges, flying, and a miscellaneous category (e.g., fear of vomiting).

Social phobias concern social situations, typically those having an element of social evaluation, such as dating, speaking in front of a group, and meeting unknown (particularly high-ranking) people, and in general, situations involving a risk of social humiliation.

The third category, agoraphobia, is not a separate diagnostic category but is diagnosed in relation to panic disorder, primarily because the fear can be construed as primarily concerning panic attacks (fear of fear). Thus, people diagnosed with agoraphobia avoid a broad set of situations (e.g., public transports, supermarkets, bridges, tunnels, and driving on freeways) in which they lack an escape route to a safe base and, hence, would be left helplessly vulnerable in the event of be a panic attack. Accordingly, they may be able to confront the feared situations if accompanied by a trusted individual.

Phobias are prevalent mental disorders with the estimated lifetime risks in the general US population for each specific and social phobia being approximately 12%. Most cases of agoraphobia are included in the 4.7% lifetime prevalence of panic disorder, but there are also 1.4% of agoraphobia without panic disorder.

Phobic Responses

Typically, phobics are hypervigilant for their phobic stimulus and automatically scan the environment for its actual presence or signs that it may be imminent.

When avoidance fails and the feared situation has to be confronted, people with phobia show a distinct fear response. Subjectively, they report feelings of intense fright and a strong urge to get out of the situation. Behaviorally, they show signs of withdrawal and flight, as well as a facial expression of fear, and this behavioral defense mobilization recruits metabolic support by strong activation of the autonomic nervous system. For example, spider and snake phobics looking at pictures of their feared animals show cardiovascular activation in the form of quick heart rate accelerations (**Figure 1(a)**) and increases in blood pressure (**Figure 1(b)**).

When frightened, we are likely to startle to any unexpected stimulus, such as when a movie audience jumps because of a sudden noise and a dramatic cut after a suspense-building film sequence. Accordingly, people with animal phobia show startle enhancement to brief noise stimuli presented against a background of pictures of their feared animals, whereas nonphobic people tend to show an inhibition of the startle response when viewing potentially fear-relevant animals (**Figure 1(c)**).

Interestingly, phobic responses can be elicited by stimuli that are not consciously perceived. For example, if pictures of snakes or spiders are presented very briefly (15–30 ms) and are immediately followed by a masking stimulus, the perceivers are aware only of the latter. Nevertheless, people with phobia show enhanced psychophysiological responses specifically to the masked presentations of their phobic stimulus. A person with snake phobia, for example, shows an enhanced response to a masked snake but not to a masked spider (**Figure 1(d)**).

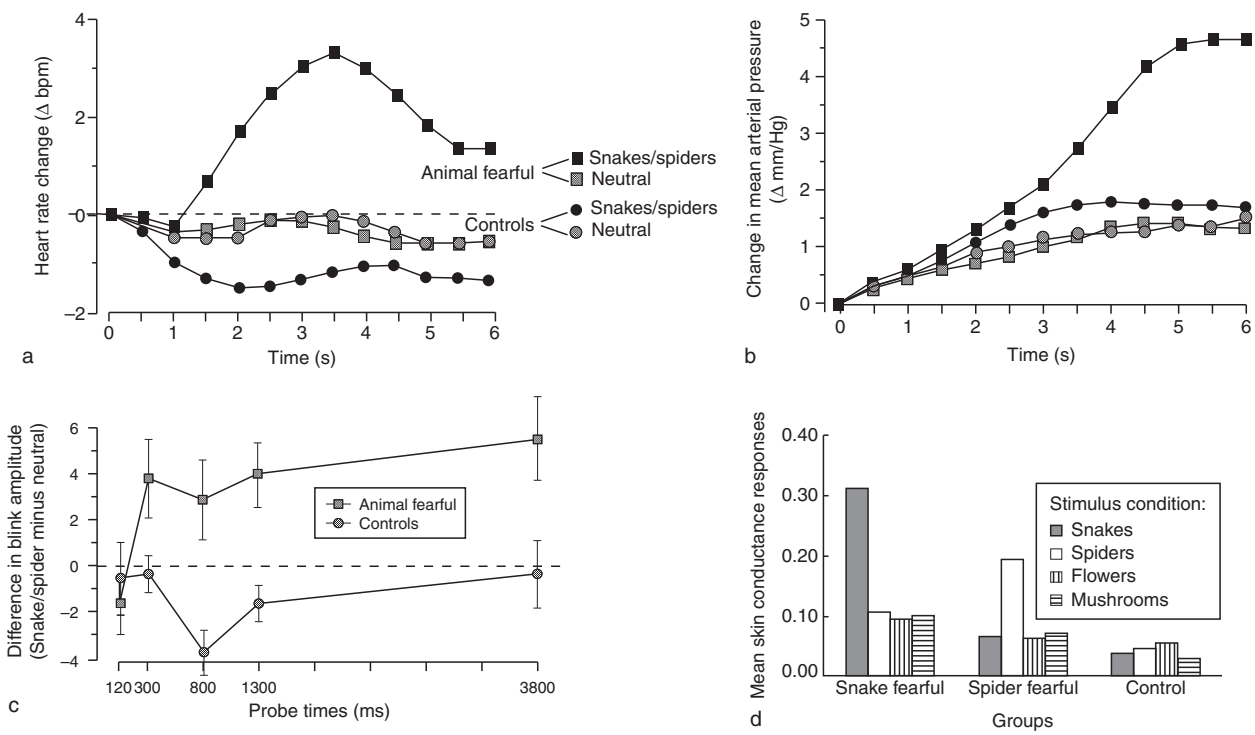

Figure 1 Responses of participants highly fearful of snakes or spiders to pictures of phobic and nonphobic content: (a) heart rate responses as changes from prestimulus level in participants selected as being highly fearful of snakes and spiders during repeated 6-s presentations of pictures of their most feared animal and or neutral control pictures; (b) blood pressure data from the same experiment; (c) difference in startle blink magnitude to brief white-noise stimuli presented at different stimulus-onset asynchronies during viewing of animal and neutral stimuli for animal fearful and control participants; (d) Skin conductance responses in participants selected to fear either snakes or spiders to masked pictures of snakes, spiders, flowers, and mushrooms. In (a), note the rapid acceleration of heart rate to the feared animals in animal fearful participants, which stand in stark contrast to the deceleration of the heart to animal stimuli in nonfearful controls; the two groups did not differ in their response to neutral pictures. In (c), note the potentiation of startle blinks to feared stimuli and the inhibition of blinks to neutral stimuli. In (d), note the elevated responses to the masked feared stimlulus. (a–c) Reproduced from Globisch J, Hamm A, Esteves F, and Öhman A (1999) Fear appears fast: Temporal course of startle reflex potentiation in animal fearful subjects. *Psychophysiology* 36: 66–75, with permission from Blackwell Publishing. (d) Reproduced from Öhman A and Soares JJF (1994) "Unconscious anxiety": Phobic responses to masked stimuli. *Journal of Abnormal Psychology* 103: 231–240, with permission from the author.

Neural Mechanisms of Fear

The neural basis of the distinct fear response shown by people with phobias when confronted with their feared object is relatively well understood. It is generated by a neural circuit centered on the amygdala, a collection of nuclei in the anterior medial temporal lobe. The central nucleus of the amygdala has efferent connections to areas in the diencephalon, midbrain, and brain stem that mediate different components of the fear response. The autonomic changes (Figures 1(a), **1(b)** and **1(d)**, for example, are controlled by pathways to the lateral hypothalamus and startle potentiation via connections between the amygdala and brain stem nuclei that control the startle reflex (Figure 1(c)). There are also pathways from the amygdala to the bed nucleus of the stria terminalis, the locus coeruleus, and the basal nucleus of the forebrain that result in the release of corticotropin-releasing hormone from the hypothalamus, activation of noradrenergic pathways, and activation of acethylcholinergic pathways, respectively. In concert, these effects activate large parts of the brain, including the cortex. Instrumental behavior (avoidance and escape), finally, is controlled by pathways to the striatum.

The amygdala receives input primarily through its lateral nucleus, which is connected to all sensory areas of the cortex as well as to the frontal lobes. In addition, there is direct input from the thalamus to the lateral amygdala, via the medial geniculate nucleus for auditory input and via a pathway through the superior colliculus and the pulvinar nucleus of the thalamus for visual input. Presumably through this direct pathway, amygdala may tune visual processing areas of the brain, such as the V1, V2, V4, and inferotemporal cortex. The lateral nucleus projects to other amygdala nuclei such as the basolateral and central nuclei. Because backward masking interrupts visual processing in the primary visual cortex, the subcortical pathway via the superior colliculi and the pulvinar may explain why masked visual stimuli can activate autonomic responses in the absence of awareness of the target stimulus (Figure 1(d)).

Early brain imaging studies of specific phobias showed strong activation of the secondary visual cortices in response to video sequences of phobic stimulation, and both activation (e.g., of the orbitofrontal cortex, anterior insula, and anterior cingulum) and deactivation (e.g., of the dorsolateral prefrontal cortex) of frontal areas but little activation of the amygdala. However, more recent studies using repeated presentations of short-duration (or masked) stimuli have consistently reported amygdala activation to phobic stimuli.

The Genetics of Phobia

The central adaptive function of fear makes it a likely product of evolutionary contingencies operating throughout mammalian evolution (roughly the last 100 My). Accordingly, both the patterns of defense behavior and its underlying neural circuitry have been evolutionarily preserved among mammals. Recent molecular work has isolated a specific gene, stathmin, which is richly expressed in the amygdala and in associated input structures of the fear network in mice and which appears to regulate both innate and learned fear.

Behavioral genetics studies of human twins provide important information about the genetic and environmental bases of the relationships within the phobias, as well as the relationships between phobias and other anxiety disorders. Studies of irrational fears and phobias in large samples of twins show that there is a common genetic factor, reflecting moderate heritability, which accounts for familial resemblance in phobias (animal, situational, blood-injection-injury, social, and agoraphobia). Specific phobias loaded more strongly than agoraphobia and social phobia in this factor. It is likely to reflect something that is common for all phobias, that is, genes coding for avoidance behavior. There are also specific factors reflecting genetic influences for each of the phobias. Although there appears to be little shared environmental influence on phobias, the data suggest that there are substantial unique environmental effects that are both general to phobias and specific to each of them.

There is substantial overlap, revealed as extensive comorbidity, between phobias and between anxiety disorders more generally. Analyses of this comorbidity across different anxiety disorders among twin pairs suggest that specific phobias reflect a unique genetic factor, whereas agoraphobia showed higher loadings in a general anxiety factor (encompassing panic disorder, generalized anxiety disorder, and posttraumatic stress disorder). Social phobia, on the other hand, had an intermediate position by showing loadings in both the specific phobia and the general anxiety factor.

Taken together the genetic findings suggest that specific phobias can be viewed as reflecting a distinct genetic factor, of which one component is shared with the other phobias, whereas another component is unique to each of the specific phobias (animal, situational and blood-injection-injury). These results suggest that specific phobias is the most typical fear disorder and probably reflects excessive activation of the amygdala fear circuit as a response to phobic stimulation.

The Evolution of Fears and Phobias

The functional nature of fear and the evidence suggesting the partly genetic origin of specific phobias imply that fear and phobias are likely to have a joint evolutionary basis. In this perspective, phobias can be viewed as prototypes of fear responses that were favored by natural selection because they provided defenses to important survival threats. These threats are likely to have a major origin in predatory pressure, which is a central force in evolution because of its direct effect on survival and procreation. Indeed, a recent theory on the forces that shaped the evolution of the primate brain gives a central role to the quintessential phobic stimulus – snakes – as an important predator and thus an evolutionary factor shaping the evolving primates. This predatory pressure integrated the expanding system for visual analysis that characterizes the primate brain with the sharpening of the fear circuitry centered on the amygdala that is a common feature of mammalian brains.

According to Lynne Isbell, predatory pressure from snakes affected primate evolution at two critical stages. First, constrictor snakes played an important role in establishing the primate order, and second, venomous snakes provided a critical pressure behind the expansion of the visual system and its integration with the fear system in the anthropoid primates (i.e., monkeys and apes).

Molecular evidence suggests that the ancestors of today's placental mammals appeared in Gondwanaland (the southern supercontinent that incorporated current South America, Africa, Madagascar, India, Australia, and Antarctica) more than 100 Ma. The early molelike mammals had to make their way in a world of reptiles. Indeed, of all the predators of primates that exist today, constrictor snakes were the first to evolve. Raptors and placental carnivores that preyed on the emerging placental mammals appeared considerably later (50–80 Ma). Thus, it is likely that the neural circuitry for defense behavior (the amygdala with associated input and output circuitry) originally was designed to deal with snakes and other reptiles rather than with attacking birds or felines. The to-become primates were, like their mammalian relatives, small and nocturnal with relatively poor vision and were dependent on olfaction for finding their way in the world.

Snakes with a very potent venom delivery system appeared in Asia or Africa about 60 Ma. An effective venom constituted a decisive step favoring snakes in the predator–prey arms race with the small primates, thus putting the burden on the latter to come up with more effective defenses. Venomous snakes are often camouflaged and difficult to detect, concealed in vegetation on the ground or in trees while waiting for prey, and their venom remained a threat even for animals too large to be prey. This put a strong and continuing premium on a visual system that rapidly could detect and activate defense to partly hidden snakes.

Interestingly, Madagascar was already separate from Africa (and Asia of course) when venomous snakes evolved. Consistent with the notion that the threat provided by venomous snakes played a critical role in stimulating primate evolution, all African anthropoids uniformly show fear of snakes, whereas the only the primates of Madagascar, the lemurs, do not. Venomous snakes migrated into South America from the north perhaps as late as 3 Ma with the formation of the Panama-bridge, whereas primates reached South America earlier, presumably directly from Africa. Accordingly, the New World monkeys appear to be more variable in their reaction to snakes than the Old World monkeys.

During the Eocene (roughly 55–35 Ma) the predation pressure from venomous snakes modified the primate visual system that provided input to the fear circuit centered on the amygdala. Because snakes are hard to detect, there was a development of greater orbital convergence, allowing better short-range stereopsis, particularly in the lower visual field. The koniocellular pathway from the retina was further developed to connect the superior colliculus with the dorsal pulvinar nucleus of the thalamus in assisting visual attention and, particularly, the preattentive detection of snakes and other fear-relevant stimuli. Furthermore, it assisted the parvocellular pathways to the lateral geneculate nucleus in allowing trichromacy and superb foveal vision, which may have served to facilitate detection and awareness of snakes on the ground or hidden in foliage. Higher-order visual areas such as parietal cortex and superior temporal sulcus, as well as the dorsolateral prefrontal cortex, also expanded for the same reason.

In addition, the increased reliance on sophisticated vision resulted in olfaction becoming less important and a shift from nocturnality to diurnality. These developments continued into the Oligocene (35–25 Ma) for the Old World anthropoids, but was disrupted for the New World anthropoids who presumably radiated through South America in the absence of venomous snakes. As a consequence, New World monkeys not only show more variable snake fear but they also have more variable visual systems. Thus, in addition to articulating decisive stages in primate evolution, Isbell's theory explains differences in the brains of different primates such as lemurs and Old and New World monkeys in terms of differential exposure to venomous snakes.

The brain that emerged from these evolutionary contingencies had much more cortex than that of nonprimate mammals, most of it devoted to vision.

Furthermore, the refined visual systems were integrated with the amygdala and its associated defense circuitry to the extent that cortical processing of visual input was tuned by the amygdala. Whereas snakes served as the most important evolutionary vehicle, the resulting changes were useful for discovering and defending against more diverse threats. Effective attention deployment and astute perceptual mechanisms, partly relying on rapid subcortical mechanisms, played key roles in the quick identification of threats, which permitted early defense mobilization and provided an adaptive edge in many situations. When anthropoids developed an extended social life, these defense mechanisms could be moved into the social arena, promoting vigilance for subtle social cues when coping with social contingencies and conflicts, a scenario that may help explain the important role of the amygdala in contemporary social neuroscience.

The Role of Learning in the Development of Phobias

The evolutionary scenario gives a perspective on how brain systems behind phobias evolved, and it implies a necessary role for genes in the development of phobias. But whether genes are not only necessary but also sufficient to account for phobias is a contentious issue. Some investigators argue that phobias originate from innate fears, reflecting an unhappy combination of alleles that are sufficient to produce intense fear and avoidance of particular objects or situations. Indeed, many phobias concern evolutionarily relevant dangerous situations, such as snakes, heights, and being confined in closed spaces. Furthermore, there are developmental fears shown by most children at particular ages, such as separation fears, stranger fears, and fears of loud noises and darkness. But it remains difficult to claim that all varieties of phobic objects or situations have their origin in evolutionary contingencies. Phobias are seen to a large and diverse, but still restricted, set of situations focused on potential threat in an evolutionary rather than contemporary perspective. The evidence for the innate view is primarily negative, reflecting the frequent failures of individuals diagnosed with different phobias to recall specific learning episodes as the start of their fear. The recall failures are hardly surprising given that most specific phobias start in early childhood.

Pavlovian Fear Conditioning as a Paradigm for the Acquisition of Phobias

There is a long tradition of understanding phobias as learned through Pavlovian fear conditioning, in which a stimulus that is in itself relatively innocuous (the conditioned stimulus (CS)) comes to signal the occurrence of a threatening stimulus (the unconditioned stimulus (US)). By associating cue to consequence, organisms may prepare themselves for the consequences by recruiting defense behavior in anticipation of the US (e.g., an attacking predator). Thus, Pavlovian fear conditioning favors prey animals in the evolutionary arms race between predator and prey, making this learning process a ubiquitous phenomenon in animal life that emerged early in evolution.

A fear-conditioning perspective on phobias proposes that a phobic stimulus acquires its fear-eliciting potency by association with other dangerous or traumatic events. For example, a dog phobia may have its origin in a childhood dog attack. However, there are several problems with this account. First, it does not explain the relative selectivity of phobias. Furthermore, it implies that phobias should primarily be seen as being to objects and situations that are threatening in a contemporary, rather than an evolutionary, perspective. Second, not all dog phobics recall being attacked by a dog, nor do members of their family. Therefore, direct Pavlovian conditioning may not be the only route to phobia. Other possibilities included observational learning (the dog phobic may have witnessed another person being attacked by a dog, in which case the fear evoked by witnessing this events may become conditioned to dogs) or verbal conditioning (e.g., from admonitions about the dangers of dogs from anxious parents).

The neural basis of fear conditioning is relatively well understood. Basically, the associative link between the CS and the US is formed in the lateral nucleus of the amygdala by *N*-methyl-D-aspartate (NMDA)-mediated potentiation of glutaminergic synaptic connections in a convergence zone of CS and US pathways from the thalamus. In this way, the CS may gain access to some of the US efference to elicit a plethora of fear-related changes mediated by the central nucleus of the amygdala, which is often seen in individuals diagnosed with phobia.

Evolutionary Influences on Pavlovian Fear Conditioning

In the natural environment, it is not any sound or smell that is likely to warn a potential prey of a lurking predator, but sound or smells produced or emitted by the predator itself. Such systematic ecological relationships between the CS and the US could be used by evolution to facilitate fear conditioning to stimuli that have provided recurrent threats in mammalian evolutionary history. Thus, the product of the evolutionary processes already discussed need not necessarily be an innate snake fear; it could be an evolved predisposition to very easily associate fear with snakes and similar threats.

This hypothesis has been confirmed by Susan Mineka and co-workers. Initially they demonstrated that wild-reared rhesus macaques at the Wisconsin Primate Research Center all showed a strong fear of snakes. Lab-reared monkeys, however, did not show snake fear until they were given the opportunity to observe a model wild-reared monkey exhibiting fear to a snake or a snakelike stimulus. For example, different groups of monkeys were exposed to videotapes in which a model monkey exhibited identical (through video splicing) fear responses either to a snake or to a flower. Only animals who had seen the model's fear response ostensibly elicited by snakes showed fear when subsequently exposed to a real snake. In contrast, seeing the same fear response ostensibly elicited by a flower did not produce a fear of flowers. Similar data have been provided from an extensive series of human conditioning studies by Öhman and co-workers. They also demonstrated that fear can be conditioned to masked stimuli, provided that the stimuli are fear-relevant (snakes or angry faces), which may help to explain failures to remember the onset of phobic fear, as well as the more general dissociation between fear and cognition in phobia.

These demonstrations suggest that evolution, rather than programming innate fears to specific stimuli, has generated predispositions to easily associate fear and avoidance to stimuli that are fear-relevant from an evolutionary perspective. Thus, not only are snakes the prototypical phobic stimulus, but phobias result from conditioning to a large but nonrandom set of situations that have in common that they involve some degree of threat (or, in some cases, a lack of escape route), and more from an evolutionary than a contemporary perspective.

Extinction and Exposure Treatment of Phobias

Pavlovian-conditioned fear responses wane if the CS is repeatedly presented in the absence of the US. This phenomenon is referred to as extinction. Conditioning and extinction are separate processes. In fact, the effect of evolutionarily prepared stimuli such as snakes or angry faces in humans typically is more evident in elevated resistance to extinction than in rapid conditioning. Thus, what makes phobia qualify as pathology is the persistence of the conditioned fear. Whereas normal conditioned fear is likely to extinguish though natural encounters with the CS, phobias resist such random exposures, perhaps because scrupulous avoidance of the phobic stimulus protects the fear from extinction.

However, once people with phobias can be persuaded and coached to confront their feared object or situation, the fear most often subsides. It is possible to treat specific phobias with accessible phobic objects (e.g., small animals or injections) in one intense sessions in which the therapist models, goads, and supports the patient to systematically approach the phobic object until it can be touched and handled. Indeed, there is an extensive research literature supporting the efficacy of different forms of exposure therapies for all types of phobias.

See also: Amygdala: Contributions to Fear; Anxiety Disorders; Aversive Emotions: Genetic Mechanisms of Serotonin; Aversive Emotions: Molecular Basis of Unconditioned Fear; Emotion: Neuroimaging; Extinction: Anatomy; Fear Conditioning and Synaptic Plasticity ; Genetics of Human Anxiety and Its Disorders; Panic Disorder; Panic Disorder as an Emotional Disorder; Pharmacology of Fear Extinction.

Further Reading

Carlsson K, Peterson KM, Lundqvist D, Karlsson A, Ingvar M, and Öhman A (2004) Fear and the amygdala: Manipulation of awareness generates differential cerebral responses to phobic and fear-relevant (but non-feared) stimuli. *Emotion* 4: 340–353.

Davey GCL (ed.) (1997) *Phobias: A Handbook of Description, Treatment, and Theory.* Chichester, UK: Wiley.

Davis M and Whalen PJ (2001) The amygdala: Vigilance and emotion. *Molecular Psychiatry* 6: 13–34.

Fanselow MS (1994) Neural organization of the defensive behavior system responsible for fear. *Psychonomic Bulletin and Review* 1: 429–438.

Globisch J, Hamm A, Esteves F, and Öhman A (1999) Fear appears fast: Temporal course of startle reflex potentiation in animal fearful subjects. *Psychophysiology* 36: 66–75.

Hettema JM, Prescott CA, Myers JM, Neale MC, and Kendler KS (2005) The structure of genetic and environmental risk factors for anxiety disorders in men and women. *Archives of General Psychiatry* 62: 182–189.

Isbell LA (2006) Snakes as agents of evolutionary change in primate brains. *Journal of Human Evolution* 51: 1–35.

Kendler KS, Myers JM, Prescott CA, and Neale MC (2001) The genetic epidemiology of irrational fears and phobias in men. *Archives of General Psychiatry* 58: 257–265.

LeDoux JE (1996) *The Emotional Brain.* New York: Simon & Schuster.

Maren S (2005) Building and burying fear memories in the brain. *Neuroscientist* 11: 89–99.

Mineka S (1992) Evolutionary memories, emotional processing, and the emotional disorders. *Psychology of Learning and Motivation* 28: 161–206.

Öhman A (2005) The role of the amygdala in human fear: Automatic detection of threat. *Psychoneurendocrinology* 30: 952–958.

Öhman A and Soares JJE (1994) "Unconscious anxiety": Phobic responses to masked stimuli. *Journal of Abnormal Psychology* 103: 231–240.

Öhman A, Dimberg U, and Öst L-G (1985) Animal and social phobias: Biological constraints on learned fear responses. In: Reiss S and Bootzin RR (eds.) *Theoretical Issues in Behavior Therapy*, pp. 123–175. New York: Academic Press.

Öhman A and Mineka S (2001) Fear, phobias and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review* 108: 483–522.

Seligman MEP (1971) Phobias and preparedness. *Behavior Therapy* 2: 307–320.

Photolysis of Caged Glutamate for Use in the CNS

G C R Ellis-Davies, Drexel University College of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Caged glutamate is proving an invaluable photochemical tool for neuroscientists. Neurotransmission at central synapses is extremely fast (~10 ms), so rapid solution changes in intact preparations, such as acutely isolated brain slices, are clearly impractical. Thus, the application of a biologically inert, photolabile pro-glutamate (i.e., caged glutamate) to living cells gives neuroscientists the ability to produce a step increase in glutamate concentration that has a similar time course to the release of glutamate by secretion from a single vesicle into the synaptic cleft (as judged by a comparison of spontaneous miniature excitatory postsynaptic current (mEPSC) with evoked EPSCs in cultured hippocampal neurons). In this article, we outline (1) the design criteria for development of these probes, (2) the structures and mechanistic properties of each of the different chromophores that has been used for caging, and (3) the thermal stability and photochemical properties of all caged glutamates.

Design Criteria

There are four important design criteria for the development of any caged neurotransmitter: (1) it must be very stable at neutral pH, (2) it must be efficiently photolyzed on irradiation, (3) it must release glutamate quickly, and (4) it must be inert before photolysis. Thermal stability of the caged glutamate is an important property because it is known that α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and *N*-methyl-D-aspartate (NMDA) receptors in the central nervous system (CNS) rapidly desensitize. If a 10 mM solution of caged glutamate is applied to a living cell which has as little as 0.5% free glutamate due to spontaneous hydrolysis, then 50 μM free glutamate is present. Caged compounds must be photolyzed with light of wavelength greater than 300 nm because DNA and proteins absorb light in the ultraviolet (UV). Further, if uncaging is effected through the microscope objective, then the caging chromophore should absorb light (the extinction coefficient) efficiently at 340–350 nm because lenses (and other microscope glass) are opaque below this range (some microscope have the rare quartz optics that passes light at 300 nm). If two-photon excitation from a mode-locked Ti:sapphire laser is used for uncaging, then 720-nm light is typically used, leaving longer wavelengths for two-photon fluorescence microscopy (800–950 nm range). The absorbed light must be also used efficiently by the photolytic process (this is the quantum yield for uncaging), otherwise phototoxic doses of light are required to release sufficient glutamate to stimulate receptors. Finally, the caged glutamate must be inert before photolysis. This last criterion is always considered to mean biologically inert, but it ought also to mean fluorescently inert. The prevalent use of fluorescent Ca probes and green fluorescent protein (GFP) technology in neuroscience places an additional design constraint on the caging chromophore; ideally the caged glutamate ought not to be fluorescent in the visible range.

Structures and Mechanistic Properties of Glutamate Caging Chromophores

Sixteen different caged glutamate molecules have been synthesized. The structures and acronyms are shown in Figure 1. Four different classes of chromophores are illustrated: (1) nitroaromatic (CNB-glu, *N*-Nmoc-glu, Noc-glu, DMCNB-glu, DMNPE-glu, DMNB-glu, MNI-glu, NI-glu, MDNI-glu, and antMNI-glu-glu), (2) coumarin (Bhc-glu and DEAC-glu), (3) quinoline (BHQ-glu), and (4) phenacetyl (pHP).

In terms of the photochemical mechanism of uncaging, these chromophores may be grouped into two simple categories: those that use photosolvolysis mechanisms for glutamate release and those that use intramolecular photoredox reactions. All the nitroaromatic-caged glutamates use some type of the latter mechanism for glutamate uncaging, whereas coumarins, quinolines, and phenacetyl caged glutamates used photochemical hydrolysis mechanisms. It must be noted, in the case of Bhc-glu, Noc-glu, and *N*-Nmoc-glu, that photolysis does not directly release glutamate. A second, pro-glutamate carbamate derivative is photoreleased by these three probes, which must undergo thermal aqueous hydrolysis of the carbamate bond (i.e., loss of CO_2) to yield the amino acid. (A detailed discussion of these mechanisms is beyond the scope of this review.)

Aqueous Stability and Purity of Caged Glutamates

Glutamate has been caged at all three of its functionalities (Figure 1). The two carboxylates have been caged as esters (Figure 1, top row). Such esters are inherently unstable to aqueous hydrolysis. CNB-glu

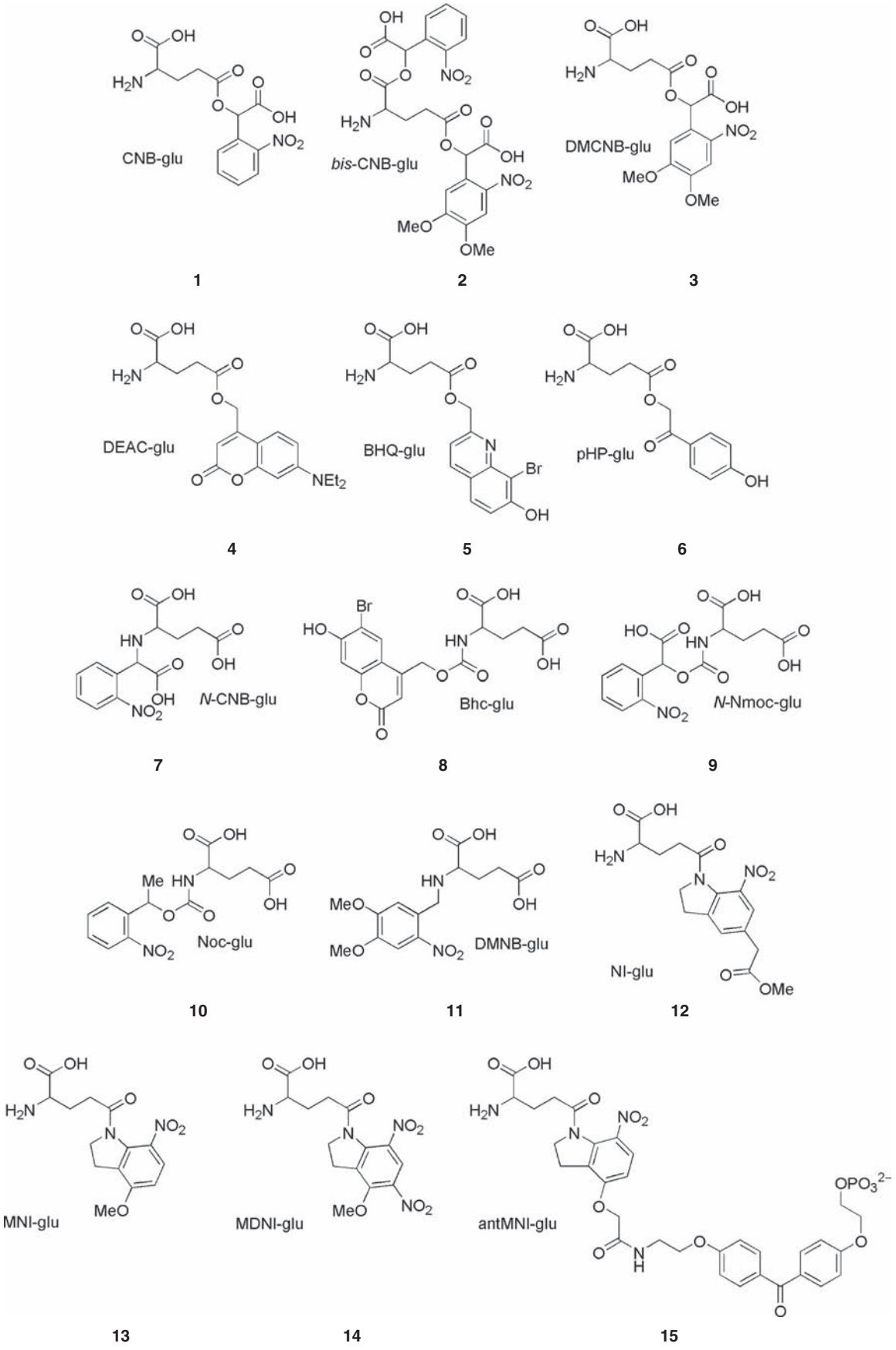

Figure 1 Continued

was the first really useful caged glutamate because it is uncaged quickly and quite efficiently with UV light and it is stable enough at physiological pH to make its use practical on a day-to-day basis. It has been reported that after 2 days at room temperature (RT) only 2% of CNB-glu hydrolyses, and frozen solutions must be quite stable. In contrast, we have found that physiological solutions of DMCNB-glu are so unstable that they are useful only for 20 min before too much free glutamate arises. BHQ-glu is so highly labile (half-life of only 14 h) that the holding current of CA1 neuron in brain slices changed so quickly as to make its use completely impractical. DEAC-glu and pHP-glu are similarly fragile. The amino group of glutamate has been caged in two ways (Figure 1, middle row), either directly as an amine (DMNB-glu and *N*-CNB-glu) or via a carbamate linkage (Noc-glu, Bhc-glu, and *N*-Nmoc-glu). All these cages are completely stable at pH 7.4. Finally, glutamate has been caged via an acylindolinyl bond (Figure 1, bottom row). We have found that solutions of MNI-glu stored at 4 °C for 13 days reveal only 1% hydrolysis, and the dinitro derivative reveal 7%. Frozen solutions of these cages are extremely stable. NI-glu has a half-life of 29 h at 30 °C. The stability of antMNI-glu has not been reported (but it must also be very stable) (see Table 1).

We examined the purity of commercial MNI-glu (from Tocris) and CNB-glu (from Molecular Probes) by high-performance liquid chromatography (HPLC) because most neuroscientists purchase these caged compounds. We found that MNI-glu was 100% pure; there was only one peak in the HPLC trace (254 and 350 nm). In contrast, CNB-glu from Molecular Probes was approximately 98% pure and had one prominent impurity in the chromatogram. Several colleagues have told us that they find batches of MNI-glu from Tocris quite toxic to cells, whereas others have used the commercial compound with considerable success. Our HPLC does not detect free glutamate (the most likely impurity) because there is no useful chromophore in the pure amino acid.

Photochemical Properties of Caged Glutamates

The introduction of CNB-glu in 1994 was a revolution for neuroscience. The cage was not only stable enough for practical use on cultured and acutely isolated neurons, but it had adequate photochemical properties for effective UV photolysis, namely, decent photochemical efficiency (i.e., $\varepsilon.\phi = 75$), accompanied by rapid photolysis (half-time of release of 25 μs). One substantial drawback of CNB-glu is that it is not uncaged by two-photon excitation. An apparent dramatic solution to this deficiency was provided in 1999 by the development of Bhc-glu. This bromocoumarin has a large two-photon cross section (~1000-fold larger than CNB-glu). However, a downside of the complete aqueous stability of the carbamate linkage is the very slow rate of release of glutamate after photolysis of Bhc-glu. This rate ($140\,s^{-1}$, or a half-time of 7 ms) makes highly localized two-photon uncaging impossible because the residence time in the focal volume is a mere 300 μs.

We tested DMCNB-glu, DEAC-glu, and BHQ-glu as possible solutions to this deficiency. As expected, DMCNB-glu did undergo efficient two-photon photolysis (Figure 2); however, as already noted, it is too unstable in aqueous solution for long-term

Figure 1 Structures of caged glutamates. antMNI, MNI with antennae; Bhc, 6-bromo-7-hydroxycoumarinylmethyl; BHQ, 8-bromo-7-hydroxyquinolino; CNB, carboxynitrobenzyl; DEAC, 7-diethylaminocoumarinylmethyl; DMNB, dimethoxynitrobenzyl; glu, glutamate; MNI, 4-methoxy-7-nitroindolinyl; NI, 5-methoxycarbonylmethyl-7-nitroindolinyl; *N*-Nmoc, *N*-(*o*-nitromandelyl)oxycarbonyl; Noc, *N*-1-(nitrophenyl)ethoxycarbonyl; pHP, *para*-hydroxyphenacetyl. Structures **1** and **7** from Denk W (1994) Two-photon scanning photochemical microscopy: Mapping ligand-gated ion channel distributions. *Proceedings of the National Academy of Sciences of the United States of America* 91: 6629–6633; structure **2** from Pettit DL, SSH Wang, Gee KR, and Augustine GJ (1997) Chemical two-photon uncaging: A novel approach to optical mapping of glutamate receptors. *Neuron* 19: 465–469; structure **6** from Park CH and Givens RS (1997) New photoactivated protecting groups, 6: p-Hydroxyphenacetyl, a phototrigger for chemical and biochemical probes. *Journal of the American Chemical Society* 119: 2453–2463; structure **8** from Furuta T, Wang SSH, Dantzer JL, et al. (1999) Bromonated 7-hydroxycoumarin-4-ylmethyls: Photolabile protecting groups with biologically useful crosssections for two photon photolysis. *Proceedings of the National Academy of Sciences of the United States of America* 96: 1193–1200; structure **9** from Rossi FM, Margulis M, Tang CM, and Kao JYP (1997) *N*-Nmoc-L-glutamate, a new caged glutamate with high chemical stability and low pre-photolysis activity. *Journal of Biological Chemistry* 272: 32933–32939; structure **10** from Corrie JET, DeSantis A, Katayama Y, et al. (1993) Postsynaptic activation at the squid giant synapse by photolytic release of L-glutamate from a 'caged' L-glutamate. *Journal of Physiology* 465: 1–8; structure **12** from Papageorgiou G, Ogden DC, Barth A, and Corrie JET (1999) Photorelease of carboxylic acids from 1-acyl-7-nitroindolines in aqueous solution: Rapid and efficient photorelease of L-glutamate. *Journal of the American Chemical Society* 121: 6503–6504; structure **13** from Matsuzaki M, Ellis-Davies GCR, Nemoto T, et al. (2001) Dendritic spine geometry is critical for AMPA receptor expression in hippocampal CA1 pyramidal neurons. *Nature Neuroscience* 4: 1086–1093; structure **14** from Fedoryak O, Sul J-Y, Haydon PG, and Ellis-Davies GCR (2004) Synthesis of a caged glutamate for efficient one and two-photon photorelease on living cells *Chemical Communications* 2005: 3664–3666; structure **15** from Papageorgiou G, Ogden DC, and Corrie JET (2004) An antenna-sensitized nitroindoline precursor to enable photorelease of L-glutamate in high concentration. *Journal of Organic Chemistry* 69: 7228–7233.

Table 1 Summary of the physicochemical properties of caging chromophores[a]

Cage	*Absorption max (nm) [ε mM^{-1}cm^{-1}]*	*Fluorescence max (nm)*	*ϕ (chem)*	*ϕ (fluo)*	*2PCS (GM, 720 nm)*	*Rate (s^{-1})*	*Thermal stability*
CNB	0.5 at 350 nm	None	0.14	None	0.001	4×10^4	Reasonable
N-CNB	0.5 at 350 nm	None	0.04	None	0.001	550	Total
DMNCB	350 [5.2]	None	0.05	None	0.01	2×10^5	Poor
bis-CNB	1.0 at 350 nm	None	0.022	None	0.001		1.3×10^4
Poor							
DMNB	350 [5.0]	None	ND	None	ND	ND	Total
DEAC	399[14.5]	484	0.22	0.0058	ND	ND (10^8)	Poor
pHP[b]	ca. 0.02 at 350 nm	ND	0.8	ND	ND	10^7–10^8	Poor
Noc	0.5 at 350 nm	None	0.65	None	0.001	140	Total
N-Nmoc	350 [5.0]	None	0.11	None	0.01	140	Total
Bhc[b]	370 [15.0]	530	0.012–0.04	0.2–0.4	0.95	140	Total
BHQ[b]	369 [2.6]	510	0.29	0.04	0.54	ND	Poor
NI	340 [1.4]	None	0.026	None	ND	10^3	Excellent
MNI	350 [4.2]	None	0.085	None	0.06	10^5	Excellent
MDNI	350 [8.2]	None	0.45	None	0.06	ND	Good
antMNI	300 [25.2]	ND	0.085	ND	ND	ND (10^5)	Excellent?

[a]2PCS, two-photon cross section at 720–740 nm; ε, extinction coefficient; ϕ, quantum yield (chemical or fluorescent); antMNI, MNI with antennae; Bhc, 6-bromo-7-hydroxycoumarinylmethyl; BHQ, 8-bromo-7-hydroxyquinolino; chem, chemical; CNB, carboxynitrobenzyl; DEAC, 7-diethylaminocoumarinylmethyl; DMNB, dimethoxynitrobenzyl; fluo, fluorescent; MNI, 4-methoxy-7-nitroindolinyl; ND, no datum; NI, 5-methoxycarbonylmethyl-7-nitroindolinyl; *N*-Nmoc, *N*-(*o*-nitromandelyl)oxycarbonyl; Noc, *N*-1-(nitrophenyl)ethoxycarbonyl; pHP, *para*-hydroxyphenacetyl. (See **Figure 1** for structures.)
[b]The absorption maxima of these cages are very pH sensitive because they are phenols that deprotonate in the physiological range (e.g., pHP has pK_a8).

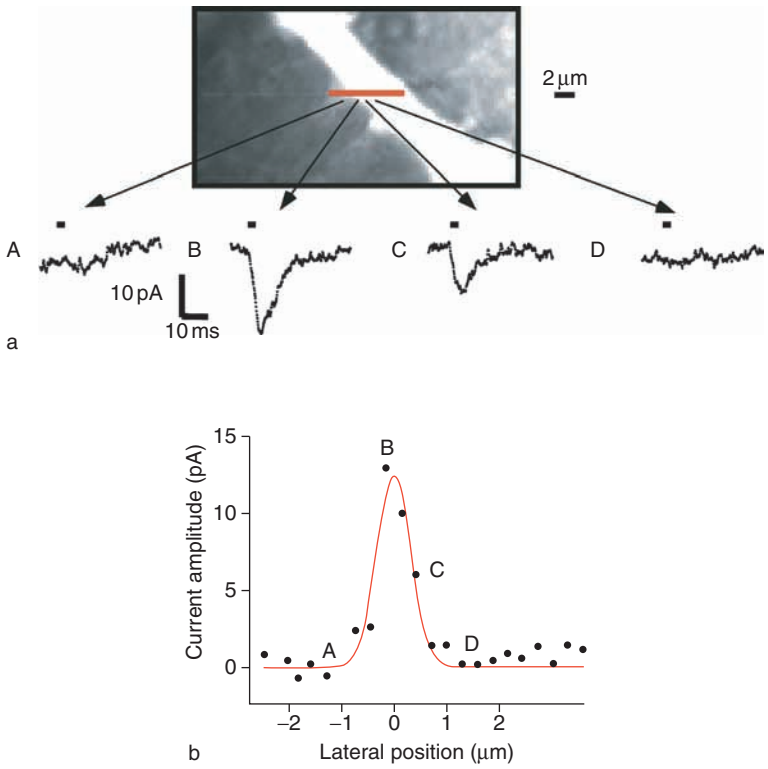

Figure 2 Two-photon uncaging of DMCNB-glu on cultured hippocampal neurons: (a) a mode-lock Ti:sapphire laser used to irradiate (720 nm for 2 ms) a solution of DMCNB-glu on cultured hippocampal neurons filled with fura-2; (b) lateral resolution of the evoked current from two-photon photolysis measured at 0.33 µm intervals. In (a), rapid, diffraction-limited currents similar to spontaneous mEPSCs are produced when the laser is focused on the synapse (current B). mEPSC, miniature excitatory postsynaptic current.

experiments in brain slices. The BHQ chromophore has a large two-photon cross section; however, to our disappointment, we found that BHQ-glu (0.5 mM) was not only highly unstable in aqueous solution but was also brightly fluorescent at both 720 and 900 nm in artificial cerebrospinal fluid (aCSF) when applied to the slice on the stage of our dual two-photon microscope. In fact, BHQ-glu was sufficiently fluorescent at 900 nm to obscure imaging enhanced GFP (eGFP)-labeled CA1 neurons in the slice culture. Bhc-glu was reported to be quite fluorescent, and DEAC is even more fluorescent than either of these chromophores. These difficulties highlight an important property that is normally overlooked in lists of design specifications of ideal caged compounds, namely that they must be nonfluorescent. Uncaging at single synapses in brain slices *in vitro* has to be at visually identified synapses. IR-DIC optics is used for this purpose, but it is almost always accompanied by fluorescent imaging (of GFP or Ca^{2+} dyes such as fluo-4). Caged glutamate is applied at very high concentrations to brain slices (~2–10 mM), so even moderately fluorescent cages (such as BHQ) can completely swamp any analytical fluorescent signal. Nitroaromatic caged compounds are nonfluorescent and have been widely used in conjunction with all possible fluorescent probes (fura-2, fluo-3, GFP, yellow fluorescent protein (YFP), cyan fluorescent protein (CFP), rhod-2, paGFP, DsRed, etc.) without any interference.

Acylnitroindolinyl-caged glutamates (**Figure 1**, bottom row) represent the best compromise currently in terms of their combination of properties. These molecules are stable enough at neutral pH for long-term experiments. Glutamate is released with excellent photochemical efficiently (i.e., $\varepsilon.\phi$ in the range of 110–4000 at 350 nm). We examined the rate of decay of the aci-nitro intermediate that has been shown to be part of the bond-breaking reaction for nitroinodlinyl-caged acids. Wan and co-workers have made a detailed kinetic study of MNI-acetate as a model compound for MNI-caged acids. They found that acetic acid is released with a rate of $5 \times 10^6\,s^{-1}$. We found that the corresponding rate for MNI-glu was $3 \times 10^5\,s^{-1}$ at pH 7.4. Significantly, NI-glu, MNI-glu, and MDNI-glu are nonfluorescent and so do not interfere with widely used probes such as GFP (the benzophenone moiety of antMNI-glu will not be fluorescent but is probably mildly phosphorescent). Finally, MNI-glu is reasonably sensitive to two-photon excitation and has been widely used to stimulate synaptic transmission at single spine heads in brain slices.

Thus, the acylnitroindolinyl caged glutamates represent the state of the art in caged glutamate probes. These caged compounds are very stable at pH 7.4, and they uncage quickly and efficiently on UV illumination. MNI-glu undergoes useful two-photon excitation that enables neuroscientists to produce three-dimensionally diffraction-limited puncta of glutamate release on the same scale as single spine heads in the CNS.

See also: Glutamate; Glutamate Regulation of Dendritic Spine Form and Function; Lipids and Membranes in Brain Aging; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Metabotropic Glutamate Receptors (mGluRs): Functions; Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology; Two-Photon Imaging.

Further Reading

Corrie JET, DeSantis A, Katayama Y, et al. (1993) Postsynaptic activation at the squid giant synapse by photolytic release of L-glutamate from a "caged" L-glutamate. *Journal of Physiology* 465: 1–8.

Denk W (1994) Two-photon scanning photochemical microscopy: Mapping ligand-gated ion channel distributions. *Proceedings of the National Academy of Sciences of the United States of America* 91: 6629–6633.

Fedoryak OD and Dore TM (2002) Brominated hydroxyquinoline as a photolabile protecting group with sensitivity to multiphoton excitation. *Organic Letters* 4: 3419–3422.

Fedoryak O, Sul JY, Haydon PG, and Ellis-Davies GCR (2004) Synthesis of a caged glutamate for efficient one and two-photon photorelease on living cells. *Chemical Communications* 2005: 3664–3666.

Furuta T, Wang SSH, Dantzer JL, et al. (1999) Bromonated 7-hydroxycoumarin-4-ylmethyls: Photolabile protecting groups with biologically useful cross-sections for two photon photolysis. *Proceedings of the National Academy of Sciences of the United States of America* 96: 1193–1200.

Kasai H, Matsuzaki M, and Ellis-Davies GCR (2005) Two-photon uncaging microscopy. In: Yuste R, and Konnerth A (eds.) *Imaging in Neuroscience and Development: A Laboratory Manual*, 2nd edn., pp. 375–384. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Matsuzaki M, Ellis-Davies GCR, Nemoto T, et al. (2001) Dendritic spine geometry is critical for AMPA receptor expression in hippocampal CA1 pyramidal neurons. *Nature Neuroscience* 4: 1086–1093.

Morrison J, Wan P, Corrie JET, and Papageorgiou G (2004) Mechanism of photorelease of carboxylic acids from 1-ayl-7-nitroindolines in solutions of varying water content. *Photochemical and Photobiological Sciences* 1: 960–969.

Papageorgiou G, Ogden DC, Barth A, and Corrie JET (1999) Photorelease of carboxylic acids from 1-acyl-7-nitroindolines in aqueous solution: Rapid and efficient photorelease of L-glutamate. *Journal of the American Chemical Society* 121: 6503–6504.

Papageorgiou G, Ogden DC, and Corrie JET (2004) An antenna-sensitized nitroindoline precursor to enable photorelease of L-glutamate in high concentration. *Journal of Organic Chemistry* 69: 7228–7233.

Park CH and Givens RS (1997) New photoactivated protecting groups, 6: p-Hydroxyphenacetyl, a phototrigger for chemical and biochemical probes. *Journal of the American Chemical Society* 119: 2453–2463.

Pettit DL, Wang SSH, Gee KR, and Augustine GJ (1997) Chemical two-photon uncaging: A novel approach to optical mapping of glutamate receptors. *Neuron* 19: 465–469.

Pelliccioli AP and Wirz J (2002) Photoremovable protecting groups: Reaction mechanisms and applications. *Photochemical and Photobiological Sciences* 1: 441–458.

Rossi FM, Margulis M, Tang CM, and Kao JYP (1997) N-Nmoc-L-glutamate, a new caged glutamate with high chemical stability and low pre-photolysis activity. *Journal of Biological Chemistry* 272: 32933–32939.

Wieboldt R, Gee KY, Niu L, Ramash D, et al. (1994) Photolabile precursors of glutamate: Synthesis, photochemical properties, activation of glutamate receptors in the microsecond time scale. *Proceedings of the National Academy of Sciences of the United States of America* 91: 8752–8756.

Photoperiodic Regulation of Reproductive Cycles

E Bittman, University of Massachusetts, Amherst, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In late February of 1921, Otto Loewi awoke from a dream in which he had imagined a way to test the idea that a chemical signal is released from the vagus nerve. He quickly collected evidence that upon electrical stimulation, a 'vagusstoff' is released that slows the beating of the frog heart. By 20 March, 1921, Loewi had submitted his manuscript to *Pflugers Archiv* and it was soon published. Others were unable to replicate Loewi's finding, however, and the ensuing controversy contributed to a drawn-out debate that delayed acceptance of the principle of chemical neurotransmission for several decades. Only much later did it become apparent that the time of year at which Loewi performed his famous experiment was critical to its success. Seasonal changes in the activity of acetylcholinesterase, and in the concentration of muscarinic receptors, make the effects of vagal stimulation far more difficult to demonstrate after the end of winter.

Many behavioral and physiological functions of animals of the temperate zone exhibit profound seasonal rhythms. Of all phases of the reproductive cycle, it is the care of the young that imposes the greatest energetic demands upon parental species. The young are typically born in the spring, and fertile matings are scheduled accordingly. This is not surprising, given the striking environmental fluctuations that exact a heavy toll on individuals who time their efforts inappropriately. Matching of the timing of reproduction to the annual phase of maximal availability and minimal expenditure of calories (i.e., the abundance of food and warmer temperatures) is thus of great selective advantage. Although fluctuations in temperature and rainfall can acutely influence reproduction, daylength provides a particularly stable and reliable predictor of future energetic demands that enables animals to initiate long-term adjustments. Thus we can readily understand why photoperiod acts as a proximate regulator of neuroendocrine function in seasonal species. In birds, the length of the incubation period dictates that eggs are laid and hatching occurs in the lengthening photoperiods of spring. Similarly, mammals with short gestation periods, such as rodents, are typically stimulated to reproduce by long days. Short days induce reproduction in ungulates, which gestate their young for approximately 6 months. In some species of bats and kangaroos and in some mustelids, mating occurs many months well in advance of implantation. In such cases seasonal cues including photoperiod terminate embryonic diapause at the appropriate time so that young are born in the spring.

Photoperiodic signals also regulate metabolism, growth, and development in a wide variety of plants, invertebrates, and vertebrates. Both annual rhythms in human birth rates and the incidence of seasonal affective disorders indicate the significance of photoperiodism in human health and disease. This article focuses on the mechanisms by which daylength influences reproduction in higher vertebrates, with the predominant emphasis on mammals.

Circadian-Based Photoperiodic Time Measurement

Many physiological events and behaviors exhibit daily periodicity even when animals are maintained in constant conditions. An endogenous circadian oscillator ensures that incompatible processes occur at different times, and that metabolic, thermoregulatory, osmoregulatory, immune, and reproductive functions and arousal are properly scheduled to meet environmental challenges. It was the remarkable insight of Erwin Bünning, one of the pioneers of the study of such rhythms, that a circadian oscillator can be used to measure photoperiod. Bünning imagined that a particular phase of a circadian oscillation might coincide with light only during the long days of summer. Photoperiods of sufficient length would thus trigger seasonally appropriate physiological events. Support for this general model of circadian-based photoperiodic time measurement was first achieved through application of protocols in which long nights were interrupted by brief pulses of light. Although it was found that the effects of long days on reproduction, flowering, diapause, etc. could be mimicked by such regimes, the results are equally consistent with the operation of an hourglass-like timer, with measurement of uninterrupted dark intervals being initiated at dusk. Full validation of the competing hypothesis, that a self-sustaining circadian oscillator underlies photoperiodic responses, awaited the application of 'resonance' experiments. In the most powerful of these designs (**Figure 1(a)**), light is presented to the organism only every second or third circadian cycle. Only a system with the capacity to begin a second or third cycle in constant darkness – that is, to free run – will experience photoinduction when

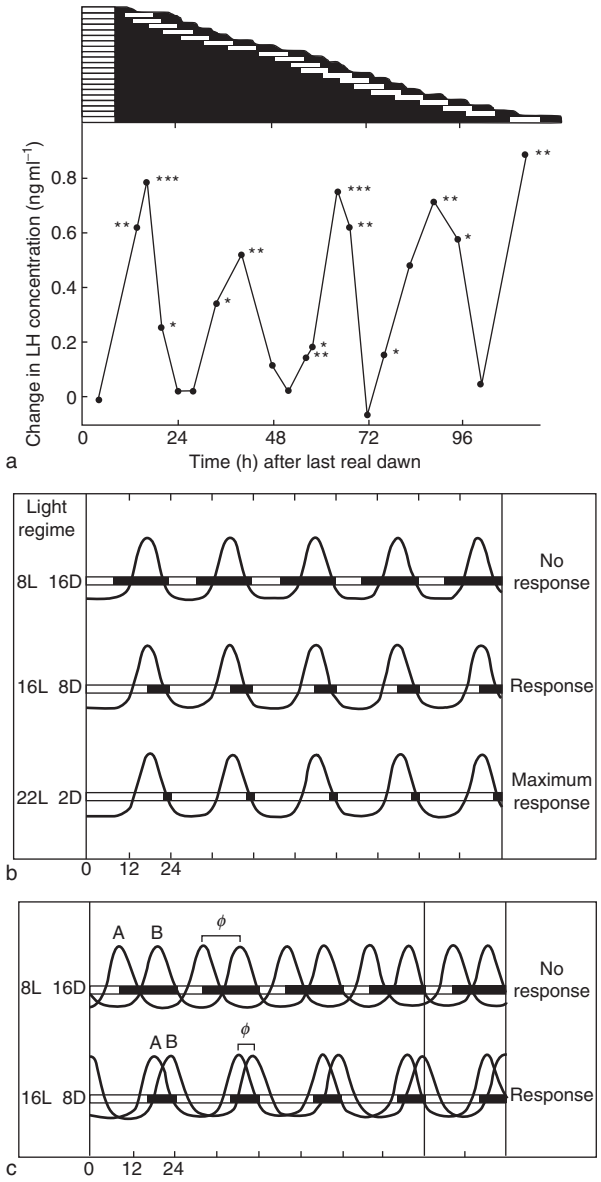

Figure 1 Experimental evidence for circadian-based photoperiodic time measurement and hypothetical models illustrating coincidence models. (a) Top, light regimens to which white crowned sparrows were exposed in a classic experiment by Follett, Mattocks, and Farner. Birds were maintained in short days (8L:16D), which inhibit reproduction, for many weeks. The last 8 h light phase of this regime is illustrated; different treatment groups are represented by horizontal lines. Groups of birds were transferred to constant darkness (black shading) which was interrupted after various intervals by an 8 h light pulse (white bars). This experiment takes advantage of the fact that reproductive induction in this species, as marked by LH secretion, occurs rapidly upon exposure to the first long day. Blood samples were collected a few hours after presentation of the 8 h light pulse. Bottom, mean serum LH response in groups of sparrows exposed to the corresponding light regimes. Note that both the peaks and nadirs of induction are observed when light pulses fall at circadian (near-24 h) intervals after the last day of the 8L:16D regimen. (b) An external coincidence model can explain the experimental results. A hypothetical rhythm of sensitivity to light is shown by the sinusoidal line. When animals are exposed to short days, the sensitive phase coincides with darkness. After transfer to summer daylengths (16L:8D, middle), part of the sensitive phase is struck by light and partial induction of the seasonal response occurs. In very long days (2L:22D, bottom), nearly the entire photosensitive phase is illuminated, and maximal reproductive induction occurs. (c) An internal coincidence model can explain the same data. It is hypothesized that two circadian oscillations participate, one of which (a) is entrained to dusk and the other of which (b) is entrained to dawn. These might be rhythms of enzyme and substrate controlling a critical process such as gonadotropin secretion. The phase relationship (ϕ) depends upon the length of the night. Only during exposure to summer daylengths does reproductive induction occur.

light is presented at the appropriate phase (during the portion of the endogenous daily cycle which normally takes place 12 or more hours after dawn, i.e., the subjective night).

In Bünning's original model, the coincidence or misalignment of an external event (light) with an internal event (the photosensitive phase of a circadian oscillation) determines whether a long- or short-day response occurs (Figure 1(b)). Light has powerful effects on the entrainment of circadian oscillators, however, and it was soon realized that a multitude of such endogenous clocks reside in multicellular organisms. Others envisioned that the phase of some of these oscillations might be set by dawn, and that of others by dusk. The hypothesis was advanced that as daylength changes through the course of the year, the phase relationship between two or more circadian oscillations might vary. At some times of year, the phase relationship would invoke a summer, and others a winter, physiological state. Such 'internal coincidence' models (Figure 1(c)) are fully consonant with Bünning's hypothesis and the results of resonance experiments.

Encoding the Photoperiodic Message

Understanding of molecular and neuroanatomical bases of circadian pacemakers provides critical insight into seasonal responses. Destruction of the suprachiasmatic nucleus (SCN) of the hypothalamus eliminates not only endogenous daily behavioral rhythms in mammals, but also the estrous cycle and circadian rhythms of pineal, pituitary, and adrenal secretion. Transplantation of fetal SCN to hamsters made arrhythmic by SCN lesions reinstates behavioral rhythms, the period of which matches that of the donor. Furthermore, transplantation of wild-type SCN to arrhythmic knockout mice (i.e., strains that are genetically engineered to be incapable of generating behavioral oscillations) instates locomotor rhythmicity. Thus the SCN must function as a master

oscillator: it is necessary and sufficient for behavioral circadian rhythms.

The SCN is critical to annual rhythms as well. Animals made arrhythmic by ablation of the SCN are unable to use daylength cues to schedule reproduction. Retinal projections to the SCN set the phase of (entrain) circadian rhythms, and photoperiod determines the length of the subjective day. Thus the photoperiod sculpts the discharge pattern of SCN neurons in hamsters (Figure 2(a)) and sets the duration, amplitude, and peak phase of SCN multiunit activity in rats. In addition, daylength sets the duration over which light signals can activate SCN neurons, as indicated by *c-fos* expression.

Just as discovery of the neuroanatomical locus of the circadian pacemaker provided insight into structures that govern photoperiodic responses, research into the molecular biology of circadian oscillations revealed biochemical processes that may underlie physiological responses to daylength. Some of the resulting models resemble internal coincidence schemes suggested by Bünning's hypothesis.

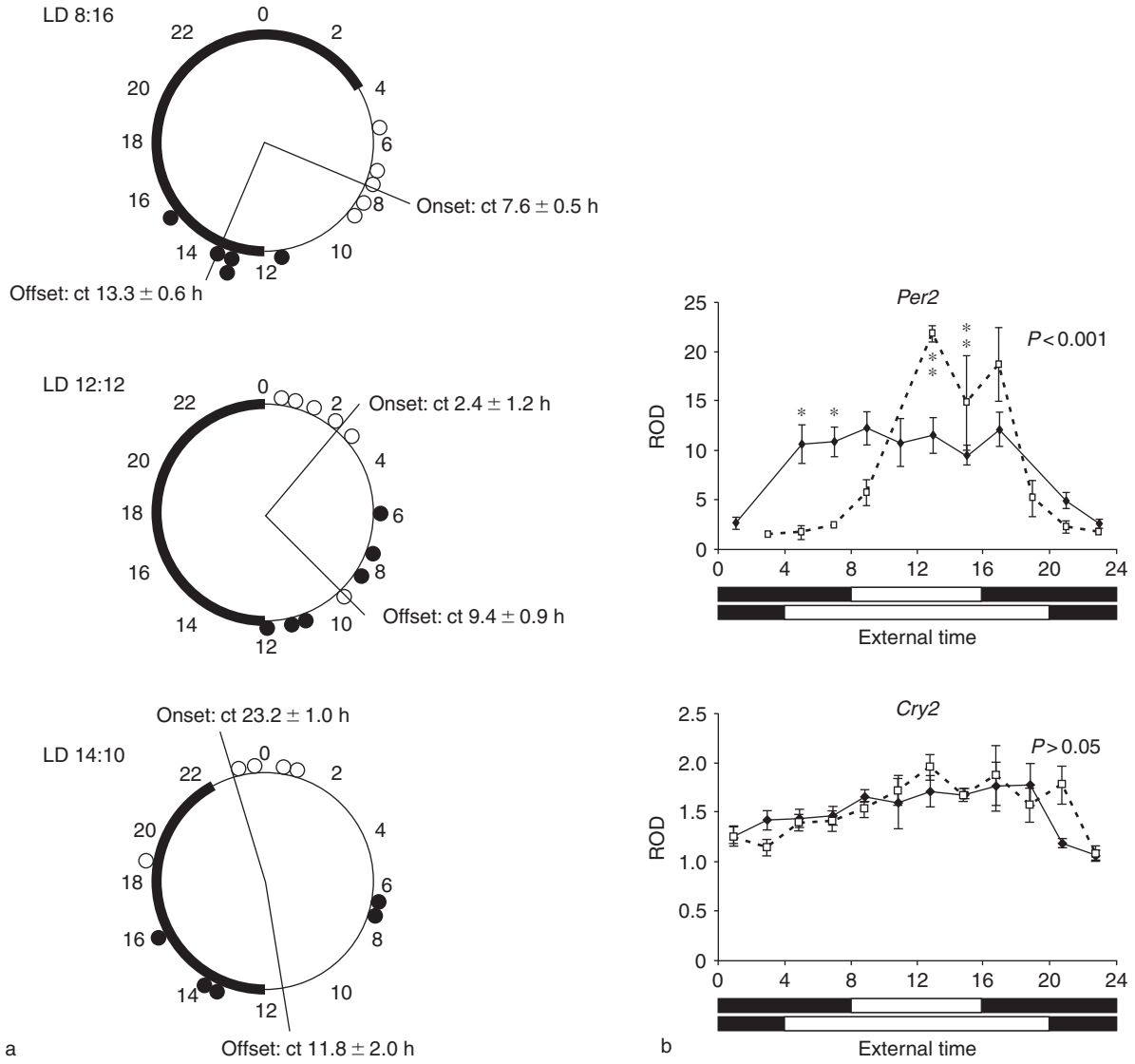

Figure 2 Photoperiod regulates electrophysiological and molecular functions of the SCN, which contains the master circadian pacemaker. (a) Duration of peak multiple unit activity in SCN slices prepared from Syrian hamsters which had been exposed to short (8L:16D), intermediate (12L:12D), or long (14L:10D) daylengths prior to sacrifice. The phase of peak activity in each experimental slice is indicated by small circles and the angle reflects the interval over which spike frequency is elevated. (b) Expression of *Per 2* and *Cry 2* in the SCN of Siberian hamsters exposed to short (open circles) or long (closed circles) daylengths as determined by *in situ* hyridization. Bars below abscissa indicate lighting regime. In short days, *Per 2* mRNA shows a sharper but briefer peak. Expression of *Cry 2* appears little affected by photoperiod. The protein products of *Per 2* and *Cry 2* dimerize to comprise the negative limb of a transcriptional–translational feedback loop that governs circadian oscillations.

Transcriptional–translational feedback loops form the basis of circadian oscillations. The negative arm of the central feedback loop that generates these rhythms is composed of dimers of the protein products of *Period* and *Cryptochrome* genes. *Per1*, *Per2*, *Cry1*, and *Cry2* are expressed in the SCN during the light period when animals are exposed to light:dark (LD) cycles, and during the subjective day when they free-run in constant darkness. The genes *Bmal1* (also known as *Mop3*) and *Clock* are also necessary for the persistence of circadian rhythms, and together their protein products dimerize to comprise the positive limb of the core feedback loop. *Bmal1* also shows a circadian pattern of expression, but its transcription in the SCN occurs during the night. The CLOCK:BMAL1 dimer acts as a positive regulator of *Per* and *Cry* transcription through its binding to E-box enhancers; PER:CRY dimers complete the loop by blocking this drive on *Per* and *Cry* transcription. The model is more complicated in at least two respects. First, PER:CRY inhibits and CLOCK:BMAL1 stimulates the transcription of *Rev-erb* α and RORα, the protein products of which inhibit and stimulate, respectively, *Bmal1* transcription. This constitutes yet an additional feedback loop within the molecular oscillator. Second, the protein products of *Bmal1*, *Cry 1* and *2*, *Per 1* and *2*, and *Clock* all serve as substrates for casein kinase 1ε. The phosphorylation of members of the protein complex determines their stability and nuclear translocation, and thus the period of the circadian oscillator.

In summary, photoperiod regulates the relative amplitudes, durations, and phases of transcription of clock genes, the interactions of which define the operation of the core circadian loops in the SCN.

Efferent projections of the SCN apprise other brain areas of photoperiod and may directly or indirectly trigger seasonal changes in physiological functions. In mammals, photoperiodically driven reproductive cycles cease upon removal of the pineal gland. Synthesis and secretion of melatonin (*N*-acetyl-5-methoxytryptamine; MEL), the principal secretory product of the pineal, is inhibited during the day through a pathway that begins with the circadian pacemaker. Principally through its GABAergic (γ-aminobutyric acid; GABA) projection to the paraventricular nucleus (PVN) of the hypothalamus, but also through glutamatergic, vasopressinergic, and VIPergic (vasoactive intestinal peptide; VIP) input during the night, the SCN controls the pineal's rhythmic secretion of MEL. Descending projections of the PVN regulate the sympathetic preganglionics in the intermediolateral cell column of the spinal cord, which in turn control the superior cervical ganglion. Ganglionic fibers provide noradrenergic input to the pinealocytes which determines the activity of *N*-acetyltransferase (NAT), the rate-limiting enzyme in melatonin synthesis. Co-released peptides and inputs to the pineal from the habenular complex may also modulate melatonin production. The activity of NAT seems principally to rely on transcriptional regulation in rodents, but upon posttranslational mechanisms in sheep and monkeys.

The duration of nightly MEL secretion depends upon daylength, reflecting photoperiodic regulation of transcriptional–translational feedback loops in the SCN pacemaker. This durational MEL signal in turn conveys information about photoperiod to the reproductive neuroendocrine axis: infusions of MEL in patterns typical of photoperiods that prevail during the breeding or nonbreeding season stimulate or inhibit reproduction in pinealectomized animals, respectively (**Figure 3(c)**). The duration of the MEL signal appears more important than its phase, and the ability of the target system to measure the length of the nightly MEL rise appears not to depend upon a circadian timer. Remarkably, the critical daylength for stimulation or inhibition of reproduction may depend upon the animal's photoperiodic history, and may even be communicated from mother to fetus through mechanisms that are not yet understood. Elucidation of this MEL duration measurement system requires description of MEL receptors and their coupling to intracellular signaling pathways.

The situation is different in birds. Although the avian pineal is sympathetically innervated, it is directly photoresponsive. Furthermore, the pineal is unnecessary for photoperiodic control of reproduction. Nevertheless, MEL may regulate seasonal changes in song and perhaps other physiological functions.

Decoding the Photoperiodic Message

The Nature of Melatonin Receptors

Although some MEL receptors may be found in peripheral tissues, it is clear that decoding of the durational MEL signal takes place in the brain and pituitary complex. The use of 2-[^{125}I]iodomelatonin (IMEL) for *in vitro* autoradiography suggests loci of MEL action, and these studies help us to interpret the effects of brain lesions on responses to MEL (**Figure 3**). This ligand clearly demonstrates *bona fide* MEL receptors: IMEL binding sites are specific and of high affinity ($K_d \sim 50$ pM, indicating substantial occupancy at nocturnal physiological serum concentrations), and their occupation elicits intracellular responses. Principal among these is inhibition of adenylate cyclase.

Several MEL receptor genes have been cloned. Three specific MEL receptor subtypes have been identified in

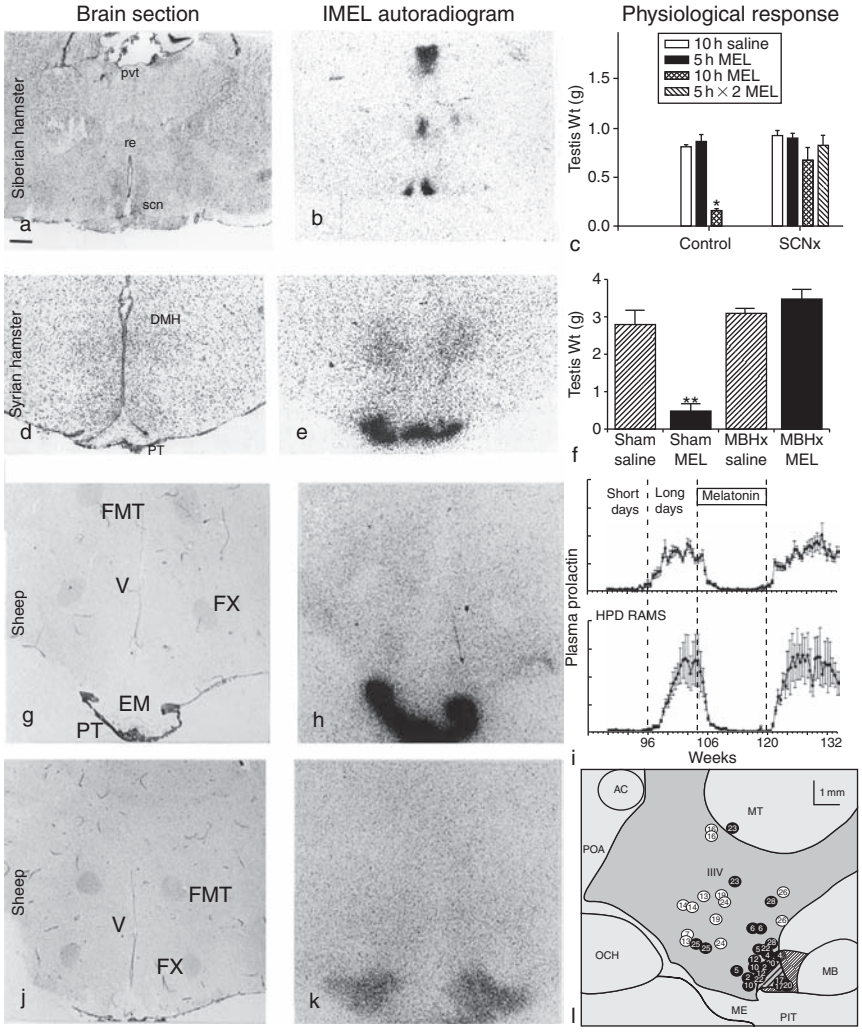

Figure 3 Anatomical and physiological evidence of localized melatonin (MEL) action in photoperiodic mammals. (a) Coronal section of Siberian hamster brain, stained for Nissl substance. Scale bar: 100 μm. (b) Autoradiogram generated by the tissue section in (a) illustrating 2-[^{125}I]-iodomelatonin (IMEL) binding. In all cases shown in this figure, nonspecific IMEL binding is negligible. (c) Combined testis weights of brain intact control (left) or SCN-lesioned pinealectomized experimental (right) Siberian hamsters given 5 weeks of nightly subcutaneous MEL infusions. Infusions with nightly durations of 10 h (mimicking short-day secretory patterns), but not those of 5 h (mimicking long-day patterns), induced reproductive collapse in brain-intact hamsters. 10 h MEL infusions were ineffective in hamsters with SCN lesions. Interruption of the 10 h of nightly MEL infusions (2 × 5 hr) makes them ineffective. (d) Nissl stained coronal section of Syrian hamster brain at the level of the medial basal hypothalamus. (e) Autoradiogram generated by tissue section in (D) showing IMEL binding. (f) Paired testis weight is reduced by nightly infusion of 10 h melatonin in sham-lesioned Syrian hamsters (Sham MEL) compared to controls (Sham Saline). Electrolytic lesions of IMEL-binding region in the dorsomedial hypothalamus(MBHx MEL) relative to lesioned saline infused controls (MBHx Saline). (g) Coronal section of sheep hypothalamus at level of the *pars tuberalis*. (h) Corresponding autoradiogram illustrating IMEL binding. (I). Plasma prolactin (PRL) in Soay rams that were sequentially exposed to winter photoperiods (short days), summer photoperiods (long days), and MEL implants. Top, intact rams; bottom, rams subjected to surgical hypothalamic–pituitary disconnection (HPD). Persistence of control of serum PRL by photoperiod and melatonin in HPD rams indicates regulation by MEL action below the level of the transection (i.e., in the PT). (j) Coronal section through caudal hypothalamus of an Ile-de-France ewe. (k) Corresponding autoradiogram illustrating IMEL binding. (l) Schematic illustration of sheep hypothalamus in the coronal plane illustrating sites in which MEL was surgically implanted. Black circles illustrate sites in which implants induced short-day effect (elevated LH secretion). Open circles illustrate sites at which MEL implants had no effect on serum LH concentrations. Abbreviations: AC, anterior commisure; DMH, dorsomedial hypothalamus; EM, median eminence; FX, fornix; FM, mamillothalamicic tract; MB, Mammillary body; MT, medial thalamus; OCh, optic chiasm; pt, pars tuberalis; PIT, pituitary; POA, preoptic area; IIIV or V, third ventricle.

vertebrates. These are structurally homologous and belong to the same superfamily of seven-membrane-spanning, G_i-protein-coupled receptors. The MT_1 (also known as mel_{1a}) receptor mRNA is of particular interest, because its expression pattern corresponds closely to the maps generated by use of IMEL for *in vitro* receptor autoradiography in mammalian brain. Furthermore, the affinity of this receptor

corresponds closely to that of the physiologically active MEL receptor in the SCN and pars tuberalis. Other members of this MEL receptor family have been identified in various vertebrates, but none can account for the critical function of MEL in reproductive seasonality. The MT_2 (mel_{1b}) sequence contains nonsense mutations in the second exon in both Siberian and Syrian hamsters, and seems not to be expressed in MEL-responsive regions of sheep brain or pituitary. Thus the MT_1 receptor is the likely mediator of photoperiodic responses in seasonal species. Nevertheless, the MT_2 receptor may participate in other putative functions of melatonin, including its capacity to shift circadian phase and to modulate long-term potentiation in hippocampus. A mel_{1c} receptor has been cloned in nonmammalian vertebrates, but is absent in mammals, and a 'MEL-related receptor' has little affinity for IMEL.

Brain Targets of Melatonin: Heterogeneity among Photoperiodic Species

The critical site(s) upon which MEL acts to regulate reproduction seems to be surprisingly variable between photoperiodic species (Figure 3). In rodents, the highest concentration of IMEL binding sites is found in the SCN. In the Siberian hamster, the SCN is critical to photoperiodic responses in two ways: not only does this structure function as a pacemaker to drive the rhythmic secretion of MEL by the pineal gland, but it is also an essential target of MEL. In this species, SCN IMEL binding is largely concentrated rostral to the main retinal input zones, but overlaps with light-induced Fos expression in the ventrolateral region. IMEL binding is regionally co-localized with arginine vasopressin (AVP) mRNA, and dual-label *in situ* hybridization confirms the cellular co-localization of MT_1 receptors with AVP mRNA. Other cell types within the SCN may also be responsive to MEL. Most importantly, destruction of the SCN eliminates the ability of MEL infusions to regulate reproduction in Siberian hamsters (Figure 3(c)). Other brain targets of MEL may also participate, but they are less sensitive than is the SCN and they fail to sustain responses to MEL after the SCN has been destroyed. Nevertheless, somata of the paraventricular and reuniens nuclei of the thalamus contain high concentrations of MEL receptors, and infusion of MEL to these sites can elicit gonadal regression. Interestingly, systemic administration of MEL can induce a second collapse after long-term implants in the SCN have ceased to maintain reproductive inhibition. Presumably, this phenomenon represents independent function of extra-SCN (possibly thalamic) MEL receptor sites.

In contrast to Siberian hamsters, Syrian hamsters remain responsive to MEL after destruction of the SCN. Other forebrain structures, including the dorsomedial hypothalamus (Figure 3(e)) and preoptic regions (including the bed nucleus of the stria terminalis), may be more important in the response to MEL in this species. Only lesion studies and IMEL autoradiography of limited resolution have thus far been used to identify these sites. Thus the phenotype of MEL-responsive cells remains unclear.

In sheep, several lines of evidence suggest that the premammillary hypothalamus functions as a particularly important target of melatonin. First, this perifornical region has a high concentration of IMEL binding sites and expresses MT_1 receptor (Figure 3(k)). In contrast, the ovine SCN contains a lower concentration of IMEL binding sites than do surrounding areas of the hypothalamus and preoptic area. Second, electrolytic lesions that encompass the caudal arcuate and premammillary region of Soay rams affect photoperiodically driven patterns of gonadotropin secretion. Third, implants of MEL in this area stimulate the secretion of luteinizing hormone (LH); (Figure 3(l)). The premammillary hypothalamus is far more sensitive to MEL implants than the anterior hypothalamic region is in this regard.

Reproduction of mustelids (including mink and ferrets) is exquisitely sensitive to photoperiod, and pineal melatonin secretion appears indispensable to this response. Nevertheless, specific MEL binding sites have not been found in the brain. The basis of the difference in the role of the SCN and other brain structures in reproductive responses of these photoperiodic species to MEL is not understood.

The *pars tuberalis* as a target of melatonin Through its control of the MEL secretory pattern, photoperiod regulates seasonal gonadal and thyroidal function. Thus it might logically be expected that lactotrophs, gonadotrophs, or thyrotrophs of the pars distalis might be MEL targets. This seems not to be the case: these cell types lack MEL receptors in adult mammals. In contrast, the *pars tuberalis* (PT) of the pituitary complex clearly plays a major role in MEL-dependent seasonal rhythms of prolactin (PRL) secretion. This structure has consistently been found to contain exceptionally dense concentrations of MT_1 receptor (Figures 3(e) and 3(h)). Importantly, local implantation of MEL into the PT (but not the *pars distalis*) regulates PRL secretion *in vivo*, although LH secretion is unaffected. Most dramatically, Soay rams continue to exhibit photoperiodically driven seasonal rhythms of prolactin (PRL) secretion after hypothalamic-portal disconnection (Figure 3(i)). These cycles persist even though section of the infundibular stalk interrupts gonadal function. Furthermore, systemic administration of MEL continues to regulate PRL secretion

in rams subjected to this operation. When applied *in vitro*, MEL inhibits release from PT of a factor that stimulates PRL secretion from cultured *pars distalis*. It thus appears that MEL acts on the PT to regulate secretion of a peptide that travels in the hypothalamo-hypophyseal portal system to regulate PRL secretion. This 'tuberalin' remains poorly characterized despite its obvious importance to our understanding of seasonal reproduction. It is also likely that MEL regulates secretion of hypothalamic signals (including dopamine, VIP, and thyrotropin-releasing hormone) that may regulate PRL secretion, but the direct action of MEL on the PT must be viewed as the principal means of photoperiodic regulation of the lactotrophs.

Within the PT, the circadian pattern of clock gene expression is regulated by MEL. This contrasts with the SCN, which responds directly to retinohypothalamic input. In rodent PT, MEL acts via the MT_1 receptor to induce rhythmic expression of *Per1*. Short days abolish the rhythm of *Per1* mRNA in the PT of Siberian hamsters. It is expected (but not yet established) that this effect of photoperiod is translated into an absence of PER1 protein in short days. Thus the fundamental operation of the circadian transcriptional–translational feedback loop in the PT is likely altered by photoperiod. The endogenous rhythm depends upon E-box-regulated expression of *Cry* and *Per*, while MEL suppresses adenylate cyclase dependent generation of transcription factors that bind Ca^{2+}/cAMP-dependent response elements (CREs). Alternatively, or in addition, MEL may also act via other signal transduction pathways to alter clock gene expression in rodent PT.

In sheep PT, *Per1/2, Bmal1*, and *Rev-erb* α expression levels continue to oscillate when MEL is suppressed by constant light. Nevertheless, MEL may have a suppressive effect on levels of their mRNAs. In contrast, expression of *Cry1* is induced by MEL. According to a current model (**Figure 4**), the phase relationship between expression of *Per1* and *Cry1* is set by MEL duration. This may provide a mechanism for daylength to act through the pineal to regulate PT function, possibly in order to drive seasonal rhythms of PRL secretion. Validation of this model awaits demonstration that manipulation of clock gene expression (other than through MEL) affects the PT's ability to regulate lactotroph function.

Responding to Photoperiodic Signals

Neural Pathways That Relay Photoperiodic Influences

Once decoded, the MEL signal must engage a neural pathway that ultimately determines the pattern of gonadotropin secretion. Nevertheless, little consistent change occurs in gonadotropin-releasing hormone (GnRH) expression as a result of exposure to inhibitory photoperiods. This has led to the suggestion that it is the secretion, rather than the synthesis, of GnRH that is photoperiodically modulated. Given the progress that has been made in identifying MEL targets, and knowledge about the multiple inputs to GnRH cells, one would think it a simple matter to identify a reasonably complete circuit by which photoperiod controls GnRH secretion. This is not the case: very little is known about the nature and identity of pathways which relay influences of MEL from its neural targets to neuroendocrine effectors. In Siberian hamsters, the SCN is a critical target of MEL and the MT_1 receptor is co-localized with AVP mRNA. Thus vasopressinergic efferents of the SCN represent a principal candidate effector of MEL action in this species. Other SCN outputs may also relay photoperiod information, and it even remains uncertain whether all of the relevant signals are neural. SCN transplants that are encased in polymer capsules which preclude axonal outgrowth are capable of driving behavioral rhythms, indicating the existence of a humoral output of the pacemaker. Although the SCN appears to regulate neuroendocrine rhythms by neuronal projections rather than humoral signals, the relative importance of humoral outputs versus axonal projections of the SCN in relaying the effects of MEL is unknown.

The neural pathways through which MEL action at non-SCN sites drives seasonal reproductive rhythms in other species are also unclear. In sheep, the premammillary region that appears to play an essential role in responses to MEL may regulate dopaminergic cells in the anterior hypothalamus (A14/15), which in turn modulate GnRH release. In sheep and hamsters, photoperiod may regulate the expression of peptides and neurotransmitters that modulate GnRH release, including neuropeptide Y (NPY), dopamine (DA), β-endorphin, and GABA. It is also likely that daylength regulates expression of receptors for these modulators either in GnRH cells or in neurons that control them. Recent identification of a gonadotropin inhibitory hormone (GnIH) in the dorsomedial hypothalamus of Syrian hamsters suggests a possible route by which MEL might regulate LH secretion, particularly in light of evidence that MEL receptors and GnIH are regionally co-localized in this species. It will be important to determine whether GnIH cells are directly responsive to MEL. Kisspeptin-1, a newly discovered modulator of GnRH neurons, is another potential seasonal regulator of reproduction whose modulation by MEL and photoperiod deserves further exploration.

In the spring of 1975, Fred J Karsch performed an experiment designed to test the role of estradiol in the

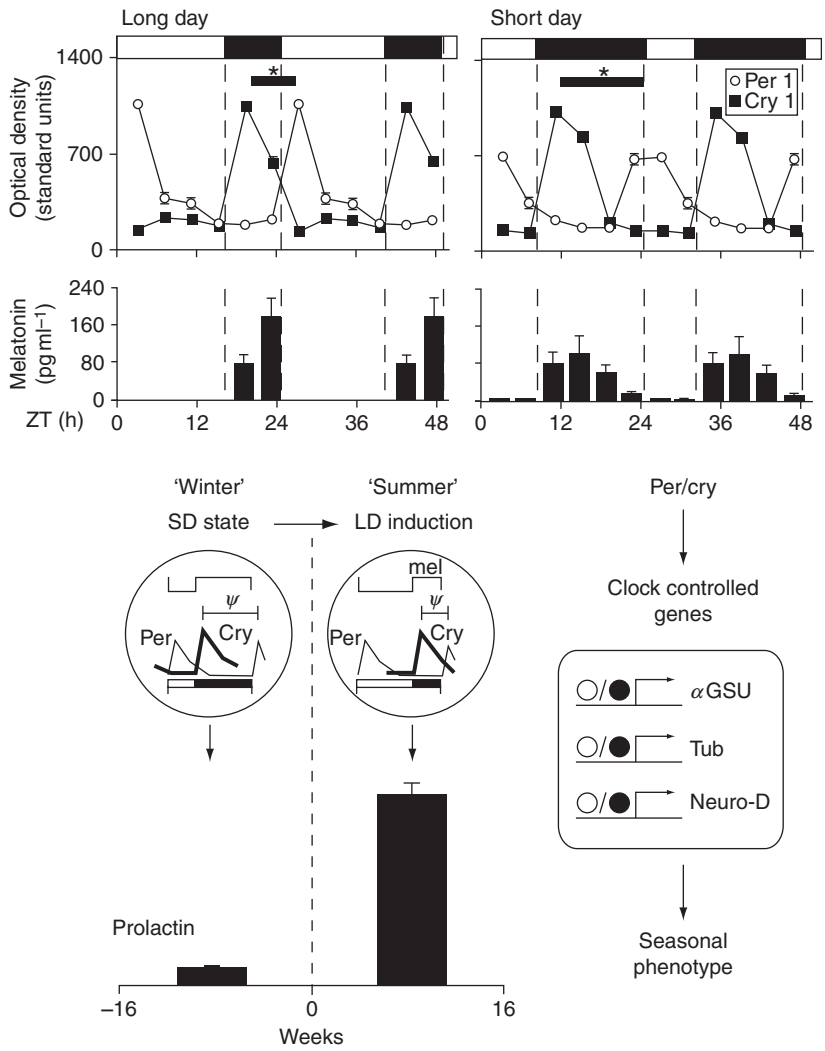

Figure 4 Experimental data supporting an internal coincidence model for photoperiodic control of prolactin secretion in sheep. Patterns of *Per1* (open circles) and *Cry1* (black squares) expression in *pars tuberalis* of Soay rams exposed to long (left) or short (right) photoperiods. Note that the interval between peak levels of these mRNA species (black bar, asterisk) depends upon daylength. Middle, serum melatonin concentrations in the same animals as a function of zeitgeber time (ZT); the duration of nightly melatonin secretion depends upon photoperiod and determines the levels of serum prolactin. Bottom, hypothetical scheme by which Per-Cry interval (ψ) determines prolactin secretion, as well as other genes (α subunit of gonadotropin, tubulin, neuro-D). The correlational data supporting this model remains to be tested by direct manipulation of *Per* and *Cry* mRNA levels.

regulation of LH secretion in the ewe. When he completed radioimmunoassays about 4 months later, he noted that estradiol had produced a marked suppression of tonic LH secretion which made the ensuing surge more obvious. He quickly repeated the experiment, but to his consternation saw no suppression of serum LH concentrations upon estradiol treatment. Repeating the experiment twice more at intervals of several months, Karsch and colleagues eventually realized that estradiol exerts a potent negative feedback effect only during the anestrous season.

Daylength determines several of the actions of gonadal steroid hormones upon the brain of seasonal breeders. During the nonbreeding season, or upon infusion of MEL in corresponding patterns, the negative feedback potency of gonadal steroid hormones upon gonadotropin secretion increases markedly. This ensures that spermatogenesis and follicular maturation are held in check, as gonadal activation is terminated before levels of gonadotropic support required for completion of these processes is reached. At the same time, the ability of testosterone to activate male sexual behavior, and that of estradiol and progesterone to activate female sexual behavior, is decreased. It is not understood how the same photoperiodic (and MEL) stimulus can simultaneously drive feedback and behavioral responses to the same gonadal hormones in opposite directions. Photoperiod

and MEL treatment can elicit small though statistically significant changes in the expression of receptors for gonadal steroid hormones, and for steroid receptor cofactors, in specific hypothalamic and preoptic regions of seasonal breeders. It seems unlikely, however, that these effects can account for dramatic seasonal swings in the sensitivity and responsiveness of the brain to steroid hormones. Furthermore, exposure of castrated animals to long or short photoperiod can elicit changes in the patterns of gonadotropin secretion (particularly GnRH pulse frequency) even when gonadal steroid hormones are not replaced. These 'steroid-independent effects of daylength suggest an underlying effect of MEL on the pulse generator that may be amplified in the presence of gonadal hormones. The precise mechanisms and circuitry by which photoperiod elicits these dramatic changes in hypothalamo-hypophyseal function remain to be explained.

Photoperiod and Neural Plasticity

The brains of some adult vertebrates are capable of substantial reorganization on a seasonal basis. This is most dramatically illustrated in several species of birds, in which the size and connectivity of nuclei critical to song (including the neostriatal high vocal center (HVc), robust nucleus of the archistriatum, and area X of the parolfactory lobe) undergo substantial volume increases every spring. Although gonadal hormones clearly participate in these changes, these structures also contain MEL receptors and the pineal hormone can suppress growth of the song nuclei even after castration.

Although less dramatic in scale, there is also clear evidence of seasonal changes in brain structure in mammals. Neurogenesis continues throughout adulthood in the subependymal zone of rodents, and newborn cells migrate to the olfactory bulb and dentate gyrus. Photoperiod has a modest but statistically significant effect upon the incorporation of new neurons in the forebrain of Syrian hamsters. Besides a potential role in neurogenesis and/or cell death, daylength may regulate neural connectivity. Production of neuronal cell adhesion molecule (N-CAM) in anterior hypothalamus and preoptic area (POA) also increases upon transfer of Siberian hamsters to long days. Ultrastructural studies reveal seasonal changes in glial ensheathment of GnRH neurons in ewes. During the breeding season, the perikarya of GnRH neurons receive more synaptic contacts from NPY and GABA-immunoreactive axons, while β-endorphin contacts decrease. The human SCN is reported to show an increase in volume and a twofold increase in vasopressin-immunoreactive neurons in autumn over summer levels. It remains uncertain whether any of these changes represent a primary response to daylength or are secondary to seasonal fluctuations of gonadal, thyroid, or other hormones.

Circannual Rhythms: Entrainment by Photoperiod

As already outlined, over the course of the year animals of the temperate zone experience alternate phases of reproductive stimulation and inhibition that are regulated by seasonal fluctuations of photoperiod. In many well-studied species, including Siberian hamsters, Syrian hamsters, deer mice, and mink, a phase of refractoriness to inhibitory photoperiods occurs. Thus gonadal regression is triggered, but reproduction resumes after several months despite continued exposure to daylengths that precipitated gonadal collapse. This recrudescence results from an endogenous mechanism and whose time course ensures that reproduction resumes in time to produce new young in the spring. In such species a second phase of reproductive inhibition cannot ensue unless refractoriness is broken, typically by exposure to daylengths of the breeding season. Interestingly, refractoriness is not accompanied by changes in the pattern of clock gene expression in SCN or PT. This may indicate that refractoriness arises 'downstream' from the mechanism of photoperiodic time measurement, and is consistent with the observation that photorefractory animals can accurately measure daylength in order to break refractoriness.

In some species, however, rhythms of reproduction and other physiological functions including metabolism, pelage, and body weight continue to oscillate with a period of approximately 9–12 months even under experimental circumstances in which daylength is held constant. If refractoriness to photoperiod can be said to exist at any phase of the reproductive cycle, it can be spontaneously broken in the absence of exposure to daylengths other than those that initially triggered reproductive collapse. In some species circannual oscillations persist only within a restricted range of unchanging photoperiods. Nevertheless, it is clear that some species have an endogenous mechanism capable of generating annual rhythmicity.

It is not clear whether a central circannual pacemaker exists in any animal at a single definable locus. Anterior hypothalamic lesions have eliminated circannual rhythms in ground squirrels, but no site has been shown consistently to produce this effect. Furthermore, the results of such lesions on circannual rhythms may depend upon ambient temperature. Evidence that

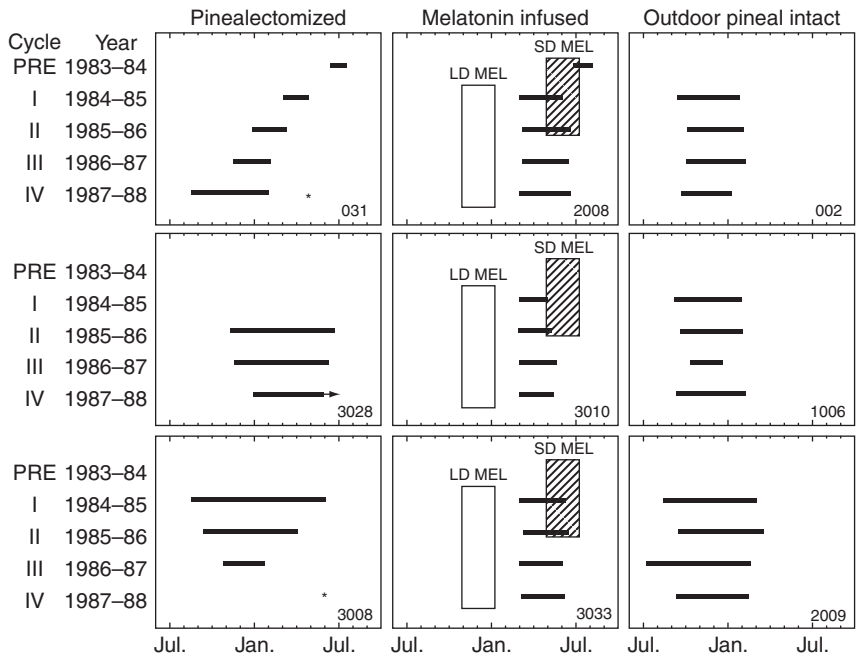

Figure 5 Circannual rhythms in sheep are entrained by melatonin secretion. Intervals of peak LH secretion, indicating low sensitivity to estradiol negative feedback typical of the breeding season, are illustrated by horizontal black bars over the course of 5 years of an experiment performed in ewes. Each box represents data from one animal. Left column illustrates negative feedback response in three pinealectomized animals maintained in constant photoperiods from 1984 to 1988. Onset and offset of peak LH secretion occurred at circannual intervals. Middle column illustrates response of three pinealectomized ewes to 2 yearly intervals of MEL infusions, one a short nightly duration (LD mel) administered from November through January and the other a long nightly duration (SD mel) administered from May through July. Note that this regimen controlled the phase and period of reproductive induction. Right column illustrates timing of LH secretion in three pineal-intact controls maintained outdoors. Note that LH secretion in the melatonin-infused ewes was out of phase with that of pinealectomized animals and outdoor controls.

the circannual oscillatory mechanism is indicates anatomically and formally distinct from the circadian clock. First, some squirrels sustaining SCN lesions that abolish circadian rhythms continue to sustain circannual rhythms, and some hypothalamic lesions that stop the circannual oscillation permit circadian rhythms of locomotor behavior to continue. Second, birds whose circadian clock was entrained to light:dark cycles whose period (T) was longer or shorter than 24 h experienced neither a lengthening nor a shortening of their circannual period. Thus the circannual rhythm is not a 'counter' of circadian intervals.

Photoperiod entrains circannual rhythms. Without such an entraining input the mismatch between the period of the circannual rhythm and the 12-month environmental cycle would result in inappropriate phasing of reproduction within a few years. This role of daylength as a circannual zeitgeber appears to depend upon MEL (Figure 5). Other endocrine cues can determine whether circannual rhythms persist, or whether animals enter refractory states. Thyroidectomy blocks circannual cycles in sheep, as animals remain in reproductive condition indefinitely. Although thyroid hormones clearly exert their action in the brain, they may nevertheless play a permissive role, in that levels that do not fluctuate seasonally appear adequate to ensure appropriately phased cessation of the breeding season. Access and processing of thyroid hormone to the brain may determine the persistence or termination of photorefractoriness in hamsters and Japanese quail, perhaps though regulation of expression of thyroid hormone-binding proteins (e.g., transthyretin) or monodeiodinase.

The relationships between circannual rhythms and seasonal cycles that require alternation of long and short days remain to be established. Rapid progress in anatomical and molecular definition of photoperiodic responses over recent years inspires hope that even this difficult question can be answered.

See also: Circadian Organization; Circadian Systems: Evolution; Circadian Rhythm Models; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Function and Therapeutic Potential of Melatonin in Humans; Clock Gene Regulation of Endocrine Function; Entrainment of Circadian Rhythms by Light; Gonadotropin, Neural and Hormonal Control; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Pineal Gland and Melatonin.

Further Reading

Goldman BD (2001) Mammalian photoperiodic system: Formal properties and neuroendocrine mechanisms of photoperiodic time measurement. *Journal of Biological Rhythms* 16: 283–301.

Hazelrigg DG and Wagner GC (2006) Seasonal photoperiodism in vertebrates: From coincidence to amplitude. *Trends in Endocrinology and Metabolism* 17: 83–91

Lincoln, GA (2006) Decoding the nightly melatonin signal through circadian clockwork. *Molecular and Cellular Endocrinology* 252: 69–73.

Lincoln GA (2006) Melatonin entrainment of circannual rhythms. *Chronobiology International* 23: 301–306.

Malpaux B (2006) Seasonal regulation of reproduction in mammals. In: Neill JD (ed.) *Knobil and Neill's Physiology of Reproduction,* 3rd edn., Chapter 41, pp. 2231–2281. Amsterdam: Elsevier.

Schwartz WJ, de la Iglesia HO, Zlomanczuk P, et al. (2001) *Le quarto stagioni* within the mammalian brain: Photoperiodic orchestration through the suprachiasmatic nucleus. *Journal of Biological Rhythms* 16: 302–311.

Photoreceptor Adaptation

G L Fain, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Visual Adaptation

Sensory systems in vertebrates have a remarkable ability to adjust their sensitivity to maintained stimulation, and this is nowhere more impressive than for the eye. The sensitivity of vision is constantly altered over a wide range of ambient illumination. Some part of this process must result from synaptic interactions in the retina and central nervous system, since adaptation in humans and other vertebrates can begin at light intensities that bleach fewer than one rhodopsin molecule per rod outer segment. At such dim intensities, adaptation cannot be mediated by the receptor itself but must take place in some cell that acts as a summation pool, effectively combining the signals of many receptors. How this occurs is still poorly understood. Over much of the range of physiological light levels, however, adaptation is now known to occur principally within the photoreceptors themselves, and the mechanism of this process has been the subject of considerable investigation.

Light Adaptation in a Rod

Figures 1(a)–1(c) show a continuous recording of photoreceptor current made with a suction electrode from the outer segment of a salamander rod. In the absence of background light (Figure 1(a)), there is a resting current in darkness of about 40 pA into the outer segment through the cyclic guanosine monophosphate (cGMP)-gated channels (see Figure 2). A flash of light closes a proportion of the channels, causing a decrease in current. The brighter the intensity of the flash, the greater the decrease in current, until bright flashes close all of the channels, and the response is saturated.

Figures 1(b) and 1(c) show the same rod in the presence of two continuous background illuminations. Presentation of the steady background light produced a maintained response, but with continued illumination the amplitude of the response to the background gradually subsided. If flashes of light were then superimposed on top of the background illumination, the responses can be seen to be smaller, and the photoreceptor less sensitive. This effect can perhaps more clearly be seen in Figure 1(d), where responses have been superimposed to a flash in darkness (largest response) and in the presence of backgrounds of increasing intensity. The backgrounds can be seen both to decrease the sensitivity of the cell and to accelerate the time course of response decay. It might seem that a decrease in sensitivity in the presence of the background would be maladaptive, but this is not the case. The decrease in sensitivity permits the photoreceptor to adjust its operating range to higher intensities in brighter ambient illumination. This makes it possible for the visual system to continue to respond with high sensitivity to differences in illumination, that is, to contrast, at a light level that is appropriate to the brightness of the visual scene.

Mechanism of Light Adaptation

Adaptation in photoreceptors is known to require a second messenger and considerable evidence now indicates that this messenger is the calcium ion. Ca^{2+} flows into the rod outer segment through the cyclic nucleotide-gated channels in darkness and is extruded by an Na^+/Ca^{2+}-K^+ exchange transporter. When the photoreceptor is illuminated (Figure 2), the active form of rhodopsin (called Rh^*) binds to the heterotrimeric G-protein transducin. This causes the exchange of guanosine triphosphate (GTP) for guanosine diphosphate (GDP) on the guanosine nucleotide binding site, and Gα-GTP then binds to an enzyme called cyclic GMP phosphodiesterase (PDE), which hydrolyzes cGMP. The reduction in cGMP concentration causes the cyclic nucleotide-gated channels to close during illumination. This in turn decreases the influx of Ca^{2+} through the channels, and the continued activity of the Na^+/Ca^{2+}-K^+ exchange transporter causes the free-Ca^{2+} concentration in the outer segment to decrease. Over most of the physiological range of the photoreceptor, the outer segment Ca^{2+} concentration is simply proportional to the fraction of channels that are open. This means that the steady decrease in current produced by background light (as in Figures 1(b) and 1(c)) produces a steady decrease in Ca^{2+}, and the amplitude of the decrease is nearly proportional to the decrease in current.

The decrease in Ca^{2+} is known to have a number of effects on the phototransduction machinery (Figure 2). One of the most important of these is the regulation of the guanylyl cyclase, the enzyme that synthesizes cGMP. The activity of the cyclase is regulated by small molecular-weight binding proteins called guanylyl cyclase-activating proteins (or GCAPs). These proteins associate with the guanylyl cyclase at cytoplasmic binding sites near the disk membrane.

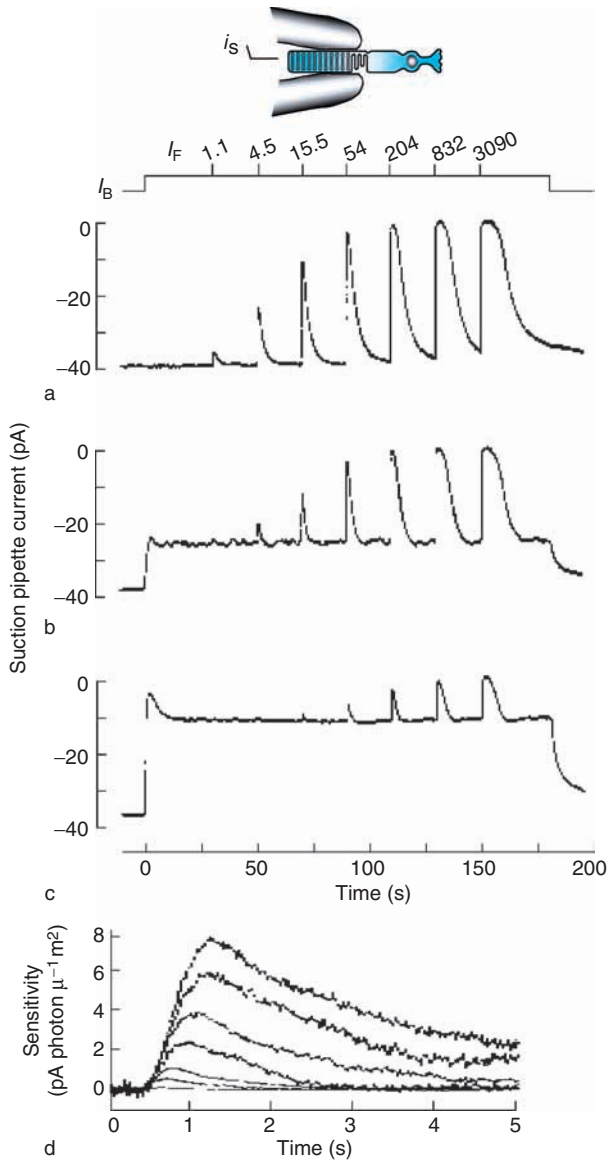

Figure 1 (a–c) Responses of same salamander rod to flashes given in darkness and in the presence of two steady background lights. I_F, flash intensity in units of photons μm^{-2}. Background intensities were 1.7 (b) and 37.2 (c) photons $\mu m^{-2}\ s^{-1}$. (d) Responses of salamander rod in darkness (largest response) and in backgrounds of progressively increasing brightness. Responses have been plotted as sensitivities by dividing the current (in pA) by the light intensity (in units of photons μm^{-2}). (a–c) Replotted from Matthews HR (1990) Messengers of transduction and adaptation in vertebrate photoreceptors. In: Herring PJ, Campbell AK, Whitfield M, and Maddock L (eds.) *Light and Life in the Sea*, pp. 185–198. Cambridge, UK: Cambridge University Press. (d) From Fain GL and Cornwall MC (1993) Light and dark adaptation in vertebrate photoreceptors. In: Shapley R and Lam DMK (eds.) *Contrast Sensitivity Proceedings of the Retina Research Foundation Symposia*, vol. 5, pp. 3–32. Cambridge, MA: MIT Press.

When the Ca^{2+} concentration is high in darkness, the GCAPs inhibit the cyclase; when the Ca^{2+} falls, the GCAPs come off the cyclase binding sites, and the rate of the cyclase is accelerated. Activation of the cyclase by low Ca^{2+} causes an increase in cGMP, which allows a fraction of the channels to reopen in spite of the maintained stimulation of the PDE by the background light. It is the cyclase activation that is responsible for the gradual decrease in the amplitude of the response to a maintained background (Figures 1(b) and 1(c)).

Ca^{2+} has also been shown to regulate the cyclic nucleotide-gated channels. This effect seems to make little contribution to adaptation in rods but may be of considerably greater importance in cones. The binding of Ca^{2+} in association with a small molecular-weight binding protein increases the effective affinity of the channels for cGMP, increasing the probability of channel opening, and allowing the current gradually to recover in the presence of maintained illumination. Finally, Ca^{2+} appears to regulate a step early in the transduction cascade, apparently at the level of the rhodopsin molecule itself. Some evidence indicates that this effect is mediated by a low-molecular-weight Ca^{2+} binding protein called recoverin, which may bind to and modulate rhodopsin kinase, the enzyme that phosphorylates rhodopsin. Modulation of kinase activity may regulate the lifetime of Rh^*.

As a result of the effects produced by the altered Ca^{2+} concentration, the steady illumination alters the sensitivity of the rod. This is primarily the result of a change in the steady level of turnover of cGMP, in the following way. The steady illumination produces a maintained increase in the rate of the PDE, which causes the channels to close. The closing of the channels decreases the free Ca^{2+} concentration, as we have seen; this in turn produces an increase in cyclase rate. The consequence is that the photoreceptor reaches a new steady state corresponding to a new level of steady current in the presence of the background (see Figures 1(b) and 1(c)), at which the rate of hydrolysis of cGMP by the PDE is equal to the rate of synthesis by the cyclase, and both are higher than in darkness. This increase in cGMP turnover produces a decrease in the sensitivity of the rod, since more light will be required to alter the PDE rate enough above the background rate to produce a significant and detectable change in the concentration of cGMP. The increase in turnover is also largely responsible for the increase in the rate of response decay in background illumination (Figure 1(d)), since increased turnover allows the change in cGMP concentration to track more faithfully the change in PDE rate produced by the transduction cascade.

Dark Adaptation

When a photoreceptor is exposed to bright light and then returned to darkness, the sensitivity of the cell

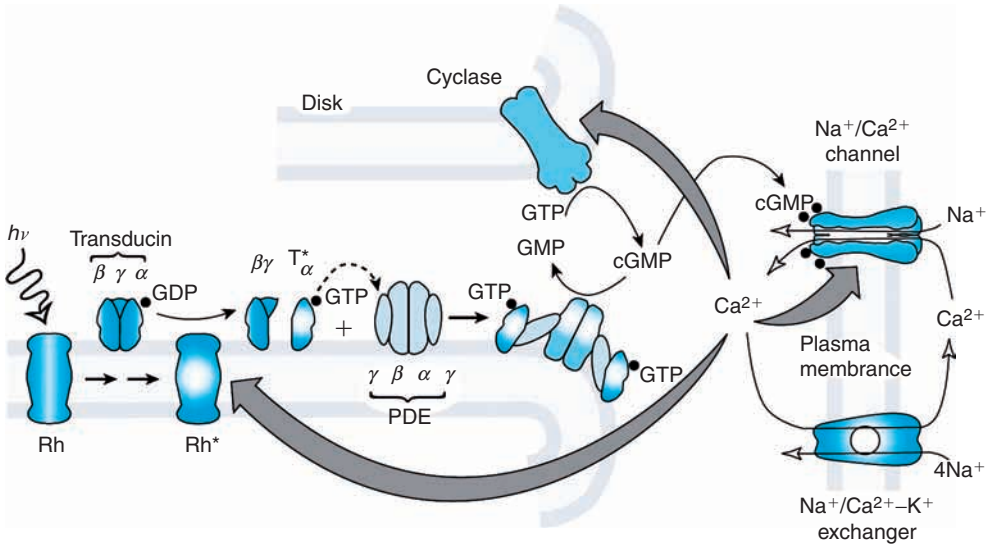

Figure 2 Visual transduction cascade in vertebrate photoreceptor showing sites of modulation by Ca^{2+}. Light activates rhodopsin to produce the active intermediate Rh^*, which in turn stimulates a heterotrimeric G-protein called transducin. Activated transducin Gα-GTP binds to the inhibitory (gamma) subunit of the cGMP phophodiesterase (PDE) and activates the enzyme; the PDE then hydrolyzes cGMP. As the cGMP concentration falls, the cyclic nucleotide-gated channels in the plasma membrane close, reducing the inflow of Na^+ as well as Ca^{2+}. Continued extrusion of Ca^{2+} by the Na^+/Ca^{2+}-K^+ exchange protein causes a decrease in free-Ca^{2+} concentration, which appears to have at least three effects on the cascade (arrows): stimulation of the cGMP cyclase to increase the cGMP concentration, alteration of the effective affinity of the channels for cGMP, and acceleration of the rate of rhodopsin kinase, effectively reducing the lifetime of Rh^*. Adapted from Fain GL (1999) *Molecular and Cellular Physiology of Neurons*. Cambridge, MA: Harvard University Press.

remains depressed for a prolonged period. Sensitivity then gradually recovers, by a process known as 'dark adaptation.' Dark adaptation is reasonably rapid if a minimal percentage of the photopigment has been isomerized, since recovery then requires only the return of the PDE and cyclase to their dark-adapted activities, and the return of the photocurrent and membrane potential to their dark-adapted values. If, however, a substantial fraction of the rhodopsin is isomerized, or bleached, recovery requires in addition the regeneration of the bleached photopigment, and 30–35 min (in humans) may be required for the complete return of sensitivity.

To understand why sensitivity recovers so slowly after a strong bleach, we must first consider what happens to rhodopsin after it receives a photon (Figure 3). The activity of Rh^* is quenched by phosphorylation and binding of arrestin, and in vertebrates all-*trans* retinal then falls off the opsin molecule. The all-*trans* retinal is converted into all-*trans* retinol within the outer segment by an enzyme called retinol dehydrogenase. The all-*trans* retinol then leaves the outer segment and travels to an adjacent cell layer, for rods predominantly to the retinal pigment epithelium (or RPE), but for cones apparently also to retinal glial cells. The transit of retinol seems to be assisted in part by interphotoreceptor retinol-binding protein (or IRBP). In the

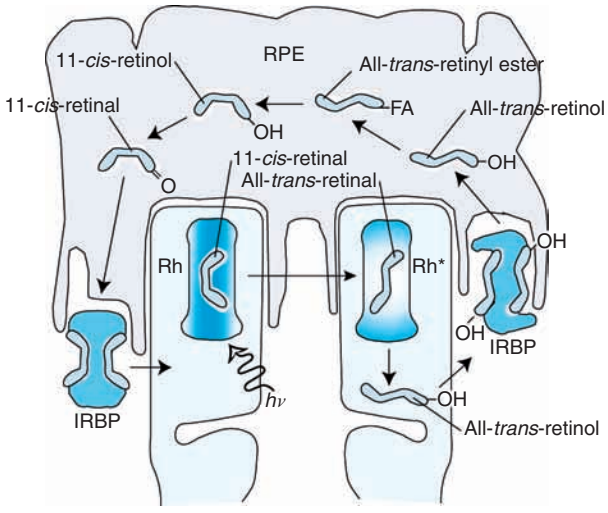

Figure 3 Pathway for regeneration of bleached photopigment in a vertebrate rod. See text. RPE, retinal pigment epithelium; FA, fatty acid; Rh, rhodopsin; IRBP, interphotoreceptor retinol-binding protein; *hν*, light. Reprinted with permission from Fain GL (2003) *Sensory Transduction*. Sunderland, MA: Sinauer.

RPE or glial cell, the all-*trans* retinol is converted back to 11-*cis* retinal by an isomerase enzyme (also known as RPE65), and the 11-*cis* retinal then diffuses or is transported back to the photoreceptors to recombine with opsin.

When a large fraction of the rhodopsin is bleached, the photoreceptor remains desensitized for a prolonged period because large concentrations of phosphorylated rhodopsin, all-*trans* retinol, and opsin accumulate within the outer segment. All of these intermediates are capable of activating the transduction cascade. The continued activation of the cascade produces an 'equivalent background,' which is in effect very much like a maintained background light, desensitizing the photoreceptor (as in Figure 1) by reducing the Ca^{2+} concentration, and modulating the components of the transduction cascade (as in Figure 2). Phosphorylated rhodopsin produces a small, continued activation of the cascade much like that produced by real light, perhaps as a result of the transient unbinding of arrestin or dephosphorylation of the protein. A much greater effect is generated by all-*trans* retinol, which can bind to opsin and actively stimulate the production of transducin Tα-GTP. For small bleaches, phosphorylated rhodopsin and all-*trans* retinal seem to be the most important contributors to the equivalent background. When a large fraction of the photopigment is bleached, the continued reduction of sensitivity seems to be mostly the result of opsin itself, which in both rods and cones can stimulate the visual cascade. Though the probability of stimulation by a single opsin molecule is low, and much lower than for Rh^*, enough opsin accumulates during large bleaches to produce an equivalent background that desensitizes the rod. Simulation by opsin continues until all of the opsin has recombined covalently with 11-*cis* retinal.

See also: Photoreceptor Mosaic; Photoreceptors and Circadian Clocks; Photoreceptors: Physiology; Phototransduction; Retinal Ganglion Cells: Receptive Fields; Rhodopsin.

Further Reading

Fain GL (1999) *Molecular and Cellular Physiology of Neurons.* Cambridge, MA: Harvard University.

Fain GL (2003) *Sensory Transduction.* Sunderland, MA: Sinauer.

Fain GL and Cornwall MC (1993) Light and dark adaptation in vertebrate photoreceptors. In: Shapley R and Lam DMK (eds.) *Contrast Sensitivity, Proceedings of the Retina Research Foundation Symposia,* vol. 5, pp. 3–32. Cambridge, MA: MIT Press.

Fain GL, Matthews HR, Cornwall MC, and Koutalos Y (2001) Adaptation in vertebrate photoreceptors. *Physiological Reviews* 81: 117–151.

Korenbrot JI and Rebrik TI (2002) Tuning outer segment Ca^{2+} homeostasis to phototransduction in rods and cones. *Advances in Experimental Medicine and Biology* 514: 179–203.

Lamb TD and Pugh EN Jr. (2004) Dark adaptation and the retinoid cycle of vision. *Progress in Retinal and Eye Research* 23: 307–380.

Matthews HR (1990) Messengers of transduction and adaptation in vertebrate photoreceptors. In: Herring PJ, Campbell AK, Whitfield M, and Maddock L (eds.) *Light and Life in the Sea*, pp. 185–198. Cambridge, UK: Cambridge University Press.

Pugh EN Jr., Nikonov S, and Lamb TD (1999) Molecular mechanisms of vertebrate photoreceptor light adaptation. *Current Opinion in Neurobiology* 9: 410–418.

Photoreceptor Mosaic

H Hofer, University of Houston, Houston, TX, USA
J Carroll, Medical College of Wisconsin, Milwaukee, WI, USA
D R Williams, University of Rochester, Rochester, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Vision begins in the retina, a delicate, multilayered tissue lining the back surface of the inside of the eye. The photoreceptor mosaic is responsible for absorbing the individual photons of light that comprise the retinal image and transforming their energy into neural signals for processing by retinal circuitry and ultimately the brain.

The mosaics of most animals contain two general types of photoreceptors – rods and cones. Although evidence shows that there are other cells in the retina that are light sensitive (e.g., certain types of ganglion cells), the majority of phototransduction occurs in the rod and cone cells. These highly specialized cells have important structural and topographical differences that subserve their unique functional roles in the visual process. In this article, we compare the structural and functional properties of rods and cones and discuss the spatial organization of the photoreceptor mosaic and the relationship between the cone mosaic and color vision.

Photoreceptor Basics

Photoreceptors consist of an outer segment, an inner segment, a cell body, and a synaptic terminal (Figure 1). The outer segments contain the various components of the visual transduction process. All proteins, including photopigments, are synthesized in the inner segment and transported to the outer segment via the connecting cilium. The outer segment of both rods and cones consists of stacks of membranous discs; in cones the discs are continuous with the plasma membrane of the cell, whereas in rods they are free-floating. Embedded throughout the outer segment membrane in each rod and cone cell are approximately 100 million photopigment molecules, which comprise approximately 85–90% of the total protein structures in the outer segments. In both rods and cones, these discs are synthesized and recycled by the retinal pigment epithelium.

Rod and Cone Specializations

Rod and cone photoreceptors have a number of important specializations that give them distinct functional properties. The primary advantage of having two classes of photoreceptors is to allow the visual system to function over a wide range of light intensities, approximately 11 orders of magnitude in the human eye. Rods are extremely sensitive: They are able to detect a single photon and thus work well at very low light intensities. Approximately ten such photon absorptions (whether across ten rods or within a single rod) are required to detect a visual stimulus under these light levels. At a rate of only one photon per rod every 85 min, a person can see well enough to just barely make out the outline of trees against the sky on a moonless night. On the other hand, cones are less sensitive and function best at high light levels. Cones start to respond when they absorb light at a rate of approximately three photons per cone per second, and they can respond well for light intensities up to 1 million photons per cone per second (the photon catch rate on a bright sunny day is approximately 300 000 photons per cone).

An important difference between rods and cones is that rods 'saturate,' whereas cones do not. At moderate light levels, rods reach a maximum response and are unable to respond to continued increases in light intensity. The point at which rods saturate is quite low, and at normal room light levels, rods are saturated. This means that most of the time your vision is based on your cones, which paradoxically comprise only 5% of the total number of photoreceptors in the human retina.

The human retina is directionally selective. That is, light entering through the center of the pupil is more effectively coupled into photoreceptors, and hence perceived as brighter, than light entering through the pupil margins. This is referred to the Stiles–Crawford effect of the first kind and is a consequence of the waveguide nature of photoreceptors. The rods and cones preferentially guide photons incident within a certain range of acceptance angles along the length of the inner and outer segments, and these photons are therefore more likely to be absorbed by photopigment. The waveguiding properties of photoreceptors are due to a larger index of refraction within the photoreceptor than in the interstitial space between them. Cones are more effective waveguides than rods, so they exhibit a greater degree of directional selectivity. This directional selectivity serves an important function in improving

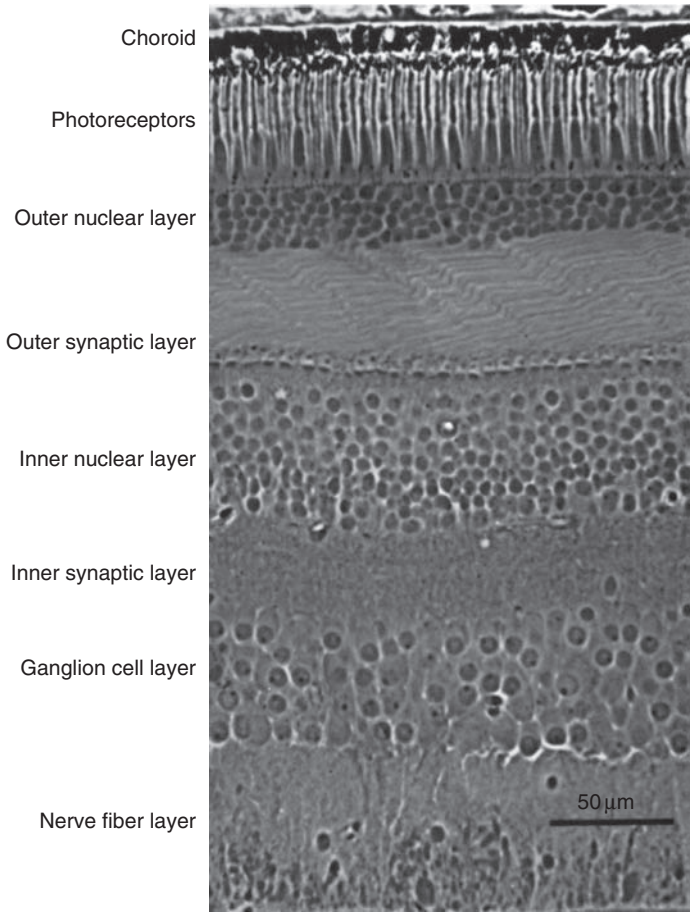

Figure 1 Cross-sectional image of the retina showing the photoreceptors and different cell layers. Reproduced from Boycott BB and Dowling JE (1969) Organization of the primate retina: Light microscopy. *Philosophical Transactions of the Royal Society of London, Series B* 255: 109–194, with permission of the Royal Society of London.

the contrast of the retinal image since obliquely incident, multiply scattered light is not able to effectively couple into cone photoreceptors. This selective rejection of scattered light by cones has a major effect; obliquely incident light is as much as 10 times less effective than light entering through the pupil center. However, this improvement in retinal image contrast comes at the expense of a cone's absolute sensitivity. This loss in cone sensitivity does not affect the sensitivity of the visual system as a whole because of the presence of the more sensitive rod system.

Another important difference between rods and cones regards their spectral response properties, which are determined by the spectral sensitivity of their respective photopigments. Retinal photopigments consist of two parts: a protein (opsin) and a chromophore (11-*cis* retinal in the human retina). Opsins belong to a larger class of proteins called G-protein-coupled receptors. Since all mammalian pigments utilize the same chromophore, it is the opsin portion of the photopigment that determines spectral sensitivity. In the human retina, there are four different photopigment types – rhodopsin (found in rods) and short-, middle-, and long-wavelength sensitive opsin (found in the S, M, and L cones, respectively). Rods have maximum sensitivity at approximately 500 nm, whereas the S, M, and L cones peak near 415, 530, and 560 nm, respectively. A curious feature of the M and L cone pigments is that there is considerable variability within each pigment class. That is, there is not a single stereotyped M (or L) pigment. The variability in spectral sensitivity of these pigments can be directly detected behaviorally or inferred through the use of molecular genetic techniques. The origin of this variation is not well understood, although it almost certainly is related to the genetic mechanisms that give rise to the high incidence of red/green color vision defects.

Topographic Arrangement of Photoreceptors

The previous section outlined the structural characteristics of rods and cones that help differentiate their

functional roles in vision. Each postreceptoral circuit in the retina invariably combines signals from multiple cones in the local patch of retina for which it is responsible. Thus, it is important to understand how rods and cones are organized in the retina in order to construct how the postreceptoral circuitry collects its signals from the mosaic. Next, we review what is known about the topography of the rod and cone mosaics within the human retina. One strategy is to tile the surface of the retina with rods as completely as possible in order to achieve the highest overall sensitivity. Conflicting with this strategy is the need to have a high number of cones in order to achieve high spatial resolution. The manner in which approximately 100 million rods and 5 million cones are interleaved into a single photoreceptor mosaic enables the visual system to function quite well over a large portion of the visual field.

Cone Topography

The cones are distributed unevenly across the retina, with a peak density of approximately 200 000 cones mm^{-2} at the fovea, which is located approximately 15 degrees temporal to the optic disc. Here, cones are packed in a regular triangular array, with the average cone diameter being approximately 2 μm (although this of course depends on the cone density, which is highly variable in central fovea). Cone density decreases precipitously moving away from the fovea in all directions (Figure 2), declining to approximately 30 000 cones mm^{-2} at approximately 400 μm eccentricity. As cone density decreases, the regularity of the mosaic declines and the average cone diameter increases. Cone distribution has a slight radial asymmetry about the fovea, with the nasal and superior retinal quadrants having higher cone density. Beyond approximately 10 degrees, the density of cones reaches a near constant value of 7000 cones mm^{-2} until the extreme edge of the neural retina, where there is an increase to nearly 15 000 cones mm^{-2}. Whether this latter feature of cone topography is an unintended consequence of retinal development or a functional specialization is uncertain.

The foveal specialization has a number of features that serve to optimize spatial vision. First, there are no rods in the central fovea, allowing the cones to be tightly packed in a locally triangular mosaic, with each cone usually surrounded by six neighbors. Second, the inner retinal cells and retinal vasculature are displaced away from the central fovea, providing a clear path for incident photons onto the foveal cone mosaic.

The developmental mechanisms behind the formation of the fovea have been well studied. Although the future location of the fovea can be molecularly identified as early as 24 weeks postgestation, the process of foveal formation is not complete until approximately 4–7 years after birth. The last steps of foveal development include the physical migration of cones toward the fovea – this process results in the characteristic spatial distribution of cones across the retina as discussed previously (Figure 3).

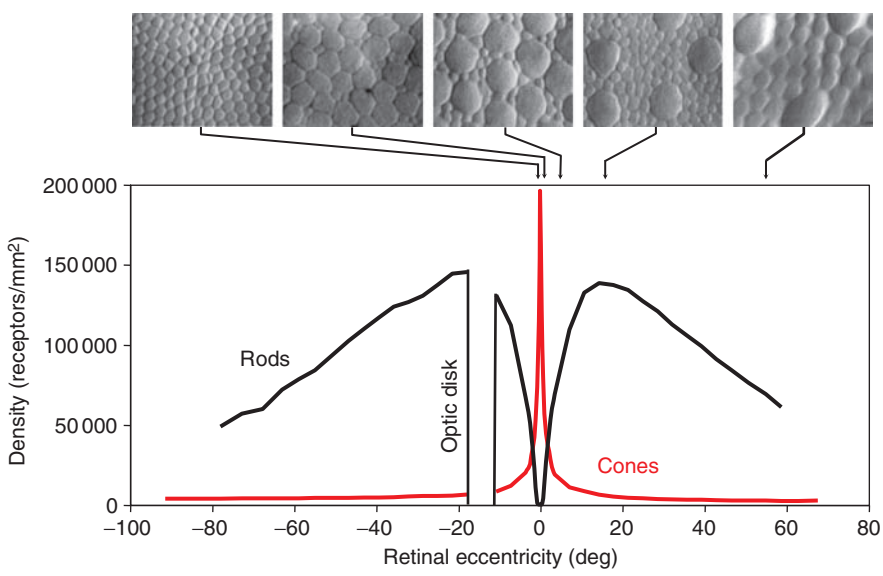

Figure 2 Plot of photoreceptor density as a function of retinal eccentricity. Top panels represent images of the photoreceptor mosaic. The image on the left is of the all-cone foveola, whereas the remaining panels contain both rods and cones (cones are the larger structures). Notice that at greater eccentricities, as the density of cones decreases and density of rods increases, the cones increase in diameter while the rods remain relatively constant in diameter. (Top panels) Reproduced from Curcio CA, Sloan KR, Kalina RE, et al. (1990) Human photoreceptor topography. *Journal of Comparative Neurology* 292: 497–523. with permission from John Wiley and Sons.

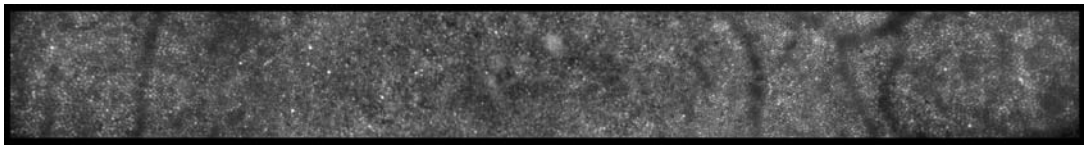

Figure 3 Montage showing the distribution of cones in the human retina across the central 5 degrees in one subject. This montage was created by adding many individual 1-degree *in vivo* images acquired with an adaptive optics retinal camera. The dark shadows are due to blood vessels overlying the cone mosaic.

Despite the involvement of the cone photoreceptors in a number of diseases (including age-related macular degeneration, which affects three in ten people older than the age of 75 years), there does not appear to be significant 'normal' loss of cones during aging.

Rod Topography

Rods are absent in the very center of the fovea; the 'rod-free zone' is on average 350 μm in diameter. Moving away from the fovea, rod density increases to a maximum of approximately 180 000 cells mm^{-2} at approximately 12-degree eccentricity. Unlike cones, whose size changes as a function of eccentricity, the size of rods remains nearly constant across the retina at approximately 2 μm in diameter. It has been shown that there is normal loss of rods with aging, although it is unclear why rods (but not cones) show this susceptibility.

Effects of the Receptor Mosaic on Spatial Vision

In addition to the topography of the receptor mosaic, the eye's optics as well as neural factors place limits on spatial vision. The optics of the typical human eye are not capable of focusing a point of light on the retina smaller than approximately 5 μm, and the highest spatial frequency that can be transmitted through the eye's optics to the retina is approximately 60–120 cycles per degree depending on pupil size. The packing of cones also places limits on spatial vision; the highest spatial frequency that the photoreceptor mosaic can accurately sample is one-half the inverse of the cone spacing. This limit, known as the Nyquist limit, is approximately 60 cycles per degree for human foveal vision. Whereas the limits imposed by the eye's optics and the typography of the mosaic are usually well matched in foveal vision, in the peripheral retina the eye's optics are able to form images on the retina with higher spatial frequencies than the peripheral mosaic can accurately sample. It is sometimes assumed that in the absence of optical factors the Nyquist limit determines acuity. However, this is not necessarily the case; for example, rods are as densely packed as foveal cones in much of the retinal mosaic and thus share the same Nyquist limit, yet the acuity of the rod system is an order of magnitude poorer than the cone system due to neural factors, specifically the postreceptoral spatial pooling of rod signals. The receptors also blur the retinal image by averaging all the light that lands within their collection area; however, at the fovea this effect impacts the retinal image minimally because the diameter of a foveal cone (and hence the averaging area) is small compared with the finest detail that can be transmitted through the eye's optics.

Cone Submosaics and Color Vision

Human color and spatial vision is mediated by three spatially distinct classes of cone photoreceptors with different but overlapping spectral sensitivities – the L, M, and S cones. The first stage of color vision is light absorption in one or more of these cone photoreceptors. Once a photon excites a cone photopigment molecule, all information about the wavelength of that photon is lost, and hue can only be recovered by comparing the relative excitations of the three different cone types. Thus, normal color vision is trichromatic since there are only three independent quantities available to the visual system from which the appearance of any stimulus may be inferred. The trichromacy of color vision was established empirically, by noting that humans can match the appearance of any patch of light by a suitable mixture of only three primary lights, long before it was possible to directly confirm that there are indeed three distinct cone types on which color vision depends. There are important differences in the relative numbers and arrangement of the three cone types that have implications for color vision.

S Cones

S cones exhibit a number of morphological and histochemical features that easily distinguish them from L and M cones. They are more cylindrical in shape, serve distinct neural circuitry, and form a spatially independent submosaic from the L and M cones. S cones are relatively sparse in the human retina, with peak density (usually less than 10% of the total cone number) near 1-degree eccentricity and very little intersubject variability. This low S cone density does not significantly impact vision since the shorter

wavelengths the S cones are most sensitive to are those that are most blurred by chromatic aberration, so fewer S cones are needed to sample the relatively coarse short-wavelength image. The density of S cones is lowest in the central fovea, with an S cone-free zone of approximately one-third of a degree existing in most, although not all, human subjects. It is not clear what governs the nonuniform topography of S cones. Whereas S cones tend to be regularly distributed throughout the macaque retina, in humans S cone placement is random near the fovea, with a more regular arrangement as eccentricity increases. This irregularity may reflect disruption during the foveal migration of L and M cones during retinal development. Some have found that S cones, presumably due to their increased size, can disrupt the otherwise normal triangular packing arrangement of the L and M cone array, although this relationship is not always observed.

L and M Cones

Number and arrangement L cones are generally more numerous than M cones in the human retina and tend to occur on average about twice as frequently. Due to the high degree of similarity between L and M cones (it is widely believed that the only difference between an L and an M cone is the particular photopigment it contains), less is known about their retinal topography and neural connectivity than for S cones. Studies that have been performed to date indicate that the L and M cone submosaics are generally randomly interleaved, at least near the fovea, and that the relative numbers of L and M cones are generally constant across the fovea and midperiphery of an individual's retina. However, there do seem to be exceptions to both these rules in at least some individuals. It has been suggested that the L and M cone submosaics may exhibit some degree of clumping in the periphery, based on observations made in macaque retina; however, no direct observations have been made of the relative numbers or arrangement of L and M cones in midperipheral human retina.

Inter- and Intraperson variability Numerous indirect estimates have suggested that there are on average approximately twice as many L cones as M cones in the human retina, with large intersubject variability. Direct information on the numbers and locations of L, M, and S cones obtained with spatially localized retinal densitometry in 10 living human subjects reveals the extent of this variability in L-to-M cone ratio across individuals with normal color vision to be more than a 40-fold range.

We have described the large variability in the ratio of L to M cones across different individuals, but we have not addressed the variability of L-to-M cone ratio within a particular individual's retina. Most studies suggest that the ratio of L to M cones is probably relatively constant across the central retina; however, evidence for this is largely inferential. Direct evidence is available for only two human subjects, one of whom exhibited a significant difference in local L-to-M cone ratio between temporal and nasal retina and the other did not. This indicates that L-to-M cone ratio need not be homogeneous across the central retina and that some individuals may exhibit a nasal–temporal asymmetry in the local L-to-M cone ratio similar to that observed in macaque retina, but how common this asymmetry is remains to be determined. There have been no direct measures of the variability of L-to-M cone ratio with eccentricity in the central and midperipheral retina; however, it is well-known that the ratio of L to M cones increases significantly in the far periphery.

Consequences for color appearance The relative numbers and organization of the three cone classes in the human retina place important constraints on color vision and the neural circuitry responsible for it. Remarkably, color experience is essentially uniform across individuals despite the huge variability in the ratio of L to M cones. For example, unique yellow, which represents the neutral point of the red-green color mechanism and is thought to be driven mainly by differences in L and M cone excitation, varies with a standard deviation of only 2–5 nm. It has been suggested that this uniformity is a consequence of a plastic mechanism that normalizes color experience based on the average spectral content of the environment. In individuals with very extreme L-to-M cone ratios, it is possible to observe some subtle effects of L-to-M cone ratio on vision. For example, there are reports that these individuals have compromised color contrast sensitivity and are more likely to have greater matching ranges on the anomaloscope than those with more balanced ratios. Due to the irregular assignment of L and M cones, these subjects would also tend to have larger patches of effectively color-blind retina (in which only one cone class is present) than those with more balanced numbers of L and M cones, meaning that these individuals would be more prone to misjudging the color appearance of very small objects.

Neural consequences The disordered arrangement of L and M cones seems to strike a balance between the needs of spatial and color vision. As can be seen in Figure 4, random cone arrangement results in patches of retina in which only the L or M cone class is present. Although this patchiness can be

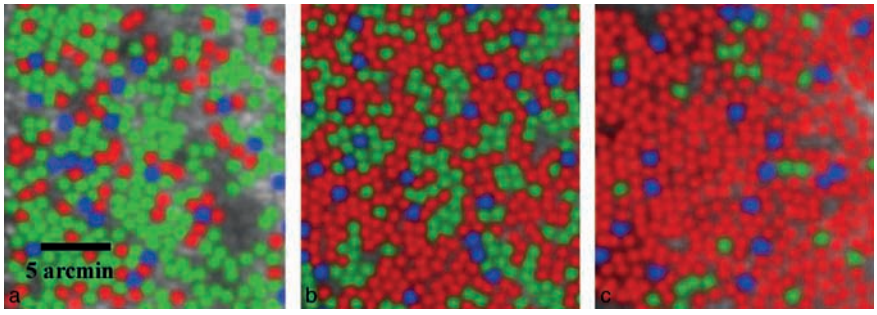

Figure 4 False color images of the trichromatic cone mosaic for three subjects illustrating the large variability in the ratio of L (red) to M (green) cones among individuals with normal color vision. S (blue) cone number is relatively uniform across subjects. The ratio of L to M cones for these subjects is (a) 2.7:1, (b) 1.9:1, and (c) 16.5:1. L, M, and S cones were identified with adaptive optics imaging combined with spatially localized retinal densitometry. Adapted from Hofer H, Carroll J, Neitz J, et al. (2005) Organization of the human trichromatic cone mosaic. *Journal of Neuroscience* 25(42): 9669–9679.

disadvantageous for foveal color vision, it is likely beneficial for high-acuity spatial vision because the neurons responsible for fine spatial vision are more likely to draw their input from cones of the same spectral type, thus avoiding the confounding of spectral and spatial information. This patchiness also creates retinal locations where the resolution of the M or L cone submosaic should be as high as that for the full mosaic, as is observed experimentally.

In the peripheral retina, there are additional advantages conferred by a disordered arrangement of L and M cones. There is a great degree of neural convergence in the periphery so that several neighboring cones drive the center of midget ganglion cells, which are believed to carry the red-green color signal. If more than one cone type contributes to the ganglion cell center response, the chromatic opponency of that cell will be weakened. Thus, red-green color vision would be expected to deteriorate in the periphery. The patchiness in the mosaic that results from random L and M cone arrangement lessens the negative effects of this convergence because it makes it more likely that some peripheral midget cone centers will be driven by cones of the same spectral type. This may explain why color discrimination in the periphery is as good as in the fovea if large enough stimuli are used, although peripheral color vision is significantly disadvantaged for smaller fields.

Despite this apparent advantage for peripheral color discrimination, patches of a single cone type can also be detrimental for color vision, especially in the fovea, because the retina is effectively color blind within them. The irregularity of L and M cone arrangement exacerbates the confounding of spatial and chromatic information on small scales, leading to errors in color appearance for small stimuli or fine spatial patterns. Chromatic aliasing is the general term that refers to perceptual errors of this type; examples are variability and misjudgment of the appearance of small points of light and the Brewster's color phenomenon in which splotches of pastel colors are erroneously seen superimposed on fine black-and-white gratings.

Cone Mosaic in Color-Blind Individuals

Color blindness – general definitions We have described the photoreceptor mosaic and its implications for vision for the normal human retina; we now discuss the photoreceptor mosaic of the color-blind retina. Color blindness refers to a variety of conditions that result in an individual requiring fewer than three primaries to match the appearance of a test light. Although color blindness can be caused by acquired dysfunction in the chromatic pathways, we restrict our discussion to the more common case in which color blindness is an inherited condition caused by the absence or malfunction of one or more of the three types of cones. Dichromacy refers to the absence of one of the cone classes, whereas monochromacy refers to the functional absence of two or even all three cone types. Dichromats can still discriminate a limited range of colors and they require two primaries to match the appearance of any test flash. Monochromats, who are extremely rare, completely lack color vision and can match the appearance of any colored test flash by adjusting the intensity of a single primary.

There are also a group of conditions known as color anomalies in which a person is still trichromatic – that is, the person still requires three primaries to match the appearance of a test light, but the matches that person makes tend to be abnormal and often the matches are made with less precision. A color anomalous individual has three different types of cones but they contain unusual or very similar pigments. Although technically referred to as color blind, these individuals often have quite good color discrimination.

Examples There are three types of dichromacy, representing the loss of L, M, or S cone function.

Loss of either L or M cone function (termed deuteranopia or protanopia, respectively) is referred to as 'red-green' color blindness and is the most common form of color blindness. This is because the pigment genes for L and M cones are found on the X chromosome. It affects approximately 2% of males and 0.03% of females, with approximately 4.5% of females with normal color vision carrying genes for red-green dichromacy. Tritanopia, the form of dichromacy resulting from the complete loss of S cone function, is much more rare and shows equal incidence across gender since the S pigment gene is located on an autosome (chromosome 7).

Historically, there has been some debate about whether the cone mosaic of red-green color-blind individuals contained a full complement of either only L or only M cones (replacement hypothesis) or whether the mosaic contained empty spaces representing the lost class of cones (loss hypothesis). We now know that both of these situations can occur depending on the specific genetic mutation underlying an individual's color blindness. For example, a protanope who does not have L cones because he lacks genes for L cones and instead has multiple genes for M cones will display a full complement of cones in his retina. On the other hand, a deuteranope, who lacks functional M cones because the M pigment gene he has does not encode a functional M pigment, displays gaps in his retinal mosaic that seem to represent the missing M cones (Figure 5). Moreover, there may be important phenotypic differences among individuals with nonfunctioning cone pigments that depend on the severity of the particular mutation.

Also in the family of red-green color vision defects is anomalous trichromacy, which is associated with the loss of L or M cone function and is referred to as deuteranomaly or protanomaly, respectively. There is no good evidence that tritanomaly exists. Deuteranomaly is characterized by the presence of two different types of functional L cones (and S cones), whereas protanomaly is characterized by the presence of two different types of M cones (and S cones). Both are a result of a rearrangement of the X-linked L/M gene array and as such are quite common. It affects approximately 5.5% of males and 0.39% of females, with approximately 11% of females with normal color vision carrying genes for red-green anomalous trichromacy. The severity of the defect has been linked to the spectral separation of the two L or two M cone pigments. Recall that there is not a single stereotyped L or M cone. Thus, for example, a deuteranomalous trichromat could have two L cone pigments that differ in maximum sensitivity by as much as 10 nm or as little as 1 or 2 nm. Likewise, protanomals can have two M pigments that differ by as much as 5 nm or as little as 1 or 2 nm. In general, the farther apart an individual's pigments, the better their residual color discrimination.

There are also three classes of monochromacy: blue cone monochromats, rod monochromats, and cone monochromats. Blue cone monochromats have rods and one functional class of cones (S cones), whereas rod monochromats have only rod function. In addition to being color blind, these individuals have low resolution, tend to be photophobic, and usually have nystagmus. Both of these conditions are very rare; rates have been estimated at 1 in 50 000 for rod monochromacy and 1 in 100 000 for blue cone monochromacy. The few histological studies that have been performed on the retinas of rod monochromats have

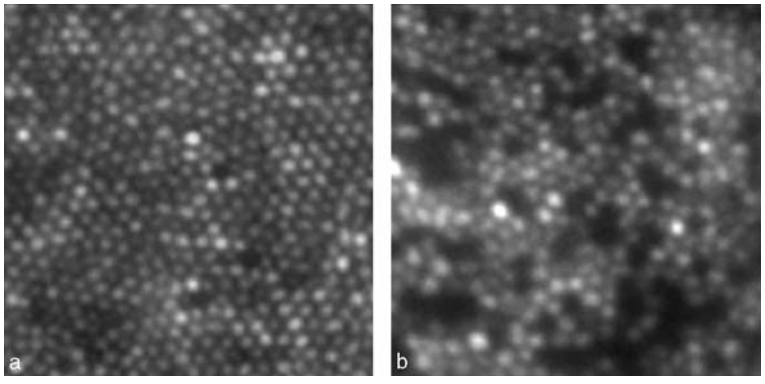

Figure 5 Images of the cone mosaic in two individuals with red-green color blindness. Each subject's mosaic contains only S cones and either L or M cones. The first subject (a) has multiple genes for M cone pigments and no genes for L cone pigment; he exhibits a full complement of cones in his mosaic. The second subject (b) has a normal gene for L cone pigment, but his M pigment gene encodes a faulty pigment; his mosaic contains numerous gaps that appear to represent the missing class of M cones. Interestingly, this loss of approximately one-third of the cone photoreceptors does not show up on any clinical test of visual function and this subject enjoys normal visual acuity. Images were acquired with a high-resolution adaptive optics ophthalmoscope. Adapted from Carroll J, Neitz M, Hofer H, et al. (2004) Functional photoreceptor loss revealed with adaptive optics: An alternate cause for color blindness. *Proceedings of the National Academy of Sciences of the United States of America* 101(22): 8461–8466.

given discrepant results for the organization of the retinal mosaic. In some cases, cones were scarce and malformed, whereas in other cases cones seemed present in normal numbers. Rod monochromacy has been shown to be caused by defects in one of two components of the visual transduction cascade in cones – the cone cyclic nucleotide-gated channel (either the α or β subunit) or the cone transducin. Numerous mutations in these proteins have been reported to cause rod monochromacy, and one idea is that the different genetic causes of the disease would lead to subtle phenotypic differences and thus might be responsible for the observed variability in retinal organization among rod monochromats. No direct information is available on the organization of the photoreceptor mosaic in blue cone monochromats. Because of the associated low visual acuity and nystagmus, and since there is no known mechanism that would allow missing L and M cone pigment to be replaced by S cone or rod pigment, the mosaics of these individuals are expected to contain rods and a sparse array of S cones with no other cones present. Cone monochromats have either only L or only M cones and rods; only a few cases have been described and none have had their authenticity completely accepted. These people have been reported to have normal visual acuity and no nystagmus, and they presumably have photoreceptor mosaics similar to regular red-green dichromats except for the lack of S cones. This disease would occur due to an individual inheriting both tritanopia and one form of red-green color blindness. The incidence of cone monochromacy is believed to be approximately 1 in 10 million, but this is based solely on the likelihood of inheriting both tritanopia and red-green dichromacy.

See also: Animal Models of Inherited Retinal Degenerations; Color Vision; Cone Photopigment Evolution; Photoreceptor Adaptation; Photoreceptors and Circadian Clocks; Photoreceptors: Physiology; Phototransduction; Representation of Color; Retina: An Overview; Retinal Color Mechanisms; Retinal Development: An Overview; Retinitis Pigmentosa; Rhodopsin.

Further Reading

Ahnelt PK and Kolb H (2000) The mammalian photoreceptor mosaic-adaptive design. *Progress in Retinal and Eye Research* 19(6): 711–777.

Boycott BB and Dowling JE (1969) Organization of the primate retina: Light microscopy. *Philosophical Transactions of the Royal Society of London, Series B* 255: 109–194.

Carroll J, Neitz M, Hofer H, Neitz J, and Williams DR (2004) Functional photoreceptor loss revealed with adaptive optics: An alternate cause for color blindness. *Proceedings of the National Academy of Sciences of the United States of America* 101(22): 8461–8466.

Curcio CA and Allen KA (1990) Topography of ganglion cells in human retina. *Journal of Comparative Neurology* 300: 5–25.

Curcio CA, Allen KA, Sloan KR, et al. (1991) Distribution and morphology of human cone photoreceptors stained with anti-blue opsin. *Journal of Comparative Neurology* 312: 610–624.

Curcio CA, Sloan KR, Kalina RE, et al. (1990) Human photoreceptor topography. *Journal of Comparative Neurology* 292: 497–523.

Gegenfurtner KR and Sharpe LT (1999) *Color Vision: From Genes to Perception*. New York: Cambridge University Press..

Hendrickson AE and Yuodelis C (1984) The morphological development of the human fovea. *Ophthalmology* 91(6): 603–612.

Hirsch J and Curcio CA (1989) The spatial resolution capacity of human foveal retina. *Vision Research* 29(9): 1095–1102.

Hofer H, Carroll J, Neitz J, Neitz M, and Williams DR (2005) Organization of the human trichromatic cone mosaic. *Journal of Neuroscience* 25(42): 9669–9679.

Nathans J, Thomas D, and Hogness DS (1986) Molecular genetics of human color vision: The genes encoding blue, green, and red pigments. *Science* 232: 193–202.

Polyak SL (1941) *The Retina*. Chicago: University of Chicago Press.

Provis JM, Diaz CM, and Dreher B (1998) Ontogeny of the primate fovea: A central issue in retinal development. *Progress in Neurobiology* 54: 549–581.

Pum D, Ahnelt PK, and Grasl M (1990) Iso-orientation areas in the foveal cone mosaic. *Visual Neuroscience* 5: 511–523.

Putnam NM, Hofer HJ, Doble N, Chen L, Carroll J, and Williams DR (2005) The locus of fixation and the foveal cone mosaic. *Journal of Vision* 5(7): 632–639.

Rodieck RW (1998) *The First Steps in Seeing*. Sunderland, MA: Sinauer.

Roorda A and Williams DR (1999) The arrangement of the three cone classes in the living human eye. *Nature (London)* 397: 520–522.

Westheimer G (2006) Specifying and controlling the optical image on the human retina. *Progress in Retinal and Eye Research* 25(1): 19–42.

Williams DR, MacLeod DIA, and Hayhoe M (1981) Foveal tritanopia. *Vision Research* 21: 1341–1356.

Williams DR, MacLeod DIA, and Hayhoe MM (1981) Punctate sensitivity of the blue-sensitive mechanism. *Vision Research* 21: 1357–1375.

Williams RW (1991) The human retina has a cone-enriched rim. *Visual Neuroscience* 6: 403–406.

Relevant Website

http://webvision.med.utah.edu – WebVision: The Organization of the Retina and Visual System.

Photoreceptors and Circadian Clocks

S N Peirson and R G Foster, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Almost all living creatures, including some prokaryotic forms of life (cyanobacteria or blue-green algae), have been shown to possess an endogenous 24 h biological clock, called a 'circadian clock' or 'circadian pacemaker.' The term 'circadian' was originally coined by Frans Halberg in 1959 and is derived from the Latin *circa*, 'about,' and *diem*, 'day.' A circadian rhythm describes an approximately 24 h oscillation that persists under conditions of light and darkness for several cycles. The circadian system provides an endogenous representation of a day and sets the times or phases at which physiological and behavioral events occur with respect to the 24 h environmental cycle. In this way, physiology is 'fine-tuned' to the varying demands of night and day. This facility will only have survival value if biological time remains synchronized or entrained to environmental time. The systematic daily change in the gross amount of light (irradiance) at dawn or dusk provides the primary indicator of the time of day. As a result, most organisms have evolved to use the twilight transition as the main zeitgeber (time giver) to adjust circadian time to local time – a process that has been termed photoentrainment.

The site of the primary circadian clock in mammals is the suprachiasmatic nuclei (SCN) of the hypothalamus; this site receives input from the retina via the retinohypothalamic tract (RHT). The clock within the SCN is a product of intracellular transcriptional–translational feedback loops, involving a number of core clock genes. Despite the fact that the SCN are essential for overt rhythmicity, with ablation of these nuclei resulting in behavioral arrhythmicity, clock genes have also been found to be rhythmically expressed in numerous other tissues, including the heart, liver, lung, and kidney. These peripheral clocks are dependent upon the SCN to maintain appropriate phasing, and in the absence of this central coordinator the individual clock cells become uncoupled or desynchronized, which results in arrhythmic non-24 h physiology (Figure 1).

Circadian Photoreception

The photoreceptors providing the light input to the mammalian circadian system are ocular. In this respect mammals are quite unlike other vertebrate classes, which possess multiple photoreceptors located in structures such as the pineal complex and deep brain. From time to time, extraretinal photoreceptors have been reported to exist in mammals. The most recent report suggested that photoreception could occur in the skin behind the knee in humans. Although widely reported as fact and even incorporated into some textbooks, this finding could never be replicated; most notably, Czeisler's group at Harvard duplicated the methodology precisely and could not achieve the reported finding.

Demonstration that the eyes provide the exclusive light input to the circadian system has been shown repeatedly by enucleation. Eye loss always completely blocks photoentrainment in those mammals examined. Although the photoentrainment process is clearly located within the eye, establishing which ocular cells bring about photoentrainment has been surprisingly difficult. Historically, the only components of the mammalian central nervous system thought to be directly light sensitive were the rods and cones (Figure 2). As a result, all light responses have been attributed to one or both of these photoreceptor classes. Rods mediate scotopic (dim light) vision, providing low resolution but high sensitivity, whereas cones are involved in photopic (bright light) vision and are found at high density in the human fovea, enabling high-resolution color vision. Photoreception by rods and cones is rapid, responding to bright, short-duration stimuli and providing a topographically mapped output via the optic nerve. In addition, signals from the outer retina are processed by the inner retinal neurons, providing the initial stages of image construction, before being relayed via the axons of the retinal ganglion cells (RGCs) to the visual centers of the brain (Figure 2).

By contrast, the RHT projection to the SCN is composed of a small number of morphologically distinct RGCs. These cells (~1% of the total) are evenly distributed across the entire retina and are not topographically organized. Unlike the visual system, the circadian system needs relatively bright, longduration stimuli to achieve photoentrainment and is capable of temporally integrating signals over extended periods of time. Rather than detecting patterns of incoming light from a specific source (radiance), the circadian system detects only gross changes in environmental light (irradiance). A summary of the key differences between circadian and visual photoreception is provided in Table 1. These marked differences between image-forming vision and circadian

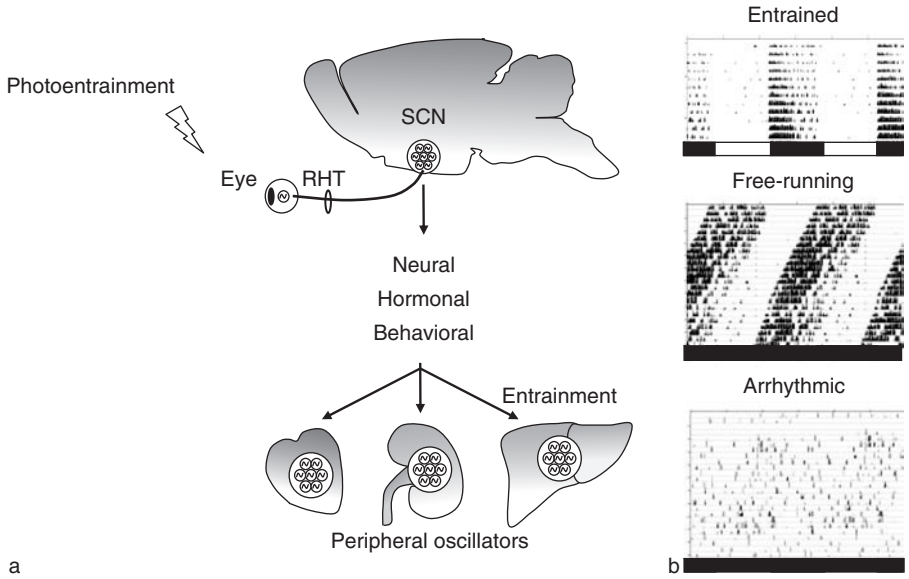

Figure 1 Schematic of the mammalian circadian system. (a) Environmental light information is detected by the eye and is transmitted to the suprachiasmatic nucleus (SCN) via the retinohypothalamic tract (RHT). The SCN is the site of the central circadian oscillator, based upon a molecular clock formed by the transcriptional–translational feedback loop of core clock genes and their proteins. Numerous additional tissues contain similar molecular clocks, and the SCN acts to coordinate these peripheral clocks to ensure internal synchronization. (b) Actograms illustrating mouse activity records. Under light–dark conditions, activity is entrained and occurs only during the dark phase. In constant darkness or following eye loss, an animal will free-run with an endogenous period of slightly less than 24 h. Following loss of the SCN, circadian rhythms are abolished, and the animal is behaviorally arrhythmic.

light detection have not been easy to explain, thus leading circadian biologists to investigate how rods and cones might be capable of mediating these two markedly different forms of light detection. The results from these studies are somewhat surprising.

A Novel Photoreceptor System

Initial studies to determine which retinal photoreceptors mediate circadian responses to light took advantage of a naturally occurring mutation in mice, termed retinal degeneration (*rd/rd*). These animals lack rods and show a greatly reduced number of cones. As might be expected, *rd/rd* mice fail to show any classical visual responses to light. Studies were thus undertaken to determine if the circadian system of *rd/rd* mice was similarly impaired. When a mouse is housed under a light–dark (LD) cycle, its circadian wheel-running behavior is entrained. As would be expected of a nocturnal species, mice are largely active during the dark and inactive in the light. Under conditions of constant darkness, entrainment is lost and circadian behavior drifts or free-runs with a period around 23.5 h. If the mouse is then exposed to a pulse of light shortly after activity onset (when the clock 'thinks' it is evening), it will delay the onset of activity the following day (Figure 3). The magnitude of this phase delay ($\Delta\phi$) is intensity dependent and can be used to determine the sensitivity of circadian responses to light. Both the wavelength and intensity of the light pulse can be systematically varied and the effect of these treatments on delaying wheel-running behavior can be assessed in a dose-dependent manner to produce a series of irradiance response curves (Figure 3).

Remarkably, the massive loss of classical photoreceptors in *rd/rd* mice had little or no effect on the ability of mice to phase-shift circadian rhythms in wheel-running behavior. Their irradiance response curves were indistinguishable from those of congenic wild-type controls. Eye loss completely blocked all effects of light on the clock. These initial studies demonstrated that the processing of light for circadian and visual responses must be different, and hinted at the fact that there may be another class of ocular photoreceptor. Such suggestions were met with considerable scepticism and the favored explanation was that, as ~5% of the cones survive in the peripheral retina of *rd/rd* mice, it is likely that only a small number of photoreceptors are necessary for photoentrainment. The development of a mouse model lacking all rods and cones, the *rd/rd cl*, resolved these issues (Figure 4) and demonstrated that both phaseshifting responses and pineal melatonin suppression in response to light were apparently normal. These data provided the conclusive evidence that an

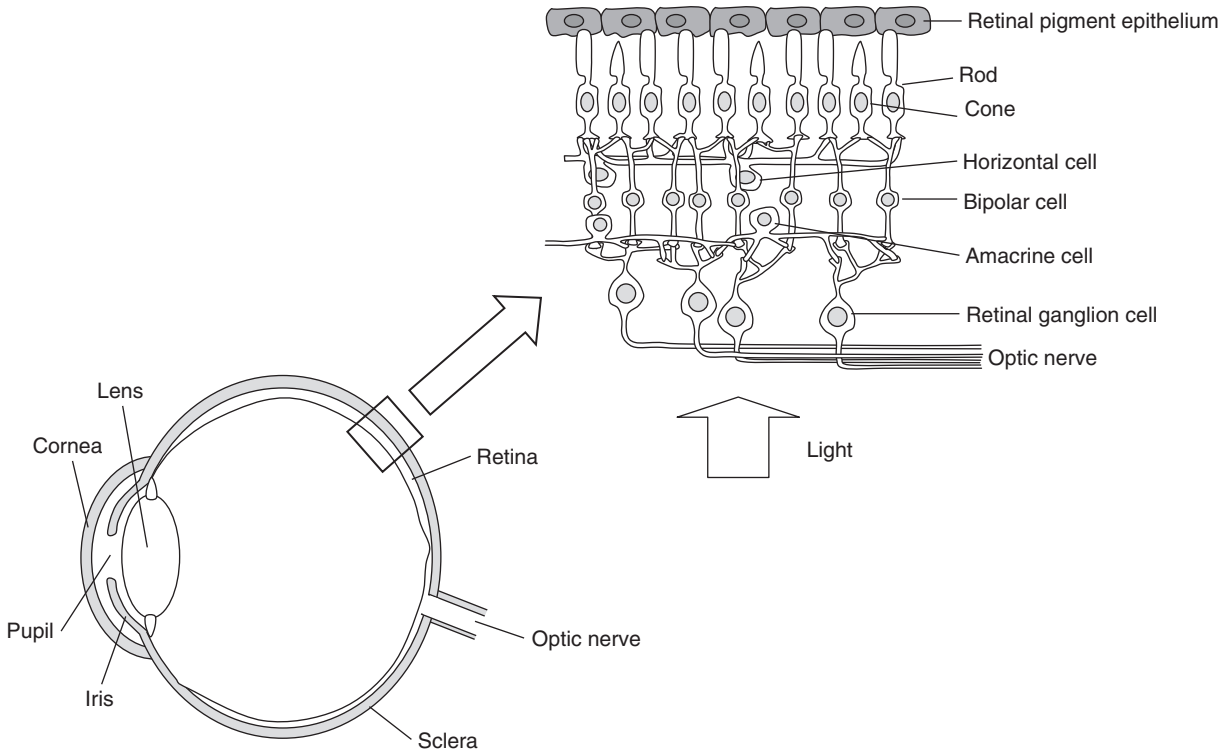

Figure 2 The mammalian eye and retina. Light entering the eye is detected by photoreceptors in the outer retina, the rods and cones. These classical photoreceptors project to bipolar cells, which in turn synapse with retinal ganglion cells. Horizontal cells and amacrine cells are involved in retinal processing. The axons of retinal ganglion cells form the optic nerve projecting to the visual centers of the brain.

Table 1 Response characteristics of image-forming vision and circadian photoreception

Image-forming vision	*Circadian photoreception*
Detects radiance (directional)	Detects irradiance (nondirectional)
Rapid response (milliseconds) to brief stimuli	Insensitive to brief stimuli
Sensitive to dim light (rods)	Relatively insensitive to dim light
Short integration time (seconds)	Long integration time (minutes)
Primary projection via optic tract to LGN and SC, then on to visual cortex	Projects via retinohypothalamic tract to suprachiasmatic nucleus
Output is topographically mapped	Output is unmapped

additional photoreceptor exists within the mammalian eye, and further studies showed that these novel receptors do more than regulate the clock.

In mammals, light-induced pupil constriction is regulated by both rods and cones, but still occurs in animals showing profound damage to these photoreceptors. Not unreasonably, it was assumed that the residual pupil light response was due to the survival of a few rod and/or cone photoreceptors. Pupil measurements were undertaken in *rd/rd cl* mice and showed that these animals maintained a pupillary light response (PLR). Although less sensitive than congenic wild-type animals, *rd/rd cl* mice retained the ability to fully constrict their pupils. Nocturnal rodents will inhibit their general activity when exposed to light during the night. This response, called masking, is thought to complement circadian entrainment by ensuring that activity is restricted to the hours of darkness or near darkness. Masking may be particularly important in environments in which day length changes rapidly and circadian behavior may have difficulty keeping up with the expanding photoperiod. Masking experiments were undertaken in *rd/rd cl* mice and demonstrated that there is marked inhibition of activity upon exposure to light presented 2 h after normal 'lights off.' Thus, phase shifting, melatonin suppression, pupil constriction, masking, and probably a number of other systems are either intact or retained at some level in mice lacking all of their rods and cones. The persistence of such diverse responses to light suggests that the novel photoreceptors might contribute to many more aspects of mammalian physiology and behavior than previously suspected. For example, the irradiance of light has been shown to modulate sleep,

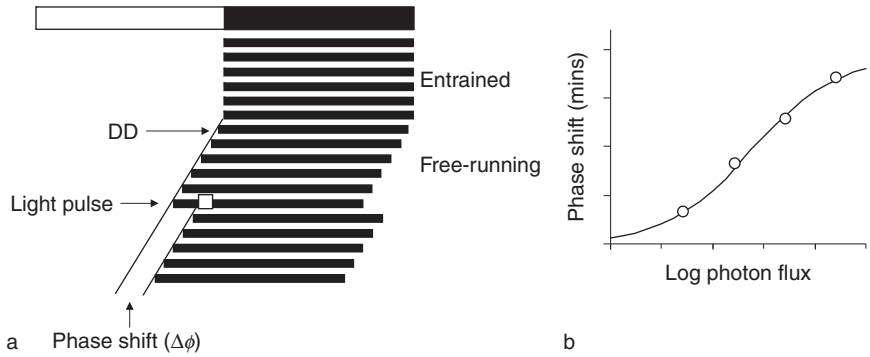

Figure 3 Schematic actogram illustrating mouse locomotor activity. (a) Under a light–dark cycle activity is confined to the dark phase of the 24-h cycle. However, in constant darkness (DD) mice free-run with a period of slightly less than 24 h. If exposed to a brief light pulse in the early evening, the animal will delay its activity onset the following day. The phase shift produced in this manner can be used to assess the sensitivity of the circadian system to light of different wavelengths and intensities. (b) The magnitude of the phase shift is proportional to the intensity of the irradiance provided by the light pulse, and can be plotted as an irradiance response curve.

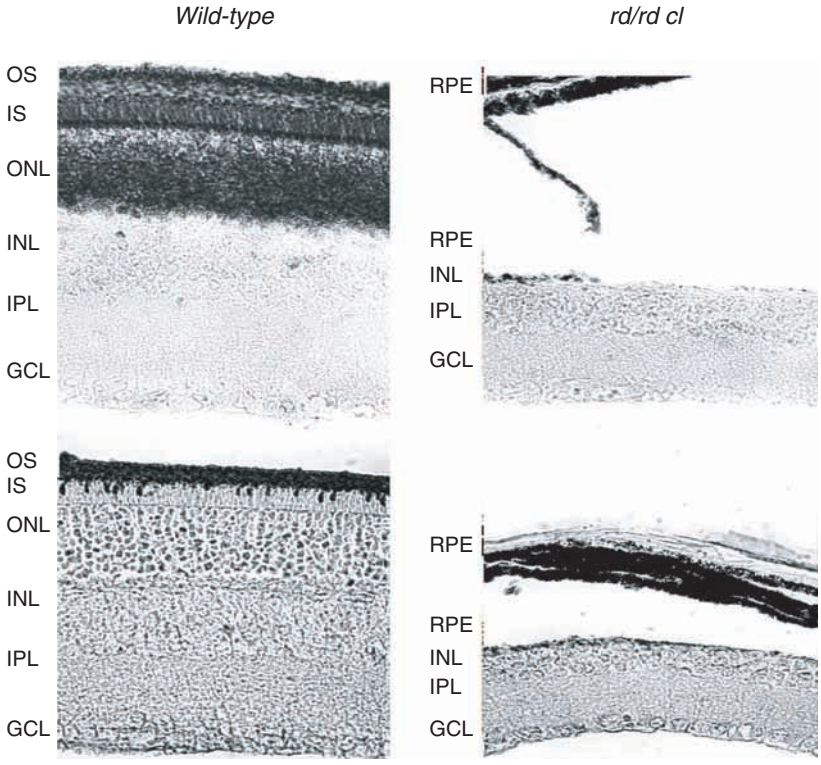

Figure 4 The *rd/rd cl* mouse provided the conclusive evidence of an additional photoreceptive system in the mammalian eye, despite lacking much of the outer retina. In the absence of rods and cones, these animals are still capable of a range on irradiance detection tasks, including phase shifting, pineal melatonin suppression, and pupillary constriction in response to light. OS, outer segment; IS, inner segment; ONL, outer nuclear layer; INL, inner nuclear layer; IPL, inner plexiform layer; GCL, ganglion cell layer; RPE, retinal pigment epithelium.

elevate cortisol, increase heart rate, and increase alertness, performance, and even mood. Whether these irradiance responses are also influenced by non-rod, non-cone ocular photoreceptors remains to be determined. Work currently underway in several laboratories around the world would suggest that this is very likely.

A Novel Photopigment

The first real insight into the nature of these novel receptors came from action spectrum studies. This powerful approach, developed by photobiologists such as the Nobel Laureate George Wald, rests upon the fact that a photopigment (light-detecting

pigment) has a characteristic absorbance spectrum or profile, which describes the likelihood of photons being absorbed at different wavelengths. As a result, a description of the spectral sensitivity profile (action spectrum) of any light-dependent response will describe the absorbance spectrum of the photopigment upon which the response is based. All known animal photopigments consist of an opsin protein, which binds a specific type of vitamin A (e.g., 11-*cis*-retinaldehyde) to form a photosensitive complex (**Figure 5**). The absorption of a photon of light by 11-*cis*-retinaldehyde converts it to an all-*trans* form. This change in shape of vitamin A alters the opsin, which in turn triggers the phototransduction cascade (**Figure 6**). This ultimately causes a change in the electrical activity of the photoreceptor cell that is transmitted by a neural pathway and ends with the perception of light. All vitamin A-based photopigments have a characteristic absorption spectrum. This means that although the peak sensitivity of the pigment may vary widely across the visible spectrum (ultraviolet at 360 nm to deep red at 750 nm), all of these pigments have the same basic shape, rather like a bell-shaped curve (**Figure 5**) – that is, they have a 'spectral fingerprint.'

The first action spectrum to be published on *rd/rd cl* mice was for pupil constriction, and the results described an opsin/vitamin A-based photopigment with peak sensitivity (λ_{max}) in the blue part of the spectrum near 480 nm (opsin photopigment; OP^{480}). The visual pigments of the mouse have λ_{max} values of 360, 508, and 498 nm for the ultraviolet-sensitive (UVS) cone, long-wavelength-sensitive (LWS) cone, and rod pigments, respectively. None of these classical photoreceptors could account for the pupillary responses to light. These studies were followed by an action spectrum in the *rd/rd cl* mouse for phase shifting. Once again, the results described an opsin/vitamin A-based photopigment with a λ_{max} very close to 480 nm. These data demonstrated that not only were rods and cone opsins not required for these irradiance detection tasks, but also that an entirely novel opsin/vitamin A-based photopigment must be present in the mouse eye.

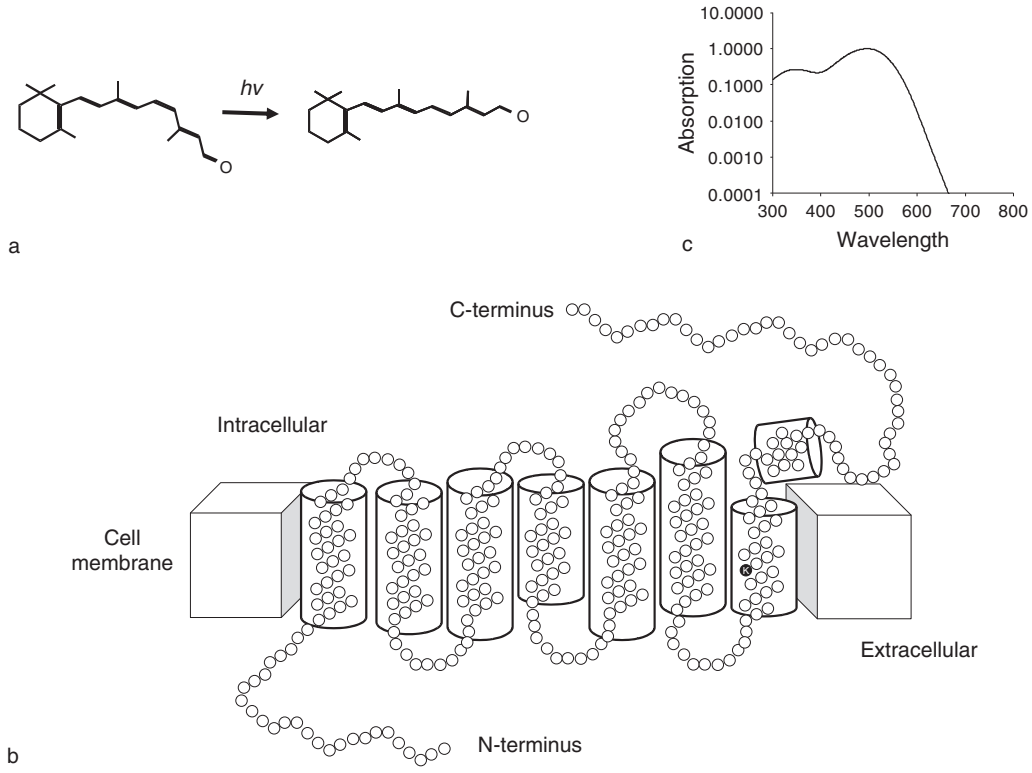

Figure 5 Vertebrate photopigments are all based upon a common plan, consisting of an opsin protein with a vitamin A-based retinal chromophore. (a) Absorption of a photon of light by the retinal chomophore results in isomerization from the 11-*cis* to the all-*trans* configuration. (b) Retinal is constitutively bound to the seven-transmembrane domain receptor opsin, via a Schiff base linkage to lysine in the seventh transmembrane domain. Following photoisomerization the opsin protein undergoes a conformational change, resulting in activation of the G-protein transducin, and initiating the phototransduction cascade. (c) All opsin/vitamin A-based photopigments have a common absorption spectrum, enabling responses mediated by this type of photopigment to be identified. Opsin schematic based upon the model of Palczewski K, Kumasaka T, Hori T, et al. (2000) Crystal structure of rhodopsin: A G protein coupled receptor. *Science* 289: 739–745.

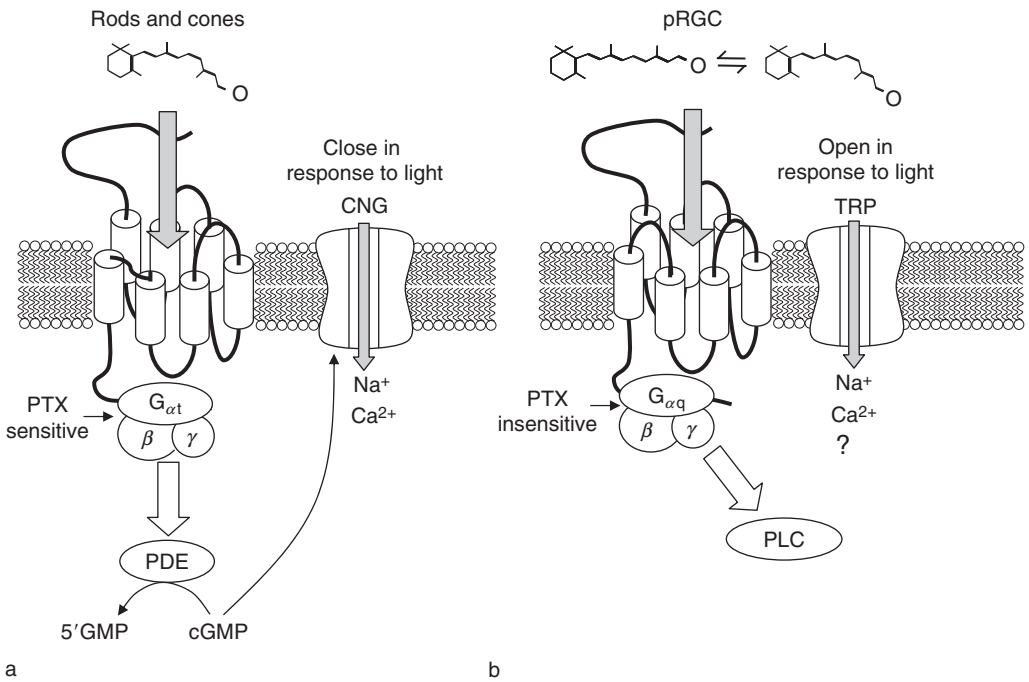

Figure 6 The process by which activation of a photopigment leads to intracellular signaling events and changes in membrane potential is termed phototransduction. (a) Hyperpolarization: Phototransduction by rods and cones has been extensively studied, and the key elements of this signaling pathway are well characterized. Following photoisomerization of retinal, the opsin protein activates the G-protein transducin (PTX, pertussis toxin). Transducin activates phosphodiesterase (PDE), which hydrolyzes cyclic guanosine monophosphate (cGMP), leading to closure of cyclic nucleotide-gated (CNG) channels and hyperpolarization of the photoreceptor cell membrane. (b) Depolarization: The phototransduction cascade used by photosensitive retinal ganglion cells (pRGCs) is poorly understood, but appears to involve a G_q-type G-protein, phospholipase C (PLC), and transient receptor potential (TRP) channels, resulting in depolarization of the pRGC.

Additional evidence for the presence of a novel blue-light-sensitive photopigment in mammals comes from three studies in humans. Two of these studies looked at the light regulation of the hormone melatonin and suggested the involvement of an opsin-based photopigment with a λ_{max} between 450 and 480 nm. The third study looked at light-driven changes in cone function in the eye and defined a pigment with a λ_{max} 483 nm. The similarity of all three human studies to each other, and correlation to mouse OP^{480}, suggest that humans possess an ortholog of mouse OP^{480}.

Photosensitive Retinal Ganglion Cells

The identification of the cells mediating the responses to light was provided by two sets of experiments. Studies on the rat used retrograde tracers injected into the SCN coupled with single-cell recordings on isolated retina in which rod and cone responses were pharmacologically blocked. Parallel studies were undertaken using calcium imaging on the *rd/rd cl* retina. Both approaches identified a population of RGCs which responded directly to light. Significantly, these photosensitive retinal ganglion cells (pRGCs) expressed an opsin called melanopsin.

The melanopsin gene family, also known as *Opn4*, was first identified in *Xenopus*, and orthologs were subsequently isolated from other vertebrate classes, including mammals. Melanopsin generated excitement almost immediately because it was shown to be expressed within RGCs, and the anatomical distribution of these RGCs was remarkably similar to that of cells projecting to the SCN and forming the RHT. More recently melanopsin-expressing RGCs have been shown to project to other nuclei in the brain associated with processing irradiance information. These include the intergeniculate leaflet (IGL), olivary pretectal nuclei (OPN), and the ventrolateral preoptic nuclei (VLPO). The essential data that melanopsin plays a critical role in the transduction of light information in pRGCs came from gene ablation studies. Melanopsin knockout mice ($Opn4^{-/-}$) exhibit attenuated phase-shifting and pupillary responses to light, as well as reduced period lengthening in constant light (LL). However, the critical involvement of melanopsin in photoreception came from triple-knockout studies of mice lacking rods, cones, and melanopsin. These animals were totally unresponsive to light, demonstrating that melanopsin is in some way essential for pRGC photosensitivity.

The possibility that melanopsin acted as a photoisomerase in these cells was eventually excluded by a series of studies in which melanopsin was functionally expressed.

The first investigation of the biochemistry of melanopsin involved expression of melanopsin in COS cells and reconstitution with 11-*cis*-retinal. The resultant pigment showed a λ_{max} between 420 and 440 nm, an absorption maximum considerably shifted away from OP^{480}. The discrepancy in λ_{max} between spectroscopy and action spectra, coupled with the low signal/noise from the absorbance spectroscopy, prompted the use of other approaches to address whether melanopsin can act as a photopigment. Quite independently, three groups combined the expression of melanopsin protein with physiological assays of cellular photosensitivity. All three studies showed that melanopsin transfection can confer photosensitivity to nonphotosensitive cell types (Neuro-2a cells, HEK293-TRPC3 cells, and *Xenopus* oocytes). In addition, these groups were able to show that specific forms of retinal (especially 11-*cis*-retinal) are needed for these responses to light, that light will ultimately trigger the release of intracellular calcium, and that this will involve some form of G-protein interaction with melanopsin. Expression studies on human melanopsin suggest that the λ_{max} of light responses is close to 420–430 nm, and in this regard the findings were similar to those obtained by Newman and colleagues. The studies on murine melanopsin, however, showed an action spectrum for light responses that exhibited a λ_{max} very close to 480 nm. The current consensus from the various groups is that something about the local environment in which melanopsin is reconstituted is important in determining its λ_{max}.

While the phototransduction cascade used by rods and cones has been characterized extensively, the mechanisms used by pRGCs remain poorly understood (**Figure 6**). Unlike rods and cones, pRGCs depolarize in response to light, and in this manner more closely resemble invertebrate photoreceptors. Pharmacological studies on pRGCs have suggested the involvement of a transient receptor potential (TRP)-like channel in these responses. Additionally, pharmacological studies on cells transfected with melanopsin suggest that their light responses are insensitive to pertussis toxin, a blocker of G_i/G_o-type G-proteins, including transducin (the G-protein of the rods and cones). However, agents which inhibit G_q/G_{11}-type G-proteins do prevent these responses. Further downstream, drugs which block phospholipase C (PLC) have been shown to block these responses. These findings suggest a phototransduction cascade more akin to that found in invertebrates. However, these findings are based upon signal transduction pathways which are activated in nonnative cell types, and whether similar processes occur in pRGCs remains to be determined.

pRGC/Rod/Cone Interactions

The pRGCs provide measures of environmental irradiance in mammals. But this does not mean that the rods and cones play no role in these light-detecting tasks. Indeed, several lines of evidence suggest that both classical and novel photoreceptors interact in the detection of brightness. This is clear in the case of the PLR, where the loss of the rods and cones leads to a marked decrease in the sensitivity of the response. In this case, we have suggested that the rods and cones allow a rapid assessment of light transitions and cause an immediate pupillary constriction or dilation. By contrast, the pRGCs are likely to provide sustained responses to prolonged periods of environmental brightness. In the case of the circadian system, the role of the rods and cones is less clear. The loss of these receptors causes only subtle effects on circadian entrainment. By contrast, loss of the melanopsin gene alone merely attenuates circadian responses to light. It is only when all the rods, cones, and melanopsin are genetically ablated that circadian responses to light are completely abolished. As a result, there seems to be a complex and semiredundant interaction between the classical and novel photoreceptor systems of the murine eye in the regulation of the biological clock. Very recent studies in primates has shown that some form of spectral opponency appears to occur in pRGC output, with excitation produced by cones sensitive to medium and long wavelengths, and inhibition from short-wavelength-sensitive cones.

Perhaps this overlap in function should be of no great surprise. There are many good examples wherein multiple photopigments have been shown to regulate circadian physiology in nonmammalian vertebrates, invertebrates, and plants. Why this should be and how these photopigments might interact remain a mystery, but must surely relate to the difficult task of extracting time-of-day information from dawn and dusk. During twilight, the quality of light changes in three important respects: (1) the amount of light (irradiance), (2) the spectral composition of light, and (3) the source of light (i.e., the position of the sun). These parameters of the light environment all change systematically and, in theory, could be used by the circadian system to detect the phase of twilight. However, each of these features is subject to subtle variations or 'noise,' the type and extent of which will depend upon both the species (e.g., sedentary or mobile) and the environment. A general feature of sensory systems is the need to extract an appropriate

signal from background noise, and much of the complexity observed in sensory systems in general is associated with noise reduction. One way to do this is to have different receptors, with slightly different response ranges, sampling the same scene. Perhaps the circadian system utilizes the pRGCs to provide measures of brightness, while it utilizes the rods and cones to detect the direction and color of environmental light. This information would be integrated to provide an assessment of the time of dawn and dusk based upon several features of the changing quantity and quality of light at twilight.

Future Perspectives

Our understanding of the photoreceptors involved in the regulation of circadian physiology has undergone a remarkable transformation in the past decade. The discovery of a previously unrecognized photoreceptor system in the mammalian eye has provided a greater understanding of how light modulates not only the circadian system, but also a whole range of non-image-forming responses to light. As with any new discovery, numerous additional questions remain unanswered. Future studies are required to address the relative contribution of rods, cones, and pRGCs to different irradiance detection tasks; the phototransduction mechanisms used by the pRGCs; whether these cells contribute to our visual perception of the environment; how the signals from the retina are integrated by the SCN to bring about photoentrainment of the biological clock; and the role of classical photoreceptors and pRGCs in the modulation of sleep. Answering these questions will provide a greater understanding of how numerous aspects of our physiology are regulated by the light environment, and may provide future therapeutic strategies for dealing with circadian disturbances associated with jet lag, shift work, sleep disorders, and a range of mental health problems.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Organization in Non-Mammalian Vertebrates; Circadian Rhythms: Influence of Light in Humans; Entrainment of Circadian Rhythms by Light; Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Ganglion Cells: Anatomy; Sleep: Development and Circadian Control.

Further Reading

Berson D, Dunn F, and Takao M (2001) Phototransduction by ganglion cells innervating the circadian pacemaker. *Investigative Ophthalmology and Visual Science* 42: S113.

Buijs RM and Kalsbeek A (2001) Hypothalamic integration of central and peripheral clocks. *Nature Reviews Neuroscience* 2: 521–526.

Foster RG (2002) Keeping an eye on the time: The Cogan lecture. *Investigative Ophthalmology & Visual Science* 43: 1286–1298.

Foster RG and Hankins MW (2002) Non-rod, non-cone photoreception in the vertebrates. *Progress in Retinal and Eye Research* 21: 507–527.

Freedman MS, Lucas RJ, Soni B, et al. (1999) Regulation of mammalian circadian behavior by non-rod, non-cone, ocular photoreceptors. *Science* 284: 502–504.

Hankins MW and Lucas RJ (2002) The primary visual pathway in humans is regulated according to long-term light exposure through the action of a nonclassical photopigment. *Currents in Biology* 12: 191–198.

Hattar S, Liao HW, Takao M, et al. (2002) Melanopsin-containing retinal ganglion cells: Architecture, projections, and intrinsic photosensitivity. *Science* 295: 1065–1070.

Hattar S, Lucas RJ, Mrosovsky N, et al. (2003) Melanopsin and rod–cone photoreceptive systems account for all major accessory visual functions in mice. *Nature* 424: 75–81.

Lucas RJ, Douglas RH, and Foster RG (2001) Characterization of an ocular photopigment capable of driving pupillary constriction in mice. *Nature Neuroscience* 4: 621–626.

Lucas RJ, Freedman MS, Munoz M, et al. (1999) Regulation of the mammalian pineal by non-rod, non-cone, ocular photoreceptors. *Science* 284: 505–507.

Peirson S and Foster RG (2006) Melanopsin: Another way of signaling light. *Neuron* 49: 331–339.

Provencio I, Rodriguez IR, Jiang G, et al. (2000) A novel human opsin in the inner retina. *Journal of Neuroscience* 20: 600–605.

Reppert SM and Weaver DR (2002) Coordination of circadian timing in mammals. *Nature* 418: 935–941.

Rodieck RW (1998) *The First Steps in Seeing*. Sunderland, MA: Sinauer.

Sekaran S, Foster RG, Lucas RJ, et al. (2003) Calcium imaging reveals a network of intrinsically light-sensitive inner-retinal neurons. *Current Biology* 13: 1290–1298.

Photoreceptors: Physiology

D-G Luo, C-Y Su, and K-W Yau, Johns Hopkins University School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Photoreceptors are cells that can detect light by virtue of a light-absorbing pigment. In the nervous system, photoreceptors serve to detect light for either image-forming or non-image-forming visual functions. Image-forming vision underlies the detection and recognition of objects and their motion. Non-image-forming vision, on the other hand, simply informs the organism about the intensity and spectral composition of the light in the ambience, often for tracking the time of day or seasonal changes.

Types of Photoreceptors

In the animal kingdom, the great majority of neural photoreceptors are readily identifiable as being either ciliary or microvillous; that is, their photosensitive structure is derived from either modified cilia or modified microvilli. In either case, the photoreceptors are often clustered in special organs (eyes), although they can also be present in small numbers in less well-defined regions of the nervous system. Occasionally, the photoreceptors may not bear any obvious morphological signs of being ciliary or microvillous.

In vertebrates, the most prominent photoreceptors are the rods and cones in the retina of the eyes (also called lateral eyes). Indeed, it was long thought (for well over 100 years) that rods and cones are the only light-sensitive neurons in the retina. However, work in the past several years has now shown that some other retinal neurons are also intrinsically photosensitive. In mammals, these consist of a small subset (about 1%) of the retinal ganglion cells (the axons of which form the optic nerve and project to the brain). Besides the two lateral eyes, the pineal gland of some vertebrates, notably birds and fish, is also light sensitive. In mammals, the pineal gland is not light sensitive, but at least in rodents it has been reported that this organ is light sensitive in neonates. Certain lizards have a structure called the parietal ('third') eye located on the forehead between the two lateral eyes (see the section titled 'Parietal-eye photoreceptors in lizard'). Some amphibians have a comparable structure called the frontal organ. Rods, cones, light-sensitive pinealocytes, and parietal-eye photoreceptors are all of the ciliary type. On the other hand, the classification of the intrinsically photosensitive retinal ganglion cells, which show no detectable ciliary or microvillous morphology, is less clear. Deep-brain photoreceptors also exist in birds, amphibians, and fishes, but not much is known about them.

In invertebrates, the great majority of photoreceptors so far studied are of the microvillous type. These include the photoreceptors in the insect compound eyes and the eyes of squid, octopus, and various mollusks. These microvillous photoreceptors are often called rhabdomeric photoreceptors because their photosensitive microvilli form a 'rhabdom.' Ciliary photoreceptors do exist in invertebrates, with one well-known example being the hyperpolarizing photoreceptor in scallop. There are also single, isolated photoreceptors in the molluskan neural ganglia that do not show any obvious ciliary or microvillous morphology, but at least some of these resemble ciliary photoreceptors in terms of the molecular components they use for phototransduction.

Retinal Rods and Cones

The rod photoreceptors across all vertebrate species are essentially the same structurally and functionally; similar is the case with cones. Rods are very sensitive to light, being able to detect and signal the absorption of a single photon. They are responsible for vision in dim light (scotopic conditions). Cones are much less sensitive to light than are rods (by about 25–100 times, or more in certain vertebrate species), and they are capable of adapting to background light much more effectively. Cones are responsible for vision in brighter light (photopic conditions). The eyes of most vertebrates are rod dominant (i.e., having a lot more rods than cones in the retina), but a few are cone dominant (e.g., those of ground squirrel and chicken). In human, although the retina has mostly rods (95% or more of the photoreceptors are rods), the center of the fovea, which is the part of the retina that mediates acute vision, has exclusively cones. Accordingly, we rely on cones for our daily vision under practically all light conditions, except for the lowest two log units of detectable light intensities.

Humans have three types of cones: red or long-wavelength-sensitive cones, green or medium-wavelength-sensitive cones, and blue or short-wavelength-sensitive cones, each with a different visual pigment. The presence of more than one type of cone with distinct spectral sensitivities is what allows us to discriminate between

monochromatic light at different wavelengths, which are perceived by us as having different colors. Each of these colors is essentially a distinct sensation evoked by a unique ratio of activation of the three cone types due to a given monochromatic light. For example, red light activates mostly red-sensitive cones, less so the green-sensitive cones, and still less so the blue-sensitive cones. Green light activates best green-sensitive cones, less so the red-sensitive cones, and still less so the blue-sensitive cones, etc. If one represents the relative excitations of the three cone types as the three axes of a three-dimensional coordinate system, the direction of a vector in this three-dimensional space then corresponds to a particular color sensation, and the length of the vector represents the brightness. In other words, color and brightness are distinct attributes of the light, with no confusion between them. The same holds when mixtures of two or three monochromatic lights are viewed; namely, each of these mixtures evokes a unique color sensation that cannot be reproduced by any other mixture of two or three monochromatic lights, regardless of their intensities. However, the situation breaks down when four or more monochromatic lights are present in a mixture. There are many mixtures of four or more monochromatic lights (with appropriate relative intensities) that will all appear to us as being identical in color. The reason for this degeneracy is that our retina has only three types of cones. If we had more types of cones, we would then be able to discriminate between light mixtures made up of more than three monochromatic components. In general, it can be shown by simple algebra that, if there are n cone types in the retina, the visual system should be able to discriminate between any two light mixtures having up to n monochromatic components (i.e., perceived by the system as having different colors). Why don't we, then, have more than three types of cones? The answer is that our visual system has to process separately many different attributes of the visual image, including bright/dark contrast, motion, acuity, etc., through complex circuitry, and color is only one of these attributes. Thus, there is inevitable compromise as to how much elaboration the system can devote to any particular visual attribute.

In contrast to cones, we have only one type of rod, all containing the same pigment, rhodopsin, which is most sensitive to blue–green light. Thus, we have no color vision under scotopic conditions, when the cones are inactive. In the vectorial representation described earlier, this situation is equivalent to saying that, with only one type of rod, the 'color perception' is reduced to one dimension, in which all vectors collapse to the same direction (hence no distinct colors), and light intensity (brightness) becomes the only variable, regardless of the complexity of the light mixture. In mesopic conditions (i.e., at light levels where the transition between rod-mediated and cone-mediated vision takes place), rods may also partake in color vision.

The number of types of cones and rods is not uniform across vertebrate species. Some lower species have more than one type of rod (e.g., amphibians) and as many as four types of cones (e.g., fishes). Other species have only two types of cones, including the New World monkeys (with only red-sensitive and blue-sensitive cones). Old World monkeys, like us, have also green cones. The green-pigment gene in us originated from duplication of the red-sensitive pigment gene (the two genes are immediately adjacent to each other). Genetic polymorphisms in the amino acid sequences of the human pigments lead to slight variations in their spectral sensitivity and therefore slight differences in color perception among individuals. The loss of one or more cone pigment genes also leads to various degrees of color blindness (i.e., a lesser ability to discriminate between different light mixtures).

Morphology of Rods and Cones

Morphologically, each rod or cone is made up of the outer segment, the inner segment, the cell body, and the synaptic terminal (**Figure 1**). The outer segment contains the visual pigment, and is where light is absorbed and transduced into an electrical signal. Developmentally, it is a modified cilium that has become highly expanded and folded, giving rise to hundreds of membrane disks tightly and orderly stacked one on top of another throughout the outer segment. In rods, almost all of the disks (except for a small number at the base) have become internalized, that is, their membranes are no longer continuous with the plasma membrane. In cones, the disk membranes remain continuous with the plasma membrane. The membranes of the disks contain the visual pigment, packed to 10^7–10^9 molecules per cell. Associated with the disk membranes are also several phototransduction proteins. The rest of the transduction proteins are either cytosolic or situated in/on the plasma membrane of the outer segment. The inner segment houses the metabolic machinery. The part of inner segment adjacent to the ciliary stalk (called the ellipsoid) is chock-full of mitochondria for handling the heavy energy demand of the outer segment. At the proximal end of the cell are the cell body and the synaptic terminal. Glutamate is the neurotransmitter.

Overlying the outer segments of rods and cones are the pigment epithelial cells, which contain black pigment granules. These cells serve multiple functions.

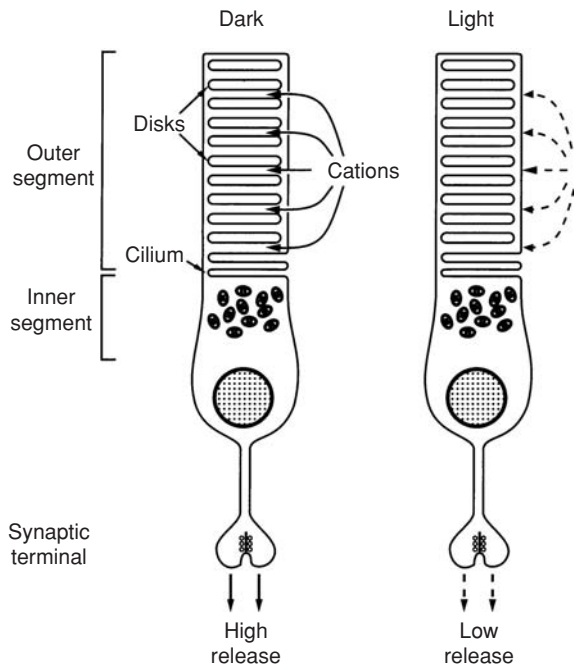

Figure 1 Morphology of rod photoreceptors. The phototransduction machinery is housed in the outer segment. The figure shows the basic mechanism underlying the light response. In darkness, a circulating current between the outer segment and the rest of the cell (with an inward direction across the outer-segment membrane) keeps the rod depolarized and maintains a high glutamate release at the synaptic terminal. In the light, the circulating current decreases, and the glutamate release is reduced. Reproduced from Yau K-W (1994) Phototransduction mechanism in retinal rods and cones. The Friedenwald Lecture. *Investigative Ophthalmology and Visual Science* 35: 10–32.

One function is to absorb any light not absorbed by the rods and cones at first pass, so as to prevent degradation of the visual image due to backscattering. Another is to mediate the transfer of metabolites, ions, water, etc., between the photoreceptors and the capillaries lying behind the pigment epithelium. A third function is to regenerate the pigment after it is bleached by light (see the section titled 'Phototransduction mechanism in rods and cones'). Finally, the outer segment of both rods and cones continue to elongate on a daily basis through the addition of new disks at the base. The overall length of the outer segment is nonetheless kept constant by the coordinated shedding of distal disks, in conjunction with the phagocytosis of these shedded disks by the pigment epithelial cells. For mouse rods, it takes about 10 days for a disk to travel from the base to the tip of the outer segment. This renewal process may be important for removing any cumulative, deleterious effect of photooxidation in rods and cones. In some nocturnal animals, such as cats and dogs, part of the pigment epithelium is less pigmented, and reflective cells lie behind the epithelium to form the tapetum. This reflective layer serves to back-reflect any light unabsorbed by the rods and cones at first pass, so as to give the light a second chance to be absorbed. The presence of a tapetum explains why the eyes of nocturnal animals glow when illuminated in the dark. The tapetum increases light sensitivity in darkness, but does degrade the visual image somewhat.

Light Response of Rods and Cones

Rods and cones have a membrane potential of about –30 to 35 mV in darkness, and they respond to light with a membrane hyperpolarization, unlike most other sensory receptor cells, which depolarize in response to stimulation. They do not have an action-potential mechanism. Because of their relatively depolarized state in darkness, they continuously release glutamate from their synaptic terminals onto bipolar cells and horizontal cells, the second-order visual neurons in the retina. The light-triggered membrane hyperpolarization reduces the glutamate release. The brighter the light, the larger is the membrane hyperpolarization in rods and cones, and therefore the greater is the reduction in glutamate release. The bipolar cell, which is the second-order throughput neuron, can respond to this decrease in glutamate release with either a hyperpolarization (OFF-bipolar cell) or a depolarization (ON-bipolar cell), depending on the nature of the postsynaptic glutamate receptor it possesses.

The rod and cone responses are relatively slow (with the cone response being the faster of the two), reflecting a second-messenger-mediated transduction mechanism (see the section titled 'Phototransduction mechanism in rods and cones'). For example, the rod response to a dim, brief flash lasts for over a second for cold-blooded animals at room temperature, though faster at the higher body temperature of mammals. The speeds of the rod and cone responses dictate how rapidly the visual system can detect image changes on the retina (temporal resolution), and therefore motion. Because cones have faster responses than do rods, they are better motion detectors. On the other hand, the slower response of rods allows these cells to sum signals slightly spread over time (temporal integration), thus enhancing their sensitivity in dim light, especially useful for static images.

Figure 2 shows the responses of an amphibian rod to a light flash at different intensities, recorded simultaneously with an intracellular voltage electrode and with a suction pipette that records membrane current at the outer segment. The suction pipette shows that an inward current flows in darkness across the outer-segment membrane (which accounts for the not-so-negative resting potential in darkness), and that light decreases this 'dark current' (which accounts for the

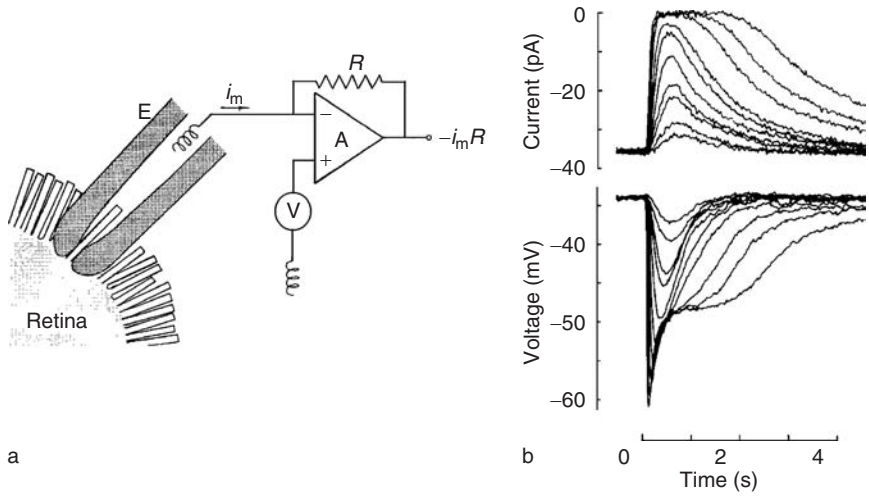

Figure 2 Electrical recordings from rod photoreceptors. (a) Diagram illustrating the suction-pipette recording method from a single rod outer segment protruding from a piece of retina. The circuitry follows the same design as patch clamp recording. (b) Response families recorded from a salamander rod simultaneously with a suction pipette (upper panel) and a sharp intracellular electrode (lower panel). Light flashes of increasing intensity (by about factor of 2) were delivered at time zero. The largest responses are saturated, resulting from a complete stoppage of the dark current. The small responses represent averages from multiple flash trials. In the upper panel, negative current means inward current at the outer segment. (a) Reproduced from Yau K-W, Lamb TD, and Baylor DA (1977) Light-induced fluctuations in membrane current of single toad rod outer segments. *Nature* 269: 78–80. (b) Reproduced from Baylor DA, Matthews G, and Nunn BJ (1984) Location and function of voltage-sensitive conductances in retinal rods of the salamander, *Ambystoma tigrinum. Journal of Physiology* 354: 203–223.

membrane hyperpolarization) until its complete suppression by bright light (see also Figure 1). With a bright flash, the voltage response shows a fast re-depolarization from the peak hyperpolarization before settling to a plateau level. This transient 'nose' (especially prominent in rods) is not observed in the response of the membrane current at the outer segment. The nose reflects the effect of a hyperpolarization-activated cation current (I_h) carried by both Na^+ and K^+, with a reversal potential close to 0 mV. This I_h current is situated at the inner segment and cell body, which explains why it is not detected by the suction-pipette recording. The suction-pipette recording method therefore monitors the phototransduction event in the outer segment more faithfully, without much contamination by the signal from other parts of the cell. This recording method has been tremendously useful in examining the details of phototransduction.

With increasing flash intensity, the response increases progressively until saturation is reached. When the transient peak amplitude of the flash response is plotted against the corresponding flash intensity (intensity–response relation), a relation fairly close to the Michaelis relation in enzyme kinetics is found, especially with cold-blooded animals, namely, $R = R_{max}[I_F/(I_F + \sigma)]$, where R is the transient peak amplitude of the response, R_{max} is the saturated response amplitude, I_F is the flash intensity, and σ is the half-saturating flash intensity. While simple, this relation does not really convey the mechanism underlying the generation of the light response. In other words, the relation is purely empirical. For one thing, with increasing flash intensity, the transient peak of the response moves earlier in time (i.e., the time between flash and transient peak of the response, or time-to-peak, gets shorter), due to the decay phase of the response cutting into the rising phase earlier. This is a sign of active adaptation by the cell to progressively brighter illumination. If a long step of light is used for stimulation instead, the response rises to a transient peak and then relaxes to a lower plateau level, with the response relaxation again reflecting adaptation of the cell to steady light. If the flash–response amplitude in Figure 2 is measured instead at a 'fixed' early time in the rising phase of the response, before any light adaptation sets in, the flash intensity–response relation is described instead by the relation $R = R_{max}[1 - \exp(-I_F/\rho)]$, where the half-saturating flash intensity is given by $\rho(\log_e 2)$. This relation bears a simple mechanistic interpretation, namely that each absorbed photon activates a spatially restricted region of the outer segment, within which the transduction effect essentially reaches saturation (i.e., all dark current in this region is suppressed).

The rod and cone photoresponses are detectable as the a-wave in the electroretinogram (ERG), which is a mass recording from the retina with an electrode placed near the cornea of the intact eye. However,

the a-wave does not provide the full time course of the photoreceptor light response because the b-wave (which largely originates from the activity of bipolar cells, especially ON-bipolar cells) of the ERG, which immediately follows the a-wave, obscures the decline phase of the a-wave. The ERG nonetheless provides a very useful clinical tool for assaying rod and cone function in human patients.

Single-Photon Response

It was mentioned earlier that rods are so sensitive to light that they can signal the absorption of a single photon. The response triggered by an absorbed photon is indeed quite large, with an amplitude close to 1 pA, and is directly observable with suction-pipette recording. In amphibian rods, the single-photon response amplitude corresponds to about 3% of the maximum (or saturated) light response. Thus, as few as about 30 absorbed photons will produce a response that is half-maximum. This large signal reflects the high amplification of the signal transduction mechanism underlying phototransduction in rods (see the section titled 'Phototransduction mechanism in rods and cones'). The single-photon response is quite constant in amplitude and highly stereotyped in kinetics: these properties make rods a very faithful photon counter.

In cones, the single-photon response is considerably (25–100 times) smaller, reflecting quantitative (although not qualitative) differences from rods with respect to the underlying phototransduction mechanism. Most of the difference in sensitivity between dark-adapted rods and cones in fact comes from this factor, rather than from the ability to capture light, which is broadly comparable in rods and cones.

Pigment Noise

A visual pigment molecule is composed of a protein moiety (opsin) and a chromophore (the light-absorbing component) covalently linked to the opsin (to be described more in detail in the following sections). Light acts by photoisomerizing the chromophore, thereby activating the pigment. It takes substantial light energy to isomerize the pigment, so the latter is very stable in darkness. Nonetheless, isomerization can, and does, occur spontaneously in darkness due to the thermal energy associated with the molecule, giving rise to false signals about light in darkness. In rods, the rate of spontaneous isomerization is about one event per minute per cell. Considering the large number of pigment molecules present in a cell (see the section titled 'Morphology of rods and cones'), the corresponding rate constant is very low, being about 10^{-10}–10^{-11} s^{-1} at room temperature (and about tenfold higher at 37 °C). This gives a half-life of rhodopsin in the neighborhood of hundreds of years! Thus, rhodopsin can be considered to be very stable, and is ideal for the task of detecting dim light. Cone pigments have a much higher tendency to isomerize thermally, especially the long-wavelength-sensitive (red) pigment. This is not as critical as in rods from the point of view of signal-to-noise ratio, because cones operate in much brighter light. On the other hand, the higher rate of thermal isomerization does mean that there is a higher level of 'dark light' (i.e., equivalent light in darkness). This dark light will trigger light adaptation in cones even in the absence of real light, causing the cones to have a lower sensitivity than they would have if the pigment were stable. In other words, part of the sensitivity difference between rods and cones in darkness may be due to the noisy cone pigment.

Another interesting difference between rod and cone pigments is that rod pigment, once formed from opsin and the chromophore, does not re-dissociate until isomerization followed by bleaching takes place (see the following sections). Cone pigment, on the other hand, has some tendency to re-dissociate into opsin and chromophore without isomerization. Thus, even in darkness, some of the opsin in cones lack chromophore and therefore incapable of signaling light (though this, again, may not matter very much, considering that cones function in bright light). Also, it turns out that opsin itself has a weak constitutive ability to activate the transduction process. This bare-opsin activity can also cause some adaptation by cones in darkness, lowering their sensitivity further. Is there any usefulness in the property of dissociation of cone pigments? Probably not much, in the context of signaling. However, cone pigments need to recycle rapidly (see the section titled 'Phototransduction mechanism in rods and cones') because they operate in bright-light conditions. This rapid recycling requires rapid dissociation of the photoisomerized chromophore from the opsin, then entering the retinal pigment epithelial cell (see the section titled 'Morphology of rods and cones') to be regenerated (reisomerized) to the dark state before returning and recombining with the bare opsin to form nascent holopigment. The rapid dissociation of the photosiomerized chromophore from cone opsin presumably requires a loose (or relatively open) chromophore-binding pocket on cone opsin, which likely impacts the stability of the dark pigment as well, hence the dark dissociation. In other words, the molecular design to achieve rapid regeneration of the bleached pigment may come with a price for the behavior of the dark pigment.

Incidentally, the same weak, constitutive activity of bare opsin underlies the phenomenon of 'bleaching adaptation.' It has been known for a long time that, if bleached pigment is not regenerated, the sensitivity of a rod or cone is suppressed. However, this suppression of sensitivity is far more severe than expected simply from the reduction in photon-capture capability due to the presence of bare opsin (the probability of photon capture varies inversely with the percentage of bare opsin). Indeed, the suppressive effect of bare opsin on sensitivity comes mostly from its constitutive activity to trigger the transduction process (and therefore adaptation) rather than the proportional decrease in holopigment.

The cGMP-Gated Channel

The inward dark current of rods and cones that is suppressed by light goes through a cyclic guanosine monophosphate (cGMP)-gated cation channel on the plasma membrane of the outer segment. This channel is unusual by being directly opened (or 'gated') by a cyclic nucleotide, cGMP in this case. It is a nonselective cation channel, passing not only Na^+ and K^+, but also Ca^{2+}. About 75% of the dark current is carried by Na^+ and the rest is largely carried by Ca^{2+}. In steady state in darkness, the Na^+ influx is balanced by an efflux via the Na^+ pump in the inner segment, and the Ca^+ influx by an equal efflux via a $Na^+/Ca^{2+}/K^+$ exchange mechanism in the outer segment. This $Na^+/Ca^{2+}/K^+$ exchange likewise depends on the maintenance of an inward-directing Na^+ electrochemical gradient by the Na^+ pump. This explains the heavy metabolic load on the rods and cones mentioned at the beginning of this article, and the heavy concentration of mitochondria next to the outer segment. Indeed, the retina has one of the highest, if not the highest, oxygen consumption in the body.

As a ligand-gated channel, the cGMP-gated channel is also unusual in that it shows no desensitization to its ligand, cGMP. While unusual, this feature is functionally important, because it allows the channel to stay open in darkness in the presence of a steady concentration of cGMP in the cytoplasm of the outer segment, and to be closed only by light via a reduction in cGMP (see next section). Rods and cones have molecularly distinct cGMP-gated channels, though these are functionally quite similar. These rod and cone channels are members of a family of cyclic nucleotide-gated (CNG) channels, which also include the channel that mediates olfactory transduction in olfactory receptor neurons in the nose.

One interesting property of the rod and cone cGMP-gated channels (and CNG channels in general) is that, apart from being permeable to divalent cations besides monvalent cations, they are also subject to constant, partial blockage by external divalent cations (Ca^{2+} and Mg^{2+}) under physiological conditions. One consequence of this (kinetically fast) blockage is that the channels have a very small effective single-channel conductance (after low-pass filtering by the membrane time constant). This small single-channel conductance means that the quantization noise associated with the opening and closing of channels is reduced, effectively increasing the signal-to-noise ratio of the light signal.

Phototransduction Mechanism in Rods and Cones

Phototransduction is the process by which the absorbed light triggers an electrical response (the neural signal) in rods and cones. As mentioned in the previous section, this process involves the closure by light of a cGMP-gated, nonselective cation channel. The mechanism by which light closes the channel is now very well understood (**Figure 3**).

As mentioned earlier, visual pigments are made up of opsin, the protein moiety, and a covalently linked chromophore, which is 11-*cis*-retinaldehyde (or simply 11-*cis*-retinal), a derivative of vitamin A. Essentially all land-based animals have the same chromophore, but each rod and cone opsin is a different protein, which tunes the absorption spectrum of the chromophore. Photoisomerization involves a conversion of 11-*cis*-retinal into all-*trans*-retinal. This change unleashes spontaneous conformational changes in the opsin, with the active intermediate, Meta II, being formed within about 1 ms. Meta II then activates transducin, a heterotrimeric G-protein (G_T, composed of α, β, and γ subunits), by prompting the substitution of guanosine triphosphate (GTP) for GDP on the α-subunit, $G_{T\alpha}$; the dissociation of $G_{T\alpha}\cdot$GTP from $G_{T\beta\gamma}\cdot G_{T\alpha}\cdot$GTP in turn stimulates a cGMP-phosphodiesterase (cGMP-PDE) to increase the hydrolysis of cGMP. This enzyme is composed of two catalytic subunits (PDE$\alpha\beta$ in rods or two copies of PDEαs in cones) and two inhibitory subunits (two copies of PDEγs in rods and in cones). $G_{T\alpha}\cdot$GTP acts by binding to an inhibitory subunit and removing its inhibitory influence on a catalytic subunit. The resulting increase in PDE activity lowers the cytoplasmic level of cGMP, and the CNG channels therefore close. This closure of the cation channels produces a membrane hyperpolarization as the light response, and reduces the glutamate release from the synaptic terminal of the cell. The high amplification of phototransduction in rods (described earlier), which allows these cells to signal single absorbed photons, arises in part from the fact that each rhodopsin molecule has a long active

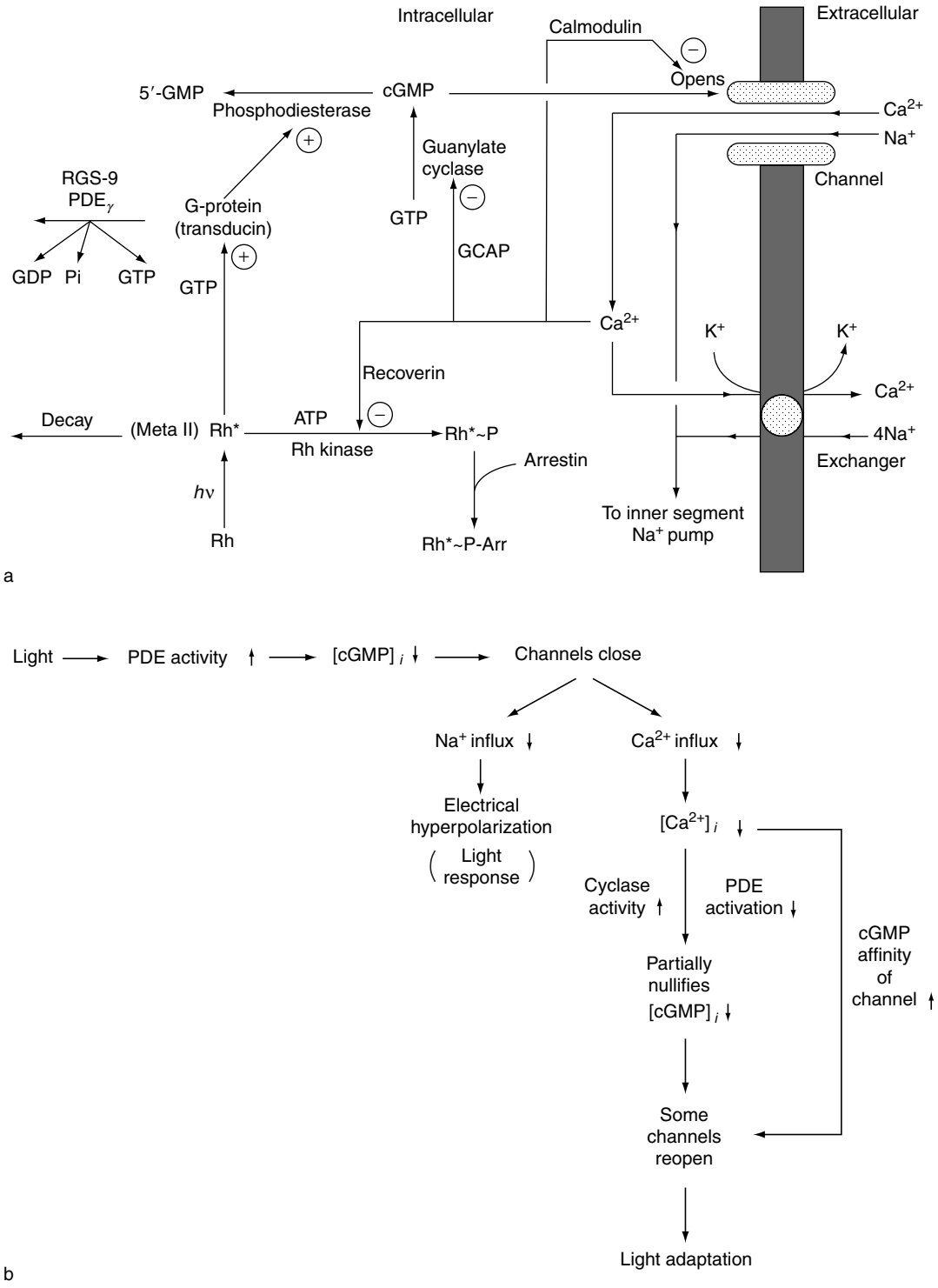

Figure 3 Phototransduction mechanism. (a) Schematic diagram of phototransduction in rods. The same applies to cones. *hν*, photon; Rh, rhodopsin; Rh*, photoactivated rhodopsin (Metarhodopsin II); Rh*∼P, phosphorylated form of rhodopsin; RGS-9, regulator of G-protein-signaling isoform 9; PGE_{γ}, inhibitory (γ) subunit of the phosphodiesterase; GCAP, guanylate-cyclase-activating protein; +, stimulation or positive modulation; –, inhibition or negative modulation. (b) Flow chart showing the cascade of events leading from light absorption to the electrical response and also light adaptation. The same applies to cones. (a) Modified from Koutalos Y and Yau K-W (1996) Regulation of light sensitivity in retinal rods. *Trends in Neuroscience* 19: 73–81. (b) Reproduced from Yau K-W (1994) Phototransduction mechanism in retinal rods and cones. The Friedenwald Lecture. *Investigative Ophthalmology Visual Science* 35: 10–32.

lifetime in the Meta II state before inactivation sets in (see following paragraph), so that it is able to activate successively a substantial number of transducin molecules.

The termination of phototransduction involves the phosphorylation of Meta II by a specific kinase (pigment kinase, a member of the G-protein-coupled receptor kinase, or GRK, family), followed by the binding of a protein called arrestin to cap the activity of Meta II. At the same time, $G_{T\alpha}$·GTP self-deactivates by its intrinsic GTPase activity, which hydrolyzes the bound GTP to GDP and consequently returns to its inactive $G_{T\alpha}$·GDP state to be ready for activation again after reassociating with $G_{T\beta\gamma}$. The GTPase activity of $G_{T\alpha}$·GTP is enhanced by PDEγ, its substrate, together with a GTPase-activating protein (GAP) complex, which consists of RGS9 (regulator of G-protein-signaling isoform 9) and $G_{\beta 5}$, another G-protein β-subunit or a G-protein β-subunit-like protein. When transducin deactivates, so does PDE by restoration of the inhibitory activity of its PDEγ. The free cGMP concentration then returns to the dark level because of ongoing activity of the cGMP-synthesizing enzyme, guanylate cyclase (GC). In short, there is steady cGMP synthesis and hydrolysis in darkness, and a light flash simply transiently tips the balance toward hydrolysis. Meta II eventually decays to an intrinsically inactive state. The pigment then dissociates into opsin and free all-*trans*-retinal, a process called 'bleaching.' The opsin also loses its bound arrestin and is dephosphorylated by a generic kinase (protein phosphatase 2A). As for the all-*trans*-retinal, it is reduced to all-*trans*-retinol ('-ol' stands for 'alcohol'). All-*trans*-retinol diffuses, perhaps helped by a carrier protein, from the photoreceptor into the pigment epithelial cell, where it is converted into all-*trans*-retinylester, then reisomerized (by the isomerohydrolase enzyme RPE65) to 11-*cis*-retinol and oxidized to 11-*cis*-retinal, before returning to the photoreceptor and combining spontaneously with opsin to re-form the holopigment. Recently, it has become known that Muller cells also possess a self-contained chromophore-regenerating mechanism, apparently specifically for the cones. This additional regeneration mechanism may be important in bright light, when cone pigment has to compete with rhodopsin for 11-*cis*-retinal but may not win out because the nondissociable rhodopsin will out-compete the dissociable cone pigment by acting as a huge sink for 11-*cis*-retinal.

Ca^{2+} has a critical role in active adaptation of the rod or cone to light, as mentioned earlier. Because the CNG channels are permeable to Ca^{2+}, there is a steady Ca^{2+} influx in darkness, balanced by a steady Ca^{2+} efflux via a $Na^+/Ca^{2+}/K^+$ exchanger, an antiporter that uses the standing inward-directed Na^+ electrochemical gradient and outward-directed K^+ electrochemical gradient to extrude Ca^{2+} against its inward-directed gradient. Like the CNG channel, this exchanger is present on the plasma membrane of the outer segment. In the light, the closure of CNG channels stops the Ca^{2+} influx, but the Ca^{2+} efflux through the exchanger continues, resulting in a decrease in the free Ca^{2+} concentration in the outer segment. This Ca^{2+} decrease triggers multiple negative-feedback pathways to regulate phototransduction by opposing the effect of light, thus producing active adaptation by the cell to light. One negative-feedback pathway acts on GC to enhance its activity. The activity of GC is facilitated by guanylate-cyclase-activating proteins (GCAPs), which in turn are controlled by Ca^{2+} in that they act well when not binding Ca^{2+} and poorly when they bind Ca^{2+}. Thus, in darkness, the GC activity is only moderate because the high intracellular free Ca^{2+} keeps the GCAPs in check, whereas in the light, the GC activity is enhanced to oppose light because the GCAPs are more active due to lower free Ca^{2+}. Another pathway apparently acts on the pigment kinase that phosphorylates Meta II to quench the latter's activity, via a Ca^{2+}-binding protein called recoverin. It is thought that recoverin in Ca^{2+}-bound form inhibits the pigment kinase. Thus, in the light, when Ca^{2+} is relatively low and recoverin loses its bound Ca^{2+}, the kinase quenches Meta II quickly to reduce amplification, again opposing light. A third pathway acts on the CNG channel via calmodulin, which in Ca^{2+}-bound form reduces the affinity of the channel for cGMP. Thus, in darkness, the Ca^{2+}-calmodulin binds to the CNG channel and lowers its probability of opening. In the light, calmodulin loses its bound Ca^{2+} and dissociates from the channel, as a result of which the CNG channel is more likely to reopen because of its now higher affinity for cGMP, thus opposing the effect of light. Calmodulin, GCAPs, and recoverin are all EF-hand-type Ca^{2+}-binding proteins. Of the three Ca^{2+}-mediated regulatory pathways, the one on GC is the most important for low and intermediate light levels, above which the one via recoverin begins to kick in, becoming increasingly important with still higher light levels. The regulatory pathway on the channel is weak and of minimal importance.

As mentioned previously, cones have a much lower sensitivity to light, and they also adapt to light much more effectively, compared to rods. The mechanisms underlying these rod–cone differences in the context of a common phototransduction process are still an area of active research. It might be pointed out here that most, though not all, of the proteins involved in transduction have distinct isoforms in rods and cones, so some quantitative differences in the process

between them are not surprising. Also, for rods, because they operate at very low light levels, the decrease in their sensitivity due to substantial bleaching of pigment by light is generally not important in their operating range of light intensities. Cones, on the other hand, operate in bright light. Thus, significant pigment bleaching is common and affects dramatically the sensitivity of cones to light in bright backgrounds. As mentioned earlier, pigment bleaching not only lowers the probability of light absorption, but also substantially reduces the signal generated by an absorbed photon because bare opsin (i.e., bleached pigment) itself has weak transducin-activating activity and triggers the negative-feedback pathways described earlier.

Diseases

Rods and cones can be severely compromised by a variety of disease conditions, genetic or otherwise, that lead to rod and cone dysfunction or irreversible degeneration, causing a partial or complete loss of vision. In the past decade or two, tremendous advances have taken place in the understanding of many of these diseases, thanks to the revelations of the cellular and molecular details of rod and cone phototransduction. These diseases include retinitis pigmentosa, Leber's congenital amaurosis, Oguchi disease, and stationary day blindness, to name just a few. Because of the intimate relationship between the retinal pigment epithelium and the photoreceptors, the degeneration of the latter will also occur when the pigment epithelium malfunctions. An example of such disease is vitelliform macular dystrophy (Best disease), which causes degeneration of the retinal fovea and its vicinity because the pigment epithelium has a defective chloride channel and presumably therefore abnormal transport function. Because the fovea is so critical for everyday vision, any disease affecting this retinal region is particularly incapacitating to the patient.

Parietal-Eye Photoreceptors in Lizard

This section briefly describes the lizard parietal-eye photoreceptor, for the reason that it shows very unusual antagonistic light responses, with underlying mechanisms that reveal the evolutionary path of rod and cone phototransduction.

Many lizards have a parietal eye, situated on the forehead between the two regular (lateral) eyes. The exact function of this eye is still not entirely clear, but the best suggestion so far is that it informs the animal about the passage of time in the course of the day by registering changes in the relative intensities of the spectral components of the ambient light as the day progresses.

The parietal eye resembles the lateral eyes by having a cornea, a lens, and a retina. The retina, however, has only photoreceptors, ganglion cells, and glial cells (i.e., no second-order and associative neurons, including bipolar, horizontal, and amacrine cells, and no retina pigment epithelium). It is also opposite in orientation to the lateral-eye retina in that the photoreceptors face the front of the eye and are therefore the first neurons in the retina to encounter light. The photoreceptors have a well-formed outer segment, with orderly stacked disks. Electron microscopy indicates that the disks, like those in cones, are continuous with the plasma membrane.

One very unusual feature of the parietal-eye photoreceptor is that each cell expresses two pigments: the blue-sensitive pinopsin (first found in the light-sensitive chicken pineal gland) and the green-sensitive parietopsin, an apparently ancient pigment not found in the lateral eyes. The coexistence of two pigments in a single cell fits with the physiology, which involves two antagonistic phototransduction pathways in the same cell. One of these, best activated by blue light, is identical to that in rods and cones, consisting of cGMP-PDE stimulation, decrease in cGMP concentration, closure of CNG channels, and membrane hyperpolarization. The other, best activated by green light, consists of inhibition of the same PDE, increase in cGMP concentration, opening of CNG channels, and membrane depolarization. Thus, it appears that pinopsin activates the hyperpolarizing pathway and parietopsin activates the depolarizing pathway. Interestingly, transducin is not present in these cells. Instead, gustducin appears to take its role, coupling to pinopsin. Although gustducin is not found in any other photoreceptor, it is considered to be a close transducin homolog, with a similar protein sequence and function. G_o, on the other hand, appears to couple to parietopsin. G_o is more different from transducin and gustducin, although still belonging to the same subfamily of G-proteins.

In vertebrates, the involvement of G_o in phototransduction is unique to the parietal-eye photoreceptor. However, in invertebrates, there is precedent for its involvement in vision, notably in the scallop hyperpolarizing photoreceptor, which as mentioned at the beginning of this article, is also of the ciliary type. In this scallop photoreceptor, G_o is coupled to SCOP2, an apparently ancient pigment, to activate a cGMP-signaling pathway analogous, though not identical, to that in rods and cones. Thus, all evidence suggests that a G_o-mediated phototransduction pathway is ancient among ciliary photoreceptors. As such, the parietal-eye photoreceptor appears to retain features

of rod and cone predecessors – a missing link between rods/cones and more primitive ciliary photoreceptors, so to speak. The notion is that a G_o-mediated phototransduction pathway existed in ciliary photoreceptors long ago, before vertebrates evolved. In the course of evolution, however, a chromatically antagonistic, gustducin/transducin-mediated phototransduction pathway was added to the cells having the G_o-mediated pathway for the purpose of analyzing spectral information. This chromatic antagonism is retained by the parietal-eye photoreceptor to the present day. In parallel during evolution, the lateral eyes made their appearance over time but retained only the more recent gustducin/transducin-mediated pathway, delegating the color opponency instead to downstream neurons – a good strategy because complex processing of color information is best provided by elaborate synaptic circuitry. As such, the chromatic opponency within a single parietal-eye photoreceptor represents a very primitive, simple form of color vision compared to, say, what goes on in the lateral-eye retina.

Incidentally, the parietal eye, like the two lateral eyes (and the pineal gland), develops as a protrusion of the diencephalon of the brain during embryogenesis, again pointing its homology to these other structures.

See also: Color Vision; Photoreceptor Adaptation; Photoreceptor Mosaic; Phototransduction; Retinal Bipolar Cells; Retinal Development: An Overview; Rhodopsin.

Further Reading

Baylor DA (1987) Photoreceptor signals and vision. The proctor lecture. *Investigative Ophthalmolgy and Visual Science* 28: 34–49.

Baylor DA, Lamb TD, and Yau K-W (1979) Responses of retinal rods to single photons. *Journal of Physiology* 288: 613–634.

Baylor DA, Lamb TD, and Yau K-W (1979) The membrane current of single rod outer segments. *Journal of Physiology* 288: 589–611.

Baylor DA, Matthews G, and Nunn BJ (1984) Location and function of voltage-sensitive conductances in retinal rods of the salamander,. *Ambystoma tigrinum*. *Journal of Physiology* 354: 203–223.

Baylor DA, Matthews G, and Yau K-W (1980) Two components of electrical dark noise in toad retinal rod outer segments. *Journal of Physiology* 309: 591–621.

Breton ME, Schueller AW, Lamb TD, and Pugh EN, Jr. (1994) Analysis of ERG a-wave amplification and kinetics in terms of the G-protein cascade of phototransduction. *Investigative Ophthalmology and Visual Science* 35: 295–309.

Burns ME and Baylor DA (2001) Activation, deactivation, and adaptation in vertebrate photoreceptor cells. *Annual Review of Neuroscience* 24: 779–805.

Dowling JE (1987) *The Retina: An Approachable Part of the Brain*. Harvard, MA: Belknap.

Dryja TP and Li T (1995) Molecular genetics of retinitis pigmentosa. *Human Molecular Genetics* 4: 1739–1743.

Fain GL, Matthews HR, Cornwall MC, et al. (2001) Adaptation in vertebrate photoreceptors. *Physiological Reviews* 81(1): 117–151.

Fesenko EE, Kolesnikov SS, and Lyubarsky AL (1985) Induction by cyclic GMP of cationic conductance in plasma membrane of retinal rod outer segment. *Nature* 313: 310–313.

Finn JT, Grunwald ME, and Yau K-W (1996) Cyclic nucleotide-gated ion channels: An extended family with diverse functions. *Annual Review of Physiology* 58: 395–426.

Koutalos Y and Yau K-W (1996) Regulation of light sensitivity in retinal rods. *Trends in Neuroscience* 19: 73–81.

Lamb TD, McNaughton PA, and Yau K-W (1981) Spatial spread of activation and background desensitization in toad rod outer segments. *Journal of Physiology* 319: 463–496.

Lamb TD and Pugh EN, Jr. (2004) Dark adaptation and the retinoid cycle of vision. *Progress in Retinal and Eye Research* 23(3): 307–380.

Pepperberg DR, Brown PK, Lurie M, et al. (1978) Visual pigment and photoreceptor sensitivity in the isolated skate retina. *Journal of General Physiology* 71: 369–396.

Rodieck RW (1998) *The First Steps in Seeing*. Sunderland, MA: Sinauer Associates, Inc.

Su C-Y, Luo D, Terakita A, et al. (2006) Parietal-eye phototransduction components and their potential evolutionary implications. *Science* 311: 1617–1621.

Yau K-W (1994) Phototransduction mechanism in retinal rods and cones. The Friedenwald Lecture. *Investigative Ophthalmology and Visual Science* 35: 10–32.

Yau K-W, Lamb TD, and Baylor DA (1977) Light-induced fluctuations in membrane current of single toad rod outer segments. *Nature* 269: 78–80.

Phototransduction

J B Hurley, University of Washington, Seattle, WA, USA

© 2009 Published by Elsevier Ltd.

Natural scenes produce increments and decrements of illumination on the retina. Phototransduction is the series of biochemical reactions by which rod and cone photoreceptors detect those changes in illumination. Photoreceptor neurons are highly compartmentalized. Phototransduction occurs in the photoreceptor outer segment, a compartment that is highly enriched with membranes and with the enzymes that catalyze the biochemical conversion of light absorption into changes in membrane potential.

Rod photoreceptors initiate scotopic vision – seeing in dim light. Rods are so sensitive that they can respond to individual photons. Cone photoreceptors initiate photopic vision – seeing in bright light. Cones are less sensitive than rods, but they are endowed with the ability to adapt to continuous stimulation even when it is very intense. Vertebrates use the high sensitivity of rods and the adaptability of cones to respond to changes in illumination on backgrounds that range over 11 log units of intensity.

Initiation of Phototransduction

Rods achieve their extraordinary photosensitivity by using a biochemical amplifier. The amplification occurs in a light-activated enzymatic cascade that controls the concentration of 3′,5′ cyclic guanosine monophosphate (cGMP), an intracellular messenger. cGMP regulates the activity of plasma membrane ion channels that allow Na^+ entry. The photoreceptor membrane potential is determined in part by its permeability to Na^+. In darkness, the concentration of cytosolic, unbound cGMP is a few micromolar, enough to partially activate the channels and depolarize the rod. Light stimulates the breakdown of cGMP to close the channels and hyperpolarize the cell.

The detector of light in rods is rhodopsin, a 36 kDa glycoprotein embedded in the disc and plasma membranes of the outer segment. Rhodopsin has seven transmembrane helices, a glycosylated N-terminus, several loops that connect the helices, and a C-terminal tail that extends into the cytoplasm. The 11-*cis* retinal chromophore on rhodopsin can absorb light. Its spectral sensitivity is broad but it is most sensitive to 500 nm photons. Absorption of light drives the chromophore into an excited state. Approximately 70% of the time, the retinal relaxes from this state to an all-trans rather than the original 11-*cis* configuration. The all-*trans*-retinal, newly formed within a hydrophobic pocket of rhodopsin, induces a series of structural changes that propagate through the protein. At one point, the photoactivated rhodopsin takes on a configuration known as Metarhodopsin II (Meta II, also referred to as R*). A representation of the protein–protein interactions that occur during the initial stages of phototransduction is shown in Figure 1. A list of the important steps in the activation of phototransduction is shown in Figure 2. The activated state of rhodopsin (R*) binds to a peripheral membrane protein called transducin ('T'), a heterotrimeric GTP-binding protein (G-protein). The Meta II–transducin interaction stimulates the 39 kDa α subunit of transducin to release guanosine 5′-diphosphate (GDP) that was bound to it and to replace it with guanosine 5′-triphosphate (GTP). Then the GTP-bound α subunit (39 kDa) dissociates from the complex. Activation of transducin in this way is a significant step of amplification. Approximately 100 transducin molecules can be activated by a single Meta II during its lifetime.

Once activated, transducin α diffuses from the site of activation. Hydrolysis of cGMP begins when transducin relieves inhibitory constraints imposed by the small (11 kDa) γ subunits on the larger (90 kDa) catalytic subunits ($\alpha\beta$) of a cGMP-specific phosphodiesterase (Pde). This is the second amplification step of phototransduction. Each Pde can hydrolyze 1000 cGMP per second in a reaction limited only by diffusion of the substrate to the enzyme. The overall amplification of phototransduction in rods is robust. A single photon can stimulate the breakdown of as many as 10^5 molecules of cGMP.

The hydrolysis of cGMP stimulated by light inactivates cGMP-gated cation channels in the plasma membrane of the outer segment (Figure 2). The cellular hyperpolarization caused by closure of these channels is detected by voltage-gated Ca^{2+} channels at the rod terminal. As the activity of these channels is reduced, less Ca^{2+} enters the synaptic terminal and release of glutamate into the synapse slows.

Restoration of the Dark State and Mechanisms of Adaptation

Most biochemical and physiological analyses of phototransduction in the laboratory have made use of brief pulses of light to initiate the photoreceptor response. A flash stimulus can reveal the initial response of a photoreceptor to light and it can also

reveal how the photoreceptor recovers from an impulse. The closure of cGMP-gated channels by a flash of light is transient. The channels reopen, reflecting processes that restore the concentration of cGMP to its dark level. The flash response also provides clues about mechanisms that can contribute to adaptation. Since rhodopsin, transducin, and phosphodiesterase all contribute to cGMP hydrolysis, each must be inactivated to return to the dark state. Resynthesis of cGMP also helps restore the rod to its dark state.

Some of the processes that restore the rod to its dark state occur independently of Ca^{2+}. Others are influenced by changes in the concentration of intracellular free Ca^{2+} that are stimulated by light. The biochemical reactions that inactivate phototransduction are summarized in Figure 3.

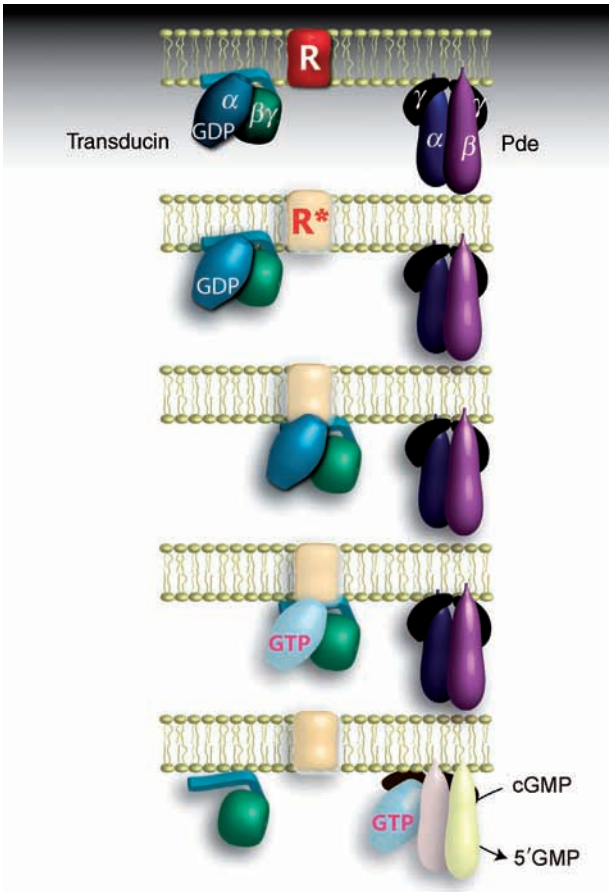

Figure 1 A representation of the protein–protein interactions that occur during the initial events of phototransduction. Light converts rhodopsin (R) into an active form, metarhodopsin II (R*), that binds the heterotrimer G-protein, transducin. Formation of that complex stimulates the replacement of GDP with GTP on the transducin α subunit. The activated transducin then activates the phosphodiesterase (Pde) catalytic subunits (α and β) by relieving inhibitory constraints imposed by the inhibitory (γ) subunits.

Ca^{2+}-Independent Mechanisms for Inactivation

Inactivation of photoactivated rhodopsin Meta II is a transient intermediate, but its lifetime is too long to account for the rapid reopening of channels that occurs following brief flashes of light. A reaction that inhibits Meta II activity more rapidly is phosphorylation of the meta II C-terminus. Multiple serines and threonines on the rhodopsin C-terminus can be phosphorylated by rhodopsin kinase (a 68 kDa soluble protein also known as G-protein-coupled receptor kinase 1 (GRK1)). Rhodopsin kinase phosphorylates rhodopsin slowly in darkness, but its activity is stimulated by binding to Meta II. Phosphorylated Meta II is a less effective activator of transducin than unphosphorylated Meta II. Phosphorylated Meta II also has a high affinity for a 48 kDa protein known as arrestin. Arrestin binds to phosphorylated Meta II and blocks further activation of transducin. The roles that rhodopsin kinase and arrestin play in restoration of the dark state were originally inferred from *in vitro* biochemical analyses, but the roles have been confirmed by physiological analyses of genetically modified animals. Rod photoresponses are prolonged in mice in which either the GRK1 or arrestin gene has been inactivated.

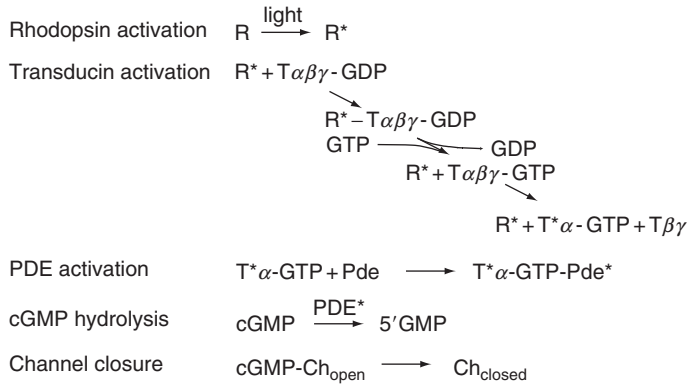

Figure 2 Steps in the activation of phototransduction. T, transducin; Pde, phosphodiesterase; Ch, cGMP-gated channel.

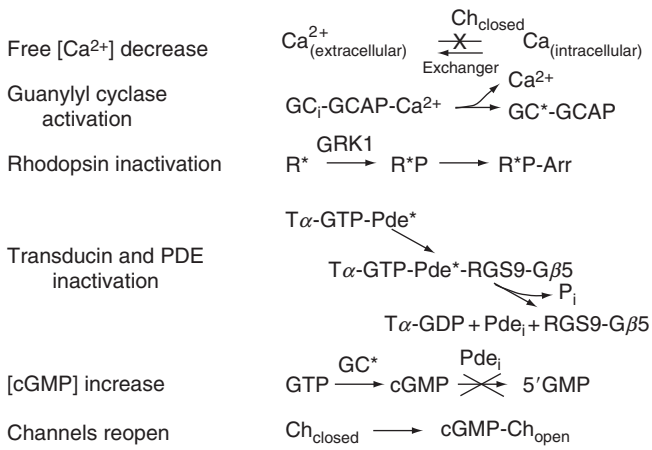

Figure 3 Steps in the deactivation of phototransduction. Arr, arrestin; GRK1, rhodopsin kinase; GC, guanylyl cyclase; GCAP, guanylyl cyclase activator proteins; P, a phosphoryl residue; RGS9-Gβ5, a complex of the regulator of G-protein signaling 5 protein with a G-protein β-subunit 5 protein.

Inactivation of transducin Transducin deactivates when it hydrolyzes the GTP bound to it. The reaction is intrinsically slow but a cluster of proteins known as the RGS-9 complex (regulator of G-protein signaling) binds transducin and stimulates it to hydrolyze its bound GTP. The RGS9 complex consists of an RGS9 subunit, a subunit similar in structure to G-protein β subunits ($G_{\beta 5}$), and a membrane-anchoring protein known as R9AP.

Inactivation of phosphodiesterase Once transducin hydrolyzes its bound GTP, the Pde inhibitor (γ) reassociates with the catalytic subunits to inhibit cGMP hydrolysis.

Ca^{2+}-Sensitive Inactivation Mechanisms

In darkness, Ca^{2+} influx and efflux are balanced. Free Ca^{2+} is flushed from the outer segment through a Na^+/Ca^{2+}; K^+ exchanger while influx proceeds through open cGMP-gated channels. The onset of illumination causes a net efflux of Ca^{2+} because the light-induced closure of channels blocks Ca^{2+} entry. Within seconds the concentration of cytosolic free Ca^{2+} decreases from ~500 nM to ~10–50 nM.

cGMP hydrolysis and synthesis also are balanced in darkness. When a flash of light stimulates phosphodiesterase activity, cGMP hydrolysis very quickly exceeds synthesis. However, the balance of hydrolysis and synthesis is restored by a Ca^{2+}-mediated feedback mechanism. cGMP is synthesized in rods by a 110 kDa guanylyl cyclase known as RetGC or ROS-GC. This dimeric membrane protein is not directly regulated by light. Instead, it is regulated by Ca^{2+} and a Ca^{2+}-binding protein known as guanylyl cyclase activator protein (GCAP). In darkness, when [Ca^{2+}] is highest, GCAP binds Ca^{2+} and inhibits guanylyl cyclase activity. When light lowers the intracellular free Ca^{2+} concentration, the apoform of GCAP stimulates the cyclase to restore cGMP and reopen channels.

Biochemical studies performed on photoreceptor homogenates and purified enzymes have revealed that Ca^{2+} can also regulate cGMP-gated channels. Ca^{2+}–calmodulin lowers the affinity of cGMP-gated channels for cGMP. When Ca^{2+} levels decline in light, the channels bind cGMP with greater affinity and reopen. The effect of Ca^{2+} and calmodulin on channel activity appears to be greater in cones than in rods. *In vitro* experiments suggest that Ca^{2+} can also affect rhodopsin phosphorylation. The binding of Ca^{2+} to a protein called 'recoverin' or 'S-modulin' can inhibit rhodopsin kinase activity. Recoverin enhances vision in dim light, but the interpretation of biochemical and physiological studies of recoverin has been controversial. The mechanism by which recoverin influences visual sensitivity is uncertain.

Effects of light on the intracellular distributions of proteins Light influences the distributions of transducin and arrestin among the intracellular compartments of photoreceptors. Transducin concentrates in the outer segments of rods and cones in darkness, but in rods it disperses throughout the cell in light. Arrestin also redistributes, but in the direction opposite to transducin. Arrestin concentrates within the inner segments in darkness, and in light it concentrates in the outer segments. Arrestin redistributes in both rods and cones; transducin redistributes only in rods. The mechanism of redistribution of these proteins is under investigation. Evidence indicates that passive diffusion and trapping by protein–protein and protein–membrane interactions are

important. However, the stoichiometry and light sensitivity of the redistribution process suggest that additional mechanisms may also contribute.

• *Phototransduction in rods versus cones.* Rods respond well only in dim light; cones function best in bright light. Flash responses of rods can last more than a second, whereas cone responses require only a few hundred milliseconds for completion. Rods adapt to light only over a limited range before they saturate, whereas cone saturation is never more than transient. The biochemical activities associated with phototransduction in rods and cones are specialized and may contribute to these differences in physiology. Many of the enzymes associated with phototransduction are expressed from distinct genes in rods versus cones. These include opsins; transducin α, β, and γ subunits; Pde catalytic and inhibitor subunits; and cGMP-gated channel subunits. Other phototransduction enzymes have expression patterns that overlap in rods and cones; the extent of overlap varies across vertebrate species. These include guanylyl cyclases, GCAPs, arrestins, and GRKs. Other proteins (e.g., recoverin and RGS9) are always expressed from the same gene in rods and cones but the levels of expression can differ.

Rods are extraordinarily sensitive to light, so the phototransduction enzymes within them are optimized for signal amplification. The highly sensitive and prolonged responses associated with rods are endowed not by the type of visual pigment expressed but instead by the downstream components of the G-protein cascade. When a cone visual pigment is used to initiate a dim flash response in rods, the response does not differ significantly from a normal response initiated by rhodopsin.

Cones are less sensitive than rods, and they generate faster responses with shorter lifetimes. Transducin is activated less efficiently, and it activates PDE less efficiently in cones than in rods. The level of RGS9 is greater in cones than in rods, so transducin is deactivated faster in cones. In some species, visual pigment kinase activity is also greater in cones than in rods. These types of biochemical differences optimize the performances of rods and cones for nighttime and daytime vision, respectively.

• *Mathematical modeling of phototransduction.* Biochemical and physiological studies suggest that phototransduction is a relatively simple signal transduction pathway. Only one type of stimulus, light, and only one type of receptor, a visual pigment, trigger phototransduction and adaptation. One cannot conclude with certainty that we currently know all of the components in the phototransduction pathway and all of the factors that influence it, but the major components of phototransduction are identified and they have been characterized extensively. Enough information is available about the biochemical activities and the physiological responses of photoreceptors that mathematical models that are consistent with the available biochemical and physiological data can be designed. A puzzling feature of the process of phototransduction in rods has been the unexpectedly small amount of variation in the amplitude and kinetics of rod responses to single photons. Mathematical modeling has been useful for making testable hypotheses that can be used to investigate this unusual phenomenon.

Structures of Phototransduction Proteins

The molecular structures of transducin, RGS9, arrestin, recoverin, and rhodopsin – important proteins that influence phototransduction – have been determined either by nuclear magnetic resonance spectroscopy of native proteins in solution or by x-ray diffraction of crystallized proteins. The structures are consistent with the biochemical activities of the proteins and have provided important insights into the mechanisms of regulation. The molecular architectures of sites that bind key molecules of phototransduction – GTP, cGMP, Ca^{2+}, and 11-*cis*-retinal – are now known and will likely be useful for understanding, and perhaps ultimately treating, the consequences of mutations that cause retinal disease.

Phototransduction and Retinal Disease

Photoreceptors are terminally differentiated neurons specialized to detect light. High metabolic demands, a highly polarized cellular morphology, demands for efficient protein synthesis, and very active vectorial transport systems make the viability of these cells extraordinarily sensitive to mutations that perturb their function.

Mutations in nearly all genes encoding phototransduction enzymes compromise specific aspects of photoreceptor function. In some cases, however, there is a more nefarious outcome. Mutations in some phototransduction genes also cause photoreceptors to die by apoptosis. For example, mutations that interfere with expression of Pde catalytic subunits cause cGMP to accumulate. Elevated cGMP stabilizes the open state of cGMP-gated channels so that Ca^{2+} flows in and accumulates to excessive concentrations in the photoreceptor. Excess Ca^{2+} has been linked to apoptosis. Mutations that raise the concentration of

Ca^{2+} required to inhibit guanylyl cyclase also cause apoptosis, probably through a similar mechanism.

Mutations in the enzyme that synthesizes cGMP also cause apoptosis, although the mechanism by which apoptosis is induced surely must be different. Without guanylyl cyclase, cGMP is not made, channels remain closed, and the inward flux of Ca^{2+} is reduced. Abnormally low levels of Ca^{2+} may also be linked to apoptosis. Continuous illumination of photoreceptors, as well as mutations in rhodopsin that cause persistent activation of the phototransduction pathway, also induce apoptosis, perhaps by a similar mechanism. Mutations in the genes encoding rhodopsin kinase and arrestin, proteins required to inactivate phototransduction, do not cause apoptosis in complete darkness. However, in normal lighting conditions these types of mutations allow unbridled phototransduction and ultimately apoptosis.

Other types of mutations are lethal to photoreceptors as well. Mutations that interfere with protein folding or transport cause rhodopsin to accumulate in the inner segment of the cell. Accumulation of protein in the wrong compartment of a photoreceptor cell may be another trigger for apoptosis. Normal expression of proteins with specific structural roles can also be critical for survival. Peripherin/RDS is a protein required for normal disc morphology in rods and for proper infolding of the plasma membrane in cones. In the absence of peripherin, rod morphology is compromised and apoptosis ensues. Photoreceptor apoptosis can also be stimulated by mutations that block the activity of transcription factors needed to express important photoreceptor-specific genes.

Not all mutations in phototransduction genes induce apoptosis. Mutations that block expression of transducins in rods and cones block the normal phototransduction process, but they do not cause photoreceptors to die. Mutations that prevent the expression of GCAP and recoverin genes also do not trigger degeneration of photoreceptors.

Conclusion

Phototransduction is a useful model for G-protein-mediated signal transduction. Biochemical investigations of the pathway have been facilitated by the abundance of phototransduction proteins, which are highly concentrated in the outer segments of photoreceptors. The ability to use genetic manipulations to investigate phototransduction has been facilitated by the nonessential nature of vision. This has empowered investigators of phototransduction with the ability to perform in-depth analyses of biochemical and physiological consequences of mutations in photoreceptor genes. Knowledge gained from biochemical, genetic, and physiological studies of phototransduction provides important insights about vision and about other G-protein-mediated signaling pathways.

See also: Animal Models of Inherited Retinal Degenerations; Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Development: An Overview; Rhodopsin.

Further Reading

Burns ME and Arshavsky VY (2005) Beyond counting photons: Trials and trends in vertebrate visual transduction. *Neuron* 48: 387–401.

Burns ME and Baylor DA (2001) Activation, deactivation, and adaptation in vertebrate photoreceptor cells. *Annual Review of Neuroscience* 24: 779–805.

Chen CK (2005) The vertebrate phototransduction cascade: Amplification and termination mechanisms. *Reviews of Physiology, Biochemistry, and Pharmacology* 154: 101–121.

Cowan CW, He W, and Wensel TG (2001) RGS proteins: Lessons from the RGS9 subfamily. *Progress in Nucleic Acid Research and Molecular Biology* 65: 341–359.

Dizhoor AM (2000) Regulation of cGMP synthesis in photoreceptors: Role in signal transduction and congenital diseases of the retina. *Cell Signal* 12: 711–719.

Fain GL (2006) Why photoreceptors die (and why they don't). *Bioessays* 28: 344–354.

Gray-Keller MP and Detwiler PB (1994) The calcium feedback signal in the phototransduction cascade of vertebrate rods. *Neuron* 13: 849–861.

Hamer RD, Nicholas SC, Tranchina D, Lamb TD, and Jarvinen JL (2005) Toward a unified model of vertebrate rod phototransduction. *Visual Neuroscience* 22: 417–436.

Hirsch JA, Schubert C, Gurevich VV, and Sigler PB (1999) The 2.8 A crystal structure of visual arrestin: A model for arrestin's regulation. *Cell* 97: 257–269.

Kennedy MJ, Lee KA, Niemi GA, et al. (2001) Multiple phosphorylation of rhodopsin and the *in vivo* chemistry underlying rod photoreceptor dark adaptation. *Neuron* 31: 87–101.

Leskov IB, Klenchin VA, Handy JW, et al. (2000) The gain of rod phototransduction: Reconciliation of biochemical and electrophysiological measurements. *Neuron* 27: 525–537.

Nair KS, Hanson SM, Mendez A, et al. (2005) Light-dependent redistribution of arrestin in vertebrate rods is an energy-independent process governed by protein–protein interactions. *Neuron* 46: 555–567.

Noel JP, Hamm HE, and Sigler PB (1993) The 2.2 A crystal structure of transducin-alpha complexed with GTP gamma S. *Nature* 366: 654–663.

Okada T, Sugihara M, Bondar AN, Elstner M, Entel P, and Buss V (2004) The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure. *Journal of Molecular Biology* 342: 571–583.

Polans A, Baehr W, and Palczewski K (1996) Turned on by Ca^{2+}! The physiology and pathology of Ca^{2+}-binding proteins in the retina. *Trends in Neuroscience* 19: 547–554.

Rivolta C, Sharon D, DeAngelis MM, and Dryja TP (2002) Retinitis pigmentosa and allied diseases: Numerous diseases, genes, and inheritance patterns. *Human Molecular Genetics* 11: 1219–1227.

Sakmar TP, Menon ST, Marin EP, and Awad ES (2002) Rhodopsin: Insights from recent structural studies. *Annual Review of Biophysics and Biomolecular Structure* 31: 443–484.

Schertler GF (2005) Structure of rhodopsin and the metarhodopsin I photointermediate. *Current Opinion in Structural Biology* 15: 408–415.

Sokolov M, Lyubarsky AL, Strissel KJ, et al. (2002) Massive light-driven translocation of transducin between the two major compartments of rod cells: A novel mechanism of light adaptation. *Neuron* 34: 95–106.

Relevant Websites

http://webvision.med.utah.edu – John Moral Eye Center, University of Utah.

http://www.sph.uth.tmc.edu – University of Texas School of Public Health.

PHR (Pam/Highwire/RPM-1)

Y Jin, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Synapse formation is a highly regulated process resulting in morphologically distinct subcellular structures in the pre- and postsynaptic sites to transmit and receive signals, respectively. At the presynaptic site, synaptic vesicles are clustered around the electron-dense projections near the synaptic plasma membrane. Although individual synaptic boutons exhibit tremendous variability in morphology and size, synapses from the same type of neurons share common characteristics and a given type of neurons often displays a stereotyped pattern of synapses. In 2000, three independent studies reported the identification of two homologous proteins in *Drosophila* and *Caenorhabditis elegans*, known as Highwire (Hiw) and RPM-1 (Regulator of Presynaptic Morphology-1), respectively. Loss of function in these genes causes specific alterations in the stereotyped synapse patterns and synapse organizations. Both invertebrate proteins are othologous to a human protein Pam (Protein associated with Myc), which was identified in 1998 as a binding protein for the Myc oncogene. Nearly all sequenced genomes of multicellular organisms contain a homologous member (see the Tree Families database website). I here refer this conserved protein family as the PHR proteins to honor the original members (Pam/Highwir/RPM-1) and the presence of a unique motif within the family. It is now established that a conserved function of the PHR proteins is to act as an ubiquitin E3 ligase. PHR proteins have been shown to regulate multiple signaling pathways involving mitogen-activated protein kinases (MAPKs), bone morphogenetic protein (BMP) pathway, adenylate cyclase, and the tuberous sclerosis complex (TSC).

PHR Protein Structure and Gene Expression

The PHR proteins are giant (Figure 1). *C. elegans* RPM-1, the smallest member, has 3766 amino acids; *Drosophila* Hiw, the largest, has 5233 amino acids. Given the size of the proteins, the degree of sequence conservation is remarkable. The N-terminus of these proteins has an RHD (RCC1 homology domain). RCC1 is a guanine exchange factor for the Ran guanosine triphosphatase (GTPase) and is composed of seven repeats, forming a seven-bladed propeller structure. The RHD domain in PHR proteins is often separated into two regions RHD1 and RHD2. Following RHD is a conserved region that contains two repeats of 90 amino acids, referred as PHR repeats because they are unique among the PHR proteins. The most conserved region is in the C-terminus, which contains a zinc finger B-box, a RING finger, and shares over 50% amino acid identity from *C. elegans* RPM-1 to human Pam. The vertebrate PHR proteins have a Myc binding domain in the middle. This region in Pam has been shown to be variable due to alternative splicing. Several additional predicted domains, including leucine zipper, nuclear localization signal, and a histone binding protein homologous domain, are also found only in the vertebrate PHR proteins.

Expression studies of mRNA and functional tagged reporter genes for Hiw and RPM-1 show that both proteins are predominantly expressed in the nervous system and are enriched in presynaptic terminals. Expression of RPM-1 or Hiw in presynaptic neurons also rescues the mutant phenotypes. Neither protein is detectable in the nucleus, even when it is overexpressed. Within the presynaptic terminals, the proteins seem to reside in a unique region because co-immunostaining with antibodies against proteins for synaptic vesicles, presynaptic active zone, and several synaptic membrane proteins exhibits very little overlap. This expression domain was initially called the presynaptic periactive zone, based on the studies in *Drosophila* neuromuscular junctions (NMJs). However, the original Hiw antibody was recently shown to recognize a different protein. Thus, the precise subcellular location for Hiw and RPM-1 at synapses remains to be characterized.

The zebra fish Esrom and vertebrate PHR proteins are broadly expressed in many tissues and are most abundant in the nervous system. In embryos, Pam and Phr1 mRNA are detected in the hippocampus, dendate gyrus, cerebellum, retina, spinal cord, dorsal root ganglia (DRG), and cranial ganglia, with high levels around the major synaptogenesis period. Pam and Phr1 mRNAs are also upregulated in the postnatal nervous system and persisted into adulthood. The data on the vertebrate PHR proteins are controversial, depending on the antibodies and cell type. The full-length Pam protein is ~510 kDa. However, two antibodies detect Pam at 450 and 350 kDa, with the latter being predominant in postnatal nervous system. Another antibody that appears to be enriched in some schizophrenic patients reacts with Pam and detects a protein of 60 kDa, which has been suggested to be a proteolytic product of Pam. In the DRG neurons, Pam is not found in the nuclei and appears to be present in the cytoplasm and axons; in the adult cortex, it is localized to nuclei and axon bundles in all layers.

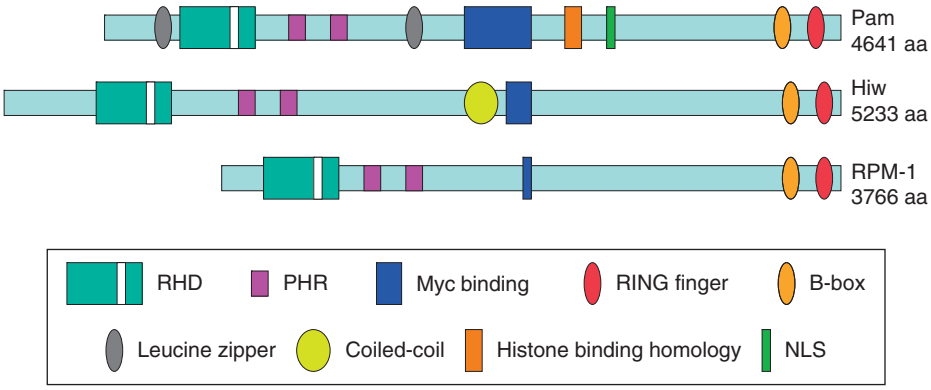

Figure 1 Schematic comparison of Pam, Hiw, and RPM-1. Domain structures are marked. NLS, nuclear localization signals; PHR, Pam/Highwire/RPM-1; RHD, RCC1 homology domain.

Mutant Phenotypes for PHR Genes

The phenotypic analysis of mutant animals provides the first important clues for the function of the PHR protein family. Given the large size and the broad expression of these proteins, it is not surprising that a range of defects are observed and that the terminal phenotypes seem to vary depending on the neuron types and the organisms.

C. elegans neurons have simple morphologies and form *en passant* synapses at defined regions along axons. The positions and the ultrastructural features of the synapses exhibit neuron type specificity. *rpm-1* mutants were isolated from several genetic screens that relied on a synaptic vesicle reporter gene to visualize the morphology and pattern of synapses of different neurons. The motor neurons innervate body-wall muscles and form regularly spaced synaptic boutons along the entire or a segment of the axon. In *rpm-1* mutants, the motor synaptic boutons are irregular in size and in distribution (Figure 2). Within a single presynaptic terminal, the total number of synaptic vesicles is reduced, and frequently, numerous presynaptic dense projections are present. The mechanosensory neurons form synapses through short axonal branches. In *rpm-1* mutants, these synaptic branches frequently do not form, leading to a failure of synapse formation (Figure 2). Moreover, the mechanosensory axons continue to grow beyond their normal termination sites, resulting in an overshooting phenotype. Similar phenotypes are seen for several types of motor and sensory neurons. Despite the synaptic defects, *rpm-1* mutant animals are viable and exhibit nearly normal movement.

Drosophila larval motor neurons elaborate terminal branches on reaching a muscle surface, and each terminal branch forms synaptic boutons of a characteristic pattern (Figure 3). During larval development, the motor terminals undergo a regulated growth in branch and bouton numbers to match the increase in muscle surface. *hiw* mutants were originally identified by screening for flies with walking deficits. In *hiw* mutant larvae, the motor terminals exhibit excessive branching, or synaptic overgrowth, increasing synaptic bouton numbers two- to fourfold and covering twice more synaptic area (Figure 3). However, individual synaptic boutons are smaller than those in wild type. The synaptic content measured by antibody staining to synaptotagmin and neurotransmitter transporter is reduced. Electrophysiological recordings also show that the quantal size and quantal content from the mutant synapses are reduced.

The zebra fish gene *Esrom* was isolated through a large-scale screen for aberrant retinal topographic mapping defects. In wild-type fish, retinal axons from the anterior eye are fasciculated until they reach the posterior tectum. In *esrom* mutants, the anterior axons branch in the anterior tectum and fail to map to the posterior tectum (Figure 4). Some retinal axons also show segregation and fasciculation defects. In addition, *esrom* mutant fish have reduced yellow pigmentation as the result of defects in the biosynthesis of pteridine.

The mouse gene *Phr1* resides on chromosome 14 within the piebald deletion complex, a region syntenic with human chromosome 13q22 that contains the Pam gene. A compound deficiency that deletes *phr1* and two neighboring genes causes animals to die at birth due to respiratory defects. In these mutant mice, the phrenic nerve is thin and has large varicosities at the axon ends. At the NMJs, nerve terminals overshoot and sprout beyond the postsynaptic receptor plaque (Figure 5). Based on the expression pattern of *phr1* and the two neighboring genes, these defects are probably due to a specific loss of *Phr1* function. The report that an antibody shows preferential reactivity with Pam in schizophrenic patients suggests a potential link between Pam and human diseases. In the adult

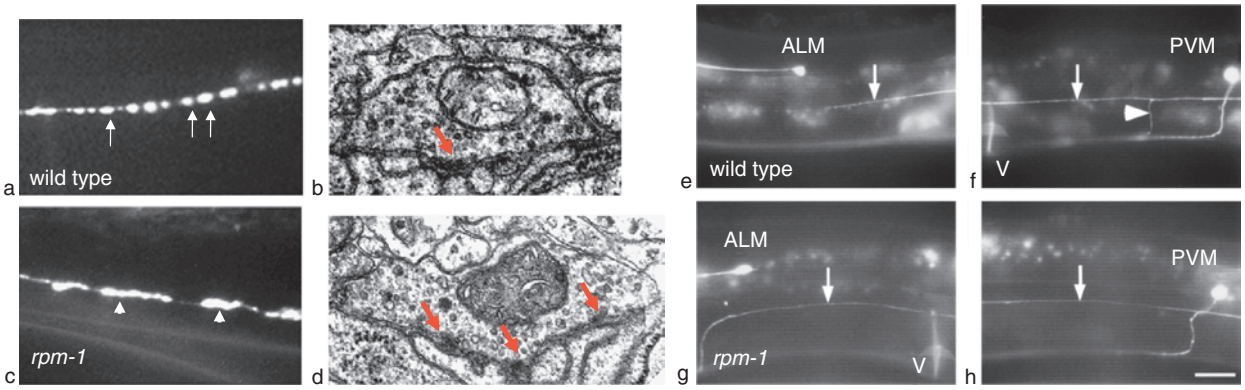

Figure 2 Synaptic defects in *C. elegans rpm-1* mutants: (a, b) synapses for the GABAergic motor neurons in the wild type; (c, d) synapses for the GABAergic motor neurons in *rpm-1* mutants; (e, f) mechanosensory neurons in the wild type; (g, h) mechanosensory neurons in *rpm-1* mutants. In wild-type animals in (a, b), the SNB-1::GFP marker forms uniformly shaped fluorescent puncta (arrows) that are evenly distributed along the nerve process (a) and in most synapses there is only one prominent electron-dense projection (arrow) within a cluster of synaptic vesicles (b). In the *rpm-1* mutants in (c, d), the SNB-1::GFP puncta are enlarged and irregular (arrowheads in (c)) and several presynaptic projections (arrows) are often present within a single cluster of synaptic vesicles (d). In the wild-type animals in (e, f), PVM neurons (arrow) terminate just posterior to the ALM cell body (e) and extend an axon branch (arrownhead in (f)) to the synapse on to the ventral cord neurons. In the *rpm-1* mutants in (g, h), PVM axons fail to terminate (g) and the synaptic branches are usually absent (h). ALM, anterior lateral microtubule; GABA, γ-aminobutyric acid; GFP, green fluorescent protein; PLM, posterior lateral microtubule; PVM, posterior ventral microtubule. (a–d) Reproduced from Zhen M, Huang X, Bamber B, and Jin Y (2000) Regulation of presynaptic terminal organization by *C. elegans* RPM-1, a putative guanine nucleotide exchanger with a RING-H2 finger domain. *Neuron* 26: 331–343, with permission from Elsevier; Nakata K, Abrams B, Grill B, et al. (2005) Regulation of a DLK-1 and p38 MAP kinase pathway by the ubiquitin ligase RPM-1 is required for presynaptic development. *Cell* 120: 407–420, with permission from Elsevier. (e–h) Reproduced from Schaefer AM, Hadwiger GD, and Nonet ML (2000) rpm-1, a conserved neuronal gene that regulates targeting and synaptogenesis in *C. elegans. Neuron* 26: 345–356, with permission from Elsevier.

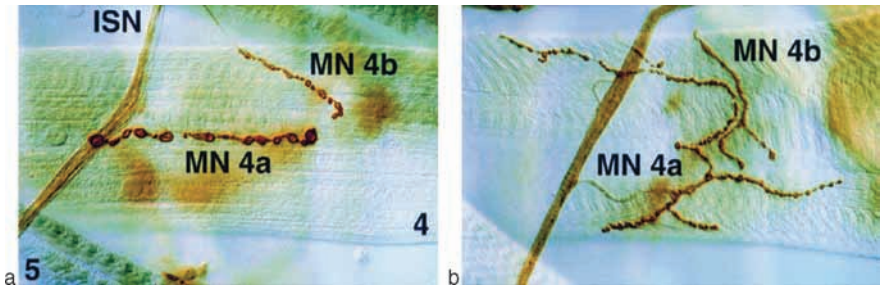

Figure 3 NMJ defects in *Drosophila Highwire* mutants wild-type larva; (b) *hiw* mutant larva. Shown are third instar larvae stained with anti-Syt to reveal the presynaptic branches and boutons. In wild-type larvae (a), muscle 4 is innervated by two motor neurons (MN4a and MN4b). In *hiw* mutants (b), both motor terminals are present on muscle 4, but contain many more branches and boutons. NMJ, neuromuscular junction. Reproduced from Wan HI, DiAntonio A, Fetter RD, Bergstrom K, Strauss R, and Goodman CS (2000) Highwire regulates synaptic growth in *Drosophila. Neuron* 26: 313–329, with permission from Elsevier.

rat spinal cord, inhibition of Pam appears to enhance nociceptive behaviors.

Thus, the analysis of the mutant phenotypes supports an important role of PHR proteins in the precision of synapse pattern and architecture. Genetic studies using temperature-sensitive assays showed that at least some of the defects arise from initial differentiation and that, in the case of *hiw*, it is also required throughout larval development. One interpretation of the mutant phenotypes is that PHR proteins are required for the stabilization of nascent synaptic architecture. In the absence of PHR proteins, nascent synapses are not developed to its maturity; different neurons then respond differently by sprouting, overgrowth, or withdraw from synaptic contacts.

Functions of PHR Proteins

Compared to many newly identified protein families, our understanding of the signal transduction pathways regulated by, or interacting with, the PHR protein has progressed remarkably rapidly. PHR proteins have been shown to function as an E3 ubiquitin ligase to mediate protein degradation. They

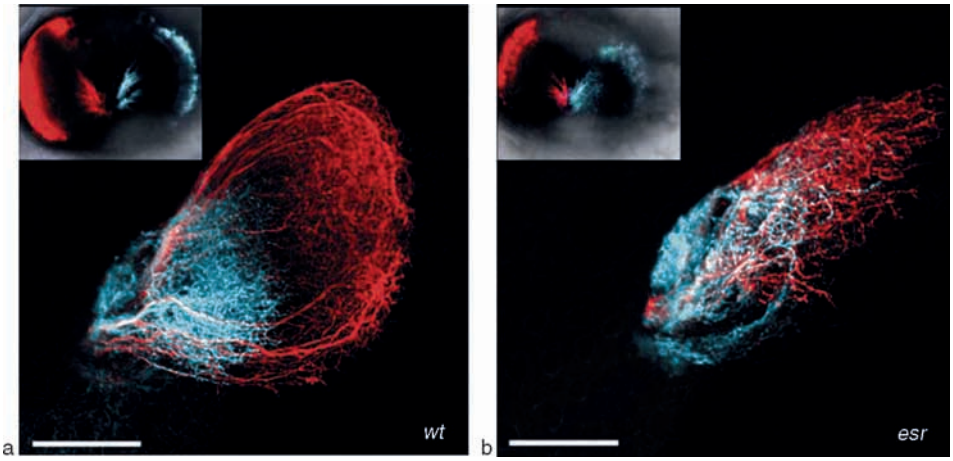

Figure 4 Retinal axon defects in zebra fish *esrom* mutants: (a) wild-type retina; (b) *esrom* mutant retina. Shown are optic tectum of 4-day-old zebra fish, with retinal axons labeled with DiI (red) and DiD (blue). In the wild-type retina (a), axons from the anterior region of the eye project to the posterior tectum, and axons from the posterior arborize in the anterior tectum. In the *esrom* mutant retina (b), anterior axons defasciculate and arborize anteriorly. *esr*, *esrom* mutant; *wt*, wild type. From D'Souza J, Hendricks M, Le Guyader S, et al. (2005) Formation of the retinotectal projection requires Esrom, an ortholog of PAM (protein associated with Myc). *Development* 132: 247–256.

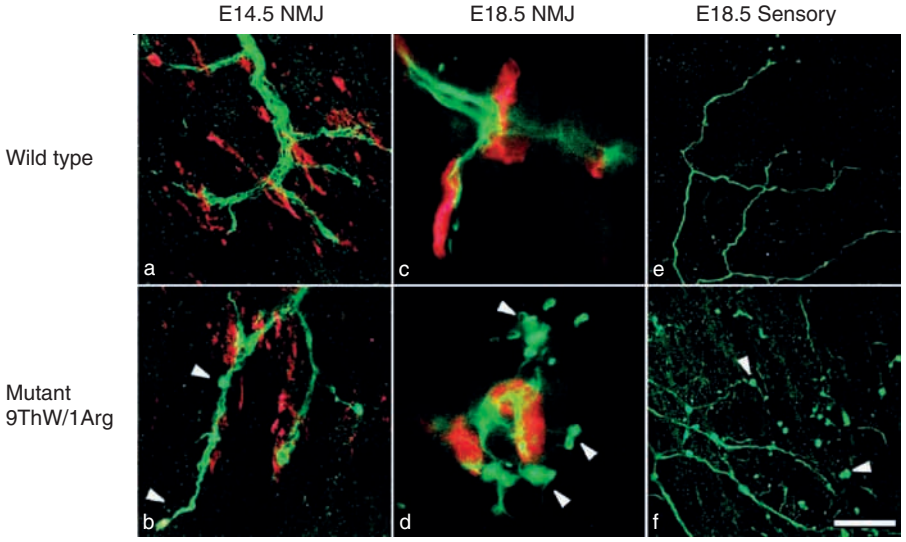

Figure 5 NMJ and sensory neuron defects in candidate Phr1 mutant mice: (a) NMJs from the phrenic nerve in wild-type E14.5 embryo; (b) NMJs from the phrenic nerve in 9ThW/1Arg mutant E14.5 embryo; (c) NMJs from the phrenic nerve in wild-type E18.5 embryo; (d) NMJs from the phrenic nerve in 9ThW/1Arg mutant E18.5 embryo; (e) sensory axon terminals in wild-type E18.5 embryo; (f) sensory axon terminals in 9ThW/1Arg mutant E18.5 embryo. At E14.5, in wild types (a), NMJs have terminals (green) overlying AChR-rich plaques (red) on the muscle fibers; in mutants (b), some terminals overgrow beyond the receptor plaques (lower arrowhead) and some have varicose morphology (upper arrowhead). At E18.5, in wild types (c), NMJs are more robust and the receptor plaques are more condensed; in mutants (d), NMJs show similar maturation, but the nerve terminals exhibit sprouting beyond receptor plaques. At E18.5, the mutants (f) show terminal sprouting and the terminal axons have numerous varicosities (arrowhead) and are either more elaborately branched or less fasiculated. Scale bar = 40 μm (a, b, e, f), 11 μm (c, d). AChR, acetylcholine receptor; E, embryonic day; NMJ, neuromuscular junction. From Burgess RW, Peterson KA, Johnson MJ, Roix JJ, Welsh IC, and O'Brien TP (2004) Evidence for a conserved function in synapse formation reveals Phr1 as a candidate gene for respiratory failure in newborn mice. *Molecular and Cell Biology* 24: 1096–1105.

negatively regulate the dual leucine zipper MAPK kinase kinase, a BMP pathway, the tuberin–hamartin complex, and adenylate cyclase.

The presence of a RING domain at the C-terminus predicts that the PHR protein may have an ubiquitin ligase activity. Several pieces of evidence support this role. First, multiple naturally occurring loss-of-function mutations disrupt or eliminate the RING domain in *rpm-1*, *hiw*, and *esrom*. A transgene expressing a Hiw protein in which the key Cys residues in the RING domain are mutated causes a dominant negative effect. Second, Hiw mutant phenotypes

can be partially suppressed by overexpression of a de-ubiquitin enzyme *fat facet*. Third, an F-box protein, FSN-1, binds RPM-1. *fsn-1* mutants exhibits a range of synapse defects similar to that in *rpm-1*. RPM-1 and FSN-1 are also present in a protein complex with Cullin and Skp, suggesting that these proteins may function in a manner similar to Skp/Cul/F-box type E3. Fourth, the RING domain of Esrom is able to catalyze ubiquitination *in vitro*. Moreover, the RING domain of RPM-1 can stimulate ubiquitination of its substrate, DLK-1.

What are the targets for the PHR E3 ligase activity? The answers to this question have come from powerful forward genetic screens in *C. elegans* and *Drosophila*. One way to interpret the defects in *rpm-1* or *hiw* mutants is that they could arise from prolonged actions of targets as the result of failure in protein degradation. Then, mutations eliminating, or impairing the action of, the substrates could restore normal synaptic pattern in *rpm-1* or *hiw* mutants. Screening for such suppressor mutations in *C. elegans* and *Drosophila* led to the identification of a dual leucine zipper MAPK (DLK), DLK-1 and Wallenda, respectively. Loss-of-function mutations in either DLK alone do not cause detectable defects but strongly suppress the synaptic defects in *rpm-1* and the synaptic overgrowth defects in *hiw*. Consistent with being a target of E3 ligase activity, the elevated expression of DLK-1 or Wallenda in otherwise wild-type background causes synaptic defects resembling those in *rpm-1* or *hiw*, respectively. In *rpm-1* or *wallenda* mutants, the levels of DLK-1 or Wallenda at synapses are upregulated. Moreover, biochemical studies in cell culture conditions show that RPM-1 can bind DLK-1 and that the RING domain of RPM-1 can stimulate DLK-1 ubiquitination. Together, these studies establish an evolutionarily conserved mechanism by which RPM-1 and Hiw act as an E3 ligase to downregulate the DLK MAPK at the synapses.

Interestingly, the MAPKs downstream of DLK-1 and Wallenda appear to have diverged in *C. elegans* and *Drosophila*. In *C. elegans*, loss-of-function mutations in the MKK-4 MAPKK and a p38 MAPK PMK-3 were isolated as suppressors of *rpm-1*. Genetic studies indicate that these two kinases act downstream of DLK-1, with PMK-3 as the downstream MAPK. On the other hand, in *Drosophila*, a c-Jun N-terminal kinase (JNK) homolog may act as the downstream MAPK, but it is not yet known which MAPKK acts downstream of Wallenda. Both *C. elegans* PMK-3 and *Drosophila* JNK MAPKs are localized in nuclei, suggesting that the interaction between RPM-1/Hiw and DLK-1/Wallenda at synapses is transduced to the nucleus. Supporting this idea, the expression of a dominant negative Fos, a known target of JNK, in Hiw mutants partially suppresses the synaptic overgrowth. Thus, the synaptic defects in *rpm-1* or *hiw* are, in part, due to improper gene expression at the transcriptional level.

Three other pathways have been implicated in PHR protein functions. Screening for Hiw binding partners by a yeast two hybrid assay identified the Smad protein Medea (Med). Med acts downstream of a presynaptic BMP signaling pathway. Loss of the BMP signaling in motor neurons causes terminal undergrowth such that the NMJ size is small and neurotransmitter release is reduced. Activation of BMP signaling in *hiw* mutants leads to further synaptic expansion, which implies a negative regulation of the BMP pathway by Hiw. However, mutants lacking both *hiw* and BMP signaling exhibit complex phenotypes that are also consistent with the interpretation that Hiw and BMP pathways function in parallel. In fact, phospho-MAD, a major readout of BMP signaling, is not upregulated in *hiw* mutants, and the neuronal expression of a constitutive active BMP receptor that leads to increased phospho-MAD does not lead to synaptic overgrowth. It is not known whether the binding of Med by Hiw leads to degradation of Med. Thus, it remains to be resolved how Hiw interacts with the BMP and Med pathway.

The C-terminal of Pam has been shown to bind TSC2/hamarin. TSC1/tuberin and TSC2 can co-immunoprecipitate with Pam from rat brain protein extracts. In primary cultured cortical neurons, Pam and TSC1/2 show co-localization along the neurites and in growth cones, although in cell bodies Pam is found in the nuclei and TSC is seen at perinuclear locations. A functional test of this interaction was conducted in the *Drosophila* eye. Overexpression of *Drosophila* TSC1 or TSC2 causes small eyes, which were enhanced in *hiw* mutants, suggesting that Hiw may negatively regulate the TSC complex. It remains to be addressed whether this regulation occurs through the E3 ligase activity or other functions of Hiw.

At present, the only known binding partner for the N-terminal of PHR proteins is adenylate cyclase V (AC V), which binds the RHD domain of Pam. Both full-length Pam and RHD domain alone can potently inhibit AC V, AC VI, and AC I. In serum-starved HeLa cells, Pam appears to reside in the endoplasmic reticulum. On treatment with sphingosine-1-phosphate (S1P), Pam translocates to the plasma membrane and inhibits AC activity. S1P is a natural ligand for G-protein-coupled receptors known as endothelial differentiation genes (EDGs). S1P-induced translocation and activation of Pam can be inhibited by pertussis toxin, implying the involvement of G_i. Pam translocation may also require protein kinase C activation and phospholipase C. The functional significance of

the interaction of Pam with AC in neurons was tested in a spinal nociceptive assay. In rat spinal cord and DRGs, Pam expression is upregulated following peripheral nociceptive stimulation by zymosan or formalin. Animals treated with an infusion of Pam antisense oligonucleotides to decrease the endogenous Pam expression exhibited an increase of nociceptive behavior following formalin injection.

Although Pam was originally identified by its binding to the transcriptional activation domain of c-myc, little is known about the role of Pam in nonneuronal cells. A microarray analysis of genes differentially expressed in quiescence and senescence cells showed that Pam is downregulated in both young quiescent and senescent cells. Until an appropriate animal model is developed, the precise roles of Pam and its mammalian counterparts remain unclear.

Conclusion and Perspectives

Within 8 years since the identification of Pam, the first member of the PHR protein family, our understanding of the functions of these conserved proteins has progressed rapidly. The genetic studies of the invertebrate members have identified important players interacting with PHR proteins. Strong evidence supports a definite function of these proteins in ubiquitin-mediated protein degradation at the synapse. Yet many questions remain. For example, the presence of RPM-1 or Hiw at synapses is quintessential for the normal regulation of synapse formation and growth. However, we know nothing about how these proteins get to synapses in the first place and how they change the micro-environment to facilitate synapse formation. The seemingly diverse effects associated with each gene in different neurons in different organisms are particularly puzzling. It is possible that PHR proteins endow synapses with morphological and functional diversity through targeting diverse sets of substrates in different neurons. Or they may act on the same signaling pathways; yet, due to other intrinsic properties, specific types of neurons respond to a common signaling in a different manner. It is clear that downregulation of DLK MAPK cascade is crucial for normal synapse formation in *C. elegans* and *Drosophila*. However, little is known about the signals activating DLK kinase; such signals could include positional cues or synaptic target recognition information. Because loss of function in the DLK pathway does not lead to overt synaptic defects in *C. elegans* or *Drosophila*, more work is needed both to define the normal roles of the DLK pathway and to understand why its inhibition by PHR proteins is so critical for normal synaptogenesis. Moreover, not all the mutant phenotypes of *hiw* or *rpm-1* are suppressed by inactivation of the DLK cascade, implying the presence of unidentified pathways. Given the giant size and multiple protein interaction motifs of PHR proteins, more signaling pathways will certainly be found to interact with these proteins. Last, but not the least, the generation of critical animal models is necessary for defining the function of mammalian PHR proteins. Understanding the synapse-specific roles of PHR proteins may hold the keys to our understanding of the sophisticated tasks that different neurons execute, in light of a uniformity of synapse structures.

See also: Comparative Biology of Invertebrate Neuromuscular Junctions; Development of Drosophila Neuromuscular Junctions; Learning and Memory in Invertebrates: *C. Elegans*; Learning and Memory in Invertebrates: *Drosophila*; Neuromuscular Junction: Neuronal Regulation of Gene Transcription at the Vertebrate; Neuromuscular Junction: Synapse Elimination; Synaptic Vesicles.

Further Reading

Burgess RW, Peterson KA, Johnson MJ, Roix JJ, Welsh IC, and O'Brien TP (2004) Evidence for a conserved function in synapse formation reveals Phr1 as a candidate gene for respiratory failure in newborn mice. *Molecular and Cell Biology* 24: 1096–1105.

Collins CA, Wairkar YP, Johnson SL, and DiAntonio A (2006) Highwire restrains synaptic growth by attenuating a MAP kinase signal. *Neuron* 51: 57–69.

D'Souza J, Hendricks M, Le Guyader S, et al. (2005) Formation of the retinotectal projection requires Esrom, an ortholog of PAM (protein associated with Myc). *Development* 132: 247–256.

Ehnert C, Tegeder I, Pierre S, et al. (2004) Protein associated with Myc (PAM) is involved in spinal nociceptive processing. *Journal of Neurochemistry* 88: 948–957.

Guo Q, Xie J, Dang CV, Liu ET, and Bishop JM (1998) Identification of a large Myc-binding protein that contains RCC1-like repeats. *Proceedings of the National Academy of Sciences of the United States of America* 95: 9172–9177.

Liao EH, Hung W, Abrams B, and Zhen M (2004) An SCF-like ubiquitin ligase complex that controls presynaptic differentiation. *Nature* 430: 345–350.

McCabe BD, Hom S, Aberle H, et al. (2004) Highwire regulates presynaptic BMP signaling essential for synaptic growth. *Neuron* 41: 891–905.

Murthy V, Han S, Beauchamp RL, et al. (2004) Pam and its ortholog highwire interact with and may negatively regulate the TSC1.TSC2 complex. *Journal of Biological Chemistry* 279: 1351–1358.

Nakata K, Abrams B, Grill B, et al. (2005) Regulation of a DLK-1 and p38 MAP kinase pathway by the ubiquitin ligase RPM-1 is required for presynaptic development. *Cell* 120: 407–420.

Pierre SC, Hausler J, Birod K, Geisslinger G, and Scholich K (2004) PAM mediates sustained inhibition of cAMP signaling by sphingosine-1-phosphate. *EMBO Journal* 23: 3031–3040.

Santos TM, Han S, Bowser M, et al. (2006) Alternative splicing in protein associated with Myc (Pam) influences its binding to c-Myc. *Journal of Neuroscience Research* 83: 222–232.

Schaefer AM, Hadwiger GD, and Nonet ML (2000) rpm-1, a conserved neuronal gene that regulates targeting and synaptogenesis in *C. elegans*. *Neuron* 26: 345–356.

Scholich K, Pierre S, and Patel TB (2001) Protein associated with Myc (PAM) is a potent inhibitor of adenylyl cyclases. *Journal of Biological Chemistry* 276: 47583–47589.

Sone M, Suzuki E, Hoshino M, et al. (2000) Synaptic development is controlled in the periactive zones of *Drosophila* synapses. *Development* 127: 4157–4168.

Wan HI, DiAntonio A, Fetter RD, Bergstrom K, Strauss R, and Goodman CS (2000) Highwire regulates synaptic growth in *Drosophila*. *Neuron* 26: 313–329.

Wang XF, Wang D, Zhu W, Delrahim KK, Dolnak D, and Rapaport MH (2003) Studies characterizing 60 kda autoantibodies in subjects with schizophrenia. *Biological Psychiatry* 53: 361–375.

Wu C, Wairkar YP, Collins CA, and DiAntonio A (2005) Highwire function at the Drosophila neuromuscular junction: Spatial, structural, and temporal requirements. *Journal of Neuroscience* 25: 9557–9566.

Yang H, Scholich K, Poser S, Storm DR, Patel TB, and Goldowitz D (2002) Developmental expression of PAM (protein associated with MYC) in the rodent brain. *Brain Research Developmental Brain Research* 136: 35–42.

Zhen M, Huang X, Bamber B, and Jin Y (2000) Regulation of presynaptic terminal organization by C. elegans RPM-1, a putative guanine nucleotide exchanger with a RING-H2 finger domain. *Neuron* 26: 331–343.

Relevant Website

http://www.treefam.org – Tree Families database.

Phrenology

C G Gross, Princeton University, Princeton, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Phrenology is the doctrine that intellectual abilities and personality traits are localized in discrete regions of the cerebral cortex and the magnitude of these abilities and traits is reflected in the overlying skull – in bumps on the head. Although viewed as quackery today, phrenology was a significant medical and popular movement in the first half of the nineteenth century. Furthermore, it had a major effect on the development of modern neuroscience.

In this article, we consider the founders of phrenology, Franz Joseph Gall (1758–1828) and his associate JC Spurzheim (1776–1852); the basic tenets of phrenology and their sources; the impact of phrenology in the lay and scientific communities; and, finally, the role of phrenology in the development of modern neuroscience.

Gall and Spurzheim

Gall was born in Germany and had a Catholic upbringing, although eventually his works were placed on the Index and he was denied a Catholic burial. He received his MD from the University of Vienna and had a successful practice with many famous patients in Vienna and Paris. He wrote and lectured extensively but never held an academic post. In fact, he often had trouble with civic, church, and medical authorities and was accused of atheism, materialism, and immorality. His lectures provided a high income and he does seem to have had a somewhat flamboyant lifestyle. Spurzheim started collaborating with Gall as a medical student in Vienna in 1800 and continued to do so for another 13 years. For the rest of his life, Spurzheim lectured and wrote on phrenology and never practiced conventional medicine.

Gall and Spurzheim made a number of significant anatomical advances that were essentially independent of their phrenological system. They viewed the nervous system, including the brain, the spinal cord, and peripheral nerves, as a hierarchy of interconnected ganglia. This was in contrast to the dominant view since Galen in which the brain was central and the spinal cord and nerves were appendages. They were the first to trace by blunt dissection sensory and motor pathways. The previous dissection technique from Vesalius onward was to make horizontal sections from rostral to caudal. This would reveal the ventricular structure thought to be so important but provide little clue as to the hierarchical organization of the nervous system or the course of neural input or output. The more empirical findings of Gall and Spurzheim included distinguishing cortical gray and white matter; differentiating projection, association, and commissural fibers; tracing the origins of most of the cranial nerves; the first description of postembryonic myelinization; establishing the pyramidal decussation; and the realization that the human brain is convoluted to save space. They produced the first accurate illustrations of the cerebral cortex.

Phrenology: Central Tenets and Methods

Gall and Spurzheim viewed the brain as an elaborately wired machine for producing behavior, thought, and emotion and the cerebral cortex as a set of organs for carrying out these functions. Their phrenological system attempted to relate these psychological functions to the organs of the cerebral cortex – to relate brain and behavior. Phrenology was based on four assumptions: (1) intellectual abilities and personality traits are differentially developed in each individual, (2) these abilities and traits reflect faculties that are localized in specific organs of the cerebral cortex, (3) the development or prominence of these faculties is a function of the activity and therefore the size of the cortical organs, and (4) the size of each cortical organ is reflected in the prominence of the overlying skull (i.e., in cranial bumps). Although they viewed the faculties as 'innate,' they thought they could be modified by experience, especially by education.

These otherwise reasonable hypotheses had one principal fatal flaw: the nature of the evidence. Gall and Spurzheim relied almost entirely on obtaining supportive or confirmatory evidence. They collected large numbers of skulls of people whose traits and abilities were known, examined the heads of distinguished savants and inhabitants of mental hospitals and prisons, and studied portraits of the high and low on various intellectual and affective dimensions. Throughout, they were seeking confirmation of their initial hypothesis usually deriving from a few cases. For example, the idea for a language organ in the frontal lobes come from Gall's experience as a school boy of a classmate who had a prodigious verbal memory and protruding eyes (being pushed out by a well-developed frontal lobe, Gall thought). The idea for an organ of destructiveness (or carnivorous instinct) came from the skulls of a parricide and of a

murderer that were sent to him, from noticing its prominence in a fellow medical student who "was so fond of torturing animals that he became a surgeon," and from examining the head of a meat-loving dog he owned. All their methods were used to seek confirmations; contradictions were explained away.

Gall and Spurzheim gave cogent reasons why the use of experimental lesions in animals (or 'mutilations' as they called them) or the study of brain-injured humans was inadequate, by them, to test their ideas. The state of anatomical knowledge was insufficient to remove a single organ. The organs were interconnected so that removal of one could alter the function of another. Animals would often live only a few days after surgery with the available methods. Generalizing from animals to humans was hazardous. Previous studies had not examined complex cognitive or emotional function in animals but only such things as 'sensibility' or irritability. Needless to say, however, when a clinical case that supported their view was brought to them, they would cite it as further confirmatory evidence.

Ideas on Brain Function before Phrenology

One source of Gall and Spurzheim's ideas was the Swiss anatomist Charles Bonnet (1720–93), who suggested that the brain might be an "assemblage of different organs." Another was the German philosopher JG Herder (1744–1803), who argued that every function must have a bodily organ responsible for it. At this time, physiognomy, the assessment of personality and character by examination of the features of the face, was a widespread practice and in a sense Gall and Spurzheim were transferring this to the top of the head.

Gall and Spurzheim's assigning of various psychological functions to organs of the cerebral cortex was variance with the prevailing view, deriving from Aristotle (384–22 BC), of the unity of the mind. There were a few intimations of punctate localization of psychological function in the brain before Gall and Spurzheim, such as by Erasistratus of Alexandria (*c*. 280 BC), Galen of Pergamon (129–200), Thomas Willis (1621–75), Pourfour du Petit (1674–1741), and Emanuel Swedenborg (1688–1772). However, these views were unknown or totally ignored in the beginning of the nineteenth century. Rather, the prevailing view was that of Albrecht von Haller (1708–77), who dominated physiology through his monumental textbooks and many followers. He concluded that all parts of the cortex were equivalent because he had found that mechanical and chemical stimulation of the cortex had the same null effect everywhere. He thought that "the greater part of the cortex consists of mere vessels."

Before Gall, the cerebral cortex was usually pictured in a uniform fashion, often looking like intestines, a comparison originally made by Erasistratus. By contrast, the illustrations of cortex by Gall and Spurzheim are strikingly accurate (**Figure 1**).

The only previous suggestion that the cerebral cortex was not uniform is structure was by Francisco Gennari (1752–97). As a medical student, he noticed and reported that a white line in the cortex was especially prominent and sometimes double in the posterior part of the cortex.

Gall and Spurzheim Go Separate Ways

The term 'phrenology' was never used or approved by Gall. Originally, he used the term 'craniology' then 'organology' and, finally, 'functions of the brain' or 'brain physiology' to describe his theory of punctate localization of psychological function in the brain. In 1813, Spurzheim separated from Gall and moved to England, where he substituted the term phrenology which seems to have been of American or British origin. There were a number of differences between Gall's organology and Spurzheim's phrenology, which continued to develop after they separated. Gall was much more tentative in naming functions, often giving several terms and long description and a more approximate location, whereas Spurzheim was satisfied with a single term and a 'precise' location. Gall had 27 organs (19 or which were found in animals), whereas Spurzheim extended them to 32–37 with sharper boundaries (cf. **Figures 1 and 2**); later phrenologists, such as Combe, added more. Gall continued to emphasize empirical observation, whereas Spurzheim became increasingly speculative and used phrenology as a way of predicting future behavior. Furthermore, Spurzheim was much more interested in applying the phrenology to social reform and to the treatment of mental disease. The term phrenology was generally applied to both sets of ideas.

Phrenology met with considerable opposition from political and religious authorities, particularly on the European continent, largely because it was viewed as implying materialism and determinism and denying the unity of the mind (and soul) and the existence of free will. On the other hand, phrenology spread widely, particularly in the United States and Great Britain, both as a medical doctrine and as a 'pop' psychology. As a medical doctrine, it influenced the diagnosis and treatment of mental and neurological disease. It generated widespread interest both among the general populace and among such writers and

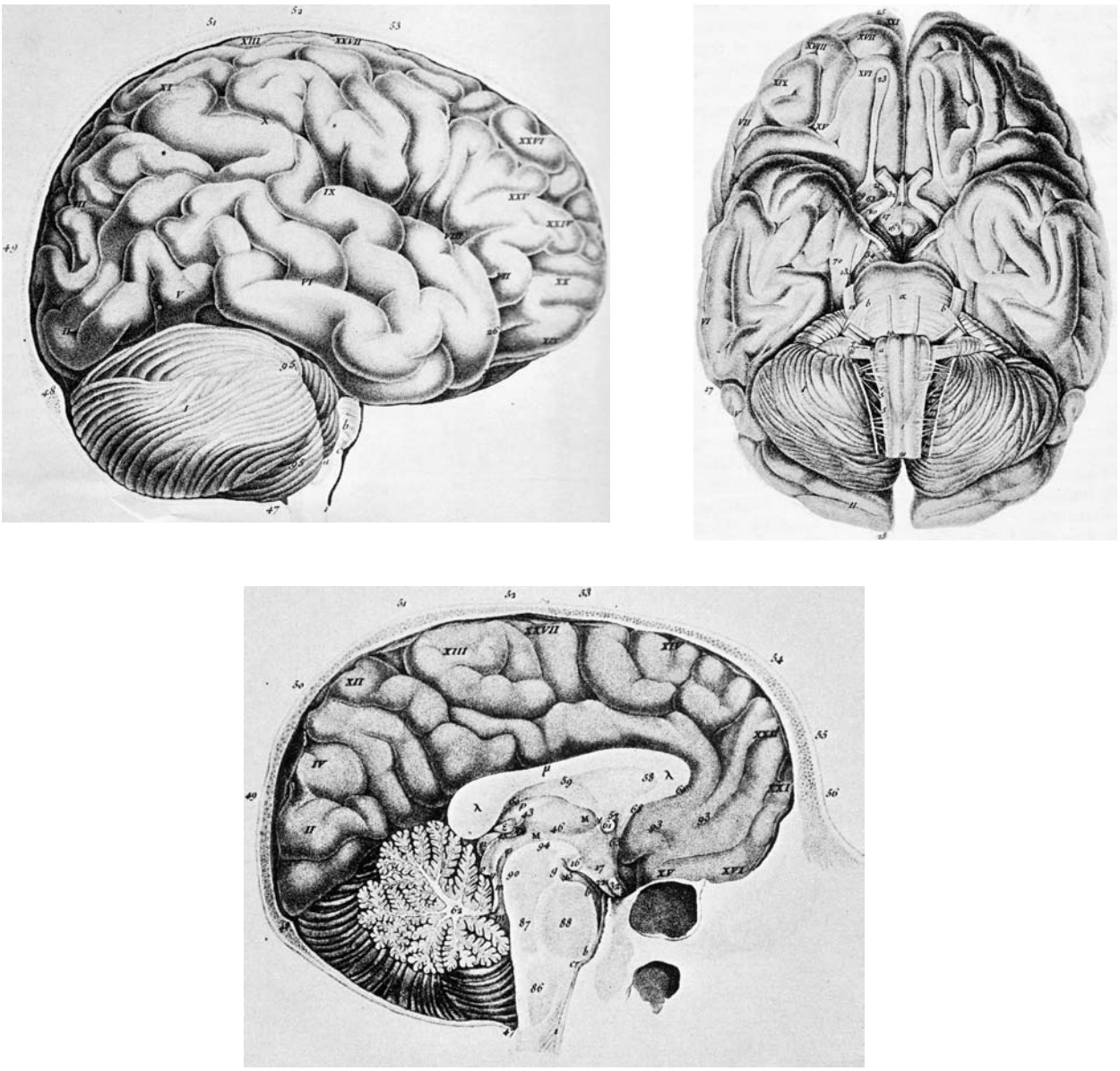

Figure 1 The Roman numbers refer to the organ's qualities as follows. Organs common to men and animals: I, instinct of reproduction; II, love of offspring; III, friendship; IV, self-defense and courage; V, carnivorous instinct, tendency to murder; VI, cunning, cleverness; VII, ownership, covetousness, tendency to steal, love of authority; VIII, pride, arrogance, haughtiness; IX, vanity, ambition, love of glory; X, caution, forethought; XI, memory of things and facts, educability; XII, sense of places and spaces; XIII, memory and sense of people; XIV, memory of words; XV, sense of language and speech; XVI, sense of color; XVII, sense of sound, music; XVIII, sense of numbers, mathematics; XIX, sense of mechanics. Organs in man alone: XX, wisdom; XXI, sense of metaphysics; XXII, satire, witticism; XXIII, poetical talent; XXIV, kindness, compassion, morality; XXV, mimicry; XXVI, religion; XXVII, firmness of purpose, obstinacy, constancy. These accurate and aesthetic drawings of the cerebral cortex are from volume 1 of Gall FJ and Spurzheim JC, *Anatomie et Physiologie du Systême Nerveux en Général, et du Cerveau en Particulier, Avec des Observations sur la Possibilité de Reconnaître Plusieurs Dispositions Intellectuelles et Morales de l'homme et des Animaux, par la Configuration de Leurs Têtes,* 4 vols. with atlas of 100 engraved plates (Schoell, Paris, 1810–19) (Gall is sole author of vols. 3 and 4).

savants as Honore de Balzac, Charles Baudelaire, George Eliot, August Comte, Horace Mann, Alfred Russell Wallace, and George Henry Lewis. In fact, it rapidly became a popular fad and drawing room amusement, particularly in Great Britain and the United States. Phrenological societies and journals continued to flourish in both countries well into the twentieth century.

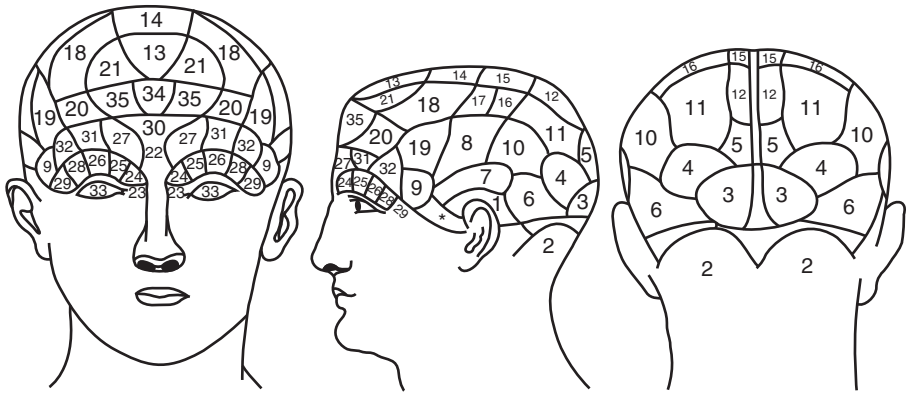

Affective faculties		Intellectual faculties	
Propensities	Sentiments	Perceptive	Reflective
Desire to live	10 Cautiousness	22 Individuality	34 Comparison
* Alimentiveness	11 Approbativeness	23 Configuration	35 Causality
1 Destructiveness	12 Self-esteem	24 Size	
2 Amativeness	13 Benevolence	25 Weight and resistance	
3 Philoprogenitiveness	14 Reverence	26 Coloring	
4 Adhesiveness	15 Firmness	27 Locality	
5 Inhabitiveness	16 Conscientiousness	28 Order	
6 Combativeness	17 Hope	29 Calculation	
7 Secretiveness	18 Marvelousness	30 Eventuality	
8 Acquisitiveness	19 Ideality	31 Time	
9 Constructiveness	20 Mirthfulness	32 Tune	
	21 Imitation	33 Language	

Figure 2 In comparison to Figure 1, note that the locations of the faculties are more specific, as are their names. Phrenological head is from Spurzheim JG (1834) *Phrenology or the Doctrine of the Mental Phenomenon*. Boston: Marsh, Capen & Lyon.

Flourens Attacks Gall

In the scientific world, the most important and influential critique of Gall came from Pierre Flourens (1794–1867). A technically brilliant experimenter, Flourens quickly rose in the French scientific establishment and at the age of 35 was elected to the French Academy of Science, which had rejected Gall. Starting in the 1820s and continuing for more than 20 years, Flourens carried out a series of experiments on the behavioral effects of brain lesions, particularly with pigeons. Flourens reported that lesions of the cerebrum in pigeons had devastating effects on willing, judging, remembering, and perceiving. However, he found that the site of the lesion was irrelevant: all regions of the cerebrum contributed to these functions. The only exception was vision in that a unilateral lesion produced only contralateral blindness, but again there was no localization within the hemisphere. These holistic results tended to eclipse Gall's ideas of punctate localization, but only in scientific circles and only temporarily.

Flourens' finding of cognitive losses after cerebral lesions was actually a confirmation of Gall's emphasis on the cognitive role of the cortex, a viewpoint that had been virtually absent before Gall. The cortex was termed a 'superadded' structure lying hierarchically and physically above the highest sensory structure, the thalamus, and the highest motor structure, the corpus striatum. The general idea that the thalamus had major sensory functions and the corpus striatum major motor functions was generally accepted by the mid-nineteenth century on the basis of a number of studies involving tracing sensory and motor tracts from the periphery and making experimental lesions in animals.

This view of the 'higher' functions of the cortex, common for the period, combines Haller's view of the insensitivity of cortex and both Gall's and Flourens' attribution of 'higher faculties' but neither sensory nor motor functions to the cortex.

Broca 'Confirms' Gall

Despite the bitter attacks of Flourens, Gall's idea of punctate localization and even many of his specific localizations, such as language in the frontal lobes and sexuality in the cerebellum, continued to be actively debated in the mid-nineteenth century.

At least in the scientific community, the supposed correlations between skull and brain morphology were quickly recognized as erroneous. Yet, Gall's ideas stimulated the search for correlations between the site of brain injury and specific psychological deficits in patients as well as in experimental animals. Reports of such correlations were published in both the phrenological and the mainstream neurological literature, and the question of the localization of psychological function in the brain was hotly debated at scientific meetings.

Thus, in 1848, JB Bouillard (1796–1881), a powerful figure in the medical establishment and a supporter of Gall, offered a cash prize for a patient with major frontal lobe damage who did not have a language deficit. The debate about localization reached a climax at a series of meetings of the Paris Societe d'Anthropologie in 1861. At the April meeting, Paul Broca (1824–80), a founder of the society, announced that he had a critical case on this issue. A patient with long-standing language difficulties, nicknamed 'Tan' because that was all that he could say, had just died. The next day, Broca displayed his brain at the meeting and indeed it had widespread damage in the left frontal lobe (actually much more widespread than today's 'Broca's area'). He presented several similar cases in the next few months. Not only did Broca's cases finally establish the principle of discrete localization of psychological function in the brain but also the discovery was hailed as a vindication of Gall. Broca himself had regarded Gall's work as "the starting point for every discovery in cerebral physiology in our century."

Conclusions

Despite its absurdities and excesses, phrenology facilitated the development of the study of the brain and behavior in several ways: by stressing that the human mind could be subdivided into specific functions and that specific brain mechanisms underlay specific mental abilities and traits, by emphasizing the importance of the cerebral cortex in mental activity, and by stimulating a great surge of research on the psychological effects of human brain damage and of experimental lesions in animals. After Gall, less radical divisions of brain function, such as those of Flourons, were much more readily accepted. The search for organs in the cerebral cortex led directly to cytoarchitectonics and myeloarchitectonics, attempts to distinguish cortical areas on the basis of structure, and to the tracing of sensory pathways to specific cortical areas. The phrenologists also influenced the development of physical anthropology: they were the first to develop methods and instruments for measuring the cranium.

The cytoarchitectonic, positron emission tomography, functional magnetic resonance imaging, and other imaging maps of the cerebral cortex that are now ubiquitous in neuroscience textbooks bear more than a coincidental resemblance to phrenological charts. They are the direct descendants of the ambitious, albeit heavily flawed, program of phrenology to relate brain structure and behavior.

See also: Brain Damage: Functional Reorganization; Broca's Area: Evolution; Cognitive Neuroscience: An Overview; History of Neuroscience: Early Neuroscience; Neurophysiology: Past and Present.

Further Reading

Ackerknecht E (1958) Contributions of Gall and the phrenologists to knowledge of brain function. In: Poynter FNL (ed.) *The Brain and Its Functions*. Oxford: Blackwell.

Brazier MAB (1988) *A History of Neurophysiology in the 19th Century*. New York: Raven Press.

Clarke E and Jacyna LS (1987) *Nineteenth-Century Origins of Neuroscientific Concepts*. Berkeley, CA: University of California Press.

Cooter RJ (1985) *The Cultural Meaning of Popular Science. Phrenology and the Organization of Consent in Nineteenth Century Britain*. Cambridge, UK: Cambridge University Press.

Gall FJ and Spurzheim JC (1810–1819) Anatomie et Physiologie du Systême Nerveux en général, et du Cerveau en Particulier, Avec des Observations sur la Possibilité de Reconnaître Plusieurs Dispositions Intellactuelles Morales de l'homme et des Animavx, par la Configuration de leurs Têtes, 4 vols. Schoell: Paris.

Gall FJ and Spurzheim JG (1835) *On the Functions of the Brain and Each of Its Parts: With Observations on the Possibility of Determining the Instincts, Propensities and Talents, or the Moral and Intellectual Dispositions of Men and Animals, by the Configuration of the Brain and Head*, 6 vols. Boston: Marsh, Capen & Lyon.

Gross CG (1998) *Brain, Vision, Memory: Tales in the History of Neuroscience*. Cambridge: MIT Press.

Pogliano C (1991) Between form and function: A new science of man. In: Corsi P (ed.) *The Enchanted Loom*. Oxford: Oxford University Press.

Spurzheim JG (1834) *Phrenology or the Doctrine of the Mental Phenomenon*. Boston: Marsh, Capen & Lyon.

Temkin O (1947) Gall and the phrenological movement. *Bulletin of the History of Medicine* 21: 275–321.

Van Wyhe J (2006) *The History of Phrenology on the Web*. http://pages.britishlibrary.net (accesed in Sep. 2008).

Walsh AA (2006) *When Heads Were Headlines*. http://www.struweb.com (accesed in Sep. 2008).

Young RM (1970) Gall, Franz Joseph. In: Gillespie CC (ed.) *Dictionary of Scientific Biography*. New York: Scribner.

Young RM (1970) *Mind, Brain and Adaptation in the Nineteenth Century*. Oxford: Clarendon.

Zola-Morgan S (1995) Localization of brain function: The legacy of Franz Joseph Gall (1758–1828). *Annual Review of Neuroscience* 18: 359–383.

Piccolo and Bassoon

J Juranek, Columbia University Medical Center, New York, NY, USA
K Mukherjee, UT Southwestern Medical Center at Dallas, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons communicate with one another and with their target cells via specialized cell–cell contact sites called synapses. The term synapse was coined in 1906 by Sherrington, and it referred only to the well-distinguished axo-dendritic junctions in central nervous system. In current terminology, synapse implies many kinds of neuronal contacts such as gap junctions (electrical synapses), end plates, and varicose axon terminals of the central nervous system (chemical synapses).

Communication at chemical synapse requires the transformation of electrical activity, the propagating action potential, into chemical signals, the neurotransmitters released from the presynapse at specialized sites. At the ultrastructural level, these presynaptic specializations are characterized by electron-dense projections referred to as active zone; a term coined by Couteaux and Pecot-Dechavassine in 1971. Active zones are formed by a large number of proteins – including cytoskeletal proteins, scaffolding proteins, voltage-gated Ca^{2+} channels and cell adhesion molecules.

Recently, active-zone-specific proteins belonging to five different families have been identified and termed CAZ (cytomatrix of active zone) proteins. These include Munc13s/Unc13 (1995), Rims (1996), Piccolo/Aczonin (1996), Bassoon (1998), Liprin αs/Syd2 (1998), and ELKS/ERC/CAST1 (2002). These proteins are thought to play an essential role in neurotransmission by defining release sites, anchoring Ca^{2+} channels, and coordinating the exo-endocytotic cycle of synaptic vesicles in the presynaptic region.

Domain Structure and Molecular Interactions of Piccolo and Bassoon

Piccolo and Bassoon were isolated in the second half of the 1990s in an expression library screen of rat brain cDNA using antibodies raised against synaptic junctional protein preparations. These are unusually large (Piccolo, 530 kDa; Bassoon, 420 kDa), structurally related, multidomain proteins that are considered central to the assembly and functional organization of the active zone in vertebrates. Unlike other CAZ proteins, these two molecules do not have any orthologs in either *Caenorhabditis elegans* or *Drosophila melanogaster*. However, several proteins containing similar domains have been found in these animals, and it is plausible that they perform similar functions to Piccolo and Bassoon in invertebrates.

The proteins share ten regions of high sequence similarity, called Piccolo Bassoon homology (PBH) domains. The PBH1 and PBH2 domains contain two double zinc-finger motifs, which have low homology with those of Rab3 interacting molecules (RIMs) and Rabphilin-3A. The PBH4, -6, and -8 domains contain coil-coiled (CC) motifs. The remaining PBH domains seem to be unique to these two proteins because they do not share homology with any known proteins. In addition, Piccolo contains one PDZ (PSD, Dlg, ZO-1) domain, two C2 domains (C2A and C2B) at the C-terminus and one glutamine heptad (Q) domain at the N-terminus. Bassoon contains one Q domain at C-terminus and is posttranslationally modified by myristoylation at its N-terminus. Due to sequence similarity, it has been suggested that Piccolo might be acylated at the N-terminus as well (**Figure 1**).

Piccolo and Bassoon, through their multiple domains, interact with various partners at the presynapse, making them ideal scaffolding molecules. The zinc finger domains of Piccolo and Bassoon are known to interact with double-prenylated rab 3A and synaptobrevin receptor PRA1. It has been speculated that such interactions *in vivo* might be important for trafficking of synaptic vesicles at the active zone. Both Piccolo and Bassoon, with their third CC motif, interact with another CAZ protein ELKS, and with two members of C-terminal binding protein (CtBP) family: CtBP1 and CtBP2. In the ribbon synapses of the mammalian retina, Bassoon interacts with RIBEYE, a CtBP2 retinal homolog. In addition, Piccolo interacts with cyclic adenosine monophosphate (cAMP)-guanidine nucleotide exchange factor II (GEFII) via its PDZ domain, with F-actin binding protein (Abp1) via its Q domain, and with G-coupled receptor kinase interactors (GIT) proteins through a region between CC1 and CC2. It interacts with profilin via a polyproline tract. The C2A and C2B domains of Piccolo interact with voltage-dependent L-type Ca^{2+} channel (α-1C). Piccolo, via its C2A domain, is capable of homodimerizing as well as heterodimerizing with RIM, another CAZ protein in a Ca^{2+}-dependent manner (**Figure 1**). The Ca^{2+} affinity of this domain, as well as its capability to dimerize is regulated by alternative splicing. The functional relevance of these interactions are not yet clear. Piccolo C2 domain exhibits a low affinity to Ca^{2+}. It has been suggested that Piccolo may act as a low-affinity Ca^{2+}

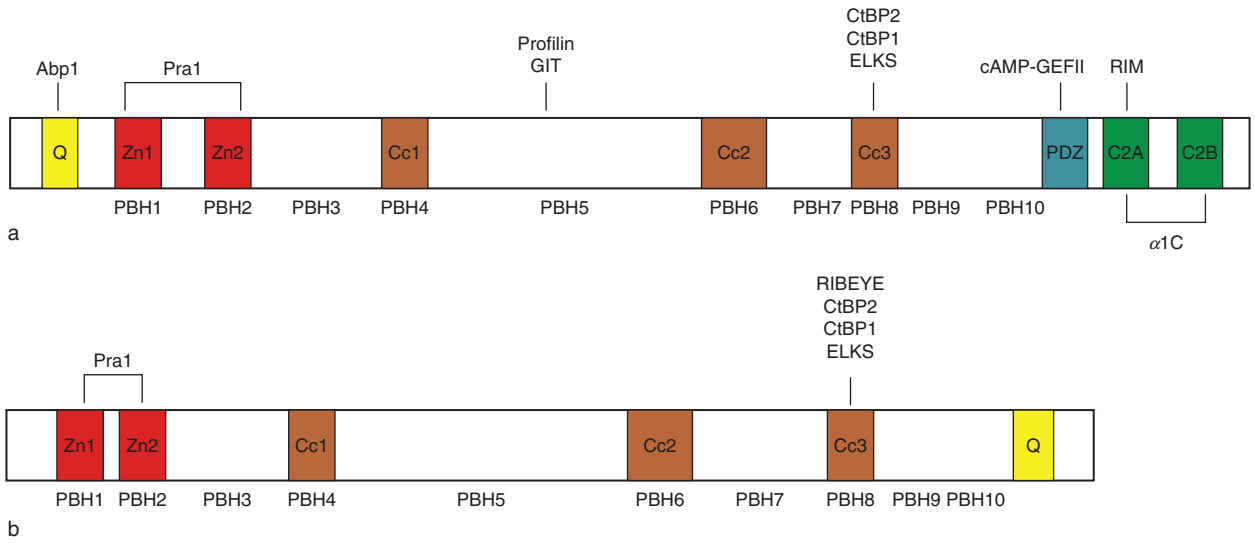

Figure 1 Domain structures and reported interactors of: (a) Piccolo and (b) Bassoon. cAMP, cyclic adenosine monophosphate; Cc, coiled coil domain; CtBP, C-terminal binding protein; ELKS, protein rich in E, L, K, and S; GEFII, guanidine nucleotide exchange factor II; GIT, G-coupled receptor kinase interactor; PBH, Piccolo Bassoon homology domain; PDZ, PSD, Dlg, ZO-1 domain; Q, glutamine-rich domain, RIM, Rab3 interacting molecule; Zn, zinc finger domain.

sensor at the presynaptic site and might regulates synaptic short-term plasticity.

Subcellular Localization of Piccolo and Bassoon in Mammalian Tissues – Morphological and Biochemical Studies

Bassoon and Piccolo are localized at the presynapse in the central nervous system. These molecules, in general, exhibit a uniform presynaptic distribution. However, bipolar ribbon synapses lack Bassoon, with only Piccolo present at the active zone. Localization of the proteins in two different systems (hippocampal cell cultures and fixed central neuronal tissues) was restricted to the limited area of presynaptic electron-dense projections. Outside the central nervous system, both proteins have been observed in somatic neuromuscular junctions (Figure 2(a)). In addition, Bassoon, along with ELKS, has been found in autonomic varicose nerve terminals (synapses *en passant*) in the vas deferens, which are devoid of any pre- or postsynaptic specializations (Figure 2(b)). Piccolo has even been detected in pancreatic β cells, where it is coexpressed with ELKS. It has been suggested that the ELKS–Piccolo/Bassoon interaction might be of functional relevance in these cells. The broader distribution pattern of these proteins suggest that their role is not limited to being scaffolding molecules at the active zones of chemical synapses; however, more comprehensive studies are required to confirm this notion.

Piccolo and Bassoon Transport and Active Zone Assembly

Presynapse can organize itself within 30–60 min of the initial axodendrite contact. Such rapid formation requires a rapid delivery of the molecular elements to the presynaptic region, specifically at the nascent active zone site. It has been postulated that active zones are preassembled in the soma of neurons, transported *en bloc* by axonal transport in the form of specific transport vesicles, and then inserted by fusion with the presynaptic plasma membrane. Indeed, the presence of such dense-core vesicles has been documented. These vesicles, also called the Piccolo transport vesicles (PTV), contain other CAZ proteins such as RIM and Munc13 but is negative for synaptic vesicle proteins such as synaptophysin. Piccolo and Bassoon co-localize with *trans*-Golgi network (TGN) markers in neuronal culture. Moreover, the inhibition of vesicle exit-form Golgi impairs the transport of Bassoon out of cell soma, indicating that PTVs are generated in the TGN. Association of Bassoon to the TGN has chiefly been attributed to an atypical TGN binding domain encompassing residues 2088–2564 of Bassoon, although N-terminal myristoylation might contribute as well. PTV vesicles exhibits some degree of heterogeneity in their composition of proteins, possibly explaining such heterogeneity in mature active zones. It is still to be elucidated whether vesicular transport of the CAZ proteins is the only way of their delivery to the synapse. Morphological studies on the developing rat cerebellar cortex

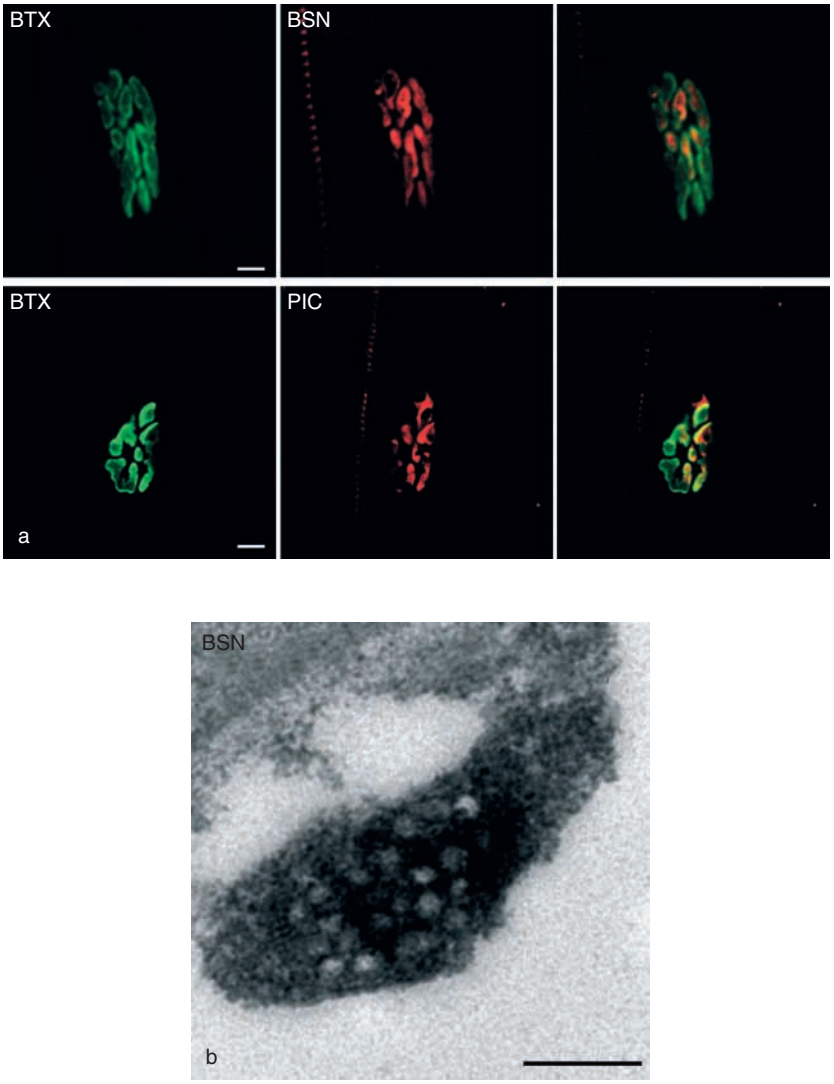

Figure 2 Localization of Piccolo and Bassoon: (a) presence of Piccolo and Bassoon in the neuromuscular junction of diaphragmatic muscle, analyzed by immunofluorescence of rat diaphragm sections. The end plates were identified by counterstaining the nicotinic receptor using bungarotoxin (green channel); (b) localization of Bassoon in varicosities of rat vas deferens using immunoelectronmicroscopy (peroxidase method), the varicosities are easily identifiable by the presence of numerous synaptic vesicle. Scale bar = 16 μm (a) and 200 nm (b). BSN, Bassoon; BTX, α-bungarotoxin; PIC, Piccolo.

revealed a striking difference between Piccolo and Bassoon distribution at postnatal day (P)5 but not P20, which suggests differential trafficking of these proteins, at least in the cerebellum. Moreover, the abundant distribution of these proteins at both the aspects (perikaryonal and axonal) of ligated sciatic nerves suggests bidirectional transportation of Piccolo and Bassoon. It is unknown whether Piccolo and/or Bassoon are upstream to the recruitment of other active zone proteins. In Bassoon mutant mice, the localization of other CAZ proteins remains unchanged. This could either mean that other CAZ proteins are recruited independently to the active zone or that the function of Piccolo and Bassoon are redundant in this context. Systematic analysis of Piccolo/Bassoon double knockout mouse would yield further information regarding their role as scaffolds at the active zone.

Genetic Studies of Piccolo and Bassoon – Overview of Their Function

The large size of Piccolo and Bassoon together with their relative insolubility are major hurdles toward understanding the biochemical and physiological role of these proteins. Downregulation of Piccolo using antisense approach in pancreatic cells pointed to a function of this protein in exocytosis. Although transfection experiments might be informative, conclusive

results will likely come from experiments targeting the genetic disruption of Piccolo and/or Bassoon.

In this context, the Bassoon mutant mouse with part of exon 4 and the entire exon 5 replaced by a *lacZ/neomycin* cassette have been of immense help. In these mice, the central portion of Bassoon (amino acids 505–2889), including an unconventional TGN binding domain, has been deleted. Although the remainder protein (180 kDa) is translated, it is mislocalized and mostly soluble. Biochemical studies revealed that the absence of Bassoon from the presynapse did not alter the levels or presynaptic localization of other CAZ proteins, except Piccolo. The compensatory increase of Piccolo level by 40% indicated a functional overlap between the two proteins. The mutant mice are viable; however, ~50% die during the first 6 months, possibly as a result of epileptic seizures. These mice suffered from general clonic seizures for up to half a minute, having up to four attacks a day. Electroencephalograph (EEG) recordings showed high-frequency (24 Hz) spikes with high amplitude and synchronicity paralleling the seizures. Overall, no anatomical changes were observed at gross light microscopic and ultrastructural levels. However, manganese-enhanced magnetic resonance imaging (MRI) revealed a significant increase in the overall volume of the brain. Most of this increase is contributed by an increase in the size of the hippocampus and cortex. Electrophysiological data demonstrated abnormalities in glutamatergic release. The readily releasable synaptic vesicle pool (RRP) and excitatory postsynaptic current (EPSC) amplitudes were significantly reduced, but the release probability, RRP refilling, and synaptic plasticity were unaltered. The functional studies from the Bassoon mutant mice suggests that a significant number of glutamtergic synapses are nonfunctional in these mice. These results suggest that Bassoon might play an important role in release from the glutamatergic nerve endings rather than affecting the actual development of the synapse or scaffolding of the active zone. In the ribbon synapse, a specialized glutamatergic nerve ending, the presynaptic organizational function of Bassoon is more apparent because the ribbons are not attached to the active zone in Bassoon mutant mice. This morphological aberration is associated with abnormal synaptic transmission and abnormal dendritic arborization of the neuron postsynaptic to ribbon synapse. The disintegration of the ribbon synapse can be traced directly to the interaction of Bassoon with the CtBP2 homolog RIBEYE. The RIBEYE–Bassoon interaction is responsible for the anchoring of the ribbons to the active zone in the ribbon synapses, enabling the proper functioning of these synapses.

The limited loss of function and 50% survival of the Bassoon mutant mouse suggest that it might be a regulatory molecule, but with no fundamental role in neurotransmission. However, it should be kept in mind that the remaining part of Bassoon, although not correctly localized, could perform some functions. In addition, Piccolo, with a shared domain structure with Bassoon, might partly compensate for the lack of this protein, as evidenced by the compensatory increase in its expression. Piccolo-null mice exhibit a Mendelian distribution of survival, with no obvious lethality or any other apparent phenotype. However, the mutant mice are considerably smaller than wild-type littermates. Bassoon/Piccolo double mutant mice exhibit a perinatal lethality, further strengthening the notion that these two proteins might have overlapping functions. Piccolo has additional domains, especially the C2 domains, which are capable of binding Ca^{2+}; therefore, the survival studies do not exclude an additional role for Piccolo. The survival of single mutant mice for both Piccolo and Bassoon does however, point out that neither of these molecules by itself plays an entirely indispensable role in either synapse formation or neurotransmission. Further characterization of the double mutants would reveal the reason for neonatal lethality and shed light on the functions of these giant proteins.

See also: Active Zone; Endocytosis and Presynaptic Scaffolds; Postsynaptic Development: Neuronal Molecular Scaffolds; Presynaptic Development and Active Zones; Ribbon Synapses.

Further Reading

Altrock WD, tom Dieck S, Sokolov M, et al. (2003) Functional inactivation of a fraction of excitatory synapses in mice deficient for the active zone protein bassoon. *Neuron* 37: 787–800.

Cases-Langhoff C, Voss B, Garner AM, et al. (1996) Piccolo, a novel 420 kDa protein associated with the presynaptic cytomatrix. *European Journal of Cell Biology* 69: 214–223.

Fenster SD, Chung WJ, Zhai R, et al. (2000) Piccolo, a presynaptic zinc finger protein structurally related to bassoon. *Neuron* 25: 203–214.

Fujimoto K, Shibasaki T, Yokoi N, et al. (2002) Piccolo, a Ca^{2+} sensor in pancreatic beta-cells: Involvement of cAMP-GEFII. Rim2. Piccolo complex in cAMP-dependent exocytosis. *Journal of Biological Chemistry* 277: 50497–50502.

Juranek J, Mukherjee K, Rickmann M, et al. (2006) Differential expression of active zone proteins in neuromuscular junctions suggests functional diversification. *European Journal of Neuroscience* 24: 3043–3052.

Schoch S and Gundelfinger ED (2006) Molecular organization of the presynaptic active zone. *Cell Tissue Research* 326: 379–391.

Shapira M, Zhai RG, Dresbach T, et al. (2003) Unitary assembly of presynaptic active zones from Piccolo–Bassoon transport vesicles. *Neuron* 38: 237–252.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Takao-Rikitsu E, Mochida S, Inoue E, et al. (2004) Physical and functional interaction of the active zone proteins, CAST, RIM1, and Bassoon, in neurotransmitter release. *Journal of Cell Biology* 164: 301–311.

tom Dieck S, Sanmarti-Vila L, Langnaese K, et al. (1998) Bassoon, a novel zinc-finger CAG/glutamine-repeat protein selectively localized at the active zone of presynaptic nerve terminals. *Journal of Cell Biology* 142: 499–509.

Wang X, Kibschull M, Laue MM, Lichte B, Petrasch-Parwez E, and Kilimann MW (1999) Aczonin, a 550-kD putative scaffolding protein of presynaptic active zones, shares homology regions with Rim and Bassoon and binds profilin. *Journal of Cell Biology* 147: 151–162.

Zhai RG and Bellen HJ (2004) The architecture of the active zone in the presynaptic nerve terminal. *Physiology* (Bethesda, MD) 19: 262–270.

Zhai RG, Vardinon-Friedman H, Cases-Langhoff C, et al. (2001) Assembling the presynaptic active zone: A characterization of an active one precursor vesicle. *Neuron* 29: 131–143.

Zhen M and Jin Y (2004) Presynaptic terminal differentiation: Transport and assembly. *Current Opinion in Neurobiology* 14: 280–287.

Pineal Gland and Melatonin

R J Reiter and D-X Tan, The University of Texas Health Science Center, San Antonio, TX, USA

© 2009 Published by Elsevier Ltd.

For decades, the pineal gland (epiphysis cerebri) was considered a functionless vestige of the brain. This concept changed when Aaron B. Lerner, in the late 1950s, discovered a new molecule, melatonin, in bovine pineal tissue. Since then, the pineal gland, via the secretion of melatonin, has been functionally linked to virtually every organ in the body. Indeed, its actions may be more ubiquitous than any other organ of internal secretion.

Macroscopic and Microscopic Anatomy

Although it derives embryologically from two anlage, the pineal gland is an azygous organ that forms a major portion of the epithalamus. In the human, the gland is situated under the corpus callosum in the free margin of the falx cerebri. It is dorsally attached to the brain stem at the junction of the diencephalon and mesencephalon and lies between the superior colliculi. The gland is surrounded by coalescing veins that drain the neighboring neural structures; these veins join to form the straight sinus. Because of its position near the center of the brain between the cerebral hemispheres and due to the close proximity of the mass of blood vessels, the gland is difficult to approach surgically. The pineal gland exhibits large variations in size in terms of its weight and averages about 120 μg. The gland size does not vary with gender.

Microscopically, the pineal gland is composed of modified neurons referred to as pinealocytes and glial cells. Postganglionic sympathetic nerves terminate near the pinealocytes. The human pineal gland typically accumulates hydroxyapatite deposits (i.e., calcifications frequently referred as acervuli or 'brain sand') throughout life. Once thought to be limited to the pineal gland of elderly individuals, these deposits, as revealed by currently available imaging techniques, begin forming in the adolescent years. The gland also may contain one or more cysts. Neither the calcified deposits nor the cysts are known to directly compromise the function of the pineal gland. Only in exceptional cases where a cyst enlarges to the point of rupture does it become problematic.

Blood Supply

The blood supply to the pineal gland is particularly perfuse. On a per-gram basis, the gland possesses a capillary bed more dense than virtually any other organ; this morphological feature portends a highly metabolically active organ. The arteries entering the pineal gland are tributaries of the posterior choroidal vessels, which are terminal branches of the bilateral posterior cerebral vessels. The venous drainage is into the large cerebral venous sinus that surrounds the gland.

Innervation

The primary functional innervation of the pineal gland is postganglionic sympathetic fibers whose cell bodies are in the superior cerebral ganglia situated deep to the bifurcation of the common carotid arteries in the neck. The postganglionic fibers follow blood vessels to reach the pineal gland where, just before they penetrate the organ, they form distinct nerves, the nervi conarii. Within the gland the fibers ramify and the terminals make contact with pinealocytes, the functional units of the pineal.

The preganglionic fibers that terminate in the superior cervical ganglia are located in the intermediolateral cell column of the first several thoracic cord segments. These neurons receive neural signals from the suprachiasmatic nuclei (SCN, or biological clock) via the paraventricular nuclei. The SCN are innervated by a specialized portion of the optic nerve, the retina hypothalamic tract, whose cell bodies are modified ganglion cells in the inner layer of the retina. Via these complex neural connections, the pineal gland is apprised of the photoperiodic state of the individual, that is, either light or dark (Figure 1).

Functionally, the pineal gland is inextricably dependent upon its sympathetic innervation. Interruption of the neural pathway at any point between the SCN and the pineal gland abolishes all known functions of the organ.

Melatonin Synthesis

The exclusive function of the pineal is the synthesis and secretion of the chemical messenger melatonin (*N*-acetyl-5-methoxytryptamine). The mechanisms and processes involved in the production of melatonin are generally well known. The dietary precursor of melatonin is the amino acid tryptophan, which is taken up into pinealocytes from the capillaries that perfuse the gland. Tryptophan, is converted to the monoamine serotonin (5-hydroxytryptamine) in a two-step process. Serotonin is then *N*-acetylated, a step catalyzed by the enzyme serotonin-*N*-acetyltransferase (NAT), to form *N*-acetylserotonin. This latter molecule is then *O*-methylated by hydroxyindole-*O*-methyltransferase

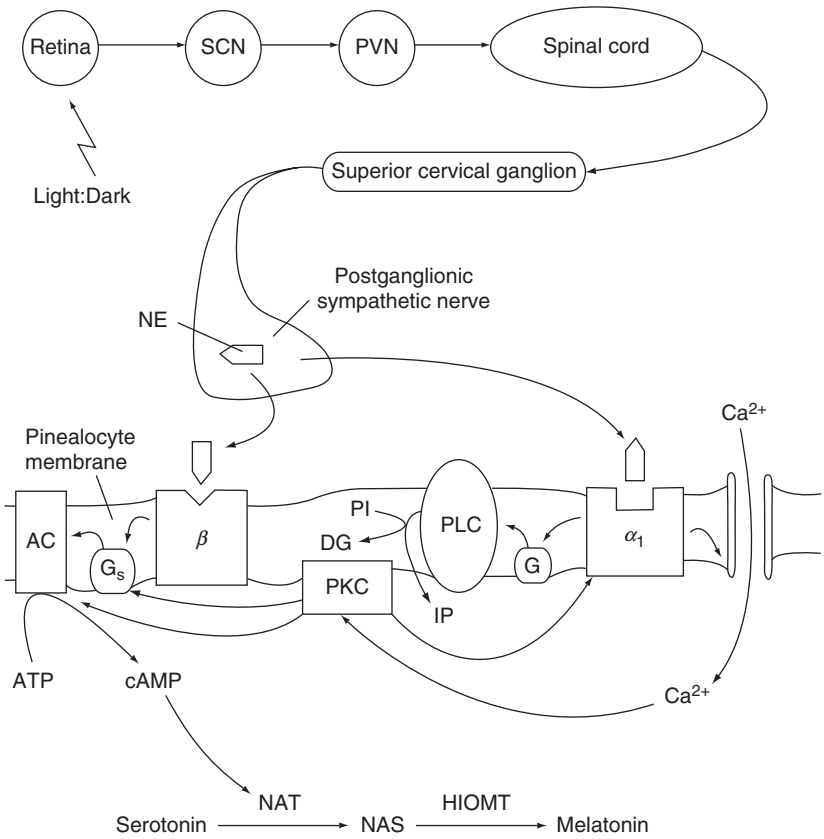

Figure 1 Neural corrections between eyes and the pineal gland and the cellular mechanisms which mediate the nighttime synthesis of melatonin from serotonin. AC, adenylate cyclase; ATP, adenosine tri phosphate; cAMP, cyclic adenosine monophosphate; DG, diacylglycerol; G, G-protein; Gs, G stimulatory protein, HIOMT, hydroxyindole-*O*-methyltransferase; NAS, *N*-acetyl serotonin; NE, norepinephrine; PI, phosphoinositol; PKC, protein kinase C; PLC, phopholipase C; PVN, paraventricular nuclei; SCN, suprachiasmatic nuclei; α_1, alpha$_1$-adrenergic receptor; β, β-adrenergic receptor.

(HIOMT) with the resultant generation of melatonin (Figure 1).

Uniquely, melatonin production in the pineal gland occurs only at night during darkness. During the night the SCN, via the nerve network described previously, signals the postganglionic neurons in the pineal to release norepinephrine (NE) onto the pinealocytes. NE acts on both β-adrenergic and α-adrenergic receptors to mediate a cascade of signal transduction events that culminate in the conversion of serotonin to melatonin (Figure 1). While the enzyme NAT was initially thought to be rate-limiting in the production of melatonin, recent data suggest this may not always be the case.

Once generated, melatonin quickly passively escapes the pinealocytes both into the capillary bed and directly, or indirectly, possibly into the cerebrospinal fluid (CSF) within the third ventricle. This latter process is achieved via the pineal recess, an outpouching of the third ventricle.

Given that melatonin is produced primarily during darkness, blood and CSF melatonin levels are significantly higher at night than during the day. The duration of elevated melatonin in these fluids is proportional to the duration of the dark period to which the individual is exposed. Likewise, light exposure at night, what is generally referred to as light pollution, if sufficiently bright and the proper wavelength, rapidly curtails pineal melatonin synthesis and fluid melatonin values drop precipitously.

Melatonin Levels in Bodily Fluids

Although there are organs in addition to the pineal gland that synthesize melatonin, only the pineal gland releases significant quantities of the indole into the blood and CSF. Because of melatonin's low production in the pineal during the daily light period, its daytime levels in bodily fluids are uniformly low (in the low pg ml^{-1} range). At night, blood levels increase markedly. The amplitude of the nighttime rise in blood melatonin varies considerably between individuals and is genetically determined. The peak melatonin levels in the blood normally range from 50 to 200 pg ml^{-1}. CSF levels of melatonin at night may be much higher. The rhythm of melatonin in bodily fluids in terms of its amplitude is highly uniform from one night to the next.

Effects of Age on Pineal Melatonin Production

In infants, the pineal does not exhibit a pronounced day/night melatonin rhythm until several months after birth. During childhood the nocturnal rise in melatonin is at its maximum. After adulthood is achieved, the production of pineal melatonin wanes and the nocturnal amplitude is diminished. This decline with age continues with the result that elderly subjects may lack a day/night difference in circulating melatonin levels. There is some evidence that certain diseases may be associated with unusually low melatonin values.

Functions of Melatonin

Once in bodily fluids, melatonin has access to every cell in the organism. The indole crosses all morpho-physiological barriers, for example, the blood–brain barrier. Melatonin is highly lipid soluble and readily enters cells where it seems to be distributed within all subcellular compartments.

Melatonin has both receptor-mediated effects as well as actions that are receptor independent. The major membrane receptors have seven transmembrane-spanning domains, are G-protein linked, have a nanomolar affinity for melatonin, and seem to be very widely distributed. The SCN have a particularly high density of melatonin receptors. The membrane receptors commonly identified in mammals including humans are classified as MT1 and MT2. The signal transduction pathways of the membrane melatonin receptors involve a pertussis toxin-sensitive G-protein which reduces intracellular cAMP. Several pharmacological agonists and antagonists of these receptors have been identified.

In addition to membrane receptors, melatonin also acts via intracellular binding sites/receptors. In the cytoplasm, melatonin binds to calmodulin, and nuclear binding sites have also been demonstrated. The specific consequences of the receptor-mediated actions of intracellular melatonin are not well defined.

Besides receptor-dependent actions, melatonin functions as a direct free-radical scavenger. This function requires only that melatonin be in the vicinity of a free radical when it is generated.

Melatonin and Seasonal Reproduction

One of the first discovered functions of the melatonin rhythm was the ability of this cycle to mediate seasonal changes in reproduction. Given that the duration of nocturnal melatonin levels varies with seasonal changes in photoperiodic length, the melatonin message functions as a calendar to signal the changing seasons. This information is utilized by animals that reproduce seasonally to alter their reproductive capability accordingly. The melatonin signal in this scheme is a passive message, with the organism using the signal to its own advantage. Some animals are functionally most competent and mate during long days, while others do so during short days of the winter. In both cases, the changing duration of the melatonin peak impels these annual cycles. Some species, such as humans, are continuous breeders.

Melatonin and Circadian Functions

One of the major functions of the melatonin rhythm seems to be to strengthen and/or synchronize circadian rhythms via an action on the SCN; in this case the melatonin rhythm functions as a clock. Whereas the loss of the rhythmic melatonin cycle may not totally disrupt circadian organization, subtle desynchrony of a variety of physiological functions as well as slower re-entrainment of activity rhythms may occur when the circadian melatonin signal is absent.

Melatonin, when administered exogenously, influences either directly or indirectly the phase and/or period of the SCN. This chronobiotic property of melatonin is used to manipulate the sleep/wake cycle and to assist individuals in recovering from jet lag. The ability of melatonin to modify circadian rhythmicity is a major function of melatonin.

Melatonin as an Antioxidant

Melatonin is a direct free-radical scavenger and an indirect antioxidant. As a scavenger it deftly neutralizes a variety of reactive oxygen species (ROS) and reactive nitrogen species (RNS). It directly detoxifies hydrogen peroxide, hypochlorous acid, singlet oxygen, nitric oxide, and the highly toxic hydroxyl radical and the peroxynitrite anion (Figure 2). Melatonin functions in this capacity by electron donation, which is a typical means of detoxifying radicals by antioxidants. When melatonin interacts with ROS/RNS, a series of metabolites are formed, among which are cyclic-3-hydroxymelatonin (c3-OHM), N^1-acetyl-N^2-formyl-5-methoxykynuramine (AFMK), and N^1-acetyl-5-methoxykynuramine (AMK) (Figure 3). Each of these products is likewise a highly efficient free radical scavenger which greatly magnifies the antioxidant activity of melatonin. This series of reactions is referred to as melatonin's free-radical scavenging cascade.

In addition to functioning as a direct free-radical scavenger, melatonin also stimulates, likely via receptor mechanisms, a number of antioxidative enzymes which are vitally important for reducing oxidative damage. Thus, melatonin stimulates gene expression and activity of the superoxide dismutases, glutathione peroxidases, and glutathione reductase. Furthermore,

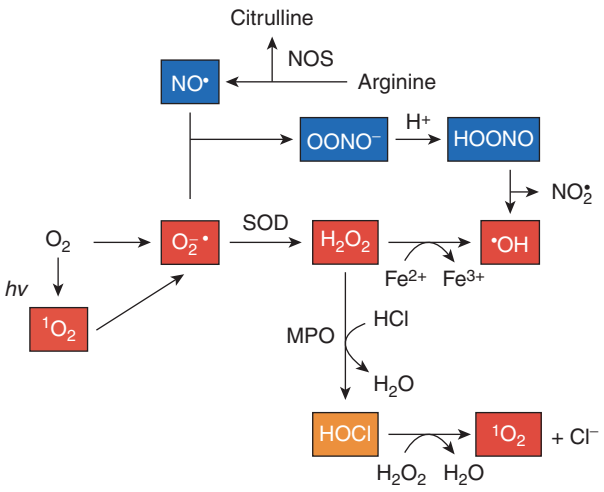

Figure 2 Formation of free radicals and radical products. Oxygen-(red), nitrogen-(blue), and chlorine (orange)-based toxic reactants are shown. The most reactive of these products are the hydroxyl radical (•OH) and the peroxynitrite anion ($OONO^-$). MPO, myloperoxidase; SOD, superoxide dismutase.

melatonin promotes the activity of the rate-limiting enzyme in glutathione (GSH) synthesis, that is, gamma-glutamylcysteine synthase. GSH is an important intracellular antioxidant. Some of these actions may be mediated via the metabolites that are formed when melatonin detoxifies radicals and radical products.

Nitric oxide synthase, which controls the formation of nitric oxide, can be a pro-oxidative enzyme. Nitric oxide couples with the superoxide anion radical to generate the devastating toxic peroxynitrite anion (Figure 2). One of the melatonin metabolites, AMF, inhibits nitric oxide synthesis, thereby reducing generation of the peroxynitrite anion and limiting oxidative damage.

Melatonin as an Oncostatic Agent

Melatonin has two potential functions in reducing the incidence of cancer: it prevents free-radical

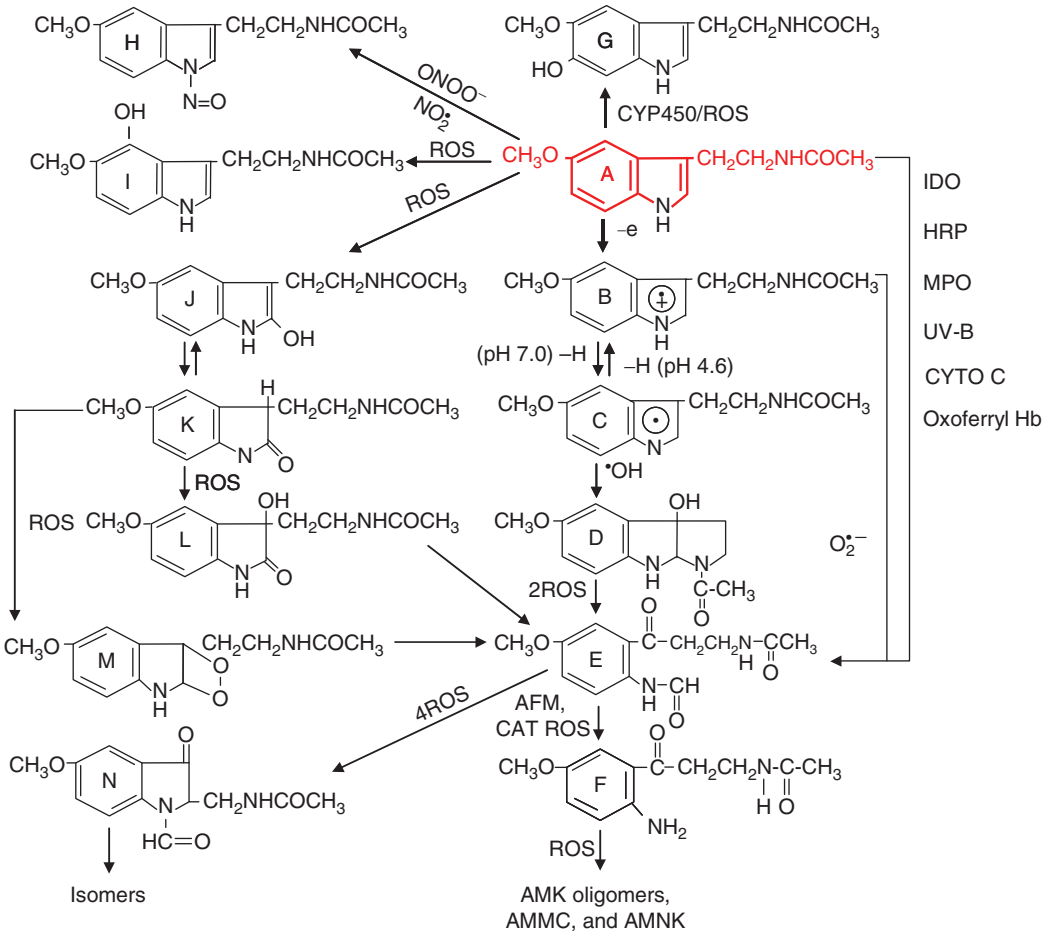

Figure 3 Melatonin (red) and the numerous metabolites that are formed when the indole reacts with free radicals and related products. AMK, N^1-acetyl-5-methoxykynuramine; AMMC, 3-acetamidomethyl-6-methoxycinnolinone; AMNK, N^1-acetyl-5-methoxy-nitrokynuramine; CAT, catalase; CYTO C, cytochrome C; Hb, hemoglobin; HRP, horse radish peroxidase; IDO, indoleamine-2,3-dioxygenase; MPO, myloperoxidase; ROS, reactive oxygen species; UV-B, ultraviolet-B.

damage to the genome (DNA), thereby limiting the initiation of cancer, and it reduces the growth rate of established tumors. In reference to the latter, melatonin has several means to stymie growth and cellular proliferation of cancer cells. Some of these actions include melatonin's ability to curtail the uptake of growth factors, for example, the omega-6 fatty acid linoleic acid, by cancer cells; this action is accomplished via membrane melatonin receptors. Other mechanisms whereby melatonin may reduce the growth of cancer cells include reducing endothelin-1 production; this molecule is produced in abundance by many cancer cells and it promotes the vascularization of tumors which they need for survival. Also, melatonin inhibits telomerase activity in cancer cells. This enzyme maintains the telomeres of linear eukaryotic chromosomes. These physical extensions are key in maintaining the integrity and stability of chromosomal structure, and, as a consequence, they also render cancer cells very hardy. Finally, antisteroidal actions of melatonin may contribute to the ability of the indole to reduce cell proliferation of hormone-dependent cancers. In particular, this action of melatonin is germane to estrogen-responsive breast cancer and testosterone-sensitive prostate cancer.

Melatonin Regulation of Immune Function

Melatonin is produced in at least some immune cells, and its interactions with this system are pleomorphic. In general, the immunomodulatory effects of melatonin are beneficial to the organism and are mediated potentially by both membrane receptors and nuclear binding sites. Melatonin has direct effects on humoral and cellular immune responses. In general, it reduces proinflammatory cytokine/chemokine production by limiting the translocation of nuclear factor kappa-B to the nucleus and its binding to DNA. Under conditions where the immune system is already hyperactive, for example, in some autoimmune diseases, melatonin's use may be counterindicated.

Concluding Remarks

The discovery of melatonin in the pineal gland occurred roughly 50 years ago. The rhythmic production and secretion of this organ are essential to some of its functions. However, melatonin's actions are far greater than originally envisaged. This functional diversity includes actions that are mediated via receptors on the membranes of many cells (and possibly via intracellular binding sites). Additionally, melatonin has receptor-independent functions, as typified by its ability to directly scavenge ROS and RNS. Some of melatonin's functions are carried out by metabolites of the indole which are generated when melatonin functions as a free-radical scavenger. It seems almost certain that the full functional repertoire of melatonin still has not yet been uncovered. Considering this functional diversity, the pineal gland must not be overlooked as an essential, pluripotent, and beneficial organ.

See also: Circadian Function and Therapeutic Potential of Melatonin in Humans; Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Melatonin Regulation of Circadian Rhythmicity in Vertebrates.

Further Reading

Acuna-Castroviejo D, Escames G, Rodriguez MI, and Lopez LC (2007) Melatonin role in mitochondrial function. *Frontiers in Bioscience* 12: 947–963.

Dubocovich ML, Rivera-Bermudez MA, Gerdin MJ, and Masana ML (2003) Molecular pharmacology, regulation and function of mammalian melatonin receptors. *Frontiers in Bioscience* 8: 1093–1108.

Jasser SA, Blask DE, and Brainard GC (2006) Light during darkness and cancer: Relationships in circadian photoreception and tumor biology. *Cancer Causes Control* 17: 515–523.

Korf HW and von Gall C (2006) Mice, melatonin and the circadian rhythm. *Molecular and Cellular Endocrinology* 252: 57–68.

Reiter RJ (1991) Pineal melatonin: Cell biology of its synthesis and physiological interactions. *Endocrine Development* 12: 151–180.

Reiter RJ (1993) The melatonin rhythm: Both a clock and a calendar. *Experientia* 49: 654–664.

Reiter RJ (2005) Melatonin. *Endocrine* 27: 87–212.

Tan DX, Manchester LC, Terron MP, Flores LJ, and Reiter RJ (2007) One molecule, many derivatives: A never ending interaction of melatonin with reactive oxygen and nitrogen species? *Journal of Pineal Research* 42: 28–42.

Wehr TA (2001) Photoperiodism in humans and other primates: Evidence and implications. *Journal of Biological Rhythms* 16: 348–364.

Pituitary Gland (Cell Types, Mediators, Development)

G V Childs, University of Arkansas for Medical Sciences, Little Rock, AR, USA

© 2009 Elsevier Ltd. All rights reserved.

Location and Anatomy

The pituitary gland, also called the hypophysis, lies in a bony depression at the base of the skull called the sella turcica (Turkish saddle, because of its shape). It is protected by dura mater, called the diaphragma sella. The location of the human pituitary gland can be pinpointed if one points to one's nose and draws an imaginary diagonal line up the nostril into the brain. This is actually the route used by neurosurgeons to access the pituitary and remove tumors.

The pituitary gland in humans is approximately the size of a pea and is made of two types of glandular tissue – the adenohypophysis and the neurohypophysis. The adenophyphysis includes a large region that faces the front of the body (anterior or ventral) called the anterior lobe, or the pars distalis. It also includes a collar of cells around the stalk called the pars tuberalis and clusters of cells attached to the neurohypophysis called the pars intermedia. Figures 1 and 2 illustrate the divisions and subdivisions of the hypophysis.

The neurohypophysis includes a region that faces the back of the body (dorsal or posterior), called the posterior lobe or the pars nervosa. It also includes the connection to the brain, which is made up of nerve fibers and vessels, called the infundibular stalk. The neurohypophysis broadens as it connects to the brain, resembling a funnel, which is how the stalk got the name. This broad area at the top of the stalk is a highly vascular region called the median eminence. The funnel-shaped anatomy inspired the Greek physician, Galen, to develop a theory of fluid circulation from the ventricles of the brain in 130–200 AD, which proposed that fluid was funneled from the third ventricle through the infundibular stalk to the pituitary. After the pituitary filtered the fluid, it was sent out the nose as mucus. This theory was believed for centuries and, in 1543, Andreas Vesalius named the pituitary after the Latin word for mucus, *pituita*.

Anterior Pituitary Cell Types

The anterior pituitary has been considered the master gland because its various cell types regulate endocrine organs and systems throughout the body. However, the pituitary cells are subject to regulation both from factors in the hypothalamus and from feedback by specific target organs.

The earliest light microscopic studies distinguished pituitary cell types by their reactions with different colored dyes. These trichrome stains differentiated three types of cells. The red cells were called acidophils, the blue cells were called basophils, and the colorless cells were called chromophobes. Historians now recognize that the term basophil was actually a misnomer because these early trichrome stains included only acid dyes. True basophilia, which is dependent on areas of ribosomes in the cytoplasm, could not be detected by these dyes. Nevertheless, the terms basophil and acidophil persist, especially in the description of different types of pituitary tumors. Later studies focused on immunolabeling to identify the hormone content in the individual cell types, which helped alleviate the confusion in terminology. The periodic acid Schiff reagent (PAS) was also used to distinguish pituitary cells that produced glycoprotein hormones. Electron microscopic studies identified the cells on the basis of secretory granule size and shape.

Somatotropes

Somatotropes produce growth hormone (GH) or somatotropin, are round–ovoid, and are missing in dwarf animals. They are distinguished by an acidophilic cytoplasm and, when viewed at the electron microscopic level, by densely packed, medium-sized granules (350 nm). In all species, they are the most abundant hormone-producing cell in the anterior pituitary, representing 30–50% of the cell population. GH cells separate in fractions of pituitary cells containing medium-sized cells.

GH is a protein hormone of 191 amino acids (22 kDa) that has anabolic and metabolic functions throughout the body. First discovered as a product of an acidophilic tumor in 1900 by Benda, Harvey Cushing established that this overactive tumor caused acromegaly, a disease associated with increases in bone mass and disfiguration. Evans and Long injected animals with pituitary extracts and caused gigantism, thus proving the hypothesis that the pituitary produces a GH.

GH is best known for its regulation of growth in length of long bones, by regulating the liver's production of insulin-like growth factor-1 (IGF-1), also known as somatomedin. IGF-1 stimulates growth in length of bones at the epiphysial plate. Gigantism results when there is a GH-secreting tumor before puberty.

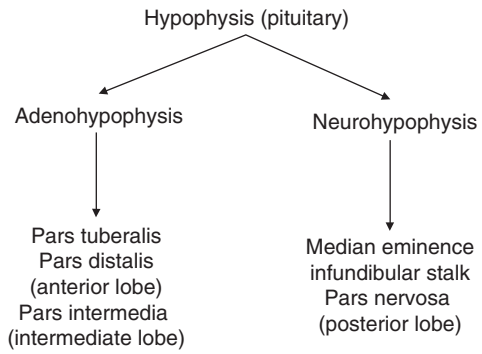

Figure 1 Flow diagram showing the two main parts of the pituitary gland (hypophysis) and their unique regions.

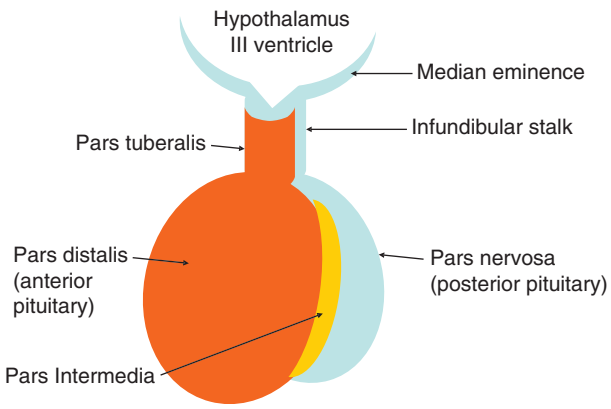

Figure 2 Pituitary gland in the human showing different regions of each lobe.

GH also plays important anabolic roles in building muscle and bone and normalizing metabolic functions, including the regulation of insulin. It is one of the few pituitary hormones to decrease dramatically during aging, which is correlated with frailty in the elderly. GH secretion is upregulated by exercise, which prevents frailty.

Lactotropes

Lactotropes, or prolactin cells, are acidophils, staining bright red with erythrosine or carmosine. The animal models in which they could best be studied were lactating or pregnant animals because their pituitaries showed enlarged, bright red prolactin cells. Lactotropes produce the polypeptide hormone prolactin, which has 199 amino acids (24 kDa). Prolactin was discovered by Oscar Riddle, who described lactation-inducing effects of anterior pituitary extracts. Prolactin is best known for the stimulation of the growth and function of the mammary gland. Its exact role in the male is poorly understood. More than 300 separate functions for this hormone have been described in a number of species.

Lactotropes have been distinguished by their angular or stellate shape and content of irregularly shaped secretion granules. In rodents, they contain the largest granules of all the pituitary cell types. The granules exhibit irregular, dumbbell shapes. In humans, the granules are smaller; however, they still retain their pleomorphic shapes. Lactotropes also display stacks of lamellar rough endoplasmic reticulum. In normal animals, lactotropes represent approximately 15–26% of the pituitary cell population, with more lactotropes in females. In cell separation studies, prolactin cells separate with the smallest cells. A subset of estrogen-sensitive somatomammotropes that contain both GH and prolactin has been identified.

Thyrotropes

Thyrotropes are among the least abundant of the pituitary cell types, representing 6–10% of the normal pituitary cell population. They produce the 28 kDa glycoprotein hormone, thyroid-stimulating hormone (TSH), which regulates thyroid follicular cells to produce thyroxine or triiodothyronine. Thus, TSH ultimately regulates the body's metabolic rate.

TSH activity from the pituitary was first described in 1926 by Eduard Uhlenhuth, who showed that injections of bovine pituitary extracts caused the enlargement of thyroid gland follicular cells. TSH, a glycoprotein, was purified in the 1960s and found to have two subunits, an α and a β subunit. The α subunit has a structure shared with both gonadotropins, and the structure of the β subunit confers the biological specificity of the TSH molecules.

Thyrotropes are basophils and stain with the PAS reagent. TSH cells are distinguished by their content of very small secretory granules (100–150 nm in diameter) and dilated profiles of rough endoplasmic reticulum. Thyrotropes separate with the fractions containing the largest cells. However, some active thyrotropes may be found among the smallest cells. Thyrotropes enlarge following removal of the thyroid (thyroidectomy) and develop vacuoles, which are dilated sacs of rough endoplasmic reticulum.

Corticotropes

In human pituitaries, corticotropes are basophils. These cells are also PAS positive because they produce a glycoprotein prohormone called pro-opiomelanocortin (POMC). POMC has 241 amino acids (30 kDa) and can be enzymatically cleaved to produce smaller polypeptides, including adrenocorticotropin (ACTH), α- or β-melanotropin (melanocyte-stimulating hormone (MSH)), β-endorphin, and met-enkephalin. ACTH

stimulates glucocorticoid production and secretion by the adrenal cortex (zona glomerulosa cells), playing a vital role in the stress response. It also stimulates the production of dehydroepiandosterone from the adrenal cortex. β-Endorphin provides our own source of analgesia, relieving pain and having a calming influence. It is the source of the runner's high experienced during exercise. MSH stimulates the dispersion of the pigment in skin cells, called melanocytes. It is more important in lower mammals, however, and is the main product of the intermediate lobe.

The early search for the corticotrope began with a description of Cushing's disease, which is caused by high levels of the adrenal hormone cortisol. Harvey Cushing associated this disease with pituitary basophilism. In 1933, Collip and co-workers isolated and characterized the hormone that regulated cortisol, ACTH. It is now known that Cushing's disease can arise directly from an adrenal tumor or indirectly from a pituitary ACTH-secreting tumor.

In humans, corticotropes are distinguished by abundant 300 nm secretory granules. They are large, prominent, and become filled with filaments in response to glucocorticoid feedback. Other large mammals have well-granulated corticotropes, similar to those in humans. In contrast, corticotropes in mice and rats have not been as easy to detect. Studies of adrenalectomized rats or mice in the early 1970s showed that corticotropes in the rat are stellate, with peripherally arranged granules approximately 200–250 nm in diameter. They represent approximately 10% of the anterior pituitary population and separate with the population of small or medium-sized pituitary cells.

Gonadotropes

Gonadotropes are classified as basophils. They also are PAS positive because their two gonadotropin hormones, luteinizing hormone (LH) and follicle-stimulating hormone (FSH), are both glycoproteins (33–38 kDa). Both gonadotropins have two subunits–an identical α subunit, which has 92 amino acids, and a β subunit that confers biological specificity to each of the gonadotropin molecules. The α subunit is identical to that in TSH as well. The LH β subunit has 121 amino acids and that for FSH has 118 amino acids.

LH bioactivity was identified when Zondek and Ascheim reported that urine extracts from pregnant women stimulated the formation of corpora lutea (thereby showing luteinizing activity). FSH bioactivity was discovered by Zondek, who reported that extracts of the urine from postmenopausal women influenced follicular development. A regulatory link between the pituitary and gonads was discovered as early as the 1920s when Evans and Long treated rats with anterior pituitary extracts and produced corpora lutea in ovaries, indicating ovulation. The pituitary origin of the stimulatory substance was confirmed in 1926 by PE Smith, who showed that ovarian atrophy, following removal of the pituitary, was reversed by pituitary transplants.

We now know that LH stimulates the final maturation of the oocyte and ovulation. FSH stimulates the development of follicles. In males, LH, also called interstitial cell-stimulating hormone, stimulates the production of testosterone by interstitial or Leydig cells. FSH is also important in the stimulation of cells that support sperm development.

In most species, gonadotropes include an ovoid or a round subset that separates with the largest pituitary cells and is distinguished by two populations of granules scattered between dilated profiles of rough endoplasmic reticulum. The granules may be 200 nm or as large as 500–600 nm. A second subset of gonadotropes is more angular and contains a population of medium-sized granules. This subset separates with medium-sized cells, is prominent in cycling females, and may also contain GH. A third subset of gonadotropes is small and sparsely granulated, resembling corticotropes; these cells may contain ACTH and gonadotropins.

The gonadotrope population can be distinguished cytochemically by its content of LH and/or FSH or by its expression of receptors for the neuropeptide gonadotropin-releasing hormone (GnRH). It can also be distinguished when it enlarges after castration, becoming filled with vacuoles of rough endoplasmic reticulum. Approximately 50–75% of gonadotropes express both gonadotropins (bihormonal) in normal animals, although there are species differences in the proportion of bihormonal and monohormonal cells. Gonadotropes detected by dual labeling for both gonadotropins or for GnRH receptors represent approximately 15% of pituitary cells in normal rats. Gonadotropes may shift their morphology with their secretory state, and after losses in stores from surge secretory activity, some cells may be detected only by their content of mRNA.

Multifunctional Anterior Pituitary Cells

Sensitive cytochemical and assay techniques have identified pituitary cell types that express multiple hormones. A well-established example of a bihormonal expression pattern is represented by the subset of cells that produce both growth hormone and prolactin, which can be regulated to produce more prolactin by estrogens. Other examples include the

co-expression of POMC by gonadotropes early in postnatal development and the transitional expression of gonadotropins by GH cells during the reproductive cycle.

Increasing evidence supports a hypothesis that not all pituitary cells are terminally differentiated. A survey of the literature will lead the reader to evidence for the co-expression of as many as four hormones by a single cell. This calls into question the original dogma that each pituitary cell type produces only one hormone, and it shows the potential of the anterior pituitary population to retain sets of cells that may serve as precursor or stem cells. This potential may allow the anterior lobe cells to support the body efficiently, when multiple systems are required, bringing online extra cells to enhance a given cell population.

Folliculostellate Cells

A network of stellate cells that are electrically coupled to one another (and perhaps to other glandular cells), called folliculostellate cells, have been described in pituitaries from many species yet they remain somewhat mysterious. They stain with a glial cell marker, glial fibrillary acid protein, and S-100 protein. They also contain a number of other hormones, growth factors, and cytokines that regulate specific pituitary cells. They form follicles with other folliculostellate cells. Their electrochemical coupling is by way of gap junctions, which suggests the capacity for rapid communication in the anterior pituitary along a cellular network.

The folliculostellate cells are chromophobic cells, showing little, if any, evidence for staining with the classical tinctorial stains. In cell separation experiments, they separate in clusters along with the largest cells. Their content of specific hormones, growth factors, and cytokines suggests that they may somehow support the growth and maintenance of hormone-bearing anterior pituitary cells. Their electrical coupling and capacity to communicate may provide a rapid means of mobilizing cells to fully support a particular function. This network could signal major secretory events, such as the LH surge from gonadotropes or the rapid responses to stress by corticotropes.

Regulation of Anterior Pituitary Cells

Blood Supply

The architecture of the blood vascular system has evolved to support the regulation of the anterior pituitary cells. Most of the anterior pituitary blood is brought in from the superior hypophysial arteries, which branch from the internal carotids and the posterior communicating branch from the circle of Willis. These vessels penetrate the stalk–median eminence before they supply the anterior pituitary and form a capillary plexus that provides increased surface area for nerve fiber endings in the median eminence. From this capillary bed, blood is collected into several long portal veins that extend down the stalk to penetrate the anterior pituitary. These portal veins then branch into another capillary bed that supplies the anterior pituitary cells. The capillaries drain into the sinus venosus (venus sinuses in the dura mater).

This portal system architecture provides a focused vascular route for the transport of hormones and other regulatory factors to pituitary cells. The hormones are secreted from nerve endings in the median eminence, and this direct route prevents them from being diluted by the entire bloodstream. Consequently, only miniscule amounts of active hypothalamic hormones are needed to regulate the cells. These regulatory hormones were identified in studies by the laboratories of Roger Guillemin and Andrew Schally, who received the Nobel Prize for their discoveries. They had to use purified fractions from 200 000 to 500 000 sheep hypothalami to obtain sufficient material for biological effects.

Regulators

The regulatory hormones are produced by neurons that are distributed in clusters (nuclei) throughout the hypothalamus. The hormones are stored in secretory granules, which are sent down axons to nerve endings that lie along the capillaries in the median eminence. There, they are released into the blood supply that is collected and transported to the anterior pituitary by the portal veins. Before their identities were confirmed, they were called releasing factors; however, most are now called releasing hormones, named after their principal anterior pituitary target cell.

Growth hormone cells (somatotropes) are stimulated by a neuropeptide produced in the arcuate nucleus of the hypothalamus, called GHRH. It is a 44-amino-acid peptide that is released in pulses into the pituitary portal circulation, bringing about pulses of GH. It was actually discovered and isolated while Michael Thorner was treating an acromegalic patient, who had a GHRH-secreting pancreatic tumor. He sent the tumor tissue to Wiley Vale and Roger Guillemin for isolation and purification of GHRH.

Somatotropes are also regulated by an inhibitor called somatostatin. A polypeptide that is produced by the cells in the periventricular nucleus of the hypothalamus, it was discovered by Guillemin in 1973 from an extract of more than 500 000 sheep hypothalami. Somatostatin inhibits GH secretion and the actions of GHRH. Peaks of somatostatin

are assayed during the lowest points of the GH and GHRH pulses.

Thyrotropes are regulated primarily by thyrotropin-releasing hormone (TRH), which is a tripeptide hormone produced in small neurons (parvocellular) in the paraventricular region of the hypothalamus. These neurons are very sensitive to temperature. Cold- and heat-sensing neurons provide information to the TRH neurons that secrete TRH into the pituitary blood supply to stimulate TSH. TSH then stimulates the thyroid to upregulate metabolism and stimulate shivering and other responses to cold.

Prolactin cells are regulated by dopamine, which is a chronic inhibitor of this hormone, providing a clamp on the cells until prolactin is needed during pregnancy and lactation. The search for a prolactin stimulatory hormone is ongoing, although researchers have shown that TRH has stimulatory actions on lactotropes. The chronic inhibition by dopamine has provided chemotherapeutic tools that help control the growth of prolactin-secreting tumors.

Corticotropes are regulated by corticotropin-releasing hormone (CRH), a 41-amino-acid peptide that is produced by parvocellular neurons in the paraventricular nucleus of the hypothalamus. Another neuropeptide with potent stimulatory actions on corticotropes is arginine vasopressin (AVP). CRH and AVP work synergistically on anterior pituitary corticotropes. CRH neurons are highly sensitive to stress and their rapid release of CRH provides the stimulus for corticotropes to produce and release ACTH and β-endorphin. ACTH, in turn, stimulates the production of adrenal cortisol, which is needed for the flight-or-fight response.

Gonadotropes are regulated by a neuropeptide called GnRH, or LH-releasing hormone, which is a decapeptide. The search for a separate hormone for FSH is ongoing, although candidates have been reported. Produced by neurons scattered in the preoptic and anterior regions of the hypothalamus stretching back to the arcuate nucleus, GnRH is secreted in pulses and stimulates both LH and FSH synthesis and release. GnRH pulses may be driven by various factors, including gonadal steroid hormone feedback and neurotransmitters such as the polypeptide products of the *KISS1* gene, called kisspeptins. Rapid pulses of GnRH favor the release of LH and slower pulses of GnRH favor the release of FSH.

The previous discussion focused on the major neuropeptides from the hypothalamus. Other factors affect pituitary cells, including feedback from target organ products as well as locally produced cytokines and growth factors. Some are produced by the pituitary cells and may represent sources of paracrine or autocrine regulation.

Pars Tuberalis

The pars tuberalis consists of anterior lobe basophils (gonadotropes and thyrotropes) that have formed an incomplete collar around the infundibular stalk. Their expression of melatonin receptors and proximity to the suprachiasmatic nuclei make them candidates for responses to light and dark stimuli (photoperiodicity). These cells may be regulated by melatonin from the pineal and they express specific clock genes. Folliculostellate cells have also been described in the pars tuberalis as having close associations with GnRH-containing nerve fiber endings. This may link responses to photoperiodic stimuli to the reproductive system via this cellular network.

Pars Intermedia

The pars intermedia forms along the region of contact between the anterior and posterior lobes. In adult humans, it remains only as nests of basophilic cells embedded in the posterior lobe. In lower mammals, it remains well developed. The cells form lobules, and nerve fibers may grow into these groups of cells and form connections. A subset of MSH-producing corticotropes is found in the intermediate lobe of most mammals. In humans, this region is well developed before birth.

Intermediate lobe cells produce POMC; however, in lower mammals their enzymatic activities cleave the molecule to produce α-MSH, which is an important regulatory hormone for pigment-bearing melanocytes. Intermediate lobe cells may also be involved in neurogenic stress responses such as that caused by sound, although the amount of biologically active ACTH is relatively low.

Posterior Pituitary

The posterior lobe of the pituitary is largely made of nerve fibers, with supportive astrocytic glial cells, called pituicytes. This lobe was formed by the expansion of the fiber tract containing axons from neurons in the hypothalamus. This downgrowth remains connected to the hypothalamus by the infundibular stalk.

The posterior lobe has a blood supply that is largely separate from that of the anterior lobe. It stems from branches of the inferior hypophyseal artery, which arise from the internal carotids. They penetrate the posterior lobe and form a capillary bed, which then empties into the venous system and the sinus venosus.

The fibers in the posterior lobe originate from neurons in two sets of nuclei in the hypothalamus. One set includes the large neurons (magnocellular) in the paraventricular nuclei, which lie along the lateral

walls of the third ventricle. The other set of magnocellular neurons are in the supraoptic nuclei, which lie over each optic tract. The neurons in both paraventricular and supraoptic nuclei produce either oxytocin or vasopressin.

Oxytocin and vasopressin are produced and stored in secretory granules in separate neurons in most cases. The granules travel down the axons to the pars nervosa. There, they may be stored in nerve endings or in expanded regions called Herring bodies along the pars nervosa capillaries.

Oxytocin and vasopressin are nonapeptides that were discovered, isolated, and synthesized in 1953 by Vincent du Vigneaud, who won the Nobel Prize in Chemistry in 1955. Oxytocin is critical for the support of uterine contractions during labor and contractions of specialized muscle-like cells in the mammary gland during lactation. Its role in the male is not certain, although it may play a role in smooth muscle actions during ejaculation or the transport of sperm. It is also known as a bonding or mothering hormone.

Vasopressin or antidiuretic hormone (ADH) has two names because of dual functions. As ADH, it stimulates water retention by actions on the kidney collecting tubules. As a blood pressure regulator (vasopressin), it constricts arterioles. Both actions ultimately raise blood pressure. Oxytocin and vasopressin were originally derived from a common gene, called vasotocin, which is still found in lower animal species.

Embryologic Development of the Pituitary (Hypophysis)

Rathke first described the origin of the pituitary gland in 1839 and the structure that gives rise to the anterior pituitary bears his name. Rathke's pouch appears at approximately the third week of embryonic life in the human from a group of cells in the anterior part of the neural plate, which are precommitted to become the anterior pituitary, the nasal cavity ectoderm, and the olfactory placode. Regions adjacent to the anterior neural ridge in the diencephalon will become the hypothalamus, infundibulum, and posterior pituitary. These contiguous regions will begin to regulate one another and send projections of their tissue to meet and form the hypophysis.

The hypothalamus produces important factors that drive the organogenesis of the pituitary. One of these is called thyroid transcription factor or T/ebp (also called Nkx1.1 or TitfI). T/ebp stimulates the ventral diencephalon and the developing infundibulum to secrete another important growth factor, fibroblast growth factor 8 (Fgf8). Fgf8 stimulates the thickening of the pituitary placode along the roof of the mouth and then stimulates the formation and upgrowth of Rathke's pouch. Figure 3 shows this process.

Another important factor needed for full Rathke's pouch development is bone morphogenetic protein 4 (Bmp4), which is also expressed in the ventral diencephalon. Bmp4 promotes the development of the pouch and expression of one of its first early gene products, Isl1. Isl1 will stimulate the production of the first anterior pituitary hormone subunit, the α subunit, which is found in thyrotropes and gonadotropes.

Thus, the formation and upgrowth of Rathke's pouch is stimulated by the hypothalamus and infundibulum. The anterior wall of Rathke's pouch then thickens to form the adenophyphysis. Figure 3(b) shows that as Rathke's pouch is growing up and filling in with new cells, structures from the neurohypophysis grow down from the hypothalamus. The expansion of the fiber tracts forms the pars nervosa, which remains connected to the brain via the infundibular stalk. Eventually, Rathke's pouch loses its connection with the oral epithelium (Figure 3(c)).

Development of Anterior Pituitary Cells

If the gene expression discussed previously is proceeding with the correct sequence and timing, the definitive Rathke's pouch grows up and its anterior wall becomes filled in by the rapid proliferation of cells, which are the precursor cells to the anterior pituitary (Figures 3(b) and 3(c)). The ventral diencephalon produces a protein called Wnt5a, which helps with the development of the shape of the pituitary. At the same time, the oral epithelium and the ventral diencephalon produce Sonic hedgehog protein (Shh) to influence early development of the pouch as well as later regional specification that allows the differentiation of each of the cell types.

The first sets of cells to form produce the α glycoprotein subunit, shared by thyrotropes and gonadotropes. These cells form the precursor population for all other pituitary cell types. They are influenced by important transcription factors, a few of which are discussed here. Pitx1 directly activates the α subunit promoter and is important for its continued function even in the adult. It is also highly expressed in the pars tuberalis, early in development, which may relate to the concentration of basophils that are found in this region.

Another gene that is vital for the normal appearance of several pituitary cell types is *Lhx*3. In mutant mice that do not make Lhx3, Rathke's pouch can develop, but cells with α subunit, TSH, and GH do not appear and the appearance of LH is also delayed.

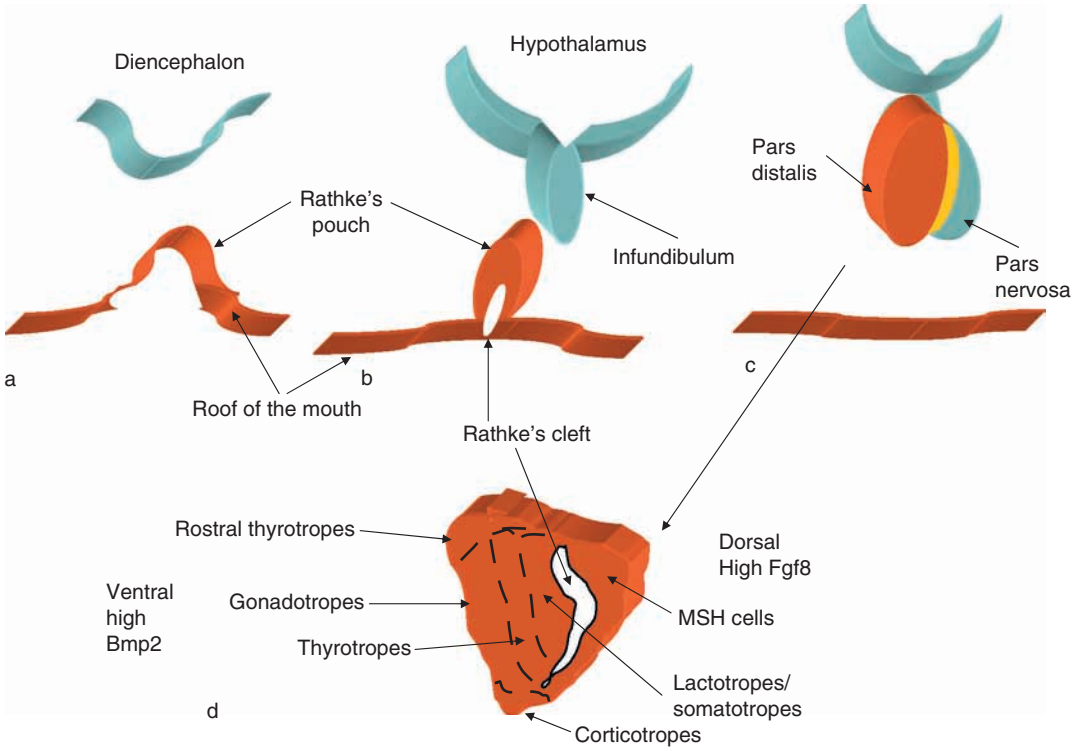

Figure 3 Embryologic development of the pituitary gland. (a) The anterior pituitary arises from the roof of the mouth as Rathke's pouch. The posterior pituitary is a downgrowth from the diencephalon. (b) Rathke's pouch continues to grow up toward the infundibulum. Its anterior (ventral) wall thickens as the anterior pituitary cells proliferate. (c) The adenohypophysis loses its connection with the roof of the mouth and lies anterior to the neurohypophysis. See **Figure 1** for the different regions of each lobe. (d) The developing pars distalis has regions of specification that allow each cell type to develop normally. The most dorsal region, attached to the pars nervosa, contains MSH cells (pars intermedia). This is separated from the remaining adenohypophysis by Rathke's cleft. The regions from ventral to dorsal contain gonadotropes, thyrotropes, and lactotropes/somatotropes. A separate population of thyrotropes (called the rostral tip thyrotropes) develop independent of Pit-1 regulation. Corticotropes develop in the opposite region. These regions are maintained by gradients of growth and transcription factors.

Humans deficient in *Lhx3* show losses in GH, prolactin, TSH, and gonadotropins.

Prop1 has been called the prophet of Pit-1 because its appearance precedes that of Pit1, another transactivation factor that is vital for the appearance of GH, prolactin, and TSH cells. The Ames mouse is an important example of a mutation in *Prop1*. Its pituitary gland is reduced by 85% and there are major reductions in gonadotropes, thyrotropes, somatotropes, and lactotropes.

Previously mentioned are some of the early genes that must be expressed for the birth and differentiation of each of the pituitary cell types. However, to make certain that each cell type develops normally, there is a regional pattern of sequestration. This allows the somatotropes, lactotropes, and most of the thyrotropes to develop in the dorsal regions of the pituitary (pointing toward the back of the body). The gonadotropes, corticotropes, and a subpopulation of thyrotropes develop in ventral regions (pointing toward the front of the body). **Figure 3(d)** shows the regions in which each of the cell types develop.

The reason for this sequestration is simple. Factors that regulate the differentiation of gonadotropes may inhibit the differentiation of somatotropes, thyrotropes, and prolactin cells. Hence, regional separation ensures that each cell type develops normally. The regions are initially formed by gradients of protein transcription factors. Early in development of the anterior pituitary there is a ventral-to-dorsal (back-to-front) graded pattern of expression of bone morphogenic hormone 2 (Bmp2), which was stimulated by the Shh secreted by cells in the roof of the mouth. In other words, regions closest to the oral epithelium have the highest concentration of Bmp2 because of their proximity to the Shh-producing cells. This Bmp2 gradient opposes a dorsal-to-ventral (back-to-front) graded pattern of Fgf8. Recall that Fgf8 was stimulated by factors coming from the infundibulum and the hypothalamus. Thus, the dorsal part of the Rathke's pouch produces more Fgf8.

The presence of high levels of Fgf8 in dorsal regions and high levels of Bmp2 in ventral regions induces specific sets of transcription factors needed for each

group of hormone-bearing cells. Fgf8 in the dorsal regions promotes the production of Prop1 and Pit1, which is important for somatotropes, prolactin cells, and a subpopulation of Pit1 responsive thyrotropes. It also represses the production of Isl1, which drives the production of α subunit.

Bmp2 in the ventral regions, along with Isl1, initially promotes the development of the α subunit and its maintenance in ventral regions. Isl1 production is also maintained in this region, and this leads to the formation of a rostral group of thyrotropes that are not dependent on Pit1 (Figure 3(d)). Eventually, Bmp2 must be inhibited to promote the development of protein factors leading to the birth of gonadotropes. One of these proteins is Gata2.

Gata2 inhibits the expression of Pit-1. Thus, to allow the development of the Pit-1-dependent somatotropes, thyrotropes, and lactotropes, Fgf8 induces the expression of Pax6, which suppresses Gata2 expression in the dorsal regions. In ventral regions, Pax6 is low and Gata2 can therefore be optimal for the development of gonadotropes. Thus, the gradient isolates the optimal environment for the selective development and differentiation of each of the cell types.

Figure 3(d) also shows that corticotropes develop from a lineage completely separate from that of other cell types. They initially appear at the ventral base of Rathke's pouch. After their numbers are enhanced by special factors, a fraction of these cells express Pitx1, which works synergistically with a factor called Tpit that is present only in POMC-expressing cells. This action initiates POMC cell differentiation and POMC transcription. Other factors involved in the enhancement of POMC cells include LIF and the NEUROD1/PanI heterodimer.

These early development events ensure that individual cell types are born and develop normally during embryonic development. Studies have demonstrated some common lineages because families of hormone-bearing cells are dependent on common transcription factors (e.g., somatotropes, thyrotropes, and lactotropes). There is interesting heterogeneity among cell types, such as the ventral- and dorsal-derived thyrotropes.

It is important to recognize that after birth, most of the anterior lobe is filled with a mixture of pituitary cells. In the human, basophils predominate in ventral and medial regions and the acidophils are more numerous in posterior and lateral regions of the pituitary. However, in other species, such as rats, these regions of concentrations are not as obvious. Furthermore, the development of multipotential cell populations suggests further modification as the animals mature after birth. Some of this modification may occur after pubertal development and may be driven by gonadal steroids. Thus, although it is clear that early developmental events support the one cell–one hormone hypothesis, it is also evident from studies of adult pituitaries that the anterior lobe retains sufficient plasticity to allow remodeling of the cell types in order to support multiple systems.

See also: Gonadotropin, Neural and Hormonal Control; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Magnocellular Neurosecretory System: Organization, Plasticity, Model Peptidergic Neurons; Neuroendocrine Aging: Pituitary Metabolism; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Neuroendocrine Aging: Pituitary-Adrenal Axis; Neuroendocrinology.

Further Reading

Antunes JL and Zimmerman EA (1978) The hypothalamic magnocellular system of the rhesus monkey: An immunocytochemical study. *Journal of Comparative Neurology* 181(3): 539–565.

Asa SL and Ezzat S (1998) The cytogenesis and pathogenesis of pituitary adenomas. *Endocrine Reviews* 19(6): 798–827.

Childs GV (1991) Multipotential pituitary cells that contain ACTH and other pituitary hormones. *Trends in Endocrinology and Metabolism* 2(3): 112–117.

Childs GV (2000) Growth hormone cells as co-gonadotropes: Partners in the regulation of the reproductive system. *Trends in Endocrinology and Metabolism* 11(5): 168–175.

Childs GV (2006) Gonadotropes and lactotropes. In: Neill J and Knobil E (eds.) *Physiology of Reproduction*, pp. 1483–1579. New York: Elsevier.

Cohen LE and Radovick S (2002) Molecular basis of combined pituitary hormone deficiencies. *Endocrine Reviews* 23(4): 431–442.

Conn PM, Huckle WR, Andrews WV, and McArdle CA (1987) The molecular mechanism of action of gonadotropin releasing hormone (GnRH) in the pituitary. *Recent Progress in Hormone Research* 43: 29–68.

Fauquier T, Lacampagne A, Travo P, Bauer K, and Mollard P (2002) Hidden face of the anterior pituitary. *Trends in Endocrinology and Metabolism* 13(7): 304–309.

Frawley LS and Boockfor FR (1991) Mammosomatotropes: Presence and functions in normal and neoplastic pituitary tissue. *Endocrine Reviews* 12(4): 337–355.

Guillemin R (2005) Hypothalamic hormones a.k.a. hypothalamic releasing factors. *Journal of Endocrinology* 184(1): 11–28.

Halmi NS (1983) Hypophysis. In: Weiss L (ed.) *Histology Cell and Tissue Biology,* 5th edn., pp. 1054–1077. New York: Elsevier.

Moriarty GC (1973) Adenohypophysis: Ultrastructural cytochemistry. A review. *Journal of Histochemistry and Cytochemistry* 21(10): 855–894.

Rivier J, Spiess J, and Vale W (1983) Characterization of rat hypothalamic corticotropin-releasing factor. *Proceedings of the National Academy of Sciences of the United States of America* 80(15): 4851–4855.

Schally AV, Coy DH, and Meyers CA (1978) Hypothalamic regulatory hormones. *Annual Review of Biochemistry* 47: 89–128.

Watkins-Chow DE and Camper SA (1998) How many homeobox genes does it take to make a pituitary gland? *Trends in Genetics* 14(7): 284–290.

Plasticity- and Activity-Dependent Regulation of Gene Expression

M R Mayford, The Scripps Research Institute, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

While we can think of plasticity in the nervous system in two broad domains, neuronal and behavioral, the two are inexorably linked in that any behavioral plasticity must arise from underlying forms of neuronal plasticity. The initial suggestion of a role for new or induced gene expression in plasticity derives from the behavioral domain. Beginning with studies in the early 1960s on the effects of pharmacological inhibitors of protein synthesis on learning and memory, it has become clear that long-lasting forms of plasticity require gene expression. The primary finding, which has been replicated in species as diverse as *Drosophila*, *Aplysia*, and mice, is that when an inhibitor of protein synthesis, such as anisomycin, is administered during the initial learning, the animals develop normal memory recall 1 h after training, but the memory is dramatically reduced when measured at 24 h. The amnestic effects of protein synthesis inhibitors have led to the suggestion that initial learning involves changes in preexisting proteins, but that the stabilization of these molecular changes over 24 h requires the synthesis of new proteins. These new proteins are presumed to be induced during the initial learning and to exert their effects of over a period of hours to consolidate the initial, transient, cellular changes that underlie short-term memory into a long-lasting or permanent form. This is supported by the fact that inhibition of protein synthesis at the later, 24 h, time point does not impact memory. The process by which short-term memories are stabilized into long-term memories, in a protein synthesis-dependent manner, is often referred to as memory consolidation, or molecular memory consolidation, to distinguish it from the more prolonged changes in the anatomy of memory that involve hippocampal function.

A number of basic questions arise from the initial observation that new gene expression is required for memory consolidation: What are the specific genes induced during synaptic or behavioral plasticity? What are the cellular mechanisms for regulating induced gene expression? What are the cellular plasticity processes that these genes act on to stabilize long-term memories? Are these induced genes required for the consolidation of long-term memory? Advances in molecular and genetic technology over the past two decades have allowed these questions to be addressed to some extent. This progress has produced an outline of transcriptional signaling mechanisms linking initial molecular and synaptic events to gene expression, neuronal plasticity, and behavioral memory.

What Are the Specific Genes Induced during Synaptic or Behavioral Plasticity?

Current models of memory formation in the mammalian brain postulate that neuronal activity-dependent forms of synaptic plasticity play a critical role. In order to look for genes that might be induced by neuronal activity, studies in the late 1980s focused on seizure-induced gene expression. The first gene identified as responding to seizure activity was the protooncogene *cfos*. The *cfos* gene is induced within minutes after the onset of seizure activity, but its expression is transient, returning to baseline levels within several hours. Subsequent screening using differential hybridization techniques or, more recently, expression profiling on DNA microarrays has identified approximately 60 genes that show rapid induction in response to high-level neuronal activity. The activity-induced genes represent a wide variety of molecular domains, from transcription factors to growth factors and components of the cytoskeleton. The identity of these genes does not provide any clear insight into common molecular mechanisms that they might subserve.

The activity-dependent genes, such as *cfos*, are induced not only by high-intensity seizure activity but also by artificial electrical stimulation and environmental stimuli. The most widely studied form of synaptic plasticity in the mammalian nervous system is long-term potentiation (LTP). LTP is an activity-dependent form of synaptic strengthening which has many features that make it an attractive candidate as a cellular mechanism of memory. LTP is typically induced by brief bouts of high-frequency stimulation, which have also been shown to effectively induce rapid expression of many of the same genes discovered in seizure screens.

At the behavioral level, many of these activity-regulated genes have been used to map neuronal activation in response to environmental stimuli. *cfos* is perhaps the most widely used marker of neuronal activity, as it has relatively low baseline expression and strong sensitivity to neuronal activity in a wide variety of brain regions. It has replaced, to a large extent, other techniques for mapping brain activation, such as 2-deoxyglucose labeling. Two other seizure-induced

genes that have been used widely for mapping activity patterns in the brain are *zif268*, encoding a zinc-finger transcription factor, and *arc*, encoding a cytoskeletal-associated protein. Both of these genes have slightly different anatomical distributions and sensitivity to neuronal activity. The properties of *arc* gene expression have allowed it to be used to assess neuronal activity at two different time points. Experiments using this approach have shown that when animals are exposed repeatedly to the same environmental contingencies they induce *arc* expression in the same neurons, indicating that the gene induction is clearly reporting environmentally relevant neuronal activity rather than nonspecific global activity.

What Are the Cellular Mechanisms for Regulating Activity-Induced Gene Expression?

The cellular mechanisms for induction of gene expression by neuronal activity have been studied extensively. While genes such as *cfos*, *zif268*, and *arc* can be induced by a variety of stimuli using different second-messenger signaling mechanisms, the primary signal in neuronal activity-induced gene expression is Ca^{2+}. The Ca^{2+} can derive from two primary sources, voltage-sensitive Ca^{2+} channels that become active when neurons are depolarized and *N*-methyl-D-aspartate (NMDA)-type glutamate receptors, which require both synaptic activity and membrane depolarization for their activity. In a typical cortical pyramidal neuron during high-frequency firing, both pathways will be active.

The transduction of this Ca^{2+} signal into transcriptional activation is significantly more complex. The majority of studies have been carried out in cell culture using artificial depolarization of neurons and using *cfos* to report gene expression. One of the difficulties in these studies is that a primary Ca^{2+} signal in neurons can activate a wide variety of second-messenger signaling pathways as shown in Figure 1. Ca^{2+} increases cAMP levels through stimulation of a Ca^{2+}-dependent form of adenylate cyclase. Ca^{2+} also activates the extracellular signal-regulated kinase/mitogen-activated kinase (ERK/MAPK) signaling pathway through ras activation. Finally, there are protein kinases that are directly activated by Ca^{2+}/calmodulin binding, such as Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) and CaMKIV. Each of these signaling pathways has been demonstrated to impact,

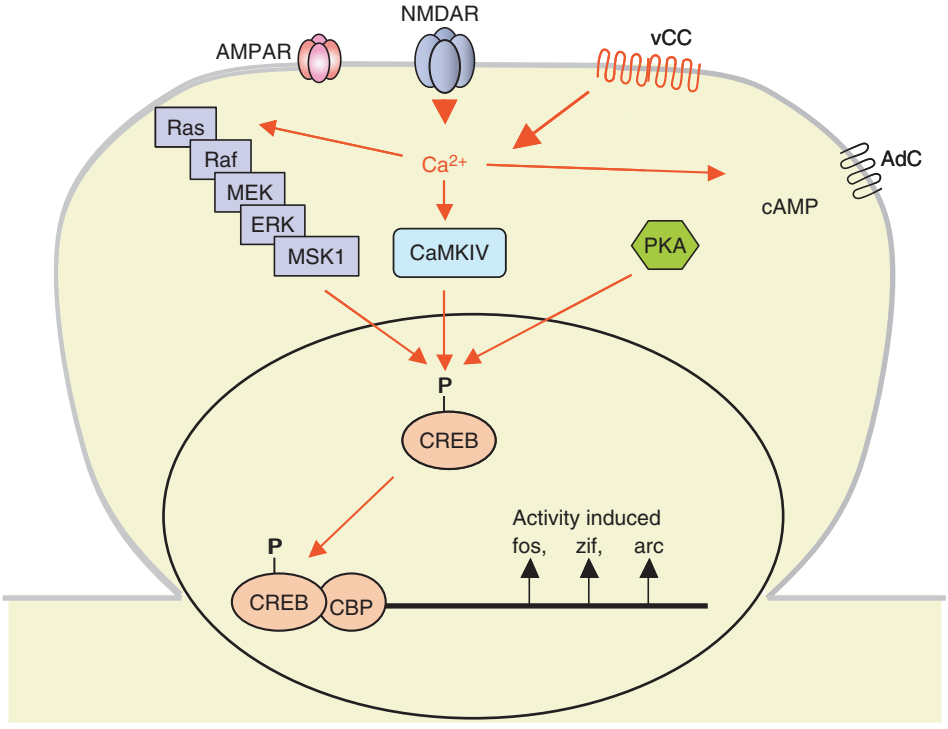

Figure 1 Molecular signaling in activity-dependent gene regulation. Neuronal activity produces an increase in Ca^{2+} primarily through the activation of *N*-methyl-D-aspartate receptors (NMDARs) and voltage-gated Ca^{2+}channels (vCCs). Ca^{2+} can be activated at least in three different second-messenger signaling pathways that converge on the nuclear transcription factor cyclic adenosine monophosphate (cAMP) response element-binding protein (CREB) to induce gene expression. AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; AdC, adenylate cyclase; CaMKIV, Ca^{2+}/calmodulin-dependent protein kinase IV; CBP, CREB-binding protein; ERK, extracellular signal-regulated kinase; MSK1, mitogen- and stress-activated kinase 1; PKA, protein kinase A.

to some degree, the induction of *cfos* by neuronal activity. The relative contribution of each of these signaling pathways to gene expression is likely dependent upon the pattern and time course of the neuronal activity-induced Ca^{2+} signal as well as the subcellular location of that signal. The interaction of these complex signaling mechanisms would in turn control the pattern of gene induction in response to given type of stimulation.

Although there are many transcription factors that respond to the various second-messenger signaling pathways, the cAMP response element-binding protein (CREB) has been studied in the most detail and plays a clear role in activity-induced gene expression. CREB is a transcription factor that was initially discovered based on its response to cAMP. Functional activation of CREB requires phosphorylation at a specific serine residue (Ser133), which can be mediated by the cAMP-responsive protein kinase A (PKA). The same serine residue is a target for phosphorylation by CaMKIV and by a downstream MAPK pathway component, mitogen- and stress-activated kinase 1 (MSK1). In this way multiple different second-messenger signaling pathways that are activated by Ca^{2+} can converge on a single target molecule to regulate transcription. The DNA-binding element for CREB is found in the promoter for the *cfos* gene, as well as in many other genes that are regulated by neuronal activity. This element alone, in a synthetic promoter construct, can confer responsiveness to neuronal activity to cis-linked genes.

What Are the Cellular Processes That These Genes Act on to Stabilize Long-Term Memories?

If new gene expression is required for the consolidation of long-term memories, what might be the cellular processes that underlie its role in memory? This is a component of the larger, and not clearly answered, question of what cellular mechanisms underlie memory. Initial progress on both of these questions has been made in invertebrate model systems such as *Aplysia*. Kandel and his colleagues have analyzed a form of simple learning in this marine mollusk at the behavioral, circuit, cellular, and molecular levels. Sensitization memory in *Aplysia* shows the same dependence on translation, for the conversion from short-term to long-term forms, that is seen in many other species and behavioral paradigms. The underlying cellular changes for this memory appear to be an increase in synaptic efficacy in the specific motor circuit that controls the behavior. In a critical series of experiments, Kandel showed that the synaptic change had the same sensitivity to protein and RNA synthesis inhibitors as did the behavioral memory. That is, when protein synthesis was inhibited during the induction of synaptic plasticity, a short-term form of the plasticity occurred but lasted only a few hours. In the absence of protein synthesis blockade, the synaptic plasticity is stabilized and maintained for more than 24 h. This striking concordance between behavioral memory and a cellular plasticity mechanism suggests that this may be the means by which memory consolidation functions at the cellular level. It also provided the first suggestion that gene expression changes may be critical for neuronal as well as behavioral forms of plasticity.

The cellular plasticity in *Aplysia* is not activity dependent but is mediated by serotonin-induced activation of adenylate cyclase. So while the plasticity in *Aplysia* does not require neuronal activity, it recruits one of the same second-messenger signaling pathways that is activated in the induction of *cfos* by action potential firing in mammalian neurons. *Aplysia* contains CREB homologs that act similarly to mammalian CREB. Blocking CREB-dependent transcription in *Aplysia* neurons produces an effect similar to blocking all transcription or translation, in that the long-lasting synaptic plasticity is inhibited while the short-lasting form is unaffected.

In mammals, the neuronal activity-dependent form of synaptic plasticity, LTP, is commonly studied in the hippocampus, either in the intact brain with implanted electrodes or *in vitro* in brain slices. LTP is induced artificially by high-frequency stimulation, resulting in a prolonged increase in the strength of the stimulated synapses that, in the intact brain, can last for weeks. Most of the molecular and pharmacological studies of LTP have focused on the brain slice preparation. In this preparation, if transcription or translation is pharmacologically blocked during the high-frequency stimulation used for induction, there is an effect similar to that discussed previously for *Aplysia* synaptic plasticity – a short-lasting form of LTP is produced, which then returns to baseline over several hours.

The nature of the transcriptional requirement for LTP has been investigated in an interesting series of experiments by Morris and Frey; their findings suggest that the products of activity-induced gene expression in the nucleus are functionally delivered only to synapses that were recently active. This process is sometimes referred to as 'synaptic tagging' and is supported by the following observations. If synapses are stimulated weakly, this induces a short-lasting potentiation that is not dependent on new gene expression. However, if those same neurons were previously stimulated strongly to induce gene expression, then the short-lasting potentiation is converted into a long-lasting LTP. The suggestion is that the

weak stimulation has molecularly altered or tagged the synapses in such a way that they can now make use of activity-induced genes, if the genes are available within the cell. The strong stimulation of the neurons, at a distinct set of synapses, provides that transcriptional signal.

Advances in techniques for genetic manipulation of the mouse, such as gene knockout and transgenesis, have allowed the functional role of specific molecules to be tested in the intact mammalian brain. The role of CREB has been tested using this approach. Mice in which the CREB gene has been deleted or inhibited through expression of a dominant-negative transgene show deficits in long-lasting LTP, while the short-lasting form remains intact. Several of the genes that are rapidly induced by neuronal activity, most notably those for the transcription factor zif268 and the cytoskeletal-associated protein arc, have been knocked out and produce similar LTP phenotypes.

These results demonstrate that in both vertebrate and invertebrate systems, forms of synaptic plasticity that are thought to underlie long-term memory formation show requirements for new gene expression that parallel what is seen at the behavioral level for long-term memory. While this provides a suggestive framework for how one might understand the role of activity-induced gene expression in long-term memory (i.e., the induced genes are necessary to stabilize the synaptic plasticity that in turn stabilizes the memory), this is certainly an oversimplification. It is clear from many studies that there is not a simple relationship between LTP and behavioral memory. LTP is an artificially induced phenomenon of which there are a number of different types, some of which do not require new gene expression. Moreover, there are a number of examples in which transcriptionally dependent forms of LTP can be impaired genetically or pharmacologically while memory remains relatively intact.

Are These Activity-Induced Genes Required for the Stabilization of Long-Term Memory?

The generation of genetically altered mice carrying mutations in the various activity-induced genes or in the signaling molecules important for their induction has allowed the possibility for linking specific genes to behavioral memory. One of the first studies of this type came from *Drosophila*, which, like the other species discussed, shows a requirement for new protein synthesis for the stabilization of long-term memory. Genetic studies in *Drosophila* have identified genes involved in the cAMP signaling pathway in the formation of memory using a simple Pavlovian conditioning paradigm. Using dominant-negative inhibitors of CREB, Tim Tully and his colleagues showed that expression of the CREB inhibitor impaired long-term memory while sparing short-term memory. Strikingly, they went on to show that expression of a CREB activator allowed training paradigms that normally produced only short-term memory to then induce a long-lasting memory. These results suggest that CREB, which, as discussed previously, is a critical mediator of neuronal activity-induced gene expression and long-lasting forms of synaptic plasticity, is also a critical regulator of the conversion of short-term into long-term forms of behavioral memory.

Similar studies have been carried out in mice. Alcino Silva and his colleagues studied a line of mice in which CREB expression was reduced through a genetic manipulation. They found that long-term memory was selectively impaired in these mice. However, the interpretation of phenotypic alterations, particularly behavioral alterations, in genetically modified mice, in which the genetic lesion is present for the entire developmental history of the animal, can be complex. For example, is the phenotype directly due to the loss of the gene of interest or to compensatory changes in the expression of other genes, or in the development of the circuitry underlying the behavior? CREB is a member of a class of related transcription factors and it was found that in the CREB mutant mice there was a compensatory upregulation in other members of this gene family. Was the phenotype indirectly due to this change in the balance of expression in CREB family members?

To address these issue, a number of groups have used inducible genetic manipulations. In this case, the animals can be allowed to develop normally and the genetic change is induced in the adult just prior to behavioral testing. This should preclude any developmental abnormalities and minimize the chances for compensatory changes that could alter or mask the phenotype. This approach was taken for CREB signaling using inducible expression of a dominant-negative transgene that targeted broadly the CREB family of transcription factors. Even when CREB was inhibited as little as 6 h prior to behavioral training, the same pattern of memory impairment, short-term intact/long-term impaired, was observed. Thus, in both *Drosophila* and mammals, using the most precise genetic manipulations available, it appears that the transcription factor CREB is a critical mediator of the stabilization of short-term into long-term memory.

Transcription is not a simple process involving single molecules, but requires the recruitment of many proteins into a complex on the DNA. Transcription is also sensitive to chromatin structure and the covalent

modification of DNA and proteins within chromatin. Among the important components of this process are the transcription coactivators, which serve to link specific transcription factors like CREB to the recruitment of the more general transcriptional machinery. In addition, these molecules can serve to alter the structure of chromatin through the covalent modification of histones to a state that facilitates transcription. One of these transcription coactivators is the CREB-binding protein (CBP). The CBP gene has drawn interest since it was identified as the underlying genetic alteration in Rubinstein–Taybi syndrome, a rare disorder in humans that is characterized by both developmental and cognitive abnormalities. Two recent studies examined CBP function in behavior using inducible genetic modification. It was found that even when the gene was inhibited only in adult animals, it caused a deficit in long-term memory similar to that seen with CREB inhibition. This suggests that some of the cognitive phenotypes observed in the human disorder may be due to an acute requirement in the adult for normal CBP function.

In addition to linking CREB to the recruitment of the general transcriptional machinery, CBP functions as an enzyme to modify chromatin structure by acetylating histones. This histone acetylation allows the chromatin to assume an open structure needed for effective transcription. Surprisingly, the behavioral effects of genetic CBP inhibition could be overcome with pharmacological agents that elevate histone acetylation nonspecifically. This demonstrates not only the critical role for chromatin structure modifications in the transcriptional regulation of long-term memory, but also suggests a possible pharmacological intervention in disorders, such as Rubinstein–Taybi, that result from disruption of this process.

The role of CREB-based transcription in the stabilization of long-term synaptic and behavioral plasticity is clear. What of the signaling pathways that activate CREB and the 60 or so activity-induced genes that have their expression mediated by CREB? Here the players are more numerous and the results less clear. Mutant mice have been engineered to disrupt each of the three major upstream signaling pathways to CREB activation (Figure 1) – Ca^{2+}-dependent adenylate cyclase, ERK/MAPK, and CaMKIV – with mixed results. Complete elimination of the Ca^{2+}-responsive cyclase leads to a phenotype in both LTP and behavioral memory that is similar to that seen with protein synthesis inhibitors, but not a complete phenocopy. Knockout of CaMKIV leads to a strong reduction in *cfos* inducibility and a deficit in protein synthesis-dependent LTP lasting more than 1 h. Surprisingly, however, the knockout fails to produce similar major changes in long-term behavioral memory. Finally, the ERK/MAPK signaling pathway contains the most components and is the most difficult to dissect. Pharmacological inhibitors of ERK/MAPK signaling have been shown in many studies to disrupt memory and synaptic plasticity. Mutation of an upstream effector of ras function (the neurofibromatosis gene, *NF1*) similarly produces deficits in long-term memory. However, mutations of some of the more downstream effectors have produced conflicting results. This type of complexity is not surprising given the diverse interactions between these very general cellular signaling pathways. As discussed previously, the induction of gene expression likely results from the contribution of each of these signaling pathways, and the impairment of one pathway might be expected to lead to compensation by the others under certain conditions.

The activity-induced genes are numerous and their individual roles in behavioral plasticity have been less well studied. Three of the activity-induced genes that appear to have an influence on memory are the transcription factor gene *zif268*, the brain-derived neurotrophic factor gene *BDNF*, and the cytoskeletal-associated protein gene *arc*. In the case of *zif268* and *arc*, mice carrying a deletion in the gene are viable to adulthood, allowing behavioral testing in which the animals show the selective impairment in long-term memory with sparing of short-term memory that is characteristic of protein synthesis inhibitors. The deletion of the *BDNF* gene produces a severe developmental phenotype. However, one recent study used antisense oligonucleotides to block *BDNF* expression specifically in adult animals and found a similar phenotype. Thus, the individual disruption of three neuronal activity-induced genes, representing distinct categories of molecular function, produces behavioral phenotypes that are similar to that produced with the general blockade of new gene expression.

Summary and Future Directions

A simple framework has been provided for understanding the role of gene expression in neuronal and behavioral plasticity. In summary, the transcription factor CREB plays a critical role, integrating multiple upstream signaling pathways into the increased transcription that is necessary for the stabilization of long-lasting synaptic plasticity and memory. In invertebrates CREB functions primarily by responding to direct cAMP signals. In the mammalian nervous system the critical role of CREB is in integrating multiple molecular signaling pathways that are activated by neuronal activity. Each of these pathways is stimulated by a primary Ca^{2+} signal that occurs when neurons are depolarized, allowing activation of both

NMDA receptors and voltage-dependent Ca^{2+} channels. The response to neuronal activity involves the transcriptional activation of about 60 different genes with diverse functions.

The signaling pathways and the molecular and temporal dynamics of activity-induced gene expression are reasonably well understood. The clear deficiency is in understanding the cellular processes on which these genes act. The questions for the future concern how the activity-induced genes effect synapses, how they stabilize synaptic changes, and whether these synaptic changes completely account for their role in long-term behavioral plasticity.

See also: Aging and Memory in Animals; Aging and Memory in Humans; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Episodic Memory: Assessment in Animals; Episodic Memory; Event-Related Potentials (ERPs); Humans; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Memory Representation; Memory: Genetic Approaches; Neuronal Plasticity after Cortical Damage; Strategic Control of Memory; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Plasticity: Neuronal Sprouting; Synaptic Plasticity: Short-Term Mechanisms; Working Memory: Capacity Limitations.

Further Reading

Davis HP and Squire LR (1984) Protein synthesis and memory: A review. *Psychological Bulletin* 96: 518–559.

Kandel ER (2001) The molecular biology of memory storage: A dialogue between genes and synapses. *Science* 294: 1030–1038.

Shaywitz AJ and Greenberg ME (1999) CREB: A stimulus-induced transcription factor activated by a diverse array of extracellular signals. *Annual Review of Biochemistry* 68: 821–861.

West AE, Griffith EC, and Greenberg ME (2002) Regulation of transcription factors by neuronal activity. *Nature Reviews Neuroscience* 3: 921–931.

Plasticity of Intrinsic Excitability

R Mozzachiodi and J H Byrne, The University of Texas Medical School, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During the past several decades, analyses of the cellular and molecular mechanisms underlying memory formation and maintenance have led to the widespread view that memory storage is largely due to persistent experience-dependent changes in synaptic efficacy. However, an accumulating body of work indicates that learning also leads to persistent changes in intrinsic neuronal excitability. These changes are due to modifications of membrane conductances and can affect electrophysiological properties, including the resting potential, the input resistance, the shape and the threshold of the action potential, and the discharge frequency.

In this article, we review some of the earlier work illustrating plasticity of intrinsic excitability produced by experience, together with some more recent findings. In addition, we discuss the implications of these results and the extent to which changes in synaptic efficacy and changes in intrinsic excitability can both contribute to memory storage.

Changes in Intrinsic Excitability Produced by Learning and Experience

Research over the past three decades has discovered modifications in the intrinsic membrane properties occurring in invertebrates and vertebrates following different forms of nonassociative and associative learning.

Invertebrate Models

Alkon and collaborators were the first to report that classical conditioning (i.e., the learned ability to associate a predictive stimulus with a subsequent salient event) altered the excitability of the photoreceptors in the nudibranch mollusk *Hermissenda crassicornis*. The positive phototaxis that the animal normally exhibits in response to illumination was reduced when the animal was trained with multiple paired presentations of light (conditioned stimulus; CS) and rotation (unconditioned stimulus; US), which activates the vestibular system. This form of learning was associative and the suppression of phototaxis, the conditioned response, persisted for weeks. Classical conditioning produced changes in several membrane properties of the type B photoreceptors, which are the primary sensory neurons of the CS pathway. The type B photoreceptors are excited by light and also by synaptic drive from the statocyst cells, which are part of the vestibular system activated by rotation. The changes produced by classical conditioning included increases in input resistance, spontaneous firing and spike activity in response to current injection, and generator potentials evoked by the CS. Importantly, these changes persisted when the synaptic input to the photoreceptors was removed. Voltage-clamp analysis of the type B photoreceptors revealed that pairing-specific changes in several ionic currents were the biophysical substrate of the altered membrane properties just mentioned. Classical conditioning decreased two K^+ currents, a rapidly activating and inactivating K^+ current (I_A) and a Ca^{2+}-activated K^+ current ($I_{K,Ca}$), and also increased a voltage-dependent Ca^{2+} current. These biophysical changes were consistent with the increase in excitability observed in the type B photoreceptors following classical conditioning, because a reduction of I_A and $I_{K,Ca}$ would tend to produce an enhanced depolarization and increased spike activity in the photoreceptors in response to the presentation of light. The pairing-specific increase in intrinsic excitability was observed for several days following conditioning. Also, the magnitude of the increase in excitability was positively correlated with the degree of phototactic suppression, thus suggesting a causal relationship between the biophysical changes in the type B photoreceptors and the expression of the conditioned response.

Following the results observed in *Hermissenda*, other learning-related changes in intrinsic excitability were identified in other invertebrates, including annelids and mollusks. In the leech *Hirudo medicinalis*, habituation and sensitization, two forms of nonassociative learning, altered in opposite directions the intrinsic excitability of an identified neuron. *Hirudo* exhibits a defensive withdrawal reflex that consists of whole-body shortening in response to light touch of the skin. This reflex can be habituated, when it is repetitively evoked, or sensitized, when a strong, noxious stimulus is delivered. Sensitization, which is mediated by serotonin (5-hydroxytryptamine; 5-HT), requires the activity of a group of electrically coupled interneurons called the S cells. Sensitization increased the input resistance of the S cells as well as their spike threshold and the number of action potentials evoked by intracellular depolarization. These effects were mimicked by exogenous application of 5-HT. Conversely, habituation produced a decrease in intrinsic excitability. These results indicate that in the leech bidirectional changes in intrinsic excitability represent

cellular correlates of two distinct forms of nonassociative learning.

The marine mollusk *Aplysia californica* has been probably the most extensively used invertebrate model to study the cellular and molecular mechanisms of learning and memory. Several examples of experience-dependent changes in intrinsic excitability have been reported in *Aplysia* following both nonassociative and associative learning of simple defensive reflexes as well as associative learning of more complex behaviors such as feeding. Long-term (24 h) sensitization of the tail-siphon withdrawal reflex was associated with facilitation of the sensorimotor neuron synapse as well as changes in the biophysical properties of the tail sensory neurons. The intrinsic changes included increases in the excitability and in the after-depolarization following either a single spike or a burst of action potentials. Voltage-clamp analysis revealed that long-term sensitization reduced the net outward current in the tail sensory neurons, which is consistent with the increased excitability observed in these neurons following learning. Interestingly, long-term sensitization training also altered the biophysical properties of tail motor neurons. These changes included a more hyperpolarized resting membrane potential and a decrease in the threshold for spike initiation. The changes produced by long-term sensitization in the sensory neurons resembled those induced by short-term sensitization, thus suggesting that the phenotype of the expression mechanism for the short-term memory is preserved for the long-term memory.

In *Aplysia*, classical conditioning also produced changes in the intrinsic excitability of sensory neurons. In a simplified preparation of the *Aplysia* siphon-withdrawal reflex, classical conditioning led to an increased input resistance and excitability of the siphon sensory neurons. Interestingly, the changes in intrinsic excitability, as well as the pairing-specific strengthening of the sensorimotor neuron synapses, were restricted to the sensory neurons that were activated by mechanical stimulation of the siphon (i.e., on-field neurons), thus providing CS specificity to the expression of the conditioned response. The presence of biophysical as well as synaptic changes following both nonassociative and associative learning suggested the intriguing hypothesis that the same cellular phenotype underlying nonassociative learning may also serve as physiological substrate for associative memory.

In *Aplysia*, the cellular mechanisms of associative learning of complex behaviors were also studied using appetitive forms of classical and operant conditioning (i.e., the learned ability to associate an expressed behavior with the consequences). Both operant and classical conditioning altered the biophysical properties of neuron B51, a key element of the feeding neural circuitry, but the changes were remarkably different. Operant conditioning increased the input resistance and decreased the burst threshold of B51, resulting in an overall increase in intrinsic excitability, whereas classical conditioning did not alter the input resistance, but increased the burst threshold of B51, resulting in a decrease in intrinsic excitability. These findings provided a biophysical substrate to the hypothesis that, at the cellular level, operant and classical conditioning are mediated by fundamentally different mechanisms. The decreased excitability in B51 was counterintuitive because classical conditioning enhances feeding in response to the CS and also strengthens the CS-evoked excitatory drive to B51. However, the analysis of the responsiveness of B51 to the CS after training revealed that classical conditioning indeed facilitated the recruitment of B51, indicating that the factors that enhanced the recruitment of B51 overpowered the diminished excitability. Although experimental evidence is needed, the pairing-specific decrease in the excitability of B51 might represent an adaptive mechanism to help shape the CS specificity produced by classical conditioning.

Although most of the effects of learning on the membrane properties involved changes in input resistance and/or altered threshold for action potential initiation resulting in modulation in intrinsic excitability, other forms of intrinsic modifications were described as a result of experience in mollusks. For example, as mentioned previously, long-term sensitization in *Aplysia* was associated with an increase in the resting potential of the motor neurons. In the pond snail *Lymnaea stagnalis*, long-term memory for appetitive classical conditioning of feeding, induced by repetitive paired presentations of mechanical stimulation of the lips (CS) and application of sucrose as a feeding stimulant (US), was associated with a persistent depolarization in the modulatory neuron CV1a involved in the initiation of feeding movements. The pairing-specific depolarization of CV1a was positively correlated with the fictive feeding response to the CS in reduced preparations from conditioned animals, but it was not associated with any change in other membrane properties such as input resistance, spike threshold, or spike frequency. The time course of the persistent depolarization also mirrored the duration of retention of long-term memory. This pairing-specific depolarization would render CV1a more likely to fire in response to the CS and allow the CS to more effectively generate feeding responses.

Vertebrate Models

The first evidence of experience-dependent changes in neuronal excitability in a vertebrate model system was reported by Brons and Woody in 1980. These authors used a conditioning paradigm in the cat in which repetitive paired presentations of an auditory click (CS) and a glabella tap (US), which evoked an eyeblink response, led to the appearance of the conditioned response consisting of a combined eyeblink and nose twitch. Cellular analysis conducted in the sensory–motor cortical areas and in the facial nucleus of awake cats revealed that classical conditioning was associated with an increased excitability as measured by the minimum amount of injected current necessary to evoke an action potential. An increase in input resistance was also measured in neurons of the facial nucleus. Although the increase in neuronal excitability lasted for at least 4 weeks following training, the difference between conditioned and control animals became significant only after 25 days. Also, the persistence of the increased excitability, once the conditioned response was extinguished, indicated that this mechanism was not directly implicated in the formation and retention of the conditioned response. A detailed analysis of the biophysical substrate of the increased excitability has not been pursued.

Several other examples of experience-dependent changes in intrinsic excitability were identified in vertebrate models. In the rabbit, delay eyelid conditioning, which is performed using a tone as CS and a coterminating air puff as US, depends on the activation of both the cerebellar cortex and cerebellar deep nuclei. Recordings from Purkinje cells revealed a reduced spike threshold and a reduced afterhyperpolarization (AHP) evoked by burst of spikes in slices from conditioned animals when compared to control animals. The AHP that follows action potentials is an important determinant for the firing activity of neurons, and a pairing-specific reduction of its amplitude is consistent with an increased excitability observed following classical conditioning. Importantly, these changes were long-lasting (up to 30 days), restricted to defined microzones of the cerebellar lobule HVI, and the degree of reduction in spike threshold positively correlated with the conditioned response in the paired animals. Another form of classical conditioning is trace eyelid conditioning, which requires the hippocampus, and occurs when a trace interval is imposed between the CS offset and the US onset. Cellular analysis of trace eyelid conditioning revealed that hippocampal pyramidal neurons in the CA1 and CA3 areas from conditioned animals exhibited a greater number of spikes evoked by depolarizing current injections and a decreased AHP due to a reduction in the Ca^{2+}-dependent K^{+} current underlying the AHP, as compared to control animals. These changes were consistent with a pairing-specific increase in excitability. Although the increased excitability was similar to that found following delay conditioning, it was not restricted to a specific area, but it was instead widespread through the dorsal hippocampus, and also could not be detected 7 days after training when the memory for trace eyelid conditioning was still retained. These findings indicate that it is unlikely that this transient increased excitability in hippocampal neurons constitutes a part of the memory trace itself, but it could represent a mechanism through which the hippocampal circuit is set in a more permissive state for input-specific synaptic modification to occur during memory formation. A similar learning-dependent reduction in AHP also occurred in CA1 pyramidal neurons in the dorsal hippocampus following a spatial learning task that depends on the hippocampus.

In rats, pairing-specific enhancements in both CS-evoked synaptic drive and neuronal excitability (i.e., increased input resistance and reduced amount of injected current necessary to elicit an action potential) were measured *in vivo* in neurons from the lateral nucleus of the amygdala of anesthetized rats trained with a conditioning paradigm during which an odor (CS) was repetitively paired with a footshock (US). The synaptic and the intrinsic changes were both prevented when dopamine signaling, which is relevant for amygdala-dependent forms of learning, was pharmacologically blocked in the lateral nucleus of the amygdala. These results are important because they indicate that both synaptic and intrinsic plasticity occur together during memory formation.

Changes in intrinsic excitability have been found also in vertebrates following operant conditioning tasks. In monkeys, operant conditioning of the spinal stretch reflex is associated with changes in the intrinsic properties of the motor neurons. The spinal stretch reflex is among the simplest vertebrate reflexes and is largely mediated by a monosynaptic pathway between Ia afferent neurons and alpha motor neurons. Because this reflex is influenced by descending activity from supraspinal structures, it can be operantly conditioned and animals can learn to gradually increase or decrease this reflex or its electrical analog, the H-reflex. In particular, operantly conditioned decrease in the H-reflex in monkeys was associated with an increase in the depolarization needed for the spinal motor neuron to fire and a decrease in its conduction velocity. These apparently intrinsic changes were consistent with the hypothesis that a positive shift in motor

neuron firing threshold and a consequent increase in the depolarization needed to reach that threshold could contribute to weaken the magnitude of the H-reflex following operant conditioning. Recent modeling studies proposed that a positive shift in the activation of voltage dependence of Na^+ channels might represent the biophysical substrate for the operantly conditioned decrease in the H-reflex.

Changes in excitability produced by operant conditioning were also reported in rats following an olfactory discrimination task. Rats were trained to discriminate pairs of odors for a water reward. Several consecutive days of training were required for a rat to successfully discriminate between a first pair of odors, but once the rat reached good performance, its ability to learn to discriminate between a new pair of odors improved dramatically and only 1 day was needed to achieve good performance. This phenomenon is called rule learning, because the animal acquires a rule to perform rapidly and efficiently in the olfactory task that is independent of the memory of any particular odor. Cellular analysis revealed that spike frequency adaptation and AHP amplitude were both reduced in pyramidal neurons of the rat olfactory cortex following olfactory discrimination training when compared to naive or pseudo-trained animals. These changes were detected 1–3 days after training, but they decayed to baseline values by 5–7 days after training. It has been speculated that this example of increased excitability is not *per se* a mechanism of memory storage, but it might be involved in rule learning, setting the neural circuits in the piriform cortex in an excitable, more permissive state for activity-dependent synaptic modifications to occur.

Activity-Dependent Modulation of Intrinsic Excitability

In addition to learning tasks *in vivo*, intrinsic excitability can be modulated also by patterns of electrical stimulation of neurons and neural pathways. Depending on the protocol of stimulation, some of these patterns of neuronal activity may induce either an enhancement or reduction of synaptic strength, termed long-term potentiation (LTP) and long-term depression (LTD), respectively. LTP, which is probably the most studied example of activity-dependent modulation of synaptic efficacy, can be induced in several areas of the vertebrate brain, including the hippocampus, the amygdala, the neocortex, and the cerebellar cortex, and is commonly considered to be a mechanism underlying aspects of learning and memory. Because several forms of LTP and LTD can be induced in isolated brain slices, the biochemical cascades underlying these forms of neuronal plasticity can be analyzed.

Although most of the work on LTP has focused on the mechanisms underlying the persistent augmentation of synaptic efficacy, several lines of evidence indicate that, at least in some cases, LTP is also accompanied by modifications in intrinsic excitability. Bliss and Lømo, who first described LTP in the dentate area of the rabbit hippocampus in 1973, also reported an increase in the population spike (i.e., the extracellularly recorded signal representing the summation of the evoked action potentials in postsynaptic neurons) accompanying LTP, which could not be entirely explained by the LTP-evoked increase in the population excitatory postsynaptic potential (EPSP). This nonsynaptic component of LTP was termed EPSP-to-spike potentiation, or E-S potentiation. The mechanisms underlying E-S potentiation are still subject to some debate. Two hypotheses have been developed to explain this phenomenon: an LTP-induced decrease in the ratio of inhibitory to excitatory drive, and an LTP-induced increase in the intrinsic excitability of the postsynaptic neurons through modulation of voltage-dependent channels. Several lines of evidence indicate that a reduction in the ratio of inhibitory to excitatory drive is not sufficient to account for the E-S potentiation, thus pointing to a role for LTP-induced intrinsic plasticity, the consequence of which would lead to a higher efficiency of EPSP-to-spike coupling. One example, which suggests the modulation of voltage-dependent channels as a mechanism underlying the increased excitability accompanying synaptic LTP, has been reported in hippocampal CA1 pyramidal neurons. In these neurons, a pattern of synaptic inputs paired with postsynaptic spikes induced long-term-increase of intrinsic excitability concurring with synaptic LTP. The increased excitability was manifested as a decrease in the action potential threshold that was attributable to a shift in the activation curve of voltage-gated Na^+ channels. This form of enduring increase of intrinsic excitability shared a similar signaling pathway with the late phase of synaptic LTP, which included the requirement of activation of glutamate receptors of the *N*-methyl-D-aspartate (NMDA) type, the influx of Ca^{2+}, the activity of Ca^{2+}/calmodulin (CaM) kinase II, and changes in protein synthesis.

Interestingly, the modulation of the EPSP-to-spike coupling appears to be bidirectional. In the CA1 region of the hippocampus, LTD, normally induced by low-frequency patterns of electrical stimulation, was accompanied by a reduction of the E-S coupling. This reduction of the E-S coupling was largely due to a decrease in the intrinsic excitability of hippocampal

cells. Therefore, opposite modulation of the synaptic strength is associated with bidirectional changes in E-S coupling, which reflect bidirectional modifications of intrinsic excitability.

Other examples of activity-dependent changes in intrinsic excitability produced by synaptic activation have been reported in different brain areas. For example, in the rat entorhinal cortex, a brain region involved in the acquisition and retention of working memory, persistent graded increases in firing frequency were induced in pyramidal neurons by repetitive excitatory synaptic activation. These sustained levels of firing could be either increased or decreased in an input-specific manner and relied on activity-dependent changes of Ca^{2+}-dependent cationic current. This intrinsic ability of the neurons in the entorhinal cortex to generate graded persistent activity has been proposed as a cellular mechanism for working memory. Another example of modulation of intrinsic excitability induced by synaptic activation has been reported in the cerebellar cortex. In this brain region, high-frequency stimulation of the mossy fiber-to-granule cells pathway led to LTP and was also accompanied by a persistent enhancement of intrinsic excitability of granule cells, which was associated with increased input resistance and decreased spike threshold. Neurons of the cerebellar deep nuclei also exhibited a rapid, synaptically driven, increase in their intrinsic excitability, which required Ca^{2+} influx through activation of NMDA-type glutamate receptors. However, in this case, the pattern of synaptic stimulation used to activate these neurons did not produce any facilitation of the synaptic input. This latest example indicates that modulation of intrinsic excitability can be expressed without the occurrence of changes in synaptic function, thus providing a contribution to the information storage independent from synaptic plasticity.

Recent findings indicate that changes in intrinsic excitability may not require synaptic activation. For example, a train of high-frequency intracellular depolarizations delivered to pyramidal neurons in layer V of the primary visual cortex produced an enduring increase in intrinsic excitability. This change in excitability required Ca^{2+} entry during neuronal activity and did not affect the passive membrane properties, thus suggesting the involvement of voltage-dependent channels. Another example of synaptic-independent modulation of intrinsic excitability has been reported in hippocampal CA1 pyramidal neurons. In these neurons, a pattern of suprathreshold repeated intracellular stimulation, called theta-burst firing, produced a decrease in somatic excitability, which was due to an upregulation of the hyperpolarization-activated cationic current I_h. This decrease in excitability was prevented by Ca^{2+} chelators, by low concentrations of tetrodotoxin (which blocked backpropagation of action potentials from the soma to the dendrites), and by NMDA receptor antagonist even in the absence of synaptic activation. These findings suggest that during theta-burst firing, backpropagating action potentials would lead to Ca^{2+} influx through the NMDA receptor, which, in turn, would increase the I_h via a mechanism mediated by CaM kinase II.

Interestingly, recent findings indicate that the synaptic drive needs not be excitatory to induce an increase in intrinsic excitability. Recordings from spontaneously firing neurons in the medial vestibular nucleus indicate that brief periods of inhibitory synaptic input, or direct membrane hyperpolarization, triggered a long-lasting increase in both spontaneous firing rate and firing responses to intracellular depolarization. This increase in excitability, termed firing rate potentiation, was due to a decrease in cytosolic Ca^{2+}, which reduced CaM kinase II activity, maximal in spontaneously active vestibular neurons, and, in turn, downregulated BK-type $I_{K,Ca}$. This novel form of neuronal plasticity might contribute to motor learning in the vestibuloocular reflex.

The intrinsic changes produced by patterns of neuronal activity are not restricted to ion channels. Modulation of the activity of membrane transporters has also been reported. For example, interneurons of the rat dentate gyrus exhibited a long-term depolarization of their resting membrane potential after high-frequency stimulation of the perforant path that was attributed to an activity-dependent change in the rate of the electrogenic Na^+ pump function. This activity-dependent depolarization, which occurred in the absence of any potentiation of the excitatory synaptic input, required the rise of intracellular Ca^{2+} and the activation of glutamate receptors of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) type, but not of those of NMDA type. As a result of the depolarization, the interneurons of the dentate gyrus responded with action potential discharge to previously subthreshold EPSPs even in the absence of synaptic potentiation, thus indicating an increase in the E-S coupling. Activity-dependent changes in the function of the electrogenic Na^+ pump have also been found in invertebrate neurons. In *Hirudo*, low-frequency repetitive stimulation of touch (T) sensory neurons led to a lasting increase in the amplitude of the AHP produced by the firing discharge, which is due to an increase in the activity of the electrogenic Na^+ pump. The modulation of the Na^+ pump activity was regulated by both the influx of Ca^{2+} during neural activity and the release of Ca^{2+} from intracellular

stores. In addition, phospholipase A2 and the downstream activation of arachidonic acid metabolites, derived from the 5-lipoxygenase pathway, were necessary for the increase of the AHP amplitude. Interestingly, the increase of the AHP amplitude was associated with a persistent depression of the synaptic connection between T cells and their follower neurons. This synaptic depression may be a cellular mechanism contributing to short-term habituation.

Plasticity of Intrinsic Excitability as a Mechanism for Memory Storage: Hypotheses and Lines of Evidence

The general motif that emerges from the preceding examples is that several different learning tasks, nonassociative and associative, as well as activation of neurons and neural pathways, induce changes in the biophysical properties of neurons in both invertebrate and vertebrate model systems. These neurophysiological correlates have been identified at the level of sensory neurons, interneurons, and motor neurons.

What is the functional relevance of these changes? In some case a direct role for the intrinsic plasticity in memory storage could be proposed. For example, in *Aplysia* and *Hermissenda,* because the loci of plasticity induced by classical conditioning were found at the level of the first central relay in the circuits controlling the responses, neuron-wide changes in intrinsic excitability in response to the conditioned stimuli appear to be an appropriate memory mechanism to express and retain the conditioned response. However, in other cases the changes in intrinsic excitability failed to correlate, or correlated poorly, with the learned behavioral changes, thus making it difficult to assess their contribution to the storage of memory. Although additional results are required, one hypothesis is that some of the intrinsic changes may not function as part of the engram itself, but they may represent either adaptive mechanisms to shape the stimulus specificity of the learned response (as in the case of classical conditioning of feeding in *Aplysia*), or mechanisms through which a neural circuit is set to a permissive state to facilitate the occurrence of the synaptic modifications necessary for memory formation (as in the case of trace eyelid conditioning in rabbits or olfactory operant conditioning in rats).

Another issue that requires more investigation is the extent to which the changes in intrinsic excitability are complementary to experience-dependent modifications in synaptic efficacy. The majority of theoretical models of memory storage have been largely based on persistent experience-driven changes in synaptic function. It is because of certain unique properties of the synapses, such as associativity and input-specificity, that systems of memory storage based on changes in synaptic strength have a potentially massive storage capacity. In contrast, global changes in intrinsic excitability would theoretically alter the throughput of all the synaptic inputs impinging to a given neuron. If this is the case, memory mechanisms based on neuron-wide altered intrinsic excitability would have a lower storage capacity and would be less versatile compared to systems of memory storage based on changes in synaptic strength.

Some recent results may help to reconcile differences between models of memory storage based on synaptic and models based on intrinsic plasticity. Patch-clamp techniques and simultaneous Ca^{2+} imaging that allow for recordings from individual dendrites have revealed that the induction of LTP was accompanied by a local increase in dendritic excitability that favored backpropagation of action potentials into that dendritic region, with a subsequent boost in the Ca^{2+} influx. A shift in the inactivation curve of a transient A-type K^+ current was found to account for the enhanced excitability. Importantly, the activity-dependent increase in dendritic excitability was localized at, or in the vicinity of, the synaptic site, which was potentiated in an input-specific manner following electrical stimulation. Local increase in dendritic excitability could facilitate enhanced propagation of synaptic potentials toward the site of action potential initiation. Therefore, enhancement of local excitability could contribute to the higher efficiency of coupling between synaptic potentials and spike as observed following LTP induction.

Although validation in *in vivo* preparations is required, these findings indicate that changes in intrinsic excitability can be induced to restricted membrane compartments and can be co-expressed with changes in synaptic efficacy in a manner that affects signal integration locally and can preserve input specificity.

Summary

Data collected over the past decades in both vertebrate and invertebrate model systems indicate that learning and memory as well as patterns of electrical stimulation of neurons and neural pathways not only alter synaptic function but also produce changes in intrinsic excitability. These changes in intrinsic excitability can be neuron wide or restricted to specific membrane compartments, such as the dendrites, thus affecting neuronal function and signal integration either globally or locally. Although ascertaining the functional relevance of certain changes in intrinsic excitability in the context of a given form of learning

has not been always successful, and in several cases the role of intrinsic plasticity still remains elusive, lines of evidence suggest that experience-dependent changes in intrinsic excitability may function as part of the engram itself, or as adaptive mechanisms to shape the stimulus specificity of the learned response, or also as mechanisms through which a neural circuit is set to a permissive state to favor the occurrence of the synaptic modifications necessary for memory formation and retrieval.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Invertebrate Models to Study Learning and Memory: Lymnaea; Learning and Memory in Invertebrates: *Aplysia*; Learning and Memory in Invertebrates: *Hermissenda*; Long-Term Depression: Cerebellum; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity.

Further Reading

Baxter DA and Byrne JH (2006) Feeding behavior of *Aplysia*: A model system for comparing cellular mechanisms of classical and operant conditioning. *Learning and Memory* 13: 669–680.

Bliss TV and Collingridge GL (1993) A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature* 361: 31–39.

Byrne JH (1987) Cellular analysis of associative learning. *Physiological Reviews* 67: 329–439.

Daoudal G and Debanne D (2003) Long-term plasticity of intrinsic excitability: Learning rules and mechanisms. *Learning and Memory* 10: 456–465.

Debanne D, Daoudal G, Sourdet V, et al. (2003) Brain plasticity and ion channels. *Journal of Physiology – Paris* 97: 403–414.

Fusi S, Drew PJ, and Abbott LF (2005) Cascade models of synaptically stored memories. *Neuron* 45: 599–611.

Lechner HA and Byrne JH (1998) New perspectives on classical conditioning: A synthesis of Hebbian and non-Hebbian mechanisms. *Neuron* 20: 355–358.

Magee JC and Johnston D (2005) Plasticity of dendritic function. *Current Opinion in Neurobiology* 15: 334–342.

Martin SJ and Morris RG (2002) New life in an old idea: The synaptic plasticity and memory hypothesis revisited. *Hippocampus* 12: 609–636.

Saar D and Barkai E (2003) Long-term modifications in intrinsic neuronal properties and rule learning in rats. *European Journal of Neuroscience* 17: 2727–2734.

Wolpaw JR (1997) The complex structure of a simple memory. *Trends in Neurosciences* 20: 588–894.

Zhang W and Linden DJ (2003) The other side of the engram: Experience-driven changes in neuronal intrinsic excitability. *Nature Reviews Neuroscience* 4: 885–900.

Pleasant Touch

A B Vallbo, H Olausson, and J Wessberg, Goteborg University, Goteborg, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Pleasant Touch – A Fundamental Affectionate Variable

It is a universal experience that bodily contact between individuals is a powerful means of nonverbal communication. Skin-to-skin touch may have a very strong emotional impact, immediate as well as long term. The affective potential of bodily contact is perhaps most apparent in erotic and parent–child relations, although touching is a potent means to convey empathy and relieve anxiety in less intimate personal contacts as well. Even a passing touch offered by a friend or an empathic nurse may give great reassurance to a person in fear, distress, or agony. Robin Dunbar stated wisely that a touch may be worth a thousand words. Further, in positive excitement we like to raise the feeling of shared joy, for instance, by hugging the teammates after a goal. On the other hand, the erotic potential of bodily contacts implies a barrier which in many situations makes us abstain from touching another person.

It was shown long ago by Harlow and co-workers that pleasant touch is essential to provide affection and security in early childhood. Studies of baby monkeys left in distress because they were deprived of their native mothers led to the conclusion that contact comfort is a very potent factor for the baby to develop affectionate responses, reassurance, and love for the mother. In fact, contact comfort was found to be even more powerful than nutrition.

Although a number of sensory mechanisms, as well as psychological factors are probably involved in most situations when we convey positive emotional signals by bodily contact, it is obvious that cutaneous touch is a significant element. The focus of the present account is on highly sensitive cutaneous sense organs which respond to light touch, particularly the unmyelinated tactile (CT) afferents which might have a specific role in this context. With regard to semantics, the term 'tactile' in this account solely refers to the finding that the afferents respond intensely to light touch, whereas the term does not refer to central effects related to haptics.

Tactile Afferents in Human Skin

The canonical view of human physiology is that light touch is signaled by large and fast conducting myelinated Aβ-afferents alone. However, in the last decades it has been demonstrated that human skin is also innervated by unmyelinated afferents responding to light touch (CT-afferents). They conduct impulses at a speed of about $1\,\mathrm{m\,s^{-1}}$ whereas Aβ-afferents are about 50 times faster (see Figure 1). Six different kinds of Aβ-afferents with separate physiological properties have been identified in human skin in addition to the CT-afferents (Table 1).

The distinct difference in receptive field size and topography is illustrated in Figure 2 for the main types of tactile afferents encountered in the hairy skin of the forearm. Corresponding illustrations for glabrous skin units may be found in Kandel, Schwartz, and Jessel's textbook. Receptive field structure as assessed with an objective method provides convincing evidence for unit identification and separation in distinct groups, although additional observations are essential, for example, conduction velocity and selective sensitivity to hair movements.

There is an interesting regional difference with regard to Aβ and CT innervation in man. CT-afferents have been demonstrated in face, forearm, hand dorsum, and thigh, suggesting that they are present all over the hairy skin. However, they have not been found in the glabrous skin of the hand in spite of numerous microenurography explorations of this area. Hence, it seems reasonable to conclude that they are lacking in the glabrous skin. Although there is no method to properly assess the proportion of CT-afferents out of total number of tactile units, it is a recurring experience in microneurography experiments that CT-afferents are encountered at least as often as Aβ-afferents in nerves innervating the forearm skin. However, there are indications that CT-afferents may be less common in the lower leg.

It should be pointed out that unmyelinated tactile afferents were demonstrated long ago in the hairy skin of various animals but they have not been found in the pad skin which corresponds to the glabrous skin in man – where CT-afferents are lacking as well. In animals the unmyelinated tactile afferents are often denoted as low threshold mechanoreceptive C-afferents (C-LTM). For long it was assumed that this system of afferents had faded away in the process of evolution from rat, rabbit, and cat to man until microneurogaphy demonstrated CT-afferents in man in the beginning of the 1990s.

Functional Properties of CT-Afferents

The CT-afferents are equally sensitive to skin deformation as many of the Aβ-afferents. They respond to

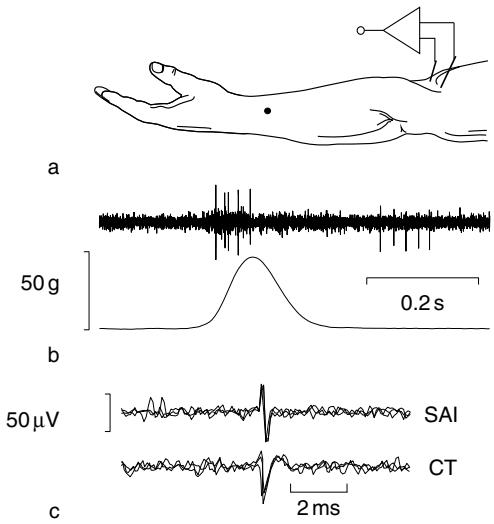

Figure 1 Microneurography recording of two single afferents to light touch, that is, a fast conducting Aβ-afferent and an unmyelinated CT-afferent. (a) Target point for touch stimulus and microneurography recording site from the lateral antebrachial cutaneous nerve in the upper arm. (b) Microneurography record and time course of tap stimulus showing an early burst of nerve impulses in the Aβ-afferent and a late burst in the CT-afferent. Distance between target area and recording site was 247 mm implying a conduction velocity below 1 m s^{-1} of the CT-afferent. The threshold of the CT-afferent as tested with von Frey monofilaments was 1.3 mN. (c) Nerve impulses from the record in (b) displayed on an expanded timescale to highlight the different impulse shape of the Aβ-afferent (SAI type) and the CT afferent. Reproduced from Wessberg J, Olausson H, Wiklund Fernstrom K, and Vallbo AB (2003) Receptive field properties of unmyelinated tactile afferents in the human skin. *Journal of Neurophysiology* 89: 1567–1575, with permission from The American Physiological Society.

Table 1 Mechanoreceptive afferents in human nerves from glabrous and hairy skin

Sensory endings and afferent unit acronyms		
Glabrous skin	*Hairy skin*	*Adaptation*
Merkel SAI	Merkel SAI	Slow
Ruffini SAII	Ruffini SA II	Slow
Pacini FA II (PC)	Pacini FA II (PC)	Fast
Meissner FAI (RA, QA)	Hair follicle unit	Fast
	Field	Fast
	C-tactile CT (CLTM)	Intermediate

The separate unit types were identified on the basis of microneurography recordings in man while types of end organ were inferred on the basis of physiological and morphological studies in various species. Slow and fast adaptation refers to the presence or lack of sustained response to stationary skin deformation. Most acronyms refer to adaptation properties (fast, rapid, quick, or slow adaptation). Acronyms within parenthesis have mainly been used in studies of nonhuman species. All afferents listed are myelinated and fast conducting (Aβ fibers) except CT units which are unmyelinated (C-fibers). The table includes highly sensitive afferents but not nociceptive afferents that may respond weakly to innouous skin deformation. The table entries refer to glabrous skin of the hand and hairy skin of the forearm and are based on recent investigations. However, there are microneurography studies suggesting a somewhat different innervation pattern in particular skin regions, the existence of subgroups, and one or two additional mechanoreceptor types.

indentation forces in the range 0.3–2.5 mN as tested with von Frey monofilaments and they produce high-frequency trains of action potentials (50–100 impulses s^{-1}) to pleasant touch stimuli. It should be recalled that maximal impulse frequency recorded from afferent C-fibers is about 100 impulses s^{-1}. In contrast, Aβ-afferents may fire close to 1000 impulses s^{-1}. It is of interest that, in addition to these two groups of sensitive tactile afferents, some of the nociceptive afferents may respond to innocuous touch as well. However, their sensitivity is low and their response to pleasant touch is minimal, usually not more than one or two impulses.

While CT-afferents share the high sensitivity with the Aβ-afferents, the physiological properties of the two groups differ in most other respects. Aβ-afferents fall in two distinct groups with regard to response to steady-state skin deformation (Table 1). The fast adapting units do not respond at all, that is, they lack static sensitivity altogether, while the slowly adapting units provide a continuous discharge for minutes or more. Actually, many Ruffini units may be firing continuously for life. In contrast, the CT-afferents have intermediate adaptation properties. They respond initially with a burst of high impulse rate, while the rate often falls to zero after a few seconds of sustained indentation. The discharge may then reassume after another few seconds, then wax and wane to finally stop altogether. CT-afferents lack the potential of the Aβ-afferents to code fast events. They respond well to slowly moving stimuli but poorly to brisk movements and vibration. Hence, their dynamic response range is limited to low frequencies. Stimuli which are particularly efficient to evoke a massive CT-input are slowly moving, caress-like stimuli over large skin areas with the hand or a broad artist paintbrush. A prominent feature of the CT-afferents is that they are highly fatigable. When identical stimuli are repeated, the first stimulus usually evokes a much larger response than the following. This feature suggests that the system is particularly interested in novel events. When a local skin deformation is released, CT-afferents may produce an after-discharge that may last for many seconds. However, the duration varies a lot even within the individual unit depending on temperature, previous history, and other nondefined factors. An additional and unique feature of the CT-afferents

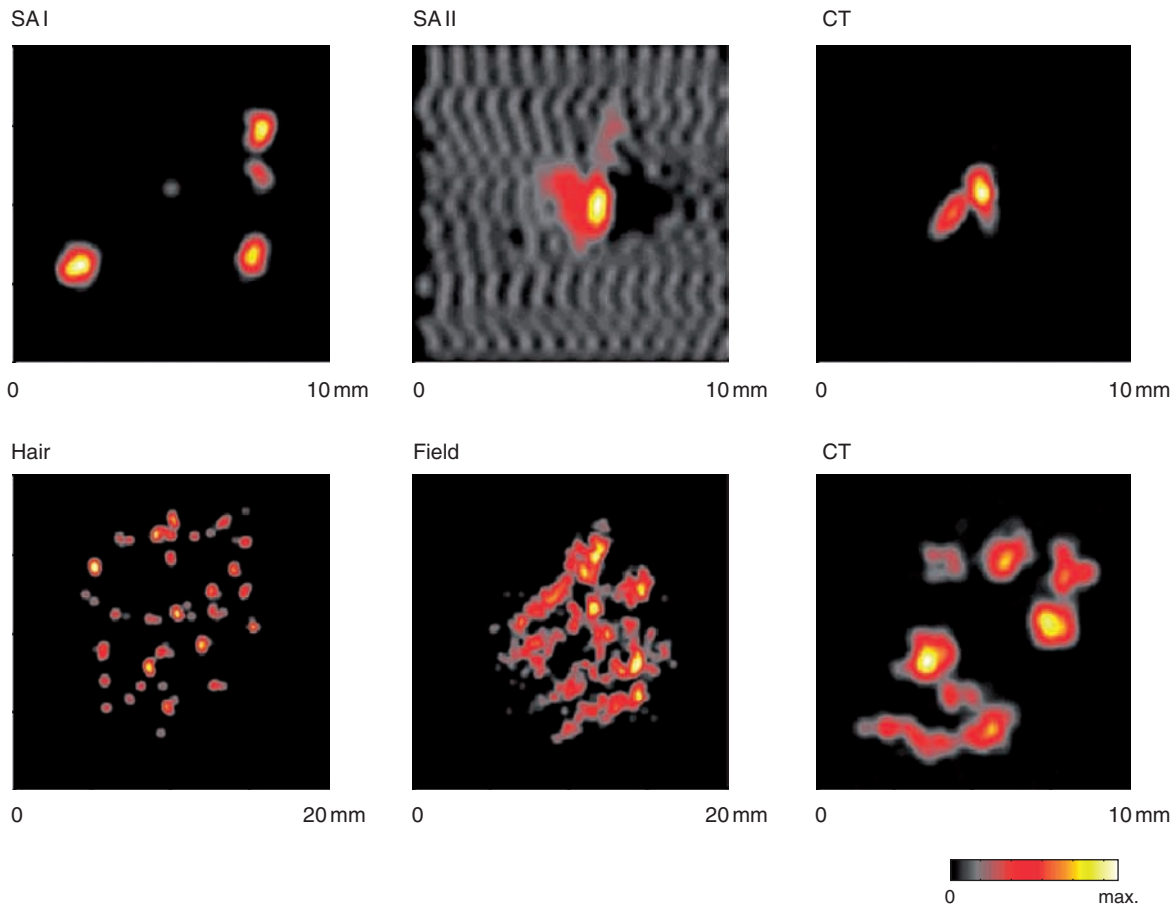

Figure 2 Receptive field topography of five kinds of tactile afferents in the hairy skin. Data from four representative Aβ-afferents and two different CT afferents (see **Table 1**). Note difference in scale for the hair and field units. The colors represent intensity of afferent firing from zero to maximal rate of the particular trial. Peak rates fall in the range 35–60 impulses s^{-1}. Gray areas which fill up the SAII box represent continuous spontaneous activity of the afferent. (Aβ afferents) Adapted from Vallbo AB, Olausson H, Wessberg J, and Kakuda N (1995) Receptive field characteristics of tactile units with myelinated afferents in hairy skin of human subjects. *Journal of Physiology* 483: 783–795, with permission from Blackwell Publishing. (CT-afferents) Reproduced from Wessberg J, Olausson H, Wiklund Fernstrom K, and Vallbo AB (2003) Receptive field properties of unmyelinated tactile afferents in the human skin. *Journal of Neurophysiology* 89: 1567–1575, with permission from The American Physiological Society.

is that the response of the individual afferent is not as consistent and predictable as with individual Aβ-afferents but may vary from one test series to the next for no obvious reason.

The response features of the CT-afferents as outlined above all converge toward the interpretation that the system is not designed to code accurately and consistently intensity and time course of skin deformations. These are functions which are essential for discriminative touch. In daily life, the tactile system continually captures details of temporally and spatially variant skin deformations. As a result we can readily perceive more complex features, for instance, the direction of a stimulus moving along the skin and we can discriminate the intensity of component indentations which are closely adjacent in time and space to assess shape and surface structure of objects. Altogether, it seems evident that the CT-afferents lack response properties required for serving discriminative touch.

Significance of conduction velocity The high conduction velocity of the myelinated Aβ-afferents is a feature which is essential for discriminative touch, as well as for guidance of motor activities. Hence, the low conduction velocity of the CT-afferents precludes these roles. First, the long delay (0.5–1.0 s) from stimulus to impulse arrival in the brain is hardly compatible with accurate integration with signals from fast Aβ-afferents and other sensory systems, for example, vision and hearing, that is, forms of integration which are often essential for guidance of

motor activities. Second, the documented interfiber variation in conduction velocity around $1\,m\,s^{-1}$ implies that the set of impulses in the sample of CT-afferents excited by a local touch would be widely dispersed in time when arriving to the brain. As a result, the CT-afferents produce a temporally blurred account of the peripheral event compared to that of the Aβ-system. Hence, the slow conduction velocity of the unmyelinated nerve fibers implies an additional argument against the CT-system having a role in discriminative touch.

CT-Afferents – Functional Hypotheses

The physiological properties of the CT-afferents as presented above seem to rule out the possibility that the system plays an essential role in discriminative and motor aspects of cutaneous mechanoreception. Hence, it is of interest to consider what other functions might justify the presence of a slow system with selective sensitivity to light touch. Which functions might afford the deficiencies of the CT-system in coding peripheral events, yet be biologically significant enough to justify the preservation of a parallel tactile system in addition to the fast and accurate Aβ-system?

Tickle Hypothesis

For long, the low threshold mechanoreceptive C-afferents were discussed as a system accounting for the sensation of tickle. This idea proposed by Zotterman in 1939 was largely based on his observation that unmyelinated tactile afferents in the cat may produce long-lasting after-discharges when a skin indentation was released. The tickle hypothesis was then reiterated in a number of successive publications as late as in the 1990s. However, in microneurogaphy studies of human CT-afferents, it has been found that the duration of after-discharge does not match the duration of tickle sensations, nor is the impulse rate of CT-afferents related to the intensity of tickle sensation. Moreover, subjects who lack Aβ-afferents but have preserved small fiber functions (see below) are not aware of being tickled in the affected skin areas. These findings indicate that the sensation of tickle is dependent on Aβ-afferents rather than on CT-afferents.

Maintenance of Cortical Maps

Studies in nonhuman subjects (cat and flying fox) have suggested that unmyelinated mechanoreceptive afferents may play an essential role in the complex processes of shaping and maintaining the map of body surface in the somatosensory cortex areas. The hypothesis is based on observations of how these maps are modified with amputations and other injuries. In relation to a possible role of CT-afferents in this context, it seems significant that the unmyelinated afferents involved in body surface mapping are sensitive to capsaicin. Since human CT-afferents are not sensitive to capsaicin, it seems reasonable to conclude that the CT-afferents are not main candidates for cortical map design.

Affective Touch Hypothesis

Although poor coding accuracy is a serious drawback in discriminative functions, an area where these deficiencies are readily affordable is the limbic realm. Considering the potential roles of light touch in relation to limbic functions, it seems reasonable to single out one function of particular interest, that is, tight physical contact with conspecifics, in the first place with your mother, later in life with your partner and children. We all strongly feel now and then that skin wants skin. These considerations led to the affective touch hypothesis, implying that the essential role of the CT-system is to convey pleasant aspects of light touch, particularly in skin-to-skin contact with affiliative human beings. It should be emphasized that this does not imply that CT-afferents are the only afferents that may contribute or give rise to limbic touch effects, as will be discussed in a later section.

Responses to Activation of the CT-System

To test the affective touch hypothesis and explore the functional role of the CT-system, it is desirable to activate the CT-afferents selectively without activating Aβ-afferents. Unfortunately, this is not possible in normal subjects although stimuli designed to give different proportions of Aβ- or CT-input have been tested.

It has been demonstrated in proper psychophysical tests that a slowly moving touch stimulus ($1–5\,cm\,s^{-1}$) which is particularly effective to activate CT-afferents is more pleasant than a fast-moving stimulus. Vibratory stimuli which activate Aβ-afferents very efficiently, but CT-afferents only weakly, are rated as less pleasant than slowly moving touch stimuli. These observations are consistent with the hypothesis that CT-afferents are more important for the feeling of pleasant touch than are Aβ-afferents. Although findings of this nature are consistent with the affective touch hypothesis, more pertinent data have been obtained from a few subjects, who suffer from a very rare neurological deficiency (sensory neuronopathy syndrome), implying that they lack Aβ-afferents from large parts of the body. As a consequence, it is reasonable to assume that their sole tactile afferents are unmyelinated CT-afferents.

Psychophysical Responses

Two subjects with sensory neuronopathy syndrome are well described in the literature. They have been studied extensively over the years particularly with regard to motor functions. It has also been reported, although more in passing, that they lack tactile sensibility. This observation was consistent with the prevailing view at that time that tactile sensibility is lacking altogether when Aβ-afferents are nonfunctioning. This view, in turn, was largely based on experiments in normal subjects which suggested that tactile sensibility is lacking when the Aβ-afferents had been selectively blocked by sustained pressure onto the nerve.

However, when it became evident that human skin is supplied with a system of unmyelinated afferents sensitive to light touch, it became pertinent to reexamine the tactile sensibility of neuronopathy subjects using more refined approaches. Rigorous psychophysical tests were pursued to study if the neuronopathy subjects were at all able to detect touch stimuli which were particularly designed to evoke massive CT-inputs. It was found, in 2-alternative-force-choice (2-AFC) tests, that subjects lacking Aβ-afferents readily detected such stimuli applied to the forearm skin where CT-afferents are abundant. In contrast, they failed altogether to detect the same kind of stimuli applied to the glabrous skin of the hand where CT-afferents are lacking. A difference of similar kind was found between the two skin regions in the neuronopathy subjects when detection threshold was assessed using von Frey monofilament. Relatively weak monofilaments which give a pure sensation of touch in normal subjects were readily detected in the hairy skin but not in the glabrous skin. Here much stronger stimuli (4–8 times stronger) were required for detection. Moreover, the neuronopathy patients reported that these strong stimuli gave rise to a sharp and marginally painful sensation, suggesting that detection in the palm was based on nociceptive input. In contrast, sensations with threshold von Frey monofilament in the hairy skin lacked all elements of pain or sharpness suggesting that detection was based on afferents sensitive to innocuous stimuli. It should be emphasized that, in normal subjects, there is no difference between the two skin regions in this respect, that is, thresholds for detection of monofilament stimuli are identical in the hairy and glabrous skin. Further, thresholds are generally lower than in the neuronopathy subjects, suggesting detection by Aβ-afferents in normal subjects. An additional highly relevant finding is that the neuronopathy subjects are unable to detect vibratory stimuli which are known to a give a poor excitation of CT-afferents but a massive activation of several kinds of Aβ-afferents.

Although these observations agree that CT-afferents may well account for detection of touch stimuli in subjects lacking Aβ-afferents, it has been more difficult to obtain a clear and consistent apprehension of the quality of the sensation perceived when CT-afferents are selectively stimulated. Still, a number of interesting features emerged, when the qualitative characteristics of the sensation were explored.

First, it was obvious that the sensation associated with a massive and selective CT-input was very weak, vague, and inconsistent when slowly stroking a broad and soft artist paintbrush along the forearm of the neuronopathy subjects. In some trials, the subject reported no sensations at all. In others, they reported a sensation of light touch which was barely detectable and difficult to mentally catch and describe in detail. In this context it should be noticed that stimulus detection in a psychophysical 2-AFC test does not prove that the stimulus is very clearly perceived. The weakness and vagueness of the sensation are further illustrated by the fact that conscious perception of CT-stimulus varied from one occasion to the other and between the two subjects as well. In fact, one of the two subjects (GL) reported that she began to feel more touch sensations in daily life, once she had had the experience of touch perception from the affected skin areas and had become aware of the possibility. Although the two neuronopathy subjects were not able to give a clear and detailed description of the sensation evoked by CT-stimulation, some consistent features of the quality emerged, namely that the sensation had no tint of pain, tickle, or itch. Further they reported that the sensation elicited by CT-stimulation was slightly or moderately pleasant.

Brain Responses

When functional magnetic resonance imaging (fMRI) was used to study brain response to touch stimuli (Figure 3) in normal and neuronopathy subjects lacking Aβ-afferents, the findings indicated that different sensory areas are activated by the fast (Aβ) and the slow (CT) tactile systems from the hairy skin. In normal subjects pleasant touch activated the classical somatosensory areas SI and SII as well as insular cortex, notably the posterior part of the contra-lateral insula. Somatosensory areas SI and SII are known to be significant for discriminative touch whereas insular cortex is a region of great interest in relation to affective mechanisms, because it is considered to house a gateway from sensory systems to the emotional systems of the frontal lobe. When similar stimuli were applied to subjects lacking Aβ-afferents, no activation was found in the somatosensory areas whereas the posterior insular region was activated.

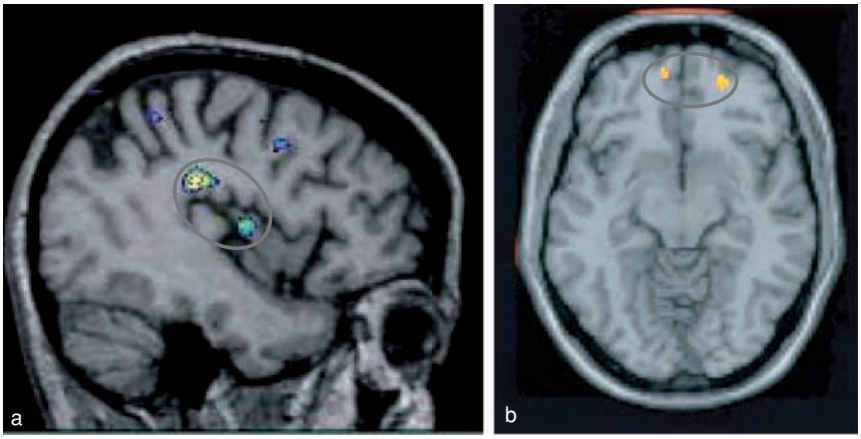

Figure 3 fMRI responses (BOLD) to pleasant touch stimuli. (a) Response in posterior and middle insular region to selective stimulation of CT-afferents in a neuronopathy subject lacking large myelinated Aβ-afferents. In this subject light touch produced no BOLD response in somatosensory areas SI and SII, whereas similar stimuli regularly activate these two areas in normal subjects. (b) BOLD response in orbitofrontal cortex to pleasant touch stimuli of the palm in normal subjects (group data). BOLD, blood-oxygen level dependent. (a) Unpublished. (b) Adapted from Rolls ET, O'Doherty J, Kringelbach ML, et al. (2003) Representations of pleasant and painful touch in the human orbitofrontal and cingulated cortices. *Cerebral Cortex* 13: 308–317, with permission from Oxford University Press.

The conclusion is that the unmyelinated CT-afferents project toward the emotional systems (insular cortex) but less or not at all toward the discriminative–cognitive systems (classical somatosensory areas SI and SII). This observation seems to provide an essential support for the affective touch hypothesis.

To what extent the tactile Aβ-afferents project toward the emotional systems along with the CT-afferents has not been widely explored. However, it seems likely that this may be the case, because pleasant touch stimuli of the palm, where CT-afferents are lacking, give rise to fMRI responses in a target area for insular efferents in orbitofrontal cortex involved in more complex emotional mechanisms. It remains to be explored whether a projection of tactile Aβ-afferents to emotional centers is unique for the palm or is true for other skin areas as well. It may be speculated that Aβ-afferents from the palm have obtained an affective role because CT-afferents are lacking here. Anyway, it seems clear that affective aspects of touch are not always dependent on CT-afferents alone.

The physiological properties of the CT-afferents as well as the psychophysical and fMRI responses to selective activation of the unmyelinated tactile system converge toward a limbic-emotional role of the system. The affective touch hypothesis implies that the essential role of the CT-system is to provide or support emotional, hormonal, and behavioral responses to skin-to-skin contact with conspecifics. On the other hand, it seems likely that a natural perceptive-emotional response to pleasant touch is dependent on the combination of afferents from the two tactile systems, because selective CT-stimulation fails to evoke anything like a full sensation of pleasant touch. The combination of CT- and Aβ-afferents is obviously required for the complete feeling of pleasant touch in the hairy skin. It is also pertinent to recall a general principle regarding emotional responses. The intensity and even the quality of the emotional response evoked by a particular stimulus is highly dependent on contextual factors, as well as state of deprivation, that is, to what extent your desire with regard to the particular emotion has recently been satisfied or not. Obviously, your sense of pleasure is intense when you hug and touch your partner after weeks of separation, whereas you are not particularly pleased when touched by an unknown person in a crowd.

It may be asked why evolution has retained a separate system of unmyelinated afferents for social-emotional touch. Why could this role not be handled by the Aβ-afferents? It is difficult to deny that they could – and, in fact, they do in the glabrous skin of the hand.A particular reason for preservation of the CT-system is that unmyelinated afferents are parsimonious because they take up little space in the peripheral nerves.

CT-System – A Branch of an Interoceptive Afferent System

In a wider perspective, the CT-system may be regarded as a branch of a large afferent system which is basically concerned with your own body rather than external events, as conjectured by Craig. The basic role of the interoceptive system is to continuously watch the condition of body tissues, as well as physiological and chemical variables within the body. Interoceptive afferents have particular access

to brain centers which control affective, hormonal, autonomic, and behavioral responses with effects to readjust adverse conditions and keep variables within desired ranges. The ultimate aim would be to keep the cell, the individual, and the species going. Included in the interoceptive system are afferents related to perception of pain, itch, temperature, air hunger, vasomotor flush, hunger, thirst, and a range of visceral sensations, as well as afferents which are essential for the subconscious control of physiological variables, such as blood pressure and concentration of blood gases.

The peripheral afferents of the interoceptive system consist of small diameter nerve fibers (thin myelinated Aδ and C fibers) with projections to the superficial lamina of the dorsal horn in the spinal cord and nucleus of the solitary tract. In this context it is interesting that touch sensitive unmyelinated afferents (C-LTM) have been shown to project to lamina I and II in several mammalian species, consistent with the interpretation that they belong to the interoceptive system. However, apart from this finding, connectivity studies of CT-afferents or homologs in other species are unfortunately lacking altogether in the literature.

At higher levels the small fiber afferents are tightly linked to amygdale and hypothalamus and, at cortical levels, to the insular region, primarily the posterior and middle parts. Moreover, there are indications that the anterior insular cortex on the nondominant side constitutes a still higher level housing a re-representation of the interoceptive image of the entire body. This representation of the physical self is postulated to constitute a basic foundation of mental consciousness.

The role of the CT-units as an afferent branch of a system guarding the well-being of the body would be to signal pleasure and reassurance as you are close to your parents, lover, kin, or friends. There are indications that pleasant bodily contacts promote hormonal responses, that is, endorphin and oxytocin, which contribute to the feeling of well-being, confidence, and calmness.

Bodily Contact

Although the affective touch hypothesis implies that the particular role of the CT-system is to promote emotional, hormonal, and behavioral responses to pleasant physical contact with conspecifics, it is obvious that a number of other sensory channels are important as well. Interesting observations of social and biological roles of tight bodily contacts in nonhuman species have been reported.

A particular kind of bodily contact is grooming. At first sight it appears that grooming has a hygienic role alone. However, Dunbar has demonstrated in monkeys that it has an affective and social role as well. Obviously, the groomee is very pleased with the procedure even though the handling is often quite tough. Grooming increases the production of endorphin in the groomee. Dunbar has also found that monkeys spend much more time in grooming than required from hygienic point of view. It is particularly interesting that the time spent on grooming is larger, the larger the social group. The bottom line conclusion is that an essential role of grooming is to promote affectionate attachment between individuals and hence to keep the group together.

The caring pattern of the mother in handling her young infants has a profound influence on the development of the child. Epidemiological studies in man have demonstrated that adverse conditions in early childhood tend to increase stressful and nonsocial behavior in the adult. Biological mechanisms that might be significant for this phenomenon have been explored in rats by Meaney and co-workers. Individual rat mothers differ with regard to caring pattern. Some are very active, spending a lot of time licking and handling their pups. Others are more passive in that respect. The caring pattern has a profound effect on epigenetic mechanisms related to the expression of stress hormone receptors in the brain. In the end, the offspring of active mothers become less stress sensitive, more curious, and exploratory than the offspring of passive mothers. Interestingly, these differences are permanent and remain in the adult, although they are partially modifiable.

Although it remains to be explored to what extent the CT-system is involved in affective response to bodily contacts in a social context, like grooming and mother's care of her baby, there is no question that the CT-system is activated in these situations. The affective touch hypothesis implies that the CT-system is one of several communication channels accounting for responses to bodily contacts which engage perceptive, hormonal, autonomic, and behavioral mechanisms and might range from immediate pleasure, joy, and reassurance in sex to the postnatal development of your children's stress coping profiles.

See also: Cross-Modal Interactions Between Vision and Touch; Mechanoreceptors; Orbitofrontal Cortex: Visual Functions; Tactile Coding in Peripheral Neural Populations; Tactile Texture.

Further Reading

Calford MB and Tweedale R (1991) C-fibres provide a source of masking inhibition to primary somatosensory cortex.

Proceedings of the Royal Society, Series B: Biological Sciences 243: 269–275.

Cole J (1995) *Pride and a Daily Marathon*. Cambridge, MA: The MIT Press.

Cole J, Bushnell MC, McGlone F, et al. (2006) Unmyelinated tactile afferents underpin detection of low-force monofilaments. *Muscle Nerve* 34: 105–107.

Craig AD (2002) How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews* 3: 655–666.

Dunbar R (1997) *Grooming, Gossip and the Evolution of Language*. London: Faber & Faber.

Harlow HF (1959) Love in infant monkeys. *Scientific American* 200: 68–74.

Kandel ER, Schwartz JH, and Jessell TM (eds.) (2000) *Principles of Neural Science*, 4th edn. New York: McGraw-Hill.

Olausson H, Lamarre Y, Backlund, et al. (2002) Unmyelinated tactile afferents signal touch and project to insular cortex. *Nature Neuroscience* 5: 900–904.

Rolls ET, O'Doherty J, Kringelbach ML, et al. (2003) Representations of pleasant and painful touch in the human orbitofrontal and cingulated cortices. *Cerebral Cortex* 13: 308–317.

Sterman AB, Schaumburg HH, and Ashbury AK (1980) The acute sensory neuronopathy syndrome: A distinct clinical entity. *Annual Neurology* 7: 354–358.

Uvnas-Moberg K (1997) Physiological and endocrine effects of social contact. In: Carter CS, Lederhendler II, and Kirkpatrick B (eds.) *The Integrative Neurobiology of Affiliation (Annals of the New York Academy of Science)*, vol. 807, pp. 146–163. New York: The New York Academy of Sciences.

Vallbo AB, Olausson H, and Wessberg J (1999) Unmyelinated afferents constitute a second system coding tactile stimuli of the human hairy skin. *Journal of Neurophysiology* 81: 2753–2763.

Vallbo AB, Olausson H, Wessberg J, and Kakuda N (1995) Receptive field characteristics of tactile units with myelinated afferents in hairy skin of human subjects. *Journal of Physiology* 483: 783–795.

Weaver IC, Cervoni N, Champagne FA, et al. (2004) Epigenetic programming by maternal behaviour. *Nature Neuroscience* 7: 847–854.

Wessberg J, Olausson H, Wiklund Fernstrom K, and Vallbo AB (2003) Receptive field properties of unmyelinated tactile afferents in the human skin. *Journal of Neurophysiology* 89: 1567–1575.

Relevant Website

http://deafferented.apinc.org – Studies on the Deafferent Patient G.L.

Population Codes: Theoretic Aspects

W J Ma and A Pouget, University of Rochester, Rochester, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons encode in their spike trains features of the external world, internal states of the organism, or motor commands. Understanding the nature of the neural code is a key step toward understanding how the brain computes. In this article, we focus on one type of neural code, known as population codes, which are used to encode many features throughout the nervous system.

A population code is a way of representing information about a feature through the simultaneous activity of a large set of neurons sensitive to the feature. This set is called a population. A population code is useful for increasing the organism's certainty about a feature as well as for encoding multiple features simultaneously.

In most studies of population codes, the activity of a neuron is quantified either by the number of spikes it emits or by the average rate of spike emission (firing) over a given time interval. The response of all neurons in a population to a stimulus is called the population response or the population pattern of activity. It can encode information about one or more stimulus features. In the case of a visual stimulus such as a moving image, a feature could, for instance, be contrast, orientation, color, shape, speed, identity, or utility. A stimulus feature of interest is itself often also referred to as a stimulus; we follow this terminology.

An important aspect of population codes is the stochastic nature of neuronal responses. This is unlike what is found in a computer, for example, in which words are mapped deterministically onto strings of zeros and ones. There is only one word that corresponds to a given string and vice versa. In contrast, in the brain, the mapping from stimulus to population response is probabilistic. When the same stimulus is presented multiple times, the population response will in general be different each time. As a consequence, different stimuli can elicit the exact same population response, although generally with different probabilities.

The probabilistic mapping from stimulus to population response is called the encoding of the stimulus. The reverse process is decoding – obtaining a stimulus estimate from a given population response.

In this article, we first discuss encoding and give examples of population codes. Modeling the process by which neurons encode information is a prerequisite for the development of decoding techniques because decoding can often be thought of as a mathematical inverse to the encoding process. Next, we describe various decoders along with their performance and their connection to the encoding models. Finally, we address recent developments and open issues. Throughout the article, we consider an example in which the stimulus feature of interest is orientation, which takes values between 0° and 180°, but none of our conclusions is specific to this particular choice of feature.

Encoding

Encoding is the mapping from stimulus to population response. Since the response is variable, this mapping is probabilistic. Describing encoding therefore requires specifying a probability distribution over responses. This distribution will change on varying the stimulus and is therefore a conditional distribution, where the conditioning is on the stimulus. Such a distribution is also called a generative model or a noise model.

Tuning Curve Plus Noise

The response of each neuron in a population to a stimulus is typically characterized as the sum of a mean response and a noise term, both of which may depend on the stimulus. The mean response of a single neuron upon varying the stimulus value is characterized by a tuning curve. Let us consider a population consisting of N neurons, labeled by an index $i = 1, \ldots, N$. We denote the mean response of neuron i to stimulus s by $f_i(s)$. On a single trial, when a stimulus s is presented, the activity of neuron i can be written as:

$$r_i = f_i(s) + \eta_i(s)$$

where $\eta_i(s)$ is zero-mean, possibly stimulus-dependent noise. Tuning curves are usually either bell-shaped or monotonic. If the tuning curve is bell-shaped, the neuron has a preferred stimulus value (i.e., one to which its mean response is highest). Figure 1(a) shows the tuning curve and the noise of a neuron responding to orientation. In this example, the tuning curve is described by a circular Gaussian (von Mises distribution),

$$f_i(s) = Ae^{B\cos(2(s-s_i))} + C$$

where A, B, and C are constants representing amplitude, narrowness, and baseline, respectively, and s_i is the preferred orientation of neuron i. Neurons often have broad tuning curves (in our example, B would be low). This might be surprising if the goal is to encode a stimulus with high precision. However, we will see that this is the strength of population coding: even

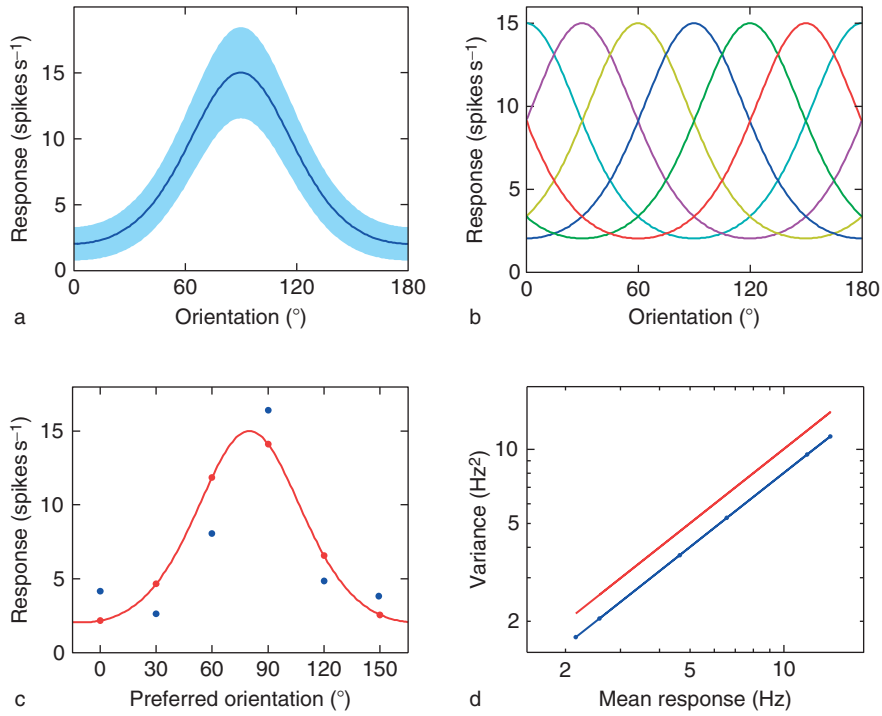

Figure 1 Example of encoding a stimulus in a population code for orientation. (a) Response of a single neuron as a function of stimulus orientation. The line represents the mean activity (tuning curve), which is described by a circular Gaussian. The maximum response was set to 15 spikes/s. This neuron has a preferred orientation of 90°. The shaded area contains responses within one standard deviation from the mean. Variability is taken to be Gaussian with variance proportional to the mean, with negative values set to zero. In this example, the constant of proportionality between mean and variance (Fano factor) is 0.8. (b) Tuning curves for all six neurons in the population. Population code typically involves thousands of neurons, but only six are shown here for clarity. The neurons' preferred orientations were chosen to be equally spaced, with a spacing of 30°. The tuning curves are broad and overlapping. (c) Population response to a contour oriented at 80°. The six neurons are labeled by their preferred orientation. The red circles indicate the mean response of the neurons to this stimulus. The red line shows the mean responses if there were many more neurons. The blue circles show an example of a response on a single trial. (d) Variance as a function of the mean response on a log–log plot. A straight line with slope 1 indicates that variance is proportional to the mean. The red line corresponds to a Poisson process, in which the variance is equal to the mean. The blue circles are the neurons in our example. Since we chose the constant of proportionality to be 0.8, these circles lie below the line of a Poisson process.

when tuning curves are broad and variability is large, stimuli can be encoded with high accuracy because every neuron contributes a piece of information. Figure 1(b) shows tuning curves for all neurons in a population of six.

Single-Neuron Response Distribution

To go beyond the description of mean response plus noise, one can describe the full response distribution of a neuron. The noise distribution is completely unrelated to the shape of the tuning curve. Common assumptions for the response distribution of a neuron are a Poisson distribution (for spike counts) and a normal distribution (for the number of spikes per second). A Poisson distribution is a one-parameter discrete distribution. The probability of a neuron i showing activity r_i in response to stimulus s is given by

$$P(r_i|s) = \frac{e^{-f_i(s)} f_i(s)^{r_i}}{r_i!}$$

In a Poisson process, the probability at each time step of a neuron firing a spike is independent of the probability at the previous time step. For large means (greater than approximately 10), a Poisson distribution is well approximated by a normal distribution with variance equal to the mean.

Normally distributed neural variability has the form

$$P(r_i|s) = \frac{1}{\sigma_i(s)\sqrt{2\pi}} \exp\left[-\frac{(r_i - f_i(s))^2}{2\sigma_i(s)^2}\right]$$

where $\sigma_i(s)$ is the possibly stimulus-dependent standard deviation. The rationale behind a normal distribution is that neuronal variability is due to the combined effect of a large number of stochastic processes. Both distributions are not fully satisfactory in describing observed neuronal variability: normal distributions assign nonzero probability to negative responses, whereas Poisson distributions are limited to a ratio between variance and mean equal to 1.

This ratio, also called the Fano factor, is experimentally found to take values between 0.3 and 1.8.

The example in Figure 1 uses for each neuron Gaussian variability with variance proportional to the mean, with negative values set to zero. Neurons are assumed to be independent. Figure 1(c) shows a sample population response and the mean population response to a stimulus oriented at 90°. Figure 1(d) shows the relation between variance and mean for the variability chosen here, compared to a Poisson process.

Population Response Distribution

To describe the probabilistic response of an entire population, we need to make assumptions about the joint responses of neurons. The simplest assumption is that all neurons respond independently from each other. Then, the population response distribution is the product of the response distributions of the neurons in the population. That is, if we denote the population response by a vector $\boldsymbol{r} = \{r_i\}_{i=1,\ldots,N}$, its distribution is

$$P(\boldsymbol{r}|s) = \prod_{i=1}^{N} P(r_i|s)$$

This is also called the likelihood function of s. However, responses of neurons in a population are usually not independent of each other but, rather, exhibit noise correlations. A positive pairwise correlation implies that when the activity of one of the neurons in the pair on a given trial is above its mean, the activity of the other neuron is likely to also be above its mean. Population responses with pairwise correlations are typically modeled by a multivariate normal distribution with a nondiagonal covariance matrix. The off-diagonal elements of this matrix measure the pairwise correlations. Experimentally, it is very difficult to measure a covariance matrix for a large number of neurons because of the large amount of data needed. Some data are available for pairwise correlations and they clearly indicate that many neurons have positive correlations. However, the form of the joint probability distribution over a large population of neurons, and particularly whether this distribution can be approximated by a multivariate Gaussian distribution, is unknown.

Higher-order correlations can be described mathematically but are experimentally even more difficult to measure. Consequently, few studies have attempted to measure or investigate higher-order correlations.

Examples of Population Codes

Population codes are believed to be widespread in nervous systems. In electrophysiological experiments, a population response is normally recorded from multiple nearby electrodes in a particular brain area. Population codes have been studied most widely in sensory systems. For instance, in primary visual cortex (V1) and area V4 of the macaque, population codes exist for orientation, color, and spatial frequency. In the hippocampus in rats, a population code exists for the animal's body location. A population code for spatial location in a visual scene or of the body of the organism is also called a topographic map. The cercal system of the cricket has a population code for wind direction. The secondary somatosensory area (S2) in the macaque has population codes for surface roughness, speed, and force. The postsubiculum in rat contains a population code for head direction. Primary motor cortex (M1) in macaque uses populations coding for direction of reach. Even abstract concepts such as number appear to be encoded by population codes in the prefrontal cortex.

Another Encoding Scheme: Encoding Probability Distributions Explicitly

When the encoding schemes described in the previous section were first developed, the central assumption was that a population code encodes a single value at any given time (e.g., the orientation of a contour). Subsequently, several authors noted that population codes might encode probability distributions instead of single values. Encoding probability distributions is a particularly intriguing idea because it would allow the nervous system to perform statistical inference, a very powerful approach when performing computations in the presence of uncertainty. As discussed later, it is in fact possible to recover a probability distribution from a population code from the 'tuning curve plus noise' encoding model described previously.

There also exist other encoding models that naturally lead to decoders that return probability distributions. These are encoding schemes in which the mean response of each neuron is directly a function of the probability that the stimulus is the preferred stimulus of the neuron. This function could, for instance, be a scaled version of the probability:

$$f_i(s) = A \cdot \Pr(s = s_i)$$

where A is a constant. It could also be a linear function of its logarithm,

$$f_i(s) = [A \log \Pr(s = s_i) + B]_+$$

where A and B are constants, and $[x]_+ = x$ if $x \geq 0$ and 0 otherwise. Such codes are sometimes called explicit codes for probability distributions. In explicit codes, neural variability is merely a nuisance and its form has no particular meaning.

Decoding

In order to assess the amount of information contained in a neural population, experimentalists recording

neural activity need to be able to obtain a stimulus estimate from the pattern of activity in that population on a single trial. This process is called decoding and can be regarded as the inverse of the encoding process.

In this section, we review several popular decoding techniques. As discussed later, these methods differ in the reliability of the stimulus estimates they produce. An optimal decoder is one that recovers all the information available in the population activity, where information is defined as Fisher information. Thus, using an optimal decoder is a way to determine the information available in a given population code.

The issue of decoding a neuronal population is not only relevant for experimentalists but also important for any computation performed by downstream neurons. Generation of a motor command is an example of obtaining a single estimate based on sensory information encoded in a population. The accuracy of such a computation is limited both by the accuracy with which sensory input is encoded (i.e., the amount of Fisher information available) and by the performance of the brain's decoding method (i.e., the amount of Fisher information extracted). Therefore, the issue of decoding population codes is central not only to analysis of neural data but also to our understanding of how neural computations are being performed.

Fisher Information

The next section reviews various decoding methods for population codes. These methods differ by two criteria: how much data they require and their performance. In general, the more data required, the better the results. However, how can one quantify the performance of a decoder? The performance is determined in part by the noise in the encoding process. Because of this noise, the neural activity is guaranteed to vary from trial to trial even when the stimulus is maintained constant. As a result, any decoder of this neural activity will return estimates that also vary from trial to trial. To evaluate the performance of a particular estimator, it is common to examine two quantities related to the probability distribution over the estimate: the bias and the variance. The bias, which is the difference between the average estimate and the actual stimulus value, indicates any systematic misestimation. The variance of the stimulus estimates quantifies the variability in the estimate from trial to trial and is inversely related to the reliability of the estimate. Ideally, a decoder should have zero bias and minimal variance. A general result, known as the Cramer–Rao bound, specifies the smallest variance that can be achieved by any estimator. It can be calculated from the bias of the estimator and an estimator-independent quantity called the Fisher information for the population code. Fisher information is defined in terms of the conditional probability distribution describing the population response:

$$J(s) = \left\langle \left(\frac{\partial}{\partial s} \log P(\boldsymbol{r}|s) \right)^2 \right\rangle$$

where $\langle \cdot \rangle$ is the expected value taken over $P(\boldsymbol{r}|s)$. $P(\boldsymbol{r}|s)$ corresponds to the encoding process, as discussed in the previous section. Note that Fisher information is not a function of a particular response since $\mathbf{r}$ is averaged out. Instead, it measures the quality of an encoding model. Fisher information provides a lower bound on the variance $\sigma_{\hat{s}}^2$ of any estimator. Specifically, the variance of an unbiased estimator $\hat{s}$ of s is guaranteed to satisfy

$$\sigma_{\hat{s}}^2 \geq \frac{1}{J(s)}$$

A generalization of this formula exists for biased estimators. The variance of an estimator also determines the smallest change in the stimulus that can be reliably discriminated. If the variance is small, the estimator can be used to detect tiny changes in s. Accordingly, there is a natural link between Fisher information and discrimination threshold: Fisher information is inversely proportional to the square of the discrimination threshold of an ideal observer of the neural activity, or equivalently, it is proportional to the square of the sensitivity d' of an ideal observer:

$$J(s) = \frac{d'^2}{\delta s^2}$$

where d' is the sensitivity (a measure of performance), and δs is the distance between the two stimuli to be discriminated. For the case of independent Poisson neurons, Fisher is easily calculated to be

$$J(s) = \sum_{i=1}^{N} \frac{f_i'(s)^2}{f_i(s)}$$

where $f_i'(s)$ is the derivative of the tuning curve. This equation implies that the neurons that fire most (for which the stimulus is near the peak of their tuning curves) contribute little to the Fisher information because the slope of the tuning curve is small near the peak. Conversely, neurons that respond with intermediate activity, on the side of the tuning curves where the slope is high, contribute maximally to Fisher information. Intuitively, it is clear that the slope of the tuning curve matters in discrimination tasks: Where the slope is high, a small change in the stimulus will cause a large change in mean activity.

Decoders

A decoder takes the population response $\boldsymbol{r}$ as input and produces an estimate $\hat{s}$ of the stimulus s as output.

Decoders can be divided into two categories: those that take the form of neural variability into account and those that do not. The performance of various decoders discussed here is compared in Figure 2 for the example from Figure 1.

Peak response decoder or winner-take-all decoder Suppose that each neuron in the population has a preferred stimulus value. Then, when the population responds to a stimulus, a simple estimator of the stimulus is the preferred stimulus value of the neuron with the highest response:

$$\hat{s}_{\mathrm{WTA}} = s_j : j = \underset{i}{\operatorname{argmax}}\ r_i$$

This very simple decoder disregards the information in all other neurons, is highly sensitive to neuronal variability, and cannot return values intermediate between the preferred stimuli of the neurons. Therefore, it will generally not perform well. However, it only requires knowledge of the preferred stimulus of each neuron, which can be measured experimentally with very few trials.

Weighted-average or population vector decoder A better decoder is obtained by computing a weighted average of the preferred stimulus values of all neurons, with weights proportional to the responses of the respective neurons. Each neuron 'votes' for its preferred stimulus value with a strength proportional to its response:

$$\hat{s}_{\mathrm{WA}} = \frac{\sum_{i=1}^{N} r_i s_i}{\sum_{i=1}^{N} r_i}$$

Although this method performs quite well in many cases, it does not take the form of neuronal variability into account. On a circular stimulus space – for instance, when the stimulus is orientation or motion direction – the equivalent of this method is called the population vector.

Template-matching decoder We can ask the question for which stimulus value s the observed population response is closest to the mean population response generated by s. That is, we match the observed population response with a set of templates (mean population responses for different s). As an error measure, we use the sum-squared difference. This gives the decoder

$$\hat{s}_{\mathrm{TM}} = \underset{s}{\operatorname{argmin}} \sum_{i=1}^{N} (r_i - f_i(s))^2$$

This decoder also does not take the form of neuronal variability into account, but it uses more than only the preferred stimulus values of the neurons.

Maximum-likelihood decoder An important method that does take the form of neuronal variability into account is the maximum-likelihood decoder. This decoder computes the probability that a stimulus value elicited the given population response, and it selects the stimulus value for which this probability is highest:

$$\hat{s}_{\mathrm{ML}} = \underset{s}{\operatorname{argmax}}\ P(\boldsymbol{r}|s)$$

Figure 2(a) shows a histogram of maximum-likelihood estimates obtained from decoding 1000 trials of the population considered in Figure 1. If the number of neurons in the population is large, the maximum-likelihood estimate is unbiased and under some conditions also achieves the Cramer–Rao lower bound. If the noise is independent and normally distributed with fixed variance, the maximum-likelihood decoder is identical to the template-matching decoder. If the tuning curve is

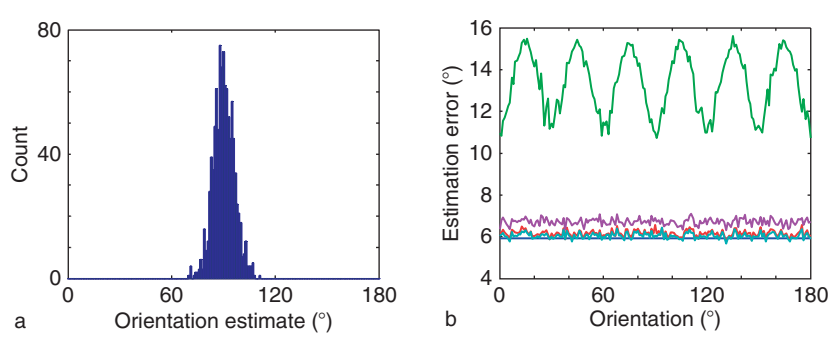

Figure 2 Decoding responses of the population of six neurons of Figure 1. (a) Histogram of maximum-likelihood estimates obtained from 1000 trials of the population responding to an orientation of 90°. The standard deviation of this histogram is the estimation error. (b) The estimation error over 1000 trials as a function of stimulus orientation for several decoders. From top to bottom, the graphs correspond to the peak response decoder (green), the template-matching decoder (purple), the population vector (red), and the maximum-likelihood decoder (cyan). The lowest graph (blue) indicates the Cramer–Rao bound. All decoders are unbiased in this example. Since actual populations usually contain many more than six neurons, estimation errors will be much smaller.

Gaussian and noise is independent Poisson, the maximum-likelihood estimate equals the weighted average.

The problem with this method is that $P(\boldsymbol{r} \mid s)$ must be known with high precision; that is, one needs to measure the tuning curves and the noise distribution of the entire neuronal population. This typically requires a large data set, especially if pairwise correlations are needed.

Bayesian decoders Bayesian decoders use Bayes' rule to express the probability of a stimulus given a response, $P(s \mid \boldsymbol{r})$, as the normalized product of the probability of this response given a stimulus, $P(\boldsymbol{r} \mid s)$, and the prior probability of the stimulus, $P(s)$:

$$P(s|\boldsymbol{r}) = \frac{P(\boldsymbol{r}|s)P(s)}{P(\boldsymbol{r})}$$

The prior probability reflects knowledge about the stimulus before the population response is elicited, and it can have been generated on the basis of previous experience. The probability distribution obtained in this way is referred to as the posterior probability distribution over the stimulus. When the number of neurons is large, this distribution is usually a narrow normal distribution. It can be collapsed onto an estimate by taking the value that has the highest posterior probability; this is called the maximum *a posteriori* (MAP) estimator:

$$\hat{s}_{\mathrm{MAP}} = \underset{s}{\operatorname{argmax}}\ P(s|\boldsymbol{r})$$

Alternatively, the posterior distribution can be collapsed onto an estimate using a loss function $L(s, s')$, which indicates the cost of reporting an estimate s' different from the true value s. The Bayesian estimate is then the value that minimizes the cost averaged over the posterior distribution:

$$\hat{s}_{\mathrm{Bayes}} = \underset{s}{\operatorname{argmin}} \int L(s, s')P(s'|\boldsymbol{r})\mathrm{d}s'$$

If the loss function is the squared error, the Bayesian estimate is the mean of the posterior distribution. For Gaussian posteriors, it is therefore equal to the MAP estimate. If the loss function is the absolute error, the Bayesian estimate is the median.

Like the maximum-likelihood estimator, this approach can be difficult to use in practice because it requires knowledge of $P(\boldsymbol{r} \mid s)$, which is data-intensive.

Neural implementation Decoding is not only of interest to scientists attempting to determine the function of a neural population but also important to the brain. The brain faces the task of generating motor output on the basis of information encoded in a population pattern of activity. The optimal strategy for many cost functions is to pick the motor command that corresponds to the MAP estimate of the stimulus. Recurrent neural networks with population codes have been shown to be able to compute these estimates in some simple cases, but there is still no general theory of Bayes-optimal computations and learning in neural circuits.

Decoding Probability Distributions

As mentioned previously, several groups have proposed that population codes may encode full probability distributions over the stimulus instead of single values $\hat{s}$. The Bayesian decoder described in the previous section falls in this category. Given an observed pattern of activity, it returns a full distribution over the encoded stimulus, denoted $P(s \mid \boldsymbol{r})$. However, this is not the only way to decode a probability distribution from a population code. Other approaches are possible depending on the nature of the encoding steps. Previously, we described several encoding models in which the mean activity of a neuron is directly related to the encoded probability of its preferred stimulus. If this is the case, the decoder should be designed to invert this encoding process. Such decoders are significantly more complicated because they ultimately involve computing a probability distribution, or an estimate, not just over s but over all possible probability distributions over s. The difficulty derives from the fact that even when s is a scalar and therefore lives in a one-dimensional space, the space of probability distribution over s has infinitely many dimensions. Several approximations have been designed to deal with this problem.

Discussion and Open Issues

We have discussed two general ways of encoding information in population codes and evidence for population codes in the brain. We described several methods for obtaining a stimulus estimate from a population response, as well as a new view in which the entire posterior distribution is computed.

An issue we have not addressed is how the performance of a population code can be improved. Population codes and their neuronal decoders can change in ways that are directly coupled to changes in behavior. For instance, attention can serve to improve the performance of a population code by increasing the amount of information that is retained between consecutive stages of processing. The phenomenon of perceptual learning, in which observers reduce the variance of their performance (i.e., their stimulus

estimates) through practice, is thought to be due in part to a similar improvement and in part to an improvement of the decoder used by the brain for the task-relevant population. The mechanisms through which learning can improve population codes are the subject of ongoing research, both empirical and theoretical.

An open issue in population coding is the impact of correlated variability in a population code on the information contained in it. Part of the debate concerns identifying which quantity is suitable to measure this information. Measures in the literature are all based on either Fisher information or Shannon information. The former has the advantage that it is computed for a particular stimulus and has an interpretation in terms of the performance of an ideal observer in a discrimination task. Another part of the debate centers around the assertion that what matters for behavior is not the information that is present in correlations but, rather, the information that would be lost if correlations are neglected in decoding. Evidence shows that neglecting correlations between pairs of neurons in decoding causes only a small loss in Fisher information. An important caveat is that even small correlations can have a significant effect in large populations. In cases in which correlations are important for decoding, a further question is how they can be modeled with a tractable number of parameters – and from an organism's standpoint, how they can be learned. Analyzing empirical population data with a range of decoders will help to elucidate the role of correlations.

Most studies of population codes ignore any information that might be present in spike timing. However, spike timing is known to be important in several perceptual tasks, such as sound localization, and simultaneous firing of large numbers of neurons – synchrony – has even been proposed as an operational principle of the brain. Studies have begun to investigate population coding for dynamic stimuli, which involves decoding population spike patterns, but much work remains to be done on this topic.

Although this article only discussed issues related to coding and decoding, it is important to keep in mind that population codes in the brain are used for computation, which eventually controls behavior. Interestingly, population codes are particularly well suited to optimal nonlinear computations because they provide basis functions. These basis functions can be recombined linearly to implement a very large class of nonlinear computations. Moreover, when embedded in recurrent networks, these computations can be nearly optimal in a Bayesian sense. The latest studies on this topic focus on whether these codes can be used to perform any kind of Bayesian inference, a type of inference that provides a robust approach to computation in the presence of uncertainty.

Despite these open questions, there is little doubt that population codes are a powerful and general way of carrying information in neuronal activity. Closely integrated experimental and theoretical work is needed to fully understand how population codes perform computations and how their decoders are implemented in neural circuitry.

See also: Bayesian Cortical Models; Bayesian Models of Motor Control; Information Coding; Natural Images: Coding Efficiency; Neural Coding of Spatial Representations; Neural Coding in Primary Motor Cortex.

Further Reading

Anderson CH and van Essen DC (1994) Neurobiological computational systems. In: Zurada JM, Marks RJ, and Robinson CJ (eds.) *Computational Intelligence: Imitating Life*, pp. 213–222. Piscataway, NJ: IEEE Press.

Averbeck BB, Latham PE, and Pouget A (2006) Neural correlations, population coding and computation. *Nature Reviews Neuroscience* 7: 358–366.

Dayan P and Abbott LF (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.

Deneve S, Latham P, and Pouget A (2001) Efficient computation and cue integration with noisy population codes. *Nature Neuroscience* 4: 826–831.

Ma WJ, Beck JM, Latham PE, and Pouget A (2006) Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9: 1432–1438.

Pouget A, Dayan P, and Zemel RS (2003) Inference and computation with population codes. *Annual Reviews in Neuroscience* 26: 381–410.

Rao RP (2004) Bayesian computation in recurrent neural circuits. *Neural Computation* 16: 1–38.

Seung S and Sompolinsky H (1993) Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences of the United States of America* 90: 10749–10753.

Zemel R, Dayan P, and Pouget A (1998) Probabilistic interpretation of population code. *Neural Computation* 10: 403–430.

Zhang K and Sejnowski TJ (1999) Neuronal tuning: To sharpen or to broaden? *Neural Computation* 11: 75–84.

Population Coding

A Berkowitz, University of Oklahoma, Norman, OK, USA

© 2009 Elsevier Ltd. All rights reserved.

Population coding is the idea that nervous systems rely on the activity of a large number of relatively imprecise neurons, rather than one or few highly precise neurons, to perform a given function. Each neuron in the large population typically is coarsely or broadly tuned to a particular parameter, rather than sharply tuned, yet the population as a whole usually generates precise outputs.

Early History of the Idea

Trichromatic Theory of Color Vision

Thomas Young was probably the first to suggest population coding. In 1802, he devised this hypothesis to explain how the retina represents the color or wavelength of light at each location:

> As it is almost impossible to conceive of each sensitive point of the retina to contain an infinite number of particles, each capable of vibrating in perfect unison with every possible undulation, it becomes necessary to suppose the number limited, for instance, to the three principal colours, red, yellow, and blue . . . each of the particles is capable of being put in motion more or less forcibly by undulations differing less or more from a perfect unison.

Although Young could not examine the responses of individual retinal 'particles,' modern studies confirm his hypothesis.

Population Coding May Increase Precision

The idea that a large number of coarsely tuned neurons can collectively encode a parameter more precisely than a small number of sharply tuned neurons arose soon after the first electrical recordings were made from individual sensory neurons by Edgar D Adrian in 1926. In 1928, Adrian wrote,

> The accuracy of localization will naturally depend on the number of receptors of the same type in the area which contains the stimulated region. . .. If there is much overlapping in the areas supplied by the terminal branches of different fibres it might be possible to localize the stimulus more exactly by comparing the relative intensities of discharge in the different fibres.

Population Coding as Democracy

A valuable analogy between population coding and political governance was developed early on: single-neuron coding is like a monarchy, whereas population coding is like a democracy. A monarch might make decisions faster, more wisely, and more consistently than a democracy (depending on the monarch), but the collective voting of a large group may be more reliable. For example, if a monarch were to go insane or be assassinated, the entire country would be in peril. Similarly, if an important neural function were dependent on the activity of one neuron, any malfunction or injury to that neuron could jeopardize the animal's life. In contrast, a democracy or a population code would be more resilient in response to injury of several members, due to the redundant inputs of remaining members. Also, in any given election, a minority of members may cast the 'wrong' vote without disrupting the outcome.

William James proposed in 1890 that our consciousness is associated with a single 'pontifical' neuron. Sir Charles S Sherrington, however, considered and rejected this idea in 1941:

> We might imagine this principle pursued to culmination in the final supreme convergence on one ultimate pontifical nerve-cell. . . . It would secure integration by receiving all and dispensing all as unitary arbiter of a totalitarian state. . . . But . . . the nervous system does not integrate itself by centralization upon one pontifical cell. Rather it elaborates a million-fold democracy whose each unit is a cell.

Localized versus Distributed Functions

There has been a long debate about whether the control of a particular behavioral function is associated with one region of the central nervous system (particularly the cerebral cortex) or instead is widely distributed. Localization of function was initially suggested in the late eighteenth century by the phrenologists, beginning with Franz Joseph Gall, who associated human mental functions with particular bumps on the skull. Later, physician-scientists who are now more respected argued for localization of function based on specific behavioral deficits in patients with brain lesions. Others, however, including Marie-Jean-Pierre Flourens in the early nineteenth century and Karl Lashley in the early twentieth century, argued for distributed function based on the lack of deficits in general tasks in many animals and humans with localized cortical lesions.

To some extent, the debate about single-neuron coding versus population coding mirrors the older debate about localization of function versus distributed function. There are important differences, however. One difference is in the level of analysis: a number of neurons versus an entire brain region. A second difference is that a functional neuronal

population is not necessarily constrained to a small and distinct brain area and may be intermixed with populations mediating other functions. A third difference is in the type of function associated with a neuronal population. Earlier studies typically associated one behavior or trait, such as spoken language or aggressiveness, with one cortical area. Recent studies typically associate a neuronal population with a function that is used in multiple behaviors and thus may term the group of neurons a module.

The Alternative: Single-Neuron Coding

To elucidate population coding, it may help to consider its opposite – single-neuron coding. This is the idea that related neuronal inputs converge massively onto a single neuron and/or diverge massively from the neuron, which signals a complex event by its activity alone.

Decision Neurons

For example, sensory neurons whose activity signals the presence of a predator could converge onto a single 'decision' neuron, which in turn could quickly and efficiently trigger a coordinated movement to move the animal away. Such an organization occurs in a marine mollusk (the dorsal ramp interneuron in *Tritonia*), in crayfish (the lateral and medial giant interneurons), and in fish (the Mauthner neuron), although crayfish and fish have an additional neuronal population that can evoke similar movements when the need is less urgent or when the decision neuron has been damaged.

Grandmother Cells

Individual neurons have also been found that respond optimally to a complex and significant object, such as a prey item (e.g., a 'bug detector' in frogs), beginning with experiments in the 1950s. Neurons in a particular part of a songbird's forebrain were later found to respond optimally to playback of that bird's own song and not nearly as well to the same song in reverse or to a conspecific's song. A subset of these cells has been found to be extraordinarily precise temporally as well, firing action potentials reliably within a period of approximately 10 ms just before the animal produces a particular song syllable. Beginning in the late 1960s, neurons in the inferotemporal cortex of monkeys and humans were found to respond optimally to a visually presented hand or face. In recent work, some appeared to respond optimally to a single person's face, regardless of the size, viewpoint, lighting, etc. Since it is not practical to present every possible visual stimulus in such experiments, however, it is possible that the selectivity of these neurons is not as limited as it appears. Such cells have been termed 'grandmother cells,' the idea being that a particular cell might be activated by the sight of one's grandmother and by nothing else.

Hybrid Coding Schemes

Population coding in its narrowest sense involves a large population of broadly tuned neurons, whereas single-neuron coding in its narrowest sense involves one highly specialized neuron for each function. If instead there are multiple grandmother cells, for example, one such cell would be sufficient to signal the sight of one's grandmother, but it would not be necessary since other grandmother cells would remain in the event of injury. Thus, this type of organization might be considered a hybrid between single-neuron and population coding, in which many cells contribute to each function but each such cell is sharply, rather than broadly, tuned and all the cells in this group have the same tuning. In 1972, Horace Barlow argued that the idea of a single, 'pontifical cell'

> should be replaced by a number of 'cardinal cells.' Among the many cardinals only a few speak at once; each makes a complicated statement, but not, of course, as complicated as that of the pontif if he were to express the whole of perception in one utterance.

This idea might be termed few-neuron or oligoneuron coding and considered intermediate between single-neuron and population coding.

The opposite of oligoneuron coding would be a hybrid organization in which a single neuron is responsible for a perception or behavior although the neuron is not highly specialized. This is essentially the idea of the lower envelope principle, which arose from studies beginning in the 1960s correlating sensory perceptions with the tuning curves of individual neurons. The idea is that the brain only uses inputs from the one neuron that is most sensitive to (and thus most strongly activated by) each stimulus. This hypothesis has been supported for perceptual discrimination of the direction of visual motion in primates, but there have been few other experimental tests of it.

Population Coding Is Widespread

Examples of Broad Tuning

A key feature of population coding, narrowly defined, is that each cell in the population is coarsely or broadly tuned. Many examples of broadly tuned neurons have been described, from sensory neurons to premotor neurons, in diverse invertebrates and vertebrates. Figure 1(a) shows an early example. This primate superior colliculus neuron fired action potentials

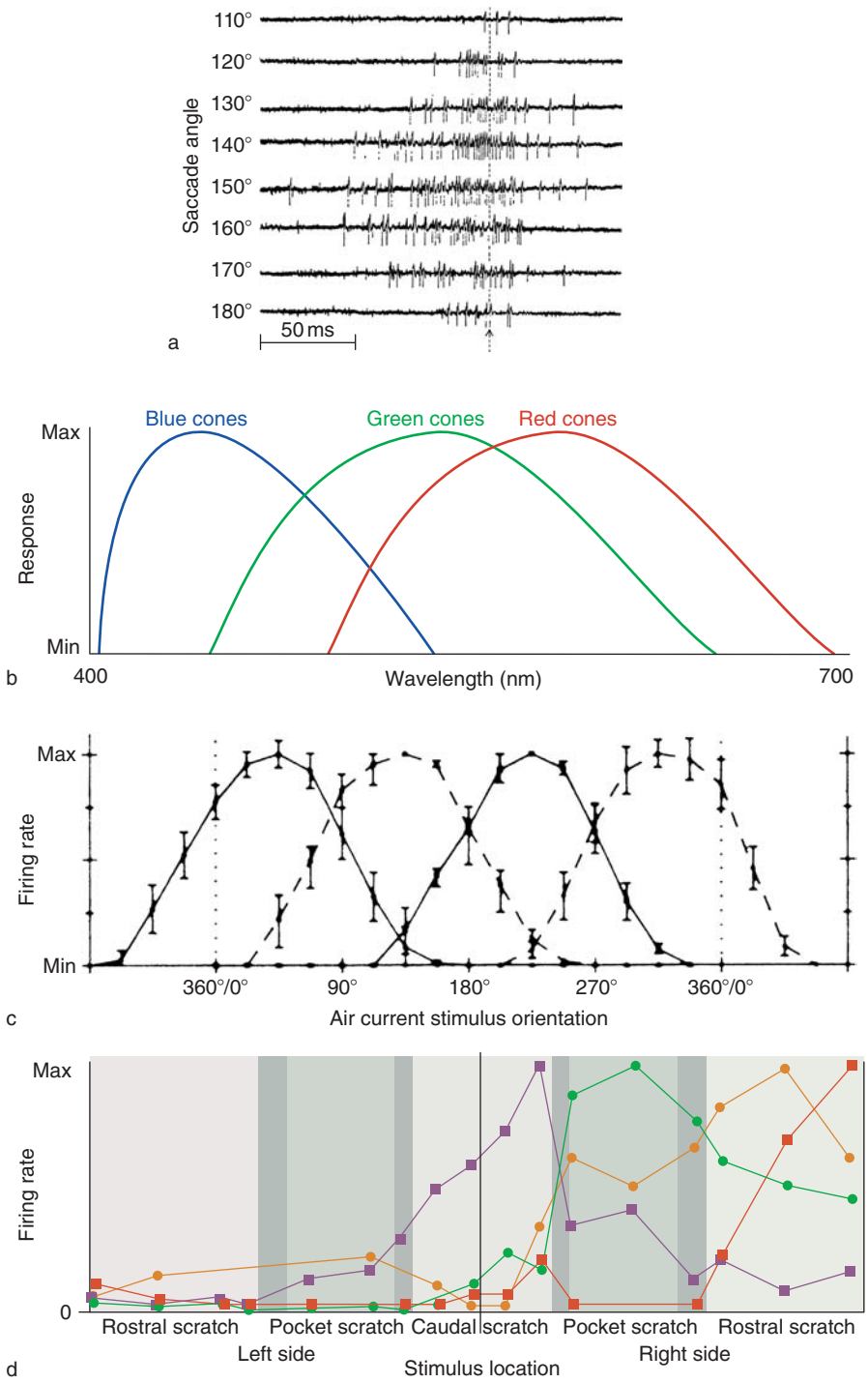

Figure 1 Examples of broadly tuned neurons. (a) Single-unit recordings from an individual neuron in the primate superior colliculus during saccades (arrow) to visual targets at varying angles. (b) Schematic diagram of responses of human cone photoreceptors to varying wavelengths of light. (c) Normalized mean responses of four identified cricket sensory interneurons to wind puffs from varying directions. (d) Normalized mean responses of four turtle scratch-activated spinal interneurons to mechanical stimulation at varying locations on the body surface. (a) Reproduced from Sparks DL, Holland R, and Guthrie BL (1976) Size and distribution of movement fields in monkey superior colliculus. *Brain Research* 113: 21–34, with permission. (c) Reproduced from Miller JP, Jacobs GA, and Theunissen FE (1991) Representation of sensory information in the cricket cercal sensory system. 1: Response properties of the primary interneurons. *Journal of Neurophysiology* 66: 1680–1689, with permission. (d) Reproduced from Berkowitz A (2001) Broadly tuned spinal neurons for each form of fictive scratching in spinal turtles. *Journal of Neurophysiology* 86: 1017–1025, with permission.

before and during saccadic eye movements in a relatively wide range of directions. Nonetheless, the cell fired fastest (and longest) for saccades directed at approximately 150° and progressively less for saccades at progressively different angles. In other words, this cell, which is typical of collicular cells, coarsely encodes a particular saccade direction. The coarseness of tuning was surprising because saccades are quite precise. The authors proposed, however, that the collective activity of the population of coarsely tuned cells results in precise and accurate saccades.

A tuning curve can be produced by graphing a cell's action potential rate as a function of any quantitative parameter. Sharply tuned cells have a sharp peak for one value of the parameter and drop to baseline levels for all larger or smaller parameter values. In contrast, broadly tuned cells respond in varying degrees to a wide range of parameter values. Sensory neurons can be either sharply or broadly tuned, and they can even be sharply tuned to one parameter and broadly tuned to another. For example, each cone photoreceptor is effectively sharply tuned to a light's location but broadly tuned to its wavelength (**Figure 1(b)**). In contrast, olfactory sensory neurons are thought to display only broad tuning to parameters of an odorant molecule (except cells that are selectively responsive to a pheromone).

Further along in sensory processing, cells may be tuned to sensory parameters that are already coded at the periphery (e.g., light wavelength or location, or sound frequency) or instead tuned to sensory parameters that must be computed. For example, neurons in the inferior colliculus of birds (especially well studied in owls) and mammals are tuned to the location of a sound, which is not coded in the cochlea but must be computed from differences in sound timing and intensity in the two ears.

Although sensory neurons can be broadly tuned (and most motor neurons are effectively broadly tuned since they are active during a wide variety of movements, but conventionally they are not described this way), most studies of population coding have focused on neurons within the central nervous system, between sensory neurons and motor neurons, which may be termed interneurons. Some examples of broadly tuned interneurons are shown in **Figures 1(c)** and **1(d)**. Four of the identified sensory interneurons on each side of the body of a cricket collectively respond to wind from all directions behind the cricket and lead to escape movements (**Figure 1(c)**). A sample of four interneurons from the spinal cord of a turtle collectively respond to tickling across one side of the body surface and contribute to hindlimb scratching movements (**Figure 1(d)**). These examples illustrate how a small number of broadly tuned interneurons can effectively cover a large parameter space.

Other examples of broadly tuned interneurons include leech somatosensory interneurons that lead to local bending movements; neurons in the midbrain of weakly electric fish and sound-producing fish that are tuned to the difference in frequency between their own electric field or vocalization and that of their neighbors; rat somatosensory cortical neurons that respond to whisker contact; bat midbrain and cortical neurons tuned to combinations of auditory parameters of the echolocation pulse and echo; cat spinocerebellar neurons that respond to the position or direction of a whole limb; and primate neurons tuned to a variety of sensory and motor parameters in several parts of the visual, somatosensory, auditory, and motor cortices, among many others.

Encoding of Multiple Parameters

Broadly tuned interneurons often encode more than one parameter simultaneously. For example, multisensory neurons in the superior colliculus of birds (especially evident in owls) and mammals encode both the visual and the auditory location of an object; these neurons are arranged in a map of space. Similarly, neurons in the rat hippocampus respond optimally when the animal is in a particular location in space, having (at least) two-dimensional coordinates. Primate motor cortical neurons are often broadly tuned to the activity of one or more muscles, as well as movement amplitude, force, velocity, and direction, to varying degrees. Thus, W Thomas Thach noted in a 1978 study that "all the types of neurons that were looked for were found, in nearly equal numbers." These parameters are normally correlated with one another, and although tasks have been designed to distinguish among them, different investigators have argued in favor of different movement parameters as primarily coded. Some cells are tuned to related sensory and motor parameters, such as movement toward a location to which there is sensory tuning.

In the previous cases, the encoded parameters are closely related conceptually. In other cases, parameters that are not intuitively related are apparently encoded by the same population. For example, many primate motor cortical neurons are broadly tuned to both the direction of an upcoming hand movement and the relative timing of a visual stimulus within a sequence of stimuli. One of these encoded parameters appears to be a component of motor planning and the other of memory. It has been argued that a population may collectively encode parameters that are not at all evident in the activity of individual neurons. Songbird and monkey studies also indicate that the parameters encoded by a population may change, and this may contribute to sensorimotor learning.

A related finding is that individual interneurons can contribute to generating multiple, distinct kinds of behaviors, such as swimming and shortening in leeches, distinct phases of rhythmic digestive movements in crustaceans, swimming and struggling in tadpoles, and scratching and swimming in turtles. Thus, multifunctional populations of neurons mediate many behaviors, especially rhythmic behaviors.

Encoding and Decoding

Whatever parameters are encoded in population activity, signals must be extracted from the population to be of any use to the animal. Thus, downstream neurons must 'read out' or decode population activity. One possibility is that decoding results in an explicit representation of a key feature in the environment (e.g., your grandmother) or a motor command (e.g., an escape movement). In this case, population coding at one stage of processing leads to single-neuron or oligoneuron coding at another stage. A second possibility is that a population encoding sensory parameters directly causes a second population to encode appropriate motor parameters, without any explicit or intermediate representation between them. A third possibility is that multiple populations interact to determine activity in yet another population. In this scenario, the brain uses a series of populations to compute new parameters, eventually producing a distributed motor command in a population of premotor interneurons. A process such as this is probably involved in coordinate transformations. Thus, a visual stimulus is initially represented in retinocentric coordinates, but subsequent processing may lead to coding in head-centered, body-centered, and even extrinsic space coordinates, prior to specifying motor output in joint angle or muscle activity coordinates.

In most cases, it is very difficult to determine whether or how the nervous system actually decodes the activity of a population of neurons. Regardless of how the nervous system does it, having a convenient decoding algorithm allows one to assess whether a given parameter is effectively encoded in a population's activity. Efficient decoding algorithms will also be critical to the success of human prostheses, the movements of which are controlled by decoding the activity of chronically monitored neuronal populations.

The Population Vector

One type of decoding algorithm is called the population vector. The idea is that each cell is broadly tuned to a certain value of the parameter of interest, for example, the direction of an upcoming hand movement. The cell's tuning curve can be approximated as a cosine function of the angle between the cell's 'preferred direction' (i.e., the direction of arm movement for which the cell fires at its highest rate) and the direction of the actual upcoming arm movement. The activity of each cell prior to each hand movement can then be represented as a vector, with the vector angle equal to the cell's preferred direction and the vector length proportional to the cell's increase in firing rate from its baseline level. The population vector is calculated as the sum or average of the vectors of all the cells recorded. It has a direction close to that of the monkey's actual hand movement (**Figure 2(a)**). This shows that hand movement direction is effectively encoded in the activity of motor cortical neurons and that, however the nervous system actually decodes this activity, the behavioral result is similar to what would be produced by population vector decoding. In later experiments, a monkey was trained to rotate a handle 90° counterclockwise from a reference angle given by a visual stimulus. In this case, the population vector gradually rotated from the stimulus direction to the movement direction (**Figure 2(b)**), perhaps revealing a neural correlate of mental rotation.

Other investigators, however, have used mathematical or computational approaches to argue that the predictive success of the population vector of cells' preferred directions does not necessarily imply that motor cortical neurons primarily encode hand movement direction. Motor cortical cells might instead encode muscle parameters, such as shortening velocity. Hand movement direction is typically correlated with muscle parameters and thus the hand-direction population vector should predict arm movements accurately in any case.

The population vector is just one of many possible decoding algorithms and is a special case of linear decoding algorithms. Modeling a neuron's tuning curve with a cosine may be inaccurate and may eliminate valuable information. Other decoding algorithms, such as maximum likelihood estimation and Bayesian inference, have been shown to extract more information from population activity and thus may approach more closely what an ideal observer could decode, but it is unclear whether real nervous systems ever implement such algorithms.

Information Theoretic Approaches

Empirical discoveries of broadly tuned neurons have inspired theoreticians to study population coding using mathematical and computational modeling approaches. Many of these studies have utilized information theory, a branch of probability and statistics, to assess quantitatively the amount of information encoded in a population as well as the efficiency of

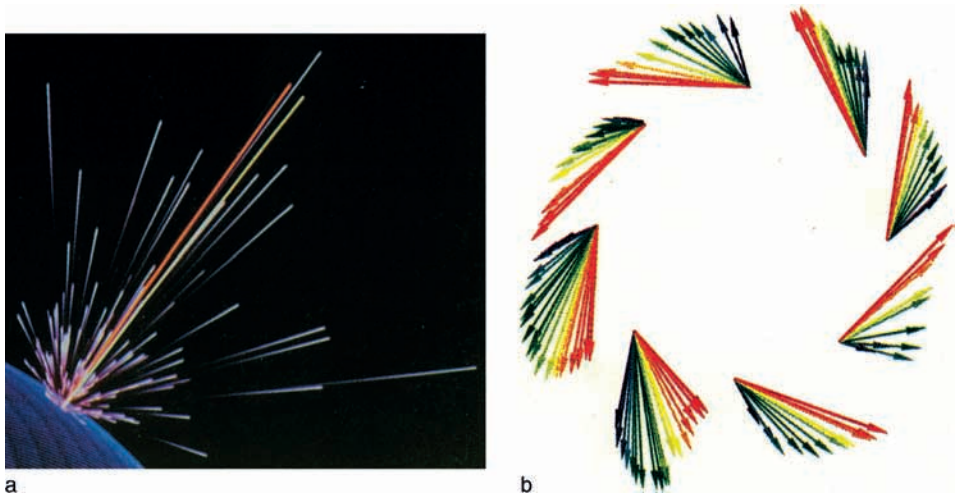

Figure 2 Population vector from monkey motor cortical recordings. (a) Single-cell vectors (light purple) for one direction of hand movement (yellow) and the population vector computed from them (orange). (b) During the reaction time of a mental rotation task, the population vector rotates from the stimulus direction (blue) to the movement direction (red). Eight stimulus directions are shown. (a) Reproduced from Georgopoulos AP, Schwartz AB, and Kettner RE (1986) Neuronal population coding of movement direction. *Science* 233: 1416–1419, with permission. (b) Reproduced from Georgopoulos AP, Taira M, and Lukashin A (1993) Cognitive neurophysiology of the motor cortex. *Science* 260: 47–52, with permission.

various decoding algorithms. These studies have led to several general conclusions, as well as several ongoing debates.

Broad tuning can increase accuracy One conclusion is that given a fixed number of neurons, broadly tuned neurons encode stimulus features more accurately than sharply tuned neurons, effectively because more neurons are activated (and hence 'vote') during each stimulus. This conclusion from mathematical and modeling approaches in the 1980s mirrored Adrian's intuition in 1928. For example, computational modeling of electrosensory processing in weakly electric fish demonstrated that the output of a population of broadly tuned sensory neurons (and thus the animal's behavior) can have greater precision than that of any individual member of the population. This phenomenon, termed 'hyperacuity', has since been demonstrated experimentally in weakly electric fish and bats. (One caveat, however, is that it is difficult or impossible to rule out experimentally that there is some neuron, not yet recorded, with a sensitivity that matches the behavioral sensitivity.)

Subsequent theoretical studies suggested that broader tuning does not necessarily lead to greater accuracy (depending on how the neurons' activity is affected by factors other than the encoded variable), but the relative accuracy of broad tuning compared to sharp tuning always increases with the number of different parameters each neuron codes for (or, in other words, the dimension of the encoded variable). In a population of neurons tuned to three or more parameters, broader tuning is predicted to provide greater accuracy. Moreover, a population in which different neurons display different kinds of tuning (e.g., different shapes or widths of tuning curves) can encode a parameter more accurately than a population with uniform tuning properties. In some cases, such as the coding of electric image amplitude and width by neurons in weakly electric fish, the optimal tuning is broad for one parameter and sharp for the other. Finally, computational modeling approaches have demonstrated that systems with widely distributed coding are more resilient, responding to increasing injury with 'graceful degradation' rather than complete failure (perhaps providing the basis for the relative insensitivity to specific cortical lesions noted by Flourens and Lashley).

The effects of noise One theoretical advantage of using a large population of broadly tuned neurons, rather than a smaller population of sharply tuned neurons, is that variability in neuronal activity unrelated to the encoded variable, often termed noise, tends to be reduced more by averaging across a larger population. In general, for neurons with minimal noise, each neuron provides maximal information for stimuli along the steep flanks of its tuning curve, but for noisy neurons, each provides maximal information for stimuli at its peak. However, noise can have multiple sources. Noise due to stochastic events (e.g., ion channel openings and closings) occurs independently in each of the neurons investigated. In this case, the larger the population averaged, the greater

the noise reduction. Other sources of noise, however, may be correlated among neurons due to common inputs not directly related to the encoded parameter. In this case, noise amplitude is not reduced by averaging, so broader tuning does not necessarily provide greater accuracy. It might be possible, however, to monitor this correlated noise via a separate population, which could then be used to subtract out correlated noise from the main population. Evidence for such a scheme has been found in experiments on weakly electric fish.

In some cases, however, correlations among neurons in the exact timing of their action potentials may themselves carry important information. This kind of population coding can only be examined by recording from multiple neurons in the population simultaneously. The extent to which nervous systems encode information through such temporal correlations is a matter of vigorous debate, with assessments of the additional information ranging from negligible to substantial.

Empirical Tests of Population Coding

Regardless of what information is encoded in the activity of a population of neurons, does the nervous system actually use this information, and if so, how? Downstream neurons might rely on the most active single neuron within the population (the lower envelope principle). There might even be competition within the population, such that the most active neuron inhibits all others (winner-take-all). In this case, the winner would effectively have a veto over other population outputs. Alternatively, the activity of each member of the population might contribute to the output (e.g., by vector averaging or vector addition). Finally, nervous systems might preferentially use signals from simultaneously active neurons so that fine-scale temporal correlations are critical. To address these issues experimentally, one must perturb population activity in a controlled manner and compare the resulting downstream activity or behavior to that predicted by each decoding scheme. This kind of experiment has been feasible in only a few systems.

One fortuitous example is the monkey superior colliculus and its control of saccades. This structure is organized into a map of saccade angles and amplitudes. If one electrically stimulates a particular site, a particular saccade results. One can then inject a local anesthetic into the same location (**Figure 3(a)**). If the nervous system uses only the activity of the most highly active neurons (winner-take-all), the local anesthetic should eliminate all visually evoked saccades toward the location previously evoked electrically (**Figure 3(a)**, arrowhead). In fact, the monkeys generated such saccades accurately. Saccades to other targets, however, were misdirected in a particular pattern: saccades to targets nearer than the electrically evoked saccade were too short, whereas farther saccades were too long, and so on. Thus, the monkeys actually use the activity of neurons in most or all of the colliculus. Movements were perturbed by the absence of 'votes' from the anesthetized region, in accordance with a vector-averaging algorithm.

Another area of the monkey brain, the medial temporal cortex (MT), contains cells selective for the direction of motion of a visual stimulus. Monkeys were given a visual stimulus moving in one direction

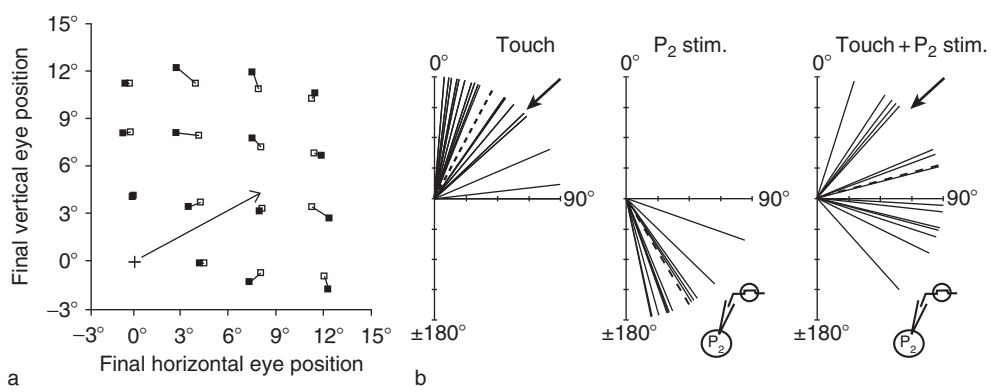

Figure 3 Experimental tests of population decoding algorithms. (a) Monkey saccades from the fixation point (+) toward target lights before (open squares) and after (solid squares) an injection of lidocaine into the superior colliculus; the injection was into the site that evoked the saccade shown by the arrow. (b) Leech bending movement directions evoked by a particular touch stimulus (left), electrical stimulation of a particular sensory neuron (middle), and these two kinds of stimulation together (right). In both cases, the resulting movement directions are consistent with vector averaging of the contributions of all the cells involved. (a) Reprinted by permission from Macmillan Publishers Ltd: *Nature*, Lee C, Rohrer WH, and Sparks DL (1988) Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature* 332: 357–360. (b) Reprinted by permission from Macmillan Publishers Ltd: *Nature*, Lewis JE and Kristan WB Jr. (1998) A neuronal network for computing population vectors in the leech. *Nature* 391: 76–79.

while experimenters electrically stimulated an area of MT that codes for a very different visual motion direction. The monkeys were trained to make a saccade toward one of eight lights to signify their perception of motion direction. Monkeys tended to choose one or the other of the directions indicated by the two simultaneous stimuli but not an intermediate direction, consistent with a winner-take-all decoding mechanism. In a different experiment, however, monkeys were trained to make eye movements to track moving visual stimuli. While the monkeys did so, the experimenters electrically stimulated a site in MT. The electrical stimulation predictably biased, but did not overrule, ongoing eye movements, consistent with vector summation or averaging but inconsistent with winner-take-all decoding. Taken together, these findings suggest that activity in a single population of neurons may be decoded differently by different downstream populations mediating different tasks.

Other fortuitous examples involve small invertebrate systems in which individual neurons can be repeatedly identified and manipulated from one animal to the next. For example, crayfish escape from a sudden touch via a tail flip. A posterior stimulus excites the lateral giant neuron, which causes a forward tail flip; an anterior stimulus excites the medial giant neuron, which causes a backward tail flip. The type of tail flip produced depends on whether the lateral giant or the medial giant fires an action potential first, because each causes inhibition of the other. Stimulation of one of these neurons alone reliably elicits a complete tail flip. Thus, this is a winner-take-all mechanism. (Nonetheless, there is also a distributed population of non-giant neurons that can mediate tail flips in other behavioral contexts.)

The case is quite different for several bilateral pairs of giant interneurons (GIs) in the cockroach, which respond to a wind puff from the right by generating an escape movement to the left. Experimenters reversed this movement decision by stimulating a left GI simultaneously with the right wind puff, causing right turns instead. Movement directions could be predicted from the GI's total number of action potentials, regardless of their exact timing, suggesting that temporal correlations do not contribute critically to the movement decision in this case. Moreover, a higher rate of action potentials in a stimulated GI shifted turning direction by a larger amount, consistent with vector averaging or summation but inconsistent with winner-take-all decoding.

Finally, four P sensory neurons in the leech, acting via interneurons, cause local bending of the body wall away from a touch stimulus. Some interneurons activated during a touch-elicited bend have effects on motor neurons that are antagonistic to the bend produced by the system as a whole (i.e., these cells may be 'outvoted' by the rest of the population). Experimenters electrically stimulated one P cell while touching a body wall location that evoked a response in a different P cell. The two stimuli individually caused bends differing by almost 180°, but the two together produced a bend that was intermediate in direction (**Figure 3(b)**), consistent with population vector decoding but not winner-take-all decoding.

See also: BDNF in Synaptic Plasticity and Memory; Color Vision; Neural Coding in Primary Motor Cortex; Population Codes: Theoretic Aspects; Tactile Coding in Peripheral Neural Populations.

Further Reading

Averbeck BB, Latham PE, and Pouget A (2006) Neural correlations, population coding and computation. *Nature Reviews Neuroscience* 7: 358–366.

Barlow HB (1972) Single units and sensation: A neuron doctrine for perceptual psychology? *Perception* 1: 371–394.

Briggman KL, Abarbanel HDI, and Kristan WB Jr. (2006) From crawling to cognition: Analyzing the dynamical interactions among populations of neurons. *Current Opinion in Neurobiology* 16: 135–144.

Fetz EE (1992) Are movement parameters recognizably coded in the activity of single neurons? *Behavioral and Brain Sciences* 15: 679–690.

Finger S (1994) *Origins of Neuroscience*. New York: Oxford University Press.

Georgopoulos AP, Taira M, and Lukashin A (1993) Cognitive neurophysiology of the motor cortex. *Science* 260: 47–52.

Gross CG (2002) Genealogy of the 'grandmother cell.' *The Neuroscientist* 8: 512–518.

Hinton GE, McClelland JL, and Rumelhart DE (1986) Distributed representations. In: Rumelhart DE and McClelland JL (eds.) *Parallel Distributed Processing: Exploration in the Microstructure of Cognition. Volume 1: Foundations*, pp. 77–109. Cambridge, MA: MIT Press.

Nicolelis MAL (ed.) (2001) Advances in neural population coding *Progress in Brain Research* 130.

Parker AJ and Newsome WT (1998) Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience* 21: 227–277.

Pouget A, Dayan P, and Zemel R (2000) Information processing with population codes. *Nature Reviews Neuroscience* 1: 125–132.

Rieke F, Warland D, de Ruyter van Steveninck R, and Bialek W (1997) *Spikes*. Cambridge, MA: MIT Press.

Sanger TD (2003) Neural population codes. *Current Opinion in Neurobiology* 13: 238–249.

Sparks DL, Kristan WB Jr., and Shaw BK (1997) The role of population coding in the control of movement. In: Stein PSG, Grillner S, Selverston AI, and Stuart DG (eds.) *Neurons, Networks, and Motor Behavior*, pp. 21–32. Cambridge, MA: MIT Press.

Porphyria: Acute Intermittent

W N Arnold, University of Kansas Medical Center, Kansas City, KS, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The various metabolic disturbances that are characterized by excessive synthesis of porphyrins or related compounds collectively are known as porphyrias. Carriers of these diseases are prone to excrete elevated concentrations of the porphyrins and related chemicals in their urine and feces. Abnormal excretion has no intrinsic medical import, but is a reflection of high concentrations of the compounds circulating within the body, and therein lies the potential for neurological abnormalities (due to porphyrin precursors), cutaneous photosensitivity (due to porphyrins), or both, depending upon the particular type of porphyria. Acute intermittent porphyria (AIP) belongs to the first subcategory. The overly produced compounds are δ-aminolevulinic acid (an amino acid) and porphobilinogen (PBG) (a pyrrole), which are intermediates on the metabolic pathway to porphyrin and eventually heme. The symptoms of AIP are all neurological. Increased sensitivity of the skin to sunlight is not observed with this type of porphyria because porphyrins do not accumulate.

'Acute' refers to the rapid onset, and abrupt cessation, of expressed symptoms. 'Intermittent' indicates the periodicity, which is typical, and emphasizes the distinct periods of normalcy, which usually intercede between medical crises. The expression of symptoms depends upon lifestyle, and the adverse effects of malnutrition and exposure to alcohol are well documented. Also, many therapeutic drugs which normally are useful and well-tolerated by the general population will initiate medical crises or provoke relapses in AIP patients. Fortunately, the disease is usually manageable by the avoidance of precipitating factors. The partial deficiency in one enzyme, PBG deaminase, is the underlying cause for the manifold derangements of the AIP patient under crisis. Most of the chemicals that are precipitating factors act by depleting the cytochrome P450 pool in the liver. When considered together, exacerbation factors and latent illness provide a prime example of the influence of the environment on the expression of symptoms due to a congenital disease.

Clinical Manifestations of AIP

Children who are AIP gene carriers remain clinically latent and symptoms rarely occur before puberty. The peak decade for onset of symptoms is from age 20 to 29 years (somewhat higher for males than females), but the disease sometimes remains latent throughout a lifetime. Among those who do express symptoms, there are patients who also have a deficiency in hepatic steroid 5-reductase. How this is acquired or related to AIP is not clear, but it is not due to abnormal porphyrin metabolism *per se*. A close correlation between elevated hormone-binding globulin and clinical expression of AIP has also been demonstrated. Indeed, it has been suggested that certain drugs may precipitate AIP attacks in part because they act through the endocrine system.

Tabulations of the most common hallmarks of AIP emphasize abdominal pain and other gastrointestinal complaints, symptoms referable to the peripheral and central nervous systems, and signs of autonomic neuropathy including tachycardia and hypertension. Porphyria-induced hypertension can cause early-onset renal failure. Bladder dysfunction may result in urinary retention. Effects on optic nerves or the occipital lobes leading to blindness (amaurosis), drooping of both upper eyelids from paralysis of the third cranial nerve (bilateral ptosis), dilation of one pupil, and involuntary rapid, rhythmic eyeball movements (nystagmus) have all been documented for AIP cases. Sexual impotence has occasionally been reported. Signs and symptoms are so varied that individual illnesses may mimic several other diseases. Premonitory symptoms include restlessness and irritability. Attacks develop rapidly. Resolution may occur in days or sometimes weeks, in an unpredictable fashion, but subsequent recovery is then often quite rapid. Convulsions occur in only 10% of all AIP cases, but when they do many of the normally effective antiseizure drugs may adversely affect the outcome. The nature of the disease with respect to both onset of crises and outcome makes an acute attack of AIP particularly frightening because patients can become almost completely paralyzed during severe episodes. The most common cause of death from AIP is respiratory paralysis.

The Biochemical Lesion in AIP

Almost any cell in the human body can engage in synthesis of heme, because heme is not only vital to hemoglobin but also for the cytochromes involved in so many aspects of metabolism. However, the organ of primary concern for AIP is the liver, where two-thirds of the heme that is produced normally is incorporated into the various types of cytochrome P450, and an even larger proportion attains during the induction of P450s. The AIP patient has a vulnerable heme

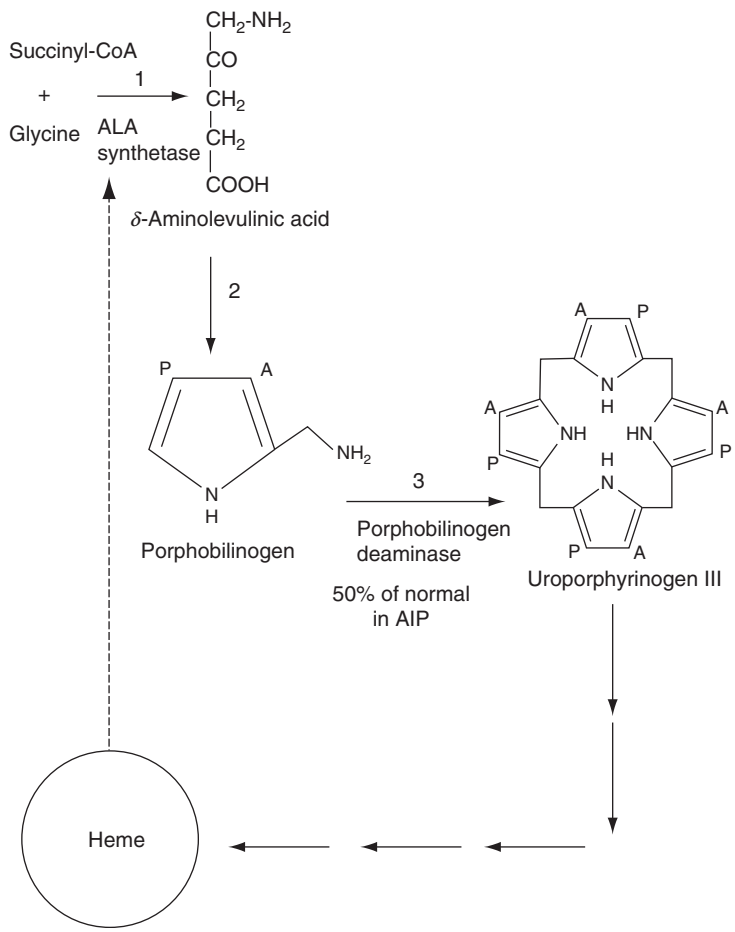

Figure 1 The heme pathway in acute intermittent porphyria (AIP). The dashed arrow from heme up to the first enzyme, δ-aminolevulinic acid (ALA) synthetase, indicates that the concentration of this enzyme, which normally limits the rate of heme synthesis, is influenced by the size of the heme pool (see text for details). In the structures depicting porphobilinogen and uroporphyrinogen III the acetate and propionate residues are indicated as (A) and (p) respectively.

pathway. The neurological problems associated with medical attacks are a consequence of disruption in the heme pathway and the transient accumulation of toxic compounds. An abbreviated representation of the metabolic pathway for heme synthesis is shown in Figure 1.

The key enzyme for our purposes is PBG deaminase, which directs and catalyzes the joining together, head to tail, of four molecules of PBG to form hydroxymethylbilane. This intermediate is subject to nonenzymatic tetrapyrrole ring closure, to form uroporphyrinogen I, but in the presence of another enzyme, uroporphyrinogen III cosynthetase, the major product is uroporphyrinogen III. The process is further complicated by a compartmentalization within cells; thus the early and final steps occur within mitochondria whereas the intermediate events proceed in the cytosol. Traffic of molecules in and out of the mitochondrion is part of the overall process and one aspect of orchestration for the overall pathway. The first enzyme, δ-aminolevulinic acid (ALA) synthetase, catalyzes the reaction which has the most influence on the actual rate of heme synthesis. This enzyme is present in lower amount than any of the subsequent enzymes. In part, this is due to rapid degradation of ALA synthetase within the cell, and it must be continually manufactured. Therein lies a major control feature, because heme, the end product of the pathway, causes both a repression (lower rate of synthesis of the enzyme) and an inhibition (lower catalytic rate of the existing enzyme) of ALA synthetase. Also, the synthesis of this key enzyme can be partially blocked by high glucose intake, thus helping to explain the ameliorating effect of a high-carbohydrate diet on AIP attacks, and the adverse effect of malnutrition or fasting.

One of the first lines of clinical treatment (whatever the precipitating factor) is concentrated glucose infusion. The molecular connection has now been

explained as mediated by a coactivator of nuclear receptors (PGC-1α), which influences both glucose metabolism and the first enzyme in hepatic heme synthesis. Thus PGC-1α is a key factor in the elevation of gluconeogenesis in the fasted liver as well as the increased expression of ALA synthetase. The activator and the enzyme are coregulated during fasting and feeding. . However, it is worth noting that some steroids will induce the synthesis of ALA synthetase as much as 40- to 50-fold and that many drugs also exert their adverse effects independently of the PGC-1α connection.

The individual with AIP has about one-half of the normal amount of PBG deaminase, the third enzyme in the heme pathway. This is the basic, inherited defect. Because PBG deaminase is not rate limiting to the overall pathway, 50% of normal is sufficient for unstressed AIP patients to supply their heme needs. This explains the lack of symptoms for latent AIP patients and the intervening periods of normalcy for patients who have experienced episodes of sickness. It also accommodates the observation that hemoglobin synthesis (in the bone marrow) is not compromised by a partial deficiency in PBG deaminase, and anemia is not a sign of AIP. Exacerbation factors exert their effect by increasing the synthesis of ALA synthetase. The combination of an increase in the first enzyme, which is normally rate limiting, and a preexisting partial block in the third enzyme, PBG deaminase, causes a large increase in the two intermediate metabolites, ALA and PBG. Under these circumstances, their excretion in the urine is significantly increased. Freshly voided urine is a normal color, but with time PBG polymerizes to form porphobilin, which imparts a reddish-brown appearance to aged specimens. Direct light accelerates this process. Urine which has aged internally due to bladder dysfunction may already be discolored when released with a catheter, although the color is sometimes mistaken for urinary bleeding. In contradistinction to the conception that 'port wine urine' is the 'telltale' sign of AIP, it is worth emphasizing that many carriers, with or without medical attacks, have never remarked upon abnormally colored urine.

Neurological Manifestations in AIP

While the overproduction of ALA and PBG is clearly a prerequisite for neurological problems, the mode of action is not completely understood at this time; at least three working hypotheses have been offered. First, ALA or PBG (or a yet-to-be-identified derivative) may be toxic to the nervous system. Increasing amounts accompany acute attacks and, even though increased excretion is sometimes seen in latent cases, neurological dysfunction *per se* never occurs when porphyrin precursor excretion is normal. ALA is structurally similar to γ-aminobutyric acid, a neurotransmitter, and ALA may interfere with γ-aminobutyric acid receptors on nerve cells. An important reservation here is that neither ALA nor PBG can easily cross the blood–brain barrier. A second suggestion arises from investigations of chemically induced hepatic heme depletion in the rat, wherein it was found that liver tryptophan 2,3-dioxygenase (a heme-containing enzyme) activity is lowered. This impairment spares tryptophan degradation, increases the concentration of tryptophan served to the central nervous system, and thus increases serotonin turnover. In the experimental system, exogenous heme reversed the effect whereas tryptophan potentiated it. The serotonin mechanism may play a role in the neurological disturbances associated with AIP. Third, if the vulnerable pathway for heme also operates in the nervous system, there may be more direct neurological effects. Regardless of mechanism, the two forms of useful therapy, namely high-carbohydrate feeding and hematin infusion, both act by suppressing the synthesis of ALA synthetase.

Chemical Exacerbations

Many drugs are metabolized in the liver by a system which involves the endoplasmic reticulum and cytochrome P450. This heme protein is subject to destruction by a free-radical reaction during the detoxification process, and the induction of P450s by drugs and xenobiotics further increases consumption to the point of becoming a drain on the heme pool. Excessive alcohol ingestion leads to a proliferation of endoplasmic reticulum in liver cells, and part of the ethanol ingested is metabolized therein via a cytochrome P450-mediated system. In this manner, alcohol and other drugs which lower the heme pool are thought to exert their influence on the overall pathway by derepressing (hence increasing) the synthesis of ALA synthetase. In the AIP patient, with half of the normal amount of PBG deaminase, the insult may be too great and toxic levels of ALA and PBG may develop. One system that was developed expressly to test potential porphyrinogenic drugs depends upon measuring the induction of ALA synthetase in fetal chicken liver cells that have been exposed to test compounds. Early examples of AIP were revealed as a response to new drugs – initially the hypnotic sulfonal, later barbiturates, and subsequently many other drugs, alcohol, and sundry organic compounds. Some steroid metabolites precipitate attacks, and endogenous changes may account for crises at puberty, for the earlier

onset with females, and for the relationships with the menstrual cycle and oral contraceptives. Other exacerbating factors include infections and malnutrition. Low-carbohydrate and low-protein diets are especially detrimental, and fasting can precipitate an attack of porphyria. A study in Scotland indicated an association between smoking and the induction of repeated attacks in patients already diagnosed with AIP.

Genetics

Barker and Estes, in 1912, were the first to note that AIP runs in families. The extensive studies of Paul Waldenström in 1957 in Sweden firmly established the inherited nature. The disease follows an autosomal dominant pattern of inheritance; if one parent is a carrier then on the average 50% of the children will bear the defective gene. However, the penetrance is variable, with only a fraction of the carriers usually expressing the disease. The enzymatic lesion is PBG deaminase and the gene for the enzyme is localized to the distal long arm of chromosome 11. Mutation within this gene leads to a defective protein product, one that has no PBG deaminase activity. This aberrant product in about 80% of cases does not survive long enough within cells to be detected by reaction with antibodies to the normal protein. Most of the remainder of cases exhibit a defective protein that cross-reacts with appropriate antibodies. A very small group apparently overproduces a different aberrant protein or (more likely) makes an altered protein that has a longer life by resisting cellular degradation. Whether this heterogeneity accounts for differences among patients in the susceptibility to exacerbation factors, and the clinical presentation, remains to be elucidated.

Famous Patients

King George III of Great Britain suffered from intermittent attacks of mental derangement which have been attributed to a form of porphyria by Macalpine and Hunter. A good case for acute intermittent porphyria was initially presented, although these authors later seemed to favor another type, variegate porphyria, because of some less well-documented claims of skin sensitivity in the subject. The illness of Vincent van Gogh has been attributed by Loftus and Arnold in the early 1990s to a toxic psychosis, and acute intermittent porphyria is the best working hypothesis. The artist suffered from severe and manifold neurological problems ranging from gastrointestinal pains to fits with hallucinations. His condition was exacerbated by his lifestyle, which was marked by inadequate nutrition, abuse of alcoholic beverages, chronic smoking, and environmental stress. His several debilitating crises were of abrupt onset and rapid resolution, and their initiation often coincided with exposure to factors now known to provoke attacks in AIP patients. The age of onset of symptoms and the incidence of mental illness in the van Gogh family also support this working hypothesis. It is clear that Vincent van Gogh was both productive and creative in spite of his illness, not because of it. Some diseases with the potential for severe neurological involvement are not incompatible with creativity.

Concluding Remarks

The intervals between presenting symptoms and the correct diagnosis of AIP have been reported to range from several months to years. Because of their psychoses, carriers may be institutionalized, and many of these individuals are never properly diagnosed. The incidence of AIP is about 1 in 50 000 within all ethnic groups (about 1 in 10 000 in Sweden). Because penetrance is low it is generally assumed that the rate for the aberrant gene is underestimated. In this context, it is also worth noting that when psychiatric institutions were surveyed for AIP, order of magnitude increases in rate compared with local populations were observed, due to the revelation of previously undiagnosed cases. Because abdominal pains are frequent, AIP patients are often referred to the surgical ward and come under the knife unnecessarily. Many otherwise useful drugs will actually precipitate attacks and add to the suffering. Lists of safe drugs and also those to be avoided by AIP patients are published, and if correctly managed the AIP patient can lead a happy and productive life. Elucidation of the mechanisms involved in this particular congenital disease serves to emphasize by contrast the importance of metabolic orchestration to health. An adenoviral vector has been used to augment a deficiency of PBG deaminase in the liver of an AIP mouse model. The expressed enzyme enabled the animal to avoid ALA and PBG accumulations which were otherwise encountered in the model in response to a phenobarbital challenge. A gene therapy for AIP in humans may be possible in the future.

See also: Autonomic Nervous System; Gamma-Aminobutyric Acid (GABA); Glial Glutamate and GABA Metabolism; Glutamatergic and Gabaergic Systems; Neuropathy: Metabolically-Induced.

Further Reading

Arnold WN (1992) *Vincent van Gogh: Chemicals, Crises, and Creativity.* Boston: Birkhäuser.

Barker LF and Estes WL (1912) Family hematoporphyrinuria in association with chronic gastroduodenal dilation, peculiar fits and acute polyneuritis: A preliminary report. *Journal of the American Medical Association* 59: 718–719.

Becker DM and Kramer S (1977) The neurological manifestations of porphyria: A review. *Medicine* 56: 411–423.

Granick S (1966) The induction *in vitro* of the synthesis of δ-aminolevulinic acid synthetase in chemical porphyria: A response to certain drugs, sex hormones, and foreign chemicals. *Journal of Biological Chemistry* 241: 1359–1375.

Handschin C, Lin J, Rhee J, et al. (2005) Nutritional regulation of hepatic heme biosynthesis and porphyria through PGC-1α. *Cell* 122: 505–515.

Johansson A, Nowak G, Möller C, et al. (2004) Adenoviral-mediated expression of porphobilinogen deaminase in liver restores the metabolic defect in a mouse model of acute intermittent porphyria. *Molecular Therapy* 10: 337–343.

Kappas A, Sassa S, Galbraith RA, et al. (1989) The porphyrias. In: Scriver CR, Beaudet AL, Sly WS, et al. (eds.) *The Metabolic Basis of Inherited Disease,* 6th edn., pp. 1305–1365. New York: McGraw-Hill.

Loftus LS and Arnold WN (1991) Vincent van Gogh's illness: Acute intermittent porphyria? *British Medical Journal* 303: 1589–1591.

Macalpine I and Hunter R (1966) The "insanity" of King George II: A classic case of porphyria. *British Medical Journal* 1: 65–72.

Meyer UA, Strand LJ, Doss M, et al. (1972) Intermittent acute porphyria: Demonstration of a genetic defect in porphobilinogen metabolism. *New England Journal of Medicine* 286: 1277–1282.

Moore MR (1980) International review of drugs in acute porphyria. *International Journal of Biochemistry* 12: 1089–1097.

Nordmann Y, Puy H, Da Silva V, et al. (1997) Acute intermittent porphyria: Prevalence of mutations in the porphobilinogen deaminase gene in blood donors in France. *Journal of Internal Medicine* 242: 213–217.

Ridley A (1969) The neuropathy of acute intermittent porphyria. *Quarterly Journal of Medicine* 38: 307–333.

Stokvis BJ (1889) Over twee zeldzame kleurstoffen in urine van zieken. *Weekblad van het Nederlandsch Tijdschrift voor Geneeskunde* 2: 409–417.

Strand LJ, Felsher BF, Redeker AG, et al. (1970) Enzymatic abnormality in heme biosynthesis in acute intermittent porphyria: Decreased hepatic conversion of porphobilinogen to porphyrins and increased δ-aminolevulinic acid synthetase activity. *Proceedings of the National Academy of Sciences of the United States of America* 67: 1315–1320.

Tishler PV, Woodward B, O'Connor J, et al. (1985) High prevalence of intermittent acute porphyria in a psychiatric patient population. *American Journal of Psychiatry* 142: 1430–1436.

Waldenström J (1957) The porphyrias as inborn errors of metabolism. *American Journal of Medicine* 22: 758–773.

With TK (1980) A short history of porphyrins and the porphyrias. *International Journal of Biochemistry* 11: 189–200.

Positron Emission Tomography (PET)

M Reivich, University of Pennsylvania, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Positron emission tomography (PET) allows quantitative measurements of regional physiological and biochemical processes to be made *in vivo* in humans. The approach consists of labeling a tracer with a positron-emitting radionuclide, the administration of this tracer to the subject, the determination of the three-dimensional distribution of the positron activity in the organ being studied by means of a positron emission tomograph, and the quantitation of the parameter under study by means of a suitable mathematical model. Frequently, the arterial input function – that is, the time course of change of the concentration of the tracer in the arterial blood – also needs to be measured in order to quantitate the parameter of interest.

The positron is a particle with a positive charge and a mass equal to that of an electron. After the emission of the positron from a radioactive nuclide, the particle interacts with an electron and the two particles undergo the process of matter–antimatter annihilation in which their masses are converted into electromagnetic radiation in the form of two photons, each with an energy of 511 keV, which travel in directions nearly 180° apart. It is the detection of this annihilation radiation by the PET scanner that is used to produce an image of the distribution of the positron activity in the tissue.

The detection and collimation of the annihilation radiation in PET devices is achieved by having an array of radiation detectors in which two detectors on opposite sides of the array are electronically connected by a coincidence circuit such that an event is registered only if both detectors sense the annihilation photons virtually simultaneously. These events are then reconstructed into an image of the distribution of these annihilation events within the organ of interest using algorithms similar to those developed for computerized tomography (CT) scanners.

Radiopharmaceuticals

The most frequently used positron-emitting radionuclides are isotopes of oxygen, nitrogen, and carbon (commonly occurring constituents of organic molecules), as well as fluorine. The specific isotopes used are oxygen-15, nitrogen-13, carbon-11, and fluorine-18 with half-lives of 2, 10, 20, and 110 min, respectively. These short half-lives permit the administration of relatively large doses of radioactivity to the subject with radiation exposures within established safe limits. They also mean that these radionuclides must be produced close to where they will be used, requiring a cyclotron or other accelerator on site.

The development of tracers for physiological and biochemical measurements in the brain with PET has focused mainly on labeled compounds to measure local cerebral blood flow, cerebral blood volume, oxygen consumption, glucose consumption, protein synthesis, neurotransmitter receptor properties, tissue proliferation, tissue pH, gene expression, and drug distribution. Table 1 provides a representative list of these tracers.

In addition to cyclotron-produced positron emitters, many potentially useful positron emitter generator systems have been developed. These systems offer convenient access to a choice of positron emitters with a wide range of half-lives and a variety of chemical properties for use in the preparation of labeled tracer molecules. The most commonly used positron emitter generator systems are ^{68}Ge/^{68}Ga and ^{82}Sr/^{82}Rb.

PET studies have added significantly to our understanding of metabolic and physiological processes in the human brain in both normal and pathologic conditions.

Sensory and Cognitive Stimuli in Normal Subjects

The effects of various sensory and cognitive stimuli on regional cerebral blood flow (rCBF) and metabolism have been examined. Monaural auditory stimulation elevates the metabolic rate for glucose in the temporal cortex contralateral to the stimulated ear. A tonal memory test has also been examined. In the tonal memory test, significant increases in glucose metabolism were found in the right posterior superior and middle temporal cortices. Those subjects employing a nonanalytic strategy for remembering the sequence of tones had higher metabolic rates in the right frontal and parietotemporal regions. In those subjects using an analytic approach, higher metabolic rates occurred in the left posterior superior temporal cortex. Similar differential effects have been seen in studies of memory with auditory presentation of facts. Subjects using the strategy of silently rehearsing the facts to themselves activated left frontal and right cerebellar regions. Subjects using the strategy of creating a vivid visual image of the facts showed activation of the prefrontal, intraparietal, and posterior superior parietal regions.

Several studies of the effects of visual stimulation on local cerebral glucose metabolism have been

Table 1 Representative PET tracers

Function	*Tracer*
Blood flow	^{11}C-iodoantipyrine
	$C^{15}O_2$
	$H^{15}O_2$
Blood volume	$C^{15}O$
	^{11}CO
Oxygen metabolism	$^{15}O_2$
Glucose metabolism	^{18}F-fluorodeoxyglucose
	^{11}C-deoxyglucose
Protein synthesis	^{11}C-leucine
	^{11}C-methionine
D_2dopamine receptors	^{11}C-raclopride
	^{18}F-*N*-methylspiroperidol
	^{11}C-spiroperidol
D_1 dopamine receptors	^{11}C-SKF 38393
	^{11}C-SCH 23390
Dopamine reuptake site	^{11}C-nomifensine
	^{11}C-cocaine
Vesicular monoamine transporter 2	^{11}C-tetrabenazine
	^{11}C-dihydrotetrabenazine
Dopamine	^{18}F-fluorodopa
Central benzodiazapine receptors	^{11}C-flumazenil
Serotonin reuptake site	^{11}C-McN 5652
Opiate receptors	^{11}C-carfentamil
	^{11}C-diprenorphine
Acetylcholine muscarinic receptors	^{11}C-quinuclidinyl benzilate methiodide
Gene expression	^{18}F-ganciclovir
	^{18}F-fluoropenciclovir
	^{18}F-fluorohydroxy propoxy methyl guanine
	^{18}F-fluoroethyl spiperone
	^{124}I-fluorodeoxy arabinofuranosyl iodouracil
Tissue proliferation	^{11}C-purines, nucleotides, and nucleosides
	^{11}C-putrescine
Tissue pH	$^{11}CO_2$
	^{11}C-dimethyloxazoladine

reported. In the visual cortex contralateral to the stimulated hemifield, a significant increase in metabolism was observed. When the visual stimulus was limited to the central 20° (macular region) of one hemifield and the peripheral 60° of the other hemifield, increases in metabolism in the macular and peripheral representations of the visual fields in the visual cortex were observed. Tactile stimulation produces an increase in metabolism in the contralateral postcentral gyrus and surrounding areas.

The effects of cognitive tasks have been studied. Verbal and spatial tasks produced the expected increase in metabolism in the left and right hemispheres, respectively, mainly in the superior temporal and inferior parietal regions. In addition, there were significant differences in the laterality of metabolism in the frontal eye fields of the prefrontal cortex which may be related to a postulated neural network that links cognitive activity with motor orientation.

Studies of episodic memory using a semantic analysis strategy have demonstrated involvement of the left inferior prefrontal cortex. In another study involving the recollection of previous events, increases in blood flow were seen in the right dorsolateral prefrontal cortex and the anterior portion of the premotor area as well as the left anterior cingulate gyrus and in the parietal lobes bilaterally. These findings may reflect the operation of a neuronal network involving prefrontal and parietal cortical regions subserving memory-related processes.

The temporary storage and manipulation of information is thought to be processed in a system termed working memory. This system consists of an attentional controller and the central executive, aided by two subsidiary systems – the visuospatial sketchpad and the phonological or articulatory loop. The visuospatial sketchpad holds and manipulates visual images, and the phonological or articulatory loop performs a similar function for speech-based information.

Verbal storage tasks result in activations in left posterior parietal cortex and three frontal sites (Broca's area, the left supplementary motor, and the premotor area). PET studies also have shown that subvocal rehearsal of phonological information is neurally implemented by left hemisphere speech areas, including Broca's area, the premotor area, and the supplementary motor area.

The central executive is assumed to be responsible for the attentional control of working memory. When a task requires processing the contents of working memory, the dorsolateral prefrontal cortex is activated.

It seems likely that visual and spatial information is handled by regions of the brain separate from those handling verbal information. Spatial information is represented more dorsally (superior parietal cortex) than object information (inferior temporal cortex). Other studies suggest that object information also appears to be more left lateralized in the ventrolateral frontal cortex (VLFC) with spatial information in the right VLFC.

Studies of patients with focal lesions have demonstrated the importance of medial temporal and diencephalic regions in long-term memory (LTM) tasks. Imaging studies have demonstrated frontal cortex activation during LTM tasks that are related to memory encoding and retrieval. The left frontal cortex (predominantly VLFC and less so dorsolateral frontal cortex (DLFC)) is activated during verbal encoding and the right frontal cortex during visuospatial encoding. Memory retrieval has been postulated to involve two main stages: specification of

search parameters and verification of memories retrieved. Imaging studies suggest that these two stages activate VLFC and DLFC, respectively, more so on the right than the left.

Cerebrovascular Disease

PET has provided important data concerning the pathophysiological events occurring after a stroke. Measurements of rCBF and metabolism have provided useful information in staging the severity of tissue injury and predicting outcome as well as showing alterations in these parameters in response to therapy.

The relationship between metabolism and flow in the acute stroke patient may also provide prognostic information. The oxygen extraction fraction (OEF), which is the ratio of oxygen consumption to blood flow, may be an index of the severity of tissue injury. Early in the patient's course, cerebral metabolism may be normal but flow reduced and manifest by an elevated OEF. This disproportional decrease in flow compared to metabolism has been termed 'misery perfusion.' There may be some critical level of increased OEF beyond which tissue ischemia leads to irreversible damage. A reduction in oxygen metabolism from the normal level of 3.75 ml $100\,g^{-1}\,min^{-1}$ to a level of 1.25 ml $100\,g^{-1}\,min^{-1}$ or a reduction of flow to approximately 12 ml $100\,g^{-1}\,min^{-1}$ has been reported to be associated with infarcted irreversibly damaged tissue. An ischemic penumbral region has been defined in which flows are at a level within the threshold of functional impairment but morphologic integrity is preserved (i.e., a flow between 12 and 22 ml $100\,g^{-1}\,min^{-1}$) and which has the capacity to recover if perfusion is improved. In the penumbral region, there is a cerebral metabolic rate of oxygen ($CMRO_2$) >1.4 ml $100\,m^{-1}\,min^{-1}$ and an OEF >0.7. The extent of the penumbra is large if measured in the first hours of ischemia and smaller if measured later. It has been demonstrated that spontaneous neurologic recovery after stroke is significantly related to the volume of penumbric tissue that escapes infarction.

In some patients within a few hours following a stroke, flow may be increased relative to oxygen metabolism, with a resulting decrease in OEF. Again, flow and metabolism are uncoupled but now in the opposite direction, which has been termed 'luxury perfusion.' A later phase of hyperemia between 10 and 20 days following a stroke may occur, which is due to capillary hyperplasia in tissue undergoing cavitation. The prognostic significance of luxury perfusion is unclear, but in some studies it has been associated with good recovery of the tissue and a good clinical outcome. The later hyperemic phase, however, seems always to be associated with infarction. In the chronic phase, approximately 1 month after the ictus, there may be recoupling of flow and metabolism but at a reduced level.

Central benzodiazepine receptor ligands have been suggested as a marker of neuronal integrity that can identify irreversibly damaged tissue irrespective of the time elapsed since the onset of ischemia. In patients with acute ischemic strokes studied within 3–16 h of onset, cortical regions with decreased ^{11}C-flumazenil (FMZ) binding predicted final infarcts seen on magnetic resonance imaging (MRI) or CT scans. The predictive value of reduced FMZ binding was comparable to that of regional $CMRO_2$ reduction less than 1.4 ml $100\,g^{-1}\,min^{-1}$. Studies with ^{18}F-fluoromisonidazole, a tracer that is trapped within hypoxic cells with still active metabolism, have demonstrated viable tissue in the peri-infarct region as long as 42 h after a stroke. This evidence of relatively long-lasting penumbral tissue suggests that the time for therapeutic intervention may be protracted in some patients and that PET studies may be able to identify those patients.

PET studies have also shed light on the mechanisms of recovery of function following a stroke. It has been a consistent finding in several studies that increased glucose metabolism and blood flow occurs in the ipsilateral sensorimotor cortex as well as in the contralateral primary and secondary sensorimotor cortices with movements of the paretic hand after recovery from stroke compared to normal hand movement (Figure 1). PET studies of rCBF have shown in patients recovering from aphasia that during speech there are areas of activation both in the right hemisphere and in spared left hemispheric regions (Figure 2). Thus, there is evidence of neural plasticity playing a role in functional recovery after stroke.

In patients with carotid artery disease, the presence of an increased OEF is a powerful predictor of future stroke. Reversal of a focal increase in OEF has been demonstrated following endarterectomy or superficial temporal artery to middle cerebral artery anastomosis in patients with transient ischemic attacks.

Hemodynamic and metabolic abnormalities may be present in the tissues distant from an ischemic area which appear normal on CT or MRI scans. These remote effects are presumed to be secondary to metabolic alterations produced by the decrease in axonal input to the region and have been termed diaschisis. Such changes have been documented in the contralateral cerebellar hemisphere, the non-ischemic hemisphere, as well as ipsilateral cortical regions outside the area of infarction. PET studies have shown that there is a relationship between the resolution of diaschisis and subsequent recovery.

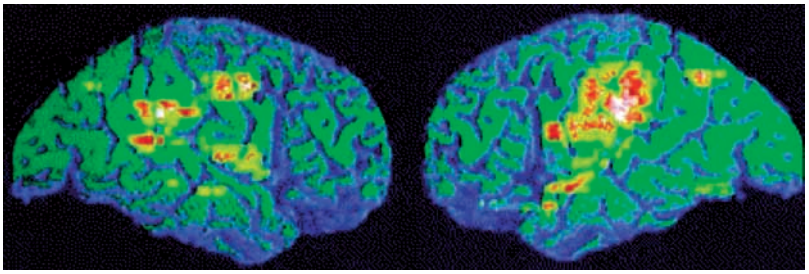

Figure 1 rCBF activation study in a patient who had recovered well after right-sided paresis due to a subcortical stroke. Shown are areas activated more than in ten normal subjects when they are all making movements with the right hand. The patient shows more activation in the left primary sensorimotor cortex and in secondary sensorimotor areas, all contralateral to the recovered hand (right). Also ipsilaterally(left), secondary sensorimotor areas are activated more than in the normal subjects. Reproduced from Rijntjes M and Weiller C (2002) Recovery of motor and language abilities after stroke: The contribution of functional imaging. *Progress in Neurobiology* 66: 109–122, with permission from Elsevier.

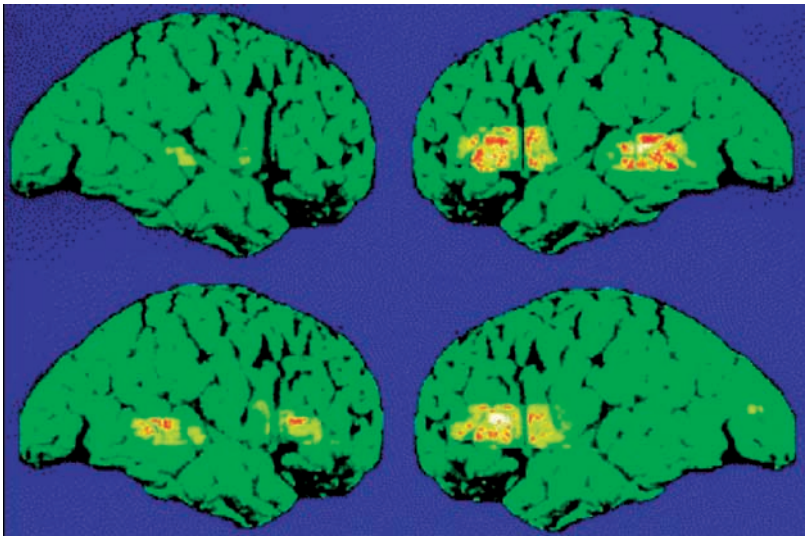

Figure 2 rCBF study showing regions activated by a verbal task in six normal subjects (top) and six patients who had recovered from Wernicke's aphasia (bottom) due to a stroke in the left posterior language regions. All subjects and patients were strongly right-handed. The right hemispheres are shown on the left and the left hemispheres on the right. In the normal subjects, there was activation in the anterior and posterior language regions in the left hemisphere, with weaker activation in homologous areas of the right hemisphere. In the patients, the left posterior language region was destroyed, so no activation was present there. There was still activation in the left frontal language region. Homologous regions in the right hemisphere are activated stronger than in the normal subjects. Reproduced from Rijntjes M and Weiller C (2002) Recovery of motor and language abilities after stroke: The contribution of functional imaging. *Progress in Neurobiology* 66:109–122, with permission from Elsevier.

Changes in flow, metabolism, and receptor binding in acute stroke patients in response to therapeutic interventions have been shown to occur and to correlate with functional recovery.

Cerebral Tumors

PET measurements can play an important role in the evaluation and management of patients with brain tumors, including the grading of tumors, determination of prognosis, and the differentiation of recurrent tumor from radiation necrosis.

Measurements of cerebral glucose metabolism have been used in an attempt to grade the malignancy of brain tumors *in vivo*. This approach is based on evidence in the literature suggesting that there is a progressive increase in the degree of anaerobic glycolysis of a glioma the higher its grade. A positive correlation between glioma grade and glycolytic activity as measured by the ^{18}fluorodeoxyglucose (FDG) technique has been observed (Figure 3).

FDG–PET has also been useful in determining prognosis in patients with brain tumors. PET can more accurately predict the degree of malignancy of a tumor than CT or MRI, and PET has also been

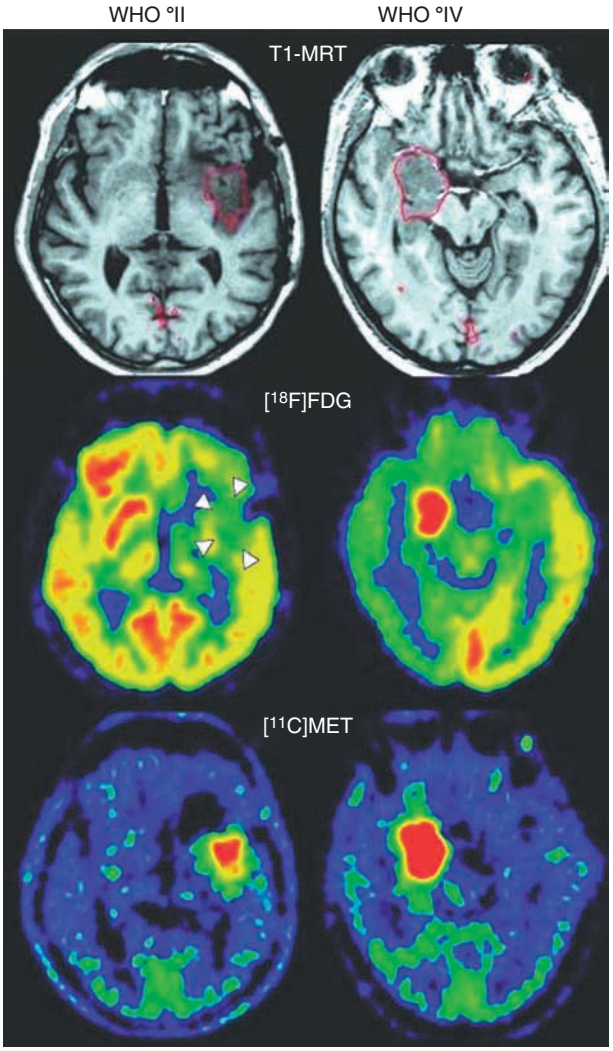

Figure 3 Differentiation between low- and high-grade glioma. In low-grade gliomas (WHO °II) glucose metabolism is similar to white matter (arrows) and amino acid uptake is only moderately increased. In high-grade gliomas (WHO °IV) both glucose metabolism and amino acid uptake are increased. Reproduced from Jacobs AH, Kracht LW, Gossmann A, et al. (2005) Imaging in neurooncology. *NeuroRx* 2: 333–347, with permission.

shown to predict survival in patients with gliomas more accurately than either CT or MRI.

FDG uptake has been shown to be useful in distinguishing between radiation necrosis and tumor recurrence, which is typically difficult with either CT or MRI. In general, areas of radiation necrosis are hypometabolic, whereas tumor recurrence appears hypermetabolic. However, accumulation of FDG in infiltrating macrophages may obscure this delineation. [^{11}C]methionine is better suited in this regard and differentiates between tumor recurrence and radiation necrosis with high sensitivity and specificity.

FDG–PET has also been used to determine tumor response to radiation and/or chemotherapy and patient survival. Patients whose tumors have decreased glucose metabolism after radiation therapy tend to have a better prognosis than those who do not show such a response (Figure 4).

[^{11}C]methionine has been used to help determine the fields for radiation therapy because the tumor volume depicted by [^{11}C]methionine is larger than that depicted by MRI. Benodiazepine receptors are increased in brain tumors compared with normal brain, and thus [^{11}C]flumazenil, a benzodiazepine antagonist, has been useful in depicting tumor volume as well.

Movement Disorders

Parkinsonism is a feature of a number of neurodegenerative diseases, including Parkinson's disease (PD), multiple system atrophy (MSA) and progressive supranuclear palsy (PSP). Separating these various parkinsonian syndromes using clinical criteria alone can be difficult. Discrimination of these various disorders is important in view of differences in prognosis and therapy. These syndromes can be differentiated partially on the basis of the involvement of different components of the dopaminergic system.

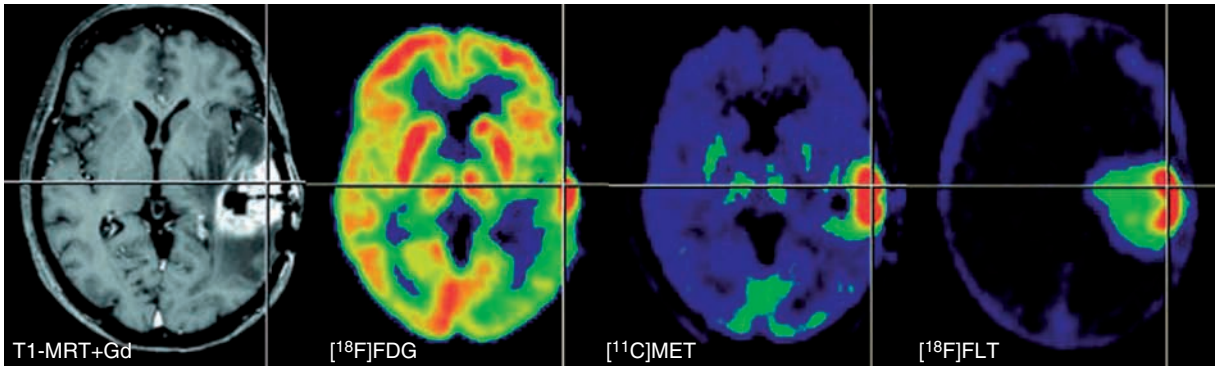

Figure 4 Parameters of interest in the noninvasive diagnosis of brain tumors. Alteration of the blood–brain barrier and the extent of peritumoral edema are detected by MRI. Signs of increased cell proliferation can be observed by means of multitracer PET imaging using [^{18}F]FDG, [^{11}C]MET, and [^{18}F]FLT as specific tracers for glucose metabolism, amino acid transport, and DNA synthesis, respectively. Secondary phenomena, such as inactivation of ipsilateral cortical cerebral glucose metabolism, may be observed and are of prognostic value. Reproduced from Jacobs AH, Kracht LW, Gossmann A, et al. (2005) Imaging in neurooncology. *NeuroRx* 2: 333–347, with permission.

The major cause of parkinsonism is PD, accounting for 60–85% of all cases. In PD presynaptic uptake of ^{18}F-dopa is significantly decreased in the striatum, being most marked in the posterior part of the putamen while being relatively preserved in the caudate (Figure 5). In patients with atypical parkinsonism (lack of resting tremor and poor response to dopaminergic drugs), both the caudate and putamen are equally affected. Studies of ^{18}F-dopa appear to be sensitive enough to detect preclinical PD. The rate of decline in ^{18}F-dopa uptake over time is significantly faster in PD than in normals. Studies of ^{18}F-dopa have been used to assess the role of genetics versus environment in PD. Studies of nigrostriatal dysfunction in monozygotic and dizygotic twins showed significantly higher concordances in the monozygotic twins which argue strongly against an environmental factor and for a genetic cause of PD.

D_2 receptor binding is decreased in the caudate in advanced PD but otherwise is usually normal. Striatal D_1 receptors have been examined and found to be unchanged. Studies of the dopamine transporter also show a more pronounced reduction of striatal binding in the putamen than in the caudate which correlates with disease severity. Presynaptic monaminergic vesicular transporters are decreased in the putamen more than the caudate in PD patients compared to elderly normal subjects.

Abnormalities in rCBF response to motor tasks have been reported in PD patients. Some of these abnormalities can be normalized after dopaminergic medication, pallidotomy, or subthalamic nucleus stimulation. A generalized decrease in rCBF has also been reported without a concomitant decrease in $CMRO_2$, which has been postulated to be due to vasoconstriction secondary to loss of dopaminergic innervation of the cerebral vessels.

In PD patients, although inconsistent results have been reported for regional glucose metabolism, most studies have shown a generalized decrease involving the cortical and subcortical regions. There is a lack of correlation between changes in glucose metabolism and clinical findings of rigidity or tremor. In patients with PD and dementia, there is a global decrease in glucose metabolism but with more severe hypometabolism in temporoparietal areas.

PET studies have been used to examine the effects of various forms of treatment of PD. For example, following intrastriatal transplantation of human embryonic mesencephalic tissue, striatal ^{18}F-dopa uptake increases and is correlated with graft survival.

MSA accounts for up to 10% of patients with parkinsonism and is poorly responsive to dopaminergic medication. In patients with MSA, presynaptic uptake of ^{18}F-dopa and dopamine reuptake sites are decreased in the putamen and caudate. Striatal D_2 receptor binding is also decreased in patients with MSA, as is vesicular transporter binding. Both ^{18}F-dopa uptake and D_2 receptor binding correlate with motor impairment. Glucose metabolism in patients with MSA is also decreased in the putamen and caudate. In addition, significant reductions in metabolism are present in the premotor and motor cortex, the thalamus, and the cerebellum.

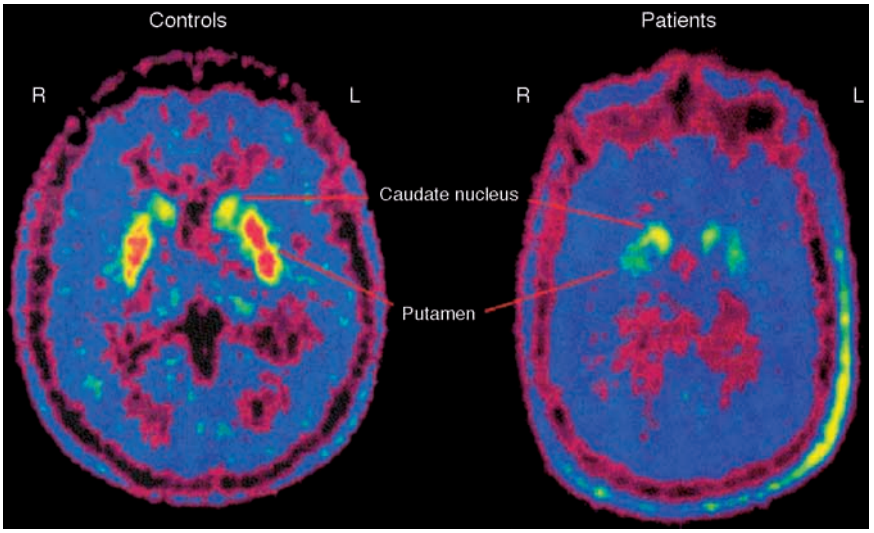

Figure 5 Reduced striatal [^{18}F]-dopa uptake is present in PD compared to normal subjects. The reduced [^{18}F]-dopa uptake is more pronounced in the putamen with relative sparing in the caudate nucleus. Reproduced from Thobois S, Jahanshahi M, Pinto S, Frackowiak R, and Limousin-Dowsey P (2004) PET and SPECT functional imaging studies in Parkinsonian syndromes: From the lesion to its consequences. *Neuroimage* 23: 1–16, with permission from Elsevier.

In patients with PSP, ^{18}F-dopa uptake is significantly reduced, involving both the caudate and the putamen. However, there is no correlation between levels of ^{18}F-dopa uptake and motor disability, unlike in PD and MSA. D_2 receptor binding sites are also decreased in the caudate and putamen, as are striatal dopamine transporters. A pattern of alterations in glucose metabolism is seen in these patients, which is similar to that seen in MSA but with less pronounced frontal lobe and striatal hypometabolism.

PD patients can be distinguished from MSA and PSP patients with PET studies of the dopamine system. In untreated PD patients, presynaptic function (F-dopa uptake and dopamine transporter and vesicular transporter binding) is impaired, whereas postsynaptic function (D_2 receptor binding) is normal or increased; in MSA and PSP, both pre- and postsynaptic functions are impaired. A decrease in striatal glucose metabolism also clearly distinguishes PD from MSA.

In Huntington's disease (HD), striatal ^{18}F-dopa levels are normal but D_2 receptor binding in the caudate is decreased. This decrease is correlated with the duration of illness. Subjects at risk for HD have normal D_2 receptor binding. Glucose metabolism in HD patients is decreased in the caudate nucleus even when the CT scan is normal and can precede the onset of symptoms in people at risk for HD. There is also evidence that glucose metabolism is reduced in the cerebral cortex in HD patients and that the severity and duration of symptoms correlate with the decrease in metabolism in cortical and subcortical regions.

Psychiatric Disorders

Depressive Disorders

Patients with manic-depressive disorders have been found to have abnormalities in glucose metabolism in the left anterolateral prefrontal cortex, inferior temporal cortex, and basal ganglia. The severity of depression correlates with the decrease in left frontal metabolism. In addition, in bipolar depressive patients compared to unipolar patients, a decrease in glucose metabolism has been observed in the frontal cortex (hypofrontality). In patients with bipolar depression, global glucose metabolism is decreased in the depressive phase, whereas in unipolar patients global metabolism is within normal limits. In addition, bipolar patients in the hypomanic phase tend to have normal glucose metabolism. Thus, FDG–PET studies might be useful in distinguishing unipolar from bipolar depression, which would be important in regard to the patient's treatment and prognosis. An increase in glucose metabolism with treatment of depression has been observed in bipolar and 'mixed' manic-depressive patients.

Studies of serotonin receptor binding in the cortex using ^{11}C-methylspiperone have shown an inverse relationship between the severity of depression in stroke patients and spiperone binding in the left temporal cortex. Serotonin transporter binding measured with PET and ^{11}C-McN 5652 was found to be reduced in the amygdala, midbrain, and ventral striatum in depressed patients in one study, whereas it has been reported to be increased in the thalamus and right cingulate and left frontal cortices in other studies. Blood flow has also been reported to be decreased in cortical regions, more so on the left, in depressed patients.

Schizophrenia

In untreated schizophrenic patients during resting conditions, decreases in glucose metabolism have been observed in frontal and temporal cortex and in central brain regions. The presence of frontal hypometabolism does not appear to be specific for this disorder because patients with affective disorders also demonstrate this finding.

Presynaptic dopamine metabolism has been studied in schizophrenic patients with ^{18}F-dopa and found to be abnormal. In one study, a significantly increased dopa uptake was found in the posterior cingulate and a significantly decreased uptake in the ventral striatum compared to normal control subjects. In other studies, dopa uptake has been reported to be increased in the striatum as well as in the prefrontal cortex. The vesicular monoamine transporter has been studied with ^{11}C-dihydrotetrabedazine and found to be normal in the striatum.

Quantitative studies of D_2 receptors in schizophrenic patients have revealed contradictory results. Differences in the ligands used (^{11}C-*N*-methylspiperone and ^{11}C-raclopride, respectively), the experimental strategy, and the patient populations as well as differences in the dopamine receptor subtype affinities of the two ligands might explain these discrepant results. Two meta-analyses have shown an increase in D_2 density in the striatum in schizophrenic patients. Further work is needed to resolve this important question. A few studies have examined striatal D_1 receptors in schizophrenic patients and found them to be unchanged from those in normal subjects. Studies of D_1 receptors in the prefrontal cortex in schizophrenic patients have produced contradictory results.

Schizophrenic patients treated with antipsychotic drugs have a significant fraction of the D_2 receptors in the striatum occupied by the drug. Some of the

antipsychotic drugs also occupy the D_1 receptors, as measured with ^{11}C-SCH 23390.

Postmortem studies have shown decreased serotonin 5-HT_{2A} receptors, increased 5-HT_{1A} receptors, and decreased serotonin transporter in the prefrontal cortex of schizophrenic patients. PET studies of serotonin receptor changes in schizophrenic patients have been performed using ^{18}F-setoperone and ^{11}C-*N*-methylspiperone and have found no significant differences from normal control subjects. However, one study using ^{18}F-setoperone did reveal a significantly lower 5-HT_{2A} binding potential in the frontal cortex in neuroleptic-naive schizophrenic patients. No PET studies have examined 5-HT_{1A} receptors or the serotonin transporter in schizophrenia.

PET studies are also being used to help determine the appropriate dose of drugs as well as in the development of new drugs to treat psychiatric illnesses.

Aging and Dementia

rCBF and oxygen metabolism measured with PET methods have demonstrated a significant decline with aging in these parameters in gray matter with relatively little change in white matter. Regional cerebral glucose metabolism has also been reported to decrease with aging, with the frontal cortex showing a greater decrease than other cortical regions. Other investigators, however, have failed to detect any statistically significant difference in regional cerebral glucose consumption between normal young subjects and normal elderly subjects.

Studies of the dopaminergic system have shown an age related decline in D_1 and D_2 receptors in the striatum, frontal cortex, temporal cortex, and thalamus of the order of 10% per decade (**Figure 6**). Studies of monoamine oxidase B have demonstrated a significant increase with age in many brain regions, including the frontal, parietal, and temporal cortices, the thalamus, basal ganglia, pons, and cerebellum. Serotonin (5-HT_{2A}) receptor binding has also been reported to decrease with aging (**Figure 7**).

In patients with Alzheimer's disease (AD), studies have shown a decrease in glucose and oxygen consumption as well as a decrease in rCBF compared to normal elderly subjects. These studies also showed a significant correlation between the degree of cognitive impairment and the reduced metabolic rate as well as a good correlation between the patient's cognitive deficits and the location of the metabolic abnormalities. The typical pattern of decreased metabolism that is seen most prominently affects the parietotemporal association and posterior cingulate cortices and later the parietal cortex with extension to the frontal cortex in advanced disease, with relative sparing of the primary sensorimotor cortices, thalamus, basal ganglia, and cerebellum. A multicenter study showed that decreased glucose metabolism in the parietotemporal, prefrontal, and posterior cingulate cortices is diagnostic of mild to moderate AD with 93% sensitivity and specificity, and even in very mild AD sensitivity was 84% and specificity was 93%. In addition, the pattern of metabolic abnormalities enables one to distinguish AD from

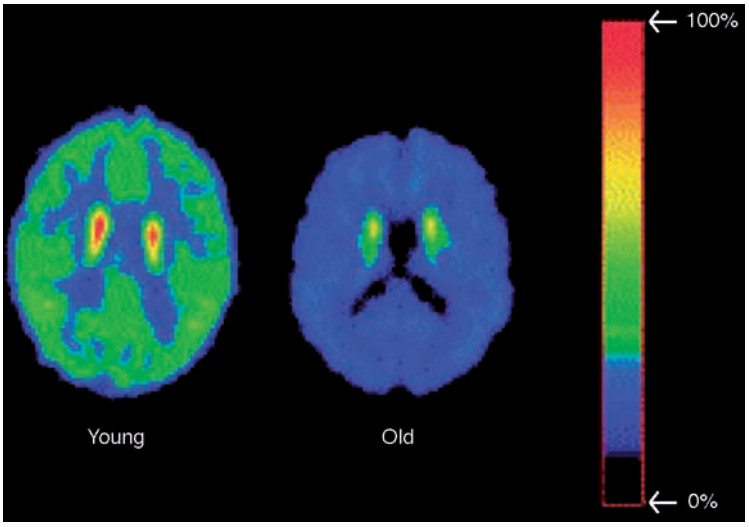

Figure 6 Loss of dopaminergic activity in aging is illustrated in the study of [^{11}C]FLB 457, a dopamine D_2 receptor ligand, in a young (age 25 years) and an old (age 70 years) healthy woman. The images are scaled to an equal radioactivity in the reference region (the cerebellum). Reproduced from Kaasinen V and Rinne JO (2002) Functional imaging studies of dopamine system and cognition in normal aging and Parkinson's disease. *Neuroscience and Biobehavioral Reviews* 26: 785–793, with permission from Elsevier.

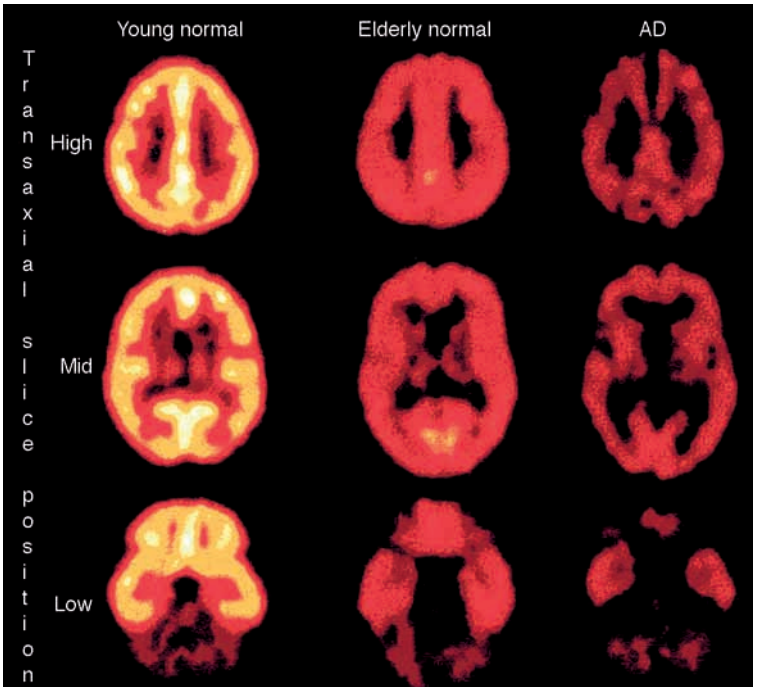

Figure 7 [^{18}F]altanserin PET images in young normal (age 20 years), elderly normal (age 66 years), and Alzheimer's disease (AD; age 67 years) subjects. Images are normalized to the cerebellum, an area with few 5-HT_{2A} receptors. Cortical 5-HT_{2A} binding is greater in the young subject compared to the elderly subject. The Alzheimer's patient exhibits diffusely lower 5-HT_{2A} binding than both the young and elderly controls. Reproduced from Meltzer CC, Smith G, DeKosky ST, et al. (1998). Serotonin in aging, late-life depression, and Alzheimer's disease: The emerging role of functional imaging. *Neuropsychopharmacology* 18: 407–430, with permission.

other forms of dementia. Frontotemporal dementia (Pick's disease) shows a more severe frontal and temporal hypometabolism, dementia with Lewy bodies demonstrates more severe occipital hypometabolism in addition to the parietotemporal hypometabolism, whereas vascular dementia is characterized by diffuse hypometabolism in cortical, subcortical, and cerebellar regions. Depression can lead to memory impairment without dementia but is associated with a more global reduction in glucose metabolism with accentuation in the frontal lobe.

Studies of protein synthesis using L-[^{11}C]methionine have shown that in patients with AD protein synthesis is also decreased compared to age-matched normal subjects and is proportional to the severity of the dementia. Detectable impairment of protein synthesis prior to any clinical signs suggests that this may be a sensitive probe for the evaluation of neuronal metabolic integrity at a very early stage of the disease.

PET neuroreceptor studies using ^{11}C-nicotine in patients with AD have shown a decrease in the low-affinity nicotine receptors in the temporal and frontal cortices and a significant correlation with the Mini-Mental State Examination Score.

Reduced acetylcholine activity has also been observed in the parietotemporal cortex in AD as well as a loss of serotonin receptors in the temporal and frontal cortex (Figure 7).

Epilepsy

Because of its ability to measure functional changes in the brain, PET has great potential in the study of epilepsy, a disorder in which changes in neural discharges usually unassociated with structural alterations are present. In patients with a generalized seizure disorder, CBF and metabolism are normal interictally. Ictally, a global increase in CBF occurs during generalized seizures, with a postictal decrease in CBF. Similar changes occur in cerebral metabolism.

In patients with partial epilepsy, it has been found that in the interictal period glucose metabolism and cerebral perfusion are reduced in the region corresponding to the epileptic focus. Ictal studies have shown an increase in both these parameters in these regions. The area of hypometabolism typically extends well beyond the area of pathologic abnormality to include the entire temporal lobe. Studies with ^{11}C-flumazenil, which binds to the γ-aminobutyric acid

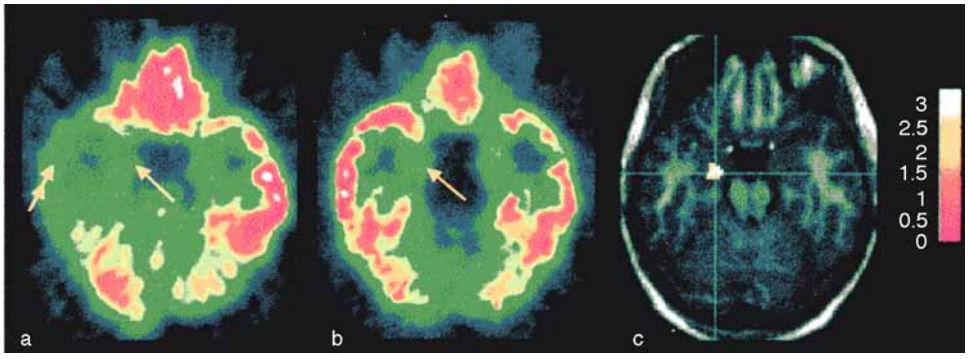

Figure 8 Comparison of ^{18}FDG PET (a) and ^{11}C-FMZ PET (b) in a patient interictally with complex partial seizures. (a) ^{18}FDG PET shows extensive hypometabolism in the left temporal lobe, affecting both the medial temporal cortex (arrow) and the inferior and lateral (double arrow) temporal neocortex. (b) ^{11}C-FMZ PET shows a circumscribed decrease of ^{11}C-FMZ volume of distribution (Vd) in the left medial temporal lobe (arrow). (c) Statistical analysis with statistical parametric mapping (SPM99) shows a significant decrease of FMZ-Vd in the left anterior hippocampus compared with 21 controls. The statistical map is overlaid on the patient's MRI; the color scale shows the *t* value. Reproduced from Tai YF and Piccini P (2004) Applications of positron emission tomography (PET) in neurology. *Journal of Neurology, Neurosurgery, & Psychiatry* 75: 669–676, with permission.

central benzodiazepine receptor complex, in these patients often show a smaller region of abnormality (Figure 8). Occasionally, the FMZ–PET study will be abnormal, whereas the FDG–PET study or MRI is normal in these patients. FMZ–PET is more sensitive and accurate in delineating the seizure focus than FDG–PET. Interictally, decreased glucose metabolism and FMZ binding have also been observed. The presence of thalamic involvement has a strong lateralizing value, particularly when cortical abnormalities are subtle. FDG–PET should be considered as the first-line PET investigation in patients with temporal lobe epilepsy with a strictly normal MRI, and FMZ–PET could be proposed in the subgroup of patients whose FDG–PET failed to reveal focal hypometabolism.

In patients with extratemporal lobe epilepsy, interictal FDG–PET, in general, has proven to be a reliable test for identifying dysfunctional cortical regions of hypometabolism that correspond to the location of epileptic foci. FDG–PET in cases of frontal lobe epilepsy has show a sensitivity and accuracy of 92–96% and 74–78%, respectively. A comparison of FDG– and FMZ–PET studies in patients with frontal lobe epilepsy showed a good correlation between location of FMZ–PET abnormalities and epileptic focus identified by surface and intracranial electroencephalogram (EEG) with subdural electrodes. Furthermore, FMZ–PET abnormalities were more restricted that those delineated with FDG–PET.

In patients with generalized forms of epilepsy, interictal PET studies usually do not demonstrate lateralized or localized cortical abnormalities. During frequent absence seizures, however, marked generalized hypermetabolism as well as a global increase in CBF may be present.

Several other PET tracers have been examined for their potential usefulness in detecting epileptic foci – for example, histamine H1 receptor binding, opioid receptor binding, *N*-methyl-D-aspartate receptor activity, and serotonin synthesis. Focal increases of these tracers have been observed in various forms of epilepsy; however, the results are preliminary.

Gene Therapy

Several approaches to gene therapy are being tested using a variety of vectors and therapeutic genes. All rely on the same rationale: genetically engineering a target tissue to produce a new therapeutic protein. This requires the selective introduction of genes into a target tissue using a vector. Such vectors include adenovirus, retrovirus, DNA–protein complexes, and liposomes. The vector has several roles: it delivers the gene across the cell membrane into the nucleus and drives expression of the gene to produce the therapeutic protein.

One major approach in cancer gene therapy that has been extensively investigated is the use of enzyme/prodrug systems, such as herpes simplex virus thymidine kinase/ganciclovir (HSVtk/GCV), which introduce a drug-activating enzyme. Phosphorylation by the viral kinase triggers the therapeutic effect of GCV. The viral kinase activates the drug to a monophosphate. Subsequent phosphorylation of the monophosphate by cellular kinases leads to inhibition of DNA synthesis and cancer cell death.

PET studies in mice using a PET scanner designed for small animal studies (2 or 3 mm in plane resolution) have demonstrated the feasibility of HSV1-tk as a reporter gene and ^{18}F-GCV as an imaging

reporter probe. The liver, which is the primary site of adenovirus uptake after intravenous administration, was visualized only in the mice receiving Ad.HSVtk compared to those receiving a control virus. There was a significant correlation between the HSVtk enzyme activity in the liver and the ^{18}F-GCV activity in the liver. Since it has been shown that there is a strong correlation between ^{18}F-GCV uptake *in vitro* and the growth inhibition in response to pharmacological doses of GCV, the uptake of ^{18}F-GCV may be not only a marker for gene transfer and expression of HSVtk in tumors but also a predictor of response to therapy.

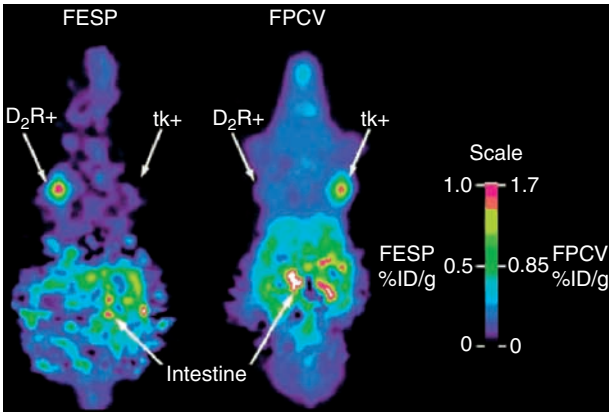

Figure 9 Imaging D_2R and HSV1-tk PET reporter genes in the same animal. A mouse carrying C6-stb-tk+ tumor (right) and D_2R (left) was imaged by micro-PET for HSV1-tk reporter gene expression using FPCV and D_2R reporter gene expression using FESP. Highly specific localization of FPCV and FESP was observed in tumors expressing HSV1-tk and D_2R genes, respectively. Reproduced from Iyer M, Barrio JR, Namavari M, et al. (2001) 8-[^{18}F]Fluropenciclovir: an improved reporter probe for imaging HSV1-tk reporter gene expression *in vivo* using PET. *Journal of Nuclear Medicine* 42: 96–105, with permission.

Fluoro-penciclovir (^{18}F-FPCV) has also been shown to be a useful imaging reporter probe. *In vitro* studies have shown that accumulation of ^{18}F-FPCV is greater compared to ^{18}F-GCV in C6 cells expressing HSVtk. The use of FPCV as a reporter probe should allow a greater imaging sensitivity.

Another tracer that has been studied is fluoro-hydroxy-propoxy methyl guanine (^{18}F-FHPG). It has been evaluated in *in vitro* and *in vivo* studies which suggest that ^{18}F-FHPG is a potential *in vivo* PET imaging agent for monitoring gene incorporation and expression. Comparison of FPCV and ^{18}F-fluro-hydroxymethyl-butyl guanine (FHBG) as reporter probes in studies of transgenic mice expressing HSV1-tk in the liver showed that there is significantly greater hepatic accumulation when using FHBG. A mutated form of HSV1-tk, HSV1sr-39tk, which phosphorylates GCV and penciclovir more efficiently than thymidine, has been shown to have improved sensitivity for imaging reporter gene expression using a probe which is a derivative of penciclovir, FHBG.

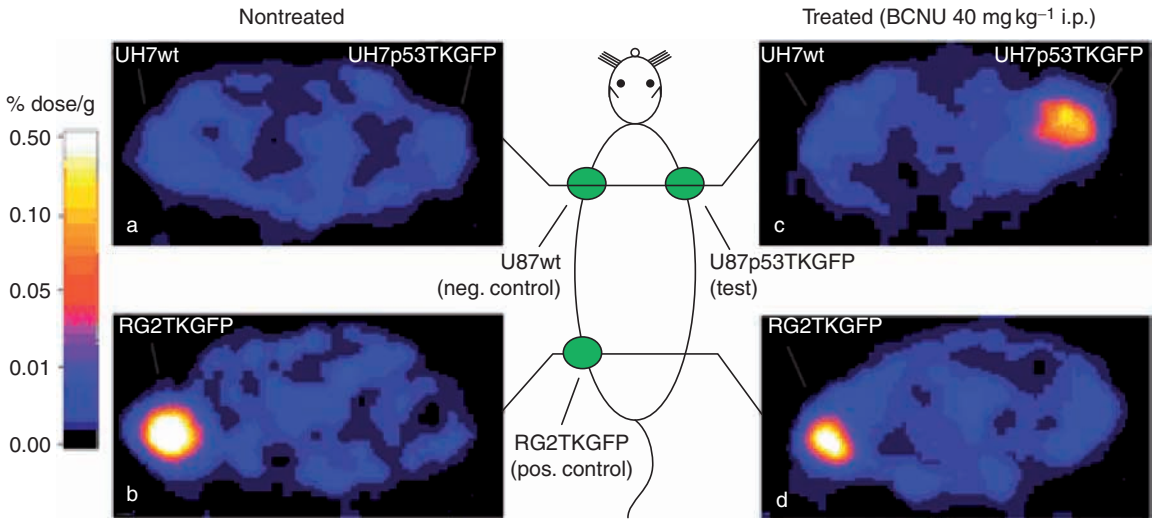

Figure 10 PET imaging of endogenous p53 activation. Transaxial PET images through the shoulder (a and c) and pelvis (b and d) of two rats are shown. A nontreated animal is shown on the left and a BCNU-treated animal on the right. Both animals had three SC tumor xenografts: U87p53TKGFP (test) in the right shoulder, U87 wild-type (negative control) in the left shoulder, and RG2TKGFP (positive control) in the left thigh. The nontreated animal on the left shows localization of radioactivity only in the positive control tumor (RG2TKGFP); the test (U87p53TKGFP) and negative control (U87 wild-type) tumors are at background levels. The BCNU-treated animal on the right shows significant radioactivity localization in the test tumor (right shoulder) and in the positive control (left thigh) but no radioactivity above background in the negative control (left shoulder). Reproduced from Doubrovin M, Ponomarev V, Beresten T, et al. (2001) Imaging transcriptional regulation of p53-dependent genes with positron emission tomography *in vivo*. *Proceeding of the National Academy of Sciences of the United States of America* 98: 9300–9305, with permission.

Another reporter gene technique is based on the expression of a transmembrane receptor as the reporter gene and its labeled receptor ligand as the marker. One application of this method uses the dopamine D_2 receptor and ^{18}F-fluoroethyl-spiperone (^{18}F-FESP) as a ligand which, unlike most HSV1-tk substrates, rapidly crosses the blood–brain barrier. The vectors used in these studies were either an adenovirus carrying the D_2R reporter gene (Ad-D_2R) or a control adenovirus expressing β-galactosidase (Ad-βGal). In studies in mice administered either Ad-D_2R or Ad-β Gal intravenously, given ^{18}F-FESP 2 days later, and then imaged, only those animals given Ad-D_2R showed significant activity in the liver. The ^{18}F activity in the liver was proportional to the levels of hepatic D_2R-dependent ^{3}H-spiperone binding. Studies in which the D_2R antagonist butaclamol was administered 2 or 3 days after Ad-D_2R administration further demonstrated the specificity of the ^{18}F-FESP reporter probe for imaging the D_2R reporter gene. An example of the use of the HSV1-tk and D_2R reporter genes is shown in Figure 9. A mutant D_2R that uncouples signal transduction while maintaining affinity for FESP has been reported. Other receptor-based reporter gene systems involve the somatostatin receptor subtype II (SSTr2) and the sodium iodide symporter (NIS).

The monitoring of patients undergoing gene therapy may be improved with the ability to determine repetitively and noninvasively where the transferred DNA is being expressed and to what extent expression occurs over time.

In addition to use in gene therapy, reporter gene imaging can be used to visualize endogenous biological processes such as transcriptional regulation of gene expression as well as specific intracellular protein interactions. For example, endogenous expression of p53 has been imaged *in vivo* using a retroviral vector bearing the *cis*-p53/TKGFP reporter system in which a reporter gene for herpes simplex virus type 1 thymidine kinase was placed under the control of an artificial *cis*-acting p53-specific enhancer and the PET reporter probe fluoro-deoxy-arabinofuranosyl-[^{124}I]iodouracil. In this study, DNA damage produced by *N*,*N*′-bis (2-chloroethyl)-*N*-nitrosourea (BCNU) administration caused upregulation of p53 transcriptional activity which could be imaged with this reporter system (Figure 10).

See also: Cognition: An Overview of Neuroimaging Techniques; Cognitive Neuroscience: An Overview; Electroencephalography (EEG); Executive Function and Higher-Order Cognition: Neuroimaging; fMRI: BOLD Contrast; Magnetic Resonance Spectroscopy; Neuroimaging; Perfusion MRI; Sleep and Sleep States: PET Activation Patterns.

Further Reading

Derdeyn CP (2005) Positron emission tomography imaging of cerebral ischemia. *Neuroimaging Clinics of North America* 15: 341–350.

Doubrovin M, Ponomarev V, Beresten T, et al. (2001) Imaging transcriptional regulation of p53-dependent gens with positron emission tomography *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* 98: 9300–9305.

Fletcher PC and Henson RNA (2001) Frontal lobes and human memory. Insights from functional neuroimaging. *Brain* 124: 849–881.

Heiss W-D (2000) Ischemic penumbra: Evidence from functional imaging in man. *Journal of Cerebral Blood Flow and Metabolism* 20: 1276–1293.

Herholz K (2003) PET studies in dementia. *Annals of Nuclear Medicine* 17: 79–89.

Iyer M, Barrio JR, Namavari M, et al. (2001) 8-[^{18}F] Fluropenciclovir: An improved reporter probe for imaging HSV1-tk reporter gene expression *in vivo* using PET. *Journal of Nuclear Medicine* 42: 96–105.

Jacobs AH, Kracht LW, Gossmann A, et al. (2005) Imaging in neurooncology. *NeuroRx* 2: 333–347.

Juhasz C, Chugani DC, Muzik O, et al. (2000) Relationship between EEG and positron emission tomography abnormalities in clinical epilepsy. *Journal of Clinical Neurophysiology* 17: 29–42.

Kaasinen V and Rinne JO (2002) Functional imaging studies of dopamine system and cognition in normal aging and Parkinson's disease. *Neuroscience and Biobehavioral Reviews* 26: 785–793.

Massoud TF (2003) Molecular imaging in living subjects: Seeing fundamental biological processes in a new light. *Genes & Development* 17: 545–580.

Meltzer CC, Smith G, DeKosky ST, et al. (1998) Serotonin in aging, late-life depression, and Alzheimer's disease: The emerging role of functional imaging. *Neuropsychopharmacology* 18: 407–430.

Mosconi L (2005) Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. *European Journal of Nuclear Medicine and Molecular Imaging* 32: 486–510.

Newberg AB, Alavi A, and Reivich M (2002) Determination of regional cerebral function with FDG PET imaging in neuropsychiatric disorders. *Seminars in Nuclear Medicine* 32: 13–34.

Parsey RV and Mann JJ (2003) Applications of positron emission tomography in psychiatry. *Seminars in Nuclear Medicine* 33: 129–135.

Rijntjes M and Weiller C (2002) Recovery of motor and language abilities after stroke: The contribution of functional imaging. *Progress in Neurobiology* 66: 109–122.

Schaller B (2004) Usefulness of positron emission tomography in diagnosis and treatment follow-up of brain tumors. *Neurobiology of Disease* 15: 437–448.

Tai YF and Piccini P (2004) Applications of positron emission tomography (PET) in neurology. *Journal of Neurology, Neurosurgery, & Psychiatry* 75: 669–676.

Thobois S, Jahanshahi M, Pinto S, Frackowiak R, and Limousin-Dowsey P (2004) PET and SPECT functional imaging studies in Parkinsonian syndromes: From the lesion to its consequences. *Neuroimage* 23: 1–16.

Wager TD and Smith EE (2003) Neuroimaging studies of working memory: A meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience* 3: 255–274.

Posterior Parietal Cortex and Arm Movement

A Battaglia-Mayer and R Caminiti, Università di Roma, Rome, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Cortical control of arm movement to visual targets (reaching) is the result of the cooperative interplay of a constellation of parietal and frontal areas and of their reciprocal corticocortical connections (Figure 1). This network is considered the main substrate for a recurrent mechanism underlying planning, execution, and control of arm movement. The posterior parietal cortex (PPC) plays a crucial role in this process. Parietal operations seem necessary for the visual-to-motor transformation underlying reaching and for the control of the congruence between motor plans and actual movement, thanks to efferent copies of motor commands from frontal cortex and reafferent signals from the sensory periphery. PPC is also crucial for the online control and adjustment of reaching, whenever requested by task demands. The premotor and motor areas of the frontal lobe, for their direct access to the spinal cord, are essential to address motor commands to the motor periphery and, through efferent cortical connections, corollary signals to other brain regions, including PPC. These frontal areas can participate in the birth of visually guided arm movements by elaborating signals about the state of the motor system to predict the sensory consequences of a motor command, thus creating the basis for an internal model of the desired movement.

These concepts stem from the cross-fertilization of results coming from different neuroscience disciplines, such as psychophysics, behavioral neurophysiology, neuroanatomy, and neuropsychology. Psychophysical studies in humans have set the background for interpreting novel neurophysiological results, mostly addressed at the elucidation of the dynamic properties of neurons in different areas of the parietofrontal system, studied in alert behaving monkeys. Neuroanatomical investigations have provided the necessary substrate for understanding the input–output relations in this cooperative distributed system, while the neuropsychological analysis of patients with brain lesions has been essential to determine whether, and to what extent, the integrity of parietal operations is essential for different aspects of reaching.

The Essential Features of Reaching in Psychophysical Analysis

Reaching consists of a coordinated eye–hand movement toward an object of interest. If target eccentricity is around 15° of visual angle, eye movement is accompanied by head movement. The onsets of activation for eye, head, and arm muscles largely coincide in time. However, because the inertias of the eye, head, and arm differ, the eye moves first, followed by the head, and then by the hand. Therefore, the initial motor command addressed to arm muscles is based on extrafoveal signals. However, at the onset of arm movement, the eye has normally landed on the target, and foveal information can be used for updating the motor command and later on for the visual control of the final hand shaping. Preventing eye or head movements results in a severe inaccuracy of reaching. In addition, eye position signals, from proprioception or efferent copies of eye movement, contribute essential information since their manipulation results in typical reaching errors.

The movement of the arm is a consequence of motion at the shoulder and elbow joints. These motions result from the torques exerted on the joints by the combined contraction of a large number of muscles acting on them, and they depend on movement trajectory and velocity, gravity, and inertial and Coriolis forces. This phase of reaching, characterized by a large amplitude movement that brings the hand close to the target, is followed by corrective adjustments of smaller amplitude, mostly at the wrist and finger joints, that depend on the orientation, geometry, and physical properties of the target. When reaching is made in different direction of space, a significant correlation exists between the activities of the different muscles set in motion, suggesting that they act in a synergic fashion, that is, through a coordinate activation with defined time sequences. The time-varying profiles of a limited number of muscle synergies well explain the overall pattern of muscle activity during fast reaching, regardless of different loads imposed or arm posture.

During arm movements, the hand occupies different positions in space at different times, and this defines the hand trajectory. The trajectory of free arm movements in three-dimensional (3-D) space is slightly curvilinear, and its shape is independent of movement speed, arm orientation, or external loads.

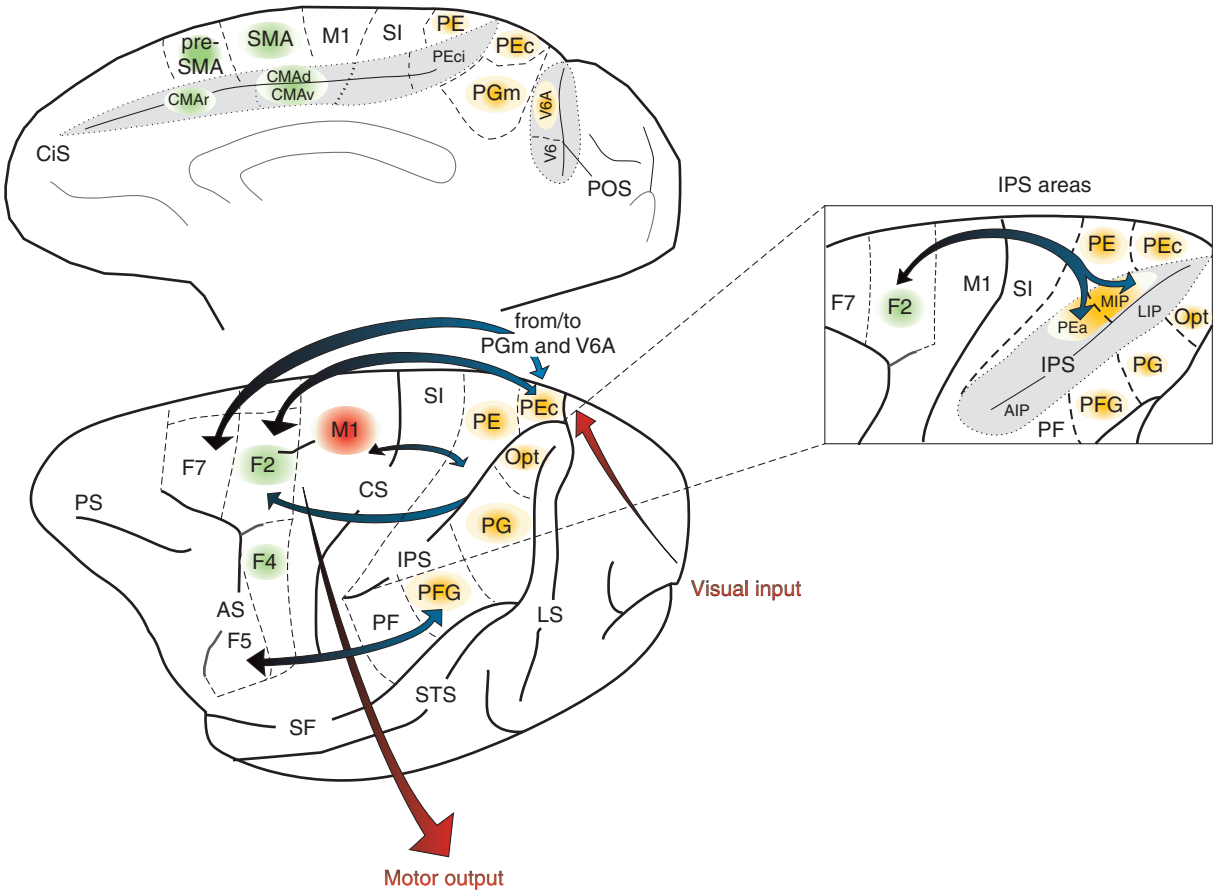

Figure 1 The parietofrontal network underlying arm movement. Brain figures of the medial (top) and lateral (bottom) views of the macaque monkey's brain, showing the regions of parietal (yellow), premotor frontal and cingulate (green), and motor (red) cortices where cell activity is related to arm movement. The inset of the brain figure on the right is an unfolding of the intraparietal sulcus (IPS), necessary to display the areas (gray shading) forming the medial and lateral banks of the sulcus. Similar unfolding is shown for the cortex of the parieto-occipital (POS) and cingulate (CiS) sulci. The blue arrows indicate the main reciprocal association connections that define the parietofrontal network. Thin broken lines indicate the architectonic borders between cortical areas. In the medial view, POS and CiS indicate parieto-occipital and cingulate sulcus, respectively. In the lateral view, LS, STS, SF, IPS, CS, PS, and AS indicate lunate, superior temporal, Sylvian, intraparietal, central, principal and arcuate sulci. Areas V6A, PGm, PEc, MIP (medial intraparietal), PEa, and PE are part of the superior parietal lobule. Areas Opt, PG, PFG, PF, LIP (lateral intraparietal), and AIP (anterior intraparietal) are part of the inferior parietal lobule (IPL). In the frontal lobe, areas F2, F4, and F5 on the lateral aspect of the hemisphere and the supplementary motor area (SMA) on the medial wall are the frontal premotor areas that project to primary motor cortex (M1). In the CiS, the dorsal (CMAd) and ventral (CMAv) cingulate motor areas are two premotor fields that project to M1. CMAr indicates the rostral cingulate motor area. S1, primary somatosensory cortex; V6, visual area VI; V6A, visual area VIA.

Therefore, it is believed that the trajectory is a movement invariant encoded by the central nervous system, in front of the exuberance of degrees of freedom that characterize movement.

Reaching movements display a typical bell-shaped velocity profile, with an accelerating and a decelerating phase. Therefore, during movement, hand velocity never remains constant. The particular shape of this otherwise invariant profile will depend on the trade-offs imposed by the different contexts in which movement is performed and by the variable that in each context has to be optimized, such as (minimizing) jerk, (maximizing) stiffness, (minimizing) energy expenditure, and so forth.

The accuracy of reaching movements is maximal when the hand starting position is located in front and around the reacher's midline, when movement is of small amplitude, and when it is directed straight ahead rather than in other directions. The faster the movement, the less it is accurate, a principle known as speed–accuracy trade-off. Changing speed also affects joint movement dynamics and requires a rescaling of the joint torques.

Visual Information Is Essential to Reaching

Arm movements are typically guided by vision, from target localization in space to visual control of movement trajectory and final shaping of the hand necessary

to match the physical properties of the target object. The visual direction of a target is mapped in retinotopic coordinates, whereas encoding its distance requires the combination of retinal disparity and fixation distance information, such as eye vergence angle, accommodation, and vertical disparity. The angle of vergence is the most important one, particularly within the near space, where the vergence range is roughly equal to the arm length. Therefore, this information might be used to specify the distance of arm movement as well. This is also suggested by ontogenetic considerations since manual reaching in infants develops at the same time as binocular control. Reaching is more accurate when vision of the arm is allowed than when it is prevented. However, this seems to depend, not only on visual feedback from the moving arm, but also from visual motion cues arising from active arm movement. These mechanisms seem to rely on peripheral vision, while the final preshaping of the hand near the target occurs under the control of central vision.

Coordinate Transformation Underlying Reaching

The vector of desired hand movement is defined by the difference (motor error) between the target and the hand location. Although hand position can be encoded in retinal coordinates, outside the visual field it is defined by proprioception and by efferent copies of the motor commands. In the process of combining information about target and hand position, the endpoint of arm movement has to be transformed from retinal to allocentric or egocentric coordinates, probably through a cascade of transformations. It is believed that these transformations carry behavioral errors of different types: (1) constant errors, measuring the systematic deviation of the mean endpoint of movement from the target; (2) variable errors, reflecting the variance of the individual movement endpoints around the mean; and (3) local distortions, indicating the degree of similitude between the configuration of the targets and that of the movement endpoints. In recent years, psychophysical studies have devoted significant attention to the analysis of reaching errors since they may reveal the internal representations and coordinate transformations used to specify the endpoint of reaching. An emerging view is that the frame of reference used to specify arm movement is task- and context-dependent. A few examples follow.

During arm movements toward visible virtual targets or to memorized locations, the specification of the final hand position occurs in terms of direction and distance from the eyes. This implies that a combination between eye and hand information enters into a common egocentric binocular frame of reference. When vision of the hand is not allowed, the axes of maximum variability are rotated around the body. When a discrepancy between the actual and the visually perceived hand position is introduced at one location in space, participants learn a visuomotor remapping over the entire workspace within a spherical coordinate system that specifies the azimuth, elevation, and distance of the hand relative to the eyes. These data support the hypothesis of a multistage transformation, from viewer-centered to head- and shoulder-centered coordinates. Furthermore, the existence of hand-centered commands for movement based on combination of target and hand signals is revealed by the additional variability observed along movement direction. When targets are located in a structured space, the endpoint position of reaching can be specified in object-centered or allocentric coordinates. Cognitive processing, such as prior information on the spatial relationships between targets that will be presented in sequence along a virtual line, also affects the representation of endpoints of reaching. In such instances, the pattern of errors reveals coexisting egocentric and allocentric representations. In conclusion, the frame of reference used to specify the endpoint for reaching is not fixed but depends on the available sensory information, task constraints, structure of visual scene, and cognitive context.

Eye–Hand Coordination

When human participants are asked to make pointing movements to visual targets presented in a sequential fashion, they cannot initiate a saccade to any given target until the hand has reached the previous one. This functional coupling between eye and hand movements is internal and does not depend on vision, since it also occurs when vision of the moving arm is prevented. In addition, saccades are faster when they occur within a coordinated eye–arm movement. When considering continuous movements, changing the gain of eye pursuit results in a change in the subsequent hand movements. Thus, in some instances the angle of gaze indicates the target direction for the hand, and in other cases hand movement affects eye movements.

Studies of eye–hand coordination in naturalistic settings confirm the importance of eye information for guiding hand movement. For instance, during tea making, in spite of the constellation of targets in the visual scene, the eyes always move to the object that will be manipulated, leading hand movement by about half a second; in addition, the eyes often move to a new location before the manipulation of the first object is completed so as to control the state of important variables such as the level of water in the teapot, the level of sugar available, and so forth.

Reaching to Moving Objects: The Neural Basis of Interception Movements

A different aspect concerns reaching to nonstationary targets, that is, interception of moving objects. To intercept a moving object, knowledge of where and when its impact with the hand will occur is essential. Temporal estimations for interception can be achieved as hypothesized by the τ model; that is, an interception movement begins when the ratio between the retinal size of an approaching object to its rate of change reaches a critical (τ) threshold. A more comprehensive formulation of this theory generalizes the model to movement control under virtually all conditions and redefines τ as the time to closure of the gap existing between the effector and the target, at the current gap closure rate. It is worth stressing that a common type of interception consists of catching free-falling objects. Studies in humans have suggested that during this task, the brain uses an internal model of gravity that takes into consideration not only the object's velocity, but also its gravitational acceleration. However, interception of a moving target on a video monitor, therefore with no real mass to interact with, occurs assuming a nonaccelerated motion, even when motion is accelerated.

Encoding Arm Movement by the PPC

General Properties of Parietal Reach Neurons

The study of the relationships between neural activity and arm movement in the PPC began in 1975 with the seminal study by Mountcastle and his collaborators. In area 5 of the superior parietal lobule (SPL) and area 7 of the inferior parietal lobule (IPL), they described the properties of a population of neurons that were active with purposeful projection of the arm toward targets of interest located within the immediate peripersonal space. These neurons ceased to fire if targets were moved out of reaching distance or when the animal was making an arm movement in absence of a specific motivation (i.e., reaching for food if it was not hungry, or for water if it was not thirsty). These reaching neurons were far less active or totally silent during passive arm movements imposed on the animal by the experimenter or during passive manipulation of skin, joints, and deep muscle receptors. Thus, the main feature of reaching neurons emerging from these early studies was the conditional and context-dependent nature of their activity, characterized by a nonobligatory relationship with arm movement. Subsequent studies, in line with those originally made in primary motor cortex, have demonstrated that the activity of reach neurons in both SPL and IPL varies in a broad fashion with direction of arm movement and that encoding of movement direction resides in a population code. Similarly, reaching neurons in PPC are influenced by arm position and by the visuospatial signal concerning target location in space. It is interesting that the activity of most parietal reach neurons is also influenced by the position and movement direction of the eye. This modulation by combined eye–hand information can be observed during reaction time and movement time of fast reaching, as well as during the delay interval occurring between the presentation of a visual signal instructing the animal about the direction of the future movement (or the location of a memorized target) and the onset of movement. Reaching-related activity in SPL is influenced by binocular eye vergence signals and is therefore well suited for encoding reaching distance in 3-D space. Therefore, the main feature of parietal reach neurons is their combinatorial nature. An example of this is offered by the different influence and weight of eye and hand signals on neural activity. On this basis, three main classes of combinatorial neurons have been identified in the IPL (Figure 2): hand-dominant, whose neural activity is mostly determined by hand signals; eye-dominant, whose activity is mostly influenced by eye signals; and eye–hand cells, with neural activity equally influenced by eye and hand information.

There exists a fundamental difference between the directional tuning properties of neurons in the SPL and IPL. In the former, the preferred directions of individual cells, obtained from different epochs of

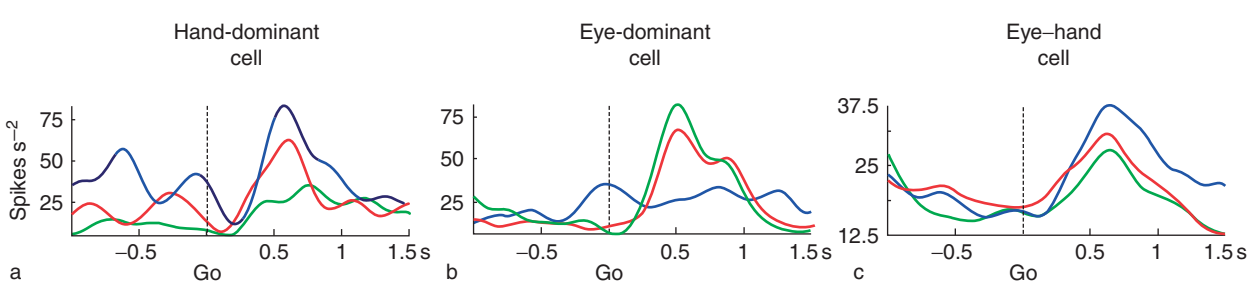

Figure 2 Combinatorial cell types in area 7a. Spike density functions (SDFs) of neural activity for three different cells, each studied during eye movement (green), hand movement (blue), and coordinated eye–hand movement (red) tasks. SDFs are aligned to the go signal (vertical interrupted line). From Battaglia-Mayer A, Mascaro M, and Caminiti R (2007) Temporal evolution and strength of neural activity in parietal cortex during eye and hand movements. *Cerebral Cortex* 17(6): 1350–1363.

different tasks, cluster within a restricted part of space, the global tuning field (Figure 3), an ideal visuomotor frame to dynamically combine eye and hand signal in a directionally congruent fashion, and therefore a possible substrate for eye–hand coordination, in its many different forms. In IPL, instead, only 25% of the cells display a global tuning field, and neuronal activity seems more intensely modulated when the hand moves toward the fixation point, therefore in a more selective form of eye–hand coordination than those apparently subserved by the SPL. At the population level, in the SPL, the distribution of the eye–hand preferred directions computed during different epochs of different tasks covers the directional continuum in a uniform fashion, whereas in the IPL it is highly skewed toward the contralateral space (Figure 4).

Understanding Actions and Motor Intentions by Parietal Neurons

There are action sequences that, although characterized by the same basic movements, have different goals, such as when one reaches out to grasp and eat an apple or to place it in a box. In the monkey IPL, a class of parietal neurons discharges differently

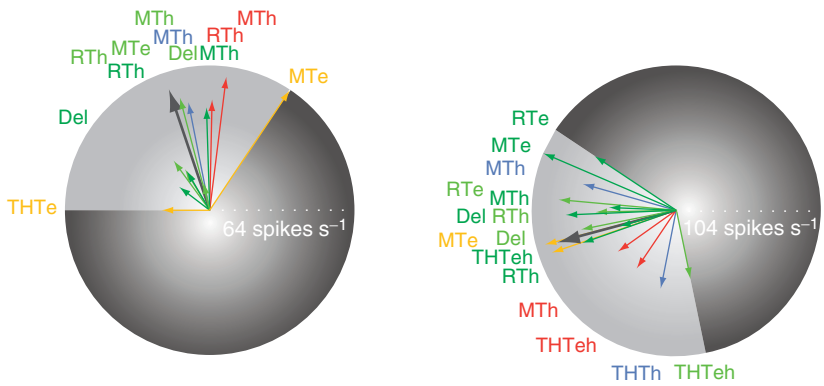

Figure 3 The global tuning field (GTF) of two parietal neurons. When eye and/or hand movements are made in eight different directions under different task conditions, the preferred directions (colored arrows) of individual parietal neurons cluster within a limited part of space, referred to as GTF. In each GTF, the thick dark grey arrow indicates the mean vector (MV), that is, the orientation in space of the neuron's tuning field. The length of the MV is directly proportional to the sharpness of the tuning field. Thus, the GTF of the cell on the left is sharper then that on the right. Yellow refers to a saccadic eye movement task (MTe, eye movement time; THTe, eye target holding time), red to a combined eye–hand movement task (RTh, hand reaction time; MTh, hand movement time; THTeh, combined eye–hand holding on target), blue to a hand movement task requiring central fixation (MTh, hand movement time; THTh, hand holding on target), green to an instructed delay task that separates eye from hand movement in time. This task was performed in light (light green) and dark (dark green) condition (RTe, MTe, eye reaction and movement time, respectively; RTh, MTh, hand reaction and movement time; Del indicates the delay period, when the eye is on the target and the hand is preparing to move; THTeh refers to combined eye–hand holding on the target). Modified from Battaglia-Mayer A, Ferraina S, Mitsuda T, et al. (2000) Early coding of reaching in the parietooccipital cortex. *Journal of Neurophysiology* 83: 2374–2391.

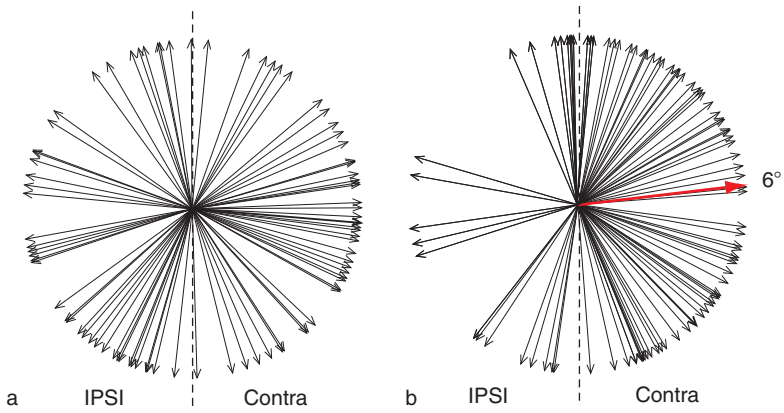

Figure 4 Spatial orientation of the mean vectors of the global tuning field of neurons in (a) the superior parietal lobule (SPL) and in (b) the inferior parietal lobule (IPL). Notice the uniform distribution of the mean vectors in SPL, while in IPL they point mainly toward the contralateral part of space (mean vector of the distribution points to 6°), relative to the hemisphere (left) of cell recording. In (b) the red arrow indicates the mean vector of this distribution. Modified from Battaglia-Mayer A, Mascaro M, Brunamonti E, and Caminiti R (2005) The overrepresentation of contralateral space in parietal cortex: A positive image of directional motor components of neglect? *Cerebral Cortex* 15: 514–525.

depending on the monkey's action goal. These neurons, located in area PFG, also discharge differently when the monkey observes the experimenter performing the same motor sequence with different goals. It has been hypothesized that these parietal mirror neurons encode the agent's motor intention and, therefore, are part of a network involved in motor intention and action understanding. This network involves the area PFG and ventral premotor area F5 (Figure 1).

Parietal Mechanisms for Online Control of Hand Movement Trajectory

Human and nonhuman primates are able to make fast online correction of hand movement trajectory if the visual information that guides movements is changed, such as when the target location is suddenly displaced. In such circumstances, the human or other primate will not reach the first target but will produce a curved trajectory toward the new target location (Figure 5(a)). Movement can also be corrected when the human or other primate is not aware of the shift in target position, that is, during an ocular saccade, or when vision of the arm is occluded, as well as in deafferented patients with loss of proprioception in all four limbs. It has been hypothesized that to perform online corrections, the brain might use a feedback mechanism based on motor outflow, as well as sensory information, such as retinal error.

In humans, transient inactivation of PPC function by transcranial magnetic stimulation (TMS) interferes with the ability to perform fast correction of hand movements when the target location changes in space and prevents learning of a new dynamic if reaches are made in a velocity-dependent force field. Lesion of the PPC leads to optic ataxia (OA), a difficulty of visual control of arm movements. A case report of one OA patient with bilateral PPC lesion documents the patient's marked inability to produce smooth corrections of hand trajectory in response to a sudden change of target location in space (Figure 5(a)).

So far, only a few cell recording studies have been devoted to the activity of cortical cells during online movement corrections. In the primary motor cortex, neural activity is influenced by signals about the change of movement direction to a visual target, consisting of a truncation of the discharge pattern typical of the movement to the first target and of its replacement by a discharge pattern associated with the

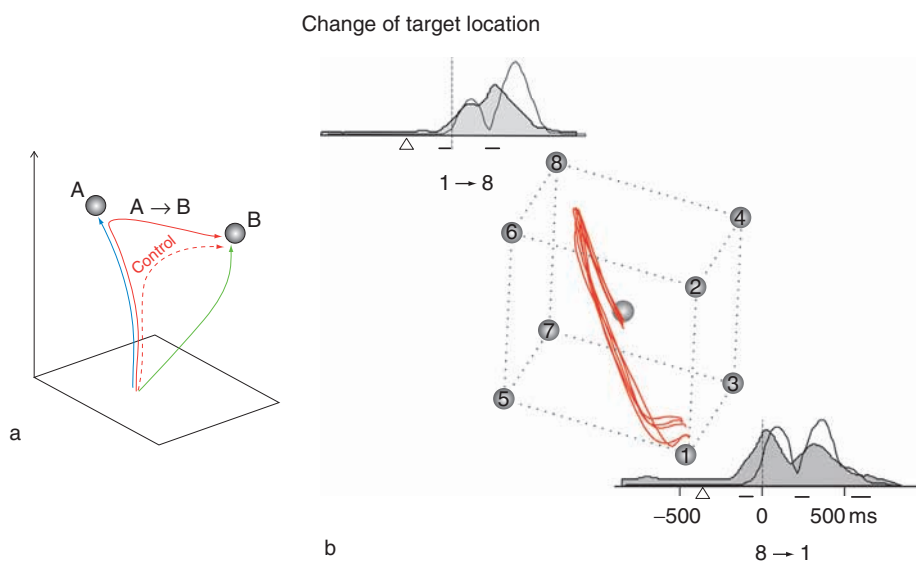

Figure 5 (a) When a visual target jumps from its original position, A, to a new location, B, during arm reaction or movement time, normal participants smoothly correct their hand movement trajectory (red broken line) from A to B, while parietal patients are incapable of this in-flight correction and complete the movement to the first target before directing the hand to second one (red continuous line). Blue and green lines are trajectories of direct movement to A and B, respectively. (b) The neural activity of a parietal neuron during a similar change of hand trajectory due to a jump from target 1 to target 8 and vice versa (gray sphere at the center of the cube is the origin of hand movement). Both the neural activity (shaded gray curves) and the hand velocity signal (solid black curves) are aligned to the beginning of hand movement (vertical dashed line). The triangles indicate the time of target presentation, and the thick horizontal lines correspond to periods of eye movements. Neural activity was modulated with the direction of hand movement and was correlated with hand velocity, leading it by an average of 65 ms. Note the presence of a double peak in the neural activity, preceding the two peaks in the hand velocity profile. Hand paths for target 8 to 1 are depicted inside the cube. This cell was not active in an eye control task (not illustrated). Modified from Battaglia-Mayer A, Archambault PS, and Caminiti R (2006) The cortical network for eye-hand coordination and its relevance to understanding motor disorders of parietal patients. *Neuropsychologia* 44(13): 2607–2620.

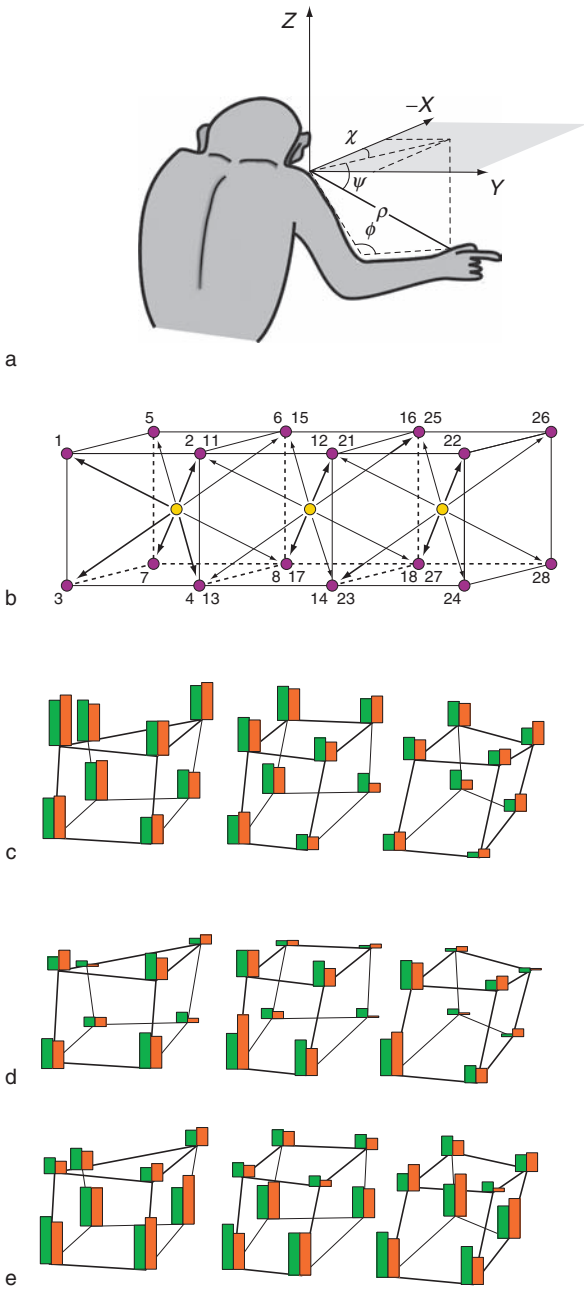

Figure 6 (a) Egocentric coding of target and hand position within a spherical coordinate system by neurons in parietal areas PE and PEa (superior parietal lobule). Azimuth is indicated by the angle χ, elevation by ψ, distance by φ or by the length of segment ρ. (b) Monkeys were trained to point to a visual target (violet) by making movements of constant amplitude from one of three different starting points (yellow) in one of eight possible directions. Targets denoted by pairs of numbers (e.g., 2, 11) could be reached from two different starting points. Each neuron is best tuned to the changes in one spatial coordinate of the target relative to the body: in (c) to the azimuth (neural activity increases monotonically from right to left); in (d) to distance (the closer the hand, the greater the activity); in (e) to elevation (increasing with downward movements). The wire frames correspond to the three workspaces depicted in (a), with the corners indicating the position of the wrist relative to the monkey at the end of the movement to the corresponding target. Red bars denote the activity averaged during movement time; green bars, the activity predicted by a linear model of final wrist position in shoulder-centered spherical coordinates. Modified from Lacquaniti F, Guigon E, Bianchi L, Johnson PB, Ferraina S, and Caminiti R (1995) Representing spatial information for limb movement: The role of area 5 in the monkey. *Cerebral Cortex* 5(5): 391–409.

movement to the second target. Motor cortex is linked to SPL areas both directly and indirectly through dorsocaudal premotor cortex. In humans, during sensorimotor adaptation, the perturbation of premotor cortex activity through TMS results in a reduction of online corrections of reaching dependent on vision. The information used by motor cortex for online control of hand movement trajectory originates, at least in part, in the SPL. Here, cell activity has been studied in alert monkeys trained to make regular reaches, as well as quick corrections of hand movement trajectories, in response to a sudden jump of a visual target in 3-D space (**Figure 5(b)**). After target jump, while the eye completes the saccade to the first target and then moves to the second one, the hand smoothly changes its movement trajectory to the first target and moves toward the second one. In spite of absence of any difference in the early phase of movement trajectory in the perturbed versus the nonperturbed trials, target jump promotes an early change of neural activity in a good proportion of SPL cells. Such change, predictive of the future correction of movement trajectory, mainly influences the activity of a population of parietal neurons encoding hand tangential velocity (**Figure 5(b)**). These neurons change their firing frequency in a way that leads by a variable time the abrupt change in hand velocity occurring shortly before the change of trajectory. Some of these cells, as well as other parietal neurons in the SPL, also fire after the change of direction, suggesting that in PPC predictive signals coexist with feedback information concerning ongoing movement. Thus, parietal neurons can control and promote fast changes of hand trajectory at all times during arm movement.

Coordinate Frames for Reaching in the PPC

An approach inspired by the results of psychophysical studies in humans has shown that SPL in monkeys can be the substrate for an egocentric representation of hand position and movement since its neural activity is tuned in a body-centered reference frame, whose coordinates defined the azimuth, elevation, and distance of the hand relative to an origin centered on the shoulder or on the eyes (**Figure 6**). The analysis of neural activity through multiple-task approaches favors the existence of allocentric representations of reaching in the most caudal part of the SPL. Other studies suggest that encoding reach plan in PPC

occurs in eye or in hybrid eye–hand coordinates and that a direct transformation between these frames might occur. During copy tasks, encoding of movement seems to occur in hand coordinates. Thus, multiple representations of reaching have so far been described in PPC. It is believed that the exact nature of these representations is task-dependent.

Parietal Cortex and Interception of Moving Targets

The cortical mechanisms of interception have been studied in parietal and motor cortex of alert monkeys trained to intercept targets moving circularly with real or apparent motion or to remain immobile during target motion. It was found that the activity of most parietal cells was influenced by the direction of stimulus motion and acceleration and by the time elapsing between the target onset and its arrival and the interception point, that is, target motion time. Neural activity types dependent on the stimulus motion predominated in parietal area 7a, and those linked to hand movement were more common in the motor cortex. Decoding of spike trains showed that embedded in area 7a's neural activity is a signal concerning the target position in real and apparent motion conditions. This signal is important for initiating an interception movement. A population of cells was found to be task-specific, being activated during interception of moving targets, and not during movement to stationary target. Therefore, the distributed system underlying interception in the primate's cerebral cortex consists of gradients in which different neural activity types relevant to the task are represented at all nodes of the network, although with different strength.

For one to better understand the neural mechanisms of reaching to moving objects, knowledge of how different types of object motion influence neural activity is essential. Optic flow exerts a significant influence not only, as already known, on PPC but also on motor cortical cell activity, with expanding flow fields as the most effective stimuli in driving neural activity. In area 7a the proportion of cells sensitive to visual stimuli is higher than in motor cortex, while in the latter the sensitivity to expanding flow fields is twice that in PPC. This type of optic flow provides information about the direction of an object moving toward an immobile observer and can also be useful when self-motion and object motion interact, such as when avoiding obstacles or intercepting visual targets during locomotion. Parietal cells sensitive to optic flow do not display sensitivity to radial motion of objects moving toward or away from the fovea. Thus, encoding of object motion in the PPC might reside on separate parallel mechanisms, segregated in different populations of neurons, one based on optic flow information and related to global motion, another on opponent vector organization and mostly devoted to local motion analysis.

Finally, concerning the above-mentioned hypothesis that interception is based mostly on an internal model of gravity, functional neuroimaging indicates that the gravitational acceleration in visual motion is encoded within a distributed network centered on the so-called vestibular parietotemporal cortex. This information probably influences the motor cortex through the premotor areas of the frontal lobe.

Multiparametric Encoding of Arm Movement Variables by Parietal Neurons

In the SPL, the study of the time-varying correlation of neural activity with different arm movement parameters has shown a significant correlation with movement direction, position, amplitude, velocity, and acceleration. Velocity encoding has also been described for slow continuous movements, such as those required in a copy task. The question of whether neural activity in PPC relates to movement kinematics only, or to movement dynamics as well, has long been debated. In the SPL, cell activity seems invariant to external loads applied to the limb, suggesting that this part of PPC is a central node of the network involved in encoding arm movement in kinematic terms. In the IPL, reaching-related activity is broadly tuned to arm movement direction and position.

The Temporal Relationships between Neural Activity and Arm Movement in Parietal Cortex

The temporal relationships between neuronal activity and behavioral events can shed light on the role and position of PPC in the chain of events leading from target localization to arm reaching, in other words, from vision to movement. It is generally assumed that neural activities leading movement onset can reflect the motor command, while those occurring around and after movement initiation reflect corollary signals and/or the sensory feedback from the moving arm.

In the majority of cells of both the SPL and IPL, neural activity precedes arm movement onset (**Figures 7(a) and 7(b)**) by a variable time, and a good proportion of cells fire around or during movement. When the same cells are studied during the saccade that brings the eye on the target, in the SPL most of them show a presaccadic nature, but in the IPL they fire after the saccade, thus reflecting an eye position signal. In both the SPL and the IPL, an intriguing result emerges when the population activities relative to eye, hand, or coordinated eye–hand movement are aligned to a common behavioral event, such as

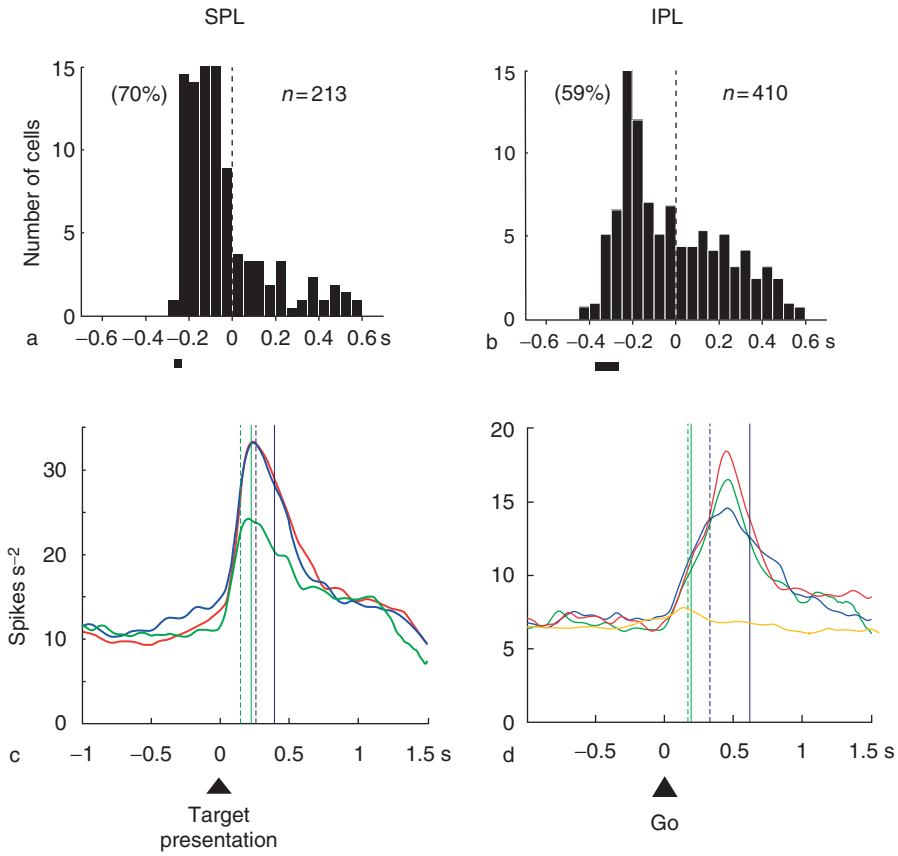

Figure 7 Temporal relationships between neural activity and arm movement in posterior parietal cortex. (a), (b) Distribution of onset times of neural activity relative to hand movement in superior parietal lobule (SPL; (a)) and inferior parietal lobule (IPL; (b)). Vertical broken lines indicate the beginning of hand movement. Percentages in brackets indicate the proportion of cells firing before the onset of hand movement. Thick horizontal bars below the x-axes refer to the beginning of reaction time ± standard deviation. (c) Population spike density functions (PSDFs) of SPL neurons during coordinated eye–hand movement (red), hand movement (blue), and eye movement (green) tasks. (d) PSDFs of IPL cells during coordinated eye–hand movement (red), hand movement (blue), eye movement (green), and in a no-go condition (yellow), in which no movement was required. In (c) and (d), the vertical lines indicate, from left to right, median values of beginning of eye movement time (MT; broken green line, in (c) at 150 ms, in (d) at 164 ms), end of eye MT (solid green, in (c) at 229 ms, in (d) at 188 ms), onset time of hand movement (broken blue line, in (c) at 265 ms, in (d) at 324 ms), end of hand MT (solid blue, in (c) at 400 ms, in (d) at 613 ms). Therefore, the two green lines encompass the median value of the duration of eye MT, and the two blue lines, the median value of the duration of hand MT. In (c) and (d), PSDFs are aligned to the presentation of the visual target, serving in both cases as the go signal for movement. Modified from Battaglia-Mayer A, Mascaro M, and Caminiti R (2007) Temporal evolution and strength of neural activity in parietal cortex during eye and hand movements. *Cerebral Cortex* 17(6): 1350–1363.

the go signal for movement (Figures 7(c) and 7(d)). In such instances, the relative curves display a common temporal evolution and peak, with SPL activity peaking at about 200 ms after go signal (Figure 7(c)), and IPL activity at about 400 ms (Figure 7(d)). In the latter, no activity is observed during a no-go task, which requires no future eye and/or hand movement. Thus, at the population level, parietal neural activity seems to be largely independent of whether a specific effector, such as the eye or the hand or both, will move but rather signals that a movement will occur or is occurring. This temporal signal can represent a common clock that can be used for different types of actions involving the eye, the hand, or their coordinated movement in different order and time sequence. It is worth noticing that there exists a strong correlation between the timing, but not the errors, of eye and hand movement under different task conditions. Interfering with parietal functions through TMS disrupts the natural correlation existing between the amplitude of eye and hand movement during coordinated eye–hand actions.

Corticocortical Circuits Underlying Reaching: Architecture of a Distributed System

In the PPC, reaching is encoded by a multiplicity of cortical areas that, thanks to their reciprocal connections with the premotor and motor areas of the frontal lobe, define a distributed parietofrontal system. The parietal reach regions so far identified, with their

frontal connections, are shown in Figure 1. In each of them, neurons combine retinal, eye, and hand information. However, in the SPL, a gradient of functional properties characterizes these different areas. In the most posterior part of the network (areas V6A and 7m), retinal, eye position, and eye movement information predominate over coexisting hand signals. At intermediate rostrocaudal levels in the SPL (areas PEc, PEa, and MIP), retinal, eye, and hand information influences cell activity in a similar fashion, while at the rostralmost level (area PE), hand position and movement information dominates over coexisting eye influences. These SPL areas project to and receive from different premotor areas and from motor cortex itself. A similar, mirror-image gradient of functional properties exists in the frontal lobe, with parietofrontal connections linking regions displaying similar functional properties. Superimposed on this functional dimension is a second one, related to the nature of the information encoded in the network. Moving from posterior to anterior in the SPL, a transition from preparatory (set- and memory-related) to movement signals occurs. The opposite is observed across premotor and motor cortices. Cells encoding eye and/or hand position information are ubiquitous at all tangential locations of the SPL, so as to form a matrix on which preparatory and movement-related information is embedded and eventually selected on the basis of task demands. Eye and hand position signals are essential under virtually all conditions for planning combined eye–hand movements.

The functional architecture of the parieto-frontoparietal network imposes constraints on the nature of the coordinate transformations underlying visually guided reaching. In fact, this cannot be regarded as the result of coordinate transformations performed by each parietal area in its own private coordinate frame but is the result of a recursive mechanism in which different signals are selected and combined in a context-dependent fashion throughout the network and further elaborated locally, thanks to intrinsic connections.

Consequences of PPC Lesions on Arm Movements

For the properties and position of PPC in the distributed network controlling arm movements, and for its lack of projections to the interneurons and motor neurons of the spinal cord, damage to the PPC results not in primary motor deficits, such as paralysis, changes of muscle tone, and so forth, but in higher-order motor disorders, interpreted as consequences of the disruption of the parietal operations and of their interplay with the premotor and motor areas of the frontal lobe.

In the neuropsychological literature, two of the main arm-movement disorders described after parietal lesion are OA and directional hypokinesia (DH). The first is more common after lesion of the SPL, while the second is classically associated with damage of the IPL.

Lesions of the SPL: OA

In 1909, Rudolf Bálint first described OA as a key feature of a more general syndrome that was a consequence of a bilateral softening of the parietal lobe. As Bálint pointed out, OA produced a profound impairment of his patient's daily activities (Figure 8). In fact, "while cutting a slice of meat . . . which he held with a fork in his left hand, he would search for it outside the plate with the knife in his right hand" (Figure 8, top panel), or "while lighting a cigarette he often lit the middle and not the end" (Figure 8, center panel, left). Bálint pointed out the systematic nature of this disorder, which was evident in the patient's

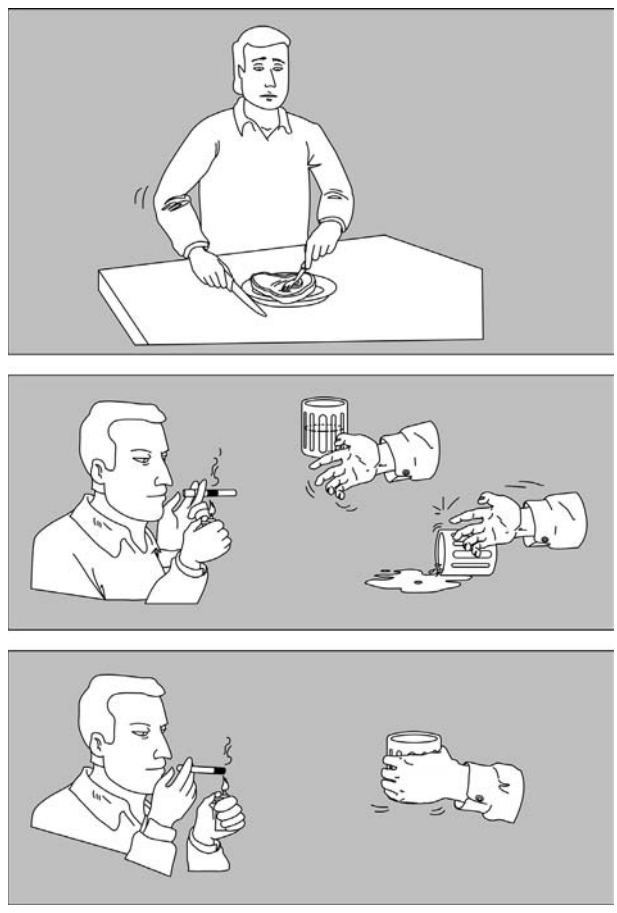

Figure 8 Schematic drawing of the impairments typical of optic ataxia, as first described by Bálint in a patient with bilateral cortical softening centered on the parietal lobes. For explanations, see text. Original drawings by C Cardone.

behavior when reaching out in his peripersonal space. "Thus, when asked to grasp a presented object with his right hand, he would miss it regularly, and would find it only when his hand knocked against it" (Figure 8, center panel, right). The patient performance in the same tasks was normal when using the left hand (Figure 8, bottom panel), which excluded disorders of visual attention. Furthermore, in the absence of visual control, he could imitate with the right hand passive postures imposed on his left hand, and this excluded apraxic disorders. In Bálint's view, the disorder was specifically dependent on a defective visual control of movement, and he named it OA, in analogy with tabetic ataxia, which was a consequence of a defective somatosensory control of movement resulting from damage of the dorsal columns of the spinal cord. OA impairs not only visual reaching but also hand grip formation and prehension under visual control. In addition, OA patients suffer from a severe disturbance of the spatial and temporal characteristics of online visual control of movement. Corrective movements prompted by changes in target location are delayed, and hand movement trajectory does not show the typical smooth curvature that brings the hand to the new target's location. In these cases, the hand completes the movement toward the location where the target was first presented and then moves toward the new one (Figure 5(a)). Obstacle avoidance during reaching is also impaired. The task-dependent nature of OA is documented by the observation that it occurs during reaching or grasping for real targets, not when pantomiming hand grasping movement of memorized objects.

OA is consequence of bilateral or unilateral parietal lesions. In all instances, arm movement errors mostly occur in the peripheral visual field. Right parietal lesions result in misreaching with one or both hands in the contralateral visual field (field effect). Left parietal lesions cause misreaching with the right hand in both ipsi- and contralateral visual fields (hand effect). When lesions mostly affect the left SPL, or when a rapid correction of movement trajectory is required, OA is not confined to the peripheral visual field but also occurs in central vision.

Lesions of the IPL: Disturbances of the Action Space and DH

In general, the disturbances of the representation of action space in neglect patients is well evidenced by their motor behavior when requested to cross out with a pencil the local components of a global letter in the hierarchical figures of Navon (Figure 9(a)).

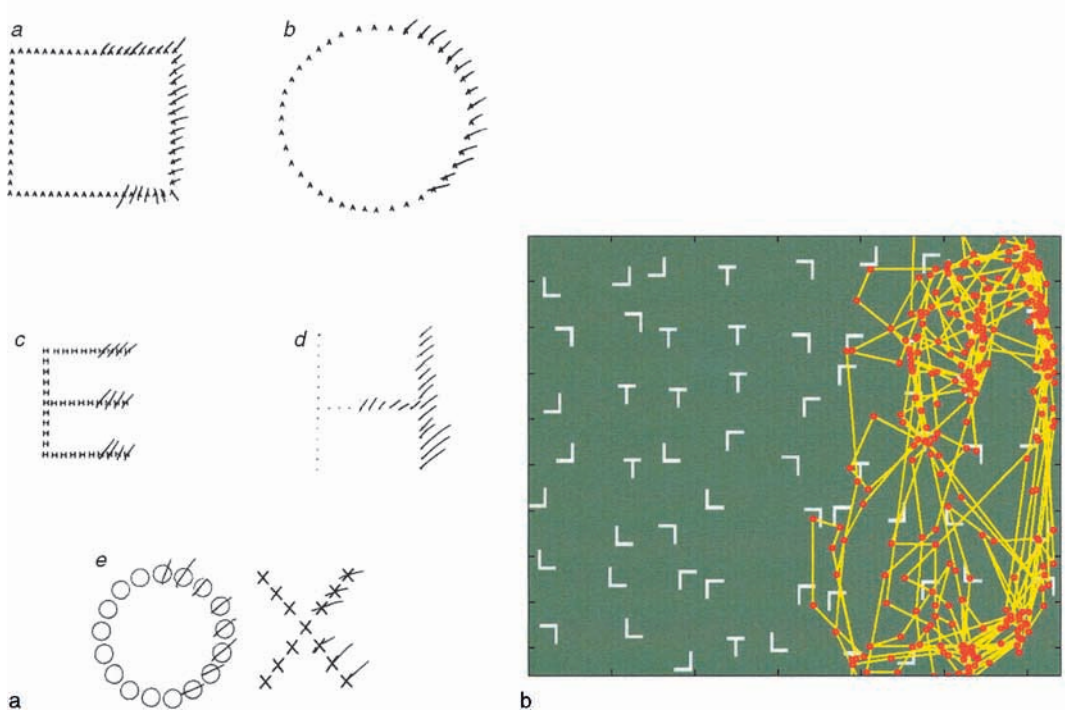

Figure 9 Movement disorders of patients with hemispatial neglect from right parietal lesion. The left part of the action space is ignored. (a) Performance when crossing letters forming the Navon figures. Reprinted by permission from Macmillan Publishers Ltd: *Nature* (Marshall JC and Halligan PW (1995) Seeing the forest but only half the trees? *Nature* 373: 521–523.), copyright (1995). (b) Records of eye movements during a visual search task in which participants are required to identify the letter T within a field of distractors. Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Neuroscience* (Husain M and Rorden C (2003) Nonspatially lateralized mechanisms in hemispatial neglect. *Nature Reviews Neuroscience* 4(1): 26–36.), copyright (2003).

Under such conditions, while being able to correctly name both local and global components of the figure, thus showing no perceptual impairments, these parietal patients cross out only the local components in the right part of each global figure, ignoring those located on the left.

DH is the hallmark of arm movement disorders of neglect, with which it is generally associated, in patients suffering from right PPC lesions. DH consists of a prolongation of hand reaction time for targets in the contralesional space, especially those requiring arm movement with leftward direction. The disorder is diagnosed when the ispilesional arm is involved. DH is often accompanied by directional bradykinesia and hypometria, characterized by reduced velocity and amplitude of arm movement, respectively. In tasks requiring patients to reach at targets within a field of distractors, the hand of patients with neglect recursively points to the same targets in the right part of the space, as a consequence of the inability to update a motor program on the basis of previous movements. The same occurs when the search is performed through eye movements (Figure 9(b)). In neglect patients, similar disorders are also common in the oculomotor domain and consist of elongation of reaction time when the eye moves toward contralesional targets, reduced saccade amplitude, and eye movement trajectories characterized by a 'staircase' shape.

A Positive Image of OA and DH: Interpreting the Movement Disorders of Parietal Patients from a Physiological Perspective

The study of the dynamic properties of PPC neurons and of their relationships with the premotor cortex via corticocortical connections provides the basis for a positive image of the arm movement disorders of parietal patients. In fact, parietal neurons are influenced by different signals concerning planning and execution of eye and hand movement. A key feature of SPL neurons is their global tuning filed, that is, the invariance of their preferred directions across tasks requiring different spatial relationships between the eye and the hand during movement to visual targets. This spatial congruence of different signals might be regarded as the neural basis of eye–hand coordination. Furthermore, these neurons are part of a network richly interconnected with premotor and motor cortices. Therefore, the reaching disorders of patients suffering from OA is today interpreted as a consequence of the breakdown of the combinatorial mechanisms of SPL parietal neurons and of their interplay with premotor cortex.

In the IPL, one of the main spatial properties of eye- and hand-related cells is the anisotropy of the distribution of their preferred directions, mostly pointing toward the contralateral space. This anisotropy of the directional motor space might explain the difficulty in movement toward the contralateral space that characterize DH in patients with hemispatial neglect.

See also: Motor Psychophysics; Posterior Parietal Cortex and Tool Usage and Hand Shape; Reaching and Grasping; Sensorimotor Control of Manipulation; Spatial Transformations for Eye–Hand Coordination.

Further Reading

Andersen RA and Buneo CA (2002) Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience* 25: 189–220.

Batista AP, Buneo CA, Snyder LH, and Andersen RA (1999) Reach plans in eye-centered coordinates. *Science* 285: 257–260.

Battaglia-Mayer A, Archambault PS, and Caminiti R (2006) The cortical network for eye–hand coordination and its relevance to understanding motor disorders of parietal patients. *Neuropsychologia* 44(13): 2607–2620.

Battaglia-Mayer A and Caminiti R (2002) Optic ataxia as a result of the breakdown of the global tuning fields of parietal neurons. *Brain* 125: 225–237.

Battaglia-Mayer A, Caminiti R, Lacquaniti F, and Zago M (2003) Multiple levels of representation of reaching in the parieto-frontal network. *Cerebral Cortex* 13: 1009–1022.

Buneo CA, Jarvis MA, Batista AP, and Andersen RA (2002) Direct visuomotor transformations for reaching. *Nature* 416: 632–636.

Caminiti R, Ferraina S, and Johnson PB (1996) The sources of visual input to the primate frontal lobe: A novel role for the superior parietal lobule. *Cerebral Cortex* 6: 319–328.

Cohen YE and Andersen RA (2002) A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience* 3: 553–562.

d'Avella A, Portone A, Fernandez L, and Lacquaniti F (2006) Control of fast-reaching movements by muscle synergy combinations. *Journal of Neuroscience* 26(30): 7791–7810.

Fogassi L, Ferrari PF, Gesierich B, Rozzi S, Chersi F, and Rizzolatti G (2005) Parietal lobe: From action organization to intention understanding. *Science* 308: 662–667.

Indovina I, Maffei V, Bosco G, Zago M, Macaluso E, and Lacquaniti F (2005) Representation of visual gravitational motion in the human vestibular cortex. *Science* 308(5720): 416–419.

Johnson PB, Ferraina S, Bianchi L, and Caminiti R (1996) Cortical networks for visual reaching: Physiological and anatomical organization of frontal and parietal lobe arm regions. *Cerebral Cortex* 6: 102–119.

Lacquaniti F, Guigon E, Bianchi L, Johnson PB, Ferraina S, and Caminiti R (1995) Representing spatial information for limb movement: The role of area 5 in the monkey. *Cerebral Cortex* 5(5): 391–409.

Land M, Mennie N, and Rusted J (1999) The role of vision and eye movement in the control of activities of daily living. *Perception* 28: 1311–1328.

Mattingley JB, Husain M, Rorden C, Kennard C, and Driver J (1998) Motor role of human inferior parietal lobe revealed in unilateral neglect patients. *Nature* 392(6672): 179–182.

Merchant H and Georgopoulos AP (2006) Neurophysiology of perceptual and motor aspects of interception. *Journal of Neurophysiology* 95: 1–13.

Pisella L, Grea H, Tilikete C, Vighetto A, Desmurget M, Rode G, et al. (2000) An 'automatic pilot' for the hand in human posterior parietal cortex: Toward reinterpreting optic ataxia. *Nature Neuroscience* 3(7): 729–736.

Shadmehr R and Wise SP (2005) *Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge, MA: MIT Press.

Wise SP, Boussaoud D, Johnson PB, and Caminiti R (1997) Premotor and parietal cortex: Corticocortical connectivity and combinatorial computations. *Annual Review of Neuroscience* 20: 25–42.

Relevant Website

w3.uniroma1.it – Laboratory of Behavioral Neurophysiology.

Posterior Parietal Cortex and Tool Usage and Hand Shape

A Iriki, RIKEN Brain Science Institute, Saitama, Japan

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Hands are the most frequently and dexterously used parts of our bodies. This is directly reflected in the disproportionately large representation of hands in the homunculi of the somatosensory and motor cortices. We manufacture and use tools with our hands. The hand is also the most extensively observed part of our own body. We look carefully at our fingertips when we manipulate objects. We look at a distant object and then extend our hand to reach and retrieve it under observation. And occasionally we explore space with our hands instead of using our eyes. Why are primates, especially humans, so intensely interested in their hands?

This article first focuses on the primate hand from the perspectives of its comparative evolution and unique functionality. Brain functions subserving manual behaviors are then discussed from an environmental point of view. Finally, the article considers how eye–hand coordination and the brain functions associated with it are essential for tool use.

The Evolution of the Hand as the Distal Extremity of the Forelimb

The hand is the distal extremity of the forelimb. First, let us consider the significance of this structure from the viewpoint of vertebrate evolution. Forelimbs are thought to originate from the pectoral fins of fish, although this view remains somewhat controversial. With some assistance from other fins, the pectoral fins are responsible for locomotion in fishes – swimming. Later in vertebrate evolution, amphibians and reptiles moved onto the land. Here, fins evolved into limbs to support their bodies against gravity and created the impulsive locomotive forces that enabled these animals to walk forward. In birds (and *Chiroptera*), forelimbs became wings, but they still had a locomotive function – flying. Finally, the major functions of limbs retained for most animals, including most mammals, were those for supporting and moving their bodies in their habitats. Thus, during the entire course of vertebrate evolution, the hands were merely the distal ends of the forelimbs, forming an interface that effectively interacts with the surface of the environment through which the animal moves. Some hands have thin webs of skin between the fingers to move water or have sharp nails at the ends of fingers to scratch and grip the soil, and still others have fluffy feathers for flapping through the air.

In these vertebrate ancestors, forelimbs have always been attached either underneath or at the sides of the trunk, the most effective locations for their primary antigravity and locomotive functions. Thus, limbs were never previously positioned in front of the face, so that these animals never, or rarely, were able to observe their hands. Hence, the control of the forelimb and hand relied exclusively on somatosensory feedback, without visual information for monitoring their motion.

However, primates evolved with dramatic changes to their forelimbs and hands. In order to adapt to the arboreal environment, primate limbs became greatly elongated compared with their trunk, enabling them to efficiently reach for branches. In addition, their hands became enlarged, with flat nails at the fingertips for more secure grasping. Accurate control of the hands as they reached for and gripped branches required that their location and shape in space had to be monitored and confirmed. The hands, at the distal extremities of the elongated forelimbs, had to be constantly moved to positions exactly in front of the face, to bring them under precise spatial control at the center of the visual field.

Eye–Hand Integration in Primates

The primate hand continued to evolve from its original functions in quadrupedal locomotion. In addition to the precise visuomotor coordination of hand and arm movements needed for locomotion in the arboreal habitat, the hand in primates developed an opposing thumb for firmly grasping branches that were visually identified and targeted.

In order to establish the visual functions required for precise reaching and grasping, vision evolved in parallel with the hand. Paired eyes that had been located on the sides of the head in earlier mammals were now on the front of the head and faced forward. This gave primates binocular vision, due to the largely overlapping visual fields of the two eyes, making precise analysis of depth information possible. Binocular vision may have evolved less in other species because it was beneficial for them to secure wider, less overlapped visual fields in two eyes located on each side of the skull. Exceptions to this tendency are the carnivores and raptors that hunt for prey. These animals are different from primates in that

they reach their prey, not by extending their forelimbs, but rather by locomotive jumps.

The hands and visual systems of primates have thus been refined through evolution for coordination, resulting in the emergence of further unique features. Flexible supination and pronation of the forearm allow primates to position the palms of their hands, which are on the bottom surfaces of the limbs, in front of their eyes. Thus, the subtle space structured by precise finger and opposing thumb manipulations is manifested at the center of the binocular visual field, which, with its maximal spatial resolution, is capable of analyzing fine, three-dimensional (3-D) structures. With this visuomanual coordination, abundant somatosensory information about hand–finger shapes can be subjected to visual confirmation. Somatosensory feedback from the hand must originally have contributed to the precise control of locomotive effectors. Moreover, by linking its functioning with vision, the hand has become capable of exploring space through its own sensations – the whole hand has become a sensory organ. Hence, we grope with our hand and thereby feel the shape of space. Alternatively, we can confirm, verify, and inspect visually perceived spatial structures by motor and tactile information acquired through moving the hand and fingers across their surfaces. Primates have now acquired, through eye–hand integration, a flexible, precise, cross-modal (visual plus somatosensory) map of space. This function, never previously possessed by any nonprimate species, reaches its highest level of elaboration in humans.

Somatosensory Traces of Hand Shape

Somatic sensory information about postures, locations, and movements of the hand is detected through millions of receptors distributed throughout the skin, joints, muscles, and other tissues. Each of these receptors responds to a specific physical energy, namely, an adequate stimulus to its receptive fields. Detected information is transformed into sequences of nerve impulses that are conducted through peripheral nerves and specific ascending pathways in the central nervous system. The trains of nerve impulses, without much modification, reach the primary somatosensory cortex (S1), located at the postcentral gyrus occupying the most anterior part of the parietal cortex. The S1 is somatotopically organized. Tactile information is projected onto Brodmann's Area (BA) 3b, whereas joint and proprioceptive information projects onto BA3a. Thereafter, with hierarchical, intracortical processing, these information fragments are integrated and transformed until they gradually become meaningful.

The response properties of BA3 neurons are generally robust and with sharp contrast, resembling those of the primary afferents that innervate their respective receptors. The receptive fields are small, and the submodalities of adequate stimuli (e.g., light palpation, joint displacement, flutter, or vibration) are simple and clear. Hence, the shape or posture of the entire hand could never be estimated merely by observing these simple neuronal responses.

Shifting toward the posterior S1 areas BA1 and BA2, neuronal response properties gradually become complex: receptive fields are larger, with multiple submodalities of adequate stimuli, as if by the gradual convergence of the elementary BA3 responses. Integration in the posterior BA2 is more complex than a summation of elements, characterized by the extraction of spatiotemporal patterns that represent characteristic features of the stimuli across wide areas of the skin surface of the hand. They detect, for example, the orientation of edges, the surface texture of objects being touched, or the direction of their movements. Posterior BA2 neurons also code particular dynamic shapes of the hand and fingers, such as scratching, pointing, or picking. Further, some neurons respond only when the hands are actively touching an object, not when being touched by the object. The function of the latter class of neurons may be to verify the sensations caused by the movement of one's own hands.

In these posterior S1 areas, neuronal response properties become less well contrasted, less robust, and less stable. In the anterior bank of the intraparietal sulcus (IPS), including BA5/7, response properties are particularly difficult to characterize. The somatosensory responses of some of these neurons can be influenced by attentional states or by intervening visual stimuli. There are neurons in the finger region of the anterior IPS area with bilateral receptive fields, via transcallosal interconnections. The receptive fields of these neurons are always strictly symmetrical. These receptive field properties cannot be explained by bottom-up mechanisms, suggesting influences from images about the shape of the hands, perhaps by feedback from higher areas of the association cortex.

Thus, as a result of S1 hierarchical information processing, integration with visual information about the location and shape of the hand should be readily prepared at the gateway to the parietal association area so that the functional representation of hand shape can be generated. Indeed, in the arm region of the parietal BA5 that is posteriorly adjacent to the S1 exist neurons that respond when monkeys assume particular arm postures, when they touch particular parts of the body with their own hands, or when the monkeys watched these actions.

Examination of the Visual Space at Hand

Visual information is first captured by retinal cells, pixel by pixel, and then projected via the medial geniculate nucleus in the thalamus to the primary visual cortex, at the posterior end of the cerebrum. The hierarchical processing of this information then proceeds forward in the cerebral cortex. The part of the visual system relevant to this review is the spatial information processing that takes place along the dorsal stream. To understand the relationships between self and external spatial structures, the processing of optical-flow information that deals with self-motion in space and the processing of 3-D depth information about the nearby space are both crucial. These abilities are well developed in primates to subserve arboreal locomotion using hands, as described above.

How does one perceive oneself moving in space? When something moves in space, the image of that object in the visual field moves. When one looks around, the whole visual image moves. The visual image, however, exhibits peculiar flowlike movements when one moves in space. Thus, in order to analyze our own 3-D movements in space, optical flow information from the wider visual field is needed. Neurons that code optical flow are first found along the dorsal stream in the middle temporal area (MT/V5) of the extrastriate cortex at the temporoparietal junction region in monkeys. However, because receptive fields of these MT/V5 neurons are generally small, the information carried by these neurons is not sufficient for analyzing self-motion.

The medial superior temporal (MST) area is adjacent to and receives direct projections from MT/V5. The receptive fields of MST neurons sensitive to optical flow are large and cover most of the visual field. Rotating movements in 3-D space with changes in depth information can also drive MST neurons. Similar response properties are observed in the ventral intraparietal area (VIP), at the fundus of the IPS. The VIP receives direct inputs from the MT/V5 and integrates these inputs with facial somatic sensations to represent movements in space relative to the face. Considering that the lower half of a monkey's whole visual field is occupied by its own face, these neurons could code self-motion by integrating motion relative to the face and environmental space, using the face as a reference point. Vestibular information also reaches the MST area and the VIP, perhaps contributing to the perception of self-motion in space.

Gross depth information based on optical flow information alone is insufficient for the precise analysis of 3-D structures in nearby space. Detailed 3-D analyses require the binocular parallax information that is generated by two forward-facing eyes. The caudal intraparietal area (CIP), the visual area that possesses neurons specialized for binocular analyses, receives input from V3A. CIP neurons code special 3-D information about features such as slope, depth, and curvature that represent the precise characteristics of the shapes of 3-D objects in the visual space of the hands. Neurons in V3A are also parallax sensitive, but their receptive fields are small and thus probably do not integrate overall shape discrimination in the central visual field. This resembles the relationship between MT/V5 and MST neurons, as described above.

Before bipedal locomotion evolved in humans, primates were able to sit, which freed their forelimbs for executing delicate handwork such as grooming. In doing so, primates position their hands and fingers just in front of their faces, where binocular vision has the highest resolution. There they rotate their hands by supination and pronation of the wrist and shape the fingers and opposing thumbs in a most efficient way. Thus, the evolution of primates has produced the dexterous hand, the delicate binocular eyes, the brain circuitry for processing detailed visual information, and the ability to precisely control hand and finger movements.

Adjustment of Hand Shape with Spatial Structures

In the primate brain, somatosensory and visual information about the hand is thus prepared across the IPS for the formation of cognitive maps of hand shape in space. One aspect of this predominantly visual process occurs in the anterior intraparietal area (AIP) on the posterolateral bank of the IPS, a posterior extension of the area of S1 finger representation. Neurons in the anteromedial bank that faces the AIP across the IPS code particular combinations of shapes of the hand and fingers, but they only sparsely integrate that representation with visual information, in naïve monkeys.

AIP neurons code hand shape when monkeys perform tasks involving precise finger actions such as pinching tiny 3-D knobs or picking up tiny objects from variously shaped small holes. In these experiments, target objects were first shown to monkeys in the light, and then the monkeys were asked to reach for and manipulate these objects either under light or in the dark. AIP neurons discharged when monkeys (1) performed hand and finger actions either in the light or in the dark, but not when just watching the target (motor-dominant neurons); (2) watched the target and executed actions under the light and, though much weaker, even in the dark (visuomotor

neurons); or (3) observed the target or made actions in the light but not in the dark (visual-dominant neurons). These neurons may contribute to shaping the hand prior to actually grasping a target, namely, preshaping. Indeed, patients with damage to this parietal region have difficulties in preshaping while reaching. If activity in this area is blocked by an injection of muscimol (a γ-aminobutyric acid agonist), monkeys are unable to make preshaping movements.

Visual-dominant AIP neurons fire when monkeys look at an object but do not make any actual hand movements. This property resembles the responses of CIP neurons to particular 3-D objects, as described above. Hence, the CIP processes precise 3-D information about objects, and this transformed information is in turn sent to the AIP through its strong, direct projections. Here it is integrated with somatosensory information processed backward from the S1 to match the hand shape to the 3-D shapes of objects that will be grasped with the hand. An efference copy for making hand actions from the ventral premotor area (PMv), through its strong mutual connections to the AIP, may also contribute to this process. Neurons exist in the AIP that appear to memorize the hand shape needed to grip a particular object. The activities of these neurons are sustained during a delay period of several seconds after the monkey has observed the object to be grasped, perhaps coordinating the memory of the shapes of the object and the hand. Thus, visual, somatosensory, and motor information regarding hand shape is integrated in this AIP area in a visually dominant way, with the visual analysis of spatial structures matched and verified through the hand and finger actions needed to trace them.

Optimization of Hand Shape for Tool Use

The ability of the brain to adjust the shape of the hand and fingers to the shape of the object to be manipulated, as described above, is very important when humans attempt to hold and use a tool with the hands. Functional imaging of the human brain reveals that intraparietal areas, including the AIP area described above, are activated when humans see the tool they intend to use, suggesting that humans automatically bring to mind the shape and meaningful hand and finger actions needed to use it. Other cortical areas also commonly activated include the PMv and the superior temporal sulcus area. It is suggested that the PMv represents the actual hand actions needed for tool usage, whereas the superior temporal sulcus represents the semantic meaning of the tool function. Thus, if the brain apparatus for matching hand shape to representations of spatial shape already existed in lower primates, humans could easily adapt these mechanisms for effectively grabbing and using external objects as tools.

Plasticity of Parietal Hand Representation on Tool Use

The mechanisms described above imply that a tool is represented in the parietal cortex through the shape and actions of the hand using it. But is this all there is to it? When we use a tool, it is not just an external object *per se*: we regard it as an object incorporated into part of our body, the externalized hand. The tool becomes a part of the body, but one that lacks direct somatic sensations. The visual appearance of the body, with the tool, must be transformed into a somatosensory representation of the body, in contrast to the AIP, where the visual image of the space is adjusted by the somatosensory representation. In monkeys trained in tool use, that transformation appears to take place in neurons situated in part of the parietal cortex close to somatosensory regions, on the anterior bank of the IPS that faces the AIP across the sulcus. This area is the posterior extension of the S1 forearm representation.

Macaque monkeys do not use tools spontaneously in nature, but they can be trained over a period of about 10 days to use a rake to retrieve distant food items that are beyond the reach of their arms. In naïve monkeys prior to training, only sparse visual inputs were found in the part of the anterior bank of the IPS discussed above. But after tool-use training, bimodal neurons that respond to both somatosensory and visual inputs were found. These neurons appeared to code the image of the hand in space, with the visual receptive field encompassing the somatosensory receptive field. When the latter moves by hand movement, the former follows accordingly.

When a monkey uses tools, these visual receptive fields extend along the axis of the tool to include the tool. This is as though the functional and internal introspection by the monkey using the tool has resulted in the extension of the representation of the distal forearm to incorporate the tool as part of the arm. The brain may do this by matching its representation with the visual appearance of the newly acquired functional forearm. This incorporation of the tool into the mental body image was never induced in monkeys that did not use the object as a tool, even though monkeys were physically holding it. Thus, it is not merely a representation of a material feature of the body but represents a cognitive map of the internal and external space, integrated in a manner

that reflects the mental image of the body shape with respect to its functional significance.

When humans play video games or use teleoperator apparatus with a TV monitor, they experience the introspective projection of their body image onto those appearing on the monitor. Thus, humans accept that the body image projected onto the monitor is part of themselves, in spite of its separation from them. Monkeys have also been successfully trained to perform an action, and even to use tools, by relying only on the image projected on the monitor and without directly viewing the image of their hand. During these behaviors, the visual receptive fields of IPS bimodal neurons, described above, became responsive to the image on the monitor, encompassing the projected image of the hand, including tools if tools were used. When images of the hand (and tools) in the monitor were expanded, shrunk, or dislocated by video effects, the visual receptive fields in the visual space of the monitor changed to match the modified images in the monitor. Thus, somatosensory images of the shapes of the body are not fixed but can be extended to include information presented on the monitor, even though these images are discontinuous and projected some distance from the body. Further, the shapes can be transformed to match an artificially modified visual appearance. These representations are found in the intraparietal cortex.

Plasticity of Parietal Space Representation with Tool Use

In the AIP, neurons match the precise shapes of 3-D space with the shape of the hand. When the shape of the arm and hand is functionally altered by holding a tool, this integration must change accordingly. When the monkey holds to use a rake, the rake becomes an extension, not of the hand itself, but of the functional length of the forearm. Indeed, the bimodal activities in the area posteriorly adjacent to the S1 area of the forearm representation, as mentioned above, were established by training. Thus, the change in the spatial representation produced by holding the rake should include the space reached and explored through forearm movements. This change in the spatial representation was found in the bimodal neurons in that same anterior IPS region.

IPS bimodal neurons with somatosensory receptive fields located on the proximal part of the body (shoulder, neck, or face) had visual receptive fields that covered the reaching distance. Before tool use, the visual receptive fields of these neurons covered only the areas within the reaching distance of hands, whereas after tool use, they expanded to include the extra distance that could be reached by using the tools. Thus, this location appears to contain a process that functionally recalibrates the perceived space, taking the shape of one's own body as a reference. This coding of near versus far space is being studied in brain-damaged patients (particularly those with damage to their parietal cortices). These patients exhibit retardations of body image and peripersonal space, also the phenomena typically observed in patients with somatosensory–visual cross-modal extinction.

Expansion of the Hominid Parietal Cortex in Evolution with Dexterous Tool Use

Higher primates have larger brains than other animals. Which parts of the brain have expanded? In lower mammals such as rodents, the primary cortices occupy most of the cerebrum. In the lower primates, perhaps due to the dramatic improvement in the visual system, especially with binocular vision, the primary visual area expanded first. This expansion was followed by the expansion of the anteriorly adjacent parietal association cortex required for visual processing. The prefrontal cortex emerges, yet remains tiny at this evolutionary stage. Then, with the evolution of the hominids, the next expansion in the cortical area was typically of the parietal cortex, as determined from comparisons of the endocast brains from primitive hominid skulls with the brain of the chimpanzee. At the same time as this parietal expansion, hominin ancestors successively invented various kinds of tools. Whether tool use itself triggered the expansion of the parietal cortex, however, is not known.

During weeks of training in tool use in monkeys, normally weak visual inputs in the IPS area became more evident. The coding of the shape of the body image is modified to include the tools. Corresponding to this phase of training, immediate early genes and neurotrophic factors are expressed in this brain region. Rich arborization and extensions of axons that project onto this area, together with synaptogenesis, are induced. These axons originate from the area of the temporoparietal junction, in the vicinity of the MT/V5 or MST area. This enhanced projection may contribute to the perception of spatial motion relative to the hands, equivalent to the VIP coding in relation to the face, as argued earlier. Hence, demand by tool use may have induced the enlargement of the parietal cortex. In monkeys, the morphological changes in this area could be induced by artificial demand, whereas in hominin ancestors, either there may have been mutational changes that automatically

induced this morphogenesis or the changes were more easily triggered during normal development. These mechanisms remain hypothetical.

Summary and Perspectives

The body structure of primates allowed them to place their hands (the distal extremity of the forelimb) in front of their face. From this position, 3-D structures could be precisely analyzed using precise binocular vision. Demand for such elaborate shaping of hands under visual supervision may have induced the expansion of the parietal cortex in primates. These mechanisms enabled higher primates to handle primitive tools, and then the tool use process *per se* may have triggered further expansion of the parietal cortex. This is the region of the cortex where not only somatosensory and visual inputs but also vestibular and auditory inputs converge. Therefore, it can be characterized as a multisensory cortex. Such parietal multimodal integration may not be limited to interpretations of represented shapes and the meanings of spatial structures. Humans express various conceptual ideas analogously to spatial processes. Humans also express the manipulation of those abstract ideas using the analogy of the body, typically by hand actions (the term 'manipulate' emphasizes this). Thus, the parietal processing of hand shape in space may be directly extrapolated to higher intellectual functions in humans.

See also: Finger Movements: Control; Motor Psychophysics; Parietal Cortex and Spatial Attention; Posterior Parietal Cortex and Arm Movement; Reaching and Grasping; Sensorimotor Integration: Models; Sensorimotor Control of Manipulation; Somatosensory Cortex; Spatial Transformations for Eye–Hand Coordination.

Further Reading

Andersen RA, Snyder LH, Bradley DC, and Xing J (1997) Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* 20: 303–330.

Castiello U (2005) The neuroscience of grasping. *Nature Review Neuroscience* 6: 726–736.

Fogassi L and Luppino G (2005) Motor functions of the parietal lobe. *Current Opinion in Neurobiology* 15: 626–631.

Hihara S, Notoya T, Tanaka M, et al. (2006) Extension of corticocortical afferents into the anterior bank of the intraparietal sulcus by tool-use training in adult monkeys. *Neuropsychologia* 44: 2636–2646.

Iriki A (2005) A prototype of *Homo faber:* A silent precursor of human intelligence in the tool-using monkey brain. In: Dehaene S, Duhamel J-R, Rizzolatti G, and Hauser M (eds.) *From Monkey Brain to Human Brain*, pp. 253–271. Cambridge, MA: MIT Press.

Iwamura Y, Iriki A, Tanaka M, Taoka M, and Toda T (2001) Processing of higher order somatosensory and visual information in the intraparietal region of the postcentral gyrus. In: Rowe MJ and Iwamura Y (eds.) *Somatosensory Processing: From Single Neuron to Brain Imaging*, pp. 101–112. Amsterdam: Harwood.

Iwamura Y, Tanaka M, Sakamuto M, and Hikosaka O (1993) Rostorocaudal gradients in neural receptive field complexity in the finger region of alert monkey's postcentral gyrus. *Experimental Brain Research* 92: 360–368.

Jeannerod M, Arbib M, Rizzolatti G, and Sakata H (1995) Grasping objects: The cortical mechanisms of visuomotor transformation. *Trends in Neurosciences* 18: 314–320.

Johnson-Frey SH (2004) The neural bases of complex tool use in humans. *Trends in Cognitive Sciences* 8: 71–78.

Maravita A and Iriki A (2004) Tools for the body (schema). *Trends in Cognitive Sciences* 8: 79–86.

Rizzolatti G, Fogassi L, and Gallese V (1997) Parietal cortex: From sight to action. *Current Opinion in Neurobiology* 7: 562–567.

Sakata H, Taira M, Kusunoki M, Murata A, and Tanaka Y (1997) The TINS lecture: The parietal association cortex in depth perception and visual control of hand action. *Trends in Neurosciences* 20: 350–357.

Post-Polio Syndrome

E Farbu, Stavanger University Hospital, Stavanger, Norway

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The first description of new muscular weakness and atrophy several years after acute paralytic poliomyelitis (polio) was given by Raymond and Charcot in 1875, when they reported a 19-year-old man with new atrophy in his shoulder following paralytic polio in childhood. In the following centenary, paralytic polio was considered to be a three-phasic illness with acute muscle weakness, recovery, and a stable phase, and Charcot's observation was not investigated further until several reports of increasing muscle weakness emerged in the 1980s. A clinical description and definition were given by Halstead in 1985 and revised in 1991, and today we have increasing knowledge about the clinical course, possible etiologies, and clinical management. This article reviews the background, features, and management of post-polio syndrome (PPS).

Acute Paralytic Polio

The term poliomyelitis originates from the Greek words *polios* (gray), *myelos* (marrow), and the Latin suffix *itis* (inflammation) and refers to the specific localization of the infection in the anterior horn (gray matter) in the spinal cord. Acute paralytic polio is caused by infection of the poliovirus, which is enterovirus of the Picornaviridae family. There are three serotypes of polio viruses; they were first identified in 1931 and later named PV-1 Mahoney, PV-2 Lansing, and PV-3 Leon. It is believed that paralysis is most common with PV-1. The polio virus is an RNA virus in which the virion consists of a single-stranded RNA genome of positive polarity in a nonenveloped icosahedral protein capsid. In the majority of cases, an immunization with no clinical symptoms takes place, although 1% of the affected persons develop fever and signs of aseptic meningitis and flaccid paralysis. The virus attacks the motor neurons in the anterior horn of the spinal cord, which can cause flaccid muscle weakness if the inflammation is severe enough. The most common localization of weakness is in the lower limbs, but all levels of the spinal cord can be affected, including the upper cervical cord and brain stem, causing respiratory difficulties. A random distribution of the weakness is common.

During the large epidemics in the United States and Europe in the 1940–50s, the mortality rate was about 7%, mainly due to respiratory weakness. Sabin's oral and active vaccine and Salk's inactivated vaccine were introduced in the 1950s, and they were quickly implemented in the industrialized world. This led to a formidable reduction of polio epidemics, and new cases of paralytic polio diminished rapidly. Europe, Australia, and the Americas have now been declared polio-free, with no reported polio cases over the past years. The last case of acute paralytic polio in Europe occured in Turkey in 1998. Despite the worldwide vaccination campaign, there are four countries in Africa and Asia – Nigeria, India, Afghanistan, and Pakistan – where the polio virus is still endemic and new epidemics occur; the number of new polio cases in 2006 was 1912. The poliovirus is highly infectious. Transmission into nonendemic countries occurs, and the mutation rate of the virus is high. Worldwide surveillance and vaccination will therefore be needed for several years before a polio-free world can be achieved.

Only humans are natural hosts for the poliovirus. Little is known about the human factors determining susceptibility for acute polio infection, but it seems that persons with Fcγ receptor IIIA V/V genotype are less susceptible for paralytic polio. Coexisting skeletal muscle injury of other causes could also be a factor, probably by introduction of retrograde axonal transport of the poliovirus.

Recovery

A recovery with improvement of muscle strength takes place after the acute illness; it is due to decrease of the inflammation which may temporarily disable infected but surviving motor neurons. In addition, there is an extensive axonal sprouting and reinnervation by surviving motor neurons and an increase in muscle fiber size among the muscle fibers with existing innervation. The degree of residual weakness may vary considerably due to the extent of neuronal damage, from complete to partial persistent weakness. Normal strength and muscle volume can be found despite a loss of 50% of the motor neurons within one segment, but subclinical signs of motor neuron affection with increased amplitudes on electromyographic (EMG) recordings can be found, confirming the motor neuron affection. Most of the recovery takes place within the first 6 months, but further improvement can be seen up to 2 years after the acute infection. The recovery phase is followed by a stable phase of neuromuscular function which can last for several decades.

Post-Polio Syndrome

Definition and Clinical Description

The first description of PPS dates back to 1875, when Raymond and Charcot reported new muscular weakness and atrophy in a 19-year-old tanner, following "*paralysie essentielle de l'enfance*." This man had marked new atrophy in his shoulder with no signs of other neurological disorder. The subject was not investigated systematically in the next decades, but anecdotal reports raised the possibility that progressive motor neuron diseases such as amyotrophic lateral sclerosis (ALS) could be a late complication following paralytic poliomyelitis. In the 1980s, three decades after the large polio epidemics in the Western world, questions about late complications after paralytic polio were raised again, and PPS was defined and implemented.

Dalakas introduced the term post-poliomyelitis atrophy in 1984 to describe patients with a confirmed history of paralytic polio, partial recovery of function, and at least 10 years with functional stability, signs of residual neurologic damage on neurologic examination characterized by muscular atrophy, weakness, areflexia, and normal sensation in groups of muscle with random distribution, and experience of new muscular symptoms. The term PPS was first used by Halstead in 1985 to cover medical, orthopedic, and psychological problems possibly or indirectly related to the long-term post-polio disability. He redefined and narrowed the description and definition in 1991:

1. Confirmed history of polio
2. Partial or fairly complete neurological and functional recovery after the acute episode
3. Period of at least 15 years with neurological and functional stability
4. Two or more of the following health problems occurring after a stable period: extensive fatigue, muscle and/or joint pain, new weakness in muscles previously affected or unaffected, new muscle atrophy, functional loss, cold intolerance
5. No other medical explanation found
6. Gradual or abrupt onset of new neurogenic weakness

Clinical Features

Patients from all parts of the world report the same symptoms of PPS; muscle weakness, new atrophy, postexercise fatigue, muscle pain, generalized fatigue, muscle cramps, fasciculations, cold intolerance, and joint pain tend to dominate. The new weakness or atrophy may occur in previously affected muscles, but also in muscles with previous subclinical motor neuron affection. The typical clinical pattern is asymmetrical and often scattered weakness, involving several segments of the spinal cord, without any signs of upper motor neuron involvement, and no rapid and severe progressive deterioration. Tendon reflexes are often weakened or absent in the same scattered pattern. Fasciculations can be observed in the affected muscles, but are not generalized. Postexercise fatigue and decreased muscular endurance during activity are frequent and may be present before new weakness or atrophy is detected on examination. This is probably due to changes in the aerobic capacity of the muscle fibers, which are often enlarged with a reduced glycolytic activity. The other symptoms are less specific and may be caused by several other disorders. Muscle pain may be present both in weak muscles and in other healthy muscles as a result of overuse due to compensatory movements. Generalized fatigue is a subjective sense of tiredness, lack of energy, and reduced endurance. It is a subjective symptom which can be caused by several other coexisting disorders which must be ruled out before attributing it to PPS.

The prevalence of PPS varies; it has been reported as affecting from 20% to 80% of previous polio patients, mostly dependent on what population was studied and what criteria applied. Today, we estimate that between 20% and 50% of polio survivors will develop PPS. Risk factors are, first of all, the severity of the acute polio, with patients with severe acute polio at greater risk for PPS. A great recovery with assumptive normal muscle function and strength also increases the risk for later PPS. Patients who acquired acute polio as adolescents or adults seem to be at greater risk than patients who had acute polio in childhood. Frequent muscle activity up to the limit of fatigue and exhaustion is another risk factor.

The diagnosis of PPS is an exclusion diagnosis; the role of the investigation is to rule out every possible cause for the new symptoms and clinical deterioration. As PPS occurs in an aging population, the frequency of other coexisting and non-polio disorders is increasing, indicating that a proper diagnostic workup is needed.

PPS patients may be more susceptible to disorders of the spine if there are weak paravertebral muscles and kyphosis or scoliosis. This could lead to other neurological problems such as radiculopathies and myelopathies, and reduced ventilatory capacity. Degenerative spine disorders without neurological complications may also occur, and can lead to pain and physical inactivity. Compensatory use of presumably healthy muscles may lead to muscle and joint pain, tendonitis, and other overuse symptoms. Degenerative joint disorders are common, particularly in weight-bearing

joints and paretic extremities. Weak muscles with following physical inactivity increase the risk of osteoporosis, and fractures in paretic limbs are not unusual. Use of walking aids such as crutches or splints and braces could increase the risk of peripheral nerve entrapment.

Pathogenesis

Despite large efforts to reveal the exact pathogenesis behind PPS, there is no single explanation which covers the clinical picture seen in PPS. Several neuropathological changes have been described, raising further questions.

Denervation and compensatory mechanisms The deterioration of the neuromuscular function among PPS patients is slowly progressive and strictly focal with scattered involvement of muscle fibers within one motor unit, and there is no generalized motor neuron disorder. Dropout of motor units does occur with a scattered pattern. In PPS, there is an imbalance between degeneration and regeneration of enlarged motor units, causing ongoing denervation, and new muscle weakness occur when the motor neuron is unable to maintain all the nerve terminals and a critical number of muscle fibers become denervated. Longitudinal studies have demonstrated that one motor unit can increase in size up to 20-fold compared to normal size before this imbalance becomes clinically evident with new weakness. Excessive use of the remaining motor units seems to contribute to the increasing size of the remaining motor units.

Healthy individuals lose 1% of their muscle strength after the third decade as a part of the normal aging process. After the age of 60, there is also a decrease in the number of motor units in the spinal cord, where the rate of loss varies between muscles. This normal aging process takes place in post-polio patients and contributes to the muscle weakness, but cannot explain the more pronounced process in PPS.

Muscle fibers Skeletal muscles consist of two types of muscle fibers. Type I fibers ('slow twitch') are dense with capillaries, myoglobin, and mitochondria and can perform aerobic activity for longer periods. Type II fibers are mainly involved in 'fast-twitch' operations, with anaerobic metabolism. Muscle biopsies from PPS patients reveal chronic and acute denervation and acute reinnervation, variable muscle fiber size, and scattered atrophic fibers, but no fiber group atrophy. It has been shown that fiber type I dominates in previous polio muscles, and these fibers have a decreased capillary density and reduced oxidative and reduced glycolytic enzyme activities. This muscle fiber transformation from type II to type I seems to depend on the functional demand on the muscle. In patients with excessive use of motor units (i.e., weight-bearing muscles in the legs), the dominant features are predominance of hypertrophic type I fibers and scattered angulated and rounded atrophic fibers, fiber splitting, internal nuclei, and rimmed vacuoles. If there is less use of the remaining motor units, the frequency of hypertrophic type I fibers is lower, and there is widespread fibrosis and atrophic fibers with vesicular central nuclei, pyknotic central nuclei, vacuoles, and cytoplasmatic bodies, but also normal fibers and hypertrophic fibers. The combination of almost total type I predominance, hypertrophic fibers, and reduced oxidative capacity seems to be unique for polio muscles.

The degree of morphologic changes depends on how extensively the muscle was affected in the acute phase, whether partial or full recovery was obtained, and the functional demand. Increased lymphocytic infiltration in perivascular and perimyseal areas has been found in certain muscle specimens but is not consistent, and there is no evidence of generalized inflammation beyond the muscle membrane in PPS.

Inflammation in cerebrospinal fluid and spinal cord It has been questioned whether a reactivation of the polio virus could be present in PPS. There are no cases of recurrent acute polio, and as the poliovirus has a rapid cytolytic cycle, it seems unlikely that the virus escapes immunologic surveillance and is harbored in motor neurons. In addition, no specific immune response against the poliovirus in the cerebrospinal fluid (CSF) in PPS patients has been confirmed. *In vitro* studies have shown that poliovirus can induce persistent infection in human neuroblastoma cells, and persistent poliovirus has also been found in mouse models, but when examining CSF from PPS patients, no complete match of the amplified sequences has been found, and there is no evidence for presence of the virus in studies employing *in situ* hybridization. There is no confirmed evidence of reactivation of the poliovirus as a cause of PPS, but it is possible that mutant virus particles are present in neural tissue without clinical significance.

The issue of a general immunologic activation has been studied, and oligoclonal IgG bands and antineurofilament antibodies in the CSF have been found in some PPS patients, but the clinical significance is unknown. An alteration in T cells in peripheral blood (CD4+ subsets) has also been demonstrated, indicating a possible immune activation.

Cytokines are proteins which can act as chemical messengers, particularly in the immune response. Inflammatory cytokines such as tumor necrosis factor-alpha

(TNF-α) are mainly produced by macrophages and play an important role in inflammation. Although their functions are not fully elucidated, it is presumed that increased volumes of TNF-α and other inflammatory cytokines such as interferon-gamma (IFN-γ) are detrimental to motor neurons and other nerve cells. Anti-inflammatory cytokines such as interleukin-10 (IL-10) are expressed on monocytes and type T2 helper cells and inhibit the actions of natural killer cells (NK cells), thereby reducing the stress response. Recent studies show that, when examining CSF and peripheral blood in PPS patients, an increased intrathecal expression of mRNA for the inflammatory cytokines TNF-α, IFN-γ, interleukin-4 (IL-4), and IL-10 compared to normal controls is found. Increased TNF-α is also seen when analyzing CSF with enzyme-linked immunosorbent assay (ELISA) tests. It has been shown that intravenous immunoglobulin reduces the expression of the proinflammatory cytokines TNF-α and IFN-γ, whereas the levels of the anti-inflammatory cytokine IL-10 remained unchanged in PPS patients. From this point of view, there is evidence of a low-grade immune response in PPS, but it is not established whether this is due to a defect immune system or is secondary to neurodegeneration, autoimmunity, or an unspecific immune response.

Management

Diagnostic assessment The first priority is to establish the diagnosis and exclude all other possible explanations for the new symptoms. There is no test or clinical sign which can confirm the diagnosis of PPS alone, and a proper diagnostic workup is mandatory, particularly if the symptoms are less specific, such as generalized fatigue, muscle pain, muscle cramps, and cold intolerance. A neurophysiological examination is useful in several aspects. At first, it is important to establish typical motor neuron involvement. This is demonstrated with neurogenic EMG findings and normal sensory findings on nerve conduction studies. Other disorders, such as more progressive motor neuron diseases such as ALS or nerve entrapment and radiculopathies, can also be detected on neurophysiological examination and should be ruled out if clinically suspected. A macro EMG examination can be useful to assess the degree of motor neuron loss, as a normal clinical picture can be found even with a loss of up to 50% of motor neurons.

Supportive therapy There is no effective specific treatment available for PPS, and the cornerstone in management of PPS is supportive therapy. Well-conducted exercises prescribed and supervised by experienced therapists can prevent further decline in muscle weakness and reduce symptoms of muscular fatigue, muscle weakness, and pain in selected PPS patients. Both isokinetic and isometric exercises can be performed, but overuse and fatiguing activities should be avoided. It is important to organize the exercises with intermittent breaks and submaximal work loads, and endurance training can be more effective than strengthening exercises. Training of severe weakened muscles should probably be avoided. If symptoms of excessive fatigue and new weakness occur during the training period, the exercises should be stopped and reevaluated. Supervision and advice from experienced therapists are usually necessary in the beginning. The long-term effects of such training have not been explored, but from a general health perspective it is advisable to perform some form of aerobic training to improve cardiovascular fitness and avoid weight gain, which could lead to worsening of the general condition and add a greater workload to already weak muscles. Training in warm water and in a warm climate is useful for amelioration of several PPS symptoms, particularly if non-swimming exercises are performed.

Reduced respiratory function because of weakness in respiratory muscles or severe kyphosis and/or scoliosis is well treated with noninvasive respiratory aids such as CPAP (continuous positive airway pressure) or BiPAP (bilevel positive airway pressure), both to prevent further decline and to treat an existing underventilation, especially if this occurs during sleep. Training of respiratory muscles can also improve pulmonary function, and general precautions such as stopping smoking, mobilization of secretions, and cough assistance are beneficial. If there is a coexisting obstructive sleep apnea, weight control and avoidance of obesity are important.

Lack of energy and general fatigue can be ameliorated with lifestyle modifications such as the introduction of assistive devices such as proper orthoses, walking sticks, and intermittent or permanent use of a wheelchair. Reorganization of daily activities with frequent rest periods, energy conserving, and work simplification is generally recommended.

If there are other coexisting orthopedic and neurological disorders like degenerative joint and spine disease, they should be treated following general guidelines for the specific disorder. Properly fitted orthoses and calipers are useful to prevent disuse and secondary joint degeneration or nerve entrapments. Surgery should be considered if the orthopedic disorder interferes with activities of daily living. If surgery is indicated, a longer recovery and rehabilitation period may be needed.

Pain in muscles and joints can be reduced by decreasing mechanical stress on joints and muscles with lifestyle changes, use of assistive devices, and

aquatic exercise training. If there is bulbar weakness with dysphagia, instruction on swallowing techniques by speech therapists can be of help.

Patients with PPS have to cope with new loss of function, and sometimes an increase in disability and handicap. This new life situation can cause emotional stress and reduced well-being. To help patients come to terms with their new situation, it is important to reassure them that PPS is not a life-threatening disorder with a rapid progression, like ALS. Regular follow-up in PPS clinics, patient education programs, and group training with other PPS patients can give a more positive experience of the 'self' and prevent a decline in mental status.

Pharmaceutical interventions Several studies on pharmaceutical interventions in PPS with regard to muscle strength, fatigue, pain, and quality of life have been performed. Pyridostigmine (an acetylcholinesterase inhibitor) was found to be effective for fatigue in two pilot studies, but this was not confirmed in larger controlled studies, and there is no convincing evidence that pyridostigmine is helpful in PPS. High doses of prednisolone have been tried to improve muscle strength and reduce fatigue, but no significant effect was seen. In order to reduce fatigue, amantadine has been tried with negative results.

It is presumed that inflammatory cytokines damage motor neurons. From this perspective, intravenous immunoglobulin (IvIg) could be an alternative in PPS, as it reduces the mRNA expression of TNF-α and IFN-γ. Recent controlled studies from Sweden showed that an IvIg-treated group of PPS patients had better performances on muscle strength and a higher vitality score after treatment. One uncontrolled Swedish study showed improvement on vitality when measured on the Quality of Life registration SF-36 after IvIg treatment. A small, controlled Norwegian study found no effect on muscle strength or fatigue after IvIg, but pain improved significantly on two measurements. Further studies on patient subgroups, doses, treatment intervals, and long-term effects are needed before any conclusion on IvIg as a therapeutic option in PPS can be drawn.

See also: Inflammation in Neurodegenerative Disease and Injury; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines.

Further Reading

Borg K and Henriksson J (1991) Prior poliomyelitis-reduced capillary supply and metabolic enzyme content in hypertrophic slow-twitch (type I) muscle fibres. *Journal of Neurology, Neurosurgery & Psychiatry* 54(3): 236–240.

Farbu E, Gilhus NE, Barnes MP, et al. (2006) EFNS guideline on diagnosis and management of post-polio syndrome. Report of an EFNS task force. *European Journal of Neurology* 13(8): 795–801.

Farbu E, Rekand T, Vik-Mo E, Lygren H, Gilhus NE, and Aarli JA (2007) Post-polio syndrome patients treated with intravenous immunoglobulin: A double-blinded randomized controlled pilot study. *European Journal of Neurology* 14(1): 60–65.

Gonzalez H, Khademi M, Andersson M, Piehl F, Wallstrom E, Borg K, and Olsson T (2004) Prior poliomyelitis-IVIg treatment reduces proinflammatory cytokine production. *Journal of Neuroimmunology* 150(1–2): 139–144.

Gonzalez H, Khademi M, Andersson M, Wallstrom E, Borg K, and Olsson T (2002) Prior poliomyelitis-evidence of cytokine production in the central nervous system. *Journal of Neurological Sciences* 205(1): 9–13.

Gonzalez H, Sunnerhagen KS, Sjoberg I, Kaponides G, Olsson T, and Borg K (2006) Intravenous immunoglobulin for post-polio syndrome: A randomised controlled trial. *Lancet Neurology* 5(6): 493–500.

Grimby G, Stalberg E, Sandberg A, and Sunnerhagen KS (1998) An 8-year longitudinal study of muscle strength, muscle fiber size, and dynamic electromyogram in individuals with late polio. *Muscle & Nerve* 21(11): 1428–1437.

Halstead LS (1991) Assessment and differential diagnosis for post-polio syndrome. *Orthopedics* 14(11): 1209–1217.

Halstead LS and Rossi CD (1985) New problems in old polio patients: Results of a survey of 539 polio survivors. *Orthopedics* 8(7): 845–850.

Howard RS (2005) Poliomyelitis and the postpolio syndrome. *British Medical Journal* 330(7503): 1314–1318.

March of Dimes (2001) *March of Dimes International Conference on Post-Polio Syndrome: Identifying Best Practices in Diagnosis and Care*. March of Dimes Birth Defect Foundation. http://www.marchofdimes.com/files/PPSreport.pdf (accessed Aug. 2007).

Relevant Website

http://www.polioeradication.org – Global Polio Eradication Initiative.

Postsynaptic Density/Architecture at Excitatory Synapses

H-C Kornau, Center for Molecular Neurobiology (ZMNH), University of Hamburg, Hamburg, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Excitatory chemical synapses of the central nervous system are highly specialized structures which rapidly convey information from one neuron to the other by means of a soluble neurotransmitter, in most cases glutamate. The neurotransmitter is released from the presynaptic terminal upon action potential-evoked calcium influx, diffuses through the synaptic cleft, and opens ligand-gated ion channels on the surface of the postsynaptic cell, resulting in excitatory postsynaptic currents. Signal reception and processing is achieved in dendrites, which are often contacted by presynaptic terminals of thousands of neurons. Excitatory synapses are mostly located on spines, little mushroom-shaped structures protruding from the dendritic shaft. Dendritic spines form a compartment for processing of the individual synaptic input.

In electron micrographs the presynaptic side is recognized by a high density of neurotransmitter-containing vesicles. On the postsynaptic side of the cleft a characteristic electron-dense thickening of the membrane, extending approximately 30 nm into the spine cytosol, is observed (Figure 1); this thickening, a structure that has attracted the interest of neurobiologists since the 1950s, is called the postsynaptic density (PSD). Given that synapses are crucial for information processing in the brain, the importance of the PSD, as the site where neurotransmitter receptors are connected to intracellular structures and where synaptic processing within each spine is initiated, is obvious. Molecular biology and biochemical approaches have identified many proteins of the PSD, including (1) adhesion molecules, receptors, and channels, (2) scaffolding, adaptor, and cytoskeletal proteins, and (3) signaling/regulatory proteins, including kinases, phosphatases, small GTPases, and similar molecules. Moreover, the studies revealed how these proteins physically interact with each other or with the postsynaptic membrane. Thus, we can draw preliminary models containing some of the major components of the PSD to get an idea of the way this macromolecular protein complex may function and how it is rearranged during development and in processes of synaptic plasticity.

How Postsynaptic Densities Are Studied

Electron microscopy initially identified PSDs as hallmarks of glutamatergic/Gray type I synapses in the central nervous system. PSDs are 25- to 50-nm-thick protein assemblies with a diameter of several hundred nanometers. High-resolution electron microscopy reveals a structured 5-nm-thick sheet, which might form the nucleus of the PSD, as well as filamentous structures which connect it to cytoskeletal elements within the cytoplasm of the dendritic spine and to organelles such as the smooth endoplasmic reticulum at the margin of the PSD. Due to their unique density and their resistance to nonionic detergents, PSDs can be purified as intact, disk-shaped structures by differential centrifugation protocols. These preparations serve as the protein sources for the molecular analysis of PSDs. Initially, individual PSD proteins were discovered by amino acid sequencing of protein bands of such PSD preparations after separation by gel electrophoresis. Alternatively, antibodies were raised against complex PSD preparations and used to screen cDNA expression libraries. Coimmunoprecipitates of known PSD components have been a source for finding new elements of this structure as well. Yeast two-hybrid screening, a genetic screen for interacting proteins, supplied a number of proteins physically binding to glutamate receptors or scaffolding proteins of the PSD. Recently, proteomics, which is generally regarded as a large-scale analysis of proteins in a given sample, involving efficient protein separation and highly sensitive protein identification techniques, has been applied to glutamate receptor complexes and to PSD preparations. The results suggest that the PSD contains several hundred different polypeptides. Mass spectrometry techniques have not only identified these proteins with great sensitivity, they have also allowed quantitative determinations. Thus, for some of the components, we know their stoichiometries. Using scanning transmission electron microscopy on a rat forebrain PSD fraction, the mass of an average PSD was measured and the number of copies of specific PSD proteins derived from their relative mass contributions was determined by quantitative gel electrophoresis (Table 1). Consistent with an independent approach using a green fluorescent protein-based calibration technique, these data suggest that a single PSD contains several hundred copies of each of its major polypeptides.

The Protein Composition of Postsynaptic Densities

Glutamate Receptor Complexes

Glutamate receptors and associated proteins were among the first components identified to be part of

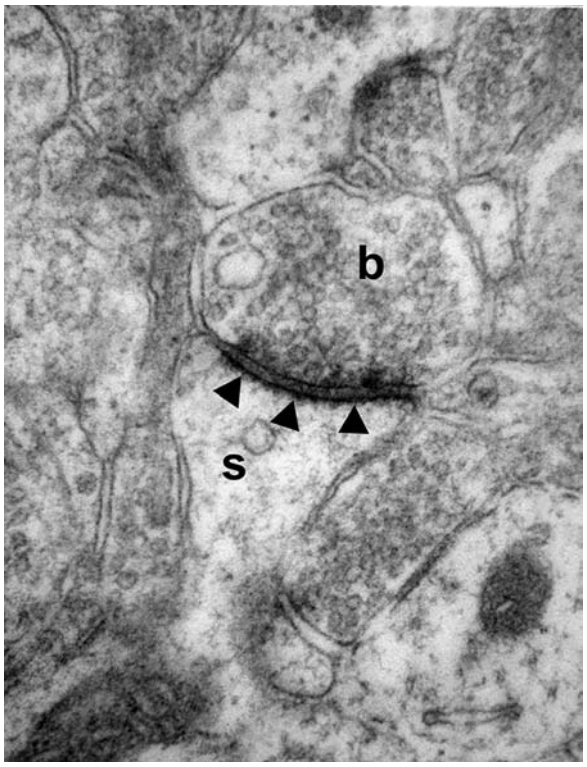

Figure 1 The PSD, a structural element of excitatory synapses. The electron micrograph shows a representative excitatory synapse (Gray type I) at a synaptic spine (s) of an adult mouse hippocampus. Major synaptic structures are the presynaptic bouton (b) filled with synaptic vesicles (diameter ~50 nm), the PSD (arrowheads), and the synaptic cleft in between. The electron micrograph was a kind gift of Dr. Michaela Schweizer, Central Service Facility of Morphology at the ZMNH, University of Hamburg, Hamburg, Germany.

Table 1 Dimensions of a PSD

Mean diameter[a]	~360 nm
Depth[b]	~25–50 nm
Mean mass[a]	~1.1 GDa
Number of PSD-95 molecules[a]	~300

[a] Data from Chen X, Vinade L, Leapman RD, et al. (2005) Mass of the postsynaptic density and enumeration of three key molecules. *Proceedings of the National Academy of Sciences of the United States of America* 102: 11551–11556.
[b] Data from Valtschanoff JG and Weinberg RJ (2001) Laminar organization of the NMDA receptor complex within the postsynaptic density. *Journal of Neuroscience* 21: 1211–1217.

the PSD. Glutamate receptor channels mediate the principal postsynaptic functions – that is, changes in membrane potential upon binding of the excitatory neurotransmitter. The different types of ionotropic glutamate receptors are named for their specific agonists. α-Amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor channels are opened upon glutamate binding alone, whereas *N*-methyl-D-aspartate (NMDA) receptor activation requires simultaneous membrane depolarization (e.g., by activation of neighboring synapses). This depolarization relieves the extracellular magnesium block of the NMDA channel, allowing influx of sodium and, importantly, calcium through the channel pore. Calcium acts as a second messenger within the cell and is important for the induction of synaptic plasticity. A very abundant PSD protein involved in these processes and activated by calcium is calcium/calmodulin-dependent protein kinase II (CaMKII). It is associated with the cytoplasmic region of the NMDA receptor, allowing it to sense the influx of calcium through this channel directly. Subsequent pathways involve a variety of structural rearrangements within the PSD and lead to a change in synaptic efficacy primarily expressed as an altered number of AMPA receptors in the postsynaptic membrane. Both native AMPA and NMDA receptor complexes are associated with a variety of scaffolding/adaptor proteins (e.g., postsynaptic density protein 95), which link the ionotropic glutamate receptors to other integral membrane proteins, to soluble signaling molecules, and to the spine cytoskeleton, and form the core lattice of the PSD. At the margin of the PSD, metabotropic G-protein-coupled receptors for glutamate (mGluRs) modulate the intraspinal calcium level. Proteomics has unraveled individual protein complexes for NMDA receptors, mGluRs, and AMPA receptors.

The NMDA receptor complex is tightly linked to the core of the PSD by a series of adaptor proteins comprising postsynaptic density protein 95 family members (PSD-95; also called synapse-associated protein 90 (SAP90), PSD-93/Chapsyn-110, SAP102, and SAP97), guanylate kinase-associated proteins (GKAPs; also called SAP90/PSD-95-associated proteins, or SAPAPs), and Shanks (also called proline-rich synapse-associated proteins, or ProSAPs) (Figure 2). PSD-95 and its relatives encompass three PDZ domains, an SH3 domain, and a guanylate kinase domain, all of which function as protein–protein interaction domains. They belong to the family of membrane-associated guanylate kinases (MAGUKs), which function primarily at sites of cell-to-cell contact. Employing its PDZ domains, PSD-95 links the NMDA receptor to neuroligins, postsynaptic cell adhesion molecules crucial for synapse formation, and to several signaling molecules, among them neuronal nitric oxide synthase (nNOS) and synaptic GTPase-activating protein (synGAP), an abundant regulator of Ras and Rap GTPases. The GKAPs connect PSD-95 to the Shank

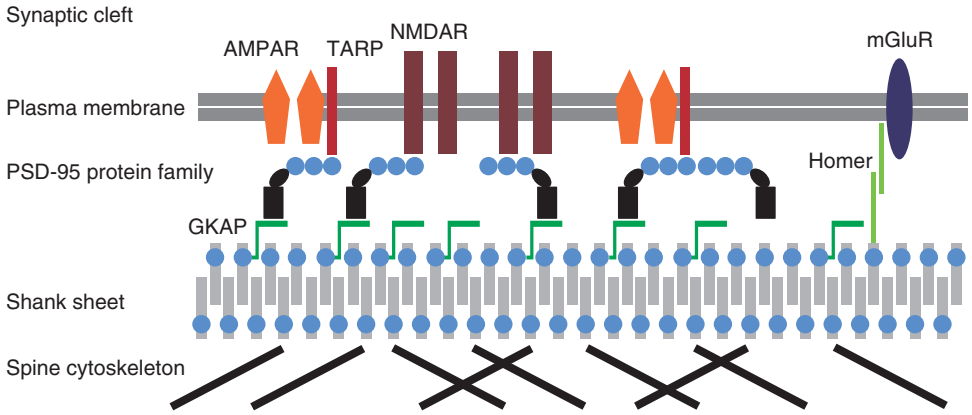

Figure 2 Simplified scheme of the molecular architecture of the PSD. The different types of glutamate receptors are linked to the Shank scaffold, a two-dimensional sheet, either by the PSD-95 family of proteins and by guanylate kinase-associated protein (GKAP) molecules, or by Homer dimers. The Shank scaffold connects the PSD with the spine actin cytoskeleton. The PSD-95 family proteins form a membrane-proximal layer of scaffolds, whereas the Shanks constitute a second, deeper layer. PDZ domains are depicted as blue circles. The C-terminal domains of the glutamate receptor subunits, the PSD-95 family members, as well as the Shanks, recruit numerous signaling molecules to the PSD. The PSD-95 family members connect the PSD with cell adhesion molecules governing the integrity with the presynaptic specialization. The Homer proteins link the metabotropic glutamate receptors (mGluRs) not only to the Shank scaffold, but also to the inositol 1,4,5-trisphosphate receptor within the smooth endoplasmic reticulum. For simplicity, all of these connections were omitted from the scheme. The figure does not reflect the real stoichiometry of the scaffolding proteins. AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; TARP, transmembrane AMPA receptor regulating protein; NMDAR, *N*-methyl-D-aspartate receptor.

core of the PSD by binding to both the guanylate kinase domain of PSD-95, on one hand, and to the PDZ domain of a Shank molecule, on the other hand. Four closely related GKAP family members show distinct spatiotemporal expression patterns in the brain. Shanks (three related proteins) contain a number of protein interaction domains, including ankyrin repeats, an SH3 domain, a PDZ domain, a proline-rich region, and a sterile alpha allowing them to bind to a variety of proteins. Moreover, by self-association of their SAM domains they are able to form large sheets within the PSD. The resulting dense platform is associated with the actin cytoskeleton, the smooth endoplasmic reticulum, the different glutamate receptor complexes, and other channels and receptors. Shanks are also early components during synaptogenesis.

The three classes of adaptor proteins play a crucial role within the PSD. They form an axis of scaffolds differing in their distance from the plasma membrane (Figure 2): PSD-95 is located approximately 12 nm cytoplasmic to the extracellular face of the postsynaptic membrane, like the C-termini of the NMDA receptor, whereas GKAP and Shank molecules are buried deeper in the cytoplasm (approximately 24–26 nm from the membrane). They recruit a number of signaling molecules to the PSD and connect the PSD with cytoskeletal elements. Proteomic analysis of the NMDA receptor complex has revealed a complexity of more than 100 proteins, including a variety of protein kinases and phosphatases, small G-proteins and their regulators, and other signaling molecules, such as nNOS, phosphatidylinositol 3-kinase (PI3K), phospholipase C-γ, Citron, and Arg3.1/Arc. In addition, several cell adhesion molecules and associated adaptors, including N-cadherins and β-catenin, which regulate dendritic spine morphogenesis, were found in the NMDA receptor complex. Finally, many cytoskeletal proteins are part of this complex.

Type I metabotropic glutamate receptors are located at the margin of the PSD. These G-protein-coupled receptors lead to inositol 1,4,5-trisphosphate (IP_3) production, which consequently activates the IP_3 receptor and induces the release of calcium from intracellular stores. Key molecules of the mGluR protein complex are the Homer proteins, which are expressed from three genes. They can dimerize by a coiled-coil interaction and these dimers connect mGluRs to the Shank backbone of the PSD on the plasma membrane side, and IP_3 receptors in the smooth endoplasmic reticulum to the Shank backbone on the cytoplasmic side, of the PSD. Thus, Homer proteins connect mGluRs and IP_3 receptors to the PSD and link mGluRs and IP_3 receptors, allowing efficient intracellular calcium release upon mGluR activation. In these assemblies the N-terminal EVH domains of the Homer dimers bind to a PPXXF motif in the receptors and to a proline-rich region in Shanks (Figure 2). (The EVH nomenclature derives from 'enabled-VASP (vasodilator-stimulated phosphoprotein) homology.') Interestingly, Homer 1a, an

isoform lacking the coiled-coil domain, which is highly expressed as an immediate-early gene, can compete with the constitutive forms of Homer and thus with the link of mGluRs and IP_3 receptors to the PSD. This may enable the cell to limit the release of intracellular calcium if required. The mGluR5 complex has been also subject to proteomic analysis. It provided a number of additional candidate proteins of mGluR complexes and confirmed the link to Homer, Shank, and IP_3 receptors delineated herein.

AMPA receptors mediate the majority of the fast excitatory neurotransmission in the brain. The efficacy of a synaptic connection is related to the number of AMPA receptors in the postsynaptic membrane. Thus, whereas NMDA receptors are mediating the induction of synaptic plasticity, AMPA receptors are crucial for its expression. The number of AMPA receptors in the postsynaptic membrane is indeed quite variable, ranging from zero to several hundred receptors per synapse, and can rapidly change upon usage of a synapse. AMPA receptors are rapidly cycling between plasma membrane and endosomal compartments. Thus, they need a unique set of associated proteins.

The best documented interaction of AMPA receptors is their binding to the integral membrane protein stargazin. Stargazin was the first member of a family of four differentially expressed TARPs (for transmembrane AMPA receptor regulatory proteins), which function as auxiliary subunits of AMPA receptors. By two independent interaction sites, TARPs regulate both the trafficking and the gating of AMPA receptors. A C-terminal interaction of TARPs with PSD-95 connects AMPA receptors with the PSD and governs the synaptic localization of AMPA receptors (Figure 2).

The rapid cycling of AMPA receptors requires the interaction with AP2 and clathrin. These and *N*-ethylmaleimide-sensitive fusion protein NSF, a multimeric ATPase playing an important role in membrane fusion, bind to specific motifs in the center of the short, intracellular C-termini of selected AMPA receptor subunits. Additional proteins interact with the very C-terminal amino acids of AMPA receptor subunits and affect the synaptic localization of AMPA receptors. These include an adaptor protein termed glutamate receptor-interacting protein (GRIP) and a protein originally identified as a protein interacting with C kinase, PICK1. Both of these proteins bind to the C-termini of AMPA subunits by PDZ domains. GRIPs (two related proteins) harbor seven PDZ domains and thereby can connect AMPA receptors to several cytoplasmic and membrane-standing proteins (e.g., the ephrin receptor). The PICK1 interaction with AMPA subunits is essential for the AMPA receptor endocytosis involved in cerebellar long-term depression.

Proteomic approaches to tackle the composition of AMPA receptor complexes identified a limited complexity as compared to NMDA receptor or mGluR complexes. However, AMPA receptor cycling between plasma membrane and endosomal locations may specifically require transient or weak interactions, which may be lost during purification of the receptor complex.

Multiple interfaces between the individual glutamate receptor multiprotein complexes appear to exist: Homer (mGluR complex) can bind to Shanks in addition to mGluRs and IP_3 receptors, and stargazin (AMPA receptor complex) binds to PSD-95 (NMDA receptor complex). Proteomic studies of PSD proteins have been performed on immunopurifications of glutamate receptors (brain protein complexes purified via peptides derived from the C-terminal sequence of the NMDA receptor subunit 2B, a target of PSD-95 family members) and on purifications of entire PSDs. The data from these different approaches have been compiled as our current view of the PSD composition. They show that in addition to the glutamate receptor complexes connected to the different scaffolding proteins we have to integrate a large number of additional integral membrane proteins, signaling/regulatory proteins, and cytoskeletal elements into our model of the PSD. Although the number of proteins identified in PSD fractions goes into the hundreds, only a few principal types of molecule classes can be covered here.

Integral Membrane Proteins

Independent analyses of the PSD proteome revealed a variety of integral membrane proteins aside from glutamate receptors. These include not only other channels and receptors, but also cell adhesion molecules (e.g., neuroligins and N-cadherin) which directly link the PSD with the juxtaposed membrane specialization on the presynaptic side termed the active zone. Cell adhesion molecules govern several steps of synaptogenesis and may keep the active zone and the PSD in register during structural changes of the synapse.

Regulatory Proteins

The importance of CaMKII as a very abundant PSD protein for the induction of synaptic plasticity has already been mentioned. However, many other types of protein kinases and phosphatases are found in the PSD as well. This is not surprising since phosphorylation and dephosphorylation events often mediate

functional and structural modulation of central nervous system synapses.

Small GTPases of the Ras superfamily, termed Ras and Rap, belong to the most important molecules involved in the integration of biochemical signals in the dendritic spine. Their activity determines the amount of AMPA receptors in the postsynaptic membrane. As GTPases they switch between an active, GTP-bound, and an inactive, GDP-bound, state. Activation requires a specific guanine nucleotide exchange factor (GEF); inactivation requires a GTPase-activating protein (GAP). PSD preparations contain several GEFs and GAPs for Ras and Rap, which themselves differ by the signals required for their activation.

Small GTPases of the Rho family, including Rac and cdc42, directly influence the state of the spine cytoskeleton. For example, adhesion molecule-mediated recruitment of specific Rac GEFs activates Rac and leads to changes in actin polymerization. These GTPases have a strong impact on the morphology of dendritic spines and their regulators are found in PSD preparations.

Cytoskeletal Proteins

The axis of scaffolds in the PSD (PSD-95 family members, GKAPs, and Shanks) is also a nucleus for interactions with the cytoskeleton in the spine. As already indicated in the section on the NMDA receptor complex, PSD preparations contain a variety of cytoskeletal proteins. They include actin, actin-binding proteins, tubulin, spectrin, cortactin, and many others.

The proteomic approaches revealed several hundred different proteins in PSD preparations. However, it should be noted that this does not necessarily reflect the complexity at a single PSD. The protein composition of PSDs in different brain regions as well as in different synapses of a single neuron may considerably differ, and thereby may contribute to the large number of identified proteins.

PDZ Domains

Like all macromolecular protein networks, the PSD is connected by modular protein–protein interactions. Analysis of the interacting protein sequences has revealed the presence of specific domains. A domain frequently found at sites of cell-to-cell contact is the PDZ domain (so named from PSD-95/discs-large/zonula occludens-1), which is present in many different proteins in humans. PDZ domains bind, in most cases, to short peptide motifs at the C-termini of other proteins and are classified according to the sequences they recognize. In addition, some PDZ domains can heterodimerize. The PDZ domain is also central in connecting different components of the PSD (Figure 2). It was first recognized in the prototype PSD scaffolding protein PSD-95 (hence, the name), which contains three N-terminal PDZ domains and uses two of them to bind to the C-termini of NMDA receptor subunits and the third PDZ and other domains to connect the receptor to a variety of PSD components. PDZ domains are essential for the function of many signaling and scaffolding proteins of the PSD. The scaffolding proteins often contain several PDZ as well as other interaction domains, each with a preferred binding partner, allowing the scaffold to coordinate large protein complexes.

Rearrangements of Postsynaptic Densities

Developmental Assembly of Postsynaptic Densities

Most of the synapse formation in the rat central nervous system occurs during the first two postnatal weeks. It is initiated by an interaction between pre- and postsynaptic cell adhesion molecules within the synaptic cleft, including neurexins/neuroligins, cadherins, synCAMs, receptor tyrosine kinases, and others. Among the earliest proteins at synapses are PSD-95 family members and Shanks, suggesting that they play a major role in the assembly of the PSD. Their interactions with the other scaffolding and signaling molecules are thought to promote the formation of the PSD very rapidly. Simultaneously, NMDA receptors are recruited into the PSD, whereas the integration of AMPA receptors into the postsynaptic membrane requires presynaptic glutamate release. Early in development many synapses in the central nervous system lack AMPA receptors and are therefore called silent synapses.

PSDs of postnatal day 2 rats contain already many of the proteins found in adult PSDs. However, during development, specific molecular changes are evident. These include a switch in the expression of NMDA receptor subunits from NR2B to NR2A and, simultaneously, of their interaction partners SAP102 to PSD-95, and an increase of AMPA receptors and CaMKII. Although during synaptogenesis the recruitment of key PSD proteins such as NMDA receptor subunits and Shanks appears to occur in a rather gradual manner, preformed transport packets containing either AMPA or NMDA receptors, or mobile clusters containing PSD-95, GKAP, and Shank, driven by actin transport, have been described. Thus, how the PSD is assembled during synapse formation requires further investigation.

Activity-Dependent Changes at Postsynaptic Densities

The size, protein composition, and structure of the PSD are altered in response to developmental and environmental cues. One cellular model for learning and memory, long-term potentiation (LTP), involves a prolonged strengthening of a synapse in response to a high-frequency input. Calcium influx through NMDA receptors, which are open under these depolarizing conditions, allows activation of a number of signaling molecules of the PSD, including CaMKII, calcineurin, nNOS, and several regulators of small GTPases. The primary result is a change in the phosphorylation pattern of PSD proteins, among them the glutamate receptor subunits. Phosphorylation leads to an altered affinity of binding partners and concomitant stabilization or destabilization of the receptor molecules at the PSD. Activation of small G-proteins can trigger endo- or exocytosis of membrane proteins. As an example, additional AMPA receptor subunits are rapidly inserted into the plasma membrane during LTP, resulting in a direct increase of the postsynaptic response. In contrast, AMPA receptors are removed from the synaptic plasma membrane during long-term depression. Other small GTPases, upon activation, alter the polymerization state of actin in the dendritic spine. Thus, synaptic plasticity involves structural plasticity at the PSD.

In addition to changes in the phosphorylation pattern, receptor localization, and cytoskeleton structure, local translation and targeted protein degradation contribute to the high capability of the PSD to rearrange. Activity-dependent translation of AMPA receptor subunits in dendrites may mediate synapse-specific modifications. The presence of mRNAs for other PSD components (e.g., Shanks) in dendrites suggests their local translation as well. Dendritic protein synthesis may contribute significantly to synaptic plasticity processes.

Posttranslational modifications aside from phosphorylation affect the PSD structure. As an example, PSD-95 family proteins are differentially palmitoylated at two cysteine residues, with important consequences for their clustering ability. Changes in the PSD structure and composition can also result from altered patterns of transcription and splicing in the nucleus. Alternative splicing events affect the interactions of various PSD proteins by adding/removing sequences encoding specific interaction domains or motifs. For example, two forms of PSD-95, a palmitoylated one (see earlier) and one with an N-terminal L27 domain which enables additional interactions of PSD-95, result from alternative splicing. Rearrangements of the PSD may not only result from alterations in the amount or modification of scaffolding proteins, but may also be influenced by their propensity to interact as a result of alterations in the ionic environment. Of interest in this context, it has been shown that the zinc concentration can, at least *in vitro,* directly influence the assembly of Shank fibers, a core element of the PSD.

The Physiological Role of Postsynaptic Densities

What may be the function of this complex protein assembly? Based on its localization and content, the clustering, localization, and regulated trafficking of glutamate receptors, the compartmentalization of different glutamate receptor complexes, the spatial organization of signaling cascades, the dynamic regulation of the cytoskeleton, and the stabilization of the synaptic structure may be some of the roles the PSD plays.

The ordered array of proteins into microdomains of the PSD governs the efficiency of the signal transduction pathways initiated by activation of the different kinds of glutamate receptors. Importantly, the PSD clusters the glutamate receptors at sites directly adjacent to the presynaptic release sites, which is crucial for fast chemical neurotransmission. Its connection to the receptors and cell adhesion molecules and to the cytoskeleton of the dendritic spine allows the PSD to coordinately respond to changes of presynaptic activity with alterations of the structure of the PSD and the morphology of the spine. Numerous links between the amount of PSD scaffolding proteins and the spine density and size as well as the glutamate receptor content of synapses have been experimentally established.

An impaired architecture of the PSD would be expected to affect the balance of excitatory and inhibitory neurotransmission in the central nervous system and, therefore, to cause fatal diseases. However, relatively little is known about the connection between neurological disorders and the dysfunction of the PSD; maybe the high molecular redundancy of scaffolding proteins secures that central synapses will function even if one of the PSD-95 family member, GKAP, or Shank genes is mutated. Nevertheless, the Shank 3 gene is linked to a neurological disorder termed 22q13.3 deletion syndrome, characterized by mental retardation, delayed speech, and dysmorphic features. Mutations and chromosomal rearrangements in loci corresponding to PSD-95 and neuroligin genes are associated with autism. Mental retardation is linked to the genes for several additional PSD components and correlates with an abnormal morphology of dendritic spines and synapses. Therefore, studying the PSD may not only help to understand

mechanisms of synaptic transmission, but may also provide insight into diseases of the central nervous system.

See also: AMPA Receptor Cell Biology/Trafficking; Dendritic Spine History; Glutamate Regulation of Dendritic Spine Form and Function; Glutamate Receptor Organization: Ultrastructural Insights; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Metabotropic Glutamate Receptors (mGluRs): Molecular Biology, Pharmacology and Cell Biology; NMDA Receptors, Cell Biology and Trafficking; Postsynaptic Development: Neuronal Molecular Scaffolds; Postsynaptic Specialization Assembly.

Further Reading

Boeckers TM (2006) The postsynaptic density. *Cell and Tissue Research* 326: 409–422.

Carlin RK, Grab DJ, Cohen RS, et al. (1980) Isolation and characterization of postsynaptic densities from various brain regions: Enrichment of different types of postsynaptic densities. *Journal of Cell Biology* 86: 831–845.

Chen X, Vinade L, Leapman RD, et al. (2005) Mass of the postsynaptic density and enumeration of three key molecules. *Proceedings of the National Academy of Sciences of the United States of America* 102: 11551–11556.

Cheng D, Hoogenraad CC, Rush J, et al. (2006) Relative and absolute quantification of postsynaptic density proteome isolated from rat forebrain and cerebellum. *Molecular & Cellular Proteomics* 5: 1158–1170.

Collins MO, Husi H, Yu L, et al. (2006) Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *Journal of Neurochemistry* 97 (supplement 1): 16–23.

Funke L, Dakoji S, and Bredt DS (2005) Membrane-associated guanylate kinases regulate adhesion and plasticity at cell junctions. *Annual Review of Biochemistry* 74: 219–245.

Gray EG (1959) Axo-somatic and axo-dendritic synapses of the cerebral cortex: An electron microscope study. *Journal of Anatomy* 93: 420–433.

Gundelfinger ED, Boeckers TM, Baron MK, et al. (2006) A role for zinc in postsynaptic density assambly and plasticity? *Trends in Biochemical Sciences* 31: 366–373.

Harris KM, Jensen FE, and Tsao B (1992) Three-dimensional structure of dendritic spines and synapses in rat hippocampus (CA1) at postnatal day 15 and adult ages: Implications for the maturation of synaptic physiology and long-term potentiation. *Journal of Neuroscience* 12: 2685–2705.

Husi H, Ward MA, Choudhary JS, et al. (2000) Proteomic analysis of NMDA receptor–adhesion protein signaling complexes. *Nature Neuroscience* 3: 661–669.

Kennedy MB (2000) Signal-processing machines at the postsynaptic density. *Science* 290: 750–754.

Kennedy MB, Beale HC, Carlisle HJ, et al. (2005) Integration of biochemical signalling in spines. *Nature Reviews Neuroscience* 6: 423–434.

Kim E and Sheng M (2004) PDZ domain proteins of synapses. *Nature Reviews Neuroscience* 5: 771–781.

Kornau HC, Seeburg PH, and Kennedy MB (1997) Interaction of ion channels and receptors with PDZ domain proteins. *Current Opinion in Neurobiology* 7: 368–373.

Nicoll RA, Tomita S, and Bredt DS (2006) Auxiliary subunits assist AMPA-type glutamate receptors. *Science* 311: 1253–1256.

Palay SL (1958) The morphology of synapses in the central nervous system. *Experimental Cell Research* 14: 275–293.

Peng J, Kim MJ, Cheng D, et al. (2004) Semiquantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry. *Journal of Biological Chemistry* 279: 21003–21011.

Scannevin RH and Huganir RL (2000) Postsynaptic organization and regulation of excitatory synapses. *Nature Reviews Neuroscience* 1: 133–141.

Scheiffele P (2003) Cell–cell signaling during synapse formation in the CNS. *Annual Review of Neuroscience* 26: 485–508.

Spacek J and Harris KM (1997) Three-dimensional organization of smooth endoplasmic reticulum in hippocampal CA1 dendrites and dendritic spines of the immature and mature rat. *Journal of Neuroscience* 17: 190–203.

Sugiyama Y, Kawabata I, Sobue K, et al. (2005) Determination of absolute protein numbers in single synapses by a GFP-based calibration technique. *Nature Methods* 2: 677–684.

Valtschanoff JG and Weinberg RJ (2001) Laminar organization of the NMDA receptor complex within the postsynaptic density. *Journal of Neuroscience* 21: 1211–1217.

Ziff EB (1997) Enlightening the postsynaptic density. *Neuron* 19: 1163–1174.

Postsynaptic Development: Neuronal Molecular Scaffolds

E Kim, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuronal synapses generally fall into two groups, excitatory and inhibitory. Excitatory synapses are mostly present in dendritic spines, which are thorn-like structures on dendrites. Inhibitory synapses are located on dendritic shafts and the cell body. Excitatory synapses are characterized by the presence of an electron-dense thickening in the postsynaptic side known as the postsynaptic density (PSD). The PSD is formed through the assembly of macromolecular postsynaptic protein complexes containing receptors, scaffolds, and signaling proteins.

Postsynaptic development at excitatory synapses is thought to involve an initial axodendritic contact, followed by localization of early postsynaptic proteins and recruitment of additional proteins to permit growth into mature postsynaptic structures. During the past decade, mechanisms underling the assembly and molecular organization of excitatory neuronal synapses have been thoroughly studied. Inhibitory neuronal synapses have received relatively less attention, partly because the wealth of proteins in the PSD has attracted scientific interest. This article focuses on postsynaptic differentiation at excitatory synapses. In particular, it discusses how postsynaptic scaffolds contribute to the assembly, organization, and plasticity of the PSD.

Postsynaptic Scaffolding Proteins

Components of the PSD have been identified by various experimental approaches including mass spectrometry and the use of protein interaction traps such as the yeast two-hybrid system. Proteomic approaches have identified hundreds of PSD components, including membrane proteins, scaffolding proteins, signaling proteins, and cytoskeletal proteins. Proteomic studies have further provided information on relative and absolute amounts the PSD proteins and their phosphorylation.

Here, postsynaptic scaffolding proteins are defined as proteins that are relatively more abundant than are other PSD proteins and that possess various domains for protein–protein interactions. The presence of such domains strongly implies that the proteins are implicated in the assembly and organization of the PSD. Many postsynaptic scaffolds contain the PDZ (PSD-95-Dlg-ZO1) domain, a 90-amino-acid-long module that interacts with the C-terminal PDZ-binding motif of other proteins. The PDZ domain is one of the most common protein domains known, being found in approximately 580 proteins encoded by the mouse genome. PDZ proteins exhibit 1–13 tandem arrays of PDZ domains. PDZ domain-binding motifs are found in a wide variety of proteins, including membrane proteins and signaling proteins. PDZ proteins, due to their ability to interact with other proteins and form macromolecular protein complexes, are found mainly in specialized cell-to-cell junctions, including neuronal synapses and tight junctions.

A well-known PDZ-containing postsynaptic scaffold is PSD-95, which is a family of proteins with four known members: PSD-95/SAP90, PSD-93/chapsyn-110, SAP97, and SAP102. PSD-95 family proteins contain various domains for protein–protein interactions, including three PDZ domains, one SH3 domain, and one GK domain. Splice variants of PSD-95, PSD-93, and SAP97 (PSD-95β, PSD-93ζ, and SAP97β) contain an additional domain, L27, at the N-terminus. Through these domains, PSD-95 binds to a wide variety of proteins and is involved in the assembly and molecular organization of the PSD. Proteomic studies have shown that PSD-95 is one of the most abundant proteins in the PSD. The four members of the PSD-95 family seem to differ in function, with PSD-95 and PSD-93 being more important in synapses and SAP97 and SAP102 playing roles in protein trafficking.

Functionally, transient overexpression of PSD-95 in dissociated neurons increases the number and size of dendritic spines and α-amino-3-hydroxy-5-methyl-isoxazole propionic acid (AMPA) glutamate receptor-mediated synaptic transmission. In hippocampal slices, PSD-95 overexpression drives GluR1 AMPA receptors into synapses, occludes long-term potentiation (LTP), and enhances long-term depression (LTD). Conversely, acute PSD-95 knockdown reduces the ratio of excitatory to inhibitory synapses and suppresses AMPA receptor-mediated synaptic transmission. Transgenic mice with truncated PSD-95 exhibit reduced LTD, enhanced LTP, and impaired spatial learning. These knockout mice show unaltered AMPA receptor-mediated currents, in contrast to acute PSD-95 knockdown, probably due to functional compensation by other PSD-95 family proteins. Indeed, double-knockout mice lacking both PSD-95 and PSD-93 show a markedly reduced ratio of AMPA/*N*-methyl-D-aspartate (NMDA) excitatory postsynaptic currents. Together, these results implicate PSD-95 in the

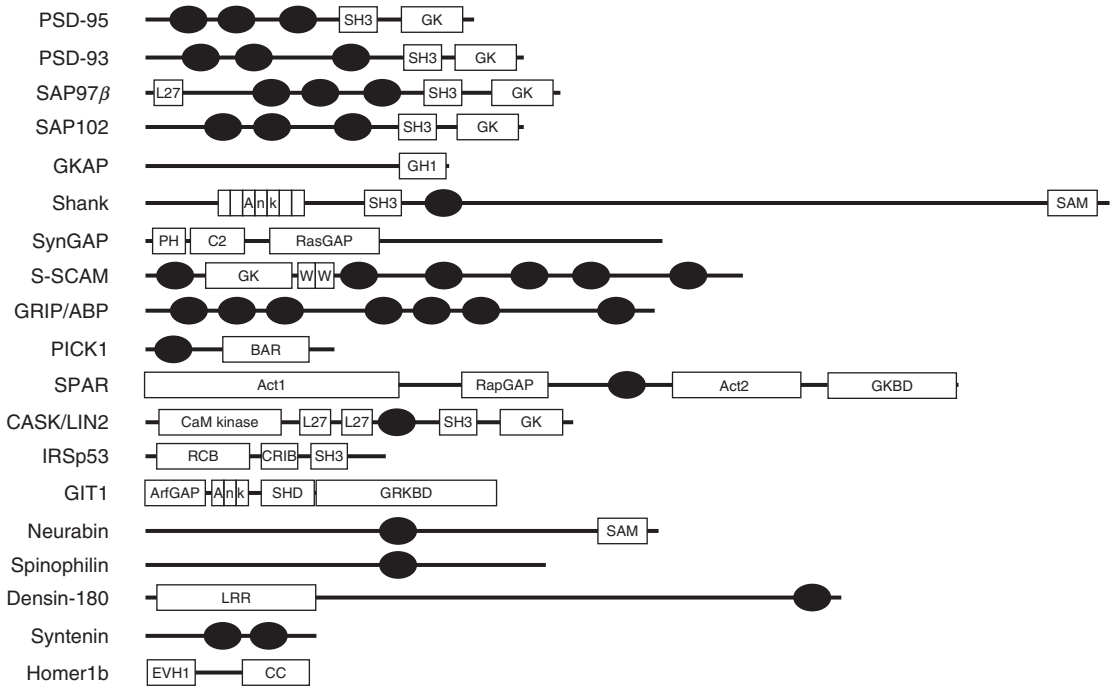

Figure 1 Schematic diagram of PSD proteins. Domain structures of selected PSD proteins. PDZ domains are shown as dark ellipses. Other domains are indicated: Act1, actin regulatory domain 1; Act2, actin regulatory domain 2; Ank, ankyrin repeats; ArfGAP, Arf GTPase-activating protein; BAR, Bin-Amphiphysin-Rvs domain; C2, calcium/lipid binding domain 2; CaM kinase, Ca^{2+}/calmodulin-dependent kinase (CaMK)-like domain; CC; coiled coil domain; CRIB, Cdc42/Rac-interactive binding; EVH1, ENA/VASP homology domain 1; GH1, GKAP homology domain 1; GK, guanylate kinase-like domain; GKBD, PSD-95 GK binding domain; GRKBD, GRK2 binding domain; kinase, serine/threonine kinase domain; L27, domain initially found in LIN2 and LIN7; PH, pleckstrin homology domain; RapGAP, Rap GTPase-activating protein; RasGAP, Ras GTPase-activating protein; RCB, Rac binding domain; SAM, sterile α motif; SH3, Src homology 3 domain; SHD, Spa2 homology domain; WW, domain with two conserved Trp (W) residues. Proteins: CASK/LIN2; vertebrate homolog of lin2; GIT1, GRK-interacting protein 1; GKAP, GK-associated protein; GRIP/ABP, glutamate receptor interacting protein/AMPA receptor binding protein; IRSp53, insulin receptor tyrosine kinase substrate p53; PICK1, protein interacting with C-kinase; PSD-93, postsynaptic density protein 93; PSD-95, postsynaptic density protein 95; SAP97, synapse-associated protein 97; SAP102, synapse-associated protein 102; Shank, SH3 and ankyrin repeat-containing protein; SPAR, spine-associated RapGAP; S-SCAM, synaptic scaffolding molecule.

regulation of excitatory synapses, dendritic spines, synaptic strength and plasticity, and learning and memory.

Other postsynaptic scaffolds implicated in postsynaptic development are listed in Figure 1. A large number of these scaffolds contain PDZ domains, as does PSD-95, suggesting that PDZ-based interaction is a widespread mechanism for postsynaptic assembly and organization.

Principles Governing Postsynaptic Differentiation

An important question in synapse formation concerns which side of the synapse initiates synapse formation. Some published studies indicate that the assembly of functional nerve terminals precedes postsynaptic differentiation. Time-lapse microscopy and retrospective immunostaining on cultured hippocampal neurons (11–14 days *in vitro* (div)) indicate that contact-induced formation of functional nerve terminals is followed by clustering of postsynaptic proteins including PSD-95 and glutamate receptors in approximately 45 min. Early stage hippocampal neurons (5–7 div) contain nonsynaptic mobile packets with postsynaptic scaffolds such as PSD-95, GKAP/SAPAP, and Shank/ProSAP. Some of these packets are fast moving and can be recruited to PSD-95/GKAP-negative nascent synapses apposed to functional nerve terminals, suggesting that the presynaptic side instructs postsynaptic assembly. However, a proportion of these mobile packets move slowly. These packets contain neuroligin-1, and when they contact axons, they induce functional nerve terminals by recruiting mobile synaptophysin clusters. This suggests that a certain degree of postsynaptic assembly also occurs prior to the formation of functional nerve terminals, and that synaptogenesis occurs in a bidirectional manner.

Presynaptic assembly is thought to be mediated by vesicular intermediates including synaptic vesicle (SV)

precursors and specialized dense-core vesicles known as Piccolo/Bassoon transport vesicles (PTVs). These two types of vesicles carry preassembled complexes of SV proteins and active zone components (scaffolds and plasma membrane proteins), respectively. Notably, active zones are assembled from a small number (typically two or three) of PTVs. A related question is whether postsynaptic assembly also occurs in this manner. Early stage (3 or 4 div) cortical neurons exhibit rapidly moving NMDA and AMPA receptor transport packets similar to the mobile packets of early hippocampal neurons (mentioned previously). This suggests that modular transport of postsynaptic proteins may play a role in early synaptogenesis. However, late-stage (8–13 div) hippocampal neurons exhibit a gradual recruitment of postsynaptic proteins, including NMDA receptors, PSD-95, and Shank. Although reasons for this discrepancy remain to be determined, a possible explanation is that neurons at different developmental stages may have distinct mechanisms of postsynaptic protein assembly.

Induction of Postsynaptic Differentiation

When the presynaptic side induces postsynaptic differentiation, what might be the initiating signals? Neurotransmitters, released from nerve terminals, are good candidates. Neurotransmitters may affect the morphology of dendritic spines and filopodia (long thin protrusions on dendrites), suggesting that the neurotransmitters may instruct synapse formation. However, transgenic mice that cannot release neurotransmitters, due to the absence of the presynaptic protein Munc13 or Munc18, exhibit normal synapse formation. In addition, blockade of neurotransmitter release in cultured neurons does not inhibit synapse development. These data suggest that neuronal activity is not required for synapse formation.

Another possible trigger is transsynaptic adhesion between presynaptic and postsynaptic cell adhesion molecules. Ideally, transsynaptic adhesions should be heterophilic to minimize homophilic adhesions between dendrites and axons. In addition, synaptic adhesion molecules need to permit coupling of adhesion events to induction of synaptic differentiation through the recruitment of various synaptic proteins to the sites of contact. A well-known example of heterophilic and synaptogenic transsynaptic adhesion is that between presynaptic neurexins and postsynaptic neuroligins. Neurexins were originally identified as receptors for α-latrotoxin, a potent neurotoxin from black widow spider venom that induces massive neurotransmitter release from nerve terminals. Subsequently, neuroligin was identified as an endogenous postsynaptic ligand of neurexins. In addition, neuroligin was found to directly interact with the PDZ domains of PSD-95 through the C-terminal tail, providing a novel mechanism of synapse formation.

The neurexin–neuroligin complex promotes synapse formation in a bidirectional manner. Neuroligin expressed in nonneural cells induces presynaptic differentiation in contacting axons of co-cultured neurons. Conversely, neurexin presented on nonneural cells or beads induces the clustering of key postsynaptic proteins in contacting dendrites. Direct aggregation of neuroligins on the surface membrane of dendrites induces similar clustering of postsynaptic proteins. Acute knockdown of neuroligins reduces the number and function of synapses. These results suggest that neuroligin is a key mediator of synapse formation. Interestingly, transgenic mice deficient in three neuroligins (mice with triple knockout mutations of neuroligin-1, -2, and -3) exhibit reductions in synaptic transmission, but their synapse number is not affected. This suggests that neuroligins regulate functional maturation of synapses rather than their formation, although further work in this area is required.

An interesting feature of the neuroligin family is that neuroligin isoforms differentially localize to excitatory and inhibitory synapses. Specifically, neuroligin-2 is mainly found at inhibitory synapses, whereas other neuroligin isoforms are detected at excitatory synapses. Consistent with these observations, direct aggregation of neuroligin-2 on dendrites induces the clustering of gephyrin, an inhibitory postsynaptic scaffold. Acute knockdown of neuroligin-1, -2, and -3 in cultured neurons results in a greater reduction in the function of inhibitory synapses than excitatory synapses. In addition, the neuroligin triple-knockout mice show an increase in the ratio of excitatory to inhibitory synapses. These results suggest that neuroligin-2 regulates inhibitory synapse formation and/or function.

SynCAM is another family of synaptic cell adhesion molecules implicated in excitatory synaptic differentiation. SynCAM family members, expressed in nonneural cells, induce the formation of functional nerve terminals in contacting axons. These terminals have release properties similar to those seen in regular synapses. The C-terminus of SynCAM contains a PDZ-binding motif and associates with synaptic PDZ proteins, including CASK/LIN2 and syntenin. This, together with data from studies on neuroligin (described previously), suggests that synaptic differentiation is mediated by synaptic scaffolds that couple synaptic adhesion events to the recruitment of various synaptic proteins.

When the great diversity of neuronal synapses is considered, synaptogenic adhesion molecules other than neuroligin and SynCAM may act in concert

with postsynaptic scaffolds. NGL, a family of adhesion molecules, interacts with PSD-95 through the C-terminus in a manner similar to the neuroligin–PSD-95 interaction. The extracellular domain of NGL associates with netrin-G/laminet, a GPI-anchored adhesion molecule. The complex of netrin-G, NGL, and PSD-95 is implicated in the regulation of excitatory synapse formation. In support of this idea, NGL presented on nonneural cells or on beads induces presynaptic differentiation in contacting axons. Direct aggregation of NGL on the dendritic surface induces postsynaptic protein clustering. The fact that both neuroligin and NGL associate with PSD-95 suggests that PSD-95 is one of the key postsynaptic scaffolds involved in adhesion-dependent postsynaptic differentiation. In addition, the dual association of PSD-95 with neuroligin and NGL suggests that these two adhesion molecules may have physical or functional interactions.

Extracellular factors capable of inducing aspects of postsynaptic differentiation at excitatory synapses include Narp, a secreted immediate early gene product upregulated by synaptic activity, and ephrin, a ligand of EphB receptor tyrosine kinases. Narp presented on nonneural cells induces the clustering of AMPA receptors, but not of NMDA receptors, in contacting dendrites. Narp interacts with the extracellular N-terminal domain of AMPA receptors. Narp induces AMPA receptor clustering at shaft synapses of aspiny neurons, and this leads to secondary NMDA receptor clustering, probably through stargazin and PSD-95. Ephrin activation of EphB receptor tyrosine kinases induces co-clustering of EphB and NMDA receptors. This interaction is mediated by their extracellular domains and does not require the kinase activity of EphB receptors.

Localization and Organization of Postsynaptic Scaffolds

If the clustering of postsynaptic adhesion molecules on dendrites is the beginning of postsynaptic differentiation, how might subsequent postsynaptic protein clustering occur? A possibility is that adhesion-induced primary clustering of neuroligin on dendrites may promote secondary clustering of PSD-95 though the C-terminal PDZ interaction, which would lead to additional recruitment of PSD-95-associated proteins. Against this, however, is the observation that while neuroligin-1 selectively binds to the third PDZ domain of PSD-95, synaptic localization of PSD-95 requires the first two PDZ domains but not the third. Notably, NGL, another adhesion molecule binding to both of the first two PDZ domains of PSD-95 but not to the third, enhances synaptic localization of PSD-95.

PSD-95 occurs in molar excess relative to other postsynaptic proteins. Therefore, instead of depending on its synaptic localization on the interaction with other proteins, PSD-95 might increase its cluster size at synapses by self-multimerization. In support of this notion, PSD-95 forms multimers through both N- and C-terminal domains. The expansion of PSD-95 multimers is more likely to be lateral than vertical because, ultrastructurally, PSD-95 mainly localizes to regions close to the postsynaptic membrane. The lateral expansion might be aided by PSD-95 interaction with membrane proteins and/or palmitoylation (a lipid modification promoting membrane attachment). Indeed, PSD-95 co-expressed with membrane proteins in heterologous cells forms large surface clusters in which both proteins are co-localized. In addition, mutations of PSD-95 that block palmitoylation eliminate membrane protein clustering by PSD-95.

PSD-95 is likely to recruit other postsynaptic scaffolds, including GKAP and Shank. The C-terminal GK domain of PSD-95 associates with GKAP, and the C-terminal PDZ-binding motif of GKAP further associates with Shank. GKAP and Shank are relatively abundant in deeper layers of the PSD and contain various domains for protein–protein interactions. Shank further associates with Homer, which is in turn linked to metabotropic glutamate receptors and IP3 receptors. Shank promotes spine maturation by mechanisms requiring synaptic Homer recruitment. In support of a possible role for PSD-95 in the recruitment of GKAP and Shank, a GKAP mutant lacking the ability to bind to PSD-95 induces the aggregation and degradation of Shank.

Synaptic Adhesion Molecules

Synaptically localized PSD-95 may reversely promote postsynaptic localization of adhesion molecules such as neuroligins or other postsynaptic adhesion molecules, further stabilizing synapse adhesion and promoting presynaptic differentiation. Indeed, overexpression of PSD-95 in cultured neurons concentrates neuroligin-1 and NGL at excitatory synapses. Importantly, PSD-95 induces the translocation of neuroligin-2 from inhibitory to excitatory synapses. Because neuroligin-2 induces presynaptic differentiation at both excitatory and inhibitory synapses, this translocation is likely to increase the number of excitatory synapses at the expense of inhibitory synapses. Therefore, the relative amounts of neuroligin-2 and PSD-95 in a single neuron may determine the ratio of excitatory and inhibitory synapses.

Although PSD-95 concentrates neuroligin at excitatory synapses, this does not seem to involve a direct interaction between the two proteins because a

neuroligin-1 mutant lacking the PSD-95-binding C-terminus is normally targeted to excitatory synapses. Instead, synaptic neuroligin-1 localization requires a membrane-proximal domain in the cytoplasmic domain. This suggests that neuroligin does not depend on its binding either to PSD-95 or to presynaptic neurexins for synaptic localization, and that there is a mechanism that precedes neuroligin localization for early postsynaptic differentiation.

Postsynaptic Receptors

Postsynaptic differentiation at excitatory synapses involves synaptic localization and local trafficking of NMDA and AMPA glutamate receptors. NMDA receptors are targeted to synapses at early stages of development, in contrast to AMPA receptors, thus forming NMDA receptor-only silent synapses. PSD-95 may contribute to synaptic localization of NMDA receptors by interaction with NR2 subunits. In support of this idea, C-terminal casein kinase II phosphorylation of NR2B on Ser1480 within the C-terminal PDZ-binding motif disrupts PSD-95 interaction with NR2B and decreases surface expression of NR2B. PSD-95 coexpression slows the internalization rate of chimeras of Tac (a surface membrane protein) containing the distal tail of NR2B. However, an NR2B mutant that lacks the ability to bind to both PSD-95 and AP2, a clathrin adaptor complex, is retained in the synapse, suggesting that PSD-95 binding may not be important for synaptic localization. In addition, mobile NMDA receptor transport packets in early stage neurons can be recruited to nascent synapses lacking PSD-95. In contrast to NR2B, NR2A does not depend on PSD-95 interaction for synaptic localization, suggesting subunit-specific rules for NMDA receptor trafficking.

Postsynaptic adhesion molecules may regulate synaptic NMDA receptor clustering. A mutant neuroligin-1 that lacks PSD-95 binding ability no longer induces PSD-95 clustering in cultured neurons but retains the ability to cluster NMDA receptors, although weaker than that afforded by wild-type neuroligin-1. This suggests that neuroligin is capable of recruiting NMDA receptors through PSD-95-independent mechanisms. SALM, a PSD-95-interacting family of synaptic adhesion-like molecules, exhibits NMDA receptor clustering activity. SALM1, a member of this family, directly associates with the NR1 subunit and promotes dendritic clustering of NMDA receptors through mechanisms requiring the PDZ-binding C-terminus. SALM2 forms a complex with both NMDA and AMPA receptors, and direct aggregation of SALM2 on dendrites induces co-clustering of NMDA and AMPA receptors.

Synaptic localization of AMPA receptors is regulated by stargazin/TARP (transmembrane AMPA receptor regulatory protein), which directly associates with both AMPA receptors and PSD-95. Stargazin traffics AMPA receptors to synapses via two distinct mechanisms. Stargazin induces surface expression of AMPA receptors and also facilitates synaptic docking of the stargazin–AMPA receptor complex. The latter mechanism depends on binding of the stargazin C-terminus to PSD-95. In support of a role for PSD-95 in synaptic localization of stargazin and AMPA receptors, a stargazin mutant that lacks PSD-95 binding rescues surface AMPA receptor responses, but not synaptic AMPA receptor responses, in cerebellar granule cells from stargazin-deficient stargazer mice.

Other postsynaptic proteins regulating synaptic localization and trafficking of AMPA receptors include GRIP/ABP, PICK1, and *N*-ethylmaleimide-sensitive factor (NSF). GRIP/ABP, a multi-PDZ protein, is implicated in the stabilization of AMPA receptors at the synaptic surface. Synaptic AMPA receptor localization is also regulated by various GRIP-associated proteins including liprin-α (a multidomain protein), GIT1 (a multidomain protein), and LAR (a receptor tyrosine phosphatase). PICK1, through its PDZ domain, associates with protein kinase Cα (PKCα) in addition to AMPA receptors, and it promotes synaptic delivery of PKCα. PKCα directly binds to GRIP, probably promoting the interaction between PICK1-bound PKCα and its substrate, GRIP-bound GluR2. PKC-dependent phosphorylation of GluR2 on Ser880 within the C-terminal PDZ-binding motif selectively disrupts GluR2 association with GRIP but not GluR2 association with PICK1, facilitating AMPA receptor endocytosis. NSF, an ATPase involved in membrane fusion, binds to a cytoplasmic region of AMPA receptors that is distinct from the GRIP/PICK1 binding site. NSF is implicated in maintaining synaptic AMPA receptors by disassembling the AMPA receptor–PICK1 complex and promoting AMPA receptor recycling/delivery. Interestingly, the NSF binding site in AMPA receptors overlaps with that of AP2, and the AP2–AMPA receptor interaction is required for regulated AMPA receptor endocytosis and LTD. In strong support of these *in vitro* results, cerebellar LTD is absent in GluR2 and PICK1 knockout mice and in two different strains of GluR2 mutant knockin mice (GluR2Δ7 and GluR2 K882A). GluR2Δ7 mice have a deletion of the last seven amino acid residues required for both GRIP and PICK1 binding, and GluR2 K882A mice carry a mutation to block PKC-dependent Ser880 phosphorylation. In addition to promotion of endocytosis, PICK1 and NSF regulate AMPA receptor exocytosis. PICK1 and NSF are required for activity-dependent

insertion of GluR2 (calcium-impermeable)-containing AMPA receptors in cerebellar granule–stellate cell synapses. In support of these *in vitro* observations, PICK1 knockout and GluR2Δ7 knockin mice lack this form of plasticity.

Other Membrane Proteins

Synaptic scaffolds concentrate and cluster interacting membrane proteins at the surface membrane. Stability of membrane proteins at the synaptic surface seems to be achieved through the inhibition of their endocytosis and promotion of their exit from intracellular pools and insertion into the plasma membrane. PSD-95 inhibits the endocytosis of NMDA receptors, Kv1.4 potassium channels, and β_1-adrenergic receptors. In addition, PSD-95 promotes the rate of membrane insertion of NMDA receptors. A C-terminal PDZ-binding motif in NR1-3, a splice variant of NR1, suppresses NR1-3 retention at the endoplasmic reticulum (ER), suggesting that a PDZ protein promotes the ER exit and surface expression of NMDA receptors. However, there are examples of the opposite situation, in which scaffolding proteins enhance endocytosis. S-SCAM/MAGI-2, a synaptic multi-PDZ protein, enhances the endocytosis of β_1-adrenergic receptors, and PICK1 is involved in AMPA receptor internalization, as described previously. Finally, functional properties of membrane proteins can be directly regulated by their interaction with scaffolds. PSD-95 suppresses single-channel conductance of the Kir2.3 potassium channel and increases the channel opening rate of NMDA receptors.

Signaling Pathways

Another key event in postsynaptic development is the establishment of signaling pathways in the PSD. A suggested role for synaptic scaffolds is to couple upstream receptor activations to downstream signaling pathways. PSD-95 associates with neuronal nitric oxide synthase, coupling NMDA receptor activation to nitric oxide generation. Similarly, PSD-95 associates with SynGAP, a neuronal GTPase activating protein (GAP) for Ras and Rap small GTPases. This interaction is not involved in synaptic localization of SynGAP but is implicated in functional coupling between NMDA receptors and the Ras-ERK signaling pathway, which regulates AMPA receptor trafficking and synaptic plasticity.

Synaptic scaffolds couple kinases and phosphatases with their specific substrates. PSD-95 and SAP97 associate with AKAP79/150, a neuronal A-kinase-anchoring protein interacting with protein kinase A (PKA), PKC, and protein phosphatase 2B (calcineurin). This interaction, in the context of SAP97 association with GluR1, promotes the PKA-dependent phosphorylation of GluR1 on Ser845, a modification implicated in AMPA receptor function and synaptic plasticity.

PSD-95 associates with Fyn, a nonreceptor tyrosine kinase, to promote Fyn-mediated tyrosine phosphorylation of NR2A. PSD-95 binds to another nonreceptor tyrosine kinase, Src, regulating NMDA receptor-dependent synaptic transmission and plasticity. Notably, this interaction is not involved in synaptic recruitment of Src but, rather, in the suppression of Src activity and Src-mediated NMDA receptor upregulation, reminiscent of the direct functional regulation of membrane proteins by scaffolding proteins.

F-Actin

F-actin is a key cytoskeletal component both in dendritic filopodia and in spines, and it is implicated in the regulation of spine morphogenesis and synaptic plasticity. F-actin is both physically and functionally associated with PSD components. For instance, stable maintenance of GKAP, Shank, and AMPA receptors at synapses requires F-actin integrity. Conversely, LTP- and LTD-inducing stimuli regulate F-actin polymerization. In this context, it is conceivable that PSD proteins may organize F-actin-regulatory signaling pathways.

F-actin polymerization in dendritic spines is regulated by small GTPases, including Rac, Cdc42, Rho, and Rap. Proteins acting upstream and downstream of these small GTPases associate with postsynaptic scaffolds, suggesting that these interactions may constitute related signaling pathways in the PSD. PSD-95 associates with and promotes synaptic localization of kalirin-7, a guanine nucleotide exchange factor (GEF) for Rac1. Shank binds βPIX, a GEF for Rac1 and Cdc42, and promotes synaptic localization of βPIX and βPIX-associated PAK, a kinase downstream of Rac1 and Cdc42 that regulates spine morphogenesis through LIMK-1 and MLC. Both PSD-95 and Shank associate with IRSp53, an abundant postsynaptic protein downstream of Rac1 that regulates spine morphogenesis. Shank associates with Abp1, α-fodrin, and cortactin, proteins that have F-actin binding, bundling, and nucleating activities, respectively. GRIP/ABP associates with EphB receptor tyrosine kinases, and ephrin activation of EphB receptors induces dendritic spine formation through the kalirin–Rac1–PAK pathway and the Cdc42 GEF intersection. Neurabin and spinophilin, which bind protein phosphatase 1 and F-actin and regulate dendritic spines, interact with Lfc, a Rho GEF. In

addition, NMDA receptors associate with and phosphorylate Tiam1, a Rac1 GEF that mediates NMDA receptor-dependent spine regulation.

Neuronal Transport

Synaptic proteins synthesized in the cell body must be transported to their target synapses. Kinesin is a microtubule (MT)-based motor protein implicated in this process. There are approximately 45 kinesin family proteins in the mouse and human genomes. Most kinesins move toward the plus end of MTs. MTs in axons are unidirectionally oriented, with their plus ends pointing toward nerve terminals, whereas dendritic MTs are bidirectionally arranged. Kinesin-dependent transport has been studied mainly in axons, although evidence supports kinesin involvement in dendritic transport.

An important question in neuron motor-dependent transport is how the limited number of kinesin proteins can transport a large number of cargoes. Studies raise the intriguing possibility that molecular scaffolds function as 'motor receptors,' linking motors to various cargoes through protein interaction domains. In accordance with this notion, synaptic scaffolds link kinesins to their specific cargoes. KIF1Bα, a kinesin motor, interacts with PSD-95 and S-SCAM, which are in turn linked to various synaptic proteins. KIF17 associates with the LIN2/7/10 PDZ protein complex that is coupled to NMDA receptors. KIF1A associates with liprin-α, a multidomain protein interacting with the GRIP–AMPA receptor complex. KIF5 interacts with GRIP/ABP, a binding partner of both AMPA receptors and EphB receptors, and disruption of the KIF5-GRIP interaction suppresses EphB trafficking and dendritic morphogenesis. Interestingly, GRIP drives KIF5 to dendrites, suggesting that a cargo regulates polarized transport of a motor.

Regulation of Postsynaptic Assembly

Synapses are dynamically formed and eliminated during development and plasticity. These processes are likely to involve rapid assembly and disassembly of the PSD. Indeed, neuronal activity regulates synaptic localization of PSD components, changing the overall molecular composition of the PSD. GluR1 and protein phosphatase 1 are delivered to synapses by LTP- and LTD-inducing stimuli, respectively. Phosphorylation of synaptic proteins regulates their synaptic localization by affecting, for example, protein–protein interactions. Phosphorylation of the N-terminal domain of PSD-95 by cyclin-dependent kinase 5 suppresses synaptic clustering of PSD-95, whereas SAP-97 phosphorylation in the L27 domain by CamKII enhances synaptic SAP97 localization. Phosphorylation of PDZ-binding ligands disrupts their binding to PDZ domains. Lipid modification also plays a role in postsynaptic assembly. Known examples include palmitoylation of PSD-95, PSD-93, GRIP, and AMPA receptors. The lipid addition regulates the trafficking and synaptic localization of these proteins. Synaptic activity depalmitoylates and disperses synaptic PSD-95 clusters. Enzymes that mediate protein palmitoylation (palmitoyl acyl transferases) of neuronal substrates have been identified. Lastly, protein degradation through the ubiquitin–proteasome pathway is involved in postsynaptic assembly. Neuronal activity regulates ubiquitination of key PSD proteins, including GKAP, Shank, and AKAP79/150. Mdm2, an E3 ubiquitin ligase, plays a role in ubiquitinating PSD-95. SNK, a polo-like kinase induced by synaptic activity, phosphorylates SPAR, a PSD-95-associated Rap GAP that regulates F-actin, and this in turn induces PSD-95 degradation and spine loss.

See also: Cell Adhesion Molecules at Synapses; Dendrite Development, Synapse Formation and Elimination; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Microtubule Associated Proteins in Neurons; Neurexins; Neuroligins; Neuromuscular Junction (NMJ): Postsynaptic Events in Neuromuscular Transmission; Postsynaptic Density/Architecture at Excitatory Synapses; Presynaptic Development and Active Zones; Retrograde Transsynaptic Influences.

Further Reading

Akins MR and Biederer T (2006) Cell–cell interactions in synaptogenesis. *Current Opinion in Neurobiology* 16: 83–89.

Barry MF and Ziff EB (2002) Receptor trafficking and the plasticity of excitatory synapses. *Current Opinion in Neurobiology* 12: 279–286.

Craig AM, Graf ER, and Linhoff MW (2006) How to build a central synapse: Clues from cell culture. *Trends in Neuroscience* 29: 8–20.

Funke L, Dakoji S, and Bredt DS (2004) Membrane-associated guanylate kinases regulate adhesion and plasticity at cell junctions. *Annual Review of Biochemistry* 74: 219–245.

Garner CC, Waites CL, and Ziv NE (2006) Synapse development: Still looking for the forest, still lost in the trees. *Cell and Tissue Research* 326: 249–262.

Huang K and El-Husseini A (2005) Modulation of neuronal protein trafficking and function by palmitoylation. *Current Opinion in Neurobiology* 15: 527–535.

Kim E and Sheng M (2004) PDZ domain proteins of synapses. *Nature Reviews Neuroscience* 5: 771–781.

Kornau HC, Seeburg PH, and Kennedy MB (1997) Interaction of ion channels and receptors with PDZ domain proteins. *Current Opinion in Neurobiology* 7: 368–373.

Levinson JN and El-Husseini A (2005) Building excitatory and inhibitory synapses: Balancing neuroligin partnerships. *Neuron* 48: 171–174.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* 25: 103–126.

Nicoll RA, Tomita S, and Bredt DS (2006) Auxiliary subunits assist AMPA-type glutamate receptors. *Science* 311: 1253–1256.

Scannevin RH and Huganir RL (2000) Postsynaptic organization and regulation of excitatory synapses. *Nature Reviews Neuroscience* 1: 133–141.

Scheiffele P (2003) Cell–cell signaling during synapse formation in the CNS. *Annual Review of Neuroscience* 26: 485–508.

Sheng M and Kim MJ (2002) Postsynaptic signaling and plasticity mechanisms. *Science* 298: 776–780.

Sheng M and Sala C (2001) PDZ domains and the organization of supramolecular complexes. *Annual Review of Neuroscience* 24: 1–29.

Sudhof TC (2001) Alpha-latrotoxin and its receptors: Neurexins and CIRL/latrophilins. *Annual Review of Neuroscience* 24: 933–962.

Tada T and Sheng M (2006) Molecular mechanisms of dendritic spine morphogenesis. *Current Opinion in Neurobiology* 16: 95–101.

Thomas GM and Huganir RL (2004) MAPK cascade signalling and synaptic plasticity. *Nature Reviews Neuroscience* 5: 173–183.

Wenthold RJ, Prybylowski K, Standley S, Sans N, and Petralia RS (2003) Trafficking of NMDA receptors. *Annual Review of Pharmacology and Toxicology* 43: 335–358.

Yi JJ and Ehlers MD (2005) Ubiquitin and protein turnover in synapse function. *Neuron* 47: 629–632.

Postsynaptic Specialization Assembly

C Sala, CNR Institute of Neuroscience, Milan, Italy
A El-Husseini[†], University of British Columbia, Vancouver, BC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Synapse are specialized contacts used by neurons to communicate with each other. Morphologically, they are cell junctions characterized by specialized membrane and submembrane regions, and their assembly takes place in various stages involving the accumulation of specific pre- and postsynaptic proteins at the contact site. Recent biochemical, genetic, and imaging studies have begun to elucidate the molecular mechanisms underlying their formation, growth, and maturation, and understanding the development and organization of neuron/neuron synapses is crucial to our understanding of the development of the nervous system, in which abnormalities contribute to most neurological and behavioral disorders. Moreover, it is now clear that the adult brain is constantly reorganizing itself in response to experience, and that at least some synaptic developmental mechanisms are used in mature brain during synaptic plasticity.

At the postsynaptic sites of excitatory synapses is localized a specific EM electron-dense organelle called the postsynaptic density (PSD) where the two main types of glutamate-receptor channels, *N*-methyl-D-aspartate (NMDA) and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors, cluster in the postsynaptic membrane. The PSD has attracted the interest of a number of scientists because it is now clear that it regulates synapse function (Figure 1).

Inhibitory synapse in brain are mainly mediated by γ-aminobutyric acid (GABA) neurotransmitters, whose receptor channels are localized on the postsynaptic membrane and are not associated with a clear PSD because the molecular composition of the postsynaptic compartment is much simpler.

Excitatory and inhibitory synapses are therefore clearly distinguishable on the basis of the morphology and molecular composition of the postsynaptic compartment. Each neuron receives both excitatory and inhibitory synapses, which are located in cross-proximity to dendrites in strictly controlled numbers and ratios in order to allow the correct functioning of brain circuits.

Our knowledge of the neuron/neuron synaptogenesis of both types of synapses has recently increased as a result of improvements in time-lapse microscopy and the molecular identification of specific proteins in neuronal synapses. This article concentrates on the structure and development of the synapses between central neurons in mammals, paying particular attention to the postsynaptic compartment. Emphasis will be placed on their molecular and structural composition of the excitatory synapses, and the early steps of synapse formation that underlie their molecular mechanisms.

Excitatory Synapses: Postsynaptic Organization

The PSD is a protein complex that forms a specialized organelle and adheres to the postsynaptic membrane in opposition to the active zones located in the axonal presynaptic terminal. A number of functions have been proposed for this structure, including the regulation of adhesion, the control of receptor clustering, and the regulation of receptor functions.

The PSD contains neurotransmitter and transmembrane receptors, various scaffold proteins, cytoskeletal elements, and regulatory enzymes, all of which are assembled together in a disk-like structure that is 30–40 nm thick and a few hundred nanometers wide.

PSD molecular architecture is designed to specialize in signal transduction, and its protein composition is highly modifiable in order to allow for the strengthening and weakening of glutamatergic synapses. Neural activity dynamically regulates both the function and composition of the PSD, and involves mechanisms such as protein phosphorylation, local translation, ubiquitination and degradation, and subcellular redistribution. Because of the central role of the PSD in synaptic transmission and plasticity, the protein composition of the PSD has been extensively studied in an attempt to understand synaptic mechanisms.

The protein components of the PSD have been gradually discovered over the years using two-hybrid yeast and biochemical approaches. The PSD is typically purified by means of differential centrifugation, sucrose gradient sedimentation, and detergent extraction, because the PSD structure cannot be solubilized using mild nonionic detergents such as Triton X-100. Kennedy et al. first discovered several PSD proteins (e.g., PSD-95, densin-180) by sequencing protein bands from one-dimensional gels of PSD preparations, and many other PSD proteins were later identified using the two-hybrid yeast system as binding partners of known postsynaptic proteins, such as the PSD-95 family of PDZ-containing scaffolds that interact with the cytoplasmic tails of NMDA

[†]Deceased

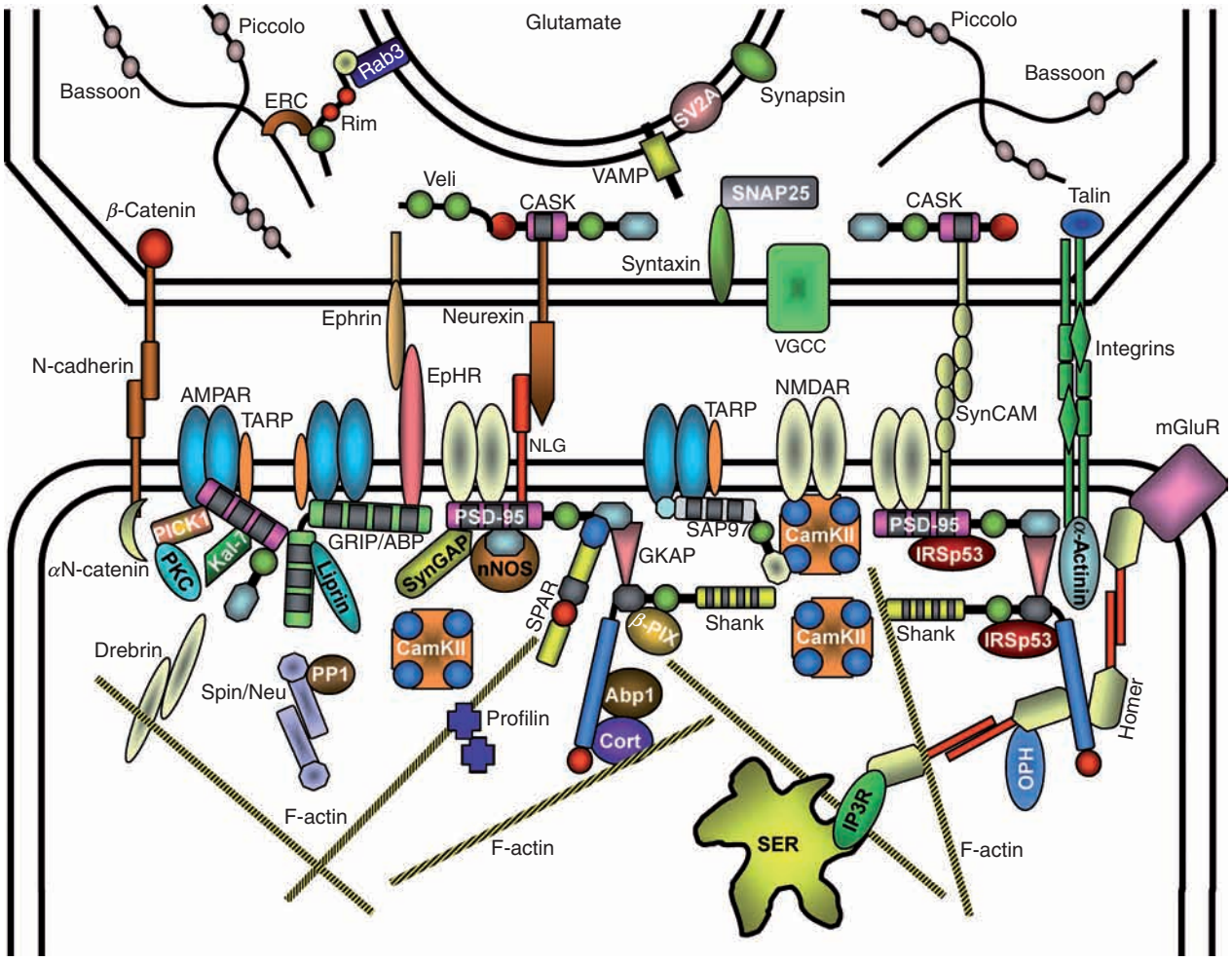

Figure 1 Molecular organization of glutamatergic synapses. Only the major components and relative interactions are shown. NLG, neuroligin; Kal-7, kalirin-7; Spin/Neu, spinophilin and neurabin I; Cort, cortactin; OPH, oligophrenin; PKC, protein kinase C; VGCC, voltage-gated calcium channels; nNOS, nitric oxide synthase; NMDAR, *N*-methyl-D-aspartate; GKAP, guanulate kinase-associated protein; IP3R, inositol-1,4,5-triphosphate; mGlur, metabotropic glutamate receptors; AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazoll propronic acid receptor.

receptors. More recently, modern proteomic and imaging techniques have probably identified all of the proteins of the PSD, and quantified the relative stoichiometry of some of them.

The components of PSD proteins can be divided into four subgroups on the basis of their functions: receptor transmembrane proteins, scaffold proteins, signaling proteins, and cytoskeletal proteins. We will here discuss the major PSD components in each subgroup.

The PSD-95 Family and Other Major PSD Scaffold Proteins

The first and still most widely studied PSD scaffold protein was PSD-95. The PSD-95 family is encoded by four genes: PSD-95/SAP90 (synapse-associated protein 90), PSD-93/chapsyn-110, SAP102, and SAP97. The structure of these proteins is characterized by the assembly of three PDZ (PSD-95, Dlg, ZO-1 homology) domains, an SRC homology (SH3) domain, and a guanylate kinase-like (GK) domain. SAP97 also has a LIN2/LIN7 (L27) domain at the N-terminal. The PSD-95 family belongs to a protein superfamily called membrane-associated guanylate kinase (MAGUK), which is characterized by the presence of at least one PDZ and one GK domain.

Immuno-EM and tomography studies have indicated that PSD-95 is localized very close to the postsynaptic membrane (a mean distance of 12 nm), a good position for interacting with postsynaptic membrane proteins such as receptors, ion channels, and cell-adhesion molecules, as well as with cytoplasmic proteins (Figure 1). It has been suggested that these interactions are important for PSD-95 clustering

and the targeting of postsynaptic membrane proteins, but there is still disagreement as to whether PSD-95 is essential for the synaptic clustering of NMDA receptors in mammals, which is certainly not eliminated by the genetic disruption of PSD-95.

It now seems clear that the most important biochemical function of PSD-95 is to organize the signaling complexes at the postsynaptic membrane. PSD-95 interacts with a wide variety of cytoplasmic signaling molecules and, by physically bringing together cytoplasmic signal-transducing enzymes and surface receptors, may therefore facilitate signal coupling in the PSD. For example, the overexpression of PSD-95 in hippocampal neurons increases the number of dendritic spines, the maturation of excitatory synapses, and the accumulation of AMPA receptors, whereas PSD-95 knockout by RNA interference in developing hippocampal neurons modifies the excitatory/inhibitory synapse ratio in favor of inhibitory synapses. PSD-95 directly binds to the proteins involved in actin remodeling: these include kalirin-7, a guanine nucleotide exchange factor (GEF) for RAC1 that promotes spine formation possibly as a downstream effector of the EphB receptor; SPAR, an inhibitory GTPase-activating protein (GAP) for RAP that promotes the growth of dendritic spines; and IRSp53, a downstream effector of Cdc42 small GTPase. The ability of PSD-95 to recruit AMPA receptors seems to be related to its binding to the tetra-spanning membrane proteins Stargazin and its relatives TARP, which are associated with AMPA receptors and essential for receptor surface expression, and synaptic accumulation and function.

Three other major scaffold components of the PSD that directly interact with each other are Shank, guanylate kinase-associated protein (GKAP), and Homer protein families. Shank, which is also known as proline-rich synapse-associated protein (ProSAP), is a large scaffold protein whose multidomain organization consists of an ankyrin repeat near the N-terminal, followed by an SH3 and a PDZ domain, a long proline-rich region, and a sterile alpha motif domain (SAM) at the C-terminus. Shank proteins (codified by the three genes *Shank1–3*) molecularly link two glutamate receptor subtypes: NMDA receptors and type-I metabotropic glutomate receptors (mGluRs). The Shank PDZ domain binds to the C-terminal of GKAP. Homer interaction with the proline-rich domain ensures the association of Shank with type-I mGluRs.

Shank interacts with a number of actin-binding proteins. It binds to cortactin, Abp1, fodrin, the Rac1 and Cdc42 exchange factor βPIX, and the Cdc42-binding protein IRSp53. Other important Shank interactors are proteins that regulate dendritic morphology and formation, such as Dasm1 and Densin-180, whose activity also seems to be regulated by Shank. Finally, Shank proteins can potentially be produced locally because of the presence of its mRNA in dendrites. The observation that the haploinsunficency of Shank3 in humans probably causes the neurological symptoms of individuals affected by the 22q13 deletion syndrome indicates that Shank proteins play a major role in synapse functions.

The GKAP/SAPAP family of proteins is less well characterized, but its four members were originally identified as proteins interacting with the GK domain of PSD-95. GKAP has five repeats of 14 amino acids involved in interacting with PSD-95 and, as it binds to *S*-SCAM, nArgBP2, Dynein light chain, and Shank, may therefore function as a scaffolding protein that links PSD protein complexes to motor proteins.

Homer proteins are encoded by three genes (*Homer 1–3*), and typically consist of an N-terminal Ena/VASP homology 1 (EVH1) domain followed by a coiled-coil domain that mediates dimerization with other Homer proteins. The EVH1 domain of Homer1 binds to a PPXXF or very similar sequence motif in Shank, mGluR1/5, inositol-1,4,5-trisphosphate (IP_3) receptor, ryanodine receptor, and to different members of the TRPC family of ion channels. Through their ability to self-associate, Homer isoforms containing the coiled-coil domain (called 'CC-Homer' or Homer1b in the case of the Homer1 gene) can physically and functionally link the proteins and receptors that bind to the EVH1 domain. Homer1a is a short-splice variant of Homer1 that contains the EVH1 domain, but lacks the coiled-coil domain; it functions as a natural dominant negative because it cannot dimerize. It is important to note that Homer1a expression is induced at mRNA level by synaptic activity.

AMPA receptors are linked to a different set of scaffolding proteins, which includes GRIP/ABP (encoded by the two distinct genes *GRIP1* and *ABP/GRIP2*), and the protein interacting with C kinase 1 (PICK1); these interactions may account for the dynamic cell biological behavior of AMPARs at synapses.

It is believed that GRIP is involved in the synaptic trafficking and/or stabilization of AMPARs and other interacting proteins. GRIP has up to seven PDZ domains by means of which it interacts with many proteins, including Eph receptors and their ephrin ligands, an RAS guanine nucleotide exchange factor (RasGEF), liprin-α, the transmembrane protein Fraser syndrome 1 (FRAS1), and the metabotropic and kainite-type glutamate receptors. It can therefore participate in synaptic function not only by interacting with AMPARs, but also by associating with Eph receptors and their ephrin ligands, which have been

found to be involved in dendrites and dendritic spine morphogenesis and hippocampal synaptic plasticity. The widespread cellular distribution of GRIP, and its interaction with motor proteins (directly with conventional kinesin KIF5 or indirectly with KIF1A via liprin-α), suggest that multiple motor proteins may contribute to the transport of AMPARs.

PICK1 is located at neuronal synaptic and nonsynaptic sites. Its PDZ domain has relatively promiscuous binding properties, and both pre- and postsynaptic partners have been found: PKCα, GluR2/3, the netrin receptor UNC5H, various metabotropic glutamate receptor subtypes, the dopamine plasma-membrane transporter, and the erythroblastic leukemia viral oncogene homolog 2 (ErbB2) receptor tyrosine kinase. In most of these cases, the subcellular localization and/or surface expression of these partners seems to be regulated by interactions with PICK1.

Receptor and Transmembrane Proteins

The major transmembrane and receptor proteins are of course the glutamate receptors which, in the PSD, are organized into supramolecular signaling complexes by interacting with specific PDZ domain-containing scaffold proteins, such as PSD-95/SAP90 and their associated proteins for the NMDA receptors, and GRIP/ABP for the AMPA receptors. The distribution of glutamate receptors and content of synapses vary during development: immature synapses preferentially contain NMDA receptors and gradually acquire AMPA receptors, an observation that has led to the identification of what have been called silent synapses because the absence of AMPA receptors makes them less responsive to glutamate. AMPA and NMDA receptors seem to be homogenously mixed and distributed inside the postsynaptic membrane, but the mGluRs are mainly distributed at the leading edge of the synapses.

Behind the receptors, an important role is played by the transmembrane adhesion molecules that regulate synapse formation and structures. Neuroligin is a synaptic adhesion molecule that probably plays an important role in synapse formation: it interacts with PSD-95 and transsynaptically with β-neurexins, which in turn bind to the PDZ domain of CASK/LIN2 (another scaffold of the MAGUK superfamily of proteins), which is enriched on both sides of the synapse and interacts with other synaptic membrane proteins such as syndecan and SynCAM. The transsynaptic neuroligin/β-neurexin interaction seems to be important for inducing pre- and postsynaptic synaptic differentiation, and the amount of PSDS-95 regulates the balance between the number of inhibitory and excitatory synapses (see below). When heterologous cells overexpressing neuroligin are co-cultured with neurons, they are able to induce morphological and functional presynaptic specializations in contacting axons.

Cadherins are found on both the pre- and postsynaptic sides of neuron synapses. Dominant-negative studies of neuronal(N)-cadherin function indicate that the cadherin cell-adhesion system is important for synapse integrity and the morphological maturation of dendritic spines; however, it is unclear whether cadherins play an initiating role in synapse formation or a supporting role in synapse growth. Through their cytoplasmic tails, cadherins bind α- and β-catenins, which in turn bind to the actin cytoskeleton and thus anchor and consolidate the cadherin cell-adhesion complex.

Integrins are heterodimeric glycoproteins that contain α- and β-subunits. Many of the 16 α- and eight β-subunits that have been identified so far are expressed in different spatial patterns in the brain. Integrins function as cell-to-matrix or cell-to-cell adhesion molecules by respectively binding to extracellular matrix proteins or immunoglobulin-superfamily receptors on other cells. They interact through their cytoplasmic domains with the actin-binding proteins, talin and α-actinin, and also activate nonreceptor protein tyrosine kinases, such as the focal-adhesion kinase that affects cytoskeletal organization and intracellular signal transduction. One immunoglobulin domain-containing protein that is found in synaptic membranes is the synaptic cell-adhesion molecule (SynCAM), which can also induce presynaptic differentiation in contacting axons when it is expressed in nonneuronal cells. SynCAM shows homophilic cell-adhesion activity through its extracellular domain, and can bind to PDZ proteins such as calcium/calmodulin-dependent serine protein kinase (CASK) through its cytoplasmic tail.

Finally, synapses contain the important tyrosine kinase EphB (ephrin B) family of RTKs and their cognate ephrinB ligands. These proteins play well-known roles in axon guidance and topographic map formation in the nervous system. The ephrinB/EphB receptor system is also involved in the development of excitatory synapses and dendritic spines by means of the phosphorylation of a cell-surface heparan/sulphate proteoglycan Syndecan and the recruitment of kalirin, a guanine-nucleotide exchange factor for the Rac GTPase. In addition, EphB2 seems to interact physically with NMDA receptors and enhance their channel activity, and ephrinB ligands can induce the clustering of NMDA receptors.

Signaling Proteins

A number of signaling proteins are associated with PSD-95. One of these is nitric oxide synthase (nNOS), a Ca^{2+}/calmodulin-activated enzyme that produces the

nitric oxide involved in regulating neurotransmission and excitotoxicity. Interestingly, the ternary NMDAR/PSD-95/nNOS complex may functionally couple NMDAR gating to nNOS activation, as is suggested by the fact that disrupting the NMDAR/PSD-95 interaction by means of a synthetic peptide mimicking the last nine residues of NR2B reduces NMDAR-induced excitotoxicity without affecting NMDAR function.

The most abundant signaling protein in the PSD is its predominant kinase CaMKII, whose multiple functions include regulating receptor targeting and function at synapses. The substrates for CaMKII have not yet been completely identified, but they include several PSD proteins.

Another abundant PSD signaling molecule that binds to PSD-95 is the synaptic Ras GTPase-activating protein (SynGAP), a GAP for the Ras small GTPase which, after activation by CaMKII, suppresses the ERK pathway regulating synaptic plasticity. PSD-95 also associates with the nonreceptor tyrosine kinases of the Src family and their upstream activator, proline-rich tyrosine kinase 2 (Pyk2), both of which are thought to be important for synaptic plasticity. It is therefore possible that it may localize the Pyk2–Src signaling cascade close to NMDARs, although the importance of PSD-95 scaffolds in synaptic regulation by tyrosine phosphorylation has not been directly investigated.

Smaller amounts of several other signaling proteins are associated with the PSD, including a number of kinases and phosphates whose functions have not been completely clarified.

Cytoskeletal Proteins

It has been demonstrated that a number of actin-binding proteins (usually Rho and Rac GTPase effectors regulated by Ca^{2+}) control the actin cytoskeleton in dendritic spines and excitatory synapses. Most of them are formed by the concatenation of different protein interaction domains and thus function as scaffolds assembled in multimolecular complexes that are enriched in the PSD.

One of the most abundant cytoskeletal proteins is drebrin, an F-actin-binding protein that is mainly expressed in neurons and highly concentrated in the PSD. In cortical neurons, drebrin overexpression increases the length of dendritic spines and, in hippocampal neurons, drebrin promotes actin assembly and the synaptic clustering of PSD-95 in the PSD. Importantly, drebrin (but also profilin II and αN-catenin) are redistributed to dendritic spines with an increased F-actin content after the induction of LTP in the dentate gyrus, or general synaptic activation and NMDA receptor activation, which suggests their importance in inducing F-actin polymerization and/or stabilization upon synaptic stimulation.

Profilin is a small actin-binding protein that promotes actin polymerization by positioning the actin monomers at the barbed end of the growing F-actin. Profilin II stabilizes spine morphology in a mature state, and suppresses dendritic spine motility by reducing actin dynamics.

αN-catenin is a cadherin-associated protein that, together with β-catenin, links adhesion molecules to the cytoskeleton and actin. In the absence of αN-catenin, dendritic spines are more motile and their filopodia rapidly protrude and retract from the spine heads, a sign of unstable synaptic contacts. Conversely, the overexpression of αN-catenin accelerates dendritic spine maturation and decreases spine motility, thus suggesting that it promotes spine morphogenesis and stabilization.

Spinophilin (or neurabin II) and neurabin I are two related F-actin-binding proteins with similar domain structures containing a PDZ and a coiled-coil domain that form homo- and heterodimers. Via its actin-binding domain, spinophilin is predominantly localized on the dendritic spines of pyramidal neurons (as its name suggests); it seems to be required for the correct maturation of dendritic spines because knockout mice have more filopodia and immature spines, and show altered glutamatergic transmission. On the contrary, neurabin I overexpression induces the formation of dendritic filopodia in immature cultured hippocampal neurons, and promotes the enlargement of dendritic spines in older neurons. It is not clear how spinophilin and neurabin I have opposite effect on spines, although they both bind a similar set of synaptic proteins.

Two other abundant actin-binding proteins localized at the PSD are α-actinin-2 and IRSp53. α-Actinin is a member of the spectrin/dystrophin family of actin-binding proteins, and binds to the NR1 and NR2B subunits of NMDA receptors, thus providing a cytoskeletal bridge between NMDA receptors and actin.

IRSp53 not only interacts with Shank, PSD-95 and PSD-93, and all of these interactions are required for its spine localization: its overexpression in cultured neurons increases the density of dendritic spines, while its siRNA-mediated knockdown reduces spine density, length, and width. IRSp53 plays a role in linking PSD-95 to activated Rac1/Cdc42 and downstream effectors of actin regulation in spines.

Inhibitory Synapses: Postsynaptic Organization

The composition of inhibitory postsynaptic components is much simpler. Fast GABAergic neurotransmission is mediated by ionotropic GABA receptors (principally

the $GABA_A$ receptor subtype), which are ligand-gated chloride channels that are segregated from glutamate receptors and concentrated in the postsynaptic membrane of inhibitory synapses. Unlike the subunits of ionotropic glutamate receptors, each subunit of a $GABA_A$ receptor has its carboxyl terminus on the extracellular side of the membrane. The only major cytoplasmic domain of $GABA_A$ receptors is the loop between the third and fourth transmembrane segments. Reasoning that this loop is likely to form the site of GABA receptor interactions with intracellular proteins, two-hybrid yeast screens have been used to search for interacting gene products using these loops, and two proteins associated with microtubules have been identified: GABARAP, which binds to the intracellular loop of the abundant $\gamma 2$ subunit of $GABA_A$ receptors, and the MAP1B heavy chain that specifically binds to the intracellular loop of the $\rho 1$ subunit of ionotropic $GABA_C$ receptors, almost exclusively in the retina. In addition to these proteins, gephyrin, neuroligin-2, and the dystroglycan/dystrophin complex are other identified components of some mature GABAergic postsynaptic sites. Gephyrin and neuroligin-2 have been respectively implicated in GABA receptor localization to synapses and the formation of inhibitory synapses, but the mechanisms underlying these activities remain elusive: for example, neuroligin-2 is mainly localized at inhibitory synapses, but a possible C-terminal interactor (like PSD-95 at excitatory synapses) has not yet been identified.

Synapse Development: A Morphological View

Over the last few years, a number of studies have described how synapses are formed morphologically: for example, the first rudimentary synapses in the rat hippocampus can be observed on postnatal day 1, after which the density of synapses in the CNS gradually increases during the following weeks with the number doubling in the second postnatal week and from day 15 to adulthood. A similar time course can be observed in neurons *in vitro*, although the maximum number is generally smaller in cultures.

As previously mentioned, adhesion molecules play an essential role in synapse development, but it is not known how many neuronal cell-surface molecules really participate in this interaction, or how they signal to effect downstream events in synapse differentiation.

Synapse assembly has been more directly visualized using cultured neurons transfected with synaptic proteins tagged with fluorescent proteins, and these time-lapse studies have shown that many components of the synaptic junction are rapidly assembled after the initial contact of axons and dendrites in a timescale of tens of minutes (Figure 2). Some studies have shown that presynaptic markers such as Bassoon (a scaffold protein of the presynaptic active zone) and synaptophysin (a synaptic-vesicle membrane protein) can accumulate at new synapses before postsynaptic NMDA receptors and PSD-95, which may indicate that presynaptic differentiation precedes postsynaptic development.

These findings support the hypothesis that presynaptic specialization is constructed from preassembled 'packets' of vesicles and/or proteins, and at least two types of precursor complexes have so far been identified, one including VAMP, synaptic-vesicle protein 2 (SV2), synapsin I, and voltage-gated calcium channels, the other containing Piccolo, Bassoon, Rab3-interacting molecule (RIM), syntaxin, and SNAP-25 and N-cadherin.

The mechanisms of assembling postsynaptic specialization are less well characterized than those on the presynaptic side. Postsynaptic assembly may depend either on the gradual recruitment of individual proteins, or on the delivery of prefabricated protein complexes to the PSD. For example, a study has identified in dendrites the presence of a class of vesicles that carry NMDA receptors associated to proteins including Veli/MALS, CASK, and the microtubule-dependent motor KIF17. In a more recent publication it has demonstrated that a complex of postsynaptic scaffold proteins consists of two populations with different content, mobility and involvement in synapse formation. One subpopulation is mobile and relies on actin transport for delivery to nascent and

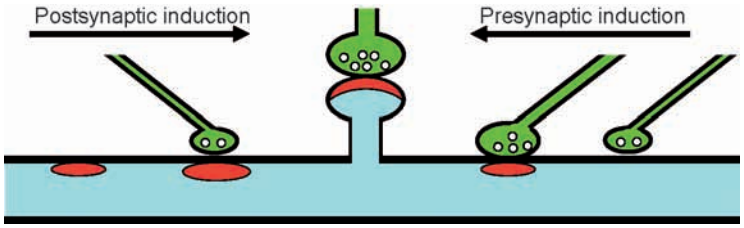

Figure 2 Schematic representation of the temporal sequence accumulation of specific pre- and postsynapse compartments based on time-lapse imaging observations of developing neurons in culture. Synapse formation is initiated by contact between dendrites and axons. These studies suggest that both synaptic sides can differentiate before the other, and that there are probably contact-induced signals triggering synapse assembly on both the pre- and postsynaptic sides.

existing synapses: these mobile clusters contain the PSD-95, GKAP, and Shank scaffolding proteins, and a slow-moving and short-traveling proportion contain neuroligin-1. The second group consists of stationary nonsynaptic scaffold complexes that mainly contain neuroligin-1, and can recruit synaptophysin-containing axonal transport vesicles that are rapidly transformed to functional presynaptic contacts that recycle the vital dye FM 4-64. These results postulate a mechanism whereby preformed scaffold protein complexes serve as predetermined postsynaptic hot spots for the establishment of new functional excitatory synapses. Interestingly, it has been found that the assembly of the scaffold protein complex requires PSD-95, because interfering with PSD-95 expression by siRNA in young hippocampal neurons disrupts the clustering of GKAP and Shank, reduces the number of excitatory synapses, and increases the number of vesicular GABA transporter (VGAT) puncta that are positive for neuroligin-1, thus indicating a shift in neuroligin-1 localization from excitatory to inhibitory contacts. The main conclusion of this study indicates that this preformed complex plays a role in inducing presynaptic maturation.

Other time-lapse imaging studies performed in dissociated hippocampal cultures indicate that PSD proteins accumulate gradually at nascent synapses with kinetics that have time constants in the range of minutes. Although not well understood, the gradual clustering of these proteins indicates that PSD proteins are recruited to nascent synaptic sites from diffuse cytoplasmic or membranal pools or, in some cases, are recruited to extrasynaptic sites, from where could then move laterally and merge with nearby synaptic clusters. It is possible to postulate that during the early stages of synaptic formation, postsynaptic preassembled protein complexes play a major role in the PSD assembly by catalyzing the initial step of synapses formation, whereas in a later stage gradual recruitment from diffuse intracellular pools regulates synaptic grow and maintenance.

Overall, these studies suggest that there are probably contact-induced signals triggering synapse assembly on both the pre- and postsynaptic sides and that both synaptic sides can differentiate before the other.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axonal Pathfinding: Extracellular Matrix Role; Cadherins and Synapse Organization; Cell Adhesion Molecules at Synapses; Dendritic Spine History; Glutamate Receptor Organization: Ultrastructural Insights; Postsynaptic Development: Neuronal Molecular Scaffolds; Postsynaptic Density/Architecture at Excitatory Synapses; Spines and Mental Disorders.

Further Reading

Cheng D, Hoogenraad CC, Rush J, et al. (2006) Relative and absolute quantification of postsynaptic density proteome isolated from rat forebrain and cerebellum. *Molecular and Cellular Proteomics* 5: 1158–1170.

Craig AM, Graf ER, and Linhoff MW (2006) How to build a central synapse: Clues from cell culture. *Trends in Neurosciences* 29: 8–20.

Ethell IM and Pasquale EB (2005) Molecular mechanisms of dendritic spine development and remodeling. *Progress in Neurobiology* 75: 161–205.

Friedman HV, Bresler T, Garner CC, and Ziv NE (2000) Assembly of new individual excitatory synapses: Time course and temporal order of synaptic molecule recruitment. *Neuron* 27: 57–69.

Gerrow K, Romorini S, Nabi SM, Colicos MA, Sala C, and El-Husseini A (2006) A preformed complex of postsynaptic proteins is involved in excitatory synapse development. *Neuron* 49: 547–562.

Kennedy MB (2000) Signal-processing machines at the postsynaptic density. *Science* 290: 750–754.

Kim E and Sheng M (2004) PDZ domain proteins of synapses. *Nature Reviews Neuroscience* 5: 771–781.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* 25: 103–126.

Sheng M and Sala C (2001) PDZ domains and the organization of supramolecular complexes. *Annual Review of Neuroscience* 24: 1–29.

Sugiyama Y, Kawabata I, Sobue K, and Okabe S (2005) Determination of absolute protein numbers in single synapses by a GFP-based calibration technique. *Nature Methods* 2: 677–684.

Post-Tetanic Potentiation (PTP)

R S Zucker, University of California, Berkeley, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

The strength of synaptic transmission can increase or decrease during repetitive presynaptic activity. These changes, usually due to variations in amount of transmitter released by each action potential, dynamically shape the information transfer characteristics of synapses.

Components of Activity-Dependent Synaptic Plasticity

Phases of Enhanced Synaptic Transmission

Chemical synapses are rarely static transmitters of information between neurons. Their effectiveness waxes and wanes, depending upon frequency of stimulation and history of prior activity. At most synapses, repetitive high-frequency action potentials (often called a tetanus) result initially in a rapid growth of successive postsynaptic potential (PSP) amplitudes, called synaptic facilitation (or sometimes frequency facilitation). This process builds to a steady-state within about 1 s and decays equally rapidly when stimulation stops. Decay is measured by single test stimuli given at various intervals after a conditioning train. Facilitation can often be divided into two exponential phases, lasting tens and hundreds of milliseconds, respectively, and called the first and second components of facilitation. It may reach appreciable levels, as high as a doubling or more of the PSP amplitude in the second of two responses at short intervals, and as much as nearly 100-fold increase in PSP amplitude in a high-frequency train in some synapses. If presynaptic activity is sustained, facilitation is followed at most synapses by a slower phase of increased efficacy called augmentation, with a characteristic time constant of several seconds. Finally, prolonged activity results in a final phase of PSP growth, called potentiation and lasting several minutes. These processes can each increase transmission a further several-fold over initial baseline levels. The combination of all three processes, at least some of which (facilitation and potentiation in particular) appear to interact multiplicatively, could in principle elevate synaptic transmission several 100-fold, but in practice saturation of the release mechanisms usually limits the overall enhancement to less than 100-fold.

Synaptic Depression

Often superimposed on these phases of enhanced synaptic strength is a phase of diminishing transmission called synaptic depression. This process is characterized by a drop in PSP amplitude during repetitive presynaptic action potentials. Depression can be absent or undetectable, quite modest, or so severe that transmission drops to zero. Depression often tends to overlap and obscure the augmentation and potentiation phases. When repetitive activity ceases, recovery from the various components of short-term synaptic plasticity occurs in the same order in which they developed during the previous activity, with facilitation decaying first, then depression and augmentation, and finally potentiation (Figure 1). Thus, potentiation is often visible as a distinct phase only long after repetitive activity or a tetanus, and so is usually called posttetanic potentiation (PTP).

Short-Term Plasticity due to Changes in Transmitter Release

Presynaptic Locus Implicated by Quantal Analysis

There are many points at which synapses can be modulated to alter the strength of synaptic transmission. The possibilities may initially be divided into presynaptic and postsynaptic processes. A good way to distinguish these two classes is to make use of the fact that synaptic transmission is 'quantized,' in the sense that each PSP represents the summed effect of the fusion of a certain number of presynaptic vesicles, each of which causes a reasonably fixed 'unitary' postsynaptic response. The number of vesicles that contribute to synaptic transmission is not fixed, even in the complete absence of synaptic plasticity, but rather is a random variable that fluctuates about some average number or mean. By statistical analysis of these fluctuations, it is possible to determine the number of unitary components in a PSP. Furthermore, even in the absence of presynaptic action potentials, vesicles occasionally fuse spontaneously with the plasma membrane, and the secretion of the contents of a single vesicle can be observed as a spontaneously occurring unitary or miniature PSP (MPSP). Comparison of the size of MPSPs to full PSPs evoked by action potentials also allows estimation of the number of vesicles contributing to a synaptic response.

This kind of analysis, called quantal analysis, has been applied to a large number of instances of short-term synaptic plasticity, including facilitation, augmentation, potentiation, and depression in humans, lower mammals, birds, amphibians, crustaceans, mollusks, insects and annelids. In practically all cases, every kind of short-term plasticity is accompanied by changes in the number of quanta released, often called quantal

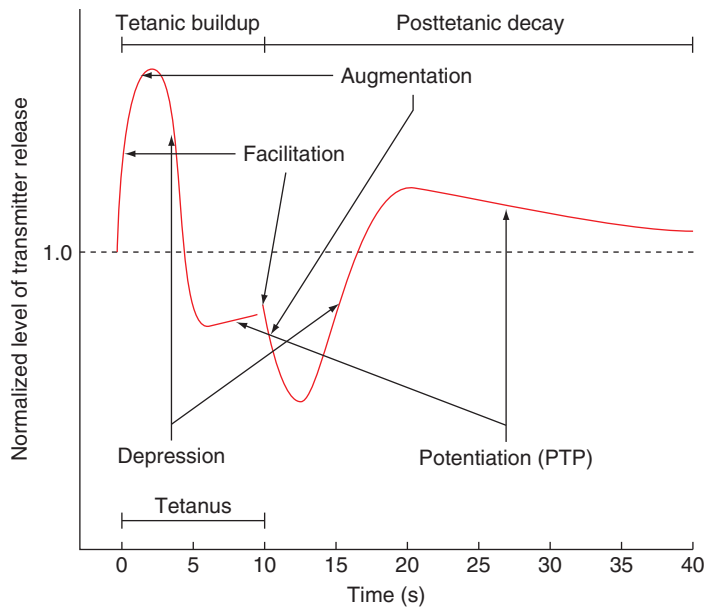

Figure 1 The red traces show the different kinetic components of short-term synaptic plasticity. On the left, transmission is shown as increasing early in a tetanus due to facilitation and then augmentation, and then as decreasing due to depression, and finally increasing again as potentiation is expressed after depression has reached a steady state. This curve represents the successive responses to action potentials during continuous stimulation. After the tetanus is stopped, recovery is measured by single action potentials at various times after the end of the tetanus, shown on the right. At very short intervals, facilitation is slightly higher than the last tetanic stimulus (because the inter-pulse interval is reduced from that during the tetanus), but facilitation decays rapidly, and augmentation more slowly. This is followed by recovery from depression, leaving only PTP. Different synapses have different mixes of facilitation, augmentation, depression, and PTP, and depression may develop and recover more or less rapidly than in this example.

content, with little or no change in the size or postsynaptic effectiveness of the unitary response to a quantum of transmitter. Thus, short-term synaptic plasticity is almost always a consequence of changes in one or more presynaptic process involved in determining the number of vesicles that action potentials cause to be secreted during neurotransmission.

Changes in Statistics of Transmitter Release

The statistics of transmitter release closely resemble those of a binomial process, in which an average of m quanta is released from a pool of n available quanta, each with a probability p of release by an action potential. The values of n obtained in such analyses are usually similar to or somewhat less than the number of anatomical synapses observed ultrastructurally. Each such synaptic contact, or active zone as it is sometimes called, consists of a cluster of presynaptic vesicles, some docked at the plasma membrane, opposed to a cluster of postsynaptic receptors across the synaptic cleft. The approximate correspondence between n and number of active zones suggests that each active zone can only release one docked vesicle in response to an action potential. If all the apparently docked vesicles could be released independently and simultaneously, n would greatly exceed the number of active zones, since it is usually clear that postsynaptic receptors are not saturated by a quantum, so that multiple quanta released at one site would summate in their effects and be readily detected. Studies of synapses consisting of a single active zone confirm that they are normally capable of releasing only a single vesicle.

Short-term synaptic plasticity is usually accompanied by parallel changes in both n and p. The likely explanation for this is that p is not uniform across all active zones (as assumed in the simple binomial model), and may be near zero for some sites that rarely release a vesicle. Synaptic enhancement increases the probability of release from active zones, and raises previously silent active zones to a level of activity in which they begin to participate in transmission. Depression has the opposite effect – fewer active zones participate effectively in the release process, and those still active have a reduced probability of releasing a vesicle.

Possible Presynaptic Mechanisms

Much experimental effort has gone into exploring presynaptic mechanisms that could underlie facilitation, augmentation, and potentiation. The first step in

synaptic transmission is the invasion of the action potential into presynaptic terminals. In a few preparations, facilitation has been found to be caused by action potential broadening, arising mainly from accumulating inactivation of K^+ channels involved in terminating action potentials. In these synapses, increases in Ca^{2+} influx in prolonged action potentials underlie facilitation of transmitter release.

At most synapses that have been studied, short-term synaptic plasticities occur in the absence of any changes in presynaptic action potential propagation or waveform in presynaptic terminals and can be elicited by trains of constant presynaptic depolarizing pulses. A second possible locus of modulation is the presynaptic Ca^{2+} channel. In a few instances, presynaptic Ca^{2+} channels have been found to show a form of facilitation, with increased duration or probability of opening to repeated depolarizations, but this process contributes little to enhanced synaptic transmission at neurons, although it appears to be involved in facilitation of hormonal secretion from adrenal chromaffin cells.

The Role of Ca^{2+} in Short-Term Synaptic Enhancement

That facilitation, augmentation, and potentiation depend upon presynaptic Ca^{2+} entry is now well established from experiments on several different preparations. The pioneering experiments of Bernard Katz and Ricardo Miledi demonstrated that an action potential that invades a nerve terminal in a Ca^{2+}-free medium not only evokes no transmitter release but also fails to facilitate release to closely following action potentials stimulated after the rapid reintroduction of Ca^{2+} to nerve terminals by local iontophoresis. Facilitation does not require that the first action potential trigger secretion, however, since action potentials that happen to cause no release (at low external $[Ca^{2+}]$ levels) are still effective in inducing facilitation. Experiments on augmentation and PTP also show that these longer-lasting forms of enhanced synaptic transmission depend on Ca^{2+} influx into nerve terminals during conditioning stimulation. Experimental elevation of presynaptic $[Ca^{2+}]_i$ levels has also been shown to substantially increase the number of quanta released by an action potential.

Much attention has focused on how Ca^{2+} ions induce these forms of plasticity. A number of possibilities exist.

The Single Site Summating Residual Ca^{2+} Hypothesis

The simplest possibility is that all forms of synaptic enhancement arise from the cooperative action of four or more Ca^{2+} ions in triggering secretion. This cooperativity is inferred from the finding that the magnitude of the phasic release of transmitter (i.e., the number of vesicles or quanta released) by an action potential depends on approximately the fourth power of external $[Ca^{2+}]$. Secretion can also be evoked by the rapid elevation of $[Ca^{2+}]_i$ liberated by intense flash irradiation of presynaptically injected photosensitive Ca^{2+} chelators; again, the maximal rate of secretion is related to the fourth power of the peak level of $[Ca^{2+}]_i$. This highly nonlinear dependence of secretory rate on $[Ca^{2+}]_i$ could naturally lead to increased transmitter release in the presence of a residual increase in $[Ca^{2+}]_i$ after action potentials. This is best seen by using a simple hypothetical calculation: suppose an increase of $[Ca^{2+}]_i$ by 1 μmol l^{-1} normally underlies an unfacilitated PSP consisting of 10 quanta, or the vesicle fusions occurring in 1 ms, and 20 ms later $[Ca^{2+}]_i$ is still elevated by 0.1 μmol l^{-1} above resting level (about 50 nmol l^{-1}). Then another action potential following the first by 20 ms might increase $[Ca^{2+}]_i$ to a peak level of 1.1 μmol l^{-1} above the resting level, which owing to the fourth-power dependence of secretion on $[Ca^{2+}]_i$ would facilitate release so that the second PSP is 46% larger (in quantal number) than the first ($1.1^4 - 1$). But the residual $[Ca^{2+}]_i$ itself would only cause secretion to occur at a rate of 1 quantum s^{-1} (10×0.1^4 ms^{-1}). Such modest increases in MPSP frequency are actually observed during facilitation, augmentation, and potentiation, and have been regarded as evidence for this hypothesis. The different phases of enhanced transmission could arise from different phases of calcium accumulation during nerve activity and its subsequent decay, with rapid phases limited by calcium diffusion away from docked vesicles near Ca^{2+} channels and slow phases representing effects of Ca^{2+} binding to intracellular buffers, Ca^{2+} being sequestered into Ca^{2+}-storing organelles such as endoplasmic reticulum and mitochondria, and Ca^{2+} extrusion from nerve terminals by pumps such as the Ca^{2+}-dependent ATPase and Na^+/Ca^{2+} exchanger at the plasma membrane.

The trouble with this simple and elegant hypothesis is that it should not be possible to obtain the very high levels of posttetanic facilitation/augmentation/potentiation sometimes observed (up to 100-fold increase) without a massive increase in MPSP frequency to levels substantially higher than the maximum rate of release to a single action potential (two times higher for an enhancement of 100: $100^{1/4} - 1$). Such predicted large increases in MPSP frequency are not observed.

The situation is even worse now that it is possible to estimate the peak $[Ca^{2+}]_i$ level needed to trigger phasic

secretion and the residual $[Ca^{2+}]_i$ present during periods of facilitation, augmentation, or potentiation. The former can be obtained from measurements of local $[Ca^{2+}]_i$ elevations using low-affinity indicators, calculations of the effects of presynaptically injected Ca^{2+} buffers on secretion, and the elevations of $[Ca^{2+}]_i$ released by photolysis of photosensitive chelators needed to mimic secretory rates triggered by action potentials. From all these methods, about 20 μmol l^{-1} $[Ca^{2+}]_i$ appears to be required to trigger a typical PSP. On the other hand, measurements of residual $[Ca^{2+}]_i$ during facilitation, augmentation, and PTP reveal levels of about 1 μmol l^{-1}, which would enhance release by only 22% maximally ($1.05^4 - 1$) – far less enhancement than any of these processes is capable of displaying at many synapses. However, at some synapses displaying only very modest facilitation, this nonlinear effect of residual $[Ca^{2+}]_i$ with the $[Ca^{2+}]_i$ transient during an action potential does appear to be responsible for the plasticity observed.

Independent Site Hypotheses

These considerations led to proposals that presynaptic Ca^{2+} acts in eliciting short-term synaptic enhancement at molecular targets different from the target involved in phasically triggering exocytosis. Two versions of this hypothesis may be distinguished. The first proposes that Ca^{2+} entering during action potentials binds to sites that enhance transmitter release to subsequent action potentials, and that the duration of each phase of enhancement reflects the binding time (dissociation rate) of Ca^{2+} from a different site. The second version proposes that synaptic enhancement is due to the continuing action of residual Ca^{2+} ions acting in equilibrium at one or more sites that enhance transmitter release to subsequent action potentials. The different phases of enhancement could reflect the kinetics of different processes removing residual Ca^{2+}. In support of the latter proposal is the finding in a number of synapses that facilitation, augmentation, and potentiation decay in parallel with phases of decay of residual Ca^{2+} in active zones.

The Residual Bound Ca^{2+} Hypothesis

One way to distinguish these possibilities is by use of exogenous Ca^{2+} buffers. Slowly binding buffers (e.g., ethylene glycol bis(*β*-aminoethyl ether)-*N*,*N* tetraacetic acid (EGTA)), or modest concentrations of fast-acting buffers (e.g., 1,2-bis(*o*-aminophenoxy) ethane-*N*,*N*,*N*′,*N*′-tetraacetic acid (BAPTA)) ought to be able to hold residual rises in $[Ca^{2+}]_i$ to a minimum, while having little effect on the peak $[Ca^{2+}]_i$ concentrations at docked vesicles near Ca^{2+} channels that trigger exocytosis. Synaptic enhancement due to residual bound Ca^{2+} would be unaffected, while effects due to the continuing action of residual free Ca^{2+} would be blocked. However, experiments with these buffers loaded into presynaptic terminals have yielded variable results. Augmentation and PTP are usually (but not absolutely always) blocked, while facilitation is sometimes blocked and sometimes not. Such variable results may reflect different degrees of success in blocking residual $[Ca^{2+}]_i$ elevations, which have rarely been monitored in these experiments, and different degrees of saturation of the exogenous buffers by the $[Ca^{2+}]_i$ levels reached during action potentials in active zones in different preparations.

The Residual Free Ca^{2+} Hypothesis

In one study, the effect of exogenous BAPTA was monitored, and shown to eliminate the accumulation of residual Ca^{2+} in short trains causing facilitation. Facilitation was greatly reduced, and what was left had a much shorter time constant (40 ms). This was probably due to the effect of Ca^{2+} entering during the action potential on the facilitation site, suggesting that facilitation is a process which itself lasts less than 100 ms and is normally rate-limited by the decay of $[Ca^{2+}]_i$ within active zones, from which Ca^{2+} will diffuse with apparent time constants of tens and hundreds of milliseconds. It is still unclear whether the two components of facilitation are due to the multi-exponential time course of Ca^{2+} removal, or to actions on different molecular targets.

In another study, residual Ca^{2+} was rapidly reduced following conditioning stimulation by brief illumination of a photosensitive buffer that takes up Ca^{2+} on photolysis. Facilitation then disappeared within 10 ms after an action potential, while augmentation and PTP were both truncated to a process lasting less than a second. Since augmentation and PTP also show a similar dependence of magnitude of enhancement of transmitter release on submicromolar levels of residual $[Ca^{2+}]_i$, augmentation and potentiation appear to be due to residual Ca^{2+} acting at a high-affinity site distinct from the sites involved in either facilitation (intrinsically decaying more rapidly) or exocytosis (having a much lower Ca^{2+} affinity than either facilitation or augmentation/ PTP).

Saturation of Endogenous Ca^{2+} Buffers – A Special Form of the Second Site Hypothesis

Calcium is buffered in cytoplasm by a variety of proteins with different capacities and affinities for Ca^{2+} ions. At some synapses, an important part of the buffering is accomplished by high-affinity buffers such as calbindin-D28k that are easily saturated by

low levels of residual $[Ca^{2+}]_i$. Following one or more action potentials, less of this buffer is available to capture Ca^{2+} ions entering through Ca^{2+} channels, so that a higher local $[Ca^{2+}]_i$ triggers a facilitated release of vesicles docked near those channels. This process seems to occur at inhibitory synapses from mouse multipolar interneurons onto cortical pyramidal cells and at excitatory mossy fiber synapses onto CA3 hippocampal pyramidal cells. However, such a mechanism failed to account for facilitation at rat somatosensory cortical pyramidal synapses onto bitufted interneurons, or at Schaffer collateral afferents onto CA1 pyramidal neurons, where a different molecular Ca^{2+} target modulating transmitter release appears to be implicated.

Molecular Targets in Short-Term Synaptic Enhancement

The sites of Ca^{2+} action in facilitation, and in augmentation/ PTP, have generally not been identified. The rapid rise of facilitation after each action potential and the similarity of its kinetics to the expected diffusion of Ca^{2+} from active zones suggest that it is activated by Ca^{2+} acting at the level of one to a few micromolar on targets on or near docked vesicles. Neuronal Calcium Sensor-1 has been identified as a strong candidate for such a target at excitatory synapses between dissociated hippocampal neurons in culture, but it is not certain whether NCS-1 acts as a Ca^{2+} buffer or a regulator of synaptic transmission. Augmentation/PTP may result from Ca^{2+} acting at more distant sites, for example, on processes that mobilize vesicles to docking sites and ready them for release. Synapses formed between cultured hippocampal neurons seem to rely on Ca^{2+} binding to calmodulin and munc-13 to produce augmentation, which is therefore a strong candidate molecular target for generating augmentation at these synapses.

The Persistence of PTP

If augmentation and potentiation are due to Ca^{2+} acting at the same site with subsecond kinetics and high affinity, why does PTP last so much longer than augmentation? The answer is that both rise with the accumulation of Ca^{2+} in nerve terminals and last until residual Ca^{2+} is removed. Two processes slow down the removal of Ca^{2+} after prolonged activity. (1) Ca^{2+} gradually accumulates in mitochondria during sustained activity, and slowly leaks out afterwards into the cytoplasm. The gradual loading of mitochondria leads to a slow rise in cytoplasmic Ca^{2+} during activity, and the long time it takes to fully unload mitochondria results in the prolonged small posttetanic residual Ca^{2+} responsible for PTP. Blocking mitochondrial uptake of Ca^{2+} eliminates PTP. (2) Presynaptic $[Na^+]_i$ rises and reduces the rate of Ca^{2+} extrusion by Na^+/Ca^{2+} exchange. This exchanger uses the normally high external-to-internal sodium gradient as the energy source for removing Ca^{2+}, and reducing this gradient retards the pump. This is why procedures that increase the loading of presynaptic terminals with Na^+ ions magnify and prolong PTP.

Modulation of Short-Term Synaptic Enhancement

Facilitation, augmentation, and potentiation can be differently affected by divalent ion substitutions for Ca^{2+}, hormone treatments, neurotrophic factors, and mutations in genes for Ca^{2+}/calmodulin-dependent protein kinase, synapsins I and II, cyclic AMP phosphodiesterase, Ca^{2+}/calmodulin-dependent adenylyl cyclase, and the type 4 metabotrophic glutamate receptor. In one case, postsynaptic hyperpolarization and $[Ca^{2+}]_i$ chelation are also reported to affect PTP, perhaps owing to some sort of retrograde messenger. These manipulations may differentially affect the Ca^{2+} targets in activating short-term synaptic plasticity, or processes involved in the clearance of residual Ca^{2+} ions. Either way, they indicate a modifiability of synaptic plasticity that could have important effects on neural circuitry with plastic synapses (see below).

The Consequences of Short-Term Synaptic Plasticity

The dynamic reshaping of synaptic strength is an essential property of synapses, and is fundamental to the way they respond to inputs, selectively transmit signals, and shape the responses of nerve circuits and the activity of neural networks. Depression makes synapses respond adaptively, reporting mainly changes in signal strength, while facilitation and augmentation allow synapses to distinguish significant signals from noise when the former are represented as sustained or high-frequency activity. These properties can be used by neural circuits to distinguish different patterns of input. The shorter forms of synaptic enhancement, as well as PTP, can also effectively adjust the gain or sensitivity of neuronal circuits following prior activity and lead to alterations in neural circuit responsiveness that last for seconds or minutes. Neuroscientists trying to understand how neural circuits analyze sensory information or produce motor patterns are now paying a good deal of attention to the important consequences of short-term synaptic plasticity for the integrative functions of the nervous system.

See also: Calcium Channels; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Plasticity and Activity-Dependent Regulation of Gene Expression; Presynaptic Facilitation; Presynaptic Inhibition; Schwann Cells and Plasticity of the Neuromuscular Junction; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Depression; Synaptic Plasticity: Short-Term Mechanisms; Voltage-Gated Calcium Channels.

Further Reading

Atluri PP and Regehr WG (1996) Determinants of the time course of facilitation at the granule cell to Purkinje cell synapse. *Journal of Neuroscience* 16: 5661–5671.

Blatow M, Caputi A, Burnashev N, Monyer H, and Rozov A (2003) Ca^{2+} buffer saturation underlies paired pulse facilitation in calbindin-D28k-containing terminals. *Neuron* 38: 79–88.

Felmy F, Neher E, and Schneggenburger R (2003) Probing the intracellular calcium sensitivity of transmitter release during synaptic facilitation. *Neuron* 37: 801–811.

Junge HJ, Rhee JS, Jahn O, et al. (2004) Calmodulin and Munc13 form a Ca^{2+} sensor/effector complex that controls short-term synaptic plasticity. *Cell* 118: 389–401.

Kamiya H and Zucker RS (1994) Residual Ca^{2+} and short-term synaptic plasticity. *Nature* 371: 603–606.

Katz B and Miledi R (1968) The role of calcium in neuromuscular facilitation. *Journal of Physiology* 195: 481–492.

Matveev V, Sherman A, and Zucker RS (2002) New and corrected simulations of synaptic facilitation. *Biophysical Journal* 83: 1368–1373.

Matveev V, Zucker RS, and Sherman A (2004) Facilitation through buffer saturation: Constraints on endogenous buffering properties. *Biophysical Journal* 86: 2691–2709.

Rosenmund C, Sigler A, Augustin I, et al. (2002) Differential control of vesicle priming and short-term plasticity by Munc13 isoforms. *Neuron* 33: 411–424.

Rozov A, Burnashev N, Sakmann B, and Neher E (2001) Transmitter release modulation by intracellular Ca^{2+} buffers in facilitating and depressing nerve terminals of pyramidal cells in layer 2/3 of the rat neocortex indicates a target cell-specific difference in presynaptic calcium dynamics. *Journal of Physiology* 531: 807–826.

Sippy T, Cruz-Martin A, Jeromin A, and Schweizer FE (2003) Acute changes in short-term plasticity at synapses with elevated levels of neuronal calcium sensor-1. *Nature Neuroscience* 6: 1031–1038.

Tang Y and Zucker RS (1997) Mitochondrial involvement in post-tetanic potentiation of synaptic transmission. *Neuron* 18: 483–491.

Zhong N, Beaumont V, and Zucker RS (2001) Roles for mitochondrial and reverse mode Na^{+}/Ca^{2+} exchange and the plasmalemma Ca^{2+} ATPase in post-tetanic potentiation at crayfish neuromuscular junctions. *Journal of Neuroscience* 21: 9598–9607.

Zucker RS (1989) Short-term synaptic plasticity. *Annual Review of Neuroscience* 12: 13–31.

Zucker RS, Kullmann DM, and Schwarz TL (2004) Release of neurotransmitters. In: Byrne JH and Roberts JL (eds.) *From Molecules to Networks: An Introduction to Cellular and Molecular Neuroscience*, pp. 197–244. San Diego, CA: Academic Press /Elsevier.

Zucker RS and Regehr WG (2002) Short-term synaptic plasticity. *Annual Review of Physiology* 64: 355–405.

Posttraumatic Stress Disorder as an Emotional Disorder

R K Pitman, Harvard Medical School, Charlestown, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Because psychopathologic phenomena often represent extreme forms of ordinarily adaptive human behavior, evolutionary neurobiologic mechanisms may be cast in higher relief in the mentally disordered. Darwin noted that the behavior of insane patients reveals the 'brute nature' within human beings. Behavior associated with mental disorders constitutes potentially useful data that are often overlooked by basic scientists. The rediscovered mental disorder known as posttraumatic stress disorder (PTSD) provides illustrations of human mechanisms of emotion, conditioning, and memory that can inform neuroscience. PTSD has also emerged as the model psychiatric disorder for studying the effect of psychological environmental events on the nervous system.

PTSD and Conditioned Fear

The current clinical definition of PTSD according to the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (DSM-IV) appears in abbreviated form in Table 1. This article discusses the basic elements of classical (Pavlovian) conditioning that it incorporates. After a traumatic event (unconditioned stimulus, or UCS) has induced an intense emotional response of fear, helplessness, or horror (unconditioned response, or UCR), external and internal cues that were present at the time of the traumatic event (conditioned stimuli, or CSs) acquire the capacity, through their association with the traumatic experience, to evoke intense emotional conditioned responses (or CRs) on subsequent occasions. Avoidance of emotionally distressing and arousing trauma-related cues reflects secondary operant conditioning. In addition to the specific symptoms of PTSD (B.1–5, C.1–2 in Table 1) that are related to conditioning and uniquely define this disorder, there are several generic psychiatric symptoms (C.4–7, D.1–5) that PTSD shares with other anxiety and/or depressive disorders.

Conditioned Psychophysiologic Responses to Trauma-Related Cues

DSM-IV PTSD criterion B.5, "physiological reactivity upon exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event," has been amply supported in the laboratory. Several studies have shown that audiovisual cues of battle situations, such as pictures of ground troops unloading from helicopters or sounds of machine gun fire, produce larger peripheral physiologic, including heart rate (HR), skin conductance (SC), blood pressure (BP), and facial electromyogram (EMG), increases in veterans with PTSD. When individuals with PTSD recall past traumatic events via internal, script-driven imagery, they also produce larger physiologic responses compared with individuals who experienced similarly stressful events but did not develop the disorder. This finding has held for various types of traumatic events including combat in Vietnam, Korea, and World War II; terrorist attacks; automobile accidents; and childhood sexual abuse. Results of a large multisite study of 1328 Vietnam veterans that combined audiovisual cues and script-driven imagery confirmed heightened physiologic reactivity to trauma-related cues in the veterans with PTSD.

Trauma-Related Brain Activation Studies

Neuroimaging studies of PTSD have directly examined changes in brain activity that presumably underlie emotional re-experiencing in PTSD. A seminal study using positron emission tomography (PET) employed personal script-driven imagery to study individuals with PTSD stemming from a variety of traumatic events. During traumatic imagery compared to neutral imagery, there were significant increases in regional cerebral blood flow (rCBF) in anterior paralimbic structures, including posterior medial orbitofrontal, insular, anterior temporal, medial temporal, and anterior cingulate cortices, as well as amygdala. However, because this study did not include a non-PTSD comparison group, it was not possible to determine the specificity of the findings to PTSD. A follow-up study examined rCBF during script-driven imagery in women with histories of childhood sexual abuse with and without PTSD. PTSD subjects showed significantly greater activation than non-PTSD subjects during personal sexual abuse imagery compared to neutral imagery in orbitofrontal cortex and anterior temporal pole. Anterior cingulate cortex (ACC), which is a component of ventromedial prefrontal cortex, however, was activated in the non-PTSD group but not in the PTSD group. In a study that employed single photon emission computed

Table 1 Abridged DSM-IV diagnostic criteria for PTSD

A. Traumatic event/response (both required)
   1. An event that threatens or inflicts serious injury to self or others
   2. A response of intense fear, helplessness, or horror
B. Re-experiencing criteria (one required)
   1. Recurrent, intrusive, distressing recollections of the traumatic event
   2. Recurrent, distressing dreams of the traumatic event
   3. Flashbacks to the traumatic event
   4/5. Intense psychological distress/physiological reactivity on exposure to traumatic event-related cues
C. Avoidance/numbing criteria (three required)
   1. Of thoughts and feelings associated with the trauma
   2. Of activities, places, or people associated with the trauma
   3. Inability to recall an important aspect of the trauma
   4. Markedly diminished interest in significant activities
   5. Feelings of detachment or estrangement from others
   6. Restricted range of affect (numbing)
   7. Sense of a foreshortened future
D. Arousal criteria (two required)
   1. Insomnia
   2. Irritability or outbursts of anger
   3. Difficulty concentrating
   4. Hypervigilance
   5. Exaggerated startle response
E. Duration at least one month (required)
F. Distress/Impairment, for example, social, occupational (required)

tomography (SPECT) in Vietnam veterans with and without PTSD and normal control subjects, the PTSD group exhibited greater activation in the left amygdaloid region in response to combat sounds.

More recently, again using PET and script-driven imagery, it was found that male and female Vietnam veterans with PTSD exhibited rCBF decreases in medial frontal gyrus in a traumatic versus neutral mental imagery comparison. Moreover, male combat veterans also exhibited rCBF increases in the left amygdala. In both males and females, rCBF changes in medial frontal gyrus were inversely correlated with rCBF changes in the left amygdala and the right amygdala/periamygdaloid cortex. Furthermore, in the traumatic condition, for both subgroups with PTSD, symptom severity was positively related to rCBF in the right amygdala and negatively related to rCBF in medial frontal gyrus. These results suggest a reciprocal relationship between medial prefrontal cortex and amygdala function in PTSD. The script-driven imagery procedure has also been adapted to functional magnetic resonance imaging (fMRI), which has revealed deficient ACC activation, as measured by blood oxygenation level-dependent (BOLD) signal, in persons with PTSD. Another fMRI study found the absence of BOLD signal increases in rostral ACC in response to the presentation of combat-related words in the context of the emotional Stroop task in combat veterans with PTSD compared to combat-exposed, non-PTSD veterans.

Conditionability

Conditioning theory provides an explanation as to how trauma-related cues might acquire the capacity to generate heightened physiologic responses in individuals with PTSD, but it does not explain why, upon exposure to comparably stressful events (i.e., comparable UCSs), some individuals develop PTSD whereas others do not. To investigate this, researchers performed a *de novo* differential conditioning study. CSs were colored circles, and the UCS was a 'highly annoying' electrical stimulus. The PTSD group showed larger differential HR, SC, and EMG responses to the reinforced versus nonreinforced stimuli (CS+ vs. CS−) compared with the non-PTSD group. Fearfulness increased for the PTSD group as the CS+ acquired signal value for predicting the UCS, whereas for individuals without PTSD, fearfulness decreased as the UCS became more predictable. Only PTSD participants continued to show differential SC responses to CS+ versus CS− during extinction trials. Results suggest that individuals with PTSD form new conditioned fear responses more strongly and extinguish them more poorly than do trauma-exposed individuals without PTSD.

Whereas the amygdala is critically involved in the acquisition of conditioned fear, a critical structure for the extinction of conditioned fear response in rodents is the infralimbic cortex, which inhibits the expression of conditioned fear by the amygdala. A homolog

of the infralimbic cortex, at least in part, is the ACC, especially its pregenual and subcallosal components. In a recent PET study of women with PTSD from childhood sexual abuse, rCBF was measured during acquisition and extinction in another *de novo* conditioning study. Compared to non-trauma-exposed controls, PTSD patients had increased left amygdala activation with fear acquisition, and decreased ACC activation during extinction. These findings implicate amygdala and ACC in the acquisition and extinction of fear responses, respectively, in PTSD.

Unconditioned Physiologic Responses

Sometimes it is difficult to identify a role for a conditioned stimulus in PTSD patients' reactions. Indeed, a problem faced by many of them is that their emotional responses are not conditional enough. Unable to confine their fear and other emotional responses to reminders of the traumatic situation, they find them spilling over into excessive, inappropriate reactions to ordinary events. A probe that has proved useful in the study of unconditioned responding in PTSD is the startle response. Indeed, 'exaggerated startle' is one of the diagnostic criteria for PTSD (Table 1, criterion D.5). In studies of the startle response in PTSD, the evocative stimulus is usually a sudden, loud pure tone or white noise that does not have any plausible associative connection to the traumatic event. Baseline measures of the muscular, for example, eyeblink, EMG response in PTSD have not consistently confirmed the presence of exaggerated startle in PTSD. More consistent EMG results have been obtained when PTSD subjects have been studied under stressful experimental conditions, suggesting an enhanced sensitivity to generally stressful contexts. Greater evidence of increased baseline startle has been provided by autonomic measures. Several studies have confirmed increased heart rate responses to sudden, loud tones in PTSD, which may indirectly reflect increased amygdala reactivity.

Two longitudinal studies have suggested disparate origins of increased autonomic startle responses in PTSD. One evaluated responses to sudden, loud tones in individuals who arrived at a hospital emergency room after psychologically traumatic events. Although those who went on to develop PTSD at 4 months posttrauma showed comparable physiologic startle responses to those who did not develop PTSD when startale was measured 1 week posttrauma, at 1 and 4 months posttrauma, the subjects with PTSD showed a greater heart rate response and required more stimulus trials to reach the criteria of skin conductance and orbicularis oculi electromyogram nonresponse. These results suggest that differences in physiologic response to startling tones incubate after the occurrence of the traumatic event. However, in a prospective study of the auditory startle response in firefighters before and after trauma exposure, pretrauma skin conductance response to startling stimuli was found to be predictive of posttrauma PTSD severity. These latter results support elevated startle response's being a vulnerability factor for PTSD. Clearly further studies are necessary. (The issue of the origin of biologic abnormalities in PTSD is also addressed below.)

Direct Studies of Amygdala Reactivity in PTSD

The 'masked faces' task has been used to measure amygdala reactivity in combat veterans with PTSD. During fMRI, investigators presented fearful and happy faces temporally masked by neutral faces. Normal subjects are typically aware of seeing only the neutral faces, although their amygdalae are activated by the fearful faces. Veterans with PTSD exhibited greater amygdala activation, as indexed by BOLD signal changes, compared to combat-exposed, non-PTSD veterans. Moreover, the magnitude of amygdala activation significantly correlated with PTSD severity. Another 'overt faces' fMRI experiment administered the same fearful and happy faces to trauma-exposed male Vietnam veterans or firefighters with and without PTSD, although the facial stimuli were not masked. The PTSD group again exhibited exaggerated amygdala responses, as well as diminished medial prefrontal cortical responses, to the fearful versus happy facial expressions. In addition, in the PTSD group, activation in the amygdala was negatively correlated with activation in the medial prefrontal cortex, and symptom severity was also negatively related to medial prefrontal cortical activation. These results provide further evidence for exaggerated amygdala responsivity, diminished medial prefrontal cortical responsivity, and a reciprocal relationship between these two regions, similar to that reported during script-driven imagery.

Event-Related Potentials

The most widely studied event-related potential component in normal and clinical populations is the parietal P300 component, or P3b. In the 'auditory oddball' procedure, the P3b response to target tones presumably reflects increased neural activity associated with effortful attention to the processing of salient information. Reduced P3b is typically interpreted as indicative of a generalized, attention-related cognitive deficit. Several studies of PTSD from various kinds of traumatic events in men and women have

reported smaller P3b responses to target stimuli. These findings have been interpreted as electrophysiologic evidence for difficulty concentrating, one of the diagnostic symptoms of PTSD (Table 1, criterion D.3).

Neurologic Soft Signs

Neurologic soft signs (NSSs) are various subtle indices of central nervous system dysfunction. It has been found that male combat veterans, as well as adult females sexually abused as children, with chronic PTSD are more neurologically compromised than their counterparts without PTSD, as evidenced by more NSSs. Even though NSSs by definition are difficult to localize to specific brain areas, their increase in PTSD suggests the presence of underlying neuroanatomic abnormalities. Structural magnetic resonance imaging (sMRI) provides an opportunity to calculate volumes of brain structures or areas by manually delimiting the boundaries of the structure on a series of slices and then integrating the areas into a volume, as well as by the use of other semi-automated and automated techniques.

Hippocampus

Volumetric studies of persons with PTSD that have employed sMRI have focused heavily on the hippocampus, a structure that has been implicated in declarative memory as well as spatial and contextual learning. Indeed, several studies have found declarative memory impairments in PTSD. The research focus on hippocampus has also been driven by animal research indicating that chronic stress can damage this brain structure. There have by now been more than a dozen studies that have reported decreased volume of the hippocampus in persons with PTSD. In some of these studies, the volumetric diminution has been accompanied by reduced declarative memory capacity, in others not. Although nearly as many other studies have failed to find smaller hippocampi, few if any have reported larger hippocampal volume in PTSD. A recent meta-analysis of hippocampal volumetric studies revealed significantly smaller volume in both right and left hippocampi in adult subjects with chronic PTSD, in comparison with both healthy controls and traumatized controls, establishing beyond doubt that diminished hippocampal volume is a bona fide feature of PTSD, albeit not a highly sensitive one. Another neuroimaging technique, magnetic resonance spectroscopy (MRS), has identified diminution of *N*-acetylaspartate (NAA), a neuronal viability marker, even in the hippocampi of persons with posttraumatic stress disorder who do not show volumetric diminution. This finding implies that there are structural hippocampal abnormalities in PTSD that may not rise to the level of gross anatomic differences. Ultimately, metabolite changes may become more sensitive indicators of PTSD neuropathology than volume losses.

Although the presence of smaller hippocampi in at least a substantial portion of persons with PTSD is no longer a matter for debate, the origin of this phenomenon is. With regard to the hippocampus, it is tempting to accept the explanation offered by findings from the animal literature, namely, that stress has damaged this brain structure in persons with PTSD. However, there are at least five possible explanations of the association between an abnormality such as diminished hippocampal volume and PTSD:

1. Preexisting hippocampal diminution may increase the risk of exposure to a traumatic event capable of causing PTSD. As noted above, the hippocampus is critically involved in declarative memory. Those with a smaller hippocampus (e.g., from inheritance, perinatal trauma, or adverse development during childhood) may have less ability to learn from experience and therefore be less able to avoid dangerous situations. Exposure to traumatic events may thus be increased, with the consequent risk of PTSD.
2. Preexisting hippocampal diminution may increase the individual's vulnerability to developing PTSD after exposure to a traumatic event. The hippocampus is critical for the encoding of context. Failure to register the context in which the traumatic event took place may lead to failure to restrict the fear response to a specific context, with resulting generalization to a variety of contexts. For example, veterans who narrowly escaped death in combat may continue to show hypervigilance and exaggerated startle in a variety of situations after redeployment, failing to recognize that they are home and no longer in a dangerous situation.
3. Traumatic exposure may damage the hippocampus, and the damaged hippocampus may lead PTSD, for example, according to the foregoing mechanism.
4. Traumatic exposure may cause the PTSD, and the PTSD may lead to the abnormality. For example, the chronic stress of having PTSD may damage the hippocampus.
5. Traumatic exposure may produce the PTSD, and the PTSD may in turn lead to sequelae or complications that produce the abnormality. For example, alcoholism is a frequent comorbid condition in PTSD, and there is evidence that ethanol damages hippocampal tissue. However, studies have found reduced hippocampal volume in subgroups of PTSD veterans with a history of lifetime alcohol abuse or dependence, in comparison to non-PTSD veterans.

Functional studies of hippocampus lag far behind structural studies. PET has revealed a failure of hippocampal activation during a verbal declarative memory task in women with sexual abuse-related PTSD compared to women with abuse but without PTSD. Another PET study used a word-stem completion task to study firefighters with and without PTSD. Relative to the non-PTSD group, the PTSD group exhibited significantly smaller rCBF increases in the left hippocampus in a high versus low recall comparison. However, this finding reflected relatively elevated rCBF in the low recall condition in the PTSD group, making the results difficult to interpret.

Anterior Cingulate Cortex

Given that decreased ACC activation, especially in its pregenual or rostral division, during exposure to trauma-related stimuli is a prominent functional neuroimaging finding in PTSD, volumetric studies of this structure are of high interest, but they are also rare. Researchers performed semi-automated cortical parcellation of sRMI data in Vietnam veteran nurses with and without PTSD. Three subdivisions of ACC were examined separately: dorsal ('cognitive'), pregenual or rostral ('affective'), and subgenual or subcallosal ('visceral'). The PTSD group exhibited substantially smaller pregenual and subcallosal, but not dorsal, ACC volumes. These reductions were relatively anatomically specific, given that a survey of all 46 other cortical parcellation units found only three that differed significantly between groups, which was consistent with chance.

The technique of voxel-based morphometry (VBM) allows an automated examination of structural brain differences throughout the brain by applying automated statistical parametric mapping to high-resolution sMRI. Researchers studied victims of the Tokyo subway sarin gas attack with and without PTSD. Results showed significant reduced gray matter density, presumably indicative of volumetric reduction, in the ACC in trauma survivors with PTSD compared with those without PTSD, although the location was fairly dorsal. PTSD severity was negatively correlated with gray matter density in the left ACC in PTSD subjects. There were no significant differences in other gray matter regions, or any white matter regions, between the two groups. Also using VBM, researchers compared the ACCs of individuals with acute PTSD with matched healthy control subjects. Results revealed significantly less gray matter density in the right pregenual ACC of the PTSD group, despite the fact that volumetric analyses of the ACC using manual tracing revealed no group differences. Probabilistic maps indicated that the difference between groups in gray matter density was due to difference in shape but not size.

Study of Identical Twins

Trauma does not occur in a vacuum. For example, abused children may also suffer from such potentially deleterious factors as neglect, socioeconomic deprivation, and malnutrition. They may also be more likely to inherit abnormalities from behaviorally disordered parents who abuse them. Soldiers with deprived backgrounds, including child abuse, may be more likely to be assigned to combat roles, and less well equipped to handle the consequent stress. Dissecting the specific pathogenic influences on the brains of persons with PTSD from a complex web of pre-, peri-, and posttraumatic factors requires scientific ingenuity and persistence.

Although successful in demonstrating the presence of brain and other biologic abnormalities in PTSD, the cross-sectional designs of the vast majority of neuroimaging and other biologic studies to date limit their ability to resolve the origin of such abnormalities. Ultimately, prospective studies will be needed that evaluate PTSD biologic markers of interest prior to and following a traumatic event. For obvious ethical and moral reasons, however, studies of the neurologic and psychiatric response to trauma must be naturalistic. Trauma cannot be scheduled in advance. This poses a challenge to prospective studies, which must screen a great many subjects in order to reap enough who will be exposed to a traumatic event during the study period, develop PTSD, and participate in follow-up evaluations.

An alternate approach to the study of the origin of abnormalities in PTSD is to study persons who serve as surrogates for what a trauma-exposed person would have been like but for the traumatic event. The best such surrogates are the identical twins of trauma-exposed persons who were not themselves exposed to the trauma. If an abnormality is genetic or due to environmental influences shared by twins during their rearing, that is, if the abnormality is a 'familial' pretrauma vulnerability factor, then it should also be found in the non-trauma-exposed co-twins of trauma-exposed twins with PTSD. On the other hand, if the abnormality results from an environmental factor unique to exposed twins, for example, a traumatic event, then their co-twins should not share the abnormality. Unfortunately, this approach is unsuitable for the study of abnormal responses to trauma-related stimuli, because such stimuli would not be salient for non-trauma-exposed co-twins.

Over the past several years, researchers have performed a series of studies of identical twins discordant for combat exposure in Vietnam, which have informed the origin of several of the neuroscientific abnormalities that have been found in PTSD discussed above, with the following results. Replicating previous findings, heart rate responses to loud tones were found to be larger in Vietnam combat veteran twins with PTSD. These larger responses were not shared by their non-combat-exposed co-twins, whose responses were similar to those of the non-PTSD combat veterans and their non-combat-exposed co-twins. These results suggest that larger heart rate responses to sudden, loud tones represent an acquired sign of PTSD, rather than a familial vulnerability factor. A similar pattern of results has emerged for diminished P3b amplitude, suggesting it too is an acquired PTSD sign.

In contrast, smaller hippocampi were found both in the Vietnam combat veteran twins with PTSD and their non-combat-exposed, 'high-risk' co-twins. PTSD severity in trauma-exposed PTSD twins was negatively correlated with both their own hippocampal volumes, as well as the hippocampal volumes of their trauma-unexposed identical twin brothers. This pattern of results suggests that diminished hippocampal volume is a pretrauma vulnerability factor for PTSD, rather than an acquired PTSD. A similar pattern of results has recently been reported for NSSs. These results do not support the suggestion that the stress of the traumatic event and/or the PTSD that may follow it cause damage to the hippocampus or other central nervous system components in this disorder.

PTSD, Hormones, and Memory

A substantial body of animal research has demonstrated that stress hormones, especially epinephrine, strengthen emotional memory or conditioned responses. This effect is eliminated by beta-adrenergic blockers such as propranolol, either given systemically or injected into the amygdala. Drawing from this line of research, it has been theorized that stress hormones and neuromodulators, especially epinephrine and norepinephrine, released in response to a traumatic event may over-consolidate memories of the event. These memories subsequently may manifest themselves as intrusive recollections and re-experiencing symptoms. Recalling the event may release additional stress hormones that further enhance the strength of the traumatic memory, creating a positive feedback loop that may eventuate in PTSD. There are several points of opportunity in this loop for secondary preventive psychopharmacological interventions (Figure 1). Based upon the frequently reported (but not consistently replicated) finding of low cortisol levels in PTSD, including

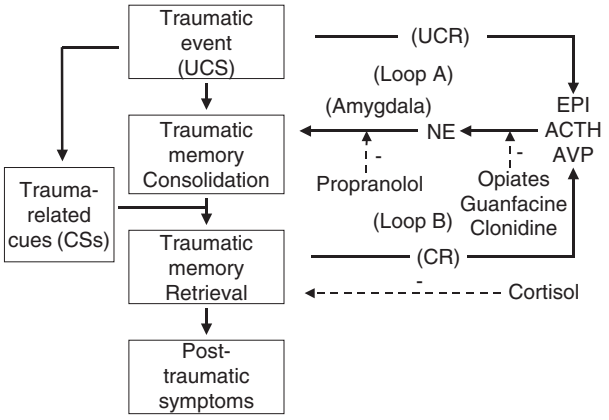

Figure 1 Conceptual model of the pathogenesis of posttraumatic stress disorder with points of opportunity for secondary preventive psychopharmacological agents.

in the immediate aftermath of the traumatic event, as well as upon work that suggests that exogenous cortisol administration has a preventive effect on PTSD, an integrative model posits that deficient cortisol activation at the time of the traumatic event fails to contain the acute catecholamine response putatively implicated in the PTSD's pathogenesis.

An implication of this theory is that pharmacologic intervention, for example, administration of a beta-blocker such as propranolol, soon after a traumatic event might interrupt this cycle of events and prevent the development of PTSD. This hypothesis was tested in a pilot study of patients presenting to a hospital emergency department (ED) after an acute psychologically traumatic event. Patients were randomized to begin, within 6 h of the event's occurrence, a 10-day course of propranolol. When tested 3 months later, the placebo subjects showed higher physiologic responding during script-driven imagery of the event that had brought them to the hospital. These pilot results suggest that acute, posttrauma propranolol has a preventive effect on the physiologic manifestations of subsequent PTSD, possibly by diminishing consolidation of the memory of the traumatic event (or associated conditioned fear response). Subsequently, researchers studied patients shortly after a motor vehicle accident or physical assault. These investigators compared patients who received a course of propranolol with those refused propranolol but agreed to participate in the study. At 2 months post-event, levels of PTSD symptoms were lower in the subjects treated with propranolol. The results of these preliminary studies offer support for the secondary prevention of PTSD with propranol, but more research is needed.

See also: Amygdala: Contributions to Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Dreams and Nightmares in PTSD; Emotion Systems and the

Brain; Fear Conditioning and Synaptic Plasticity; Posttraumatic Stress Disorder: Overview; Posttraumatic Stress Disorder: Neurobiology.

Further Reading

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: American Psychiatric Press.

Charney DS (2004) Psychobiological mechanisms of resilience and vulnerability: Implications for successful adaptation to extreme stress. *The American Journal of Psychiatry* 161: 195–216.

Liberzon I and Phan KL (2003) Brain-imaging studies of posttraumatic stress disorder. *CNS Spectrums* 8: 641–650.

McGaugh JL (2000) Memory – a century of consolidation. *Science* 287: 248–251.

Nutt DJ and Malizia AL (2004) Structural and functional brain changes in posttraumatic stress disorder. *The Journal of Clinical Psychiatry* 65(supplement 1): 11–17.

Orr SP, Metzger LJ, and Pitman RK (2002) Psychophysiology of post-traumatic stress disorder. *The Psychiatric Clinics of North America* 25: 271–293.

Pitman RK, Shalev AY, and Orr SP (2000) Post-traumatic stress disorder: Emotion, conditioning, and memory. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 1133–1147. Cambridge, MA: MIT Press.

Pitman RK, Shin LM, and Rauch SL (2001) Investigating the pathogenesis of posttraumatic stress disorder with neuroimaging. *The Journal of Clinical Psychiatry* 62(supplement 17): 47–54.

Sapolsky RM (2002) Chickens, eggs and hippocampal atrophy. *Nature Neuroscience* 5: 1111–1113.

Taber KH, Rauch SL, Lanius RA, and Hurley RA (2003) Functional magnetic resonance imaging: Application to posttraumatic stress disorder. *The Journal of Neuropsychiatry and Clinical Neurosciences* 15: 125–129.

Vermetten E and Bremner JD (2002) Circuits and systems in stress. II: Applications to neurobiology and treatment in posttraumatic stress disorder. *Depression and Anxiety* 16: 14–38.

Yehuda R (2002) Post-traumatic stress disorder. *New England Journal of Medicine* 346: 108–114.

Yehuda R (2005) Neuroendocrine aspects of PTSD. *Handbook of Experimental Pharmacology* 169: 371–403.

Posttraumatic Stress Disorder: Neurobiology

P Runnels, A S New, and D S Charney, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

The physiologic response to threatening and stressful situations serves as an important mechanism for increasing survival and fitness. However, when stressful cues overwhelm the body's natural homeostatic ability to terminate this response once the stressor has resolved, the results can produce long-term psychological and physiological distress. One way to conceptualize posttraumatic stress disorder (PTSD) is as the persistence of heightened fear despite the cessation of a threatening stimulus, resulting in persistent, vivid recall of traumatic events, autonomic hyperarousal, avoidance of nontraumatic stimuli, and emotional numbing.

While trauma is necessary for the development of PTSD, recent evidence from twin studies suggests that neurobiologic characteristics present before the onset of trauma may play an equally important role. One such study showed a significant decrease in IQ, verbal memory, attention, and executive function in combat veterans who developed PTSD compared to those who did not develop PTSD. However, the novel finding in this study was that these findings were also present in the non-combat-exposed identical twins of those veterans who developed PTSD, while the non-combat-exposed twins of those who did not develop PTSD resembled their combat-exposed twins who did not develop PTSD. Thus, these cognitive abnormalities appeared to track with the genetic factors predisposing to PTSD, rather than arising from trauma exposure. Two other studies showed that combat-exposed Vietnam veterans who developed PTSD had significant reductions in hippocampal volume, increased prevalence of a cavum septum pellucidum, and significantly increased neurologic soft signs when compared with combat veterans who did not develop PTSD. Again, the non-combat-exposed twins of the PTSD veterans more closely resembled their combat-exposed twin with PTSD and were markedly different from the non-combat-exposed twins of the veterans who did not develop symptoms. This suggests that there are genetic factors in individuals which may predispose them to the development of PTSD when they are exposed to traumatic experiences. As new evidence arises, our understanding of stress and PTSD continues to expand. This article explores the neural pathways involved in the fear response, the neurotransmitters that modulate this response, and the most recent evidence for the physiologic processes that are disordered both before and after the onset of PTSD.

Neural Mechanisms of Fear

Classic fear conditioning can be conceptualized as the development of a physiologic reaction to a previously neutral environmental cue (conditioned stimulus) which has become associated with a traumatic event (unconditioned stimulus), such that repeated exposure to this cue consistently elicits anxiety and fear. Physiologically, the conditioned stimulus is perceived by external and visceral pathways and is transferred to the thalamus, from which the signal is transmitted to the basolateral nucleus of the amygdala by means of two circuits: a rapid, direct, subcortical pathway from the dorsal (sensory) thalamus and a slower, regulatory cortical pathway that passes through the primary somatosensory cortices, the insula, and the anterior cingulate/prefrontal cortex. Through this indirect pathway, the stimulus eventually reaches the hippocampus and from there is projected to the lateral amygdala. In passing through this long pathway, the stimulus is exposed to centers of cognition, which interpret the significance of the stimulus and provide a cognitive mechanism for modulating the fear response based on past exposure to similar stimuli. While the basolateral amygdala is primarily responsible for voluntary emotional behavior based upon aversive emotional events, the central nucleus of the amygdala is more involved in conditioned responses to fear-inducing stimuli. Once signals reach the lateral amygdala, they proceed to the central amygdala, which then projects to areas of the hypothalamus and brain stem that mediate the conditioned autonomic, endocrine, and behavioral responses associated with fear.

Emotional Memories and Fear

Brain imaging studies have consistently demonstrated that the amygdala plays a central role in fear conditioning and associated emotional responses to environmental cues. Specifically, the basolateral amygdala has been shown to be integral to the storage of emotionally salient memories. When a traumatic event occurs, memory for emotionally salient aspects of this event remain vivid and intact, while memory for other, less salient information is often impaired. Beta-adrenergic

activity in the amygdala is thought to play a central role in this process. Numerous studies involving propranolol (a centrally acting beta-adrenergic blocker) demonstrate that beta-blockade impairs the explicit memory of emotionally traumatic events while having no effect on emotionally neutral events. This suggests that beta-adrenergic hyperactivity in the amygdala might be an underlying mechanism for the exaggerated and prolonged intrusive memories characteristic of PTSD.

Of course, the exact nature of amygdala involvement in PTSD is still being characterized. Multiple studies have demonstrated amygdala hyperactivity in patients with PTSD when they are exposed to traumatic imagery, and one functional magnetic resonance imaging (fMRI) study of patients with PTSD, compared with controls, showed increased amygdala activity when patients were presented with subliminal, fearful stimuli (in particular, fearful faces). However, other studies using similar paradigms have failed to demonstrate a clear difference. This remains an active area of investigation.

The role of the amygdala in processing emotional memories is further modulated by other agencies. Located in the medial prefrontal cortex (mPFC), the subgenual anterior cingulate gyrus (ACG) appears to play a central role in cognitively processing the significance of emotional stimuli. Multiple reciprocal neuronal pathways have been defined between the amygdala and the PFC, particularly in the ACG. Specifically, studies have demonstrated that lesions of the PFC produce longer extinction times for conditioned aversive stimuli, that the degree of cerebral blood flow predicts subjective experience of negative emotion, and that glucose metabolism in the PFC and amygdala are inversely related.

Such findings suggest a central role for the mPFC and specifically the ACG in controlling consolidation of emotional memories, prompting studies of the ACG in patients with PTSD. Volumetric studies of the ACG in PTSD have been mixed. Two separate studies showed decreased ACG volume in patients with PTSD compared with controls who have experienced similar traumas, but a more recent study using more robust imaging techniques failed to confirm these findings.

Other studies have examined overall activity of the ACG using imaging techniques. Several studies using both positron emission tomography (PET) and fMRI showed that activation of the ACG in response to stressful or negative images and sounds is decreased in subjects with PTSD compared with controls, and that, conversely, symptoms of PTSD are reduced in patients with greater mPFC activation. Another study using PET to measure blood flow demonstrated that patients with PTSD deactivated the ACG while controls deactivated the mPFC instead.

Finally, the role of the hippocampus in emotional memories and PTSD symptoms has also been explored. A meta-analysis of nine studies representing a total of 133 patients compared with 148 controls confirms that hippocampal volume is decreased in patients with chronic PTSD. One specific study of a population of police officers diagnosed with PTSD demonstrated these findings as well. That same study also showed that increased reexperience symptoms are negatively correlated with hippocampal volume. While some have hypothesized that cortisol damages hippocampal cells slowly over time, as an explanation for decreased hippocampal volume in patients with PTSD, the patients in this study had only experienced symptoms for a mean of 1.7 years, hardly long enough for chronic brain damage. This suggests that decreased hippocampal volume is a predisposing factor for, rather than a consequence of, PTSD.

The Hypothalamic–Pituitary Axis in the Normal Stress Response

The normal stress response involves a wide range of neurotransmitters. In response to acute and chronic stress, the paraventricular nucleus (PVN) of the hypothalamus secretes corticotropin-releasing factor (CRH), which, in turn, stimulates both the anterior pituitary gland to synthesize and release adrenocorticotropin (ACTH) and the locus coeruleus to release norepinephrine in the PVN, hippocampus, and prefrontal cortex. ACTH then stimulates the synthesis and release of adrenal glucocorticoids, principally cortisol. Cortisol mobilizes and replenishes energy stores, inhibits growth and reproductive systems, contains the immune response, and affects behavior through actions on multiple neurotransmitter systems and brain regions. Cortisol has important regulatory effects on the hippocampus, amygdala, and prefrontal cortex, including overall enhancement of amygdala activity, increased CRH mRNA concentrations in the central amygdala, an enhanced effect of CRH on conditioned fear, and improved encoding of emotion-related memory.

Recently, the role of cortisol in memory consolidation has been more closely examined. In one set of patients, increased exogenous cortisol doses did not affect alertness and attention, but did impair long-term memory consolidation. Subsequently, another study demonstrated that patients with PTSD are more sensitive to working/declarative memory impairment in the setting of increased cortisol levels. The suspected

target for the effects of cortisol on memory is the hippocampus, where specific glucocorticoid receptors are thought to downregulate working memory when activated. Interestingly, a study of patients with PTSD demonstrated a smaller reduction in working memory compared with controls when dexamethasone was administered, suggesting that the chronic elevation of cortisol in PTSD results in downregulation of hippocampal glucocorticoid receptors.

Abnormalities in the Hypothalamic–Pituitary Axis

Hypothalamic–pituitary axis (HPA) abnormalities in PTSD have been explored in numerous studies. Elevated CRH has been demonstrated in chronic PTSD, suggesting hyperactivity of hypothalamic stress response. Other studies have shown a decreased peripheral stress-response, suggesting an adrenocortical insufficiency in PTSD, a finding which has been further supported by multiple studies showing increased glucocorticoid receptor binding in peripheral mononuclear cells in combat PTSD and hypersuppression of cortisol in response to low-dose dexamethasone administration. These findings, taken together, suggest that patients with PTSD have increased feedback inhibition of cortisol release secondary to enhanced secretion of CRH. Recent work examining the possibility that specific glucocorticoid receptor gene polymorphisms might be responsible for enhanced response to cortisol, and, therefore, a greater risk for developing PTSD, have failed to identify any clear candidate genes at this point.

CRH is among the most important mediators of the stress response. The CRH-containing neurons are located throughout the brain beyond the PVN of the hypothalamus, including the prefrontal and cingulate cortices, the central nucleus of the amygdala, the bed nucleus of the stria terminalis, the nucleus accumbens, the periaqueductal gray matter, the locus coeruleus, and even the dorsal and median raphe nuclei, normally associated with serotonin. Fear-related behaviors are increased when CRH neurons in the amygdala are activated, while reward expectation is reduced when similar CRH neurons in the cortex are excited. Like cortisol, CRH inhibits multiple neurovegetative functions, including food intake, sexual activity, growth, and reproduction. CRH levels are also negatively correlated with percentage of time spent in delta wave sleep.

Increased levels of CRH in cerebrospinal fluid (CSF) and saliva have been linked to PTSD. CRH hyperfunction might be influenced by early-life stress, a negative social environment, past trauma history, and behavioral dominance. Conversely, the biologic propensity to restrain the initial CRH response to an acute stressor may be related to psychobiological resilience to developing PTSD. Furthermore, two subtypes of CRH receptors, CRH-1 and CRH-2, have been found distributed widely throughout the neocortex, amygdala, and hippocampus of the primate brain and appear to have contrasting roles in mediating the stress response. CRH-1-deficient mice display decreased anxiety-like behavior and an impaired stress response, while CRH-2-deficient mice display increased anxiety-like behavior and are hypersensitive to stress. This evidence suggests that regulation of the relative contribution of the two CRH receptor subtypes may be essential to coordinating psychological and physiological responses to stressors. To date, no studies have been able to successfully evaluate CRH-1 and CRH-2 receptors in living humans, though avenues using CRH ligands in PET scans are being explored. This represents an avenue of research targeting pharmacologic interventions that might enhance stress resilience in at-risk or even symptomatic PTSD patients.

Dehydroepiandrosterone

Along with cortisol, dehydroepiandrosterone (DHEA) is another adrenal steroid that is released under stress, and its actions appear to be antagonistic to those of cortisol. Like cortisol, in response to fluctuating levels of ACTH, DHEA is also released synchronously and episodically from the adrenal glands. In the brain, DHEA has antiglucocorticoid and antiglutaminergic activity; however, regionally specific metabolism of DHEA appears to be the primary control mechanism by which DHEA's effects on cognition and behavior are mediated. For instance, 7-hydroxylated metabolites of DHEA in the hippocampus interfere with the normal uptake of activated glucocorticoid receptors. This action may be an intrinsic neuroprotective mechanism aimed at curbing the development of physiological and psychological changes that persistently interfere with function in the absence of trauma or stress. Data supporting this include a negative correlation between DHEA reactivity (in response to ACTH administration) and severity of PTSD symptoms, and a negative correlation between DHEA/cortisol ratios and stress-induced levels of dissociation among elite special forces soldiers undergoing intense survival training. DHEA also restores cortisol-induced suppression of long-term potentiation in hippocampal neurons, thereby enhancing the consolidation of nonemotionally salient memories which are inhibited by the amygdala. Aside from the antiglucocorticoid actions

of DHEA, effects γ-aminobutyric acid (GABA)-A receptors and *N*-methyl-D-aspartic acid (NMDA)-based neurotransmission may be involved in the behavioral effects of DHEA.

Locus Coeruleus and Norepinephrine

The locus coeruleus is a major focus of norepinephrine production and release in the brain, particularly during times of stress. Stressors (physical or emotional, extrinsic or intrinsic) activate the locus coeruleus to release norepinephrine to certain projection sites, including the amygdala, the hippocampus, and the prefrontal cortex, while activating other sites through different neurotransmitters, most notably the sympathetic nervous system (SNS) and the HPA axis. While activating these systems, the locus coeruleus concomitantly inhibits the parasympathetic system, magnifying the decreased need for sleep and food, and, at high levels of activation, even the prefrontal cortex, predisposing an organism to instinctual responses over more complex cognition – to react before thinking.

Encoding of traumatic or emotionally charged memories at the level of the amygdala is heightened by activation of the HPA and locus coeruleus–norepinephrine systems. Once activated, the amygdala activates a feedback loop that further inhibits the prefrontal cortex and stimulates both hypothalamic CRH release and brain stem autonomic centers, further increasing HPA and locus coeruleus activity. Unchecked, such feedback loops presumably lead to persistent hyperresponsiveness of the locus coeruleus–norepinephrine system, thereby contributing to the chronic anxiety, fear, and intrusive memories seen in patients suffering from PTSD.

The relative magnitude of this stress response by the sympathetic system varies from person to person, such that some individuals are prone to 'over-react' and develop PTSD, while others have a more tightly controlled reaction and are less likely to develop PTSD. In fact, in support of this concept, some studies have demonstrated heightened locus coeruleus–norepinephrine activity in patients with panic disorder and PTSD, while other studies suggest that elevated noradrenergic function resulting in total body elevations of norepinephrine increases overall levels of anxiety in response to stress.

Neuropeptide Y

Neurons containing the amino acid neuropeptide Y (NPY) are densely concentrated in five brain areas: the locus coeruleus, the PVN of the hypothalamus, septohippocampal neurons, the nucleus of the solitary tract, and the ventral lateral medulla. Significant levels are also found in the amygdala, hippocampus, cerebral cortex, basal ganglia, and thalamus. NPY is released with norepinephrine when the sympathetic nervous system is activated and appears to have strongly anxiolytic effects. Along with preventing sympathetic overactivation, NPY administration has also been shown to result in a decrease of HPA activation. Microinjection of NPY into the central nucleus of the amygdala reduced anxious behaviors, while NPY mRNA levels were upregulated in the amygdala after chronic stress.

A series of studies on humans confirm that acute stress results in increased levels of cortisol, norepinephrine (NE), and plasma NPY. US Army soldiers as well as US Navy personnel who have participated in survival school training that included capture and interrogation demonstrate an increased level of plasma NPY positively correlated with elevation in plasma cortisol and NE; specifically, stress-induced increases of NPY were significantly and negatively correlated with subjective reports of distress. In another study involving military personnel, soldiers undergoing special operations training consistently produce significantly higher levels of NPY secretion in response to acute stress and had higher baseline levels of NPY in the absence of stress than did soldiers who did not undergo special forces training. These data suggest that military personnel with decreased capacity for stress-induced NPY release and increased dissociation are at a greater risk for the development of stress-related illnesses such as PTSD. The anxiolytic effects of NPY also involve the locus coeruleus, possibly by means of the neuropeptide Y-Y2 receptor. NPY reduces the firing of neurons in the locus coeruleus, thereby reducing the overall levels of norepinephrine, which is anxiogenic. Furthermore, brain regions that express CRH and CRH receptors also contain NPY and NPY receptors, and the functional effects are often opposing, particularly in the locus coeruleus, amygdala, and periaqueductal gray matter. NPY has been shown to counteract the anxiogenic effects of CRH, while CRH antagonists have been shown to block the anxiogenic effects of a neuropeptide Y-Y1 antagonist, leading to the hypothesis that the balance between NPY and CRH neurotransmission is important to a balanced emotional response to stress.

Finally, NPY has been shown to be involved in the consolidation of fear memories. Robust evidence suggests NPY exerts action on the amygdala, probably by means of the neuropeptide Y-Y1 receptor. In rats, injection of NPY into the amygdala impairs memory retention in a foot-shock avoidance paradigm. NPY also has behaviorally relevant effects on memory in the hippocampus. Transgenic rats with hippocampal

NPY overexpression adapt more slowly to the behavioral consequences of stress.

Serotonin

Serotonin (hydroxytryptamine; HT) induces both anxiogenic and anxiolytic effects, depending on the brain region and receptor subtype involved. While the 5-HT_{2A} receptor may induce anxiety, the 5-HT_{1A} receptors are anxiolytic and may even be involved in adaptive responses to stress. The role of the 5-HT_{1A} receptor in the genesis and modulation of anxiety and stress has been most clearly defined. The 5-HT_{1A} receptors are found in the superficial cortical layers, the hippocampus, the amygdala, and the raphe nucleus (primarily presynaptic). 5-HT_{1A} knockout mice show increases in anxiety-like behaviors mediated by postsynaptic 5-HT_{1A} receptors in the hippocampus, amygdala, and cortex.

An interesting recent finding shows that embryonic and early postnatal knockout of 5-HT_{1A} receptors produces an anxiety phenotype that cannot be restored with replacement of 5-HT_{1A} receptors once individuals reach adulthood. By contrast, the same knockout and replacement models conducted in mice entirely during adulthood show a clear ability to regain appropriate anxiety modulation, demonstrating a necessary but poorly understood contribution to anxiety regulation in early development. Postsynaptic 5-HT_{1A} receptor gene expression can also be inhibited by adrenal steroids, demonstrated by the reduction of 5-HT_{1A} receptor density and mRNA levels in response to stress, a reduction which was then prevented by adrenalectomy. These results taken together suggest that early-life increases in CRH and cortisol levels secondary to stress may in turn downregulate 5-HT_{1A} receptors, resulting in a persistently lowered threshold for stressful life events to produce anxiety. Alternatively, genetic decreases in 5-HT_{1A} receptors might be expected to produce a similar phenotype. The density of 5-HT_{1A} receptors is reduced in depressed patients when they are depressed as well as when in remission.

Galanin

Galanin, a peptide containing 30 amino acids, is involved in a number of physiological and behavioral functions, including learning and memory, pain control, food intake, neuroendocrine control, cardiovascular regulation, and anxiety. Galanin is found in high concentrations in the locus coeruleus, where approximately 80% of noradrenergic cells co-express galanin. Furthermore, a dense galanin fiber system originates in the locus coeruleus and innervates forebrain and midbrain structures, including the hippocampus, hypothalamus, amygdala, and prefrontal cortex. Galanin-1 receptor mRNA levels are high in the amygdala, hypothalamus, and bed nucleus of the stria terminalis, and galanin-1 receptor-deficient mice show increased anxiety-like behavior. Studies have shown that galanin reduces the firing rate of the locus coeruleus, possibly by stimulating a galanin-1 autoreceptor. In separate studies, mice that overexpress galanin are unresponsive to the anxiogenic effects of the α-2 receptor antagonist yohimbine. Yohimbine itself increases galanin release in the central nucleus of the amygdala in normal mice. Consistent with these observations, galanin administered directly into the central nucleus of the amygdala of normal mice blocked the anxiogenic effects of stress. Separately, galanin administration or galanin overexpression in the hippocampus results in deficits in fear conditioning.

These results suggest that the noradrenergic response to stress includes induction of galanin release to the central nucleus of the amygdala and prefrontal cortex, and along with NPY, galanin acts as a buffer to the anxiogenic effects of norepinephrine. Interestingly, a recent study examining galanin receptor subtypes demonstrated that treatment with orally administered galanin-3 receptor specific antagonists also reduced anxious behavior in rats, suggesting that galanin activity varies based on receptor subtype. To this point, galanin has not been a focus of studies in patients exposed to traumatic stress or patients with PTSD.

Testosterone

Testosterone levels are reduced in the setting of psychological stress. The mechanism by which physical and psychological stress reduce plasma testosterone remains unclear. The involvement of luteinizing hormone-releasing hormone (LH-RH) synthesis at the hypothalamus level or luteinizing hormone (LH) secretion in the pituitary also remains unclear and has, until now, not been well defined. One possible mechanism involves a recently characterized hypothalamic-testicular pathway, independent of the pituitary, that travels through the spinal cord – this pathway appears to reduce testosterone levels secondary to being acted upon directly by CRH. There is a recent report of reduced CSF testosterone levels in PTSD patients; the finding was negatively correlated with CSF CRH concentrations, but, in the same study, no correlation between plasma and CSF testosterone levels was found. The data from studies measuring plasma testosterone levels in PTSD patients are mixed.

Estrogen

The role of estrogen in anxiety and in the development of PTSD is even less clear. A survey found that

31% of women, compared with 19% of men, develop PTSD when exposed to major trauma; the role of estrogen in the development of PTSD, however, has not been directly investigated in humans. Several studies in female rats lacking endogenous sources of estrogen suggest that estradiol enhances the stress response in female rats and that this effect was dose related. However, a more recent investigation in female rats with intact ovaries found that estradiol reduced the ability of female rats to restrain stress. Taken together, these results indicate that short-term exposure to low doses of estrogen can suppress HPA axis responses to stress, but higher doses and more prolonged treatment potentiate HPA axis responses. The mechanism underlying these effects remains to be elucidated. Greater increases in corticosterone and ACTH levels in female rats in response to both acute and chronic stressors is thought to result from the role of gonadal steroids in activating elements of the HPA axis in females. The effects on the noradrenergic and serotonergic systems are less well understood.

In humans, estrogen has been shown to blunt HPA axis responses to psychological stress in postmenopausal women and to blunt the ACTH response to CRH in postmenopausal women with high levels of body fat. In addition, 8 weeks of estrogen supplementation to perimenopausal women blunted systolic and diastolic blood pressure, cortisol, ACTH, plasma epinephrine and norepinephrine, and norepinephrine responses of the entire body to stress.

See also: Amygdala: Contributions to Fear; Appetitive Systems: Amygdala and Striatum; Aversive Emotions: Molecular Basis of Unconditioned Fear; Aversive Emotions: Genetic Mechanisms of Serotonin; Dreams and Nightmares in PTSD; Emotion: Neuroimaging; Posttraumatic Stress Disorder: Overview; Posttraumatic Stress Disorder as an Emotional Disorder; Stress, the HPA Axis and Depressive Illness.

Further Reading

Bremner JD, Narayan M, Staib LH, et al. (1999) Neural correlates of memories of childhood sexual abuse in women with, without posttraumatic stress disorder. *American, Journal of Psychiatry* 156(11): 1787–1795.

Bremner JD, Staib LH, Kaloupek D, et al. (1999) Neural correlates of exposure to traumatic pictures and sound in Vietnam combat veterans with and without posttraumatic stress disorder: A positron emission tomography study. *Biological Psychiatry* 45(7): 806–816.

Bremner JD, Vythilingam M, Vermetten E, et al. (2004) Effects of dexamethasone on declarative memory function in posttraumatic stress disorder. *Psychiatry Research* 129(1): 1–10.

Britton JC, Phan KL, Taylor SF, et al. (2005) Corticolimbic blood flow in posttraumatic stress disorder during script-driven imagery. *Biological Psychiatry* 57(8): 832–840.

Charney DS (2004) Psychobiological mechanisms of resilience and vulnerability: Implications for successful adaptation to extreme stress. *American Journal of Psychiatry* 161(2): 195–216.

Geracioti TD Jr., Baker DG, Ekhator NN, et al. (2001) CSF norepinephrine concentrations in posttraumatic stress disorder. *American Journal of Psychiatry* 158(8): 1227–1230.

Gurvits TV, Metzger LJ, Lasko NB, et al. (2006) Subtle neurologic compromise as a vulnerability factor for combat-related posttraumatic stress disorder: Results of a twin study. *Archives of General Psychiatry* 63(5): 571–576.

Lanius RA, Williamson PC, Hopper J, et al. (2003) Recall of emotional states in posttraumatic stress disorder: An fMRI investigation. *Biological Psychiatry* 53(3): 204–210.

Morgan CA III, Southwick S, Hazlett G, et al. (2004) Relationships among plasma dehydroepiandrosterone sulfate and cortisol levels, symptoms of dissociation, and objective performance in humans exposed to acute stress. *Archives of General Psychiatry* 61(8): 819–825.

Morgan CA III, Wang S, Rasmusson A, et al. (2001) Relationship among plasma cortisol, catecholamines, neuropeptide Y, and human performance during exposure to uncontrollable stress. *Psychosomatic Medicine* 63(3): 412–422.

Rauch SL, Shin LM, Segal E, et al. (2003) Selectively reduced regional cortical volumes in post-traumatic stress disorder. *NeuroReport* 14(7): 913–916.

Rauch SL, Whalen PJ, Shin LM, et al. (2000) Exaggerated amygdala response to masked facial stimuli in posttraumatic stress disorder: A functional MRI study. *Biological Psychiatry* 47(9): 769–776.

Shin LM, Wright CI, Cannistraro PA, et al. (2005) A functional magnetic resonance imaging study of amygdala and medial prefrontal cortex responses to overtly presented fearful faces in posttraumatic stress disorder. *Archives of General Psychiatry* 62(3): 273–281.

Strange BA, Hurlemann R, and Dolan RJ (2003) An emotion-induced retrograde amnesia in humans is amygdala- and beta-adrenergic-dependent. *Proceedings of the National Academy of Science of the United States of America* 100(23): 13626–13631.

Yehuda R (2002) Current status of cortisol findings in posttraumatic stress disorder. *Psychiatric Clinics of North America* 25(2): 341–368, vii.

Posttraumatic Stress Disorder: Overview

R Yehuda, Mount Sinai School of Medicine and Bronx Veterans Affairs, New York, NY, USA

Published by Elsevier Ltd.

Posttraumatic stress disorder (PTSD) is an anxiety disorder that can develop in persons who experience exposure to life-threatening traumatic experiences. Approximately 15–25% of those exposed to such events develop this disorder, but the prevalence rates may be higher or lower depending on the nature and severity of the event. Trauma survivors are also at increased risk for the development of other mental disorders, such as major depression, panic disorder, generalized anxiety disorder, and substance abuse, as well as persistent anxiety and distress that does not meet criteria for a specific psychological disorder. They are also at risk for developing more somatic symptoms and physical illnesses, particularly hypertension, asthma, chronic pain syndromes, and other psychosomatic illnesses which can occur separately or in tandem with PTSD. However, it is far more common for persons who undergo traumatic stress exposure to not develop any mental illness, including PTSD.

In the immediate aftermath of trauma exposure, it is natural for persons to experience symptoms such as depression, anxiety, nightmares, or repeated, distressing images of the event. However, in the vast majority of cases, these symptoms abate over time, and the presence of such symptoms does not constitute psychiatric impairment. It is only in a minority of persons that the PTSD symptoms that develop in the immediate aftermath of a trauma do not abate. If symptoms progress for several months (or years), chronic PTSD may be diagnosed. Some develop PTSD weeks, months, or even years after an event occurs even if they did not have PTSD immediately after the event. For such persons, PTSD can be a chronic, disabling illness that results in substantial costs to individuals and to society. Longer term consequences of PTSD include absenteeism, reduced productivity, loss of employment, and increased utilization of health care systems due to both mental and physical disability. This article provides relevant information about the conception, phenomenology, psychology, and biology of PTSD, as well as about the therapeutic options for this disorder, which will likely become increasingly relevant in medical settings.

Epidemiology

Studies of the incidence of PTSD in the United States demonstrate that approximately 10% of men and 18% of women had developed PTSD at some time in their lives, making PTSD the fourth most common psychiatric disorder. Events that give rise to PTSD typically involve interpersonal violence, (e.g., rape, assault, violent crime, and torture) or exposure to life-threatening accidents or disaster (e.g., fires and earthquake). These events can result in severe physical injury; however, the defining characteristic of a traumatic event, as described in the context of PTSD, is its capacity to provoke a psychological response of fear, helplessness, or horror owing to the threat of injury or death. Indeed, PTSD can also occur in individuals who witness or learn about traumatic events. This aspect of traumatic stress has become particularly relevant because many people report becoming symptomic to coverage of horrific events in the media. This was particularly true in the aftermath of the September 11, 2001 terrorist attacks. Even those who did not directly witness the collapse of the World Trade Center watched the attacks and their aftermath on television and became extremely anxious and symptomatic as a result. True PTSD is far more likely to occur following direct than indirect exposure to trauma; nonetheless, observing a violent injury or unnatural death of another person, or even learning about such an event occurring to a loved one, can result in the development of PTSD, provided there is a sufficient associated terror or shock.

Different types of traumatic experiences are associated with different probabilities for the development of PTSD. Generally, events that involve interpersonal violence more often given rise to PTSD than those that do not. For example, the prevalence of rape trauma is 0.7% in men and 9.2% in women. Correspondingly, the prevalence of PTSD to rape is 65% in men and 45.9% in women. By comparison, the prevalence of accidents is 35.0% for men and 13.8% for women, with a rate of PTSD of 3.2% for men and 8.8% for women. The witnessing of traumatic events is also associated with relatively low prevalence of PTSD. Thus, events that are more common are associated with lower prevalence rates. Regarding the gender differences in prevalence, this may reflect that women are more vulnerable to the development of PTSD, or it may reflect an increased incidence of violence toward women that might preferentially affect them compared to men.

Most of those who experience a traumatic event – even those who experience it directly – will not meet the criteria for PTSD even if they experience symptoms for several days or weeks following exposure to an event. PTSD is a condition that is only first diagnosed at least 1 month after exposure to a traumatic

event. This time delay between the event and declaration of disorder suggests that symptoms observed in the earlier aftermath should not be conceptualized as mental illness but, rather, as part of universal responses that are expected to abate. Additionally, many of those initially diagnosed with PTSD will recover, either spontaneously or with treatment.

However, the diagnosis of PTSD implies that it is certainly possible to have a prolonged recovery process. This recognition has constituted a major conceptual shift in how stress responses are viewed. Prior to the establishment of the diagnosis of PTSD, stress-related symptoms were generally viewed as transient and not requiring intensive treatment. This notion reflected a still-held view that persons should be able, with time, to 'get over' the effects of a traumatic experience and move on without noticeable impairment. One of the tensions inherent in educating both patients and the public about the effects of trauma is managing the dialectic that requires an emphasis on recovery and resilience as the most common outcomes of trauma while simultaneously validating the potency of its detrimental consequences.

The experience of trauma survivors in the immediate aftermath of a trauma may be quite intense, but it should be interpreted to the survivor as reflecting a normal stage of recovery. It is the failure of PTSD symptoms to abate with time which culminates in multiple mental and physical consequences. Yet, it is often difficult to determine the timeline for recovery, and substantial differences among individuals are present in the time it takes to recover from trauma, even for those demonstrating a normal course of symptom resolution.

An important question is whether the survivor's immediate distress and symptoms in response to trauma are a condition that requires mental health treatment. Treatment should be sought if there is distress and functional impairment since it is currently difficult to predict whether any individual's symptoms will resolve over time. Immediate treatment may forestall a more chronic and debilitating disease. Often, the survivor communicates treatment preference by seeking or not seeking medical or mental health treatment in the immediate weeks following trauma. Typically, trauma survivors rely on natural supports such as family members, clergy, and community during this time. As time goes on and symptoms do not resolve, trauma survivors can become more reluctant to seek treatment. In cases of public disasters, the passage of time results in the fading of interest in the survivor's ordeal or aftermath, creating a shortage of treatment options, embarrassment for not having recovered, and doubts in the survivor's mind about whether ongoing symptoms are in fact related to trauma exposure. It is therefore critically important to emphasize that for those who develop PTSD, the course of illness can be extremely chronic.

Definition and Description of PTSD

In the immediate aftermath of a traumatic event, the biological and psychological components of the response serve to modify one another, creating a dynamic network. The initial fear response to a traumatic event is inherently biological, but it can be modified by the body's interpretation of this biologic state. Moreover, the initial biological and psychological responses to trauma are influenced by prior history and other individual characteristics that together influence whether the initial responses to trauma are transformed on a pathway to recovery or transduced into a chronic disorder. Because PTSD is an evolving dynamic of biological, physiological, psychological, and cognitive connections, it is also highly amenable to therapeutic intervention of both a psychological and a biological nature.

PTSD defines a syndrome with multiple symptoms, most of which reflect ways in which the experience of trauma dominates a person's life and adversely affects mood, cognitive functioning, interpersonal relationships, and quality of life. The three symptom clusters are (1) re-experiencing symptoms, which refer to the ways in which the traumatic events repeatedly intrude into consciousness in the form of distressing images, unwanted memories, nightmares, or flashbacks; these are typically associated with emotional distress and attendant physical symptoms such as palpitations, shortness of breath, and other panic-like symptoms; (2) avoiding reminders of the event, including persons, places, or things associated with the trauma, and becoming emotionally numb and constricted; and (3) physiologic symptoms reflecting hyperarousal, such as insomnia, irritability, impaired concentration, hypervigilance, and increased startle responses.

The symptoms are readily identifiable by a primary care physician. However, the diagnosis of PTSD cannot be made unless one asks about the occurrence of a traumatic event. Often, practitioners are reluctant to ask their patients to talk about events that might be distressing or that might involve shame or secrecy. By providing the opportunity for disclose about the presence of trauma, the practitioner breaks down an important barrier to treatment by legitimizing trauma exposure as a valid explanation for symptoms. The survivor's concern of being misunderstood and pathologized is further allayed when his or her experience of PTSD is clearly understood by the provider.

The Psychological Experience of PTSD

The experience of trauma profoundly alters the survivors' thoughts and beliefs about the world and the self by challenging the illusion of personal safety and causing feelings of vulnerability and powerlessness. If one had previously felt that the world was a safe place, or at least that the United States was a safe place, and that it was possible to protect oneself from adversity, being exposed to a situation in which there was no ability to protect oneself challenged that belief. Recovering from trauma involves accepting or accommodating a new reality about the world in a way that allows for learning, mastery, and empowerment. However, such accommodations are difficult to achieve in the immediate aftermath of a trauma because the body's biologic response perpetuates a physical state of fear that is more conducive to exaggerating a negative supposition about one's own safety. When people feel afraid in response to having experienced an uncontrollable threat to their lives, they may be more compelled to conclude that they can never be safe. Such inaccurate or extreme views perpetuate fear. Many theorists believe that PTSD emerges because of the development of a fear network in memory that compels individuals to avoid thoughts and feelings associated with the traumatic experience. Avoidance prevents exposure to information that can disconfirm inaccurate cognitions about the world.

Other responses, such as horror, anger, sadness, humiliation, or guilt, also occur following trauma. Failure to recover from traumatic experiences is typically due not to emotions that are direct outcomes of traumatic exposure (e.g., horror) but, rather, to those that result from faulty interpretations made by the survivor. For example, many people blame themselves for how they respond during trauma, and such reactions promote shame or embarrassment. Traumatic grief is also typically associated with feelings of blame and guilt. These reactions are not necessarily part of the biologic fear response to the traumatic event but, rather, appear to represent thoughts and interpretations which lead to faulty beliefs.

The process of recovering from trauma requires active engagement with the traumatic experience. To heal from trauma, a person must allow him- or herself to think about what has happened and not run away from the distressing emotions that invariably result from thinking about the event. If one can confront one's feelings in the early aftermath of an event, this will promote a process of habituation. Over time, thoughts about the event and emotional responses will become less distressing. Once a survivor can feel that he or she has achieved some control over the emotional distress of the memory, it will be possible to challenge and correct cognitive misinformation that is erroneous and has led to overgeneralization and further distress.

Biological Aspects

The biologic component of PTSD has become increasingly important as findings delineate that this essentially psychological condition is associated with distinct, often circumscribed physical alterations in the same hormonal, immune, neurochemical, and physiological systems involved in other medical illnesses. In tandem with observations of the phenomenology and psychology of PTSD, neurobiological examinations of trauma survivors have also supported the possibility that the development of PTSD is facilitated by a failure to contain the normal stress response at the time of the trauma, resulting in a cascade of biological alterations that lead to intrusive, avoidance, and hyperarousal symptoms. This failure represents an alternative trajectory to the normal process of adaptation and recovery observed in most people who undergo traumatic experiences.

The normal fear response is characterized by a series of biological reactions that help the body cope with, and gradually recover from, stress. Activation of the sympathetic nervous system (SNS) and the release of adrenaline allow the organism to increase its physiological capacity for 'fight-or-flight' in response to threat and facilitate consolidation of the threat memory. The simultaneous activation of the hypothalamic–pituitary–adrenal (HPA) axis, culminating in the release of cortisol, helps contain the neural, defensive stress reactions. As stress-activated biological reactions are restricted, elevated cortisol levels also suppress the further release of cortisol through negative feedback inhibition on the pituitary, hypothalamus, hippocampus, and amygdala.

Prospective biologic studies have demonstrated that individuals who develop PTSD or PTSD symptoms appear to have attenuated cortisol increases in the acute aftermath of a trauma compared to those who do not develop this disorder. Moreover, persons who develop PTSD show higher heart rates in the emergency room and at 1 week posttrauma compared to those who ultimately recover, suggesting a greater degree of SNS activation. These findings imply that the actual biologic response to acute trauma may be altered in persons who develop PTSD, some of which are typical of and others atypical of the 'normal' stress response.

Low cortisol levels have also been observed in chronic PTSD, even decades after trauma exposure, and therefore appear to be sustained throughout the course of the illness. However, corticotropin-releasing

factor (CRF) levels, measured in cerebrospinal fluid, appear to be increased in this disorder. The pattern of lower cortisol and higher CRF is in contrast to the findings of acute and chronic stress, and of major depressive disorder, which are typically associated with both increased cortisol and increased CRF. That the HPA axis alterations in PTSD are distinct from these conditions has also been shown in studies that have demonstrated an increased negative feedback sensitivity. This enhanced sensitivity of the HPA axis is reflected by an exaggerated suppression of cortisol in response to DEX administration and an increased concentration and sensitivity of lymphocyte glucocorticoid receptors. These findings contrast with the well-demonstrated phenomenon of reduced cortisol suppression in response to DEX, which has been routinely observed in approximately 50% of depressed patients, as well as findings of reductions in glucocorticoid receptor number and sensitivity in major depressive disorder.

The origin and onset of the HPA axis alterations in PTSD have not been elucidated. Low cortisol levels may represent a risk factor for the disorder rather than either a consequence of trauma exposure or a reflection of the pathophysiologic expression of the disorder. Persons who have a parent with PTSD have lower cortisol levels than those who do not. Lower cortisol levels are also associated with childhood emotional abuse, which is itself a risk factor for PTSD. In a study examining the cortisol response in the hours following the trauma of rape, the strongest predictor of low cortisol levels was the risk factor of prior rape or assault.

Since cortisol inhibits its own release through negative feedback at the level of the pituitary and the hypothalamus, lower levels may disrupt the process of stress recovery by failing to inhibit the activation of the pituitary, resulting in increased CRF stimulation in synergy with other neuropeptides, such as arginine vasopressin. This would result in a higher magnitude ACTH response, which might further stimulate the SNS. Moreover, since glucocorticoid release inhibits norepinephrine secretion from the sympathetic nerve terminals, lower cortisol levels may have the consequence of prolonging norepinephrine availability to synapses both in the periphery and in the brain. Ultimately, low cortisol levels might result in an upregulation or increased sensitivity of glucocorticoid receptors in response to the detection of a greater internal demand by the pituitary, and this would further strengthen negative feedback inhibition. Alternatively, an enhanced negative feedback inhibition may be present at the time of the trauma, and it may contribute to the premature suppression of ACTH and cortisol.

There may be consequences of increased catecholamine levels in the acute aftermath of the trauma for promoting the consolidation of the traumatic memory. Under such conditions, the memory of the event would not only be strongly encoded but also be associated with extreme subjective distress. The distress, in turn, would likely facilitate the development of altered cognitions in the aftermath of the event, particularly those associated with one's own perception of danger or ability to cope with threat. These altered cognitions would serve to further delay recovery by leading to a failure to quell fearful responses, which would further serve to strengthen, rather than reduce, both maladaptive cognitions in response to trauma and fear responses, thus perpetuating the intrusive, avoidance, and hyperarousal symptoms as described previously.

The previously mentioned cascade can lead to substantial impairments in brain function. It is well established that persons with PTSD do not show the normal circuitry between the amygdala and the anterior cingulate cortex, which is generally one in which the latter region exerts an inhibition on the amygdala, particularly during exposure to fear-producing triggers. Alterations in hippocampal volume have also been noted in PTSD, although it has been more rare to observe changes in hippocampus brain metabolism in response to traumatic triggers; alterations in this region have been observed in persons with PTSD in association with cognitive deficits. There is evidence that, like lower cortisol levels, smaller hippocampal volumes may constitute early risk factors for PTSD rather than occur as a result of trauma exposure. Thus, the biology of PTSD appears to be one in which there may be a set of pretraumatic biological risk factors that serve to impede the individual's ability to contain the stress response and facilitate a biological sensitization to subsequent traumatic events by allowing for a heightened level of physiologic arousal and consequently distress at the time of the traumatic event. PTSD has been associated with numerous other biologic alterations, including immune, catecholaminergic, psychophysiologic, and in sleep physiology, and these are likely also relevant to the expression of the disorder. The biologic findings in PTSD have contributed to our knowledge substantially by providing an important framework for understanding how similar events could result in biologically heterogeneous outcomes related to individual differences in the presence and severity of PTSD.

Treatment Options for Chronic PTSD

Counseling

An important component of treating trauma survivors is the provision of education, safety, and support,

which can be done in primary care. Education helps the survivor understand the nature of his or her condition and what is involved in the process of recovery. An important element of education involves explaining posttraumatic symptoms in the context of a normal biologic reaction to overwhelming stressors so as to help the survivor understand that such reactions do not represent flaws in character or personal weakness. It is essential to listen noncritically and to provide the frequent message to the survivor that he or she is not alone.

Such intervention can serve the critical role of engaging the trauma survivor and highlighting the need for therapy. Trauma survivors are notoriously reluctant to seek help for their symptoms, particularly from mental health practitioners. Sometimes this may result from a lack of awareness on the part of the survivors that their symptoms are linked to a traumatic event. Often, however, survivors do have an awareness of the relationship between trauma and their symptoms but are concerned about the stigma associated with being portrayed as 'damaged' or emotionally unstable. For many survivors, there may be consequences of disclosure of trauma or treatment (e.g., a police officer may lose the right to carry a weapon and a politician may be affected by disclosure of past trauma or psychiatric treatment). Supportive counseling in primary care can help address many of these barriers to care and reassure patients of confidential treatment options.

Alleviating the core symptoms of PTSD, however, sometimes requires the use of specialized techniques that help the survivor confront fears and emotional responses to the traumatic event in a more structured format, without running away or being overwhelmed. PTSD treatment fundamentally involves reducing the distress associated with traumatic memories and quelling the resultant physiological reactions. Research has demonstrated the effectiveness of a variety of cognitive behavioral techniques, such as exposure therapy, cognitive processing therapy, and eye movement desensitization, which help the person confront painful memories and feelings. Cognitive therapy, in particular, helps the individual modify his or her thoughts and beliefs about safety after a traumatic event, anxiety management, and interpersonal therapies (which focus on how the impact of the traumatic event on relationships and current problems). Group therapy may also be helpful since persons with similar experiences can often support each other in a particularly effective way. Although there are less empirical studies to support treatments such as massage therapy, acupuncture, and other alternative body-focused approaches, many persons seek these avenues of treatment in the aftermath of trauma and report feeling benefited by them.

Pharmacology

Medications can also benefit trauma survivors and can be prescribed by a primary care physician. Results of randomized clinical trials demonstrate that medications such as selective serotonin uptake inhibitors (SSRIs), tricyclic antidepressants, and MAOIs appear to alleviate PTSD symptoms and are also associated with improvements in global functioning quality of life. SSRIs are considered a superior first-line medication because they are safer and better tolerated. Two SSRIs, sertraline, and paroxetine, have been approved by the Food and Drug Administration for the treatment of PTSD. Prasozin, an α_1 agonist, has been demonstrated to reduce nightmares in PTSD patients, and if doses are carefully monitored by a physician, it can achieve very good effects, particularly in improving sleep. There is consensus that benzodiazepines should be avoided or used very judiciously in PTSD. This is based not only on clinical anecdotal wisdom but also on a randomized, double-blind controlled trial that failed to show improvement of alprazolam over placebo in chronic PTSD, as did a study using either alprazolam or clonazepam in the acute aftermath of trauma.

Most experts agree that prescribing medication may not provide a complete treatment response for PTSD, and it is best done as an adjunct to psychotherapy. The partial response to medication may be due to the fact that pharmacologic agents targeting the specific constellation of biologic alterations in PTSD have not been developed. Rather, pharmacotherapy for PTSD has involved administering agents with proven efficacy in other mood and anxiety disorders. In contrast, there has been a focused attempt to develop specialized psychological treatments for PTSD.

There is great interest in medications that may prevent the development of PTSD if administered early enough in the aftermath of a traumatic event. Approaches that focus on reducing distress and the consolidation of traumatic memories are currently being investigated. These include direct reduction of SNS activity (e.g., with medications such as propranolol) or indirect reduction of SNS activity by augmenting HPA axis functioning (e.g., administration of cortisol). Both of these treatments may be somewhat efficacious in preventing the formation of PTSD or traumatic memories.

Summary

PTSD is a serious condition that results from exposure to traumatic stress and appears to represent a failure to recover from what is almost a universal set of emotions and reactions. PTSD typically manifests in distressing memories or nightmares that replay the traumatic event, attempts to avoid reminders of the trauma, and

a heightened state of physiological arousal. The disorder is very common, and it has been shown to occur in 8–10% of the general population at some point in their lives. Studies of the biology of PTSD have delineated circumscribed alterations in brain regions such as the amygdala and hippocampus that are associated with fear and memory, as well as changes in hormonal, immune, neurochemical, and physiological systems involved in coordinating the body's response to stress. The treatment of PTSD involves educating the patient about the nature of the disorder, providing a safe and supportive environment for discussing traumatic events and their impact, and attending to the distress associated with experiencing and avoiding traumatic memories and reminders. A variety of modalities, such as exposure therapy, cognitive therapy, and pharmacotherapy, have been found effective in the treatment of PTSD.

See also: Dreams and Nightmares in PTSD; Posttraumatic Stress Disorder: Neurobiology; Posttraumatic Stress Disorder as an Emotional Disorder; Stress and Suicide; Stress Response: Neural and Feedback Regulation of the HPA Axis.

Further Reading

Aerni A, Traber R, Hock C, et al. (2004) Low-dose cortisol for symptoms of posttraumatic stress disorder. *American Journal of Psychiatry* 161(8): 1488–1490.

Bromet E, Sonnega A, and Kessler RC (1998) Risk factors for DSM-III-R posttraumatic stress disorder: Findings from the National Comorbidity Survey. *American Journal of Epidemiology* 147(4): 353–361.

Christopher M (2004) A broader view of trauma: A biopsychosocial–evolutionary view of the role of the traumatic stress response in the emergence of pathology and/or growth. *Clinical Psychology Review* 24(1): 75–98.

Ducrocq F and Vaiva G (2005) From the biology of trauma to secondary preventive pharmalogical measures for posttraumatic stress disorders. *Encephale* 31(2): 212–226.

Kessler RC, Sonnega A, Bromet E, Hughes M, and Nelson CB (1995) Posttraumatic stress disorder in the National Comorbidity Survey. *Archives of General Psychiatry* 52(12): 1048–1060.

Schelling G, Roozendaal B, and De Quervain DJ (2004) Can posttraumatic stress disorder be prevented with glucocorticoids? *Annals of the New York Academy of Sciences* 1032: 158–166.

Solomon EP and Heide KM (2005) The biology of trauma: Implications for treatment. *Journal of Interpersonal Violence* 20(1): 51–60.

Weiss SJ (2007) Neurobiological alterations associated with traumatic stress. *Perspectives in Psychiatric Care* 43(3): 114–122.

Yehuda R (2006) Advances in understanding neuroendocrine alterations in PTSD and their therapeutic implications. *Annals of the New York Academy of Sciences* 1071: 137–166.

Yehuda R and Ledoux J (2007) Response variation following trauma: A translational neuroscience approach to understanding PTSD. *Neuron* 56(1): 19–32.

Potassium Channel Regulation

N G Gravagna and A B Ribera, University of Colorado Health and Sciences Center, Aurora, CO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Voltage-Gated Potassium Channels

Voltage-dependent potassium channels (Kv) play essential roles in establishing membrane excitability and belong to a large, diverse family of membrane proteins. The nomenclature for voltage-gated potassium channels consists of 'K' (potassium), 'v' (voltage gated), 'number' (subfamily), and '.number' (subfamily member). Each Kv gene encodes a channel subunit containing a signature set of motifs, many of which are targets for developmental or activity-dependent regulation.

In the nervous system, neurons display cell-type-specific expression patterns for Kv genes. Although neurons express multiple Kv genes, many Kv genes are detected in only discrete, restricted populations of neurons. Thus, the repertoire of Kv genes expressed within a neuron limits and defines, in a cell-specific manner, the mechanisms available for modulation of potassium current function and membrane excitability. Here, we will focus on recent progress regarding two of the more thoroughly understood subfamilies of potassium channels, Kv1 (Shaker) and Kv2 (Shab). We compare structural motifs predicted by primary sequence and review evidence implicating subunit-specific domains in developmental or activity-dependent regulation.

The Kv1 and Kv2 Families of K^+ Channels

Kv1 subfamily members have been the focus of numerous studies investigating *in vivo* mechanisms of channel regulation. In contrast, much less is known about the regulation of Kv2 subfamily members, especially Kv2.2. Recent work regarding Kv2.1 channels, however, has provided novel insights. The emerging view indicates that substantially different mechanisms regulate Kv1 and Kv2 channels.

Structural comparisons of Kv1 and Kv2 channels Although the primary sequences of Kv1 and Kv2 genes predict proteins that share many structural features, important subfamily and subunit-specific motifs exist (Figure 1). Overall transmembrane topologies include similar features, such as six transmembrane-spanning (S) domains, a P-loop between the fifth (S5) and sixth (S6) transmembrane domains (pore, ion selectivity filter), and an S4 domain with positively charged residues at every third position (voltage sensor). Each Kv1 or Kv2 gene codes for a single pore-forming (α) subunit of a potassium channel that contains four α-subunits and additional auxiliary subunits. In both Kv1 and Kv2 subunits, the amino (N)- and carboxy (C)-termini are thought to be cytoplasmic (Figure 1). However, closer examination of structural domains highlights important differences.

N-terminal domains Each Kv1 and Kv2 subunit contains a cytoplasmic N-terminal tetramerization domain (T1; Figure 1). However, the T1 domains of Kv1 and Kv2 differ significantly and mediate subfamily-specific co-assembly of subunits. The Kv1 T1 domain also contains the interaction site for accessory (Kvβ) subunits. Kvβ subunits influence surface expression, localization, and kinetic properties of Kv1 complexes.

Heterologous expression studies indicate that Kv1 and Kv2 heteromultimers function as channels with novel properties, indicating that the molecular diversity produced by subunit heteromultimerization has functional consequences. Kv1 heteromultimers form *in vitro* as well as *in vivo*. In contrast, evidence for Kv2 heteromultimerization exists for channels formed *in vitro* but not in neurons.

A distinguishing feature of some Kv1 subunits is an inactivation 'ball' motif contained within the N-terminus (Figure 2). The inactivation ball mediates fast (N-type) inactivation and converts sustained outward currents (I_{Kv}) into transient potassium currents (I_{KA}). The two current types have dramatically different consequences on neuronal excitability. In some Kv1 channel complexes, the inactivation ball resides not within the Kv1 α-subunit but instead on the auxiliary Kvβ subunit, which interacts with the T1 domain (Figure 1). Kvβ1 and Kvβ3, but not Kvβ2, subunits contain inactivation balls. Thus Kv1 channel complexes can carry either sustained or transient currents, depending on the identity of their pore-forming and/or accessory subunits.

There also exist auxiliary proteins that modify the effects of Kvβ subunits. Co-assembly of Kv1.1 and Kvβ1.1 results in formation of a channel that carries inactivating current. An additional protein, leucine-rich glioma inactivating gene 1 (Lgi1), exists in Kv1.1 channel complexes identified *in vivo*. Lgi1 alters channel function by removing fast inactivation. Some forms of temporal lobe epilepsy are due to mutations in Lgi1, highlighting the functional significance of regulation of channel inactivation properties.

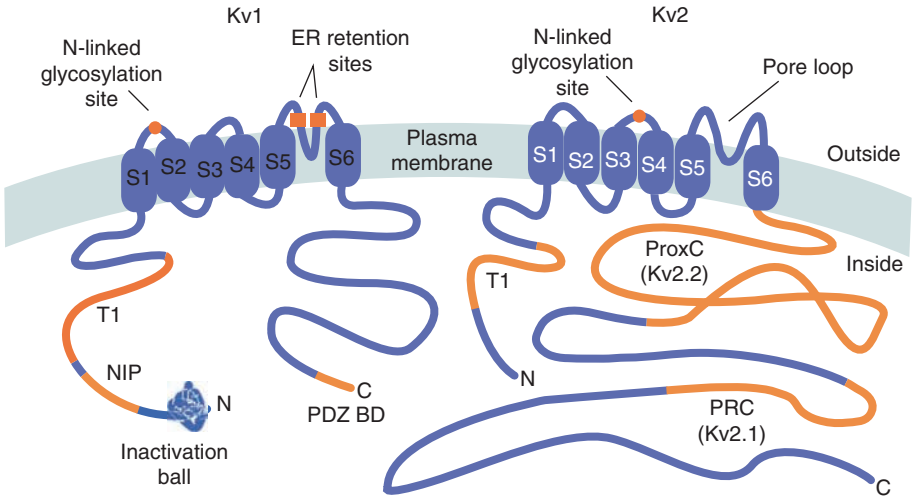

Figure 1 The primary sequences of Kv1 and Kv2 genes predict pore-forming subunits with similar transmembrane topologies. The channel will contain four α-subunits plus additional β-subunits. The β-subunits interact with the T1 domain of Kv1 α-subunits and exist in channel complexes detected *in vivo*. Subunits interacting *in vivo* with Kv2 subunits have yet to be identified. N, amino-terminus; C, carboxy-terminus; S, transmembrane domain; NIP, N-type inactivation prevention domain; T1, tetramerization domain; PDZ BD, PDZ-binding domain; ProxC, proximal carboxy-tail regulatory region; PRC, proximal restriction and clustering signal.

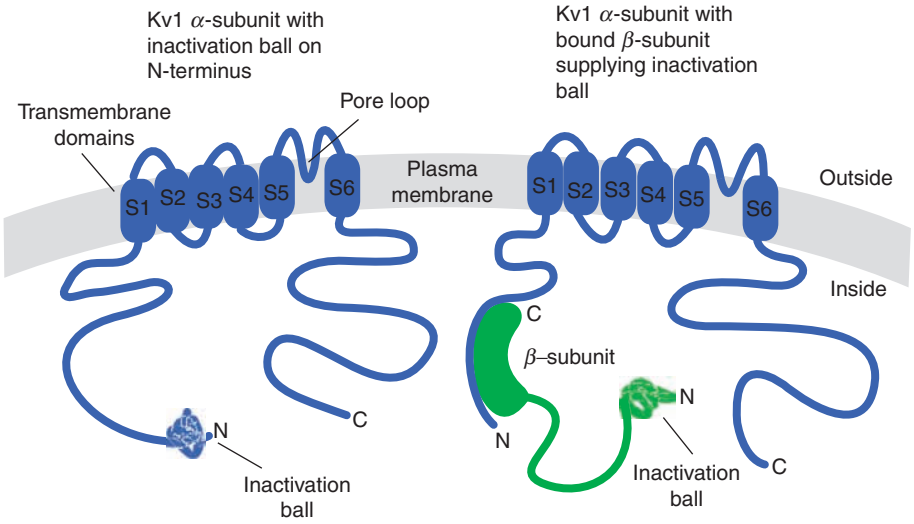

Figure 2 Kv1 channel complexes can carry sustained noninactivating or inactivating currents. Channel complexes containing inactivation balls underlie currents that inactivate rapidly. The inactivation ball can reside on either the pore-forming α-subunit or the auxiliary β-subunit.

In heterologous systems, Kv2 isotypes co-assemble with KChAP (a putative chaperone protein), Kv6, Kv8, or Kv9 subunits (Table 1). Co-expression of KChAP with Kv2 subunits increases the densities of surface membrane Kv2 protein and Kv current. In contrast, Kv6, Kv8, and Kv9 subunits share high amino acid identity with pore-forming Kv2 subunits. However, they do not generate current when expressed on their own. When co-expressed with Kv2 subunits, Kv6, Kv8, or Kv9 subunits decrease current amplitudes. While it is tempting to speculate that Kv6, Kv8, or Kv9 subunits may function as endogenous dominant-negative subunits, direct evidence demonstrating co-assembly with Kv2 subunits *in vivo* is lacking.

C-terminal domains The cytoplasmic C-terminal regions of Kv1 and Kv2 subunits differ substantially. Most obviously, the regions have different amino acid lengths, being much longer in Kv2 subunits (Figure 1). In addition, the C-termini of Kv1 and Kv2 channels mediate different interactions and functions. For example, the C-terminal amino acids of Kv1 subunits bind to modules in neuronal scaffold proteins, known as postsynaptic density/disc/zonula

Table 1 Known Kv1 and Kv2 auxiliary proteins and the effect of interaction

Auxiliary protein	*Associates with*	*Effect of association*[a]
Cytoplasmic		
Kvβ1.1	Kv1.1	Induces inactivated state
	Kv1.3	–
Lgi1	Kv1.1 and Kvβ1.1	Removes fast N-type inactivation, producing sustained currents
Kvβ1.2	Kv1.1	–
	Kv1.2	Increases channel density in surface membrane
	Kv1.4	Increases the rate of inactivation
	Kv1.5	Increases the rate of inactivation, decreases current amplitude
Kvβ1.3	Kv1.5	Binds to the N-terminus of Kv1.5 to induce N-type inactivation, change voltage dependence; alters kinetics of activation and slows the rate of inactivation
	Kv1.6	Immune to Kv1.6 NIP domain, results in fast inactivation
Kvβ2.1	Kv1.1	–
	Kv1.5	Increases the rate of inactivation
	Kv1.4	Increases the rate of inactivation
Kvβ2.2	–	–
Kvβ3.1	Kv1.1	–
Kvβ3.2	Kv2.2	Chaperones channel integration into membrane
KChAP	Kv1.5	–
	Kv2.1	Binds to N-terminus and increases surface expression
	Kv2.2	Binds to N-terminus and increases surface expression
Transmembrane		
Kv5.1	Kv2.1 or Kv2.2	Decreases current amplitude, slows inactivation
Kv6.1	Kv2.1	Decreases current amplitude, slows inactivation
Kv8.1	Kv2.1 or Kv2.2	Modifier or suppressor of activity, depending on concentration; modifies properties of inactivation process
Kv9.1	Kv2.1 or Kv2.2	Modifies kinetics and expression: shifts one-half activation potential to more hyperpolarized values
Kv9.2	Kv2.1 or Kv2.2	Modifies kinetics and expression: shifts activation $V_{1/2}$ to more hyperpolarized values

[a]The designation '–' in some columns indicates that this information is unknown.

occludens (PDZ) proteins (e.g., PSD-95). Scaffold proteins play multiple roles in neurons, including organization of signaling complexes and subcellular localization of membrane proteins. Evidence exists implicating a clustering function for the PDZ-binding domain (PDZ BD) of Kv1 subunits.

Kv2 subunits do not contain any standard PDZ-binding domains or other recognized sites for interactions with scaffold proteins. However, Kv2.1-containing potassium channels form discrete clusters in the soma and dendrites. For Kv2.1 subunits, a novel domain, the proximal restriction and clustering (PRC) signal, directs the characteristic clustering of the channel in the neuronal soma and proximal dendrites. The region corresponding to the PRC in Kv2.2 subunits is poorly conserved, consistent with no clusters of Kv2.2 channels in the somatodendritic region. The carboxy-tail region of Kv2.2 does possess its own signature sequence, called ProxC, because it resides within the proximal portion of the large cytoplasmic carboxy-tail region. The analogous region in Kv2.1 subunits shows little sequence similarity with the Kv2.2 ProxC, suggesting a subunit-specific function. In Kv2.2 subunits, the presence of the ProxC sequence limits potassium current density in mature neurons, ensuring that current density does not exceed a limit referred to as the set point, even when overexpression is attempted empirically by loading neurons with excess Kv2.2 mRNA. In contrast, Kv2.1 subunits can be functionally overexpressed in mature neurons. In addition to subunit specificity, ProxC displays developmental specificity. In neurons that have recently left the postmitotic cycle and are initiating postmitotic differentiation, the ProxC does not limit functional overexpression of Kv2.2 channels.

Additional structural motifs In addition to the structural motifs just described, Kv1 and Kv2 sequences contain several different consensus sites for posttranslational modification. For example, both Kv1 and Kv2 subunits contain an N-linked glycosylation site in an extracellular loop. Glycosylation occurs co-translationally and influences protein folding and stability, often by preventing premature degradation. In Kv1 subunits, the N-glycosylation site resides in an extracellular loop between S1 and S2. Glycosylation at this site contributes to surface expression and membrane stability of channels. In Kv2 subunits, the extracellular loop between S3 and S4 contains a potential N-glycosylation site, but the role of this

posttranslational modification for Kv2 subunits has not yet been tested.

Sites regulating intracellular trafficking have been identified in several different Kv1 subunits. Endoplasmic reticulum (ER) retention sites ensure that, unless appropriately modified, proteins are retained in the ER after translation. Kv1.1 and Kv1.2 subunits, which are abundantly expressed in neurons and heteromultimerize with each other, have ER retention sites flanking the P-loop. Kv1.4 subunits, in contrast, lack ER retention signals in the P-loop but instead have a sequence, VKESL, which promotes surface insertion of channels.

Consensus sites for protein phosphorylation abound in Kv1 and Kv2 subunits. Further, the majority of sites are subfamily, if not subunit, specific. For Kv1 and Kv2 subunits, evidence exists implicating phosphorylation in trafficking of channels through intracellular membrane compartments, control of surface channel density via endocytosis, and modulation of channel functional properties such as voltage dependence. The *in vivo* roles of these mechanisms are discussed in more detail in the following sections.

Regulation of Kv1 and Kv2 Channels

In the mature central nervous system, a majority of changes that occur in response to activity arise through regulation of existing cells rather than by replacement of cell populations. During development, changes in cell number and identity specification also occur, suggesting that nonoverlapping sets of mechanisms may regulate a neuron's electrical phenotype in the developing versus mature nervous systems. For example, on the basis of *in situ* hybridization studies, developmental regulation of ion channel gene transcription is inferred. However, the molecular mechanisms directing expression of specific channel genes in coordinated temporal sequences have yet to be identified.

Mechanisms regulating Kv1 and Kv2 channel function have been identified at the molecular level. The majority of the mechanisms target posttranslational events, although translational control also has emerged as a powerful way to regulate electrical membrane properties in discrete compartments of a neuron in an activity-dependent manner.

Activity-Dependent Translational Regulation of Kv1.1

mTOR mediated Activity-dependent changes in dendrites result in structural remodeling and changes in synaptic strength of synapses. Recent evidence indicates that synaptic activity can also regulate the density of Kv1.1-containing channels at postsynpatic sites. Specifically, *N*-methyl-D-aspartate (NMDA) receptor activation triggers a dendritic signaling pathway that uses the mTOR serine/threonine kinase to specifically block local translation of Kv1.1. If Kv1.1 current density were to decrease, excitatory synaptic stimulation would produce larger dendritic depolarizations, thereby functionally and selectively tagging recent sites of NMDA receptor by an increase in membrane excitability. The mTOR pathway regulates synthesis of Kv1.1 but not Kv1.4 protein, even though dendrites contain mRNAs coding for both.

Posttranslational Regulation

Trafficking and localization Regulation of channel density and location in the surface membrane will dramatically affect the consequences of channel activity on neuronal function. Both pore-forming and auxiliary subunits of a channel complex influence the trafficking of a channel from point of synthesis in the ER to the surface. Surface dwell time reflects the balance between surface membrane insertion and removal, the latter typically achieved by endocytosis.

Studies in heterologous systems have revealed an ER retention sequence near the pore domain of Kv1.1 subunits. Heterotetramerization of Kv1.1 with different Kv1 subunits or binding of dendrotoxin K to the pore region masks the ER retention sequence, resulting in increased trafficking of Kv1.1 channel complexes to the surface membrane. The effect of dendrotoxin K raises the possibility of an endogenous protein that similarly masks the ER retention site, regulating surface insertion of Kv1.1 channels.

Co-assembly of Kv1 with Kvβ2 subunits also promotes increased surface insertion of complexes. However, whether the effect involves masking of ER retention motifs or another mechanism is not yet known. Removal of channels from the surface membrane, known as retrograde trafficking, regulates Kv1.5 channel density in the surface membrane. The mechanism involves dynein, a motor protein that associates channel-containing in-membrane vesicles with microtubules, and results in molecule transport toward the minus, or inward, direction.

Axonal targeting of Kv1.2 channel complexes relies upon Kvβ2 subunits. In addition to the T1 domains of Kv1 subunits, Kvβ2 binds to two different microtubule (MT)-associating proteins called EB1 and kinesin II (Figure 3). EB1, an MT plus-end tracking protein, associates with the MT developing end and promotes addition of tubulin. Kinesin II is a plus-end-directed motor protein. Similar to EB1, the plus-end-directed motor protein kinesin II moves toward the cell surface. However, unlike EB1, kinesin

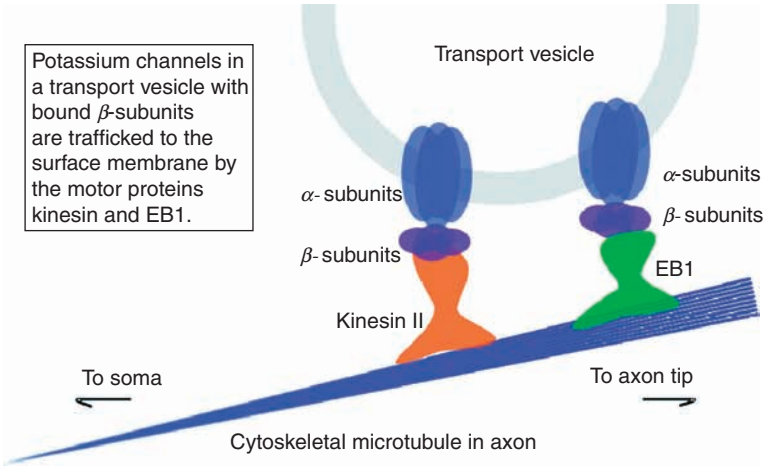

Figure 3 Kvβ2 subunits mediate axonal targeting of potassium channels. The microtubule proteins EB1 and kinesin II direct transport of Kv1.2-containing membrane vesicles via interactions with the auxiliary Kvβ2 subunits. These interactions result in targeted delivery of Kv1.2–Kvβ2 channel complexes to the axon.

II does not necessarily remain associated with the MT end. Interactions of Kvβ2 with EB1 and kinesin II direct trafficking of vesicles containing intracellular pools of Kv1 channel complexes down the MT cytoskeleton of the axon.

As mentioned, Kv2.1 channels localize to the soma and dendrites of neurons in discrete clusters. Interestingly, channels are inserted into the surface membrane at cluster sites. Thus, random insertion of Kv2.1 channels followed by redistribution into clusters does not seem to occur. Further, although channels are restricted to clusters, they move freely within a cluster. The underlying mechanism has been suggested to resemble a fence rather than a scaffold. Channel clustering is not universally observed. For example, Kv1.3 and Kv1.4 channels are highly mobile and evenly distributed throughout the surface membrane. Further, when Kv1.3, Kv1.4, and Kv2.1 channels are present in the same cell, the three subunits exist in separate compartments.

Phosphorylation Ion channel phosphorylation reversibly and dynamically modulates channel function and, consequently, membrane excitability. Phosphorylation often regulates channel density, but the underlying mechanism can differ significantly. For example, phosphorylation can trigger endocytosis of potassium channels from the surface membrane and thereby reduce potassium current density. Alternatively, phosphorylation can alter biophysical properties of the channel and regulate current density elicited by physiological levels of membrane depolarization.

Three Kv1 family channels, Kv1.2, Kv1.3, and Kv1.5, contain consensus sites for tyrosine phosphorylation. Phosphorylation of Kv1.3 subunits by the tyrosine kinase v-src leads to a dramatic reduction in current density. Similarly, Kv1.5 contains an SH3 consensus site for tyrosine phosphorylation that plays a role in mechanisms that reduce current density and/or change biophysical properties of the channel. As expected, modulation of the channel occurs upon phosphorylation of the targeted tyrosine. However, modulation also occurs in the absence of phosphorylation, upon binding of the tyrosine kinase to the SH3 consensus site. Channel removal involves endocytosis via motifs that are regulated by phosphorylation of tyrosine residues. This type of phosphorylation-dependent endocytosis occurs for Kv1.2 and a distantly related potassium channel, the inward rectifier potassium channel, Kir1.1. Thus, tyrosine phosphorylation-dependent endocytosis has been conserved as mechanism for regulation of surface membrane density for potassium channels with significantly different functions.

Phosphorylation of serines in the N- and C-terminal tail regions of Kv2.1 underlies an activity-dependent regulation of channel function. Further, similar to the positive-feedback activity-dependent regulation of Kv1.1, synaptic activity leads to a positive-feedback regulation of Kv2.1 that in turn mediates negative-feedback regulation of activity. Activity promotes release of calcium from the ER that, in turn, activates the phosphatase calcineurin. Upon dephosphorylation of serines by calcineurin, Kv2.1 clusters disperse. Further, the voltage-sensitivity of Kv2.1 channels changes so that the $V_{1/2}$ for activation becomes more negative, allowing a larger percentage of the channels to be activated by physiological levels of membrane depolarization. Thus, neuronal activity leads to generation of larger potassium currents in response to

physiological levels of membrane depolarization; this is positive-feedback regulation of the current. However, increases in potassium conductance will lead to a decrease in membrane excitability and reduction in calcium influx, and thus to negative-feedback regulation of neuronal activity. These findings implicate Kv2.1 in homeostatic mechanisms that are important for maintaining appropriate responses to inputs during periods of plasticity.

Sumoylation Sumoylation, the covalent attachment of small ubiquitin-like modifier (SUMO) to lysine residues of a protein, is known to affect the function of transcription factors and other developmental genes. However, recent results indicate that ion channels also serve as targets for this form of posttranslational modification. The first channel to be recognized as a target for sumoylation was K2P1, a leak 2-pore-type potassium channel. When K2P1 is sumoylated, the conductance of the channel drops essentially to zero. Recently, sumoylation of Kv1.5 subunits was detected *in vivo*. In Kv1.5, SUMO groups attach to lysines in two regions flanking the six transmembrane-spanning domains and shift the voltage-dependence of steady-state inactivation to more negative membrane potentials. Thus, fewer channels are available for activation when the membrane potential is at rest. Interestingly, KChAP, a protein previously identified to interact with Kv1.5, shows up in a Basic Local Alignment Search Tool (BLAST) search as E3 SUMO-protein ligase/PIAS3. Moreover, KChAP also interacts with other Kv1 subunits as well as with both Kv2 subunits, raising the possibility that sumoylation regulates the function of several Kv1 and Kv2 channels.

Conclusions

Studies of Kv1 and Kv2 channels have provided insights into how potassium channel regulation allows neurons to adjust appropriately and dynamically to changes in inputs and activity. During development, these challenges can be extreme because the intrinsic excitable membrane properties change rapidly, from the time when neuronal cells first emerge to maturity. In addition, the number of inputs (e.g., synapses) also changes dramatically, from zero to sometimes hundreds of thousands. In the mature nervous system, such changes are thought to underlie learning and memory. Both the pore-forming components and the auxiliary subunits are targets for regulatory mechanisms. Modification of channel inactivation properties has profound effects on neuronal excitability. Inactivation kinetics are determined in large part by the subunit composition of the channel. The channel density in the surface membrane will affect many critical properties of neuronal excitability, such as the threshold for action potential generation. Regulation of forward or reverse trafficking of Kv1 and Kv2 channel complexes plays a major role in setting surface membrane channel density levels. Moreover, and of particular relevance to the mature nervous system, highly dynamic regulation of channel function also occurs. Signature motifs within each subunit, such as ones for phosphorylation in addition to newly discovered functional modules, determine the set of available regulatory mechanisms and allow an impressive degree of fine-tuning of neuronal excitability. Control of neuronal excitability can occur in temporally and spatially specific manners, depending upon expression pattern. However, little is known about transcriptional mechanisms that control ion channel gene expression, either during development or in the mature nervous system. Thus, identification of transcription factors and RNA-binding proteins that regulate channel expression will provide important new information about potassium channel regulation.

See also: Calcium Channel and Calcium-Activated Potassium Channel Coupling; Glial Influence on Synaptic Transmission; Inwardly Rectifying Potassium Channels; Ion Channel Localization in Axons; Ion Channel Localization in Cell Bodies and Dendrites; Large Conductance Calcium-Activated Potassium Channels; Potassium Homeostasis in Glia; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Clark E, Vacher H, and Trimmer JS (2006) Kv1.1 takes a deTOR from the axon to the dendrite. *Neuron* 52: 399–401.

Gu C, Zhou W, Puthenveedu MA, et al. (2006) The microtubule plus-end tracking protein EB1 is required for Kv1 voltage-gated K^+ channel axonal targeting. *Neuron* 52: 803–816.

Gutman GA, Chandy KG, Grissmer S, et al. (2005) International Union of Pharmacology. LIII. Nomenclature and molecular relationships of voltage-gated potassium channels. *Pharmacology Reviews* 57: 473–508.

Heusser K and Schwappach B (2005) Trafficking of potassium channels. *Current Opinion in Neurobiology* 15: 364–369.

Kaczmarek LK (2006) Policing the ball: A new potassium channel subunit determines inactivation rate. *Neuron* 49: 642–644.

Lai HC and Jan LY (2006) The distribution and targeting of neuronal voltage-gated ion channels. *Nature Reviews Neuroscience* 7: 548–562.

Misonou H and Trimmer JS (2004) Determinants of voltage-gated potassium channel surface expression and localization in mammalian neurons. *Critical Reviews in Biochemistry and Molecular Biology* 39: 25–145.

Misonou H, Mohapatra DP, Menegola M, et al. (2005) Kv2.1: A voltage-gated K^+ channel critical to dynamic control of neuronal excitability. *NeuroToxicology* 26: 743–752.

Moody WJ (2004) Subtype-specific mechanisms for regulating K^+ channel density during development. *Journal of Neurophysiology* 92: 3169–3170.

O'Connell KM, Rollig AS, Whitesell JD, et al. (2006) Kv2.1 potassium channels are retained within dynamic cell surface microdomains that are defined by a perimeter fence. *Journal of Neuroscience* 26: 9609–9618.

Nesti E, Everill B, and Morielli AD (2004) Endocytosis as a mechanism for tyrosine kinase-dependent suppression of a voltage-gated potassium channel. *Molecular Biology of the Cell* 15: 4073–4088.

Pongs O, Leicher T, Berger M, et al. (1999) Functional and molecular aspects of voltage-gated K^+ channels. *Annals of the New York Academy of Sciences* 868: 344–355.

Vacher H, Mohapatra DP, Misonou H, et al. (2007) Regulation of Kv1 channel trafficking by the mamba snake neurotoxin dendrotoxin K. *FASEB Journal* 21: 1–8.

Wilson VG and Rosas-Acosta G (2005) Wrestling with SUMO in a new arena. *Science STKE* 290pe32 (doi: 10.1126/stke.2902005pe32).

Potassium Homeostasis in Glia

P Kofuji and E A Newman, University of Minnesota, Minneapolis, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Potassium in the Extracellular Environment of the Central Nervous System

Potassium concentration in the extracellular environment of the central nervous system (CNS) is kept at relatively low levels (2.5–5 mM) compared to inside neurons (100–130 mM). The transmembrane concentration gradient for Na^+ ions is reversed, with higher Na^+ outside. These transmembrane gradients for K^+ and Na^+ ions are exploited by neurons for fast electrical signaling and synaptic communication. During the repolarizing phase of action potentials and during excitatory synaptic potentials, K^+ ions exit neurons via voltage-dependent K^+ channels and various ionotropic receptors. Even modest neuronal K^+ efflux generates considerable changes in the extracellular K^+ concentration ($[K^+]_o$), given the low initial K^+ concentration and the tortuosity and relatively small volume of extracellular space in the CNS. Small increases in $[K^+]_o$ bring about neuronal depolarization and raised excitability. Larger increases in $[K^+]_o$, on the other hand, inactivate voltage-gated channels and therefore suppress excitability. Thus, it is not surprising that diverse cellular mechanisms are present in the CNS of vertebrates and invertebrates for tightly controlling $[K^+]_o$.

Glial Cells

Numerous functions have been attributed to glial cells in the CNS, including neurotransmitter clearance from synapses, guidance during neuronal migration and, more recently, active control of neuronal synaptic transmission. Additionally, it has long been accepted that glial cells help to maintain the optimal extracellular ionic environment required for communication between neurons. Direct experimental evidence indicates that glial cells, most notably astrocytes, can effectively minimize fast (timescale of seconds) changes in $[K^+]_o$. Astrocytes are well suited for this task because they have processes that are in close association to neuronal synapses and blood vessels (Figure 1) and are abundant in the CNS.

Astrocytes have plasma membranes that are almost exclusively and highly permeable to K^+. The high membrane permeability to K^+ in astrocytes accounts for their relatively hyperpolarized membrane potentials, which are in the range of −70 to −90 mV, close to the predicted K^+ equilibrium potential (E_K). Astrocytes are also connected to each other by gap junctions, which are composed of connexins. Connexins of two adjacent cells can dock with each other to form the gap junction channel, allowing the exchange of small molecules and ions. Astrocytes are extensively coupled to each other, forming a functional electrical syncytium.

Overview of Extracellular K^+ Regulatory Mechanisms

Potassium buffering mechanisms work efficiently in the CNS; increases in $[K^+]_o$ do not surpass a plateau or ceiling level of approximately 10–12 mM, even during intense and prolonged neuronal activity. When these clearance mechanisms are weakened in instances such as spreading depression or ischemia, $[K^+]_o$ can reach values as high as 60 mM. Neuronal function is then severely suppressed.

Glial cells play a key role in regulating extracellular K^+ in the CNS. Glial mechanisms of $[K^+]_o$ buffering are broadly categorized as either net K^+ uptake or K^+ spatial buffering (Figure 1). In net K^+ uptake, excess extracellular K^+ released from active neurons is temporarily sequestered in neighboring glial cells by the activity of K^+ transporters and pumps. To preserve electroneutrality, K^+ influx into glial cells is accompanied by either influx of anions such as Cl^- or efflux of cations such as Na^+ (Figure 1). With net K^+ uptake there is an initial gain of K^+ in glial cells which is usually followed by water influx and swelling. Eventually, the K^+ accumulated in glia is released back into the extracellular space and the overall distribution of K^+ across different cellular compartments in the CNS is restored.

The K^+ spatial buffering mechanism was first proposed by Orkand and colleagues in their landmark experiments on the optic nerve of amphibians. They stated, "If a glial cell becomes depolarized by K^+ that has accumulated in the clefts, the resulting current carries K^+ inward in the high $[K^+]_o$ region and out again, through electrically coupled glial cells in low $[K^+]_o$ regions." Localized increases in $[K^+]_o$ in the CNS are dissipated through transfer of K^+ through glial cells from areas of elevated $[K^+]_o$ to areas in which $[K^+]_o$ is low.

Net K^+ Uptake by Glial Cells

Net K^+ uptake is mediated mainly by glial Na, K-ATPase pumps and Na–K–Cl cotransporters. The

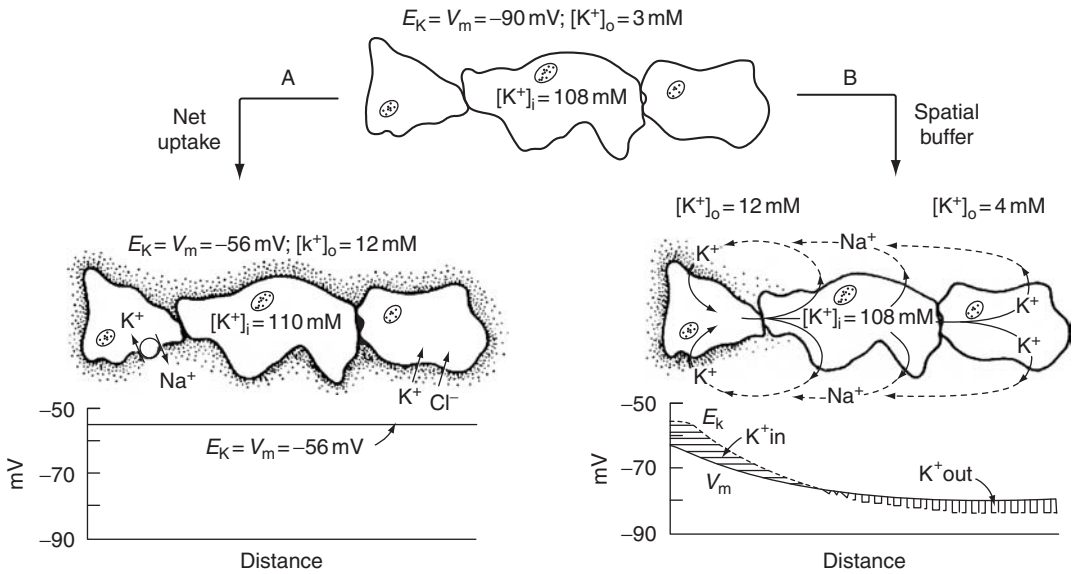

Figure 1 Diagram depicting the role of glial cells in $[K^+]_o$ homeostasis. (Top) Glial cells are electrically coupled via gap junctions forming a functional syncytium. With $[K^+]_o$ equaling 3 mM, the glial syncytium has a membrane potential of approximately −90 mV. (A) Net K^+ uptake mechanism. When $[K^+]_o$ is increased, glial cells accumulate K^+ either by the activity of an Na,K-ATPase or by a pathway in which K^+ is cotranported with Cl^-. In this mechanism of $[K^+]_o$ regulation, the membrane potential in the glial syncytium is spatially uniform at −56 mV. (B) K^+ spatial buffering mechanism. Local increases of $[K^+]_o$ produce a glial depolarization that spreads through the glial syncytium. The local difference in glial syncytium membrane potential (V_m) and the K^+ equilibrium potential (E_K) drives the K^+ uptake in regions of elevated $[K^+]_o$ and K^+ outflow at distant regions. The intracellular currents are carried by K^+ and extracellular currents are mediated by other ions such as Na^+. Reproduced from Orkand RK (1986) Glial–interstitial fluid exchange. *Annals of the New York Academy of Sciences* 481: 269–272, with permission from Blackwell Publishing Ltd.

Na, K-ATPase pump is an integral membrane protein that utilizes the energy from ATP hydrolysis to transport Na^+ out of cells and K^+ into cells, against their concentration gradients. The Na, K-ATPase is a member of the P-type family of ATPases and is closely related to Ca^{2+}-ATPases and H^+–K^+-ATPases. Upon local $[K^+]_o$ increases, the activity of the Na, K-ATPase rises and K^+ is transported into cells, in exchange for Na^+ ions.

Several lines of evidence demonstrate that Na, K-ATPase pumps contribute to $[K^+]_o$ buffering in the CNS. For instance, a transient accumulation of K^+ ions and a simultaneous depletion of Na^+ ions are detected in glial cells of the guinea pig cortex following intense electrical stimulation. A substantial fraction of this K^+ accumulation is prevented by pharmacological blockade of Na, K-ATPase pumps. Likewise, in rat hippocampal slices, blockade of Na, K-ATPase pumps increases baseline $[K^+]_o$ and prevents the rapid clearance of extracellular K^+ ions following neuronal stimulation. In the rat optic nerve, clearance of K^+ accumulation following axonal stimulation is highly temperature dependent, as expected for a carrier-mediated process, and is largely blocked by Na, K-ATPase pump inhibitors.

Transient accumulation of K^+ in cultured astrocytes has also been linked to the activity of Na–K–Cl cotransporters. To date, two Na–K–Cl cotransporter isoforms have been identified: NKCC1, which is present in a wide variety of secretory epithelia and nonepithelial cells, and NKCC2, which is present exclusively in the kidney. Both NKCC isoforms are members of a diverse family of cation–chloride cotransporter proteins that share a common predicted membrane topology and are sensitive to loop diuretics such as bumetanide and furosemide. In cultured astrocytes, intracellular K^+ accumulation following increased $[K^+]_o$ can be partially prevented by furosemide or bumetanide or by removal of external Na^+ and Cl^-. The role of Na–K–Cl cotransporters in $[K^+]_o$ homeostasis has also been demonstrated by optical methods. In the rat optic nerve, intrinsic optical signals reveal that increased $[K^+]_o$ induces astrocyte swelling. This swelling is reversibly depressed by furosemide and bumetanide. Monoclonal antibody labeling of the NKCC1 form of the Na–K–Cl cotransporter shows that NKCC1 is expressed in astrocytes from optic nerves, suggesting the involvement of this particular Na–K–Cl cotransporter isoform in $[K^+]_o$ buffering.

Net K^+ accumulation could, in theory, also occur by K^+ influx through Kir channels accompanied by Cl^- influx through Cl^- channels. However, the Cl^- permeability of nonreactive astrocytes is low, and little K^+ accumulation is believed to occur by this mechanism.

K^+ Spatial Buffering

During K^+ spatial buffering, K^+ enters and exits glial cells via Ba^{2+}-sensitive inwardly rectifying K^+ (Kir) channels. Kir channels have a high open probability at the normal glial resting membrane potential and their slope conductance increases with elevations in $[K^+]_o$. This unique property of Kir channels allows K^+ conductance to increase in glial cells when $[K^+]_o$ is raised, enhancing K^+ clearance rates. Transcellular movement of K^+ ions in the glial syncytium is mediated by various connexins. Gap junctions in astrocytes are primarily composed of Cx43. Astrocytes also express Cx30, Cx47, Cx40, Cx45, Cx46, and Cx26.

During K^+ spatial buffering, the transcellular transfer of K^+ ions from areas of elevated $[K^+]_o$ to areas of lower $[K^+]_o$ generates return current loops in the extracellular space, giving rise to extracellular field potentials. These activity-induced slow extracellular field potentials are generated in various CNS regions, including the neocortex and retina. In the retina, slow field potentials (slow PIII and M waves) are generated upon light stimulation. Current source density analysis indicates that these waves originate from K^+ spatial buffering by retinal glial cells. The transfer of K^+ ions by retinal glial cells, which generates the slow PIII wave, buffers the photoreceptor-based light-evoked decrease in $[K^+]_o$ in the outer retina. As expected, these K^+ spatial buffering fluxes are abolished by blocking retinal glial cell K^+ channels with Ba^{2+}.

Measurements of activity-dependent $[K^+]_o$ changes in the neocortex and cerebellum show that $[K^+]_o$ varies with depth and time in a manner consistent with transcellular transfer of K^+ ions. These $[K^+]_o$ changes are not abolished by sodium pump inhibitors, indicating that they are likely due to a passive K^+ transport mechanism such as K^+ spatial buffering. Similar results are obtained in the drone bee retina, in which it is estimated that approximately 10 times more K^+ ions move as a result of spatial buffering than by simple diffusion through the extracellular space.

Potassium Siphoning in the Retina

In the retina, the Müller cell is the principal type of glial cell. Müller cells display morphological polarization with an end foot process in close apposition to the vitreous humor and apical microvilli projecting into the subretinal space. Like many other glial cells, Müller cells are distinguished from their neuronal counterparts by a resting membrane potential that closely follows the E_K with changes in $[K^+]_o$. This behavior demonstrates a high selectivity of Müller cell membranes to K^+. High membrane permeability is due to the high density and localized expression of Kir channels. Potassium channels are highly concentrated in the end foot process, with more than 90% of the total K^+ conductance localized to this relatively small subcellular domain in amphibian Müller cells.

The highly nonuniform distribution of Kir channels in Müller cells and the lack of extensive gap junctional coupling led to the hypothesis by Newman of a specialized form of the spatial buffering mechanism in the retina called K^+ siphoning. In this hypothesis, excess K^+ from the inner plexiform layer of retina, where most retinal synapses are located, is 'siphoned' by Müller cells to the vitreous humor (Figure 2). The asymmetric subcellular distribution of Kir channels in Müller cells directs clearance of K^+ into large reservoirs such as the vitreous humor.

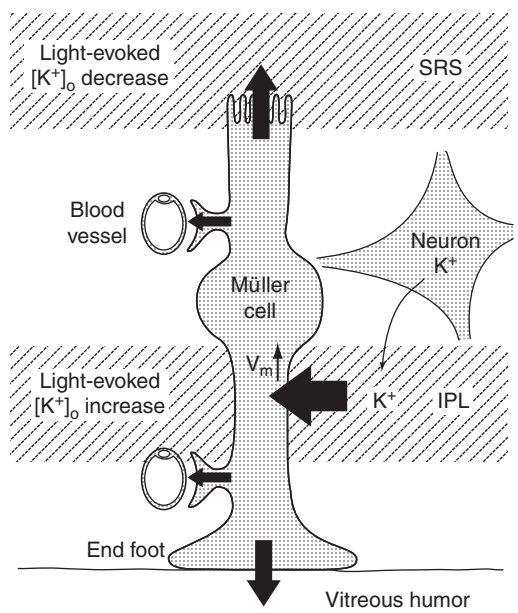

Figure 2 K^+ siphoning in the retina. K^+ released in the inner plexiform layer (IPL) upon neuronal stimulation enters the principal glial cell of the retina, the Müller cell. K^+ leaves the glial cell preferentially from the end foot processes enveloping blood vessels and the vitreal end foot since K^+ conductance is maximal in these membrane regions. Light-evoked $[K^+]_o$ decreases in the subretinal space (SRS) lead to K^+ efflux from apical processes. Reproduced from Newman EA (1996) Regulation of extracellular K^+ and pH by polarized ion fluxes in glial cells: The retinal Muller cell. *Neuroscientist* 2: 109–117, with permission.

Potassium siphoning has been observed in the amphibian retina, in which light flashes evoke increases in $[K^+]_o$ in the vitreous humor. This light-evoked K^+ increase is mediated by Kir channels expressed in Müller cells, as shown by the observation that vitreal increases are blocked by superfusion of retinas with Ba^{2+}, a Kir channel blocker. Similarly, in cat retinas, light-evoked $[K^+]_o$ increases in the inner plexiform layer are threefold larger following blockade of Kir channels, confirming that glial-mediated K^+ spatial buffering currents play an important role in curtailing large changes in $[K^+]_o$.

Glial Cells and K^+ Channels

The Kir channels have recently been cloned, and more than 20 genes are currently known to encode different Kir channel subunits. The basic Kir channel design is of two transmembrane domains and a reentry loop (P-loop), with intracellular N- and C-termini. The Kir channel subunits are usually categorized into seven major subfamilies (Kir1–Kir7) that are diversely regulated by intracellular and extracellular factors.

Of these Kir family members, immunocytochemical and *in situ* hybridization studies demonstrate that the Kir4.1 subtype is broadly expressed in glial cells in the CNS. Kir4.1 immunoreactivity can be demonstrated in cultured and *in situ* astrocytes. Immunogold microscopic examination reveals that Kir4.1 channels are enriched in the processes of astrocytes enveloping synapses and blood vessels. In the retina, highly concentrated expression of Kir4.1 channels is found in Müller cell end feet and on the processes enveloping blood vessels (Figure 3), a distribution that correlates well with electrophysiological studies. Genetic ablation of Kir4.1 channels in mice decreases the membrane conductance of Müller cells by more than tenfold, indicating that the Kir4.1 channel accounts for more than 90% of the cell membrane conductance.

Kir Channel Accessory Proteins in Müller Cells: Localization and Function

The water channel aquaporin 4 (AQP4) is also highly enriched in the end feet and perivascular processes of Müller cells. Transfer of K^+ by K^+ siphoning in the retina will generate osmotic imbalances. Parallel water fluxes are needed to dissipate such imbalances. The highly asymmetric distribution of AQP4 channels allows funneling of water to the appropriate intracellular and extracellular compartments in the retina.

Although the Kir4.1 and AQP4 channels are highly divergent in their primary sequences, they share a key-S-X-V-COOH motif in their C-termini. This sequence binds to PDZ domains, which are modular amino acid motifs implicated in many protein–protein interactions. Proteins possessing these domains are abundantly expressed in the nervous system and include postsynaptic density protein-95 (PSD-95), Chapsyn-110/PSD-93, SAP-102, and hDlg/SAP97.

The PDZ domain-containing adapter protein, α-syntrophin, is crucial for the clustering and subcellular localization of AQP4 and Kir4.1 channels in Müller cells. α-Syntrophin is localized to the cell membrane by its association with the multiprotein dystrophin glycoprotein complex (DGC). DGC spans the cell membrane, forming a molecular bridge between basal lamina proteins in the extracellular space and an array of signaling molecules in the intracellular domain. Immunolocalization studies in retina reveal that the DGC components α-dystroglycan and the short dystrophin isoform, Dp71, are localized in a manner very similar to that of Kir4.1 in Müller cells. The role of Dp71 in the localization of Kir4.1 has been investigated in the dystrophin null mutant mouse, mdx^{3Cv}. The polarized subcellular distribution of Kir4.1 is altered in Müller glial cells from mdx^{3Cv} mice, displaying a more homogeneous distribution pattern (Figure 3). Immunoblotting and whole cell patch clamp experiments demonstrate that the channel is expressed at normal levels at the plasma membrane and its electrophysiological properties are unchanged. Only the cellular distribution of Kir4.1 is changed in the transgenic animals.

In astrocytes, the direct participation of DGC proteins in the targeting of Kir4.1 and AQP4 channels has been demonstrated in an α-syntrophin knockout mouse line. In astrocytes of wild-type animals, as in Müller cells, AQP4 and Kir4.1 are strongly expressed in glial membranes (end feet) that are in direct contact with capillaries and the pia. Quantitative immunoelectromicroscopy shows that in the hippocampus of α-syntrophin knockout mice, the expression of AQP4 in astrocyte end feet is greatly diminished, although the expression of Kir4.1 channels is less affected. These results indicate that α-syntrophin is critical for the targeting and clustering of AQP4 channels to astrocytic end feet but is perhaps not as vital for targeting of Kir4.1 channels.

Despite the apparent lack of major rearrangements in Kir4.1 channel localization and expression in hippocampus astrocytes of α-syntrophin knockout mice, the clearance of extracellular K^+ following neuronal stimulation is significantly impaired in hippocampus slices. Such studies provide the first evidence for a direct functional coupling between AQP4 and Kir4.1 channels in the regulation of $[K^+]_o$.

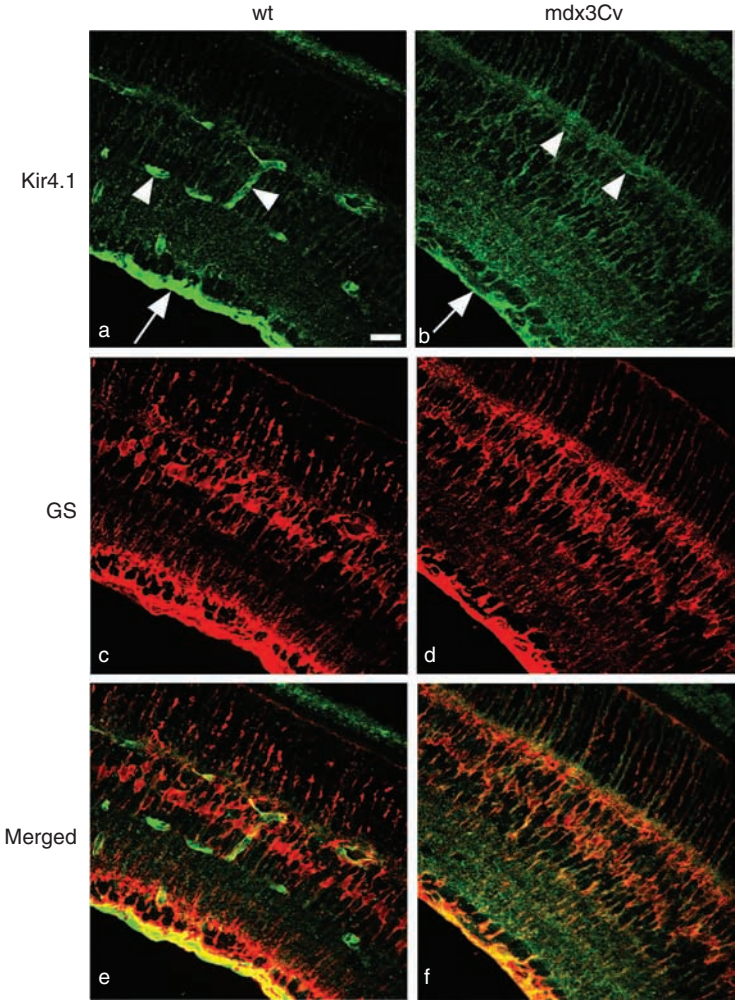

Figure 3 Kir4.1 channel localization in wild-type and mdx3Cv (dystrophin knockout) mouse retinal sections. (a) Kir4.1 is concentrated at the inner limiting membrane (arrow) and to processes around blood vessels (arrowheads) in wild-type retina. (b) In the mdx3Cv mouse, Kir4.1 is more evenly distributed throughout the retina, with a reduction in staining at the inner limiting membrane (arrow) and no apparent enrichment of Kir4.1 around blood vessels (arrowheads). The Müller-specific marker glutamine synthetase (c and d) and merged images (e and f) suggest that Kir4.1 is localized to Müller cells. Scale bar = 25 μm. Reproduced from Connors NC and Kofuji P (2002) Dystrophin Dp71 is critical for the clustered localization of potassium channels in retinal glial cells. *Journal of Neuroscience* 22: 4321–4327. Copyright 2002 by the Society for Neuroscience, with permission.

Potassium Buffering and Pathology

In conditions of increased neuronal activity such as epilepsy there is an elevation of $[K^+]_o$ in the CNS which may further contribute to seizure generation and spread. In the hippocampus of patients with temporal lobe epilepsy, there is substantial reduction in Kir channel expression in astrocytes. Similar findings are reported in animal models of epilepsy. Such reductions of Kir conductance in astrocytes will impair K^+ spatial buffering and contribute to focal hyperexcitabilty and seizure spread.

In animal models of ischemia and diabetes in the retina, there is downregulation of Kir4.1 channel expression in Müller cells as assessed with immunocytochemical and electrophysiological assays. Such decreases in Kir4.1 conductances will impair retinal K^+ siphoning and perhaps contribute to neuronal insult observed in these pathological conditions.

See also: Calcium Homeostasis in Glia; Glial Ion Homeostasis: A Fluorescence Microscopy Approach; Inwardly Rectifying Potassium Channels; Ionic Channels in Glia; Potassium Channel Regulation; Retinal Pharmacology: Inner Retinal Layers; Retinal Glia.

Further Reading

Amedee T, Robert A, and Coles JA (1997) Potassium homeostasis and glial energy metabolism. *Glia* 21: 46–55.

Connors NC and Kofuji P (2002) Dystrophin Dp71 is critical for the clustered localization of potassium channels in retinal glial cells. *Journal of Neuroscience* 22: 4321–4327.

D'Ambrosio R, Gordon DS, and Winn HR (2002) Differential role of KIR channel and Na^+/K^+-pump in the regulation of extracellular K^+ in rat hippocampus. *Journal of Neurophysiology* 87: 87–102.

Holthoff K and Witte OW (2000) Directed spatial potassium redistribution in rat neocortex. *Glia* 29: 288–292.

Karwoski CJ, Lu HK, and Newman EA (1989) Spatial buffering of light-evoked potassium increases by retinal Müller (glial) cells. *Science* 244: 578–580.

Kofuji P, Ceelen P, Zahs KR, Surbeck LW, Lester HA, and Newman EA (2000) Genetic inactivation of an inwardly rectifying potassium channel (Kir4.1 subunit) in mice: Phenotypic impact in retina. *Journal of Neuroscience* 20: 5733–5740.

Kofuji P and Newman EA (2004) Potassium buffering in the central nervous system. *Neuroscience* 129: 1045–1056.

Kuffler SW, Nicholls JG, and Orkand RK (1966) Physiological properties of glial cells in the central nervous system of amphibia. *Journal of Neurophysiology* 29: 768–787.

Leis JA, Bekar LK, and Walz W (2005) Potassium homeostasis in the ischemic brain. *Glia* 50: 407–416.

Newman EA (1996) Regulation of extracellular K^+ and pH by polarized ion fluxes in glial cells: The retinal Muller cell. *Neuroscientist* 2: 109–117.

Newman EA and Reichenbach A (1996) The Müller cell: A functional element of the retina. *Trends in Neuroscience* 19: 307–312.

Orkand RK (1986) Glial–interstitial fluid exchange. *Annals of the New York Academy of Sciences* 481: 269–272.

Orkand RK, Nicholls JG, and Kuffler SW (1966) Effect of nerve impulses on the membrane potential of glial cells in the central nervous system of amphibia. *Journal of Neurophysiology* 29: 788–806.

Seifert G, Schilling K, and Steinhauser C (2006) Astrocyte dysfunction in neurological disorders: A molecular perspective. *Nature Reviews Neuroscience* 7: 194–206.

Somjen GG (2002) Ion regulation in the brain: Implications for pathophysiology. *Neuroscientist* 8: 254–267.

Prader–Willi Syndrome

D C Bittel and M G Butler, University of Missouri–Kansas City School of Medicine and Children's Mercy Hospitals and Clinics, Kansas City, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Clinical Description and Natural History

Prader–Willi syndrome (PWS) is a neurodevelopmental disorder characterized by infantile hypotonia with reduced muscle tone, feeding difficulties, hypogonadism and hypogenitalism, hyperphagia and obesity, small hands and feet, mild mental deficiency, behavioral problems, and a characteristic face (small upturned nose, narrow bifrontal diameter, dolichocephaly, downturned mouth, and almond-shaped eyes) (Figure 1). Hyperphagia in early childhood leads to life-threatening obesity, if uncontrolled. Short stature, scoliosis, and dental and endocrine problems are common (Table 1). In addition, diabetes mellitus, sleep apnea, and other comorbidities associated with obesity may be present. Approximately 50% of children with PWS develop temper tantrums and stubbornness between 3 and 5 years of age and depression during adolescence and adulthood; behavioral problems may be precipitated by withholding food. Intolerance to changes in routine can be a severe problem for many adolescents and adults with PWS, as well as poor peer interaction, immaturity, and inappropriate social behavior. PWS can be divided into two clinical stages: stage 1 is recognized in early infancy with hypotonia, feeding difficulties, and failure to thrive; stage 2 begins in early childhood with onset of hyperphagia and obesity.

PWS arises from lack of expression of paternally inherited genes known to be imprinted in the chromosome 15q11–q13 region. Genomic imprinting is an epigenetic phenomenon whereby gene expression is modified depending on the sex of the parent contributing the allele. It arises from epigenetic changes (e.g., methylation) of genes during gametogenesis which is reversible and parent-specific monoallelic expression of either the maternal or the paternal allele of a particular imprinted gene locus. Approximately 70% of PWS cases are caused by a deletion of the paternal chromosome 15q11–q13 region; approximately 25% have maternal disomy 15 (i.e., two maternal chromosome 15s and no paternal chromosome 15); and the remaining cases arise either from defects of the imprinting center (IC) located in the 15q11–q13 region or from chromosome 15 translocations. The IC is a *cis*-acting sequence involved in establishing and maintaining epigenetic modifications which determine gene expression patterns based on parental origin of the chromosome (Figure 2).

PWS occurs in approximately 1 in 20 000 individuals, with all ethnic groups involved but a reportedly higher occurrence in Caucasians. Most cases are sporadic, with recurrence less than 1%; however, in rare cases the father carries an imprinting defect on his maternal chromosome 15 resulting in a risk of 50%. This occurs due to an inability to replace the female imprint pattern with a male pattern, thus interfering with gene expression in the 15q11–q13 region in the offspring.

As a result of better awareness and advances in genetic testing, the diagnosis of PWS is made earlier than in the past. Early diagnosis is important since nutritional intervention and growth hormone (GH) therapy is frequently prescribed, which improves stature, body composition, energy level, and metabolism. Specific serotonin reuptake inhibitors have been helpful in treating skin picking and depression. If weight is adequately controlled, life expectancy is similar to that of mildly retarded individuals, with an approximate death rate of 3% per year. Obesity and related complications are the most common causes of death in PWS. Caloric diet restriction (approximately 60% of normal) and increased physical activity are important to control the obesity and comorbidities.

Imprinting Center: Characteristics, Function, and Regulation

PWS critical region (PWSCR) is a 2 Mb region within 15q11–q13 containing the IC. It is a bipartite structure composed of five domains which impart maternal- and paternal-specific epigenetic patterns of gene expression, and it spans 35 kilobases (kbs) containing a 4.3 kb sequence (PWS–IC), implicated in PWS, including the *SNRPN* promoter/exon 1 and an 880 bp sequence (AS–IC) located just upstream of the *SNRPN* transcription start site. Deletions of the AS–IC sequence is implicated in Angelman syndrome (AS), an entirely different clinical syndrome generally due to loss of maternally expressed genes in the 15q11–q13. These modules are short sequences which bind different protein complexes depending on gametogenic origin. The modules interact and constitute the *cis*-regulatory elements of the imprinting center that control imprinting of the entire 2 Mb AS/PWS domain (Figure 3). These elements include two copies of a *de novo* methylation sequence (DMS1 and -2) that establish the maternal imprint during oogenesis, an allele discrimination sequence (ADS) that establishes the paternal imprint, and two

elements that maintain the paternal imprint (MPI1 and -2). All five elements bind specific proteins that are presumably the factors acting in *trans* in the imprinting process. The DMS binds a homodimer protein (DMSP) which induces methylation during oogenesis in the CpG island of the promoter, exon 1, and enhancer of *SNRPN*. The ADS binds an allele discrimination protein complex (ADP) which prevents methylation of the CpG islands during spermatogenesis. The MPI elements bind proteins (MPIP1 and MPIP2) together with the ADP complex maintain the unmethylated state of the paternal sequences after fertilization. The *SNRPN* enhancer (SE), which is necessary for *SNRPN* transcription, remains unmethylated on the paternal allele and allows binding of an enhancer protein complex (SEP) which promotes transcription.

Methylation status is apparently constant in somatic tissues throughout life. In addition to *SNRPN*, three other genes (*NDN*, *MAGEL3*, and *MKRN3*) are also maternally imprinted and paternally expressed. The *SNRPN* transcript contains a second protein coding sequence, *SNURF*, as well as multiple copies of small nucleolar RNAs (snoRNAs) thought to participate in RNA processing, all of which are under the IC control. Furthermore, two genes (*UBE3A* and *ATP10A)* are paternally methylated and maternally expressed at least in some tissues. Interestingly, *UBE3A* causes AS but is monoallelically (maternally) expressed in the brain and biallelic in other tissues.

Figure 1 An 8-year-old male with Prader–Willi syndrome showing the facial appearance, obesity, and small hands typically seen in this condition.

Molecular Biology: Candidate Genes for Prader–Willi Syndrome

Imprinted Genes

Candidate genes causing PWS should be paternally expressed and maternally silenced, located within the chromosome 15q11–q13 region and involved directly or indirectly in brain development and function. Genes that meet these criteria include the *SNURF–SNRPN*–snoRNA complex, *NDN*, *MKRN3*, and *MAGEL2*. The most extensively studied paternally expressed gene that is active in brain tissue is the bicistronic gene *SNURF–SNRPN*. Exons 4–10 (termed *SNRPN* for small nuclear ribonucleoprotein N) encode a core spliceosomal protein (SmN) involved in messenger RNA (mRNA) splicing in the brain, whereas exons 1–3 (termed *SNURF* for '*SNRPN* upstream reading frame') encode a 71-amino-acid protein enriched in arginine residues without known function. The promoter and first exon of *SNURF–SNRPN* contain integral sequences of the IC that controls the regulation of imprinting or gene activity throughout the chromosome 15q11–q13 region. A disruption (i.e., paternal deletion, maternal disomy 15, imprinting defects, or chromosome translocations) of this complex locus

Table 1 Clinical characteristics of Prader–Willi syndrome

Stage 1 (0–3 years)	
Pregnancy	Reduced fetal activity, breech delivery, pre- or postterm delivery
Neonatal period and infancy	Narrow forehead, undescended testicles, small genitals and testicles, poor muscle tone (hypotonia), feeding problems, poor suck, sticky saliva, weak cry, temperature instability, developmental delay
Stage 2 (>3 years)	
Childhood	Short stature, small hands and feet, light skin and eye color (in deletion subjects), almond-shaped eyes, strabismus, myopia, skin picking, dental caries, excessive appetite/food foraging, obesity, cognitive impairment, behavioral problems, obsessive–compulsive behavior
Adolescence and adulthood	Short stature, excessive appetite, obesity, cognitive impairment, small hands and feet, scoliosis, osteoporosis, delayed puberty, diabetes mellitus, depression, excessive sleepiness

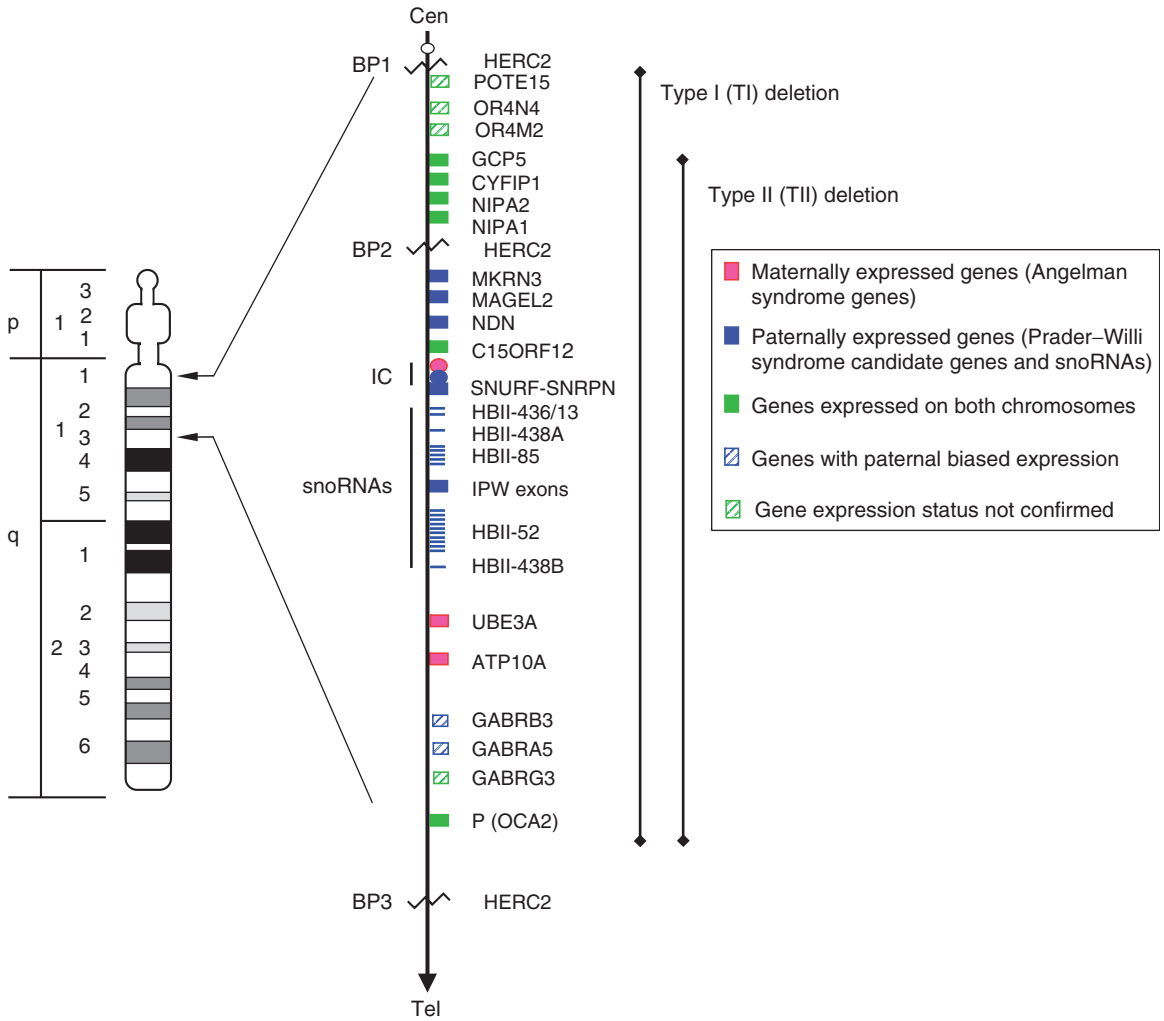

Figure 2 Chromosome 15 ideogram showing the deletion breakpoints (BP1–BP3) and the order of genes within the 15q11–q13 region. Adapted from Bittel DC and Butler MG (2005) Prader–Willi syndrome: Clinical genetics, cytogenetics and molecular biology. *Expert Reviews in Molecular Medicine* 7: 1–20, by permission of Cambridge University Press.

will cause loss of function of paternally expressed genes leading to PWS.

Necdin (*NDN*), a member of the melanoma-associated antigen gene (MAGE) family of proteins, is expressed only from the paternal allele in the 15q11–q13 region. The function of the MAGE family of proteins remains unclear, but they may play a role in cell cycle regulation and apoptosis. Necdin is detected in all developing neurons of the embryonic mouse, in both the central and peripheral nervous system, with the highest expression levels in the diencephalon and hindbrain. After embryonic day 13, necdin has been shown to be essential for axonal outgrowth and is expressed in specific brain structures (hypothalamus, thalamus, and pons), suggesting a developmental role. Mice deficient for necdin have shown hypothalamic deficiency, neonatal lethality, and respiratory and behavioral problems similarly seen in PWS.

The three breakpoint sites (BP1–BP3) within the 15q11–q13 region are 450–500 kb sequences containing a large transcribed gene (*HERC2*) with many partially duplicated copies, some of which are transcribed. The functional *HERC2* gene located at BP3 encodes a highly conserved large protein distantly related to HERC1, a guanine nucleotide exchange factor (GEF) implicated in vesicular trafficking. The mouse genome contains a single *Herc2* locus, located in the juvenile development and fertility-2 (jdf2) interval of chromosome 7C. Mutations in *Herc2* lead to defects in neuromuscular secretory vesicles and sperm acrosomes, other developmental abnormalities, and juvenile lethality of jdf2 mice. These findings suggest that *HERC2* is an important gene encoding a GEF involved in protein trafficking and degradation pathways in the cell. *HERC2* repeated sequences have also been implicated in unequal crossing over in meiosis and may participate in the typical 15q11–q13

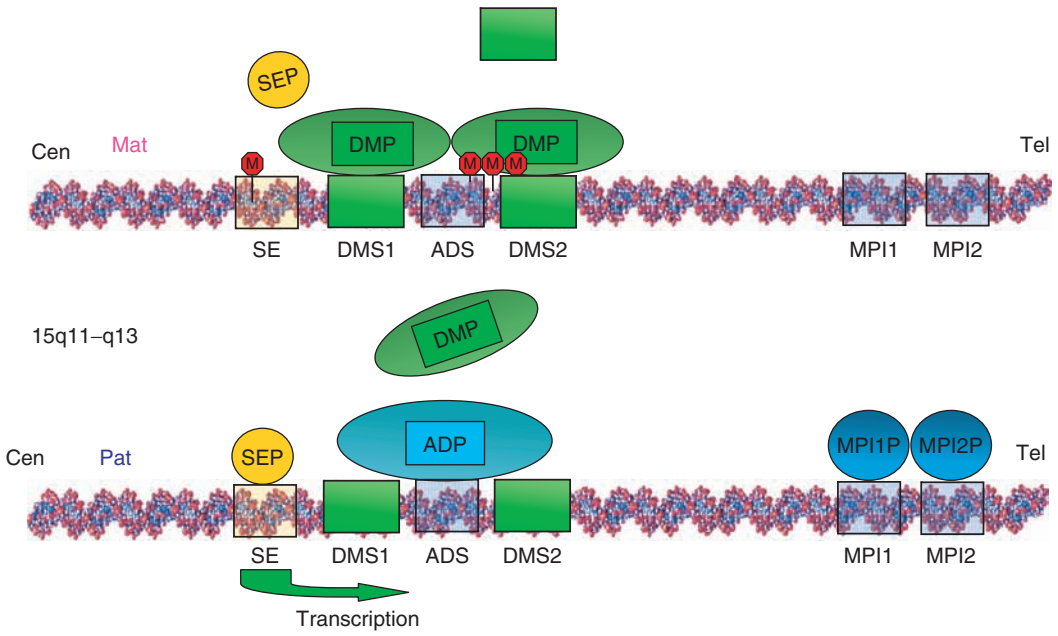

Figure 3 Imprinting elements within the SNRPN promotor. The *de novo* methylation protein (DMP) binds the maternal (Mat) allele at the *de novo* methylation sequence (DMS1 and -2) and induces methylation (M) at the CpG islands, including at the SNRPN enhancer (SE) during oogenesis. The unmethylated state is established on the paternal (Pat) allele when the allele discrimination protein (ADP) displaces DMP during spermatogenesis. The unmethylated state is protected by the maintenance of paternal imprint protein (MPIP1 and -2) binding to the maintenance of paternal imprint sequence (MPI1 and -2), protecting the paternal allele from methylation and allowing transcription to proceed. Adapted from Kantor B, Makedonski K, Green-Finberg Y, Shemer R, and Razin A (2004) Control elements within the PWS/AS imprinting box and their function in the imprinting process. *Human Molecular Genetics* 13: 751–762, by permission of Oxford University Press.

deletion of two sizes (type I and type II; Figure 2). In addition, a PWS female with the typical deletion has been reported with a duplication of the proximal end of the 15q11–q13 region inherited from her father.

Other genes implicated in PWS include *IPW* ('imprinted in Prader–Willi'), *MAGEL2*, and a series of highly repeated sequences encoding snoRNAs. *IPW* is a component of the large *SNURF–SNRPN* transcript and is spliced and polyadenylated but apparently does not encode a protein. However, it may be a functional RNA, similar to *H19* and *XIST*. Also encoded within the *SNURF–SNRPN* transcript are multiple snoRNAs. Two snoRNAs – HBII-52 and HBII-85 – are encoded in a tandemly repeated array of 47 and 24 units, respectively. These snoRNAs were absent from the brain cortex of an individual studied with PWS and from a PWS mouse model, demonstrating paternal expression and a possible causal role in PWS. Importantly, the snoRNA, HBII-52, has been shown to exhibit sequence complementarity to an alternatively spliced exon of the serotonin receptor 5-HT(2C)R, thus regulating alternative splice forms of the receptor. PWS individuals lacking HBII-52 have different 5-HT(2C)R mRNA isoforms than healthy individuals, suggesting that the chromosome 15 defect contributes to misregulation within the serotonin response pathway. Finally, the orthologous mouse gene (*Magel2*) is paternally expressed predominantly in late developmental stages and in adult brain. In humans, the loss of expression of *MAGEL2* might contribute to brain abnormalities and possibly other dysmorphic features seen in PWS.

Comparing overlapping deletion regions among cases of PWS has allowed the identification of a minimal critical region for PWS. This critical region includes an approximately 121 kb segment within the >460 kb *SNRPN* locus which contains only the PWCR1/HBII-85 cluster of snoRNAs and a single HBII-438A snoRNA, further implicating snoRNAs in PWS.

Paternal Bias (GABA Receptor Subunit Genes)

Little consideration has been given to the potential contribution to the PWS phenotype by altered gene expression of nonimprinted genes. Gene expression analysis by microarray technology suggested that several genes/transcripts thought to be biallelically expressed within or proximal to the PWSCR had expression but it was reduced by more than 50% in lymphoblasts from PWS subjects with 15q11–q13 deletions compared to controls. Furthermore, cells from maternal disomy subjects had reduced expression of these genes compared to controls, even though two alleles were present. Most notable of these genes

with paternal bias (i.e., greater expression from the paternal allele) were receptor subunit genes (*GABRA5* and *GABRB3*) for the inhibitory neurotransmitter gamma aminobutyric acid (GABA) located in the 15q11–q13 region. Loss of the paternal allele results in reduction of expression greater than 50% since the paternal allele accounts for a larger percentage of gene expression. GABAergic mechanisms have been implicated in a number of symptoms associated with PWS, including hunger, obsessive–compulsive disorder, and altered visual perception and memory.

Maternally Expressed Genes (*UBE3A* and *ATP10A*)

UBE3A and *ATP10A* are both expressed from maternal alleles in the brain. *UBE3A*, a ubiquitin ligase required for long-term synaptic potentiation, has been identified as the gene causing AS. Studies of the molecular neurophysiology of AS subjects suggest aberrant electroencephalographic patterns resulting from dysregulation of synaptic GABAergic neurotransmission by deficient recruitment of functional $GABA_A$ receptors related to reduced *UBE3A* gene expression. Furthermore, AS subjects with deletions may have decreased amounts of the β_3 subunit for $GABA_A$ receptors due to reduced *GABRB3* gene expression. In addition, the mouse *Atp10a* gene shows tissue-specific maternal expression in the hippocampus and olfactory bulbs overlapping the region of expression of *Ube3a*, suggesting an association with neurological disorders, including AS and autism. Both of these genes have been shown to be overexpressed in PWS subjects with maternal disomy relative to controls. Although the effect of overexpression of these genes has not been studied, it is possible that one or both may contribute to unique features associated with PWS due to maternal disomy (e.g., increased psychosis). Other genes from chromosome 15 may also be differentially expressed depending on parental origin.

Physiology of Eating Behavior and Obesity

Abnormal levels of multiple neuropeptides and hormones involved in eating and neurological function, such as ghrelin, growth and sex hormones, and GABA, have been reported in PWS but the molecular mechanisms are not understood. The peptide ghrelin is produced by the stomach and stimulates eating. Conversely, peptide YY released by the intestine inhibits eating. PWS subjects have higher ghrelin levels compared to lean or obese controls, but expression studies for both genes have shown activity throughout the brain in both PWS and control subjects. A third peptide, obestatin, cleaved from preproghrelin, is thought to inhibit eating.

A 30% reduction in the number of GH-releasing hormone (GHRH) neurons was found in the arcuate nucleus in the brain of PWS subjects, which may explain why GH responsiveness to GHRH is impaired in PWS. Furthermore, PWS has been suggested as a genetic model of starvation due to abnormal hypothalamic function that results in the body interpreting the absence of satiation as a state of starvation. In addition, 7B2 peptide (secretory granule, neuroendocrine protein 1, also known as SGNE1), a neuroendocrine chaperone protein that interacts with prohormone convertase 2 (PC2), is involved in the regulation of secretory pathways in the brain including the release of hormones. SGNE1 is detected in the supraoptic and paraventricular nucleus of the hypothalamus in control subjects but not in the majority of PWS individuals.

GABA is widely distributed throughout the central nervous system and an estimated 40% of the brain and spinal cord neurons utilize GABA as their neurotransmitter, making it the most important inhibitory brain neurotransmitter. Plasma GABA levels are increased three- or fourfold in individuals with PWS compared to control subjects. Furthermore, three genes encoding subunits for $GABA_A$ receptors (*GABRA5*, *GABRB3*, and *GABRG3*) are located in the 15q11–q13 region. The results of a study in which positron emission tomography was used suggested a reduction in GABA receptors in the cingulate, frontal, and temporal neocortices and insula in six adult PWS patients as measured by reduced benzodiazepine binding. An increase in GABA level may be a direct response to the reduction of subunit expression in PWS subjects. Abnormal function of the genetic material on chromosome 15 might alter synthesis, release, metabolism, binding, intrinsic activity, or reuptake of specific neurotransmitters or alter the receptor numbers and/or distribution involved in modulating appetite. Evidence strongly implicates a significant role of altered GABAergic pathways in PWS.

As mentioned previously, most PWS subjects have reduced GH secretory capacity and hypogonadotropic hypogonadism, suggesting hypothalamic–pituitary dysfunction. Replacement of GH and/or sex hormones has been shown to be beneficial in children with PWS. Characteristics such as decreased growth velocity despite the onset of obesity, reduced lean body mass in the presence of adiposity, small hands and feet, relatively low insulin-like growth factor-1 (IGF-1), and relatively low insulin levels are probably the direct result of the hypothalamic GH deficiency in PWS.

Leptin, a peptide produced by adipose tissue, is involved in regulation of appetite and fat storage.

An association has been found between genetic variants of the human obesity gene (*OB*) or leptin and body mass index (BMI) or weight in PWS and age- and gender-matched lean and obese subjects. However, there is no evidence for reduced leptin in PWS. Furthermore, the functional leptin receptor gene (*OBRb*) was reported to be active in lymphoblasts; however, no *OBRb* gene activity was found in human hypothalamus. Reduction of leptin levels after GH treatment and associated reduction of body fat in PWS patients further suggest that leptin is not misregulated in PWS. There is no evidence of abnormalities of leptin receptors, neuropeptide Y, or agouti-related protein in PWS subjects.

Adiponectin is another peptide produced by adipose tissue that plays a role in regulating adiposity. Serum adiponectin levels are significantly lower in PWS subjects compared to lean controls but significantly higher than in obese controls. In addition, no differences in adiponectin levels were found between PWS subjects with 15q deletions compared to those with maternal disomy. No correlation was found between adiponectin and anthropometric parameters or measures of insulin sensitivity or between adiponectin and IGF-binding protein 1 or IGF-1 levels in PWS. Furthermore, adiponectin levels did not change during GH treatment in PWS. Evidence suggests that PWS individuals are less likely to develop diabetes than comparison subjects without PWS but with comparable BMI. However, this may be a consequence of altered fat distribution in PWS resulting in greater adiponectin levels than in obese controls.

Oxytocin, another important neuropeptide produced by the hypothalamus and involved with behavior, has been reported to have significantly higher levels in the cerebral spinal fluid of PWS individuals compared to controls. However, circulating oxytocin levels were abnormally low in relationship to obesity after GH treatment of PWS. A decrease of oxytocin neurons and leptin resistance in PWS might also induce hyperphagia and obesity. However, diabetes mellitus is not a diagnostic criterion for PWS but is often present. The etiology for diabetes mellitus in PWS might be related to the morbid obesity and consequent insulin resistance.

Genetic Subclass Comparison in Prader–Willi Syndrome and Relationship to Imprinting

Gene Expression and Behavior Correlates

The typical deletion, consisting of two classes (type I and type II) of the 15q11–q13 region, is the most common cause of PWS presumably due to unequal crossing over in meiosis at repeated transcribed DNA sequences (i.e., *HERC2* genes) located at the proximal and distal ends of the 15q11–q13 region. The typical deletion is of two sizes and involves a distal breakpoint (BP3) at the end of the 15q11–q13 region and either of two proximally positioned breakpoints (BP1 or BP2) resulting in two different-sized deletions – type I or type II. The type I deletion is larger and involves breakpoints BP1 and BP3; it is approximately 6 Mb in size. The type II deletion involves breakpoints BP2 and BP3 and is approximately 1 Mb smaller based on array comparative genomic hybridization (aCGH). Maternal disomy 15 is found in about 25% of PWS subjects and may occur from nondisjunction during maternal meiosis resulting in a trisomy 15 embryo. Through trisomy rescue of the fetus and loss of the father's chromosome 15, the pregnancy is salvaged and not spontaneously aborted. The fetus is delivered at term having PWS with normal chromosomes but with maternal disomy 15.

Maternal disomy 15 is of two types – heterodisomy or isodisomy. Maternal heterodisomy occurs when the infant inherits each of the mother's chromosome 15s but no chromosome 15 from the father. Maternal isodisomy results when two identical chromosome 15s are inherited from the mother as a result of nondisjunction in meiosis II or from nondisjunction in meiosis I with crossing over occurring in the proximal long arm of chromosome 15. Maternal isodisomy presents the possibility of additional genetic disorders in the PWS subject if the mother is a carrier of an autosomal recessive gene mutation on chromosome 15 (e.g., Bloom syndrome at 15q26.1).

PWS individuals with the typical chromosome 15 deletion are more homogeneous in their clinical presentation and have hypopigmentation due to the deletion of the *OCA2* gene involved with pigment formation. PWS subjects with maternal disomy have fewer typical facial features and are less likely to have certain behavioral findings such as skin picking, unusual skill with jigsaw puzzles, a high pain threshold, and articulation problems. The diagnosis of PWS is often delayed in those with maternal disomy 15 compared to the 15q deletion, reflecting a milder phenotype in the maternal disomy subjects.

Behavioral problems including obsessive–compulsive and self-injurious behavior are frequently reported in PWS. The incidence of self-injurious behavior in individuals with intellectual impairment, autism, and developmental disabilities without PWS ranges from 5% to 60%. However, self-injurious behavior is reported in approximately 70% of PWS adolescents. Self-injurious behavior has been analyzed in a cohort of PWS individuals and skin picking was found to be the most common form of self-injury, particularly in those

PWS individuals with the typical 15q11–q13 deletion. Additionally, PWS individuals hoard and arrange items excessively. Compulsive symptoms have been found in approximately 60% of PWS patients. Those with the deletion recorded higher scores on the Compulsive Behavior Checklist, indicating a greater number of compulsive symptoms and more symptom-related distress than for those with maternal disomy 15. In addition, acute psychosis has been reported in PWS, particularly adults with maternal disomy 15, with possibly more seizures in PWS deletion subjects. PWS children with the deletion showed relative strengths in standardized visual–spatial tasks such as object assembly compared with age- and IQ-matched controls with mixed mental retardation. Interestingly, the PWS group outperformed normal peers in proficiency at jigsaw puzzle placement.

Measures of intelligence and academic achievement in PWS have shown that those with maternal disomy had significantly higher verbal IQ scores than those with the typical deletion. However, PWS deletion subjects also scored higher than the maternal disomy subjects on object-assembly subtests, supporting specific visual perceptual skills as being relative strengths for the deletion group. Additionally, discrimination of shape and motion testing showed that the kinetic form visual performance of the maternal disomy group was significantly worse than that of the comparison or typical deletion PWS groups. Poorer performance in the maternal disomy group could be due to retinal function and visual processing differences related to excessive GABA levels and abnormal GABA subunit receptors. However, PWS individuals with maternal disomy performed better than either typical deletion or comparison subjects in visual recognition and memory tasks. Superior visual recognition memory in the maternal disomy subjects further supports the possible role of two active alleles of maternally expressed genes from chromosome 15 (Table 2).

Although PWS individuals share common characteristics, deletion subjects have poorer outcomes compared to maternal disomy individuals. Although paternally expressed genes are inactive in both subtypes, the increased severity associated with the deletion must be due to reduced expression of hemizygous genes and compensated when two maternal copies are present or there is disruption of interactive genes outside chromosome 15 by altered expression of chromosome 15 genes. In addition, some of the differences might be due to changes in expression of genes distal to the PWSCR that could be affected by the deletion process, repositioning of the genes on chromosome 15, or maternal disomy.

Clinical Findings Associated with Type I versus Type II Typical Deletions

Clinical findings associated with longer (type I) versus shorter (type II) typical deletions have been compared (Table 3). Clinical, anthropometric, and behavioral data were analyzed in PWS individuals with type I or type II deletions. PWS individuals with type I deletions scored significantly worse in self-injurious and maladaptive behavior assessments compared to PWS

Table 2 Differences among subjects with Prader–Willi syndrome due to typical deletion or uniparental maternal disomy (UPD)

15q11–q13 deletion	*Maternal disomy 15 (UPD)*
Hypopigmentation	Shorter course of gavage feeding (in females)
Homogeneous clinical findings and more typical facial appearance	Delayed diagnosis
Lower birth weight	Higher verbal IQ scores
Greater birth length (in males)	Greater numeric calculation skills, attention, word meanings, factual knowledge, and social reasoning
More self-injurious behavior (skin picking)	Superior visual memory
Higher pain threshold	Poorer object assembly and visual perceptual skills
Greater jigsaw puzzle (spatial) skills	Poorer random element stereopsis
	Increased psychosis and autistic behaviors

Table 3 Comparison of genetic subtypes in Prader–Willi syndrome

Genetic subtype	*Behavioral differences*
Type I deletion	Increased maladaptive and compulsive behavior relative to type II and maternal disomy; poorer academic performance relative to type II and maternal disomy
Type II deletion	Better adaptive behavior relative to type I; better social skills relative to type I or maternal disomy
Maternal disomy	Higher verbal IQ scores, greater numeric calculation skills, superior visual memory, poorer object assembly and visual perceptual skills, and increased psychosis relative to typical deletion

subjects with shorter (type II) deletions. Obsessive–compulsive behavior was also more evident in PWS subjects with type I deletions, along with more impairment of visual perception.

Interestingly, individuals with shorter deletions exhibited significantly more self-injury than those with maternal disomy but less than those with longer deletions. The disassociation of compulsivity and skin picking is consistent with our previous factor analytic study revealing that skin picking does not factor with compulsivity using the Compulsive Behavior Checklist. Psychobehavioral characteristics of PWS individuals with type I deletions involving BP1 were similar in several aspects to those of individuals with uncharacterized typical deletions but differed from those with type II deletions involving BP2, the latter group resembling PWS individuals with maternal disomy.

Several academic achievement scores differed between those with shorter and those with longer typical deletions, reflecting a difference in intellectual functioning and visual perceptional skills affecting reading. Therefore, loss of genetic material between BP1 and BP2 apparently increases the severity of behavioral and psychological problems seen in PWS.

Four genes identified in the chromosomal region between BP1 and BP2 (*NIPA1*, *NIPA2*, *CYFIP1*, and *GCP5*) are implicated in compulsive behavior and lower cognition in PWS individuals with type I versus type II deletions. mRNA of these four genes have been compared between subjects with type I and type II deletions and correlated with psychological and behavioral outcomes. mRNA from the four genes was reduced in PWS subjects with the type I deletion, supporting biallelic expression. Generally, mRNA levels were positively correlated with assessment parameters, indicating a direct relationship between mRNA levels and better assessment scores, with the highest correlation for *NIPA2*. Covariate analysis indicated that the quantity of mRNA of the four genes explained a considerable amount (e.g., >50%) of the variation in behavioral and academic parameters measured. More research is needed to illuminate the subtle interactions between the genes on chromosome 15 and interactive networks to clarify the role of these genes in regulating brain development and function.

Genetic Research in Progress in Prader–Willi Syndrome

Microarrays: Pathways and Systems Analysis

The loss of function of genes from the 15q11–q13 region does not directly explain the phenotype of PWS. Genes in *cis* (located on chromosome 15) are affected by the disturbance of chromosome 15, but with the exception of *NDN*, none have been demonstrated to have a phenotype observed in PWS. Therefore, *trans* effects, caused by disturbance of genes not located on chromosome 15 may be responsible for many of the phenotypic consequences associated with the disruption of 15q11–q13 gene expression. Using custom-made microarrays to examine expression of genes/transcripts from 15q11–q13, a higher level of regulatory mechanisms were perturbed in PWS. For example, several genes/transcripts acting in *cis* (e.g., *GABRA5* and *GABRB3*) showed increased expression in PWS lymphoblastoid cell lines established from individuals with maternal disomy compared with those with the typical deletion. However, these genes showed less expression in PWS individuals with maternal disomy than in controls, whereas two maternal copies of the genes produced less RNA than one maternal and one paternal copy of the same gene. This indicated paternal bias of expression. In addition, several transcripts outside the PWSCR had increased expression in individuals with typical deletions relative to controls or PWS individuals with maternal disomy. These observations were confirmed using similar methodology in AS. This expression pattern is presumably a consequence of repositioning of these genes on the proximal long arm region of the deleted chromosome 15. These results warrant further investigations since they indicate that interconnected mechanisms can produce subtle and unexpected changes in gene expression that may relate to phenotypic differences observed among the genetic subtypes of PWS.

Microarray analysis of a transgenic mouse model of PWS equivalent to the deletion genetic subtype demonstrated that genes located in a single chromosomal domain on mouse chromosome 18 are modestly upregulated by the loss of gene expression from the PWSCR (mouse chromosome 7), suggesting a transacting regulatory mechanism. It appears that a gene(s) from the PWSCR interacts with a regulatory element within this domain to reduce expression from this group of genes under normal conditions. We reported transcriptome studies of an IC knockout PWS mouse model and found that out of more than 45 000 probes examined, only 69 had a significant change in expression of at least 1.5-fold between PWS and control mice, 8 of which were from the PWS critical region. Two genes with increased expression in the PWS IC deletion mice were *Pomc* and *Mc5r*; both were validated by quantitative reverse-transcriptase polymerase chain reaction (RT-PCR). *Pomc* gene products are known to be anorexigenic, whereas *Mc5r* directly responds to upregulation of *Pomc* and is involved with thermoregulation, which

is abnormal in PWS infants. PWS model mice generally do not survive more than a few days presumably as a consequence of neonatal failure to thrive reminiscent of the human condition. Imbalances in the Pomc/melanocortin pathway may be an important genetic factor in the survival of these mice because this may negatively affect energy homeostasis.

Whole genome microarrays have been used to analyze mRNA levels in lymphoblastoid cells from PWS subjects and matched controls. Significant differences of at least 1.5 times were seen in 323 genes in PWS lymphoblastoid cells relative to comparison cells, 14 of which were related to neurodevelopment and function. Alterations in expression of serotonin receptor genes (e.g., *HTR2B*) and genes involved in eating behavior and energy homoeostasis (*ADIPOR2*, *MC2R*, *HCRT*, and *OXTR*) were noted. Other genes of interest with reduced expression in PWS subjects included *STAR* (a key regulator of steroid synthesis) and *SAG* (an arrestin family member which desensitizes G-protein-coupled receptors). Quantitative RT-PCR validated differences were seen in microarrays using RNA from both lymphoblastoid cells and brain (frontal cortex) from separate PWS individuals and controls. Our analysis further identified previously unappreciated changes in gene expression which may contribute to the clinical manifestations seen in PWS.

Microarray analyses suggested a cumulative effect of modest changes in expression of many genes, especially those involved in energy metabolism which may contribute to failure to thrive in PWS infants. Subtle transactivational control of gene activity by genes within the PWSCR suggests that changes in the transcriptome and/or proteome remain to be discovered to explain dramatic phenotypic effects of loss of expression of 15q11–q13 genes.

Behavior and Brain Imaging

Brain imaging studies using functional magnetic resonance imaging from a cohort of PWS individuals and controls revealed postprandial hyperfunction in the limbic and paralimbic regions of the brain in PWS. Conversely, brain imaging of healthy weight individuals indicated increased activity of the limbic and paralimbic regions prior to consumption of a meal. After eating, these regions of the brain were normalized and neuronal activity decreased substantially. In contrast to healthy weight individuals, PWS subjects failed to normalize these regions of the brain after meal consumption and instead developed a dramatic increase in neurological activity (Figure 4). Thus, the failure of satiation following a meal appears to have a neurological basis in PWS.

Future Directions

PWS individuals exhibit reduced selectivity in choosing food items for consumption (i.e., they may consume inappropriate food items). There is a known correlation between smell and taste and/or selection

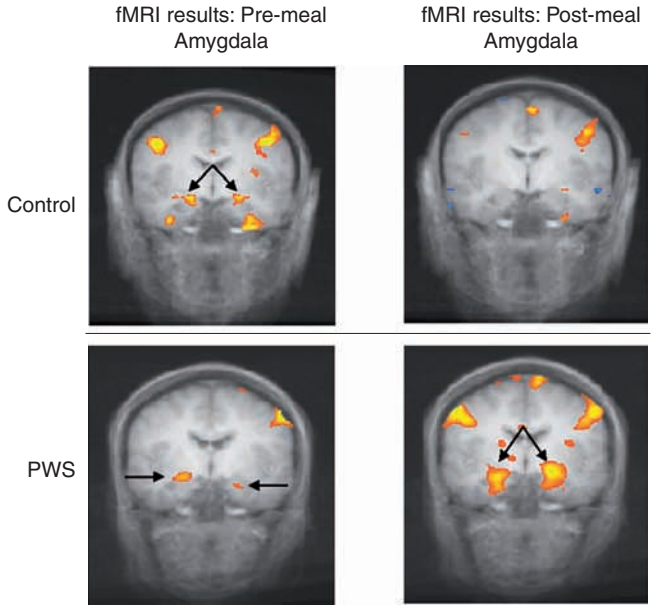

Figure 4 Functional magnetic resonance imaging (fMRI) of the limbic region (arrows indicate the amygdala) in Prader–Willi syndrome (PWS) and control subjects while viewing food stimuli pictures before and after a meal. Courtesy of Cary Savage and Laura Holsen, Kansas University Medical Center (Kansas City, KS).

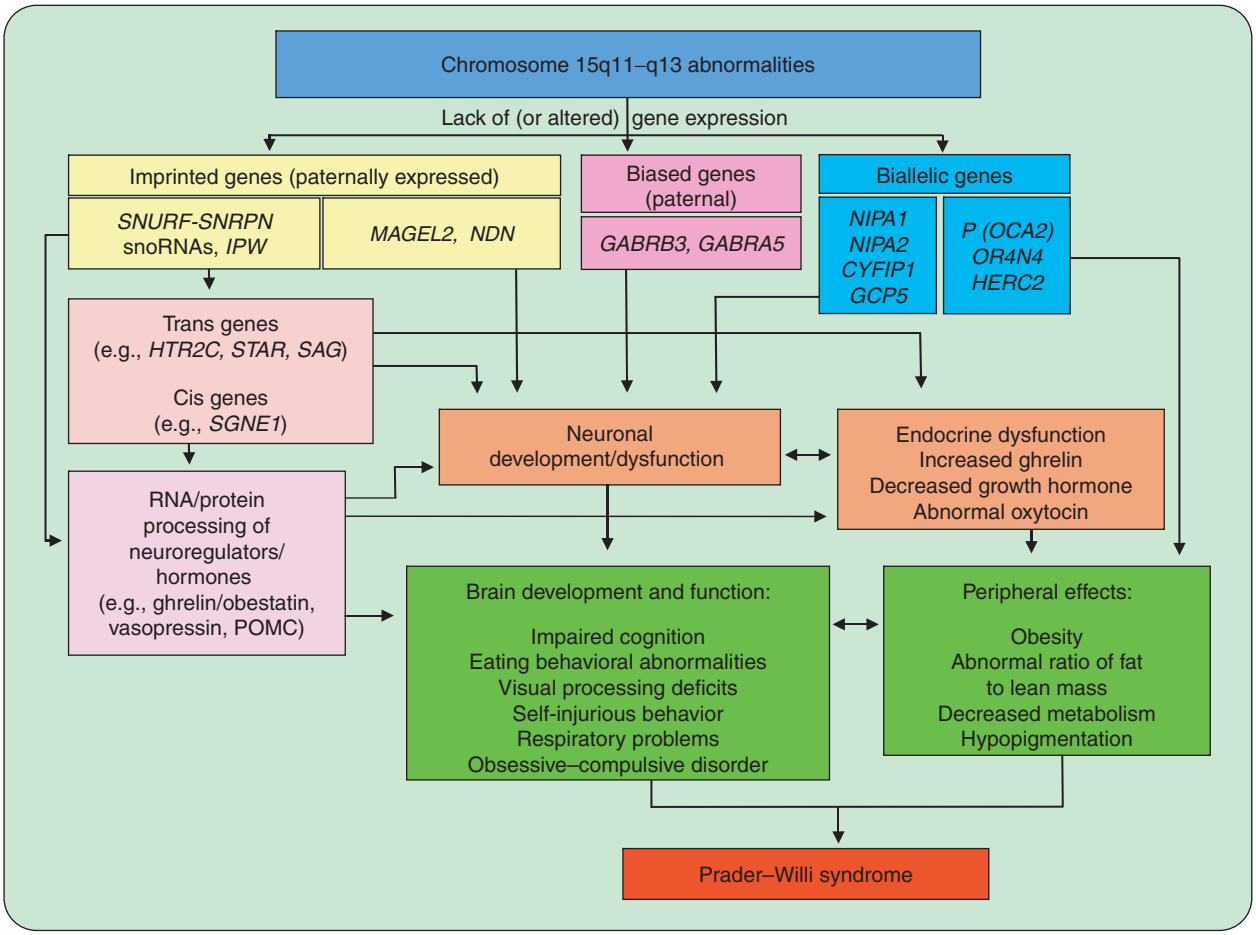

Figure 5 Speculative diagram illustrating possible interconnected and interactive mechanisms leading to Prader–Willi syndrome. Adapted from Bittel DC and Butler MG (2005) Prader–Willi syndrome: Clinical genetics, cytogenetics and molecular biology. *Expert Reviews in Molecular Medicine* 7: 1–20, by permission of Cambridge University Press.

of food for consumption. Interestingly, two olfactory receptor genes (*OR4M2* and *OR4N4*) are located proximally at the boundary of the PWSCR on chromosome 15 (UCSC Genome Browser). If these genes are misregulated or perturbed in PWS due to genetic rearrangements, they might contribute to reduced olfactory sensitivity or normal olfaction. This disturbance could lead to failure to discriminate between appropriate and inappropriate food items for consumption. Therefore, the expression of the olfactory receptor genes might be affected more in the larger type I deletion. One could propose to investigate olfactory receptor gene expression patterns in PWS and identify differences that correlate with food-seeking and food-preference behavior. In addition, proteomic technology in brain and peripheral tissues may help in elucidating the role of deleted or perturbed genes in neurobehavioral outcomes in PWS.

There is increasing evidence that the loss of expression of *SNRPN* and snoRNAs causes some or all features recognized in PWS, although these genes do not appear to be directly causal. As described previously, these genes are involved in processing/splicing of other genes, particularly in the hypothalamus, where they are known to be paternally expressed. There are a number of nonimprinted genes acting in the hypothalamus that are involved in appetite and behavior regulation with complex processing and splicing required prior to translation (e.g., ghrelin, ghrelin receptor, and vasopressin receptor). It is possible that faulty production of mature mRNA of these and/or other downstream genes as a result of the loss of *SNRPN* function and/or snoRNAs could produce PWS (Figure 5). Genomic and proteomic tools and bioinformatics will undoubtedly identify more interactive genes outside the 15q11–q13 region that are affected by incorrect posttranscriptional or post-translational processing in PWS, possibly leading to treatment modalities.

See also: Angelman Syndrome; Food and Water Intake: Regulation; Hormonal Signaling to the Brain for the

Control of Feeding/Energy Balance; Neuropeptides: Food Intake.

Further Reading

Bittel DC and Butler MG (2005) Prader–Willi syndrome: Clinical genetics, cytogenetics and molecular biology. *Expert Reviews in Molecular Medicine* 7: 1–20.

Bittel DC, Kibiryeva N, and Butler MG (2006) Expression of 4 genes between chromosome 15 breakpoints 1 and 2 and behavioral outcomes in Prader–Willi syndrome. *Pediatrics* 118: e1276–e1283.

Butler MG (2002) Imprinting disorders: Non-Mendelian mechanisms affecting growth. *Journal of Pediatric Endocrinology & Metabolism* 15(supplement 5): 1279–1288.

Butler MG, Bittel DC, Kibiryeva N, Talebizadeh Z, and Thompson T (2004) Behavioral differences among subjects with Prader–Willi syndrome and type I or type II deletion and maternal disomy. *Pediatrics* 113: 565–573.

Butler MG, Lee PDK, and Whitman BY (2006) *Management of Prader–Willi Syndrome,* 3rd edn. New York: Springer.

Cassidy SB and Morris CA (2002) Behavioral phenotypes in genetic syndromes: Genetic clues to human behavior. *Advances in Pediatrics* 49: 59–86.

Dimitropoulos A, Feurer ID, Roof E, et al. (2000) Appetitive behavior, compulsivity, and neurochemistry in Prader–Willi syndrome. *Mental Retardation and Developmental Disabilities Research Reviews* 6: 125–130.

Goldstone AP (2004) Prader–Willi syndrome: Advances in genetics, pathophysiology and treatment. *Trends in Endocrinology and Metabolism* 15: 12–20.

Hanel ML and Wevrick R (2001) The role of genomic imprinting in human developmental disorders: Lessons from Prader–Willi syndrome. *Clinical Genetics* 59: 156–164.

Kantor B, Makedonski K, Green-Finberg Y, Shemer R, and Razin A (2004) Control elements within the PWS/AS imprinting box and their function in the imprinting process. *Human Molecular Genetics* 13: 751–762.

Kantor B, Shemer R, and Razin A (2006) The Prader–Willi/Angelman imprinted domain and its control center. *Cytogenetic and Genome Research* 113: 300–305.

Nicholls RD and Knepper JL (2001) Genome organization, function, and imprinting in Prader–Willi and Angelman syndromes. *Annual Review of Genomics and Human Genetics* 2: 153–175.

Ohta T, Gray TA, Rogan PK, et al. (1999) Imprinting-mutation mechanisms in Prader–Willi syndrome. *American Journal of Human Genetics* 64: 397–413.

Roof E, Stone W, MacLean W, Feurer ID, Thompson T, and Butler MG (2000) Intellectual characteristics of Prader–Willi syndrome: Comparison of genetic subtypes. *Journal of Intellect Disability Research* 44(part 1): 25–30.

Swaab DF (2004) Neuropeptides in hypothalamic neuronal disorders. *International Review of Cytology* 240: 305–375.

Relevant Websites

http://www.GeneTests.org – GeneTests.

http://www.ipwso.org – International Prader–Willi Syndrome Organization.

http://www.ncbi.nlm.nih.gov – National Center for Biotechnology Information.

http://www.pwsausa.org – Prader–Willi Syndrome Association.

http://genome.ucsc.edu – University of California at Santa Cruz Genome Browser.

Prediction Errors in Neural Processing: Imaging in Humans

T Lohrenz and P R Montague, Baylor College of Medicine, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the mid-1990s several studies strongly suggested that recordings from single midbrain dopamine neurons in monkeys engaged in conditioning tasks could be explained by a computational model. This model, temporal difference (TD) learning, imported from the machine learning community, was used to model the phasic firing of single midbrain dopamine neurons in alert monkeys harvesting rewards in a Pavlovian conditioning task. In this task there were several salient features (Figure 1). Before training, the receipt of unpredicted reward – the unconditioned stimulus (UCS) – caused dopamine neuron firing. Also before training, the conditioned stimulus (CS) caused no firing in the dopamine neurons, while the receipt of reward still did. After training, the CS caused dopamine firing, but the receipt of the UCS no longer did, while if the predicted reward was omitted, the firing rate fell. How does TD learning capture these features of the data?

Brief Review of the TD Algorithm

TD learning is a solution method for a class of statistical learning models called Markov decision problems (MDPs). More formally, the basic setup for a Markov decision problem is given by a set of states S, a set of actions A, transition functions $P^a_{ss'}$ giving the probability of moving from state s to state s' given action a, and the expected return $R^a_{ss'}$ from taking action a and transitioning from s to s'. A policy π is a function from states to actions (or more generally a probability distribution on actions). The value of a given state s under a certain policy π is given by the expected discounted reward to be obtained from starting at state s and following policy π:

$$V^\pi(s) = E_\pi\left[\sum_{i=1}^{\infty}\gamma^{i-1}r_i\right] \qquad [1]$$

The goal is to learn the function V^π. TD learning is a way to learn the value function while executing the policy π. Suppose in state s the action is given by $a = \pi(s)$ and we transition from $s \to s'$. Then if V denotes our approximation to the value function V^π, an improved approximate value function at state s is then given by

$$V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)] \qquad [2]$$

The expression inside the brackets is the TD error term, δ. Two parameters appear here: α, the learning rate, and γ, the discount factor ($0 < \gamma < 1$).

Under assumptions on α this process will converge to V^π. Once one has a value function V^π one can improve the policy π, since the real aim is to find the best policy π and its value function V^*. A convenient change in point of view is to look at the value of a state action pair rather than just the state value. The letter Q is usually reserved for this function and is defined by

$$Q^\pi(s,a) = E_\pi\left[\sum_{i=1}^{\infty}\gamma^{i-1}r_i \,\middle|\, s_0 = s, a_1 = a\right] \qquad [3]$$

In words, $Q(s,a)$ is the value of taking action a in state s and then following policy π.

This formulation leads directly to the method of policy improvement. If we have in some state that for some action a, $Q^\pi(s,a) > V^\pi(s)$, then we can modify the policy π to a policy π' so that $\pi'(s) = a$ (and equals π for all other states) and in fact $V^{\pi'}(s) \geq V^\pi(s)$ for all s. A general method to find an optimal value function is generalized policy iteration: start with an arbitrary policy π, and approximation to $\hat{V}^\pi(s)$. Take the action under policy π, receive the reward r, and make the state transition $s \to s'$. Compute the TD error as given previously, update the value function, and then make the policy improvement step to get an improved policy π'. Now repeat. This process will converge to the optimal policy π^* and value function V^*.

There are a few variations on this idea that are important to consider. The first uses the Q values directly to find an optimal policy. This algorithm is called SARSA (state, action, reward, state, action) and goes as follows. Start with an arbitrary Q. Decide on a way to go from Qs to an action. (There are two popular ways to pick an action: (1) ε-greedy, in which the action a with largest $Q(s,a)$ is chosen with probability $1 - \varepsilon$, and the rest of the actions are chosen with probability $\varepsilon/n - 1$ if there is a total of n possible actions; and (2) softmax, or Boltzmann, in which the probability of choosing an action a in state s is given by

$$P(a,s) = \frac{e^{\beta Q(s,a)}}{\sum_{\tilde{a}} e^{\beta Q(s,\tilde{a})}} \qquad [4]$$

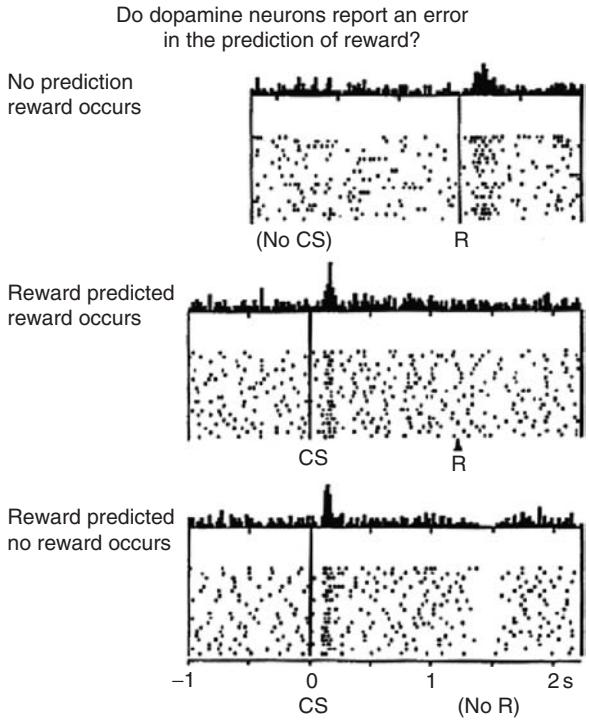

Figure 1 Raster plots of recordings from monkey dopamine neurons during a Pavlovian conditioning task. Top: Before training, no increase in spiking at cue (conditioned stimulus; CS), but increase at delivery of juice reward (R; unconditioned stimulus). Middle: After training, increased spiking at presentation of CS, and no increase at reward. Bottom: After training, decreased spiking at unpredicted omission of unconditioned stimulus (no R). From Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

where β is a parameter (the inverse temperature) that determines how sensitive the probability is to difference in the Qs. As $\beta \to 0$, the distribution becomes uniform; as $\beta \to \infty$, the distribution localizes on the action with the biggest Q.) Pick the action, receive the reward r, make the transition to the new state s', and update Q using the TD algorithm:

$$\begin{aligned} Q(s,a) &\leftarrow Q(s,a) + \alpha\delta \\ \delta &= r + \gamma Q(s',a') - Q(s,a) \end{aligned} \qquad [5]$$

where the action a' is the action that is determined by the policy determined by Q. SARSA is an on-policy algorithm because the Qs are updated using the current policy.

An off-policy algorithm using the Q functions is called Q-learning. In this algorithm, the policy used to explore the state space is divorced from the policy determined by the Qs. All that matters is that the sampling policy samples the state-action space 'well enough.' The actual algorithm is given as follows: choose action a according to sampling policy, receive reward r, transition to state s', and update Q according to

$$\begin{aligned} Q(s,a) &\leftarrow Q(s,a) + \alpha\delta \\ \delta &= r + \gamma \max_{\tilde{a}} Q(s',\tilde{a}) - Q(s,a) \end{aligned} \qquad [6]$$

Another reinforcement learning algorithm worth mentioning at this point is the 'actor–critic' architecture. In this architecture, the value function is maintained by the critic, while the policy is maintained by a separate actor. The standard TD error signal is used both to update the value function estimate and to criticize the policy. For example, if the policy is encoded by a preference function $p(s,a)$, the preference for action a while in state s, and the probability of taking in action a while in state s is given by the softmax formula, then after an action a is taken and a reward r is received we have the updates:

$$\begin{aligned} \text{critic}: \quad & V(s) \leftarrow V(s) + \alpha_1\delta \\ & \delta = r + \gamma V(s') - V(s) \\ \text{actor}: \quad & p(s,a) \leftarrow p(s,a) + \alpha_2\delta \end{aligned}$$

One final algorithm which has been used in functional magnetic resonance imaging (fMRI) work is advantage learning. In advantage learning, the advantage of an action a in a state s is the key quantity:

$$A(s,a) = Q(s,a) - V(s) \qquad [7]$$

After taking an action a, $V(s)$ is updated as in the basic TD algorithm, while the advantage is updated by the error signal:

$$\begin{aligned} \delta_A &= r + \gamma V(s') - Q(s,a) \\ &= r + \gamma V(s') - V(s) - [Q(s,a) - V(s)] \\ &= \delta_{TD} - A(s,a) \end{aligned} \qquad [8]$$

One can see that this is the correct update by observing that $r + \gamma V(s')$ is the new estimate of $Q(s,a)$ after taking action a and then receiving reward r. δ_A is the change in the advantage coming from the update to $Q(s,a)$, while the rest of the update is handled by updating V. In the simple Pavlovian task described earlier, the model is given as follows. States are represented by the vector $\vec{x}(t)$, the components of which represent the presentation of the CS in time. For example, if the CS is presented at $t = 0$,

$$\begin{aligned} \vec{x}(0) &= (1,0,0,0,0,0) \\ \vec{x}(1) &= (0,1,0,0,0,0) \\ \vec{x}(2) &= (0,0,1,0,0,0)\ldots \end{aligned} \qquad [9]$$

The value function is given by a linear representation in weights: $\vec{w}\,(t) = (w_1, w_2, \ldots, w_n)$,

$$V(t) = V(\vec{x}(t)) = \vec{w}(t) \cdot \vec{x}(t) \qquad [10]$$

The TD error is given by

$$\delta_t = r_t + \gamma V(\vec{x}(t+1)) - V(\vec{x}(t)) \qquad [11]$$

and the value function update is given by updating the linear weights:

$$\vec{w} \leftarrow \vec{w} + \alpha \cdot \vec{x}(t) \cdot \delta_t \qquad [12]$$

The following example illustrates two important points of TD learning: (1) The TD error response moves backward to the CS, and (2) the (TD error) response at the delivery of the UCS decreases as it becomes predicted. We leave it as an exercise for the reader to continue these calculations until the response occurs at the CS, and is predicted away at the delivery of the UCS.

Suppose that each trial lasts 6 time units, and that the UCS is delivered at $t = 5$. Set the weights initially at $\vec{w} = 0$. The CS appears at $t = 0$, so that the states are

$$[1] \quad \begin{aligned} \vec{x}(0) &= (1,0,0,0,0,0) \\ \vec{x}(1) &= (0,1,0,0,0,0) \\ \vec{x}(2) &= (0,0,1,0,0,0) \\ \vec{x}(3) &= (0,0,0,1,0,0) \\ \vec{x}(4) &= (0,0,0,0,1,0) \\ \vec{x}(5) &= (0,0,0,0,0,1), \end{aligned}$$

and since $\vec{w} = 0$, and no reward is obtained until $t = 5$, $V(\vec{x}(t)) = 0$, $t = 0, 1, 2, 3, 4, 5$.

On the first trial the only interesting thing happens at $t = 5$. Here $r = 1$, so that

$$[2] \quad \begin{aligned} \delta_5 &= r + \gamma V(\vec{x}(6)) - V(\vec{x}(5)) \\ &= 1 + \gamma \vec{w} \cdot \vec{x}(6) - \vec{w} \cdot \vec{x}(5) \\ &= 1, \end{aligned}$$

and thus

$$[3] \quad \begin{aligned} \vec{w}_{new} &= w_{old} + \alpha \cdot \vec{x}(5) \cdot \delta_5 \\ &= \vec{0} + \alpha \cdot (0,0,0,0,0,1) \cdot 1 \\ &= (0,0,0,0,0,\alpha). \end{aligned}$$

On the second trial, something interesting happens at $t = 4$: the signal 'moves back' toward the CS. At $t = 4$, no reward occurs, but the TD update is given by

$$[4] \quad \begin{aligned} \delta_4 &= r + \gamma V(\vec{x}(5)) - V(\vec{x}(4)) \\ &= 0 + \gamma \cdot (0,0,0,0,0,1) \cdot (0,0,0,0,0,\alpha) \\ &\quad - (0,0,0,0,1,0) \cdot (0,0,0,0,0,\alpha) \\ &= \gamma \cdot \alpha. \end{aligned}$$

The weights are updated accordingly:

$$[5] \quad \begin{aligned} \vec{w}_{new} &= \vec{w}_{old} + \alpha \cdot \vec{x}(4) \cdot \delta_4 \\ &= (0,0,0,0,0,\alpha) + \alpha \cdot (0,0,0,0,1,0) \cdot \gamma \cdot \alpha \\ &= (0,0,0,0,\gamma \cdot \alpha^2, \alpha). \end{aligned}$$

At $t = 5$, a reward is again received, and we get the TD error:

$$[6] \quad \begin{aligned} \delta_5 &= 1 + \gamma V(\vec{x}(6)) - V(\vec{x}(5)) \\ &= 1 - (0,0,0,0,\gamma \cdot \alpha^2, \alpha) \cdot (0,0,0,0,0,1) \\ &= 1 - \alpha. \end{aligned}$$

TD Error Signals in Humans: fMRI Evidence

While the original impetus for the TD error theory came from single-unit recordings in nonhuman primates, the advent of sophisticated fMRI paradigms has allowed researchers to hunt for TD error signals in the human brain. In this section we review the results some of these experiments.

Early Developments

A relatively early study that captured an aspect of the TD error signal in the human brain was undertaken by Berns and colleagues in 2001. The TD error algorithm is able to distinguish errors in both timing and type or amount of reward. Hence, rewards that are unpredictable in both timing and type should generate a larger reward signal than that generated by predictable rewards. Berns et al. used a 2×2 design to test this hypothesis. More precisely, study participants received squirts of either juice or water (and were queried afterward as to which they preferred), and received the squirts either in a regular sequence or in an unpredictable fashion (randomized over both type and time). The main finding of this study supported the hypothesis: the contrast of unpredictable–predictable showed significant activity in the bilateral nucleus accumbens/ventral striatum, areas known to be direct projections of dopamine neurons in the substantia nigra pars compacta (SNc) and the ventral tegmentum area (VTA). It should be stressed at this point that none of these experiments (save one discussed in the last section) will provide any direct evidence that these activations are the result of the phasic firing of dopamine neurons in the SNc or VTA. The nonhuman primate evidence suggests that the midbrain dopamine (DA) signals encode TD error signals (and we will examine some recent results from single-unit recordings in monkeys later), and the dopamine neurons in these regions project into the striatum, but it is conceivable that TD error signals found in the striatum (or elsewhere) could be the result of the combination of signals from the DA system (e.g., for positive errors) and, for example, the serotonin system (for negative errors), as has been suggested (and see later).

Another early result was that of Pagnoni and co-workers in 2002. In an instrumental design,

study participants were shown a visual cue, then pushed a button, and then 4 s later, juice was delivered. In a small percentage of the trials the juice squirt was omitted until 8 s later, thus setting up the conditions for a prediction error at the time of the omitted juice squirt. In the fMRI analysis, the authors examined a region of interest (ROI) centered on the nucleus accumbens (NAcc). Indeed, they found a significant difference between the activations in the delivered and undelivered conditions at the time of the juice delivery (more precisely, 6–8 s after the juice delivery time, the approximate peak hemodynamic response time for an event at the delivery time).

The experiments of Berns et al. and Pagnoni et al., while suggestive, had shortcomings. The Berns et al. design changed two things at one time: both the type of reward and the reward time were changed from predictable to unpredictable. In the Pagnoni study, the instrumental nature of the design (the button press) makes the interpretation less clear as compared to a passive task. McClure and colleagues, in 2003, attempted to address these issues in a passive design that incorporated only unpredictability in time. In their design a light cue came to predict a squirt of juice (see Figure 2). After a series of normal training trials in which juice was delivered 6 s after the cue, six catch events were inserted randomly among normal trials. In these catch trials, the juice expected 6 s after the light cue was omitted, and then was delivered 4 s later. In this design, two hypotheses are generated from a prediction-error model: (1) activity in the striatum at unexpected delivery of juice 4 s after the omission of juice at the expected time should be significantly higher than the activity at the expected delivery of juice (evidence of positive prediction error); (2) activity in the striatum at the time of omission of expected juice should be greater than that 4 s after the expected delivery of juice (evidence of negative prediction error). This is exactly what the authors found.

While the preceding studies give convincing evidence that the signal in the ventral striatum is consistent with a prediction error signal, they do not explicitly model the TD error signal during the conditioning process, and regress this signal with the neural activity. In 2003, O'Doherty et al. addressed these issues in a Pavlovian conditioning task (see Figure 3). Their paradigm trials consisted of the presentation of one of three visual stimuli followed 3 s later by a squirt of liquid that was either sweet (positive), saliva-like (neutral), or of no taste (negative). After a number of pairings of the positive cue with reward, subsequent positive cues were sometimes followed by the omission of the reward. Additionally, sometimes after the negative cue, a reward was delivered. A TD model of the each participant's experience using a model is almost identical to that described at the end of the introductory modeling section (see previous section, 'Brief review of the TD algorithm'). Two learning rates (previously denoted α) were chosen as 'biologically plausible, $\alpha = 0.2, 0.7$ (since no decisions were made, there are no behavioral data to fit and optimize α; this will not be true in the following discussion), and the discount factor was set at $\gamma = 0.99$ (the future really matters a lot). The principal finding of this study is that the TD signal obtained from a computational model correlated significantly with the actual blood oxygen level-dependent (BOLD) signal in the ventral striatum and the orbitofrontal cortex. In particular, the authors tested both a signed prediction error model and a model using the absolute value of the prediction error. The signed version produced the results already reported, while

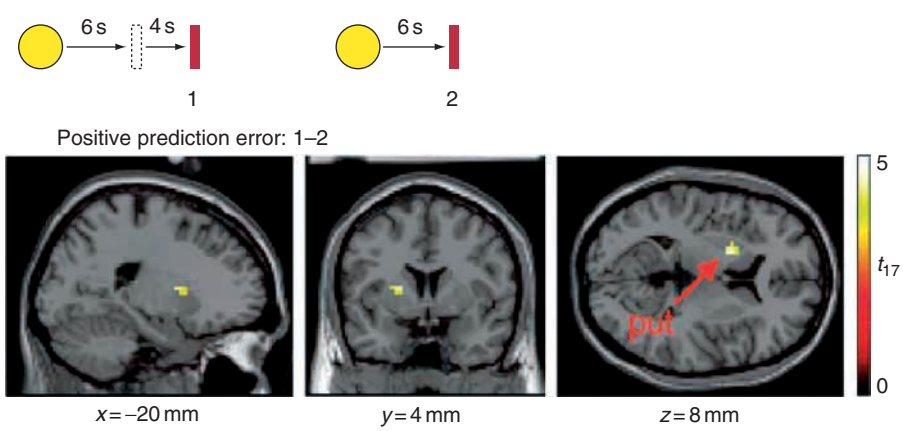

Figure 2 Functional magnetic resonance image, showing activation from temporal difference error signal in a Pavlovian conditioning task: area in ventral striatum showing differential activation from unexpected juice delivery versus expected juice delivery (put, putamen). See text for discussion. From McClure SM, Berns GS, and Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38: 339–346.

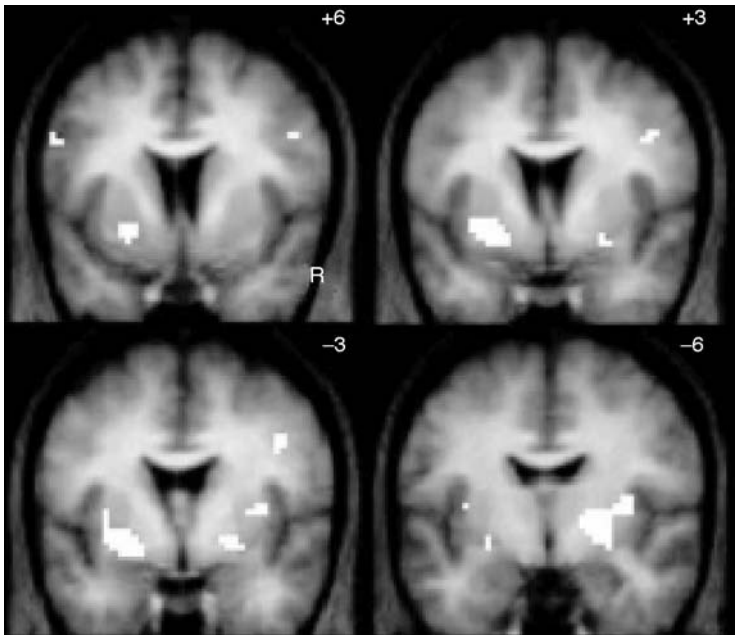

Figure 3 Activation in ventral striatum from temporal difference error signal in a Pavlovian conditioning task: coronal sections, moving from anterior to posterior ($y = 6, 3, -3, -6$). From O'Doherty J, Dayan P, Friston K, et al. (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38: 329–337.

the absolute value produced significant activation only in the inferior prefrontal gyrus. The different learning rates produced qualitatively similar results. It should be noted that the results of this study seem to contradict those of Pagnoni et al. In particular, this study (as well as that of McClure et al.) shows that an undelivered expected reward produces a negative prediction error in the ventral striatum. As the authors of the study under consideration have noted, the Pagnoni study focused solely on the NAcc, an area with no reported activations. However, Knutson and colleagues, in 2001, did report a deactivation on omitted monetary reward.

Refinements and Questions

Aside from the Pagnoni study, none of the above paradigms involves an actual decision. Even in the Pagnoni task the choice was only about when the participant was ready for the juice delivery. Presumably the DA system is not just about reward prediction, but is also about making choices. In the aforementioned computational models, discussed in the review of the TD algorithm, the analysis split between the value of a state under a certain policy V^{π} and the policy π itself. Indeed, one architecture discussed, the actor–critic, explicitly splits the functions of state value estimation (critic) and policy (actor). Previous evidence (by O'Doherty, in 2004) had suggested a division along these lines in the striatum – dopaminergic projections to the ventral striatum corresponding to the critic, and dopaminergic projections to the dorsal striatum corresponding to the actor (or error signals to the actor). O'Doherty and co-workers performed a study to explicitly test this hypothesis (see Figure 4). In their study, a Pavlovian conditioning task was intermixed with an instrumental task. In the instrumental task, participants had to choose between two fractal images: in the reward condition, one image represented a high probability of a juice reward (60%) while the other represented a low probability of reward (30%). The unrewarded case was similar except that the stimuli were different, and the reward was no longer a sweet juice but was instead a neutral liquid. The Pavlovian trials were randomly distributed throughout the experiment and were identical except that the computer, rather than the participant, made the choice between the two stimuli.

For the Pavlovian condition, a TD model similar to the one discussed earlier was constructed. For the instrumental condition, an advantage-learning model was fit. The result was that the ventral striatum was activated in both the Pavlovian and the instrumental conditions, while the dorsal striatum was active only in the instrumental condition. Interestingly, the authors selected the advantage-learning algorithm versus the actor–critic architecture on the basis of behavioral fits: the advantage-learning model fit actual behavior better on the basis of the log-likelihood of the various fits.

An important question is to what extent these previous results hinge on the fact that the rewards were

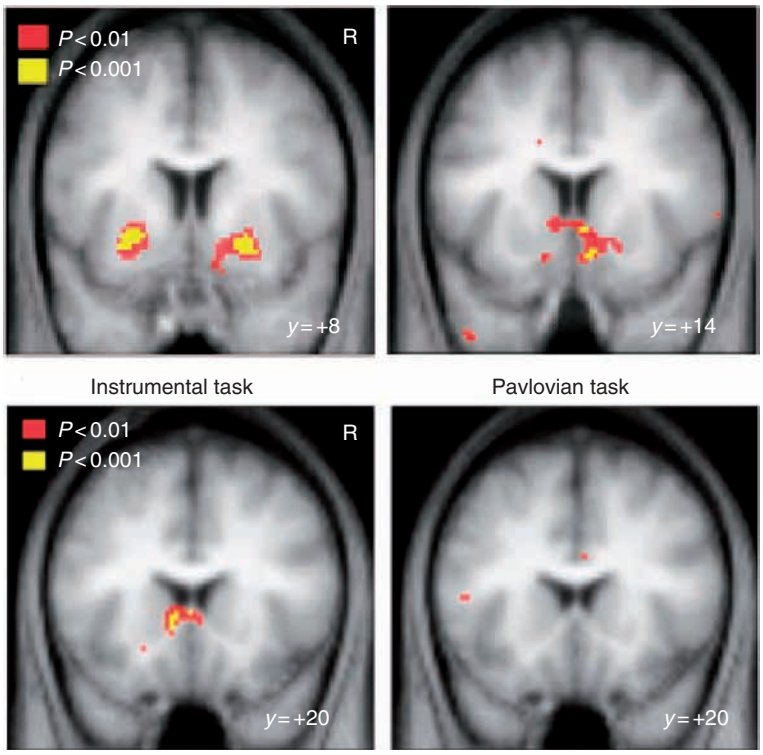

Figure 4 Functional magnetic resonance image, showing activation in ventral and dorsal striatum in an instrumental task. Top: Activation of ventral striatum in a Pavlovian task (left) and an instrumental task (right). Bottom: Activation in dorsal striatum in an instrumental task (left) and a Pavlovian task (right). See text for discussion. From O'Doherty JP, Dayan P, Schultz J, et al. (2004) Dissociable roles of dorsal and ventral striatum in instrumental conditioning. *Science* 304: 452–454.

all in fact appetitive. Does TD learning (or a variant) work for aversive stimuli? Additionally, the previous studies were first-order conditioning paradigms. Does TD learning work for higher order conditioning? Seymour and colleagues address this question, in 2004, in a higher order Pavlovian design with pain as the unconditioned stimulus. The design was higher order because there were two sequential visual cues, followed by a high- or low-intensity shock. If the first two possible cues are denoted A and C, and the second two are denoted B and D, then the sequences led to $A, B \rightarrow$ high pain, $A, D \rightarrow$ low pain, $C, B \rightarrow$ high pain, $C, D \rightarrow$ low pain. A basic (i.e., value function) TD learning algorithm was used to model the brain responses. Six states for three time points were used: s_1 for cue A at time 1, s_2 for cue C at time 1, s_3 for cue B at time 2, s_4 for cue D at time 2, s_5 for high pain ($r = 1$) at time 3, and s_6 for low pain ($r = 0$) at time 3. The prediction error signal (taken at the three time points) was used as a regressor in the fMRI analysis, and showed activation in the bilateral ventral putamen, right head of caudate, left substantia nigra, right insula, and bilateral cerebellum, areas broadly consistent with the findings from the appetitive cause. Notably, the predicted prediction error signal included both positive and negative activations, and these (suitably normalized) predictions were borne out in ROI analysis.

A related question is whether avoiding an aversive UCS carries a reward signal similar to an actual appetitive reward. Avoidance learning poses a puzzle for reinforcement learning, since, on the face of it, successfully avoiding an aversive experience means the absence of a UCS to reinforce behavior. Kim and colleagues tackled this issue in 2006, using an advantage learning model for analysis of the fMRI and behavioral data. In their paradigm, study participants are presented with two visual stimuli patterns which represent choices. In the reward condition, one pattern represents a 60% reward probability (60% chance of earning \$1, 40% no-change outcome) and the other represents a 30% chance. In the avoidance condition, one pattern represents a 60% chance of a neutral outcome (60% chance of no-change, 40% chance of losing \$1) and the other represents a 30% chance of a neutral outcome. There were also two control conditions in which participants made a choice to get either neutral feedback or no feedback at all. The first finding was that medial orbitofrontal complex/ventral medial prefrontal cortex (OFC/VMPFC) showed activity that correlated with reward receipt and loss

avoidance (i.e., activity was positive for received rewards and avoided losses, and activity was decreased for missed rewards or losses received). This finding is in line with other studies of the outcome phase of reward paradigms. The second finding concerns the prediction error signal. Separate advantage learning models were estimated from the behavioral data by subject and for the rewarding, aversive, and neutral conditions, and were entered as regressors at the time of outcome. The reward prediction error signal correlated significantly with activity in the ventral striatum, whereas the aversive prediction error signal (for the case when the receipt of an unexpected loss gave a positive signal) correlated strongly with activity in the insula. The authors found no significant activity correlating with a prediction error signal such that an unexpected aversive signal predicted decreased activity. The OFC/VMPFC result is positive, in the sense that previously identified valuation areas treat an avoided loss as a reward. On the other hand, the prediction error result (as the authors noted) is in contradiction to other results, including the work by Seymour et al. in 2004, as discussed previously. Also, since the authors used an advantage learning model in an instrumental task, one would expect to see, as in the work of O'Doherty et al., activity in the dorsal striatum associated with the 'actor' in the actor–critic architecture. Why this activity is not present in this experiment is not readily apparent.

Yacubian did an fMRI study in 2006, investigating the differences between gain and loss conditions (see Figure 5). In this work, individuals were essentially presented with high- and low-probability conditions of making either 1 or 5 euros, with the conjugate probabilities of losing the same amount. For example, in the high-probability (5 euros) condition, the probability of winning 5 euros was 0.66, while the probability of losing 5 euros was $1 - 0.66 = 0.33$. In any

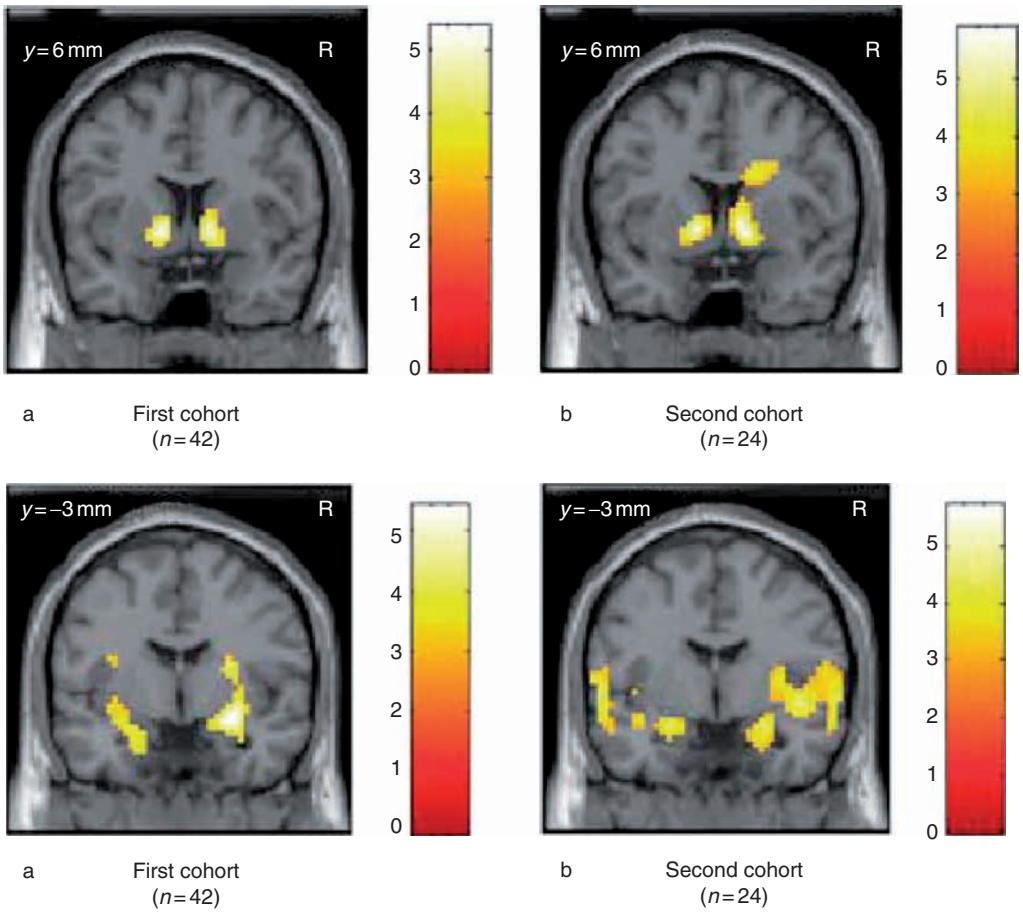

Figure 5 Dissociable functional magnetic resonance image activation from positive and negative parts of expected value and prediction errors during a guessing task. Top: Activation from the positive part of expected value during the anticipation phase (a) and the positive part of prediction error during the outcome phase (b). Bottom: Activation from the negative part of expected value during the outcome phase (a) and the negative part of prediction error during the outcome phase (b). From Yacubian J, Glascher J, Schroeder K, et al. (2006) Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *Journal of Neuroscience* 26: 9530–9537.

trial, the expected value can be expressed as the sum of a positive part and a negative part:

$$EV = p_{\text{win}} \cdot V - (1 - p_{\text{win}}) \cdot V \equiv EV_+ + EV_-$$

where V is the amount at stake (1 or 5 euros). Yacubian also defined positive and negative prediction errors:

$$\delta_+ = R - EV_+$$
$$\delta_- = R - EV_-$$

where R is the actual outcome of the trial. The main result is that EV_+ (during an anticipation phase) and δ_+ (during the outcome phase) are localized in the ventral striatum, while EV_- and δ_- are localized in the amygdala.

The TD algorithm hypothesis of reinforcement learning makes a strong qualitative prediction about the presentation of a second stimulus after a first conditioned stimulus has been strongly associated with reward. Since the appearance of the reward is already fully predicted by the first stimulus, the second stimulus acquires no meaning, since no prediction error is generated. This prediction, blocking, has previously been investigated in single-unit recordings of dopamine neurons in monkeys. In particular, the firing patterns of the individual dopamine neurons supported the predictions of the TD learning algorithm. Tobler et al. investigated this prediction of TD learning theory, in 2006, in an fMRI experiment in humans. In this paradigm, individuals were trained to expect reward after a visual stimulus A, and to expect nothing after B. Next they paired a new stimulus X with A, still predicting reward, and another new stimulus Y with B, now followed by reward. Test trials were then introduced, whereby X and Y were presented by themselves and were not followed by reward. Blocking predicts that the response to X will basically be flat (it is blocked), while Y will show a response at the presentation of the stimulus, and a decrease (negative prediction error) at the expected time of the reward (which is omitted). This is the main finding of Tobler's work, and provides strong evidence in favor of TD learning which does not rely on the details of a particular model.

Discussion

While the preceding discussions paint a picture that is supportive of the TD learning hypothesis, we have seen some ambiguities and contradictions. There are competing models (value learning, actor–critic, Q-learning, advantage learning), the Pagnoni result for prediction error is the wrong sign (perhaps the NAcc carries a different signal), and the Kim study contradicts the Seymour study, while not reproducing the result of O'Doherty. These difficulties we see getting worked out as more refined experiments are conducted and repeated. However, on the modeling front there are two particular issues that seem worth discussing. The first is incorporating a model of the task into the reinforcement learning (RL) algorithm. Hampton et al. took a step in this direction in a 2006 study of choice in a reversal learning task (see Figure 6). In this work, they compare a variant of the Q-learning model with a Bayesian hidden Markov model (BHMM). The BHMM incorporates a hidden state variable with Bayesian updating to model an individual's choices. One key prediction of this model is that value signals of both alternatives are updated at the same time (to capture the constraint of the model, which is that, if one alternative has x probability of being 'good,' then the other choice has $1 - x\%$ chance of being good). Thus if your current choice is 'bad,' and you decide to switch, the value goes up, which is not the same as RL, since the value of a choice not recently selected has not been updated. In this model, the posterior, prior difference becomes the error signal.

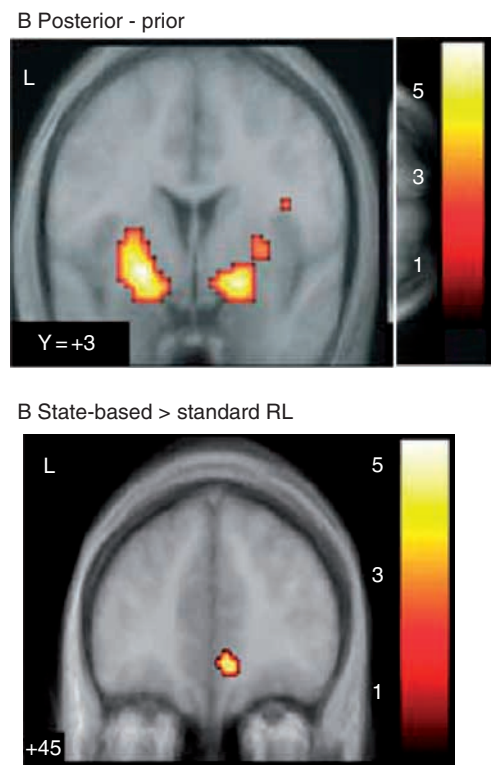

Figure 6 Top: Error signal. Bottom: Activation in the medial prefrontal cortex is greater in the state-based model, compared to standard reinforcement learning (RL). See text for discussion. From Hampton AN, Bossaerts P, and O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision-making in animals. *The Journal of Neuroscience* 26: 8360–8367.

Another source of information that could easily be incorporated in to the RL framework is what could be called fictive errors: the difference between what could have been obtained and what was actually received, rather than the difference between what was received and what was expected. In many of the tasks that have been considered experimentally, only the outcome of the choice actually made is revealed, but in reality this is often not the case: we often have access to outcome that we could have received if we had made a different choice. Indeed, regret is an emotion we often associate with this putative error signal. Several experiments have touched upon this topic. In 2005, Kuhnen and Knutson used a form of fictive error as an fMRI regressor to analyze choice behavior in a gambling task in which more than one outcome is revealed. In the same year, Coricelli et al. also considered a regret or fictive error term in their fMRI analysis of a choice task in which multiple outcomes were revealed. In the behavioral realm, in 1999, Camerer and Ho used foregone alternatives to update the value of choices in their experiential weighted attraction model.

Exciting possibilities remain for the study of RL, and in particular TD models, in human choice behavior. Incorporating more detailed task structure into the models, as well as information from foregone choices, in a coherent model that matches both choice and fMRI data remains a challenge. Furthermore, the issue of how gains and losses are encoded still remains to be settled.

See also: Conditioning: Theories; Learning, Action, Inference and Neuromodulation; Procedural Learning: Classical Conditioning; Reinforcement Models; Reward and Learning; Reward Decision-Making.

Further Reading

Baird LC (1995) Advantage learning. *Technical Report, WL-TR-93–1146*. Wright-Patterson Air Force Base, OH: Wright Laboratory.

Berns GS, McClure SM, Pagnoni G, et al. (2001) Predictability modulates human brain response to reward. *Journal of Neuroscience* 21: 2793–2798.

Camerer CF and Ho T (1999) Experience – weighted attraction and learning in normal form games. *Econometrica* 67: 827–874.

Daw N, Kakade S, and Dayan P (2002) Opponent interactions between serotonin and dopamine. *Neural Networks* 15: 603–616.

Hampton AN, Bossaerts P, and O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision-making in animals. *The Journal of Neuroscience* 26: 8360–8367.

Kim H, Shimojo S, and O'Doherty JP (2006) Is avoiding an aversive outcome rewarding? The neural substrates of avoidance learning in the human brain. *PLoS Biology* 4(8): 1453–1461.

Knutson B, Fong GW, Adams CM, et al. (2001) Dissociation of reward anticipation and outcome with event-related fMRI. *NeuroReport* 12: 3683–3687.

Kuhnen CM and Knutson B (2005) The neural basis of financial risk taking. *Neuron* 47: 763–770.

McClure SM, Berns GS, and Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38: 339–346.

Montague PR, Dayan P, and Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience* 16(5): 1936–1947.

Montague PR, Hyman SE, and Cohen JD (2004) Computational roles for dopamine in behavioural control. *Nature* 431: 760–767.

O'Doherty J, Dayan P, Friston K, et al. (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38: 329–337.

O'Doherty JP, Dayan P, Schultz J, et al. (2004) Dissociable roles of dorsal and ventral striatum in instrumental conditioning. *Science* 304: 452–454.

Pagnoni G, Zink CF, Montague PR, et al. (2002) Activity in human ventral striatum locked to error of reward prediction. *Nature Neuroscience* 5: 97–98.

Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

Seymour B, O'Doherty JP, Dayan P, et al. (2004) Temporal difference models describe higher-order learning in humans. *Nature* 429: 664–667.

Suttton RS and Barto AG (1998) *Reinforcement Learning*. Cambridge, MA: MIT Press.

Tobler TN, O'Doherty JP, Dolan RJ, et al. (2006) Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology* 95: 301–310.

Waelti P, Dickinson A, and Schultz W (2001) *Nature* 412: 43–48.

Yacubian J, Glascher J, Schroeder K, et al. (2006) Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *Journal of Neuroscience* 26: 9530–9537.

Prefrontal Contributions to Reward Encoding

E T Rolls, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

The (primate and human) orbitofrontal cortex receives taste, olfactory, visual, somatosensory, and auditory inputs from cortical regions which are toward the end of the unimodal cortical processing streams that build representations of what stimulus is present in each of these sensory modalities. In this article, I review evidence from human functional neuroimaging and neuropsychology showing that the orbitofrontal cortex and a region to which it projects, the anterior cingulate cortex (ACC), are involved not only in rewards produced by olfactory and taste stimuli but also in much more abstract types of reward, such as monetary reward, and in altering behavior when rewards change. This indicates that the orbitofrontal and cingulate cortices are involved in emotion, and this is confirmed by many changes in emotion described here that can result from damage to the human orbitofrontal or ACC.

Functional Neuroimaging of the Human Orbitofrontal Cortex: Reward and Nonreward

Investigations with functional magnetic resonance imaging (fMRI) show that taste and olfactory stimuli can activate the orbitofrontal cortex and that the activation is related to the pleasantness or unpleasantness of the stimuli, that is, to their reward value. For example, activation of a part of the human orbitofrontal cortex is related to the pleasantness of food odor, in that the activation measured with fMRI produced by one food odor, banana, decreased after banana was eaten for lunch to satiety but remained strong to another food odor, vanilla, not eaten in the meal. Further evidence that pleasant odors are represented in the orbitofrontal cortex is that three pleasant odors (linalyl acetate (floral, sweet), geranyl acetate (floral), and alpha-ionone (woody, slightly food-related)) had overlapping activations in the medial orbitofrontal cortex in a region not activated by three unpleasant odors (hexanoic acid, octanol, and isovaleric acid). Moreover, activation of the medial orbitofrontal cortex was correlated with the subjective pleasantness ratings of the odors, and activation of the lateral orbitofrontal cortex was correlated with the subjective unpleasantness of the odors. Pleasant touch has also been shown to activate the medial orbitofrontal cortex, and pain has been shown to activate the lateral orbitofrontal cortex.

Visual stimuli that activate the human orbitofrontal cortex include those associated with pleasant taste. But it has been possible to extend the types of visual conditioned reinforcers to quite abstract reinforcers, such as monetary reward. In an fMRI study, we used a visual discrimination task in which one stimulus was associated with monetary reward and a different visual stimulus with monetary loss (punishment). The actual amounts of money won on reward trials and lost on punishment trials were probabilistic. This part of the design, and the fact that unexpected visual discrimination reversals occurred so that there were trials on which money was lost, enabled us to show that the magnitude of the activation of the medial orbitofrontal cortex was correlated with the amount of money won on each trial and that the magnitude of the activation of the lateral orbitofrontal cortex was correlated with the amount of money lost on each trial. Activation of the medial orbitofrontal cortex is also correlated with face attractiveness. Auditory stimuli may have similar representations in the orbitofrontal cortex related to their affective value. For example, a correlation has been found between subjective ratings of dissonance and consonance of musical chords and the activations produced in the orbitofrontal cortex. Another type of reward, the administration of amphetamine to naive human subjects, also activates the medial prefrontal cortex.

Cognition can reach down into the orbitofrontal cortex to influence affect, as shown by the effects that word descriptors (such as cheese vs. body odor) have on the activations produced by olfactory stimuli in the medial orbitofrontal cortex. Related to the orbitofrontal cortex error neurons, we have also been able to obtain evidence that nonreward used as a signal to reverse behavioral choice is represented in the human orbitofrontal cortex. Kringelbach and Rolls used the faces of two different people; if one face was selected, then that face smiled, and if the other was selected, the face showed an angry expression. After good performance was acquired, there were repeated reversals of the visual discrimination task. It was found that activation of a lateral part of the orbitofrontal cortex in the fMRI study was produced on the error trials, that is, when the human chose a face and did not obtain the expected reward. Control tasks showed that the response was related to the error and the mismatch between what was expected and what was obtained, in that just showing an angry

face expression did not selectively activate this part of the lateral orbitofrontal cortex. An interesting aspect of this study that makes it relevant to human social behavior is that the conditioned stimuli were faces of particular individuals and the unconditioned stimuli were face expressions. Moreover, the study reveals that the human orbitofrontal cortex is very sensitive to social feedback when it must be used to change behavior.

The activations found in functional neuroimaging studies by many types of reward appear to involve relatively medial parts of the human orbitofrontal cortex and unpleasant stimuli or nonreward appear to involve more lateral parts of the human orbitofrontal cortex. Single-neuron recording studies in macaques show that neurons can be exquisitely tuned differently not only to reinforcers in different sensory modalities (taste, smell, touch, etc.) but also to combinations of these, and even within any one modality (e.g., with neurons tuned to different tastes). Nevertheless, what account might we give for why so many different types of reward are represented in the human medial orbitofrontal cortex? I suggested in *Emotion Explained* that part of the functional utility of this is that there can be a comparison of the magnitudes of what may be quite different types of reward, implemented by the local lateral inhibition mediated via inhibitory interneurons. The result of the mutual inhibition is that the relative magnitude of the different rewards available can be compared and that the most strongly firing neurons after the competition reflect the strongest reward.

Neuropsychology of the Human Orbitofrontal Cortex

Having considered evidence from neuroimaging on the functions of the human orbitofrontal cortex in emotion and motivation, we now consider complementary evidence from the effects of damage to the human orbitofrontal cortex.

Reward Reversal and Its Relation to Other Changes in Behavior

It is of interest that a number of the symptoms of frontal lobe damage in humans appear to be related to the functions of representing primary (unlearned) reinforcers and of altering behavior when stimulus–reinforcement associations alter. Thus, humans with frontal lobe damage can show impairments in a number of tasks in which an alteration of behavioral strategy is required in response to a change in environmental reinforcement contingencies. It is of interest that frontal patients may be able to verbalize the correct rules yet be unable to correct their behavioral sets or strategies appropriately.

Some of the personality changes that can follow frontal lobe damage may be related to a similar type of dysfunction in processing rewards and punishers. For example, the euphoria, irresponsibility, lack of affect, and lack of concern for the present or future that can follow frontal lobe damage may be related to a dysfunction in altering behavior appropriately in response to a change in reinforcement contingencies. Indeed, in so far as the orbitofrontal cortex is involved in the disconnection of stimulus–reinforcer associations, and such associations are important in learned emotional responses, it follows that the orbitofrontal cortex is involved in emotional responses by correcting stimulus–reinforcer associations when they become inappropriate.

The hypotheses about the role of the orbitofrontal cortex in the rapid alteration of stimulus–reinforcer associations and the functions more generally of the orbitofrontal cortex in human behavior have been investigated in recent studies in humans with damage to the ventral parts of the frontal lobe. (The description ventral is given to indicate that there was pathology in the orbitofrontal or related parts of the frontal lobe, and not in the more dorsolateral parts of the frontal lobe.) A task that was directed at assessing the rapid alteration of stimulus–reinforcer associations was used because the findings indicated that the orbitofrontal cortex is involved in this type of learning. The task used was visual discrimination reversal, in which patients could learn to obtain points by touching one stimulus when it appeared on a video monitor but had to withhold a response when a different visual stimulus appeared, otherwise a point was lost. After the subjects had acquired the visual discrimination, the reinforcement contingencies unexpectedly reversed. The patients with ventral frontal lesions made more errors in the reversal (or in a similar extinction) task, and completed fewer reversals, than control patients with damage elsewhere in the frontal lobes or in other brain regions (see **Figure 1**).

An important aspect of the findings was that the reversal learning impairment correlated highly with the socially inappropriate or disinhibited behavior of the patients (see **Figure 1**), and also with their subjective evaluation of the changes in their emotional state since the brain damage. The patients were not impaired at other types of memory task, such as paired associate learning. It is of interest that the patients can often verbalize the correct response yet commit the incorrect action. This is consistent with the hypothesis that the orbitofrontal cortex is normally involved in executing behavior when the behavior is performed by evaluating the

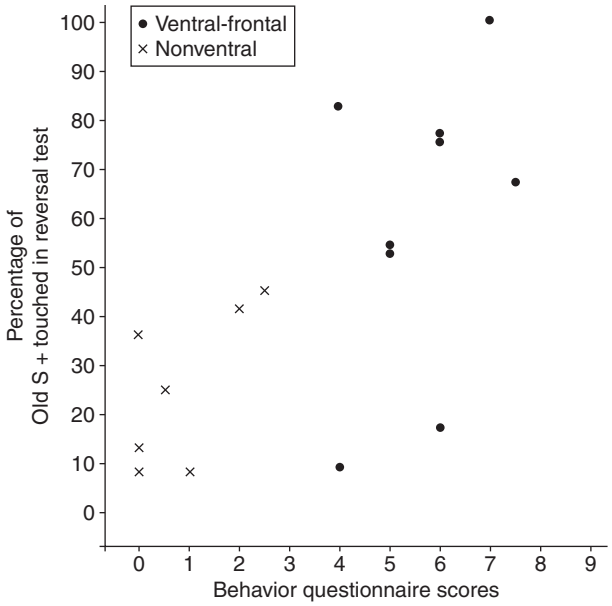

Figure 1 Visual discrimination reversal performance in humans with damage to the ventral part of the frontal lobe. The task was to touch the screen when one image, the S+, was shown in order to obtain a point and to refrain from touching the screen when a different visual stimulus, the S−, was shown in order to obtain a point. The scattergraph shows that during the reversal the group with ventral damage were more likely to touch the previously rewarded stimulus (Old S+) and that this was related to the score on a behavior questionnaire. Each point represents one patient in the ventral-frontal group or in a control group. The behavior questionnaire rating reflected high ratings on at least some of the following: disinhibited or socially inappropriate behavior, misinterpretation of other people's moods. impulsiveness. unconcern about or underestimation of the seriousness of the individual's condition. and lack of initiative. Reproduced from Rolls ET, Hornak J, Wade D, and McGrath J (1994) Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery and Psychiatry* 57: 1518–1524.

reinforcement associations of environmental stimuli. The orbitofrontal cortex seems to be involved in this in both humans and nonhuman primates when the learning must be performed rapidly, for example, in acquisition and during reversal.

To seek positive confirmation that effects on stimulus–reinforcer association learning and reversal were related to orbitofrontal cortex damage rather than to any other associated pathology, a new reversal-learning task was used with a group of patients with discrete, surgically produced lesions of the orbitofrontal cortex. In the new visual discrimination task (the same as that used in the O' Doherty et al. in 2001 monetary-reward functional neuroimaging task, two stimuli are always present on the video monitor and the patient obtains monetary reward by touching the correct stimulus and loses money by touching the incorrect stimulus. This design controls for an effect of the lesion in simply increasing the probability that any response will be made. The new task also uses probabilistic amounts of reward and punishment on each trial to make it harder to use a verbal strategy with an explicit rule. The task also had the advantage that it was the same as that used in our human functional neuroimaging study that showed the activation of the orbitofrontal cortex by monetary gain or loss. It was found that a group of patients with bilateral orbitofrontal cortex lesions were severely impaired at the reversal task, in that they accumulated less money. The investigation showed that the impairment was only obtained with bilateral orbitofrontal cortex damage, in that patients with unilateral orbitofrontal cortex (or medial prefrontal cortex) lesions were not impaired in the reversal task. These patients often failed to switch their choice of stimulus after a large loss; and often did switch their choice even though they had just received a reward; this was quantified in a more recent study.

It is of interest that the patients with bilateral orbitofrontal cortex damage who were impaired at the visual discrimination reversal task had high scores on parts of a social behavior questionnaire in which the patients were rated on behaviors such as emotion recognition in others (e.g., their sad, angry, or disgusted mood), in interpersonal relationships (e.g., not caring what others think and not being close to the family), emotional empathy (e.g., when others are happy, they are not happy for them), interpersonal relationships (e.g., does not care what others think and are not close to the family), public behavior (is uncooperative), antisocial behavior (is critical of and impatient with others), impulsivity (does things without thinking), and sociability (is not sociable, and has difficulty making or maintaining close relationships) – all of which could reflect less behavioral sensitivity to different types of punishment and reward. Further, in a subjective emotional change questionnaire, in which the patients reported on any changes in the intensity and/or frequency of their own experience of emotions, patients with bilateral orbitofrontal cortex lesions with deficits in the visual discrimination reversal task reported a number of changes, including changes in sadness, anger, fear, and happiness.

These results on the effects of brain damage to the orbitofrontal cortex and the complementary neuroimaging results already described provide evidence that at least part of the function of the orbitofrontal cortex in emotion, social behavior, and decision making is related to representing reinforcers, detecting changes in the reinforcers being received, using these changes to rapidly reset stimulus–reinforcer associations, and rapidly changing behavior as a result.

An idea of how such stimulus–reinforcer learning may play an important role in normal human behavior, and may be related to the behavioral changes seen clinically in patients with ventral-frontal lobe damage, can be provided by summarizing the behavioral ratings given by the carers of these patients. The patients were rated high on at least some of the following: disinhibition or socially inappropriate behavior, violence, verbal abusiveness, lack of initiative, misinterpretation of other people's behavior, anger or irritability, and lack of concern for their own condition. Such behavioral changes correlated with the stimulus–reinforcer reversal and extinction learning impairment. The suggestion, thus, is that the insensitivity to reinforcement changes in the learning task may be at least part of what produces the changes in behavior found in these patients with ventral-frontal lobe damage.

The more general impact on the behavior of these patients is that their irresponsibility tended to affect their everyday lives. For example, if such patients had received their brain damage in a road traffic accident and compensation had been awarded, the patients often tended to spend their money without appropriate concern for the future, sometimes, for example, buying a very expensive car. Such patients often find it difficult to invest in relationships, too, and are sometimes described by their family as having changed personalities, in that they care less about a wide range of factors than before the brain damage. The suggestion that follows from this is that the orbitofrontal cortex may normally be involved in much social behavior and the ability to respond rapidly and appropriately to social reinforcers which is, of course, an important aspect of primate social behavior.

Bechara and colleagues also have findings in patients with frontal lobe damage when they performed a gambling task that are consistent with those already described. In the Iowa gambling task, subjects were asked to select cards from four decks of cards and maximize their winnings. During the task, the electrodermal activity (skin conductance responses (SCR)) of the subject was measured as an index of somatic activation. After each selection of a card, facsimile money was lost or won. Two of the four decks produced large payouts with larger penalties (and can thus be considered high risk), whereas the other two decks produced small payouts but smaller penalties (low risk). The most profitable strategy was, therefore, to consistently select cards from the two low-risk decks, which is the strategy adopted by normal control subjects. Patients with damage to the ventromedial part of the orbitofrontal cortex, but not the dorsolateral prefrontal cortex, persistently drew cards from the high-risk packs and lacked anticipatory SCRs while they pondered risky choices. The task was designed to mimic aspects of real-life decision making that patients with orbitofrontal cortex lesions find difficult. Such decisions typically involve choices between actions associated with differing magnitudes of reward and punishment in which the underlying contingencies relating actions to relevant outcomes remain hidden.

Most known cases of human orbitofrontal damage occurred in adulthood, but two cases of damage acquired in early life have been reported. The two patients showed life-long behavioral problems, which were resistant to corrective influences. But more importantly, the patients appeared completely to lack knowledge about moral and societal conventions. Interestingly, other patients with late acquired orbitofrontal lesions retained knowledge of such matters, even if they did not always act in accordance with this explicit knowledge. The lack of this moral knowledge and subsequent reckless behavior in the two patients with early life damage to the orbitofrontal cortex is consistent with the hypothesis that the orbitofrontal cortex is crucial for stimulus–reinforcer association learning. The implication seems to be that the orbitofrontal cortex is necessary for the development of personal moral-based knowledge based on the processing of rewards and punishers.

Face and Voice Expression

To investigate the possible significance of face-related inputs to the orbitofrontal visual neurons, we also tested the responses of these patients to faces. We included tests of face (and also voice) expression decoding because these are ways in which the reinforcing quality of individuals is often indicated. Impairments in the identification of facial and vocal emotional expression were demonstrated in a group of patients with ventral-frontal lobe damage who had socially inappropriate behavior. The expression identification impairments occurred independently of perceptual impairments in facial recognition, voice discrimination, or environmental sound recognition. The face and voice expression problems did not necessarily occur together in the same patients, providing an indication of separate processing. Poor performance on both expression tests was correlated with the degree of alteration of emotional experience reported by the patients. There was also a strong positive correlation between the degree of altered emotional experience and the severity of the behavioral problems (e.g., disinhibition) found in these patients. A comparison group of patients with brain damage outside the ventral-frontal lobe region, and without these behavioral problems, was unimpaired on the face expression identification test, was significantly less impaired at vocal expression

identification, and reported little subjective emotional change.

These findings were extended, and it was found that patients with face expression-decoding problems do not necessarily have impairments of visual discrimination reversal, and vice versa. This is consistent with some topography in the orbitofrontal cortex.

To obtain clear evidence that the changes in face and voice expression identification, emotional behavior, and subjective emotional state were related to orbitofrontal cortex damage itself, and not to damage to surrounding areas which is present in many closed-head injury patients, we performed these assessments in patients with circumscribed lesions made surgically in the course of treatment. This study also enabled us to determine whether there was functional specialization within the orbitofrontal cortex and whether damage to nearby and connected areas (such as the ACC), in which some of the patients had lesions, could produce similar effects. We found that some patients with bilateral lesions of the orbitofrontal cortex had deficits in voice and face expression identification and that the group had impairments in social behavior and significant changes in their subjective emotional state. The same group of patients had deficits on the probabilistic monetary reward task. Some patients with unilateral damage restricted to the orbitofrontal cortex also had deficits in voice expression identification, and the group did not have significant changes in social behavior or in their subjective emotional state. Patients with unilateral lesions of the anteroventral part of the ACC and/or medial prefrontal cortex area BA9 were in some cases impaired on voice and face expression identification, had some change in social behavior, and had significant changes in their subjective emotional state. Patients with dorsolateral prefrontal cortex lesions or with medial lesions outside the ACC and medial prefrontal BA9 areas were unimpaired on any of these measures of emotion. In all cases in which voice expression identification was impaired, there were no deficits in the control tests of the discrimination of unfamiliar voices and the recognition of environmental sounds.

Relation to Psychiatric Symptoms

It is also becoming possible to relate the functions of the orbitofrontal cortex to some psychiatric symptoms. Berlin et al. compared the symptoms of patients with a personality disorder syndrome, borderline personality disorder (BPD), with those of patients with lesions of the orbitofrontal cortex. The symptoms of the self-harming BPD patients include high impulsivity, high affective instability, high emotionality, and low extraversion. It was found that orbitofrontal cortex and BPD patients performed similarly in that they were more impulsive; reported more inappropriate behaviors in the frontal behavior questionnaire; and had more BPD characteristics, more anger, and less happiness than control groups (either normals or patients with lesions outside the orbitofrontal cortex).

Both the orbitofrontal and BPD groups also had a faster perception of time (i.e., they underproduced time) than normal controls. This may be one factor underlying their increased impulsiveness; they feel that sufficient time has elapsed to initiate action. It was of considerable interest that the BPD group, as well as the orbitofrontal group, scored highly on the frontal behavior questionnaire, which assessed inappropriate behaviors typical of orbitofrontal cortex patients including disinhibition, social inappropriateness, perseveration, and uncooperativeness. Both groups were also less open to experience (i.e., less open-minded), a personality characteristic.

The orbitofrontal and BPD patients performed differently on other tasks. BPD patients were less extraverted and conscientious, and were more neurotic and emotional, than all other groups. Patients with orbitofrontal cortex lesions had more severe deficits in reversing stimulus–reinforcer associations than all other groups and had a faster perception of time (overestimated time) than normal controls. These deficits were not related to spatial working-memory functions, which are impaired by dorsolateral prefrontal cortex damage.

Thus, some but not other symptoms of self-harming BPD patients are similar to those of patients with orbitofrontal cortex damage. The symptoms the groups have in common include impulsiveness and the inappropriate behaviors typical of frontal patients. This could imply that in BPD patients some aspects of the operation of the orbitofrontal cortex are occurring (whatever their etiology, which could include innate or acquired changes and might involve different expression or operation of neurotransmitters) differently from the way they operate in normal control subjects. Part of our interest in this is that it may help to point toward new concepts that may be useful in the treatment of some of the symptoms of patients with BPD. On the other hand, other aspects of BPD do not appear to be related to orbitofrontal cortex functions, including the more neurotic and more emotional personality characteristics of the BPD group, together with their lower extraversion and conscientious.

Frontotemporal Dementia

Another case in which it is possible to relate psychiatric types of symptom to orbitofrontal cortex function

is frontotemporal dementia, which is a progressive neurodegenerative disorder attacking the frontal lobes and producing major and pervasive behavioral changes in personality and social conduct resembling those produced by orbitofrontal lesions. Patients appear either socially disinhibited, with facetiousness and inappropriate jocularity, or apathetic and withdrawn. Many patients show mental rigidity and an inability to appreciate irony or other subtle aspects of language. They tend to engage in ritualistic and stereotypical behavior, and their planning skills are invariably impaired. The dementia is accompanied by gradual withdrawal from all social interactions. Memory is usually intact but patients have difficulties with working memory and concentration. Interestingly, given the anatomy and physiology of the orbitofrontal cortex, frontotemporal dementia causes profound changes in eating habits, with escalating desire for sweet food coupled with reduced satiety, which is often followed by enormous weight gain.

The ACC and Affect

The perigenual cingulate area occupies approximately the anterior one-third of the cingulate cortex (see Figure 2) and is involved in emotion; it may be distinguished from a midcingulate area (i.e., further back than the perigenual cingulate region and occupying approximately the middle third of the cingulate cortex), which has been termed the cingulate motor area and may be involved in response selection. Both the perigenual and midcingulate areas are part of what anatomically is frequently termed the ACC; but, for convenience, when using the term anterior cingulate cortex I refer here to the perigenual cingulate cortex and distinguish this from the midcingulate cortex described elsewhere.

Vogt and colleagues showed that pain produced an increase in regional cerebral blood flow (rCBF, measured with positron emission tomography (PET)) in an area of perigenual cingulate cortex which included parts of areas 25, 32, 24a, 24b, and/or 24c. Vogt et al. suggested that the activation of the anterior part of the cingulate area is related to the affective aspect of pain. There are direct projections to the cingulate cortex from medial thalamic areas that relay pain inputs, including the parafascicular nucleus. In terms of other connections, the ACC is connected to the medial orbitofrontal areas, parts of lateral orbitofrontal area 12, the amygdala (which projects strongly to cingulate subgenual area 25), and the temporal pole cortex; it also receives somatosensory inputs from the insula and other somatosensory cortical areas (see Figure 2). The ACC has output projections to the periaqueductal gray in the midbrain (which is implicated in pain processing), to the nucleus of the solitary tract and dorsal motor nucleus of the vagus (through which autonomic effects can be elicited), and to the ventral striatum and caudate nucleus (through which behavioral responses could be produced).

Consistent with the anterior cingulate region being involved in affect, it is close to (just above) the area activated by the induction of a sad mood in studies made by Mayberg and colleagues. Also consistent with this region being involved in affect, Lane and colleagues found increased regional blood flow in a 1997 PET study in a far anterior part of the cingulate cortex, where it adjoins the prefrontal cortex, when humans paid attention to the affective aspects of pictures they were being shown which contained pleasant images (e.g., flowers) and unpleasant pictures (e.g., a mangled face and a snake). The perigenual ACC (approximately the anterior one-third of the cingulate cortex and including parts of areas 32, 24, and 25) is thus related to affect.

The perigenual cingulate cortex may be part of an executive output response-selection system for some emotional states, in which the responses can include autonomic responses. Part of the basis for this suggestion is its inputs from somatosensory cortical areas, including the insula, the orbitofrontal cortex, and the amygdala (see Figure 2), and its outputs to brain stem areas such as the periaqueductal (or central) gray in the midbrain, the ventral striatum, and the autonomic brain stem nuclei. Anterior cingulate lesions in humans can produce apathy, autonomic dysregulation, and emotional instability. We have shown that patients with circumscribed, even unilateral, surgical lesions of the ACC/perigenual cingulate cortex may be impaired for voice and face expression identification; have impaired subjective emotional states; and have some changes in social behavior, including being less likely to notice when other people were angry, not being close to his or her family, and doing things without thinking.

Functional neuroimaging studies showed that there appear to be separate representations of aversive and positively affective stimuli in the ACC/perigenual cingulate cortex. The area activated by pain is typically 10–30 mm behind and above the most anterior (i.e., pre- or perigenual) part of the ACC. Pleasant touch was found to activate the most anterior part of the ACC, just in front of the (genu, or knee, of the) corpus callosum (i.e., the pregenual cingulate cortex). Oral somatosensory stimuli such as viscosity and fat texture also activate this most anterior part of the ACC. More than just somatosensory stimuli are represented, however, in that (pleasant) sweet taste also activates the most anterior part of the

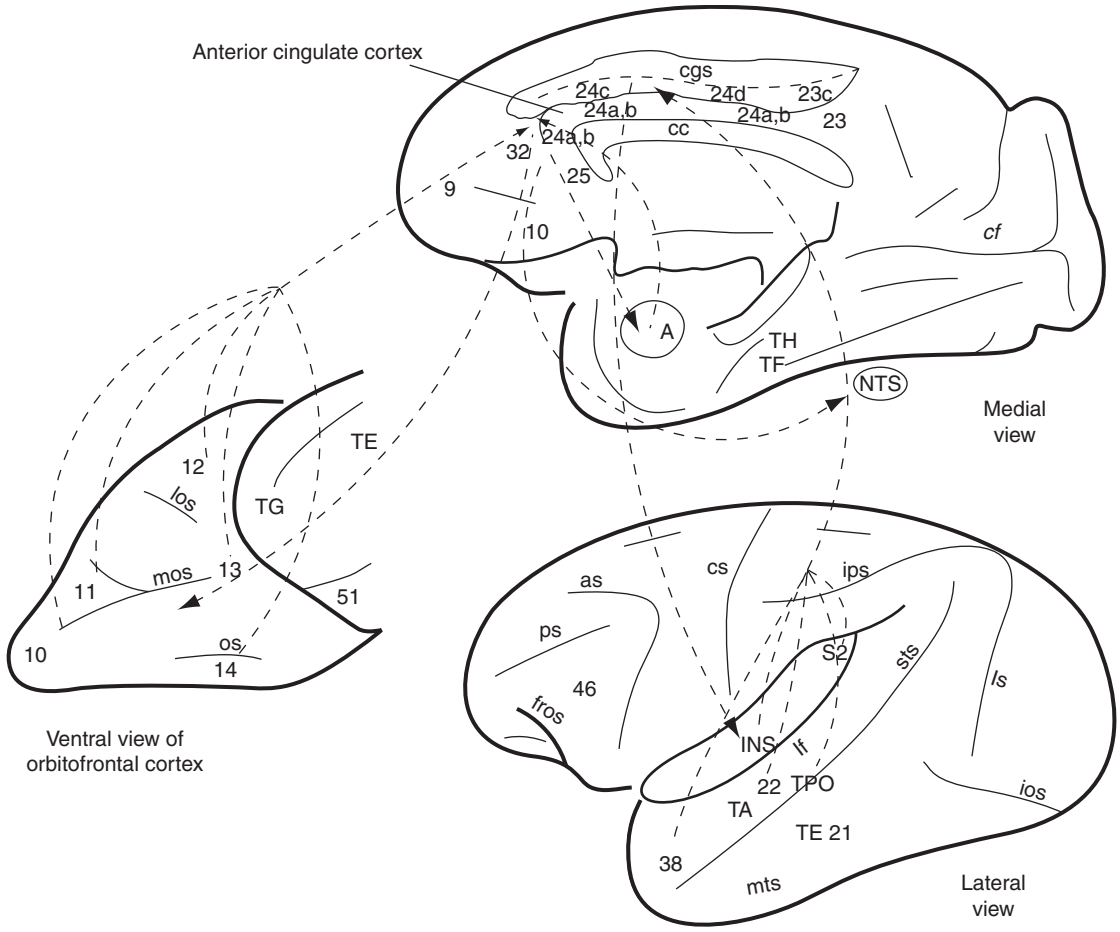

Figure 2 Connections of the perigenual and midcingulate cortical areas (shown on views of the primate brain). The cingulate sulcus (cgs) has been opened to reveal the cortex in the sulcus, with the dashed line indicating the depths (fundus) of the sulcus. The cingulate cortex is in the lower bank of this sulcus and in the cingulate gyrus, which hooks above the corpus callosum and around the corpus callosum at the front and the back. The perigenual cingulate cortex extends from cingulate areas 32, 24a, and 24b to subgenual cingulate area 25. (The cortex is called subgenual because it is below the genu (knee) formed by the anterior end of the corpus callosum (cc).) The perigenual cingulate cortex tends to have connections with the amygdala and orbitofrontal cortex, whereas area 24c tends to have connections with the somatosensory insula (INS), the auditory association cortex (22, TA), and the temporal pole cortex (38). The midcingulate areas include area 24d, which is part of the cingulate motor area. 12, 13, 11, orbitofrontal cortex; 38, TG, temporal pole cortex; 51, olfactory (prepyriform and periamygdaloid) cortex; A, amygdala; as, arcuate sulcus; cc, corpus callosum; cf, calcarine fissure; cgs, cingulate sulcus; cs, central sulcus; fros, fronto-orbital sulcus; ios, inferior occpital sulcas; ips, intraparietal sulcus; mts, middle temporal sulcus; INS, insula; ios, inferior occipital sulcus; lf, lateral (or Sylvian) fissure (which has been opened to reveal the insula); ls, lunate sulcus; mos, medial orbital sulcus; NTS, autonomic areas in the medulla, including the nucleus of the solitary tract and the dorsal motor nucleus of the vagus; os, orbital sulcus; ps, principal sulcus; sts, superior temporal sulcus; S2, somatosensory cortical area S2; TG, temporal pole cytoarchitectomic area; TA (22), superior temporal auditory association cortex; TE (21), inferior temporal visual cortex; TF and TH, parahippocampal cortex; TPO, multimodal cortical area in the superior temporal sulcus.

ACC, as do pleasant odors and cognitive inputs that influence the pleasantness of odors. Unpleasant odors activate further back in the ACC. Activations in the anterior/perigenual cingulate cortex are also produced by the taste of water when it is rewarding because of thirst, by the flavor of food, and by monetary reward. The locations of some of these activations are shown in Figure 3, from which it is clear that many positively affective stimuli are represented in the most anterior part of the perigenual cingulate cortex, with some less positively affective or negatively affective stimuli activating a region of the cingulate cortex which is just posterior to this, above the corpus callosum.

In addition to the pregenual cingulate sites which are activated in many of our studies of affective stimuli, it was also frequently found that activation in a region at the intersection of the medial prefrontal cortex, the subgenual cingulate cortex, and the orbitofrontal cortex is activated by positively affective

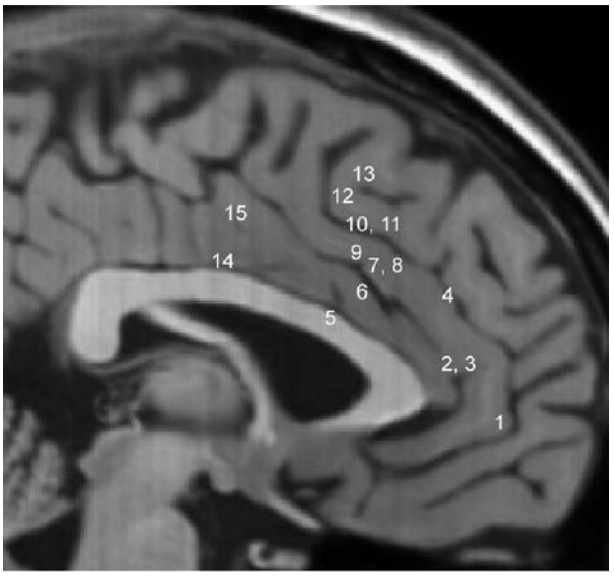

Figure 3 Activations of the most anterior part of the perigenual cingulate cortex by different positively affective stimuli, with less positively affective or negatively affective stimuli producing activations centered just behind this area. Touch: 3 pleasant, 14 unpleasant/pain, 5 unpleasant/pain; Odor: 7 pleasant, 2 correlation with pleasantness ratings, 9 unpleasant; monetary loss: 8; Water: 6 correlation with pleasantness, 15 correlation with pleasantness; Taste: 10 glucose, 11 salt, 4 salt; sucrose and fat: 1; nonreward in a social reversal task: 13, 12. Reproduced from Rolls ET (2005) *Emotion Explained*, figure 4.61, p. 183. Oxford: Oxford University Press by permission of Oxford University Press.

stimuli. The facts that this anterior cingulate region is activated by many rewarding stimuli (as just summarized), affects emotion if lesioned, and is connected with the orbitofrontal cortex indicate that it is important in emotion, perhaps as a region linking the orbitofrontal cortex to outputs. It is in this context very interesting that chronic electrical stimulation in or close to this region (in the subgenual cingulate cortex) relieved symptoms of treatment-resistant depression in some patients. This provides a close link between evidence from neuroscience about the functions of this medial region and the therapeutic potential of these concepts for helping to understand and treat depression.

Overall, it is thus notable that what can be termed an affective part of the ACC, the perigenual cingulate cortex, appears to have a most anterior part where activations are produced by a number of different pleasant stimuli. But further back toward the junction with the midcingulate area and perhaps sometimes including parts of the midcingulate area several affectively negative stimuli produce activations.

An investigation in patients with selective surgical lesions showed that patients with unilateral lesions of the anteroventral part of the ACC and/or medial BA9 were in some cases impaired for voice and face expression identification; had some change in social behavior, such as inappropriateness; and had significant changes in their subjective emotional state. Unilateral lesions were sufficient to produce these effects, and there were no strong laterality effects. Consistent with the effects of the anterior cingulate lesions in humans on recognizing vocal (and, in some cases, facial) emotional expression, neuroimaging studies concerned with vocal expression identification have reported orbital and medial activation. For example nonverbal sounds expressing fear, sadness, and happiness, when compared to a neutral condition, activated BA11 (orbital cortex) bilaterally and the medial BA9 on the left. Fear-related increases in activity were also found on the right only in BA11. In another study, it was found that fearful sounds activated the medial BA32 and BA24 (the anterior cingulate at/below the level of the genu of the corpus callosum), again on the right side only.

There is also neuroimaging evidence that complements the effects of lesions in suggesting a role for certain medial regions in the subjective experience of emotion. In neuroimaging studies by Lane and colleagues with normal human subjects, bilateral activations in the medial BA9 were found as subjects viewed emotion-laden stimuli, and in both the medial BA9 and ventral ACC during self-generated emotional experience (i.e., in the absence of a stimulus) as subjects recalled emotions of sadness or happiness. On the basis of a review of imaging studies which consistently emphasize the importance of anterior and ventral regions of the ACC for emotion, Bush et al. argued that the ACC can be divided into a ventral affective division (which includes the subcallosal region and the part anterior to the corpus callosum) and a dorsal cognitive division, a view strengthened by the demonstration of reciprocally inhibitory interactions between these two regions.

The subgenual part (area 25) of the anterior cingulate cortex is, via its outputs to the hypothalamus and brain stem autonomic regions, involved in the autonomic component of emotion. The ACC is also activated in relation to autonomic events, and Critchley and colleagues showed that there is a correlation with skin conductance, a measure of autonomic activity related to sympathetic activation, in the ACC and related areas.

A current working hypothesis is that the affective part of the ACC receives inputs about expected rewards and punishers, and about the rewards and punishers received, from the orbitofrontal cortex and amygdala. There is some segregation of the areas that receive these inputs. The ACC may compare these signals, and use them in functions such as affective decision making and in producing autonomic

responses. As such, the perigenual cingulate cortex may act as an output system for emotional responses and actions.

See also: Connectivity of Primate Reward Centers; Cortical Processing of the Reward Value of Food; Flavor Physiology; Orbitofrontal Cortex: Visual Functions; Reward and Learning; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Bechara A, Tranel D, Damasio H, and Damasio AR (1996) Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cerebral Cortex* 6: 215–225.

Berlin H, Rolls ET, and Kischka U (2004) Impulsivity, time perception, emotion, and reinforcement sensitivity in patients with orbitofrontal cortex lesions. *Brain* 127: 1108–1126.

Bush G, Luu P, and Posner MI (2000) Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Science* 4: 215–222.

Hornak J, Bramham J, Rolls ET, et al. (2003) Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain* 126: 1691–1712.

Hornak J, O'Doherty J, Bramham J, et al. (2004) Reward-related reversal learning after surgical excisions in orbitofrontal and dorsolateral prefrontal cortex in humans. *Journal of Cognitive Neuroscience* 16: 463–478.

Hornak J, Rolls ET, and Wade D (1996) Face and voice expression identification in patients with emotional and behavioral changes following ventral frontal lobe damage. *Neuropsychiatry* 34: 247–261.

Kringelbach ML and Rolls ET (2003) Neural correlates of rapid reversal learning in a simple model of human social interaction. *NeuroImage* 20: 1371–1383.

Kringelbach ML and Rolls ET (2004) The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology* 72: 341–372.

Mayberg HS (2003) Positron emission tomography imaging in depression: A neural systems perspective. *Neuroimaging Clinics of North America* 13: 805–815.

Nagai Y, Critchley HD, Featherstone E, Trimble MR, and Dolan RJ (2004) Activity in ventromedial prefrontal cortex covaries with sympathetic skin conductance level: A physiological account of a "default mode" of brain function. *NeuroImage* 22: 243–251.

O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, and Andrews C (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience* 4: 95–102.

Phillips ML, Drevets WC, Rauch SL, and Lane R (2003) Neurobiology of emotion perception. I: The neural basis of normal emotion perception. *Biological Psychiatry* 54: 504–514.

Rolls ET (2004) The functions of the orbitofrontal cortex. *Brain and Cognition* 55: 11–29.

Rolls ET (2005) *Emotion Explained*. Oxford: Oxford University Press.

Rolls ET, Hornak J, Wade D, and McGrath J (1994) Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery and Psychiatry* 57: 1518–1524.

Vogt BA and Sikes RW (2000) The medial pain system, cingulate cortex, and parallel processing of nociceptive information. *Progress in Brain Research* 122: 223–235.

Prefrontal Cortex

J M Fuster, University of California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

The cortex of the anterior pole of the brain can be subdivided, histologically and physiologically, into three major parts. They have been most clearly identified and extensively investigated in the frontal lobe of the primate: (1) the primary motor cortex, a strip of agranular cortex rostral and adjacent to the central sulcus; (2) the premotor cortex, a conglomerate of cytoarchitectonically transitional areas, including the supplementary motor area (SMA); and (3) the prefrontal cortex, which constitutes the largest and most anterior of the three components and which in the primate has a prominent internal granular layer (IV) and for this reason is designated as the granular frontal cortex. The three components can be well distinguished in the external convexity of the frontal lobe; in the cingulate and inferior (orbital) aspects of the lobe, the first two components are absent and the granular cytoarchitecture of the prefrontal cortex blends more or less gradually into that of limbic cortex behind it.

All three parts of the frontal cortex are functionally involved in behavior. The premotor and motor areas participate in the preparation and execution of specific movements, whereas the prefrontal cortex participates in cognitive aspects of behavior. Although its functions have not been completely elucidated, current knowledge strongly indicates that the prefrontal cortex is essentially engaged in the organization of goal-directed behavioral sequences. Furthermore, the prefrontal cortex seems essential for goal-directed rational thinking. Thus, the functional integrity of this cortex is necessary for the temporal organization of behavioral actions as well as of the 'internal actions' that constitute logical reasoning.

Development

Phylogenetically, the prefrontal cortex is the most rostral portion of the neocortex. It is one of the cortical regions to undergo the latest and most expansive development in the course of evolution (Figure 1); this is especially true for the cortex of the external prefrontal convexity, which is the part of frontal cortex most heavily involved in cognitive functions. Such development reaches a maximum in the human brain, where the prefrontal cortex constitutes nearly one-third of the neocortex. Ontogenetically, the prefrontal cortex also develops relatively late. In the human, its myelinization is not complete until the second decade of life.

Connections

The prefrontal cortex is connected with many other cerebral structures. Its extrinsic connectivity has certain peculiarities that distinguish it from other neocortical regions. They are as follows:

1. Reciprocal connections with anterior and dorsal nuclei of the thalamus; the prominent links of the prefrontal cortex with the nucleus medialis dorsalis are conventionally used as the criterion for the definition of this cortex in all species.
2. Reciprocal connections with several sensory processing areas of posterior association cortex; afferent inputs from various sensory systems (somatic, visual, auditory, and olfactory) arrive in the prefrontal cortex, which therefore may be considered cortex of polymodal sensory convergence and association.
3. Reciprocal connections with limbic formations, especially the hypothalamus, the amygdala, and the hippocampus, which are known to be involved in certain aspects of emotion, motivation, and memory.
4. Efferent connections to subcortical structures involved in motor control, such as the basal ganglia; further, through the basal ganglia, the cerebellum, and the thalamus, the prefrontal cortex has indirect connective access to the neighboring motor areas of the frontal lobe.

In addition, the prefrontal cortex, as in other associative regions of the neocortex, is reciprocally connected with diencephalic and mesencephalic nuclei and reticular formations with diffuse cortical projection. As in other regions, some afferent (corticocortical) connections in the prefrontal cortex have been noted to terminate in columnar transcortical patterns of distribution.

Neurotransmitters

All major types of neurotransmitter terminals and receptors are represented in the prefrontal cortex. The principal monoaminergic afferent systems from the brain stem and limbic structures innervate the prefrontal cortex. Conversely, the prefrontal cortex sends descending monoaminergic efferents to the thalamus, hypothalamus, subthalamus, and the core of the lower brain stem. Norepinephrine and dopamine seem to be synthesized and concentrated in

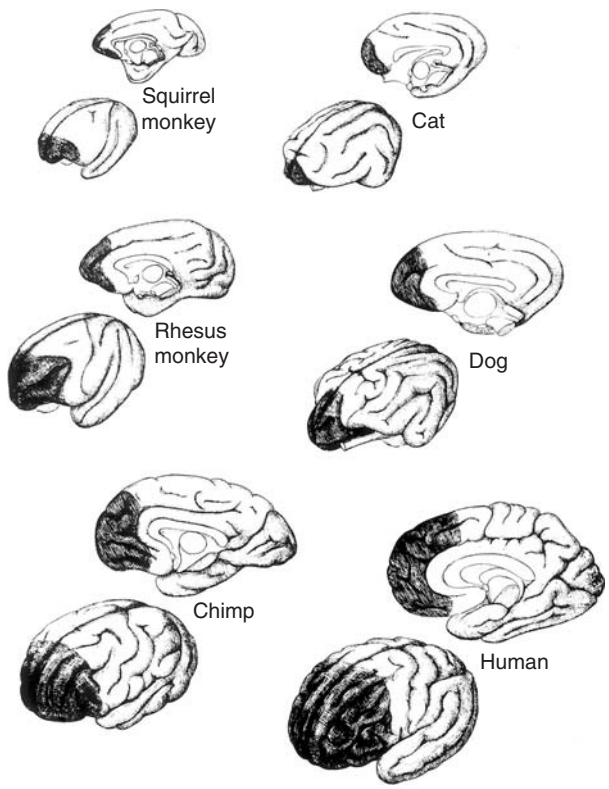

Figure 1 Lateral and medial views of the brains of several species. The prefrontal cortex is indicated by dark shading.

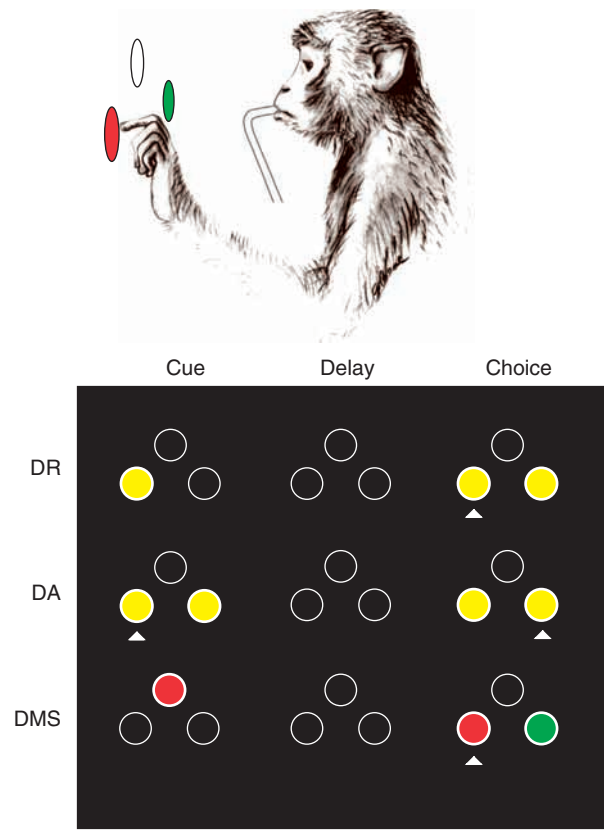

Figure 2 Monkey facing a panel with stimulus–response buttons for performance of delay tasks (mouth spigot for delivery of liquid reward). Below, the sequence of events in three such tasks: delayed response (DR), delayed alternation (DA), and delayed matching-to-sample (DMS). A DR trial begins with brief illumination of a button (right or left, randomly changing); after a delay of a few seconds, two buttons are simultaneously lit, and the animal must press the one lit at the start of the trial. In DA, the animal must alternately press the right and the left of a pair of buttons that are simultaneously lighted between delays. In DMS, the trial begins with brief presentation of a sample color, red or green (upper button); after the delay, both colors appear in the lower buttons, and the animal must press the one matching the sample color; both that color and the positions of the two colors in the lower buttons are changed randomly between trials. (White triangles mark the site of correct response).

higher quantities in the prefrontal cortex than in posterior sensory and associative cortex. The reverse is true for serotonin. The prefrontal region is the recipient of abundant cholinergic fibers from the basal nucleus of Meynert (or its equivalent in the rat and the cat). This nucleus is a relay in a cholinergic system of ascending connections that originates in the brain stem reticular core and diffusely projects to the cerebral cortex. Both catecholamine and acetylcholine systems have been postulated to mediate, in a yet unspecified manner, the participation of the prefrontal cortex in cognitive processes, memory in particular.

Neuropsychology

Lesions of the prefrontal cortex cause changes in various aspects and forms of behavior. The magnitude of the changes depends on the location and extent of the lesion. This has been most systematically studied in primates, but the general characteristics of the behavioral changes observed are the same in all species in which they have been investigated.

The best known alterations resulting from lesions of the dorsolateral prefrontal convexity are deficits in learning and performance of various structured behavioral tasks that have one common feature: the necessity to execute motor acts in accord with sensory information no longer present in the environment. The temporal discontinuity between information and action (cross-temporal contingency) is an essential element of these tasks (Figure 2). Thus, the deficit is neither sensory nor motor but has to do with the ability of the animal to integrate, across a span of time, sensory input with appropriate action. However, the degree of deficit depends not only on the length of that time but also on the novelty and complexity of the sensory stimulus or stimuli before it.

It also depends on whether the stimulus–response coupling required for the task is merely one of two or more similar and alternative couplings in the behavioral repertoire of the animal. If only one specific signal calls for one specific act, without the competing possibility of similar signals and acts, no deficit can be seen after prefrontal lesion, however long the time between the signal and the action.

Most typical and documented are the deleterious effects of prefrontal lesions, especially if they involve the lateral convexity, on the performance of delay tasks (e.g., delayed response, delayed alternation, delayed matching to sample) and discrimination reversal tasks. All such tasks impose the need to bridge a cross-temporal contingency between information and action, as well as the need to reject similar alternatives.

Lesions of inferior (orbital) or medial prefrontal cortex may also disrupt to some extent the performance of behavioral tasks, especially if they require the inhibition or suppression of a prepotent and well-established tendency or response alternative. Inferior lesions severely disrupt such tasks as conditioned inhibition and successive discrimination, where that requirement is critical for successful performance but the bridging of cross-temporal contingencies is not.

Summing up, cortical lesion experiments show that the prefrontal cortex, unlike other parts of the cerebral cortex, is essential for the formation of goal-directed structures of behavior that demand cross-temporal integration and the rejection of inappropriate alternatives. The behavioral effects of selective cortical ablations and reversible lesions (by cooling) in the monkey suggest that this is accomplished by the joint action of a temporally retrospective function of short-term memory – also called 'working memory' – and a function of control of internal interference. The former seems mainly based in dorsolateral, the latter in inferior (orbital) prefrontal cortex.

Although clinical observations of the effects of frontal lobe pathology were made long before the relevant animal lesion studies had been carried out, the data from prefrontal lesions in the human have been less systematically investigated and are more difficult to evaluate than those from nonhuman primates. Nevertheless, in recent years, rigorous neuropsychological studies of patients with frontal lobe lesions have shed new light on the nature of symptoms that are commonly encountered in such patients (e.g., difficulties in attention, recent memory, planning). These studies also reveal a remarkable degree of similarity between the prefrontal syndromes of humans and animals. It has now been shown that humans with lesions of the prefrontal convexity, like monkeys with such lesions, exhibit deficits in performance of short-term memory tasks (e.g., delayed matching) and tasks that require the ordering of items in time. In the human, as in the monkey, lesions of orbital prefrontal cortex disrupt the ability to perform tasks demanding the inhibition of strong response tendencies (e.g., discrimination reversal, go/no go). Thus, also in humans, prefrontal lesions result in deficits of short-term memory and control of interference.

As might be expected from these observations, the human subject with a prefrontal lesion is likely to be deficient in the temporal organization of sequential behaviors, such as the spoken language, that rely heavily on cognitive operations. The most dramatic language difficulty resulting from prefrontal pathology is Broca's aphasia, a severe deficit in the construction and articulation of speech resulting from lesions of a small area of the left inferolateral prefrontal cortex. Lesions of more anterior and superior cortex induce more subtle deficits, characterized by a general impoverishment of language and trouble in the construction of elaborate propositions; there is a diminution of spontaneous speech production and of the use of dependent clauses.

In humans, as in animals, extensive bilateral lesions involving most or all areas of the prefrontal cortex result in a generalized blunting of affect and emotions. Apathy and neglect characterize the subject's attitude toward others and the environment. Circumscribed lesions of dorsolateral prefrontal cortex induce cognitive deficits but only rarely changes in emotional behavior. On the other hand, in the monkey, orbital lesions result in fear and loss of normal aggressiveness; in the human, the effects are more variable and frequently include disinhibition, abnormal excitability, irritability, and poor ethical judgment. It is worth noting that the high incidence of emotional abnormalities deriving from ventromedial (in contrast to dorsolateral) prefrontal lesions is probably related to the anatomical and functional relationships between the ventromedial prefrontal cortex and the limbic structures of the hypothalamus and the temporal lobe.

Electrophysiology

Slow electrical potentials and sustained changes of neuronal discharge can be recorded from the prefrontal cortex in the time elapsing between a stimulus and a consequent delayed response to it. A surface-negative potential occurring between these two mutually contingent events is commonly known as the contingent negative variation (CNV) or expectancy wave. It may be observed in both humans and animals. Cellular

discharge shows important correlations with this field potential. By microelectrode recording in the monkey, large numbers of prefrontal cells can be seen to undergo altered discharge throughout the delay (seconds or minutes) that separates cue from choice in a delay task. The sustained activation of some prefrontal cells during that delay depends on the behavioral significance and the physical attributes of the cue. That of others depends on the characteristics of the impending motor response. Cells of the first type appear to participate in the short-term memory process; those of the second appear to do it in the preparation for movement. The two types of cells are found interspersed in certain areas of dorsolateral prefrontal cortex.

The posterior fringe of the dorsolateral prefrontal cortex bordering the premotor cortex contains the so-called frontal eye fields. Electrical stimulation of sites in these cortical fields elicits conjugate eye movements. The cells within them show increases of discharge in temporal relation to ocular movements. In an animal conditioned to move its eyes in response to a visual stimulus, some cells are activated right after the stimulus, some right before the movement, and still others through the time that separates the two. Thus, like the neurons in the dorsolateral prefrontal cortex at large, but on a shorter timescale, eye field neurons seem to be involved in the temporal integration of sensory stimuli with their consequent motor responses.

Metabolism and Blood Flow

Modern methods yielding images of the working human brain by the use of computer tomography and radioactive tracers reveal that blood flow and metabolic activity increase in prefrontal areas of the cortex during the ideation and programming of speech and motor sequences. Premotor and motor areas mainly show activation during the performance of speech and skeletal movements. In the past few years, neuroimaging has revealed the activation of the human dorsolateral prefrontal cortex during performance of short-term memory tasks. Increased prefrontal metabolism has been demonstrated during memorization of verbal or visual material. The activation of dorsolateral prefrontal and premotor cortices has also been found related to the preparation of movements that are dependent on the material memorized. Therefore, the metabolic studies in the human confirm the basic inferences from neuropsychological and neurophysiological studies in nonhuman primates, namely, that the neuronal populations of dorsolateral prefrontal cortex are involved in the retention of sensory information for prospective action, as well as in the preparation of motor systems for that action. Both are essential for organizing behavior in the time domain.

General Conclusions

In conclusion, the functions of the prefrontal cortex are not yet fully understood, but the evidence from human and animal studies points to a critical role of this cortex in the temporal organization of behavior. The prefrontal cortex may play that role by the coordinated works of at least three subordinate functions, each with a somewhat different topographic representation within the prefrontal region: (1) short-term memory, (2) short-term set for action, and (3) control of interference. Thus, the various sectors of prefrontal cortex, heterogeneous as to connectivity and apparent behavioral function, may cooperate with one another and with other parts of the brain in providing the cognitive support that the organism needs to bridge cross-temporal contingencies and to form goal-directed and temporally extended structures of behavior.

See also: Aphasia: Sudden and Progressive; Brodmann's Areas; Cerebral Cortex; Neocortex: Origins; Prefrontal Cortex: Structure and Anatomy; Prefrontal Contributions to Reward Encoding; Premotor Cortex in Primates: Dorsal and Ventral; Premotor Areas: Medial; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Fuster JM (1995) *Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate*. Cambridge, MA: MIT Press.

Fuster JM (1997) *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe*. Philadelphia, PA: Lippincott-Raven.

Fuster JM (2001) The prefrontal cortex – an update: Time is of the essence. *Neuron* 2: 319–333.

Goldman-Rakic PS (1984) Modular organization of prefrontal cortex. *Trends in Neuroscience* 7: 419–429.

Swartz BE, Halgren E, Fuster JM, Simpkins F, Gee M, and Mandelkern M (1995) Cortical metabolic activation in humans during a visual memory task. *Cerebral Cortex* 3: 205–214.

Prefrontal Cortex: Structure and Anatomy

H Barbas, Boston University, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

The prefrontal cortex has been associated with central executive functions, involved in selecting relevant information and ignoring irrelevant information for the task at hand. To accomplish these functions the prefrontal cortex must have access to information from structures associated with sensory perception, memory, and emotions. Indeed, the prefrontal cortex receives a wealth of information from a large variety of cortical and subcortical structures. In turn, the prefrontal cortex issues projections to the same structures, targeting both excitatory and inhibitory neurons, a pattern that likely underlies the selection of relevant signals and the suppression of irrelevant signals. Within the prefrontal cortex there is division of labor and specialization, so that distinct areas receive and issue topographically organized projections. The participation of different prefrontal sectors in complex behavioral settings appears to be complementary. A key distinction of prefrontal cortex among high-order association areas is its orientation for action, facilitated by its rich and diverse connections with motor control systems.

The prefrontal areas commonly associated with central executive functions are the lateral areas (**Figure 1(b)**), particularly the areas around the posterior principal sulcus and prearcuate region. These areas have a role in cognitive tasks, requiring the selection, retrieval, and temporary holding of information long enough to accomplish a task at hand. The prefrontal cortex, however, has additional components on the medial and orbitofrontal surfaces, which have a key role in the function of the prefrontal cortex (**Figures 1(a) and 1(c)**). The posterior parts of medial and orbitofrontal cortices belong to the cortical component of the limbic system by virtue of their architecture and connections (**Figure 1**, dark blue). Damage to prefrontal limbic areas in primates has detrimental consequences on emotional behavior and social interactions. The limbic components of the prefrontal cortex have strong connections with the rest of the prefrontal cortex, providing the anatomic basis for the merging of pathways associated with cognitive and emotional processes.

Location of the Prefrontal Cortex in Primates

The frontal lobe extends from the central sulcus to the frontal pole. This large cortical expanse includes about half of the entire cortex in primates. The prefrontal cortex occupies roughly the anterior half of the frontal lobe in macaque monkeys, and about two-thirds in humans. The primary motor cortex, also known as M1 (or area 4 of Brodmann), marks the posterior extent of the frontal lobe, and the premotor cortex (also known as area 6) is interposed between the primary motor and prefrontal cortices (**Figures 1 and 2**). Species other than primates have a prefrontal cortex, and at least in mammals it is found in front of the premotor cortex. In birds, a region thought to be analogous to the prefrontal cortex, on the basis of physiological properties and connections, occupies the posterior pole of the telencephalon.

The prefrontal cortex is greatly expanded and specialized in primates, and its architectonic organization and connections have been studied most widely in nonhuman primates, and more specifically, in macaque monkeys. In this species, the lateral aspect of the prefrontal cortex extends from the arcuate sulcus to the frontal pole (**Figure 1(b)**, light blue). The medial prefrontal cortex lies in front of the medial parts of the premotor cortex –the supplementary motor area (SMA), or MII – and the cingulate motor areas, and extends from the anterior part of the corpus callosum in the anterior cingulate to the medial part of the frontal pole (**Figure 1(a)**). The orbitofrontal cortex occupies the basal aspect of the brain. In macaque monkeys it is located behind the eye socket, and forms a concave shape around the orbit (**Figure 1(c)**). This large cortical region extends from the olfactory areas and the anterior insula to the basal surface of the frontal pole.

Global and Local Architecture

The prefrontal cortex is composed of 13 principal architectonic areas in the human in the classic map of Brodmann (**Figure 2**), and 12 in the nonhuman primate rhesus monkey, as depicted in the map of Walker (**Figure 3**). These classic maps have been modified by several investigators, and nearly all architectonic areas in the macaque monkey prefrontal cortex have been further subdivided into two or more subdivisions, now numbering over 20 in this species. The

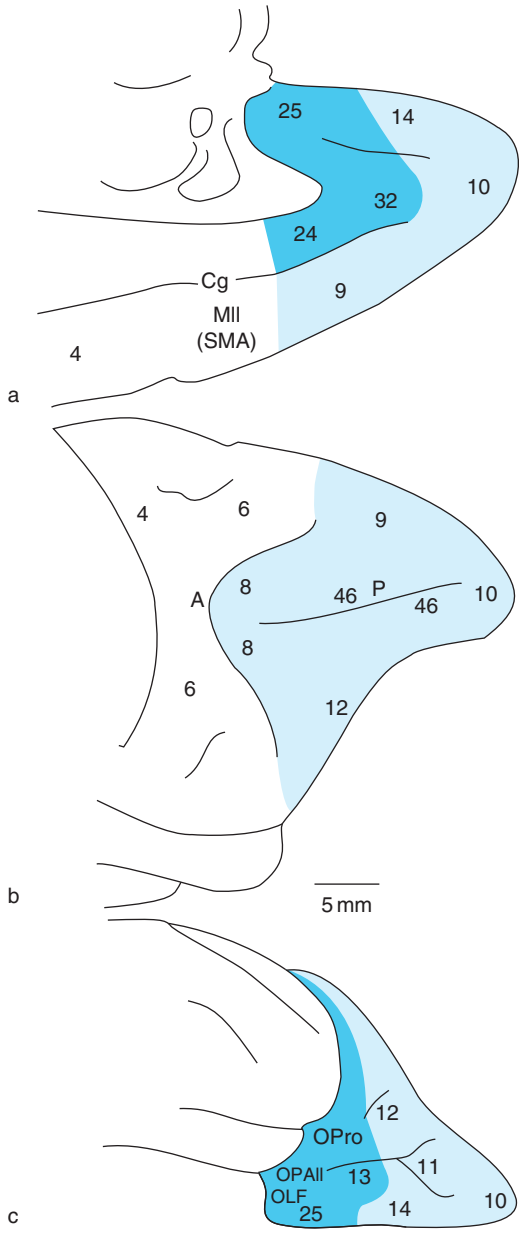

Figure 1 The principal architectonic areas of the prefrontal cortex of the rhesus monkey brain (shaded areas), depicting the (a) medial, (b) lateral, and (c) orbital surfaces. Light blue shading indicates eulaminate areas, with six cortical layers and a distinct granular layer IV. Dark blue shading indicates areas with fewer than six layers (noneulaminate) and which have a poorly developed layer IV (dysgranular) or no granular layer IV (agranular). The prefrontal cortex is bounded posteriorly by the premotor cortices (supplementary motor area MII; SMA) on the medial surface (a), and area 6 on the lateral surface (b). A, arcuate sulcus; Cg, cingulate sulcus; MII SMA, supplementary motor area; OLF, olfactory area; OPAII, orbital periallo cortex; OPro, orbital proisocortex; P, principal sulcus.

classic maps were largely based on brain sections stained for Nissl substance, which labels the cell bodies of neurons and glia. More recent maps have relied on an increasing number of specific neurochemical and molecular markers that aid in delineating architectonic borders by labeling distinct classes of neurons. Nevertheless, the Nissl stain has provided the most reliable information about the organization of the cortex (because architectonic areas vary reliably in overall density) and the relative distribution of neurons in different cortical layers.

The popular name 'frontal granular cortex' is based on the presence of a well-delineated granular layer IV, which is particularly prominent in lateral prefrontal cortex. Most prefrontal areas, including the entire lateral surface, the anterior medial surface, and the anterior orbitofrontal region, have six layers, including a granular layer IV, and thus are called eulaminate (**Figure 1**, light blue). However, the width and density of neurons in layer IV vary widely among prefrontal areas. Layer IV is densest and widest in lateral prefrontal area 8 and posterior area 46 (prearcuate region) of macaque monkeys (**Figure 4**), and least developed in anterior medial and anterior orbitofrontal areas (**Figures 1(a) and 1(c)**, light blue). But not all prefrontal cortex has a granular layer IV. The most posterior parts of the medial and orbitofrontal cortex have fewer than six layers and are either agranular, meaning that they lack granular layer IV, or dysgranular, meaning that they have a poorly developed layer IV (**Figures 1(a) and 1(c)**, dark blue). In the posterior orbitofrontal and anterior cingulate areas the deep layers (V and VI) predominate in cell density in comparison with the upper layers (II–III) (**Figure 5**). In contrast, all eulaminate prefrontal areas, and particularly the lateral areas, have a well-developed granular layer IV and a broad layer III, which is densely populated with neurons. The geographically and structurally distinct sectors of the prefrontal cortex have some common connections as well as some specific connections that underlie their specialization.

Common and Specialized Thalamic Connections of Prefrontal Cortices

All prefrontal cortices have bidirectional connections with the mediodorsal (MD) thalamic nucleus. In fact, the prefrontal cortex has often been defined as the cortical region that is connected with the MD thalamic nucleus. This definition is only approximate, because the prefrontal cortex is also connected with thalamic nuclei other than the MD nucleus, and conversely, the MD thalamic nucleus projects to areas outside the classic geographic extent of the prefrontal cortex, albeit more sparsely. Nevertheless, the MD thalamic nucleus is robustly connected with all prefrontal cortices in a topographic manner. Lateral prefrontal cortices are connected with the lateral

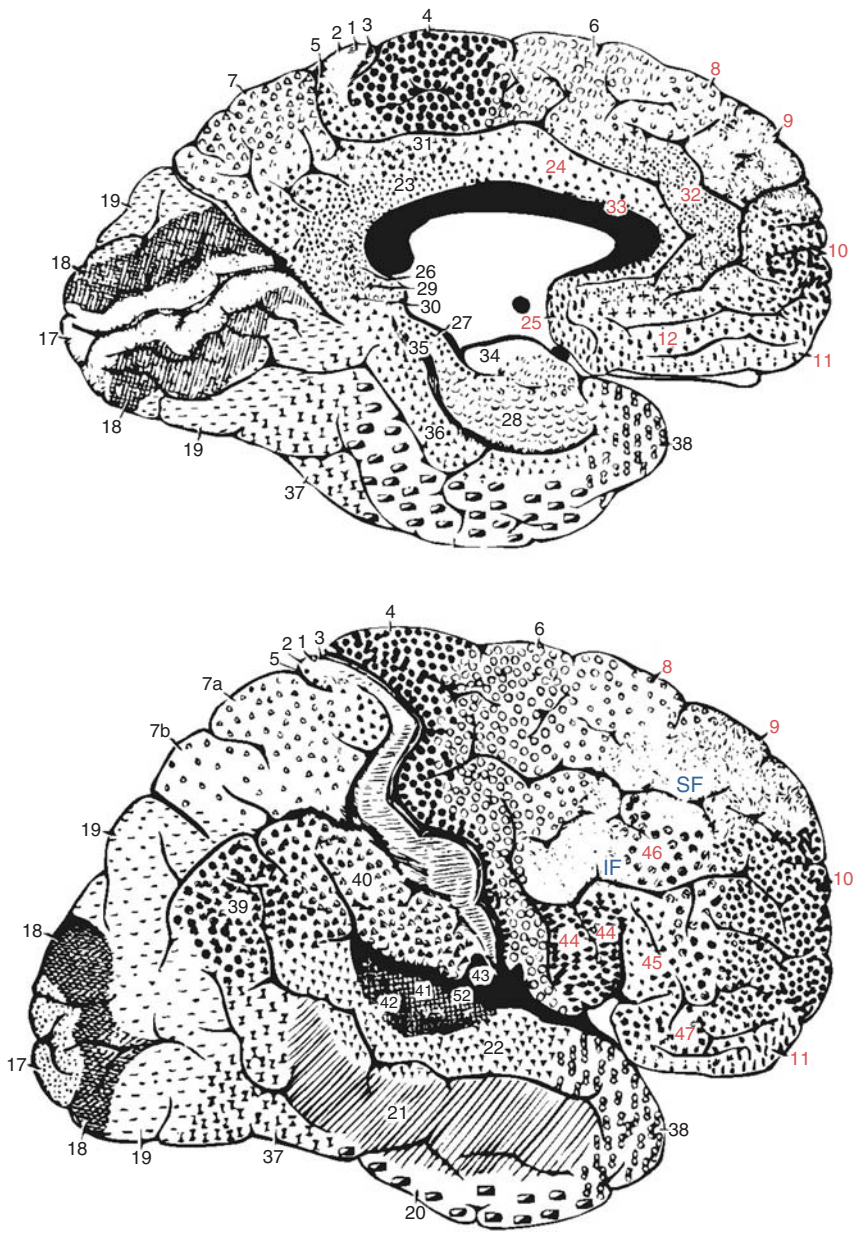

Figure 2 The architectonic map of the human cortex by Brodmann. Prefrontal areas are shown by numbers in red. The medial surface is shown on the top, and the lateral surface is shown on the bottom. IF, inferior frontal sulcus; SF, superior frontal sulcus.

parts of the MD thalamic nucleus, including its most lateral or multiform sector, and with the adjacent parvicellular sector. 'Multiform' refers to the various shapes of neurons, and 'parvicellular' refers to the small size of neurons. Medial and orbitofrontal cortices are connected mainly with the medial, or magnocellular (large cell), sector of the MD thalamic nucleus.

In addition to their connections with the MD thalamic nucleus, the prefrontal cortex is connected with other thalamic nuclei, including the ventral anterior, the medial pulvinar, intralaminar, medial, and anterior nuclei. Lateral prefrontal areas are connected primarily with the MD thalamic nucleus, and have only sparse connections with midline nuclei, but they have connections with some intralaminar nuclei and the medial pulvinar. On the other hand, orbitofrontal and posterior medial prefrontal areas have significant connections with midline and intralaminar nuclei, anterior nuclei, and the medial pulvinar, but still have substantial thalamic connections with the MD thalamic nucleus.

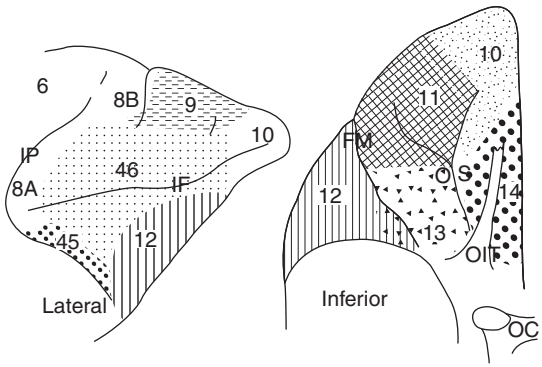

Figure 3 Architectonic areas of the rhesus monkey prefrontal cortex according to the map of Walker. The lateral surface is on the left, the orbital surface (inferior) is on the right; the medial surface is not shown.

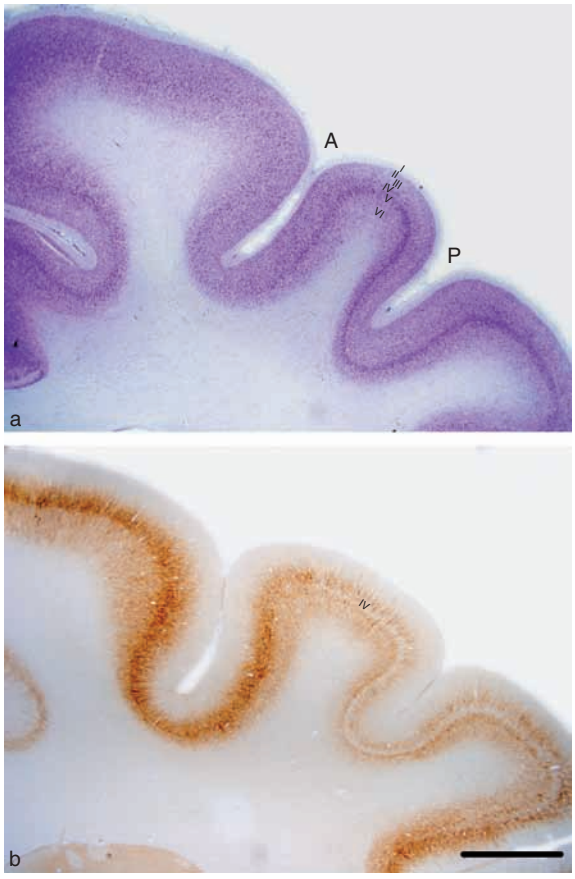

Figure 4 Example of granular prefrontal cortices. (a) Cross section (coronal) through the posterior part of prefrontal cortex of a macaque monkey (*Macaca nemestrina*), showing granular prefrontal cortex within the lateral bank of the arcuate sulcus (A) and adjacent cortex (to the right), and within the banks of the principal sulcus (P). The section was stained with cresyl violet to show the cell bodies and glia. Roman numerals indicate cortical layers. (b) A section at about the same level through the prefrontal cortex of a rhesus monkey brain stained with an antibody to neurofilament protein (SMI-32), which labels neurons in layers II–III and V–VI but not in layer IV. The empty zone in this frame shows clearly the position of granular layer IV. Medial is to the left. Scale bar = 3 mm.

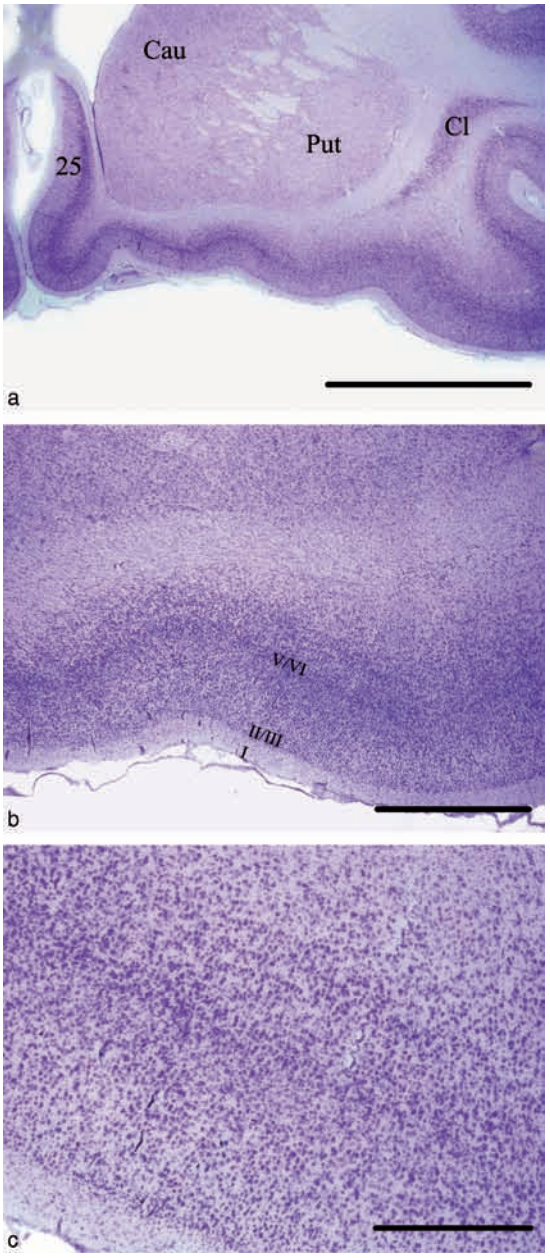

Figure 5 Example of agranular and dysgranular prefrontal cortices. (a) Low-power brightfield photomicrograph showing agranular- and dysgranular-type cortex on a coronal section through the orbitofrontal region of a macaque monkey (*Macaca nemestrina*) stained for Nissl substance. Area 25 on the medial surface is also dysgranular. (b, c) The orbitofrontal cortex at higher magnifications. Medial is to the left. Scale bar = 5 mm (a), 1 mm (b), 400 μm (c). Cau, caudate; Cl, claustrum; Put, putamen.

Specialized Connections of Prefrontal Cortices with Other Cortical Systems

Projections from Sensory Association Areas to Prefrontal Cortices

One of the key sources of information to the prefrontal cortex comes from most cortical sensory and

polymodal association areas. The primary koniocortices ('dustlike'), which include visual area V1, somatosensory area 3b, and auditory area A1, do not project directly to the prefrontal cortex. On the lateral surface of the prefrontal cortex, projections from visual association cortices reach several areas in the prearcuate region, targeting robustly area 8, which includes the frontal eye fields, and the posterior extent of the principal sulcus (area 46) and lateral area 12. Auditory association cortices project to several sites within the lateral prefrontal cortex, targeting robustly the frontal polar cortex (area 10), several sites of the dorsal bank of the principal sulcus (area 46), and the anterior tips of the limbs of the arcuate sulcus, including the anterior part of dorsal area 8, and area 9. Projections from somatosensory cortices reach the central portion of the principal sulcus (ventral area 46) and the subadjacent area 12. Cortices which are not unimodal, identified as those with neurons that respond to stimuli from two or more modalities, project to prefrontal cortices as well. On the lateral surface these projections originate mostly from the depths of the upper bank of the superior temporal sulcus, and to a lesser extent from the inferior parietal lobule.

Medial prefrontal areas receive projections from auditory association cortices, from polymodal areas in the depths of the upper bank of the superior temporal sulcus and perirhinal region, and only sparse projections from either visual or somatosensory association cortices. The orbitofrontal cortex, on the other hand, receives projections from all unimodal sensory cortices, including the primary olfactory cortices, and visual, auditory, somatosensory, and gustatory association cortices, as well as from polymodal association areas in the superior temporal and perirhinal regions.

There is a big difference in the topography of sensory projections to prefrontal cortices. The prearcuate cortex receives projections from posterior visual association areas, which represent relatively early stages of visual processing. The frontal eye fields seem to be specialized for processing signals from a large variety of visual association cortices, extending from as far back as V2 in the occipital region, up to the anterior parts of the inferior temporal visual association cortices. In contrast, orbitofrontal cortices receive projections from anterior inferior temporal cortices, where the visual receptive fields are global.

There is a comparable organization in the projections from auditory association areas to prefrontal cortices. The prearcuate cortex (area 8; at the tip of the upper limb of the arcuate sulcus) and posterior area 46 receive projections from posterior auditory association cortices in the superior temporal gyrus; these cortices are comparatively close to the primary auditory cortex. In contrast, orbitofrontal and medial prefrontal areas receive projections from the anterior superior temporal gyrus, from areas that process signals associated with species-specific vocalizations in nonhuman primates. Similarly, projections from other sensory association areas to the posterior orbitofrontal cortex originate in late-processing sensory association cortices. Visual input to orbitofrontal cortex, for example, originates in the anterior inferior temporal cortex, where neurons have large visual receptive fields and have a role in visual memory. The only exception is the input from primary olfactory areas to posterior orbitofrontal cortices.

Interface of Prefrontal Projections with Excitatory and Inhibitory Systems

One of the fundamental functional attributes of the prefrontal cortex is the selection of relevant signals and the suppression of irrelevant signals in behavior. This function can be illustrated by the interaction of prefrontal cortices with temporal auditory association cortices. Damage to the lateral prefrontal cortex after a stroke, or in pathological aging, impairs the ability of humans to ignore irrelevant auditory stimuli and thus to follow a conversation in a noisy environment. This function is likely mediated through robust projections from prefrontal cortices to auditory association cortices. Long-distance prefrontal pathways to sensory cortices are excitatory in nature, but they interface with both excitatory and inhibitory systems in auditory association areas. These pathways may form the anatomic substrate for choosing relevant signals and ignoring irrelevant signals. Similar interactions occur between medial prefrontal areas and auditory association areas. Interestingly, medial prefrontal areas in the anterior cingulate are hypoactive in schizophrenia, which likely affects medial prefrontal to temporal cortical communication. The same types of interactions with both excitatory and inhibitory systems occur between prefrontal and parietal cortices, as well as with other sensory and high-order association cortices.

Pathways Related to Specific Aspects of Memory Processing in the Prefrontal Cortex

Lateral Prefrontal Areas and Working Memory

Posterior lateral prefrontal areas are robustly and bidirectionally connected with lateral intraparietal areas, a region associated with visuomotor functions. Connections with the visuomotor parts of the lateral intraparietal cortex are restricted to the prearcuate region, including the cortex within the anterior bank

of the arcuate sulcus and the adjacent cortex (area 8), and to the posterior quarter of the principal sulcus (caudal area 46) in macaque monkeys. The prearcuate and lateral intraparietal cortices have in common a role in selective attention, required in cognitive operations. The strong and bidirectional connections that link these distant cortices suggest that they may function in concert in cognitive operations. The posterior cingulate cortex, which has a role in oculomotor responses and attentional processes, is also connected with the prearcuate cortex.

Posterior Medial and Orbitofrontal Areas and Long-Term Memory

Medial and orbitofrontal cortices are robustly connected with structures associated with long-term memory. These areas include the perirhinal region (areas 35 and 36), the entorhinal cortex (area 28), and the parahippocampal cortex (areas TF and TH). In addition, posterior orbitofrontal areas, and especially medial prefrontal cortices in the anterior cingulate, receive projections from the hippocampal formation, originating from the subiculum, presubiculum, and CA1 fields. Posterior medial prefrontal areas share long-term memory functions with the hippocampus, and their damage in humans results in memory deficits.

Pathways for Emotional Processing

Among prefrontal cortices, the posterior orbitofrontal and posterior medial cortices have the heaviest connections with the amygdala, a group of nuclei in the temporal lobe associated with emotional processing. Most of these connections involve the basal complex of the amygdala, especially the basolateral nucleus, and to a lesser extent the cortical and lateral nuclei. The projections of the posterior orbitofrontal cortex with the amygdala are particularly specialized, as they terminate heavily onto a group of nuclei known as the intercalated masses, which are composed of inhibitory neurons and have a role in the internal processing of the amygdala. The connections of posterior orbitofrontal areas are with the posterior half of the amygdala, which, in turn, receives projections from all sensory association cortices (Figure 6). One hypothesis is that the rich direct and indirect sensory inputs to posterior orbitofrontal cortices underlie the process through which events acquire emotional significance.

Posterior medial prefrontal cortices also have substantial connections with the amygdala. Projections from posterior medial prefrontal cortices reach heavily the output nuclei of the amygdala to hypothalamic and brain stem autonomic centers, and also terminate onto hypothalamic autonomic centers (Figure 7). In monkeys, posterior medial prefrontal cortices in the

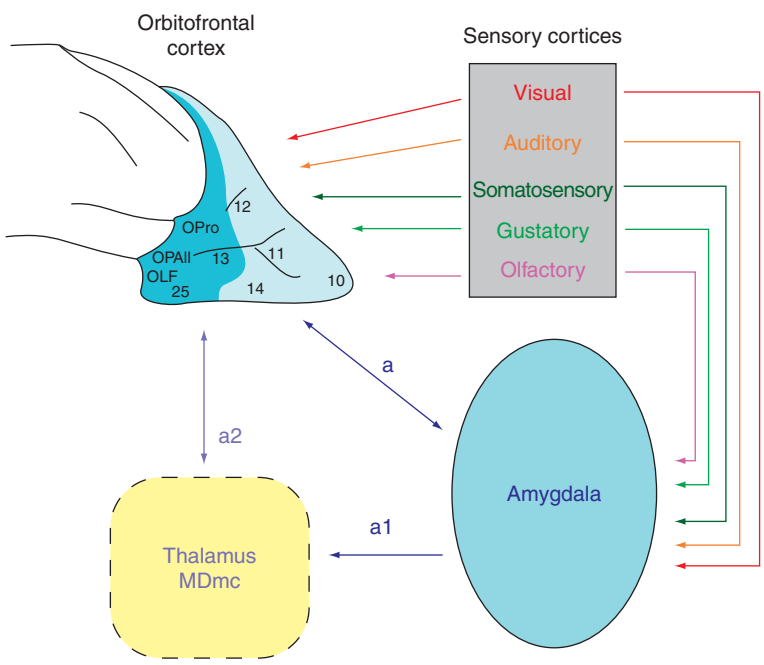

Figure 6 Multiple sources of projections to orbitofrontal cortices. Projections from sensory association cortices reach orbitofrontal cortex as well as the amygdala, which has bidirectional connections with orbitofrontal cortex (pathway a). The amygdala also projects to the medial (magnocellular) part of the mediodorsal thalamic nucleus (MDmc, pathway a1), which has bidirectional connections with the orbitofrontal cortex (pathway a2). Connections of sensory association cortices with orbitofrontal cortex, or with the amygdala, are also bidirectional (not shown). Architectonic areas indicated by letters are: OLF, olfactory area; OPAll, orbital periallo cortex; OPro, orbital prosiocortex.

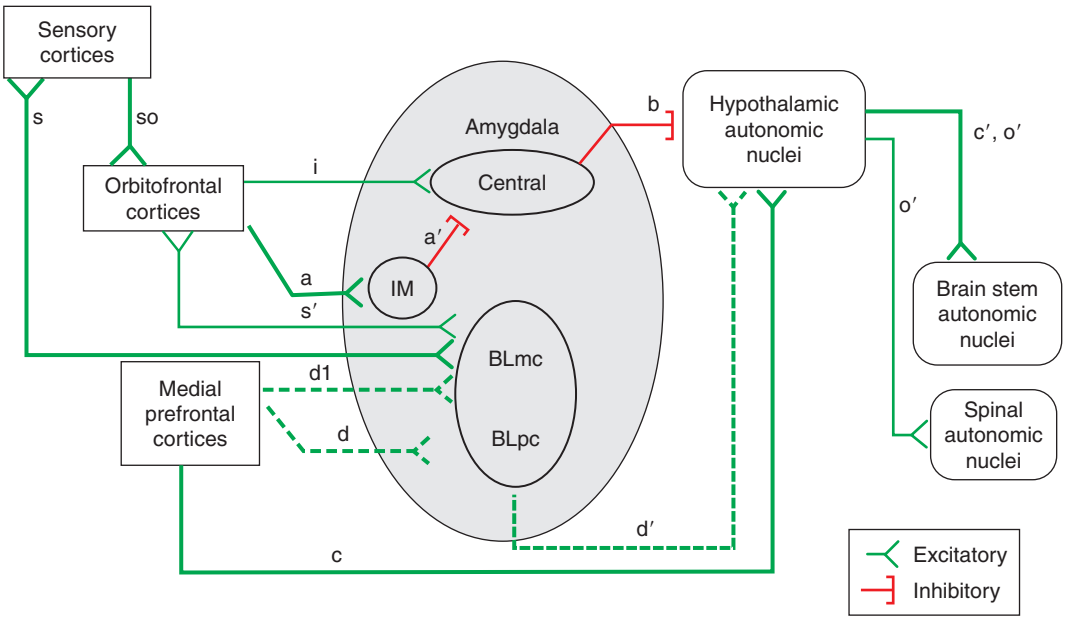

Figure 7 Summary of the connections of orbitofrontal and medial prefrontal cortices associated with perception and expression of emotions. The orbitofrontal cortex and the amygdala receive projections from every sensory modality through the cortex (pathways so, s), and are robustly connected with each other, providing the structural basis for direct (so) and indirect (s′) sensory information to the orbitofrontal cortex. Orbitofrontal axons terminate heavily in the intercalated masses (IMs) of the amygdala (pathway a), which then project to the central nucleus of the amygdala (pathway a′). The central nucleus projects to hypothalamic autonomic nuclei (pathway b). A direct pathway (c) from posterior medial prefrontal cortex reaches hypothalamic autonomic nuclei, which project to brain stem autonomic nuclei (pathway c′) and to spinal autonomic nuclei (pathway o′). Medial prefrontal cortices also project to the parvicellular basolateral nucleus of the amygdala (BLpc, pathway d), and to the magnocellular basolateral nucleus of the amygdala (BLmc, pathway d1), which project to hypothalamic autonomic nuclei (pathway d′), providing sequential pathways through which posterior medial prefrontal areas may influence hypothalamic autonomic nuclei (dotted pathways). A lighter pathway from orbitofrontal cortex terminates directly on the central nucleus of the amygdala (pathway i). Thickness of the lines represents the density of the projections. Red, inhibitory pathways; green, excitatory pathways. Reproduced from Barbas H, Saha S, Rempel-Clower N, and Ghashghaei HT (2003) Serial pathways from primate prefrontal cortex to autonomic areas may influence emotional expression. *BMC Neuroscience* 4:25, copyright 2003 Barbas H et al; licensee BioMed Central Ltd.

anterior cingulate region have a direct role in vocalization associated with emotional communication, emitted in fear or distress, such as at the sight of a predator.

Connections of Prefrontal Cortices with Neurotransmitter-Specific Systems

Like other cortical systems, the prefrontal cortex receives projections from modulatory neurotransmitter systems, including cholinergic projections from the basal forebrain, noradrenergic projections from the locus coeruleus, dopaminergic projections from the substantia nigra/ventral tegmental area, and serotonergic projections from the raphe nuclei. These systems have a role in the attentional and cognitive functions of the prefrontal cortex. In macaque monkeys and rats, the strongest interactions of the basal forebrain are with orbitofrontal and posterior medial prefrontal cortices and are distinct for being bidirectional, in contrast with other prefrontal cortices, which receive, but do not send, projections to the basal forebrain.

In the lateral prefrontal cortex a delicate balance in the availability of dopamine and noradrenaline appears to be crucial in optimal performance on working memory tasks in nonhuman primates. These functions are thought to be mediated through projections from neurotransmitter-specific systems, which target dopamine receptors of the D1 type, and adrenergic receptors of the α-2A type. Disruption of the optimal levels of these neurotransmitters in stress impairs performance on working memory tasks. Imbalance in the neurotransmitter equilibrium in prefrontal cortex may underlie attention-deficit/hyperactivity disorder in humans.

Another system with widespread connections to all sectors of the prefrontal cortex is the hypothalamus, which projects widely to all prefrontal cortices, but has bidirectional connections only with posterior orbitofrontal and posterior medial prefrontal cortices (Figures 1(a) and 1(c), dark blue). Some of these projections from the hypothalamus to the cortex may be histaminergic, constituting yet another neurochemically distinct pathway.

Common Connections of Prefrontal Cortices with Motor Control Systems

Linkage of Prefrontal Cortices with the Basal Ganglia

The prefrontal cortex is in an ideal position to exercise executive control through its multiple links with motor control systems. All prefrontal cortices project to the neostriatum (caudate and putamen), the input center of the basal ganglia. The neostriatum receives projections from all cortical areas, but the output of the basal ganglia (the globus pallidus and substantia nigra reticulata) reaches selectively the frontal cortex, including the motor/premotor and prefrontal cortices. Information from the basal ganglia reaches the frontal cortex indirectly, through specific thalamic nuclei. The principal nuclei involved in this system, as it pertains to the prefrontal cortex, are the mediodorsal nucleus and the ventral anterior nucleus, and to a lesser extent the anterior medial nucleus, all of which have bidirectional connections with the prefrontal cortex. Distinct prefrontal areas are influenced by the basal ganglia through parallel pathways, suggesting that they have specific roles in behavior. However, the segregation of pathways through the basal ganglia is relative rather than absolute, so that the pathway from lateral prefrontal cortices to the basal ganglia through the thalamus reaches not only lateral prefrontal cortices, but the adjacent premotor cortex as well. These connections suggest that lateral prefrontal cortices have multiple interactions with premotor cortices for action.

The projections of the prefrontal cortex are directed to the anterior part of the neostriatum, terminating mostly in the caudate nucleus. These projections have a certain degree of topographic organization, so that lateral prefrontal areas reach mostly the dorsal part of the caudate nucleus, medial areas project to a more medial position within the caudate, and posterior orbitofrontal cortices project to the ventral striatum. The latter is also known as the limbic striatum, and is also connected with the amygdala and the hippocampus. The limbic striatum projects through a relay in the ventral pallidum to the medial (magnocellular) part of the mediodorsal thalamic nucleus, which has strong bidirectional connections with the posterior orbitofrontal cortex (Figure 6) and the anterior cingulate.

Specialized Connections of Prefrontal Areas with Motor Control Systems

Lateral Prefrontal Cortices Are Linked with Brain Stem and Cortical Premotor Systems

Prearcuate prefrontal cortices are connected with neighboring premotor areas, which, in turn, are connected with the motor cortex, suggesting that the prefrontal cortex has close access to cortical motor control systems. Another specialized motor circuit is the projection of the frontal eye fields to the superior colliculus and brain stem oculomotor control systems, which are involved in eye movements. The frontal eye fields, situated within cortical area 8, were so named after the observation that damage to the area in humans or monkeys temporarily shifted the eyes toward the side of the lesion. It subsequently became clear that the frontal eye fields are not necessary for the execution of eye movements. It is now known that the involvement of the frontal eye fields in eye movement is related to attentional processes and to orienting the eyes to relevant stimuli in the environment.

Posterior Parts of Medial and Orbitofrontal Cortices Have Access to Central Efferent Autonomic Structures

The medial and orbitofrontal cortices have a different role in motor control. The posterior orbitofrontal cortex and, especially, the anterior cingulate cortex are connected with hypothalamic and brain stem autonomic structures (Figure 7), which innervate peripheral organs, such as the heart, the lungs, and the gut, the activities of which are markedly changed in emotional arousal. These prefrontal areas, therefore, have a role in motor control associated with emotional expression. These circuits are consistent with evidence that damage to orbitofrontal cortex in humans impairs the ability to respond appropriately in emotional situations. Conversely, overactivity of the orbitofrontal cortex is seen in patients with anxiety disorder and obsessive–compulsive disease.

Medial prefrontal cortices have an additional specialization within the motor domain, namely emotional vocalization, effected through a projection to brain stem structures that innervate laryngeal muscles necessary for phonation. This pathway has a role in emotional communication, such as the characteristic cries emitted by monkeys in emotional distress.

There are also other projections from orbitofrontal and, especially, anterior cingulate cortices to central autonomic structures targeting noncholinergic parts of the basal forebrain. One of these basal forebrain regions is the ventral striatopallidum, found in the dorsal half of the basal forebrain, and another is the extended amygdala, which is made up of groups of neurons scattered from the central nucleus of the amygdala through the basal forebrain to the bed nucleus of the stria terminalis. These specialized pathways provide parallel avenues from anterior cingulate and posterior orbitofrontal cortex to central autonomic structures.

Functional Relatedness of Cortical and Subcortical Connections of Prefrontal Cortices

Structures that are anatomically connected collaborate in distinct functions. For example, lateral prefrontal cortices in the posterior part of the principal sulcus and the adjacent prearcuate cortex have a role in working memory, as do the parts of the MD thalamic nucleus with which they are bidirectionally linked. Experimental inactivation of the lateral prefrontal cortex disrupts neuronal responses in the MD thalamic nucleus, observed during working memory tasks, suggesting the interdependence of these structures in behavior. The frontal eye fields have additional thalamic connections with intralaminar neurons, and in both structures neurons have visual and visuomotor properties. The prearcuate and lateral intraparietal regions are also strongly connected bidirectionally, and both have a role in oculomotor responses related to cognitive operations.

Posterior orbitofrontal and anterior cingulate areas are connected with the magnocellular part of the MD thalamic nucleus and with midline and anterior thalamic nuclei, and all of these structures have a common role in long-term memory and emotional memory. Similarly, the polymodal nature of projections to the posterior orbitofrontal region is matched by an equally polymodal input from the same high-order sensory association cortices to the amygdala (Figure 6). In turn, the amygdala and the orbitofrontal cortices have strong bidirectional connections, involving the same parts of the amygdala that receive input from sensory and polymodal cortices (Figure 7). Finally, medial prefrontal cortices have the strongest connections with central autonomic structures in several distinct brain regions. These pathways include hypothalamic and brain stem autonomic centers, basal forebrain structures, and nuclei of the amygdala that are preferentially connected with central autonomic structures.

Summary

Specialized sectors of the prefrontal cortex have a distinct set of connections and complementary roles in cognition, memory, and emotion. Cortical input to each prefrontal sector is complemented by connections with functionally related subcortical structures. The connections of the prefrontal cortex with the MD thalamic nucleus appear to unify them as one cortical region, though they occur with topographically and functionally distinct sectors of the MD thalamic nucleus, consistent with the specialization of each region in cognitive, mnemonic, and emotional operations.

Notwithstanding the distinctiveness of the set of connections of each prefrontal sector, there are connections between prefrontal sectors, suggesting that they function synergistically in behavior. Moreover, the different sectors of the prefrontal cortex target specific layers of each other, in an intricate but highly ordered communication system. The posterior orbitofrontal and anterior cingulate areas, for example, issue projections through their deep layers, and their axons terminate in the upper layers of eulaminate prefrontal cortices, especially layer I, suggesting 'feedback' communication. Conversely, eulaminate prefrontal areas communicate with anterior cingulate and orbitofrontal cortices through their supragranular layers (layers II–III), and their axons reach the middle–deep layers of noneulaminate cingulate and posterior orbitofrontal cortices. Reciprocity of connections is a common feature in cortical systems and appears to be fundamental to normal function. In their participation in bidirectional circuits, posterior orbitofrontal and the anterior cingulate cortices may function as feedback systems to the neuraxis. The frequent involvement of posterior medial and orbitofrontal cortices in neurologic and psychiatric diseases may be related to disconnection of feedback systems necessary for the synthesis of cognition, memory, and emotions, as these areas are linked with other parts of the prefrontal cortex.

See also: Emotion Systems and the Brain; Frontal Cortex Evolution in Primates; Frontal Eye Fields; Orbitofrontal Cortex: Visual Functions; Prefrontal Cortex; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Alexander GE, Delong MR, and Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience* 9: 357–381.

Arnsten AF and Li BM (2005) Neurobiology of executive functions: Catecholamine influences on prefrontal cortical functions. *Biological Psychiatry* 57: 1377–1384.

Barbas H, Ghashghaei H, Rempel-Clower N, et al. (2002) Anatomic basis of functional specialization in prefrontal cortices in primates. In: Grafman J (ed.) *Handbook of Neuropsychology*, pp. 1–27. Amsterdam: Elsevier Science B.V.

Brodmann K (1909) *Vergleichende Lokalizationslehre der Grosshirnrinde*. Leipzig: Barth.

Chao LL and Knight RT (1997) Prefrontal deficits in attention and inhibitory control with aging. *Cerebral Cortex* 7: 63–69.

Damasio AR (1994) *Descarte's Error: Emotion, Reason, and the Human Brain*. New York: G. P. Putnam's Sons.

DeOlmos JS and Heimer L (1999) The concepts of the ventral striatopallidal system and extended amygdala. *Annals of the New York Academy of Sciences* 877: 1–32.

Foote SL and Morrison JH (1987) Extrathalamic modulation of cortical function. *Annual Review of Neuroscience* 10: 67–95.

Fuster JM (1989) *The Prefrontal Cortex*. New York: Raven Press.

Goldman-Rakic PS (1988) Topography of cognition: Parallel distributed networks in primate association cortex. *Annual Review of Neuroscience* 11: 137–156.

Joel D and Weiner I (1994) The organization of the basal ganglia–thalamocortical circuits: Open interconnected rather than closed segregated. *Neuroscience* 63(2): 363–379.

Lynch JC and Tian JR (2005) Cortico-cortical networks and cortico-subcortical loops for the higher control of eye movements. *Progress in Brain Research* 151: 461–501.

McFarland NR and Haber SN (2002) Thalamic relay nuclei of the basal ganglia form both reciprocal and nonreciprocal cortical connections, linking multiple frontal cortical areas. *Journal of Neuroscience* 22: 8117–8132.

Nauta WJH (1979) Expanding borders of the limbic system concept. In: Rasmussen T and Marino R (eds.) *Functional Neurosurgery*, pp. 7–23. New York: Raven Press.

Neafsey EJ (1990) Prefrontal cortical control of the autonomic nervous system: Anatomical and physiological observations. *Progress in Brain Research* 85: 147–166.

Petrides M (1996) Lateral frontal cortical contribution to memory. *Seminars in the Neurosciences* 8: 57–63.

Rose J and Colombo M (2005) Neural correlates of executive control in the avian brain. *PLoS Biology* 3: e190.

Usher M, Cohen JD, Servan-Schreiber D, et al. (1999) The role of locus coeruleus in the regulation of cognitive performance. *Science* 283: 549–554.

Uylings HB, Groenewegen HJ, and Kolb B (2003) Do rats have a prefrontal cortex? *Behavioral Brain Research* 146: 3–17.

Walker AE (1940) A cytoarchitectural study of the prefrontal area of the macaque monkey. *Journal of Comparative Neurology* 73: 59–86.

Relevant Website

http://brainmaps.org – Interactive high-resolution brain atlas (Regents of the University of California, Davis campus).

Premenstrual Dysphoric Disorder and Postpartum Major Depression: Chronobiology

B L Parry, C J Meliska, A M López, L F Martínez, D L Sorenson, S Nowakowski, and R T Loving, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Women have at least twice the risk of developing a major depressive episode (MDE) compared with men. They are prone to develop depressive episodes during times of reproductive hormonal change during the menstrual cycle, the postpartum period, and the menopause transition. This chapter focuses on depressive mood changes that occur during the menstrual cycle, termed premenstrual dysphoric disorder (PMDD) in the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (DSM-IV), and during the postpartum period. Some 3–5% of menstruating women experience PMDD, characterized by moderate to severe changes in mood (depression, anxiety, irritability, affective lability, anhedonia), cognition (difficulty concentrating, sense of being overwhelmed), and physical functioning (sleep, energy, appetite, pain). These disturbances interfere with work, school, or usual social activities and relationships with others. The PMDD diagnosis must be confirmed by prospective daily ratings which document that at least five of the symptoms are present most of the time during the last week of the luteal (premenstrual) phase, and begin to remit within a few days after the onset of the follicular phase (menses). Premenstrual mood symptoms usually begin during the teens to late 20s; women often seek treatment in their 30s. Selective serotonin reuptake inhibitor (SSRI) antidepressants, fluoxetine and sertraline, are approved by the Food and Drug Administration (FDA) for treatment of PMDD.

Postpartum depression is categorized as an onset specifier for mood disorders in the DSM-IV. Postpartum psychiatric syndromes have three types of presentation. (1) The maternity or 'baby blues,' occurring in 50–80% of postpartum women, is not considered a disorder. It is characterized by crying and rapid mood shifts after the third postpartum day, often remitting 7–10 days postpartum. (2) An MDE postpartum, occurring in 10–15% of mothers, 25% of whom may require hospitalization, generally has an insidious onset characterized by anxious, agitated features with sleep disturbance (even when the child is sleeping) within 4–6 weeks postpartum and sometimes progressing to symptoms of melancholia that peak 3–5 months postpartum. (3) Postpartum psychosis occurs in 1/500–1/1000 primiparous women; with subsequent deliveries, the risk increases to 1/3 to 1/5. Onset is typically acute, occurring within the first 2 weeks postpartum. It is characterized by hallucinations and delusions and represents a medical emergency that may require hospitalization to prevent infanticide, which occurs in 1% of women with postpartum psychosis. When treated early and aggressively with antipsychotic (but not atypical) medication for at least 6–8 weeks, with close follow-up for 6–12 months postpartum, prognosis is relatively good. Untreated, postpartum depression and psychosis become more severe with time, more likely to recur (one-third of episodes outside of the postpartum period), and more refractory to intervention, causing extensive disability to the woman and her family, and potentially impairing neurocognitive development of the child.

Alterations in reproductive hormones may exert pathogenic effects in predisposed individuals by influencing the circadian system. When estrogen and progesterone decline postpartum, or during the luteal phase of the menstrual cycle, they may destabilize the normal temporal relationship between the circadian rhythms of melatonin, cortisol, thyroid hormones, prolactin (PRL), and sleep. To elucidate the role of these hormones in depressive disorders linked to the reproductive cycle in women, it is important to examine the diurnal secretion of these hormones in relation to the light/dark cycle of the environment, to sleep, and to each other, as each of these factors influences the expression of biological rhythms.

Studies of the amplitude, phase, and temporal organization of the circadian rhythms of these hormones during reproductive epochs reveal differences between depressed patients (DPs) and matched healthy normal control (NC) women. Women with PMDD have decreased morning secretion of melatonin in the symptomatic luteal, compared with the asymptomatic follicular, menstrual cycle phase. In postpartum major depressive disorder (MDD), subjective and objective measures of sleep disturbance distinguish DP from NC women. Further study of circadian measures in well-defined diagnostic groups is needed to identify common versus distinctive neuroendocrine profiles for DP during each reproductive epoch. This information could serve as a basis for developing more specific treatment modalities.

Menstrual Cycle Circadian Neuroendocrine Studies

Menstrual Cycle Sleep Studies

Healthy women. Subjective sleep reports in six healthy women show sleep duration was an hour longer and most disturbed premenstrually (5 days preceding the onset of menses), and shortest at ovulation. In case histories, two adolescent females reported hypersomnia associated with menstruation. Small sample sizes and lack of objective measures of sleep, mood, or reproductive hormones to define menstrual cycle phase limit interpretations of these studies.

Objective sleep recordings in 13 healthy women showed that rapid-eye-movement (REM) sleep latency was shorter during the luteal, compared with the follicular, menstrual phase. Three study subjects, however, used oral contraceptives, and the follicular and luteal phase measures were not randomized or sequential.

In a careful study of objective and electroencephalogram (EEG) power density throughout one menstrual cycle (phase determined by temperature and hormonal measures), nine healthy women without symptoms of premenstrual syndrome (PMS) showed no significant variations across the menstrual cycle in subjective mood and sleep quality, or in objective measures of total sleep time, sleep efficiency, sleep latency, REM, or slow-wave sleep. Spectral analysis of non-REM (NREM) sleep, however, showed a large variation in power density (14.25–15.0 Hz band of sleep spindles) across the menstrual cycle, with a maximum in the luteal phase that paralleled core body temperature activity. In contrast, the time course of EEG slow-wave activity (reflecting homeostatic sleep regulatory mechanisms) did not vary across the menstrual cycle.

To investigate diurnal fluctuations of sleep propensity in the follicular and luteal phases, excluding masking effects of light and behavior, investigators conducted multiple nap tests using an ultrashort sleep–wake cycle in eight healthy women (20–23 years of age) without sleep disorders. Although there were no nighttime sleep differences between the two menstrual cycles, daytime (09:00–16:30 h) subjective sleepiness and napping increased from the follicular to the luteal phase, suggesting a relation to mechanisms controlling slow-wave sleep.

Depressed women. In one study, increased REM sleep premenstrually was correlated with severity of 'premenstrual tension' in four normal individuals and three inpatients with diagnoses that included depression and schizophrenia. Total sleep time and number of awakenings were unchanged; however, the fact that the subjects were taking medications, including oral contraceptives and stimulants, compromises the interpretation of these findings. While collecting sleep EEG data from four female students for eight nights over a single menstrual cycle, investigators found an increase in stage 2 and a decrease in delta (deep) sleep at the time of menses in the one subject who reported dysphoric mood premenstrually. Similarly, researchers compared sleep EEG, temperature, and activity two nights per week over the course of a menstrual cycle in eight women with moderate to severe premenstrual depression, with that of eight age-matched controls, after adaptation to the sleep laboratory. Depressed women had more stage 2 sleep and less REM sleep than healthy controls. Stage 3 sleep and the frequency of intermittent awakenings, but not temperature or motor activity, varied with menstrual cycle phase. Other investigators showed that women with premenstrual negative effect had less delta sleep than asymptomatic women during both follicular and luteal menstrual cycle phases. In contrast, another group of scientists observed no significant polysomnographic (PSG) changes related to the menstrual cycle in nine sleep parameters or differences between three patients with PMS (diagnosed by interview and a visual linear analog scale) and six control subjects.

Menstrual Cycle Melatonin Studies

Healthy women. One study showed a phase delay of the nocturnal peak of melatonin secretion during the mid-menstrual period (mid-cycle), although melatonin was sampled only every 2–4 h at two points in the cycle and without documentation of estradiol or progesterone levels. Other investigators found melatonin circadian rhythms were relatively stable and resistant to hormonal influences during the menstrual cycle in NC women. As noted above, eight healthy women who underwent an ultrashort sleep–wake cycle schedule had decreased serum melatonin secretion in the luteal, compared with the follicular, menstrual cycle phase.

Depressed women. In one study, melatonin secretion had an earlier (phase-advanced) offset, shorter secretion duration and a decreased area under the curve (AUC) in eight women with PMDD. In contrast, melatonin rhythms in eight NC women remained unchanged during four phases of the menstrual cycle. In a larger follow-up study of 21 PMDD and 11 NC subjects, melatonin onset time was delayed, duration was compressed, and AUC, amplitude, and mean levels were decreased in PMDD subjects, during the symptomatic luteal compared with the asymptomatic follicular menstrual cycle phase. Melatonin rhythms remained unchanged with the menstrual cycle in NC subjects.

Menstrual Cycle Cortisol Studies

Healthy women. Using ultrashort sleep–wake cycles in eight healthy women, scientists observed that the amplitude of cortisol rhythms was decreased in the luteal, compared with the follicular menstrual cycle phase.

Depressed women. In eight women with PMDD who had multiple samplings for cortisol during mid-cycle (late follicular) and the late luteal (LL) menstrual cycle phase, investigators reported that serum cortisol levels were increased during the mid-cycle phase. In 20 women with LL phase dysphoric disorder (LLPDD) and 11 NC subjects, in whom cortisol levels were measured every 30 min from 18:00 to 09:00 h, during the mid-follicular (MF) and LL menstrual cycle phases, in NC, but not in LLPDD, subjects, the cortisol peak was significantly delayed in the LL compared with the MF menstrual cycle phase. In a study of 15 women with PMDD and 15 NC subjects, timing but not quantitative measures of cortisol secretion were altered in PMDD versus NC: in the LL versus MF phase, the cortisol acrophase occurred 1 h earlier, on average, in NC, but not in PMDD subjects.

Menstrual Cycle TSH Studies

Healthy women. In the study of eight healthy women undergoing ultrashort sleep–wake cycle schedules, thyroid-stimulating hormone (TSH) rhythm amplitude decreased in the luteal, compared with the follicular menstrual cycle phase.

Depressed women. In a study of the circadian rhythms of TSH in 23 PMDD and 18 NC subjects in which samples for TSH were measured every 30 min from 18:00 to 09:00 h, TSH acrophase (time of the highest point of a fitted cosine curve) and peak time (based on absolute values) occurred earlier in PMDD than in NC subjects during MF and LL menstrual cycle phases.

Menstrual Cycle PRL Studies

In a study of 20 LLPDD and 11 NC women, investigators measured PRL every 30 min from 18:00 to 09:00 h during MF and LL menstrual cycle phases. PRL peak and amplitude were higher, and acrophase earlier, in LLPDD patients, than in NC subjects. In a related study of 23 PMDD and 18 NC subjects, PRL during MF and LL menstrual cycle phases was higher in PMDD than in NC subjects, consistent with previous findings.

Postpartum Circadian Neuroendocrine Studies

Postpartum Sleep Studies

Healthy women. Using subjective ratings, 30 primiparous women reported more evening awakenings, more time awake after retiring, and more naps than 28 nonpostpartum mothers, but sleep times were similar. Controlling for time awake at night indicated dysphoric mood in the first week in postpartum women was due to sleep disruption. Other investigators compared 12 primigravidae with 19 multigravidae. Primigravidae experienced more fatigue and disturbed sleep, third trimester sleep efficiency falling from 90% to 77% postpartum, while multigravidae had only a minor reduction in sleep efficiency. In a study of seven Japanese women, based on subjective sleep logs, wake time at night gradually decreased from the 5th to the 12th week postpartum. Hormonal measures were not specified in these studies.

Most PSG assessments suggest REM sleep decreases after delivery. Researchers noted only a slight increase in REM sleep in seven women in the postpartum, compared with the prepartum period. Sleep efficiency increased, and wake after sleep onset decreased by 3–5 months postpartum. A study of 31 women showed REM latency was shorter at 1 month postpartum, when mood state was most negative compared with the third trimester of pregnancy. In a cross-sectional study, other investigators compared 12 nonpregnant Japanese women with 10 primipara women at home from 9–12 weeks postpartum. Postpartum women had more interrupted sleep, lower sleep efficiency, decreased total sleep time, decreased stage 2 and increased stage 4 sleep. Another group found that 12 postpartum women who breast-fed babies (from 4 to 30 weeks) had increased slow-wave sleep, and a compensatory reduction in light NREM sleep compared with seven women who bottle-fed infants, or with 12 age-matched control subjects studied during the follicular menstrual cycle phase. Other investigators found total sleep time and sleep efficiency deteriorated in 10 women from the 5th prepartum to the 15th postpartum week, compared with late pregnancy, for 12 weeks after delivery. Increased wake after sleep onset persisted for 10 weeks postpartum. Studying arousal to auditory stimulation in six women showed that compared with pregnancy, there was a significant lowering of awakening threshold postpartum during stage 2, REM, and slow-wave sleep.

Depressed women. Compared with 28 women without pregnancy-related depressive episodes, 24 women with postpartum episodes that began during pregnancy had significantly longer REM time and more REM activity, suggesting a greater biological vulnerability to depression. Studies comparing women with ($n = 14$) and without ($n = 20$) a history of mood disorder show childbearing was associated with greater changes in total sleep time, and with REM reduction, apparently earlier in the childbearing course, that persisted throughout the 8th postpartum month in

women with prior history, supporting the association between sleep abnormalities and a vulnerability to mood disorders. Sleep disturbances diagnostic for MDE are strongly associated with the risk of future depression. As demonstrated for MDE, sleep disturbance may antedate subsequent postpartum depression. In a PSG study of 15 postpartum DP and eight NC subjects, DP had a longer sleep latency, later sleep onset time, lower sleep efficiency, and less total sleep time than NC subjects.

Postpartum Melatonin Studies

A study of 11 women with an MDE postpartum and five matched NC women showed DP had a trend toward higher mean melatonin peak and AUC compared with NC subjects. When age, weeks postpartum, and body mass index (BMI) were applied as covariates in the analysis, melatonin offset was significantly later in DP than in NC (09:23 vs. 07:52 h). Other timing measures (onset, duration) were not significantly different between groups.

Although not systematically studied in postpartum women, melatonin levels were reported lower in most studies comparing DP with NC subjects, consistent with replicated studies demonstrating lower melatonin circadian rhythms in patients with PMDD and in combined pregnant and postpartum DP compared with NC.

Postpartum Cortisol Studies

Investigators measured cortisol and progesterone from 2 weeks before delivery to day 35 postpartum in 120 primiparous women. Seven women developed major depression postpartum. Lower levels of evening cortisol (but not progesterone) in the immediate peripartum period were associated with postnatal depression. Another study, however, reported no significant differences in serum cortisol between 11 DP and five NC postpartum subjects.

Postpartum PRL Studies

PRL concentrations obtained at 20 min intervals for 12–24 h remained elevated in eight nonlactating healthy postpartum women, 20–26 days postpartum, with the diurnal pattern of secretion conserved. Serum PRL concentrations declined with each successive postpartum week. Another examination of plasma PRL between 08:00 and 09:00 h just prior to delivery and 3 days postpartum in 25 NC women showed the rise in PRL plasma levels postpartum had a negative and significant correlation with scores on the Hamilton Anxiety Scale, suggesting that high PRL plasma levels associated with milk production reduced anxiety in lactating women. When a group of 43 lactating women was randomly assigned to either breast-feed or merely hold their infants, both conditions decreased anxiety, as well as lowering adrenocorticotropin (ACTH) and plasma cortisol. Breast-feeding (but not holding) reduced cortisol responses to a brief psychological stressor.

In postpartum DP, researchers observed that of 147 postpartum mothers, 14.9% of whom were depressed, plasma PRL levels were inappropriately low in depressed women who breast-fed. In other studies, however, hyperprolactinemia has been associated with depressive symptoms. In a study of 11 postpartum DP and five NC subjects, in which serum PRL was sampled at 30 min intervals from breast-feeding mothers, PRL was elevated from 04:00 to 11:00 h in DP relative to NC, when weeks postpartum and BMI were applied as covariates in the analyses. In contrast, other investigators found lower plasma PRL levels in postpartum depressed women when only a single morning sample was obtained.

Reproductive Hormones

Estradiol. In 15 DP and eight NC postpartum women, estradiol was significantly lower in DP than NC when age, weeks postpartum, BMI, and breast-feeding status were applied as covariates in the analyses. DP and NC groups did not differ in progesterone, luteinizing harmone (LH), and follice stimulating harmone (FSH) levels. When applied as covariates, levels of reproductive hormones did not affect the outcomes of statistical analyses of other neuroendocrine measures.

Comparisons of Sleep and Circadian Neuroendocrine Hormones across Reproductive Events

Sleep in Menstrual Cycle versus Postpartum Women

Longer sleep latencies, later sleep onset times, and reduced stage 3 percentage sleep distinguish PMDD versus postpartum DP, while menstrual NC women show less stage 3 percentage and delta percentage sleep compared to postpartum NC women.

Neuroendocrine Hormones in Menstrual Cycle versus Postpartum Women

As expected, postpartum NC and DP women who breast-fed had higher PRL levels than nonlactating menstrual cycle women. In contrast, TSH levels were reduced in postpartum breast-feeding women compared with menstrual cycle women, when controlling for age and weeks postpartum. Nonbreast-feeding postpartum women did not differ from

menstrual cycle women in any of these neuroendocrine measures. A comparison across reproductive events showed that serum TSH levels of menstrual cycle women were marginally higher than postpartum women in NC and DP combined; DP and NC women did not differ reliably.

Summary

Menstrual Cycle Studies

While some studies suggest lighter, more disturbed sleep premenstrually and increased REM sleep and decreased REM latency in DPs, we must consider the marked inconsistency in the results, which can be attributed to variability of methods, small sample sizes, subjects on different medications, or lack of hormonal measures to define menstrual cycle phase and no standardized criteria to diagnose depression.

For menstrual cycle studies of serum or plasma melatonin, those that sampled more frequently, for example, at 30 min intervals, suggest that melatonin circadian rhythms are relatively stable and resistant to hormonal influences during the menstrual cycle in NC subjects. In contrast, women with well-defined mood symptoms during the menstrual cycle have decreased amplitude (quantitative) measures of melatonin, which are associated with changes in phase markers, in the symptomatic luteal phase. The findings suggest DPs have decreased amplitudes of the circadian pacemaker that impairs their ability to regulate biological rhythms.

With regard to cortisol, TSH, and PRL circadian rhythms during the menstrual cycle, the data suggest greater differences between healthy and depressed women in the timing, rather than in the quantitative measures of cortisol secretion. TSH levels are marginally higher in menstruating than in postpartum women, but do not differentiate DP and NC women. In contrast, increased PRL amplitude measures differentiate DP versus NC during the menstrual cycle. Sleep architecture studies in DP versus NC women generally have inconsistent findings, although most women with PMDD report more disturbed sleep premenstrually.

Postpartum Studies

In postpartum sleep studies, the general findings suggest that postpartum sleep is disrupted and does not return to baseline levels until at least 12 weeks postpartum (at which time the infant's sleep and melatonin rhythms tend to establish diurnal patterns).

Postpartum studies of melatonin, cortisol, TSH, and PRL circadian rhythms suggest melatonin and PRL are elevated in postpartum compared with healthy women. In contrast, in healthy postpartum women, higher PRL levels were associated with decreased anxiety in lactating women. Estradiol levels were lower in DP compared with NC postpartum women when age, weeks postpartum, BMI, and breast-feeding status were applied as covariates in the analyses. Impaired sleep, as measured primarily by objective (i.e., PSG) quality of sleep indices (e.g., decreased total sleep time and sleep efficiency, increased sleep latency, and delayed sleep onset) characterize postpartum DP compared with NC women.

Conclusions

Certain circadian neuroendocrine features distinguish DPs from healthy controls across reproductive epochs. Elevated PRL levels characterize menstruating and postpartum depressed women. Melatonin levels are lower in menstruating depressed women and higher in postpartum depressed women compared with healthy women. In future work, neuroendocrine variables should be examined for the influence of modulating factors such as age, season of the year studied, BMI, menstrual cycle phase, week postpartum, breast-feeding status, and estradiol and progesterone levels in well-defined diagnostic groups. With more precise measurements and larger sample sizes, further work in each of these reproductively related depressive disorders would serve as the basis for developing treatment modalities targeted to each specific population. These endeavors also have the potential to yield valuable information on how changes in reproductive endocrine function affect the brain's regulation of circadian rhythms and mood.

See also: Circadian Organization; Circadian Function and Therapeutic Potential of Melatonin in Humans; Circadian Rhythms: Influence of Light in Humans; Depression and the Brain; Estrus and Menstrual Cycles: Neuroendocrine Control; Hormones and Behavior; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Neuroendocrine Control: Maternal Behavior; Prolactin and its Neuroendocrine Control; Stress Response: Sex Differences.

Further Reading

Berga SL and Yen SS (1990) Circadian pattern of plasma melatonin concentrations during four phases of the human menstrual cycle. *Neuroendocrinology* 51: 606–612.

Driver HS, Dijk DJ, Werth E, Biedermann K, and Borbely AA (1996) Sleep and the sleep electroencephalogram across the menstrual cycle in young healthy women. *Journal of Clinical Endocrinology and Metabolism* 81: 728–735.

Meliska CJ, Maurer EL, Martinez LF, et al. (2005) Neuroendocrine correlates of sleep in premenstrual, pregnant, postpartum, and menopausal depression. In: Cardinali DP and Pandi-Perumal SR (eds.) *Neuroendocrine Correlates of Sleep/Wakefulness*, pp. 391–411. New York: Springer.

Parry BL (1999) Postpartum depression in relation to other reproductive cycle mood changes. In: Miller LJ (ed.) *Postpartum Mood Disorders*, pp. 21–46. Washington, DC: American Psychiatric Press.

Parry BL (2003) Management of depression and psychoses during pregnancy and the puerperium. In: Creasy RK and Resnick R (eds.) *Maternal–Fetal Medicine. Principles and Practice,* 5th edn., pp. 1193–1200. Philadelphia: Elsevier Science.

Parry BL, Berga SL, Kripke DF, et al. (1990) Altered waveform of plasma nocturnal melatonin secretion in premenstrual depression. *Archives of General Psychiatry* 47: 1139–1146.

Parry BL, Berga SL, Mostofi N, Klauber MR, and Resnick A (1997) Plasma melatonin circadian rhythms during the menstrual cycle and after light therapy in premenstrual dysphoric disorder and normal control subjects. *Journal of Biological Rhythms* 12: 47–64.

Parry BL, Gerner RH, Wilkins JN, et al. (1991) CSF and endocrine studies of premenstrual syndrome. *Neuropsychopharmacology* 5: 127–137.

Parry BL, Hauger R, LeVeau B, et al. (1996) Circadian rhythms of prolactin and thyroid-stimulating hormone during the menstrual cycle and early versus late sleep deprivation in premenstrual dysphoric disorder. *Psychiatry Research* 62: 147–160.

Parry BL, Hauger R, Lin E, et al. (1994) Neuroendocrine effects of light therapy in late luteal phase dysphoric disorder. *Biological Psychiatry* 36: 356–364.

Parry BL, Javeed S, Laughlin GA, Hauger R, and Clopton P (2000) Cortisol circadian rhythms during the menstrual cycle and with sleep deprivation in premenstrual dysphoric disorder and normal control subjects. *Biological Psychiatry* 48: 920–931.

Parry BL, Sorenson DL, Meliska CJ, et al. (2003) Hormonal basis of mood and postpartum disorders. *Current Women's Health Reports* 3: 230–235.

Parry BL, Udell C, Elliott JA, et al. (1997) Blunted phase-shift responses to morning bright light in premenstrual dysphoric disorder. *Journal of Biological Rhythms* 12: 443–456.

Wirz-Justice A, Benedetti F, Berger M, et al. (2005) Chronotherapeutics (light and wake therapy) in affective disorders. *Psychological Medicine* 35: 939–944.

Wirz-Justice A, Terman M, Oren DA, et al. (2004) Brightening depression. *Science* 303: 467–469.

Relevant Website

http://www.postpartumdepression.net – Mother-to-Mother: Postpartum Depression Network.

Premotor Areas: Medial

J Tanji and E Hoshi, Tamagawa University Research Institute, Tokyo, Japan

© 2009 Elsevier Ltd. All rights reserved.

Overview

At least five premotor areas are present in the medial part of the frontal cortex (see Figure 1). The supplementary motor area (SMA) is located immediately anterior to the hindlimb representation area of the primary motor cortex (M1). The SMA occupies the medial wall of the cortical hemisphere above the cingulate sulcus, extending several millimeters into the dorsal surface. The pre-SMA is located rostral to the SMA. Although both of these areas correspond to the medial agranular cortex (area 6), the anatomical connectivity of these two areas is strikingly different, suggesting different functional roles. Cingulate motor areas (CMAs) are located in the banks of the cingulate sulcus and should be distinguished from the cingulate gyrus (occupying the medial wall below the cingulate sulcus), which belongs to the limbic cortex. The rostral part, CMAr, corresponds to area 24 occupying the upper and lower banks of the cingulate sulcus. The caudal part, CMAc, is often subdivided into two parts corresponding to the upper and lower banks of the sulcus. The supplementary eye field (SEF) is an oculomotor area located anterolateral to the pre-SMA. The organizational principles of these medial motor areas are similar in humans and nonhuman primates.

Clinical investigations in human patients have discovered that unlike the primary motor cortex, lesions in the medial premotor areas cause little if any motor paralysis. Instead, the clinical signs include impairment in the initiation of voluntary movements or the release of reflexive hand movements (forced grasping in response to sensory stimuli). Moreover, the release of certain voluntary motions takes mysterious forms, including the alien hand sign, forced tool use, or the release of well-learned, routine motions. Furthermore, impairments in the temporal organization of sequential movements, the correct sequencing of multiple movements, the coordination of bilateral limb movements, or the rhythmic organization of motor routines have been the focus of attention in the literature describing the effects of lesions in the medial motor areas. Corresponding to these intriguing clinical findings, basic studies in system neuroscience have made discoveries that characterize each of the medial motor areas. Above all, the analysis of cellular activity in monkeys performing motor tasks has revealed the preferential use of individual areas for various aspects of motor behavior.

SMA

The SMA was first described by Penfield and Welch as the secondary motor area in humans. The term 'supplementary' was selected because of subtle lesion effects and the less striking effects of electrical stimulation than those observed in the M1. However, subsequent studies revealed that this area is involved in important aspects of motor control in a way very different from the M1. For many years since its discovery, the SMA was viewed as a single motor area. Based on anatomical and physiological studies, it is now established that the traditionally defined SMA should be divided into two different areas: the caudal area, now redefined as the SMA, and the rostral area, called the pre-SMA. These two areas, which have different anatomical connections with other brain areas, are now viewed as separate areas with distinct functional properties. In addition, a separate area involved in oculomotor control (the SEF) has also been identified at a cortical site situated rostrolateral to the pre-SMA.

Basic Organization and Anatomical Connectivity

In both humans and monkeys, the SMA is organized somatotopically. Its subareas representing the face, forelimb, trunk, and hindlimb are arranged in essentially an anterior-to-posterior order in such a way that the face area is immediately posterior to the pre-SMA and the hindlimb area is rostral to the hindlimb area of the M1. The distal and proximal parts of the limb are represented in the SMA, indicating that this area is not merely the area for postural control. The SMA sends its output to other motor areas in the cortex, the M1 and premotor cortex, and to the striatum, brain stem, and spinal cord. It receives afferent inputs from the primary and secondary somatosensory areas, as well as from the M1, premotor cortex, and CMAs. Through afferents coming from the thalamus, the SMA receives more information from the basal ganglia than from the cerebellum.

Effects of Lesions or Inactivation

In humans, the clinical consequences of ablation of the medial frontal cortex, including the SMA, have been reported as deficits in initiating movements or speech, the release of pathological reflexive movements (forced grasping), and at later stages, the involuntary release of learned movements. It is important

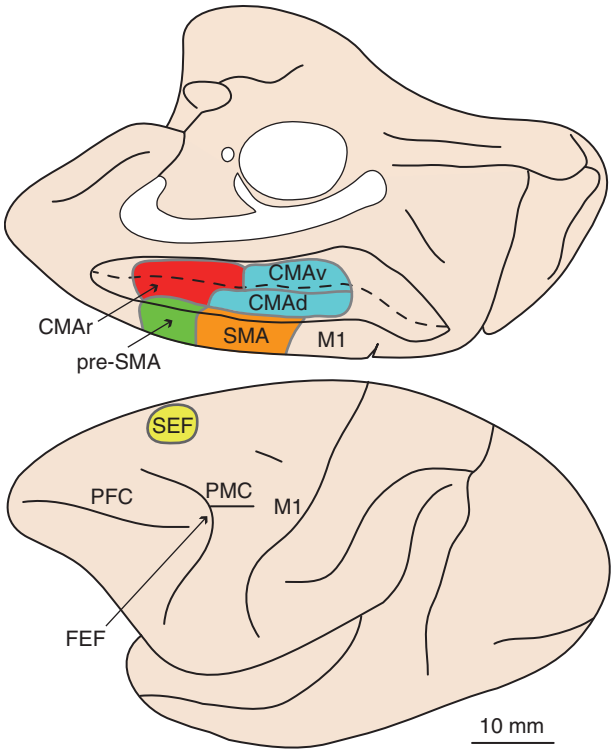

Figure 1 Schematic drawing of the cortical areas in the macaque frontal cortex. The lateral view is illustrated in the bottom panel. The top panel shows the medial view in an inverted manner. The cingulate sulcus is opened; its lips are depicted with solid lines, and its fundus is depicted with a broken line. CMAd, dorsal cingulate motor area; CMAr, rostral CMA; CMAv, ventral CMA; FEF, frontal eye field; M1, the primary motor cortex; PFC, prefrontal cortex; PMC, premotor cortex; pre-SMA, presupplementary motor area; SEF, supplementary eye field; SMA, supplementary motor area.

to note that as long as the lesions do not include the M1, the limbs are not paralyzed. Instead, patients are reported to be impaired in organizing sequential movements temporally, exerting rhythm control, or coordinating the limbs bilaterally. In nonhuman primates, the effects of ablation of the SMA have been reported to be subtle, unlike the effects of M1 lesions. No gross changes in general motor behavior are apparent, and effort is required to specify the nature of the deficits in motor behavior after ablation. The lesion effects include the inappropriate use of both hands in retrieving food and impairment in selecting between movements or changing between movements without external instructions. Transient inactivation of the SMA by the local injection of an inhibitory substance also demonstrated that the effects are subtle. If movements are triggered by visual signals, the reaction time is not altered. In contrast, SMA inactivation leads to impairment in arranging multiple movements in the correct order on the basis of memorized information.

Functional Brain Imaging Studies in Humans

Functional imaging of the human brain has provided evidence that the medial frontal cortex is involved in multiple aspects of motor control. Early cerebral blood flow studies emphasized differences in the participation of the medial frontal cortex and M1 in various aspects of motor control. Subsequent studies revealed the need to delineate separate areas in the medial wall. The level of the anterior commissure (the VCA line) is used as the border between the SMA and pre-SMA. The SMA, located caudally, has been found to be active in association with relatively simple motor tasks, such as key pressing, joystick manipulation, tracking, and fixed-direction finger movements, as well as in sequential finger movements. Imaging studies have also hinted at somatotopy in the SMA. In contrast, the pre-SMA is active during the performance of relatively complex motor tasks, such as sequential key pressing, conditional pointing movements, and playing a musical score. During speech tasks, the pre-SMA is activated more often during demanding verbal tasks, whereas the SMA is active during the simple repetition of words. The acquisition of new motor skills leads to more activation of either the SMA or the pre-SMA.

Functional Role of the SMA Suggested from the Analysis of Cell Activity

Cells in the SMA are often active when one is initiating or planning to initiate limb movements, regardless of whether the movements are triggered or self-initiated. Activity is observed when an individual is initiating either simple or complex organized movements. These observations led to the view that the SMA is similarly involved in the initiation or preparation of limb movements. However, just because cells in both the SMA and M1 fire on initiation of limb movements, it does not necessarily follow that the functional meaning of the cellular activity is similar. To address this issue, researchers trained monkeys to select three different distal movements: to press a key with the digits of the right or left hand alone or both together. Cells in the M1 are typically active in association with contralateral hand movements. Therefore, the activity accompanying a right-hand key press or bilateral (simultaneous) key press was similar. However, a majority of SMA cells behaved differently. SMA cells were typically active before the initiation of the right-hand key press only (and not before the bilateral key press), or the left-hand key press only, or exclusively when both hands pressed the key together (Figure 2). This implies that the SMA cells are involved in selecting which hand to use rather than selecting muscles in the contralateral

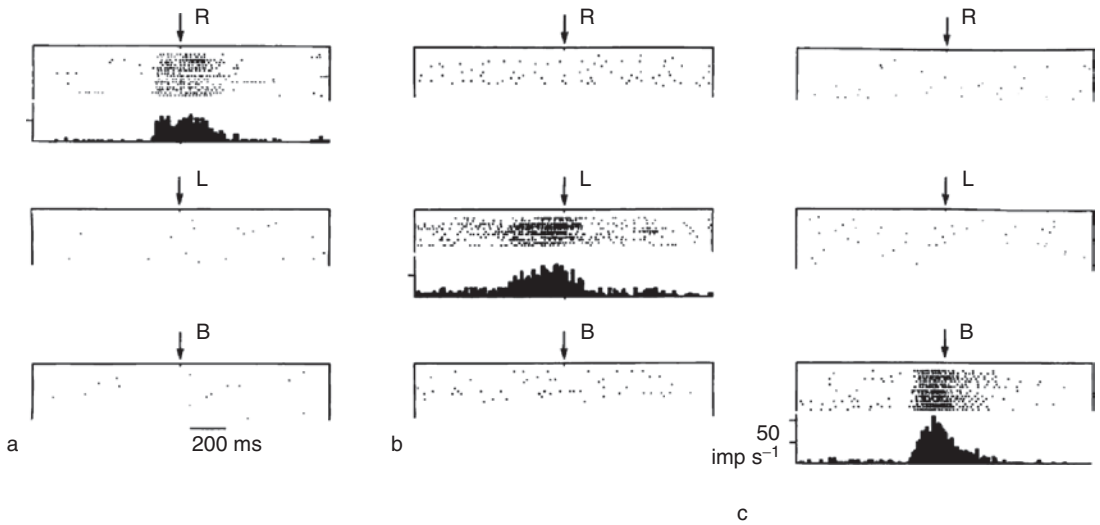

Figure 2 Three examples of activity in supplementary motor area neurons exhibiting specific relationships to the use of digits in the right, left, or both hands. In the raster displays, dots represent individual discharges of a single cell. The discharges are aligned at the onset of the key press movement (↓). In the histograms below, the discharges are summed across trials. The neuron shown in (a) was active only when the animal pressed the key with digits in the right hand, but not with the left hand or both hands together. The neurons in (b) and (c) were exclusively active when the animals used the left hand and both hands together, respectively. From Tanji J, Okano K, and Sato KC (1988) Neuronal activity in cortical motor areas related to ipsilateral, contralateral, and bilateral digit movements of the monkey. *Journal of Neurophysiology* 60: 325–343.

limb (as is applicable for M1 cells). It is important to note that an observation of this sort is possible only when experiments are designed rigorously to pursue the aim of the investigation. The proposal that the SMA plays a role in selecting body parts to be used in intended actions was substantiated in a recent report. A separate report proposed a role of the SMA in the temporal organization of multiple movements; SMA cells were found to take part in retrieving the sequences of multiple movements and in temporally connecting one movement to another.

Pre-SMA

It is now established that the pre-SMA, which corresponds to cytoarchitectonic area 6aβ of Vogt, is a distinct entity separate from the SMA, which corresponds to area 6aα. Although these two adjacent areas were originally treated as a single area named the SMA in the old terminology, an anatomical and physiological basis exists for regarding these areas as separate. The pre-SMA, but not the SMA, receives substantial input from the prefrontal cortex. The SMA, but not the pre-SMA, sends output to the M1. In addition, the input–output relationships with other cortical areas, the thalamus, and the basal ganglia differ. Intracortical microstimulation of the pre-SMA gives rise to limb movements. However, the stimulation effects are more complex and variable than those observed in the SMA. Positive stimulus effects require stronger currents or a long pulse duration.

Brain Imaging Studies in Humans

Although considerable doubts remain whether the pre-SMA and SMA have been delineated convincingly using brain imaging techniques that have limited spatial resolution, some indications that the pre-SMA participates in aspects of behavioral control have emerged from numerous reports. First, concerning the control of motor behavior, this region seems to be effector-independent because similar regions are activated in manual motor and oculomotor tasks. Second, the pre-SMA is often active under conditions that activate the lateral prefrontal cortex, such as during working memory tasks, attentional shift tasks, and conditional association tasks. These reports indicate that the pre-SMA is under powerful control of inputs from the lateral prefrontal cortex. More recent studies using functional magnetic resonance imaging (fMRI) revealed activity focused in the pre-SMA that was attributable to task switching or attention to intention.

Pre-SMA Studied by Analyzing Single-Cell Activity

The first attempt to analyze pre-SMA cells found that the activity was much less strongly associated with movement initiation than that of SMA cells. This means that the pre-SMA is not very involved in specifying movements themselves. Instead, responses to visual signals and delay-period activity reflecting visual instructions were dominant. Subsequent studies revealed the role of the pre-SMA in regulating the

temporal sequences of multiple actions. In a behavioral task requiring control of the temporal sequence of multiple arm movements, pre-SMA cells were found to reflect the ordinal number in the task sequence (i.e., the first, second, or third movement; see Figure 3), as well as the order of different movements (e.g., push, pull, and turn a handle, in that order). It is interesting that the activity of pre-SMA cells also reflects the temporal structure of the oculomotor sequence during planning. This finding is consistent with the view that the pre-SMA is involved in organizing the temporal structure of forthcoming actions irrespective of the effectors (body parts) to be used. A recent study also found that the pre-SMA is not very involved in selecting the limb to be used. Instead, pre-SMA cells were found to participate in specifying a motor target in space, in a manner commonly found in cells in the lateral prefrontal cortex. Another dominant factor characterizing the activity of SMA cells is the transitional process involved in changing from one behavioral routine to another. In the first study of this, pre-SMA cells were found to be particularly active when a change occurred in the plan to reach toward a target. In a subsequent study, monkeys were trained to perform three different movements sequentially in a temporal order. The correct order was planned based on visual information before it was executed. A population of pre-SMA cells was active during the process when monkeys were required to discard the current motor plan and develop a plan appropriate for the next sequential movements. Such activity was not common in the SMA and not found in the primary motor cortex. This suggests a role of pre-SMA cells in updating motor plans for subsequent temporally ordered movements. These findings, together with the aforementioned fMRI reports, indicate a role of the pre-SMA in selecting motor behavior to conform to alterations in behavioral demands.

CMAs

In the dorsal and ventral banks of the cingulate sulcus, three distinct motor areas have been identified: the rostral CMA (CMAr) in the banks of the rostral part of the cingulate sulcus (corresponding to area 24c), the dorsal CMA (CMAd) in the dorsal bank (area 6c) caudal to the CMAr, and the ventral CMA (CMAv) in the ventral bank (area 23c) facing the CMAd. These areas all project to brain structures involved in the control of movements, such as the primary motor cortex and motor centers in the brain stem and spinal cord. This section deals with the CMAs in the human brain separately because they are often named in a manner different from traditional nomenclature.

CMAr of Nonhuman Primates

The CMAr has interconnections with limbic areas in the brain, such as the rostral and caudal parts of the cingulate gyrus, parahippocampal gyrus, and perirhinal cortex. It is also interconnected with the prefrontal cortex, premotor cortex in the lateral surface of the frontal cortex, and primary motor cortex. These wide-ranging connections suggest that the CMAr both integrates diverse sets of limbic and cognitive information and sends its output signals to motor areas. In line with the anatomical connectivity placing the CMAr at a pivotal position between the limbic and motor centers, a variety of behavioral roles have been inferred, on the basis primarily of the analysis of neuronal activity in this area. Main aspects of such behavioral factors are summarized below.

Motor control The CMAr is viewed as a motor area because this area projects to the M1, intracortical microstimulation evokes body movements, and neurons exhibit changes in activity during the execution and preparation of limb movements. It is important to note, however, that the premovement or preparatory activity, although profound, does not reflect motor parameters such as the location of motor targets or body parts to use. Therefore, the participation of the CMAr in motor control should differ in nature from that of the M1. It is conjectured that the long-lasting and prominent preparatory activity represents a general intention for action.

Processing of reward information CMAr neurons change their activity in response to reward delivery or the omission of an expected reward, consistent with the anatomical connectivity of this area with limbic structures. Furthermore, activities of the CMAr neurons reflect the expectancy of, or proximity to, a forthcoming reward. These findings suggest that the CMAr is involved in information processing relevant to a reward in the context of behavioral planning or the motivational modulation of behavior.

Performance monitoring Neurons in the CMAr are active in close temporal relation to the occurrence of a variety of behavioral events, including sensory signals, instruction cues, motor performance, and the outcomes of motor responses (correct or error). Neurons are also active in differentiating exploratory or

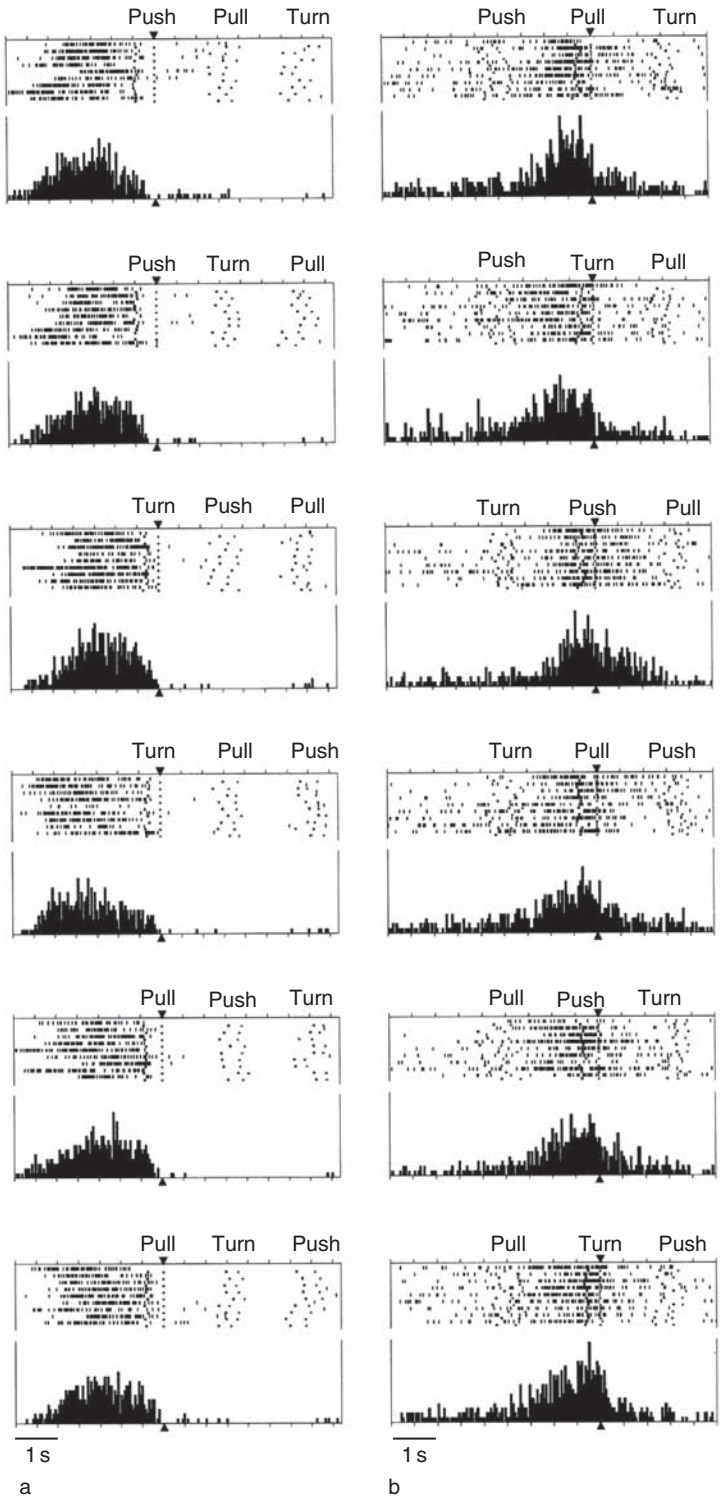

Figure 3 Two examples of neuronal activity during the execution of sequential arm movements, exhibiting rank order selectivity. (a) Discharges of a presupplementary motor area (pre-SMA) neuron that was most prominently active during the preparatory period before initiating the first movement. (b) In this pre-SMA neuron, activity was most prominent during the preparatory period before the second movement, irrespective of the sequence of movements. ▼, the start of the movement with which the data are aligned. From Shima K and Tanji J (2000) Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology* 84: 2148–2160.

repetitive selection tasks, or behavioral sequences. These reports suggest that the CMAr is involved in monitoring the performance of ongoing behavior.

Reward-based selection of action To investigate the role of the CMAr in aspects of behavioral control based on reward information, Shima and Tanji trained monkeys to select and perform two different arm movements triggered by a visual signal. The monkeys selected the motor task themselves, based on the amount of reward. During a series of constant-reward trials, monkeys had to keep selecting a particular movement. If they detected a reduction in the reward, they chose to perform the alternate movement because if they did, the reward returned to its full amount. Muscimol injection revealed that CMAr inactivation led to the impairment of proper selection based on the reward amount. Furthermore, CMAr cells were found to exhibit changes in activity during the interval between the detection of the reduced reward and the initiation of a new movement (Figure 4). A group of cells responded shortly after the occurrence of the reduced reward, and others were active progressively before selecting the alternate movement. These findings support the view that CMAr cells take part in selecting an action based on the reward obtained as a result of the action. A subsequent study examined the effect of ablation of the CMAr and its adjacent region in monkeys performing a reward-conditional response selection task. The results of this study led the authors to conclude that the CMAr is concerned with action–reward associations and is not limited to just detecting when actions lead to errors, and that the role of the CMAr is in action–reinforcer associations but not stimulus–reward associations.

Caudal CMAs of Nonhuman Primates

The CMAd and CMAv have dense connections with the primary motor cortex and spinal cord. Intracortical microstimulation of these areas evokes limb movements with less current than required for activating the CMAr. Neurons in these areas show distinct activity changes in association with the execution of movements, with apparent selectivity for such motor parameters as the direction of arm movements. Neuronal activity specifying the precision grip by the fingers has also been reported. These findings suggest that the caudal CMAs are much more involved in movement control than the CMAr is. However, more studies are necessary to determine the specific functional roles of these areas. Although the differences in the cytoarchitecture and anatomical connectivity of the CMAd and CMAv suggest the functional specificity of these two areas, no reports have validated this contention.

CMAs in the Human Brain

In the human brain, the complexity of the cingulate sulcus precludes straightforward delineation of the CMAs. A currently prevailing view holds that the CMAs contain two major regions: the rostral cingulate zone (RCZ) and the caudal cingulate zone (CCZ). The RCZ is further divided into anterior (RCZa) and posterior (RCZp) subdivisions. Functional brain imaging studies found that the CCZ is activated primarily in relation to movement execution and that the RCZ is involved more in willed control of action, suppressing unwanted and facilitating wanted actions. The RCZ occupies a portion of an area often called the anterior cingulate cortex, which refers to a broad, ill-defined area in the anterior, inferior part of the mesial cortex. One current hypothesis emphasizes the role of RCZa in conflict monitoring and the role of RCZp in action selection.

Supplementary Eye Field

The dorsomedial frontal cortex rostral and lateral to the pre-SMA contains an area in which intracortical microstimulation evokes eye movements. This area, located inside Brodmann's area 6, is called the supplementary eye field (SEF). The SEF is interconnected with structures controlling eye movements, such as the superior colliculus in the brain stem and the frontal eye field in the frontal cortex, providing the basis for its involvement in oculomotor control. Moreover, the SEF is interconnected with the lateral prefrontal cortex and pre-SMA, suggesting the participation of this area in cognitive aspects of oculomotor behavior. Aspects of the functional roles played by this area, listed below, have been revealed primarily by studies in nonhuman primates. However, clinical studies and functional brain imaging studies in humans have confirmed that the basic principles of its morphological organization and functional properties are similar in humans and nonhuman primates.

Oculomotor Control

As commonly found in oculomotor areas, electrical stimulation of the SEF with low currents evokes eye movements. Detailed analyses have revealed that the properties of the evoked saccades are complex and vary greatly depending on the behavioral conditions. Evoked saccades could have either constant-vector or converging properties, in which the eyes tend to converge toward a certain goal in space or in the orbit.

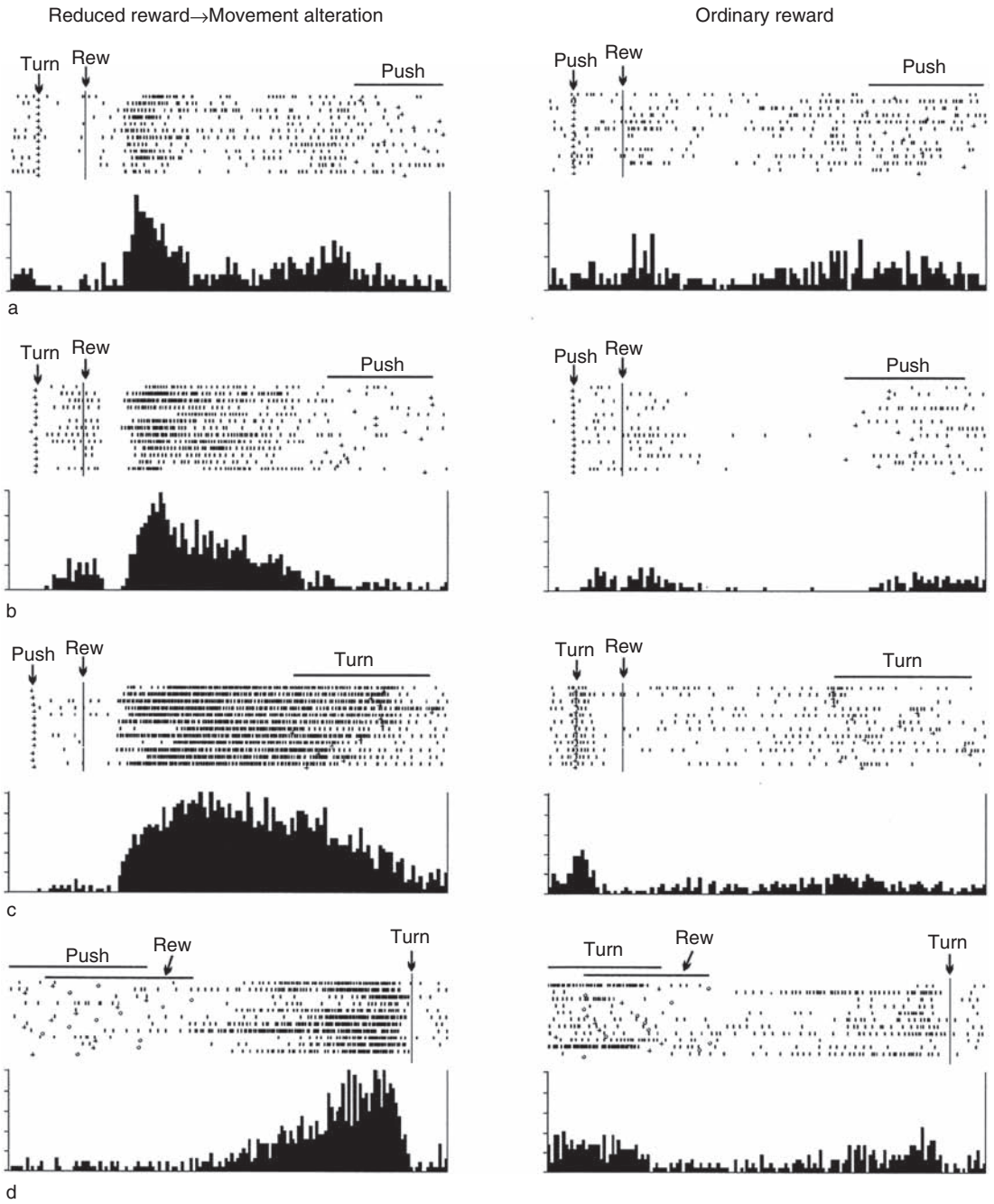

Figure 4 Left panels: The activity of four types of rostral cingulate motor area cells (a–d) showing increased discharges after the reward (Rew) was reduced and before a newly selected movement was initiated. Each raster corresponds to a single trial when the animal switched from Turn to Push (a, b) or Push to Turn (c, d), with 5 to 12 trials in between. (a) Short-lasting activity after the reward. (b) Long-lasting, but decaying, activity. (c) Continuous activity. (d) Activity that increases progressively before the next movement. Right panels: The same four cells were not as active under the constant reward condition when the animals did not select an alternative movement. In the raster displays, small crosses denote the onset of movement. In the histograms, discharges over 11 trials are aligned with the onset of the reward and summed, except for (d), which is aligned with the onset of the next movement. The horizontal bar in the top right of a raster display indicates the time range of occurrences of the movement. From Shima K and Tanji J (1998) Role for cingulate motor area cells in voluntary movement selection based on reward. *Science* 282: 1335–1338.

It is interesting that if a stimulus is applied during an oculomotor task, the properties of the evoked saccades vary in a task-dependent manner. For example, during the period of saccade preparation, electrical stimulation can trigger saccades in the direction of a target that an animal is intending to capture in advance. An analysis of neuronal activity has also reflected the complexity of the manner in which this

area is involved in a variety of oculomotor tasks. SEF neurons typically discharge before or after the initiation of saccades, or in repose to the appearance of visual signals. Some neurons also show a buildup of activity before a spontaneous/self-paced saccade. Usually, these neurons have a preferential response field in space. It is important to note that the properties of presaccade activity could vary depending on such behavioral factors as attention, reward condition, and task training. Furthermore, an interesting hypothesis has been proposed concerning the reference frame controlling saccade direction. Some SEF neurons were selectively active preceding saccades capturing the right (or left) end of an object, suggesting that the SEF tunes saccade direction using an object-centered, spatial frame of reference.

Monitoring Oculomotor Behavior

As stated above, the activity of SEF neurons is modulated in relation to diverse aspects of oculomotor behavior. A possible interpretation of the SEF neuronal activity is monitoring, rather than generating, oculomotor behavior. For example, when monkeys are engaged in an eye movement-countermanding task that requires withholding a saccade immediately before its prepared initiation, distinct groups of neurons were active after errors, after successful withholding, and in association with reinforcement. These results suggest that the SEF plays a role in monitoring the performance of oculomotor behavior.

Control of Sequential Oculomotor Performance

Electrical stimulation of the SEF affects the performance of successive saccades that are made in conforming to two successive target flashes. The activity of SEF neurons reflects either the sequence or rank order of multiple saccades performed in predetermined sequences. These findings suggest that the SEF is involved in the executive control of sequential saccades, as well as in the cognitive control of sequential oculomotor behavior.

Summary and Conclusions

The motor areas in the medial frontal cortex are key structures constituting either a somatomotor or an oculomotor network in the cortex. With the interconnections depicted in Figure 5, each of the medial premotor areas plays a crucial role in processing information provided by perceptual or memory systems, providing a basis for generating output signals that are sent primarily from the M1 or frontal eye field. Obviously, broader anatomical networks including other cortical and subcortical areas, such as the parietal cortex, basal ganglia, and cerebellum, enable the medial premotor areas to play their part in each aspect of organizing motor behavior.

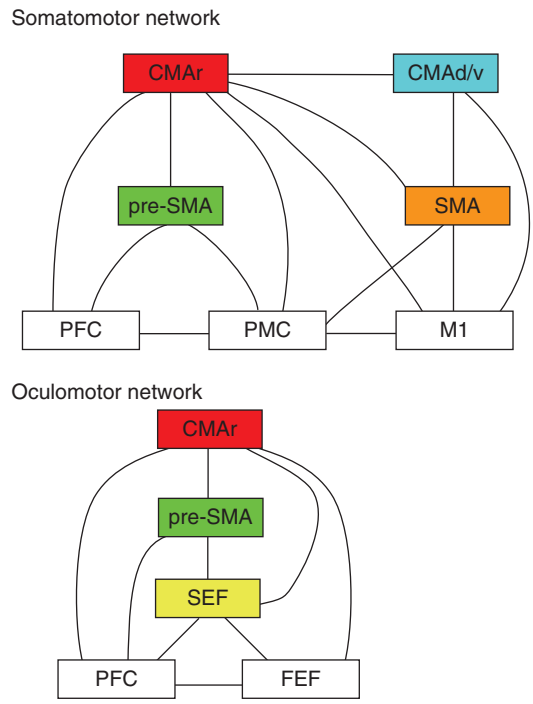

Figure 5 Diagrams representing the cortico-cortical networks involving the medial premotor areas in the frontal cortex. The top panel illustrates a network processing information necessary for generating output signals controlling body movements (somatomotor network). The bottom panel illustrates a network for oculomotor behavioral control before generating eye movements (oculomotor network). CMAd, dorsal cingulate motor area; CMAr, rostral CMA; CMAv, ventral CMA; FEF, frontal eye field; M1, the primary motor cortex; PFC, prefrontal cortex; PMC, premotor cortex; pre-SMA, presupplementary motor area; SEF, supplementary eye field; SMA, supplementary motor area.

See also: Attention and Eye Movements; Cerebellum and Oculomotor Control; Cortical Control of Eye Movements; Frontal Eye Fields; Oculomotor System: Models; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Premotor Cortex in Primates: Dorsal and Ventral; Saccadic Eye Movements; Supplementary Eye Fields.

Further Reading

Brinkman C (1984) Supplementary motor area of the monkey's cerebral cortex: Short- and long-term deficits after unilateral ablation and the effects of subsequent callosal section. *Journal of Neuroscience* 4: 918–929.

Chen YC, Thaler D, Nixon PD, Stern CE, and Passingham RE (1995) The functions of the medial premotor cortex. II. The timing and selection of learned movements. *Experimental Brain Research* 102: 461–473.

Dum RP and Strick PL (2005) Motor areas in the frontal lobe: The anatomical substrate for the central control of movement.

In: Riehle A and Vaadia E (eds.) *Motor Cortex in Voluntary Movements: A Distributed System for Distributed Functions*, pp. 3–47. Boca Raton, FL: CRC Press.

Hadland KA, Rushworth MF, Gaffan D, and Passingham RE (2003) The anterior cingulate and reward-guided selection of actions. *Journal of Neurophysiology* 89: 1161–1164.

Laplane D, Talairach J, Meininger V, Bancaud J, and Orgogozo JM (1977) Clinical consequences of corticectomies involving the supplementary motor area in man. *Journal of the Neurological Sciences* 34: 301–314.

Luppino G, Matelli M, Camarda R, and Rizzolatti G (1993) Corticocortical connections of area F3 (SMA-proper) and area F6 (pre-SMA) in the macaque monkey. *Journal of Comparative Neurology* 338: 114–140.

Matsuzaka Y, Aizawa H, and Tanji J (1992) A motor area rostral to the supplementary motor area (presupplementary motor area) in the monkey: Neuronal activity during a learned motor task. *Journal of Neurophysiology* 68: 653–662.

Olson CR (2003) Brain representation of object-centered space in monkeys and humans. *Annual Review of Neuroscience* 26: 331–354.

Paus T (2001) Primate anterior cingulate cortex: Where motor control, drive and cognition interface. *Nature Reviews Neuroscience* 2: 417–424.

Penfield W and Welch K (1951) The supplementary motor area of the cerebral cortex: A clinical and experimental study. *A. M. A. Archives of Neurology and Psychiatry* 66: 289–317.

Picard N and Strick PL (2001) Imaging the premotor areas. *Current Opinion in Neurobiology* 11: 663–672.

Procyk E, Tanaka YL, and Joseph JP (2000) Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nature Neuroscience* 3: 502–508.

Schlag J and Schlag RM (1987) Evidence for a supplementary eye field. *Journal of Neurophysiology* 57: 179–200.

Shidara M and Richmond BJ (2002) Anterior cingulate: Single neuronal signals related to degree of reward expectancy. *Science* 296: 1709–1711.

Shima K and Tanji J (1998) Role for cingulate motor area cells in voluntary movement selection based on reward. *Science* 282: 1335–1338.

Shima K and Tanji J (2000) Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology* 84: 2148–2160.

Stuphorn V, Taylor TL, and Schall JD (2000) Performance monitoring by the supplementary eye field. *Nature* 408: 857–860.

Tanji J (1996) New concepts of the supplementary motor area. *Current Opinion in Neurobiology* 6: 782–787.

Tanji J (2001) Sequential organization of multiple movements: Involvement of cortical motor areas. *Annual Review of Neuroscience* 24: 631–651.

Tanji J, Okano K, and Sato KC (1988) Neuronal activity in cortical motor areas related to ipsilateral, contralateral, and bilateral digit movements of the monkey. *Journal of Neurophysiology* 60: 325–343.

Premotor Cortex in Primates: Dorsal and Ventral

G Rizzolatti and M Fabbri-Destro, Universitá di Parma, Parma, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In primates, the motor cortex occupies the caudal part of the frontal lobe. Histologically, it is characterized by a complete, or almost complete, lack of granular cells ('agranular frontal cortex'). According to Brodmann, the motor cortex consists of two main cytoarchitectonic areas: area 4, or primary motor cortex, and area 6, or premotor cortex. Modern anatomical techniques and functional data have shown that this subdivision is inadequate. Today, there is general agreement that the agranular frontal cortex of primates is composed of seven functionally distinct motor areas. Their locations are shown in Figure 1.

Of the seven areas forming the motor cortex, two are located on the mesial surface (area F3, or supplementary motor area, and area F6, or presupplementary motor area) and four (F2, F4, F5, and F7) on the lateral convexity. F4 and F5 are often collectively indicated as the ventral premotor cortex, whereas F2 and F7 are referred to as the dorsal premotor cortex. In this article, we review the properties of the four areas located on the lateral cortical surface.

Area F5 (Rostral Ventral Premotor Cortex)

General Properties

Area F5 forms the rostral part of the ventral premotor cortex. Electrical stimulation studies reveal that area F5 contains a movement representation of the hand and the mouth. The two representations overlap to a considerable extent.

Single neuron recordings showed that most F5 neurons code specific motor acts – that is, movements with a specific goal (e.g., grasping) – rather than the individual movements forming the acts (e.g., finger movements) (Figure 2). Using the effective motor act as the classification criterion, F5 neurons have been subdivided into various classes, such as grasping (the most represented), holding, tearing, and manipulating. In each class, many neurons code specific types of hand shaping, such as precision grip, whole hand prehension, and finger prehension. Finally, whereas some neurons discharge during the whole action (e.g., opening and closing of the hand), others fire only during a certain part of the action (e.g., the last part of the grasping).

It therefore appears that F5 contains a storage ('vocabulary') of motor acts. This motor vocabulary is composed of 'words,' each of which is represented by a population of F5 neurons. Some words code the general goal of a motor act (e.g., grasping and holding). Others code how, within a general goal, a specific motor act must be executed (e.g., precision grip). Finally, other words specify the temporal aspects of the action to be executed (e.g., opening of the hand, closing it, and holding).

Approximately 20% of F5 neurons respond to visual stimuli. F5 visuomotor neurons are grouped into two classes: 'canonical' neurons and 'mirror' neurons. Canonical neurons discharge to the presentation of three-dimensional (3-D) objects. Mirror neurons discharge when the monkey observes another individual performing an action similar to that encoded by those neurons.

Canonical Neurons

The functional properties of canonical neurons have been studied using a behavioral paradigm that separates activity elicited by object presentation from that related to action execution (Figure 3). The results showed that the majority of canonical neurons respond selectively to the presentation of objects of a certain size, shape, and orientation. Typically, congruence was found between the object's physical characteristics that activate a given neuron and its motor properties.

On the basis of these results and control experiments, which ruled out the possibility that the object-triggered neuron discharges were due to motor preparation, attention, or intention to act, it was concluded that the discharge of canonical neurons, in the absence of movement, represents a potential motor act. This potential motor act may be executed or not, according to the external contingencies and intentions.

Canonical circuit Area F5 receives its main parietal input from the rostral part of the inferior parietal lobule (IPL) and in particular from the anterior intraparietal area (AIP; Figure 1). AIP neurons have been subdivided into three main classes: motor-dominant neurons, visual-dominant neurons, and visual-and-motor neurons. Motor-dominant neurons discharge during grasping in both light and dark. They are silent during object fixation. Visual-dominant neurons discharge during grasping in light and during object fixation but not during grasping in dark. Visual-and-motor neurons discharge both during grasping in light and grasping in dark. They also discharge during object fixation. Most

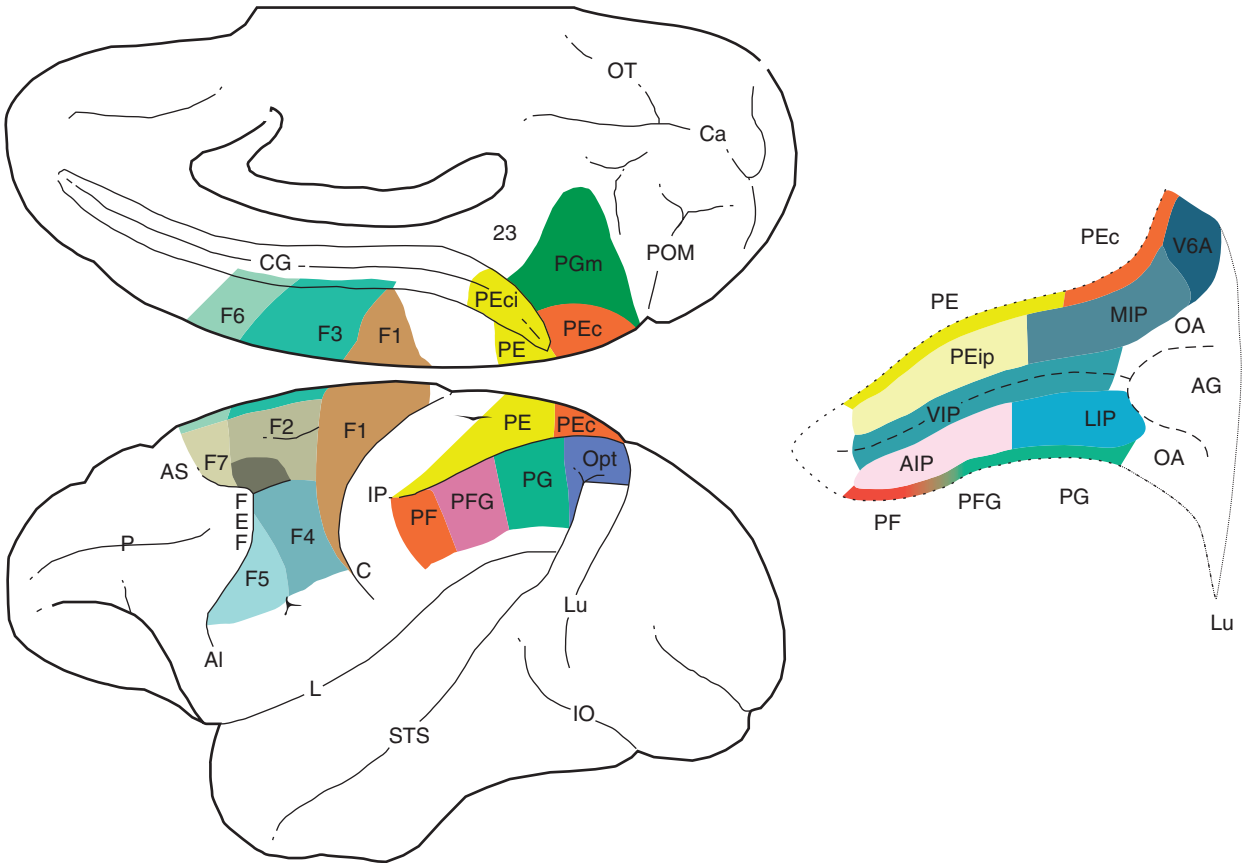

Figure 1 Mesial and lateral views of the macaque brain showing the cytoarchitectonic parcellation of the frontal motor cortex (areas indicated with 'F' and numbers) and of the parietal lobe (areas indicated with 'P' and progressive letters). Areas buried within the intraparietal sulcus are shown in an unfolded view of the sulcus. AI, inferior arcuate sulcus; AIP, anterior intraparietal area; AS, superior arcuate sulcus; C, central sulcus; Ca, calcarine fissure; CG, cingulate cortex; FEF, frontal eye field; IP, intraparietal sulcus; L, lateral sulcus; LIP, lateral intraparietal area; Lu, lunate sulcus; MIP, medial intraparietal area; P, principal sulcus; POS, parieto-occipital sulcus; STS, superior temporal sulcus.

AIP neurons, regardless of the class, respond preferentially or even selectively to the presentation of 3-D objects of a certain size and shape.

The neural properties of AIP and F5 neurons suggest that these two areas, which are richly connected with each other, play an important role in visuomotor transformations for object grasping. Strong evidence supporting this hypothesis derives from inactivation studies of AIP and F5 with a GABA agonist. Inactivation of each of these areas produces a dramatic deficit in the hand-shaping necessary to grasp objects.

The grasping circuit: Functional interpretations How do AIP and F5 interact? The properties of parietal and premotor grasping neurons, previously summarized, enable the proposal of a theoretical model of how visual object representations are transformed into hand movements appropriate for grasping those objects. As previously discussed, a fundamental property of AIP visual-dominant and visual-and-motor neurons is that they respond to the presentation of 3-D objects. The functional meaning of these responses can be easily accounted for if one considers the notion of affordance introduced by James Gibson. According to Gibson, the observation of an object entails an immediate selection of those of its properties that allow one to interact with it. These properties are not the figural properties of the object but, rather, the pragmatic opportunities that it offers to the observer.

It is well known that the visual stimuli features necessary for acting on objects are extracted in the dorsal visual stream. AIP visual-dominant neurons and visual-and-motor neurons represent the stage in object processing in which, following dorsal stream elaboration, object affordances become neuron triggering features. Information on affordances is then sent to F5 neurons, which code potential motor acts. Because of the congruence between F5 neuron responses elicited by a specific object affordance and the motor act they code, a given object affordance

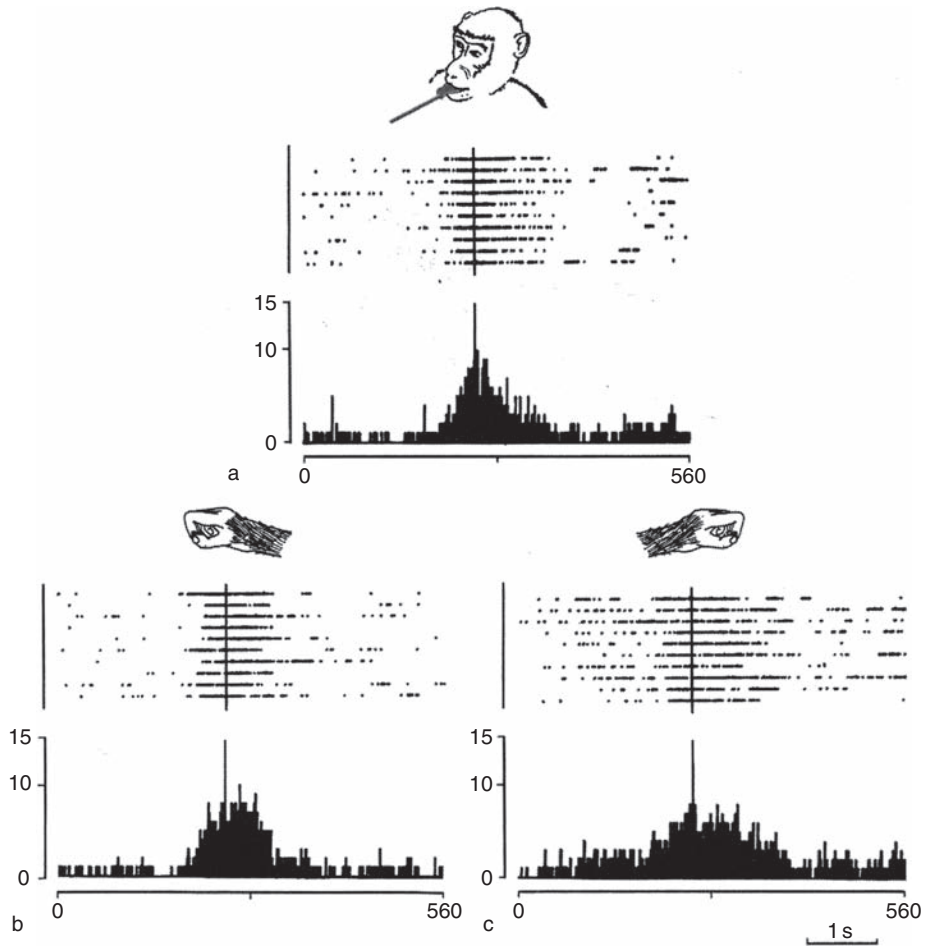

Figure 2 Example of a 'grasping' F5 neuron. (a) Neurons discharge during grasping with mouth. (b) Neurons discharge during grasping with the right hand. (c) Neurons discharge during grasping with the left hand. Rasters and histograms are aligned with the moment at which the monkey touched the food. Horizontal scale =1 s. Vertical scale = spikes/bin. Bin width = 10 ms.

is transformed into the appropriate potential motor act. Object becomes action.

The model discussed up to now takes into consideration a very simple situation, namely that each object has only one affordance. In real life, however, objects, especially those that humans use, have typically more than one affordance and may be grasped in several ways. How can this problem be solved?

Behavioral analysis of prehension showed that when objects possess several affordances, other nonvisual factors determine the selection of the affordances and ultimately how an object will be grasped. These factors are the object semantics (what is that object) and what the individual, who acts on the object, wants to do with it at that specific moment. A simple example will clarify these points. Imagine a cup. The cup has three major affordances: the handle, the top, and the body. If the cup is recognized as a cup, and the agent wants to use it in the way cups are commonly used, he or she will grasp it by the handle. However, if one wants to move the cup, the cup will be taken not by the handle but by its body or its top (Figure 4).

Thus, a more complete model of the F5–AIP grasping circuit should assume that this circuit automatically extracts not one but all the affordances of a given object and that only one of them is then selected, according to the object semantics and the agent's intention on how to use the object.

The neural substrate for selection on the basis of object semantics is most likely represented by the connections between the inferotemporal lobe (i.e., the final stage of ventral visual stream – the 'what' circuit) and AIP, whereas the neural substrate for affordance selection related to the current goal of the agent is represented by the connections between AIP and the prefrontal lobe, the cortical region in which long-term action plans are elaborated. When necessary, the pre-

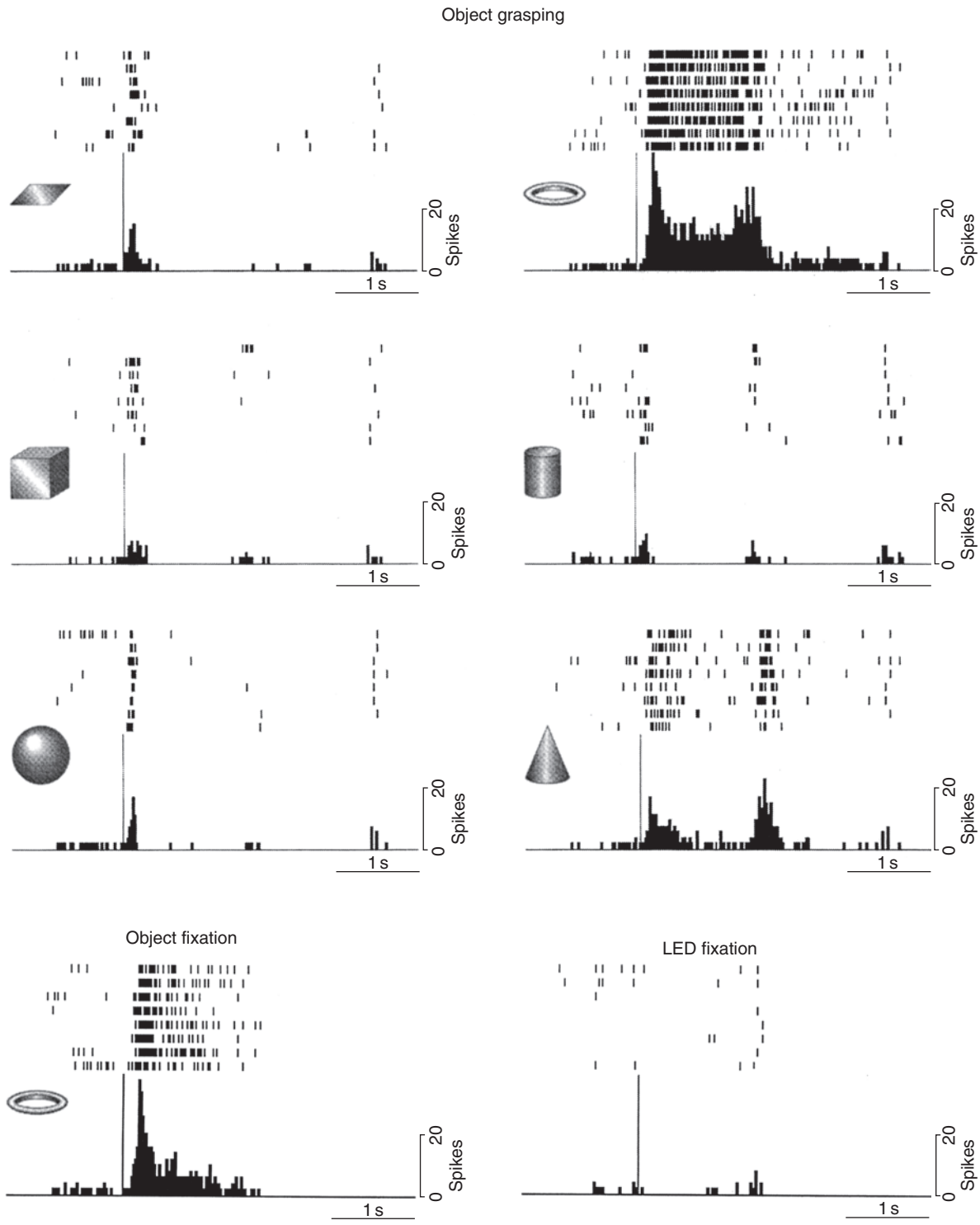

Figure 3 Example of an F5 canonical neuron. Panels show neural activity recorded during three different experimental conditions: grasping in light (object grasping), object fixation, and LED fixation. Note the neuron's response selectivity for the ring. Note also that the neuron has two response peaks, one at stimulus presentation and the second during grasping. Rasters and histograms are aligned (vertical bar) with key press (onset of object presentation). Horizontal scale = 1 s. Vertical scale = spikes/bin. Bin width = 20 ms.

frontal lobe input can override the selection of affordances based on the object standard use and can select those affordances that are congruent with the agent's intention at that moment.

Mirror Neurons

Mirror neurons are a particular class of visuomotor neurons that discharge both when the monkey does

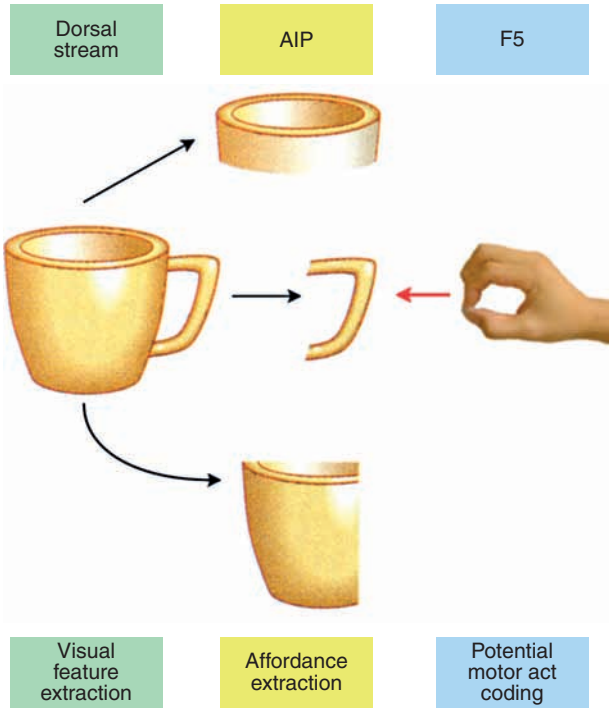

Figure 4 Schematic illustration of the steps involved in the transformation of an object (a cup) into its grasping. The neural centers involved in the process are indicated at the top of the figure, whereas the operations carried out by them are summarized at the bottom.

a particular action and when it observes another individual doing a similar action (Figure 5).

In order to be triggered by visual stimuli, mirror neurons require an interaction between an effector and an object. The sight of an object alone is ineffective. Mirror neurons show a large degree of visual generalization. Different visual stimuli representing the same action are equally effective. For example, the same grasping mirror neuron responding to a human hand grasping an object also responds when the grasping hand is that of a monkey. Similarly, the response is typically not affected if the action is done near or far from the monkey, despite the fact that the size of the observed hand is obviously different in the two conditions.

An important functional aspect of mirror neurons is the relation between their visual and motor properties. Virtually all mirror neurons show congruence between the motor responses they code and the observed action, done by another, to which they respond. Mirror neurons for which the effective observed and the effective executed actions correspond in terms of goal and means for reaching the goal have been termed 'strictly congruent.' Mirror neurons that do not require the observation of exactly the same action that they code motorically to be triggered have been termed 'broadly congruent.' They represent the majority of F5 mirror neurons.

Besides area F5, mirror neurons have also been described in the rostral part of the IPL (Figure 1). This region receives input from the cortex of the superior temporal sulcus (STS) and sends output to ventral premotor cortex including area F5. STS neurons respond to the observation of actions done by others but are not endowed with motor properties. Thus, the cortical mirror neuron circuit is formed by two main regions: the ventral premotor cortex and the rostral part of the inferior parietal lobule.

Functional role of the mirror neurons It is generally accepted that the basic functional role of mirror neurons is that of understanding actions done by others. How do mirror neurons mediate this function? The proposed mechanism is that each time an individual sees an action done by another individual, neurons which represent that action are activated in the observer's premotor cortex. This automatically induced motor representation corresponds to the one that is spontaneously generated during active action and the outcome of which is known to the acting individual. Thus, the mirror neuron system transforms visual information into knowledge.

This hypothesis was tested by experiments in which the monkey could not see the action done by another individual but had clues for understanding it. The rationale of the experiments was the following: If mirror neurons mediate action understanding, they should discharge even in the absence of action visual features.

On the basis of these considerations, two series of experiments were carried out. In the first, it was tested whether F5 mirror neurons were able to recognize actions from their sound. The results showed that mirror neurons that respond to the vision of a noisy action, such as ripping a piece of paper, also respond to its sound, without the vision of it. In the second experiment, grasping neurons were tested in two conditions. In one condition, the monkey saw a complete reaching-to-grasp action, and in the other a screen was placed between the monkey and the object to be grasped and the monkey saw only the reaching phase of it. The results showed that grasping mirror neurons started to discharge during the reaching and continued firing during the grasping, although the monkey could not see the hand–object interaction. This suggests that the observing monkey generated a mental representation of the action done behind the screen.

Correlation between Anatomical and Functional Organization of F5

F5 is not a homogeneous area, but cytoarchitectonically it consists of three sectors: area F5 convexity (F5c), the posterior part of the posterior bank of

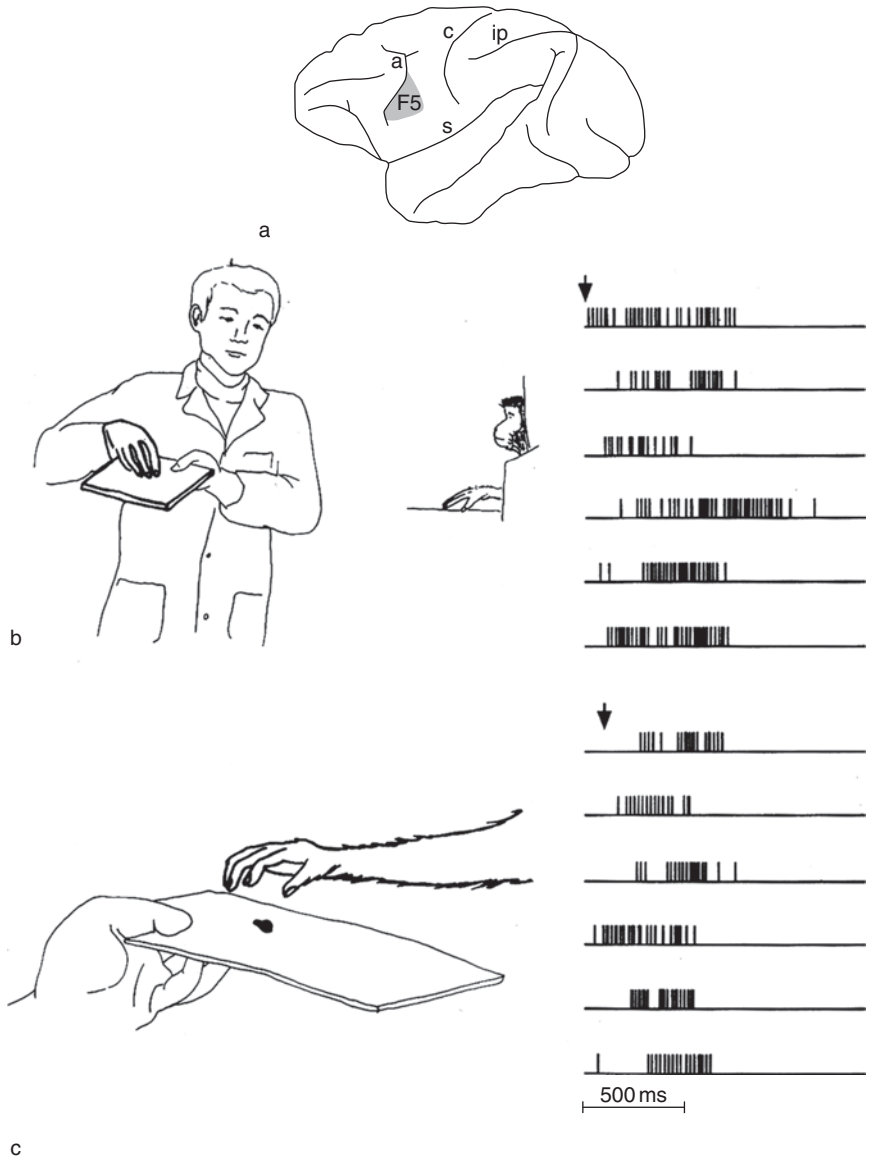

Figure 5 Example of an F5 mirror neuron selectively discharging during monkey grasping movements and during observation of a grasping movement done by the experimenter. (a) Lateral view of the brain with the location of F5 indicated.

the arcuate sulcus (F5p), and the anterior part of the same bank (F5a) (Figure 6).

Although no differences in the motor properties of the neurons of the different F5 sectors have been described, differences have been observed in the visual properties of neurons located in the various sectors. Mirror neurons are located mostly in F5c, whereas canonical neurons are frequently found in F5p. No neurophysiological data exist on F5a. A functional magnetic resonance imaging study of awake monkeys showed an interesting difference between F5c, on one side, and F5a on the other. Whereas a full view of the action's agent activated both these sectors of area F5, the vision of a grasping hand, without the vision of the agent, activated only F5a. These data suggest that the action coding done by the rostral part of F5 enables action understanding, even when only a part of its elements are present. The interpretation of the responses of the caudal part is more difficult. One possibility is that the use of information provided by the full view of the agent is used differently by the two F5 sectors. F5a uses it for action understanding, whereas F5c uses it for preparing future actions according to who is doing the action.

Area F4 (Caudal Ventral Premotor Cortex)

Area F4 forms the dorsocaudal part of the ventral premotor cortex. Its main parietal input originates from area VIP, an area located in the fundus of

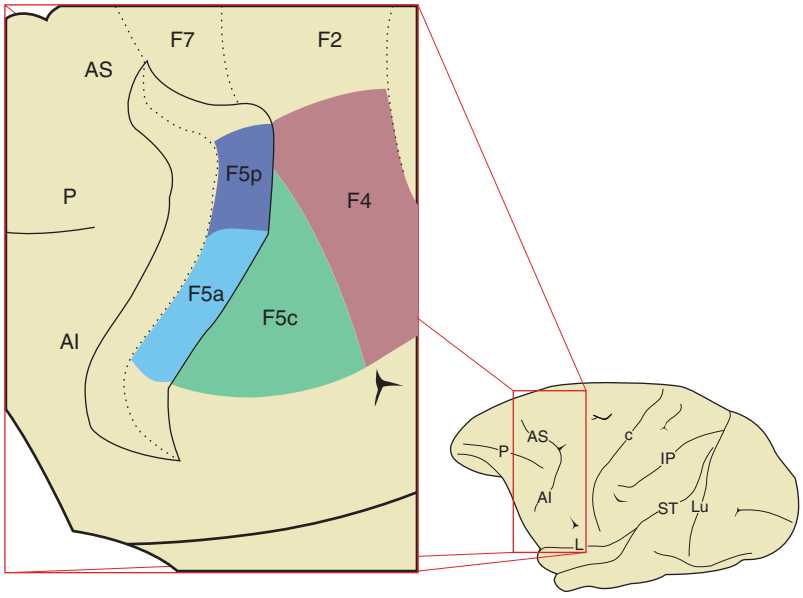

Figure 6 Areas forming monkey ventral premotor cortex (PMv). Area F5 consists of three different subareas: F5c, F5p, and F5a. The large rectangle showing PMv areas is an enlarged projection of the rectangle indicated on the lateral view of the monkey brain.

the intraparietal sulcus (Figure 1). Area VIP is the target of various visual areas of the dorsal stream. It also receives somatosensory input from the parietal cortex.

Electrical stimulation studies showed that F4 contains a representation of arm, neck, and face movements. Single neuron recordings confirmed these findings. They also showed that, similar to F5 neurons, F4 neurons do not discharge in association with elementary movements but, rather, in relation to the goal of the motor act defined, in this case, as the achievement of a spatial location.

Most F4 neurons respond to sensory stimuli. According to their effective stimulus, F4 neurons have been subdivided into somatosensory, bimodal (somatosensory and visual), and polymodal (somatosensory, visual, and auditory) neurons.

Somatosensory F4 neurons are typically activated by light tactile stimuli. Their receptive fields (RFs), which are generally large, are located on the face, chest, arms, and hands. Examples are shown in Figure 7. The preferred visual stimuli of F4 neurons are 3-D (especially moving) objects. These neurons respond when visual stimuli are presented in a specific space portion located around the monkey (spatial visual RF). In most F4 neurons, the position of the RF does not change when the gaze moves. This indicates that unlike the visual responses of visual cortical areas, which are coded in retinal coordinates, F4 visual responses are coded in spatial coordinates. This conclusion is confirmed by the observation that visual RFs, anchored to different body parts, do not change location when the body part on which the tactile RF is located is moved.

There are no systematic studies in which sensory and motor properties of F4 neurons have been investigated in the same neuron. In those rare neurons in which this relation was examined, a good congruence was found between the neuron's visual RF location and the motor field of that neuron.

The description of the properties of F4 neurons has strongly influenced the conceptualization of how space is represented in the brain. First, the presence of RFs anchored to the body demonstrated that space around the body ('peripersonal space') is coded in structures that are different from those coding space far from it ('extrapersonal space'). Second, the response properties of F4 neurons showed that spatially directed movements, such as reaching movements, are organized not according to a single reference system (e.g., head) but, rather, according to a multiplicity of reference systems anchored to different body parts. Finally, lesion experiments showed that damage to F4 results in profound deficits in space perception (spatial neglect). This indicates that action organization in space also plays an important role in space perception.

Area F2 (Caudal Dorsal Premotor Cortex)

Area F2 occupies the caudal part of the dorsal sector of area 6. F2 is not homogeneous in terms of its cortical connections. The part located around the superior precentral dimple (F2d) receives its major projections from the parietal areas PEc and PEip, two areas mostly involved in the processing of somatosensory stimuli. In contrast, the rostroventral part of F2 (F2vr; Figure 1) receives its major projections from

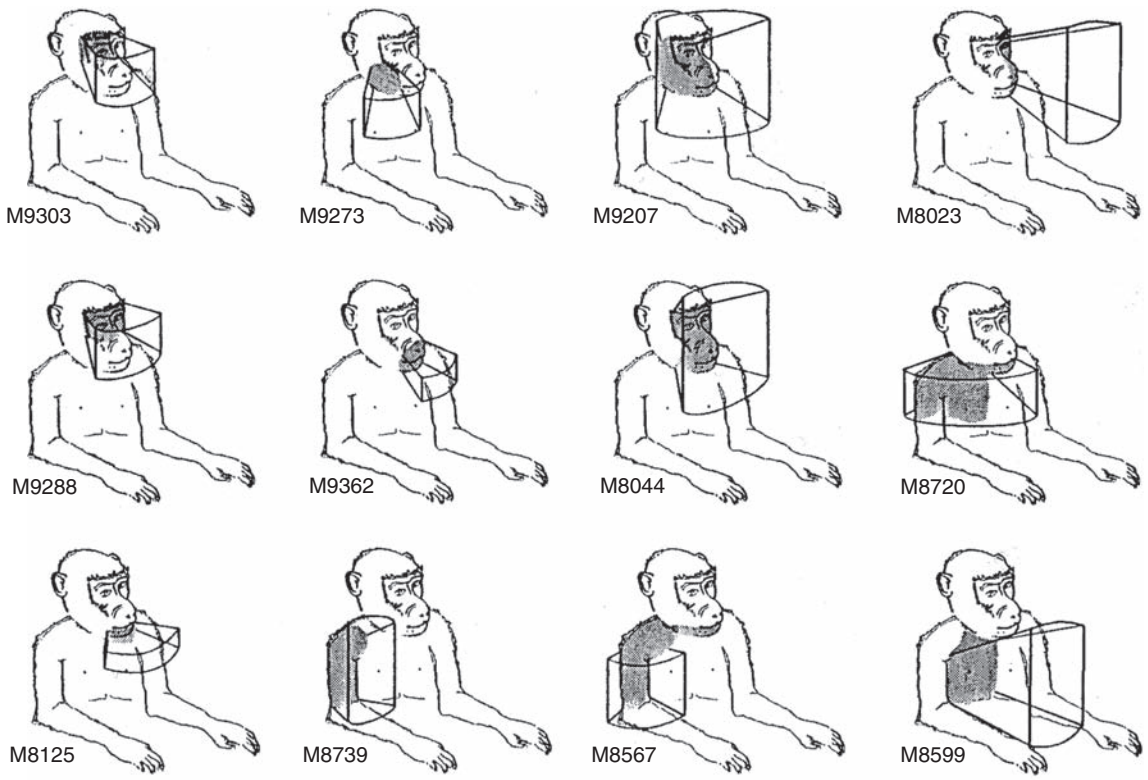

Figure 7 Examples of different types of tactile and visual receptive fields of F4 bimodal neurons. Gray shaded areas indicate tactile receptive fields. Solid lines around the tactile receptive fields indicate visual receptive fields.

areas MIP and V6A, two areas in which neurons have visual, somatosensory, or both visual and somatosensory responses. This anatomical organization has been confirmed by physiological data showing that whereas in F2d most neurons respond to proprioceptive stimuli, neurons in F2vr also respond to somatosensory as well as visual stimuli.

Electrical stimulation of F2 revealed that this area has a somatotopic organization, with a leg representation located dorsal to the superior precentral dimple and an arm representation located ventral to it. In addition, single neuron studies showed that within the arm field there is a particular sector where distal movements (hand and wrist) are represented. Some of these hand and wrist neurons respond to 3-D visual stimuli showing congruence between visual responses and motor properties, similar to those described for F5 neurons.

F2 neurons have been classified according to their activity before and during movements. Some F2 neurons discharge in association with movement onset (movement-related activity), some become active during the delay period when the monkey is waiting for a go signal (set-related activity), and others discharge at the appearance of the instruction stimulus (signal-related activity).

Motor properties of F2 neurons have been further investigated in two studies. The main result of the first study was that in F2vr there are a large number of neurons that code both the arm that the monkey uses to reach a target and the position in space that the monkey has to reach. In the other study, it was shown that F2 neurons respond when the monkey prepares to reach a target inside a restricted region of space, the neuron's reach field. This reach field is coded neither in retinal coordinates nor in hand coordinates, but it is a function of the spatial relationship between the target, the hand, and the eyes.

In summary, F2d appears to be involved in the control and preparation of arm and leg movement on the basis of somatosensory information. F2vr appears to have more complex properties and to be involved in the organization and control of arm movement in space.

A new dimension to the processing occurring in F2 has been identified in a study in which monkeys observed sensory events (a cursor moved on the screen) associated with a well-learned motor task. The results showed that most F2 task-related neurons exhibited the same activity pattern when the monkey actively performed the task and when

it merely observed the cursor motion. It was suggested that this F2 activity might be the neuron correlate of the monkey's internal rehearsal of the task.

Area F7 (Rostral Dorsal Premotor Cortex)

Area F7 is functionally heterogeneous. Its dorsal part contains the supplementary eye field (SEF). This field, which is richly interconnected with the frontal eye field (FEF), is primarily involved in the control of eye movements. The remaining part of F7 (ventral F7) receives its major input from the prefrontal lobe and specifically from dorsal area 46 and area 8b. A minor input to ventral F7 originates from parietal areas PGm and V6A. On the basis of this prominence of frontal connections with respect to the parietal ones, typical of the premotor areas, it has been suggested that F7 is a functional component of the prefrontal cortex rather than of premotor areas.

In accordance with this view are the results of electrical stimulation, which in contrast with F2, do not evoke consistent motor responses. Recording studies showed, however, that in F7 there are neurons that discharge in relation to arm movements. Some of them fire before movement onset; others are active also in response to visual stimuli even in the absence of a subsequent motor action.

In summary, the functions of this area, except for the part related to eye movements, are largely unknown. The available data suggest that F7 is involved in coding space and in planning actions according to information on locations of targets either present in the space around the animal or kept in memory.

Organization of the Premotor Cortex in Humans

There are no detailed physiological data on the functional properties of human premotor cortex comparable to those of the monkey. Important new findings, however, on the functional organization of human premotor areas have been obtained by brain imaging studies. These studies showed that all premotor areas are active during action execution. Some of them are also active during motor imagery and action observation.

We will not review all human studies in detail but, rather, we examine whether the general organization of the premotor region found in the monkey is maintained in humans. We first discuss the possible anatomical homologies between human and monkey premotor areas and then examine the human functional data within this anatomical framework.

Homologies between Human and Monkey Premotor Areas

The monkey lateral premotor cortex is formed by two major parts: the dorsal (PMd) and ventral (PMv) premotor cortex. The border between these two parts roughly corresponds to the spur of the arcuate sulcus. An anatomical marker that subdivides PMd from PMv is absent in humans. However, a border between these two parts may be tentatively drawn in humans on the basis of embryological data. These data suggest that the monkey superior limb of the arcuate sulcus plus the spur represents the human homolog of the superior frontal sulcus plus the superior precentral sulcus. If this homology is accepted, the border between human PMd and PMv should lie approximately at *Z* level 50 in Talairach coordinates (Figure 6).

This hypothesis is supported by the location of the FEF in the two species. In monkeys, the FEF lies in the anterior bank of the arcuate sulcus, bordering posteriorly the sector of PMv (area F4) where head and proximal arm movements are represented. In humans, the FEF is located in the anterior bank of the superior precentral sulcus. Thus, the position of the FEF appears to be a good indicator of the border between PMd and PMv.

A second important issue concerns the presence and location in humans of the two main areas of PMv (F4 and F5) and of the three subareas (F5c, F5p, and F5a) that form F5. In humans, there is no obvious histological distinction between the upper part and the lower part of PMv. The only part of premotor cortex that may be clearly differentiated from the rest of it is area 44. Classical anatomical studies suggested that the whole F5 (area FCBm of von Bonin and Bailey) should be the macaque homolog of area 44. Recent data indicate, however, that this homology is valid only for the rostral part of area F5–area F5a.

Figure 6 illustrates the possible localization of the different areas found in the monkey on the lateral premotor region of humans, as suggested by gross anatomy and the considerations previously described. In the next section, we examine the functional validity of these subdivisions.

Different Functional Roles of the Various Sectors Forming the Human Premotor Cortex

As previously mentioned, both PMd and PMv are active during action execution. However, differences in the activation of the two areas have been observed during the phase that precedes overt action execution. When individuals are asked to prepare a movement that they subsequently have to execute, PMd is strongly

activated, whereas such an activation is typically lacking in PMv. Both PMd and PMv are active during motor imagery. During action observation, PMv is strongly and consistently activated, whereas activation of PMd is only occasionally observed.

These findings suggest a clear functional similarity between human PMd and monkey F2, on the one side, and human PMv and monkey F5 on the other. This conclusion has been supported by a repetitive transcranial magnetic stimulation study in which human PMd and PMv were stimulated during grasping and lifting of a manipulandum. Disrupting the neural processing in PMv impaired positioning of the fingers on the manipulandum, whereas disrupting PMd affected the generation of proximal arm movements necessary for lifting the manipulandum. These data provide clear evidence for a dissociation between the functional roles of PMv and PMd, with PMv shaping the hand before lifting and PMd initiating motor commands for lifting.

Other imaging studies, exploiting PMv activation during the observation of actions done by others, have investigated whether this area is somatotopically organized. Actions done with different effectors (mouth, arm/hand, and leg/foot) were used as stimuli. The results showed that the observation of mouth movements determined activation of the lower part of PMv and of area 44 bilaterally. Observation of hand/arm movements determined two areas of activation: one in area 44 and the other located in the precentral gyrus, dorsal to that found during the observation of mouth movements. Finally, the observation of foot/leg actions determined an activation of a dorsal sector of the precentral gyrus.

These data should be considered with caution, however. In fact, although it is possible that PMv is organized in terms of effectors, like the primary motor cortex, it might also be possible that given its higher functional complexity, PMv is organized, at least in part, in terms of action goal. This possibility has not been studied. Despite this, the available data allow some important conclusions.

First, a somatotopic trend is undoubtedly present in PMv, with the mouth located ventrally and the foot in the dorsalmost part of PMv. Second, a representation of hand/arm movements is present in the dorsal part of PMv. It is likely that this representation corresponds to that found in monkey F4 (Figure 8). This possibility is supported by brain imaging studies that showed polymodal responses, a characteristic of monkey F4, in dorsal PMv. Third, there is a representation of hand actions in area 44. This finding may seem surprising. Area 44 is part of Broca's area, an area traditionally considered to be a speech area. In recent years, however, a large amount of evidence

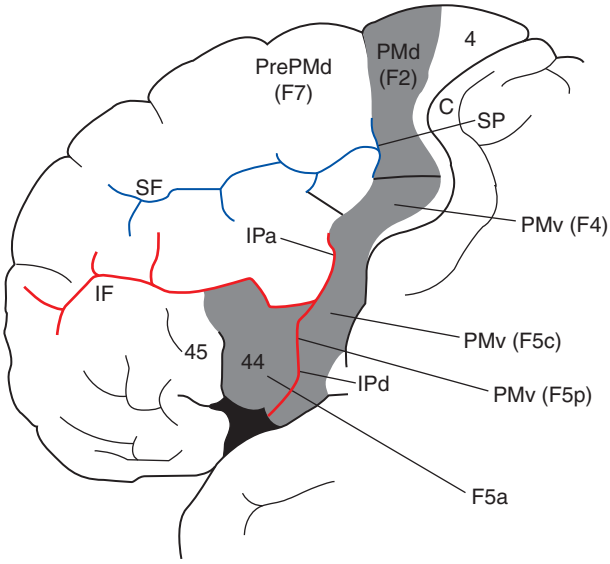

Figure 8 Lateral view of human frontal lobe showing a parcellation of the motor cortex according to the proposed homologies with the monkey motor cortex. In gray: areas 6aα and 44. PrePMd (F7) corresponds to area 6aβ of Vogts. In blue: superior frontal sulcus (SF) and the upper part of the superior precentral sulcus (SP). In red: inferior frontal sulcus (IF), the ascending branch of the inferior precentral sulcus (IPa), and the descending branch of the inferior precentral sulcus (IPd). C, central sulcus.

has shown that human area 44, in addition to being a speech area, also contains a representation of finger movements and is strongly active during object manipulation. Precision grip is especially represented. These properties, also present in monkey F5, confirm the cytoarchitectonic homology between monkey F5 (particularly F5a) and area 44.

In summary, the comparison of human and monkey premotor regions clearly indicates that the basic functional organization of this region has been maintained in evolution. The most dramatic modifications have been those due to the appearance of speech and related evolution of areas mediating this function. Yet, even in this case, the presence of hand-action representation in human area 44 indicates continuity with monkey F5, where mouth and hand actions are also represented. Thus, it is likely that the development of area 44 from monkey area F5 is not only in part related to the necessity for a sophisticated motor control of the orolaryngeal tract but also, in part, to its mirror properties that before speech evolution mediated facial and gesture communication.

See also: Corticomotoneuronal System; Finger Movements: Control; Map Plasticity and Recovery from Stroke; Neural Coding in Primary Motor Cortex; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Premotor Areas: Medial; Synaptic Mechanisms of Learning.

Further Reading

Buccino G, Binkofski F, Fink GR, et al. (2001) Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience* 13: 400–404.

Cisek P and Kalaska JF (2005) Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron* 45: 801–814.

Davare M, Andres M, Cosnard G, Thonnard JL, and Olivier E (2006) Dissociating the role of ventral and dorsal premotor cortex in precision grasping. *Journal of Neuroscience* 8: 2260–2268.

Gallese V, Fadiga L, Fogassi L, and Rizzolatti G (1996) Action recognition in the premotor cortex. *Brain* 119: 593–609.

Gibson JJ (1979) *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Graziano MSA and Gross CG (1995) The representation of extrapersonal space. A possible role for bimodal, visual–tactile neurons. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Hoshi E and Tanji J (2006) Differential involvement of neurons in the dorsal and ventral premotor cortex during processing of visual signals for action planning. *Journal of Neurophysiology* 95: 3596–3616.

Jeannerod M, Arbib MA, Rizzolatti G, and Sakata H (1995) Grasping objects: The cortical mechanisms of visuomotor transformation. *Trends in Neurosciences* 18: 314–320.

Kakei S, Hoffman DS, and Strick PL (2001) Direction of action is represented in the ventral premotor cortex. *Nature Neuroscience* 4: 1020–1025.

Nelissen K, Luppino G, Vanduffel W, Rizzolatti G, and Orban GA (2005) Observing others: Multiple action representation in the frontal lobe. *Science* 310: 332–336.

Pesaran B, Nelson MJ, and Andersen RA (2006) Dorsal premotor neurons encode the relative position of the hand, eye, and goal during reach planning. *Neuron* 51: 125–134.

Rizzolatti G and Craighero L (2004) The mirror–neuron system. *Annual Review of Neuroscience* 27: 169–192.

Rizzolatti G and Luppino G (2001) The cortical motor system. *Neuron* 31: 889–901.

Sakata H, Taira M, Murata A, and Mine S (1995) Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. *Cerebral Cortex* 5: 429–438.

Wise SP (1985) The primate premotor cortex: Past, present and preparatory. *Annual Review of Neuroscience* 8: 1–9.

Prepulse Inhibition of Startle in Humans and Laboratory Models

N R Swerdlow, University of California at San Diego School of Medicine, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Mechanisms within the central nervous system (CNS) normally serve to inhibit responses to sequential or repetitive stimuli. This inhibition can be demonstrated in a variety of ways, from the millisecond refractory period that follows an action potential to the few milliseconds of suppressed cortical excitability that follow a pulse of transcranial magnetic stimulation (TMS) to the inhibition of P50 and N100 cortical event-related potentials (ERPs) when auditory click pairs are spaced at 500 ms intervals. At a clinical level, a number of neuropsychiatric disorders are characterized by failures of inhibition – deficits in mechanisms that normally inhibit the intrusion into consciousness of unwanted repetitive sensations, images, or thoughts or the expression of unwanted repetitive movements or sounds. Significantly, the neural bases for these inhibitory deficits in patients can be studied through the use of laboratory measures that assess the amount of neuronal, cognitive, sensory, or motor inhibition elicited by specific types of controlled stimuli.

Prepulse inhibition (PPI) is the normal reduction in startle magnitude that occurs when a startling stimulus is preceded 30–300 ms by a weak prestimulus (Figure 1). The inhibitory processes activated by the weak prepulse and the resulting decrement in startle amplitude are used in an operational definition of sensorimotor gating – the degree to which a motor response (the startle reflex) is inhibited (gated) by a sensory event (the prepulse) is a measure of the amount of sensorimotor gating. PPI occurs whether the prepulse and startling stimuli are in the same or different sensory modalities. It is not a form of conditioning because it occurs on the first exposure to the prepulse and pulse stimuli. Whereas the inhibitory effect of the prepulse on startle is exerted in the pons, descending forebrain influences regulate the inhibitory tone within the pons and determine the degree to which the prepulse inhibits the subsequent motor response. PPI thus appears to reflect the activation of ubiquitous hardwired centrally mediated behavioral gating processes that are regulated by forebrain circuitry.

In a conceptual model of PPI, a weak stimulus activates neural processes that blunt responsivity to sensory events during a brief temporal window. The gating window is empirically determined to be approximately 20–300 ms in duration in both rats and humans; in humans, inhibition during the initial 30–60 ms after a prepulse appears to be automatic (preconscious), whereas inhibition at longer intervals (e.g., 120–240 ms) appears to be sensitive to attentional mechanisms and conscious manipulation. Some of the earliest experimental demonstrations of PPI came in studies of the time order error in the 1930s, in which the presentation of paired S1–S2 acoustic stimuli, separated by 177 ms, resulted in a reduction in the perceived intensity of S2 and a suppression of the blink response to that stimulus. More recent studies have shown that prepulses reduce the perceived intensity of a forehead tap, a 105 dB acoustic noise, a 40 psi air puff, and a 170 V cutaneous shock. It has been suggested that the period of reduced responsivity after the prepulse momentarily protects information contained in this weak stimulus so that it can be processed without interference from subsequent stimuli.

In humans, PPI appears to increase from childhood to adulthood, reaching maximal levels by the end of the second decade, although the development of PPI prior to the age of 6 or 7 years in humans has not been fully studied. In adulthood, PPI exhibits robust test–test reliability in normal comparisons subjects over intervals of several weeks. Thus, intraclass correlations exceed 0.9 for PPI elicited over 3 consecutive weeks, using relatively intense prepulses and long prepulse intervals (e.g., 120 ms). PPI is sexually dimorphic, with higher levels in men than in women under many experimental conditions, and has also been reported to vary across the normal menstrual cycle.

In the laboratory, PPI is studied under controlled conditions. In this environment, PPI is highly sensitive to experimental conditions, including stimulus modalities and parameters, background noise type and levels, test instructions, attentional demands, and methods of response acquisition. For example, although they may generalize to many other conditions, the findings of human PPI described in this article refer primarily to uninstructed PPI acquired using discrete acoustic noise bursts as prepulses and startle pulses with 60–120 ms stimulus onset asynchronies (SOAs) and with eyeblink reponses assessed via electromyographic (EMG) measures of orbicularis oculi (OO). This artificial and very specific sampling of inhibitory CNS tone is used as a basis for understanding the more complex construct of sensorimotor gating, which, under natural circumstances, is conceptualized as being continuously active in waking humans, triggered differentially by the wide range of stimuli that constitute the natural environment. Under these

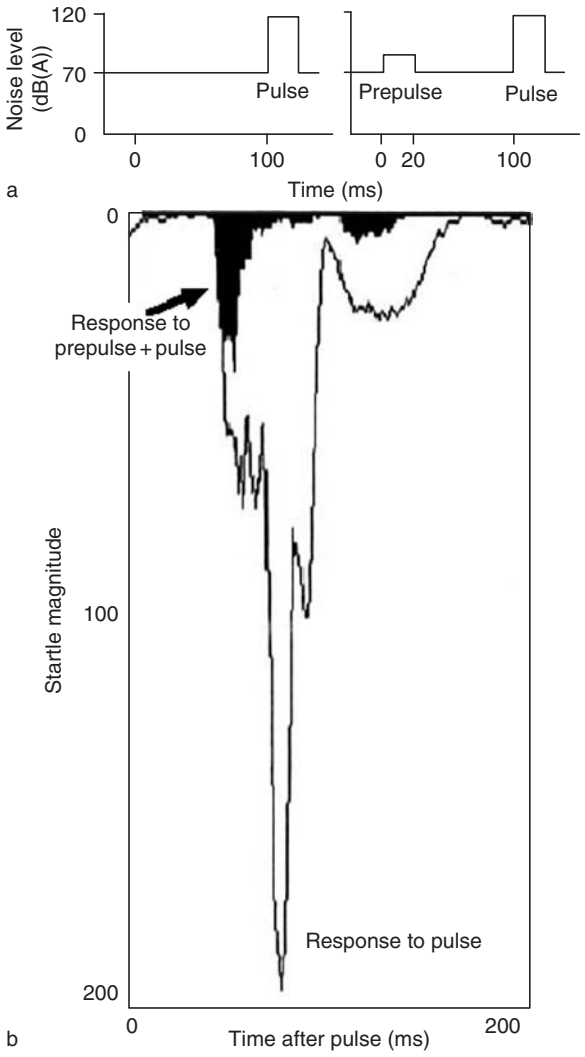

Figure 1 Stimuli for and measurements of PPI: (a) schematic representation of stimuli used to elicit PPI in laboratory measures; (b) superimposed tracings of EMG of the right OO in a normal adult male subject, from sequential trials that included either a prepulse (20 ms 4 dB noise burst over a 70 dB(A) background) followed 100 ms later by a 118 dB(A) 40 ms startle noise pulse (black shaded area) or the startle pulse alone (open area). The amount of inhibition generated by the prepulse can be appreciated visually by subtracting the shaded area from the open area. Also note the latency facilitation (shorter peak reflex latency on trial with prepulses).

natural conditions, PPI is thought to contribute adaptive value by enhancing the ability to segregate a continuous stream of sensory and cognitive information and to selectively allocate attentional resources to salient stimuli.

Specific characteristics of an individual's gating processes are viewed to be plastic, shaped by genetic and developmental forces, but also to be sensitive to changes in environment, stressors, or the neurochemical milieu of the nervous system. There is some evidence that, in clinically normal individuals, levels of PPI correlate positively with performance on certain neurocognitive tasks and, conversely, that PPI deteriorates with increasing mental fatigue. A more substantial pathological breakdown in the gating window might be expected to impair the orderly, hierarchical processing of sensory events and result in flooding of sensory information, with a fragmenting impact on cognition.

PPI in Animal Models and Preclinical Studies

One of the major advantages of PPI as a measure of CNS function and dysfunction is that it can be measured across species. Studies of PPI in rodents have been particularly valuable in delineating the neural control of PPI and in starting to identify genes and gene products that impact this neural circuitry to influence the PPI phenotype.

Measurement and Interpretation of Prepulse Inhibition

As noted earlier, PPI is most often measured in humans using EMG of the OO muscle to capture the blink component of startle. In rodents, including mice, rats, and guinea pigs, PPI is generally assessed using measures of whole-body startle. Here, the startling stimulus triggers a contraction of the skeletal muscles, which shortens the torso and elevates the back, causing the equal and opposite force to be directed through the feet to the floor of a test chamber. This force is then captured by a sensor, often integrated across a 50–200 ms window, and quantified as a peak amplitude or an area-under-the-curve magnitude. Prepulses with appropriate stimulus characteristics reduce or inhibit this force. Both whole-body and OO EMG measures have been used to measure PPI in infrahuman primates.

Whereas most measures of PPI are based on a ratio of startle magnitude elicited by a pulse under conditions with versus without a prepulse, other approaches use a measure of startle probability to quantify both the startling effect of pulses and the startle-modifying effects of prepulses. One complexity introduced by the latter approach is that it requires the categorical assessment that any nonzero level of motor activity constitutes a startle response. In practice, the sensitivity of many startle measurements is such that low levels of background motor activity (<1% of a startle signal) are often detected, even on no-stim trials (when no stimuli are presented to the subject). The interpretation of the response probability measures of PPI depends greatly on the distinction between spontaneous and stimulus-evoked motor

activity; at very low (nonzero) levels, this distinction is quite difficult.

Under certain conditions (e.g., weak prepulses or very short or long prepulse intervals) or after certain experimental manipulations (e.g., high doses of direct dopamine (DA) agonists in some strains of rats), prepulses can actually increase startle response magnitude. There is no clear consensus as to whether PPI and prepulse potentiation (PPP) reflect distinct neurobiological processes or, instead, are simply part of a continuum of prestimulus effects, with an arbitrary zero point. From a practical standpoint, many clinical and preclinical data sets ignore this possible distinction and combine trials that elicit PPI and those that elicit PPP into a single mean value; similarly, group means often reflect the combination of subjects who exhibit PPI with those who exhibit PPP. Clearly, if these reflect two distinct biological processes, an observed mean reduction of PPI might actually reflect an enhancement of PPP, and vice versa.

Except at extreme floor and ceiling levels, PPI is relatively insensitive to differences in startle magnitude to the pulse (S2) stimulus. For example, it is often the case that drug effects on startle magnitude and PPI can be dissociated experimentally, either via the use of both weak and intense startle stimuli or via analytic approaches that control for levels of startle magnitude. However, at extremely low startle values, it becomes impossible to reliably detect PPI (i.e., you cannot inhibit something that is not there), and the resulting artificial values of percentage of PPI (%PPI) can introduce substantial variability into a data set. For example, under conditions in which a startling stimulus typically elicits a response magnitude of 100 arbitrary units, a deaf rat (e.g., resulting from the use of inappropriate ear bars during stereotaxic surgery) might register a response magnitude of 1 or 2 units simply by shifting body position within the 100 ms recording window. If, by chance, this value registers on a trial that includes a prepulse, the calculated (artificial) value of %PPI might be quite extreme (e.g., −500%). Even among normal subjects, it becomes prudent to exclude a small percentage from analysis based on nonresponse criteria (which can be derived empirically from a normal distribution), as long as it can be determined that such an approach does not introduce a systematic error into the analysis.

Above certain intensities, prepulses (S1s) can elicit startle responses, and the sensitivity to this S1-elicited motor responsivity is influenced by a number of experimental factors. The relationship between the paired-pulse inhibition and PPI elicited by nonstartling cues is an area of ongoing study. At short prepulse intervals in humans (e.g., 30 ms), prepulses also reduce the latency to reflex onset and peak amplitude. This latency facilitation provides useful evidence of sensory registration. Thus, several clinical conditions characterized by a loss of prepulse effects on startle magnitude (i.e., PPI) exhibit normal latency facilitation. In these cases, the loss of PPI cannot be attributed to a failure of the CNS to detect or register the prepulse because the prepulse clearly produces a physiological change in reflex latency.

Under specific conditions, prepulses are known to inhibit the perceived intensity of the startling stimulus in addition to the motor response elicited by that stimulus. The PPI of perceived stimulus intensity (PPIPSI) has been demonstrated using startling stimuli ranging from noise bursts to air puffs to intense cutaneous shocks. PPI and PPIPSI exhibit both similarities and differences in their responses to experimental manipulations, probably reflecting both convergence and divergence in their underlying neurobiological substrates. The ability of prepulses to inhibit the perceived intensity of painful stimuli suggests that they may have clinical utility in the delivery of therapeutic shocks.

Pharmacology of PPI

PPI can be studied using laboratory measures that are largely automated, objective, and under tight parametric control and that can be adapted across several mammalian species. Because of this, PPI is a measure that is well suited to rapid throughput studies, and an active laboratory with a few test chambers can generate agonist–antagonist dose–response curves in a single day. It is thus not surprising that the pharmacology of PPI has been extensively studied in mice, rats, guinea pigs, pigs, infrahuman primates, and humans. Most of these studies have been conducted in rats, and almost all have focused on acute drug effects on the amount of PPI across a limited set of stimulus parameters. For example, relatively less is known about the pharmacology that regulates the sensitivity of PPI to different prepulse intervals, frequencies, or background noise levels; under common experimental conditions, the amount of PPI is regulated by many different neurotransmitters and neuropeptides. Although there are clear species differences, the most common findings reveal that the amount of PPI is reduced by systemic administration of DA or serotonin (5-HT) agonists or *N*-methyl-D-aspartate (NMDA) antagonists; under specific conditions, the amount of PPI is increased by atypical antipsychotics and nicotine. Neuropeptides most often implicated in the regulation of PPI are neurotensin, cholecystokinin, and corticotropin-releasing hormone (CRH).

Of the major neurotransmitters, perhaps most studied is the regulation of PPI by DA. The PPI-disruptive effects of direct DA agonists (e.g., apomorphine or

quinpirole) are generally more robust than those of indirect agonists (e.g., amphetamine or cocaine). Among direct DA agonists, PPI-disruptive effects in rats appear to require the stimulation of D2-family receptors, although such effects may be augmented by concomitant D1 stimulation. Interestingly, PPI can also be disrupted by the systemic administration of the D1 antagonist, SCH 23390. The PPI-disruptive effects of the direct D1–D2 agonist apomorphine are opposed by clinically effective antipsychotics, and the potency of antipsychotics in this model predicts their clinical antipsychotic potency ($r = 0.991$). Related models have demonstrated the ability of acute administration of atypical (but not of typical) antipsychotics to prevent the PPI-disruptive effects of the NMDA antagonist phencyclidine and the ability of typical and atypical antipsychotics to restore normal levels of PPI in adult rats that had been reared in social isolation.

Neural Substrates of PPI

PPI is mediated at the level of the pons and remains intact in the acutely decerebrate rat. Preclinical studies have demonstrated that PPI in rodents is regulated by cortical structures (mesial temporal and medial prefrontal cortex) and subcortical structures (striatum, pallidum, and pontine tegmentum). This limbic cortico-striato-pallido-pontine circuitry converges with the primary startle circuit at the level of the nucleus reticularis pontis caudalis (NRPC). Representative examples of regionally specific neurochemical manipulations that disrupt PPI in rodents are listed in Table 1.

Studies of the neurobiology of PPI in humans reveal both similarities and differences from findings in animal models. For example, evidence of profound PPI deficits in patients with Huntington's disease (HD) supports the role of the striatum in the regulation of PPI in humans, and parallel findings demonstrate PPI deficits after excitotoxic lesions of the striatum in rats and in mice transgenic for the HD gene. Neuroimaging studies in schizophrenia patients and normal comparison subjects confirm both the role of limbic cortico-striato-pallido-thalamic (CSPT) circuitry in the normal regulation of PPI (Table 1) and the association between CSPT abnormalities and PPI deficits in patients. Also consistent with a number of preclinical findings is the fact that PPI deficits are observed in other clinical populations, ranging from Tourette syndrome to seizure disorders, with known or presumed pathology in limbic CSPT circuitry.

On the other hand, pharmacological studies of PPI in normal humans raise questions as to whether the PPI neural circuit blueprint in rodents can be easily translated across species. Thus, although the indirect DA agonist D-amphetamine reduces PPI in both normal humans and some rodent strains, the NMDA antagonist ketamine and the serotonin releaser, 4-methylenedioxy-*N*-methylamphetamine (MDMA) have opposite effects on PPI across species – disrupting PPI in rodents but increasing PPI in clinically normal humans. A similar pattern is observed with the mixed DA agonist–NMDA antagonist amantadine. Direct DA agonists potently disrupt PPI in rats, but their effects in humans are not as convincing. Both atypical antipsychotics and nicotine have been reported to increase PPI across species but perhaps only under certain experimental conditions. Clearly, substantial cross-species differences in drug doses, routes of administration, and drug disposition (based on species differences in pharmacokinetics and -dynamics) complicate the comparison of PPI drug effects in rodents versus humans. Even simple mechanical differences, such as the relative distances required for impulse conduction through PPI-regulatory circuitry, might be expected to impact the temporal sensitivity of prepulse effects to drug challenges. Indeed, drugs can have opposite effects on PPI with short (10–30 ms) versus longer (60–120 ms) prepulse intervals in rodents. An additional level of complexity is added by findings that some drug effects on PPI in both rodents and humans depend substantially on the basal levels of PPI; for example, normal humans and rats with high basal levels of PPI appear to be most sensitive to the PPI-disruptive effects of DA agonists, whereas those with low basal PPI levels appear to be most sensitive to the PPI-enhancing effects of atypical antipsychotics. Last, given the considerable strain differences in PPI drug sensitivity in rodents, it is clear that different mouse and rat strains might exhibit more or less homology to humans in terms of PPI sensitivity to various drugs.

Table 1 Examples of neurochemical manipulations that disrupt PPI in rats

Brain region	*Manipulation that disrupts PPI*
Medial prefrontal cortex	D1 blockade
Basolateral amygdala	NMDA receptor blockade
Ventral hippocampus	NMDA
Nucleus accumbens core	D2 stimulation
Ventral pallidum	GABA receptor blockade
Dorsomedial thalamus	GABA receptor stimulation
Pedunculopontine nucleus	GABA receptor stimulation

D1, D2, dopamine receptors; GABA, γ-aminobutyric acid; NMDA, *N*-methyl-D-aspartate; PPI, prepulse inhibition.

Genetics of PPI

Because PPI is regulated by brain circuitry, it follows that genes that control the function of that brain

circuitry should also impact the level of PPI. Although this relationship must hold for all brain-mediated processes, the relative simplicity of PPI and the fact that it can be quantified via automated measures in laboratory animals in rapid throughput screening in controlled environments with precise parametric control have made it a popular focus for studies exploring gene–behavior relationships. Furthermore, because PPI is aberrant in several brain disorders, the identification of the genes associated with PPI carries the additional potential that PPI genes might contribute to the vulnerability for, or expression of, these disorders.

Heritability Very convincing evidence for the genetic control of PPI comes from studies in rodents. For example, Francis et al. reported that mice generated via embryonic transplants into a different maternal strain exhibited PPI phenotypes of the embryo and not of the maternal uterine or rearing environments; this contrasted with several other behavioral phenotypes, which exhibited characteristics linked to the pre- and postnatal environments. Thus, PPI appears to be linked closely to genotype in mice. PPI levels differ substantially across mouse strains, and strain-specific PPI phenotypes are stable over multiple generations.

Both basal and drug-modified PPI also exhibit profound differences across several rodent strains. For example, compared to Harlan Sprague Dawley (SD) rats, Harlan Brown Norway (BN) rats exhibit significantly greater levels of PPI at short prepulse intervals and significantly lower levels of PPI at long prepulse intervals. SD rats are also significantly more sensitive than Harlan Long–Evans (LE) rats to the PPI-disruptive effects of DA agonists; such a difference is also observed between SD and Harlan Wistar rats. This strain difference in PPI sensitivity between SD and LE rats has been found to be (1) neurochemically specific to DA agonists (i.e., not seen with NMDA antagonists or 5-HT agonists); (2) stable across breeding and testing facilities; (3) evident on or before day 18 of life and after systemic or central drug administration; and (4) inherited through F1 (SD × LE) and N2 backcross (F1 × SD) generations via an orderly pattern of DA agonist sensitivity (SD > N2 > F1 > LE), which suggests the additive effects of multiple genes. The latter heritable differences are unaffected by cross fostering and do not appear to be related to presynaptic forebrain processes but, rather, to D2-receptor-linked G-protein functions.

Only one published study has directly assessed heritability of PPI in humans. Anokhin et al. assessed PPI in 40 monozygotic and 31 dizygotic female twin pairs and used structural equation modeling to demonstrate significant heritability of PPI that accounted for over 50% of PPI variance. Menstrual cycle phase was not matched in these twin pairs, suggesting that heritability may have actually been higher given the likely added variance in this sample based on the differential hormonal state influences on PPI.

Specific genes Based on the neurobiology of PPI, it is reasonable to hypothesize that genes that impair the normal development or function of limbic CSPT circuitry should be associated with reduced PPI. Three types of evidence have been used to implicate specific genes or gene regions in the regulation of PPI. First, PPI deficits in both HD and 22q11-deletion syndrome suggest that the genes affected in both of these disorders modify brain circuitry that regulates PPI. In both of these instances, animal models with homologous genetic defects also exhibit PPI deficits. Second, quantitative trait loci (QTLs) have been identified, either through interval mapping in inbred rat strains (on chromosomes 2 and 18) or recombinant congenic mouse strains (five QTLs across chromosomes 3, 5, 7, and 16) or through the use of chromosome substitution strains in mice (two QTLs on chromosome 16). Third, reverse genetic approaches have identified a long list of genes (and molecules that they encode) that are associated with a reduction in PPI or alterations in the effects of a drug on PPI when inactivated via constitutive or conditional knockout techniques (Table 2).

Significantly, most genes and molecules associated with lower versus higher levels of PPI in mutant animals are most likely unrelated to reduced PPI in schizophrenia or other disease states. The most potent physiological influence on acoustic PPI is hearing threshold because an organism that cannot detect an acoustic prepulse does not exhibit acoustic PPI. Many studies reporting PPI deficits in mice after genetic alterations test neither hearing threshold nor levels of nonacoustic PPI. Similarly, PPI gene-mapping strategies in inbred and recombinant rodents might be expected to identify genes associated with hearing threshold. Beyond the level of sensory registration, the most potent physiological influence on PPI is exerted at the level of the pedunculopontine nucleus (PPTg), which mediates PPI via its impact on the NRPC. For the same reasons noted for hearing threshold, genetic studies of PPI are likely to be influenced strongly by genes coding for the normal function of the PPTg. And although the PPTg has been hypothesized to be abnormal in some individuals with schizophrenia, it does not play a central role in models for the pathophysiology of this disorder (although it may be relevant to other conditions characterized by PPI deficits, e.g., nocturnal enuresis). In contrast, the prefrontal cortex (PFC) – which is suspected to be a critical substrate for some symptoms of schizophrenia – is likely to be three or four synapses removed from the

Table 2 Examples of genes and molecules that have been implicated in the regulation of PPI through reverse genetic strategies in animal models

α2A-Adrenergic receptors	HD gene
Calcineurin	Homer-1
Catna2	Interleukin-2
Cell adhesion molecule L1	Lysophosphatidic acid receptor-1
CRH1 and 2 receptors	mGluR-1 and -5 receptors
DA transporter	NCAM-180
DA D2-S receptor	Nicotinic receptor β3 subunit
Dvl1	NMDA receptor subunits NR1, 2A, 2B, 3A
Fmr1	Npas3
G(s)α	Reelin
GABA transporter subtype 1	Superoxide dismutase-2
$GABA_A$ α3 and -B1 subunits	Stable tubule-only polypeptide
Glutamic acid decarboxylase 65	Tbx1
Glycine transporter subtype 1	Telencephalin
Glycogen synthase kinase 3 β	Zic2
Gnb1l	

CRH, corticotropin-releasing hormone; DA, dopamine; HD, Huntington's disease; mGluR, metabotropic glutamate receptor; NCAM, neural cell adhesion molecule; PPI, prepulse inhibition.

primary startle circuit, and in a normal human or rodent, genes controlling the PFC probably contribute only weakly to any gene mapping a signal based on levels of PPI.

Conversely, it is known that the HD gene exerts a powerful control on the viability of striatal circuitry that regulates PPI and that the pathology effected by this gene is strongly associated with low or absent levels of PPI in humans and animal models. Yet, despite this, it is almost certain that no gene-mapping efforts based on the PPI phenotype in normal rodent strains or reverse genetic models based on constitutive or conditional knockout strategies (not including the insertion of the HD gene) will ever link the HD gene to low levels of PPI. Significantly, PPI is a measure that reflects the normal function of specific brain circuits, and genes associated with PPI deficits in neuropsychiatric disorders are likely to be ones that contribute uniquely to dysfunction in those brain circuits. Such genes will be most effectively identified through the use of affected and at-risk individuals, in whom deficient PPI is used as an endophenotype – a surrogate marker for neural circuit dysfunction encoded by disease genes that contributes to the vulnerability for the expression of the full clinical phenotype.

PPI in Clinical Populations

PPI has been studied extensively as an example of normal CNS plasticity and an index of early information processing, but it is also used as a means to investigate CNS dysfunction in neuropsychiatric disorders. Deficits in PPI suggest abnormal neural activity either within the primary PPI circuitry within the pons or, more often, within the forebrain circuitry that regulates PPI. Under different testing conditions, PPI has been reported to be deficient in schizophrenia, schizotypal personality disorder, 22q11-deletion syndrome, mania, Tourette syndrome, obsessive–compulsive disorder, panic disorder, HD, blepharospasm, autism, Asperger's syndrome, fragile-X syndrome, and noturnal enuresis. Normal levels of PPI have been reported in major depressive disorder, euthymic bipolar disorder, borderline personality disorder, Alzheimer's disease, and stuttering; conflicting reports have described either reduced or normal levels of PPI in different populations of patients with posttraumatic stress disorder (PTSD).

Schizophrenia

Among these patient groups, PPI has been most extensively studied in schizophrenia. Deficient PPI has been reported not only in schizophrenia patients but also in their clinically unaffected relatives, suggesting that these deficits may be a trait marker for individuals at-risk for developing the disorder (rather than a marker of schizophrenia *per se*). The relationship between PPI deficits and clinical features of schizophrenia is not well understood. Studies report weak or no significant correlations between PPI and symptom severity. Some evidence links deficient PPI to the core schizophrenia trait (rather than symptoms) or cognitive and functional impairment. Other studies report deficient PPI only in symptomatic patients, which partially or completely resolves with medication treatment. In a study of more than 100 schizophrenia probands, we failed to detect significant relationships between PPI deficits and a number of neuropsychological measures, despite the fact that these PPI deficits did correlate significantly with measures of long-term functional impairment (e.g., Global Assessment of Function (GAF) score and level of independent living).

Relatively few studies have assessed the longitudinal stability of PPI deficits in schizophrenia patients. Whereas several groups reported PPI deficits in relatively stable, medicated outpatient schizophrenia cohorts, others reported that PPI deficits occurred only in patients who are in an acute symptomatic phase of a psychotic episode and resolve in these same patients as symptoms subside. One complexity in interpreting such state-related changes in PPI relates to the role of antipsychotic medications in both symptomatic improvement and potentiation of

PPI. Thus, it is not clear whether PPI is sensitive to the state of psychotic symptoms, the state of antipsychotic medications, or both.

Factors That Commonly Impact PPI in Clinical Studies

The interpretation of PPI levels in clinical populations is complicated by a number of factors that impact PPI in humans.

Antipsychotics In preclinical models, antipsychotics oppose the PPI-disruptive effects of pharmacological (e.g., DA agonists and NMDA antagonists), surgical (e.g., adult lesions of the basolateral amygdala or neonatal lesions of the ventral hippocampus), or developmental (e.g., isolation rearing) challenges. Some atypical antipsychotics (e.g., quetiapine and clozapine) are capable of increasing basal levels of PPI in rats and mice under conditions that elicit submaximal levels of PPI (e.g., weak prepulses and prepubertal age). Significantly, this suggests that, under some physiological conditions, antipsychotics modify processes responsible for the normal regulation of sensorimotor gating and not simply its disruption by pathological states.

Relatively little is known about the effects of antipsychotics on PPI in normal human subjects. Published studies using the typical antipsychotic haloperidol reported either no effect or PPI-reducing effects of this drug, consistent with observations in infrahumans. These reports used different sets of experimental conditions, including pure tones versus noise bursts and a range of prepulse intensities. We reported that 12.5 mg of quetiapine has no significant effects on PPI (10–120 ms intervals) in normal men; this was also observed by Graham et al., with both 12.5 and 25 mg doses of quetiapine. However, we also demonstrated that quetiapine (12.5 mg) significantly increases short-interval (20 ms) PPI in clinically normal, low-gating men (defined by basal PPI levels in the lowest quartile of a normal distribution). This latter finding, together with similar findings in normal and low-gating rodent strains, suggests that elevations in PPI by antipsychotics may not reflect processes linked specifically to the pathophysiology of schizophrenia *per se*.

The initial report of PPI deficits in schizophrenia patients predated the use of atypical antipsychotics. Preclinical studies later demonstrated that, in addition to preventing the PPI-disruptive effects of various experimental manipulations, atypical antipsychotics also enhance baseline PPI in some rodent strains and species. Subsequent clinical reports suggested that atypical antipsychotics are associated with greater and potentially normalized PPI levels in schizophrenia patients. In our recent study of over 100 schizophrenia patients, only 17% of patients were not being treated with atypical antipsychotics. Clearly, if atypical antipsychotics increase PPI in some groups of patients, the result of such pervasive use of atypical antipsychotics in schizophrenia patients (or other populations, e.g., those with Tourette syndrome) will be to diminish or eliminate the ability to detect PPI deficits between patients and normal controls and to diminish estimates of PPI heritability in any sample in which antipsychotics are used. Some studies also suggest that these schizophrenia-linked deficits are partially normalized by typical antipsychotics, and our recent findings are consistent with these reports. Because the overwhelming number of schizophrenia patients are currently treated with atypical antipsychotics, it is possible that PPI deficits in this population are a vanishing biomarker. Alternative strategies for understanding the biology and clinical implications of deficient sensorimotor gating have become increasingly important, including the use of optimized stimulus features (e.g., 60 ms prepulse intervals), broadband stimuli and prominent background noise, unmedicated schizophrenia and schizophrenia spectrum patients, and populations of low-gating controls.

Nicotine The pro-gating effects of nicotine have been reported in measures of both PPI and P50 event-related potential (ERP) suppression. If nicotine use *per se* is associated with increased PPI, then the higher rates of smoking among schizophrenia patients would be expected to diminish the patient- versus control-group differences in PPI measures and to diminish measures of heritability. However, nicotine has a complex pharmacology, characterized by rapid shifts in both blood levels and states of receptor sensitization that can influence PPI. Because of the difficulty in controlling for the time interval between PPI testing and smoking, the amount smoked, smoking history, or the interindividual differences in nicotine receptor sensitivity, the impact of smoking on measures of PPI is effectively to increase uncontrolled variance. Studies with adequate power to compare nonsmoking patients (particularly males) with normal controls revealed group PPI differences (controls > patients) with greater group separation than is observed across the inclusive group of smokers and nonsmokers. Furthermore, among patients, PPI is significantly greater among heavier smokers.

Sex differences Sex differences (male > female) and menstrual cyclicity (peak levels in early follicular phase, and trough levels in mid-luteal phase) in PPI have been reported by several different groups and may have important implications for the interpretation

of PPI differences in schizophrenia versus control populations. Typically, open subject recruitment in schizophrenia studies favors ascertainment of male patients and female controls. Because men exhibit more PPI than women, this ascertainment bias artificially diminishes the control- versus schizophrenia-group differences. Menstrual cyclicity of PPI adds uncontrolled variance that may differentially affect control versus patient samples based on the hormonal effects of neuroleptics, and without carefully timed measurements, this state-sensitive cyclicity of PPI must diminish apparent heritability. Nevertheless, PPI deficits have been reported in both male and female schizophrenia patients. Lower PPI in female versus male patients has implications for studies attempting to assess the clinical or functional correlates of PPI in schizophrenia; sex differences in PPI among patients may obscure the impact of many other moderating variables or correlates.

One group has reported that PPI differs among normal subjects based on sexual orientation. In this study, heterosexual adults demonstrated the previously reported male > female difference in PPI levels. In contrast, among self-identified homosexuals, PPI did not differ between men and women. Specifically, PPI among female homosexuals was significantly elevated compared to female heterosexuals but was comparable to levels exhibited by heterosexual males.

Other subject characteristics PPI appears to be sensitive to a number of state and trait features. As noted earlier, PPI declines with mental fatigue. It is unclear how this change relates to the inverse correlation of PPI and spontaneous blink rate because the latter also increases with fatigue. Certainly, the potential impact of mental fatigue on PPI should be considered when it is used as part of an extended phenotyping battery in patient populations. In terms of traits, PPI does not seem to be related in any robust way with personality, although some significant (negative) correlations have been reported between PPI and personality features that are conceptually linked to psychosis. Differences in PPI have been reported between normal Caucasian and Asian individuals that appear to reflect significantly lower levels of eyeblink startle magnitude in Asian subjects, associated with relatively reduced mass of the OO muscle in the Asian sample. The potential effects of botulinum toxin injections of OO on startle magnitude and PPI may also be worth considering as a source of artifact, but they have not yet been evaluated.

Summary

PPI is a simple form of behavioral plasticity that serves as an operational measure of sensorimotor gating. It is easily quantified using laboratory measures that are largely automated, objective, and under tight parametric control and that can be adapted across several mammalian species. Studies of PPI have demonstrated its regulation by forebrain circuitry, which is of relevance to a number of neuropsychiatric disorders, and clinical studies have detected PPI deficits in several of these disorders. PPI is perfectly suited to rapid throughput screening, and this, together with a well-described pharmacology, make it an ideal model for drug discovery. Our evolving understanding of the genetic control of PPI suggests that measures of PPI may play an important role in identifying the genes that convey a vulnerability to certain brain disorders, including schizophrenia. Despite the great promise of this measure to be highly informative about a number of basic and clinical issues, its sensitivity to experimental parameters, subject characteristics, and potential analytic and behavioral artifacts make it a measure that is both deceptively complex and demanding of careful, systematic study and interpretation.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Antipsychotic Drugs; Dopamine – CNS Pathways and Neurophysiology; Mood Stabilizers; NMDA Receptors and Disease + C464; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Schizophrenia: Genetics; Startle Response.

Further Reading

Blumenthal TD, Cuthbert BN, Filion DL, Hackley S, Lipp OV, and van Boxtel A (2005) Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology* 42: 1–15.

Braff D, Stone C, Callaway E, Geyer M, Glick I, and Bali L (1978) Prestimulus effects on human startle reflex in normals and schizophrenics. *Psychophysiology* 15: 339–343.

Braff DL, Geyer MA, and Swerdlow NR (2001) Human studies of prepulse inhibition of startle: Normal subjects, patient groups, and pharmacological studies. *Psychopharmacology* 156: 234–258.

Fendt M, Li L, and Yeomans JS (2001) Brain stem circuits mediating prepulse inhibition of the startle reflex. *Psychopharmacology* 156: 216–224.

Geyer MA, Krebs-Thomson K, Braff DL, and Swerdlow NR (2001) Pharmacological studies of prepulse inhibition models of sensorimotor gating deficits in schizophrenia: A decade in review. *Psychopharmacology* 156: 117–154.

Geyer MA, McIlwain KL, and Paylor R (2002) Mouse genetic models for prepulse inhibition: An early review. *Molecular Psychiatry* 7: 1039–1053.

Graham F (1975) The more or less startling effects of weak prestimuli. *Psychophysiology* 12: 238–248.

Hoffman HS and Searle JL (1968) Acoustic and temporal factors in the evocation of startle. *Journal of the Acoustical Society of America* 43: 269–282.

Ison JR, McAdam DW, and Hammond GR (1973) Latency and amplitude changes in the acoustic startle reflex produced by

variations in auditory prestimulation. *Physiology & Behavior* 10: 1035–1039.

Swerdlow NR, Braff DL, and Geyer MA (2000) Animal models of deficient sensorimotor gating: What we know, what we think we know, and what we hope to know soon. *Behavioural Pharmacology* 11: 185–204.

Swerdlow NR, Braff DL, Taaid N, and Geyer MA (1994) Assessing the validity of an animal model of deficient sensorimotor gating in schizophrenic patients. *Archives of General Psychiatry* 51: 139–154.

Swerdlow NR, Geyer MA, and Braff DL (2001) Neural circuit regulation of prepulse inhibition of startle in the rat: Current knowledge and future challenges. *Psychopharmacology* 156: 194–215.

Swerdlow NR, Light GA, Cadenhead KC, Sprock J, Hsieh MH, and Braff DL (2006) Startle gating deficits in a large cohort of patients with schizophrenia: Relationship to medications, symptoms, neurocognition and level of function. *Archives of General Psychiatry* 63: 1325–1335.

Presynaptic Development and Active Zones

N E Ziv, Rappaport Family Institute for Research in the Medical Sciences and Technion Faculty of Medicine, Haifa, Israel
C C Garner, Stanford University, Palo Alto, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Chemical synapses are specialized sites of cell–cell contact designed for the transmission of signals between neurons and their targets – muscles, glands, or other neurons. Synaptic transmission depends on the tightly regulated secretion of neurotransmitters by the presynaptic cell and the reception of this signal by postsynaptic neurotransmitter receptors. The directional nature of synaptic transmission is manifested in the asymmetric structure of pre- and postsynaptic compartments. Structurally, presynaptic compartments are characterized by the presence of hundreds to thousands of neurotransmitter-filled synaptic vesicles (SVs) and by active zones (AZs) – specialized regions of the presynaptic plasma membrane where SVs dock, fuse, and release neurotransmitter into the synaptic cleft (Figure 1). The AZ is characterized by an electron-dense matrix or lattice of proteins (the cytoskeleton of the active zone (CAZ)), arranged into regular arrays of electron-dense tufts linked together by fine filamentous material. This structure, known as the presynaptic web or grid, is thought to define the AZ as the site of SV docking and fusion. The postsynaptic reception apparatus is also characterized by an electron-dense thickening referred to as the postsynaptic density (PSD), the central function of which is to confine receptors of the appropriate type beneath the AZ. The PSD and CAZ are held in register by transsynaptic cell adhesion molecules and extracellular matrix proteins.

Although presynaptic specializations can contain hundreds to thousands of SVs (the so-called 'reserve pool'), at any given moment only a few are 'tethered' or 'docked' at the AZ. A depolarizing action potential and the consequential opening of voltage-dependent calcium channels leads to an influx of Ca^{2+}, triggering the fusion of some docked vesicles with the presynaptic plasma membrane. SV membrane proteins are then retrieved by clathrin-mediated endocytosis in a periactive zonal region or possibly by direct retrieval of vesicles at AZs. Endocytosed vesicles are refilled with neurotransmitter, possibly after passage through specialized endosomal compartments, and returned to the reserve pool.

The formation of a new functional presynapse is a demanding process that calls for the recruitment of numerous components. These include molecules involved in SV docking, fusion and endocytosis; structural AZ proteins; Ca^{2+} channels; proteins that confine SVs to the presynaptic bouton; and the SVs themselves. In addition, presynapse formation must be tightly coordinated with the formation of the appropriate postsynaptic receptive apparatus. Here, the current state of knowledge on cellular and molecular mechanisms involved in presynapse development is summarized.

Presynaptic Development: Initial Events

Most presynaptic sites are formed during development as axons grow out and establish connections with their targets. Much of our understanding of presynapse formation comes from studies on the vertebrate neuromuscular junction (NMJ). These synapses, formed between motor neurons and skeletal muscle cells, begin to form within a few hours after the closure of the neural tube. Motor neuron axons, guided by growth cones at their tips, establish contacts with muscle targets in their final stages of differentiation. These axons grow a short distance down the midline of the differentiating muscle, giving rise to side branches in the process. Eventually, these growth cones develop into bulbous enlargements that possess a rudimentary capacity for spontaneous and evoked release of neurotransmitter. With time (many hours to days), these structures differentiate into typical presynaptic terminals and gradually acquire mature functional characteristics.

Studies concerning NMJ formation have traditionally emphasized the roles of axonal growth cones in setting off the series of events that ultimately lead to NMJ differentiation. This neurocentric view seemed natural given the motility of axonal growth cones compared to the relative immobility of developing muscles, their ability to release neurotransmitter before they contact their targets, and their capacity for inducing postsynaptic receptor clustering when placed in nerve–muscle co-cultures. Recent evidence, however, reveals that not all instructive cues are axonally derived: it has been known for many years that small neurotransmitter receptor clusters form spontaneously in muscle cells even in the absence of innervation, although the role of such receptor clusters was not clear. Recent *in vivo* experiments, however, reveal that elongating axons tend to grow toward preformed receptor clusters concentrated along the midline of differentiating muscles, and

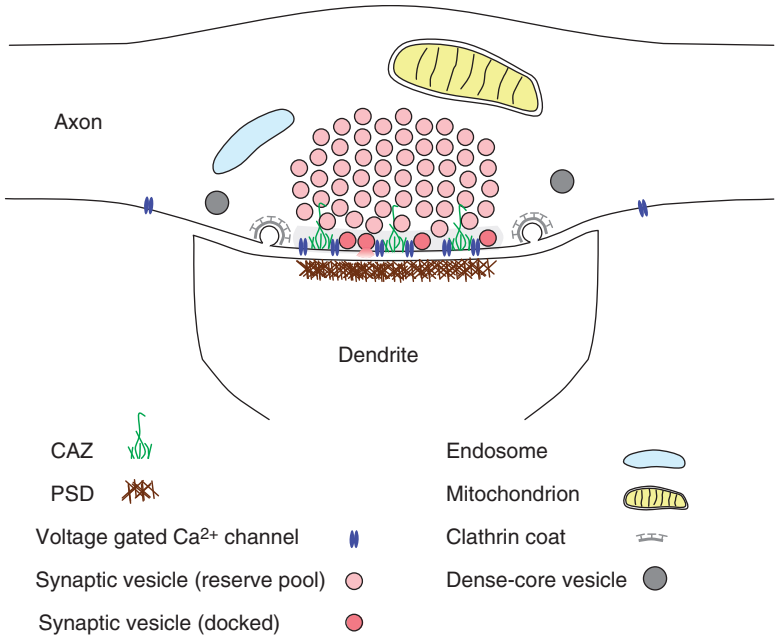

Figure 1 A schematic illustration of a vertebrate CNS, *en passant* spine glutamatergic synapse. The active zone, including CAZ molecules, voltage-gated Ca^{2+} channels, and docked vesicles, is shaded in gray. Clathrin-dependent endocytosis of synaptic vesicles is shown to occur along the margins of the active zone (the periactive zonal area). Distal to the active zone is the reserve pool, composed of numerous synaptic vesicles enmeshed and bound to a matrix of filamentous actin (cytoskeletal elements are not illustrated for purposes of clarity). Presynaptic compartments often contain mitochondria, a small number of dense core vesicles, and additional organelles such as endosomes.

these axons subsequently develop presynaptic specializations at these sites. This phenomenon serves to indicate that presynaptic differentiation is most likely specified by postsynaptic membrane factors. Interestingly, these factors are probably not the receptors themselves. Instead, studies suggest that components of the extracellular matrix, confined to muscle midline regions, may facilitate this process. This phenomenon also serves to highlight the fact that signals that specify pre- and postsynaptic differentiation are bidirectional, and thus the historic question (that is still raised occasionally) whether synaptic differentiation is induced by pre- or postsynaptic members is somewhat irrelevant.

In contrast to motor neuron axons which do not form synapses en route to their final destinations, axons of the vertebrate central nervous system (CNS) establish presynaptic sites along their lengths in the form of small swellings, or varicosities, known as presynaptic boutons. This arrangement allows one axon to form *en passant* synaptic connections with many dendrites along its route. *En passant* presynaptic boutons are often formed behind advancing axonal growth cones as they weave their way through the developing neuropil. However, the potential to form presynaptic boutons is not limited to the terminal growth cones and seems to be distributed along the entire axonal membrane. *In vitro* and *in vivo* studies indicate that presynaptic boutons may form along existing axonal segments in response to axodendritic contacts initiated by dendritic protrusive structures, namely growth cones at the tips of elongating dendritic branches or filopodia extended laterally from dendritic shafts. Axons also extend highly protrusive lateral structures that might also give rise to presynaptic boutons at positions far behind leading-edge growth cones. Thus, new presynaptic boutons can be formed from axonal growth cones, immediately behind advancing axonal growth cones and along established axonal segments.

Recent progress in imaging techniques, such as the introduction of green fluorescent protein (GFP) and the advent of multiphoton confocal microscopy, has provided unprecedented information on the cellular dynamics associated with presynapse development, both in cell culture and in the intact developing CNS of tadpoles, zebra fish, and mice. A consistent observation made in such imaging studies is that presynapse development is an extremely dynamic process. Specifically, these studies reveal that (1) many contacts formed between axons and dendrites are transient, resulting from exploratory-like growth and retraction

of axonal and dendritic processes; and (2) many, perhaps most, nascent presynaptic structures are transient in nature and often disappear within hours or less of their initial formation. These dynamics may be taken to indicate that the formation of a persistent presynaptic site is at odds with opposing mechanical forces that act to 'tear apart' the nascent axodendritic junctions – forces generated by axons and dendrites as these extend through the developing neuropil and by dendritic and axonal filopodia as they undergo cycles of extension and retraction. In this respect, it is interesting to note that *in vivo* imaging of axon and synaptic vesicle dynamics of retinal ganglion neurons in developing zebra fish embryos and *Xenopus* tadpoles indicates that the formation of a presynaptic site along an axonal branch stabilizes that branch and protects it from subsequent retraction, as if synapses form 'spot welds' between axons and dendrites. Furthermore, new axonal branches were observed to emerge preferentially from sites of presynaptic specializations. It was proposed that iterative rounds of branch extension from existing presynaptic sites and selective branch stabilization by synapse formation at tips of new branches would guide axon arbor growth in a manner reminiscent of the 'synaptotrophic hypothesis' proposed by James Vaughn several decades ago (Figure 2).

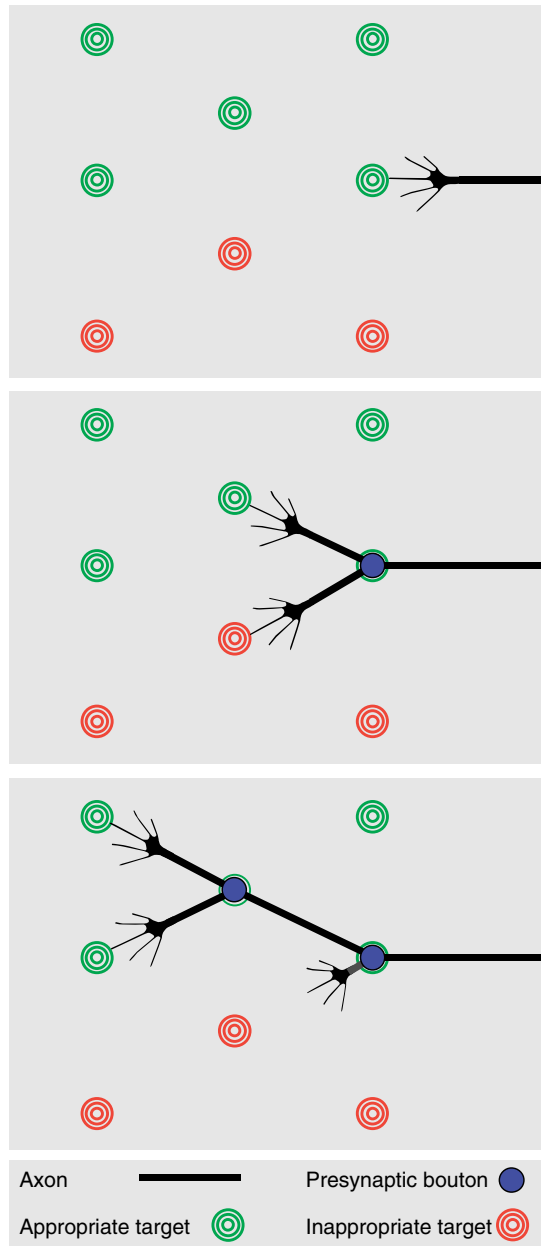

Figure 2 The synaptotrophic hypothesis as applied to axonal growth and presynapse formation. Axonal growth is governed by two principles: (1) the formation of a presynaptic specialization by an elongating axonal branch stabilizes this branch, whereas the failure to form such a connection entails subsequent branch retraction; and (2) new branches form preferentially from sites of newly formed presynaptic specializations. Iterative rounds of growth/retraction and branching according to these principles consequently result in the preferential establishment of presynaptic specializations within appropriate postsynaptic target regions.

Maturation of Nascent Presynaptic Sites

The differentiation of growth cones or patches of axonal membrane into well-formed functional presynaptic compartments is associated with significant structural changes. Unfortunately, these are mostly beyond the resolving power of light microscopy, including the cutting-edge microscopy techniques that have been so informative in documenting the choreography of synaptogenesis. Electron microscopy (EM) has therefore been used to examine the fine structure of synapses in samples from developing tissue in an attempt to determine the sequence of structural rearrangements that takes place as nascent contacts evolve into mature presynaptic specializations. Collectively, these studies suggest that nascent CNS synapses have parallel, apposed membranes, with varying degrees of structural specializations (electron-dense thickenings). Initially, their presynaptic compartments are very simple, essentially composed of a patch of electron-dense membrane (considered to be the AZ) associated with only a small number (two or three) of SVs. Reserve pools of SVs are virtually nonexistent, as are the mitochondria that are commonly observed in mature presynaptic boutons. On the other hand, pleomorphic vesicular structures, dense core vesicles, some with spicules projecting from their surface, as well as coated vesicles are often observed at nascent presynaptic sites. With development, SV numbers increase and boutons become larger, and the presynaptic membrane becomes more complex.

In addition to structural changes, presynaptic maturation is also associated with changes in functional characteristics. For example, SV exocytosis becomes confined to the presynaptic AZ region; changes occur in the types and subunit composition of voltage-dependent calcium channels involved in evoked neurotransmitter release. Moreover, the sensitivity of SV exocytosis to tetanus toxin is increased, whereas the sensitivity to the vesicle budding inhibitor Brefeldin A is reduced. Finally, changes in the probability of neurotransmitter release are observed. Although details of these maturational changes differ from one type of synapse to another and from one organism to another, it is of note that such changes are protracted, taking place over days or even weeks, at least in vertebrates.

Relationships between structural and functional aspects of synapse maturation are not well understood. However, a combined analysis based on optical methods (recycling of styryl dyes), electrophysiological recordings, and EM indicates that presynaptic boutons of rat hippocampal neurons in primary culture pass through three distinct states. At the onset of detectable synaptic function (about 5 days after plating), synapses lack readily releasable vesicles, although they possess a pool of recycling vesicles that can release neurotransmitters under strong stimulation. During the next 2 days, small pools of docked, readily releasable vesicles begin to appear (presumably reflecting the formation of AZs). Subsequently, reserve pools of SVs start forming, increasing their size concomitantly with a parallel increase in the docked, readily releasable pool size over the rest of the maturation period (several more days).

The long maturation process described previously is seemingly at odds with light microscopy-based observations made in developing neurons that indicate that functional presynaptic sites can form surprisingly fast – within 1 h or less of the establishment of a new axodendritic contact. In these studies, however, new presynaptic sites were recognized as such on the basis of (1) morphological criteria, such as the clustering of a fluorescently tagged presynaptic or SV proteins at the prospective presynaptic site; (2) functional assays, usually the capacity for evoked endocytosis and exocytosis of fluorescent endocytosis tracers such as the styryl dye FM4–64; (3) cytochemical data – that is, retrospective immunohistochemistry using antibodies against synaptic molecules; or (4) combinations of these. These criteria are rather crude in comparison to EM-based ultrastructural analysis, and so, in most of these studies, it remains unknown if the new presynaptic sites had acquired the structural characteristics typical of mature presynaptic terminals or, perhaps, represent nascent structures that are yet to undergo a prolonged maturation process over the days and weeks to come.

Ideally, one would like to monitor the formation of individual synapses in living neurons, determine their functional characteristics, and then examine their ultrastructure by EM. To date, such experiments have been performed in model systems: Experiments performed in *Xenopus* spinal neurons co-cultured with muscle cells revealed that excitatory postsynaptic potentials as well as increases in intra-axonal calcium concentrations can be recorded within seconds to minutes of nerve–muscle contact. During the next few hours, marked increases were observed in the frequency of spontaneous synaptic currents and in the amplitude of both spontaneous and evoked synaptic currents. Surprisingly, practically no well-formed synaptic specializations were found when the same nerve–muscle contact sites were scrutinized by EM, despite their marked capacity for synaptic transmission. In particular, no obvious AZ-like structures were detected. Instead, presynaptic regions contained scattered SVs, some dense core vesicles, and endocytotic profiles. More mature-like presynaptic specializations were observed only a day or so after nerve–muscle contact. Similar experiments performed in nascent CNS synapses formed between hippocampal neurons in primary culture reveal a similar situation – the absence of well-formed AZs or numerous typical SVs at nascent presynaptic sites within 2 or 3 h of their formation, even though a capacity for stimulation-evoked vesicle recycling was demonstrated at the same sites before fixation. Instead, numerous pleomorphic and tubulovesicular structures as well as dense core vesicles were observed. These findings, as well as others, indicate that the ultrastructure of nascent CNS presynaptic sites might be significantly different from that of mature presynaptic boutons.

When considering the issue of presynaptic maturation, it is important to remember that periods of synaptogenesis can last days, weeks, or even months (depending on the organism). As discussed previously, this period is characterized by enormous cellular dynamics during which axons and dendrites grow, branches extend and retract, resulting in the formation and elimination of synaptic contacts at great rates. This means that the population of presynaptic sites is heterogeneous in terms of synaptic age, making it unlikely that all new synapses progress synchronously through well-defined developmental stages such as those described previously. It is more likely that much of this maturation process reflects genetically programmed global changes in neuronal maturational states. The maturational state would dictate the availability and cellular distribution of synaptic building blocks, the dynamics exhibited by

axons and dendrites, the competence of axons and dendrites to evolve into mature synaptic structures, and the functionality of molecular pathways that are involved in coordinating and assembling mature presynaptic structures.

This point is illustrated by several examples. *Xenopus* spinal neurons form rudimentary synapses with co-cultured muscle cells within 1 day in culture. However, synapsin I, a very prominent presynaptic molecule, is not detected before day 3 in culture, nor is it detected *in vivo* during early phases of synaptogenesis. It was therefore suggested that the onset of presynaptic maturation is causally related to the onset of synapsin I expression. Similarly, the expression of dynamin and amphiphysin, two molecules that have crucial roles in SV endocytosis, is delayed considerably in comparison to SV proteins such as synaptotagmin and SV2. Given their roles in presynaptic endocytosis, it is conceivable that the delayed expression of these molecules could be related to the late formation of SV reserve pools at new presynaptic sites.

Another global change that occurs with development is a gradual reduction in axonal and dendritic motility: In fact, *in vivo* imaging in mice (and in macaque monkeys as well) reveals that in adult animals, overall dendritic and axonal structure is remarkably stable, particularly when compared to the incredible dynamics observed during periods of intense synaptogenesis. Conceivably, this reduction in cellular dynamics may promote presynaptic maturation by increasing the likelihood that nascent presynapses will persist and evolve into mature presynaptic structures.

A phenomenon often observed during synaptogenesis is the presence of presynaptic structures that exhibit many ultrastructural or morphological hallmarks of bona fide presynaptic sites but exhibit practically no capacity for evoked neurotransmitter release. The reasons for this dysfunctional state are not known, although it is likely to reflect the presence of incompletely formed AZs or the lack of components essential for SV exocytosis. Interestingly, these presynaptically silent or mute synapses can be converted rapidly to functional presynaptic sites by elevating cAMP levels, by manipulating extracellular Ca^{2+} and Mg^{2+} levels, or by high-frequency stimulation of presynaptic neurons. This unsilencing seems to involve rapid actin polymerization mediated by brain-derived neurotrophic factor (BDNF) and the small GTPase cdc42 (a critical regulator of cytoskeletal rearrangements). Interestingly, the fraction of presynaptically silent synapses gradually diminishes over time, at a rate that parallels the stabilization of axonal and dendritic arbors, indicating that structural stabilization and presynaptic maturation are tightly interconnected developmental processes.

Cellular Mechanisms of Presynaptic Differentiation

Studies based on live imaging microscopy techniques have led to the realization that presynaptic development is a highly dynamic process, occurring over a time scale of minutes to hours. Some of the cellular mechanisms that underlie these dynamics have been characterized and will be discussed next.

As mentioned previously, axons of immature neurons display a capacity for evoked SV recycling along axonal segments, even in the absence of obvious targets. Closer examination has revealed the presence of mobile packets of vesicles or vesicular material that travel along axons at rates of up to 0.1–1 $\mu m\,s^{-1}$. These packets move intermittently in both directions, often pausing, sometimes splitting into smaller packets or merging into larger clusters that sometimes exhibit a capacity for depolarization-evoked exocytosis and endocytosis similar to bona fide synapses. In one study, carried out in cultured hippocampal neurons, mobile SVs (visualized by means of a GFP-tagged variant of the SV molecule VAMP2/Synaptobrevin2) were observed to accumulate rapidly at new axodendritic contacts sites that subsequently exhibited a capacity for stimulation-evoked SV recycling. Intriguingly, retrospective immunohistochemical analysis revealed that these packets were associated with additional presynaptic molecules, including voltage-dependent calcium channels, SV2, synapsin I, and amphiphysin. These findings have led to suggestions that the mobile packets observed in axons might be preassembled precursors ('prototerminals') involved in the rapid formation of true presynaptic structures.

The concept that vesicular structures might be used for presynapse formation was originally put forward several decades ago. As mentioned previously, electron micrographs of developing spinal cord synapses revealed the presence of dense core vesicles with spicules projecting from their surface. On basis of their structural similarity to AZ material and their proximity to presynaptic membranes, it was suggested that these might be involved in the delivery of materials necessary for presynaptic differentiation. In agreement with this idea, a previously unknown 80 nm dense core vesicle was shown to specifically contain multiple AZ components, including the CAZ scaffolding molecules Bassoon, Piccolo, and Rim, but not characteristic SV molecules. Later studies showed that these Piccolo/Bassoon transport vesicles (PTVs) are Golgi derived and that the formation of new functional SV release sites is preceded by the recruitment of two to five mobile Piccolo/Bassoon packets to these sites. These findings suggest that at

least some AZ proteins are incorporated into nascent presynaptic sites in a unitary fashion.

Time-lapse microscopy of SVs labeled with styryl dyes such as FM1–43 revealed that functional release sites occasionally give rise to small mobile vesicle clusters that moved along axons, sometimes merging with nearby presynaptic sites. Interestingly, such mobile clusters often display a capacity for SV release en route, even when they are not juxtaposed to postsynaptic structures or even dendrites, although they do contain the CAZ molecule Bassoon. These features suggested that these mobile structures might represent 'orphan' SV release sites. Intriguingly, orphan release sites were sometimes observed to give rise to new presynaptic sites at axodendritic contacts, indicating that new presynaptic compartments might sometimes be formed from 'units' of SV release machinery that 'bud off' preexisting synapses.

Presynaptic proteins are synthesized in the cell body and then 'shipped out' to axons by means of various axonal transport mechanisms. In principle, components of new presynaptic sites could arrive directly from somatic biosynthetic sources. However, studies such as those mentioned previously indicate that such components can also be 'pilfered' from nearby synapses. Intriguingly, studies show that typical SV membrane proteins (e.g., synaptobrevin-2/VAMP-2 and synaptotagmin-1) are also exchanged among neighboring presynapses by another route that involves SV fusion with the presynaptic membrane, lateral diffusion along the axonal membrane, and endocytosis at nearby presynaptic locations. These findings suggest that presynaptic proteins, SVs, and SV membrane proteins are, in fact, shared resources that are dynamically redistributed among or even competed over by neighboring presynaptic boutons, and that these pools may be drawn upon for the establishment of new presynaptic sites.

The finding that SV proteins are present in significant amounts on the axonal plasma membrane is relevant to another facet of presynaptic differentiation – the source of the numerous SVs found in presynaptic compartment. There seem to be two distinct sources for SVs. The first is the Golgi apparatus in the cell body, where SVs are formed and then transported into axons. These vesicles differ in several respects – both morphologically and in their molecular contents – from the typical clear core vesicles found at mature presynaptic sites, and they have thus been called SV precursors. It seems that such vesicles fuse with the axonal plasma membrane in growth cones and along its entire length in a relatively nonspecific manner. These fusion events, possibly mediated by the exocyst (a conserved protein complex essential for trafficking secretory vesicles to the plasma membrane), deposit SV membrane proteins in the axonal plasma membrane. This pool of SV membrane proteins is then used for the biogenesis of typical SVs by rounds of membrane protein sorting, clathrin-mediated endocytosis, endosomal processing, and local confinement to presynaptic compartments. This scenario may explain the presence of pleomorphic and tubulovesicular structures (SV precursors?) and the paucity of typical SVs observed at nascent presynaptic sites (as described previously) and also explain why the delayed expression of presynaptic molecules involved in SV endocytosis, biogenesis, and confinement would delay the formation of large SV pools at developing presynaptic sites.

The cellular events involved in presynapse formation may be summarized as follows (**Figure 3**): During early stages of neuronal differentiation, as axons are growing toward their targets, an inherent capacity for rudimentary SV recycling exists at axonal growth cones and along the axonal plasma membrane. Mobile packets of SVs, SV precursors, and AZ precursors travel up and down axons. As axons reach their target regions, fleeting axodendritic contacts, initiated by axonal and dendritic growth cones and filopodia, induce the clustering of such mobile packets at contact sites. Most of these contacts break up, and the transport packets disperse and renew their migration. As neurons mature, an increasing proportion of transient contacts do not break up and instead are stabilized by cytoskeletal rearrangements and cell adhesion interactions with postsynaptic membranes. These subsequently develop into functional presynaptic boutons, presumably by the fusion of PTVs with the presynaptic membrane, the recruitment of SV precursor transport packets, the recruitment of SV packets, and the local biogenesis of SVs. The new presynaptic sites are not completely stable, however, and occasionally 'orphan release sites' are uncoupled from their nascent postsynaptic partner and migrate to adjacent presynaptic sites or participate in the formation of new ones, presumably at sites of contact with dendritic structures. With time, genetically controlled processes stabilize axonal and dendritic arbors, reduce protrusive activity, and promote the structural and functional maturation of presynaptic structures.

Molecular Mechanisms of Presynaptic Differentiation

Mature presynaptic specializations are almost always found at sites of contact with postsynaptic targets, implying an exchange of signals that either induce the formation or promote the stabilization of presynaptic structures at these locations. Decades of research has resulted in the identification of numerous molecules

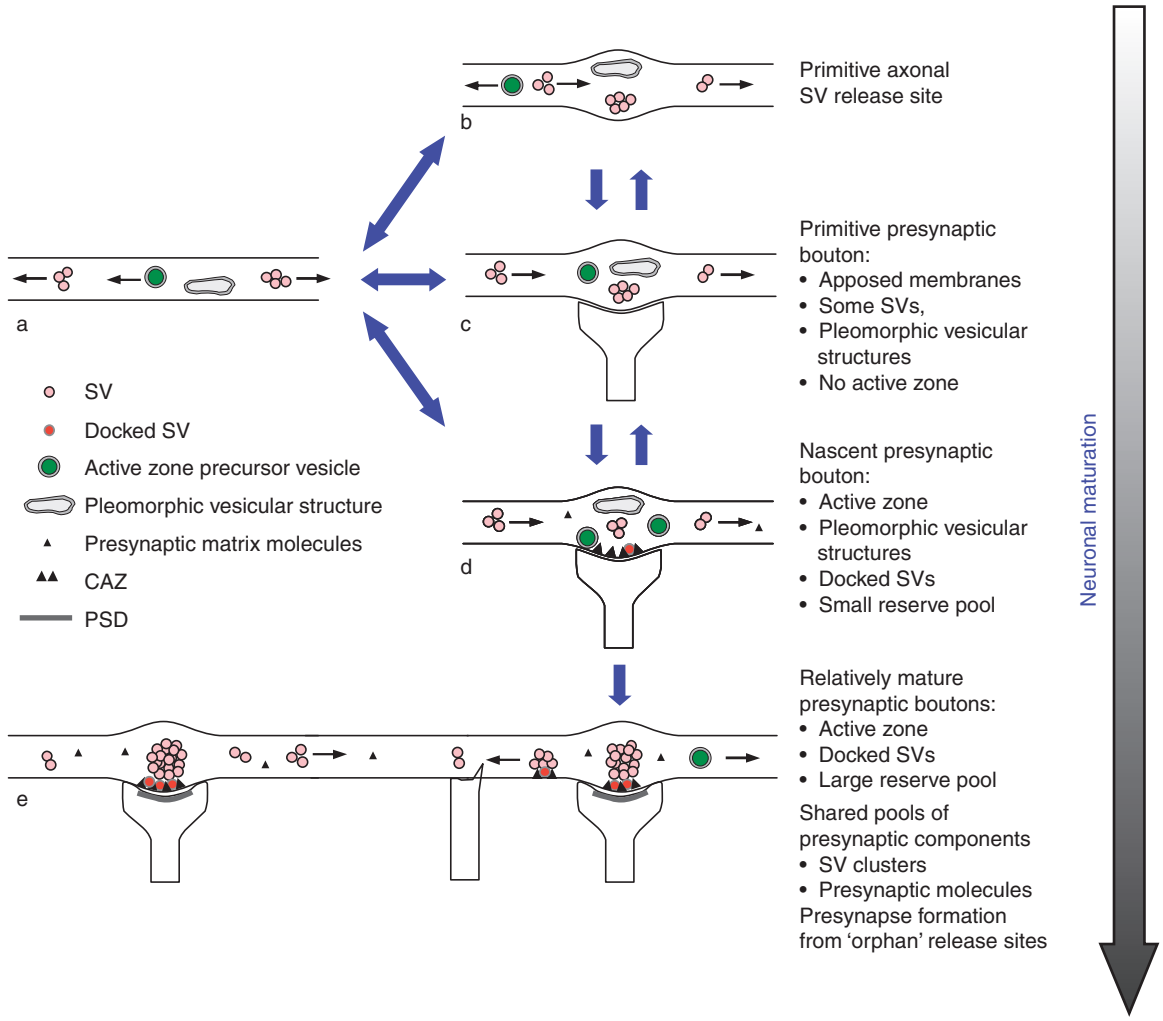

Figure 3 A model for presynaptic development. The formation of presynaptic specializations is envisioned to occur by multiple processes that take place over several time scales. (a) The axons of developing neurons contain several types of transport packets that are used for the assembly of nascent presynaptic structures. These include synaptic vesicle packets, pleomorphic tubulovesicular structures (synaptic vesicle precursors?), and active zone precursor vesicles such as PTVs. (b) In immature axons, primitive sites of synaptic vesicle recycling form spontaneously along axonal segments. (c) Contacts between axons and targets lead to the accumulation of various transport packets at contact sites, resulting in the formation of primitive presynaptic boutons. These sites display some capacity for synaptic vesicle recycling but lack the ultrastructural features of mature presynaptic boutons. (d) Alternatively, the establishment of an axodendritic contact might lead to the rapid formation of an active zone by the fusion of active zone precursor vesicles (such as PTVs) and the subsequent recruitment of synaptic vesicles. It is possible that this scenario becomes more common as neurons mature. (e) With time (days), synapses form larger synaptic vesicle reserve pools and gradually acquire the structural and functional characteristics of mature synapses. During this period, synaptic vesicles, plasma and synaptic vesicle membrane molecules, presynaptic matrix molecules, and even units of active zone material with cognate synaptic vesicles clusters (orphan release sites) are exchanged among nearby presynapses, sometimes giving rise to new presynaptic sites at nascent axodendritic junctions. It should be noted that uncertainty exists concerning some details of this scheme. For example, it is not known whether presynaptic sites form with a precise temporal order of events. Primitive presynaptic sites (c) could form first, followed by active zone formation (arrow from (c) to (d), or primitive sites could fall apart (arrow from (c) to (a) and reform later according to scheme (d) (arrow from (a) to (d). Similarly, axonal release sites (b) might form first and then become synaptic (arrow from (b) to (c) or eventually fall apart (arrow from (b) to (a). Adapted from Ziv NE and Garner C (2004) Cellular and molecular mechanisms of presynaptic assembly. *Nature Reviews Neuroscience* 5(5): 385–399.

involved in such bidirectional signaling. Yet, at the same time, it has become apparent that presynaptic differentiation is a dynamic, multistep process controlled by a host of molecular pathways. Thus, not all of the molecules identified so far are directly involved in specifying presynapse differentiation at specific locations and times, even though they do play roles in processes essential for the ultimate formation of presynaptic specializations, such as appropriate timing of presynaptic protein expression, axonal pathfinding, target recognition, arresting axonal growth motility, or stabilizing initial axon–target cell contacts. Unfor-

tunately, a comprehensive description of all molecules identified to date is beyond the scope of this article, and thus the focus is on a small number of relatively well-characterized molecules that do seem to be directly involved in presynapse formation induction.

Earlier work on NMJ formation led to the identification of Agrin, a large proteoglycan secreted by motor neurons that exhibits an ability to induce acetycholine receptor clustering beneath presynaptic axon terminals. This molecule was thus considered to be a major inducer of NMJ postsynaptic differentiation. Less is known about the reciprocal pathway – that is, the molecules responsible for inducing AZ formation in motor neuron terminals. However, laminin β_2, a component of the NMJ basal lamina, was shown to play an important role in presynaptic organization. Furthermore, it was shown to exert its effects via interactions with extracellular domains of presynaptic voltage-gated calcium channels, particularly P/Q and N type, the major subtypes found in NMJ AZs. Interestingly, however, laminin β_2 appears to be dispensable for the initial formation of AZs but is required for their subsequent stabilization. In this respect, it is also interesting to note that recent work implies that Agrin too is not required for acetylcholine receptor clustering *per se*, but it is required for counteracting neurotransmission-induced dispersion of these receptors. These findings highlight the importance of studying synaptic differentiation in a framework of molecular dynamics and stabilizing forces, rather than rigid molecular interactions and assembly processes.

Whereas secreted molecules clearly play crucial roles in NMJ formation, there are good reasons to believe that many of the bidirectional signals that pass between prospective pre- and postsynaptic partners rely on interactions between extracellular domains of axonal and target cell membranes proteins, particularly interactions mediated by various classes of cell adhesion molecules (CAMs). The cadherin family of calcium-dependent adhesion molecules has long been suspected of playing important roles in inducing presynaptic development, and there is evidence that this may be the case for some invertebrate synapses, such as those formed between photoreceptor cell axons and their targets in *Drosophila*. At vertebrate CNS synapses, however, classical cadherins do not seem to possess the capacity to induce presynaptic specialization on their own and appear to be more important for the stabilization of nascent presynaptic structures.

Several other classes of CAMs, however appear to perform similar functions. The best characterized is neuroligin, a member of the Ig superfamily of CAMs. Its presynaptic binding partners include members of the α- and β-neurexin families of cell surface proteins. Of these, the α-neurexins have been implicated in mediating the recruitment of voltage-gated calcium channels into the presynaptic AZ, whereas β-neurexins (essentially truncated α-neurexins) have been implicated in inducing AZ formation and SV clustering by actin polymerization promoted by their cytoplasmic, C-terminal tails or by interactions of these domains with presynaptic scaffolding molecules, such as CIPP, CASK/Lin-2, and Mint/Lin-10/X11α. When expressed in nonneuronal cells, Neuroligin can induce the formation of functional SV release sites in axons growing along these cells and this activity appears to require β-neurexin. When overexpressed in cultured hippocampal neurons, a large increase in morphologically identified synapses is observed. Conversely, RNAi-based knockdown of Neuroligin1–3 (the three functional forms in mice) in cultured hippocampal neurons reduces the number of morphologically identified synapses. Surprisingly, however, such increases or reductions in morphologically identified synapse number are not always associated with similar changes in functional synapse numbers as determined by miniature excitatory/inhibitory synaptic potential frequency. This puzzling discrepancy may indicate that additional molecules are required for inducing the formation of functional presynaptic compartments, a conclusion supported by the finding that normal numbers of synapses are formed in triple knockout mice lacking all three forms of murine neuroligin. It is worth mentioning, however, that the mice die soon after birth due to respiratory failure, which seems to be a consequence of reduced GABAergic/glycinergic and glutamatergic synaptic transmission in brain stem centers that control respiration.

A second molecule that exhibits a capacity to induce presynaptic specialization when expressed in nonneuronal cells is SynCAM, a member of the Ig superfamily of CAMs. Here, induction is via homotypic interactions with SynCAM present on both pre- and postsynaptic membranes. Interestingly, SynCAM overexpression in cultured hippocampal neurons leads to a large increase in functional synapse number but not an increase in morphologically identified synapse number – exactly the opposite of the phenotype observed for Neuroligin1. Furthermore, SynCAM effects seem to be specific for excitatory synapses, whereas Neuroligins seem to have a larger effect on inhibitory synapse formation.

Other cell adhesion molecule families, such as N-CAM and γ-protocadherins, have been implicated in controlling various aspects of presynapse development in vertebrate synapses. However, their roles and mode of action are not completely clear. In contrast, fascinating information concerning cell adhesion

molecules and presynaptic development has been obtained in invertebrate models, namely *Drosophila* and *Caenorhabditis elegans*, by harnessing the powerful genetic methods these systems offer. In one striking example from *C. elegans*, the expression of Syg2 in vulval epithelial cells (which are neither pre-nor postsynaptic cells) was shown to instruct hermaphrodite-specific neurons (HSNs) to form presynaptic specializations on specific motor neurons at the site where the three cells intersect. The effects of Syg2 on the HSNs is mediated by interactions with Syg1, both members of the Ig superfamily of CAMs, that are similar to the vertebrate molecules Nephrin and NEPH1. However, apparently normal presynaptic specializations still form in the absence of Syg1 or Syg2, albeit at erroneous locations, pointing to the difficulty in separating presynapse induction from more subtle roles such as a confinement of presynaptic structures to preferred locations.

Cell surface molecules that induce presynaptic differentiation do so by means of direct interactions with intracellular scaffolding/cytoskeletal molecules and by setting off various intracellular signaling cascades. Progress in identifying such molecular pathways has come from genetic screens in *Drosophila* and *C. elegans*. Such screens have led to the identification of several genes from several families that affect the morphology of presynaptic structures during development. For example, sad-1, a member of the microtubule affinity-regulating kinase (MARK) family, was found to be important for confining SVs to active zonal regions, whereas α-liprins were shown to control AZ size and complexity. Similarly, Pam/Highwire/Regulator of Presynaptic Morphology-1 (RPM-1) was found to play essential roles in organizing AZ and presynaptic structure, perhaps by direct interactions with other molecules but more likely by regulating multiple signaling cascades, including ubiquitination-dependent degradation of synaptic growth-promoting molecules and the MAP kinase pathway. Despite this progress, a satisfactory understanding of presynaptic differentiation in molecular terms is yet to be attained.

In summary, it seems that the induction of presynaptic differentiation involves multiple molecules and molecular cascades that act in parallel but probably differ in subtle ways. Even for molecules intimately linked to individual presynapse formation, it is not always clear if these actually trigger new presynapse formation or, rather, selectively promote the stabilization of presynaptic specializations that form spontaneously or migrate up and down axons in partially of fully assembled form. Further work is necessary to clarify the roles, interactions, and functional overlaps of these signaling cascades.

See also: Cell Adhesion Molecules at Synapses; Dendrite Development, Synapse Formation and Elimination; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Postsynaptic Development: Neuronal Molecular Scaffolds; Presynaptic Facilitation; Presynaptic Events in Neuromuscular Transmission; Receptor Trafficking; Spine Plasticity; Synapse Formation: Competition and the Role of Activity; Synaptic Precursors: Filopodia.

Further Reading

Ahmari SE, Buchanan J, and Smith SJ (2000) Assembly of presynaptic active zones from cytoplasmic transport packets. *Nature Neuroscience* 3: 445–451.

Antonov I, Chang S, Zakharenko S, and Popov SV (1999) Distribution of neurotransmitter secretion in growing axons. *Neuroscience* 90: 975–984.

Biederer T, Sara Y, Mozhayeva M, et al. (2002) SynCAM, a synaptic adhesion molecule that drives synapse assembly. *Science* 297: 1525–1531.

Buchanan J, Sun Y, and Poo MM (1989) Studies of nerve–muscle interactions in*Xenopus* cell culture: Fine structure of early functional contacts. *Journal of Neuroscience* 9: 1540–1554.

Dai Z and Peng HB (1996) Dynamics of synaptic vesicles in cultured spinal cord neurons in relationship to synaptogenesis. *Molecular and Cellular Neuroscience* 7: 443–452.

Jin Y (2002) Synaptogenesis: Insights from worm and fly. *Current Opinion in Neurobiology* 12: 71–79.

Kraszewski K, Mundigl O, Daniell L, Verderio C, Matteoli M, and De Camilli P (1995) Synaptic vesicle dynamics in living cultured hippocampal neurons visualized with CY3-conjugated antibodies directed against the lumenal domain of synaptotagmin. *Journal of Neuroscience* 15: 4328–4342.

Krueger SR, Kolar A, and Fitzsimonds RM (2003) The presynaptic release apparatus is functional in the absence of dendritic contact and highly mobile within isolated axons. *Neuron* 40: 945–957.

Kummer TT, Misgeld T, and Sanes JR (2006) Assembly of the postsynaptic membrane at the neuromuscular junction: Paradigm lost. *Current Opinion in Neurobiology* 16: 74–82.

Meyer MP and Smith SJ (2006) Evidence from *in vivo* imaging that synaptogenesis guides the growth and branching of axonal arbors by two distinct mechanisms. *Journal of Neuroscience* 26: 3604–3614.

Mozhayeva MG, Sara Y, Liu X, and Kavalali ET (2002) Development of vesicle pools during maturation of hippocampal synapses. *Journal of Neuroscience* 22: 654–665.

Nakata K, Abrams B, Grill B, et al. (2005) Regulation of a DLK-1 and p38 MAP kinase pathway by the ubiquitin ligase RPM-1 is required for presynaptic development. *Cell* 120: 407–420.

Nishimune H, Sanes JR, and Carlson SS (2004) A synaptic laminin–calcium channel interaction organizes active zones in motor nerve terminals. *Nature* 432: 580–587.

Ruthazer ES, Li J, and Cline HT (2006) Stabilization of axon branch dynamics by synaptic maturation. *Journal of Neuroscience* 26: 3594–3603.

Scheiffele P, Fan J, Choih J, Fetter R, and Serafini T (2000) Neuroligin expressed in nonneuronal cells triggers presynaptic development in contacting axons. *Cell* 101: 657–669.

Shapira M, Zhai RG, Dresbach T, et al. (2003) Unitary assembly of presynaptic active zones from Piccolo–Bassoon transport vesicles. *Neuron* 38: 237–252.

Shen K, Fetter RD, and Bargmann CI (2004) Synaptic specificity is generated by the synaptic guidepost protein SYG-2 and its receptor, SYG-1. *Cell* 116: 869–881.

Shen W, Wu B, Zhang Z, et al. (2006) Activity-induced rapid synaptic maturation mediated by presynaptic cdc42 signaling. *Neuron* 50: 401–414.

Takahashi T, Nakajima Y, Hirosawa K, Nakajima S, and Onodera K (1987) Structure and physiology of developing neuromuscular synapses in culture. *Journal of Neuroscience* 7: 473–481.

Varoqueaux F, Aramuni G, Rawson RL, et al. (2006) Neuroligins determine synapse maturation and function. *Neuron* 51: 741–754.

Vaughn JE (1989) Fine structure of synaptogenesis in the vertebrate central nervous system. *Synapse* 3: 255–285.

Yoshihara M, Rheuben MB, and Kidokoro Y (1997) Transition from growth cone to functional motor nerve terminal in *Drosophila* embryos. *Journal of Neuroscience* 17: 8408–8426.

Zhai RG, Vardinon-Friedman H, Cases-Langhoff C, et al. (2001) Assembling the presynaptic active zone: A characterization of an active one precursor vesicle. *Neuron* 29: 131–143.

Ziv NE and Garner CC (2004) Cellular and molecular mechanisms of presynaptic assembly. *Nature Reviews Neuroscience* 5: 385–399.

Presynaptic Development: Functional and Morphological Organization

D Atasoy and E T Kavalali, The University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Structural Features of Central Synapses

Chemical synapses are intercellular junctions critical for information transfer and processing in the nervous system. They consist of two compartments physically juxtaposed within several nanometers of each other: presynaptic terminals and postsynaptic specializations. Presynaptic terminals store and release neurotransmitter substances in membranous organelles named synaptic vesicles, whereas postsynaptic structures contain signaling molecules responsible for generation of neuronal responses to released neurotransmitters. Presynaptic terminals are highly organized subcellular structures. At the electron microscopic level, clusters of synaptic vesicles around the plasma membrane regions called active zones can readily distinguish them from other structures within a neuron. Synaptic vesicle exocytosis is thought to take place exclusively at the active zone, whereas synaptic vesicle endocytosis may occur within the vicinity of this region. Active zones are characterized by enrichment of scaffolding molecules, and enable assembly of proteins required for regulated vesicle fusion and recycling. Juxtaposed to the presynaptic terminal, the postsynaptic site is characterized by electron-dense material called the postsynaptic density. Postsynaptic density is enriched in scaffolding molecules that anchor neurotransmitter receptors and organize signaling in response to second messenger cascades activated by the neurotransmitter receptors. The pre- and postsynaptic sides of the synapse are held together with adhesion molecules spanning the synaptic cleft (Figure 1). The size of the active zone and the number of docked vesicles are critical determinants of the functional responses of a presynaptic terminal. These structural markers are continually modified during synapse maturation. Depending on the type of synapse, presynaptic terminals in a given synapse contain varying number of synaptic vesicles, some of which are physically attached or docked at the plasma membrane.

The most striking difference of synapses from other cell–cell junctions is the asymmetry of structures on both sides of the synaptic junction. Such asymmetry implies that two compartments must respond differently to the signal(s) that initiate synaptogenesis. This asymmetry is partially achieved through differential distribution of synaptic components to axonal and dendritic compartments within a neuron. Asymmetric interaction of cell adhesion molecules can also account for triggering divergent cascades of downstream events and induction of pre- and postsynaptic sites. However, it is important to note that despite this asymmetry the sizes of the structures on both sides of the synaptic cleft are all correlated, suggesting that the structural synapse assembly is significantly coordinated across the cleft.

In a mature presynaptic terminal, vesicles can be divided into two pools. The first pool contains a relatively small fraction of vesicles close to release sites. These vesicles can be released by brief Ca^{2+}-dependent stimuli or by hypertonic stimulation, which is Ca^{2+} independent. This release-ready pool of vesicles is referred to as the immediately releasable pool or the readily releasable pool (RRP). RRP vesicles are considered to be in a morphologically docked state, although not all morphologically docked vesicles are necessarily release competent at any given time. A priming step in addition to the morphological docking is required to make vesicles fully release competent. A secondary pool of vesicles, the reserve pool, is spatially distant from the release sites and constantly replaces the vesicles in the RRP that have been exocytosed. The rate of replenishment of RRP vesicles from the reserve pool is a critical parameter that determines the response of synapses to repetitive stimulation. Recent evidence indicates that intrasynaptic Ca^{2+} can facilitate the rate of replenishment. The number of vesicles contained in the RRP is a critical parameter that regulates the probability of release, which is defined as the probability that a presynaptic action potential can result in an exocytotic event. Therefore, the number of vesicles in the RRP and the rate and pathways by which they are replenished is a crucial determinant of presynaptic efficacy and of several forms of short- and long-term synaptic plasticity. Several lines of evidence support the presence of the non-recycling pool of vesicles in the synapse. Mechanisms that can render this pool functional remain to be determined.

Multiple Stages of Synapse Assembly

In the mammalian central nervous system (CNS), synapse formation is a precisely timed process. Synapses appear within days in a given brain region. At later stages of development, synapse proliferation

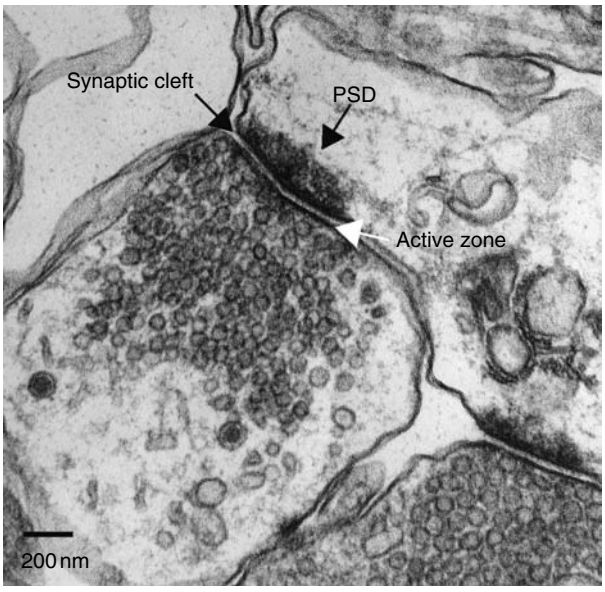

Figure 1 Structural organization of a central synapse. In electron micrographs, synapses can be readily distinguished by the cluster of synaptic vesicles around protein-rich (thus electron-dense) regions called the active zone. The size of the active zone is tightly correlated with the size of postsynaptic density, another protein-rich region associated with the postsynaptic dendrite. Synaptic cell adhesion molecules structurally linking the two sides of the synapse span the synaptic cleft between the presynaptic and postsynaptic regions. These molecules are thought to coordinate synapse maturation in pre- and postsynaptic compartments and also enable functional interactions between the two sides of the synapse. PSD, postsynaptic density.

is thought to be balanced by synapse elimination and pruning of synaptic contacts through activity-dependent mechanisms. Several lines of evidence suggest that synapse formation *per se* does not require neuronal activity. Initial events that establish immature synaptic contacts in neuromuscular junctions involve the interaction of axonal growth cones with target muscle membrane. In the central synapses, however, initial synapse formation is thought to take place between the axonal shaft and filopodial processes that extend from the dendrites. This type of interaction results in the formation of *en passant* synaptic boutons along the axonal shaft, which is a common feature of most CNS synapses. Retraction and stabilization of these filopodial processes together with contacted axonal regions or nascent presynaptic terminals marks the beginning of synapse maturation. Further maturation involves structural modifications that increase the anatomical complexity of the synaptic boutons, including an increase in the number of synaptic vesicles, the size of the synaptic boutons, and, in some cases, the number of active zones. An interesting aspect of synapse maturation is the matching change in the pre- and postsynaptic regions that results in a strong correlation between the size and complexity of both sides of a synapse.

The sequence of events leading to synapse formation has recently been studied in detail using time-lapse imaging techniques in dissociated hippocampal cultures. These studies took advantage of action potential-dependent, rapid synaptic vesicle recycling as the earliest indicator of synaptogenesis following the initiation of axo-dendritic contacts. These findings support the scenario that prepackaged presynaptic molecules are rapidly released at sites of axo-dendritic contact, forming functional presynaptic terminals. However, it has been previously shown that isolated synaptic vesicles in axons can also recycle in an activity-dependent manner prior to target contact. This immature form of synaptic vesicle recycling proceeds with slower kinetics compared to mature synapses. In some cases, this immature form of synaptic vesicle recycling has been shown to be resistant to tetanus toxin implicating the requirement for a vesicular SNARE (soluble *N*-ethylmaleimide-sensitive factor attachment protein receptors) protein other than synaptobrevin-2 (VAMP-2), which is tetanus toxin sensitive. In the developing neuromuscular junction, unlike mature junctions, synaptic vesicle recycling is highly sensitive to Brefeldin A, which disrupts synaptic vesicle trafficking thorough endosomal intermediates.

Maturation of presynaptic terminals occurs in structurally and functionally distinguishable stages (**Figure 2**). During early synapse maturation, synapses are unresponsive to action potential stimulation or hypertonicity (which normally induces swift fusion of docked vesicles), although they can release neurotransmitters and recycle synaptic vesicles during strong stimulation such as induced by elevated extracellular potassium. This form of vesicle recycling detected in these nascent synapses is mechanistically analogous to the form observed in the absence of critical components of the synaptic SNARE machinery. This is consistent with the observation that synaptic vesicle recycling at this stage is tetanus toxin insensitive. Following this initial stage, synapses undergo a transition to become responsive to action potential stimulation and rapidly recycle synaptic vesicles. Examination of electron micrographs of nascent synapses reveals a strong correlation between this functional switch and the formation of the active zone, leading to the assembly of the RRP.

Indeed electron microscopic analysis of immature hippocampal cultures shows that synaptic vesicles are not as closely associated with plasma membrane as they are in mature synapses. These vesicles recycle in a calcium-dependent manner as they travel in the axon, indicating that basic machinery needed for

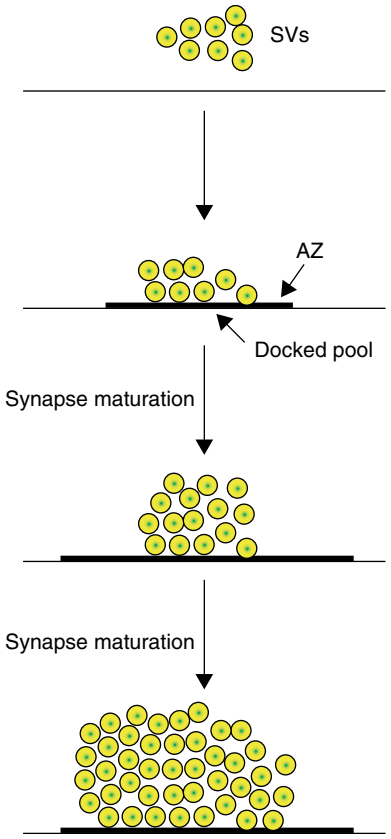

Figure 2 Sequence of events leading to maturation of presynaptic terminals during synaptic development. Nascent synapses typically contain synaptic vesicles that recycle only in response to strong stimulation. However, at these early stages of synapse assembly, vesicles are not tightly associated with the plasma membrane, presumably due to the immature state of active zones. These synapses lack a set of readily releasable vesicles; thus, they do not effectively respond to presynaptic action potentials. Formation of an active zone coincides with functional maturation of a presynaptic terminal. This stage may also involve actin cytoskeleton increasing association of the synaptic vesicle clusters with the active zone and the surface membrane release machinery. After these initial stages, synapse maturation involves a gradual increase in the size of the total vesicle pool, which also reflected an increase in the number of vesicles available for release. SV, synaptic vesicle; AZ, active zone.

docking and fusion is distributed loosely along the axon. Nevertheless, these hot spots of glutamate release sites may have an important role during the initial stages of synaptogenesis. The released glutamate can stimulate filopodial motility of both dendrites and axons, thereby increasing the chance of axo-dendritic encounter. However, in more mature cultures, glutamate strongly inhibits filopodial motility and stabilizes connections. These contradictory observations can be reconciled if during development changes in protein expression make filopodia less responsive to glutamate, or, alternatively, there is a concentration threshold for glutamate above which its activity has opposite effects.

Besides glutamate, many other secreted molecules increase the number and motility of dendritic and axonal filopodia prior to contact. These include brain-derived neurotrophic factor (BDNF) and neurotrophin-3 (NT-3), which have been shown to promote axonal arborization, dendritic growth, and synapse maturation. BDNF-coated beads can increase Ca^{2+} levels and trigger neurotransmitter release in contacting axons in a protein synthesis-dependent manner. Some members of the fibroblast growth factor (FGF) family of proteins (e.g., FGF22, FGF7, FGF10) can induce presynaptic organization via FGF receptor-2. In a recent study, Sanes and colleagues used a combination of chromatographic steps to isolate factors that cause axon branching and vesicle aggregation in chick motorneurons and purified FGF22 as an active component. Secreted Wnt proteins (Wnt-7a, Wnt-3) can also induce remodeling of growth cones and accumulation of synaptic vesicles. Wnt-7 knockout mice show a delay in maturation of multisynaptic glomerular rosettes formed between mossy fibers and granule cells in the cerebellum. Deficiency of a Wnt homolog in *Drosophila*, Wingless (Wg), causes abnormal pre- and postsynaptic differentiation indicating possible conservation of function.

The peak of synapse formation in the mammalian brain coincides with extensive proliferation of glia, in particular astrocytes. Astrocytic protrusions engulf potential contact sites between axonal and dendritic processes, implicating their possible role in synaptogenesis. Astrocytes release several factors that impact synaptic differentiation. Initial studies by Pfrieger and Barres showed that in the absence of glia the number of synapses formed between retinal ganglion cells in culture are reduced and individual synapses are less efficient in neurotransmitter release. Later work by Pfrieger and colleagues led to isolation of Apolipoprotein E (Apo-E) as the factor upregulated in the presence of glia. Apo-E is a carrier for cholesterol. In these experiments, application of cholesterol to retinal cultures caused a massive increase in spontaneous excitatory postsynaptic currents and the number of presynaptic nerve terminals. In parallel, Barres and colleagues identified thrombospondins 1 and 2 as additional glia-derived synaptogenic molecules, which specifically cause an increase in the number of synapses. Thrombospondins are multidomain extracellular matrix proteins, initially identified in platelet activation.

Role of Synaptic Cell Adhesion Molecules in Synapse Assembly

Whereas soluble factors play an important role in mediating the initial stages of vesicle aggregation and

priming axons for synapse assembly, synaptic cell adhesion molecules mediate the physical contact and functional communication between axonal and dendritic protrusions leading to the formation of synaptic junctions. These molecules are composed of several large families, which include N-cadherins, protocadherins, neural cell adhesion molecules (NCAMs), nectins, neurexins, and neuroligins. In most cases, extensive alternative splicing and differential glycosylation patterns create enormous variability in the possible repertoire of protein products. This high level of variety in individual protein products and the large number of combinatorial possibilities for intermolecular interactions between these molecules may contribute to the specificity of synaptic connections in the brain. However, the process of synapse formation itself seems to be somewhat promiscuous as evidenced by three observations. First, injury or degeneration in the brain can trigger extensive synaptic rewiring, which leads to formation of ectopic synapses between cells that do not normally make synapses with each other. Second, neurons in dissociated cultures form synapses rather promiscuously, where, in some conditions, cells can even form autapses with themselves regardless of their proper *in vivo* partners. Finally, axonal contacts onto polylysine-coated glass beads can induce assembly of presynaptic specializations. Taken together, these results indicate that the basic mechanism for synapse formation is inherent to all neurons. Furthermore, these observations suggest that synaptogenesis is not an event triggered by a single molecule, but rather a vast repertoire of molecular interactions that can lead to synaptogenesis between neurons. Activity-dependent processes test these connections over time and direct stabilization of the most resilient connections.

Neuroligin is the first adhesion molecule identified to be a direct inducer of presynaptic terminal assembly. When expressed in nonneuronal cells, they can cause an accumulation of presynaptic vesicle clusters on contacting axons. These synapses are fully functional as demonstrated by fluorescence imaging of synaptic vesicle recycling and electrophysiological detection of neurotransmitter release. The soluble extracellular domain of neurexin can mask this effect of neuroligin, indicating that neurexin is the mediator of the synapse-inducing activity of neuroligins within the presynaptic axon. An attractive part of the neurexin–neuroligin system is that it satisfies the necessary asymmetry required to induce different signaling events on pre- and postsynaptic sides of the synapse. Neurexins are primarily associated with the presynaptic site and intracellularly bind to calcium/calmodulin-dependent serine protein kinase (CASK) and syntenin. In contrast, neuroligins are located on the postsynaptic site and their C-termini interact with PSD-95.

SynCAM is the only other molecule shown to be sufficient to induce synapse formation when expressed in nonneuronal cells. It is a member of the immunoglobulin (Ig) superfamily, and mediates homophilic interaction through extracellular Ig domains. Similar to neuroligin, synapses induced *in vitro* by SynCAM are fully functional. Overexpression of SynCAM in dissociated cultures has a dominant positive effect in the functioning and formation of synapses. In contrast, the overexpression of dominant negative SynCAM lacking extracellular domain impairs presynaptic assembly. Unlike the neurexin–neuroligin interaction, SynCAM mediates a homophilic interaction, indicating that different downstream cues are present on different sides of the synapse (**Figure 3**).

How does the interaction of cell adhesion molecules translate into changes within the cell? Following initial contact, assembly of synapses takes 1–2 h and may occur in either of two ways. First, following contact, each molecule could be captured from the stream of axonal cytoplasm, and a synapse can be built depending on protein–protein interactions. A second model suggests that synaptic molecules are pre-assembled in small units in other parts of the neuron and transported to the axon. The speed of assembly favors a pre-assembled trafficking model. The evidence for the existence of cytoplasmic transport packages came from studies of Garner and colleagues. They have been able to isolate large dense vesicles from developing axons containing active zone proteins such as piccolo, bassoon, and RIM. Pre-assembled packages of active zone vesicles fuse with the axonal plasma membrane to create a scaffolding framework for other components. In addition to the active zone components, synaptic vesicles also assemble as clusters in a unitary fashion. Nevertheless, these two components do not seem to travel together along the axon. There is also evidence that at the end of this initial stage of synapse assembly, synaptic vesicles switch to a docked state and associate more closely with the plasma membrane in a synapsin-dependent manner.

Analysis of the intracellular interactions of neurexin and SynCAM reveals CASK as a converging downstream target. CASK is a member of the membrane-associated guanylate kinase family (MAGUK) and strongly interacts with neurexin and SynCAM cytoplasmic tail. Upon interaction with neuroligin, neurexin oligomerizes and recruits CASK. Given the multidomain structure of CASK, it is usually envisaged as a recruiter to the newly formed contact sites for both neurexin and SynCAM. CASK also

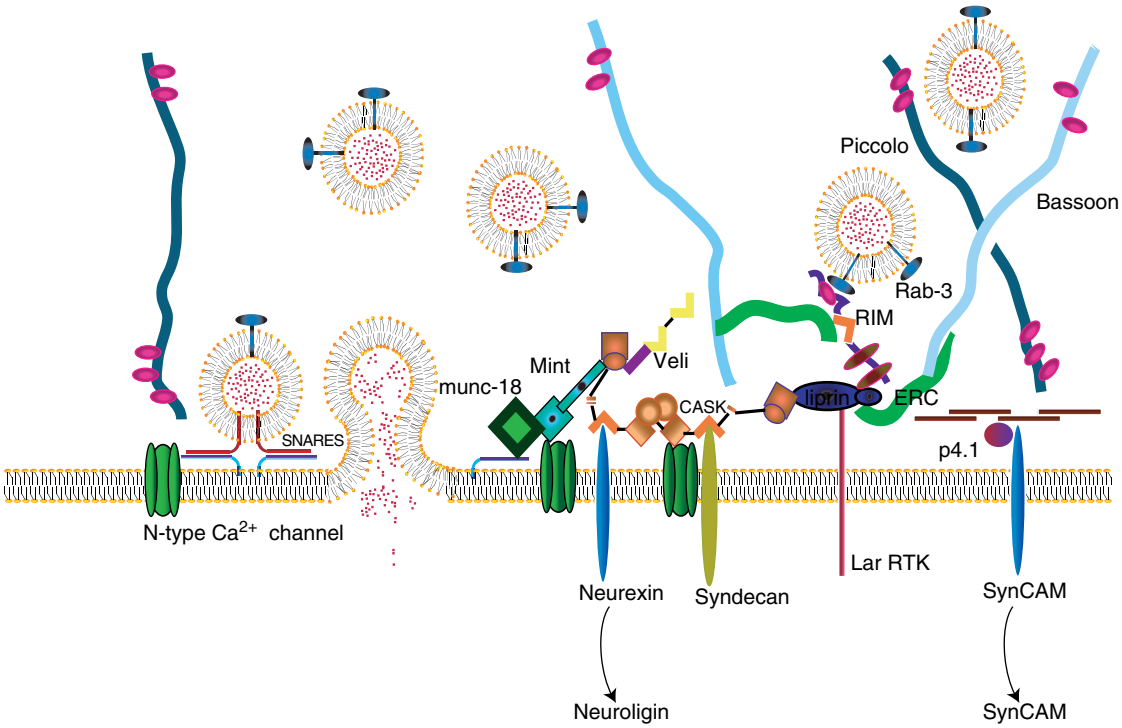

Figure 3 Putative molecular interactions that link synaptic cell adhesion molecules to the active zone and to the neurotransmitter release machinery. Cytoplasmic tails of neurexins and SynCAM can interact with a modular adaptor protein CASK which in turn binds another adaptor protein Mint interacting with the release machinery (munc-18 which interacts with syntaxin) as well as the voltage-gated calcium channels. Active zone components such as RIM can interact with synaptic vesicle-associated molecules, which include rab3, thus bridging synaptic junctional scaffold to the synaptic vesicle cluster.

forms a well-conserved tripartite complex with Mint and Veli, multidomain PDZ molecules. This complex is proposed to be responsible for the recruitment of vesicle fusion machinery. CASK has also been shown to interact with liprin, which organizes the presynaptic active zone in *Caenorhabditis elegans*. By tightly interacting with the active zone proteins RIM and ERC, liprins constitute the insoluble backbone of the active zone.

As mentioned above, synaptic vesicles are tethered at the vicinity of the active zone by the actin cytoskeleton. Actin depolymerizing agents have a strong disruptive effect on nascent synapses but not on mature synapses, implying a role for actin during synapse formation. CASK can polymerize actin on the neurexin C-tail and stabilize it by interacting with protein 4.1. In this way, synaptic adhesion molecules neurexin and SynCAM (and possibly other CASK-interacting adhesion molecules such as syndecans) can induce local polymerization of actin at contact sites and trap traveling synaptic components.

As discussed above, free-moving vesicle clusters have different cycling properties than mature ones. How maturation changes vesicle identity is not known. It could be achieved through transport of mature vesicles or conversion of the identity of existing vesicles. One can speculate that these initial synaptic vesicles are still present in mature synapses but in a reduced capacity for synaptic vesicle recycling. This may explain the presence of the enormous number of vesicles at synapses, while only a fraction of them are functional.

There are a multitude of possible pathways that can lead to eventual assembly of synaptic terminals. This redundancy can increase the robustness of the synapse assembly process and also contribute to the functional and structural versatility of synapses.

Stabilization of Synapses

Synapse formation is rather error prone at the initial stage, and a certain degree of mismatch often occurs. Therefore, initial promiscuous synapses either are usually secured by help of additional and more specific adhesion molecules, or are eliminated. Like previous stages, this process also occurs in a hierarchical manner. The synaptogenic molecules neurexin and neuroligin could be good candidates for helping to achieve this specificity. Neurexins undergoes extensive splicing, and recently Boucard and colleagues demonstrated that they interact according to a splice code. However, several other cell adhesion molecules

are also postulated to play a role in late stages of synaptic development.

One of the best-characterized synaptic cell adhesion proteins is N-cadherin. Similar to neurexins, they link the extracellular adhesive function to actin cytoskeleton via α- and β-catenins. Even though the function of cadherins at the synapse is not clear, several recent experiments provide significant insights. Overexpression of an N-cadherin construct lacking the extracellular domain while maintaining the ability to bind cytosolic partners markedly reduced the number of presynaptic boutons, indicating the importance of the adhesive function. In another set of experiments, expression of mutant α-N-catenin prevented the interaction of cadherin with the actin cytoskeleton but did not strongly affect presynaptic assembly. These experiments imply that cadherins have a rather adhesive function during early synaptogenesis. Therefore, they could act at an intermediate stage between initial contact and final maturation by prolonging the brief lifetime of axo-dendritic contacts. Disruption of cadherin function in mature synapses, however, does not have a strong effect. This finding raises the possibility that some of the cadherin functions are redundant with protocadherins, a subset of the cadherin superfamily. Protocadherins are composed of nearly 60 members expressed by three gene clusters that are expressed in distinct patterns in the nervous system and undergo extensive splicing.

Molecular Components of the Presynaptic Active Zone and the Cytomatrix

Active zones are the principal sites of synaptic vesicle fusion in synapses. The molecular components of the active zone are thought to serve a structural role by clustering synaptic vesicles around the active zone and increasing proximity between molecules on the synaptic vesicle membrane and the plasma membrane. Active zone proteins are also involved in priming the vesicles for release and perhaps in vesicle retrieval after fusion. Proteins, such as Bassoon and Piccolo, are recruited to activate the synapses during synaptogenesis. For instance, in experiments conducted by Garner, Ziv, and colleagues, Bassoon was detected in nascent synapses capable of action-potential-dependent uptake and release of FM dyes. In addition, dense core active zone precursor vesicles contain multiple synaptic proteins, including the active zone proteins, Bassoon and Piccolo. Fusion of these vesicles with the plasma membrane can rapidly assemble active zone. Despite extensive data on their localization, the functional properties of these active zone proteins are still unclear. However, for RIM1 and CASK, there are several well-characterized biochemical interactions with multiple proteins. As discussed above, in the case of CASK, these molecular interactions suggest a central role in the bridging of neurexins to munc-18, a critical component of presynaptic release machinery. Recent mouse knockouts of munc-13 and RIM1 uncovered critical functional roles for these molecules. Synapses deficient in munc-13-1 are severely impaired in their function. The remaining munc-13-2-dependent synaptic transmission displays marked synaptic facilitation. RIM1 knockout mice, on the other hand, have a less severe but significantly altered properties of short- and long-term plasticity. Interestingly, loss of these molecules does not lead to structural alterations in the synapse, which are presumably due to the redundancy of molecular interactions that assemble synapse structure.

How do the active zone proteins regulate synaptic function? This regulation is likely achieved by the ability of active zone proteins to recruit the components of fusion machinery, such as SNAREs and munc-18. An important step in the chain of events leading to vesicle fusion is the formation of the SNARE core complex between target membrane SNARE proteins (i.e., syntaxin and SNAP-25) and the synaptic vesicle SNARE, synaptobrevin/VAMP. Active zone proteins can exert significant functional effects by regulating the formation and dissociation of SNARE complexes. Replenishment of vesicles released at the active zone requires SNARE core complex assembly and disassembly. This assembly process in the synapse is much faster than the rates of SNARE core complex assembly *in vitro*. Therefore, the assembly process is most likely facilitated by protein–protein interactions between the components of the presynaptic active zone and synaptic vesicles. For instance, munc-18, a protein required for fusion, could be recruited to the active zone through its interaction with Mint (munc-18 interacting protein), which in turn binds to CASK.

Functional Maturation of Presynaptic Terminals and the Role of Activity

Following the initial assembly of synaptic terminals, a large number of synapses are functionally silent. In some cases, these functionally silent synapses can be rendered operational in response to activity. The most commonly studied models of silent synapses propose a postsynaptic mechanism that underlies this silence. According to this model, a fully functional presynaptic terminal may exist but the postsynaptic site does not possess α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors although it contains *N*-methyl-D-aspartate (NMDA) receptors. Activity, in turn, induces the insertion of functional AMPA receptors, making silent synapses functional under physiological conditions. In contrast,

studies in dissociated hippocampal cultures have also identified a developmental stage where synapses are presynaptically silent after assembly. This model possesses the same apparent features of an NMDA-only synapse, but the NMDA-only nature of these synapses is explained by a presynaptic mechanism and not by lack of postsynaptic AMPA receptors. According to this scheme, activation of postsynaptic AMPA or NMDA receptors can be determined by the kinetics of fusion pore opening and the release profile of glutamate. In young nerve terminals, neurotransmitter release occurs through a narrow fusion pore leading to exclusive activation of NMDA receptors as they have a higher affinity for glutamate. Synapse maturation in turn leads to an increase in preponderance of full fusion events, thus activating NMDA as well as AMPA receptors. An alternative model suggests that immature synapses do not readily respond to action potential stimulation leading to a full presynaptic neurotransmitter failure due to some inadequacy in fusion competence or localization of synaptic vesicles. This model is consistent with the previously discussed findings on the gradual reorganization of synaptic vesicle clusters after synaptogenesis. In a recent study, Shumin Duan and colleagues showed that a burst of action potentials can rapidly awaken these silent synapses by increasing the availability of synaptic vesicles for fusion through BDNF-triggered presynaptic actin remodeling mediated by the small GTPase Cdc42.

In most synapses, the initial structural assembly and functional unsilencing is followed by a gradual maturation process that typically involves alterations in short-term plasticity. Studies conducted in acutely isolated brain slices have described functional alterations that are solely of presynaptic origin. One finding of these experiments was an apparent decrease in release probability during synaptic development. This result is rather surprising, given the prevalent structural observation that the number of vesicles within a synapse increases during maturation implying an increase in synaptic reliability and release. Another interesting observation in cortical as well as hippocampal mossy fiber synapses is target-dependent alterations in short-term and some long-term forms of plasticity during the course of development. Cellular mechanisms underlying these developmental changes in short-term plasticity are postulated to involve altered Ca^{2+} dependence of fusion and regulation of vesicle mobilization in presynaptic terminals. It is tempting to speculate that synaptic cell adhesion molecules either individually or in combination may regulate these target specific functional alterations in the output of single neurons.

As exemplified by the presynaptic unsilencing process discussed above, several aspects of synaptic functional maturation during early development can be influenced by activity and neuromodulators. Most neuronal networks exhibit spontaneous action potential firing patterns and synaptic potentials in the absence of extrinsic influences. The background activity that arises from the properties of individual neurons and their characteristic synaptic connections has been shown to be critical for the refinement of synaptic connectivity within the nervous system. Most of the signaling cascades that play a role in synapse maturation can be physiologically activated or regulated by the background activity. These include several signal transduction pathways, including Ca^{2+}-signaling mechanisms and the activation of protein kinase C (PKC) and protein kinase A (PKA). For example, direct involvement of cyclic adenosine monophosphate (cAMP)-dependent signaling in synaptic development was demonstrated in hippocampal slices, as well as at the level of individual synapses in culture. Activation of Ca^{2+}, cAMP, or diacylglycerol second messenger cascades can be triggered either directly by neuronal activity through Ca^{2+} influx or indirectly by the release of glutamate and activation of metabotropic glutamate receptors.

Recent studies have shown that chronic alterations in spontaneous activity levels modify several synaptic properties including the size of postsynaptic responses, probability of neurotransmitter release, as well as the number of synapses. These experiments strongly support a role for background activity in regulating the proper functional maturation of individual synapses. Although activity is an indispensable component of synaptic development, the mechanism through which it influences synapse maturation and the elimination process is still elusive.

See also: Activity-Dependent Remodeling of Presynaptic Boutons; Cadherins and Synapse Organization; Cell Adhesion Molecules at Synapses; Endocytosis and Presynaptic Scaffolds; Exocytosis: Ca^{2+}-Sensitivity; Neurexins; Neuroligins; Piccolo and Bassoon; Postsynaptic Density/Architecture at Excitatory Synapses; Presynaptic Development and Active Zones; Presynaptic Regulation by Liprins; SNAREs; Synaptic Vesicles; Vesicle Pools.

Further Reading

Ahmari SE, Buchanan J, and Smith SJ (2000) Assembly of presynaptic active zones from cytoplasmic transport packets. *Nature Neuroscience* 3: 445–451.

Akins MR and Biederer T (2006) Cell–cell interactions in synaptogenesis. *Current Opinion in Neurobiology* 16: 83–89.

Biederer T, Sara Y, Mozhayeva M, et al. (2002) SynCAM, a synaptic adhesion molecule that drives synapse assembly. *Science* 297: 1525–1531.

Craig AM, Graf ER, and Linhoff MW (2006) How to build a central synapse: Clues from cell culture. *Trends in Neurosciences* 29: 8–20.

Goda Y and Davis GW (2003) Mechanisms of synapse assembly and disassembly. *Neuron* 40: 243–264.

Mozhayeva MG, Sara Y, Liu X, and Kavalali ET (2002) Development of vesicle pools during maturation of hippocampal synapses. *Journal of Neuroscience* 22: 654–665.

Scheiffele P (2003) Cell–cell signaling during synapse formation in the CNS. *Annual Review of Neuroscience* 26: 485–508.

Shen W, Wu B, Zhang Z, et al. (2006) Activity-induced rapid synaptic maturation mediated by presynaptic cdc42 signaling. *Neuron* 50: 401–414.

Sudhof TC (2001) The synaptic cleft and synaptic cell adhesion. In: Cowan WMST and Stevens CF (eds.) *Synapses*, pp. 275–313. Baltimore: The Johns Hopkins University Press.

Ziv NE and Garner CC (2004) Cellular and molecular mechanisms of presynaptic assembly. *Nature Reviews Neuroscience* 5: 385–399.

Presynaptic Endosomes

O Shupliakov, Karolinska Institutet, Stockholm, Sweden
R Fernández-Chacón, Universidad de Sevilla, Sevilla, Spain

© 2009 Elsevier Ltd. All rights reserved.

Endocytic Intermediates in Presynaptic Terminals

Endosomes are distinct membrane-bound endocytic organelles inside cells. Classically, receptors and their ligands are taken up into clathrin-coated vesicles from the plasma membrane surface during endocytosis. They are sorted to the early endosome and then targeted to various intracellular destinations. They may be either transferred to the recycling endosome and then exposed back to the surface, or directed to degradation by transport to multivesicular bodies and then to late endosomal compartments and lysosomes. Each of these compartments is characterized by a distinct protein–lipid composition, morphology, and intraluminal pH, which become more acidic as the degradation pathway progresses.

Like other peripheral cellular compartments, nerve terminals contain structures involved in the renewal and recycling of plasma membrane components and the constitutive and/or regulated internalization of receptors and ion channels. Some pathogens, such as viruses and toxins, also accumulate in these membrane compartments. During synaptic activity the nerve terminal contains a variety of endocytic intermediates derived from fused synaptic vesicles (Figure 1).

Endocytic intermediates are represented by clathrin-coated pits of various sizes, vesicles and by larger irregular membrane compartments, such as vacuoles, cysternae, and tubular invaginations. Large endocytic compartments, which are temporarily observed in nerve terminals during stimulation, represent a functionally heterogeneous population. Most of the vacuoles are generated by bulk endocytosis from the plasma membrane. The membrane of these vacuoles is structurally and functionally equivalent to the plasma membrane. Accordingly, these structures are disassembled into vesicles in resting synapses via clathrin-mediated endocytosis. Some of the organelles, however, are indeed represented in classical sorting endosomes downstream of clathrin-coated vesicle formation. It remains a matter of debate whether early endosomes are composed of a single population of organelles, in which recycling of the membrane components may intersect with synaptic vesicle cycling, or if they represent a multiple population, specialized for a particular pathway. Recent studies of epidermal growth factor (EGF) receptor trafficking in nonneuronal cells have identified proteins which are confined only to a subpopulation of peripheral endosomes, thus indicating the possibility for multiple sorting pathways. Late endosomes (multivesicular bodies) may also be present at synapses. These organelles are eventually delivered to lysosomes in the soma-dendritic compartment of the neuron.

Sorting of Receptors

An important process that depends on endosomal trafficking is the uptake, processing, and recycling of presynaptic receptors. Little is known about the pathways directly linked to presynaptic terminals, mostly because of the limited size of this neuronal compartment, but it is evident that these endocytic events may play a crucial role in synaptic plasticity. Trafficking of several receptor subtypes, including alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) glutamate receptors, has been followed in the postsynaptic compartment. They showed that recycling endosomes, rather than *trans*-Golgi network (TGN)-derived vesicles, supply AMPA receptors during long-term potentiation (LTP) in hippocampal neurons. Blocking exit from recycling endosomes by expression of dominant-negative mutants of the GTPase Rab11a or the EH-domain containing accessory protein EHD1/Rme1 trapped internalization of the receptors in recycling endosomes and prevented expression of LTP in hippocampal neurons. It should be noted, however, that the postsynaptic endosomal compartments in neurons are significantly larger compared to the ones in presynaptic terminals, which suggests less significant physiological effects of receptor turnover in the presynapse.

Thus, the degree of similarity between these mechanisms in the presynaptic and postsynaptic compartments still needs to be clarified. It seems generally important that the presence of certain subunits determines the chosen sorting pathway. Homomeric GluR1 receptors, for example, are retained in recycling endosomes, whereas GluR3 homomers enter the late endosomal/lysosomal pathway regardless of stimulation. In the case of heteromers, endosomal sorting is apparently governed by GluR2, which exerts dominant effects, perhaps by recruiting adaptor proteins after undergoing posttranslational modifications including tyrosine phosphorylation and

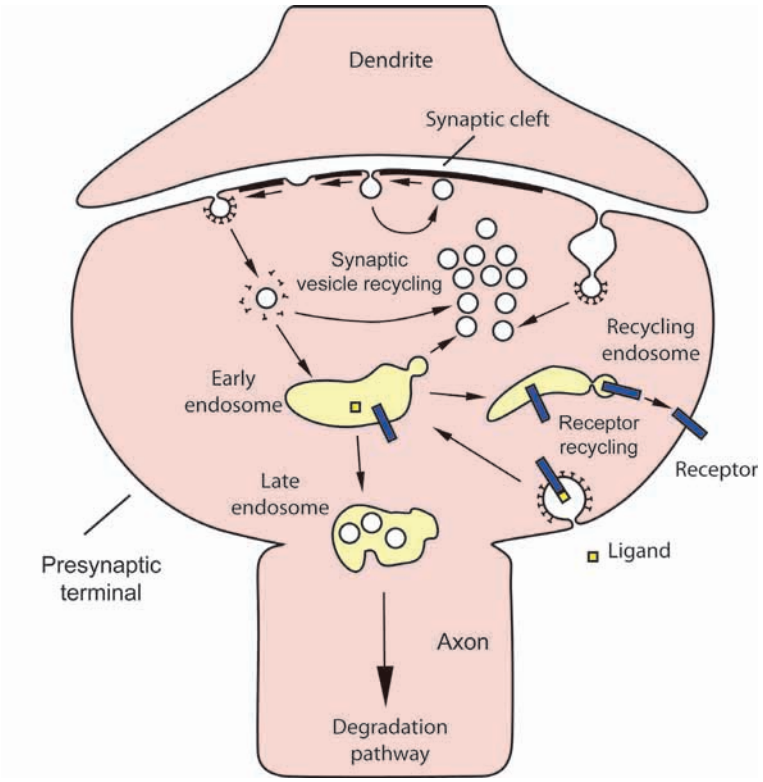

Figure 1 Membrane-trafficking pathways in presynaptic terminals. Synaptic vesicles releasing neurotransmitter into the synaptic cleft at the active zone (thick line) may be retrieved via the 'kiss and run' mechanism and via clathrin-mediated endocytosis, which occurs lateral to the active zone. During intense activity and massive fusion of synaptic vesicles deep plasma membrane folds are formed. During the synaptic vesicle cycle, the vesicles can be reformed directly from uncoated clathrin-coated vesicles or through an endosomal compartment. The 'classical' endosomal trafficking may regulate internalization and surface expression of receptors, transporters, and ion channels within the presynaptic compartment. The recycling pathway includes early and recycling endosomes. It may also involve late endosomal compartments (also referred to as multivesicular bodies) if the cargo is designated for degradation.

ubiquitination. In nonneuronal cells, sorting of ubiquitinated cargo is achieved by ubiquitin-interaction motif (UIM) and ubiquitin-associated domain (UBA) containing proteins, which include the phospholipid-binding protein epsin and the endocytic accessory protein Eps15. Both epsin and Eps15 are highly expressed in presynaptic terminals and thus may serve functions in presynaptic receptor sorting within the endosomal system in addition to their involvement in synaptic vesicle recycling.

Role of Endosomes in Synaptic Vesicle Recycling

The role of endosomes in synaptic membrane recycling also remains a matter of debate. A large proportion of vesicles in synapses are reformed via clathrin-mediated endocytosis, which occurs in the regions around active synapses referred to as periactive zones. Clathrin-coated intermediates formed in this region have the same diameter as functional synaptic vesicles. Thus, vesicles formed at the periactive zone may miss the endosomal compartment during the synaptic vesicle cycle, and thus originate directly from clathrin-coated vesicles pinched off from the plasma membrane or by the closing of a transient fusion pore ('kiss-and-run'). Studies in hippocampal synapses using the fluorescent membrane dye FM1–43 to trace the membrane of a recycled vesicle have shown that most of the vesicles recycle through the pathways independent of endosomal compartments. They have demonstrated that the amount of dye per vesicle taken up by endocytosis equals the amount of dye a vesicle releases upon exocytosis. It was thus concluded that the internalized vesicles participating in endo–exo recycling do not communicate with intermediate endosomal compartments during the recycling process. These experiments, however, did not exclude the possibility that a remaining population of vesicles not participating in exo–endocytic recycling could exchange membrane with an endosome or that this organelle could be recruited upon a certain activity demand. The sorting endosomal compartment could also be critical for long-term

maintenance of synaptic vesicle trafficking, since vesicles may become fusion incompetent following depletion of their protein content during repeated rounds of release. Evidence for an involvement of endosomes in synaptic vesicle cycling has recently been provided by studies using specific molecular endosomal markers.

The classical case of functionally distinct cellular microdomains is defined by the presence of specific Rab GTPases and their effector molecules. Three GTPases, Rab5, Rab4, and Rab11, are present on early endosomes. Cargo destined for degradation or recycling enters the early endosome via the Rab5 domain. Subsequently, cargo is sorted either for recycling via the Rab4 and Rab11 domains or for degradation via the Rab7 domain on late endosomes.

Rab5, initially delivered to the endosomal membrane in an inactive, GTP-bound form, undergoes activation catalyzed by specific effectors (GEFs), such as the Rabex5–Rabaptin5 complex, Ras interference 1 (RIN1), RIN2, or RIN3. Active Rab5 recruits two phosphatidylinositol-3-kinases, PI(3)-kinases p110/p85 and VPS34/p150, which trigger a local enrichment of phosphatidylinositol-3-phosphate, PI(3)P, in the endosomal membrane. PI(3)P specifically binds to the FYVE zinc finger domain (named after Fab1p, YOTB, Vac1p, and EEA1) of endosomal factors such as EEA1 and Rabenosyn-5 or Rabankyrin 5, which ultimately mediate endocytic vesicle tethering and fusion with early endosomes. Both Rab5 and the FYVE domain, which binds to PI(3)P within an intact lipid bilayer, can thus be considered markers for PI(3)P-containing endosomes.

The presence of Rab5-positive, PI(3)P-containing endosomes at the presynaptic terminal of *Drosophila* neuromuscular junctions has been determined using expression of GFP-tagged FYVE domains and antibodies against Rab5. Under conditions in which the synaptic vesicle pool was depleted, endosomes were drastically reduced in size and recovered by dynamin-mediated endocytosis. Interfering with Rab5 function using a dominant-negative version of Rab5 caused a reduction in the number of released quanta during synaptic transmission, whereas elevated levels of Rab5 increased the quantal content. These data strongly indicated that Rab5-dependent trafficking pathway plays a role in presynaptic vesicle cycling.

Live imaging of EGFP-Rab5a was also achieved in cultured neurons from mammalian hippocampus. For a number of years it has been known that the small GTPase Rab5 is present in vesicles isolated from synaptosomes. Moreover, overexpression of Rab5a has been shown to target the protein to synaptic terminals and to vesicular and endosome-like structures. The imaging data confirmed that EGFP-Rab5a is accumulated in clusters along the axon, forming green puncta, which colocalized with synaptic markers. Upon stimulation, Rab5a showed a minimum dispersion, far less than that seen with EGFP-Rab3a, which undergoes a cycle of association and dissociation with synaptic vesicles. These data suggested that like in *Drosophila,* mammalian Rab5 is confined to a specific population of vesicular structures most probably representing a recycling compartment.

Fusion of vesicles with the endosomal compartment would require a soluble N-ethylmaleimide-sensitive factor attachment protein (SNAP) receptor (soluble N-ethylmaleimide-sensitive factor attachment protein receptors (SNARE)) complex. The proteins Vti1b and Vti1a-β have been found on small vesicles isolated from nerve terminals. Vti1a-β was co-purified with synaptobrevin and was able to interact with N-ethylmaleimide-sensitive factor (NSF) and aSNAP, suggesting that it might form a SNARE complex. The protein complex, containing Vti1a-β, did not contain synaxin 1 and SNAP-25, and thus was different from the SNARE complex involved in fusion of synaptic vesicles with the plasma membrane. This Vti1a-β-containing protein complex is currently suggested as a potential SNARE for fusion with the endosomal compartment.

The formation of vesicles from the endosomal compartment also assumes a budding step. This budding step is thought to be dependent on the adaptor complex AP-3. The neuron-specific isoform of AP-3 (AP-3B) as well as ADP ribosylation factor 1 (ARF1) are required for the biogenesis of synaptic-like microvesicles budding from PC12 cell endosomes, but a similar function in nerve terminals has not yet been confirmed. It is known, however, that mutations in the AP-3 gene are linked to a variety of neurological defects. The *mocha* mouse, a null mutation of the δ-subunit of AP-3, exhibits balance and hearing problems, is hyperactive, and is prone to seizures. Mice in which the neuron-specific AP-3B subunit μ3B has been genetically deleted show specific defects related to the biogenesis of GABA-containing SSVs, suggesting a particularly important function for AP-3B at inhibitory synapses.

Another potential function for endosome-derived synaptic vesicles is in the recovery of membrane components of dense-core vesicles (DCVs) that have just undergone exocytosis. Membrane retrieval of this type has been detected in PC12 cells transfected with a chimeric P-selectin. Thus, neuronal AP-3B may recapture protein components of DCV proteins. It is believed that a recapture step could sequester selected DCV proteins from degradation and allow them to be incorporated into the synaptic vesicle cycle.

Signaling Endosomes

A population of 'signaling endosomes' has been recently described. These organelles avoid the classical lysosomal targeting system but deliver compounds to the cell body in biologically active form. Interestingly, some of these membrane structures may orient the intracellular domains of receptors to face the cytoplasm, making the receptors available to interact with cytoplasmic proteins and thus increasing the signaling possibilities (Figure 2).

Nerve growth factor (NGF) binds to and activates presynaptic neurotrophin receptors TrkA and p75 and triggers their endocytosis and transport to the endosomal compartments. In sciatic nerve preparations, the stimulation with NGF induces early endosomes, positive for Rab5B and EEA1. These endosomes become signaling endosomes that are retrogradely trafficked to the cell soma to initiate local signal transduction cascades and transcription of specific genes. On the other hand, NGF induces the formation of signaling endosomes. The EH domain protein Pincher/EHD4 regulates the formation and cellular fate of these endosomes. Pincher/EHD4 overexpression enhances NGF-induced internalization, sorting and trafficking of TrkA-containing vesicles. In contrast, a dominant negative version of Pincher/EHD4 blocks TrkA internalization. Thus, Pincher/EHD4-containing Trk endosomes represent a pathway opposite to lysosomal degradation and ensure persistent signaling through extracellular signal-regulated kinase 5 (Erk5), an important effector of retrograde survival signaling in neurons.

Transport of the clostridial tetanus toxin (TnT) to the somatodendritic compartment of motoneurons displays many similarities to the transport of NGF receptors. The TnT is composed of two chains: a heavy chain (TnT-HC) and a light (TnT-LC) chain. The TnT-LC is a zinc-endopetidase that specifically cleaves synaptobrevin/VAMP, and the TnT-HC is responsible for the presynaptic uptake and retrograde transport of the toxin. The toxin is then taken up into the neuromuscular junction by endocytosis and is sequestered into signaling endosomes. The endosomes undergo fast axonal retrograde transport and carry the toxin to the soma and dendrites of motor neurons located in the spinal cord. Once there, the endosomes fuse with the plasma membrane releasing the toxin. It reaches the presynaptic terminals of inhibitory interneurons and at this level incorporates into the synapses through recycled synaptic vesicles.

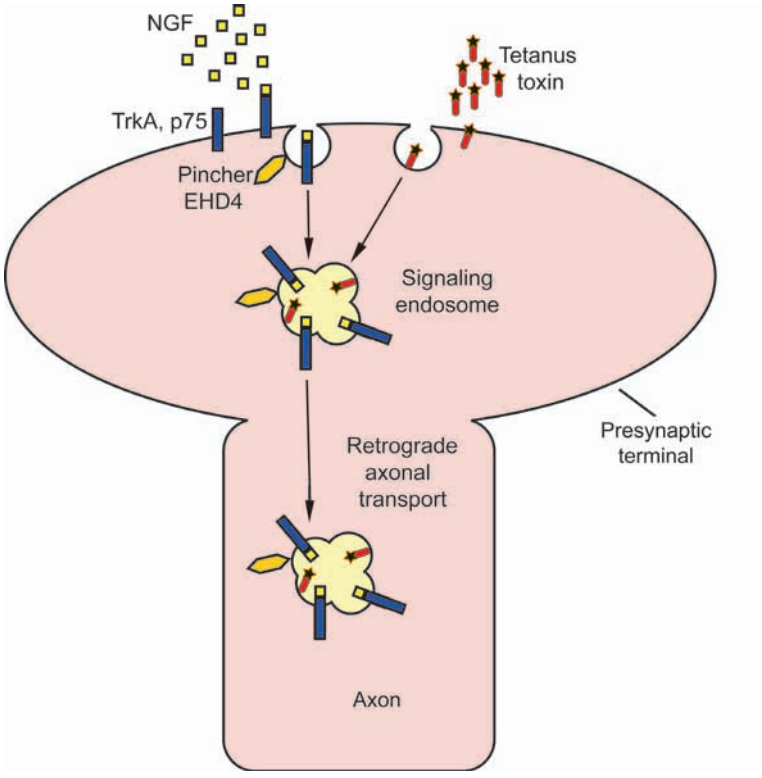

Figure 2 'Signaling' endosomal compartment involved in the retrograde transport of neurotrophin receptors and tetanus toxin. Target released nerve growth factor (NFG) binds and activates TrkA or p75 neurotrophin receptor. Pincher/EHD4 activates internalization of the complexes to signaling endosomes. The tetanus toxin binds to the presynaptic membrane and is endocytosed to the presynaptic endosome to be delivered to the somatodendritic neuronal compartment.

This process is referred to as transcytosis. The tetanus toxin blocks neurotransmission in synapses through the cleavage of synaptic vesicle protein synaptobrevin/VAMP, which is essential for vesicle fusion. Importantly, the TnT-HC and NGF appear to share the same retrograde transport organelle, which is characterized by the presence of the neurotrophin receptor p75. This coincidence implies that during evolution, bacteria have adapted their toxins to benefit from an essential neuronal membrane trafficking pathway in order to invade the nervous system.

See also: Clathrin and Clathrin-Adaptors; Endocytosis and Presynaptic Scaffolds; Large Dense Core Vesicles (LDCVs); Lysosome and Endosome Organization and Transport in Neurons; Presynaptic Receptor Signaling; Synapsins and Regulation of the Reserve Pool; Synaptic Vesicles; Vesicle Pools.

Further Reading

Antonin W, Riedel D, and von Mollard GF (2000) The SNARE Vti1a-beta is localized to small synaptic vesicles and participates in a novel SNARE complex. *Journal of Neuroscience* 20: 5724–5732.

Ashby MC, De La Rue SA, Ralph GS, Uney J, Collingridge GL, and Henley JM (2004) Removal of AMPA receptors (AMPARs) from synapses is preceded by transient endocytosis of extrasynaptic AMPARs. *Journal of Neuroscience* 24: 5172–5176.

Howe CL and Mobley WC (2005) Long distance retrograde neurotrophic signaling. *Current Opinion in Neurobiology* 15: 40–48.

Kavalali ET (2006) Synaptic vesicle reuse and its implications. *Neuroscientist* 12: 57–66.

Lalli G, Bohnert S, Deinhardt K, Verastegui C, and Schiavo G (2003) The journey of tetanus and botulinum neurotoxins in neurons. *Trends in Microbiology* 11: 431–437.

Montesinos ML, Castellano-Munoz M, Garcia-Junco-Clemente P, and Fernandez-Chacon R (2005) Recycling and EH domain proteins at the synapse. *Brain Research and Review* 49: 416–428.

Murthy VN and Stevens CF (1998) Synaptic vesicles retain their identity through the endocytic cycle. *Nature* 392: 497–501.

Shupliakov O and Hauke V (2006) Synaptic endosomes. In: Dikic I (ed.) *Endosomes*, p. 36. Georgetown, TX: Landes Bioscience.

Wenk MR and De Camilli P (2004) Protein–lipid interactions and phosphoinositide metabolism in membrane traffic: Insights from vesicle recycling in nerve terminals. *Proceedings of the National Academy of Sciences of the United States of America* 101: 8262–8269.

Wucherpfennig T, Wilsch-Brauninger M, and Gonzalez-Gaitan M (2003) Role of *Drosophila* Rab5 during endosomal trafficking at the synapse and evoked neurotransmitter release. *Journal of Cell Biology* 161: 609–624.

Presynaptic Events in Neuromuscular Transmission

H Teng and R S Wilkinson, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Because of its accessibility, the neuromuscular junction (NMJ) is the most studied vertebrate synapse. The NMJ is also the simplest synapse – a relay, so called because its function is to relay information, traveling on motor axons originating from within the central nervous system, to the membrane of muscle fibers. The information is digital, in the form of all-or-nothing depolarizing pulses, or action potentials (APs). Each pulse instructs the muscle fibers to contract or (for most fibers) twitch. Thus the NMJ amplifies a weak control signal (current supplied by the axonal AP) into a stronger one capable of depolarizing the muscle fiber, which is much larger than the axon. In contrast, and as described elsewhere in this encyclopedia, synapses within the brain and spinal cord are thought to serve far more complex functions – those that underlie behavior, such as learning, memory, emotion, and addiction. To do so, brain synapses utilize various kinds of use-dependent behavior, such as depression, potentiation, and facilitation. They are also able to integrate, or combine, multiple inputs (APs originating from many neurons) and arrive at a single decision – namely, whether or not to generate an output AP.

With such disparate roles, it seems remarkable that the simple NMJ can serve as the model synapse, as it is called, for understanding synaptic function throughout the vertebrate nervous system. There are two reasons. First, many of the characteristics of brain synapses are found at the NMJ, albeit in more rudimentary form, and indeed were discovered there. Second, and most relevant for what follows herein, is the fact that much still remains to be learned. One key aspect of all synapses – how the presynaptic nerve terminal, or presynapse, releases its chemical signal – remains poorly understood. The importance of gaining this knowledge becomes clear if one considers the output side of the synapse, where the postsynaptic receptors convert an increased concentration of the chemical neurotransmitter, in the synaptic cleft, into the depolarization that triggers muscle excitation. Here, a great deal is known. Early information came exclusively from the NMJ, but many classes of receptors at different synapses are now studied in detail. Knowledge of precisely how these receptors work has advanced substantially in recent years, and continues to advance. As a result, pharmacological intervention is possible, and many successful drugs have postsynaptic receptors as their target. A similar arsenal of presynaptic drugs, if available, could alter neurotransmitter release directly, providing additional avenues for intervention. Such drugs would have a profound impact on the treatment of disease, as have those that modify synaptic function by acting on receptors. But, unlike receptors, the basic physiology of nerve terminals remains to be elucidated. If history serves, the synapse most likely to reveal these secrets is the simplest and most accessible one, the NMJ.

Anatomy of Motor Nerve Terminals

Presynaptic function derives from extreme evolutionary specialization. Cells in early animals communicated with each other via hormones, which are chemicals released by certain cells into the extracellular milieu; these chemical signals are detected by other cells, often some distance away, that express the appropriate receptors. One problem with this scheme is that each signal to be communicated requires a unique hormone. Another is that chemical communication over long distances is slow. Though still extant today (the endocrine system), some hormonal communication evolved into neural communication precisely to avoid these problems. Neural communication occurs only between adjacent cells. The cells are so close (~50 nm for NMJs) that only a tiny amount of chemical released by one is sufficient to rapidly influence receptors of the other. After the signal is delivered, the chemical diffuses away and is harmlessly diluted, influencing no other cells. Alternatively, it is degraded (as at the NMJ) or taken up for reuse. The most remarkable aspect of this specialization is that, during development, one cell, the presynaptic neuron, extends a thin process, its axon, over a distance of up to 1 m or so (in humans) and finds its target muscle fibers with precision. Thus a motor neuron in the spinal cord becomes, in effect, adjacent to a muscle cell in the foot (or more than one cell – the axon can branch). The specialization at which the terminal of the axon comes in near-contact with the receptors of the muscle cell is the NMJ. As a result, many muscle cells can be independently and rapidly controlled by their innervating motor neurons, using only one chemical neurotransmitter. In vertebrates, this transmitter is acetylcholine (ACh).

A consequence of this system is that each motor nerve terminal undertakes its function without much

help from the motor neuron's distant cell body. Although the latter contains the nucleus and protein-making machinery, axonal transport of molecules from cell body to the motor terminal of the NMJ, or vice versa, can require up to several months. In particular, the need for independent function requires that motor terminals supply their own energy and continuously resynthesize and package ACh for reuse (see later, vesicle recycling).

Nerve Terminals at NMJs of Different Vertebrate Species

The most fundamental, unitary synapse comprises a presynaptic bouton (button), a narrow cleft, and a cluster of membrane-bound postsynaptic receptors (Figure 1). The bouton contains, principally, mitochondria (M), synaptic vesicles (SVs) that contain neurotransmitter, and additional transmitter within the cytoplasm. Active zones (AZs), sites of neurotransmitter release, are located within the presynaptic membrane, the portion of the plasma membrane that apposes postsynaptic receptors. Interestingly, solitary boutons are common within the central and autonomic nervous system, but rarely serve as NMJs. However, the structure of NMJs of various species is essentially a cluster of boutons, each with its own postsynaptic receptor specialization. In amphibians (frog), thousands of boutons are linearly arranged, each containing a single AZ. In reptiles, ~60 boutons are clustered, much like a bunch of grapes, and each contains ~60 AZs (Figure 2). In mammals, such as mice, the NMJ resembles a tangle of worms. Each worm appears to be formed by the fusion of several reptilian boutons, and indeed, during development, the not-yet-fused individual boutons are evident. Interestingly, human NMJs retain separate boutons to some extent and thereby resemble those of reptiles more than rodents! Thus the NMJs of different species are homologous, with reptilian (and human) NMJs most resembling the discrete bouton synapses found in the mammalian central nervous system.

Anatomical Correlates of Synaptic Strength

The strength, or efficacy of a motor terminal (amount of transmitter released) is reflected in its size, which in turn is matched to the requirements of its target muscle fiber. Thus larger fibers receive larger terminals. So do faster ones, such as those involved in fright-or-flight response. These are important for survival and therefore receive more transmitter than is minimally required for initiating contraction (the safety factor). Conversely, smaller fibers, as well as slow, chronically used muscle fibers that are relatively unimportant to acute survival, receive smaller terminals (e.g., eye muscles, anal sphincter). In the snake and probably other species as well, postsynaptic receptors are more efficient in fibers that are larger, faster, or more important for survival. Thus nerve terminal size is not the only determinant of synaptic strength. Boutons in small and large terminals are roughly the same size – it is the number of boutons that changes.

The Active Zone

In the 1950s, when Sir Bernard Katz and his colleagues elucidated the basis of synaptic function at the frog NMJ, transmission electron microscopy had recently become available to provide images of

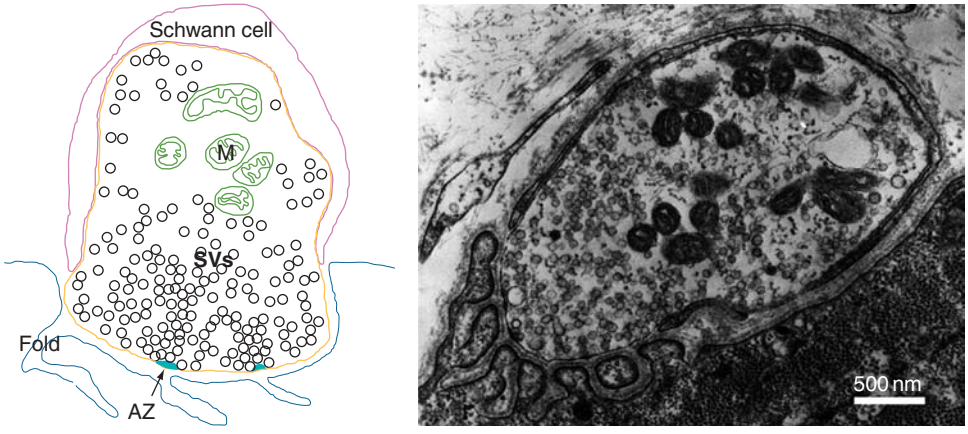

Figure 1 Cross section of typical presynaptic bouton. (Left) Diagrammatic representation showing bouton (yellow) capped by Schwann cell (purple) and innervating the postsynaptic membrane (endplate) of muscle fiber (blue). Mitochondria (M; green) reside toward the back of the bouton, away from the presynaptic membrane. Synaptic vesicles (SVs) are clustered throughout the bouton, but particularly near active zones (AZ; blue). Deep invaginations, or folds, within the postsynaptic muscle membrane precisely appose AZs. (Right) Electron micrograph thin section of a bouton at the snake neuromuscular junction. Contractile proteins in the muscle fiber (seen along the bottom of the micrograph) appear in cross section. Folds are continuous along the postsynaptic surface with active zones spaced along them (see Figure 2). Thus some (or all) folds may lack active zones in a particular electron micrograph thin section.

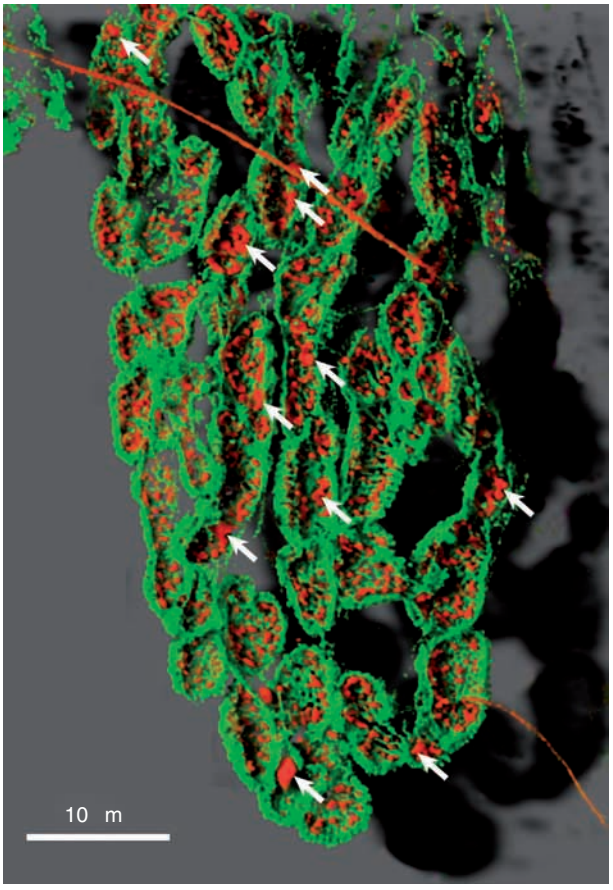

Figure 2 Light-level image of snake neuromuscular junction from the single-fiber-thick transversus abdominis muscle. The postsynaptic membrane of the muscle fiber has been stained with a lectin (fluorescein–*Vicia villosa* agglutinin; green) to reveal the location of acetylcholine receptors (AChRs). Each concave, bowl-like cluster of AChRs corresponds to the dimpling of the endplate membrane where it contacts boutons (blue in diagram of **Figure 1**). The repeated stripes within each bowl, resembling fingerprints, are postsynaptic folds where AChR density is highest. Boutons are labeled with sulforhodamine 101 (red), a fluorescent marker of endocytosed structures (see **Figure 6**). Larger structures are endosomes (arrows); small dots are vesicle clusters near active zones. Particularly near the edges of boutons, it can be appreciated that the clusters appose folds. Confocal 3-D reconstruction from 115 image planes.

synapses at high magnification. The numerous small (~50 nm) vesicles that filled boutons were thought to be the anatomical correlate of the singular electrical events, or quanta, observed with intracellular electrophysiological recording. Similarly, electron-dense structures located in the presynaptic membrane, precisely opposite infoldings of the postsynaptic membrane and often surrounded by vesicles, were thought to be likely sites of transmitter release. Two more recent advances in electron microscopy (EM) – the freeze-fracture technique used by J Heuser and colleagues and electron tomography used by UJ McMahan and colleagues – have provided additional detail about the structure of AZs (**Figure 3**). Freeze-fracture EM replicas reveal that AZs contain arrays of large intramembraneous particles that are thought to be calcium channels, or calcium-dependent potassium channels, or both. The arrangement of particles is in striking quadruple or double rows of various lengths. Associated with the particles are rows of vesicles are evidently captured or docked, awaiting release. The specific arrangement and number of vesicles docked at the AZ (~4–40, the latter in frog) depend not only on species (frog, lizard, rat, etc.) but also on fiber type (twitch, tonic) within a particular species. Tomography has revealed further structural components, named pegs, ribs, and beams, whose functions are still unknown. This technique promises to relate high-resolution anatomical images of AZs with structures (e.g., the SNARE complex, so named for the soluble NSF (*N*-ethylmaleimide-sensitive factor) attachment receptor) conceived from the molecular biological study of transmitter release. Eventually, molecules known to participate in docking vesicles for release at AZs will be visible, along with a detailed view of the vesicles.

EM provides only a static picture of the AZ and associated vesicles, but (as discussed later), vesicle release is a transient phenomenon. An ingenious technique – rapid, or slam, freezing – was developed by Heuser in an attempt to overcome the static nature of EM. In lieu of chemical fixation, living tissue is instantaneously frozen by slamming it against a cold metal block. By stimulating a frog NMJ at the time of freezing, Heuser hoped to capture a dynamic snapshot of vesicles just as they were released from AZs. This proved unsuccessful, presumably because two very brief yet independent transient events, vesicle release and the freezing of the NMJ, were unlikely to coincide. Only by utilizing a chemical agent to prolong and enhance transmitter release was it possible to capture such a snapshot, albeit under nonphysiological conditions.

Although long suspected, actual proof that AZs release transmitter under normal conditions came only recently. Zefirov and colleagues, using three recording electrodes to triangulate the source of synaptic currents at an active frog NMJ, found that the loci of all such currents (spontaneous and evoked) correspond precisely to the anatomical positions of AZs.

Classical (Katz) Theory of Transmitter Release

The standard model of transmitter release is based on the work of Katz and his colleagues, mainly over the period 1950–80 that followed the introduction of intracellular recording in 1949. In a series of experiments, they established the following events:

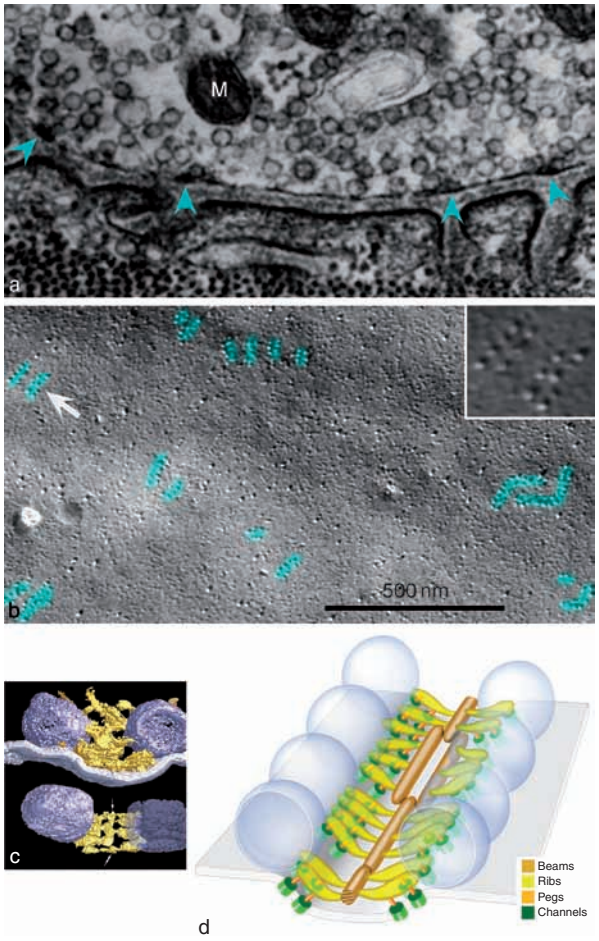

Figure 3 Electron microscopic visualization of active zones at the vertebrate neuromuscular junction. (a) Transmission electron micrograph of thin (~70 nm) sections of snake neuromuscular junction reveals electron-dense structures (active zones; arrows) surrounded by vesicles. The active zones are within the bouton (note vesicles and mitochondria (M)) and are associated with its presynaptic membrane. Each active zone precisely apposes a fold in the postsynaptic membrane of the muscle fiber. (b) Freeze-fracture electron microscopy separates regions of the presynaptic lipid bilayer membrane into outer and inner leaflets, thereby allowing imaging of intramembraneous proteins. Shown are several active zones (blue) within the presynaptic membrane of a bouton at the snake neuromuscular junction. The orientation is as if the bouton in micrograph a were viewed looking in the direction of the arrows. Each active zone appears as two double rows of particles (Ca^{2+} channels; rows are colored blue). One active zone (arrow) is shown at higher magnification in the insert (upper right; scale bar applies to both micrographs (a) and (b). (c) Electron tomography produces high-resolution surface models; shown is a portion of one active zone at the frog neuromuscular junction. The upper image shows docked synaptic vesicles (dark blue), presynaptic membrane (pale blue), and active zone material (gold). The active zone material is composed of a network of filamentous macromolecules, many of which contact the docked vesicles and presynaptic membrane. The lower image is of the superficial 15 nm of active zone material, showing beams (arrows), which lie in the midline of the active zone material, and ribs, which extend from the beams to the docked vesicles. (d) Schematic of the frog active zone derived from electron tomography and freeze-fracture electron microscopy. Shown is the presynaptic membrane (pale blue-gray), docked synaptic vesicles (darker blue), and the superficial 15 nm of the active zone material, consisting of beams (brown gold), ribs (yellow-gold), and pegs (orange-gold). Positions of macromolecules, which include calcium channels, in the presynaptic membrane are shown in green (compare to (b)). (a, b) Image provided by J Heuser. (c) Reprinted by permission from MacMillan Publishers Ltd: *Nature*, Vol. 409, Harlow ML, Ress D, Stoschek A, et al. The architecture of active zone material at the frogs neuromuscular junction, 479–484, Copyright 2007. (d) Reproduced from Ress DB, Harlow ML, Marshall RM, et al., (2004) Methods for generating high resolution structural models from electron microscope tomography data. *Structure* 12: 1763–1774, with permission from Elsevier.

1. Activation of frog motor terminals by nerve stimulation produced potentials that could be recorded from muscle fibers near the endplate (i.e., endplate potentials, or EPPs).
2. Depolarizing electrical events (miniature endplate potentials, or mEPPs) occurred spontaneously at endplates; mEPPs (~0.5 mV) resembled EPPs but were smaller.
3. The amplitude of EPPs evoked by nerve stimulation varied randomly, and the occurrence of spontaneous mEPPs was random in time.
4. The concentration of Ca^{2+} in the bath determined the range of EPP amplitudes: EPPs were absent when Ca^{2+} was zero, fluctuated among amplitudes corresponding to 0, 1, 2, 3, 4, etc. mEPPs when Ca^{2+} was low (~0.2 mM), and rose to ~20–40 mV at physiological bath Ca^{2+} levels (2–3 mM).

These experiments, together with subsequent ones by AR Martin (who had been Katz's student), resulted in a series of hypotheses that comprise the classical theory.

The Quantal Hypothesis

Katz's observations indicated that motor terminals secrete relatively uniform packages of ACh called quanta. Individual quanta are released randomly, and create mEPPs. Classical experiments conducted by Kuffler and Yoshikami, using the snake NMJ, demonstrated that one quantum does not correspond to a single ACh molecule but rather to 5–10 thousand molecules, synchronously released. This number is consistent with what might be packed into a single vesicle, furthering evidence that the quantum and the synaptic vesicle are one and the same. With nerve stimulation, several quanta are released nearly simultaneously, so that the postsynaptic currents they cause add together into a summed endplate current (the EPC). The EPC, in turn, develops a potential, the EPP, across the muscle fiber's membrane resistance (Ohm's law). To activate the muscle fiber, some

minimum number of quanta is required, either to directly activate contraction in tonic fibers (which we do not consider further) or to drive the fiber's membrane well above the threshold for action potential initiation (twitch fibers). The actual number of quanta released with each stimulus is in principle random, but it varies little and is statistically quite unlikely to be below threshold (see the section titled 'Binomial statistics with normal extracellular calcium').

The Calcium Hypothesis

Rise in intracellular Ca^{2+} is now well understood as a ubiquitous signaling mechanism in animals. Calcium entry (or its release from intracellular stores) triggers hormone secretion, muscle contraction, and fertilization of eggs, as well as transmitter release. Thus Katz's fundamental observation regarding the dependence of transmitter release on calcium entry into the nerve terminal was seminal far beyond synaptic physiology. At the NMJ, entry of Ca^{2+} via voltage-dependent Ca^{2+} channels causes a transient increase in the intracellular level of Ca^{2+} in the bouton, particularly near AZs where the channels are heavily concentrated. The increase triggers the fusion with the plasma membrane of some of the vesicles already poised for release at the AZ. With the vesicle lumens now continuous with the extracellular face of the plasma membrane, their vesicular contents, including ACh, diffuse into the synaptic cleft.

In motor terminals, probability of release is a steep (~fourth power) function of intracellular Ca^{2+}. At rest, low levels of Ca^{2+} (~40 nM) are associated with the release of an occasional vesicle every second or so (mEPPs) from an entire terminal. There is some evidence that mEPPs may serve a trophic or tropic function, or they may simply represent the inability of the system to turn off release completely. With the arrival of an AP from the motor neuron, the voltage-dependent channels open, and extracellular Ca^{2+} flows passively into the terminal (down both concentration and electrical gradients). In the snake, intracellular levels rise to ~300 nM (Figure 4), and hot spots near AZs are probably at micromolar concentrations. Just as rapidly, buffers within the terminal (including mitochondria) return Ca^{2+} levels to near the resting level. The result is a rapid, transient increase in release probability. A terminal (containing, say, 3000 AZs) releases ~50–100 vesicles in the ~1 ms during which Ca^{2+} was elevated, rather than a single vesicle per second – a rise in release probability by a factor of ~10^5!

Poisson Statistics with Low Extracellular Calcium

Lowering bath Ca^{2+} by a factor of ~5–10, and/or increasing the concentration of magnesium or other divalent cations that compete with it, will decrease the activity of extracellular Ca^{2+} and decrease the gradients that support its entry. Instead of releasing 50–100 vesicles per AP, a terminal releases a much smaller number. Katz estimated this number by dividing the EPP amplitudes he observed under these conditions by the mean amplitude of spontaneous mEPPs. The number fluctuated dramatically! For

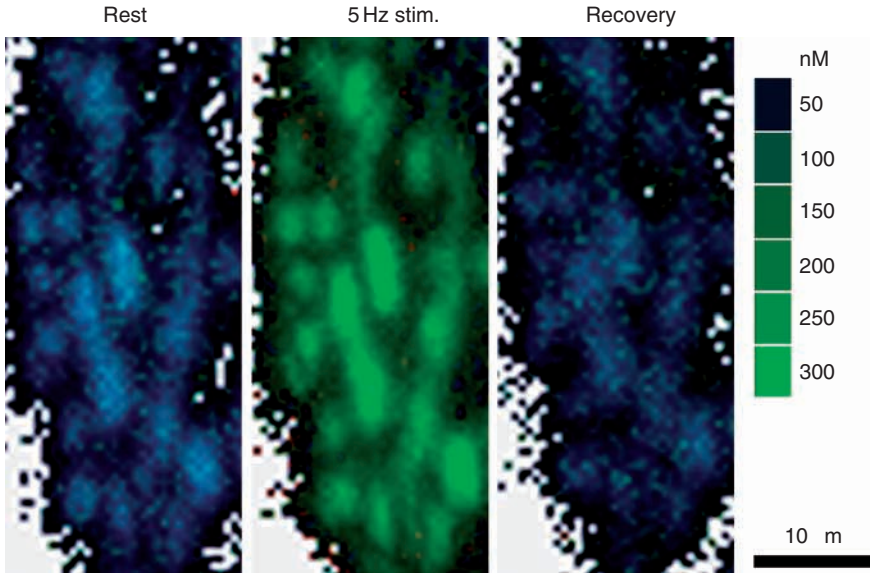

Figure 4 Intracellular Ca^{2+} transient in a snake motor terminal during a 5 Hz, 30 s stimulation. The Ca^{2+}-sensitive dye fura-2 was loaded into the muscle nerve to permit measurement of bulk intrabouton Ca^{2+} concentration. Stimulation increased Ca^{2+} level from 50 nM at rest (left) to 300 nM during activity (center). Levels remained elevated for ~30 s after the end of stimulation, after which they rapidly returned to baseline (right).

example, a terminal might release on average 5 quanta per stimulus (called the mean quantal content, m). But the actual number varies about that mean with each AP delivered – 8, 4, 1, 7, 0, 10, etc. Katz, and later Martin, showed that the EPPs generated by frog motor terminals with low bath calcium obeyed Poisson statistics. Poisson's formulation allows one to assess the probability of transmitter release, p, by recording the distribution of EPP amplitudes after many trials. It was found that the distribution fit Poisson's theory (meaning that quantal events appeared to be independent and therefore randomly distributed about their mean), p was low ($\sim$0.01), and p increased with increasing bath Ca^{2+}.

Binomial Statistics with Normal Extracellular Calcium

Poisson's statistical treatment is applicable only for probabilities below $\sim$0.02, which was indeed the case in early experiments where bath Ca^{2+} was deliberately lowered. The more general statistical model of random two-state phenomena (i.e., a vesicle is either released with the arrival of an AP or it is not) is binomial statistics, which must be used under physiological conditions where p is known to be higher (normal bath Ca^{2+}). Unfortunately, p is not as easily measurable as it is with Poisson statistics. The binomial distribution depends on p plus a second parameter, n, the binomial number of release sites. For binomial statistics to apply, the release sites must be presumed independent; that is, release from one must not influence release from another. The mean quantal content is $m = np$. Events are said to be simply binomially distributed (and therefore truly random, or independent) if and only if their fluctuations about the mean, m, correspond to those predicted by a particular pair of n and p. The expected fluctuations are given by

$$N_x = N_{\text{tot}} n! p^x (1-p)^{n-x} / (n-x)! x!$$

where N_{tot} is the number of stimuli delivered and N_x is the number of responses having the amplitude $x = 0, 1, 2, \ldots, n$ quanta (the symbol ! means factorial). For example, if p is 1 (certainty), then each of n sites releases one vesicle and $m = n$; there are no fluctuations. If p is 1/2 (the same probability as that of obtaining heads in a coin flip), then $m = n/2$. If a bouton contained, say, four binomial release sites and p were 1/2, on average $m = 2$ vesicles would be released per stimulus, just as flipping four coins at once would on average produce two heads per flip. But how often would one expect to see exactly two heads? Our equation answers that question, giving the number of trials for which one would expect 0 (all tails), 1, 2, 3, or 4 (all) heads (1, 4, 6, 4, and 1 out of 16 trials, respectively). Note that release of the mean number of quanta (2) is the most probable, but this particular outcome is expected in only 6 of 16, or $\sim$38% of the trials.

There are several other general features of binomial statistics. Fluctuations about the mean increase with decreasing p, as can be inferred from the preceding comments. The variance (standard deviation squared) is given by $np(1-p)$. Variance is zero if p is 1 and increases until, with $p < 0.02$, it approaches equaling the mean, m (as it does with Poisson statistics). Fluctuations also depend upon n, but in a different way. If 1000 coins are flipped simultaneously, it is unlikely that exactly 500 will be heads, just as in the preceding example (500 is the most likely single outcome, but there are very many other possibilities). However, it is also unlikely that the distribution will deviate far from the mean. Fluctuations over the range from 480 to 520 heads are feasible, but the probability of 499 heads and only one tail would be akin to winning the lottery.

The properties of the parameters p and n as discussed here make difficult the statistical analysis of quantal release at NMJs under physiological conditions. Quantal fluctuations are quite small in normal bath calcium, presumably because n is so large. EPP fluctuations due to thermal 'noise,' change in membrane properties, etc. mask EPP fluctuations due to varying quantal release. One way to overcome this difficulty, although it is not completely satisfactory, is to attempt to isolate small portions of the NMJ in order to lower n. Because snake NMJs comprise $\sim$60 discrete boutons, about 2% of the nerve terminal is studied by activating a single isolated bouton. At this small synapse, quantal fluctuations are indeed evident in normal Ca^{2+} bath (Figure 5), and binomial analysis suggests that p is $\sim$0.4 and n is $\sim$4. Mean quantal content, m, is $\sim$1.6 vesicles per bouton. These data are for low-frequency (1–5 Hz) continuous stimulation; m is often higher for the first stimulus or two after rest, and lower after sustained stimulation at high frequency. In other NMJ preparations (human, frog), p has usually been reported in the range from 0.1 to 0.3, measured or estimated using a variety of techniques.

The Binomial *n* and the Countable Number of AZs

If the Katz theory is correct, the binomial n should have an anatomical correlate corresponding to the number of vesicles that are independently available for release upon arrival of an AP. EPPs are generally sufficient to depolarize the muscle fiber above threshold with a safety factor of $\sim$2–4, meaning that m is $\sim$100. Thus, with p in the range from 0.1 to 0.4,

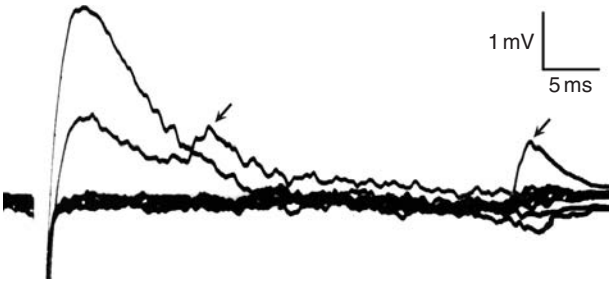

Figure 5 Quantal fluctuations recorded from a snake motor bouton in normal Ca^{2+} concentration bath. Endplate potentials were recorded intracellularly from a muscle fiber endplate, the entire nerve terminal of which, save a single bouton, was removed. Shown superimposed are the responses to five stimuli (negative-going artifact at left) delivered at 1 s intervals. Either zero, 1–2, or 2–4 quanta were released. Two spontaneous miniature endplate potentials (arrows) are contained in the records, showing approximate quantal size.

n is ~250–1000. In contrast, by direct count, an average snake terminal comprising ~60 boutons contains ~3600 AZs. A similar number of AZs is present at the frog NMJ, and probably at other NMJs as well. These rough estimates ignore important systematic variations among NMJs of various species, and among different types of muscle within one animal. Nevertheless, it seems likely that *N*, the anatomical number of AZs, exceeds *n*, the number of release sites calculated from binomial statistics, perhaps by as much as an order of magnitude. Moreover, as already discussed, each AZ contains two or more apparently releasable vesicles. If each such vesicle acts independently (which is unknown), *N* is still larger.

There is at present no conclusive explanation for the disparity between *n* and *N*, although several models have been proposed. One is that some AZs have higher *p* than others, and thus form a functional pool of release sites whose number approximates *n*, not *N*. Presumably, these AZs are employed under conditions of normal use, with the others awaiting recruitment when needed. A second model is that the mEPP, or quantum, actually comprises several vesicles that are synchronously released. Thus the anatomical correlate of *n* is not one AZ, but rather several AZs. A third explanation is that transmitter release viewed at the level of single boutons, not whole NMJs, is partly deterministic, not random. Put another way, the AZs within one bouton act cooperatively, not independently. For example, an AZ that has recently released a vesicle may wait its turn before it releases again. Any sort of systematic cooperation among AZs would reduce the observed variance, and, consequently, the calculated *n*. There is currently some support, but no conclusive evidence, for each of these models.

Factors Underlying the Binomial *p*

The dependence of *p* on Ca^{2+} levels indicates that one contributing factor is the opening probability of Ca^{2+} channels and K^+ channels associated with the AZ. This probability could change dynamically, influencing release from particular AZs. A second factor, operating on a slower time scale, might originate in the postsynapse. Retrograde feedback from the effect of activating receptors apposing an AZ could regulate that AZ's efficacy. Alternatively, vesicle membrane proteins might regulate an AZ once the proteins are inserted into the plasma membrane – signaling that transmitter release from that AZ has occurred. Two vesicle proteins, SV2 and synaptotagmin, have already been found to interact with other proteins within the synaptic cleft. All of the preceding mechanisms have the potential to add a deterministic component to transmitter release, thereby explaining the disparity between *n* and *N*.

Both *p* and *n* vary systematically with various paradigms of use, although the mechanisms are not fully understood. A recent finding is that use regulates release by influencing the rate at which vesicles are made available for re-release. Also, the response of NMJs to various paradigms of use is highly dependent on species and type of innervated muscle. Of the many factors involved, some potentiate *m*, while others depress it; the behavior of a particular synapse is the algebraic sum of many genetic and use-dependent influences.

Asynchronous Transmitter Release

Ca^{2+} levels, and consequently the probability of vesicle fusion and transmitter release, increase dramatically when a terminal is invaded by an AP, as already discussed. But intracellular Ca^{2+} does not return to resting levels immediately. When several APs arrive in rapid sequence, residual elevated Ca^{2+} in the terminal can increase *p* between APs (as shown in Figure 4) and for a period (seconds to minutes) after the train is completed. Among the consequences is an increase in mEPP frequency. This release is in addition to that ascribed to the quantal content, *m*. It is evoked, but not synchronized, with the stimulus. At NMJs, the role of asynchronous release is unknown. At certain other synapses, it may be the dominant form of release under certain conditions.

Vesicle Processing

The number of vesicles stored within a typical motor terminal is in the range from 10^5 to 10^6. There is evidence that not all are releasable; for example, measurements in snake set the releasable number at

130 000 (the functional releasable pool). As mentioned previously, motor terminals must rely on their own machinery to remake spent vesicles. The process includes reuptake of vesicular membrane, formation of new vesicles, separation of vesicle proteins from those endogenous to the plasma membrane, reintroduction of these proteins into the vesicular membrane, reuptake of choline (ACh is hydrolyzed into choline and acetate), synthesis of new ACh, packaging of ACh (and other molecules) into vesicles, and, finally, mobilization of reformed vesicles into the releasable pool.

Reuptake of Vesicular Membrane by Endocytosis

Endocytosis is used by cells for many purposes – for example, to sample and obtain nutrients from the extracellular milieu. There are several variants, including bulk membrane invagination (phagocytosis and pinocytosis, which mean, respectively, to eat and to drink the contents of the extracellular fluid) and clathrin-mediated endocytosis (CME), which is often associated with specific receptors. At NMJs, both bulk membrane invagination (which we call macroendocytosis) and CME are utilized. The function of endocytosis at nerve terminals (and other secretory cells) is unique, however, and is called compensatory – compensating for vesicular exocytosis and thereby maintaining a relatively constant plasma membrane area. Without endocytosis (as, e.g., in muscles treated with black widow spider venom, nerve terminals swell and are no longer congruent with the postsynaptic membrane.

Methods for Observing Endocytosis

Compensatory endocytosis was documented in 1973 at the frog NMJ by two independent groups, led by J Heuser, and B Ceccarelli. Both groups utilized transmission EM and a bath-applied marker (horseradish peroxidase or dextran) which is endocytosed and thereby labels the internalized endocytic structures (see Figure 6). Heuser observed CME (evidenced by a clathrin coat surrounding endocytosed vesicles loaded with horseradish peroxidase) at some distance from AZs, while Ceccarelli observed virtually the opposite: labeled vesicles located near AZs, and lacking clathrin coats. Endosomes (called cisternae), now known to be the products of macroendocytosis (ME), were also labeled. These were initially thought to be formed by fusion of vesicles internalized by CME, just as similar endosomes are formed in nonneuronal cells.

An optical method for visualizing endocytosis at NMJs, particularly in living terminals, emerged with the development of sensitive fluorescent endocytic probes (see Figure 7) by Betz and colleagues in 1992. Use of optical probes, and other recent techniques, has confirmed Heuser's observations and possibly Ceccarelli's as well. The current model of endocytosis (Figure 6) includes three different endocytic mechanisms. The first is consistent with Ceccarelli's model and is called kiss-and-run transmitter release – the brief fusion of a vesicle with the plasma membrane to release its contents, followed by immediate withdrawal. This occurs at the AZ, and clathrin is not required. Kiss-and-run has been observed in central synapses but remains putative at vertebrate NMJs. The second is CME (Heuser's model), which can occur, depending on the species, either near or away from the AZ. The third is ME, which can also occur both near and away from the AZ. Macroendocytosis directly produces most, if not all, of the endosomes, that were first thought to be created by vesicle fusion. These endosomes in turn dissipate, via a clathrin-dependent process, into vesicles. While the three processes may differ in speed, location, and other details, their product is the same: vesicles of the proper size (~50 nm) for refilling with transmitter.

Refilling of Vesicles for Reuse

An enzyme within the nerve terminal, choline acetyltransferase (ChAT), resynthesizes ACh from choline that has been taken up from the cleft. Loading of vesicles involves acidification of the vesicle's lumen via a proton pump, and use of the resulting proton gradient to drive a proton–ACh exchanger, or transporter. Both the pump and the ACh transporter are among the proteins resident in the vesicle membrane. The time required for refilling is ~1 min or more in snake (at room temperature), and probably much faster in mammals.

Mobilization of Reformed Vesicles to AZs

Pathways for recycling vesicles have been studied extensively at the frog and snake NMJ, each providing a different model. In frog, there are two distinct recycling strategies, which are differentially recruited according to levels of use. Each strategy employs its own vesicle pool, either a small (local) pool associated with a particular AZ, or a global (reserve) pool that occupies much of the bouton (see Figure 6). The first strategy, associated with low levels of use (e.g., brief low-frequency stimulation), comprises exocytosis at the AZ, endocytosis near the AZ (either CME or putative kiss-and-run), mobilization of spent vesicles into the local pool (which is not anatomically distinct, but distributed among vesicles of the reserve pool), and return of the same vesicles to the AZ for preferential reuse. Thus a particular vesicle is locally

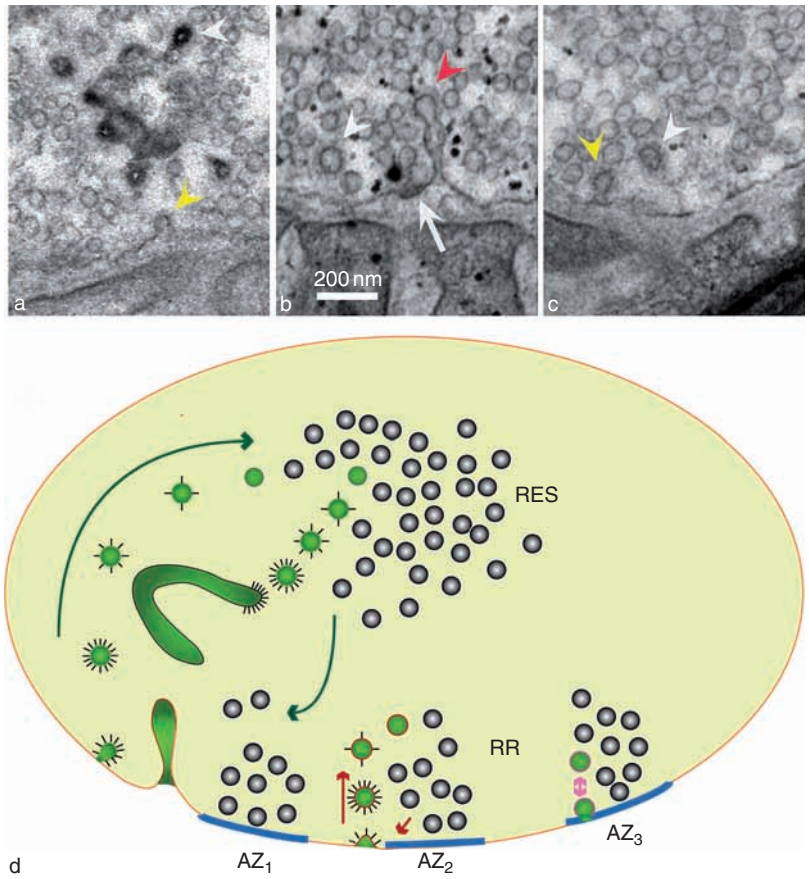

Figure 6 Current models of vesicle processing at the vertebrate neuromuscular junction. (a–c) Electron micrographs of snake boutons, illustrating structures common to all models. Endocytosed structures are labeled with horseradish peroxidase (a) or photoconverted fluorescent endocytic probe FM1-43 (b, c). Fuzzy clathrin coats are visible on vesicles that have been recently endocytosed by clathrin-mediated endocytosis (white arrowheads in (b) and (c) as well as on those still continuous with the presynaptic membrane (yellow arrowheads). A large, recently internalized endosome in a has budded several vesicles, some still clathrin coated (white arrowhead in (a). In (b), note the active zone (white arrow), local clathrin-mediated endocytosis (labeled, coated vesicle just left of the active zone), and macroendocytosis (deep membrane invagination already in the process of budding; red arrowhead). (d) Diagram depicting three types of endocytosis. On the left, macroendocytosis and clathrin-mediated endocytosis occur away from active zone 1 (AZ_1); vesicles budded from the membrane, and from the endosome, enter the reserve pool (RES), which then repopulates the readily releasable (RR) pool near AZ_1. In the center, clathrin-mediated endocytosis occurs locally near AZ_2, and reformed vesicles immediately repopulate the RR pool of AZ_2. On the right, a vesicle undergoes partial membrane fusion at AZ_3, then returns to the RR pool of AZ_3 via kiss-and-run transmitter release.

exo- and endocytosed repeatedly, without involvement of the reserve pool, and without involvement of plasma membrane distal to the AZ. The second strategy, associated with high levels of use (e.g., prolonged high-frequency stimulation), comprises exocytosis at the AZ but endocytosis some distance away, followed by mobilization of the vesicle to the reserve pool, where it remains for some time before it makes its way to the AZ for reuse (possibly via a myosin–actin motor). Endocytosis in this case is via ME (followed by budding of vesicles from the endosome) and probably CME directly from the membrane as well. In effect, mechanisms very similar to those proposed early on by Heuser and Ceccarelli are both used, albeit under different conditions. This two-tiered system of recycling has been observed at other synapses – for example, the glutamatergic NMJs of most *Drosophila* muscles. In contrast, snake NMJs utilize the same endocytic strategies as in frog (CME, particularly near the AZ, and ME), but both operate concomitantly at high and low levels of use. Moreover, both strategies supply vesicles exclusively to the reserve pool; local recycling has not been observed at the snake NMJ.

Vesicle Processing as a Presynaptic Determinant of Synaptic Strength

Endocytosis, filling with transmitter, and mobilization of vesicles to release sites require time and energy (hydrolysis of ATP or GTP). Consequently, these

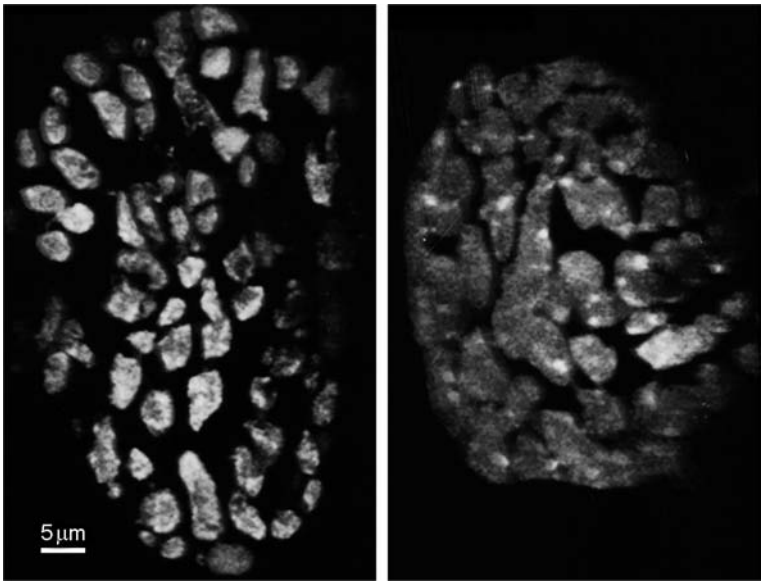

Figure 7 Snake motor nerve terminals stained with the fluorescent endocytic probe FM1-43. (Left) Typical twitch fiber terminal. (Right) Terminal from same animal treated with black widow spider venom, which inhibits compensatory endocytosis. Note severe swelling of boutons.

processing steps have the potential to limit the availability of reformed vesicles for release, particularly with high levels of use. Recent studies indicate that this is indeed the case. Thus *m*, the mean quantal content, depends not only on the exocytotic mechanisms discussed herein but on recycling as well. Once viewed as simple housekeeping, vesicle processing can be rate limiting for transmitter release. Vesicle processing is therefore a subject of considerable research interest, particularly as a potential new substrate for the pharmacological regulation of synaptic strength.

See also: Motor Neuron Specification in Vertebrates; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Acetylcholinesterases; Neuromuscular Connections: Vertebrate Patterns of; Neuromuscular Junction (NMJ): Postsynaptic Events in Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Non-Quantal Release of Transmitter; Presynaptic: Mitochondria and Presynaptic Function.

Further Reading

Bennett MR and Kearns JL (2000) Statistics of quantal release at nerve terminals. *Progress in Neurobiology* 60: 546–606.

del Castillo J and Katz B (1954) Quantal components of the end-plate potential. *Journal of Physiology* 124: 560–573.

Fatt P and Katz B (1951) An analysis of the end-plate potential recorded with an intra-cellular electrode. *Journal of Physiology* 115: 320–370.

Harlow ML, Ress D, Stoschek A, et al. (2001) The architecture of active zone material at the frog's neuromuscular junction. *Nature* 409: 479–484.

Martin AR (1955) A further study of the statistical composition of the end-plate potential. *Journal of Physiology* 130: 114–122.

Rizzoli SO and Betz WJ (2005) Synaptic vesicle pools. *Nature Reviews Neuroscience* 6: 57–69.

Royle SJ and Lagnado L (2003) Endocytosis at the synaptic terminal. *Journal of Physiology* 553: 345–355.

Zefirov AL, Benish T, Fatkullin N, et al. (1995) Localization of active zones. *Nature* 376: 393–394.

Presynaptic Facilitation

R D Hawkins, **E R Kandel, and C H Bailey,** College of Physicians and Surgeons of Columbia University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Synaptic plasticity has long been thought to be a fundamental mechanism of learning and memory. However, the cellular and molecular changes that underlie memory have proven difficult to study in the mammalian brain because the effects are often modest and the contribution of individual synapses to the learning process is not yet well defined. To bridge this gap, the more tractable nervous systems of higher invertebrates have proven useful for the analysis of behavioral problems and have enhanced our knowledge about the synaptic loci and mechanisms that underlie various elementary forms of learning and memory. One such model system has been the gill- and siphon-withdrawal reflex of the marine invertebrate *Aplysia californica*. This reflex exhibits several forms of learning including dishabituation, sensitization, and classical conditioning that have many of the behavioral features of learning in mammals, suggesting that learning in *Aplysia* and mammals may share common mechanisms.

Recent studies of a variety of memory processes, ranging in complexity from simple forms of memory in invertebrates to more complex forms in mammals, have supported the view that changes in the strength and structure of synaptic connections contribute critically to these diverse forms of memory storage. Studies of *Aplysia* have focused on presynaptic changes, whereas studies on mammals have focused primarily on postsynaptic changes. However, there is evidence for both pre- and postsynaptic mechanisms in both experimental systems, and recent experiments suggest that they have additional similarities. In particular, for both invertebrates and mammals, two general types of storage mechanisms have been described: short-term memory lasting minutes and long-term memory lasting days, weeks, or longer. This temporal distinction in behavior is reflected in specific forms of synaptic plasticity that underlie each form of behavioral memory as well as specific molecular requirements for each of these two forms of synaptic plasticity. The short-term forms involve the covalent modifications of preexisting proteins by a variety of kinases and are expressed as alterations in the effectiveness of preexisting connections. By contrast, in addition to protein phosphorylation, the long-term form also requires CREB-mediated gene expression and new mRNA and protein synthesis. Moreover, the long-term form often is associated with the growth of new synaptic connections, which is thought to represent the final and self-sustaining change that stabilizes the long-term process. In addition to short- and long-term memory, a family of intermediate processes that last one or more hours and often require translation but not transcription can be produced by various training protocols using repeated or prolonged stimulation.

In this article, we discuss and compare the underlying cellular and molecular mechanisms of short- and intermediate-term (Figure 1) and long-term (Figure 2) memory storage that have been identified by neurobiological studies of elementary forms of learning in *Aplysia*.

Sensitization and Classical Conditioning of the Gill-Withdrawal Reflex in *Aplysia*: Two Elementary Forms of Learning

The nervous system of *Aplysia* contains only approximately 20 000 large, identifiable nerve cells, clustered into 10 major ganglia. The ability to identify individual neurons and record their activity has made it possible to define the major components of the neuronal circuits of specific behaviors and to delineate the critical sites and underlying mechanisms used to store memory-related representations.

The molecular mechanisms contributing to memory storage have been extensively studied for the gill- and siphon-withdrawal reflex of *Aplysia*. As is true for other types of defensive reflexes, the gill- and siphon-withdrawal reflex can be modified by several different forms of learning. We begin by focusing on sensitization, an elementary form of nonassociative learning by which an animal learns about the properties of a single noxious stimulus and enables the formation of a learned fear response. When a light touch is applied to the siphon of the snail, the snail responds by withdrawing its gill and siphon. This response is enhanced when the animal is given a noxious, sensitizing stimulus. As with other forms of defensive behaviors, the memory for sensitization of the withdrawal reflex is graded, and repeated tail shocks lead to a longer-lasting memory. A single tail shock produces short-term sensitization that lasts for minutes, whereas repeated tail shocks given at spaced intervals produce long-term sensitization that lasts for up to several weeks. The reflex also exhibits

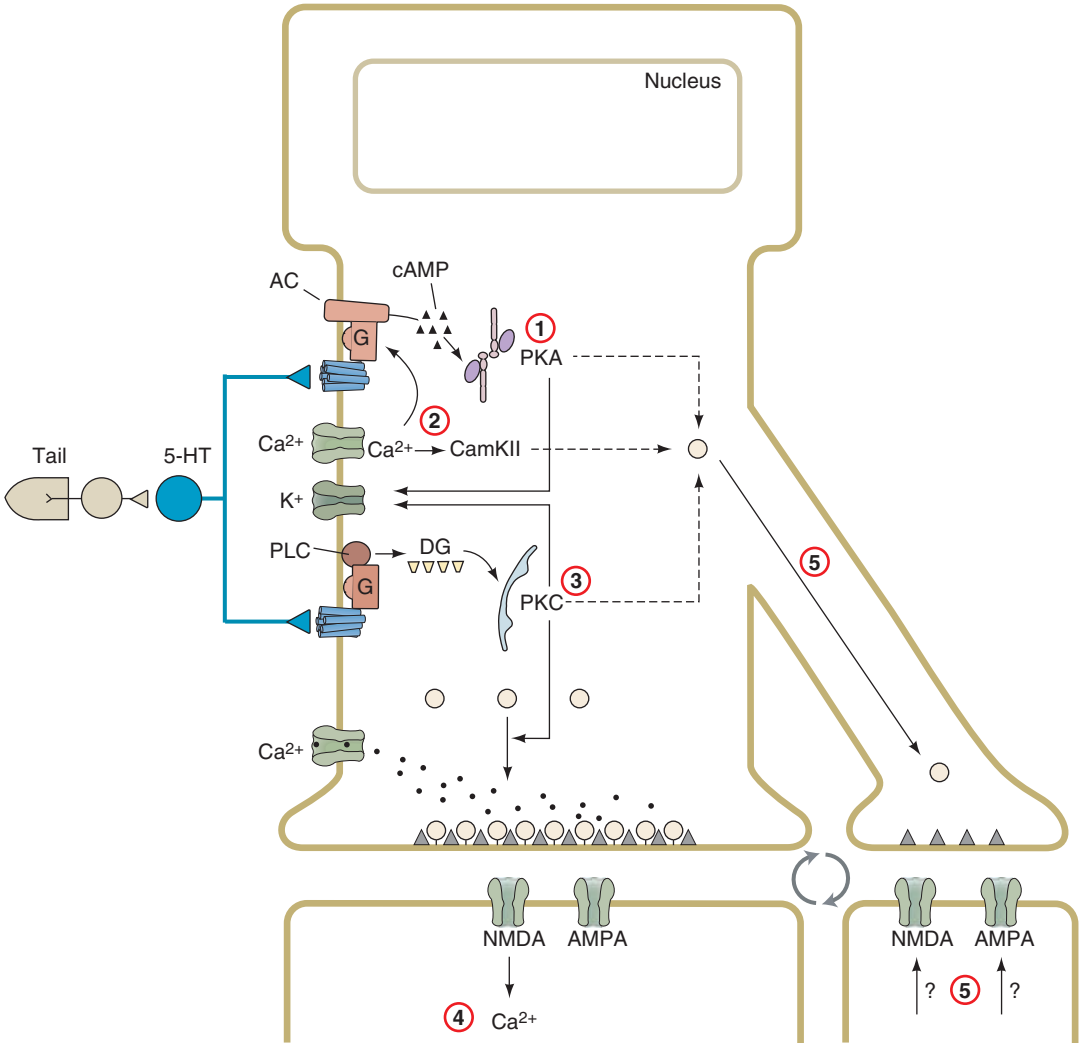

Figure 1 Mechanisms of short- and intermediate-term memory formation in *Aplysia*. Different forms of short- and intermediate-term synaptic plasticity contributing to learning in *Aplysia* involve different combinations of molecular pathways including (1) presynaptic PKA, (2) presynaptic Ca^{2+} and CamKII, (3) presynaptic PKC, (4) postsynaptic Ca^{2+}, and (5) recruitment of pre- and possibly postsynaptic molecules to new sites. See the text for details.

classical conditioning, an associative form of learning by which an animal learns about the predictive relationship between two stimuli. Enhancement of the withdrawal reflex is greater and longer lasting if the siphon is touched just before the noxious, sensitizing stimulus (paired training), compared to unpaired training or training with either stimulus alone.

The simplicity of the neuronal circuit underlying these behavioral modifications – including direct monosynaptic connections between identified mechanoreceptor sensory neurons and their follower cells – has allowed reduction of the analysis of the short- and long-term memory for sensitization to the cellular and molecular level. This monosynaptic sensory to motor neuron connection, which is thought to be glutamatergic, can be reconstituted in dissociated cell culture. A number of studies have demonstrated that this simplified *in vitro* model system reproduces what is observed during behavioral training by replacing the tail shocks with brief applications of serotonin (5-HT), a modulatory transmitter normally released by sensitizing stimuli in the intact animal. A single brief application of 5-HT produces a short-term change in synaptic effectiveness (short-term facilitation or STF), whereas repeated and spaced applications produce changes in synaptic strength that can last for more than a week (long-term facilitation, or LTF). The facilitation is also larger and longer lasting if the presynaptic sensory neuron fires action potentials just before the 5-HT application, analogous to classical conditioning.

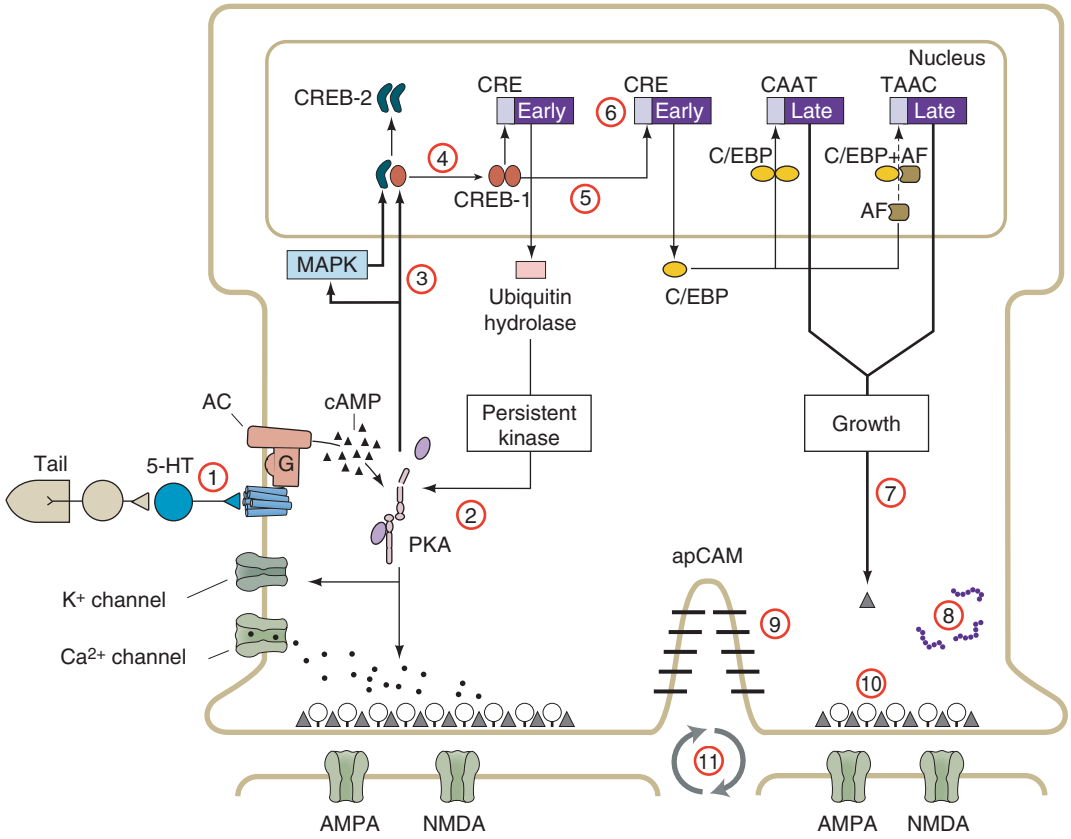

Figure 2 Mechanisms of long-term memory formation. Long-term synaptic plasticity contributing to learning involves a sequence of cellular and molecular mechanisms including (1) neurotransmitter release and short-term strengthening of synaptic connections, (2) equilibrium between kinase and phosphatase activities at the synapse, (3) retrograde transport from the synapse to the nucleus, (4) activation of nuclear transcription factors, (5) activity-dependent induction of gene expression, (6) chromatin alteration and epigenetic changes in gene expression, (7) synaptic capture of newly synthesized gene products, (8) local protein synthesis at active synapses, (9) synaptic growth and the formation of new synapses, (10) activation of preexisting silent synapses, and (11) self-perpetuating mechanisms and the molecular basis of memory persistence. The location of these events, which may act in part to stabilize some of the changes that occur during short- and intermediate-term plasticity, moves from the synapse (1–2) to the nucleus (3–6) and then back to the synapse (7–11). Molecular details are discussed in the text.

Cellular and Molecular Mechanisms Underlying Short- and Intermediate-Term Forms of Memory Storage

STF of Synaptic Connections Involves Enhanced Transmitter Release from the Sensory Neurons

S-HT released *in vivo* during sensitization or applied directly to cultured neurons binds to cell surface receptors on the sensory neurons and promotes the production of the diffusible second messenger cAMP by activating the enzyme adenylyl cyclase (AC). This increase in intracellular concentration of cAMP results in short-term behavioral sensitization lasting minutes and correlates with an increase in synaptic strength of the sensory to motor neuron connection referred to as STF. This facilitation is due to the enhanced release of the transmitter glutamate by the sensory neuron onto its follower cells and is accompanied by an increase in excitability of the sensory neuron attributable to the depression of specific sets of potassium channels. In addition, the changes in cAMP and Ca^{2+} levels triggered by the activation of S-HT receptors and ionic channels regulate the activity of different kinase and phosphatase activities that control the duration and strength of the changes in synaptic efficiency, as we discuss later. These second messengers mediate the transient reinforcement of synaptic connections by covalent modifications of channel activation and the enhancement of neurotransmitter release at presynaptic terminals.

Different *In Vitro* Training Protocols Elicit Facilitation That Involves Different Molecular Mechanisms

Previous studies have shown that facilitation of the sensory to motor neuron synaptic connection

involves different mechanisms depending on the duration of exposure to 5-HT and the state (rested or depressed) of the synapse. Facilitation at rested synapses by a relatively brief exposure to 5-HT involves activation of AC and cAMP-dependent protein kinase A (PKA) in the sensory neuron, leading to reduced K^+ current, increased action potential duration, increased Ca^{2+} influx, and increased transmitter release, as described above. A longer exposure to 5-HT recruits activation of protein kinase C (PKC), which can also increase the duration of presynaptic action potentials. Facilitation at depressed synapses involves PKC as well, but in this case it acts through a mechanism that is independent of spike broadening and is thought to involve vesicle mobilization. Facilitation by short or longer exposure to 5-HT at rested synapses may also involve Ca^{2+}/calmodulin-dependent protein kinase (CamKII). In addition, longer exposure to 5-HT can induce intermediate-term facilitation, which requires protein synthesis and involves it mitogen-activated protein (MAP) kinase as well as PKA or (under some circumstances) PKC, and can involve postsynaptic mechanisms as well.

Behavioral Dishabituation and Sensitization Involve Different Molecular Mechanisms of Facilitation

A major question in psychology is how different forms of learning such as dishabituation, sensitization, and classical conditioning are related. The sensory–motor neuron synapses exhibit cellular analogs of these forms of learning with temporal parameters similar to behavioral learning, making it possible to examine how the analogs are related at the molecular level. However, these studies of synaptic changes in the isolated nervous system have not been able to address directly the contribution of the different molecular mechanisms of plasticity to behavior. That question can be addressed in a simplified preparation for studying the siphon-withdrawal reflex of *Aplysia*, with which it is relatively easy to record the activity of identified neurons and their synaptic connections simultaneously with behavioral learning. This preparation undergoes dishabituation, sensitization, and classical conditioning that are similar parametrically to learning in intact animals. The neural circuit mediating the withdrawal reflex in the simplified preparation has been well characterized, and consists in part of monosynaptic connections from LE siphon sensory neurons to LFS siphon motor neurons.

During behavioral dishabituation and sensitization of siphon withdrawal, monosynaptic postsynaptic potentials (PSPs) from LE neurons that either fire during the siphon tap (on-field) or do not (off-field) change approximately in parallel with changes in the behavior, supporting the idea that both forms of learning involve heterosynaptic facilitation of sensory–motor neuron PSPs. However, pharmacological studies suggest that sensitization involves both presynaptic PKA and CamKII whereas dishabituation involves PKC, similar to facilitation at nondepressed and depressed synapses *in vitro*.

Activity-Dependent Plasticity in *Aplysia* Involves Both Pre- and Postsynaptic Mechanisms

In addition to these nonassociative, neuromodulatory forms of plasticity, *Aplysia* sensory–motor neuron synapses also exhibit two associative, activity-dependent cellular mechanisms: *N*-methyl-D-aspartate (NMDA)-dependent long-term potentiation (LTP), which occurs when presynaptic spike activity is temporally paired with postsynaptic depolarization (Pre–Post), and activity-dependent facilitation, which occurs when presynaptic spike activity is temporally paired with 5-HT (Pre–Mod). During activity-dependent facilitation, the spike activity is thought to cause an influx of Ca^{2+} that 'primes' the AC, leading to enhanced activation of the PKA pathway. Consistent with that idea, activity-dependent facilitation can be blocked by injecting either the slow Ca^{2+} chelator EGTA or a peptide inhibitor of (PKAi) into the sensory neuron. However, it can also be blocked by injecting the fast Ca^{2+} chelator BAPTA into the motor neuron, suggesting that activity-dependent facilitation involves both pre- and postsynaptic mechanisms.

Classical Conditioning Involves Both Activity-Dependent Facilitation and Hebbian LTP

The siphon-withdrawal reflex in the simplified preparation also exhibits an associative form of learning, classical conditioning. Paired training in which the conditioned stimulus (CS; mechanical stimulation of the siphon) occurs just before the unconditioned stimulus (US; a shock to the tail) produces a greater increase in the withdrawal reflex than unpaired, CS-alone, or US-alone training. Compared with unpaired training, paired training also produces greater facilitation of monosynaptic PSPs from LE neurons that fire during the siphon tap (on-field), supporting the idea that conditioning involves associative, activity-dependent plasticity at sensory–motor neuron synapses.

In addition to changes in the PSP, there are also associative, activity-dependent increases in both the evoked firing and input resistance of the LE sensory neurons during conditioning, consistent with the idea that the synaptic plasticity involves activity-dependent presynaptic facilitation. However, it could also involve Hebbian LTP because the CS causes firing of LE

sensory neurons and the US causes firing of both facilitatory interneurons and the LFS motor neurons. Consistent with that idea, exposing the abdominal ganglia to either the PKA inhibitor KT5720 or the NMDA antagonist 2-amino-5-phosphonopentanoic acid (APV) blocks behavioral conditioning. Similarly, either injecting several LE sensory neurons with a PKAi or injecting two to three LFS motor neurons with BAPTA significantly reduces behavioral conditioning. These results suggest that both activity-dependent presynaptic facilitation and Hebbian LTP are required for the conditioning.

Furthermore, either injecting an LE neuron with PKAi or an LFS neuron with BAPTA significantly reduces pairing-specific facilitation of the PSP between the neurons during conditioning, providing the strongest evidence to date that either activity-dependent presynaptic facilitation or Hebbian LTP contributes to synaptic plasticity underlying behavioral learning. Moreover, the facilitation is almost completely blocked by either injection, suggesting that the pre- and postsynaptic mechanisms are more than additive. Surprisingly, either injecting an LE neuron with PKAi or injecting an LFS motor neuron with BAPTA also significantly reduces the effects of pairing on evoked firing and membrane resistance of the LE sensory neuron, suggesting that the pre- and postsynaptic neurons interact through retrograde signaling. Collectively, the results of these experiments suggest that the synaptic plasticity during conditioning does not simply involve separate and independent pre- and postsynaptic mechanisms. Rather, they fit better with the idea that facilitation of the PSP during conditioning involves both pre- and postsynaptic changes coordinated by transsynaptic interactions.

Cellular and Molecular Mechanisms Underlying Long-Term Forms of Memory Storage

Gating Signals at the Synapse: A Balance between the Activities of Protein Kinases and Phosphatases

Synaptic stimulation of *Aplysia* sensory neurons leads to a local increase in cAMP and the activation of the cAMP-dependent PKA by causing the catalytic subunits of this enzyme to dissociate from the regulatory subunits. The catalytic subunits can then phosphorylate different substrates in the synaptic terminals, such as potassium channels and proteins involved in exocytosis, leading to enhanced transmitter availability and release as described above for the storage of short-term memory. When synaptic stimulation reaches a given threshold or is repeated a number of times, it causes a persistent increase in the level of cAMP and leads to longer-lasting forms of synaptic plasticity. At the molecular level, this more robust pattern of stimulation causes the catalytic subunit of PKA to recruit p42 MAPK and both then move to the nucleus where they phosphorylate nuclear targets, including other kinases that, in turn, can phosphorylate transcription factors and activate gene expression required for the induction of long-term memory.

In addition to protein kinases, synaptic protein phosphatases also play a key role in regulating the initiation of long-term synaptic changes. Various protein phosphatases, such as PP1 and calcineurin, counteract the local activity of PKA acting as inhibitory constraints of memory formation. For example, recent experiments in cultured *Aplysia* neurons indicate that calcineurin may act as a memory suppressor for sensitization in this organism. Thus, an equilibrium between both kinase and phosphatase activities at a given synapse gates the synaptic signals that reach the nucleus and, thus, can regulate both memory storage and retrieval.

Retrograde Signaling from the Synapse to the Nucleus

One of the features that fundamentally distinguishes the storage of long-term memory from short-term cellular changes is the requirement for the activation of gene expression. Given this requirement at the nucleus, one might expect that LTF would have to be cell wide. However, experiments using local applications of 5-HT in an *Aplysia* bifurcated sensory neuron–two motor neuron culture preparation, as well as parallel experiments in the hippocampus, demonstrated that each synapse could be modified independently in a protein synthesis-dependent manner. Thus, LTF and the associated presynaptic changes are synapse specific. This implies that there must be not only anterograde signaling from the nucleus to the synapse, but also retrograde signaling from the synapse back to the nucleus. 5-HT stimulation which produces LTF triggers the nuclear translocation of importins, proteins involved in carrying cargoes through nuclear pore complexes, suggesting that importins may play important roles in the nuclear signaling during LTF.

Activation of Nuclear Transcription Factors and Induction of cAMP-Responsive Genes

Long-term memory is represented at the cellular level by activity-dependent modulation of both the function and the structure of specific synaptic connections that, in turn, depends on the activation of specific patterns

of gene expression. As mentioned above, the inhibition of transcription or translation blocks the formation of long-term memory in a variety of model systems, but does not affect short-term memory.

Studies in *Aplysia* revealed the participation of the cAMP/PKA-signaling pathway in LTF and sensitization. In mammalian cells, PKA activates gene expression by the phosphorylation of transcription factors that bind to the cAMP-responsive element (CRE). The CRE is one of the DNA response elements contained within the control region of a gene. The binding of different transcription factors to these response elements regulates the activity of RNA polymerase, thereby determining when and to what level a gene is expressed. One of the major transcription factors that recognizes the CRE is a protein called CRE-binding protein (CREB1), which functions as a transcriptional activator only after it is phosphorylated by either PKA, MAPK, or CamK. The role of CREB1 in *Aplysia* LTF was examined by microinjecting CRE oligonucleotides into sensory neurons cocultured with motor neurons. This oligonucleotide inhibits the function of CREB1 by binding to the CREB1 protein within the cell, thereby preventing it from binding to CRE sites in the regulatory regions of cAMP-responsive genes and activating gene expression. While injection of the CRE oligonucleotide had no effect on STF, it selectively blocked LTF. Several studies by a number of laboratories have now revealed that different members of the CREB family of transcription factors participate in the molecular switch that regulates LTF formation. Both the CREB activator ApCREB1 and the repressor ApCREB2 contribute to this process. The formation of LTF requires the activation of ApCREB1 by PKA and the concomitant downregulation of ApCREB2 by MAPK. Injection of anti-ApCREB2 antibodies into *Aplysia* sensory neurons causes a single pulse of 5-HT, which normally induces STF lasting minutes, to evoke LTF that lasts several days. Conversely, the injection of pApCREB1 can by itself trigger facilitation lasting 24 h and this can be stabilized by a single pulse of 5-HT.

The activity of ApCREB1 leads to the expression of several immediate-response genes, such as ubiquitin hydrolase that stabilizes STF, and the transcription factor CCAAT-box-enhanced binding-protein (C/EPB), whose induction has been shown to be critical for LTF. This induced transcription factor (in concert with other constitutively expressed molecules such as ApAF) activate a second wave of downstream genes that lead to the growth of new synaptic connections. These genes represent only two of a family of physiological relevant examples of gene products generated by CREB activity.

Chromatin Alteration and Epigenetic Changes in Gene Expression with Memory Storage

Although epigenetic mechanisms were widely known to be involved in the formation and long-term storage of cellular information in response to transient environmental signals during development, the discovery of their putative relevance in adult brain function is relatively recent. The epigenetic marking of chromatin, such as histone modification, chromatin remodeling, and the activity of retrotransposons, may have long-term consequences in the transcriptional regulation of specific loci involved for long-term synaptic changes.

Histone tail acetylation, a modification that favors transcription and is associated with active loci, contributes to LTF formation in *Aplysia* neurons. Both facilitatory and inhibitory stimuli bidirectionally alter the degree of acetylation and the accessibility of promoters driven by the expression of genes involved in the maintenance of LTF, such as C/EBP. In addition, enhancing histone acetylation with deacetylase (HDAC) inhibitors facilitates the induction of LTF. These results indicate that critical chromatin changes occur during the formation of long-term memory and that these changes are required for the stable maintenance of these memories.

Synaptic Capture of Learning-Induced Gene Products

Following transcriptional activation, newly synthesized gene products, both mRNAs and proteins, have to be delivered specifically to the synapses whose activation originally triggered the wave of gene expression. The synaptic capture hypothesis has been proposed to explain how this specificity can be achieved in a biologically economical way in spite of the massive number of synapses in a single neuron. This hypothesis, also referred to sometimes as synaptic tagging, suggests that the products of gene expression are delivered throughout the cell, but are only functionally incorporated in those synapses that have been tagged by previous synaptic activity. The 'synaptic tag' model has been supported by a number of studies both in the rodent hippocampus and *Aplysia*.

Studies of synaptic capture at the synapses between the sensory and motor neurons of the gill-withdrawal reflex in *Aplysia* have demonstrated that to achieve synapse-specific LTF, more than the production of CRE-driven gene products in the nucleus is necessary. One also needs a PKA-mediated covalent signal to mark the stimulated synapses and local protein synthesis to stabilize that mark. Thus, injection into the cell body of phosphorylated CREB-1 gives rise to

LTF at all the synapses of the sensory neuron by seeding these synapses with the protein products of CRE-driven genes. However, this facilitation is not maintained beyond 24–48 h unless the synapse is also marked by the short-term process, a single pulse of 5-HT.

Local Protein Synthesis at Active Synapses

A number of distinct mRNAs have now been localized in the axons of *Aplysia* and the dendrites of mouse hippocampal neurons. The molecular mechanisms that carry these mRNAs to the synapse are largely unknown, but they are thought to involve the recognition of *cis*-acting elements in their 3′ untranslated region by specific RNA-binding proteins that interact with the cytoskeleton. Once translocated to synaptic compartments, these mRNAs are efficiently translated only after docking at active synaptic sites, a process frequently referred to as synaptic or local protein synthesis. Recent studies indicate that regulation of local protein synthesis may play a major role in the control of synaptic strength.

The role of local protein synthesis in the maintenance of synapse-specific LTF in *Aplysia* has been investigated using a culture system in which a single bifurcated sensory neuron was plated in contact with two spatially separated motor neurons. In this culture system, repeated application of 5-HT to one synapse produces a CREB-mediated, synapse-specific LTF that can be blocked by the local application of inhibitors of translation, suggesting that local protein synthesis at the synapse is required for the initiation of synapse-specific LTF.

The control of translation at the synapse is likely to be complex and involves several different mechanisms, including mRNA transport and docking, cytoplasmic polyadenylation, and the phosphorylation of different translation factors. Many of the molecules contributing to the regulation of this process have been found to be required for both LTF in *Aplysia* and L-LTP in the hippocampus. This is the case for the kinase mTOR (mammalian target of rapamycin), which is the target of the selective protein synthesis inhibitor rapamycin, the brain-derivate neurotrophic factor (BDNF), which promotes local protein synthesis, and the cytoplasmic polyadenylation element binding protein (CPEB), which activates dormant mRNAs.

Synaptic Growth and the Formation of New Synapses

Synaptic remodeling and the growth of new synaptic connections have been found to accompany various forms of long-term memory. This is particularly well documented at the connections of identified sensory neurons in *Aplysia*. The storage of long-term memory for sensitization (lasting several weeks) is accompanied by an increase in the number of sensory neuron active zones (transmitter release sites) as well as an increase in the total number of presynaptic varicosities (boutons) per sensory neuron, both of which parallel the duration of the behavioral memory. This synaptic growth can be reconstituted in dissociated sensory–motor neuron co-cultures by repeated applications of 5-HT, and requires altered gene expression and new protein synthesis. The 5-HT-induced synaptic growth is also associated with the modulation of neural cell adhesion molecule (NCAM)-related cell adhesion molecules and reorganization of the presynaptic actin network.

Activation of Silent Synapses

To monitor continuously the functional and structural changes during LTF, recent studies have combined time-lapse confocal imaging of individual presynaptic sensory neuron varicosities labeled with different fluorescent markers. Repeated pulses of 5-HT induce two distinct classes of presynaptic changes: (1) a rapid (3–6 h) activation of silent presynaptic terminals through the filling of preexisting empty varicosities with synaptic vesicles, which requires translation but not transcription, and (2) somewhat slower (12–18 h) generation of new synaptic varicosities, which requires both transcription and translation. The rapid activation of silent presynaptic terminals suggests that in addition to its role in LTF, this modification of preexisting synapses may also contribute to the intermediate phase of synaptic plasticity and memory storage. The activation of silent synapses also plays a major role in LTP in the hippocampus, although in that case the term refers to membrane insertion of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors at postsynaptic sites previously lacking them. A similar regulation involving the insertion and clustering of postsynaptic neurotransmitter receptors has recently been described during LTF at sensory–motor neuron synapses in *Aplysia*.

Self-Perpetuating Mechanisms and the Molecular Basis of Memory Persistence

What is the nature of the enduring molecular changes that underlie long-term memory storage? One possible answer to this elusive question is that some type of self-sustained alteration at the synaptic level is required to confer stability to memory storage. Several molecular mechanisms have been proposed as potential candidates for this process. We discuss here one recent model to account for the stable maintenance and persistence of LTF in *Aplysia*.

Prion-like proteins represent autoreplicative structures that may serve as a persistent form of information. It has recently been proposed that the persistence of long-term synaptic plasticity might be explained by the prion-like properties of *Aplysia* neuronal CPEB. Neuronal CPEB can activate the translation of dormant mRNAs through the elongation of their poly-A tail. *Aplysia* CPEB has two conformational states: one is inactive or acts as a repressor, while the other is active. In a naïve synapse, the basal level of CPEB expression is low and its state is inactive or repressive. According to the model, 5-HT induces an increase in the amount of neuronal CPEB and, if a given threshold is reached, causes the conversion to the aggregated prion-like state, which is more active and lacks the inhibitory function of the basal state. Once the prion state is established at an activated synapse, dormant mRNAs, made in the cell body and distributed cell wide, would be translated only at the activated synapses. Because the activated CPEB can be self-perpetuating, it could contribute to a self-sustaining synapse-specific long-term molecular change and provide a mechanism for the stabilization of learning-related synaptic growth and the persistence of memory storage.

An Overall View

Long-lasting forms of learning-related synaptic plasticity are often associated with changes in both the number and structure of synapses. Some of the best evidence for changes in synapse number has come from studies of long-term (days to weeks) plasticity in *Aplysia*, where long-term sensitization *in vivo* and LTF *in vitro* are both accompanied by increases in the number of presynaptic sensory neuron varicosities. The late (hours) protein synthesis-dependent phase of hippocampal LTP is also accompanied by increases in both postsynaptic and presynaptic structures, consistent with an increase in synapse number.

By contrast, short-term (up to about 1 h) plasticity and the early phases of long-term plasticity in *Aplysia* and hippocampus have been thought to involve a completely different type of mechanism, covalent modification of existing proteins in either presynaptic (*Aplysia*) or postsynaptic (hippocampus) structures. However, the discovery of intermediate (2–3 h) forms of plasticity that involve elements of the short-term and long-term mechanisms in both *Aplysia* and hippocampus questioned this dichotomy, and in recent years it has been challenged by two additional types of evidence. First, at least under some conditions, pre- as well as postsynaptic mechanisms have been found to contribute to early-phase LTP in hippocampus and to a variety of types of short-term plasticity in *Aplysia* including homosynaptic potentiation, heterosynaptic facilitation by longer exposure to 5-HT, activity-dependent facilitation, and the synaptic plasticity during behavioral conditioning. In many cases, inhibitors injected into the pre- and postsynaptic neurons have more than additive effects, suggesting that the pre- and postsynaptic mechanisms are not independent but rather act synergistically. Furthermore, in some cases, postsynaptic inhibitors have been shown to affect presynaptic properties, suggesting transynaptic signaling.

Second, imaging studies have shown that even the earliest phases of hippocampal LTP are accompanied by pre- and postsynaptic structural alterations. Tens of minutes after the induction of LTP there is an outgrowth of new pre- and postsynaptic processes, and even earlier (minutes) there are increases in spine size, clusters of postsynaptic glutamate receptors, and clusters of presynaptic vesicle-associated proteins and sites where the pre- and postsynaptic clusters co-localize. Moreover, the presynaptic alterations appear to depend on retrograde signaling from the postsynaptic cells. Similar to clustering of synaptic proteins during LTP, intermediate-term facilitation in *Aplysia* is accompanied by filling of preexisting empty presynaptic varicosities with synaptic vesicles. However, these intermediate-term changes do not persist for 24 h unless they are stabilized by additional molecular events (including the machinery for translational activation) recruited during LTF. LTF is also accompanied by clustering of postsynaptic proteins including the *Aplysia* homologs of NMDA and AMPA receptors. It is not yet known whether intermediate-term facilitation is accompanied by similar postsynaptic changes. However, presynaptic microstructural changes can occur during even earlier phases of learning-related synaptic plasticity in *Aplysia*: homosynaptic potentiation induced by moderate tetanic stimulation of the presynaptic neuron is accompanied by rapid (less than 10 min) aggregation of the vesicle-associated protein synaptophysin into new clusters or puncta, as occurs during early-phase LTP in hippocampal neurons and 5-HT-induced intermediate-term facilitation in *Aplysia*.

Combined, these results suggest that even the early phases of learning-related synaptic plasticity can involve pre- and postsynaptic microstructural alterations that may lead to the formation of new functional synapses, which can then be stabilized or maintained by additional long-term training. This emerging view is likely to supplement (rather than replace) the more established hypothesis that early-phase plasticity exclusively involves either pre- or postsynaptic covalent modifications. In some ways, this new appreciation represents a paradigm shift in thinking about the

cellular and molecular bases of learning and memory, because it suggests that aspects of the pre- and postsynaptic mechanisms that underlie both early- and late-phase plasticity may not be fundamentally different and independent. Rather, these mechanisms appear to have certain structural features in common that may contribute to the transition between short-term and long-term memory storage.

See also: Activity-Dependent Remodeling of Presynaptic Boutons; Cyclic AMP (cAMP) Role in Learning and Memory; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Hebbian Plasticity; Learning and Memory in Invertebrates: *Aplysia*; Long-Term Depression: Cerebellum; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Plasticity: Short-Term Mechanisms.

Further Reading

Bailey CH, Kandel ER, and Si K (2004) The persistence of long-term memory: A molecular approach to self-sustaining changes in learning-induced synaptic growth. *Neuron* 44: 49–57.

Byrne JH and Kandel ER (1996) Presynaptic facilitation revisited: State and time dependence. *Journal of Neuroscience* 16: 425–435.

Hawkins RD, Kandel, ER, and Bailey CH (2006) Molecular mechanisms of memory storage in *Aplysia*. *Biological Bulletin* 210: 174–191.

Hawkins RD, Kandel ER, and Siegelbaum SA (1993) Learning to modulate transmitter release: Themes and variations in synaptic plasticity. *Annual Review of Neuroscience* 16: 625–665.

Jessell TM and Kandel ER (1993) Synaptic transmission: A bidirectional and self-modifiable form of cell–cell communication. *Cell 72/Neuron* 10: 1–30.

Kandel ER (2001) The molecular biology of memory storage: A dialogue between genes and synapses. *Science* 294: 1030–1038.

Lamprecht R and LeDoux J (2004) Structural plasticity and memory. *Nature Reviews Neuroscience* 5: 45–54.

Levenson JM and Sweatt JD (2005) Epigenetic mechanisms in memory formation. *Nature Reviews Neuroscience* 6: 108–118.

Luscher C, Nicoll RA, Malenka RC, and Muller D (2000) Synaptic plasticity and dynamic modulation of the postsynaptic membrane. *Nature Neuroscience* 3: 545–550.

Malinow R, Mainen ZF, and Hayashi Y (2000) LTP mechanisms: From silence to four-lane traffic. *Current Opinion in Neurobiology* 10: 352–357.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* 25: 103–126.

Martin SJ, Grimwood PD, and Morris RG (2000) Synaptic plasticity and memory: An evaluation of the hypothesis. *Annual Review of Neuroscience* 23: 649–711.

Roberts AC and Glanzman DL (2003) Learning in *Aplysia*: Looking at synaptic plasticity from both sides. *Trends in Neuroscience* 26: 662–670.

Steward O and Schuman EM (2001) Protein synthesis at synaptic sites on dendrites. *Annual Review of Neuroscience* 24: 299–325.

Waddell S and Quinn WG (2001) Flies, genes, and learning. *Annual Review of Neuroscience* 24: 1283–1309.

Presynaptic Inhibition

S Alford and E Schwartz, University of Illinois at Chicago, Chicago, IL, USA

© 2009 Published by Elsevier Ltd.

Presynaptic Molecular and Morphological Factors

Synaptic Transmission

Action potential invasion and depolarization of presynaptic terminals induces a Ca^{2+} influx through voltage-gated Ca^{2+} channels. The resulting increase in cytosolic Ca^{2+} concentration triggers vesicle fusion and neurotransmitter release. Neurotransmitter diffusing across the synaptic cleft binds to and activates postsynaptic receptors, usually leading to postsynaptic currents that may be recorded electrophysiologically. Although the tight Ca^{2+}-dependency of vesicle fusion and neurotransmitter release was demonstrated over 40 years ago, the specific mechanisms and the proteins that mediate vesicle fusion are still under investigation.

Recent work has demonstrated that a group of presynaptic proteins constituting the soluble *N*-ethylmaleimide-sensitive factor attached protein receptor (SNARE) complex is necessary for exocytosis. This complex is composed of a bundle of four α-helices thought to bridge synaptic vesicle and cell membranes. These α-helices are donated by three proteins: syntaxin and synaptosomal-associated protein (SNAP)-25, located in the synaptic active zone; and synaptobrevin, in the synaptic vesicle membrane. Synaptotagmin interacts with the C-terminal region of the SNARE complex during Ca^{2+}-dependent fusion and represents a likely candidate Ca^{2+} sensor for transmitter release. Synaptic vesicles fuse with the plasma membrane in microseconds. Consequently, they are located very near to the point of fusion, the active zone, at the presynaptic terminal. This organization is termed docking. It is believed that there is a pool of ready-to-fuse synaptic vesicles that have undergone a further maturation step, priming, that requires adenosine triphosphate (ATP), submicromolar Ca^{2+} concentrations, and alterations in membrane lipids by lipid transferases and kinases. Evoked synaptic fusion is thought to require high (hundreds of micromolar) local concentrations of Ca^{2+} following the action potential invasion of the nerve terminal; and small modifications of this Ca^{2+} entry may cause large changes in neurotransmitter release.

Neurotransmitter release Seminal work by Del Castillo and Katz analyzing synaptic transmission at the frog neuromuscular junction combined with electron microscopic studies demonstrating compensatory endocytosis of vesicle membrane formed the vesicle theory of neurotransmitter release. This theory predicts that vesicle size, and the amount of neurotransmitter contained within the vesicle, determines the lowest unitary response (1 quantum). Thus, in response to increases in Ca^{2+} concentration, vesicles fuse and release neurotransmitter in an all-or-none event. The predicted uniform nature of vesicles and their neurotransmitter cargo led to the interpretation that presynaptic plasticity must be due to changes in the number of quanta (or vesicles) released from the presynaptic terminal in response to a single action potential (a change in probability). Furthermore, changes in the amplitude of individual quantal events must be due to changes in the number or properties of postsynaptic receptors. However, recent evidence suggests that presynaptic factors may also regulate changes in quantal size.

Presynaptic changes in the mean excitatory postsynaptic current (EPSC) amplitude might, therefore, result from a change either in neurotransmitter concentration at each synaptic site or in the number of vesicles released distributed across all of these sites (Figure 1). This distinction is significant because synaptic strength and variability may be determined presynaptically by (1) the quantal size of individual fusion events, (2) the number of active zones within a synaptic connection, or (3) the probability of vesicle fusion at individual synapses. All these starting conditions are affected by a range of variables across synapses.

Before we discuss the mechanisms underlying presynaptic inhibition we must, therefore, consider a range of synaptic substrates on which it plays out.

Synapse Structure

A synapse is most commonly understood as an individual contact between electron-dense pre- and postsynaptic membranes. This is most generally a small contact area with rapid diffusional loss of neurotransmitter from the synapse, although the synapse may have a more complex topography in which the neurotransmitter is temporarily trapped. An extreme example of the latter may be found at the unipolar brush cell in the cerebellum, but similar examples may be found in sensory inputs (e.g., to the olfactory bulb). Synaptic responses from recordings between pre- and postsynaptic neurons typically represent a summed synaptic current from a number of such release sites.

The structure of the axon may profoundly affect presynaptic inhibitory mechanisms. Although many synapses exist as varicosities along a single axon

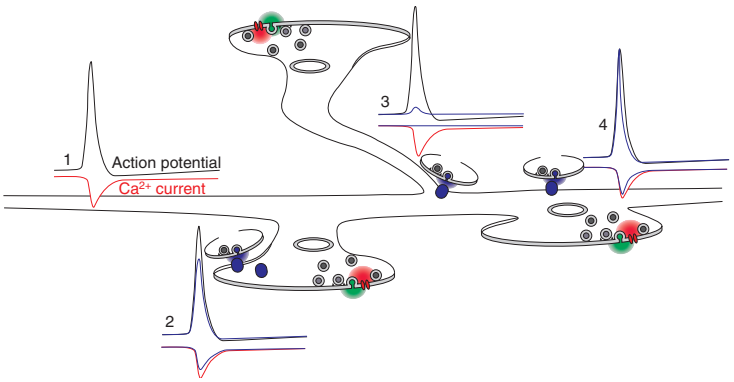

Figure 1 Many different mechanisms exist for ionotropic receptors to mediate presynaptic inhibition. An action potential (1) reaching axonal synaptic terminals may be modified by, for example, GABA receptors activated by axo-axonic inputs (blue). $GABA_A$ receptors in isolated presynaptic boutons (2) may reduce the action potential amplitude. Receptors at branch points (3) may cause a failure of the action potential to invade large or small areas of the axon terminal arborization. Receptors at *en-passant* boutons (4) may reduce the action potential duration. All these effects ultimately modify Ca^{2+} entry (red) to the presynaptic terminal and consequently reduce neurotransmitter release. GABA, γ-aminobutyric acid. From Dunlap K and Fischbach GD (1978) Neurotransmitters decrease the calcium component of sensory neuron action potentials. *Nature* 276: 837–839.

trunk, others are found at the terminations of elaborate axon branches (Figure 1). The complexity of these structures may impede the action potential invasion of distal synaptic active zones. Depending on the location within the structure of the axon, presynaptic inhibition might therefore selectively reduce or completely prevent action potential depolarization of active zones within distinct axon branches.

Voltage Dependence of Release

Because the relationship between presynaptic Ca^{2+} entry and neurotransmitter release is supralinear, changes in the duration or magnitude of action potential depolarization ultimately alter the neurotransmitter-release probability by altering Ca^{2+} entry. Changes in presynaptic depolarization may result from impeding action potential invasion of the axon, as already mentioned. Alternatively, the distribution and molecular properties of voltage-gated Na^+ channels in the presynaptic terminal also determine the duration or magnitude of action potential-induced depolarization and subsequently voltage-dependant Ca^{2+} entry. Similarly, the molecular properties of voltage-operated Ca^{2+} channels and their location relative to vesicular release apparatus determine the Ca^{2+} cooperativity of release and, therefore, probability.

Kiss-and-Run versus Full Vesicle Fusion

Many of the underlying assumptions concerning presynaptic inhibition have been predicated on the quantal nature of synaptic transmission itself. That is, if synaptic transmission requires the fusion of a synaptic vesicle and a synaptic vesicle necessarily releases all of its neurotransmitter when it fuses with the presynaptic membrane, then presynaptic inhibition must involve a reduction in release probability. However, synaptic neurotransmitter release may be more subtle and complex because complete fusion may not always occur. Three hypotheses have been proposed to account for the recycling of vesicle membrane after exocytosis. Heuser and colleagues proposed that vesicles are recycled through an endosome. Concern that this process is too slow to account for recycling observed experimentally led De Camilli and colleagues to propose a model by which the vesicle completely fuses with the presynaptic membrane and is recovered directly into function vesicles. However, it is apparent that large dense core vesicles can fuse with the membrane and then return to a pool of vesicles available for release without fully collapsing into the presynaptic membrane, and this may lead to variability in transmitter release (kiss-and-run fusion). Measurements of catecholamine release from adrenal chromaffin large dense core vesicles provide direct evidence that variable amounts of hormone may be released during vesicle fusion. More recently, it was proposed that changes in interactions between synaptotagmin (a putative Ca^{2+} sensor) and SNAP-25 may govern a mode of fusion shift. It has been suggested that synaptotagmin-I may govern complete fusion and synaptotagmin-IV incomplete fusion (i.e., kiss-and-run). The role of kiss-and-run fusion in the central nervous system (CNS) has proved contentious, but it is becoming more accepted.

Presynaptic Receptors Inhibiting Neurotransmitter Release

Synaptic transmission is presynaptically inhibited by both ionotropic (receptors with an integral ion

channel) and metabotropic receptors (G-protein-coupled receptors (GPCRs)) within the nervous system. These receptors have been identified variously as autoreceptors (activated by the release of transmitter from the terminal on which they are located) or axo-axonic (in which the transmitter is released from a second presynaptic terminal).

The mechanism by which presynaptic receptors modify neurotransmitter release remains open to some question, although some insights have been derived from the study of the activation of the γ-aminobutyric acid (GABA) receptor $GABA_A$ as a component of primary afferent depolarization. The concept that a distal or perhaps precentral mechanism might inhibit sensory inputs to the spinal cord was first proposed as early as 1957 by Frank and Fuortes. Dubbed remote inhibition, reflecting what was then known of the phenomenon, the theory was based on the finding that sensory excitatory potentials could be depressed with no visible effect on the membrane potential of the recorded spinal motor neuron receiving that input. This early work in the mammalian nervous system was paralleled by a series of markedly similar findings within arthropod preparations first outlined by Eckert. Thus, in both vertebrate and invertebrate preparations, it was found that presynaptic inhibition modifies sensory inflow to the CNS. Later recordings from primary afferent nerve terminals gave the name primary afferent depolarization (PAD) to this phenomenon, in which the sensory nerve terminal is depolarized causing presynaptic inhibition.

It is clear that a large component of PAD is caused by the activation of presynaptic $GABA_A$ receptors that evoke a Cl^-current. This current is depolarizing because Cl^-pumps at the primary afferent terminal actively raise the intracellular Cl^- concentration above its equilibrium. Since these seminal studies, a number of additional mechanisms have been proposed as underlying PAD. Recently, ionotropic glutamate receptors were identified at the presynaptic terminals in a number of systems, including the lamprey. Ionotropic glutamate receptors, including kainate receptors, *N*-methyl-D-aspartate (NMDA) receptors, and α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors have been identified at presynaptic terminals throughout the vertebrate neuroaxis. It has been proposed that these receptors may contribute to PAD at spinal afferent terminals, perhaps by spillover of glutamate from neighboring synapses. Metabotropic receptors (e.g., serotonin (5-HT) receptors) may also contribute to PAD and certainly contribute to the central inhibition of afferents to the spinal cord.

How Do Presynaptic Ionotropic Receptors Modify Release?

Ionotropic presynaptic receptors How does the activation of an ionotropic receptor on the presynaptic axon or terminal lead to a reduction in neurotransmitter release from that terminal? Perhaps the simplest answer is that regardless of whether GABA depolarizes or hyperpolarizes the terminal, a conductance shunt occurs in the terminal to decrease the amplitude of the action potential. Put more simply, the resting conductance of the axon is raised, and therefore during an action potential the current flowing through Na^+ channels leads to less potential change and a smaller action potential (Figure 1). This has been recorded directly in the lamprey spinal cord, although not at primary afferents. The smaller action potential may reduce the amount of Ca^{2+} that enters the terminal through Ca^{2+} channels. However, the precise effects may be more complex. For example, depolarizing the terminal may serve to inactivate Ca^{2+} channels partially, but also may reduce the driving force for Ca^{2+} entry at the start of the action potential. Finally, it is possible that $GABA_A$ receptor activation can prevent the action potential invasion of entire terminal regions of primary afferents (Figure 1). There is good evidence for this in a number of invertebrate preparations, although the evidence in vertebrates is less compelling.

Presynaptic mechanisms may also have effects more distal to the presynaptic terminals themselves to alter sensory inflow to the CNS. Ionotropic receptor activation can inhibit sensory receptor potentials in vertebrates and invertebrates alike, depending on the location of the receptor with respect to the CNS. In the crayfish, mechanosensitivity is modified by peptides and 5-HT; in the vertebrate lamprey spinal cord it is modified by GABA and peptidergic transmission. In the mammalian cochlear, outer hair-cell receptors can be activated by efferent output to drive sufficient motility to cause detectable noise output from the ear.

Metabotropic presynaptic receptors It is perhaps not overstating the case that every synaptic terminal possesses a complement of GPCRs. GPCRs, the largest known family of cell-surface receptors are integral membrane proteins comprising seven membrane-spanning α-helices, connected by extracellular and intracellular loops. Following activation by a GPCR, heterotrimeric G-proteins dissociate into activated G_α-guanosine triphosphate (GTP) and free $G_{\beta\gamma}$ subunits. Given the overwhelming number of pathways that G_α-GTP and free $G_{\beta\gamma}$ can activate, it is far from surprising that a similarly overwhelming number of

mechanisms have been proposed to explain presynaptic inhibition by G-proteins (see Figure 2(a)). Uncertainty over mechanisms by which G-proteins alter transmitter release reflects both the variety of effector targets from both $G_{\beta\gamma}$ and G_{α}-GTP and also the difficulty in gaining experimental access to presynaptic terminals, which, in most cases, are small and particularly difficult to record from in real time.

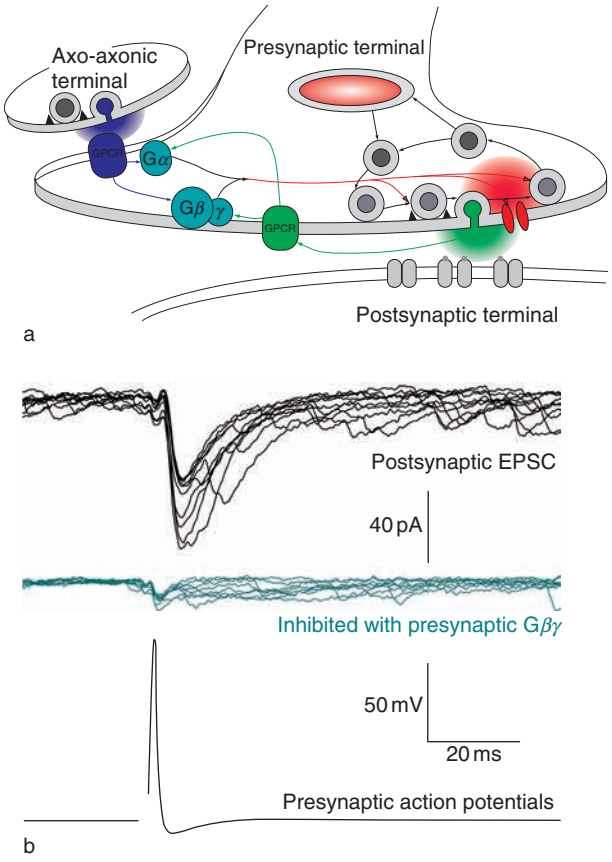

Figure 2 GPCR-mediated presynaptic inhibition: (a) mechanisms; (b) effectiveness. In (a), mirroring the complexity of GPCR signaling, many possible means of altering synaptic transmission have been proposed in the presynaptic terminal. Presynaptic inhibition may be mediated by axo-axonic inputs to the presynaptic terminal (blue) or by autoreceptors activated by release of transmitter from the presynaptic terminal itself (green). After the receptor is activated, either $G_{\beta\gamma}$ or G_{α} signaling may lead to the modification of neurotransmitter release, although $G_{\beta\gamma}$ has been more often implicated in presynaptic inhibition. Downstream in G-protein signaling pathways, the targets may be presynaptic Ca^{2+} channels, the release machinery (e.g., the SNARE complex), vesicle recycling or neurotransmitter loading of vesicles. In (b), an example of the effectiveness of presynaptic inhibition is shown. In this case, action potentials in one presynaptic axon were stimulated through a recording electrode (bottom trace, 10 in sequence). These action potentials evoked postsynaptic currents (top traces) that vary in amplitude depending on the number of vesicles released following each stimulus. Injection of $G_{\beta\gamma}$ into the presynaptic terminal, to mimic the receptor activation of a G-protein, almost abolishes this synaptic transmission (middle trace). EPSC, excitatory postsynaptic current; GPCR, G-protein-coupled receptor; SNARE, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor.

The seminal finding of Dunlap and Fischbach set the stage for a long-standing theory on the means by which all these receptors might function in the presynaptic terminal, whether activated by peptides, monoamines, or amino acids. In the cell body of sensory nerve cells in the dorsal root ganglion, a number of neurotransmitters reduce Ca^{2+} entry through voltage-gated Ca^{2+} channels. Although this finding was repeated and expanded on in many nerve cells, it was not until Forsythe and colleagues developed the calyx synapse of Held as a model system in the 1990s that definitive proof was provided that GPCRs directly alter presynaptic Ca^{2+} entry. GPCRs reduce Ca^{2+} entry by $G_{\beta\gamma}$ binding to the Ca^{2+} channel responsible for release – principally, P/Q, N, or R type channels (now named Cav2.1–Cav2.3) in which $G_{\beta\gamma}$ may bind to one or more sites and change gating properties. In a number of examples, this modification of channel properties has also been linked to a channel relationship with the SNARE protein syntaxin 1A that in turn is necessary for synaptic vesicle fusion. The efficacy of this channel inhibition may depend to a great extent on the very recent history of synaptic activity at the presynaptic terminal.

In addition to direct effects of the $G_{\beta\gamma}$ on activation of voltage-gated Ca^{2+} channels responsible for neurotransmitter release, Ca^{2+} entry may be altered by activation of K^{+} channels (G-protein-activated inwardly rectifying potassium channels (GIRKs)). G-proteins may also modulate presynaptic Na^{+} channels. It should be stressed, however, that at present little evidence supports or refutes these mechanisms at the presynaptic terminal.

$G_{\beta\gamma}$ may also inhibit synaptic transmission without altering Ca^{2+} entry. Spontaneous release, which occurs independently of Ca^{2+} entry, can be detected by recording miniature postsynaptic currents (mPSCs). These mPSCs are modulated by many GPCRs. G-proteins can inhibit exocytosis after cell permeabilization, suggesting a role very late in exocytosis that does not rely on Ca^{2+} entry. In addition, exocytosis in pancreatic β cells, mast cells, chromaffin cells, PC12 cells, and secretory granules are regulated independently of Ca^{2+} entry by G-proteins, and $G_{\beta\gamma}$ interacts directly with the fusion machinery in rat mast cells. In the lamprey giant synapse, 5-HT-mediated synaptic inhibition does not reduce Ca^{2+} entry to the synapse, and its effect is blocked by presynaptic intracellular scavenging of free $G_{\beta\gamma}$. $G_{\beta\gamma}$ can interact with the core vesicle fusion machinery; it directly binds SNARE proteins syntaxin and SNAP-25 and the ternary SNARE complex.

Intriguingly, $G_{\beta\gamma}$ interacts with the C-terminal region of the SNARE complex, a region that may play a role in linking synaptotagmin, a putative Ca^{2+} sensor, to SNARE at the C-terminal region of SNAP-25. Interaction of SNAP-25 and syntaxin with synaptotagmin links Ca^{2+} entry with vesicle fusion, but may also control the stability of the fusion pore leading to a kiss-and-run form of fusion.

This finding essentially parallels findings of the effect of $G_{\beta\gamma}$ in the lamprey and in chromaffin cells in culture, suggesting a vital role for synaptotagmin–SNARE complex interactions in fusion pore stability and neurotransmitter release. Thus, it is possible that presynaptic inhibition involves not just a change in the probability of neurotransmitter release but a change in the vesicle fusion mechanism that leads to neurotransmitter release.

Finally, it may be possible to inhibit synaptic transmission by lowering the amount of glutamate in synaptic vesicles prior to fusion. Indeed, quantitative immunogold electron microscopy of glutamatergic Schaffer collateral and parallel fiber terminals reveals that some synaptic vesicles contain G_{β} subunits. Furthermore, G-protein activation inhibits glutamate uptake into mouse and rat brain vesicular preparations. In addition to changing the concentration of neurotransmitter, it also appears possible to alter the volume of the vesicle and, therefore, alter the total amount of neurotransmitter within the vesicle.

Patterns of Neuronal Activity That Are Likely to Activate These Receptors, Autoreceptors, Heteroreceptors, and Spillover

Presynaptic inhibition can have very different functional implications for the synapse depending on the source of the inhibition. Broadly, we may break down presynaptic inhibition into two categories: (1) inhibition mediated by axo-axonic or dendro-axonic projections (i.e., a third party interferes with the transmission of synaptic responses between a synaptically coupled pair of two neurons), which we may describe as heterosynaptic inhibition; and (2) inhibition mediated by the release of neurotransmitter from the active presynaptic terminal itself, or autoreceptive inhibition.

Heterosynaptic presynaptic inhibition The best example of this phenomenon is undoubtedly PAD. Animals are subject to a constant barrage of sensory inflow, only a small fraction of which we pay attention to at each moment. Thus, somewhere in the passage of information from the periphery to the higher centers of cognition, animals must curtail the flow of nonessential information. It is believed that presynaptic inhibition provides this role. For example, when an animal walks, its response to a touch on the foot varies depending on whether that foot is raised or on the ground. Phasic primary afferent depolarization, driven by central pattern generators in the spinal cord, is thought to represent a component of such sensory gating. Thus, heterosynaptic presynaptic inhibition can allow one central nervous function (motor drive in this case) to modify a related but separate function (sensory inflow). A corollary to heteroreceptor-mediated presynaptic inhibition has also been identified in a variety of synapses. Spillover, or cross-talk between synapses, can occur in which neurotransmitter released from one terminal may reach neighboring terminals, albeit at a reduced concentration. The resultant lower neurotransmitter concentration may allow the selective activation of higher-affinity receptors at the affected terminal, both presynaptically and postsynaptically.

Autoreceptor-mediated presynaptic inhibition Inhibition mediated by receptors activated by release of neurotransmitter from the same presynaptic terminal necessarily implies a very different function. This form of presynaptic inhibition induces short-term use dependent effects at the presynaptic terminal in question. The effect of the activation of the presynaptic receptors are frequency- and pattern-dependent. Because the release of transmitter from earlier stimuli activates presynaptic receptors that modify future release, the time delay between stimuli can significantly alter the subsequent release of neurotransmitter. However, we must also consider the properties of the receptors activated and their downstream effect.

Ionotropic autoreceptors. Examples of these at the presynaptic terminal are GABA receptors, NMDA receptors, kainate receptors, and AMPA receptors. However, there is not a great deal of consensus on the mechanisms of action of presynaptic glutamate receptors. Both kainate and AMPA receptors appear to show convergence of signaling with metabotropic receptor effects in the presynaptic terminal to inhibit release, although it is not yet clear how this occurs.

Metabotropic autoreceptors. By far the greater attention has been paid to GPCRs acting as autoreceptors. We have already discussed the variety of possible mechanisms by which GPCRs may modify release. However, the mechanism also has important implications for the function of these receptors during the repetitive stimulation that is necessary to activate the autoreceptor under conditions at which it might actually inhibit neurotransmitter release. For example, GPCRs inhibit Ca^{2+} entry to the presynaptic terminal, but this inhibition is attenuated by depolarization associated with action potential firing because the GPCR-mediated inhibition of Ca^{2+} dependence is transiently relieved by membrane depolarization. Thus, paradoxically, depending on the initial state

of GPCR activation, paired pulse facilitation or depression may follow autoreceptor activation and the inhibition of presynaptic Ca^{2+} channels. In fact, at central glutamatergic synapses known to contain inhibitory metabotropic glutamate receptors, pharmacological block of the receptor may have little effect on neurotransmitter release during trains of stimuli, but may profoundly alter recovery of neurotransmitter release shortly after the train ends.

More recent work on the interaction between $G_{\beta\gamma}$ and the SNARE complex also opens the possibility for complex interactions between inhibition of release by direct effects on the fusion machinery and effects on Ca^{2+} entry to the terminal. We might expect synergistic effects of $G_{\beta\gamma}$ because it can both reduce the efficacy of Ca^{2+} to cause synaptic vesicle fusion and reduce the Ca^{2+} entry that is necessary to drive vesicle fusion.

Conclusion

The extent of presynaptic inhibition following receptor activation is determined by (1) the morphology of the axon, (2) the molecular properties of the proteins involved in vesicle fusion, and (3) recent activity of the axon. It therefore appears that unique forms of presynaptic inhibition exist that may be activated by distinct patterns of neuronal activity. For example, presynaptic $GABA_A$ receptor depolarization decreases Ca^{2+} influx and therefore inhibits synaptic transmission by lowering release probability (decreasing the number of vesicle that fuse in response to an action potential). Nevertheless, the amount of neurotransmitter released in to the synaptic cleft from the remaining vesicles that fuse remain the same. An alternative mechanism has, however, been proposed in which the activation of presynaptic GPCRs inhibits synaptic transmission downstream of Ca^{2+} entry through a direct interaction with vesicle fusion proteins. Recent work demonstrates that $G_{\beta\gamma}$ alters the mode of vesicle fusion, resulting in a lesser release of neurotransmitter, perhaps from each vesicle. This form of synaptic modification may, in turn, lead to a differentially greater inhibition of lower-affinity receptors compared to receptors with a higher affinity at the same synapse.

Finally, we must ask why presynaptic inhibition exists? Ultimately, the presynaptic terminal provides one more source of computational ability in the nervous system. However, we may break these mechanisms down into some clearly important uses:

1. *Sensory gating*. Presynaptic inhibition represents an early step in the modification of sensory inflow, depending on the state of motor output or related sensory inputs.
2. *Receptor selectivity*. If presynaptic inhibition can modify the fusion properties of a synaptic vesicle sufficiently to reduce synaptic concentrations of neurotransmitter, then this mechanism can readily alter the population and activity of activated postsynaptic receptors.
3. *Selective activation of synapses*. If one axon can selectively activate different synaptic terminals, depending on the state of presynaptic inhibitory inputs, then the computational power of the circuit in which the axon is embedded will be increased.

See also: Calcium Channels and SNARE Proteins; Long-Term Depression: Cerebellum; Neuromuscular Junction (NMJ): Presynaptic Non-Quantal Release of Transmitter; Presynaptic Facilitation; SNAREs.

Further Reading

Augustine GJ, Adler EM, and Charlton MP (1991) The calcium signal for transmitter secretion from presynaptic nerve terminals. *Annals of the New York Academy of Sciences* 635: 365–381.

Brunger AT (2005) Structure and function of SNARE and SNARE-interacting proteins. *Quarterly Review of Biophysics* 38: 1–47.

Brunk I, Holtje M, von Jagow B, et al. (2006) Regulation of vesicular monoamine and glutamate transporters by vesicle-associated trimeric G proteins: New jobs for long-known signal transduction molecules. In: Sitte HH and Freissmuth M (eds.) *Handbook of Experimental Pharmacology, Vol. 175: Neurotransmitter Transporters*, pp. 30–325. New York: Springer.

Catterall WA (2000) Structure and regulation of voltage-gated Ca^{2+} channels. *Annual Review of Cell and Developmental Biology* 16: 521–555.

Del Castillo J and Katz B (1954) Quantal components of the end-plate potential. *Journal of Physiology* (London) 124: 560–573.

Dunlap K and Fischbach GD (1978) Neurotransmitters decrease the calcium component of sensory neuron action potentials. *Nature* 276: 837–839.

Frank K, Fuortes MG, and Nelson PG (1959) Voltage clamp of motoneuron soma. *Science* 130: 38–39.

Gerachshenko T, Blackmer T, Yoon EJ, Bartleson C, Hamm HE, and Alford S (2005) Gbetagamma acts at the C terminus of SNAP-25 to mediate presynaptic inhibition. *Nature Neuroscience* 8: 597–605.

Hamm HE and Gilchrist A (1996) Heterotrimeric G proteins. *Current Opinion in Cell Biology* 8: 189–196.

Han X, Wang CT, Bai J, Chapman ER, and Jackson MB (2004) Transmembrane segments of syntaxin line the fusion pore of Ca^{2+}-triggered exocytosis. *Science* 304: 289–292.

Harata NC, Aravanis AM, and Tsien RW (2006) Kiss-and-run and full-collapse fusion as modes of exo-endocytosis in neurosecretion. *Journal of Neurochemistry* 97: 1546–1570.

Heuser JE, Reese TS, Dennis MJ, Jan Y, Jan L, and Evans L (1979) Synaptic vesicle exocytosis captured by quick freezing and correlated with quantal transmitter release. *Journal of Cell Biology* 81: 275–300.

Rudomin P and Schmidt RF (1999) Presynaptic inhibition in the vertebrate spinal cord revisited. *Experimental Brain Research* 129: 1–37.

Silinsky EM (1984) On the mechanism by which adenosine receptor activation inhibits the release of acetylcholine from motor nerve endings. *Journal of Physiology* 346: 243–256.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Presynaptic Receptor Signaling

M Casado, Ecole Normale Supérieure, CNRS, Paris, France
P Ascher, Université Paris Descartes, CNRS, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Fifty Years of Presynaptic Receptors

The notion of presynaptic receptors dates from the discovery of presynaptic inhibition, which itself followed the understanding of postsynaptic inhibition and postsynaptic excitation. In the 1950s, it was established that postsynaptic potentials (PSPS) are produced by a change of ionic conductances. The demonstration was based on the measure of the reversal potential of the PSPS. It was shown that inhibitory PSPSs differed from excitatory synaptic potentials by the fact that their reversal potential was more negative than the threshold of the action potential, whereas for EPSPs the reversal potential was more depolarized than the threshold. The functional result of coincident excitatory and inhibitory inputs was seen as determined by the algebraic sum of the changes they produce in membrane potential.

These concepts had barely been established before exceptions appeared. In some preparations, the EPSP produced by an excitatory axon A was reduced by stimulation of an inhibitory axon B, which by itself did not produce a change of conductance in the postsynaptic cell. The explanation, first proposed by Frank and Fuortes in 1957 and then developed by the groups of Eccles and Kuffler, was that the inhibitory axon terminal synapsed with the excitatory terminal and produced a change in conductance in this terminal. Soon, both Gray and Kidd provided electron microscopy pictures which established the existence of axo-axonic synapses ('serial' synapses) and gave a structural basis to the prediction of the physiologists (Figure 1). Presynaptic inhibition was then observed in a variety of other systems. The concept of presynaptic facilitation soon followed (it was first proposed by Tauc and Kandel in 1964).

Presynaptic inhibition and presynaptic facilitation were initially viewed as heterosynaptic and based on serial synapses. In 1971, four laboratories proposed that in the peripheral nervous system, norepinephrine can inhibit its own release by acting on 'autoreceptors.' It was later found that in other cases presynaptic autoreceptors facilitate the release of transmitter. Autoreceptors and heteroreceptors can nowadays be visualized by combining immunochemistry and electron microscopy, as illustrated in Figure 2. However, this simple classification became insufficient when retrograde signals were discovered and when it became clear that presynaptic receptors can also be activated by glial signals and by long-range messages.

In the 1960s, the mechanisms proposed for the action of presynaptic receptors all assumed that the receptor activation opens ionic channels. Later, it was realized that presynaptic receptors could modulate channels that they did not open, particularly Ca channels, which, although they carry only small currents, play a major role in transmitter release. The key experiment, performed by Dunlap and Fischbach in 1978, showed that in sensory ganglion cells, a variety of transmitters reduced the Ca current activated by a depolarization and shortened the plateau of Ca action potentials. This effect could not be explained by the opening of ionic channels by the transmitter. It required receptors modulating the Ca channels activated by depolarization.

In 1979, Eccles and McGeer formalized the need for two types of receptors by introducing the distinction between ionotropic receptors and metabotropic receptors. Ionotropic receptors corresponded to the classical view of a receptor opening an ionic channel for short durations (in the millisecond range). Metabotropic receptors were defined slightly vaguely as receptors with a slow mode of action. At that time, G-proteins had not been identified, and it was only in approximately 1985 that it was realized that G-protein-coupled receptors (GPCRs) can be coupled to ion channels. It took a few more years to accept that in many cases this coupling is 'membrane delimited' and does not require second messengers. In this article, we use the distinction proposed by Eccles and McGeer to analyze how ionotropic receptors and metabotropic receptors control transmitter release. We then discuss how presynaptic receptors are activated.

Technical Problems Associated with the Study of Presynaptic Receptors

The postsynaptic somatodendritic compartment is usually accessible to electrical measurement and control through electrodes placed in the soma. In contrast, presynaptic terminals are often too small to be recorded from and too far from the soma to be accurately recorded by electrodes placed in the soma. These morphological considerations explain why understanding the mechanisms of presynaptic control has benefited disproportionately from experimental preparations with very large presynaptic elements, such as invertebrate synapses, the synapses made by mossy fibers in the hippocampus and the cerebellum, and the synapse between the calyx of Held and the soma of neurons

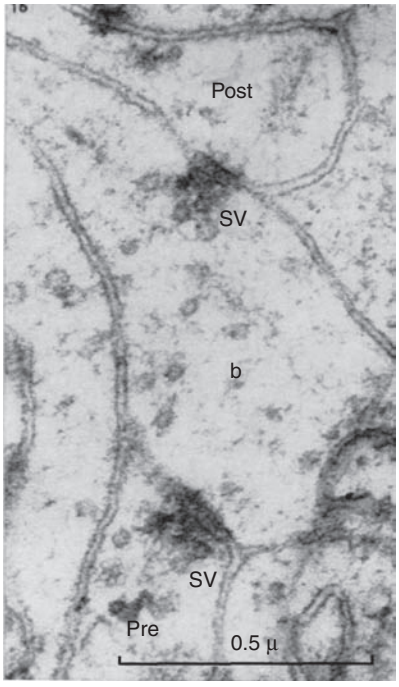

Figure 1 The first observation of a serial synapse made by M Kidd in 1962 in the retina of the pigeon. The postsynaptic process labeled 'post' is contacted by a presynaptic process labeled 'b.' The presynaptic side is identified by the presence of synaptic vesicles (sv; modern techniques would make these appear much more clearly, but they are nevertheless unmistakable). The vesicles identify a typical chemical synapse, a well-established notion since 1955. The novelty here was the second chemical synapse, in which the presynaptic process with synaptic vesicles, labeled 'pre,' contacts the structure b. b is therefore postsynaptic for 'pre' and presynaptic for 'post.' Reprinted from Kidd M (1962) Electron microscopy of the inner plexiform layer of the retina in the cat and the pigeon. *Journal of Anatomy (London)* 96: 179–187, plate 3, figure 16.

from the medial nucleus of the trapezoid body (MNTB) in vertebrate brain stem. In this latter synapse, the postsynaptic cell body is enveloped by a 'giant' presynaptic element (the calyx). In these giant synapses, it is possible to record simultaneously the presynaptic events (Ca currents and Ca concentration changes) and the postsynaptic response which makes it possible to evaluate transmitter release (Figure 3). However, even in these systems it is difficult to measure the concentration and the distribution of the presynaptic modulators. This has been achieved in only a few cases, such as by the study of the modulation by dopamine of glutamate release in the striatum, in which it has been possible to follow the dopamine concentration by amperometry.

Presynaptic Ionotropic Receptors and Transmitter Release

Presynaptic receptors were first described in the crayfish neuromuscular junction and in the spinal cord. In both cases, the inhibition involved ionotropic receptors opening Cl channels. It has since been found that the opening of Cl channels can also lead to facilitation, and that some presynaptic ionotropic receptors open cationic channels.

Ionotropic Receptors Opening Presynaptic Cl Channels Can Produce Inhibition or Excitation

Figure 4(a) illustrates the mechanism which was found to operate at the crayfish neuromuscular junction. The presynaptic terminal of the axon bears $GABA_A$ receptors which, when activated by GABA, open Cl channels. Intracellular recording from the presynaptic excitatory axon showed that its resting potential is approximately −80 mV, whereas the reversal potential of the GABA current is −85 mV. Hence, the opening of Cl channels induces a net entry of Cl and a hyperpolarization. This hyperpolarization is likely to impede the propagation of the action potential in the terminal part of the axon.

Figure 4(b) illustrates the case of presynaptic inhibition of the primary afferents of the spinal cord. Here again, the inhibitory transmitter, GABA, opens Cl channels, but in these axon terminals the Cl concentration is high and the GABA reversal potential is less negative than the resting potential. As a result, the terminal is depolarized by GABA. Because the phenomenon affects a large number of presynaptic axons, the depolarization can be recorded extracellularly as a relatively large signal known as the primary afferent depolarization. A similar situation is found in the posterior pituitary, where it was possible to record intracellularly from presynaptic terminals. It was found that the GABA reversal potential was approximately −48 mV, whereas the resting potential was approximately −60 mV. Shunting by GABA had relatively little effect on transmitter release, with the main effect being caused by the depolarization-induced inactivation of Na channels. A depolarization from −60 to −45 mV in voltage clamp inactivated 90% of the Na channels. As a result, the propagation of action potential in the terminal was reduced, as was transmitter release.

Figure 4(c) illustrates a case in which opening of Cl channels can lead to facilitation. The calyx of Held (Figure 3) bears ionotropic glycine receptors. Activation of these receptors opens Cl channels. In this terminal, the Cl concentration is high, as in the primary afferents of the spinal cord. The glycine reversal potential was found to be at −50 mV, whereas the resting potential was approximately − 60mV. Thus, the effect of glycine is depolarizing, as in Figure 4(b). However, the functional effect of glycine is a facilitation likely due to the opening of Ca channels by the depolarization and to the subsequent increase in the basal level of Ca

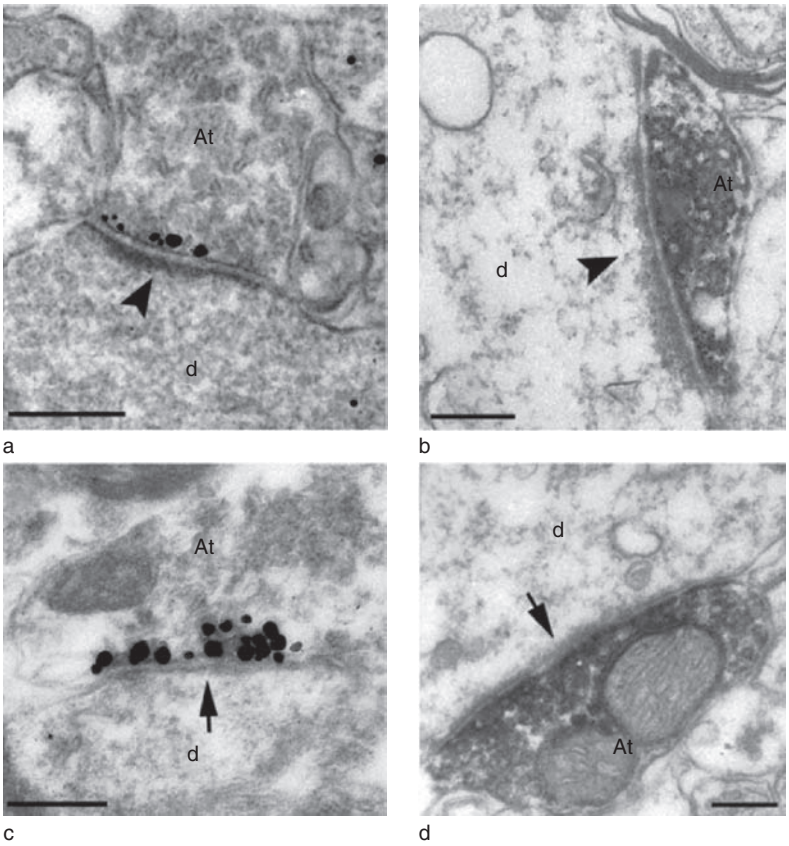

Figure 2 Direct visualization of presynaptic receptors. Electron micrographs showing immunoreactivity for mGluR8a and mGluR8b in the stratum oriens of the CA1 area of hippocampus in rat. (a, c) Immunogold/silver labeling for mGluR8a is concentrated in the presynaptic active zone facing the postsynaptic membrane specialization of type I (arrowhead in (a)) or type II (arrow in (c)) synapses on dendritic shafts. (b, d) Immunoperoxidase labeling for mGluR8b spreads from the presynaptic membrane specialization in boutons forming a type I (arrowhead in (b)) or a type II (arrow in (d)) synapse on dendritic shafts. Type I synapses are glutamatergic, whereas type II synapses are GABAergic. Thus, the glutamate presynaptic receptors are autoreceptors in the case of (a) and (b) and heteroreceptors in the case of (c) and (d). Scale bars = 0.2 μm. At, axon terminal; d, dendritic shaft. Reprinted from Ferraguti F, Klausberger T, Cobden P, et al. (2005) Metabotropic glutamate receptor 8-expressing nerve terminals target subsets of GABAergic neurons in the hippocampus. *Journal of Neuroscience* 25: 10520–10536, with permission.

(which was documented using Ca indicators). The relation between Ca_i and release is known to be very steep, and a small increase in basal Ca can give rise to a significant increase in transmitter release.

The two effects described in Figures 4(b) and 4(c) may coexist at the same synapse, as has been suggested for the effects of presynaptic $GABA_A$ receptors on the mossy fibers in hippocampus. Globally, the balance between facilitation and inhibition is difficult to predict because it depends on multiple factors, such as the presynaptic resting potential, the Ca dependence of transmitter release, the voltage dependence of Na and Ca channel activation and inactivation, and the activity of Cl transporters.

Ionotropic Receptors Opening Cationic Channels Can Facilitate or Inhibit Transmitter Release

Cationic channels are defined here as channels which allow the flow of both Na and K ions and which have similar permeabilities for these two ions. In such channels, the current through the channels reverses at approximately 0 mV and the opening of the channels is therefore depolarizing. Some cationic channels are permeable to Ca and/or Mg, and in these channels the flow of divalent cations, although quantitatively much smaller than that of monovalent cations, also contributes to the depolarization.

Figures 4(b) and 4(c) show that depolarization of the presynaptic terminal by opening of Cl channels can lead to either inhibition or facilitation. Similarly, when the presynaptic transmitter opens cationic channels, depolarization can inactivate Na channels and thus reduce the presynaptic spike (Figure 4(d)) and/or it can open voltage-dependent Ca channels and thus increase Ca_i, which will lead to facilitation of release (Figure 4(e)). Inhibition by inactivation of Na channels has been proposed in the case of central terminals (primary afferents) of dorsal root ganglion

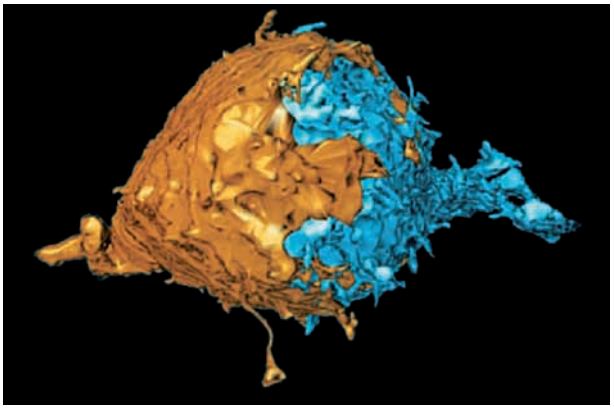

Figure 3 The calyx of Held. Top view of the three-dimensional reconstruction of a calyx of Held and of the cell body of the MTNB principal neuron in a P9 rat. Three-dimensional reconstruction was made by alignment of serial electron microscopy images. The calyx (left) is shown in yellow and the principal neuron (right) in blue. The nerve terminal forms a cuplike structure covering approximately 40% of the surface area of the principal neuron. The final part of the preterminal axon that gives rise to the giant terminal is seen on the left. The cell body measures 18 μm in the vertical axis and 23 μm in the horizontal axis. Reprinted from Sätzler K, Söhl LF, Bollman JH, et al. (2002) Three-dimensional reconstruction of a calyx of Held and its postsynaptic principal neuron in the medial nucleus of the trapezoid body. *Journal of Neuroscience* 22: 10567–10579, with permission.

neurons, in which activation of AMPA receptors produces the same effect as that of neighboring $GABA_A$ receptors. Facilitation of release has been observed for kainate receptors on GABA–glycine terminals in the dorsal horn and for nicotinic receptors.

It must also be mentioned that if the presynaptic axon contains K channels which show a voltage-dependent inactivation, depolarization could inactivate them and facilitate release by prolonging the action potential. This explanation has been proposed, but not demonstrated, in the case of the kainate receptors in the presynaptic terminals of mossy fibers in the CA3 region of hippocampus.

Ionotropic Receptors Opening Cationic Channels Permeable to Ca Can Facilitate Release by Allowing Direct Entry of Ca

Ionotropic cationic receptors show variable degrees of Ca permeability, from very low (as for some glutamate AMPA-Rs) to very high (for some nicotinic receptors and glutamate NMDA receptors). At some presynaptic terminals bearing ionotropic receptors with a high Ca permeability, the influx of Ca through these receptors can facilitate release without requiring a substantial depolarization and the recruitment of voltage-dependent Ca channels.

An example of presynaptic involvement of such receptors is that of the nicotinic presynaptic receptors. Many glutamatergic or GABAergic terminals have been shown to increase their release of transmitter under the action of nicotine. In a fraction of cases, the effect of nicotine disappeared when analyzed in the presence of TTX, and it was therefore concluded that the mechanism involved the depolarization induced by nicotine, which led to the opening of voltage-dependent Ca channels in the axon and increased the intracellular Ca concentration. This mechanism is therefore of the same type as that described previously for other transmitters depolarizing the terminal. However, in other synapses, the facilitation of release persisted in the presence of TTX, as did the increase in Ca in the presynaptic terminal (Figure 4(f)). In such cases, it was concluded that the changes in transmitter release are dependent on Ca entry through the ionotropic nAChRs.

A similar situation has been found for NMDA receptors and some AMPA receptors. The terminals of the cerebellar inhibitory interneurons (INs) synapsing with the Purkinje neuron (PN) bear NMDA receptors. These receptors are activated by glutamate released by the depolarization of the PN, and this inhibits the release of their transmitter (GABA). Similar effects have been observed in the spinal cord, cortex, and amygdala. A presynaptic facilitation involving Ca-permeable AMPA receptors has been described in a retinal neuron which does not elicit action potentials.

Presynaptic Metabotropic Receptors and Transmitter Release

Although not all metabotropic receptors are GPCRs, all known examples of presynaptic metabotropic receptors are GPCRs. GPCRs can control ion channels either directly, by a 'membrane delimited' process, or indirectly through second messengers. Presynaptic GPCRs can also act on the release machinery without involving ion channels. It is not always known by which way the control is exerted. We briefly examine a few well-understood cases.

Most Presynaptic Metabotropic Receptors Inhibit Release by Inhibiting the Activation of Ca Channels

The α-adrenergic inhibition of catecholamine release, which was the observation that led to the notion of autoreceptors in 1971, was reanalyzed in 1989 in frog sympathetic neurons by combining single-channel electrophysiology and measurement of radiolabeled transmitter release. Norepinephrine was found to act specifically on one class of Ca channels, the N-type Ca channels, without an effect on L-type channels. It had no effect on transmitter release when the activity of Ca channels was bypassed by a Ca ionophore.

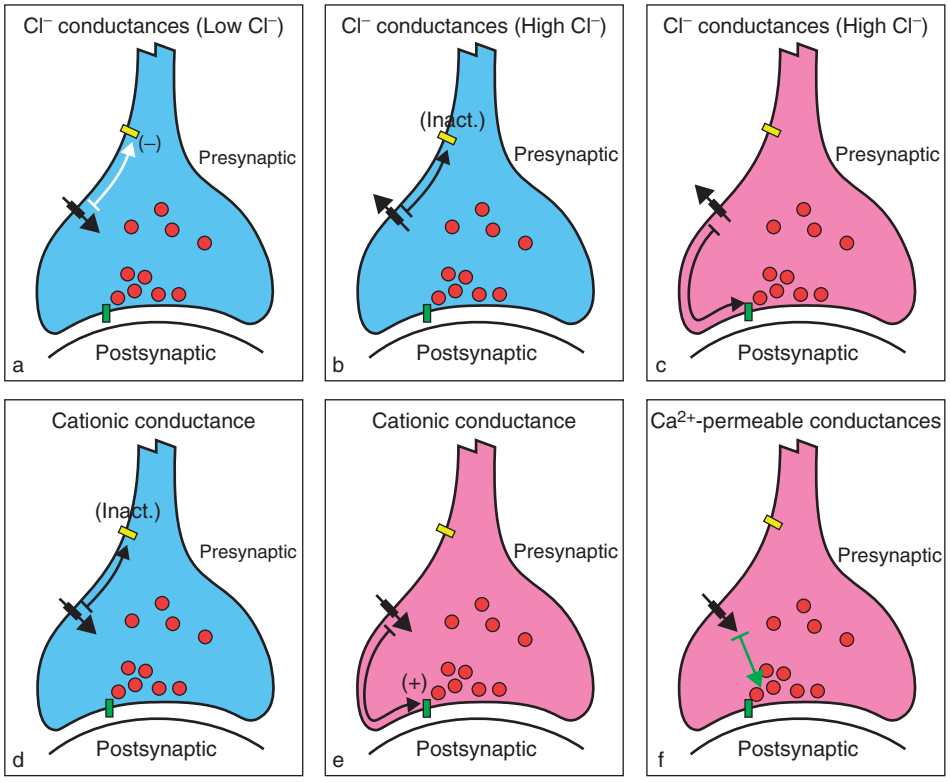

Figure 4 Presynaptic ionotropic receptors. (a) Chloride conductances hyperpolarize the presynaptic element when its internal chloride concentration is low. Hyperpolarization decreases the excitability and hence inhibits the synapse. (b, c) Chloride conductances depolarize the presynaptic element when its internal chloride concentration is high. This may result in inactivation of sodium channels and inhibition of transmitter release (b) or in activation of calcium currents and facilitation of release (c). (d, e) Cationic conductances depolarize the synaptic bouton. Again, this may result in inactivation of sodium channels and inhibition of transmitter release (d) or in activation of calcium currents and facilitation of release (e). (f) Activation of calcium-permeable conductances may enhance presynaptic calcium levels and thus facilitate transmitter release. The white arrow under the membrane (a) indicates hyperpolarization. The black arrows under the membrane (b–e) indicate depolarization. The green arrow (f) indicates that calcium entry affects vesicle recruitment. A pink presynaptic element indicates facilitation of release. A blue presynaptic element indicates inhibition of release. Yellow rectangles, sodium channels; green rectangles, calcium channels; red circles, synaptic vesicles; short arrows, ion fluxes; +, increased channel activation; −, decreased channel activation; Inact., channel inactivation.

The calyx of Held bears several kinds of presynaptic metabotropic receptors, including those activated by glutamate (mGluR) and GABA ($GABA_B$), which both inhibit release by acting on Ca channels. The action of GABA has been particularly well studied. Presynaptic loading with analogs of guanosine diphosphate or with various G subunits indicated that presynaptic inhibition occurs by a direct interaction of G_o β/γ subunits with presynaptic Ca channels. The biophysical analysis of the effects of GABA on Ca currents showed that this interaction led to a slowing of the Ca channel activation kinetics at moderate depolarizations, whereas strong depolarization relieved the inhibition.

A recently identified mechanism of presynaptic modulation through Ca channel inhibition is that of endocannabinoids (CBs). The synthesis of CBs is triggered in the postsynaptic cell by a rise in Ca_i. The enzymatic machinery producing the cannabinoids seems to be localized preferentially at the bottom of the postsynaptic spines. CBs, which are amphiphilic, diffuse out of the postsynaptic cell and can reach CB receptors on presynaptic terminals. CB receptors inhibit release by inhibiting the presynaptic Ca channels.

Presynaptic Metabotropic Receptors Can Inhibit Release by Activating K Channels

In the spinal cord of the lamprey, Cochilla and Alford recorded simultaneously from the presynaptic terminals of reticulospinal neurons and from the postsynaptic neurons. They obtained evidence that presynaptic mGluRs depress transmitter release by activating a presynaptic outwardly rectifying K current without altering the presynaptic Ca current. This is a mirror image of what happens at the calyx of Held, in which mGluRs reduce the Ca current without altering the K current.

Cochilla and Alford noted that the fact that the K current was rectifying

> is ideal for the decrement of axonal action potentials and synaptic transmission. Activation of a hyperpolarizing response alone might increase the amplitude of the presynaptic action potential, whereas an outward current that does not hyperpolarize the axon because it rectifies at resting membrane potentials will be much more effective in shunting the series resistance of the axon during the depolarizing action potential.

Despite the perfection of this mechanism, it has not been observed at other synapses.

Presynaptic Metabotropic Receptors Can Increase Release by Blocking K Channels

When first proposed by Kandel and Tauc in 1964, the concept of presynaptic facilitation had problems associated with the fact that, at that time, opening presynaptic monovalent ion channels seemed to always lead to inhibition of the action potential and thus presynaptic inhibition. The discovery of Ca channels permitted to envision a facilitation of the Ca channel activation, and it has been shown that ionotropic receptors can produce such a facilitation. However, metabotropic receptors inducing a facilitation by direct modulation of Ca channels have not been described, and the historical example described by Kandel and Tauc seems to involve a broadening of the action potential caused by the closure of K channels. The mechanism of this broadening has been particularly well studied in a sensory neuron activated by peripheral mechanical stimuli in *Aplysia*. Serotonin increases release by increasing the duration of the action potential in the presynaptic terminal. This results from a decrease in the slowly activating outwardly rectifying K current. A small change in duration of the action potential (10–20%) is sufficient to increase release by 1.5- to 2-fold. Spike broadening was found to be necessary for facilitation. The effect of serotonin appears to involve the activation of both protein kinase A and protein kinase C, with each of these pathways playing a role at different timescales. No similar system has been identified at vertebrate synapses.

Metabotropic Receptors Can Influence Vesicle Recruitment to the Readily Releasable Pool

Rapid transmitter release involves an extraordinarily high number of steps, but studies, particularly by the group of Erwin Neher, suggest the existence of two limiting processes: Ca entry and the recruitment of synaptic vesicles to sites where Ca channels cluster. Much evidence indicates that presynaptic receptors control the second step as well as the first.

In the calyx of Held, Sakaba and Neher (Figure 5) analyzed the effects of $GABA_B$ receptor activation on transmitter release. They found that in addition to the inhibition of Ca currents discussed previously, there

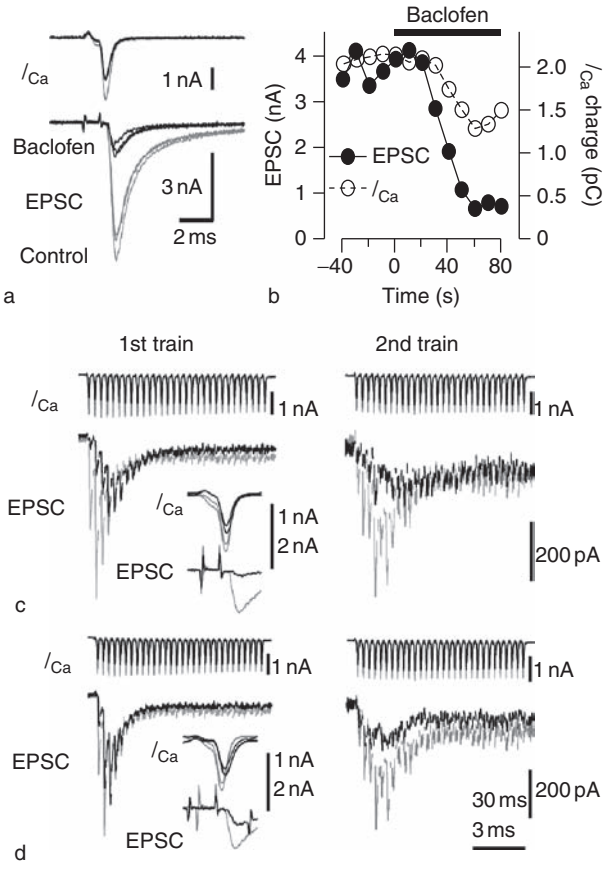

Figure 5 Presynaptic metabotropic receptors can act on more than one target. The experiments illustrate two effects produced by the activation of presynaptic $GABA_B$ receptors on transmitter release in the calyx of Held. The records were obtained in brain stem slices from young rats. The presynaptic element (the calyx of Held) and the postsynaptic element (the cell body of the MNTB neuron) were both whole cell clamped. The current labeled I_{Ca} is the presynaptic current recorded after a 1 ms depolarizing pulse to +40 mV was applied to the calyx. The voltage-dependent Na and K currents were blocked pharmacologically. The current labeled EPSC is the excitatory postsynaptic current, which measures the amount of glutamate released by the calyx. Gray traces were obtained during a control period, and black traces were obtained in the presence of baclofen (20 μM), a selective agonist of $GABA_B$ receptors. (a) Activating $GABA_B$ receptors reduces the presynaptic Ca current, and this in turn induces a reduction of EPSC. (b) The time course of the baclofen effect. Baclofen was applied at time 0. (c) Baclofen influences the recovery of the depression observed after applying a train of presynaptic pulses. During a 300 Hz train of action potential-like stimuli (30 pulses), the postsynaptic response decreases. Applying a second train after 200 ms reveals that the response has partially recovered (gray traces). However, in the presence of baclofen (black traces), the recovery is much reduced. This reduced recovery was not due the reduced Ca influx (not shown). Reprinted from Sakaba T and Neher E (2003) Direct modulation of synaptic vesicle priming by GABA(B) receptor activation at a glutamatergic synapse. *Nature* 424: 775–778, with permission.

was also a slowing of the recovery from depression after a 300 Hz train. This was due to modulation of vesicle recruitment. Addition of cAMP to the patch pipette, which was known to increase release, prevented the effect of $GABA_B$ receptors on vesicle recruitment. This suggests that the $GABA_B$ receptors slow down vesicle recruitment by decreasing cAMP concentrations.

Studies have shown that cAMP produces both an increase in the number of readily releasable quanta and an increase in the release probability, with no effect on Ca or K currents. The main target of cAMP appears to be the guanine nucleotide exchanging factor Epac, which interacts with RIM proteins at the active zone. Epac was also found to be involved in a presynaptic facilitation induced by serotonin at the crayfish neuromuscular junction. In this system, serotonin induces the production of cAMP, and cAMP appears to activate Epac.

Vesicle recruitment may not be the only intracellular presynaptic process involved in the modulation of transmitter release by GPCRs. GPCRs can modulate the two intracellular compartments storing intracellular Ca: that responsible for Ca-induced Ca release (CICR) and the IP3 sensitive stores. In many synapses, the increase in Ca following the action potential is amplified by CICR. Modulation of this amplification is likely to alter release. Many GPCRs, particularly some mGluRs, activate IP3 production, which releases Ca from intracellular stores, and the resulting increase in Ca_i can therefore modulate release. It has also been reported that in some cases the activation of protein kinase C associated with the same receptors is capable of modulating transmitter release.

Cooperation between Ionotropic and Metabotropic Receptors

Ionotropic and metabotropic receptors may operate in succession. In one interesting case studied by Kerchner and colleagues, presynaptic kainate receptors facilitate GABA and glycine release in a transient manner by an ionotropic mechanism (which requires Na entry and activation of voltage-dependent Ca channels). GABA then activates presynaptic $GABA_B$ metabotropic receptors which inhibit the release of GABA and glycine. A biphasic effect is observed if kainate is applied exogenously, but in physiological conditions the long-lasting process is inhibition due to the homosynaptic metabotropic $GABA_B$ receptors.

Ionotropic and metabotropic presynaptic receptors may operate synchronously, as has been observed in the study of the long-term depression of the excitatory synapses between pyramidal cells in the L5 layer of the cerebral cortex. This depression was found to depend on the simultaneous activation of cannabinoid receptors and glutamate (NMDA) receptors. In the model proposed by Sjöstrom and colleagues, release of glutamate by the presynaptic neuron depolarizes the postsynaptic neuron. This triggers postsynaptic Ca entry, which induces the production of endocannabinoids. The cannabinoids diffuse in the retrograde direction and activate presynaptic CB1 receptors. At the same time, the glutamate released by the presynaptic terminal activates presynaptic NMDA receptors (autoreceptors). This coincident activation triggers a long-lasting depression of glutamate release.

Where Does the Agonist Come From?

As mentioned previously, the agonists of the presynaptic receptors can come from multiple sources. Figure 6 indicates the six main sources which have been described: the presynaptic cell, the postsynaptic cell,

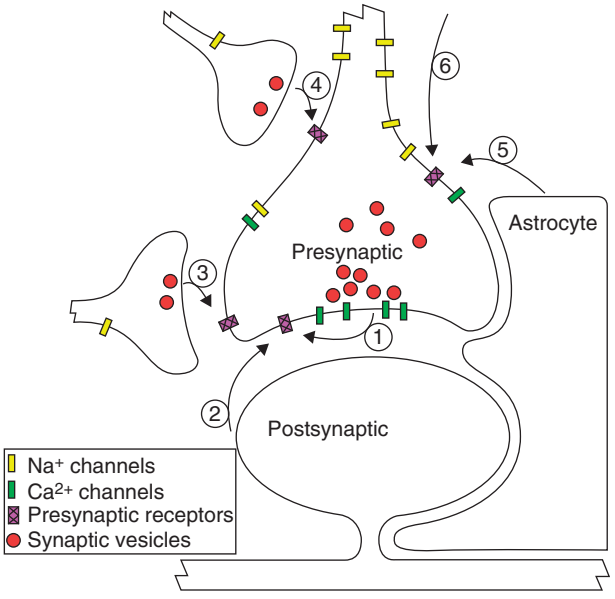

Figure 6 The sources of presynaptic receptor agonists. The figure illustrates six possible sources of the agonists activating presynaptic receptors. (1) The transmitter released by the presynaptic terminal activates presynaptic autoreceptors. (2) The agonist originates in the postsynaptic target cell. (3 and 4) Serial synapses. The transmitter originates from a presynaptic terminal which makes a synapse with the presynaptic terminal bearing the presynaptic receptors. The receptors can be autoreceptors (if the two presynaptic elements release the same transmitter) or heteroreceptors (if the two transmitters are different). The synapse can either be close to release sites and control Ca fluxes into the presynaptic terminal (3) or be situated on the axon (preterminal receptors) and act by modulating the action potential invasion of the terminal (4). (5) The agonist is released by a neighboring glial cell. (6) The agonist comes from a distance and acts like an 'endocrine' signal.

terminals of a third neuron contacting either the presynaptic terminal or the preterminal axon, glial cells, and distant sources (humoral, vascular, endocrine, etc.). These different pathways are likely to have different functional meanings.

Autoreceptors suggest the presence of a feedback loop, which can be negative or positive depending on the sign of the effect of the presynaptic receptors. However, such schematic expectations are probably oversimplified, as illustrated by the example of the presynaptic mGluRs at the calyx of Held. These presynaptic mGluRs are autoreceptors which depress transmitter release when activated by an exogenous agonist. They were expected to depress release during a repetitive stimulation, and they do, but this depression is surprisingly small compared to that produced by exogenous glutamate agonists. It turns out that the decrease in the release probability is masked by a simultaneous increase in the readily releasable pool, which passively increases when the release probability decreases. Billups et al. proposed that "mGluR autoreceptors function to change the synaptic state and distribute metabolic demand, rather than to depress synaptic amplitude."

Both autoreceptors and heteroreceptors may depend for their activation on the process of spillover, by which transmitters reach the receptors from another terminal. For example, in the amygdala, the presynaptic glutamate receptors found on the cortical terminals can be activated by glutamate released from thalamic afferents. Inside the cerebellar glomerulus, where mossy fibers and Golgi cell axons synapse on the dendrites of granule cells, glutamate released by the mossy fibers activates presynaptic mGluRs on the Golgi terminals and, conversely, GABA released by the Golgi axons activates $GABA_B$ receptors on the mossy fiber terminals. Because spillover implies a dilution of the transmitter, it usually activates receptors with high affinity for the transmitter (i.e., receptors with slow kinetics).

Retrograde signaling may involve classical transmitters such as glutamate and GABA but also liposoluble compounds, such as arachidonic acid, nitric oxide, and endocannabinoids. Retrograde signaling has been well characterized at the cerebellar synapses between GABAergic INs and the PN. Depolarization of the PN triggers the production of endocannabinoids which inhibit the GABA release by the INs. At this synapse, there is also evidence that glutamate can be released by the depolarization and act as a retrograde messenger which acts on presynaptic NMDA receptors of the INs and potentiates release.

Finally, signaling from glial cells or more remote sources may represent a more distributed form of modulation capable of integrating slow signals (or even steady signals) on the global level of network activity or even the general state of the organism.

The receptors involved in the six pathways can in principle be ionotropic or metabotropic, but the different pathways will privilege one type of receptor over another depending on their kinetics. Classical ionotropic receptors (nicotinic ACh receptors, AMPA receptors, and $GABA_A$ receptors) are short-acting receptors. They have a low affinity for the neurotransmitter, which thus rapidly dissociates from its binding site. The decay time constants of such conductances are typically in the millisecond range. These receptors are ill-equipped to trigger long-lasting changes. If such changes are detected, there must be secondary intracellular processes triggered by the initial conductance change. In contrast, metabotropic receptors all have slow kinetics, which is linked with the high affinity of the agonist (the dissociation of the agonist can last tens of milliseconds to seconds). This is also the case for some ionotropic receptors such as the NMDA receptors, some kainate receptors, and some GABA receptors. Such slow receptors are well adapted to the integration of slow signals but also to the detection of small concentrations of agonists. This is why they are the most frequently involved when the agonist comes from a distance, as is the case not only when there is spillover, but also when there is activation by an 'ambient' level of the agonist.

See also: Calcium Channels; Neuromodulation of Sodium Channels; Neuromodulation of Calcium Channels; Presynaptic Facilitation; Presynaptic Inhibition; Presynaptic Events in Neuromuscular Transmission; Sodium Channels; Synaptic Vesicles; Vesicle Pools.

Further Reading

Bamford NS, Zhang H, Schmitz Y, et al. (2004) Heterosynaptic dopamine neurotransmission selects sets of corticostriatal terminals. *Neuron* 42: 653–663.

Cochilla AJ and Alford S (1998) Metabotropic glutamate receptor-mediated control of neurotransmitter release. *Neuron* 20: 1007–1016.

Diana MA and Marty A (2004) Endocannabinoid-mediated short-term synaptic plasticity: Depolarization-induced suppression of inhibition (DSI) and depolarization-induced suppression of excitation (DSE). *British Journal of Pharmacology* 142: 9–19.

Duguid I and Sjöstrom PJ (2006) Novel presynaptic mechanisms for coincidence detection in synaptic plasticity. *Current Opinion in Neurobiology* 16: 312–322.

Engelman HS and MacDermott AB (2004) Presynaptic ionotropic receptors and control of transmitter release. *Nature Reviews Neuroscience* 5: 135–145.

Fuchs PA and Getting PA (1980) Ionic basis of presynaptic inhibitory potentials at crayfish claw opener. *Journal of Neurophysiology* 43: 1547–1557.

Lipscombe D, Kongsamut S, and Tsien RW (1989) Alpha-adrenergic inhibition of sympathetic neurotransmitter release mediated by modulation of N-type calcium-channel gating. *Nature* 340: 639–642.

Malenka RC and Siegelbaum SA (2001) Synaptic plasticity. Diverse targets and mechanisms for regulating synaptic efficacy. In: Cowan WM, Südhof TC, and Stevens C (eds.) *Synapses*, pp. 393–453. Baltimore, MD: Johns Hopkins University Press.

McGehee DS, Heath MJ, Gelber S, Devay P, and Role LW (1995) Nicotine enhancement of fast excitatory synaptic transmission in CNS by presynaptic receptors. *Science* 269: 1692–1696.

Mitchell SJ and Silver RA (2000) Glutamate spillover suppresses inhibition by activating presynaptic mGluRs. *Nature* 404: 498–502.

Sakaba T and Neher E (2003) Direct modulation of synaptic vesicle priming by GABA(B) receptor activation at a glutamatergic synapse. *Nature* 424: 775–778.

Schneggenburger R and Forsythe ID (2006) The calyx of Held. *Cell and Tissue Research* 326: 311–337.

Takahashi T, Forsythe ID, Tsujimoto T, Barnes-Davies M, and Onodera K (1996) Presynaptic calcium current modulation by a metabotropic glutamate receptor. *Science* 274: 594–597.

Turecek R and Trussell LO (2001) Presynaptic glycine receptors enhance transmitter release at a mammalian central synapse. *Nature* 411: 587–590.

Presynaptic Regulation by Liprins

K Miller, Michigan State University, East Lansing, MI, USA
D Van Vactor, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The synapse is the fundamental structure that allows neuronal networks and circuits to form and function. Presynaptic neurotransmission requires the formation and maintenance of a pool of synaptic vesicles through axonal transport, formation of specialized vesicular release sites, and local recycling of synaptic vesicle proteins. Postsynaptically, neuronal transmission involves the targeting and the clustering of synaptic receptors followed by receptor activation through neurotransmitters. The process of organization and localization of pre- and postsynaptic proteins is thought to depend on a variety of scaffolding proteins.

Scaffolding proteins are thought to organize the constituents of the active zone and to provide physical links between the active zone and synaptic vesicle proteins. Several proteins have been identified that may function as synaptic scaffolds: Rab3-interacting molecules (RIMs), which regulate synaptic vesicle exocytosis; Unc13 and Munc13, which are thought to be involved in priming; and three classes of proteins that are speculated to organize the structure of the active zone – the large presynaptic proteins Bassoon and Piccolo/aczonin, found in higher organisms, the well-conserved CAST/CAZ/ERC/Brp family of proteins, and the well-conserved liprin-α family.

Liprin-α (synonyms: LAR-interacting protein 1, LIP.1, SYD-2, Dliprin-α (CG11199), and 'out of step' (oos)) is a 130 kDa protein that is important for efficient neurotransmission and for shaping the structure of neuromuscular junctions (NMJs), synaptic boutons, and active zones. It contains an N-terminal coiled-coil domain that mediates homo- and heteromultimerization and a C-terminal region that contains three sterile alpha motif (SAM) domains that mediate binding to an array of partner proteins (Figure 1(a)). At the sequence level, liprin-α is highly conserved, from vertebrates to *Caenorhabditis elegans* and *Drosophila*. Liprin-α interacts with itself and with proteins of at least eight other classes: liprin-βs 1 and 2, receptor protein tyrosine phosphatases (leukocyte antigen-related protein (LAR), PTPσ, PTPδ), the anaphase-promoting complex/cyclosome (APC/C), RIM, ERC, kinesin-1 and -3, G-protein-coupled receptor kinase-interacting protein (GIT), CASK, and glutamate receptor-interacting protein (GRIP) (Figure 1(b)), all of which appear to be involved in normal synaptic function (Table 1).

Liprin Expression, Localization, and Diversity

In addition to its localization in neurons, liprin-α is widely expressed. In mammals, the mRNAs for liprin-α 1 and -β 1–2 have been found in all tissues that have been examined (heart, brain, placenta, lung, liver (very low levels of β 1 in liver), muscle, kidney, and pancreas). In contrast, liprin-αs 2–4 may play more specialized roles because their expression is restricted. The mRNA for liprin-α 4 is found only in heart, brain, and muscle, and the mRNAs for liprin-αs 2 and 3 are solely expressed in brain tissue. In all cell types, liprin-α has been implicated in the formation of cell–cell contacts.

In nonneuronal cells, liprin-α is enriched at focal adhesions and co-localizes with the receptor protein tyrosine phosphatase (RPTP) LAR, phosphotyrosine, talin, vinculin, paxillin, and pp125^{FAK}. It has been suggested that liprin-α is involved in the disassembly of focal adhesions, recruitment of LAR. Analysis of liprin-α may give insights into parallels between synaptic and focal adhesion assembly. In neurons, liprin-α is found both pre- and postsynaptically and at both excitatory and inhibitory types of synapses. In *Drosophila*, liprin-α localizes to synapses, with a broad distribution. However, synapse-defective *C. elegans* liprin-α (SYD-2) has been shown to cluster at the active zone.

While there are four isotypes of liprin-α in mammals, there is only a single version of liprin-α in *C. elegans* and in *Drosophila*; furthermore, little work has been done on liprin−αs 2–4. In addition, there are different splice forms of liprin-α 1 in mammals, with isoform 1B being the longest (version 1A lacks the last 30 amino acids which contain the GRIP-binding domain and accordingly does not bind to GRIP). Thus, unless otherwise specified, when we use the term liprin-α it will refer to the splice form liprin-α 1B in mammals and the single versions that are the most similar to liprin-α 1B in flies and worms.

Function of Liprin-α at the Synapse

Although liprins were first characterized in nonneuronal cells, the majority of functional data on liprin-α and its binding partners has been obtained from the nervous systems of worms and flies. Liprin-α function at the synapse was first appreciated in *C. elegans*,

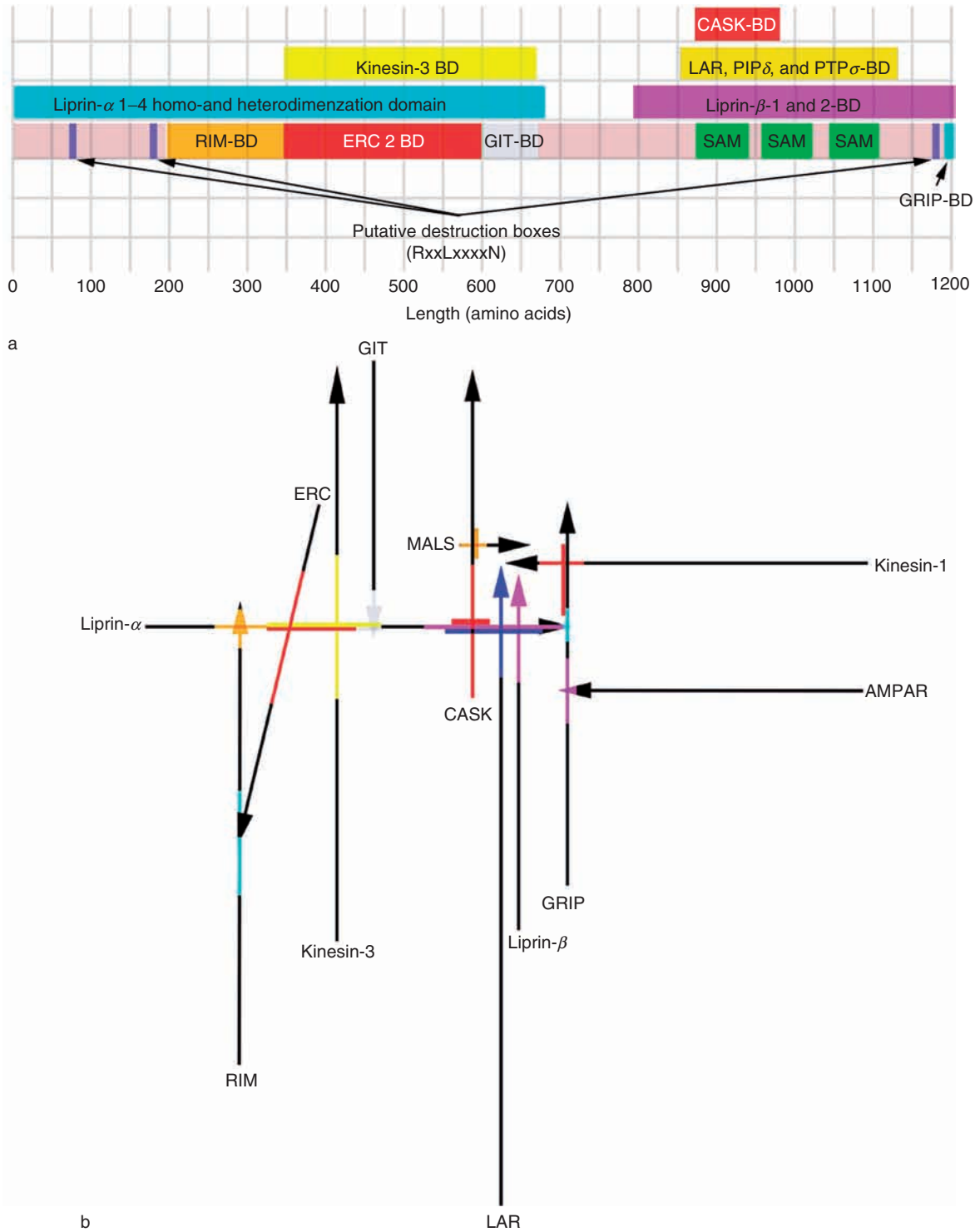

Figure 1 (a) Overview of protein-binding sites of liprin-α. (b) Interaction map of liprin interactions. All proteins are drawn to scale based on the number of amino acids. Arrowheads are the C-terminus of the protein. Mapped sites of protein–protein interaction are in matching color. AMPAR, α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid receptor.

where SYD-2 was identified in a screen for altered distribution of a synaptic vesicle marker at neuromuscular synapses. Although, like its vertebrate counterpart, the *C. elegans* NMJ is cholinergic, it is formed by muscle projections that reach into the neuropil to find motor axons, thus requiring morphogenetic expansion of presynaptic varicosities along the axon shaft. Behavioral assays, pharmacology, and

Table 1 Protein–protein interactions relevant to Liprin-α function

Protein	*Protein*
CASK	MALs
CASK	Syndecan
ERC2	RIM1
GIT	Piccolo–Bassoon
GRIP	AMPA receptor
GRIP	Kinesin-1
LAR	β-catenin
LAR	Syndecan
Liprin-α	APC/C complex
Liprin-α	CASK
Liprin-α	ERC2
Liprin-α	GITs
Liprin-α	GRIP1
Liprin-α	LAR
Liprin-α	Liprin-α
Liprin-α	Liprin-β
Liprin-α	Kinesin-1
Liprin-α	Kinesin-3
Liprin-α	RIM1
RIM	Piccolo–Bassoon

ultrastructural characterization showed that neuromuscular synapses function poorly and contain larger presynaptic active zones in mutants lacking liprin-α. Behaviorally, *C. elegans* mutants lacking liprin-α are sluggish in movement and retain twice as many eggs. These defects are fully rescued when liprin-α is selectively expressed in neurons, but not muscles. Since egg laying can be induced by treatment with serotonin (but not imipramine) and SYD-2 worms are resistant to aldicarb (an inhibitor of cholinesterase, which causes wild-type worms to hypercontract and die), these defects are both genetically and pharmologically consistent with a primary defect in presynaptic transmission.

In *Drosophila*, where NMJs are glutamatergic, liprin-α was identified as a binding partner of the RPTP LAR. In this context, liprin-α was shown to localize to the NMJ and was shown to be required for the overall morphological growth of the nerve terminal. In the fly, like in vertebrates, NMJs are branched expanses at the ends of axons where motor neurons make connections with muscles. Synaptic boutons, which are 0.5–2 μm in diameter, lie along NMJs and resemble beads on a string. Each bouton contains multiple active zones, typically between 50 and 100 nm in length. As observed in *C. elegans*, *Drosophila* NMJs lacking liprin-α display larger active zones (~2.5-fold higher mean area), suggesting that the role of liprins in the presynaptic cytomatrix is well conserved.

When liprin-α is disrupted, synaptic transmission is reduced in both flies and worms. Yet, there is an apparently paradoxical increase in the average size of active zones and a proportional increase in the number of presynaptic vesicles. Reassuringly, at the electron microsopic (EM) level, active zones appear less electron dense, and at the light microcopic level the intensities of staining for synaptobrevin, synaptotagmin, syntaxin, and Rab3 are decreased. As disruption of liprin-α has the functional consequence of disrupting synaptic transmission, it is possible that active zone size is increased in a futile attempt to compensate for decreases in evoked release. If so, the observed decreases in protein density at active zones could occur because a fixed amount of material is spread over a larger area. Alternatively, the absence of a key synaptic scaffold may simply reduce the cross-linking within the cytomatrix and result in a loose network of active zone proteins. A failure to recruit all the necessary presynaptic components or a decreased concentration of the remaining proteins could reduce the efficiency of neurotransmitter release.

One great advantage of *Drosophila* as a genetic model organism for studies of synapse biology is the accessibility of the NMJ to electrophysiological characterization. Disruption of *Drosophila* liprin-α (and Dlar (*Drosophila* LAR)) results in a ~30% decrease in evoked postsynaptic potentials and a ~60% decrease in quantal content (excitatory/mini-excitatory junction potentials (EJP/mEJP)). The amplitude and frequency of mini-excitatory junction potentials in these mutants are comparable to those in wild type, inferring that synaptic vesicles are present at the active zones, but there is a defect in evoked release. Similar defects in evoked release are also seen when Dlar, CAST, RIM, and APC/C are disrupted in different organisms. In combination with work on these proteins, this implies that liprin-α directly or indirectly determines the size of the releasable pool. Based on these results, liprin-α may be important in some aspect of docking or priming of presynaptic vesicles. Consistent with this hypothesis, liprin-α mutants display defects in vesicle trafficking.

Disruption of liprin-α results in ectopic accumulation of synaptic vesicle markers along the motor axon shaft (often called 'clogs') and a 25% reduction in both the number of synaptic boutons and the NMJ branches. One hypothesis to explain the effects seen in liprin-α mutants is that they result from an impairment of axonal transport. Indeed, a quantitative analysis of axon transport in living animals reveals that liprin-α positively regulates anterograde transport and negatively regulates retrograde transport. In *Drosophila* cells, liprin-α can associate with a complex including the heavy chain of kinesin-1 (Khc), consistent with the fact that *khc* mutants display similar axon clogs. Thus, liprin-α plays an interesting

role in the delivery of synaptic material. Whether the regulation of axon transport is liprin's primary role or whether this is one of many functions remains to be elucidated.

Partners of Liprin-α

APC/C and Liprin-α

The APC/C is an E3 ubiquitin ligase that regulates the concentration of liprin-α and modulates NMJ morphology in *Drosophila*. It controls the abundance of proteins by regulating their targeting to the ubiquitin-proteasome degradation system. While it has been established that APC/C is a critical mediator of cell-cycle transitions, recent work has demonstrated that it also functions in postmitotic cells. In *Drosophila* muscles, APC/C regulates synaptic transmission by controlling the concentration of a postsynaptic glutamate receptor. While the change in postsynaptic response may be due to an alteration in liprin-α-dependent clustering of glutamate receptors, the observation that mutations in liprin-α do not affect quantal size at the *Drosophila* NMJ suggests that APC/C may be regulating the degradation of other postsynaptic proteins. In neurons, APC/C regulates synaptic growth and synaptic transmission by controlling the concentration of the downstream effector liprin-α. When APC/C is disrupted, liprin-α intensity is increased presynaptically and the number of synaptic boutons is also increased. These changes are consistent with previous work which demonstrated that disruption of liprin-α leads to a decrease in the number of boutons in this system. Together, these findings suggest that the morphology of the NMJ is regulated through the control of liprin-α concentration downstream of APC/C.

MALS/CASK and Liprin-α

Liprin-α has now been shown to be a component of the MALS/Veli–CASK–Mint-1 complex of PDZ proteins. MALS has been implicated in presynaptic vesicle cycling. When all three isoforms of MALS proteins are disrupted in mice, the mice die soon after birth; they have difficulty breathing and show impaired excitatory synaptic transmission. Of particular interest, the MALS complex seems to be important for determining the size of the releasable pool of synaptic vesicles and may be important for replenishing this pool from the reserve pool. The lack of GRIP or α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA) receptors in the MALS complex suggests no role for liprin-α in the MALS complex postsynaptically. The lack of liprin-β suggests that this protein is not critical for the MALS/Veli–CASK–Mint-1/liprin-α complex. The Veli–CASK–Mint-1 complex, a homologous counterpart of the LIN-2–LIN-7–LIN-10 complex of *C. elegans*, has also been implicated in several other synaptic protein interactions, such as with voltage-gated N-type calcium channels (Table 1). This work is consistent with a model where a MALS/Veli–CASK–Mint-1/liprin-α complex couples extracellular synaptic interactions to the intercellular organization of the presynaptic secretory machinery.

Liprin-α and CASK

CASK (CAKI or CAMGUK in flies) is a membrane-associated guanylate kinase-associated (MAGUK) protein that has a catalytically inactive N-terminal CaM-kinase domain that interacts with Veli and Mint-1, an SH3 domain that binds N-type Ca^{2+} channels, and a central PDZ domain that interacts with other membrane proteins such as glycophorins, neurexins, and syndecan. The observation of an interaction between CASK and syndecan is particularly interesting in light of the recent demonstration that another liprin-α-interacting protein, LAR, also binds to syndecan. In addition, LAR and CASK both bind liprin-α in the same region (see Figure 1). We raise the possibility that the interaction between these three proteins may be relevant for their function, but determining how these proteins interact will take further investigation.

Disruption of *caki* results in an increase in spontaneous neurotransmitter release and an impairment in the response of the giant fiber pathway to continuous stimulation. These defects appear to involve the loss of presynaptic control of exocytosis, which is similar to the defect seen in liprin-α mutants. The defect in neurotransmission seen in *caki* mutants is one of many possibilities to explain how disruption of liprin-α leads to defects in neurotransmission.

Liprin-α and LAR

LAR family receptor protein tyrosine phosphatases (LAR, PTPδ, and PTPσ) bind liprin-α. Disruption of *Drosophila* LAR results in axonal pathfinding defects and defects in synaptic morphology and neurotransmission that are similar to those seen in liprin-α mutants. The extracellular regions of LAR-RPTP contain three N-terminal immunoglobulin-like domains and a variable number of fibronectin type III-like domains that are well conserved, from insects to humans. This suggests that these phosphatases may function as receptors for cell surface or extracellular matrix proteins. The heparan sulfate proteoglycans (HSPGs) syndecan and dallylike have recently been

shown to bind to the extracellular domains of LAR and to regulate synaptic growth and active zone form and function, respectively. The intracellular region of LAR contains two tandem phosphatase domains: the membrane proximal domain (D1), which has phosphatase activity, and the distal phosphatase domain (D2), which is devoid of catalytic activity but may regulate the substrate specificity of LAR. LAR has been shown to dephosphorylate β-catenin *in vitro*, but not liprin-α, although recent data suggest that liprin-α is tyrosine phosphorylated in some contexts. Liprin-α binds the D2 domain of LAR and co-localizes with it at focal adhesions and at NMJs. Although the functional relationship between liprin-α and LAR is not completely clear, is possible that LAR and liprin-α simply anchor each other at the active zone. In support of this hypothesis, expression of liprin-α 2 by transfection alters the distribution of LAR in cultured cells. Disruption of LAR in *Drosophila* reveals that it is not required for localization of liprin-α to the synapse; however, the reverse may be true. These results suggest that proteins other than Dlar can maintain synaptic liprin-α localization. Work in mammalian neurons suggests that the LAR–liprin-α interaction may contribute to synapse morphogenesis by regulating the synaptic delivery of a protein complex that includes AMPA receptors, GRIP, liprin-α, LAR, cadherin, and β-catenin. How the interaction between LAR and liprin-α relates to cadherin/β-catenin, syndecan, and dallylike function is an important future direction of research. In other contexts, such as epithelial cell polarity and axon guidance, LAR and its signaling partners regulate the structure of actin cytoskeleton, further raising the possibility that synaptic LAR and liprin-α play some role in cytoskeletal remodeling to control presynaptic structure.

Liprin-α, GRIP, and GluR2

GRIP is a multi-PDZ domain-containing protein that binds liprin-α, as well as both ephrin receptors and ligands and the kinesin-3 motor protein family member KIF1A. GRIPs have been suggested to mediate the transport of glutamate receptors, the stabilization of receptors within postsynaptic densities, and the sorting of internalized receptors. GRIP1 is found in synaptic preparations and intracellular compartments. This includes putative transport vesicles for glutamate receptors. Genetic analysis in mice shows that GRIP1 is required early during development, because a GRIP1 knockout is embryonic lethal at day 12 and the embryos suffer from defects in junction formation between dermis and epidermis.

Liprin-α isotypes 1, 3, and 4 interact with GRIP through the last seven amino acids of liprin-α, and liprin-α 1 immunoprecipitation pulls down GRIP and GluR2/3. This interaction is selective because the postsynaptic density 95 kDa protein (PSD-95) and *N*-methyl-D-aspartate (NMDA) do not associate with liprin-α 1. Of note, LAR interaction in this system is weak, as LAR does not immunoprecipate with liprin-α 1 unless it is cross-linked with dithiobissuccinimidylpropinate. In young neurons, GRIP and liprin-α are both present in the axons of cultured mouse hippocampal neurons, but in mature presynaptic regions these proteins do not co-localize, suggesting that the GRIP–liprin interaction may be more important for the assembly of synapses than for their maintenance. Although liprin-α seems to perform both pre- and postsynaptic functions, the interaction with GRIP seems to be most important postsynaptically, since *Drosophila* GRIP is only expressed in muscles and mouse liprin-α and GRIP interactions primarily involve postsynaptic proteins. Transfection with liprin-α 1A that lacks the C-terminus (TYSC) in hippocampal neurons reduces the number of AMPA receptors that cluster along dendrites. In addition, an interaction between GRIP, GIT1, and liprin-α is important for AMPA receptor targeting. This suggests that liprin-α is involved in the localization of AMPA receptors through its interactions with these proteins at postsynaptic sites.

Given the strong evidence that (1) liprin-α interacts with GRIP, (2) DGRIP is found in muscles, and (3) interaction between liprin-α and GRIP has been implicated in the clustering of glutamate receptors, it is surprising that disruption of liprin-α does not result in a decrease in postsynaptic sensitivity (i.e., the amplitude of mini-excitatory junction potentials). Nonetheless, as full rescue of synaptic function can be done by the selective expression of liprin-α (SYD-2) in neurons, it is possible that liprin-α may play a minor role in postsynaptic functioning in invertebrates, and a more important role in vertebrate synapses.

Liprin-α and GIT

GIT1 was originally isolated as a protein interacting with G-protein-coupled receptor kinases. The GIT family of proteins contains two known members, GIT1/Cat-1/p95-APP1 and GIT2/Cat-2/PKL/p95-APP2/KIAA0148. GIT proteins contain a GTPase-activating protein (GAP) domain for ADP-ribosylation factors (ARFs), small GTP-binding proteins implicated in the regulation of membrane traffic, and the actin cytoskeleton. In addition, GIT proteins contain various domains for protein interactions: ankyrin repeats, the Spa2 homology domain (SHD), and the G protein-coupled receptor kinase-binding domain (GRKBD). GIT proteins regulate endocytosis of various membrane proteins and regulate the assembly of focal adhesion

complexes by interacting with the Rho-type guanine nucleotide exchange factor (GEF) PIX, focal adhesion kinase (FAK), the focal adhesion adaptor protein paxillin, and liprin-α When GIT is disrupted, dendritic synapse number and AMPA receptor targeting are decreased. Piccolo has been proposed to play a role in GIT localization at active zones, but given the number of proteins that interact with GIT it seems likely that multiple proteins are influencing its distribution. As liprin-α and GIT are both found at focal adhesions and synapses, interact, and share a role in intracellular trafficking of AMPA receptors (AMPARs), it is tempting to speculate that the same machinery is used for 'trans-Golgi' trafficking, axonal transport, dendritic trafficking, and synaptic vesicle trafficking, but in a context-specific manner.

Liprin-α and ERCs

The ERC family of 'cytomatrix at the active zone' (CAZ) proteins contains two known members and both interact with liprin-α: ERC1/ELKS/Rab6IP2/CAST2 and ERC2/CAST1. Recent analysis of mutations in a *Drosophila* CAST ortholog, bruchpilot (NC82), reveals a requirement for CAST in synaptic development and active zone morphogenesis, consistent with the highly specific localization of this protein to the active zone. In addition, ERC interacts with other CAZ proteins, including RIM, Piccolo, Bassoon, and syntenin. ERC2 has been implicated in modulating liprin-α levels at the synapse: co-expressing ERC2 leads to increased levels of liprin-α. Functionally, the interaction of ERC2 with RIM1 and Bassoon/Piccolo has been shown to regulate neurotransmitter release. As liprin-α binds both ERC2 and RIM, the defect in synaptic evoked release that occurs when liprin-α is disrupted may occur through the disruption of these interactors. Since ERC2 and GIT interact with Rab6 and ARF, it has been suggested that liprin-α might mediate the integration of the Rab6 and ARF signaling pathways for the regulation of membrane traffic.

Liprin-α and RIM

RIM1 is an active zone protein that was initially identified as a putative effector for the synaptic vesicle protein Rab3A. RIM interacts with several active zone molecules, including Munc13–1 and liprin-α, to form a protein scaffold in the presynaptic nerve terminal. RIMs also bind to RIM-BPs through a C-terminal sequence, to cAMP-GEFII through an unknown site, and to Ca^{2+} channels and synaptotagmin 1 through both C2 domains. Abolishing the expression of RIM1 in mice shows that RIM1 is essential for maintaining normal probability of neurotransmitter release and for regulating release during short-term synaptic plasticity. This presumably occurs through a disruption of either vesicle priming or fusion, or both. Based on these observations, it seems likely that the presynaptic defect in evoked release seen when liprin-α is disrupted in *Drosophila* is caused in part through a disruption of RIM localization and/or function; however, mutations in *Drosophila* RIM have not been described.

Note Added in Proof

Sheng and colleagues have a very exciting story showing that the levels of liprin-α are activity-regulated in hippocampal neurons and appear to control the delivery of key components like LAR (consistent with the transport defects observed in Miller and co-authors). Also, the work by Treisman and colleagues on axon guidance is noteworthy.

Although much of the focus on liprin family proteins has been on synaptogenesis, recent work from Treisman and colleagues reveals that *Drosophila* Liprin-α is also necessary for the correct synaptic targeting of specific R7 photoreceptor cells to their targets in the adult medulla; this study challenges the notion the liprin serves only to localize LAR, suggesting that there are LAR-independent functions of liprin-α (Hofmeyer et al., 2006).

The latest work on liprin-α and LAR in mammalian hippocampal synapses suggests that liprin-α links neural activity to the remodeling of synapse structure through Ca(2+)/calmodulin-dependent protein kinase II (CamKII) regulation of liprin-α stability and its ability to deliver LAR to functional dendritic sites (Hoogenraad et al., 2007). This exciting study implies that the regulation of LAR activity will be important to our understanding of synaptic plasticity.

Kinesins, Liprin, and Presynaptic Vesicle Precursor Transport

There are two major subdivisions of axonal transport: fast and slow. Soluble cytoskeletal proteins such as tau, kinesin, dynein, myosin, and tubulin are transported at a rate of approximately $1\,mm\,day^{-1}$ by slow axonal transport. Membrane-bound proteins, associated with organelles such as presynaptic precursor vesicles, are transported at a rate of approximately $1\,\mu m\,s^{-1}$ by fast axonal transport. The kinesin family of molecular motors moves soluble proteins and organelles away from the cell body, and dynein moves these materials toward the cell body. When kinesin-1 is disrupted, axonal clogs that contain synaptic components occur along axons and synaptic transmission is disrupted.

While substantial evidence suggests that liprin-α plays a direct role in determining the form and function of synapses, recent data also reveal that liprin-α is involved

in the fast axonal transport of presynaptic vesicle precursors (PVPs). PVPs contain the raw materials for the assembly of synapses, and while they are thought to be transported primarily by kinesin-3 (Unc-104, Kif1a), kinesin-1 (conventional kinesin) family members may also play a role in their transport. Since axonal accumulations of endogenous synaptotagmin similar to those found in kinesin heavy chain (Khc) alleles are also observed in mutants lacking *Drosophila* liprin-α, we speculate that liprin-α is involved in the transport of a subset of synaptic components and that this subset may be transported by more than one family of kinesin.

At one time is was thought that there were molecules whose sole purpose was to link kinesins to cargos, but it now appears that most adaptors are scaffolding proteins that interact with multiple proteins and do not serve as simple linkers. Liprin-α binds to kinesin-1 and kinesin-3. Yet, the balance of the evidence is against liprin-α acting as a simple linker between PVPs and kinesins. When it is disrupted, the transport of PVPs is not halted. Instead, it shifts the probability of transport from anterograde (kinesin-mediated) to retrograde (dynein-mediated) transport. Thus, liprin-α is not required for the movement of PVPs *per se*. Instead of a simple linker, liprin-α appears as a modulator of axonal transport.

One of the most basic questions in neurobiology is how some types of neuronal components are trafficked to dendrites, and others are selectively transported into axons. Liprin-α is part of a complex that targets postsynaptic components to dendrites and involves GRIP, AMPARs, and kinesin-1. Of note, there are splice variants of liprin-α that lack the GRIP-binding domain. It is tempting to speculate that this differential targeting between axons and dendrites may involve variations in splicing. Yet further work will be required to determine if and how liprin-α could help to selectively guide components into either axonal or dendritic compartments. While several lines of evidence suggest that liprin-α regulates transport, a great deal of work remains to provide a satisfying molecular answer as to how this occurs.

Overview

Throughout this article we have outlined liprin-α function at the synapse and we find three reoccurring themes: (1) liprin-α binds multiple proteins that are required for efficient neurotransmitter release; (2) liprin-α regulates intercellular transport; and (3) liprin-α is involved in the postsynaptic clustering and targeting of AMPARs. A great deal of effort has been devoted to identifying which proteins bind to liprin-α; now the problem is understanding the mechanism of normal functioning.

See also: Active Zone; AMPA Receptor Cell Biology/Trafficking; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Comparative Biology of Invertebrate Neuromuscular Junctions; Development of Drosophila Neuromuscular Junctions; Liprins, ELKS, and RIM-BP Proteins; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Presynaptic Events in Neuromuscular Transmission; Rab3A Interacting Molecules (RIMs); Synaptic Vesicles; Ubiquitin–Proteasome System and Plasticity.

Further Reading

Ackely BD and Jin Y (2004) Genetic analysis of synaptic target recognition and assembly. *Trends in Neuroscience* 27: 540–547.

Collingridge GL, Isaac JT, and Wang YT (2004) Receptor trafficking and synaptic plasticity. *Nature Reviews Neuroscience* 5: 952–962.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews Neuroscience* 6: 201–214.

Hoefen RJ and Berk BC (2006) The multifunctional GIT family of proteins. *Journal of Cell Science* 119: 1469–1475.

Hofmeyer K, Maurel-Zaffran C, Sink H, and Treisman JE (2006) Liprin-α has LAR independent functions in R7 photoreceptor axon targeting. *Proceedings of the National Academy of Sciences of the United States of America* 103: 11595–11600.

Hoogenraad CC, Felice-Mojer MI, Spangler SA, et al. (2007) Liphrinalphal degradation by calcium/calmodulin-dependent protein kinase II regulates LAR receptor tyrosine phosphatase distribution and dendrite development. *Development Cell* 12: 587–602.

Johnson KG, Tenney AP, Ghose A, et al. (2006) The HSPGs Syndecan and Dallylike bind the receptor phosphatase LAR and exert distinct effects on synaptic development. *Neuron* 49: 517–531.

Kaufmann N, DeProto J, Ranjan R, et al. (2002) Drosophila liprin-alpha and the receptor phosphatase Dlar control synapse morphogenesis. *Neuron* 34: 27–38.

Miller KE, DeProto J, Kaufmann N, et al. (2005) Direct observation demonstrates that liprin-alpha is required for trafficking of synaptic vesicles. *Current Biology* 15: 684–689.

Olsen O, Moore KA, Nicoll RA, et al. (2006) Synaptic transmission regulated by a presynaptic MALS/liprin-alpha protein complex. *Current Opinion in Cell Biology* 18: 223–227.

Schnapp BJ (2003) Trafficking of signaling modules by kinesin motors. *Journal of Cell Science* 116: 2125–2135.

Schoch S, Castillo PE, Jo T, et al. (2002) RIM1alpha forms a protein scaffold for regulating neurotransmitter release at the active zone. *Nature* 415: 321–326.

Presynaptic: Mitochondria and Presynaptic Function

G David, University of Miami Miller School of Medicine, Miami, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Motor nerve terminals must be prepared for sudden increases in energy demand associated with vesicular release and recycling. This article summarizes studies using functional fluorescence imaging and electrophysiological recording that demonstrate that vertebrate motor terminals rely on their abundant mitochondria not only for adenosine triphosphate (ATP) production but also for efficient, sustained sequestration of large stimulation-induced calcium loads (Figure 1). Ca^{2+} influx into the mitochondria is essential for sustaining neuromuscular transmission during repetitive stimulation and is a major stimulant for mitochondrial respiration in motor terminals. Their heavy reliance on mitochondria may make motor terminals especially vulnerable to stresses and diseases that impair mitochondrial function.

Distribution of Mitochondria in Motor Nerve Terminals

A general theme governing the distribution of mitochondria in polarized cells (e.g., motor neurons) is their increased density in regions of high metabolic activity. This is especially evident in mature vertebrate motor terminals in which approximately 20–30% of the cytoplasm is occupied by mitochondria. In these terminals, mitochondria tend to cluster away from the active zone regions, which are mostly occupied by synaptic vesicles. This concentration of mitochondria in terminals is thought to be achieved by the complex regulation of their axonal transport by local signals of increased demand for mitochondrial function. Such signals, which include cytosolic adenosine diphosphate (ADP) and Ca^{2+}, are thought to trap mitochondria by arresting/slowing their movement and by inducing their binding to as yet unidentified docking proteins. Axonal transport of mitochondria occurs along microtubules and microfilaments and is powered by the ATP-using molecular motors kinesin and dynein. Anterograde transport of mitochondria from the cell body to axons and terminals uses the adaptor protein Milton, which links the mitochondrial protein Miro to the heavy chain of kinesin.

ATP Demands of Motor Nerve Terminals

Motor nerve terminals are activated by trains of action potentials, which produce influx of Na^+ and Ca^{2+}, and efflux of K^+. Long-term maintenance of the resting transmembrane electrochemical gradients for these ions and their rapid restoration after activity poses a large energetic challenge for the small (2–3 μm in diameter) motor nerve-terminal boutons (in mammals and reptiles) or long processes (in frogs). Because of the terminals' large surface-to-volume ratio, ion influx/efflux can rapidly produce large changes in cytosolic ionic composition. Also, many spinal motor neurons (and probably their motor terminals) have relatively low concentrations of Ca^{2+}-binding proteins to mitigate the elevation of cytosolic [Ca^{2+}] during stimulation-induced influx.

Maintenance of Ionic Electrochemical Gradients

ATP energizes the long-term maintenance of Na^+ and K^+ gradients by a plasma membrane Na^+/K^+-ATPase, which uses the energy derived form ATP hydrolysis to extrude 3 Na^+ from the cytosol and import 2 K^+ from the extracellular fluid. This Na^+/K^+ pump is the main ATP-consuming process in neurons. The Na^+ gradient maintained by this pump also contributes indirectly to extrusion of Ca^{2+} from the cytosol via a plasma membrane Na^+/Ca^{2+} exchanger. ATP also directly fuels a plasma membrane Ca^{2+}-ATPase that extrudes Ca^{2+} from the cytosol.

Neurotransmitter (Acetylcholine) Synthesis, Storage, and Release

The loading of acetylcholine (ACh) into synaptic vesicles, to concentrations of ~0.5 M, is accomplished by the vesicular ACh transporter in the vesicle membrane. This transporter exchanges cytoplasmic ACh for vesicular H^+ and is driven by the H^+ gradient (pH ~6.7 in the vesicle, ~7.2 in cytosol). This H^+ gradient is established by proton extrusion by a vesicular ATPase.

The uptake of choline from the synaptic cleft (where it is generated by the rapid degradation of released ACh by ACh esterase) is essential for maintaining ACh synthesis. The Na^+ gradient established by the Na^+/K^+-ATPase drives choline uptake via a high-affinity Na^+/choline co-transporter in the plasma membrane.

The ATP requirement for exo/endocytosis has not been studied directly in vertebrate motor nerve terminals; but in the ribbon synapse of fish bipolar cells, cytosolic ATP is required to prime synaptic vesicles for Ca^{2+}-dependent fusion. In *Drosophila* motor nerve terminals, ATP is required for the mobilization of the reserve pool of synaptic vesicles.

Signaling Roles of ATP

ATP is present in synaptic vesicles (~0.2 M) and is co-released with ACh. In cultured neuromuscular

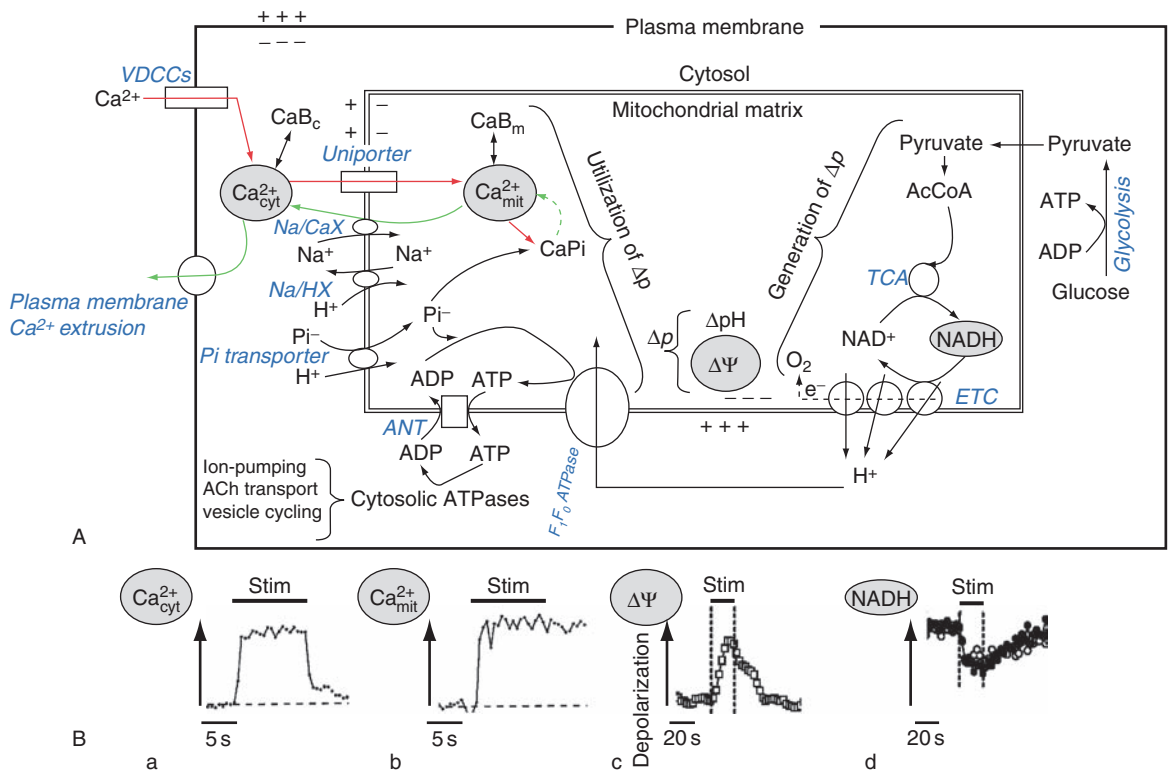

Figure 1 Mitochondrial processes: (A) simplified diagram of mitochondrial processes involved in generating and using the proton electromotive force (Δp) and in sequestering and extruding Ca^{2+}; (B) representative experimental recordings of stimulation-induced changes in the shaded components. In (A), Δp is composed of electrical ($\Delta\Psi_m$) and chemical (ΔpH) components. Depicted at the right are the processes involved in generating Δp; and at left are the processes that use Δp (predominantly its $\Delta\Psi$ component) to generate ATP and to temporarily sequester Ca^{2+} entering the cytosol during stimulation. Pathways for Ca^{2+} influx into the mitochondria and extrusion from the mitochondria are shown with red and green arrows, respectively. Shown in (B) are experimental recordings of changes in the components in (A) in response to 50 Hz stimulation of lizard motor nerve terminals. Cytosolic and mitochondrial Ca^{2+} (Ca^{2+}_{cyt} and Ca^{2+}_{mit}) are measured with Ca^{2+}-sensitive fluorescent dyes loaded selectively into the cytosolic or mitochondrial compartments, $\Delta\Psi_m$ is measured with the potentiometric dye rhodamine-123, and NADH is measured by imaging changes in its intrinsic fluorescence. $\Delta\Psi_m$, mitochondrial membrane potential; AcCoA, acetyl-coenzyme A; ACh, acetylcholine; ADP, adenosine diphosphate; ANT, adenine nucleotide translocator; ATP, adenosine triphosphate; CaB_c, calcium bound to cytosolic Ca^{2+}-binding proteins/sites; CaB_m, calcium bound to mitochondrial Ca^{2+}-binding proteins/sites; CaPi, osmotically inactive calcium–phosphate complex; ETC, electron transport chain; NADH, NAD^+, nicotinamide adenine dinucleotide; Pi, inorganic phosphate; stim, stimulation; TCA, tricarboxylic acid cycle; VDCCs, voltage-dependent Ca^{2+} channels. (B) (Ba, Bb) Reproduced from David G (1999) Mitochondrial clearance of cytosolic Ca^{2+} in stimulated lizard motor nerve terminals proceeds without progressive elevation of mitochondrial matrix [Ca^{2+}]. *Journal of Neuroscience* 19: 7495–7506, with permission from Society for Neuroscience. (Bc, Bd) Reproduced from Talbot JD, Barrett JN, Barrett EF, and David G (2007) Stimulation-induced changes in NADH fluorescence and mitochondrial membrane potential in lizard motor nerve terminals. *Journal of Physiology* 579(3): 783–798, with permission from Wiley-Blackwell Publishing Ltd.

junctions, the retrograde effects of this co-released ATP potentiate spontaneous vesicular ACh release and might play a role in synaptic development. It has been suggested that endogenously released ATP modulates intracellular [Ca^{2+}] in perisynaptic Schwann cells of amphibian neuromuscular junctions. It has also been suggested that presynaptic actions of adenosine, derived from hydrolysis of released ATP, contribute to the depression of transmitter release in frog neuromuscular junctions. Although exogenously applied ATP depresses transmitter release in mammalian neuromuscular junctions, the endogenous release of ATP (or the accumulation of adenosine) might not be sufficient to alter evoked transmitter release under physiological conditions.

ATP Supply by Cytosolic Glycolysis and Mitochondrial Oxidative Phosphorylation

Glucose is the major initial substrate for both these pathways, present at concentrations of approximately 4 mM in peripheral axons. Glucose transport from the extracellular fluid has been demonstrated in mammalian and amphibian motor nerve terminals and is found to proceed faster in terminals of fast (compared to slow) motor units. Because the glycolytic pathway is more prominent in fast terminals, this faster glucose transport might compensate for the smaller yield of ATP production from glycolysis compared to oxidative phosphorylation (see the

section titled 'Mitochondrial oxidative phosphorylation'). An additional source of glucose is glycogen, present at a concentration of approximately 1 mM in some peripheral nerves. Glycogen-phosphorylase breaks down glycogen and delivers glucose-1-phosphate to the glycolytic pathway; this enzyme is present in the somata, axons, and terminals of motor neurons.

Glycolysis

The glycolytic pathway employs cytosolic enzymes to catalyze the net generation of two ATP molecules from ADP via a multistep anaerobic conversion of glucose to lactate. Significantly, two ATPs are actually consumed in the first stage of glycolysis (from glucose to glyceraldehyde 3-phosphate), and four ATPs are produced in the second stage (3-phosphoglycerol phosphate to pyruvate). This need for initial priming of glycolysis by ATP suggests that, under conditions of severe ATP depletion, ATP production by glycolysis might be inhibited. Glycolytically supplied ATP has been shown to delay the rundown of transmitter release in frog motor terminals stimulated by prolonged Ca^{2+} influx under conditions in which mitochondrial ATP generation was inhibited by oligomycin. Glycolytic enzyme activity detected in spinal cord mammalian α-motor neurons did not differ for motor neurons innervating fast versus slow motor units.

Mitochondrial Oxidative Phosphorylation

Under aerobic conditions, pyruvate derived from glycolysis (or amino acids) is imported into the mitochondrial matrix, converted to acetyl-coenzyme A and enters the Krebs (or tricarboxylic acid (TCA)) cycle. Here several reactions result in the reduction of NAD^+ to nicotinamide adenine dinucleotide (NADH), which enters the respiratory chain at complex I. In the respiratory chain complexes, located in the inner mitochondrial membrane, a series of redox reactions involving electron carriers result in the transfer of electrons from NADH (at complex I) and succinate (at complex II) to O_2 (reduced to H_2O at complex IV). This linear series of electron transfer reactions is driven by the large difference between the redox potential of the NAD^+/NADH (~ -300 mV) couple and the O_2/H_2O couple ($\sim +800$ mV). Part of this energy is converted into an electrochemical gradient of H^+ ions across the inner mitochondrial membrane by the parallel extrusion of H^+ from the matrix at complexes I, III, and IV of the respiratory chain. This proton gradient, termed the proton motive force (Δp) is composed of two components: an electrical component ($\Delta\Psi_m$), derived from the negative potential in the matrix (compared to cytosol), and a chemical component ΔpH, derived from the lower concentration of protons (alkalinity) in the matrix compared to cytosol. The relationship between these two components (when converted to millivolts at 25 °C) is embodied in the equation $\Delta p = \Delta\Psi_m - 59\Delta pH$. The respiratory chain is capable of producing a Δp of ~200 mV, which under physiological conditions is mostly (~90%) composed of $\Delta\Psi_m$ (−180 mV in the matrix). Virtually all the metabolic and transport functions of the mitochondria are energized by the Δp (Figure 1). For example, ATP synthesis in the matrix from inorganic phosphate (Pi) and ADP uses the Δp in the following steps:

1. Cytosolic Pi (as $H_2PO_4^-$) is symported into the matrix with an H^+ (electroneutrally) using ΔpH.
2. Cytosolic ADP^{3-} is exchanged with matrix ATP^{4-} by the adenine nucleotide transporter; this net transfer of one positive charge into the matrix is thus facilitated by $\Delta\Psi_m$.
3. ADP phosphorylation to ATP by the F_1,F_0-ATP synthase (complex V) is energized by H^+ influx (driven mainly by $\Delta\Psi_m$) via the F_0 region of this enzyme. This combination of reactions displaces the cytosolic ratio of [ATP]/([ADP][Pi]) well above its equilibrium, creating a phosphorylation potential that supplies the cell's energetic demands. The number of ATP molecules produced from glucose when oxidative phosphorylation operates is approximately 10–20 times greater than the ATP yield from anaerobic glycolysis. Glycolysis can be accelerated under anaerobic conditions. Interestingly, cytosolically generated ATP can act to preserve $\Delta\Psi_m$; this is accomplished by reversing the reaction catalyzed by the F_1,F_0-ATP synthase from its physiological direction (using $\Delta\Psi_m$ to generate ATP) to the ATPase direction (using (and perhaps depleting) ATP to generate/maintain $\Delta\Psi_m$).

The relative contributions of mitochondrial oxidative phosphorylation and glycolysis have not been studied directly in motor nerve terminals, but histochemical measurements of oxidative and glycolytic enzymes in spinal motor neurons indicate that their metabolic profile closely matches that of their target muscle fibers. A classification of skeletal muscle fibers, based on their ability to maintain contraction force during continuous activation, includes slow-twitch (type I) and fast-twitch (type II) fibers. A motor neuron innervates one or more muscle fibers of the same type, forming motor units with type-specific patterns of action potential activity. Slow motor units are activated at low frequencies (10–20 Hz) for sustained periods, whereas fast motor units are activated in higher frequency bursts (50–80 Hz) for briefer periods. In general, ATP production in fast muscle fibers relies more on glycolysis,

whereas that in slow fibers relies more on oxidative phosphorylation. There is evidence that these metabolic pathways are also employed differentially in fast and slow motor neurons and their terminals. For example, oxidative enzyme activity is higher in slow motor neurons, and mitochondrial density in slow motor neurons and their terminals is higher than that in their fast counterparts.

Motor Terminal Function in the Absence of Mitochondrially Supplied ATP

Transmitter release in response to nerve stimulation is abolished after approximately 20 min hypoxia (rat diaphragm, 33–38 °C) in the presence of normal glucose. This neuromuscular block is initiated by decreased evoked transmitter release rather than by axonal conduction block. It is clear that long-term motor nerve terminal function cannot be sustained by anaerobic glycolysis alone, but the deterioration that occurs during normoglycemic hypoxia is accelerated by the depletion of glycolytically supplied ATP by the reversed operation of the F_1,F_0-ATPase previously mentioned. When mitochondrial ATP synthesis is blocked without depolarizing mitochondria (using oligomycin), evoked release persists for longer times (>1 h).

Mitochondria as Regulators of Cytosolic [Ca^{2+}] during Repetitive Stimulation

The concept that mitochondria not only supply ATP but also powerfully sequester Ca^{2+} originated in studies led by Albert Lehninger. The transport mechanisms by which mitochondria accomplish this task were further refined by David Nicholls' laboratory. Consideration of the K_ds for various Ca^{2+} transport mechanisms suggests that in motor terminals the resting cytosolic [Ca^{2+}] concentration (~0.05–0.1 μM) is controlled mainly by high-affinity transport mechanisms such as the plasma membrane Ca^{2+} ATPase. When motor terminals are stimulated repetitively, the resulting elevation of cytosolic [Ca^{2+}] recruits additional Ca^{2+} transport mechanisms, especially those associated with mitochondria. Stimulation at 50 Hz (within the physiological range of discharge frequencies) causes cytosolic [Ca^{2+}] to rise rapidly at first (Figures 2(b) and 3(c)), but then the rate of rise slows dramatically, even though Ca^{2+} continues to enter the terminals (as evidenced by the continued evoked transmitter release). (In most imaging protocols, cytosolic Ca^{2+}-sensitive dyes report a temporal and spatial average of [Ca^{2+}] in the stimulated terminal rather than the larger, more localized [Ca^{2+}] elevations near open Ca^{2+} channels in the plasma membrane.) The plateauing behavior of cytosolic [Ca^{2+}] has been documented in motor terminals from crayfish, lizard, frog, and mouse over frequencies of 20–100 Hz, reaching a range of 0.5–4 μM [Ca^{2+}]. What mechanism limits the elevation of cytosolic [Ca^{2+}] during stimulation? Treatments that inhibit plasma membrane (Ca^{2+}-ATPase and Na^+/Ca^{2+} exchanger) and endoplasmic reticular Ca^{2+} transport have relatively little effect on the magnitude of the [Ca^{2+}] responses in lizard and mouse vertebrate motor terminals, but when mitochondrial Ca^{2+} uptake is inhibited, cytosolic [Ca^{2+}] rises to much higher levels (Figure 2(c)). In frog motor nerve terminals, Ca^{2+} uptake by the endoplasmic reticulum also collaborates with mitochondria in clearing the Ca^{2+} load. Thus, mitochondria are critical to understanding how motor terminals handle large Ca^{2+} loads, as well as the associated energy demands.

Mechanisms of Mitochondrial Ca^{2+} Uptake

Much of our understanding of mitochondrial Ca^{2+} handling has come from studies of isolated mitochondria. Ca^{2+} enters mitochondria passively, down a large electrochemical gradient created mainly by H^+ extrusion by complexes I, III, and IV of the electron transport chain (ETC). The membrane potential ($\Delta\Psi_m$) in resting mitochondria is estimated to be −150 to −200 mV. In spite of this large electrical gradient favoring Ca^{2+} influx, the free Ca^{2+} concentration in the matrix of resting mitochondria is low (~0.1 μM) because the small Ca^{2+} influx at resting conditions is fully balanced (or exceeded) by Ca^{2+} extrusion via the mitochondrial Na^+/Ca^{2+} exchanger. Elevation of cytosolic [Ca^{2+}] opens a Ca^{2+} uniporter channel in the inner mitochondrial membrane. Opening of this uniporter demonstrates a steeper-than-linear (second- to third-power) dependence on cytosolic [Ca^{2+}]. Some studies in isolated mitochondria suggested that large [Ca^{2+}] elevations in the micromolar range were necessary to initiate channel opening. But cytosolic components such as spermine lower the threshold for channel opening, and studies of mitochondria within motor terminals suggest uniporter opening with [Ca^{2+}] elevations of only a few hundred nanomolars.

Regulation of [Ca^{2+}] in the Mitochondrial Matrix

Mitochondria can sequester large amounts of Ca^{2+}, especially when the Ca^{2+} load is delivered gradually. This has been demonstrated in isolated mitochondria by adding known amounts of Ca^{2+} to the surrounding bath and measuring both total mitochondrial Ca^{2+} uptake and free matrix [Ca^{2+}] with Ca^{2+}-sensitive dyes. Total Ca (free + bound) in mitochondria within

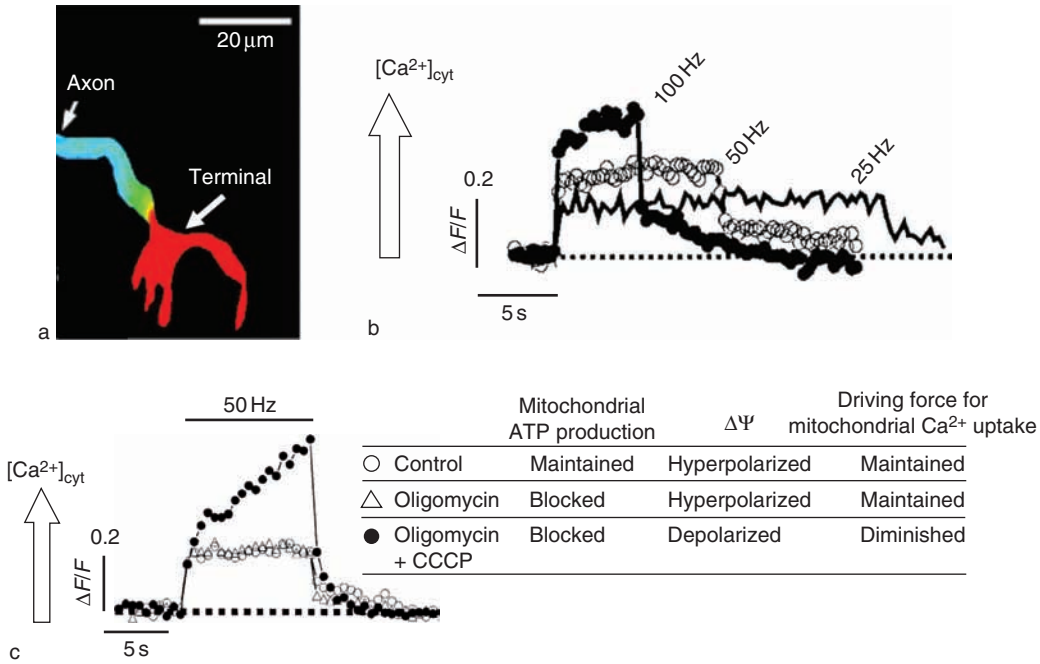

Figure 2 Stimulation-induced elevations of cytosolic $[Ca^{2+}]$ in motor terminals stimulated at different frequencies and in the presence of oligomycin and a proton carrier (CCCP): (a) pseudo-color image of an indicator-filled motor terminal and preterminal axon during 50 Hz stimulation of the motor nerve; (b) Changes in cytosolic $[Ca^{2+}]$ evoked in a motor terminal by trains of action potentials (delivered at 100, 50, and 25 Hz); (c) stimulation-induced increase in cytosolic $[Ca^{2+}]$ with and without CCCP. In (a), the image shows an elevation of $[Ca^{2+}]$ (red) localized to the motor terminal. The fluorescent Ca^{2+}-indicator dye was injected via a microelectrode into a distant internode of the axon and allowed to diffuse into the terminal. Image in (a) came from a motor terminal innervating the mouse levator auris longus muscle, injected with the Ca-sensitive dye Oregon green-BAPTA-1. In (b), during stimulation there is a rapid rise followed by a plateau whose amplitude increases with frequency. Following stimulation, there is a rapid decrease followed by a slow decay to baseline. As shown in (c), the stimulation-induced increase in cytosolic $[Ca^{2+}]$ is much greater when mitochondrial Ca^{2+} uptake is inhibited by depolarizing $\Delta\Psi_m$ with CCCP. This effect is due to the blockade of mitochondrial Ca^{2+} uptake rather than to the inhibition of mitochondrial ATP synthesis because oligomycin, which blocks ATP synthesis without depolarizing $\Delta\Psi_m$, did not (in the short term) alter the $[Ca^{2+}]$ response. Recordings in (b) and (c) came from motor terminals innervating the mouse internal oblique muscle using Oregon green-BAPTA-5N. Muscle contractions were inhibited with D-tubocurarine, which blocks muscle acetylcholine receptors; 32–34 °C. F, fluorescence; ATP, adenosine triphosphate. (b, c) Reproduced from David G and Barrett EF (2000) Stimulation-evoked increases in cytosolic $[Ca^{2+}]$ in mouse motor nerve terminals are limited by mitochondrial uptake and are temperature-dependent. *Journal of Neuroscience* 20(19): 7290–7296, with permission from Society for Neuroscience.

cells can also be measured by electron-probe microanalysis. Both techniques demonstrate that mitochondria can accumulate millimolar amounts of Ca. Several mechanisms allow sustained mitochondrial Ca^{2+} uptake. The first mechanism is that the $\Delta\Psi_m$ depolarization produced by Ca^{2+} uptake accelerates H^+ extrusion by the ETC by reducing the electrical gradient against which H^+ extrusion occurs. This acceleration of ETC activity limits the extent of $\Delta\Psi_m$ depolarization, thus maintaining an electrical gradient favoring Ca^{2+} entry. However, the accelerated H^+ efflux via this mechanism would be expected to hinder Ca^{2+} uptake by alkalinizing the matrix and converting the $\Delta\Psi_m$ that drives Ca^{2+} uptake into ΔpH. The second mechanism offsets this potential limitation of Ca^{2+} uptake due to matrix alkalinization by importing H^+ electroneutrally via the mitochondrial Pi/H^+ symporter. The third mechanism, also dependent on Pi influx, limits the increase in free matrix $[Ca^{2+}]$ to only 1–2 μM via the reversible formation of an (insoluble) Ca–phosphate complex (probably amorphous calcium phosphate, $Ca_3(PO_4)_2$) in the matrix, thereby removing Ca^{2+} from solution when the solubility product for this complex is exceeded. It appears that matrix $[Ca^{2+}]$ remains clamped near this level as long as this complex is present, that is, until sufficient Ca^{2+} is extruded to bring matrix $[Ca^{2+}]$ below the critical value required for complex formation. This powerful matrix-buffering mechanism requires metabolic activity because the formation of the CaPi complex is favored by the alkaline pH provided by ETC H^+ extrusion. The buffering of free-matrix $[Ca^{2+}]$ protects this organelle from several deleterious consequences of high free-matrix $[Ca^{2+}]$, including osmotic swelling and the release of cytochrome C, activation of mitochondrial calpains, inhibition of complex I of the ETC, inhibition of cytochrome oxidase by nitric oxide generated by Ca^{2+}-activated mitochondrial nitric oxide synthase,

and possibly opening of the mitochondrial permeability transition pore (MPTP).

How Does Sequestered Calcium Leave the Mitochondria?

Two major mechanisms have been identified in neuronal mitochondria. One is extrusion via the mitochondrial Na^+/Ca^{2+} exchanger (which is distinct from the Na^+/Ca^{2+} exchanger in the plasma membrane). The Na^+ that enters the matrix via this exchanger is, in turn, extruded by a Na^+/H^+ exchanger, so this extrusion mechanism uses the energy of the transmembrane $[H^+]$ gradient produced by ETC activity. The other mechanism is the opening of the MPTP, which can rapidly release Ca^{2+} from the matrix, but at the cost of completely depolarizing $\Delta\Psi_m$ and releasing other mitochondrial components (e.g., cytochrome C) that can trigger cell death. Measurements of mitochondrial $[Ca^{2+}]$ in mouse motor nerve terminals indicate that the predominant method of removing stimulation-induced Ca^{2+} loads from the matrix is the mitochondrial Na^+/Ca^{2+} exchanger. After 50 Hz stimulation under physiological conditions, matrix $[Ca^{2+}]$ declines slowly following stimulation (Figures 3(a4) and 3(b)). After the addition of CGP-37157 to inhibit the mitochondrial Na^+/Ca^{2+} exchanger, matrix $[Ca^{2+}]$ remains elevated (Figure 3(b)); the decrease in Ca^{2+} efflux from the mitochondria is also manifested by a faster decay of poststimulation cytosolic $[Ca^{2+}]$ (Figure 3(c)). In contrast, application of cyclosporin A, which inhibits MPTP opening, had no significant effect on mitochondrial or cytosolic $[Ca^{2+}]$ responses. The Ca^{2+} released from mitochondria to the cytosol after stimulation ends is eventually extruded from the cytosol

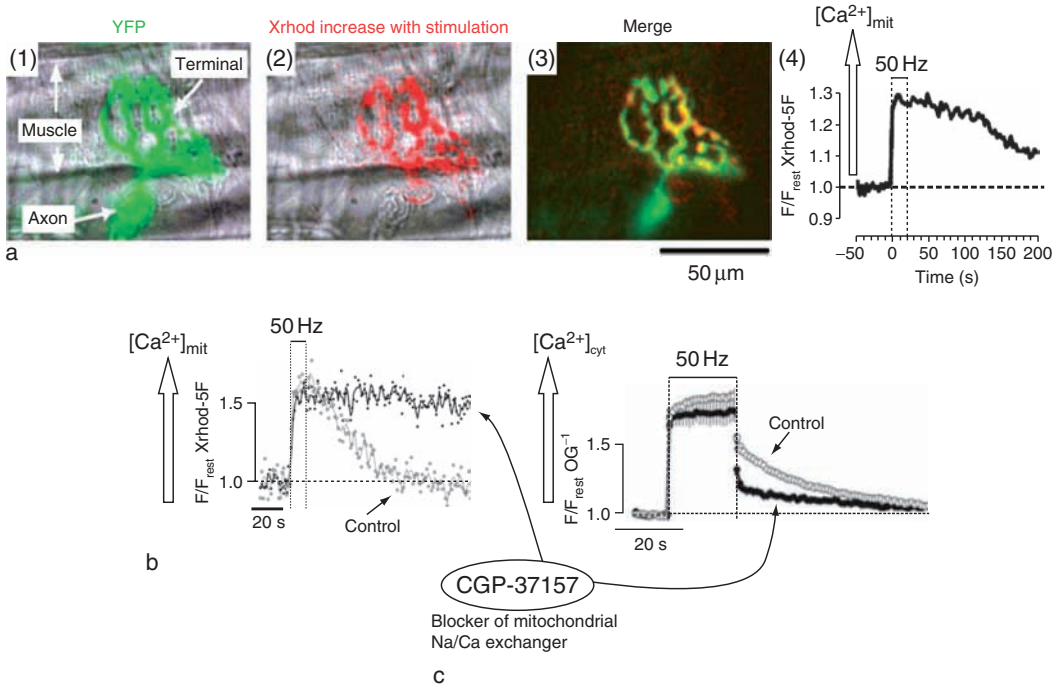

Figure 3 Extrusion of mitochondrially sequestered Ca^{2+} into the cytosol: (a) micrographs (1)–(3) showing a motor terminal and axon in a mouse that transgenically expresses yellow fluorescent protein (YFP; shown as green) in neurons and a graph (4) of mitochondrial $[Ca^{2+}]$ during stimulation; (b) mitochondrial $[Ca^{2+}]$ decay after the addition of CGP-37157 to inhibit Ca^{2+} extrusion via the mitochondrial Na^+/Ca^{2+} exchanger; (c) cytosolic $[Ca^{2+}]$ after the addition of CGP-37157 to inhibit Ca^{2+} efflux from the mitochondria. In (a), the YFP image is superimposed on a phase image (1) showing the underlying muscle fibers. Image (2) shows in red the regions in which mitochondrial $[Ca^{2+}]$ increased during stimulation at 50 Hz, indicated by a Ca^{2+}-sensitive dye, Xrhod-5F. Yellow regions in the merged image in (3) show that the increase in mitochondrial $[Ca^{2+}]$ occurred within the terminal arborization. The graph (4) shows that during stimulation mitochondrial $[Ca^{2+}]$ in this terminal exhibited a rapid increase followed by a plateau; when stimulation stopped, mitochondrial $[Ca^{2+}]$ decreased slowly. As shown in (b) mitochondrial $[Ca^{2+}]$ decays even more slowly after the addition of CGP-37157 to inhibit Ca^{2+} extrusion via the mitochondrial Na^+/Ca^{2+} exchanger. In both (a4) and (b), the plateau of matrix $[Ca^{2+}]$ during stimulation and the early posttetanic interval were not due to the saturation of the fluorescent indicator but, rather, to a powerful matrix Ca^{2+}-buffering mechanism. As shown in (c), inhibiting Ca^{2+} efflux from the mitochondria with CGP-37157 decreased the posttetanic tail of cytosolic $[Ca^{2+}]$, demonstrating that the extrusion of sequestered Ca^{2+} from the mitochondria elevates posttetanic $[Ca^{2+}]$ in the cytosol. Recordings are from motor terminals innervating the mouse levator auris longus muscle, in the presence of D-tubocurarine. F, fluorescence. Reproduced from Garcia-Chacon LE, Nguyen KT, David G, and Barrett EF (2006) Extrusion of Ca^{2+} from mouse motor terminal mitochondria via a Na^+/Ca^{2+} exchanger increases post-tetanic evoked release. *Journal of Physiology* 574: 663–675, with permission from Wiley-Blackwell Publishing Ltd.

via the plasma membrane Na^+/Ca^{2+} exchanger and the plasma membrane Ca-ATPase.

Impact of Mitochondrial Ca^{2+} Handling on Transmitter Release

Measurements of transmitter release confirm the importance of mitochondrial Ca^{2+} sequestration and subsequent extrusion to motor terminal function.

When mitochondrial Ca^{2+} uptake is inhibited by depolarizing $\Delta\Psi_m$, phasic transmitter release, which produces the endplate potential recorded in the underlying muscle, depresses rapidly during stimulation (Figure 4(a)). These $\Delta\Psi_m$-depolarizing drugs do not (within the first hour) alter the endplate potential recorded at the onset of stimulation, so the extra depression is not due to the acceleration of the initial phasic release. Rather, the accelerated depression of phasic release appears to be due to the acceleration of asynchronous release, which increases as cytosolic $[Ca^{2+}]$ increases to high levels (Figure 4(b)). This excessive asynchronous release depletes the pool of vesicles available for phasic release. Thus, by limiting the elevation of cytosolic $[Ca^{2+}]$, mitochondrial Ca^{2+} uptake helps preserve the phasic release necessary to sustain muscle contraction during repetitive stimulation. Because mitochondrial $[Ca^{2+}]$ uptake is initiated only after approximately 25–50 action potentials (25–50 Hz), the amplitude and recovery rate of cytosolic $[Ca^{2+}]$ after a single stimulus relies mainly on fast diffusion and binding of subplasmalemmal Ca^{2+} to cytosolic buffers. Thus, mitochondrial Ca^{2+} sequestration does not terminate transmitter release after a single action potential; rather, mitochondrial Ca^{2+} uptake limits the buildup of cytosolic $[Ca^{2+}]$ during repetitive activity, thus protecting the vesicular pool from massive depletion via Ca^{2+}-driven asynchronous release.

Extrusion of Ca^{2+} from mitochondria into the cytosol following stimulation also has important effects on transmitter release (Figure 4(c)). When mitochondrial Ca^{2+} extrusion is inhibited by CGP-37157, posttetanic phasic transmitter release exhibits more depression at high quantal contents and less potentiation at low quantal contents. Also consistent with a contribution of mitochondrial Ca^{2+} extrusion to posttetanic release is the observation that prolonging the tetanus results in the prolongation of both the posttetanic changes in release and the posttetanic decay of mitochondrial $[Ca^{2+}]$. Thus, the rapid sequestration of cytosolic Ca^{2+} by the mitochondria during stimulation protects phasic transmitter release from depletion via asynchronous release; the slow extrusion of this Ca^{2+} back into the cytosol after stimulation augments phasic release, thus speeding its recovery from posttetanic depression. These recent findings thus confirm the hypotheses proposed by Rami Rahamimoff and later by Robert Zucker that mitochondrial Ca^{2+} handling is a critical regulator of transmitter release from motor nerve terminals. Although mitochondria are abundant in most nerve terminals, it is as yet unclear whether mitochondrial Ca^{2+} transport is as important in other terminals as it is in vertebrate motor nerve terminals. Analysis of plasma membrane and organellar Ca^{2+}-handling mechanisms in a large central nervous system (CNS) synapse, the calyx of Held, suggests a less dominant role for mitochondrial Ca^{2+} transport. *Drosophila* mutants whose motor terminals lack mitochondria did not exhibit larger-than-normal cytosolic $[Ca^{2+}]$ responses during short stimulus trains, but in that study stimulation frequencies were limited to 10 Hz. At this relatively low stimulation frequency, mitochondrial Ca^{2+} transport also plays a relatively small role in vertebrate motor terminals. Thus, in assessing mitochondrial contributions to presynaptic Ca^{2+} handling, it is important to consider the magnitude of the Ca^{2+} load. It is also important to consider temperature because mitochondrial contributions to Ca^{2+} handling in mouse motor terminals decrease at lower temperatures. The relative importance of mitochondrial Ca^{2+} handling may also diminish in diseases characterized by impaired mitochondrial function.

Interaction between Mitochondrial Ca^{2+} Uptake and Mitochondrial Metabolism

The information summarized so far demonstrates that nerve stimulation and the ensuing vesicular recycling impose energy demands on the motor terminal and, thus, would be expected to alter mitochondrial metabolism. Two measures of mitochondrial metabolism, [NADH] and $\Delta\Psi_m$, change during stimulation of lizard motor nerve terminals (Figure 5). NADH autofluorescence decreases during the stimulus train, demonstrating a reversible oxidation of fluorescent NADH to nonfluorescent NAD^+. A partial, reversible depolarization of $\Delta\Psi_m$ occurs over roughly the same time course. Both these changes require Ca^{2+} influx into the motor terminal; they disappear when Ca^{2+} is removed from the bath or when the main voltage-dependent Ca^{2+} channels in these motor terminals are blocked with ω-conotoxin GVIA. Surprising to us, the stimulation-induced changes in [NADH] and $\Delta\Psi_m$ persisted when vesicular release was blocked by botulinum toxin A.

What Causes These Changes in [NADH] and $\Delta\Psi_m$?

An unlikely cause of the decrease in [NADH] is the deceleration of the tricarboxylic acid (TCA) cycle, which produces NADH. A more likely cause is the

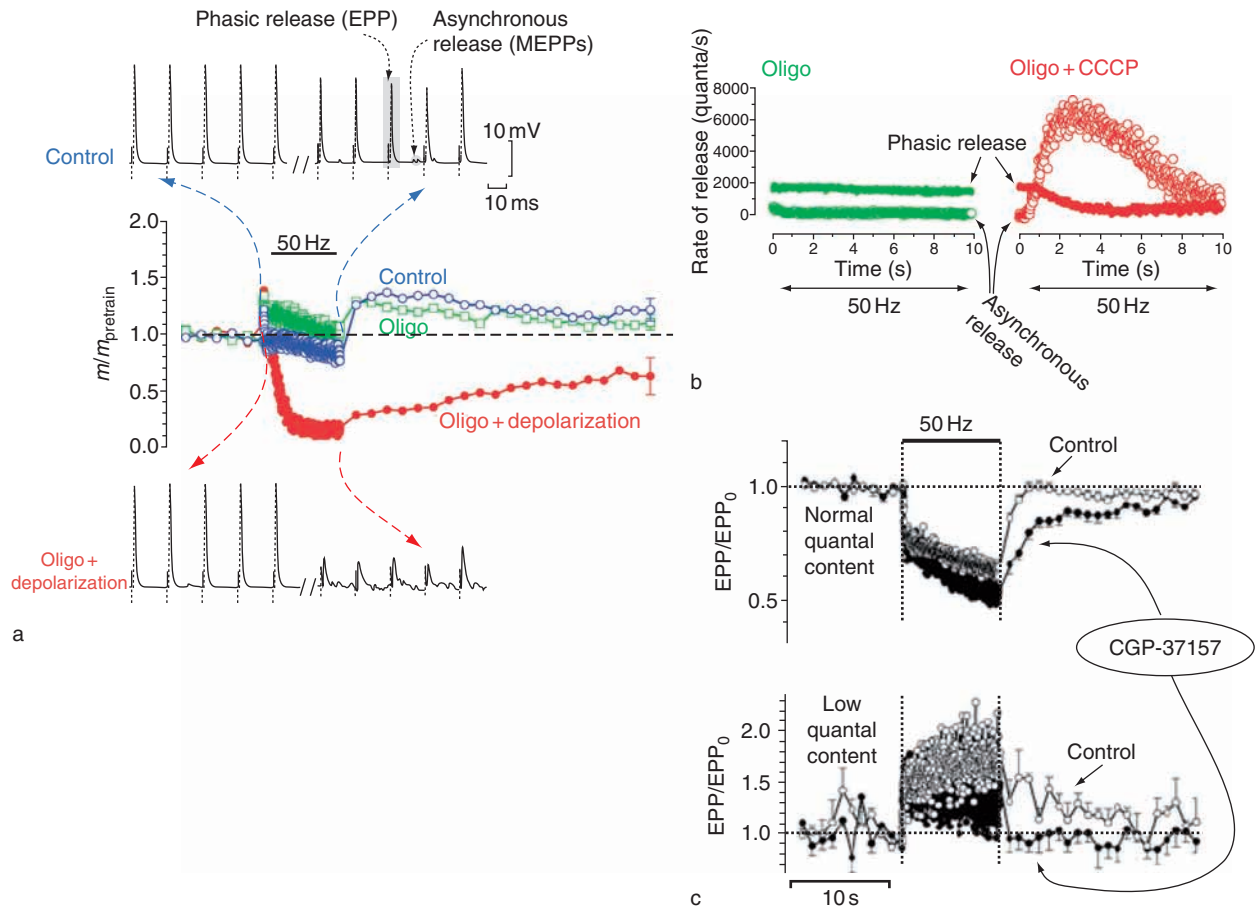

Figure 4 Mitochondrial Ca^{2+} uptake and subsequent extrusion help to sustain phasic transmitter release during and following tetanic stimulation: (a) five sequential endplate potentials (EPPs) recorded at the beginning and end of the stimulus train before (upper traces) and after (lower traces) $\Delta\Psi_m$ depolarization and the quantal content (*m*, central plot) normalized to the value recorded at the onset of stimulation ($m_{pretrain}$); (b) comparison of the rates of phasic and asynchronous release during stimulus trains; (c) results of tetanic stimulation in normal extracellular [Ca^{2+}] and thus normal quantal content (top) and results of a similar experiment performed when extracellular [Ca^{2+}] was reduced from 1.8 to 0.4 mM (bottom). In (a) the depression of phasic transmitter release during a 50 Hz stimulus train is greater when mitochondrial Ca^{2+} uptake is blocked by depolarizing $\Delta\Psi_m$ (depolarization, red). Phasic release was measured as the peak amplitude of the EPP recorded by a microelectrode inserted into the muscle fiber. In the central plot *m* is calculated as EPP amplitude divided by the average miniature EPP (MEPP) amplitude. The tetanic and posttetanic depression of *m* were not increased when mitochondrial ATP production was blocked without $\Delta\Psi_m$ depolarization (with oligomycin (oligo), shown in green). In (b) the comparison of the rates of phasic (= EPP) and asynchronous (= MEPP) release during stimulus trains show that, when mitochondrial Ca^{2+} uptake is inhibited by depolarizing $\Delta\Psi_m$ (red), the decrease in phasic release is accompanied by a dramatic increase in the rate of asynchronous release, driven by the large increase in cytosolic [Ca^{2+}] recorded in this condition (**Figure 2(c)**). Calculations (not shown) indicate that the increased asynchronous release at the beginning of the train is sufficient to deplete the pool of releasable vesicles. Similar experiments performed using oligomycin alone (green) show sustained phasic release and little asynchronous release, demonstrating that the dramatic increase in asynchronous release during $\Delta\Psi_m$ depolarization was not due to inhibiting mitochondrial ATP production. As shown in (c), following tetanic stimulation in normal extracellular [Ca^{2+}] and thus normal quantal content, the EPP amplitude returns rapidly to control levels (top). This recovery is slowed when mitochondrial Ca^{2+} extrusion is inhibited by CGP-37157. When extracellular [Ca^{2+}] is reduced from 1.8 to 0.4 mM (bottom), posttetanic release is temporarily increased but CGP-37157 blocks this increase. EPP amplitudes are normalized to prestimulation values (EPP_0) and were recorded from endplates in mouse internal oblique (a, b) and levator auris longus (c) muscles; contractions were blocked using μ-conotoxin GIIIB, which blocks muscle, but not axonal, Na^+ channels. The $\Delta\Psi_m$ depolarization in (a) was achieved with either the proton carrier CCCP or the complex III blocker antimycin A1. $\Delta\Psi_m$ depolarizing agents were applied with oligomycin to prevent the rebuilding of $\Delta\Psi_m$ by reverse operation of the F_1,F_0-ATPase. Ψ_m, mitochondrial membrane potential; ATP, adenosine triphosphate. (a, b) Reproduced from David G and Barrett EF (2003) Mitochondrial Ca^{2+} uptake prevents desynchronization of quantal release and minimizes depletion during repetitive stimulation of mouse motor nerve terminals. *Journal of Physiology* 548: 425–438, with permission from Wiley-Blackwell Publishing Ltd. (c) Reproduced from Garcia-Chacon LE, Nguyen KT, David G, and Barrett EF (2006) Extrusion of Ca^{2+} from mouse motor terminal mitochondria via a Na^+/Ca^{2+} exchanger increases post-tetanic evoked release. *Journal of Physiology* 574: 663–675, with permission from Wiley-Blackwell Publishing Ltd.

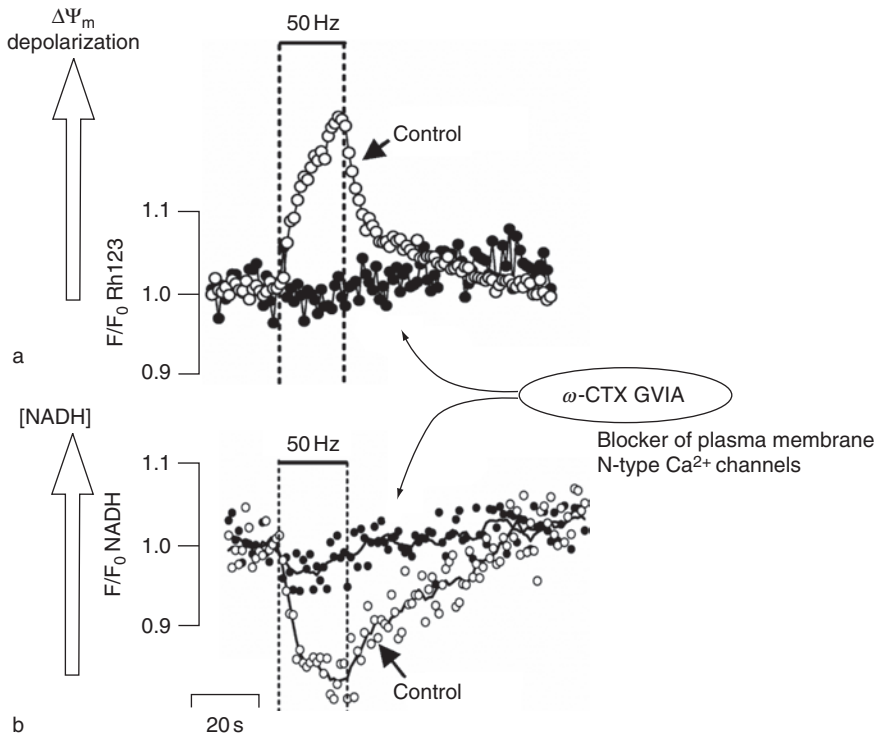

Figure 5 Stimulation-induced changes requiring Ca^{2+} influx into motor nerve terminals: (a) changes in mitochondrial membrane potential (Ψ_m); (b) changes in NADH fluorescence. Ψ_m depolarization is indicated by the increased fluorescence of rhodamine [Rh]-123. Both Ψ_m and NADH changes were blocked by ω-conotoxin (CTX) GVIA, which blocks the voltage-dependent Ca^{2+} channels (N-type) in the plasma membrane of these lizard motor terminals. Both plots are normalized to pretrain fluorescence (F_0). Recordings are from motor terminals innervating lizard intercostal muscles. NADH, nicotinamide adenine dinucleotide. Reproduced from Talbot JD, Barrett JN, Barrett EF, and David G (2007) Stimulation-induced changes in NADH fluorescence and mitochondrial membrane potential in lizard motor nerve terminals. *Journal of Physiology* 579(3): 783–798, with permission from Wiley-Blackwell Publishing Ltd.

acceleration of ETC activity, which oxidizes NADH. Studies in isolated mitochondria demonstrate that $\Delta\Psi_m$ depolarization accelerates ETC activity by reducing the magnitude of the energy barrier that opposes proton extrusion. One possible cause of the measured $\Delta\Psi_m$ depolarization is increased H^+ influx into the mitochondria via the F_1F_0-ATPase when the cytosolic ATP/ADP ratio decreases (a mechanism called respiratory control). However, this mechanism seems unlikely during short stimulation trains because blocking F_1F_0-ATPase activity with oligomycin did not inhibit the recorded changes in [NADH] or $\Delta\Psi_m$. Another possible cause of $\Delta\Psi_m$ depolarization is the direct Ca^{2+} influx into mitochondria. This mechanism seems likely because of the Ca^{2+} dependence already mentioned and because the measured $\Delta\Psi_m$ depolarization has a time course similar to that expected for Ca^{2+} influx into the mitochondria, occurring during stimulation and terminating quickly thereafter. Thus an overall hypothesis concerning how these changes are interrelated (see Figure 1) is that Ca^{2+} influx depolarizes $\Delta\Psi_m$, leading to easier mitochondrial proton extrusion and, thus, accelerating ETC activity and the oxidation of NADH to NAD^+. The ability of the ETC to accelerate during Ca^{2+} influx limits the magnitude of the $\Delta\Psi_m$ depolarization. When motor terminals are stimulated in the presence of ETC blockers, the depolarization of $\Delta\Psi_m$ is greater and mitochondrial Ca^{2+} uptake becomes less efficient, as indicated by a much larger elevation in cytosolic $[Ca^{2+}]$. Mitochondria in repetitively stimulated motor nerve terminals thus appear to achieve a compromise between their ATP-producing and Ca^{2+}-buffering roles. Both roles require a hyperpolarized $\Delta\Psi_m$. The necessity for Ca^{2+} uptake produces $\Delta\Psi_m$ depolarization (and thus a temporary decrease in ATP synthesis), but this depolarization is kept relatively small due to their ability to accelerate ETC activity. Motor nerve terminals thus appear to tolerate a temporary decrease in mitochondrial ATP production (due perhaps to other energy reserves such as glycolysis) in order to avoid an excessive elevation of cytosolic

[Ca^{2+}] that would rapidly disrupt their main physiological function (i.e., maintaining phasic transmitter release at levels that exceed the threshold for activating muscle contraction).

Other Tissues

The impact of the described Ca^{2+} influx into the mitochondria on [NADH] may be different in other tissues. Measurements of stimulation-induced changes in [NADH] in other tissues (including the CNS) usually show that the most prominent change is not the brief decrease we have described but, rather, a pronounced increase that persists long after stimulation. Such a mechanism would accelerate ATP synthesis to match the increased demand signaled by matrix [Ca^{2+}] elevation. Only a fraction of motor terminals exhibited any poststimulation increase in NADH, and the magnitude of any such increase was small. One way to understand these differences is by considering that the direction of change in [NADH] when the mitochondria take up Ca^{2+} depends on the relative contribution of the two outcomes of this uptake: $\Delta\Psi_m$ depolarization due to Ca^{2+} current into the matrix and the elevation of matrix [Ca^{2+}].

1. $\Delta\Psi_m$ depolarization pulls the ETC toward expending more NADH to accelerate H^+ extrusion. This immediate response allows the mitochondrial Ca^{2+} influx to proceed without progressive $\Delta\Psi_m$ depolarization and seems to dominate in motor terminals.
2. The elevation of matrix [Ca^{2+}] pushes the ETC by stimulating NADH supply via the Krebs cycle. Here, matrix [Ca^{2+}] serves as a signal for the anticipated increase in cytosolic ATP demand which requires NADH-supported H^+ extrusion. This mechanism may dominate in tissues in which mitochondria are not the main mechanism for buffering Ca^{2+} loads.

Mitochondrial Involvement in Nerve Terminal Pathology

Ischemia/Reperfusion

Acute limb ischemia in humans originates from a sudden decrease in arterial blood flow to an extremity, commonly caused by peripheral arterial occlusive disease (PAD), leading to intermittent and repeated episodes of limb ischemia during walking or exercise. Muscle denervation is one of the earliest manifestations of PAD. Animal models of hindlimb ischemia/reperfusion (I/R) injury demonstrate that motor nerve terminals are the most vulnerable neuromuscular component. In the first 4 weeks following an episode of ischemia experimentally applied to rat hindlimbs, morphological studies detected no changes in axons, Schwann cells, or muscle fibers but multiple signs of degeneration in motor nerve terminals. These signs included the disruption of presynaptic membranes and the swelling and degeneration of the mitochondria. Functional measurements also show that the neuromuscular junction is a major site of I/R injury because in postischemic muscle contractions elicited by the activation of axons and their motor terminals recover less than contractions elicited by direct muscle stimulation. Mechanisms for ischemic injury to motor nerve terminals include mitochondrial Ca overload during ischemia and the subsequent oxidative stress during reperfusion, which might collaborate to induce the irreversible opening of the MPTP. The heavy reliance of motor nerve terminals on the mitochondria for Ca^{2+} sequestration (see the section entitled 'Mitochondria as regulators of cytosolic [Ca^{2+}] during repetitive stimulation') might predispose them to degeneration after stresses (e.g., I/R) that damage this organelle.

Amyotrophic Lateral Sclerosis

This adult-onset disease is characterized by the death of upper and lower spinal motor neurons, especially those innervating the limbs. Most cases of amyotrophic lateral sclerosis (ALS) arise sporadically, but approximately 10% are familial. Some familial cases involve mutations in superoxide dismutase I (SOD1), an enzyme that contributes to defense against oxidative stress. These SOD1 mutations produce disease by a dominant gain-of-function mechanism that appears unrelated to whether the mutant SOD1 retains dismutase activity. It is as yet unclear why motor neurons are differentially vulnerable to the mutant protein, which is expressed in all body cells. Mice that transgenically express some of these ALS-inducing mutant human SOD1s develop motor neuron death and thus constitute animal models for familial ALS. Work from multiple laboratories showed that in these mice the first structural signs of motor neuron damage are found in motor nerve terminals and their mitochondria, changes that occur months before the motor neuron somata begin to die. An early functional change in the motor neuron mitochondria is reduced ability to sequester Ca^{2+} loads, which might contribute to the early degeneration of motor terminals. Motor terminals that innervate fast limb muscles degenerate earlier than those innervating slow limb muscles, and these fast motor terminals exhibit increased vulnerability to I/R injury at an early age. Future studies will clarify the link among mutant SOD1, mitochondria, and motor terminal damage and will perhaps suggest strategies for treating this fatal disease.

See also: Adenosine Triphosphate (ATP); Amyotrophic Lateral Sclerosis (ALS); Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Neuromuscular Junction: Synapse Elimination; Neuromuscular Junction (NMJ): Presynaptic Stretch Effects on Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission.

Further Reading

Alnaes E and Rahamimoff R (1975) On the role of mitochondria in transmitter release from motor nerve terminals. *Journal of Physiology* 248: 285–306.

Bacman SR, Bradley WG, and Moraes CT (2006) Mitochondrial involvement in amyotrophic lateral sclerosis: trigger or target? *Molecular Neurobiology* 33: 113–131.

De Graaf F, Van Raamsdonk W, Van Asselt E, and Diegenbach PC (1991) Histochemical profiles of motoneurons innervating muscle fibres with different activity patterns in the zebrafish, *Brachydanio rerio*. *Histochemistry Journal* 23: 273–280.

David G (1999) Mitochondrial clearance of cytosolic Ca^{2+} in stimulated lizard motor nerve terminals proceeds without progressive elevation of mitochondrial matrix [Ca^{2+}]. *Journal of Neuroscience* 19: 7495–7506.

David G and Barrett EF (2003) Mitochondrial Ca^{2+} uptake prevents desynchronization of quantal release and minimizes depletion during repetitive stimulation of mouse motor nerve terminals. *Journal of Physiology* 548: 425–438.

David G, Barrett JN, and Barrett EF (1998) Evidence that mitochondria buffer physiological Ca^{2+} loads in lizard motor nerve terminals. *Journal of Physiology* 509: 59–65.

Duchen MR (1992) Ca^{2+}-dependent changes in the mitochondrial energetics in single dissociated mouse sensory neurons. *Biochemistry Journal* 283(1): 41–50.

Fiskum G, Reynafarje B, and Lehninger AL (1979) The electric charge stoichiometry of respiration-dependent Ca^{2+} uptake by mitochondria. *Journal of Biological Chemistry* 254: 6288–6295.

Garcia-Chacon LE, Nguyen KT, David G, and Barrett EF (2006) Extrusion of Ca^{2+} from mouse motor terminal mitochondria via a Na^{+}/Ca^{2+} exchanger increases post-tetanic evoked release. *Journal of Physiology* 574: 663–675.

Hollenbeck PJ and Saxton WM (2005) The axonal transport of mitochondria (review). *Journal of Cell Science* 118: 5411–5419.

Nakatani T, Nakashima T, Kita T, and Ishihara A (2003) Cell size and oxidative enzyme activity of type-identified fibers in rat hindlimb muscles: A review. *Acta Histochemica et Cytochemica* 36: 105–114.

Nicholls DG and Chalmers S (2004) The integration of mitochondrial calcium transport and storage (review). *Journal of Bioenergetics and Biomembranes* 36: 277–281.

Nicholls DG and Ferguson SJ (2001) *Bioenergetics*, 3rd edn. London: Academic Press.

Tang Y and Zucker RS (1997) Mitochondrial involvement in post-tetanic potentiation of synaptic transmission. *Neuron* 18: 483–491.

Werth JL and Thayer SA (1994) Mitochondria buffer physiological calcium loads in cultured rat dorsal root ganglion neurons. *Journal of Neuroscience* 14: 348–356.

Zucker RS and Regehr WG (2002) Short-term synaptic plasticity (review). *Annual Review of Physiology* 64: 355–405.

Primate Communication: Evolution

A A Ghazanfar, Princeton University, Princeton, NJ, USA
Y E Cohen, Dartmouth College, Hanover, NH, USA

© 2009 Elsevier Ltd. All rights reserved.

Vocal Perception in Nonhuman Primates

By comparing the vocal behavior of extant primates with that of humans, one can deduce the behavioral capacities of extinct common ancestors, allowing the identification of homologies and providing clues as to the adaptive functions of such behaviors. By exploring how living nonhuman primate (hereafter, 'primate') species use their vocalizations, numerous investigators are building a rigorous, testable framework for how speech might have evolved. Much of this work has been carried out in the domain of vocal perception (as opposed to production). There are numerous psychophysical investigations of how primates, such as macaques, baboons, vervets, and chimpanzees, process human speech sounds, best for the most part, these studies show that with the appropriate amount of training (usually quite extensive), these species perceive speech sounds in a manner similar to that of humans. A more insightful line of inquiry includes research into how primates perceive their own vocalizations and whether the underlying patterns and processes are similar to those of human speech perception. This article highlights the insights made by this line of inquiry.

Referential Communication

The species-typical vocalizations of many primates, as well as other animals, contain information that can be used by listeners as a source of information about objects, events, and the status of peers in their environment. Of particular interest is the capacity of primate vocalizations to transmit functional ('referential') information about an event or the presence of an object in the environment. That is, based on signal morphology alone, and regardless of the mechanisms that might underlie vocalization production, listeners are able to extract information about a vocalization's putative meaning.

A classic example of referential communication is the use of predator-alarm calls by vervet monkeys (*Cercopithecus aethiops*). Vervets produce unique alarm calls for three different predators: a snake, a leopard, and an eagle. The information that is transmitted to the listeners initiates specific patterns of species-typical behaviors. For example, when vervets hear an eagle-alarm call, they scan the sky for visual cues of the airborne predator, and, in some cases, run to locations that provide overhead coverage. In contrast, when they hear a snake-alarm call, they stand up and scan the ground. Finally, a leopard-alarm call initiates a third distinct behavior: vervets run up the nearest tree while scanning the horizon for the leopard.

The capacity to successfully process referential signals also allows animals to use the referential information that is transmitted by the vocalizations of other species. For example, female Diana monkeys (*Cercopithecus diana*) elicit a predator-alarm call when they hear a male Diana monkey producing a leopard alarm or when they hear the leopard-alarm call of a crested guinea fowl. This observation is important since it suggests that Diana monkeys can form abstract categorical representations of a vocalization's functional meaning that is independent of acoustics and the species generating the signal.

Another example of categorization is the food-associated calls of rhesus macaques (*Macaca mulatta*). When free-ranging rhesus monkeys encounter low-quality food, they elicit coos or grunts. In contrast, when they encounter rare, high-quality food, they elicit harmonic arches or warbles. Both free-ranging and laboratory-housed rhesus monkeys categorize these vocalizations based on differences in their function as opposed to their acoustics. For example, harmonic arches and warbles, despite having distinct acoustic properties, are categorized as being functionally equivalent.

Syntactic Processing of Vocal Sequences

Many primate species produce bouts of vocalizations containing sequences of similar acoustic units and/or different-sounding acoustic units, but we know very little about whether each of these units separately codes meaningful information (like words in a sentence) or whether units can be combined to form only a single meaningful utterance (like syllables in a word). For our purposes here, we will refer to all temporally distinct acoustic pulses as syllables. This criterion is arbitrary in the sense that it is based solely on an acoustic measure, not a behavioral one.

Most of our insights into the units of perception come from studies of the orderly arrangement of syllables within primate 'long calls'. Long calls serve as localization cues for conspecifics and are often produced in the context of territorial encounters, mate attraction, and isolation/group cohesion. These multisyllabic calls are produced by a range of different

primate species, such as the 'whoop–gobble' of mangabeys (*Cercocebus albigena*), the 'chirrup–pant–bellow–pump–honk' call sequences of titi monkeys (*Callicebus moloch*), the 'chirp–trills–squaws' of wedge-capped capuchins (*Cebus olivaceus*), the 'twitters' of squirrel monkeys (*Saimiri sciureus*), the 'whinny' of spider monkeys (*Ateles geoffroyi*), the 'pant–hoot' of chimpanzees (*Pan troglodytes*), the 'songs' of gibbons (*Hylobates* spp.), and the combination long calls of cotton-top tamarins (*Saguinus oedipus*).

In these vocal signals, which consist of multiple syllables, both individual syllables and combinations of syllables may mediate specific aspects of behavior. As a result, for any single vocalization, information may be coded at multiple levels. For example, the chimpanzee 'pant–hoot' consists of a series of 'hoot' calls followed by a series of 'screams.' Since both hoots and screams are produced individually in other contexts, the pant–hoot could either be a single vocalization or a bout of several vocalizations.

To determine the relative importance of single syllables versus combinations of syllables in eliciting behaviors from listeners, one can use playback experiments. For example, early studies of the long calls of titi monkeys and gibbons demonstrated that individual renditions of these calls have stereotypical arrangements of syllables; individual males use these long calls to defend/demarcate their territories. Playback experiments, in which these syllables were artificially rearranged, revealed that both primate species recognized the novel arrangement of syllables and responded as though a new male was in the adjacent territory. Gibbons produced significantly more 'squeak' calls (given during intergroup encounters) when hearing these novel stimuli, whereas titi monkeys produced significantly more 'moaning' responses (also given in response to interspecies and intergroup encounters). These data suggest that, at least in these species, there is recognition of the global order of syllable sequences.

It is also important to understand the smallest invariable acoustic unit that can influence behavior. Cotton-top tamarins proved to be a useful model system for exploring this issue. In response to social isolation, tamarins will emit a long call beginning with one to two chirp syllables and ending with two to five whistle syllables (**Figure 1**). Upon hearing such vocalizations, conspecifics often respond with their own long calls – a behavior called 'antiphonal calling.' Although tamarins can produce chirps and whistles independently, it was hypothesized that the combination of both syllable types would function as a more effective stimulus (i.e., the unit of perception) for eliciting antiphonal long calls than either unit alone. Playback experiments using whole long calls (i.e., both chirps and whistles) versus isolated chirps versus isolated whistles confirmed this hypothesis: whole calls were much more effective in eliciting antiphonal long calls than either single whistles or single chirps. These data suggest that, in this species, the whole call is the most meaningful unit from the perspective of socially isolated receivers.

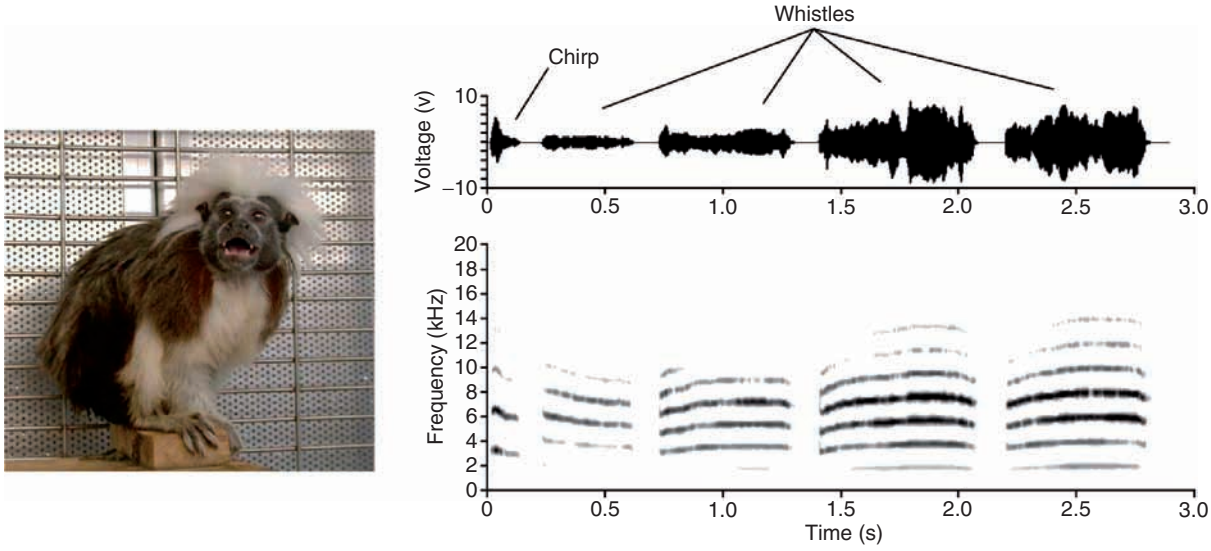

Figure 1 Cotton-top tamarin combination long call. Time–amplitude waveform (*y*-axis represents voltage in volts) and spectrogram (*y*-axis represents frequency in kilohertz) of a representative combination long call recorded in captivity. Chirps and whistles are labeled. Adapted from Ghazanfar AA, Flombaum JI, Miller CT, et al. (2001) The units of perception in the antiphonal calling behavior of cotton-top tamarins (*Saguinus oedipus*): Playback experiments with long calls. *Journal of Comparative Physiology A* 187: 27–35.

The data described here represent examples of 'phonological syntax,' whereby smaller units are assembled to form larger ones. Are there any examples of 'lexical syntax,' where the different combinations of sounds produce different 'meanings'? Studies of African forest monkeys provide a tentative 'yes' to this question. There are two such examples. First, Diana monkeys seem to comprehend not only the two different alarm calls of a sympatric species (the leopard- and eagle-alarm calls of Campbell's monkeys (*Cercopithecus campbelli*)) but they also seem to understand that if these alarm calls are preceded by a Campbell's monkey's 'boom' call, the threat is less urgent. Importantly, if a Campbell's monkey's boom is placed before a Diana monkey's own alarm calls, it has no effect on the Diana monkeys' behavior. The second example is from the putty-nosed monkey (*Cercopithecus nictitans*). These monkeys also produce two different alarm calls, one for leopards ('pyows') and one for eagles ('hacks'). When heard, each of these calls produces a stereotypical escape response from listeners. However, when the two calls are combined into 'pyow–hack' sequences by the males of a group, this sequence of calls elicits general group movement, not an alarm response. Thus, both of these examples suggest that at least some primates comprehend lexical syntax.

Multisensory Processes

One aspect of speech that is overlooked by comparative biologists seeking to understand its origins is the fact that speech is not a purely auditory phenomenon. Human speech is a multisensory function, perceived through both the visual and auditory channels. Multisensory speech perception is evident even at the earliest stages of human cognitive development (at 2 months of age), its integration across the two modalities is ubiquitous and automatic, and even at the neural level, audiovisual speech integration is evident at the 'earliest' stages of cortical processing. One example of this integration is the integration of speech signals and speech-related facial motion. For example, human adults link high-pitched sounds to facial postures producing an /i/ sound and link low-pitched sounds to faces producing an /a/ sound. Thus, multisensory speech may be the 'primary' mode of speech perception and not simply a capacity that is 'piggybacked' on to auditory speech perception. If its multisensory nature is a fundamental feature of human speech, whereby visual/facial and vocal signals are inextricably linked, then do other primates show similar links in their vocalizations?

Data from a variety of primate species suggests that there is a similar link between acoustic output and facial dynamics. For example, different rhesus monkey vocalizations are produced with unique lip configurations and mandibular positions and the motion of such articulators influences the acoustics of the signal. Coo calls, like /u/ in speech, are produced with the lips protruded, while screams, like the /i/ in speech, are produced with the lips retracted (Figure 2). Moreover, as in humans, the articulation of these expressions has visible consequences on facial motion beyond the oral region. Grimaces, produced during scream vocalizations, for instance, cause the skin folds around the eyes to increase in number. Chimpanzees, though less studied in the domain of multisensory communication, also link

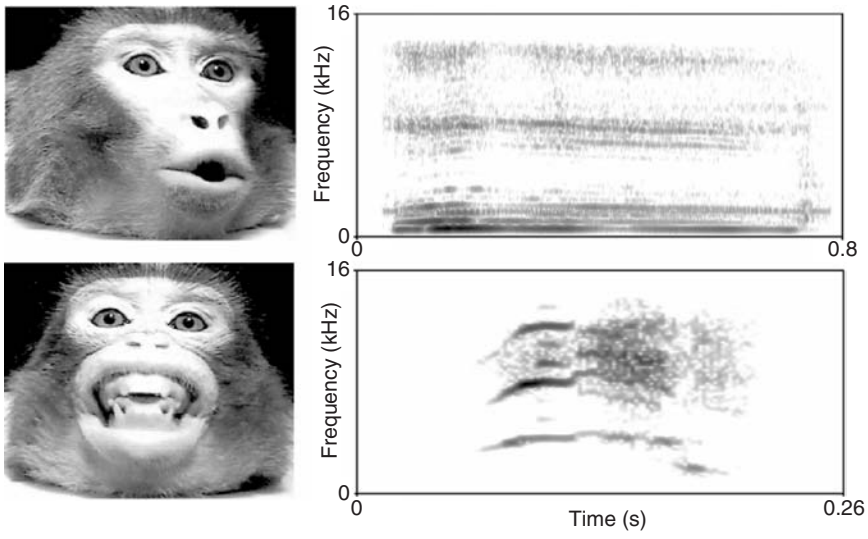

Figure 2 Exemplars of rhesus monkey facial expressions produced concomitantly with vocalizations. Top panels represent a coo call taken at the midpoint of the expression and the corresponding spectrogram. Bottom panel represents a scream call.

facial expressions with vocal acoustics. Video and audio analyses of common chimpanzee vocalizations produced in the wild show that a decline in the fundamental frequency (F0) occured when submissive screams transitioned into aggressive barks. These changes in F0 were correlated with changes in visible articulators such as lip and teeth opening. Taken together, these data suggest that it is likely that many of the facial motion cues that humans use for speech reading are present in at least some apes and monkeys as well.

Given that both humans and other extant primates use both facial and vocal expressions as communication signals, it is unlikely that only humans recognize the correspondence between the visual and auditory components of vocal signals. In behavioral situations similar to those used to study the face/voice-matching abilities of human infants, rhesus monkeys demonstrated that they recognize the auditory–visual correspondences between their 'coo' and 'threat' calls. A similar study using capuchin monkeys (*Cebus apella*) as subjects showed that they could match the appropriate facial expressions to heard vocalizations as well. Chimpanzees can also match familiar voices to familiar faces as well as facial expressions with their corresponding vocalizations (albeit using a training-induced behavioral measure).

Temporal Cues for Vocal Recognition

Humans use temporal cues such as the duration, interval, and order of acoustic features to distinguish among speech sounds. For example, humans distinguish /pa/ from /ba/ on the basis of voice onset time, and /sa/ from /sta/ on the basis of the silent time between consonants and vowels. It has been suggested that speech perception is based on the rapid processing of temporal information.

To what extent do primates use temporal cues to recognize their own vocalizations? Field and lab studies of the coo calls of Japanese macaques (*Macaca fuscata*) suggested that they could distinguish among different coo calls based upon temporal cues (the temporal position of a peak frequencies), and calls defined by this feature were produced in different behavioral contexts. Later studies, however, showed that although they could use temporal cues to discriminate among their own coo calls, this may have been an artificial classification scheme inadvertently imposed by the experimenters.

Studies from two other species provide additional support for the notion that primates use temporal cues for vocal recognition. First, as already mentioned, cotton-top tamarins antiphonally call preferentially to whole calls, compared to signals composed of one of the two syllable types present in their long calls. Further studies revealed that while tamarins did not distinguish between normal calls and time-reversed or pitch-shifted long calls, normal response rates did require the species-typical temporal structure of the amplitude envelope. Furthermore, there was suggestive evidence that the number of syllables and the syllable rate may also influence antiphonal calling responses. Second, in rhesus monkeys, playback experiments using three different call types – shrill barks, grunts, and copulation calls – demonstrated that rhesus monkeys responded differently to shrill barks and grunts than to copulation calls when the intersyllable interval was expanded or contracted beyond the species-typical range. In another experiment, rhesus monkeys' sensitivity to the shape of the amplitude envelope in the shrill barks and harmonic arch calls was tested by using time-reversed calls. This manipulation changes the calls' temporal structure without affecting their spectral content. Rhesus monkeys responded to these calls in a fashion suggesting that they did not recognize these stimuli as species-typical vocalizations.

Overall, the referential, syntactic, multisensory, and temporal features of primate vocalizations, from many different species, suggest striking parallels with human speech processing. Are these processes mediated by homologous neural circuits?

Neural Bases of Primate Vocal Communication

Hemispheric Asymmetries

A classic feature of human speech processing is its lateralization. Although there is no clear dichotomy of function between the cerebral hemispheres, perceptual experiments, studies of brain-damaged patients, and functional imaging studies have indicated that speech perception is usually lateralized to the left temporal lobe (**Figure 3(a)**). A behavioral indication of left-hemispheric specialization for speech processing in intact humans is the performance advantage exhibited by the right ear for the identification of speech sounds, and the left-ear advantage for nonspeech sounds. Behavioral experiments under laboratory and field conditions reveal that primates also exhibit similar asymmetries in the perception of their vocalizations. Psychophysical data from macaques also revealed a right-ear advantage when discriminating conspecific calls, and the orienting asymmetry assay alluded to in the previous section, whereby rhesus monkeys orient with the right ear leading toward conspecific calls, also suggests a lateralization of perceptual processes toward the left cerebral hemisphere. Similar right-ear orienting

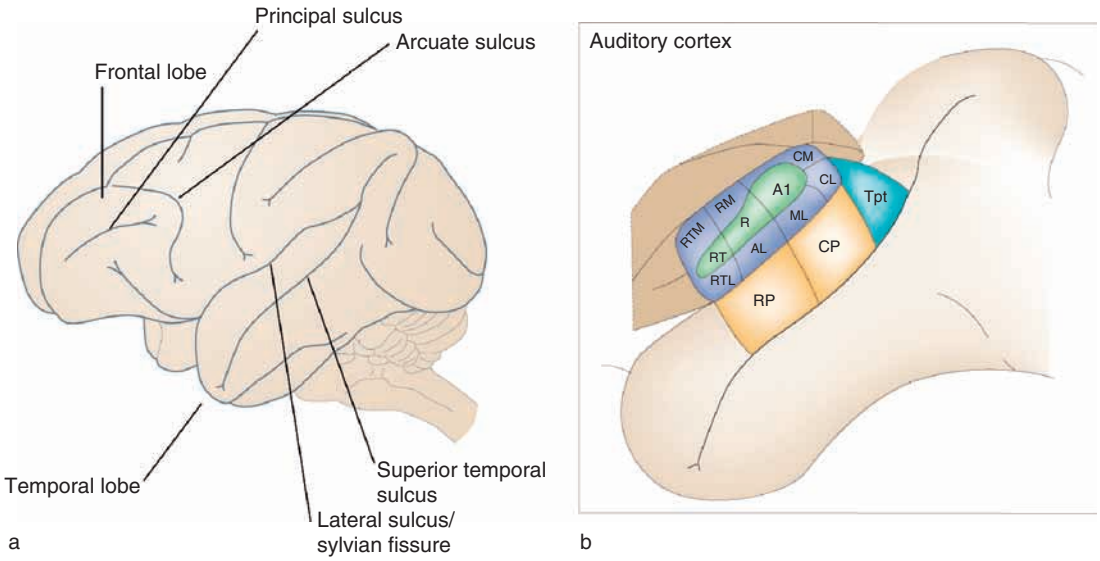

Figure 3 (a) The macaque cerebral cortex. (b) Auditory areas within and around the lateral sulcus. The core regions (A1, R, and RT) are thought to be 'primary-like' areas, responding best to pure tones, while the surrounding lateral belt areas (CL, ML, and AL) are more responsive to complex sounds. The response properties of other areas have not been systematically investigated. A1, auditory area 1; AL, anterior lateral belt; CL, caudolateral belt; CM, caudomedial belt; CP, caudal parabelt; ML, middle lateral belt; R, rostral area; RM, rostromedial belt; RT, rostrotemporal area; RP, rostral parabelt; RTL, rostrotemporolateral belt; RTM, rostrotemporomedial belt; Tpt, temporoparietal area. Adapted from Ghazanfar AA and Santos LR (2004) Primate brains in the wild: The sensory bases for social interactions. *Nature Reviews Neuroscience* 5: 603–616.

biases have been reported for chimpanzees and mouse lemurs (*Microcebus murinus*).

More direct evidence for hemispheric asymmetries in processing vocalizations comes from anatomical, lesion, and imaging studies. For humans that show functional left-hemispheric biases for language processing, it has been shown that the Sylvian fissure (bordering auditory cortex) is significantly longer in the left hemisphere than in the right (Figure 3(a)). Using this measurement, potential anatomical asymmetries have similarly been measured in several species of primates. Left Sylvian fissure length was found to be significantly greater than in the right hemisphere in apes (*Pan, Gorilla*, and *Pongo*), Old World macaques (*M. fascicularis* and *M. mulatta*), and New World Callichtrids (cotton-top tamarins and marmosets). These cerebral asymmetries might represent the 'specialized' neural circuitry that mediates the behavioral asymmetries to conspecific vocalizations.

Lesion and imaging studies support this notion. Following lesions of the left auditory cortex, Japanese macaques exhibit a selective impairment for discriminating species-specific vocalizations but not for other types of auditory stimuli. Their performance was unimpaired by similar lesions in the right hemisphere. In rhesus monkeys, a positron emission tomography (PET) imaging study showed that conspecific vocalizations activate the anteriormost portion of the left upper temporal lobe to a greater degree than do other sounds. These studies naturally lead one to wonder how the neurophysiological signals of the neocortex encode vocalizations.

Two key neocortical nodes involved in processing human speech (both auditory and auditory–visual) are the superior temporal gyrus, within which lies the auditory cortex, and the prefrontal cortex. The human imaging data consistently show that the auditory cortex and prefrontal cortex seem to play important roles in speech perception. These two regions are, furthermore, densely interconnected. Neurophysiological evidence suggests that these areas play a similar role in monkeys; this observation is expanded upon in the following sections.

Auditory Cortex

Despite the importance of vocal communication in the lives of primates, we lack a complete understanding of how biologically relevant features of complex sounds are processed at the level of single neurons or small populations of neurons. For many years, the squirrel monkey was the only primate model for investigating the role of auditory cortex in processing species-specific vocalizations. Studies using this species found that many cells in the superior temporal gyrus responded to species-specific vocalizations. However, one of the drawbacks of these studies was that recordings were made across the superior

temporal gyrus without reference to any cytoarchitectural or any other anatomical subdivisions.

Recent anatomical and neurophysiological experiments in rhesus monkeys have, for the most part, eliminated this confound. A series of experiments in anesthetized rhesus monkeys have delineated three cochleotopic areas along the superior temporal gyrus (Figure 3(b)). They were labeled the caudolateral, middle lateral, and anterolateral belt areas (CL, ML, and AL, respectively). Unlike the primary auditory cortex (A1), neurons in these areas responded best to complex sounds, such as band-passed noise and species-specific vocalizations. In fact, neurons in all three lateral belt areas preferred rhesus monkey calls to energy-matched pure-tone stimuli. Furthermore, among the seven call exemplars tested, 90% of lateral belt neurons responded better to certain call types over others, that is, they were 'call selective.'

The lateral belt areas are not the only regions with neurons selectively responsive to species-specific vocalizations. Recent studies reveal that A1 neurons, despite their exquisite pure-tone sensitivity, can also be selective to complex time-varying stimuli as well. In common marmosets (*Callithrix jacchus*) and squirrel monkeys, A1 neurons can even be selective for conspecific calls. In both species, A1 neuronal firing patterns appear to phase-lock to the temporal envelope of the species-specific vocalizations used as stimuli. This phase locking appears to be driven by the global syllabic structure of the vocalizations rather than by the finer acoustic features of the vocalizations. Also, many neurons in the primary auditory cortex and the surrounding belt and parabelt areas of the marmoset are sensitive to the pitch of an auditory stimulus; this information may be used to infer the affective content of a vocalization. Based on these studies of the marmoset and squirrel monkey auditory cortex, it appears that even at early stages of cortical processing, there is selectivity for species-specific communication sounds that may be based on experience-dependent and/or species-specific neural mechanisms.

Recently, auditory cortex has also been shown to integrate multisensory communication signals. Local field potentials and single units in the auditory cortex are capable of integrating facial and vocal signals in monkeys (Figure 4). The vast majority of responses were specific to 'face + voice' integration, and such integration could take the form of either enhancement or suppression. The specificity of these integrative

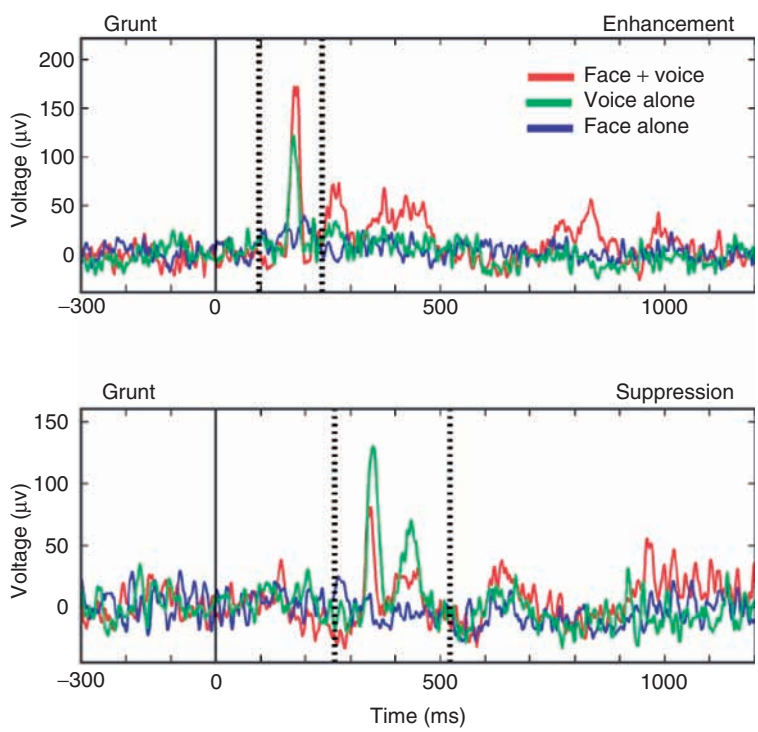

Figure 4 Auditory cortical responses to multimodal vocalizations from the lateral belt region. Rectified local field potential responses to 'face + voice' (F + V), voice-alone (V), and face-alone (F) components were compared. The solid vertical line indicates the onset of the 'face' signal. Dotted vertical lines indicate the onset and offset of the 'voice' signal. Graphs represent the mean of ten repetitions, with the mean baseline activity subtracted on a trial-by-trial basis. Top panel shows an enhanced response to a multisensory grunt vocalization, while the bottom panel shows a suppressed response to a different grunt vocalization. Adapted from Ghazanfar AA, Maier JX, Hoffman KL, et al. (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience* 25: 5004–5012.

responses was tested by replacing the dynamic faces with dynamic disks which mimicked the aperture and displacement of the mouth. In human psychophysical experiments, such artificial dynamic stimuli can still lead to enhanced speech detection in noise, but not to the same degree as a real face. When cortical sites were tested with dynamic disks, far less integration was seen when compared to the real monkey faces.

Prefrontal Cortex

The frontal lobes contain an auditory responsive region that responds robustly to vocalizations. This region, the ventrolateral prefrontal cortex (vPFC), lies below the principal sulcus and receives connections from the anterolateral belt region (Figure 3). The vPFC, an area thought to be part of a circuit involved in auditory–object processing, has been hypothesized to play an important role in processing the more abstract components of vocalizations. Specifically, it has been suggested that the vPFC plays an important role in processing the referential information transmitted by a vocalization.

In one set of experiments, vPFC neurons were hypothesized to be modulated more by the referential information transmitted by a vocalization than by the acoustic features of a vocalization (Figure 5). Consistent with this hypothesis, the activity of vPFC neurons was not modulated by transitions between presentations of food vocalizations that transmitted the same referential information (Figure 5(a)), even though the vocalizations had significantly different acoustic structures. vPFC activity, however, was modulatedby transitions between presentations of food vocalizations that conveyed different types of referential information (high-quality vs. low-quality food; Figure 5(b)). These data suggested that, on average, vPFC neurons

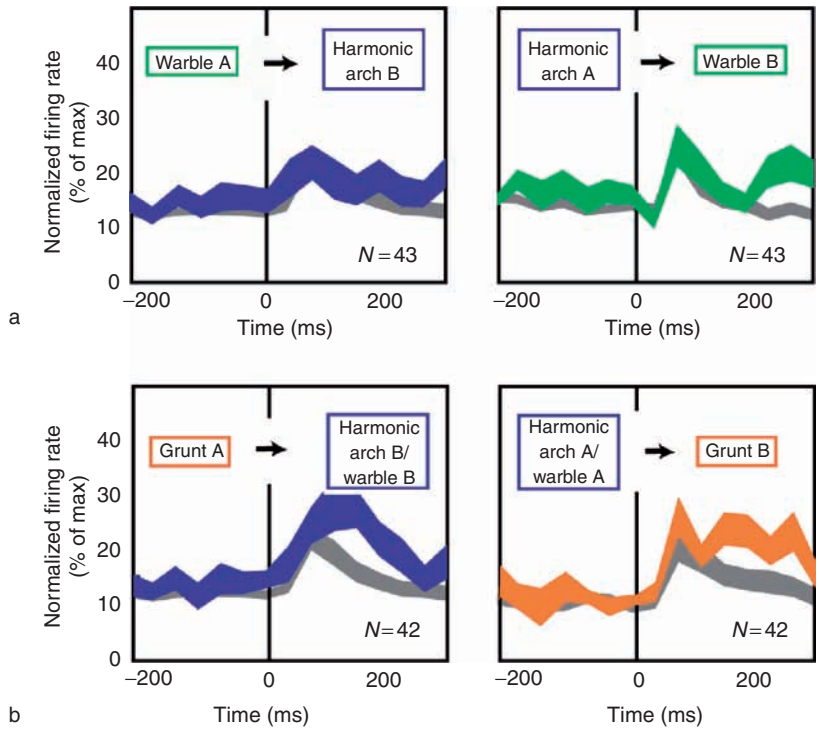

Figure 5 Population responses of vPFC activity in response to transitions between vocalization exemplars. Vocalizations were presented during the repetitive-presentation context and during a baseline condition. During the repetitive-presentation context, a vocalization was presented repeatedly (the 'repeated' vocalization) 7–9 times. This was followed by the presentation of a different vocalization (the 'test' vocalization). During the baseline condition, the test and repeated vocalizations and other stimuli were presented in a random order. The thick gray strip in each panel represents the normalized population firing rates in response to a test vocalization during the baseline condition. The colored strip in each panel represents normalized population firing rates in response to a test vocalization during the repetitive-presentation context. The thickness of the strips indicates activity ranging two standard deviations from the mean response for the population. All data are aligned relative to onset of the test vocalization; the vertical black line indicates stimulus onset. In each panel's schematic, the repeated and test vocalizations are, respectively, denoted to the left and right of the vertical black line indicating stimulus onset. (a) Normalized vPFC population activity to a test vocalization that belonged to a different acoustic class but the same functional category as the repeated vocalization. (b) Normalized vPFC population activity to a test vocalization that belonged to a different acoustic class and different functional category than the repeated vocalization. Adapted from Gifford GW, MacLean KA, Hauser MD, et al. (2005) The neurophysiology of functionally meaningful categories: Macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. *Journal of Cognitive Neuroscience* 17: 1471–1482.

are modulated preferentially by transitions between presentations of food vocalizations that belong to functionally meaningful and different categories.

Consistent with the proposed role of vPFC in categorical processing, vPFC neurons respond in the same way to different vocalizations that transmit information about different types of food quality (i.e., high-quality and low-quality food). However, vPFC neurons responded 'differently' to different vocalizations that transmitted information about different nonfood events. Importantly, this difference in response patterns cannot be attributed wholly to differences between the acoustics of the food and nonfood vocalizations: the acoustic variability in the food vocalizations was comparable to the variability in the nonfood vocalizations. vPFC neurons also respond similarly to different vocalizations that transmit information about high-quality food, low-quality food, and high-quality or low-quality food. These observations are consistent with the hypothesis that vPFC neurons respond to food and nonfood vocalizations based on their category membership but do not differentiate between categories of food quality.

Together, these two studies provide the first evidence that in the absence of operant training, the vPFC is involved in the recognition and discrimination of natural categories. Also, these studies suggest that the vPFC is part of a circuit involved in contextually dependent vocalization production. These results are also consistent with data from the human literature which posit a role for this area in the maintenance and retrieval of abstract rules and possibly a role in semantic processing and syntactic working memory.

See also: Auditory Cortex Structure and Circuitry; Contextual Interactions in Visual Perception; Language Evolution; Language: Auditory Processes; Language: Nonhuman Animals; Multisensory Convergence and Integration; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex.

Further Reading

Arnold K and Zuberbuhler K (2006) Semantic combinations in primate calls. *Nature* 441: 303.

Bieser A and Muller-Preuss P (1996) Auditory responsive cortex in the squirrel monkey: Neural responses to amplitude-modulated sounds. *Experimental Brain Research* 108: 273–284.

Cohen YE, Hauser MD, and Russ BE (2006) Spontaneous processing of abstract categorical information in the ventrolateral prefontal cortex. *Biology Letters* 2: 261–265.

Ghazanfar AA, Flombaum JI, Miller CT, et al. (2001) The units of perception in the antiphonal calling behavior of cotton-top tamarins (*Saguinus oedipus*): Playback experiments with long calls. *Journal of Comparative Physiology A* 187: 27–35.

Ghazanfar AA and Logothetis NK (2003) Facial expressions linked to monkey calls. *Nature* 423: 937–938.

Ghazanfar AA, Maier JX, Hoffman KL, et al. (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience* 25: 5004–5012.

Ghazanfar AA and Santos LR (2004) Primate brains in the wild: The sensory bases for social interactions. *Nature Reviews Neuroscience* 5: 603–616.

Gifford GW, Hauser MD, and Cohen YE (2003) Discrimination of functionally referential calls by laboratory-housed rhesus macaques: Implications for neuroethological studies. *Brain Behavior and Evolution* 61: 213–224.

Gifford GW, MacLean KA, Hauser MD, et al. (2005) The neurophysiology of functionally meaningful categories: Macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. *Journal of Cognitive Neuroscience* 17: 1471–1482.

Heffner HE and Heffner RS (1986) Effect of unilateral and bilateral auditory-cortex lesions on the discrimination of vocalizations by Japanese macaques. *Journal of Neurophysiology* 56: 683–701.

Poremba A, Malloy M, Saunders RC, et al. (2004) Species-specific calls evoke asymmetric activity in the monkey's temporal poles. *Nature* 427: 448–451.

Rauschecker JP, Tian B, and Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory-cortex. *Science* 268: 111–114.

Romanski LM (2003) Anatomy and physiology of auditory–prefrontal interactions in non-human primates. In: Ghazanfar AA (ed.) *Primate Audition: Ethology and Neurobiology*, pp. 259–278. Boca Raton, FL: CRC Press.

Romanski LM and Goldman-Rakic PS (2002) An auditory domain in primate prefrontal cortex. *Nature Neuroscience* 5: 15–16.

Seyfarth RM, Cheney DL, and Marler P (1980) Vervet monkey alarm calls – semantic communication in a free-ranging primate. *Animal Behaviour* 28: 1070–1094.

Wang XQ, Merzenich MM, Beitel R, et al. (1995) Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: Temporal and spectral characteristics. *Journal of Neurophysiology* 74: 2685–2706.

Zuberbuhler K (2000) Referential labelling in Diana monkeys. *Animal Behaviour* 59: 917–927.

Zuberbuhler K (2002) A syntactic rule in forest monkey communication. *Animal Behaviour* 63: 293–299.

Primate Interneurons

S I Perlmutter, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

More than a century of experimentation has revealed that spinal interneurons perform complex processing of sensory and supraspinal information for all types of movements, from the most basic somatosensory reflexes to highly skilled voluntary movements of the hand. Yet interneurons have been studied directly for relatively few behaviors, unlike neurons in supraspinal structures. The spinal cord is not as accessible in awake animals as the brain is, and the mobility of the spinal column discourages use of microelectrodes. Consequently, most information about interneurons comes from anesthetized or reduced animal preparations or from indirect measurements in animals and humans. Recently, several labs have overcome the technical difficulties and are recording the activity of spinal neurons in awake animals to provide new insights into interneuron function. This article summarizes recent findings from the cervical cord of macaque monkeys performing trained behaviors of the forelimb. Peter Soja, at the University of British Columbia, also studies spinal neurons in awake animals, elucidating state-dependent processing in somatosensory pathways in cats during sleep and wakefulness.

Methodologies

Classical techniques for chronic recording from brain neurons in behaving animals were first adapted for the cervical spinal cord of the monkey in the lab of Eberhard Fetz, at the University of Washington. Access to the cord is achieved by cementing a chamber to the vertebral column over a laminectomy of several vertebrae. Mobility of the spine during recording is reduced by securing the chamber to the recording apparatus. Monkeys are trained to make hand movements in response to visual cues presented on a video screen or to reach to and grasp food pellets. While the monkey performs the tasks, microelectrodes are advanced into the chamber, penetrate the dura mater, and record the activity of individual spinal neurons.

Correlational linkages of neurons to muscles are detected with spike-triggered averages (STAs) of electromyographic (EMG) activity. Distinctive postspike effects, along with location within the cord and characteristic firing properties, identify motor neurons. Postspike facilitation or suppression of EMG activity at latencies consistent with synaptic actions on motor neurons identify premotor neurons (Figure 1). Such identification is difficult when muscles are inactive and is an advantage of the behaving animal model. Output effects from multiple neurons and fibers within a small region are elicited with intraspinal microstimulation and stimulus-triggered averages of evoked EMG activity.

Traditional classifications of spinal interneurons are based on patterns of input from peripheral afferents and supraspinal structures. In awake monkeys, natural stimuli can be delivered only if the forelimb is removed from the experimental device and manipulated, jeopardizing recording stability. Electrical stimulation of peripheral nerves must be limited to low current strengths to avoid aversive sensations. Thus, interneurons usually cannot be identified by the traditional criteria. Another caveat of identification is important. Even though linkages with muscles can be detected with STAs, other projections of recorded neurons are unknown. Thus, the population of spinal neurons sampled from the awake monkey undoubtedly contains local interneurons, commissural interneurons, propriospinal neurons, and ascending tract neurons (collectively called interneurons in this article, to distinguish them from motor neurons and primary afferents). Nevertheless, response patterns during active movement and STAs of EMG activity, supplemented by responses to passive displacements of joints and electrical stimulation of peripheral nerves at group I intensities, establish a set of distinguishing characteristics for interneurons in the behaving monkey.

In the future, functional MRI will probably allow study of interneuron activity during natural movements in humans. To date, however, technical difficulties have limited the success of spinal fMRI. Recent advances suggest that these will be overcome soon.

Neurons in the Cervical Enlargement

Activity during Movement Execution

Response profiles during hand movement The activity of many interneurons is modulated during hand movements made isometrically or against elastic loads (Figure 1). The responses of premotor neurons and those without identified effects on muscles are similar. At least under the conditions studied to date, the activity of premotor neurons does not have clear distinguishing characteristics.

Many task-related interneurons fire at a constant, background rate when the monkey sits quietly before

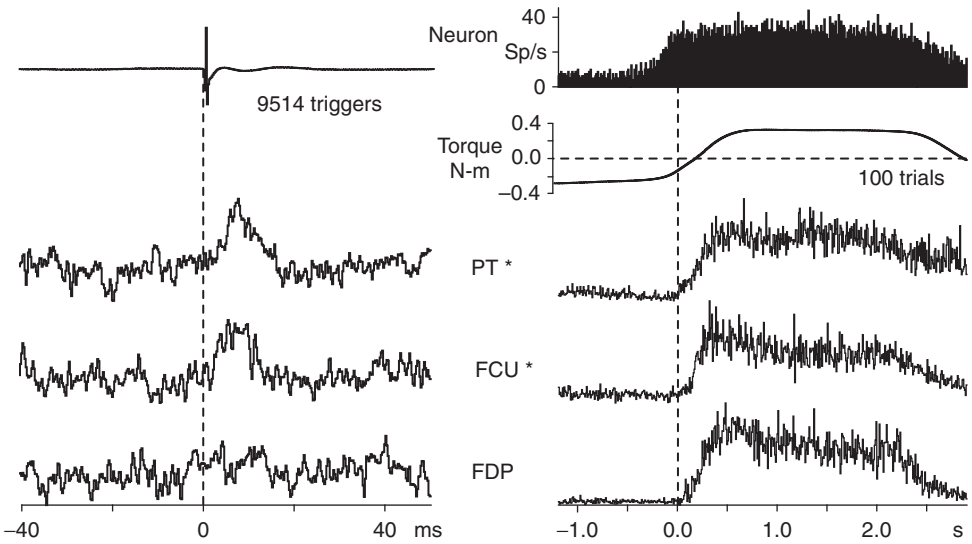

Figure 1 Spike-triggered averages (STAs) and response averages of a premotor interneuron located in the C8 segment. Triggering action potential is shown in top left, above the STAs of electromyographic (EMG) activity in three independently recorded coactive muscles. Asterisks indicate muscles with significant postspike effects. The STAs are plotted with the variances in baseline activity normalized to the STA of the muscle with the largest postspike effect (PT). The average activity of the neuron during flexion torques is shown in the discharge histogram at top right, above the average torque trajectory (positive torque in flexion, negative torque in extension) and average responses of the coactive muscles. The neuron exhibited tonic firing during static torques, at a higher rate for flexion. Reproduced from Perlmutter SI, Maier MA, and Fetz EE (1998) Activity of spinal interneurons and their effects on forearm muscles during voluntary wrist movements in the monkey. *Journal of Neurophysiology* 80: 2475–2494, used with permission of The American Physiological Society.

a cue to move is given, and other neurons are silent. They exhibit a variety of activity patterns when ramp-and-hold torque is generated at the wrist (**Figure 2**). Most neurons fire at a steady rate above their background rate during constant torques. This type of activity is called a tonic response component. During the ramp phase of the task, when torque is changing quickly, firing rate most often either tracks the torque trajectory or transiently increases in a burst, falling to the baseline or tonic level near the end of the torque ramp. These bursts are called phasic response components. Tonic, phasic-tonic, or phasic responses are exhibited by nearly 90% of the task-related neurons. Phasic activity of premotor neurons helps bring motor neurons to firing threshold, perhaps overcoming strong inhibition. Tonic activity contributes to the maintenance of a steady level of muscle contraction.

For the majority of neurons, phasic and tonic response components are correlated with movement parameters. Mean tonic firing rate is linearly related to the magnitude of static torque, and the peak firing rate of the phasic burst is linearly related to the maximum rate of change of torque. These parametric relations suggest that interneurons contribute to the coding of movement amplitude and speed. It is worth noting that interneurons have not been studied during tasks without an applied load, or in which force, position, and muscle activity are uncoupled, so there are no data on which movement parameters are best correlated with neuron discharge. Kazuhiko Seki, at the National Institute of Physiological Sciences in Japan, is beginning to address this issue by recording from spinal interneurons during performance of precision grip.

The activity of interneurons is also related to the direction of hand movement, but only one-fourth reflect motor neuron tuning. Flexor and extensor muscles of the forearm are active for wrist flexion and extension, respectively, and are silent for movements in the opposite direction. One-third of interneurons increase their discharge above the baseline rate for both flexion and extension torques, and another third increase their rate for movement in one direction but remain active for the other.

Premotor neurons in primary motor cortex and red nucleus exhibit similar properties, although there are significant differences. For example, corticomotoneuronal (CM) cells exhibit predominantly tonic and phasic-tonic response components, but rarely purely phasic activity, and are almost always silent for movement opposite to their preferred direction. A study from Yifat Prut's laboratory in Israel shows that the sensitivity of spinal and motor cortex neurons to torque direction and amplitude evolves with similar time courses during the task. Activity related to movement direction appears after cue presentation and

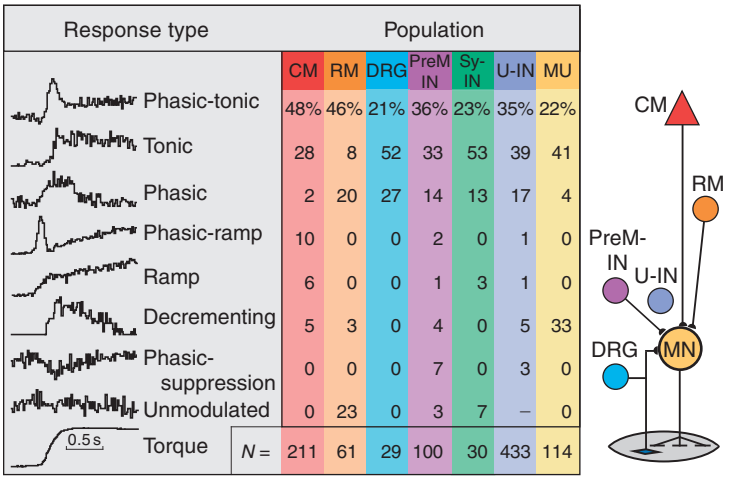

Response type	CM	RM	DRG	PreM IN	Sy-IN	U-IN	MU
Phasic-tonic	48%	46%	21%	36%	23%	35%	22%
Tonic	28	8	52	33	53	39	41
Phasic	2	20	27	14	13	17	4
Phasic-ramp	10	0	0	2	0	1	0
Ramp	6	0	0	1	3	1	0
Decrementing	5	3	0	4	0	5	33
Phasic-suppression	0	0	0	7	0	3	0
Unmodulated	0	23	0	3	7	–	0
N =	211	61	29	100	30	433	114

Figure 2 Summary of response patterns in the preferred direction for different populations of neurons during generation of flexion and extension torques at the wrist. Examples of each pattern are illustrated on the left, and schematic of populations on the right. Proportions are given for corticomotoneuronal (CM) and rubromotoneuronal (RM) cells, premotor afferents in dorsal root ganglia (DRG), spinal premotor interneurons (PreM-IN), unidentified spinal neurons (U-IN), spinal interneurons with synchrony effects (Sy-IN), and motor neurons (MU). From Fetz EE, Perlmutter SI, Prut Y, and Votaw S (2002) Roles of primate spinal interneurons in preparation and execution of voluntary hand movement. *Brain Research Reviews* 40: 53–65. CM data from Cheney PD, Fetz EE, and Mewes K (1991) Neural mechanisms underlying corticospinal and rubrospinal control of limb movements. *Progress in Brain Research* 87: 213–252; RM data from Mewes K and Cheney PD (1994) Primate rubromotor neuronal cells: Parametric relations and contribution to wrist movement. *Journal of Neurophysiology* 72: 14–30; DRG data from Flament D, Fortier PA, and Fetz EE (1992) Response patterns and postspike effects of peripheral afferents in dorsal root ganglia of behaving monkeys. *Journal of Neurophysiology* 67: 875–889.

increases further after movement onset. In contrast, amplitude-related activity is not exhibited during movement preparation.

In another recent experiment, Prut has used electrical stimulation in primary motor cortex to identify functionally connected sites in the spinal cord. Spinal and cortical neurons from connected sites, but not from unconnected sites, express similar directional tuning and high trial-to-trial covariability of firing at movement onset. This suggests facilitated information transfer between cortex and spinal cord at this time, when directional signals, at least, are transmitted. The experiments from Prut's lab are showing how motor output is shaped by continuous interactions between motor cortex and spinal cord.

Spatial tuning of interneurons Bidirectionally active premotor interneurons produce excitatory postsynaptic potentials in motor neurons of forearm flexor and extensor muscles at times when their target muscles are silent, and inhibitory postsynaptic potentials when their muscles are active. Thus, despite the strong relation between firing rate and wrist torque, the activity of spinal premotor interneurons cannot be accounted for entirely by the pattern of forearm muscle activation. Instead, interneuron activity reflects functionally broader inputs.

Studies of the spatial tuning of interneurons for a wider range of hand movements further elucidate this property. In this experiment, monkeys generate torques in flexion–extension (FE), radial–ulnar deviation (RU) and pronation–supination (PS), and perform a power grip (G) that cocontracts wrist flexor and extensor muscles. Activity varies as a cosine function of torque direction in the FE–RU plane for many interneurons but is more broadly tuned for others. Preferred directions are distributed throughout the FE–RU plane, although more neurons prefer radial to ulnar movements. Most neurons with FE–RU responses are also modulated strongly during PS and/or G (Figure 3), and a few respond only for G. On average, the tuning of interneurons is broader than that of forearm muscle EMG activity. The spatial properties of some premotor interneurons are well matched to those of their target muscles, but many have different preferred directions or significantly broader tuning (Figure 3). These data again suggest that spinal premotor neurons are not exclusively related to the activation patterns of the primary wrist movers.

What function could premotor interneurons play beyond specifying the pattern of muscle activation needed to execute a movement? Broad directional tuning and bidirectional activity of interneurons for wrist movements can be produced by the convergence of inputs that are related to limb stabilization, that set reflex thresholds, that indicate the current posture of the limb, or that participate primarily in other movements for which the same muscles are recruited; these inputs are probably not related to the reciprocal activation of forearm muscles for FE–RU movements.

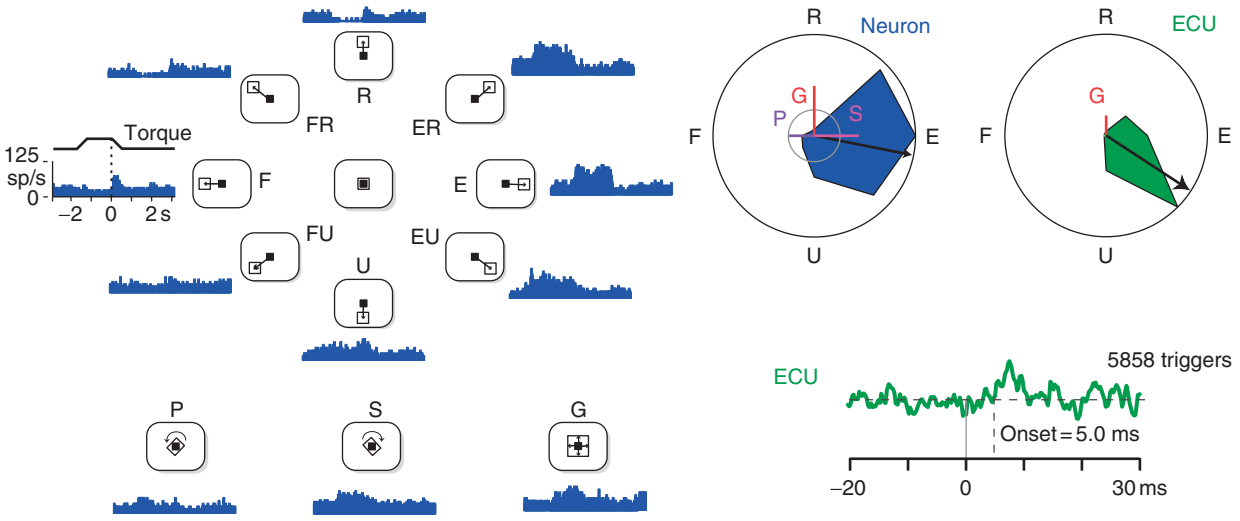

Figure 3 Responses and Spike-triggered average (STA) for a premotor interneuron during performance of a multidirectional hand task. Left: Discharge histograms of the neuron are plotted alongside pictorial representations of the visual display shown to the monkey during isometric torque/force in 11 directions (open box is target, filled box is cursor controlled by monkey's behavior): radial (R), ulnar (U), flexion (F), extension (E), flexion–radial (FR), flexion–ulnar (FU), extension–radial (ER), and extension–ulnar (EU) directions plus pronation (P), supination (S), and power grip (G). Histograms are aligned at the end of torque/force hold period (schematic of torque shown above histogram for F) and show average firing for 10–16 trials. Upper right: Response tuning curves (polygons, normalized to maximum response of the cell or muscle) and preferred directions in the FE–RU plane (vectors, based on best cosine fit to data) are presented for the neuron and the extensor carpi ulnaris (ECU) muscle. Lines indicating responses to P, S, and G overlie the polygons. Tuning curves show responses for the last second of the static hold period for each movement. Lower right: STA of electromyographic (EMG) activity in the ECU muscle showing a significant postspike facilitation, identifying this cell as a premotor interneuron.

Spinal interneurons are well situated to transform this information, from multiple descending and sensory feedback pathways with competing spatial and task-dependent properties, into motor neuron activity appropriate for the current situation. This interpretation is consistent with the widely held view that transmission through spinal pathways is flexible and dynamically modulated during behavior.

Inhibitory control of interneuron firing One way to modulate transmission through spinal pathways is to regulate their excitability by adjusting the balance of excitation and inhibition, even during behaviors for which they might be expected to alternate. The existence and operation of this balance can be explored by blocking one set of inputs with local application of neurotransmitter antagonists. In the behaving monkey, this has been accomplished with a seven-barrel recording-iontophoresis electrode. The central pipette contains a tungsten electrode for extracellular recording and provides the stiffness needed to penetrate the dura. The surrounding barrels are filled with neuroactive drugs that are applied to neurons near the tip of the electrode with iontophoresis.

Iontophoresis of the inhibitory neurotransmitters glycine or γ-aminobutyric acid (GABA) during wrist movements decreases the activity of most interneurons in a dose-dependent manner. It is interesting that the vast majority of cells tested with both glycine and GABA are inhibited by both.

Ejection of the glycine antagonist strychnine and the GABA antagonist bicuculline increases the discharge of many interneurons, even during strong activation in their preferred direction. However, the spatial tuning of about three-quarters of the neurons is not changed by the antagonists. The antagonists increase the tuning width of some neurons, probably by elevating their excitability above firing threshold for some directions, without a qualitative change in directional preference. The inhibitory input to most interneurons is either independent of movement direction or has the same directional tuning as the excitatory input.

Thus, interneurons appear to be under inhibitory control, even when strongly activated during movement, that does not contribute to directional tuning. Instead inhibition may allow fine adjustments in excitability, related to the specific conditions of the task, that alter transmission through premotor pathways. This hypothesis is consistent with current interpretations of overlapping excitatory and inhibitory drive to cortical neurons, as well as with recent experiments from Jorn Hounsgaards's lab in Denmark showing that excitation and inhibition covary in turtle spinal motor neurons during the scratch reflex.

Sensory Responses and the Modulation of Sensory Input

Many interneurons that are active during hand movements also respond to passive displacements of the wrist. They exhibit three stereotypical response patterns to brief perturbations. A C-type response is characterized by short bursts of firing during the onset, and usually the offset, of displacements in both flexion and extension. C-type responses are consistent with input from fast-adapting cutaneous afferents. B-type responses consist of excitation for passive displacement in one direction and inhibition for the other (Figure 4). This pattern produces responses of opposite polarity at the onset and offset of each perturbation. B-type responses are consistent with input from either muscle or joint receptors. A smaller percentage of neurons show monophasic responses that are difficult to ascribe to muscle or cutaneous afferents alone.

Neurons with B- and C-type responses have different activity patterns during active wrist movement. C-type cells have low firing rates at rest and during maintenance of static torque and bidirectional increases in activity during torque onset and offset. B-type cells tend to have higher firing rates at rest and tonic activity during maintained torques (Figure 4). During generation of dynamic torque, B-type cells are phasically excited for one direction and phasically inhibited for the other. Most B-type cells are excited by perturbations in a direction opposite to the preferred direction for active movement, suggesting a consistent involvement in the recruitment of agonist muscles for both active and passive movements. Centrally and sensory-driven response components during active movement can be deduced for some neurons from their activity during voluntary wrist movements (Figure 4).

What type of sensory processing is performed by spinal interneurons during voluntary movements? The first function studied directly in the primate is the modulation of transmission along sensory pathways. It is known that spinal reflexes and cortical somatosensory potentials are reduced, and psychological thresholds for detecting tactile stimuli are

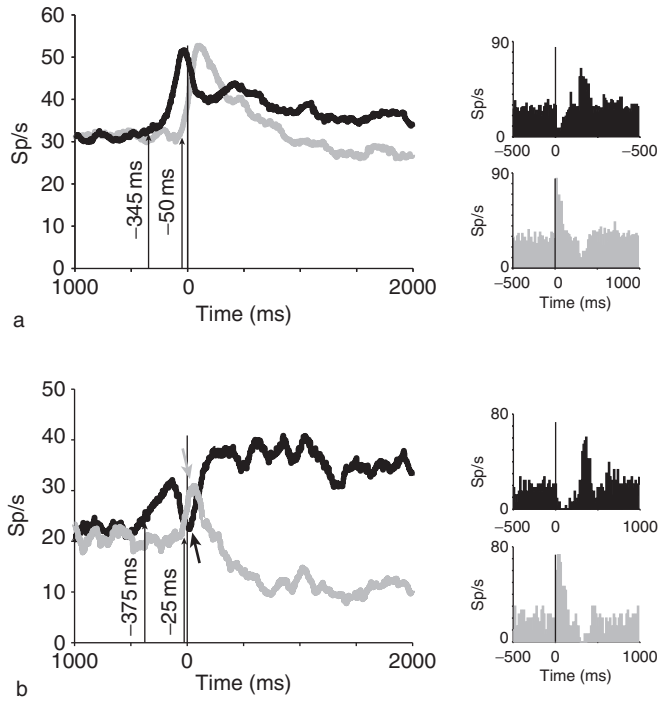

Figure 4 Histograms of firing during active generation of torque (left) and passive wrist displacement (right) for two interneurons (INs; a, b) with B-type responses to wrist perturbations. Black histograms show activity for the direction of movement that produced the largest response during the active-hold period (i.e., preferred direction for active movements). Gray histograms show activity for the opposite direction. Histograms are aligned on movement onset (left) or stimulus onset (right). Both INs have opposite preferred directions for active and passive movements. Small arrowheads indicate onset times for active movement. Both neurons had earlier onset times for active movements in the preferred than in the nonpreferred direction. In (b), phasic responses (large gray arrow) and decrease (large black arrow) in firing near movement onset probably reflect the neuron's sensory response to wrist displacement (note directional and temporal similarity with passive responses shown to the right). Reproduced from Prut Y and Perlmutter SI (2003) Firing properties of spinal interneurons during voluntary movement. I. State-dependent regularity of firing. *Journal of Neuroscience* 23: 9600–9610. Copyright 2003 by the Society for Neuroscience.

increased, before and during movement. The earliest site at which sensory inflow can be regulated is the synapse between primary afferents and second-order neurons in the dorsal horn of the spinal cord. The responses of second-order neurons to electrical stimulation of the superficial radial (SR) nerve, which contains exclusively cutaneous afferents, can be used to monitor transmission along one somatosensory pathway.

Second-order neurons are identified by SR-evoked responses at monosynaptic segmental latencies, relative to the incoming volley recorded in the cord dorsum potential. The SR-evoked response of 66% of second-order neurons is significantly reduced during

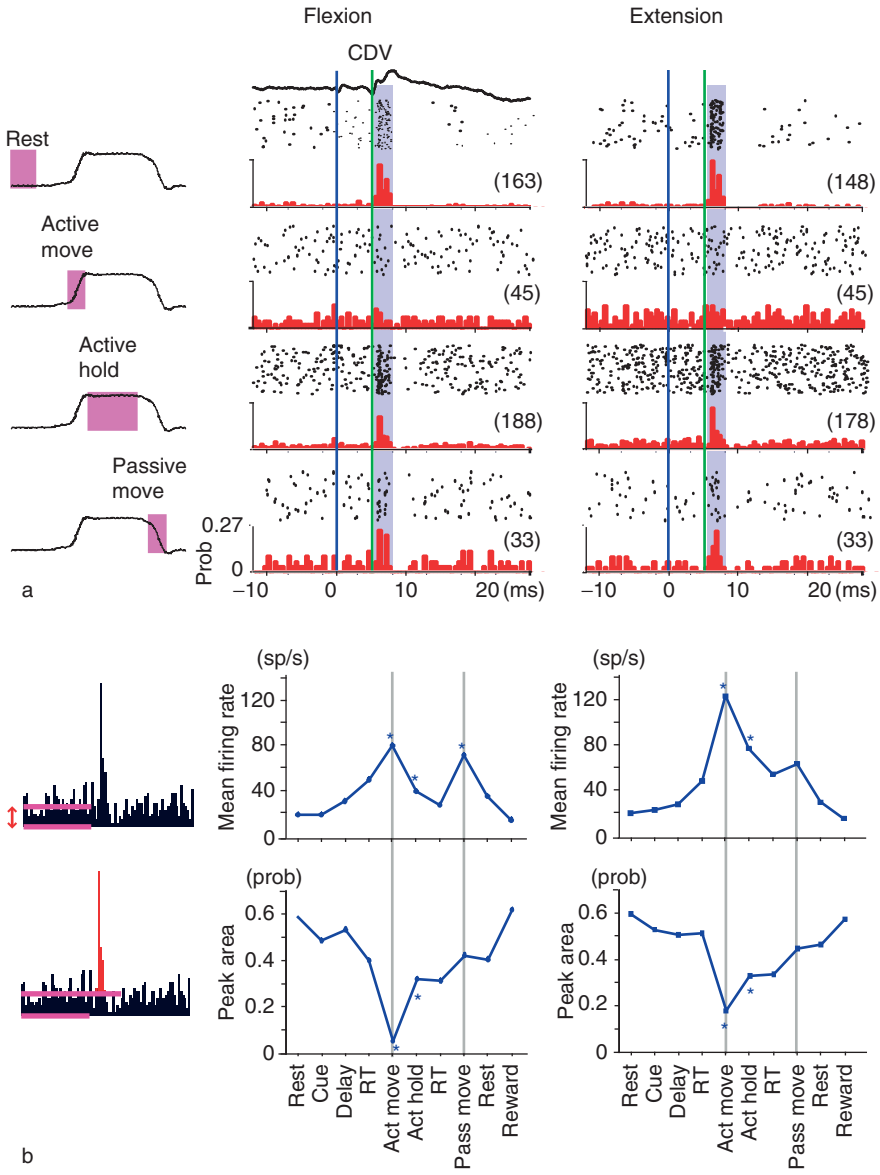

Figure 5 Suppression of monosynaptic, superficial radial (SR)-evoked response during active wrist flexion (left) and extension (right). (a) Peristimulus time histograms below raster plots of the interneuron's activity following SR stimulation (3 Hz, 1.6 times threshold for the afferent volley) aligned with SR stimuli (blue line); trace above shows the cord dorsum potential with afferent volley, CDV, green line). Height of each histogram bin represents the normalized probability of spike occurrence per stimulus. From top to bottom, responses compiled during different task epochs (shading on torque traces to left). Number of stimuli delivered in each epoch shown above each histogram. (b) Averages (±SE) for all second-order neurons of mean firing rate (top) and peak area of SR-evoked response (bottom) for each task epoch, calculated as shown by left insets. Mean firing rate was obtained from mean bin height before (50–1 ms) stimuli. Peak area was sum of bin heights above the mean firing rate during peak (red). Peak duration was time when the bin counts exceeded 2 standard deviations of the mean firing rate. Asterisks: significant differences with rest using a Student's *t*-test. Reproduced from Seki K, Perlmutter SI, and Fetz EE (2003) Sensory input to primate spinal cord is inhibited presynaptically during voluntary movement. *Nature Neuroscience* 6: 1309–1316, with permission from Nature Publishing Group.

voluntary wrist movements (Figure 5). The mean firing rate of most neurons increases during active movement in both directions, suggesting that the suppression of cutaneous input is not related to a decrease of the neurons' excitability due to postsynaptic inhibition. This implicates a presynaptic mechanism, which is also supported by excitability testing of the primary afferents. The magnitude of antidromic volleys recorded in the SR after intraspinal stimulation at sites of second-order neurons increases during active movements, likely due to increased excitability of SR terminals by primary afferent depolarization. The suppression of SR-evoked responses starts before movement onset for more than half of the neurons and does not occur when the wrist is passively moved. Thus, descending commands are the likely source of the presynaptic inhibition.

Electrical microstimulation within the cord at sites of second-order neurons typically produces strong inhibition of task-related EMG activity in forearm muscles, consistent with known inhibitory effects of cutaneous reflexes. Thus, output from these sites could partially suppress ongoing muscle activity. The reduction of cutaneous transmission during movement prevents the sensory-evoked activation of inhibitory pathways to agonist muscles. This might act to reduce inhibition due to expected cutaneous stimulation, including self-induced activation of sensory receptors.

Linkages with Muscles

Postspike effects in forearm muscles are seen for a significant number of spinal neurons active during hand movements. Most premotor interneurons produce effects in only one or two muscles. These 'muscle field' sizes are smaller than those of premotor neurons in the motor cortex or red nucleus. The distribution of effects is also different for premotor supraspinal and spinal neurons. STA features from spinal interneurons are of equal magnitude in flexor and extensor muscles but occur twice as often in flexors than in extensors. Premotor neurons in primary motor cortex produce effects in flexor and extensor muscles with equal frequency, but their magnitude is larger in extensors. Red nucleus neurons produce more effects in extensors. The preferential linkages of spinal interneurons to small numbers of forearm flexors may reflect a greater involvement of spinal processing in the control of individual or small groups of muscles during diverse grasping movements of the hand and less influence over more-stereotyped hand opening and wrist extension movements.

STAs also reveal effects in forearm muscles that have onset latencies too early to be caused by synaptic connections from the recorded neuron to motor neurons. Most of these are synchrony effects, which are typically longer in duration and larger in amplitude than postspike effects and often straddle the trigger time. Such effects are similar to the zero-delay peaks found in cross-correlograms of the activity of neighboring neurons in many nervous system structures. These central peaks are evidence of common synaptic input that increases above chance the probability that the two neurons will discharge synchronously. Analogously, synchrony peaks (or troughs) in STAs of EMG activity reflect common input to the recorded cell and to other neurons with excitatory (or inhibitory) synaptic connections to motor neurons. The frequency with which synchrony effects are produced by interneurons suggests a premotor network of interconnected neurons that receives common supraspinal and feedback signals.

The organization of spinal linkages to muscles has also been studied with intraspinal microstimulation (5–50 μA). In both anesthetized and awake animals, stimulus-triggered averages of EMG activity show short-latency activation of forelimb muscles in response to stimuli delivered at numerous sites. Stimulation usually produces effects in multiple muscles, often in antagonists, and frequently evokes overt movements. Although EMG activity can be evoked from sites throughout the gray and white matter, there is little somatotopic organization within the cervical enlargement for movement or muscle activation, probably due to the excitation by the stimulus of complex mixtures of fibers and cell bodies.

Activity during Preparation for Movement

Cerebral cortex is considered the place in the brain where higher-order computations are processed. However, the cortex and spinal cord are highly interconnected, and interneurons appear likely to carry signals related to some of these functions. One manifestation of higher-order motor processing is activity related to the preparation for movement. Monkeys can be trained to perform isometric wrist flexion–extension with an instructed delay period. At the beginning of this period, a transient visual cue instructs the monkey to generate either flexion or extension torque. However, the monkey is trained to withhold this movement until a second, later visual cue is presented. This task is used to study motor preparation because neural activity can be recorded during a period when the brain is planning, but not executing, a movement. About one-third of spinal interneurons show significant modulation of activity during the delay period, when there is no EMG activity in forearm muscles (Figure 6). Firing rate decreases relative to baseline for 60% of the interneurons and

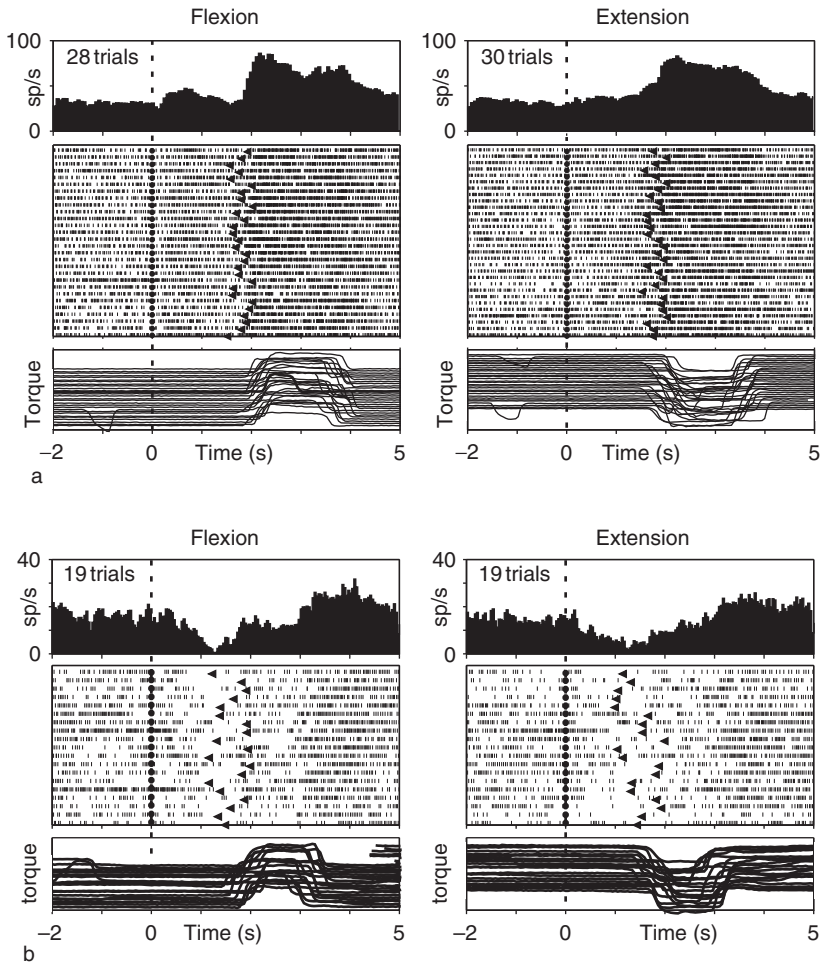

Figure 6 Two spinal interneurons with significant modulation of activity during an instructed delay period. Discharge histograms are shown above single-trial raster plots for isometric flexion (left) and extension (right) torques at the wrist. Diamonds superimposed on rasters indicate torque onset. Torque trajectories during single trials are shown below. Data are aligned on cue onset. Neuron in (a) showed an increased discharge during the delay period before flexion and a bidirectional increase in rate during torque generation in both directions. Neuron in (b) exhibited a decrease in delay period activity in flexion and extension trials. Copyright (© 2001) from Transformation of descending commands into muscle activity by spinal interneurons in behaving primates. In: Cope TC (ed.) *Motor Neurobiology of the spinal cord*, pp. 193–213. London: CRC Press by Perlmutter SI and Prut Y (2001). Reproduced by permission of Routledge/Taylor & Francis Group, LLC.

increases for 40%. Most interneurons with delay period activity also respond during the movement itself, usually with elevated discharge. The magnitude of the responses during the delay period is usually smaller than that during movement execution.

The visual cue gives the monkey information about both the direction and amplitude of the upcoming movement. However, delay activity rarely predicts either parameter of the upcoming movement, in contrast to the activity during the movement (as described above). Furthermore, delay activity of many interneurons is not related to the neuron's subsequent response during active torque, often having the opposite polarity (facilitation vs. suppression).

Thus, spinal interneurons are involved in preparing the motor system to move. Two independent processes appear to be at work. First, neurons that have similar delay- and movement-related changes in activity can prime motor neurons with a subthreshold buildup of postsynaptic potentials in the direction needed for movement execution, perhaps shortening reaction times. Second, the widespread inhibition of neuronal activity during the delay period may suppress the expression of any movement, including those elicited by sensory stimuli, until motor preparation is complete. In fact, premotor interneurons are significantly more likely to exhibit delay period activity than are interneurons with no functional linkages to muscles, and this response is usually inhibitory.

Propriospinal Neurons

In the cat, propriospinal neurons in the C3–C4 segments mediate disynaptic excitation of forelimb

motor neurons from the motor cortex. Lesion of this pathway disrupts muscle coordination during forelimb target reaching. It has been hypothesized that the feline motor cortex controls specific forelimb muscle synergies by selective activation of C3–C4 propriospinal neurons and inhibitory networks which focus their outputs. Electrophysiological experiments performed by Tadashi Isa and Bror Alstermark have demonstrated that similar propriospinal neurons exist in monkeys. However, the functional importance of this system in primates is still unclear. In humans, Emmanuel Pierrot-Deseilligny and his colleagues have provided support for a significant role from studies using conditioned H-reflexes and transcranial magnetic stimulation of motor cortex. However, disynaptic cortical excitation of arm motor neurons has been difficult to demonstrate in monkeys without the systemic administration of the glycine receptor antagonist strychnine.

The existing data have led to two current hypotheses. First, as argued by Roger Lemon and his colleagues, propriospinal neurons in the primate exert weak influences on arm motor neuron excitability and serve a relatively minor role in comparison with direct CM connections. Second, as argued by Alstermark and Isa, propriospinal neurons mediate a substantial corticospinal excitation to arm motor neurons that is under strong feed-forward and feedback inhibitory control. Recent experiments on the activity of neurons in the C3–C4 segments of behaving monkeys have added new data to this debate.

Activity of C3–C4 Neurons during Hand and Reaching Movements

Many C3–C4 neurons are strongly modulated during wrist flexion–extension and a reach-and-grasp task. Activity correlated with wrist extension and arm reaching predominates, with far fewer neurons modulated during wrist flexion or hand grasping. Some C3–C4 neurons are activated during wrist movements only, others during reach only, and others during both the wrist and the reach task. This is consistent with involvement in specific movements or the activation of specific combinations of arm muscles, as proposed for propriospinal neurons in the cat. At least some task-related neurons are activated monosynaptically by electrical stimuli delivered to the motor cortex, and antidromically by stimulation in motor nuclei of the cervical enlargement. These cells can mediate disynaptic cortical excitation of arm muscle motor neurons.

Output Effects Evoked with Microstimulation

Intraspinal microstimulation at widespread sites in C3–C4 produces robust output effects in stimulus-triggered averages of arm muscle EMG activity. Most output sites in C3–C4 activate muscles at more than one joint. Qualitatively, the effects from C3–C4 are similar to those evoked from the cervical enlargement, although there are differences in the frequency, distribution, and mean latency of effects from the two levels. Most interesting, effects evoked from C3–C4 have amplitudes that are significantly larger in extensor than in flexor muscles, whereas the effects evoked from C6–T1 are larger in flexors. The neural activity patterns and the stimulation-evoked output effects suggest that non-CM components of motor commands for different movements are routed through different premotor pathways in the spinal cord. Flexion movements of the hand, like grasping, may primarily involve activation of circuits in the cervical enlargement, and arm reaching and hand extension may primarily engage those in the upper cervical segments.

Linkages with Muscles

Neurons that exert strong control on motor neurons should produce postspike effects in STAs of EMG activity. However, the majority of C3–C4 neurons do not produce such effects during the wrist flexion–extension and reach-and-grasp tasks, although many synchrony effects are produced.

Strong responses for reaching and wrist extension, subpopulations with activity correlated with different combinations of movements, and strong stimulation-evoked output effects to extensor muscles in multiple joints are all consistent with an important role for C3–C4 propriospinal neurons in the activation of arm muscles during voluntary movement. However, the paucity of functional linkages to arm muscle motor neurons argues against direct premotor projections from C3–C4. Future studies will have to resolve this discrepancy before the function of upper cervical propriospinal neurons in the primate is completely understood.

See also: Motor Primitives; Neonatal Circuits; Neural Coding in Primary Motor Cortex; Operant Conditioning of Reflexes; Reaching and Grasping; Representation of Movement; Reticulospinal System; Sensorimotor Plasticity and Control of Movement Following Spinal Cord Injury +D18; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Regeneration and Functional Recovery: Strategies; Spinal Cord Injuries; Spinal Motor Neurons: Properties.

Further Readings

Alstermark B and Isa T (2002) Premotoneuronal and direct corticomotoneuronal control in the cat and macaque monkey. *Advances Experimental Medicine Biology* 508: 281–297.

Berg RW, Alaburda A, and Hounsgaard J (2007) Balanced inhibition and excitation drive spike activity in spinal half-centers. *Science* 315: 390–393.

Cheney PD, Fetz EE, and Mewes K (1991) Neural mechanisms underlying corticospinal and rubrospinal control of limb movements. *Progress in Brain Research* 87: 213–252.

Fetz EE, Perlmutter SI, Prut Y, and Seki K (2002) Functional properties of primate spinal interneurons during voluntary hand movements. In: Gandevia SC, Proske U, and Stuart DG (eds.) *Sensorimotor Control of Posture and Movement*, pp. 265–271. New York: Kluwer Academic.

Hultborn H (2006) Spinal reflexes, mechanisms and concepts: From Eccles to Lundberg and beyond. *Progress in Neurobiology* 78: 215–232.

Jankowska E (1992) Interneuronal relay in spinal pathways from proprioceptors. *Progress in Neurobiology* 38: 335–378.

Jankowska E and Edgley SA (2006) How can corticospinal tract neurons contribute to ipsilateral movements? A question with implications for recovery of motor functions. *Neuroscientist* 12: 67–79.

Lemon RN and Griffiths J (2005) Comparing the function of the corticospinal system in different species: Organizational differences for motor specialization? *Muscle Nerve* 32: 261–279.

Perlmutter SI, Maier MA, and Fetz EE (1998) Activity of spinal interneurons and their effects on forearm muscles during voluntary wrist movements in the monkey. *Journal of Neurophysiology* 80: 2475–2494.

Perlmutter SI and Prut Y (2001) Transformation of descending commands into muscle activity by spinal interneurons in behaving primates. In: Cope TC (ed.) *Motor Neurobiology of the Spinal Cord*, pp. 193–213. London: CRC Press.

Pierrot-Deseilligny E and Burke D (2005) *The Circuitry of the Human Spinal Cord*. Cambridge, UK: Cambridge University Press.

Prut Y and Fetz EE (1999) Primate spinal interneurons show premovement instructed delay activity. *Nature* 401: 590–594.

Seki K, Perlmutter SI, and Fetz EE (2003) Sensory input to primate spinal cord is inhibited presynaptically during voluntary movement. *Nature Neuroscience* 6: 1309–1316.

Soja PJ (2007) Modulation of prethalamic sensory inflow during sleep and wakefulness. In: Lavigne G, Choiniere M, Sessle B, and Soja P (eds.) *Sleep and Pain*, pp. 45–76. Seattle: International Association for the Study of Pain Press.

Stroman PW (2005) Magnetic resonance imaging of neuronal function in the spinal cord: Spinal fMRI. *Clinical Medicine Research* 3: 146–156.

Takei T and Seki K (2008) Spinomuscular coherence in monkeys performing a precision grip task. *Journal of Neurophysiology* 99: 2012–2020.

Yanai Y, Adamit N, Harel R, Israel Z, and Prut Y (2007) Connected corticospinal sites show enhanced tuning similarity at the onset of voluntary action. *Journal of Neuroscience* 27: 12349–12357.

Priming

R Henson, MRC Cognition and Brain Sciences Unit, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Priming refers to a change in behavioral response to a stimulus, following prior exposure to the same, or a related, stimulus. Examples include faster reaction times to make a decision about the stimulus, a bias to produce that stimulus when generating responses, or the more accurate identification of a degraded version of the stimulus. From the memory perspective, the most exciting aspect to priming concerns evidence that it can occur in the absence of conscious memory for the prior exposure (an example of so-called 'implicit' or 'nondeclarative' memory). For instance, amnesic patients typically show intact levels of priming, despite their impairments in conscious ('explicit' or 'declarative') memory. Evidence like this led to the proposal that the brain regions supporting priming are different from the medial temporal lobe (MTL) regions believed to support declarative memory. Indeed, it is often assumed that priming is the by-product of prior processing of a stimulus, that is, arising from plasticity in multiple cortical regions whose primary role is perceptual/conceptual processing. As such, priming is likely to be one of the most basic expressions of memory, influencing how we perceive and interpret the world.

A Working Example

Priming can be measured using a range of different stimuli and tasks. To illustrate the breadth of neuroscientific data on priming, we initially focus on one specific paradigm – the word-stem completion paradigm – which has provided some of the most convincing evidence that priming involves memory systems, and associated brain regions, that differ from those involved in declarative memory. In the 'study' phase of this paradigm, participants perform an incidental task on a list of words (normally unaware that their memory for these words will be tested). At some time later, they are presented with three letters that form the start of words (word-stems) (see Figure 1(a)). Some of these stems can be completed with the words from the study phase, though in the 'indirect' version of this task, participants are not informed of any relationship between this test phase and the previous study phase. They are simply asked to complete the stem with the first word that comes to mind. Priming is indexed by the increased probability of completing stems with studied words, relative to the baseline probability of completing the stems when those words were not studied. Importantly, participants normally show this priming without reporting any conscious reference to the prior study phase.

Neuropsychological Evidence

When this task is given to amnesic patients, they typically show priming of equivalent magnitude to controls. However, when the instructions at test are changed, so that participants are asked to use the stems explicitly to recall studied words ('word-stem cued recall,' a 'direct' version of the task), amnesic patients typically perform worse than controls (Figure 1(b)). This dissociation is appealing because the stimuli are kept constant across the indirect and direct versions of the task; all that differs are the instructions. While amnesic patients are also known to have sparing of other forms of memory, such as motor skill-learning, the intriguing aspect of priming is that it can occur after a single stimulus presentation ('one-shot learning'), rather than requiring sustained training. Furthermore, a few patients with more posterior lesions (e.g., in occipital cortex) have shown the reverse pattern of impaired word-stem completion for visually studied words (indirect test), despite intact performance on a recognition memory test (direct test). This double dissociation between the two types of memory test is strong support for different memory systems underlying priming and declarative memory. Further support comes from reports that word-stem completion priming is less susceptible than is cued recall to the effects of healthy aging, and to some dementias like Alzheimer's disease.

Psychological Evidence

This dissociation of priming from declarative memory following brain damage is supported by functional dissociations between direct and indirect versions of the word-stem completion paradigm in healthy individuals. For example, performance on the direct version (word-stem cued-recall) is higher when the corresponding words are generated from semantic cues at study relative to when they are simply read, yet performance on the indirect version (word-stem completion) is higher when the words are read relative to being generated from cues. Similarly, presenting words auditorily rather than visually at study tends to impair performance on visual word-stem completion to a greater extent than on visual word-stem cued

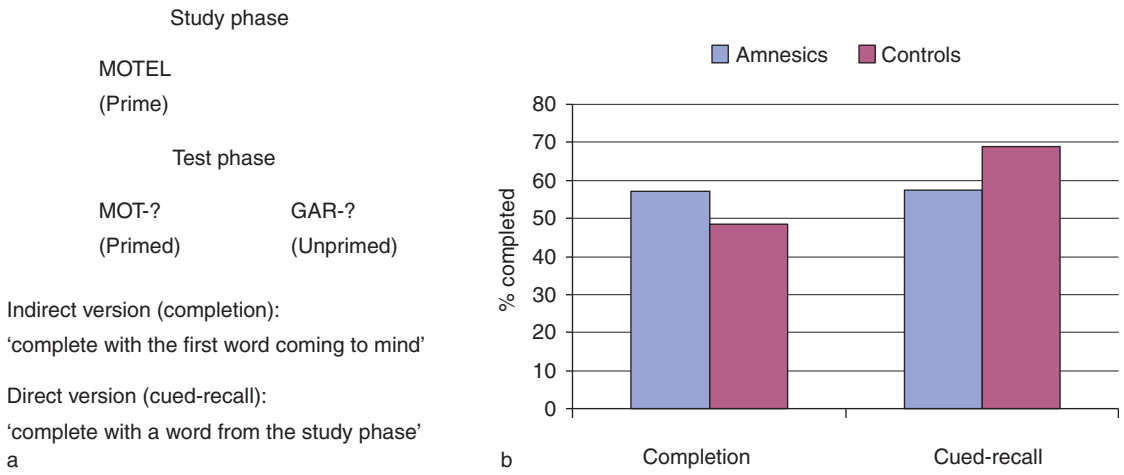

Figure 1 (a) Schematic of the word-stem completion paradigm. (b) Performance of amnesic patients and matched controls on indirect (completion) and direct (cued-recall) versions of the task. (b) Reprinted with permission from Graf P, Squire LR, and Mandler G (1984) The information that amnesic patients do not forget. *Journal of Experimental Psychology*: *Learning, Memory and Cognition* 10: 164–178, APA publishers.

recall, whereas use of a study task that engages semantic versus phonological processing of words tends to improve performance on word-stem cued recall to a greater extent than on word-stem completion. The reduced performance found on word-stem completion when words are generated or heard at study, rather than read, suggests that a component of priming in this task relates to the visual overlap between words and their stems, or more precisely, to overlap in the processes involved in visual word reading.

Nonetheless, though dissociations have been found between direct and indirect tests of memory, care must be taken when using performance on indirect tests as a pure measure of implicit memory: even though an indirect memory test does not refer participants to previous encounters with stimuli, participants may voluntarily, or involuntarily, recollect such encounters. For example, some participants may guess the relationship between the study and test phase of a word-stem completion task, and hence attempt to recall studied items in order to complete the stems. Moreover, even if a completion comes to mind involuntarily, participants may subsequently recognize that word as having come from the study phase. Though the latter eventuality may not affect behavioral measures of priming, it is important when measuring the neural correlates of priming, as discussed in the next two sections. As a consequence, considerable effort has been devoted to developing methods that minimize the contribution of explicit memory to priming.

Neuroimaging Evidence

Some of the first positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) studies of priming used the word-stem completion paradigm. These studies consistently found reduced hemodynamic responses in extrastriate visual cortex for studied relative to unstudied word-stems (often in both direct and indirect versions of the task). This reduction is attributed to 'more efficient' or 'facilitated' visual processing of the stems corresponding to studied words, owing to the prior exposure to those words. Similar reductions have also been found in a number of regions including left inferior temporal and left inferior frontal cortices. Given that these regions are also generally activated by the task (e.g., vs. passive fixation), the reductions are consistent with facilitated processing within a language-related network that is engaged when reading words and completing word-stems, possibly including phonological, lexical, and semantic processes.

Some imaging studies have also reported hemodynamic increases in MTL cortices, even in indirect versions of the task. This illustrates the problem highlighted earlier that explicit memory may contaminate indirect tasks. In particular, even if this explicit memory is involuntary and does not affect the behavioral response, the poor temporal resolution of hemodynamic imaging techniques means that imaging data may include explicit memory processes arising subsequent to that response. More recent imaging experiments have therefore used variants of the word-stem completion paradigm that attempt to minimize contamination by explicit memory. For example, participants can be asked to complete stems with words from the study phase (i.e., word-stem cued recall), but if they cannot, they are asked to produce the first word coming to mind. Having produced a completion (through either approach), they then indicate whether they remember that word from the study

phase. The critical trials used to isolate the neural correlates of priming are those stems completed with studied words but which participants indicated they did not remember. These trials are likely to provide a purer measure of implicit memory (despite occurring in the context of a direct memory task). When compared to a baseline condition (stems that did not correspond to studied words, but which were completed with a word that was correctly indicated as unstudied), event-related fMRI has replicated the reduced responses found in extrastriate, left inferior temporal, and left inferior frontal cortices (Figure 2(a)). Equivalent reductions in these regions were also found for 'remembered' trials (stems that were completed with a studied word and that word was indicated as studied), unlike in MTL cortex, where greater activity was found for remembered trials relative to primed trials (suggesting that explicit memory had been successfully minimized in primed trials). Furthermore, the reductions for primed and remembered versus baseline trials in these regions were also found when using an indirect version of the task akin to the standard word-stem completion task. Taken together, these results suggest that such hemodynamic reductions are not only implicit, but also automatic, consistent with a role for these regions in the priming observed in healthy individuals and in amnesic patients.

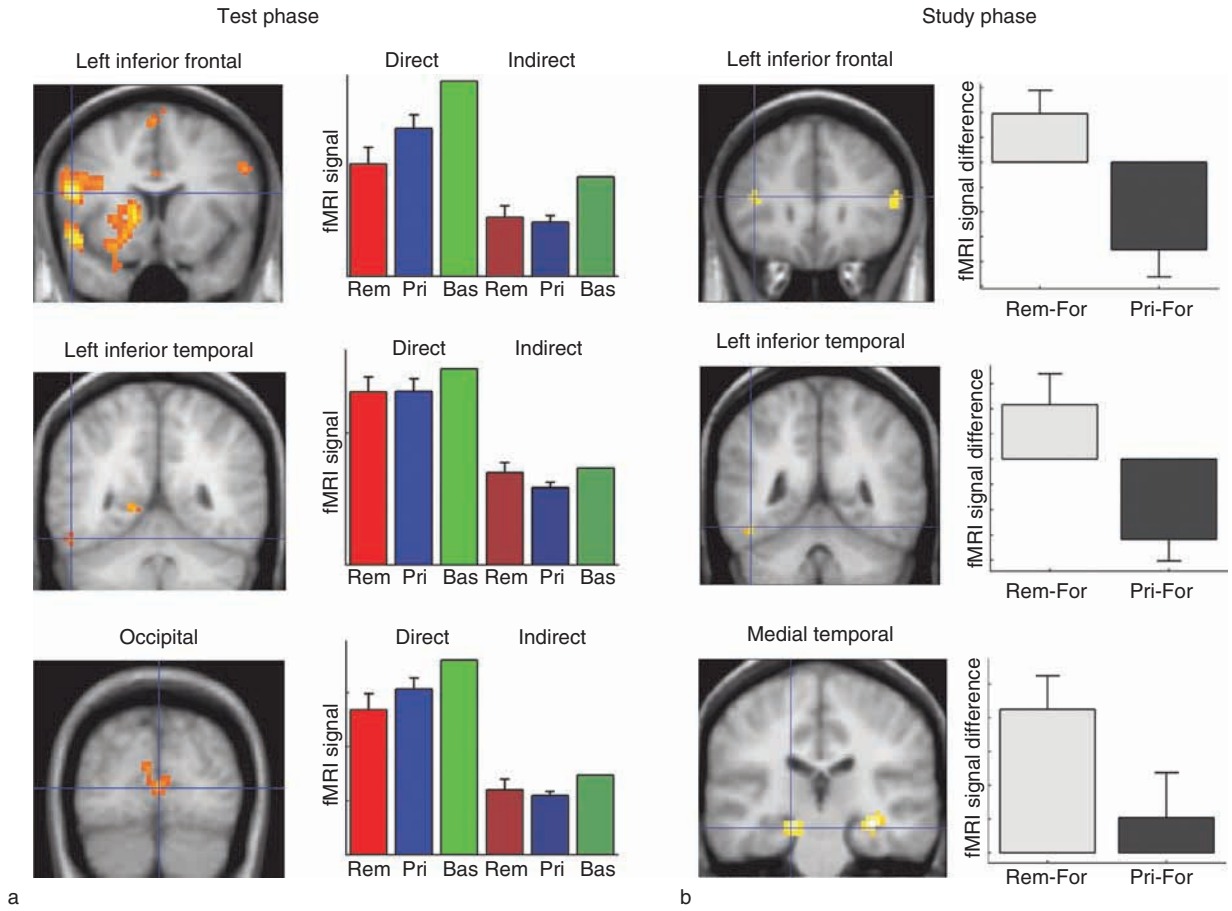

Figure 2 Event-related fMRI data from a word-stem completion paradigm. In the test phase (a), a number of regions including those shown in left inferior frontal, left inferior temporal and occipital cortices, exhibited reduced responses to primed stems (blues), and to remembered stems (reds), relative to baseline stems (greens), regardless of whether participants performed a direct (lighter colors) or indirect (darker colors) version of the paradigm. In the study phase (b), a number of regions including those shown in left inferior frontal and inferior temporal cortices showed reduced responses for subsequently primed relative to subsequently forgotten words (darker bar), whereas regions in medial temporal lobe showed increased responses for subsequently remembered (but not subsequently primed) relative to subsequently forgotten words (lighter bar). Rem, remembered; Pri, primed; Bas, baseline; For, forgotten (see text). (a) Adapted with permission from Schott BH, Henson RN, Richardson-Klavehn A, et al. (2005) Redefining implicit and explicit memory: The functional neuroanatomy of priming, remembering and control of retrieval. *Proceedings of the National Academy of Sciences of United States of America* 102: 1257–1262. (b) Adapted with permission from Schott B, Richardson-Klavehn A, Henson R, et al. (2006) Neuroanatomical dissociation of encoding processes related to priming and explicit memory. *Journal of Neuroscience* 26: 792–800. Copyright (2005) National Academy of Sciences, USA.

One can also use event-related imaging methods to examine differences at the time stimuli are encoded (i.e., during the study phase) as a function of the degree to which those stimuli show later priming (or the degree to which they show explicit memory). Using the same paradigm as above, for example, hemodynamic responses to the words at study differed for those words subsequently primed (completed with the studied word but not remembered) than those subsequently forgotten (completed with a different word). Such differences were found in a number of cortical regions, including the left inferior temporal and frontal regions that showed priming-related reductions at test. Again, the pattern in these regions differed from that in the MTL, which showed greater activity for words subsequently remembered versus words subsequently forgotten (Figure 2(b)). These findings suggest that priming is affected by how stimuli are encoded (on their initial presentation), possibly in addition to how that information is accessed at test. Yet further work has combined event-related fMRI with pharmacological interventions: for example, scopolamine and lorezapam, which are believed to modulate synaptic plasticity, have been shown to attenuate repetition suppression in the same inferior temporal and frontal regions during word-stem completion.

Electrophysiological Evidence

The time course of word-stem completion priming has also been investigating using electroencephalographic (EEG) and magnetoencephalographic (MEG) methods. For example, in a standard indirect completion task, event-related potentials (ERPs) to primed word-stems have been reported to diverge from those to baseline (unstudied) word-stems within 200 ms of the onset of the word-stem (Figure 3(a)). The same ERP effect was reported for words remembered during

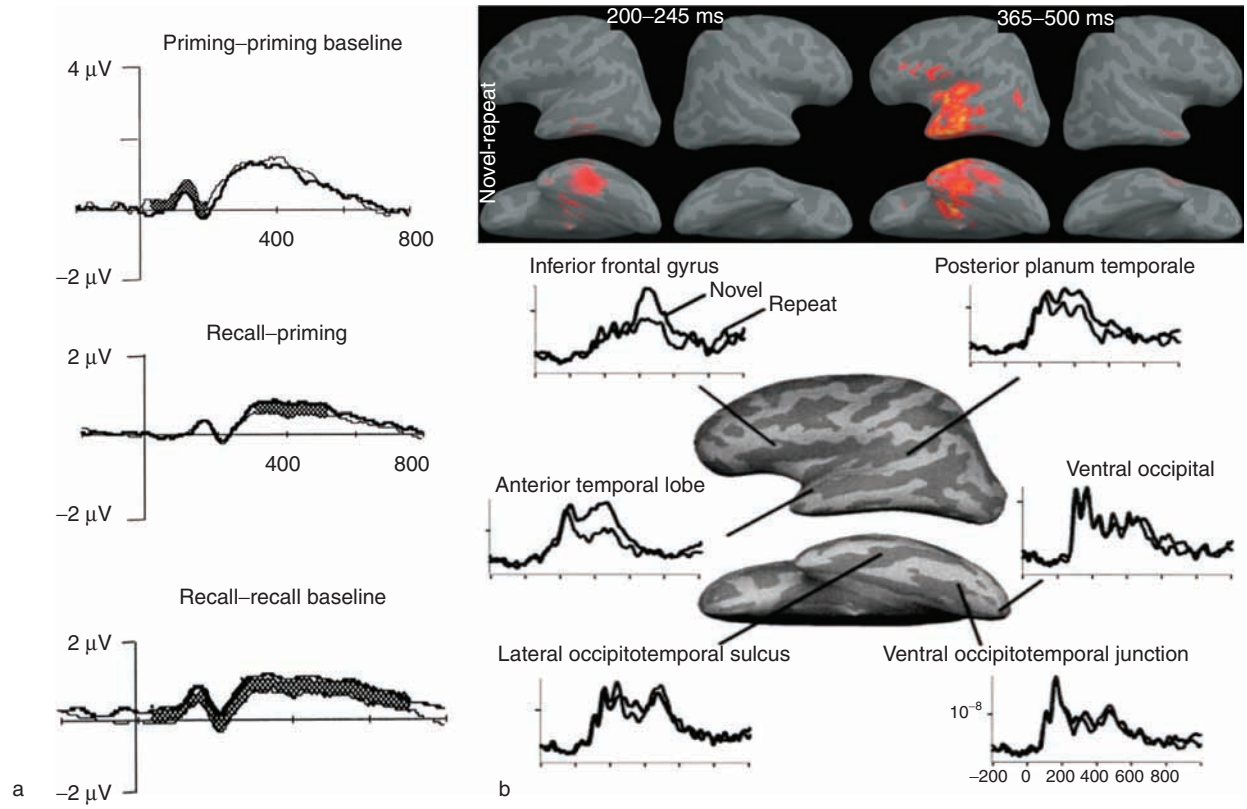

Figure 3 Electrophysiological data from a word-stem completion paradigm. In (a), the event-related potentials (ERPs) over right parietotemporal electrodes recorded by EEG diverge between primed and baseline stems (upper plot), and between recalled and baseline stems (lower plot), within 200 ms, whereas the ERPs to primed and recalled stems only differ from 300 ms onwards (topographies also differ). In the upper panel of (b), regions with reliable differential source-reconstructed event-related activity, inferred from MEG data, for repeated versus novel presentation of stems are shown for two time windows. In the lower panel, the reconstructed responses are shown for a number of regions of interest. (a) Data adapted with permission from Badgaiyan RD and Posner MI (1997) Time-course of cortical activations in implicit and explicit recall. *Journal of Neuroscience* 17: 4904–4913. (b) Data adapted with permission from Dhond RP, Buckner RL, Dale AL, Marinkovic K, and Halgren E (2001) Spatiotemporal maps of brain activity underlying word generation and their modification during repetition priming. *Journal of Neuroscience* 10: 3564–3571.

a direct cued-recall version of the task. However, when contrasting primed and remembered word-stems directly, differences only emerged after approximately 300 ms. These data suggest that priming occurs prior to explicit memory retrieval. Furthermore, and consistent with fMRI data described above, qualitative differences have been found in the study between the differential ERPs to words subsequently primed versus those subsequently remembered. More recently, related experiments (in which word-stems were repeatedly completed) have used the superior spatial resolution of MEG (combined with structural information from MRI) in an attempt to localize the evolution of repetition-related effects over the first few hundred milliseconds from stem onset. Interestingly, an initial repetition-related increase in electrical activity has been reported early in left inferior temporal cortex (lateral occipitotemporal sulcus), followed by a later decrease in both this region and left inferior frontal cortex (Figure 3(b)). Such spatiotemporal data suggest rapid interactions between brain regions, which would be invisible to hemodynamic techniques like fMRI. Indeed, such data afford greater opportunity to look at priming-related changes in the connectivity between different brain regions, for example, as a function of plasticity following initial processing of stimuli.

Summary

The above studies using the word-stem completion paradigm provide convergent evidence that priming reflects a form of memory dissociable from conscious memory, and can arise from a number of cortical regions outside the MTL. These regions are normally a subset of those engaged by the specific stimuli and task (in the above cases, a number of left inferior frontal, left inferior temporal, and occipital regions associated with word reading), suggesting that priming reflects a form of plasticity in these regions. The expression of priming is generally associated with reduced neural activity in these regions, possibly reflecting faster and/or more efficient processing (e.g., a 'greasing of the tracks'). This expression appears to occur very rapidly (around 200 ms) and automatically (regardless of precise task). Differences in neural activity have also been found during the initial processing of stimuli as a function of whether they are subsequently primed. This implicates important roles for factors such as attention and the type of processing in the nature of neural/synaptic changes assumed to underlie priming.

Similar conclusions have been reached using a number of other priming paradigms, though the specific network of cortical regions involved may differ. Indeed, the general finding of decreased hemodynamic responses to repeated stimuli (which has been called 'repetition suppression') has been observed repeatedly across a range of stimuli and indirect memory tasks. Interestingly, however, not all regions activated by initial presentations of stimuli show such decreases: early sensory regions, such as V1, for example, do not normally show repetition suppression. One possibility is that not all of the processing stages in a given task show stimulus-specific effects of prior processing (or alternatively, such effects of prior processing have lifetimes shorter than is typically tested). An interesting theoretical question is therefore which stages (and associated brain regions) show the largest effects of prior processing? Are they, for example, stages that involve greatest competition between alternative outputs/interpretations?

Other Common Measures of Priming

Another common measure of priming is the latency to identify (e.g., name) pictures of objects. This measure has shown priming effects that last up to a year. In a common neuropsychological version, in which degraded pictures are gradually clarified (the 'Gollin figures'), amnesic patients show strong priming (i.e., earlier identification for pictures identified previously). Researchers have also used such priming to investigate the representation of visual objects in the brain, by examining the extent to which priming-related effects (e.g., in different brain regions) generalize across different views of an object.

Another common type of priming is semantic priming, where the response to a probe word (e.g., HEAVEN) is speeded when preceded by a related prime word (e.g., HELL) relative to a nonrelated word (e.g., HILL). Such priming tends to be short-lived, often not surviving more than a second between the prime and probe. As with visual object priming, it has also been used as a tool, in this case to investigate the organization of semantic memory. Yet another type of priming that has been influential on a theoretical level is associative priming, where the priming of a probe stimulus is modulated by whether or not it has previously been paired with a context stimulus. According to theories that hypothesize a key role for MTL in forming associations, patients with MTL damage should not show such priming, even if such priming is shown to be implicit. At the moment, the data appear mixed.

There is also much interest in subliminal priming, in which the prime is presented so briefly (around 50 ms; often forward and backward masked) that participants show little to no awareness of the prime, and yet still show behavioral priming to a probe stimulus

presented shortly afterwards (typically within a few hundred milliseconds). Indeed, this type of priming has been recommended for imaging studies of implicit processing, since it is easier to rule out contamination by explicit memory for the prime.

Theories of Priming

A broad distinction has been made between 'perceptual' and 'conceptual' priming. The main difference is that perceptual, but not conceptual, priming is affected by differences in the physical features of the prime and probe stimuli, whereas conceptual, but not perceptual, priming is affected by differences in the degree of semantic processing of the stimuli. This distinction is supported by dissociations in, for example, Alzheimer's patients, who reportedly show intact perceptual priming but impaired conceptual priming, presumably because the early sensory areas believed to be important for perceptual priming are less affected by the disease than anterior temporal regions believed to be important for conceptual priming. A related distinction assumes that priming is supported by specialized 'perceptual representation systems' that are distinct from the MTL memory system. In principle, this allows for dissociable forms of perceptual priming as a function of modality (e.g., visual vs. auditory).

Though useful heuristics, such broad distinctions do not seem sufficient for the varieties of priming that have been investigated. For example, the word-stem completion task discussed above is difficult to fit into the perceptual/conceptual dichotomy, since it is affected by variations in some physical properties, yet residual priming can still occur across modalities (even when voluntary explicit retrieval has been excluded). Furthermore, word-stem completion can be unaffected by the degree of semantic processing of primes (e.g., semantic vs. phonological), yet be reduced following very superficial (e.g., graphemic) processing of primes. These data are consistent with a contribution from a lexical component, in addition to perceptual components. Furthermore, other dimensions along which indirect tasks differ can also produce behavioral dissociations, such as the distinction between competitive and noncompetitive access to conceptual information, for example.

An alternative perspective is the 'component process' view of priming, according to which there are typically several processes involved in a given task that may be facilitated by prior processing. As with the related concept of 'transfer-appropriate processing,' the amount of priming will depend on the degree of overlap between the processes performed on the prime and those performed on the probe (**Figures 4(a) and 4(b)**). These processes may derive from interactions between brain regions that become configured according to the task demands. Therefore, "performance will depend on the interplay of components and the processes they mediate" (Moscovitch, 1994: 301). A behavioral measure like reaction time is likely to reflect the conglomeration of the times taken to perform each component process. Only by varying the stimuli and task can the key components in behavioral priming be isolated (**Figures 4(b) and 4(c)**). Alternatively, the concurrent acquisition of hemodynamic or electrophysiological data allows the opportunity to examine activity within the 'network' of brain regions (and the interactivity between those regions) that are configured for the current task.

Another important theoretical dimension is that between 'structural' and 'episodic' theories of priming. According to structural (or 'abstractionist') theories, priming reflects some modification, such as lowered thresholds of, or residual activity in preexisting representations. According to episodic (or 'instance') theories, any exposure to a stimulus can, in principle, leave some residual trace of its processing. Attempts to test these theories have used priming of novel stimuli (e.g., nonwords, or impossible objects) or examined the effects of context (e.g., novel associations) on priming. In general, this distinction is hard to evaluate in the absence of more detailed theories of the nature of representation in the brain (e.g., even nonwords can consist of familiar bigrams of letters). Nonetheless, one important class of episodic theories assumes that each stimulus and response within a task can become associated together, or bound into a single representation (or 'event record'). In such cases, repetition of the stimulus can directly cue retrieval of the associated response. Such stimulus–response associations have been shown to influence behavior, particularly in the case of subliminal (masked) priming, but also in the case of 'negative priming,' when the response to a probe stimulus is hindered if the same stimulus has previously been presented in the context of a different response (even if an implicit response). Such associations have also been shown to have dramatic effects on imaging data. Though stimulus–response associations can be minimized by ensuring that the response made to the probe is unrelated to the response made to the prime, it is clear that when present, such associations can have important consequences, possibly 'by-passing' several stages of processing that occurred on the prime (**Figure 4(d)**).

Future

One important future direction that provides strong evidence for the direct involvement of circumscribed

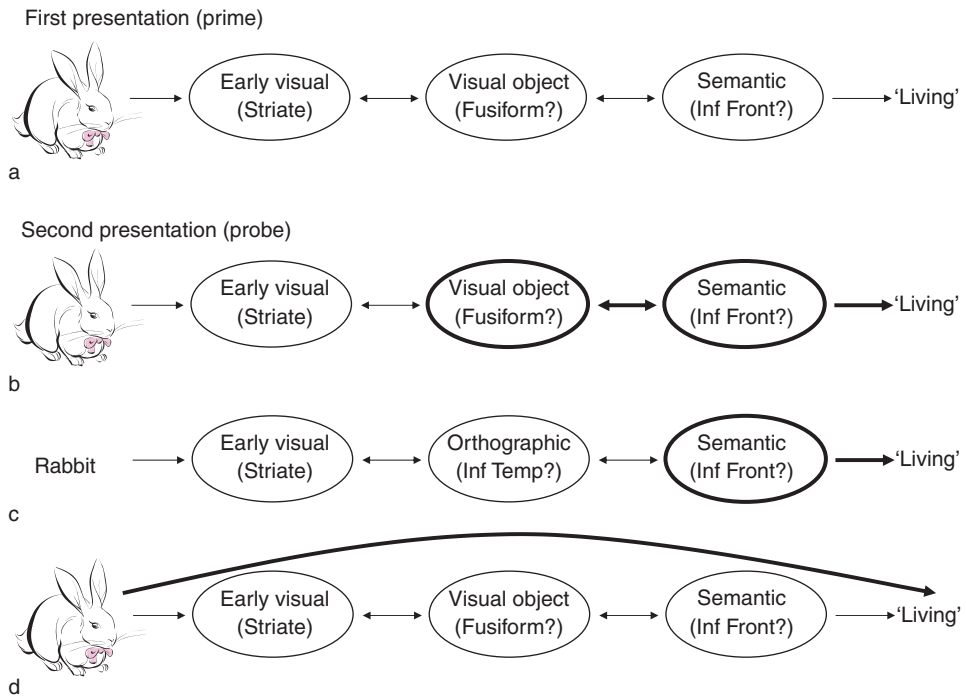

Figure 4 Schematic of the component process view of priming. (a) Initial presentation of an object in a semantic classification task (living vs. nonliving) involves interactions between a number of processing stages between the stimulus (left) and response (right). Putative brain regions associated with each stage are shown. (b) Repetition of the same object in the same task causes faster classification by virtue of facilitated processing in some (but not necessarily all) of the stages (facilitation indicated by darker lines). (c) A change in the stimulus involves re-processing in only one of the original stages (in this example), leading to reduced priming (and potential isolation of semantic component). (d) The formation of direct a stimulus–response association (event record) can also cause priming, even though it bypasses the initial processing stages. Inf, inferior; Front, frontal cortex; Temp, temporal.

brain regions in priming is the use of transcranial magnetic stimulation (TMS). In the only such study to date, either left inferior frontal cortex or a control site was stimulated approximately 300 ms after the initial presentation of a visual object. Frontal stimulation reduced the amount of priming for semantic classification of that object when it was repeated later (Figure 5(a)). Moreover, repetition suppression for that object, when later repeated, was abolished in left inferior frontal cortex, but not in occipital cortex (Figure 5(b)). This supports the existence of at least two components to priming in this task: for example, a visual component in occipital cortex and a verbal component in left inferior frontal cortex.

Another important future direction is the combination of multiple imaging modalities, such as the combination of spatial information in fMRI with the temporal information in MEG. Once the detailed time course of activity can be estimated in different brain regions, one can go further and explore changes in the covariance between these time courses, in order to infer directionality and possible causality. It will also be useful to acquire more fMRI and EEG/MEG data on amnesic patients, to confirm that priming-related changes in brain activity persist despite compromised activity in MTL.

From a theoretical perspective, more detailed models of the causes of priming are needed. According to the 'component process view' mentioned above, an important question is: What is the nature of processes that lead to facilitation when they are re-engaged, in a stimulus-specific manner, in relation to other processes that may not show such facilitation? And how does this component process view incorporate other evidence suggesting a role for direct stimulus–response associations, or episodic records? Some computational models have been used to fit behavioral priming; such models now need to be extended to the neural level, to make contact with the imaging and electrophysiological data described above. One important question for these models is: Is there a single neural mechanism underlying priming, or multiple mechanisms operating on different spatial and temporal scales (e.g., long-term potentiation/depression, or synaptic depression, or even firing-rate adaptation)?

A final important question is how priming and explicit memory, if truly dissociable, nonetheless interact in many circumstances. Some have suggested that the cause of priming (often described as increased 'fluency' of processing) can, under some conditions, be utilized by people in order to make an explicit memory judgment. In other words, people can

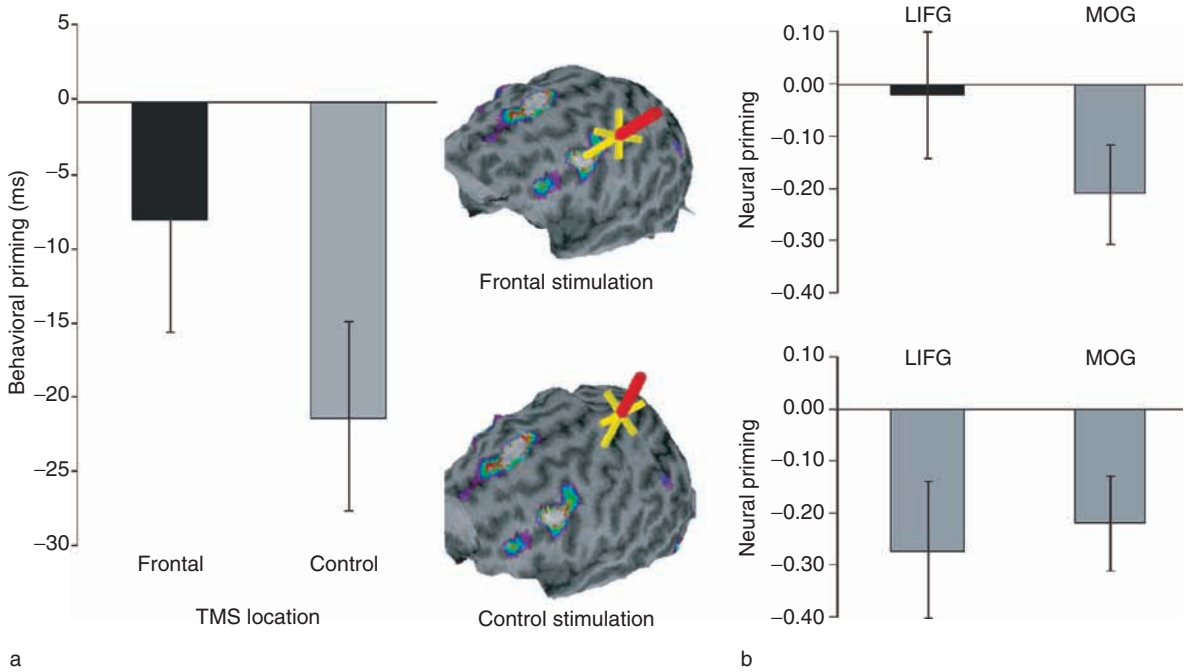

Figure 5 A transcranial magnetic stimulation (TMS) study of priming. (a) Using a semantic classification task like in **Figure 4**, TMS of left inferior frontal cortex (but not of a control region in left motor cortex) shortly after initial presentation of visual objects reduced subsequent behavioral priming. (b) TMS of the frontal region abolished fMRI repetition suppression in that region when the object was subsequently repeated but did not abolish repetition suppression in occipital cortex. LIFG, left inferior frontal gyrus; MOG, middle occipital gyrus. Adapted by permission of Macmillan Publishers Ltd: *Nature Neuroscience*. (Wig GS, Grafton ST, Demos, and Kelley WM (2005) Reductions in neural activity underlie behavioral components of repetition priming. *Nature Neuroscience* 9: 1228–1233), copyright (2005).

attribute the feeling of faster/easier processing of a stimulus to a past encounter with that stimulus. A current controversy is whether amnesic patients can be encouraged or even taught to use such fluency cues to improve their memory judgments.

See also: Cognition: An Overview of Neuroimaging Techniques; Declarative Memory System: Anatomy; Electroencephalography (EEG); Memory Representation; Memory Disorders; Multiple Memory Systems.

Further Reading

Badgaiyan RD and Posner MI (1997) Time course of cortical activations in implicit and explicit recall. *Journal of Neuroscience* 17: 4904–4913.

Bowers JS and Marsolek CJ (2003) *Rethinking Implicit Memory*. Oxford: Oxford University Press.

Buckner RL, Petersen SE, Ojemann JG, et al. (1995) Functional anatomical studies of explicit and implicit memory retrieval tasks. *Journal of Neuroscience* 15: 12–29.

Dhond RP, Buckner RL, Dale AL, Marinkovic K, and Halgren E (2001) Spatiotemporal maps of brain activity underlying word generation and their modification during repetition priming. *Journal of Neuroscience* 10: 3564–3571.

Gabrieli JDE (1998) Cognitive neuroscience of human memory. *Annual Review of Psychology* 49: 87–115.

Graf P, Squire LR, and Mandler G (1984) The information that amnesic patients do not forget. *Journal of Experimental Psychology: Learning, Memory and Cognition* 10: 164–178.

Grill-Spector K, Henson R, and Martin A (2006) Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Science* 10: 14–23.

Henson RNA (2003) Neuroimaging studies of priming. *Progress in Neurobiology* 70: 53–81.

Jacoby LL (1983) Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory and Cognition* 9: 21–38.

Logan GD (1990) Repetition priming and automaticity: Common underlying mechanisms? *Cognitive Psychology* 22: 1–35.

Moscovitch M (1994) Memory and working with memory: Evaluation of a component process model and comparisons with other models. In: Schacter DL and Tulving E (eds.) *Memory Systems*, pp. 233–268. Cambridge: MIT Press.

Roediger HL and McDermott KB (1993) Implicit memory in normal human subjects. In: Boller F and Grafman J (eds.) *Handbook of Neuropsychology*, vol. 8, pp. 63–161. Amsterdam: Elsevier.

Schacter DL (1987) Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory and Cognition* 13: 501–508.

Schott BH, Henson RN, Richardson-Klavehn A, et al. (2005) Redefining implicit and explicit memory: The functional neuroanatomy of priming, remembering and control of retrieval. *Proceedings of the National Academy of Sciences of United States of America* 102: 1257–1262.

Schott B, Richardson-Klavehn A, Henson R, et al. (2006) Neuroanatomical dissociationof encoding processes related to priming and explicit memory. *Journal of Neuroscience* 26: 792–800.

Squire LR (1987) *Memory and Brain*. New York: Oxford University Press.

Tenpenny PL (1995) Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin and Review* 2: 339–363.

Wig GS, Grafton ST, Demos K, and Kelley WM (2005) Reductions in neural activity underlie behavioral components of repetition priming. *Nature Neuroscience* 9: 1228–1233.

Relevant Website

http://en.wikipedia.org – Memory. In Wikipedia: The Free Encyclopedia.

Prion Diseases

A Aguzzi, University Hospital of Zurich, Zurich, Switzerland

© 2009 Elsevier Ltd All rights reserved.

Introduction

In the 1980s, the prion was defined as the infectious agent that causes transmissible spongiform encephalopathies (TSEs), or prion diseases, following decades of work by Tikvah Alper and JS Griffith in the 1960s and by then by Nobel Prize winner Stanley B Prusiner, whose team announced in 1982 that they had purified material containing the infectious agent, which was a protein. According to all evidence available to date, the prion (from 'proteinaceous infectious only') is devoid of informational nucleic acids and consists mainly of a prion protein (PrP), a cellular protein of basically unknown function, highly expressed on cells of the nervous system as well as on many other cells of the body, such as lymphocytes. The human gene is designated *PRNP*; its cognate gene, *Prnp*, was identified in hamster by Charles Weissmann and, independently, in mice by Bruce Chesebro. PrP exists in at least two different isoforms. The fundamental event in prion diseases seems to be a conformational change which occurs during the conversion of one form, named PrP^{C} (cellular, normal host prion protein), into another form, named PrP^{Sc} (scrapie-associated, pathological, potentially 'infectious' prion protein).

PrP^{C} has been identified in all mammals and birds examined to date, and it is anchored to the external surface of cells by a glycolipid moiety. Its function is unknown. All attempts to identify a posttranslational chemical modification that distinguishes PrP^{Sc} from PrP^{C} have been unsuccessful to date. PrP^{C} contains ~45% α-helix and is virtually devoid of β-sheet. Conversion to PrP^{Sc} creates a protein that contains ~30% α-helix and 45% β-sheet. The mechanism by which PrP^{C} is converted into PrP^{Sc} remains unknown, but PrP^{C} seems to bind to PrP^{Sc} and to form an intermediate complex during the formation of nascent PrP^{Sc}. Transgenic (Tg) mouse studies have demonstrated that PrP^{Sc} in the inoculum interacts preferentially with homotypic PrP^{C} during the propagation of prions (Tg mice are genetically modified mouse strains expressing at least one foreign or altered gene).

Prnp is transcribed in brain and many other cells of the body in both prion-infected and noninfected animals. Weissmann demonstrated in the 1990s that removal of *Prnp* from the mouse genome leads to complete immunity from prions. The only organ system in which histopathological damage can be demonstrated as a consequence of infection with prions – causing corresponding clinical signs – is the nervous system. This applies to all human TSEs, such as Creutzfeldt–Jakob disease (CJD), Gerstmann–Sträussler–Scheinker (GSS) syndrome, kuru, and fatal familial insomnia (FFI), and to all known prion encephalopathies of animals. The latter diseases comprise scrapie in sheep and goats, bovine spongiform encephalopathy (BSE; formation of sponge-like vacuoles within the gray matter of the brain is a highly characteristic morphological correlate in all TSEs), and chronic wasting diseases of mule, deer, and exotic ungulates. While the prototype of all prion diseases, scrapie, has been known from more than two centuries, BSE was first recognized as an epizootic phenomenon in 1986. The emergence of a new variant form of Creutzfeldt–Jakob disease (vCJD) in young people in the UK has since raised the possibility that BSE has spread to humans by dietary exposure. Experimental evidence claiming that the agent causing BSE is indistinguishable from the vCJD agent has recently supported this view.

The 'Protein-Only' Hypothesis

Although the exact physical nature of the transmissible agent is still controversial, a very large body of experimental data supports the 'protein only' hypothesis, first outlined in general terms by Griffith in 1967 and enunciated in an updated and detailed form by Prusiner. This hypothesis posits that the agent is devoid of nucleic acid and consists solely of an abnormal conformer of the cellular prion protein, PrP^{C}.

Conversion of PrP^{C} into PrP^{Sc}

Accumulation of an abnormal isoform (PrP^{Sc}) of the host-encoded PrP^{C} in the central nervous system is a hallmark of prion diseases. Because no chemical differences have been found between PrP^{C} and PrP^{Sc}, the two species are believed to differ in their conformation. The fine 3-D structure of PrP^{C} has been resolved at the atomic level using both nuclear magnetic resonance spectroscopy and X-ray diffraction crystallography. Only low-resolution structural data existed for PrP^{Sc} until recent work extended the analyses to electron crystallography. The β-sheet content of PrP^{Sc} is high, whereas PrP^{C} contains only a very short β-sheet stretch. Conversion of PrP^{C} into PrP^{Sc} in prion-infected cells is known to be a late posttranslational process, occurring after PrP^{C} has reached its normal extracellular location.

In order to explain the mechanism by which a misfolded form of PrP might induce the refolding of 'native,' normal PrP molecules into the abnormal conformation, two distinct models have been postulated: (1) the template assistance, or 'refolding,' model and (2) the nucleation–polymerization, or 'seeding,' model. In the first model, the conformational change is kinetically controlled; a high activation energy barrier prevents spontaneous conversion at detectable rates. Interaction with exogenously introduced PrP^{Sc} causes PrP^{C} to undergo an induced conformational change to yield PrP^{Sc}. This reaction may involve extensive unfolding and refolding of the protein to explain the postulated high energy barrier and could be dependent on an enzyme or chaperone. In the second model, PrP^{C} and PrP^{Sc} are in equilibrium strongly favoring PrP^{C}. PrP^{Sc} is stabilized only when it adds onto a crystal-like aggregate of PrP^{Sc}, acting as a seed in a nucleation-dependent polymerization process. Cell-free conversion studies suggest that PrP^{Sc} aggregates are able to convert PrP^{C} into a protease-resistant PrP isoform.

The Prion Protein

The cellular prion protein PrP^{C} contains two N-linked complex-type oligosaccharides at positions 181 and 197. Accordingly, Western blot analysis of PrP reveals three major bands, reflecting PrP that has two, one, or no glycosylation signals occupied. Upon translocation of PrP into the endoplasmic reticulum, an NH_2-terminal secretory signal peptide of 22 amino acids is cleaved from PrP, and 23 COOH-terminal residues are processed during addition of the glycosylphosphatidylinositol (GPI) anchor to the amino acid residue serine at position 231. Mature mouse PrP contains 209 amino acids.

PrP^{Sc}, which accumulates in TSE-affected organisms, is a modified form of PrP^{C}. PrP^{C} and PrP^{Sc} have an identical amino acid sequence and share the same posttranslational modifications as assessed by currently available methodology, but differ in their secondary, tertiary, and (presumably) quaternary structure. The physiological isoform PrP^{C} is protease sensitive (consequently sometimes designated as PrP^{sen}), while the pathological isoform PrP^{Sc} is partially protease resistant (thus also called PrP^{res}). The protease-resistant core of PrP^{Sc}, designated PrP^{27-30} according to its molecular weight on sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE), provides a specific and reliable molecular marker for the presence of the infectious agent.

Because of its location at the outer surface of cells, anchored by phosphatidylinositol glycolipid, PrP is a candidate for signaling, cell adhesion, or perhaps even for some transport functions. PrP is expressed on many cell types, including neurons, glial cells, and lymphocytes, and appears to be developmentally regulated during mouse embryogenesis. All prion proteins investigated hitherto contain four to five 'octa repeats,' repetitive sequences of eight amino acids. Amplification of the number of octa repeats has been found in hereditary prion diseases such as familial Creutzfeldt–Jakob disease (fCJD) and GSS syndrome. Several studies have established that mammalian and chicken prion protein octa repeats bind up to four copper atoms. It has been suggested from studies with peptides based on the octa repeat region of PrP that the four highly conserved histidines in the repeat region coordinate four copper atoms, indicating that binding of copper may be essential for the function of PrP^{C}. PrP refolded with copper has been claimed to have an antioxidant activity similar to that of superoxide dismutase, but these claims were not independently reproduced and there is no evidence that PrP^{C} acts as a superoxide dismutase *in vivo*.

Although PrP is predominantly found in brain tissue, high levels are also present in heart, skeletal muscle, and kidney, whereas it is barely detectable in the liver. Several candidate proteins that bind PrP^{C} have been reported. Among them are a member of the amyloid precursor protein (APP) family (the amyloid precursor-like protein 1), the antiapoptotic protein bcl-2, and the 37-kDa human laminin receptor precursor. The observation that PrP binds to amyloid precursor-like protein 1 (APLP1) is intriguing in light of potential analogies of transmissible spongiform encephalopathies to Alzheimer's disease. Evidence that any of these interactions are of physiological relevance is, however, still lacking.

The Gene Encoding PrP

The gene encoding the cellular isoform of the prion protein is located on chromosome 20 in humans and on chromosome 2 in mice, and is termed *Prnp* in mouse and *PRNP* in humans. In all known mammalian and avian genes, PrP^{C} is encoded by a single open reading frame. The *Prnp* genes of mouse, and *PRNP* of sheep and cattle, contain three exons, while these genes in hamsters and humans span over two exons. PrP expression seems to be controlled by the SP1 transcription factor that binds to G+C-rich nonamer regions upstream of exon 1. Within the genome, *Prnp* is present as a single copy gene. A partially homologous gene termed *Prnd*, and encoding the Dpl protein ('downstream of the *Prnp* locus') was identified approximately 16 kb 3′ of the *Prnp* locus. To date, no evidence has surfaced that would incriminate Dpl in the pathogenesis of transmissible prion diseases,

and the physiological function of Dpl seems to be related to male reproduction.

Prnp mRNA is constitutively expressed in the adult organism and is developmentally regulated during mouse embryogenesis. Mutations in the host-encoded PrP gene were shown to be genetically linked to human prion diseases. In addition, polymorphisms in the PrP gene control many features of prion diseases, including incubation time, species barrier, and specificity of recovered prion strain. Mice that lack PrP^C (designated $Prnp^{0/0}$ or $Prnp^{-/-}$) are resistant to infection with prions.

$Prnp^{0/0}$ mice with an N-proximally truncated transgene encoding PrP devoid of the octa repeats and the conserved 106–126 region suffer from ataxia and degeneration of the cerebellar granule cells shortly after birth. Similar (but not identical) phenotypes are observed in mice overexpressing Dpl; these mice experience degeneration of cerebellar Purkinje cells. Introduction of a single intact PrP allele largely rescues this phenotype. These findings may be explained by postulating that PrP interacts with a ligand to obtain an essential signal. An alternative explanation posits that Dpl or N-terminally truncated PrP^C can exist as transmembrane proteins and form oligomers. Lack of the N terminus may create a toxic channel that would be patent for ions and lead to osmotic or excitotoxic damage. Single molecules of PrP^C containing its N-terminal tail may block such channels in a transdominant fashion, and thereby abrogate neurotoxicity.

The Species Barrier

Transmission of prions from one species to another is usually accompanied by a prolongation of the incubation period in the first passage and incomplete penetrance ('attack rate') of the disease. This phenomenon is called the 'species barrier.' Subsequent passage in the same species occurs with high frequency and shortened incubation times. In the case of prion transmission from hamster to mice, this species barrier was overcome by introducing hamster *Prnp* transgenes into recipient wild-type mice. Crucially, the properties of the prions produced were compatible with the prion species used for inoculation: infection with hamster prions led to production of hamster prions, whereas infection with mouse prions gave rise to mouse prions. With respect to the protein-only hypothesis these findings can be interpreted as follows: hamster PrP^C but not murine PrP^C (the latter differing from the former by 10 amino acids) is an appropriate substrate for conversion to hamster PrP^{Sc} by hamster prions, and vice versa. Intriguingly, susceptibility of mice to prions from other species is increased when the corresponding PrP transgene is introduced into a PrP knockout mouse, suggesting that the resident murine gene inhibits the propagation of the alien prions (knockout mice are genetically engineered mice lacking the ability to express a specific protein due to a deletion of a gene of interest).

Understanding Prion Strains

In 1961, distinct isolates or strains of prions were first described in scrapie-diseased goats, in which two dissimilar clinical manifestations ('scratching' and 'drowsy,' respectively) were identified. These strains differ in their incubation times in various inbred mouse lines and by their lesion pattern in the brain. Strikingly, distinct strains of prions can be propagated in an inbred mouse strain that is homozygous with respect to *Prnp* – a perplexing finding with regard to the protein-only hypothesis, meaning that an identical polypeptide chain is able to mediate different strain phenotypes. Both the 'refolding' and the 'seeding' models propose that each strain is associated with a distinct conformation of PrP^{Sc} and that each of these can convert PrP^C of its host into a likeness of itself. Indeed, PrP^{Sc} species associated with two hamster-adapted scrapie strains, hyper (HY) and drowsy (DY), proved to display characteristic clinical and histopathological properties as well as distinct biochemical patterns with respect to proteinase K digestion – which may be explainable by the presence of different conformations of PrP^{Sc}. Analogous findings have been made with other prion strains propagated in mice. In addition, PrP^{Sc} of certain strains differ in the ratio of the diglycosylated to the monoglycosylated form. Based on the fragment size and the relative abundance of individual bands, three distinct patterns (PrP^{Sc} types 1–3) were defined for sporadic and iatrogenic CJD cases. In contrast, all cases of vCJD displayed a novel pattern, designated type 4 by Collinge, and type III by Parchi and Gambetti. Interestingly, brain extracts from BSE-infected cattle as well as from BSE-inoculated macaques exhibited a type 4 pattern. Even more intriguingly, transmission of BSE or vCJD to mice produced mouse PrP^{Sc} with a type 4 pattern indistinguishable from the original inoculum – a finding that impressively supports the hypothesis that vCJD is the human counterpart to BSE. An additional tool, which may eventually serve to characterize prion strains, utilizes the differential affinity of certain antibodies for properly folded and denatured isoforms of PrP, thereby differentiating as many as eight different conformers. This provides further circumstantial evidence for the assumption that strain specificity is encrypted within the physical structure of PrP.

Studies of scrapie in inbred mice led to identification of a single autosomal gene, termed *Sinc* (*scrapie incubation*), controlling incubation time of mice infected with mouse prion strains. Soon after, a link between *Sinc* and a polymorphism in the mouse *Prnp* gene was discovered. Elegant knockin experiments, in which *Prnp* alleles of inbred mice were swapped, have provided strong support for the contention that *Sinc* is identical to *Prnp*.

Spread of Prions

Prion pathogenesis can be broken down into spatially and temporally distinct phases: (1) infection and peripheral replication, (2) migration from the periphery to the central nervous system (CNS; neuroinvasion), and (3) neurodegeneration. The resistance to prions by mice that lack PrP^C expression is amply documented. PrP^C expression is required for transporting the infectious agent from the peripheral sites to the CNS (as monitored by PrP^C-expressing neurografts) and within the CNS. However, reconstitution of $Prnp^{0/0}$ mice with wild-type (wt) bone marrow is insufficient to restore neuroinvasion in engrafted $Prnp^{0/0}$, mice although the capacity of the spleen to accumulate prions of the Rocky Mountain Laboratory (RML) strain (derived from sheep with scrapie) is reconstituted. This suggests that hematopoietic cells transport prions from the entry site to the lymphoreticular system (LRS), which accumulates and replicates prions, but that PrP^C expression in an additional compartment, presumably the peripheral nervous system, is required. B lymphocytes (not necessarily expressing PrP^C) are crucial for peripheral prion spread and neuroinvasion.

The dependence on lymphotoxin (LT)-mediated signaling by B cells may explain – at least in part – the requirement for B cells in peripheral pathogenesis: follicular dendritic cells (FDCs) accumulate PrP^{Sc} following scrapie infection, and maturation of FDCs requires signaling by B cells expressing LT-α/LT-β trimers on their surface. Indeed, blockade of LT-β signaling via administration of soluble LT-β receptor immunoglobulin (LT-βR–Ig) fusion protein ablates mature FDCs and significantly impairs neuroinvasion and accumulation of peripheral PrP^{Sc} and infectivity. FDCs are crucial to disease progression after oral scrapie challenge, but only within a short time window.

FDCs play a role in antigen trapping and in binding opsonized antigens to the CD21/CD35 complement receptors. Two studies have demonstrated that the complement system is relevant to prion pathogenesis. Mice genetically engineered to lack complement factors or mice depleted of the C3 complement component exhibited enhanced resistance to peripheral prion inoculation. Because FDCs are most likely immobile cells, they are unlikely to be responsible for prion transport into the CNS.

But just which cell types are involved in neuroinvasion? The innervation pattern of lymphoid organs is primarily sympathetic. Sympathectomy delays the onset of scrapie, while sympathetic hyperinnervation enhances splenic prion replication and neuroinvasion, suggesting that innervation of secondary lymphoid organs is the rate-limiting step to neuroinvasion. Although there is no physical contact between FDCs and sympathetic nerve endings, the distance between FDCs and splenic nerves affects the velocity of neuroinvasion. It remains to be determined whether this results from passive diffusion of prions or whether mobile cells (e.g., germinal center B cells) are involved in an active transport process.

Inflammation: License to Replicate?

Because lymphoid infectivity is found in most prion diseases, and proinflammatory cytokines and immune cells are involved in lymphoid prion replication, it has been of highest interest to test whether chronic inflammatory conditions in nonlymphoid organs can affect prion distribution. Inclusion body myositis, which is an inflammatory disease of muscle, is reported to lead to strong PrP^{Sc} deposits in muscle.

Indeed, many chronic inflammatory conditions, some of which are very common, including rheumatoid arthritis, type 1 diabetes, and Crohn's, Hashimoto's, and chronic obstructive pulmonary diseases, show organized inflammatory foci of B lymphocytes, FDCs, dendritic cells (DCs), macrophages, and other immune cells associated with germinal centers. Extranodal metastases of Hodgkin's disease and non-Hodgkin's lymphomas may contain neoplastic follicles with FDCs. Moreover, the meninges can also develop ectopic lymphoid follicles under conditions of chronic inflammation. Most importantly, tertiary follicles can be induced in nonlymphoid organs by naturally occurring infections in free-ranging ruminants, and are surprisingly prevalent in nonlymphoid organs.

To investigate whether inflammatory diseases influence prion pathogenesis, mice with a variety of five inflammatory diseases of kidney, pancreas, or liver were inoculated with the RML strain. In all cases and in all organs tested, chronic lymphocytic inflammation enabled prion accumulation in otherwise prion-free organs. Inflammatory foci consistently correlated with upregulation of LTs and ectopic induction of PrP^C-expressing FDC-$M1^+$ cells. Inflamed organs of mice lacking LT-αRs or LT-βRs neither accumulated PrP^{Sc} nor showed infectivity upon prion inoculation. Hence scrapie infection of mice suffering from

nephritis, hepatitis, or pancreatitis induces unexpected prion deposits at the sites of inflammation.

The aforementioned results raised concerns that analogous phenomena might occur also in farm animals, since these are in constant contact with inflammogenic pathogens. Indeed, we found that naturally scrapie-sick sheep suffering from mastitis contain PrP^{Sc} in their mammary glands. Therefore, also in sheep, inflammatory conditions induce accumulation and replication of prions in an organ believed to be previously free of prions. This finding may have an impact on the risk assessment of biologicals (e.g., milk) and could lead to a readjustment of current rules.

Outlook

There is a daunting amount of experimental and theoretical evidence supporting the proposition that the prion (i.e., the transmissible entity of TSEs) is composed of a self-propagating isoform of PrP, and that informational nucleic acids are not required for transmissibility. Many questions remain concerning the portals of entry of prions into the body of the host, and many processes still need to be elucidated, such as the mechanisms and the requirements for the conversion reaction, the transport of prions from the periphery to the CNS, and the mechanisms of pathogenesis. From a clinical point of view, early diagnosis of prion disease and treatments aiming to arrest or cure prion disease in humans are important goals.

See also: Infectious Agents in Neurodegenerative Disease; Prion Transport; Protein Folding and the Role of Chaperone Proteins in Neurodegenerative Disease; Variant Creutzfeldt–Jakob Disease.

Further Reading

Aguzzi A (2004) Understanding the diversity of prions. *Nature Cell Biology* 6: 290–292.

Aguzzi A and Polymenidou M (2004) Mammalian prion biology. One century of evolving concepts. *Cell* 116: 313–327.

Aguzzi A and Weissmann C (1997) Prion research: The next frontiers. *Nature* 389: 795–798.

Basler K, Oesch B, Scott M, et al. (1986) Scrapie and cellular PrP isoforms are encoded by the same chromosomal gene. *Cell* 46: 417–428.

Behrens A and Aguzzi A (2002) Small is not beautiful: Antagonizing functions for the prion protein PrP(C) and its homologue Dpl. *Trends in Neurosciences* 25: 150–154.

Behrens A, Genoud N, Naumann H, et al. (2002) Absence of the prion protein homologue Doppel causes male sterility. *EMBO Journal* 21: 3652–3658.

Bessen RA and Marsh RF (1994) Distinct PrP properties suggest the molecular basis of strain variation in transmissible mink encephalopathy. *Journal of Virology* 68: 7859–7868.

Blättler T, Brandner S, Raeber AJ, et al. (1997) PrP-expressing tissue required for transfer of scrapie infectivity from spleen to brain. *Nature* 389: 69–73.

Brandner S, Isenmann S, Raeber A, et al. (1996) Normal host prion protein necessary for scrapie-induced neurotoxicity. *Nature* 379: 339–343.

Castilla J, Saa P, Hetz C, et al. (2005) *In vitro* generation of infectious scrapie prions. *Cell* 121: 195–206.

Castilla J, Saa P, and Soto C (2005) Detection of prions in blood. *Nature Medicine* 11: 982–985.

Chesebro B, Race R, Wehrly K, et al. (1985) Identification of scrapie prion protein-specific mRNA in scrapie-infected and uninfected brain. *Nature* 315: 331–333.

Griffith JS (1967) Self-replication and scrapie. *Nature* 215: 1043–1044.

Heikenwalder M, Zeller N, Seeger H, et al. (2005) Chronic lymphocytic inflammation specifies the organ tropism of prions. *Science* 307: 1107–1110.

Hill AF, Desbruslais M, Joiner S, et al. (1997) The same prion strain causes vCJD and BSE [letter]. *Nature* 389: 448–450.

Jarrett JT and Lansbury PT Jr. (1993) Seeding "one-dimensional crystallization" of amyloid: A pathogenic mechanism in Alzheimer's disease and scrapie? *Cell* 73: 1055–1058.

Klein MA, Frigg R, Flechsig E, et al. (1997) A crucial role for B cells in neuroinvasive scrapie. *Nature* 390: 687–690.

Klein MA, Kaeser PS, Schwarz P, et al. (2001) Complement facilitates early prion pathogenesis. *Nature Medicine* 7: 488–492.

Ligios C, Sigurdson CJ, Santucciu C, et al. (2005) PrP(Sc) in mammary glands of sheep affected by scrapie and mastitis. *Nature Medicine* 11: 1137–1138.

Meier P, Genoud N, Prinz M, et al. (2003) Soluble dimeric prion protein binds PrP(Sc) *in vivo* and antagonizes prion disease. *Cell* 113: 49–60.

Montrasio F, Frigg R, Glatzel M, et al. (2000) Impaired prion replication in spleens of mice lacking functional follicular dendritic cells. *Science* 288: 1257–1259.

Oesch B, Westaway D, Walchli M, et al. (1985) A cellular gene encodes scrapie PrP 27-30 protein. *Cell* 40: 735–746.

Prinz M, Heikenwalder M, Junt T, et al. (2003) Positioning of follicular dendritic cells within the spleen controls prion neuroinvasion. *Nature* 425: 957–962.

Prusiner SB (1982) Novel proteinaceous infectious particles cause scrapie. *Science* 216: 136–144.

Prusiner SB, Scott M, Foster D, et al. (1990) Transgenetic studies implicate interactions between homologous PrP isoforms in scrapie prion replication. *Cell* 63: 673–686.

Riek R, Hornemann S, Wider G, et al. (1996) NMR structure of the mouse prion protein domain PrP(121–321). *Nature* 382: 180–182.

Saborio GP, Permanne B, and Soto C (2001) Sensitive detection of pathological prion protein by cyclic amplification of protein misfolding. *Nature* 411: 810–813.

Shmerling D, Hegyi I, Fischer M, et al. (1998) Expression of amino-terminally truncated PrP in the mouse leading to ataxia and specific cerebellar lesions. *Cell* 93: 203–214.

Stahl N, Borchelt DR, Hsiao K, et al. (1987) Scrapie prion protein contains a phosphatidylinositol glycolipid. *Cell* 51: 229–240.

Weissmann C and Aguzzi A (1999) Perspectives: Neurobiology. PrP's double causes trouble. *Science* 286: 914–915.

Prion Transport

S E Encalada and L S B Goldstein, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Axonal transport of vital cargo within neuronal cells is indispensable for the proper maintenance of cell viability and function. This process involves the active movement along neurites of essential materials such as signaling molecules, RNAs, synaptic vesicle precursors, vesicular and cytoskeletal components, and growth factors. These diverse components arrive at dendritic or axonal termini while certain neurotrophic and other signals travel back from synapses to the cell body in response to various stimuli. The proteins involved in executing fast transport are a number of microtubule-based kinesin and dynein molecular motor proteins which use ATP hydrolysis to transduce chemical energy into kinetic energy and movement along microtubules.

A number of neurodegenerative diseases have been linked to deficits in axonal transport. For example, disruptions in the transport of the amyloid precursor protein (APP) and huntingtin (Htt) result in accumulations of these and other proteins within axons and in phenotypes that are consistent with axonal transport defects in Alzheimer's and Huntington's diseases, respectively. Likewise, in transmissible spongiform encephalopathies (TSEs) or prion diseases such as mad cow disease, scrapie in sheep, and Creutzfeldt-Jacob disease in humans, abnormal intracellular accumulations are numerous. Whether axonal transport defects are also involved in prion pathologies is unknown. This article reviews evidence that the normal endogenous prion protein transports along axons and the putative role of this transport in synaptic function. Transport of infectious prions to the central nervous system (CNS) is necessary for disease pathology, thus this article also reviews current knowledge on the routes of neuroinvasion used by the infectious agent (PrP^{Sc}) to reach the CNS from extracerebral sites and a putative model for PrP^{Sc} entrance into cells.

Localization and Axonal Transport of the Mammalian Prion Protein, PrP^C

Prion diseases cause sporadic, familial, or infectious illnesses that are invariably fatal. In addition to spongiosis and astrocytosis, these diseases are also characterized by neuronal cell death. The main culprits, prions, are protein particles that exist in two forms. The normal nonpathological cellular prion protein (PrP^C) is a glycosyl-phosphatidylinositol (GPI) anchored protein that is widely expressed in a number of neural and nonneural tissues. This endogenous protein contains a number of α-helical domains that, by mechanisms that are still poorly understood, can posttranslationally convert to a β-sheet, protease-resistant, and infectious PrP^{Sc} (scrapie) conformation. It is thought that conversion from normal to scrapie forms necessitates the direct interaction of both types of proteins.

While both cellular and infectious isoforms localize to axonal termini and PrP^C has been shown to transport along axons by fast anterograde transport, the mechanisms underlying these processes have not been elucidated. Studies on the detailed localization of PrP^C in the brain include immunohistochemical analyses and quantitative studies of the ultrastructural localization of PrP^C. With these techniques, PrP^C particles have been found in a number of subcellular locations in the hippocampus, including the endoplasmic reticulum, Golgi complex, endosomes and lysosomes, tubules, vesicles without coats, clathrin-coated vesicles, plasma membrane, mitochondria, and nucleus, among others. In addition, these studies showed a distinctive presence of PrP^C in the neuropil of axonal and dendritic membranes as well as in synaptic and perisynaptic membrane profiles. Immunogold electron micrograph (EM) studies have revealed PrP^C localization mainly at presynaptic sites, and notably at asymmetric excitatory synapses in several brain regions.

The presence of PrP^C at the presynapse suggests that this protein is transported along the axon after it is produced in the cell body. Understanding the role and mechanisms of transport of the PrP^C along axons is important for understanding prion diseases because PrP^C is indispensable for the initiation and maintenance of prion infectivity. Moreover, while PrP^C localizes at synapses, its function remains unclear. Knowledge of PrP^C transport might clarify the role of PrP^C at synapses.

To date, studies on the axonal transport of PrP^C have been performed *in vivo* in the peripheral nervous system (PNS) and CNS using radiolabeled PrP^C. To determine whether PrP^C travels anterogradely at the whole nerve level, [^{35}S]methionine was injected near the dorsal root ganglia (DRG) of the sciatic nerve (in the PNS) and into layer II/III neurons of the entorhinal cortex in rats (in the CNS). These injection sites are located in the cell body layers of the respective tissues. Four hours after injection, radiolabeled PrP^C was immunopurified from distal axons of sciatic nerves and from the most rostral region of

the dentate gyrus. Sciatic nerves are composed primarily of sensory and motor neuron axons, but do not contain neuronal cell bodies, while neurons of layers II/III project to the outer molecular layer of the dentate gyrus in the perforant pathway. In both cases, radiolabeled PrP^C was detected in these remote axonal regions. To verify that detection of PrP^C in axons is the result of transport and not due to diffusion of [^{35}S] methionine, a ligature was applied adjacent to the DRG following injection of radiolabeled PrP^C. This ligation prevented the appearance of labeled PrP^C in remote parts of axons. The 4 h time interval between the injection of [^{35}S]methionine and the collection of distant axonal tissues with radiolabeled PrP^C is consistent with fast anterograde transport which is estimated at 200–400 mm day^{-1} (2.3–4.6 μm s^{-1}).

Using an accumulation paradigm (Figure 1), PrP^C has also been shown to move anterogradely and retrogradely in hamster sciatic nerves. The sciatic nerve was severed at the midsection (Figure 1). Six to 24 h later, proximal and distal regions of the nerve were collected. Proteins moving in a fast anterograde axonal pathway generally accumulate on the proximal side, while those moving in a fast retrograde fashion will accumulate on the distal side of the injury. The level of PrP^C from basal, proximal, and distal regions was analyzed using a two-site immunometric assay. In this assay, sciatic nerve homogenates are applied to wells pre-coated with an anti-PrP^C antibody. After an incubation period, another anti-PrP^C antibody coupled with acetylcholinesterase was added and incubated. Absorbance of wells was measured and PrP^C levels recorded. The absorbance for PrP^C in both the proximal and distal nerve segments was significantly higher as compared to that of the basal segment. This indicated that PrP^C moved in both antero- and retrograde directions. The isoform of the prion protein traveling along axons was a larger form (38 kDa vs. the normal 31–36 kDa size of the protein) not usually detected in axons under basal conditions. Interestingly, the accumulation of PrP^C at the distal end of the sciatic nerve was not readily detectable under standard tissue extraction conditions. An enhanced detection method involving the use of SDS buffer during tissue

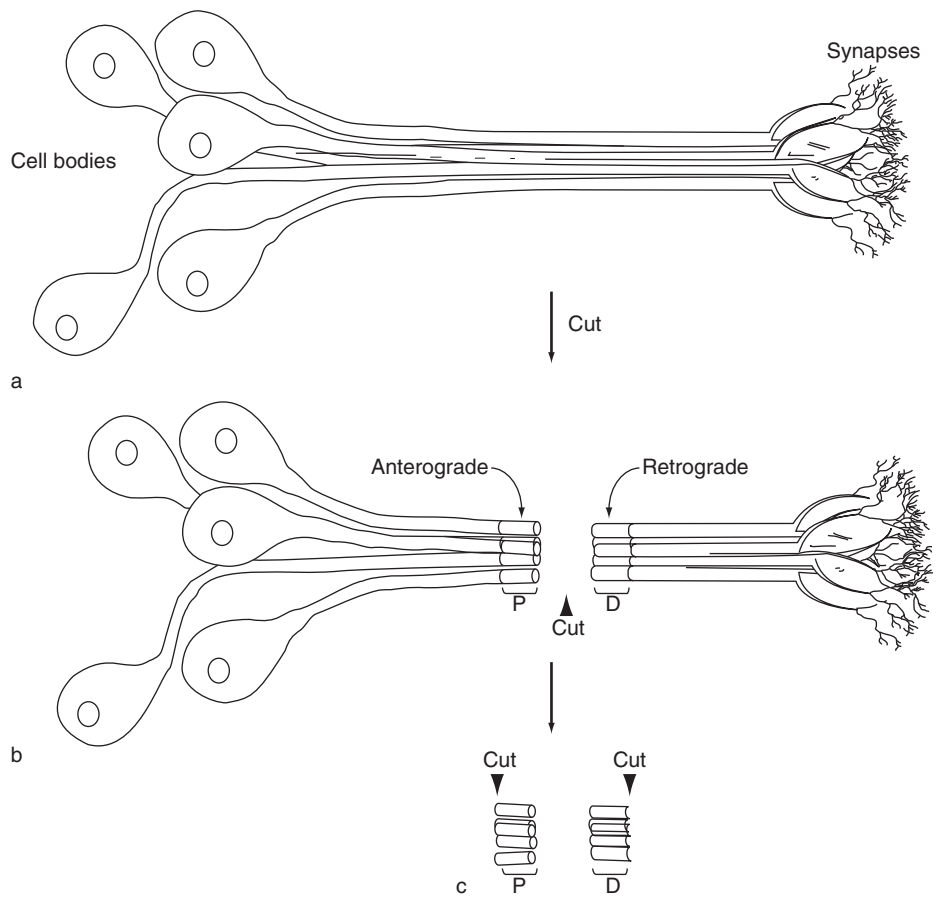

Figure 1 Schematic of sciatic nerve accumulation paradigm. (a) An intact nerve bundle prior to cutting. (b) Sciatic nerve after cut at midsection. Shown are the proximal (P) and distal (D) segments where anterograde and retrograde moving proteins accumulate after cut. (c) The proximal (P) and distal (D) segments are removed and analyzed.

homogenization allowed the detection of PrP^C in the distal site, suggesting lower levels of retrogradely traveling PrP^C as compared to that traveling anterogradely. Whether the retrogradely traveling PrP^C is derived from the uptake of PrP^C by presynaptic endings near muscle or other regions or whether anterograde PrP^C reverses direction and returns to the cell body is presently unknown.

Although PrP^C is expressed ubiquitously throughout the brain, its function has not been determined. The presence of PrP^C at synapses suggests a role there, but this role is unclear. Further studies of the axonal transport of PrP^C during nerve regeneration experiments suggest that this protein might be involved in axonal growth. In experiments in which the sciatic nerve of hamsters was crushed at a single position and then allowed to regenerate, the levels of PrP^C expression at the site of injury were found to increase during the period of regrowth of axons. PrP^C levels subsequently decreased to basal levels at the time of complete functional nerve recovery. The isoform of PrP^C with increased levels at the site of injury was the higher molecular weight 38 kDa form, suggesting that only axonally transported PrP^C might be involved in the growth of axons *in vivo*.

Fast anterograde and retrograde transport suggest that a microtubule-based process is involved in transporting the prion protein along axons. Treatment of cultured neuroblastoma cells with the microtubule depolymerizing drug nocodazole caused PrP^C to concentrate in the cytosol with little or no subcellular organelle distribution. Generalized disruption of kinesins and dyneins with agents such as AMP-PNPs and vanadate or with antibodies against unspecified kinesin and dyneins also appears to disrupt movement of PrP^C particles inside neuroblastoma cells. However, whether these effects are the result of specific interactions between PrP^C and microtubule-dependent motor proteins or are a secondary consequence of global disruption of the entire microtubule cytoskeleton has not been determined.

Axonal Transport of Scrapie, PrP^{Sc}

The spread of scrapie infectivity presupposes the presence of PrP^C. Thus, mice that lack PrP^C are immune to prion infections. Understanding the mechanisms of axonal transport of the PrP^C might illuminate how scrapie reaches the CNS and becomes pathological. However, the partially protease-resistant PrP^{Sc} has a distinctively different set of biological properties that distinguish it from PrP^C. For example, its turnover rate, subcellular localization, and distribution are different, and thus it is feasible that the trafficking of PrP^{Sc} might not necessarily involve a mechanism similar to that for PrP^C.

To date, only one study has followed the subcellular localization of fluorescently tagged PrP^{Sc} in cultured cells. Magalhães and colleagues purified infectious scrapie from mice brains and labeled it with a fluorescent probe. The labeled preparation was subsequently applied to the surface of SN56 cultured cells and of dissociated primary cortical cells from hamster brains. SN56 cultured cells are of septal neuronal origin and can be differentiated into neuron-like cells that have projections. Labeled PrP^{Sc} accumulated on the surface of SN56 and cortical primary cells as aggregates. To unequivocally determine the source of infection, small amounts of labeled PrP^{Sc} were directly applied to the surfaces of cells without application to cells in the immediate vicinity. Starting 3 days after exposure, fluorescently labeled PrP^{Sc} was internalized by both types of cells into vesicle-like compartments that moved within the cell body and within neurites at a rate of 0.3–5 $\mu m\, s^{-1}$, a rate consistent with fast axonal transport. These PrP^{Sc} puncta co-localized in fixed SN56 cells with a number of organelle markers, such as LysoTracker, which marks acidic compartments and lysosomes, and dextran, a marker for fluid-phase endocytosis. In contrast, PrP^{Sc} puncta did not co-localize with markers of lipid rafts (cholera toxin) or with early and recycling endosomal markers (transferrin). These data indicate that externally applied PrP^{Sc} sorts into lysosomal and late endosomal compartments within neuronal cells. Interestingly, application of labeled PrP^{Sc} in PrP knockout mice did not affect the internalization of PrP^{Sc}, indicating that an interaction between PrP^C and PrP^{Sc} is not necessary for internalization.

Neuroinvasion by the Scrapie Agent

While PrP^{Sc} has been detected in a number of extracerebral tissues, such as in skeletal muscle and lymphoid organs, histopathological lesions have been identified only in organs of the CNS. Thus, in the case of peripherally acquired prion infection (such as through the ingestion of contaminated tissue), PrP^{Sc} has to migrate to the CNS for pathological signs to appear. The movement of scrapie from extra-CNS sites to the CNS is termed neuroinvasion. Interestingly, in experimental laboratory models, the most effective method for transmission of scrapie infection is through the injection of the scrapie inoculate directly in recipients' brains. If transmission of scrapie is initiated by injection of inoculate extracerebrally, incubation times are usually significantly lengthened. This suggests that the scrapie agent takes time to travel from intake sites

(such as gastrointestinal sites), perhaps via the PNS, to tissues and organs of the CNS.

Transport of PrP^{Sc} in the CNS has been investigated *in vivo* by injecting prions intraocularly in mice. After injection, the first prion lesions were observed in the contralateral projections of the retino-tectal pathway. Based on these and other studies, the rate of scrapie spread was calculated to be around 1 mm/day. This rate is markedly slower than that mediated by fast axonal transport, which is estimated at 200–400 mm day^{-1}. However, this apparent discrepancy does not preclude a role for the axonal transport of PrP^{C} in the spread of scrapie because infectivity might be spread at lower-than-detectable levels through axons before reaching accumulation sites in the CNS where it would become more easily detectable. In fact, the rapid anterograde and retrograde axonal transport of PrP^{C} might suggest a possible mechanism for the spread of scrapie infectivity to the CNS. Because PrP^{C} and PrP^{Sc} must interact for conversion to occur, it is feasible that conversion could occur during transport. Whether spread of scrapie infectivity is mediated via axonal transport or through a domino-like effect by which PrP^{C} is converted to PrP^{Sc} in a step-wise fashion (with or without axonal transport) is still unclear.

Routes of Neuroinvasion of PrP^{Sc}

Neuroinvasion of PrP^{Sc} from peripheral tissues to the PNS and then to the CNS can be roughly classified as a two-step process. The first stage usually requires the accumulation and replication of PrP^{Sc} in 'reservoirs' of the lymphoreticular system (LRS). The second stage is transport of PrP^{Sc} within the PNS and then to the CNS. This is a loose classification, however, since it presumes the existence of the scrapie agent in or near the LRS, while the mechanism of transport to lymphoid organs from extraneuronal tissues such as gastrointestinal sites to LRS is not yet clear.

Although accumulation of PrP^{Sc} in lymphoid organs has not been found in all models of prion disease, this is a common early step in prion infectivity. The cells in the LRS responsible for accumulation and replication of prions are stromal cells such as the follicular dendritic cells (FDCs). Prions appear to accumulate specifically on the plasmalemma of FDCs. Thus, FDCs of the spleen of mice inoculated with prions have been observed to accumulate PrP^{Sc}, while nonstromal cells remain intact. Moreover, mice with impaired FDCs do not show signs of PrP^{Sc} infectivity in the spleen.

From the LRS, the mechanism by which prions reach the PNS has not been determined. Perhaps the location of PrP^{Sc} on the plasmalemma of FDCs positions scrapie near nerve fibers. However, expression of PrP^{Sc} only in neuronal cells or specific application of the infectious agent to nerve terminals results in active neuroinvasion, suggesting that accumulation in stromal cells can be bypassed. These observations suggest that a mechanism of transport independent of accumulation and replication in the LRS could operate during prion infection.

Within the PNS, the mechanisms of PrP^{Sc} transport are unclear. Several studies point to a requirement of the presence of PrP^{C} for scrapie transport. However, whether conversion of PrP^{C} to scrapie occurs during transport or in a domino-step manner is still to be determined. Also, PrP^{Sc} transport within the PNS and CNS has been studied *in vivo* only indirectly, by inoculation of mice with scrapie and then examination of different tissues at later times, or by estimating and comparing incubation times of mice that were inoculated intracerebrally versus those inoculated intraperitoneally. As mentioned earlier, slow rates of spread of prion infectivity have been found (of approximately 1 mm day^{-1}), suggesting that nonaxonal modes of transport of PrP^{Sc} might also be conceivable.

There is still much to be deciphered about the nature of scrapie infectivity and conversion. One of these unknown aspects is how PrP^{Sc} reaches the brain to produce disease. Whether via axonal or other routes, the mechanism of movement of scrapie within and between cells will provide us with a clearer understanding of infection routes and with possible methods to counteract the spread of infection. Knowledge of the function of endogenous PrP^{C} and why and how it travels along axons might also tell us something about scrapie infectivity.

See also: Axonal Transport Disorders; Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Axonal Transport and Neurodegenerative Diseases; Infectious Agents in Neurodegenerative Disease; Prion Diseases.

Further Reading

Bailly Y, Haeberle AM, Blanquet-Grossard F, et al. (2004) Prion protein (PrP^{C}) immunocytochemistry and expression of the green fluorescent protein reporter gene under control of the bovine PrP gene promoter in the mouse brain. *Journal of Comparative Neurology* 473: 244–269.

Borchelt DR, Koliatsos VE, Guarnieri M, et al. (1994) Rapid anterograde axonal transport of the cellular prion glycoprotein in the peripheral and central nervous systems. *Journal of Biological Chemistry* 269: 14711–14714.

DeArmond SJ, Mobley WC, DeMott DL, Barry RA, Beckstead JH, and Prusiner SB (1987) Changes in the localization of brain prion proteins during scrapie infection. *Neurology* 37: 1271–1280.

Ford MJ, Burton KJ, Morris RJ, and Hall SM (2002) Selective expression of prion protein in peripheral tissues of the adult mouse. *Neuroscience* 113: 177–192.

Fournier J-G, Escaig-Haye F, Billette de Villemeur T, and Robain O (1995) Ultrastructural localization of cellular prion protein (PrP^C) in synaptic boutons of normal hamster hippocampus. *C. R. Academy of Science Paris III* 318: 339–344.

Fraser H and Dickinson AG (1985) Targeting of scrapie lesions and spread of agent via the retino-tectal projection. *Brain Research* 346: 32–41.

Glatzel M, Giger O, Braun N, and Aguzzi A (2004) The peripheral nervous system and the pathogenesis of prion diseases. *Current Molecular Medicine* 4: 355–359.

Gunawardena S and Goldstein LSB (2003) Cargo-carrying motor vehicles on the neuronal highway: Transport pathways and neurodegenerative disease. *Journal of Neurobiology* 58: 258–271.

Hachiya NS, Watanabe K, Yamada M, Sakasegawa Y, and Kaneko K (2004) Anterograde and retrograde intracellular trafficking of fluorescent cellular prion protein. *Biochemical and Biophysical Research Communications* 315: 802–807.

Magalhães AC, Baron GS, Lee KS, et al. (2005) Uptake and neuritic transport of scrapie prion protein coincident with infection of neuronal cells. *Journal of Neuroscience* 25: 5207–5216.

Moya KL, Hässig R, Breen KC, Volland H, and Giamberardino LD (2005) Axonal transport of the cellular prion protein is increased during axon regeneration. *Journal of Neurochemistry* 92: 1044–1053.

Moya KL, Hässig R, Créminon C, Laffont I, and Di Giamberardino L (2004) Enhanced detection and retrograde axonal transport of PrP^C in peripheral nerve. *Journal of Neurochemistry* 88: 155–160.

Rodolfo K, Hässig R, Moya KL, Frobert Y, Grassi J, and Di Giamberardino L (1999) A novel cellular prion protein isoform present in rapid anterograde axonal transport. *Neuroreport* 10: 3639–3644.

Procedural Learning in Animals

M G Packard, Texas A&M University, College Station, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

According to the multiple memory systems hypothesis, the mammalian brain contains relatively independent neural systems that differ in terms of the type of memory they mediate. This hypothesis is supported by extensive neurobiological research, and several theories have been proposed describing the putative psychological operating principles that distinguish different types of memory. One prominent dual-memory theory draws a distinction between declarative and procedural memory. Declarative memory involves the acquisition, consolidation, and retrieval of relational information for facts and events that allows for flexibility in the expression of learned behaviors. Declarative memory processes appear to rely on a neuroanatomical system composed primarily of various structures in the medial temporal lobe, including the hippocampus. In contrast, procedural memory involves the acquisition, consolidation, and retrieval of invariant information that is expressed in an inflexible manner. Note that in lower animals, the distinction between declarative and procedural memory is similar in many aspects to that drawn by experimental psychologists who were strident proponents of either cognitive or habit theories of animal learning, respectively.

In lower animals, dissociable types of procedural memory are mediated by different neuroanatomical structures, including, for example, the dorsal striatum and amygdala. The dorsal striatum is critically involved in procedural learning and memory involving the formation of stimulus–response habits, whereas the amygdala may selectively mediate procedural learning involving the formation of stimulus–affect associations. This article briefly describes neurobiological evidence in lower animals supporting a role for the dorsal striatum and amygdala in these two forms of procedural learning.

The Dorsal Striatum and Procedural Learning: Lesion Studies

In rats, lesions of the dorsal striatum, a primary anatomical component of the basal ganglia, impair acquisition of several tasks that theoretically can be acquired using a procedural learning mechanism. Examples include one-way and two-way active avoidance behavior; auditory, tactile, and visual simultaneous discrimination learning; operant bar-press behavior; and straight-alley maze behavior. Moreover, consistent with a multiple systems approach to memory organization, evidence indicates that the dorsal striatum selectively mediates procedural learning. An early study that employed dissociation methodology to demonstrate a selective role of the dorsal striatum in procedural learning that involves the formation of stimulus–response habits involved an experiment using two eight-arm radial maze tasks. In a version of the task involving declarative learning and memory, rats obtained food rewards by visiting each arm of the maze once within a daily training session, and reentries into maze arms that were previously visited were scored as errors. Performance in the declarative radial maze task requires rats to remember those arms that have been previously visited within a daily training session. In a different version of the task, involving procedural learning and memory, rats obtained food rewards by visiting four randomly selected and illuminated maze arms twice within a daily training session, and visits to unlit maze arms were scored as errors. Performance in the procedural radial maze task requires rats to learn to approach lit maze arms, and this task is essentially a simultaneous visual discrimination that may involve acquisition of a stimulus–response habit. Lesions of the dorsal striatum severely impair acquisition of the procedural task and do not affect acquisition of the declarative task. Note that the two radial maze tasks used to dissociate the role of the dorsal striatum in procedural and declarative memory share the same motivational (appetitive), sensory (visual), and motor (maze running) characteristics, indicating that the selective impairment observed in the procedural memory task was not due to an influence on nonmnemonic factors.

An additional study used a plus maze task to investigate the selective role of the dorsal striatum in procedural memory. The plus maze apparatus is arranged so that a goal box (e.g., east or west), can be approached from one of two start boxes (e.g., north or south). In a dual-solution version of the task originally introduced by Edward Tolman and colleagues in the mid-1940s, rats were released from the same start box (e.g., south) and trained to obtain food from a consistently baited goal box (e.g., west). Following acquisition of the task, a probe trial was administered in which rats were released from the opposite start box (e.g., north); the choice behavior on this trial was used to assess the type of learning strategy the animals were employing. For example, rats trained in this task might learn the spatial location of the reinforcer and use this declarative or relational information to guide

approach behavior to the baited goal box. Alternatively, rats might learn to approach the baited goal box by acquiring a response tendency (i.e., a specific body turn) at the choice point of the maze and use this form of procedural learning during task acquisition. Note that although either a declarative or a procedural learning mechanism can be used to acquire this task successfully, the use of these types of mnemonic strategies predicts different choice behavior on the critical probe trial. Thus, rats that used declarative knowledge of the spatial location of the reinforcer (i.e., 'place' learning) should continue to approach the baited goal box on the probe trial, whereas rats using a specific body turn response, or procedural memory (i.e., 'response' learning), should choose the opposite goal box on the probe trial. The hypothesis that the dorsal striatum may play a selective role in the expression of procedural, or response, learning in the plus maze has been examined using a reversible brain lesion technique. Specifically, following the rats' acquisition training in the plus maze, and just prior to administration of a probe trial, investigators produced a reversible brain lesion via intracerebral injection of the sodium channel blocker lidocaine. On a probe trial given after 1 week of training in the plus maze, rats receiving saline vehicle injections into the dorsal striatum were predominantly place learners. Moreover, infusions of lidocaine into the dorsal striatum did not impair the expression of declarative, or place, learning. The lack of effect of neural inactivation of the dorsal striatum on the use of place learning is consistent with the radial maze findings described above. However, with extended training in the dual-solution plus maze task, rats eventually switched from the use of place learning to a response learning tendency. Thus, on a second probe trial administered after an additional week of training, rats receiving vehicle injections into the dorsal striatum were then predominantly response learners. In contrast, rats receiving intradorsal striatal injections of lidocaine prior to the second probe trial exhibited place learning, indicating a blockade of the expression of response learning. Again, consistent with the radial maze findings, these results indicate that the dorsal striatum is selectively involved in the expression of procedural learning. In addition, they suggest that the shift or transition from the use of place learning to response learning in a dual-solution plus maze task involves the gradual recruitment of a dorsal striatal-based procedural learning system to guide behavior. Lesions of the dorsal striatum also impair acquisition of a single-solution plus maze task that requires rats to use procedural response learning by varying the start point on each trial and reinforcing the same body turn at the choice point.

The selective role of the dorsal striatum in stimulus–response procedural learning and memory generalizes to aversively motivated behavior. For example, in a two-platform water maze task that requires rats to learn to approach a visible escape platform on the basis of either procedural memory (i.e., a consistent visual appearance but differing spatial location across trials) or declarative memory (i.e., a consistent spatial location but differing visual appearance across trials), lesions of the dorsal striatum selectively impair acquisition of the procedural memory task. In addition, in a single-platform water maze task in which rats are allowed swim to a visible escape platform that is always located in the same spatial location, the relative use of declarative and procedural learning can be assessed on a probe trial in which the cued platform is moved to a new spatial location. On a probe trial administered following extended training in this task, control rats exhibit a procedural memory strategy and swim directly to the visible platform in its new location. In contrast, rats with lesions of the dorsal striatum exhibit a declarative memory strategy and swim to the spatial location that the platform was previously located in.

An important aspect of dorsal striatal involvement in stimulus–response procedural learning concerns the existence of functional heterogeneity in the mnemonic functions of this brain structure. Specifically, lateral, but not medial regions of the dorsal striatum may be particularly critical in stimulus–response habit learning. For example, lesions of the medial dorsal striatum in rats can in some cases produce deficits in acquisition of declarative learning and memory tasks, although these deficits often appear transient in nature. In addition, lesions of the dorsomedial, but not the dorsolateral, striatum block the ability of reinforcer devaluation to reduce instrumental responding, suggesting that the dorsomedial striatum selectively mediates expression of behavior based on action-outcome associations in instrumental conditioning rather than stimulus–response habit formation.

In addition to evidence dissociating the mnemonic functions of medial and lateral regions of the dorsal striatum, other findings indicate that the role of the lateral dorsal striatum in habit learning is organized on the basis of the nature of the sensory information provided by cortical input. Thus, lesions of the ventrolateral dorsal striatum, an area that receives olfactory cortical input, impair procedural leaning involving olfactory, but not visual, sensory information. In contrast, lesions of the posteroventrolateral dorsal striatum, an area that receives visual cortical input, impairs procedural leaning involving visual, but not olfactory, sensory information. The impairing

effect of dorsolateral striatal lesions on response learning in the plus maze may reflect a role for the vestibular and kinesthetic information that this striatal region receives from the somatosensory cortex. Thus, two levels of functional heterogeneity in the mnemonic functions of the dorsal striatum may be present. A medial-to-lateral anatomical gradient may correspond to a differential involvement in declarative and procedural memory, respectively. Second, within the lateral dorsal striatum, stimulus–response procedural learning is organized on the basis of the nature of the specific sensory input that is received from different cortical regions.

Finally, it is important to note that in each of the radial maze, plus maze, and water maze tasks described above, lesions of the dorsal striatum and hippocampal system produce opposite effects. Specifically, in contrast to the effects of dorsal striatal lesions, damage to the hippocampal system selectively impairs behavior in the declarative versions of the tasks. The double dissociations that have been observed in several studies following lesions of the dorsal striatum and hippocampus provide compelling evidence in favor of the hypothesis that these two brain structures are components of memory systems that selectively mediate procedural and declarative memory, respectively.

The Dorsal Striatum and Procedural Learning: Neurochemical Studies

A second line of evidence indicating a selective role for the dorsal striatum in procedural learning is derived from studies investigating the neurochemical bases of memory. For example, corticostriatal projections use glutamate as a neurotransmitter and are hypothesized to provide the sensory stimulus information contained in stimulus–response associations. Dopaminergic projections to the dorsal striatum are thought to provide a reinforcing signal that is critical in the initial formation of stimulus–response associations. Finally, acetylcholine is contained within a population of dorsal striatal interneurons and may interact with dopaminergic function to modulate procedural learning and memory.

Extensive evidence implicates each of these three neurotransmitters in procedural learning and memory. Thus, impairments in the acquisition of various conditioned avoidance behaviors have been observed following depletion of dopamine in the nigrostriatal pathway, and posttraining electrical stimulation of the nigrostriatal bundle enhances consolidation of procedural memory. An enhancement of procedural memory is also observed following posttraining intradorsal striatal injections of indirect (e.g., amphetamine) and direct (e.g., quinpirole) dopamine agonists. The memory-enhancing effects of intradorsal striatal administration of dopamine agonists are task specific: injections facilitate memory consolidation in procedural versions of the radial and water maze tasks described above but have no effect on declarative memory. It is interesting that the role of dorsal striatal dopamine in procedural learning and memory may be relatively restricted to early stages of task acquisition. Consistent with this suggestion, depletion of nigrostriatal dopamine, or administration of dopamine receptor antagonist drugs, does not appear to impair the expression of previously acquired stimulus–response habits. Thus, dopamine release in the dorsal striatum may function during initial procedural learning as a reinforcing signal that acts primarily to 'stamp' together stimulus–response associative bonds.

Findings of several studies also indicate a role for dorsal striatal glutamate in procedural learning and memory. For example, procedural memory consolidation is enhanced by intradorsal striatal injection of glutamate in rats trained to swim to a visible escape platform that is located in a different spatial location on each trial. Moreover, infusions of the glutamatergic *N*-methyl-D-aspartate (NMDA) receptor antagonist AP5, or metabotropic glutamate receptor antagonists, impair memory consolidation in this procedural memory water maze task. Similar to the dissociation observed with dopaminergic drugs, posttraining intradorsal striatal infusions of glutamatergic agonist or antagonist drugs have no effect on memory in a declarative version of the water maze task in which rats are trained to swim to a hidden escape platform that is located in the same spatial location across trials. Additional evidence of a selective role for dorsal striatal glutamate in procedural learning and memory was obtained using the dual-solution plus maze task. Specifically, the transition to procedural response learning that normally occurs with extended training can be facilitated by posttraining intradorsal striatal injections of glutamate during early training. Moreover, infusions of the NMDA receptor antagonist AP5 into the dorsolateral striatum impair acquisition of response learning in a single-solution plus maze.

Several studies examining the effects of posttraining injections of cholinergic drugs also implicate dorsal striatal cholinergic activity in procedural memory consolidation. In general, posttraining intrastriatal infusions of cholinergic agonists (e.g., oxotremorine, choline) enhance, whereas cholinergic antagonists (e.g., atropine, scopolamine) impair, consolidation of procedural memory. Experiments using intracerebral microdialysis to monitor neurotransmitter release in the hippocampus and dorsolateral striatum during performance of the dual-solution plus maze task

suggest that the relative amount of acetylcholine release in these two brain regions influences the use of striatal-dependent procedural learning. For example, the transition from hippocampus-dependent place learning to dorsal-striatal response learning that occurs with extended training in the dual-solution plus maze tasks coincides temporally with an asymptotic value of acetylcholine release in the dorsolateral striatum. A similar relationship between gradual increases in acetylcholine release and transition from a spatial strategy to a response strategy was observed in rats trained in a food-rewarded Y-maze task. In addition, following training in the dual-solution plus maze task, rats that exhibit response learning on a subsequent probe trial display a higher ratio of acetylcholine release in the dorsolateral striatum relative to the hippocampus, whereas the opposite pattern of acetylcholine release is observed in rats displaying place learning.

In sum, a selective role for dorsal striatal dopamine, glutamate, and acetylcholine in procedural learning and memory has been observed across several studies. One challenge for future research is to determine the manner in which these neurotransmitters interact to contribute to the various forms of neural plasticity that have been demonstrated in the dorsal striatum (i.e., long-term potentiation and long-term depression) and which presumably form the basis of synaptic changes that underlie procedural learning.

The Amygdala and Stimulus–Affect Conditioning

Whereas the dorsal striatum has been implicated in procedural learning involving stimulus–response habit formation, other research suggests that the mammalian amygdala mediates a form of procedural learning in which stimulus–affect associations may be acquired. A role for the amygdala in 'emotional' procedural memory has been extensively described by researchers employing various fear conditioning paradigms including contextual, discrete cue, and potentiated startle tasks. Thus, lesions of amygdala severely impair acquisition of these tasks, and several elegant experiments have provided a circuit analysis describing the manner in which efferent amygdala projections can influence specific brain stem regions involved in the expression of different conditioned autonomic fear responses. It should be noted that some investigators have argued that a role for the amygdala in learned fear responses should be offered cautiously in view of evidence that damage to this structure may impair some types of unconditioned fear responses (e.g., freezing behavior).

Brain lesion experiments in which the mnemonic functions of the basolateral amygdala and dorsal striatum have been dissociated are of particular relevance to the hypothesis that these two brain structures selectively mediate stimulus–response and stimulus–affect procedural memory, respectively. For example, the functional integrity of the basolateral amygdala, but not the dorsal striatum, is necessary for acquisition of conditioned place preference behavior for both natural rewards and addictive drugs. In a conditioned place preference task, rats are confined, on alternating days, to one distinct environmental context paired with natural or drug rewards and a second context not paired with the rewarding treatment. In a reward-free test session given following training, the amount of time spent in the two environments is measured, and rats demonstrate a reliable preference for the environment previously paired with the rewarding stimulus. The conditioned place preference paradigm has been used primarily by researchers interested in the neurobiological bases of drug addiction and reward. However, as the animals are in a drug-free state on the critical test day, approach behavior to the environmental cues that have been previously paired with reward presumably involves stimulus–affect memory. Pretraining lesions of the basolateral amygdala impair acquisition of conditioned place preference behavior for natural rewards (food and sex) and for drugs of abuse (e.g., amphetamine and cocaine). In addition, memory consolidation underlying the acquisition of conditioned place preferences for food and amphetamine is blocked by posttraining intrabasolateral amygdala infusions of the cholinergic antagonist scopolamine and lidocaine, respectively. In contrast, lesions of the basolateral amygdala have no effect on acquisition of stimulus–response procedural learning in the radial or water maze.

Finally, activation of efferent basolateral amygdala projections can modulate consolidation of procedural memory occurring in other brain structures, including the dorsal striatum. The memory modulatory role of the basolateral amygdala is related in part to the effects of hormone release on emotional arousal. In addition, emotional state may interact with amygdala function to influence the relative use of multiple memory systems. For example, peripheral and intrabasolateral amygdala injections of anxiogenic drugs result in a predominant use of dorsal striatal-dependent procedural learning in the dual-solution plus maze task. It is important to note, however, that the modulatory influence of the basolateral amygdala has been described in multiple memory systems and is therefore not limited to procedural learning. Moreover, although memory storage and modulation theories of amygdala function have been contrasted as mutually exclusive and competing ideas, a time-limited modulatory role

for the basolateral amygdala on some types of learning and memory does not necessarily preclude a long-term role for this structure in memory underlying stimulus–affect procedural learning.

Conclusions

In sum, significant progress has been made in identifying the neurobiological bases of procedural learning and memory. Specifically, extensive evidence implicates the dorsal striatum and basolateral amygdala in stimulus–response and stimulus–affect forms of procedural learning, respectively. Although this article focuses on findings from brain lesion and pharmacological studies, it is important to note that similar evidence for the selective roles of the dorsal striatum and amygdala in procedural learning is also provided by studies employing electrophysiological and molecular approaches in lower animals, including nonhuman primates. Furthermore, findings of human studies employing neuroimaging techniques, as well as neuropsychological investigations of various neurodegenerative diseases and psychopathologies, also indicate that the dissociable roles of the dorsal striatum and amygdala in procedural learning and memory generalize across mammalian species.

See also: Amnesia: Declarative and Nondeclarative Memory; Animal Intelligence: The Search for Animal Intelligence; Multiple Memory Systems; Perirhinal Cortex: Neural Representations; Procedural Learning: Classical Conditioning; Procedural Learning: Striatum; Procedural Learning in Humans; Procedural Learning: Cerebellum Models.

Further Reading

Cador M, Robbins TW, and Everitt BJ (1989) Involvement of the amygdala in stimulus-reward associations: Interaction with the ventral striatum. *Neuroscience* 30: 77–86.

Cahill L and McGaugh JL (1998) Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neuroscience* 21: 294–299.

Eichenbaum H and Cohen NJ (2001) *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. New York: Oxford University Press.

Gold PE (2004) Coordination of multiple memory systems. *Neurobiology of Learning and Memory* 82(3): 230–242.

Graybiel AM (1998) The basal ganglia and chunking of action repertories. *Neurobiology of Learning and Memory* 70: 119–136.

Kesner RP, Bolland BL, and Dakis M (1993) Memory for spatial locations, motor responses, and objects: Triple dissociation among the hippocampus, caudate nucleus, and extrastriate visual cortex. *Experimental Brain Research* 93: 462–470.

Knowlton BJ, Mangels JA, and Squire LR (1996) A neostriatal habit learning system in humans. *Science* 273: 1399–1402.

LeDoux JE (1995) Emotion: Clues from the brain. *Annual Review of Psychology* 46: 209–235.

Packard MG, Hirsh R, and White NM (1989) Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *Journal of Neuroscience* 9: 1465–1472.

Packard MG and Knowlton BJ (2002) Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience* 25: 563–593.

Packard MG and McGaugh JL (1996) Inactivation of the hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory* 65: 65–72.

Squire LR, Knowlton B, and Musen G (1993) The structure and organization of memory. *Annual Review of Psychology* 44: 453–495.

White NM (1997) Mnemonic functions of the basal ganglia. *Current Opinion in Neurobiology* 7: 164–169.

White NM and McDonald RJ (2002) Multiple parallel memory systems in the brain of the rat. *Neurobiology of Learning and Memory* 77(2): 125–184.

Yin HH and Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* 7(6): 464–476.

Procedural Learning in Humans

K Foerde and R A Poldrack, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Multiple Memory Systems

Human long-term memory consists of multiple distinct types, which appear to rely on separable mechanisms that can be localized to distinct regions of the brain. Early evidence for multiple forms of long-term memory came from studies of patients with neurological damage to the medial temporal lobe (MTL) or related diencephalic structures. Such individuals suffer from a global amnesia characterized by the inability to remember new events or learn new facts (anterograde amnesia), as well as a loss of memory for events prior to the neurological insult (retrograde amnesia), without any deficits in attention, perception, language, or reasoning. This form of memory has been termed 'declarative' or 'explicit' in recognition of the fact that its contents are generally available to conscious awareness and require attention for learning. However, beginning in the late 1960s it became clear that many forms of learning were normal even in the face of dense amnesia. The term 'procedural' was applied to these spared forms of memory because they appeared to reflect changes in how a task was performed ('knowing how') without the necessity of conscious memory for how this learning occurred ('knowing that'). More recently, the term 'nondeclarative memory' has been proposed to cover a broader set of memory phenomena that are spared in amnesia.

Nondeclarative memory encompasses a number of heterogeneous mnemonic functions, one of which is procedural learning, a term that we will use to refer specifically to the acquisition of new skills or abilities which are general to a task domain. As opposed to declarative memory, the knowledge acquired through procedural learning is not necessarily available to conscious awareness, but is instead tied to the routines or procedures involved in task performance. Whereas declarative memory is generally tested through direct assessment about details of a specific episode, procedural learning is observed indirectly as facilitation of performance after training.

Types of Procedural Learning

In the following discussion, we divide procedural learning into the domains of motor, perceptual, and cognitive skills for the sake of organization. These distinctions mainly arise as a result of the tasks used to assess procedural learning. However, it should be recognized that these categories are primarily heuristic and may not necessarily reflect fundamental neurobiological distinctions.

Motor Skill Learning

Motor skill learning involves improvement in the speed and/or accuracy of motor movements with practice. Across the wide range of tasks that have been examined, motor skill learning has commonly been dissociated from declarative memory by showing that amnesic patients can learn the skills normally.

Sequence learning Motor sequence learning is perhaps the most studied motor skill task. In this paradigm, study participants respond manually to the spatial location of a visual stimulus. In implicit sequence-learning paradigms, the stimuli appear in a regular sequence during training, but the participant is not alerted to this regularity. With practice, participants respond more quickly to these sequences compared to randomly ordered stimuli. A substantial proportion of neurologically intact persons routinely report being unaware of the underlying sequence; although the role of awareness remains controversial, awareness of the sequence does not appear to be necessary for learning the skill. Amnesic patients can also show normal learning in this paradigm. In contrast, patients with basal ganglia damage exhibit specific deficits in motor sequence learning. Most neuroimaging studies of implicit motor sequence learning have observed striatal activity, consistent with a role for the striatum in sequence learning. In addition, imaging studies have reported activity in the supplementary motor area (SMA), which is likely involved in motor planning, and the parietal lobe, which appears to be involved in spatial aspects of the task (Figure 1).

Explicit learning of sequences has been examined using a procedure in which participants learn a sequence through trial and error. This type of learning engages regions similar to implicit learning as well as additional prefrontal regions. However, with extended practice, sequence performance becomes more automatic and relies less on prefrontal mechanisms.

With increasing practice, a shift in activity from anterior to posterior cortical and subcortical regions has been observed in several functional magnetic resonance imaging (fMRI) studies. Neurophysiological studies in monkeys have also observed such anterior–posterior shifts in the striatum. Anterior striatal regions are critical for processing feedback, whereas posterior regions appear to be related to performance

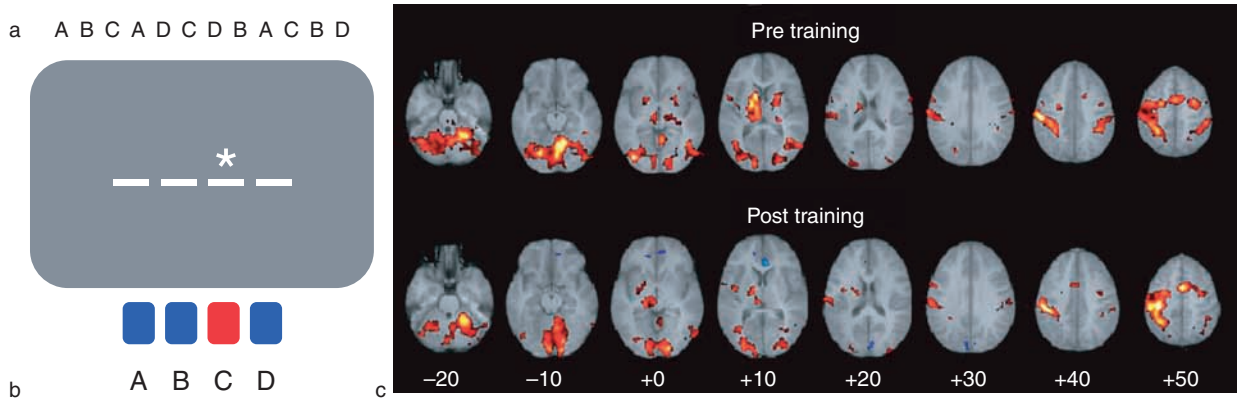

Figure 1 (a) Example of repeating sequence used in a serial reaction time (SRT) task. (b) Participants press a button corresponding to the spatial location of the onscreen asterisk. (c) Results from a fMRI study of SRT performance. Participants in this study were trained on an SRT task for three 1 h sessions (top row, brain images prior to the training session; bottom row, brain images after training). Reproduced from Poldrack RA, Sabb F, Foerde K, et al. (2005). The neural correlates of automaticity during motor skill learning. *Journal of Neuroscience* 25: 5356–5364, with permission. Copyright 2005 by the Society for Neuroscience.

of the learned skill *per se*. Findings regarding contributions of the cerebellum are currently mixed. Based on work with patients, neuroimaging studies, and neurophysiological studies, it has been suggested that the cerebellum may be important for the execution of well-learned motor sequence skills rather than the acquisition.

Rotary pursuit In rotary pursuit tasks the participant tracks a small target on a rotating platter with a stylus. The profile of impairment in patients is mixed, with intact performance in amnesics, impairment in Huntington's disease, and some evidence of deficits in Parkinson's disease (PD). According to neuroimaging studies the SMA and posterior parietal cortex are involved in learning; these regions are similar to those engaged in sequence-learning tasks. Striatum and cerebellum activity has been observed in relation to motor execution on a rotary pursuit task, but thus far increases or decreases related to learning have not been observed in these regions.

Mirror tracing Mirror tracing requires subjects to trace the outline of geometric figures while viewing the figures reflected in a mirror. Tracing becomes faster and less error prone with practice. Amnesic patients and PD patients exhibit normal learning on the task, whereas patients with cerebellar damage do not.

Motor adaptation Motor adaptation tasks generally involve adjusting motor behavior in response to an artificial perturbation. In prism adaptation tasks, participants, fitted with prism spectacles that displace the visual field, must learn to reach for objects. In another paradigm, they are required to move a cursor to a target while interacting with a manipulandum that applies a predictable force. They learn to adjust their movements to compensate for the displacement. The parietal cortex appears to be involved in learning in these paradigms, consistent with the importance of spatial translation for accurate behavior. Furthermore, the cerebellum may be important for the acquisition of these types of motor skills. Through the cerebellum's involvement in error correction for movements, specifically the ability to use execution error information to adapt internal models, it allows improvement on subsequent trials. The striatum also appears to play a role in processing errors online, but is perhaps more involved in adapting current behavior and movement goals without affecting adaptation of internal models (Figure 2).

Perceptual Skill Learning

Perceptual skills involve the development of new abilities to recognize or distinguish perceptual stimuli. We do not discuss perceptual learning for low-level perceptual features or expertise in object and face recognition, but limit our discussion to forms of perceptual skill that have been directly dissociated from declarative memory.

Mirror reading Mirror reading has been used extensively to test the acquisition of perceptual skills. Participants are exposed to spatially transformed (e.g., mirror-reversed) text. Amnesic patients show normal acquisition of mirror-reading skill without the ability to remember specific words. There is some evidence that the basal ganglia are important for learning the mirror-reading task, but the results are inconsistent. Neuroimaging studies have implicated both parietal regions and the striatum in mirror-reading skill learning.

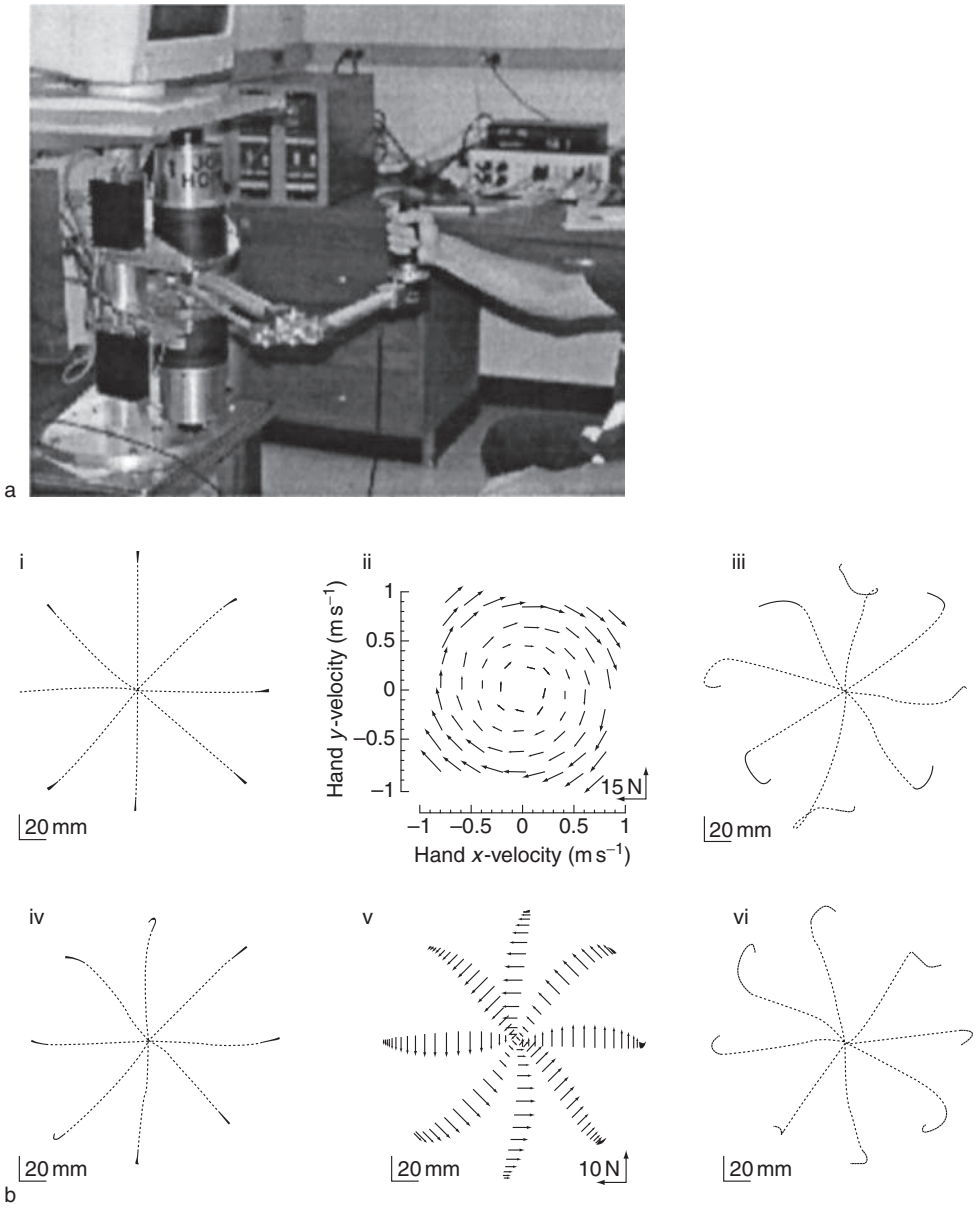

Figure 2 (a) Experimental setup for a motor adaptation task. The manipulandum is a very low-friction, planar mechanism powered by two high-performance torque motors. The study participant grips the handle of the robot. The handle houses a force transducer. The video monitor facing the participant displays a cursor corresponding to the position of the handle. A target position is displayed, and the participant makes a reaching movement. With practice, the participant learns to compensate for the forces produced by the robot. (b) (i) Hand path of a typical participant in the null field (the points in all hand paths are 10 ms apart). (ii) An example of a force field produced by the robot. (iii) Hand path of an untrained participant in the field. (iv) Hand path after 300 movements in the field. The trajectory in the field converges to the trajectory observed in the null field. (v) Forces produced by a typical trained participant to counter the effect of the force field as a function of hand position for each movement. These forces are the projection of the forces measured at the interaction point between the participant and robot onto a direction perpendicular to the direction of target. (vi) While training in the field, random targets are presented with null field conditions. The results are aftereffects. (a) Adapted from Shadmehr R and Brashers-Krug T (1997) Functional stages in the formation of human long-term motor memory. *Journal of Neuroscience* 17: 409–419, with permission. Copyright 1997 by the Society for Neuroscience.

Perceptual categorization In perceptual categorization tasks, participants must learn to categorize multidimensional visual stimuli (e.g., Gabor patches varying in spatial frequency and orientation; see Figure 3(a)). In so-called information integration categorization tasks, categories are delineated by a linear or nonlinear decision boundary involving multiple perceptual dimensions, and no simple verbal rule describes the category. Amnesics can learn perceptual categories even with difficult nonlinear decision bounds.

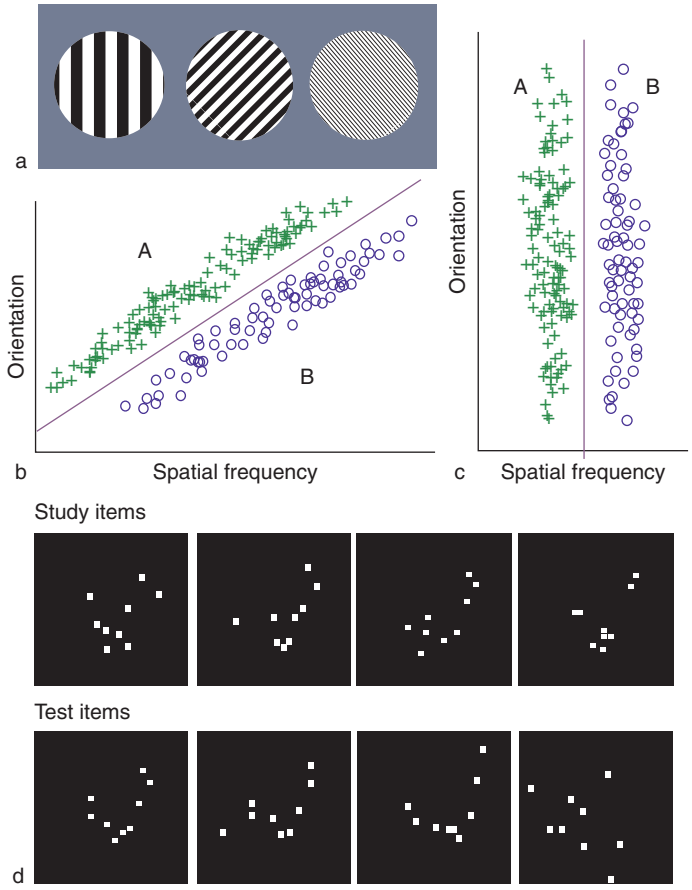

Figure 3 (a) Gabor patches used in perceptual categorization experiments. Spatial frequency and orientation vary from trial to trial. (b) Plot of stimuli that might be used in an information-integration category-learning task. Each plus symbol denotes the spatial frequency and orientation of an exemplar from category A and each circle denotes a category B exemplar. The diagonal line indicates the optimal decision bound, which requires integrating information about orientation and spatial frequency prior to making a decision. (c) Plot of stimuli that might be used in a rule-based category-learning task. The vertical line is the optimal category boundary, a rule that can easily be verbalized. (d) Examples of study items and test items used to assess classification learning of dot patterns. Study items were high distortions of a prototype dot pattern. The test items, illustrated left to right, were presentations of the training prototype, low and high distortions of the training prototype, and random dot patterns. (a–c) Adapted from Ashby FG and Maddox WT (2005) Human category learning. *Annual Review of Psychology* 56: 149–178, with permission. (d) Reproduced from Reber PJ, Stark CEL, and Squire LR (1998) Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences of the United States of America* 95: 747–750, with permission. Copyright 1998 by the National Academy of Sciences, USA.

Basal ganglia disorders consistently impair learning of categories with nonlinear bounds, with inconsistent deficits in learning of linear-bound categories.

Visual prototype learning In prototype learning, participants study a set of dot patterns (see Figure 3(b)), which are distortions of a prototype, without being told about the presence of a category. Participants are then required to decide for subsequent test items whether the items are from the same category as the study items. Amnesic patients are able to normally classify new distortions of the prototype-among distortions of different prototypes, but are significantly worse than neurologically intact controls at recognizing studied patterns. Patients with basal ganglia damage are not impaired at dot pattern classification. Imaging studies have suggested that this type of learning may rely upon early perceptual systems, as classifying new distortions of the learned prototype is associated with decreased activity in occipital cortex (BA 17/18) compared to distortions of unstudied prototypes. This seems specific to incidental learning, as classification of new category members after intentional learning is associated with increased activity in the MTL and prefrontal regions.

Cognitive Skill Learning

Cognitive skill learning involves the development of new task procedures or classification abilities, which are not closely tied to new perceptual or motor abilities.

Category learning Cognitive skills are often acquired through performance of tasks requiring classification of novel stimuli; whereas the examples discussed earlier involved development of new perceptual categories, the tasks discussed here are more abstract. In some tasks, participants simply learn by observing a number of category exemplars, whereas other tasks involve learning from feedback on a trial-by-trial basis.

Probabilistic classification learning In this commonly used paradigm, cues are probabilistically associated with categories. The most common cover story (the 'weather prediction task') is that participants must predict the weather based on a set of cards. Participants are shown individual cues or combinations of the cues and must predict the outcome (rain or sunshine) on each trial, based on feedback provided after each response; however, this feedback is probabilistic. The complexity of the task (as well as the probabilistic feedback) makes it difficult to simply memorize the cue-outcome association on each trial, and participants routinely report feeling unable to learn the task. However, they customarily attain around 70–80% optimal responses within less than 100 training trials. Amnesics are sometimes able to perform this task as well as normal controls do, but they have poorer memory for the details of the testing episode. In contrast, patients with basal ganglia damage are consistently impaired at probabilistic classification, but have intact memory for the testing episode. This deficit appears to be related to the need to process feedback, which relies upon dopaminergic prediction error signals. Neuroimaging data have confirmed the involvement of the basal ganglia in classification learning, while also suggesting that, sometimes, normal persons might rely upon the MTL to perform the task. Learning under dual-task conditions can reduce the reliance upon the MTL and drive study participants to rely on the basal ganglia to perform the task instead (Figure 4).

Artificial grammar learning In the artificial grammar paradigm, participants incidentally process letter strings produced using an underlying set of rules (usually a finite-state grammar). After a short delay participants are told that the items were produced according to a set of rules and they must classify a new set of items according to whether or not they follow the same rules. Amnesics show normal grammar learning on this task, yet have poor recognition for individual letter strings, whereas basal ganglia pathology does not appear to lead to impaired artificial grammar learning (AGL) performance. There are multiple forms of knowledge that can underlie learning in this task, including knowledge about specific features (such as chunks of letter strings) as well as more abstract rules for generating the strings. One fMRI study found activity in the caudate associated with rule knowledge and in the hippocampus in association with chunk knowledge. However, amnesics can acquire both of these forms of knowledge normally, suggesting that neither requires the MTL, even though it is active in imaging studies. Neuroimaging studies have found activation in prefrontal and occipital regions during classification of grammatical versus ungrammatical items, suggesting that those regions may be involved in expressing grammar knowledge in the task.

Tower puzzle tasks Procedural learning of a planning skill has been assessed in neurologically impaired populations by using several variations of a problem-solving task (e.g., Towers of Hanoi, London, and Toronto). The Tower puzzles involve three pegs and disks of varying size (or color). The goal is to move all of the disks from the right peg to the left peg, such that the disks end up in the same configuration as on the starting peg. Disks must be moved one at a time and a larger (or darker) disk may not be placed upon a smaller (or lighter) disk. This class of tasks has been used extensively to test cognitive procedural learning, and initial reports indicated that amnesics could solve this problem, whereas patients with basal ganglia damage were impaired. Follow-up studies in amnesics have not consistently confirmed intact learning on these tasks, whereas the deficits in patients with basal ganglia pathology have been largely confirmed. Imaging has shown increased activity in both basal ganglia and prefrontal regions as task complexity increases, suggesting that the involvement of those regions may be closely related to the planning demands of the task rather than learning *per se*.

Characteristics of Procedural Learning

The heterogeneity of behaviors that are supported by procedural learning mechanisms makes it difficult to formulate general principles regarding procedural learning. We focus herein on a number of general features of procedural learning, though there seem to be exceptions to each potential principle.

Automaticity

One general characteristic of procedural learning may be that skills can become automatic with extensive practice, where automaticity refers to the ability to perform a task without the need for executive control. Automaticity is often defined in terms of dual-task performance: if performance of a skill is insensitive to interference from a secondary task, then it is considered automatic. It has long been suggested that the

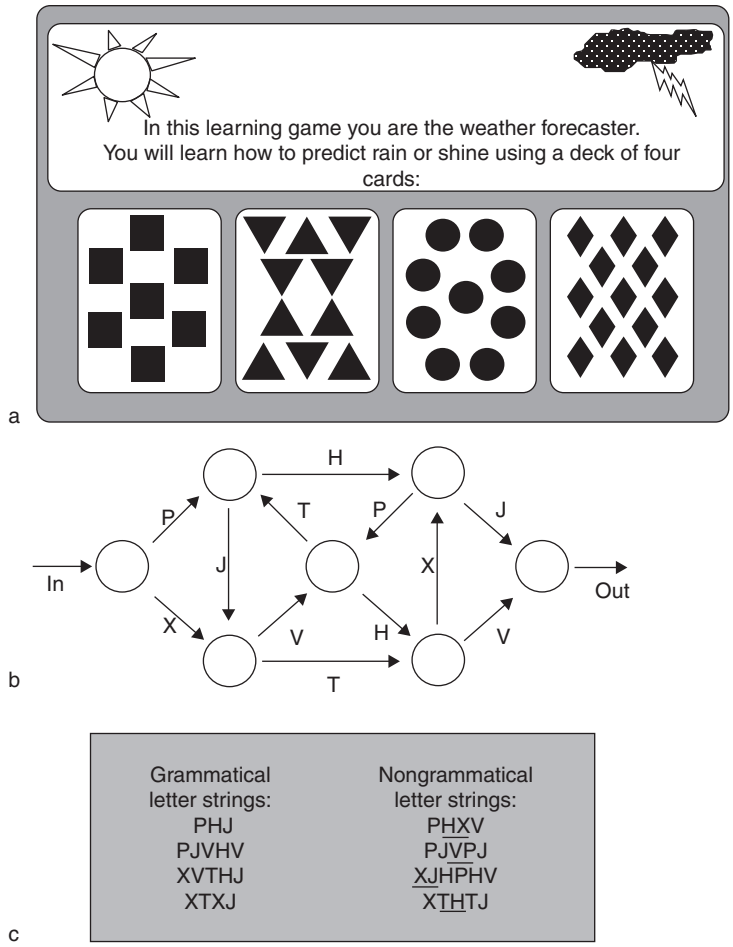

Figure 4 (a) Example of stimuli used in probabilistic classification tasks. One or more of the cards are presented on each trial and a prediction of sunny or rainy outcome is made. (b) Finite-state grammar used to generate letter strings. (c) Examples of grammatical and nongrammatical letter strings that participants would attempt to classify. (a) Reproduced from Knowlton BJ, Mangels JA, and Squire LR (1996) A neostriatal habit learning system in humans. *Science* 273: 1399–1402, with permission. Copyright 1996 by AAAS. (b, c) Reproduced from Skosnik PD, Mirza F, Gitelman DR, et al. (2002) Neural correlates of artificial grammar learning. *NeuroImage* 17: 1306–1314, with permission. Copyright 2002 by Elsevier.

need for executive processes differs between early and late stages of learning. Many studies find that executive control is necessary for learning a skill, although in some domains it appears that executive control is not necessary at any point in learning. Other studies find that interference with executive control processes may affect only the performance, but not the acquisition of a skill. Whatever the role of executive control in early learning, it is clear that procedural learning leads to automaticity in a way that is qualitatively different from declarative memory.

Awareness

The role of awareness in procedural learning is perhaps the most controversial issue surrounding the concept of multiple memory systems. The first difficulty is that the term can refer either to the online awareness of what is being learned (e.g., the grammatical rules in the artificial grammar paradigm, or the particular sequence in the motor sequence-learning task) or to later declarative memory for what was learned. The ability of amnesics to learn skills in the absence of declarative memory for the learning event or insight into their newly acquired abilities provides strong evidence that at least some forms of skill are independent of later declarative memory. Disentangling the role of online awareness from procedural learning is more difficult, since any test for awareness during learning will potentially contaminate the person's strategy for performing the task. Imaging has proved useful in this domain, by showing that the neural basis of performance differs for skills that are learned with awareness versus those learned without awareness. Thus it appears that there are differences in the degree to which procedural learning requires online awareness of task contingencies, but this issue will remain controversial until there

is an adequate account of the functional role of awareness in learning.

Long-Term Retention

Although rates of retention in procedural learning have not been extensively investigated, in some cases procedural learning has been characterized by remarkably long-term retention. Studies of mirror tracing, mirror reading, and AGL have all reported skill retention in excess of 1 year, in amnesics as well as normal controls. Studies of motor sequence learning have reported retention after 1–2 weeks, though one study reported a failure to retain sequence knowledge at 1 year. Whereas forgetting of declarative memories is thought to occur primarily due to interference, this factor may be less relevant for procedural learning. Study participants are unlikely to have much experience outside of the laboratory with the kinds of tasks involved in procedural learning studies (e.g., categorizing dot patterns) and thus there may be less possibility for interference than for declarative memories. Some studies also suggest that procedural memories may become resistant to interference over time due to a consolidation process that requires some number of hours following learning. Interference has mainly been studied in the context of motor adaptation. Other skills (e.g., sequence learning) rely on different neural substrates and the representations may change differently over time.

Neural Bases of Procedural Learning

Much knowledge regarding the neural basis of procedural learning initially came from neuropsychological studies. Early studies of amnesic patients revealed that procedural learning does not require intact MTL structures. Studies in Parkinson's and Huntington's disease patients with damage to the basal ganglia, and other patients with damage to the cerebellum, have shown that both of these brain structures may be important substrates of procedural learning. Neuroimaging has provided additional evidence regarding the neural basis of procedural learning and has highlighted the fact that learning involves dynamic changes in brain activity with practice. There has sometimes been a lack of convergence between neuropsychological data and imaging data, which may reflect the fact that activation in neuroimaging does not indicate that a region is necessary for task performance. On the other hand, imaging studies in healthy populations have revealed that healthy individuals may approach tasks differently than damaged populations do because they have multiple representations or types of knowledge available. Here we outline the major brain structures known to be involved in procedural learning.

Striatum

The striatum appears to be the most widely involved brain structure in procedural learning, based on both lesion evidence and functional imaging results. One potential role for the striatum is in the acquisition of stimulus–response (S–R) associations, likely dependent upon dopamine-modulated synaptic plasticity at corticostriatal synapses. Tasks such as classification learning, which involve learning of responses to specific stimuli based on feedback, are impaired in patients with Huntington's disease (damaging striatal neurons) and PD (disrupting the dopaminergic system). Further, the impairment in PD on classification tasks appears to be specific to learning from feedback. Thus, the reinforcement learning mechanisms supported by the striatum may play an important role in procedural learning.

Feedback does not play a critical role in sequencing tasks, but the basal ganglia also appear to be important for developing unitized action sequences. Through repeated performance of a motor sequence, individual responses become chained into a single chunk or an action sequence. In this way common actions can be executed quicker and more reliably. For example, a forgotten telephone number or password might be successfully completed if one knows the first elements. A chunking mechanism might account for improved performance of motor sequences and perhaps also on Tower puzzles, where optimal performance requires multiple moves in a particular sequence. The best evidence for basal ganglia involvement in such a chunking mechanism comes from recordings from neurons in rodent striatum. During repeated performance of a task, the pattern of neural activity changed from activity occurring throughout the task procedure to activity occurring only at the onset and the offset of the task procedure. A more efficient representation of repeated behaviors could allow more automatic or habitlike performance.

The role of the striatum in procedural learning is increasingly being parsed into more specific functions and related to subregions within the striatum. For example, processing feedback in response to actions is consistently tied to the head of the caudate nucleus, whereas motor learning appears to rely upon the putamen. Separate parallel corticostriatal loops are tied to performance of tasks in various domains, but these loops also communicate throughout the striatum. Different tasks probably rely on the various functions of the basal ganglia to a different degree. Generally, it has been suggested that the basal ganglia may be critical in open-loop skill learning, where behavior is initiated by internally generated plans and subsequently received feedback guides future behavior.

Several neuroimaging studies have found that striatal activity during procedural learning is associated with decreased MTL activity. The meaning of this MTL decrease is not fully understood, but it has been suggested that regions involved in learning declarative knowledge and those involved in procedural learning may be in competition to control behavior. Interestingly, some studies of classification learning and sequence learning using neuroimaging in patients with striatal pathology have observed MTL activity during learning rather than the striatal activity found in normal controls. Thus, for some of these tasks it appears that other strategies may be available when procedural learning is compromised.

Cerebellum

In contrast to the striatum, the cerebellum has been proposed as critical in closed-loop skill learning, where continuous external feedback guides behavior (e.g., mirror tracing), or when novel mappings between motor behavior and visual cues are required (e.g., motor adaptation tasks). As such, a function of the cerebellum may also be thought of as processing feedback, but specifically monitoring and optimizing movements in response to online sensory feedback.

Cerebral Cortex

Several other regions are consistently associated with procedural learning. Whereas the necessity of cerebellum and basal ganglia structures has been studied in various patient populations, the involvement of other regions is generally observed using neuroimaging. Although imaging does not allow inference about necessity, imaging results have generally suggested that procedural learning involves changes in the same areas that are necessary for task performance. For example, the SMA, premotor areas, and primary motor cortex are involved in learning motor skills. Occipital lobe and temporal lobe visual regions are involved in visual perceptual tasks, such as prototype learning. Parietal regions are engaged when tasks require processing of spatial information. These results are consistent with the notion that procedural learning involves online changes in the neural networks involved in task performance. Additionally, prefrontal regions are generally involved during the early stages of learning, likely providing executive control over novice performance.

Conclusion

Although much is still to be learned about how procedural learning is acquired and develops with training, it is clear that humans have a remarkable capacity for acquiring a wide range of skills. These skills can sometimes be retained for years without continued practice and often generalize beyond the studied materials. Conscious memory for the learning event is not necessary, as numerous demonstrations of intact procedural learning by amnesics have shown. The striatum and cerebellum appear to play general roles in procedural learning and, depending on the task, various sensory–motor cortical regions might also be important. Future studies will be needed to further explore how activity changes with practice and better understand what these changes mean.

See also: Basal Ganglia: Habit; Cognition: Cerebellum Role; Motor Skill Learning; Multiple Memory Systems; Perceptual Learning: Neural Mechanisms; Perceptual Learning and Sensory Plasticity; Procedural Learning: Striatum; Procedural Learning in Animals.

Further Reading

Ashby FG and Maddox WT (2005) Human category learning. *Annual Review of Psychology* 56: 149–178.

Cohen NJ and Squire LR (1980) Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science* 210(4466): 207–210.

Corkin S (1968) Acquisition of motor skill after bilateral medial temporal-lobe excision. *Neuropsychologia* 6: 255–265.

Eichenbaum HE and Cohen NJ (2001) *From Conditioning to Conscious Recollection: Memory Systems of the Brain.* New York: Oxford University Press.

Foerde L, Knowlton BJ, and Poldrack RA (2006) Distraction modulates the engagement of competing memory systems. *Proceedings of the National Academy of Sciences of the United States of America* 103: 11778–11783.

Karni A, Meyer G, Jezzard P, et al. (1995) Functional MRI evidence for adult motor cortex plasticity during motor skill learning. *Nature* 377: 155–158.

Knowlton BJ, Mangels JA, and Squire LR (1996) A neostriatal habit learning system in humans. *Science* 273: 1399–1402.

Poldrack RA, Sabb F, Foerde K, et al. (2005) The neural correlates of automaticity during motor skill learning. *Journal of Neuroscience* 25: 5356–5364.

Poldrack RA and Willingham DB (2006) Functional neuroimaging of skill learning. In: Cabeza R and Kingstone A (eds.) *Handbook of Neuroimaging of Cognition,* 2nd edn., pp. 113–148. Cambridge, MA: MIT Press.

Reber PJ, Stark CEL, and Squire LR (1998) Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences of the United States of America* 95: 747–750.

Saint-Cyr JA, Taylor AE, and Lang AE (1988) Procedural learning and neostriatal dysfunction in man. *Brain* 111: 941–959.

Salmon D and Butters N (1995) Neurobiology of skill and habit learning. *Current Opinion in Neurobiology* 5: 184–190.

Shadmehr R and Brashers-Krug T (1997) Functional stages in the formation of human long-term motor memory. *Journal of Neuroscience* 17: 409–419.

Shadmehr R and Holcomb HH (1997) Neural correlates of motor memory consolidation. *Science* 277(5327): 821–825.

Skosnik PD, Mirza F, Gitelman DR, et al. (2002) Neural correlates of artificial grammar learning. *NeuroImage* 17: 1306–1314.

Squire LR (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* 99(2): 195–231.

Procedural Learning: Cerebellum Models

R D Seidler, University of Michigan, Ann Arbor, MI, USA
J Ashe, University of Minnesota, Minneapolis, MN, USA; and Minneapolis VA Medical Center, Minneapolis, MN, USA

© 2009 Published by Elsevier Ltd.

Introduction

Procedural learning is the acquisition of skills, or knowledge regarding how to perform tasks. In contrast, declarative learning is the acquisition of facts (what, where, and when knowledge). Procedural, or motor, learning was defined by Schmidt in 1988 as "a set of processes associated with practice or experience leading to relatively permanent changes in the capability for responding." Such learning can be either explicit, in which the learner is aware of the learning conditions and the goal to acquire the task, in addition to being able to describe the nature of the task, or implicit, in which learning occurs below the level of awareness of the learner. Under most natural settings, skill learning involves both explicit and implicit components. In the laboratory, however, the two can be isolated and studied independently. Although important debates related to these distinctions are ongoing, they will not be emphasized in this article because the cerebellum is not believed to be an important neural substrate.

Researchers who study skill learning tend to classify motor tasks into at least two categories: sequence learning and sensorimotor adaptation. When acquiring new sequences of action, the learner combines isolated elements of movement into a smooth, coherent action – for example, when a gymnast combines tumbling movements into a floor routine. Sensorimotor adaptation requires a remapping of sensory and motor representations due to either altered sensory input or motor output characteristics. One example is learning the relation between the display of cursor movements on a computer screen and hand–mouse movements. These two types of learning rely on distinct cognitive strategies and engage different neural circuitry. We focus on whether and how the cortico-cerebellar pathways contribute to sequence learning and sensorimotor adaptation. The literature shows that the role played by the cerebellum during procedural learning is task dependent.

Motor Performance versus Motor Learning: A Potential Confound

One potential confound that needs to be dealt with when studying the neural basis of skill acquisition is that the sensory and motor systems reflect motor performance levels, and performance changes as a function of learning. For example, it has been shown that neural activity in the sensorimotor system changes with muscle force, movement rate, movement size, movement duration, reaction time, and the extent of motor errors. These measures are typically altered as a function of practice, with movements becoming faster, more efficient, and more accurate with learning. Therefore, when comparing brain activity between early and late learning, it is difficult to disentangle whether changes in neural activation, particularly in the sensorimotor system, are a true reflection of the learning process or simply reflect the implementation of faster, more accurate movements.

Researchers studying skill learning have attempted to address this confound in one of two ways. The first is to build into the experiments controls for change in performance. For example, participants in experiments on skill learning can be asked to make movements under varying performance levels in the absence of learning. They may make movements at different rates in response to a pacing signal or under conditions of low and high difficulty. The level of error may be manipulated by requiring subjects to make movements to targets of different size; movements at a certain speed to smaller targets will have a larger error than those to large targets at the same speed. Brain activity can be recorded under these conditions and compared to the changes that occur during learning with similar performance changes to determine whether the patterns during learning reflect encoding of new movement representations or simply executing movements at differing performance levels. Another experimental approach is to interfere with changes in performance without affecting learning using manipulations such as dual tasking or varying the level of attention. The extent of learning is then assessed by retesting subjects without the interfering condition. Such manipulations can be used to limit the performance changes that typically occur during learning, allowing one to isolate the neural substrates associated with learning *per se*.

Studies on the role of the cerebellum in procedural learning are particularly susceptible to the performance/learning confound. This is because neural activity in the cerebellum may vary parametrically with changes in motor performance in the absence of learning. For example, cerebellar Purkinje cells in lobules IV–VI produce complex spikes both early and late in a movement. It is thought that the early complex spikes encode the absolute reach destination, or the distance between the start and end points of an

action. The later complex spikes have been interpreted as encoding the relative error at the end of a reaching movement. Understanding cerebellar sensitivity to motor error is important because errors systematically decrease during learning. Additionally, when participants make movements under conditions of high difficulty (aiming to small targets), functional magnetic resonance imaging (fMRI) activation is seen bilaterally in the medial cerebellum. Increasing cerebellar activity was found for slower movements and larger errors. Taken together, these data suggest that cerebellar activity would be high early in learning, when errors are being made, and would then decay as a function of practice, as is typically seen. However, it is not clear whether this pattern simply reflects the steady decrease in motor errors or, rather, indicates a role in encoding of new motor memories.

The cerebellum is also thought to play a role in short-duration (hundreds of milliseconds) motor timing, although there is continuing debate about this in the literature. Its role in timing may be task dependent. For example, the cerebellum seems to be important for the control of discontinuous movements (repetitive tapping on a table top). It has been suggested that such movements require the explicit representation of a temporal goal. Continuous movements, on the other hand, are those that do not have discrete events, such as repeatedly drawing circles. It has been proposed that timing is an emergent feature of movement specification in these tasks. Patients with unilateral cerebellar lesions have great difficulty in performing discontinuous motor tasks with the ipsilesional hand but not the contralesional one. In contrast, they have no difficulty with continuous movements. In addition, patients with damage to the cerebellum have difficulty with other behaviors that require precise timing, such as timing the opening of the hand when throwing a ball or in conditioning the eye-blink response to an auditory tone. Timing demands and temporal coordination may vary during the time course of learning, again complicating the interpretation of the contribution of the cerebellum to skill learning.

In summary, the cerebellar system codes for many variables of movement, including reaction time, motor error, movement speed, and timing. These representations have made it difficult to parse out the specific contributions of this system to procedural learning because performance varies along these dimensions as a function of motor practice.

Cerebellar Contribution to Sensorimotor Adaptation

Sensorimotor adaptation tasks are classified into two general categories: dynamic (or kinetic), in which subjects adapt to an altered motor environment such as a force field, and kinematic, in which subjects adapt movements to a mismatch in the visual and proprioceptive environments such as with the use of wedge prisms glasses that cause refractive displacement. It is thought that sensorimotor adaptation occurs through updating the internal model of the behavior. An internal model is used to predict the effects of a particular object or environment on the body and to calculate the motor output required to perform certain actions. Because the cerebellum appears to contribute to both forms of sensorimotor adaptation, much effort has been expended searching for reflections of internal models in this structure.

Some early positron emission tomography (PET) studies were unable to document activation of the cerebellum during prism adaptation after the motor error effects of performance had been accounted for. However, subsequent fMRI and PET studies of both kinematic and dynamic adaptation have repeatedly shown cerebellar activation associated with the acquisition, storage, and consolidation of new internal models, even after adjusting for performance effects such as detecting and correcting errors and varying movement speeds. Such activation is typically found in the lateral cerebellum in hemispheres V and VI, in the region surrounding the superior posterior fissure. In addition, patients with cerebellar damage do not show adaptive performance changes across trials when making pointing movements during force-field adaptation. The reason for the discrepancy between these recent studies and earlier studies remains unclear. A neural recording study in behaving monkeys also raises questions about the role of the cerebellum in updating internal models in response to a perturbing force field. Purkinje cell activity was recorded while monkeys made arm movements under viscous and elastic loads. There was no relation between the rate of simple spike discharge in Purkinje cells and the force conditions, as would be required for storage of the newly acquired internal model; rather, these neurons showed activity that was related to the kinematics of the arm movements.

Current efforts are focusing on determining the specific role of the cerebellar circuitry during adaptation. Its contribution appears to be in updating internal models from one trial to the next (predictive control), as opposed to performing on line (within a single trial) movement corrections. For example, in one experiment, cerebellar patients walked on a split-belt treadmill. When the two belts were driven at different speeds, the patients showed reactive changes to their locomotor pattern within a few strides of the perturbation. In contrast, they could not make predictive changes over time and did not adapt to the perturbation in the same manner as healthy control

participants. A similar finding has been observed for cerebellar patients in terms of anticipatory postural adjustments (APAs). Typically, when a person holds an object in his or her hand and it is removed by another person, the hand moves upward. This occurs because the person holding the object was not able to plan in advance for the other's actions. In contrast, when the object is removed by the holder's other hand, the hand remains stable because of an APA that compensates for the lifting action. Patients with cerebellar damage do show such APAs during the bimanual manipulation of objects. However, they cannot learn to modify and properly time APAs in response to predictable lifting of an object from their hand by an external agent (robotic device), whereas healthy control participants can do so.

Interestingly, the basal ganglia thalamocortical pathways appear to play a role in error correction that is complementary to that of the cerebellum. Patients with Huntington's disease, a basal ganglia pathology, are able to adapt arm movements to an opposing force field. They show adaptive modifications to their performance from one trial to the next. However, they are not able to show corrective adjustments for motor errors within a single trial in the manner in which cerebellar patients do.

Another form of sensorimotor adaptation that appears to be an important function of the cerebellum is updating of the vestibulo-ocular reflex (VOR). This response allows one to maintain eye gaze on a target during head movement. Head motions, detected by the vestibular system, are canceled out via reflexive, compensatory eye movements. The gain of this response (ratio of head velocity to eye velocity) can be altered by having experimental participants wear minifying or magnifying visual lenses during combined eye and head movements. Animals are unable to acquire new VOR gains after removal of the cerebellum, and human patients with cerebellar damage exhibit similar deficits. Studies have shown that floccular Purkinje cell complex spikes change their sensitivities to head velocity, eye position, and eye velocity during the course of VOR adaptation, providing the necessary substrate for VOR adaptation.

In summary, with the exception of the very special case of the VOR, the exact contribution of the cerebellum to sensorimotor adaptation remains unclear. The focus of future work will be on determining the extent to which internal models are reflected in the activity of the cerebellum.

Cerebellar Contributions to Sequence Learning

A large number of studies have also implicated the cerebellum in learning new sequences of action, although there is evidence to the contrary. The most commonly used sequence learning paradigm for investigating cerebellar contributions is the serial reaction time (SRT) task. In this task, subjects make manual button presses in response to stimuli presented on a computer screen. For explicit SRT learning, participants are told about the existence of the sequence ahead of time and encouraged to use explicit search strategies. For implicit SRT learning, participants are not informed about the presence of the sequence, and it is typically embedded among randomly presented stimuli in an effort to prevent participants from gaining explicit awareness. Patients with cerebellar damage exhibit impaired implicit SRT sequence learning, and neuroimaging investigations of this task have found cerebellar activation occurring during the course of sequence encoding. It has been hypothesized that the cerebellum contributes to sequence learning through facilitating the increase in sensorimotor coordination and temporal precision occurring with practice. Interestingly, work with patients with cerebellar damage and repetitive transcranial magnetic stimulation experiments that inhibit cerebellar function have shown that the right cerebellum is associated with SRT sequence learning regardless of the hand used for performance. One interpretation of this finding is that the right cerebellum interfaces with the parietal and motor regions of the left hemisphere, which have been shown to play a dominant role in motor control and sequence learning for both hands.

Evidence against cerebellar contributions to sequence learning derives from an experiment in which the performance changes that typically occur during the time course of learning were suppressed. Participants performed a secondary, distractor task simultaneously with the SRT sequence learning task. This secondary task served to keep response time and accuracy constant, even though participants were continuing to practice the sequence. This secondary task did not impact sequence encoding, however, because once it was removed, participants showed immediate expression of learning, the magnitude of which was comparable to that seen during single-task SRT studies. In this fMRI study, cerebellar activity was not observed during the sequence learning process in the absence of performance changes; rather, cerebellar activation in the lateral hemispheres (V/VI) was correlated with the response time and error savings observed at expression of learning. These data raise the possibility that cerebellar contributions to sequence learning reported in other studies may be a reflection of changing performance during the time course of learning rather than making a specific contribution to the encoding processes *per se*.

Because of the clear involvement of the cerebellum in timing discussed previously, it may indeed contribute to sequence learning tasks that have a specific temporal component. A PET study that involved both spatial and temporal sequencing components showed early learning activation in the lateral cerebellar hemispheres. This activity was positively correlated with activation in the right dentate nucleus and the left primary motor cortex. As learning progressed, cerebral blood flow in the dentate decreased and the primary motor cortex showed corresponding increases, possibly reflecting storage and consolidation of temporal sequence representations. Additionally, in a manual tracking study, patients with cerebellar damage showed improvements in spatial tracking accuracy with practice that paralleled changes shown by healthy controls. The patients showed no difference in temporal precision, however, between tracking a random and repeating target pattern following 3 days of task practice. Combined, these data imply that although the cerebellar circuitry may support sequence learning that has temporal requirements, it does not appear to play a role in spatial sequence learning.

See also: Cerebellum: Clinical Pathology; Cerebellum: Evolution and Comparative Anatomy; Cerebellum: Models; Cognition: Cerebellum Role; Motor Skill Learning; Motor Sequences; Motor Timing; Procedural Learning in Animals; Procedural Learning in Humans; Vestibulo-Ocular Reflex.

Further Reading

Bastian AJ (2006) Learning to predict the future: The cerebellum adapts feedforward movement control. *Current Opinion in Neurobiology* 16: 645–649.

Imamizu H, Miyauchi S, Tamada T, et al. (2000) Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature* 403: 192–195.

Kitazawa S, Kimura T, and Yin PB (1998) Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature* 392: 494–497.

Morton SM and Bastian AJ (2006) Cerebellar contributions to locomotor adaptations during splitbelt treadmill walking. *Journal of Neuroscience* 26: 9107–9116.

Pasalar S, Roitman AV, Durfee WK, and Ebner TJ (2006) Force field effects on cerebellar Purkinje cell discharge with implications for internal models. *Nature Neuroscience* 9: 1404–1411.

Seidler RD, Purushotham A, Kim SG, Ugurbil K, Willingham D, and Ashe J (2002) Cerebellum activation associated with performance change but not motor learning. *Science* 296: 2043–2046.

Spencer RM, Zelaznik HN, Diedrichsen J, and Ivry RB (2003) Disrupted timing of discontinuous but not continuous movements by cerebellar lesions. *Science* 300: 1437–1439.

Wolpert DM, Miall RC, and Kawato M (1998) Internal models in the cerebellum. *Trends in Cognitive Sciences* 2: 338–347.

Procedural Learning: Classical Conditioning

R E Clark, University of California at San Diego, La Jolla, CA, USA
R F Thompson, University of Southern California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Why Classical Conditioning Is a Form of Procedural Memory

Memory is composed of various abilities that depend on different brain systems. Declarative memory provides the capacity for conscious recollection of facts and events. Procedural memory is a type of nondeclarative memory in that it is not accessible to conscious recollection (i.e., it cannot be brought to mind and declared). Rather, procedural memory is expressed through performance. Classical conditioning is a type of procedural memory because learning and memory are indexed by the expression/generation of learned response. Under most circumstances (but not all, see below), declarative knowledge is superfluous to the expression of the classically conditioned response (CR). Accordingly, this form of conditioning is considered a quintessential form of procedural learning and memory.

Classical conditioning is a basic form of associative learning where the organism learns something about the causal fabric of the environment. Stimuli can be arranged so that one stimulus provides the organism with information concerning the occurrence of another stimulus. This type of associative learning is most commonly referred to as classical conditioning, but has also been commonly referred to as Pavlovian conditioning. In the most basic form of classical conditioning, the stimulus that predicts the occurrence of another stimulus is termed the conditioned stimulus (CS). The predicted stimulus is termed the unconditioned stimulus (US). The CS is a relatively neutral stimulus that can be detected by the organism, but does not initially induce a reliable behavioral response. The US is a stimulus that can reliably induce a measurable response from the first presentation. The response that is elicited by the presentation of the US is termed the unconditioned response (UR). The term 'unconditioned' is used to indicate that the response is 'not learned,' but rather it is an innate or reflexive response to the US. With repeated presentations of the CS followed by US (referred to as paired training) the CS begins to elicit a CR. Here the term 'conditioned' is used to indicate that the response is 'learned.'

The Distinction between Classical and Instrumental/Operant Conditioning

In classical conditioning, no contingency exists between the CR and the presentation or omission of the US. In other words, it does not matter if a CR is made or not made; the presentation of the US still occurs and is experienced and processed by the subject. This has been termed stimulus-contingent reinforcement or the 'law of contiguity.' In instrumental or operant conditioning, by contrast, a contingent relationship is arranged between the subject's response and the presentation or nonpresentation of a reinforcer. In other words, this form of conditioning involves response-contingent reinforcement or the 'law of effect.'

Origin of the Term 'Classical Conditioning'

Before the 1940s, the process of classical conditioning was referred to most often as the conditioned reflex, Pavlovian conditioning, or simply as conditioning. However, in the 1930s, scientists began to understand that the laws governing learning, in paradigms where reinforcement was contingent upon the organism's behavior appeared to be fundamentally different from the laws governing the conditioned reflex. The latter type of learning would come be known as instrumental or operant conditioning. This created a need to distinguish different forms of conditioning. It appears that the term classical conditioning developed as a contraction of the descriptive phase 'classical Pavlovian conditioning' which was used to denote the 'well-known' (i.e., classical) type of conditioning used by Pavlov (i.e., Pavlovian).

Different Types of Classical Conditioning based on the Measured Response

Different types of classical conditioning can be characterized by the measured response. Here we focus on two different types of classical conditioning paradigms. First we describe classical conditioning of the emotional response of fear. In the case of fear conditioning the CR to the CS involves a relatively broad change in emotional state. Next we describe classical conditioning of the eyeblink response that is an example of an aversively conditioned somatic motor response. The response is a highly specific motor movement that becomes adaptively timed to the presentation of the US.

Fear Conditioning

Fear in most cases represents the perception and recognition of danger, the learning and remembering of a dangerous experience, and the coordination of defensive behaviors to an environmental threat.

In the laboratory (as with rats) fear typically refers to emotional responses learned with an aversive stimulus. Presentation of an unpleasant stimulus like an electric shock typically elicits a variety of behaviors that are used to infer a central state of fear (Figure 1). After several pairings (or even one) of a several-second tone or light, terminating with a paw-shock, the CS now elicits behaviors initially elicited by the shock US (Figure 1), the CR. The expression of conditioned fear in the rat involves the synchronization of defensive reactions which include autonomic increases in blood pressure, respiration and heart rate, neuroendocrine glucocorticoid release, behavioral immobilization or freezing, and potentiated startle (Table 1). Interestingly, as will be discussed later in more detail, reexposure to the original training context alone is sufficient to trigger conditioned fear.

Brain Substrates of Fear Conditioning

Work in a number of laboratories using lesions, pharmacological and electrophysiological techniques, demonstrates that the amygdala is essential for the acquisition, storage, and expression of conditioned fear.

Anatomical evidence reveals that both CS and US afferents converge on the amygdala. Lesions of the amygdala prevent the acquisition of conditioned fear to all CS modalities (tone, light, and odor). Lesions limited to medial geniculate nucleus, an afferent of

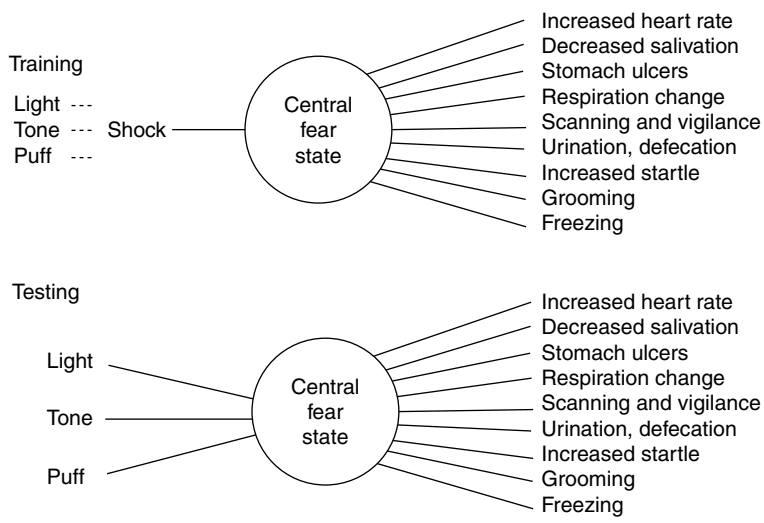

Figure 1 General scheme believed to occur during classical conditioning with an aversive CS. During training, the aversive stimulus (e.g., shock) activates a central fear system that produces a constellation of behaviors generally associated with aversive stimuli (URs). After consistent pairings of some neutral stimulus such as light or tone or puff of air with shock during the training phase, the neutral stimulus is capable of producing a similar fear state and hence the same set of behaviors (conditioned reponses), formerly only produced by the shock. Reproduced from Davis M (1992) The role of amygdala in fear and anxiety. *Annual review of Neuroscience* 15: 353–375.

Table 1 Comparison of measures in animals typically used to index fear and those in the DSM-III manual to index generalized anxiety in people

Measures of fear in animal models	*DSM-III criteria – generalized anxiety*
Increased heart rate	Heart pounding
Decreased salivation	Dry mouth
Stomach ulcers	Upset stomach
Respiration change	Increased respiration
Scanning and vigilance	Scanning and vegilance
Increased startle	Jumpiness, easy startle
Urination	Frequent urination
Defecation	Diarrhea
Grooming	Fidgeting
Freezing	Apprehensive expectation – something bad is going to happen

Reproduced from Davis M (1992) the role of amygdala in fear and anxiety. *Annual review of Neuroscience* 15: 353–375.

the amygdala, block the acquisition of conditioned fear to a tone, but leave conditioning to light intact, suggesting that the amygdala is not part of the sensory pathway in the formation of a memory trace. Furthermore, in untrained animals, amygdala-specific infusions of the *N*-methyl-D-aspartate (NMDA) receptor antagonist AP5 prevent the acquisition of conditioned fear, whereas similar infusions in a well-trained animal show little effect. Alternatively, appropriate lesions of the amygdala in the well-trained animal completely and permanently abolish the expression of all forms of conditioned fear (freezing, tachycardia, changes in respiration and blood pressure). These studies (and others) strongly suggest that memory traces for conditioned fear are established and maintained in the amygdala.

Anatomically, the amygdala is a rather complex system containing several nuclei. In brief, auditory information projects from the medial geniculate nucleus of the thalamus to the dorsal portion of the lateral amygdala (LA), which projects to the basolateral amygdala and on to the major efferent nucleus, the central nucleus of the amygdala (see Figure 2). More complex stimuli, for example, about environmental cues, reach the amygdala via the hippocampus and entorhinal cortex.

As indicated in Figure 2, there are two other aspects to fear conditioning in addition to simple pairing of tone and shock, termed delay conditioning. Animals (and people) develop conditioned fear to the environment (context) where they experience strongly aversive events. Contextual fear conditioning procedures typically involve placing a rat in a novel training box for fixed period of time, then delivering a series of spaced footshocks. The following day the rat is returned to the original context (without shocks) and the total amount of time the rat freezes is recorded as an index of contextual fear conditioning. Lesions of the amygdala abolish all forms of conditioned fear including fear classical conditioning to context. However, unlike delay conditioned fear, lesions of the hippocampus prevent the acquisition and expression of contextual fear conditioning. Lesions of the hippocampus made 1 day after conditioning completely abolished conditioned fear to context, whereas lesions made 28 days later had no effect on this conditioned fear.

Still another form of fear conditioning involves the trace procedure, where a tone is separated for several seconds from the presentation of an aversive footshock. Appropriate lesions of the hippocampus prevent the acquisition of trace-conditioned fear as well.

A Few Caveats

There is an interesting apparent discrepancy between effect of amygdalar lesions in classically conditioned

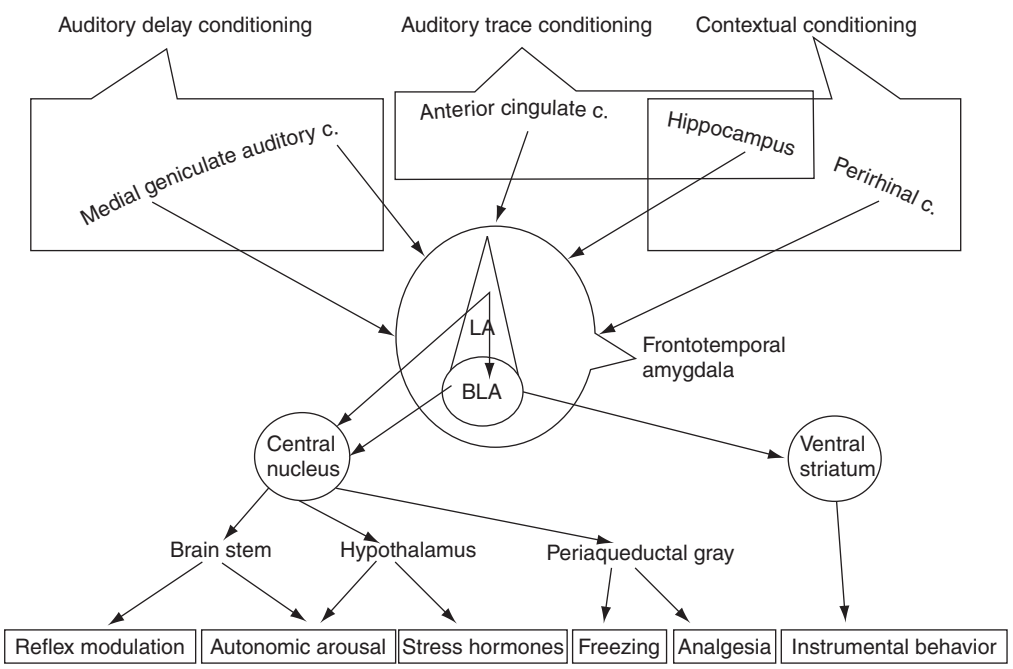

Figure 2 The basic circuit for fear conditioning organized according to various types of conditional stimuli at the top and conditional fear responses at the bottom. LA refers to the lateral nucleus of the amygdala and BLA refers to the basolateral nucleus of the amygdala. The BLA and the LA are parts of the frontotemporal amygdala and comprise the neuroanatomical hub for learned fear. Reproduced from Fanselow MS and Poulos AM (2005) The neuroscience of mammalian associative learning. *Annual Review of Psychology* 56: 207–234.

fear and instrumental learning of fear. In the instrumental situation (e.g., passive avoidance), on day 1 the rat is placed in a lighted compartment and runs into a dark compartment, where it receives a footshock. When placed in the lighted compartment the next day, the animal typically waits a very long time before entering the dark compartment. Amygdala lesions immediately after training abolish this learning but if done a week after training, do not. When animals are given classically conditioned fear training and tested 16 months later they still showed robust conditioned freezing. Subsequent amygdala lesions dramatically impaired the 16-month-old fear memory. Interestingly, animals with amygdala lesions that completely abolished previously trained conditioned fear (freezing) are able to reacquire normal conditioned fear with additional training. Thus, brain structures other than the amygdala, as yet unknown, would seem to become engaged if the amygdala is no longer available.

Classical Conditioning of the Eyeblink Response

Eyeblink conditioning is the most widely studied form of procedural learning in mammals. In standard delay conditioning, a tone or a light CS precedes and coterminates with a reflex-eliciting US such as periorbital shock or a puff of air to the cornea. In trace conditioning, a period of no stimulation separates CS offset from US onset. Extensive investigation into the neural substrates of this associative memory has resulted in perhaps the most complete description of mammalian memory formation to date.

Brain Substrates of Eyeblink Conditioning

Electrophysiological recordings of neural activity in eyeblink conditioning indicate that a number of brain areas and systems become engaged, most prominently motor nuclei, the hippocampus, and the cerebellum. In all these systems neuronal activity in the CS period increases over the course of learning and precedes the onset of the behavioral CR within trials. Indeed, this increased pattern of action potential discharges in the hippocampus and cerebellum predicts the occurrence and the actual temporal form of the CR (but not the reflex response). In standard delay conditioning when the CS and the US overlap, the cerebellar system is critical but the hippocampal system is not. However, in trace conditioning both the cerebellar and hippocampal systems are critical. Within the cerebellar system, the following regions developed the predictive neuronal response: relevant motor nuclei, a region of the fifth nucleus, various reticular brain stem regions, the cerebellar cortex, the cerebellar interpositus nucleus, the pontine nuclei, and the red nucleus.

The Cerebellar System

Lesions A series of studies have been able to identify a nuclear region in the anterior lateral interpositus nucleus ipsilateral to the trained eye that is essential to the acquisition and retention of the CR. Very large lesions of the cerebellar cortex that did not damage the interpositus nucleus did not abolish the behavioral CR, although CR latency was altered such that the eye closed and opened before the onset of the US – the timing of the behavioral response was no longer adaptive.

In these initial studies it was shown that the interpositus lesion effect was ipsilateral – it abolished the eyeblink CR on the side of the lesion but did not impair eyeblink conditioning of the contralateral eye, providing a control for possible nonspecific or state variables. Furthermore, if the lesion is made before training, learning is completely prevented. These lesions had no effects at all on performance of the UR. Importantly, lesions of the interpositus 1 month after training completely abolish the eyeblink CR – the interpositus is the site of long-term storage of memory.

It is important to emphasize that the cerebellar interpositus nucleus is essential for the learning of all discrete movements trained with an aversive US: eyeblink, limb flexion, head turn, etc. To the extent tested, the memory traces for all those learned movements are stored in the interpositus nucleus in separate loci.

The Pathways

A large literature is in general agreement in identifying the essential circuitry for classical conditioning of eyeblink and other discrete responses (see **Figure 3**). This circuitry follows closely the well-established anatomy of the cerebellar system and is in general accord with classical theories of cerebellar learning.

The UR pathway The eyeblink reflex is a coordinated response involving simultaneous and perfectly correlated external eyelid closure, eyeball retraction, and resulting passive extension of the nictitating membrane, as noted above. In terms of the reflex pathways, there are direct projections from neurons in regions of the trigeminal nucleus to eyeblink motor nuclei.

The CR pathway Neurons in a localized region of the interpositus nucleus ipsilateral to the trained eye develop a neuronal model of the learned behavioral

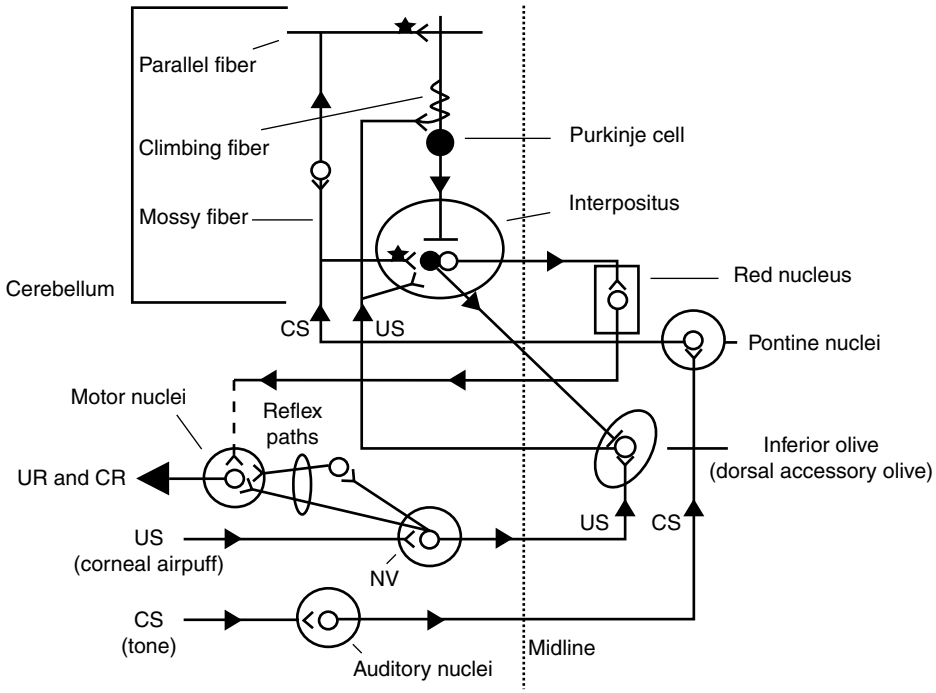

Figure 3 Simplified schematic (most interneurons omitted) of the putative essential circuitry for delay classical conditioning of eyeblink (and other discrete responses) learned with an aversive US. (The sensory and motor nuclei activated depend of course on the nature of the CS and US – the more central portions of the circuit appear to be general.) The reflex US–UR pathway involves direct and indirect projection from the trigeminal nucleus (NV) to the motor nuclei (for the eyeblink UR and CR, primarily accessory 6 and 7). The tone CS pathway projects from auditory nuclei to the pontine nuclei and to the cerebellum as mossy fibers. The US pathway includes projections from the trigeminal to the inferior olive (IO) and to the cerebellum as climbing fibers. The CR pathway projects from the interpositus to the red nucleus and on to premotor and motor nuclei. There is also a direct GABAergic inhibitory projection from the interpositus to the IO. Omitted from this diagram are feedback projections from the cerebellum to several afferent structures. These projections have generally been described physiologically, not anatomically, and may well be involved in some aspect of CR modulation. See text for discussion. Solid cell bodies and bar terminals indicate inhibitory neurons, open cell bodies and fork terminals indicate excitatory neurons. Stars indicate sites of plasticity based on current evidence. See text for details. Reproduced from Thompson RF (1986) The neurobiology of learning and memory. *Science* 233: 941–947.

CR; lesions of this region selectively abolish the CR with no effect in the UR; electrical stimulation of this region evokes the eyeblink response before training (as noted above). This region of the interpositus projects via the superior cerebellar peduncle (SCP) to a region of the contralateral magnocellular red nucleus. Lesions of the peduncle abolish the CR, as do lesions of the red nucleus, where neurons also show a model of the learned response. Stimulation of the red nucleus also elicits the eyeblink response. The descending rubral pathways project contralaterally to premotor and motor nuclei for external eyelid closure. Overwhelming evidence identifies this circuit as the efferent CR pathway for the conditioned eyeblink response.

The CS pathway The pontine nuclei send axons as mossy fibers directly to the cerebellar cortex and interpositus nucleus, mostly contralaterally. The pontine nuclei in turn receive projections from auditory, visual, somatosensory, and association systems, both cortical and subcortical. Appropriate lesions of the pontine nuclei can abolish the CR established to a tone CS but not a light CS, that is, can be selective for CS modality (interpositus lesions abolish the CR to all modalities of CS). Extensive lesions of the middle cerebellar peduncle (MCP), which conveys mossy fibers from the pontine nuclei and other sources to the cerebellum, abolish the CR to all modalities of CS. Electrical stimulation of the pontine nuclei serves as a 'supernormal' CS, yielding more rapid learning than does a tone or light CS and stimulation of the MCP itself is an effective CS.

The US pathway Neurons in the inferior olive (IO) send climbing fiber projections contralaterally directly to cerebellar cortex and interpositus nucleus. The critical region of the IO for eyeblink conditioning is the dorsal accessory olive (DAO), which receives predominantly somatosensory input relayed from the spinal cord and appropriate cranial nuclei. Lesions of the critical region of the IO, the face representation

in the DAO, completely prevent learning if made before training and result in extinction of the CR if made after training. Neurons in this critical DAO region respond only to US onset, show no learning-related activity, and the US-evoked response decreases as animals learn. Electrical microstimulation of this region serves as a very effective US. All these data argue that the DAO-climbing fibers system is the essential US reinforcing pathway for the learning of discrete responses but the DAO is not the locus of the memory trace.

The interpositus nucleus sends direct GABAergic projections to the DAO. Hence, as learning-induced increases in interpositus neuron activity develop, inhibition of the DAO neurons will increase. This accounts for the fact that US-evoked activity in the DAO decreases as learning develops. This also appears to serve as a part of the neural circuit essential for the behavioral learning phenomenon of 'blocking,' where prior training to one CS, for example, tone, prevents subsequent learning to a light CS when it is then presented together with the tone in paired compound stimulus training. Infusion of picrotoxin in the DAO to block the γ-aminobutyric acid (GABA) inhibition from the interpositus during compound stimulus training completely blocks the development of behavioral blocking.

Conjoint activation of CS and US pathways If the above hypotheses concerning the identities of the CS and US pathways are correct, it should be possible to train behavioral CRs by conjoint stimulation of these pathways. Stimulation of the pontine nuclei – mossy fibers as a CS (below movement threshold) and DAO-climbing fibers as a US yielded normal learning of the response provoked by the US. The US-elicited movements included eyeblink, head turn, and limb flexion. These movements were learned to the mossy fiber stimulation CS, just as was the case for peripheral CSs.

Reversible inactivation Although the evidence cited above is consistent with the interpositus nucleus being the locus of the conditioned eyeblink memory trace, it does not prove this hypothesis. Predictive neuronal models of the CR develop in the interpositus nucleus, the red nucleus, and the motor nuclei, as noted. Further, lesions of the interpositus, of the SCP (conveying all the output from the interpositus), and of the red nucleus, all abolish the CR with no effect on the UR. Lesions of the motor nuclei of course abolish both the CR and the UR. The technique of reversible inactivation was used to determine where within the circuit the memory trace was formed and stored.

Several parts of the cerebellar circuit, illustrated in Figure 4, have been reversibly inactivated for the duration of training (eyeblink conditioning) in 'naïve' animals using muscimol or cooling for neuron somas and tetrodotoxin. In brief, inactivation of the SCP, the

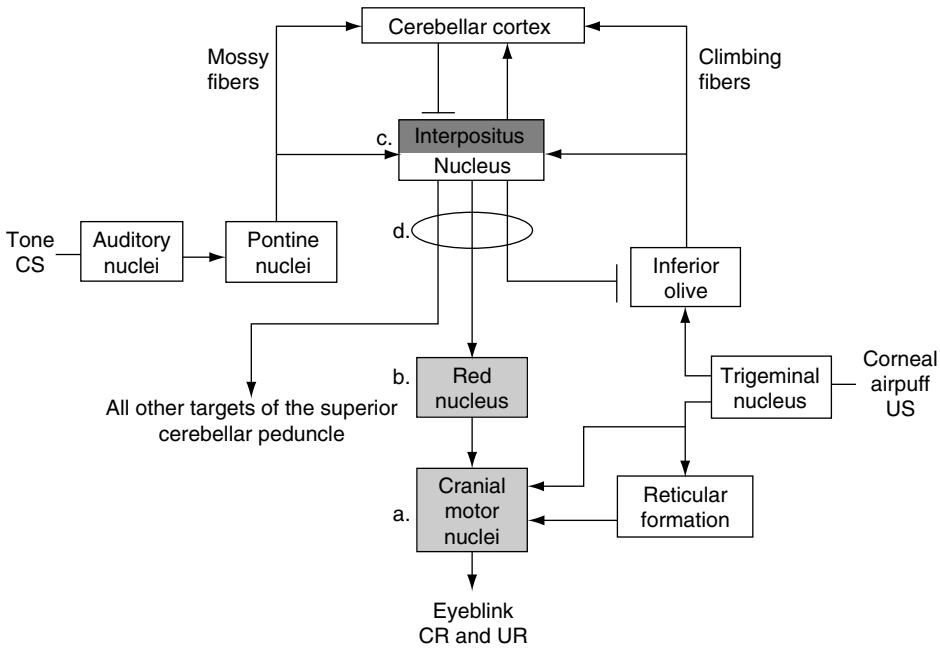

Figure 4 Highly simplified schematic of the cerebellar memory circuit to indicate reversible inactivation during training experiments. a, motor nuclei; b, red nucleus; c, interpositus; d, superior cerebellar peduncle (SCP). See text for details. Modified from Thompson RF (2005) In search of memory traces. *Annual Review of Psychology* 56: 1–23.

red nucleus, and the motor nuclei prevent expression of the CR but do not prevent learning at all. In contrast, inactivation of the anterior interpositus nucleus not only prevents expression of the CR but also completely prevents learning. Collectively, these data strongly support the hypothesis that the memory trace is formed and stored in the interpositus nucleus.

Mechanisms of memory storage in the interpositus Extremely important direct evidence for a strengthening of the mossy fiber – interpositus neuron synapses – has been demonstrated using eyeblink conditioning in the rat. Thus, highly significant increase in the number of excitatory synapses in the interpositus nucleus was found following eyeblink conditioning with no change in inhibitory synapses. Further, infusion of GABA antagonists during training completely prevents learning. Infusion of AP5 (which blocks the NMDA receptor) into the interpositus markedly impaired acquisition, but not retention of the eyeblink CR. Finally, infusion of protein synthesis inhibitors (both transcription and translation) into the region of the interpositus nucleus impaired acquisition of the eyeblink CR but had no effect on expression of the CR. Thus, protein synthesis in the interpositus nucleus is necessary for learning of the eyeblink CR, but not for its expression once learned. All these results argue for the formation of neuronal/synaptic plasticity, a 'memory trace,' in the interpositus itself.

In sum, the evidence is now very strong from behavioral, physiological, pharmacological, anatomical, and reversible inactivation studies that the basic associative memory trace in eyeblink conditioning is established and maintained in the interpositus nucleus.

Cerebellar cortex It is clear that the cerebellar cortex plays a very important role in normal learning – animals with large cortical lesions learn more slowly than normal and not as well. Indeed, animals with no functional cerebellar cortex (Purkinje cell degeneration mutant mice) do learn, albeit more poorly than normal. A clear result of cortical damage is altered timing of the CR. In normal animals, the CR peaks at the time of onset of the UR; it is maximally adaptive. After cortical lesions, the CR typically peaks prior to US onset and is not adaptive.

Eyeblink Conditioning and the Hippocampus

In eyeblink conditioning, neuronal unit cluster recordings in hippocampal fields increase in discharge frequency during training trials very rapidly, shift forward in time as learning develops, and form a predictive 'temporal model' of the learned behavioral response, both within trials and over the trials of training. To summarize a large body of research, the growth of the hippocampal unit response is, under normal conditions, an invariable and strongly predictive concomitant of subsequent behavioral learning. This increase in neuronal activity in the hippocampus becomes significant by the second or third trial of training, long before behavioral signs of learning develop, as would be expected of a declarative memory system. This initial hippocampal unit increase is in the US period; increases in the CS period appear at about the time point in training when behavioral CRs appear. With continued training, the neuronal model eventually declines and the hippocampus appears to disengage.

Brain Substrates Trace Eyeblink Conditioning

Trace conditioning was first described by Pavlov; the CS terminates and there is a period of no stimulation between CS offset and US onset (as Pavlov stressed, the organism must maintain a 'trace' of the CS in the brain in order for the CS and the US to become associated). As noted above, the acquisition and retention of delay eyeblink conditioning require the cerebellum and associated brain stem structures. No forebrain structures, including the hippocampus, are required. Thus, decerebrate rabbits with no remaining forebrain tissue (removal of cerebral cortex, basal ganglia, limbic system, thalamus, and hypothalamus) exhibited normal retention of delay eyeblink conditioning. Findings in humans are consistent with the animal work. Thus, delay eyeblink conditioning was impaired in patients with cerebellar or brain stem lesions, but was intact in amnesic patients with damage that includes the hippocampus.

Successful trace eyeblink conditioning, like delay conditioning, requires the cerebellum. However, trace conditioning differs from delay conditioning in that it also requires the hippocampus and portions of neocortex. Thus, acquisition and retention of trace conditioning were severely disrupted in rabbits and rats when the hippocampus was damaged and trace conditioning in rabbits was disrupted by damage to prefrontal cortex. Again, findings in humans are consistent. In amnesic patients with damage that includes the hippocampus, trace eyeblink conditioning was mildly impaired with a trace interval of 500 ms, and severely impaired with a trace interval of 1000 ms. Another distinctive feature of trace conditioning is that the importance of the hippocampus is time-limited. When hippocampal lesions were made in rabbits 1 day after acquisition, trace

conditioning was abolished, but lesions made 30 days after acquisition had no effect. Thus, trace conditioning exhibits two important characteristics of declarative memory. First, trace conditioning requires the hippocampus. Second, in trace conditioning, as in other forms of declarative memory, the hippocampus has a time-limited role.

Awareness, Expectancy, and Human Eyeblink Conditioning

The difference between delay and trace eyeblink conditioning has recently come into sharper focus by studies that have explored the importance of awareness. Declarative memory typically includes knowledge (or awareness) about what has been learned. In contrast, nondeclarative memory does not require awareness of any memory content, and when awareness is present it appears to be unrelated to task performance. With respect to eyeblink conditioning, the questions of interest are: when conditioning occurs, do individuals become aware of the relationship between the CS and US, namely, that the CS precedes and predicts the US? Is awareness of the stimulus contingencies necessary for successful conditioning to occur? Does the importance of awareness differ for delay and trace conditioning?

Awareness and Eyeblink Conditioning

When participants were given trace conditioning while watching a silent movie or while performing an attention-demanding digit-monitoring task the participants who watched the movie emitted more CRs, and performed better on a seven-item postconditioning questionnaire that assessed awareness, than did participants who were distracted by the digit-monitoring task. Other participants watched the silent movie during conditioning and took the seven-item questionnaire after the first 10 conditioning trials, again after 60 trials, and finally after all 120 trials were completed. Those who became aware of the CS–US relationship early in the session (after 10 trials) conditioned to a greater extent than those who became aware of the relationship later in the session or who never became aware at all. Further, performance on the questionnaire after the first 10 conditioning trials predicted the magnitude of trace conditioning over the entire 120-trial conditioning session.

In contrast to these findings for trace conditioning, awareness was unrelated to delay conditioning. Thus, participants who became aware of the stimulus contingencies early in the conditioning session conditioned no better than participants who became aware later in the session or did not become aware. Further, there was no relationship between awareness scores obtained early in conditioning and the magnitude of conditioning across the 120-trial conditioning session.

Expectancy of the US and Delay and Trace Conditioning

Individuals who develop awareness of the stimulus contingencies may be successful at trace conditioning because they have acquired a representation that allows them to expect the US when the CS is presented. During the course of conditioning, expectation of the US can be changed in predictable ways by capitalizing on a phenomenon known as 'the gambler's fallacy.' Using this method, expectancy of the US increases during strings of CS-alone trials and decreases during strings of CS + US trials. Accordingly, before each trial, participants are asked to rate how much they 'expect' the US (airpuff) to appear on the next trial.

In one study the performance of participants given delay conditioning was compared with the performance of those given trace conditioning. For both groups, expectancy of the US varied in accordance with the gambler's fallacy. For delay conditioning, the probability of a CR was related not to expectancy but to the associative strength of the CS and US. The results for trace conditioning were opposite. The probability of a CR was positively related to expectancy of the airpuff US. That is, CR probability was high when expectancy of the US was high and low when expectancy of the US was low. These findings show that expectation of the US has a different role in trace conditioning than it does in delay conditioning and support the idea that trace conditioning and delay conditioning are fundamentally different in their dependence on higher-order cognitive processes like awareness.

Summary and Conclusions

The findings outlined above can be understood in terms of the declarative and nondeclarative memory systems that support eyeblink classical conditioning. In both delay and trace conditioning paradigms, individuals sometimes develop declarative (conscious) knowledge about the stimulus contingencies and sometimes do not. For the most commonly studied forms of delay conditioning, declarative knowledge is superfluous to the acquisition of the CR, and cerebellar and brain stem circuits can support performance. Trace conditioning is fundamentally different. Trace conditioning resembles delay conditioning in that it also depends on the cerebellum, but trace conditioning is additionally dependent on the hippocampus

and neocortex. Further, unlike delay conditioning, trace conditioning is strongly related to the acquisition of declarative knowledge (awareness) of the CS–US contingencies and to the degree to which the US is expected.

For both trace and delay conditioning the cerebellum is responsible for the acquisition, storage, and generation of the CR. Trace conditioning may additionally require declarative knowledge because the trace interval makes it difficult for the cerebellum to process the CS and the US and form a motor memory. In trace conditioning, because the US follows the CS by as much as 1000 ms the cerebellum may not be able to maintain a representation of the CS across the trace interval. If, however, the hippocampus and neocortex have represented the stimulus contingencies, then perhaps processed information about the CS can be transmitted to the cerebellum at a time during each trial that is optimal for cerebellar plasticity (immediately before and during the US). Thus, an interesting possibility is that in trace conditioning, the cerebellum receives reformatted information, such that CS and US information arrive at the cerebellum in the temporally overlapping fashion that the cerebellum can use. By this view, awareness of the contingencies, while not directly driving the acquisition of the CR by the cerebellum, serves as an indicator that the hippocampus and related structures are effectively engaged by the task and working with the cerebellum so that the trace-conditioned response can be acquired.

See also: Conditioning: Theories; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Eyeblink Conditioning; Motor Skill Learning; Procedural Learning in Animals; Procedural Learning: Cerebellum Models.

Further Reading

Christian KM and Thompson RF (2003) Neural substrates of eyeblink conditioning: Acquisition and retention. *Learning and Memory* 10(6): 427–455.

Clark RE (2004) The classical origins of Pavlov's conditioning. *Integrative Physiological and Behavioral Science* 39(4): 279–294.

Clark RE, Manns JR, and Squire LR (2002) Classical conditioning, awareness, and brain systems. *Trends in Cognitive Science* 6(12): 524–531.

Clark RE and Squire LR (2000) Awareness and the conditioned eyeblink response. In: Woodruff-Pak DS and Steinmetz JE (eds.) *Eyeblink Classical Conditioning: Human*. Boston, MA: Kluwer Academic Publishers.

Davis M (1992) The role of amygdala in fear and axiety. *Annual review of Neuroscience* 15: 353–375.

Fanselow MS and Poulos AM (2005) The neuroscience of mammalian associative learning. *Annual Review of Psychology* 56: 207–234.

Gormezano I, Prosy WF, and Thompson RF (eds.) (1987) *Classical Conditioning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kleim JA, Freeman JH Jr., Bruneau R, et al. (2002) Synapse formation is associated with memory storage in the cerebellum. *Proceedings of the National Academy of Sciences of the United States of America* 99(20): 13228–13231.

Maren S (2001) Neurobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience* 24: 897–931.

McGaugh JL (2004) The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience* 27: 1–28.

Ohyama T, Nores WL, Murphy M, and Mauk MD (2003) What the cerebellum computes. *Trends in Neuroscience* 26(4): 222–227.

Thompson RF (1986) The neurobiology of learning and memory. *Science* 233: 941–947.

Thompson RF (2005) In search of memory traces. *Annual Review of Psychology* 56: 1–23.

Voneida TJ (2000) The effect of brachium conjunctivum transection on a conditioned limb response in the cat. *Behavioural Brain Research* 109(2): 167–175.

Woodruff-Pak DS and Steinmetz JE (eds.) (2000) *Eyeblink Classical Conditioning*, 2 vols. Boston, MA: Kluwer Academic.

Procedural Learning: Striatum

J D Berke, University of Michigan, Ann Arbor, MI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Following a century of debate over the rules governing learning and memory, it is now broadly accepted that there are distinctly different kinds of learning, reflecting the operation of multiple memory systems in the brain. In part, this reflects the fact that experience-dependent plasticity is widespread (if not ubiquitous) in the mammalian central nervous system. Beyond this, it is clear that we place on memory a variety of demands, which require different, even conflicting aspects of our experience to be stored. For example, for some purposes, it is important to remember what happened on a particular, unique occasion (e.g., where did I park my car this morning?), but for other purposes, we wish to generalize across different episodes of experience (e.g., which parking area is most likely to have free spots?). Just as the brain uses different visual pathways for processing object identification versus object location, it seems to use distinct brain circuits to preserve or discard different features of our ongoing actions and experiences. The overall function of neural circuits that involve the striatum seems to be to resolve competition between different possible choices (decision making) on the basis of how well they have worked in the past. Choices that lead to desirable outcomes, especially if they are unexpected, provoke plasticity within the striatum, which increases the likelihood that these choices will be selected again. This core algorithm is termed reinforcement learning.

The term procedural (or implicit) learning is defined negatively – it includes a broad range of experience-dependent improvements in behavioral task performance that are not declarative. This means that procedural performance is not dependent on the subject's either being able to recall specific training episodes or being able to explicitly state the acquired knowledge. The best-known forms of procedural learning are within the motor domain, but procedural learning can also include performance on cognitive tasks and emotional/motivational features of memory. Some aspects of procedural learning that have been commonly ascribed to the striatum are habits and skills. Both terms refer not to the specific procedures that are learned but rather to certain attributes of learned performance that develop incrementally with extended practice. Actions become habits when they have been performed so often that they can be done automatically, i.e., without effortful attention. To continue with automotive examples, driving a car along a highly familiar route can easily be done while focusing on a conversation with a passenger. No goal needs to be in mind to perform this habit. Rather, we often need to pay attention to avoid habitual/automatic responding – if engrossed in conversation, we may find ourselves having driven this very familiar route even if another destination had been initially intended. As habits are reactions to familiar configurations of cues, they are sometimes referred to as stimulus–response (S-R) associations, in contrast to action-outcome associations. Because habitual behavior is not goal-directed, it is somewhat inflexible when circumstances change. Habits are relatively persistent even when the consequences of actions are removed (extinction) or made less appealing (outcome devaluation). Finally, practicing a behavior leads to more-skilled performance, in which the components of an action sequence become more fluidly and accurately joined together.

The following paragraphs summarize in turn some of the key anatomical and physiological features of the striatum relevant to learning, the effects of striatal damage and drugs on learning in humans and animals, and results from neuroimaging and electrophysiological investigations of the striatum during learning. Although plasticity within portions of the striatum is indeed responsible for certain forms of procedural learning such as habits, the overall function of striatal plasticity appears to be far more elaborate than simply connecting stimuli with responses.

Basal Ganglia Circuitry and Plasticity

Diverse Information in Afferent Inputs

The striatum is one component of large-scale cortical-basal ganglia loop circuits (Figure 1). Striatal subregions make use of rich patterns of information contained in inputs from almost all cortical (and thalamic) regions. These include sensory association areas that might contribute to the stimulus of S-R habits, motor cortical areas that provide information about current motor commands in the corticospinal pathway (efference copy), and prefrontal areas concerned with cognitive and emotional/motivational information. The striatum projects to other nuclei of the basal ganglia, which in turn project back to thalamus and thence to anterior (motor/premotor/prefrontal) cortical regions. As a rough approximation, the 'motor' loop involves the putamen (lateral/dorsal striatum), the

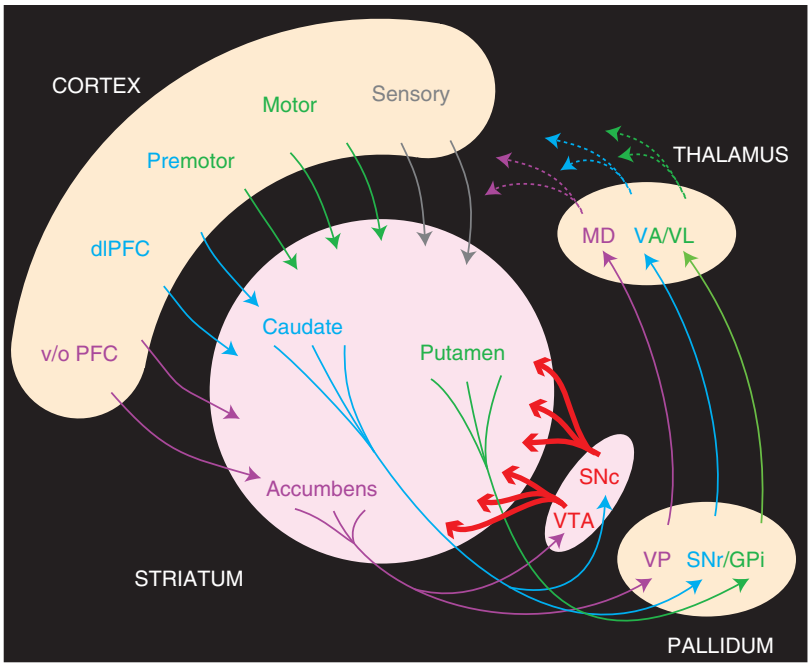

Figure 1 Simplified anatomy of cortex–basal ganglia loop circuits. The striatum is a convergence zone for a wide range of cortical and thalamic inputs, carrying rich patterns of information of diverse types. The motor loop is shown in green, cognitive loop in blue, and motivational loop in purple. Dopamine (in red) from midbrain neurons acts in the striatum to modulate all loops. For clarity, the connections from thalamus back to cortex and to striatum are truncated, and important additional basal ganglia pathways are omitted altogether. VTA is ventral tegmental area; SNc, substantial nigra pars compacta; VP, ventral pallidum; dlPFC, dorsolateral prefrontal cortex; v/oPFC, ventral/orbital prefrontal cortex; SNr, substantial nigra pars reticulata; GPi, globus pallidus pars interna; MD, medial dorsal thalamic nucleus; VA/VL, ventral anterior/ventral lateral thalamic nucleus.

'cognitive' or 'associative' loop involves the caudate (medial/dorsal striatum) and the 'limbic' loop involves the nucleus accumbens (ventral/medial striatum). However, there are no sharp neurochemical or histological boundaries within striatum that demarcate the different loop circuits, and the number of different loops is essentially arbitrary. The point is that all kinds of information reaches the striatum and can potentially form part of the content of learning.

Convergence of Corticostriatal Information

The architecture of striatal afferents allows the integration of information from different sources. First, the terminals of projections from different cortical regions show considerable intermingling within striatum. In other words, loop circuits bearing different types of information are not strictly segregated to distinct striatal subregions. Second, striatal (principal) neurons each receive thousands of inputs, provided by nearly as many different axons (because each axon makes only a very small number of terminals on any given striatal neuron). Such convergence of information potentially allows striatal neurons to serve as pattern recognition devices, in which their activity is highly dependent on specific combinations of external and internal cues and contexts. Convergence is also implied by the relative number of neurons in cortex and basal ganglia nuclei. There are approximately ten times more cortical than striatal neurons, and the ratio of striatal neurons to their pallidal targets is even larger.

Dendritic Spines and Striatal Plasticity

More than 90% of striatal neurons are the medium spiny cells, which both receive incoming afferents and themselves project out of the striatum. The spines that densely cover the dendrites of these cells (Figure 2) are believed to be sites of synaptic plasticity that support striatum-based learning. It is worth noting that spiny cells are comparably spiny in all striatal subregions – hence, the striatal components of cognitive and/or motivational loops likely have as much capacity to support learning as motor regions of the striatum. Long-term potentiation (LTP) and long-term depression of corticostriatal synapses have been demonstrated both *in vitro* and in anesthetized *in vivo* preparations. In addition to the spiny cells, there are

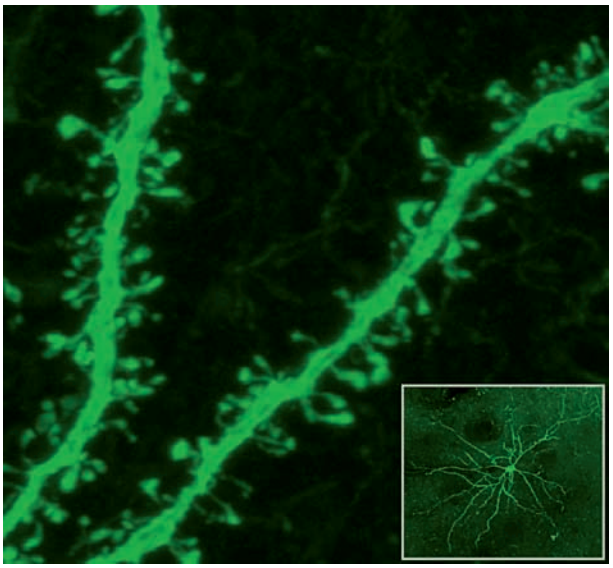

Figure 2 The dendrites of striatal projection neurons are densely coated in spines. Shown are examples of dendritic segments labeled with green fluorescent protein. Inset: example of whole dendritic field of a medium spiny neuron. Image from Jakub Jedynak and Terry Robinson, University of Michigan.

at least four distinct populations of interneurons, and some of these have been shown to regulate plasticity at spiny cell synapses.

Dopamine Input

A relatively small number of midbrain dopamine cells branch out to innervate wide areas of forebrain. This projection seems especially important to striatal function as the striatum is packed with dopamine terminals and receptors to a far greater extent than other forebrain areas. For the most part, dopamine neurotransmission and intracellular signaling appear set up not for conveying spatially rich patterns of information but rather for providing a more global signal that influences broad regions of striatum in a similar fashion. Most (though not all) investigators believe that this widespread dopamine signal has a central role in striatum-dependent learning (see the section titled 'Dopamine and striatal learning'). This idea is supported by at least some of dopamine's actions within striatum. Dopamine enhances long-term potentiation of corticostriatal synapses, probably at least in part because dopamine D1 receptors are coupled to the intracellular cyclic adenosine monophosphate (cAMP)/protein kinase A/cAMP responsive element-binding protein pathway, leading to increased transcription of plasticity-related genes. Manipulations of striatal dopamine (including chronic administration of addictive drugs) cause changes in striatal dendritic spines that likely reflect alterations in synaptic connectivity. Hence the machinery for dopamine-modulated learning is present and is almost certainly used.

Consequences of Striatal Damage for Learning

It is worth distinguishing between several quite different ways in which the striatum can contribute to learning. First, neural changes within the striatum itself – such as synaptic plasticity at spines – can affect behavior. These mechanisms could be the permanent repository of learning or have a more time-limited role as consolidation proceeds elsewhere. Second, the striatum is an essential part of circuits involved in the cognitive manipulation of information. Striatal damage can thereby affect encoding/plasticity in other structures and also the way information stored in other structures is employed for acute task performance. These different senses of striatum-dependent learning can confound investigators' ability to use striatal damage as a clue to specific striatal functions, and this difficulty has contributed to confusion in the literature.

Humans

Two relatively common neurological disorders have often been used to study striatal learning: Parkinson's disease (PD) involves a progressive degeneration of the dopamine input to the striatum, while Huntington's disease (HD) features progressive degeneration of the striatal spiny neurons themselves. The most obvious behavioral deficits in patients with advanced cases of either disorder are not learning difficulties but rather severe motor difficulties, often extending into the cognitive domain, and the resulting problems in task performance have often complicated the interpretation of learning studies. Both types of patient are relatively unimpaired in tests of declarative memory that stump amnesiacs with medial temporal lobe damage, and both types also learn normally in some procedural tasks. However, on some other tasks considered to be procedural, they have more difficulty. These include select forms of motor skill learning, category–classification learning, and sequence learning.

A well-known example is the weather prediction task, which involves probabilistic classification. On each trial, participants select the outcome (rain or shine) they think most likely after the presentation of a set of stimuli, each of which is associated with a certain outcome probability. The use of combinations of probabilistic cues makes it hard to consciously infer and calculate the most likely answer. Patients with PD or HD show retarded learning in such tasks;

they also often make use of suboptimal strategies, such as being guided by the presence or absence of one cue rather than making use of the full set. An analogous pattern of deficits is observed in information integration forms of category learning, in which good performance requires participants to use multiple aspects of visual patterns (e.g., color, shape, number) when classifying them. PD patients appear to have particular trouble making use of feedback about whether their responses were correct, and (in unmedicated patients) this deficit seems more severe for feedback about correct responses than about incorrect responses. Such deficits tie in well with the anatomy and physiology described above. Much of category–classification learning involves, in essence, using feedback to develop a mapping between a complex stimulus space and a simpler set of responses. PD patients would have a compromised dopamine reinforcement learning signal acting on spiny neurons that integrate multiple sources of information in the selection of actions.

Since an important part of the inputs that striatal spiny neurons receive are axon collaterals from the corticospinal tract, these cells are well positioned to use current motor commands as part of the information used to compute or anticipate the next movement. This may be why striatal damage has often been reported to interfere with sequence learning. Many tests of PD and HD patients have shown deficits in serial reaction time (SRT) tasks, in which participants are sequentially cued to perform a set of actions, and learning is measured as lower reaction times during trained sequences compared with random sequences. However, there is considerable inconsistency in this literature, with a number of reports of preserved SRT learning in PD patients. Recent studies have attempted to break down the demands of SRT tasks, for example by dissociating the spatial and temporal aspects of sequence presentation or the spatial and motor response demands. These studies have suggested that, once again, it is the integration of different forms of information in the service of action selection that presents particular problems for patients with striatal damage.

Striatal damage also frequently produces deficits similar to those seen in patients with damage to the frontal cortex. Both PD and HD patients have difficulty with executive function (broadly construed), attentional set shifting, and the learning and application of abstract rules to guide behavior. For more-focal forms of striatal damage, it has been noted that the behavioral effects of lesions to striatal subregions typically resemble those of focal lesions to the interconnected frontal cortical regions. To follow the distinction drawn earlier, these frontal deficits likely reflect both a problem with engaging intrastriatal plasticity using information provided by frontal cortex and the fact that loop circuits through the basal ganglia play a vital, continuous role in the ongoing manipulation of information in frontal cortical regions.

Experimental Animals

The first suggestion that the striatum is specifically involved in procedural learning came from Mishkin and colleagues in 1984. They had observed that monkeys with medial temporal lobe damage performed very poorly when required to store trial-unique information, such as which of two objects had been baited with food just a few seconds previously. But when given repeated presentations each day of the same pair of objects, these monkeys could progressively learn to select the consistently rewarded object. Comparable results for humans with amnesia due to medial temporal lobe damage have been reported recently. Mishkin proposed that this slow mode of learning relied on information flow from visual object identification areas in temporal cortex to the striatum, and that "this system of [corticostriatal] projections therefore provides a mechanism through which cortically processed sensory inputs could become associated with motor outputs generated in the pallidum and so yield the S-R bonds that constitute habits." Subsequent lesion studies confirmed that this progressive learning to select the consistently rewarded object is impaired by damage to ventrocaudal neostriatum, the portion of monkey striatum that receives input from cortical areas involved in visual object identification (but note results discussed under the section 'Electrophysiology of single striatal neurons and learning'). The suggestion that the striatum can serve as a switchboard connecting patterns of sensory inputs to motor outputs has an attractive simplicity and seems at least superficially consistent with the human deficits in integrating information for action selection that are produced by striatal damage.

The effects of striatal lesions on learning have been most thoroughly investigated in rats, where impressive double dissociations with the effects of hippocampal damage have broadly confirmed the importance of striatum for procedural learning. The deficits produced by striatal damage have been interpreted as problems with S-R learning, although what exactly this means has varied in substantial ways, as the following examples demonstrate.

Consider first performance in a four-arm, or plus, maze, in the classic situation devised by Tolman to distinguish between S-R habits and the use of cognitive maps. Hungry rats are repeatedly placed on the

end of one arm (call it south) with food available at the end of another arm (west). They progressively learn to rapidly move to the food, which involves making (in this example) a left turn to get to the goal. After certain amounts of training, they are given probe trials in which they start at the opposite arm (north) and must choose east or west. Going to the spatial location (west) where food has been presented in the past is called a 'place' strategy, while making the same left turn as before and ending up in the east arm is called a 'response' strategy. A standard finding is that with a few days' training in the task, most rats use a place strategy, but with extended training, most rats use a response strategy. Evidence that these distinct strategies rely on distinct brain systems comes from experiments involving intracerebral infusions of the local anesthetic lidocaine to temporarily block activity in the infused area. On probe trials, lidocaine into dorsal hippocampus causes rats with little pretraining to switch early to the response strategy, while lidocaine into lateral striatum causes even rats with extended training to still use the place strategy. The striatum is therefore important for incrementally acquiring an 'egocentric' (body-centered) response tendency, which seems consistent with an S-R learning role (even if the stimulus in this case is a bit vague).

A clear dissociation between striatal and hippocampal contributions to learning has also been demonstrated between spatial and cued variants of the circular water maze. Dorsal striatal damage interferes with the ability to learn to approach a specific cue that marks a platform allowing escape from the water, but not with the ability to learn that the escape platform is in a consistent spatial location. This has also been interpreted as S-R learning although the nature of the response is quite different. In the plus-maze example, the specific egocentric turn is an arbitrary response that must be learned. By contrast, we have an innate tendency to orient toward and approach stimuli that are attractive (i.e., have motivational significance), making the interpretation of approach responses more complicated.

Another example of learning an approach-type response is in the win-stay version of the radial maze. Striatal damage impairs learning of this task, in which rats are rewarded for choosing arms that are marked with a consistent cue (e.g., lights). By contrast, hippocampal damage causes deficits when rats must make use of trial-unique information, such as which arms have been recently visited (a win-shift task). Such results emphasize the role of the striatum in learning invariances – constant rather than shifting features of the world. In the same vein, rats with striatal damage are impaired in reference memory maze tasks, in which the same set of arms is consistently baited, while hippocampus damage instead causes problems with working memory versions, in which the rat must keep track of and avoid arms that have already been foraged (**Figure 3**).

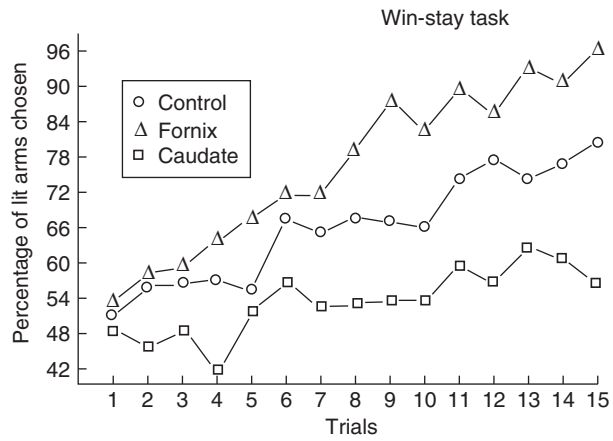

Figure 3 Striatal damage interferes with win-stay learning. Compared to the control group, rats with damage to dorsal/lateral striatum (here labeled as caudate) are slower to learn to pick the lit arms of a radial maze to receive food rewards. By contrast, rats with damage to the fornix (an important set of hippocampal connections) learn faster than intact rats. This is most likely because the hippocampus helps keep track of which arms have been recently visited as part of a normal foraging strategy, which is unhelpful in this task. Reproduced from Packard MG, Hirsh R, and White NM (1989) Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *Journal of Neuroscience* 9: 1465–1472, copyright 1989 by the Society for Neuroscience, with permission.

From these studies and others that have compared the roles of particular striatal subregions, it seems reasonable to conclude that sensorimotor (dorsal/lateral) zones of striatum are important for the incremental acquisition and performance of specific responses to stimuli. A small number of studies have gone further and provided evidence that such striatum-dependent strategies are indeed habits, in the sense of being insensitive to outcome devaluation. By contrast, damage to more medial/ventral striatal regions has more cognitive effects, including difficulties with spatial learning, working memory, and behavioral flexibility. These deficits presumably reflect anatomical relationships to hippocampus and frontal cortex, although investigators currently lack a clear notion about what is particularly striatal about the striatal contributions to more cognitive forms of learning. At least some forms of striatum-dependent learning appear to involve synaptic plasticity within the striatum itself, as suggested by a requirement for functional *N*-methyl-D-aspartate receptors and protein synthesis in the key striatal subregions.

Evidence for a role for the basal ganglia in organizing plasticity elsewhere comes from songbirds. In avian brains, a neural circuit called the anterior forebrain pathway (AFP) has many physiological similarities to the mammalian basal ganglia. The AFP is necessary for initial song learning but not for the production of adult song. If normal adult birds are deafened, their song slowly deteriorates; however, this deterioration does not occur in birds whose AFP has been lesioned. From such results, it has been argued that the AFP is essential for guiding plasticity in motor pathways, comparing sensory feedback to an internally stored target template. At first glance, this seems rather different from the mammalian situation, in which damage to the basal ganglia causes overt problems, not just with learning, but also adult task performance. However, some effects of basal ganglia damage can be paradoxical. For example, observations that the motor problems of PD patients are at least partly alleviated by surgical destruction of regions of pallidum and thalamus suggest that dysregulated basal ganglia may be worse than none at all. Such results have led some to argue that mammalian basal ganglia circuits also are not as critical for routine behavior as for adapting to changing circumstances.

Dopamine and Striatal Learning

Dopamine acting in striatum has an acute influence on behavior; this influence might perhaps be summarized as altering decision thresholds for action selection, but this subject is not the focus of this chapter. The learning deficits of PD patients suggest that dopamine must also be present in the striatum for normal striatum-dependent learning, and this proposition is supported by animal studies. For example, rats without dopamine in dorsal/lateral striatum can still learn to press a lever in response to a cue in order to obtain a food reward (though they learn slowly). However, unlike intact rats, despite extensive overtraining, they stop responding if the food is devalued – indicating that habit formation is disrupted. Complementary studies have shown that drugs that enhance striatal dopamine release can facilitate striatum-dependent learning. For example, systemic or intrastriatal amphetamine injections can accelerate the rate of win-stay learning in rats. This effect is seen even if drugs are given shortly after the training session – indicating that dopamine is able to affect the consolidation of striatum-dependent learning.

Drugs that are addictive in humans share the common property of enhancing striatal dopamine release, and it has been proposed that dopamine-induced striatal plasticity is a key mechanism underlying drug addiction. Drug addiction is characterized by compulsion, the loss of control over drug intake, and this may reflect the development of increasingly strong drug-taking habits – that is, learned actions not under deliberate volitional control. In experimental animals, repeated exposure to psychostimulants and other addictive drugs can alter striatal dendritic spines, with a particular increase in spine density in the dorsal/lateral regions of striatum that are critical for habits.

Brief electrical stimulation of dopamine pathways can produce more-transient striatal dopamine increases that are still highly reinforcing in the sense that animals will learn to perform actions that are followed by such stimulation. The speed with which rats learn to press a lever for electrical stimulation is proportional to the extent that such stimulation enhances striatal LTP. The idea that transient changes in striatal dopamine might be a critical learning signal under normal circumstances as well has gained great traction. Wolfram Schultz and others have recorded from likely midbrain dopamine neurons and observed that they typically show brief increases in firing rate in response to unexpected rewards or cues that predict rewards (e.g., see **Figure 4(b)**). Reward-related cues result in transient increases in striatal dopamine, as detected by voltammetry. Dopamine cells fire less or not at all to predicted rewards or aversive stimuli (though novel stimuli of all kinds continue to provoke firing). In many respects, the activity of dopamine neurons may therefore encode a reward prediction error signal, a key component of reinforcement learning models. The use of error signals seems to be a common property of several brain systems performing various forms of procedural learning, including the striatum and cerebellum.

However, the relationship between dopamine fluctuations and learning is still disputed. Among the remaining problems, it is not clear why striatal dopamine increases on the timescale of minutes can be evoked by a wide variety of arousing events, including stressful stimuli that are not obviously positively reinforcing. Also, the latency of dopamine neurons to respond to stimuli is so short that it has been argued that they must be responding to simple inputs from the superior colliculus rather than from cortical regions that should be better able to judge the current motivational significance of the particular stimulus. Although the activity of dopamine neurons is clearly modulated as the result of prior experience, critical experiments have not yet been devised that would clearly establish that the striatum makes use of their activity specifically as a learning signal, in addition to or instead of more immediate purposes such as dynamically adjusting psychomotor arousal.

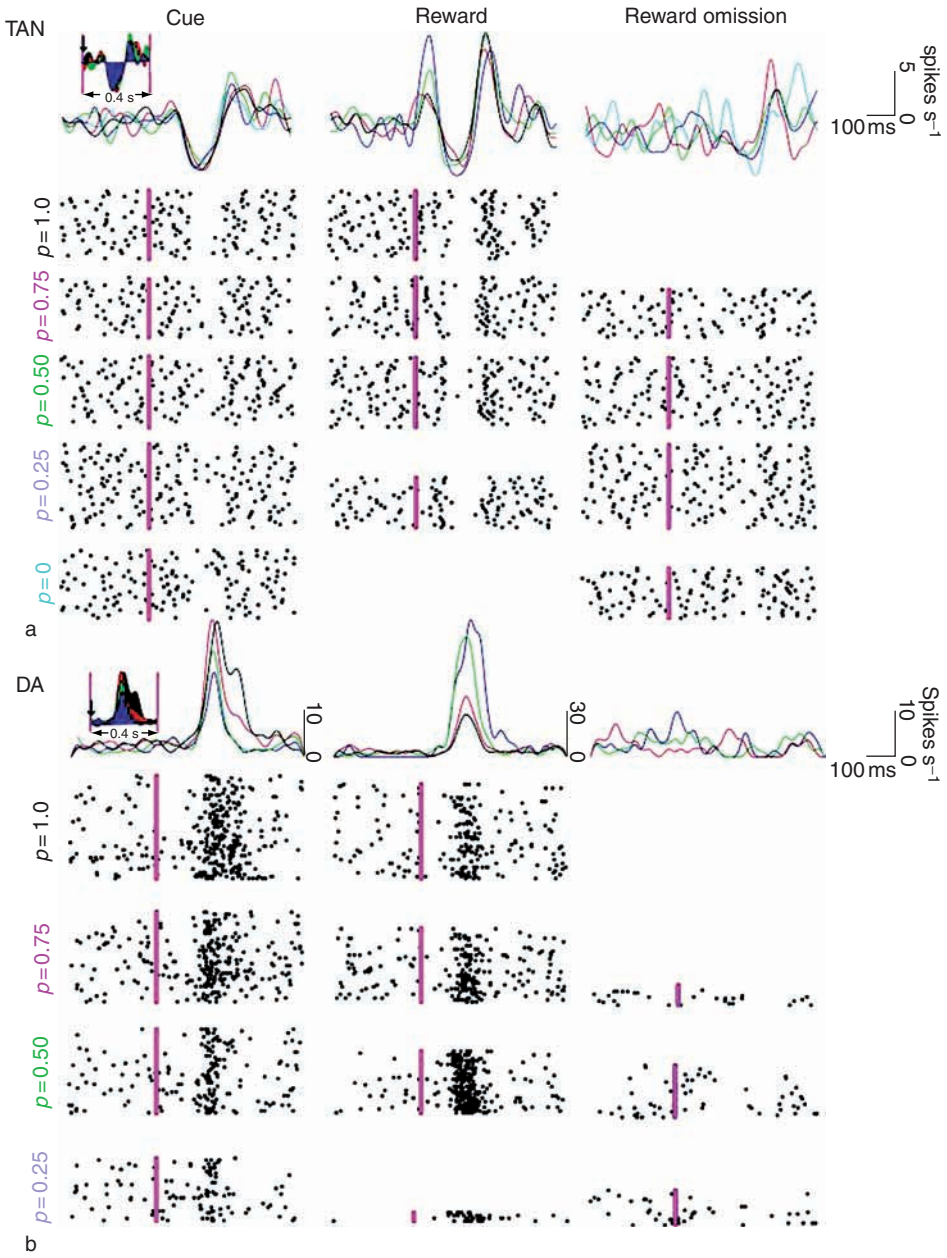

Figure 4 Comparison of the firing patterns of a monkey presumed striatal cholinergic interneuron (TAN) and midbrain dopamine cell (DA) in response to visual cues and the juice rewards that they probabilistically predict. Shown are raster plots (black ticks indicate spikes) and above them the averaged responses for each cue (e.g., $p = 0.5$ indicates that this visual cue is followed by reward on half of its trials). Note that the dopamine cell fires more to cues that predict likely reward and to unexpected rewards. By contrast, the cholinergic interneuron shows a pause in firing that does not systematically vary with reward probability. From Morris G, Arkadir D, Nevet A, et al. (2004) Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron* 43: 133–143.

Neuroimaging Studies of Striatal Learning

Recent years have seen a slew of interesting functional magnetic resonance imaging (fMRI) investigations of the human striatum, especially in reinforcement learning tasks. Before addressing the results, it is appropriate to sound one note of caution, namely, that we have essentially no idea what is responsible for activity as assessed by changes by the fMRI signal (using blood oxygenation level as contrast). In sensory cortex, blood oxygenation level signal seems to correlate with the local field potential, which in turn seems to reflect patterns of synaptic input more than the firing rates of local neurons. But the equivalent studies have not been performed

in the striatum, and the relative contributions of glutamatergic synapses, dopaminergic neurotransmission, interneuronal activity, and projection neuron spike firing are simply unknown.

Striatal fMRI activity changes can be observed in the putamen in response to simple, untrained movements, such as self-paced finger or toe flexion–extension. Many groups have examined how striatal fMRI activity changes as a result of procedural learning, often employing the SRT (sequence) tasks. The most common finding is that performing the sequence of finger movements engages more cognitive/associative parts of striatum early in learning, with a progressive shift to sensorimotor regions with further practice. This shift may parallel a change in the encoding of the sequence – while newly learned sequences are performed equally well with the hand used in training or the untrained hand, task performance becomes effector-specific with more practice. It is not clear whether striatal fMRI activity continues to persist even with very extended training; in one study, it was found that as actions become highly automatized (as assessed by sequence performance no longer causing interference with the performance of other tasks), activity in putamen disappeared. A progressive decrease in activity makes some intuitive sense, in that one might expect actions that are performed efficiently, accurately, and nonattentively to also be efficient in terms of neural energy requirements. It is less clear whether these changes in fMRI activity reflect (for example) more efficient neural representations of information stored in the striatum itself, versus (for example) the reduction in performance errors registered by the striatum and used to adjust neural representations elsewhere.

Inspired by the reinforcement learning paradigm, many groups have looked for, and found, correlations between error signals and striatal fMRI signal. All parts of striatum seem to be responsive to feedback, though the particular regions engaged can depend on the task involved. For example, accumbens activity correlates with reward prediction error in both Pavlovian (i.e., no choice) and instrumental (active choice) conditions whereas more-dorsal striatum responds to prediction error only in the instrumental condition. Paralleling the monkey dopamine cell studies, there is some evidence that the accumbens responds more to prediction errors about rewards than prediction errors about losses and that it also responds to salient nonrewarding stimuli. The caudate also responds to salient nonrewards, but only if they require some behavioral response. Prediction errors can activate striatum even if the participant is not aware of them. For example, when a participant is performing a learned probabilistic sequence task, an unsignaled and unnoticed sudden shift in the underlying probabilities causes signal change in the accumbens.

In addition to responding after feedback, striatal fMRI activity shows anticipatory changes that also vary between subregions. While expecting monetary rewards, accumbens activity increases in proportion to the expected reward (at least in nonblock designs in which the size of the signaled reward changes from trial to trial). The accumbens does not respond in the same way to anticipated financial penalties, though the caudate responds to both. However, the idea that the accumbens fundamentally cares about rewards is not supported by other studies, showing accumbens activation during anticipation of an unpleasant electrical current applied to the skin.

Overall, these studies provide good evidence that the striatum is actively involved both in anticipating motivationally significant events and in responding to errors in these expectations, even if the difficulty of knowing the nature of the striatal fMRI signal limits our ability to understand exactly what changes in striatum during learning.

Electrophysiology of Single Striatal Neurons and Learning

An important achievement of striatal electrophysiology has been the ability to distinguish medium spiny projection neurons, which tend to fire infrequently and only at specific parts of a behavioral task, from the much smaller populations of interneurons that are more constantly active – notably the tonically active neurons (TANs) which are thought to correspond to the giant cholinergic interneurons. As in other brain areas, the larger population of projection cells is believed both to carry more-complex patterns of information and to have greater capacity to store new information. Among other functions, interneurons influence plasticity in projection cells, for example, by controlling their spike timing or spritzing them with neuromodulators.

TAN cells do not respond to neutral stimuli, but with conditioning they develop a brief pause or pause–rebound in firing as stimuli become motivationally relevant. Unlike midbrain dopamine cells, they do not discriminate between appetitive and aversive stimuli and do not show responses that are proportional to the expected magnitude of reward (Figure 4). Although their firing pattern is quite distinct from that of dopamine neurons, they require a certain level of dopamine stimulation to show their pause response. They tend to respond as a single, synchronous population, and although some discriminate between, for example, different locations of stimuli or different forthcoming movements, most do not.

Like the dopamine cells, the TANs therefore seem to be a more global, scalar signal rather than conveying complex patterns of information. Why the striatum needs multiple such scalar signals is not currently clear. The effects of acetylcholine (ACh) on striatal plasticity are complex, but blocking striatal muscarinic receptors enhances LTP at corticostriatal synapses. The ACh released by TANs is very rapidly degraded by striatal enzymes, so a synchronized pause in TAN firing may provide an even more temporally precise signal than the increase in dopamine cell activity. It has therefore been suggested that the TAN population instructs striatal neurons precisely 'when' to learn.

The striatal spiny neurons have a much more diverse set of responses – so much so that it seems at least a small proportion of these cells increase firing at each phase of any task studied. Many show activity related to specific ongoing or upcoming movements, though typically in a highly context-specific manner: similar movements outside the task are not associated with increased firing. The context specificity of spiny cell firing likely reflects the wide diversity of inputs that each cell is receiving. Consistent with the fMRI results above, the activity of spiny neurons is very often strongly modulated by reward expectation (though this is also true for neurons in many other brain regions). The combination of movement selectivity and reward expectation may indicate that striatal neurons are encoding action values that could potentially be useful for deciding between different actions. Although some reports have found reward expectancy to be more common in the accumbens, it is a frequent attribute of neurons in more-dorsal striatal regions too.

How might the formation of a procedural memory be expected to alter striatal projection cell firing? A hallmark of procedural memory is smoother, faster, less-error-prone performance, that is, a less diverse range of actions than is seen early in learning. Consistent with fewer different actions would be fewer active neurons and more consistency in neuronal firing, and this has indeed been observed – for example, in the striatum of rats running a maze task. Yet it is very difficult to know whether the observed neurons are the actual site of the key plastic events involved in procedural learning or whether their change in activity simply mirrors the change in behavior, as would be observed, for example, in a set of recordings from different muscles. Such considerations indicate a general conundrum in studying the changing neural representations underlying improved procedural learning. If the behavior changes, then it is hardly surprising that the representation of that behavior changes. Conversely, if neurons were to change firing without behavioral change, we would require some additional form of evidence to be confident that those neurons are actually involved in the behavior.

More specifically, what would an S-R association look like at the single-neuron level? It seems reasonable to assume that it should involve changes in striatal neuron responses to the stimulus that are also selective for the type of response made by the animal. Several studies have examined this possibility. Following the evidence that visual inputs to ventrocaudal striatum may be important for learning to select an object, recordings were made from this part of striatum while monkeys learned to make or withhold a hand movement depending on specific visual stimuli. The striatal neurons showed only visual responses that were generally similar to those in temporal cortex and were not modulated by the learned movement response or expected reward (unlike most parts of striatum). Hence, no evidence was found for neuronal responses related specifically to learning of S-R associations, despite the lesion data suggesting this area was involved.

A set of similar monkey electrophysiology studies examined other parts of striatum during the learning of arbitrary associations between visual stimuli and particular eye or hand movements, using enforced delay periods between stimulus and response to help distinguish between distinct aspects of neural coding. Over the course of the initial learning and beyond, many striatal spiny neurons do develop (or lose) firing increases at specific parts of each trial, though the overall diversity of these learned responses makes it hard to tell a specific story about how the association is encoded. Comparing the time course of activity changes in striatal regions with their afferents, it appears that neurons in putamen and in premotor cortex change at approximately the same rate, consistent with these structures' being part of the same cortex–basal ganglia loop circuit. One recent study compared caudate to dorsolateral prefrontal cortex (dlPFC) during reversal learning of S-R associations and found that the caudate neurons changed their direction selectivity in fewer trials. This finding was interpreted as evidence that the basal ganglia learn before the PFC, although given the wide diversity of inputs to caudate neurons, it might simply indicate the dlPFC is not the right place to look for the fastest changes during reversal learning (especially as orbitofrontal regions are more implicated in reversal).

Although striatal neurons change activity patterns during learning of this S-R association task, the task itself appears not to be a clear example of procedural learning. In contrast to the slow mode of learning proposed by Mishkin to involve the striatum, intact monkeys learn new associations in just a few trials, typically trying out each response in turn until they

find the one that works, then sticking with it. Lesions to hippocampus severely retard learning of these arbitrary visual-motor associations, and hippocampal neurons show some of the fastest learning-related responses. This is strong evidence that this form of learning is at least initially not procedural at all but may instead rely (for example) on the monkeys' remembering the specific episodes in which they tried out the other options. It would be very interesting if future studies examined striatal responses while discouraging a declarative strategy, such as by making the S-R mapping probabilistic or through inactivation of hippocampus. Although it might be hard to train animals, procedural learning could also be usefully assessed by testing interference with performance of a second task, as in some of the human fMRI studies previously mentioned.

Conclusions and Speculations

The striatum, especially dorsal/lateral subregions, is closely involved in certain forms of procedural learning, including the development of S-R habits. Yet the idea of associations is likely too limited to capture the role of striatal plasticity in learning. Borrowing concepts from engineering, recent theoretical work has begun to emphasize the function of cortical-basal ganglia loops as control systems that employ internal models of the organism's interactions with its environment. Internal models involve temporally structured expectations (explicit or implicit) about the consequences of actions, and striatal plasticity might best be described as allowing internal models to be refined when these expectations are not met. The diverse, context-dependent firing of striatal spiny cells during task performance may be a key aspect of such internal models, representing the progression through a learned, structured sequence of actions and expected events. Especially in sensorimotor striatum, such sequences may involve specific cues and movements, while more cognitive zones, such as caudate neurons, may encode more the abstract rules that govern overall task performance (i.e., learning sets). Compared with associations, such richer concepts of basal ganglia function may be essential for progress in the design and interpretation of critical experiments.

Some key unresolved questions include the following: Are the spines of striatal spiny neurons the permanent repository of learned action and expectation sequences, or are the basal ganglia involved primarily in influencing plasticity elsewhere (especially the cortex)? How do we make use of internal models to evaluate distinct courses of action without actually performing them? How can experimenters distinguish between psychomotor arousal and learning aspects of dopamine modulation of striatum, and over what timescales do these functions operate? These are questions that physiology and imaging investigations have just begun to tackle.

See also: Basal Ganglia: Functional Models of Normal and Disease States; Cognition: Basal Ganglia Role; Perceptual Learning: Neural Mechanisms; Procedural Learning in Animals; Procedural Learning in Humans; Striatum: Internal Physiology.

Further Reading

Berke JD and Hyman SE (2000) Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* 25: 515–532.

Berke JD (2003) Learning and memory mechanisms involved in compulsive drug use and relapse. In: Wang J (ed.) *Drugs of Abuse: Analysis of Neurological Effects*. Totowa, NJ: Humana Press.

Berridge KC (2006) The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology* 191(3): 391–431.

Daw ND, Niv Y, and Dayan P (2006) Actions, values, policies and the basal ganglia. In: Bezard E (ed.) *Recent Breakthroughs in Basal Ganglia Research*, pp. 111–130. New York: Nova Science.

Doya K (2000) Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology* 10: 732–739.

Graybiel AM (1998) The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory* 70: 119–136.

Morris G, Arkadir D, Nevet A, et al. (2004) Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron* 43: 133–143.

Packard MG, Hirsh R, and White NM (1989) Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *Journal of Neuroscience* 9: 1465–1472.

Redgrave P and Gurney K (2006) The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews Neuroscience* 7: 967–975.

Schultz W and Dickinson A (2000) Neuronal coding of prediction errors. *Annual Review of Neuroscience* 23: 473–500.

Wise SP, Murray EA, and Gerfen CR (1996) The frontal cortex-basal ganglia system in primates. *Critical Reviews in Neurobiology* 10: 317–356.

Yin HH and Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* 7: 464–476.

Programmed Cell Death

R Oppenheim and C Milligan, Wake Forest University School of Medicine, Winston-Salem, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During the embryonic, fetal, larval, and early postnatal stages of development, there is a massive loss of undifferentiated and differentiating cells in most tissues and organs, including the central and peripheral nervous systems. This cell loss is a normal part of development, and it occurs by a highly regulated process known as programmed cell death (PCD). Developmental PCD is defined as the spatially and temporally reproducible, tissue and species-specific loss of cells whose occurrence serves diverse functions and is a process required for normal development. Perturbation of this normal process can be maladaptive, resulting in pathology.

History

Although the occurrence of developmental PCD was first reported in the middle of the nineteenth century, it was not until the middle of the twentieth century that the occurrence and significance of PCD in the nervous system began to be appreciated by embryologists and developmental biologists. In a series of seminal papers by Viktor Hamburger and Rita Levi-Montelcini in the 1930s and 1940s, it was shown that the sensory and motor neurons in the spinal cord of the chick embryo are generated in excess during neurogenesis, followed by the PCD of approximately one-half of the original population. The period of cell loss was found to occur as sensory and motor neurons were establishing synaptic connections with their peripheral targets (e.g., muscle in the case of motor neurons). In a conceptual *tour de force*, Hamburger and Levi-Montalcini proposed that developing neurons compete for limiting amounts of target-derived survival-promoting signals (the winners survive and the losers undergo PCD) and in this way neurons are thought to optimize their innervation of targets by a process known as systems matching. It was this conceptual framework that led to the discovery of target-derived neurotrophic molecules and to the formulation of the neurotrophic theory that has been a major factor in fostering progress in this field for over 50 years. According to the neurotrophic theory, neurons that compete successfully for neurotrophic molecules avoid PCD, owing to the expression of survival promoting intracellular molecular genetic programs. The discovery of molecular genetic programs for both survival and death has revolutionized the study of PCD, resulting in the publication of thousands of articles each year since the early 1990s.

Evolution

The occurrence of massive cell death during development is, on the one hand, counterintuitive, whereas, on the other hand, the fact of its existence in many vertebrate and invertebrate species argues for its significance as an adaptively significant and fundamental part of embryogenesis. The death of occasional cells during development is to be expected in biological systems in which accidental deleterious events may be lethal to individual cells. However, because the stereotyped death of large numbers of developing cells in all members of a species cannot be explained in this way, it raises two basic questions regarding the evolution of PCD: (1) because PCD occurs by a metabolically active, genetically regulated process (see the sections titled 'Molecular regulation of cell death and survival by neurotrophic factors' and 'Intracellular regulation of cell death'), how and why did the molecular mechanisms involved arise during evolution; and (2) what are the adaptive reasons for massive developmental cell death in the nervous system (see the section titled 'The functions of PCD in the nervous system')? A plausible answer to the first question is that PCD arose as a defense mechanism for eliminating abnormal cells that threatened the survival of the organism. According to this argument, the death-promoting intracellular machinery arose in host cells to defend against the spread of viral infection and, in response, viruses evolved survival-promoting genetic mechanisms to block or counter the host cell death-promoting defenses. Because these death- and survival-promoting mechanisms have been found to mediate PCD in a diversity of organisms (see the section titled 'Intracellular regulation of cell death'), they are considered to represent an evolutionarily conserved core molecular genetic program that has been coopted by the nervous system for regulating cell numbers.

The Functions of PCD in the Nervous System

PCD in the nervous system involves both neurons and glia, and it occurs in the central and peripheral nervous system, the brain, and spinal cord and in virtually all neuronal and glial cell types. Accordingly, it is

Table 1 Some possible functions of developmental PCD in the nervous system

Differential removal of cells in males and females (sexual dimorphisms)
Deletion of some of the progeny of a specific sublineage that are not needed
Negative selection of cells of an inappropriate phenotype
Pattern formation and morphogenesis
Deletion of cells that act as transient targets or that provide transient guidance cues for axon projections
Removal of cells and tissues that serve a transient physiological or behavioral function
Systems matching by creating optimal quantitative innervation between interconnected groups of neurons and between neurons and their targets (e.g., muscles, sensory receptors)
Systems matching between neurons and their glial partners by regulated glial PCD (e.g., Schwann cells and axons)
Error correction by the removal of ectopically positioned neurons or of neurons with misguided axons or inappropriate synaptic connections
Removal of damaged or harmful cells
Regulation of the size of mitotically active progenitor populations
Production of excess neurons may serve as an ontogenetic buffer for accommodating mutations that require changes in neuronal numbers in order to be evolutionary adaptive (e.g., increases or decreases in limb size may require less or more motor neuron death for optimal innervation)
Regulated survival of subpopulations of adult-generated neurons as a means of experience-dependent plasticity

not surprising that the diversity of adaptive roles for PCD differs according to the animal species, cell type, nervous system region, and stage of development. A list of some of the most common reasons for PCD in the nervous system is provided in Table 1. Of these, the one that has received the most attention because of its central role in establishing optimal functional connectivity in neuronal circuits is systems matching (Figure 1). In birds and mammals, the evidence is compelling that for postmitotic neurons that are establishing interconnections with targets and afferents, systems matching, occurring in the framework of the neurotrophic hypothesis (see the section entitled 'Introduction'), is the primary reason for why a significant proportion of neurons (20–80%) undergo PCD. An analogous kind of systems matching adjusts myelinating glial cell numbers to the number of available axons.

PCD by Autonomous versus Conditional Specification

PCD occurring during systems matching represents a paradigm example of the conditional specification of cell fate during development. Developmental biologists have identified two kinds of pathways that cells use for specifying their differentiated fate or phenotype. One pathway, autonomous specification, involves the differential segregation of cytoplasmic signals into daughter cells following mitosis. In this way, cells become different from one another by the presence or absence of these cytoplasmic signals with little if any contribution from signals from neighboring cells. The other pathway, conditional specification, requires signals from other cells (cell–cell interactions) to progressively restrict differentiation and determine whether a cell lives or dies. These cell–cell interactions can be of four types: (1) juxtacrine (direct cell–cell or cell–matrix contact); (2) autocrine (a secreted signal acts back on the same cell type from which the signal arose); (3) paracrine (a secreted diffusible signal from one cell that acts locally on a different cell type); and (4) endocrine (signaling via the bloodstream). In the developing vertebrate nervous system, neuronal and glial survival is largely dependent on conditional specification involving paracrine interactions that use neurotrophic factors (NTFs). For neurons, the most commonly used NTFs are members of three major families: (1) neurotrophins, (2) the glial cell line-derived family ligands (GFLs), and (3) ciliary NTFs. Each of these contains several distinct members that act preferentially on specific types of neurons via membrane-bound receptors. By contrast, the survival of glial cells depends on different families of trophic factors such as the neuregulins and insulin-like growth factors. Neurons and glial cells use paracrine signaling to promote survival by ligand–receptor interactions. However, there are some situations in which paracrine interactions can signal death versus survival.

Molecular Regulation of Cell Death and Survival by NTFs

The dependence of neurons on NTFs for survival intuitively leads to the conclusion that the inactivation of NTF receptors, and the signal transduction associated with them, leads to the activation of cell death events (Figure 2). This has been best studied with neurotrophins and their receptors, notably the

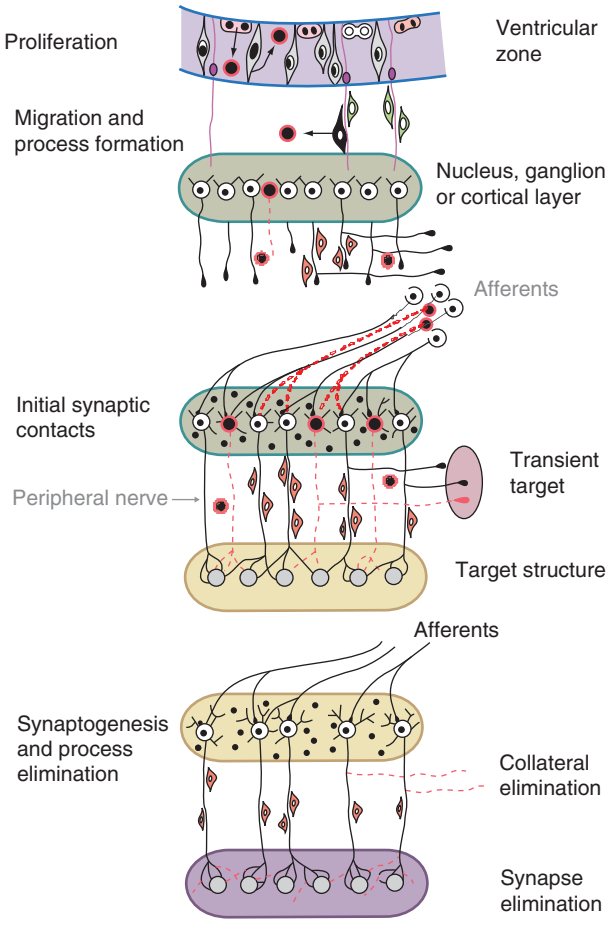

Figure 1 Schematic illustration of some key steps in neuronal development. Neurons undergoing PCD (●) are observed during neurogenesis in the ventricular zone, during migration and while establishing synaptic contacts. Schwann cells in developing nerves also undergo PCD. represents peripheral glial (Schwann) cells; ● represents CNS glia (astrocytes and oligodendrocytes); ⊙ represents surviving, differentiating neurons (motor neurons) whose targets are skeletal muscle.

Trk receptors and p75 (Table 2). Activation of the trk receptors results in the activation of the phosphatidylinositol-3 kinase (PI3-K) and mitogen-activated protein kinase (MAPK) pathways. Akt, a substrate of PI3-K, has been reported to phosphorylate Bad, promoting its association with 14-3-3 and preventing the inactivation of Bcl-2 and Bcl-x. Activation of PI3-K/Akt, extracellular signal-regulated kinase-1/2 (ERK1/2), protein kinase (PK)C, or PKA promotes the serine phosphorylation of Bad and Bcl-2, and Gsk-3B phosphorylates and inactivates Bax – events that are associated with cell survival. The survival-promoting activity of phosphorylated Bcl-2, however, is controversial, being associated with motor neuron survival-induced trophic support, whereas in neurons treated with microtubule destabilizing agents it appears to promote death. The c-jun N-terminal kinase (JNK) pathway is reported to exhibit increased activity in neurons triggered to die. The activation of one of its substrates, c-jun, is also thought to play a role in mediating neuronal death. For example, the JNK activation of the BH3 proteins Bim and DP5 has been associated with neuronal death, and a JNK–p53–Bax pathway appears to be critical for death in specific cell types. Nonetheless, the JNK pathway has also been reported to be a critical mediator of survival-promoting events such as neurite outgrowth and may serve as a double-edged sword in neurons through changes in intracellular localization or specific activation or inactivation of individual isoforms.

The role of the p75 nerve growth factor (NGF) receptor in promoting neuronal survival or death is complicated. For example, complexes of p75 and the trk receptors bind neurotrophins and promote survival, while complexes of p75 and sortilin appear to promote binding of the proneurotrophins resulting in the death of neurons. Activation of p75 in the absence of Trk activation has most notably been associated with the death of neurons, but this appears to be developmentally regulated. The precise mechanism by which p75 promotes death is unclear. However, p75 is a member of the tumor necrosis factor receptor family, and another member of this family better known for its death promoting activity is Fas/CD95. Engagement of Fas by Fas ligand leads to the activation of caspase 8. This event is dependent on the formation of a death-induced signaling complex (DISC). Homotrimerization is the first step in this process. Engagement of Fas by its ligand can only occur when it is homotrimerized. This family of receptors is characterized by the presence of a death domain (DD) on the cytoplasmic region of the receptors. The Fas-associated death domain-containing protein (FADD) can then bind to Fas. In addition to the DD, FADD also contains a death effector domain (DED). Procaspase 8 binds the DED. The autolytic nature of caspases leads to active capsase 8 that can then go on to directly activate caspase 3 (in type 1 cells, e.g., thymocytes) or cleave Bid leading to changes at the mitochondria (in type 2 cells, e.g., hepatocytes and neurons) and cell death.

Intracellular Regulation of Cell Death

Many of the intracellular mechanisms mediating neuronal cell death were investigated following initial work that suggested that new gene expression was required for the process. Horvitz and coworkers provided some of the first evidence that there was indeed a genetic component of PCD from their work in the 1980s with the free-living nematode *Caenorhabditis elegans*, although it was two decades later

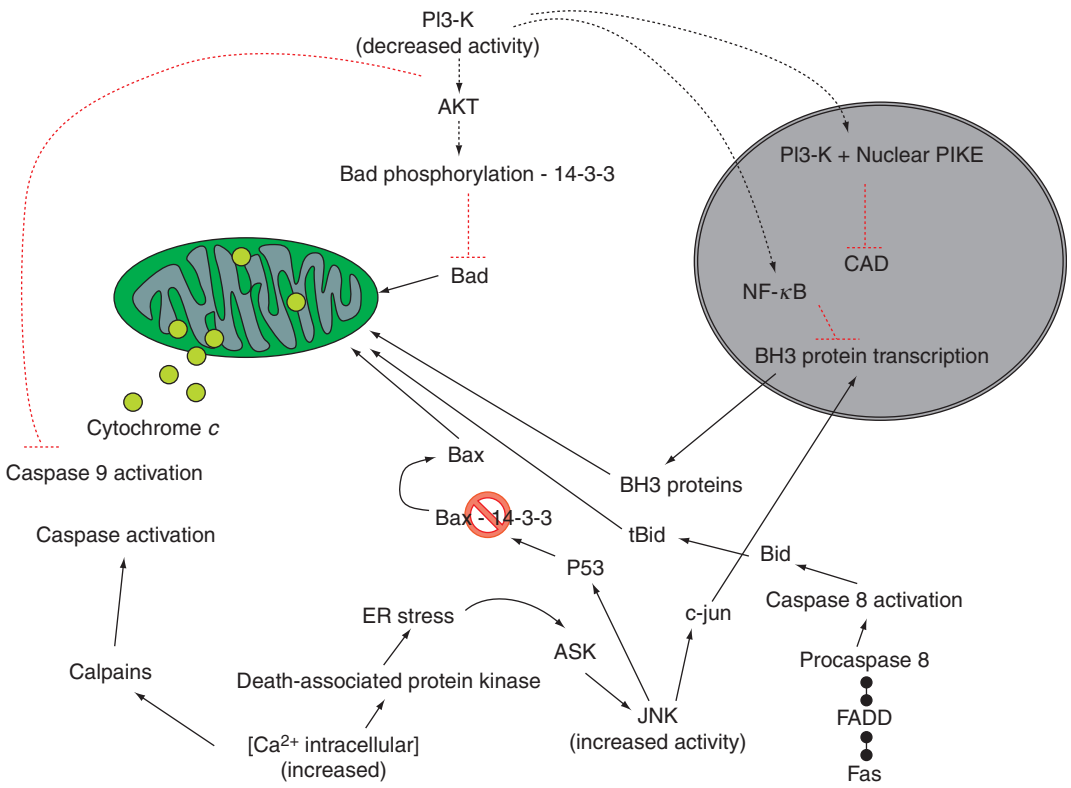

Figure 2 Signals to die. Changes in signal transduction pathways, activation of receptors containing a DD, or changes in intracellular calcium concentrations have been identified as events that can lead to the activation of cell death-specific events. Loss of trophic support results in decreased activation of the phosphatidylinositol-3 kinase (PI3-K), thereby releasing inhibitory factors linked to this pathway. Alternatively, increases in intracellular calcium concentrations, activation of the FAS receptor, and increased c-jun terminal kinase activation are associated with the mitochondrial changes and caspase activation that occurs in cell death. ASK, apoptosis signal-regulating kinase; ER, endoplasmic reticulum; FADD, Fas-associated death domain-containing protein; JNK, c-jun N-terminal kinase; NF-κB, nuclear factor κB; PIKE, phosphatidylinositol-3 kinase enhancer.

that the significance of this work was fully appreciated (Figure 3). As these genes were identified, the sequence of events leading to the death of cells was pieced together.

The presence of Bcl-2 family proteins appears critical for mediating the survival or death of nervous system cells. During development, Bcl-2 expression is correlated with neuronal survival. With continued maturation and development, this expression declines while Bcl-x expression increases. The specific mechanisms responsible for this change in expression are currently not known. On the other hand, the expression of pro-death Bcl-2 family proteins such as Bid or Bax appears to be consistent throughout development, whereas the intracellular localization of these proteins appears to change in cells undergoing death. In healthy cells, Bax is localized more in the cytoplasm, whereas in dying cells the majority of Bax localizes to organelle membranes, including the mitochondria. The localization of Bax to the mitochondria corresponds to the release of cytochrome *c* into the cytoplasm. Once in the cytoplasm, cytochrome *c* binds Apaf-1, causing a conformational change in Apaf-1 to reveal a caspase recruitment (CARD) domain. In the presence of ATP, a heptamer of the cytochrome *c*–Apaf-1 complex is formed. Procaspase 9 has a high affinity for the CARD domain and localizes to the heptamer. This complex is referred to as the apoptosome. With the increased local concentration of procaspase 9, the autolytic property of caspases leads to the generation of active caspase 9. As an initiator caspase, active caspase 9 can cleave procaspase 3, resulting in the active form of this protease (Figure 4). The activation, inactivation, or destruction of specific substrates significantly contributes to the rapid degeneration of the cell. Caspase activity also leads to the activation of endonucleases that may play a role in the changes in nuclear morphologies that are observed in may cell types.

The removal of the dying cell is most likely just as critical an event in the cell death process as any of those already discussed. The event is accomplished by adjacent cells that can be nonprofessional phagocytes (e.g., Schwann cells) but more often involves

Table 2 Specific molecules associated with cell death in neurons[a]

Receptors	Fas
	P75
	Trk receptors
Signal transduction pathways	PI3K–Akt
	ERK1,2
	ERK 5
	JNK
	PLC
	PKC
	PKA
Bcl-2 proteins	Bcl-2
	Bcl-x
	Bid
	Bik
	Bag
	Bax
	Bad
	Dp5
	Bim
	Puma
Mitochondrial proteins	Cytochrome *c*
	SMAC/Diablo
	XIAP
	AIF
	Endonuclease G Omi/HtrA2
Caspases	
Initiator caspases	Caspase 8
	Caspase 9
Effector caspases	Caspase 2
	Caspase 3
Inflammatory caspases	Caspase 1
Other proteins	Apaf-1
	Amyloid precursor protein/ β-amyloid
	PARP
	p53
	p35
	Ubiquitin
	E2F
	Cyclin D1
	Cyclin E
	Cyclin B
	Cdk4/6
	Cdk2
	Cdk1
	Rb
	p130
	HIF1-α
	CAD
	iCAD

[a]AIF, apoptosis-inducing factor; CAD, caspase-activated DNase; HIF1-α, hypoxia-inducing factor; iCAD, inhibitor of CAD; JNK, c-jun N-terminal kinase; PARP, poly(ADP-ribose) polymerase; PKA, protein kinase A; PKC, protein kinase C; PLC, phospholipase C; XIAP, x-linked inhibitor of apoptosis protein.

dedicated phagocytes such as tissue macrophage (microglia) or circulating monocytes. During central nervous system (CNS) development, resident microglia are not present in large numbers among dying cell populations. Intuitively, some signal must be sent by the dying cells to recruit the phagocytes into the area. One such chemotactic factor has been identified in nonneuronal cells, and interestingly, this factor is the phospholipid lysophosphatidylcholine. Caspase 3 activation appears to be necessary for the release of this factor. Once the phagocytic cells are in the region, they must be able to distinguish dying from healthy cells. Changes on the dying cell's surface include revealing of thrombospondin 1 binding sites, exposure of phosphatidyl serine, ATP-binding cassette (ABC-1) molecules, and carbohydrate changes. The phagocyte, in turn, expresses the phosphatidylserine receptor, as well as thrombospondin receptors that recognize the thrombospondin bound to the binding site on the dying cell, lectins that bind to the carbohydrate changes, and ABC-1 molecules that bind to like molecules. Phagocytosis occurs only when multiple changes are recognized, and this phagocytosis is limited and does not result in macrophage secretion of cytokines that would normally induce an inflammatory or immune response if foreign antigens were phagocytosed.

Different Types of Neuronal Death

Many individuals studying cell death often rely on the critical papers of Currie, Kerr, and Wyllie to define and characterize the different types of cell death. These investigators described in detail two distinct morphological changes associated with cell death. An active process, apoptosis, was characterized by specific morphological changes that included condensation of nuclear chromatin, shrinking of the cytoplasm, and a breaking up of the cell into membrane-bound particles that were phagocytosed. Necrosis, on the other hand, was characterized by a swelling of the cytoplasm with eventual bursting of the cell (Figure 5). In this case, intracellular components are spilled into the extracellular space where they could initiate inflammatory and immune responses. However, electron microscopy studies of the developing nervous system indicated that neuronal death could not be so easily defined by only these two modes of death. During the naturally occurring death of some neurons, initial changes in dying cells are observed in the cytoplasm where there is an increase in the diameter of the cisternae of the rough endoplasmic reticulum (RER). Mitochondrial swelling was also observed, although it was not clear if these changes occur in the same cell. Nonetheless, during normal development initial changes are observed in the cytoplasm with little nuclear alterations in dying cells. The cell then appears to round up and break into pieces that are phagocytosed. Although cytoplasmic cell death appears to be more prominent

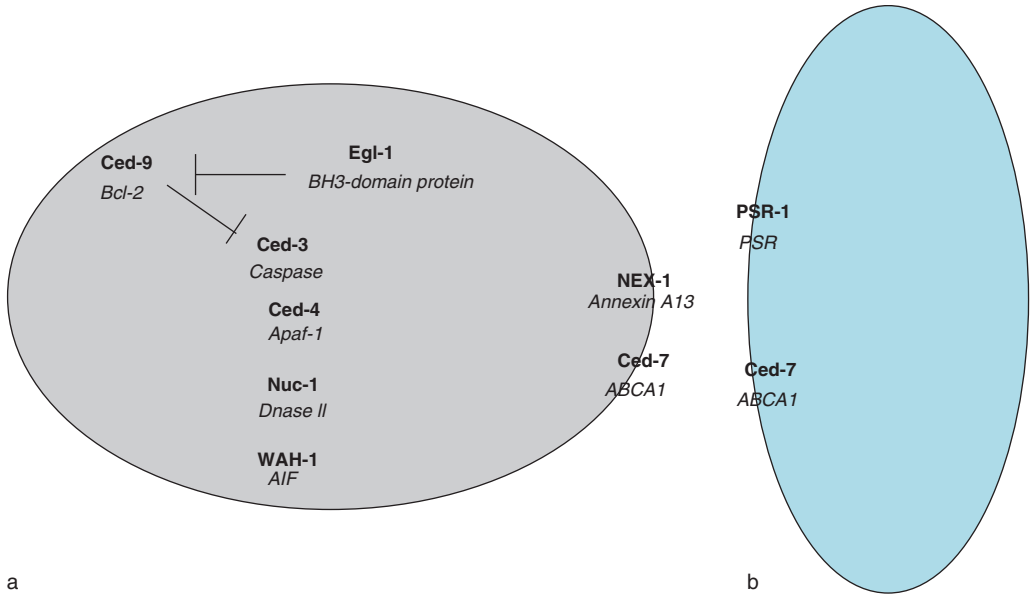

Figure 3 Genes for the key regulators of cell death: (a) dying cell; (b) engulfing cell. Many of the key regulators of cell death are evolutionarily conserved. Many of the genes for these proteins were initially identified through genetic mutations in the free living nematode (bold type), with homologs later being identified in *Drosophila* and mammals (italics).

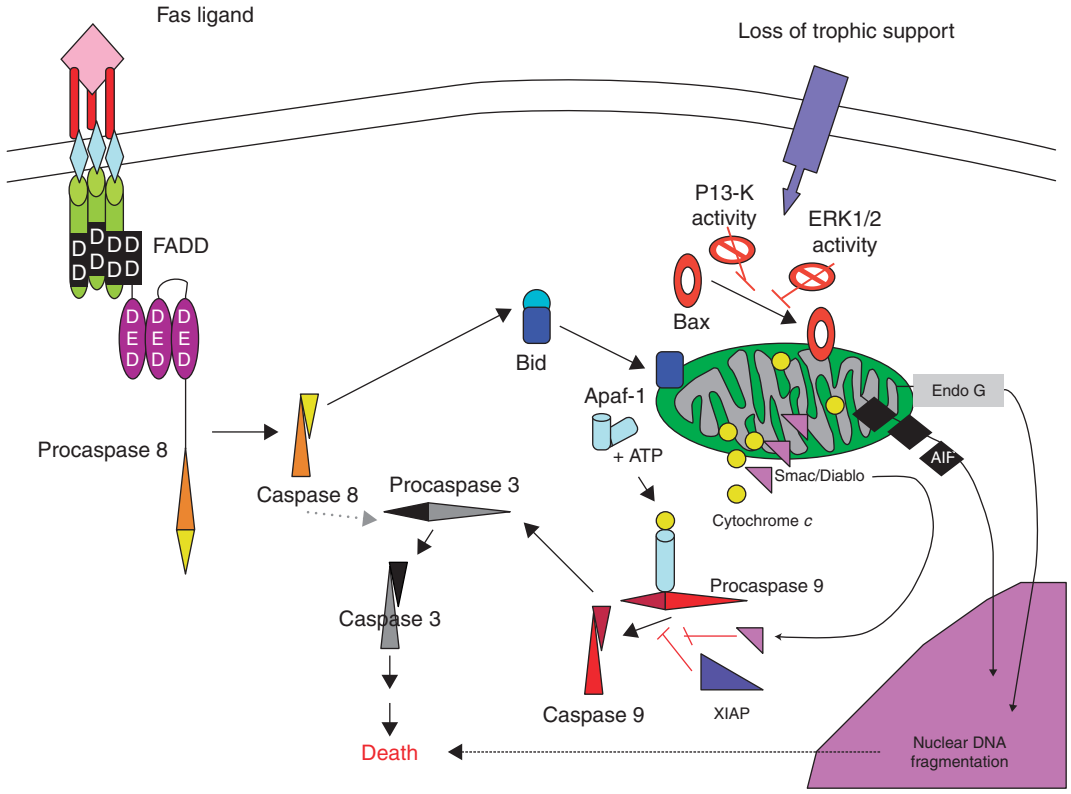

Figure 4 Many of the critical components of neuronal death and their apparent sequence of activation during development. The mitochondria release many regulators that can lead to death with or without caspase activation. Neurons are considered type 2 cells where FAS engagement results in minimal caspase 8 activation, resulting in the cleavage of Bid and its subsequent movement to the mitochondria. The localization of Bax to the mitochondria appears to be a critical event in the death of neurons. AIF, apoptosis-inducing factor; ATP, adenosine triphosphate; DD, death domain; DED, death effector domain; ERK, extracellular signal-regulated kinase; FADD, Fas-associated death domain-containing protein; PI3-K, phosphatidylinositol-3 kinase; XIAP, x-linked inhibitor of apoptosis protein.

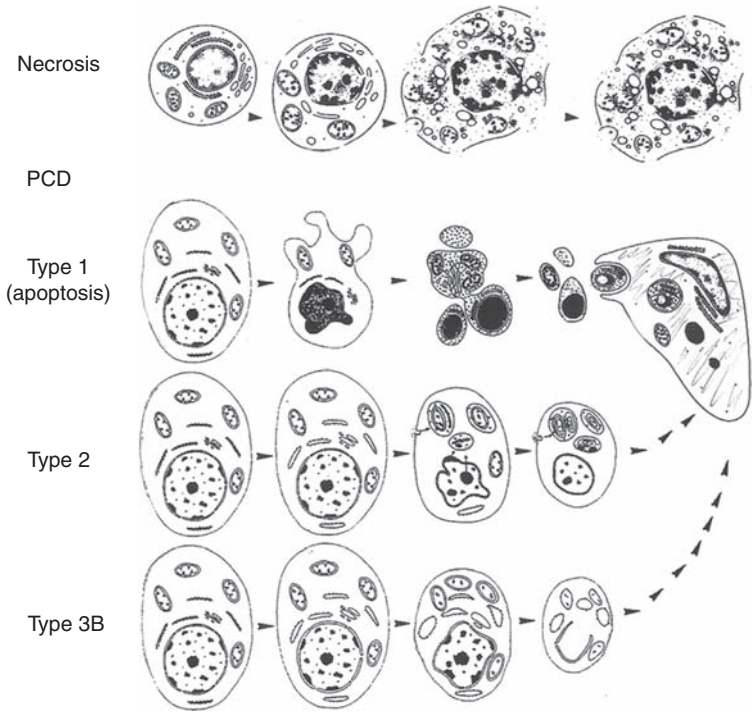

Figure 5 Necrotic death and the three most common types of programmed cell death (PCD) observed in the nervous system. Type 1 meets the criteria for apoptosis. In PCD, phagocytes remove the resulting corpses, whereas necrosis in many cases results in inflammation.

during development, nuclear or apoptotic death also occurs. A third type of death is autophagy, characterized by the formation of numerous autophagic vacuoles. Nuclear changes may also occur in this type of death.

It is important to note that all three types of neuronal death appear to reflect a metabolically active process. Accordingly, it appears that a homogeneous mode of suicide does not occur but, rather, that processes leading to death appear to be context and developmentally dependent. One example is a neuronal population undergoing apoptosis when nuclear condensation is a prominent feature. On the other hand, in animals in which caspases have been inhibited or deleted, the same neuronal population undergoes a delayed death with major changes occurring in the cytoplasm.

Pathological Neuronal Death

Many of the mechanisms that have been identified as playing a role in neuronal death during development are reported to be reactivated in the mature pathological nervous system. Alterations in the expression of Bcl-2 proteins are observed in animal models and in postmortem human tissue in Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's chorea (HC), amyotrophic lateral sclerosis (ALS), and stroke. Caspase activation has also been reported in these and other disorders. Interestingly, although reports of apoptotic cells are present in these conditions, the majority of degenerating cells exhibit morphological changes reminiscent of cytoplasmic or autophagic death. This is most notable in the mutant SOD1 mouse model of ALS. During development, dying motor neurons exhibit many features of apoptosis, including possible Fas activation, Bax translocation, mitochondrial dysfunction, and caspase activation and condensation of nuclear chromatin. By contrast, in the adult spinal cord of SOD1 mice, degenerating motor neurons exhibit cytoplasmic vacuolization, mitochondrial dilation, and protein aggregation. Nonetheless, some of the cell death-associated events that occur during development are also observed. It is becoming increasingly recognized that neuronal dysfunction is most likely the event responsible for clinical symptoms in ALS and other neurodegenerative diseases. In experiments in which motor neuron death is inhibited in the SOD1 mouse, disease progression and survival of the animal are only very modestly affected; muscle dennervation still occurs, and the animals die prematurely. Results such as this call into question the practical application of inhibiting cell death as a therapeutic approach for these disorders.

See also: Apoptosis in Nervous System Injury; Apoptosis in Neurodegenerative Disease; Autonomic and Enteric Nervous System: Apoptosis and Trophic Support During Development; Autophagy and Neuronal Death.

Further Reading

Akhtar RS, Ness JM, and Roth KA (2004) Bcl-2 family regulation of neuronal development and neurodegeneration. *Biochimica et Biophysica Acta* 1644(2-3): 189–203.

Bronfman FC and Fainzilber M (2004) Multi-tasking by the p75 neurotrophin receptor: Sortilin things out? *EMBO Reports* 5(9): 867–871.

Brunet A, Datta SR, and Greenberg ME (2001) Transcription-dependent and-independent control of neuronal survival by the PI3K-Akt signaling pathway. *Current Opinion in Neurobiology* 11(3): 297–305.

Buss RR, Sun W, and Oppenheim RW (2006) Adaptive roles of programmed cell death during nervous system development. *Annual Review of Neuroscience* 29: 1–40.

Clarke PGH and Clarke S (1996) Nineteenth century research on naturally occurring cell death and related phenomena. *Anatomy and Embryology* 193: 81–99.

Fischer U, Janicke RU, and Schulze-Osthoff K (2003) Many cuts to ruin: A comprehensive update of caspase substrates. *Cell Death and Differentiation* 10(1): 76–100.

Greene LA, Biswas SC, and Liu DX (2004) Cell cycle molecules and vertebrate neuron death: E2F at the hub. *Cell Death and Differentiation* 11(1): 49–60.

Horvitz HR (2003) Worms, life, and death (Nobel lecture). *ChemBioChem* 4(8): 697–711.

Krantic S, Mechawar N, Reix S, and Quirion R (2005) Molecular basis of programmed cell death involved in neurodegeneration. *Trends in Neuroscience* 28(12): 670–676.

Lettre G and Hengartner MO (2006) Developmental apoptosis in *C. elegans*: A complex CEDnario. *Nature Reviews Molecular and Cell Biology* 7(2): 97–108.

Lindholm D and Arumae U (2004) Cell differentiation: Reciprocal regulation of Apaf-1 and the inhibitor of apoptosis proteins. *Journal of Cell Biology* 167(2): 193–195.

Morrison RS, Kinoshita Y, Johnson MD, Ghatan S, Ho JT, and Garden G (2002) Neuronal survival and cell death signaling pathways. *Advances in Experimental Medicine and Biology* 513: 41–86.

Oppenheim RW and Johnson JE (2002) Programmed cell death and neurotrophic factors. In: Zigmond MJ, et al. (eds.) *Fundamental Neuroscience*, pp. 499–532. New York: Academic Press.

Polster BM and Fiskum G (2004) Mitochondrial mechanisms of neural cell apoptosis. *Journal of Neurochemistry* 90(6): 1281–1289.

Ryan CA and Salvesen GS (2003) Caspases and neuronal development. *Biological Chemistry* 384(6): 855–861.

Sathasivam S and Shaw PJ (2005) Apoptosis in amyotrophic lateral sclerosis – What is the evidence? *Lancet Neurology* 4(8): 500–509.

Savill J and Fadok V (2000) Corpse clearance defines the meaning of cell death. *Nature* 407(6805): 784–788.

Stefanis L (2005) Caspase-dependent and- independent neuronal death: Two distinct pathways to neuronal injury. *Neuroscientist* 11(1): 50–62.

Wyllie AH, Kerr JF, and Currie AR (1980) Cell death. The significance of apoptosis. *International Review of Cytology 68: 251–306.*

Yuan J, Lipinski M, and Degterev A (2003) Diversity in the mechanisms of neuronal cell death. *Neuron* 40(2): 401–413.

Prolactin and its Neuroendocrine Control

N Ben-Jonathan and C R LaPensee, University of Cincinnati, Cincinnati, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Prolactin (PRL) is produced by the pituitary lactotrophs, which constitute 20–40% of all pituitary cells. Lactotrophs exhibit many distinguished properties, including late ontogenic development, the capacity for proliferation during adulthood, expression of spontaneous electrical activity, the presence of receptors for neurotransmitters, responsiveness to many secretagogues, substantial morphological and functional heterogeneity, the existence of dual hormone secreting cells, and a high intrinsic synthesis and secretory activity. Lactotrophs are under tonic inhibition by hypothalamic dopamine, and this inhibitory activity is counteracted by multiple PRL releasing/regulating factors (PRFs). This article summarizes salient features of PRL and its receptors, highlights comparative aspects of PRL physiology, and evaluates hypothalamic, peripheral, and paracrine/autocrine regulation of PRL production.

Prolactin and the Prolactin Receptor

The Prolactin Gene and Protein

The PRL gene, located as a single copy on human chromosome 6, is evolutionally and functionally related to both growth hormone (GH) and placental lactogen (PL). The rat PRL gene contains five exons and is 10 kb long. It is regulated by a proximal promoter composed of a 2.5 kb sequence at the 5′ flanking region. Several *cis*-acting elements within the promoter bind Pit-1, a pituitary-specific POU-homeodomain transcription factor which confers cell-specific PRL expression. The human PRL gene is composed of six exons and is 18 kb long. An extra noncoding exon 1a is located 5.8 kb upstream of the pituitary start site (Figure 1). In extrapituitary sites such as the decidua, myometrium, and lymphoid cells, exon 1a is spliced to exon 1b, generating an mRNA transcript that is 150 bp longer in the 5′ untranslated region than the pituitary transcript. An alternative, superdistal promoter upstream of exon 1a regulates PRL expression at extrapituitary sites. As judged by biochemical, biological, and immunological criteria, the pituitary and extrapituitary PRL proteins are identical.

The PRL protein is made of a single chain of 199 residues stabilized by three disulfide bridges. PRL adopts a three-dimensional configuration as four antiparallel α-helices organized in an 'up-up-down-down' manner (Figure 1). The four-helix bundle motif is shared with GH, hematopoietic factors, and several cytokines. Alternative splicing, posttranslational modifications, and proteolysis generate PRL variants that differ in size or in functional groups. Larger forms result from aggregation, whereas smaller forms are generated by proteolysis. Some variants retain PRL-like activities, whereas others have unique properties or no known functions.

Structure and Distribution of the Prolactin Receptor

The gene encoding the human PRL receptor (PRL-R) is located on chromosome 5 and consists of 10 exons encompassing 100 kb. PRL-R belongs to the non-tyrosine kinase cytokine/GH/PRL receptor superfamily. These receptors are characterized by a single-pass transmembrane stretch that divides the receptor into an extracellular ligand-binding domain and an intracellular domain. The extracellular domain contains two fibronectin-like regions named S1 and S2. The S1 region has two glycosylation sites and two pairs of disulfide bridges and contains most of the PRL binding sites. The S2 region has a conserved WS motif (Trp-Ser-x-Trp-Ser) that is involved in receptor folding and dimerization (Figure 2). Two binding sites within PRL – site 1 composed of helices 1 and 4 and site 2 composed of helices 1 and 3 – are required for induction of receptor homodimerization and formation of an active trimeric complex. The cytoplasmic domain of the receptor has a proline-rich motif ('box 1') adjacent to the plasma membrane that couples to intracellular signaling molecules.

The PRL-R is expressed in almost all tissues, with the highest expression in the liver, mammary gland, adrenal, and choroid plexus. Human GH and placental lactogens also bind to the PRL-R and mimic some actions of PRL. In humans, alternative splicing generates five PRL-R isoforms that differ in the length and composition of the cytoplasmic domain. Rats express a long, an intermediate, and a short isoform (Figure 2), whereas mice express one long and three short isoforms. A soluble PRL-binding protein which contains only the extracellular domain is generated by proteolysis. The long isoform is responsible for most of the signal transduction properties of the receptor. Short isoforms are expressed at a variable ratio in different tissues, but their precise action is unclear.

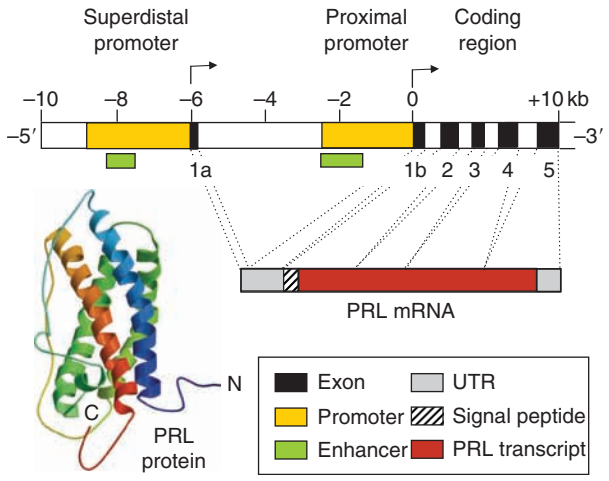

Figure 1 Diagram of the human PRL gene, PRL mRNA transcript, and mature PRL protein. Arrows designate transcriptional start sites for the proximal pituitary promoter and the superdistal extrapituitary promoter. Transcription from either promoter produces mRNAs with identical protein coding sequences but differing in the 5′ untranslated sequences (5′ UTR). The mature PRL protein shows the four-helix bundle configuration.

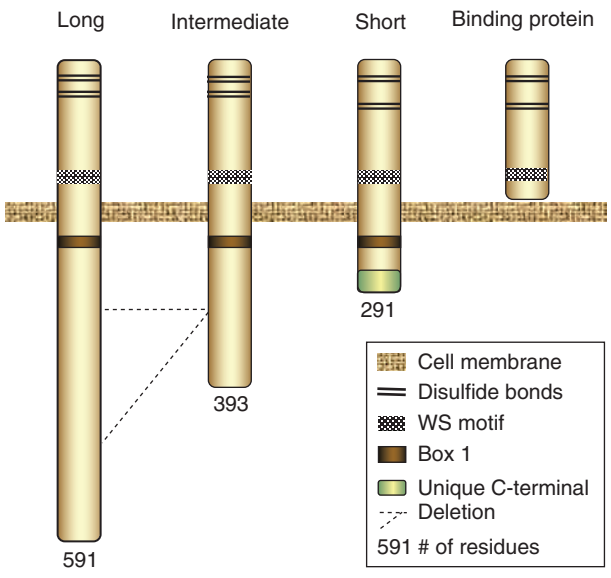

Figure 2 Schematic presentation of the different isoforms of the rat PRL receptor. All isoforms have an identical extracellular binding domain.

Activation of Multiple Signaling Pathways by Prolactin

PRL-R is devoid of intrinsic tyrosine kinase activity and utilizes the Jak2/Stat pathway as its main signaling cascade (Figure 3). Receptor-associated Jak2 is rapidly phosphorylated upon PRL binding and induces phosphorylation of the receptor, associated

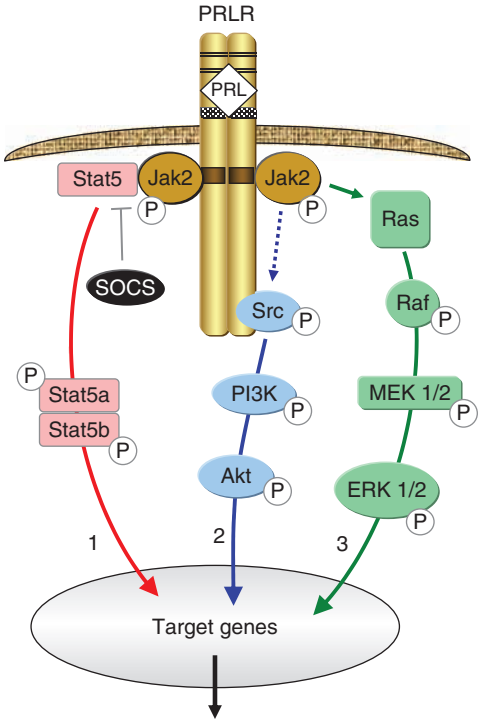

Figure 3 Signaling pathways that mediate activation of the PRL receptor. Jak2 binds to box 1 within the intracellular domain of the receptor. The Jak2/Stat5 pathway (1) is considered the main pathway, but phosphatidylinositol-3-kinase (2) and mitogen-activated protein kinase (3) pathways are also stimulated by ligand binding to the PRLR in a cell-specific manner. Phosphorylation is designated by circled P's.

kinases, and Stat proteins. Of the seven known Stat proteins, Stats 5a and 5b serve as the primary mediators of PRL action. Upon phosphorylation, Stat proteins dimerize and translocate to the nucleus, where they bind to consensus elements on target genes. Milk proteins are the best-characterized PRL-R-regulated genes involving the Jak/Stat pathway.

The Ras/Raf/MAPK pathway is also activated by PRL (Figure 3). Phosphorylated Jak2 recruits adaptor proteins such as Ras and Raf which then activate the MAP kinase pathway. Activation of PRL-R also facilitates docking of Src family kinases and activation of the phosphatidylinositol 3 kinase/Akt pathway. Both the strength and the duration of the PRL-induced signaling are regulated by suppressors of cytokine signaling (SOCS) proteins, whose expression is rapidly induced upon receptor activation. The SOCS proteins interact with the PRL-R or with Jak2 and inhibit further signaling. In summary, signal transduction activated by the PRL-R does not represent a linear progression but involves interacting pathways which differ in predominance among the various PRL-R-expressing cells.

Comparative Physiology of Prolactin

Rodents

PRL is a modulating/adaptor hormone which has assumed multiple functions throughout evolution. Consequently, its overall regulation and the dynamics of its secretion vary among species. In rodents, PRL is essential for the maintenance of the corpus luteum (luteotropic activity), and its release is altered during the reproductive cycle and pregnancy by an interplay between dopamine-releasing hormones, ovarian steroids, and placental lactogens. Concomitant with the preovulatory surge of luteinizing hormone (LH) during late proestrous in rats, serum PRL levels also surge by an estrogen-dependent mechanism. The preovulatory PRL surge is nonsymmetrical but has a prolonged plateau phase which subsides toward estrous. If mating occurs, the corpus luteum is rescued from disintegration by two daily PRL surges whose timing is controlled by the superchiasmatic nucleus. In the case of a nonfertile mating (pseudopregnancy), the PRL surges persist for 12 days, followed by resumption of cyclicity. In the case of pregnancy, the daily PRL surges terminate on day 10 in response to the rising levels of placental lactogens. During mid- to late pregnancy, serum PRL is low, until a large surge occurs just before parturition.

Suckling during lactation initiates a classical neuroendocrine reflex and is the most potent physiological stimulus for PRL release. Within minutes of the initiation of suckling, and in direct proportion to the intensity of the stimulus, PRL release increases manifold and is sustained as long as nursing endures. Cessation of suckling rapidly terminates PRL release, with the decline in its serum levels determined by the metabolic clearance rate. The control of the suckling-induced PRL release is complex, involving a coordinated action of several PRL-releasing factors from both the hypothalamus and the posterior pituitary, superimposed on a markedly reduced dopaminergic activity.

Humans

The physiology of PRL in humans differs in many respects from that in rodents. During development, fetal plasma PRL levels are low between 10 and 20 weeks and increase progressively to peak levels before birth. Another increase occurs immediately after birth and persists for several days. Fetal PRL is regulated independent of the mother and may respond to dopaminergic inhibition. PRL in the maternal circulation also increases gradually toward term, reaching seven- to tenfold higher levels than in nonpregnant women. Elevated PRL is essential for preparation of the breasts for lactation, which is prevented during pregnancy by high progesterone levels. Pregnancy also causes pituitary enlargement due to lactotroph hypertrophy and hyperplasia. Labor activates cervical neural signals that trigger a PRL surge which does not occur during cesarean delivery.

During pregnancy, large amounts of PRL are independently produced by the decidua. Neither dopamine nor estrogen affects decidual PRL release. Decidual PRL is secreted into the amniotic fluid, where its concentrations peak between 18 and 23 weeks of gestation, attaining 50- to 100-fold higher levels than in either the fetal or the maternal circulation. Basal PRL levels remain elevated during the first 2 or 3 weeks postpartum in women who do not breastfeed, and then they decline. Similar to rodents, suckling is the most potent stimulus for PRL release. The magnitude of the suckling-induced PRL increase is robust during early lactation but wanes thereafter. Tactile stimuli of the breast can increase serum PRL in nonlactating women. Coitus for both men and women also causes an increase in serum PRL levels.

During childhood, there are no sex differences in serum PRL levels, with PRL becoming twice as high in women than men after puberty. In contrast to rats, estrogens have little effect on acute PRL release in humans. Indeed, circulating PRL levels are largely unchanged throughout the menstrual cycle and do not increase during the preovulatory LH surge. Although estradiol increases PRL release from dispersed adult human pituitary cells, it does not reverse the dopamine-mediated inhibition, indicating that it may be effective when dopaminergic inhibition is absent or reduced. Stress from anesthesia, surgery, strenuous exercise, and insulin-induced hypoglycemia also stimulates PRL release in both men and women. Small increases in PRL also occur in response to high-protein meals. When sampled at high frequency, serum PRL levels display an episodic release pattern whose amplitude increases after the onset of sleep. The nocturnal increase in PRL release is seen in both sexes and is sleep induced rather than driven by an intrinsic diurnal rhythm.

Hypothalamic Dopaminergic Systems

Anatomy and Relative Contribution to Prolactin Release

Studies using rats identify two main hypothalamic dopaminergic systems that regulate PRL: the tuberoinfundibular (TIDA) and the tuberohypophysial (THDA). The perikarya of TIDA neurons are located in the arcuate nucleus and send short axons that terminate in the external zone of the median eminence. These do not form synapses and function as

true neurosecretory neurons whose product is released into perivascular spaces and carried by portal blood to the anterior pituitary. THDA neurons from the rostral arcuate nucleus send axons that course through the pituitary stalk and terminate in the neural (NL) and intermediate (IL) lobes of the pituitary. They are joined by the periventricular hypothalamic neurons with perikarya in the periventricular nucleus and terminals in the IL lobe. The NL terminals are neurosecretory, whereas those in the IL synapse on melanotrophs. Short portal vessels connect the NL and anterior lobe (AL) and provide a vascular route for dopamine delivery to the lactotrophs. Of the three systems, the TIDA plays the predominant inhibitory role in PRL release. However, a reciprocal relationship between TIDA activity and PRL release is not always apparent since it is often masked by other PRL regulatory factors.

Regulation of Hypothalamic Dopamine by Prolactin and Estrogen

In the absence of target gland hormones to provide feedback control over PRL release, PRL is used as the primary regulator of the dopaminergic systems via a short loop negative feedback mechanism (Figure 4). TIDA neurons are activated by both acute and chronic elevations in PRL, with few exceptions. These include pregnancy, lactation, and prolactinomas, when the dopaminergic neurons become refractory to the elevated PRL, enabling the maintenance of physiological or pathological hyperprolactinemia. The rat brain expresses both the long and short PRL-R isoforms, and their expression is altered under various endocrine conditions. After the choroid plexus, the hypothalamus contains the highest density of the PRL-R within the brain. The PRL-R often co-localizes with neurons expressing tyrosine hydroxylase (TH), the rate-limiting enzyme in dopamine synthesis. Similar to peripheral tissues, binding of PRL to hypothalamic neurons activates the Jak2/STAT signaling pathway. PRL reaches the arcuate nuclei via the cerebrospinal fluid following receptor-mediated uptake at the choroid plexus.

Estrogens affect PRL homeostasis at several sites, including the hypothalamus, posterior pituitary, and anterior pituitary (Figure 4). The action of estrogen is mediated by two receptors, ERα and ERβ, that are the translation products of different genes. Both have a similar modular structure of five functional domains, with high homology in the DNA- and ligand-binding domains but distinct regulatory domains. They bind estradiol with similar affinity, and after homo- or heterodimerization they activate an estrogen response element located in the regulatory region of estrogen-responsive genes. Many of the TH-positive

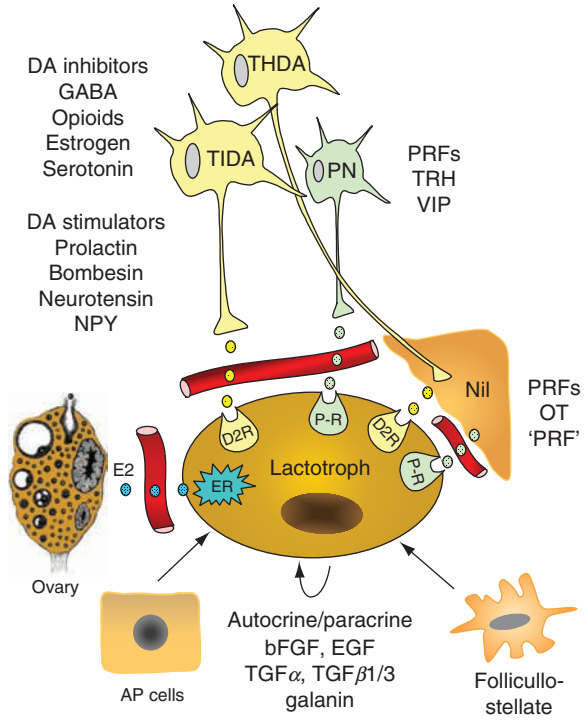

Figure 4 Model of the regulation of lactotroph function by hypothalamic, peripheral, and paracrine/autocrine factors. Dopamine from the TIDA and THDA neurons reaches the anterior pituitary (AP) via the long and short portal vessels, respectively, and serves as the primary PRL inhibiting factor. Compounds that inhibit dopamine cause stimulation of PRL release, whereas those that stimulate dopamine result in its inhibition. Direct stimulators of the lactotrophs are designated as PRFs (PRL releasing factors) and originate either from hypothalamic peptidergic neurons (PN) or from the neurointermediate lobe (NIL). D2R and P-R designate dopamine type 2 receptors and peptidergic receptors, respectively. Estrogen (E2) affects PRL directly or indirectly by modulating the dopaminergic neurons. Local peptides and growth factors are produced by neighboring AP cells, folliculostellate cells, or the lactotrophs.

neurons in the arcuate nucleus express ERα, whose expression changes during the estrous cycle. The importance of estrogen in the regulation of TIDA neurons is underscored by the higher basal TIDA activity in females than in males, the suppression of their activity by ovariectomy, and their reversal by estrogen.

Interactions of the Dopaminergic Neurons with Neuropeptides and Neurotransmitters

Opioids are the best-established neuropeptides that stimulate PRL release by inhibiting the TIDA neurons (Figure 4). The hypothalamus contains perikarya of the three opioid classes – that is, proopiomelanocortin (POMC), endorphins, and enkephalins. Although the arcuate nucleus contains POMC-positive cells, only a few TIDA perikarya receive contacts from β-endorphin terminals, with most interactions occurring at their median eminence terminals. In contrast,

most TIDA neurons make synaptic contacts with dynorphin terminals, with proenkephalin- and prodynorphin-related peptides acting as paracrine regulators of TIDA neurons via κ- and μ-opioid receptors. Interactions between opioid peptides and TIDA neurons are especially extensive during the proestrus PRL surge, its nocturnal rise in pregnant rats, and during suckling, indicating that they comprise an integral component of the PRL regulatory system. Several neuropeptides, including neuropeptide Y, neurotensin, angiotensin II, calcitonin, bombesin-like peptides, and atrial natriuretic peptides, decrease PRL release by stimulating the dopaminergic systems (Figure 4).

Among neurotransmitters, the stimulatory effect of serotonin on PRL release is well recognized (Figure 4). Serotonergic neurons from the raphe nucleus terminate in the superchiasmatic and arcuate nuclei. Drugs that impair serotonergic neurotransmission or lesions of the raphe nucleus prevent PRL increases in response to stress and suckling. However, a consistent effect of serotonin agonists on TIDA neuronal activity has not been observed and serotonin may affect PRL secretion primarily by stimulating the release of PRF. γ-Aminobutyric acid (GABA) has a dual effect on PRL release: a direct inhibitory action at the level of the anterior pituitary and a central stimulation. There is also an endogenous rhythm of cholinergic neurons which appears to contribute to the circadian changes of TIDA neuronal activity.

Mechanism of Action of Dopamine

Dopamine exerts its action on the lactotrophs by binding to dopamine D_2 receptor, a seven domain G-protein-coupled receptor. Based on sequence and structural similarities, the five dopamine receptors are subdivided into two groups. The D1-like receptors (D1 and D5) increase adenylyl cyclase activity in response to dopamine, whereas the D2-like receptors (D2–D4) inhibit adenylyl cyclase. Alternative splicing of the D2 receptor yields two isoforms, long (D_2L) and short (D_2S). The short isoform has a 29-amino-acid deletion in the third intracytoplasmic domain, where interactions with G_0/G_i proteins occur. The long isoform is predominant in the pituitary and its expression is regulated by sex steroids. Despite a similar binding affinity to dopamine, the two isoforms differ in their coupling to second messengers and account for the different responsiveness of lactotroph subpopulations to dopamine.

The diverse effects of dopamine on the lactotrophs are best exemplified by temporal events. Within seconds, dopamine increases potassium conductance and inactivates voltage-sensitive calcium channels. This results in membrane hyperpolarization, reduced intracellular calcium, and inhibition of PRL release. Since PRL release from storage granules is coupled to calcium availability, the relatively high intracellular calcium accounts for the high basal release of PRL typical of cultured lactotrophs in the absence of dopamine and the immediate suppression of PRL release upon exposure to dopamine. Within minutes to hours, dopamine suppresses adenylyl cyclase activity as well as inositol phosphate metabolism and arachidonic acid release, followed by downregulation of the PRL gene. Within days, dopamine suppresses lactotroph proliferation.

Prolactin Releasing/Regulating Factors

Thyrotropin-Releasing Hormone, Vasoactive Intestinal Peptide, and Prolactin-Releasing Peptides

In addition to compounds that affect PRL release by altering the dopaminergic system, several peptides affect PRL gene expression/release by acting directly on the lactotrophs (Figure 4). Following the discovery of the hypothalamic releasing/inhibiting factors, strict criteria have been established for this class of compounds. These include localization in the hypophysiotropic area of the hypothalamus, the presence in portal blood, binding to specific receptors on lactotrophs, and alterations in their expression/release that reflect changes in PRL secretion. Because many of these compounds were later found to also be produced within the pituitary, broadening of the criteria for PRL regulating factors became necessary.

Thyrotropin-releasing hormone (TRH) was initially defined by its ability to regulate thyrotropin-stimulating factor, but it is also an effective stimulator of PRL secretion. TRH is a tripeptide with a C-terminal amide and a pyro-Glu residue at the N-terminus. The prohormone is cleaved by prohormone convertases to yield five copies of TRH. TRH-producing neurons are located in the paraventricular nuclei and have terminals in the median eminence, where their product is secreted into portal blood. TRH binds to membrane receptors, present on both thyrotrophs and lactotrophs, that belong to the family of seven domain G-protein-coupled receptors. TRH stimulates PRL release both *in vivo* and *in vitro*, especially when the dopaminergic input is absent or reduced. It induces a rapid, biphasic increase in intracellular free calcium, leading to increased PRL release. TRH also induces the PRL gene via protein kinase C- and calcium-dependent activation of MAP kinase pathways. Although TRH is an important modulator of PRL release, it is not considered the primary PRF. A TRH stimulation test is used by clinicians to distinguish prolactinomas from other causes of hyperprolactinemia in patients.

Vasoactive intestinal peptide (VIP) is a 28-amino-acid peptide present at high concentrations in the superchiasmatic and paraventricular nuclei and in hypophysial portal blood. VIP and related peptides, peptide histidine–isoleucine and pituitary adenylate cyclase-activating polypeptide, stimulate PRL release both *in vivo* and *in vitro*. This is supported by passive immunization studies and the increase in hypothalamic VIP during sexual maturation, lactation, and following estradiol treatment. Hyperprolactinemia causes a reduction in VIP concentrations in portal blood, suggesting negative feedback by PRL. VIP is also produced by anterior pituitary cells, including the lactotrophs, suggesting that it also acts as a paracrine/autocrine agent that maintains elevated basal PRL release. VIP binds to two subclasses of G-protein-coupled receptors named VIP-R1 and VIP-R2. Despite their high homology, each subtype is coupled to specific signaling pathways and its expression is tissue- and cell-specific. VIP acts by increasing the production of cyclic AMP, followed by activation of protein kinase A. On a molar basis, VIP is more sluggish and less potent than TRH as a PRL secretagogue. Similar to TRH, VIP stimulates PRL gene expression, but it also enhances PRL mRNA stability.

Two peptides, consisting of 20 and 31 amino acids, were named PRL-releasing peptides (PrRP20 and PrRP31) based on their ability to increase PRL release *in vitro*. Both peptides originate from a larger precursor molecule by proteolytic cleavage and have amidated C-terminals. It is currently debated whether they truly deserve their assigned name. Upon incubation with rat pituitary cells, only PRL release was increased, whereas other pituitary hormones were unchanged. However, this required pharmacological rather than physiological doses. In addition, peptide expression is highest in the nucleus of the solitary tract and the adrenal medulla, with low to undetectable levels in the ventral hypothalamus or median eminence, raising the question of how they can reach the pituitary gland. Since both the peptides and their receptors are abundant in rat and human pituitaries, one possibility is that they regulate PRL release in an autocrine/paracrine manner. If so, they do not differ from several locally produced peptides/growth factors capable of stimulating PRL secretion but are not classified as hypothalamic releasing factors.

Prolactin Regulators within the Pituitary

The posterior pituitary in most species is composed of two adjacent tissues, the NL and IL, which differ in embryonic origin, innervation, vascularization, cell types, and hormonal products. The NL extends from the median eminence and is composed of nerve endings and pituicytes, or astroglial cells. The IL shares its origin and many features with the glandular AL. The main cell type within the IL is the POMC-producing melanotroph. The short portal vessels that connect the NL and AL lobes do not penetrate the IL, and this layer of cells is practically avascular. Two compounds, oxytocin from the NL and an unidentified substance from the IL, act as PRFs (**Figure 4**).

Oxytocin is a nonapeptide produced by the magnocellular neurons of the paraventricular and supraoptic nuclei. Posttranslational processing of prooxytocin occurs inside secretory granules during their axonal transport to the NL. Both oxytocin and its carrier protein, neurophysin I, are stored in nerve terminals in the NL and are released in equimolar amounts in response to action potentials which originate in the hypothalamic perikarya. Oxytocin can reach the AL either from the median eminence via the long portal vessels or from the NL via the short portal vessels, and it binds to receptors present on lactotrophs and other cell types. Oxytocin is released at times of elevated PRL secretion, such as during suckling, the afternoon of proestrus, following estradiol administration, and during immobilization stress. Since oxytocin and PRL have different thresholds of activation and dissimilar kinetics, their co-release may be coincidental or functionally related. The consensus is that oxytocin is not a major physiological PRF. Relative to TRH, the potency of oxytocin is much lower, but it is much more abundant and thus it can serve as a modulator of PRL release under some conditions.

Another PRF was serendipitously discovered in the posterior pituitary when posterior pituitary lobectomized (LOBEX) rats were used to examine the role of THDA neurons in the suckling-induced PRL rise. Basal PRL release was elevated in LOBEX lactating rats, confirming that dopamine from THDA neurons contributes to the tonic inhibition of PRL release. However, suckling, with or without oxytocin administration, failed to increase plasma PRL levels, indicating the presence of PRF. Additional studies have shown that this PRF participates in the generation of the proestrus PRL surge, mediates the acute estradiol-induced increase in PRL, contributes to the nocturnal increase in PRL during early pregnancy, but is not involved in the stress-induced PRL release. This PRF activity is localized in the IL but not in the NL, hypothalamus, or cerebral cortex. Posterior pituitary extracts stimulated PRL release in a rapid, dose-dependent and hormone-specific manner both *in vivo* and *in vitro*. Co-culturing posterior pituitary cells and lactotrophs resulted in a dramatic increase in PRL mRNA levels. Attempts to isolate this PRF from posterior pituitaries from several species were unsuccessful.

Many locally produced peptides and growth factors affect PRL production/release as well as lactotroph proliferation (Figure 4). Unlike the classical releasing/inhibiting factors characterized by a very rapid action on PRL release, local factors primarily affect the PRL gene with a subsequent effect on PRL release. The intrapituitary regulatory system involves not only hormone-producing cells, such as lactotrophs and gonadotrophs, but also folliculostellate cells. The latter form an intricate cellular network via slender processes and are involved in the coordination and synchronization of neighboring cells. Local PRL regulatory factors include epidermal growth factor, basic fibroblast growth factor, transforming growth factor-α (TGF-α), TGF-β, calcitonin, and galanin. The role of TGF-β is particularly intriguing since TGF-β_1 and TGF-β_3, two isoforms encoded by different genes, stimulate and inhibit PRL expression and lactotroph proliferation, respectively. Most of the local factors also interact with dopamine and/or estrogen by affecting their receptors or signaling pathways.

See also: Dopamine Neurons: Reward and Uncertainty; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Pituitary Gland (Cell Types, Mediators, Development); Prolactin-Releasing Peptide.

Further Reading

Ben-Jonathan N and Hnasko RM (2001) Dopamine as a prolactin inhibitor. *Endocrine Reviews* 22: 724–763.

Ben-Jonathan N, Khurana S, and Hnasko R (2002) Brain prolactin. *Hormones and Behavior* 5: 97–120.

Ben-Jonathan N and Liu JW (1992) Pituitary lactotrophs: Endocrine, paracrine, juxtacrine, and autocrine interactions. *Trends in Endocrinology and Metabolism* 3: 254–258.

Bole-Feysot C, Goffin V, Edery M, et al. (1998) Prolactin (PRL) and its receptor: Actions, signal transduction pathways and phenotypes observed in PRL receptor knockout mice. *Endocrine Reviews* 19: 225–268.

Denef C (2003) Paracrine control of lactotrope proliferation and differentiation. *Trends in Endocrinology and Metabolism* 14: 188–195.

Freeman ME, Kanyicska B, Lerant A, and Nagy G (2000) Prolactin: Structure, function, and regulation of secretion. *Physiological Reviews* 80: 1523–1631.

Goffin V, Binart N, Touraine P, and Kelly PA (2002) Prolactin: The new biology of an old hormone. *Annual Review of Physiology* 64: 47–67.

Gourdji D and Laverriere JN (1994) The rat prolactin gene: A target for tissue-specific and hormone-dependent transcription factors. *Molecular and Cellular Endocrinology* 100: 133–142.

Horseman ND (2001) *Prolactin*. Boston: Kluwer.

Hu Z-Z, Zhuang L, and Dufau ML (1998) Prolactin receptor gene diversity: Structure and regulation. *Trends in Endocrinology and Metabolism* 9: 94–102.

Moore KE and Lookingland KJ (2000) Dopaminergic neuronal systems in the hypothalamus. In: Meador-Woodruff JH (ed.) *Psychopharmacology*, pp. 1–14. Nashville, TN: American College of Psychoneuropharmacology.

Pan J-T (1996) Neuroendocrine functions of dopamine. In: Stone TW (ed.) *CNS Neurotransmitters and Neuromodulators: Dopamine*, pp. 213–231. Boca Raton, FL: CRC Press.

Schwartz J (2000) Intercellular communication in the anterior pituitary. *Endocrine Reviews* 21: 488–513.

Sinha YN (1995) Structural variants of prolactin: Occurrence and physiological significance. *Endocrine Reviews* 16: 354–369.

Prolactin-Releasing Peptide

A Mochiduki, S Devnath, and K Inoue, Saitama University, Saitama, Japan

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The human genome project was completed in 2003, and the draft of human genome sequences was published. One of the most surprising findings of the human genome sequences was the existence of a large number of G-protein-coupled receptor (GPCR) candidates in the genome. Most of these receptors are classified as odor receptors; however, about 170 GPCRs were considered orphan receptors, and their natural ligands were not known. Therefore, many studies have sought to discover the endogenous ligands for these orphan receptors. Prolactin-releasing peptide (PrRP) was found by a reverse pharmacological search for an endogenous ligand of an orphan receptor (hGR3/UHR-1/GPR10) highly expressed in the anterior pituitary gland. PrRP was first reported as a good candidate for prolactin-releasing hypophysiotropic factor; however, much recent research suggests that besides prolactin (PRL) release, PrRP has a wide range of functions in the central nervous system, such as metabolic homeostasis, stress responses, cardiovascular regulation, gonadotropin secretion, growth hormone (GH) secretion, and sleep regulation.

Chemical Structure of PrRP and Its Receptor

PrRP has two isoforms, PrRP31 and PrRP20, which are composed of 31 and 20 amino acid residues and have molecular weights of 3.6 and 2.2 kDa, respectively. Bovine PrRP complementary DNA (cDNA) isolated from hypothalamus was found to encode a precursor protein with 98 amino acid residues. This protein has a typical N-terminal secretory signal sequence of 22 amino acid residues. PrRP31 starts from Ser-23 just after the signal sequence, whereas PrRP20 starts from Thr-34. PrRP has turned out to be RFamide peptide, which contains the RF (Y) amide motif in its C terminal. PrRP also undergoes some enzymatic processes to be amidated like other RFamide peptides. First, signal peptide is removed from the bovine PrRP precursor protein, followed by cleavage in the basic amino acid of Arg-55, Arg-56, and Arg-57. After these basic amino acids are removed, Phe-53 at the C-terminus of the mature PrRP is amidated by reacting with Gly-54 as an amide donor. Human and rat PrRP precursor proteins, with amino acid lengths of 87 and 83, respectively, have essentially the same structure as the bovine precursor protein and have sequence similarity of 90% and 83%, respectively, with bovine PrRP31 (Figure 1).

PrRP receptor is a seven-transmembrane-domain receptor, that span the cell membrane seven times. PrRP is coupled to guanine nucleotide-binding protein (G-protein), a heterotrimer consisting of α, β, and γ subunits as machinery to transduce intracellular signals, and the result is called G-protein-coupled receptor. In GH_3 rat pituitary tumor cells and primary cultures of rat anterior pituitary cells, PrRP rapidly and transiently activated extracellular signal-regulated protein kinase (ERK). This PrRP-induced ERK activation was completely blocked by pertussis toxin, which inactivates G_i/G_o proteins. This finding suggests that the PrRP receptor couples to either of the G-proteins.

Distribution of PrRP

Among the variety of tissues studied, however, major PrRP distribution was noted in the brain. An *in situ* hybridization study has shown PrRP messenger RNA (mRNA) to be localized in the caudal portion of the dorsomedial nucleus of hypothalamus (DMH), nucleus of the tractus solitarius (NTS), and ventral and lateral reticular nucleus (VLRN) in the brain, with the strongest PrRP mRNA detected in the NTS area. PrRP mRNA has also been detected in some peripheral tissues of the adrenal gland, pancreas, testis, and placenta. Immunohistochemical assay has shown that PrRP-producing cells/PrRP-cells are located in the NTS, VLRN, and DMH in the brain. PrRP-producing cells have also appeared in the area postrema (AP) of adrenalectomized rats.

In contrast, PrRP fibers have been detected in a wide range of the brain, such as the paraventricular hypothalamic nucleus (PVN), basolateral amygdaloid nucleus, bed nucleus of the stria terminalis, supraoptic nucleus, and paratenial thalamic nucleus. However, the external region of the median eminence, site of the classical hypophysiotropic hormones, was devoid of immunopositive PrRP fibers. This distribution profile indicates that PrRP may have a wide range of biological functions.

Distribution of PrRP Receptor

To elucidate the specific sites of action of PrRP, one should know the distribution of its receptor. Various advanced molecular techniques, that is, *in situ* hybridization, polymerase chain reaction (PCR), reverse transcription PCR (RT-PCR), Northern blotting, and autoradiography, have made it possible to know

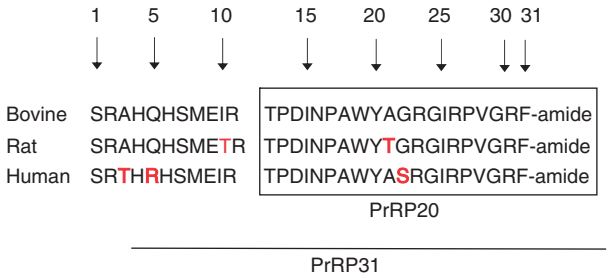

Figure 1 Amino acid sequence of PrRP20 and PrRP31. The numbers and arrows indicate the position of amino acid from N terminal in the mature peptide. Here we additionally represent the human PrRP sequence to compare with other two species' PrRP (similarity and dissimilarity). Red letters show the amino acids different from bovine PrRP.

the tissue distribution of PrRP receptor and PrRP receptor mRNA. Both RT-PCR and *in situ* hybridization have confirmed a dense expression of PrRP receptor mRNA in rat pituitary gland, indicating that this gland is the major site for PRL secretion in the body. High expression of human PrRP receptor mRNA in the pituitary gland, among the tissue examined by Northern blot analysis, has strengthened this concept. *In situ* hybridization has also found PrRP receptor mRNA expressed in the brain (Figure 2). This receptor is expressed abundantly in the reticular thalamic nucleus, periventricular hypothalamus, dorsomedial hypothalamus, NTS, AP, and adrenal medulla. These areas of the brain are important for the control of neuroendocrine and autonomic functions such as cardiovascular regulation, homeostasis, gonadotropin secretion, and stress, which suggests the involvement of PrRP in the mediation of these functions.

Functions of PrRP

PRL Secretion of PrRP

As for biological activities, PrRP was first reported to stimulate PRL secretion from rat pituitary adenoma-derived cell line RC-4B/C in a dose-dependent manner and from dispersed rat anterior pituitary cells. In addition, PrRP31 was not found to promote the secretion of the other anterior pituitary hormones, such as GH, follicle-stimulating hormone (FSH), luteinizing hormone (LH), thyroid-stimulating hormone, or adrenocorticotropin (ACTH), but PrRP31 immediately promoted the secretion of PRL from rat anterior pituitary cells, indicating that PrRP might directly stimulate lactotroph cells to secrete PRL. As revealed from other studies, however, the stimulating effect of PrRP on PRL release is less than that of thyrotropin- releasing hormone, which has been shown to have more potential in inducing PRL secretion. More interesting, a quite opposite result has been reported, that is, that PrRP had no direct stimulating effect on PRL secretion from the dispersed pituitary cells. It has also been reported that PrRP has a much weaker effect on PRL secretion in rats in stress or suckling conditions.

On the other hand, following PrRP injection, PRL secretion has been found to vary according to gender, estrous cycle, and dose. A relatively high dose (500 nmol PrRP31 kg^{-1} IV) was found to be effective for a significant increase in plasma PRL level in male rat whereas only 50 nmol PrRPkg^{-1} IV enhanced the plasma PRL level in female rat within 25 min. Female rat, especially at the proestrous and estrous stage, proved more sensitive than male rat to PrRP-induced PRL secretion. In addition, because PRL release is regulated by estrogen, dopamine, and several other hypothalamic factors, these factors might have some influencing effects on PrRP.

By immunocytochemical study, it has been found that immunoreactive PrRP fibers project only onto the internal layer of median eminence, whereas other classical hypophysiotropic hormone-producing cells project onto the external layer of median eminence. These results suggest that the mechanism and pathway for PrRP's action on the anterior pituitary differ from those of the classic hypophysiotropic peptides regulating the secretion of pituitary hormones. Actually, this stimulating mechanism is not straightforward. Pharmacologically, PrRP is shown to stimulate PRL secretion, but it has no physiological basis; physiological investigation is needed. The PRL secretion-stimulation mechanism is described on the basis of two possible assumptions. On the one hand, the dense existence of PrRP receptor in the anterior pituitary gland suggests that PrRP stimulates PRL release in an autocrine fashion. On the other hand, PrRP is transported to the posterior pituitary gland through the blood vessel between the hypothalamus and posterior pituitary gland. This PrRP from the posterior pituitary gland might be transported via the short loop portal vein system into the anterior pituitary, where lactotroph cells exist. However, this seems rather unlikely, and additional research is required to resolve this discrepancy.

Stress Responses of PrRP

These investigations demonstrate that PrRP fibers and cells are widely distributed throughout the brain. Actually, this distribution pattern of PrRP has broadened the way for scientists studying whether PrRP stimulates the stress mechanism in animals. Initially, it was found that PrRP-like immunoreactivity (PrRP-LI) colocalized with tyrosine hydroxylase-like immunoreactivity (TH-LI) in A1 (VLRN) and A2 (NTS)

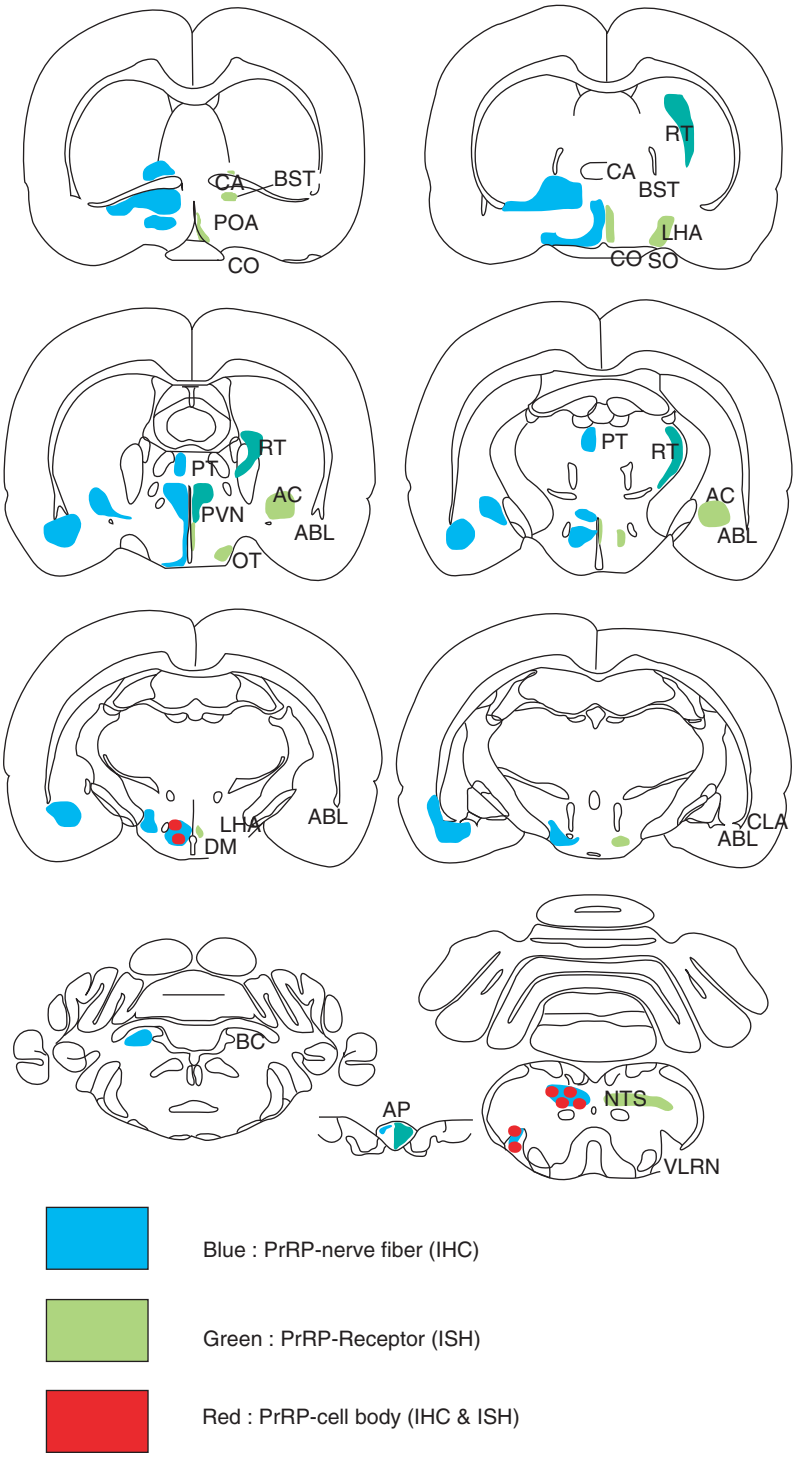

Figure 2 Distribution of PrRP and UHR-1 in the brain. Blue, PrRP nerve fiber (immunohistochemistry); green, PrRP-Receptor (*in situ* hybridization); red, PrRP-cell body (immunohistochemistry and *in situ* hybridization). ABL, basolateral amygdaloid nucleus; AC, central nucleus of amygdala; AP, area postrema; BC, superior cerebellar peduncle; BST, bed nucleus of the stria terminalis; CA, anterior commissure; CLA,; CO, optic chiasm; DM, dorsomedial hypothalamic necleus; LHA, lateral hypothalamic area; NTS, nucleus tractus solitarius; POA, preoptic area; PT, paratenial thalamic nucleus; PVN, paraventricular hypothalamic nucleus; RT, reticular thalamic nucleus; SO, supraoptic nucleus; VLRN, ventral and lateral reticular nucleus.

noradrenergic neurons of the rat caudal medulla. However, some of the TH-LI neurons appeared to be PrRP-negative in the A2 region. Noradrenergic neurons in A1 and A2 cell groups appear to send their axons to innervate various regions of the rat hypothalamus, including median eminence, arcuate, dorsomedial, ventromedial, periventricular, and paraventricular nuclei. Also, PrRP neuron cells in NTS and VLRN project

their fibers/axons to corticotropin-releasing hormone (CRH)-positive parvocellular neurosecretory cells in the PVN and then make synapselike contact with CRH cell bodies. This suggests that PrRP might have some effect on CRH secretion. In addition, plasma ACTH level has been found to increase dramatically after intracerebroventricular (icv) administration of PrRP. On the other hand, animals pretreated with a potent CRH antagonist (α-helical CRH) showed attenuated plasma ACTH levels after icv PrRP31 application. These results strongly suggest that PrRP affects the hypothalamic–pituitary–adrenal (HPA) axis as a potent stimulator of CRH neurons in the PVN. Furthermore, colocalized stay with a marker enzyme (tyrosine hydroxylase) of catecholaminergic neurons is strong evidence for PrRP in stress-responsive signal transduction through the HPA axis. In any case, activation of the HPA axis results in the secretion of CRH from the CRH cells located in hypothalamic PVN, which in turn enhances ACTH secretion from the pituitary gland (Figure 3).

Regulation of PrRP cell function is important for understanding the physiological response of PrRP cells. It has been reported that nicotine stimulates PrRP cells. In addition, c-Fos expression in the A2 PrRP neurons is a marker of neural activation and c-Fos-positive cells in the PrRP neurons have been found to be lowest in estrus. Investigators have also found that estrogen receptor α is positive in A2 PrRP neurons, which suggests that PrRP cells may regulate with estrogen. In addition, they studied the effect of estrogen on the stress response of A2 PrRP neurons and found that stress responses of PrRP cells are suppressed by estrogen.

Earlier studies revealed that different types of stressors (e.g., water-immersion, foot shock, hemorrhage, and conditioned fear) activate PrRP/TH double positive cells/neurons in the A1 and A2 regions of medulla oblongata, resulting in the elevation of ACTH in plasma, which in turn stimulates the release of glucocorticoids and catecholamines from the adrenal glands.

Also, synapic contact of PrRP neurons with oxytocin cell bodies in the hypothalamus has been demonstrated by means of double immunohistochemical study. It has also been reported that icv administration of PrRP stimulates oxytocin release from the pituitary. PrRP-producing neurons in the medulla oblongata are known to produce both PrRP and noradrenaline simultaneously. These morphological features of PrRP neuron suggest that PrRP and noradrenaline might synergistically affect the CRH neurons in the secretion of ACTH. On this point, it has been reported that icv administration of low doses of PrRP (1 nmol per rat) and noradrenaline (0.1 nmol per rat) could not significantly stimulate ACTH secretion through CRH, and vice versa. On the other hand, simultaneous administration of these subactive doses of PrRP and noradrenaline markedly stimulated ACTH secretion. This result suggests that the combined effect of PrRP and noradrenaline, even at a subactive dose, is stronger than that of a separate single dose, either high or low, in stimulating secretion of ACTH.

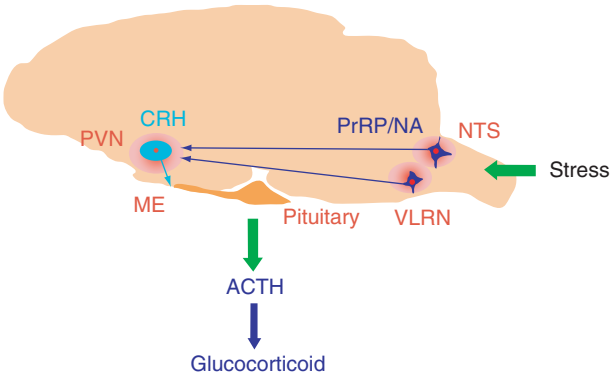

Figure 3 Activation of hypothalamic–pituitary–adrenal axis. ACTH, adrenocorticotropin; CRH, corticotropin-releasing hormone; ME, Median eminence; NTS, nucleus tractus solitarius; PrRP/NA, prolactin-releasing peptide/noradrenaline; PVN, paraventricular hypothalamic nucleus; VLRN, ventral and lateral reticular nucleus.

Food Intake Regulation of PrRP

PrRP mRNA is expressed in three regions of brain: DMH, NTS, and VLRN. These regions are also implicated in appetite control and autonomic regulation. Therefore, PrRP might have some role in the regulation of food intake. Previous study has suggested that PrRP may act as a homeostatic regulator of food intake. It has been found that 4 nmol PrRP significantly reduces the initial refeeding response in fast-induced female rats. This result indicates that PrRP has a physiological role to play in appetite and weight control. This reduction in food intake was not owing to visceral illness or injury after icv injections of PrRP (2–4 nmol). However, water intake was unaffected by the treatment. In animal physiological systems, there are some special types of endogenous, homeostatic satiety factors (cholecystokinin, for instance) that signal to the brain via vagal nerve afferents being released by the gut during a meal. Also, some exogenous, nonhomeostatic factors such as lithium chloride cause visceral illness or injury and reduce food intake thereby. Although these two factors have a different origin, they seem to exert their action on ingestive behavior by the same action pathway, via vagal nerve afferents. Likewise, on satiation, PrRP causes a pattern of brain/neural activation similar to that produced by cholecystokinin. However, PrRP does

not cause conditioned taste aversion but instead enhances the normal behavioral satiety sequence. Moreover, PrRP has been found to activate some additional areas in the brain, such as lateral hypothalamus and hypothalamic PVN. It is interesting that both of these areas of brain have been implicated in appetite regulation. Actually, PrRP inhibits food intake via the DMH because this area of brain contains a high level of neurotensin, a potent food intake inhibitor. It has also been found that PrRP can stimulate the release of neurotensin from hypothalamic explants in a dose-dependent manner. The increase in neurotensin by PrRP could be an important mediator of the anorectic effect of PrRP. In addition, PVN injection of PrRP does not appear to change food intake in male rat.

Coadministration of PrRP and leptin, an anorexigen, resulted in additive reductions in nocturnal food intake and body weight gain and an increase in core body temperature compared with either peptide alone. Using immunohistochemical study, it has been shown that more than 90% of PrRP cells in all regions (DMH, NTS, and VLM) contain leptin receptors. This result suggests that the effect of PrRP on food intake is mediated, at least in part, by leptin.

On the other hand, icv administration of PrRP increases food intake in the chicken. This result is completely opposite to those obtained in mammals. Further extensive research is required in this area.

Other Functions of PrRP

Morphological studies have shown that PrRP fibers are distributed in different parts of the brain. As the brain controls almost all mechanisms in the body, either directly or indirectly, PrRP might have some involvement in those mechanisms. Beyond PRL release, stress response, and food intake inhibitory functions, PrRP is also known to play some important physiological roles in cardiovascular regulation, gonadotropin secretion, growth hormone secretion, sleep regulation, and others. The stress response of PrRP has been reported already. During stress conditions, circulating catecholamine increases the heartbeat, producing high blood pressure. In experiments with rats, arterial blood pressure was significantly increased by the icv administration of PrRP. Elevation of arterial pressure was observed in response to 0.4 and 4 nmol PrRP31. The effect of PrRP continued 15 min following administration of 0.4 nmol and 20 min following administration of 4 nmol PrRP. It is also known that GPR10 and PrRP are associated with blood pressure.

Gonadotropin-releasing hormone (GnRH) neuronal perikarya exist in medial preoptic area (MPOA), and GnRH neuronal fibers terminate in arcuate nucleus (ARC). It has been found that PrRP is released in the MPOA prior to the GnRH release, whereas in the ARC area the release of PrRP is not changed significantly. This suggests that PrRP may play a role in mediating the steroid-induced LH surge by activating GnRH neurons in the MPOA. It has also been reported that plasma LH and FSH were increased on central administration of PrRP in male rat.

PrRP-positive nerve fibers make synapselike contact with somatostatin neurons in the hypothalamic periventricular nucleus (PerVN). Some parts of somatostatin neurons in PerVN express PrRP receptor; therefore, they are considered to be directly regulated by PrRP via PrRP receptor. In addition, icv administration of PrRP decreases plasma GH level, but rat pretreated with somatostatin depleter, cysteamine, does not show any decrease in GH. These results strongly suggest that plasma somatostatin level increases on PrRP administration, which in turn inhibits growth hormone secretion from the anterior pituitary gland.

PrRP increases sleep time in a dose-dependent manner. Low-dose PrRP enhances only rapid eye movement (REM), whereas high-dose PrRP (10 nmol) enhances non-REM sleep with a febrile response. PrRP receptors express prominently in reticular thalamic nucleus (RTN) of brain. This region of brain is critical for sleep regulation. Icv injection of PrRP was found to suppress sleep oscillation in sleeping rats and promote rapid and prolonged awakening. Given the critical role of the RTN neurons in the generation and control of thalamic oscillatory activity, activation of PrRP receptors in this region could be used to influence sleep–wake state and absence seizures.

The endogenous opioid system is an innate pain-relieving system in animals. It produces opioids that diminish pain and produce analgesia. PrRP has been found to destroy these opioids, resulting in increased pain. PrRP receptor knockout mice were found to have higher nociceptive thresholds and stronger stress-induced analgesia than wild-type mice. Icv administration of PrRP in wild-type mice increased hyperalgesia and reversed morphine-induced anti-nociception, whereas PrRP receptor knockout mice did not show any change, suggesting that PrRP is a strong modulator of the opioid system.

Localization of PrRP mRNA in the adrenal gland has also been reported. The function of PrRP cells in the adrenal gland remains to be elucidated.

Conclusion

As implied by its name, PrRP releases PRL; more specifically, it stimulates lactotroph cells to release PRL. However, recent investigations have modified this concept, and its other physiological roles are beginning to receive more emphasis, although even

more is needed. PrRP's broad-spectrum roles in the body include metabolic homeostasis, stress response, cardiovascular regulation, gonadotropin secretion, GH secretion, and sleep regulation. Despite, we think, the presence of its receptors outside the HPA axis, PrRP may have many other functions yet to discover. In fine, PrRP's research horizon is vast.

See also: Growth Hormone; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Prolactin and its Neuroendocrine Control; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress, the HPA Axis and Depressive Illness; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load; Sympathoadrenal System: Neural Arm of the Stress Response.

Further Reading

Ellacott KLJ, Lawrence CB, Rothwell NJ, and Luckman SM (2002) PRL-releasing peptide interacts with leptin to reduce food intake and body weight. *Endocrinology* 143(2): 368–374.

Hinuma S, Habata Y, Fujii R, et al. (1998) A prolactin-releasing peptide in the brain. *Nature* 393: 272–276.

Laurent P, Becker JAJ, Valverde O, et al. (2005) The prolactin-releasing peptide antagonizes the opioid system through its receptor GPR10. *Nature Neuroscience* 8(12): 1735–1741.

Lawrence CB, Celsi F, Brennand J, and Luckman SM (2000) Alternative role for prolactin-releasing peptide in the regulation of food intake. *Nature Neuroscience* 3: 645–646.

Lawrence CB, Ellacott KLJ, and Luckman SM (2002) PRL-releasing peptide reduces food intake and may mediate satiety signaling. *Endocrinology* 143(2): 360–367.

Maruyama M, Matsumoto H, Fujiwara K, et al. (1999) Immunocytochemical localization of prolactin-releasing peptide in the rat brain. *Endocrinology* 140: 2326–2333.

Maruyama M, Matsumoto H, Fujiwara K, et al. (2001) Prolactin-releasing peptide as a novel stress mediator in the central nervous system. *Endocrinology* 142: 2032–2038.

Matsumoto H, Noguchi J, Horikoshi Y, et al. (1999) Stimulation of prolactin release by prolactin-releasing peptide in rats. *Biochemical and Biophysical Research Communications* 259: 321–324.

Samson WK, Resch ZT, and Murphy TC (2000) A novel action of the newly described prolactin-releasing peptides: Cardiovascular regulation. *Brain Research* 858: 19–25.

Seal LJ, Small CJ, Dhillo WS, et al. (2001) PRL-releasing peptide inhibits food intake in male rats via the dorsomedial hypothalamic nucleus and not the paraventricular hypothalamic nucleus. *Endocrinology* 142(10): 4236–4243.

Proopiomelanocortin

R Day, University of Sherbrooke, Sherbrooke, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, © 2004 Elsevier B.V.

Proopiomelanocortin (POMC) is a complex precursor protein that is proteolytically cleaved to a variety of biologically active and important neuroendocrine peptides. These are the highly potent opioid peptide β-endorphin (β-END) and its intermediate β-lipotropin (β-LPH), the stress peptide adrenocorticotropic hormone (ACTH), and various melanotropic peptides including α-, β-, and γ-melanocyte-stimulating hormone (MSH). Other peptides that are produced include corticotropin-like intermediary peptide (CLIP) and joining peptide (JP). The cells which principally express POMC (Figure 1) are the pituitary anterior lobe corticotrophs and intermediate lobe melanotrophs. POMC is highly synthesized in neurons of the arcuate nucleus which project to the median eminence and to a wide spectrum of other brain structures. POMC is also found in neurons of the medullary nucleus tractus solitarius (NTS). In each of these cells, the size and structure of the POMC protein precursor (241 amino acids) and mRNA (1.1 kilobases) are identical; however, a tissue-specific posttranslational processing occurs which results in the formation of a different set of end products by each POMC-producing cell. Therefore, from a single POMC precursor protein, multiple biological activities are obtained depending on which cleavages occur.

Specific enzymes generally known as the prohormone or proprotein convertases (PCs) are responsible for these processing events. Seven such enzymes have been characterized and include PC1 (also known as PC3), PC2, furin, PACE4, PC4, PC5 (also known as PC6), and PC7 (also known as LPC). These enzymes are members of a family of enzymes known as the kexin/subtilisin-like serine proteinases. The PCs specifically cleave only at the C-terminus of single or multiple basic residues, and while POMC is a substrate for the PCs, other proteins can also serve as substrates. The amino acid sequence lysine-arginine (KR) is the most frequently observed cleavage site by the PCs. Other cleavage sites include KK, RR, and RK as well as multiple basic residues. Each PC recognizes basic residue motifs, but cleavage specificity is also determined by the surrounding amino acids. The differential cleavage catalysis by each PC explains how a single precursor such as POMC can be processed into different peptides in different tissues.

The differential processing of POMC has been studied thoroughly in the pituitary. The anterior lobe corticotrophs cleave POMC principally into ACTH and β-LPH. The intermediate lobe melanotrophs also express high levels of POMC, but the main products obtained are α-MSH and β-END. The differences in processing are principally explained by a differential expression of PCs, in particular that of PC1 and PC2 (Figure 2). The anterior lobe corticotrophs express PC1 but rarely express PC2. On the other hand, the intermediate lobe melanotrophs express both PC1 and high levels of PC2. Various experiments, including cellular expression studies and *in vitro* analysis have come to the conclusion that PC1 can only cleave POMC into ACTH and β-LPH (Figure 3). The further processing of these POMC products requires the actions of PC2, which can produce β-END and α-MSH. Since PC2 is generally not expressed in the anterior lobe corticotrophs, ACTH and β-LPH are the major peptides found in these cells. In neuronal POMC-expressing cells, such as those of the ARC, both α-MSH and β-END are observed. Within these POMC ARC neurons, PC1 and PC2 are both co-localized. The role of the other members of the PC family in the processing of POMC is not clear and remains to be investigated.

While it is clear that differential processing of POMC will result in different biological activities, the most striking switching of biological activity of this precursor is observed during anterior pituitary development. Adult anterior lobe corticotrophs normally produce ACTH, but in the first 2 weeks of postnatal development, a large number of these cells produce an α-MSH-like peptide. During this same period, these same corticotrophs also produce high levels of the enzyme PC2, which is important for the production of α-MSH. It is only when the PC2 levels are reduced, after 2 weeks of postnatal development, that MSH-like peptide levels are then reduced, and afterward only occasional MSH cells can be observed. The switching of POMC biological activity (from ACTH to MSH-like peptide) during development may be functionally related to some of the observed effects of melanocortins as growth factors.

The pituitary POMC system is a major component of the physiological stress-regulating circuitry. Anterior lobe corticotrophs release ACTH which acts on the biosynthesis and release of adrenal steroids. Adrenal steroids exert a negative feedback control either directly on the anterior lobe corticotrophs or via the

hypothalamus, through corticotroph-releasing hormone (CRH) cells synthesized in the parvocellular cells of the paraventricular nucleus. CRH neurosecretory neurons project to the median eminence and CRH is released into the portal vessels. Removal of the negative feedback by adrenal steroids can be produced in animals by adrenalectomy, a procedure which stimulates ACTH release and increases anterior lobe POMC mRNA dramatically. Steroid replacement, using dexamethasone or corticosterone, prevents these effects in adrenalectomized animals.

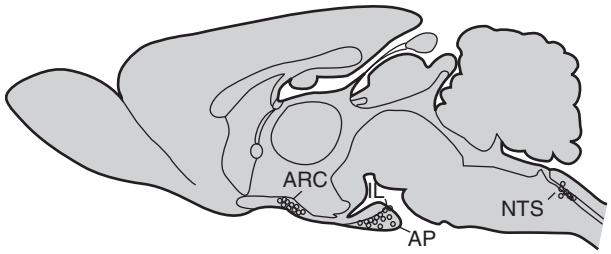

Figure 1 Sagittal diagram describing the location of POMC-synthesizing cells in the brain and pituitary. ARC, arcuate nucleus; IL, pituitary intermediate lobe; AP, pituitary anterior lobe; NTS, nucleus tractus solitarius. Figure outline is based on Swanson LW (1992) *Brain Maps: Structure of the Rat Brain*. Amsterdam: Elsevier.

Glucocorticoids may directly inhibit POMC expression at the level of the corticotroph as well as modulate hypothalamic input to the corticotroph.

In the intermediate lobe, the synthesis and secretion of POMC-derived peptides are tonically inhibited by hypothalamic axons that release dopamine. Adrenocortical steroids from the adrenal gland which feedback and inhibit the anterior lobe corticotrophs do not appear to regulate the POMC cells of the intermediate lobe. The role of the intermediate lobe POMC cells is not clear; however, there is some evidence which suggests that α-MSH secreted from the melanotrophs may act on the neighboring anterior lobe cells to recruit additional lactotrophs to release stored prolactin. It is also noted that β-END produced in the intermediate lobe melanotrophs is mostly acetylated, resulting in poorly opioid receptor active peptides. These weak opioid peptides may be secreted into the blood, but there is no evidence that they exert biological effects in the periphery.

In the brain mainly two cells groups of POMC neurons are observed in the ARC nucleus and the NTS. Both groups of cells have extensive projections in the CNS. The POMC ARC neurons, in the posterior hypothalamus, are homogeneously distributed over the entire rostrocaudal length, with some

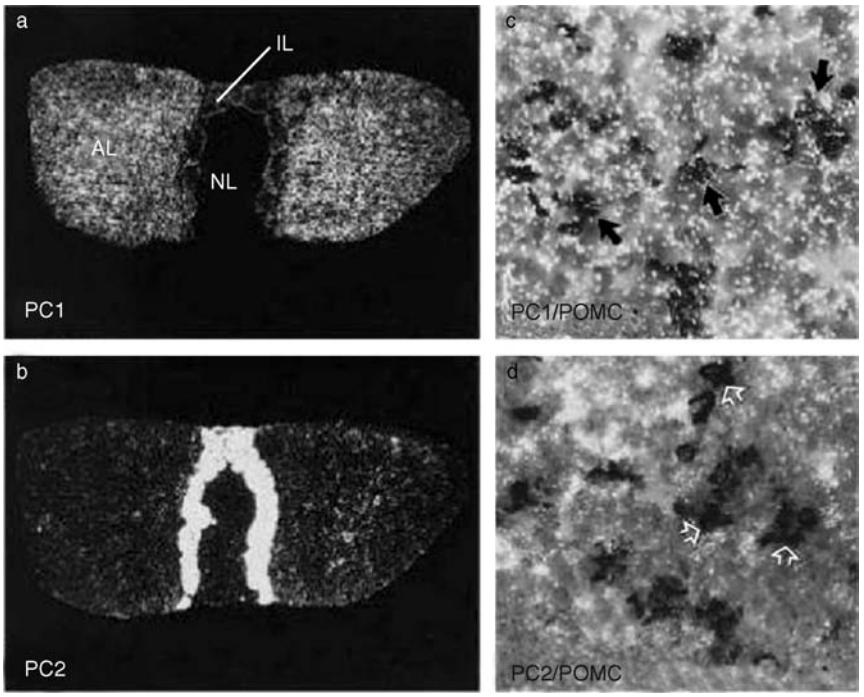

Figure 2 Distribution of PC1 and PC2 mRNA in the rat pituitary using *in situ* hybridization histochemistry. (a) PC1 and (b) PC2 mRNA distribution as displayed by silver grains which appear in white in these dark-field low-power photomicrographs. PC1 mRNA is heterogeneously expressed in many anterior lobe (AL) cells but also in all intermediate lobe (IL) melanotrophs which homogeneously express POMC. PC2 mRNA is very highly expressed in the IL, but expressed in fewer AP cells. To better define the sites of PC1 and PC2 expression in AL POMC cells, an mRNA co-localization experiment is shown in (c) and (d). In (c) PC1 mRNA (white grains) is colocalized in AL corticotrophs that also express POMC mRNA (dark staining cells) as indicated by black arrows. In (d) PC2 mRNA is not co-localized (white grains) with POMC mRNA (dark staining cells) as indicated by open arrows.

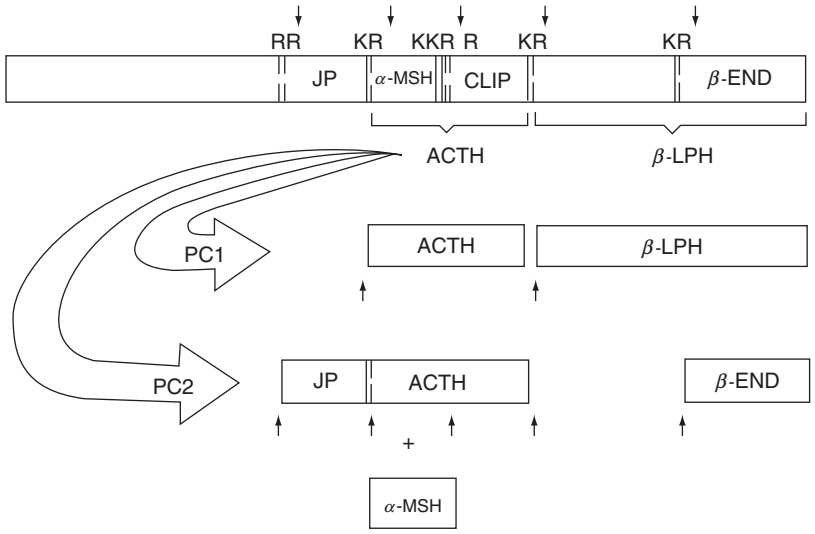

Figure 3 Differential processing of POMC by proprotein convertase (PC)1 and PC2. PC1 cleaves POMC at basic residues (as indicated by small arrows) to yield adrenocorticotropic hormone (ACTH) and β-lipotropin (β-LPH). PC2 cleaves POMC into α-melanocyte-stimulating hormone (α-MSH) and β-endorphin (β-END). This processing profile correlates with the POMC peptide products found in the pituitary corticotrophs and melanotrophs and the expression of PC1 and PC2 (see also **Figure 2**). JP, joining peptide; CLIP, corticotropin-like intermediary peptide.

scattered cells in the medial basal hypothalamus. Rostral projections pass through periventricular and diencephalic regions and innervate hypothalamic nuclei and limbic structures, including the septum, bed nucleus of the stria terminalis, and the preoptic area. Laterally, POMC fibers extend through the medial basal hypothalamus ventrally and enter the amygdala. Caudal projections course through the dorsal diencephalon into the mesencephalon and brain stem, innervating the periventricular thalamus and the periaqueductal gray, areas associated with nociception and sensory integration. Projections to the brain stem regions include, ventrally, the reticular formation and, dorsally, the nuclei parabrachialis and ambiguus, the NTS, and the regulation of the respiratory and cardiovascular systems. Areas which show no or extremely few POMC-containing fibers are striatum, cortex, and hippocampus.

The second group of POMC neurons is observed in the caudal part of the of the NTS and exhibits more local and caudal projections extending down to the spinal cord. The small POMC cells in the NTS show a more caudal orientation forming a fiber bundle between the NTS and the ventrolateral medulla. The POMC fibers observed in the NTS seem to be of local origin, whereas the catecholamine-containing brain stem nuclei and the baroreceptor areas of the lower brain stem receive a double innervation from the arcuate nucleus and the NTS. POMC fibers observed in the spinal cord originate, at least in part, from the NTS, although the presence of local POMC circuits in the cord cannot be excluded.

While the physiological role of POMC-derived peptides in the pituitary is clearly characterized by its involvement in the stress-regulating circuitry, their functional role in the CNS is less clear. Given the potent analgesic properties of β-END, it is likely to play a role in pain modulation, including phenomena such as stimulation-produced analgesia. Electrical stimulation of the periaqueductal gray, an area rich in POMC fibers, can produce profound behavioral analgesia, while stimulation of the medial thalamus, along the POMC bundle, can relieve pain. In addition, the best sites of stimulation-produced analgesia represent a pattern which coincides almost perfectly with the β-END pathway.

See also: Neuropeptides: Endocrine Cells; Opioid Peptides and Receptors.

Further Reading

Herz A (ed.) (1993) *Opioids*. Berlin: Springer.

Marcinkiewicz M, Day R, Chrétien M, and Seidah NG (1993) Ontogeny of the prohormone convertases PC1 and PC2 in the mouse hypophysis and their colocalization with ACTH and α MSH. *Proceedings of the National Academy of Sciences of the United States of America* 90: 4922–4926.

Schäfer MKH, Day R, Watson SJ, and Akil H (1990) Distribution of opioids in brain and pepipheral tissues. In: Almeida OFX and Shippenberg TS (eds.) *Opioid Peptides and Receptors: Biochemistry, Physiology and Pharmacology*, pp. 53–71. Berlin: Springer.

Seidah NG, Chrétien M, and Day R (1994) The family of subtilisin/kexin like pro-protein and pro-hormone convertases: Divergent or shared functions. *Biochimie* 76: 197–209.

Swanson LW (1992) *Brain Maps: Structure of the Rat Brain*. Amsterdam: Elsevier.

Proprioception

J L Taylor, Prince of Wales Medical Research Institute, Sydney, NSW, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Bastian, in 1887, described 'kinaesthesia' as "the sense of movement" whereby "we are made acquainted with the position and movements of our limbs, we are enabled to discriminate between different degrees of 'resistance' and 'weight,' and by means of it the brain also derives much unconscious guidance in the performance of movements generally." The term 'proprioception' is used synonymously with kinesthesia. It was first used by Sherrington in 1906 and can be thought of as referring to the sensations of the body. Although proprioception is a sense like vision, hearing, touch, taste, and smell, these more commonly recognized sensory systems give us information about the outside world, whereas proprioceptors are the sensory receptors which tell us what the body itself is doing. As Bastian suggested, kinesthesia encompasses a complex of sensations which include a sense of joint movement, a sense of joint position, a sense of muscle force or tension, and a sense of effort. Sensations derived from the vestibular apparatus are also sometimes included in proprioception, but these signals of the position and movement of the head relative to the external world will not be discussed here.

The sense of joint movement (movement sense) is the process by which we perceive movements of parts of the body relative to one another. It includes detection of movements and perception of their direction, velocity, distance, and timing. Both passive movements, which are imposed by an external force, and active movements, generated by the muscles, are perceived. The sense of joint position (position sense) is the process by which we perceive the current positions of parts of the body relative to one another. Again, relative positions can be perceived when the body parts are at rest and when they are held in position by muscle contraction. The sense of force or tension is the process by which we perceive forces generated through the muscles, whereas the sense of effort gives a perception of the strength of muscle contraction relative to the total strength of the muscle. Both of these signals are used in perception of applied forces, and in judging heaviness. All of these sensations are used separately and together to give knowledge of the body's own actions and its interactions with the environment. Thus, proprioceptive sensations provide feedback of the consequences of motor output and are vital to the control of voluntary movement.

Sensory Receptors Contributing to Proprioception

Sensory receptors in the muscles, joints, and skin are all involved in proprioception. In muscle, the major receptors for proprioception are muscle spindles and Golgi tendon organs. Muscle spindles are complex receptors which lie in parallel with the muscle fibers. They are fluid-filled fusiform (spindle-shaped) capsules of connective tissue with bundles of small (intrafusal) muscle fibers inside them. On the intrafusal fibers are two kinds of sensory nerve endings: the spiral endings of the primary muscle spindle (Ia) afferents and the flower-spray endings of the secondary muscle spindle (II) afferents. Both classes of muscle spindle endings fire with stretch of the muscle and so give a signal of muscle length. The primary ending has the more dynamic response. It fires a burst at the beginning of a stretch and signals velocity and acceleration as well as muscle length. Muscle spindles also have their own motor nerve supply to the intrafusal muscle fibers. As contraction of the intrafusal fibers stretches the sensory endings, muscle spindle firing can result from motor output as well as from muscle stretch. Thus, interpretation of muscle spindle signals during active muscle contractions is complex.

Golgi tendon organs are situated at the musculotendinous junction. They consist of several strands of collagen attached to the tendon at one end and to a number (3–25) of muscle fibers at the other end. The strands of collagen are covered by a connective tissue capsule and are innervated by a Ib sensory neuron. Although tendon organs can fire in response to passive stretch of the whole muscle, they are much more sensitive to active forces generated by the specific muscle fibers that are attached to them.

In the joints, sensory endings are located in the joint capsule and ligaments. Ruffini endings are slowly adapting afferents which are sensitive to stretch of the joint capsule. They fire most at the ends of joint range, but a few afferents can fire across the midrange of joint movement. Larger tendon-organlike receptors are found in the ligaments. These also respond to stretch but have a higher threshold compared to the joint capsule receptors.

In the skin, the receptors most directly associated with proprioception are the slowly adapting type II (SAII) receptors. These fire in response to stretch of the skin in specific directions. They can signal changes

in joint angle when the skin is stretched over one side of a joint and compressed on the other side.

Neural signals from all of these sensory receptors are carried via large-diameter myelinated sensory neurons with cell bodies in the dorsal root ganglia and projections which enter the spinal cord through the dorsal roots. Cutaneous afferents from both the upper and the lower limbs ascend directly to the medulla, where they synapse. Muscle afferents from the upper limb follow the same route whereas those from the lower limb have an intermediary synapse in Clarke's column in the spinal cord. From the medulla, proprioceptive signals cross the midline to the contralateral thalamus and from there project to the cortex.

Sensations of Movement and Position

Position sense and movement sense give perceptions about the relative locations and movements of parts of the body. These can be perceptions of angle or movement at a single joint (e.g., bend of the knee), about many joints (e.g., perception of the distance between the thumb and a fingertip), or even of position or movement that does not involve joints (e.g., of facial expressions or the direction of the eyes). The sense of movement and sense of position can be difficult to separate, as movement is a change of position and each position is reached through movement. However, it is possible to impose movements that are too slow to feel and yet, when displacement is great enough, a new position is perceived. This suggests that there is a sense of position separate from movement.

Testing Movement Sense

The sensitivity of movement sense is often tested by determining the smallest imposed movement that can be detected with the muscles relaxed. The size of this threshold movement depends on velocity, with faster movements detected more easily than slower movements (Figure 1). With velocities of $>1^\circ\,s^{-1}$, very small movements of 0.1–0.5° can be detected at limb joints and for rotation of the neck and trunk. With slower velocities, larger movements of ~1–3° are required. With very slow velocities ($2^\circ\,min^{-1}$), study participants no longer report any sensation of movement. At the interphalangeal joints of the fingers and toes, much larger angular movements are needed for detection. For the big toe, detection thresholds of ~20° at $1^\circ\,s^{-1}$ are reported. It is not clear how active muscle contraction changes the detection of imposed movements, as both impairment and improvement have been reported. The strength of contraction and the task performed may be important.

Perception of the velocity of movement is another facet of movement sense that has been tested. Discrimination between velocities of imposed movement about a joint is better for slower velocities. For movements at $\sim15^\circ\,s^{-1}$, individuals perceive differences of $2–4^\circ\,s^{-1}$, whereas for movements at $\sim50^\circ\,s^{-1}$, differences of $10^\circ\,s^{-1}$ are detected. Individuals can actively match the velocity of an imposed movement to within $\pm5^\circ\,s^{-1}$ over velocities from ~15° to $80^\circ\,s^{-1}$.

Testing Position Sense

Position sense also has a number of facets that can be tested in different ways. As noted, movements can be imposed so slowly that individuals do not

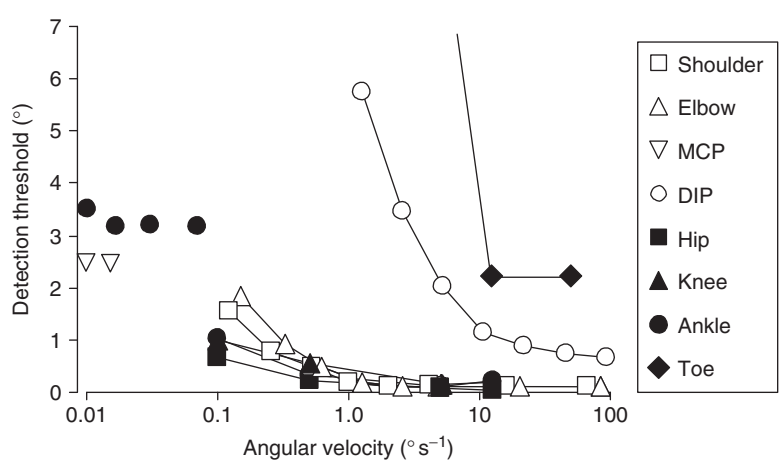

Figure 1 The magnitudes of movement required for detection of movements of different velocities are plotted for joints of the upper limb – shoulder, elbow, distal interphalangeal joint (DIP) and metacarpophalangeal joint (MCP) of the finger – and for joints of the lower limb – hip, knee, ankle, and the interphalangeal joint of the big toe (toe). Movement detection is better at faster velocities. If velocities are slow enough ($0.1^\circ\,s^{-1}$), detection is independent of velocity. Adapted from Refshauge KM, Chan R, Taylor JL, et al. (1995) Detection of movements imposed on human hip, knee, ankle and toe joints. *Journal of Physiology* 488: 231–241, with permission.

perceive them. With these very slow displacements, changes of position of ~2–5° of the ankle, shoulder, and finger are discriminated.

The ability to match joint positions is tested in a variety of ways. The simplest is for the person to move one limb to match the angle of a joint of the contralateral limb which was moved passively and is supported in position. Under such conditions, individuals can match joint angle within 5–10°. Some variants on this test include repositioning the same limb rather than matching with the other limb, using a pointer or indicator rather than using a matching limb, and giving a verbal estimate of angular position. Other variations focus on changes in the limb being matched – for example, varying the time between arrival at the position and matching, having the person move actively to position, or hold the position actively. When muscle activity is used to maintain position, accuracy is improved with errors of ~2°.

Another aspect of position sense is the ability to perceive the instantaneous position of the arm during an ongoing movement. During an imposed movement at the elbow at velocities of 70–200° s^{-1}, individuals can indicate with an accuracy of ±1° when the arm passes through a particular joint angle. This contrasts with the poorer performance in detection of imperceptibly slow movements, and in position-matching tasks.

Evidence for the Mechanisms Underlying Movement and Position Sense

It is generally accepted that sensory inputs from muscle, skin, and joints all contribute to movement and position sense, although belief in the relative importance of each contribution has varied over the past century. The evidence that supports the contribution from each type of afferent includes the findings that (1) the afferents fire in response to joint movement or joint position, (2) activation of the afferents leads to the perception of joint movement or altered joint position, and (3) elimination of the firing of the afferents impairs the ability to perceive joint movement or position.

Joint receptors The activity of joint afferents has been recorded from awake people through microneurography, in which a fine needle can record the activity from single nerve fibers. In general, the afferents tend to fire at the ends of joint range and do not give an unambiguous signal of joint position or the direction of movement. However, occasional afferents do fire throughout the whole range in a way related to position, and this may be sufficient to give a useful signal. With microneurography it is possible to stimulate a single nerve fiber electrically. For ~30% of joint afferents, this stimulation is perceived as small joint movements. With other afferents, individuals report pressure in the joint. When joint afferents are not active, either during testing after total joint replacements or with injection of local anesthetic into the joint, there is little impairment. Thus, although joint afferents contribute to perception of joint movement and position, other inputs can compensate for their loss.

Skin receptors Recording from skin afferents shows that the slowly adapting afferents which respond to stretch of the skin can fire with changes in joint angle. For example, afferents with receptive fields over the back of the hand increase firing with flexion of the metacarpophalangeal joints. Thus, these afferents can signal joint movement and position. Activation of skin afferents can produce perceptions of movement of the fingers and limb joints. If the skin is stretched on one side of a joint by pulling on tape stuck to the skin, many people report feeling movements about the joint. The illusion is enhanced if the skin on the other side of the joint is compressed at the same time. When input from the skin is removed by local anesthetic, perceptions of movement and position are impaired but not abolished at finger joints. However, impairment does not occur with anesthesia of a sleeve of skin around the knee. Thus, sensory input from the skin contributes to movement and position sense. It may be easier to compensate for its loss at the larger limb joints than in the hand.

Muscle spindles The muscle afferents associated with movement and position sense are those from the muscle spindles. These fire when the muscle spindle is stretched during lengthening of the muscle. Recording in humans and in animals shows that muscle spindles can signal joint movement and position during passive movements. The Ia afferents signal velocity (with a position component) whereas the group II afferents primarily signal position. Muscle spindle activity increases with muscle lengthening, but most spindles fall silent during passive shortening of the muscle. Thus, muscles on both sides of a joint are needed to give adequate signals of movement and position. Stimulation of a single muscle spindle afferent causes no sensation, but when a population of muscle spindles is activated by vibration applied over the tendon of the muscle, joint movement and altered joint position are perceived. Illusory movements are felt in the direction consistent with lengthening of the vibrated muscle. Higher frequency vibration (~100 Hz) generates a strong illusion of movement with a velocity related to the frequency of vibration. This illusion is attributed to discharge of Ia afferents. With lower frequency vibration, the

movement illusion is diminished but an illusion of altered position remains. This is attributed to group II afferent activity. The anatomy of muscle attachments to the phalanges allows the muscles to be disengaged functionally from the distal interphalangeal joint of the middle finger by positioning the hand with the middle finger fully flexed at the proximal interphalangeal joint and all the other fingers extended. Because no muscles act at the distal interphalangeal joint in this posture, movement and position sense can be tested without muscle spindle activity. The absence of muscle spindle firing impairs, but does not abolish, detection of joint movement and position.

During active movements, signals from joint and skin afferents are the same as during passive movements. Muscle spindle activity in the antagonist (non-contracting) muscle should also be unchanged. In contrast, the firing of muscle spindles in the active muscle depends on fusimotor drive, in addition to changes in muscle length. In humans, fusimotor drive to the muscle fibers inside the muscle spindles usually occurs concurrently with voluntary activation of the muscle. Therefore, during an isometric contraction, spindle firing increases from fusimotor drive even though the muscle is not stretched. During a shortening contraction, spindles will keep firing if the shortening is slow, but will fall silent if the shortening velocity of the muscle exceeds the ability of the intrafusal fibers to contract quickly. During a lengthening contraction, both fusimotor drive and muscle stretch will add to spindle discharge. It has been postulated that a signal related to fusimotor output must be used by the brain to interpret muscle spindle input in terms of joint movement and position during voluntary contraction.

Corollary discharge A final mechanism that may give rise to sensations of movement or position is a central signal related to motor output. Such a signal is known as a corollary discharge. This idea is well established for the eyes. Because the muscles of the eyes always have to move the same load, a set amount of drive to the muscles should always result in the same eye movement. That is, you should know which way your eyes are pointing because you told them to go there. Individuals' reports that the visual world moves when passive movements are imposed on the eyes are consistent with this proposal. In contrast, there is no fixed relationship between motor command and the resultant movement or position for the limb muscles, because they work under ever-changing loads. However, a recent study, in which individuals experienced illusions of altered joint position when they tried to move a paralyzed and anesthetized hand, suggests that a signal related to motor command may also have a role in position sense in the limbs.

Sensations of Force and Effort

The sense of muscle force or tension and the sense of effort are difficult to separate. One provides a perception of the absolute force being generated by the muscle. The other gives a perception of the proportion of total muscle strength being used. Both contribute to judgments of applied forces and of heaviness. The mechanisms underlying the sensations differ. The sense of muscle force is attributed to sensory input whereas the sense of effort is attributed to a central signal related to motor command.

Testing Force and Effort

Sensations of force are often tested by asking persons to make judgments about the weights of objects. The heavier an object, the more muscle force is required to move or lift it. People can discriminate between weights which differ by about 6%. This ability is independent of the muscle group which is used to support the weight. When individuals give numerical estimates of the magnitude of force or weight, perceived force doubles if exerted force is increased by 50%. However, if they lift a weight with one muscle group and then choose a matching weight with the contralateral muscles, the relationship between the two weights is linear, although lower weights (<40% maximal force) are overestimated and higher ones (>80% maximal force) are underestimated. The linearity of weight-matching suggests that the relationships between force production and perception are similar in the two muscle groups. People can match isometric forces in the same way as weights, with an average error of ~4% of maximal force. People can also scale forces based on their own effort. When individuals produce designated proportions of maximum effort, they produce more force than expected for weak efforts and less for strong efforts. This is consistent with the overestimation and underestimation in matching light and heavy weights.

Evidence for the Mechanisms Underlying Sensations of Force and Effort

The main sensory inputs that underlie the sense of muscle force are the Golgi tendon organs. Because each tendon organ is activated by the contraction of only a few muscle fibers, the firing of single tendon organs does not represent the force generated by the whole muscle. In contrast, the response from a population of tendon organs is a good measure of overall force, although a dynamic component to the

firing accentuates changes in force. It is not clear whether stimulation of tendon organs gives a perception of muscle force, although people do experience sensations of muscle force when muscle contractions are produced by electrical stimulation.

Sensory signals of muscle force and central signals of effort are both directly related to the force generated by a muscle. However, when a muscle becomes fatigued, it cannot produce as much force, and sensations of muscle force and effort are dissociated. It is a common experience to hold an object which gets heavier and heavier as the muscles holding it get tired. The increase in perceived heaviness despite the object's constant weight is an indication that the sensation is not conveyed by the tendon organs, which continue to give a true signal of muscle force. In this situation, progressive fatigue of the muscle fibers means that more of the muscle needs to be recruited to support the object. That is, the motor drive or command to the muscle is increased, and the sense of effort, which is related to the proportion of the muscle that is being used, also increases. Thus, the increase in heaviness of objects during fatigue suggests that the sense of effort has a large influence on perception of weight and force. This is confirmed when individuals match weights lifted by two muscle groups of different strength. Heavier weights lifted with the stronger muscles are matched to lighter weights lifted with the weaker muscles (Figure 2). However, the sense of effort is not the sole contributor to perceptions of heaviness. Although a weight supported by a fatigued muscle feels heavier, the increase in heaviness is not directly proportional to the reduction in force-generating capacity of the muscle. That is, if the maximal force of the muscle is reduced by half, the weight does not feel twice as heavy. This suggests that some other signal, presumably sensory input from tendon organs, also has a role. Similar results are reported when individuals' muscles are weakened by partial curarization.

Finally, in most judgments of force, tactile information is also important. Pressure-sensitive receptors are not truly proprioceptors, because they give information about the body's interaction with the environment rather than conveying sensations about the body itself. However, skin receptors can signal the magnitude and direction of forces. When muscle receptors and central signals of effort cannot contribute, such as when objects are placed on a hand which is resting on a table, people can discriminate changes in weight of ~30%. Although discrimination is improved when the object is supported actively, the elimination of tactile input alters force sensations even during active conditions.

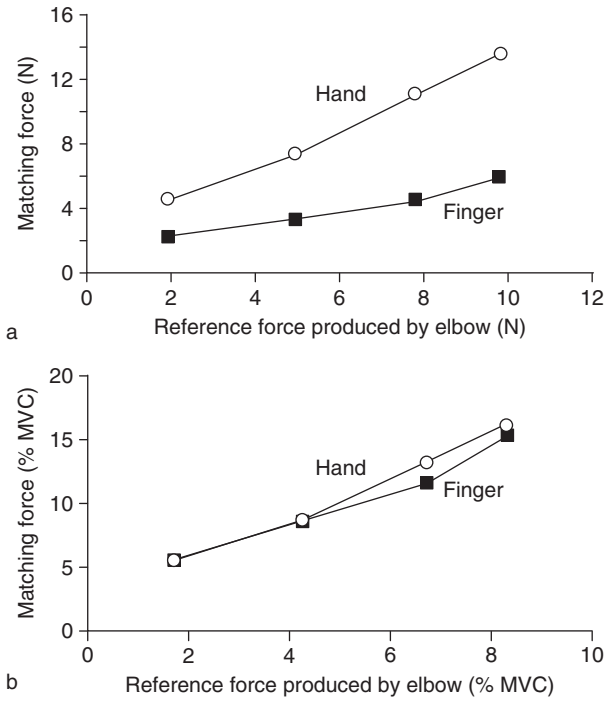

Figure 2 Individuals matched forces produced by elbow flexion (reference) with forces produced by a finger (■) or by a pinch grip of the hand (○). (a) The same force produced by the elbow was matched with a lower force produced by the finger than by the hand. As the finger could produce a lower maximal force, compared to the hand, the difference in matching is consistent with matching effort. (b) The same data as in panel (a) are shown expressed as percentages of maximal voluntary force (MVC). For a force produced by the elbow, individuals matched the same proportion of muscle force for the finger and hand. Note that these weak forces produced by the elbow (~2–8% MVC) were overestimated by both the finger and the hand, which matched with ~5–15% MVC. Reproduced from *Experimental Brain Research*, Vol. 151, 2003, 197–203, Perceptual constancy and the perceived magnitude of muscle forces, Jones LA, figures 2 and 3, Copyright Springer. With kind permission of Springer Science and Business Media.

Interactions of Senses

Some sensations require interactions between the proprioceptive senses. For example, heaviness of an object can only be gauged if the object can be moved. Movement signals the time when the perceived force is a measure of the object's heaviness. The perceptions of qualities such as stiffness, which is a combination of force and distance, and viscosity, which combines force and velocity, need continuous interaction of sensations of movement and force. The senses of movement, position, force, and effort can all contribute to the perception of limb position under different circumstances. If the limb is resting passively, only position sense is available to give a perception of joint angle. In contrast, if the limb is supported against gravity by active muscle contraction, then the muscle force

needed to hold the limb in position will also vary with joint angle. Thus, the sense of effort associated with the level of muscle contraction, as well as signals of muscle force from tendon organs, will be able to contribute (Figure 3). Finally, if the joint position was reached by a movement, then movement sense may also have a role.

Interactions between proprioception and other senses provide orientation in the environment. Perception of where the body is in space relies on the sensations of movement and position of the neck, trunk, and limbs, combined with sensory input from the vestibular apparatus, which signals movements of the head in space. If the head rotates 90° to the right, it might mean that the neck has turned, that the trunk has turned or that the whole body has turned. Sensations of movement and position from the neck and trunk allow us to differentiate these conditions which give similar vestibular signals. Sensations of position and movement of the eyes are essential for the perception of the location of visual objects relative to the head, with proprioceptive input from the neck needed to identify the location relative to the body. Similarly, the location of sounds relative to the body requires sensations of movement and position of the head on the neck.

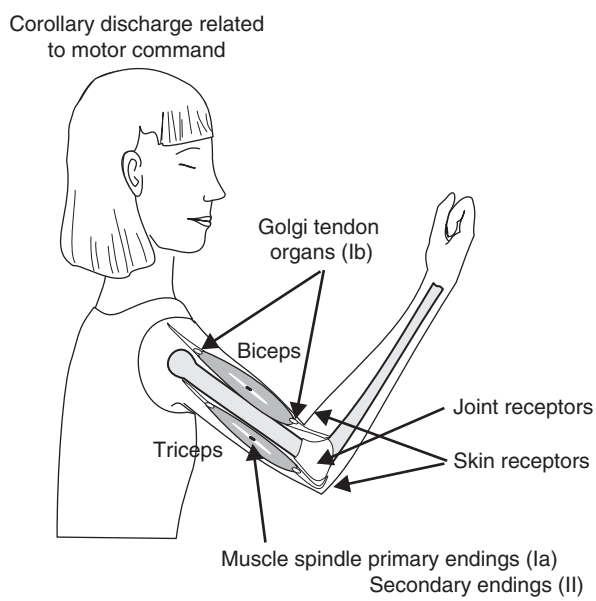

Figure 3 Diagram showing the proprioceptive signals available to contribute to the perception of elbow angle (limb position) while the forearm is supported against gravity by active contraction of biceps brachii. Sensory signals are available from muscle spindles and tendon organs in the agonist (biceps) and antagonist (triceps) muscles. Sensory input is also available from joint receptors and from skin receptors on both sides of the joint. A central signal related to the motor output (corollary discharge) is also available.

The contribution of proprioceptive sensations to awareness of the size and shape of the body is demonstrated by illusions that can be induced by tendon vibration. If vibration is applied to the biceps brachii tendon while a person touches the nose with the finger, the nose feels like it grows. As discussed previously, the vibration activates muscle spindles, which give the illusion that the arm is extending at the elbow (consistent with biceps lengthening). However, tactile sensations insist that the finger is still touching the nose. The brain's resolution to this dilemma is that the nose is getting longer. Thus, proprioceptive signals influence the body image.

Changes in Proprioception with Pathology

Proprioceptive function changes with normal events such as fatigue as well as with injury and disease. As discussed, muscle fatigue alters judgment of force and heaviness. It can also impair the detection of imposed movement and judgment of limb position when the limb is actively held against gravity.

Musculoskeletal injuries such as ankle sprains and knee ligament injuries are associated with impaired movement and position sense at the injured joint. Similarly, impaired proprioception is reported with arthritis and low back pain. Disturbance of proprioceptive input from the neck may lead to the dizziness and impaired balance in patients with whiplash. The mechanism of such impairments is not clear. Possibilities include changes in the periphery, including inflammation or altered tissue compliance, which could change receptor firing, and more central effects, such as interaction between pain and proprioceptive sensations.

Interruption of the neural pathway from the sensory receptors to the brain will also result in impairment of proprioception. This could happen in the periphery through neural injury or neuropathy, in the spinal cord with injury or diseases such as multiple sclerosis, and in the brain stem or at the cortex through injury or stroke. Very rarely, patients who have lost all large myelinated nerve fibers have no sensory input to provide proprioceptive signals and must rely on central signals related to muscle contraction and on vision to perform motor tasks successfully. They are unable to learn new motor skills.

Apart from lesions in the proprioceptive pathway to the cortex, movement disorders (such as Parkinson's disease, focal dystonia, Huntington's chorea) and cerebellar disorders have associated impairments of various proprioceptive sensations. For example, detection of passive movements and joint position is

impaired in Parkinson's disease, whereas in cerebellar disorders, the discrimination of velocity and duration of movements is impaired.

Contribution of Proprioception to Motor Control

Proprioception and motor control are so entwined that it is often impossible to separate where the sensory signals subserve a direct motor function and where the same signals evoke proprioceptive sensations which might then be used in control of movement. Proprioceptive signals contribute to movement at every level. Proprioceptive afferents have roles in reflex responses at the spinal and cortical levels and provide feedback that allows the control of all purposeful movements. We can attend to some of this feedback and report sensations related to it, but we more commonly employ it to guide our actions. Although we usually pay no attention to the current position of a hand before we use it to pick something up, we could, if asked, report the hand's location. Proprioception is essential to motor learning. New patterns of movement, such as driving a car or hitting a ball, cannot be learned in its absence. Without proprioceptive sensations there is no direct link between the motor output and its consequences.

Summary and Conclusions

Proprioception is the sense which allows perception of the location, movement, and action of parts of the body relative to each other. Proprioceptive sensations are derived from combined sensory input from muscle, joint, and skin receptors, plus central signals related to motor output. Sensations of movement and position and force and effort can be tested separately, but are used in conjunction with each other during many perceptual and motor tasks. Proprioception is so closely tied to the control of voluntary movement that impairments are more likely to be noticed as deficits in the performance of motor tasks than as impaired sensations.

See also: Evolution of Sensory Receptor Specializations in the Glabrous Skin; Mechanoreceptors; Motor Skill Learning; Multisensory Convergence and Integration; Neural Circuitry in the Somatosensory System; Neuromuscular Junction (NMJ): Presynaptic Stretch Effects on Neuromuscular Transmission; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Tactile Texture.

Further Reading

Bastian HC (1887) The 'muscular sense': Its nature and cortical localisation. *Brain* 10: 1–144.

Clark FJ, Burgess RC, Chapin JW, et al. (1985) Role of intramuscular receptors in the awareness of limb position. *Journal of Neurophysiology* 54: 1529–1540.

Collins DF and Prochazka A (1996) Movement illusions evoked by ensemble cutaneous input from the dorsum of the human hand. *Journal of Physiology* 496: 857–871.

Cordo P, Carlton L, Bevan L, et al. (1994) Proprioceptive coordination of movement sequences: Role of velocity and position information. *Journal of Neurophysiology* 71: 1848–1861.

Donaldson IM (2000) The functions of the proprioceptors of the eye muscles. *Philosophical Transactions of the Royal Society of London, Series B. Biological Sciences* 355(1404): 1685–1754.

Edin BB and Abbs JH (1991) Finger movement responses of cutaneous mechanoreceptors in the dorsal skin of the human hand. *Journal of Neurophysiology* 65: 657–670.

Gandevia SC (1996) Kinesthesia: Roles for afferent signals and motor commands. In: Rowell LB and Shepherd JT (eds.) *Handbook on Integration of Motor, Circulatory, Respiratory and Metabolic Control during Exercise*, pp. 128–172. Bethesda, MD: American Physiological Society.

Gandevia SC, Smith JL, Crawford M, et al. (2006) Motor commands contribute to human position sense. *Journal of Physiology* 571: 703–710.

Goodwin GM, McCloskey DI, and Matthews PBC (1972) The contribution of muscle afferents to kinaesthesia shown by vibration induced illusions of movement and by the effects of paralysing joint afferents. *Brain* 95: 705–748.

Jones LA (1986) Perception of force and weight: Theory and research. *Psychological Bulletin* 1000: 29–42.

Jones LA (2003) Perceptual constancy and the perceived magnitude of muscle forces. *Experimental Brain Research* 151: 197–203.

Lackner JR (1988) Some proprioceptive influences on the perceptual representation of body shape and orientation. *Brain* 111: 281–297.

Naito E, Roland PE, and Ehrsson HH (2002) I feel my hand moving: A new role of the primary motor cortex in somatic perception of limb movement. *Neuron* 36: 979–988.

Proske U (2005) What is the role of muscle receptors in proprioception? *Muscle & Nerve* 31: 780–787.

Refshauge KM, Chan R, Taylor JL, et al. (1995) Detection of movements imposed on human hip, knee, ankle and toe joints. *Journal of Physiology* 488: 231–241.

Prosopagnosia

D Tranel and N L Denburg, University of Iowa Hospitals and Clinics, Iowa City, IA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Prosopagnosia is a neurological term of Greek origin (Greek *prosop*, face, + a + Greek *gnosis*, to know), used to denote a condition in which patients lose the ability to recognize familiar faces. The disorder is confined to the visual realm, and the affected patient cannot arrive at the meaning of some or all previously known faces, despite normal or near-normal visual perception and intact alertness, attention, intelligence, and language. Most patients manifest a comparable defect in the anterograde compartment; that is, they cannot learn new faces despite adequate exposure. Prosopagnosia is usually caused by cerebrovascular disease, head injury, or cerebral tumors. In some instances, it can occur in the setting of a degenerative condition such as Alzheimer's disease, although in these cases it is usually part of more widespread neuropsychological dilapidation. Finally, prosopagnosia can arise during early childhood, with deficits presumably present from birth.

Prosopagnosia (or face agnosia) is the most frequent and well established of the visual agnosias. The phenomenon has been noted since the nineteenth century, and it has been the focus of considerable scientific inquiry. The fascination with prosopagnosia stems from the oddity of the disorder, along with the opportunity it provides to investigate perception, learning, and recall of complex and unique knowledge. From a clinical standpoint, the condition is also important, and given its devastating effect on its victims, accurate diagnosis and management are essential. Investigation of prosopagnosia has been marked by controversy regarding the relative importance of perceptual versus mnestic factors in the development of the condition and a dispute about whether prosopagnosia requires bilateral lesions or can be caused by unilateral damage.

Neuropsychological and Neuroanatomical Correlates of Prosopagnosia

The face recognition defect in prosopagnosia typically covers both the retrograde and anterograde compartments. Patients can no longer recognize the faces of previously known individuals and are unable to learn new ones. They are unable to recognize the faces of family members, close friends, and in the most paradigmatic instances, even their own face in a mirror. On seeing those faces, the patients experience no sense of familiarity, no inkling that those faces are known to them; that is, they fail to conjure up consciously any pertinent information that would trigger recognition. However, the recognition disorder is confined to the visual modality, and recognition via other modalities (e.g., hearing recognizable voices) is unaffected.

As with other forms of agnosia, prosopagnosia must be distinguished from disorders of naming, that is, it is not an inability to name faces of persons who are otherwise recognized as familiar. There are numerous examples of face naming failure, from both brain-injured populations and the realm of normal everyday experience, but in such instances, the unnamed face is invariably detected as familiar, and the precise identity of the possessor of the face is usually apprehended accurately. In prosopagnosia, however, the defect sets in at the level of recognition. Needless to say, the patients will also manifest a face naming impairment (they will not name faces they cannot recognize), but this is artifactual and not a true naming defect. Several major types of prosopagnosia can be distinguished.

Pure Associative Prosopagnosia

In this variety, the recognition impairment is relatively pure in the sense that it is confined to the visual modality and occurs in the setting of normal or near-normal visual perception. Associative prosopagnosia largely conforms to the strict definition of agnosia, that is, "a normal percept stripped of its meaning" (Teuber's classic definition of agnosia; Teuber, 1968: 274–328). The patients perform normally on standard neuropsychological tests of visuoperceptual discrimination and visuospatial judgment. As noted, recognition via other sensory modalities is unaffected, and on hearing the voices of individuals whose faces go unrecognized, the patients will instantly recognize the identities of those individuals. Even within the visual modality, the defect is highly circumscribed. For instance, patients may be able to recognize individuals on the basis of a distinctive feature (e.g., hairstyle) or gait or posture.

Most patients with pure associative prosopagnosia also have achromatopsia, an acquired impairment of color perception. The combination of the defects is due to the contiguity of the neural processing systems for form and color and their sweeping damage by one lesion. The color perception defect *per se*, however, does not account for the face recognition impairment, and normal individuals can easily recognize faces in black-and-white. The ability to read may or may not

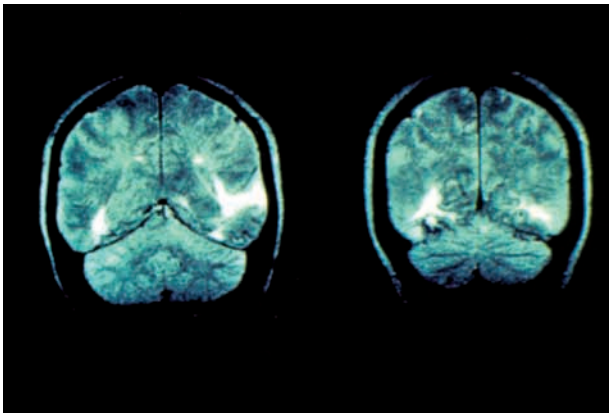

Figure 1 Illustration of the neuroanatomic findings of a patient with pure associative prosopagnosia.

be affected in prosopagnosia patients, depending on the location of the lesion in the left occipital region. When the lesion encompasses both the left occipito-temporal region and the left periventricular region (the white matter beside, beneath, and behind the occipital horn), reading impairment (alexia) coexists with prosopagnosia.

Pure associative prosopagnosia is caused by bilateral damage in inferior occipital and temporal visual association cortices, that is, in the inferior component of cytoarchitectonic areas 18 and 19 and part of the nearby cytoarchitectonic area 37 (see Figures 1 and 2 for an illustration of the neuroanatomic findings in such a patient). Most cases are due to cerebral infarctions caused by occlusion in posterior cerebral artery branches. Head injury and cerebral tumors, especially gliomas originating in one occipital lobe and traversing into the opposite hemisphere via the splenium of the corpus callosum, can also produce prosopagnosia. Prosopagnosia in connection with lesions (bilateral or unilateral) located exclusively above the calcarine fissure, in superior visual association cortices, has never been reported, although there are cases of face agnosia in which on one or both sides, the inferior lesions extend upward above the calcarine fissure.

Apperceptive Prosopagnosia

Face agnosia in the setting of significant visuoperceptual disturbance has been termed 'apperceptive' prosopagnosia. Apperceptive face agnosics have defects in basic visual perception, demonstrable on neuropsychological testing, which compromise abilities such as matching of unfamiliar faces, judgment of line orientations, and mental manipulation of pictures and picture fragments. It should be noted, however, that such defects are not so severe as to compromise most aspects of basic form vision; thus, the condition conforms in a broad sense to the definition of agnosia. As in other forms of recognition, perceptual processes on the one hand and recognition–recall processes on the other cannot be rigidly compartmentalized. Even in 'pure' cases of associative prosopagnosia, there may be fairly subtle, albeit important, disturbances of high-level integrative perceptual abilities which cannot be detected by available probes. Also, any explanation of face agnosia tied to perceptual factors must reconcile the fact that most patients with visual perceptual defects, even severe ones, do not lose their ability to recognize faces. That is, visuoperceptual disturbance as detected by neuropsychological probes does not necessarily cause face agnosia. For these reasons, it is useful to maintain a distinction between associative and apperceptive prosopagnosia because generally it is possible to determine whether the condition is largely due to memory defects (associative) or to high-level perceptual defects (apperceptive).

Apperceptive face agnosia is most often associated with bilateral damage to 'early' visual association cortices or with damage in right visual association cortices within the occipital and parietal regions. It appears that the damage must involve both the inferior and superior components of posterior visual association cortices (areas 18 and 19), mesially and laterally, for severe and lasting face agnosia to develop. In most cases, parts of areas 39 and 37 on the right will also be damaged. Figure 3 illustrates the neuroanatomic findings in such a patient.

Developmental Prosopagnosia

Some individuals never develop a normal capacity for learning faces. They have a lifelong deficiency in learning and recognizing faces that ought to have been readily mastered, and since the problem begins in childhood, it is appropriate to call it 'developmental' prosopagnosia. Developmental prosopagnosia has not been widely reported or studied, but it may be far more frequent than suspected, its apparent rarity being because affected persons tend to conceal their disability. It is probably associated with learning disability for other classes of visual stimuli which, like faces, require individual identification and have many similar exemplars. Neural correlates for this condition are in the early stages of being identified, and the findings point to the potentially important role of the midfusiform gyrus and the inferior occipital gyrus. The youngest documented case of developmental prosopagnosia was recently reported. Briefly, patient AT was studied at approximately 5 years of age. AT has both a negative medical history and a negative MRI of the brain. Intellect falls in the very superior range, and he came to medical attention during preschool because of problems recognizing individuals who were well known to him. AT's difficulties are mainly

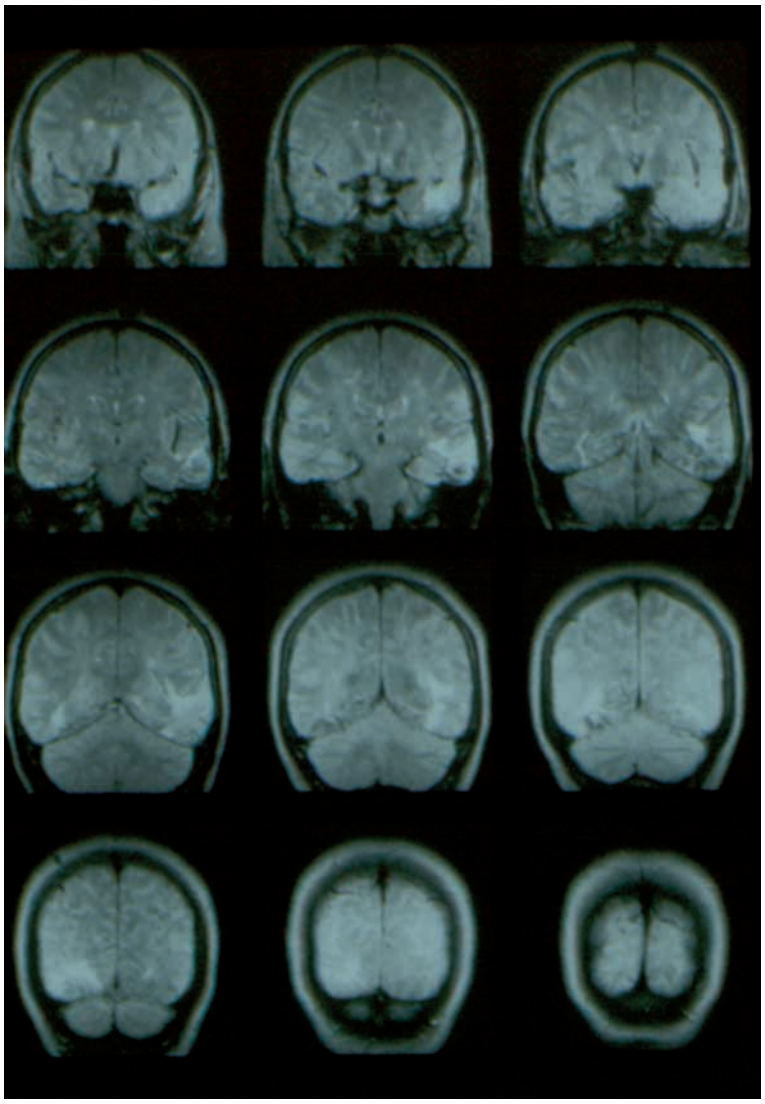

Figure 2 Illustration of the neuroanatomic findings of a patient with pure associative prosopagnosia.

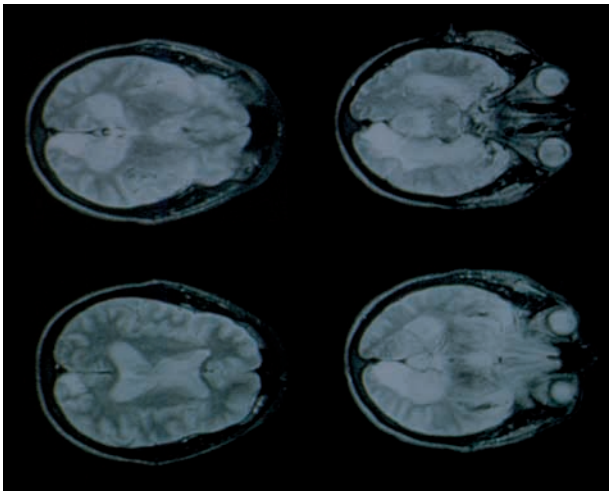

Figure 3 Illustration of the neuroanatomic findings of a patient with apperceptive face agnosia.

associative in nature, although he does manifest aspects of apperceptive prosopagnosia, as demonstrated by defects in certain aspects of visuoperception. Interesting parallels with the experimental research in acquired prosopagnosia in adults are evident in this case and are noted in the section titled 'Nonconscious discrimination of familiar faces.'

A recent study identified a subgroup of patients with Asperger's disorder who manifested prosopagnosic-like defects, suggesting the possibility of occipitotemporal dysfunction in patients with social developmental disorders.

Other Aspects of Face Processing

Nature and Extent of the Defect

In prosopagnosia, the recognition impairment occurs at the most subordinate taxonomic level, that is, at

the level of identification of unique faces. Prosopagnosics are fully capable of recognizing faces as faces; that is, performance is normal at the superordinate taxonomic level. Also, most prosopagnosics can recognize facial expressions, and can make accurate determinations of gender and age based on face information. These dissociations highlight the fact that recognizing faces at the level of unique identity is a highly demanding task which requires the brain to distinguish between numerous exemplars that bear a high degree of resemblance to one another.

Although face agnosics recognize any number of visual entities at basic object level, such as cars as cars, buildings as buildings, and dogs as dogs, they often fail to recognize these items at the subordinate level of unique identity. Thus, similar to the problem with faces, they are unable to recognize the specific identity of a particular car or building. These recognition impairments are common in prosopagnosia and underscore the notion that the core defect in face agnosia is the inability to disambiguate fully individual visual stimuli. In fact, cases have been reported in which the most troubling problem for the patient was in classes of visual stimuli other than human faces, such as a farmer who lost his ability to recognize individual dairy (Holstein) cows or a birdwatcher who became unable to tell apart various subtypes of birds.

Prosopagnosic patients can often recognize identity from movement. This means not only that their perception of movement is intact but also that they can evoke appropriate memories from the perception of a unique pattern of movement. Combined with the finding that recovery of identity from motion can be impaired by lesions in superior occipitoparietal regions, these findings underscore the separable functions of the 'dorsal' and 'ventral' visual systems, the dorsal one being specialized for spatial placement, movement, and other 'where' capacities and the ventral one being specialized for form detection, shape recognition, and other 'what' capacities.

Nonconscious Discrimination of Familiar Faces

Some prosopagnosic patients show accurate covert or nonconscious discrimination of familiar faces despite their complete inability to recognize those faces at overt level. For example, it has been shown that prosopagnosics generate large, discriminatory electrodermal responses to familiar faces that are otherwise unrecognized. Preserved covert face discrimination has been demonstrated in other experimental paradigms, such as reaction time tasks and forced choice procedures. In the electrodermal paradigm, covert face discrimination has even been demonstrated for faces from the anterograde compartment, indicating that the brain can continue to learn new visual information even without conscious influence. In the earlier mentioned developmental case, it was found that AT's covert face discrimination based on an autonomic index was normal, suggesting that the brain is capable of acquiring some information about familiar faces, even without conscious recognition.

Recovery

Prosopagnosic patients often become quite adroit at using nonface information to recognize the persons around them. For example, they rely on voice, or gait, or a distinctive visual feature. Regarding the latter, it is often helpful to add a distinctive feature in situations in which recognition is highly demanded but difficult. At a crowded social gathering, for example, the patient's spouse might wear a special hat or other article of clothing, which would facilitate rapid and accurate identification by the patient. There is also evidence to suggest that the prosopagnosic individual may benefit from a local, components-based examination of the face rather than a holistic, configural approach. Generally speaking, though, the earlier in the course of recovery that compensatory strategies such as these can be taught, the better the chances for healthy adaptation to the disability.

Unanswered Questions

An issue which has remained central in the study of prosopagnosia is the mechanism of the defect and in particular, the reason the recognition of faces is affected so much more severely than the recognition of other visual stimuli. It has been proposed that faces, more so than many other types of items, have two important qualities: First, they are highly similar in terms of physical structure, that is, they have high visual resemblance or what can be termed homomorphism. Second, they are numerous, and it is necessary to tell them apart at the unique, individual level. It does not suffice, for example, to recognize a good friend as simply a female person or as a female person who works at the hospital. Rather, recognition must take place at the level of unique identity. Together, these factors make faces a 'visually crowded' category, which may explain the vulnerability of familiar face recognition to acquired brain disease.

Other questions remaining in the study of prosopagnosia concern the reasons some prosopagnosic patients are able to generate covert responses to familiar faces while other patients cannot. A related question concerns the relationship among different indices of covert discrimination – for example, are skin conductance responses, reaction times, and eye movement patterns all tapping into the same discrimination mechanisms?

On another point, much still remains unknown about the neural mechanisms of developmental prosopagnosia. Finally, a recent finding regarding covert discrimination of faces in the setting of 'childhood amnesia' raises many interesting questions. Specifically, it has been shown that when school-age children (e.g., ages 10–15) are shown faces of former classmates who were known only in the early developmental years for which much information is not readily retrievable into declarative memory (especially ages 2–5), the children generate discriminatory skin conductance responses even for faces that are entirely 'forgotten' by overt measures. This finding warrants further investigation.

See also: Agnosia; Recognition Memory; Representation of Color.

Further Reading

Barton JJS, Cherkasova MV, Hefter R, Cox TA, O'Connor M, and Manoach DS (2004) Are patients with social developmental disorders prosopagnosic? Perceptual heterogeneity in the Asperger and socio-emotional processing disorders. *Brain* 127: 1706–1716.

Baudouin J-Y and Humphreys GW (2006) Compensatory strategies in processing facial emotions: Evidence from prosopagnosia. *Neuropsychologia* 44: 1361–1369.

Behrmann M and Avidan G (2006) Congenital prosopagnosia: Face-blind from birth. *Trends in Cognitive Sciences* 9: 180–187.

Benton AL (1990) Facial recognition. *Cortex* 26: 491–499.

Damasio AR, Damasio H, and Van Hoesen GW (1982) Prosopagnosia: Anatomic basis and behavioral mechanisms. *Neurology* 32: 331–341.

Damasio AR, Tranel D, and Damasio H (1990) Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience* 13: 89–109.

De Renzi E (1986) Prosopagnosia in two patients with CT scan evidence of damage confined to the right hemisphere. *Neuropsychologia* 24: 385–389.

Hadjikhani N and de Gelder B (2002) Neural basis of prosopagnosia: An fMRI study. *Human Brain Mapping* 16: 176–182.

Jones RD and Tranel D (2001) Severe developmental prosopagnosia in a child with superior intellect. *Journal of Clinical and Experimental Neuropsychology* 23: 265–273.

Landis T, Cummings JL, Christen L, Bogen JE, and Imhof H-G (1986) Are unilateral right posterior cerebral lesions sufficient to cause prosopagnosia? Clinical and radiological findings in six additional patients. *Cortex* 22: 243–252.

Meadows JC (1974) The anatomical basis of prosopagnosia. *Journal of Neurology, Neurosurgery, and Psychiatry* 37: 489–501.

Sergent J and Villemure J-G (1989) Prosopagnosia in a right hemispherectomized patient. *Brain* 112: 975–995.

Teuber HL (1968) Alteration of perception and memory in man: Reflections on methods. In: Weiskrantz L (ed.) *Annals of Behavior Change*, pp. 274–328. New York: Harper & Row.

Tranel D and Damasio AR (1985) Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics. *Science* 228: 1453–1454.

Tranel D, Damasio AR, and Damasio H (1988) Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology* 38: 690–696.

Tranel D, Damasio H, and Damasio AR (1995) Double dissociation between overt and covert face recognition. *Journal of Cognitive Neuroscience* 7: 425–432.

Proteasome Role in Neurodegeneration

L Korhonen, N Putkonen, and D Lindholm,
Minerva Foundation Institute for Medical Research, Helsinki, Finland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Recent studies have shown the critical role of the proteasome system in different human neurological disorders. Herein, we review the current knowledge of the role of the proteasome in neuronal and axonal degeneration and discuss different ubiquitin E3 ligases, deubiquitinating enzymes (DUBs), and signaling pathways controlling the activity of the proteasome in these disorders.

The Proteasome

Proteasomes are the main proteolytic machinery in cells and exhibit trypsin-like, chymotrypsin-like, and peptidyl-glutamyl peptide hydrolyzing enzymatic activity. The proteasomes are present as 26S and 20S particles in the cell. The 26S proteasome consists of a catalytic 20S core and two 19S regulatory subunits (19RS) at both ends of the 20S subunit forming a barrel-like structure. The 19RS consists of several proteins that mediate substrate recognition and delivery to the core complex for proteolytic degradation into short peptides. The 19RS also removes ubiquitin from proteins ensuring optimal levels of free ubiquitin in the cell. The 20S proteasome cleaves peptides and unfolded proteins without the regulatory unit. Apart from protein degradation, additional nonproteolytic functions have been ascribed to the proteasome. Thus, it was shown in *Saccharomyces cerevisiae* that the 19RS regulates gene transcription through an interaction with the activator protein, Gal4. In addition, 19RS promotes the Spt-Ada-Gcn5-acetyltransferase (SAGA) histone acetyltransferase complex to target promoters during gene activation.

In neurons as well as in other cells, the proteasomes are present in the nucleus, in the cell soma, as well as in the dendrites and the axons. Proteins destined for proteasomal degradation are marked with a chain of ubiquitin molecules in a three-step reaction catalyzed by the E1, E2, and E3 enzymes (see Figure 1). In mammalian cells, there is one E1 and a couple of E2's, but many more E3 ubiquitin ligases that can be divided into four major classes: the homologous to E6-AP C-terminus (HECT), Really Interesting New Gene (RING), plant-homeodomain (PHD), and UFD2-homology domain (U-box) families. The HECT domain ligases share a conserved C-terminal catalytic domain that forms a thioester bond with ubiquitin during the transfer from E2 to the substrate. The N-terminal regions of HECT domain ligases are involved in substrate specificity and recognition. The RING-finger ubiquitin ligases exist either as multisubunit complexes or as monomers and can interact directly with the E2 and the protein substrate. The PHD and U-box domain containing proteins, as well as the RING-finger E3 ligases, facilitate ubiquitin conjugation by acting as a bridge between the E2's and the protein substrates. In addition to the E3 ligases, the ubiquitin chain elongation factor E4 is involved in the efficient elongation of polyubiquitin chains. These proteins contain a protein motif known as the U-box and may act in concert with the E3 ligases. The C-terminus of Hsc0 interacting protein (CHIP) is a U-box containing protein important for the degradation of misfolded proteins, such as the Pael receptor in PD (see below).

Ubiquitin protein has seven lysine sites to which ubiquitin can be enzymatically linked. Attachment of multiple ubiquitin molecules to lysine 48 (called K48 ubiquitinylation) mediates mainly protein degradation, whereas K63 ubiquitinylation is involved in endocytosis, DNA repair, and in cell signaling. However, proteins can also become monoubiquitinated at different lysine sites as shown, for example, for receptor tyrosine kinases (RTKs), such as the epidermal growth factor receptor (EGFR) (see Figure 1).

The levels of ubiquitin are tightly regulated in the cells, among others by the action of the DUBs that remove and recycle ubiquitin from ubiquitin-conjugated proteins (see Figure 1). There are approximately 80 DUBs in the human genome that can be divided into five classes based on the ubiquitin protease domain: ubiquitin-specific proteases (USPs), ubiquitin C-terminal hydrolases (UCHs), Otubain proteases (OTUs), and Machado–Joseph disease proteases (MJDs) and JAMM (JAB1/MBN/Mov34). The latter is a metalloprotease whereas the others are cysteine proteases.

During neuronal differentiation the levels of ubiquitin and its conjugates, as well as protein ubiquitination and the activity of the proteasome, all increase suggesting an important function of these processes during development. In the mature nervous system, mutations in different components of the ubiquitin–proteasome system have been linked to the process of neurodegeneration, occurring in many human neurological disorders.

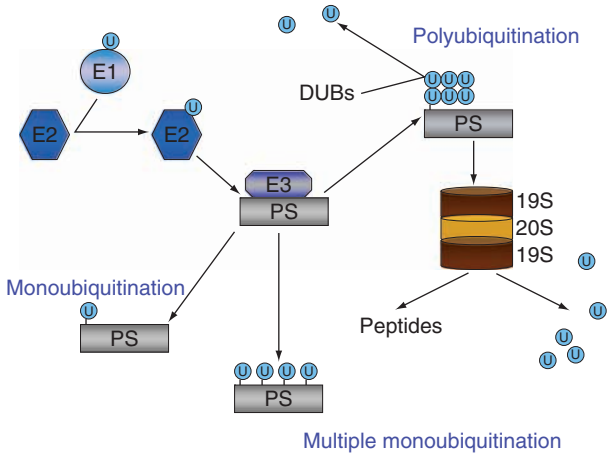

Figure 1 Schematic view of the function of the ubiquitin–proteasome system (UPS) in protein turnover. Cellular proteins are ubiquitinated by the action of various enzymes acting in concert to add either poly- or monoubiquitinated species to the substrates. Polyubiquitinated proteins are usually doomed for degradation, while monoubiquitinated ones can have other destinies. The regulation of the UPS is complex and dysfunctions may occur in various human diseases. These aspects of the UPS are covered in great detail in recent reviews. U, ubiquitin; E1, ubiquitin-activating enzyme; E2, ubiquitin-conjugating enzyme; E3, ubiquitin ligase; 20S, catalytic core; 19S, regulatory particle; PS, protein substrate; DUB, deubiquitinating enzyme.

Neuronal Degeneration and the Proteasome

Neurodegenerative diseases, such as Alzheimer's, Parkinson's, and Hungtinton's diseases (AD, PD, and HD), are characterized by the selective loss of neurons in specific brain areas. Distinct insoluble protein aggregates and inclusion bodies are found in the affected brain areas in form of tau-positive neurofibrillary tangles and amyloid β (Aβ) peptide in the plaques in AD, α-synuclein containing Lewy bodies in PD, and intranuclear inclusions of polyQ proteins in HD. The deposits contain misfolded proteins, molecular chaperones, and various components of the ubiquitin–proteasome system, such as ubiquitin, various E3 ligases, the 19RS and the 20S proteasome subunit. The accumulation of misfolded proteins in neurons can disturb cell metabolism in different ways and also cause dysfunction of the proteasome.

Normally in the cell, ubiquitin is generated from the ubiquitin precursor protein consisting of tandem ubiquitin molecules that are cleaved into monomeric ubiquitin by the UCH. However, a mutant form of ubiquitin (Ubiquitin^{+1}) has been identified in some neurodegenerative diseases. The ubiquitin^{+1} is generated by a dinucleotide deletion in the mRNA encoding for the ubiquitin B precursor, resulting in a +1 frame shift in the C-terminus of the first encoded ubiquitin molecule. This molecular misreading leads to a 19-amino-acid-long extension of the protein that is resistant to the cleavage by the UCHL. Ubiquitin^{+1} in high concentrations accumulates in neurons and glial cells and can inhibit the proteasome. Studies with human autopsy material and transgenic *ubiquitin*$^{+1}$ mice crossed with transgenic AD and other mouse models have shown that aberrant ubiquitin^{+1} accumulates in AD and in some polyglutamine diseases. Interestingly, ubiquitin^{+1} is not detected in synucleinopathies, such as in Lewy body disease and in multiple system atrophy.

Decreased proteasome activity has been reported in brain autopsy material from PD, AD, and HD patients. In sporadic PD, the 20S proteasome activity is impaired and various components, such as the 19RS and PA28 protein regulator, are expressed at lower levels compared to the controls. In addition, proteasome activity is decreased in neuronal cells after expression of the mutant forms of α-synuclein and of other disease-causing proteins. Recent developments in fluorescent-based reporter systems monitoring proteasome activity *in vivo* has provided further evidence for an impaired proteasome function in neurodegenerative diseases. Treatment with chemical compounds inhibiting the proteasome have been reported to induce degeneration of dopaminergic neurons and cause PD-like symptoms in rats, as well as decrease the viability of nerve cells expressing the mutant amyloid precursor protein. In contrast, the proteasome inhibitors are shown to be partly neuroprotective in ischemic brain injuries, showing a complex relationship between the control of neuronal death and survival and the function of the proteasome.

Direct evidence for the role of the proteasome in neurodegeneration comes from studies of some familial and genetic forms of diseases. In PD, deletions and point mutations have been found in the protein Parkin in approximately 50% of patients with the autosomal recessive form of PD (AR-JP). Parkin is a 465-amino-acid-long protein with an N-terminal UBL domain and two C-terminal RING motifs that can function as an E3 ligase. Parkin is a component of the SCF-like E3 complex together with the F-box/WD repeats protein hSel-10 and Cullin. Mutations in Parkin are thought to reduce its ability to degrade substrates and decrease the overall proteasome activity in cells. Several substrates have also been identified for Parkin, such as the Parkin-associated endothelial-like (Pael) receptor, synphilin-1, and the Cell Division Control related protein (CDCrel-1). Pael-1 is a putative G-protein-coupled transmembrane polypeptide that can trigger the unfolded protein response in the endoplasmic reticulum, and Parkin can ubiquitinate and promote degradation of the insoluble receptor.

Synphilin-1 has an unknown function but is associated with α-synuclein. CDCrel-1 is a member in the serpin family and has a role in synaptic vesicle dynamics. The mechanism by which loss-of-function of Parkin mutations causes PD is currently under intensive investigation.

Like Parkin, Dorfin is an E3 ligase with the ability to protect motor neurons against degeneration caused by mutant human copper/zinc superoxide dismutase 1 (SOD1) protein in amyotrophic lateral sclerosis (ALS). Dorfin ubiquitinates mutant SOD1 and reduces the occurrence of inclusion bodies of this protein in neurons. Dorfin also interacts with the AAA-ATPase valosine-containing protein (VCP) that can bind both the proteasome and polyubiquitinated proteins. The role of such shuttle proteins in different neurodegenerative diseases warrants more studies. Similarly, the precise role of Dorfin in sporadic ALS that represents the major form of the disease remains to be clarified.

For other diseases, it was shown that the ubiquitin E3 ligase, Ube3A, is mutated in Angelman's syndrome, characterized by mental retardation. Ube3A is crucial for long-term potentiation in mouse hippocampus, suggesting important protein targets in the nerve cells. In addition, it is known that anti-apoptotic proteins, such as the X chromosome-linked inhibitor of apoptosis protein (XIAP), exhibit ubiquitin ligase activity. XIAP is highly expressed in the nervous system and inhibits various caspases and further facilitates ubiquitinylation and degradation of pro-apoptotic proteins, such as Smac/DIABLO and Omi//HrtA2 released from the mitochondria during cell death. XIAP is an important regulator of neuronal cell death as shown in excitotoxicity, ischemia, and after nerve cell axotomy.

Synaptic Dysfunction and the Proteasome

Apart from the ubiquitin E3 ligases, decreased activity of some DUBs has been linked to neuronal dysfunction. Usp14 is a USP acting on ubiquitinated target proteins, and mutations in the *usp14* gene cause ataxia in mice. This is the result of defects in synaptic transmission observed both at the neuromuscular junction and in hippocampal CA3–CA1 synapses, suggesting a presynaptic action of the Usp14 protein. Many of the DUBs are involved in the regulation of protein trafficking and endocytic pathways and are therefore expected to have important roles in vesicle transport and in the control of neuronal viability. The precise functions of the different DUBs in the nervous system and in neurological diseases remain to be studied.

Synapses undergo large structural changes during maturation and in response to neuronal activity. Recent studies have further shown that the proteasome is involved in the degradation of molecules belonging to the postsynaptic densities (PSDs) that contain various receptors and scaffolding proteins. In addition, synaptic activity influences the ubiquitination and turnover of a subset of PSD proteins, important for the control of synapse function, and maintenance. As shown by studies in models, systems such as *Aplysia* and *Drosophila* have shown that protein ubiquitination can influence synaptic development and plasticity. In the fruit fly, mutations in the *fat facets* (*faf*) gene, encoding a DUB enzyme, cause severe presynaptic overgrowth and defects in neurotransmitter release. Faf most probably acts by stabilizing the *liquid facets* (*laf*, *Drosophila* epsin) gene product, thereby influencing endocytosis and cell receptor signaling. Recent studies on the neuromuscular system in *Drosophila* revealed a relatively short time window for the degradation of synaptic proteins by the proteasome. Chemical compounds inhibiting the proteasome, as well as genetic manipulations in the fly, subsequently identified the *Drosophila* UNC protein, DUNC-13, as a selective target for the proteasome in the presynaptic terminal. Data showed that proteasome inhibition locally increases neurotransmitter release and the levels of DUNC-13, and enhances synaptic efficacy. Members of the UNC protein family are involved in synaptic vesicle priming and regulation of neurotransmitter release in different species, suggesting similar mechanism for the control of synaptic activity by the proteasome.

As indicated above, monoubiquitination regulates protein trafficking, involving endosomes and clathrin-mediated endocytosis. RTKs, such as the EGFR, are monoubiquitinated upon ligand binding. This leads to RTK internalization and degradation in lysosomes, thereby preventing receptor recycling to the plasma membrane. Similarly, several other receptors such as the neurotrophic factor receptors, TrkA and TrkB, are ubiquitinylated after receptor activation. Very recently, it was shown that the accumulation of mutant Aβ peptide disturbs multivesicular body sorting by inhibiting the ubiquitin–proteasome system and thereby causes defects in receptor cycling. Current data suggest that receptor activation triggered by ubiquitinylation may serve as an important control step for signals mediating neuronal survival and degeneration.

Apart from this, several signaling pathways in the nervous system are regulated by the proteasome. In this context, the activation of the nuclear factor-kappa B (NF-κB) pathway after proteasome degradation of the inhibitor-κB protein mainly mediates neuronal survival. The nature of specific genes regulated by

the different NF-κB proteins in brain tissue, however, is only partly understood. In addition, the activities of the c-Jun-N-terminal kinase (JNK)-pathway can, depending on the cellular context, either increase nerve cell death or promote survival. The c-Jun-specific E3 ligase, Fbw7, is required for proper degradation of this protein, and mice deficient in Fbw7 exhibit tremor and loss of specific neuronal subtypes in the brain. The precise roles of the proteasome in controlling various signaling proteins and cascades in different neuronal populations afflicted by neurodegeneration need to be studied more in the future.

Axonal Degeneration and the Proteasome

Apart from changes in the cell soma, various neurological diseases are often characterized by early disturbances in neuronal connectivity and in axonal functions. Particularly in AD, in ALS, and in multiple sclerosis, axons degenerate that leads to changes in neuronal signaling and cell communication. The axon and the synapse can also undergo degeneration due to genetic causes or disturbed metabolism by various chemicals and toxins, as well as following different types of nerve injuries. In these conditions, the neuronal soma will ultimately also degenerate by mechanisms involving disturbed axonal transport and signaling leading to a 'dying-back' disorder that follows the degeneration of the axon.

One model to study axonal reactions has been mechanical transections of a peripheral nerve in rodents that leads to a degeneration of the distal nerve stump. In these conditions, characteristic changes occur in the cellular composition and molecular interactions in the nerve following the injury, described under the term Wallerian degeneration. Studies of the mouse mutant (*WLd*^{s}) with a delay in Wallerian degeneration for about 2–3 weeks has helped to identify a gene product on chromosome 4 that is important for preservation of axonal integrity following the injury. The unique 85 kDa fusion protein in the *WLd*^{s} mice is produced by gene rearrangements and consists of 70-amino-acid-long N-terminal portion of the ubiquitination factor E4b (UbE4b) connected by an 18-amino-acid stretch to the full-length nicotinamide mononucleotide adenylyl transferase (NMNAT) gene. Expression of the fusion protein in transgenic mice and using different vectors has been shown to delay axonal degeneration after traumatic and toxic injuries, as well as those due to genetic mutations.

The mechanisms by which WLd^s protein affords neuroprotection and delays axonal degeneration has been intensively studied using both *in vitro* neuronal systems and transgenic mice. The *WLd*^{s} mice protein in fusion gene is involved in the synthesis of nicotinamide adenine dinucletoide (NAD) that takes part in different cellular reactions. Increased levels of NAD have been shown to be neuroprotective acting either locally in the axons or indirectly in the nucleus by enhanced activity of the NAD-dependent deacetylase, SIRT1. SIRT1 in turn may induce transcription of genes involved in neuroprotection that then acts on the axons, and the effect of NAD can be blocked or enhanced using chemical compounds acting on SIRT1. In line with the presumed effect on gene transcription is the observation that the WLd^s fusion protein is abundantly expressed in nuclei. However, studies on SIRT1 gene deleted mice as well as others suggest that other mechanisms and genes may also be activated in the *WLd*^{s} mice.

Recent findings using transgenic mice indicate that apart from the NMNAT gene the N-terminal region in the *WLd*^{s} gene is also important for counteracting axonal degeneration. The truncated UbE4b protein in this region is part of an E4 enzyme that can function in polyubiquitination and proteasome-mediated degradation of proteins. It may be that for full neuroprotection the whole molecule of WLd^s is required. In the future, it is important to define the precise mechanisms and downstream gene products by which the WLd^s protein can counteract axonal degeneration and increase neuronal survival in different systems.

Results obtained with the crush lesions in the optic nerve in rodents and with injured cultured neurons have shown that inhibition of the proteasome can retard the onset of axonal degeneration and specifically inhibit the fragmentation of microtubule that occurred early after the injury. Preserved axonal integrity was also observed after the expression of an ubiquitin protease that can reverse ubiquitination of protein substrates. These studies thus show that the local activity of proteasome is involved in axonal degeneration and certain drugs targeting this complex can inhibit this. Studies on the Gracile axonal dystrophy (*Gad*) mutant mice support further the involvement of the proteasome in axonal degeneration. Dysfunctional UCH-L1 leads to the degeneration of the gracile tract of the spinal cord and this protein can act both as a DUB and as an ubiquitin ligase. The exact mechanism causing axonal degeneration in the *Gad* mice is however so far unknown.

Studies in *Drosophila* have shown that the proteasome activity is involved in axonal degeneration of occurring during development. In axonal pruning, neuronal connections in the nervous system are refined and mature to meet requirements of functional activity. Interestingly, expression of an ubiquitin

protease or a mutant form of the *Drosophila* E1 ubiquitin-activating enzyme, in the mushroom bodies, was found to inhibit pruning in the fly. This was also observed with mutations in two of the subunits in the 19S regulatory particle. These results indicate a crucial role for local proteasome activity in axon pruning and degeneration in *Drosophila*, supporting studies made on Wallerian degeneration in rodents. However, the identity of ubiquitination proteins as targets for the proteasome in the axons and in axonal degeneration remains to be studied.

In conclusion, available data support the view that the mechanisms underlying axonal degeneration and degeneration of the nerve cell body are inherently different. Death of the soma can occur through both caspase-dependent and caspase-independent mechanisms involving the mitochondria, whereas axonal degeneration occur likely via different mechanisms. However, both neuronal and axonal degeneration involves alterations in local activity of the proteasome including changes in ubiquitination of certain target proteins. As discussed above, data on the function of the WLd^s fusion protein have shown that nuclear reactions with change in gene expression cause a delay in axonal degeneration and can contribute to nerve repair. Similar neuroprotective mechanisms with activation of specific gene programs also underlie the action of different neurotrophic factors that counteract neurodegeneration and promote neuronal survival.

Conclusions

Ample evidence suggests that dysregulated proteasome activity can contribute to different neurodegenerative diseases involving both the cell soma and the axons. The discovery of the E3 ligase, Parkin, has given important insights into the role of proteasome in PD, and similar mechanisms or disturbances in protein turnover may prevail in other types of neurological disorders. Studies of the proteasome activity in axonal reactions and in the slow Wallerian degeneration mice (Wlds) have revealed mechanisms governing axonal degeneration and repair. A summary of the current knowledge about the involvement of the proteasome in axonal and nerve cell degeneration is given in Table 1. Increased insights into the function of the proteasome and its disturbances will deepen our understanding about the cellular mechanisms and proteins underlying different neurological disease and help us to design drugs to modify the cellular reactions involved and to counteract disease and axonal degeneration.

Table 1 Summary of the evidence suggesting a role for the proteasome in neuronal and axonal degeneration

Defect in	*Gene*	*Diseases/models*
Ubiquitin gene		
	Ubiquitin^{+1}	AD, HD: accumulation and cell death
Ubiquitination enzymes and related proteins		
E3 ligases	UBE3A	Angelman syndrome
	Parkin	PD
	Dorfin	ALS (mouse model)
Ubiquitin hydrolase	UCH-L1	PD, GAD
Ubiquitin proteases	usp14	Ataxia (mouse model), synaptic defects
Fusion protein	NMNAT	Slow Wallerian degeneration (Wld^s): axonal degeneration
Proteasome inhibition		
Synthetic inhibitors		PD (rat model), cell models: decreased cell viability; axons: decreased axonal degeneration
Decreased proteasomal activity		PD, HD (autopsy material, cell and mouse models)
Aggregates		
PolyQ proteins	Different	HD, SCA, DRPLA, SBMA
Amyloid plaques	Aß	AD
Lewy bodies	α-Synuclein	PD
Bunina bodies	SOD-1	ALS
Others		
Axonal pruning		*Drosophila* models with synthetic proteasome inhibitors
Signaling		NF-κB
Anti-apoptotic proteins	XIAP	Smac/DIABLO
Receptor endocytosis	RTK, Glu6R TrkB, TrkA	Receptor internalization

See also: Alzheimer's Disease: Neurodegeneration; Huntington's Disease: Neurodegeneration; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Stress and Neuronal Plasticity; Ubiquitin–Proteasome System and Plasticity; Wallerian Degeneration.

Further Reading

Conforti L, Fang G, Beirowski B, et al. (2007) NAD(+) and axon degeneration revisited: Nmnat1 cannot substitute for Wld(S) to delay Wallerian degeneration. *Cell Death Differentiation* 14: 116–127.

Fischer DF, DeVos RA, VanDijk R, et al. (2003) Disease-specific accumulation of mutant ubiquitin as a marker for proteasomal dysfunction in the brain. *FASEB Journal* 17: 2014–2024.

Glickman MH and Ciechanover A (2002) The ubiquitin–proteasome proteolytic pathway: Destruction for the sake of construction. *Physiological Reviews* 82: 373–428.

Hegde AN (2004) Ubiquitin–proteasome-mediated local protein degradation and synaptic plasticity. *Progress in Neurobiology* 73: 311–357.

Korhonen L and Lindholm D (2004) The ubiquitin proteasome system in synaptic and axonal degeneration: A new twist to an old cycle. *Journal of Cell Biology* 165: 27–30.

Lee D, Ezhkova E, Li B, Pattenden SG, Tansey WP, and Workman JL (2005) The proteasome regulatory particle alters the SAGA coactivator to enhance its interactions with transcriptional activators. *Cell* 123: 423–436.

Lindsten K, Menendez-Benito V, Masucci MG, and Dantuma NP (2003) A transgenic mouse model of the ubiquitin/proteasome system. *Nature Biotechnology* 21: 897–902.

Millard SM and Wood SA (2006) Riding the DUBway: Regulation of protein trafficking by deubiquitylating enzymes. *Journal of Cell Biology 173*: 463–468.

Moore DJ, West AB, Dawson VL, and Dawson TM (2005) Molecular pathophysiology of Parkinson's disease. *Annual Reviews of Neuroscience* 28: 57–87.

Nateri AS, Riera-Sans L, DaCosta C, and Behrens A (2004) The ubiquitin ligase SCFFbw7 antagonizes apoptotic JNK signaling. *Science* 303: 1374–1378.

Niwa J, Ishigaki S, Hishikawa N, et al. (2002) Dorfin ubiquitylates mutant SOD1 and prevents mutant SOD1-mediated neurotoxicity. *Journal of Biological Chemistry* 277: 36793–36798.

Wilson SM, Bhattacharyya B, Rachel RA, et al. (2002) Synaptic defects in ataxia mice result from a mutation in Usp14, encoding a ubiquitin-specific protease. *Nature Genetics* 32: 420–425.

Yi JJ and Ehlers MD (2005) Ubiquitin and protein turnover in synapse function. *Neuron* 47: 629–632.

Protein Folding and the Role of Chaperone Proteins in Neurodegenerative Disease

P J Muchowski, Gladstone Institute of Neurological Disease, University of California, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Many neurodegenerative disorders are characterized by the accumulation of misfolded proteins in or adjacent to neurons and glia. Misfolded proteins can adopt distinct conformational states that exert abnormal and even pathological activities. Molecular chaperones are proteins that mediate the proper folding of other proteins and appear to have other important functions in the healthy and diseased brain. One essential function is to ensure that misfolded proteins do not accumulate in the brain. Increased expression of molecular chaperones is protective in numerous animal models of neurodegenerative disease. A small molecule that induces the expression of molecular chaperones is being evaluated in a clinical trial for amyotrophic lateral sclerosis (ALS). The molecular mechanisms that enable molecular chaperones to combat the toxicity of misfolded proteins *in vivo* are not well understood. However, their ability to modulate conformational states of disease-related proteins suggests that they prevent early aberrant protein interactions that trigger pathogenic cascades.

Protein Misfolding, Aggregation, and Molecular Chaperones

The amino acid sequence of a protein contains all of the information necessary for it to fold into a functional, three-dimensional structure *in vitro*. Within cells, however, protein folding occurs in a complex environment. The cellular cytoplasm is chock full of proteins, nucleotides, sugars, lipids, and organelles. As they attempt to fold in this dense milieu, proteins can easily misfold and aggregate.

Because the proper folding (or conformation) of proteins is critical to life, a class of 'helper' proteins called molecular chaperones has evolved. These proteins prevent inappropriate interactions within and between nonnative polypeptides, enhance the efficiency of *de novo* protein folding, and promote the refolding of proteins that have become misfolded as a result of cellular stress. Molecular chaperones function at essentially any cellular location where a folding or unfolding protein chain may interact inappropriately (for example, as a protein chain is translocated into mitochondria). Under certain conditions, chaperones interact with components of the ubiquitin–proteasome pathway and lysosome-mediated autophagy to degrade proteins they cannot refold.

What exactly are molecular chaperones? Many are 'heat shock' proteins (HSPs). When cells are stressed, such as with increased temperature, the transcription of HSPs is quickly induced, while generalized protein synthesis is turned off. HSPs are classified into six main families based on their approximate molecular weight. The number of human diseases linked to mutations in molecular chaperones is large and growing.

Protein Aggregation and Neurodegeneration

Molecular chaperones, together with the ubiquitin–proteasome pathway and lysosome-mediated autophagy, constitute a 'quality control' system that normally prevents the accumulation of misfolded proteins. However, since misfolded proteins are thought to accumulate as a result of normal aging, additional insults, such as mutations, can easily saturate the quality control system. Interestingly, neurons and other postmitotic cells appear to be especially vulnerable to the effects of misfolded proteins. It is now abundantly clear that saturation of the protein quality control machinery is a factor in many neurodegenerative diseases, including Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease (HD) and related polyglutamine (polyQ) disorders, ALS, and the prion encephalopathies.

The major pathological feature of neurodegeneration associated with misfolded proteins is the accumulation of aggregated protein in intra- or extracellular deposits, generically referred to as amyloid and known as senile plaques and neurofibrillary tangles in AD, Lewy bodies and neurites in PD, and inclusion bodies in HD and other diseases in which the protein aggregates inside cells. The aggregated proteins are composed of fibrillar material, as shown by electron microscopy, and have similar chemical structures, consisting of β-sheets perpendicular to the fibril axis, that produce a characteristic cross-β X-ray diffraction pattern. They also share biochemical properties, such as insolubility in detergents, resistance to proteases, and the ability to bind lipophilic dyes such as congo red.

Diverse neurodegenerative diseases are characterized by accumulation of amyloid-like fibrils and likely have common pathogenic mechanisms. But are the amyloid fibrils themselves toxic? Increasing evidence suggests that fibrillar aggregates are inert or perhaps

even protective, rather than being directly pathogenic. For example, senile plaques are found in persons without clinical symptoms of AD, and in cases of AD, the severity of dementia does not correlate well with plaque density. In PD, neuropathological analyses show that neurons that contain Lewy bodies are healthier than cells that do not, indicating that inclusion bodies might be protective. In support of this hypothesis, the formation of inclusion bodies by proteins with expanded polyQ repeats is a regulated cellular process that requires an intact microtubule cytoskeleton. This process might have evolved as a protective mechanism to sequester toxic, misfolded protein entities that would otherwise disrupt cellular homeostasis.

If amyloid fibrils do not trigger pathogenic cascades in neurodegenerative diseases, then what is the toxic culprit? In the past decade, detailed structural and biochemical studies of purified proteins have shown that the assembly of aggregating proteins into amyloid fibrils is a complex process, involving a number of different intermediates. Intense research has focused on the pathogenic role of these so-called prefibrillar or protofibrillar amyloid intermediates, also called soluble oligomers or, in the case of AD, amyloid β-peptide (Aβ)-derived diffusible ligands (ADDLs). Atomic force microscopy has been used to characterize several metastable structures, such as spherical oligomers, protofibrils, and porelike annular structures, which might be on a pathway to fibril formation. These small, potentially diffusible assemblies, rather than mature amyloid fibrils, might trigger neuronal dysfunction by initiating a cascade that culminates in neuronal death. For example, injecting soluble oligomers of Aβ into the brain decreases long-term potentiation, and it is thought that the oligomers specifically target proteins localized to synapses. Antibodies have been generated that specifically recognize aggregating proteins solely in their oligomeric forms. Incubation of oligomers with these antibodies blocks their subsequent toxicity to cultured cells, supporting the hypothesis that soluble oligomers are the toxic culprit. However, our understanding of soluble oligomers is largely based on *in vitro* and cell culture studies. Little is known about their *in vivo* role for the simple reason that we lack the tools to visualize their structure *in vivo*. The development of such tools will allow one to carefully characterize from a temporal and spatial perspective the role of oligomers in pathogenesis.

Molecular Chaperones and Neurodegeneration

Since molecular chaperones normally function in protein folding, it is not surprising that they can be found, together with components of the ubiquitin–proteasome system, by immunohistological analysis of sections from diseased brains at sites of aggregate deposition, such as in senile plaques and inclusion bodies. In several studies, sophisticated fluorescence imaging techniques showed that molecular chaperones are not bound irreversibly to intracellular aggregates. Rather, they quickly cycle on and off the aggregates, apparently in an effort to refold the misfolded protein, to transfer it to the ubiquitin–proteasome system, or to help sequester oligomeric intermediates into fibrillar inclusion bodies. Future studies will be required to discriminate among these and other possibilities.

Compelling evidence to support a critical role for chaperones in neurodegenerative diseases comes from recent *in vivo* studies that show that overexpression of the chaperone Hsp70 or Hsp40 suppresses neurodegeneration in fruit fly and transgenic mouse models of HD and PD. While initial studies were based on a candidate gene approach, unbiased genetic screens in *Saccharomyces cerevisiae, Drosophila melanogaster,* and *Caenorhabditis elegans* models of polyQ aggregation have further substantiated the importance of molecular chaperones and protein folding in regulating polyQ aggregation and toxicity. Importantly, overexpression of chaperones also suppresses neurodegeneration *in vivo* in transgenic mouse models of PD and polyQ diseases.

Mechanisms of Chaperone-Mediated Neuroprotection

Although chaperones are clearly protective in animal models of neurodegenerative disease, the mechanisms that underlie protection are unclear. They can directly influence the folding state or conformation of disease-associated proteins that aggregate. They might alter a particular conformational state that is highly toxic, exerting a protective effect simply by decreasing its abundance. Another possibility is that molecular chaperones may cooperate with the ubiquitin–proteasome and lysosome degradation systems to help degrade aggregation-prone proteins; however, experimental evidence in animal models that directly supports this mechanism of action is lacking. The chaperones likely facilitate neuroprotection through mechanisms that are both dependent on, and independent of, their protein folding activity (**Figure 1**). In addition to cooperating with the protein degradation machinery, they can also block apoptotic signaling pathways. Ultimately, the collective effects of multiple direct and indirect chaperone actions might be essential to impede disease pathogenesis.

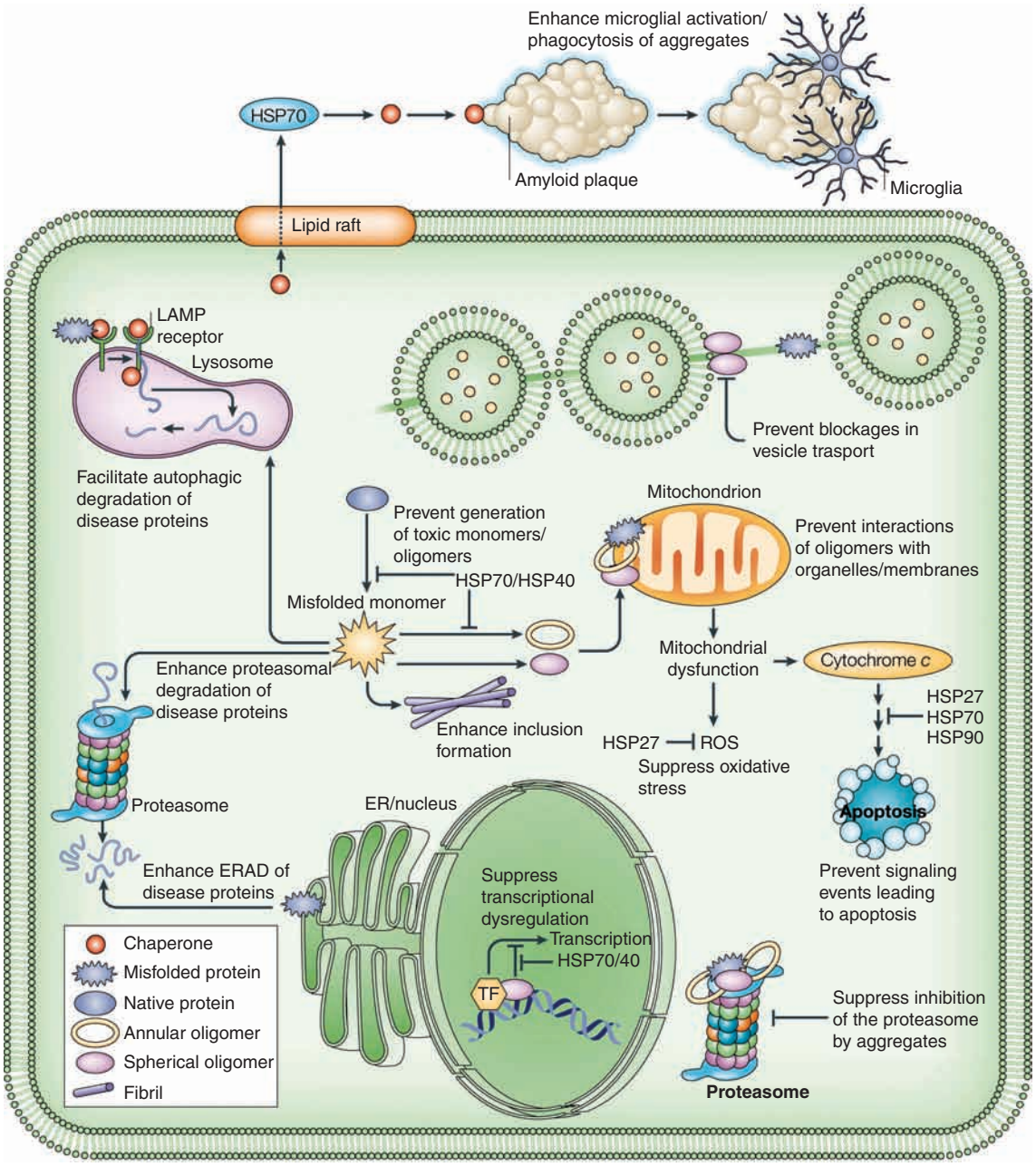

Figure 1 Direct and indirect effects of molecular chaperones on disease protein toxicity. Molecular chaperones might prevent toxicity by preventing inappropriate protein interactions, by facilitating disease protein degradation/sequestration, and/or by blocking downstream signaling events that lead to neuronal dysfunction and apoptosis. HSP, heat shock protein; LAMP, lysosome-associated membrane protein; ROS, reactive oxygen species; ER, endoplasmic reticulum; ERAD, endoplasmic-reticulum-associated degradation. Reproduced from Muchowski PJ and Wacker JL (2005) Modulation of neurodegeneration by molecular chaperones. *Nature Reviews Neuroscience* 6: 11–22, with permission.

Molecular Chaperones as Potential Drug Targets

It is clear from animal studies that increased expression of molecular chaperones might prove beneficial for the treatment of neurodegenerative diseases. However, finding drugs that increase the expression of proteins in specific cells and tissues will be a major challenge. Furthermore, continuous upregulation of molecular chaperones might cause undesirable side effects, including cancer. Nonetheless, a small molecule that induces expression of HSPs has shown promise in an animal model of ALS and is being evaluated in a phase II clinical trial.

Conclusions

Protein folding in neurons mediated by molecular chaperones plays a critical role in normal brain function and in neurodegenerative diseases. The mechanism by which chaperone activity facilitates neuroprotection is being studied intensively in many laboratories, but deep insights are still lacking. In principle, molecular chaperones are an attractive target for therapeutic intervention in neurodegenerative diseases, because they presumably can act at the earliest change of the disease process, by preventing aggregation-prone proteins from adopting a toxic conformation. Interestingly, most studies of molecular chaperones in animal models have shown that chaperones facilitate neuroprotection in the absence of a visible effect on inclusion-body formation, suggesting that protection might arise from decreasing formation of small, diffusible aggregate assemblies or perhaps of misfolded conformations of monomers.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Blood-Brain Barrier and Neurovascular Mechanisms of Neurodegeneration and Injury; Huntington's Disease: Neurodegeneration; Neuroprotection: Pharmacological Approaches; Oxidative Damage in Neurodegeneration and Injury; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Bukau B and Horwich AL (1998) The Hsp70 and Hsp60 chaperone machines. *Cell* 92: 351–366.

Caughey B and Lansbury PT (2003) Protofibrils, pores, fibrils, and neurodegeneration: Separating the responsible protein aggregates from the innocent bystanders. *Annual Review of Neuroscience* 26: 267–298.

Cuervo AM and Dice JF (1998) Lysosomes, a meeting point of proteins, chaperones, and proteases. *Journal of Molecular Medicine* 76: 6–12.

Dobson CM (2003) Protein folding and misfolding. *Nature* 426: 884–890.

Glabe CG (2006) Common mechanisms of amyloid oligomer pathogenesis in degenerative disease. *Neurobiology of Aging* 27: 570–575.

Hartl FU and Hayer-Hartl M (2002) Molecular chaperones in the cytosol: From nascent chain to folded protein. *Science* 295: 1852–1858.

Lindquist S (1986) The heat-shock response. *Annual Review of Biochemistry* 55: 1151–1191.

Morimoto RI and Santoro MG (1998) Stress-inducible responses and heat shock proteins: New pharmacologic targets for cytoprotection. *Nature Biotechnology* 16: 833–838.

Muchowski PJ and Wacker JL (2005) Modulation of neurodegeneration by molecular chaperones. *Nature Reviews Neuroscience* 6: 11–22.

Selkoe DJ (2004) Cell biology of protein misfolding: The examples of Alzheimer's and Parkinson's diseases. *Nature Cell Biology* 6: 1054–1061.

Sherman MY and Goldberg AL (2001) Cellular defenses against unfolded proteins: A cell biologist thinks about neurodegenerative diseases. *Neuron* 29: 15–32.

Psychiatric Disorders Associated with Disturbed Sleep and Circadian Rhythms

E K Simon, Rosalind Franklin University of Medicine and Science, North Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sleep disturbance is one of the most common complaints associated with the psychiatric syndromes described in the diagnostic and statistical manual of mental disorder-4th edition-text revision (DSM-IV-TR). The mood disorders, unipolar and bipolar together with anxiety disorders (generalized anxiety disorder (GAD) and posttraumatic stress disorder (PTSD)), have sleep disturbance in their criterion. Schizophrenia is also associated with poor sleep quality. Sleep and circadian disturbances influence the presentation and course of psychiatric illnesses, together with the quality of life, and in many cases help in predicting the outcome. Models developed from studies of sleep and circadian rhythms in psychiatric patients help us gain insight into the disease process itself. Proper diagnosis and treatment with emphasis on sleep and circadian rhythm lead to better understanding of the disorders, as well as an improved outcome.

Mood Disorders

Major Depressive Episode

The diagnostic criteria for major depressive episode (MDE) are shown in Table 1.

Sleep disturbance in MDE MDE is characterized by depressed mood, anhedonia, fatigue, psychomotor retardation or agitation, insomnia or hypersomnia, weight loss, guilt, sense of worthlessness, and suicidal ideations, plans, or acts (Table 1). Insomnia manifests as difficulty in falling asleep or staying asleep or waking up earlier than the usual schedule. Self-reports, questionnaires, and structured interviews from several studies reveal poor sleep quality and impaired daytime functioning in MDE. Hypersomnia is seen primarily in younger patients and in bipolar depressed patients.

According to the (DSM-IV-TR), 90% of inpatients and 40–60% of outpatients with MDE have polysomnography (PSG) abnormalities. They have delayed sleep onset, disturbed sleep continuity, and poor sleep efficiency (Figure 1). Slow-wave sleep (SWS and non-rapid eye movement (NREM) stages 3 and 4) is reduced particularly in the first half of the night. Rapid eye movement (REM) sleep shows decreased REM latency, increased REM density, and increased duration of REM sleep early in the night and increased duration of REM throughout the night.

Quantitative electroencephalography (EEG) shows decreased delta wave power and delta wave counts during sleep. The abnormalities in delta sleep are seen more in men than in women. PSG abnormalities are not prominent in adolescents and prepubertal children. In psychotic depression, the PSG findings are most marked and REM latency is very short. The REM sleep abnormalities are associated more with the first episode of depression, or seen earlier during the course of a depressive episode or in recurrence. The patients with significant life stressors preceding the onset of a depressive episode is less likely to have REM sleep abnormalities compared to patients without stressors. Reduced REM latency is associated with increased response to pharmacotherapy and electroconvulsive therapy (ECT) but not psychotherapy. Response to cognitive behavioral therapy (CBT) or interpersonal therapy (IPT) is significantly reduced in patients with reduced REM latency, increased REM density, and disturbed sleep continuity. Reduced REM latency and reduced SWS are trait markers of depression, as they persist after remission and are seen in twins and first-degree relatives of the patients.

Functional neuroimaging findings in MDE Functional neuroimaging studies of depressed patients during wakefulness show hypometabolism in the dorsolateral prefrontal cortex (DLPFC) and anterior cingulate regions and hypermetabolism in the amygdala, hippocampus, and the orbitofrontal cortex (ventral and medial) (Figure 2). The functional neuroimaging studies during sleep reveal differences between controls and patients during both NREM and REM sleep. An 18-fluorodeoxyglucose positron emission tomography ([^{18}F] FDG PET) study focusing on the first NREM sleep period in MDE showed a higher metabolic rate of glucose globally in comparison to controls. The glucose metabolic rate was high in the posterior cingulate, amygdala, hippocampus, occipital and temporal cortices, and the pons. Reduced glucose metabolism was found in the anterior cingulate, caudate, and medial thalamus compared to controls. Another [^{18}F] FDG PET study examining the transition from wakefulness to NREM sleep in MDE found that the metabolic rate remained relatively elevated in the frontoparietal regions and the thalamus (Figure 2). Examination of the relationship

Table 1 Diagnostic criteria for major depressive episode

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.

Note: Do not include symptoms that are clearly due to a general medical condition or mood-incongruent delusions or hallucinations.

1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful). *Note:* In children and adolescents, can be irritable mood.
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others).
3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. *Note:* In children, consider failure to make expected weight gains.
4. Insomnia or hypersomnia nearly every day.
5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
6. Fatigue or loss of energy nearly every day.
7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

B. The symptoms do not meet criteria for a Mixed Episode.

C. The symptoms cause clinically significant distress or impairment in social,occupational, or other important areas of functioning.

D. The symptoms are not due to the direct physiological effects of a substance (e.g., a drug of abuse or a medication) or a general medical condition (e.g., hypothyroidism).

E. The symptoms are not better accounted for by bereavement (i.e., after the loss of a loved one), the symptoms persist for longer than 2 months or are characterized by marked functional impairment, morbid preoccupation with worthlessness, suicidal ideation, psychotic symptoms, or psychomotor retardation.

Source: American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington, DC: American Psychiatric Association.

between regional cerebral glucose metabolism assessed by [^{18}F] FDG PET and beta EEG power, a marker of arousal, in patients with MDE and controls during NREM sleep showed that beta power is negatively correlated with subjective sleep quality. Both groups showed significant correlation between beta power and regional cerebral glucose metabolism in venteromedial PFC and right lateral inferior occipital cortex. The studies taken together suggest a dysfunction of arousal in the depressed patients.

Functional neuroimaging of REM sleep in the normal subjects show a pattern resembling wakefulness in the cortical metabolism (Figure 3). There is also increased metabolism in the limbic and paralimbic structure during REM sleep. This has been referred to as the anterior paralimbic REM activation axis. In patients with MDE, [^{18}F] FDG PET reveals an activation of the anterior paralimbic structures greater in area, spatially, compared to controls. They also show greater activation in bilateral DLPFC, left premotor, primary sensorimotor, and left parietal cortices, together with midbrain reticular formation. The increased paralimbic activation may be related to affective dysregulation and the increased activation of the DLPFC may be linked to cognitive dysfunction in depressed patients.

Circadian rhythm disturbance in MDE *Methodological issues*. Most of the studies on circadian rhythm in MDE are of patients following a regular diurnal rhythm. Constant routine or other protocols used to avoid masking are seldom used. When employed, sleep deprivation, a necessary part of the procedure, can lead to remission or a switch into hypomania or mania, thus making interpretation difficult. The disorder itself can influence the circadian rhythm and lead to differences between controls and patients. The differences noted may be epiphenomena or they can actually precipitate or perpetuate the depression.

Standardized sleep–wake schedules that are used prior to protocols, such as constant routine, being therapeutic, alter the phenomenology of the illness. Even with these limitations, this strategy can eliminate the possibility that the observed differences are epiphenomena from the influence of the disordered sleep–wake cycle on the patient's circadian rhythms. Most studies recruit patients via ads or from tertiary centers, thereby reducing the generalizability of the findings. There is also the bias of self-selection. Use of a heterogeneous patient population has led to confounding of results in circadian studies (e.g., mixing bipolar and unipolar patients, inpatients and outpatients, melancholic and nonmelancholic patients). Homogenous patient samples, correctly matched for age and sex, with control groups, lead to better understanding of the disease and to valid, reproducible results.

Core body temperature rhythm in MDE. Early studies had reported a phase advance of core body

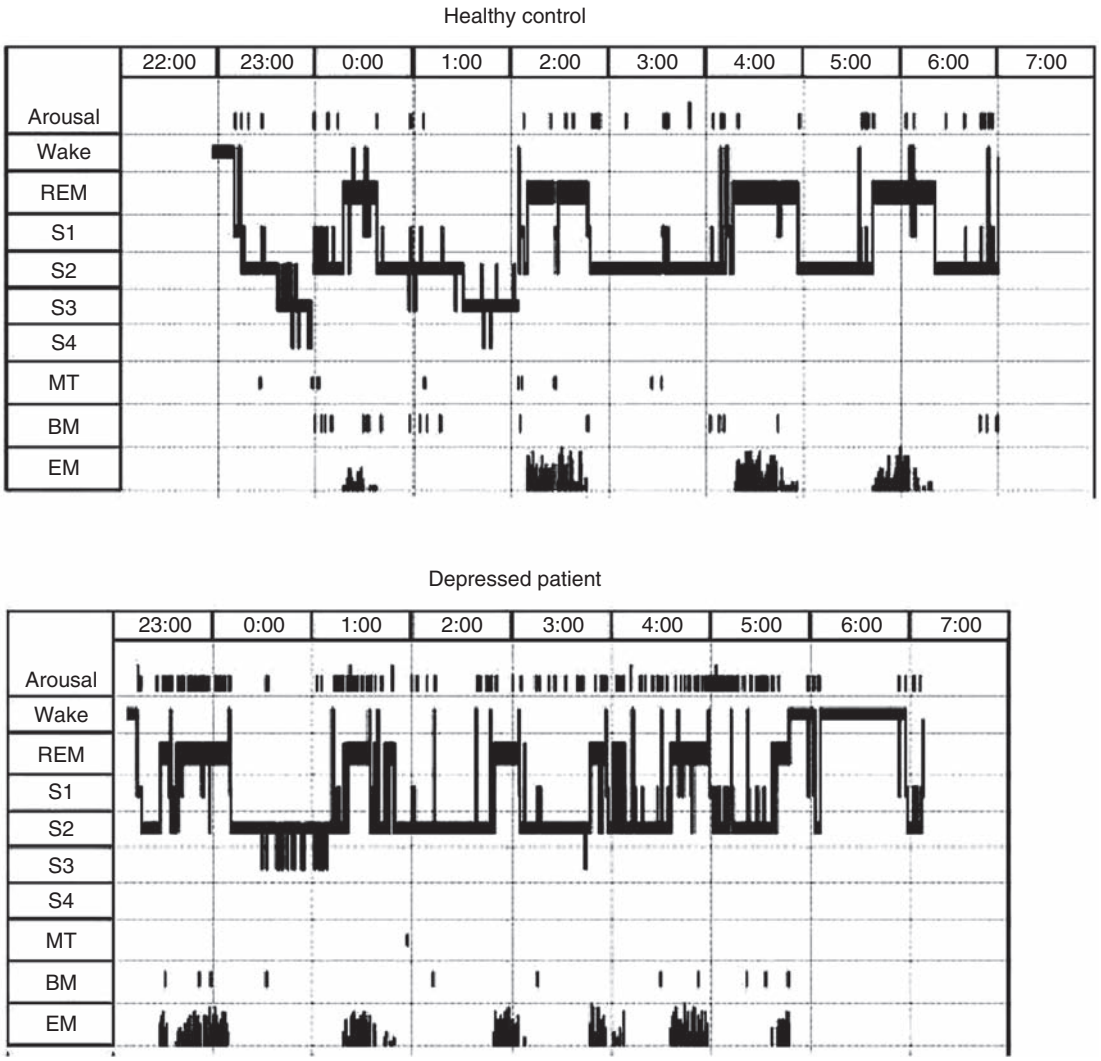

Figure 1 Sleep profile of a medication-free depressed female patient and a healthy female control. Compared with the healthy subject, the sleep profile of this patient shows many of the typical features of sleep in depression: impaired sleep continuity, disinhibition of REM sleep, and reduction of slow-wave sleep. W, wake; REM, rapid eye movement sleep; S1–4, sleep stages 1–4; MT, movement time; BM, body movement; EM, rapid eye movements. From Berger M, van Calker D, and Riemann D (2003) Sleep and manipulations of the sleep–wake rhythm in depression. *Acta Psychiatrica Scandinavica* 108(supplement 418): 84. Used with permission from Blackwell Munksgaard.

temperature in MDE, but it was not consistently replicated by later studies. The most consistent finding was the reduced amplitude of the core body temperature (Figure 4). This was related to the reduced fall of temperature during sleep. The amplitude was lower among patients with reduced REM latency compared to patients with longer REM latency. The period and amplitude of core body temperature rhythm normalizes while in remission.

Melatonin in MDE. Melatonin exhibits blunted amplitude, as already noted, with temperature. Studies with highly specific assays did not validate the finding already discussed that resulted from older studies. Melatonin is also influenced by light, posture, and medications, factors not taken into account by older studies. Reduced concentration of melatonin in patients with MDE has been reported, but the studies have methodological flaws, such as small sample size, poor matching, and infrequent sampling. At reported temperature, the amplitude of melatonin normalizes with remission. Differing levels of nocturnal melatonin secretion have been observed during two different episodes of depression in the same patient. This finding should serve as a cautionary note while we attempt to interpret cross-sectional studies.

Hormonal rhythms in MDE. Among the rhythms studied, cortisol linked to the circadian rhythm has shown a decrease in amplitude and phase advance. Growth hormone secretion shows a reversal of its usual pattern, leading to more secretion in the day as opposed to the night. Thyroid-stimulating hormone also shows decreased amplitude. To sum up, decreased

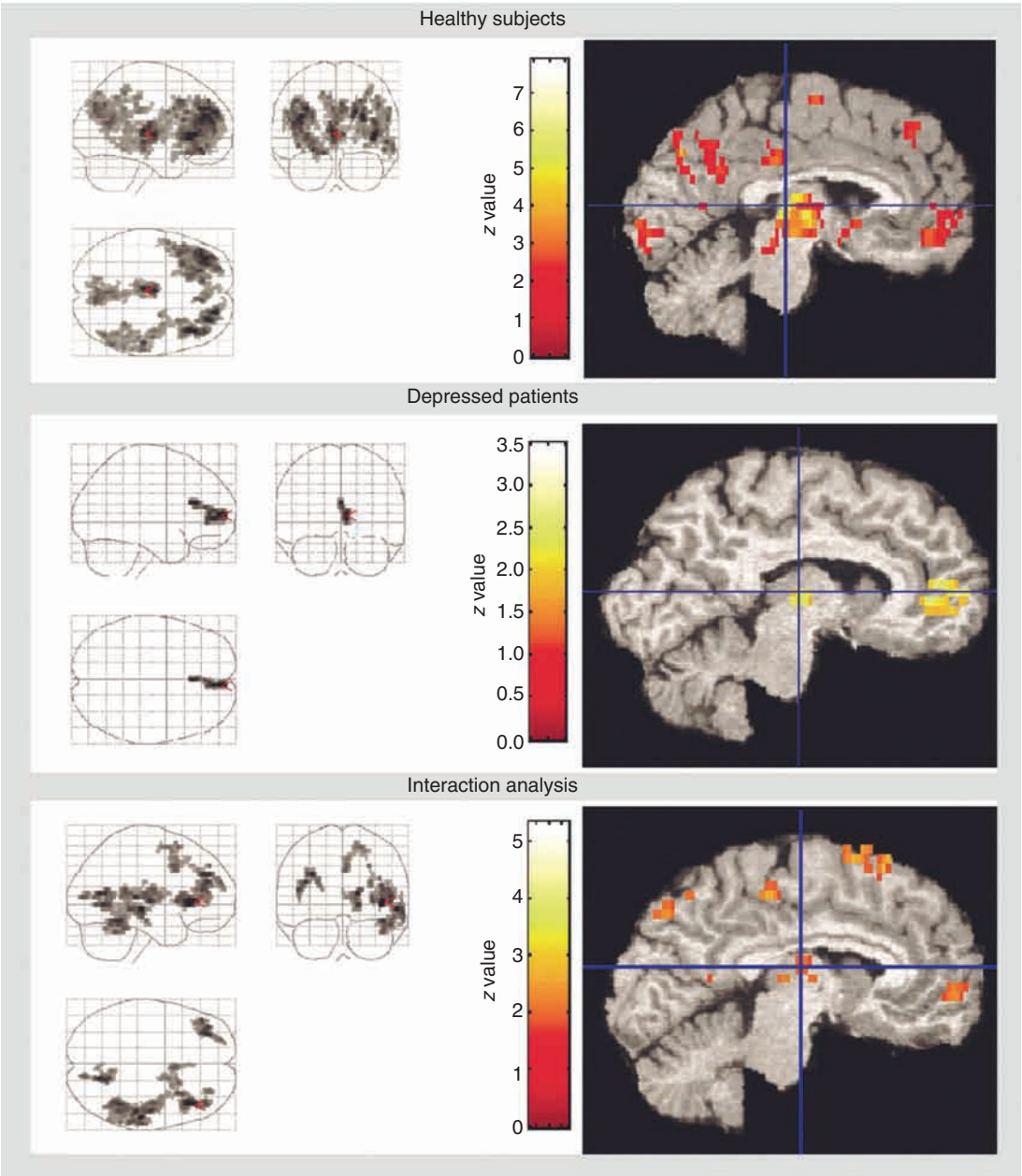

Figure 2 Areas of significant reduction in cerebral glucose metabolism from presleep wakefulness to NREM sleep in healthy subjects and depressed patients and between group differences in metabolic change. Region-of-interest analyses with small volume corrections show a reduction in regional cerebral glucose metabolism in the thalamus from presleep wakefulness to NREM sleep in both study groups. A bilateral metabolic decline in the thalamus was seen in the healthy subjects (left thalamus shown in figure; blue lines intersect at Talairach coordinates $x = -4$, $y = -22$, $z = 8$), whereas the metabolic decline was limited to the right thalamus in depressed patients (blue lines intersect at Talairach coordinates $x = 8$, $y = -16$, $z = 8$). Areas where depressed patients exhibited less of a decline in regional cerebral glucose metabolism from presleep to NREM sleep relative to healthy subjects are presented in the bottom image: depressed patients showed less of a metabolic decline in the thalamus relative to healthy subjects (blue lines intersect at Talairach coordinates $x = -2$, $y = -14$, $z = 4$). From Germain A, Nofzinger EA, Kupfer DJ, and Buysee DJ (2004) Neurobiology of non-REM sleep in depression: Further evidence for hypofrontality and thalamic dysregulation. *American Journal of Psychiatry* 161: 1859. Used with permission from American Psychiatric Publishing, Inc.

amplitude of various rhythms appears to be the most consistent observation in depression (Figure 4).

A recent study comparing patients with MDE with and without psychotic features suggests that each subtype is associated with distinct patterns of hypothalamic–pituitary–adrenal (HPA) axis activity. The patients with psychosis had significantly higher evening cortisol levels than healthy controls

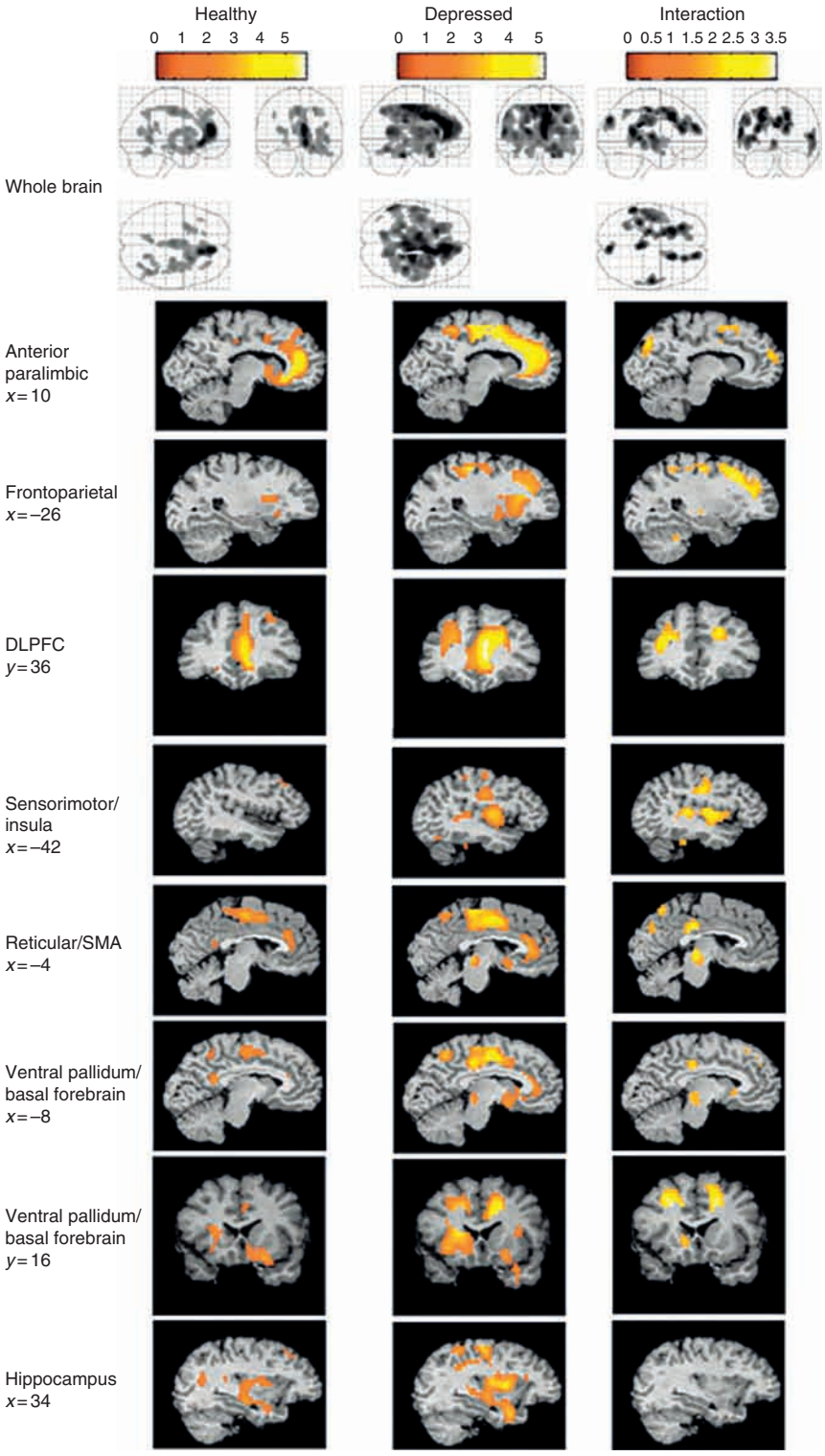

Figure 3 Waking to REM sleep activation, comparison between patients with MDE and healthy controls. Waking to REM sleep activations in healthy subjects (column 1), depressed subjects (column 2), and interactions showing regions where the depressed subjects' waking to REM activations are greater than those of healthy subjects (column 3). DLPFC indicates dorsolateral prefrontal cortex; SMA, supplementary motor area; and *x* and *y*, Talairach *x* and *y* coordinates, respectively. From Nofzinger EA, Buysee DJ, and Germain A, et al. (2004) Increased activation of anterior paralimbic and executive cortex from waking to rapid eye movement sleep in depression. *Archives of General Psychiatry* 61: 699. Used with permission from American Medical Association.

or nonpsychotic patients. This is consistent with earlier literature and is thought to be linked to reduced feedback inhibition of HPA axis in psychotic depression.

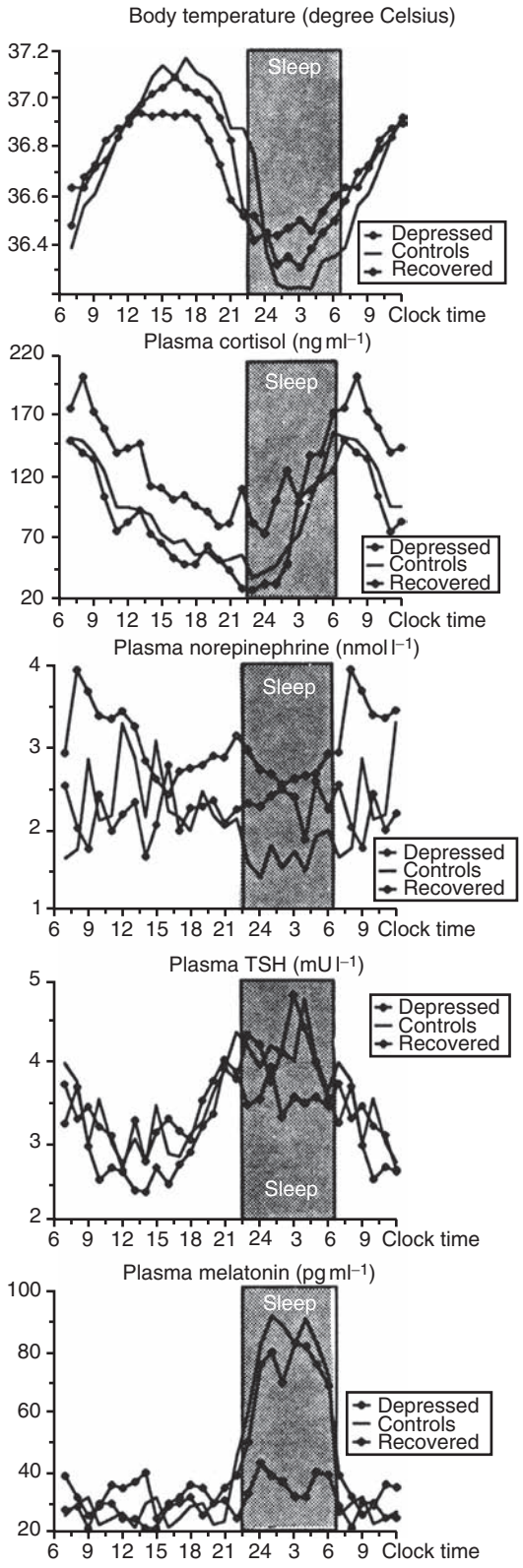

Figure 4 Comparison of circadian rhythms of temperature, cortisol, norepinephrine (NE), thyroid stimulating hormone (TSH), and melatonin between patients with endogenous depression, recovered patients, and healthy subjects. Circadian profiles approximating that of controls are noted in recovered patients. Adapted from Souetre E, Salvati E, Belugou J-L, et al. (1989) Circadian rhythms in depression and recovery: Evidence for blunted amplitude as the main chronobiological abnormality. *Psychiatry Research* 28: 266. Used with permission from Elsevier Scientific Publishers Ireland, Ltd.

Diurnal variation in MDE. It has been known since antiquity that the symptoms of depression have a distinct diurnal pattern: worse in the morning and improving as the day progresses. This is especially true for the melancholics. Some studies suggest that the diurnal variation is also seen in other types of depression, other psychiatric disorders, and even in healthy subjects. It has been demonstrated that every pattern of diurnal variation can be observed in a population of depressed individuals on long-term follow-up (i.e., morning low, evening low, and indifferent type).

A recent study of patients with MDE compared with controls, matched for age, gender, body mass index (BMI), and menstrual cycle, found significant mean elevation of interleukin-6 (IL-6) in the patient group from 1000 to 1200 h and at 1500 h (**Figure 5**). They also found a 12 h shift in the circadian rhythm of IL-6 but not of cortisol in the MDE patients with the zenith during wake hours and nadir during sleep hours, as opposed to the rhythm in normal subjects (zenith during sleep hours (0100–0500 h) and nadir during morning hours (0800–1000 h)). A striking finding of the study was the significant correlation found between fatigue and poor concentration with IL-6 levels in depressives, which could explain the diurnal variation of symptoms.

Social rhythms in MDE. The social *zietgeber* hypothesis posits that in patients who are vulnerable to affective disorder, changes in social *zeitgebers* can lead to disruption of circadian rhythms and eventually precipitate a mood episode, manic, or depressive (**Figure 6**). The event in question may be a *zeitstorer*, an active disrupter of social and familial life, such as the death of a spouse or the birth of a new baby. The *zeitstorer* leads to lack of regular activity or to uncoupling phenomena, where the regular activity fails to entrain the circadian rhythm and presentation of the disorder.

A recent study found that depressed and control groups were similar in levels of regular daily activity, done alone or with others. The regular pattern of daily activity was entraining the cortisol rhythm of healthy controls but not of the depressed patients,

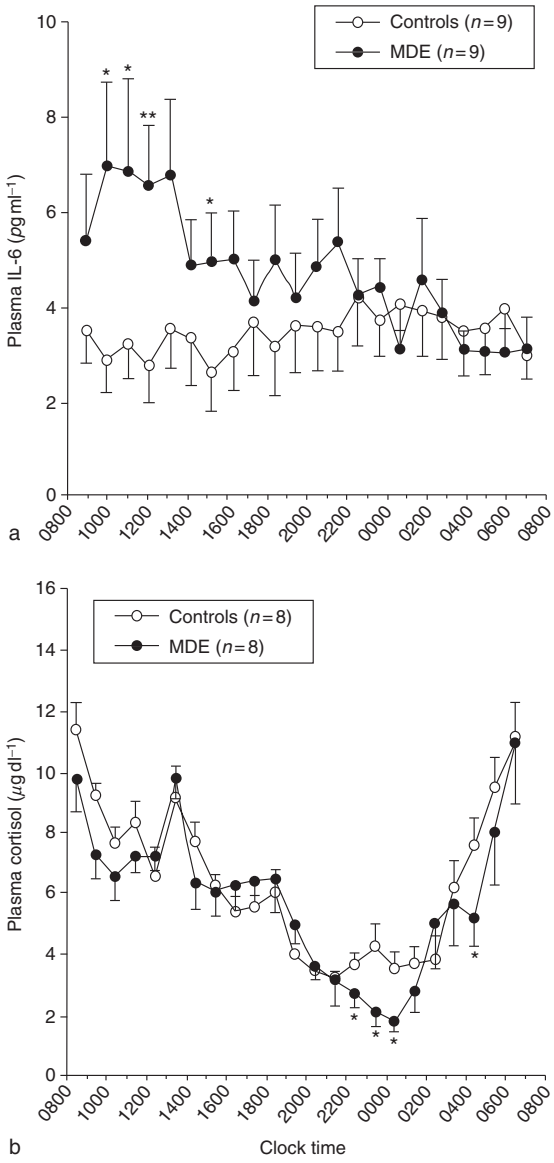

Figure 5 Mean chronograms of IL-6 (a) and cortisol levels (b) measured over 24 h in patients with MDE and age-, gender- and BMI-matched controls. A significant group-by-time interaction in repeated-measures ANOVA, followed by *post hoc t*-tests at each time point, showed significant IL-6 elevation from 1000–1200 h and at 1500 h, and cortisol decrease from 2230–0030 h and at 0430 h in MDE patients, compared with controls. Results are presented as mean ± SEM in metric units (conversion factors to SI units: IL-6, 0.131; cortisol, 27.5862). *, $P < 0.05$; **, $P < 0.01$. From Alesci S, Martinez PE, Kelkar S, et al. (2005) Major depression is associated with significant diurnal elevations in plasma interlukin-6 levels, a shift of its circadian rhythm, and loss of physiological complexity of its secretion. *Journal of Clinical Endocrinology and Metabolism* 90: 2525. Used with permission from the Endocrine Society.

and the controls had a more normative decline in cortisol rhythm on the days they reported more activities involving others. In depressive patients the daily activity was not entraining the cortisol rhythm

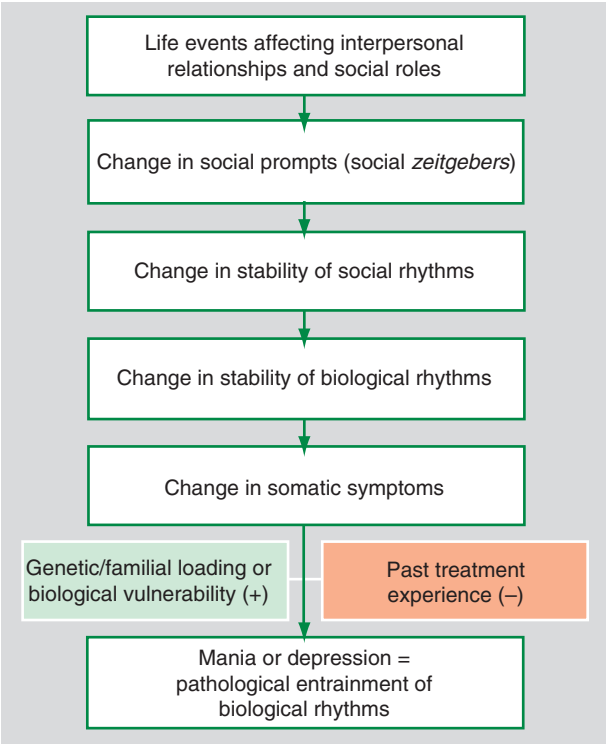

Figure 6 Social *zeitgeber* hypothesis. Adapted from Frank E (2005) *Treating Bipolar Disorder: A Clinician's Guide to Interpersonal and Social Rhythm Therapy*. Used with permission from Guilford Press.

and an uncoupling was observed. It should be noted that the patients were outpatients who had a score of 23 on Beck Depression Inventory without meeting the DSM criteria for MDE.

Psychotherapy helps re-entrain the disrupted rhythm of depressed patients by establishing regular routines and providing exposure to *zeitgebers*, both photic and otherwise. With cognitive behavior therapy, activity scheduling, in which the patient is asked to schedule and engage in preplanned activities, was found to be one of the critical components of the therapeutic action.

Seasonal Affective Disorder

In seasonal affective disorder (SAD), episodes of depression are temporally connected to a particular time of the year (e.g., depression in winter and remission in summer) (Table 2). Winter SAD shows prominent anergy, hypersomnia, overeating, weight gain, and carbohydrate craving. Less common is presentation of depression in the summer; summer SAD presents with insomnia, irritability, loss of appetite, and loss of weight. SAD can be a part of a unipolar or bipolar disease. It is more common in Bipolar II Disorder, in which the summer leads to hypomania. The prevalence of the disease increases with higher latitudes, young are at higher risk, and women comprise 60–90% of patients with SAD.

Self-report questionnaires and prospective sleep diaries show that the symptomatic SAD patients have significant increase in sleep duration and difficulty awakening in winter (Figure 7). Polysomnographic studies of SAD show reduced sleep efficiency, reduced NREM stage 3 and 4, and increased REM density, but less evidence of reduced REM sleep latency (SL). Actigraphic studies have found dimin-

Table 2 Diagnostic criteria for seasonal pattern specifier

Specify if:

With Seasonal Pattern (can be applied to the pattern of Major Depressive Episodes in Bipolar I Disorder, Bipolar II Disorder, or Major Depressive Disorder, Recurrent)

A. There has been a regular temporal relationship between the onset of Major Depressive Episodes in Bipolar I or Bipolar II Disorder or Major Depressive Disorder, Recurrent, and a particular time of the year (e.g., regular appearance of the Major Depressive Episode in the fall or winter). *Note:* Do not include cases in which there is an obvious effect of seasonal-related psychosocial stressors (e.g., regularly being unemployed every winter).

B. Full remissions (or a change from depression to mania or hypomania) also occur at a characteristic time of the year (e.g., depression disappears in the spring).

C. In the last 2 years, two Major Depressive Episodes have occurred that demonstrate the temporal seasonal relationships defined in criteria A and B, and no nonseasonal Major Depressive Episodes have occurred during that same period.

D. Seasonal Major Depressive Episodes (as described above) substantially outnumber the nonseasonal Major Depressive Episodes that may have occurred over the individual's lifetime.

Source: American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington, DC: American Psychiatric Association.

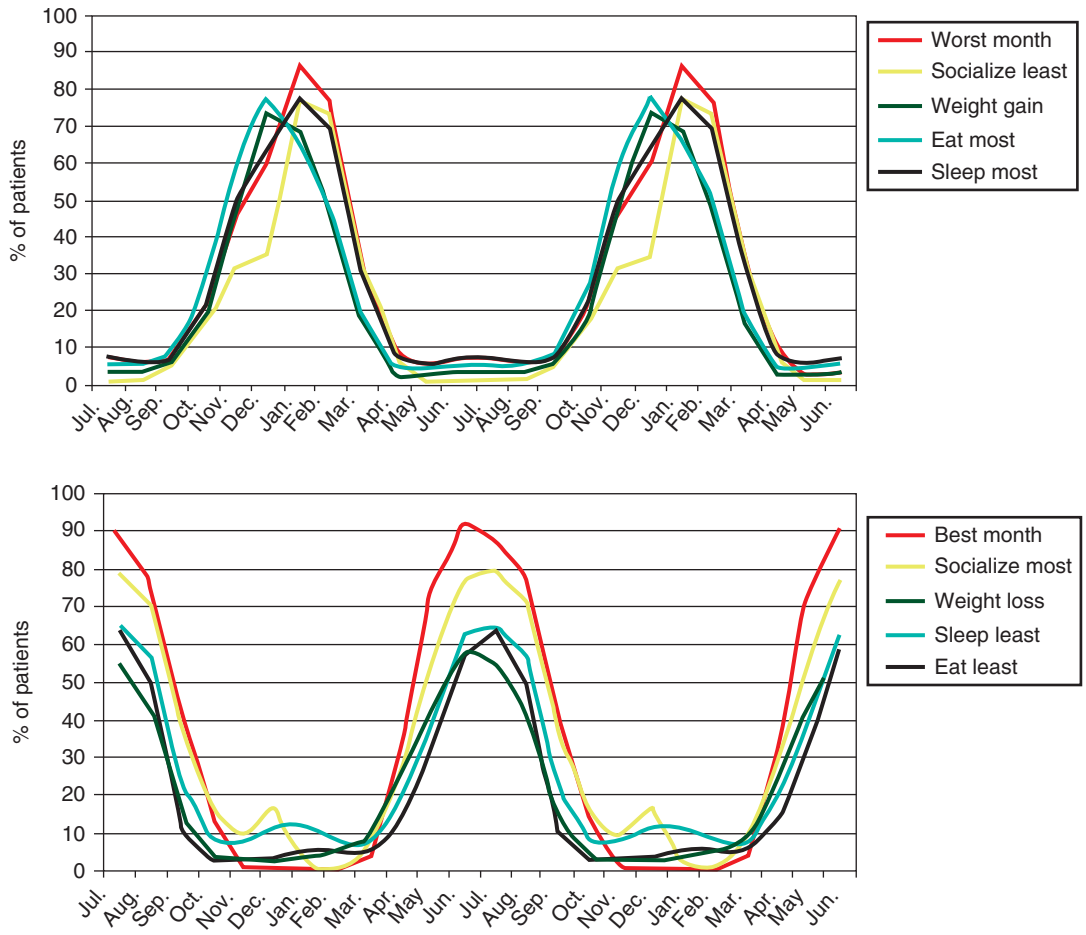

Figure 7 Seasonal change in symptoms of SAD ($n = 1042$). Adapted from Modell JG, Rosenthal NE, and Harriet AE et al. (2005) Seasonal affective disorder and its prevention by anticipatory treatment with bupropion XL. *Biological Psychiatry* 58: 663. Used with permission from Society of Biological Psychiatry.

ished levels of activity in both seasonal and nonseasonal depression.

A recent trial of bright light for those with SAD comparing patients who fulfilled the Rosenthal and DSM-IV-TR criteria for SAD with age and sex-matched healthy subjects found that the former had 33% lower total activity and 43% lower daylight activity in week 1 of the trial, as documented by actigraphy. The SAD patients also exhibited attenuation of the amplitude of sleep–wake cycle by 6%, a phase delay of almost 1 h, and lowered sleep efficiency. The bright light by week 4 led to increase in the total and daylight activity in the patients. The reduced amplitude of the circadian rhythm also increased and the delayed circadian rhythm advanced with improved sleep efficiency.

The above view is supported by another study that demonstrated that the prototypical SAD patient is phase delayed (71% in the study) whereas a subgroup may be phase advanced (29% in the study). The same study also found that the phase-delayed patient group responded to low-dose melatonin in the afternoon. The effect was pronounced as the phase angle difference (PAD) between dim light melatonin onset (DLMO) and midsleep approached the ideal of 6 h. The phase-advanced SADs showed a similar antidepressant response to low-dose melatonin when it was administered in the morning.

The treatment for SAD is bright light therapy, at an intensity of up to 10 000 lux, for a period of 30–60 min in the morning. Studies suggest that morning light is superior to evening light. The bright lights used currently are polychromatic. One recent study reports that blue light with short wave length (468 nm) from light-emitting diodes (LED) is comparable in efficacy to bright white light and is better than dim long-wavelength red light.

FDA approved bupropion extended release in 2006 for prevention of SAD. The pivotal studies were three randomized placebo controlled trials of patients ($n = 1042$) who had history of SAD, on average 13 prior episodes, who were treated with Bupropion XL 300 mg day^{-1}, in autumn prior to the onset of depression. At the end of 4–6 months, the treatment groups had 16% patients who had depression as opposed to the 28% in the placebo group. It should be noted that there was a significant placebo response, which was attributed to the close attention the placebo group received during the trial. Historically, the placebo response has been increasing in efficacy trials of depression.

Bipolar Disorder

Bipolar disorder is a severe disorder of mood manifested by recurrent episodes of mania and depression. The mania is characterized by persistently elevated, expansive, or irritable mood, with grandiosity, decreased need for sleep, pressure of speech, distractibility, flight of ideas, and increased involvement in goal-directed or high-risk activity (Table 3). Patients with bipolar disorder exhibit depression that resembles unipolar depression. Bipolar disorder is divided into slow cycling (fewer than four episodes a year) and rapid cycling (more than four episodes) forms. Rapid cycling bipolar disorder (RCBD) tends to be more common in women than in men. Bipolar disorder in which the switch happens multiple times a day has been called ultra rapid cycling.

Table 3 Diagnostic criteria for manic episode

A. A distinct period of abnormally and persistently elevated, expansive, or irritable mood, lasting at least 1 week (or any duration if hospitalization is necessary).

B. During the period of mood disturbance, three (or more) of the following symptoms have persisted (four if the mood is only irritable) and have been present to a significant degree:
   1. Inflated self-esteem or grandiosity
   2. Decreased need for sleep (e.g., feels rested after only 3 h of sleep)
   3. More talkative than usual or pressure to keep talking
   4. Flight of ideas or a subjective experience that thoughts are racing
   5. Distractibility (i.e., attention too easily drawn to unimportant or irrelevant external stimuli)
   6. Increase in goal-directed activity (either socially, at work or school, or sexually) or psychomotor agitation
   7. Excessive involvement in pleasurable activities that have a high potential for painful consequences (e.g., engaging in unrestrained buying sprees, sexual indiscretions, or foolish business investments)

C. The symptoms do not meet criteria for a Mixed Episode.

D. The mood disturbance is sufficiently severe to cause marked impairment in occupational functioning or in usual social activities or relationships with others, or to necessitate hospitalization to prevent harm to self or others, or there are psychotic features.

E. The symptoms are not due to the direct physiological effects of a substance (e.g., a drug of abuse, a medication, or other treatment) or a general medical condition (e.g., hyperthyroidism).

Note: Manic-like episodes that are clearly caused by somatic antidepressant treatment (e.g., medication, electroconvulsive therapy, and light therapy) should not count toward a diagnosis of Bipolar I Disorder.

Source: American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington DC: American Psychiatric Association.

Sleep disturbance in bipolar depression It has been suggested that bipolar depressed patients have hypersomnia, characterized by increased duration of sleep during the night and excessive daytime somnolence. Studies using measures such as the multiple sleep latency test (MSLT) found no signs of objective hypersomnia in bipolar depression as opposed to hypersomnia in obstructive sleep apnea (OSA) or narcolepsy.

PSG studies of sleep in bipolar depression have generally found results that are no different from those in unipolar depression (Figure 8). A recent review of PSG studies, seven in total, found no difference between unipolar and bipolar participants. The authors went on to reanalyze their own data of unipolar and bipolar patients, exactly matched for age, gender, and severity of depression, and found that bipolar depressed patients had more pronounced sleep abnormalities. They showed significant decrease in sleep efficiency, increase in number of wake periods, longer duration of early morning awakenings, and increased REM density, both in the first REM period and during the total duration of REM.

Sleep disturbances in mania There are few studies of sleep disturbance in manic patients compared to those with depression because of the nature of the disorder, the difficulty in getting consent for a study, and the ethical issues involved in keeping patients unmedicated for the study period. Subjectively, manic patients report reduced need for sleep, with few hours of sleep at night and no deleterious effect on daytime functioning or energy. Some patients have total insomnia during the manic phase.

Studies using PSG in unmedicated patients have found more pronounced sleep continuity disturbances in mania compared to depressives. REM sleep findings were similar to depression, but the SWS was not found to be reduced.

Circadian rhythm disturbances in bipolar disorder Among the different types of bipolar disorder, rapid cycling bipolar disorder (RCBD) has been most associated with abnormality of the circadian rhythm. In patients with RCBD, the decrease in sleep duration heralds the onset of mania or the switch from depression into mania (Figure 9). Decreased sleep duration is more consistently associated with a shift to an earlier wake-up time. Diurnal variation, which is discussed prominently in depression literature, is also seen in RCBD. The patients with RCBD tend to switch up during the day from depression or euthymia into hypomania or mania, and switch down from mania or hypomania into depression or euthymia after overnight sleep. It is also observed in a handful of cases that DLMO is phase delayed early in the course of a depression episode in patients with RCBD. The phase advances slowly as the illness progresses, and the most advanced position is reached prior to a switch to hypomania or mania.

Social rhythm in bipolar disorder The social *zeitgeber* hypothesis mentioned earlier in the case of MDE holds good for bipolar disorder too (Figure 6). It has

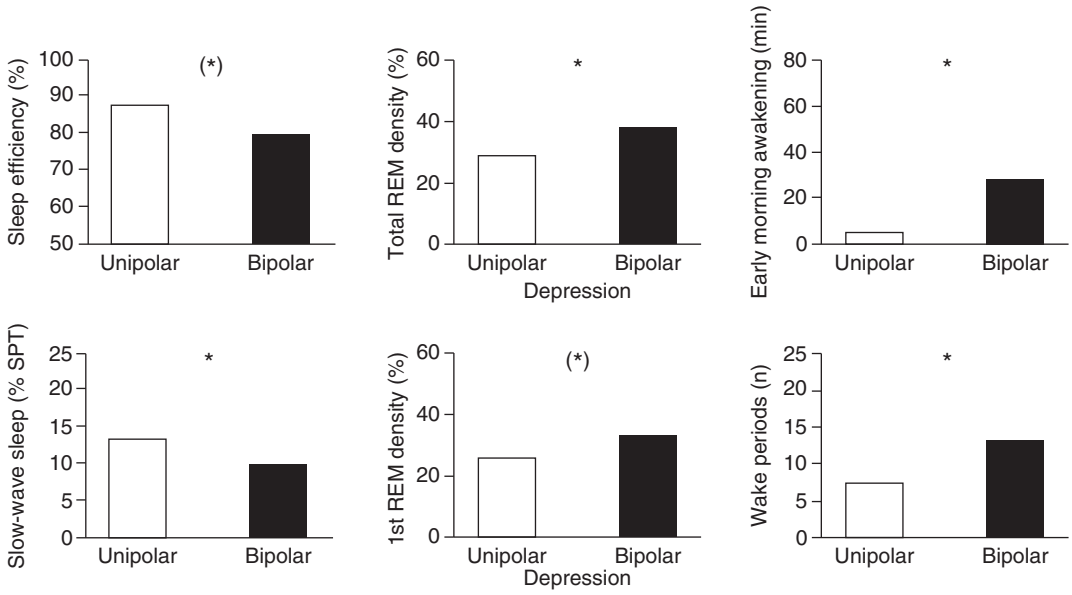

Figure 8 Polysomnographically recorded sleep variables in unipolar vs. bipolar depression ($n = 27$ patients per group). (*) $p < 0.10$, * $p < 0.05$. SPT; Sleep period time. From Riemann D, Voderholzer U, and Berger M (2002) Sleep and sleep–wake manipulations in bipolar depression. *Neuropsychobiology* 45(supplement 1): 9. Used with permission from Karger.

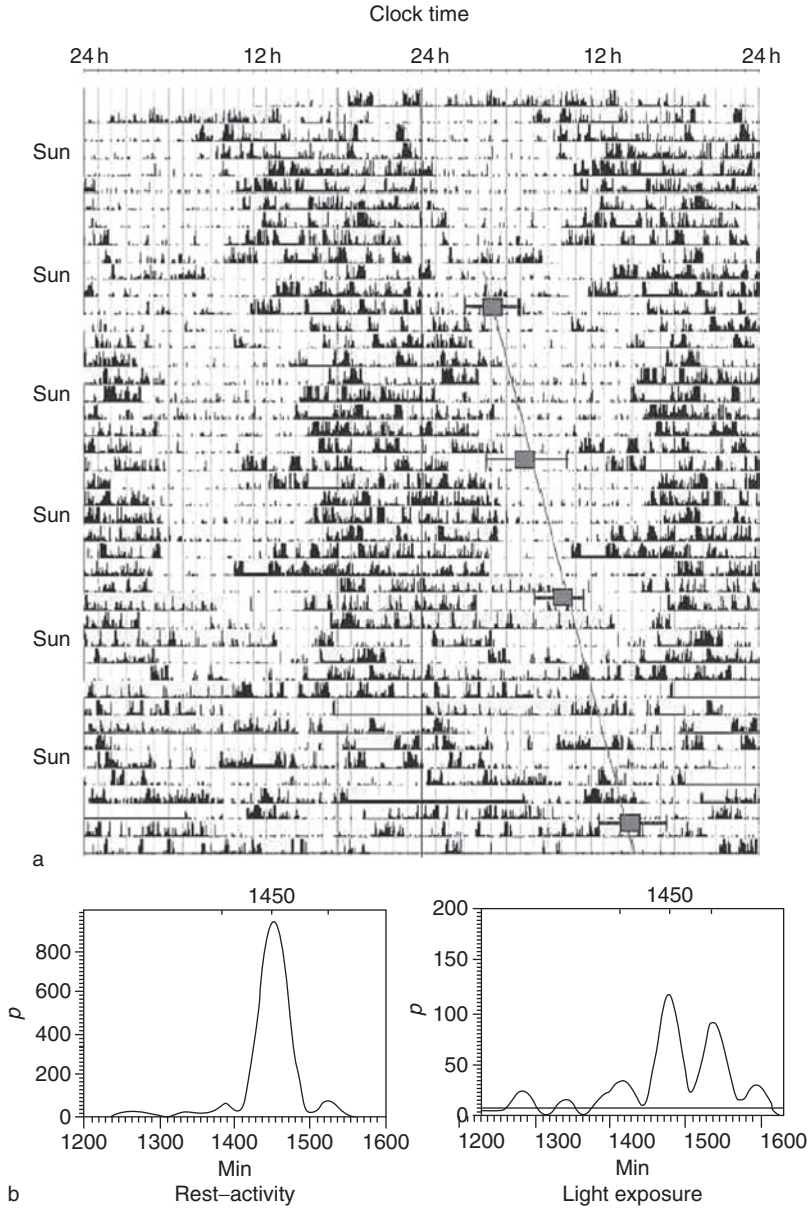

Figure 9 (a) Double-plotted actogram (consecutive days next and beneath each other) of a schizophrenic patient (27-year-old man), whose rest–activity patterns reveal extremely delayed rest and activity onsets, reversed day–night activities, and remarkably prolonged activity and rest phases. This person's weekly sleep onset times delayed from 00:00 to 06:40 h between week 1 and week 5, and his wake-up times delayed from 10:38 to 16:52 h. On average, he slept about 8:30 h (SD + 2:43 h) per 24 h, but sleep efficiency was low (75.9%, SD + 12.9%). Light–dark exposure (gray) also coincides with his rest–activity behavior, indicating an irregular light input to the clock. The phase of his melatonin peak (dark gray squares with 95% confidence interval) changes in synchrony with activity. Abscissa 1/4 days of measurements, Ordinate: clock time in hours. (b) Lomb-Scargle periodograms of the above data, detecting the dominant period at 1450 min (24.17 h) for both activity–rest and light–dark cycles. *p*, power; — indicates missing data. Medication: olanzapine, sodium valproate. From Wulff K, Joyce E, Middleton B, Dijk D-J, and Foster RJ (2006) The suitability of actigraphy, diary data, and urinary melatonin profiles for quantitative assessment of sleep disturbances in schizophrenia: A case report. *Chronobiology International* 23: 489. Used with permission from Taylor & Francis.

been noted that disruptive events (*zeitstorer's*) happen more frequently prior to manic or hypomanic episodes when compared with euthymic periods. In RCBD, significantly decreased social rhythm has been reported. Reduced social rhythm has been found in groups of adolescents who were at high risk of developing bipolar disorder in a sample of bipolar II patients compared to a control group. In a recent study conducted in Germany that examined social rhythm instabilities in adolescents and young adults at risk for affective disorders, participants from the bipolar risk group were found to have

reduced social rhythm scores as measured by the Social Rhythm Metric when compared with a control group. The Social Rhythm Metric quantifies the stability of an individual's daily routine and lists activities that are identified as central to everyday life. The group with risk for unipolar depression defined by the high rigidity scores from the rigidity subscale of the Munich Personality Test did not show reduced social rhythm scores.

Schizophrenia

The diagnostic criteria for schizophrenia are shown in Table 4.

Sleep Disturbance in Schizophrenia

Schizophrenia is a chronic psychotic disorder characterized by delusions, hallucinations, disorganized speech, grossly disorganized behavior, and negative symptoms such as affective flattening, alogia, and avolition.

The sleep of the schizophrenic is characterized by increased latency and increased wake time after sleep onset. The latter may be a sign of impending relapse of the psychosis. In a study of 145 patients with schizophrenia, patients who reported poor subjective sleep quality had lower perceived quality of life, more depressive symptoms, more psychological distress, and adverse events to medications. Some schizophrenics complain of total or near total insomnia although PSGs document objective sleep. The subjective perception of insomnia may be incorporated by schizophoenic patients into their delusional belief system.

Patients with schizophrenia show a decrease in continuity of sleep, SWS, REM sleep, and reduced REM latency. A recent review of never-medicated or previously medicated schizophrenics found that the majority of the studies reported reduced SWS and REM latency

Table 4 Diagnostic criteria for schizophrenia

A. *Characteristic symptoms:* Two (or more) of the following, each present for a significant portion of time during a 1-month period (or less if successfully treated):
1. Delusions
2. Hallucinations
3. Disorganized speech (e.g., frequent derailment or incoherence)
4. Grossly disorganized or catatonic behavior
5. Negative symptoms, i.e., affective flattening, alogia, or avolition

Note: Only one criterion A symptom is required if delusions are bizarre or hallucinations consist of a voice keeping up a running commentary on the person's behavior or thoughts, or two or more voices conversing with each other.

B. *Social/occupational dysfunction:* For a significant portion of the time since the onset of the disturbance, one or more major areas of functioning such as work, interpersonal relations, or self-care are markedly below the level achieved prior to the onset (or when the onset is in childhood or adolescence, failure to achieve expected level of interpersonal, academic, or occupational achievement).

C. *Duration:* Continuous signs of the disturbance persist for at least 6 months. This 6-month period must include at least 1 month of symptoms (or less if successfully treated) that meet criterion A (i.e., active-phase symptoms) and may include periods of prodromal or residual symptoms. During these prodromal or residual periods, the signs of the disturbance may be manifested by only negative symptoms or two or more symptoms listed in criterion A present in an attenuated form (e.g., odd beliefs, unusual perceptual experiences).

D. *Schizoaffective and mood disorder exclusion:* Schizoaffective Disorder and Mood Disorder with Psychotic Features have been ruled out because either (1) no major depressive, manic, or mixed episodes have occurred concurrently with the active-phase symptoms; or (2) if mood episodes have occurred during active-phase symptoms, their total duration has been brief relative to the duration of the active and residual periods.

E. *Substance/general medical condition exclusion:* The disturbance is not due to the direct physiological effects of a substance (e.g., a drug abuse or a medication) or a general medical condition.

F. *Relationship to a pervasive developmental disorder:* If there is a history of Autistic Disorder or another Pervasive Developmental Disorder, the additional diagnosis of schizophrenia is made only if prominent delusions or hallucinations are also present for at least a month (or less if successfully treated).

Classification of longitudinal course (can be applied only after at least 1 year has elapsed since the initial onset of active-phase symptoms):

Episodic with Interepisode Residual Symptoms (episodes are defined by the reemergence of prominent psychotic symptoms); also specify if with Prominent Negative Symptoms

Episodic with No Interepisode Residual Symptoms

Continuous (prominent psychotic symptoms are present throughout the period of observation); also specify if with Prominent Negative Symptoms

Single Episode in Partial Remission; also specify if with Prominent Negative Symptoms

Single episode in Full Remission

Other or Unspecified Pattern

Source: American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington, DC: American Psychiatric Association.

together with unchanged duration of REM sleep. The never-medicated group had increased latency to stage 2, wake time after sleep onset (WASO), and total number of awakenings and reduced total sleep time (TST) and sleep efficiency (SE), with pronounced decrease in REM latency but unchanged REM duration. Similar findings were found for patients in first episode or acute exacerbations. Two studies did not find a decrease in SWS. Findings for chronic schizophrenics were also similar, except for a significant increase in the latency to stage 2 sleep. To sum up, sleep in patients with schizophrenia, irrespective of exposure to psychotropics or clinical course, shows initial and middle insomnia and reduced SWS.

Reduced SWS is consistently seen in studies using quantitative methods of EEG analysis but not always in studies using visual. SWS decrease is associated with the negative symptoms and has been consistently replicated in many studies. It is stable over the course of the disease (1 year follow-up) and predicts poor outcome at years 1 and 2. It also is associated with impaired attention in schizophrenia. The findings regarding REM sleep are harder to interpret, as studies have variously reported an increase, decrease, and no change in the total REM sleep. Only previously treated patients show an increase in total REM, which may be due to the effects of withdrawal from medication or related to the state of the disease (acute exacerbation). Two studies have noted an association between increased REM sleep and suicidality in schizophrenia. The negative symptoms have also been inversely associated with the REM SL.

Methodological Issues

Most studies follow a small number of schizophrenic patients who are heterogeneous regarding the course and subtype of the illness. The control groups include patients with neurological disorders or other medical disorders. The use of differing criteria for diagnosis and lack of clarity regarding the process of diagnosis makes it difficult to compare results across studies. Some studies have not used the latest scoring criteria for sleep, thus putting the validity of findings at risk. The age and gender of the patients also influence the results (increased incidence of OSA or periodic limb movement disorder (PLMD) with age or insomnia in women). Many studies fail to rule out the primary sleep disorder, thereby leading to abnormal findings falsely attributed to schizophrenia. Patients exposed to psychotropics, especially on a chronic basis, have changes in sleep continuity, NREM, and REM sleep that can act as confounding factors. Another factor noted is the first-night effect that reduces the REM and SWS sleep, particularly stage 4, leading to a lack of difference between patient and control group. This is obvious when the values of the control group are compared with similar controls.

Circadian Rhythm Disturbance in Schizophrenia

The studies in this area are few, follow a small number of patients, and have methodological issues as already described in the section on PSG. Studies of schizophrenia have not used techniques such as temporal isolation or constant routine. Body temperature and hormonal rhythms generally show a pattern similar to that of the normal subjects, as opposed to the findings in MDE. The cortisol levels in one study showed higher levels in early sleep but a normal overall pattern. Growth hormone secretion shows inconsistent results across studies, with both normal and decreased values reported.

A recent case report of a 27-year-old man with schizophrenia (monitored by wrist actigraphy, light detection, diary records, and urinary 6-sulfatoxymelatonin (aMT6s) measurements) showed phase shifts, highly delayed sleep-onset and offset, and irregular rest activity phases (Figure 9). The period of the rest activity rhythm, melatonin rhythm, and light–dark (L-D) cycle were longer than 24 h. In another study, the second-generation antipsychotic (SGA), clozapine, led to highly ordered rest–activity cycles, when compared to first-generation antipsychotics (FGA) (e.g., haloperidol; Figure 10). This was the first study that reported the differential effect of FGA and SGA on the circadian rhythm.

Anxiety Disorders

The anxiety disorders that are described in the DSM-IV-TR are panic disorder, phobias (specific and social), obsessive compulsive disorder (OCD), PTSD, and GAD. Among these disturbed sleep is mentioned as a criterion for PTSD and GAD. Epidemiological studies have noted the links between insomnia and anxiety disorders. The treatments, pharmacologic and psychotherapeutic, used for anxiety disorders are also useful for insomnia, thus implying a common basis for these disorders.

Sleep and Circadian Rhythm Disturbance in Panic Disorder

Panic disorder is characterized by recurrent panic attacks and anticipatory anxiety, together with avoidance (Tables 5 and 6). It may present with or without agoraphobia. Sleep is of poor quality, interrupted, or restless as reported in surveys of patients with panic disorder when compared to controls. Among patients with panic disorder, 68% have difficulty falling

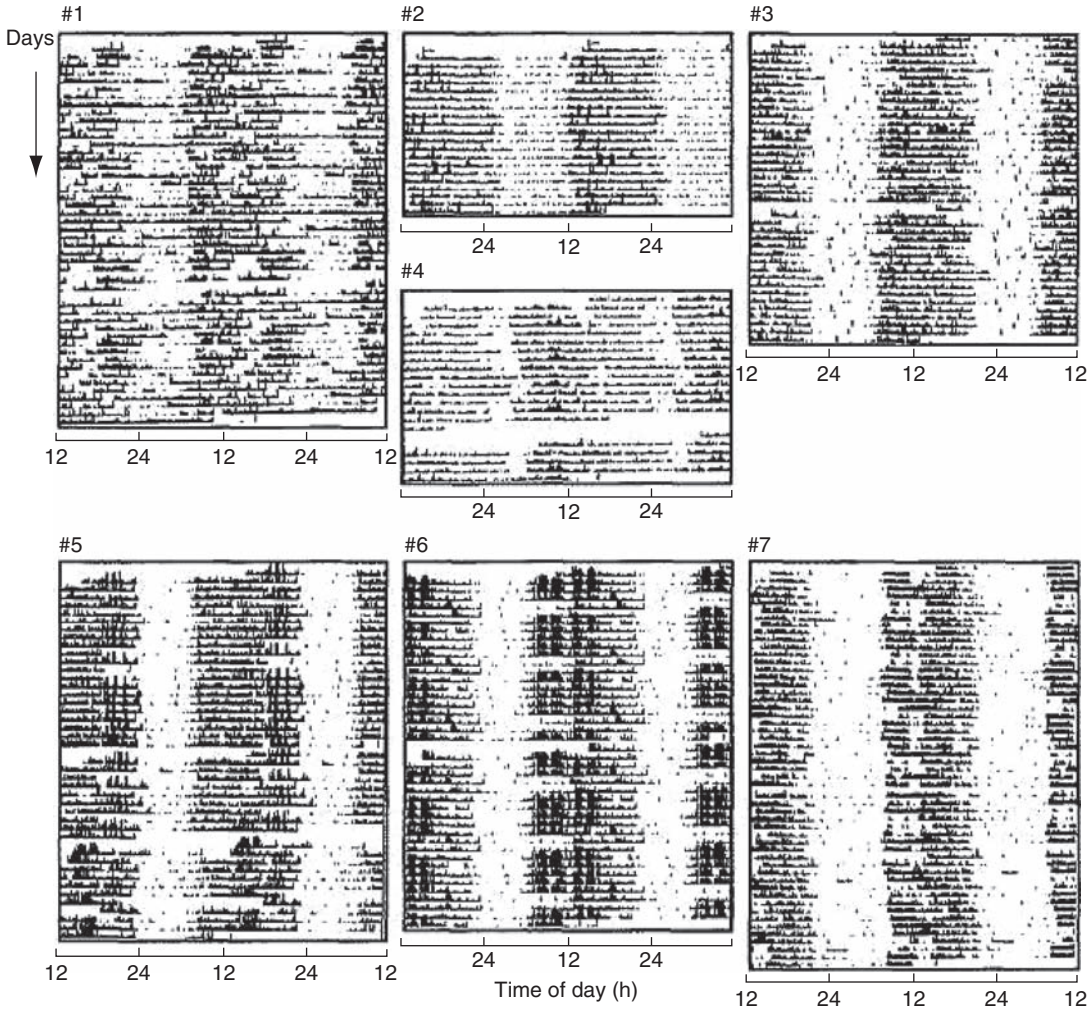

Figure 10 Actograms of seven schizophrenia patients on monotherapy for at least a year. All patients were hospitalized except patient 3 (day patient) and patient 4 (ambulatory). Patients 1–4 were treated with classical neuroleptics (haloperidol or flupentixol); patients 5–7 were treated with clozapine. Motor activity is registered in black (1-min bins) and graphed as a double 24 h plot over 48 h (first line, days 1 and 2, second line, days 2 and 3, etc.). Occasional missing data are left blank. From Wirz-Justice A, Haug H-J, and Cajochen C (2001) Disturbed circadian rest–activity cycles in schizophrenia patients: An effect of drugs? *Schizophrenia Bulletin* 27: 499. Used with permission from Oxford University Press.

asleep and 77% have restless sleep. Nocturnal panic attacks are reported by patients (33–71%) that resemble the daytime attacks. Patients report waking up "like a jolt" and experience a full-blown attack. Patients with nocturnal panic have early age of onset, comorbid depression, severe symptomatology, and increased suicidal ideation. It has been noted in case reports that panic disorder is associated with sleep-related breathing disorder.

PSGs show decreased sleep efficiency and total sleep duration. However, studies that excluded patients with comorbid depression did not show these changes. Normal amounts of REM and SWS are seen. The REM latency is reported either as normal or reduced. Increased movement time is also seen with panic disorder. The panic attacks appear while transitioning from stage 2 to SWS in the first half of night. Nocturnal panic attacks can be induced by infusing caffeine during SWS. Two groups have reported cardiopulmonary instability in patients with panic disorder, one by infusing lactate during sleep and the other by demonstrating greater variability in the heart rate during NREM sleep.

There have been few studies regarding panic disorder and circadian rhythm. Circadian rhythms of cortisol and adrenocosticotropin (ACTH) may be abnormal in panic disorder. Elevated nocturnal cortisol secretion has been reported, as well as increased ultradian secretory episodes. The CRF has variable phase and changes in mean level has been noted.

Table 5 Diagnostic criteria for panic attack

Note: A Panic Attack is not a codable disorder. Code the specific diagnosis in which the panic attack occurs (e.g., **300.21** Panic Disorder with Agoraphobia).

A discrete period of intense fear or discomfort, in which four (or more) of the following symptoms developed abruptly and reached a peak within 10 min.

1. Palpitations, pounding heart, or accelerated heart rate
2. Sweating
3. Trembling or shaking
4. Sensations of shortness of breath or smothering
5. Feeling of choking
6. Chest pain or discomfort
7. Nausea or abdominal distress
8. Feeling dizzy, unsteady, lightheaded, or faint
9. Derealization (feelings of unreality) or depersonalization (being detached from oneself)
10. Fear of losing control or going crazy
11. Fear of dying
12. Paresthesias (numbness or tingling sensations)
13. Chills or hot flushes

Source: American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington, DC: American Psychiatric Association.

Table 6 Diagnostic criteria for panic disorder

Diagnostic criteria for panic disorder without agoraphobia

A. Both (1) and (2):
   1. Recurrent unexpected Panic Attacks.
   2. At least one of the attacks has been followed by 1 month (or more) of one (or more) of the following:
      a. Persistent concern about having additional attacks
      b. Worry about the implications of the attack or its consequences (e.g., losing control, having a heart attack, 'going crazy')
      c. A significant change in behavior related to the attacks

B. Absence of Agoraphobia.

C. The Panic Attacks are not due to the direct physiological effects of a substance (e.g., a drug or abuse a medication) or a general medical condition (e.g., hyperthyroidism).

D. The Panic Attacks are not better accounted for by another mental disorder, such as Social Phobia (e.g., occurring on exposure to feared social situations), Specific Phobia (e.g., on exposure to a specific phobic situation), Obsessive–Compulsive Disorder (e.g., on exposure to dirt in someone with an obsession about contamination), Posttraumatic Stress Disorder (e.g., in response to stimuli associated with a severe stressor), or Separation Anxiety Disorder (e.g., in response to being away from home or close relatives).

Diagnostic criteria for panic disorder with agoraphobia

A. Both (1) and (2):
   1. Recurrent unexpected Panic Attacks.
   2. At least one of the attacks has been followed by 1 month (or more) of one (or more) of the following:
      a. Persistent concern about having additional attacks
      b. Worry about the implications of the attack or its consequences (e.g., losing control, having a heart attack, and 'going crazy')
      c. A significant change in behavior related to the attacks

B. The presence of Agoraphobia.

C. The Panic Attacks are not due to the direct physiological effects of a substance (e.g., a drug abuse or a medication) or a general medical condition (e.g., hyperthyroidism).

D. The Panic Attacks are not better accounted for by another mental disorder, such as Social Phobia (e.g., occurring on exposure to feared social situations), Specific Phobia (e.g., on exposure to a specific phobic situation), Obsessive–Compulsive Disorder (e.g., on exposure to dirt in someone with an obsession about contamination), Posttraumatic Stress Disorder (e.g., in response to stimuli associated with a severe stressor), or Separation Anxiety Disorder (e.g., in response to being away from home or close relatives).

Source: American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington, DC: American Psychiatric Association.

Sleep and Circadian Rhythm Disturbance in Social and Specific Phobia

Social and specific phobia are both characterized by fear and avoidance (Tables 7 and 8). Subjective reports show poor sleep quality, daytime impairment, and increased SL, but these findings were not reported by one study that used PSG. In another study, sleep architecture was no different between patients having MDE with or without phobias. The TST, SL, and SE were similar between patients and controls. The measure of REM SL, density, duration, and distribution was also normal in phobias.

Table 7 Diagnostic criteria for specific phobia (formerly known as simple phobia)

A. Marked and persistent fear that is excessive or unreasonable, cued by the presence or anticipation of a specific object or situation (e.g., flying, heights, animals, receiving an injection, seeing blood).
B. Exposure to the phobic stimulus almost invariably provokes an immediate anxiety response, which may take the form of a situationally bound or situationally predisposed Panic Attack.
Note: In children, the anxiety may be expressed by crying, tantrums, freezing, or clinging.
C. The person recognizes that the fear is excessive or unreasonable.
Note: In children, this feature may be absent.
D. The phobic situation(s) is avoided or else is endured with intense anxiety or distress.
E. The avoidance, anxious anticipation, or distress in the feared situation(s) interferes significantly with the person's normal routine, occupational (or academic) functioning, or social activities or relationships, or there is marked distress about having the phobia.
F. In individuals under age 18 years, the duration is at least 6 months.
G. The anxiety, Panic Attacks, or phobic avoidance associated with the specific object or situation, are not better accounted for by another mental disorder, such as Obsessive–Compulsive Disorder (e.g., fear of dirt in someone with an obsession about contamination), Posttraumatic Stress Disorder (e.g., avoidance of stimuli associated with a severe stressor), Separation Anxiety Disorder (e.g., avoidance of school), Social Phobia (e.g., avoidance of social situations because of fear of embarrassment), Panic Disorder With Agoraphobia, or Agoraphobia without History of Panic Disorder.

Specify type:
Animal Type
Natural Environment Type (e.g., heights, storms, and water)
Blood-Injection-Injury Type
Situational Type (e.g., airplanes, elevators, and enclosed places)
Other Type (e.g., fear of choking, vomiting, or contracting an illness; in children, fear of loud sounds or costumed characters)

Source: American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington, DC: American Psychiatric Association.

Table 8 Diagnostic criteria for social phobia (social anxiety disorder)

A. A marked and persistent fear of one or more social or performance situations in which the person is exposed to unfamiliar people or to possible scrutiny by others. The individual fears that he or she will act in a way (or show anxiety symptoms) that will be humiliating or embarrassing. *Note*: In children, there must be evidence of the capacity for age-appropriate social relationships with familiar people and the anxiety must occur in peer settings, not just in interactions with adults.
B. Exposure to the feared social situation almost invariably provokes anxiety, which may take the form of a situationally bound or situationally predisposed Panic Attack. *Note*: In children, the anxiety may be expressed by crying, tantrums, freezing, or shrinking from social situations with unfamiliar people.
C. The person recognizes that the fear is excessive or unreasonable. *Note*: In children, this feature may be absent.
D. The feared social or performance situations are avoided or else are endured with intense anxiety or distress.
E. The avoidance, anxious anticipation, or distress in the feared social or performance situation(s) interferes significantly with the person's normal routine, occupational (academic) functioning, or social activities or relationships, or there is marked distress about having the phobia.
F. In individuals under age 18 years, the duration is at least 6 months.
G. The fear or avoidance is not due to the direct physiological effects of a substance (e.g., a drug abuse or a medication) or a general medical condition and is not better accounted for by another mental disorder (e.g., Panic Disorder with or without Agoraphobia, Separation Anxiety Disorder, Body Dysmorphic Disorder, a Pervasive Developmental Disorder, or Schizoid Personality Disorder).
H. If a general medical condition or another mental disorder is present, the fear in criterion A is unrelated to it, e.g., the fear is not of Stuttering, trembling in Parkinson's disease, or exhibiting abnormal eating behavior in Anorexia Nervosa or Bulimia Nervosa.

Specify if:
Generalized: if the fears include most social situations (also consider the additional diagnosis of Avoidant Personality Disorder)

Source: American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington, DC: American Psychiatric Association.

Sleep and Circadian Rhythm Disturbance in Generalized Anxiety Disorder

GAD is characterized by excessive worry (Table 9). Patients have sleep disturbance as one of the criterion in the DSM-IV-TR, namely initial or middle insomnia with restless or unsatisfying sleep. The latter occurs in 60–70% of patients. Patients may report a variety of sleep complaints clinically. PSG studies show sleep initiation and maintenance difficulties. NREM stage 2 is increased while SWS is decreased in GAD. GAD patients do not show the decreased REM latency seen in MDE. This may help differentiate one from the other. Patients with GAD have lower levels of activity as documented by SRM (social rhythm metric) in one study.

Table 9 Diagnostic criteria for generalized anxiety disorder (includes overanxious disorder of childhood)

A. Excessive anxiety and worry (apprehensive expectation), occurring more days than not for at least 6 months, about a number of events or activities (such as work or school performance).
B. The person finds it difficult to control the worry.
C. The anxiety and worry are associated with three (or more) of the following six symptoms (with at least some symptoms present for more days than not for the past 6 months). *Note:* Only one item is required in children.
   1. Restlessness or feeling keyed up or on edge
   2. Being easily fatigued
   3. Difficulty concentrating or mind going blank
   4. Irritability
   5. Muscle tension
   6. Sleep disturbance (difficulty falling or staying asleep, or restless unsatisfying sleep)
D. The focus of the anxiety and worry is not confined to features of an Axis I disorder, e.g., the anxiety or worry is not about having a Panic Attack (as in Panic Disorder), being embarrassed in public (as in Social Phobia), being contaminated (as in Obsessive–Compulsive Disorder), being away from home or close relatives (as in Separation Anxiety Disorder), gaining weight (as in Anorexia Nervosa), having multiple physical complaints (as in Somatization Disorder), or having a serious illness (as in Hypochondriasis), and the anxiety and worry do not occur exclusively during Posttraumatic Stress Disorder.
E. The anxiety, worry, or physical symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.
F. The disturbance is not due to the direct physiological effects of a substance (e.g., a drug abuse or a medication) or a general medical condition (e.g., hyperthyroidism) and does not occur exclusively during a Mood Disorder, a Psychotic Disorder, or a Pervasive Developmental Disorder.

Source: American Psychiatric Association (2000): *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington DC: American Psychiatric Association.

Table 10 Diagnostic criteria for posttraumatic stress disorder

A. The person has been exposed to a traumatic event in which both of the following were present:
   1. The person experienced, witnessed, or was confronted with an event or events that involved actual or threatened death or serious injury, or a threat to the physical integrity of self or others
   2. The person's response involved intense fear, helplessness, or horror. *Note:* In children, this may be expressed instead by disorganized or agitated behavior.
B. The traumatic event is persistently reexperienced in one (or more) of the following ways:
   1. Recurrent and intrusive distressing recollections of the event, including images, thoughts, or perceptions. *Note:* In young children, repetitive play may occur in which themes or aspects of the trauma are expressed.
   2. Recurrent distressing dreams of the event. *Note:* In children, there may be frightening dreams without recognizable content.
   3. Acting or feeling as though the traumatic event were recurring (includes a sense of reliving the experience, illusions, hallucinations, and dissociative flashback episodes, including those that occur on awakening or when intoxicated). *Note:* In young children, trauma-specific reenactment may occur.
   4. Intense psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event
   5. Physiological reactivity on exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event
C. Persistent avoidance of stimuli associated with the trauma and numbing of general responsiveness (not present before the trauma), as indicated by three (or more) of the following:
   1. Efforts to avoid thoughts, feelings, or conversations associated with the trauma
   2. Efforts to avoid activities, places, or people that arouse recollections of the trauma
   3. Inability to recall an important aspect of the trauma
   4. Markedly diminished interest or participation in significant activities
   5. Feeling of detachment or estrangement from others
   6. Restricted range of affect (e.g., unable to have loving feelings)
   7. Sense of a foreshortened future (e.g., does not expect to have a career, marriage, children, or a normal life span)
D. Persistent symptoms of increased arousal (not present before the trauma), as indicated by two (or more) of the following:
   1. Difficulty falling or staying asleep
   2. Irritability or outbursts of anger
   3. Difficulty concentrating
   4. Hypervigilance
   5. Exaggerated startle response
E. Duration of the disturbance (symptoms in criteria B, C, and D) is more than 1 month.
F. The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

Specify if:
Acute: duration of symptoms is less than 3 months
Chronic: duration of symptoms is 3 months or more
Specify if:
With Delayed Onset: onset of symptoms is at least 6 months after the stressor

Source: American Psychiatric Association (2000): *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., Text Revision. Washington DC: American Psychiatric Association.

Sleep and Circadian Rhythm Disturbance in Posttraumatic Stress Disorder

PTSD is characterized by persistent reexperiencing of the trauma, with avoidance and hyperarousal (Table 10). Nightmares and initial and middle insomnia are reported subjectively. Nightmares are found in 59–68% of patients. There are anxious arousals that occur in relation to both REM and NREM sleep. The arousals lead to excessive motor activity and awakenings. PSG documents decreased sleep efficiency and decreased TST with increased movement time. REM sleep in PTSD is reported as resembling depression in some studies. Reduced or prolonged latency is shown, depending on the studies reviewed, and reduced duration of REM sleep. Comorbid MDE, SRBD (sleep-related breathing disorder), or SUD (substance use disorders) may confound results. Although nightmares are reported in both REM and NREM sleep, the majority of studies support an association with REM sleep. Increased REM density has been described in combat veterans with chronic PTSD.

Patients with PTSD have increased excretion of 24 h catecholamines. The ratio of nocturnal to diurnal urinary catecholamines is increased relative to controls. The hypersecretion of cortisol during the night is not seen in the PTSD, as opposed to patients with depression.

Conclusion

There is strong evidence that disturbances of sleep are central to many psychiatric disorders. The strongest evidence is with mood disorders, both unipolar and bipolar. The REM and SWS abnormalities noted are biologic markers, both state and trait. They are also strong predictors of response to treatment and outcome in MDE. SAD is increasingly being recognized, and the new treatment modalities that are being developed will go a long way in treating and preventing it. Treatments of bipolar disorder, which take into account the circadian component, would bolster the current therapies. Studies of sleep in schizophrenia help us understand the disease and improve the quality of life in patients by managing sleep and circadian rhythms optimally. Anxiety disorders present hand in hand with sleep disturbance, and findings regarding sleep help us delineate anxiety from depression and devise better management strategies.

See also: Anxiety Disorders; Circadian Rhythm Models; Circadian Rhythms: Influence of Light in Humans; Depression and the Brain; Genetics of Circadian Disorders in Humans; Posttraumatic Stress Disorder: Neurobiology; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Sleep Deprivation and Brain Function; Sleep Deprivation: Neurobehavioral Changes.

Further Reading

Alesci S, Martinez PE, Kelkar S, et al. (2005) Major depression is associated with significant diurnal elevations in plasma interlukin-6 levels, a shift of its circadian rhythm, and loss of physiological complexity of its secretion. *Journal of Clinical Endocrinology and Metabolism* 90: 2522–2530.

Benca RM (2005) Mood disorders. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, 4th edn., pp. 1311–1327. Elsevier Saunders: Philadelphia.

Boivin DB (2000) Influence of sleep–wake and circadian rhythm disturbances in psychiatric disorders. *Journal of Psychiatry and Neuroscience* 25: 446–458.

Buysse DJ, Nofzinger EA, Keshavan MS, Reynolds III CF, and Kupfer DJ (1999) Psychiatric disorders associated with disturbed sleep and circadian rhythms. In: Turek FW and Zee PC (eds.) *Regulation of Sleep and Circadian Rhythms*, pp. 597–641. New York: Marcel Dekker.

Halaris A (1987) *Chronobiology and Psychiatric Disorders*. New York: Elsevier.

Leibenluft E and Frank E (2001) Circadian rhythm in affective disorders. *Handbook of Behavioral Neurobiology* 12: 625–644.

Lewy AJ, Lefler BJ, Emens SJ, and Bauer VK (2006) The circadian basis of winter depression. *Proceedings of the National Academy of Sciences of the United States of America* 103: 7414–7419.

Mellman TA (2003) Sleep aspects of anxiety disorders. In: Nutt DJ and Ballenjer JC (eds.) *Anxiety Disorders*, pp. 125–133. Cambridge, Mass: Blackwell Publishing.

Meyer TD and Maire S (2006) Is there evidence for social rhythm instability in people at risk of affective disorders? *Psychiatry Research* 141: 103–114.

Monti JM and Monti D (2005) Sleep disturbance in schizophrenia. *International Review of Psychiatry* 17: 247–243.

Nofzinger EA (2005) Functional neuroimaging of sleep. *Seminars in Neurology* 25: 9–18.

Nofzinger EA and Keshavan MA (2002) Sleep disturbances associated with neuropsychiatric disease. In: Davis KL, Charney D, Coyle JT, and Nemeroff C (eds.) *Neuropsychopharmacology: The Fifth Generation of Progress*, pp. 1945–1959. Philadelphia: Lippincott Williams & Wilkins.

Stetler C, Dickerson SS, and Miller GE (2004) Uncoupling of social zeitgebers and diurnal cortisol secretion in clinical depression. *Psychoneuroendocrinology* 29: 1250–1259.

Wehr TA and Goodwin FK (1983) Circadian rhythms in psychiatry. *Psychobiology and Psychopathology*, vol. 2. Pacific Grove, CA: Boxwood Press.

Wirz-Justice A (1994) Biological rhythms in mood disorders. In: Bloom FE and Kupfer DJ (eds.) *Psychopharmacology: The Fourth Generation of Progress*, pp. 999–1017. New York: Raven Press.

Relevant Websites

http://www.aasmnet.org – American Academy of Sleep Medicine.
http://www.acnp.org/default.aspx?Page=Home – American College of Neuropsychopharmacology.
http://www.psychiatryonline.com – American Psychiatric Publishing, Inc.

http://www.northwestern.edu – Center for Sleep and Circadian Biology at Northwestern University.
http://www.hhmi.org – Howard Hughes Medical Institute.
http://www.sleepmedtext.com – *Principles and Practice of Sleep Medicine* (the authoritative text on sleep medicine).
http://www.npi.ucla.edu – Siegel Lab at the University of California Los Angeles.
http://www.wpic.pitt.edu – Sleep Neuroimaging Research Program, University of Pittsburgh Medical Center, Western Psychiatric Institute and Clinic, Pittsburgh.
http://www.sltbr.org – Society for Light Treatment and Biological Rhythms.
http://med.stanford.edu – Stanford University Center for Excellence for the Diagnosis and Treatment of Sleep Disorders.

Psychiatric Disorders: Functional Genomics

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

There are about 30 000 genes in the human genome, approximately 20 000 of which are expressed in the central nervous system and approximately 3000 of which are brain specific. Diseases and disorders of the nervous system, including psychiatric disorders, result from a combination of genetic and environmental effects, so it is reasonable to assume that many of the 20 000 genes expressed in the central nervous system (CNS), and most of the 3000 genes that are not expressed anywhere else, might be functionally linked to important neural processes whose disruption or perturbation could contribute to the disease phenotype or at least to a form of susceptibility when there is a significant environmental component. Furthermore, since the products encoded by these genes will have direct functional roles in the nervous system, it is also likely that many of them would represent useful biomarkers and/or drug targets.

The key to identifying genes involved in human disease, including psychiatric disorders, is to establish each gene's function and show how the disruption of its activity has a pathogenic effect. Through conventional methods, the gene–disease or gene–response relationship has been unraveled for only about 1500 genes, and in each case the path to enlightenment has been slow and laborious. The aim of functional genomics is to determine the functions of genes on a large scale using high-throughput screening technologies. The overall goal is to learn the precise relationships between genes, or more accurately between the proteins they encode, in coordinating the activities of the healthy brain.

Functional Genomics: Major Technology Platforms

Functional genomics is not a specific technique but an umbrella term to describe any approach for large-scale functional annotation of uncharacterized genes. Therefore, many different approaches can be used in functional genomics, ranging from those carried out completely *in silico* to complex, large-scale bench experiments in which many genes are analyzed simultaneously.

In Silico Methods

Computers are an essential component of all functional genomics technology platforms because of the need to handle, analyze, and store large data sets. However, it is possible to find out a large amount of information about the function of a gene without doing any bench experiments, and these bioinformatics-based approaches are usually the best place to start. The computer programs that are used to search for genes in genomic DNA often incorporate algorithms that attempt to find sequences matching known genes by applying the principle of similarity searching. Search algorithms such as BLAST, FASTA, and their derivatives allow new sequences to be compared with all the stored sequences in an attempt to find matches. It is important to note that the sequence databases contain not only sequences but also information that may relate to gene function. If the function of a human gene is unknown, it is often the case that a related gene has been studied in another species and that some information about its function is known. Therefore, the quickest way to establish the function of a new gene is often to search the databases and attempt to find related sequences that have already been annotated.

Functional annotation by sequence comparison can be applied on a whole genome scale, but this is not sufficient to determine the functions of all genes. For example, when the baker's yeast (*Saccharomyces cerevisiae*) genome was sequenced in 1996, 30% of the genes were already known, and functions had been assigned by bench experiments, a further 30% of the predicted genes could be assigned tentative functions on the basis of homology to other genes already in databases, and 10% were regarded as uncertain predictions. This left researchers with no hint as to the likely function of 30% of the genes, nearly 2000 in total. The functional annotation of the human genome is complicated by the greater difficulty in predicting genes accurately, and the proportion of uncharacterized genes in the human genome is therefore even higher, 40%. Based on the numbers cited at the beginning of the article, this would approximate to about 8000 genes expressed in the CNS and 1200 genes expressed solely in the brain. Many of these genes are likely to underpin susceptibility to diseases of the nervous system, including psychiatric disorders.

Standard search algorithms such as BLAST and FASTA can reliably detect protein sequences that are >30% identical, but below this level are many relationships that are not found. More-sophisticated search techniques based on patterns and profiles compiled from large protein families can help in some circumstances, but these methods are not always reliable. A relatively new approach is to use protein structures rather than sequences for functional annotation.

Protein structures are much more highly conserved than sequences, and it is possible to match protein structures showing less than 10% identity at the primary sequence level by the use of structural comparison algorithms such as DALI, VAST, and COMPARER. As for sequences, primary databases of protein structures have been solved by X-ray crystallography and/or nuclear magnetic resonance spectroscopy. Several international collaborative efforts are ongoing to solve protein structures in a systematic and high-throughput manner so that examples of every protein family in existence may be stored in the databases. In the future, it may be possible to identify the functions of uncharacterized genes by expressing the proteins, solving the protein structures, and comparing these to structures in the databases.

Another sequence-based method for functional annotation is comparative genomics. The principle of comparative genomics is that similarities between species extend far beyond the level of individual genes and include whole genomes. In closely related species (e.g., mouse and human), comparative genomics can help identify genes and their regulatory elements since only sequences with an evolutionarily conserved function would be found in both genomes whereas others would have diverged significantly. There is also a surprising degree of conservation at the level of whole molecular pathways and networks, even between distantly related organisms. An example is the extraordinary degree of conservation in hormone and neurotransmitter signaling pathways and the general processes involved in the development of the nervous system between mammals and invertebrate animals such as the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*, which are much more amenable to large-scale experiments. Indeed, nearly 60% of known human disease genes have counterparts in the worm and/or the fruit fly, and 30% of them have counterparts in yeast.

Expression Profiling

Many diseases are characterized by the dysregulation of gene expression. Traditional methods that look at the expression of single genes in an attempt to identify disease biomarkers are now being replaced with global profiling strategies which look at all genes simultaneously. In the context of functional genomics, expression profiling can help to determine a gene's function through the principle of guilt by association. If a gene is switched on in response to neurotransmitter release, then it is likely that the function of that gene has something to do with the generation, detection, propagation, or downstream impact of that neurotransmitter, and thus without any other information, it is possible to imply that the gene is involved in one of those functions. By studying all genes simultaneously and identifying those whose expression is modulated, it becomes possible to select them for further, more detailed experiments. In the case of psychiatric disorders, as with any other disturbance to the normal function of the body, useful information can be obtained by comparing the expression profiles of hundreds or thousands of genes in the disease state and the unaffected state, that is, by searching for differential expression specific for the disease.

Two major technologies are used in expression profiling, one based on the high-throughput sampling of complementary DNA (cDNA) sequences and one based on miniature devices called microarrays, which allow messenger RNA (mRNA) levels to be measured by hybridization. Sequence sampling is based on the principle that a strongly expressed gene will be abundantly represented in the mRNA population of a particular cell or tissue and will be similarly represented in an unbiased cDNA library or sequence database derived from that population. Therefore, random sampling will tend to pull out clones or sequences of strongly expressed genes much more frequently than those of poorly expressed genes. Relative expression levels can be determined simply by counting the score for each sequence pulled out during sampling, and because these data are in the form of real numbers, they are digital. A method called serial analysis of gene expression accelerates the sampling process by reducing each cDNA to a short tag 9–15 bp in length and allowing up to 200 such tags to be concatemerized prior to sequencing. Microarrays are miniature devices comprising a collection of immobilized DNA sequences arranged in a grid. Each address on the grid corresponds to a unique gene, or at least to a unique mRNA if the gene undergoes alternative splicing. These arrayed targets are hybridized (interrogated) with a complex mRNA or cDNA probe, that is, a probe comprising many different sequences representing the entire population of mRNAs from the cell or tissue under investigation. In this case, quantitative data are obtained, not by counting, but by measuring the strength of the hybridization signal at each address. If the source mRNA/cDNA probe is universally labeled, the most abundant mRNA/cDNA molecules in the population will generate quantitatively stronger signals on the array, as long as hybridization does not reach saturation. Both sequence sampling and microarrays can be used for comparative expression profiling, in one case by comparing the representation of different sequences in two different tissues (e.g., prefrontal cortex from normal brain and schizophrenic brain) and in the

other case by using two complex probes, labeled with different fluorophores, and comparing the signal intensity of each fluorophore for each gene.

Expression profiling is also possible at the protein level, although the procedure is more complex because proteins have a wider range of physicochemical properties than do DNA and RNA. Although all mRNA/cDNA molecules will hybridize to a microarray under the same conditions, it is more difficult to establish conditions in which all proteins in a sample will bind uniformly to, for example, an array of antibodies or other capture agents. Therefore, while protein microarrays have been developed, the more widely used technology is the display of proteins on two-dimensional (2D) gels. The proteome of a typical human cell is extremely complex, containing up to 100 000 different protein molecules (this takes into account the many different forms of posttranslational modification) differing in abundance over four or five orders of magnitude. Since there is no procedure for amplifying rare proteins equivalent to the polymerase chain reaction (PCR), expression profiling at the protein level will always lack sensitivity compared with mRNA profiling, a particular disappointment because the interesting proteins (regulators, signaling components) are often those that are least abundant, sometimes only a few molecules per cell. Because the proteome is so complex, a separation process based on one property (e.g., size, charge, solubility) would not provide sufficient resolution. The advantage of 2D gel electrophoresis is that proteins are separated orthogonally, that is, in one dimension on the basis of charge and then in the second dimension on the basis of mass. It is possible to run pairs of gels using different samples (e.g., schizophrenia vs. unaffected control) or to use the technique of difference gel electrophoresis, which, analogous to the DNA arrays, involves the use of two protein samples, universally labeled with different fluorophores, and coseparating them, then identifying protein spots where the two fluorophores have different intensities. There are also a number of ways to quantify proteins directly by mass spectrometry, and these can also be used to compare multiple samples. Such methods involve the use of pairs of labels carrying different isotopes that can be distinguished by a mass spectrometer. The heavy and light labels are applied to the different samples (e.g., healthy and schizophrenic brain tissue), and then the samples are mixed prior to protein isolation so that purification losses are equivalent. The proteins are separated by two-dimensional gel electrophoresis (2DGE) or high-performance liquid chromatography (HPLC) and the fractions analyzed by mass spectrometry, with each labeled peptide appearing as a doublet, one carrying each isotope. The relative amounts of the two isotopes are measured in each fraction, and a significant difference indicates that the protein in that fraction is either more abundant or underrepresented in the disease tissue.

Interaction Analysis

Another guilt-by-association method for ascertaining gene functions is to look at protein interactions directly. This can be useful when one gene is known to be involved in a particular psychiatric disorder but nothing is known about the pathway or regulatory hierarchy to which it contributes. The principle is that proteins physically interacting with this initial lead protein are almost certainly involved in the same function. It is also likely that interference with the activity of such interacting proteins may cause a similar spectrum of disease symptoms.

Methods such as affinity chromatography and coimmunoprecipitation have been used for many years to study the interactions of individual proteins. More recently, they have been used more systematically to identify the components of protein complexes, in some cases on a global scale. Complexes are isolated by lysing cells gently and using antibodies that recognize one of the complex's components to pull down the interactors. The complexes may then be analyzed directly by mass spectrometry or may first be separated into individual components by electrophoresis. This approach can define all the proteins that function in a complex and may thus result in the functional annotation of several uncharacterized genes. Furthermore, since many signaling pathways involve the formation of transient complexes, this is an excellent way to probe the roles of different signaling pathways in psychiatric disease.

Interaction screening on an even larger scale can be carried out by the yeast two-hybrid system, an *in vivo* assay in which a bait protein is expressed in yeast as a fusion (a hybrid) with the DNA-binding domain of a transcription factor. All the candidate interactor (prey) proteins are expressed as fusions with the transcription factor's transactivation domain. The bait yeast strain is mated either systematically or randomly to a prey library. The products of the mating are diploid cells containing the bait construct and one of the many prey constructs in the library. If the bait protein happens to interact with the prey, a functional transcription factor is assembled that can activate a test gene which is also introduced into the cell. However, if the bait and prey do not interact, the transcription factor remains as two separate hybrid proteins, and the gene remains inactive. Large-scale interaction screening using the yeast two-hybrid system has been used to study the interactions of many human proteins and has recently been carried out on the entire mouse proteome.

Mutational Genomics

All the functional genomics technologies discussed above help establish molecular/biochemical or cellular functions but do not provide direct links to disease. The direct approach is to mutate genes in model organisms, particularly the mouse, and look for phenotypes that resemble human psychiatric diseases. Disease models can be generated by sporadic mutations and by deliberate gene knockout techniques on individual targets, but more recently there has been increasing interest in carrying out whole-genome mutagenesis programs with the aim of generating comprehensive mutant libraries representing hits in every single gene. These can be used as a central resource so that when researchers find evidence from another experimental approach that a particular gene is involved in a psychiatric disease, the corresponding mutant can be 'ordered' and used to study the molecular pathology of the disease and to develop and test new therapies.

Whole genome mutagenesis (saturation mutagenesis) has been used for many years to recover mutants affecting particular biological processes. The difference in the functional genomics approach is that whereas in the past scientists discarded mutants that did not interest them, now every mutant is interesting. Recent mutagenesis screens in mice using the potent chemical mutagen ethylnitrosourea (ENU) involved the analysis of some 40 000 lines for medically relevant phenotypes, including allergy, immunology, physiology, developmental defects, and behavioral and clinical biochemistry, with the behavioral phenotypes being particularly relevant in terms of psychiatric disorders in humans.

In addition to chemical mutagens, further mouse mutagenesis programs using DNA-based mutagens have been initiated. These are DNA sequences that are introduced into the genome in a deliberate effort to disrupt genes and cause disease-like phenotypes. The advantage of this approach over chemical mutagenesis is that the interrupted gene is tagged with an alien sequence that can be detected by hybridization or PCR, which permits the rapid isolation of the gene and its identification.

One of the most popular current methods for functional analysis in cells and whole organisms is RNA interference (RNAi), in which double-stranded RNA is expressed in or introduced into a cell and processed into short interfering fragments that induce the rapid degradation of any mRNA with the same sequence. This method has been used to inactivate nearly 95% of the genes in the nematode, which as stated above contains relatives of 60% of the known human disease genes and provides useful models for a number of disorders. RNAi can also be used directly in human cells and indeed represents perhaps the only current method that allows the large-scale and direct analysis of the functions of human genes. Although RNAi has yet to be used specifically to study psychiatric disorders, it has been widely used to study gene function in fundamental neurobiological processes such as intracellular signaling, embryonic development, formation of cellular junctions, cell cycle, programmed cell death, neural plasticity, axon guidance, and membrane trafficking, all of which may contribute in some small way to the disease phenotype.

Case Studies

Expression Profiling

Expression profiling using sequence sampling and microarrays has been applied to the study of several psychiatric disorders, with particularly promising results from the analysis of schizophrenia, bipolar disorder, and manic depressive disorder (MDD).

In MDD, expression profiling has shown that levels of mRNA encoding neuropeptide Y (*NPY* mRNA) and mRNA encoding thyrotropin-releasing hormone (*TRH* mRNA) fall, but that the levels rise again in response to electroconvulsive therapy, correlating with the antidepressant effects of NPY and TRH injections in the forced-swim model.

Several common themes have been identified in schizophrenia, including the dysregulation of genes involved in energy metabolism, oligodendrocyte function, and the development and maintenance of synapses. The general downregulation of mitochondrial genes encoding enzymes and structural components of the respiratory chain has been replicated in several studies, but similar trends are seen in other disorders, including Alzheimer's disease and multiple sclerosis, indicating that reduced energy metabolism in schizophrenia may be secondary to another, primary defect.

The dysregulation of critical oligodendrocyte development and maturation genes is seen in the prefrontal cortex and other brain areas, which supports evidence from magnetic resonance imaging studies showing reduction in white matter volume and decreased oligodendrocyte density in schizophrenic patients. Another common feature of schizophrenia expression profiling studies is the dysregulation of genes with roles in the presynaptic release machinery in the prefrontal cortex. It is interesting that while many of the same transcripts have been identified in independent studies, it has been shown that the pattern of affected transcripts differed from individual to individual, suggesting diversity at the molecular level, which might underpin the similarly diverse spectrum of disease symptoms. It is also notable that many of

the transcripts underrepresented in disease tissue have been shown to increase in abundance in response to haloperidol or clozapine, drugs that are used to treat schizophrenia.

Protein expression profiling has also identified several genes involved in psychiatric disorders. Brain proteome analysis of mice selectively bred for either high or low anxiety-related behavior revealed quantitative and qualitative protein expression differences, with the enzyme glyoxalase-I consistently expressed to a higher extent in several brain areas in low-anxiety mice. Additionally, enolase phosphatase displayed a different mobility in 2D gel electrophoresis, probably reflecting different degrees or forms of posttranslational modification in the different samples. The presence of these protein markers in red and white blood cells, respectively, should allow anxiety to be diagnosed by blood screening.

Protein Interactions

The yeast two-hybrid system has been employed successfully to find interaction partners for a protein named Disrupted-In-Schizophrenia 1 (DISC1), which has been identified as a susceptibility factor in schizophrenia. The first such protein to be identified was fasciculation and elongation protein zeta-1 (FEZ1) which plays a role in the control of axonal outgrowth and fasciculation. The two proteins were subsequently shown to be co-localized in the growth cones of hippocampal axons, and evidence suggests that the interaction is required for neurite outgrowth. Follow-on studies identified additional binding partners Kendrin and DISC1-binding zinc-finger protein, the latter of which was shown to be expressed strongly in the hippocampus, olfactory tubercle, cerebral cortex, and striatum. This pattern was similar to that of pituitary adenylate cyclase, suggesting a role for the pituitary adenylate cyclase-activating polypeptide (PACAP) pathway in providing further insight into the basis of psychiatric disorders.

Mutational Genomics and RNAi

Mouse genetic models have played an important role in the analysis of psychiatric disorders since candidate human susceptibility genes can be evaluated through the behavioral analysis of mice carrying mutations in orthologous genes. Recently, the use of large-scale ENU mutagenesis followed by phenotypic analysis for mutants with anomalies in anxiety behavior, rest–activity patterns, sensorimotor gating, and learning and memory has considerably expanded the potential for functional annotation of human genes associated with psychiatric disorders. The identification of behavioral phenotypes is more challenging than are mutations affecting visible or physiological traits, which can be measured directly, and as in human populations, an additional layer of complexity is introduced by the variable genetic background and environmental effects. Even so, the strategy has been successful in identifying the mouse *Clock* gene, encoding a transcription factor that plays a fundamental role in establishing the circadian rhythm. Since sleep deficit and disrupted circadian rhythms underlie a variety of psychiatric disorders, including bipolar disorder and depression, the identification of the *Clock* gene provides an important example of the use of mutational genomics to identify and characterize genes involved in psychiatric disorders.

RNAi has also been used to study psychiatric disorders following the development of an osmotic minipump to facilitate widespread siRNA vector infusion into the brain. Silencing has been achieved for several genes, including *DAT* (encoding the dopamine transporter) and *SERT* (encoding the serotonin transporter). DAT controls the temporal and spatial activity of dopamine by allowing reuptake into presynaptic terminals, thus regulating dopamine actions on emotion, reward, and cognition and playing a role in many neurological diseases, including mood disorders like depression. *DAT* mRNA levels were reduced by more than one-third using a 2-week minipump, resulting in various behavioral phenotypes including hyperlocomotor activity, as seen in *DAT* knockouts. SERT is implicated in the etiology of several psychiatric disorders, including anxiety and depression, and a similar 2-week siRNA infusion into the dorsal third ventricle also reduced mRNA levels by one-third in the dorsal and median raphe nuclei. There was a concomitant reduction in protein levels, and the resulting antidepressant behavior in a forced swim test was very similar to that obtained using a selective SERT inhibitor.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Cognitive Dysfunction in Psychiatric Disorders; Depression and the Brain; Executive Functions: Eye Movements and Neuropsychiatric Disorders; Gene Expression Regulation: Steroid Hormone Effects; Psychiatric Genomics and Expression Profiling; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Schizophrenia: Genetics; Stress Response: Genetic Consequences.

Further Reading

Brown SD and Nolan PM (1998) Mouse mutagenesis: Systematic studies of mammalian gene function. *Human Molecular Genetics* 7: 1627–1633.

Ditzen C, Jastorff AM, Kessler MS, et al. (2006) Protein biomarkers in a mouse model of extremes in trait anxiety. *Molecular and Cellular Proteomics* 5: 1914–1920.

Genc S, Koroglu TF, and Genc K (2004) RNA interference in neuroscience. *Molecular Brain Research* 132: 260–270.

Hattori T, Baba K, Matsuzaki S, et al. (2007) A novel DISC1-interacting partner DISC1-Binding Zinc-finger protein: Implication in the modulation of DISC1-dependent neurite outgrowth. *Molecular Psychiatry* 12: 398–407.

Hoyer D (2007) RNA interference for studying the molecular basis of neuropsychiatric disorders. *Current Opinion in Drug Discovery and Development* 10: 122–129.

Hrabe de Angelis M and Balling R (1998) Large scale ENU screens in the mouse: Genetics meets genomics. *Mutation Research* 400: 25–32.

Mirnics K, Levitt P, and Lewis DA (2006) Critical appraisal of DNA microarrays in psychiatric genomics. *Biological Psychiatry* 60: 163–176.

Miyoshi K, Honda A, Baba K, et al. (2003) Disrupted-In-Schizophrenia 1, a candidate gene for schizophrenia, participates in neurite outgrowth. *Molecular Psychiatry* 8: 685–694.

Morris CM and Wilson KE (2004) High throughput approaches in neuroscience. *International Journal of Developmental Neuroscience* 22: 515–522.

Schimenti J and Bucan M (1998) Functional genomics in the mouse: Phenotype-based mutagenesis screens. *Genome Research* 8: 698–710.

Stoltenberg SF and Burmeister M (2000) Recent progress in psychiatric genetics: Some hope but no hype. *Human Molecular Genetics* 9: 927–935.

Tarantino LM and Bucan M (2000) Dissection of behavior and psychiatric disorders using the mouse as a model. *Human Molecular Genetics* 9: 953–965.

Thakker DR, Hoyer D, and Cryan JF (2006) Interfering with the brain: Use of RNA interference for understanding the pathophysiology of psychiatric and neurological disorders. *Pharmacological Therapy* 109: 413–438.

Psychiatric Genomics and Expression Profiling

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

Certain aspects of gene function can be determined by the analysis of gene and protein expression. For example, if a gene is expressed specifically in response to growth factor signaling, it is likely the gene's function is in some way linked to the control of cell proliferation. Genes that are expressed in a disease state but not in healthy tissue are likely in some way to reflect the pathogenesis of that disease or the body's response to it. This is one way to identify and characterize genes that are modulated in psychiatric disorders or whose abnormal expression underlies susceptibility to such disorders.

Traditionally, gene expression profiles have been studied individually or in small groups, using techniques such as northern blotting, nuclease protection, or the reverse transcriptase–polymerase chain reaction (RT-PCR). This has provided useful information about specific candidate genes but has limited the scope for functional annotation on a wider scale. In the mid-1990s, however, the first experiments were carried out based on high-throughput expression profiling, in which hundreds or even thousands of gene expression profiles could be studied simultaneously. The ultimate goal of examining the entire transcriptome, that is, the sum of all mRNA molecules synthesized in a cell, tissue, or organism, became possible. Such comprehensive analysis can help identify all the genes involved in any medically relevant process and fill in some of the gaps in our knowledge of signaling pathways and regulatory networks. Already, transcriptome analysis has helped identify new disease markers, drug targets, and potentially therapeutic gene products for psychiatric disorders.

Expression Profiling Technology

Two major strategies have been developed for global gene expression profiling, one based on high-throughput sequencing and the other on miniature devices called microarrays or gene chips. Both methods are quantitative, but one produces digital data and the other analog data which must be converted into a digital readout. It is also important to understand that while the terms transcriptional profiling and expression profiling are widely used to describe these methods, what they really measure is not the amount of transcription or expression that occurs but the steady state mRNA level for each gene. The steady state mRNA level takes into account RNA degradation and turnover as well as transcription, and both processes are highly regulated and differ considerably between gene products. A prodigiously expressed gene whose mRNA is very unstable and breaks down quickly will have the same steady state level as a weakly expressed gene with long-lasting mRNA, the principle being that the actual amount of functional gene product produced (the protein) will depend on the availability of mRNA in the cell for translation, not solely on how quickly it is produced or how long it lasts.

Sequence Sampling

The first large-scale gene expression profiling studies involved the random sampling of sequences from complementary DNA (cDNA) libraries, and later the *in silico* equivalent – unbiased sampling of partial cDNA sequences, known as expressed sequence tags (ESTs), stored in DNA databases. The principle behind such techniques is that a strongly expressed gene will be abundantly represented in a cDNA library or equivalent database as long as it remains unnormalized. Normalization is the process by which redundant sequences are removed so that genes are equally represented. In an unnormalized library or database, then, redundancy remains, and any sampling experiment will tend to pull out clones or sequences of strongly expressed genes much more frequently than those of poorly expressed genes. Relative expression levels can be determined simply by counting the score for each sequence pulled out during sampling, and because these data are in the form of real numbers, they are digital.

Although the sampling of cDNA libraries and EST databases has been used with success, the amount of sequencing required to achieve statistically relevant data is very high, making such experiments lengthy and expensive. An improved method called serial analysis of gene expression (SAGE), developed in 1995, circumvents this problem by using very short tags to represent each gene. The principle is that even in a large genome, it should still be possible to identify each gene uniquely with a short sequence signature of 9–15 nucleotides. SAGE works by joining hundreds of such sequence signatures (SAGE tags) in a concatemer, which can be sequenced in a single reaction. Whereas one sequencing reaction is required to identify one gene in traditional sampling, one sequencing reaction can identify 200–300 genes in SAGE. However, the data are still digital representations of transcript abundance, allowing direct comparison between expression profiling experiments.

A more recent development in this field is massively parallel signature sequencing (MPSS), which replaces the SAGE concatemer with a flow cell full of latex beads, each carrying out an individual sequencing reaction. Although the basis of MPSS is complex, essentially what happens is that restriction enzymes are used to chew away cDNA sequences that are covalently attached to microscopic beads. The enzyme chews away four bases of DNA in each cycle and always leaves a four-base overhang. The 16 possible combinations of four bases can be identified by hybridization with complementary oligonucleotides attached to a unique fluorescent label. Therefore, each sequencing cycle sees four bases exposed on thousands of different cDNA sequences attached to beads and the identification of those bases by a flash of light at a particular wavelength. Since each bead is kept in a unique position, 10–50 bp of sequence can be determined very quickly simply by logging the different flashes of color on each bead as the cDNA is progressively digested away. Just like SAGE, each cDNA sequence is identified and counted, and more abundant mRNA species will be represented by more abundant cDNAs. Just like SAGE, the data are digital scores of sampling, and the advantage is that hundreds or even thousands of sequences can be sampled in one experiment.

Microarrays

Microarrays are miniature devices comprising a collection of immobilized DNA sequences arranged in a grid. Each address on the grid corresponds to a unique gene, or at least to a unique mRNA if the gene undergoes alternative splicing. These arrayed targets are hybridized (interrogated) with a complex mRNA or cDNA probe, that is, a probe comprising many different sequences representing the entire population of mRNAs from the cell or tissue under investigation. In this case, quantitative data are obtained, not by counting, but by measuring the strength of the hybridization signal at each address. If the source mRNA or cDNA probe is universally labeled, the most abundant mRNA/cDNA molecules in the population will generate quantitatively stronger signals on the array as long as hybridization does not reach saturation. Since each individual feature on the array generally comprises 10^6–10^9 identical DNA molecules, only a small proportion are occupied during hybridization even in the case of abundant RNAs. The use of arrays allows simultaneous measurement of the relative levels of thousands of mRNAs, but it should be borne in mind that the information derived from the experiment is restricted by which sequences are represented on the array. If any gene is not represented, then its expression level cannot be measured. This is a disadvantage compared with SAGE, in which the system is completely open and any sequence can be sampled, although this disadvantage can be overcome by using comprehensive arrays uniquely representing every possible transcript.

There are two major types of microarrays used in expression analysis: spotted DNA arrays and printed oligonucleotide chips. They differ in many ways, including the manufacturing process, experimental principles, and data extraction, but they do provide data of approximately equal sensitivity. Spotted arrays are produced by printing DNA samples onto microscope slides treated to facilitate covalent attachment. The features are printed in dense grids (usually about 5000 features per square centimeter) using specially designed arraying robots that can transfer minute amounts of fluid from microtiter dishes onto the array surface. The DNA sequences transferred to the device must therefore be prepared beforehand with the polymerase chain reaction, and they are double-stranded molecules varying in length from 100 to 300 bp.

When using spotted arrays to study psychiatric disorders, samples are obtained from diseased and normal tissue, and the complex probes from each source are extracted and labeled with different fluorophores. They are applied to the array simultaneously to allow the measurement of expression levels in both tissues on the same device. The most common strategy is to use Cy3 and Cy5, which fluoresce at different wavelengths, and then use false color to provide a readout of which genes are differentially expressed and which are expressed at the same level in the two different tissues. If a particular cDNA is present only in the Cy3-labeled population, the spot on the array can be colored to appear red. If another cDNA is present only in the Cy5-labeled population, the spot on the array can be colored to appear green. The cDNAs that are equally represented in both populations contain equivalent proportions of each label and appear yellow.

In contrast to spotted arrays, oligonucleotide chips consist of short single-stranded oligonucleotides (20–70 bases in length) which are manufactured on the chip *in situ*. The procedure involves the attachment of single nucleotides at the correct positions on a hydroxylated and silanized glass or silicon wafer and the obstruction of the 3′-hydroxyl group that accepts the next incoming nucleotide with a photolabile protecting group. A chromium mask is then applied to the surface of the chip and determines which areas are exposed to light. Under illumination, the protecting groups in these areas are destroyed, allowing the addition of a single nucleotide, which is also blocked with a photolabile protecting group. If this process is repeated using a series of different masks and different nucleotides, a precise grid can be generated

containing millions of defined and precisely arrayed oligonucleotides. More recently, alternative technologies have been developed which reduce the capital costs involved in chip manufacture by employing virtual masks, such as miniature arrays of mirrors, to control oligonucleotide synthesis. Oligonucleotide chips are operated similarly to spotted arrays, but dual hybridization is not used. Instead, probes for chip hybridization are made from cleaved, biotinylated RNA, and comparative expression analysis is carried out by hybridization of alternative RNA samples to identical chips, followed by comparison of signal intensities. The specificity of hybridization to spotted DNA arrays is relatively high, due to the length of each target. In contrast, the specificity of short oligonucleotides is lower, and to get round this problem, each gene is represented by up to 20 nonoverlapping oligonucleotides to reduce false positives. Also, the chips are designed to contain both 'perfect match' and 'mismatch' oligonucleotides for each specific target, the former being exactly complementary to the target and the latter containing a central mismatching base to control for cross-hybridization. The signal obtained from the mismatch control is subtracted from that of the corresponding perfect match to reveal the actual level of specific hybridization.

Experimental Design and Data Processing

Expression profiling to study psychiatric disorders necessarily involves the use of postmortem brain material, and one factor that must be considered carefully is the cellular diversity and hence the potential for gene expression heterogeneity in small subpopulations of brain cells. This will affect comparisons of tissue samples from different parts of the same brain, as well as comparisons between different brains. In contrast to cancer profiling, in which it is relatively simple to obtain diseased and healthy tissue from the same individual, the study of psychiatric disorders is complicated by potential variance introduced by the diverse genetic background of different brain tissue samples, the impact of lifestyle and drug treatments on gene expression, diversity introduced by age and race, comorbidity, and the fact that many psychiatric disorders actually represent a spectrum of symptoms rather than a single disease. In addition, since some psychiatric disorders involve the depletion of certain populations of brain cells, differences found in gene expression may reflect the differing populations of particular cells rather than the disease itself.

If all these issues can be addressed using carefully matched postmortem brains, sequence sampling experiments such as SAGE will generate digital data in the form of hits on specific genes that can be counted and converted into percentage scores that can be compared directly between experiments without further standardization or normalization. In contrast, microarrays generate analog data (spots of light at different wavelengths) which must be corrected for background, nonspecific hybridization and spurious signals and then normalized against the expression of housekeeping genes (whose expression should be the same in all samples) before cross-comparisons are possible. This is achieved by converting the signal from each area of the array into a numerical readout by measuring the intensity measurement of each of the microarray pixels and combining these into a single intensity value that corresponds to individual gene targets. Normalization is necessary because the overall microarray signal intensity may vary greatly across arrays within the same experiment, and the manual elimination of spurious signals, such as from dust particles, may also be required. Microarray targets that do not yield a uniform signal are flagged, as are gene targets that do not reach a certain signal-to-background ratio or signal intensity. As a result of this procedure, microarray targets are divided into a group of measurable and nonmeasurable genes. In a typical brain microarray experiment, this approach can reduce the data set to about 50% of the total starting amount.

In most profiling experiments, there is some form of comparative analysis, which means the differentially expressed genes need to be identified, and this is usually the province of specialized data analysis software that calculates the likelihood that observed differences reflect a true difference in gene expression as opposed to natural variation. It is difficult to apply stringent correction methods without limiting the discovery of genuine differential gene expression, so a typical approach is to pick candidate genes showing at least a 20% change in expression level and then use a data set-specific false discovery rate assessment, which involves multiple statistical testing to ensure that differential expression results can be repeated.

To help researchers interpret data from various transcriptional profiling experiments, most peer-reviewed journals have implemented a standardized process for presenting microarray data. This process is based on the Minimum Information about a Microarray Experiment guidelines devised by the Microarray Gene Expression Data Society.

Expression Profiling in Specific Psychiatric Disorders

Anxiety

Although human studies have not been carried out, insight into the changes in gene expression that

accompany human anxiety disorders has been gained by studies in rats. PVG hooded rats and Sprague–Dawley rats react quite differently in response to cats, showing different degrees of anxious behavior. Comparative analysis has revealed significant changes in the abundance of mRNAs for Arc, glyoxalase-I, NGF-I, and PLP, as well as genes involved in the synthesis and transduction of γ-aminobutyric acid (GABA)ergic signals, suggesting these may also be involved in human anxiety.

Autism

Autism was one of the first conditions tested by expression profiling of brain tissue. Using cDNA microarrays and a range of other techniques to compare brains from ten autistic patients and 23 normal individuals, it was established that autism might involve differences in the mRNA levels for certain α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type glutamate receptors and glutamate transporters in the cerebellum. Another study using cortical tissue from 28 brains suggested that differences in MeCP2 expression could be linked to autism as well as a variety of other neurodevelopmental disorders, such as pervasive developmental disorder, Prader–Willi syndrome, and Angelman syndrome. However, these findings have yet to be repeated.

Bipolar Disorder

Another early study involved the analysis of 1200 genes in the frontal cortices of ten postmortem brains from people with bipolar disorder. This analysis showed clearly increased levels of *CASP8* and *ERBB2* mRNA, whereas the levels of TGFβ1 decreased. TGFβ1 has a neuroprotective role, and it has been suggested that its reduced expression may reflect neurotoxicity. Other studies have shown that bipolar disorder involves significant disruption to transcripts encoding enzymes and other components of mitochondrial energy metabolism, causing downstream deficits in adenosine triphosphate. There are also some common links with schizophrenia, in that studies of both bipolar disorder and schizophrenia show that neuropeptide Y (*NPY*) mRNA levels are lower in the frontal cortex, and there is also downregulation of certain oligodendrocyte and myelination genes, including transcription factors that regulate their expression.

Major Depression Disorder

Few studies have shown reproducible transcriptome data for major depression disorder (MDD), although there have been some interesting findings when therapeutic strategies for treating the disorder have been investigated. Oligonucleotide arrays were used to test dorsolateral prefrontal cortex and anterior cingulate samples from MDD patients and unaffected controls, revealing dysregulation affecting several components of the fibroblast growth factor (FGF) signaling pathway. Among the ligands affected were FGF1, FGF9, and FGF12, and at least two receptors were also involved: FGFR2 and FGFR3. The comparison of 12 patients with MDD and 14 closely matched controls also found evidence that genes involved in cell communication and signal transduction were dysregulated, leading to oligodendroglial abnormalities and loss of synaptic function.

Studies of MDD therapies have pointed to the activation of several signaling pathways by electroconvulsive therapy, which is used to treat drug-resistant MDD patients. Of particular interest was the finding that *NPY* mRNA levels increased, since NPY levels are known to fall in MDD patients and can be increased by antidepressants. Similarly, the level of thyrotropin-releasing hormone (TRH) mRNA was shown to increase in response to therapy, which correlates with the antidepressant effects of intraperitoneal TRH injections in the forced-swim model. Notably, NPY also has antidepressant effects in this model. The data from drug therapy experiments have been less rewarding. For example, only four genes were shown to be downregulated in response to treatment with the antidepressants paroxetine and mirtazapine, and it is not clear how these genes are related, if at all, since two were uncharacterized ESTs, one encoded a ribosomal protein, and the other a conserved gene in the major histocompatibility complex.

Schizophrenia

Schizophrenia has been profiled more than any other psychiatric disorder, but analysis has been complicated by the fact that the disease presents a spectrum of symptoms rather than a single, unique phenotype. Perhaps for this reason, different studies have generated diverse findings, although there have also been some interesting cross-study correlations, particularly focusing on the dysregulation of genes involved in energy metabolism, oligodendrocyte function, the GABA and glutamate systems, and the development and maintenance of synapses.

The impact of energy metabolism on schizophrenia is one of the most clear-cut findings from microarray studies, with one team of investigators showing that the expression of genes related to energy metabolism and oxidative stress differentiated almost 90% of schizophrenia patients from controls, in agreement with proteomic and metabolomic data. The general downregulation of mitochondrial genes encoding

enzymes and structural components of the respiratory chain has been replicated in several studies, along with indications that schizophrenia may also affect the transcarboxylic acid cycle, amino acid metabolism, polyamine metabolism, and ubiquitin metabolism. However, similar trends are seen in other disorders, including Alzheimer's disease and multiple sclerosis, indicating that reduced energy metabolism in schizophrenia may be the secondary consequence of a distinct primary defect.

Another consistent finding across expression profiling studies is the reduction in mRNA levels corresponding to myelination and oligodendrocyte function. In one of the largest such studies, prefrontal cortices from older, hospitalized schizophrenia patients were analyzed against matched controls, with the use of an oligonucleotide chip containing 6000 probe sets. This study revealed 35 genes that were specifically downregulated in schizophrenia, and with these genes alone, it was possible to differentiate 100% of the schizophrenia patients from the controls. The repeated finding of oligodendrocyte and myelination genes in profiling studies using sequence sampling and microarrays has been backed up by other experimental approaches (e.g., differential display PCR, RT-PCR). This provides compelling evidence that the dysregulation of critical oligodendrocyte development and maturation genes is an important component of schizophrenia, not only in the prefrontal cortex but also in other brain areas, such as the temporal gyrus. This is concordant with studies based on magnetic resonance imaging, which show reduction in white matter volume and decreased oligodendrocyte density in schizophrenic patients.

Schizophrenia is also characterized by the loss of presynaptic markers. The first profiling study of human postmortem brain tissue from patients with schizophrenia established that transcripts related to the presynaptic release machinery in the prefrontal cortex were present at lower levels than normal. However, it was found that the pattern of affected transcripts differed from individual to individual, suggesting diversity at the molecular level, which might underpin the spectrum of disease symptoms. In this and subsequent investigations, a large number of underrepresented transcripts have been identified, including those encoding synaptobrevin, synaptogyrin 1, synaptophysin, synaptotagmin I and IV, SNAP23 and SNAP25, *N*-ethylmaleimide-sensitive factor, and various components of the clathrin-dependent endocytosis pathway. While most of the dysregulated transcripts were underrepresented in the disease state, the mRNA for the plasma membrane protein syntaxin was shown to be overexpressed in schizophrenia. It is also notable that many of the transcripts identified above were shown to increase in abundance in response to haloperidol or clozapine, which are used to treat schizophrenia. However, many other genes are also affected, and synaptic genes can be repressed as well as induced by these drugs, so the picture revealed by these studies remains complex.

While many of the genes shown to be modulated in schizophrenia fall into the functional groups discussed above, others have been identified which fall without this classification, and the specific way in which they contribute to the disease remains to be elucidated in detail. Microarray studies support previous findings that schizophrenia may involve alterations to the GABA and glutamate systems. One study showed a reduction in the mRNA for GAD67, an interneuronal enzyme, and in the mRNA for several postsynaptic GABA-A receptor subunits. Other studies have shown reduced expression of AMPA2 receptor mRNA in the temporal cortex of patients with schizophrenia. The mRNA for 14-3-3 protein, a regulatory protein with roles in many signaling pathways, the cell cycle, and neurotransmitter metabolism, is underrepresented in schizophrenics, according to several independent studies, as is the mRNA for the heat shock protein and molecular chaperone HSPA12A and for the signaling regulator RGS4, which has been identified in linkage studies as a susceptibility factor for the disease. Conversely, transcripts for several members of the apolipoprotein L family are upregulated in the prefrontal cortex, and the *chondrex* mRNA, encoding a protein involved in neuronal migration, is upregulated in the hippocampus.

In addition to studies profiling neural tissue, there has been some effort to use the profiling of nonneural tissue for disease diagnosis. In one notable study, the microarray-based analysis of peripheral blood from schizophrenics, individuals with bipolar disorder, and unaffected controls was used successfully to distinguish between the three groups. This effort involved the analysis of linear and nonlinear combinations of eight putative biomarker genes (*ADSS*, *APOBEC3B*, *ATM*, *CLC*, *CTBP1*, *CXCL1*, *DATF1*, and *S100A9*) and achieved an accuracy of more than 95%.

When all the results from the schizophrenia transcriptome studies are combined, it appears to be that the affected transcripts in schizophrenia occur in multiple cell types and involve impairments related to processes of synaptic release, GABA- and glutamate-related neurotransmission, cell signaling, second messenger systems, energy metabolism, protein turnover, and myelination. There is little doubt that these changes are interrelated, although the precise causal relationship among them is poorly understood.

Summary

Expression profiling has been used since 1995 to identify genes whose dysregulation underlies human disease and drug responses. The use of expression profiling to study psychiatric disease is a more recent development, yet already these studies have provided useful data on a range of disorders, particularly schizophrenia. Several principles have already emerged, including the tendency for certain transcriptome disturbances to occur in multiple psychiatric disorders and the tendency for diseases often to reflect a multitude of molecular disturbances with a destabilizing cumulative effect on certain processes (such as oligodendrocyte function in the case of schizophrenia) rather than very clear and definitive disturbances in the expression of individual genes. This fits in well with the spectrum of symptoms seen in many psychiatric diseases. Data from expression profiling have provided a number of new leads in terms of identifying disease susceptibility genes, particularly where changes in expression levels are due to regulatory mutations (i.e., mutations in promoters, enhancers, or *cis*-acting sites required for RNA splicing) and can thus be identified in at-risk populations by scanning for single nucleotide polymorphisms in target genes. The identification and characterization of such mutations may also lead to improved therapies, as drugs can be developed with the specific aim of restoring the expression levels of key genes to normal or of selectively destroying overexpressed transcripts and normalizing the phenotype.

See also: Anxiety Disorders; Autism; Decoding Neuron Transcriptome by SAGE; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; Gene Expression Regulation: Activity-Dependent; Gene Expression Regulation: Chromatin Modification in the CNS; Gene Expression Dysregulation in CNS Pathophysiology; Genetic Influence on CNS Gene Expression: Impact on Behavior; Psychiatric Disorders: Functional Genomics; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Schizophrenia: Genetics; Stress Response: Genetic Consequences.

Further Reading

Johnston-Wilson NL, Bouton CM, Pevsner J, Breen JJ, Torrey EF, and Yolken RH (2001) Emerging technologies for large-scale screening of human tissues and fluids in the study of severe psychiatric disease. *International Journal of Neuropsychopharmacology* 4: 83–92.

Luo Z and Geschwind DH (2001) Microarray applications in neuroscience. *Neurobiological Disease* 8: 183–193.

Marcotte ER, Srivastava LK, and Quirion R (2003) cDNA microarray and proteomic approaches in the study of brain diseases: Focus on schizophrenia and Alzheimer's disease. *Pharmacological Therapy* 100: 63–74.

Mirnics K, Levitt P, and Lewis DA (2004) DNA microarray analysis of postmortem brain tissue. *International Reviews in Neurobiology* 60: 153–181.

Mirnics K, Levitt P, and Lewis DA (2006) Critical appraisal of DNA microarrays in psychiatric genomics. *Biological Psychiatry* 60: 163–176.

Mirnics K, Middleton FA, Lewis DA, and Levitt P (2001) Analysis of complex brain disorders with gene expression microarrays: Schizophrenia as a disease of the synapse. *Trends in Neuroscience* 24: 479–486.

Mirnics K and Pevsner J (2004) Progress in the use of microarray technology to study the neurobiology of disease. *Nature Neuroscience* 7: 434–439.

Sun Y, Zhang L, Johnston NL, Torrey EF, and Yolken RH (2001) Serial analysis of gene expression in the frontal cortex of patients with bipolar disorder. *British Journal of Psychiatry* 41: s137–s141.

Taylor JG, Choi E-H, Foster CB, and Chanock SJ (2001) Using genetic variation to study human disease. *Trends in Molecular Medicine* 7: 507–512.

Relevant Website

http://www.mged.org – Microarray and Gene Expression Data Society.

Psycholinguistics

N B Ratner, University of Maryland, College Park, MD, USA
J B Gleason, Boston University, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition
© 2004, Elsevier B.V.

Historical Background

The work of the German philosopher and scientist Wilhelm Wundt (1832–1920) laid the foundation for psycholinguistics, although Wundt did not have the benefit of the linguistic insights that give the modern field its power. At his laboratory in Leipzig, he developed an early theory of speech production, and used reaction time and other experimental measures that are still basic to psycholinguistic research. The premise that underlies the measurement of reaction time is that the time taken to respond to an experimental task reflects the degree of mental complexity involved in processing it. For instance, in some contemporary studies the participants are asked to respond as quickly as possible as to whether a string of letters flashed in front of them on a tachistoscope comprise a real word; these tasks are called lexical decision tasks. Typically, it takes much longer to respond to letters representing rare words than it does to those representing common ones. Thus, it takes much longer to reach a lexical decision about the string of letters 'a w r y' compared to a decision about the string 'n e a t.' This difference in reaction time presumably mirrors the greater amount of cognitive activity required to search through the mental lexicon to find the rarer word, match it to the tachistoscopic stimulus, and decide that the string of letters is, indeed, a word. Studies of this nature have consistently shown that there is a direct relation between the usage frequency of a word and the latency of response to it: the more common the word, the more rapidly it can be accessed.

Until the 1950s most psychologists studying language followed learning theories that emphasized serial patterning in behavior. Psychologists working within the framework of behaviorism, such as Hull, Watson, and Skinner, differed in their specific accounts, but they all viewed language as the product of various forms of learning. For instance, classical conditioning could be invoked to explain the acquisition of meaning. According to this model, an infant learns the meaning of a word such as 'bottle' through repeatedly hearing adults say the word while presenting the actual bottle. Ultimately the word alone evokes a response in the infant that is very similar to the response to the bottle itself. That response (pleasure, anticipatory eating actions, etc.) can be thought of as the meaning of the word. Instrumental learning occurs if parents train infants to speak through shaping and reinforcement of their early vocalizations until the infants' successive approximations of language become closer and closer to actual language. Finally, social learning occurs if children acquire language through imitation of the speech of others. The belief that children acquire language through imitation is widespread, even though psycholinguistic research has challenged this view. Behaviorist accounts of sentence production based on transitional probabilities (or the statistical likelihood that one word is likely to follow another) were advanced by some researchers in the 1960s. In his 1957 book *Verbal Behavior*, Skinner presented a behavioral model of language functioning that became particularly controversial as the revolution in cognitive science began. Within 2 years (in 1959), Chomsky forcefully argued that conditioned stimulus–response associations could not adequately explain the infinite productivity of language, its rapid acquisition by children, and the lack of one-to-one correspondence between the meanings of utterances and their grammatical structure. His work spurred the development of numerous accounts specifying the domain-specific mental representation and acquisition of linguistic rules, including proposals that the mechanisms that enable language development and use were distinct from those used in general learning, and innately enabled by principles of universal grammar. Psycholinguistic experimentation seeking to explore the functional reality of evolving linguistic theories dominated the field until the mid-1980s when an alternative framework, connectionism, emerged to compete with linguistic accounts of language acquisition and use. Connectionism views language knowledge as the product of generalized, rather than domain-specific, patterns of neural network activity that establish and then rely on spreading activation patterns distributed across the network.

The Structure of Language and Its Mental Representation

Every human language is hierarchically organized and can be described in terms of its component subsystems. These systems include phonology, the lexicon, morphology, syntax, and pragmatics and discourse.

Phonology

Phonology includes the significant sounds of the language and the rules for their combination. The words of a language are divisible into sound sequences, and part of language knowledge is an understanding of the particular sounds used in a language, and the rules for how they can be combined and ordered. There are many speech sounds available to the world's languages. Any single language employs a subset of these sounds, typically about 23 consonants and 9 vowels. There is, however, substantial diversity in the world's languages, with numbers of consonants in a given language ranging between 6 and 95, and numbers of vowels ranging between 3 and 46. The distinctive sounds used by a language are its phonemes. Phonemes are contrastive; changing from one to another within a word produces either a change in meaning or a nonword. For example, the /p/ in pit serves to contrast pit from other English words, such as bit, sit, and kit, which are similar in all respects except that they begin with other English phonemes. Psycholinguistic research has shown why speakers often have difficulty trying to learn a second language that has a different phonemic inventory. For example, the initial sound of the English word *this* is rare in the world's languages, and poses particular difficulties for those learning English as a second language, whereas English speakers have difficulty learning sounds not in English, such as the initial sound of the French word *rue*. Research studies using innovative techniques, such as nonnutritive sucking and conditioned head turn responses to stimuli, have shown that human infants can initially discriminate among the sounds of all languages, but around their first birthdays, as they acquire their native language, they begin to lose the ability to make fine distinctions among the phonemes of other languages.

The importance of knowing the rules for the combination of sounds within one's language (phonotactics) becomes apparent in ordinary conversation, because the identity of individual speech sounds is often unclear. Experimental research shows that listeners are often confronted with an auditory puzzle: They hear some sounds, but not all, and must 'fill in the blanks.' The effortless and accurate solution of such puzzles is mediated by adult listeners' knowledge of their phonemic inventory, phonotactic constraints, and lexicon, and also by inferences that can be drawn from the context. In addition, spoken language is characterized by a great deal of acoustic variability in the representation of phonemes, both within and across speakers. Research suggests that speech sound processing is characterized by categorical perception, in which acoustically variant tokens are forced into binary percepts. Infants' abilities to use such skills in segmenting and decoding the ambient speech signal to induce the lexicon and grammar of their language is obviously less apparent, leading to large numbers of current studies and models that seek to understand how they 'bootstrap' and expand early linguistic discoveries.

The phonological system of a language also includes rules for the interpretation of prosody, or intonation and stress patterns. In English, prosodic cues can distinguish between grammatical contrasts, such as the difference between statements and questions (You're going. vs. You're going?). Prosody can also convey emphasis and emotion in language. The status of prosody as a separate system is upheld by the finding that some patients with right hemisphere brain damage lose the ability to distinguish between happy and sad productions of the same sentence. Spoken word recognition is thus enabled by a number of perceptual biases that operate from the bottom up in analyzing the input signal, combined with the use of top-down knowledge of language-specific phonological and prosodic rules, vocabulary, and syntax, along with generalized use of context to derive an appropriate interpretation. Such rapid and complex integration of cues to meaning poses a continuing challenge to the development of computerized spoken word recognition systems that can successfully mimic human speech perception.

The Lexicon

A capable user of a language possesses a vast and complicated mental lexicon, or dictionary, the nature of which is of great interest to psycholinguists. Most adults know the meanings of 60 000–80 000 words in their language, and can readily comprehend or access them for communicative purposes. Psycholinguists investigate both how these words are represented in the mental lexicon as well as how they are rapidly accessed during language production and comprehension.

It is not easy to explain exactly what a word 'means.' How is it that one identifies certain objects as chairs, no matter how they may vary in construction? When does an object deserve the label cup, rather than glass? The difficult problem of explaining how people categorize and name concrete concepts has been the subject of much research. Proposals that the association of names with objects is constrained by innate biases in child language acquisition have been an object of recent psycholinguistic research. There is also clear evidence, through studies in which lexical access is speeded by priming the listener with semantically related items, that meaning is used in the word recognition process. Researchers also

contrast the knowledge of concrete words with abstract notions such as 'friendship' or 'patriotism,' and relative words such as 'good' or 'short.' There is evidence that factors such as abstractness and relativity of meaning increase task difficulty and influence the order in which children acquire the meanings of various classes of words. Some types of Freudian slips, in which the speaker retrieves the opposite of the intended lexical target, also suggest an underlying organization of the mental lexicon along semantic dimensions.

Some words in the language do not easily lend themselves to definition: In the sentence 'John is frustrated by calculus,' what does 'by' mean? In 'John wants to sleep,' what does 'to' mean? Linguists categorize words such as 'John,' 'frustrated,' and 'calculus' differently from words like 'is,' 'by,' and 'to.' Words in the first group are content words and those in the second group are function words. Content words have external referential meaning; function words serve particular functions within the sentence by making the relations between the content words clearer. The distinction between content and function words is psychologically meaningful. It is, among other things, possible for psycholinguists to distinguish between the processes used to retrieve these two types of words in both comprehension and production. The distinction between content and function words in mental representation is further upheld by studies of the language of patients suffering from damage to the language areas of the brain (aphasia). Patients with lesions in Broca's area frequently exhibit agrammatism, a type of aphasia in which content words are easily produced, whereas function words and other grammatical markers, such as plural endings, may be difficult or impossible to produce.

Normal conversational phenomena also provide insight into the possible nature of the mental dictionary. In a 'tip of the tongue state,' when we cannot recall a word, why do we often know what its beginning sound is? Speech production errors (slips and spoonerisms), such as saying 'slickery' rather than 'slippery,' or the Reverend Spooner's, "You have hissed all my mystery lectures. You have tasted the whole worm," are of potential value in psycholinguistic modeling. Such performance errors provide evidence for possible stages in and models of the way we use our mental lexicon in constructing messages. They suggest, among other things, that words may not be stored as wholes, and that to access some words, speakers must retrieve and assemble their components. Some errors suggest that the meanings of words and their phonological representations are stored separately. Moreover, the fact that the vast majority of speech errors result in real, albeit inappropriate, words of the language suggests a prearticulatory stage in speech production in which output is scrutinized for lexical well-formedness, with revision of ill-formed plans as necessary.

Morphology, or Word Formation

In English and other languages, one can readily identify meaningful items that are separable parts of words. In a word like 'cats' we recognize that the final '-s' signifies the notion of plurality; it can thus be appended to other words we wish to pluralize. Other similar word endings in English are -ing, as in 'jumping,' '-s' as in 'boy's', and -ed, as in 'dropped.' These suffixes are grammatical morphemes. A morpheme is the smallest unit of a language that carries definable meaning or grammatical function.

In English, some words consist of only one morpheme. One cannot divide 'table' or 'the' into smaller units that have meaning. Words like 'cats,' however, comprise two separable but unequal notions. Cat can stand by itself in a sentence; '-s' cannot. A free morpheme is one that can stand by itself, whereas a bound morpheme, or affix, cannot stand alone. Affixes serve two distinct functions. Some affixes change a word into a new word with a different meaning or grammatical function; these are called derivational morphemes because they are used to derive a new word. For instance, the affix '-ness' may be added to the adjective 'kind' to derive the noun 'kindness.'

A second type of bound morpheme provides additional information about a word without changing its part of speech or meaning. These are inflectional morphemes; in English they mark number in nouns (cat vs. cats), possession (John vs. John's), verb tense (eat/eating; study/studied), and subject–verb agreement (I/you/we/they understand vs. he/she/it understands).

The distinctions between content and function words and between free and bound morphemes appear psychologically significant when one examines language understanding and production. There is evidence that morphologically complex items are decomposed into their constituent units during the process of lexical access. In addition, slips of the tongue show that these different classes of morphemes are retrieved at different stages of the sentence formulation process during the process of speaking, because they may exchange and have their pronunciation adjusted to their new environments in the errored output (e.g., 'he run outs,' in which the transposed inflection assumes the allophonic realization of /s/ rather than /z/ appropriate to its intended position at the end of the word 'run').

Morphological rule learning is also an intense topic of scrutiny in connectionist accounts of language learning, because connectionist models have been relatively successful in replicating children's patterns of learning and overregularization (e.g., the transient and quite normal use of 'holded' for 'held').

Syntax

Syntax includes rules for combining words into meaningful propositions. Phonological, lexical, and morphological systems are only part of the competent language user's armamentarium. It is important to know how to combine words into sentences that convey meaning, and to be able to interpret the grammar of the language heard.

Young children soon learn the differences signaled by syntax. For example, the sentences 'John loves Mary' and 'Mary loves John' have very different meanings, although the words are the same in the two utterances. Their order differs, and in English these different orders signal major differences in meaning. In English, word order is extremely important: In parsing or processing utterances, we expect the first noun in a sentence to be the subject, and the next to be its object. We also expect subject nouns to be followed closely by verbs. For these reasons, English is called an S–V–O (subject–verb–object) language. Approximately 75% of the world's languages are either S–V–O languages or S–O–V (subject–object–verb) languages.

Word order in English aids in the interpretation of the relationships among words in a sentence. However, not all sentences of English are arranged in S–V–O order. For example, the sentences 'Mary is loved by John' (a passive construction), 'Mary showed John the old man', (a dative construction) and 'The man who lives next to my sister collects antique cars', (a center-embedded relative clause) all violate the typical ordering of subject–verb–object in English. The expectation that words follow a conventional order in English is a processing heuristic that helps a listener arrive at a rapid and accurate interpretation of many sentences. We all use typical word order patterns as a comprehension strategy in English. But if word order deviates from the norm, problems in interpretation can arise; for example, the sentence 'The invasion surprised many people' is easier to process than its synonymous passive 'Many people were surprised by the invasion.' Experimental research with children has also shown that initially children learning English employ a S–V–O word order strategy that leads them to misinterpret passive sentences: Preschoolers, when asked to select from an array of pictures the one called 'The mommy is kissed by the daddy' typically point to a picture showing a woman kissing a man. Sentence processing heuristics vary among languages according to their specific word order typologies. Moreover, the lack of explicit parsing cues in a sentence, such as 'The man felt the dog was poorly trained,' may cause readers expecting subjects, verbs, and, objects to follow each other to initially assume that the subject was touching his pet. It also appears that the mental lexicon assists the language user by setting up syntactic expectations that accompany word meaning. Current theory holds that the mental lexical entry for a verb such as 'greet,' for example, includes information about what greet 'means,' and also specifies the requirement that it be followed by a noun phrase in a sentence. Such a model predicts the psycholinguistic finding that sentences having verbs with a variety of potential arguments appear to be more difficult to process, compared to those with limited options.

Sentence ambiguity is another focus of psycholinguistic research. A sentence such as 'Visiting relatives can be a nuisance' has more than one meaning; a complex interaction of parsing preferences and context appears to aid listeners or readers in selecting the appropriate interpretation. Recent research using functional brain imaging demonstrates that complex and atypical constructions elicit greater cortical activity; additional recent techniques such as eye-tracking and event-related potentials permit psycholinguists to track the parsing strategies used in comprehension and the time course for resolving difficult structures.

It is increasingly clear that language processing is affected by interactions between linguistic systems and the situational context. Work that examines the role of context in interpreting ambiguous speech signals and in making logical assumptions in understanding sentences suggests that many kinds of information contribute to the process of language comprehension. Not all of these sources of information are adequately described by linguistic theory alone.

Pragmatics and Discourse

Language users must evaluate what they hear or read within a context. If we are told by a police officer, for instance, 'This is a no-parking zone,' the message is typically taken as more than a simple comment. It may be a warning or a threat. Language has many uses: It can, among other things, inform, promise, request, query, blame, and even lie. The way language is used in social context is the domain of pragmatics. Pragmatics influences our choice of wording and our interpretation of language within the social context. Just as there are rules for creating grammatical sentences, there are also linguistic conventions for the appropriate use of language in various contexts.

The situational setting or context of spoken messages is often crucial to their successful interpretation.

A woman who says to her friend standing near the window, 'It's hot in here,' is probably asking her friend to open the window. This indirect statement contrasts with the more direct utterance, 'Open the window.' Although it may appear to complicate conversation, indirect phrasing is often used to make requests of others, because it is more deferential or polite. Politeness, like grammar, is a complex and rule-governed linguistic system. However, use and comprehension of such indirect language may pose particular problems for children and adults with language disorders.

Spoken discourse or text comprehension demands more than successful identification and parsing of words and phrases. Every speaker of a language must understand connected discourse and the conventions that relate sentences to one another. Much of our language use involves verbal or written interactions that are longer than a single utterance. Conversation and written text are governed by discourse conventions that dictate the appropriateness of content, structure, and turn-taking. For example, determining the reference for a pronoun (anaphoric reference) may require successful processing and retention of large quantities of prior linguistic information. Conversely, use of a pronoun implies that its antecedent has been previously established. In English, a noun must be specified before a pronoun can be used. Thus, 'I found it!' is a poor opening gambit in conversation with a stranger on a bus, but it makes perfectly good sense if you have just told her that you have lost your ticket. This skill may be difficult for children to master and may result in frustrating conversational gambits for parents (e.g., 'He hit me! Who hit you?'). There is ample evidence that memory for verbal material is rather fleeting; the specific words and sentence structures heard or read decay rather rapidly, leaving only propositional and more general semantic content. Recall of text usually results in substitution of lexical and syntactic paraphrases for the original material. Additionally, much of what we take to be the meaning of text or discourse is inferred rather than explicitly stated (as in mistakenly responding that the phrase 'John used a hammer' was seen in a series of sentences that included the phrase, 'John nailed the boards together').

Successful use of language in social settings also requires appropriate modifications when addressing different types of listeners. Speech styles can vary according to the subject being discussed, the characteristics of the addressee, or the characteristics of the speaker, as well as according to the medium of discourse (written or spoken, telephoned, or face-to-face, etc.). These specially marked ways of speaking are called 'registers.' The way we talk to babies, the language of a sportscaster, and the speech of a minister are examples of different registers. Some registers typify casual interaction, whereas others are more appropriate to formal events. Recent research has shown that the pragmatic system may be less developed than other parts of the language in children suffering from certain types of developmental disabilities, particularly children with autism.

Major Themes in Psycholinguistic Research

Rather than characterizing psycholinguistic inquiry according to the structural properties of language, we can also identify major themes that require integration of linguistic and paralinguistic information across levels of language structure.

How Individuals Understand Spoken and Written Language

The broad area of investigation of understanding language involves scrutiny of the comprehension process at many levels, including investigation of how speech signals are interpreted by listeners (speech perception), how the meanings of words are found (lexical access), how the grammatical structure of sentences is analyzed to obtain larger units of meaning (sentence processing), and how longer conversations or text are appropriately evaluated (discourse). Concerns specifically related to how written language is processed are also in this domain.

Psycholinguistics also explores language processing in a wide variety of disordered populations, including individuals with aphasia, acquired reading disorders and dyslexia/alexia, split-brain functioning, deafness, autism, mental retardation, stuttering, and Williams syndrome. In particular, study of language use following brain damage, and in children with developmental disorders, has the potential to greatly inform and constrain theories and models of normal language functioning.

How Individuals Produce Language: Speech Production

It is easier to evaluate the comprehension process than the production process. We can analyze patterns of accuracy and error, response time, and other behaviors to arrive at a rough estimate of how language stimuli are comprehended by listeners. However, it is more difficult to gain insight into how ideas are put into linguistic form; the process is largely hidden from observation, and speakers' verbal expressions, even in response to controlled eliciting stimuli, vary considerably. Researchers are currently developing complex models of the production process. Typical

models of language production contain multiple stages, beginning with the meaning that is generated; they then proceed to the selection of syntactic structures and intonation contours, the specification of words from the lexicon, and, finally, phonological rules and commands to the motor system for speech. Major sources of information about the probable nature of the speech production process can be found in speakers' speech errors or false starts and in hesitations and pauses that break the rhythm of ongoing speech.

How Language Is Acquired

The major focus in the domain of language acquisition has been on developmental psycholinguistics, or the acquisition of a first language by children. Early findings in developmental psycholinguistics countered behaviorist claims about imitation as a basis for language learning. They showed that young children's language often contains utterances that do not replicate anything the child has heard from adults (e.g., 'I falled down and hurted myself'); in the earliest stages of language development children also do not use articles such as 'a' and 'the', even though these are the most frequent words in English. In an early study (1958) of children's acquisition of the English morphological system, Berko showed that preschool-aged children can correctly form the plurals and past tenses of nonsense words they have never before heard: When shown a picture of a creature called a wug, even very young children could say that two such creatures were wugs. This was strong evidence that children were acquiring a rule-governed internal representation of the language, and not simply imitating the adults around them.

There has been active and enduring controversy over the respective roles of the environment and innate (or inborn) abilities in helping the child master language. Do adults teach language to children by using special kinds of language with them, and providing them with feedback when they have used the language well or poorly? Or do children come into the world with a special, uniquely human talent that enables them to extrapolate the grammar of a language without overt instruction or correction?

Chomsky noted two problems that have spurred enduring controversy in the field of child language development. The first is what has become known as the degeneracy problem: the assumption that the language overheard by young children contains many incomplete and ungrammatical sentences, and that it does not contain the full range of structures used by the language. Thus, there may not be sufficient positive evidence available to the child to allow adequate, competent language development. The second problem relates to negative evidence. This argument holds that children are not taught that some structures are not permissible in the language, either through parental correction of their errors or through literal instruction (e.g., 'You can say X in English, but you can't say Y'). Both the validity and interpretation of such claims are a matter of spirited inquiry and debate within the field of child language acquisition. Many child language researchers have proposed that environmental influences play a substantive role in the acquisition of language by children. The study of language development in children has become a formidably large discipline, with a wide array of journals, texts, and monographs specifically devoted to developmental psycholinguistics.

See also: Connectionist Models of Language Processing; Language Evolution; Language Development; Language Following Congenital Disorders (not SLI); Language: Nonhuman Animals; Language: Aphasia; Language: Learning Impairments; Sentence Production; Sentence Comprehension; Speech Production: Adult; Speech Production: Development; Speech Perception: Adult; Speech Perception: Development; Word Production; Word Recognition; Word Learning.

Further Reading

Altmann G (2001) The language machine: Psycholinguistics in review. *British Journal of Psychology* 92: 129–170.

Berko J (1958) The child's learning of English morphology. *Word* 14: 150–177.

Brown R (1970) Psycholinguistics. New York: The Free Press.

Chomsky N (1959) A review of *Verbal Behavior*, by B.F. Skinner. *Language* 35: 26–58.

Chomsky N (1986) *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

Elman J, Bates E, Johnson M, et al. (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Gernsbacher M (1994) *The Handbook of Psycholinguistics*. San Diego, CA: Academic Press.

Gleason JB and Ratner NB (1998) *Psycholinguistics*. 2nd edn. New York: Harcourt Brace Jovanovich.

Levelt W (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

MacWhinney B and Bates E (1989) *The Cross-Linguistic Study of Sentence Processing*. Cambridge: Cambridge University Press.

Markman E (1989) *Categorization and Naming in Children*. Cambridge, MA: Bradford.

McQueen J and Cutler A (1998) Morphology in word recognition. In: Spencer A and Zwicky M (eds.) *The Handbook of Morphology*, pp. 406–407. Oxford: Blackwell.

Miller G (1990) Linguists, psychologists, and the cognitive sciences. *Language* 66: 317–322.

Sachs J (1967) Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics* 2: 437–442.

Seidenberg M and McClelland J (1989) A distributed, developmental model of word recognition and naming. *Psychological Review* 96: 523–568.

Psychopharmacology of Reward and Appetite in Rats

E M Bowman, University of St. Andrews, St. Andrews, Fife, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The order Rodentia includes nearly all of the species of small mammals used in neuroscience, including mice, rats, hamsters, gerbils, and guinea pigs. Considered as a group, rodents are found in a wide variety of ecological niches and are so successful that they are estimated to include more than 40% of mammal species. In this article, the neurobiology of mice and rats will form the core of the material since these two species are studied most frequently in neuroscience and since behavioral studies can take advantage of their dietary and behavioral flexibility.

First, however, it is important to define the terms reward and appetite. In common language, 'reward' is used to indicate pleasure, gratification, or satiation. In this article, the term reward will refer to the hedonic experience an organism feels after fulfilling a biological or social need. Reward is a latent variable that describes a central psychological state that cannot be observed directly but must be inferred from behavior. In this regard, reward is often confounded with reinforcement and motivation. Reinforcement refers to the ability of the outcome of an action to increase the relative frequency of the action. In many circumstances, reward and reinforcement covary, but reinforcement can occur following the avoidance of aversive outcomes; it is possible to disrupt reinforcement without disrupting the emotional experience of reward; and the conditioned behaviors induced by reinforcement can become independent of reward after habit formation.

Reward is also easily confounded with motivation, which for purposes of this article is defined as the desire to obtain a reward. Perhaps the most concise statement regarding the difference between motivation and reward is the distinction between 'wanting' and 'liking' that was drawn by Berridge and Robinson. It is also important to note that reward occurs during a positive behavioral event rather than in anticipation of it.

The measures used to infer reward in rodents are categorized into three classes. First, unconditioned behaviors, such as orofacial reactions to tastes or ultrasonic vocalizations to play, can be quantified. This is perhaps the approach in which reward is least confounded with motivation and reinforcement. Second, the reinforcing efficacy of a putative reward can be used, assuming that varying the magnitude of reward will influence the relative vigor of operant responding in animals. Third, a behavioral economic approach can be taken in which the cost that the organism is willing to pay for a reward is viewed as an indicator of its magnitude. The costs can be quantified by the amount of work required to earn the reward or by the degree of aversive stimulation an animal is willing to endure to obtain it.

'Appetite' in common parlance refers to hunger, although in many scientific contexts appetite also refers to enduring patterns of consumption. Like reward, appetite is a latent psychological variable that is inferred rather than directly measured. In scientific and medical literature, the specificity of the term appetite varies from author to author, with some referring only to hunger and others referring to ingestive behaviors (both thirst and hunger). The focus of this article is on hunger because it has been most thoroughly researched. Short-term measures of appetite include the amount of food ingested, meal duration, and food preference, whereas long-term measures are based mostly on weight changes.

Psychopharmacology of Reward in Rodents

Dopamine: Pleasure Pathway or an Adaptive Foraging System?

Perhaps the first method of studying the neural correlates of pleasure was intracranial self-stimulation (brain stimulation reward). Discovered in 1954 when Olds and Milner were mapping the functions of the rat reticular activating system, the rewarding effects of localized electrical stimulation were first noted as a conditioned place preference for the quadrant of the test chamber in which the stimulation was delivered. Later, the degree of reward was quantified by varying the stimulation parameters and its locus and subsequently measuring the rate at which rats would press a bar to self-administer it. Mapping studies in rats found that among the most effective sites for intracranial self-stimulation is the medial forebrain bundle, which links the dopaminergic tegmentum with the ventral striatum (nucleus accumbens). It is noteworthy that the title of Olds and Milner's paper refers to the stimulation in terms of reinforcement, whereas the discussion describes the effects of the stimulation in terms of reward. The difficulty in distinguishing reinforcement from reward with respect to the functions of the dopamine system persists to this day.

The rat mesolimbic system, which includes the ventral tegmental area and the ventral striatum, responds to a variety of naturally rewarding events, although the functional significance of these responses is still debated. Increases in extracellular dopamine levels in the ventral striatum have been observed after hungry rats consume food. The dopamine response is lessened by conditioned taste aversion, in which rats learn to avoid food that makes them ill, and by sensory-specific satiety, in which rats devalue food that they have eaten to satiety. Neurophysiological responses evoked by food and water have been observed from individual neurons in the rat ventral striatum and ventral tegmental area. Social rewards evoke dopamine release in the ventral striatum as well. In male and female rodents, increases in dopamine levels in the ventral striatum occur during mating behavior. Increases in extracellular levels of dopamine have been also observed in rat dams suckling their pups. The breadth of dopamine responses to biological and social rewards suggests that they do not encode the specific sensory aspects of reward but, rather, its hedonic valence.

The range of natural rewards to which the rat dopamine system responds is matched by its responsiveness to a variety of artificial rewards. Given the research of Olds and Milner, it is not surprising that intracranial self-stimulation of the medial forebrain bundle leads to increases in extracellular dopamine levels in the ventral striatum, increased expression of c-*fos* in the ventral striatum and the ventral tegmental area, and neurophysiological responses of ventral striatal neurons. Additionally, nearly all recreational drugs, including psychostimulants, opiates, cannibinoids, alcohol, and nicotine, cause increases in extracellular dopamine in the ventral striatum, despite the fact the some of these drugs are stimulants, whereas others are depressants. Drug abstinence in rats results in psychological withdrawal and is associated with decreased dopamine levels in the ventral striatum and elevated intracranial self-stimulation thresholds in the medial forebrain bundle. One hypothesis supported by research on artificial reinforcers in rats is that they substitute for natural rewards but do not undergo satiety because of their direct neural effects. As noted by DiChiara, this notion has obvious application to human addiction.

The dopamine responses described previously do not irrefutably link dopamine with reward since the data are correlational. In order to make statements regarding causality, it is necessary to manipulate the dopamine system by administering dopamine antagonists, dopamine agonists, dopamine reuptake blockers, or compounds that alter dopamine synthesis and metabolism. In general, drugs that increase stimulation of dopamine receptors in the striatum are self-administered by rats, and they potentiate self-administration of other drugs. Dopamine antagonists have been used to disrupt the rat mesolimbic system. The results vary by the selectivity for dopamine receptor subtype, dose, route of administration, and behavioral assay of reward. For instance, administration of neuroleptics in rats leads to response reduction in operant tasks in which the reinforcer is food or intracranial self-stimulation. This led Wise to propose that neuroleptics produce anhedonia. This apparent anhedonia is also seen in unconditioned behaviors, as shown by the reduction in rat ultrasonic vocalizations evoked by reward following administration of the dopamine antagonist flupenthixol. However, there are counterexamples. For instance, administration of the D2 antagonist pimozide spares the orofacial reaction of rats to sweet tastes, which Berridge and colleagues used as a measure of the degree to which the rats like the reward. Given the latter results, Robinson and Berridge proposed that dopamine is involved in incentive motivation (wanting, the process whereby stimuli are tagged as those that should be approached) rather than pleasure (liking).

The involvement of the dopamine system in reward-related processes has also been explored through behavioral phenotyping in mice. The results are globally consistent with psychopharmacological studies in rats. Genetic manipulations that interfere with the lifetime expression of dopamine receptors (D1–D3) or of the dopamine transporter alter the reaction of mice to rewarding stimuli, including addictive drugs. However, dopamine-deficit mice retain preferences for sweet tastes, although they do not feed normally unless given L-DOPA. They also form conditioned place preference for morphine, which demonstrates that dopamine is not necessary for behavioral reactions to reward in these animals. It should be noted that generalizing these results to 'normal' outbred mice should be done cautiously since the result of the genetic manipulation can cause compensatory changes in the nervous system. It is likely that the use of acute antisense RNA or interference RNA will minimize such changes in future mouse research on the dopamine system.

In summary, research in rodents has implicated the mesolimbic dopamine system in reward-related processes, but the precise function is controversial. Two hypotheses have been supported by the rodent work: the first is that the mesolimbic dopamine system signals reward (liking), and the second is that it is involved in incentive salience (wanting). However, research in monkeys by Schultz and others has raised the possibility that the dopamine system is involved in

associating sensory stimuli and environmental contexts with reward (learning). Like the neurophysiological work in monkeys, single-neuron recording studies in rats have shown that neurons in the nucleus accumbens and the ventral tegmental area respond not only to primary reward (both ingestive reward and addictive drugs) but also to stimuli that have become associated with reward through Pavlovian (classical) conditioning. Indeed, manipulating the dopamine system in rats changes the degree to which conditioned stimuli act as reinforcers or enhance operant responding for primary rewards. Thus, reward-related processing within the rodent dopamine system displays plasticity based on experience, and it is possible that the system is involved with the adaptive foraging that has made rodents so successful.

Opiates: The Neural Basis of 'Liking'?

The role that opiates play in reward is obvious from the self-administration of these compounds in both humans and animals. The addictive potency of opiate drugs is interpretable in light of the widespread distribution of opiate receptors in heavily interconnected structures in reward-related neural pathways. Opiate receptors in rodents are found in the ventral tegmental area, the striatum, prefrontal cortex, and the ventral pallidum (a target of ventral striatal projections). Considered as a group, there is strong evidence that these structures are involved in opiate self-administration and reinstatement after abstinence. For instance, rats will self-administer morphine directly into the ventral striatum and into the ventral tegmental area. Rats will also administer selective mu and delta opiate agonists into the ventral tegmental area. The mu agonists are far more potent in this regard, and therefore there is evidence that different opiate receptor subtypes are not equally involved in reward-related processes.

Research on opiate self-administration has linked the opiate system to gamma-aminobutyric acid (GABA) and endogenous cannabinoid neurotransmission within 'limbic' striatopallidal pathways. Heroin self-administration is reduced strongly by compounds that elevate GABA levels in the ventral tegmental area and the ventral pallidum. In the ventral striatum, the activation of the cannabinoid receptor CB1 appears to be important for the reinforcing effects of opiates. Mice lacking the CB1 receptor do not form conditioned place preference to heroin, and in rats CB1 receptor antagonists reduce heroin self-administration and lessen the transient decreases in GABA levels in the ventral pallidum that are evoked by morphine. Thus, it appears that opiate self-administration causes a chain of events in which the activation of CB1 receptors reduces the activity of ventral pallidal GABA receptors, thereby disinhibiting neural activity in the ventral pallidum. Opiate antagonism by naloxone also prevents cannabinoid place preference and self-administration in mice. Thus, the opiate and cannabinoid systems are very closely linked, leading some authors to suggest that the point of intersection lies within the ventral striatum, where cannabinoid and mu opiate receptors are closely juxtaposed.

Given the addictive nature of opiates, it is reasonable to ask whether the endogenous opiate system plays any role in the processing of natural rewards. It is well known that opiate agonists increase the intake of palatable foods and antagonists reduce food intake. Microinjection of the opiate mu agonist DAMGO into the ventral striatum nearly quadruples the amount of fat that hungry rats eat and nearly doubles the amount of carbohydrates they ingest. Microinjections in the ventral striatum of the opiate antagonists naltrexone and naloxone reduce the consumption of sucrose solution in satiated rats. These studies suggest that endogenous opiates selectively influence consumption based on palatability rather than hunger. Thus, the most reasonable conclusion at this point is that endogenous opiates within the mesolimbic system are involved in 'liking.'

In summary, evidence is accumulating that endogenous opiates are used in the encoding of the hedonic aspects of reward. This is based on an analysis of the role that activation of opiate receptors plays during self-administration and during consumption of natural ingestive reward. The mu opiate receptor has been confirmed to play a role in reward, but there are reports that other opiate receptor subtypes are also involved.

Acetylcholine: Rapid Sensory Activator of the Dopamine System?

One of the most addictive drugs in humans is tobacco. The self-administration of nicotine by smoking tobacco is based on an activation of endogenous cholinergic systems. Given the role of the dopamine system in reward, research has focused on the anatomical connections between the dopamine and cholinergic systems. One of the problems that arises when attempting this hodological approach is the complicated anatomy of the Ch5 and Ch6 cholinergic cell groups in the pedunculopontine and lateral dorsal tegmentum, in which both cholinergic and noncholinergic neurons are located. However, it is known that cholinergic neurons in Ch5 and Ch6 project to the ventral tegmental area and to the substantia nigra pars compacta. Dopamine neurons in these sites

express both nicotinic and muscarinic receptors, and cholinergic stimulation of dopamine neurons alters their firing rates. Additionally, pontine cholinergic neurons project to thalamic nuclei, including the mediodorsal nucleus, that are allied with the mesolimbic system. Finally, there are cholinergic interneurons within striatum, one of the major targets of midbrain dopamine projections.

Cholinergic responses are seen in a variety of artificial and natural rewards. Substantial acetylcholine release ($\geq$75% of baseline) in the rat ventral tegmental area is evoked by intracranial self-stimulation, feeding during hunger, and drinking during thirst. Changes in acetylcholine levels in the mediodorsal nucleus of a similar magnitude occur when hungry rats eat food or when rats are injected with amphetamine. The concentration of acetylcholine increases in the ventral striatum during cocaine self-administration, and there is a strong correlation ($r^2 > 0.90$) between *fos* expression in ventral striatal cholinergic interneurons and the vigor with which rats self-administer cocaine. Thus, there appear to be at least two cholinergic systems related to reward processing: the ascending Ch5/Ch6 system and the intrinsic striatal system.

Regarding the ascending system, there has been one study in rats in which neurons in the pedunculopontine system were recorded during appetitive Pavlovian conditioning. It was found that the most neurons responded to the onset lights and tones without regard to whether they were paired with reward, although some neurons responded tonically or to the offset of the stimuli. Transient inactivation of the pedunculopontine region eliminated conditioned dopamine responses in midbrain dopamine neurons. This pattern of results implies that the ascending cholinergic system always broadcasts a signal induced by salient sensory stimuli. Given that dopamine neurons selectively respond to conditioned stimuli, it is likely that cholinergic inputs to dopamine neurons are countermanded by other inputs. Thus, the apparent development of conditioned dopamine responses might reflect the removal of inhibitory influences on the dopamine neurons rather than the addition of new excitatory input.

Norepinephrine (Noradrenaline): Attention Signal Related to Reward?

The functions of norepinephrine neurons in the locus coeruleus seem to closely parallel those of dopamine neurons. For instance, both the dopamine and the norepinephrine (NE) systems project to the rat medial prefrontal cortex and parts of the ventral striatum. In the medial prefrontal cortex, both NE and dopamine are released by conditioned stimuli associated with food reward, food itself, and novelty and putative mild stressors such as human handling. Like the dopamine neurons, locus coeruleus neurons may play a role in plasticity. Disruption of prefrontal NE neurotransmission impairs conditioned place preference to morphine, amphetamine, and food. Moreover, rat locus coeruleus neurons respond to conditioned stimuli associated with reward much in the same way that neurons in the ventral tegmental area do. They respond to conditioned stimuli with a shorter latency than neurons in their prefrontal projection areas, and they exhibit plasticity in their conditioned responses earlier in training than prefrontal neurons. Finally, like drugs that enhance dopamine neurotransmission, the adrenergic agonist clonidine supports both self-administration and conditioned place preference in rats.

Research in mice also leads to the conclusion that the NE system interacts with the dopamine system. Depleting prefrontal NE reduces dopamine release evoked by food and cocaine in both the medial prefrontal cortex and the ventral striatum. However, the same depletion also reduces the dopamine response in those two areas evoked by lithium chloride, which causes a highly aversive malaise. It is noteworthy that the NE depletion also blocks the formation of conditioned place aversion evoked by lithium chloride.

Given the close interrelationship between the NE and dopamine systems, it is reasonable to ask whether there are substantive differences in their reward-related processing. There are indications that this is the case. First, sham feeding results in increases in dopamine within the ventral striatum while reducing NE release. Second, although locus coeruleus neurons respond to conditioned stimuli associated with reward, these neurons respond to both paired and unpaired stimuli presented within an aversive conditioning task. It appears that the NE system responds to a wider range of activating stimuli and therefore its functions are likely to be closely related to attentional state. Attention to environmental stimuli is an important determinant of Pavlovian conditioning in learning theory. It is plausible, therefore, that the effects of NE on appetitive processes are related more closely to reinforcement than to pleasure.

Serotonin: A Role in Modulating Reward?

The role of serotonin in reward-related process has been investigated extensively in rodents. Two themes have emerged from this research. First, serotonin modulates the activity of the dopamine system. Second, serotonin is involved when there is a conflict between choosing immediate small reward and choosing delayed large reward. The basis of these serotonergic

effects appears to be changes in processing of both local and extrinsic inputs to midbrain neurons and direct effects in the prefrontal cortex, respectively.

In rats, anatomical studies have localized serotonin receptors on GABA neurons within the ventral tegmental area as well as on dopamine neurons in both the substantia nigra and the ventral tegmental area. Tract-tracing studies in rats have confirmed that afferents to the dopaminergic mesencephalon arise from the raphe nuclei. Moreover, manipulating serotonin alters the efficacy of dopamine neurotransmission in the ventral striatum and the prefrontal cortex, and it changes baseline neural activity within the ventral tegmental area. Since the nucleus accumbens and prefrontal cortex project to the ventral tegmental area, it is possible that serotonin modulates the dopamine system indirectly as well as directly.

Reward-related effects of serotonergic compounds are numerous. It is difficult to formulate generalizations given the number of serotonin receptor subtypes and the varying degrees to which these subtypes can be selectively manipulated by drugs. One obvious connection between serotonin and reward is the use of serotonin reuptake inhibitors (SSRIs) in depression, a central diagnostic feature of which is anhedonia. The clinical use of SSRIs is matched by the finding that long-term SSRI administration in rats lowers thresholds for intracranial self-administration. In another condition affecting global serotonin function, mice lacking the serotonin reuptake transporter (SERT) exhibit less operant responding for food than control mice. Similar findings have been found using drug manipulations that are selective for specific serotonin receptor subtypes: in both rats and mice, serotonin 5HT-1B receptors are important for modulating the reinforcing effects of cocaine and amphetamine. In rats, 5HT-2C agonists reduce operant responding for both food and intravenous nicotine. Antagonism of the 5HT-3 receptor in rats does not affect operant responses for conditioned stimuli (conditioned reinforcers) that have been paired with reward, but mice overexpressing the 5HT-3 receptor gene show reduced conditioned place preference for contexts associated with cocaine, and they exhibit a lower preference for ethanol than control mice.

Serotonin appears to play a role in decision-making processes when different action–reward possibilities confront rats. Typically in these tasks, animals are confronted with delayed large reward and immediate small reward. Both animals and humans discount the delayed reward, resulting in a pattern of choices that is suboptimum. In rats, serotonin is released in the medial prefrontal cortex in such tasks. Serotonergic lesions of the dorsal raphe increase impulsive choices for the immediate small reward under some conditions, as does systemic administration of a partial 5HT-1A agonist. Since serotonin is released when rats are forced to choose between the two alternatives rather than when rats experience a yoked procedure in which the same pattern of reinforcement is given, it has been proposed that serotonin is not involved in the direct valuation of the rewards but, rather, in the neural estimate of elapsed time.

Psychopharmacology of Appetite in Rodents

In rodents, the homeostatic regulation of nutrients is essential. As might be expected from the complexity of digestion and metabolism, the biological processes that determine how much rodents eat are numerous. It simplifies matters to broadly categorize these processes into stages: monitoring of metabolic and digestive state, triggering of appetitive approach behavior, initiation of feeding, and termination of meals due to satiety.

In terms of behavioral measures, the vigor of approach behavior, the amount of food eaten per unit time, the length of meals, and changes in weight all have been used to quantify changes in appetite. Although these measures relate directly to appetite under normal conditions, confounding processes can come into play when psychopharmacological manipulations are undertaken. For instance, a drug that appears to stimulate satiety might actually produce malaise, whereas a drug that appears to stimulate hunger might do so indirectly by altering energy consumption through psychomotor stimulation.

Research on homeostatic monitoring of nutrients in rodents has revealed several peripheral and central mechanisms. Sensory information regarding the taste and quantity of food ingested is sent from the periphery to an ascending central system that includes the nucleus of the solitary tract and the parabrachial nucleus. Additionally, several neuropeptides act as messenger molecules from the body to the brain. Among the most well known are insulin, leptin, cholecystokinin, and ghrelin. Finally, within the hypothalamus there are glucose-sensing neurons in the arcuate nucleus that project to the other hypothalamic nuclei involved in feeding: the paraventricular, dorsomedial, lateral, and ventromedial nuclei.

Neuropeptides appear to control when rodents eat. It has been hypothesized that ghrelin increases appetite, whereas leptin lowers it. Ghrelin is released by enteroendocrine cells in the stomach, subsequently crossing the blood–brain barrier and strongly activating neurons in the arcuate nucleus. Ghrelin also causes arcuate projection neurons to increase production of neuropeptide Y (NPY) and agouti-related peptide (AGRP). It is believed that release of NPY

and AGRP within the lateral and paraventricular hypothalamus stimulates the onset of eating. In most reviews of appetite, leptin is viewed as part of a negative feedback loop that limits the actions of ghrelin. The gene for leptin was discovered following investigation of mutant *ob/ob* mice, which are hyperphagic and develop severe obesity. Mice and rats that have mutations for the leptin receptor are similarly affected. Leptin is released by adipose tissue and causes the production of cocaine and amphetamine regulated transcript (CART) and pro-opiomelanocortin (POMC). CART and POMC, in turn, counteract the effects of ghrelin, NPY, and AGRP. Thus, the balance between leptin and ghrelin can be viewed as a mechanism that influences the set point for appetite and weight.

It should be noted that the neural control of appetite is far too complex to summarize fully here. Many aspects of appetite control have been left out, including multiple metabolic effects of the neuropeptides listed previously, the appetite-enhancing and -suppressing actions of dozens of neurotransmitters and hormones, the coordination of brain stem and hypothalamic structures with the forebrain, the role of the hypothalamus in influencing diverse feeding-related psychological functions, and the complexity of peripheral mechanisms that sense metabolic states. However, it is clear that considerable progress has been made during the past few years and that modern molecular and biochemical techniques have enabled a renaissance of research on feeding control.

See also: Appetitive Systems: Amygdala and Striatum; Cortical Processing of the Reward Value of Food; Dopamine: Cellular Actions; Electrical Self-Stimulation; Goal-Directed Behavior Theories; Neuropeptides: Food Intake; Neuropsychology of Primate Reward Processes; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex.

Further Reading

Arora S and Anubhuti (2006) Role of neuropeptides in appetite regulation and obesity – a review. *Neuropeptides* 40: 375–401.

Barbano MF and Cador M (2007) Opioids for hedonic experience and dopamine to get ready for it. *Psychopharmacology (Berlin)* 191: 497–506.

Berridge KC (2007) The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology (Berlin)* 191: 391–431.

Burdakov D, Luckman SM, and Verkhratsky A (2005) Glucose-sensing neurons of the hypothalamus. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 360: 2227–2235.

Burgdorf J, Wood PL, Kroes RA, Moskal JR, and Panksepp J (2007) Neurobiology of 50-kHz ultrasonic vocalizations in rats: Electrode mapping, lesion, and pharmacology studies. *Behavioural Brain Research* 182: 274–283.

Di Chiara G, Bassareo V, Fenu S, et al. (2004) Dopamine and drug addiction: The nucleus accumbens shell connection. *Neuropharmacology* 47(supplement 1): 227–241.

Everitt BJ, Dickinson A, and Robbins TW (2001) The neuropsychological basis of addictive behaviour. *Brain Research: Brain Research Reviews* 36: 129–138.

Kelley AE, Baldo BA, Pratt WE, and Will MJ (2005) Corticostriatal–hypothalamic circuitry and food motivation: Integration of energy, action and reward. *Physiology & Behavior* 86: 773–795.

Levine AS (2006) The animal model in food intake regulation: Examples from the opioid literature. *Physiology & Behavior* 89: 92–96.

Neary NM, Goldstone AP, and Bloom SR (2004) Appetite regulation: From the gut to the hypothalamus. *Clinical Endocrinology (Oxford)* 60: 153–160.

Pan WX and Hyland BI (2005) Pedunculopontine tegmental nucleus controls conditioned responses of midbrain dopamine neurons in behaving rats. *Journal of Neuroscience* 25: 4725–4732.

Robinson TE and Berridge KC (2003) Addiction. *Annual Review of Psychology* 54: 25–53.

Wise RA (2006) Role of brain dopamine in food reward and reinforcement. *Philosophical Transactions of the Royal Society of London, Series B: Biological Science* 361: 1149–1158.

Wynne K, Stanley S, McGowan B, and Bloom S (2005) Appetite control. *Journal of Endocrinology* 184: 291–318.

Psychophysics of Attention

H E Egeth, Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The amount of information available to our senses greatly exceeds the information processing capacity of our brains. How we deal with the overload is the topic of this article – attention. More specifically, this article focuses on the study of visual attention. The purpose of the article is not to summarize the neuroscience of attention, but to introduce the basic behavioral findings that have motivated interest in the concept.

Perhaps the most fundamental points about attention are that it is limited, and that it is selective. It allows an organism to select the information that is most relevant to its current goals. With attention we are not simply passive recipients of stimuli; instead, we play an active role in our interactions with the environment.

Before proceeding, it is worth making an important philosophical point. When we say something like "it allows an organism to select…" there is a real danger that we may reify the 'it.' We should remember that there is not a homunculus directing traffic in our heads; referring to attention as an active 'it' is a convenient verbal simplification. In fact, the selectivity of attention may be better thought of as an emergent property of our nervous system. In this brief overview of the field of attention we will touch on (1) evidence of limited capacity and (2) the nature of attentional selectivity.

Attention Is Limited

Enumeration

Some of the earliest experiments in psychology amounted to demonstrations of the limits of attention. For example, Jevons, a nineteenth-century researcher, picked up handfuls of beans and threw them into a box, and after glancing at them briefly, attempted to say how many there were. He found that he was essentially perfect when there were up to four beans, but beyond four made progressively more errors. Many variations of this experiment have been carried out. When reaction time is measured, the time required to respond increases only slightly from one to about four, and increases more steeply thereafter. The ability to enumerate small numbers of items quickly and accurately has been referred to as 'subitizing.' The widely, but not universally, accepted explanation is that a few items (up to about four to six, depending on the methodology) can be apprehended simultaneously; beyond that elements are presumably enumerated one at a time, or, if there is time pressure, their numerosity may be estimated.

Multiple-Item Tracking

Imagine an array of, say, ten gray disks appearing on a computer screen. They are all moving in erratic, random trajectories. One of them is marked as a target by blinking several times, but it then reverts to gray and keeps moving along with the other nine disks. Your task is to follow it for, say, 12 s. At that time the disks stop moving and a probe (a small box) appears around one of the gray disks. Your task is to indicate whether or not the probed item was the one that had been marked by flashing. Could you do it? Yes, indeed. This is a trivial task. Now imagine a trial on which not one but two of the ten disks are marked by flashing, and then after 12 s a single disk is probed and you are asked if this is one of the two that had been marked by flashing. Could you do this? Yes; this is a bit more difficult, but not much. However, as the number of marked disks increases there comes a point at which you simply cannot keep track of additional elements. That point varies across observers, but is approximately three to five disks.

Span of Immediate Memory

In a classic experiment by Sperling, when individuals were shown a 3×4 matrix of letters (four letters in three rows) briefly (say for about 0.1 s) and were asked to report as many as possible, they typically reported only four to five letters. In another condition, individuals were presented with a high-, medium-, or low-frequency tone, the tone indicating which one row (top, middle, or bottom) of the matrix they were to report. In this case they were able to report all four letters in a row when the tone came before or very shortly after the visual display. As the tone was selected randomly on each trial, this means that individuals briefly had access to all twelve letters in the display. However as the tone was delayed, performance declined, until, after as little as a 0.25 s delay, performance reached the same level as the whole report technique (i.e., about four or five letters). Thus, this research provides evidence of a high-capacity short-lived sensory memory (often referred to as iconic memory), followed by a small-capacity short-term memory store.

Psychologists have wondered whether the numerical similarity of the capacity limits (three to four items) shown in the these kinds of studies (enumeration, multiple-item tracking, and immediate-memory span) is evidence of a common limitation, or is just a coincidence. A strong case has been made for a common limit, due ultimately to the ability of attention to hold just three to four items at once. However, the issue is not yet settled. A related point is whether the limitation is properly described in terms of a fixed number of 'slots,' with one 'item' assigned to each slot. This is far from a trivial issue, as the definition of an item is by no means straightforward. For example, is the word 'attention' a single visual object, or are the nine letters composing it the proper units of analysis, or are the line segments composing those letters the proper units?

Visual Search

A popular method of studying perceptual processes is to have an individual look for a designated target item in a display that contains several distractors. Typically, the number of distractors is varied, although the number of targets has also been manipulated. Performance is assessed by measuring reaction time and/or accuracy of performance. Visual search provides a remarkably versatile tool for exploring many aspects of perception and attention with a task that has considerable ecological validity, as it is ubiquitous in everyday life.

One important characteristic of visual search is its efficiency, where by efficiency we refer to the slope relating total search time to the number of stimuli in the display. When search requires a simple discrimination, such as a circle among triangles, a 4 among Cs, or a red item among blue items, the process is highly efficient; the slope of the search function is often near zero. As adding additional items does not incur a cost in search time, such a result implies parallel processing of all elements in the display. Although a search based on a simple feature such as color or line orientation can be done in parallel, when a search target is defined as a conjunction of features (e.g., the target is a red horizontal line segment and distractors are a mixture of black horizontals and red verticals), then search time is often found to increase linearly with the number of stimuli in the display. One interpretation of such results is that efficient 'feature' searches are performed by preattentive mechanisms that can analyze the entire display at once. Searches for conjunctively defined targets, however, require sequential inspection of items, as attention is responsible for conjoining the separate features that constitute a visual object. (Although it is common to take linearly increasing reaction time functions to imply that processing is serial, it has been known for many years that this is not necessarily a safe assumption.)

The feature-integration theory of attention has been very influential. One appealing aspect of this theory is that it is based on known functions of the brain. It seems to suggest that elementary features detectable by early visual processing areas of the brain can be processed over the entire visual field at once, whereas dealing with combinations of elementary features, which presumably requires higher level cortical processing, must be done one item at a time, to avoid a 'combinatorial explosion.' However, over time, some problems with the theory have appeared. One difficulty is that several studies have demonstrated that there are feature combinations that can be processed in parallel. For example, efficient search has been found for brick-shaped objects drawn in such a way that their apparent orientation depended solely on their shading (i.e., which surface looked like it was directly illuminated). That is, a misoriented brick could be found in a background of uniformly oriented distractor bricks independent of the number of those distractors. It is possible to consider such demonstrations as either embarrassments to feature-integration theory or as discoveries of surprising new elementary features (although that claim seems inelegant and unparsimonious, as the number of such new 'features' has become substantial). Worse yet for the theory, there is clear evidence that a feature/shape conjunction can be processed in parallel. There is now even some doubt about the idea that simple features are processed preattentively; this is covered in the next section, in the discussion of the attentional blink. Finally, there are also some interesting effects of familiarity that seem inconsistent with the theory. For example, individuals have been asked to search for a single backward N among several normal Ns or a single normal N among several backward Ns. The latter search was fast and parallel, while the former search was slower and serial. The physical features of the targets and nontargets in these two searches were essentially identical; it is not obvious how the theory can account for the results. Indeed, these data suggest that familiarity itself may be serving as a feature. Could familiarity be detected by early visual areas? There is functional magnetic resonance imaging (fMRI) evidence of widespread reorganization of activity in the visual pathway after extensive training on a perceptual task. Retinotopic cortex is involved, although this is presumably activated in a top-down fashion.

As a result of these problems, several competing interpretations have been proposed. For example, in the 'guided search' model, attention is guided by

preattentive processes to the location of likely target items. The ultimate identification of an item as the target or not requires attentive scrutiny of the item. The efficiency of a search depends on the amount of preattentive information available to guide search. This varies from the great efficiency of searching for a red patch in an array of green patches, through conjunction search, to the extremely inefficient search that occurs when the target and distractors share the same features (e.g., looking for a randomly rotated L among several randomly rotated Ts).

Doing Two Things at Once

In the preceding examples individuals may have had to deal with multiple stimuli, but the task in each case was in some sense unitary (count the objects, repeat back all of the letters, etc.) There has also been a great deal of research in which quite different tasks have been combined; what has been at issue is the extent to which tasks can be performed without mutual interference. The general expectation based on everyday experience is that it is difficult to do several things at once (e.g., pay attention to two simultaneous conversations). In contrast, multitasking is possible in some cases (e.g., driving while listening to the radio). Thus the research effort has largely been directed at finding out just which tasks are difficult to combine, and why. One part of this effort has been a search for special conditions in which tasks might be combined without loss, as such conditions might have implications for real-world scenarios, such as using a cell phone while driving.

Some early research with continuous tasks suggested that even fairly complex pairs of tasks that initially interfered with one another – such as typing visually presented text and shadowing (i.e., repeating back) a message played through earphones – could, after sufficient practice, be done as well together as separately. However, with complex continuous tasks it is often possible to interleave components of the tasks so as to minimize or even eliminate the amount of temporal overlap between components of the two tasks. (An extreme example: a centerfielder could probably read a book during the 'slack time' when he is not actively engaged in the game.) Scientists wishing to achieve a more fine-grained analysis of dual-task performance have used simplified tasks to permit a closer analysis of stimulus–response relations. Perhaps the most well-developed analyses come from the psychological refractory period paradigm. In this paradigm two stimuli are presented, S1 and S2, with a variable stimulus-onset asynchrony (SOA). Typically, study participants are instructed to respond as quickly as possible to each stimulus with the appropriate response, R1 and R2, respectively. When the two tasks are widely separated in time there should be no interference between them and we can obtain baseline reaction times (RTs). As the stimuli are presented closer together in time, the reaction time to the second stimulus (RT2) will tend to increase. Most strikingly, in some cases, as the time between stimuli becomes less than about 300 ms, RT2 increases about 1 ms for each millisecond that the SOA is decreased. It is this kind of result, suggesting a clear inability to produce R2 at the same time as R1, that led to the nomenclature of a refractory period, by analogy to the refractory period of neurons.

Another example of the difficulty of responding to two stimuli presented close together in time is known as the attentional blink (AB). In a typical AB task, a rapid series of stimuli (e.g., digits or letters) is presented at fixation, typically at a rate around 10 items per second, and either one or two targets can appear within the stream. The AB refers to a decrement in the detection or identification of the second target (T2) when it occurs soon after the presentation of the first target (T1), to which a response is required. Early models of the attentional blink assumed that T2 processing is impaired because attention is temporarily fully occupied by processing of T1.

In an instructive example, a stream of letters was presented at fixation. There were two targets. The first was a green letter; all of the other letters in the stream were black. The green letter was followed after a variable delay (the lag) by the second target, which was a ring of Gabor patches (a Gabor patch is, essentially, a small patch of parallel stripes) in the periphery surrounding one of the later letters in the stream. Participants had to name the colored letter and also indicate if all of the Gabor patches were oriented in the same direction or if one was misoriented by 90°. In a control condition, individuals could ignore the stream of letters and just indicate if the ring of Gabor patches contained an orientation oddball. Performance in the control condition was about 90% correct and did not vary with the lag between the green letter and the Gabor patches. In the experimental condition, performance was poor (about 60% correct) when the Gabor patches were simultaneous with the green target letter, and improved to nearly 90% correct when the lag between the green letter and the Gabor patches was 700 ms.

What makes these results interesting are the implications they may have for the study of visual processing. The detection of a Gabor patch differing by 90° in orientation from an otherwise uniformly oriented set of patches should be handled preattentively according to feature-integration theory, and, indeed, the reaction time to detect a Gabor patch

misoriented by 90° is independent of the number of other patches – which is precisely the diagnostic that has been taken to indicate preattentive processing in discussions of feature-integration theory. Why then should there be a large dual-task decrement? One possible answer is that it takes a really difficult competing task, such as is presumably provided by the attentional blink paradigm, to show that orientation discrimination, while easy, is not completely attention free. However, it is also possible that the deficit is not due to T1 stressing attentional capacity for feature processing; it may be that the attentional blink reflects difficulty with one or more higher level functions, such as maintaining executive control of a task set in the face of a fast-moving sequence of targets and nontargets. The attentional demands of various kinds of perceptual discriminations remain a topic of great interest and the final word has not yet been written.

Attention Is Selective

The preceding evidence points to a variety of limitations on attentional capacity. It is clear that we cannot process everything, at least not at once. What is it that determines just what it is that we do process? William James distinguished between passive and active attention. Today, the corresponding terms for passive attention are usually 'bottom-up,' 'stimulus-driven,' or 'exogenous,' while corresponding terms for active attention are 'top-down,' 'goal-directed,' or 'endogenous.' What we attend to is determined both by properties of the stimulus array (e.g., a sudden loud sound, or a movement in the periphery) and by our goals and intentions. Much research has examined the nature of the interaction between, and relative importance of, stimulus-driven and goal-directed sources of control.

The brute fact of selectivity has been demonstrated many times. In the technique known as shadowing, used in studies of dichotic listening, different messages are played to individuals over left and right earphones. Study participants are instructed to repeat back the message played to one ear. They are able to do this quite well. What is especially interesting is the analysis of what information individuals extract incidentally from the unattended message in such circumstances. They can report a change from speech to a pure tone, or a change from a male to a female voice. However very few persons notice a change from English to German or a change from normal to reversed speech when these changes were in the same voice. Most important is the near-total lack of comprehension of the message in the unattended ear. Related experiments have been performed in vision. Participants were shown two episodes on optically superimposed video screens. One was a ballgame and one was a handgame. Participants counted unpredictable events (e.g., the number of times the ball was passed from one player to another). In some conditions individuals were instructed to attend to just one of the episodes; in some conditions they were asked to attend to both. One finding concerned capacity; participants were reasonably accurate when they had to monitor one episode, but they could not deal with both simultaneously. The other finding concerned selectivity. The episodes occasionally contained distinctly odd events (e.g., a woman carrying an umbrella walking through the ballgame). After seeing the episodes, participants were asked about these odd events; they were often entirely unaware of those that occurred in the unattended episode. This kind of result has since been dubbed inattentional blindness and has been the subject of several more analytic studies.

The interaction of stimulus-driven and goal-directed attention is illustrated in a series of studies concerned with the capture of attention by featural singletons (e.g., the lone red element in a display of green elements) or by sudden onsets of a new stimulus object in a field of persisting objects. In one experiment participants searched for a target letter in an array of three or six other letters. This is a demanding search task in which reaction time increases with the number of letters in the display. Stimuli started out as placeholders (like the figure 8 of a digital clock), and at a certain point in time selected segments were deleted, revealing letters. In some conditions a new object was added to the display at the moment when the letters were being revealed (i.e., a letter was displayed in a location that had not contained a placeholder). Importantly, the location of the sudden onset was in no way predictive of the location of the target. This new object attracted attention; if it was the specified target character, reaction times were fast and independent of display numerosity, but if it was not the target, reaction times increased with display numerosity. Thus the sudden onset of a new object appears to capture attention in a stimulus-driven fashion. A follow-up experiment was conducted to determine if there is something special about sudden onsets. In this case the displays contained a single character that was unique with respect to color or brightness (e.g., a red element when other elements were green). Strikingly, reaction time was the same whether or not the display contained a singleton. In other words, unlike a sudden onset, color and brightness singletons did not automatically attract attention. This study raised at least two questions that

have been the subject of further investigation: (1) Is it really the case that feature singletons do not capture attention? (2) Is the sudden onset of a new object really different from other kinds of featural singletons with respect to its ability to capture attention?

Do Feature Singletons Automatically Capture Attention?

In some studies, performance on displays containing an irrelevant singleton is compared to performance when displays contain no irrelevant singleton. For example, in one study participants were presented with displays consisting of colored circles or diamonds arranged around an imaginary circle. A line segment of a variable orientation appeared inside each shape; the task was to determine the orientation of the line segment within a target item. The target item was defined as the unique green diamond among green circles. Search was parallel – that is, independent of display size – but was significantly slower when one of the circles was red than when all elements were green. When the color singleton was made less salient than the shape singleton by changing the colors in the display, the color singleton no longer slowed search for the shape singleton; indeed when the roles of the original color and shape dimensions were reversed (color target and shape distractor), the more salient shape singleton now slowed search for the color target. Note that the color singleton provided no information about the target. This study was interpreted as showing that attention is captured automatically by the most salient element in the display.

An alternative to this analysis is that what an individual first attends to depends not just on the salience of stimuli, but also on the individual's task set. It has been suggested, for example, that salient stimuli may only attract attention when individuals adopt the strategy of searching for a feature discontinuity (singleton-detection mode). When individuals adopt the strategy of searching for a specific, known-to-be-relevant, feature (feature-search mode), bottom-up capture can be overridden. In tasks where both strategies are available to the individual (e.g., where the target is a specific feature and also a singleton, such as a green circle among green triangles), both search strategies are available. Therefore, an irrelevant singleton may or may not cause interference, depending on which strategy the individual adopts. In support of this claim it has been shown that when singleton-detection mode is discouraged, and individuals are forced to resort to feature-search mode, a salient distractor no longer interferes with search.

There has been an ongoing debate about the extent to which the most salient item captures attention automatically, as opposed to capturing attention only when an appropriate strategy has been selected. Be that as it may, the consideration of search strategies has led to new findings. For example, circumstances can be arranged to force individuals to search in time rather than in space. Individuals viewed a central rapid serial visual presentation stream (just as in the attentional blink paradigm) in which a target letter was defined by being a particular color (e.g., red). On critical trials, an irrelevant color stimulus could appear in the periphery during one frame of the stream. Interestingly, the irrelevant stimulus reduced the probability of an individual identifying the target letter, when that irrelevant stimulus preceded the target by 200 ms. That is, even though the peripheral stimulus was irrelevant and did not have to be responded to, it had the same effect as a first target in the attentional blink paradigm. That sounds like automatic capture by the irrelevant stimulus, but there is a critical additional finding: the existence of this interference effect was dependent on whether the individual was in singleton-detection mode or feature-search mode. In one experiment, to induce singleton-detection mode, the letters were all gray except for one that was red. (In fact, half the individuals were tested with red, half with green.) Note that when there is just one colored character in a stream of otherwise all gray characters, either feature search or singleton detection could have been applied. In this experiment, regardless of whether an individual was in the search-for-red or search-for-green group, both red and green peripheral distractors produced substantial, and equal, interference with target detection. In a second experiment, multiple colors were used in constructing the letter stream. Again the target was either red or green (different groups of individuals), but now the individuals had to look for a particular color, whereas in the first experiment they could have searched for 'anything chromatic.' Now, individuals who searched for a red target only suffered interference when the distractor was red, and individuals who searched for a green target only suffered interference when the distractor was green. Follow-up experiments established that the interference was due to a spatial capture of attention.

Are Sudden Onsets Special?

An influential account of how performance in search tasks might be affected by nominally irrelevant singletons is that participants in search experiments always adopt one or another sort of attentional control setting that determines what will be selected. (We have already come across this notion; singleton detection and feature search can be considered specific examples of attentional control settings.) To test this

idea, participants searched for a target character that could be either a color or an onset singleton across blocks. Immediately preceding each search display another display appeared that contained a cue that was itself either a color or an onset singleton in one of the (four) potential target locations. It was explained to the participants that the location of the cue was independent of the location of the target. The chief result was that target reaction times were strongly dependent on the cue location, but only when the cue matched the target dimension. For example, if the task was to identify a color singleton, then reaction time was faster when the target was in the same location that the cue had occupied than when it was in another location. However, when participants had to identify a color singleton, the location of an onset cue had no effect on reaction time. This finding is contrary to the notion that the sudden onset of a new object is unique in that it captures attention automatically (i.e., in a stimulus-driven fashion). The possibly special status of sudden onsets is controversial and remains a topic of lively current interest.

Conclusion

The study of attention has been of keen interest to psychologists for over a century. A great deal of progress has been made in understanding the limitations of our perceptual and cognitive processes, and the selectivity these limitations force on us. Most of this research has used psychophysical methods solidly within the behavioral tradition. However over the past 20–30 years the conceptual and methodological armamentarium available to study attention has been expanded in scope by the growth of the neurosciences.

See also: Attention: Models; Attentional Networks; Attentional Networks in the Parietal Cortex; Attentional Functions in Learning and Memory; Human Methods: Psychophysics; Short Term and Working Memory; Visual Attention.

Further Reading

Burnham BR (in press) Displaywide visual features associated with a search display's appearance can mediate attentional capture. *Psychonomic Bulletin & Review.*

Cowan N (2001) The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24: 87–185.

Enns JT and Rensink RA (1990) Sensitivity to three-dimensional orientation in visual search. *Psychological Science* 1: 323–326.

Folk CL, Remington RW, and Johnston JC (1992) Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception & Performance* 18: 1030–1044.

Jonides J and Yantis S (1988) Uniqueness of abrupt visual onset in capturing attention. *Perception and Psychophysics* 43: 346–354.

Joseph JS, Chun MM, and Nakayama K (1997) Attentional requirements in a 'preattentive' feature search task. *Nature* 379: 805–807.

Leber AB and Egeth HE (2006) It's under control: Top down search strategies can override attentional capture. *Psychonomic Bulletin & Review* 13: 132–138.

Lien M-C, Ruthruff E, and Johnston JC (2006) Attentional limitations in doing two tasks at once: The search for exceptions. *Current Directions in Psychological Science* 15: 89–93.

Mandler G and Shebo BJ (1982) Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General* 11: 1–22.

Neisser U and Becklen R (1975) Selective looking: Attending to visually specified events. *Cognitive Psychology* 7: 480–494.

Olivers CNL, Stigchel S van der, and Hulleman J (2005) Spreading the sparing: Against a limited-capacity account of the attentional blink. *Psychological Research* 8: 1–14.

Pashler H (1998) *The Psychology of Attention.* Cambridge, MA: MIT Press.

Pylyshyn ZW and Storm RW (1988) Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision* 3: 179–197.

Sperling G (1960) The information available in brief visual presentations. *Psychological Monographs* 74(no. 11): 1–30.

Theeuwes J (2004) Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review* 11: 65–70.

Treisman A and Gormican S (1988) Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review* 95: 15–48.

Wang Q, Cavanagh P, and Green M (1994) Familiarity and pop-out in visual search. *Perception & Psychophysics* 56: 495–500.

Wolfe JM (2003) Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences* 7: 70–76.

Psychotherapeutic Approaches to Psychiatric Disorders

G O Gabbard and E Poa, Baylor College of Medicine, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Definitions

Despite impressive advances in the neurosciences and psychopharmacology over the last several decades, psychotherapy continues to be a major treatment modality in the field of psychiatry. A growing body of efficacy research demonstrates that psychotherapeutic interventions can produce substantial improvements in a great many psychiatric disorders. Psychotherapy can be conducted in a one-to-one context (i.e., individual psychotherapy), or it may occur with one or two therapists in the room with a family, a couple, or a group of unrelated individuals. Many schools of psychotherapy are practiced. In this overview, we focus on those that are most commonly used and for which a greater research base has been established.

Psychodynamic or psychoanalytic psychotherapy is the grandfather of all psychotherapies. Growing out of Freud's development of psychoanalysis as a treatment, this form of therapy is based on a set of fundamental principles. These include the notion that much of mental life is unconscious and that the psychotherapist helps patients gradually accept how unconscious influences affect their experiences of themselves and others. Psychodynamic therapists view the patient's transference to the therapist as a primary source of understanding. In other words, a patient repeatedly attributes qualities, thoughts, and feelings to the therapist that actually are connected with figures from the patient's past. Analogously, psychodynamic therapists also focus on their own emotional reactions to the patient – the so-called countertransference – as another important tool in understanding the patient and the process. A key conceptual point in dynamic therapy is that patients construct a set of psychological defense mechanisms early in life to avoid the experience of emotional pain. These defenses emerge as resistances to the psychotherapy process when the patient enters treatment. Hence, a succinct definition of psychodynamic therapy is a process that involves careful attention to the therapist–patient interaction with carefully timed interpretation of transference and resistance embedded in a sophisticated appreciation of the therapist's contribution to the two-person field.

Cognitive behavioral therapy (CBT) is the psychotherapy with the largest empirical base to support its efficacy. It is based on two core principles: (1) the way people act or behave strongly influences their feelings and thoughts, and (2) cognitions themselves exert a powerful influence on behavior and feelings. Cognitive therapists look at levels of cognitive processing and devote a great deal of their attention to identifying automatic thoughts and schemas. They commonly assign homework for their patients. CBT may also involve a daily log of thoughts or prescribed exposure to feared situations. Cognitive therapists frequently use Socratic questioning to stimulate inquisitiveness and curiosity in the patient.

Interpersonal psychotherapy (IPT) was developed as a treatment for depression, but it has since been expanded to other conditions. At its basic level, IPT was targeted to improve the quality of a depressed patient's interpersonal world. In other words, IPT therapists help patients clarify and understand the common interpersonal themes that are involved in the onset or maintenance of a depressive disorder. Four common theme areas that are the focus of IPT are unresolved grief, social role disputes, social role transitions, and interpersonal deficits. Direct interventions are identified based on which of the four themes is most prominent in a particular patient's situation. This therapy is also reeducative in the sense that it explains diagnosis and treatment to the patient. Based on this explanation, the therapist collaborates with the patient on a plan that addresses the interpersonal issues that are pivotal in the pathogenesis of the disorder.

Supportive psychotherapy (SP) is a widely practiced treatment modality that often lacks systematic rigor. SP is not simply offering the patient support in the way that a friend might help someone who is suffering. It is a treatment that involves a careful assessment of the patient's strengths and weaknesses and the use of direct measures to restore or improve self-esteem, ameliorate symptoms, and improve adaptive functioning. Defense mechanisms are questioned only when they are maladaptive, in contrast to the more exploratory version of psychodynamic therapy, where defenses are understood and confronted. Transference phenomena are generally not discussed but simply noted silently by the therapist. A great deal of attention is placed on establishing a solid therapeutic alliance so that the patient trusts the therapist and feels helped. In addition, the therapist tries to reinforce sensible thinking and behavior, and to praise, reassure, encourage, refrain, and advise. An important subcomponent of SP is psychoeducational therapy, which employs education of the patient and family as a major thrust.

These definitions have considerable overlap. Many practitioners use combinations of techniques geared to the patient's specific individual needs. Psychotherapy

research has demonstrated that the therapeutic relationship, particularly when the therapist is flexible and responsive, accounts for most of the outcome variance. Technique probably accounts for only 12–15% of the variance across different kinds of therapies. Nevertheless, for purposes of discussion we consider these therapies separately and examine the research base that supports them. We also present some preliminary research in the area of how psychotherapy may affect the brain.

Efficacy and Indications

In this section, we examine the efficacy data and indications for each of the major therapeutic modalities. Although we focus on randomized controlled trials (RCTs), we also acknowledge that there is considerable controversy regarding whether psychotherapy as a practice is well suited to RCTs. Because of the requirements for manualization and standardization, the treatments delivered in psychotherapy studies often do not reflect clinical practice. Similarly, the patient sample selected for such studies is often unrepresentative of the patients seen by clinicians in community practice settings. Because of the cost of psychotherapy research, most of the treatments delivered in RCTs are unrepresentatively brief. Research also has demonstrated that the researcher's allegiance to one brand of treatment or another is probably the best predictor of outcome. The evidence-based approach used in the rest of psychiatry is problematic because it does not take into account individual differences among therapists. Finally, the studies seeking to prove the efficacy of a particular psychotherapeutic approach are almost always disorder-based, even though may patients wish to discuss problems of living that are inextricably intermingled with specific disorders.

Psychodynamic Psychotherapy

Short-term psychodynamic therapy

Individual therapy. Although psychodynamic therapists have been slow to build a research base, a growing number of RCTs now document the efficacy of short-term psychodynamic psychotherapy (STPP), usually defined as lasting no more than 6 months or 24 sessions, in the treatment of axis I psychiatric disorders. Four RCTs examined the use of STPP in major depression, finding it to be effective for depressive symptoms, general psychiatric symptoms, and impaired social functioning. Combined treatment studies examining the addition of psychodynamic psychotherapy to medication regimens for depression have revealed greater improvement as well as decreased dropout rates, an effect which increases with time.

One study has shown STPP to be useful in decreasing trauma-related symptoms in posttraumatic stress disorder (PTSD), with continued improvement at the 3-month follow-up. Research evidence also supports the use of STPP in eating disorders, with two studies showing significant and lasting improvements in bulimia nervosa and one study demonstrating increases in weight and body mass index changes in anorexia nervosa.

Another area in which STPP has been used to advantage is in substance-use disorders. In one study, STPP was effective in measures of drug-related and psychiatric symptoms among opiate-dependent subjects. In a second study, this modality led to significant declines in alcohol abuse, which remained at the 15-month follow-up.

Group therapy. STPP delivered in a group setting has also shown significant benefits in borderline personality disorder. One RCT showed that this treatment reduced the ratings of borderline symptomatology while also demonstrating improvements in general psychiatric and depression symptoms.

Long-term therapy

Individual therapy. Evidence also exists for the use of long-term psychodynamic psychotherapy (LTPP) in both axis I and II disorders. One RCT showed LTPP to be efficacious in treating generalized social phobia. LTPP has also yielded significant symptomatic improvements in both bulimia nervosa and anorexia nervosa. The value of LTPP in substance-use disorders was confirmed in one study, which showed it to be significantly superior to drug counseling in opiate dependence.

One area in which LTPP has been highly beneficial is in the treatment of personality disorders. In one study, a psychoanalytically oriented partial hospitalization program for borderline personality disorder lasting 18 months led to dramatic decreases in self-harm behavior, suicide attempts, inpatient days, medication use, and self-reported psychiatric symptoms. The main difference between the treatment as usual control group and the experimental group was that the patients in the latter group received individual and group psychotherapy of a psychoanalytic nature. The benefits of the treatment were clear at 6 months and continued to accrue up to the time of termination. In fact, compared to the controls, the patients in the experimental group showed continued improvement during the postdischarge period. In another study, psychodynamic psychotherapy was compared to SP and dialectical behavior therapy (DBT) in a 12-month randomized controlled trial of borderline personality disorder treatment. All three treatments were equally effective, although the dynamic

treatment made statistically significant gains in increasing mentalization compared to the other two.

Evidence also exists for the efficacy of psychodynamic therapy for cluster C personality disorders. In a Norwegian study, patients with these conditions were randomly assigned to 40 weeks of either dynamic or cognitive therapy. Both groups showed significant improvements that remained stable at a 24-month follow-up, and those who received dynamic therapy demonstrated continued posttreatment improvement at a 2-year follow-up.

Group therapy. There has also been some limited evidence for the use of STPP in the treatment of alcoholism; fewer people completed the psychodynamic therapy compared to a behavioral group therapy, but of those who did, a greater number were abstinent than those who completed the behavioral process.

Meta-analyses Two recent meta-analyses have looked at psychodynamic psychotherapy. One meta-analysis of 14 studies of STPP and LTPP in personality disorders showed significant effect sizes in both self- and observer-rated measures. Included were two RCTs that also exhibited significant effectiveness over control conditions. In these studies, there appeared to be a correlation between treatment duration and effect size.

Another meta-analysis evaluated the efficacy of psychodynamic psychotherapy, with durations ranging from 7 to 40 sessions. This investigation examined its use in psychiatric disorders in general, due to an inadequate number of studies focused on individual disorders for the meta-analysis. Overall, it was shown that psychodynamic therapy led to significant improvements in target problems, general psychiatric symptoms, and social functioning. A summary of evidence-based indications for psychodynamic psychotherapy is in Table 1.

Table 1 Evidence-based indications for psychodynamic psychotherapy

Therapy	*Indications*
Short-term	Major depression
	Posttraumatic stress disorder
	Bulimia nervosa
	Anorexia nervosa
	Substance-use disorders
	Borderline personality disorder
Long-term	Social phobia
	Bulimia nervosa
	Anorexia nervosa
	Substance-use disorders
	Borderline personality disorder
	Cluster C personality disorders

Cognitive Behavioral Therapy

Individual therapy CBT lends itself to manualized treatments, so there has been an abundance of experimental data. Extensive evidence supports the use of CBT in axis I disorders.

In depression, CBT has been shown to be effective in nonpsychotic, unipolar depression and in decreasing rates of relapse. These findings are particularly robust in mild to moderate major depression. It has also demonstrated some benefit in patients with a poor response to medication intervention. Some studies have indicated that the combination of CBT and medications may provide a benefit surpassing that of CBT or medications alone, especially when the rates of relapse are examined. A limited number of studies, including both psychotic and nonpsychotic depressed patients, showed considerable benefit from the addition of CBT to medications.

Studies have shown CBT to be effective for panic and anxiety symptoms, indicating that it may be more effective than medications in some instances. In studies of panic disorder, CBT has proven to be more effective than both placebo and medication in terms of symptom reduction as well as tolerability. The use of behavioral therapy in obsessive–compulsive disorder (OCD) has been widely accepted in the form of exposure and ritual prevention (ERP), and has shown significant benefit both for symptom relief and for relapse prevention, even in comparison to medication treatment. Evidence also exists for its use in social phobia. A recent meta-analysis of CBT for generalized anxiety disorder shows this treatment to be highly effective and better tolerated than medications.

CBT has also been effective in the treatment of another anxiety disorder, PTSD. One meta-analysis of CBT found that a brief course produced substantial improvement, with a 56% recovery rate for those entering the studies. Moreover, 67% of those completing the protocol no longer met the criteria for PTSD.

Multiple studies have shown significant benefits from CBT in acute and chronic schizophrenia. These improvements were noted in overall symptomatology as well as in positive and negative symptoms. On follow-up, these gains were sustained across studies. Investigators have noted that the use of CBT may also be able to decrease the likelihood of schizophrenia progressing in patients with prodromal symptoms.

A growing area of research in the use of CBT is in the treatment of personality disorders. As previously noted, CBT appears to be efficacious with cluster C patients. One particular outgrowth of CBT has been DBT, originally designed for borderline personality disorder. This approach, consisting of one weekly individual session and one skills group per

week, has been shown to decrease suicidal behaviors, self-harm, substance abuse, and binge eating. Moreover, greater numbers of patients tend to remain in treatment, and those that do require less inpatient treatment. Studies have also shown promising benefits for borderline patients with comorbid substance abuse.

A smaller body of research suggests that CBT is useful for other conditions as well. CBT has significant evidence supporting it as perhaps the most effective treatment for bulimia nervosa. Finally, CBT has shown benefits for alcohol and cocaine abuse.

Although CBT has proven its utility as a primary treatment, researchers have also examined its use in promoting treatment compliance. Patients suffering from bipolar disorder have greater medication compliance and decreased relapses/hospitalizations when a short course of CBT is added to their medications. CBT also increases adherence to medication when used in patients with schizophrenia.

Group/family therapy The use of CBT-oriented family therapy has proven to be a helpful adjunct in the treatment of disorders with traditionally high relapse rates. This applies to schizophrenia, in which certain types of family involvement with a patient who has schizophrenia has led to decreased rates of relapse in the identified patient. These same findings hold true for patients with comorbid schizophrenia and substance-use disorders. In addition to a decline in relapse, there was a decrease in overall symptoms. Similar results have been found with bipolar disorder.

Patients with alcohol abuse found CBT-oriented groups useful in decreasing alcohol consumption. CBT groups have been effective in the treatment of marijuana abuse as well. Patients with both bipolar disorder and substance-use disorders have gained benefits, as shown in measures of addiction severity and increased rates of abstinence. Behaviorally oriented couples and family therapy have also been effective in the treatment of substance abuse. Finally, marital therapy using behavioral interventions has shown benefit for a variety of marital problems. A summary of the indications for CBT is in Table 2.

Table 2 Evidence-based indications for cognitive behavioral therapy

Therapy	*Indications*
Individual	Major depression
	Bipolar disorder
	Schizophrenia
	Panic disorder
	Obsessive–compulsive disorder
	Social phobia
	Generalized anxiety disorder
	Posttraumatic stress disorder
	Bulimia nervosa
	Substance-use disorders
	Borderline personality disorder
	Treatment compliance problems
Group/family	Schizophrenia
	Bipolar disorder
	Substance-use disorders
	Marital conflict

Interpersonal Therapy

Individual therapy IPT is effective for a number of different psychiatric illnesses. The majority of the studies have examined IPT in affective disorders. These have shown clearly that IPT is efficacious in major depression and may be as effective as medication treatment in some forms of mild to moderate depression. In addition to reducing overall depression ratings, it has been particularly beneficial in increasing self-esteem, social functioning, and overall functioning. Research with IPT targeting elderly patients suggests that it improves treatment response to pharmacotherapy. When that same population enters remission, findings indicate that it is also effective as a maintenance therapy.

IPT has also been applied to bipolar disorder. This research has shown that IPT patients are more likely to remain in a euthymic state, compared to those receiving standard clinical management. In addition, IPT has also been effective when applied to bulimia nervosa.

Group/family therapy Less research has accrued on IPT applied in group and family therapies, but a small body of findings shows promise. IPT-based family therapy demonstrated benefit when treating younger patients with recent onset of anorexia nervosa. Group therapy based on IPT was successful in decreasing bulimic symptoms in individuals suffering from non-purging bulimia nervosa.

IPT focused groups have been used for the treatment of adult victims of childhood sexual abuse as well. In this population, it has led to a decrease in depression, anxiety, trauma-related symptoms and a decrease in the global severity of psychiatric symptoms, as well as an increase in social adjustment.

Pilot studies show that IPT-based couples therapy improves marital satisfaction and decreases depression. Another study combined IPT and supportive interventions in a group of heroin-dependent mothers, leading to decreased opioid use and child maltreatment as well as increased parent–child interaction and overall adjustment. Table 3 summarizes the indications for IPT.

Table 3 Evidence-based indications for interpersonal therapy

Therapy	*Indications*
Individual	Major depression
	Bipolar disorder
	Bulimia nervosa
Group/family	Anorexia nervosa
	Bulimia nervosa
	Substance-use disorders
	Childhood sexual trauma
	Marital conflict

Table 4 Evidence-based indications for supportive interventions

Therapy	*Indications*
Individual	Schizophrenia
	Bipolar disorder
	Borderline personality disorder
Group/family	Schizophrenia
	Bipolar disorder
	Substance-use disorders

Supportive and Psychoeducational Therapy

Individual interventions SP has often been denigrated by psychiatrists and other mental health professionals, but recent empirical research suggests that it is a modality that is of substantial benefit in psychiatric illness. In schizophrenia, for example, psychoeducational strategies have led to increased patient knowledge and medication compliance. SP has also been shown to increase occupational functioning in this same group. Vocational rehabilitation has displayed consistent results in improving rates of competitive employment, which is often a source of considerable gratification to the patient.

In a recent RCT involving SP, DBT and psychodynamic therapy for borderline personality disorder, SP did just as well as the other two modalities on most measures. In addition, psychoeducation has proven to be useful in improving social and occupational functioning in patients with bipolar disorder, as well as increasing the time before a subsequent manic episode.

Group/family interventions SP family and group interventions are highly useful in the overall treatment of schizophrenia. Especially in families with high expressed emotion, psychoeducational family therapy helps to decrease psychotic symptoms, prevent relapse, and improve social and vocational functioning in the person with schizophrenia. Levels of expressed emotion are also decreased. A majority of studies looking at the use of assertive community treatment programs using intensive and proactive case management in schizophrenia have shown a decreased need for inpatient hospitalization, as well as beneficial effects in symptomatology, housing stability, and subjective quality of life. Social skills training groups have led to improved behavioral social skills, but have not necessarily translated to better functioning overall in patients with schizophrenia.

Family psychoeducational therapy for patients with bipolar disorder has led to decreased relapse rates in multiple studies. It also provided a higher quality of family life in one study, whereas another investigation showed decreased symptoms and improved global functioning, particularly in female patients. Similarly, marital psychoeducation showed substantial gains in medication compliance and global assessment scales for individuals with bipolar illness.

SP group interventions may be extremely useful in problems of substance abuse. They have been shown to be efficacious in smoking cessation. Social skills training groups for alcoholics have led to a decrease in alcohol consumption and an increase in the number of abstinent days. Table 4 summarizes the indications for SP.

The Impact of Psychotherapy on the Brain

The therapeutic action of psychotherapy is now being investigated with the overarching assumption that successful psychotherapy must work by changing the brain. Kandel has demonstrated how synaptic connections can be permanently altered and strengthened through the regulation of gene expression based on learning from the environment. Neurobiological studies of the sea slug *Aplysia* have shown that the number of synapses double or triple as a result of learning. Kandel postulates that psychotherapy should be considered a form of learning about oneself and one's relationship to others. The learning process that occurs in psychotherapy may produce alterations of gene expression and thereby alter the strength of synaptic connections. Kandel clarifies that the sequence of a gene (the template function) cannot be affected by environmental experience but that the ability of the gene to direct the manufacture of specific proteins (the transcriptional function) is quite responsive to environmental factors and regulated by these influences.

Neural imaging studies have allowed investigators to document structures and functions within the brain that are affected by psychotherapy. One investigation of patients with OCD showed that both behavior therapy and fluoxetine appear to produce similar decreases in cerebral metabolic rates in the head of the right caudate nucleus. Another study compared

24 subjects with major depression who received either paroxetine or IPT. At baseline, subjects had higher normalized metabolism than normal controls in the prefrontal cortex, caudate, and thalamus and lower metabolism in the temporal lobe. After treatment with either paroxetine or IPT, positron emission tomography (PET) scans showed that the subjects had metabolic changes in the direction of normalization in these regions. Regional metabolic changes appeared similar in the two forms of treatment.

In other studies, medication and psychotherapy appear to have separate and distinct influences on the brain. Goldapple and her colleagues scanned 17 unmedicated, unipolar depressed patients before and after a 15–20-session course of CBT using PET. They compared the findings to a separate group of 13 depressed patients who responded to paroxetine. The psychotherapy appeared to alter brain regions that medication did not touch. The psychotherapy was associated with increases in metabolic activity in the anterior cingulate and the hippocampus, with decreases in metabolic activity in the dorsal, ventral, and medial-frontal cortices. By contrast, paroxetine showed increases in metabolic activity in the prefrontal cortex and decreases in the brain stem and subgenual cingulate. In brief, therapy seemed to work in a top-down manner, whereas medication worked in a bottom-up manner.

Cognitive therapy appears to influence thyroid hormone levels in those who suffer from major depression. In one study, patients who responded to CBT showed striking decreases in their levels of thyroxin (T_4), whereas nonresponders showed increases in T_4. In another study of depressed patients, CBT seemed to affect the biology of sleep. In fact, the biological changes in sleep architecture were identical to those produced by antidepressant medication.

Psychotherapy may also make modifications in procedural memory. What unfolds in the relationship with the therapist is the patient's stereotyped, automatic, habitual mode of relatedness shaped by attachment relationships in the first years of life. These relational configurations become encoded in procedural memory and operate outside conscious awareness. Psychotherapy may provide a new attachment relationship that is able to restructure attachment-related, implicit procedural memory. Stored prototypes of self and other representations encoded in implicit procedural memory may be modified by new interactions with the therapist that are internalized by the patient. In other words, old neural networks of self and other from childhood may be progressively weakened in psychotherapy while new neural networks of self and other representations are simultaneously strengthened.

Freud's dream was that the changes brought about by psychoanalytic treatment could ultimately be understood in terms of the neurobiological substrate of the brain. Although a great deal of research is needed to clarify the exact nature of these changes, we now stand on the threshold of an era when the technology is available to fulfill Freud's dream.

See also: Emotional Disorders: Treatment; Obsessive–Compulsive Disorder; Posttraumatic Stress Disorder: Overview; Posttraumatic Stress Disorder: Neurobiology; Stress Response: Genetic Consequences.

Further Reading

Bateman A and Fonagy P (1999) Effectiveness of partial hospitalization in the treatment of borderline personality disorder: A randomized controlled trial. *American Journal of Psychiatry* 156: 1563–1569.

Baxter LR, Shortz JM, Bergman KS, et al. (1992) Caudate glucose metabolic rate changes with both drug and behavior therapy for obsessive-compulsive disorder. *Archives of General Psychiatry* 49: 618–689.

Beck AT (2005) The current state of cognitive therapy: A 40-year retrospective. *Archives of General Psychiatry* 62: 953–959.

Fairburn CG and Harrison PJ (2003) Eating disorders. *Lancet* 361: 407–416.

Gabbard GO (2004) *Long-term Psychodynamic Psychotherapy: A Basic Text*. Arlington, VA: APPI.

Gaudiano BA, Beevers CG, and Miller IW (2005) Differential response to combined treatment in patients with psychotic versus nonpsychotic major depression. *Journal of Nervous and Mental Disease* 193: 625–628.

Goldapple K, Segal Z, Garson C, et al. (2004) Modulation of cortical-limbic pathways in major depression: Treatment-specific effects of cognitive behavior therapy. *Archives of General Psychiatry* 61: 34–41.

Gunderson JG and Gabbard GO (1999) Making the case for psychoanalytic therapies in the current psychiatric environment. *Journal of the American Psychoanalytic Association* 47: 679–704.

Kandel ER (1998) The new intellectual framework for psychiatry. *American Journal of Psychiatry* 155: 457–469.

Leichsenring F (2005) Are psychodynamic and psychoanalytic therapies effective?: A review of empirical data. *International Journal of Psycho-Analysis* 86: 841–868.

Leichsenring F, Rabung S, and Leibing E (2004) The efficacy of short-term psychodynamic psychotherapy in specific psychiatric disorders. *Archives of General Psychiatry* 61: 1208–1216.

Mitte K (2005) Meta-analysis of cognitive-behavioral treatments for generalized anxiety disorder: A comparison with pharmacotherapy. *Psychological Bulletin* 131: 785–795.

Rector NA and Beck AT (2001) Cognitive behavioral therapy for schizophrenia: An empirical review. *Journal of Nervous and Mental Disease* 189: 278–287.

Robertson M, Humphreys L, and Ray R (2004) Psychological treatments for posttraumatic stress disorder: Recommendations for the clinician based on a review of the literature. *Journal of Psychiatric Practice* 10: 106–118.

Shadish WR and Baldwin SA (2005) Effects of marital therapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology* 73: 6–14.

Wampold BE (2001) *The Great Psychotherapy Debate: Models, Methods, and Findings*. Mahwah, NJ: Laurence Erlbaum.

Westen D, Novotny CM, and Thompson-Brenner H (2004) The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin* 130: 631–663.

Winston A, Rosenthal RN, and Pinsker H (2004) *Introduction to Supportive Psychotherapy.* Arlington, VA: APPI.

Wright JH, Basco MR, and Thase ME (2006) *Learning Cognitive--Behavior Therapy: An Illustrated Guide*. Arlington, VA: APPI.

Pulsatility in Neuroendocrine Systems

G Leng, N Sabatier, and C Caquineau,
The University of Edinburgh College of Medicine and Veterinary Sciences, Edinburgh, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A major role of the neuroendocrine hypothalamus is to generate pulsatile patterns of activity that are appropriate for the secretion of each of the pituitary hormones and to regulate these patterns according to specific physiological requirements. How the hypothalamus generates the various patterns varies from system to system; there is no common 'pulse generator' into which different systems can tap to generate their individual rhythms. However, what has been learned of the three best studied systems – oxytocin, the gonadotrophins, and growth hormone (GH) – indicates that there are commonalities as well as differences in the underlying mechanisms.

Oxytocin and vasopressin are secreted from the posterior pituitary as a direct result of the electrical activity of the hypothalamic magnocellular neurons that produce these hormones; thus, pulsatile secretion of oxytocin reflects synchronized burst activation of the oxytocin neurons. However, the process is more complex for the hormones secreted from the anterior pituitary gland. GH, prolactin, thyroid-stimulating hormone, ACTH, follicle-stimulating hormone (FSH), and leutinizing hormone (LH) are all secreted in a markedly pulsatile manner, in at least some physiological circumstances. For the most part, their secretion is regulated by releasing factors and release-inhibiting factors. These factors are released into the hypothalamo-hypophysial portal circulation from the nerve terminals of hypothalamic neuroendocrine neurons that project to the median eminence. In general, the secretion of a pulse of a hormone from the anterior pituitary gland follows release of a pulse of a releasing factor into the portal circulation, and it can be inferred that the pulse of releasing factor release in turn reflects synchronized pulsatile activation of the neurons that make the releasing factor. However, this inference has not been demonstrated directly; the only identified neuroendocrine neurons in which synchronized electrical activation has been directly observed during reflex hormone secretion are the magnocellular oxytocin neurons. These play a key role in forming the paradigm by which we understand the origins and significance of pulsatile hormone secretion.

Oxytocin

In a lactating rat, oxytocin is secreted into the blood in short-lived pulses while the mother is being suckled by the young. These pulses are the result of a brief, synchronized burst discharge of all the magnocellular neurosecretory oxytocin neurons in the supraoptic and paraventricular nucleus of the hypothalamus. Each burst lasts 1–3 s, and during a burst the firing rate of the neurons can exceed 100 action potential/s. The normal spontaneous firing rate of these neurons between bursts is just 1–3 action potential/s. These 'milk ejection' bursts occur at intervals of typically 3–5 min while the young are suckling. The fact that the milk ejection reflex involves intermittent pulsatile response to a continuous suckling stimulus appears to be a general feature in mammals, including humans (Figure 1).

In the rat, each milk ejection burst releases approximately 1 mU of oxytocin into the blood from the neurosecretory nerve terminals in the posterior pituitary gland. The released hormone initially travels as a bolus in the blood, and it disperses quickly as it is mixed with the total blood volume. The disappearance rate of the bolus is just a few seconds, much less than the half-life of oxytocin in the blood. This fast disappearance rate makes it very difficult to 'catch' the pulses for measurement by radioimmunoassay of blood samples, but the pulses can readily be detected by online bioassay – measurement of intramammary pressure changes by cannulation of the mammary gland.

The burst discharge of oxytocin cells is still not fully understood. Between bursts, the neurons show little, if any, sign of coordination in their spontaneous activity. However, oxytocin, acting within the hypothalamus, plays a key role in the burst behavior. Oxytocin that is secreted into the blood does not reenter the brain in significant amounts because of the blood–brain barrier, and oxytocin injected peripherally has no effect on the activity of oxytocin neurons, but very large amounts of oxytocin are released from the dendrites of the oxytocin neurons. This dendritic release is regulated semi-independently of peripheral, axonal secretion; dendritic release can be triggered, for instance, by peptides that act on oxytocin neurons to mobilize intracellular calcium stores. Oxytocin injected in small amounts into the brain facilitates the suckling-induced reflex, whereas oxytocin antagonists injected into the brain block it.

Oxytocin is released from the dendrites in response to suckling before any burst discharge is seen. The oxytocin neurons have oxytocin receptors, and activation of these induces calcium mobilization and

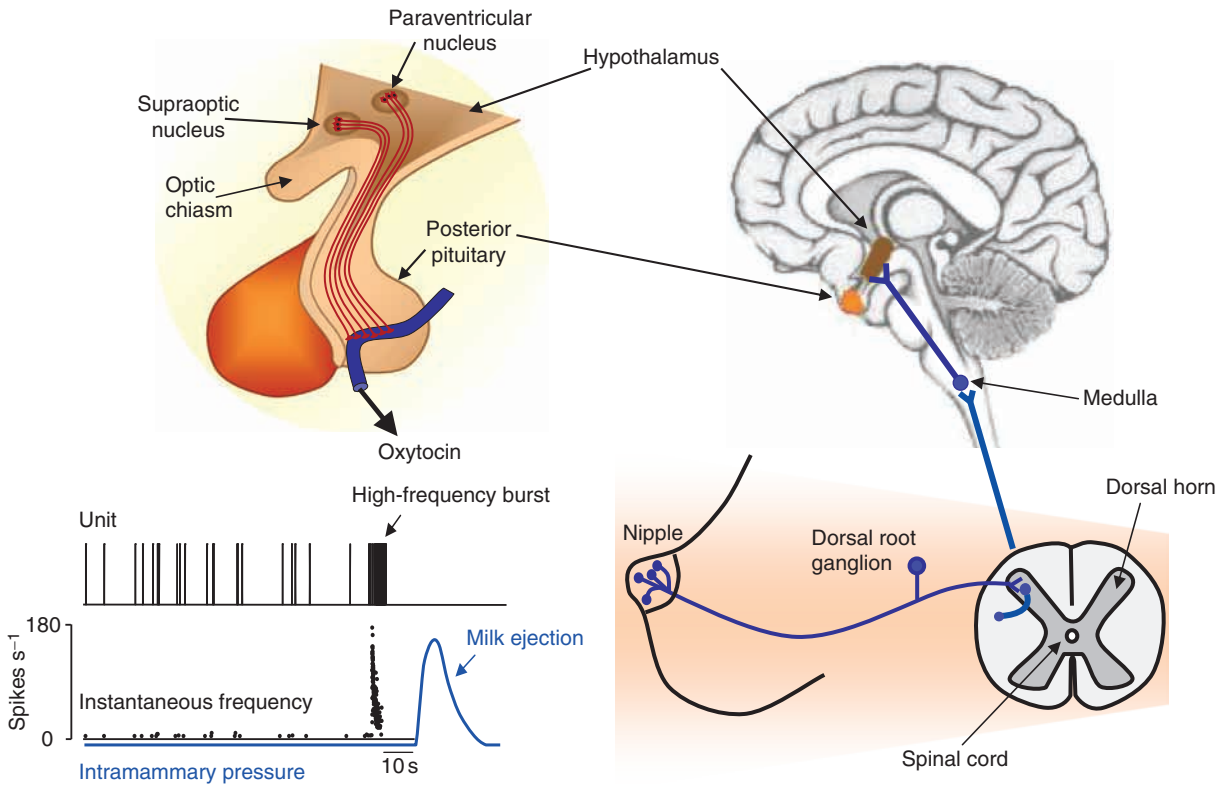

Figure 1 The milk ejection reflex. At the posterior pituitary gland, oxytocin is released from the neurosecretory nerve terminals of magnocellular neurosecretory neurons whose cell bodies are in the supraoptic and paraventricular nuclei of the hypothalamus. When young suckle at the nipples of a lactating mother, the sensory signals travel via the spinal cord and brain stem, resulting in synchronized bursting of the oxytocin neurons. This behavior, first observed by Jonathan Wakerley and Dennis Lincoln in the 1970s, leads to a brief pulse of oxytocin which triggers milk let-down and a transient increase in intramammary pressure. The inset shows the burst from one oxytocin cell preceding an intramammary pressure rise. See Lincoln DW, and Wakerley JB (1974) Electrophysiological evidences for the activation of supraoptic neurones during the release of oxytocin. *Journal of Physiology* 242: 533–554; Wakerley JB and Lincoln DW (1973) The milk-ejection reflex the rat: A 20- to 40-fold acceleration in the firing of paraventricular neurones during oxytocin release. *Journal of Endocrinology* 51: 477–493.

further oxytocin release; thus, there is a positive feedback loop. Oxytocin affects the electrical activity of oxytocin neurons in several ways: It suppresses afferent input to the oxytocin neurons, tending to reduce their spontaneous activity, but it also has a slight direct depolarizing action. In addition, oxytocin can 'prime' dendritic stores of oxytocin, making them more available for subsequent activity-dependent release. The consequence of these factors appears to be that during suckling, oxytocin cells become less responsive to external inputs but increasingly capable of mutual excitation.

The amount of oxytocin secreted as a result of a burst is very large even considering the high rate of spike activity. This is because the nerve terminals in the posterior pituitary show a marked 'frequency facilitation' of stimulus-secretion coupling. An action potential within a burst triggers approximately 100 times as much oxytocin secretion as an action potential that arises spontaneously between bursts. The frequency facilitation is a result in part of the nonlinear effects on exocytosis of the summation of calcium entry through voltage-dependent calcium channels opened by the depolarizing effects of action potentials at nerve terminals. In addition, action potential propagation in the axonal fields in the neural lobe is more efficient in axons that are partly depolarized as a result of potassium efflux from neighboring depolarized axons; thus, action potentials within a burst are less likely to 'fail' at branch points, and they therefore invade more of the neurosecretory terminals.

Synchronized bursting is thus a very efficient way of secreting a large bolus in a brief time. The resulting pulsatile hormone signal is also very efficient for stimulating milk let-down at the mammary gland. When exposed to continuous high concentrations of oxytocin, the myoepithelial cells of the mammary gland on

which oxytocin acts desensitize rapidly and become refractory. Accordingly, pulsatile secretion is necessary for the biological efficacy of oxytocin at this site.

Oxytocin is also secreted in pulses during parturition in the rat and other mammals, including humans. Oxytocin is not essential for normal delivery in rodents because transgenic mice without oxytocin can deliver young normally. However, in the rat, the progress of parturition can be interrupted by giving oxytocin antagonists or by pharmacologically blocking oxytocin secretion. In the rat, pulsatile secretion is important for the efficient progress of delivery of the young; when endogenous oxytocin secretion is blocked, parturition is interrupted, but it can be restored by giving exogenous oxytocin. This is achieved much more efficiently by delivering oxytocin in pulses rather than by continuous infusion.

In humans, the role of oxytocin in normal delivery is less clear. In pregnancy, an enzyme that degrades oxytocin, human pregnancy oxytocinase, is highly expressed. When first appreciated, this fact was taken as evidence that oxytocin plays no physiological role in parturition in humans, although its pharmacological potency in inducing uterine contractions has long been known. However, this view has undergone some reappraisal as the importance of pulsatile patterning of secretion has become recognized; the presence of oxytocinase effectively reduces the half-life of a secreted bolus with little effect on its amplitude, and thus oxytocinase might enhance pulsatile patterning of oxytocin rather than eliminate oxytocin completely.

Although pulsatile secretion of oxytocin appears to be very important in lactation and during parturition, there are other circumstances in which oxytocin is secreted continuously. In the rat, oxytocin is secreted continuously when the plasma concentration sodium is high, and the significance of this appears to be that in the rat (but not in all mammals) oxytocin promotes sodium excretion both by a direct action at the kidney and indirectly by activating the secretion of atrial natriuretic factor from the heart. Interestingly, pulsatile oxytocin secretion is an ineffective trigger for natriuresis, apparently because this reflex requires sustained activation of the receptors. Accordingly, continuous secretion of oxytocin and pulsatile secretion of oxytocin can fulfill independent physiological roles – roles that can be combined without conflict.

Vasopressin

Vasopressin neurons, although similar in many respects to oxytocin neurons, do not generate pulsatile vasopressin release in the rat. However, they do generate bursts of activity. These 'phasic bursts' are longer, slower, and more frequent than the bursts of oxytocin cells but are not synchronized among the vasopressin cells and so give rise to continuous secretion rather than intermittent secretion of vasopressin from the posterior pituitary. The significance of the bursting appears to be that it optimizes the efficiency of vasopressin secretion from the nerve terminals of individual vasopressin neurons. Generally, in mammals, vasopressin is a cofactor regulating the secretion of ACTH (acting synergistically with corticotropin-releasing factor), but in most species this regulation is achieved by parvocellular vasopressin neurons of the paraventricular nucleus that project to the median eminence. These parvocellular neurosecretory neurons are functionally independent of the magnocellular vasopressin neurons that project to the posterior pituitary gland, and in most species the concentration of vasopressin secreted from the pituitary is too low for significant action at corticotrophs. However, in some species, notably the horse, vasopressin can be released in large pulses from the posterior pituitary, and these large pulses have an important role in regulating ACTH secretion.

Gonadotrophic Hormones

LH and FSH, coproduced in pituitary gonadotrophs, are secreted in a prominent pulsatile manner in both males and females in response to stimulation by luteinizing hormone-releasing hormone (LHRH), which is released from neurosecretory neurons that project to the median eminence. Whereas magnocellular oxytocin cells are compactly localized in the supraoptic and paraventricular nuclei, the LHRH neurons, of which there are only approximately 1000, are widely dispersed through the preopic/septal area. Nevertheless, these dispersed LHRH neurons are in direct contact with each other via contacts between their dendrites, which in the rat can extend for more than 1 mm. Just as the oxytocin cell dendrites are packed with neurosecretory vesicles, so are those of LHRH neurons, which also express LHRH receptors. Thus, it seems that, as for the oxytocin cells, the pulsatile release of LHRH is a result in part of mutual excitatory interactions between LHRH cells mediated by dendritic release of LHRH.

Again, as for the oxytocin cells, although a basic mechanism for synchronization and pulse generation resides within the LHRH network, afferent inputs to the LHRH cells determine the amplitude and frequency of the LHRH pulses. The gonadal steroids – estrogen and progesterone in the female and testosterone in the male – are powerful modulators of the patterning. The varying steroid environment of the cycling female means that the pattern of

LHRH secretion and hence LH and FSH secretion is different at different stages of the ovarian cycle. The ovarian steroids mediate negative feedback as a result of the gonadotrophic effects of LH and FSH. These negative feedback actions are exerted in the brain but apparently not on the LHRH cells; the LHRH cells express estrogen receptor-β but not estrogen receptor-α (at least not in abundance), but the inhibitory effects of estrogen appear to be mediated mainly by local GABA interneurons and by adrenergic inputs from the brain stem. These two inputs are strongly modulated by direct actions of steroids and are critically important in pulse generation (Figure 2).

The LHRH pulses are not transformed 'passively' into LH and FSH pulses by the pituitary gonadotrophs. First, the response of the pituitary gonadotrophs to LHRH is nonlinear: below a critical amplitude, fluctuations in LHRH release will thus be filtered out, whereas above that threshold the dose–response relationship is steep, so the pituitary will thus tend to 'digitize' the LHRH signal. Second, the synthesis of LH and FSH is regulated differentially by the pattern

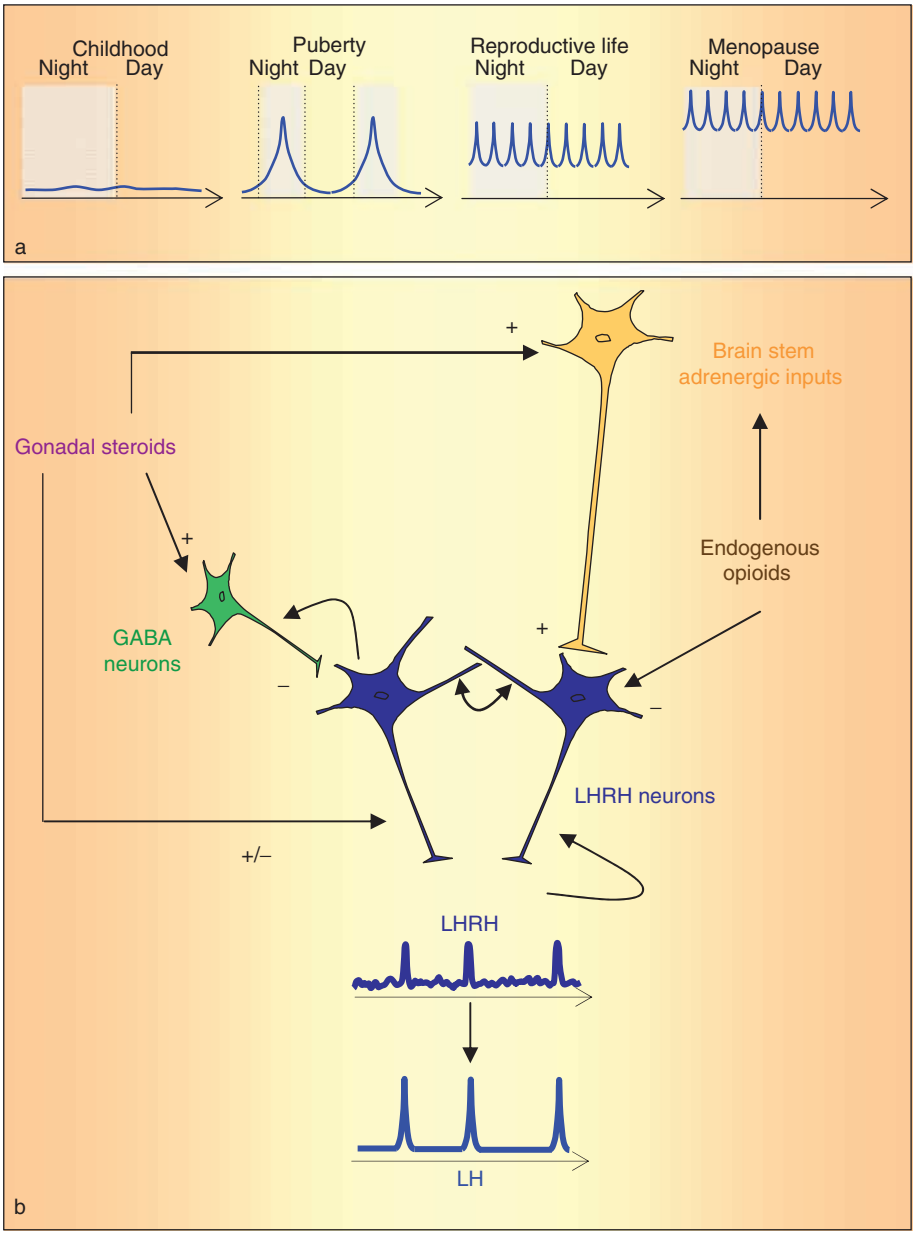

Figure 2 Pulsatile gonadotrophin secretion. (a) The pattern of LH and FSH secretion in mammals varies at different developmental stages, and in women it also varies during the ovarian cycle. (b) LHRH neurons are interconnected via their dendrites and are directly regulated by local GABA neurons, which are targets for estrogen actions. This hypothalamic mini-network is modulated by adrenergic inputs from the brain stem (among others). The resulting LHRH output is filtered and amplified by the pituitary gonadotrophs.

of stimulation by LHRH, so the relative proportions in which these two hormones are secreted will depend on the pattern of LHRH stimulation. Third, estrogen, for example, also has effects at the pituitary; it both stimulates the synthesis of LH and alters how the gonadotrophs respond to a pulsatile LHRH signal. After prolonged exposure to high concentrations of estrogen, the effects of LHRH on LH secretion display 'self-priming'; this has close parallels to the priming of oxytocin cells by oxytocin, which potentiates activity-dependent oxytocin release from the dendrites of oxytocin cells. In both cases, this priming involves a delayed and long-lasting recruitment of neurosecretory vesicles into sites from which their release by exocytosis can be readily evoked. Self-priming in estrogen-primed gonadotrophs means that successive pulses of LHRH can lead to progressively larger pulses of LH secretion, culminating in a dramatic surge release of LH at proestrous. This surge is a prerequisite for ovulation in mammals.

Pulsatile secretion of LH and FSH is essential for normal gonadal function in both males and females. The actions of LH in particular require pulsatile delivery to avoid desensitization of the peripheral targets. Thus, continuous administration of an LHRH agonist results in a functional castration, and this is exploited both for prolonged male contraception and for androgen-deprivation therapy as an adjunct to treatment of prostate cancer.

Growth Hormone

In the male rat, GH secretion from the somatotroph cells of the pituitary gland is dominated by large, 3 h pulses separated by 'troughs' of very low levels of secretion. GH is secreted in a similar pulsatile pattern in most mammals, including humans. In rats and mice, this pulsatile pattern is important for the efficiency of its growth-promoting effects; the sexual dimorphism in body size in rats is largely accounted for by a sexually dimorphic pattern of GH secretion. In the female rat, GH pulses are smaller and more frequent than in the male and the trough levels are higher, so the pulsatile pattern is less marked (Figure 3).

GH secretion is controlled by two hypothalamic peptides: growth hormone-releasing factor (GRF), produced by neurosecretory neurons of the arcuate nucleus, and somatostatin, produced by neurosecretory neurons of the periventricular nucleus. It is thought that the underlying pattern of GRF activity is similar in males and females, with the sexually differentiated patterns being produced by different patterns of somatostatin release. Estrogen influences somatostatin neurons, although probably not directly because somatostatin neurons do not express estrogen receptor-α.

Our understanding of pulsatile GH secretion has been limited by the paucity of information about the electrophysiological properties of these hypothalamic neurons and especially about their behavior *in vivo*. In part this is because pulsatile GH secretion does not persist under general anesthesia, which might allow the underlying processes to be observed more directly. However, it is possible to measure GRF and somatostatin release into portal blood and correlate this with GH secretion. These are invasive procedures with limited time resolution, but information from such studies, especially in sheep, has been important in supporting our understanding of GH secretion. However, our understanding mostly comes from measurements of GH concentrations in blood after experimental interventions that affect the release or actions of GRF, somatostatin, and other regulatory factors, including GH, which has negative feedback actions on the hypothalamus.

Much is known about the pituitary somatotrophs and how they respond to somatostatin, GHRH, and ghrelin, which plays an important role in regulating GH secretion by actions at both the hypothalamus and the pituitary. GRF stimulates GH secretion via G-protein-coupled receptors; the transduction pathway involves activation of adenylate cyclase and protein kinase A, elevated levels of cAMP, Ca^{2+} entry through voltage-gated channels, and Ca^{2+}-dependent exocytosis. Somatostatin inhibits GH secretion directly, but it is also a functional GRF antagonist; when somatostatin is present, GRF is less potent at stimulating GH secretion. Somatostatin acts (at separate G-protein-coupled receptors) to oppose the actions of GRF; it reduces the intracellular concentration of cAMP and hyperpolarizes the somatotrophs by increasing potassium conductance. In male rats, GRF and somatostatin are released cyclically and reciprocally to produce the peaks and troughs of GH secretion.

The following features of the pituitary response to GRF and somatostatin are particularly relevant to understanding pulsatile secretion of GH:

- GRF stimulates dose-dependent GH secretion. The sigmoidal nature of this dose–response relationship means that the pituitary acts as a 'band pass filter' for the hypothalamic signal. Small fluctuations of GRF around a low average concentration will be filtered out as below the threshold for activating GH secretion, whereas high concentrations will tend to produce an all-or-nothing secretory response.
- GH secretion in response to GRF desensitizes in a dose-dependent manner. This means that sustained high concentrations of GRF will tend to produce a transient GH response.

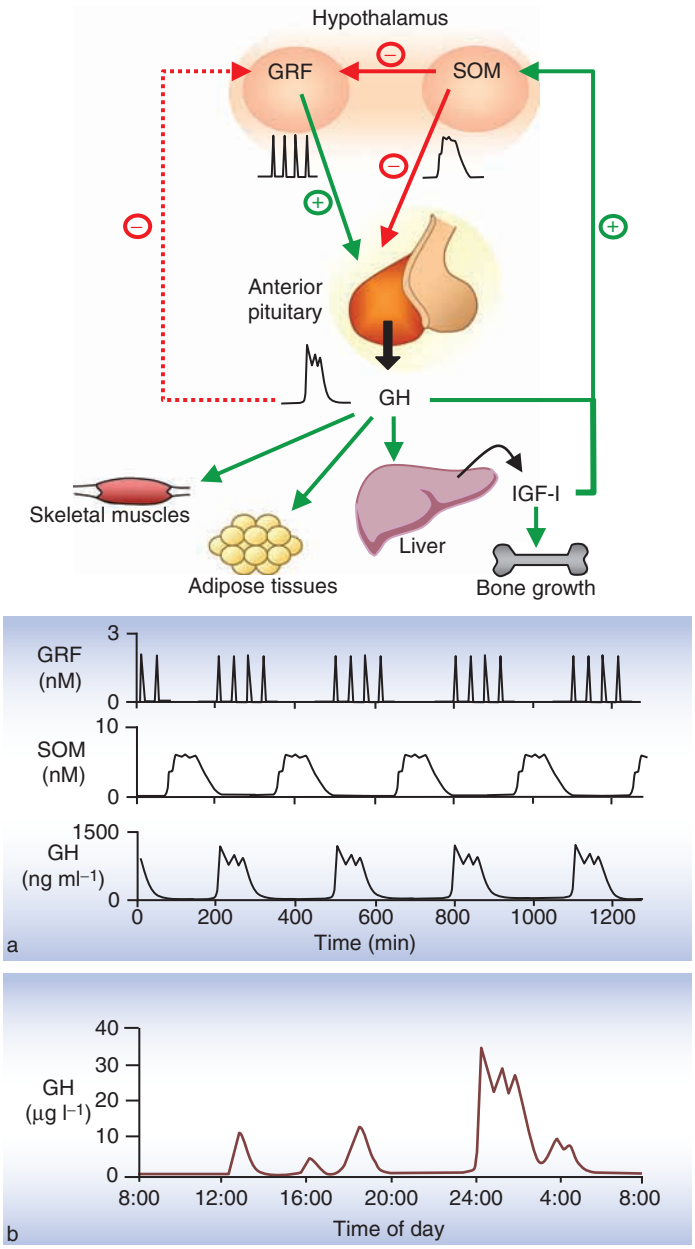

Figure 3 Pulsatile GH secretion. In humans, large pulses of GH are secreted at 3 h intervals. The magnitude of these pulses is maximal at puberty and declines slowly but progressively with age, often leading in old age to increased adiposity, bone frailty, and muscle weakening. The pulsatile pattern of GH secretion is important for its biological efficacy, and it arises as a result of a reciprocal interaction between neurosecretory somatostatin and GRF neurons, regulated by negative feedback from GH and somatomedins (particularly insulin-like growth factor 1 (IGF-I)). (a) The behavior of a computational model of the GH system. Adapted from Macgregor DJ and Leng G (2005) Modelling the hypothalamic control of growth hormone secretion. *Journal of Neuroendocrinology* 17: 788–803. (b) The typical GH profile of an adult man with 3 h pulses and a conspicuously large nocturnal pulse.

- Somatostatin inhibits GH secretion in a dose-dependent manner and can prevent desensitization to GRF; rapid removal of somatostatin causes a GRF-dependent rebound secretion of GH. This means secretion of somotostatin will tend to eliminate low levels of GH secretion and prepare the pituitary to respond to a subsequent GRF pulse.

Thus, GH secretion *in vivo* reflects the activity of hypothalamic neuroendocrine systems, acting via the pituitary, which integrates and filters these signals. These two hypothalamic systems interact; exogenous somatostatin inhibits GRF mRNA expression, and somatostatin has a fast inhibitory effect on GRF neuronal activity and release. Therefore, it seems that

GRF neurons are inhibited following increased activity in somatostatin neurons, and there might also be a complementary influence of GRF on somatostatin neurons.

In addition, GH and somatomedins feed back on the hypothalamus to inhibit the GRF neurons. Administration of GH *in vivo* inhibits GH secretion; this effect develops over approximately 1 h, and it takes another 1 or 2 h for normal secretion to recover. GH suppresses GH secretion in both male and female rats, but females continue to respond to GRF after GH injection, whereas males are unresponsive. This suggests that GH induces a prolonged, delayed release of somatostatin in the male rat but suppresses GH secretion through a different mechanism in the female. Males normally show a more variable response to GRF injections than females that is apparently related to more variable somatostatin release generally.

There is evidence that these negative feedback actions of GH are mediated by effects of GH on the somatostatin neurons; GH appears to trigger a prolonged release of somatostatin, lasting much longer than the original GH challenge. This is also a delayed effect; up to 1 h is required before GH stimulates an increase in somatostatin release. This delay seems too short to reflect changes in synthesis but too long to reflect classic electrophysiological mechanisms.

There are two other documented examples of delayed activation in peptide systems. The first is the self-priming action of LHRH at pituitary gonadotrophs, whereby LHRH potentiates its subsequent effectiveness in estrogen-primed pituitaries, and this involves a delayed mobilization of the secretory pool of LH. The second example is the priming effect of oxytocin on dendritic oxytocin release. There is no direct evidence for priming in somatostatin neurons. However, for both oxytocin neurons and gonadotrophs, priming is a delayed, long-lasting effect that is apparently mediated by mobilization of releasable pools of one peptide in response to the actions of another. A similar effect might underlie the delayed, long-lasting potentiation of somatostatin release by GH.

Synthetic GH secretagogues (GHS), like GH-releasing peptide, also promote GH secretion as GRF does but via different receptors. There are GHS receptors both on somatotrophs and in the hypothalamus, where their actions include activation of the GRF neurons. *In vivo*, GRF and GHS act synergistically; GH secretion is greater when GRF and GHS are applied together than when applied separately. The endogenous ligand for the GHS receptor, ghrelin, is secreted mainly from the stomach and is present in the circulation at high concentrations. Ghrelin is also expressed in a subpopulation of arcuate neurons. The physiological role of ghrelin in regulating GH secretion is unclear, but the potential value of GHS in promoting pulsatile GH secretion has been noted.

Conclusions

Although neuronal networks in the hypothalamus can function as pattern generators, the anterior pituitary gland also plays a major role in filtering, amplifying, and modulating the resulting pattern of hormone secretion. It is increasingly recognized that pulsatile patterns of hormone secretion are near ubiquitous, and even hormones once considered classic endocrine hormones with little neural regulation also display biologically significant pulsatile secretion. Also, underlying the generation of these patterns are mechanisms that appear to be very similar to those which govern pituitary hormone release. The biological importance of pulsatile hormone signals has many implications, including for therapeutic interventions in conditions of hormonal insufficiency. It is difficult, if not impossible, to reproduce a complex pattern of hormonal delivery by exogenous administration. In some cases, drugs that act in the central nervous system show promise for their ability to boost the function of hypothalamic pulse generator systems to 'rejuvenate' dysfunctional systems. Other approaches include developing controlled means of intermittent drug release.

See also: Gonadal Steroid Actions on Brain; Gonadotropin, Neural and Hormonal Control; Gonadotropin-Releasing Hormone: GnRH-1 System; Magnocellular Neurosecretory System: Organization, Plasticity, Model Peptidergic Neurons; Neuroendocrine Aging: Pituitary Metabolism; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Neurosecretion (Regulated Exocytosis in Neuroendocrine Cells); Oxytocin (Peripheral/Central Actions and their Regulation); Vasopressin/Oxytocin and Receptors.

Further Reading

Alexander SL, Irvine CH, and Donald RA (1996) Dynamics of the regulation of the hypothalamo-pituitary-adrenal (HPA) axis determined using a nonsurgical method for collecting pituitary venous blood from horses. *Frontiers in Neuroendocrinology* 17: 1–50.

Clarke IJ (2002) Two decades of measuring GnRH secretion. *Reproduction Supplement* 59: 1–13.

Clarke IJ and Cummins JT (1987) Pulsatility of reproductive hormones: Physiological basis and clinical implications. *Baillieres Clinical Endocrinology and Metabolism* 1(1): 1–21.

Goldbeter A and Claude D (2002) Time-patterned drug administration: Insights from a modeling approach. *Chronobiology International* 19: 157–175.

Hauffa BP (2001) Clinical implications of pulsatile hormone signals. *Growth Hormone and IGF Research* 11(supplement A): S1–S8.

Kikuchi A and Okano T (2002) Pulsatile drug release control using hydrogels. *Advanced Drug Delivery Reviews* 54: 53–77.

Leng G, Caquineau C, and Sabatier N (2005) Regulation of oxytocin secretion. *Vitamins & Hormones* 71: 27–58.

Leng G and Ludwig M (2006) Jacques Benoit Lecture. Information processing in the hypothalamus: Peptides and analogue computation. *Journal of Neuroendocrinology* 18: 379–392.

Ludwig M and Leng G (2006) Dendritic peptide release and peptide-dependent behaviours. *Nature Reviews Neuroscience* 7: 126–136.

Macgregor DJ and Leng G (2005) Modelling the hypothalamic control of growth hormone secretion. *Journal of Neuroendocrinology* 17: 788–803.

McNeilly AS, Crawford JL, Taragnat C, Nicol L, and McNeilly JR (2003) The differential secretion of FSH and LH: Regulation through genes, feedback and packaging. *Reproduction Supplement* 61: 463–476.

Porksen N (2002) The *in vivo* regulation of pulsatile insulin secretion. *Diabetologia* 45: 3–20.

Robinson IC, Gevers EF, and Bennett PA (1998) Sex differences in growth hormone secretion and action in the rat. *Growth Hormone & IGF Research* 8(supplement B): 39–47.

Russell JA, Leng G, and Douglas AJ (2003) The magnocellular oxytocin system, the fount of maternity: Adaptations in pregnancy. *Frontiers in Neuroendocrinology* 24: 27–61.

Terasawa E (2001) Luteinizing hormone-releasing hormone (LHRH) neurons: Mechanism of pulsatile LHRH release. *Vitamins & Hormones* 63: 91–129.

Veldhuis JD (1998) Neuroendocrine control of pulsatile growth hormone release in the human: Relationship with gender. *Growth Hormone & IGF Research* 8(supplement B): 49–59.

Young EA, Abelson J, and Lightman SL (2004) Cortisol pulsatility and its role in stress regulation and health. *Frontiers in Neuroendocrinology* 25: 69–76.

Pulvinar Structure and Circuitry in Primates

S Shipp, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Pulvinar Structure and Circuitry

The term pulvinar derives from the Latin word for a cushion. Being the thalamic nucleus situated adjacent to the lateral geniculate ('knee-like') nucleus (LGN), the term is quite apt, although the humor may not have been intentional. The LGN and pulvinar constitute the twin thalamic elements of the visual system, but they differ significantly in structure and connectivity. The LGN has significant connections external to its primary cortical target (area V1), but the scope of pulvinocortical circuitry is far more extensive. In fact, there is no known region of ipsilateral visual cortex that fails to interact with the pulvinar, and parts of somatosensory, auditory, cingulate, and frontal cortices are also connected to it. Also unlike the LGN, the pulvinar receives only a very sparse retinal input and lacks the layered structure associated with segregated ocular inputs and monocular response properties. Instead, it appears that a specific subset of cortical inputs to the pulvinar mimics the role of retinal afferents to the LGN – both are known as 'primary afferents.' On this basis, the pulvinar is classified as a higher-order thalamic relay (and the LGN is termed a first-order relay). It therefore follows that the prime purpose of the pulvinar is to operate a set of corticothalamocortical (CTC) loops because the pulvinar has no anatomical output other than its return projections to the cortex.

Clearly, in order to probe the functional role of such a CTC loop, it is helpful to have a clear picture of the detailed anatomical organization of pulvinocortical circuitry: Which cortical areas are interconnected through which zones of the pulvinar? Although we have a workable system of cortical parcellation, a suitable reference system for the pulvinar is not well established. Indeed, the traditional morphological subdivisions found in standard brain atlases are nothing but a hindrance. By analogy, it is like trying to devise a chart of international flight routes indexed to river basins; the task may not be impossible, but hydrology is hardly the natural frame of reference for airborne transport. The following section outlines the basis of the traditional subdivisions and a more satisfactory alternative parcellation based on chemoarchitecture. However, the simplest means to analyze patterns of connectivity within the pulvinar is to allow those patterns to set their own boundaries. This can be achieved by carefully mapping the topography of pulvinocortical connections and using reversals or other forms of topographic discontinuity to establish structural subunits. The exercise is facilitated by reference to visual maps (in ventral pulvinar) which obey similar principles.

The final element for discussion is to consider the neural elements of pulvinocortical circuitry because different subpopulations of neurons may participate in circuits with similar topography while performing different functional roles. At the cellular level, synaptic contacts may be driving or modulatory in effect. It is not clear how these properties are integrated at a systems level, but a hypothetical, hierarchical model for pulvinocortical interaction provides a useful means of organizing existing observations and a source of hypotheses for experimental investigation.

Pulvinar Architecture

Myeloarchitecture and Cytoarchitecture

The traditional subunits of the pulvinar are its medial, lateral, and inferior nuclei, defined on the basis of cell and fiber stains in the 1930s. A fourth nucleus, the oral pulvinar, lying anterior to the other three, was added in 1952. With such methods, some boundaries are more cleanly observed than others, and the pulvinar is no exception. There is little to distinguish the lateral pulvinar nucleus from the neighboring lateral posterior nucleus, for instance. The inferior pulvinar is the most morphologically distinct subunit since it is isolated from the main body of the pulvinar by a substantial fiber tract known as the brachium of the superior colliculus. Many of these fibers are collaterals branching from the optic tract, bound for the superior colliculus, and they course through the gray matter of the pulvinar without making any functional contribution to it. The lateral pulvinar is distinguished from the medial pulvinar by the presence of numerous fiber bundles that perforate its lateral surface, adopting a roughly parallel arrangement within a coronal plane of section. The medial pulvinar nucleus is held to commence at the point where these bundles eventually end, approximately halfway to the medial limit of the pulvinar, but the transition between the two zones is not especially abrupt. Having staked out these zones, there are some corroboratory differences in cellular architecture to be observed, but given the nature of the morphological criteria, it should not come as a surprise if the traditional nuclei bear an arbitrary relationship to the functional substructure of the pulvinar.

Chemoarchitecture

The alternative to traditional cell and fiber stains is a variety of histological treatments that can be collectivized under the term chemoarchitecture. These reveal the distribution of various enzymes or other protein molecules that regulate cell physiology, or transmitter neurochemistry. For instance, fluctuations in the density of stains for the mitochondrial enzyme cytochrome oxidase, and the neurotransmitter degradation enzyme acetylcholinesterase, are often highly correlated. So too is the pattern revealed by an antibody to the calcium-binding protein calbindin, except that this is the complementary pattern – light where the acetylcholinesterase stain is dark and vice versa. Figure 1 contrasts the traditional substructure of the pulvinar with a more recent scheme based on

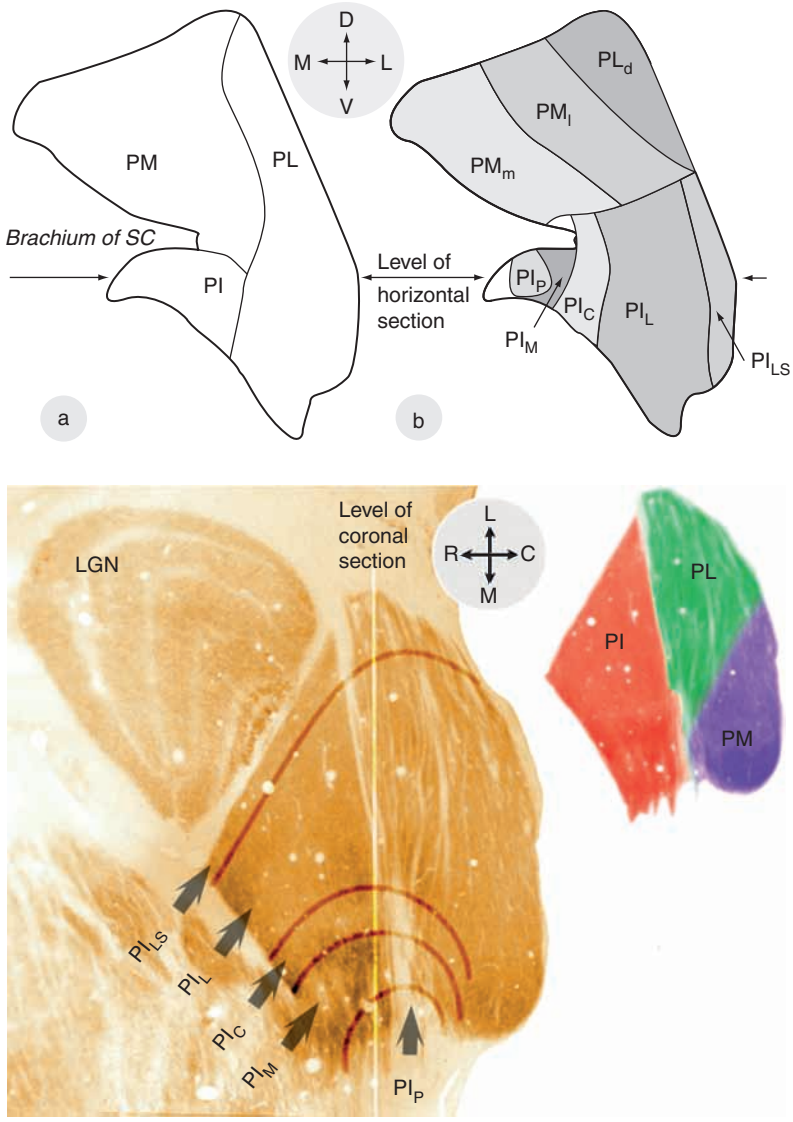

Figure 1 Comparison of traditional morphological and neurochemical subdivisions of the pulvinar, depicted in a coronal plane of section (top) and in a photograph of a horizontal section stained for cytochrome oxidase (bottom). Architectural subunits (a) are the inferior, lateral, and medial nuclei (PI, PL, and PM, respectively); as coded red, green, and violet in the (miniature) horizontal section, the inferior nucleus is sharply demarcated from the lateral and medial nuclei, whose subdivision is structurally more arbitrary. Neurochemical subunits (b) are the concentrically arranged posterior, medial, central, lateral, and lateral shell divisions of an enlarged inferior pulvinar (PI_P, PI_M, PI_C, PI_L, and PI_{LS}, respectively), and three subdivisions of the dorsal pulvinar that, aiming to retain continuity with the traditional terminology, were named the dorsal division of lateral pulvinar (PL_d) and lateral and medial subdivisions of medial pulvinar (PM_l and PM_m, respectively). The shading in (b) represents relative density of staining for acetylcholinesterase. Borders of the neurochemical subunits of PI indicated on the horizontal cytochrome oxidase section are as inferred from diagrams such as (b), with limited correspondence to the intrinsic staining pattern. Each plane of section (coronal/horizontal) carries an indication of the approximate level of the other; mediolateral, dorsoventral, and rostrocaudal anatomic axes are indicated by the two dials.

chemoarchitecture. The newer scheme retains elements of the older terminology, although none of its subregions are exact equivalents (Figure 1). For instance, the chemoarchitectonic inferior pulvinar (PI) is enlarged to include the ventral half of the traditional lateral pulvinar. It is subdivided into several segments – posterior, medial, central, lateral, and lateral shell (PI_P, PI_M, PI_C, PI_L, and PI_{LS}, respectively; Figure 1) – based on a pattern of alternating densities in calbindin and acetylcholinesterase stains (Figure 2). The dorsal border of PI is placed where this alternation in density subsides, just above the level of the brachium of the superior colliculus. The remainder of the pulvinar, termed the dorsal pulvinar, is also subdivided according to gradations in staining density, with the terminology PL_d, PM_l, and PM_m reflecting approximate correspondence to the traditional lateral and medial nuclei (Figure 1).

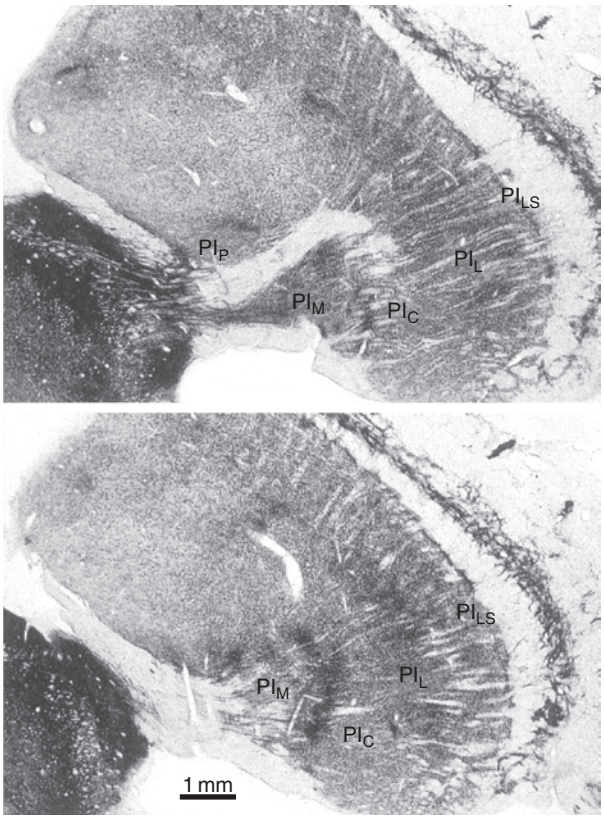

Figure 2 A pair of coronal sections through macaque thalamus, stained for acetylcholinesterase, showing subdivisions of inferior pulvinar. The lower section is taken from a level just caudal to the posterior inferior pulvinar (PI_P). PI_M, medial inferior pulvinar; PI_C, central inferior pulvinar; PI_L, lateral inferior pulvinar; PI_{LS}, lateral shell of an inferior pulvinar. Reproduced from Cola MG, Gray DG, Seltzer B, and Cusick CG (1999) Human thalamus: Neurochemical mapping of inferior pulvinar complex. *Neuroreport* 10(18): 3733–3738, with permission.

Having a basis in cellular physiology, it is more realistic that chemoarchitectural zones should reflect functional differentiation within the pulvinar, for example, by demarcating some specific input/output pathways. The functional correspondence of chemoarchitecture is not well understood theoretically, but some empirical correlation with afferent pathways has emerged. For instance, the subunit PI_M (also known as the calbindin 'hole' due to its pale appearance in calbindin-stained tissue) is the main site of retinal terminations. By contrast, an indirect source of retinal input, relayed by the superficial layers of the superior colliculus, avoids PI_M but terminates most densely in the neighboring subunits PI_P and PI_C. PI_M is also the prime focus for reciprocal connections with cortical area V5/MT. Differential relationships with a number of other cortical areas are known to exist in a variety of primate species, but it is also clear that the (known) chemoarchitectural plan is not the sole agent configuring the pattern of corticothalamic connectivity. It is therefore useful to examine a complementary system of codifying the pattern of cortical connections.

Topology and Topography of Cortical Connectivity

Global Organization of Connectivity

As illustrated in many textbooks, there is a topographic relationship between the thalamus as a whole and the cortex of the same hemisphere. To envision this, imagine an image of a flattened cortical sheet being mirror reversed, rotated 90°, and shrink-wrapped around the thalamic ovoid. Figure 3 illustrates, in principle, that the same overall relationship holds for the topography of cortical connectivity within the pulvinar. In the ventral pulvinar, an occipitotemporal cortical gradient is expressed along a rostrolateral–caudomedial axis. Along roughly the same axis, in dorsal pulvinar, there is a parietotemporal cortical gradient. This is consistent, in schematic terms, with a mirror image of the cortical sheet being shrink-wrapped around the posterior surface of the pulvinar (i.e., around the posterior pole of the thalamus); in general, the visual cortex (or the ventral visual pathway) is mapped into the ventral pulvinar and the dorsal visual pathway into the dorsal pulvinar. Neighboring regions of visuosomatic and visuoauditory association cortex are mapped into rostral sectors of the pulvinar, including the anterior pulvinar nucleus and part of the medial pulvinar nucleus.

The 'shrink-map' model is simply a first approximation because there are many topographic irregularities

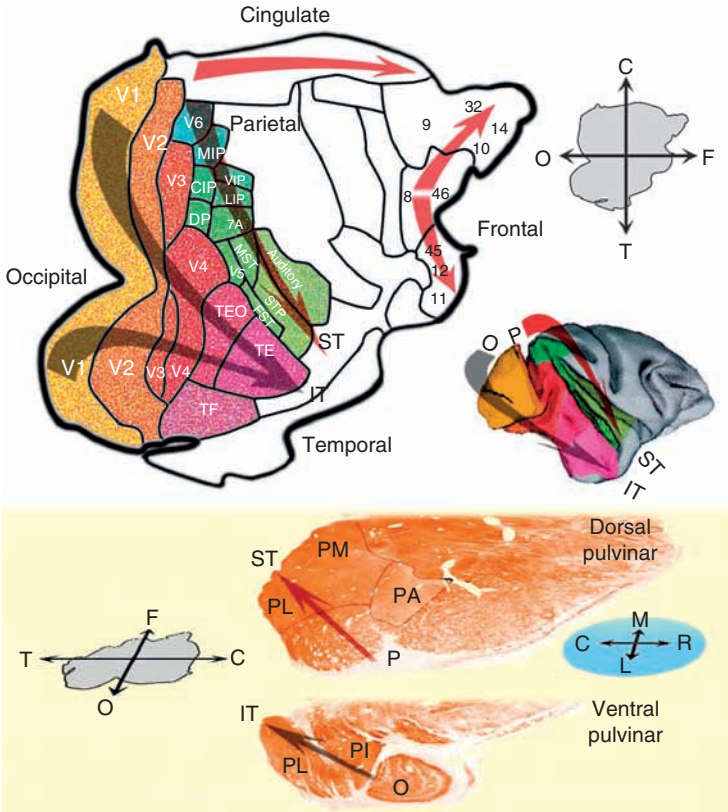

Figure 3 Topographic relationships between the cortical sheet and the pulvinar. Two principal vectors, parietal–superotemporal (P-ST; red) and occipito-inferotemporal (O-IT; gray) are shown on a whole brain view and on the flattened cortical sheet taken from the same right hemisphere (top left). The constituent individual areas are represented by blue-green and orange-violet color gradients, respectively. In the bottom panel, the same two vectors are marked on a perspectively viewed pair of horizontal sections through the dorsal and ventral pulvinar of the right hemisphere, in which they follow roughly parallel rostrolateral to caudomedial trajectories. The additional red vectors in the cortical sheet represent topographic gradients in cingulate and frontal cortex that are expressed in parallel to P-ST in dorsal pulvinar. The gray arrow is shown with dual components in the cortical sheet to reflect the discontinuity (or tear) which is made in the calarine region of V1 to facilitate flattening without distortion. The form of the cortical sheet is reproduced as a 'compass' (top right) showing cortical axes O-F (occipitofrontal) and C-T (cingulotemporal). The compass in perspective view (bottom left) has been mirror reflected and rotated by 90°. It is now topographically consistent with the P-ST and O-IT vectors marked on the pulvinar sections. Mediolateral and rostrocaudal anatomic axes through the pulvinar are shown on the blue dial.

that it fails to recognize, including superimpositions, duplications, and topological reversals. Furthermore, topography can only be used to rationalize two dimensions of organization within the three-dimensional volume of the pulvinar; there are numerous instances of local inhomogeneities and mosaic or patchy connectivities that could be considered part of a 'third dimension' of pulvinar organization.

Ventral Pulvinar

Ventral pulvinar is territorially equivalent to the expanded, histochemically defined PI and can be subdivided, on topographic criteria, into three subzones termed VP1–VP3 (these terms are adapted from P1 to P3, which were originally used to describe the topography of connections with cortical area V5/MT alone). VP1 are VP2 are coregistered with a pair of visual retinotopic maps, whereas VP3 is much less retinotopic.

Visual topography in VP1 and VP2 The organization of the visual maps of the ventral pulvinar is analogous to the visual topography expressed by cortical areas V1 and V2 in several respects: (1) the 2° map is a mirror image of the 1° map and surrounds it concentrically; (2) the border is formed by a common representation of the vertical meridian; (3) the 2° map has a split representation of the horizontal meridian from 3° eccentricity onward, such that the inferior field quadrant is found dorsally and the superior quadrant ventrally (Figure 4). The 1° and 2° pulvinar maps were originally named the inferior map and the lateral map because they respectively occupy much of the traditional inferior and lateral subnuclei. However, the visual maps do not exactly respect the architectural divisions since part of the 1° map invades lateral pulvinar, and the 2° map only occupies the remaining sector of the ventral half of the lateral pulvinar.

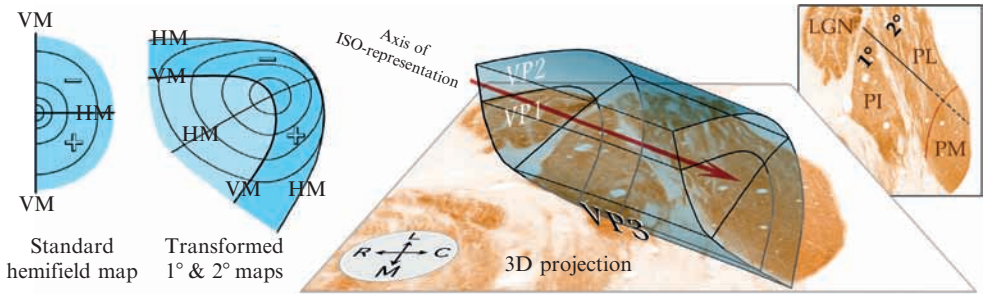

Figure 4 Schematic illustration of the 1° and 2° visual maps in VP1 and VP2 subdivisions of ventral pulvinar. The three-dimensional mapping volume (translucent blue) is superimposed on a perspective view of a horizontal slice through the right pulvinar (at approximately the same level illustrated in Figure 1). The dial in the slice indicates rostrocaudal and mediolateral anatomic axes. (Inset) The same slice in normal view showing how the border between the 1° and 2° maps intersects the anatomical border between the traditional inferior and lateral pulvinar (PI and PL, respectively). (Far left) A left visual hemifield map, as imaged on the retina (i.e., with inversion of superior and inferior quadrants). (Middle left) Transformed 1° and 2° maps (labeled VP1 and VP2), as found in any slice through the mapping volume made in a plane perpendicular to the axis of isorepresentation.

Again, this confirms the functional irrelevance of the traditional subnuclei.

To form a two-dimensional visual map in a three-dimensional body, there has to be one dimension along which the same point in space is continually re-represented. This dimension is well known to be perpendicular to the intrinsic lamination in structures such as visual cortex and the LGN, and it is known as the axis of isorepresentation (AIR). The alignment of the AIR in the pulvinar is approximately rostrolateral–caudomedial (Figure 4). Importantly, the axes of cortical topography and of visual isorepresentation are coincident. The alignment of the AIR can be directly visualized in horizontal sections through ventral pulvinar showing the connections made by focal injections of tracer into cortical areas with good visual topography (e.g., V1–V4), as illustrated in Figure 5. The characteristic product is a column of labeled neurons and/or axon terminals within ventral pulvinar, tracing out the line of the AIR. If the tracer is deposited sufficiently far from the vertical meridian in the visual map of the target cortical area, there are two such columns. One, in VP1, is oriented roughly perpendicular to the planar interface of the inferior pulvinar with the adjacent part of the LGN; the other is within VP2, and it adopts a roughly parallel trajectory. Such anatomical experiments therefore demonstrate that cortical areas link to the pulvinar maps in a point-to-point fashion, and thus that the cortical topographies of VP1 and VP2 have the same mirror relationship to each other as the underlying visual maps. There is no known direct reciprocal connection between VP1 and VP2.

In essence, the connection zones of successive cortical areas along the occipitotemporal gradient within ventral pulvinar constitute a set of parallel slabs, all containing co-registered visual maps. These slabs, or connection zones, are heavily overlapping such that the 1° and 2° pulvinar maps can be considered as global composite maps that emerge from the superimposed cortical (and collicular) inputs. The degree of retinotopic order diminishes along the AIR as the source of cortical input is derived from areas toward the temporal pole of the occipitotemporal gradient, such as TEO and TE, which are known from cortical recording studies to possess large receptive fields and little inherent visual topography. Consequently, the distinction between VP1 and VP2 is more provisional approaching the caudomedial pole of the AIR. However, given the connectional continuity, there is little reason to regard the caudomedial pole of ventral pulvinar as a functionally discrete region.

VP3 The third subunit of ventral pulvinar, VP3, adjoins VP1 on its rostromedial boundary. In contrast to the topological distinction between VP1 and VP2 (i.e., the mirror relationship of visual and cortical topographies), the boundary with VP3 may be recognized as a topographical discontinuity. VP3 is also roughly (perhaps precisely) coincident with the PI_P, PI_M, and PI_C zones of the histochemical scheme. A major source of input to VP3 is the cortical motion area, V5/MT. Connections with this cortical area are focused on PI_M but also invade PI_P and PI_C. The adjacent part of VP1 is connected to area V4, but the connection zones of these two areas (V4 and V5) display little, if any, direct overlap, despite the fact that they are near neighbors in visual cortex; this observation contrasts readily with the high degree of overlap that exists between the connection zones of V4 and other cortical neighbors, such as areas V3 and TEO. Hence, the 'fault line' between the V5 and V4 zones is a topographic discontinuity that defines the boundary of VP1 with VP3. Finally, in both physiological and anatomical studies, VP3 has little visuotopic organization (although there are some indications that the central field is mapped

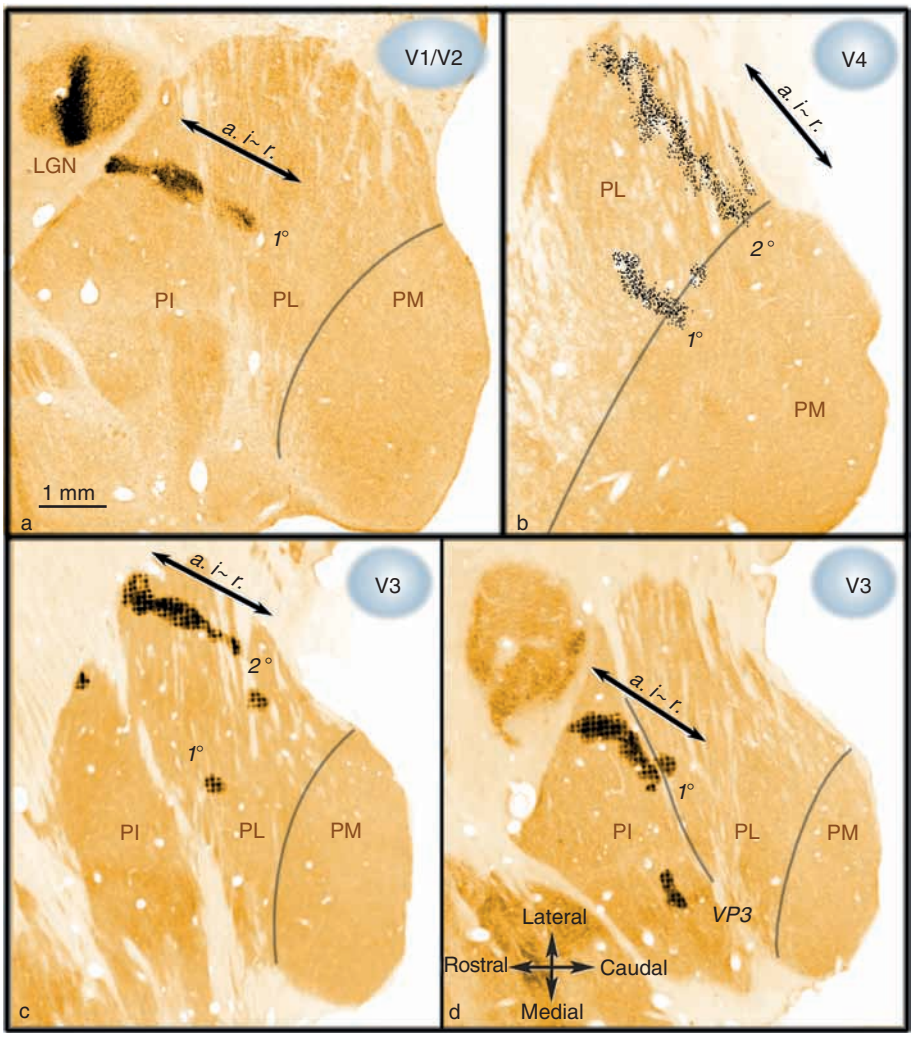

Figure 5 Anatomical demonstration of the axis of isorepresentation (a. i.-r.) as observed in horizontal sections through macaque pulvinar, demonstrating patterns of connection from occipital visual areas. (a) Connections from a site straddling the V1/V2 border. In visual terms, this was a locus on the inferior vertical meridian, predicting a single column of connections at the representation of this meridian in the pulvinar, coincident with the border between VP1 and VP2. A bidirectional neural tracer was used, demonstrating retrogradely filled cell bodies and orthogradely filled axon terminals in the pulvinar. (b) Orthograde projections from area V4, at a locus in the inferior contralateral visual quadrant. Note the presence of two parallel columns of axon terminals at mirror locations in the respective 1° and 2° visual maps of VP1 and VP2. (c, d) Connections with area V3 from a similar visual locus to that in V4. The two columns of label occur at different levels, with the main part of the 1° column being located more ventral (d). Note that section d also shows connections between area V3 and pulvinar zone VP3. All images show sections stained for cytochrome oxidase with neural connections from adjacent sections superimposed or artificially rendered with graphics software tools.

caudally in VP3). Overall, VP3 may be recognized as a zone distinct from VP1 by a combination of several histochemical and topographic criteria.

Further correspondence (or not) of histochemical and topographic subunits of ventral pulvinar The subdivision between VP1 and VP2 has no known histochemical correlate. This is not necessarily a surprise, since the cortical and subcortical territories connected to VP1 and VP2 are, in general, highly similar. The most prominent difference concerns input from cortical area V1, which distributes throughout VP1 but only within the central 7° of VP2. Hence, VP1 and VP2 together correspond to the outer two histochemical zones PI_L and PI_{LS}. The latter of these (the outermost or 'shell' region) occupies the lateral margin of the ventral pulvinar and has a number of unique characteristics. In calbindin-stained tissue, PI_{LS} is distinguished from PI_L by mildly darker neuropil and sporadic intensely calbindin-positive neurons with extensive dendritic fields. Golgi studies have also described a unique type of relay cell in this region, with an elongated bipolar dendritic field, aligned in parallel with the lateral margin. Two specific

projections are also known to concentrate at the lateral margin of the pulvinar, as if coextensive with the shell region. One of these derives from the superficial layers of the superior colliculus, and the other one derives from V5/MT.

The second V5 zone is similar to the one in VP3 in that both are reciprocally connected to V5. However, the shell zone is much more topographic and, because it receives collicular input (unlike PI_M within VP3), can presumably relay retinotopic collicular signals to V5. Pulvinocortical axons from this region terminate in V5 with remarkably small arbors (0.1–0.2 mm diameter) – approximately half the size of arbors with similar origin observed in V2 and V3 and an order of magnitude smaller than observed within V4.

The shell V5 zone is also a topographic anomaly in that it is inserted at the very rostrolateral pole of the occipitotemporal cortical gradient within VP1 and VP2. In other words, the shell V5 connection zone overlaps inputs from V1 (and V2 and V3) but is of narrower width. A related observation is that it abuts with virtually zero overlap against the V4 connection zone, which fails to extend as far as the lateral margin of ventral pulvinar. It is noteworthy, therefore, that the V4 connection zone within VP1 and VP2 appears to be segregated from the V5 connection zones investing its medial and rostrolateral boundaries, both of which have candidate but dissimilar histochemical correlates. Figure 6 shows a schematic illustration of the geometry of these various relationships; the precise three-dimensional configuration has yet to be fully established.

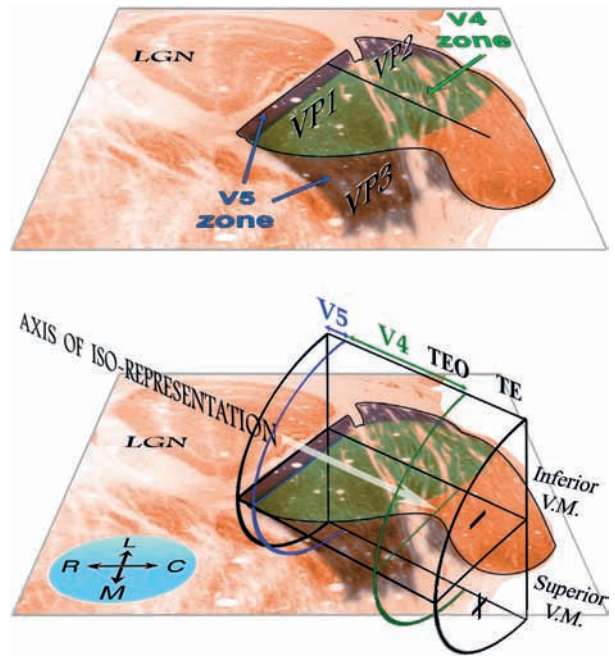

Figure 6 A model of the relationship between cortical connection topography, visual topography, and histochemical zones in ventral pulvinar. (Top) The two V5 connection zones and the single V4 zone are tinted blue and green, respectively. The subdivision into VP1 and VP2 is dictated by visual mapping (as described in the text and the legend to Figure 4). The border between VP1 and VP3 is demarcated by the transition between the two connection zones. The blue zone within VP1 and VP2 is approximately equivalent to histochemical zone PI_{LS} and the green zone (plus the adjacent untinted region) to PI_L. VP3 corresponds to the sum of PI_C, PI_M, and PI_P. (Bottom) A three-dimensional visual map in VP1 is superimposed on the same slice. The half-cylinder shape is purely schematic to illustrate the principle that the offset between cortical zones is aligned with the visual axis of isorepresentation, and hence that adjacent cortical connection zones can have registered visual maps. Although the V4 zone is separate from that of V5, it overlaps caudomedially with the connection zones of subsequent areas in the occipitotemporal gradient, namely areas TEO and TE. Similarly, the V4 zone overlaps rostrolaterally with the zones of V1–V3 (not shown). The true shape of the mapping volume is closer to that illustrated in Figure 4; clearly, it must be confined within the anatomical body of the ventral pulvinar and in some way distorted by the exclusion of VP3.

Dorsal Pulvinar

There are a number of topographic features by which the dorsal pulvinar is distinguished from the ventral pulvinar. Most obviously, it connects with a separate set of areas, expressing a cortical topography from superior parietal to superior temporal along a rostrolateral–caudomedial axis. Experiments which have simultaneously traced connections from parietal and inferotemporal locations in the same hemisphere have accordingly found a dorsoventral separation of the respective pulvinar connection zones. What overlap, or proximity, does exist is due to the presence of small foci of inferotemporal connections within the territory of dorsal pulvinar. Many of the cortical areas linked to dorsal pulvinar are visual but poorly retinotopic. Consequently, there is little evident influence of visual topography in dorsal pulvinar, and nothing comparable to the topological duplication and mirror inversion shown by VP1 and VP2. Another major difference is the presence of superimposed cortical gradients in dorsal pulvinar. As indicated in Figure 3, the rostrolateral–caudomedial axis also expresses a caudorostral sequence of cingulate cortex connections and perhaps two separate gradients within frontal cortex. There are also connections with the amygdala. The strength of these additional connections is not as great as the primary, parietotemporal gradient; however, their presence implies that a small locale in dorsal pulvinar can connect to widely separate parts of the cortex (i.e., different lobes), unlike anything known for ventral pulvinar.

Although the identification of an overall gradient helps to present a global picture of pulvinocortical connections, it is hardly a full description. Anatomical experiments reveal that the connection zones of

individual areas are broadly overlapping and frequently broken up into separate patches, columns, or disclike aggregates of labeled neurons. One detailed study has also demonstrated that the topographic relationship is not perfectly reciprocal in that a proportion of relay cells whose axons terminate at a particular point within area 7a (i.e., the site of injection of a neural tracer) appear to lie outside the terminal field of corticopulvinar axons emanating from that site. There is also an uncertain relationship vis-á-vis topography and the histochemical substructure of dorsal pulvinar. The neurochemical axis (i.e., a line perpendicular to the three identified subunits shown in Figure 1) is similar but not necessarily identical to the primary topographic axis.

The Replication Principle

What factors influence the organization of CTC loops and their topographic precision? It is worth noting that if the topography of thalamocortical circuitry were perfectly precise, there would be little or no scope for CTC loops to achieve transcortical communication (there would be a set of pinpoint reciprocal connections, with the thalamus acting as a kind of seventh cortical layer). Clearly, an imprecise topography is the signature of the fact that cortical areas do communicate via the thalamus. The 'replication principle' refers to the generalization that indirect CTC circuits tend to replicate direct corticocortical circuits. In other words, if two cortical areas are interconnected, they are likely to have overlapping, or at least partially overlapping, thalamic connection zones, whereas a pair of unconnected cortical areas will maintain segregated thalamic connection zones. Depending on the areas in question, the segregated zones might be totally separate, or interdigitating.

A cortical area forms most of its connections with nearby areas. Hence, a simple blurring of cortical topography in the thalamus will serve to replicate this network of connections. The connections of the ventral visual pathway with VP1 are a suitable example: neighboring areas, such as V4 and TEO, are heavily interconnected by direct and indirect (i.e., CTC) circuits, whereas the areas at opposite poles of the pathway, V1 and TE, do not connect directly and do not overlap in their connection zones. Longer-range cortical networks are replicated by superimposed cortical gradients; for example, the convergent cortical gradients in the dorsal pulvinar (as described previously) replicate the triangular network formed by inferior parietal, frontal, and posterior cingulate cortices. Also notable are reverse examples of the replication principle – that is, instances in which the mutual exclusion of connection zones appears to sculpt the local pattern of connections. An example described previously concerns areas V5 and V4, whose segregated connection zones in ventral pulvinar are matched by a low level of transcortical interconnectivity, in comparison to other areas in their neighborhood. A similar example in dorsal pulvinar concerns areas 7A and LIP, whose thalamic connections are mediated by separated, sandwiched clusters of cells.

Perhaps the largest-scale organization of the visual system is its bifurcation into dorsal and ventral subsystems. The two are certainly not separate but are known to operate semiautonomously (e.g., maintaining parallel representations of object size and shape). Does the replication principle extend to this level of organization? As already mentioned, the ventral visual pathway is represented intact by VP1 and VP2, but the dorsal sequence – for example, V1–V2–V3–V5 leading to parietal and superior temporal cortex – is not so easily recognized. It could exist, however, in a gradient leading from VP3 into dorsal pulvinar. There appears to be topographic continuity between VP3 and dorsal pulvinar, mediated by the connections of areas MST and FST. The connection zones of this pair of areas (both cortical neighbors of V5) overlap the V5 zone in VP3 but also extend more dorsally into dorsal pulvinar. Crucially, VP3 also receives input from V1 and V3. Although the relationships of the V1, V3, and V5 connection zones in VP3 have yet to be established, existing knowledge of pulvinar topography is reasonably consistent with a faithful 'replication' of dorsal and ventral visual subsystems.

Elements of Pulvinar Circuitry

The pulvinar, like other thalamic nuclei, consists of relay cells with cortical projections and also GABAergic (i.e., inhibitory) interneurons. There are two main classes of relay cells: core relay cells, whose axons terminate in the middle cortical layers (layers 3 and 4), and matrix relay cells, whose axons terminate superficially. The two types are distinguished histochemically by being positive for calcium-binding proteins parvalbumin and calbindin, respectively. Matrix relay cells are slightly smaller and are so named because they are relatively evenly distributed throughout the thalamus. Core cells are typically more aggregated. In the LGN, for instance, core relay cells are represented by the well-formed magno- and parvocellular layers, and matrix cells are represented by the koniocellular cells that are distributed within and between these layers. Golgi studies of pulvinar (and of thalamus in general) have identified several subtypes of interneurons with different dendritic characteristics. Interneuron cell bodies are

smaller than relay cells, but the dimensions of the dendritic field can be just as large (150–600 μm). Corticothalamic afferent axons originate from either layer 5 or layer 6 and have several distinguishing characteristics. As visualized in Golgi-stained tissue, or by using an orthograde tracer such as the *Phaseolus vulgaris* lectin, R-type afferents have large, beaded terminals within a rounded, discrete arbor; E-type afferents are finer in caliber, with spinous terminations over a more elongated zone. It is generally accepted (although yet to be directly demonstrated) that E afferents derive from layer 6 corticothalamic neurons, and R afferents derive from sparser, larger layer 6 corticothalamic neurons. R axons have collaterals directed toward typical targets of layer 5, such as the superior colliculus. Both R and E axons may give off collateral branches as they pass through the inhibitory thalamic reticular nucleus ensheathing the thalamus.

The synaptic terminals of R and E axons can also be distinguished at the ultrastructural level. RL and RS terminals, formed by R and E axons, respectively, are both asymmetric, and both contain round synaptic vesicles (the well-established ultrastructural characteristic of excitatory terminals). However, RL terminals are larger, and they contact larger caliber relay cell dendrites than the RS terminals. RL terminals are also observed to form synaptic complexes including dual pre- and postsynaptic elements of interneuron dendrites. A summary diagram of synaptic connectivity is shown in Figure 7.

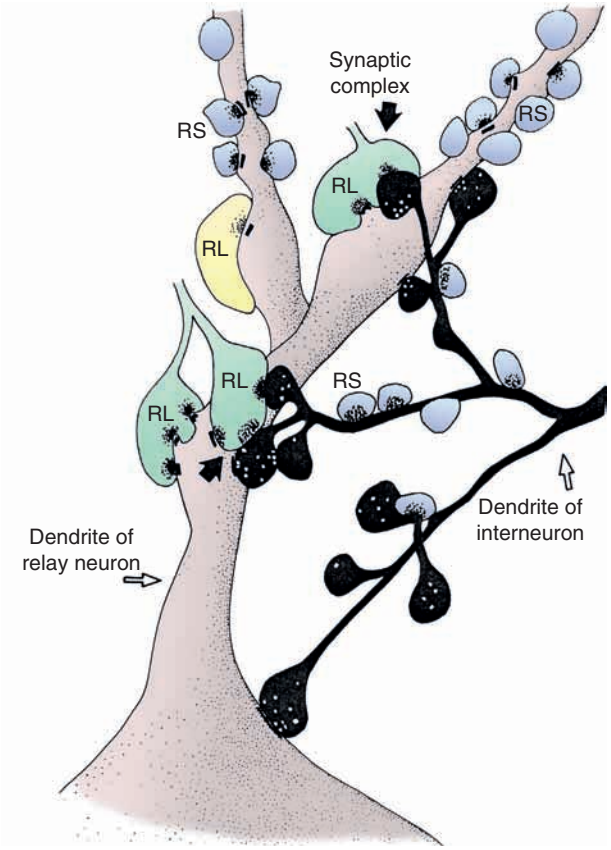

Figure 7 Synaptic circuitry in ventral pulvinar, as inferred from ultrastructural examination. RS and RL are terminals of corticothalamic axons (blue and green). Both contain round vesicles and make asymmetric contacts, which are hallmarks of excitatory transmission. RL terminals contact larger caliber dendrites (i.e., proximal dendrites), frequently taking part in synaptic complexes. Some RLs (yellow) were thought to be of collicular origin, but this was not directly ascertained. RS terminals are more numerous, contacting finer dendrites, but rarely taking part in synaptic complexes. RS terminals also synapse upon dendritic shafts of inhibitory interneurons. The terminal appendages of interneuron dendrites form synaptic complexes in which they are postsynaptic to an RL terminal and presynaptic to a relay neuron dendrite. The latter synapses are symmetric in form and contain pleomorphic vesicles (hallmarks of inhibitory transmission). Interneurons also occasionally contact other interneurons (not shown). Redrawn from Ogren MP and Hendrickson AE (1979) The morphology and distribution of striate cortex terminals in the inferior and lateral subdivisions of the Macaca monkey pulvinar. *Journal of Comparative Neurology* 188: 179–200, with permission.

Are Corticothalamic Circuits Organized According to Hierarchical Principles?

The morphological characteristics of R axons and RL terminals are similar to those of retinal afferents within the LGN. Despite forming less than 10% of all afferents to the LGN, retinal afferents drive the response of LGN relay cells and confer the receptive field properties. Other afferents (e.g., from layer 6 of V1) modulate the responsiveness of LGN relay cells, acting through metabotropic as well as ionotropic synaptic receptors. The physiology of pulvinar synaptology is much less studied, but by analogy it is reasonable to assume that layer 5 and layer 6 afferents adopt equivalent driving and modulatory roles. It is because the excitatory drive to pulvinar derives from cortex, rather than from the sensory periphery, that the pulvinar is classified as a higher order (or associative) nucleus. Driving and modulatory roles in corticocortical interactions are associated with ascending and descending pathways within a hierarchical scheme of area relationships. However, ascending and descending connections are more rigorously defined anatomically with respect to their differing layers of origin and termination. It is therefore possible to outline a provisional scheme for hierarchical relationships between visual cortex and pulvinar involving eight identified anatomical characteristics (Table 1).

The core and matrix relay cells are the putative drivers and modulators of the pulvinocortical pathways by virtue of the layers of their cortical terminations. Rather like cortical feedback (descending) connections, the matrix relay cells terminate diffusely within layer 1, in contrast to the more focal

Table 1 Anatomical elements of an integrated corticothalamic hierarchy

		Corticothalamic	*Thalamocortical*
Forward or driving	Source	Layer 5	Core relay neurons; parvalbumin positive
	Termination	R-type axons; RL terminals	Layers 3 and 4
Feedback or modulatory	Source	Layer 6	Matrix relay neurons; calbindin positive
	Termination	E-type axons; RS terminals	Layers 1 and 2

terminations of the core relay cells within the middle cortical layers that resemble ascending cortical projections.

The guiding principle of a notional, integrated corticothalamic hierarchy is that any reciprocal connection between specific cortical and pulvinar locations should be driving in one direction and modulatory in the other. In other words, a direct reciprocal loop should not be formed by driving connections in both directions (known as the 'no strong loops' hypothesis). Figure 8 depicts a hypothetical scheme in which the ventral pulvinar is subdivided into a hierarchical sequence of discrete zones, modeled on its topographic relationship with the hierarchical ventral visual pathway. The model makes certain predictions for the topography of the elements listed in Table 1 that can be tested against observational data.

Notably, the anatomical relationship with area V1 is entirely consistent with such a scheme since the corticopulvinar output from V1 is formed largely by layer 5 efferents, and the return projection is to layer 1 (presumed to arise from matrix relay cells). Hence, V1, at the base of the hierarchy, can be interpreted to provide a driving output to pulvinar, reciprocated by modulatory feedback. In the model, this is stage 'a' in the pulvinar, which in turn provides driving outputs to extrastriate cortex. The model envisions that a second pulvinar stage ('b') is driven by low-order extrastriate cortex (e.g., V2) and provides a driving output to higher-order extrastriate cortex (e.g., V4).

If organized in this fashion, anatomical tract-tracing experiments should have predictable outcomes. For instance, if tracer were placed in zone b of the pulvinar, there should be a signature pattern of cortical labeling, with back-labeled cells present in layer 5 of V2, and layer 6 of V4; orthograde tracer should reveal projections to layer 1 of V2 and layers 3 and 4 of V4. Such a result, with clear offsets in the zones of labeling of different layers, has never been reported. One good reason is undoubtedly that the deposits of tracer in the pulvinar are frequently large, potentially covering two or more zones (a, b, c, etc.) in the pulvinar. In this case, regions of driving and modulatory connection with the cortex would be expected to overlap, perhaps extensively. However, the hierarchical model would still predict particular

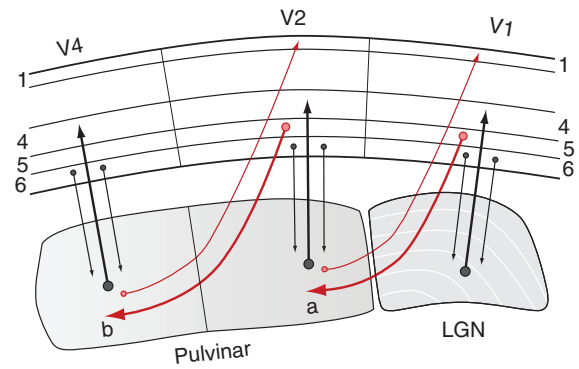

Figure 8 A hypothetical scheme for an integrated corticothalamic hierarchy. There are two kinds of corticothalamic loops. (Black) Core relay cells (parvalbumin positive) provide driving output to cortical layer 4, with modulatory feedback from layer 6. (Red) Layer 5 cortical neurons provide driving outputs to the pulvinar, reciprocated by modulatory feedback from matrix relay cells (calbindin positive) to superficial cortical layers. The physiological basis for this scheme relies on an analogy with the LGN (in which, for instance, the primary retinal afferents are isomorphic to the layer 5 afferents to pulvinar). The division of the pulvinar into discrete zones ('a,' 'b,' etc.) coordinated with hierarchical levels in the cortex (V2, V4, etc.) is purely conjectural.

laminar patterns at the fringes of the cortical region demonstrating tracer-labeled cell bodies and axon terminals – specifically, a fringe area toward the lowest tier in the cortical hierarchy with cells in layer 5 and terminals in layer 1, alone, and a fringe area toward the highest tier in the cortical hierarchy with cells in layer 6 and terminals in layer 4, alone. It is due to a lack of any such observational details in the literature that the model, as shown in Figure 8, is not well supported by experimental data.

A similar case could (in principle) be made by considering the results of anatomical experiments tracing the distribution of R and E axons from identified extrastriate visual areas. One crucial fact is that V1 is not the only source of R-type afferents: all areas so far examined (V2, V4, V5/MT, TEO, and TE) provide both R and E afferents, although V1 is the only known area whose corticopulvinar output is predominantly R type. R and E afferents from the same site of origin (i.e., the few cubic millimeters where tracer was deposited in cortex) may terminate separately or in overlapping fields. A systematic pattern with strictly segregated terminal zones, as envisaged

by the model in Figure 8, is not in evidence. Perhaps, if there is an integrated thalamocortical hierarchy, it might not require such regular stratified zones in the pulvinar; a mosaic organization would provide an alternative, in which the formation of strong loops is avoided by patchy local arrangements of driving and modulatory circuitry. A final avenue of study would be to chart the distribution of core and matrix relay cells projecting to the same site in cortex, but this has yet to be performed for visual cortex and pulvinar.

What Are the Functions of Corticopulvinocortical Circuitry?

Theoretical ideas relating to the function of second-order thalamic nuclei are sketchy at best. A recurring notion in regard of the pulvinar is that it is in some way associated with attentional processes. Typically, such ideas address the medial pulvinar in relation to its involvement with frontoparietal cortical circuits known to mediate attention. However, adopting an anatomical approach, it is equally possible to apply this idea to the entire pulvinar. In a neural context, attention must address the means by which brain areas interact cooperatively to select certain forms of processing over others, as dictated by the current behavioral goal. In vision, for example, this may require scrutinizing one particular object. The pulvinar, with its extensive links to the entire visual cortex, is well equipped to take part in this process. The logic of the replication principle may be that the pulvinar acts to facilitate or regulate transmission within existing cortical networks – a kind of remote hub for controlling network activity. The quandary, therefore, is how information transmitted via a corticopulvinocortical (CPC) loop differs in nature from that passing over the equivalent direct corticocortical pathway. The remainder of the article offers a couple of general formulations.

The Pulvinar as a Salience System

Neural models of attention frequently invoke competition. Attending to one object in a busy scene is a process whereby neurons representing that object outcompete other neurons representing rival objects. Because any object is multiply represented across several different visual areas (with each area emphasizing a different kind of feature), there are several simultaneous competitions. An integrated competition model proposes a mechanism to ensure that the same object wins out in all areas. Anatomically, CPC circuitry is an attractive candidate: The winning location emerges from a pooling of all relevant information within the visual map(s) of the pulvinar and is rebroadcast back to the relevant location in each cortical map. In this scheme, pulvinar activity is related to the salience of each object. Salience is a composite signal, having a basis in the raw, visual conspicuity of each object but also modified by behavioral contingency (if bananas are the object of search, anything yellow and/or crescent shaped has enhanced salience). It is also possible to attend relatively featureless locations in space, and such a spatial attentional bias signal could be injected into pulvinar from frontoparietal cortex, either directly or by means of a reentrant pathway via the superior colliculus. Several features of pulvinar physiology are compatible with this concept, such as the lack of feature selectivity in comparison to cortical neurons. In addition, the characteristic bursting property of thalamic relay cells, when they are reactivated from a quiescent state, is a mechanism that lends itself to signaling the salience of a freshly significant location. In summary, the concept of salience illustrates how CPC circuitry might replicate the pattern of cortical circuitry without necessarily duplicating it in function.

The salience model derives from topographic considerations and is rather less informed by knowledge of the local circuitry. A key prediction is that the pulvinar is itself a site of neural competition (e.g., a winner-take-all salience map as envisaged by some models). The architecture of its intrinsic inhibitory circuits is unknown, but the prediction is that there should be long-range lateral inhibition within the plane of mapping (i.e., the plane perpendicular to the AIR) and facilitation (perhaps disinhibition) along the AIR in order to propagate a winning location across the connection zones of different cortical areas. Either process might be augmented by recurrent connections with the TRN. The primary source of activity for the representation of each object within a salience map would be driving (R-type) input from V1. Reentrant collicular afferents might also require a driving action if they are expected to influence activity at a featureless location. The shape of subsequent interactions with higher visual areas is only poorly constrained by existing details of driving and modulatory circuits.

The Pulvinar as a Switchable Central Processor

Another feature associated with the operation of the CTC loop is the generation of rhythmic activity, in which corticothalamic inputs have been demonstrated to exert a synchronizing role. The matrix relay cells, with their diffuse anatomical connectivity, are thought to be an important means by which coherent activity broadens its distribution. Synchronized activity in the cortex, although broadly distributed across areas, is also locally restricted (e.g., to particular columns). The binding-by-synchrony theory posits that rival objects may be represented by distributed cortical networks,

each united by synchronized activity but distinguished by phase. Hence, phase of synchronicity is an alternative (perhaps complementary) domain for hosting an attention-related competition. The crux of the idea is that an incipient transcortical network requires thalamic access in order to grow and distribute across multiple areas, and that networks representing rival objects would compete to recruit pulvinar circuits that, lacking feature selectivity, have no initial affiliation. In this scenario, there is less requirement for competition within the pulvinar. Once the pulvinar, or much of it, has been entrained to the same phase, this phase can be rebroadcast to a wider expanse of cortex. These pulvinocortical signals would not necessarily carry stimulus-specific information but can be expected to sensitize their recipients to corticocortical signals of the same phase, emitted by an increasingly dominant, object-specific network. Acting in this way, CPC circuitry would be (one) anatomical analog for the limited-capacity but task-switchable processing machinery typically envisaged by cognitive accounts of attention.

See also: Corticothalamic Connections: Ultrastructure; Corticothalamic Connections: Structure and Function; Presynaptic Facilitation; Presynaptic Inhibition; Superior Colliculus; Visual Cortex in Humans.

Further Reading

Adams MM, Hof PR, Gattass R, Webster MJ, and Ungerleider LG (2000) Visual cortical projections and chemoarchitecture of macaque monkey pulvinar. *Journal of Comparative Neurology* 419: 377–393.

Bender DB (1981) Retinotopic organization of macaque pulvinar. *Journal of Neurophysiology* 46: 672–693.

Benevento LA and Standage GP (1983) The organization of projections of the retinorecipient and nonretinorecipient nuclei of the pretectal complex and layers of the superior colliculus to the lateral pulvinar and medial pulvinar in the macaque monkey. *Journal of Comparative Neurology* 217: 307–336.

Crick F and Koch C (1998) Constraints on cortical and thalamic projections: The no-strong-loops hypothesis. *Nature* 391: 245–250.

Darian-Smith C, Tan A, and Edwards S (1999) Comparing thalamocortical and corticothalamic microstructure and spatial reciprocity in the macaque ventral posterolateral nucleus (VPLc) and medial pulvinar. *Journal of Comparative Neurology* 410: 211–234.

Gutierrez C, Cola MG, Seltzer B, and Cusick C (2000) Neurochemical and connectional organization of the dorsal pulvinar complex in monkeys. *Journal of Comparative Neurology* 419: 61–86.

Jones EG (2002) Thalamic circuitry and thalamocortical synchrony. *Philosophical Transactions of the Royal Society of London, Series B* 357: 1659–1673.

Robinson DL and Petersen SE (1992) The pulvinar and visual salience. *Trends in Neurosciences* 15: 127–132.

Rockland K (1996) Two types of corticopulvinar terminations: Round (type 2) and elongate (type 1). *Journal of Comparative Neurology* 368: 57–87.

Sherman SM and Guillery RW (1996) Functional organization of thalamocortical relays. *Journal of Neurophysiology* 76: 1367–1395.

Sherman SM and Guillery RW (2002) The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society of London, Series B* 357: 1695–1708.

Shipp S (2001) Corticopulvinar connections of areas V5, V4 and V3 in the macaque monkey: A dual model of retinal and cortical topographies. *Journal of Comparative Neurology* 439: 469–490.

Shipp S (2003) The functional logic of cortico-pulvinar connections. *Philosophical Transactions of the Royal Society of London, Series B* 358: 1605–1624.

Shipp S (2004) The brain circuitry of attention. *Trends in Cognitive Sciences* 8: 223–230.

Ungerleider LG, Galkin TW, and Mishkin M (1983) Visuotopic organization of projections from striate cortex to inferior and lateral pulvinar in rhesus monkey. *Journal of Comparative Neurology* 217: 137–157.

Purinergic Receptors

G E Knight, Royal Free and University College Medical School, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Receptors for adenine nucleosides and nucleotides have been divided into P1 (those responsive to adenosine and AMP) and P2 (those responsive to purine and pyrimidine di- and triphosphates). P2 receptors are further divided into P2X (ligand-gated cation channels) and P2Y (G-protein-coupled) receptors. There are four subtypes of adenosine receptor – A_1, A_{2A}, A_{2B}, and A_3 receptors – and all members are G-protein-coupled receptors. A_1 and A_3 receptors mediate inhibition of adenylate cyclase, whereas A_{2A} and A_{2B} receptors stimulate this enzyme. To date, the P2X receptor family has seven members ($P2X_{1–7}$) and the P2Y family has eight ($P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, $P2Y_{11}$, $P2Y_{12}$, $P2Y_{13}$, and $P2Y_{14}$). The structure, molecular biology, and second messenger signaling of these receptors have been discussed in detail in some excellent reviews.

This article is concerned with the distribution and function of these receptors on autonomic neurons and neuroeffector cells. However, considering the ubiquitous nature of purines and pyrimidines as signaling molecules, the diversity of receptor subtypes, and the fact that in almost all systems of the body ATP, ADP, and/or adenosine are signaling molecules, this article is not able to provide a comprehensive review.

Nervous System

The nervous system expresses multiple P1 and P2 receptor subtypes (Table 1). Peripheral sympathetic neurons have been extensively investigated and shown to express all the P2X receptor subtypes, although functionally the responses of sympathetic neurons can be accounted for by the presence of $P2X_2$ and $P2X_3$ subunits in varying proportions of homomeric and heteromeric receptors, in addition to $P2X_1$ and $P2X_5$ receptors. The dominant P2Y receptors expressed by sympathetic neurons appear to be $P2Y_1$, $P2Y_2$, $P2Y_4$, and $P2Y_6$ receptors. The predominant adenosine receptor is of the A_1 subtype. In contrast, there have been fewer studies of the expression of P1 and P2 receptors on parasympathetic neurons. To date, $P2X_2$, $P2X_4$, $P2Y_2$, and $P2Y_4$ receptor subunits have been identified, although functional responses of parasympathetic neurons are mediated by $P2X_2$ and $P2X_3$ subunits in varying proportions of homomeric and heteromeric receptors and by $P2Y_1$, $P2Y_2$, and $P2Y_{11}$ receptors. As with sympathetic neurons, A_1 receptors are the predominant subtype of adenosine receptor expressed by parasympathetic neurons.

Peripheral sensory neurons have been shown to also express all the P2X receptor subtypes, and $P2Y_1$, $P2Y_2$, and $P2Y_4$ receptors have been identified. Functional responses of sensory neurons are a mixture of $P2X_2$ and $P2X_3$ subunits as homomeric and heteromeric receptors and also $P2Y_1$ and $P2Y_2$ receptors. The principal adenosine receptor expressed by sensory neurons is of the A_{2A} subtype.

Enteric neurons express multiple P2X and P2Y receptor subtypes, with $P2X_3$, $P2X_4$, $P2X_7$, $P2Y_1$, $P2Y_6$, and $P2Y_{12}$ subtypes being abundant, and although there are multiple functional P2X receptors, the $P2Y_1$ receptor is the dominant functional P2Y subtype. With respect to adenosine receptors, A_1, A_{2A}, and A_{2B} receptors have been identified.

Musculoskeletal System

Skeletal Muscle

All of the P2X receptor subtypes are expressed by skeletal muscle; although the functional receptor has not been fully characterized for developing and young animals, there is evidence for the involvement of the $P2X_2$ receptor in the development of the neuromuscular junction. In contrast, in the adult, $P2X_1$, $P2X_3$, and $P2X_5$ receptors have been described. ATP, in both early development and adults, has a role in the upregulation of acetylcholine receptors and inhibition of acetylcholine release. Mammalian motor nerve terminals have been shown to express $P2X_7$ receptors. Developing skeletal muscle expresses $P2Y_1$ receptors, as does the neuromuscular junction, whereas in adults the dominant receptors expressed are the $P2Y_2$ and A_{2A} subtypes.

Bone and Cartilage

Multiple P2X and P2Y receptor subtypes are expressed in osteoblasts, although functionally the dominant receptors are $P2X_7$, $P2Y_1$, and $P2Y_2$, which regulate bone formation (Table 2). Osteoclasts have been found to express a similar distribution of P2X and P2Y receptors, with the addition of a functional $P2X_4$ receptor. Adenosine receptors are also expressed by osteoblast-like cells, inducing proliferation, although

Table 1 Principal P1 and P2 receptors expressed by excitable tissues

Neuronal	
Sympathetic neurons	$P2X_{1-7}$, $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, A_1
Parasympathetic neurons	$P2X_2$, $P2X_3$, $P2X_4$, $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_{11}$, A_1
Sensory neurons	$P2X_{1-7}$, $P2Y_1$, $P2Y_2$, $P2Y_4$, A_{2A}
Enteric neurons	$P2X_3$, $P2X_4$, $P2X_7$, $P2Y_1$, $P2Y_6$, $P2Y_{12}$, A_1, A_{2A}, A_{2B}
Muscle cells	
Smooth muscle	$P2X_{1-7}$, $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, A_1, A_{2A}, A_{2B}, A_3
Skeletal muscle	
Developing	$P2X_2$, $P2X_5$, $P2X_6$, $P2Y_1$, $P2Y_2$
Adult	$P2X_{1-7}$, $P2Y_2$, A_{2A}
Cardiac muscle	$P2X_{1-6}$, $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, A_1, A_3

Table 2 Principal P1 and P2 receptors expressed by nonneuronal cells

Osteoblasts	$P2X_7$, $P2Y_1$, $P2Y_2$
Cartilage	$P2X_2$, $P2Y_1$, $P2Y_2$, A_{2A}, A_{2B}
Keratinocytes	$P2X_5$, $P2X_2$, $P2X_3$, $P2X_7$, $P2Y_1$, $P2Y_2$, $P2Y_4$, A_{2B}
Fibroblasts	$P2X_7$, $P2Y_1$, $P2Y_2$, A_{2A}
Adipocytes	$P2X_1$, $P2Y_1$, $P2Y_2$, $P2Y_4$, A_1
Epithelial cells	$P2X_5$, $P2X_6$, $P2X_4$, $P2X_7$, $P2Y_1$, $P2Y_2$, $P2Y_{11}$, A_1, A_{2A}, A_3
Hepatocytes	$P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, $P2Y_{13}$, A_{2A}, A_3
Glial cells	$P2X_{1-7}$, $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, $P2Y_{11}$, $P2Y_{12}$, $P2Y_{13}$, A_1, A_2
Sperm	$P2X_2$, $P2X_7$, $P2Y_2$, A_1
Endothelial cells	$P2X_1$, $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, A_1, A_{2A}
Erythrocytes	$P2X_2$, $P2X_4$, $P2X_7$, $P2Y_1$
Platelets	$P2X_1$, $P2Y_1$, $P2Y_{12}$, A_{2A}
Immune cells	$P2X_4$, $P2X_7$, $P2Y_1$, $P2Y_2$, A_{2A}, A_3
Exocrine secretary cells	$P2X_1$, $P2X_4$, $P2X_7$, $P2Y_1$, $P2Y_2$, $P2Y_4$, A_1, A_{2A}
Endocrine secretory cells	$P2X_{1-7}$, $P2Y_2$, $P2Y_4$, A_1, A_{2A}, A_{2B}, A_3
Special senses	
Inner ear	$P2X_1$, $P2X_2$, $P2X_3$, $P2X_7$, $P2Y_2$, $P2Y_4$, A_1
Eye	$P2X_2$, $P2X_7$, $P2Y_2$, A_1, A_2, A_3
Tongue	$P2X_2$, $P2X_3$, $P2Y_1$, A_1
Olfactory organ	$P2X_2$, $P2X_4$, $P2Y_1$, $P2Y_2$, A_{2A}, A_3

the subtype has not been fully characterized. All four adenosine subtypes are expressed by the human osteoprogenitor cell line HCC1.

The dominant P2X receptor expressed by cartilage is the $P2X_2$ subtype, whereas the dominant P2Y receptors are $P2Y_1$ and $P2Y_2$. A_{2A} and A_{2B} receptors have been shown to be expressed by human articular chondrocytes.

Skin

Mammalian keratinocytes express $P2Y_1$, $P2Y_2$, and $P2Y_4$ receptors, and both $P2Y_1$ and $P2Y_2$ receptors have been identified functionally, mediating proliferation. In addition, $P2X_5$ receptors mediating differentiation and $P2X_7$ receptors, involved in cell death, are expressed. $P2X_2$ and $P2X_3$ receptors are also expressed by keratinocytes. Epidermal Langerhans cells express $P2X_7$ receptors, which are believed to have a role in the skin immune response. Adenosine has been shown to enhance the growth of a keratinocyte cell line via an A_{2B} receptor.

Connective Tissue

Fibroblasts express predominately $P2X_7$ receptors, involved in cell death, in addition to $P2Y_1$ and $P2Y_2$ receptors, shown to be involved in the control of proliferation. Fibroblasts express all of the adenosine receptor subtypes; in particular, the A_{2A} receptor has been shown to be involved in stimulating wound healing.

Brown adipocytes express multiple P2X and P2Y receptors, although the dominant functional receptors belong to the $P2Y_1$ and $P2Y_2$ or $P2Y_4$ subtypes. Although a functional P2Y receptor has been described in white adipocytes, it has yet to be characterized. A_1 receptors have been shown to be expressed by both brown and white adipocytes.

Gastrointestinal and Related Systems

Gastrointestinal System

The dominant P2 receptors expressed by intestinal smooth muscle are $P2X_1$, $P2X_2$, $P2X_7$, and $P2Y_1$ and $P2Y_2$ receptors. The P2X receptor subtypes mediate contraction of intestinal smooth muscle and the P2Y receptor subtypes mediate relaxation. In addition, adenosine A_1, A_{2A}, and A_{2B} receptors have been shown to be expressed (Table 2).

Epithelial cells of the intestine express multiple P2X and P2Y receptors, but the dominant receptor subtypes are $P2X_5$, $P2X_7$, $P2Y_1$, and $P2Y_2$ receptors, in addition to adenosine A_{2B} receptors. Interstitial cells of Cajal express $P2X_2$ and $P2X_5$ receptors, and it is thought that ATP may provide a feedback mechanism for pacemaker activity.

Liver and Biliary System

Multiple P2Y receptor subtypes are expressed by hepatocytes and cholangiocytes, the two principal liver epithelial cell types, and functional $P2Y_1$, $P2Y_2$, and $P2Y_4$ receptors have been identified. ATP has a role in regulating hepatocyte gluconeogenesis and mediating bile release from cholangiocytes. Similarly, $P2Y_2$, $P2Y_4$, and $P2Y_6$ receptors expressed by bile duct epithelial cells also modulate bile release. In rat liver, all four adenosine receptor subtypes are

expressed, although the dominant receptor is the A_3 subtype, which is involved in glucose release, particularly following ischemia. Adenosine A_{2A} receptors are expressed in rat and human hepatic stellate cell lines.

Enteric Glial Cells

There are few studies of the expression of P1 and P2 receptors on enteric glial cells, and only $P2X_7$ and $P2Y_2$ or $P2Y_4$ receptors have been shown to be expressed. No evidence for the expression of P1 receptors has been put forward.

Respiratory System

Lung

$P2Y_2$, $P2Y_4$, and $P2Y_6$ receptor subtypes are the predominant receptor subtypes expressed by airway smooth muscle and epithelial cells. In addition, the dominant functional P2X receptor subtypes are of the $P2X_4$ and $P2X_5$ subtypes, where ATP is involved in mucociliary clearance and surfactant release. All of the P1 receptor subtypes are expressed by the lungs. A_1, A_{2B}, and A_3 receptors are involved in adenosine-induced bronchoconstriction, whereas A_{2A} receptors are involved in bronchodilation.

Trachea

The dominant receptors expressed by tracheal smooth muscle and epithelial cells are of the $P2Y_1$ and $P2Y_2$ subtypes. Functional $P2X_1$ receptors are also expressed by smooth muscle mediating contraction, and $P2X_4$ and $P2X_7$ receptors by epithelium activating ciliary function and increasing $[Ca^{2+}]_i$, respectively.

Neuroepithelial Bodies

Only $P2X_3$ and $P2X_2$ receptors have been identified on neuroepithelial bodies, and it has been suggested that the functional receptor is a $P2X_{2/3}$ heteromeric receptor, possibly involved in ATP mechanosensory transduction and O_2 sensing.

Nasal Respiratory Epithelium

$P2X_3$ receptors are expressed by primary olfactory neurons located both in the olfactory epithelium and on subepithelial nerve fibers in the respiratory region. $P2X_5$ and $P2X_7$ receptors are expressed by epithelial cells of the rat nasal mucosa. $P2Y_1$, $P2Y_2$, $P2Y_6$, and $P2Y_{11}$ receptors are found to be expressed by respiratory epithelium. Although adenosine receptors are known to be expressed by apical epithelial cells, they have only been characterized as A_2 receptors.

Urinogenital System

Bladder and Urethra

The expression of all of the known P2X receptor subtypes has been shown in smooth muscle and urothelium of the bladder, although functionally the $P2X_1$ receptor, mediating detrusor contractility, is dominant. The bladder smooth muscle also contains a functional P2Y receptor mediating detrusor relaxation, although this has not been fully characterized, whereas the urothelium has been shown to express $P2Y_1$, $P2Y_2$, and $P2Y_4$ receptors. All four adenosine receptors are expressed in the urothelium. A_1 receptors are prominently localized to the apical membrane of the umbrella cell layer, whereas A_{2A}, A_{2B}, and A_3 receptors are localized intracellularly or on the basolateral membrane of umbrella cells and the plasma membrane of the underlying cell layers. One function of these receptors is thought to be the modulation of exocytosis in umbrella cells.

Urethral smooth muscle cells have been shown to express $P2Y_1$, $P2Y_2$, $P2Y_4$, and $P2Y_6$ receptors, and functional $P2Y_1$ and $P2Y_2$ receptors have been identified mediating relaxation, together with a P2Y subtype responsive to α,β-methylene ATP and an uncharacterized P1 receptor. Urethral epithelial cells are known to express $P2X_5$, $P2X_6$, and $P2X_7$ receptors, with the $P2X_5$ receptor being involved in differentiation of stratified squamous epithelial cells of the terminal urethra.

Ureter

Multiple P2X receptor subtypes are expressed by the ureter, where ATP is involved in smooth muscle contraction and also nociception, following release of ATP in response to distension. A_{2B} receptors have been shown to be expressed in pig intravesical ureter.

Kidney

P2 receptors are widely expressed in the kidney, including most of the P2X and P2Y receptor subtypes currently known. Structures within the glomerulus, including mesangial cells, podocytes, and glomerular endothelial and epithelial cells, express predominantly $P2X_4$, $P2X_5$, $P2X_6$, $P2Y_1$, $P2Y_2$, and $P2Y_4$ receptors. The descending and ascending limbs of the loop of Henle express predominantly $P2Y_1$, $P2Y_2$, and $P2Y_6$, whereas collecting ducts express multiple P2X and P2Y receptor subtypes.

All adenosine receptor subtypes are expressed in the cortex and medulla of the kidney to varying degrees, with differential expression on apical and basolateral surfaces such that proximal tubule

membranes express A_1, A_{2A}, and A_3 receptors, whereas basolateral membranes express A_1 and A_{2A} receptors.

Male Genital System

Functional $P2Y_1$ and $P2Y_2$ receptors have been shown to be present on penile corpus cavernosum, with ATP mediating relaxation of the cavernosum. $P2X_1$, $P2X_2$, and A_{2A} receptors are also expressed, which are believed to be involved in detumescence. Endothelial cells lining the lacunar space express $P2Y_1$ receptors.

In the excretory ducts (ductus epididymidis) of the testis, the dominant P2 receptors are $P2X_1$ and $P2X_2$ subtypes, expressed predominantly in the membranes of the smooth muscle layer, suggesting that these receptors are involved in the process of sperm transport and ejaculation. Sertoli cells express $P2X_4$, $P2X_7$, $P2Y_1$, and $P2Y_2$ receptors. The testes express high levels of A_3 receptors, and adenosine is thought to play an important role in regulating male reproduction. Seminiferous tubule myoid cells express multiple P2X receptor subtypes and $P2Y_2$ receptors. Similarly, multiple P2X receptor subtypes and A_1 receptors have been identified on sperm, and although functional P2X and P2Y receptors have been demonstrated, these remain uncharacterized.

Multiple P2X receptor subtypes are expressed in vas deferens smooth muscle, particularly $P2X_1$, $P2X_2$, $P2X_4$, and $P2X_7$, although functionally the smooth muscle receptor has been identified as being the $P2X_1$ subtype, mediating contraction of the vas deferens smooth muscle. The presence of a P2Y receptors, most likely of the $P2Y_2$ subtype, has also been suggested, although there is evidence for the presence of $P2Y_{12}$ and/or $P2Y_{13}$ receptors. Whereas A_1 receptors are expressed along the whole length of the rat vas deferens, the expression of A_2 receptors differs according to the portion of the vas deferens, with A_{2B} receptors being expressed by the prostatic portion and A_{2A} receptors by the epididymal portion.

In the epididymal epithelium and smooth muscle, $P2X_1$, $P2X_2$, $P2X_4$, $P2X_7$, $P2Y_1$, and $P2Y_2$ receptors are expressed. Functional A_{2A} receptors have been identified, but functional P2X and P2Y receptors are uncharacterized.

Smooth muscle $P2X_1$ and $P2Y_1$ receptors have been identified in the prostate, and $P2X_7$ receptors are expressed by prostate epithelial cells. The expression of $P2X_3$ receptors in interstitial connective tissue has been shown, probably associated with sensory nerves. Prejunctional A_1 receptors have been identified on prostatic smooth muscle, and A_1 and A_{2A} receptors have been identified on prostatic cell lines.

Female Genital System

Uterine smooth muscle expresses predominantly $P2X_2$ receptors when in a nonpregnant state, whereas both $P2X_1$ and $P2X_2$ receptors are expressed in pregnancy. In contrast, multiple P2X and P2Y receptors are expressed by both nonpregnant and pregnant uterine epithelium, particularly $P2Y_4$. Strong A_{2B} receptor expression has also been observed in the myometrium.

Axons in the cervix express $P2X_3$ receptors, presumed to be associated with sensory neurons, whereas cervical epithelium expresses $P2X_4$, $P2X_7$, and $P2Y_2$ receptors. Vaginal smooth muscle expresses $P2X_2$ and A_{2A} receptors, and vaginal epithelial cells, in common with cervical epithelium, express $P2Y_2$ receptors in addition to $P2X_5$ and $P2X_7$ receptors.

The ovary expresses $P2X_1$, $P2X_2$, and $P2Y_1$ receptor subtypes, with $P2X_2$ receptors being dominant in the nonpregnant state, switching to $P2X_1$ receptors during pregnancy. A_2 receptors have been identified in the ovary, but the subtype has not been fully characterized. Oviduct and fallopian tube epithelial cells have functional $P2Y_2$ receptors, and adenosine is known to inhibit noradrenaline release in human fallopian tubes, but the receptor subtype has not been characterized.

The placenta has a functional $P2Y_2$ receptor, although $P2Y_6$ and $P2Y_{11}$ receptors are also expressed. Multiple P2X receptor subtypes are expressed, but the $P2X_7$ receptor is dominant functionally. High- and low-affinity adenosine receptors have been described in the placenta, although these have not been fully characterized.

Cardiovascular System

Vascular Smooth Muscle

Vascular smooth muscle expresses multiple P2X and P2Y receptor subtypes, and this is true for both arteries and veins (Table 1). Although functionally the $P2X_1$ receptor subtype involved in vascular smooth muscle contraction is dominant, $P2Y_1$, $P2Y_2$, or $P2Y_4$ and $P2Y_6$ receptors have been shown to be functional. All the adenosine receptor subtypes are expressed and have functional effects, including prejunctional inhibition of neurotransmission via A_1 receptors on perivascular sympathetic nerves and vasodilatation via A_{2A} and A_{2B} receptors.

Vascular Endothelium

Similar to vascular smooth muscle, the endothelium expresses multiple P2X and P2Y receptor subtypes in addition to functional $P2Y_1$, $P2Y_2$ or $P2Y_4$, and $P2Y_6$

receptors. ATP and UTP are important vasodilators, but they also have roles in endothelial cell proliferation. Adenosine A_1 and A_{2A} receptors are expressed by endothelia cells mediating vasodilatation.

Heart

Expression of all of the P2X and P2Y receptor subtypes has been shown in the heart, although functionally $P2X_1$ and $P2Y_2$ receptors are dominant. In addition to the expression of adenosine receptors on the cardiac vasculature, all the adenosine receptor subtypes have been shown to be expressed in the heart, with functional effects ranging from cardioprotection to cardiac remodeling. In particular, A_1 receptors mediate cardiac depression, and A_1 and A_3 receptors have also been reported to mediate the protective effects of adenosine in preconditioning and during ischemia or reperfusion injury.

Erythrocytes

Erythrocytes express predominantly $P2X_7$ receptors and, at a much lower intensity, $P2X_2$ receptors, whereas erythroid progenitor cells also express $P2X_4$ and $P2Y_1$ receptors. From a functional standpoint, $P2X_7$ and $P2Y_2$ receptors are dominant.

Platelets and Megakaryocytes

Platelets express $P2X_1$ receptors, which are important in platelet activation, and $P2Y_1$ and $P2Y_{12}$ receptors, which are involved in ADP-induced platelet aggregation. In contrast, adenosine inhibits platelet aggregation via A_{2A} receptors.

Megakaryocytes express functional $P2X_1$, $P2Y_1$, $P2Y_2$, and $P2Y_{12}$ receptors. In contrast, megakaryocyte cell lines express multiple P2X and P2Y receptor subtypes, although only $P2Y_1$, $P2Y_2$, $P2Y_{11}$, and $P2Y_{12}$ subtypes have been found to be functional in addition to A_{2A} receptors.

Immune System

Immune Cells

Although multiple P2X and P2Y receptor subtypes are expressed in macrophages, neutrophils, eosinophils, dendritic cells, monocytes, lymphocytes, and hematopoietic cells, the dominant functional subtypes are $P2X_7$ receptors, involved in the formation of pores and cell death, and $P2Y_2$ receptors, involved in the release of inflammatory mediators. Similarly, all four adenosine receptor subtypes are expressed by immune cells, with the A_{2A} receptor being abundantly expressed; the A_3 receptor of eosinophils is thought to be involved in apoptosis.

Thymus

Multiple P2X receptor subtypes have been found to be expressed by both thymocytes and epithelial cells, but the $P2X_1$, $P2X_2$, and $P2X_7$ receptors are dominant, in addition to $P2Y_2$ receptors. Adenosine has been shown to induce apoptosis in immature thymocytes via adenosine A_{2A} receptors.

Spleen

The dominant P2 receptor subtypes expressed by the spleen are $P2X_1$, $P2Y_1$, $P2Y_2$, and $P2Y_{13}$, although functionally the dominant subtypes are $P2X_1$ and $P2Y_1$ receptors, in addition to A_1 and A_{2A} receptors.

Exocrine Glands

Pancreatic ducts express $P2X_1$, $P2X_4$, and $P2X_7$ receptors as well as $P2Y_1$, $P2Y_2$, and $P2Y_4$ receptor subtypes. The dominant P2 receptor expressed by pancreatic epithelial and acini cells is of the $P2Y_2$ subtype, although functional A_{2A} receptors have also been shown. Eccrine sweat glands have functional $P2Y_1$, $P2Y_2$, and $P2Y_4$ receptors, whereas the dominant receptors expressed by salivary glands are A_1, $P2X_4$, $P2X_7$, $P2Y_1$, and $P2Y_2$ receptors. The $P2X_7$ receptor is located on the luminal membrane, and the $P2Y_2$ receptor is typically situated on the basolateral membrane. Only the $P2Y_2$ receptor has been found to be expressed by lachrymal gland acinar cells.

Endocrine Glands

Endocrine Pancreas

Various P2 receptors have been found to be expressed in the pancreas, including $P2X_1$, $P2X_2$, $P2X_7$, $P2Y_1$, $P2Y_2$, and $P2Y_4$ subtypes, whereas functionally, $P2X_7$ and $P2Y_4$ are dominant on α cells, and $P2X_1$, $P2Y_1$, and $P2Y_2$ receptors are dominant on β cells. Adenosine acts on β cells to inhibit insulin secretion via an action on A_1 receptors.

Functional $P2Y_2$ receptors have been demonstrated on pancreatic epithelial cells, although $P2Y_1$ and $P2Y_4$ receptors are also expressed, in addition to $P2X_1$, $P2X_4$, and $P2X_7$ subtypes.

Adrenal Gland

All of the P2X receptor subtypes are expressed in adrenal tissues, although the dominant functional receptor is of the $P2X_2$ subtype, which modulates catecholamine and aldosterone secretion. The dominant functional P2Y receptor linked to steroidogenesis in bovine adrenocortical fasciculata cells is of the

$P2Y_2$ subtype. Bovine adrenal zona reticularis cells express at least three adenosine receptor subtypes – A_1, A_{2A}, and A_3 receptors – involved in the coordination of stress-induced secretion of corticosteroids and catecholamines from the adrenal gland.

Thyroid Gland

Thyroid follicle cells (cuboid epithelia) and thyroid epithelial cell lines express all of the P2X receptor subtypes with the exception of $P2X_1$ and $P2X_2$. $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, and $P2Y_{11}$ receptor subtypes are also expressed by thyrocytes, although $P2Y_2$ receptors are dominant functionally, involved in the modulation of cytokine release. Adenosine A_1 receptors have been found to be expressed by the thyroid cell line FRTL-5, which is believed to be involved in a regulatory feedback mechanism with thyroid-stimulating hormone receptors.

Pituitary Gland and Pineal Gland

The anterior pituitary expresses multiple P2X receptor subtypes, and this is reflected in the presence of multiple functional P2X receptors, with $P2X_2$ and $P2X_7$ receptors being dominant. In contrast, the $P2Y_2$ receptor is the only subtype identified, and it is involved in the release of luteinizing hormone. The posterior pituitary expresses $P2X_2$ and $P2X_6$ receptors, with the $P2X_2$ receptor being dominant functionally. The presence of functional A_1 and A_{2B} receptors in primary anterior pituitary cell cultures and in pituitary folliculostellate cell lines has been demonstrated, with the A_{2B} receptor being functionally dominant in the anterior pituitary. A_{2A} receptor gene expression has been reported to occur transiently during the embryological development of the anterior and intermediate lobes of the pituitary gland.

Only limited studies of the expression of P2 receptors have been carried out on the pineal gland, and only the expression of $P2Y_1$, $P2Y_2$, and $P2Y_4$ receptors has been demonstrated. It has been suggested that $P2Y_1$ receptors present in rat pinealocytes mediate the potentiation of noradrenaline-induced *N′*-acetyl-5-hydroxytryptamine production (the precursor of melatonin), whereas one of the functions of $P2Y_2$ receptors on the pineal gland is the selective inhibition of β-adrenergic receptor-mediated signaling pathways. ATP is thought to be a cotransmitter with noradrenaline, acting on an uncharacterized P2X receptor (although probably $P2X_1$) to stimulate *N′*-acetyl-5-hydroxytryptamine and melatonin release. In contrast, all of the adenosine receptor subtypes have been identified in pinealocytes, and a role for adenosine in the regulation of melatonin synthesis has been postulated, with the A_{2B} receptor being dominant.

Endocrine Ovary

Granulosa–luteal cells express functional $P2Y_2$ receptors, whereas granulosa cells express functional $P2Y_1$ and $P2Y_2$ receptors. There is no evidence for the expression of P2X receptors. Adenosine has been shown to enhance gonadotropin stimulation in granulosa and luteal cells; this effect has been found to be via an A_2 receptor, although the exact subtype has not been identified.

Special Senses

Inner Ear

Structures within the inner ear express multiple P2X receptor subtypes. The cochlear hair cells express $P2X_1$, $P2X_2$, $P2X_3$, and $P2X_7$ receptors, with the $P2X_2$ receptor being dominant functionally. However, $P2X_3$ receptors are believed to be important in cochlear synaptic reorganization and establishment of neurotransmission, which occurs just prior to the onset of hearing function.

The dominant receptor subtypes expressed by epithelial cells of the inner ear are of the $P2X_2$ subtype, whereas for vestibular dark cells, the dominant receptors are $P2Y_2$ and $P2Y_4$ receptors. Both Hensen and Deiters' cells of the organ of Corti express $P2X_2$ receptors involved in the regulation of ion and H_2O balance of cochlear fluid and purinergic modulation of cochlear micromechanisms, respectively.

A_1 receptors have been identified in the cochlear, where they are thought to have a protective effect.

Eye

Structures within the retina have been shown to express multiple P2X and P2Y receptor subtypes, although little is known about the functional role of these receptors in retinal signaling. $P2X_7$ receptors expressed by the inner and outer plexiform layers have been shown to have a significant role in neurotransmission within the retina and may modulate both photoreceptor and rod bipolar cell responses. $P2X_2$ receptors expressed by large ganglion cells as well as in a subset of amacrine cells may be involved in neuronal signaling in the retina, particularly within cone pathways. Activation of $P2Y_2$ receptors stimulates retinal pigment epithelium active transport. $P2Y_2$ receptors are dominant in the ciliary body, where they are involved in the regulation of aqueous humour secretion. Adenosine receptors are also expressed by structures of the eye. Adenosine, via A_1 receptors, has been shown to protect cultured retinal neurons against *N*-methyl-D-aspartate-induced cell death. Photoreceptors express a functional A_2-like

adenosine receptor, involved in the inhibition of L-glutamate release from rod photoreceptors. Ganglion cells have been shown to express A_1 and A_3 receptors, which are believed to be involved in protection from ischemic injury.

Tongue

The expression of $P2X_2$ and $P2X_3$ receptor subtypes has been shown on the nerve fibers innervating taste buds. The $P2X_2$ receptor is probably associated with taste sensation, whereas the $P2X_3$ receptor may play a role in nociception. $P2Y_1$ receptor expression has been detected in a subset of taste bud cells, where they may play a role in ATP-mediated signal transduction between taste bud cells and afferent taste fibers. The presence of an A_1 adenosine receptor has been suggested, which is believed to be important in the modulation of taste perception.

Olfactory Organ

The dominant P2 receptors expressed by the olfactory bulb are $P2X_2$, $P2X_4$, $P2Y_1$, and $P2Y_2$ receptors. Although functional P2X and P2Y receptors have been identified in olfactory receptor neurons, the subtypes have not been characterized. Adenosine A_{2A} and A_3 receptors have been identified in the olfactory bulb, but their exact roles in olfaction are unknown.

Conclusion

This article provided an overview of the distribution of P1 and P2 receptors on autonomic neurons and neuroeffector cells. It should be stressed that this distribution is extensive, with most systems in the body expressing multiple receptor subtypes, having diverse effects such as vasoconstriction, platelet aggregation, nociception, and apoptosis. This article concentrated on the distribution of P1 and P2 receptors in mammalian tissues; however, there is increasing evidence that purinergic signaling is a very primitive system. Evidence has shown that receptors for ATP and adenosine are expressed in animals from primitive phyla such as protozoa (in which ATP inhibits amoeboid movement), mollusks (in which adenosine modulates an excitatory acetylcholine response of *Helix aspersa* subesophageal neurons), and insects (in which ATP stimulates the gorging response of mosquitoes) to lower vertebrates (in which ATP constricts and adenosine dilates the frog aorta). Although there are similarities between these more primitive receptors and those of mammalian receptors, there are many differences, and until these primitive receptors are cloned, a full characterization remains difficult. The first such cloning and characterization of a nonvertebrate ATP-gated ion channel has been carried out for the parasitic blood fluke *Schistosoma mansoni*, which has been characterized as $P2X_4$-like. The field is wide open to the characterization of many more nonmammalian receptors.

A final feature of purinergic signaling that should be noted is the expression of multiple receptor subtypes on single cells. Why are so many different receptor subtypes necessary? A review has attempted to put forward a theory as to why multiple P2 receptors are expressed by cells. It suggests that P2 receptor signaling is an example of increased receptor complexity, either because newly synthesized P2 receptor proteins are casually located on the cell surface or because distinct subunits are expressed by different cell phenotypes determined by genetic control and/or there is selective recruitment under pathophysiological conditions. The idea behind this concept is that the presence of multiple receptor subtypes allows a 'fine-tuning' of the response. Although this idea requires further investigation, it does open a new area of research into purinergic signaling.

See also: Adenosine; Adenosine Receptor Mediated Functions; Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Gastrointestinal Control; Autonomic Nervous System: Respiratory Control; Autonomic Nervous System: Urogenital Control; Autonomic Nervous System: Central Urogenital Control; Calcium Waves: Purinergic Regulation; Immune System–Neuroendocrine Interactions; Pharmacology of Sleep: Adenosine; Purines and Purinoceptors: Molecular Biology Overview.

Further Reading

Abbracchio MP, Burnstock G, Boeynaems J-M, et al. (2006) International Union of Pharmacology. Update and subclassification of the P2Y G protein-coupled nucleotide receptors: From molecular mechanisms and pathophysiology to therapy. *Pharmacological Reviews* 58: 281–341.

Burnstock G (1996) Purinoceptors: Ontogeny and phylogeny. *Drug Development Research* 39: 204–242.

Burnstock G (2001) Purinergic signalling in development. In: Abbracchio MP and Williams M (eds.) *Handbook of Experimental Pharmacology, Volume 151/I. Purinergic and Pyrimidinergic Signalling I: Molecular, Nervous and Urinogenitary System Function*, pp. 89–127. Berlin: Springer.

Burnstock G (2003) Purinergic receptors in the nervous system. In: Schwiebert EM (ed.) *Current Topics in Membranes, Vol. 54. Purinergic Receptors and Signalling*, pp. 307–368. San Diego: Academic Press.

Burnstock G (2006) Pathophysiology and therapeutic potential of purinergic signalling. *Pharmacological Reviews* 58: 58–86.

Burnstock G (2007) Physiology and pathophysiology of purinergic transmission. *Physiological Reviews* 87: 659–797.

Burnstock G and Knight GE (2004) Cellular distribution and functions of P2 receptor subtypes in different systems. *International Review of Cytology* 240: 31–304.

Dunn PM, Zhong Y, and Burnstock G (2001) P2X receptors in peripheral neurones. *Progress in Neurobiology* 65: 107–134.

Fredholm BB, Ijzerman AP, Jacobson KA, Klotz KN, and Linden J (2001) International Union of Pharmacology. XXV: Nomenclature and classification of adenosine receptors. *Pharmacological Reviews* 53: 527–552.

Gever J, Cockayne DA, Dillon MP, Burnstock G, and Ford APDW (2006) Pharmacology of P2X channels. *Pflugers Archiv, European Journal of Physiology* 452: 513–537.

Illes P and Ribeiro JA (2004) Neuronal P2 receptors of the central nervous system. *Current Topics in Medicinal Chemistry* 4: 831–838.

Jacobson KA and Gao ZG (2006) Adenosine receptors as therapeutic targets. *Nature Reviews Drug Discovery* 5: 247–264.

Khakh BS, Burnstock G, Kennedy C, et al. (2001) International Union of Pharmacology. XXIV: Current status of the nomenclature and properties of P2X receptors and their subunits. *Pharmacological Reviews* 53: 107–118.

North RA (2002) Molecular physiology of P2X receptors. *Physiological Reviews* 82: 1013–1067.

Ralevic V and Burnstock G (1998) Receptors for purines and pyrimidines. *Pharmacological Reviews* 50: 413–492.

Purines and Purinoceptors: Molecular Biology Overview

G Burnstock, Royal Free and University College Medical School, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Early Studies

A seminal paper describing the potent actions of adenine compounds was published by Drury and Szent-Györgyi in 1929. Many years later, ATP was proposed as the transmitter responsible for nonadrenergic, noncholinergic transmission in the gut and bladder, and the term 'purinergic' was introduced by Burnstock in 1972. Early resistance to this concept appeared to stem from the fact that ATP was recognized first for its important intracellular roles in many biochemical processes, and the intuitive feeling was that such a ubiquitous and simple compound was unlikely to be utilized as an extracellular messenger, although powerful extracellular enzymes involved in its breakdown were known to be present.

Implicit in the concept of purinergic neurotransmission was the existence of postjunctional purinergic receptors, and the potent actions of extracellular ATP on many different cell types also implicated membrane receptors. Purinergic receptors were first defined in 1976, and 2 years later a basis for distinguishing two types of purinoceptor, identified as P1 and P2 (for adenosine and ATP/ADP, respectively), was proposed. At about the same time, two subtypes of the P1 (adenosine) receptor were recognized, but it was not until 1985 that a pharmacological basis for distinguishing two types of P2 receptor (P2X and P2Y) was made. In 1993, the first G-protein-coupled P2 receptors were cloned and a year later two ion-gated receptors were cloned, and in 1994 Abbracchio and Burnstock, on the basis of molecular structure and transduction mechanisms, proposed that purinoceptors should belong to two major families: a P2X family of ligand-gated ion channel receptors and a P2Y family of G-protein-coupled purinoceptors. This nomenclature has been widely adopted and currently seven P2X subtypes and eight P2Y receptor subtypes are recognized, including receptors that are sensitive to pyrimidines as well as purines.

It is widely recognized that purinergic signaling is a primitive system involved in many nonneuronal as well as neuronal mechanisms, including exocrine and endocrine secretion, immune responses, inflammation, pain, platelet aggregation, and endothelial-mediated vasodilatation. Cell proliferation, differentiation, and death that occur in development and regeneration are also mediated by purinergic receptors.

P1 Receptors

Four different P1 receptor subtypes, A_1, A_{2A}, A_{2B}, and A_3, have been cloned. All are G-protein-coupled receptors (GPCRs). At 318 amino acids in length, the A_3 subtype is the shortest, while A_{2A} is the longest (412 residues). Their N-termini are relatively short (7–13 residues in length), as are their C-termini (32–120 residues). In the transmembrane domains (TMI–TMVII), human adenosine receptors share 39–61% sequence identity with each other and 11–18% identity with P2Y receptors. Each of the four human P1 receptor genes contains an intron within the coding region, located immediately after the end of the third transmembrane domain (see Figure 1(a)).

P1 receptors couple principally to adenylate cyclase. A_1 and A_3 are negatively coupled to adenylate cyclase through the $G_{i/o}$ protein α subunits, whereas A_{2A} and A_{2B} are positively coupled to adenylate cyclase through G_s. The human A_{2B} receptor has also been observed to couple through $G_{q/11}$ to regulate phospholipase C activity, and the A_3 receptor may interact directly with G_s.

A number of P1 subtype-selective agonists and antagonists have been identified (see Table 1). Alteration or opening of the ribose ring drastically reduces affinity. The hydroxyl group at the 2′ position is needed for both affinity and activity. The most selective agonist for the A_1 subtype is 2-chloro-N^6-cyclopentyladenosine (CCPA). CGS 21680 is the most selective A_{2A} agonist; NECA is the most potent A_{2B} receptor agonist. 2-Cl-IB-MECA is 11-fold selective for the human A_3 receptor and about 1400-fold selective for the rat A_3 receptor. In general, methylxanthines such as caffeine and theophylline are weak P1 receptor antagonists. DPCPX (8-cyclopentyl-1,3-dipropylxanthine) is an A_1 receptor antagonist with subnanomolar affinity. The most selective A_{2B} receptor antagonist is MRS1754. MRE3008-F20 is the most selective human A_3 receptor antagonist. The diverse physiological effects mediated by the different P1 receptor subtypes, particularly modulation of the cardiovascular, immune, and central nervous systems, have been confirmed by transgenic knockout mice for A_1, A_{2A}, and A_3 receptors. In contrast to knockout studies, overexpression of either A_1 or A_3 subtypes in transgenic mice has a cardioprotective effect.

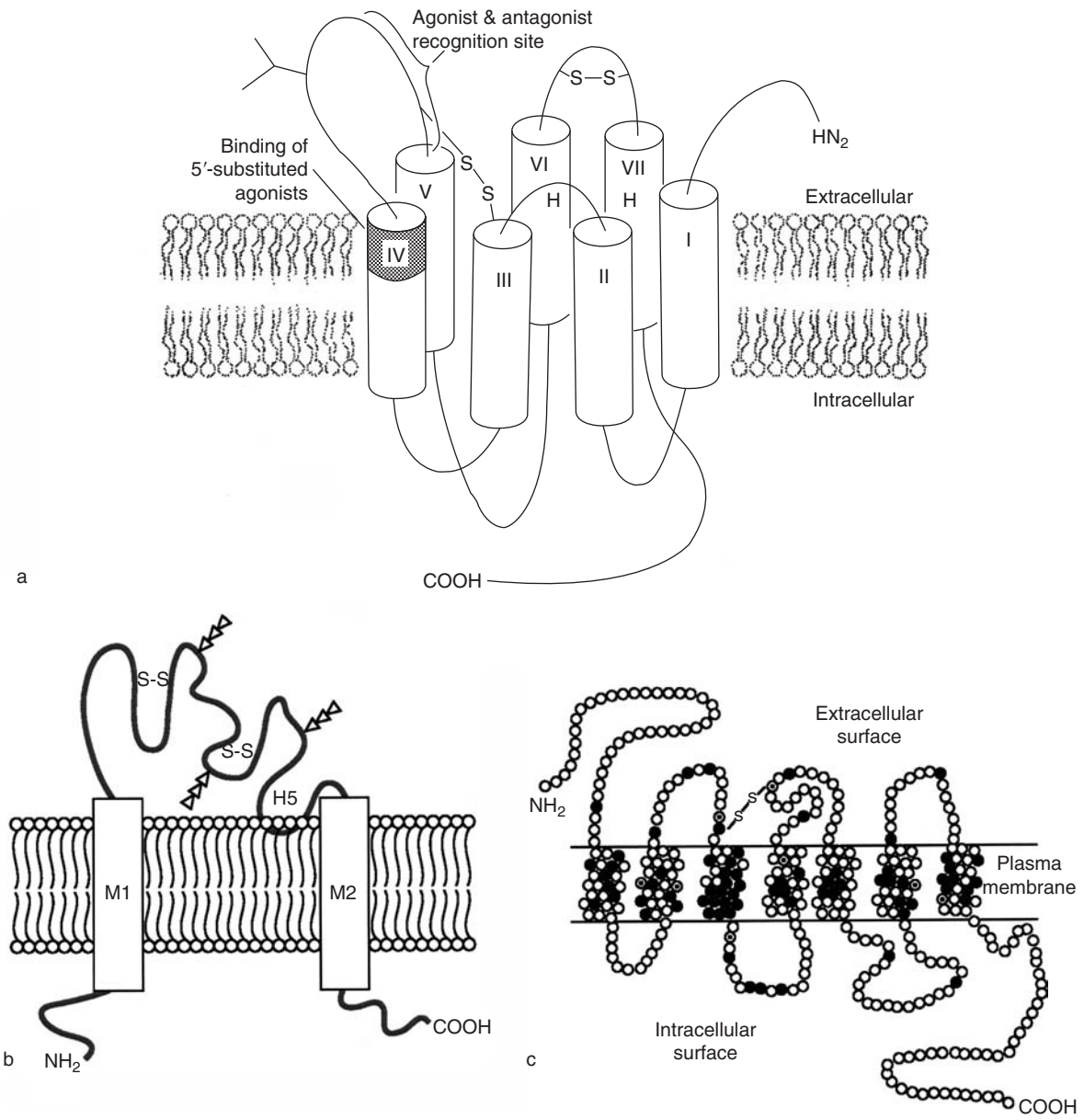

Figure 1 Membrane receptors for extracellular ATP and adenosine. The P1 family of receptors for extracellular adenosine comprises G-protein-coupled receptors signaling by inhibiting or activating adenylate cyclase (a). The P2 family of receptors binds extracellular ATP or ADP, and comprises two types of receptors (P2X and P2Y). The P2X family receptors are ligand-gated ion channels (b), and the P2Y family members are GPCRs (c). (a) Reproduced from Ralevic V and Burnstock G (1998) Receptors for purines and pyrimidines. *Pharmacological Reviews* 50: 413–492, with permission from the American Society for Pharmacology and Experimental Therapeutics. (b) Reprinted by permission from Macmillan Publishers Ltd: [Nature] (Brake AJ, Wagenbach MJ, and Julius D (1994) New structural motif for ligand-gated ion channels defined by an ionotropic ATP receptor. *Nature* 371: 519–523), copyright (1994). (c) Adapted from Barnard EA, Burnstock G, and Webb TE (1994) G protein-coupled receptors for ATP and other nucleotides: A new receptor family. *Trends in Pharmacological Sciences* 15: 67–70, with permission from Elsevier.

P2X Receptors

Molecular Structure

The first cDNAs encoding P2X receptor subunits were isolated in 1994. Members of the family of ionotropic $P2X_{1-7}$ receptors show a subunit topology of intracellular N- and C-termini possessing consensus binding motifs for protein kinases; two transmembrane-spanning regions (TM1 and TM2), the first involved with channel gating and the second lining the ion pore; a large extracellular loop, with ten conserved cysteine residues forming a series of disulfide bridges; a hydrophobic H5 region close to the pore vestibule, for possible receptor/channel modulation by cations; and an ATP-binding site, which may involve regions of the extracellular loop adjacent

Table 1 Characteristics of purine-mediated receptors

Receptor	*Main distribution*	*Agonists*	*Antagonists*	*Transduction mechanisms*
P1 (adenosine)				
A_1	Brain, spinal cord, testis, heart, autonomic nerve terminals	CCPA, CPA, S-ENBA	DPCPX, N-0840, MRS1754	$G_{i/o}$ ↓cAMP
A_{2A}	Brain, heart, lungs, spleen	CGS 21680, HENECA	KF17837, SCH58261, ZM241385	G_S ↑cAMP
A_{2B}	Large intestine, bladder	NECA (nonselective)	Enprofylline, MRE2029-F20, MRS1754, MRS1706	G_S ↑cAMP
A_3	Lung, liver, brain, testis, heart	IB-MECA, 2-Cl-IB-MECA, DBXRM, VT160	MRS1220, L-268605, MRS1191, MRS1523, VUF8504	$G_{i/o}$ $G_{q/11}$ ↓cAMP ↑IP_3
P2X				
$P2X_1$	Smooth muscle, platelets, cerebellum, dorsal horn spinal neurons	α,β-meATP = ATP = 2-MeSATP (rapid desensitization), L-β,γ-meATP	TNP-ATP, IP_5I, NF023, NF449	Intrinsic cation channel (Ca^{2+} and Na^+)
$P2X_2$	Smooth muscle, CNS, retina, chromaffin cells, autonomic and sensory ganglia	ATP ≥ ATPγS ≥ 2-MeSATP ≫ α,β-meATP (pH + zinc sensitive)	Suramin, isoPPADS, RB2, NF770	Intrinsic ion channel (particularly Ca^{2+})
$P2X_3$	Sensory neurons, NTS, some sympathetic neurons	2-MeSATP ≥ ATP ≥ α,β-meATP ≥ Ap_4A (rapid desensitization)	TNP-ATP, PPADS, A317491, NF110	Intrinsic cation channel
$P2X_4$	CNS, testis, colon	ATP ≫ α,β-meATP, CTP, ivermectin	TNP-ATP (weak), BBG (weak)	Intrinsic ion channel (especially Ca^{2+})
$P2X_5$	Proliferating cells in skin, gut, bladder, thymus, spinal cord	ATP ≫ α,β-meATP, ATPγS	Suramin, PPADS, BBG	Intrinsic ion channel
$P2X_6$	CNS, motor neurons in spinal cord	(Does not function as homomultimer)		Intrinsic ion channel
$P2X_7$	Apoptotic cells in, for example, immune cells, pancreas, skin	BzATP > ATP ≥ 2-MeSATP ≫ α,β-meATP	KN62, KN04, MRS2427, Coomassie brilliant blue G	Intrinsic cation channel and a large pore with prolonged activation
P2Y				
$P2Y_1$	Epithelial and endothelial cells, platelets, immune cells, osteoclasts	2-MeSADP = ADPβS > 2-MeSATP = ADP > ATP, MRS2365	MRS2179, MRS2500, MRS2279, PIT	G_q/G_{11}; PLC-β activation
$P2Y_2$	Immune cells, epithelial and endothelial cells, kidney tubules, osteoblasts	UTP = ATP, UTPγS, INS37217	Suramin > RB2, AR-C126313	G_q/G_{11} and possibly G_i; PLC-β activation
$P2Y_4$	Endothelial cells	UTP ≥ ATP, UTPγS	RB2 > suramin	G_q/G_{11} and possibly G_i; PLC-β activation
$P2Y_6$	Some epithelial cells, placenta, T cells, thymus	UDP > UTP ≫ ATP, UDPβS	MRS2578	G_q/G_{11}; PLC-β activation
$P2Y_{11}$	Spleen, intestine, granulocytes	AR-C67085MX > BzATP ≥ ATPγS > ATP	Suramin > RB2, NF157, 5′-AMPS	G_q/G_{11} and G_S; PLC-β activation
$P2Y_{12}$	Platelets, glial cells	2-MeSADP ≥ ADP ≫ ATP	CT50547, AR-C69931MX, INS49266, AZD6140, PSB0413, ARL66096, 2-MeSAMP	$G_{i/o}$; inhibition of adenylate cyclase
$P2Y_{13}$	Spleen, brain, lymph nodes, bone marrow	ADP = 2-MeSADP ≫ ATP and 2-MeSATP	MRS2211, 2-MeSAMP	$G_{i/o}$
$P2Y_{14}$	Placenta, adipose tissue, stomach, intestine, discrete brain regions	UDP glucose = UDP-galactose		G_q/G_{11}

BBG, brilliant blue green; BzATP, 2′- and 3′-*O*-(4-benzoyl-benzoyl)-ATP; cAMP, cyclic AMP; CCPA, chlorocyclopentyl adenosine; CPA, cyclopentyl adenosine; CTP, cytosine triphosphate; IP_3, inosine triphosphate; IP_5I, diinosine pentaphosphate; 2-MeSADP, 2-methylthio-ADP; 2-MeSATP, 2-methylthio-ATP; NECA, 5'-*N*-ethylcarboxamido adenosine; NTS, nucleus tractus solitarius; PLC, phospholipase C; RB2, reactive blue 2.

Adapted and reproduced from Burnstock G (2003) Introduction: ATP and its metabolites as potent extracellular agonists. *Current Topics in Membranes* 54: 1–27, with permission from Elsevier.

Table 2 Potential coassembly of P2X receptor subunits[a]

	P2X$_1$	*P2X$_2$*	*P2X$_3$*	*P2X$_4$*	*P2X$_5$*	*P2X$_6$*	*P2X$_7$*
P2X$_1$	+	+	+		+	+	
P2X$_2$		+	+		+	+	
P2X$_3$			+		+		
P2X$_4$				+	+	+	
P2X$_5$					+	+	
P2X$_6$							
P2X$_7$							+

[a]P2X receptor subunits carrying either one of two epitope tag units were expressed in pairs of HEK293 cells. +, Subunits immunoprecipitated with antibody to one epitope could be detected with an antibody to the second epitope.
Reproduced from Torres GE, Egan TM, and Voigt MM (1999) Hetero-oligomeric assembly of P2X receptor subunits. Specificities exist with regard to possible partners. *Journal of Biological Chemistry* 274: 6653–6659, with permission from the American Society for Biochemistry and Molecular Biology.

Table 3 Chromosomal localization of human P2X receptors[a]

Subunit	*Chromosome*	*Accession number*
P2X$_1$	17p13.2	X83688
P2X$_2$		AF190826
P2X$_3$	11q12	Y07683
P2X$_4$	12q24.31	Y07684
P2X$_5$	17p13.3	AF016709
P2X$_6$	22q11	AB002059
P2X$_7$	12q24.31	Y09561

[a]Accession numbers are those for the original submission of cDNA sequences. Chromosomal localizations are from human genome databases. P2X$_2$ chromosomal location is not yet determined. The mouse gene is located on chromosome 5, in a region that is syntenic with the extreme end of the long arm of human chromosome 12 (some 6 MB from the P2X$_4$ and P2X$_7$ genes).
Reproduced from North RA (2002) Molecular physiology of P2X receptors. *Physiological Reviews* 82: 1013–1067, with permission from the American Physiological Society.

to TM1 and TM2 (see Figure 1(b)). The P2X$_{1-7}$ receptors show 30–50% sequence identity at the peptide level. The stoichiometry of P2X$_{1-7}$ receptors is thought to involve three subunits, which form a stretched trimer or a hexamer of conjoined trimers. All of the P2X receptor subunits have consensus sequences for N-linked glycosylation.

The pharmacology of the recombinant P2X receptor subtypes expressed in oocytes or other cell types is often different from the pharmacology of P2X-mediated responses in naturally occurring sites. Several contributing factors may account for these differences. First, heteromultimers as well as homomultimers are involved in forming the trimer ion pore. P2X$_{2/3}$, P2X$_{1/2}$, P2X$_{1/5}$, P2X$_{2/6}$, P2X$_{4/6}$, and P2X$_{1/4}$ receptor heteromultimers have been identified (Table 2). P2X$_7$ does not form heteromultimers, and P2X$_6$ will not form a functional homomultimer. Second, spliced variants of P2X receptor subtypes might play a part.

There are seven genes for P2X receptor subunits. P2X$_4$ and P2X$_7$ subunit genes are located close to the tip of the long arm of chromosome 12. P2X$_4$ and P2X$_7$ subunits are among the most closely related pairs in amino acid sequences. P2X$_1$ and P2X$_5$ genes are also very close together on the short arm of chromosome 13. The remaining genes are on different chromosomes (Table 3). The genes vary considerably in size (e.g., mP2X$_3$ = 40 kb; hP2X$_6$ = 12 kb). The full-length forms have 11–13 exons, and all share a common structure, with well-conserved intron/exon boundaries. Many spliced forms of the receptor subunits (or fragments) have been described. Several full-length nonmammalian vertebrate sequences have been identified. There are no reports of homologous sequences from invertebrate species, although there is considerable functional evidence that extracellular ATP and other nucleotides can directly gate ion channels in invertebrates.

Recent advances have been made by the preparation of knockout mice for P2X$_1$, P2X$_2$, P2X$_3$, P2X$_4$, and P2X$_7$ receptors, and transgenic mice that overexpress the P2X$_1$ receptor.

P2X Receptor Subtypes

P2X$_1$ receptors A cDNA encoding the P2X$_1$ receptor was isolated by direct expression in *Xenopus* oocytes, beginning with a cDNA library made from rat vas deferens. Human and mouse cDNAs have also been cloned and expressed. The homomeric P2X$_1$ receptor is a cation-selective channel that shows little selectivity for sodium over potassium. It has a relatively high permeability to calcium.

A major property of the P2X$_1$ receptor is the mimicry of the agonist actions of ATP by α,β-methylene ATP (α,β-meATP), which distinguishes P2X$_1$ and P2X$_3$ receptors from the other homomeric forms. 2′,3′-O-(benzoyl-4-benzoyl)-ATP (BzATP) is also an effective agonist. P2X$_1$ receptors are blocked by suramin and pyridoxal-phosphate-6-azophenyl-2′,4′-disulfonic acid (PPADS), but there are newer antagonists that are more P2X$_1$-selective (see Table 1). A valuable antagonist at P2X$_1$ receptors is 2′,3′-O-(2,4,6-trinitrophenyl)-ATP (TNP-ATP), which has an IC$_{50}$ of about 1 nM.

Desensitization means the decline in the current elicited by ATP during the continued presence of ATP. In some P2X receptors this decline occurs in milliseconds (fast desensitization: P2X$_1$, P2X$_3$), and in others it occurs 100–1000 times more slowly (slow desensitization: P2X$_2$, P2X$_4$). Recovery from desensitization is extremely slow. Treatment with

apyrase allows $P2X_1$ receptors to recover from desensitization. Adenoviral expression of a $P2X_1$ receptor–green fluorescent protein construct in vas deferens shows the receptor to be localized in clusters, with larger ones apposing nerve varicosities.

$P2X_2$ receptors The rat $P2X_2$ receptor cDNA was isolated from a library constructed from nerve growth factor (NGF)-differentiated PC12 cells by testing pools for functional expression in *Xenopus* oocytes. The human receptor cDNA was amplified from pituitary gland. There are no agonists currently known that are selective for $P2X_2$ receptors. However, $P2X_2$ receptors are potentiated by protons and by low concentrations of zinc and copper. There are no selective antagonists for $P2X_2$ receptors. The $P2X_2$ receptor is generally described as nondesensitizing, compared with the $P2X_1$ and $P2X_3$ receptors.

When oocytes are injected with RNAs encoding $P2X_2$ receptors, and also the $\alpha3$ and $\beta4$ subunits of nicotinic receptors, they show responses to both ATP and acetylcholine; these can be selectively antagonized with appropriate receptor blockers. However, with concomitant application of both agonists, the resultant current is less than the expected sum of the two independent currents, indicating an interaction between the two receptors.

Heteromeric $P2X_{1/2}$ receptors $P2X_1$ and $P2X_2$ receptor subunits have been co-expressed in defolliculated *Xenopus* oocytes and the resultant receptors were studied under voltage clamp conditions. Co-expression yielded a mixed population of homomeric and heteromeric receptors, with a subpopulation of novel pH-sensitive P2X receptors showing identifiably unique properties that indicate the formation of heteromeric $P2X_{1/2}$ ion channels. It has been claimed that trimeric $P2X_{1/2}$ receptors incorporate one $P2X_1$ and two $P2X_2$ subunits.

$P2X_3$ receptors $P2X_3$ receptor subunit cDNAs were isolated from rat dorsal root ganglion cDNA libraries, from a human heart cDNA library, and from a zebra fish library. The mimicry of ATP by α,β-meATP makes these receptors similar to $P2X_1$ and distinct from the other homomeric forms. 2-MethylthioATP is as potent as or more potent than ATP at $P2X_3$ receptors. The antagonists suramin, PPADS, and TNP-ATP do not readily distinguish between $P2X_1$ and $P2X_3$ receptors, but NF023 is about 20 times less effective at $P2X_3$ than at $P2X_1$ receptors. Similar to $P2X_1$ receptors, desensitization is fast and recovery is very slow. $P2X_3$ receptors are prominently expressed on nociceptive sensory neurons.

Heteromeric $P2X_{2/3}$ receptors Direct association between $P2X_2$ and $P2X_3$ receptor subunits has been shown by co-immunoprecipitation. $P2X_{2/3}$ heteromeric channels can be defined on the basis of a sustained current elicited by α,β-meATP. $P2X_{2/3}$ receptor channels and, like homomeric $P2X_2$ receptors, are potentiated by low pH, and do not desensitize rapidly. The $P2X_{2/3}$ heteromer, like the homomeric $P2X_3$ receptor, is blocked by TNP-ATP, as well as PPADS and suramin. IP_5I is much more potent for blocking $P2X_1$ and $P2X_3$ homomers than for blocking the $P2X_{2/3}$ heteromers and is therefore useful to distinguish between $P2X_3$ and $P2X_{2/3}$ receptors. $P2X_{2/3}$ receptors have been identified in subpopulations of sensory neurons, sympathetic ganglion cells, and brain neurons.

$P2X_4$ receptors cDNAs for the rat $P2X_4$ receptor were isolated independently from superior cervical ganglion, brain, hippocampus, and pancreatic islet cells. Human, mouse, chick, and *Xenopus* cDNAs have also been isolated. Homomeric $P2X_4$ receptors are activated by ATP, but not by α,β-meATP. The most useful distinguishing feature of ATP-evoked currents at $P2X_4$ receptors is their potentiation by ivermectin.

When the application of ATP is of short duration, $P2X_4$ receptors operate as cation-selective channels; the calcium permeability is relatively high. When the application of ATP is continued for several seconds, the $P2X_4$ receptor channel becomes increasingly permeable to larger organic cations such as *N*-methyl-D-glucamine (NMDG). Desensitization at $P2X_4$ receptors is intermediate between that observed at $P2X_1$ and $P2X_2$.

The rat $P2X_4$ receptor is unusual among the P2X receptors in its relative insensitivity to blockade by the conventional antagonists suramin and PPADS. Currents evoked by ATP at the mouse $P2X_4$ receptor are actually increased by PPADS and suramin, probably because of their ectonucleotidase inhibitory activity.

Heteromeric $P2X_{1/4}$ receptors Co-injection of $P2X_1$ and $P2X_4$ subunits into *Xenopus* oocytes showed that both subunits were present in trimeric complexes of the same size. Voltage clamp experiments revealed functional P2X receptors with kinetic properties resembling those of homomeric $P2X_4$ receptors and a pharmacological profile similar to that of homomeric $P2X_1$ receptors. Preliminary results show that the $P2X_1$ receptor from the vas deferens and the $P2X_4$ receptor from salivary gland form complexes of the same size as the recombinant trimeric complexes expressed in oocytes.

$P2X_5$ receptors The $P2X_5$ receptor cDNA was first isolated from cDNA libraries constructed from rat celiac ganglion and heart. A P2X receptor was also cloned from embryonic chick skeletal muscle. The only human cDNAs reported are missing exon 10 ($hP2X_{5a}$) or exons 3 and 10 ($hP2X_{5b}$).

A feature of the currents elicited by ATP in cells expressing the rat $P2X_5$ receptor is their small amplitude, compared with the currents observed with $P2X_1$, $P2X_2$, $P2X_3$, or $P2X_4$ receptors expressed under similar conditions. The currents otherwise resemble those seen at $P2X_2$ receptors: they show little desensitization, are not activated by α,β-meATP, and are blocked by suramin and PPADS. $P2X_5$ mRNA is highly expressed in developing skeletal muscle.

Heteromeric $P2X_{1/5}$ receptors $P2X_1$ and $P2X_5$ subunits can be co-immunoprecipitated and the defining phenotype of this heteromer is a sustained current evoked by α,β-meATP, which is not seen for either of the homomers when expressed separately. Cells expressing the heteromeric receptor are very sensitive to ATP, concentrations as low as 3 or 10 nM evoking measurable currents. The sensitivity to the antagonist TNP-ATP is intermediate between the sensitive homomeric $P2X_1$ receptor and the insensitive homomeric $P2X_5$ receptor.

$P2X_6$ receptors The rat $P2X_6$ receptor was cloned from superior cervical ganglion cDNA and from rat brain. The human equivalent was isolated from peripheral lymphocytes as a p53-inducible gene. This was originally designated P2XM to reflect its abundance in human and mouse skeletal muscle. The $P2X_6$ receptor appears to be a 'silent' subunit, in the sense that no currents are evoked by ATP when it is expressed as a homomultimer in oocytes or HEK293 cells. It appears that the $P2X_6$ subunit is only functionally expressed as a heteromultimer.

Heteromeric $P2X_{2/6}$ receptors $P2X_2$ and $P2X_6$ receptors have been found to co-immunoprecipitate after expression in HEK293 cells. Oocytes expressing this combination have subtly different responses to ATP as compared to oocytes expressing only $P2X_2$ receptors. The most convincing of these differences is the fact that at pH 6.5 the inhibition of the current by suramin is clearly biphasic; one component has the high sensitivity of homomeric $P2X_2$ receptors, whereas the other component is less sensitive. $P2X_{2/6}$ receptors are prominently expressed by respiratory neurons in the brain stem.

Heteromeric $P2X_{4/6}$ receptors $P2X_4$ and $P2X_6$ receptors form a heteromeric channel when co-expressed in oocytes. The subunits can be co-immunoprecipitated from oocytes and HEK293 cells. The principal functional evidence for co-expression is that currents elicited by ATP are larger in oocytes 5 days after injection of mRNAs for $P2X_4$ and $P2X_6$ than after injection of $P2X_4$ alone. However, the phenotype of the heteromer differs only in minor respects from that of $P2X_4$ homomers. $P2X_{4/6}$ receptors are prominent in adult trigeminal mesencephalic nucleus and in hippocampal CA1 neurons.

$P2X_7$ receptors A chimeric cDNA encoding the rat $P2X_7$ receptor was first constructed from overlapping fragments isolated from superior cervical ganglion and medial habenula; full-length cDNAs were subsequently isolated from a rat brain cDNA library. Human and mouse cDNAs were cloned from monocyte and microglial cells, respectively. The main feature of the $P2X_7$ receptor is that in addition to the usual rapid opening of the cation-selective ion channel, with prolonged exposure to high concentrations of ATP it becomes permeable to larger cations (e.g., NMDG) and later undergoes a channel-to-pore conversion to allow the passage of large dye molecules such as ethidium and YO-PRO-1, and this usually leads to cell death. Evidence for $P2X_7$ receptor activation includes inward currents and increase in $[Ca^{2+}]_i$; other end points involve uptake of YO-PRO-1 or similar fluorescent dyes which bind to nucleic acid, and structural changes in the cell, such as membrane blebbing.

BzATP is a potent agonist at the $P2X_7$ receptor. There are five main types of blockers (see Table 1): ions (calcium, magnesium, zinc, copper, and protons), the suramin analog NF279, Coomassie brilliant blue G (which is most effective at rat $P2X_7$ receptors), oxidized ATP, and KN62, which is selective for the human $P2X_7$ receptor.

ATP or BzATP induces remarkable changes in the appearance of HEK293 cells transfected with the rat $P2X_7$ receptor. After continuous application of BzATP (30 μM) for about 30 s, the plasma membrane begins to develop large blebs, and after 1 or 2 min, these become multiple and sometimes coalesce. Blebs are usually preceded by the appearance of smaller vesicles (<1 μm in diameter), which are shed from the cell and appear to release inflammatory cytokines.

P2Y Receptors

Molecular Structure

The first P2Y receptors were cloned in 1993. Since then several other subtypes have been isolated by homology cloning and are assigned a subscript on the basis of cloning chronology ($P2Y_4$, $P2Y_6$, $P2Y_{11}$). The long-awaited G_i-coupled ADP receptor ($P2Y_{12}$) of platelets was finally isolated by expression

cloning in 2001, while $P2Y_{13}$ and $P2Y_{14}$ receptors were characterized later during a systematic study of orphan receptors. At present, there are eight accepted human P2Y receptors: $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, $P2Y_{11}$, $P2Y_{12}$, $P2Y_{13}$, and $P2Y_{14}$ (see Table 1). The missing numbers represent either nonmammalian orthologs, or receptors having some sequence homology to P2Y receptors, but for which there is no functional evidence of responsiveness to nucleotides. p2y3 may be a chicken ortholog of $P2Y_6$, while p2y8 and tp2y could be the *Xenopus* and turkey orthologs of $P2Y_4$, respectively. p2y7 is a leukotriene B4 receptor. p2y5 and p2y10 are considered as orphan receptors. A p2y8 receptor cloned from the frog embryo appears to be involved in the development of the neural plate. p2y9 was reported to be a novel receptor for lysophosphatidic acid, distant from the Edg family. $P2Y_{15}$ was recently introduced to designate the orphan receptor GPR80/GPR99 on the basis that it would be a receptor for adenosine 5′-monophospahte (AMP), but it is now firmly established that it is not a P2Y receptor, but rather a receptor for α-ketoglutarate, as recently underlined in a report by the International Union of Basic and Clinical Pharmacology (IUPHAR) P2Y Subcommittee.

In contrast to P2X receptors, P2Y receptor genes do not contain introns in the coding sequence, except for the $P2Y_{11}$ receptor. Site-directed mutagenesis of the $P2Y_1$ and $P2Y_2$ receptors has shown that some positively charged residues in TM3, TM6, and TM7 (see Figure 1(c)) are crucial for receptor activation by nucleotides. From a phylogenetic and structural (i.e., protein sequence) point of view, two distinct P2Y receptor subgroups characterized by a relatively high level of sequence divergence have been identified. The first subgroup includes $P2Y_{1,2,4,6,11}$ subtypes and the second subgroup encompasses the $P2Y_{12,13,14}$ subtypes (see dendrogram in Figure 2).

Most of the P2Y receptor subtypes are still lacking potent and selective synthetic agonists and antagonists. However, ADPβS is a potent agonist of $P2Y_1$, $P2Y_{12}$, and $P2Y_{13}$ receptors. 2-Methylthio-ADP (2-MeSADP) is a potent agonist (EC_{50}, in nM) at $P2Y_1$, $P2Y_{12}$, and $P2Y_{13}$ receptors. Selective antagonists have been identified for some P2Y receptor subtypes (see Table 1). P2Y receptor-mediated responses occur in nonneuronal and nonmuscular cell types, as well as in the nervous system, involved in both short- and long-term signaling.

Second Messenger Systems and Ion Channels

$P2Y_1$, $P2Y_2$, $P2Y_4$, and $P2Y_6$ receptors couple to G-proteins to increase inositol trisphosphate (IP_3) and cytosolic calcium. Activation of the $P2Y_{11}$ receptor by ATP leads to a rise in both cAMP and in IP_3, whereas activation by uridine 5′-triphosphate (UTP) produces calcium mobilization without IP_3 or cAMP increase. The $P2Y_{13}$ receptor can simultaneously couple to G_{16}, G_i, and, at high concentrations of ADP, G_s. The activation of several P2Y receptors is commonly associated with the stimulation of several mitogen-activated protein (MAP) kinases, in particular extracellular signal-regulated protein kinase-1/2.

In recent years, GPCRs in neurons and other excitable cells have been found to modulate the activity of voltage-gated ion channels in the cell membrane through certain actions of activated G-proteins. Such actions are now well established in closing (or, in certain cases, in opening or potentiating) various classes of K^+ channels and voltage-gated Ca^{2+} channels. ATP (or UTP, or their products ADP or UDP) present at synapses, plus ATP diffusing from astrocytes, activates P2Y receptors on distinct subsets of brain neurons, regulating their activities by the coupling of those receptors to specific ion channels. While ion channel couplings of P2Y receptors are primarily of importance in neurons, they have in a few cases been detected also in various other tissues (e.g., in cardiac muscle cells). Among the channels with which the superior cervical ganglion cell membrane is well endowed are two types of voltage-gated channels, which are important in receptor-based regulation of neuronal activity, the Ca^{2+} channel of the N-type and the M-current K^+ channel.

P2Y Receptor Subtypes

$P2Y_1$ receptors Human, rat, mouse, cow, chick, turkey, and *Xenopus* $P2Y_1$ receptors have been cloned and characterized. In most species, ADP is a more potent agonist than ATP is and their 2-methylthio derivatives are even more potent. UTP, UDP, CTP, and GTP are inactive. At present, the most potent and selective agonist known is the *N*-methanocarba analog of 2-MeSADP, MRS2365 (EC_{50} of 0.4 nM). The most effective antagonists to display selectivity for the $P2Y_1$ receptor are MRS2179, MRS2279, and MRS2500 (see Table 1).

Site-directed mutagenesis studies on the human $P2Y_1$ receptor have shown that amino acid residues in TM3, TM6, and TM7 are critical determinants in the binding of ATP. Four cysteine residues in the extracellular loops, which are conserved in P2Y receptors, are essential for proper trafficking of the human $P2Y_1$ receptor to the cell surface. $P2Y_1$ mRNA expression is highest in various regions of the brain, prostate gland, and placenta, and has also been detected at varying levels in other organs.

$P2Y_1$ receptor knockout mice have been generated. These mice are viable with no apparent abnormalities affecting their development, survival, and reproduction. Platelet counts are normal, but shape change is abolished. Transgenic mice overexpressing the $P2Y_1$

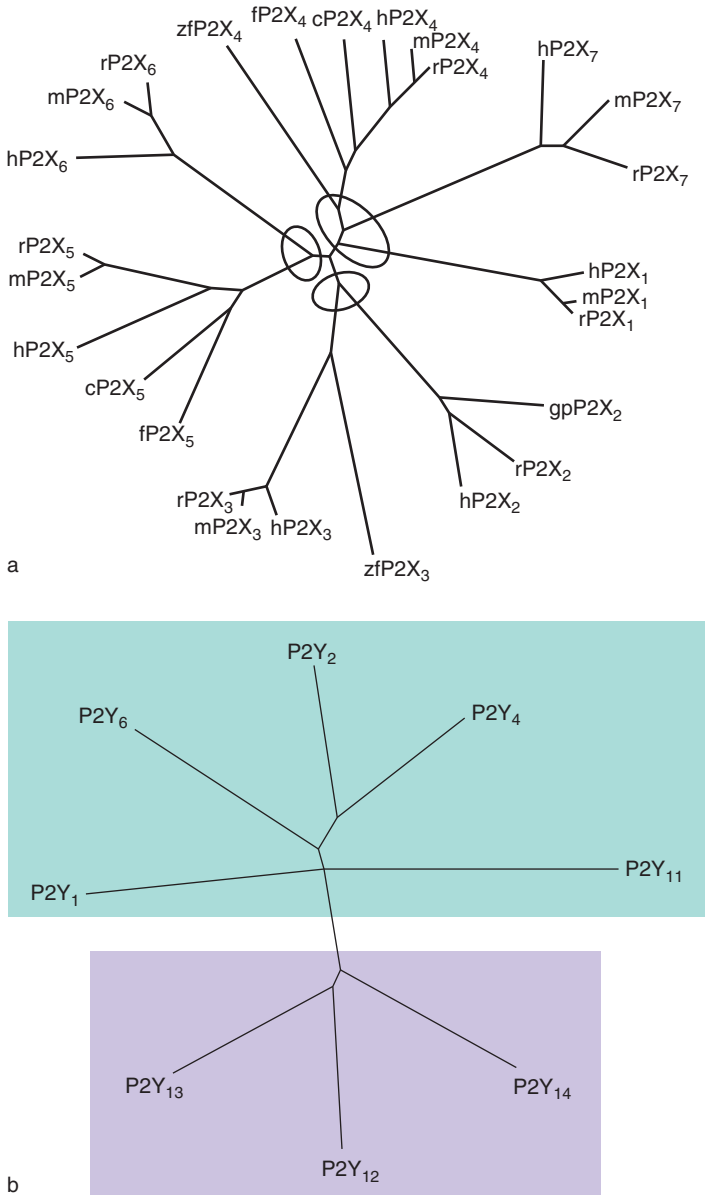

Figure 2 (a) Dendrogram to show relatedness of 29 P2X receptor subunits. Full-length amino acid sequences were aligned with Clustal W software using default parameters. The dendrogram was constructed with Tree View. h, human (*Homo sapiens*); r, rat (*Rattus norvegicus*); m, mouse (*Mus musculus*); gp, guinea pig (*Cavia porcellus*); c, chicken (*Gallus gallus*); zf, zebra fish (*Danio rerio*); f, fugu (*Takifugu rubripes*). The ovals indicate the apparent clustering by relatedness into subfamilies. (b) A phylogenetic tree (dendrogram) showing the relationships among the current members of the P2Y receptor family (human $P2Y_1$, $P2Y_2$, $P2Y_4$, $P2Y_6$, $P2Y_{11}$, $P2Y_{12}$, and $P2Y_{13}$ receptors) and the human UDP-glucose receptor (here indicated as the $P2Y_{14}$ receptor). The P2Y receptors can be divided into two subgroups, shown with blue and purple backgrounds. Sequences were aligned using Clustal X and the tree was built using the Tree View software. (a) Reproduced from North RA (2002) Molecular physiology of P2X receptors. *Physiological Reviews* 82: 1013–1067, with permission from the American Physiological Society. (b) Reproduced from Abbracchio MP, Boeynaems J-M, Barnard EA, et al. (2003) Characterization of the UDP-glucose receptor (re-named here the $P2Y_{14}$ receptor) adds diversity to the P2Y receptor family. *Trends in Pharmacological Sciences* 24: 52–55, with permission from Elsevier.

receptor specifically in the megakaryocytic/platelet lineage have also been generated.

$P2Y_2$ receptors $P2Y_2$ receptors have been cloned and pharmacologically characterized from human, rat, mouse, canine, and porcine cells or tissues. $P2Y_2$ receptors are fully activated by ATP and UTP, whereas ADP and UDP are much less effective agonists. The γ-thiophosphate, UTPγS, has been shown to be a potent hydrolysis-resistant agonist of $P2Y_2$

receptors, as is the recently developed $P2Y_2$ receptor agonist, INS37217 (Up_4dC). Suramin acts as a weak competitive antagonist of human and rat $P2Y_2$ receptors. AR-C126313 and the related aminotetrazole derivative AR-C118925, flavanoids, and tangeretin have been claimed recently to be effective antagonists.

$P2Y_2$ receptors can directly couple to $PLC\beta_1$ via $G_{\alpha q/11}$ protein to mediate the production of IP_3 and diacylglycerol, which are second messengers for calcium release from intracellular stores and protein kinase C (PKC) activation, respectively.

Expression of $P2Y_2$ receptor mRNA and protein has been detected in many tissues. $P2Y_2$ receptor activation increases the synthesis and/or release of arachidonic acid, prostaglandins, and nitric oxide. $P2Y_2$ receptor expression in smooth muscle cells is upregulated by agents that mediate inflammation. $P2Y_2$ receptors have been shown to play a role in the wound healing process. $P2Y_2$ receptor activation increases Cl^- secretion and inhibits Na^+ absorption in epithelial cells. A $P2Y_2$ receptor knockout mouse has been produced that is defective in nucleotide-stimulated ion secretion in airway epithelial cells. $P2Y_2$ receptors have been shown to inhibit bone formation by osteoblasts, and N-type calcium currents in neurons.

$P2Y_4$ receptors Human, rat, and mouse $P2Y_4$ receptors have been cloned and characterized. UTP is the most potent activator of the recombinant human $P2Y_4$ receptor. In contrast, the recombinant rat and mouse $P2Y_4$ receptors are activated equipotently by ATP and UTP. Up_4U (INS365) and dCp_4U (INS37217) are agonists of the human $P2Y_4$ receptor. Reactive blue 2 effectively blocks rat $P2Y_4$ receptors, but only partially blocks human $P2Y_4$ receptors. Suramin is a weak antagonist at the $P2Y_4$ receptor. The structural determinants of agonism versus antagonism by ATP are located in the N-terminal domain and the second extracellular loop.

In the human and the mouse, $P2Y_4$ mRNA and protein was most abundant in the intestine, but was also detected in other organs. $P2Y_4$-null mice have apparently normal behavior, growth, and reproduction, but the chloride secretory response of the jejunal epithelium to apical UTP and ATP is abolished.

$P2Y_6$ receptors The mouse, rat, and human $P2Y_6$ receptors are UDP receptors. The uridine β-thiodiphosphate, $UDP\beta S$, and Up_3U are selective agonists of the $P2Y_6$ receptor and are more stable to degradation. INS48823 is also a potent $P2Y_6$ agonist. A 1,4-di-(phenylthioureido)butane derivative (MRS2578) has recently been shown to selectively inhibit UDP-induced PLC activity. A unique feature of the $P2Y_6$ receptor is its slow desensitization and internalization. A wide tissue distribution of $P2Y_6$ mRNA and protein has been demonstrated, with the highest expression in spleen, intestine, liver, brain, and pituitary.

$P2Y_{11}$ receptors Among P2Y receptors, the human $P2Y_{11}$ receptor has a unique profile. Its gene is the only P2Y receptor gene that contains an intron in the coding sequence. The potency of its natural agonist ATP is relatively low and it is dually coupled to PLC and adenylyl cyclase upon stimulation. $ATP\gamma S$ is a more potent agonist than ATP is. The $P2Y_{12}$ antagonist AR-C67085MX acts as a potent agonist at the $P2Y_{11}$ receptor. Suramin behaves as a competitive antagonist of the $hP2Y_{11}$ receptor.

$P2Y_{12}$ receptors The human, rat, and mouse $P2Y_{12}$ receptors have been identified and characterized. ADP is the natural agonist of this receptor. The $P2Y_{12}$ receptor is heavily expressed in platelets, where it is the molecular target of the active metabolite of the antiplatelet drug clopidogrel. Potent direct competitive $P2Y_{12}$ antagonists are also available, including the 5′-triphosphate derivative AR-C69931MX compound, named cangrelor. The $P2Y_{12}$ receptor has also been shown to be expressed in subregions of the brain, glial cells, brain capillary endothelial cells, smooth muscle cells, and chromaffin cells. The $P2Y_{12}$ knockout mice that have been generated display the phenotype of clopidogrel-treated animals.

$P2Y_{13}$ receptors The human, mouse, and rat $P2Y_{13}$ receptors have been identified and characterized. ADP and Ap_3A are naturally occurring agonists of the $P2Y_{13}$ receptor. The $P2Y_{13}$ receptor is primarily coupled to a $G_{i/o}$ protein. However, cangrelor, which is an antagonist of the $hP2Y_{12}$ receptor, is also an antagonist of the human and rat $P2Y_{13}$ receptors. Recently MRS2211, a derivative of PPADS, was shown to selectively antagonize the human $P2Y_{13}$ receptor. The $P2Y_{13}$ receptor is strongly expressed in the spleen, followed by placenta, liver, heart, bone marrow, monocytes, T cells, lung, and various brain regions. $P2Y_{13}$-null mice have been generated recently, but no phenotype has been characterized to date.

$P2Y_{14}$ receptors From a phylogenetic and structural point of view, the $P2Y_{14}$ receptor (previously known as GPR105 or UDP-glucose receptor) is 47% identical to the $P2Y_{12}$ and $P2Y_{13}$ receptors. The gene for this receptor has been found in human chromosome 3q24-3q25, where a cluster of other related GPCRs, consisting of $P2Y_1$, $P2Y_{12}$, and $P2Y_{13}$ receptors and the orphan receptors GPR87, GPR91, and H963, have been found. The $P2Y_{14}$ receptor couples to the $G_{i/o}$ family of G-proteins and is activated by

UDP-glucose as well as UDP-galactose, UDP-glucuronic acid, and UDP-*N*-acetylglucosamine. At present, no selective antagonists are available. $P2Y_{14}$ mRNA is widely distributed in the human body. Both chemoattractant and neuroimmune functions have been claimed for the $P2Y_{14}$ receptor.

Receptor Dimerization and Cross-Talk

It is now recognized that interactions between GPCRs can take place through the formation of oligomers, or downstream of the receptor through the action of second messengers. The former process is commonly referred to as receptor dimerization. The latter process is known as receptor cross-talk.

There is evidence that the human $P2Y_2$ receptor forms homodimers. An example of dimerization involving P2Y receptors with non-P2Y receptors is rat $P2Y_1$ and adenosine A_1 receptors co-expressed in HEK293 cells. It has also been shown that the $P2Y_1$ and A_1 receptors are co-localized in neurons of the rat cortex, hippocampus, and cerebellum. The formation of oligomers by P2Y receptors is likely to be widespread and to greatly increase the diversity of purinergic signaling. P2X receptors and P2Y receptors are often expressed in the same cells. Thus, there is the possibility of bidirectional cross-talk between these two families of nucleotide-sensitive receptors. For example, the $P2X_1$ receptor may have a priming role in activation of $P2Y_1$ receptors during platelet stimulation.

See also: Adenosine; Adenosine Triphosphate (ATP); Adenosine Triphosphate (ATP) as a Neurotransmitter; Adenosine Receptor Mediated Functions; Calcium Waves: Purinergic Regulation; P2X Receptors; Pharmacology of Sleep: Adenosine; Purinergic Receptors.

Further Reading

Abbracchio MP, Boeynaems J-M, Barnard EA, et al. (2003) Characterization of the UDP-glucose receptor (re-named here the $P2Y_{14}$ receptor) adds diversity to the P2Y receptor family. *Trends in Pharmacological Sciences* 24: 52–55.

Abbracchio MP, Burnstock G, Boeynaems J-M, et al. (2006) International Union of Pharmacology. Update and subclassification of the P2Y G protein-coupled nucleotide receptors: From molecular mechanisms and pathophysiology to therapy. *Pharmacological Reviews* 58: 281–341.

Brake AJ, Wagenbach MJ, and Julius D (1994) New structural motif for ligand-gated ion channels defined by an ionotropic ATP receptor. *Nature* 371: 519–523.

Barnard EA, Burnstock G, and Webb TE (1994) G protein-coupled receptors for ATP and other nucleotides: A new receptor family. *Trends in Pharmacological Sciences* 15: 67–70.

Boeynaems J-M, Communi D, Gonzalez NS, et al. (2005) Overview of the P2 receptors. *Seminars in Thrombosis and Hemostasis* 31: 139–149.

Burnstock G (1978) A basis for distinguishing two types of purinergic receptor. In: Straub RW and Bolis L (eds.) *Cell Membrane Receptors for Drugs and Hormones: A Multidisciplinary Approach*, pp. 107–118. New York: Raven Press.

Burnstock G and Knight GE (2004) Cellular distribution and functions of P2 receptor subtypes in different systems. *International Review of Cytology* 240: 31–304.

Fredholm BB, Ijzerman AP, Jacobson KA, et al. (2001) International Union of Pharmacology. XXV: Nomenclature and classification of adenosine receptors. *Pharmacological Reviews* 53: 527–552.

Hollopeter G, Jantzen H-M, Vincent D, et al. (2001) Identification of the platelet ADP receptor targeted by antithrombotic drugs. *Nature* 409: 202–207.

Jacobson KA, Jarvis MF, and Williams M (2002) Purine and pyrimidine (P2) receptors as drug targets. *Journal of Medicinal Chemistry* 45: 4057–4093.

King BF and Burnstock G (2002) Purinergic receptors. In: Pangalos M and Davies C (eds.) *Understanding G Protein-Coupled Receptors and Their Role in the CNS*, pp. 422–438. Oxford: Oxford University Press.

Neary JT, Rathbone MP, Cattabeni F, et al. (1996) Trophic actions of extracellular nucleotides and nucleosides on glial and neuronal cells. *Trends in Neurosciences* 19: 13–18.

Nicke A, Baumert HG, Rettinger J, et al. (1998) $P2X_1$ and $P2X_3$ receptors form stable trimers: A novel structural motif of ligand-gated ion channels. *EMBO Journal* 17: 3016–3028.

North RA (2002) Molecular physiology of P2X receptors. *Physiological Reviews* 82: 1013–1067.

Ralevic V and Burnstock G (1998) Receptors for purines and pyrimidines. *Pharmacological Reviews* 50: 413–492.

Surprenant A, Rassendren F, Kawashima E, et al. (1996) The cytolytic P_{2Z} receptor for extracellular ATP identified as a P_{2X} receptor ($P2X_7$). *Science* 272: 735–738.

Torres GE, Egan TM, and Voigt MM (1999) Hetero-oligomeric assembly of P2X receptor subunits. Specificities exist with regard to possible partners. *Journal of Biological Chemistry* 274: 6653–6659.

Webb TE, Simon J, Krishek BJ, et al. (1993) Cloning and functional expression of a brain G-protein-coupled ATP receptor. *FEBS Letters* 324: 219–225.

Pursuit Eye Movements

U J Ilg, University of Tübingen, Tübingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

General Properties of the Pursuit System

Compared to a modern digital camera, the human eye has a surprisingly low number of pixels, only approximately 1 million axons of ganglion cells transmit the information from the retina to the brain. This low number of pixels follows from a restriction of the high spatial resolution to the foveal field of vision. We are able to see up to 60 lines per 1° of visual angle in the fovea, and this resolution drops massively with eccentricity. The size of our visual field is approximately 90° in temporal direction and, due to our nose, approximately 60° in nasal direction. The advantage of the reduced amount of transmitted data is bought by the necessity of a very precisely acting oculomotor apparatus. We need to direct our gaze toward an important detail in our environment which requires vision at very high spatial resolution. If this detail moves, there is need for an eye movement command that maintains the image of this detail on, or at least close to, the fovea. In addition, retinal image motion introduces substantial image blur; movement of only $1^\circ\,s^{-1}$ results in image blur equivalent to myopia of 2 diopters. To sum up, smooth pursuit eye movements (SPEMs) are necessary for two reasons: (1) they keep the image of an important moving object on or close to the fovea, and (2) they prevent motion blur of this object. SPEMs warrant the high fidelity of the visual input necessary for higher cognitive processing such as face recognition or the recognition of the emotional state of another person. A hallmark of SPEMs is the excellent adjustment of eye speed to target speed. This can be easily demonstrated. Just move a pencil in front of a friend's face and ask him or her to track the tip of the moving pencil. The eyes move as though they are physically connected to the pencil.

Traditionally, pursuit is modeled as a negative feedback loop minimizing retinal image motion. More precisely, position, velocity, and acceleration of the retinal image of the pursued target are determined and transformed with different weights into an eye movement command. With this in mind, we might ask whether fixation of a stationary target is equivalent to pursuit of a target moving at zero velocity. There are two observations that point out the difference between fixation and pursuit. First, only during the execution of pursuit does the connection between the motion processor and the oculomotor control run at a high sensitivity state. Imagine a small perturbation in target position over time, a sine wave at a periodicity of 100 ms. In the case of a stationary fixation target, no response in the oculomotor behavior of a subject can be elicited by this stimulus. In contrast, if the target is moving, a huge oculomotor response is elicited by exactly the same perturbation. The neuronal substrate of this sensitivity boost is most likely a subarea of the frontal eye field (FEF), the frontal pursuit area (FPA). Electrical microstimulation of this area in rhesus monkeys results in a boost of the motion sensitivity similar to the engagement of the pursuit system. Second, many single-unit studies in rhesus monkeys revealed that different neurons are activated during the execution of fixation or pursuit.

During testing, everybody can imagine a moving target. As easy this is, it is surprisingly impossible to generate SPEMs in this condition. Even highly trained subjects are able to generate only a sequence of small saccades. In other words, we cannot produce SPEMs voluntarily. This inability emphasizes the negative feedback loop that operates on retinal image motion signals during the execution of SPEMs. However, it must be added that we are able to track a nonvisual moving target. If we close our eyes and move our left index finger on our right forearm, we can track this somatosensory perception smoothly. Even if another person touches our forearm, we can track touch in total darkness. Finally, if a sound source moves in darkness, we can track it also. Although there are no retinal slip signals available in these conditions, these situations have in common that there is a moving target. In contrast, if the target is only imagined, we cannot move our eyes smoothly.

Independently of the nature of the tracked target, SPEMs can be separated into two different temporal phases: pursuit initiation (open-loop), followed by steady-state pursuit (closed-loop). In the first phase, the eyes are initially stationary and retinal image movement of the target is directly related to target movement in space. In the second phase, target movement in space has to be computed by adding retinal image motion, eye speed, and, if applicable, head and body velocity. These two phases of SPEMs also differ with respect to their sensitivity for moving distractors. During pursuit initiation, vector averaging occurs if two moving targets are presented. In other words, the initial direction of the eye movement is given by the vector average of the target and distractor trajectory. During steady-state pursuit, a winner-takes-all mechanism is active and prevents

the moving distractor from affecting the ongoing pursuit eye movements.

During the execution of steady-state pursuit, two problems related to our subjective motion perception exist which indicate that our motion perception is not exclusively based on the processing of retinal image motion. First, we perceive the target as moving, even in the absence of any retinal image motion signal, as a consequence of perfectly adjusted SPEMs. Second, we do not perceive the stationary background as moving despite its retinal image. Having said this, it must be noted that our perception of a stationary environment is not perfect. If you move a pencil in front of this text, the text will appear to move in opposite direction while you track the pencil. This apparent motion is called Filehne illusion (for the German physiologist Wilhelm Filehne). Note that you are not able to read the text in this tracking condition. If you actually read the text, your eye movements consist in a series of small rightward saccades with intermingled fixations and huge leftward saccades at the end of each line.

As already indicated, a substantial part of our present knowledge of SPEMs results from experiments using rhesus monkeys. It was shown that the SPEMs of monkeys are very similar to those of human subjects. In addition to the analysis of the executed eye movements, single-unit responses from various brain areas were recorded during the execution of pursuit. These responses were used to pinpoint the neuronal substrate of SPEMs.

Initiation of Pursuit Eye Movements

During steady-state pursuit, eye speed is usually very precisely matched to target speed. But how are SPEMs initialized? Due to the latency of the visual system, the moving target inevitably leaves the foveal field of vision before the eyes could possibly move. Figure 1 shows the stereotypical buildup and maintenance of SPEMs in a human subject. There is a presaccadic initiation during which the eyes start to move. However, due to the latency, the eyes are lagging the target. This position error is compensated by the initial saccade. After the end of this saccade, steady-state pursuit is executed. Having said this, it is clear that pursuit initiation has a shorter latency than saccades.

It was shown that the initial acceleration of the eye is proportional to target speed and inversely proportional to target eccentricity. However, eye velocity is clearly below target velocity just before the initial saccade, and it is quite precisely matched to target velocity immediately after the initial saccade. Note that due to the high speed during saccades, the retinal image velocity of the target during the saccade cannot be used to adjust postsaccadic eye velocity. This

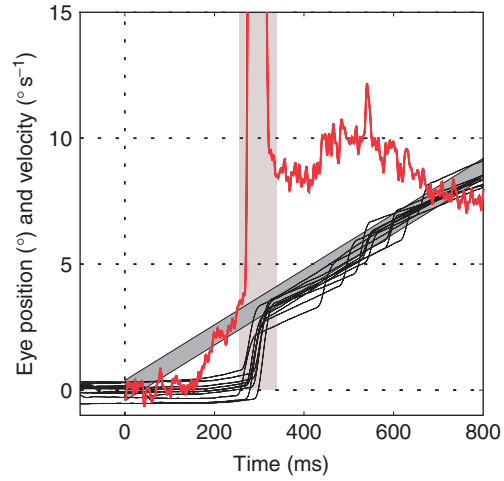

Figure 1 Typical smooth pursuit eye movements of a human subject. A subject was asked to track a moving (11° s^{-1}) target as precisely as possible. Target position is given by the oblique gray line, its width corresponding to the actual size of the target. Note that the eye-position traces of every single trial are very similar. The resulting median eye velocity is given by the red profile. The vertical light gray bar gives the time interval in which all initial saccades occurred.

boost in eye velocity before and after the initial saccade was named postsaccadic enhancement. At the onset of the saccade, the system has already decoded successfully target velocity and now uses this information to compute saccade amplitude adjusted to the target velocity. The saccade terminates exactly at the target's momentary position. These additional computations might explain why saccades directed to moving targets have longer latencies than saccades directed toward stationary targets.

Presaccadic acceleration of the eye can be an example of motor learning or adaptation. In the learning paradigm, the target starts to move at a given velocity and changes its velocity after approximately 100 ms (i.e., the pursuit latency). At the beginning, the initial acceleration of the eye is adjusted to the initial velocity of the target. Consequently, this maneuver produces a severe position error that needs to be corrected by an additional saccade. After close to 100 repetitions of this procedure, the initial acceleration is adjusted to the final target velocity and the huge position error is avoided.

The initiation phase of SPEMs is characterized by a need-for-speed behavior. It is most important to achieve a coarse estimate of target velocity as soon as possible. Because there is a direct relationship between time and position error, longer latencies yield bigger position errors. Due to this need-for-speed behavior, there are specific shortcomings during pursuit initiation. If subjects are asked to track

a second-order motion stimulus, the eyes initially follow the dot movement. In this experiment, let us say, the second-order motion stimulus consists of a rectangular area in which individual dots are moving to the left. The rectangle itself moves to the right. Later, during steady-state pursuit, the subjects' eyes follow the moving rectangle. Another shortcoming of pursuit initiation can be observed if subjects are asked to track a moving edge. The initial eye movements are always perpendicular to the edge, independent of the actual direction of edge movement. This misdirection of pursuit initiation can be explained by the aperture problem illustrated in Figure 2. Without visible line endings, we are able to perceive only motion perpendicular to a moving edge. After approximately 300 ms, these directional errors are corrected and pursuit at high fidelity is generated in the true direction of target movement. With respect to the moving edge experiment, it was nicely shown that the misdirection in eye movements is accompanied by an initial miscoding in the middle temporal area (MT), a bottleneck in the processing of visual motion. The initial error in the direction of eye movements is observed only if subjects track a real contour. If subjects are asked to track an illusionary contour, as shown in Figure 2, the initial misdirection perpendicular to the contour is absent. Nevertheless, we clearly perceive the illusionary contour. The pursuit onset latency is lengthened in this condition compared to the tracking of a real contour.

Although the execution of pursuit is tightly connected to the processing of retinal image motion, there is also a substantial influence of anticipation or prediction, especially during pursuit initiation. If the target moves consistently in the same direction at the same speed, human subjects and rhesus monkeys display pursuit eye movements due to anticipation, without any physical stimulus. This anticipatory pursuit is specific for direction and speed of the predicted target.

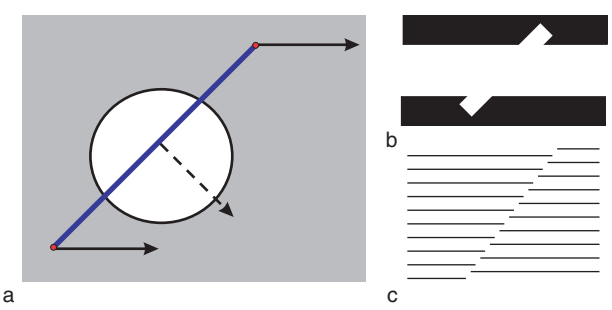

Figure 2 Aperture problem and illusionary contours: (a) the aperture problem; (b–c) two different types of illusory contours. In (a), imagine that the tilted blue bar moves to the right. Thanks to the line endings (marked by red circles), we perceive the horizontal movement as indicated by the black arrows. However, if our vision is restricted to the circular area, the bar seems to move 45° to the right and down, as indicated by the dashed arrow. In (b) and (c), the illusory contours do not produce a directional error during pursuit initiation.

In addition to the benefit of anticipation, there is another possibility for reducing pursuit latencies, similar to the way saccadic latencies are reduced. Initially, subjects have to fixate a stationary target. If this fixation target disappears 200 ms before the pursuit target appears (gap paradigm), pursuit onset latency is reduced. This reduction can be explained by the presence of a fixation system, which might consist of the rostral part of the superior colliculus and cortical areas such as the parietal cortex and the FEF. In case of the gap paradigm, this fixation system is switched off as a consequence of the disappearance of the target. In case of the regular paradigm, the fixation system has to be actively switched off to allow the execution of pursuit or saccade, respectively.

Brief Insights into the Anatomy of the Pursuit System

Because SPEMs can be executed only in the presence of a moving target, there is an ultimate connection between motion processing and the SPEM system. As a consequence, many properties of motion perception are in common with properties of the pursuit system. Various experiments revealed very similar characteristics of oculomotoric and psychometric functions. For instance, expectance influences eye movement control and motion perception in the same way.

The neuronal substrate of SPEMs starts with the processing within the retina and ends with the generation of an eye movement command by the extraocular motor neurons as shown in Figure 3. There is general agreement that motion processing in primates is a cortical function. Primates do not have directionally selective retinal ganglion cells like rabbits, for instance. Directionally selective neurons are primarily found in the hierarchy of visual processing in the primary visual cortex (V1) at an abundance between 10% and 20% of all neurons. There is experimental evidence that these directionally selective neurons exclusively provide direct input to the MT in the superior temporal sulcus (STS). In the MT, almost all neurons respond directionally selectively to visual motion. In turn, the MT provides the input of middle superior temporal area (MST) in the same sulcus. MST consists of two parts: a dorsal (MSTd) and lateral (MSTl) subdivision. Object movement is decoded in MSTl, whereas ego motion is decoded in MSTd. The receptive field size increases parallel to

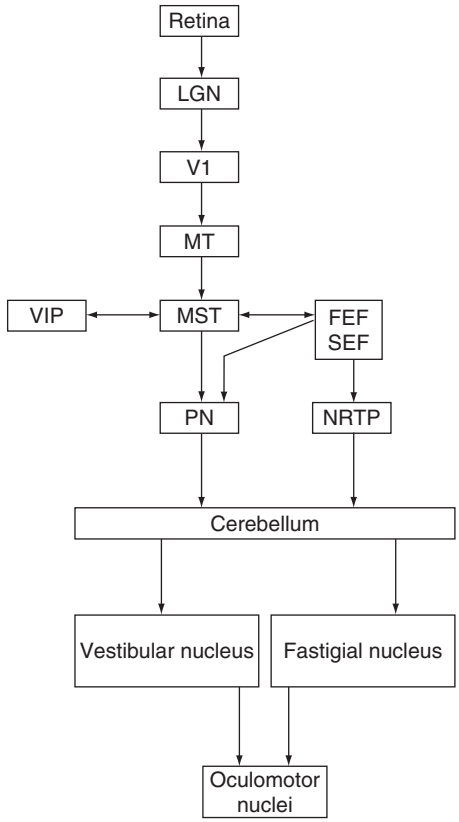

Figure 3 Schematic presentation of the pathways for pursuit. FEF, frontal eye field; LGN, lateral geniculate nucleus; MST, middle superior temporal area; MT, middle temporal area; NRTP, nucleus reticularis tegmentis ponti; PN, pontine nuclei; SEF, supplementary eye field; V1, primary visual cortex; VIP, ventral intraparietal area.

the cortical processing from the V1 via the MT to the MST. At all stages, neurons with similar response properties are clustered together, and there is a columnar organization perpendicular to the surface of these areas. With respect to the generation of SPEMs, the cortical processing seems to be finished at the level of the MST because a massive projection from the MST to the pontine nuclei (PN) exists. It was earlier thought that this cortical processing is confined to the processing of retinal image motion. Position, velocity, and acceleration of the pursuit target is extracted from the retinal image. Extra-retinal signals are thought to be added in subsequent subcortical stages, mainly at the level of the cerebellum. However, the cortical processing is not at all confined to retinal signals, there is now evidence that cortical neurons combine retinal image motion and eye- and head-movement-related signals to compute target motion in space. Most likely, this information is sent to the brain stem. The PN provide an input to the cerebellum. In turn, the output of the cerebellum is directed to the vestibular and fastigial nuclei, whose neurons are presynaptic to the extra-ocular motor neurons. This cortico-ponto-cerbellar system is the direct path underlying the execution of SPEMs.

In parallel to this direct path, the neuronal substrate of the optokinetic system also contributes to the generation of SPEMs. This substrate consists of the accessory optic system (AOS), the nucleus of the optic tract (NOT) located in the pretectal area, and its projection to the inferior olive, which provides the climbing fiber input to the cerebellum. The AOS in primates also receives its directionally selective input from motion-sensitive areas in STS, different from the AOS of other vertebrates which receives directionally selective input from the retinal ganglion cells.

The function of the FEF and its subarea FPA, as well as the supplementary eye field (SEF), in the generation of SPEMs is not totally clear at present time. Some neurons in these areas respond during the execution of SPEMs. Microstimulation at specific sites in the FPA elicits pursuitlike eye movements. More recently, it was suggested that the main function of FPA might be the boost of motion processing during SPEMs. With respect to the importance of the SEF, it was proposed that neurons from this area provide an anticipatory signal related to the execution of predictive pursuit. The frontals areas project to the nucleus reticularis tegmentis ponti (NRTP), which provides another input to the cerebellum.

Finally, it was shown that some neurons in the ventral intraparietal area (VIP) respond directionally selectively during the execution of SPEMs. Because VIP contains a representation of near space, these neurons display a preference for targets moving at high velocity.

How to Differentiate between Retinal and Extra-Retinal Signals

The sensory–motor borderline underlying SPEMs is still a matter of debate. Because it is impossible to execute pursuit without a moving target, there is no easy way to disentangle sensory and motor processing. So far, two attempts have been made to solve this problem. The first attempt consists of a brief temporal removal of pursuit target for on the order of 100–200 ms. Human subjects reported that these brief disappearances of the target yield the perception that the target has briefly moved into the screen, like a toy train moves into a tunnel. Contrary to the assumption that pursuit would continue, eye speed was substantially decreased, even during short tunnel intervals. If the subjects received feedback about the velocity of the executed eye movement (for instance, by a change in the pitch of a sound), the drop in eye speed was less pronounced. A second attempt to disentangle visual and extra-retinal signals consisted of the use of

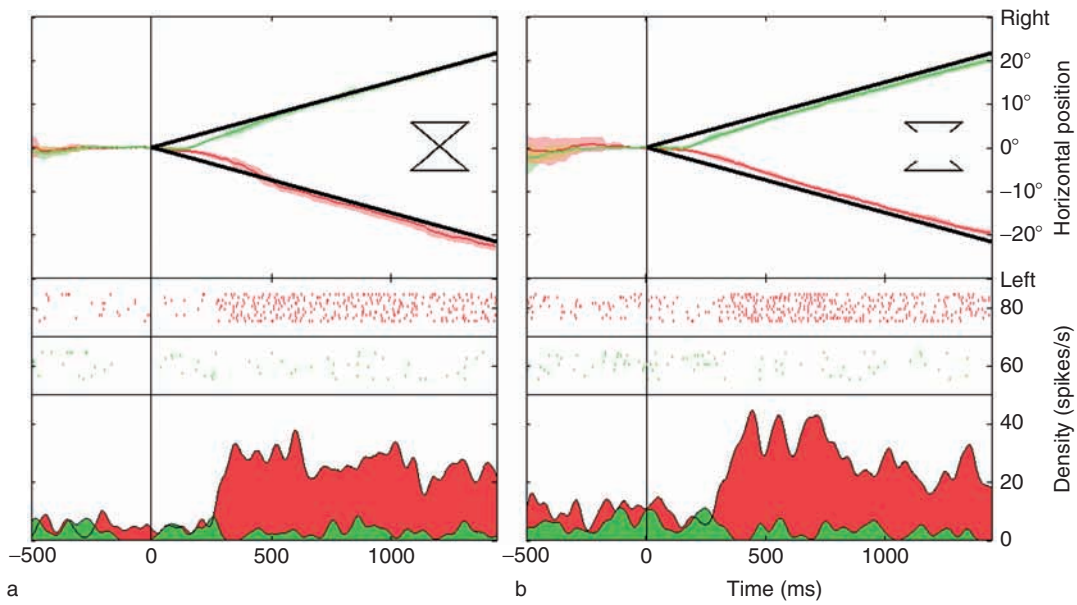

Figure 4 Middle superior temporal area (MST) activity in a monkey during the tracking of a target: (a) real target; (b) imaginary target. A monkey was trained to track a moving at 15° s^{-1} in the preferred (red) and nonpreferred (green) direction of the recorded neuron. The activity is shown as raster display and as spike density functions. Mean eye position profiles are displayed together with the resulting standard deviation (ten trials each). Note that the eye movements are very similar for both targets, as are the neuronal responses. The experiment was controlled so the receptive field of this neuron was not stimulated by parts of the imaginary figure; thus, the response in this condition can only be explained by extra-retinal signals related to the executed eye movements.

an imaginary figure, as shown in Figure 4(b). Here, the pursuit target was removed not temporally but spatially. During pursuit of such an imaginary target, no visual stimulation occurred in the part of visual field in which the receptive field of the neuron under study is located.

The results obtained from both attempts are consistent – MT turns out to process exclusively visual signals related to retinal image motion. If the pursuit target is either temporally or spatially removed, MT neurons stop responding. In contrast, MST neurons continue to respond in the absence of stimulation while the monkeys execute SPEMs. In addition, it was shown that MST neurons receive vestibular information related to head movements. Based on these results, a hypothesis was developed that MST neurons combine retinal image motion with extra-retinal signals, such as eye and head movement signals, to generate a signal related to the target movement in space.

The signal of the head velocity originates from the vestibular organ, but what is the origin of the extra-retinal signal related to the ongoing eye movement? This question was answered by experiments in which monkeys were trained to track a moving target either by isolated eye or by combined eye and head movements, as shown in Figure 5.

Because this neuron responds very similarly during both types of pursuit, we can dismiss the hypothesis that the extra-retinal signal is a copy of the motor command from the extra-ocular motor neurons; this command is quite different in both tracking conditions. Based on the result of this experiment and others consisting of a combination of a tracking task and vestibular stimulation, it was concluded that the activity of these neurons in the MST represents the trajectory of a moving target in space.

An important benefit of this finding is that the same representation of target movement can explain action (i.e., the generation of SPEMs) and motion perception. As mentioned earlier, motion perception and execution of pursuit have many properties in common. In addition, the signal related to target movement in space is also important to achieving the perception of a stationary environment despite self-induced retinal image motion.

Finally, the MST is not a motion processor exclusively for the generation of SPEMs. In several studies, it was shown that the neuronal activity in the MST is closely related to the animal's perception of motion. In addition, there is experimental evidence that the MST is also involved in the generation of goal-directed hand movements directed toward a moving visual target. Microstimulation and transient inactivation of this area yield modifications of SPEMs as well as modifications of goal-directed hand movements.

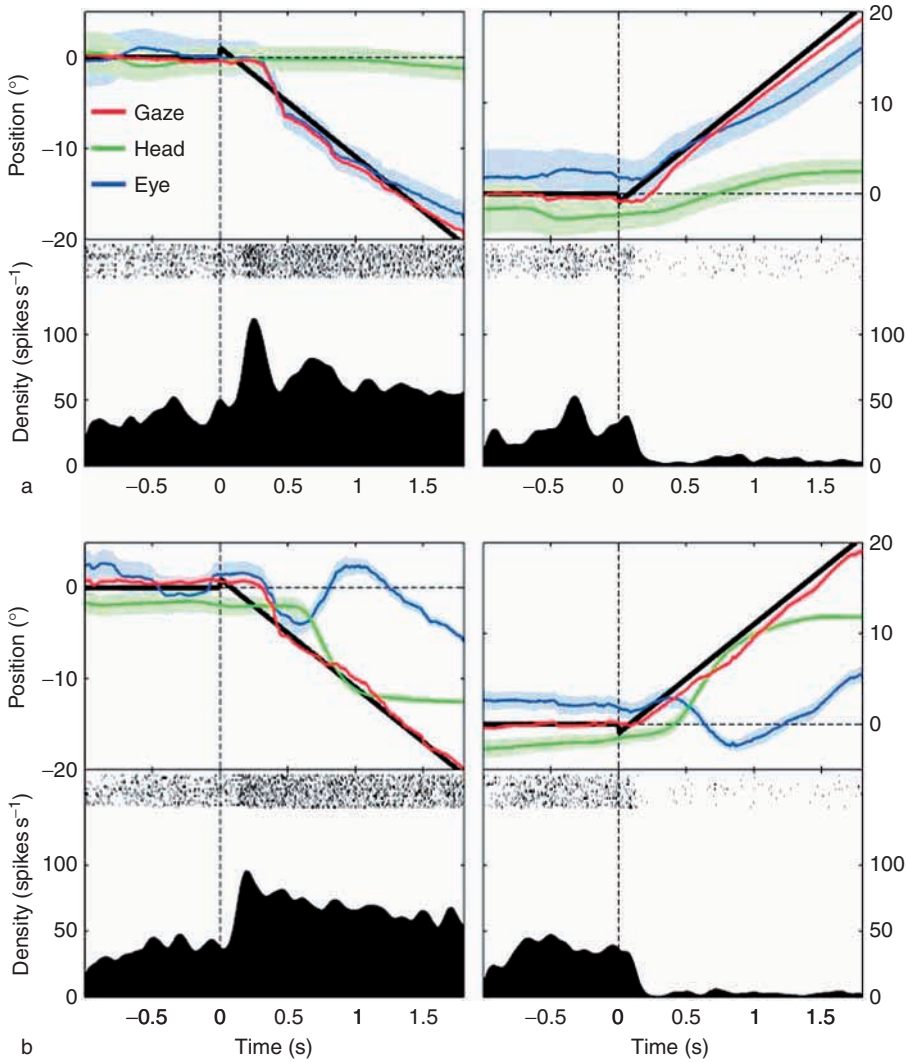

Figure 5 Middle superior temporal area (MST) activity in a monkey: (a) during eye movements; (b) during combined eye–head movements. The left panel represents preferred direction, right panel gives nonpreferred direction. A monkey was trained to track a moving target (10° s^{-1}) with either only eye movements or with combined eye and head movements. The neuronal activity is shown as raster displays and spike-density functions. Although the eye-in-head movements were very different in both conditions, the neuronal responses were very similar. This indicates that the extra-retinal component is not related to the eye-in-head movement (i.e., efference copy of the raw motor command) but, rather, to a signal coding for the movement of the target into an external frame of reference.

Conclusion

The execution of pursuit eye movements is a logical consequence of the restriction of the high spatial resolution to the fovea. These eye movements consist of two phases: the an initial open-loop phase, which is dominated by the fast, in some cases incorrect, processing of retinal image motion; and the subsequent closed-loop phase, which is dominated by the presence of a representation of target movement in space based on the combination of retinal and extra-retinal signals.

See also: Frontal Eye Fields; Parietal Cortex and Spatial Attention; Saccade–Pursuit Interactions; Sensorimotor Integration: Models; Shape Representation in Inferotemporal Cortex; Target Selection for Pursuit and Saccades; Visual Motion Models; Visual Cortical Models of Orientation Tuning; Visual Motion Detection.

Further Reading

Born RT and Pack CC (2002) Integration of motion signals for smooth pursuit eye movements. *Annals of the New York Academy of Sciences* 956: 453–455.

Chou IH and Lisberger SG (2004) The role of the frontal pursuit area in learning in smooth pursuit eye movements. *Journal of Neuroscience* 24: 4124–4133.

Fukushima K, Yamanobe T, Shinmei Y, Fukushima J, Kurkin S, and Peterson BW (2002) Coding of smooth eye movements in three-dimensional space by frontal cortex. *Nature* 419: 157–162.

Heinen SJ, Badler JB, and Ting W (2005) Timing and velocity randomization similarly affect anticipatory pursuit. *Journal of Vision* 5: 493–503.

Ilg UJ and Churan J (2004) Motion perception without explicit activity in areas MT and MST. *Journal of Neurophysiology* 92: 1512–1523.

Ilg UJ, Schumann S, and Thier P (2004) Posterior parietal cortex neurons encode target motion in world-centered coordinates. *Neuron* 43: 145–151.

Krauzlis RJ (2004) Recasting the smooth pursuit eye movement system. *Journal of Neurophysiology* 91: 591–603.

Krauzlis RJ, Dill N, and Kornylo K (2002) Activity in the primate rostral superior colliculus during the "gap effect" for pursuit and saccades. *Annals of the New York Academy of Sciences* 956: 409–413.

Lisberger SG, Morris EJ, and Tychsen L (1987) Visual motion processing and sensory-motor integration for smooth pursuit eye movements. *Annual Review of Neuroscience* 10: 97–129.

Masson GS and Stone LS (2002) From following edges to pursuing objects. *Journal of Neurophysiology* 88: 2869–2873.

Newsome WT, Wurtz RH, and Komatsu H (1988) Relation of cortical areas MT and MST to pursuit eye movements, II: Differentiation of retinal from extraretinal inputs. *Journal of Neurophysiology* 60: 604–620.

Schlack A, Hoffmann KP, and Bremmer F (2003) Selectivity of macaque ventral intraparietal area (area VIP) for smooth pursuit eye movements. *Journal of Physiology* 551: 551–561.

Tanaka M and Lisberger SG (2001) Regulation of the gain of visually guided smooth-pursuit eye movements by frontal cortex. *Nature* 409: 191–194.

Thier P and Ilg UJ (2005) The neural basis of smooth-pursuit eye movements. *Current Opinion in Neurobiology* 15: 645–652.

Relevant Websites

http://www.keck.ucsf.edu – Neuroscience and Physiology at UCSF.
http://www.inma.ucl.ac.be – Université Catholique de Louvain, Départment d'ingénierie mathématique.

Rab3

O M Schlüter, Göttingen University Medical School and Max-Planck-Society, Göttingen, Germany

© 2009 Published by Elsevier Ltd.

Introduction

Rab3A, 3B, 3C, and 3D constitute a subfamily of small guanosine triphosphate (GTP)-binding proteins among the Rab family members. The Rab family is the largest family of small ras-like GTPases. Rab proteins are essential molecular switches in regulating virtually all membrane trafficking steps in eukaryotic cells. Rab proteins bind two operationally defined classes of proteins. The first class, the auxiliary proteins, regulates the Rab cycle and the switching between the active GTP-bound form and the inactive guanosine diphosphate (GDP)-bound form. The second class is defined as the Rab effectors, which specifically interact with the active Rab protein and define the function of the Rab protein in the membrane-trafficking step.

During evolution, the Rab family expanded from 7 or 11 members in yeast to more than 60 members in mammals to account for the special requirements of differentiated tissues in mammalian multiorgan systems. The steps involved in membrane traffic are (1) vesicle budding from the source membrane, (2) cargo transport along the cytoskeletal elements, (3) target membrane recognition and attachment, and (4) fusion with the target membrane. Rab proteins are involved in all these events by controlling the assembly and disassembly of effector complexes involved in these processes.

Rab3 paralogs regulate stimulated secretion, in which the final events of membrane traffic are further subdivided into (4a) vesicle docking (i.e., hooking the vesicle onto the release machinery), (4b) priming (when the vesicle becomes fusion competent for Ca^{2+}-triggered release), and (4c) Ca^{2+}-triggered fusion. In synaptic vesicle fusion, these events are tightly coordinated and account for the fastest membrane-trafficking event in eukaryotic cells; it takes 50 μs to transduce the Ca^{2+} influx after excitation to vesicle fusion, and this is approximately 10 times faster than other regulated fusion events, such as granule fusion in adrenal chromaffin cells.

The efficacy of triggered fusion is dynamic, which results in synaptic plasticity. Changes in synaptic efficacy on a short timescale of milliseconds to minutes are defined as short-term synaptic plasticity, important for shaping the signal flow during trains of action potentials. Conversely, changes on a long timescale of hours and perhaps longer are defined as long-term synaptic plasticity, important for mechanisms of synaptic memory. Synaptic plasticity is regulated both presynaptically, in terms of changes in release probability, and postsynaptically, in terms of strength of the neurotransmitter-triggered receptor response. Rab3 is important for regulating the release probability at some synapses and is essential for at least one form of long-term synaptic plasticity expressed presynaptically.

Cloning and Structure of Rab3

Rab3A (initially named smg p25A) is the most abundant GTP-binding protein in the brain and was first purified and cloned (together with its paralogs Rab3B and -3C) from bovine brain membranes. The nucleotide sequence was independently pulled from a rat cDNA library during a ras GTPase homologs screen. The three Rab3 paralogs 3A, 3B, and 3C are enriched in brain, neuroendocrine, and endocrine tissues, where they localize to synaptic vesicles and secretory granules. Rab3D was later cloned from an adipocyte cell line, in which it associates with insulin-sensitive glucose transporter (GluT4) vesicles. Rab3D was also identified on secretory granules from mast cells in the lung and peritoneum, zymogen granules of the pancreas, and post-*trans*-Golgi network vesicles of osteoclasts.

The amino acid identity within the Rab3 subfamily amounts to 77–85%. This value is in the range of other Rab subfamilies, which averages 78.3% amino acid identity, whereas the Rab family average identity is only around 33%. The highest identity is found in the four Rab subfamily sequence (RabSF) motifs, which are characteristic for each subfamily. Whereas the RabSFs are solely determined from primary sequence comparisons, other characteristic motifs, which specifically interact with an effector protein, were derived from the crystal structure of active Rab3A bound to rabphilin. Three regions, named Rab complementary-determining region (RabCDR), form a binding pocket that specifically interacts with a conserved SGAWF(F/Y) structural element of the Rab3 effectors rabphilin, Rab3-interacting molecule (RIM)1α and 2α, and NOC2. RabSF1, RabSF3, and RabSF4 correspond approximately to RabCDR1, -2, and -3, respectively, indicating that RabSF motifs might build effector interaction surfaces of Rab subfamily members. RabSF2 maps to the first α-helix and loop 2 of Rab3A, which do not interact with rabphilin and might therefore form an additional subfamily-specific surface that could bind unrelated effectors or auxiliary proteins.

A key feature of small GTP-binding proteins is that they bind only in the GTP-bound active conformation to effector proteins and they could be switched off by GTP hydrolysis. The structural correlate for this mechanism in Rab3A is the interactions of the switch I and switch II region with rabphilin. On GTP hydrolysis, these two motifs change their conformation and retract the additional binding interface, loosening the strength of the Rab–effector interaction.

Rab3A is attached to synaptic vesicle membranes by posttranslationally modified C-terminal cysteines, which are isoprenylated by a Rab-specific geranyl–geranyl transferase (GGT) and are, furthermore, carboxy-methylated at the terminal residue. The isoprenylation is a common modification of Rab proteins and enables them to bind to intracellular membrane compartments. The lipid attachment is facilitated by Rab escort protein (REP), which binds to GGT and buries the hydrophobic modification, before Rab proteins are hooked onto membranes or bind to GDP dissociation inhibitor (GDI), an additional Rab auxiliary protein.

Rab3 Cycle and the Role of Auxiliary Proteins

As molecular switches, Rab proteins cycle between GTP-bound (active) and GDP-bound (inactive) forms. Membrane association and dissociation parallel this cycle. These features enable Rab proteins to specifically control single steps in a directed membrane-trafficking pathway. In their active state, they recruit effector proteins (or complexes) to the specific membrane-trafficking step. Afterward they are switched off and released into the cytosol to be reloaded for a new step. The membrane dissociation ensures that the protein does not need to recycle with the membrane it was loaded on; instead, it can take a short cut to be specifically used in its target process again. The Rab cycle requires multiple auxiliary proteins, which assist Rab proteins in membrane targeting, activation, inactivation after its performed task, and retrieval from the membrane. This is a multistep procedure, and the tracks of multiple Rab proteins overlap partially. This allows for a reduced set of auxiliary proteins, which assist multiple Rab proteins. However, compartmentalization, noncomplementary sets of auxiliary proteins, and specific effectors define specificity of the Rab protein function.

At the first step, the solubility of GDP-bound Rab proteins in the cytosol is increased. For this purpose, Rab proteins form a complex with GDI, in which the hydrophobic isoprenyl groups are enveloped in a groove of GDI. Only three GDI paralogs are expressed in mammals. GDIβ is expressed ubiquitously, whereas the other two mammalian paralogs have restricted expression patterns; GDIα is predominantly expressed in brain, whereas GDIδ is expressed specifically in muscle and adipocyte cells.

The mechanisms of the next step, the targeting of Rab proteins to their membrane compartment, is just beginning to evolve. A small family of GDI-displacement factors (GDFs) potentially catalyzes the initial entry of Rab proteins to membranous compartments. The mammalian homologs, prenylated Rab acceptor protein (PRA)1 and -2 bind promiscuously multiple Rab proteins, including Rab3A and, in addition, other prenylated small GTPases. Altogether 14–16 PRA (also named Yip) paralogs exist in mammals, of which most are ubiquitously expressed, indicating that no specific Rab protein is assisted by a single GDF. This is advantageous because increased compartmentalization in eukaryotic cells requires at least two entry points for Rab proteins. Rab3 needs to enter transport vesicles to the synapse after biosynthesis in the neuron soma and later has to reversibly attach to synaptic vesicles or the synaptic vesicle precursors from the endosome compartment. Consistent with this hypothesis, both PRA1, localized in the Golgi/endosome compartment, and PRA2, localized in the endoplasmic reticulum, can bind to Rab3 and might serve as the Rab3 GDFs.

After entering the membrane compartment, Rab3 is activated by a GDP/GTP exchange protein (GEP). It has not been directly shown whether the activation happens immediately after Rab3 enters the membrane or after Rab3 reaches the target membrane. Evidence for the first scenario is provided by the following observations. The synaptic targeting and binding on synaptic vesicles of the Rab3 effector rabphilin is strictly dependent on Rab3, whereas Rab3 synaptic targeting does not require rabphilin. In Rab3A single and Rab3 quadruple knockout mice, rabphilin protein (but not mRNA) levels are decreased and rabphilin accumulates in the neuron soma. This indicates that Rab3 is activated in the neuron soma before it reaches synaptic vesicles. However, during local recycling, Rab3 is also activated in synaptic terminals, either on synaptic vesicles or their precursors.

A cytosolic Rab3 GEP was isolated from rat brain, which is the ortholog of the human DENN/ mitogen-activated protein kinase (MAPK)-activating death domain (MADD) protein involved in blocking stress-induced apoptotic pathways in Alzheimer's and other neurodegenerative diseases. Synaptic targeting of Rab3 is dependent on Rab3 GEP function because in mutants of the *Caenorhabditis elegans* Rab3 GEP ortholog aex-3, Rab3 levels are reduced in synapses. Aex-3 serves as a GEP for an additional

Rab. It also catalyzes the GDP/GTP exchange of Rab27, a synaptic Rab protein in *C. elegans*. In aex-3 mutants, synaptic-rich structures are depleted of both Rab3 and Rab27. In addition to the GDP/GTP exchange, aex-3 has another function, whose connection to Rab function is currently unclear. It interacts with Cab-1, a protein involved in neural cell proliferation and differentiation. These multiple functions of the Rab GEP indicate why, in both *C. elegans* and mice, the phenotype of the Rab3 GEP/aex-3 mutants is more severe than in corresponding Rab3 mutants. Both Rab3 quadruple knockout and Rab3 GEP knockout mice die shortly after birth due to respiratory failure. However, whereas in Rab3-lacking neurons the evoked synaptic transmission is reduced modestly by 30%, synaptic transmission in Rab3 GEP knockouts is reduced by 80%. Furthermore, the size of the synaptic vesicles and shape of the active zone were altered in synapses lacking Rab3 GEP activity, whereas both structures were unchanged in Rab3-lacking neurons. In *C. elegans*, the aex-3 phenotype could be mimicked by combining either the Rab3/Rab27 mutations or the Rab3/Cab-1 mutations because the single Rab27 and Cab-1 mutants have similar phenotypes. After activation, Rab3 binds to specific effectors, which define the function of each individual Rab protein.

To terminate Rab3 function, the Rab3 intrinsic GTPase activity is enhanced by Rab3 GTPase activating protein (GAP), catalyzing the conversion of GTP to GDP. Rab3 GAP is a heteromeric complex consisting of a 130-kDa catalytic subunit and a 150-kDa noncatalytic subunit. The human genome encodes more than 50 putative Rab GAPs, which could indicate a high degree of specificity in the termination of Rab activity. However, the specificity of some GAPs is not restricted to one small GTP-binding protein family and acts on the Rab family and on Rho and other families, so we need more comprehensive data on the specificity of the Rab GAPs to draw general conclusions.

Rab3 GAP is ubiquitously expressed and enriched in the synaptic soluble fraction of the brain. Mutations in p130 have been reported to cause Warburg Micro syndrome, a condition with severe mental retardation, indicating an important function for Rab3 GAP in synaptic function. The function of Rab3 GAP was further tested in p130-deficient mice. From all proteins tested, only the levels of p150 were severely reduced, emphasizing the importance of the complex *in vivo*. Significantly, the fraction of GTP-bound Rab3A was increased, indicating that Rab3 GAP is the predominant *in vivo* GAP for Rab3A. Analysis of the synaptic function revealed that the elevated level of active Rab3A led to an increased paired-pulse ratio of two shortly spaced evoked synaptic responses in hippocampal CA1 pyramidal cells. Changes in the paired-pulse ratio are a sensitive measure of functional presynaptic alteration and are often caused by a decrease in synaptic release probability. However, the results of testing release probability were not conclusive because the use-dependent block of *N*-methyl-D-aspartate (NMDA) receptor responses with MK-801, a measure of synaptic release probability, was not altered in the Rab3 GAP-lacking neurons, whereas potassium-induced glutamate release from synaptosomes was greatly attenuated in the absence of Rab3 GAP. More detailed experiments are required to address the consequences of an increase in active Rab3A on synaptic function.

After synaptic vesicle fusion, Rab3A and -3C and their effector rabphilin dissociate from synaptic vesicles. This dissociation requires both synaptic vesicle fusion and Ca^{2+} entry. Osmotic or α-latratoxin-induced release in the absence of Ca^{2+} prevents Rab3 dissociation. These results indicate that synaptic vesicle fusion and Rab3 dissociation are normally coupled but are independent reactions. A tempting question remains of whether Rab3 GAP activity and GTP hydrolysis are involved in the Ca^{2+}-induced retrieval of Rab3 from synaptic vesicles because for some Rab GTP hydrolysis is not absolutely required for function termination.

To close the Rab3 cycle, brain-specific GDIα retrieves Rab3A from synaptic vesicle membranes using a membrane-associated heat shock protein (Hsp)90 chaperone complex.

Rab3 Function in Membrane Traffic and the Role of the Effectors

In neurons, Rab3 has two classes of effector proteins, rabphilin and Rim1α/2α. Rabphilin binds via Rab3 onto synaptic vesicles, whereas RIM1α and -2α are independent components of the active zone at the synapse. Because RIM resides close to the target membrane of synaptic vesicles, an anticipated function of the Rab3–RIM complex was to regulate the tethering of synaptic vesicles at the active zone; a function multiple Rab proteins have in common. It was surprising that the most abundant Rab protein in the brain, Rab3A, was not essential for the tethering of synaptic vesicles. The distribution of synaptic vesicles at the active zone is not convincingly changed to indicate an essential function for Rab3A in synaptic vesicle tethering. Furthermore, there is no obvious change in the distribution of synaptic vesicles in mice lacking all four Rab3 paralogs or Rim1α, indicating that the Rab3–RIM complex has no essential

function in synaptic vesicle tethering. Interestingly, the number of synaptic vesicles at the neuromuscular junction active zone is reduced in the *C. elegans* Rab3 mutant. However, it is not clear whether this change is due to a defect in synaptic vesicle tethering or results from a defect later in the synaptic vesicle cycle.

Rab3 dissociates from synaptic vesicles after Ca^{2+}-triggered vesicle fusion, indicative of a function after docking. Whereas the size of the readily releasable vesicle pool (primed vesicle pool) is unaltered in Rab3 lacking hippocampal neurons, the evoked synaptic responses are reduced by 30%. This reduction in vesicular release probability could be readily explained by the reduced apparent Ca^{2+} affinity of the release machinery in the Rab3-deficient neurons. However, the key finding explaining how Rab3 is involved in this process was an accelerated repriming rate in the Rab3-deficient neurons. This finding indicates that Rab3 is involved in an additional maturation step of primed synaptic vesicles. Synaptic vesicles in this superprimed state, whose conversion is rate-limiting in the priming reaction, have an elevated release probability and are the ones to be released initially during an action potential train. This has immediate consequences for short-term synaptic plasticity, which was altered in Rab3-deficient neurons to a greater extent, compared to a mean change in release probability across all synaptic vesicles. The importance of the additional maturation step in synaptic vesicle priming, at least in mammals, is demonstrated by the lethal phenotype of the Rab3 quadruple knockout mice.

Consistent with the hypothesis of two vesicle populations with different release probabilities, one study analyzed the retrieval of green fluorescent protein (GFP)-tagged Rab3A from synaptic vesicles during trains of action potential and found that the fluorescence dispersion ends before the vesicle pool turns over.

How do the Rab3 effectors contribute to the maturation step in priming? The synaptic localization of rabphilin relies completely on Rab3. Therefore, Rab3 quadruple knockout mice are also deficient in synaptic rabphilin. Conversely, Rab3A is targeted to synaptic vesicles independent of rabphilin, so the function of rabphilin could be analyzed in rabphilin knockout mice in isolation. Surprisingly, no obvious alterations in synaptic transmission and plasticity were detectable in the hippocampus of rabphilin-deficient mice, whereas rabphilin mutants in *C. elegans* are lethargic. This phenotypic difference could indicate either functional redundancy in mice or additional functions of the *C. elegans* homolog. The latter explanation appears to account for the difference because in *C. elegans* rabphilin interacts with two Rab proteins, Rab3 and Rab27, whereas in mammalian synapses rabphilin is dependent on Rab3 alone, despite interacting with Rab3 and Rab27 *in vitro*. The mammalian effectors of Rab27 have a similar structure compared to rabphilin and belong to the family of synaptotagmin-like proteins (Slps), including granuphilin. In mammals, Rab27 and its effectors are localized primarily on secretory granules. Therefore, *C. elegans* rabphilin might serve a dual function as a Rab3 and Rab27 effector, whereas later in evolution more specialist effectors might have developed.

In an analysis of the function of RIM1α, the only α RIM paralog in mouse hippocampal neurons, significant differences to Rab3 mutants were described. The changes were much more complex, compared to changes in the Rab3-deficient neurons. This is probably due to additional interactions of RIM1α with α-liprin, an essential component for active-zone differentiation; ERC; RIM- and liprin-binding proteins of the active zone; and Munc13–1, an essential priming factor, whose levels are reduced in the RIM1α knockouts but not the Rab3 knockouts. In RIM1α knockout mice, the readily releasable pool was reduced by 50%, which resulted in a similar reduction of evoked synaptic responses. Consistent with this result, in hippocampal slices the synaptic release probability was reduced, whereas the vesicular release probability in hippocampal neurons was unaltered. The synaptic release probability measures whether an action potential at a synapse can elicit neurotransmitter release, whereas the vesicular release probability depends on the number of released vesicles at all synapses. Thus, in RIM1α-deficient neurons, the likelihood is that, after triggering, a vesicle fuse is unchanged, but the number of primed vesicles and therefore fusion events is reduced. Consistent with a sole reduction of fusion-competent vesicles, the apparent Ca^{2+} affinity of the release apparatus and the repriming rate were unchanged. This resulted in a change of paired-pulse ratios similar to changes induced by a mean reduction in release probability. The increase was biggest for short interstimulus intervals, in contrast to the change in Rab3-deficient neurons, in which the increase was biggest for intermediate interstimulus intervals. Altogether, these changes in RIM1α-deficient neurons could best be explained by a reduction in synaptic vesicle priming, supported by the reduction of priming factor Munc13–1 levels in RIM1α-deficient synapses. A similar conclusion was drawn from the analysis of *C. elegans* mutants of the RIM homolog unc-10.

In addition to the priming defect in RIM1α-deficient neurons, the release kinetics of synaptic vesicles are changed. After Ca^{2+} influx, the majority

of synaptic vesicles are released instantaneously (termed synchronous release), whereas release of a small fraction of vesicles is delayed (termed asynchronous release). Surprisingly, in RIM1α-deficient neurons, the asynchronous component was preferentially reduced, indicating that the release properties of the primed vesicles are altered.

RIM1α interacts with both Rab3 and Munc13–1 in a trimeric complex and this complex might regulate synaptic vesicle priming. However, the role in priming of Rab3 and RIM1α appear very separated. Whereas RIM1α controls primarily the number of fusion-competent vesicles, Rab3 controls the release probability of fusion-competent vesicles. Surprisingly, these functions do not appear to be interconnected because the most characteristic changes are not shared in the corresponding knockout mice. The number of fusion-competent vesicles is not changed in Rab3-deficient neurons, and the vesicular release probability and priming rate are unchanged in RIM1α-deficient neurons. The functional analysis of both Rab3 effectors RIM1α and rabphilin indicates that both are not responsible for defining the essential role of Rab3 on synaptic vesicle priming. It is interesting to identify the effector which defines this function.

Despite its independent role on synaptic vesicle priming, RIM1α is the effector of Rab3 mediating its function on hippocampal mossy fiber long-term potentiation. This form of plasticity requires protein kinase A (PKA) activation and is expressed presynaptically. In both Rab3A- and RIM1α-deficient mice, this form of plasticity is absent. Surprisingly the short-term alterations, characteristic of both mutants in CA1 pyramidal cells (previously discussed) were absent in the granule cell CA3 pyramidal cell synapses. It will be important to test whether in the mossy fiber terminals additional RIM and Rab3 paralogs are expressed, which are functionally redundant to mask the functional loss of Rab3A and RIM1α, and whether the mechanisms of priming are different. The mechanism of RIM1α function in long-term plasticity was further analyzed in a similar form of plasticity in cerebellar neurons. It was shown that RIM1α phosphorylation is required for this form of plasticity. Thus, these forms of long-lasting increase in transmitter release require both Rab3A–RIM1α interaction and phosphorylation of RIM1α. Because RIM1α can modulate the number of fusion-competent vesicles, a plausible expression mechanism for this form of plasticity could be a long-lasting increase in the readily releasable vesicle pool.

See also: Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Munc13 and Associated Molecules; Priming; Synaptic Vesicles; Vesicle Pools.

Further Reading

Calakos N, Schoch S, Südhof TC, and Malenka RC (2004) Multiple roles for the active zone protein RIM1alpha in late stages of neurotransmitter release. *Neuron* 42: 889–896.

Castillo PE, Schoch S, Schmitz F, Südhof TC, and Malenka RC (2002) RIM1alpha is required for presynaptic long-term potentiation. *Nature* 415: 327–330.

Dulubova I, Lou X, Lu J, et al. (2005) A Munc13/RIM/Rab3 tripartite complex: from priming to plasticity? *EMBO Journal* 24: 2839–2850.

Fukuda M (2003) Distinct Rab binding specificity of Rim1, Rim2, rabphilin, and Noc2: Identification of a critical determinant of Rab3A/Rab27A recognition by Rim2. *Journal of Biological Chemistry* 278: 15373–15380.

Fukui K, Sasaki T, Imazumi K, Matsuura Y, Nakanishi H, and Takai Y (1997) Isolation and characterization of a GTPase activating protein specific for the Rab3 subfamily of small G proteins. *Journal of Biological Chemistry* 272: 4655–4658.

Grosshans BL, Ortiz D, and Novick P (2006) Rabs and their effectors: Achieving specificity in membrane traffic. *Proceedings of the National Academy of Sciences of the United States of America* 103: 11821–11827.

Iwasaki K and Toyonaga R (2000) The Rab3 GDP/GTP exchange factor homolog AEX-3 has a dual function in synaptic transmission. *EMBO Journal* 19: 4806–4816.

Mahoney TR, Liu Q, Itoh T, et al. (2006) Regulation of synaptic transmission by RAB-3 and RAB-27 in *Caenorhabditis elegans*. *Molecular Biology of the Cell* 17: 2617–2625.

Matsui Y, Kikuchi A, Kondo J, Hishida T, Teranishi Y, and Takai Y (1988) Nucleotide and deduced amino acid sequences of a GTP-binding protein family with molecular weights of 25,000 from bovine brain. *Journal of Biological Chemistry* 263: 11071–11074.

Nonet ML, Staunton JE, Kilgard MP, et al. (1997) Caenorhabditis elegans rab-3 mutant synapses exhibit impaired function and are partially depleted of vesicles. *Journal of Neuroscience* 17: 8061–8073.

Oishi H, Sasaki T, Nagano F, et al. (1998) Localization of the Rab3 small G protein regulators in nerve terminals and their involvement in $Ca2^{+}$-dependent exocytosis. *Journal of Biological Chemistry* 273: 34580–34585.

Pereira-Leal JB and Seabra MC (2001) Evolution of the Rab family of small GTP-binding proteins. *Journal of Molecular Biology* 313: 889–901.

Schlüter OM, Basu J, Sudhof TC, and Rosenmund C (2006) Rab3 superprimes synaptic vesicles for release: Implications for short-term synaptic plasticity. *Journal of Neuroscience* 26: 1239–1246.

Schoch S, Castillo PE, Jo T, et al. (2002) RIM1alpha forms a protein scaffold for regulating neurotransmitter release at the active zone. *Nature* 415: 321–326.

Südhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Rab3A Interacting Molecules (RIMs)

P S Kaeser and C M Powell, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

RIM Genes and Proteins

RIM Discovery and Isoforms

RIM1α is the founding member of the RIM (Rab3-interacting-molecule) protein family. It was first detected as a binding partner of Rab3A in a yeast-two hybrid screening. Rab3A is a small guanosine triphosphatase (GTPase) that is covalently linked to synaptic vesicles via geranylgeranyl moieties. Thus, RIMs are named after this interaction. RIM1α is tightly and selectively associated with the presynaptic active zone, the electrondense material that is attached to the presynaptic membrane opposite postsynaptic densities (Figure 1) and that regulates synaptic vesicle docking and priming.

RIM1α is a large multidomain protein (Figure 2(a)) that consists of an N-terminal zinc finger domain, a relatively large linker sequence that contains a protein kinase A phosphorylation site (serine 413), a central PSD-95/Dlg/ZO-1 (PDZ) homology domain, and two C-terminal C_2domains (called C_2A and C_2B). Between the C_2 domains, there is a relatively long linker region that can be alternatively spliced at two splice sites and that also contains a proline-rich region.

Since the initial identification of RIM1α, investigators have found three additional RIM genes that produce six isoforms and many splice variants (Figures 2(b) and 2(c)) in mammals. The RIM1, -3, and -4 genes each produces only a single protein isoform (RIM1α, RIM3γ, and RIM4γ, respectively). The RIM2 gene encodes a full-length α-isoform, a β-isoform in which the N-terminal zinc finger is replaced by a short β-specific sequence, and a short γ-isoform that consist of only the C-terminal C_2domain with a short N-terminal γ-specific peptide (Figure 2(b)). The RIM1 and RIM2 genes are large genes (Figure 2(c)) of approximately 0.5 Mb. Their chromosomal locations are indicated in Figure 2(c).

In both *Drosophila* and in *Caenorhabditis elegans,* RIM homologs have been identified. Whereas there is no functional data available about the *Drosophila* RIM gene, mutants for the *C. elegans* gene unc-10 that encodes RIM have been analyzed in detail. Although some differences in the molecular and cellular functions of mammalian RIM and *C. elegans* RIM have been suggested, the basic function of RIM as being necessary for the efficient regulation of presynaptic release is conserved between invertebrates and vertebrates.

N-Terminal Zinc Finger Domain

The N-terminal zinc finger domain of α-RIMs is very similar to the zinc finger domain of rabphilin, another Rab3 binding protein that is soluble and bound to synaptic vesicles via Rab3 isoforms. Two elements of the N-terminal zinc finger are important for Rab3 binding: (1) N-terminal of the zinc-binding cysteine residues, α-RIMs and rabphilin contain a long α-helix called α1, and (2) C-terminal of the zinc finger, they include a SGAWFF/Y motif at the end of helix α2. The structures of the complexes of Rab3 with RIM1α or rabphilin have been described in detail based on nuclear magnetic resonance (NMR) analysis and crystallography.

In addition to Rab3s, the RIM N-terminus binds to the active-zone protein Munc13. This interaction occurs between the RIM zinc finger domain and the Munc13–1 C_2A domain. Evidence for this interaction is convincing. First, this binding was characterized by use of glutathione *S*-transferase (GST) pull-down experiments and co-immunprecipitation. Later, it was strongly supported by the fact that brain Munc13–1 levels are reduced by approximately 50% in a mouse line lacking RIM1α. Initially, *in vitro* experiments suggested that the binding of either Rab3A or Munc13 to the N-terminus of RIM1α exclude one another. However, more recent data suggest that RIM, Rab3, and Munc13 form a tripartite complex that is involved in priming synaptic vesicles for release and that potentially modulates synaptic plasticity. Interestingly, the SGAWFF/Y motif that is C-terminal to the Munc13–1 binding site is released from Rab3A when the RIM–Rab3A complex binds to Munc13–1. The RIM1 gene (but not the RIM2 gene) contains an alternatively spliced exon between the Rab3 interacting portion of helix α1 and the zinc finger domain. It has been suggested that alternative splicing at this splice site A (SSA) might regulate the binding of either Rab3s or Munc13s, but the experimental evidence for this is inconclusive.

Zinc Finger/PDZ Domain Linker Sequence

Several putative phosphorylation sites have been identified in the linker sequence between the zinc finger and the PDZ domain, and one – the serine residue 413 (S413) – has been characterized in detail. It is phosphorylated by the cyclic adenosine monophophate

Figure 1 Electron micrograph of a presynaptic terminal (pr) synapsing onto a postsynaptic neuron (po) stained for RIM1α by silver-enhanced immunogold labeling showing that RIM1α is localized at presynaptic active zones in the spinal motor cord neurons. RIM1α signal appears as black deposits found only in presynaptic nerve terminals close to the active zone. Scale bar = 0.2 μm. Adapted by permission from Macmillan Publishers Ltd: *Nature,* Wang Y, Okamoto M, Schmitz F, Hofmann K, and Sudhof TC (1997) Rim is a putative Rab3 effector in regulating synaptic-vesicle fusion. 388: 593–598, copyright 1997.

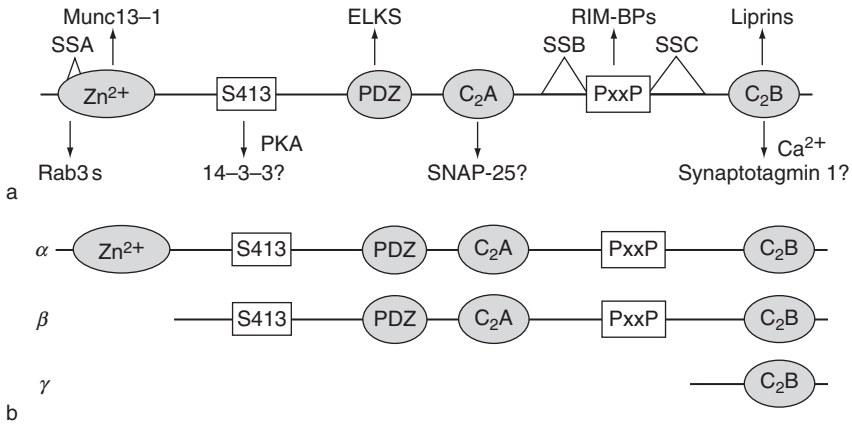

Gene	Protein isoforms	Mouse/human chromosome	Mouse/human size (kb)
Rim1 (Rims1)	RIM1α	1A3/6q14	491.3/512.7
Rim2 (Rims2)	RIM2α	15A2/8q23	485.4/747.9
	RIM2β		
	RIM2γ		
Rim3 (Rims3)	RIM3γ	4D2.2/1p34.1	13.7/15.4
Rim4 (Rims4)	RIM4γ	2H3/20q13.3	53.9/54.5

c

Figure 2 Domain structure and genetic organization of RIM proteins: (a) domain structure of the RIMα isoform showing RIMs interaction partners and sites of alternative splicing; (b) domain structure of RIM isoforms; (c) genetic representation of RIM proteins. C_2A and C_2B, C-terminal C_2domains; cAMP, cyclic adenosine monophosphate; ELKS, protein rich in E, L, K, and S; PDZ, PSD-95/Dlg/ZO-1; PKA, cAMP-dependent protein kinase A; PxxP, proline rich region; RIM-BPs, Rab3A interacting molecule-binding proteins; S413, serine 413; SNAP-25, synaptosomal-associated protein of 25 kDa; SSA, splice site A; SSB, splice site B; SSC, splice site C; Zn^{2+}, zinc finger domain. Reproduced with permission, from Kaeser PS and Sudhof TC (2005), *Biochemical Society Transactions* 33: 1345–1349, © the Biochemical Society.

(cAMP)-dependent protein kinase A (PKA), and phosphorylation of S413 is developmentally regulated. In rats, it is highly phosphorylated at birth, and the percentage of phosphoserine 413 RIM1α slowly decreases to reach a relatively low steady state at around postnatal day 30. *In vitro* experiments have further suggested that the S413-phosphorlylated peptide binds to 14–3–3 adaptor proteins, but there is no *in vivo* data to support such an interaction. This PKA phosphorylation site is also present in RIM2β. The physiological relevance of PKA phosphorylation of S413 is a subject of intense investigation.

PDZ Domain

α- and β-RIMs contain a central PDZ domain with unusual binding specificities. They bind to the C-terminal peptide of a class of proteins called ELKS (for protein rich in E, L, K, and S). These active-zone proteins have also been variously referred to as cytomatrix at the active-zone-related structural protein (CAST), Rab6-interacting molecule 2, and ELKS/Rab6IP2/CAST (ERCs). The hydrophobic C-terminal peptide of ELKS (with the amino acid sequence GIWA) matches the profile of a class II PDZ binding motif, but even conserved mutations in any of the four residues abolish binding. This may be because the RIM PDZ domain has an unusually narrow and deep binding groove. Piccolo, another active zone protein, contains a closely related PDZ domain, but it does not bind to the ELKS C-terminus. The high specificity of this interaction might suggest a role of ELKS in organizing active zones, especially because ELKS proteins interact with several other active zone proteins such as Piccolo and α-liprins via their coiled-coil regions.

C_2 Domains

At the C-terminus, α- and β-RIMs contain two C_2 domains (called C_2A and C_2B), whereas γ-RIMs only contain the C_2B domain. C_2 domains were originally described as Ca^{2+}-dependent phospholipid binding modules. The RIM C_2 domains are atypical because they do not bind to Ca^{2+} and they have not been shown to bind to phospholipids. In fact, their interactions are poorly understood, with little *in vivo* data to support *in vitro* findings. The C_2B domain of RIM binds to α-liprins *in vitro*, a class of proteins which are located at presynaptic active zones and which have been shown to regulate the size and organization of active zones in invertebrates. In addition, it has been suggested that the RIMs bind to synaptosomal-associated protein of 25 kDa (SNAP-25), synaptotagmin 1, and cAMP-guanidine nucleotide exchange factor II (GEFII). While tenuous, these possible interactions with the key effectors of neurotransmitter release provide provocative potential mechanisms for RIM function in modulating release.

The linker region between the C_2 domains contains a relatively large amino acid sequence with two alternative splice sites called SSB and SSC. There is a proline-rich region between these two canonical splice sites that binds to RIM-binding proteins (RIM-BPs), another class of active-zone proteins of unknown function.

The localization and biochemical interactions of RIM directly or indirectly with many additional active-zone and synaptic vesicle proteins puts RIM at the core of neurotransmitter release machinery at the synapse. Thus, RIMs are a central component of a protein scaffold that is involved in the organization of the presynaptic release machinery and that controls the docking and priming of synaptic vesicles (Figure 3). In the next section, we discuss the available functional data that have been obtained from a gene knockout (KO) mouse line for RIM1α, one of the key RIM isoforms.

RIM Function in Synaptic Transmission

To characterize the *in vivo* function of RIM1α, a mouse gene KO line has been generated. This mouse line has delivered key insights into the function of RIMs at a number of central synapses. Investigators found that RIMs are crucial for the normal regulation of neurotransmitter release at central nervous system

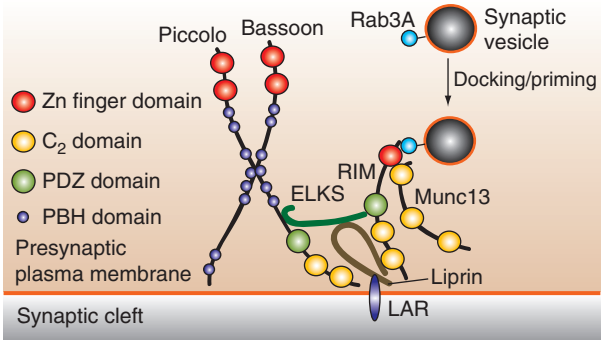

Figure 3 Schematic of the presynaptic active zone components that assemble in close proximity to the presynaptic plasma membrane. The available data put RIM in the center of this electrondense protein meshwork in terms of scaffolding structure and of function. RIM proteins form a tripartite complex with Rab3A on the synaptic vesicles and with Munc13–1 at the active zone, potentially coupling synaptic vesicles to the active zone and preparing them for Ca^{2+}-mediated release, a process referred to as priming. In addition, RIMs interact directly or indirectly with all other active zone proteins. LAR, LAR protein-tyrosine phophatases, a family of transmembrane proteins that is known to bind to liprins; ELKS, protein rich in E, L, K, and S; PBH, piccolo bassoon homology domain; PDZ, PSD-95/Dlg/ZO-1.

synapses. Interestingly, the loss of RIM1α leads to very different phenotypes at different types of synapses.

Excitatory synapses have been analyzed in RIM1α-KO mice with electrophysiological recordings from acute slices and cultured neurons of the hippocampus (Figure 4(a)) and the cerebellum. At different hippocampal synapses, very different phenotypes were observed.

Presynaptically Expressed Forms of Long-Term Potentiation

Some central synapses exhibit a presynaptically expressed, PKA-dependent form of long-term potentiation (LTP). These synapses are hippocampal mossy fiber to the CA1 pyramidal neuron synapses, cerebellar parallel fiber to Purkinje neuron synapses, corticostriatal synapses, and corticothalamic synapses. Both hippocampal mossy fiber LTP (mfLTP) and cerebellar parallel fiber LTP (pfLTP) are deficient in mice that lack RIM1α (Figures 4(d) and 4(e)). mfLTP is also absent in mice lacking the RIM1α-interacting molecule Rab3A. In spite of the lack of presynaptic mfLTP in RIM1α-KO mice, short-term plasticity and forskolin-induced potentiation are normal at these synapses. Interestingly, the RIM1α PKA phosphorylation site, S413, has been implicated in presynaptically expressed pfLTP in the cerebellum. *In vitro* experiments on cultured cerebellar neurons suggest that wild-type RIM1α can rescue the lack of pfLTP, whereas a mutant form of RIM1α lacking the S413 phosphorylation site cannot rescue pfLTP. These experiments suggest that PKA phosphorylation of RIM1α may be involved in presynaptically expressed forms of LTP. However, this observation in cultured cerebellar neurons *in vitro* remains to be validated *in vivo*.

Probability of Release and Short-Term Plasticity

In contrast, other synapses are affected in more basic parameters of neurotransmitter release. In Schaffer collateral to CA1 pyramidal cell synapses, paired pulse ratios, posttetanic potentiation (PTP), and

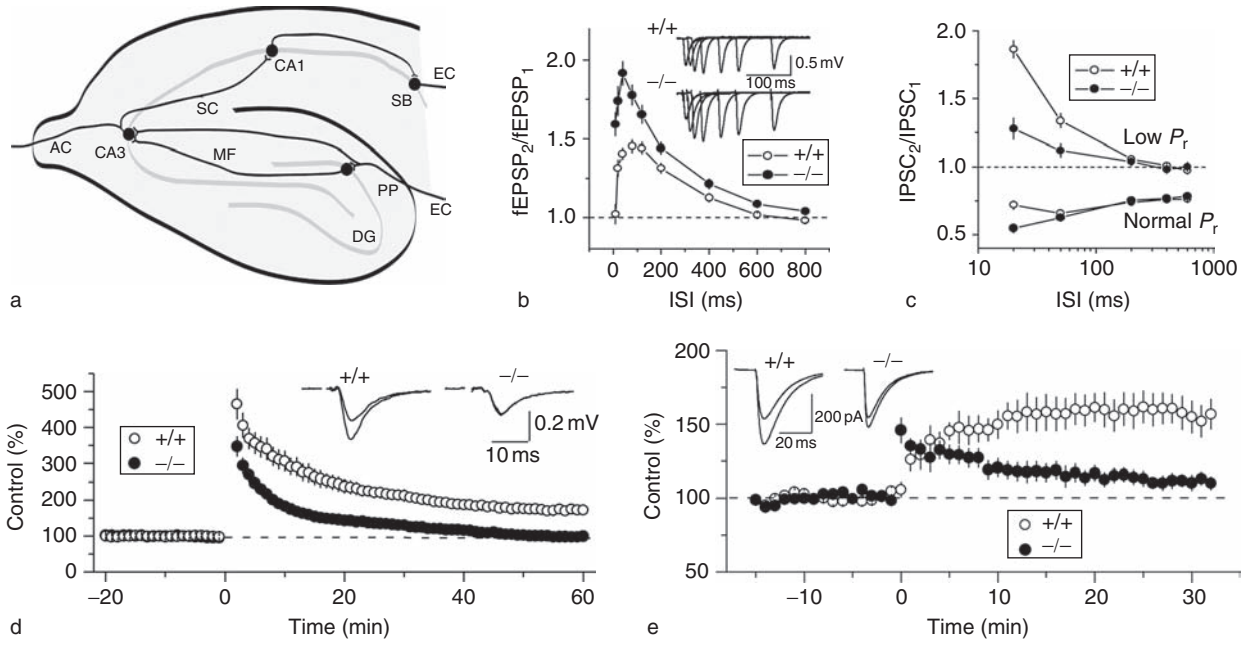

Figure 4 RIM function in short-term and long-term synaptic plasticity: (a) model of the unidirectional hippocampal pathways which were used to analyze several forms of synaptic plasticity in RIM1α-knockout mice; (b) increased paired pulse facilitation in RIM1α-knockout mice compared to littermate wild-type controls; (c) increased paired pulse depression at inhibitory synapses at short interstimulus intervals when the initial release probability is normal in RIM1α-knockout mice, and decreased facilitation when the initial release probability was reduced by lowering the extracellular calcium concentration (note that this effect is reversed compared to excitatory synapses (b)); (d) long-term potentiation at hippocampal mossy fiber synapses is absent in RIM1α-knockout mice; (e) long-term potentiation at cerebellar parallel fiber to Purkinje cell synapses is significantly decreased when RIM1α is absent. AC, associational commissural pathway; DG, dentate gyrus; EC, entorhinal cortex; EPSP, excitatory postsynaptic potential; IPSC, inhibatory postsynaptic current; ISI, interstimulus interval; MF, mossy fibers; PP, perforant path; P_r, release probability; SB, subiculum; SC, Schaffer collateral pathway. (a) Reproduced with permission, from Kaeser PS and Sudhof TC (2005), *Biochemical Society Transactions* 33: 1345–1349, © the Biochemical Society. (b, c) Reprinted with permission from Macmillan Publishers Ltd: *Nature,* Schoch S, Castillo PE, Jo T, et al. (2002) RIM1alpha forms a protein scaffold for regulating neurotransmitter release at the active zone. 415: 321–326, copyright 2002. (d, e) Reprinted from Macmillan Publishers Ltd: *Nature,* Castillo PE, Schoch S, Schmitz F, Sudhof TC, and Malenka RC (2002) RIM1alpha is required for presynaptic long-term potentiation. 415: 327–330, copyright 2002.

response to prolonged 14 Hz stimulation are significantly increased when RIM1α is absent (Figure 4(b)). These observations suggest a decrease in release probability (P_r). Indeed P_r at these synapses is decreased by 50% when measured indirectly using the rate of blockade of *N*-methyl-D-aspartate (NMDA) receptor-mediated synaptic responses with successive synaptic activations.

These recordings in acute hippocampal slices from RIM1a-KO mice also demonstrated that the loss of function phenotypes are not simply due to RIM1α's role as a Rab3A effector protein. Rab3A-KO mice show a similar decrease in paired pulse ratios, but no change in posttetanic stimulation, P_r, or in response to 14 Hz stimulation.

A subsequent study of RIM1α-deficient hippocampal autapses in culture revealed that the RIM1α-KO mice exhibit decreased P_r via a decrease in the readily releasable pool of synaptic vesicles. Curiously, the decreased P_r in RIM1α-KO hippocampal autapses is observed only during the initial four to five stimuli during sustained 14 Hz stimulation. Wild-type synapses exhibit a depression of P_r during 14 Hz stimulation, whereas RIM1α-KO synapses do not. Thus, the relationship between P_r and the frequency of synaptic activation in the RIM1α-KO mice is complex.

Long-Term Potentiation at Schaffer Collateral to CA1 Synapses

LTP at the Schaffer collateral to CA1 synapse is NMDA receptor-dependent and expressed in part via postsynaptic mechanisms. The early phase of LTP is normal in both RIM1α- and Rab3A-KO mice. However, more intense LTP induction protocols induce a longer-lasting form of plasticity that is dependent on protein synthesis (called late LTP). This late phase of LTP is deficient in RIM1α- and Rab3A-KO mice, implicating presynaptic expression in this form of LTP. Although the mechanism of this involvement remains a mystery, a requirement for presynaptic PKA activation has been suggested. Further study is warranted to confirm and extend this initial observation.

In summary, the loss of RIM1α at excitatory synapses leads to a relative decrease in neurotransmitter release and alterations in short-term and long-term plasticity. This suggests that RIM1α is a positive regulator of excitatory synaptic transmission and plasticity.

Inhibitory Synaptic Transmission

In contrast to its function at excitatory synapses, RIM1α may lead to a decrease in neurotransmitter release at inhibitory synapses. Inhibitory synapses in the CA1 region of the hippocampus show paired pulse depression (PPD) rather than paired pulse potentiation, putatively due to their high initial P_r. This depression can be reversed to facilitation by decreasing the extracellular Ca^{2+} concentration to lower the P_r. In acute hippocampal slices, the loss of RIM1α causes an increase in PPD at the CA1 inhibitory synapse at short stimulus intervals (Figure 4(c)). When the initial P_r is lowered to a level that is comparable to excitatory Schaffer collateral synapses by lowering extracellular Ca^{2+}, RIM1α-KO mice show a decrease in paired pulse facilitation (PPF), which is the opposite effect of what was observed at excitatory synapses. Thus, RIM1α may act as a negative regulator of inhibitory synaptic transmission and plasticity, and this is in strong contrast to the function of RIM at excitatory synapses.

This dichotomy of RIM function in excitatory and inhibitory synaptic transmission and the divergent functions of RIM molecules in different excitatory synapses present an interesting conundrum that leaves many questions open. What are the fundamental molecular mechanisms of RIM function? Which RIM binding partners are mediating the different effects? Do RIM binding partners differ at individual synapses? In addition, it is critical to address how the various RIM isoforms work together or separately in these functions. Can other RIM isoforms compensate for RIM1α? Do different combinations of RIM isoforms lead to different functional properties? These questions and many others are the subject of intense investigation.

Clearly, RIM1α is critical to normal synaptic function and plasticity. The relevance of these functions to the whole animal has been investigated in behavioral studies of RIM1α-KO mice, considered next.

RIM Proteins Are Critical for Normal Learning and Memory

The abnormalities in both short- and long-term plasticity in RIM1α-KO mice raised the issue of whether RIM1α might play a critical role in learning and memory. A broad behavioral analysis of RIM1α-KO mice revealed abnormalities in learning and memory in both the Morris water maze and fear conditioning. In spite of the various synaptic physiology abnormalities and the widespread expression of RIM1α throughout the brain, RIM1α-KO mice revealed normal motor coordination and anxiety-like behaviors.

The abnormalities in Morris water maze behavior imply that RIM1α function is important for hippocampus-dependent learning. In contrast, Rab3A-KO mice with mfLTP deficits, increased PPF, and abnormal synaptic depression exhibited normal learning and memory in fear conditioning and classic water maze behavior. Thus, altered mfLTP, PPF, and synaptic depression

alone are unlikely to play a role in the severe RIM1α learning and memory deficits. In other studies of Rab3A-KO mice, additional behavioral testing revealed deficits in episodic-like/working memory on a delayed match-to-place version of the water maze (Table 1). It will be of great interest to examine additional mutant mice selectively lacking mfLTP to confirm its role in this form of memory.

Another deficit that seems a likely candidate for the behavioral abnormalities in the RIM1α-KO mice is the decrease in P_r in area CA1. A decrease in P_r in isolation, however, does not necessarily alter learning and memory behavior. Synaptotagmin 1 R233Q point mutant knockin mice have a similar decrease in P_r in CA1 as the RIM1α-KO mice, yet they exhibit normal learning and memory on the same tasks. Regardless of

Table 1 Summary of plasticity and behavioral abnormalities in RIM1α and related mutant mice[a]

	Short-term plasticity	*Long-term plasticity*	*Behavior*	*Sources*[b]
PKA-KO	PPF normal	E-LTP↓, L-LTP normal or ↓, mfLTP ↓, CA1-LTD ↓	Normal fear conditioning Normal spatial memory (water maze and Barnes maze) Normal response to novelty	1, 7, 10
Rab3A-KO	PPF ↑, PTP normal, TS normal, P_r normal	E-LTP normal, L-LTP ↓, mfLTP ↓	Normal fear conditioning Normal spatial memory (water maze) Impaired reversal learning (delayed match to place) Normal response to novelty	2–4, 6, 9, 11
RIM1α-KO	PPF↑, PTP↑, TS ↑, P_r ↓	E-LTP normal, L-LTP ↓, mfLTP ↓, CA1-LTD ↑	Impaired fear conditioning Impaired spatial memory (water maze) Increased locomotor activity in response to novelty	3, 8, 9, 11
Syt1 R233Q	PPF ↑, P_r ↓	N.D.	Normal fear conditioning Normal spatial memory Normal response to novelty	5, 9

[a]Summary of behavioral deficits observed in mice that lack PKA (two lines have been analyzed lacking either the catalytic subunit Cβ_1 or the regulatory subunit RIβ; data are pooled), Rab3A, or RIM1α or mice that carry a arginine 233 to glutamine point mutation in Synaptotagmin 1 that decreases the neurotransmitter release probability to 50%. CA1-LTD, long-term depression in area CA1 of hippocampus; E-LTP, early long-term potentiation; KO, knockout; L-LTP, late long-term potentiation; mfLTP, mossy fiber long-term potentiation; N.D., not done; PKA-KO, cAMP-dependent protein kinase A knockout mice; PPF, paired pulse facilitation; P_r, release probability; PTP, posttetanic potentiation; Syt1 R233Q, synaptotagmin 1 arginine 233 to glutamine knockin mice; TS, response to 14 Hz train stimulation.

[b]Data from:

1. Brandon EP, Zhuo M, Huang YY, et al. (1995) Hippocampal long-term depression and depotentiation are defective in mice carrying a targeted disruption of the gene encoding the RI beta subunit of cAMP-dependent protein kinase. *Proceedings of the National Academy of Sciences of the United States of America* 92: 8851–8855.
2. Castillo PE, Janz R, Südhof TC, Tzounopoulos T, Malenka RC, and Nicoll RA (1997) Rab3A is essential for mossy fibre long-term potentiation in the hippocampus. *Nature* 388: 590–593.
3. Castillo PE, Schoch S, Schmitz F, Sudhof TC, and Malenka RC (2002) RIM1alpha is required for presynaptic long-term potentiation. *Nature* 415: 327–330.
4. D′Adamo P, Wolfer DP, Kopp C, Tobler I, Toniolo D, and Lipp HP (2004) Mice deficient for the synaptic vesicle protein Rab3a show impaired spatial reversal learning and increased explorative activity but none of the behavioral changes shown by mice deficient for the Rab3a regulator Gdi1. *European Journal of Neuroscience* 19: 1895–1905.
5. Fernandez-Chacon R, Konigstorfer A, Gerbe, SH, et al. (2001) Synaptotagmin I functions as a calcium regulator of release probability. *Nature* 410: 41–49.
6. Hensbroek RA, Kamal A, Baars AM, Verhage M, and Spruijt BM (2003) Spatial, contextual and working memory are not affected by the absence of mossy fiber long-term potentiation and depression. *Behavioral Brain Research* 138: 215–223.
7. Huang YY, Kandel ER, Varshavsky L, et al. (1995) A genetic test of the effects of mutations in PKA on mossy fiber LTP and its relation to spatial and contextual learning. *Cell* 83(7): 1211–1222.
8. Huang YY, Zakharenko SS, Schoch S, et al. (2005) Genetic evidence for a protein-kinase-A-mediated presynaptic component in NMDA-receptor-dependent forms of long-term synaptic potentiation. *Proceedings of the National Academy of Sciences of the United States of America* 102: 9365–9670.
9. Powell CM, Schoch S, Monteggia L, et al. (2004) The presynaptic active zone protein RIM1alpha is critical for normal learning and memory. *Neuron* 42: 143–153.
10. Qi M, Zhuo M, Skålhegg BS, et al. (1996) Impaired hippocampal plasticity in mice lacking the Cbeta1 catalytic subunit of cAMP-dependent protein kinase. *Proceedings of the National Academy of Sciences of the United States of America* 93(4): 1571–1576.
11. Schoch S, Castillo PE, Jo T, et al. (2002) RIM1alpha forms a protein scaffold for regulating neurotransmitter release at the active zone. *Nature* 415: 321–326.

any molecular accommodations in the synaptotagmin 1 R233Q-knockin mice, the decreased P_r remains, suggesting that synaptotagmin 1 R233Q-knockin mice can accommodate at the synaptic or circuit level to dramatic alterations in their P_r set point.

Other possibilities remain to explain the behavioral abnormalities in RIM1α-deficient mice. These include increased PTP at excitatory synapses and altered PPD at inhibitory synapses. Furthermore, the decreased P_r in RIM1α-KO mice may differ from that of synaptotagmin 1 R233Q-knockin mice. For example, it is possible that the synaptotagmin 1 mutation affects inhibitory and excitatory synapses equally, whereas the loss of RIM1α may affect excitatory out of proportion to inhibitory synapses. Such differential regulation of neurotransmitter release between excitatory and inhibitory synapses has been described for other presynaptic proteins such as Munc13–1 and Munc13–2. In addition, it may be that the multiple synaptic abnormalities in RIM1α-KO mice interact to lead to the behavioral deficits, although this conclusion requires excluding individual synaptic abnormalities as a cause. Finally, although LTP in area CA1 of the hippocampus was reportedly normal in RIM1α-KO mice, this experiment was performed in the presence of γ-aminobutyric acid (GABA) receptor antagonists using a very strong 3 s 100 Hz LTP induction paradigm. Thus, it will be of interest to examine earlier phases of LTP and long-term depression (LTD) in area CA1 of the hippocampus in RIM1α-KO mice more thoroughly before examining additional possible physiological explanations for the hippocampus-dependent learning deficits.

In fear conditioning, RIM1α-KO mice were abnormal in both context-dependent and cue-dependent fear conditioning. The abnormal cue-dependent fear conditioning implicates RIM1α in the amygdala function in addition to its role in hippocampal learning. Thus, it is of particular interest to examine LTP in the amygdala of the RIM1α-KO mice. This is a particularly interesting proposition in light of recent data suggesting that some forms of heterosynaptic LTP in amygdala are postsynaptically induced and presynaptically expressed. Indeed, the initial studies of cortico-amygdala LTP in RIM1α-KO mice suggested a trend toward decreased LTP expression in these mutants.

The loss of RIM1α leads to learning and memory deficits, and many questions remain. Is RIM1α modulated during learning *in vivo*? Which RIM1α binding proteins might be involved in learning and memory that is mediated by RIM1α? And could their function in learning and memory be controlled by modifications such as phosphorylation? Does the loss of RIM1α in a particular brain region correlate with the learning and memory deficits that we observed in the RIM1α-KO mice? These are questions currently under active investigation.

RIM1α-KO mice also display increased locomotor responses when exposed to a novel environment. Locomotor habituation was normal, suggesting that RIM1α-KO mice are not constitutively hyperactive but that they have abnormal responses to the mild stress of a novel environment. This is a phenotype often observed in rodent models of schizophrenia and may mimic the psychomotor agitation seen in some human schizophrenic patients. Additional behavioral characterization of RIM1α-KO mice in other schizophrenia-relevant behaviors such as sensorimotor gating, response to psychotomimetics, and social interaction are under investigation.

In summary, the behavioral data show that RIM1α is essential for normal learning and memory, but it remains unclear which of the physiological deficits in these mice causes the behavioral abnormalities. To understand RIM function *in vivo*, it will be crucial to understand how the various RIM isoforms work together. Furthermore, it will be important to develop strategies that allow for conditional deletion of RIM molecules in selective brain regions. With the strong deficits in synaptic transmission that are observed when RIM1α is absent, it might be possible in future to use RIMs as a tool to gain further insight into how certain areas of the central nervous system work together and how the performance of the central nervous system is affected when RIM-dependent parameters of neurotransmission fail at specific subclasses of synapses.

Overview

RIM proteins are key presynaptic active zone molecules important for normal synaptic transmission and behavior. Future studies will yield additional insights into how RIM proteins function differently in synaptic transmission and behavior at different synapse types. A more precise understanding of the physiological roles of the various RIM protein interactions as well as the functional redundancy across RIM isoforms is a critical frontier in the study of synaptic function.

See also: Calcium Channel Subtypes Involved in Neurotransmitter Release; Cortical Plasticity and Learning: Mechanisms and Models; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Learning and Memory in Invertebrates: *C. Elegans*; Learning and Memory in Invertebrates: *Drosophila*; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Neurotransmitter Release from Astrocytes;

Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Plasticity: Short-Term Mechanisms; Synaptic Vesicles.

Further Reading

Calakos N, Schoch S, Sudhof TC, and Malenka RC (2004) Multiple roles for the active zone protein RIM1alpha in late stages of neurotransmitter release. *Neuron* 42: 889–896.

Castillo PE, Schoch S, Schmitz F, Sudhof TC, and Malenka RC (2002) RIM1alpha is required for presynaptic long-term potentiation. *Nature* 415: 327–330.

Kaeser PS and Sudhof TC (2005) RIM function in short- and long-term synaptic plasticity. *Biochemical Society Transactions* 33: 1345–1349.

Koushika SP, Richmond JE, Hadwiger G, et al. (2001) A post-docking role for active zone protein Rim. *Nature Neuroscience* 4: 997–1005.

Powell CM (2006) Gene targeting of presynaptic proteins in synaptic plasticity and memory: Across the great divide. *Neurobiology of Learning and Memory* 85: 2–15.

Powell CM, Schoch S, Monteggia L, et al. (2004) The presynaptic active zone protein RIM1alpha is critical for normal learning and memory. *Neuron* 42: 143–153.

Schoch S, Castillo PE, Jo T, et al. (2002) RIM1alpha forms a protein scaffold for regulating neurotransmitter release at the active zone. *Nature* 415: 321–326.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Wang Y, Okamoto M, Schmitz F, Hofmann K, and Sudhof TC (1997) Rim is a putative Rab3 effector in regulating synaptic-vesicle fusion. *Nature* 388: 593–598.

Radial Glial Cells: Brain Functions

P Rakic, Yale University School of Medicine, New Haven, CT, USA

© 2009 Published by Elsevier Ltd.

Introduction

Radial glial cells (RGCs) are a morphologically, biochemically, and functionally distinct nonneuronal cell class that, during development, radially spans the entire width of the developing cerebral wall from the ventricular cavity to the pial surface. They play several pivotal roles in the production and placement of neurons during vertebrate brain development. Since the discovery of RGCs in the human fetal cerebral wall, this specialized cellular population has been identified in most regions of the developing vertebrate brain, ranging from the spinal cord to the telencephalon (Figure 1). RGCs have a bipolar shape with oval nuclei and one shorter, inner (apical) process terminating at the ventricular surface and the other, much longer outer (basal) process terminating at the cerebral surface. The phenotypic distinction of RGCs from the more specialized neuronal progenitors in the proliferative zones and from the postmitotic migrating neurons in the intermediate zone was initially based on cytological and ultrastructural analysis of the primate fetal forebrain. This distinction has been supported by comparative studies of their molecular and physiological characteristics in other species. In addition, *in vivo* and *in vitro* approaches revealed that these specialized embryonic cells generate neuronal cell lines that either immediately or after several divisions migrate along radial shaft processes of the mother cells which span the expanding cerebral wall. However, the size, longevity, and significance of RGCs for guiding migrating neurons have increased with the evolutionary expansion of the neocortex, reaching a peak in the size and phenotypic differentiation in the gyrencephalic human forebrain. The stability of the scaffolding of RGCs which contains transiently a subpopulation of nondividing RGCs may be an essential evolutionary adaptation to enable proper allocation of neurons to the expanded and convoluted primate cerebral cortex. In most mammals and, with some exceptions, other vertebrates during the late stages of ontogenetic development, the majority of RGCs disappear by transformation into various forms of astrocytes. A subpopulation of RGCs enter the dormant stage in the subependymal zone and can be activated by various insults to produce reactive gliosis or generate glial brain tumors (glioma). The multiple functions of RGCs and their species-specific adaptations indicate a pivotal role in evolution, development, and pathology of the cerebral neocortex.

History and Definition

RGCs impregnate well with most variants of the Golgi silver impregnation method, leading to their early discovery in the human fetal brain during the late nineteenth century. Although similar cells have been observed in most subdivisions of the vertebrate central nervous system, most attention has been given to the elongated RGCs of the telencephalon. In initial studies of the developing telencephalon in various mammalian species, Ramón y Cajal and his contemporaries recognized the distinct characteristics of these cells, with an apical endfoot firmly embedded at the ventricular surface and basal processes that at later stages become studded with lamellate expansions and terminate with branches into several endfeet, forming the outer, subpial cerebral surface, also called the 'glia limitans.' Because of these morphological characteristics, they were classified as a separate cell class that is different from the 'pyriform' neuroblasts which give rise to bipolar neurons migrating to the cerebral cortex. In addition, Ramón y Cajal suggested that RGCs eventually transform into 'spiderlike' cells that differentiate into fibrillary or protoplasmic types of astrocytes of the white and gray matter of the cerebrum, respectively. The investigators of that era also identified additional subclasses of modified RGCs, such as 'tanycytes,' 'faserglia,' 'Müller cells,' and 'Bergmann glia,' that persist in some regions of the adult brain. Although the methods available at the time were not adequate to suggest a specific function or provide definitive phenotypic classification, based on the morphological criteria, most investigators have considered them as 'primitive ependyma' or embryonic or fetal 'glia,' and this term has endured the test of time.

It is not surprising that the initial discoveries of the RGCs, as well as their classification into the glial phenotype, were made in the embryonic human brain because in primates, these cells have distinct ultrastructural and biochemical properties, including the much earlier expression of glial acidic fibrillary protein (GFAP), compared to rodents (Figure 2). Furthermore, a subset of RGCs in the fetal primate telencephalon do not divide for several months while serving as a guide for the curvilinear migratory pathway in the large, gyrencephalic primate forebrain. Thus, RGCs may have evolved to mature early in order to provide stable scaffolding for the formation of the convoluted primate cortex. Although it may

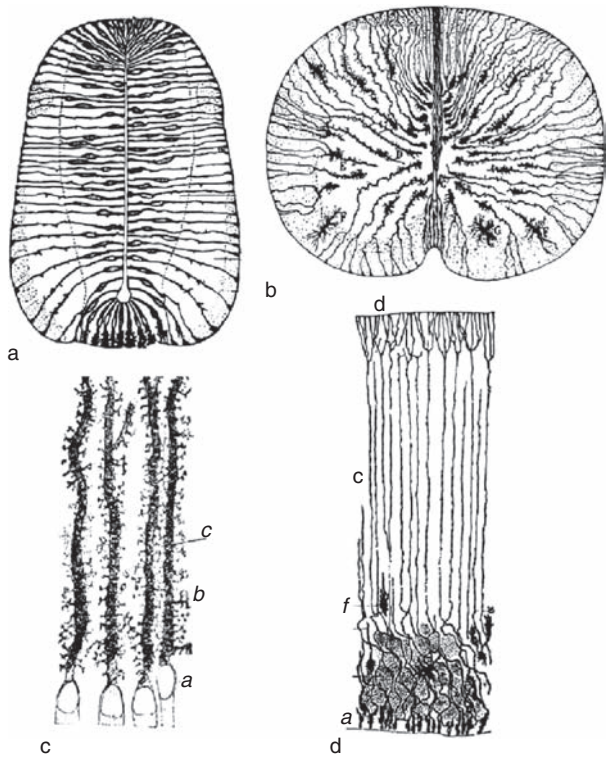

Figure 1 (a) Primordial epithelium, including spongioblast (radial glial) cells of the spinal cord of the chick embryo on the third day of incubation when, according to Ramón y Cajal, they become stainable by the Golgi method. (b) Epithelial cells (radial glial cells) and displaced astroglial precursors and more mature astrocytes in the spinal cord of the newborn mouse stained by the Golgi method display different phases of morphogenetic transformation of a more differentiated glial cell line. (c) Characteristic lamellate expansion on the radial shafts of epithelial (radial glial). (d) Epithelial (radial glial) and neuroglial cells of the cerebral cortex in neonatal rabbit stained using the Golgi method. Adapted from Ramón y Cajal S (1909) *Histologie du Système Nerveux de l'Homme et des Vertébrés*, vol. 1. Paris: A. Maloine. (Reprinted in 1952 by Consejo Superior de Investigaciones Cientificas, Instituto Ramón y Cajal, Madrid).

seem surprising that nonneuronal cells play such a pivotal role for placing cortical nerve cells into their proper locations, it should be recognized that the vertebrate central nervous system is an elaboration of the ectoderm of the dorsal surface of the embryo that initially consists of the epithelial cells, and only after closure of the neural tube do they transform into neuroepithelial cells lining the ventricular cavity.

The major advance in defining the nature of RGCs came with the introduction of new methods in developmental, cellular, and molecular biology, such as electron microscopy, ^{3}H-thymidine autoradiography, immuncytochemistry, transgenic technology, and retroviral gene transfer methods which provide ever higher resolution of cellular and molecular events, allowing for more reliable identification of cell classes. These studies have confirmed the glial phenotype of RGCs in all vertebrates, including humans and nonhuman primates, by their immunoreactivity to a variety of cell markers, such as GFAP, RC1, RC2, vimentin, Rat 401, and Ran-2.

Radial Glial Cells as Neuronal Stem Cells

Classical histological methods indicated that RGCs which arise from nuroepithelial cells retain their capacity to divide and may serve as precursors that directly or indirectly give origin to all major classes of neurons and astrocytes. Studies using more advanced *in vivo* and *in vitro* methods have shown that RGCs can generate not only neuronal progenitors but also, directly, postmitotic neurons (Figure 3). Modern studies also confirmed the existence of dedicated neuronal progenitors in the ventricular zone (VZ)/subventricular zone (SVZ) that are derived from the parent RGCs and do not inherit the basal fiber that terminates at the pial surface. Use of the retroviral gene transfer method, which enables tracing cell lineages in the developing mammalian telencephalon, shows a complex divergence of basic cell types. The use of specific cellular markers, combined with computer-assisted serial electron microscopy (EM) cell reconstruction, *in utero* electroporation, and time-lapse multiphoton imaging, revealed that the embryonic cerebral wall contains multiple cell classes, including dividing and nondividing RGCs, postmitotic, radially and tangentially migrating cells, and stationary neuronal and glial cells that are organized into transient layers (Figure 3). Various classes of neural stem cells and neuronal and glial progenitors can be isolated from the VZ/SVZ of both the rodent and the human cerebrum, and their migratory pattern and phenotypes can be followed in slice preparations of mouse, monkey, and human fetal tissue as they transform into different classes of projections and local circuit neurons. Eventually, fetal RGCs that remain attached to the ventricular surface transform into ependymal cells, and those which detach become astrocytes and/or glioblasts (Figure 3). Thus, RGCs give rise to both neuron and glial progenitors that can each produce several generations of dedicated progenitors before their terminal differentiation.

Role in Neuronal Migration

Another important role of the RGCs is to provide scaffolding and guides for neurons migrating from the proliferative VZ/SVZ to the increasingly distant cortical plate. Initially, the reconstruction of the relationship of the RGCs and migrating neurons was based on the use of a combination of Golgi impregnation, ^{3}H-thymidine autoradiography, and reconstruction

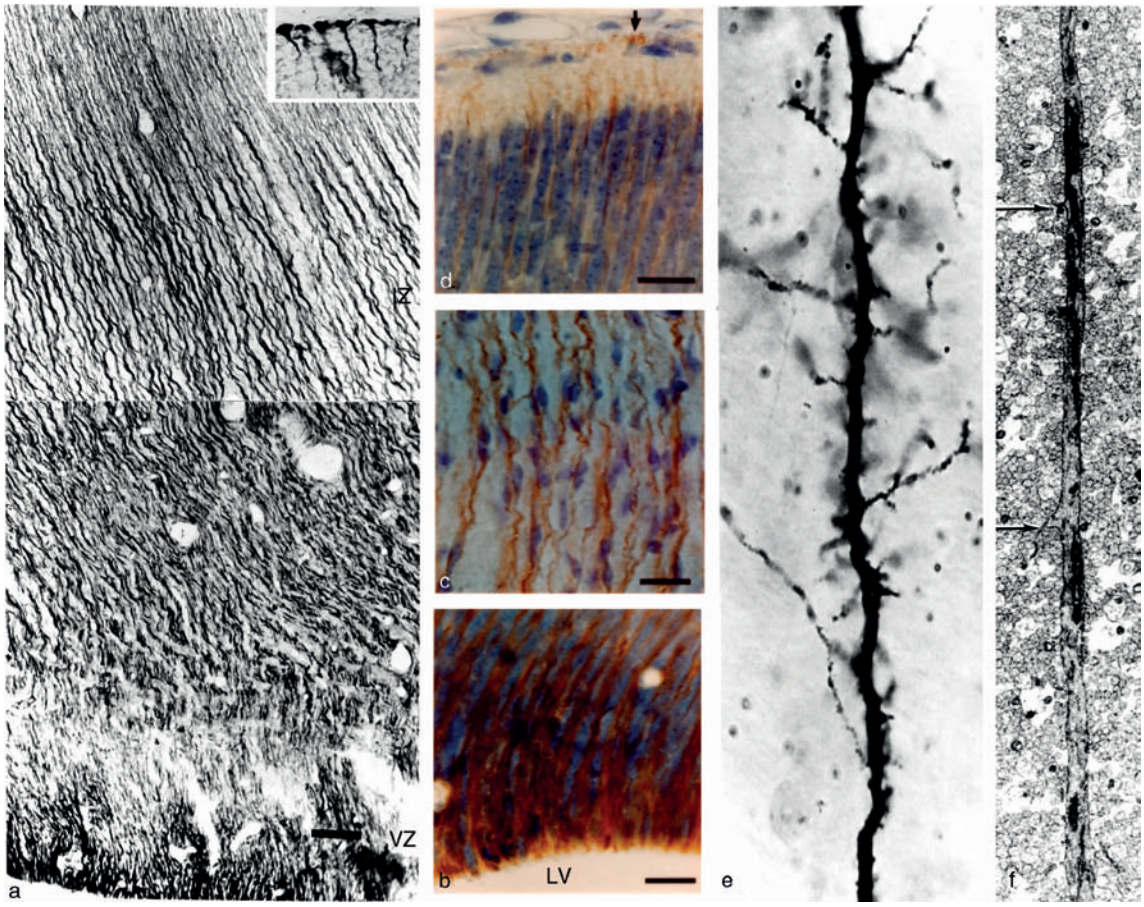

Figure 2 (a–e) Radial glial (RG) cells in the E70 monkey fetal cerebrum immunostained with the antibodies to GFAP. (a) Photomicrograph of a portion of the fetal cerebral wall exposes the richness of the RG fiber scaffolding at the midpoint of cortical neurogenesis. The radial fibers continue to run from the ventricular zone near the lateral cerebral ventricle (LV) across the intermediate zone (IZ) all the way to the cortical plate (not shown), where many of them terminate with the typical endfeet at the pial surface (inset). (b–d) Verichrome photographs of the GFAP-stained RG fibers taken from the three levels of the cerebral wall, starting from the surface of the lateral ventricle (LV) in (b), across the intermediate zone (c), to the pial surface (d). The GFAP-negative migrating neurons (blue) are distinguishable from the GFAP-positive RG cells (brown). (e) RG fibers impregnated using the Golgi method and viewed with the electron microscope. (f) Numerous elaborate lamellate expansions that become more prominent at later stages of corticogenesis. Adapted in part from Levitt P and Rakic P (1980) Immunoperoxidase localization of glial fibrillary acid protein in radial glial cells and astrocytes of the developing rhesus monkey brain. *Comparative Neurology* 193: 815–840.

from EM serial sections in the macaque cerebrum which revealed that cohorts of bipolar postmitotic cells originating in the same sites within the proliferative mosaic of the VZ follow a radial pathway consisting of single or, more often, multiple RGC fibers. Since newly born neurons migrate along the radial fiber of the mother cell, which remains attached to the ventricular surface, the daughter cells are guided by the radial fibers of their mother's cells to the appropriate location in the developing cortical plate (Figure 4). In the human fetal cerebrum, during midgestation a single RG fiber may simultaneously guide several generations of migrating neurons (Figure 5). Although postmitotic neurons in the large gyrencephalic primate brain may require several weeks to reach their final destinations, clonally related cells that originate at the same area in the VZ/SVZ follow the same guide and eventually settle within the same cortical column. The regularity of the well-defined migratory streams (Figure 4) suggests both the stability of the RGCs and the existence of a differential binding affinity mediated by heterotypic, 'gliophilic' adhesion molecules present on opposing neuronal and RG cell surfaces in contrast to cells which move tangentially along axonal tracts using 'neurophilic' adhesion.

Time-lapse imaging of migrating neurons in slice preparations shows that the leading process extends along the shafts of the RGCs. It was shown that specific and separate sets of molecular classes may be involved in the recognition, selective adhesion, and maintenance of neuron–glial interactions during extension of the leading process. Among the proliferating number of

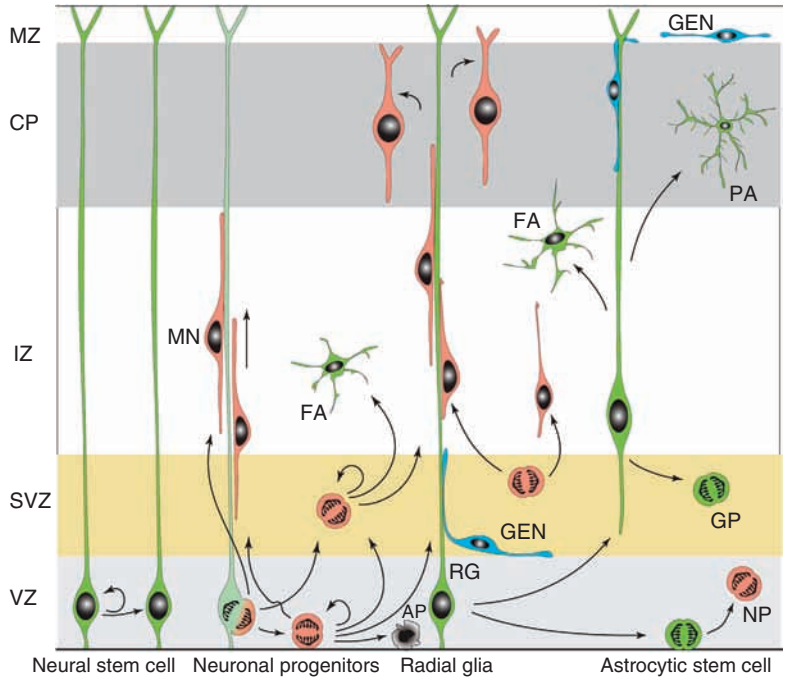

Figure 3 Schematic diagram of the composition, cellular events, and the relationship in the developing cerebral wall, based on data obtained from species ranging from the mouse to human and nonprimates. Initially, a homogeneous population of neural stem cells (green), which divide symmetrically, transform into radial glial cells (RG), which divide asymmetrically and over time produce migrating neurons (MN; red) and dedicated neuronal progenitors with short processes (red mitotic divisions) that populate the ventricular zone (VZ) and subventricular zone (SVZ) and produce all projection neurons as well as the majority of interneurons in humans and a small fraction in rodents. In addition, a population of tangentially migrating neurons arrives to the dorsal telencephalon from the ganglionic eminence mostly via the SVZ and marginal zone (MZ) to supply the majority of interneurons in rodents and approximately one-third in humans. Eventually, RG undergo apoptosis (AP) or directly or indirectly generate ependymal cells, fibrillary astrocytes (FA), protoplasmic astrocytes (PA), glial progenitors (GP), or astrocytic stem cells that retain a neurogenic potential (NP).

molecules considered to be 'candidate' molecules involved in radial migration are neuregulin, which binds to the glial surface via ErbB2 and -4, and integrins, which provide the optimal level of basic neuron–glial adhesion needed to maintain neuronal migration on the shafts of the RGCs. However, a gliophilic-to-neurophilic switch in preference of adhesive interactions of the developing cortical neurons occurs in the absence of functional α_3 integrins.

Dissolution and Transformation

As originally suggested by Ramon y Cajal, the fetal type of RGCs in most regions of the mammalian brain disappears soon after birth. Studies of GFAP immunohistochemistry in the monkey and human fetal cerebrum confirmed that RGCs become transformed into fibrillary and/or protoplasmic astrocytes. The timetable of the disappearance of RGCs in the primate neocortex, hippocampus, and cerebellum correlates with the emergence of protoplasmic and fibrillary astrocytes (Figure 6). The cellular or molecular events that underlie this transformation are not known. In primary cultures, neuronal cells have been shown to exert an inhibitory effect on glial cell proliferation and, additionally, appear to regulate changes in the astroglial cell shape from epithelial-like to radial or stellate cells. The morphological transformation of GFAP-positive RGCs into classical astrocytic cell forms, as well as alternative forms of RGCs, appears to coincide with the loss of RC1, RC2, and Rat-401 antigens, which are not expressed in adulthood. A correlative analysis of antigenic and morphologic transformation *in vivo* and *in vitro* demonstrates that the morphological transformation of RGCs in rodents occurs concomitantly with a gradual acquisition of GFAP immunoreactivity and with a corresponding loss of RC1 immunoreactivity.

Modified RGCs may be found in some regions of the adult central nervous system where a selective microenvironment may allow for their maintenance. In some structures of the mammalian brain, RGCs adapt to the local functional requirements and spatial conditions and transform into specialized astrocytic cell types (e.g., the Bergmann glial cells of the cerebellum, tanycytes of the hypothalamus, or Müller cells of the retina). These specialized cells retain the basic morphological, immunological, and biochemical features of

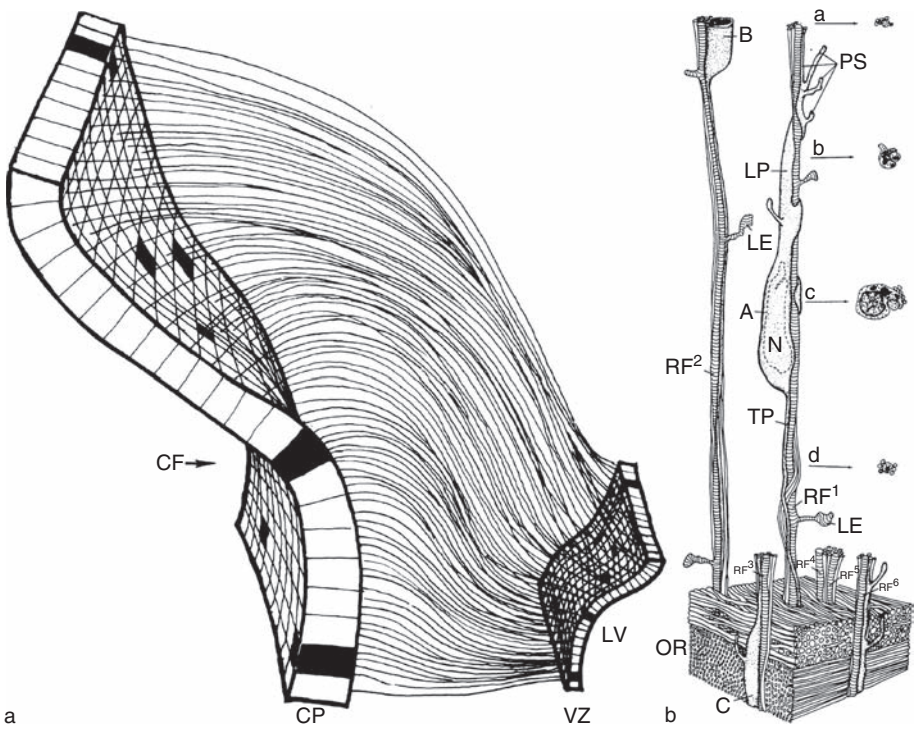

Figure 4 Schematic three-dimensional reconstruction of the portion of the medial cerebral wall at the level of the incipient calcarine fissure (CF) in the 80-day-old monkey fetus. The reconstruction illustrates how the corresponding points in the ventricular zone (VZ) are connected by the array of elongated RGC fibers which span the full thickness of the cerebral wall to the increasingly distant cortical plate (CP) situated below the convoluted pial surface. (b) Three-dimensional reconstruction of migrating neurons based on electron micrographs of semi-serial sections of the occipital lobe of the monkey fetus. The reconstruction was made at the midlevel of the 2500 μm-wide intermediate zone. The bottom portion of the diagram contains uniform, parallel axons of the optic radiation (OR), whereas the irregularly disposed fiber systems occupying the top part of the diagram are deleted to expose the RGC fibers (striped vertical shafts RF_{1-6}) and their relations to the migrating neurons A–C and other vertical processes. The soma of migrating cell A, with its nucleus (N) and leading process (LP), is situated within the reconstructed space, except for the terminal part of the attenuated trailing process and the tip of the vertical ascending pseudopodium. The perikaryon of cell B is cut off at the top of the reconstructed space, whereas the leading process of cell C is shown just penetrating between axons of the optic radiation (OR) on its way across the intermediate zone. LE, lamellate expansions; PS, pseudopodia. Reproduced from Rakic P (1978) Neuronal migration and contact guidance in primate telencephalon. *Postgraduate Medical Journal* 54: 25–40, with permission.

RGCs that include expression of GFAP. Thus, they may be considered as genealogical, morphological, and biochemical descendents of the transient fetal RGCs. However, in some nonmammalian vertebrate species, RGCs may persist throughout the adult life span. Thus, the fate of the RGCs depends on the context and functional requirements, which differ between regions and between species.

Remnants of RGCs in the adult rodent forebrain can produce glial cells throughout life as well as specialized neuronal cell lines that migrate into the olfactory bulb and as hippocampal formations where they are capable of transformation into neurons. The gain-of-function approach has revealed that Notch activity may be involved in cell decision making by promoting glial differentiation and/or inhibiting neurogenesis. The available data also indicate that both cell–cell interactions and cell lineages are likely to determine the fate of RGCs, and under certain conditions they can serve as a multipotent neural stem cell. Thus, while during their transient existence, fetal RGCs contribute actively to brain construction, and the remnants of their progeny may be involved in brain functioning in health and disease during the entire life span.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Enteric Nervous System: Glial Cells and Interstitial Cells of Cajal; Neural Stem Cells and CNS Diseases; Neural Stem Cells: Adult Neurogenesis; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Anton ES, Marchionni MA, Lee K-F, and Rakic P (1997) Role of GGF/neuregulin signaling in interactions between migrating neurons and radial glia in the developing cerebral cortex. *Development* 124: 3501–3510.

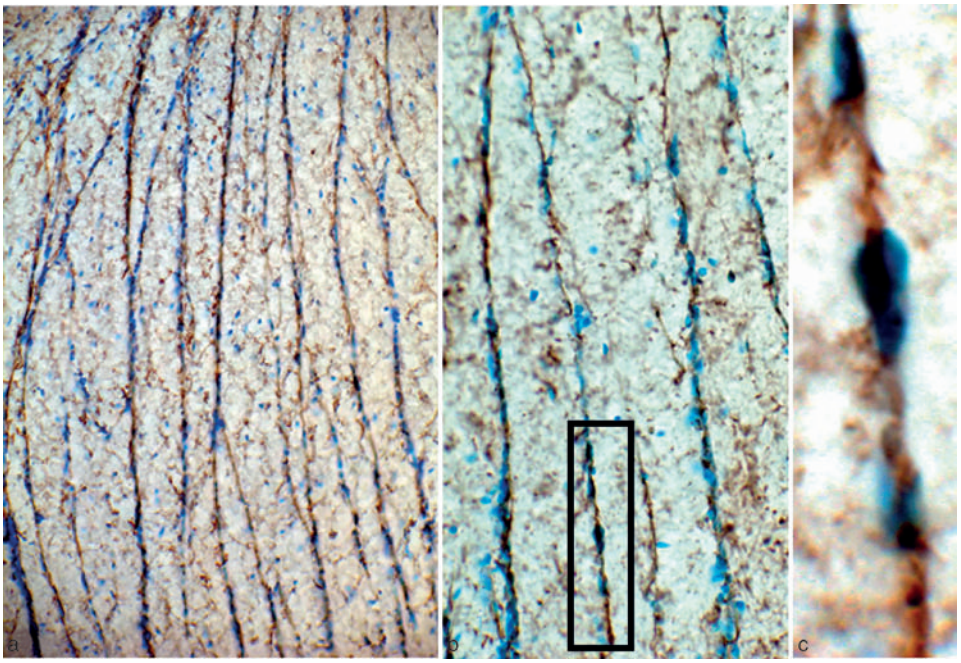

Figure 5 (a) Cohorts of migrating neurons visualized by toluidine blue stain (blue) are aligned along immunostained, GFAP-positive RGC shafts (brown) crossing the portion of the intermediate zone of the frontal lobe of the cerebral wall in the 18-week-old human fetus. (b) The adjacent section counterimmunostained with vimentin (brown) showing rows of similarly aligned vimentin-negative neuron aligned along vimentin-positive RGC shafts. (c) A higher magnification of the segment of the migratory pathway outlined with a rectangle in (b). The image illustrates that the nuclei of the migrating neurons are not enclosed within the cytoplasm of the RG shaft. The migrating cells cannot be reproduced in the same focal plane because they are situated at the opposite sides of the RG shaft. Reproduced from Rakic P (2003) Elusive radial glial cells: Historical and evolutionary perspective. *Glia* 43: 19–32, with permission.

Bentivoglio M and Mazzarello P (1999) The history of radial glia. *Brain Research Bulletin* 49: 305–315.

Cameron RS and Rakic P (1991) Glial cell lineage in the cerebral cortex: Review and synthesis. *Glia* 4: 124–137.

Choi BH (1986) Glial fibrillary acid protein in radial glia of early human fetal cerebrum: A light and electron microscopic immunocytochemical study. *Journal of Neuropathology & Experimental Neurology* 45: 408–418.

Dahl D, Crosby CJ, Sethi JS, and Bignami A (1985) Glial fibrillary acidic (GFA) protein in vertebrates: Immunofluorescence and immunoblotting study with monoclonal and polyclonal antibodies. *Journal of Comparative Neurology* 239: 75–88.

de Azevedo LC, Fallet C, Moura-Neto V, Daumas-Duport C, Hedin-Pereira C, and Lent R (2003) Cortical radial glial cells in human fetuses: Depth-correlated transformation into astrocytes. *Journal of Neurobiology* 55: 288–298.

Gaiano N and Fishell G (2002) The role of notch in promoting glial and neural stem cell fates. *Annual Review of Neuroscience* 25: 471–490.

Hatten ME (1985) Neuronal regulation of astroglial morphology and proliferation *in vitro*. *Journal of Cell Biology* 100: 384–396.

Kadhim HJ, Gadisseux J-F, and Evrard P (1988) Topographical and cytological evolution of the glial phase during prenatal development of the human brain: Histochemical and electron microscopic study. *Journal of Neuropathology & Experimental Neurology* 47: 166–188.

Laywell ED, Rakic P, Kukekov VG, Holland EC, and Steindler D (2000) Identification of a multipotent astrocytic stem cell in the immature and adult mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* 97: 13883–13888.

Levitt P, Cooper ML, and Rakic P (1981) Coexistence of neuronal and glial precursor cells in the cerebral ventricular zone of the fetal monkey: An ultrastructural immunoperoxidase analysis. *Journal of Neuroscience* 1: 27–39.

Levitt P and Rakic P (1980) Immunoperoxidase localization of glial fibrillary acid protein in radial glial cells and astrocytes of the developing rhesus monkey brain. *Comparative Neurology* 193: 815–840.

Malatesta P, Hack MA, Hartfuss E, et al. (2003) Neuronal or glial progeny: Regional differences in radial glia fate. *Neuron* 37: 751–764.

Misson J-P, Edwards MA, Yamamoto M, and Caviness VS (1988) Mitotic cycling of radial glial cells of the fetal murine cerebral wall: A combined autoradiographic and immunohistochemical study. *Developmental Brain Research* 38: 183–190.

Nieuwenhuys R, ten Donkelaar HJ, and Nicholson C (1998) *The Central Nervous System of Vertebrates*. Berlin: Springer.

Noctor SC, Flint AC, Weissman TA, Dammerman RS, and Kriegstein AR (2001) Neurons derived from radial glial cells establish radial units in neocortex. *Nature* 409: 714–720.

Onteniente B, Kimura H, and Maeda T (1983) Comparative study of the glial fibrillary protein in vertebrates by PAP immunohistochemistry. *Journal of Comparative Neurology* 215: 427–436.

Rakic P (1971) Neuron–glia relationship during granule cell migration in developing cerebellar cortex. A Golgi and electron microscopic study in Macacus rhesus. *Journal of Comparative Neurology* 141: 283–312.

Rakic P (1972) Mode of cell migration to the superficial layers of fetal monkey neocortex. *Journal of Comparative Neurology* 145: 61–84.

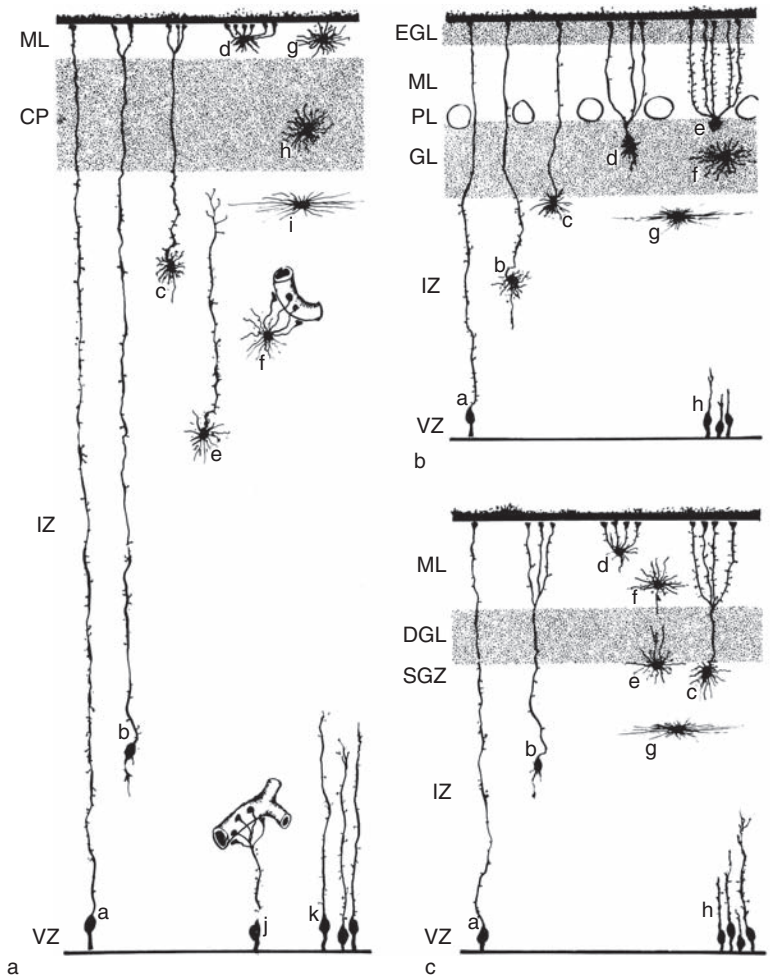

Figure 6 Semischematic diagram illustrating the morphogenetic transformation of fetal RGCs into various astrocytic forms in the cerebral hemisphere (a), cerebellar hemisphere (b), and dentate gyrus of the hippocampus (c). CP, cerebral cortical plate; EGL, external granule layer; GL, granular layer (internal) of the cerebellum; DGL, dentate gyrus–granular layer of the hippocampal region; IZ, intermediate zone; ML, molecular layer; PL, Purkinje cell layer; SGZ, subgranular zone of the dentate gyrus; VZ, ventricular zone. Further explanation and designation of various transitional cell forms (a–k) is provided in the text. Reproduced from Rakic P (1984) Emergence of neuronal and glial cell lineages in primate brain. In: Black IB (ed.) *Cellular and Molecular Biology of Neural Development*, pp. 29–50. New York: Plenum.

Rakic P (1988) Specification of cerebral cortical areas. *Science* 241: 170–176.

Rakic P (1995) A small step for the cell – a giant leap for mankind: A hypothesis of neocortical expansion during evolution. *Trends in Neurosciences* 18: 383–388.

Rakic P (2003) Developmental and evolutionary adaptations of cortical radial glia. *Cerebral Cortex* 13: 541–549.

Ramón y Cajal S (1909) *Histologie du Système Nerveux de l'Homme et des Vertébrés* vol. 1. Paris: A. Maloine. (Reprinted in 1952 by Consejo Superior de Investigaciones Cientificas, Instituto Ramón y Cajal, Madrid.)

Retzius G (1894) Die Neuroglia des Gehirns deim Meuschen und fei Saugethieren. *Biol Untersuchungen* 6: 1–24.

Schmechel DE and Rakic P (1979) A Golgi study of radial glial cells in developing monkey telencephalon: Morphogenesis and transformation into astrocytes. *Anatomy and Embryology* 156: 115–152.

Schmechel DE and Rakic P (1979) Arrested proliferation of radial glial cells during midgestation in rhesus monkey. *Nature* 227: 303–305.

Sidman RL and Rakic P (1973) Neuronal migration with special reference to developing human brain: A review. *Brain Research* 62: 1–35.

Steindler DA, Cooper NGF, Faissner A, and Schachner M (1989) Boundaries defined by adhesion molecules during development of the cerebral cortex: The J1/tenascin glycoprotein in the mouse somatosensory cortical barrel field. *Developmental Biology* 131: 243–260.

Tramontin AD, Garcia-Verdugo J, Lim DA, and Alvarez-Buylla A (2003) Postnatal development of radial glia and the ventricular zone (VZ): A continuum of the neural stem cell compartment. *Cerebral Cortex* 13: 580–587.

Varone SS and Somjen GG (1979) Neuronal–glial interactions. *Neurosciences Research Program Bulletin* 17: 27–84.

Weissman T, Noctor SC, Clinton BK, Honig LS, and Kriegstein AR (2003) Neurogenetic radial glial cells in reptile, rodent and human; from mitosis to migration. *Cerebral Cortex* 13: 550–559.

Reaching and Grasping

U Castiello and A C Pierno, University of Padua, Padua, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Although this article is presented in two main sections – reaching and grasping – this dichotomist organization does not mirror a strict anatomical and functional distinction between reaching and grasping. Early evidence suggested that prehension movements under visual guidance are controlled by two parallel visuomotor channels, one for reaching, the other for grasping. According to this idea, reaching movements entail the medial parietofrontal circuit that comprises the superior parietal lobule (SPL) and the dorsal premotor cortex (PMd), whereas grasping movements entail the lateral parietofrontal circuit involving the inferior parietal lobule (IPL) and the ventral premotor cortex (PMv). Recently, however, this basic dichotomy has been reviewed by several authors. Not only it has been shown that the pathways that link posterior parietal cortex to PMd and PMv partially overlap, but also that both PMd and PMv contain cells encoding the arm direction of movement, as well as proximal and distal representations of the arm. Thus, this new evidence questions the existence of two separate circuits for reaching and grasping and suggests that both circuits are likely to be involved in both actions.

Reaching

Reaching toward objects requires the unraveling of a complex spatial problem, given that different reference systems are involved in the coding of the position of the reaching hand and of the position of the to-be-reached object. In order to perform a successful reaching movement toward an object, signals about limb starting position, eye position, and target location have to be combined and integrated into common, distributed spatial representations. A central role for such integration, played by a neural circuit involving the frontal and the parietal cortex for the control of reaching movements, has been revealed by neurophysiological, neuropsychological, and neuroimaging evidence. This circuit comprises motor areas and regions located in the frontal lobe as well as in posterior parietal regions. The evidence emphasizes the role of the posterior parietal cortex in space perception and guiding actions, indicating that the sensory control of action depends crucially on this region. This fundamental role ascribed to the posterior parietal cortex may be due to its anatomical location. The posterior parietal cortex is localized between the visual cortex in the occipital lobe and the somatosensory cortex in the postcentral gyrus. This makes it very well positioned to receive visual and somatosensory information as well as to send output to premotor and motor areas in the frontal cortex, which is also fundamental for the control of reaching movements.

Reaching in the Monkey

By means of single-unit recording techniques, a number of studies have indicated the presence of fronto-parietal reach-related neurons that have the capacity to integrate different spatial variables relative to the structure of the external world, such as spatial location, distance, physical dimension, and to perform some of the associated visuomotor transformations required to organize successful visually guided movements (Figure 1(a)).

Studies conducted on macaque monkeys have also shown that a number of subareas of the posterior parietal cortex contain neurons that respond during the execution of reaching movements. For instance, activity within a visual area (area V6A; Figure 1(a)) located in the superior parietal lobe of the monkey was found when the animal was instructed to push or pull a lever positioned outside the field of view. Another important finding suggested that eye-position signals influence reaching-related activity not only in parietal areas but also in regions of the frontal cortex such as PMd and PMv, suggesting that frontal areas do contribute to the process leading to target localization in space, and therefore to the transformation from visuospatial to motor coordinates necessary for reaching. In line with this evidence an earlier investigation indicated that frontal and posterior parietal areas share similar functional properties, thus suggesting that the parietal region is not the sole area involved in the generation of motor commands appropriate for carrying the hand to a visual to-be-reached target.

Collectively, results from these studies indicate that coding visually guided reaching requires the integration of multisensory information into common or comparable frames of reference, and that this multimodal fusion relies on the combinatorial properties of parietal and frontal reach neurons.

Another interesting issue that has been taken into account by researchers pertains to the study of the 'early' mechanisms of reaching. Of relevance is the question of at which temporal stage of the

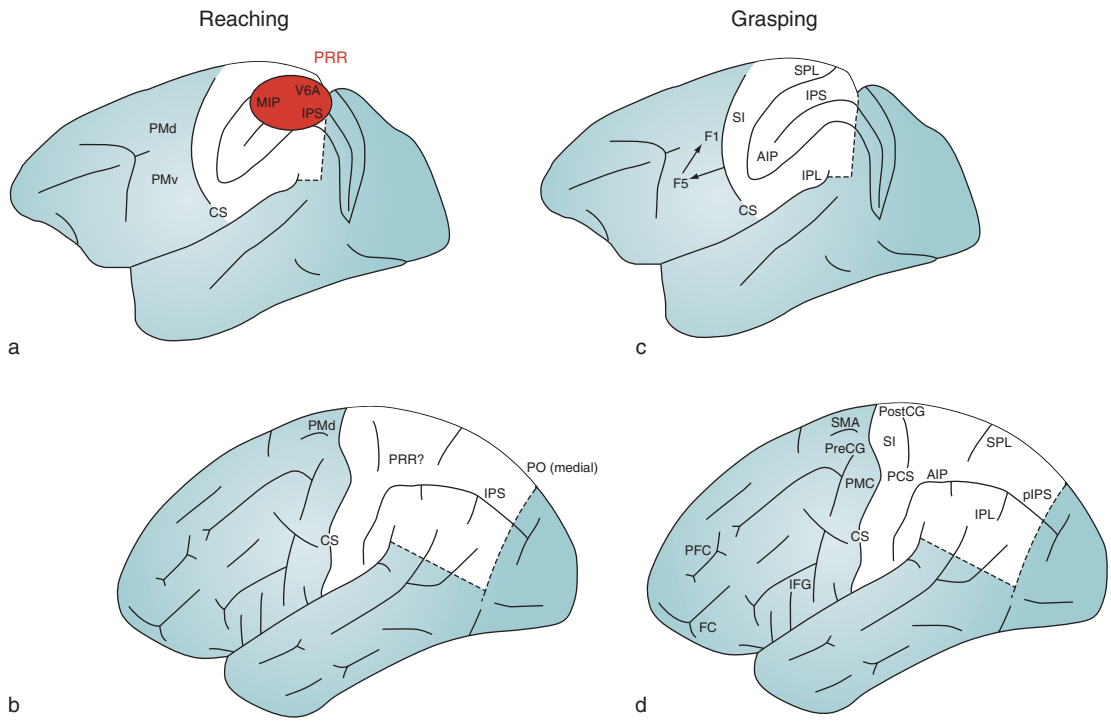

Figure 1 Comparison between neural circuits for reaching and grasping in the macaque monkey and humans. Lateral view of the monkey and human cerebral cortex for both reaching (a and b, respectively) and grasping (c and d, respectively) circuits. (a) Representation of the parietal reach region for reaching as identified in monkeys. (b) Reaching areas in humans as identified by neuroimaging studies. (c) Representation of the visuomotor stream for grasping as identified in monkeys. (d) Grasping areas as identified in humans by neuroimaging studies. Cortical areas that control grasping are also connected with basal ganglia and cerebellar circuits (these circuits, although involved in grasping, are not depicted here). AIP, anterior intraparietal area; CS, central sulcus; FC, frontal cortex; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; IPS, intraparietal sulcus; MIP, medial intraparietal area; PCS, postcentral sulcus; PFC, prefrontal cortex; PreCG, precentral gyrus; PostCG, postcentral gyrus; pIPS, posterior intraparietal sulcus; PMC, premotor cortex; PMd, dorsal premotor cortex; PMv, ventral premotor cortex; PO, parietooccipital; PRR, parietal reach region; SI, primary somatosensory cortex; SMA, supplementary motor area; SPL, superior parietal lobule. Adapted from Culham JG and Kanwisher NG (2001) Neuroimaging of cognitive functions in human parietal cortex. *Current Opinion* in *Neurobiology* 11: 157–163.

information processing flow leading from vision to movements does the convergence of multimodal reach-related signals occur. In a recent study, mechanisms underlying the early formation of motor commands for reaching in the parietal cortex were investigated. Results not only confirmed previous evidence suggesting that neurons within the posterior parietal cortex encode visual and eye-related signals, but also indicated that neuronal activity is modulated by arm-related information at very early stages of the coding of reaching. All in all, these findings suggest that many parameters connected to visually guided reaching, such as visual stimuli, oculomotor behavior, and arm motor behavior, affect neuronal activity at an early stage of the visuomotor processing.

Reaching in Humans

Neuropsychological evidence The role of the posterior parietal cortex in reaching has also been documented by neuropsychological studies conducted on patients suffering from optic ataxia, a specific disruption of visually guided hand movements in which misreaching occurs for targets presented in the peripheral visual field, but not for targets presented in central vision. Nevertheless, these patients are usually able to saccade to targets or to accurately reach toward targets presented on their own bodies, suggesting that optic ataxia is not a sensory or a motor disorder *per se*. The lesions underlying optic ataxia were classically localized within the parietal cortex, in particular along the intraparietal sulcus (IPS) and sometimes extending into the IPL or SPL. A recent neuropsychological study, however, in which parietal lesions of optic ataxic patients were compared with parietal lesions of patients who were not diagnosed with optic ataxia, indicates that specific parietal areas associated with this disorder are the IPS, the medial occipitoparietal junction, the superior occipital gyrus, and the SPL (Figure 1(b)).

Neuroimaging and transcranial magnetic stimulation evidence Various studies designed to examine the neural basis of reaching had participants reaching to touch an object with the index finger, reaching to touch an object with the knuckles, or reaching to grasp an object. On the basis of this variability, it may be hypothesized that these tasks differ in their demands and therefore elicit results that are not directly comparable. Although characterized by these methodological problems, results from human neuroimaging studies appear to nicely fit with the aforementioned reported neurophysiological and neuropsychological results. A number of positron emission tomography (PET) studies reported parietal activation contralateral to the hand used to perform the reaching action.

Functional magnetic resonance imaging (fMRI) studies mainly have focused on pointing rather than reaching, and activation in both the medial intraparietal sulcus and the precuneus has been reported. In one study, participants used a joystick to move a cursor to a visual target; a region of activity was found within the medial intraparietal sulcus. Other work provided evidence in favor of a general functional network shared by pointing actions independently of effector (i.e., hand vs. eye). An fMRI experiment examined both execution and preparation of hand-based pointing movements and eye-based pointing (via saccades) movements. Data indicated that similar left hemisphere intraparietal areas were recruited during movement preparation for both eye- and hand-based cued pointing movements, although this activation was slightly greater for hands as compared to eyes. Other recent studies have instead focused on the role of the medial occipitoparietal junction (Figure 1(b)). For instance, in an experiment, participants were requested to reach toward body parts with their eyes closed. Results showed that the level of activity in the medial occipitoparietal junction was maximum for the first executed movements and decreased for subsequent movements, suggesting that this region not only activates in response to reaching toward visual targets but also toward not-visually-available bodily targets.

A more recent study suggested the existence of distinct neural circuits for central and peripheral reaching. Brain activity was measured while healthy participants were requested to reach toward a central or a peripheral visual target. Results indicated the existence of two systems. The first is specific for reaching in central vision and involves a restricted network, including the medial intraparietal sulcus and the caudal part of the PMd (Figure 2). The second system is specific for reaching in peripheral vision and involves the same network, with the addition of two areas: the parietooccipital junction and the rostral part of the PMd (Figure 2). These results indicate that reaching to the peripheral visual field engages a more extensive cortical network than does reaching to the central visual field, and well explain the deficits of peripheral vision showed by patients with optic ataxia.

Using transcranial magnetic stimulation (TMS), a technique that allows temporary and circumscribed virtual lesions to be induced in healthy humans, to

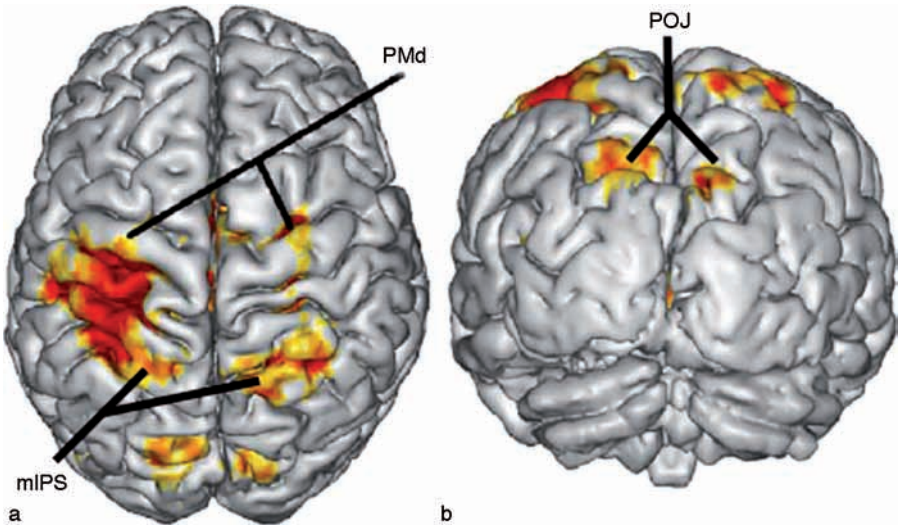

Figure 2 (a) Areas of brain activity within the human neural circuit specific for reaching in central vision, as identified by Prado and colleagues. This circuit comprises the medial intraparietal sulcus (mIPS) and the dorsal premotor cortex (PMd). (b) Areas of brain activity within the parietooccipital junction (POJ), an area specific for reaching in peripheral vision. Adapted from Prado J, Clavagnier S, Otzenberger H, Scheiber C, Kennedy H, and Perenin M (2005) Two cortical systems for reaching in central and peripheral vision. *Neuron* 48: 849–858, with permission from Elsevier.

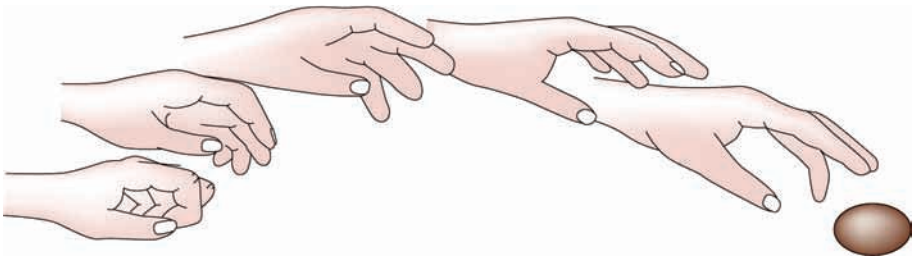

Figure 3 Example of how the hand preshapes during reaching to the target object.

momentarily disrupt the human posterior parietal cortex, indicated a tendency to reach closer toward fixation and the inability to correct (i.e., to compensate for jumps in target location and to learn new movement trajectories). Again this confirms the prominent role played by the parietal cortex for the planning and execution of reaching.

Grasping

Grasping is coded in terms of changes in grip aperture – the separation between the thumb and the index finger and hand-shaping, or the posture assumed by all digits along the reaching action. During a reach-to-grasp movement, there is first a progressive opening of the grip, with straightening of the fingers, followed by a gradual closure of the grip until it matches the object's size (Figure 3). As discussed in the following sections, a surge in research on human prehension is forming the basis for the unraveling of the neural circuits underlying this action.

Grasping in the Monkey

By means of single-cell recordings three specific areas involved in grasping have been uncovered in the cortex of the monkey brain: the primary motor cortex (F1), the premotor cortex (F5), and the anterior intraparietal sulcus (AIP; Figure 1(c)). In terms of neural mechanisms, performing a successful grasping action depends primarily on the integrity of the primary motor cortex (F1). Lesion of this area in monkeys produces a profound deficit in the control of individual fingers and consequently disrupts normal grasping.

Another fundamental process for a successful grasp involves a transformation of the intrinsic properties of the object, visually described, into motor actions. As for reaching, a frontoparietal neural circuit seems to be involved in visuomotor transformations for grasping. This circuit is composed of area F5 and AIP. Area F5 forms the rostral part of the monkey PMv. AIP is a small zone within the rostral part of the posterior bank of the IPS that is directly connected with area F5.

On the basis of the functional roles of neurons in areas AIP and F5, a model has been developed in which area AIP provides multiple descriptions of three-dimensional objects for the purpose of manipulation, whereas area F5 is mainly involved in selecting the most appropriate motor prototype from a motor vocabulary – for example, the type of grip that is effective in interacting with a target object. Confirmation that the AIP/F5 circuit is likely to be relevant for grasping has also been derived by reversible inactivation of each of these areas independently in the monkey. Inactivation of either AIP or F5 markedly impaired hand shaping during reaching, and the hand posture was inappropriate for the object size and shape.

Somatic receptors in muscles, joints, and skin provide information regarding the current posture of the hand and its location and orientation with respect to potential targets for grasping. To investigate the conjunction of visual and somatic processing, one study compared the timing of spike trains recorded by single-unit recording in the somatosensory cortex (SI) and the anterior intraparietal cortex of the same animals during a reach-to-grasp task. The response of cells in AIP was influenced by the shape of the target object. Neurons in SI typically responded later than did those in AIP, showing a significant increase in firing rates only after the hand touched the object, and peaking when grasping was secure. Further, SI neurons rarely differentiated the shape of the grasped object in the manner that occurred in AIP neurons.

Grasping in Humans

Neuropsychological evidence The evidence for specialized circuits for grasping in the human brain comes mainly from the neuropsychological literature. As in the monkey, lesions of the human primary motor cortex or corticospinal fibers profoundly disrupt grasping. Such lesions typically lead to grasping movements that are initially characterized by the loss of independent finger movement, although synergistic movements of all fingers (a power grip) remain intact.

In terms of the human–monkey homology, there is also some evidence that the AIP has a specific role in human grasping. In this respect it has been suggested

that the area responsible for grasping is confined to the anterior part of the human IPS, contralateral to the impaired hand (Figure 1(d)). In line with the results obtained from monkeys in which AIP was inactivated, human patients with lesions in AIP had deficits in grasping, whereas reaching remained relatively intact.

It has been found that in reaching to grasp an object, the finger grip aperture of patients with optic ataxia was abnormally large, and the usual correlation between maximum grip aperture and object size was missing. Subsequently, various patients have been described that show specific deficits in the control of grasping after damage to the SPL. Patient V.K., for example, showed an apparently normal early phase of hand opening during attempts to grasp an object, but her on-line control of grip aperture quickly degenerated, resulting in numerous secondary peaks in the grip aperture profile, rather than the single peak typical of a healthy human. Another patient, A.T., with extensive damage to the superior parietal lobe and secondary visual areas, and some damage to the IPL, showed exaggerated anticipatory opening of the fingers with poor correlation with object size, resulting in awkward grasps. However, this deficit was much less marked if neutral 'laboratory' objects, such as wooden blocks, were replaced with familiar objects, such as a lipstick. Thus, in the case of commonly used objects, cognitive cues can be used to determine the size of the object, based on prior knowledge about these objects, suggesting that the meaning attached to an object might modulate classic grasping circuits.

Evidence from neuroimaging and TMS Brain imaging experiments have investigated the existence and localization in humans of cortical circuits for grasping similar to those described in monkeys. Early attempts to use brain imaging to identify the functional anatomy underlying the generation of goal-directed arm movements during the action of reaching and grasping movements in humans were conducted with PET. When participants were asked to reach and grasp illuminated cylindrical objects or to point to the same targets with their right hand, increases in regional cerebral blood flow (compared with looking at the target) were found in many cortical and subcortical areas, including the contralateral motor cortex, premotor cortex (PMC), ventral supplementary motor area (SMA), cingulate cortex, SPL, dorsal occipital cortex, and left parietal operculum.

In another study, participants were asked to touch or grasp one of five cylinders with their right hand. Compared with touch, grasping increased activity in extensive regions of the bilateral PMC, the posterior parietal cortex (PPC), and the prefrontal cortex (PFC; Figure 1(d)). Other investigations have used 'grasping,' 'pointing,' and 'matching' conditions. In the matching condition, participants compared the shape of the presented object with that of the previous one. The comparison between grasping and pointing showed an increased regional cerebral blood flow in the anterior part of the PPC. The comparison between grasping and matching showed activity increase in the cerebellum, the left frontal cortex (FC) around the central sulcus (CS), the medial frontal cortex, and the left IPL (Figure 1(d)).

In an early PET study, participants were tested in a 'grasping' condition that was compared with two 'observation' conditions. In one observation condition, participants observed grasping movements of common objects performed by the experimenter. In the other they simply looked at the object. Grasping with the right hand significantly activated the left supramotor cortical areas, the left SPL, the cuneus bilaterally, the left putamen, and the cerebellum bilaterally.

The results from a study comparing a precision grip action toward a sweet food with the right hand, grasping with the mouth, and simply observing the sweet showed an extensive hand grasping network including the left pre- and postcentral gyri (PreCG and PostCG, respectively), along the midline in the cingulate gyrus, the precuneus, and the IPL bilaterally (Figure 1(d)). When addressing the selection and execution for particular types of grasp, it emerged that for power grips, using the whole hand, and for a precision task, using the thumb and index finger, the PreCG and the PostCG and two sites in the parietal cortex, the SPL and AIP (Figure 1(d)), were activated.

Studies conducted with fMRI provide spatial resolution superior to that achieved in PET imaging, eliminating many of the possible confounds that characterized the earlier PET studies. One fMRI study used an alternating task design, in which grasping of a rectangular object, the orientation of which could vary, was used as the activation condition and pointing toward the object was used as the baseline condition. The results showed a specific activation of the lateral bank of the AIP in grasping tasks, together with activations including the contralateral sensorimotor cortex, bilateral PMC, and the PPC. The activated part of the AIP is the same area in which lesions in humans and chemical deactivation in monkeys lead to grasping deficits (Figure 1(d)). This study was further confirmed by studies that used a region-of-interest approach and an event-related design. For example, participants were presented with a diverse and unpredictable sequence of objects (rectangular shapes of varied length and orientation) using custom equipment termed 'the grasparatus,' and had to reach

toward the long axis of the objects and grasp them using a precision grip. The results provided strong evidence that the AIP contributes to the ability to execute grasping actions toward objects.

Other studies sought to find a neural network for grasping in the human brain similar to those in area F5 of macaques. In three execution conditions, participants executed a type of grasp appropriate for the object that they viewed; imitated a pantomime action they viewed; or imitated a viewed hand grasping an object. In the baseline condition, participants executed a power grip on all trials while viewing a stationary background. One of the objects used was a large object that would normally be grasped with a power grip; the other was small and would normally be grasped by a precision grip. The areas that were activated were consistent with the monkey AIP/F5 visuomotor circuit. The results showed activation of the AIP, the inferior frontal gyrus (IFG) and the premotor cortex during grasping movements (**Figure 1(d)**).

Similarly, when comparing brain activity during a precision grip and a power grip, the activity recorded in the contralateral primary sensorimotor cortex was higher during a power grip than during a precision grip. By contrast, the activity in the ventral PMC, the rostral cingulate motor area, and at several locations in the posterior parietal cortex and the PFC was stronger during the precision grip than during the power grip (**Figure 1(d)**). The power grip was associated predominately with contralateral activity, whereas the precision-grip task involved extensive activation in both hemispheres. These findings indicate that in addition to the primary motor cortex, premotor and parietal areas are important for control of fingertip forces during precision grip. All together, imaging studies indicate that a grasping circuit similar to that of the monkey can be broadly identified in humans, but also that a much wider network of areas is engaged.

Recent work has shown that TMS has the potential to reveal important facts about the neural networks that mediate grasping. For instance, repetitive TMS stimuli applied to the left superior parietal lobe as participants reached to grasp objects that could suddenly change in size indicate that this area might be crucial for the on-line control of grasping. The future might lie in combining TMS with activation data to study functional connectivity during grasping.

See also: Cognition: An Overview of Neuroimaging Techniques; Evoked Potentials: Recording Methods; Finger Movements: Control; fMRI: BOLD Contrast; Multisensory Convergence and Integration; Neural Prostheses for Reaching; Neural Circuitry in the Somatosensory System; Positron Emission Tomography (PET); Posterior Parietal Cortex and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape; Somatosensory Perception; Somatosensory Cortex; Spatial Cognition and Executive Function; Spatial Transformations for Eye–Hand Coordination.

Further Reading

Bennett KMB and Castiello U (1994) *Insights into the Reach to Grasp Movement*. Amsterdam: Elsevier.

Castiello U (2005) The neuroscience of grasping. *Nature Reviews Neuroscience* 6: 726–736.

Cattaneo L, Voss M, Brochier T, et al. (2005) A cortico-cortical mechanism mediating object-driven grasp in humans. *Proceedings of the National Academy of Sciences of the United States of America* 102: 898–903.

Culhan JC and Kanwisher NG (2001) Neuroimaging of cognitive functions in human parietal cortex. *Current Opinion in Neurobiology* 11: 157–163.

Culham JC and Valyear KF (2006) Human parietal cortex in action. *Current Opinion in Neurobiology* 16: 1–8.

Fadiga L and Craighero L (2004) Electrophysiology of action representation. *Journal of Clinical Neurophysiology* 21: 157–169.

Fogassi L, Gallese V, Buccino G, et al. (2001) Cortical mechanism for the visual guidance of hand grasping movements in the monkey: A reversible inactivation study. *Brain* 124: 571–586.

Frey SH, Vinton D, Norlund R, et al. (2005) Cortical topography of human anterior intraparietal cortex active during visually guided grasping. *Cognitive Brain Research* 23: 397–405.

Galletti C, Kutz DF, Gamberini M, et al. (2003) Role of the medial parieto-occipital in the control of reaching and grasping movements. *Experimental Brain Research* 153: 158–170.

Glover S, Miall RC, and Rushworth MFS (2005) Parietal rTMS selectively disrupts the initiation of on-line adjustments to a perturbation of object size. *Journal of Cognitive Neuroscience* 17: 124–136.

Jeannerod M, Arbib A, Rizzolatti G, et al. (1995) Grasping objects: The cortical mechanisms of visuomotor transformation. *Trends in Neuroscience* 18: 314–320.

Karnath HO and Perenin MT (2005) Cortical control of visually guided reaching: Evidence from patients with optic ataxia. *Cerebral Cortex* 15: 1561–1569.

Porter R and Lemon RN (1993) *Corticospinal Function and Voluntary Movement*. Oxford: Oxford University Press.

Prado J, Clavagnier S, Otzenberger H, Scheiber C, Kennedy H, and Perenin M (2005) Two cortical systems for reaching in central and peripheral vision. *Neuron* 48: 849–858.

Rizzolatti G and Luppino G (2001) The cortical motor system. *Neuron* 31: 889–901.

Wing AM, Haggard P, and Flanagan JR (1996) *Hand and Brain: The Neurophysiology and Psychology of Hand Movements*. San Diego: Academic Press.

Reading

A Pollatsek, University of Massachusetts, Amherst, MA, USA
K Rayner, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Reading is clearly one of the most important human activities. Indeed, Abraham Lincoln gave a speech on inventions in which he deemed that reading was the greatest human invention in history because it made civilization and democracy possible. In addition, reading is an important venue in which to study human cognition because it is simultaneously an activity of great importance and one that can be examined in the laboratory with great precision and with minimal perturbation of the normal activity that occurs outside the psychological laboratory. The tool that allows this to happen is a machine that allows one to monitor the location of what the eyes are looking at on a moment-to-moment basis as someone reads text. Thus, although we allude to other laboratory paradigms for studying the processing of written and printed language, this article is largely limited to discussing what has been learned about reading through the study of eye movements.

First, however, we need to make clear how the eyes move as one reads. Although one has the impression that they move smoothly across the page, this is an illusion. That is, when one reads, the eyes stay relatively fixed for periods of time that are typically between approximately 200 and 300 ms and then quickly jump (in approximately 20–30 ms) to a new location. The stable periods are called fixations, and these quick, ballistic, eye movements are called saccades ('saccade' is French for 'jump'). Moreover, vision is essentially suppressed during saccades so that the only useful information is obtained during the fixations. Note that the eyes are not entirely still during fixations and the two eyes are not always pointing at exactly the same letter, but it is useful to regard both of them as fixated on a single letter during a fixation. Any corrections to what we conclude later, where we ignore these two factors, are extremely minor. As one might expect, the large majority of saccades move forward through the text, although regressions (eye movements going backward through the text) are not rare and account for approximately 10–15% of the eye movements in reading. When the reader comes to the end of a line, a large rightward, although not regressive, eye movement (called a return sweep) is made to near the beginning of the next line of text. This process is somewhat complex because the second or third word of the next line is often fixated initially (since large motor movements are often undershoots), quickly followed by leftward saccades to the first word of the next line. For this reason, the detailed studies of reading discussed in the following sections rarely involve analyses of words near the beginning and end of a line of text, and often participants read a single line of text for comprehension.

Eye movement recording has progressed to the point at which it is quite easy to record eye movements extremely accurately. The most precise eye trackers are spatially accurate so that they are no more than approximately one-half a character off and provide a millisecond-by-millisecond record in the time domain. More widely used trackers that are cheaper and easier to use are still accurate to approximately a character and are accurate in the temporal domain up to approximately 5 ms. Thus, with either type of tracker used today, one can be quite sure of what word someone is fixating and can measure the duration of a fixation with an accuracy of approximately 3–5 ms. This precision would be irrelevant, of course, if the location where the eye is pointing had little relation to the cognitive processing occurring in reading. However, the eye position is in fact quite closely linked to the underlying processes of reading.

The Perceptual Span in Reading

The most direct way in which a tight linkage between the location of an eye fixation and the underlying processing in reading has been demonstrated is through the moving window paradigm that was first used in classic experiments by McConkie and Rayner. In this paradigm, the text displayed on the video monitor depends on where the eye is looking. For example, in a one-word window condition, the word that is currently being fixated would be normal, but all the letters in the other words would be replaced by 'garbage' (e.g., random letters). However, when the eyes move and the reader fixates a new word, that word would now be normal on the next fixation and all the other words (including the previously fixated word) would have their letters replaced by random letters (Figure 1). Thus, one can think of this paradigm as one in which the reader is in a custom-tailored 'slide show,' in which he or she sees a new slide (tailored to where he or she is fixating) on each fixation. An interesting fact about this paradigm is that if the equipment is working correctly (and provided that the window is not too small), people are rarely, if ever, aware of the window; it just seems like normal reading.

The logic of the paradigm implies that if on a fixation people were encoding only the word being

```
Reading is x xxxx xxxxxxxxx xxxxx xx xxxxxxx xx xxx xxxxxxx.
    *

Xxxxxxx xx x very important xxxxx xx xxxxxxx xx xxx xxxxxxx.
             *

Xxxxxxx xx x xxxx important skill xx xxxxxxx xx xxx xxxxxxx.
                  *

Xxxxxxx xx x xxxx xxxxxxxxx skill to xxxxxxx xx xxx xxxxxxx.
                             *

Xxxxxxx xx x xxxx xxxxxxxxx skill to acquire in xxx xxxxxxx.
                                       *
```

Figure 1 An illustration of the moving window paradigm. The size of the window in the illustration is two words: the fixated word and the word to the right. The asterisks below the lines indicate the letter that is being fixated on a sequence of successive fixations. As indicated in the text, the size of the window can be defined in terms of either the number of letters or the number of words. In addition, the location of the window can be manipulated so that, for example, a two-word window could include the fixated word and the word to the left.

fixated, then they would read normally with a one-word window. This is not the case, however; in the one-word window condition, people tend to read approximately 200 words per minute, which is approximately 60% of the normal reading rate. However, this result does indicate that processing the word that is fixated is an important part of reading. This raises the question of how large the window has to be in order for reading to be normal. The answer from the original studies by McConkie and Rayner was that if the window extended 3 or 4 character spaces to the left of fixation to 14 or 15 character spaces to the right of fixation, reading was normal (i.e., normal comprehension and no loss in speed). Later experiments, in which the window was defined by words, indicated that in a two-word window condition (the window contained the fixated word and the word to the right) reading was approximately 90% of normal, and in a three-word window condition (the window contained the fixated word and the two words to the right) reading was normal. It thus appears that readers are spending most of their time during a fixation processing the fixated word but spending some time processing the next word and occasionally processing the one after that.

The outer limits of the window to the right are probably defined by visual acuity. That is, although one has the illusion that one can see clearly over the whole visual field, in fact, acuity is greatest in central vision (the fovea) and drops off quite quickly. However, there is no simple way to characterize the limits imposed by visual acuity, because the window is smaller if the text is quite difficult to decode (e.g., if it is very faint or in an unusual font). Moreover, it would depend on the distance of the text from the viewer and the size of the font. However, it is probably best to think of the window as being defined by the number of characters rather than in terms of a measure such as visual angle because there is probably a trade-off in acuity if the distance of the text or the size of the font changes. That is, if one makes the text larger, the letters would be easier to process (all things being equal); however, making the letters larger also forces them to be further from the fovea, and these two factors would compensate (within limits). Finally, it is interesting to note that the perceptual span for readers of languages that are more densely packed than English tends to be smaller than that for English. Thus, in the most extreme comparison, the perceptual span for Chinese readers is approximately one character to the left of fixation to two or three characters to the right.

Another fact about the perceptual span in reading is that it is asymmetric since it extends much further to the right of fixation. However, this is not a hard-wired 'biological' constraint because an experiment performed with bilingual Hebrew–English readers indicated that asymmetry of the window depended on the language being read (Hebrew is written from right to left). When these readers read Hebrew, the perceptual span was asymmetric to the left, but when they read English the span was asymmetric to the right, as with other readers of English. It is also the case that the span for beginning readers and dyslexic readers tends to be smaller than that for skilled readers. This is generally assumed to be due to the fact that beginning readers and dyslexic readers have more difficulty encoding the words in the center of vision than do more skilled readers.

Eye Fixations Reflect Ongoing Cognitive Processing

The results of the moving window experiments described previously suggest that a major part of the

agenda on a fixation is to encode the word that is fixated. A large amount of evidence supports this assertion, most notably that characteristics of the fixated word, such as its length (number of letters), frequency in the language, and predictability from the prior context in a sentence, all affect the amount of time spent on a word. Perhaps the best measure of word encoding is gaze duration, which is the sum of the duration of all fixations on a word from the first time it is fixated until it is exited. This measure, like the others discussed later, is conditional on the word being fixated on the first pass through the text (i.e., it is fixated before a word to the right of it is fixated). Other frequently used measures are first-fixation duration (the duration of the first fixation on a word), single-fixation duration (the duration of the fixation on a word conditional on the word being fixated once on the first pass), and total time (the sum of all fixation durations on a word including regressions back to the word). Another important measure is the probability of skipping a word.

As one might suspect, the properties of words indicated in prior paragraph – length, frequency, and predictability – are highly correlated. Short words tend to be frequent in the language and predictable, especially function words (articles, prepositions, conjunctions, and pronouns). However, experiments can unconfound them. For example, in an experiment examining word frequency, two words matched in length and predictability but varying in frequency are inserted into the same sentence frame. In the typical experiment, a given participant reads one of the two sentences, but the design is counterbalanced so that equal numbers of participants read each sentence. In addition, the sentences are closely matched on 'naturalness' (as judged by other participants). Moreover, even low-frequency words are selected so as to be easily within the vocabulary of all participants. The standard finding is that the average gaze durations on low-frequency words are between 30 and 50 ms longer than those on high-frequency words, with the difference depending on the difference in frequency between the two sets of words. This difference does not merely reflect some very long times on the low-frequency words. When one plots the histogram of gaze durations for the two sets, the difference is seen largely as a shift of the entire distribution. Similar but smaller effects are observed for first- and single-fixation durations on the word.

Similar findings are observed for word length and predictability (when the other factors are equated). That is, more predictable words are fixated for shorter periods of time than less predictable words, and shorter words are fixated for shorter periods of time than longer words. In addition, these variables also influence the probability of skipping a word. Shorter, more frequent, and more predictable words are skipped more often than longer words. The influence of word length on these measures may not seem very important since it makes sense that when there is a longer word to perceive, it takes more time to do so. However, this again shows that eye movements are registering moment-to-moment processing in reading. The skipping data deserve additional comment. The effect of word length on skipping may be largely due to factors other than ongoing cognitive processing. For example, it may be that the eyes may be biased to move forward a distance approximately equal to the average saccade length in reading, so a saccade that is planned to be short will tend to be longer than planned and a saccade that is planned to be long will tend to be shorter than planned. (This idea was touched on when we discussed return sweeps.) Thus, shorter words may tend to be skipped largely because of errors in motor programming rather than because of the ongoing cognitive processing. However, the same reasoning does not apply to skipping more frequent and/or predictable words. The fact that these are skipped more often indicates that the word to the right of fixation is sometimes processed and that this processing influences where the eye moves. We return to this point when we outline our model of reading.

Our discussion to this point has documented that the time to identify a word (as indexed by variables such as frequency, predictability, and length) has a strong influence on how long the word is fixated, and that these variables also influence whether a word is skipped. This suggests that word identification may be the primary engine that drives the eyes forward in reading and also may be the process that largely determines reading speed. We think that this suggestion is reasonable; however, it is an oversimplification. The next two sections discuss research that indicates that much more is going on in reading. The first discusses whether word identification can be viewed as a simple process in which the letters are encoded and readers merely 'look up' a word in a mental dictionary in a process that only involves letters and words or whether other coding is involved, such as encoding of sound or of morphemes. The second discusses the influences of 'higher level' factors such as syntactic parsing; some of these influences are surprisingly immediate.

The Processing of Words

From approximately 1970 to 1990, the literature in cognitive psychology was dominated by a view in which word encoding was a process by which printed

words made contact with an orthographic lexicon – a mental dictionary that was accessed by a particular sequence of letters. The idea was that other codes – such as the sound of the word, the meaning of the word, and the knowledge of what part of speech it is – were accessed only after this initial look-up process. Much of the work in the past 20 years, however, has indicated that the process is more complex.

First, we consider the issue of where and when phonological coding (i.e., access to the spoken form of language) occurs. It is a common belief that skilled readers go directly from print to meaning without accessing the sound of the language. However, the data indicate that this is false and that access to sound occurs quite early in the process of identifying printed words and influences the process quite strongly. In a classic experiment, participants were asked to make meaning judgments on whether a single word presented in isolation was a member of a semantic category (e.g., whether 'maple' is a tree). The participants were both slow and made large numbers of errors (>10%) in judging, for example, that 'beach' is not a tree. Indeed, other experiments demonstrated effects that were almost as large using nonword homophones (e.g., judging that 'tayble' is not furniture). If readers went directly from print to meaning, and only later encoded the sound of a word, this should not happen.

The importance of phonological coding in early stages of word recognition has also been demonstrated in silent reading of text. Many of these experiments use the boundary paradigm, which is similar to the moving window paradigm but only involves a single display change. In it, there is an invisible boundary just to the left of a target word (**Figure 2**), but before the reader crosses the boundary, there is typically a preview different from the target word. However, when the eyes cross the boundary, the preview is replaced by the target word. The sentence always makes sense with the target word and, as with the moving window experiments described previously, readers are almost always unaware of the identity of the preview and of the display change. In early experiments involving the boundary technique, it was demonstrated that if there was letter overlap between the preview and the target, gaze durations were shorter on the target word than if the preview shared no letters with the target. This indicates that some partial information from the preview is extracted that facilitates processing of the word when it is fixated. However, this could either be orthographic (letter) information or phonological information. A later experiment demonstrated that at least part of the information was phonological. It employed previews that were either homophones of the target word or control words that were matched on the number of letters they had in common with the target word (**Figure 2**). It was found that fixation times on the target word were shorter when the homophone was the preview. This indicates not only that phonological coding occurs in reading but also that it occurs early – even before a word is fixated.

```
                    |
John walked by the beech on his way to the store.
        *

                    |
John walked by the beech on his way to the store.
             *

                    |
John walked by the beach on his way to the store.
                      *
```

Figure 2 An illustration of the boundary paradigm. The vertical mark indicates the boundary (which is not visible to the participant). The display change involves only a single word location, and the only display change occurs when the boundary is first crossed. In this illustration, a homophone of the target word is the preview. Other possible previews could be 'bench,' 'broad,' 'bsfhfs,' or 'xxxxx.' In fact, fixation time on the target word is shorter when 'beech' is the preview rather than 'bench' (indicating phonological coding of the preview), but 'bench' is a better preview than any of the others, indicating orthographic coding of the preview.

A second issue in the encoding of words is whether one goes directly to a single 'lexical entry' (be it orthographic or phonological) for all words or whether the meaningful parts of complex words are part of the encoding process. For example, when one sees the word 'unlikely,' does one treat it as a unit, or do the morphemes un- (meaning not) and -likely play a role in the encoding process? One way this has been investigated is to determine whether the frequency of the component morphemes plays a role in the speed of word encoding as well as the frequency of the entire words. In the typical experiment using this paradigm, either the frequency of an individual morpheme is varied or the frequency of the whole word is varied while holding the frequency of the other constant.

The original experiments using this paradigm were done with long (>10 letters) two-component Finnish compound words (e.g., vaniljakastike = vanilla sauce), and quite large effects (>70 ms) were obtained from varying both the frequency of the components and the frequency of the entire compound word. Later experiments using this paradigm with shorter Finnish compounds, however, obtained much smaller effects of the components. Consistent with these findings, only small effects of the components have been found with five- to nine-letter English words. It thus appears that both the whole word and its morphemic components are active

in accessing a complex word, but there is a 'race' between the process that is looking up the word and the process that is constructing it from the parts, and the process that looks it up directly tends to win the race unless the word is fairly long. Similar results have been obtained with English prefixed words, where the frequency of the whole word is most important for shorter prefixed words, and the frequency of the root (e.g., 'likely' in 'unlikely') plays a major role for longer prefixed words.

Finally, it is important to note that other properties of the fixated word influence how long readers look at it. That is, to simplify our previous exposition, we have used the frequency of a word in the written language as the sole indicator of difficulty. However, this is clearly confounded with a number of important variables, such as the frequency in the spoken language and the age at which the word was acquired. Several studies have shown that the age of acquisition is an important variable and has a significant effect even when word frequency is controlled. Another important variable is whether a word has several distinct meanings (lexical ambiguity). Readers tend to look longer at balanced lexically ambiguous words (those whose meanings are approximately equal in frequency) than control words matched on length and frequency in neutral contexts. However, when the context instantiates one of the meanings, there is no difference between the ambiguous word and the control word. Conversely, readers do not look longer at biased ambiguous words (those in which one of the meanings is dominant (i.e., much more frequent)) than control words in a neutral context. Here, the reader apparently takes the dominant meaning and goes with it, but if the subsequent context instantiates the less frequent meaning there is considerable disruption. However, when the preceding context instantiates the subordinate meaning, readers look longer at the ambiguous word than at the control word. These findings clearly demonstrate that the number of meanings of a word and the nature of the context interact in influencing how long readers look at such words.

Processing of Syntax and Global Meaning

Although written text has some ambiguity at the lexical and phonological level, there are also ambiguities at the level of syntax. Thus, for example, a sentence such as, 'While Mary was mending the sock fell off her lap' is ambiguous in that 'the sock' can initially be taken as the object of 'was mending' (in this case, an incorrect parse of the sentence) or as the beginning of a new clause (in this case, a correct parse of the sentence). Numerous studies using syntactically ambiguous sentences have examined readers' parsing strategies, and the resulting eye movement data have been very informative. In these studies, readers are often led 'down the garden path' in that they expect one interpretation, but an alternative interpretation turns out to be correct. It appears that with these types of ambiguity, readers are committed to one interpretation (hence, they can easily be 'garden-pathed'), but when they encounter disambiguating information, they revise their initial parse. Having to deal with the disambiguating information typically produces either very long fixations or regressions back to where the reader would have been led astray (or both).

The global meaning of a passage can also influence eye movements. Thus, readers often need to engage in anaphoric reference (tying together an anaphor such as 'he,' 'she,' or 'they' to a prior antecedent). When there is ambiguity regarding the antecedent, fixation times increase and readers are also sometimes forced to make a regression back in the text. Likewise, clever experimental manipulations have revealed that when readers make inferences as they read, the nature of the inferences can be inferred from the pattern of eye movements.

Models of Eye Movements during Skilled Reading

As should be clear from the previous discussion, there is still much to be learned about how people go from the marks in the page to understanding a novel, an encyclopedia article, or even a relatively simple paragraph. Hence, current models of reading have a more modest goal: they are primarily interested in trying to understand how the cognitive processing of the text is related to the movement of the eyes through the text. As a result, most of the models proposed in the past 10 years have basically assumed that most of the control of eye movements is related to identifying individual words and that there is a secondary role for the higher-order processes discussed in the previous section. That is, in most models, the role for such secondary processes is largely to stop processing when difficulty is encountered (such as in the syntactic misparsing example discussed previously), and that the normal, forward, movement of the eyes is largely guided by word identification. Empirically, this assumption is reasonable given that the characteristics of individual words, such as their frequency in the language, do produce large effects on the fixation durations on those words. In addition, it seems unlikely that the eyes could move three or four times per second if the reader is waiting on each fixation to make sure that the fixated word was a good fit into the ongoing parse of the sentence and into the ongoing discourse structure.

As a result, most models have striven to explain how word identification moves the eye through the text. We concentrate on the E-Z Reader model, which we think is the simplest and illustrates most of the major points. The E-Z Reader model makes the simplest possible assumption about word identification: Words are identified serially – one word at a time. When the identification of one word is completed, covert attention shifts to the next word and processing of that word begins so that processing of most words starts before they are fixated and continues when they are fixated. However, one complication of the model is that there is a slightly different signal to make an eye movement to the next word; a slightly earlier stage of word identification is the signal to fixate the next word. The target of a fixation is the center of the next word. A major motivation for positing the two stages of word processing is that the assumption allows one to explain why some frequency effects spill over to the next word.

One can then easily predict that the frequency and predictability of a word affect the time to fixate that word if one assumes that word identification time is influenced by these variables. However, such a model would predict that all words would be fixated. The mechanism for skipping in the E-Z Reader model was influenced by experiments that indicated that saccades can be canceled and replaced by other saccades if the signal for the second saccade occurs early enough. In the model, this is captured by positing that a program for a saccade is in two stages. That is, after the signal to make a saccade enters the motor system, there is an earlier labile stage in which the program can be canceled, followed by a later nonlabile stage in which the earlier program cannot be canceled. This explains skipping in the following way: If the reader is fixating word_n and has programmed a saccade to word_{n+1} and word_{n+1} is easy to process (e.g., if it is a short function word), then completion of the first stage of processing of word_{n+1} (which produces a command to fixate word_{n+2}) may follow quickly enough after the command to fixate word_{n+1} so that the first program is canceled and the reader goes directly to word_{n+2}.

There are other aspects of the model that handle refixations on words and other phenomena, but we only have space to outline the basic idea of the E-Z Reader model. Most important, the model can not only explain the phenomena on word processing described in the previous section in the qualitative manner but also produces good quantitative fits of the data and has made several nonobvious predictions that have been confirmed. As a result, we think it is a sufficient explanation of the reading process and, if true, allows for a relatively simple interpretation of eye movement data. For example, even in this simple model, one cannot assume that the mean gaze duration time on a word is equal to the mean time to identify the word, but under the assumptions of the model, one can infer that they are closely related. However, these simple assumptions may not be correct, and other models, which can also model eye movement data fairly well, have assumed that words are processed in parallel: two to four words may all be simultaneously processed at a point in time. Some experiments have attempted to distinguish between these assumptions. None, however, have been conclusive.

Summary

Reading is a complex skill that involves coordination of the language processing system and the eye movement system. Data indicate that forward movement of the eyes through the text is largely driven by the identification of words but modulated by other processes such as sentence parsing. On some level, this appears to be amazing because fixations in reading are usually slightly less than 0.25 s, and within this time, a word has to be encoded and a signal sent out to the eye movement system to execute a saccade based on this encoding. This can be accomplished, in part, because the processing of words begins even before they are fixated. Moreover, several models of eye movements can account for many of the phenomena of eye movements in reading while making reasonable assumptions about the time to encode words and the time to execute eye movements.

See also: Language Development; Saccadic Eye Movements; Sentence Production; Sentence Comprehension; Word Recognition; Word Learning.

Further Reading

Andrews S (ed.) (2006) *From Inkmarks to Ideas: Current Issues in Lexical Processing*. Hove: Psychology Press.

Crowder RG and Wagner RK (1992) *The Psychology of Reading*. Oxford: Oxford University Press.

Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124: 372–422.

Rayner K and Pollatsek A (1989) *The Psychology of Reading*. Englewood Cliffs, NJ: Prentice Hall.

Reichle ED, Rayner K, and Pollatsek A (2003) The E-Z Reader model of eye movement control in reading: Comparison to other models. *Brain and Behavioral Sciences* 26: 445–476.

Snowling MJ and Hulme C (2005) *The Science of Reading: A Handbook*. Oxford: Blackwell.

Reasoning and Problem Solving: Models

A K Barbey and L W Barsalou, Emory University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neuroscience research has increasingly provided evidence that informs theories of reasoning. We begin by defining classic forms of reasoning; and summarizing psychological contemporary theories. We then review empirical evidence from neuroscience that bears on these theories. Finally, we summarize current challenges and identify promising areas for future research.

Definition and Types of Reasoning

Reasoning is a hallmark of human thought, supporting the process of discovery that leads from what is known or hypothesized, to what is unknown or implicit in one's thinking. Reasoning can take the form of deductive inference, whereby the evidence guarantees the truth of the conclusion. Alternatively, when reasoning depends on conditions of uncertainty, it takes the form of inductive inference, whereby the evidence provides only limited support for the truth of the conclusion. The deductive inference that a storm has emerged is supported by evidence that it is raining, whereas observing lightning clouds provides only limited support for the inductive inference that it will rain. Many forms of reasoning typically depend on conditions of uncertainty, including problem solving, causal reasoning, and analogical inference.

Problem solving refers broadly to the inferential steps that lead from a given state of affairs to a desired goal state. For example, deciding who to vote for in a presidential election, diagnosing a patient on the basis of observed symptoms, or preparing for a mountain climbing expedition all depend on problem solving. Problem solving often requires the process of planning, namely, formulating a method for attaining a desired goal state. Planning a mountain climbing expedition, for example, requires scheduling the season of the expedition by predicting the consequences of taking the trip during different seasons. This is accomplished by modeling the situation and observing the consequences of possible actions (e.g., the occurrence of snow in winter introduces new challenges).

Prediction and explanation further depend on causal reasoning (i.e., the ability to infer causal relations). For example, knowledge that cold winter weather can cause an accumulation of snow in the mountains, which in turn can cause an avalanche, supports the prediction that an avalanche is possible during winter.

Finally, new predictions and explanations can be generated by analogical reasoning, whereby relations from one domain are mapped onto another domain. For example, students can conceptualize particle motion in an atom through an analogy to the solar system.

Theoretical Perspectives on Reasoning in the Psychological Literature

The psychological literature contains various theories that address the cognitive architecture and symbolic representations that underlie reasoning. After reviewing these theories, we turn to neuroscience evidence that bears on them.

Cognitive Architecture

One major distinction can be drawn between theories that view the mind as containing specialized reasoning modules, versus theories that view the mind as containing general-purpose reasoning systems. According to the modular view, the mind consists of specialized modules that are unavailable to conscious awareness and deliberate control (cognitively impenetrable), and that are only able to process specific types of information (information encapsulation). Advocates of this view have proposed a diversity of modules that underlie reasoning, including modules for semantic inference, communicative pragmatics, social exchange, intuitive numbers, spatial relations, naive physics, and biomechanical motion. If cognitive architecture emerges from an underlying neural architecture, then strong modular views predict that the neural systems for reasoning should be relatively localized, implementing modules that are cognitively impenetrable and informationally encapsulated.

Alternatively, dual-process theorists propose that reasoning is based on two general-purpose systems: an associative system and a rule-based system. The associative system uses basic cognitive operations such as association, similarity, and memory retrieval to produce primitive judgments quickly and unconsciously. The rule-based system reflects more evolutionarily advanced mechanisms that implement reasoning procedures deliberately and consciously. For example, inductive reasoning largely depends on the retrieval and evaluation of world knowledge, whereas deductive reasoning depends on rule-based, formal procedures.

The dual-process theory further motivates the cognitive demand hypothesis: when people have little time and limited processing resources, incentives,

and/or external aids available for making a judgment, they only use the associative system when reasoning. Conversely, when people have more time and greater processing resources, incentives, and/or external aids available, they also use the rule-based system. The dual-process model predicts that reasoning recruits different neural systems, depending on cognitive demand. When tasks are easy, the associative system is sufficient for correct reasoning and primarily recruits left inferior frontal gyrus, especially Broca's area, in addition to the temporal lobes and posterior parietal association cortex. As tasks become more difficult, the rule-based system becomes required for correct reasoning and recruits the prefrontal cortex, especially the ventrolateral subregion, which has been implicated in rule maintenance.

A second theory that provides a general-purpose account of reasoning is grounded in a somewhat different pair of systems: the linguistic system and the conceptual system. According to this view, the brain's language system initially produces relatively superficial information about a reasoning problem, such as word associates, syntactic structures, etc. As linguistic forms become generated, their meanings become increasingly represented in the conceptual system. A key assumption of this approach is that superficial processing based on linguistic representations may often be sufficient for adequate reasoning performance. When it is not sufficient, conceptual representations must be generated to produce more sophisticated reasoning. Thus, consistent with the dual-process model, this view predicts that differences in the cognitive demands of a reasoning task (i.e., superficial vs. deep processing) will differentially engage neural systems.

In general, both dual coding frameworks predict that reasoning will recruit neural systems that support two forms of coding. One important set of systems underlies language processing, including the left frontotemporal language system. A second set of important systems underlies conceptual processing, mental simulation, and imagery, including bilateral sensorimotor areas.

Representation in Reasoning

A second major distinction can be drawn between theories of reasoning that are based on amodal versus modality-specific representations. According to standard theories, the modality-specific states that are active while perceiving an entity are redescribed into amodal representations that bear no correspondence to the neural systems producing them. For example, the concept 'chair' is represented by redescribing the modality-specific states that underlie the perception of a chair into amodal representations (e.g., feature lists, semantic nets), processed by classic symbolic mechanisms (e.g., predication, argument binding). Thus, amodal theories predict that conceptual processing will recruit brain regions outside sensorimotor areas that underlie language and rules, including left prefrontal and superior temporal regions implicated in formal, rule-based operations.

In contrast, embodied theories of knowledge propose that knowledge and meaning are grounded in modality-specific representations. Increasing empirical evidence from both the cognitive and the neuroscience literatures suggests that modality-specific representations underlie higher level cognition. According to this framework, concepts are represented by simulating the modality-specific states that were initially activated during perception, action, and interoception. Embodied theories propose that simulations are organized at a higher level by simulators that integrate information across a category's instances. Over time, for example, visual information about how cakes look becomes integrated in a 'cake' simulator, along with gustatory information about how cakes taste, somatosensory information about how they feel, motor programs for interacting with them, emotional responses to experiencing them, and so forth. The result is a distributed system throughout the brain's modality-specific areas that establishes conceptual content for the general category of 'cake.' Thus, embodied theories predict that conceptual processing will recruit a broadly distributed system of modality-specific brain regions.

Embodied theories further motivate the task specificity hypothesis. According to this hypothesis, the neural areas underlying a particular type of reasoning, such as deduction, may show little in common, as the specific materials and tasks vary. Because different materials and tasks produce different patterns of modality-specific activation, the same general form of reasoning does not show a single, stable pattern, and no common areas may emerge. In contrast to the 'cognitive demand hypothesis,' this view predicts that significant differences in the task and/or materials – even those that do not affect cognitive demand – are likely to dominate neural activation more than the type of reasoning performed.

Mental Models versus Mental Logic

A third major distinction can be drawn between theories of reasoning that are based on visuospatial models versus theories that are based on logical operations. Mental-model theory proposes that deductive and inductive reasoning depend on spatially organized mental models. According to this view, an argument is

evaluated by generating alternative models of its premises, where each model represents possible circumstances that render the premises true. If all the models constructed from the premises are consistent with the truth of the conclusion, then the argument is judged to be deductive or valid. If, however, only a limited number of models render the conclusion true, then the argument is judged to be inductive or probabilistic.

Given the proposed role of mental models in both deductive and inductive reasoning, this theory predicts that these forms of reasoning will recruit common neural systems. Specifically, mental-model theorists have predicted that deductive and inductive reasoning will both primarily recruit right-hemisphere regions. Furthermore, to the extent that mental models represent situations spatially, this view predicts that parietal and occipital regions implicated in visuospatial processing will be engaged.

In contrast, mental logic theory offers an account of deductive reasoning that is based on the application of formal deductive rules according to formal syntactic operations. Thus, rule theorists have predicted that reasoning is likely to recruit left prefrontal and superior temporal regions implicated in formal, rule-based operations.

Empirical Evidence on Reasoning from the Neuroscience Literature

We next present empirical evidence from neuroscience that bears on the psychological theories just reviewed. We begin with research on deductive reasoning, and also address research that compares deductive and inductive reasoning. We then review research on problem solving, causal reasoning, and analogical reasoning.

Deductive Reasoning

The brain systems that implement deductive reasoning depend on whether the reasoning problem consists of familiar versus unfamiliar semantic content. When reasoning about familiar semantic content (e.g., All dogs are pets / All poodles are dogs / Therefore, all poodles are pets), a left frontotemporal system is engaged, including left inferior frontal cortex (BA 47), left middle/superior temporal cortex (BA 21/22), and left temporal pole (BA 21/38). Previous research has implicated this system, not only in deductive reasoning, but also in memory and language tasks that employ familiar semantic content. In general, linguistic processing appears central to all these tasks.

In contrast, reasoning about unfamiliar semantic content (e.g., All P are B / All C are P / Therefore all C are B) activates a bilateral frontoparietal system, including bilateral dorsal (BA 6) and inferior (BA 44) frontal lobes, bilateral superior and inferior parietal lobes (BA 7), and bilateral occipital lobes (BA 19). This pattern of activation is also found during the processing of spatial information, and is similar to the neural activity observed while people make transitive inferences about geometrical shapes.

Thus, deductive reasoning recruits left and right prefrontal cortex asymmetrically as a function of familiarity. Across both familiar and unfamiliar deduction problems, left prefrontal cortex is generally active, suggesting that this region is necessary for deductive inference. Conversely, right prefrontal cortex is engaged only when a problem involves unfamiliar semantic content or a conclusion that conflicts with prior beliefs (e.g., No harmful substances are natural / All poisons are natural / Therefore, no poisons are harmful). Whereas language may often dominate familiar reasoning, language and spatial/visual processing may often be central for unfamiliar reasoning.

The brain systems that implement deductive reasoning also depend on whether a reasoning problem produces correct versus incorrect conclusions. Research on this issue has used inhibitory belief problems, namely, problems whereby individuals must inhibit a highly accessible belief that could interfere with correct reasoning (e.g., No addictive things are inexpensive / Some cigarettes are expensive / Therefore, some cigarettes are not addictive). Drawing a correct conclusion on these problems requires that individuals (1) detect the conflict between their prior beliefs and the logical inference, (2) inhibit the prepotent response associated with their belief bias, and (3) engage the appropriate reasoning mechanisms. In contrast, drawing an incorrect conclusion on these problems results from failing to detect the conflict between beliefs and logical inference, and/or failing to inhibit the prepotent response associated with a belief bias.

When people draw correct conclusions on inhibitory belief problems, right inferior prefrontal cortex becomes active. When they draw incorrect conclusions, ventromedial prefrontal cortex is active instead. Activation in right inferior prefrontal cortex when drawing correct conclusions appears to reflect the detection and/or resolution of the conflict between belief and logic. Conversely, activation in ventromedial prefrontal cortex when drawing incorrect conclusions appears to reflect the role of nonlogical mechanisms, perhaps associated with greater affective processing.

In summary, the neural systems that underlie deduction vary considerably, depending on task factors and

cognitive demand. Consistent with the task specificity hypothesis, the areas that support deduction vary with the familiarity of the materials, and with whether belief violations occur and are detected. Consistent with the cognitive demand hypothesis, more neural areas are recruited for difficult unfamiliar problems than for easier familiar ones. Consistent with the modularity view, left prefrontal cortex generally appears active across most deduction paradigms, suggesting that it is essential for deductive inference.

Deductive versus Inductive Reasoning

Many experiments have asked individuals to perform both deduction and induction in the same experiment, so that the neural circuits underlying these two types of reasoning could be distinguished. One neuroimaging experiment used a categorical syllogism task consisting of three conditions – deduction, induction, and baseline – to assess this issue. In the deduction condition, individuals received a valid or invalid categorical syllogism (e.g., None of the bakers play chess / Some of the chess players listen to opera / Therefore, some of the opera listeners are not bakers). The task was to indicate whether the conclusion was valid (i.e., whether the conclusion was guaranteed by the truth of the premises). In the inductive reasoning condition, individuals only received invalid categorical syllogisms (e.g., Some of the computer programmers play the piano / No one who plays the piano watches soccer matches / Therefore, some computer programmers watch soccer matches). The task was to indicate whether the conclusion was probable or improbable. The deduction and induction conditions were each compared to a baseline condition, in which a categorical syllogism with anomalous semantic content was presented (e.g., All the engineers own a computer / None of the engineers has been to school / Therefore, all the people who own computers are married). Importantly, the categorical syllogisms were fully counterbalanced across individuals so that the same materials occurred in every condition.

Deductive reasoning (deduction minus baseline) produced activation in the left dorsolateral frontal cortex (BA 6), broadly consistent with the left frontal activation observed by other researchers for familiar semantic content. When deduction was compared directly to induction, however, a different pattern emerged, namely, activation in bilateral posterior regions, with a right-hemisphere prevalence, including associative visual cortex (e.g., cuneus, precuneus, middle and superior occipital gyri), as well as right superior parietal lobe (BA 7) and thalamus. These areas have been reported for visuospatial tasks that require form discrimination and imaginative operations. These areas have also been reported for deductive reasoning tasks that employ visuospatial materials. Such activations suggest that people use visuospatial representations, such as Venn diagrams or Euler circles, to support deductive reasoning on categorical syllogisms. In addition, activation was also found in right anterior cingulate (BA 24/32), implicating attention and executive control in deductive reasoning. Inductive reasoning (induction minus deduction) revealed activation in left dorsolateral frontal (BA 8 and 10) and right insular cortices. These regions are known to be involved in probabilistic reasoning tasks that require the estimation of relative frequencies and other quantities (e.g., How fast do race horses gallop?).

A subsequent neuroimaging study extended the methods and design of the study just described to a deduction task that employed conditional statements instead of categorical syllogisms (e.g., If he is an electrician, then he spent two years in night school / He is an electrician and owns a computer / Therefore, he spent two years in high school). In contrast to the categorical syllogisms used in the previous study, visuospatial processing did not seem relevant to these conditional syllogisms. Thus, these researchers predicted that the new task and materials would not recruit brain regions that perform visuospatial processing. Consistent with this prediction, deductive reasoning (deduction minus induction) revealed a major focus of activation in right inferior frontal cortex (BA 44), in right anterior cingulate (BA 24), and in right middle temporal cortex (BA 21). These researchers concluded that this frontotemporal system constitutes a logic-specific network in the right hemisphere that is comparable to the language-specific network in the left hemisphere. Specifically, these researchers argued that this right-hemisphere system implements a calculus of mental transformations that underlie formal deduction.

In a further comparison, inductive reasoning (induction minus deduction) revealed large intense activations in the left inferior frontal (BA 47) and left insular cortices, in addition to left posterior cingulate (BA 31), parahippocampal (BA 36), left medial temporal (BA 35), and superior and medial prefrontal cortex (BA 9). These areas are broadly consistent with the left frontal (BA 8 and 10) and insular areas found in the previous study, known to be involved in the recall and evaluation of familiar world knowledge. Consistent with standard theories of inductive reasoning, the recall and evaluation of familiar world knowledge appear to play central roles during induction.

Another neuroimaging study developed a category learning task to assess the component processes of inductive reasoning. Of primary interest were the neural systems that support rule application versus

rule inference during category learning. Stimuli consisting of novel animals were presented to individuals, who were asked to judge whether all the animals in a set were from the same category. In the rule application condition, a rule was provided that specified the criteria for category membership. In the rule inference condition, individuals had to infer the rule with no instruction. Each condition was further divided into an easy and difficult condition based on the computational demands of the task. These researchers found that rule inference (rule inference minus rule application) preferentially recruited bilateral hippocampus, an area in which activation is modulated by stimulus novelty. In contrast, rule application (rule application minus rule inference) revealed activation in the presupplementary motor area (BA 8). This area is implicated in the anticipation of motor activity and likely reflects an anticipatory response to category exemplars, a response that is absent when the categorization rule is unknown.

In addition to the episodic encoding of novel stimuli, inductive reasoning requires the generation and testing of hypotheses. For example, inferring the basis for category membership requires generating and testing possible rules, such as "has spots on the abdomen" or "has only two appendages" for a category of fictional animals. To assess the neural systems that underlie hypothesis selection, these researchers evaluated the task by difficulty interaction: {hard rule induction minus hard rule application} minus {easy rule induction minus easy rule application}. This comparison assessed the effects of increased difficulty due to subtle variations in the stimulus features of animals that satisfied or violated the category membership rule. The result was an activation in right lateral orbital prefrontal cortex (BA 47 and BA 11), an area implicated in complex reasoning tasks such as analogical and metaphorical transfer. Across different types of reasoning, difficult problems often activate this area.

In summary, current findings from neuroimaging studies of deduction and induction again suggest that the neural bases of reasoning are highly sensitive to the particular tasks and materials employed. For example, deductive reasoning from conditional statements engages a right frontotemporal system, thought to support the application of formal deduction rules, whereas deductive reasoning from categorical syllogisms activates a parieto-occipital system that supports visuospatial processing. Furthermore, the latter result differs from the findings of other researchers, who have found that a left frontotemporal system supports deductive inference in a categorical syllogism task. The observed differences on the same fundamental reasoning process – deduction – probably reflect methodological differences between the tasks and materials used across studies. Furthermore, the fact that different materials are often used on deduction and induction tasks suggests that different activations for these two forms of reasoning may often be driven by differences in materials rather than by differences in reasoning.

Consistent with the task specificity hypothesis, the neural processes that underlie a given type of reasoning (e.g., deduction) may depend more on the tasks and materials used than on the type of reasoning *per se*. Although left prefrontal cortex is often active on deduction tasks (as described in the previous section), it was not active for some of the studies discussed in this section. Perhaps the one consistent finding so far is that reasoning about familiar materials tends to utilize left-hemisphere language and knowledge networks. Conversely, reasoning about less familiar problems tends to utilize bilateral systems that include right-hemisphere mechanisms. This pattern is consistent with theories that postulate two different reasoning systems, one that processes language, and one that processes spatial/visual information.

Problem Solving and Planning

Depending on whether a problem-solving task is well structured or ill structured, different brain systems are engaged. On well-structured problems, such as the Tower of Hanoi, the starting state, the goal state, and possible transformations are specified completely. For example, the starting state consists of three pegs mounted on a platform and three disks of varying sizes stacked in descending order on the first peg. The goal state is to stack the disks in descending order on the third peg. The possible transformations are restricted to moving disks such that (1) only one can be moved at a time, (2) any disk that is not removed must remain on a peg, and (3) a larger disk cannot be placed on a smaller disk. In such tasks, well-structured planning typically recruits left prefrontal cortex, including frontopolar, dorsolateral, and ventrolateral regions.

In contrast, an ill-structured planning problem is specified incompletely. In the Multiple Errands Task, for example, individuals are taken to an unfamiliar neighborhood and asked to complete errands, such as buying a loaf of bread, according to the following rules: you are to spend as little money as possible (within reason) and take as little time as possible (without rushing excessively). You are not to use anything not bought on the street (other than a watch) to assist you. You may perform task steps in any order. In such tasks, ill-structured planning typically activates right dorsolateral prefrontal cortex

(DLPFC), in contrast to the left prefrontal regions that support well-structured planning.

The neural systems underlying planning are further differentiated into subsystems for plan formation versus plan execution. One finding is that DLPFC supports sequential operations during plan formation, whereas medial ventral prefrontal cortex (MVPFC) plays a motivational role in plan execution. More specifically, DLPFC appears to support the generation of hypotheses and the construction of plan steps, whereas the MVPFC appears to support the affective processing required for plan execution (e.g., initiative and determination).

Lesion studies complement the neuroimaging studies just reviewed. One representative study evaluated problem solving in patients with frontal lobe lesions (FLLs) on the Water Jug problem. Patients had to construct unique action sequences that transferred specific quantities of water between jugs of different sizes (e.g., Jar A = 8 units of water, Jar B = 5 units of water, Jar C = 3 units of water). Specifically, individuals had to construct an action sequence that achieved a goal state given by the experimenter (e.g., Jars A and B must contain 4 units of water, and Jar C must contain 0 units). The researchers found that patients with FLLs struggled to make counterintuitive moves that were required to solve the task but that appeared to deviate from the desired goal state. Furthermore, left and bilateral FLL patients were more impaired than right FLL patients were, suggesting that poor performance in the Water Jug task was primarily linked to left DLPFC damage.

In summary, these findings demonstrate that problem solving and planning generally depend on prefrontal systems. Because DLPFC appears important for most problem-solving tasks, modular views receive support. Differences in task conditions modulate the active areas of the prefrontal cortex, however, providing support for the task specificity hypothesis. For example, well-structured versus ill-structured problems rely on left versus right prefrontal regions, respectively. This pattern is also consistent with theories that assume two different systems, such as dual-process and dual-code theories, support reasoning. Finally, distributed theories receive support, given that multiple systems typically underlie problem solving. For example, problem solving requires both planning and execution, and also both reasoning and motivation, with different neural systems supporting each component process.

Causal Reasoning

Several studies have evaluated the neural systems that support the perception of mechanical causation in the classic Michotte launching event. In a launching event, a ball travels horizontally across a computer screen and collides with a ball located in the center. The collision results in the second ball 'launching' away from the first, horizontally, across the screen, thereby eliciting the perception that the first ball caused the second to move. One study compared the neural response produced by the launching event to the response elicited by a control event in which the first ball passed below the second ball without a collision (noncausal condition). Of primary interest were the neural systems engaged when individuals judged either (1) the presence or absence of causation versus (2) the direction of the ball's motion. A reliable increase in medial frontal activation occurred for judgments of causality relative to judgments of ball movement. Moreover, this increase occurred during both the causal and noncausal conditions, suggesting that the signal increase was specifically associated with the process of making a causal judgment, not with the perception of actual causality.

Another study evaluated whether causal perception and causal inference rely on common or distinct hemispheric regions. Two callosotomy (split-brain) patients and a group of neurologically intact patients were tested. Of primary interest was neural activity in the left versus right hemispheres during (1) the perception of causal events (i.e., the Michotte launching event) and (2) causal inference tasks when the relation between a candidate cause and an observed effect had to be inferred (rather than perceived directly). Perception of causality and causal inference depended on different hemispheres of the divided brain. Whereas causal perception engaged the right hemisphere, causal inference engaged the left-hemisphere.

Another study assessed the brain systems for processing evidence that was either consistent or inconsistent with an individual's existing causal beliefs. Individuals received evidence on the effectiveness of drugs designed to relieve depressive symptoms. Two factors were manipulated: the plausibility of the theory that explained the drug's action, and the consistency between theory and data. When individuals reasoned with evidence that was consistent with existing causal beliefs, a network of brain regions widely associated with learning and memory was engaged, including the caudate and the parahippocampal gyrus. In contrast, when they reasoned with inconsistent evidence, a different pattern of activation occurred that is widely associated with error detection and conflict resolution, including the anterior cingulate cortex (BA 24/32), posterior cingulate, and precuneus (BA 7). The researchers concluded that people's beliefs and expectations act as a filter during evidence evaluation. When

evidence is consistent with existing causal beliefs, the neural systems underlying those beliefs implement causal reasoning. When evidence is inconsistent with existing beliefs, a different neural system detects this inconsistency and triggers the construction of a novel causal explanation.

The observed findings demonstrate that causal reasoning is supported by a broadly distributed neural system that is highly sensitive to the causal reasoning task (i.e., causal perception versus causal inference) and the consistency of causal evidence with existing beliefs. These findings provide support for theories of reasoning that are (1) based on distributed rather than localized representations (e.g., dual-process, dual-code, embodied theories), (2) incorporate the role of existing knowledge rather than operating on the basis of purely logical representations (e.g., dual-process, dual-code, embodied, mental-model theories), and (3) advocate the task specificity hypothesis (e.g., dual-process, dual-code, embodied theories).

Analogical Reasoning

Several studies have found that analogical reasoning engages frontopolar cortex. Furthermore, different components of analogical reasoning appear to differentially engage frontopolar versus dorsolateral prefrontal areas. Whereas dorsolateral areas are recruited for processing externally generated information (e.g., the monitoring and manipulation of presented facts), frontopolar areas are recruited additionally for the evaluation and manipulation of internally generated information.

One study assessed the neural systems that support analogical reasoning in the Raven's Progressive Matrices task. Study participants received a 3×3 matrix of figures with the bottom right figure missing, and had to infer the missing figure by selecting one of four possible alternatives. Participants received three types of problems that differed in their degree of relational complexity (0-relational, 1-relational, 2-relational). The 0-relational problems involved no relation of change and thus required no relational processing. The 1-relational problems involved one relation of change in either the horizontal or vertical dimension, and thus required relational reasoning. Finally, the 2-relational problems involved two relations of change, in both the horizontal and vertical directions, and thus required even more relational reasoning.

A region-of-interest analysis was performed to assess the role of prefrontal cortex in processing multiple relations simultaneously (i.e., relational integration). This analysis produced two main findings. First, activation occurred in frontopolar prefrontal cortex (BA 10), reflecting the internal generation of relations required to form complex analogies. Interestingly, this activation only occurred for 2-relational problems, not for 1-relational problems, suggesting that frontopolar cortex is important for processing complex relational structures. Second, activations also occurred in right DLPFC (BA 46), which reflected greater manipulation of externally presented information in more complex problems. Other studies using the Raven's Progressive Matrices task have found similar results, and have also reported bilateral posterior parietal activations (BA 7).

In another study, individuals received a source picture of colored geometric shapes, followed by a target picture of colored geometric shapes. Pictures that did not share similar geometric shapes but that did share the same system of abstract visuospatial relations were also presented. Individuals judged whether each source-target pairing was analogous (analogy condition) or identical (literal condition). Analogical reasoning (analogy minus literal) recruited the dorsomedial frontal cortex (BA 8) and left-hemisphere regions, including frontopolar (BA 10), inferior frontal (BA 44, BA 45, BA 46, and BA 47), and middle frontal (BA 6) cortices, and also inferior parietal cortex (BA 40). These findings suggest that analogical reasoning is mediated by a predominantly left-hemisphere frontoparietal system.

Another study systematically evaluated the component processes of analogical reasoning. Specifically, this study assessed the neural systems that underlie (1) the storage of abstract relations in working memory and (2) the process of integrating abstract relations to form analogies. These researchers also found that analogical reasoning activated a left frontoparietal system, with some regions of this circuit mediating working memory processes and others mediating abstract relational integration. In particular, left frontopolar regions (BA 9/10) were again central to the processing of relations that underlie analogical reasoning.

In summary, the observed findings for analogical reasoning are consistent with theories that advocate distributed rather than localized representations (e.g., dual-process, dual-code, embodied theories). As we saw, distributed frontal and posterior systems appear to play a wide variety of roles as representations are retrieved, stored, and integrated. Findings on analogical reasoning further support the cognitive demand and task specificity hypotheses (e.g., dual-process, dual-code, embodied theories). Different kinds of information recruited different prefrontal areas (e.g., external vs. internal information), and the harder the reasoning, the more areas recruited (e.g., 1-relation

vs. 2-relation problems). The general importance of frontopolar cortex for generating relations internally suggests that this area is especially important for processing complex analogies.

Summary, Conclusions, and Future Directions

The findings discussed in this article support the following conclusions about the neural bases of reasoning. First, deductive reasoning does not appear to recruit a unitary neural system but instead engages different brain regions based on the particular reasoning task and materials employed. Second, inductive inferences drawn from familiar categorical syllogisms and conditional statements engage a left-hemisphere language and knowledge network implemented in frontal and temporal regions. Third, problem solving generally recruits prefrontal regions, including bilateral DLPFC and MVPFC. Fourth, causal reasoning does not engage a single neural system but instead recruits different systems based on the causal reasoning task and the consistency of causal evidence with background beliefs. Finally, analogical inference selectively recruits frontal and parietal regions, with frontopolar cortex becoming increasingly important as task complexity increases.

The neuroscience evidence reviewed helps evaluate current psychological theories. First, reasoning typically recruits broadly distributed neural systems, providing evidence inconsistent with the relatively localized predictions of modularity theory. Possible exceptions include the importance of certain frontal regions for problem solving (left DLPFC) and analogical reasoning (frontopolar cortex). Second, the results demonstrate that reasoning does not solely recruit left-hemisphere regions for language and rule-based operations. Instead, reasoning often engages bilateral and posterior regions beyond those implicated by amodal and mental logic theories of reasoning. Third, the predominantly right-hemisphere system predicted by the mental-models theory across both deduction and induction is inconsistent with many patterns of left-hemisphere and bilateral activation observed across studies. Fourth, dual-process, dual-code, and embodied theories are generally consistent with the reviewed findings. These theories receive support because they predict the presence of (1) broadly distributed neural systems and (2) multiple reasoning systems (e.g., associative vs. rule, language vs. knowledge). These theories also receive support from the many studies that exhibit effects of task specificity and cognitive demand. As task conditions change, so do the neural systems that represent and process the relevant information.

In general, reasoning tends to recruit broadly distributed and diverse neural systems. The difficulty of establishing specific neural systems for a given type of reasoning (e.g., deduction) strikes us as one major challenge for future research. Does a particular type of reasoning consistently activate a specific neural circuit across wide variation in tasks and materials? If so, what is this circuit? Another major challenge is that the multifaceted nature of the neurobiological evidence and cognitive theories results in a many-to-many mapping: neural regions often serve multiple cognitive functions that can be mapped onto multiple cognitive theories. Sorting out the functional roles of particular brain areas and their roles in psychological theories of reasoning should be another major goal in this research area.

Many other challenges also await future research in this area. For example, do the results from the laboratory tasks reviewed here generalize to everyday reasoning tasks? Rather than occurring in a vacuum, everyday reasoning often occurs in social situations and is associated with emotional affect. What are the neural bases of reasoning under these conditions? How do social and emotional processes modulate reasoning? Future research should also address the important distinction between reasoning to a conclusion versus recognizing a conclusion. Because many real-world situations require that people generate valid conclusions (not just recognize them), future neuroscience research should assess the neural bases of this process. Future research should also continue to address the role of task difficulty in reasoning. Does task difficulty result in the recruitment of more brain regions, or are the same regions activated more intensely? Future research should also address the effect of learning on the neural systems that underlie reasoning. Do novices and experts engage similar neural systems during reasoning?

Finally, researchers should develop psychological theories that motivate fine-grained neurobiological predictions, and should design experiments that distinguish between theories, rather than simply attempting to confirm one. Although much progress has been made in framing the issues and establishing preliminary evidence, neuroscience research on reasoning is still in its infancy. We anticipate major advances in coming years.

Acknowledgments

This work was supported by National Science Foundation Grants DGE-0536941 and DGE-0231900 to A K Barbey, and by National Science Foundation Grant BCS-0212134 and Defense Advanced Research Programs Agency Contract FA8650–05-C-7256 to L W Barsalou.

See also: Decision-Making and Vision; Executive Function and Higher-Order Cognition: Assessment in Animals; Games in Monkeys: Neurophysiology and Motor Decision-Making; Memory Representation; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Referentiality and Concepts in Animal Cognition; Reward Decision-Making.

Further Reading

Barbey AK and Sloman SA (in press) Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*.

Barsalou LW (1999) Perceptual symbol systems. *Behavioral and Brain Sciences* 22: 577–660.

Braine MDS (1990) The 'natural logic' approach to reasoning. In: Overton WF (ed.) *Reasoning, Necessity, and Logic: Developmental Perspectives*, pp. 135–158. Mahwah, NJ: Erlbaum Press.

Christoff K and Gabrieli JDE (2000) The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology* 28: 168–186.

Christoff K, Prabhakaran V, Dorfman J, et al. (2001) Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage* 14: 1136–1149.

Colvin MK, Dunbar K, and Grafman J (2001) The effects of frontal lobe lesions on goal achievement in the water jug task. *Journal of Cognitive Neuroscience* 13: 1129–1147.

Flombaum JI, Santos LR, and Hauser MD (2002) Neuroecology and psychological modularity. *Trends in Cognitive Sciences* 6: 106–108.

Fonlupt P (2003) Perception and judgment of physical causality involve different brain structures. *Cognitive Brain Research* 17: 248–254.

Fugelsang J and Dunbar K (2004) A cognitive neuroscience framework for understanding causal reasoning and the law. *Philosophical Transactions of the Royal Society of London. Series B* 359: 1749–1754.

Goel V (2002) Planning: Neural and psychological. In: Nadel L (ed.) *Encyclopedia of Cognitive Science*, vol. 3, pp. 697–703. New York: Macmillan.

Goel V (2005) Cognitive neuroscience of deductive reasoning. In: Holyoak K and Morrison R (eds.) *Cambridge Handbook of Thinking and Reasoning*, pp. 475–492. Cambridge University Press: New York.

Goel V, Buchel C, Frith C, et al. (2000) Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage* 12: 504–514.

Goel V and Dolan RJ (2003) Explaining modulation of reasoning by belief. *Cognition* 87: B11–B22.

Green AE, Fugelsang JA, Kraemer DJM, et al. (2006) Frontopolar cortex mediates abstract integration in analogy. *Brain Research* 1096: 125–137.

Holyoak KJ (2005) Analogy. In: Holyoak KJ and Morrison RG (eds.) *Cambridge Handbook of Thinking and Reasoning*, pp. 117–142. Cambridge, UK: Cambridge University Press.

Johnson-Laird PN (1983) *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, UK: Cambridge University Press.

Novick LR and Bassok M (2005) Problem solving. In: Holyoak KJ and Morrison RG (eds.) *Cambridge Handbook of Thinking and Reasoning*, pp. 321–349. New York: Cambridge University Press.

Osherson D, Perani D, Cappa S, et al. (1998) Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia* 36: 369–376.

Over DE (2003) From massive modularity to meta-representation: The evolution of higher cognition. In: Over DE (ed.) *Evolution and the Psychology of Thinking: The Debate*, pp. 121–144. New York: Psychology Press.

Paivio A (1986) *Mental Representations*. New York: Oxford University Press.

Parsons LM and Osherson DN (2001) New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex* 11: 954–965.

Roger ME, Fugelsang JA, Dunbar KN, et al. (2005) Dissociating processes supporting causal perception and causal inference in the brain. *Neuropsychology* 19: 591–602.

Wharton CM, Grafman J, Flitman SS, et al. (2000) Toward neuroanatomical models of analogy: A positron emission tomography study of analogical mapping. *Cognitive Psychology* 40: 173–197.

Whitaker H, Savary F, Markovits H, et al. (1991) Inference deficits after brain damage. Paper presented at the International Neuropsychological Meeting, San Antonio, 1991.

Receptor Trafficking

R L Stornetta and J J Zhu, University of Virginia School of Medicine, Charlottesville, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Glutamate receptor trafficking plays a key role in synapse formation and elimination. Soon after the formation of presynaptic and postsynaptic contacts, *N*-methyl-D-aspartate (NMDA)-sensitive glutamate (Glu) receptors (-Rs) arrive in excitatory synapses, resulting in silent synapses. Activation of NMDA receptors then triggers synaptic delivery of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-sensitive Glu-Rs into these synapses, transforming the silent synapses into mature or functional synapses. Synapse elimination reverses the processes: AMPA-Rs are removed from synapses prior to removal of NMDA-Rs. Silent synapses are abundant in young animals. Their number declines with increasing age and stabilizes in the adult. Synaptogensis occurs rapidly during early postnatal development, with a much slower rate in the mature brain. Synapse elimination reduces the number of synapses, which is initially far greater than the number retained in the adult brain. The relatively stable numbers of total and mature synapses in adult animals reflect the balanced processes of synaptic delivery and removal of NMDA-Rs and AMPA-Rs.

AMPA-R Trafficking

AMPA-Rs are tetrameric proteins which are composed of GluR1, GluR2, GluR2L, GluR3, GluR4, or GluR4c – six subunits that are coded by four genes, GluR1–4. The cytoplasmic C-termini of the constituent subunits, which can be either long or short, govern the synaptic trafficking characteristics of AMPA-Rs. At least five distinct synaptic AMPA-R trafficking events exist at synapses (Figure 1). AMPA-Rs with long cytoplasmic termini (i.e., GluR1-, GluR2L-, or GluR4-containing AMPA-Rs) are driven into synapses by synaptic activity. This process mediates synaptic potentiation (e.g., long-term potentiation (LTP)). Soon after synaptic potentiation, some of the newly delivered AMPA-Rs with long cytoplasmic termini are removed from synapses in an activity-dependent manner during depotentiation. Other synaptic GluR1-, GluR2L-, or GluR4-containing AMPA-Rs are eventually exchanged with AMPA-Rs with only short cytoplasmic termini (i.e., GluR2-containing AMPA-Rs). GluR2-containing AMPA-Rs constitutively cycle between synaptic and nonsynaptic sites. Synaptic AMPA-R exchange and cycling processes require no synaptic activity. These two processes are important for maintaining synaptic strength despite continuous protein turnover. GluR2-containing AMPA-Rs can be removed from synapses during activity-induced synaptic depression (e.g., long-term depression (LTD)).

Multiple mitogen-activated protein kinase (MAPK) signaling pathways function to relay intracellular signaling that controls three activity-dependent AMPA-R trafficking events. During LTP, activation of NMDA-Rs and an influx of calcium stimulate the small GTPase Ras–extracellular signal-regulated kinase kinase (MEK)–extracellular signal-regulated kinase (ERK) and phosphoinositide 3 kinase (PI3K)–protein kinase B (PKB; also called Akt) signaling pathways, which phosphorylate AMPA-Rs with long cytoplasmic termini and drive them into synapses. During depotentiation, activation of NMDA-Rs and an influx of calcium stimulates the small GTPase Rap2–c-Jun amino-terminal kinase (JNK)–protein phosphatase 2B (PP2B) signaling pathway, which dephosphorylates AMPA-Rs with long cytoplasmic termini and triggers synaptic removal of these receptors. During LTD, activation of NMDA-Rs and an influx of calcium stimulates the small GTPase Rap1–p38 MAPK signaling pathway, which phosphorylates AMPA-Rs with only short cytoplasmic termini and leads to synaptic removal of these receptors.

Expression of AMPA-R subunits, regulated by a number of factors, such as developmental stage, brain region, and cell type, affects synaptic AMPA-R trafficking. For example, in hippocampal CA1 neurons, GluR4 is expressed only in the first postnatal week. The expression of GluR2L peaks at the end of the second postnatal week and declines by half in the adult. GluR1 expression increases with increasing age and reaches a maximal expression level after the third postnatal week. In addition, distinct patterns of synaptic activity seem to stimulate different signaling pathways and drive synaptic delivery of different AMPA-Rs. Whereas experience-independent or spontaneous synaptic activity is sufficient to drive synaptic delivery of GluR4- and GluR2L-containing AMPA-Rs, sensory experience-dependent activity and/or coactivation of other physiological factors (e.g., neuromodulators, hormones, and neurotrophic factors) is required for synaptic insertion of GluR1-containing AMPA-Rs in the intact brain. Thus, dependent on particular populations of synapses and developmental stages, transformation of silent synapses to functional synapses or synaptic delivery of AMPA-Rs requires different types of activities. Expression of other molecules, particularly those associated with the postsynaptic density (PSD), also influences synaptic AMPA-R trafficking. The synaptic functions of these proteins, including transmembrane AMPA-R regulatory proteins, PSD95,

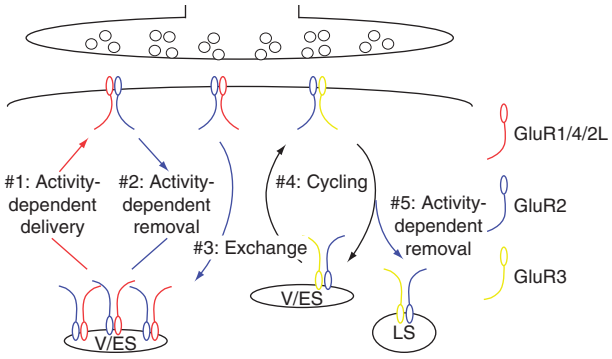

Figure 1 Model for synaptic AMPA-R trafficking. LS, lysosome; V/ES, vesicles/endosomes.

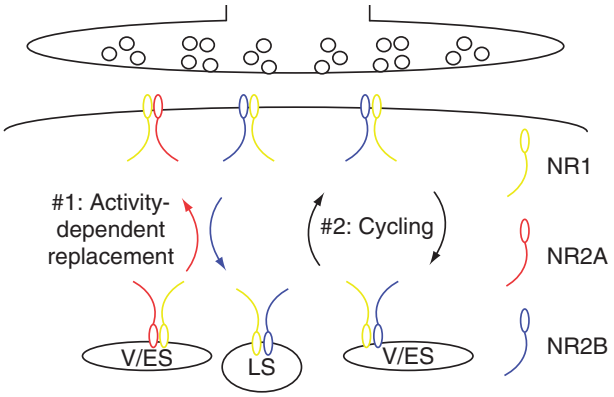

Figure 2 Model for synaptic NMDA-R trafficking. LS, lysosome; V/ES, vesicles/endosomes.

PSD93, and SAP97, are covered in more detail elsewhere in this encyclopedia.

Another factor that influences synaptic trafficking of AMPA-Rs is the egress of AMPA-Rs from the endoplasmic reticulum (ER). The ER has a rigorous quality control process to ensure that only properly assembled AMPA-Rs exit. As is the case with AMPA-R trafficking at synapses, the cytoplasmic C-termini of AMPA-R subunits define the properties of intracellular AMPAR trafficking. AMPA-Rs with long cytoplasmic tails mature and leave the ER rapidly, but AMPA-Rs with short cytoplasmic tails are largely confined within the ER; they mature and travel out of the ER at a much slower rate. Interestingly, the ER retention of GluR2-containing AMPA-Rs is controlled by Arg607 at the Q/R editing site in the GluR2 channel pore region: in GluR1, GluR3, and GluR4, there is a glutamine (Q) at this critical pore site, whereas in GluR2 the uncharged glutamine is replaced by a charged arginine (R). Reversion to Gln (R607Q) results in relief of the ER retention and rapid release of GluR2 from the ER. Low expression in neonatal neurons, reduced ER exit, or increased demand in young and adult neurons can result in an undersized intracellular pool of GluR2 subunits, which may be responsible for physiological and pathological expression of calcium-permeable GluR2-lacking AMPA-Rs at synapses.

NMDA-R Trafficking

NMDA-Rs are tetrameric complexes consisting of two obligatory NR1 subunits and up to two of four NR2 subunits (NR2A–D). The NR2A and-2B subunits are the primary NR2 subunits and are fairly widespread, whereas the expression of NR2C and NR2D is regionally and developmentally restricted in the brain. The NR2 subunits govern synaptic trafficking properties of NMDA-Rs. Two major synaptic NMDA-R trafficking events take place at synapses (Figure 2). NR2B-containing NMDA-Rs constitutively cycle between synaptic and nonsynaptic sites, and this process requires no synaptic activity. In contrast, NR2A-containing NMDA-Rs traffic into synapses by replacing synaptic NR2B-containing NMDA-Rs, and this process requires synaptic activity. The replacement is regulated by synaptic activity and is triggered by ligand binding to NR2B-containing NMDA-Rs. NR2A- and NR2B-containing NMDA-Rs differ significantly in their biophysical properties. These properties, together with the precise subcellular localization of NMDA-Rs, may determine which downstream signaling pathways (e.g., Ras–ERK, Rap2–JNK, or Rap1–p38 MAPK) are stimulated and which synaptic AMPA-R trafficking events are initiated.

Synaptic trafficking of NMDA-R is regulated by other factors. For instance, development has a significant impact on synaptic delivery of NMDA-Rs. Synapses in neonatal neurons have primarily NR2B-containing NMDA-Rs. However, at mature synapses, NR2A-containing NMDA-Rs predominate and NR-2B-containing NMDA-Rs are mostly extrasynaptic. Unlike AMPA-Rs, the number of NMDA-Rs per synapses remains approximately constant during postnatal development. This is important for maintaining synaptic efficacy and plasticity within the homeostatic range.

The ER selectively retains unassembled NMDA-R subunits and ensures only the forward trafficking of NMDA-Rs with certain subunit combinations, which determines the available intracellular pool size of functional NMDA-Rs. A large fraction of endogenous NR1 resides in the ER, waiting to be assembled with the NR2 subunits. By regulating NR1 mRNA splicing and NR1 subunit phosphorylation, neuronal activity and intracellular signaling bidirectionally modulate

the speed of forward trafficking of NMDA-Rs from the ER to synapses. For example, low neuronal activity results in synthesis of NR1 subunits with a divaline-based ER-export motif that speeds up the exit of NMDA-Rs from the ER, whereas high activity results in synthesis of NR1 subunits without the motif that conversely slows down the assembly of NMDA-Rs and their ER egress. Moreover, protein kinases, such as protein kinase A and protein kinase C, phosphorylate NR1 subunits and facilitate the assembly of NMDA-Rs and their ER exit.

Conclusion

AMPA-R and NMDA-R trafficking, which plays a key role in controlling synapse formation and elimination and in regulating synaptic transmission and plasticity, has been the subject of an intense and diverse investigation in the past decade. A considerable amount of information is now available on membrane trafficking of AMPA-Rs and NMDA-Rs at synapses. Only a limited amount is known about AMPA-R and NMDA-R assembly, quality control, and exit from the ER. Even less is known about how AMPA-Rs and NMDA-Rs traffic toward synapses after exiting the ER. For example, given that different populations of synapses in single neurons have distinct AMPA-R and NMDA-R compositions, how AMPA-Rs and NMDA-Rs recognize different types of synapses and how synaptic activity or intracellular signaling direct different groups of AMPA-Rs and NMDA-Rs to specific synapses after departing the ER remains unclear.

See also: AMPA Receptors: Molecular Biology and Pharmacology; AMPA Receptor Cell Biology/Trafficking; AMPA Receptors: Disease; NMDA Receptors and Disease + C464; NMDA Receptors, Cell Biology and Trafficking; NMDA Receptor Function and Physiological Modulation; Postsynaptic Development: Neuronal Molecular Scaffolds.

Further Reading

Collingridge GL, Isaac JT, and Wang YT (2004) Receptor trafficking and synaptic plasticity. *Nature Reviews Neuroscience* 5: 952–962.

Cull-Candy S, Kelly L, and Farrant M (2006) Regulation of Ca (2+)-permeable AMPA receptors: Synaptic plasticity and beyond. *Current Opinion in Neurobiology* 16: 288–297.

Dingledine R, Borges K, Bowie D, and Traynelis SF (1999) The glutamate receptor ion channels. *Pharmacology Reviews* 51: 7–61.

Greger IH, Khatri L, Kong X, and Ziff EB (2003) AMPA receptor tetramerization is mediated by Q/R editing. *Neuron* 40: 763–774.

Holtmaat A, Wilbrecht L, Knott GW, Welker E, and Svoboda K (2006) Experience-dependent and cell-type-specific spine growth in the neocortex. *Nature* 441: 979–983.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Reviews Neuroscience* 25: 103–126.

McCormack SG, Stornetta RL, and Zhu JJ (2006) Synaptic AMPA receptor exchange maintains bidirectional plasticity. *Neuron* 50: 75–88.

Perez-Otano I and Ehlers MD (2005) Homeostatic plasticity and NMDA receptor trafficking. *Trends in Neuroscience* 28: 229–238.

Qin Y, Zhu Y, Baumgart JP, et al. (2005) State-dependent Ras signaling and AMPA receptor trafficking. *Genes & Development* 19: 2000–2015.

Rosenmund C, Stern-Bach Y, and Stevens CF (1998) The tetrameric structure of a glutamate receptor channel. *Science* 280: 1596–1599.

Tada T and Sheng M (2006) Molecular mechanisms of dendritic spine morphogenesis. *Current Opinion in Neurobiology* 16: 95–101.

Thomas GM and Huganir RL (2004) MAPK cascade signalling and synaptic plasticity. *Nature Reviews Neuroscience* 5: 173–183.

Turrigiano GG and Nelson SB (2004) Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience* 5: 97–107.

Waites CL, Craig AM, and Garner CC (2005) Mechanisms of vertebrate synaptogenesis. *Annual Reviews Neuroscience* 28: 251–274.

Wenthold RJ, Prybylowski K, Standley S, Sans N, and Petralia RS (2003) Trafficking of NMDA receptors. *Annual Review of Pharmacology and Toxicology* 43: 335–358.

Recognition Memory

C E Stern and M E Hasselmo, Boston University, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Recognition memory refers to the capacity to accurately identify a particular item or stimulus as having been encountered previously. For example, a witness to a crime might be asked to identify a crime suspect or a weapon using recognition memory. The ability to accurately recognize an item requires that the subject first be able to encode stimuli into memory, and subsequently be able to discriminate between new and previously viewed stimuli. In the crime example, the witness would first need to accurately encode the information at the time of the crime, and then at a later time would need to be able to accurately identify the suspect or select the suspect from a line-up of other subjects who the witness had not encountered before. In research on memory function, recognition is often tested by having subjects distinguish between previously presented stimuli and stimuli that were not previously presented (termed lures).

There is a long history of research suggesting that recognition memory is composed of multiple components. In the 1970s, Atkinson and Juola described two processes important for recognition, an initial fast familiarity process and a slower process necessary for retrieving specific detailed information. For example, if you pass someone in an airport or supermarket, you may instantly recognize that person as familiar, followed by a slower search through memory as you retrieve detailed information about who that person is, his or her name and relationship to you, and where you last met that person. In some cases this 'retrieval attempt' is successful. In other cases, you might simply recognize the person as being familiar but be unable to generate details such as the person's name or where you met him or her. In 1983, Endel Tulving proposed that recognition could be divided into 'recollection' (or 'remembering') and 'familiarity' (or 'knowing'). The recollection process is proposed to involve retrieval of the prior episodic experience of encountering the previously presented stimulus, which includes remembering contextual details associated with that event. In contrast, familiarity is proposed as a process that does not rely on any explicit remembering of the details of the spatial or temporal context of the prior event.

In addition to dissociating the processes of recollection and familiarity, memory researchers make a distinction between the processes of recognition and recall. Both recognition and recall require that stimuli be encoded, stored, and retrieved from memory. However, recall places a much greater demand on the retrieval system, whereas recognition tests can assess whether information was encoded while placing less demand on memory retrieval systems by simply requiring that the subject distinguish previously learned information from new information. One analogy is that recognition memory tests resemble 'multiple choice' exams, where the answer is present but must be identified from false 'lures,' whereas a recall test resembles a 'fill in the blank' exam, where the subject must retrieve the learned information from long-term memory in order to generate the answer. When memory declines, such as during aging and in diseases such as Alzheimer's disease, recall performance declines earlier and faster than recognition performance. A person may therefore be unable to recall someone's name, but if he or she is provided with multiple choices, he or she can recognize the correct one.

Recall can be divided into two types, free recall and cued recall. Both free recall and cued recall involves the retrieval or recall of a previously presented stimulus without the stimulus being presented during the test period. In a free recall task, stimuli are presented during the sample period, and during the test period the subject is requested to generate as many stimuli as possible. Usually answers are provided verbally. In a cued recall test, the experimenter provides a cue, such as "name the animals that were on the list," that assists the subject in recalling the stimuli. In one specific type of cued recall task, known as a paired associates test, stimulus pairs are presented during the sample period, and during the test period only a single stimulus in a pair is presented, and the subject must generate the previously paired stimulus. Free recall and cued recall can only be performed for stimuli that can be generated by the subject, and therefore usually requires the use of verbal stimuli in humans. Because recognition can be tested without requiring the subject to generate the previously presented stimulus, it proves very important in neuroscience studies of memory function in animals.

In the following sections, work on the neuroscience of recognition memory in humans and in animals is described in greater detail. First, an overview of recognition memory tests is provided, followed by sections describing lesion studies of recognition memory

in humans and animals, electrophysiological studies of recognition memory in animals, imaging studies of recognition memory in humans, and computational studies of recognition. The human and animal studies indicate an important role for medial temporal lobe structures in the encoding of stimuli for subsequent recognition.

Recognition Memory Paradigms

Recognition memory has been studied extensively in human subjects using a number of different paradigms. In one classic recognition memory task, the experimenter first presents subjects with a series of stimuli to be learned. After a variable delay period, the experimenter then presents the subject with a list of previously presented stimuli mixed with new ones (lures). The subject simply responds 'yes' or 'no' to each stimulus. This recognition paradigm is known as a simple 'yes/no recognition test.' In another common recognition test, a single stimulus or set of stimuli are presented during a sample period, followed by a delay period of different durations, after which recognition is tested by presenting both a previously viewed sample stimulus and a novel lure at the same time during the test period. The subject must generate a response indicating which stimuli (target stimuli) were previously presented and which stimuli (lures) were not previously presented. This method is known as a 'forced choice recognition memory test.' The simple 'yes/no recognition test' cannot be used effectively with patients with memory disorders, as they will often respond "no" to all stimuli. The method is not often used because of this response bias. The forced choice recognition memory test is less sensitive to response bias, and can more sensitively assess recognition memory impairments and measurements of response bias can be generated. There are a number of variations of these two basic behavioral designs that have been created and are described here.

More recent studies have elaborated on the simple 'yes' or 'no' response to more accurately characterize the two processes that combine to support recognition memory, recollection and familiarity. Both recollection and familiarity would generate a 'yes' response in a yes/no or forced choice recognition memory paradigm. By modifying the task, the experimenter can distinguish between recollective and familiarity responses. The distinction between a strong feeling of recollection and a weaker sense of familiarity are sometimes referred to as the 'remember' versus 'know' distinction. In experimental tests of these processes, a subject may be requested to identify whether they have a distinct recollection of the event of observing the stimulus ('recollection') or whether they only have a general feeling of familiarity about the stimulus ('familiarity'). They may also be asked to retrieve information about contextual information related to the stimulus. Alternatively, in another twist on the yes no recognition test, subjects may be requested to use an arbitrary scale to provide information about the strength of their memory or their level of confidence in remembering an item or event. Some behavioral and imaging studies have used a simple three-point scale, corresponding to high-confidence yes (recollection), low-confidence yes (familiarity), and new (for the lures). Others have used five or six point scales ranging from a high-confidence 'yes' (identify target) response (numerical value five or six) to a high-confidence 'no' (reject lure) response (numerical value one). Studies of behavioral confidence rankings have been used to support a two-process theory of recognition memory in work by Yonelinas. This work uses signal detection theory-based receiver operator characteristics (ROC) curves to analyze the nature of recognition memory. This work used the distribution of proportions of responses to argue that recognition memory includes a graded mechanism of familiarity recognition, and a high threshold process of recollection-based recognition. The Yonelinas model assumes that previously presented items give rise to a range of confidence ratings based on their familiarity, while new items (lures) give rise to some familiarity but not as much as the previously presented items. In contrast, when recollection is triggered for an individual item, the recollection process is assumed to provide sufficient confidence for subjects to respond with certainty in the highest 'yes' category (allowing the responses to cross a high threshold criterion). In the dual-process model, familiarity is a variable phenomenon and recollection is an all or nothing response. There have been studies that suggest distinct medial temporal lobe regions support recollection and familiarity-based recognition. However, the studies remain controversial. Some researchers support the idea that the hippocampus is required for recollection, whereas parahippocampal structures support familiarity; however other data suggest that recollection and familiarity are part of a unified single process (see Wixted, 2007), supported by the same neuroanatomical structures within the medial temporal lobe. Similar controversy surrounds efforts to make distinctions between regions supporting encoding during the sample period versus separate regions mediating recognition during the testing period.

Another common recognition memory task is known as the 'delayed match to sample task' or DMS task. The DMS task involves presentation of single sample stimuli followed by a delay period,

followed by a pair of test stimuli, where one stimulus matches the presentation stimulus and one does not match the sample stimulus. The subject must select the stimulus that matches the sample stimulus. In a variant of this task known as a delayed nonmatch to sample or DNMS task, the subject (human or animal) must select the nonmatching 'novel' stimulus. A further variation involves continuous presentation of stimuli, in which the subject must respond on the basis of whether the current stimulus matches the stimulus presented *N* trials previously (an '*N*-back' task) which can be extended to longer intertrial intervals. Since it typically uses very short delays, the *N*-back task is considered to be a working memory task.

The delayed match to sample or delayed nonmatch to sample task design have been used most extensively in studies of the neural basis of recognition memory in animals. In these paradigms, the animal encounters the sample stimulus, and after a delay period is presented with a pair of stimuli and must respond to either the match (target) or nonmatch (lure). Alternatively, the sample stimulus could be followed sequentially by single lure stimuli (during which the animal withholds its response) until the target stimulus appears and the animal must generate a response. Performance of these tasks require recognition memory, but the process of recognition could be interpreted as involving either short-term active maintenance of the stimulus (working memory) or long-term synaptic encoding of the stimulus, depending on the length of the delay.

Other paradigms used in animals include presentation of a longer list of sample stimuli, which could include a series of visual stimuli presented to monkeys, or a set of odor stimuli presented to rats. After a delay period, the animals are then tested with a forced choice design, in which the animal encounters pairs of stimuli each of which includes one sample stimulus and one lure stimulus. In rats, recognition memory is commonly tested using the nonmatch to sample paradigm, to capitalize on the natural tendency of rats to explore novel stimuli more extensively than familiar ones. The rat can be presented with a sample stimulus, and subsequently when presented with an identical sample stimulus and a novel lure, it will spend more time exploring the novel lure.

Lesion Studies of Recognition Memory

There is extensive lesion evidence supporting the idea that the medial temporal lobe structures, including the hippocampus and neighboring parahippocampal regions, are critical for encoding information into long-term memory. Research on the neuroanatomical structures critical for memory function focused in on the medial temporal lobe following the case of patient HM, first described by Scoville and Milner in 1957 and subsequently studied by Milner and Corkin. Patient HM showed a striking impairment of declarative memory encoding following bilateral removal of his medial temporal lobes in an attempt to control epileptic seizures. Initial descriptions of patient HM's impairment focused on his almost total loss of free recall and severe impairment of cued recall performance. Similarly, other hippocampal amnesics show strong impairments of free recall and cued recall. There is some sparing of recognition memory in hippocampal amnesics, and the amount of impairment appears to be associated with the magnitude of damage of parahippocampal structures including the entorhinal cortex, perirhinal cortex, and parahippocampal gyrus.

Lesion studies in humans have also addressed the question of whether the processes of recollection and familiarity are subserved by different regions within the medial temporal lobes. However, the results continue to be controversial. Researchers including Yonelinas have argued that hippocampal lesions cause a greater impairment of recollection-based recognition than familiarity-based recognition. On the other hand, studies by Manns and Squire have demonstrated impairments of both recollection and familiarity-based recognition associated with medial temporal lesions.

In animals, lesions have been demonstrated to impair recognition memory in delayed nonmatch to sample tasks. Researchers including Squire and Zola-Morgan have demonstrated that lesions of the hippocampus alone cause some impairments in recognition of stimuli after delays of 8 s to many minutes. These studies also demonstrate that lesions that include parahippocampal structures, such as the perirhinal cortex and parahippocampal gyrus, cause much stronger impairments in recognition performance in delayed nonmatch to sample tasks, supporting an important role for parahippocampal structures in recognition memory. Impairments of performance in object or odor recognition tasks have also been shown with inactivation or lesions of the perirhinal cortex in rats. Selective impairments of the encoding of stimuli for subsequent recognition have been induced with manipulations that block acetylcholine receptors or remove the cholinergic innervation of the perirhinal cortex. Recent work by Fortin and Eichenbaum has used modified variants of tasks designed for human studies of recollection and familiarity to examine these processes in rats, demonstrating that hippocampal lesions appear to selectively reduce the high threshold recollection-based recognition, in contrast to the more variable familiarity-based

recognition. However the role of the hippocampus and perirhinal cortex to recollection and familiarity remains controversial (see Wais et al., 2006).

Electrophysiological Studies of Recognition Memory

As noted above, most studies of memory function in animals use tasks that test recognition memory, since it is not feasible to have animals perform recall tasks. Thus, most studies of electrophysiological activity associated with memory function have been performed during short- or long-term recognition memory paradigms. In both monkeys and rats, unit recordings studies analyze the generation of action potentials by individual neurons during performance of such tasks. The activity associated with recognition can be evaluated by comparing the response during the test period to a previously presented target stimulus with the response to a nonpresented lure, or by comparing the response to the target stimulus with the response to the same stimulus during the sample period. A number of studies have shown increases or decreases in neural activity for a previously presented target stimulus in comparison to nonpresented lures. Studies by researchers including Malcolm Brown and Earl Miller demonstrate a significant decrease in response to a previously presented stimulus in comparison to the sample response for neurons in regions of temporal lobe including parahippocampal cortices. This effect is commonly referred to as match suppression or repetition suppression. Miller demonstrated that repetition suppression occurs even for repeated distractor stimuli presented during a delay period, whereas enhancement of the target test stimulus relative to the sample stimulus (match enhancement) only occurs for the target stimulus and not for distractor stimuli. Increased responses to matching stimuli were shown in the anterior thalamus in studies of a serial recognition task.

Across short delays, the process of recognition could be supported by active maintenance of information during the delay period. In support of active maintenance, a number of unit recording studies performed during DMS by researchers such as Niki and Fuster have demonstrated that some neurons activated by the sample stimulus will maintain their activity during the delay period until the test stimulus is presented. This delay period activity could reflect active maintenance of the stimulus for performance of the recognition task. The persistent activity could arise from intrinsic mechanisms within individual neurons, or from excitatory feedback between neurons in a population. This type of persistent spiking activity could contribute to the increase in cerebral blood flow measures seen during the delay period of DMS or two-back tasks in human subjects.

Functional Imaging of Recognition Memory

Considerable research has focused on imaging changes in blood flow and metabolism associated with performance of recognition memory tasks. Despite considerable evidence from human and animal lesion studies regarding the importance of the hippocampus and medial temporal lobes for encoding and recognition, early imaging studies using positron emission tomography (PET) primarily identified changes in activity in cortical structures outside of the medial temporal lobe, most notably within the prefrontal cortex. The tasks used in these earlier studies focused on the recognition of verbal stimuli learned prior to scanning. When the task design was altered to focus on examining functional changes during the encoding period, Stern and colleagues identified robust functional magnetic resonance imaging (fMRI) activation of the hippocampus and parahippocampal structures. As fMRI techniques developed, it became possible to focus on event-related activity during individual trials involving presentations of single stimuli. Researchers designed studies to focus on how the event-related activity during presentation of single stimuli during the encoding period would correlate with performance for those individual stimuli during a subsequent recognition memory task. These studies demonstrated a subsequent memory effect, in which increases in fMRI activity in the hippocampus and parahippocampal gyrus were noted for stimuli that were remembered, but not for stimuli that were subsequently forgotten. In addition to activity in medial temporal lobe sites, imaging studies have provided data that support the idea that encoding and subsequent recognition requires widespread activity including activity in the prefrontal cortex and parietal lobes. Research is ongoing to determine whether the prefrontal and parietal activity is related to the successful retrieval of information from memory, or due to control mechanisms driving recollection attempts.

Imaging studies have also addressed the hypothesis that recollection and familiarity are dependent on different regions within the medial temporal cortex. Based on animal studies and anatomical differences between the hippocampus and parahippocampal regions, several researchers have suggested that the hippocampus might be necessary for recollection, whereas the parahippocampal gyrus might be sufficient and necessary for supporting familiarity. Proponents of this account suggest that the parahippo campal

regions are associated with successful item recognition, whereas hippocampal activity is necessary for successful source memory, which includes information about contextual details. However, the imaging studies are inconclusive at this time, and the issue remains controversial.

The recognition of stimuli in working memory tasks has been examined using the *N*-back task and the DMS. These tasks probably require active maintenance of previous stimuli rather than modification of synapses. The *N*-back task typically causes extensive activity in prefrontal cortex and parietal cortex, which increases when the memory load is increased by increasing the value of 'N' to require matching with a stimulus presented three or four stimuli previously. When familiar stimuli are used in a delayed matching task, it does not cause significant medial temporal lobe activity, but when novel stimuli are used it causes a significant increase in medial temporal lobe activity, which is correlated with subsequent recognition memory for the stimuli. This could reflect a greater role for medial temporal lobe structures in active maintenance and encoding of novel stimuli.

Computational Modeling of Recognition Memory

The processes involved in recognition memory have been modeled on a number of different levels. Many modeling studies have focused on the mechanisms of active maintenance, in which persistent spiking activity of a population of neurons during the delay period is proposed to form the basis for retaining information about the stimulus presented during the sample stimulus. Most models of active maintenance focus on the role of excitatory recurrent connections within cortical structures such as the prefrontal cortex. These excitatory recurrent connections can cause a population activated by the sample stimulus to continue to reactivate its component neurons, causing the population to maintain activity for the full delay period, retaining the information until the test period when a comparison can be made. In contrast, other recent models of active maintenance focus on intrinsic single-cell mechanisms, in which a single neuron can maintain persistent spiking activity without requiring excitatory synaptic feedback between neurons. This mechanism would be particularly appropriate for novel stimuli for which a representation has not previously been formed.

For recognition over longer periods, it is not possible to maintain persistent spiking representing all previously viewed stimuli. In this case, some representation of stimuli based on changes in synaptic connections between neurons would be more appropriate. Computational models have demonstrated how presentation of a sample stimulus could induce Hebbian synaptic potentiation, in which synapses are strengthened dependent upon presynaptic activity coupled with postsynaptic activity. This synaptic modification results in greater activity of neurons during the test period if they were previously active during the sample period (match enhancement). Match suppression or repetition suppression can be obtained by competitive long-term depression of synapses for nonactivated neurons. Alternately, researchers have proposed that repetition suppression might arise from Hebbian long-term depression which provides a mechanism for recognition memory with higher capacity than Hebbian long-term potentiation. The above processes focus on responses to the item alone and could be seen as reflecting familiarity processes. The separate theoretical processes of recollection-versus familiarity-based recognition have been modeled as different processes in different structures. In contrast to familiarity, which could involve an increase or decrease in activity of the neurons representing the sample stimulus, recollection may involve a test item reactivating a full representation of contextual features associated with a specific episode. This recollection process could involve more complex neuroanatomical circuits required for reactivation of the temporal or spatial context of the item being recognized.

See also: Amnesia: Declarative and Nondeclarative Memory; Memory: Computational Models; Executive Function and Higher-Order Cognition: Assessment in Animals; Executive Function and Higher-Order Cognition: Neuroimaging; Hebbian Plasticity; Neuroimaging; Working Memory: Capacity Limitations.

Further Reading

Bogacz R and Brown MW (2003) Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus* 13: 494–524.

Brown MW and Xiang JZ (1998) Recognition memory: Neuronal substrates of the judgement of prior occurrence. *Progress in Neurobiology* 55: 149–189.

Fortin NJ, Wright SP, and Eichenbaum H (2004) Recollection-like memory retrieval in rats in dependent on the hippocampus. *Nature* 431(7005): 188–191.

Hasselmo ME and Wyble BP (1997) Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research* 89: 1–34.

Manns JR, Hopkins RO, Reed JM, Kitchener EG, and Squire LR (2003) Recognition memory and the human hippocampus. *Neuron* 37: 171–180.

Miller EK, Erickson CA, and Desimone R (1996) Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience* 16: 5154–5167.

Norman KA and O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychology Review* 110: 611–646.

Stern CE, Corkin S, Gonzalez RG, et al. (1996) The hippocampal formation participates in novel picture encoding: Evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America* 93: 8660–8665.

Stern CE and Hasselmo ME (1999) Bridging the gap: Integrating cellular and functional magnetic resonance imaging studies of the hippocampus. *Hippocampus* 9: 45–53.

Stern CE, Sherman SJ, Kirchhoff BA, and Hasselmo ME (2001) Medial temporal and prefrontal contributions to working memory tasks with novel and familiar stimuli. *Hippocampus* 11: 337–346.

Wagner AD, Schacter DL, Rotte M, et al. (1998) Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science* 281: 1188–1191.

Wais PE, Wixted JT, Hopkins RO, and Squire LR (2006) The hippocampus supports both the recollection and the familiarity components of recognition memory. *Neuron* 49: 459–468.

Wixted JT (2007) Dual-process theory and signal-detection theory of recognition memory. *Psychological Review* 114(1): 152–176.

Yonelinas AP (2001) Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Human Learning and Memory* 20: 1341–1354.

Yonelinas AP, Otten LJ, Shaw KN, and Rugg MD (2005) Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience* 25: 3002–3008.

Young BJ, Otto T, Fox GD, and Eichenbaum H (1997) Memory representation within the parahippocampal region. *Journal of Neuroscience* 17: 5183–5195.

Red Nucleus

L E Miller, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
A R Gibson, Barrow Neurological Institute, Phoenix, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Location and Subdivisions of the Red Nucleus

The red nucleus (RN) comprises a large group of cells located in the ventral medial mesencephalon. Caudally, the RN is bounded by the decussation of the superior cerebellar peduncle, and rostrally the RN reaches approximately to the level of the habenulo-interpeduncular tract (fasciculus retroflexus). The primate RN is distinguished by two characteristics: one is the presence of extremely large cells at its caudal pole, and the second is the presence of a large body of small cells rostral to the giant cells. The nucleus is thus commonly divided into a caudal magnocellular division, RNm, and a rostral parvicellular (or parvocellular) division, RNp.

The defining characteristics of the RN are inputs from the cerebellar nuclei and cerebral cortex and projection to either the contralateral spinal cord via the rubrospinal tract (RST) or ipsilateral inferior olive (IO) via the central tegmental tract. The very largest cells of RNm project to the contralateral spinal cord, and the smallest cells of RNp project to the ipsilateral IO. The projections are mutually exclusive, but both cell populations receive input from the cerebellar nuclei as well as from the cerebral cortex.

Connections of the RN do not precisely correlate with cell size. Early studies of the RN recognized that axons from many small and medium-size cells decussate and join the RST. However, since the smaller cells that decussate do not all project to the spinal cord, neither cell size nor spinal projection alone provides an adequate basis for dividing the nucleus into functional groupings.

This article uses as much of the traditional nomenclature as possible: RNp will be used to identify small cells that project to the ipsilateral IO. RNm will be used to identify large cells that project to the contralateral spinal cord (Figure 6(b)). Thus RNm and RNp do not include the small and medium-size cells that project to the contralateral brain stem and spinal cord. An additional term, parvicellular contralateral red nucleus (RNpc; Figure 6(a)), is introduced to identify this group of cells. To a first approximation, RNm comprises the most caudal portion of the nucleus, RNp comprises a rostral-medial portion of the nucleus, and RNpc comprises a rostral-lateral portion of the nucleus. An additional population of small interneurons, which are probably present throughout the RN, is not addressed by this nomenclature.

Species Variation of the Red Nucleus

The RN varies between species. Much of the variability is due to RNp, which is a distinct nucleus. The size of RNp parallels that of several other brain stem areas that are connected to the motor cortex, such as the pontine nuclei, the principal division of the IO, and the lateral cerebellum. Figure 1 illustrates parasagittal sections through the brain stem of a cat and a monkey following injections of wheat germ agglutinin–horseradish peroxidase (WGA-HRP) into motor cortex. RNp is much larger in the monkey than it is in the cat – massive terminations outline the monkey RNp. A similar expansion of terminals is apparent in the pontine nuclei. RNm, on the other hand, is roughly the same size in cat and monkey. In neither species is there a dense input from motor cortex to either RNm or RNpc.

The size of RNm varies with the size of its major input source, the interpositus nucleus (NI) of the cerebellum. NI does not share the progressive and extreme enlargement that is seen for the cerebellar dentate nucleus in monkeys, apes, and humans. In humans, RNm is greatly reduced in size, and the RST may not descend below the cervical cord. The loss of RNm in humans is mystifying since its evolution appears to be associated with the presence of limbs; snakes and fishes do not have the equivalent of an RNm. The smaller RNm in humans may simply reflect a shift in emphasis from large cells projecting to lumbar cord to smaller cells projecting to upper cervical cord and brain stem.

It is likely that RNp, RNm, and RNpc exist in all mammalian species. However, in species below primates, RNp has been confused with RNpc. In primates, there has been little effort to distinguish small cells projecting to the IO from small cells contributing to contralateral projections. RNp is greatly expanded in mammals with a large motor cortex, whereas RNm and RNpc have probably remained relatively constant across species.

The RNm

RNm is defined by the presence of very large cells (Figure 6(b)), but there is actually a gradation in cell size within RN. The gradation follows the general principle that neurons supporting long axons tend to have larger cell bodies than those supporting short

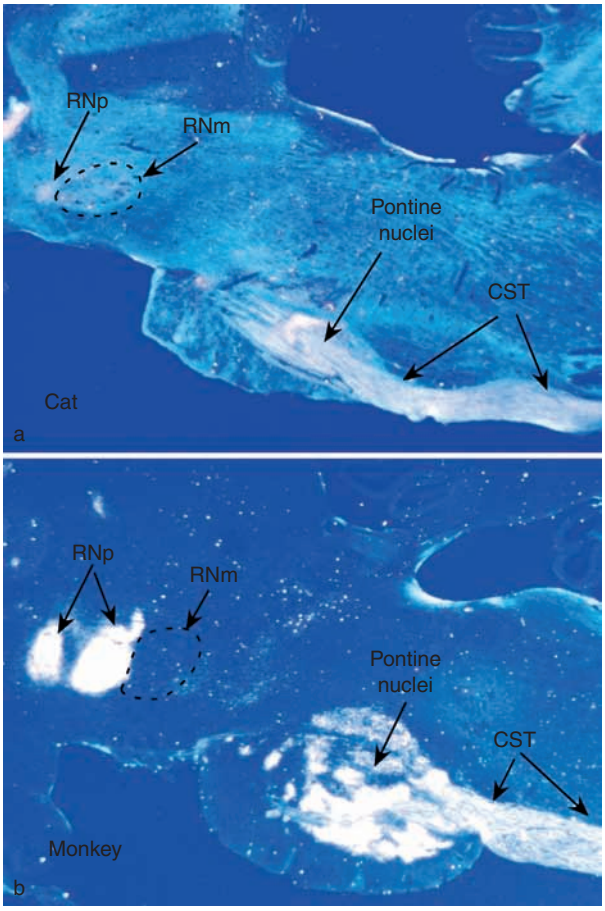

Figure 1 Comparison of monkey and cat red nucleus (RN). Parasagittal brain stem sections following injection of wheat germ agglutinin–horseradish peroxidase into primary motor cortex (area 4) in cat (a) and monkey (b). While the size of magnocellular RN (RNm) is unchanged across species, note the tremendous expansion of the olivary projecting parvicellular (or parvocellular) red nucleus (RNp) in monkey. RNp is divided at this laterality by the fascicularis retroflexus. The pontine nuclei, which relay cortical input to cerebellar cortex, are also greatly enlarged in the monkey. Darkfield polarized light illumination was used in this and following photographs. Labeled areas are light. Magnification ×4.3 (a); ×5.1 (b). CST, corticospinal tract.

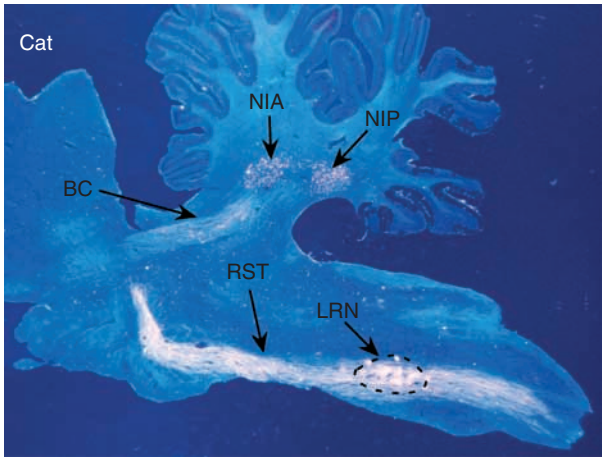

Figure 2 Input–output connections of magnocellular red nucleus (RNm). Parasagittal section through the brain stem and cerebellum of a cat following wheat germ agglutinin–horseradish peroxidase injection into contralateral RNm. Anterograde labeling of fibers highlights the rubrospinal tract (RST) and terminals within the lateral reticular nucleus (LRN). Retrograde labeling of cells in the cerebellar anterior interpositus (NIA) and posterior interpositus (NIP). Magnification ×3.6. BC, brachium conjunctivam. Adapted from Robinson FR, Houk JC, and Gibson AR (1987) Limb specific connections of the cat magnocellular red nucleus. *Journal of Comparative Neurology* 257: 553–577, with permission.

axons. Therefore, the very largest RNm cells, which are located ventrally in the nucleus, project to the lumbar cord. Cells projecting to the cervical cord tend to be somewhat smaller and lie dorsally to those projecting to lumbar cord. Even smaller cells, prevalent at rostral levels of the nucleus, project to the contralateral brain stem and also contribute to spinal projections. The topography within the RN is quite sharp, and very few cells project to both lumbar and cervical levels.

RNm Anatomy

Cerebellar input to the RNm arises from the anterior NI (NIA) and medial regions of posterior NI (NIP). Figure 2 illustrates a parasagittal section through the cat brain stem after an injection of WGA-HRP into the contralateral RNm. Large numbers of retrogradely labeled cells can be seen in NIA and NIP, and anterogradely labeled fibers can be seen forming the RST. The projection from interpositus to RNm is somatotopically organized: cells that discharge during hind limb movements are located medially in NIA and NIP and terminate in ventral and lateral RNm. Cells that discharge during forelimb movements are located more laterally in NIA and NIP and terminate in dorsal and medial RNm. The most lateral cells of NIP project to the superior colliculus rather than to the RN.

Cerebral cortical input to RNm is relatively limited, arising from cells deep in layer V of posterior motor cortex (area 4), including some Betz cells. Presumably, the input from Betz cells is a collateral input from corticospinal neurons. The fine caliber of collateral projections may contribute to the relatively light labeling seen in RNm after injection of anterograde tracers in motor cortex. However, physiological studies also indicate that the input from motor cortex to RNm is relatively weak in comparison with the input from interpositus.

Other significant inputs to RNm arise from the cuneate, trigeminal, and gracile nuclei, including inputs from the spinal cord. All these inputs conform to the RNm somatotopy, and they are likely to account for sensory properties displayed by RNm neurons. Trigeminal connections are located in a dorsolateral location and

indicate that RNm probably has a smaller-celled projection to the brain stem related to jaw movements. This projection would not be identified by retrograde labeling from the spinal cord, so it is likely that the size of RNm has been somewhat underestimated.

The axons of RNm neurons decussate and travel immediately ventral to the superior cerebellar peduncle to form the RST. The RST turns ventrally at the level of the pons and travels along the ventral-lateral edge of the brain stem to join fibers in the dorsolateral funiculus of the spinal cord. Most of the spinal terminations of RNm lie in Rexed's laminae V–VII, but at C8-T1 a small patch of terminations is located laterally among motor neuronal pools of lamina IX. Lamina IX terminations have been confirmed in monkey, cat, and rat. In all species, the terminations correspond to the location of motor neuronal pools that innervate muscles that move the digits of the forelimb. Figure 3 illustrates frontal sections of the cat spinal cord at C8. Panel (a) illustrates RNm terminations, and panel (b) illustrates retrogradely labeled motor neurons after injection of a digit extensor muscle, extensor digitorum communis. Although the direct projections to spinal motor neurons are fewer than those to interneuronal projections, the importance of RNm and its cerebellar input nucleus, NIA, for the control of digit movements is further emphasized by the results of physiological experiments.

In addition to its spinal projections, RNm (together with the motor cortex) has prominent projections to ventral regions of the gracile, cuneate, and trigeminal nuclei, among cells that project to the IO. These projections may provide a substrate for modulation of olivary activity by descending motor signals. Beyond these focused projections, there are less well-defined, finer projections to large regions of brain stem reticular areas.

RNm has both direct and indirect connections back to the cerebellum. A prominent indirect feedback pathway is provided by RNm terminations in the contralateral lateral reticular nucleus (Figure 2), which supplies mossy fiber input to the cerebellar cortex, as well as to interpositus. Terminations within lateral reticular nucleus are topographic, such that RNm terminations are separate from cortical or spinal terminations. It is likely that the RNm terminations are collaterals of spinally projecting axons. Along with the cortical projections through the pontine nuclei, these projections may provide efference copy signals to the cerebellum.

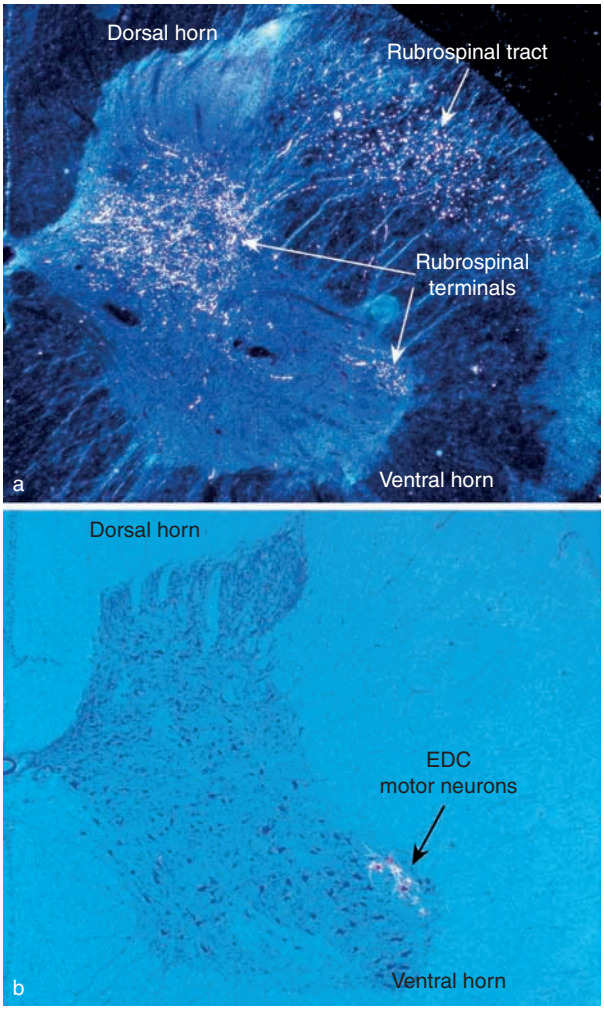

Figure 3 Spinal projections of magnocellular red nucleus (RNm). Frontal sections through the cervical enlargement at spinal segment C8 of the cat. (a) Anterograde labeling of fibers demonstrates a dense plexus among interneuronal regions with a small focused termination in lateral motor neuronal pools. (b) Labeled motor neurons following horseradish peroxidase injections into a digit extensor muscle, extensor digitorum communis (EDC). The location of these motor neurons matches RNm ventral horn terminations shown in (a). Magnification ×25.

RNm Function

Chronic lesions and reversible inactivation of RNm

In cats and monkeys, permanent lesions of the RNm or rubrospinal tract cause only relatively mild deficits that are predominantly related to the animal's ability to grasp objects. Temporary inactivation of the cat RNm produces deficits in grasping. The deficit appears to be related to particular muscles rather than the grasping movement *per se*, since the cats also drag the dorsum of the paw during walking. Movement deficits are much greater and more persistent if RNm lesions are combined with lesions of the pyramids, thereby also removing the influence of the corticospinal tract. These animals lose virtually all ability to use the affected limb independently, although they are still able to maintain a reasonably normal posture and to locomote. This remaining

capability is probably due to corticobulbar projections to several other brain stem motor nuclei that project bilaterally to the spinal cord (collectively referred to as the ventromedial brain stem system). The ventromedial system is primarily concerned with control of midline and axial musculature.

Electrophysiological evidence of spinal projections from RNm Electrophysiological observations are consistent with the anatomical connections of the RNm and with the behavioral deficits resulting from lesions. In dorsolateral regions of the nucleus, cells discharge at high rates during face and mouth movements. Cells within dorsomedial regions, which project to cervical cord, discharge at high rates prior to and during movements of the contralateral forelimb. Cells in ventrolateral RNm, which projects to lumbar cord, discharge during movements of the contralateral hind limb.

As in the primary motor cortex, small movements can be evoked by stimulating within RNm, using brief trains of electrical pulses. Within the forelimb regions of RNm, this microstimulation typically causes brief twitches of the digits, generally in the extension direction, rather than flexion. These effects can be quantified by recording the electrical activity related to muscle contractions and averaging the tiny fluctuations with respect to the time of these stimulus pulses (stimulus-triggered averaging). Similarly, one can average electromyographic activity relative to individual action potentials recorded from RNm neurons (spike-triggered averaging). This approach allows one to survey the connections to the muscles made by individual neurons. Such studies have shown that RNm neurons activate groups of three to six muscles of the limb, with a strong bias toward the distal extensor musculature. An analogous region ventrolateral in the nucleus contains neurons projecting to the lumbar cord and discharging during movements of the hindlimb, especially extension movements of the toes. At the lowest currents, stimulation in hindlimb areas of RNm produces toe extension (dorsiflexion).

Many textbooks state that the RN is involved in flexion movements of the proximal limb, but this statement is based on results from stimulation in decerebrate preparations and is not consistent with results from intact animals. Neurons within RNm do influence proximal muscles, but their major emphasis is on distal limb muscles, especially extensors. It may be that the influence of RNm on proximal limb musculature primarily serves to support digit movements.

Signals recorded from RNm Averaging methods, together with anatomical tracing and lesion studies, provide a rather consistent picture of the spatial organization of RNm output. By studying the modulation of signals produced by RNm neurons during movement, one can relate the timing of these control signals to the movement. Typically, an experimental animal is trained to control the movement of a cursor on a screen by moving a manipulandum, analogous to a video game joystick. Particular movements can be required by displaying a sequence of targets on the screen and by the use of different types of manipulanda. Figure 4 shows several different devices operated by various movements of the arm and hand. Of these three devices, the throttlelike device (panel C), which requires a coordinated action of digits and wrist, is one of the most reliable means of evoking discharge from forelimb RNm neurons. The same cells show large bursts of activity as a monkey reaches toward and grasps a small piece of food, such as a raisin. However, discharge modulation shows little dependence on the direction of reach. If the monkey makes a limb movement without a grasp, such as when holding a piece of food or a device handle, there is little discharge modulation in RNm (or in interpositus, which provides RNm input). In general,

Figure 4 Movement specificity of magnocellular red nucleus (RNm). Three different devices used to elicit specific movements from monkey subjects during recording from neurons in the forelimb area of RNm. (a) Manipulandum used to restrict movements to the elbow joint. Pivot joints are indicated by +. E, extension; F, flexion. Arrows indicate directions of movements. (b) Manipulandum used to restrict to metacarpal phalangeal joints. The largest discharge typically occurred during the coordinated wrist and digit movements necessary to operate the device in (c). Adapted from Gibson AR, Houk JC, and Kohlerman NJ (1985) Magnocellular red nucleus activity during different types of limb movement in the macaque monkey. *Journal of Physiology* 358: 527–549.

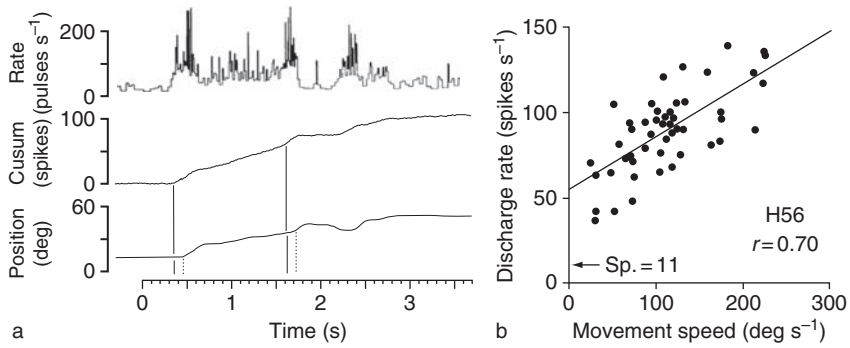

Figure 5 Integrated magnocellular red nucleus (RNm) discharge matches movement. (a) Summed discharge of an RNm neuron during movements of the device in Figure 4(c). Increases in firing rate (solid vertical lines) occurred slightly more than 100 ms prior to changes in movement speed (dashed lines). (b) Mean discharge rate during these movements was well correlated with the speed of the corresponding movement. Sp. indicates spontaneous discharge rate. H56b is designation of subject. Cusum, cumulative sum of spikes. Adapted from Gibson AR, Houk JC, and Kohlerman NJ (1985) Relation between red nucleus discharge and movement parameters in trained macaque monkeys. *Journal of Physiology* 358: 551–570, figures 2 and 6.

movements requiring use of the digits, perhaps in coordination with other limb parts, are the most important factor for eliciting high discharge rates from RNm.

Discharge from RNm neurons is typically phasic: discharge onset occurs approximately 100 ms before movement onset, and discharge offset leads movement offset by a somewhat shorter latency. The lead in discharge onset and offset suggests that RNm discharge is a movement command signal, not primarily driven by movement feedback. For some cells, as in the example in Figure 5, the timing and magnitude of discharge mirror the timing and magnitude of movement with remarkable fidelity. This is reflected in the similarity between a cumulative sum of spikes (Cusum) and the position of the manipulandum. In fact, there is a highly significant linear relation between the speed of these movements and the rate of discharge, as shown in the scatter plot in panel B.

The correlation between discharge rate and movement speed has led to the proposition that the signals from RNm encode kinematic properties. However, subsequent work has also found correlations with force, the first time derivative of force, and muscle activity, particularly among distal extensors. Much of the phasic nature of the RNm signals may actually reflect the significant acceleration- and velocity-related forces that occur during limb movement rather than the speed of the movement *per se*.

The Contralaterally Projecting RNpc

At more-rostral levels of the cat RN, a dorsolateral extension of the nucleus containing cells of mixed sizes becomes prominent. The axons of some of these cells decussate and join with other RST fibers at the level of the pons. Due to the crossed descending projection to the facial nucleus and spinal cord, RNpc is more closely associated with RNm than with RNp. However, RNpc has considerably more extensive connections with other mesencephalic regions than does RNm. Although most of the anatomical data relevant to RNpc is based on cat studies, a smaller-celled rostral-dorsolateral extension of the RN has also been noted for the rat, monkey, and human.

RNpc Anatomy

RNpc of the cat receives cerebellar input from rostral dentate and a caudal lateral segment of NIA. The same region of NIA receives input from a region of the IO that contains cells responsive to cutaneous stimulation of the face. Unlike RNm, RNpc also receives a large input from other areas of the mesencephalon. The most prominent of these inputs include regions immediately rostral to the RN, such as the fields of Forel, the dorsal half of the zona incerta, and the anterior pretectal nucleus. These nuclei, in turn, receive inputs from many other areas of the brain, including the cerebellar dentate nucleus and the entopeduncular nucleus, the feline equivalent of the internal segment of the globus pallidus. Cerebral cortical inputs to RNpc have not been well investigated. For the cat, RNpc receives input from rostral regions of the suprasylvian gyrus, but the monkey dorsolateral extension of RN, which may be a component of RNpc, does not appear to receive cortical input.

Prominent brain stem terminations of RNpc include the contralateral facial and spinal trigeminal nuclei. RN cells projecting to the facial nucleus are located in the rostral-dorsal extension of the nucleus. The labeled cells in Figure 6(a) illustrate the location of RNpc cells projecting to the intermediate division of the contralateral facial nucleus. In the rat, the population of RNpc cells projecting to the facial

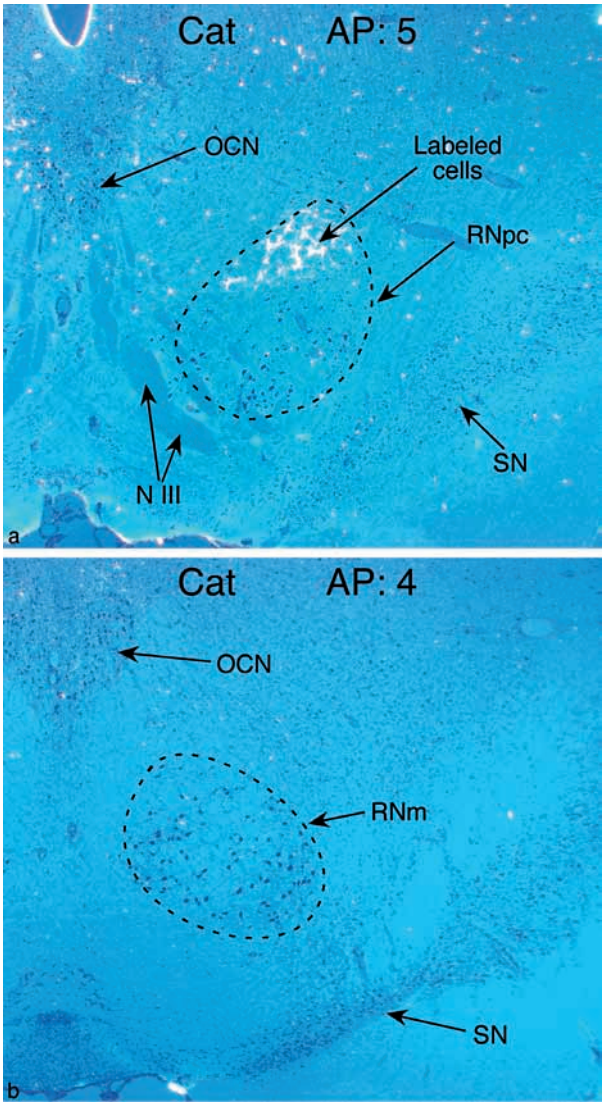

Figure 6 Facial nucleus projections of parvicellular contralateral red nucleus (RNpc). (a) Retrogradely labeled cells (white) in dorsolateral RNpc following injection of wheat germ agglutinin–horseradish peroxidase centered on the intermediate subnucleus of the contralateral facial nucleus (AP 5.0). This subnucleus innervates the orbicularis oculi muscle, which is important for making eyelid movements for facial expression and protection. (b) The large cells of magnocellular red nucleus (RNm; AP 4.0) have no projections to the facial nucleus. Magnification ×14. AP, anterior–posterior plane; NIII, oculomotor nerve (third nerve); OCN, oculomotor nucleus; SN, substantia nigra.

nucleus is distinct from the population of cells projecting to the spinal cord, and it is likely that this is true for other species. The strong projection of RNpc to facial motor neurons is an exception for the RN, since most terminations of RNm are on spinal interneurons rather than motor neurons.

The most rostral dorsal regions of RNpc project to the upper cervical cord. The projection is heaviest to interneuronal regions of C1 and largely disappears at levels below C4. There are no terminations within spinal motor neuronal pools. The larger cells in ventral areas of RNpc discharge during hind limb movements and probably project to lumbar levels, although their terminations have not been specifically traced. The cat RNpc also has reciprocal projections to several of the mesencephalic nuclei that provide input including zona incerta, fields of Forel, and the anterior pretectal nucleus. Many studies in species below primates have attributed these connections to RNp; however, the extent that cells projecting to the IO share these connections is not known.

RNpc Function

Only a few studies have been aimed at determining the functional characteristics of RNpc. In the monkey, cells in the rostral dorsolateral extension of the RN show relatively large increases in discharge rate during movements of the contralateral limbs or mouth. The discharge rates are somewhat lower than those seen for RNm, but the cells fire in a consistent fashion during movement, which is similar to cells in RNm.

In the cat, stimulation of fibers exiting RNpc produces poststimulus facilitation in muscles of the contralateral limb. In contrast to RNm, muscles of the proximal limb are more likely to be facilitated than are those of the distal limb. The discharges of cat RNpc cells during reaching are tightly related to movement, but the peak discharge rates are lower than those seen for RNm. Chemical inactivation or activation of the cat RNpc produces disturbances of reaching as well as digit use during a prehension task and can also produce torsion of the neck.

There is a growing body of evidence that the cerebellum is a necessary structure for motor adaptation and the formation of classically conditioned responses. As seen in Figure 6(a), RNpc has a significant projection to the intermediate division of the facial nucleus, which innervates muscles surrounding the eye that are important for protective responses. Inactivation of the RN impairs the expression of the conditioned eyeblink response, and it is likely that projections to the facial nucleus are part of the pathway carrying conditioned responses. Conditioned withdrawal responses of the limbs are eliminated by section of either the superior cerebellar peduncle or the RST. RNpc or RNm or both parts of the RN may be important for relaying learned motor activity from the cerebellum to the spinal cord.

The Ipsilaterally Projecting RNp

RNp is prominent only in primates, and this, coupled with the fact that large RN cells are reduced in

number in apes and humans, has led to the impression that RNp has subsumed function of RNm. Although the properties of RNp have not been studied extensively, it is clear that the nucleus has little commonality with RNm. The expansion of RNp is more likely to represent an expansion in its functionality rather than a trade-off with existing structures.

RNp Anatomy

In contrast to RNm, RNp receives an extensive cortical input from primary motor cortex (Figure 1(b)), as well as from other cortical areas, including the supplementary motor area, anterior cingulate cortex, parietal cortex, and, possibly, frontal eye fields. Some of the projections, such as those from supplementary motor area, are bilateral, although the ipsilateral projections are stronger. The cortical projections to RNp are somatotopic, with face being located medially, arm centrally, and leg laterally. The motor cortical projections to RNp and RNm arise from different populations of cells. The cells projecting to RNp are located dorsally in layer V and do not include Betz cells, which are deep in layer V.

Although cortical inputs to RNp have been reasonably well studied, little is known about other potential inputs. The cerebellar dentate nucleus projects to RNp, but the extent and topography of the input are not known. Essentially nothing is known about other potential brain stem inputs to RNp. The literature on RNp afferents in species below primates, typically, has not distinguished RNpc from RNp.

Efferent projections from RNp to the ipsilateral IO are well documented. The projection is largely confined to the principal nucleus of the IO, which projects to lateral cerebellar cortex and dentate. Other efferents include projections to the pontine reticular formation, but these have not been investigated in detail.

RNp Function

Cells in the monkey RNp discharge at low, irregular rates and show no obvious relation to movement. At first, this seems paradoxical, since the largest input to RNp derives from motor cortex. However, the cortical cells that project to RNp are not Betz cells and may not have movement-related discharge (even a significant percentage of corticospinal cells in motor cortex lack movement-related discharge). IO cells also fire at low irregular rates and have little relation to active movement, so many of the discharge properties of the IO may be present in its afferent sources.

One popular hypothesis about IO function is that it compares actual movement with intended movement to generate an error signal for cerebellar processing. RNp is considered a key element in providing efference copy signals to IO from the cerebellum and motor cortex. Clearly, the properties of RNm and RNpc fit this function better than do those of RNp. If error information is supplied by RNp, then the comparison is likely to be made within or prior to RNp.

Summary

Cells within RN are associated with one of two distinct neural pathways. RNp is part of a major pathway connecting cerebral cortical and dentate output to the ipsilateral principal division of the IO. RNp is well developed only in primates and has often been confused with RNpc in lower species. During movement, RNp cells discharge at low irregular rates that do not relate to movement, which is similar to their target neurons in IO. RNm and RNpc are part of a major pathway that connects cerebellar output to motor and premotor structures in the contralateral brain stem and spinal cord. The cells in these parts of the RN discharge consistently and at high rates when the corresponding parts of the contralateral body are moved. The largest cells in RNm focus on control of movements that involve the digits of the forelimb and hind limb. RNpc projects to the facial nucleus, upper cervical segments, and probably the lumbar cord. It is thought to be important for control of head and face movements, as well as movements of the proximal limbs.

See also: Brainstem Control of Eye Movements; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Cerebellum and Oculomotor Control; Corticomotoneuronal System; Eye and Head Movements; Finger Movements: Control; Kinematics and Dynamics; Motor Sequences; Neural Coding in Primary Motor Cortex; Posterior Parietal Cortex and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape; Sensorimotor Integration: Models.

Further Reading

Asanuma C, Thach WT, and Jones EG (1983) Brainstem and spinal projections of the deep cerebellar nuclei in the monkey, with observations on the brainstem projections of the dorsal column nuclei. *Brain Research Reviews* 5: 299–322.

Burman K, Darian-Smith C, and Darian-Smith I (2000) Macaque red nucleus: Origins of spinal and olivary projections and terminations of cortical inputs. *Journal of Comparative Neurology* 423: 179–196.

Catman-Berrevoets CE, Kuypers HG, and Lemon RN (1979) Cells of origin of the frontal projections to magnocellular and parvocellular red nucleus and superior colliculus in cynomolgus monkey: An HRP study. *Neuroscience Letters* 12: 41–46.

Chapman PF, Steinmetz JE, Sears LL, and Thompson RF (1990) Effects of lidocaine injection in the interpositus nucleus and red nucleus on conditioned behavioral and neuronal responses. *Brain Research* 537: 149–156.

Flumerfelt BA, Otabe S, and Courville J (1973) Distinct projections to the red nucleus from the dentate and interposed nuclei in the monkey. *Brain Research* 50: 408–414.

Fujito Y, Imai T, and Aoki M (1972) Monosynaptic excitation of motoneurons innervating forelimb muscles following stimulation of the red nucleus in cats. *Neuroscience Letters* 127: 137–140.

Gibson AR, Houk JC, and Kohlerman NJ (1985) Magnocellular red nucleus activity during different types of limb movement in the macaque monkey. *Journal of Physiology* 358: 527–549.

Gibson AR, Houk JC, and Kohlerman NJ (1985) Relation between red nucleus discharge and movement parameters in trained macaque monkeys. *Journal of Physiology* 358: 551–570.

Holstege G, Van Ham JJ, and Tan J (1986) Afferent projections to the oribularis motoneuronal cell group. An autoradiographical tracing study in the cat. *Brain Research* 374: 306–320.

Humphrey DR, Gold R, and Reed DJ (1984) Sizes, laminar and topographic origins of cortical projections to the major divisions of the red nucleus in the monkey. *Journal of Comparative Neurology* 225: 75–94.

Humphrey DR and Rietz RR (1976) Cells of origin of corticorubral projections from the arm area of primate motor cortex and their synaptic actions in the red nucleus. *Brain Research* 110: 162–169.

Kennedy PR, Gibson AR, and Houk JC (1986) Functional and anatomic differentiation between parvicellular and magnocellular regions of red nucleus in the monkey. *Brain Research* 364: 124–136.

Lawrence DG and Kuypers HGJM (1968) The functional organization of the motor system in the monkey. II. The effects of lesions of the descending brain-stem pathways. *Brain* 91: 15–36.

Matano S, Baron G, Stephan H, and Frahm HD (1985) Volume comparisons in the cerebellar complex of primates. II. Cerebellar nuclei. *Folia Primatologica* 44(3–4): 182–203.

McCurdy ML, Hansma DI, Houk JC, and Gibson AR (1987) Selective projections from the cat red nucleus to digit motor neurons. *Journal of Comparative Neurology* 265: 367–379.

McKiernan BJ, Marcario JK, Karrer JH, and Cheney PD (1998) Corticomotoneuronal postspike effects in shoulder, elbow, wrist, digit, and intrinsic hand muscles during a reach and prehension task. *Journal of Neurophysiology* 80: 1961–1980.

Mewes K and Cheney PD (1994) Primate rubromotoneuronal cells: Parametric relations and contribution to wrist movement. *Journal of Neurophysiology* 72: 14–30.

Miller LE, van Kan PLE, Sinkjaer T, Andersen T, Harris GD, and Houk JC (1993) Correlation of primate red nucleus discharge with muscle activity during free-form arm movements. *Journal of Physiology London* 469: 213–243.

Miller RA and Strominger NL (1973) Efferent connections of the red nucleus in the brainstem and spinal cord of the Rhesus monkey. *Journal of Comparative Neurology* 152: 327–345.

Morcuende S, Delgado-Garcia JM, and Ugolini G (2002) Neuronal premotor networks involved in eyelid responses: Retrograde transneuronal tracing with rabies virus from the orbicularis oculi muscle in the rat. *Journal of Neuroscience* 22: 8808–8818.

Nathan PW and Smith MC (1982) The rubrospinal and central tegmental tracts in man. *Brain* 105: 223–269.

Oscarsson O (1980) Functional organization of olivary projection to the cerebellar anterior lobe. In: Courville J, De Montigny C, and Lamarre Y (eds.) *The Inferior Olivary Nucleus: Anatomy and Physiology,* pp.279–289. New York: Raven Press.

Padel Y, Angaut P, Massion J, and Sedan R (1981) Comparative study of the posterior red nucleus in baboons and gibbons. *Journal of Comparative Neurology* 202: 421–438.

Padel Y, Rathelot JA, and Vinay L (1995) Analysis of potentials induced in red nucleus neurones from the somaesthetic pathway stimulated at the bulbar level. *Experimental Brain Research* 106(3): 377–390.

Papez JW and Stotler WA (1940) Connections of the red nucleus. *Archives of Neurology and Psychiatry* 44: 776–791.

Pompeiano O and Brodal A (1957) Experimental demonstration of a somatotopical origin of rubrospinal fibers in the cat. *Journal of Comparative Neurology* 108: 225–252.

Pong M, Horn KM, and Gibson AR (2002) Spinal projections of the cat parvicellular red nucleus. *Journal of Neurophysiology* 87: 453–468.

Robinson FR, Houk JC, and Gibson AR (1987) Limb specific connections of the cat magnocellular red nucleus. *Journal of Comparative Neurology* 257: 553–577.

Rutherford JG, Anderson WA, and Gwyn DG (1984) A reevaluation of midbrain and diencephalic projections to the inferior olive in rat with particular reference to the rubro-olivary pathway. *Journal of Comparative Neurology* 229: 285–300.

Strominger NL, Truscott TC, Miller RA, and Royce GJ (1979) An autoradiographic study of the rubroolivary tract in the rhesus monkey. *Journal of Comparative Neurology* 183: 33–46.

Sybirska E and Gorska T (1980) Effects of red nucleus lesions on forelimb movements in the cat. *Acta Neurobiologiae Experimentalis* 40: 821–841.

Ten Donkelaar HJ (1988) Evolution of the red nucleus and rubrospinal tract. *Behavioural Brain Research* 28: 9–20.

van Kan PL and McCurdy ML (2002) Discharge of primate magnocellular red nucleus neurons during reaching to grasp in different spatial locations. *Experimental Brain Research* 142: 151–157.

van Kan PLE, Horn KM, and Gibson AR (1994) The importance of hand use to discharge of interpositus neurones of the monkey. *Journal of Physiology* 480(Part 1): 171–190.

Voneida TJ (1999) The effect of rubrospinal tractotomy on a conditioned limb response in the cat. *Behavioural Brain Research* 105: 151–162.

Referentiality and Concepts in Animal Cognition

K Zuberbühler, University of St. Andrews, St. Andrews, UK

© 2009 Elsevier Ltd. All rights reserved.

> No name belongs to a particular thing by nature, but only because of the rules and usage of those who establish the usage and call it by that name. (Plato, *c*. 360 BC)

Introduction

Communication happens when individuals exchange information by means of a common system of behavior, an inherently social event. Human language is a particularly remarkable example, and the quest for its evolutionary origins is a scientific problem of continued interest. One popular approach is to examine the cognitive capacities of nonhuman animals, especially those of primates, during natural acts of communication. The comparative approach is based on the premise that human language is built on a number of cognitive capacities with independent evolutionary histories, some with phylogenetic roots deep in the primate lineage.

Precisely what capacities are required for language? An article by Hockett in 1960 has been particularly influential. He identified 13 different design features that are common to all languages. Some are shared with animal communication systems; others appear to be uniquely human. The distinction between human speech and primate vocal behavior is particularly interesting since the two differ in only four design features (Table 1).

Referential Communication and Semanticity

A key property of language is that it allows speakers to use distinct utterances in order to convey information about objects and events in the outside world. This by itself is not a uniquely human trait, however. Honeybees are able to communicate about the location of distant food sources, and many animal species produce acoustically different alarm calls for different predators. For nonhuman primates, the classic example for such referential signals is the vervet monkey alarm calls. These primates produce acoustically distinct vocalizations to some of their predators, such as eagles, leopards, or pythons. When others hear a group member's alarm call to a python, they respond in ways that suggest the call is meaningful to them, for instance, by scanning the surrounding area for the snake they assume is present. Another example is the West African Diana monkeys, which produce acoustically different alarm calls to leopards and crowned eagles. These calls convey the biological class of the predator and are not simple responses to situational circumstances or perceived threat (Figure 1).

Eavesdropping is common in alarm situations, both within and across species. For example, yellow-casqued hornbills live sympatrically with Diana monkeys in the Taï Forest, Ivory Coast, and occasionally feed on the same trees. Crowned eagles threaten both species, but only leopards also attack the monkeys. Field playback experiments have shown that the hornbills not only recognize these two predators by their calls but also discriminate between the eagle and leopard alarm calls of the Diana monkeys, despite the fact that acoustic differences are very subtle (Figure 2).

Although the relation between monkey alarm calls and their meanings is arbitrary, the ontogenetic processes leading to the link are not. Infant vervet monkeys give eaglelike alarm calls to numerous flying objects, such as storks or falling leaves, and only with experience do they learn to restrict call use to the genuinely dangerous raptor species. The main alarm-call types develop fairly rigidly, as do the underlying psychological states that elicit them. Social learning subsequently leads to fine-tuning of the range of events that warrant call production, within the predetermined categories. Although widely accepted, the empirical basis for this model is not very strong, suggesting that further research could alter our understanding of how primates acquire semantic competence.

What sorts of mental processes do animals experience when responding to predator-specific alarm calls? Are these calls truly meaningful to them, in the sense that they are linked with specific mental representations, or are call production and comprehension possible without complex processing – a kind of cognitive knee-jerk reflex? These are difficult problems and some researchers thus prefer the term 'functional reference' to acknowledge that what superficially resembles human referential communication may be the product of different cognitive processes. For nonhuman primates, at least, there is relatively good evidence that individuals experience some fairly specific mental representations when hearing each other's calls.

In one study, the cognitive processes experienced by primates when listening to their own eagle or leopard alarm calls were investigated. The question was whether monkeys simply responded to the physical

Table 1 Design features in which human speech is thought to differ from primate communication

Design feature	*Human language*	*Primate communication*
Displacement	Information about past, present, and future events	Communication about present events only
Productivity	Generation of novel meanings; familiar pieces are reassembled into new utterances	No semantic creativity; closed call systems
Transmission	Socially learned and culturally transmitted	Genetically hard-wired signals
Duality of patterning	Small stock of meaningless phonemes to build meaningful morphemes	Holistic utterances, no permutations

Adapted from Hockett CF (1960) The origin of speech. *Scientific American* 203: 89–96.

features of the calls or whether they were capable of generating a mental representation of the corresponding predator (i.e., the semantic content of their calls). In the experiment, different groups of Diana monkeys were exposed to one of three different prime stimuli: the shrieks of a crowned eagle, a male's alarm calls to an eagle, or a male's alarm calls to a leopard (Figure 3). Five minutes later, the same monkey group heard a second stimulus, the shrieks of a crowned eagle. Hence, depending on the condition, the two playback stimuli did or did not correspond semantically. Results showed that eagle shrieks did not elicit new alarm calls when monkeys were previously primed with eagle shrieks or monkey eagle alarm calls, in contrast to when they were primed with monkey leopard alarm calls (Figure 3). Therefore, when responding to each other's vocalizations, nonhuman primates activate mental representations, which are responsible for subsequent behavioral patterns.

The Semiotic Triangle

In humans, the relation between speech utterances and world entities is mediated by mental concepts in the form of a semiotic triangle (Figure 4). A number of studies suggest that this model may also be useful in explaining some animal communication systems. For phylogenetic reasons, the focus is on nonhuman primates, although an increasing number of nonprimate species have demonstrated that some of the crucial cognitive capacities are not uniquely primate.

Thoughts

Accessing mental processes in animals is not a trivial task – the 'other mind' problem in philosophy. Although it is possible to systematically study an animal's behavioral responses to a stimulus, the only way to access underlying cognitive processes is by inference. A common custom is to select the most parsimonious mechanisms from a selection offered by human psychology, but this practice habitually causes discomfort and ongoing debate. Nevertheless, much empirical progress has been made in describing how animals mentally represent their worlds – that is, the kinds of conceptual structures they build and maintain to interact with their environment. It is important to know that in humans mental concepts can exist independently of language. Some mental concepts are well established in prelinguistic infants well before language is acquired. In some cases, infants have demonstrated competence in dealing with abstract mental concepts in ways that exceeded that of adults.

One important technique to study concept formation in animals is to train individuals to respond to a number of real-world exemplars, which are members of a particular category. In a subsequent testing phase, subjects are then asked to make choices in response to novel stimuli. In one pioneering study, pigeons learned to distinguish between unfamiliar photographs that did or did not contain trees, suggesting they possessed a natural concept for tree. An intriguing example is honeybees, which have been shown to form concepts of sameness and oddity. Thus, complex abstract concepts are widespread throughout the animal kingdom and are not restricted to primates or even vertebrates.

Human intelligence is thought to have evolved in the social realm, however, and research on social concepts is therefore of particular interest. One study of long-tailed macaques has become somewhat of a classic. Monkeys were trained to select photographs showing pairs of familiar group members, either mother–offspring dyads or other pairings. The study animals successfully mastered novel dyads, demonstrating that they utilized a concept analogous to the human mother–child affiliation. It has also been claimed that chimpanzees are able to match black-and-white photographs with faces of unfamiliar mothers and their sons, providing further support for mother–offspring social concepts in primates but also suggesting that phenotypical markers play a role during categorization. In another study, rhesus monkeys watched video clips showing an agonistic interaction between two unfamiliar conspecifics. Not only was it

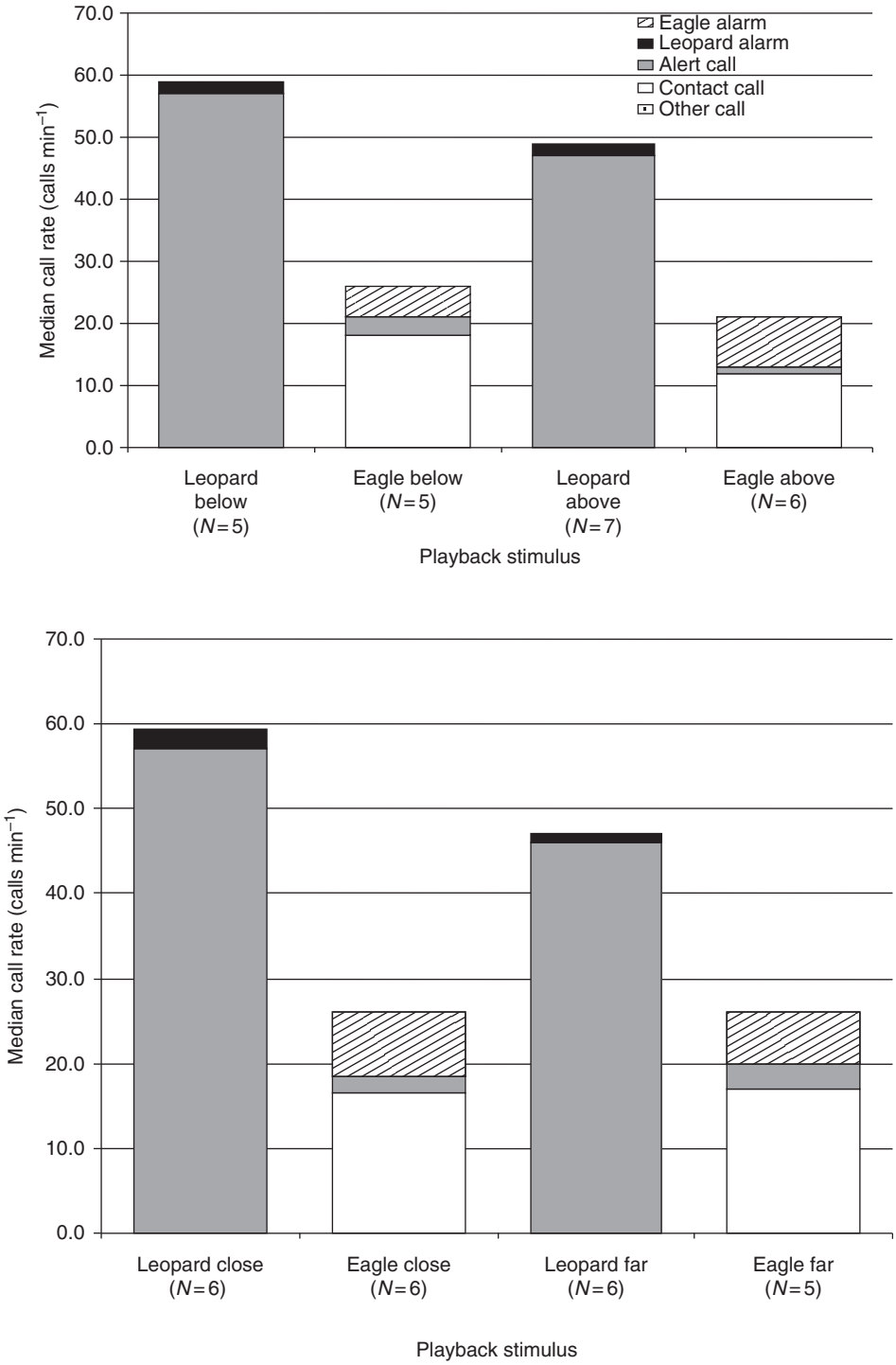

Figure 1 Diana monkey alarm calls convey information about the biological class of the predator, not its distance or direction of attack. Nearby group members respond with the same vocal behavior to the growls of a leopard (or shrieks of a crowned eagle, respectively) regardless of whether or not the calls were broadcast at close distance or from within the canopy above or on the ground. Adapted from Zuberbühler K (2000) Referential labelling in Diana monkeys. *Animal Behaviour* 59: 917–927.

possible to train subjects to select the dominant individual in the interaction but also the subjects continued to do so when watching novel video clips – a demonstration that they possessed a social concept for dominance.

Despite these examples of primate social intelligence, there is a widespread consensus that nonhuman primates are unable to attribute mental states to one another, an activity that humans engage in from early childhood. However, empirical work has led to some

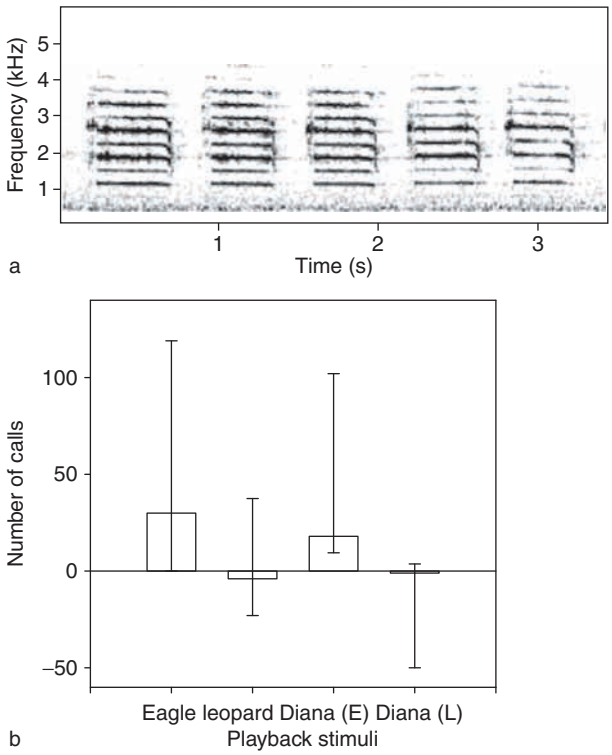

Figure 2 Responses of yellow-casqued hornbills of Tai forest to various predator-related acoustic stimuli. (a) Spectrogram of typical 'long' calls. Rate of production of this call type increased after playback of eagle calls and Diana monkey eagle alarm calls. Five individual calls are presented here. (b) Median and interquartile ranges of differences between the number of hornbill calls before and after playback of different stimulus types. Hornbills increased numbers of calls after playback of eagle shrieks significantly more than after playback of leopard growls. The same differences were found after playback of Diana monkey eagle and leopard alarm calls (Diana E and L). Adapted from Rainey HJ, Zuberbühler K, and Slater PJB (2004) Hornbills can distinguish between primate alarm calls. *Proceedings of the Royal Society of London Series B-Biological Sciences* 271: 755–759.

interesting amendments. Chimpanzees have been found to be capable of predicting the movements of more dominant competitors based on assessing the competitor's visual perspective. Related to this, when chimpanzees and other great apes interact with a human experimenter who is looking behind a barrier, they tend to position themselves such that they can see what the experimenter is looking at, suggesting that they attempt to take the experimenter's visual perspective. A number of non-ape species have shown comparable skills. When given the choice between two experimenters who were capable or incapable of seeing a food item, rhesus monkeys preferentially retrieved the food items from the visually unaware experimenter. In an experiment with ravens, the view of individuals during caching was manipulated, thereby creating competitors that were either knowledgeable or ignorant of the cache location. Results

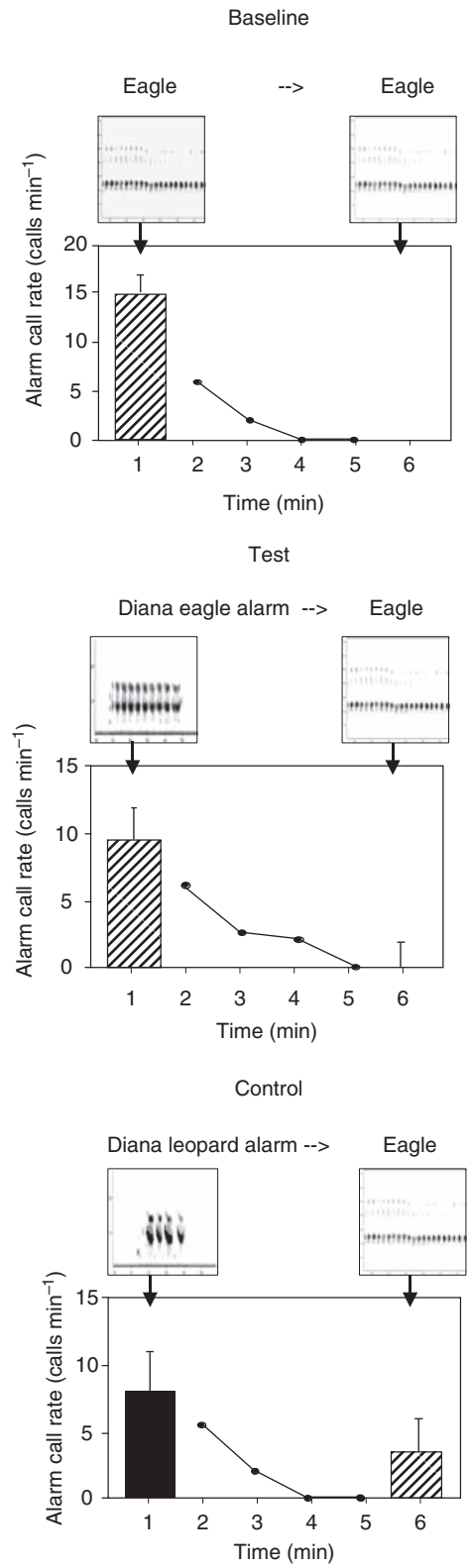

Figure 3 Design and results of a prime probe experiment to investigate the cognitive processes underlying Diana monkey alarm-call production. Analogous results were obtained if leopard growls were used as predator stimuli. Adapted from Zuberbühler K, Cheney DL, and Seyfarth RM (1999) Conceptual semantics in a nonhuman primate. *Journal of comparative Psychology* 113: 33–42.

demonstrated that ravens adjusted their subsequent behavior depending on a competitor's knowledge – that is, whether or not it previously had witnessed the caching event. In addition, ravens can follow gaze around barriers similar to primates, further suggesting that they make judgments of the visual perspective of other individuals.

In summary, although primates and other animals are able to understand what others see or have seen in the past (i.e., that they know), there is currently no evidence that they can reflect on what they know or that they recognize and share each other's intentions, which has potentially significant consequences, particularly for cultural development.

Symbols

Vocal repertoires Primate vocal repertoires, and those of many other species, are often said to be genetically predetermined, suggesting that individuals have little control over vocal production, similar to some nonverbal utterances in humans, such as laughing or crying. Although the main thrust of this argument remains unchallenged, some empirical studies have provided a number of clarifications.

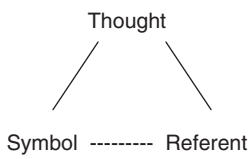

Figure 4 The semiotic triangle of referential communication.

First, several studies have shown that adult primates can change the acoustic structure of some of their calls, usually following significant social events. In Campbell's monkeys, adult females exchange contact calls as part of their daily social interactions. Each individual produces a certain number of acoustic variants, but these patterns collapse after changes in group composition. Playback of females' currently produced call variants reliably triggered vocal responses from other group members, whereas the same female's former, no longer used call variants caused long-term cessation of vocal behavior, indicating that the calls form part of a long-term social memory. Chimpanzee pant hoot vocalizations are a further example of vocal behavior that can undergo acoustic modifications depending on the social circumstances.

Second, the apparatus used by nonhuman primates for vocal production is not fundamentally different from that of humans. Sound is generated by the larynx, from which it radiates into the supralaryngeal vocal tract that acts as an acoustic filter. During speech, the shape of the vocal tract changes rapidly due to the movements of the various articulators, such as the lips, tongue, jaw, or larynx. The same basic processes have been found in Diana monkeys. Males produce alarm calls that contain formants with variable structure. Crowned eagles elicit calls with little or no formant transition, whereas leopards elicit calls with strong transitions (Figure 5). These acoustic modifications are the result of constrictions in their vocal tract formed by various articulators, similar to what takes place during speech production.

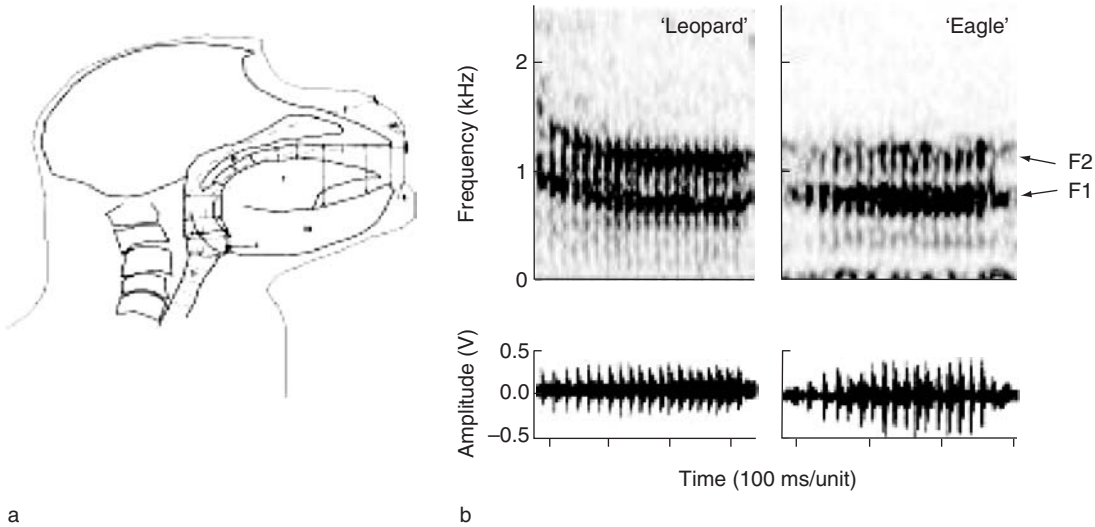

Figure 5 Vocal production mechanisms in a nonhuman primate. (a) Cross section of a Diana monkey vocal tract highlighting locations where acoustic filtering takes place due to constrictions. T, tongue; M, mandible; uL, upper lip; lL, lower lip; P, palate; Tr, Trachea. (b) Formant structure and transitions in Diana monkey eagle and leopard alarm calls. F1, first formant; F2, second formant. Adapted from Riede T, Bronson E, Hatzikirou H, and Zuberbühler K (2005) Vocal production mechanisms in a non-human primate: morphological data and a model. *Journal of Human Evolution* 48: 85–96.

Despite these similarities, there are vast differences between human and nonhuman vocal production. Humans are enormously skillful in producing rapid and precise articulatory movements, particularly with the tongue. As a result, humans are able to create a rich array of sounds, whereas sound production in nonhuman primates seems to be confined to a limited number of call types. For example, whereas modern English operates with approximately 40 different phonemes, male Diana monkeys possess only two (Figure 5). It has been suggested that this human specialization may be linked to a particular gene, *FoxP2*, since damage to this gene causes a severe speech disorder due to a lack of orofacial control. The human *FoxP2* is slightly different from those of the other nonhuman primates, suggesting that human–primate difference in orofacial control could be caused by this genetic difference.

Combinations of calls In contrast to speech, animal communication is considered to be holistic and lacking features that mark particular relationships between one call and another – Hockett's notion of duality of patterning. Although some animals combine acoustic units into patterned sequences, this behavior usually only serves to advertise identity, status, or quality, with no evidence of any sort of referential function.

A number of studies have provided empirical evidence that may require revision of current theory. At least two species of nonhuman primates have been found to combine calls into structurally more complex units, thereby modifying the semantic content of the entire utterance. An example is provided by free-ranging Campbell's monkeys. Like the Diana monkeys described previously, male Campbell's monkeys produce acoustically distinct alarm calls to leopards and eagles. However, if males witness other disturbing events that are not directly linked with these predators (e.g., a falling tree or a fleeing forest duiker), they usually produce a pair of low-sounding 'booms' before giving a series of alarm calls. Playback experiments have shown that other monkeys ignore such boom-introduced alarm calls, suggesting that the booms selectively change the meaning of subsequent alarm calls from predator-specific alarm signals to more general sign of disturbance (Figure 6).

Second, work on free-ranging putty-nosed monkeys in Nigeria has revealed similar results. In this species, the males produce two basic alarm call types, the 'hacks' and 'pyows.' These calls are not predator specific because males produce them to both eagles and leopards. Instead, callers concatenate the two calls into different sequence types, and these sequences are predator specific. In other words, putty-nosed monkeys encode predator class at the level of the call sequence and not at the level of the individual call. In addition, the males produce a peculiar pyow–hack sequence used to instigate group progression.

In summary, some primates do assemble calls into higher-order sequences and thereby generate information that is unrelated to the meaning conveyed by constituent elements. Although these combinatorial

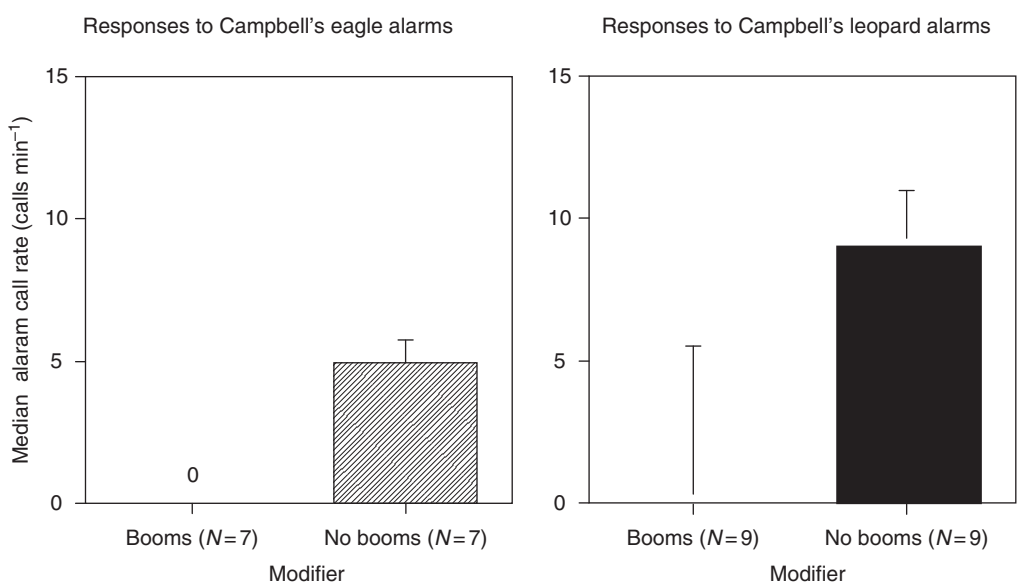

Figure 6 Diana monkey responses to Campbell's monkey eagle or leopard alarm calls in the presence or absence of preceding boom calls. Artificially adding boom calls prior to an alarm-call sequence deletes the predator-specific meaning of these calls. The hatched bar indicates the median number of Diana monkey eagle alarms in the first minute after playback, the solid bar indicates the median number of Diana monkey leopard alarm calls. Adapted from Zuberbühler K (2002) A syntactic rule in forest monkey communication. *Animal Behaviour* 63: 293–299.

systems may lack the generative power of recursion, they provide an efficient means of increasing message repertoire without the need to invent new calls.

Referents

The final element in the semiotic triangle is the referent – that is, the kinds of entities that trigger vocal behavior in nonhuman primates. Most progress has been made in two contexts that are of particular evolutionary relevance: finding food and avoiding predators.

Food discovery Many primates vocalize when finding food. Although the underlying mental processes are not well understood, two factors appear to be particularly relevant: the social setting during call production and the food. Food-associated calls attract nearby listeners, and in some species callers exert some control over this process by adjusting call production. For example, chimpanzees are more likely to produce food-associated calls, the so-called rough grunts, if the food source is divisible and anticipated competition is likely to be low, a pattern also found in bonobos. These processes are not limited to primates. Male chickens produce specific calls upon discovering food, more likely so in the presence of a potential mate. Hens respond to the calls with anticipatory feeding movement, suggesting that the calls contain basic referential information. In ravens, there is a relation between food type encountered by the caller and call rates produced.

In chimpanzees, food quality is unrelated to call rates. In one study, calls from captive individuals were recorded in response to nine different food items, ranked by the individuals as high, medium, or low preference. There was no evidence for food labels with the low and medium preference class; acoustic differences were only significant between the three preference classes. Surprisingly, however, calls given to the high preference foods were acoustically distinct, suggesting that referential labels for food items can emerge in captive conditions. In a related study, it was investigated whether these call variants are meaningful to recipients. Rough grunts, given in response to bread and apples, were played back to recipients, and the subsequent amount of search effort underneath an apple and a bread tree was measured. Results demonstrated that the focal animal spent significantly more time and search effort underneath the correct tree (i.e., the tree corresponding to the call type played before), demonstrating that chimpanzees are able to attend to the subtle acoustic differences inherent in their food calls and that they are meaningful to them (Figure 7).

Predator detection Predator-specific alarm calls are a common feature of primate and nonprimate vocal behavior. Monkeys convey the biological category of the predator, but other information is sometimes also encoded. For example, some species, such as white-browed scrubwrens or California ground squirrels, adjust their calling behavior as a function of predator distance. Domestic chickens mainly distinguish between events on the ground or in the air, regardless of predator species. Gunnison's prairie dogs' alarm calls differ depending on the color of

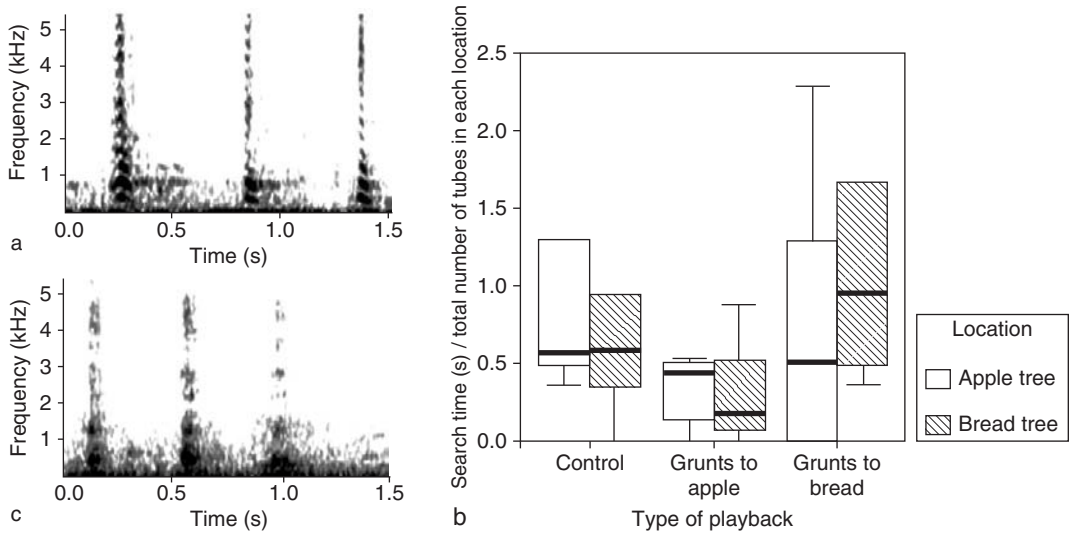

Figure 7 Chimpanzee food grunts are referential signals. (a) Rough grunts given in response to two different food items – bread (top) and apples (bottom). (b) Search behavior of the focal animal in response to playbacks of apple grunts and bread grunts given by different group members. The subject devoted more search effort underneath the food tree, indicated by the food grunts played back earlier. Adapted from Slocombe KE and Zuberbühler K (2005) Functionally referential communication in a chimpanzee. *Current Biology* 15: 1779–1784.

clothes and general shape of human experimenters, and black-capped chickadees' alarm calls convey the size of potential predators with the number of D-notes in the 'chick-a-dee' alarm calls, a relevant source of information for recipients.

Similar to food calls, alarm call production in primates is under significant social control. In vervet monkeys, high-ranking individuals alarm-called at higher rates than low-ranking ones. The presence of offspring also increased alarm call rates in females, whereas males called at higher rates when with an adult female than when with other adult males. Primates may even keep track of the alarm-calling behavior of other group members. A field study reported that male Thomas langur responded with alarm calls to a model tiger and continued to do so until every member in their group had give at least one alarm call, suggesting that males keep track of which group members have and have not yet called.

Social events Primates routinely use vocalizations during social interactions, and in some cases the acoustic structure of calls varies with the social situation. The social calls of baboons have been studied in depth. During social interactions, subordinate animals tend to produce fear barks to approaching dominant group members, whereas dominants give grunts to signal their willingness to interact peacefully. These types of call exchanges have been used to experimentally mimic different types of social interactions to nearby listeners. In one experiment, individuals not only recognized the rank orders of the different group members but also recognized to which matrilineal kin group the individuals belonged.

Similarly, chimpanzees often scream when participating in agonistic interactions. In one study at Budongo Forest, Uganda, the vocal behavior of a large number of different individuals was recorded during agonistic interactions, such that each individual contributed the same number of calls both as a victim and as an aggressor. Results showed that the screams given in the two roles differed in a number of subtle acoustic features, providing reliable cues of the identity of the individuals involved and their social role. The degree to which listeners take advantage of this rich set of acoustic information needs to be explored.

Conclusion

The long-standing notion that the meaning of an utterance is nothing more than its use also has some relevance in animal communication. Nonhuman primates heavily rely on vocal communication, particularly those species whose primary habitat is the forest, where visual contact is limited and unreliable. Primates actively attend to each other's calls, and they are able to extract multifaced levels of information from individual calls and the way in which calls are arranged into different sequences. There is also evidence that primates understand something about the causal structure of the events that were responsible for the production of the various vocal signals produced by conspecifics and other animals. At the same time, primates have only limited control over their articulators and the vocal signals they can produce, but many of the fundamental principles of speech production are already well established.

A neglected area of research relates to the problem of how much primates take their audience into account and whether callers intentionally inform each other about events they have just witnessed rather than responding to the events directly. Anecdotal observations with chimpanzees suggest that individuals utilize vocalizations to affect the behavior of particular group members, although there is no clear evidence that this is for the sake of sharing information only. Chimpanzees are able to make judgments about other's intentions, even without being specifically trained to do so. For example, individuals behave differently depending on whether a human experimenter is unwilling or unable to give them food. Similarly, free-ranging baboons recognize when a call is directed at themselves rather than at other individuals. After hearing a threat grunt of a more dominant female, subjects moved away more quickly after having received aggression than grooming from that female, suggesting that baboons make inferences about the target of the caller's attention – a first crucial step toward the recognition of other individuals' intentions and motives.

One central Hockettian claim that has received comparably little attention throughout the years is that animal communication is thought to take place in the here and now only (Table 1). This point is interesting because the past is clearly an important psychological variable for many animal species. For example, there is evidence that scrub jays are able to recall particular food caching events in the past, including where and when they have occurred. Free-ranging gray-cheeked mangabeys of Kibale Forest, Uganda, can take the weather condition over the previous few days into account when searching for figs. The monkeys were more likely to revisit fig trees following several days of warm and sunny weather compared to after cooler, cloudier periods, presumably because they knew that temperature and solar radiation affect the ripening rate of figs. Given that animals are able to process information about past events, particularly in the context of foraging,

why is there no evidence that they also communicate about them? Perhaps there are deep reasons for this; perhaps future empirical work will reveal unexpected communication capacities.

Human language is a culturally transmitted behavior. It is maintained by children who try to understand and mimic utterances produced by relevant social figures, a process that starts very early in life. A number of studies have shown that nonhuman animals, including dogs, birds, and apes, can be brought to understand spoken language to some degree. Social processes, particularly the ability to share attention and intention, appear to play a crucial role in language acquisition in humans, an ability that emerges early on during human ontogeny. The degree to which joint attention and joint intention and related social processes play a role in the acquisition of meaning in animal communication is not well understood. The current state of the art is that even human enculturated apes show a total lack of attempts to share attention with others. This peculiar lack of social intelligence is widely regarded as the main reason why humans' common system of behavior to exchange information, spoken language, is so much more complex and powerful than those of any animal species. As usual, however, future work may tell a different story.

See also: Animal Communication: Honesty and Deception; Facial Expression in Primate Communication; Game Theory and the Economics of Animal Communication; Language: Nonhuman Animals; Primate Communication: Evolution; Signal Design Rules in Animal Communication; Social Interaction; Social Interaction Effects on Reward and Cognitive Abilities in Monkeys.

Further Reading

Cheney D and Seyfarth R (1992) How monkeys see the world. *Behavioral and Brain Sciences* 15: 135–182.

Crystal D (1997) *The Cambridge Encyclopedia of Language*. Cambridge, UK: Cambridge University Press.

Emery NJ and Clayton NS (2005) Evolution of the avian brain and intelligence. *Current Biology* 15: R946–R950.

Fitch WT (2000) The evolution of speech: A comparative review. *Trends in Cognitive Sciences* 4: 258–267.

Hockett CF (1960) The origin of speech. *Scientific American* 203: 89–96.

Hurford JR (2007) *The Origins of Meaning*. Oxford, UK: Oxford University Press.

Lieberman P (2002) On the nature and evolution of the neural bases of human language. *Yearbook of Physical Anthropology* 45: 36–62.

Marcus GF and Fisher SE (2003) FOXP2 in focus: What can genes tell us about speech and language? *Trends in Cognitive Sciences* 7: 257–262.

Marler P, Evans C, and Hauser MD (1992) Animal signals: Reference, motivation, or both? In: Papoucek H, Jürgens U, and Papoucek M (eds.) *Nonverbal Vocal Communication: Comparative and Developmental Approaches*, pp. 66–86. Cambridge, UK: Cambridge University Press..

Pepperberg IM (1999) *The Alex Studies*. Cambridge, MA: Harvard University Press.

Segerdahl P, Fields W, and Savage-Rumbaugh S (2005) *Kanzi's Primal Language*. New York: Palgrave.

Seyfarth RM and Cheney DL (2003) Signalers and receivers in animal communication. *Annual Review of Psychology* 54: 145–173.

Tomasello M (1999) *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Tomasello M and Call J (1997) *Primate Cognition*. New York: Oxford University Press.

Zuberbühler K (2003) Referential signalling in non-human primates: Cognitive precursors and limitations for the evolution of language. *Advances in the Study of Behavior* 33: 265–307.

Zuberbühler K (2006) Alarm calls – Evolutionary and cognitive mechanisms. In: Brown K (ed.) *Encyclopedia of Language and Linguistics*. Oxford: Elsevier.

Relevant Websites

http://www.budongo.org – Budongo Conservation Field Station.

http://www.psychol.cam.ac.uk – Cambridge University, Department of Experimental Psychology.

http://www.gsu.edu – Georgia State University; see Language Research Center.

http://www.wjh.harvard.edu/~mnkylab – Harvard University, Cognitive Evolution Laboratory.

http://www.haskins.yale.edu – Haskins Laboratories (Yale University).

http://www.eva.mpg.de – Max Plank Institute for Evolutionary Anthropology; see Wolfgang Köhler Primate Research Center.

http://psy.st-andrews.ac.uk – School of Psychology, University of St. Andrews; see Research – Origins of Mind.

http://www.csl.sony.fr – Sony Computer Science Lab Paris.

http://nationalzoo.si.edu – Smithsonian National Zoological Park; see Think Tank.

http://www.alexfoundation.org – The Alex Foundation.

Reflex Circuits

T R Nichols, Emory University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The reflex has long been a central concept in physiology and psychology. During the late nineteenth and early twentieth centuries, the reflex was considered a unit of function and was used as the basis for the experimental investigation of the central nervous system. However, as recognized by Sir Charles Sherrington, the simple reflex arc is an abstract concept since the associated pathways are embedded in a larger system of neural circuits. Sherrington recognized that reflexes vary in complexity and that coordination results in part from a compounding of reflexes that may work together or in opposition. The philosopher John Dewey criticized the concept of the reflex arc in psychology for similar and for several additional reasons. A simple stimulus and resulting response may be a convenient experimental or clinical tool but does not capture the closed-loop nature of most sensorimotor processes, nor does this notion recognize that most reactions occur as components of ongoing behaviors. The term 'reflex arc' implies a stimulus and response but makes no provision in itself for communication from response back to stimulus. An exemplary notion in the motor control literature is that of clasp knife inhibition. This response is a useful clinical sign, but the functional interpretation common in textbooks that this response normally serves a protective function in the musculoskeletal system was taken out of the context of normal function. The implicated group III and group IV receptors in fact do not require excessive forces for stimulation, and the profound inhibition observed under pathological circumstances has not been observed in the normal state. How the associated spinal pathways operate in the normal animal in a way that is consistent with stability remains unknown.

The above issues have been extended in contemporary work on spinal circuits in several ways. First, it is generally recognized that reflex circuits are organized hierarchically. Sensory feedback may influence motor neuronal output directly, in a distributed fashion to motor neurons controlling other muscles directly or through interneuronal circuits, or through pattern-generating networks. Second, sensory feedback represents a number of different variables, and the composition of a set of such signals is dependent on behavioral or mechanical context. Third, reflex pathways are embedded in closed-loop neuromechanical circuits. For example, while signals representing muscular length are processed and distributed by circuits in the spinal cord, muscular force is transformed mechanically by the architecture and dynamic properties of the musculoskeletal system into changes in length. These biomechanical transformations are essential components of the control system. Rather than merely providing different responses to different stimuli, this control system endows the motor actuator with appropriate mechanical properties and promotes coordination and stability among the segments of the body. Each of these points will be considered in turn.

The Hierarchical Organization of Spinal Circuits

The Regulator

An important structural basis for the integration of reflex circuits in the spinal cord is the hierarchical design of the associated pathways. At the most elemental level is the monosynaptic reflex that arises from muscle spindle receptors, excites motor neurons through a single synaptic connection, and generally results in force production and then changes in lengths of muscle units, completing a loop back to the muscle spindle. Although this pathway is usually viewed as a private pathway, and is therefore modular in structure, the pathway arising from force receptors is far from private, as will be argued in the next section. The feedback from muscle spindles and other receptors is also distributed to other muscles through interneuronal networks. The muscle targets may cross the same or different joints in the limb or may be located in other parts of the musculoskeletal system. At this hierarchical level (the regulator, Figure 1), monosynaptic and polysynaptic pathways form distributed networks in the spinal cord.

Pattern Generating Networks

It has been established that the spinal cord contains pattern-generating networks that are capable of rhythmic activity in the absence of sensory feedback. Additional evidence suggests that the motor commands are organized in the form of synergies of muscles or motor units. It has been established that sensory feedback from muscles and skin impinges on these central pattern generators and is capable of modulating the magnitude of the output signals or resetting the rhythmic activity. On the basis of more-recent evidence, it has

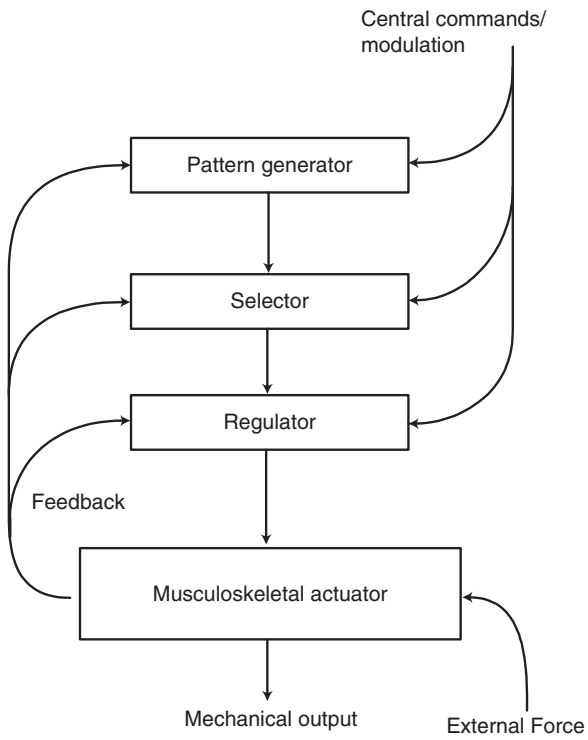

Figure 1 The organization of circuits of the spinal cord is hierarchical. The musculoskeletal actuator consists of the musculoskeletal system. This actuator has intrinsic properties that provide an initial mechanical response to perturbations. Feedback from a variety of receptors reaches the spinal cord and influences a variety of spinal pathways. The pattern generator sets the timing of muscular activity during locomotion. The selector determines which groups of muscles or motor units are active during a given task. The regulator is a network of pathways that receives sensory feedback. This module regulates the stiffness of the limb and promotes interjoint coordination. It is capable of mediating basic postural responses in conjunction with sensory feedback from other limbs. Central commands or signals for modulation from the brain stem or other areas in the spinal cord affect these hierarchical stages.

been suggested that the pattern generators are composed of two levels of processing. The first generates the appropriate timing signals, and the second selects the appropriate synergies on the basis of the behavioral context. For example, in walking on ramped surfaces, different patterns of muscular activity are required to achieve the appropriate balance of forces. Feedback that reflects the orientation of the animal with respect to gravity may influence the selection level appropriately. Other networks mediate responses to postural disturbances, although it is not clear whether these networks are located in the spinal cord, brain stem, or both. These networks are driven by feedback from skin and muscles, and the outputs appear to be organized into synergies as well. These hierarchical levels are shown schematically in Figure 1. Sensory feedback is essential to the operation of each of these levels.

Sensory Feedback Is Multivariable and Context-Dependent

Reflexes are usually defined in an experimental context. One or more channels of sensory information can be stimulated either electrically or mechanically. The response will result from the actions of one or more reflex pathways associated with any of the above hierarchical levels. The reflex thus defined might or might not be composed of a natural combination of individual pathways in view of the artificial or isolated nature of the stimulus. During natural movements, sensory pathways will be recruited or modulated up or down in a variety of combinations and engage different combinations of the hierarchically arranged pathways. Resulting alterations in motor neuron firing will bring about mechanical changes in the musculoskeletal system, closing the loop back to the sensory receptors. The sensory pathways can be modulated through neuronal mechanisms in the spinal cord. They can also be modulated by virtue of the influence of mechanical context on sensory transduction.

Golgi Tendon Organs

The fundamental mechanical variables, length and force, are sensed by muscle spindle receptors and Golgi tendon organs, respectively. Golgi tendon organs are particularly sensitive to active contractile forces and provide interneurons in the spinal cord with a measure of total muscular force although the feedback may fade as a result of presynaptic inhibition in the spinal cord. Increased force can occur during active shortening when muscles contract concentrically, during active lengthening due to intrinsic mechanical properties of contracting motor units, or during active lengthening when the stretch reflex recruits additional motor units. There is electrophysiological evidence of oligosynaptic inhibition of parent motor neurons from tendon organs, but this connection is rather weak and has not been observed in muscle stretch. Pathways arising from these receptors can, however, excite motor neurons of the parent muscle during locomotion but generally inhibit motor neurons of muscles spanning joints or axes of rotation different from the parent muscle's over a variety of behavioral conditions.

Muscle Spindle Receptors

Muscle spindle receptors respond in a nonlinear fashion to changes in the lengths of muscle fibers in large part because of the mechanical properties of the contractile intrafusal muscle fibers. There are two types of these receptors, primaries and secondaries. The nonlinear properties and dynamic responsiveness are similar in the two types but are more extreme in the

primary receptors. A functional consequence of these nonlinear properties is that the response properties of the receptors depend on mechanical context, as illustrated by the following examples. First, muscle spindles increase firing to stretch more than they decrease firing to release and so are more sensitive during eccentric than concentric contractions. Second, these receptors have a decreasing sensitivity to velocity as velocity increases. Third, muscle spindle receptors appear to be more sensitive to length change during quiet standing than ongoing locomotion. Responses to increases in muscle length occurring after a period of isometric loading are characterized by an initial burst and a large dynamic response. Responses to increases in length occurring during ongoing movements are characterized by a smaller dynamic response, lack of an initial burst, and overall more linear behavior. The nonlinear properties described above may be more important during transitions between quiet standing and movement or during changes in movement direction.

Modulation of discharge characteristics of muscle spindle receptors also occurs through neural inputs by virtue of the gamma motor innervation of intrafusal fibers. Some gamma motor neurons apparently fire according to a template of the expected movement, while others bias the firing of the spindle receptors. A third type can modulate the sensitivity of the spindle receptors during certain phases of movement or perhaps in the process of learning a novel task.

Since muscle spindles signal changes in length of muscle fibers, increases in muscular length that can occur during some active contractions can occur mostly in the tendon, resulting in little change in firing rate of the spindle receptors. This situation can occur during walking down a ramp, a task that features eccentric muscular contractions. Under these conditions, force feedback from Golgi tendon organs would predominate, illustrating the importance of context in the composition of sensory feedback.

Primary receptors of muscle spindles project back to the parent muscle, to other levels in the hierarchy of spinal circuits, and also to the cells of origin of projection pathways to the cerebellum, cortex, and other supraspinal structures. Although individual muscle spindles code the length changes of parent muscle units, the projection cells of the spinocerebellar tract to which they project code more global variables related to limb length and orientation, indicating that considerable synthesis of proprioceptive information occurs at the level of the spinal cord. This information could be used for the local regulation of balance and stability. Muscle spindles in muscles of the neck also have specialized projections to regulate muscular tone and influence the selection modules. These pathways interact with feedback from the vestibular apparatus to regulate patterns of muscular activity during locomotion on irregular and sloped surfaces.

Secondary spindle receptors project to interneurons in the spinal cord. Pathways arising from secondaries are more divergent in the spinal cord than are pathways from the primaries but include excitatory actions on motor neurons of the parent muscle. The projection interneurons fire rhythmically during locomotion. These interneurons also receive inputs from the vestibulospinal and reticulospinal tracts, indicating a potential role in postural regulation. Muscles also contain many other varieties of receptors, such as those associated with lightly myelinated or unmyelinated sensory nerve fibers. These group III and group IV receptors code mechanical signals as well as pain and metabolic feedback. Pathways from these receptors are less well understood than those from muscle spindles and tendon organs, but evidence exists that feedback from these receptors does impinge on spinal circuits and can influence motor unit firing.

Cutaneous Receptors

The other major sources of feedback for the reflex circuits associated with skeletal muscles are the skin and connective tissue elements associated with joints. Feedback from the skin is mediated through interneuronal circuits to influence muscular activity through the hierarchical levels described above. These pathways may exhibit local sign, in which muscular responses tend to diminish the effects of a noxious stimulus on the same patch of skin. For example, when the receptive field of the feline caudal cutaneous sural nerve (lateral side of ankle) is stimulated, muscles are activated which plantarflex and abduct the ankle. These actions tend to remove the limb from the stimulus. Cutaneous pathways also have more global actions on the animal, including flexion–withdrawal responses and adjustments in muscular tone of the four limbs. It has been speculated that cutaneous information from the paws in cats may provide directional information to the circuits generating responses to postural perturbations. The functions of feedback from receptors in joints and connective tissue elements such as the anterior cruciate ligament are not well understood, but the associated pathways do excite or inhibit gamma motor neurons and therefore influence proprioceptive feedback from muscle spindle receptors.

Autonomic Reflex Circuits

The spinal cord also contains reflex circuits arising from receptors located in the bladder, bowel, and sexual organs. Pathways from stretch receptors in

the bladder wall project to micturition control centers in the brain stem. Spinal mechanisms promote storage of urine (contraction of sphincter and relaxation of the detrusor muscle) through sympathetic and somatic pathways, while brain stem mechanisms that inhibit sympathetic output and enhance parasympathetic output are required for reflex emptying of the bladder.

Reflexes Are Embedded in Neuromechanical Circuits

Mechanical Component

The mechanical state of the musculoskeletal system will determine the constellation of sensory feedback that arrives in the spinal cord. The musculoskeletal system consists of linked segments that are spanned by muscles and connective tissue elements. Some muscles span a single joint while others span more than one joint. Many joints rotate about more than one axis, and some muscles contribute to more than one axis of rotation. Muscular contractions cause mechanical changes in the musculoskeletal system by exerting forces that are transformed into torques about the spanned joints. These forces are also communicated by way of physical forces to other joints and axes of rotation and to other limbs and body segments. These physical forces consist of inertial coupling and the viscoelastic properties of connective tissue and muscle. Muscles exert forces through tendons of insertion and also through fascia that provides mechanical coupling across several joints. The distribution and extent of these mechanical effects depend on the architecture of the musculoskeletal system as well as the intensive and extensive properties of the tissues involved. These properties are in turn influenced by the configuration of the skeleton and the distribution of muscular activity. According to this mechanical description, the musculoskeletal system can be seen as a circuit based on the variables length and force. This circuit determines how mechanical variables are transformed into the neural representations in the spinal cord.

Mechanical perturbations applied to the body are resisted by mechanisms that are intrinsic to the musculoskeletal system in addition to reflex mechanisms. The intrinsic responses arise from the viscoelastic properties of muscles and connective tissue elements, as well as inertial properties of the system. These intrinsic properties provide a fundamental basis for stability and coordination on which the actions of reflex circuits are superimposed. These intrinsic properties provide reactions to mechanical disturbances that are not subject to the time delays of neural pathways.

Neuromechanical circuits are organized in a distributed fashion rather than a modular plan by virtue of the structure of both neuronal and musculoskeletal components, as illustrated by the schema shown in Figure 2. The stretch reflex is often depicted as a single loop in which muscular force is transformed into length changes through the load supported by the muscle. As stated above, however, muscular action results in mechanical changes throughout the musculoskeletal system by way of mechanical coupling. For example, forces exerted by biarticular muscles will exert torques about both spanned joints and cause length changes in a number of different muscles. The load that transforms muscular force into length changes is actually an extensive, distributed system. The length changes occurring in the muscle are due to mechanical interactions throughout the musculoskeletal system and not

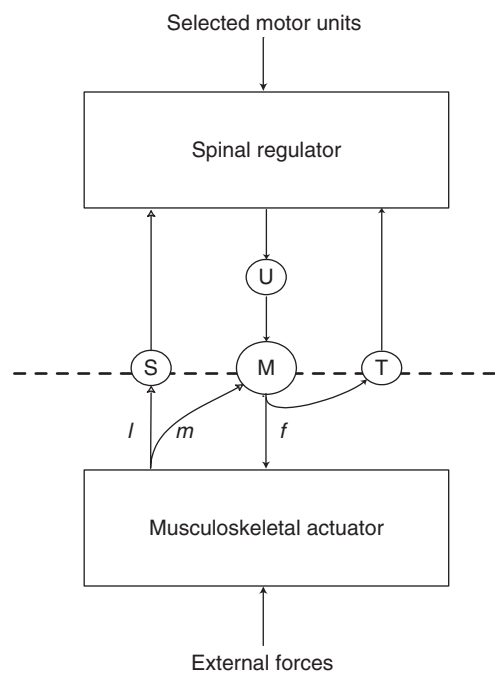

Figure 2 The interface between the musculoskeletal actuator and regulator consists of muscle units and sensory receptors. Excitation–contraction mechanisms of the muscle unit (M) convert neural information from spinal motor neurons (U) into mechanical signals, and receptors do the reverse. Muscle spindle receptors (S) are influenced by length changes (*l*) resulting from dynamics of the musculoskeletal system. Muscle spindles project to the regulator by way of monosynaptic and polysynaptic connections to motor neurons. Muscle length also provides direct mechanical feedback to the muscle unit (*m*). Golgi tendon organs (T) respond primarily to active contractile forces (*f*) in the muscle that contains them. These receptors project to interneurons in the spinal regulator. During locomotion, excitatory projections to the parent motor nucleus are opened. Under a wide variety of conditions, tendon organs project to muscles that exert torques about other joints or axes of rotation than the parent muscle by inhibitory connections.

merely the forces exerted by the muscle. Although the feedback from the spindle receptor may exert major effects on the motor neurons of the parent muscle, the signals encoded by the spindle receptors in the muscle are not private to the muscle. Furthermore, the preferential sensitivity of muscle spindle receptors for active lengthening indicates that they respond more vigorously to external perturbations or stretches due to contractions of other muscles than the shortening contractions of the parent muscle.

Neural Component

At the level of the spinal cord, the feedback from muscle receptors is distributed not only to the parent muscle but to others as well. In the cat, muscle spindle receptors project mainly to the parent muscle, but muscle spindles from the vastus intermedius muscle project monosynaptically to the soleus muscle as well. Therefore, knee communicates to ankle not only mechanically through biarticular muscles but also through the neuronal pathways of length feedback. On the output side, activity in one motor nucleus communicates with other motor nuclei through the distribution of Renshaw inhibition. Renshaw cells are excited by recurrent collaterals of motor axons and project back not only to the parent motor neurons but to other motor neurons as well. Signals from muscle spindle receptors also impinge on the selection and oscillator components of the spinal networks and therefore are distributed widely through the musculoskeletal system. At all hierarchical levels, reflex circuits become distributed by virtue of both mechanical and neural components.

Muscle spindle receptors are stimulated by length changes brought about by musculoskeletal dynamics, and their responses are not simply related to the contractile output of the muscles containing them. In contrast, the firing rates of Golgi tendon organs follow the active contractile forces exerted by the parent muscle, whether in active lengthening or shortening. As noted above, feedback from Golgi tendon organs is distributed widely to motor neuron pools in the limb and presumably the trunk. The cross-joint inhibitory actions constitute neural linkages that parallel the direct mechanical linkages of multiarticular muscles. Force feedback is itself force-dependent, so that the extent of neural coupling across joints and axes of rotation increases at higher forces, reducing the degrees of freedom of the corresponding musculoskeletal elements.

The myotatic unit can be used to illustrate how elements in the musculoskeletal and nervous systems combine to form a neuromechanical circuit. Although the myotatic unit is not a universal building block of musculoskeletal control, it can be excerpted experimentally in the distal limb musculature. The neuronal component consists of excitatory feedback from muscle spindle receptors to the parent motor neurons (the stretch reflex) and inhibitory feedback to antagonists. The mechanical component consists of the reciprocal mechanical linkage of the two muscles. An external perturbation that causes movement of the joint so as to stretch the antagonist muscle is immediately met by the intrinsic resistance of both muscles. The agonist resists by increasing its force, and the antagonist resists by decreasing its force. The further excitation of the agonist muscle through the stretch reflex and inhibition of the antagonist muscle adds to the resistance of this neuromechanical apparatus to the perturbation.

Modularity

Two ideas proposed recently concern ways in which neuromechanical circuits might be modular. The first is the concept of motor primitives. According to this hypothesis, the sensorimotor apparatus in the spinal cord is subdivided into a small number of modules. Each module consists of a subset of interneuronal pathways, sensory feedback, and musculoskeletal elements. Movements are made by selection of different combinations of these modules. The modules can be recruited by centrally derived commands or by sensory feedback. In the scratch reflex, appropriate modules would be selected to achieve the local sign of this triggered behavior. The second hypothesis is that the motor apparatus consists of synergies of muscles or motor units that are brought into activity in groups. These synergies are presumably specified by the selection module as a result of either feedback or central commands. These synergies are flexible in that a given motor unit or muscle can be a member of more than one synergy. Activation of a synergy would lead to the recruitment of a particular population of Golgi tendon organs. The mechanisms by which modularity can be achieved anatomically or physiologically are unknown.

Sensory Feedback Regulates the Mechanical Properties of the Musculoskeletal System and Promotes Coordination and Stability

Sensory feedback links the mechanical and neuronal components of the circuit and, in parallel with the intrinsic mechanical properties of the musculoskeletal system, determines the mechanical properties of the complete system. As described in the section titled 'Neural component,' length feedback is organized quite differently from force feedback. Mechanical changes in the musculoskeletal system evoke a constellation of signals from muscle spindles that depend

on the length changes of the muscles. The feedback would be particularly strong in muscles undergoing active lengthening as a result of the perturbation or commanded muscular contractions. Due to the matching of intrafusal and extrafusal mechanics, the monosynaptic autogenic connections will compensate for certain muscle nonlinearities by recruiting new motor units. This integration of intrinsic and reflexive mechanisms and additional contributions of reciprocal inhibition will set the springlike mechnical characteristics of the joint. In this case, the springlike characteristics differ from those of an ideal spring in that stiffness increases with force and length. This dependence of stiffness on force has an important function in mediating the dependence of joint stiffness on the level of cocontraction, according to the lambda model of the equilibrium point hypothesis. Mechanical coupling between adjacent joints would be promoted by the stiffness of reflexive biarticular muscles.

By virtue of more divergent projections in the spinal cord, inhibitory force signals from Golgi tendon organs would interact with length signals and intrinsic properties in two ways. First, the balance of force and length feedback would determine the stiffness of the limb measured at the endpoint, as originally suggested by James Houk for single muscles. Second, force feedback would promote interjoint coordination and stability by providing neural coupling across joints and axes of rotation. Should forces in the limb increase, the gain of force feedback and therefore coupling would increase as well. Given the redundancies inherent in multisegmented limbs, these sensory pathways would promote stability across the joints of the limb in parallel with the mechanical coupling provided by multiarticular muscles.

During locomotion, a pathway of excitatory force feedback is opened. This positive force feedback appears to be primarily autogenic and has the effect of increasing stiffness and force in the muscles of propulsion. This excitatory autogenic feedback is particularly strong in the medial gastrocnemius muscle of the cat, a muscle that contributes in important ways to propulsion during walking up sloped surfaces. Positive force feedback is important in maintaining force during active lengthening in this muscle when much of the stretch is absorbed in the tendon. Under these conditions, muscle spindle receptors would provide modest feedback because the muscle fibers would be nearly isometrically loaded.

This regulatory network would be capable of mediating basic postural responses in the whole limb. Muscles would be activated according to pulling directions by pathways arising from muscle spindle receptors. Communication through propriospinal pathways and the local selector could provide the appropriate integration of responses of the supporting limbs. More complex postural responses and integration with vestibulospinal and proprioceptive feedback from the neck would involve participation of brain stem areas as well.

Reflex Circuits Are Subject to Peripheral and Central Modulation

Neuromechanical circuits are subject to both short- and long-term modulation. Short-term effects allow adaptation to changing behavioral and environmental contexts, and long-term changes occur with growth, learning, and disease. Short-term modulation has been observed in both neural and mechanical components of the motor apparatus. The dependence of the responses of muscle spindle receptors on movement history is matched to the dependence of the mechanical properties of muscle on movement history to preserve springlike behavior. Mechanical responses of muscle to increases in length are mediated largely by length feedback under quasi-static conditions. During ongoing movements, the responses include a larger contribution from intrinsic mechanics. This feature contributes to stability of the motor system. A mechanical perturbation is initially resisted by the elastic properties provided by reflex recruitment. Any ensuing oscillations are effectively damped out by the predominance of viscous intrinsic properties of muscle. This automatic change in gain is mediated by the interaction of neural and mechanical components. A second example of modulation that results from the nonlinear properties of the muscle spindle concerns the termination of a voluntary movement. Since the velocity dependence of muscle spindle receptors contributes to damping, damping is greater toward the termination of movement as the limb slows during its approach to its target position. The increased damping as well as the attraction of the final equilibrium point helps maintain the limb on target.

In the central nervous system, mechanisms for modulation are manifold. For example, the sensitivity of muscle spindle receptors can be modulated by gamma motor neuronal activity, while modulation of force feedback from Golgi tendon organs occurs centrally in interneuronal pathways. Presynaptic inhibition of sensory afferents can also regulate the gain of incoming feedback. The gain of primary spindle receptors is apparently reduced at the stance to swing transition of the step cycle to allow rapid clearance of any obstacles by the foot. Automatic postural responses can be downregulated within a few trials if inappropriate

for stability, perhaps mediated by presynaptic inhibition. Descending pathways from the brain stem can also promote bistable firing of motor neurons through serotonergic receptors. Bistable firing may help maintain upright posture for extended periods. Finally, modifications of motor behavior evoked by stimulation of the dorsum of the foot in walking cats are altered in a phase-dependent manner. During the swing phase, stimulation results in withdrawal of the foot and ankle extension to clear an obstacle. During the stance phase, the same stimulus can increase extension or increase flexion during the subsequent swing phase. Long-term changes in neuromechanical circuits also occur. For example, the gain of the stretch reflex can increase or decrease over a period of days or weeks by appropriate operant conditioning. The cellular mechanisms of these changes are complex but appear to involve both intrinsic properties of motor neurons and changes in synaptic strength.

See also: Central Pattern Generators: Sensory Feedback; Operant Conditioning of Reflexes; Pattern Generation; Reticulospinal System; Spinal Motor Neurons: Properties.

Further Reading

Bosco G and Poppele RE (2001) Proprioception from a spinocerebellar perspective. *Physiological Reviews* 81(2): 539–568.

Duysens J, Clarac F, and Cruse H (2000) Load-regulating mechanisms in gait and posture: Comparative aspects. *Physiological Reviews* 80(1): 83–133.

Feldman AG and Orlovsky GN (1972) The influence of different descending systems on the tonic stretch reflex in the cat. *Experimental Neurology* 37: 481–494.

Giszter SF, Mussa-Ivaldi FA, and Bizzi E (1993) Convergent force fields organized in the frog's spinal cord. *Journal of Neuroscience* 13(2): 467–491.

Hultborn H (1999) Plateau potentials and their role in regulating motoneuronal firing. *Progress in Brain Research* 123: 39–48.

Jami L (1992) Golgi tendon organs in mammalian skeletal muscle: Functional properties and central actions. *Physiological Reviews* 72: 623–666.

Jankowska E (1992) Interneuronal relay in spinal pathways from proprioceptors. *Progress in Neurobiology* 38: 335–378.

Lafreniere-Roula M and McCrea DA (2005) Deletions of rhythmic motoneuron activity during fictive locomotion and scratch provide clues to the organization of the mammalian central pattern generator. *Journal of Neurophysiology* 94(2): 1120–1132.

Nichols TR (1994) A biomechanical perspective on spinal mechanisms of coordinated muscular action: An architecture principle. *Acta Anatomica* 151(1): 1–13.

Nichols TR and Cope TC (2004) Cross-bridge mechanisms underlying the history-dependent properties of muscle spindles and stretch reflexes. *Canadian Journal of Physiology and Pharmacology* 82(8–9): 569–576.

Pearson KG (2000) Neural adaptation in the generation of rhythmic behavior. *Annual Review of Physiology* 62: 723–753.

Prochazka A (1996) Proprioceptive feedback and movement regulation. In: Rowell LB and Shepherd JT (eds.) *Handbook of Physiology. Section 12: Exercise: Regulation and Integration of Multiple Systems*, pp. 89–127. New York: Oxford.

Quevedo J, Stecina K, and McCrea DA (2005) Intracellular analysis of reflex pathways underlying the stumbling corrective reaction during fictive locomotion in the cat. *Journal of Neurophysiology* 94(3): 2053–2062.

Sherrington CS (1947) *The Integrative Action of the Nervous System*. New Haven, CT: Yale University Press.

Sinkjaer T (1997) Muscle, reflex and central components in the control of the ankle joint in healthy and spastic man. *Acta Neurologica Scandinavica* 96(supplement 170): 1–28.

Taylor A (2002) Give proprioceptors a chance. In: Gandevia S, Proske U, and Stuart DG (eds.) *Sensorimotor Control of Movement and Posture*, vol. 508, pp. 327–334. New York: Kluwer Academic/Plenum.

Regulation of Cell Volume in Neural Cells

A A Mongin, Albany Medical College, Albany, NY, USA
H K Kimelberg, Albany Medical College and Ordway Research Institute, Albany, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Osmosis, Osmotic Pressure, and Steady-State Cell Volume Homeostasis

Animal cells are surrounded by semipermeable plasma membranes. Water moves across this membrane freely, whereas the majority of solutes require special transport systems to cross the membrane barrier. Any change in the concentration of solutes on one side of the membrane causes water movement along the new concentration gradient by a process termed osmosis. Such water movement can be prevented by applying to the membrane a hydrostatic force, called osmotic pressure, which may be calculated using the van't Hoff equation:

$$\pi = RT\Delta C$$

where π is the osmotic pressure, R is the gas constant, T is the absolute temperature, and ΔC is the difference in solute concentrations across the semipermeable membrane.

Since animal cells do not possess a rigid cell wall, their membranes cannot maintain osmotic gradients, and therefore any accumulation or loss of solutes is immediately followed by water flow so that the osmotic gradient dissipates. Since animal cells contain the large proportion of impermeant macromolecules such as proteins, which are predominantly anionic, that would lead to an unstable situation in which the cell would inexorably swell until the impermeant macromolecules were infinitely diluted so that the cell no longer effectively contained them. Long before this, of course, the cell would no longer be able to function as a cell. This process of cell swelling due to the presence of impermeable macromolecules in the cytoplasm is termed Donnan swelling after the Harvard physiologist who first recognized its importance.

To maintain their volume and viability, animal cells evolved a system that neutralizes their tendency to swell. The first part of this process is the accumulation of K^+ inside the cell. This is achieved by the ubiquitous Na^+, K^+ pump, which directly uses ATP to achieve a high $[K^+]$ in cells ~130 mM and a low $[Na^+]$ in cells 10–20 mM. For a membrane containing many more K^+ than Na^+ channels, the membrane potential is almost entirely a K^+ diffusion potential of ~ -90 mV with the normal low extracellular K^+ of 3 or 4 mM. If the only transport route for Cl^- is channels, intracellular chloride will be driven out of the cell by the internal negative charge and will be ~4 mM, as given by the Nernst potential with $[Cl^-]$ outside being ~130 mM plus a high $[Na^+]_o$ of an equivalent concentration. The low intracellular $[Cl^-]$ compensates for the presence of organic anions in the cytoplasm for electrical neutrality. With a low density or activity of Na^+ channels making Na^+ permeability much less than that of K^+, the whole system is in a steady state, called the Gibbs–Donnan equilibrium, which can be maintained with a minimum expenditure of energy.

Water Movement across Plasma Membranes

The processes of ion transport and osmotic regulation require that the cell membranes are freely permeable to water relative to the fastest ion movements so that the movement of water does not become rate limiting. Despite the basic characteristics of cell membranes being those of a lipid bilayer or, more strictly, a phospholipid bilayer, they are quite permeant to water. The upper limit of the permeability coefficient (P_f) for water of phospholipid bilayers is 5×10^{-3} cm s^{-1}, 10^8 higher than for Cl^- and approximately 10^{10} higher than that for K^+ or Na^+.

Nonetheless, a special class of membrane proteins accelerating water movement has been identified in many cell membranes. These proteins are termed aquaporins and are known to consist of at least 11 different members termed AQP0–10. AQP1, – 4, and – 9 are present in the brain, with AQP4 and – 9 specifically found in astrocytes and AQP1 present in the epithelial cells of the choroid plexi.

For cultured astrocytes, genetic deletion of AQP4 reduced P_f from 0.05 to 0.007 cm s^{-1} for cells prepared from wild-type and AQP4$^{-/-}$ mice, respectively. The latter value is very similar to the P_f for lipid bilayers. AQPs may be needed for water transport to keep pace with the transport of ions, which is very rapid through the various channels and transporters. Alternatively or in addition, some biological membranes can be relatively water impermeable, and the water permeability can then be regulated by AQPs. This is important in water-controlling organs such as the kidney. However, the relative significance of aquaporins in the brain is presently less understood.

Physiological and Pathological Perturbations of Cell Volume

A large variety of physiological reactions and pathological states can disrupt the Gibbs–Donnan

equilibrium and cause changes in cell volume. Changes in extracellular osmolarity represent the simplest reason for alterations in cell volume. However, in mammals only blood cells passing through kidney capillaries and a few types of epithelial cells are exposed to substantial osmotic gradients. Osmolarity of the extracellular environment throughout the body is tightly regulated, except under certain pathological conditions. For example, hyponatremia – a decrease in plasma sodium concentration – is typically caused by hormonal disorders leading to alterations in kidney function. Causes of hyponatremia include overproduction of antidiuretic hormone (ADH), glucocorticoid deficiency, and renal failure. Conversely, acute dehydration, osmotic diuresis, and decreased ADH production may initiate hypernatremia – an increase in plasma sodium concentration.

A far more common reason for cell volume alterations is changes in intracellular osmolarity. Net Na^+ uptake during activation of neuronal ionotropic excitatory amino acid receptors, such as AMPA receptors, can cause neuronal volume increase. Na^+-dependent uptake of glutamate and other neurotransmitters has been linked to cell swelling seen mainly in astrocytes. Sustained neuronal activity or pathological neuronal depolarization leads to a buildup of extracellular K^+ levels, which promote slow cell swelling because of the plasma membrane depolarization followed by $(K^+ + Cl^-)$ influx according to the new Gibbs–Donnan equilibrium. *In situ*, this latter type of swelling is also predominantly found in astrocytes.

Inhibition of energy metabolism, as seen in a number of neuropathologies, generally causes cell swelling resulting from a loss of Na^+, K^+-ATPase function, dissipation of K^+ gradient, and membrane depolarization with concomitant accumulation of Na^+ and Cl^- and cell swelling.

An opposite process of solute efflux and cell shrinkage may result from the stimulation of K^+ channels and/or Cl^- channels. Such shrinkage is thought to occur at the initial stages of apoptosis in both neuronal and nonneuronal cells.

Methods of Cell Volume Measurements

A number of methods allow for monitoring cell volume changes in different experimental systems, ranging from suspensions of roughly spherical cells to cells in the intact tissue. There is essentially an increase in complexity and decrease in precision as one advances from the dispersed cell suspensions through substrate attached cell cultures to different cells mixed in a single tissue. We cannot describe these methods in detail but will restrict our discussion to briefly mentioning the main ones.

The isotope-based methods, such as the formerly extensively used ^3H 3-*O*-methyl glucose (3-O-MG) measurements, calculate intracellular water space occupied by the isotope-labeled compound. The volume of distribution of 3-O-MG can be calculated but only based on the assumption that the concentrations of 3-O-MG are the same inside and outside the cell. In a glucose-containing medium, the 3-O-MG will trace the glucose concentrations and it requires that the transporter is of the equilibratory type and there is no concentration gradient.

Spectroscopic methods can provide a dynamic readout of the time course of swelling. These methods use changes in the concentration and therefore the signal of a fluorescence compound that remains in the cells and does not leak out with the volume changes. However, the measuring device, a fluorescence microscope, will have to record from a partial, constant volume of the cell so that the dilution of the compound is registered. The measurement is more reliable if a confocal microscope is used, in which a fixed volume can be measured in the horizontal focal plane. With rapid sampling of the cells from top to bottom and the appropriate software, entire cells can be reconstructed and there is the possibility of measuring total cell volume.

Measuring the volume of total tissue is straightforward; it is done by measuring its wet weight and referencing this to the weight after the sample has been thoroughly dried. This method has been used since early studies on tissue volume and water and ion content in animal tissues.

More dynamic but indirect methods for measuring the time course of cell volume changes involve measuring the extracellular impedance electrically using electrodes, as pioneered by van Harreveld. Since cells do not conduct current compared to the extracellular space (ECS), this can be applied to tissue slices or the intact organ *in vivo*. The volume (V) of the brain is constant:

$$V_{\text{blood}} + V_{\text{ESC+CSF}} + V_{\text{brain intracell}} = \text{constant}$$

where CSF is cerebrospinal fluid.

Using alternating current to avoid polarization effects and neuronal excitation, and Maxwell's equations, one can measure extracellular space and dynamic changes in cell volume based on the fact that membrane-delimited cells are nonconductors of current compared to the ESC.

Another method for measuring dynamic changes in the ECS has been pioneered by Nicholson and colleagues. Ion-specific electrodes are used to measure the changes in activities of a membrane-impermanent species which is restricted to the ECS. Thus, as cells swell, the species' electrochemical activity will increase, and it will decrease when cells shrink.

The light scattering method, in which cells swell when there is a decrease in light scattering, can be used for cell suspensions or tissue slices. This is also referred to as a change in the intrinsic optical signal or IOS.

A newer, noninvasive method to measure cell volume changes *in vivo*, which is widely used clinically, is the diffusion-weighted spin echo sequence with magnetic resonance imaging. This method allows one to calculate the apparent diffusion coefficient of water (ADC). A decline in this parameter means the bulk diffusion of water is more hindered. This has been interpreted in brain tissue as principally due to the movement of water from the extracellular space to intracellular space due to cell swelling. In animal models of stroke and epilepsy, the decrease in ADC correlated with increased swelling of astrocytes measured by electron microscopy.

Electron microscopy, after rapid perfusion fixation to avoid postmortem fixation artifacts, is in fact the only truly reliable method for studying the swelling of specific cells *in situ*. However, generally the cross-sectional areas of the cells in the thin sections are measured. It is time consuming, especially if three-dimensional reconstructions are attempted, and it is nondynamic.

Ionic Mechanisms Responsible for Regulatory Volume Decrease in Swollen Cells

In response to swelling, the overwhelming majority of animal cells, including neural cells, activate K^+ and Cl^- efflux through independent K^+ and anion channels. These channels are functionally coupled via the plasma membrane potential. Efflux of K^+ immediately leads to membrane hyperpolarization, increasing the driving force for Cl^- efflux. Since intracellular Cl^- concentration is typically low, Cl^- efflux is supplemented by the release of bicarbonate ions and small organic anions through the same low-selectivity anion channels.

Among the potassium channels known to be activated or modulated by cell swelling are Kv1.3, Kv1.5, mini-K, and TASK-2 (KCNK5). Although these channels are expressed in the central nervous system (CNS), their relative contribution to volume regulation in neuronal and glial cells is not known. Maxi-K channels, which are opened in response to cell swelling in many types of epithelial cells, are activated indirectly via increases in intracellular $[Ca^{2+}]$.

The molecular nature of the volume-regulated Cl^-/anion channels (VRACs), which are ubiquitous in invertebrate and vertebrate cells, is unclear. Several previously suggested candidates for the role of VRAC, including pICln, CFTR, ClC-2, phospholemman, and porin-like voltage-dependent anion channels, have been ruled out. These channels either show different biophysical properties compared to the endogenous VRAC or their respective gene knockouts do not affect endogenous VRAC currents in swollen cells. The ClC-3 gene product is still discussed as a VRAC candidate; however, ClC-3 gene knockouts did not affect the VRAC current and therefore have placed this concept in doubt.

The electroneutral K^+, Cl^- cotransporter is an alternative, nonchannel system which contributes to regulatory volume decrease (RVD) in red blood cells and some types of epithelial cells which have a relatively high intracellular Cl^- content. To date, four members of this family, KCC-1 through -4, have been cloned. Although KCC-1 is most common, KCC-2 is highly expressed in the brain, where it plays a major role in the regulation of intracellular $[Cl^-]$ in neuronal cells. The relative contribution of KCC transporters to cell volume regulation in neural cells is not known.

Ionic Mechanisms Responsible for Regulatory Volume Increase in Shrunken Cells

In the majority of cell types, regulatory volume increase (RVI) is mediated by the electroneutral Na^+, K^+, $2Cl^-$ cotransporter and/or Na^+/H^+ exchanger. The NKCC1 isoform of the Na^+, K^+, $2Cl^-$ cotransporter is ubiquitously expressed. This transporter is activated by cell shrinkage, and it contributes to compensatory accumulation of Na^+, K^+, and Cl^- and subsequent cell volume recovery. In some cell types, however, hyperosmotic activation of NKCC1 is not accompanied by RVI because of the absence of an inwardly directed net gradient of Na^+, K^+, and Cl^-. Activation of the Na^+, K^+, $2Cl^-$ cotransporter by isoosmotic cell shrinkage (e.g., due to a metabolic decrease of osmolytes in the cytoplasm) typically leads to effective RVI due to the preserved inwardly directed electrochemical gradient for the transported ions.

Volume-dependent activation of the Na^+/H^+ exchanger mediates accumulation of osmotically active Na^+ in exchange for osmotically inactive H^+ (because the latter does not exist in a free form but as a hydroxonium ion). Cytoplasmic acidification then stimulates the Cl^-/HCO_3^- exchanger, which is itself insensitive to cell volume changes. As a consequence of the cooperative work of these two transporters, cells accumulate Na^+ and Cl^- and extrude H^+ and HCO_3^-. Bicarbonate is converted to CO_2 and diffuses back to the cell, replenishing pools of H^+ and HCO_3^-. Six mammalian isoforms of the Na^+/H^+ exchanger (NHE-1 through -6) have been cloned. NHE-1 is highly and ubiquitously expressed in all tissues including the

CNS, where it contributes to cell volume regulation and cytosolic pH maintenance. Other isoforms are differentially distributed among various cell types. Interestingly, in striking contrast to all other Na^+/H^+ exchangers, the epithelial exchanger NHE-3 is inhibited by cell shrinkage.

Gene Transcription Changes in Response to Cell Volume Alterations

RVD and RVI reactions serve as emergency responses protecting cells against acute changes in volume. However, these regulatory processes cause severe disturbances in cellular ion composition, which need to be compensated for in order to return cells to their normal functioning capacities. This is typically achieved through long-term adjustments in intracellular levels of organic osmolytes resulting from the upregulation or downregulation of genes involved in osmolyte synthesis and/or osmolyte uptake. Key organic osmolytes involved in chronic adaptation to osmotic stress include *myo*-inositol, sorbitol, betaine, glycerophosphorylcholine, and taurine.

The osmotic regulation of gene expression in shrunken cells is a complex multistage process. It is initiated by volume-dependent activation of stress-related protein kinases of the mitogen-activated protein kinase (MAPK) family, which in turn triggers the expression of immediate response genes (c-*fos*, c-*jun*, CREB, Egr-1 in epithelial cells, and others). This is followed by changes in the expression of osmolyte transporters and the enzymes involved in osmolyte biosynthesis, which typically peak between 6 and 24 h after the onset of hyperosmolarity.

Osmolyte transporters upregulated by hypertonicity include the Na/*myo*-inositol transporter (SMIT), Na/Cl/betaine transporter (BGT-1), and taurine transporter. An increase in osmolyte biosynthesis is mediated by the upregulation of aldose reductase (converts glucose to impermeable osmolyte sorbitol), choline dehydrogenase (contributes to betaine production), and cysteine deoxygenase (one of the enzymes of taurine biosynthesis). Conversely, chronic hypoosmolarity downregulates expression of organic osmolyte transporters and enzymes involved in osmolyte biosynthesis, causing a long-lasting decrease in total cytoplasmic osmolyte content.

Volume-dependent regulation of gene expression has been studied in detail in renal tissue, but it also occurs in the brain.

The Cell Volume Sensor

Despite our advanced knowledge on volume-dependent ion transporters involved in RVD and RVI, the central question remains unresolved: how do cells sense changes in their volume? To explain the phenomenon of cell volume autoregulation, the concept of a volume sensor has been introduced. The volume sensor is a hypothetical molecule or intracellular structure which undergoes conformational transitions in swollen or shrunken cells. Such structural transition in volume sensor turns on signaling cascades coupled to the activation of ion transporters mediating RVD and RVI.

Cells adapted over a long time to hyperosmotic or hypoosmotic media return to their initial 'set point' volume and continue to respond to acute changes in osmolarity and cell volume. Therefore, cells seem to sense a signal or signals triggered by changes in cell volume rather than changes in cytoplasmic osmolarity.

The most commonly discussed candidates for the role of volume sensor/volume-sensing processes are mechanosensitive ion channels, the cortical cytoskeleton, and changes in intracellular concentration of ions (ionic strength) or macromolecules ('macromolecular crowding').

Mechanosensitive nonselective cation channels are directly activated in response to membrane stretching or compression and allow for membrane depolarization and Ca^{2+} influx. Elevation in intracellular $[Ca^{2+}]$ may then serve as a secondary volume signal in some but not all cell types. Two of the mammalian mechano-/volume-sensitive cation channels, swelling-activated TRPV4 (OTRPC4, VR-OAC, TRP-12) and shrinkage-activated splice variant of TRPV1, have been cloned. However, since mechanosensitive cation channels are not ubiquitously expressed, it seems unlikely that they play a universal role of cell volume regulation.

The cortical cytoskeleton supports the plasma membrane and protein compartmentalization within the plasma membrane. It is composed of spectrin heterodimers, actin bundles, and proteins involved in spectrin–actin interactions, and this network is anchored to integral proteins and acid phospholipids. Any changes in cell shape or volume are associated with reorganization of cortical cytoskeleton and trigger extensive changes in signaling molecules associated with this structure. In rat erythrocytes, disruption of the spectrin network completely abolishes volume-dependent regulation of the Na^+, K^+, Cl^-, and K^+, Cl^- cotransporters. In neural cells, various agents, which depolymerize actin filaments, modulate but do not abolish cell volume regulation. Thus, there is inconclusive evidence for the contribution of cortical cytoskeleton in cell volume regulation in neural cells.

Changes in the total activity of electrolytes (intracellular ionic strength) have also been proposed to play a key role in volume signal transduction. Several studies have demonstrated that a decrease in

cytoplasmic ionic strength serves as a signal for the indirect (via intracellular signaling) activation or modulation of volume-sensitive anion channels. In a mathematical model developed by Voets and collaborators, changes in ionic strength rather than cell volume accurately predict the time course of anion channel activation.

Variations in cell volume directly affect the concentrations of all cellular components. The total concentration of macromolecules in the cytoplasm is so high that it behaves as a nonideal solution that limits the diffusion of small and especially large molecules. The resulting nonideal interaction of macromolecules, known as macromolecular crowding, appears to crucially influence the activity of many enzymes. For example, high concentrations of macromolecules decrease the activity of lactate dehydrogenase and hyaluronate lyase but potently stimulate DNA polymerase I and DNA ligase. Thus, changes in macromolecular crowding may principally serve as a primary volume signal and activate intracellular signaling proteins. This hypothesis is supported by strong experimental evidence in erythrocytes, in which manipulating the concentration of cytoplasmic proteins and/or inert macromolecules altered the volume sensitivity of the K^+, Cl^- cotransporter and Na^+/H^+ exchanger. However, the role of macromolecular crowding in nonerythroid cells remains untested.

Mechanisms of Volume Signal Transduction

Although the precise nature of the volume sensor in unknown, it is well established that cell volume alterations activate a variety of intracellular signaling cascades. Extensive experimental evidence suggests that some of these cascades contribute to volume signal transduction, whereas activation of others seem to be epiphenomena resulting from cross-talk between intracellular signaling systems. Chemical transduction of the cell volume signal appears to be cell-type specific and varies depending on the type of volume-activated transporters.

In most animal cells studied, cell swelling leads to elevations in intracellular Ca^{2+} concentrations due to activation of membrane permeability pathways and/or release of Ca^{2+} from intracellular stores. Cell shrinkage is generally not associated with changes in intracellular Ca^{2+} concentrations. The omission of extracellular Ca^{2+} or buffering of intracellular Ca^{2+} with membrane-permeable Ca^{2+} chelators inhibit RVD in some but not all cell types. A detailed literature review by Pasantes-Morales and Morales-Mulia suggests that the high Ca^{2+} sensitivity of RVD is observed predominantly in epithelial cells and is mediated by Ca^{2+}-sensitive maxi-K^+ channels. In other cells, including neural cells, an increase in Ca^{2+} is not obligatory for swelling-induced activation of ion transporters.

In addition to intracellular Ca^{2+}, cell swelling may activate ion transport pathways via receptor and nonreceptor tyrosine kinases. Genetic deletion of the $p56^{LCK}$ kinase (a member of the *src* family of nonreceptor tyrosine kinases) blocks both anion channel activation and cell volume regulation in Jurkat cells. When injected, $p56^{LCK}$ stimulates anion channels in nonswollen cells. Broad-spectrum tyrosine kinase inhibitors and inhibitors of the MAPK cascade block activation of volume-dependent K^+ and anion channels in many cell types. In glial cells, tyrosine phosphorylation is thought to modulate volume sensitivity of ion channels rather than serve as a primary transducer of the volume signal.

Several other serine/threonine protein kinases and kinases of mixed specificity have also been implicated in cell volume transduction in a cell-type-specific manner. Kinases involved in volume regulation in swollen cells include protein kinase C (PKC), myosin light-chain kinase (MLCK), phosphatidylinositol 3-kinase (PI-3K), focal adhesion kinase $p125^{FAK}$, and a variety of kinases belonging to the cascade of MAPK, such as Jun-NH_2-terminal kinase (JNK), p38 PK, and extracellular signal-regulated kinases (ERK-1 and ERK-2).

Cell shrinkage-related signaling may involve Ca^{2+}/calmodulin-dependent kinase II (Ca/CM-PK); PKC; MLCK; tyrosine kinases Pyk2, Fgr, Syk, and Hck; and a number of the MAPK cascade members, including MKK4, MEK, JNK, p38 MAPK, ERK-1, ERK-2, and SAPK. Some of these protein kinases regulate the activity of ion transporters, whereas others likely contribute to shrinkage-induced regulation of gene expression.

Physiological and Pathological Significance of Cell Volume Regulation in the CNS

Under physiological conditions, swelling of cells is not observable in the CNS. However, in the empirical sciences, the absence of evidence is not evidence of absence because further work with different approaches or more sensitive measurements may reveal what had hitherto not been seen. It is clear that detection of small swelling of cells in brain tissue is technically very difficult. In addition, small changes in cell volume may be amplified by certain signaling events. Thus, ATP can markedly augment the release of amino acids from nonswollen or moderately swollen cultured astrocytes via a transport route that is robustly activated by much larger swelling under pathological conditions.

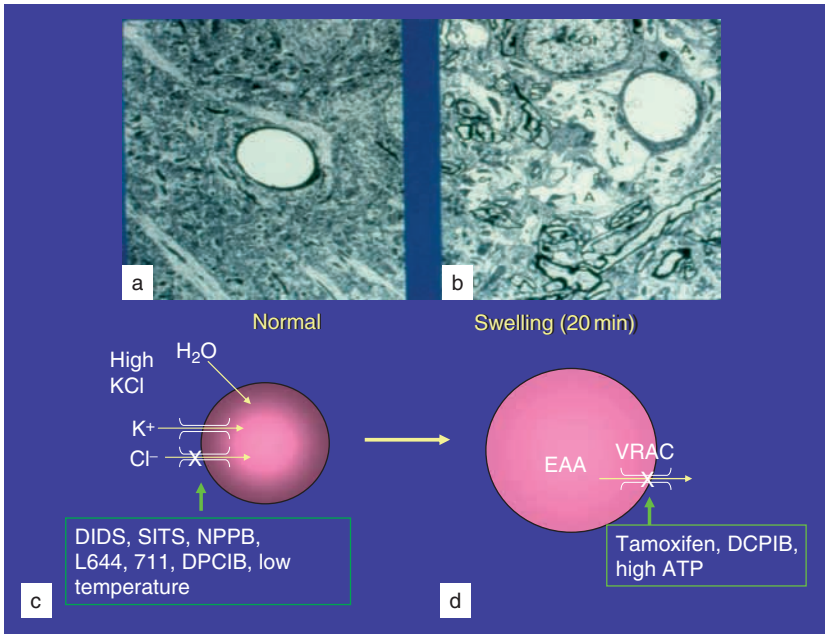

Figure 1 (a) Electron micrograph of a sham cat sensorimotor cortex. (b) Electron micrograph from the same region 20 min after an acceleration deceleration injury. Note the marked perivascular and neuropilar astrocytic swelling labeled A in figure. (c) A mechanism of swelling by Donnan forces due to raised $[K^+]_o$. This needs an open chloride channel, and it is proposed that the anion channel blockers indicated can block astrocytic swelling by blocking this channel. The swelling then opens volume-regulated anion channels (VRAC), which are permeable to excitatory amino acids (EAAs) and these are blocked by some of the anion channel blockers, as shown. L644,711, R(+) [(5,6 dichloro 9a-propyl 2,3,9,9a-tetrahydro 3-oxo-1*H* fluoren-7-yl)oxy] acetic acid; DCPIB, 4-[(2-butyl-6,7-dichloro-2-cyclopentyl-2,3-dihydro-1-oxo-1*H*-inden-5-yl)oxy] butanoic acid; SITS, 4-acetamido-4′-isothiocyanostilbene-2,2′-disulfonic acid; NPPB, 5-nitro-2-(3-phenylpropylpropylamino) benzoate; DIDS, 4,4′-diisothiocyanostilbene-2,2′-disulfonic acid.

Hussy and colleagues have shown that small changes in osmolarity can release taurine from specialized astroglial cells to regulate vasopressin release from neurons in the rat supraoptic nucleus by activating their glycine receptors.

Marked astrocytic swelling is seen in a number of pathological states, such as stroke, traumatic CNS injury, hepatic encephalopathy, and seizures. Such swelling, termed cellular edema, needs to be distinguished from total brain swelling, which primarily involves ECS swelling. However, both tissue swelling and cell swelling likely coexist and lead to life-threatening effects of raised intracranial pressure and shifts of brain tissue. Figure 1 shows the marked perivascular and neuropilar astrocytic swelling seen 20 min after experimental closed-head injury. Shown below the micrographs is one mechanism of swelling due to the documented rise of K^+ in the ECS which, with the opening of Cl^- channels, can lead to a Donnan-type swelling. Such marked swelling leads to activation of VRACs, which are permeable to excitatory amino acids, aspartate and glutamate. Thus, part of the excess glutamate and aspartate release in brain trauma and stroke, and generally accepted to be the initiating event for further damage, can come from swollen astrocytes. Also shown in Figure 1 are various compounds that can inhibit astrocytic swelling if given early enough and block volume-regulated anion channels. Some of these compounds (L644, 711 and tamoxifen) have been shown to be markedly effective in preventing death, disability, or brain infarct volume in animal models of ischemia and traumatic brain injury.

See also: Glial Ion Homeostasis: A Fluorescence Microscopy Approach; Neuron-Glia pH Regulation; Osmoregulation.

Further Reading

Amiry-Moghaddam M and Ottersen OP (2003) The molecular basis of water transport in the brain. *Nature Reviews Neuroscience* 4: 991–1001.

Burg MB, Kwon ED, and Kultz D (1996) Osmotic regulation of gene expression. *FASEB Journal* 10: 1598–1606.

Hussy N, Deleuze C, Desarmenien MG, and Moos FC (2000) Osmotic regulation of neuronal activity: A new role for taurine and glial cells in a hypothalamic neuroendocrine structure. *Progress in Neurobiology* 62: 113–134.

Kimelberg HK (1995) Current concepts of brain edema: Review of laboratory investigations. *Journal of Neurosurgery* 83: 1051–1059.

Kimelberg HK (2005) Astrocytic swelling in cerebral ischemia as a possible cause of injury and target for therapy. *Glia* 50: 389–397.

Lang F, Busch GL, Ritter M, et al. (1998) Functional significance of cell volume regulatory mechanisms. *Physiological Reviews* 78: 247–306.

Mongin AA and Kimelberg HK (2005) Astrocytic swelling in neuropathology. In: Kettenmann HO and Ransom BR (eds.) *Neuroglia*, pp. 550–562. New York: Oxford University Press.

Mongin AA and Orlov SN (2001) Mechanisms of cell volume regulation and possible nature of the cell volume sensor. *Pathophysiology* 8: 77–88.

Popp AJ, Feustel PJ, and Kimelberg HK (1996) Pathophysiology of traumatic brain injury. In: Wilkins RH and Renbachary SS (eds.) *Neurosurgery,* vol. 2, pp. 2623–2637. New York: McGraw-Hill.

Solenov E, Watanabe H, Manley GT, and Verkman AS (2004) Sevenfold-reduced osmotic water permeability in primary astrocyte cultures from AQP-4-deficient mice, measured by a fluorescence quenching method. *American Journal of Physiology – Cell Physiology* 286: C426–C432.

Strange K (2004) Cellular volume homeostasis. *Advances in Physiology Education* 28: 155–159.

Reinforcement Models

M Kawato, ATR Computational Neuroscience Laboratories, Kyoto, Japan

© 2009 Elsevier Ltd. All rights reserved.

The reinforcement learning algorithms proposed and developed by Andy Barto and Rich Sutton in the early 1980s have provided some of the most influential computational theories in neuroscience. The original theories were developed for explaining adaptive animal behaviors, but during the following decade they had a larger impact in the machine learning, computer science, and robotics communities than within the biological disciplines. In the mid-1990s, Wolfram Schultz produced epoch-making neurophysiological data on dopamine neuron activities, which led several theorists to propose computational learning models of basal ganglia and dopamine neurons based on reinforcement learning algorithms. This interaction between experiments and theories generated a steady and influential research trend, which has encompassed many previous and ongoing studies. Here, the temporal difference error plays the most important role in both experimental and theoretical studies of reinforcement learning. In brief, it is computed as the difference among the following three terms: the reward obtained at the current timing, the discounted expectation of the cumulative reward at the neighboring future timing, and the subtracted expectation of the cumulative reward at the current timing. In a class of reinforcement learning algorithms collectively called 'temporal difference learning,' the temporal difference (TD) error plays a central role in guiding acquisition of the predicted reward and behaviors. In the context of classical conditioning, the TD error is predicted to be positive at the time of the primary reward in early learning trials, but as learning progresses it is predicted to move at the time of a cue that indicates a future reward. If the expected reward from the cue is omitted, the TD error becomes negative at the time of the expected but nondelivered reward. Schultz and others confirmed all these predictions in studies of dopamine neuron activities. Because dopamine neurons innervate not only the basal ganglia but also a wide area of the cerebral cortex, learning in both regions might be able to be understood by reinforcement learning algorithms. Many experimental results supporting this hypothesis have been obtained from different levels (cell, circuit, brain, and behavior) utilizing different techniques (slice experiment, extracellular unit recording in behaving animals, lesion, and human imaging). It seems quite safe to state that at least some aspects of learning that are dependent on reward and penalty can be understood within the framework of reinforcement learning theory.

However, at least three difficult issues remain unresolved and should be studied from both theoretical and experimental standpoints. First, the plain reinforcement learning algorithm is too slow for practical problems to be considered as a realistic model of brain learning, which is not extremely slow. Second, although it is highly probable that dopamine neurons encode the TD error, the neural circuits and neural mechanisms needed to compute the TD error are still unknown. Third, although some functional aspects of the neural circuits consisting of dopamine neurons and the striatum of the basal ganglia might be understood by plain reinforcement learning algorithms, behavioral learning, which is dependent on global brain networks involving the cerebral cortex, cerebellum, basal ganglia, and midbrain, seems to necessitate much more complicated algorithms. This article discusses theoretical explorations for more efficient reinforcement learning algorithms and related experimental studies to resolve these difficulties.

Slow Learning of Plain Reinforcement Learning

The plain and basic reinforcement learning algorithm is attractive because it can learn new behaviors from only reward and penalty information without using explicit and quantitative models of controlled objects or environments. The trade-off is that the learning is extremely slow for any practical problem. Plain reinforcement learning algorithms can work reasonably well in a toy problem such as a discrete maze or in a game such as backgammon; however, even in such problems, tens of thousands of learning trials are often required, and millions of learning trials are not rare. Moreover, these algorithms fail badly for real-world problems such as those encountered in robotics, with continuous time and continuous state space. Performing even hundreds of learning trials is prohibitive for animal or robot learning. The slowness of the plain and basic reinforcement learning algorithm can be explained by either a spatial reason or a temporal reason. First, up-to-date humanoid robots often possess 50 degrees of freedom. Even if each degree of freedom is partitioned very roughly into five intervals, the total number of discrete regions necessary to cover the entire state space is $5^{(50\times 2)} = 5^{100} \approx 10^{70}$, which is astronomical since each degree of freedom necessitates position and velocity representations. Optimal (or suboptimal) solutions to the reinforcement learning problem are paths connecting some of these discrete regions;

these should be found by trial and error. Even if each region was examined with sub-millisecond duration, the solution would not be found within the life of the universe because of the enormous number of discrete regions. Human bodies possess at least 1000 degrees of freedom because the number of small muscles around the spine, important for posture control, is enormous. In the brain, neurons with population coding and overlapping tuning curves should replace the previously mentioned discrete partition of the state space, and 1000 degrees of freedom needs to be handled. Thus, a simple reinforcement learning algorithm cannot be expected to perform any faster than an artificial humanoid robot. Second, in most real-life examples of reinforcement learning, a reward is given to an animal only at the end of a long behavioral sequence. For example, meats can be eaten only at the end of a hunting tour lasting a few days, rice can be harvested only several months after seeds are sown, and entrance to university can be attained only after several years of hard study. The reward obtained at the terminal state of a sequence of behavioral decisions should be backtracked and temporal credit assignment to each decision should be computed. Some slow synaptic plasticity processes should propagate the representation of reward prediction at one time step to its preceding representations. Fully propagating the reward prediction from the terminal state to the initial state usually requires a very long time.

The three difficulties associated with plain reinforcement learning – slowness, computation mechanism for TD error, and incompatibility with brain structure – all necessitate specific structures, frameworks, and neural mechanisms in the brain for efficient learning algorithms, if the brain still adopts the essence of plain reinforcement learning algorithms. In computational and engineering studies aimed at finding efficient reinforcement learning algorithms, the following factors, at least, have been proposed and examined as accelerating mechanisms: hierarchy, modularity, blending with supervised learning, internal models, imitation learning, and preconditioning of search space. The first three factors are reviewed with several new models to resolve the difficulties.

Hierarchy and Modularity

If the number of regions to be explored for finding optimal solutions is astronomical, then solutions cannot be found within a reasonable amount of time, as stated previously. One possible solution to resolve this difficulty is to introduce a hierarchy in reinforcement learning algorithms. Consider a simple two-level hierarchy. In the upper hierarchy, the state space for the problem is very coarsely represented by a reasonable number of discrete regions, cells, or neurons. Because the number is not astronomically large, an optimal solution can be found in a reasonable amount of time at the upper hierarchy, but the representation of the solution is very approximate due to the coarse representation. In the lower hierarchy, the state space is finely represented by an enormous number of discrete regions, cells, or neurons. The approximate solution obtained at the upper level can constrain the possible range of the space in which the optimal solution at the fine level is explored. Thus, the reinforcement learning problem could be solved within a reasonable amount of time and with the required accuracy. The large-scale versus small-scale contrast between the higher and lower hierarchies applies not only to spatial representation but also to temporal representation. Multiple levels of hierarchy can also be introduced. However, the following theoretical issues arise from the introduction of a hierarchy. First, how to introduce coarse representations in the upper hierarchies is not at all trivial, especially if this process should be executed automatically. Second, some consistency should be maintained between different levels of the hierarchy regarding the state space representation, the trajectories and optimal solutions, the reward prediction, the reward prediction error, and the reward. These theoretical issues were studied in several proposals of hierarchical reinforcement learning algorithms (Composite Q, Feudal Q, HQ-learning, HAM, and Option). In most of these proposals, the researchers needed to specify coarse representations of the state space in the upper hierarchy. In some of them, reinforcement learning does not take place in either the upper or the lower hierarchy; that is, behaviors (policy) and reward expectation (value function) need to be prespecified by researchers. In many of the proposals, even local optimality of the obtained solutions is not theoretically guaranteed due to the lack of consistency between higher and lower hierarchies. Because of these limitations, most of the examples examined in previous proposals were discrete time and discrete space problems, such as navigation in a two-dimensional maze. The shortcomings of previous hierarchical reinforcement learning algorithms were quite serious obstacles to viewing the algorithms as possible computational theories of brain learning. In particular, the fact that researchers need to specify coarse representations in the upper hierarchy, or need to predetermine motor primitives or subgoals in the lower hierarchy, is nearly equivalent to assuming the existence of a homunculus or that those factors are all predetermined genetically, both of which are unacceptable from the viewpoint of neuroscience. One of the very rare exceptions among the continuous time and continuous state space examples was a hierarchical

reinforcement learning algorithm for standing up by a real multijoint robot.

One of the strategies common to the military, politics, and engineering is to deal with complicated and large-scale problems in a 'divide and conquer' manner. That is, to solve a difficult large-scale problem, it is first partitioned into several smaller and more tractable subproblems. Then, a responsible expert deals with each subproblem to obtain a partial solution. Finally, the obtained partial solutions are combined and a global solution to the original problem is obtained. For example, if a given problem is to move from a southern area of Kyoto prefecture in Japan to an Italian village in Sicily, taking an airport limousine to Kansai International airport is a subproblem and should be much more easily solved than the original problem. This modular approach is another promising strategy that has been explored in the context of reinforcement learning. Actually, most hierarchical reinforcement learning methods either explicitly or implicitly adopt the 'modular' approach. However, not all modular approaches are hierarchical. The most critical and difficult issue in modular reinforcement learning is how to divide or modularize the original problem into subproblems. Again, it is unsatisfactory from the viewpoint of neuroscience if researchers deal with this manually. Another difficult and important theoretical issue is to guarantee the global optimality of the entire problem since each expert deals mainly with a single subproblem and is essentially narrow-scoped. Several theoretical studies have addressed this global optimality issue.

Dividing the original problem based on space or time seems to be the most natural and efficient method for a modular approach. In a discrete state and discrete time example, such as a maze, this approach is frequently used because the division point is often very obvious, such as doors segregating different rooms and/or corridors, and the time division could follow the space division. However, note that the division of even discrete examples is still executed by researchers. Continuous state and continuous time examples, with which brains must deal, are much more complicated to modularize. In the continuous case, there is no apparent division point in either space or time. Furthermore, for neurons in the brain, which are involved in reinforcement learning of some problem, the coordinate system appropriate for the problem and its subdivisions is not explicitly provided. Each neuron receives sensory, motor, and reward information related to the problem, but there is no global picture of the whole problem and the coordinate system available to each neuron. Even in this difficult situation, modularity should spontaneously emerge via self-organization within neural networks.

One promising approach for self-organization of modularity is a modular selection and identification control (MOSAIC) model. MOSAIC was originally proposed for supervised learning in sensorimotor integration. Doya and colleagues extended MOSAIC to reinforcement learning, and further developments have been made. In the most advanced form of reinforcement MOSAIC, three different sets of expert networks self-organize through reinforcement learning (**Figure 1**). One set of experts includes predictors of state transitions. These can formally be called internal forward models. An internal forward model receives the current state and the current motor command and predicts the next time step state. In interpretations of neuroscience, forward models receive sensory feedback and an efference copy of motor commands and then predict the sensory feedback from the next time step. In MOSAIC, many forward models compete and cooperate to predict the state change based on the goodness of prediction of each forward model. That is, many forward models are different from each other; thus, they make different predictions despite receiving identical sensory feedback and motor command. Their predictions are compared with the actual sensory feedback at the next time point, and a responsibility signal for each forward model is computed. The responsibility signal for a forward model is larger when its prediction is better. The entire prediction by MOSAIC is the linear weighted sum of predictions made by its forward models, which are weighted by their responsibility signals. Two other sets of experts are for approximators of actual rewards and reinforcement learning controllers. The responsibility signals and linear weighted summation are computed similarly for these two sets of experts; however, the prediction error for the approximators of rewards involves errors in reward approximation, whereas that for the controllers involves reward prediction errors such as the TD error. The responsibility signal also gates the learning of forward models, reward approximators, and controllers by regulating learning rate by its amplitude. That is, if the responsibility signal of one expert is large, then the expert learns more. On the other hand, if its responsibility signal is small and close to zero, the expert learns little. Within this framework, three sets of experts self-organize while guided by three different kinds of error signal: prediction error in dynamics, approximation error in rewards, and reward prediction error. Consequently, modularity self-emerges without ad hoc manual tuning by researchers.

If the original reinforcement learning problem is rather simple, with a simple linear dynamics and a simple reward function shape, then MOSAIC will utilize only one forward model and only one reward

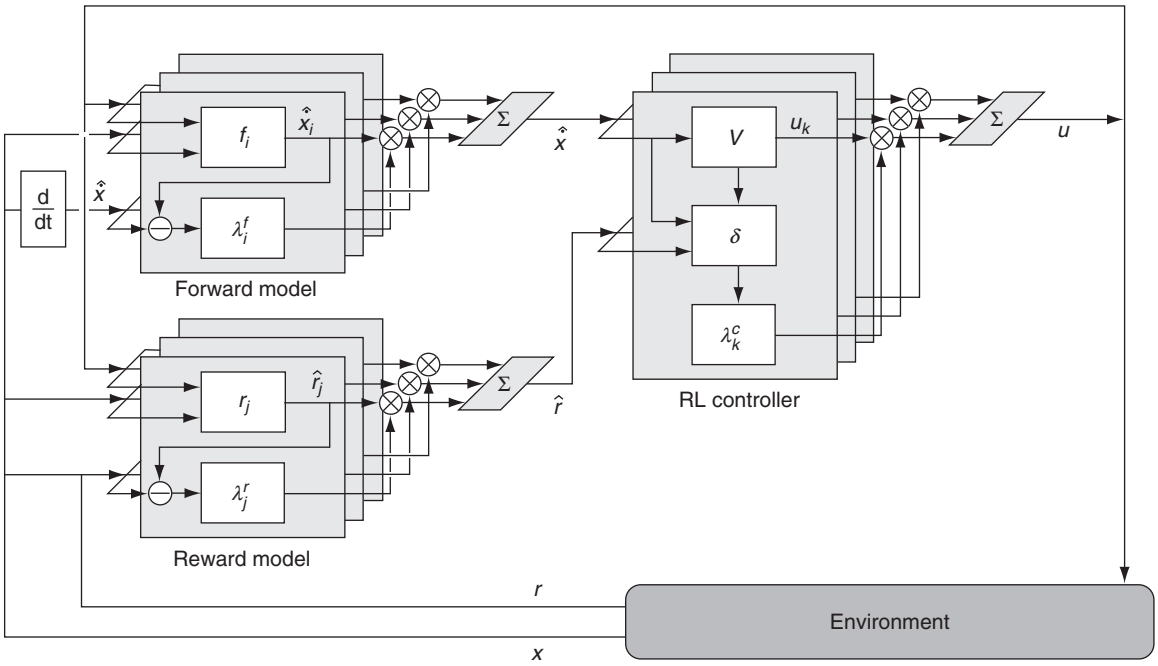

Figure 1 Overall organization of the combinatorial module-based reinforcement learning (CMRL) architecture. The architecture consists of forward modules (top left), reward modules (bottom left), and reinforcement learning (RL) controllers (right). The forward modules receive the state and action, and they segregate the environmental dynamics into several subdynamics based on the prediction error of the dynamics. Each reward module represents the local reward function, conditioned on the state variable (position and velocity) and the error of approximated reward. Then, an RL controller approximates the local value function for each configuration of a forward module and a reward module and outputs the local action. Finally, the local actions are weighted based on the squared TD errors of each RL controller. In real-world control problems, reinforcement learning has to handle nonstationary environments, in which both the dynamics of control and how the consequence of the control is rewarded change over time. One of promising strategies to deal with nonstationary environments is to decompose a complex task into several subtasks and adapt multiple RL modules for each of these subtasks. A basic assumption of this modular architecture is that the environment can be approximated as stationary within a single subtask. Previous studies of modular architecture for control attempted to separate a complex task based on the prediction error of the environmental dynamics and reported fairly good results for nonstationary dynamics. However, none of these studies addressed the dual nonstationary problem consisting of the dynamics and reward. To make RL applicable to such situations, the CMRL decomposes a complex task based on both the prediction error of the environmental dynamics and the approximation error of the reward function. First, the forward module (f_i) predicts the environmental dynamics ($\dot{x}$) from the current state and action (x, u). Second, the reward module (r_j) represents the local reward function conditioned on the state. The RL module, receiving outputs from forward and reward modules, approximates local value functions (V_k) and outputs the local greedy action (u_k). Finally, the CMRL selects actions by the sum of these local actions weighted by the responsibility signal of each RL module ($\lambda^c{}_k$), a posterior probability obtained from the squared TD errors ($\delta^2{}_k$) of each RL module (likelihood) and the usage history of component forward and reward modules (priors). These priors can be recomputed as the responsibility signals of forward and reward modules ($\lambda^f{}_i$ and $\lambda^r{}_j$) by applying Bayes rules to the prediction error of the environmental dynamics and reward approximation error, respectively. The CMRL can successfully cope with nonstationary problems such as the swinging of a pendulum, in multiple dynamics and reward conditions, and the bipedal locomotion of robots.

approximator. Accordingly, only one controller suffices to compute the necessary control commands. By contrast, if the dynamics of the original problem are highly nonlinear and consist of several qualitatively different dynamics, a set of forward models is recruited to approximate different segments of the original dynamics. If the dynamics are different, the necessary controllers are also different, even for the same reward function. Thus, a set of reinforcement learning controllers is recruited with corresponding forward models. In this nonlinear case, modular architecture dealing with nonlinear dynamics self-organizes. If the objective of the entire task changes from time to time, such as earning money and then spending it on a fan, then different reward approximators are recruited for different times, and different controllers are also needed. In this nonstationary case, modular architecture dealing with different task goals self-organizes. Thus, if the original problem contains nonlinearity and/or nonstationarity, MOSAIC automatically recruits its modular architecture to cope with the situation and self-organizes a set of controllers. A reinforcement MOSAIC model has also been extended to a hierarchical model.

Hierarchy and modularity are generally conceived as a common design principle for many different areas of the brain, based on an enormous number of previous anatomical and neurophysiological studies,

but direct support of their significance in the context of reinforcement learning mainly comes from neuroimaging studies on humans. For example, Haruno and colleagues found that the activities of many brain areas, such as the prefrontal cortex, orbitofrontal cortex, premotor cortex, supplementary motor area, cerebellum, and basal ganglia, are correlated with important variables in reinforcement learning, such as accumulated reward, learned behavior, short-term reward, and behavioral learning, even in a very simple Markov decision process with monetary reward. Furthermore, whereas the putamen is correlated with the reward prediction by action representation, which is consistent with a monkey study by Samejima and colleagues, the caudate is more correlated with the reward prediction error. Tanaka and colleagues found topographic representations of reward prediction in the medial prefrontal cortex and the insular cortex, and they found topographic representations of reward prediction error in the basal ganglia, with different values of discount factor, which is a very important parameter balancing future and immediate rewards in reinforcement learning algorithms.

How Can TD Error be Computed?

TD error plays the most critical role in solving the temporal credit assignment problem in reinforcement learning theory. Thus, Houk et al. proposed an explicit neural circuit model for computing the TD error, as shown in Figure 2(b), when they first proposed the reinforcement learning model of the basal ganglia and dopamine neurons shown in Figure 2(a). Doya then proposed the modified model shown in Figure 2(c). In both models, the primary reward information, $r(t)$, is assumed to be carried to the substantia nigra compacta (SNc) dopamine neurons by direct excitatory inputs. By contrast, the reward expectation of the next time step, $V(t+1)$, minus the reward expectation at the current time step, $-V(t)$, is assumed to be computed within the basal ganglia network utilizing double-inhibitory and inhibitory connections, respectively, from the common source of the value function $V(t)$ representation in the striatum. These models are attractive from a theoretical standpoint because the same value function $V(t)$ is the common source for positive and negative terms in the TD error, thus satisfying the most basic assumption of reinforcement learning algorithms: consistency of Bellman equation is the source of the error signal that drives learning. Moreover, these models utilize the known neural network within the basal ganglia. However, no direct physiological evidence has been obtained to support these models. Furthermore, double inhibition may not be able to explain the burst firing of dopamine neurons in response to a cue signal predicting a future reward, as in Schultz's experiment, due to the membrane properties of dopamine neurons.

Previous anatomical studies have shown that major excitatory synaptic inputs to SNc originate from the pedunculopontine tegmental nucleus (PPN), as shown in Figure 3(a). Neurophysiological studies by Kobayashi and colleagues further suggest that excitatory inputs from the PPN to the SNc might play the most central role in computing TD error. In their experiments, monkeys performed saccade tasks with variable reward amounts, which were indicated by the shape of a fixation spot. Two populations of neurons whose firing rates varied with reward amount were observed. One population of neurons seemed to encode the primary or current reward $r(t)$, whereas the other population of neurons started to fire upon the presentation of the fixation spot and maintained their firing until the end of saccade or the delivery of the primary reward, even after the fixation spot was extinguished. In addition, the firing rates predicted the amount of the future reward, thus seeming to encode the predicted reward or the value function $V(t)$. These remarkable findings directly suggest two possible neural mechanisms for computing the TD error, as shown in Figures 3(b) and 3(c). In Figure 3(b), some intranuclear circuits within the PPN or SNc, or some membrane properties of dopamine neurons in SNc, execute either temporal difference or temporal differentiation (in the case of Doya's continuous time reinforcement learning) as shown by the box in Figure 3(b). In this case, the PPN provides all of the components necessary for computing the TD error. By contrast, the model shown in Figure 3(c) predicts that the primary reward information $r(t)$ and the expected reward at the next time step $V(t+1)$ are carried by excitatory inputs from the PPN to the SNc, whereas the inhibitory input from the striatum conveys the subtracted and predicted reward information at the current time-$V(t)$. The model shown in Figure 3(c) seems more plausible because (1) no cellular or circuit mechanisms for temporal difference or differentiation are known that involve the PPN or SNc and (2) the inhibitory inputs from the striatum to the SNc are well established and it would be quite odd if they did not contribute to computing the TD error.

Heterarchical Reinforcement Learning Model

The fact that $V(t)$ and $V(t+1)$ arise from different sources (Str and PPN) in the model of Figure 3(c) casts doubt on its fidelity to reinforcement learning theory: consistency of the Bellman equation. By contrast, in the models shown in Figures 2(b), 2(c), and 3(b),

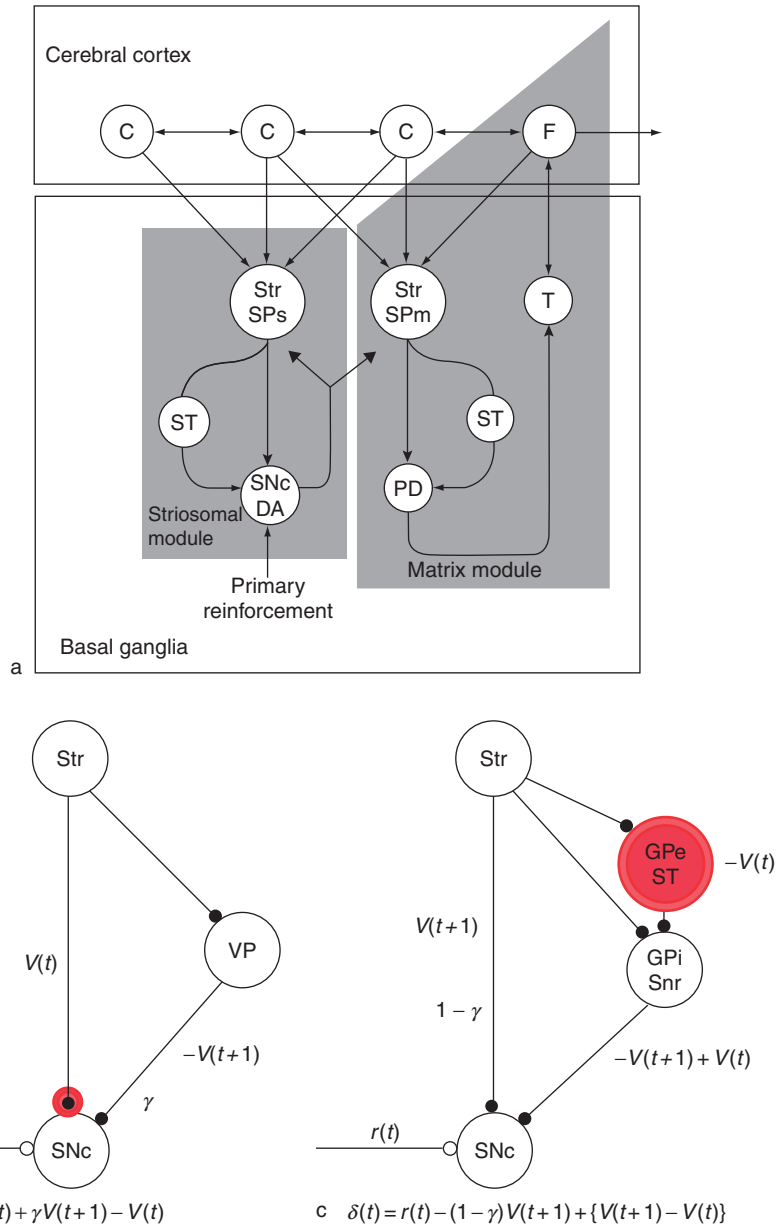

Figure 2 (a) Reinforcement learning model of the basal ganglia The modular organization of the basal ganglia, including both striosomal and matrix modules, is shown. C, cortical column composed of corticocoritical, corticothalamic, and other projection neurons; F, cortical column in the frontal cortex; Str, striatum; SPm, spiny neurons in the matrix compartment of the striatum; SPs, spiny neurons in the striosomal compartment of the striatum; SNc, substantia nigra pars compacta; DA, dopamine neurons; ST, subthalamic nucleus neuron; T, thalamocortical neuron; PD, pallidal neurons. (b, c) Houk versus Doya models of TD error computations. Possible mechanisms for TD computation of the value function represented in the basal ganglia are shown. Open circles ○ represent excitatory connections, whereas black dots • represent inhibitory connections. The large red circle represents the source of possible time delay. SNr, substantia nigra pars reticulata; GPi and GPe, internal and external segments of the globus pallidus, respectively; VP, ventral pallidum. (a) From Houk JC, Adams JL, and Barto AG (1995) A model of how the basal ganglia generates and uses neural signals that predict reinforcement. In: Houk JC, Davis JL, and Beiser DG (eds.) *Models of Information Processing in the Basal Ganglia*, pp. 249–270. Cambridge, MA: MIT Press.

the TD error is computed from the same source of the value function; thus, they conform to the most basic principle of temporal difference learning. However, deviation from this principle may lend efficiency and power to the model in Figure 3(c). If subtraction can be computed in the SNc between PPN and striatal inputs, we can introduce a supervised learning aspect into reinforcement learning. That is, the learner $V(t)$ tries to approximate the teaching signal $r(t) + V(t+1)$ and the error signal is conveyed by dopamine neurons. Indeed, many robotics applications of reinforcement learning algorithms have been forced to incorporate

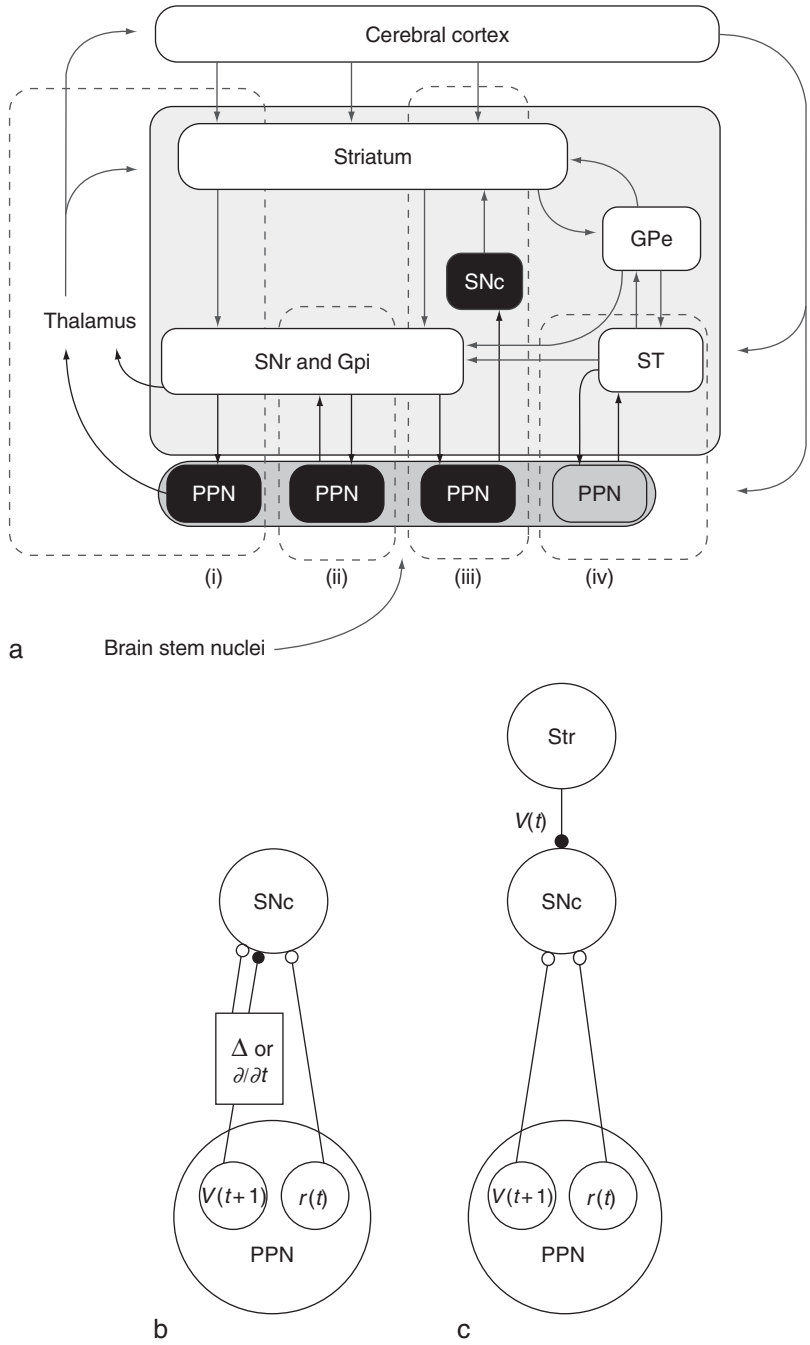

Figure 3 (a) Connections of the pedunculopontine tegmental nucleus (PPN) with the basal ganglia and cortex. (i) The PPN sends projections to the thalamic cells that project to the striatum; (ii) there are reciprocal connections between the PPN and the basal ganglia output nuclei (substantia nigra pars reticulata (SNr) and internal segment of the globus pallidus (GPi)), with the latter nuclei providing a dense inhibitory projection to the PPN; (iii) a substantial projection from the PPN innervates dopaminergic neurons of the substantia nigra pars compacta (SNc), which in turn modulate striatonigral and striatopallidal pathways; and (iv) a reciprocal, putative excitatory loop is formed between the PPN and the subthalamic nucleus neuron (ST). These complex interconnections imply that the PPN has a profound and widespread influence on basal ganglia activity and, similarly, that the basal ganglia are in a position to modulate PPN activity at many different levels. (b, c) Two possible neural circuit models for computing the TD error with an emphasis on excitatory inputs from the PPN to the SNc. Str, striatum. (a) From Mena-Segovia J, Bolam JP, and Magill PJ (2004) Pedunculopontine nucleus and basal ganglia: Distant relatives or part of the same family? *Trends in Neuroscience* 27: 585–588.

aspects of supervised learning to attain allowable learning time. The model shown in Figure 3(c) might represent a bridge between neuroscience and these efforts in robotics; furthermore, this model led to the proposal of the 'heterarchical' reinforcement learning model shown in Figure 4. In this figure, there exist two independent closed loops: the prefrontal cortex–caudate–ventral tegmental area–PPN cognitive loop

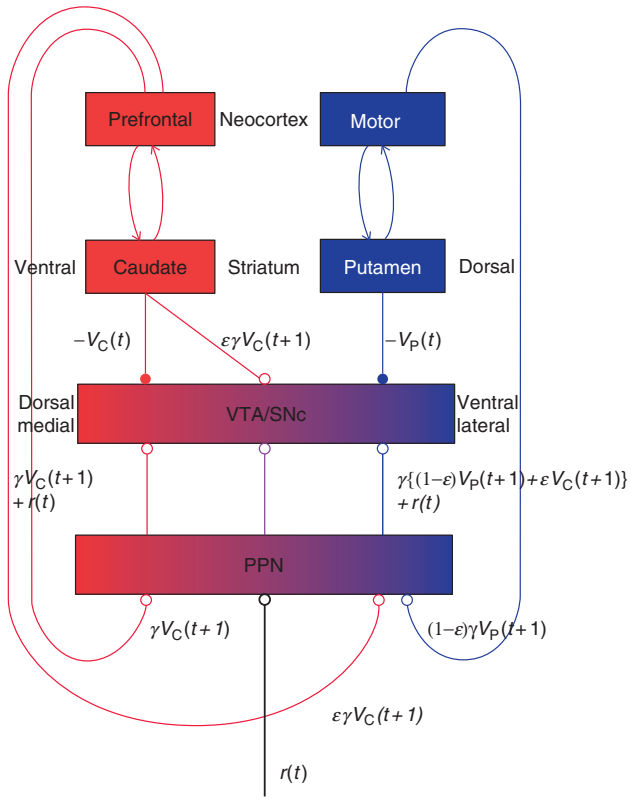

Figure 4 A basic heterarchical reinforcement learning model consisting of caudate–prefrontal and motor–putamen loops, the ventral tegmental area (VTA), the substantia nigra pars compacta (SNc), and the pedunculopontine tegmental nucleus (PPN). Open and solid circles show excitatory (disinhibitory) and inhibitory projections, respectively. V_C, a coarse value function in the caudate nucleus; V_P, a fine value function in the putamen; r, a reward; γ, a discount factor for future rewards; ε, a time-varying weighting factor between V_C and V_P.

and the motor cortex–putamen–SNc–PPN sensorimotor loop. If these two closed loops are entirely independent, then they can provide neural mechanisms for computing the TD error in a strictly reinforcement learning manner. However, the heterarchical structure shown in Figure 4 allows the cognitive loop value function to spread to the sensorimotor loop via neural connections from the caudate to the ventrolateral part of SNc, as well as through connections from the prefrontal cortex to the part of PPN innervating the ventrolateral part of SNc. This heterarchical model captures some aspects of the hierarchical reinforcement learning models introduced previously, but it does not have a strict hierarchy; thus, the name heterarchy was adopted. Representations of the state and actions in the cognitive loop are much coarser than those in the sensorimotor loop, so reinforcement learning progresses much faster in the cognitive loop than in the sensorimotor loop, just as with the higher and lower levels of hierarchical reinforcement learning algorithms. Then, the value function $V_C(t+1)$, which was learned earlier within the cognitive loop, plays the role of a teaching signal for the sensorimotor loop. That is, the sensorimotor (putamen) loop value function $V_P(t)$ is trained to approximate $V_C(t+1)$ during the early stages of learning. However, a much more fine-grained value function will be ultimately learned using the proper reinforcement learning implemented by the sensorimotor loop. Thus, fast but accurate learning is expected in the heterarchical reinforcement learning model, and no distinct boundaries should be manually set by researchers or homunculus. For simplicity, we explained the model using only two distinct closed loops, but a full model is rather continuous and realistic. This model was motivated by the neurophysiological studies of Kobayashi and colleagues and also by the anatomical studies of Haber on the spiral connections of corticostriatal and striatonigral loops. Kobayashi and colleagues characterized PPN neural firings as encoding intrinsic motivation or internal rewards. Because the PPN also receives inputs from brain stem nuclei, some portion of the future expected reward $V(t+1)$ could be represented based on quite coarse and endogenous factors, almost excluding sensorimotor or cognitive inputs. The coarse $V(t+1)$ could still be very helpful for developing more fine-tuned value functions based on sensorimotor and cognitive inputs, in a way similar to the technique known

as value function shaping in computational studies, and it could be a clue to understand the computational advantages of intrinsic motivation.

See also: Conditioning: Theories; Conditioning: Simple Neural Circuits in the Honey Bee; Delayed Reinforcement: Neuroscience; Delayed Reinforcement: Economics; Neuropsychology of Primate Reward Processes; Prediction Errors in Neural Processing: Imaging in Humans; Reward and Learning; Reward Processing: Human Imaging; Segawa Dopa Responsive Dystonia; Visual Cortical Models of Orientation Tuning.

Further Reading

Barto AG, Sutton RS, and Anderson CW (1983) Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics* 13: 835–846.

Dayan P and Hinton G (1993) Feudal reinforcement learning. *Advances in Neural Information Processing Systems* 5: 271–278.

Doya K (2000) Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology* 10: 732–739.

Haruno M and Kawato M (2006) Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops; fMRI examination in stimulus–action–reward association learning. *Neural Networks* 19: 1242–1254.

Houk JC, Adams JL, and Barto AG (1995) A model of how the basal ganglia generates and uses neural signals that predict reinforcement. In: Houk JC, Davis JL, and Beiser DG (eds.) *Models of Information Processing in the Basal Ganglia*, pp. 249–270. Cambridge, MA: MIT Press.

Kawato M (1999) Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* 9: 718–727.

Kobayashi Y, Inoue Y, Yamamoto M, et al. (2002) Contribution of pedunculopontine tegmental nucleus neurons to performance of visually guided saccade tasks in monkeys. *Journal of Neurophysiology* 88: 715–731.

Kobayashi Y, Okada K, Inoue Y, et al. (2002) Reward predicting activity of pedunculopontine tegmental nucleus neurons during visually guided saccade tasks. In: *Abstract of the 35th Annual Meeting of the Society for Neuroscience,* No. 890.5.

Mena-Segovia J, Bolam JP, and Magill PJ (2004) Pedunculopontine nucleus and basal ganglia: Distant relatives or part of the same family? *Trends in Neuroscience* 27: 585–588.

Morimoto J and Doya K (1999) Hierarchical reinforcement learning for motion learning: Learning 'stand up' trajectories. *Advances in Robotics* 13: 267–268.

Parr R and Russell S (1997) Reinforcement learning with hierarchies of machines. *Advanced in Neural Information Processing Systems* 10: 1043–1049.

Samejima K, Veda Y, Doya K, et al. (2005) Representation of action-specific reward values in the striatum. *Science* 310: 1337–1340.

Schultz W and Dickinson A (2000) Neuronal coding of prediction errors. *Annual Review of Neuroscience* 23: 473–500.

Singh S (1992) Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning* 8: 323–339.

Sugimoto N (2006) *Hierarchical Reinforcement Learning: Computational Model of Communication.* Nara Institute of Science and Technology, PhD Dissertation, No. NAIST-IS-DD0361013.

Sutton RS (1991) Planning by incremental dynamic programming. In: Brinbaum LA, Sutton RS, and Collins GC (eds.) *Proceedings of the Eighteenth International Workshop on Machine Learning*, pp. 353–357. San Mateo, CA: Morgan Kaufmann.

Sutton RS and Barto AG (1998) *Reinforcement Learning.* Cambridge, MA: MIT Press.

Sutton R, Precup D, and Singh S (1999) Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112: 181–211.

Tanaka SC, Doya K, Okada K, et al. (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience* 7: 887–893.

Wiering M and Schmidhuber J (1997) HQ-learning. *Adaptive Behavior* 6: 219–246.

Representation of Color

G D Horwitz, University of Washington, Seattle, WA, USA
T D Albright, The Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Wavelength and Color

Intuition tells us that color is a property of objects in the physical world, but in fact, color exists only in our internal, perceptual worlds. The physical world contains lights, which are composed of many wavelengths, and it contains surfaces, which reflect and absorb these lights to varying degrees. The wavelength composition, or spectrum, of light entering our eyes thus contains information about light sources and surface reflectances, and our visual systems analyze this information so that we can interact productively with our environment. Color is the quality of visual perception associated with this analysis.

The spectrum of a light is closely related to its color. A bright light with a flat spectrum looks white and the absence of light looks black. Colored lights have nonflat spectra: Long-wavelength lights tend to look reddish and short-wavelength lights tend to look bluish (Figure 1).

If we take a light with a fixed spectrum and double its intensity (doubling the photon flux at every wavelength) it looks brighter, but its color changes little, if at all. In this case, we say that we have doubled the luminance of the light but kept its chromaticity constant. Changing chromaticity without changing luminance is a more complex operation because luminance depends on both the spectrum of the light and the spectral sensitivity of the human visual system. By convention, two lights have the same luminance if they appear equally bright to the average human observer when they are shown in rapid alternation.

It is tempting to ignore luminance when considering the neural basis of color vision. This would be a mistake because the processing of luminance and chromaticity are tightly linked in the early visual system, as evidenced by the first stage of visual processing, the photoreceptors.

Cone Photoreceptors

In daylight, vision begins with the absorption of photons by three classes of cone photoreceptor at the back of the retina. This section describes how light is encoded by the cones and how cone signals are combined into the channels that provide the basis of color vision.

Light Coding in Cone Photoreceptors

The probability that a photoreceptor of a given class absorbs a photon depends on the wavelength of the photon. The three classes of cone photoreceptor are named after the wavelengths to which they are maximally sensitive: Long-wavelength cones, or L-cones, are maximally sensitive to long-wavelength photons (560 nm); M-cones are maximally sensitive to medium-wavelength photons (530 nm); and S-cones are maximally sensitive to short-wavelength photons (420 nm) (Figure 2).

There exists an infinite array of physically different lights that stimulate the cones identically. Pairs of such lights are called metamers and are visually indistinguishable. Metamers are the foundation of many common visual displays. For example, the lights generated by a color television have spectra that differ profoundly from those in the natural world, but because television-produced and natural lights can affect the cones in the same way, they give rise to indistinguishable color percepts. Lights that look reddish tend to excite L-cones well, and lights that look blueish tend to excite S-cones well. It is thus tempting to think of each cone class as leading directly to a corresponding color percept (e.g., L-cone excitation = red). It is easy to see that this is incorrect. A flat spectrum light, which looks white, excites L-cones moderately well. Doubling the intensity of this light doubles the excitation of the L-cones, and yet the light continues to appear white, not reddish. Our perception of redness is thus not strictly a function of L-cone excitation alone.

To a first approximation, changing the intensity of a light affects all three cone classes identically; increasing light intensity increases the number of photons caught by each cone. Changes in light intensity thus preserve the relative levels of excitation across cone classes and, as mentioned previously, the color of the light. These observations suggest that color perception depends on a comparison of excitation across the cone classes.

Cone Opponency: Theory

Comparing excitation across cone classes makes good sense from a computational perspective. L- and M-cones have broad and overlapping spectral sensitivities, so most lights excite them to similar degrees. The visual system emphasizes the small differences that exist between the L- and M-cone signals, however, by subtracting them. This subtraction facilitates

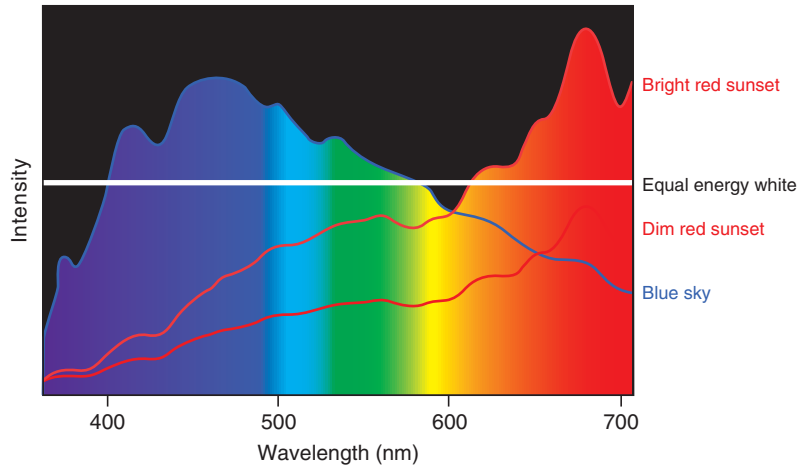

Figure 1 Spectra of a bright red sunset, a dim red sunset, a blue sky, and a light with equal energy at all wavelengths (equal energy white). Adapted from Rechtsteiner GA and Ganske JA (1998) Using natural and artificial light sources to illustrate quantum mechanical concepts. *The Chemical Educator* 3: 3, figure 3, with kind permission of Springer Science and Business Media.

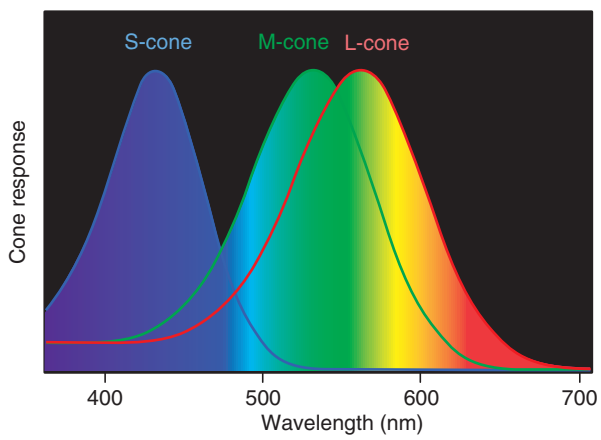

Figure 2 Spectral absorption functions of the three cone photoreceptors.

the discrimination of very small differences in wavelength around ~550 nm, where the L- and M-cone absorption spectra overlap.

Cone difference signals carry important information about the world around us. For example, the nutritional value of a leaf is better indicated by the difference in L- and M-cone signals than by either one alone (this fact has been suggested as a significant pressure in the evolution of opponency between L- and M-cone signals in Old World primates). This signal, however, is quite small with respect to the absolute levels of L- and M-cone excitation, which depend on ancillary factors such as whether the leaf is in direct sunlight or shadow. The differential signal is immune to the absolute levels of cone excitation and is therefore useful for finding nutritious leaves irrespective of the illumination.

Cone Opponency: Implementation

The biological substrate for the comparison of cone signals is a group of cone-opponent retinal neurons, which are excited by some cones and inhibited by others. Midget bipolar cells receive antagonistic input from L- and M-cones so their output reflects the difference between L- and M-cone excitation. Small bistratified ganglion cells are excited by S-cones and inhibited by a combination of L- and M-cones.

Cone-opponent signals in the retina are transmitted along the optic nerve to the lateral geniculate nucleus of the thalamus (LGN), where they are preserved faithfully. Neurons in the LGN are arranged in layers, each of which receives input from a different population of retinal ganglion cells. In an individual layer, neuronal cell bodies have a characteristic size after which they are named. Parvocellular neurons ('parvo' meaning 'small') receive input from the midget pathway and are thus described as having 'L versus M' or 'L – M' tuning. At least some koniocellular neurons ('konio' meaning 'dustlike') receive input from small bistratified ganglion cells and are thus described as having 'S – (L + M)' tuning. A third population of cells, the magnocellular neurons ('magno' meaning 'large'), is not cone opponent. These cells are excited by both L- and M-cones and are thus described as having 'L + M' tuning.

Magnocellular neurons have been proposed to be responsible for our sensation of luminance, and parvocellular and koniocellular neurons have been proposed to be responsible for our sensation of chromaticity. This model is supported by three observations. First, magnocellular neurons respond more vigorously to changes in luminance than do either parvocellular or koniocellular neurons. Second,

luminance flicker can be detected at much higher frequencies than chromatic flicker, and magnocellular neurons are particularly sensitive to high-frequency flicker. Third, at least two cone-opponent channels are required to account for human color perception, and neurons with L − M and S − (L + M) tuning are attractive candidates for these channels.

The truth appears to be more complex, however. Neurons in all three categories respond to changes in chromaticity as well as to changes in luminance, albeit to different degrees. Furthermore, the spatial resolution of these cell types conflicts with predictions from human psychophysics. Human observers can easily see fine spatial patterns defined by luminance modulations but have difficulty seeing similarly fine chromatically defined spatial patterns. One might thus expect that neurons with L + M tuning would have small receptive fields, implying high spatial resolution, and neurons with L − M tuning would have large receptive fields, implying low spatial resolution. The truth is exactly the opposite: Magnocellular (L + M) neurons have larger receptive fields than parvocellular (L − M) neurons.

This apparent paradox is elegantly resolved by the receptive field organization of a subclass of parvocellular neuron. Parvocellular type I neurons are excited by one cone class in the receptive field center and suppressed by another cone class in the receptive field surround. Type II neurons, in contrast, are excited by one cone class and suppressed by another cone class throughout the receptive field. A type I neuron with an L-cone-dominated excitatory center and an M-cone-dominated inhibitory surround responds well to a change in light that increases L-cone excitation and decreases M-cone excitation, and this is true throughout the receptive field. The same neuron, however, responds differently to lights that modulate L- and M-cone excitations together. Such a light drives the cell poorly if it covers the entire receptive field because center and surround contributions cancel, but a small L + M spot restricted to the center drives the cell quite well. Type I neurons thus carry information about chromaticity at coarse spatial scales and luminance at fine spatial scales, illustrating again the close relationship between luminance and chromatic processing in the early visual system (Figure 3).

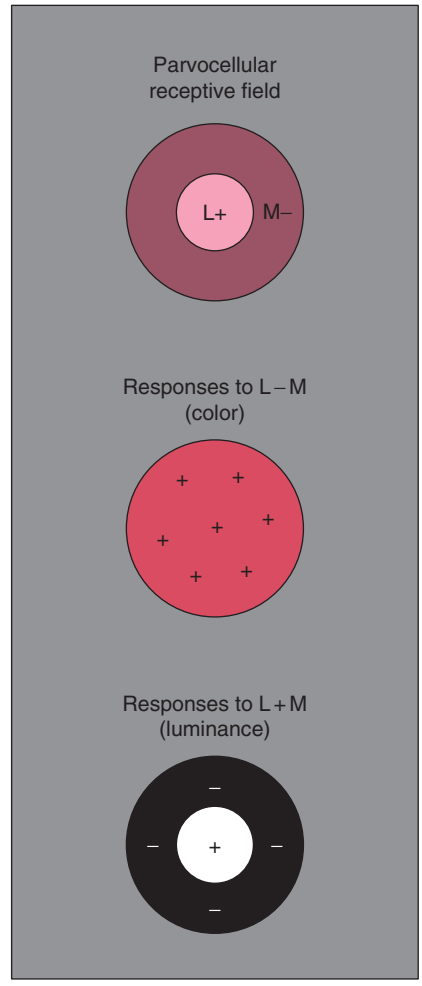

Figure 3 Schematic receptive field structure of a type I parvocellular LGN neuron (top). L − M modulations excite the cell throughout the receptive field (middle), whereas L + M modulations are excitatory in the center and suppressive in the surround (bottom). A relatively small M-component to the receptive field center has been omitted from the figure for clarity.

Cone Opponency: Relationship to Opponent Colors

Although wavelength is naturally represented on a line, our perception of hue wraps around on itself, creating the familiar color circle. Red and green appear at opposite sides of the color circle because they are mutually exclusive hues: A single light never appears at once reddish and greenish. The same can be said of blue and yellow. On the other hand, bluish-reds (purple), yellowish-reds (orange), bluish-greens (turquoise), and yellowish-greens (lime) are easily seen.

In 1878, the German physiologist Ewald Hering hypothesized that the perception of any color derives from a combination of four elemental sensations: red, green, blue, and yellow. According to Hering's model, these sensations are the result of two independent physiological processes: a red–green process and a blue–yellow process. The red–green process cannot be in the red and green states simultaneously, and the blue–yellow process cannot be in the blue and yellow states simultaneously, which accounts for the mutual exclusivity of opposite hues. Hering's idea, that color vision is founded on four primary sensations, was thought originally to be incompatible with the view championed by Hermann von Helmholtz

and Thomas Young that color vision was based on three sensors in the eye. Today, the ideas can be reconciled: The three sensors of Helmholtz and Young correspond to the cones, and the two opponent processes of Hering correspond (roughly) to the cone-opponent neurons in the retina and LGN.

Cone-opponent neurons in the LGN behave similarly to the opponent process channels predicted by Hering. L- and M-cone signals cannot be dominant simultaneously in $L-M$ neurons just as the red–green process cannot be in the red and green states simultaneously. Likewise, S- and $(L+M)$ cone signals cannot be dominant simultaneously in $S-(L+M)$ neurons just as the blue–yellow process cannot be in the blue and yellow states simultaneously. Because of these relationships, $L-M$ neurons are sometimes referred to as 'red–green cells' and $S-(L+M)$ neurons are sometimes referred to as 'blue-yellow cells.' It is important to note, however, that LGN neurons differ in important ways from the processes predicted by Hering. LGN cells are not tuned to the unique hues, which are the particular shades of red, green, blue, and yellow that human observers agree are pure and uncontaminated by other hues. Thus, a change in the blueness of a light affects activity in both the $S-(L+M)$ and $L-M$ pathways. Together, these observations suggest that cone opponency at the level of the retina and LGN is important for the mutual exclusivity of complementary hues, but additional processing of color signals in the cortex may also play an important role.

Color Processing in the Cortex

Color processing in the cerebral cortex is an active area of research. This section describes some of the better understood aspects of cortical color processing.

Signal Mixing in V1

The three major cell types in the LGN project to distinct layers of the primary visual cortex (area V1, which is also called striate cortex). Magnocellular neurons project primarily to layer 4Cα, parvocellular neurons project primarily to layer 4Cβ, and koniocellular neurons project primarily to layers 2/3. Lower layer 3 receives information about increases in S-cone excitation, and upper layer 4A receives information about decreases in S-cone excitation. Whether these signals have a koniocellular or parvocellular origin is unclear.

The segregation of LGN inputs to V1 is not maintained long because intracortical connections within V1 mix signals across pathways. The nature of this mixing is beyond the scope of this article, but it is worth noting that some of the underlying intracortical circuitry is highly stereotyped and some of this signal mixing is likely important for color vision.

In principle, the mixing of LGN signals in V1 could account for the differences between the color tuning of LGN cells and perceptually opponent unique hues. For example, by mixing $L-M$ and $S-(L+M)$ signals in the appropriate proportions, the brain could create discrete classes of V1 neurons whose activity mimics the human perception of redness, greenness, blueness, and yellowness. Again, however, the truth appears to be more complicated: The distribution of preferred colors in V1 appears to be more or less random, implying indiscriminate mixing of LGN signals.

The apparent randomness of preferred colors in V1 may be a manifestation of the complexity of processing that occurs here. Neurophysiologists make simplifying assumptions about the computations that neurons perform because the full range of possibilities cannot be tested during an experiment of realistic duration. One common assumption is that V1 neurons add and subtract LGN inputs but do not perform more complex operations. LGN neurons, in turn, can be modeled as adding and subtracting cone signals, so this assumption is formally equivalent to asserting that V1 neurons add and subtract cone signals. In this case, the color tuning of a V1 neuron can be described by a set of weights with which it combines cone signals. Cone weights estimated from the activity of V1 neurons are nearly random, consistent with actual random color tuning or, alternatively, with complex, orderly color tuning that is not yet understood.

Color Processing in V1

Several important specializations for color processing have been discovered in V1. First, V1 is the earliest stage of the primate visual system to contain cells that are excited by some wavelengths and inhibited by others in the center of their receptive fields and have the reverse spectral sensitivity in the surround. These cells are called double opponent because they exhibit both wavelength and spatial opponency. The receptive fields of some double-opponent cells are nearly circularly symmetric, rendering them sensitive to spatial color contrast irrespective of orientation. Other double-opponent cells have receptive fields with elongated, side-by-side subunits, rendering them sensitive to edges of a particular orientation and chromaticity. Many double-opponent cells respond to both color and luminance modulations, again illustrating the close relationship between color and luminance processing in the early visual system (Figure 4).

A second specialization for color processing in V1 derives from the location of the cone-opponent neurons. Layers 2/3 of V1 contain small, regularly spaced

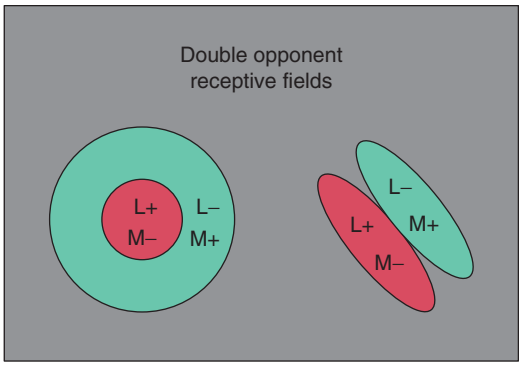

Figure 4 Schematic receptive field structure of double-opponent cells.

regions that stain darkly for the enzyme cytochrome oxidase (these regions are called cytochrome oxidase blobs). Many neurons in the blobs are cone opponent and many receive a strong koniocellular input. Koniocellular input is consistent with a role in color processing because S-cone signals, at least some of which are carried by the koniocellular pathway, play a greater role in chromatic vision than they do in luminance vision. Many neurons in the blobs project to a subcompartment of area V2, which is the next stage of cortical visual processing.

Color Processing in V2

Area V2 can be divided into three subcompartments, which are named the thick, thin, and pale stripes after their appearance when stained for cytochrome oxidase. Neurons in the thin stripes receive input preferentially from V1 neurons in the cytochrome oxidase blobs, suggestive of a role in color vision.

Our understanding of color processing in V2 is less developed than our understanding of color processing in V1. Nevertheless, a comparison between these areas reveals two interesting distinctions. First, color processing in V2 is even less linear than in V1; many V2 neurons respond to an extremely narrow range of colors that is inconsistent with a weighted sum of cone signals. Second, strongly color-responsive V2 neurons are spatially arranged according to their color tuning. For example, V2 neurons that are excited by purple are physically close to those that are excited by blue and others that are excited by red. This physical arrangement of neurons thus reflects our perception of hue similarity.

Color Processing in V4

Neurons in the thin stripes of V2 project to area V4, which was once considered a color center in the monkey brain. Early studies reported an unusually high proportion of color-tuned neurons in V4, but recent studies have reported unexceptional proportions. Lesions of V4 impair a variety of visual functions including, but not restricted to, color perception. V4 is thus unlikely to be specialized for color processing exclusively. On the other hand, some V4 neurons compensate for the spectrum of illumination in a manner similar to human color perception (human observers compensate effortlessly for surprisingly large changes in illumination, as described later). V4 is unlikely to be single-handedly responsible for this important feature of color perception but may contribute to it.

Color Constancy

Two factors largely determine the cone excitations produced by illuminated objects. Consider an apple; the first factor is the reflectance of the skin of the apple. Red Delicious and Granny Smith apples reflect different wavelengths of light with different efficiencies. Red Delicious apples reflect long-wavelength light better than do Granny Smith apples, making Red Delicious apples appear red and Granny Smith apples appear green under normal illumination.

The second factor affecting cone excitation is the spectrum of the illumination. Incandescent bulbs produce more long-wavelength light than do fluorescent bulbs. Thus, apples of both varieties reflect more long-wavelength light when lit by an incandescent bulb than when lit by a fluorescent bulb. Nevertheless, Red Delicious apples continue to look red and Granny Smith apples continue to look green under either illuminant. This consistency of color perception, which is called color constancy, implies a mechanism in the visual system that discounts the spectrum of the illumination.

Color constancy is likely a consequence of several physiological processes. Cone responses adapt over time, and this adaptation can partially account for color constancy. Some V1 neurons have been shown to exhibit a degree of color constancy which may be a consequence of double-opponent receptive field organization. Cognitive factors, such as an observer's knowledge of the physical layout of a visual scene, also play a role in color constancy. Elucidating the neural mechanisms underlying color constancy remains an important goal.

Color Processing in MT

The middle temporal area (area MT) of primate cerebral cortex is specialized for the analysis of visual motion. Neurons in this area respond well to moving luminance patterns but poorly to moving patterns defined solely by chromatic contrast. Perhaps as a consequence of this fact, moving luminance patterns are more easily seen than moving chromatic patterns.

The chromatic sensitivity of MT neurons may derive from at least three sources. First, a heterogeneous population of magnocellular neurons provides the dominant input to MT. On average, magnocellular neurons integrate L- and M-cone signals with roughly equal weight, but individual magnocellular neurons weight L- and M-cone inputs unequally. As a result, stimuli that modulate L- and M-cones with equal strength and opposite sign minimize the magnocellular population response while still driving some individual magnocellular neurons reasonably well. Second, many individual magnocellular neurons respond well to chromatic edges without regard to the colors on either side of the edge. The signals transmitted by these neurons are thus sufficient to mediate the computation of motion direction of a chromatic edge, but they are not sufficient to identify the colors on either side of the edge. Third, MT neurons receive modest inputs from koniocellular and parvocellular pathways which may provide overtly cone-opponent signals.

Color signals in different parts of the brain are probably used to perform different tasks. Color signals in MT are likely used to determine the locations of object boundaries for the analysis of visual motion but are unlikely to be important for the identification of surface properties. This task is likely performed by neurons in the ventral temporal lobe.

Color Processing beyond V4/MT

Human patients with damage to a small region of the ventral temporal lobe suffer from a disorder called acquired cerebral achromatopsia. These patients report that the colors they see in their day-to-day lives are washed out. In the laboratory, they perform poorly on tasks that require color naming or arranging stimuli according to their colors.

The part of the brain damaged in acquired cerebral achromatopsia was originally thought to be the human homolog of monkey area V4 but is now believed to be several centimeters more anterior. Currently called VO or V8, this area is strongly modulated by colored stimuli in functional imaging experiments, further suggesting its participation in color vision.

Determining whether a VO (V8) homologue exists in the monkey brain is an important direction for future research. Anatomical considerations suggest that such an area may exist in the inferior temporal cortex. Some neurons in this part of the brain respond to single hues. Others respond to multiple perceptually dissimilar hues. Some respond to highly saturated colored stimuli – in some cases irrespective of hue. Lesions of this area create a syndrome in monkeys that is similar to cerebral achromatopsia in humans. The weight of the current evidence thus suggests that a cortical area in monkeys and humans anterior to V4 is specialized for the analysis of color, but we do not yet know what processing occurs here, how this processing is implemented, and what functions it serves.

Conclusion

Color processing in the early visual system has been reasonably well described, but many fundamental questions remain: How are LGN signals combined in the cortex and what are the functional consequences of these combinations? How is color constancy achieved? What is the physiological basis for the unique hues?

The difficulty in understanding how the brain processes color information arises in part from the fact that color and other aspects of visual processing are closely related. As discussed in this article, the processing of chromaticity and luminance is tightly linked in the early visual system. Modulations of chromaticity and luminance often derive from different physical sources, and they give rise to different perceptual experiences. Nevertheless, both types of information are processed by the same neural hardware in the early visual system, and fixing one while varying the other in an experiment can lead to a confusing pattern of results.

Other examples abound. Psychophysical studies in humans have shown that the spatial, temporal, and chromatic context of a light can have dramatic effects on its color. This complicates the study of color processing: The spatial and temporal properties of the stimuli used in neurophysiological experiments, which are usually treated as nuisance parameters and fixed arbitrarily, affect color tuning measurements. This complication has a silver lining, however, because the relationships among visual attributes may provide valuable leverage into the problem. For example, the neural signals that underlie color perception should vary with the spatial and temporal properties of a stimulus in the same way that color perception does. Exploiting these relationships may help reveal how electrical signals in the brain give rise to color perception.

See also: Cerebral Cortex; Color Vision; Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Color Mechanisms.

Further Reading

Gegenfurtner KR (2003) Cortical mechanisms of colour vision. *Nature Reviews Neuroscience* 4: 363–386.

Komatsu H (1998) Mechanisms of central color vision. *Current Opinion in Neurobiology* 8: 220–231.

Lennie P and Movshon JA (2005) Coding of color and form in the geniculostriate visual pathway (invited review). *Journal of the Optical Society of America A: Optics, Image Science, and Vision* 22: 2013–2033.

Martin PR (2004) Colour through the thalamus. *Clinical and Experimental Optometry* 87: 249–257.

Rechtsteiner GA and Ganske JA (1998) Using natural and artificial light sources to illustrate quantum mechanical concepts. *The Chemical Educator* 3: 3.

Representation of Movement

N S C Price and R T Born, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Motion sensitivity is a fundamental aspect of vision for most animals. While the most obvious use for representing motion is to determine if an object is moving, analysis of motion serves a number of other roles. Assessing movements in the environment is critical to track, avoid, or intercept objects that may be food, friends, or foes. Motion analysis can be used for trajectory prediction, including determining if a moving object can be intercepted and the time to contact with the object. It also enables the resolution of ambiguities in two-dimensional relationships between objects. For example, depth information can be recovered through motion parallax, while motion segmentation can be used for camouflage breaking as occurs when a hidden animal only becomes distinct from its surroundings when it moves.

Motion can also be self-generated by eye, head, or body movements. Analyzing self-motion helps calibrate proprioception and locomotion and can be used to stabilize the retinal image to prevent motion-induced blur. The significance of analyzing motion is highlighted by the large proportion of motion-sensitive neurons in all animals that have been studied to date and the finding that while many of these species are insensitive to other visual modalities, such as disparity or color, no species lack motion sensitivity.

The focus here is on the neural representation of visual motion in primates, drawing briefly on results from models and psychophysics. The assessment addresses how the representation of movement changes along a hierarchy of brain regions that all contain motion-sensitive neurons and how this neural encoding ultimately affects perception and behavior.

Fundamentals of Motion Detection

In order to detect motion in a given direction a scene must be sampled at a minimum of two points in space at two different times. For example, the motion sensor in Figure 1(a) is sensitive to motion to the right if the time taken to traverse the distance Δd between the two input channels matches the temporal delay Δt. In addition, there must be a nonlinear interaction between the inputs from the two channels to ensure that the time-averaged output of the detector is direction selective. Early studies of the retina suggested that inhibition of responses to motion in the opposite direction to the preferred (the 'antipreferred,' or null, direction) is critical for generating physiological direction selectivity. This inhibition could either act linearly as subtraction or, more likely, nonlinearly as a shunting inhibition to suppress responses to motion in the antipreferred direction.

An early model incorporating the three basic requirements was Reichardt's cross-correlator, which was inspired by studies of the optomotor response of beetles (Figure 1(b)). The Reichardt detector is best stimulated by spatial wavelengths four times the separation of the input channels, since this leads to a 90° phase difference between the intensity variations at the two inputs (spatial quadrature). Further, for wavelengths less than twice the input channel separation the detector can suffer from spatial aliasing, as predicted by Shannon's sampling theorem. An important variant, the spatiotemporal energy model, incorporates simple linear spatial and temporal prefilters that reduce aliasing and help constrain the range of stimuli to which the detector responds (Figure 1(c)). This model highlights the idea that motion-sensitive neurons are filters tuned along a set of stimulus dimensions such as spatial and temporal frequency, rather than simply feature detectors, as suggested by the model in Figure 1(a).

With only two input channels, it is obvious that an 'apparent motion' stimulus, in which the stimulus appears at one location, and then the next, without physically moving between the two locations, could stimulate the detectors described in the preceding paragraphs as effectively as true, continuous motion. Similarly, motion sensors like the Reichardt detector can also falsely detect motion in random luminance flicker. These observations raise the obvious problem of determining whether the two input channels have been sequentially activated by one moving object or by two different objects – the so-called correspondence problem. A special form of this problem exists even for single objects that contain extended contours – as is the rule in natural scenes – due to the limited size of the sampling regions. The true direction of a moving edge viewed through a circular aperture cannot be resolved, a dilemma known as the 'aperture problem.' For example, in Figure 2(a), the edge could be moving slowly to the right, or more rapidly along a diagonal up and right, or down and right. Related to this is the issue of determining whether two adjacent local motion cues belong to one or two objects (Figure 2(b)). This problem is resolved if the vertex of an object is visible through an aperture, since it provides an unambiguous two-dimensional (2D) motion

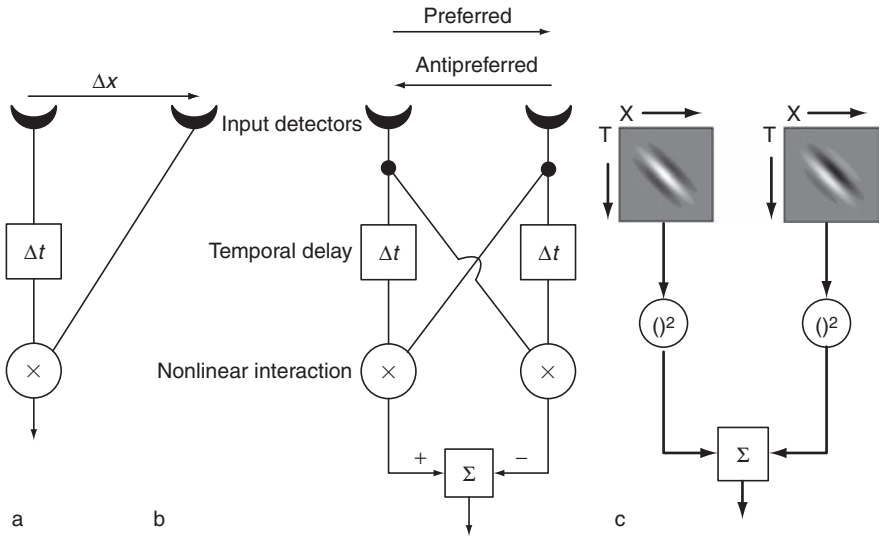

Figure 1 Models of motion detection. (a) An elementary motion detector that prefers motion to the right incorporates spatial separation of input channels (Δx), a temporal delay (Δt) in the left channel, and a nonlinear interaction between the input channels (in this case, multiplication, '×'). (b) In the Reichardt cross-correlator, the outputs of two elementary motion detectors preferring motion in opposite directions are subtracted. (c) The spatiotemporal energy model calculates motion energy by combining the squared outputs of two oriented space and time (X, T) filters that are in quadrature. The slant in the contour plots of the space–time filters indicates that they prefer motion to the right.

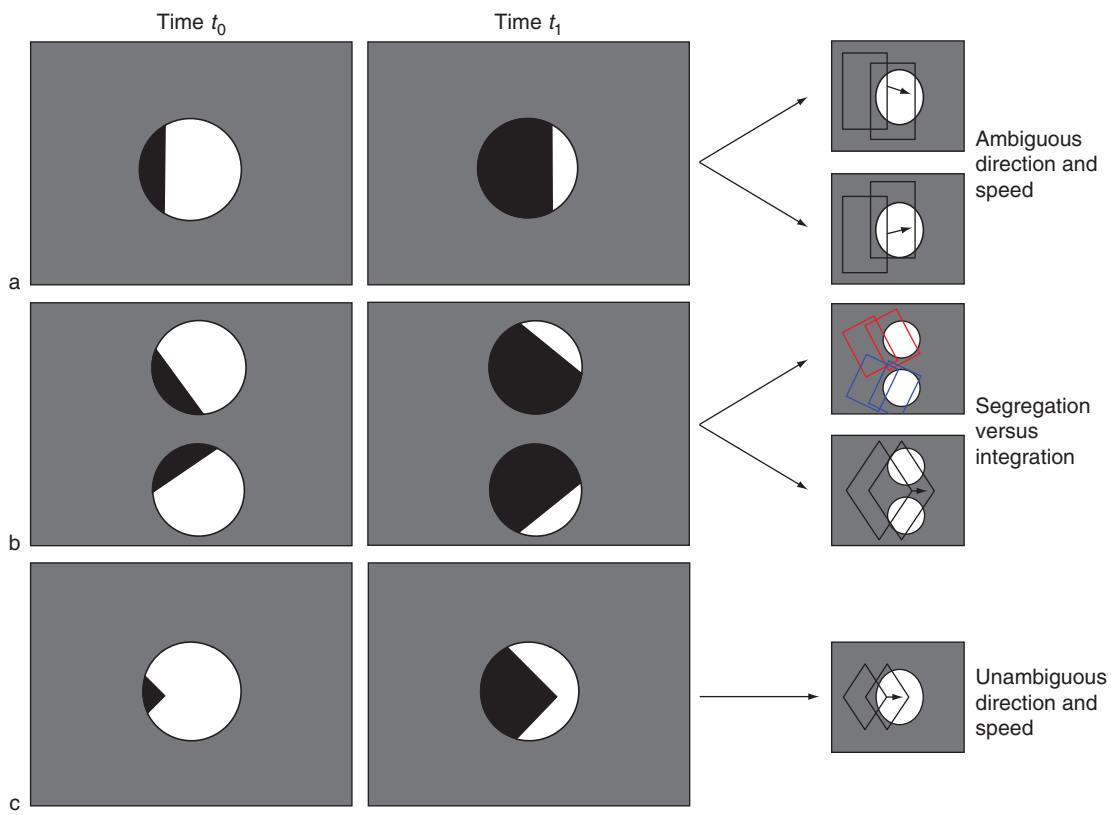

Figure 2 The aperture problem and motion segregation and integration. Each row shows the view through one or two apertures at two points in time and on the right, the possible object motions that could account for these views. (a) The direction and speed of a single moving edge are ambiguous, thus creating the aperture problem. (b) The motion of two edges viewed through two apertures can be segregated, meaning that they are treated as independent objects with ambiguous direction and speed, or integrated, resulting in an unambiguous percept. (c) An object vertex provides two-dimensional information and thus an unambiguous percept of speed and direction.

cue (Figure 2(c)). Thus segmentation (determining which local motion cues belong together and which do not) and integration (combining the cues that do belong together) are important, interdependent processes for determining what global movements are actually happening in the environment.

One reason for pooling the responses of local motion detectors is to create robust responses in the face of low signal-to-noise ratios. In this context, noise could arise from local motion detectors signaling directions contrary to the true global motion direction due to flicker or spatial aliasing. Alternatively, the strength or certainty of a motion signal may be weak because of low luminances or contrasts. Two primary methods exist for decreasing noise sensitivity: first, signals from many detectors can be averaged over a large spatial region or over an extended period of time; second, a detector can incorporate motion opponency, whereby preferred and antipreferred direction motions respectively excite and suppress the detector's response. Motion opponency is explicit in the Reichardt detector (Figure 2(c)) and serves to reduce the response to stationary flicker or simultaneous motion in the preferred and antipreferred directions.

Introducing noise to stimuli has been a powerful method for probing the neural basis of motion perception. Changing stimulus strength, or dot coherence (the proportion of dots that move in a single direction versus those that move in random directions), affects both the perception and neural response to motion. Thus variable coherence dot stimuli have allowed investigators to measure the relationship between a neuron's activity and the perception of motion. The results of such studies have provided the most compelling links between neural activity and perceptual decisions to date.

Psychophysics

A student of physics can explain that motion is represented by a velocity vector, consisting of a speed and a direction. Since speed represents the change in position over time, it is conceivable that the visual system is not explicitly sensitive to motion, *per se*, but generates motion sensitivity by judging changes in position over time. However, there is ample evidence demonstrating that position and motion are independently represented. The most compelling demonstration of this independence is the motion aftereffect (MAE; also known as the 'waterfall illusion'): after observing a moving pattern, a stationary pattern appears to move in the opposite direction. This produces a strong sense of motion without the perception that the pattern changes position over time. Akinetopsia (visual motion blindness) also demonstrates the segregation of motion and position processing. In this condition, which can be caused by damage to the parieto-occipital cortex, moving objects do not appear to move smoothly and continuously but rather suddenly jump to a new location. Further, the condition is only associated with deficits in motion processing, without changes in acuity or localization abilities.

A number of fundamental psychophysical observations are critical for understanding the physiological work that will be described in subsequent sections. First, humans and primates are exquisitely sensitive at distinguishing the direction and, to a somewhat lesser degree, the speed of spatially or temporally adjacent motion stimuli. Intriguingly, the perceived speed of an object is dependent on many factors, including the contrast, size, and spatial frequency content of the stimulus, but also on how long the object has been viewed. The influence of viewing duration is also highlighted by the MAE, in which the percept of a stimulus depends on what had been viewed previously.

Psychophysical studies have shown that we can be 'fooled' by apparent motion. Thus a succession of short displacements is perceived as smooth motion, provided these displacements are less than 15–20 arc min. Watching television provides a clear demonstration of this: while objects appear to move smoothly across the screen, they are constrained spatially and temporally by the screen's pixels and refresh rate. In addition to short-range apparent motion, 'phi,' or long-range apparent motion, has been described, in which two flashes of light separated by much more than 20 arc min are interpreted as a single moving flash provided they have a suitable temporal separation. In addition, the long-range process can be distinguished from the short-range process by its sensitivity to shape and color, its failure to reverse directions for inverting contrast sequences (reverse phi), and its ability to integrate sequences across the two eyes (interocular transfer). These differences indicate the existence of multiple mechanisms in the visual system for detecting and analyzing motion.

The role of vision in guiding locomotion is evident from vection – the visually induced illusion of self-motion. This may be experienced when looking through the window of a stationary vehicle at another nearby stationary vehicle. If the other vehicle moves forward, one can experience a vivid sensation of moving backward. This arises because motion of the entire visual scene is generally attributable to eye, head, or body movements and is typically interpreted as such. This suggests that there are specialized mechanisms for detecting and interpreting optic flow fields characteristic of those generated during self-motion. These same mechanisms are also important for controlling

eye movements that serve to keep the image stable on the retina during self-motion. In the absence of such mechanisms, image quality would be severely degraded due to motion blur.

Anatomy

The basic anatomical organization of the visual system has been described as a processing hierarchy with extensive feed-forward and feedback connections between discrete areas (Figure 3). Motion processing is strongly associated with the dorsal stream, and is relatively segregated from the ventral form pathway as early as the retina and lateral geniculate nucleus (LGN). In primates, motion information is primarily carried by the so-called magnocellular stream, beginning with parasol retinal ganglion cells in the retina, which project to the more ventral, magnocellular layers of the LGN and subsequently to layers 4Cα and 6 of primary visual cortex (V1 but also known as 'striate cortex' or area 17). Subsequently, information reaches the middle temporal area (MT) predominantly through direct projections from layers 4B and 6 in V1, but also via indirect inputs from the thick stripes of V2 as well as V3. MT can be considered a motion-processing gateway in the cortex: it is the earliest of the visual areas in the hierarchy to contain predominantly motion-sensitive neurons, and the majority of these neurons are highly direction selective.

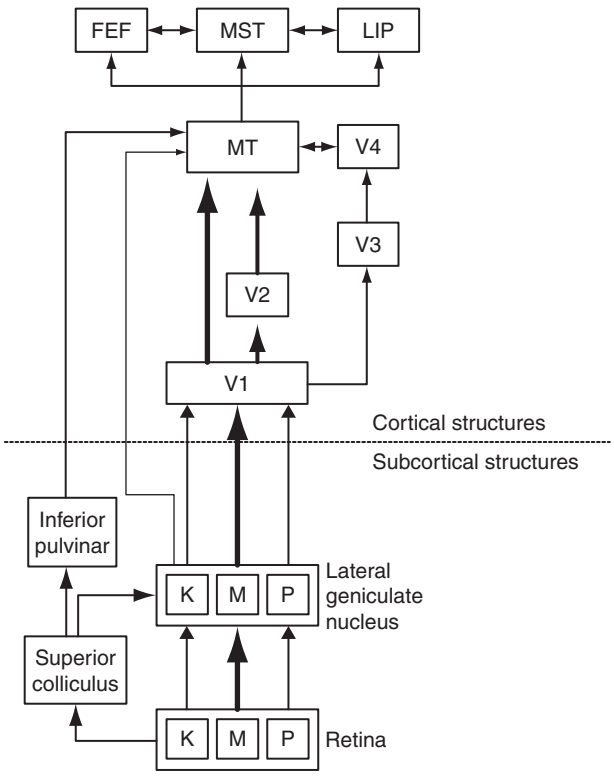

Figure 3 Major motion processing structures along the dorsal stream. Retinal information comprising konicellular (K), magnocellular (M), and parvocellular (P) processing channels originates in subsets of retinal ganglion cells and is passed to the superior colliculus and lateral geniculate nucleus. The primary stream for motion information is via the M cells in the retina and lateral geniculate nucleus through V1, then directly to the middle temporal area (MT). An indirect cortical stream also exists via V2, and much weaker inputs to MT exist via the superior colliculus and inferior pulvinar and directly from the K cells in the lateral geniculate nucleus. Information is also shared between MT and V4, and also reaches higher-order areas such as the frontal eye field (FEF), middle superior temporal area (MST), and lateral intraparietal area (LIP).

MT passes information to a number of additional cortical areas, including the dorsal and lateral compartments of the middle superior temporal area (MST), where neurons are sensitive to wide-field optic flow generated by body movements (MSTd) and pursuit eye movements (MSTl). Both MT and MST project to a range of higher-order cortical areas such as area 7a, the lateral intraparietal area (LIP), and frontal eye fields (FEFs) to guide decision making and eye movements; and to subcortical structures such as the accessory optic system, pontine nuclei, superior colliculus, and pulvinar to guide eye movements such as pursuit and optokinetic nystagmus. Although all the cortical and subcortical regions are highly interconnected, we will focus mainly on the properties of neurons in the feed-forward cortical hierarchy of the early dorsal stream, from the retina to MST. This provides a useful method for describing how the representation of movement changes from the retinal input channels sampling the visual world to neural codes more closely associated with perception.

Direction Selectivity

As early as the retina, neurons are found that prefer motion in one direction over another. Retinas in the turtle and rabbit have been studied more extensively than have those of other species, with up to one-third of all retinal ganglion cells (RGCs) in the turtle estimated to be direction selective, with similar proportions of directionally biased amacrine and bipolar cells. However, despite their presence in most vertebrate species, the role and prevalence of direction-tuned retinal cells in the primate are not clear, probably because of the large expansion of the midget (parvocellular) system in primates compared to other species. Furthermore, it appears that virtually all direction-selective RGCs in the primate retina project to brain stem structures, such as the nuclei of the accessory optic system, where they are involved in reflexes for gaze stabilization.

Both the proportion of direction-tuned neurons and the strength of their direction tuning (how much they prefer motion in the preferred direction over other directions) increase in areas along the dorsal stream. The direction selectivity underlying primate motion perception first arises in V1, where 20–30% of all cells are directionally biased. These cells are mainly confined to output layers 4B and 6, tend to respond to a narrow range of orientations, and have high contrast sensitivity and relatively small receptive fields. In V2, only ~15% of neurons are highly direction selective; however, such cells are prominent in layers 3B and 4 of the thick stripes, which receive significant projections from layer 4B in V1 and which subsequently project to MT.

In area MT, nearly all neurons are highly selective for the direction of motion. In addition, many neurons also show inhibition by motion in the null direction, a property which may give rise to the MAE. Another important feature of MT is that direction-selective neurons are organized into a map, with neurons that are tuned to the same direction of motion grouped together in columns, and adjacent columns representing different directions of motion. At subsequent processing stages, such as MST, essentially all cells are motion sensitive, but many prefer more complex types of motion rather than simple translation. In MSTd, approximately half of the cells are sensitive to translation, but many are sensitive to rotation in the fronto-parallel plane, rotation in depth, or expansion or contraction. These optic flow patterns are consistent with those produced by locomotion and, indeed, many MSTd neurons receive extraretinal inputs associated with locomotion. In contrast, neurons in MSTl are sensitive to motion produced by objects moving against a background. This requires a center-surround receptive field organization (see the section titled, 'Spatial aspects of receptive fields').

Temporal Frequency and Speed Tuning

Early quantitative studies of the visual system treated neurons as filters tuned along multiple stimulus dimensions, such as direction or speed. The full spatiotemporal frequency tuning of neurons can be tested using sine wave grating stimuli with a range of spatial frequencies (SFs) and moved with a range of temporal frequencies (TFs). Grating speed is then given by $v = \mathrm{TF/SF}$. A spatiotemporal frequency (STF) tuning plot can be generated from the responses to individual gratings (Figure 4). Simple cells in V1 are predominantly temporal frequency tuned, meaning that contours in their STF plots are oriented such that the TF giving the largest response is constant regardless of the grating's spatial frequency

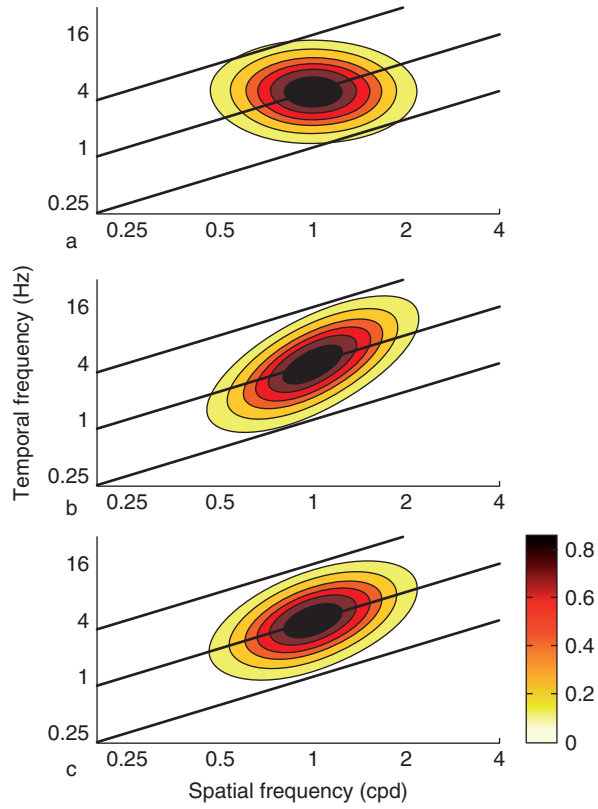

Figure 4 Spatiotemporal frequency (STF) tuning. STF plots are generated by plotting a neuron's mean spiking rate in response to moving gratings with a range of spatial and temporal frequencies. Temporal frequency (TF) tuning (a) implies that contours in the STF align along a single TF, regardless of the grating spatial frequency (SF). Speed tuning (c) implies that the STF contours align with an iso-speed line, such that the ratio of preferred TF to preferred SF is a constant. (b) STF plots can also show oriented STF contours that are indicative of neither TF nor speed tuning.

(Figure 4(a)). While one-fourth of V1 complex cells and MT cells also show temporal frequency tuning, a similar proportion show form-invariant speed tuning, meaning that contours in their STF plots are oriented along an iso-speed line (Figure 4(c)). The remaining half of V1 complex and MT cells show oriented STF plots that are not aligned with a constant TF or constant speed (Figure 4(b)). The major difference between complex cells and MT neurons is that MT neurons show nonlinear interactions when tested with stimuli containing multiple spatial frequencies. Thus speed-tuning bandwidths calculated using textured stimuli are smaller for MT neurons than for V1 complex cells, and these bandwidths are narrower than would be predicted based on linear predictions from the responses to single spatial frequencies.

Temporal frequency tuning and speed tuning confer different advantages. TF tuning allows a detector to reliably judge the motion of a given object, regardless of the viewing distance. Speed tuning allows a

detector to reliably judge the motion of an object, regardless of its form. Perceptually, speed judgments are strongly dependent on the nature of the stimulus, with stimuli having low SF perceived as moving faster than those with high SF. This effect is consistent with the observation of neurons with slanted STF plots that do not show true speed tuning.

Spatial Aspects of Receptive Fields

A neuron's motion sensitivity is dependent on the spatial structure of its receptive field – its size, shape, and subdomains. In V1, V2, and MT, receptive fields have a bias toward the central 5° of the visual field and tile the entire contralateral hemisphere; however, they do not extend more than 10–15° ipsilateral of the vertical meridian. In V1, foveal receptive field sizes are extremely small (<1° diameter) but increase with eccentricity. At a given eccentricity, receptive fields are approximately two to four times larger in V2 than in V1, and larger still in MT. This changes in MST, where the tiling of receptive fields (RFs) extends 20–40° into the ipsilateral field, and foveal RFs are 10–15° in diameter. Further, the relationship between eccentricity and RF size is much weaker in MST than in V1–MT.

Receptive field size limits the separation of a motion detector's inputs, thus affecting the neuron's spatial frequency tuning bandwidth and the maximum resolvable step size that can be detected as motion. However, this maximum step size does not systematically change along the cortical hierarchy. Thus, despite MT receptive fields being ~10 times larger than those in V1 at a given eccentricity, neurons in the two areas are sensitive to a similar range of step sizes, and these step sizes match those that have been determined psychophysically. This suggests that MT does not calculate direction selectivity *de novo*, but elaborates on the motion analysis performed by output neurons from V1. Perceptually, long-range apparent motion can be perceived for displacements larger than those that stimulate V1 and MT neurons. Although not explored physiologically, this may be achieved through the attentive tracking of stimulus features that change position over time. However, the perception of long-range apparent motion suggests that motion processing may occur independently of V1 and MT at higher cortical levels than those studied so far.

From the level of the retina to MST, neurons display 'extraclassical' receptive fields or modulatory surrounds. For example, MT neurons respond well to motion in their preferred direction presented only in a small region at the center of their receptive fields. If the stimulus is enlarged to cover a suppressive surround, the neuron ceases to respond to the motion. This type of center-surround opponency for motion is likely to contribute to the extraction of local motion contrast that is created when objects move against a textured background or during pursuit eye movements, which produce a stationary object on the fovea surrounded by background motion in the direction opposite the eye movement. Center-surround receptive fields can also help assign border ownership and may contribute to motion segmentation, by determining if local motion cues belong to the same or different objects.

The opposite problem, motion integration, requires combining multiple local motion signals to determine the true direction of a larger object. The aperture problem described previously (see the section titled 'Psychophysics') provides a good demonstration of this: only by assuming that two different local motions belong to the same object can the information be combined to decode the correct direction of the object (**Figure 2(b)**). It is currently unclear by what mechanisms the visual system solves this problem. Some models propose that the visual system first extracts only the local one-dimensional (1D) motion perpendicular to the orientation of moving edges. Subsequently, this local motion information is nonlinearly combined using an 'intersection of constraints' calculation. Indeed, there is psychophysical and physiological evidence for the existence of orientation-selective filtering at early stages of motion processing followed by pattern-sensitive neurons in MT that can signal motion directions that are not orthogonal to an object's orientation. Many of these results have come from studies employing plaid stimuli, which contain two overlapping gratings with different orientations. Under some circumstances, the stimulus appears to move coherently, so that the perceived direction of motion is not orthogonal to the orientation information present in the pattern.

However, recent studies have also suggested that an alternative strategy may be employed, in which intrinsically 2D visual features, such as line endings or vertices, are first identified through nonlinearities, such as end-stopping, as found in some V1 cells. Motion information about these features is then selectively combined to calculate object motion. It is quite likely that multiple strategies are available to the visual system, with different ones predominating depending on viewing conditions. This is consistent with the concept of the visual system as a 'bag of tricks,' in which multiple mechanisms have evolved to take advantage of different environmental cues.

In MT and MST, spatial subdomains or asymmetries are evident in receptive fields, possibly as a result of combining inputs from spatially separated

V1 neurons with different preferred speeds, directions, or disparities. In addition to the preference of MST neurons for complex motion such as rotation and expansion, asymmetries within MT receptive fields appear to create neurons tuned for spatial gradients in speed or disparity. Such receptive field substructure may contribute to the perception of 3D surfaces or motion in depth.

Temporal Aspects of Motion Processing

When presented with an unchanging motion stimulus, the responses of most neurons change over time. The most commonly described changes are attributable to neural adaptation, in which the neuron's firing rate declines approximately exponentially over time. Adaptation was originally attributed to a kind of 'neural fatigue,' such as might be produced by depletion of neurotransmitter, but this cannot be the sole mechanism since adaptation is observed even when stimuli are presented briefly or when they do not elicit large responses. Further, adaptation rates and levels are highly stimulus dependent, being influenced by contrast, spatial frequency, temporal frequency, speed, and direction. Interestingly, adaptation does not simply change the amplitude of a neuron's response to a specific stimulus, but it can temporarily change a neuron's tuning. For example, after prolonged exposure to a grating, the tuning curves of V1 neurons shift away from the adapting grating's TF, SF, and direction. Conversely, attractive shifts have been observed in MT neurons, in which the preferred direction of many neurons shifts toward the direction of the adapting motion.

Neural adaptation is a very general mechanism present in multiple levels of the visual system as well as in other sensory systems. For example, photoreceptors are faced with light intensity variations spanning over 10 orders of magnitude, but the information that a neuron employing a rate code can transmit is limited by its maximum spiking rate. By adapting to the mean luminance in a scene, photoreceptors can increase the range of stimuli to which they can respond and also signal only changes in light intensity, which are of greater behavioral relevance than is the absolute light intensity (Figure 5). Similarly, a motion-sensitive neuron can adapt to the mean speed of a stimulus, reducing redundancy in the information encoded by its spikes. Perceptually, this dynamic range optimization is evident in studies of speed perception. During prolonged viewing of motion, the perceived speed of the stimulus declines approximately exponentially with viewing duration. This is accompanied by an enhanced ability to detect changes in speed, indicative of the functional role of adaptation in increasing motion sensitivity. An interesting by-product of adaptation, in combination with the directional opponency described above, is that prolonged viewing of a constant direction of motion can produce illusory motion of stationary patterns, as in the MAE.

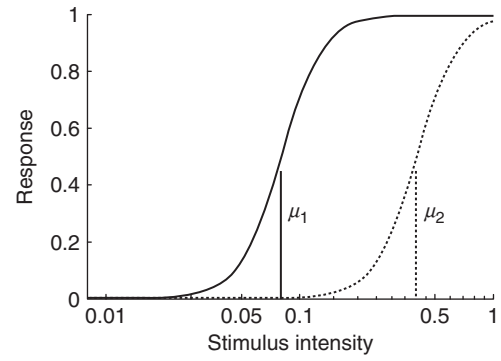

Figure 5 The functional role of adaptation. The sigmoid curves represent the normalized stimulus–response relationship for a hypothetical neuron. In a given adaptation state (e.g., with mean stimulus intensity μ_1) the neuron's spiking rate cannot cover the full range of stimulus intensities. If the mean stimulus intensity shifts to μ_2, then the neuron dynamically shifts its tuning curve so that it can better represent changes in stimulus parameters.

Changes over time in neural responses to a constant stimulus can also reflect ongoing processing rather than adaptation. A striking example of this is that over ~60 ms, the responses of MT neurons change from reflecting the component of motion perpendicular to the orientation of a moving line segment to reflecting its true direction of motion. Similarly, when presented with plaid stimuli, consisting of two superimposed gratings, the responses of pattern motion-sensitive neurons evolve from initially being component motion sensitive. This demonstrates the dynamic coding of visual stimuli and provides further evidence that motion information may be processed collaboratively between a range of visual areas.

Higher-Order Aspects of Motion Representation

In addition to the more basic neural representation of movement that encodes the direction, speed, and location of motion within a scene, the visual system performs an array of higher-order motion computations. Areas along the dorsal stream progressively incorporate a greater proportion of extraretinal information into their coding of the retinal signal. This takes two main forms: copies of motor commands and attentional signals. So-called efference copy of eye movements and body movements can be used to help distinguish self-motion from object motion, and it can also be used to prepare the visual system for sudden changes in the visual inputs, such as those accompanying saccades. In such situations the motor

signal may allow neurons to transiently shift their spatial receptive fields or to suppress processing of the rapid self-generated motion. Attentional modulation of neural signals typically results in increased responsiveness to stimuli in the in the region of visual space to which attention is directed, or to distant objects that share features of the attended object, such as its direction of motion. Attention can also change receptive field shapes and response timings. While the precise mechanisms underlying these phenomena are unclear, many of the effects can be attributed to simple changes in response gain. Since areas further along the dorsal stream have larger receptive fields, the effect of attention acting in lower cortical areas could be to provide the higher-order areas with a representation of the visual world that is relatively dominated by the attended stimuli.

Another higher order aspect of motion representation is the integration of information between the form and motion pathways. The three-dimensional structure of objects can be extracted from motion and parallax cues generated by either object motion or movements of the observer. For example, during locomotion, objects at different distances generate distinct patterns of optic flow, leading to discontinuities in the local motion field. This allows the visual system to segment local motion vectors at the discontinuity and integrate local motion vectors on either side of the discontinuity. The perception of biological motion, in which observers recognize a moving animal or human form using only sparse information about the positions of the joints over time, may represent the highest form of the integration of local motion cues. However, it is also possible that moving animal forms are recognized predominantly from multiple template-matching operations of the static form across time, and that motion *per se*, as discussed previously (see the section titled 'Psychophysics'), is irrelevant.

Conclusion

Over four decades of physiological investigation of motion processing and over a century of quantitative psychophysical studies have given us great insight into the representation of motion at the single-neuron level and at the perceptual level. However, it is still not clear how the responses of a population of neurons are combined to generate perception, or which brain areas are associated with the perception of specific stimuli. The most basic link between a brain region and perception is that the activity in the area is correlated with the strength of the percept. In addition, microstimulation studies and lesion studies help link anatomical regions with behavior and perception. A wealth of studies has led to the strong association of MT with the perception of stimulus speed, direction, and depth, and the association of area MST with heading detection and object tracking. However, it is still not clear which neurons in these areas contribute to perception, how different neurons may be recruited for different tasks, and how their responses are combined. Determining how motion is represented at the population level for guiding perception and behavior remains one of the great technical and analytical challenges for future studies of vision.

See also: Canal–Otolith Interactions; Motor Psychophysics; Neural Coding of Spatial Representations; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Spatial Transformations for Eye–Hand Coordination; Spatial Cognition; Spatial Cognitive Maps.

Further Reading

Allman JM, Miezin F, and McGuinness E (1985) Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local–global comparisons in visual neurons. *Annual Reviews Neuroscience* 8: 407–430.

Barlow HB and Levick WR (1965) The mechanism of directionally selective units in rabbit's retina. *Journal of Physiology (London)* 178(3): 477–504.

Born RT and Bradley DC (2005) Structure and function of visual area MT. *Annual Reviews Neuroscience* 28: 157–189.

Borst A and Egelhaaf M (1989) Principles of visual motion detection. *Trends in Neurosciences* 12(8): 297–306.

Braddick O (1993) Segmentation versus integration in visual motion processing. *Trends in Neuroscience* 16(7): 263–268.

Britten KH, Shadlen MN, Newsome WT, et al. (1992) The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience* 12(12): 4745–4765.

Clifford CW and Ibbotson MR (2002) Fundamental mechanisms of visual motion detection: Models, cells and functions. *Progress in Neurobiology* 68(6): 409–437.

Felleman DJ and VanEssen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1 (1): 1–47.

Hubel DH (1959) Single unit activity in striate cortex of unrestrained cats. *Journal of Physiology (London)* 147: 226–238.

Maunsell JH and Newsome WT (1987) Visual processing in monkey extrastriate cortex. *Annual Reviews Neuroscience* 10: 363–401.

Movshon JA, Adelson EH, Gizzi MS, et al. (1985) The analysis of moving visual patterns. In: Gattass R and Gross CG (eds.) *Pattern Recognition Mechanisms*, pp. 117–151. Berlin: Springer.

Nakayama K (1985) Biological image motion processing: A review. *Vision Research* 25(5): 625–660.

Pack CC, Livingstone MS, Duffy KR, et al. (2003) End-stopping and the aperture problem: Two-dimensional motion signals in macaque V1. *Neuron* 39(4): 671–680.

Parker AJ and Newsome WT (1998) Sense and the single neuron: Probing the physiology of perception. *Annual Reviews Neuroscience* 21: 227–277.

Reid RC, Soodak RE, and Shapley RM (1991) Directional selectivity and spatiotemporal structure of receptive fields of simple cells in cat striate cortex. *Journal of Neurophysiology* 66(2): 505–529.

Representation of Reward

B J Richmond, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Definition of Reward

Rewards are those conditions, events, or objects that cause approach behavior, or elicit feelings of pleasure, such as might happen with the accumulation of art or wealth. The value of the reward governs the intensity of the reaction, and, conversely, in experiments or observations the reward value is inferred from the intensity of the reaction. This latter inference is generally correct, but there are examples wherein the logic of arguments based on this inference breaks down, thus experimentalists and theoreticians studying reward-elicited effects must be alert to possible difficulties in modeling reward/response behavior.

Behavior and Behavioral Modeling

There are two classes of paradigms used to study the effects of reward. The first is classical conditioning. In classical conditioning a stimulus – say, a picture of Abraham Lincoln – is paired with an outcome occurring after a fixed interval, here a reward. After a few exposures to Abraham Lincoln's picture, responses that might have previously occurred just when the reward was delivered, such as licking, increased heart rate, or salivation, will begin to occur when the picture first appears. Thus, Lincoln's picture has become a conditioned stimulus (CS), predicting the later appearance of the unconditioned stimulus (US), which is the reward itself. This is a strong effect, so classical conditioning procedures provide a powerful means to study how learning takes place under the influence of rewards without any action on the part of the individual. All reward-related learning is probably driven, at least to some degree, through the mechanisms underlying classical conditioning.

The second class of behavior is instrumental; in this case the individual must perform an action to receive a reward. Rewards increase the likelihood of the activity that leads to the reward. Punishments, in contrast, lead to avoidance. These are the appetitive and aversive aspects of responses. Actions (except for simple reflexes) are generally driven by the desire to obtain a reward, and/or the effort to avoid punishment. Instrumental procedures require the animal in a laboratory study to interact with the environment – that is, it must do something such as press a lever, or must peck (pigeons), to obtain the reward or avoid punishment. In experiments with simple instrumental procedures such as lever pressing or pecking, the rate of responding is taken as a measure of the value of the action (the reward). If more presses or pecks are required, the time to begin pressing or pecking is related to the number of presses or pecks required to obtain the reward.

The degree of motivation used to obtain a reward depends on the value of the reward to the individual. The intrinsic value of any individual reward is a matter of biological need and individual preference. Thus, when the effects of different rewards are being compared, the individual reward preference must be assessed. Experimentally, relative reward values are frequently assessed using choice behavior – that is, experiments in which individuals choose between alternative actions, each of which gives rise to a different reward. The action chosen most often is taken to indicate the preferred reward. Through this procedure rewards can be ranked on a relative scale for the individual under study, thereby providing relative values of different rewards, different amounts of the same reward, or conditions that modify reward value.

The value of a reward can be modified by a number of factors. In general, larger amounts of reward are more valuable than smaller amounts. Rewards that are more distant in time are less valuable, and there are good descriptions of how rewards are discounted by delay to delivery (see later). Rewards can also be devalued by satiation for the particular reward. The first piece of chocolate is most appealing, and, for most people, the chocolate becomes less appealing as the succeeding pieces are eaten. Rewards that are more difficult to obtain are also devalued.

Through behavioral and theoretical analyses the concept of value has been generalized to cover the values of states that in themselves do not yield reward. These states are important to consider because so much of behavior depends on succeeding in intermediate conditions that do not inherently yield rewards. The values of these states are assessed on the same scale as for rewards, and, hence, the term 'value' does not apply only to the reward. Reward values are part of a more general concept of a value leading to action system.

Rewards are devalued or discounted by time: the value of a reward is decreased as the time to its delivery increases. This can be seen in both classical conditioning (Pavlovian) procedures and in instrumental tasks. Classical conditioning occurs when a stimulus is paired with an outcome, a reward, or punishment. The animals show an autonomic or involuntary response, such as change in heart rate or blood pressure, or change in

pupil diameter. Frequently licking is monitored as a sign of classical conditioning.

In choice experiments, individuals are offered a choice between a smaller immediately available reward and a larger, delayed reward. Experiments such as these were analyzed brilliantly by Herrnstein, who developed the 'matching law,' which says that when there is a choice, the proportion of choices will match the accumulated rewards or reinforcements in the two conditions – that is, $p_1/(p_1+p_2)=r_1/(r_1+r_2)$. This says that the probability, p, of a choice is equal to the probability with which the choice is reinforced, r. This has been shown to be generally true, but, with the increased interest in reward behavior in neuroscience and economics, the details of this relation are being examined closely.

Choice experiments have been generalized to find the relative values of the reward contingencies. In one particularly dramatic demonstration of the matching law, Oxford students stood at opposite sides of the Oxford arboretum pond, throwing bread into the two sides at different rates. The ducks in the pond distributed themselves in proportion to the amount of bread available at the two sites. Thus, not only are the ducks sensitive to the availability of reward, but the relation that could in principle be seen by testing one duck with reward probability can be seen immediately by observing a large population. The results of this experiment are interpreted in a game theoretic context. Game theory provides the mathematical framework for analyzing choices in competing situations. This particular example shows that efficacy of the so-called Nash equilibrium, the point at which competing options reach an optimal balance, so that the global gains and losses are maximized. It is consideration of these types of observations that is bringing economists, experimental psychologists, and neuroscientists together in the burgeoning study of what is being called neuroeconomics.

It should be possible to measure the equivalence between an immediate reward and a delayed, larger reward. In a large number of experiments, it has been shown that temporal discounting is well described by $V=R/(1+kD)$, where V is the state value, R is the reward (usually its size), D is the delay, and k is a constant that controls the rate of discounting with respect to time. This relation, with V directly related to the reward, and inversely related to a function of the delay, is both intuitive and amazingly accurate.

A major issue regarding reward value, or action-state values, is how the values of different states are learned and computed. In the 1970s Rescorla and Wagner, in working on this problem, developed a theoretical construct to describe how reward, and more generally state values, might be learned. A somewhat more sophisticated version of this model appeared in the work of Sutton and Barto, who were trying to solve a problem in machine learning. They wanted to find a general theoretical algorithm that would define a machine trying to learn a task from successes and failures. The general idea was to define a reward (i.e., success), and when the machine did something that provided some reward, that state would be reinforced. What they did was to optimize the total reward over time. The model, the temporal difference algorithm, is elegantly simple, and works well in machine learning. The value V at time t is described by $\delta(t)=r+\gamma V(t)-V(t-1)$, where r is the reward and γ is a discount factor. The change in value is $\Delta V(t)=\alpha\ \delta(t)$. This model also describes temporal discounting of rewards, as mentioned earlier, but it would have the reward discounted according to $1/e^{-kD}$. This exponential form has some theoretical advantages in that this form behaves better mathematically in the limit of infinite reward discounting. However, the hyperbolic form just described fits most behavioral data better. This difference, while important for theoretical modeling, is not critical here. The temporal difference model learns the values of states and simultaneously adjusts them to encourage behavior to change so as to maximize the average reward over time.

As animals and people repeatedly receive a particular reward, the value of the reward becomes less – that is, the reward is devalued. In monkey experiments, for example, this can be assessed in two steps. First, the monkey is offered choices between variously flavored food pellets. The outcome of this procedure is a ranking from most to least preferred flavor. Then, at some time, say 30 min, before some sessions, the monkey is fed a number of pellets of a usually preferred reward (perhaps as many as it will take). The monkey then can be tested either with foods directly, or it can be tasked to associate a stimulus with each of the two rewards. In this latter case, normal monkeys will select objects associated with the more preferred stimulus if there had been no devaluation procedure, or with the lower ranking stimulus if the reward had been sufficiently devalued. In another step, the experimenter can present different stimuli which, when touched or displaced, provide the monkey with a food pellet. This procedure allows the experimenter to assess whether the animal adjusts the choice of the associated stimulus under the influence of the devaluation. In other words, was the value of the stimulus or its associated action influenced by the devaluation procedure?

The value of reward is modified by the context in which it is presented. Kahnemann and Tversky showed that framing has a profound influence on

the choices that humans make. This work, which is about how reward values are modified, led to a Nobel Prize in economics. Two principles have been central to defining so-called rational behavior. The first is that reward values are transitive: if the value of A is greater than B, and B is greater than C, A should be greater than C. The other is equivalence: two conditions or states that have equivalent (not necessarily identical) outcomes should have equal value. However, Kahnemann and Tversky showed that humans, rather than showing rational behavior, are greatly influenced by the manner in which a choice is presented or framed. They showed that humans will consider two circumstances that have identical objective values as having different reward values. In one of their examples, they found that ticket-holders en route to the theater, but who lost their $10 theater ticket on the way, were less likely to buy a (replacement) ticket than were ticketless theatergoers who lost $10 in cash while en route to the theater to buy a ticket for the first time. Both conditions result in a $10 loss, but a majority of individuals will not replace the lost ticket, whereas a majority will still buy a $10 ticket after having lost $10 cash. It appears that the $10 left in the pocket has different values after the different types of loss, even though the actual values are identical.

In addition, Kahnemann and Tversky showed that by framing a problem in which the reward is stochastic in different ways again influences the choices that are made. This work has provoked a lot of study of risk aversion, and its distribution in the population. It is now clear that there is a wide spectrum of risk taking, with some individuals trying to avoid risk in almost every way, many who will take some risk, and a few for whom risk seems to have a rewarding effect. This latter group probably includes pathological gamblers and those who seem to be rewarded by engaging in activities with large physical risk.

Brain Mechanisms for Reward Evaluation

Over the past 50 years quite a bit has been learned about which brain mechanisms play a role in reward evaluation. As yet, however, there is no good overall model for how reward evaluation is carried out in the brain. The discovery that rats with electrodes placed in the hypothalamus would press a lever seemingly forever, to deliver electrical stimulation, gave great impetus to studies of the circuitry and mechanisms underlying reward-seeking behavior.

Brain regions that are involved in assessing and acting on rewards include the limbic system, the frontal and posterior parietal cortices, the medial temporal cortex, the cingulate cortex, the anterior insula, and the basal ganglia, especially many parts of the striatum. Behavioral evidence comes from observing animals with selective inactivations through ablation or other means, and humans with damage from disease or injury. Recently, functional imaging in human subjects has been making large contributions by identifying brain regions that show changes in blood flow (activations) during reward-related conditions. Finally, electrical neuronal recordings reveal signals carried by single neurons or populations related to reward evaluation.

The transmitter dopamine plays a central role in current thinking about reward, and by extension, all of those brain areas that receive dopaminergic input are under consideration. At this point, it is widely agreed that dopamine and dopamine neurons play some important role in reward learning and, for this discussion, reward value. However, it is obvious from all of the motor symptoms of Parkinson's disease, a human condition in which the dopamine neurons inexplicably die over a number of years, that dopamine must play roles in the brain other than just coding for some aspect of reward. Nonetheless, the evidence that dopamine plays an important role in reward-related behavior is compelling. Thus, the basal ganglia, which are extremely rich in dopamine, and other areas, such as the frontal cortex, amygdala, the medial part of the temporal lobe, and the posterior parietal cortex, have all been implicated in reward-seeking behavior. In the primate, where dopamine is found in at some level throughout virtually all of the cortex, including primary sensory areas, it might even turn out that reward-related learning occurs, to some degree, just about everywhere.

Dopamine neurons are located in the high midbrain in the substantia nigra pars compacta, and in the ventral tegmental area. These neurons have a slow, regularly irregular pattern of firing, and they emit a couple of extra spikes in rapid succession when a surprising event occurs, when a reward occurs, or when a stimulus predicting a reward a short time later occurs. When the effects of dopamine are decreased by depletion or receptor blockade, the ability to associate reward with predictive stimuli is impaired, showing that dopamine plays a critical role in this type of associative learning. Thus, it appears that dopamine is critical for learning the values of the associations or the states that they signify.

A strong current hypothesis about dopamine neurons is that they might carry a signal related to reward prediction error – that is, a signal related to the difference between what reward is expected and what is actually delivered, or not. This hypothesis grew from the observation that when a stimulus is paired with a reward occurring a few seconds later in a

classical conditioning experiment, the response of the dopamine neurons, which had been locked to the reward, 'moves' to the time when the stimulus-predicting reward occurs. This reward-prediction error hypothesis is very intriguing, for if it were correct, it would suggest that the brain is computing an explicit signal, like the temporal difference signal used in the actor–critic temporal difference model of Sutton and Barto. The work and debate about the exact nature of the signal carried by dopamine neurons continues, but the debate about the hypothesis has been leading to exciting research at many levels of biophysics, single-neuron recording, and human imaging.

Most of the brain regions that play important roles in reward, in addition to their complex interconnections, are generally ones that are rich in dopamine fibers and receptors. These areas have dense interconnections that certainly are important for the signal processing underlying reward evaluation. However, the roles of the connections are poorly understood. For ease, rather than trying to elucidate all of the possible interactions, the discussion here now focuses on a partial list of important findings, relying on the presence of dopamine mechanisms as a common feature.

The basal ganglia, especially the striatum, have a large role in reward-related behavior. It has been proposed that the striatum implements something akin to the actor–critic components of the temporal difference model. From lesion data on learning of reward values, a case can be made that the ventral striatum acts like a critic, evaluating whether a response leads to the predicted outcome, and the dorsal striatum acts like an actor, guiding the actions. Whether or not this model is exactly correct, it offers a clear interpretation for several ablation experiments and provides an important framework for further experiments.

Several lines of evidence suggest that orbitofrontal cortex is critical for evaluating reward values. Experimental ablations of orbitofrontal cortex impair judgment of reward value and reward contingency. Ablations of the amygdala also damage the judgment of reward value, but not of reward contingency; that is, ablations of both orbitofrontal cortex in monkeys reduces or eliminates reward devaluation, but object reversal learning is only damaged after orbitofrontal lesions. Single neurons in orbitofrontal cortex seem to encode information about reward desirability or value. Human imaging shows activations or orbitofrontal cortex when the individuals are engaged tasks requiring planning and judging the value of a stimulus or situation as it pertains to the value of actions or rewards related to outcome. Thus, orbitofrontal cortex seems to be critical for the computation or representation of actual current value.

The lateral prefrontal cortex has dopamine receptors throughout it. Lesions of this tissue interfere with learning or changing choices in response to new rules. Physiological studies including single-neuron recording and imaging show activations related to upcoming choice of action, so rather than computing value or reward, this tissue appears to be critical for using the value to guide an upcoming action.

The anterior cingulate cortex is activated in humans when they are engaged in risk-taking tasks, and in tasks when they must monitor their own activity, especially the successes and the errors, as a means of improving performance. Behavioral experiments suggest that the cingulate is important for evaluating the trade-off between an immediate reward and a delayed, but bigger, reward. Normal rats might be willing to climb a big barrier to get a large reward even though a smaller reward is available by surmounting a smaller barrier; after cingulate ablation the rats choose the less demanding route for the smaller reward. In single-neuron recordings in monkeys, the neurons in the anterior cingulate cortex are sensitive to reward contingency and the neurons in the posterior cingulate are sensitive to different reward sizes and respond during the preparation to switch motor activity so as to increase reward, a function that seems once again similar to the long-time goals that form the base of the theoretical treatments of reward (e.g., the temporal difference models).

Recent work has shown activation of the insula and ventral striatum during some risk-taking tasks. From earlier work in animals, the insula has been related to gustatory and/or taste activation. The posterior part is more connected with cortical motor areas. The insula seems well positioned to play a role in reward evaluation, especially with regard to actual reward delivery.

Conclusions

Motivated behavior arises through expectation of reward or avoidance of punishment. Many factors contribute to the computation of the current value of a particular state. These include whether a reward will be delivered, how soon and how desirable a reward that might be delivered will be, and how aversive the effort to gain the reward is likely to be. Over the past 30 years some theoretical constructs have been developed that provide structures for organizing and evaluating results from behavioral, pharmacological, and physiological experiments. There seem to be connections between these biological experiments and the factors that these theories demonstrate are needed to influence motivation. The next few

years will provide the data to refine the theories and help choose among them.

See also: Connectivity of Primate Reward Centers; Dopamine: Cellular Actions; Electrical Self-Stimulation; Goal-Directed Behavior Theories; Perceptual Learning and Sensory Plasticity; Prefrontal Contributions to Reward Encoding; Psychopharmacology of Reward and Appetite in Rats; Reward and Learning; Reward Systems: Human; Reward Decision-Making; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex; Reward Processing: Human Imaging; Social Interaction Effects on Reward and Cognitive Abilities in Monkeys; Transcription and Reward Systems.

Further Reading

Dickinson A (1980) *Contemporary Learning Theory*. Cambridge, UK: Cambridge University Press.

Gintis H (2000) *Game Theory Evolving*. Princeton, NJ: Princeton University Press.

Glimcher P (2002) Decisions, decisions, decisions: Choosing a biological science of choice. *Neuron* 36: 323–332.

Herrnstein RJ (1997) *The Matching Law*. New York: Russell Sage Foundation.

Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology* 57: 87–115.

von Neumann J and Morgenstern O (1944) *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Wise RA (2004) Dopamine, learning and motivation. *Nature Reviews Neuroscience* 5: 483–494.

Respiration

G S Mitchell and T L Baker-Herman, University of Wisconsin–Madison, Madison, WI, USA
D R McCrimmon, Northwestern University, Chicago, IL, USA
J L Feldman, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

An essential regulatory function of the nervous system is to maintain homeostasis of respiratory gases. The neural system regulating breathing must coordinate rhythmic contractions of the respiratory muscles to generate appropriate lung ventilation for oxygen and carbon dioxide exchange. Nonhomeostatic functions of the respiratory system must also be served, including speech, sniffing, behavioral displays, and airway defense. This system is functional at birth and adapts continuously throughout life, maintaining homeostasis despite changing environmental conditions, activity levels, and physical properties of the lung/chest wall system during development, exercise, pregnancy, aging, and disease. In this article, we provide a brief introduction to features of the respiratory control system that enable its regulatory functions, followed by more extensive discussion of chemoreceptors, respiratory plasticity, ventilatory control disorders that challenge homeostasis, and integrated ventilatory responses to physiological stimuli.

Fundamental Elements of Ventilatory Control

Breathing Is Automatic and Not Autonomic

The control of breathing is an automatic process that works without conscious intervention when asleep, anesthetized, or awake and not specifically thinking about breathing. Conscious factors can override or modify automatic functions of the respiratory control system for a limited period. For example, an individual can voluntarily speak, smell, hyperventilate, or hold their breath. However, automatic functions ultimately mandate a return to normal breathing. Because the control of breathing is largely automatic and the regulation of breathing is intimately associated with autonomic functions, the respiratory control system is often confused as part of the autonomic nervous system. However, ventilatory control is in many respects more similar to systems controlling somatic motor functions, such as walking.

Ventilatory Control Processes

Figure 1 is a schematic of neural processes involved in the control of breathing. Breathing rhythm begins in the medulla. Supramedullary brain structures, including the pons, cerebellum, and forebrain, modulate breathing, but breathing continues without them. The ventral respiratory column (VRC) is a long, bilateral column of respiratory-related neurons in the ventrolateral medulla; the VRC contains functionally distinct neuron populations that are active during different phases of the respiratory cycle. Recent hypotheses suggest that two neuron clusters of the VRC underlying respiratory rhythm generation – the pre-Bötzinger complex (preBötC) and the retrotrapezoid nucleus/parafacial respiratory group (RTN/pFRG) – represent interactive (functionally coupled) respiratory rhythm generators that collectively establish breathing rhythm.

VRC neurons also craft the spatiotemporal details of a breath, a process known as respiratory pattern formation. VRC inspiratory and expiratory premotoneurons integrate a variety of CNS inputs, and differentially distribute respiratory synaptic drive to inspiratory and expiratory motoneurons innervating respiratory muscles of the thorax, abdomen, and upper airway. Phrenic motoneurons in the cervical spinal cord innervate the diaphragm and drive inspiration. During quiet breathing, the diaphragm actively contracts during inspiration and relaxes during expiration, which is passive in most mammals. During elevated respiratory drive or lung disease, accessory inspiratory muscles (e.g., external intercostals) and expiratory muscles (e.g., internal intercostals, abdominal muscles) contribute more to ventilation. Brain stem motoneurons innervating the upper airway muscles (e.g., hypoglossal) preserve airway patency during breathing. During each inspiratory effort, synaptic inputs activate hypoglossal motoneurons, flattening and stiffening the tongue to prevent collapse into the airway. Respiratory motoneurons integrate multiple synaptic inputs and exhibit neuroplasticity (see later), thereby contributing to respiratory pattern formation.

Rhythm generation and pattern formation are modified by feedback, feed-forward, and adaptive control mechanisms that confer flexibility and plasticity to this critical regulatory system. Sensory receptors provide feedback concerning chemical (O_2 and CO_2) or mechanical conditions (e.g., lung inflation) and initiate reflexes to regulate these variables within reasonable limits. More rapid responses are possible via feed-forward regulation. Longer lasting or repeated disturbances require long-lasting adaptations, or plasticity. Modulatory neurons modify the intrinsic respiratory

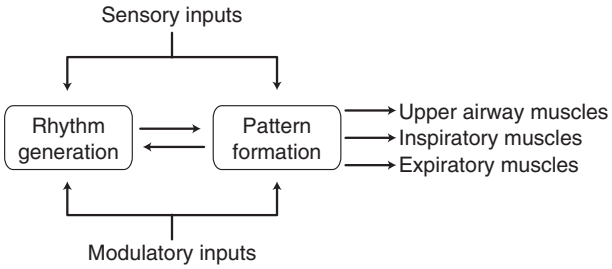

Figure 1 Schematic representation of neural processes fundamental to the control of breathing. Respiratory rhythm generation establishes the timing of a breath (i.e., breathing frequency) and originates in specialized neurons of the medullary ventral respiratory column, including the pre-Bötzinger complex and the retrotrapezoid nucleus/parafacial respiratory group. Respiratory pattern formation establishes the detailed spatiotemporal outflow of respiratory motoneuron pools that innervate different respiratory muscles. Neurons participating in pattern formation are diffuse and arise from rhythm-generating neurons, inspiratory and expiratory premotoneurons in the ventral respiratory column, and respiratory motoneurons of the brain stem and spinal cord. The intrinsic respiratory rhythm and burst pattern are modified extensively by sensory inputs (i.e., chemoreceptors and mechanoreceptors) and modulatory inputs, such as monoaminergic or peptidergic inputs from multiple sources. Other modulatory inputs arise from cortical processes, and cardiovascular and pain networks. The respiratory control system exhibits the capacity for plasticity. Plasticity can occur at multiple neuroanatomical sites in this system. To date, plasticity has been studied most extensively in chemoreceptors and in respiratory motoneurons. The motor output is distributed to respiratory pump muscles of the thorax and abdomen as well as the upper airway muscles that regulate airway resistance and stability.

rhythm and pattern. For example, brain stem serotonergic and noradrenergic neurons have been implicated in central chemoreception and/or the initiation and orchestration of respiratory plasticity. Modulation and plasticity can occur at multiple points in the respiratory control system, including sensory receptors, feed-forward neural pathways, VRC neurons, and respiratory motoneurons.

Chemoreceptors and Chemoreflexes

Two essential goals of the respiratory system are to obtain adequate O_2 into the body for aerobic metabolism, and to eliminate the resultant CO_2. As CO_2 is a major determinant of blood and tissue pH, it is tightly regulated to maintain short-term acid/base homeostasis (the kidneys contribute to longer term pH homeostasis). Arterial CO_2 is the primary regulated variable in ventilatory control under normal conditions. For example, when arterial CO_2 pressure (normally ~40 mmHg) is elevated by only 1 to 3 mmHg, ventilation doubles. In contrast, arterial O_2 pressure (normally ~100 mmHg) changes up to 20 mmHg are virtually ignored by the system. Oxygen-sensitive chemoreceptors stimulate breathing acutely only when O_2 partial pressure falls by 30 mmHg or more. However, over longer periods (hours or days), O_2 is more effective at triggering respiratory plasticity (see later).

Sources of Chemoreceptor Feedback

Chemoreceptors provide sensory feedback concerning CO_2 or O_2 changes and initiate chemoreflexes, adjusting breathing to minimize deviations in blood gas levels via negative feedback. In mammals, O_2-sensitive chemoreceptors are found in the peripheral nervous system, whereas CO_2-sensitive chemoreceptors are found both peripherally and in the central nervous system (CNS). CO_2-sensitive chemoreceptors in the lungs of birds and reptiles also play a prominent role in ventilatory control, but similar receptors are not found in mammals.

Central Chemoreceptors

The predominant CO_2-sensitive receptors controlling breathing in mammals are the central chemoreceptors, but there is considerable debate regarding their location and cellular identity. Central chemoreceptors sense intracellular and/or extracellular brain pH, indirectly sensing changes in CO_2. Neurons in multiple brain stem regions show CO_2/pH sensitivity (Figure 2), including (1) retrotrapezoid nucleus (RTN) ventral to the medullary facial nucleus, (2) serotonergic and γ-aminobutyric acid (GABA)ergic neurons in the raphe nuclei, (3) noradrenergic neurons of the locus coeruleus, (4) nucleus of the solitary tract (NTS), (5) preBötC, and (6) cerebellar fastigial nucleus. Additional CO_2-sensitive neurons are found in disparate regions, including the caudal hypothalamus and the hippocampus. The relative sensitivity of these neuronal populations to CO_2/pH *in vivo,* and the contribution they make to the hypercapnic ventilatory response are topics of active debate.

One hypothesis is that a single group of central chemoreceptors in a discrete location with highly specialized cells detecting slight changes in CO_2/pH and eliciting robust ventilatory responses dominates the CO_2 chemoreflex; this hypothesis implies that other CO_2-sensitive neurons play different roles related to arousal or behavioral responses. The RTN is one candidate site for a prominent role in the CO_2 chemoreflex; it contains neurons with high CO_2/pH sensitivity with strong projections to the VRC. An alternate hypothesis for the bewildering array of central chemosensitive sites is that different cell groups assume more

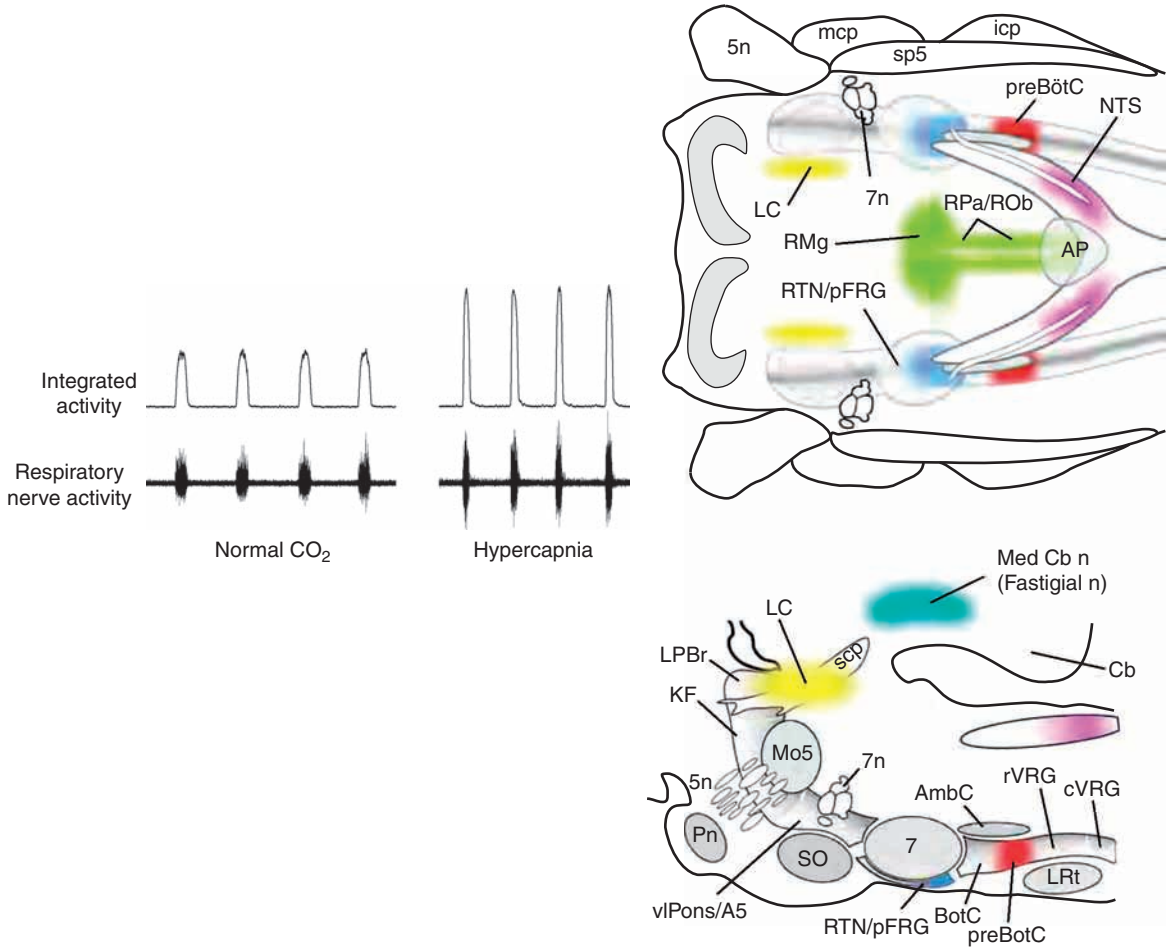

Figure 2 Central chemoreceptors are distributed throughout the brain stem and modify respiratory motor output. Left: Phrenic neurogram and integrated phrenic nerve activity during normal and hypercapnic conditions in an anesthetized, vagotomized, and paralyzed rat. Increased CO_2 elevates respiratory drive, largely due to actions of central chemoreceptors. Right: Horizontal (upper) and sagittal (lower) diagrams of rhombencephalon of the rat, showing the location of different regions where chemosensory neurons have been identified. The ventral respiratory column is the gray shaded area of the ventrolateral medulla. In the nucleus of the solitary tract (NTS; purple), the distribution of chemosensitive neurons has not been accurately determined, but includes portions of the ventral and lateral NTS coextensive with the area postrema. In the retrotrapezoid nucleus (RTN; darker blue), chemosensitive neurons are concentrated under the caudal third of the facial nucleus. The caudal raphe nuclei (RMg/RPa/Rob; green) extend from the rostral medulla to the caudal pons and provide inputs to the brain stem and spinal cord. Raphe magnus (RMg) is the main source of medullary serotonergic projections. Other CO_2-sensitive neurons are found in the pre-Bötzinger complex (preBötC; red), the pontine locus coeruleus (LC; yellow), and the cerebellar fastigial nucleus (Fastigial n; turquoise) of the cerebellum. Additional CO_2-sensitive neurons are found in the caudal hypothalamus and hippocampus (not shown), although these regions are not normally thought to be involved in respiratory motor control. The relative role of each chemosensitive site in the hypercapnic ventilatory response is under active debate. 5n, trigeminal nerve; 7, facial nucleus; 7n, facial nerve; A5, lateral pontine A5 noradrenergic neuronal group; AmbC, compact part of nucleus ambiguus; AP, area postrema; BötC, Bötzinger complex; Cb, cerebellum; cVRG, caudal division of ventral respiratory group; icp, inferior cerebellar peduncle; KF, Kölliker–Fuse nucleus; LPBr, lateral parabrachial region; LRt, lateral reticular nucleus; mcp, medial cerebellar peduncle; Med Cb n., medial cerebellar nucleus (also called fastigial nucleus); Mo5, motor nucleus of the trigeminal nerve; pFRG, parafacial respiratory group; Pn, basilar pontine nuclei; Rob, raphe obscurus; RPa, raphe pallidus; rVRG, rostral division of ventral respiratory group; scp, superior cerebellar peduncle; SO, superior olive; sp5, spinal trigeminal tract; vlpons, ventrolateral pontine region. Diagrams provided courtesy of Dr. G. Alheid.

prominent roles in different physiological states, such as during sleep versus during wakefulness. Alternately, central chemoreflexes may reflect the collective input of many neuronal groups, each contributing partially to the overall ventilatory response. Such a distributed chemoreceptor system would more accurately represent average brain CO_2/pH levels versus the limited spatial domain of a single site.

Peripheral Chemoreceptors

In adult mammals, the most important O_2-sensitive chemoreceptors are the carotid body chemoreceptors, located at the bifurcation of the internal and external carotid arteries. Carotid body chemoreceptors also sense changes in arterial CO_2 and pH. Although the mechanism of carotid chemosensory transduction is largely unknown, recent evidence implicates ATP acting through purinergic P2 receptors (particularly P2X2). Carotid body chemoreceptors relay signals concerning arterial O_2 via the carotid sinus nerve to the nucleus of the solitary tract (NTS), where second-order projections diverge to many CNS regions, including the VRC, RTN, cerebellum, and cortex. When increased chemoafferent neuron activity stimulates breathing, the ventilatory response is complicated by feedback interactions with CO_2 chemoreceptors. For example, increased ventilation during hypoxia minimizes the arterial O_2 decrease, but concurrently decreases arterial CO_2. Consequently, CO_2 chemoreceptors are inhibited and restrain the hypoxia-induced increase in breathing.

O_2-sensitive receptors are also found in the aortic arch (aortic body), but these receptors appear to play little role in ventilation under normal conditions; their greater contribution is in cardiovascular regulation. However, aortic body chemoreceptors increase their contribution to ventilatory chemoreflexes if the carotid bodies are removed, particularly during development. Additional O_2-sensitive neurons are found in the CNS, but their role in ventilatory control is not yet clear.

Integrated Responses

Time Domains of the Hypoxic Ventilatory Response

The hypoxic response is a complex interplay amongst multiple neural mechanisms that differ in their direction (facilitation vs. depression), time domain (seconds to years), and expression during versus following hypoxic exposures (see Figure 3). The acute response during hypoxia is an immediate (within one breath) increase in tidal volume and frequency due to the immediate chemoreflex effects of peripheral chemoreceptor stimulation. Following the acute response, tidal volume gradually increases over several minutes, an effect known as short-term potentiation (STP). On termination of hypoxia, the reversal of STP is observed as a slow return to normal tidal volume. The mechanism of STP is not yet clear but may involve presynaptic calcium accumulation in NTS relay neurons or premotoneurons.

Inhibitory mechanisms are also apparent in the hypoxic ventilatory response, such as hypoxic ventilatory decline (HVD) and posthypoxia frequency decline (PHFD). HVD is the 'roll off' in ventilation (primarily tidal volume) when moderate hypoxia is sustained for many minutes at constant CO_2. HVD may result from decreased peripheral chemoreceptor sensitivity and/or inhibitory mechanisms in NTS relay neurons. PHFD is observed following acute hypoxia, and is a transient

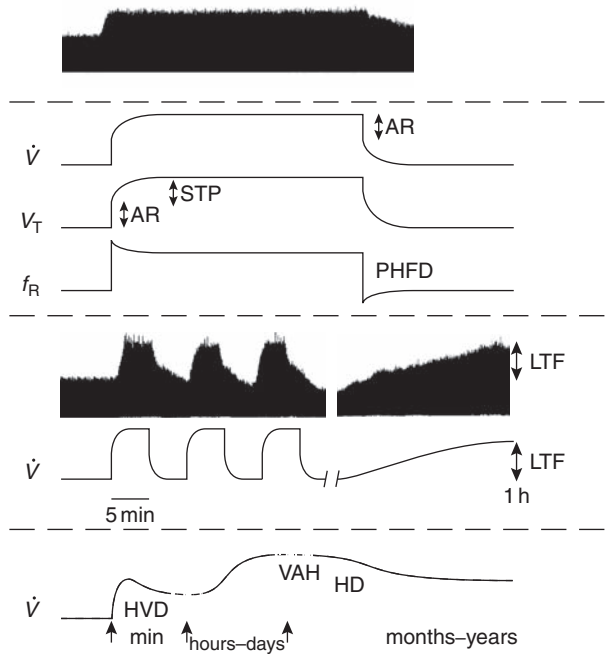

Figure 3 Time-domains of integrated ventilatory responses during and following hypoxia. Top: Integrated phrenic motor output in an anesthetized rat challenged with a single episode of 11% O_2 (5 min). The time scale is compressed, merging successive breaths to highlight changes in the amplitude of respiratory motor output. Below the experimental trace are schematized representations of the time-dependent changes in ventilation ($\dot{V}$) and its components – tidal volume (V_T) and frequency (f_R) – during this short-term hypoxic response. The immediate increase and decrease in ventilation at the beginning and end of the hypoxic episode represent the acute hypoxic response (AR). The AR is followed by a relatively slow and progressive increase in VT, an effect known as short-term potentiation (STP). The overshoot and undershoot in respiratory frequency at the beginning and end of hypoxia, respectively, are forms of short-term depression; the off-response is referred to as the posthypoxia frequency decline (PHFD). When hypoxia is presented in an intermittent pattern (third panel from top; three hypoxic episodes, 5 min in duration, 5-min intervals), long-term facilitation (LTF) is revealed as a slow, progressive increase in respiratory amplitude lasting for several hours posthypoxia. LTF is not revealed by a single hypoxic episode of similar cumulative duration (25 min). When hypoxia is sustained for many hours to years (lower trace), a series of ventilatory responses are observed in sequence, including hypoxic ventilatory decline (HVD; 15–60 min), ventilatory acclimatization to chronic hypoxia (VAH; hours to days), and hypoxic desensitization (HD; months or years). Each time domain of the hypoxic ventilatory response arises from unique mechanisms. See text for further explanation.

(minutes) decrease in respiratory frequency below prehypoxic levels. Although the mechanism of PHFD is not clear, it requires neurons in the ventrolateral pons, and is modulated by noradrenergic (α-2) and serotonergic (5-hydroxytryptamine; 5-HT_{2A}) receptor activation. Prolonged hypoxia (days) and intermittent hypoxia induce additional mechanisms representing forms of respiratory plasticity (reviewed later).

Exercise Ventilatory Response: Feed-forward Control

The ventilatory response to mild or moderate physical activity (e.g., walking) is the most frequent and largest ventilatory adjustment during normal life. To preserve arterial CO_2 (and O_2) homeostasis, ventilation must increase in proportion to increased metabolism. Just how that ventilatory response is achieved is one of the fundamental, yet unresolved, questions of respiratory neurobiology. Most mammalian species increase breathing prior to any perceptible arterial CO_2 change (e.g., prior to or during the first step). Indeed, the steady-state ventilatory response to exercise is typically in excess of increased metabolic CO_2 production, slightly decreasing arterial CO_2 from rest to exercise. Thus, the predominant drive to breathe during modest exercise is feed-forward with respect to arterial CO_2 regulation. Chemofeedback provides inhibitory constraints on ventilation in most mammals, preventing excessive decreases in arterial CO_2 during exercise. In contrast to most other terrestrial vertebrates, the CO_2 change during exercise in humans is a function of exercise intensity. During mild exercise, the feed-forward response is slightly inadequate, and arterial CO_2 increases slightly; during moderate exercise, the feed-forward response precisely matches increased metabolic rate, and arterial CO_2 is regulated at its resting level. Speculations concerning the feed-forward exercise stimulus include parallel activation of respiratory and somatic motor output ('cortical irradiation'), or an unidentified receptor in exercising muscles that couples breathing with increased metabolic rate (Figure 4).

Adaptive Control Strategies: Respiratory Plasticity

With repeated or long-lasting perturbations, the neural system controlling breathing exhibits plasticity, an alteration in behavior based on experience. Respiratory plasticity enables adjustments to longer lasting physiological changes, such as ascent to altitude, normal development and aging, changes in weight, pregnancy, etc. Examples of plasticity in respiratory motor control include (1) plasticity induced by hypoxia or hypercapnia, (2) developmental plasticity, (3) modulation and plasticity of the exercise ventilatory response, (4) respiratory plasticity following neural injury, and (5) classical conditioning (e.g., sensitization of gill-withdrawal reflex in *Aplysia*). Hypoxia-induced respiratory plasticity is the most thoroughly studied and best understood form of respiratory plasticity. The specific mechanisms of hypoxia-induced plasticity depend on the pattern, severity, and duration of hypoxic exposure.

Respiratory Long-Term Facilitation following Acute Intermittent Hypoxia

Repetitive hypoxic episodes (seconds to minutes) elicit respiratory long-term facilitation (LTF), a progressive increase in respiratory motor output lasting several hours posthypoxia (Figures 3 and 5). LTF is primarily expressed as increased respiratory amplitude, with small and inconsistent effects on frequency. Motoneurons innervating respiratory pump muscles (e.g., phrenic and inspiratory intercostals) and upper airway muscles (e.g., hypoglossal) exhibit LTF following intermittent hypoxia. LTF is pattern sensitive, since it is observed following intermittent, but not sustained, hypoxia. LTF is frequently studied in respiratory nerve recordings in anesthetized rodent preparations, and ventilation in unanesthetized mammals. Ventilatory LTF in unanesthetized animals is state dependent, since it is more prominent in sleep versus wakefulness.

The cellular/synaptic mechanisms of LTF are under active investigation. Here, we limit discussion to phrenic LTF (pLTF), although similar mechanisms are proposed in other respiratory motoneuron pools. Our working model of pLTF is that intermittent hypoxia triggers intermittent serotonin release onto phrenic motoneurons, thereby activating 5-HT_{2A} receptors and initiating pLTF. 5-HT_{2A} receptor activation stimulates new protein synthesis, including new synthesis of brain-derived neurotrophic factor (BDNF) from existing mRNA. New BDNF synthesis is necessary for pLTF, since translational inhibition of BDNF mRNA via RNA interference with small interfering RNAs abolishes pLTF. Further, BDNF applied to the cervical spinal cord is sufficient to elicit phrenic motor facilitation with a magnitude and time course similar to pLTF following intermittent hypoxia. BDNF activates the receptor tyrosine kinase, TrkB, thereby activating downstream signaling molecules such as extracellular signal-regulated mitogen-activated protein kinases (ERK-1/2) and protein kinase B (Akt). These kinases may phosphorylate glutamate receptor subunits, increasing glutamate-induced currents and strengthening the synapse between medullary premotoneurons and phrenic motoneurons.

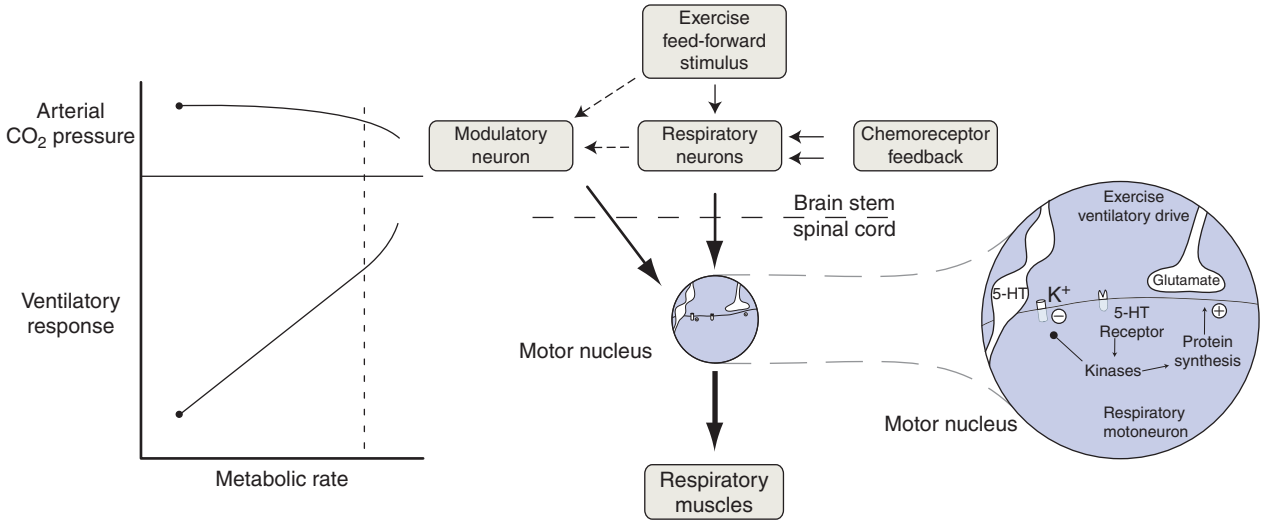

Figure 4 Modulation and plasticity of the exercise ventilatory response during mild or moderate physical activity in a nonhuman mammalian species. Left: Ventilation increases linearly with increased metabolic rate during mild to moderate physical activity; however, the ventilatory response is greater than necessary to maintain homeostasis, causing a slight decrease in arterial CO_2 pressure during exercise. During heavy exercise, the ventilatory response increases disproportionately, accelerating the decrease in CO_2. Middle: The primary drive to breathe during exercise results from a feed-forward mechanism, and is restrained by inhibitory chemoreceptor feedback to prevent excessive hypocapnia during exercise. Although the exercise ventilatory response is isocapnic or slightly hypercapnic in humans, these differences appear to be quantitative versus qualitative. The net result in mammals is good (not perfect) regulation of arterial CO_2 during exercise. When conditions are altered, such as with increased respiratory dead space, modulatory neurons (such as raphe serotonergic neurons) are activated, increasing respiratory motoneuron excitability and augmenting the exercise ventilatory response by a mechanism referred to as short-term modulation. Short-term modulation may reflect serotonergic modulation of potassium channels on respiratory motoneurons (detailed in the enlargement of the blue-shaded motoneuron). When exercise and increased dead space are explicitly paired over many exercise trials, long-term modulation, a form of plasticity, causes persistent augmentation of the exercise ventilatory response, possibly by synthesizing new proteins that increase the synaptic strength of inputs to respiratory motoneurons. Since long-term modulation is also serotonin dependent, it may reflect synaptic enhancement similar to long-term facilitation following intermittent hypoxia (see **Figure 5**).

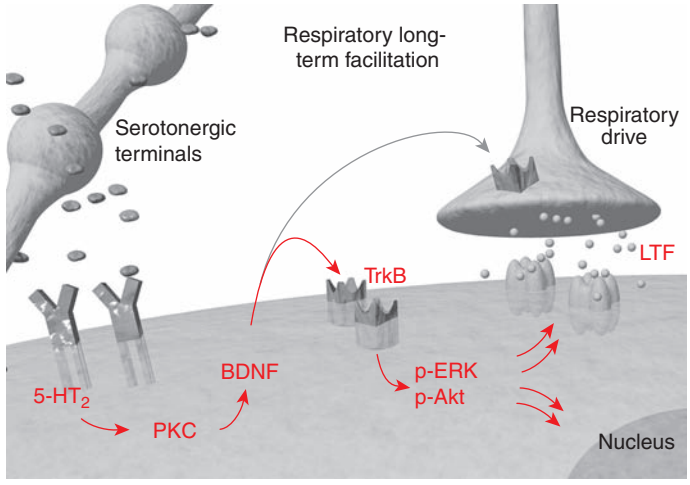

Figure 5 Working model of cellular/synaptic mechanisms underlying respiratory long-term facilitation (LTF). In the schema, inputs from premotoneurons (respiratory drive) form glutamatergic synapses on phrenic motoneuron dendrites (gray shaded region). Raphe serotonergic neurons with varicosities near phrenic motoneurons are coactivated with respiratory drive during intermittent hypoxia, releasing serotonin in an episodic pattern and stimulating postsynaptic 5-hydroxytryptamine ($5\text{-}HT_2$) receptors on motoneurons. Intermittent $5\text{-}HT_2$ receptor activation initiates translation and new protein synthesis of brain-derived neurotrophic factor (BDNF) within respiratory motoneurons (PKC, protein kinase C). Subsequent BDNF release activates presynaptic and/or postsynaptic TrkB receptors, which increase synaptic strength between premotoneurons and phrenic motoneurons through activation of extracellular signal-regulated mitogen-activated protein kinases (p-ERKs) and/or protein kinase B (p-Akt). Increased synaptic strength likely results from increased phosphorylation and trafficking of postsynaptic glutamatergic receptors, although effects on presynaptic glutamate release cannot be ruled out. Phosphorylated ERK and/or Akt may also translocate to the nucleus, initiating transcription and leading to longer lasting forms of plasticity and metaplasticity in respiratory motor control.

In *in vitro* preparations from neonatal rodents, intermittent applications of 5-HT_{2A} receptor agonists or serotonin to hypoglossal and/or spinal respiratory motoneurons elicit long-lasting facilitation of respiratory motor output (*in vitro* LTF). *In vitro* LTF is pattern sensitive, and results from increased α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor currents. AMPA-induced currents in hypoglossal motoneurons are enhanced by episodic serotonin receptor activation in preparations in which action potential generation has been blocked, thus demonstrating (1) that *in vitro* LTF at least partially results from postsynaptic effects in motoneurons and (2) that *in vitro* LTF is an activity-independent form of synaptic plasticity.

Although the physiological significance of LTF remains uncertain, it may stabilize breathing during sleep. Apneas, with consequent hypoxic episodes, are frequent during sleep; if multiple apneas occur in succession, intermittent hypoxia may induce LTF and prevent subsequent apneas by increasing respiratory motor output and stiffening the upper airway. This hypothesis is consistent with the observation that 15–25 s hypoxic episodes are sufficient to elicit LTF – a duration consistent with apnea-induced hypoxia. Regardless of its specific role, LTF indicates the capacity to alter synaptic strength of inputs to respiratory motoneurons. This capacity may impart the ability to shift the balance of motoneuron pools contributing to a breath. For example, during pregnancy, when the growing fetus compromises diaphragmatic function, a shift toward greater activation of accessory inspiratory muscles alleviates functional breathing deficits.

Chronic Intermittent Hypoxia

When intermittent hypoxia is maintained for prolonged periods, as would occur with sleep apnea, additional mechanisms of plasticity are observed. For example, resting breathing and the hypoxic ventilatory response are increased in rats exposed to chronic intermittent hypoxia (CIH). Increased carotid body sensitivity and central neural integration of carotid chemoafferent neurons are both implicated in these forms of plasticity. Another unique form of plasticity revealed in rats pretreated with CIH is sensory LTF. Following acute intermittent hypoxia in normal rats, carotid body activity increases during hypoxic episodes, but returns to baseline activity shortly thereafter; thus, acute intermittent hypoxia does not elicit long-lasting facilitation of carotid chemosensory activity in normal rats. However, acute intermittent hypoxia induces a long-lasting facilitation of carotid chemosensory activity in rats pretreated with CIH (i.e., sensory LTF). Sensory LTF requires reactive oxygen species generation during CIH and is a serotonin-dependent mechanism.

CIH also elicits metaplasticity, or plasticity in the ability to express plasticity. For example, CIH pretreatment attenuates PHFD and enhances phrenic LTF following acute intermittent hypoxia. Enhanced LTF following CIH is a serotonin-dependent mechanism, but different serotonin receptor subtypes are involved; enhanced LTF requires the activation of both 5-HT_{2A} and 5-HT_7 receptors. Enhanced LTF may reflect combined mechanisms of sensory LTF and normal phrenic LTF, although this hypothesis is untested.

Ventilatory Acclimatization to Chronic Sustained Hypoxia

At the start of sojourn to altitude or chronic sustained hypoxia, the acute increase in ventilation is followed by a slow, time-dependent increase over hours to days, a phenomenon known as ventilatory acclimatization to chronic hypoxia (VACH; Figure 3). VACH mitigates the severity of hypoxia at altitude, but the costs are worsening hypocapnia and respiratory alkalosis. The primary mechanism of VACH appears to be carotid body plasticity. During chronic hypoxia, the carotid body hypertrophies and increases its hypoxic sensitivity, contributing to increased ventilation. When hypoxia persists for days to weeks, NTS relay neurons also exhibit plasticity, accentuating an already elevated input from carotid chemoafferent neurons. When normal O_2 is restored following chronic hypoxia, ventilation only gradually returns to baseline levels, with a time course similar to the onset of VACH, most likely reflecting the reversal of VACH.

Developmental Plasticity in Ventilatory Control

Similar to other neural systems, experiences during development can have lifelong impact on ventilatory control. These effects are often unique to a critical developmental period. For example, when newborn rats are exposed to modest hyperoxia (30–60% O_2) during the first 4 weeks of life, normal carotid body development is suppressed, causing (1) carotid body hypoplasia, (2) decreased chemoafferent neuron counts in the carotid sinus nerve, (3) blunted carotid body responses to stimulation, and (4) blunted hypoxic ventilatory responses in adult rats (Figure 6). There is no evidence for abnormal central neural integration of chemoafferent inputs. Long-lasting effects are not observed when adult rats receive similar hyperoxic exposures, demonstrating that the plasticity is unique to development. With time, rats exposed to perinatal hyperoxia experience slow, spontaneous recovery or lifelong functional deficits, depending on the duration of hyperoxic exposure; rats exposed to 1 week of perinatal hyperoxia recover, whereas those

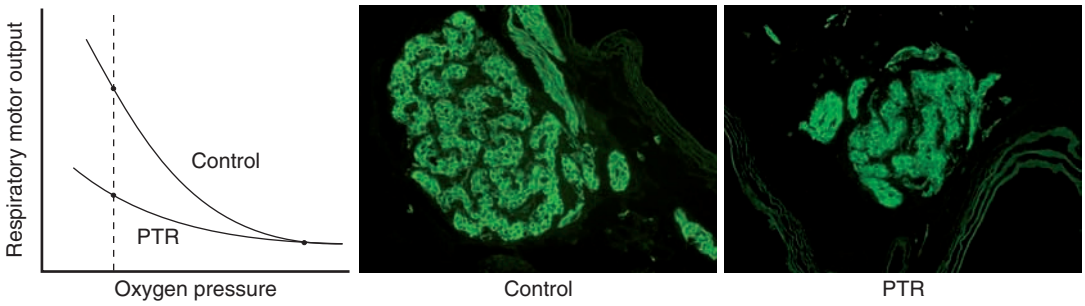

Figure 6 Developmental plasticity of the hypoxic ventilatory response and carotid chemoreceptors in rats exposed to perinatal hyperoxia. Left: The hypoxic ventilatory response is diminished in adult rats that received moderate hyperoxia (60% O_2) during the first four postnatal weeks. Suppression of the adult hypoxic ventilatory response arises from abnormal carotid body chemoreceptor development. Photomicrographs: Carotid bodies in a control rat (middle) and a rat exposed to hyperoxia for the first two postnatal weeks (right; PTR, perinatal treated rat). Images were captured at 200 . Green fluorescence indicates staining for tyrosine hydroxylase, which identifies carotid body glomus cells. Carotid body hypoplasia is associated with decreased axon counts from chemoafferent neurons in the carotid sinus nerve, and diminished carotid sinus nerve responses to hypoxia or intravenous cyanide. Image's provided courtesy of Wang ZY, collected during the study of Wang ZY and Bisgard GE (2005) Postnatal growth of the carotid body. *Respiratory Physiology & Neurobiology* 149: 181–190; graph drawn after Ling L, Olson EB, Vidruk EH, et al. (1997) Developmental plasticity of the hypoxic ventilatory response. *Respiration Physiology* 110: 261–268.

exposed for 4 weeks do not. The mechanism of impaired carotid chemoreceptor development with perinatal hyperoxia is not known, but it does not appear to be overt oxygen toxicity. Suppressed carotid body activity during hyperoxia may suppress the synthesis of neurotrophic factors necessary for developmental cellular hypertrophy and the attachment of chemoafferent neurons to glomus cells.

Modulation and Plasticity of the Exercise Ventilatory Response

The exercise ventilatory response undergoes plasticity. The best studied example is with increases in respiratory dead space, as would occur with impaired pulmonary gas exchange. Increased dead space diminishes alveolar ventilation and, consequently, increases resting ventilation via negative feedback from CO_2 chemoreceptors. However, with the onset of exercise, the ventilatory response is augmented by a mechanism that does not require changes in chemoreceptor feedback from rest to exercise. The augmented exercise ventilatory response is known as short-term modulation, and requires spinal serotonin receptor activation. During exercise with increased dead space, spinal serotonin receptor activation may increase respiratory motoneuron excitability, thereby enhancing the feed-forward exercise stimulus at the level of respiratory motoneurons (Figure 4). If the dead space is removed after a single exercise trial, the exercise ventilatory response reverts quickly to normal. However, with repeated trials of exercise with increased dead space, a form of plasticity known as long-term modulation is observed. Long-term modulation is a prolonged augmentation of the feed-forward exercise stimulus that lasts for many trials (or hours) after the dead space has been returned to normal. Long-term modulation is serotonin dependent, and may reflect a more robust response to the same stimuli initiating short-term modulation.

Age, Sex, and Genetics Influence Ventilatory Control

Important changes occur in ventilatory control from neonatal development through old age, and these changes are modulated by sex hormones. For example, sex hormones modulate resting ventilation and chemoreflexes during pregnancy. Respiratory plasticity is also influenced by sex hormone changes with aging. In female rats, respiratory LTF increases with age, from young adults to middle age, and then decreases in geriatric rats that have become acyclic. In contrast, LTF is greatest in young male rats, but decreases considerably by middle age; this effect appears to result from age-associated decreases in testosterone-derived CNS estrogen levels.

Ventilatory control is also influenced by genetics. Indeed, transgenic and knockout mice often exhibit interesting respiratory control phenotypes. Of particular interest are those with known associations to human disease (e.g., MeCP2, SOD-1, and leptin). Even normal animals exhibit variability in respiratory control. For example, different rodent strains vary in their hypoxic and hypercapnic ventilatory responses and their expression of respiratory plasticity. Genetic influences on the neural control of breathing remain an important area for future research.

Disorders of Ventilatory Control

Ventilatory control disorders have major human health consequences and are the cause of severe morbidity and

mortality. Some disorders are specific deficits in the ventilatory control system. Other ventilatory control disorders are associated with less specific neurological diseases or injury, and may exacerbate their clinical presentation and/or end life.

Sleep-Disordered Breathing

Sleep-disordered breathing is characterized by repetitive periods without breathing (apnea) and may be particularly prevalent during specific stages of development or aging. Morbidities due to intermittent hypoxia associated with sleep apnea include pulmonary and systemic hypertension, insomnia, depression, learning deficits, hippocampal apoptosis, and hyperactivity in children. Apneas during sleep result from (1) central apneas due to lack of effort and (2) obstructive apneas due to upper airway collapse, preventing breathing despite continued effort. While the mechanistic basis of central sleep apnea is unclear, deficits in brain stem respiratory neurons are often indicated. Central sleep apneas result from neurological diseases (e.g., tumors, traumatic brain stem injury, infectious diseases), circulatory disorders (e.g., congestive heart failure, arteriosclerosis), and sleep at altitude. The latter is characteristic of even normal humans and contributes to acute mountain sickness.

Obstructive sleep apnea (OSA) is the most common ventilatory control disorder in both children and adults. Snoring and excessive sleepiness are important clinical signs indicative of OSA; however, most cases of OSA remain undiagnosed. A strong influence of sex hormones is indicated, since OSA occurs in at least twice as many men as women. OSA is most frequent in middle-aged men, postmenopausal women, and children. Although obesity and congenitally narrowed airways are associated with OSA, diminished modulatory influences on upper airway motoneurons may predispose the upper airway to narrowing or collapse.

Developmental Disorders

Apnea of prematurity and sudden infant death syndrome (SIDS) are relatively frequent in the pediatric population. Apnea of prematurity is characterized by both central and obstructive apneas and is due to an immature respiratory control system. SIDS is characterized by a complete cessation of breathing during the first year of life (particularly during sleep) and is strongly associated with abnormalities in the raphe serotonergic system.

Traumatic and Ischemic Neural Injury

Traumatic or ischemic injury may impair neurons or synaptic inputs to the VRC or respiratory motoneurons. For example, spinal injury interrupts descending projections from brain stem premotoneurons to spinal motoneurons or interneurons. Of particular importance, high cervical spinal injuries interrupt the ability of phrenic motoneurons at cervical spinal levels C3–C6 to drive the diaphragm (and other respiratory muscles innvervated by more caudal spinal segments). Thus, patients with high cervical spinal injuries often die of respiratory insufficiency. Patients with spinal injuries below the cervical spinal cord generally have a normal breathing pattern at rest, but may have impaired breathing during increased respiratory drive or in respiratory defense reflexes such as coughing. Since most spinal injuries are incomplete, one goal of current research is to strengthen spared synaptic pathways to induce at least partial functional recovery.

Neurodegenerative Diseases

Ventilatory insufficiency and sleep apnea are significant complications in many neurodegenerative diseases, including amyotrophic lateral sclerosis (ALS), Parkinson's disease, Huntington's disease, multiple sclerosis, multiple system atrophy, and Alzheimer's disease. Motoneuron disease causes the progressive death of respiratory motoneurons, leading to ventilatory failure; to preserve function, surviving motoneurons must carry a greater load. Less frequently appreciated, more than 50% of patients with Parkinson's, Alzheimer's, and other neurodegenerative diseases experience severe sleep apnea.

Genetic Diseases

Patients with congenital central hypoventilation syndrome (CCHS), Rett syndrome, and Down syndrome have severe abnormalities in the control of breathing. CCHS is characterized by poor or complete lack of automatic breathing mechanisms during sleep and is associated with mutations in the homeobox gene, *Phox2B*. Rett syndrome is associated with mutations in MeCP2, a regulator of gene expression (including BDNF). Patients with Rett syndrome experience alternating episodes of hyperventilation and central apneas (i.e., breath holding) during wakefulness. MeCP2 mutations cause abnormalities in noradrenergic neurons, an effect that may explain at least some ventilatory control abnormalities experienced by these patients. Down syndrome is caused by an extra copy of chromosome 21 and is associated with a high (>50%) incidence of OSA, likely due to craniofacial abnormalities and/or deficits in the serotonergic system. Repetitive intermittent hypoxia during development may contribute to learning deficits in these patients.

Psychiatric Disorders

Anxiety hyperventilation disorder is a frequent affliction in the pediatric population (~6%) and in patients with prior psychological stress. Although the mechanisms of anxiety hyperventilation are not known, most treatments are based on manipulation of the serotonergic nervous system. Neonatal stressors, such as maternal separation in rodents, can have a lifelong impact on adult chemoreflexes and ventilatory control.

Ventilatory Control during Lung Disease

During lung and/or chest wall diseases, mechanics of the respiratory system change, requiring major adjustments in respiratory motor output. With increased lung compliance or lower chest wall compliance (e.g., emphysema and scoliosis), shifts in the balance between diaphragmatic and inspiratory intercostal muscle contractions are necessary to achieve adequate ventilation. With increased airway resistance, greater respiratory muscle contractions are necessary. When mechanical disturbances exceed the ability to compensate, ventilatory failure results.

Common Features of Ventilatory Control Disorders

Ventilatory control adjustments during disease require a robust and flexible neural control system. Since similar network and cellular mechanisms compensate during multiple breathing disorders, common therapeutic approaches to seemingly disparate diseases may be possible. For example, respiratory disorders caused by cervical spinal injury, ALS, and OSA share (relative) deficits in respiratory motoneuron output. Since they share a common need for increased motor output to respiratory muscles, therapeutic approaches designed to strengthen synaptic inputs to respiratory motoneurons may be useful. Disorders leading to central apneas suggest defects in respiratory rhythm generation. Thus, drugs that stimulate respiratory rhythm generation may prove useful. Although our understanding of detailed cellular, synaptic and network mechanisms is rudimentary at best for any ventilatory control disorder, an understanding of promising molecular targets may lead to the development of new therapeutic strategies for devastating respiratory control disorders for which there are few effective treatments and no known cures.

See also: Autonomic Nervous System: Respiratory Control; Autonomic Nervous System: Central Respiratory Control; Brainstem Respiratory Circuits; Evolution of Vertebrate Respiratory Control; Network Control; Sleep Apnea.

Further Reading

Carroll JL (2003) Developmental plasticity in respiratory control. *Journal of Applied Physiology* 94: 375–389.

Feldman JL and Del Negro CA (2006) Looking for inspiration: New perspectives on respiratory rhythm generation. *Nature Reviews Neuroscience* 7: 232–242.

Feldman JL, Mitchell GS, and Nattie EE (2003) Breathing: Rhythmicity, plasticity, chemosensitivity. *Annual Review of Neuroscience* 26: 239–266.

Gaultier C, Matrot B, and Gallego J (2006) Transgenic models to study disorders of respiratory control in newborn mice. *ILAR Journal* 47: 15–21.

Gaultier C and Gallego J (2005) Development of respiratory control: Evolving concepts and perspectives. *Respiratory Physiology & Neurobiology* 149: 3–15.

Goshgarian HG (2003) The crossed phrenic phenomenon: A model for plasticity in the respiratory pathways following spinal cord injury. *Journal of Applied Physiology* 94: 795–810.

Gozal D (2004) New concepts in abnormalities of respiratory control in children. *Current Opinion in Pediatrics* 16: 305–308.

Guyenet PG, Stornetta RL, Bayliss DA, et al. (2005) Retrotrapezoid nucleus: A litmus test for the identification of central chemoreceptors. *Experimental Physiology* 90: 247–253.

Katz DM (2005) Regulation of respiratory neuron development by neurotrophic and transcriptional signaling mechanisms. *Respiratory Physiology & Neurobiology* 149: 99–109.

Ling L, Olson EB, Vidruk EH, et al. (1997) Developmental plasticity of the hypoxic ventilatory response. *Respiration Physiology* 110: 261–268.

Mahamed S and Mitchell GS (2006) Is there a link between intermittent hypoxia-induced respiratory plasticity and obstructive sleep apnoea? *Experimental Physiology* 92: 27–37.

Mantilla CB and Sieck GC (2003) Mechanisms underlying motor unit plasticity in the respiratory system. *Journal of Applied Physiology* 94: 1230–1241.

Mitchell GS, Baker TL, Nanda SA, et al. (2001) Physiological and genomic consequences of intermittent hypoxia: Invited review: Intermittent hypoxia and respiratory plasticity. *Journal of Applied Physiology* 90: 2466–2475.

Mitchell GS and Johnson SM (2003) Neuroplasticity in respiratory motor control. *Journal of Applied Physiology* 94: 358–374.

Mitchell GS, Babb TG, et al. (2006) Layers of exercise hyperapnea: Modulation and plasticity. *Respiratory Physiology & Neurobiology* 151: 251–266.

Powell FL, Milsom WK, and Mitchell GS (1998) Time domains of the hypoxic ventilatory response. *Respiration Physiology* 112: 123–134.

Prabhakar NR and Peng YJ (2004) Peripheral chemoreceptors in health and disease. *Journal of Applied Physiology* 96: 359–366.

Prabhakar NR (2006) O_2 sensing at the mammalian carotid body: Why multiple O_2 sensors and multiple transmitters? *Experimental Physiology* 91: 17–23.

Putnam RW, Filosa JA, and Ritucci NA (2004) Cellular mechanisms involved in CO(2) and acid signaling in chemosensitive neurons. *American Journal of Physiology – Cell Physiology* 287: C1493–C1526.

Reeves SR and Gozal D (2005) Developmental plasticity of respiratory control following intermittent hypoxia. *Respiratory Physiology & Neurobiology* 149: 301–311.

Richerson GB, Wang W, and Hodges MR (2005) Homing in on the specific phenotype(s) of central respiratory chemoreceptors. *Experimental Physiology* 90: 259–266.

Wang ZY and Bisgard GE (2005) Postnatal growth of the carotid body. *Respiratory Physiology & Neurobiology* 149: 181–190.

Restless Leg Syndrome

S L Clardy and J R Connor, Pennsylvania State University Milton S. Hershey Medical Center, Hershey, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Restless legs syndrome (RLS) is characterized by dysthesias and an irresistible urge to move the lower limbs. RLS sensations generally occur when the sufferer is at rest, and they are relieved by voluntary movement. Attempting to suppress the dysthesias only increases patient discomfort, often making activities such as plane travel or dining out extremely unpleasant. A circadian pattern characterizes RLS, with symptoms worsening in the evening. The prevalence of RLS in Western countries is estimated to be 7.2%, making it a common disorder of the adult population. The prevalence is approximately twice as high for women as for men, and the percentage of the population with RLS increases with age.

RLS can be broadly classified into two categories, as outlined in (Figure 1): primary, with no known relation to another disorder, or secondary, with the symptoms attributable to another medical condition. Major causes of secondary and reversible RLS include pregnancy, iron deficiency, and end stage renal disease both before and after dialysis treatment. Approximately 25–40% of patients with any of these three conditions report RLS, which is approximately three to five times the expected occurrence for adults. Generally, RLS dramatically improves with the effective treatment or resolution of these primary conditions. It is unclear if there is an association between RLS and neuropathy or Parkinson's disease. Some studies also suggest that RLS may be associated with essential tremor in some families. Limited studies have also demonstrated that attention-deficit hyperactivity disorder (ADHD) also appears to be associated with RLS in children and adults and possibly with RLS in the parents of children with ADHD.

Primary RLS onset can occur at almost any age, with symptom onset classified as early onset if symptoms begin before age 45 years and as late onset if symptoms begin after the age of 45 years. When symptoms begin before the age of 45 years, they are generally slowly progressive, with the expression of the most severe symptoms occurring mainly in adults older than age 50 years. The first-degree relatives of patients whose onset of symptoms occurs before age 45 years are five times more likely to have RLS than the general population.

The accepted diagnostic criteria for primary RLS are based on the following clinical symptoms: (1) an urge to move the extremities, usually the legs, frequently accompanied by dysthesias; (2) onset or exacerbation of symptoms when at rest; (3) relief of the urge to move or unpleasant sensations with movement; and (4) symptoms have a marked circadian pattern, becoming worse at night or only occurring at night. The circadian cycle of RLS symptoms is a primary effect and not secondary to more rest activity during the night. Symptom severity varies widely, and for more severely affected patients, the agony of the sensations may eventually continue through both the day and the night (but the worsening at night must have been previously present to meet the diagnostic criteria). Total daily sleep times are often reduced to 5 h or less, with marked reduction of sleep efficiency.

Sleep deprivation is a common feature of the RLS population, an observation consistent with reports of a circadian rhythm to the symptoms onset and severity (e.g., worsening of symptoms at night). Indeed, a significant number of RLS sufferers are initially diagnosed based on interviews with the patient's bed partner. RLS sufferers are sleep deprived not only because they have difficulty in sleep maintenance but also because they generally have a prolonged latency to sleep onset. This delayed sleep onset is not unexpected because attempting to relax triggers symptoms. Some RLS sufferers may also exhibit involuntary jerking movements, commonly of the legs. Such movements are referred to as dyskinesias while awake (DWA) and typically involve flexion of the leg, hip, knee, or ankle. The presence of DWA may be more specific for RLS than periodic leg movements. The majority of individuals with RLS exhibit periodic limb movements of sleep (PLMS). These are repetitive movements that occur during the period of nonrapid eye movement sleep. When severe, PLMS can disrupt sleep. PLMS increase in prevalence with age, and thus a finding of PLMS in a younger patient is more suggestive of RLS than such a finding in an elderly patient. PLMS occur in 70–90% of RLS patients, although at most, only 30% of patients with PLMS will have RLS.

A specific gene mutation has not been found in RLS despite a considerable amount of research in recent years. Three different loci for RLS have been reported on chromosomes 12q, 14q, and 9p and are referred to as RLS-1, RLS-2, and RLS-3, respectively, as outlined in (Figure 2). RLS-1 was mapped to a French–Canadian family assuming a recessive model of inheritance with a possible pseudodominant appearance, whereas RLS-2 (found in a north Italian family) and RLS-3 (discovered in an American family) mapping was based on an autosomal dominant mode of inheritance. The significance of these three loci is not fully

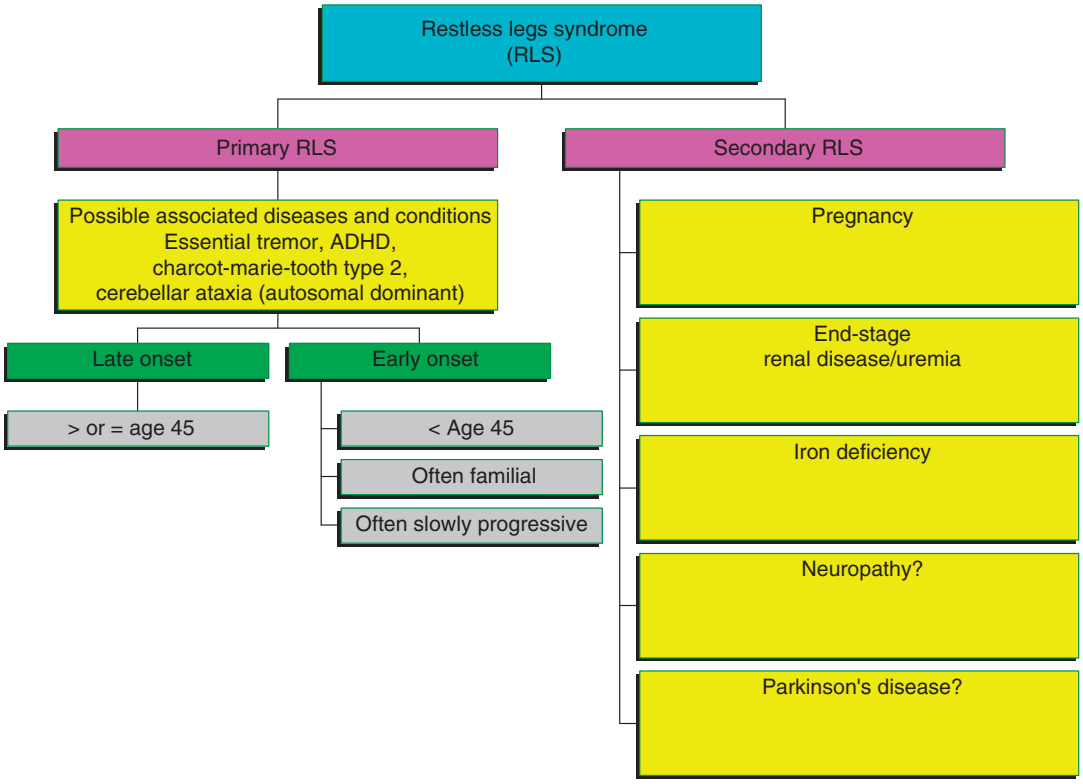

Figure 1 Primary and secondary restless legs syndrome.

understood, and further investigations of different RLS families are needed to confirm the significance of the loci and possibly identify additional loci in RLS. The genetics of RLS are further complicated by reports of RLS in familial forms of other diseases, such as Charcot–Marie–Tooth type 2 and autosomal dominant-type cerebellar ataxia, as well as case reports and studies suggesting a predilection for mother-to-child transmission.

The current clinical data clearly support a role for the dopaminergic system in this disorder. Dopamine antagonists can aggravate RLS or induce a focal akathesia clinically similar to RLS. Early clinical studies indicated that RLS was dopamine responsive. Subsequently, low-dose levodopa treatment was found to be significantly more effective than the relatively high doses of opiates that had previously been used to treat RLS. Levodopa was one of the first drugs to be proven efficacious in placebo-controlled trials. Treatment with levodopa is greatly limited, however, by augmentation of symptoms, including onset of RLS symptoms earlier in the evening compared to pretreatment onset, involvement of other body parts, and/or increased severity. In 2005, the dopamine agonist ropinirole (Requip) became the first Food and Drug Administration-approved medication for the treatment of moderate to severe primary RLS (15 or more episodes monthly). Consistent with the clinical observations, some brain imaging studies of RLS patients have shown striatal dopamine abnormalities. Not all studies have found striatal abnormalities, however, and none of the studies were done in the evening, when subjects are symptomatic. It is possible that variable symptom expression could account for the conflicting findings, underscoring the need for evaluation of subjects during the symptomatic period. An autopsy report has identified the first direct evidence for a dopaminergic involvement in RLS.

Opioids, benzodiazepines, carbamazepine, valproic acid, gabapentin, clonidine, iron, and magnesium have all been reported to be beneficial in the treatment of RLS, but experience with these drugs is limited. Nonpharmacologic treatments include improved nutrition, exercise, and sleep hygiene. Additionally, many patients tend to find that engaging in a mental activity requiring concentration, such as painting or computer programming, alleviates symptoms. Caffeine, antihistamines, and cold preparations have been reported to aggravate symptoms, whereas alcohol may offer a brief reduction in symptoms but with rebound worsening of symptoms.

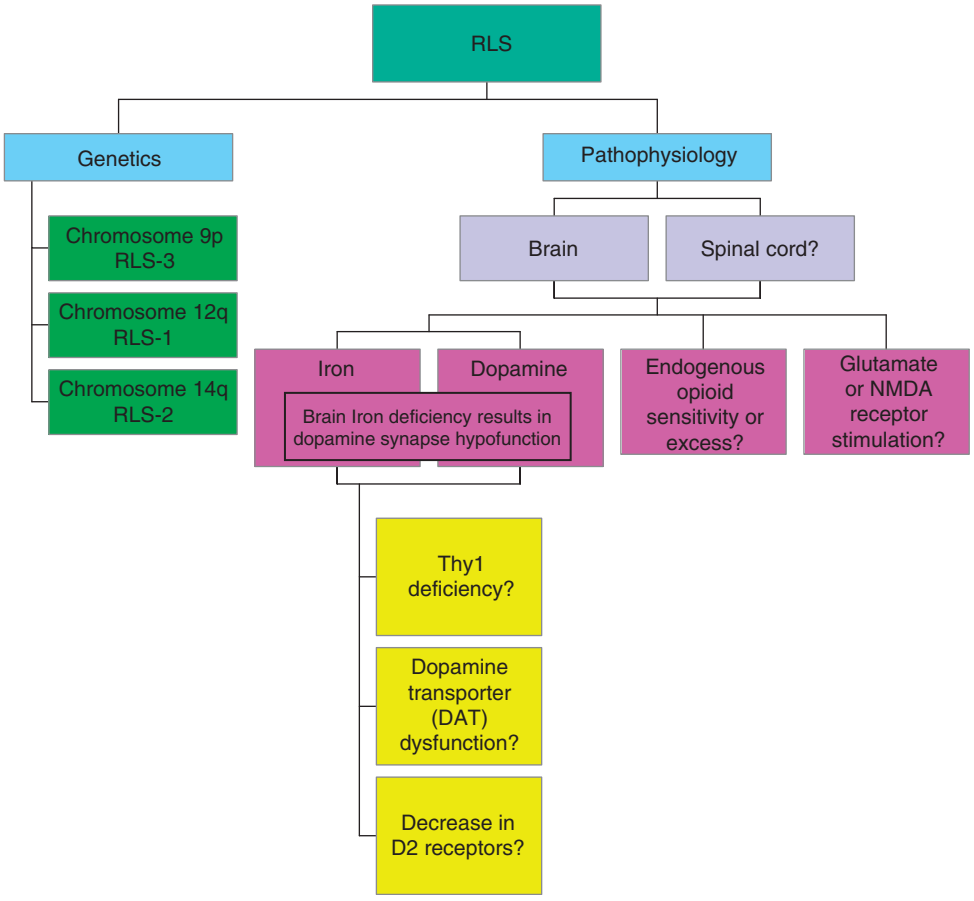

Figure 2 Genetics and pathophysiology of restless legs syndrome.

Although much of the clinical intervention for RLS has focused on the dopaminergic system, there is a growing body of literature that suggests a significant role for iron. Iron is a cofactor for tyrosine hydroxylase, which in turn is the rate-limiting enzyme for dopamine metabolism. Iron deficiency has been recognized as a significant contributing cause of RLS for approximately 50 years. Oral iron supplementation has been shown to improve symptoms in some RLS patients, but the uptake of iron is limited by the rate of absorption from the gut. Intravenous iron injection eliminates this barrier to absorption, and as shown by Nordlander, multiple, low-dose, intravenous iron therapy brought about complete relief of symptoms in 21 of 22 RLS patients for periods of up to 3–9 months, even though the majority had normal iron status prior to treatment. Earley et al. demonstrated symptomatic improvement for an average of 6 months in a small number of RLS patients after a 1000-mg injection of iron dextran. Furthermore, they demonstrated that repeated infusions of 450 mg iron gluconate provided additional months of relief. The rate of serum ferritin excretion in the subjects was much greater than predicted after the first ferritin injection, but the excretion rate decreased with subsequent injections. The elevated loss of ferritin in RLS patients may suggest part of the mechanism by which iron contributes to the biology underlying this disorder. In support of this idea, a strong negative correlation between serum ferritin levels and RLS symptom severity has been observed.

Low iron in the central nervous system in RLS may be more important than low iron levels in serum in RLS pathophysiology because low central nervous system iron storage may directly result in reduced dopamine synthesis and/or receptor function. Early onset RLS is associated with significant decreases in several iron-related proteins, including whole-molecule ferritin, H-ferritin, L-ferritin, and pro-hepcidin in cerebrospinal fluid (CSF). Late-onset CSF values in RLS subjects, however, are not significantly decreased compared to controls. Analyses of the CSF in RLS have provided compelling evidence for disruption of brain iron homeostasis in RLS.

Additionally, there is a strong correlation between the age of symptom onset and CSF ferritin values such that the earlier the age of RLS onset, the lower the CSF ferritin level at the time of the study. These studies suggest that early and late-onset phenotypes may define two distinct disease processes.

Direct evidence of brain iron deficiency derives from a magnetic resonance imaging analysis that indicates brain iron levels in the substantia nigra are below normal in individuals with RLS. Both the substantia nigra and the putamen showed iron losses in proportion to RLS severity. Histopathological analyses and biochemical studies on autopsy material also provide support that the substantia nigra in RLS is iron deficient. Two autopsy studies have been performed, and neither found histopathological alterations that suggest a structural basis (e.g., plaques and Lewy bodies) for RLS. An intriguing observation is that a neuronal surface protein, Thy1, which is important for maintaining structural integrity of the synapse and promoting interaction of presynaptic vesicles with the synaptic membrane, is decreased in the substantia nigra of individuals with RLS and in animal models of iron deficiency. Thus, the working hypothesis for the biological basis for RLS is that it is a result of a hypofunctionality of the dopaminergic system that stems from insufficient amounts of brain iron. Autopsy studies on the dopaminergic parameters are under way.

A discussion on the biological basis would be remiss in not including the possibility of a spinal origin for RLS. Reports exist that individuals with RLS may have enhanced spinal excitability or diminished central inhibition. Animal experiments have clearly demonstrated that the spinal cord is intrinsically able to generate motor activity without descending input through a central pattern generator (CPG), thus raising the possibility that the pathophysiology of RLS may be independent of the brain. The existence of RLS in paraplegic patients supports the possible role of CPGs in humans. In RLS patients and spinal cord injury patients, there is also lower threshold and greater spatial spread of the flexor reflex. In RLS patients, this difference is more prominent during sleep, perhaps contributing to the nocturnal worsening of symptoms in RLS. Further evidence for spinal cord involvement in RLS derives from studies of task-dependent modulation. Motor activity modifies spinal reflex pathways, and this presynaptic control of sensory afferent inflow outlasts the period of motor activity by a few minutes. This could help to explain the temporary relief of dysthesias in RLS patients after movement. A study that lesioned the A11 dopaminergic diencephalic spinal neurons in rats found an increased number of standing episodes and increased total standing time (behaviors suggestive of RLS in an animal model) and a subsequent reduction in the total standing episodes and standing time in the lesioned animals after intramuscular pramipexole treatment. Within the spinal cord, there is an extensive dopaminergic system with roles in motor functions and sensorimotor integration, although the exact relevance is not fully understood, and thus it is not understood how levodopa and dopamine agonists affect spinal cord networks in RLS.

RLS is a common and likely underdiagnosed disorder that ranges in severity from irritating to debilitating. Current evidence suggests that primary RLS is a consequence of iron- and dopamine-related dysfunction in one or more locations in the central nervous system. Additional research is necessary to elucidate the specific pathophysiologic cause and the primary or initial location of the iron- and dopamine-related dysfunction, as well as the genetic inheritance and chromosomal linkage patterns. Nonetheless, data strongly support a biological basis for RLS and evidence is building against the early notion that RLS is a psychosomatic disorder.

See also: Attention Deficit Hyperactivity Disorder; Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine; Dopamine in Perspective; Dopamine – CNS Pathways and Neurophysiology; Genetics of Circadian Disorders in Humans; Sleep Architecture; Sleep Deprivation and Brain Function.

Further Reading

Allen RP, Barker PB, Wehrl F, Song HK, and Earley CJ (2001) MRI measurement of brain iron in patients with restless legs syndrome. *Neurology* 56: 263–265.

Allen RP and Earley CJ (2000) Defining the phenotype of the restless legs syndrome (RLS) using age-of-symptom-onset. *Sleep Medicine* 1: 11–19.

Allen RP, Picchietti D, Hening WA, Trenkwalder C, Walters AS, and Montplaisi J (2003) Restless Legs Syndrome Diagnosis and Epidemiology workshop at the National Institutes of Health; International Restless Legs Syndrome Study Group. Restless legs syndrome: Diagnostic criteria, special considerations, and epidemiology. A report from the Restless Legs Syndrome Diagnosis and Epidemiology workshop at the National Institutes of Health. *Sleep Medicine* 4(2): 101–119.

Allen RP, Walters AS, Montplaisir J, et al. (2005) Restless legs syndrome prevalence and impact: REST general population study. *Archives of Internal Medicine* 165(11): 1286–1292.

Bara-Jimenez W, Aksu M, Graham B, Sato S, and Hallett M (2000) Periodic limb movements in sleep: State-dependent excitability of the spinal flexor reflex. *Neurology* 54(8): 1609–1616.

Bonati MT, Ferini-Strambi L, Aridon P, Oldani A, Zucconi M, and Casari G (2003) Autosomal dominant restless legs syndrome maps on chromosome 14q. *Brain* 126(part 6): 1485–1492.

Chen S, Ondo WG, Rao S, Li L, Chen Q, and Wang Q (2004) Genomewide linkage scan identifies a novel susceptibility locus

for restless legs syndrome on chromosome 9p. *American Journal of Human Genetics* 74(5): 876–885.

Clardy SL, Earley CJ, Allen RP, Beard JL, and Connor JR (2006) Ferritin subunits in CSF are decreased in restless legs syndrome. *Journal of Laboratory and Clinical Medicine* 147(2): 67–73.

Clardy SL, Wang X, Boyer PJ, Earley CJ, Allen RP, and Connor JR (2006) Is ferroportin–hepcidin signaling altered in restless legs syndrome? *Journal of the Neurological Sciences* 247(2): 173–179.

Connor JR, Boyer PJ, Menzies SL, et al. (2003) Neuropathological examination suggests impaired brain iron acquisition in restless legs syndrome. *Neurology* 61(3): 304–309.

Connor JR, Wang XS, Patton SM, et al. (2004) Decreased transferrin receptor expression by neuromelanin cells in restless legs syndrome. *Neurology* 62(9): 1563–1567.

Desautels A, Turecki G, Montplaisir J, Sequeira A, Verner A, and Rouleau GA (2001) Identification of a major susceptibility locus for restless legs syndrome on chromosome 12q. *American Journal of Human Genetics* 69(6): 1266–1270.

Earley CJ, Connor JR, Beard JL, Clardy SL, and Allen RP (2005) Ferritin levels in the cerebrospinal fluid and restless legs syndrome: Effects of different clinical phenotypes. *Sleep* 28(9): 1069–1075.

Earley CJ, Heckler D, and Allen RP (2005) Repeated IV doses of iron provides effective supplemental treatment of restless legs syndrome. *Sleep Medicine* 6(4): 301–305.

Hogl B, Kiechl S, Willeit J, et al. (2005) Restless legs syndrome: A community-based study of prevalence, severity, and risk factors. *Neurology* 64(11): 1920–1924.

Kotagal S and Silber MH (2004) Childhood-onset restless legs syndrome. *Annals of Neurology* 56(6): 803–807.

Nordlander NB (1953) Therapy in restless legs. *Acta Medica Scandia* 145: 453–457.

Ondo WG and Lai D (2006) Association between restless legs syndrome and essential tremor. *Movement Disorders* 21(4): 515–518.

Picchietti DL, Underwood DJ, Farris WA, et al. (1999) Further studies on periodic limb movement disorder and restless legs syndrome in children with attention-deficit hyperactivity disorder. *Movement Disorders* 14(6): 1000–1007.

Wang X, Wiesinger J, Beard J, et al. (2004) Thy1 expression in the brain is affected by iron and is decreased in restless legs syndrome. *Journal of the Neurological Sciences* 220(1–2): 59–66.

Winkelmann J, Lichtner P, Putz B, et al. (2006) Evidence for further genetic locus heterogeneity and confirmation of RLS-1 in restless legs syndrome. *Movement Disorders* 21(1): 28–33.

Relevant Website

http://www.rls.org – Restless Legs Syndrome Foundation.

Reticular Activating System

E Garcia-Rill, University of Arkansas for Medical Sciences, Little Rock, AR, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sleep and arousal states are characterized in the electroencephalogram (EEG) by synchronization of slow rhythms during slow-wave sleep and synchronization of fast rhythms during waking and rapid eye movement (REM) sleep. These rhythms are generated by reverberating oscillations between the thalamus, hypothalamus, and basal forebrain and the cortex. The reticular activating system (RAS) serves to modulate these oscillations (ascending inhibition or decreased excitation will serve to slow these rhythms, whereas ascending excitation or disinhibition will speed up rhythms) depending on sensory inputs and ongoing activity. In addition, descending modulation of motor neurons serves to decrease muscle tone during REM sleep (atonia) or prime the system to fight or flee during waking by modulating the startle response.

Wiring Diagram

RAS is composed of three main cell groups: the pedunculopontine nucleus (PPN) (and its medial partner, the laterodorsal tegmental nucleus, both of which are mainly cholinergic), the locus coeruleus (LC; mainly noradrenergic), and the raphe nucleus (RN; mainly serotonergic). RAS receives input from all afferent sensory systems in parallel to primary afferent sensory projections. That is, the 'nonspecific' projection system through RAS relays 'arousal' information via the intralaminar thalamus (ILT) to the cortex. This system functions in parallel with the flow of 'specific' sensory input through the primary sensory thalamic nuclei that is then relayed to the cortex. It is the temporal summation of ILT nonspecific inputs with primary sensory specific inputs at the level of the cortex that is thought to participate in the conscious perception of a sensory event. Disturbances in RAS driving of the ILT and thalamocortical oscillations thus can be expected to lead to disturbances in perception.

RAS also sends descending projections to posture and locomotion systems that act as the 'fight-or-flight' system, simultaneously activating higher centers while priming motor systems to respond appropriately to sudden stimuli. These descending projections modulate the pontine inhibitory area (PIA) that controls the atonia of REM sleep, pontine neurons that generate the startle response (a flexor response that primes the motor system), and reticulospinal systems that drive locomotion. RAS, especially cholinergic, projections to the dorsal subcoeruleus region also trigger the generation of ponto-geniculo-occipital (PGO) waves during REM sleep. This region generates high-frequency bursts of activity (such as those required for long-term potentiation) that have been proposed to participate in sleep-dependent plasticity.

The hypothalamic sleep–wake modulatory system is composed mainly of the tuberomammillary nucleus (TMN; mainly excitatory histaminergic), the lateral hypothalamus (LH; mainly excitatory orexinergic), and the ventrolateral preoptic nucleus (VLPO; mostly inhibitory GABAergic). These hypothalamic systems are thought to help stabilize sleep–wake states. Excitatory LH orexin projections can be thought of as an 'on' switch promoting waking, particularly through activation of excitatory TMN neurons, which are tonically active during waking (and decrease firing during sleep), and of excitatory basal forebrain and RAS neurons. The basal forebrain cholinergic system is mainly active during waking and serves to raise the excitability of the cortex, although acetylcholine release from the basal forebrain is greater during REM sleep than during waking. Inhibitory VLPO GABAergic projections can be thought of as an 'off' switch promoting slow-wave sleep through inhibition of RAS, LH, basal forebrain, and cortex.

Figure 1 outlines the main projections described previously. The dorsal ascending cholinergic (PPN) and monoaminergic (LC and RN) projections from RAS to ILT serve to activate the cortex via thalamocortical projections. There is also a ventral projection system that bypasses the thalamus to terminate diffusely throughout the cortex that originates in RAS (noradrenergic LC and serotonergic RN) and ascending projections from TMN, LH, and VLPO (Figure 1, VP), as well as from the basal forebrain (not shown). In turn, TMN, LH, and VLPO send descending projections to RAS that may act reciprocally to stabilize sleep–wake states.

In addition, adenosine (Figure 1, A; diffusely localized) is a ubiquitous homeostatic factor believed to be involved in sleep–wake regulation. High metabolic activity or prolonged wakefulness leads to a buildup of adenosine, which decreases subsequent to sleep. Adenosine may also modulate sleep–wake states via its inhibitory actions on most cells. Adenosine injections into RAS are known to decrease waking, whereas adenosine levels in the basal forebrain (but not in the thalamus) progressively increase during

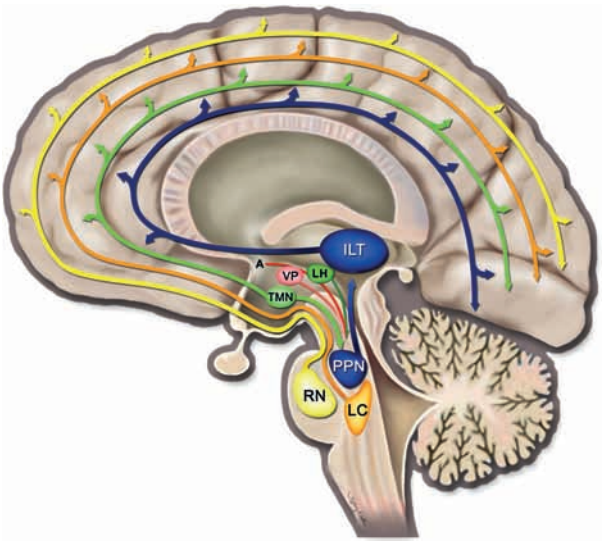

Figure 1 Main ascending projections of the reticular activating system. A, adenosine; ILT, intralaminar thalamus; LC, locus coeruleus; LH, lateral hypothalamus; PPN, pedunculopontine nucleus; RN, raphe nucleus; TMN, tuberomammillary nucleus; VP, ventrolateral preoptic nucleus.

prolonged wakefulness and decrease during subsequent recovery of sleep.

Figure 2 shows the connectivity of neurons within RAS and their firing patterns during sleep and waking. RN inhibits PPN and LC, and LC inhibits PPN, whereas PPN activates LC. During waking, all three cell groups are active, but in slow-wave sleep the cholinergic cells decrease firing while monoaminergic cells remain active. However, in REM sleep, the cholinergic cells are highly active while monoaminergic cells decrease their firing rates markedly, thereby disinhibiting cholinergic cell groups. Thus, cholinergic RAS neurons are active during waking and REM sleep (i.e., during synchronization of fast cortical rhythms) but slow their firing during synchronization of slow cortical rhythms in slow-wave sleep.

Intrinsic Properties

Neurons in the region of PPN have been reported to show 'wake-on,' 'REM-on,' and 'wake-REM-on' firing patterns whose properties are implied by their names. It is not clear how these patterns correspond to intrinsic membrane properties. There are three types of PPN neurons based on intrinsic membrane properties: type I have low-threshold spike (LTS) calcium-dependent inward currents, are noncholinergic, and tend to burst especially when released from hyperpolarization; type II have Ia transient outward potassium currents, are mostly cholinergic, and tend to fire slowly and tonically; and type III have both LTS and Ia currents, are mainly noncholinergic, and fire tonically and in bursts. A large proportion of these cells also have hyperpolarization-activated Ih currents that tend to restore the membrane potential following hyperpolarization. Evidence suggests that at least some neurons in PPN, possibly GABAergic, are electrically coupled, promoting ensemble activity. Cholinergic RAS neurons are active tonically during waking depending on the level of depolarization. In a slightly hyperpolarized state, some of these cells show LTS bursts, perhaps corresponding to the bursting activity observed during REM sleep. Then, further hyperpolarization leads to cessation of firing, as in slow-wave sleep. During moderate inhibition, PPN neurons thus may undergo a change from tonic to bursting activity such as that observed in waking versus REM sleep.

LC and RN cells fire relatively slowly during slow-wave sleep, cease firing during REM sleep (disinhibiting PPN cells), and are most active and variable during waking, particularly in response to novel stimuli. LC neurons show electrotonic coupling, which allows them to act as a group. Therefore, a novel stimulus could elicit a synchronized response from this population of neurons. However, recurrent inhibition follows such activation so that the initial response is quickly followed by inhibition of further output. RN cells are similar to LC neurons in that many of the conductances of these neurons tend to limit the firing rate (e.g., long-duration action potential and high-amplitude afterhyperpolarizations). It has been postulated that during waking serotonin inhibits REM-on neurons without affecting wake-REM-on neurons, thereby allowing synchronization of fast rhythms; however, these slow down during REM sleep, thus disinhibiting REM-on, presumably bursting, neurons to promote REM sleep. In general, noradrenergic and serotonergic neurons have high input resistance so that relatively small currents (such as those elicited by marginal inputs) will elicit considerable changes in membrane potential. There are also both excitatory and inhibitory interneurons (GABAergic) and projection (glutamatergic) neurons intermingled within all RAS cell groups.

Descending cholinergic projections to the subcoeruleus region affect several types of neurons, including cells with LTS currents, Ia currents, and LTS and Ia currents – similar to the types found in PPN. Electrical coupling appears to be more common in this region, mainly among GABAergic neurons that are induced to fire rhythmically at theta rhythm by cholinergic agonists. Such rhythmic coherent activation may underlie the mechanism behind REM sleep and perhaps PGO wave generation.

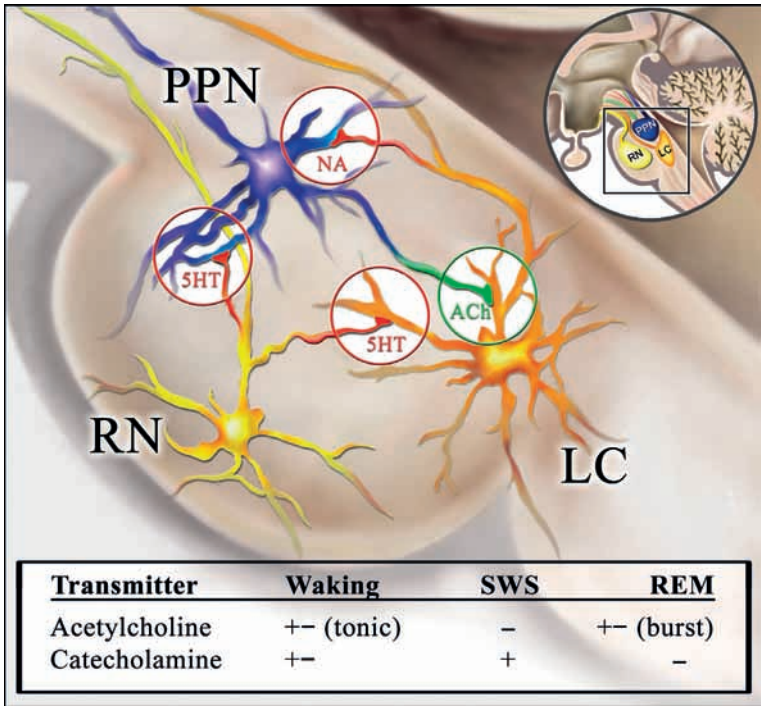

Transmitter	Waking	SWS	REM
Acetylcholine	+− (tonic)	−	+− (burst)
Catecholamine	+−	+	−

Figure 2 Wiring diagram within the main cell groups of the reticular activating system. LC, locus coeruleus; PPN, pedunculopontine nucleus; RN, raphe nucleus; SWS, slow-wave sleep.

Ascending Effects

Waking and the 40 Hz Rhythm

The waking state is characterized by fast EEG oscillations in the 15–60 Hz range. The waking EEG is of low amplitude and composed of synchronization of fast rhythms or of high frequencies >15 Hz. Faster rhythms appear upon arousal and occur mainly in the 35–45 Hz range during selective attention to sensory input, referred to as the '40 Hz rhythm.' This rhythm is thought to be generated by cortical neurons receiving inputs from thalamocortical specific and nonspecific nuclei, such that the specific inputs to cortex provide the content of cognition (via input through primary sensory pathways), whereas the nonspecific inputs provide the context (via ILT), with their coherence providing the temporal binding required for the unity of cognitive experience.

The mechanism behind the 40 Hz rhythm appears to include layer IV cortical inhibitory interneurons, which exhibit a sequential activation of a persistent low-threshold sodium current followed by a subsequent potassium conductance. These cells can transmit the 40 Hz rhythm onto thalamic cells, allowing the entrainment of thalamocortical resonance. This 'ringing' is thought to be the basis of conscious perception so that for a transient sensory input to be perceived or consciously appreciated, it must generate a lasting reverberation or reentrant signaling in thalamocortical systems. RAS modulates these oscillations through its ascending projections (Figure 1).

The process of awakening entails two stages: a rapid (5 min) reestablishment of consciousness that is marked by increases in cerebral blood flow in RAS and thalamus, followed by a slower (15 min) increase in cerebral blood flow, primarily in anterior cortical regions. It should also be noted that cholinergic RAS neurons have some of the highest concentrations of nitric oxide synthase in the brain. Therefore, wherever PPN projects, a corollary effect may include vasodilation and other processes modulated by nitric oxide.

Slow-Wave Sleep

This state is characterized by high EEG with synchronization of low frequencies <15 Hz. These oscillations include spindles (7–14 Hz), delta waves (1–4 Hz), and slow oscillations (<1 Hz). This state is marked by decreases in cerebral blood flow throughout the brain, especially in RAS, thalamus, hypothalamus, and basal forebrain. It is characterized by decreased activity of cholinergic RAS, basal forebrain neurons, LH orexinergic cells, and TMN histaminergic cells.

The main excitatory projection systems decrease their outputs during slow-wave sleep.

Spindles Spindles are present early in sleep and the inhibition they promote basically deafferents the cerebral cortex (primary sensory pathways are gated), which may be a prerequisite for falling asleep. The reticular nucleus of the thalamus is thought to be the main cell group generating spindles, which result from bursting in a syncitium of electrically coupled reticular nucleus neurons that produce synchronized inhibitory postsynaptic potentials in thalamocortical neurons. These lead to postinhibitory rebound bursts that are relayed to cortex, producing slow, rhythmic, excitatory postsynaptic potentials. Spindles begin to occur as brain stem cell groups, particularly cholinergic RAS neurons, diminish their firing rates, disinhibiting reticular nucleus neurons.

Delta waves Delta waves increase gradually as slow-wave sleep deepens. These waves are generated by cortical and thalamocortical neurons that are synchronized by inhibitory reticular nucleus neurons and scattered dorsal thalamic cells. They represent waveforms elicited at a more hyperpolarized membrane potential than spindles so that as slow-wave sleep deepens, the occurrence of spindles decreases, whereas that of delta waves increases. These waves mark the virtual cessation of activity in mesopontine cholinergic cells.

Slow waves Slow waves appear to be generated spontaneously in cortical neurons, but this rhythm is also present in thalamic reticular nucleus neurons. All major classes of cortical cells display slow waves. The slow depolarization of these neurons corresponds to the depth-positive EEG wave, whereas their hyperpolarization corresponds to the depth-negative EEG wave. Because of the widespread synchronization of slow waves in cortical neurons, it is likely that thalamic neurons, specifically inhibitory reticular thalamic cells, help entrain these slow oscillations.

REM Sleep

Every 90 min during sleep, slow rhythms are interrupted by fast oscillations triggered by cholinergic mesopontine cell groups that induce REM sleep. The EEG displays low-amplitude synchronization of fast oscillations identical to those in the waking state. This 'paradoxical' state is accompanied by eye movements, contains most of our dreaming time, and leads to a decrease in skeletal muscle tone or atonia. The atonia is particularly strong on antigravity muscles, whereas the diaphragm and extraocular muscles retain substantial tone. REM sleep is characterized by increased blood flow in RAS, thalamus, anterior cingulate, and other cortical areas, with decreases in blood flow in the dorsolateral prefrontal cortex. It has been proposed that the unregulated (by other RAS nuclei) activity of RAS cholinergic cells is responsible for REM sleep and, via unknown mechanisms, for decreased frontal lobe blood flow. The 'hypofrontality' of REM sleep may account for the lack of critical judgment during dreaming (and during hallucinations).

In advance of and during REM sleep, paroxysmal events called PGO waves can be detected in the brain stem, the lateral geniculate nucleus, and occipital cortex, and they appear to be due to bursting activity in brain stem nuclei, including the PPN and especially the subcoeruleus. Interestingly, the 40 Hz cortical oscillation evident in waking is also present during REM sleep; however, it does not appear to be reset by sensory input, despite the fact that specific primary sensory thalamic systems are not gated. This suggests that the intrinsic activity of the brain does not place sensory input in the context of the functional state being generated by the brain. That is, REM sleep may be a state of intrinsic activity in which external sensory input cannot access thalamocortical reverberation or conscious experience.

It should be noted that transitions in state from waking to slow-wave sleep, or from slow-wave sleep to REM sleep, are gradual. During these transitions, waveforms characteristic of more than one state appear intermingled in the EEG. The frequency of these events increases until the new state is fully expressed. It appears that new states are gradually recruited rather than elicited instantaneously.

Muscle Tone and Locomotion

RAS is a phylogenetically conserved system that modulates fight-or-flight responses. During waking, our ability to detect predator or prey is essential to survival. RAS is linked to the motor system in order to optimize attack or escape. During REM sleep, atonia keeps us from acting out our dreams. Only our diaphragm and eye muscles appear to be acting out dream content. During waking, RAS can modulate muscle tone and locomotion via reticulospinal systems. For example, in a standing individual, there is tonic activation of antigravity, mainly extensor, muscles. Before the first step can be taken, there must be flexion (i.e., a release from extensor bias). The startle response, a rapid response to a supramaximal stimulus, is basically an induced flexor response, placing the body in a 'ready' position. The startle response is composed of a short latency

activation of muscle activity (the ready condition), followed by a brief inhibition (the 'reset' state) and then a long latency activation (the 'go' condition). The intermediate latency inhibition is thought to be part of the modulation of the startle response by cholinergic RAS neurons, and it may represent a 'resetting' of motor programs which allow the subsequent selection of response strategies – the triggering of attack or escape movements.

There appear to be multiple descending pathways by which RAS modulates muscle tone and locomotion. PIA can be activated by cholinergic agonists to induce REM sleep with atonia and PGO waves. Lesions of this region produce an animal exhibiting REM sleep without atonia, although such lesions may damage passing axons. Presumably, outputs from this region activate reticulospinal systems that lead to profound hyperpolarization of motor neurons, which is the mechanism responsible for the atonia of REM sleep. Cholinergic projections to the medioventral medulla appear to elicit increases in muscle tone and controlled locomotion. Outputs from this region activate reticulospinal systems that trigger spinal pattern generators to induce stepping. Electrical stimulation of the pontine and medullary reticular formation is known to induce decreased muscle tone at some sites while producing stepping movements at other sites. This suggests the presence of a heterogeneous, distributed system of muscle tone and locomotor control.

Neuropharmacology of Sleep and Arousal

Alcohol

Direct effects of alcohol on sleep–wake control regions appear to involve potentiation of GABAergic transmission. Significant impairment in motor performance, such as driving, occurs at very low blood alcohol concentrations, an effect potentiated by sleep deprivation or sleep loss. Most alcoholic patients suffer from insomnia, which is clinically important since alcoholism can exacerbate the adverse consequences of insomnia, such as mood changes and anxiety, and because insomnia has been associated with alcohol relapse.

Anesthetics and Sedatives

The proposed mechanisms of action of anesthetics have typically involved a number of cellular effects at different sites. Evidence suggests that the primary site of action of most anesthetics may be the sleep–wake control system. The parallel manifestations between sleep and anesthesia suggest that anesthetics basically 'hijack' the sleep–wake control system to induce anesthesia.

Most anesthetics, including barbiturates, etomide, propofol, neuroactive steroids, and volatile anesthetics, act on $GABA_A$ receptors, among others. Sedation and natural sleep occur mostly as a result of enhanced GABAergic transmission, which in turn affects the release of a number of excitatory transmitters, such as acetylcholine, excitatory amino acids, and histamine. These actions may take place specifically in such regions as RAS, TMN, and basal forebrain (all of which have local circuit GABAergic neurons and receive GABAergic input from VLPO), thereby regulating the level of arousal.

Interestingly, halothane and propofol appear to block gap junctions, and oleamide, which promotes sleep, is also known to block electrical coupling. Anandamide, which enhances adenosine levels to promote sleep, can block gap junctions. Carbenoxolone, a putative gap junction blocker, is known to decrease gamma band oscillations (high-frequency EEG), whereas connexin 36 (the neuronal gap junction protein) knockout mice exhibit low gamma band power. These findings all support a role for electrical coupling via gap junctions in the control of sleep–wake states and their modulation by certain anesthetic agents.

The benzodiazepines act by binding to a site that modulates GABA receptors, especially $GABA_A$ receptors. These agents produce sedative, hypnotic, anxiolytic, and anticonvulsant activities. They act generally by amplifying GABAergic transmission, such that short-acting agents have been used to promote sleep in insomnia patients.

Antihistamines

Pathology and lesions of TMN cause hypersomnia (these neurons are highly active during waking). Histaminergic inputs to RAS suppress slow-wave sleep and promote waking, although they do not affect REM sleep significantly. Administration of antihistamines (histamine receptor blockers) results in sedation due to blockade of histaminergic inputs to RAS, basal forebrain, and/or LH.

Stimulants

The popularity of caffeine is related to its stimulant properties, which are mediated by its ability to reduce adenosine release in the brain. Caffeine appears to block adenosine A1 and A2a receptors, producing a psychomotor stimulant effect.

Amphetamine induces the release of monoamines, especially dopamine, but also blocks reuptake and may have neurotoxic effects on nigral neurons. Methylphenidate does not appear to have such neurotoxic

effects. Amphetamine psychosis occurs in two forms – acute intoxication after a single large dose (characterized by confusion and disorientation) and chronic abuse that produces a schizophrenia-like syndrome. The increased release of dopamine by amphetamine is considered a model of schizophrenia, and it led to the 'dopamine theory' of schizophrenia.

Modafinil is a newer stimulant that does not appear to act through dopaminergic mechanisms, and it has been found to be effective in the treatment of daytime sleepiness in patients with narcolepsy. Modafinil has also been reported to alleviate spatial neglect and to increase vigilance in stroke patients. This suggests that sensory neglect arising from stroke can be treated successfully by 'waking up' the involved cortex, implying that the consequences of stroke may be to decrease activity, blood flow, and/or metabolism, one or more of which modafinil may counteract.

Development

The human newborn sleeps 16 h per day – 8 h in REM sleep and 8 h in slow-wave sleep. Between birth and the end of puberty, there is a developmental decrease in REM sleep such that the adult has 1 h of REM sleep and 6 or 7 h of slow-wave sleep per day. That is, the decrease in REM sleep is replaced by a postpubertal increase in waking. During development, REM sleep may serve to direct brain maturation, providing endogenous stimulation at a time when the brain has little or no exogenous input. However, REM sleep is not essential for survival in the adult. REM sleep suppression does not lead to death or disease. Evidence suggests that a number of transmitter systems change during the developmental decrease in REM sleep and may modulate that decrease. In addition, levels of the gap junction protein connexin 36 are high before the developmental decrease (10 days in the rat) and are considerably lower at the end of the decrement (30 days in the rat). It thus appears that there is a large syncitium of coupled neurons which segregates into distinct functional networks with age; however, the capacity to revert to a developmental condition may be present in the adult, leading once more to increased REM sleep drive.

Arousal and Sleep Disorders

The connectivity of RAS shown in Figure 2 can be used to deduce the potential sites of dysregulation related to hypervigilance and increased REM sleep drive that are evident in depression, anxiety disorder, and schizophrenia.

The treatment of depression previously included tricyclic antidepressants that mainly blocked reuptake of noradrenaline and serotonin and blocked histamine and acetylcholine release, thus accounting for increased sleepiness. The selective serotonin reuptake inhibitors more selectively affect RAS by increasing the inhibition of PPN and LC neurons (Figure 2). It is not clear if the etiology of depression is related to disinhibition of PPN and LC by a decrement in serotonergic tone, although this would seem a likely origin for the sleep–wake disturbances seen in depression.

The treatment of anxiety disorder has involved the use of benzodiazepine amplification of GABAergic inhibition. In addition, the use of the α_2 receptor agonist clonidine produces anxiolytic effects, probably by inhibiting autoreceptors in LC and postsynaptic receptors in PPN, thus downregulating vigilance. The etiology of anxiety disorder has been proposed to include downregulation or degeneration of LC outputs (possibly induced by stress hormones), which would act to disinhibit PPN neurons.

The treatment of schizophrenia previously involved the use of the dopaminergic receptor blocker haloperidol, which induced tardive dyskinesia, among other side effects. Newer antipsychotics appear to block dopaminergic, noradrenergic, and serotonergic receptors. More striking antipsychotic effects resulted from the use of clozapine, which was designed as a muscarinic cholinergic blocker for the treatment of Parkinson's disease. A later generation agent that has maintained anticholinergic activity without this side effect is olanzapine. These agents appear to have partial penetrance at serotonergic, cholinergic, and dopaminergic receptors, basically reducing muscarinic cholinergic activation of the substantia nigra (SN), as well as partially blocking dopaminergic actions in the striatum. The etiology of schizophrenia has been suggested to include increased PPN output, accounting for marked hypervigilance and hallucinations (REM sleep intrusion into waking). Excessive PPN output would overactivate the SN and, in turn, increase striatal release of dopamine, complying with the dopamine theory of schizophrenia.

Disturbances in RAS function are evident in a number of parasomnias that are discussed in other articles in this encyclopedia. Suffice it to say that RAS dysregulation is part of the psychopathology in a host of disorders and may even presage the onset of some of these disorders.

See also: Dopamine Control of Arousal; Narcolepsy; Pharmacology of Sleep: Adenosine; Reticulospinal System; Sleep and Sleep States: Thalamic Regulation; Sleep and Sleep States: Hypothalamic Regulation; Sleep and Sleep States: Histamine Role; Sleep Oscillations and PGO Waves; Sleep in Adolescents; Sleep in Aging;

Startle Response; Stimulant and Wake-Promoting Substances; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Chase MH and Morales FR (1994) The control of motoneurons during sleep. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, pp. 163–176. London: Saunders.

Datta S and Patterson EH (2003) Activation of phasic pontine wave (p-wave): A mechanism of learning and memory processing. In: Maquet J, Stickgold R, and Smith C (eds.) *Sleep and Brain Plasticity*, pp. 135–156. Oxford: Oxford University Press.

Garcia-Rill E, Kobayashi T, and Good C (2003) The developmental decrease in REM sleep. *Thalamus & Related Systems* 2: 115–131.

Greene R and Siegel J (2004) Sleep. A functional enigma. *Neuro-Molecular Medicine* 5: 59–68.

Heister DS, Hayar A, Charlesworth A, Yates C, Zhou Y, and Garcia-Rill E (2007) Evidence for electrical coupling in the SubCoeruleus (SubC) nucleus. *Journal of Neurophysiology* 97: 3142–3147.

Jouvet M (1999) *The Paradox of Sleep. The Story of Dreaming*. Cambridge, MA: MIT Press.

Kilduff TS and Peyron C (2000) The hypocretin/orexin ligand-receptor system: Implications for sleep and sleep disorders. *Trends in Neurosciences* 23: 359–365.

Kleitman N (1963) *Sleep and Wakefulness*. Chicago: University of Chicago Press.

Llinas R, Ribary U, Joliot M, and Wang XJ (1994) Content and context in temporal thalamocortical binding. In: Buzsaki G (ed.) *Temporal Coding in the Brain*, pp. 251–272. Berlin: Springer.

Lydic R, Baghdoyanm HA, and McGinley J (2003) Opioids, sedation and sleep. Different states, similar traits, and the search for common mechanisms. In: Malviya S, Naughton N, and Tremper KK (eds.) *Contemporary Clinical Neuroscience: Sedation and Analgesia for Diagnostic and Therapeutic Procedures*, pp. 1–31. Totowa, NJ: Humana Press.

Moruzzi G and Magoun HW (1949) Brain stem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology* 1: 455–473.

Siegel JM (1994) Brainstem mechanisms generating REM sleep. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, pp. 125–144. London: Saunders.

Steriade M and McCarley RW (2005) *Brain Control of Wakefulness and Sleep*, 2nd edn., New York: Springer.

Reticular Formation and the Brain Stem

A B Scheibel, Brain Research Institute UCLA, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

The reticular formation of the brain stem is the most ancient part of the brain and the one part that is absolutely necessary to the life of the organism. It consists of large numbers of neurons and fibers and – exclusive of the adjacent cranial nerve nuclei, long ascending and descending tracts, and a few large nuclear masses such as the olives, nucleus ruber, and substantia nigra – it constitutes the core of the brain stem. It is generally thought of as extending from the caudal medulla to the interface between mesencephalon (midbrain) and diencephalon (thalamus–hypothalamus). However, a series of cell masses clustered in the medial portions of the thalamus continues forward, forming the thalamic nonspecific system and, as such, can be thought of as the reticular formation of the thalamus.

In addition, a thin, sheet-like layer of neurons surrounding the lateral surfaces of the thalamus is known as the nucleus reticularis thalami or reticular nucleus of the thalamus. This functionally distinct cell ensemble should not be confused with the others.

Terminology for the system is far from satisfactory. Reticular refers to the net-like pepper-and-salt appearance of these cell-fiber fields as they were visualized in tissue stains prepared by neuroanatomists over a century ago. More recent terms include nonspecific system and ascending reticular activating system (ARAS). The term nonspecific differentiates these cell groups from the specific, earlier-recognized ascending sensory relay systems such as the dorsal column-medial lemniscal systems of the brain stem. Activation refers to the functional capability of significant portions of the system to activate (desynchronize) electrocortical rhythms. The latter is obviously a more functionally meaningful term, based as it is on important experimental studies performed in the 1940s and 1950s. However, presently developing knowledge of the functional repertoire and range of structural patterns argues for the inadequacy of any and all of these names. I suggest the use of the functionally noncommittal term brain stem reticular core, combining, as it does, aspects of geographic position with historical precedent.

Histologic Structure

The core is made up of interneurons of various sizes and shapes, most of which are multipolar in appearance, with radiating, relatively unramified dendrites, a pattern which has been called idiodendritic. No typical short-axoned cells have been identified, although many of the more lateral-lying core neurons have axons with rather limited trajections. The majority of cells send their axons appreciable distances up or down the brain stem and beyond. The range and pattern of these projections provide some measure of the enormous scope of reticular core influence over the remainder of the central nervous system (Figure 1).

A number of subnuclei have been delineated on purely morphologic and cytoarchitectonic grounds. The significance of such structural parcellation is not entirely clear, especially in view of the mutual interpenetration of such cell fields by dendrites from adjacent, or even distant, ensembles. In addition, more meaningful characterization of cell groups on the basis of enzyme or neurotransmitter content and intramembraneous receptor arrays is now possible with fluorescence and immunocytochemical techniques. Selective microstimulation of small cell groups in the unanesthetized animal or in tissue slabs maintained under *in vitro* conditions provides further evidence as to the functional role of cell assemblies in the core. Within this broad frame of possibilities, it seems advisable to review the major cell groups as structurally and chemically identifiable entities, and then to sketch in their functional roles wherever possible.

Small to medium neurons characterize posterior medulla, upper pons, and mesencephalon, whereas large and even giant reticular cells (nucleus reticularis gigantocellularis and nucleus pontis caudalis) are centrally located in the upper medulla and lower pons. These large neurons (somata up to 80 μm diameter) often possess large bifurcating axons whose branches may ascend as far as diencephalon or descend into spinal cord. Scattered cell groups found along the midline of medulla, pons and mesencephalon constitute the raphe nuclei of the reticular core. Although these large bidirectional axon systems are dramatically presented in Golgi silver impregnations, newer retrograde double labelling studies with fluorescent tracers and microelectrode experiments using antidromic invasion techniques indicate that not more than 5% of reticular neurons generate such axon

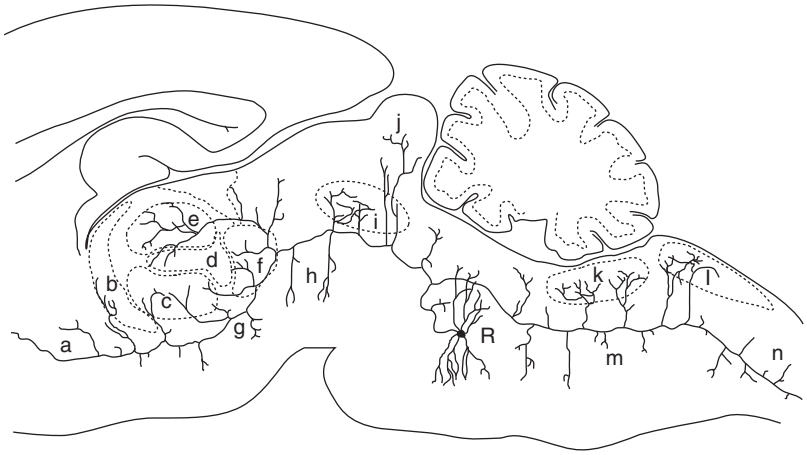

Figure 1 The axonal path of a single large reticular neuron R in the nucleus reticularis magnocellularis of the lower pons. Collateral and terminal branches reach synaptic targets in the basal forebrain region, a; the nucleus reticularis thalami, b; central nuclear masses of the thalamus, c, d, e; the centre median-parafascicular complex, f; subthalamus, g; mesencephalic tegmentum, h; cranial nerve nuclei III and IV, i; inferior colliculus, j; cranial nerve nucleus XII, k; nucleus cuneatus, l; medullary reticular formation, m; spinal cord, n. Reproduced from a drawing in Brazier M (1968) *The Electrical Activity of the Nervous System*, 3rd edn., Battimore: Williams and Wilkins, with permission.

systems. Many of these cells, which are closely applied to small blood vessels, project their axons to the dorsal horn of the spinal cord, where they are believed to modulate incoming pain stimuli. Cell groups flanking the raphe system at pontine levels and grouped under the general rubric of pararmedian pontile nuclei have been found to enter complex synaptic relationships with the optic tectum (midbrain) and those cranial nerve nuclei (III, IV, and VI) that mediate eye movements. As a substrate for visual tracking behavior, this reticular zone is of obvious importance in maintaining reflex orientation in a rapidly changing world. Operating in a somewhat different physiologic mode, it is also involved in the rapid eye movements characteristic of REM sleep.

A potentially delimitable family of nuclei consists of those whose major axonal projections are directed primarily toward the cerebellum, rather than up and down the brain stem. These include the lateral reticular nucleus of the caudal medulla, the nuclei prepositi hypoglossi of the dorsomedial medulla, and the nucleus reticularis tegmenti pontis of the ventral pons. In view of the widespread distribution of most axonal systems of reticular cells, it seems unwarranted to single out any cell group on the basis of axonal trajectory alone.

Groups of reticular neurons have been characterized chemically at virtually all levels of the reticular core. One group of aminergic nuclei exerts profound effects over the entire neuraxis. Scattered along the midline from caudal medulla to rostral mesencephalon are groups of raphe cells (reticular cell groups B1–B9) rich in the indoleamine serotonin. These neurons, which appear to constitute one of the major central nervous system sources for this neurotransmitter, project widely upon the diencephalon and forebrain. Descending projections to the spinal cord terminate in significant measure on neurons of laminae 1, 5, and 6, where they are believed to exert suppressive influences on incoming pain impulses.

Equally potent ensembles of catecholaminergic neurons (reticular cell groups A1–A12) containing norepinephrine or dopamine are also found at various stations throughout the reticular core from the caudal medulla through the midbrain. High concentrations of dopamine and one of its synthesizing enzymes, tyrosine hydroxylase (TH), are found in neurons of the substantia nigra and in perinigral cells of the surrounding brain stem core, the ventral tegmental area of Tsai (cell groups A9–A10). While nigral neurons project primarily to the corpus striatum, axons of the tegmental cells reach many parts of the limbic system, including the nucleus accumbens septi, nucleus of the stria terminalis, and olfactory tubercle. Highest concentrations of norepinephrine are found in a compact group of deeply pigmented cells, the locus coeruleus, situated in the dorsolateral portion of the pontine reticular core and known as cell group A6. From this relatively small cell cluster (only 30 000 cells per side) a remarkably rich axon system spreads out through virtually the entire central nervous system, including diencephalon, limbic system, parts of neocortex, brain stem, cerebellum, and spinal cord. Significant and largely caudally directed projections tend to characterize the more caudally situated noradrenergic cell groups of the medulla and lower pons. The locus coeruleus has proved of particular interest because of its apparent functional relationship to modulation of cortical activity pattern and behavioral attentiveness and to sensory and motor neurons

throughout the brain stem and spinal cord. Its putative relationship to the capillary bed of the neuraxis, coactively with the dopaminergic and serotonergic systems, has recently become a topic of interest.

Several ascending, acetylcholine (ACh)-rich systems have also been identified within the reticular core. Parent cell clusters in the dorsal midbrain supply much of the tectum, thalamus, and globus pallidus. More ventrally placed cell groups in the midbrain and pons project at least as far rostrally as the thalamus, hypothalamus, preoptic area, and the basal forebrain, the last constituting another significant site of cholinergic neurons. The role of these acetylcholine-rich systems in cortical activation, orientation, and memory processing is receiving increasing attention, especially in light of recent evidence suggesting progressive declines in ACh concentration in aging and more extreme loss of the neurotransmitter in senile and presenile dementias.

Although not strictly located within the reticular core, the histaminergic projection system of the tuberomamillary portion of the hypothalamus should also be included with this group of neurotransmitters because of its apparent conjoint role in cortical activation and arousal.

It seems increasingly likely that the cortical projections of noradrenergic, dopaminergic, and cholinergic systems are bilaterally asymmetric. Dopaminergic and cholinergic projections appear denser in the left hemisphere, whereas noradrenergic systems seem to have greater representation on the right. It is tempting to hypothesize that this type of pattern may be responsible for the now well-recognized affectual depression and concern that usually accompanies cerebrovascular accidents in the left anterior hemisphere compared with the lack of apparent concern and general sense of equanimity that comes with similar episodes in the right hemisphere.

Physiologic Effects of Reticular Cell Activity

From a phenomenologic point of view, it appears that catecholamines such as norepinephrine and dopamine may change the level of gain of cortical cells. Some studies suggest that both act to diminish the responsiveness of pyramidal and stellate cells, with greater decrease being described in 'spontaneous firing' than in evoked discharge. The net result of this would be to improve the signal-to-noise ratio and, accordingly, the fidelity of cortical cell response. Although more work is necessary in this important area, one recent study suggests that dopamine exerts this effect on pyramidal cells and norepinephrine on stellate and granule cells.

The actual cellular mechanisms substrating these effects are also not completely clear. However, there are data that indicate that the 'arousal system' alters activity of thalamic and cortical neurons by blocking potassium conductance. In the process, slow depolarization may be effected and with it a change from burst firing mode to single spike firing. However, it also seems increasingly clear that there is continuous interaction among the multiple neurotransmitters, so that postulation of single cause-and-effect relationships in the intact cortex (as opposed to *in vitro* preparations) may prove overly simplistic. Experimental evidence suggests that neurotrophins operating at a pontine level may also be involved in changing cortical – and therefore behavioral – states.

As one example of neuromodulator interaction, locus coeruleus noradrenergic neuronal activity is usually considered to accompany high attention levels. However, sustained high levels of attention have recently been found to be accompanied by low levels of locus coeruleus activity and enhanced levels of cholinergic activity. It is conceivable that the initial peak of noradrenergic activity accompanying the original arousal component of attention rapidly attenuates as cholinergic activity levels increase to sustain the attentive mode. My colleagues and I suggest that a number of synergistic patterns of this type characterize the interplay of reticular core modulatory systems on the rest of the central nervous system.

Input–Output Patterns

Cell masses of the reticular core receive a broad range of information from impinging collateral and terminal fibers. Microphysiologic analyses of single-cell responses emphasize the heterogeneity of the convergent afferent terminals, which may include ascending somethetic and interoceptive sources and descending cortical and subcortical, visual, auditory, reticular, and cerebellar sources. However, preference patterns for certain afferent inputs have been identified. For instance, the face is the body part most heavily represented in the lower two-thirds of the brain stem, and its distribution pattern is usually ipsilateral. Hindlimbs, trunk, and tail are least frequently represented and are usually found more rostrally and in diffuse patterns with bilateral representation. Reticular neurons also show remarkable adaptive or habituative response to iterative stimuli. Experimental studies of spike responses of single cells to repetitive stimulation of the sciatic nerve, at $2\,\mathrm{V\,s^{-1}}$, show a decline in responsiveness to 10% of initial values within 60s. This type of plastic response contrasts strongly with that of visual cortical cells, which show response habituation patterns of no more than

5–15%. It emphasizes the role of reticular cells in the recognition of the immediacy and relevance of stimulus patterns.

Powerful descending (reticulospinal) fiber systems modulate ongoing activity of spinal (and brain stem) motor neurons. The stimulation studies of Magoun and Rhines identified inhibitory zones in the medial portion of the medulla and facilitatory areas in the dorsolateral medulla and caudal pons. More detailed investigations indicate that stimulation of the pontine reticular formation evokes monosynaptic excitatory postsynaptic potentials in spinal motoneurons, whereas stimulation of the caudal half of the medullary reticular formation results in large, rather long-latency inhibitory postsynaptic potentials in motoneurons. This pontomedullary tissue continuum is a mosaic of functional zones, in that most caudal areas seem to act selectively on axial musculature of the neck and back, whereas more rostral and central zones exert increasingly diffuse effects on spinal motor outflow. Reasonably localized 'centers' for control of respiration, licking, locomotor patterns and 'species-specific calling' have been described.

Ascending reticular projections diverge at the junction of mesencephalon and diencephalon. A dorsal leaf enters the thalamus and terminates among the cell masses of the thalamic nonspecific system and as far rostrally as the nucleus reticularis thalami. The ventral leaf projects through the subthalamus and portions of the dorsal-lateral hypothalamus, reaching the septum and basal forebrain. Approximately 10% of these fibers continue on into the cerebral cortex, terminating in greatest concentration in the anterior third of the cerebral hemisphere. Converging lines of evidence point to the conjoint operation of the dorsal leaf with thalamic nonspecific neurons and their thalamocortical projections as the most likely source of cortical synchronous wave activity, such as the alpha rhythm and spindle waves. Conversely, the ventral leaf seems primarily responsible for driving the cortex at faster rhythms (i.e., low-voltage fast activity characteristic of the attentive conscious state and of REM sleep). Ascending projections from the raphe complex of the reticular formation via the medial septum-diagonal band of Broca exert control over hippocampal theta frequencies and thus can influence hippocampal information processing in response to particular behavioral states.

Thalamic Extensions of Reticular Formation

Surrounding the third ventricle is a central core of thalamic cells with distinctive structural and functional properties that differentiate it from the mass of the surrounding thalamus. The largest and most caudal elements are the paired center median and parafascicular nuclei (CMPF complex), which constitute a prominent landmark of the posterior thalamus. A group of thin, vertically oriented nuclei (the medial and midline nonspecific nuclei) and their dorsolateral wing-like extensions in the internal medullary lamina, the intralaminar nuclei, make up the remainder of the thalamic nonspecific system. Complex synaptic patterns with cerebral cortex, basal ganglia, diencephalon, and brain stem mark these still-enigmatic cell fields. A rather extensive but as yet unsatisfying literature points to their role as pacemaker for thalamocortical electrical patterns and, in a larger sense, for the fine-tuning of conscious states.

The nucleus reticularis thalami, an embryologic derivative of the subthalamus, consists of a thin layer of elongated cells lying within the external medullary lamina along the lateral surface of each thalamus. In this situation, the nucleus sits astride almost all connections between upper brain stem and cortex, and projects its axons back upon mesencephalic and thalamic sources of these fibers. Electrophysiologic and immunocytochemical studies reveal it to be a γ-aminobutyric acid (GABA)-rich feedback system whose facilitation by ascending thalamic volleys results in characteristic high frequency bursts of activity in reticularis cells which, in turn, inhibit the thalamic sources of these volleys. Intrinsic activity of ensembles of reticularis cells appears to serve as source of thalamic spindle bursts consonant with drowsing and light slow wave sleep activity in the organism. The blockade of these synchronized oscillations upon arousal appears to result from increased activity in cholinergic midbrain reticular neurons, which not only stimulate relay and thalamic reticular cells but also suppress depolarizing oscillations of reticularis neurons by inducing long-lasting hyperpolarization associated with enhanced K^+ conductance. Connectional specificity among the mesencephalic, thalamic and cortical loci playing upon this sheet of cells enhances the precision of gating effects exerted by the nucleus reticularis. It is of further interest that the prefrontal cortex also modulates the activity of reticularis neurons by way of the thalamic 'nonspecific' nuclei, thereby providing a putative override at the highest cognitive levels. Here, perhaps, lie mechanisms that enable not only selective focusing of the conscious state, but also the suprarogative control over body functions and painful states achieved at certain levels of concentration, meditation, or hypnosis.

See also: Brainstem and Cranial Nerves; Brainstem Control of Eye Movements; Reticular Activating System;

Reticulospinal System; Spinal Cord Pain Systems; Thalamus: Evolution in Vertebrates.

Further Reading

Berlucchi G (1997) One or many arousal systems? Reflections on some of Giuseppe Moruzzi's foresights and insights about the intrinsic regulation of brain activity. *Archives Italiennes de Biologie* 135(1): 5–14.

Brazier M (1968) *The Electrical Activity of the Nervous System*, 3rd edn. Baltimore: Williams and Wilkins.

Brodal A (1957) *The Reticular Formation of the Brain Stem; Anatomical Aspects and Functional Correlations*. London: Oliver and Boyd.

Denham MJ and Borisyuk RM (2000) A model of theta rhythm production in the septal-hippocampal system and its modulation by ascending brain stem pathways. *Hippocampus* 10(6): 698–716.

Jones BE (1991) The role of noradrenergic locus coeruleus neurons and neighboring cholinergic neurons of the pedunculopontine tegmentum in sleep wake states. *Progress in Brain Research* 88: 533–543.

Marrocco RT, Witte EA, Davidson MC, et al. (1994) Arousal systems. *Current Opinion in Neurobiology* 1: 166–170.

Moruzzi G and Magoun HW (1949) Brain stem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology* 1: 455–473.

Scheibel AB (1984) The brain stem reticular core and sensory function. In: Darian-Smith F (ed.) *Handbook of Physiology Vol. 1, The Nervous System III*, pp. 213–256. Bethesda, MD: American Physiological Society.

Sirota MG, Di Prisco GV, Dubuc R, et al. (2000) Stimulation of the mesencephalic locomotor region elicits controlled swimming in semi-intact lampreys. *European Journal of Neuroscience* 12(11): 4081–4092.

Steriade M and Llinas RR (1988) The functional states of the thalamus and the associated neuronal interplay. *Physiology Reviews* 68: 649–742.

Yingling CD and Skinner JE (1977) Gating of thalamic input to cerebral cortex by nucleus reticularis thalami. In: Desmedt JR (ed.) *Progress in Clinical Neurophysiology, Vol. I, Attention, Voluntary Contraction and Event-Related Cerebral Potentials*, pp. 70–96. Basel: Karger.

Reticulospinal System

J A Buford, Ohio State University, Columbus, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

The reticulospinal system descends from neurons in the brain stem reticular formation to the spinal cord. The most commonly accepted divisions are the medial and lateral reticulospinal tracts. The medial reticulospinal tracts originate mainly from nucleus reticularis pontis caudalis (NRPc) and pontis oralis (NRPo), and from the dorsal part of nucleus reticularis gigantocellularis (NRGc). The lateral reticulospinal tracts originate mainly from the more ventral parts of NRGc and from nucleus reticularis magnocellularis (NRMc). The general locations of these nuclei in the monkey are illustrated in Figures 1(a) and 1(b). Taken together, these regions are called the pontomedullary reticular formation (PMRF). The reticular formation of the brain stem also includes sections in the rostral pons and mesencephalic reticular formation that project to the PMRF as well as to the cervical spinal cord, although these spinal projections are sparse. The focus of this article is the reticulospinal fibers from the PMRF. These fibers are the main sources of the reticulospinal system's effects on motor control of the neck, trunk, and proximal limb muscles.

As a whole, the reticulospinal system is notable for it influence over whole-body movement patterns. For example, postural responses engaging the trunk and limbs and the ultimate whole-body movement pattern, locomotion, depend heavily on reticulospinal control. Head and neck movements are also heavily dependent on reticulospinal control. In addition, the reticulospinal system is involved in the support of skilled voluntary movements such as reaching. Just as there are widespread outputs from the reticulospinal system, there are also widespread inputs. The cerebral cortex, cerebellum, vestibular system, auditory system, somatosensory systems, and other brain stem nuclei, notably the mesencephalic locomotor region (MLR), are the major inputs to the reticulospinal system.

Anatomy of the Reticulospinal System

Medial and Lateral Reticulospinal Tracts – Origins

The medial and lateral reticulospinal tracts are named for the locations of their descending axons in the white matter of the spinal cord, as illustrated in Figure 1(c). The medial reticulospinal tract is concentrated in the ventromedial funniculus, somewhat ventral to the medial vestibulospinal tracts. The lateral reticulospinal tracts descend in the ventrolateral funniculus, mostly dorsal to the lateral vestibulospinal tracts, and also to some extent in the dorsolateral funniculus. There is not an absolute segregation of the cells of origin for the medial and lateral reticulospinal tracts, but there is general separation. The medial reticulospinal tracts originate mainly from the NRPc and dorsal NRGc, with some from the caudal part of the NRPo. The lateral reticulospinal tracts originate mainly from the ventral part of the NRGc and NRMc. Brodal referred to this latter combination as the nucleus reticularis ventralis and did not specifically recognize the NRMc. Even with these rough designation, the cells of origin of the two tracts are somewhat spread out and overlap with one another in the reticular formation. To state this another way, just as the nuclear boundaries of the parts of the reticular formation such as the NRPc and NRGc are somewhat indistinct, so are the exact origins of the medial and lateral reticulospinal tracts.

Part of the ventromedial system The reticulospinal system is a ventromedial system according to Kuypers's classification. This is a general nomenclature for systems descending in the ventral and medial white matter of the primate spinal cord, in contrast with the lateral descending systems (e.g., the lateral corticospinal tract) descending in the lateral funniculus. Contrasting the general characteristics of the ventromedial versus lateral descending systems provides a useful framework for understanding the reticulospinal tracts.

Consider the lateral systems for a moment. A typical axon of the lateral descending system terminates in a few spinal segments concentrated at a certain level of the spinal cord. Although most descending axons use interneurons, the lateral descending systems tend to make strong contacts on premotor interneurons that are only one synapse away from the motor neurons. In higher primates, lateral systems terminate heavily in the lateral part of the ventral horn, where distal muscles are represented. Monosynaptic contacts from corticospinal and rubrospinal cells of the lateral system are common for the distal muscles of fine control in primates. This gives the lateral systems the anatomical substrate for specific control over individual movement components, such as a relatively independent control of finger movements.

The reticulospinal system is virtually the opposite. Individual reticulospinal axons tend to make contacts at multiple levels of the spinal cord, sometimes including the cervical, thoracic, lumbar, and sacral segments. These terminations tend to be medial in the gray matter

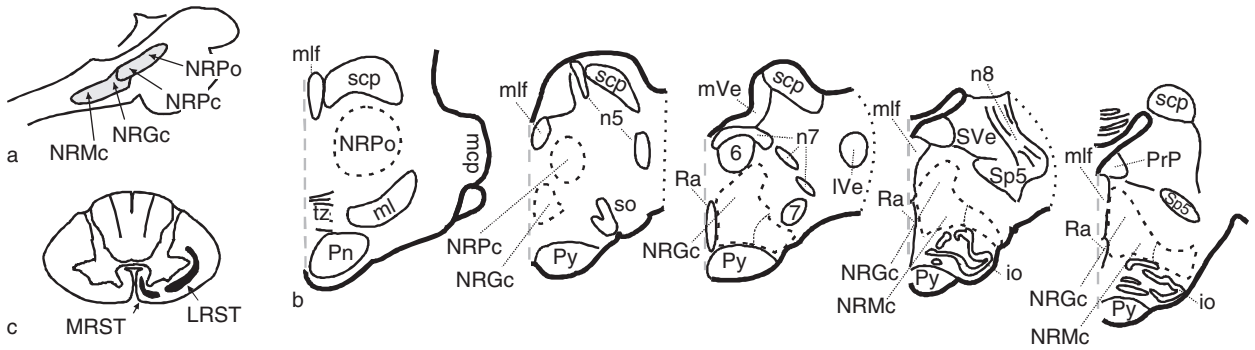

Figure 1 Anatomy of the monkey reticulospinal systems: (a) sagittal view of the brain stem sowing the general locations of the cells of origin for the reticulospinal tracts; (b) frontal sections showing the locations of the key nuclei; (c) the location of descending fibers in the white matter of the spinal cord for the medial reticulospinal tract (MRST) and lateral reticulospinal tract (LRST). These diagrams are from the monkey, but data from the cat are similar. 6, nucleus of 6th cranial nerve (abducens); 7, nucleus of 7th cranial nerve (facial); io, inferior olive; lVe, lateral vestibular nucleus; ml, medial lemniscus; mlf, medial longitudinal fasciculis; mVe, medial vestibular nucleus; n5, 5th nerve; n7, 7th cranial nerve; NRGc, nucleus reticularis gigantocellularis; NRMc, nucleus reticularis magnocellularis; NRPc, nucleus reticularis pontis caudalis; NRPo, nucleus reticularis pontis oralis; Pn, pontine nuclei; PrP, nucleus prepositus; Py, medullary pyramids; Ra, Raphe nuclei; scp, superior cerebellar peduncle; so, superior olive; Sp5, spinal nucleus of 5th cranial nerve; SVe, spinal vestibular nucleus; tz, trapezoid body.

or in the intermediate zone. Although there are some monosynaptic contacts on motor neurons, most reticulospinal contacts tend to be on interneurons, presumably providing polysynaptic control over the motor system. When present, the monosynaptic contacts on motor neurons are probably mostly on the distal dendrites. Hence, the anatomical substrate of the reticulospinal system tends to favor whole-body movement patterns. The muscles around the spinal column (axial muscles) and in the hip and shoulder girdles tend to be most strongly influenced by the reticulospinal systems, but there are also projections affecting muscles of the limbs. Thus, not only can the reticulospinal system contribute to the control of posture, it can also produce movements of the limbs.

Laterality and terminations of the medial versus lateral reticulospinal tracts Before defining the laterality of the reticulospinal tracts, it is important to be clear about what is meant by the laterality of a descending system. Anatomy is typically used to define laterality. If a reticulospinal cell on the left side of the brain stem produces an axon that descends and makes synaptic contacts exclusively on the left side of the spinal cord, the cell is called ipsilateral; however, if those contacts are predominately on interneurons that cross the midline to exert their influence on muscles of the other side, then the functional effects could be contralateral. For the developmental biologist, the anatomical feature may be most important, but for the systems physiologist, the function may be the key issue. Ultimately, when controlling axial muscles, bilateral influences at a functional level are probably the norm for the reticulospinal tracts; it makes little functional sense to regulate muscles on the one side of the spinal column without coordinated control of the muscles on the other side. As the details of anatomical and physiological data are considered, it is important to keep this in perspective.

Medial Reticulospinal Tracts

Neurons contributing to the medial reticulospinal tract make synaptic contacts throughout the spinal cord from the cervical through lumbosacral levels, often making contacts ipsilaterally and contralaterally through axon collaterals that decussate. Cells originating in the pons, especially those originating rostrally, descend ipsilaterally. Those that originate in the caudal pons and rostral medulla can descend either ipsilaterally or contralaterally. Those that descend contralaterally decussate in the medulla around the level of the soma itself. The axons of these neurons have a variety of sizes (3–10 μm) and conduction velocities (10–120 m s^{-1}).

The terminals of the medial reticulospinal tracts are in lamina VII and especially VIII of the spinal gray matter, where axial and proximal limb motor control is concentrated. Many of contacts are on commissural interneurons in the spinal gray. Although cells descending ipsilaterally or contralaterally can both contact commissural interneurons, reticulospinal cells descending contralaterally are somewhat more likely to engage the commissural interneurons. These commissural interneurons decussate and recruit motor outputs on the contralateral side.

Lateral Reticulospinal Tracts

The lateral reticulospinal tract is a bilateral system. Like the medial reticulospinal tract from the NRGc,

neurons of the lateral reticulospinal tract either descend ipsilaterally or decussate in the medulla and descend contralaterally. The lateral reticulospinal tract appears to make more limited contacts at only a few spinal segmental levels at a time. Its contacts are in lamina VI, VII, and VIII. In general, the projections of the lateral reticulospinal tract are slightly more dorsal in the gray matter of the spinal cord, suggesting more indirect control over movement. Axons of the lateral reticulospinal tract are mainly small (<3 μm in diameter) and relatively slowly conducting (10–20 m s^{-1}).

In sum, both anatomically and functionally, the reticulospinal tracts have bilateral influence over motor circuits controlling the axial and proximal limb muscles on both sides of the body. The substrate at a single-cell level on with this system operates is whole-body movement engaging both sides of the body, often engaging the upper and lower limbs at once.

Inputs to the Reticulospinal System

The reticulospinal system receives inputs from the cortical, tectal, somatosensory, vestibular, and cerebellar centers. Thus, vestibular systems can recruit reticulospinal outputs as a part of whole-body movement patterns necessary in postural responses, and tectal influences may use this avenue as part of the head and neck movements associated with gaze control. Midbrain projections to the reticulospinal system also originate from the MLR. Cerebellar influences may also engage trunk and proximal limb control, as well as anticipatory postural adjustments in advance of movement and locomotion. Cortical projections come mainly from the motor and premotor cortical areas. Ascending influences from somatosensory systems are bilateral in origin and are strongest for proprioceptors.

Cortical Inputs

Corticoreticular projections to the reticulospinal system emanate from the primary motor cortex (area 4) and the premotor cortical areas (area 6). Some corticoreticular projections are collateral branches of corticospinal neurons. These collaterals ascend from the pyramids up into the PMRF. Other corticoreticular projections are direct projections that are not part of the corticospinal tract, and these also ascend from the pyramids. In cat and monkey, anterograde labeling indicates that pure corticoreticular projections come from the more rostral parts of the cortical motor areas, with the strongest contributions from area 6 (Figure 2). The area richest with corticospinal neurons' delivering collaterals to the reticulospinal systems appears to be the transitional zone at the rostral boundary of area 4, where the distinction between area 4 and area 6 is sometimes difficult. In the more caudal regions of area 4, where corticospinal control of distal muscles is concentrated, there are fewer cortical projections to the reticulospinal system.

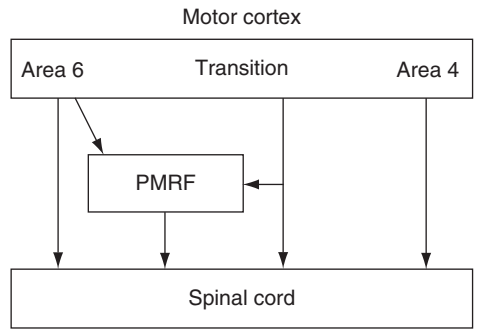

Figure 2 Corticoreticular projections from the motor cortex in relation to corticospinal fibers and cortical motor areas. Area 6 includes the supplementary motor area and the premotor cortex. This area contributes corticospinal fibers and corticoreticular fibers, but these corticoreticular fibers are relatively small and are not usually collaterals of corticospinal fibers. The transition is the region at the rostral end of area 4, the primary motor cortex, and the caudal end of area 6 where the anatomical distinction between these areas can be difficult. Corticospinal tract fibers originating from this transitional area are likely to deliver collaterals to reticulospinal neurons. From the caudal part of area 4 in the monkey, corticospinal fibers tend not to send collaterals to the reticulospinal system. PMRF, pontomedullary reticular formation.

In the most rostral regions of the PMRF, in the NRPo, the corticoreticular projection may be mostly ipsilateral and tends to be from smaller-diameter fibers originating in area 6. But for the majority of the reticulospinal system, in the NRPc, NRGc, and NRMc, corticoreticular projections are bilateral and come from both area 4 and area 6. Projections from the upper limb region of the motor cortex appear slightly more numerous than projections from the lower limb area, suggesting that the corticoreticulospinal system may be somewhat biased in favor of the upper limb. The largest-diameter corticoreticular fibers are collaterals from corticospinal cells, although these collaterals are smaller than their main-stem axons. Direct corticoreticular projections that are not from collaterals of corticospinal cells tend to be the smallest axons of the corticoreticular system. From all sources, the corticoreticular projections terminate mainly on the dendrites of the reticulospinal cells.

Cerebellar Inputs

The cerebellar output to the reticulospinal system mainly originates from the fastigial nucleus and is crossed. These synaptic contacts tend to be closer to the soma and exert strong influences over reticulospinal neurons than corticoreticular projections. As an

output nucleus for the spinocerebellum, where motor control of posture and movement is concentrated, it is not surprising that the fastigial nucleus are a strong source of cerebellar influence over the reticulospinal system. The cerebellar locomotor region appears to work largely through projection to the reticulospinal system, especially to the pontine part, which in turn activates spinal cord central pattern-generator circuitry.

Inputs for Initiation of Locomotion

In addition to the cerebellar locomotor region already described, the MLR and other rostral portions of the brain stem associated with control of locomotion project to the reticulospinal system. These projections are strongest to the more caudal regions of the reticulospinal system, where reticulospinal cells projecting to the lumbosacral spinal cord are more numerous.

Tectal Inputs

The deep layers of the superior colliculus, around the origins of the tectospinal tracts, project to the reticulospinal system, especially the pontine regions. Head and neck movements are easy to elicit from electrical stimulation of the NRPo and NRPc, so it seems likely that these tectal projections to the reticulospinal system are part of the gaze-control system. To the extent that head movements require accompanying postural shifts to maintain stability, integration of control of head movement with control of gross upper limb movements and posture in the brain stem makes sense. Indeed, as any parent of a new teenage driver recognizes, the tendency to move the arms (and hence the steering wheel) when turning the head is a natural reaction that new drivers must learn to suppress.

Somatosensory Inputs

Reticulospinal neurons can be excited by mechanical stimulation of the skin and deep tissues (muscles and tendons) and by electrical stimulation of peripheral nerves. Light stroking of the skin or hair does not seem to affect reticulospinal cells. Proprioceptors such as muscle spindles and Golgi tendon organs (GTOs) sending their signals up the dorsal columns project strongly to reticulospinal neurons. High-threshold cutaneous afferents, particularly nociceptors, appear more effective for exciting reticulospinal neurons than discriminative touch modalities from the skin. Interestingly, just as reticulospinal outputs are bilateral, stimulation of muscle and cutaneous afferents from either side of the body can activate a reticulospinal neuron. In most cases, stimulation of a peripheral receptor results in the excitation of the reticulospinal neuron, but a sizable minority can produce inhibition.

The cutaneous and muscle receptors appear most likely to influence reticulospinal neurons rostrally. We can speculate that these types of inputs are more likely to influence the outputs of the medial reticulospinal tracts, which in turn appear to have more rapidly conducting fibers and more direct connections to the motor outputs. The higher-threshold afferents, including pain fibers that ascend in the spinothalamic tracts, appear more likely to influence reticulospinal neurons more caudally in the brain stem, from the NRGc. Again, this suggests that these more slowly conducting fibers tend to influence the more slowly conducting outputs in the lateral reticulospinal tracts.

Auditory Inputs

The acoustic startle response appears to be mediated through reticulospinal outputs. In response to a loud auditory tone, the head turns toward the sound and a postural adjustment accompanies the response. The NRPc appears to be the most critical part of the reticulospinal system for conveying the acoustic startle.

Functions of the Reticulospinal System

At the level of behavioral neurophysiology, it becomes difficult to distinguish the medial and lateral reticulospinal tracts. Despite the differences in their origins and projections, there is no clear distinction in the literature between the task-related neural activity for these two different kinds of reticulospinal tract neurons.

Initiation and Regulation of Locomotion

Reticulospinal projections to the spinal cord appear to be the principal system for activating the central pattern generators of the spinal cord for initiation and regulation of locomotion. The evidence for this role comes from multiple experimental approaches and multiple species, has been replicated in different labs, and is overwhelming. As already indicated, the reticulospinal system is a primary target of outputs from the MLR and the cerebellar locomotor region. Reticulospinal neurons are active in phase with locomotion and respond to somatosensory inputs that alter the timing of the locomotor cycle. Stimulation of the PMRF during locomotion produces phase-dependent alterations of the step cycle in intact and reduced preparations. Injection of certain neurochemicals into the PMRF can also initiate locomotion. Partial lesions of the spinal cord white matter in the cat are most disruptive to locomotion when they section the ventral funniculus, where the reticulospinal tracts descend. In the rat,

subjects who recover better locomotion after incomplete spinal cord injury have stronger retrograde labeling of reticulospinal cells from sites distal to the spinal cord injury.

During locomotion in the awake, behaving cat, reticulospinal tract neurons show a variety of response patterns. In many cases, these cells have modulation of activity with an increase during swing (or stance) for each limb. In other words, the cells appear to have firing patterns related to both limbs in an integrated way, not just one limb or the other. On a gross time scale, some of these neurons may appear to be muscle related in that their general period of activation mimics the general period for selected limb muscles. However, when techniques such as spike-triggered averaging are applied, the reticulospinal cells with a pattern apparently mimicking that of a certain muscle may not demonstrate a postspike effect indicative of a direct projection to that muscle. This probably indicates that (1) reticulospinal projections are largely mediated by interneurons that, in turn, project to multiple motor pools and (2) surface similarities in long-term activation patterns for neurons in widely separated parts of the brain cannot be taken as evidence of connections, even when there is anatomical and physiological evidence for tracts connecting those two parts of the brain. Nonetheless, reticulospinal neurons are powerfully modulated during locomotion, and their patterns of modulation can be correlated with critical events in locomotion such as swing onset, stance onset, and the bursting of selected muscles important in walking. There can be little doubt the reticulospinal neurons are a key source of brain stem drive to the spinal cord for supraspinal control and coordination of locomotion.

Preparatory and Anticipatory Aspects of Motor Control

Recent evidence from the cat and primate suggests that, in addition to motor patterns related to the control of posture and limb movements during locomotion, reticulospinal neurons are active in relation to skilled reaching. In the cat, two basic patterns of activity have been described. One group of cells begins firing around the time of an anticipatory postural adjustment preceding the lifting of the limb from a quadrupedal stance to reach for food. The other group of cells is best related to the beginning of the reach itself. In this situation, the cat self-selects the timing between the beginning of the weight shift and the beginning of the reach, providing enough difference in onset times for these two components of the behavior to allow recognition of these two patterns of activity.

In a slightly different experimental setup in the monkey, a similar result has been described. Here, the subject receives an instruction for where to reach, but there is a delay before a go cue indicates the reach should begin. This is the instructed delay paradigm popularized in early studies of the premotor cortex. A substantial portion (35–50%) of neurons sampled in the PMRF have activity that begins with the instruction, well before reaching begins. An example of such a cell is illustrated in Figure 3. In this study, it was possible for the monkey to make subtle adjustments in its posture without failing the trial, so it is possible that the preparatory activity was associated with a postural adjustment or contralateral arm movement that was not measured. Nonetheless, hand position on the reaching side was constant during the instructed delay interval. In addition to cells with this kind of preparatory activity, many of the cells had activity that began with reaching, displaying no response to the instruction stimulus.

This dichotomy between cells in the PMRF of the monkey is highly reminiscent of the two kinds of cells described in the cat. Although the tasks are different, the element of motor control associated with anticipatory postural adjustments seem to be a good candidate for the kinds of functions occurring during instructed delay intervals. In other words, it seems reasonable to speculate that the cells in the cat showing activity in association with anticipatory postural adjustments might be in the same class as the cells in the monkey showing instructed delay-period activity. The corticoreticular projections from area 6 and the rostral part of the motor cortex in area 4, where cortical activity associated with preparatory activity is concentrated, are consistent with the finding of this type of activity in the PMRF.

Movements and Muscle Recruitment Evoked by Electrical Stimulation

Application of brief stimulus trains and averaging electromyography (EMG) responses to a series of single-pulse stimuli produces results that are consistent across the years, across research methods, and across species (cat and monkey). Stimulation of the PMRF usually causes the head to turn to the side ipsilateral to the stimulating electrode in the brain stem. In the upper limbs, reticulospinal stimulation facilitates flexors and inhibits extensors in the ipsilateral upper limb and evokes the opposite response in the contralateral upper limb (Figure 4). About 80% of upper limb responses fit this pattern. In the lower limbs, responses are more variable. In some cases, the responses of the lower limbs are in concert with the responses of the upper limbs, such that both ipsilateral limbs flex and both contralateral limbs extend. This pattern is reminiscent of pacing in a quadrupedal gait. In other cases, the pattern is diagonal between the limb girdles, such

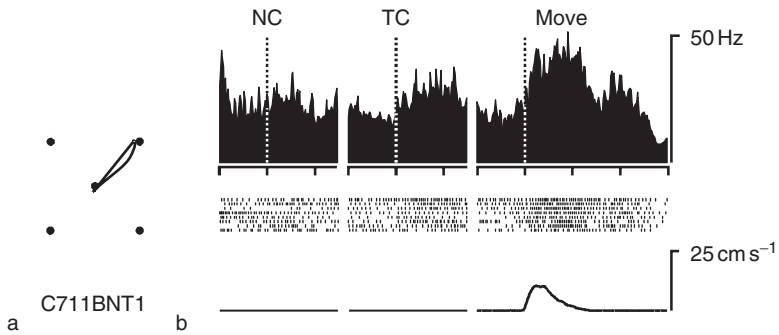

Figure 3 Example of monkey neural activity from a neuron in the PMRF located in the NRGc, just ventral to the abducens: (a) average trajectory from the central while the monkey is waiting spot to one of the targets (the monkey was reaching his hand to one of four targets located on a horizontal work surface); (b) average firing rates for three periods in the task and hand speed. In (b), NC marks the time when a neutral cue was given, alerting the monkey to a particular target that was not the target he would reach to, TC indicates presentation of the true cue, the target for reaching, and Move indicates the time of movement to the target. Hand speed is shown at the bottom of (b). The hand only moves during the movement period. Dot rasters show the times of action potentials for each trial, one trial per row. This neurons activity level increased when the monkey saw which target he was supposed to reach for and also when he actually reached. Such activity is called preparatory activity. PMRF, pontomedullary reticular formation; NC, neural cue; NRGc, nucleus reticularis gigantocellularis; TC, true cue.

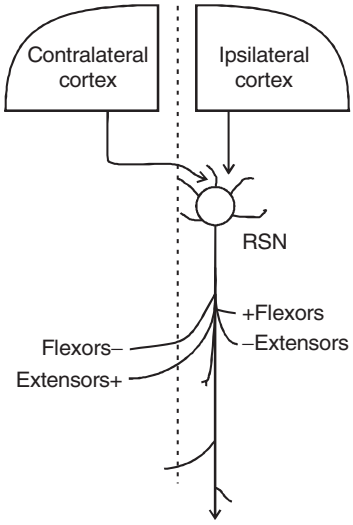

Figure 4 Bilateral basis for functions of reticulospinal neurons. Single reticulospinal cells, especially those from the NRPc and dorsal/rostral parts of the NRGc, deliver collaterals to multiple levels of the spinal cord and often decussate at the segmental level. This, combined with corticoreticular projections from both sides of the motor cortex, makes the reticulospinal system a bilateral system for motor control and a likely alternate pathway for control of movement when one side of the cortex is damaged by an injury such as a cerebrovascular accident. NRGc, nucleus reticularis gigantocellularis; NRPc, nucleus reticularis pontis caudalis; RSN, reticulospinal neuron.

that the ipsilateral upper limb flexes, the ipsilateral lower limb extends, and the opposite combination occurs contralaterally. Not every stimulating site evokes an effect in all four limbs; there is some degree of separation between sites where upper and lower limb movements can be evoked. Most sites, however, produce effects bilaterally within a limb girdle, and sites evoking responses in three or four limbs are not uncommon.

Evidence from electrical stimulation with stimulus trains (12 pulses at approximately 300 Hz) in the cat suggests that the cells most likely to have bilateral influence originate closest to the midline in the reticular formation. However, this result could be confounded by stimulation of decussated fibers that originated on the other side and then descend medially toward the spinal cord. Retrograde labeling studies indicate that cells with origins ipsilaterally and contralaterally are spread through the nuclei of origin on both sides of the brain stem in similar locations, with no medial concentration of the contralateral cells.

The strongest and most prevalent effects of electrical stimulation are observed in the muscles around the spinal column and in the proximal muscles of the limbs. In the cat, there have not been responses documented in muscles distal to the elbow. In the monkey, some effects have been noted in wrist muscles, but these are much less prevalent than responses in the neck, shoulder, and elbow muscles. Likewise, in the lower limb, reticulospinal outputs are mainly directed to proximal muscles. This is consistent with the termination of the reticulospinal axons in the medial part of the ventral gray matter of the spinal cord, where proximal and axial motor pools are located.

When stimulus trains are used in the cat, latencies from the onset of stimulation to the beginning of the EMG response range from around 7 ms in the neck and upper limbs to around 16 ms in the lower limbs. When single-pulse stimuli are used in stimulus-triggered averaging in the monkey, upper limb EMG

onsets can be observed in as fast as 3 ms. This is consistent with the fastest observed reticulospinal conduction velocities, which are reported in the 100–120 m s^{-1} range.

The gross somatotopic organization of the PMRF in the cat is that sites associated with neck movements are found throughout the PMRF; the upper limb is represented in the dorsal aspects of the PMRF about midway along its rostral–caudal extent, in the region just ventral to abducens; and the lower limb is represented in the ventral regions and at the rostral or caudal ends of the PMRF but not as much in the middle. No results indicate that there is a muscle-by-muscle somatotopic map. Broad regions of the PMRF best associated with effects in the broad regions of the body can be found, as already described, but even this organization includes considerable overlap. This overlap combined with broad output effects provides an intrinsic substrate for whole-body movements.

Although intense electrical stimulation suggests there is an inhibitory region in the caudal and ventral PMRF, around the NRMc, and an excitatory region just lateral to the NRPo, more specific electrophysiological methods such as stimulus-triggered averaging do not reveal separation between the course of inhibition and excitation. Inhibition is probably a consequence of projection to inhibitory interneurons in the spinal cord.

These studies of output effects by electrical stimulation have not distinguished whether the effects were conveyed by the medial or lateral reticulospinal tracts. However, based on the bilateral nature of the responses and the rapid latencies between brain stem stimulation and EMG responses in the muscles, it seems likely that most of the results of electrical stimulation that produce direct responses through stimulus-triggered averaging or brief stimulation trains reflect the output of the medial reticulospinal tracts.

Lesion Studies

Lesions of the ventromedial systems in primates, including the reticulospinal and vestibulospinal tracts, leave the animals profoundly debilitated, unable to maintain their posture or move their limbs away from their body. Although hand movement remained possible, limb movement was so limited by the lack of proximal strength that the hand use was of little functional benefit. Lesions of the corticospinal tract in this species, in contrast, were compensated for by the subjects, who could still move about in the cage and feed themselves, albeit clumsily, with gross motor control of the limbs. This result emphasizes the crucial importance of a stable postural base for the effective use of the limbs in skilled movement.

Summary and Implications

The reticulospinal systems originating from the PMRF are one of the major descending systems for motor control. They can receive inputs from cutaneous, tectal, vestibular, auditory, cerebellar, and cortical sources, often from either side of the body. The MLR also projects to the reticulospinal system. The reticulospinal system's outputs are equally widespread and, to a large extent, produce stereotyped responses. Most studies agree that the prevalent outputs from the reticulospinal system are ipsilateral upper limb flexion and contralateral upper limb extension, with turning of the head to the ipsilateral side. In the lower limbs, responses can match those in the upper limbs or be the opposite. The strongest influences are on axial and proximal limb muscles. These widespread outputs are represented at the level of single reticulospinal axons as well as in the organization (or lack thereof) of reticulospinal neurons within the PMRF.

Hence, the reticulospinal tracts are in a position to integrate influences from all over the nervous system and produce responses throughout the body. As we consider lower animals such as the goldfish and lamprey eel, the reticulospinal system is the principal descending system and is mainly aimed at the initiation and regulation of locomotion (swimming). This role in initiation and regulation of locomotion has been retained in the higher animals, and the reciprocal actions between the limbs and, in some cases, between limb girdles is obviously a motor pattern with its roots in the neural control of locomotion. But, in the higher animals, there is now a new integration required between reticulospinal and corticospinal functions for voluntary movement. An interesting aspect of many skilled movements is that they are preceded by preparatory activity in the premotor cortical areas and by anticipatory postural adjustments prior to movement. The corticoreticular projections engaged in skilled movements seem well suited to produce these early aspects of motor control. Through the proper activation of postural and proximal muscles as a first part of movement, a stable basis can be provided on which to superimpose the skilled distal movements that are so uniquely human and that so capture our imagination. It is difficult to imagine how we could perform any useful movement without the integrated control of posture and proximal limb stability provided by the reticulospinal system.

Interestingly, the typical output pattern resulting from the stimulation of the reticulospinal system mimics the upper limb components of the asymmetric tonic neck reflex. This same pattern of upper limb movements dominates the impaired limb after cerebrovascular accidents affecting the motor cortex in the

human. This supports the long-held idea of the reticulospinal system as an alternative pathway for voluntary control of limb movements after stroke. Through projections from the contralesional cortex to both sides of the PMRF (Figure 4) and with the bilateral nature of output from each side of the PMRF, the corticoreticulospinal system is well positioned to provide for the control of upper limb movement even when the normal outputs from the contralateral motor cortex are damaged. Mechanisms behind this potential route for recovery and the possibilities for extending the movement repertoire possible through this system offer exciting areas for future investigation.

See also: Anterior-Posterior Spinal Cord Patterning of the Motor Pool; Balance and Posture Control: Human; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Motor Neuron Specification in Vertebrates; Neural Coding in Primary Motor Cortex; Primate Interneurons; Reaching and Grasping; Reticular Formation and the Brain Stem; Segmentation: Spinal Cord Segmentation and A–P Somite Patterning; Spinal Motor Neurons: Properties.

Further Reading

Brodal A (1969) *Neurological Anatomy*, 2nd edn. New York: Oxford University Press.

Buford JA and Davidson AG (2004) Movement-related and preparatory activity in the reticulospinal system of the monkey. *Experimental Brain Research* 159: 284–300.

Davidson AG and Buford JA (2006) Bilateral actions of the reticulospinal tract on arm and shoulder muscles in the monkey: Stimulus triggered averaging. *Experimental. Brain Research* 173: 25–39.

Davis M, Gendelman DS, Tischler MD, and Gendelman PM (1982) A primary acoustic startle circuit: Lesion and stimulation studies. *Journal of Neuroscience* 2: 791–805.

Drew T, Prentice S, and Schepens B (2004) Cortical and brainstem control of locomotion. In: Mori S, Stuart DG, and Weisendanger M (eds.) *Progress in Brain Research,* vol. 143, pp. 251–262. New York: Elsevier.

Drew T and Rossignol S (1990) Functional organization within the medullary reticular formation of intact unanesthetized cat, I: Movements evoked by microstimulation. *Journal of Neurophysiology* 64: 767–781.

Drew T and Rossignol S (1990) Functional organization within the medullary reticular formation of intact unanesthetized cat, II: Electromyographic activity evoked by microstimulation. *Journal of Neurophysiology* 64: 782–795.

Eccles JC, Nicoll RA, Schwarz DWF, Taborikova H, and Willey TJ (1975) Reticulospinal neurons with and without monosynaptic inputs from cerebellar nuclei. *Journal of Neurophysiology* 38: 513–530.

Homma Y, Nonaka S, Matsuyama K, and Mori S (1995) Fastigiofugal projection to the brainstem nuclei in the cat: An anterograde PHA-L tracing study. *Neuroscience Research* 23: 89–102.

Kuypers HG (1981) Anatomy of descending pathways. In: Brookhart JM and Mountcastle VB (eds.) *Handbook of Physiology, Sect. 1: The Nervous System, Vol. 2: Motor Control,* part 1, pp. 597–666. Bethesda, MD: American Physiological Society.

Matsuyama K, Mori F, Nakajima K, Drew T, Aoki M, and Mori S (2004) Locomotor role of the corticoreticular–reticulospinal–spinal interneuronal system. In: Mori S, Stuart DG, and Weisendanger M (eds.) *Progress in Brain Research,* vol. 143, pp. 231–250. New York: Elsevier.

Peterson BW (1979) Reticulospinal projections to spinal motor nuclei. *Annual Review of Physiology* 41: 127–140.

Wolstencroft JH (1964) Reticulospinal neurons. *Journal of Physiology (London)* 174: 91–108.

Retina: An Overview

J E Dowling, Harvard University, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

The vertebrate retina is a thin (~0.25 mm) layer of neural tissue lining the back of the eye. It is a true part of the brain displaced into the eye during development. In addition to the light-sensitive photoreceptor cells, the retina contains four basic classes of neurons and one principal type of glial cell, the Müller cell. The neurons are organized into three cellular (nuclear) layers, which are separated by two synaptic (plexiform) layers. Virtually all the junctions (synapses) between the retinal neurons are made in the two synaptic layers, and all visual information passes across at least two synapses – one in the outer plexiform layer and another in the inner plexiform layer – before it leaves the eye.

Analysis of visual information occurs in both plexiform layers of the retina. The outer plexiform layer separates visual information into ON and OFF channels and carries out a spatial-type analysis on the visual input. The inner plexiform layer is concerned more with the temporal aspects of light stimuli. Many cells receiving input in this layer respond with transient responses and respond better to moving stimuli than to static spots of light. The output neurons of the retina, the ganglion cells, reflect the processing of information occurring in either the outer or the inner plexiform layers.

Figure 1 is a light micrograph of a piece of human retina. The photoreceptors are located furthest from the front of the eye, at the top of the micrograph. Light, entering the eye, passes through the transparent retina and is captured by the visual pigment-containing, light-sensitive outer segments of the photoreceptors. Overlying the photoreceptors is the pigment epithelium, which serves to absorb stray light and to prevent backscatter of light into the retina, as well as to play an essential role in the regeneration of the visual pigments after light exposure. The cell bodies of the photoreceptors are located in the outer nuclear layer, whereas the cell bodies of three of the basic classes of retinal neurons – horizontal, bipolar, and amacrine cells – are in the inner nuclear layer. The cell bodies of the ganglion cells make up the most proximal cellular layer. The outer and inner plexiform layers are interspersed, respectively, between the outer and inner nuclear layers and the inner nuclear and ganglion cell layers.

In many primates including humans, and in some birds and reptiles, a small region of the retina is specialized for high-acuity vision. It is called the fovea and is centrally located (i.e., on the visual axis of the eye). The layers of the retina below the photoreceptor inner and outer segments are displaced aside from the fovea so that light can impinge directly on the photoreceptors. Only cones are present in the center of the fovea, and these cones are the thinnest and longest photoreceptors in the retina. In humans, the central rod-free foveal area is approximately 0.3 mm in diameter and contains approximately 10 000 cones.

Photoreceptors

Most vertebrate retinas contain two types of photoreceptors, rods and cones. Rods mediate dim-light vision, whereas cones function in bright light and are responsible for color vision. The light sensitivity of the photoreceptors results from the presence of visual pigment molecules contained within their outer segments. One rod pigment, called rhodopsin, and three cone pigments are in the human retina; nonmammalian species may have four or more cone visual pigments. In humans, there is one type of rod cell and three cone types which absorb light preferentially in the red, green, and blue regions of the spectrum.

All visual pigments have a similar chemistry. They consist of two components – retinal (vitamin A aldehyde), termed a chromophore, bound to a protein called opsin. Different visual pigments have different opsins, and this accounts for the variations in their color sensitivity. The light sensitivity of the visual pigments is due to the retinal chromophore. When a visual pigment molecule absorbs a quantum of light, several conformational (shape) changes occur, first in the chromophore and then in the protein (opsin) part of the molecule, which lead, via a cascade of molecular interactions, to excitation of the photoreceptor cell.

All vertebrate photoreceptors hyperpolarize in response to light (i.e., the inside of the cell becomes more negative) as follows: in darkness, the membrane of the outer segment is permeable (leaky) to Na^+. Because Na^+ levels are higher outside the cell than inside, positively charged Na^+ ions enter the cell in darkness, causing the cell to be partially depolarized (i.e., the cell's inside is more positive than is typically the case for neurons at rest). Light decreases the permeability of the outer segment membrane to Na^+, thereby decreasing the flow of positive ions into the cell and causing the inside of the cell to become more negative (i.e., to hyperpolarize).

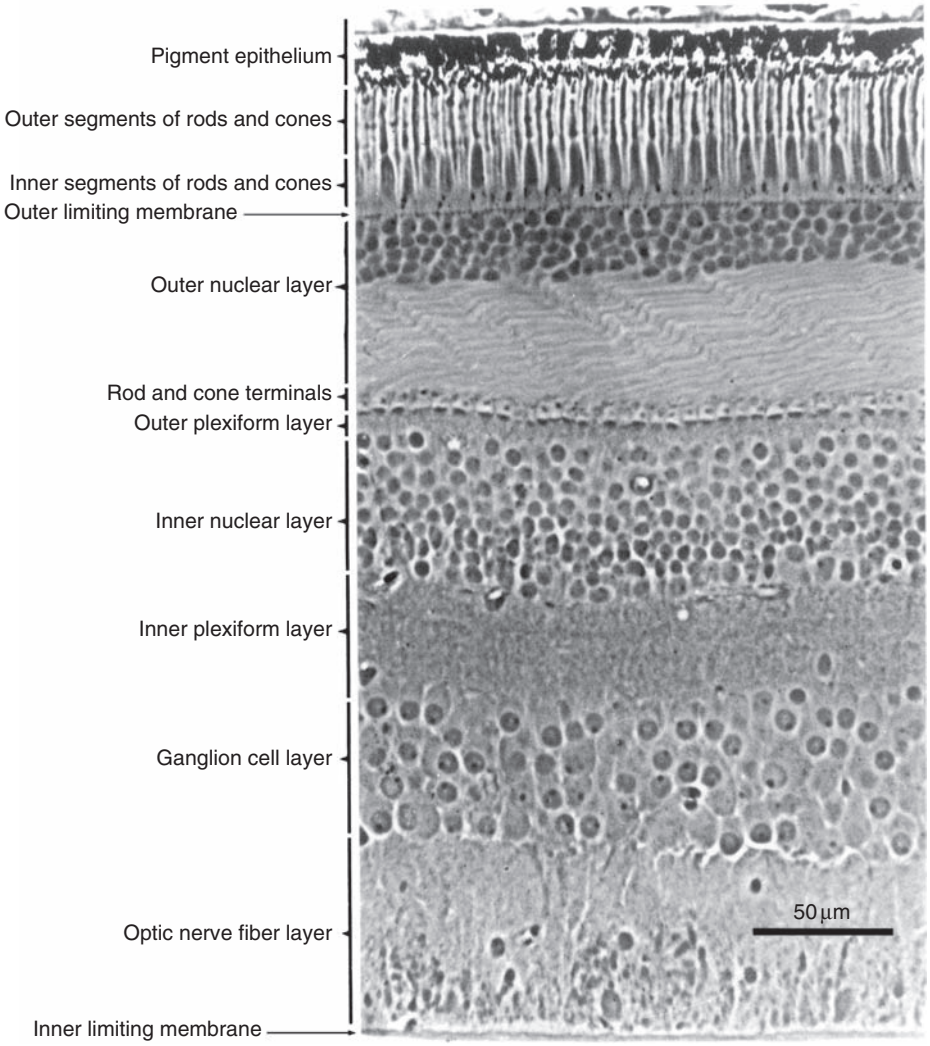

Figure 1 Vertical section through the human retina. Micrograph shows an area approximately 1.25 mm from the center of the fovea. In the foveal region of the retina, the inner layers of the retina are pushed aside so that light can impinge more directly on receptors. Thus, around the fovea and for some distance away (as shown here), receptor terminals are displaced laterally from the rest of the photoreceptor cells. From Boycott B and Dowling JE (1969) Organization of the primate retina: Light microscopy. *Philosophical Transactions of the Royal Society of London, series B: Biological Sciences* 255: 109–184.

Retinal Neurons

Cellular Organization

There are four major classes of retinal neurons in the retina; however, there are many morphological subtypes in each major cell class. The focus here is on the major subtypes of cells in the mammalian and primate retinas (Figure 2).

In the outer plexiform layer, two cell classes – horizontal and bipolar cells – receive input from the photoreceptors. The bipolar cells are the output neurons for the outer plexiform layer; all information passes from outer to inner retina via these neurons. Horizontal cells, on the other hand, extend processes (branches) widely in the outer plexiform layer, but their processes are confined to this layer. Their role principally is to mediate lateral interactions within this first synaptic zone.

Bipolar cell terminals provide the input to the inner plexiform layer, and two cell classes – amacrine and ganglion cells – are ultimately activated there. Amacrine cells, like horizontal cells in the outer plexiform layer, spread processes widely in the inner plexiform layer, and their processes are usually confined to this layer. Ganglion cells are the output neurons for the retina; their axons run along the margin of the retina, collect at the optic disc to form the optic nerve, and carry all visual information to higher visual centers.

Outer plexiform layer: Horizontal and bipolar cells
Most mammalian retinas contain two basic subtypes

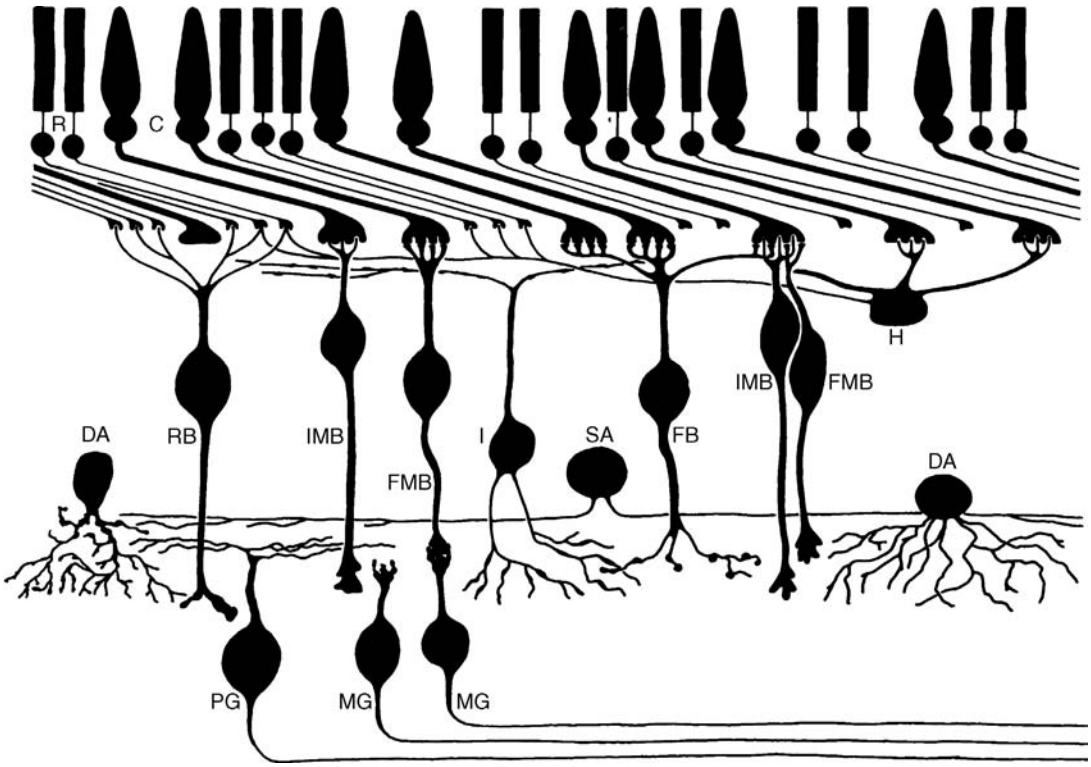

Figure 2 Major cell types found in primate retina as viewed in vertical sections of Golgi-stained retinas. C, cones; DA, diffuse amacrine cell; FB, flat bipolar cell; FMB, flat midget bipolar cells; H, horizontal cell; I, interplexiform cell; IMB; invaginating midget bipolar cell; MG, midget ganglion cell; PG, parasol ganglion cell; R, rods; RB, rod bipolar cells; SA, stratified amacrine cell. From Dowling JE (1991) Retina. In: *Encyclopedia of Human Biology*, vol. 6. New York: Academic Press.

of horizontal cells – a cell with a short axon that ends in a prominent terminal expansion and an axonless cell. The axonless cell has not been seen in the primate retina; rather, the second horizontal cell type in primates has a short axon, although its synaptic connections are like those of the axonless cell. In mammalian retinas, the processes of the axonless cell and the proximal dendritic branches of the short axon cell connect exclusively with cone synaptic terminals, whereas the axon terminals of the short axon cell end exclusively in the rod terminals. Why the input to the short axon cells is organized this way is not clear; an appealing suggestion is that this allows for segregation of rod and cone responses in different regions of the cells. In fish, for example, there are separate rod and cone horizontal cells.

In many vertebrates, up to 11 subtypes of bipolar cells have been distinguished. In mammals, one type is exclusively connected to rods and the rest are connected to cones. In other species (i.e., fish), there are bipolar cells that receive input from both rods and cones as well as pure cone bipolar cells. The rod-related bipolar cells in mammals extend their dendrites into the rod synaptic terminals, and their axon terminals end deep in the inner plexiform layer. These rod bipolars contact between 15 and 50 rod terminals depending on eccentricity; more peripherally located rod bipolar cells contact more rods.

Two of the cone-related bipolar cell types in primates contact only a single cone. These cells, called midget bipolar cells, make different kinds of synaptic contacts with the cone terminals. Their axon terminals end at different levels within the inner plexiform layer, and they appear to be related to the generation of either ON or OFF responses to light in the retina. Every cone terminal in the primate retina probably makes connections with both kinds of midget bipolar cells (Figure 2). In addition to the cone midget bipolar cells in primates, there are also bipolar cells that contact several cones, probably as many as six or seven. They are called flat or invaginating bipolars based on the type of connection they make with the photoreceptors.

Inner plexiform layer: Amacrine and ganglion cells

Most amacrine cells have no axons; all their processes look similar. Some amacrine cells do show two kinds of branches emerging from the cell body. The thinner of the two has a wider extent and has been termed an axon. Many such branches can emerge from a single cell body and it is not known whether they have axon-like properties (i.e., generate and conduct action

potentials). Amacrine cells are diverse in terms of the extent and distribution of their processes, and it is possible to describe a large number of amacrine cell subtypes in most species – up to 20–30. Investigators often classify amacrine cells into two major types: diffuse and stratified amacrine cells. Diffuse amacrine cells extend their processes throughout the thickness of the inner plexiform layer, whereas the stratified cells extend their processes on one or a few levels in the layer (Figure 2). This simple classification scheme can be expanded to include narrow- and wide-field diffuse or stratified amacrine cells, depending on how far their processes extend, and mono-, bi-, or multi-stratified cells, depending on whether their processes are confined to one, two, or several levels in the inner plexiform layers.

A cell termed the interplexiform cell deserves special mention. Some investigators classify these cells as a separate class of retinal neuron, but others believe they are a subtype of amacrine cell. Their cell bodies sit among the amacrine cells, but unlike typical amacrine cells, they send processes to both plexiform layers. Input to these cells is in the inner plexiform layer, whereas most of their output is in the outer plexiform layer. Hence, they appear to be mainly a centrifugal type of neuron, carrying information from the inner to the outer plexiform layer.

Ganglion cells, like amacrine cells, are also diverse in their morphology. In some vertebrate retinas, as many as 10–15 morphological subtypes can be distinguished. They are mainly stratified cells. In primates, particularly in the central region of the retina, ganglion cells are observed with limited dendritic fields. These cells, termed midget ganglion cells, receive input from one midget bipolar cell. The dendrites of these midget ganglion cells ramify either in the upper or the lower part of the inner plexiform layer so that they receive input from one or the other of the two midget bipolar cells described previously. This implies that one cone in the central (foveal) part of primate retina can send two separate messages to the rest of the brain via two midget ganglion cells. One cell signals increases in illumination of the cone (i.e., it is an ON cell), and the other signals decreases in illumination (i.e., it is an OFF cell). The midget ganglion cells are the most common ganglion cell subtype found in the primate retina.

The other major ganglion cell subtype in the primate retina is called the parasol ganglion cell. It receives most of its input from the bipolar cells, receiving input from several cones; is stratified, extending its processes in either the ON or OFF laminae of the inner plexiform layer; and is physiologically either an ON or OFF cell.

Ganglion Cells and Receptive Field Organization

As noted previously, ganglion cells signal brightness (ON cells) or darkness (OFF cells) information to the rest of the brain, but they also communicate more. Indeed, two basic kinds of analyses appear to be occurring in the retina – one carried out mainly in the outer plexiform layer and the other in the inner plexiform layer. Ganglion cells convey information to the rest of the brain that reflects these two stages of visual information processing.

Ganglion cells typically respond to illumination of a restricted but relatively large region of the retina. This region is called the receptive field of the cell and is typically approximately 1 mm in diameter. The most common ganglion cell in the mammalian retinas shows evidence of spatial processing of visual information in its responses. Ganglion cells of this type are often called contrast-sensitive cells, and they are subdivided into two mirror image classes: ON-center, OFF-surround cells and OFF-center, ON-surround cells. Each has a receptive field that is organized into two concentric zones that are antagonistic to each other (Figure 3(a)).

Most ON-center cells respond to an increase in the illumination of the receptive field center with a sustained burst of nerve impulses (action potentials), whereas illumination in the surround inhibits the firing of nerve impulses by the cell for as long as the light is on. OFF-center cells respond in the opposite way; illumination of the receptive field center inhibits the cell, whereas surround illumination provides a sustained excitation of the cell. In both cases, the center and surround zones are antagonistic; if center and surround areas are simultaneously illuminated, the cell responds only with a weak response that reflects the central response. Those ganglion cells that respond with sustained responses to appropriately positioned spots of light have relatively small receptive fields, and their responses mainly reflect the processing that occurs in the outer plexiform layer of the retina. In cat, these ganglion cells are termed X cells; in primates, they are termed P cells.

Other types of ganglion cells respond with more transient responses to retinal illumination, regardless of the position of the illuminating spot in the receptive field. These cells typically have large receptive fields and respond with more vigorous responses to moving stimuli than to static spots of light. The receptive fields of some of these cells are organized into antagonistic center and surround regions like the sustained contrast-sensitivity cells described previously. The transient ganglion cells with a center–surround organization are termed Y cells in cat and M cells in

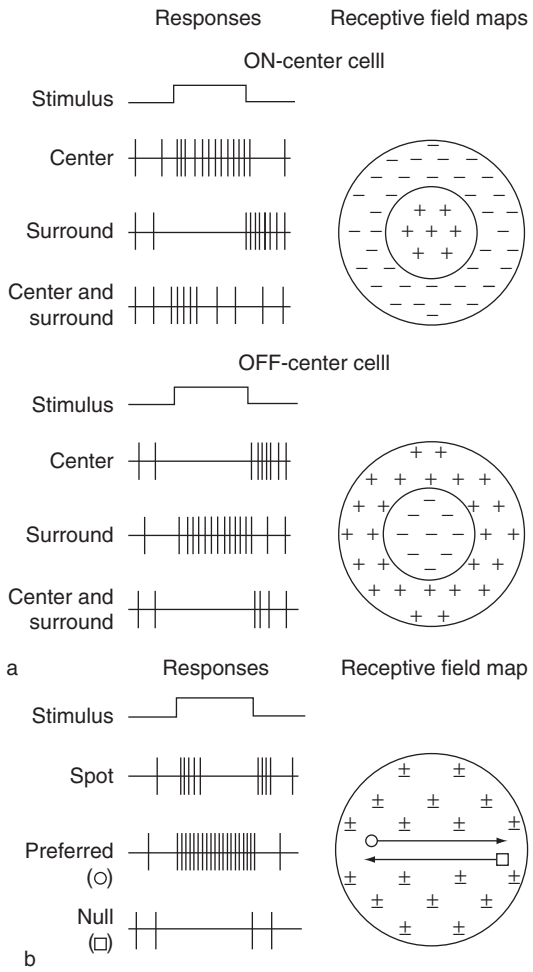

Figure 3 (a) Idealized responses and receptive field maps for ON-center (top) and OFF-center (bottom) contrast-sensitive ganglion cells. Drawings on the left represent hypothetical extracellularly recorded action potential responses to a spot of light presented in the center of the receptive field, in the surround of the receptor field, or in both center and surround regions of the receptive field. +, An increase in firing rate of the cell (i.e., excitation); −, a decrease in firing rate (i.e., inhibition). (b) Idealized responses and a receptive field map for a direction-sensitive ganglion cell. Such cells respond with a burst of action potentials at both onset and termination of a spot of light presented anywhere in the cell's receptive field. This response is indicated by + symbols all over the map. Movement of a spot of light through the receptive field in the preferred direction (open circles) elicits action potentials from the cell that last for as long as the spot is within the field. Movement of a spot of light in the opposite (null) direction (open squares) causes inhibition of the cell's maintained activity for as long as the spot is within the receptive field. From Dowling JE (1991) Retina. In: *Encyclopedia of Human Biology*, vol. 6. New York: Academic Press.

primates. Other ganglion cells, termed ON–OFF cells, give short on–off bursts of impulses to static spots of light positioned anywhere in the receptive field. They respond vigorously to moving spots of light, often have no antagonistic surround, and are particularly numerous in nonmammalian species. Some of these show direction-selective properties (Figure 3(b)). Movement of a light spot in one direction vigorously excites the cell, whereas movement in the opposite direction inhibits the cell.

As noted previously, midget ganglion cells receive input from a single midget bipolar cell, which in turn receives input from a single cone. The center of the receptive field corresponds to this direct pathway and is therefore cone (color) specific. The antagonistic surround responses reflect horizontal and/or amacrine cell activity, and these cells receive input from many cones. The center of the receptive field, therefore, is very small and has a spectral sensitivity different from that of the surround. Such ganglion cells are termed color-opponent cells and are involved in the early stages in color vision processing.

Retinal Synapses

Two types of chemical synaptic contacts are observed in both plexiform layers. One of these is similar in morphology to known chemical synapses seen throughout the brain, and it is termed a conventional synapse. It is characterized by an aggregation of synaptic vesicles in the presynaptic terminal clustered close to the membrane. In the retina, conventional synapses are made by horizontal and amacrine cells. The other type of synapse is characterized by an electron-dense ribbon or bar surrounded by synaptic vesicles and is called a ribbon synapse. The ribbon is thought to act like a conveyer belt, directing and allowing a large number of synaptic vesicles to discharge their contents into the synaptic cleft at the base of the ribbon (Figure 4). Photoreceptor and bipolar cells make ribbon synapses in the retina. Electrical (gap junction) synapses are also observed in both plexiform layers of the retina, and they mediate direct electrical interactions between certain retinal neurons.

Figure 4(a) shows a drawing of a bipolar cell ribbon synapse and an amacrine cell conventional synapse in the inner plexiform layer. Typically, there are two postsynaptic processes at the ribbon synapses of bipolar cells, whereas at conventional synapses only one postsynaptic element is found. In this drawing, the amacrine cell synapse is made back onto the bipolar terminal that is making a synapse on it. Thus, a reciprocal or feedback synaptic arrangement is suggested, which is commonly seen between bipolar cell terminals and amacrine cell processes. Amacrine cell synapses are also made on the processes and cell bodies of ganglion and other amacrine cells.

Photoreceptor cell synapses are particularly complex, and a summary drawing of part of a cone synaptic terminal is shown in Figure 4(b). The synaptic

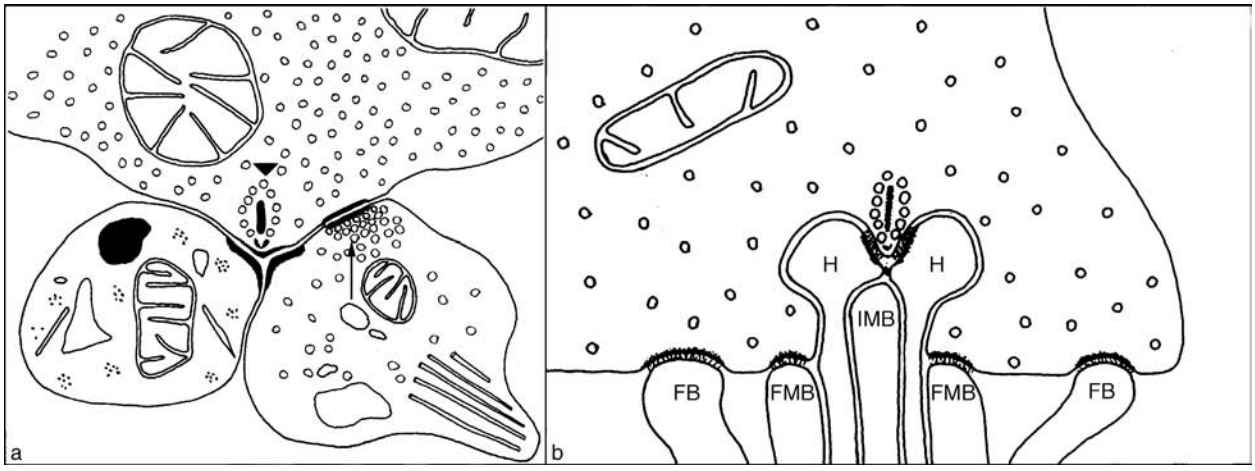

Figure 4 (a) Schematic drawing of a bipolar cell ribbon synapse (arrowhead) and a conventional amacrine cell synapse (arrow) back onto the bipolar cell terminal. One postsynaptic process at the ribbon synapse is an amacrine cell process (right); the other is a ganglion cell dendrite (left), a typical arrangement at cone bipolar cell terminals in primate. (b) Schematic drawing of synapses made by cone terminals in primates. FB, flat bipolar cell dendrite; FMB, flat midget bipolar cell dendrite; H, horizontal cell process; IMB, invaginating bipolar cell dendrite. From Dowling JE (1991) Retina. In: *Encyclopedia of Human Biology*, vol. 6. New York: Academic Press.

ribbons are found above invaginations of the basal surface of the terminal. Dendrites from horizontal cells and invaginating bipolar cells penetrate into the invaginations. Horizontal cell dendrites lie lateral to the synaptic ribbons, whereas the dendrites of the invaginating bipolar cells are centrally positioned. In addition to the ribbon synapse, the cone photoreceptor terminals make a second, unusual synaptic contact with the flat bipolar cells, called a flat or basal junction. No ribbon or vesicle cluster is associated with this junction, but some specializations of the membranes on both sides of the synapse are seen. The flat midget bipolar dendrites are typically found immediately adjacent to the invaginating midget bipolar dendrites, whereas the dendrites of the other flat bipolar cells are positioned away from the invaginations. Rod photoreceptor terminals do not make basal junctions; all rod bipolar dendrites penetrate into invaginations of the rod terminal.

Figure 5 is a simplified summary diagram of the synaptic organization of the primate retina. Each cone in the primate retina makes connections with two midget bipolar cells, one that makes invaginating-type junctions and another that make flat junctions (Figure 5, left). In the central region of the primate retina, these two bipolar cell types synapse on separate midget ganglion cells. In the periphery of the primate retina, midget bipolar cells synapse on ganglion cells that receive input from a few to many midget bipolar terminals.

All cone terminals in primates also make synapses with other bipolar cell types and the dendrites of horizontal cells. The axonal branches of the horizontal cells usually extend to the rod terminals. Synapses made by the horizontal cells have been observed on bipolar cells in many species but very rarely back onto the photoreceptor terminals. However, there is good physiological evidence for feedback from horizontal cells onto cone photoreceptors in many species. Another interesting feature of the cone photoreceptor terminals is that they make small electrical synapses with each other and with adjacent rod terminals.

Rod bipolar terminals in the mammalian retina do not contact ganglion cells directly (Figure 5, right). Rather, amacrine cell processes are always postsynaptic at the ribbon synapses of the rod bipolar terminals. One of these makes a feedback synapse onto the terminal, whereas the other belongs to a special subtype of amacrine cell, called the AII cell, that makes both electrical and conventional chemical synapses with cone bipolar terminals. These cone bipolar terminals then contact the ganglion cells. Thus, rod information in the mammalian retina usually passes through an amacrine cell before it is transmitted to the ganglion cells via the cone bipolar cells. It is believed that the amacrine cells serve to amplify and quicken the rod signal.

Finally, interplexiform cells receive their input from amacrine cells, and they make some synapses in the inner plexiform layer on amacrine and ganglion cell processes. Most of their synapses, however, are made in the outer plexiform layer on bipolar and horizontal cells.

It should be noted that Figure 5 is highly simplified. No one amacrine cell makes the variety of contacts shown for either of the amacrine cells drawn in the figure.

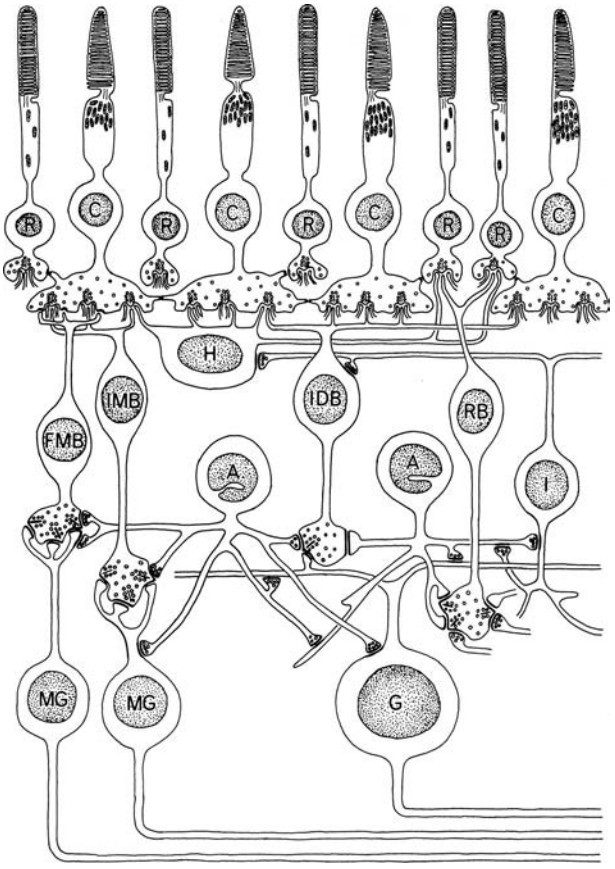

Figure 5 Summary diagram of the synaptic organization of the primate retina. See text for details. A, amacrine cell; C, cones; FMB, flat midget bipolar cells; G, parasol ganglion cell; H, horizontal cell; I, interplexiform cell; IDB, invaginating bipolar cell; IMB, invaginating midget bipolar cell; MG, midget ganglion cells; R, rods; RB, rod bipolar cell. From Dowling JE (1997) Retina. In: *Encyclopedia of Human Biology*, vol. 7. New York: Academic Press.

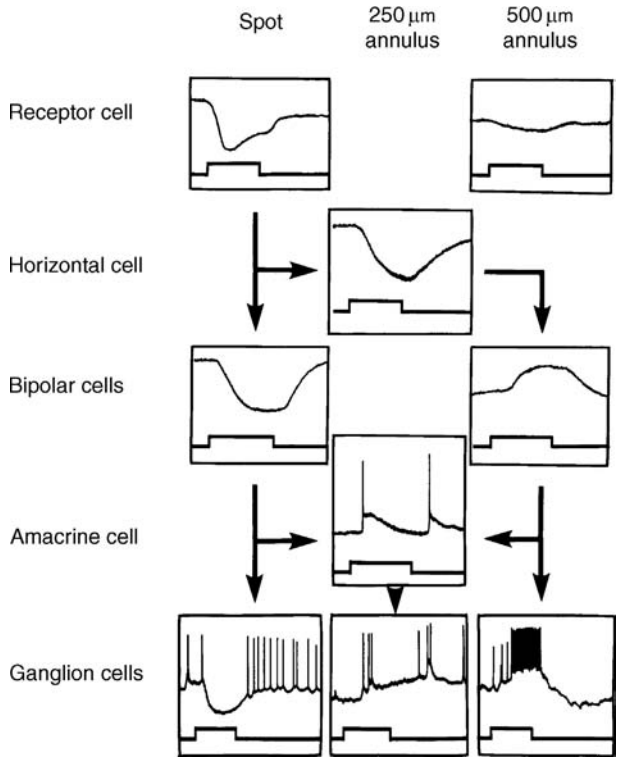

Figure 6 Intracellular responses from receptor (R), horizontal (H), bipolar (B), amacrine (A), and ganglion (G) cells of the mudpuppy retina. Distal retinal neurons (receptor, horizontal, and bipolar cells) respond to illumination with sustained graded potentials; proximal retinal neurons show both sustained and transient potentials and action potentials. Receptor, bipolar, and ganglion cells respond differently to center (spot) and surround (annular) illumination. Horizontal and amacrine cells usually respond similarly to spot and annular illumination; here, responses to a small annulus (250 μm) are shown that would stimulate both the center and the surround of their receptive fields. The bipolar cell illustrated is a center-hyperpolarizing cell, the amacrine cell shown is a transient amacrine cell, and the ganglion cell responses shown on the left and right are from an OFF-center cell. The ganglion cell response in the center is from an ON–OFF cell. The figure also shows how the receptive field organization of the hyperpolarizing bipolar cells, OFF-center ganglion cells, and ON–OFF ganglion cells may be established. The OFF-center bipolar cell responds to direct central illumination (left) by hyperpolarizing and to annular illumination (right) by depolarizing. Note that the switch from hyperpolarizing to depolarizing potentials along the surround illumination pathway occurs at the horizontal–bipolar cell junction. The OFF-center, ON-surround ganglion cells receive much of their input directly from the bipolar cells, whereas the ON–OFF ganglion cell (middle) receives substantial input from the transient amacrine cell. From Dowling JE (1991) Retina. In: *Encyclopedia of Human Biology*, vol. 6. New York: Academic Press.

Neuronal Responses

The distal retinal neurons – receptors, horizontal cells, and bipolar cells – respond to light with sustained and graded membrane potential (voltage) changes (Figure 6). Unlike most neurons found elsewhere in the brain, they do not generally generate action potentials. This is the case because these neurons have relatively short processes, and they do not need to transmit information over long distances; in other words, passive spread of potential along the cell membrane is sufficient to transmit information from one end of the cell to the other. A second reason that the distal retinal cells may function with graded potentials is that such potentials are capable of discriminating a wider range of signals than can all-or-none events (i.e., action potentials). Many amacrine cells and all ganglion cells, on the other hand, generate action potentials.

Another unusual feature of the distal retinal neurons is that most of them respond to light with hyperpolarizing potentials during illumination; that is, the inside of the cell becomes more negative. Elsewhere in the nervous system, hyperpolarizing

potentials are usually associated with inhibition because they prevent neurons from generating action potentials. In the distal retina, the neurons do not generate action potentials; thus, hyperpolarizing potentials can reflect excitation, and the photoreceptors are a good example of this.

Photoreceptors Both rod and cone photoreceptors only hyperpolarize in response to illumination, but rods are approximately 25 times more sensitive to light than are cones. Why photoreceptors hyperpolarize in response to light and why the membrane potential of the cell is partially depolarized in darkness were discussed previously. A typical resting value for the membrane potential of photoreceptors in the dark is −30 mV, and in response to bright illumination the membrane potential decreases to −60 mV, which is a typical resting potential for most neurons at rest. Thus, in the distal retina, photoreceptors and other neurons behave as though darkness is the stimulus and light turns them off.

As noted previously, there are small electrical synapses between photoreceptors. They may relate to the functioning of the photoreceptors or as a pathway for information flow. For example, they can provide an alternative pathway for rod signals to reach the cone bipolar cells. On the other hand, electrical coupling between photoreceptors can reduce membrane noise and hence improve signal detection. The electrical coupling between photoreceptors does slightly increase the receptive field size of the cell. That is, the receptors respond to illumination over a wider area of the retina than that of a single receptor. Nevertheless, photoreceptors typically have the smallest receptive fields of any of the retinal neurons (Figure 6).

Horizontal cells Horizontal cells, like photoreceptors, have relatively low resting membrane potentials in the dark (−30 mV), and most often they only hyperpolarize in response to light (Figure 6). Some horizontal cells respond with depolarizing responses to certain wavelengths of light and with hyperpolarizing responses to other wavelengths. These cells appear to be involved in color processing and are termed chromaticity (C) cells. Horizontal cells that only hyperpolarize to illumination are called luminosity (L) cells.

The receptive fields of horizontal cells are characteristically large, often several millimeters in diameter. Thus, the receptive field size usually exceeds the dendritic spread of the cells. Horizontal cells typically make extensive electrical junctions with each other, and the large receptive fields of these cells can be explained on the basis of the extensive electrical coupling between horizontal cells.

Bipolar cells Bipolar cells, like receptors and horizontal cells, respond to light mainly with sustained graded potentials (Figure 6). In the mammalian retina, all rod bipolar cells depolarize to light. However, two physiological types of cone bipolar cells are found in all species: those that depolarize in response to central spot illumination (ON-center cells) and those that hyperpolarize to such stimuli (OFF-center cells; Figure 6). Furthermore, the cone bipolar cell receptive field is organized into antagonistic zones such that illumination of the surround antagonizes the response to spot illumination. Thus, cone bipolar cells show a center–surround receptive field organization.

Amacrine cells In most retinas, two basic types of amacrine cell responses – transient and sustained – are observed. Transient amacrine cells usually give ON and OFF depolarizing responses to illumination presented anywhere in their receptive field (Figure 6), but there are also transient amacrine cells that respond only at the onset (ON) of illumination. Transient amacrine cells typically generate action potentials, and these potentials are observed on the transient ON and OFF depolarizations. Often, only one to three action potentials are observed on the transient depolarization, and thus these action potentials may serve more as an amplifying mechanism for the amacrine cell potentials rather than as a signal transmitted along the cell, as is the case for most neurons.

Sustained amacrine cell responses often resemble horizontal cell responses. They may be either hyperpolarizing or depolarizing in polarity, and the amplitudes of the two response types are comparable. Usually, sustained amacrine cells, like transient amacrine cells, give similar responses to spot illumination anywhere in their receptive field. There are also amacrine cells whose responses are a mix of transient and sustained potentials.

Ganglion cells The receptive field properties of ganglion cells were discussed previously, and it was noted that many ganglion cells typically show sustained ON-center, OFF-surround responses or sustained OFF-center, ON-surround responses (Figure 6). These ganglion cell responses resemble bipolar cell responses, the major difference being that ganglion cells fire action potentials when they depolarize. Other ganglion cell responses give transient ON or OFF responses or ON–OFF responses. These ganglion cell responses resemble transient amacrine cell activity.

Functional Organization of the Retina

Figure 6 also suggests in a simplified way how receptive fields of the retinal neurons may be produced

by the synaptic interactions occurring within the retina. The figure correlates the basic connections of the retinal cells with intracellular responses recorded from an amphibian (mudpuppy) with spots or annuli (rings) of light.

Photoreceptors have small receptive fields. They respond well to spots of light centered over them but not to light that does not directly strike the cell (annuli). Bipolar and horizontal cells are both activated by the receptors, and Figure 6 shows that both cell types respond with sustained graded potentials that roughly resemble in waveform the photoreceptor response. The figure further shows that the horizontal cells interact with bipolar cells and that this interaction is opposite in terms of polarity from the receptor–bipolar interaction. In Figure 6, the receptor causes the bipolar cell to hyperpolarize, whereas the horizontal cell causes the bipolar cell to depolarize. Horizontal cells have a much larger lateral extent than do bipolar cells; thus, a center–surround receptive field organization is observed in the bipolar cell response. The center response of the bipolar cell (Figure 6, left) is mediated by direct receptor–bipolar cell interactions, whereas the surround response (Figure 6, right) reflects input to the bipolar cell via the horizontal cells.

How horizontal cells mediate the surround response in bipolar cells is not well understood. In most species, feedback from horizontal cells to cone photoreceptor cells has been observed, and it has been proposed that the surround antagonism is generated in this way. Feedforward inhibition of horizontal cells onto bipolar cells has been seen in some species, and no feedback inhibition has been observed on rod photoreceptor cells. Thus, the surround response of the bipolar cells probably results mainly from feedback interactions, but feedforward interactions may contribute as well. The mechanism underlying feedback interaction is unclear. It may represent classic chemical synaptic inhibition or complex electrical interactions between cells in the outer plexiform layer.

As noted previously, two types of synapses made by the cone terminals have been observed: invaginating and flat junctions. Generally, the flat junctions mediate the center-hyperpolarizing responses (OFF responses) in bipolar cells, whereas the invaginating junctions result in the center-depolarizing bipolar cell responses (ON responses).

The bipolar cell terminals carry the visual signal from the outer to inner plexiform layers. Depolarizing or ON bipolar cell terminals are found in the inner half of the plexiform layer, whereas hyperpolarizing (OFF) bipolar cell terminals are in the outer half of the layer. Thus, there is a division of the inner plexiform layer into ON and OFF laminae. ON responses of amacrine and ganglion cells are generated in the lower part of the layer, whereas OFF responses are generated in the upper part. Thus, ON-center ganglion cells have their processes in the lower or ON layer of the inner plexiform layer, OFF-center cells in the upper or OFF layer, and ON–OFF ganglion cells spread processes in both halves of the layer. In addition to the ON and OFF laminae, there is evidence for as many as 10–12 sublaminae in the inner plexiform layer. The cells contributing processes to these sublaminae have distinct properties with regard to characteristics such as rod or cone dominance, response waveform, and excitatory and inhibitory input. Each sublamina thus appears to process visual information somewhat differently.

The responses of the two basic types of ganglion cells found in the mudpuppy (and other retinas) appear to relate to the responses of the major input neurons to the ganglion cells. The sustained ON- or OFF-center ganglion cells (left and right recordings in Figure 6) receive much of their synaptic input directly from the bipolar cells; their responses resemble those of the ON or OFF bipolar cells, and their receptive fields thus reflect primarily the processing that occurs in the outer plexiform layer. The ON–OFF transient ganglion cells, on the other hand, resemble in their properties the transient amacrine cells, and they receive more of their input from amacrine cells (middle recording in Figure 6). These cells reflect more the processing that occurs in the inner plexiform layer. As noted previously, the transient ganglion cells respond better to moving than to static stimuli, and some of these cells show complex receptive field properties such as direction selectivity. Substantial anatomical evi dence indicates that complex ganglion cell receptive field properties, such as motion and direction sensitivity, are mediated in the inner plexiform layer, predominately as a result of interactions between amacrine and ganglion cells. In species that have many motion- and direction-selective cells, there are more amacrine cell synapses per unit area of the inner plexiform layer than there are in species that have relatively few motion- and direction-selective ganglion cells.

In addition to the two basic types of ganglion cells described here, many ganglion cells have a mix of transient amacrine and bipolar cell characteristics. The Y-type ganglion cell of the cat and M-type primate ganglion cell are examples. They show a center–surround receptive organization like the bipolar cells, but their responses to both center and surround stimulation are quite transient and they are sensitive to moving stimuli. These cells appear to receive more of a mix of bipolar and amacrine cell synaptic input than the more sustained ON-center or OFF-center ganglion cells of the cat and primate – the X and P cells.

Pharmacology

The retina, like other regions of the brain, contains a large number of neuroactive substances. At least 17 substances are believed to be released from retinal neurons during retinal activity (Table 1), and, like in other brain regions, these substances may be classified into two general categories – neurotransmitters and neuromodulators – based on their mode of action. Relatively few substances appear to serve as neurotransmitters in the retina. L-Glutamate and acetylcholine mediate fast excitatory pathways in the retina, whereas GABA and glycine mediate fast inhibitory pathways. The bulk of the neuroactive substances released from retinal neurons appear to be neuromodulatory in nature, although little is known about the action or role of most of these substances, especially the peptides.

Both photoreceptors and bipolar cells employ L-glutamate as their transmitter. L-Glutamate depolarizes both horizontal cells and OFF-center (hyperpolarizing) bipolar cells, whereas it hyperpolarizes the ON-center (depolarizing) bipolar cells. Neurotransmitter is released when neurons are depolarized, and because photoreceptors are maintained in a depolarized state in the dark, they release L-glutamate in the dark. When illuminated, photoreceptors hyperpolarize, and transmitter release is decreased. Thus, the light responses of the horizontal and bipolar cells reflect the withdrawal of transmitter from the cell. Horizontal cells are depolarized in the dark because of a dark release of L-glutamate from the photoreceptors; their light response is a hyperpolarization, reflecting the decrease in transmitter release from the photoreceptors in light. The same explanation holds for the OFF-center bipolar cells. These cells are depolarized in the dark; they hyperpolarize in light as transmitter release from the photoreceptor decreases. The ON-center cell, however, is hyperpolarized in the dark, and it depolarizes in the light in response to the decreased release of L-glutamate from the photoreceptor.

Table 1 Substances released from retinal neurons during retinal activity

Amino acids
γ-Aminobutyric acid (GABA)
L-Glutamate
Glycine
Amines
Acetylcholine
Dopamine
Serotonin
Peptides
Cholecystokinin
Enkephalin
Glucagon
Neurotensin
Neuropeptide Y
Somatostatin
Substance P
Vasoactive intestinal peptide
Unconventional modulators
Nitric oxide
Retinoic acid
Zinc

The amino acids GABA and glycine serve as the main inhibitory neurotransmitter agents in both inner and outer plexiform layers. Many horizontal cells contain GABA, and approximately 80% of the amacrine cells in many retinas contain either GABA or glycine, equally divided. Both GABA and glycine powerfully inhibit ganglion cells by opening channels in the cell membrane that cause hyperpolarization of the cell and inhibition of action potential generation. GABA and glycine act on many retinal neurons, mediating specific inhibitory effects. For example, by blocking the effects of GABA with pharmacological agents, direction-sensitive ganglion cells lose their directional selectivity. Such cells now respond to spots of light moving in any direction across the retina. It is believed, therefore, that direction sensitivity is mediated by GABAergic inhibition in the inner plexiform layer.

Most amines and neuropeptides appear to function as neuromodulatory agents in the retina. The exception is acetylcholine, which functions in the retina as an excitatory neurotransmitter in the inner plexiform layer. Acetylcholine is found in a specific type of amacrine cell, the starburst cell, which provides excitatory input to the transient ON–OFF ganglion cells. Interestingly, the starburst amacrine cells also contain GABA, and both acetylcholine and GABA are released by this neuron. The starburst amacrine cells appear to play a critical role in the generation of direction-selective ganglion cell responses.

The other amines, dopamine and serotonin, and the neuropeptides have also been localized principally to amacrine cells. These cells are relatively rare, accounting for no more than a few percent of the total number of amacrine cells. These cells usually spread their processes widely in the inner plexiform layer, and so they are capable of exerting extensive modulatory effects. It is also the case that coexistence of neuroactive substances occurs in many amacrine cells. It has been reported that two peptides can coexist in the same amacrine cell, that a peptide and a monoamine coexist, or that a peptide or monoamine and an inhibitory amino acid (usually GABA) are in the same cell. The significance of the co-localization of two or more neuroactive agents in a single retinal neuron is not yet understood, although an obvious idea is that it permits multiple effects to be exerted on postsynaptic neurons.

The best-understood neuromodulatory substance in the retina is dopamine, which exerts effects throughout the retina via its action on neurons and synapses. For example, it modulates the strength of electrical and chemical synapses in both outer and inner plexiform layers, resulting, for example, in an alteration in strength of the center–surround interactions in bipolar and ganglion cells. Dopamine also modulates the relative strength of rod and cone pathways in many retinas by altering transmitter-gated and voltage-gated channels in neurons. Dopamine exerts these effects by activating receptors on retinal neurons that are linked to the enzyme adenylate cyclase, which catalyzes the formation of the second messenger cyclic AMP in the cells. Cyclic AMP activates a specific kinase which, via phosphorylation, alters the properties of synaptic channels and receptors. Thus, dopamine does not initiate activity in the retina but, rather, modulates retinal function by altering the properties of the retinal neurons and their synapses.

Conclusion

The retina is a relatively well-understood part of the central nervous system, yet its complexity is overwhelming. Whereas we have long recognized that it consists of just photoreceptors and four major classes of neurons, we now realize that each class of neuron consists of many cell subtypes, perhaps as many as 20–30 in the case of amacrine cells. Furthermore, although we can classify the responses of most retinal neurons as ON-center, OFF-surround cells, OFF-center, ON-surround cells, or ON–OFF cells, and as giving principally sustained or transient responses, we now recognize that a variety of subtle image representations are being sent to the brain via the optic nerve and that this variety reflects complex synaptic excitatory and inhibitory synaptic interactions occurring among various neuronal subtypes at different levels of the retina, particularly in different laminae in the inner plexiform layer. Finally, whereas relatively few substances appear to serve as the principal excitatory and inhibitory neurotransmitters in the retina, a much larger number of neuromodulatory substances are known to exist in the retina, including several unconventional substances that have been shown to have significant effects on the retinal cells. Thus, the responses of individual retinal neurons reflect not only the direct excitatory and inhibitory synaptic interactions between neurons but also an extensive modulation of these synaptic interactions by various substances acting in different ways on the neurons and their channels.

See also: Color Vision; Fovea: Primate; Photoreceptor Adaptation; Photoreceptors: Physiology; Phototransduction; Retinal Amacrine Cells; Retinal Bipolar Cells; Retinal Ganglion Cells: Anatomy; Retinal Ganglion Cells: Receptive Fields; Retinal Horizontal Cells; Retinal Pharmacology: Inner Retinal Layers; Retinal Color Mechanisms; Retinal Development: An Overview; Retinal Glia; Rhodopsin; Ribbon Synapses.

Further Reading

Baylor DA (1987) Photoreceptor signals and vision. *Investigative Ophthalmology & Visual Science* 28: 34–49.

Boycott BB and Dowling JE (1969) Organization of the primate retina: Light microscopy. *Philosophical Transactions of the Royal Society of London, Series B* 255: 109–184.

Djamgoz MBA, Archer SN, and Vallerga S (eds.) (1995) *Neurobiology and Clinical Aspects of the Outer Retina*. London: Chapman & Hall.

Dowling JE (1987) *The Retina: An Approachable Part of the Brain*. Cambridge, MA: Harvard University Press.

Dowling JE, Stryer L, and Wiesel T (eds.) (1999) Colloquium on vision: From photon to perception. *Proceedings of the National Academy of Sciences of the United States of America* 93: 557–639.

Fried SI, Münch TA, and Werblin FS (2005) Directional selectivity is formed at multiple levels by laterally offset inhibition in the rabbit retina. *Neuron* 46: 117–127.

Kolb H (1994) The architecture of functional neural circuits in the vertebrate retina. *Investigative Ophthalmology & Visual Science* 35: 2385–2404.

Kolb H, Ripps H, and Wu S (eds.) (2001) *Concepts and Challenges in Retinal Biology*. Amsterdam: Elsevier.

Polyak SL (1941) *The Retina*. Chicago: Chicago University Press.

Rodieck RW (1998) *The First Steps in Seeing*. Sunderland, MA: Sinauer.

Roska B, Molnar A, and Werblin FS (2006) Parallel processing in retinal ganglion cells: How integration of space–time patterns of excitation and inhibition form the spiking output. *Journal of Neurophysiology* 95: 3810–3822.

Sterling P and Demb JB (2004) Retina. In: Shepherd GM (ed.) *The Synaptic Organization of the Brain*, 5th edn., pp. 217–269. Oxford: Oxford University Press.

Wässle H and Boycott BB (1991) Functional architecture of the mammalian retina. *Physiological Reviews* 71: 447–480.

Werblin FS and Dowling JE (1969) Organization of the retina of the mudpuppy, *Necturus maculosus*: II. Intracellular recording. *Journal of Neurophysiology* 32: 339–355.

Retinal Amacrine Cells

S A Bloomfield, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The amacrine cells are the major interneurons in the proximal retina of vertebrates. They form the most diverse type of neuron in the retina, with 20–30 morphological subpopulations described in mammals, based on differences in somatic and dendritic architecture (Figure 1). The dendritic branches of these subtypes stratify within restricted levels of the inner plexiform layer (IPL), indicating selectivity in the synaptic connections made with bipolar cell axon terminals as well as the dendritic processes of neighboring amacrine and ganglion cells. It is believed that the structural diversity within the amacrine cell population reflects an equally large range of physiological properties and the corresponding functional roles played in formulating the retinal output signals expressed by the postsynaptic ganglion cells.

Amacrine Cell Morphology

The amacrine cells were named by Ramón y Cajal after a Greek derivation of 'neurons lacking long fibers,' based on the typically circumscribed space occupied by their dendritic arbors. The morphological classification of amacrine cells has been based on a number of structural parameters, including (1) size, position, and shape of the perikaryon; (2) size and symmetry of dendritic arbors, including the shape and caliber of branches and density of arborization; and (3) the stratification pattern of dendrites within the IPL. Beginning with the Golgi method, a wide variety of techniques have been employed to elucidate the different morphological subtypes of amacrine cell in the retina. It is now clear that 20–30 amacrine cell subtypes exist. A significant number of them are conserved across mammalian species and even show correspondence to amacrine cell subtypes in lower vertebrate retinas. Within a given retina, the amacrine cell subtypes form regular mosaics, but these show tremendous variability in intercellular spacing, indicating differences in their sampling of visuotopic space. It is interesting that any single subtype of amacrine cell forms only a small percentage of the total population. While the AII subtype, which subserves the pathway that carries rod information, constitutes about 13% of all amacrine cells, no other subtype totals more than 5%. This division of amacrine cell subtypes suggests that each plays a separate, yet equally important, role in visual processing.

The somata of amacrine cells lie mainly within the proximal portion of the inner nuclear layer (INL), but those of many subtypes are displaced to the ganglion cell layer (GCL) or lie within the IPL. In some retinas, displaced amacrine cells even outnumber the ganglion cells in the GCL. On the basis of dendritic arbor size, amacrine cells can be divided generally into small-field cells, thought to play a role in local processing and higher acuity vision, and wide-field or long-range cells, involved in more-global retinal processing. Some of these long-range amacrine cells have very e xtensive dendritic arbors that span hundreds of micrometers laterally across the retina. Further, a number of amacrine cell subtypes, described recently in mammals, have very long, narrow, axonlike processes distinct from a conventional dendritic arbor (Figure 2). These fibers show the ultrastructure of typical axons in the central nervous system (CNS), including the expression of high-molecular-weight neurofilament proteins. In contrast to ganglion cells, these so-called polyaxonal amacrine cells bear multiple axonlike fibers that emerge from the soma or dendritic processes and course for several millimeters through the IPL but do not exit the retina via the optic nerve.

The synapses made by amacrine cells are confined to the IPL. Chemically mediated synapses are made with ganglion cell dendrites, bipolar cell axonal processes, or dendrites of other amacrine cells. The dendrites of amacrine cells can maintain both presynaptic and postsynaptic specializations. These synapses can occur juxtaposed along the same dendritic segment to form complex circuits. For example, amacrine cell processes that are postsynaptic to bipolar cells at ribbon synapses can make nearby reciprocal synapses back to the bipolar cell axon. This feedback synapse, which is inhibitory, is thought to modulate the release of neurotransmitter from the bipolar cell. A second type of complex amacrine cell circuit is called a serial synapse; in it, one amacrine cell makes contact with a second amacrine cell, which, in turn, synapses onto a third cell, all within a short distance.

It is interesting that the ratio of the number of synapses that ganglion cells make with amacrine cells and with bipolar cells varies across species. This ratio can also vary for different ganglion cell subtypes within the same retina. It has been suggested that ganglion cells with simple, concentric receptive fields receive synaptic inputs mainly from bipolar cells, whereas those with complex response properties,

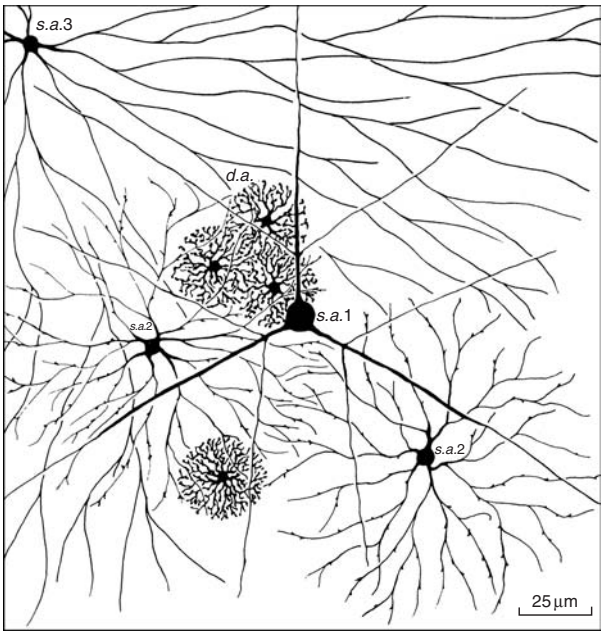

Figure 1 Amacrine cell subtypes in the central primate retina. Drawings are of Golgi-impregnated cells in the rhesus monkey retina as they appear in flat mount. The cells designated *s.a.*1, 2, and 3 are different subtypes of amacrine cells with unistratified dendritic arbors in the IPL. The amacrine cells labeled as *d.a.* display diffusely stratifying dendritic arbors. Note the overlapping fields of the different amacrine cell subtypes as well as differences in their intercellular spacing. From figure 96 in Boycott BB and Dowling JE (1969) Organization of the primate retina: Light microscopy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 255: 109–184, with permission.

such as direction or orientation selectivity, receive more amacrine cell synaptic input. This suggests that amacrine cell circuitry in the IPL underlies the processing of complex visual responses seen in some ganglion cells.

Electrical coupling via gap junctions occurs for almost all the subtypes of amacrine cell in mammalian retinas (**Figure 3**). This includes homologous coupling between adjacent amacrine cell neighbors or heterologous coupling between amacrine cells and ganglion cells or bipolar cells. These electrical synapses are thought to play a variety of roles in information processing. For example, AII amacrine cells receive inputs from rod bipolar cells and thereby process information during nighttime vision. The AII amacrine cells are extensively coupled to one another via gap junctions composed of connexin36 protein. The coupling between AII cells is believed to sum synchronous signals and reduce background noise, thus increasing the fidelity of the rod signals they carry. Further, the conductance of these gap junctions is affected by dopamine, which acts as a light-mediated neuromodulator. Thus, the size of the network formed by AII cells changes under different ambient, background light conditions. The result is that AII cell coupling is constantly changing to ensure the highest fidelity of responses to maintain the very sensitive rod signals in the proximal retina (see the section titled 'AII amacrine cells'). In support of this notion, the sensitivity of postsynaptic ganglion cells is reduced by about one log unit of illumination in the connexin36 knockout mouse retina, in which the AII amacrine cells are uncoupled.

Amacrine Cell Neurotransmitters

With one exception, all amacrine cells are inhibitory interneurons and typically use either γ-aminobutyric acid (GABA) or glycine as their neurotransmitter. The responses of virtually all ganglion cells are affected by exogenous application of these transmitters, either directly or through the complex circuitry of the IPL described above. The GABA receptors appear to be segregated, whereby $GABA_A$ and $GABA_B$ receptor subtypes subserve feedforward inhibition of amacrine cells onto other amacrine cells and ganglion cells, and $GABA_C$ receptors mediate the feedback inhibition onto bipolar cell axon terminals. The combination of GABA effects via the different receptors is believed to influence the sustained or transient characteristics of amacrine and ganglion cell responses. In general, glycine is the transmitter used by small-field amacrine cells, whereas GABA is used by medium- and wide-field cells. Acetylcholine-containing amacrine cells, also called starburst amacrine cells, consist of mirror-image pairs with somata in the GCL and INL and dendrites monostratified in the inner and outer IPL, respectively. It is interesting that starburst amacrine cells are found in all vertebrate retinas, suggesting that they play an important role in visual processing (see the section titled 'Starburst amacrine cells'). The acetylcholine release from starburst cells occurs by two methods: a small, tonic release that is independent of calcium and light and a second, larger acetylcholine release that is light dependent. It is interesting that starburst cells also release GABA, and so this neuron has mixed excitatory and inhibitory effects.

Dopaminergic amacrine cells form a distinctive subtype that is also found in all vertebrate retinas. Dopaminergic amacrine cells in the mammal have polyaxonal morphology, consisting of a relatively small dendritic arbor and numerous long axonal processes that extend more than 1 mm. Most processes stratify in the IPL, where they encircle the AII amacrine cells. As discussed below, dopaminergic amacrine cells control the electrical coupling between AII amacrine cells. The axons of some dopaminergic cells, particularly in nonprimate species, extend outward past the INL to terminate in the outer plexiform layer. The dendrites of

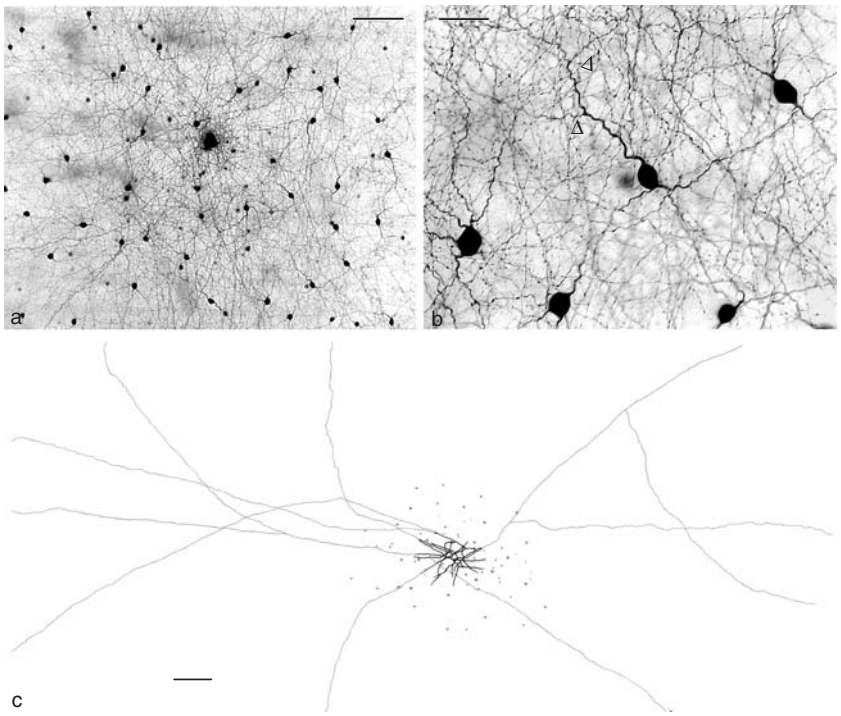

Figure 2 Morphology of a subtype of polyaxonal amacrine cell in the rabbit retina. (a) Photomicrograph showing the extensive tracer coupling pattern following injection of Neurobiotin into a type I polyaxonal amacrine cell (asterisk). (b) Somata of tracer-coupled type I polyaxonal amacrine cells and the overlapping thick dendritic processes and thin axonal processes. Arrowheads indicate axonal processes emerging from proximal dendrites. (c) Camera lucida drawing providing a flat-mount view of a type 1 polyaxonal amacrine cell. The dendritic arbor is presented in black, and the axonal arbor and somata of tracer-coupled amacrine cells are shown in gray. Scale bars = 100 μm (a), 25 μm (b), 200μm (c). From figures 1 and 2 in Völgyi B, Xin D, Amarillo Y, and Bloomfield SA (2001) Morphology and physiology of the polyaxonal amacrine cells in the rabbit retina. *Journal of Comparative Neurology* 440: 109–125.

dopaminergic amacrine cells receive most if not all the excitatory synaptic input from bipolar cells, whereas the release of dopamine occurs at varicosities within the axonal arbor. Thus, there is a polarity to the structure of dopaminergic amacrine cells in terms of their input and output synaptic circuits.

Another well-studied subtype is the indoleamine-accumulating or serotonergic amacrine cell. In the mammal, the serotonergic amacrine cells maintain very long and thin dendritic processes with numerous varicosities. Computational models suggest that varicosities act to electrically isolate individual dendritic segments, thereby allowing for independent processing of synaptic inputs. These cells receive input mainly from rod bipolar cells, but rather than transmit the signals forward to ganglion cells, the serotonergic amacrine cells make reciprocal synapses with the same rod bipolar axon terminals. These amacrine cells are thus thought to modulate the release of bipolar cell neurotransmitter under scotopic light conditions. Little is known about the effects of serotonin on ganglion cells, but application of serotonin antagonists appears to reduce the responses of on-center ganglion cells and to increase the responses of off-center cells.

In addition to conventional neurotransmitters, amacrine cells express a large number of neuropeptides, including substance P, enkephalin, somatostatin, neurotensin, glucagon, vasoactive intestinal peptide, neuropeptide Y, and cholecystokinin. The peptidergic amacrine cells have been best studied in the bird retina, in which they show differences in dendritic structure, particularly in the IPL strata they occupy. In the mammal, peptidergic amacrine cells are mainly subtypes of polyaxonal cells, suggesting that peptides are released over large areas of the retina and serve a neuromodulatory role. Many of these peptidergic amacrine cells accumulate or express more than one peptide and/or express the conventional inhibitory transmitters GABA and glycine. Individual amacrine cells may thereby have multiple roles in both local and global visual processing.

There have been relatively few studies of the effects of peptides on the responses of retinal neurons, but they appear to have general excitatory or inhibitory actions. For example, substance P excites retinal ganglion cells in the fish retina, whereas enkephalin appears to inhibit their activity in the amphibian. In mammals, ablation of neuropeptide Y amacrine cells results in a reduction of the receptive field size of certain ganglion cells that

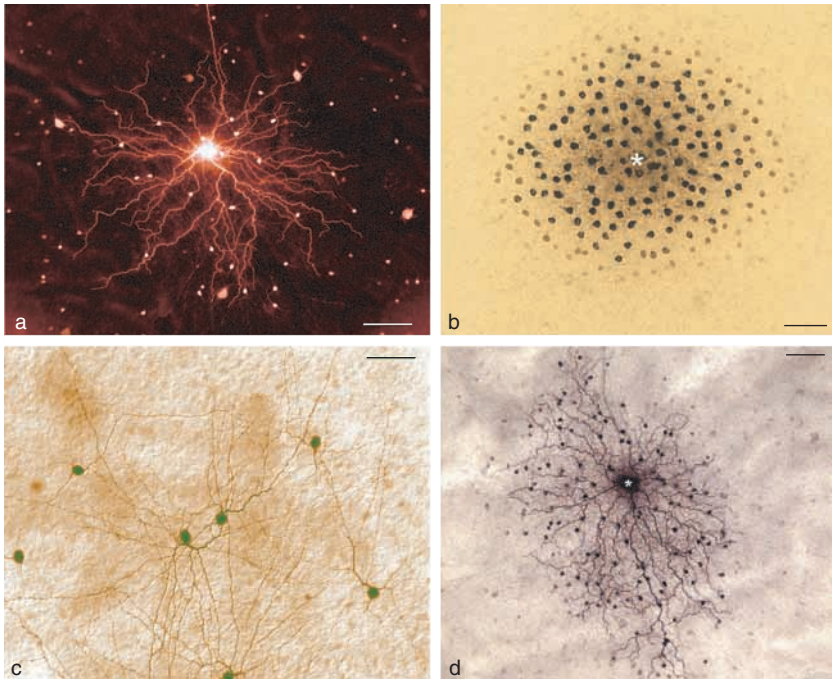

Figure 3 Tracer coupling patterns of amacrine cells in the rabbit retina. (a) Injection of Neurobiotin tracer into an alpha ganglion cell revealed homologous coupling to a ring of nearest-neighbor alpha cells (arrowheads) as well as extensive heterologous coupling to an extensive array of amacrine cells (arrows). (b) Injection of Neurobiotin into an AII amacrine cell (asterisk) reveals extensive homolgous coupling to more than 100 neighboring AII cells. (c) Homologous tracer coupling between a subtype of polyaxonal amacrine cell. (d) Injection of an ON subtype of directional selective ganglion cell (asterisk) reveals that it is extensively coupled to an array of amacrine cells. These amacrine cells show separate dendritic and axonal systems, identifying them as a subtype of polyaxonal amacrine cell. Scale bars = 50 μm (a–d). Panel (c) from figure 4 in Völgyi B, Xin D, Amarillo Y, and Bloomfield SA (2001) Morphology and physiology of the polyaxonal amacrine cells in the rabbit retina. *Journal of Comparative Neurology* 440: 109–125. Panel (d) from Figure 1 in Ackert JM, Wu SH, Lee JC, et al. (2006) Light-induced changes in spike synchronization between coupled ON direction selective ganglion cells in the mammalian retina. *Journal of Neuroscience* 26: 4206–4215.

respond preferentially to low acuity scenes. This peptide is thus thought to have a role in the circuitry underlying ganglion cell spatial tuning.

Amacrine Cell Physiology

Although there are 20–30 subtypes of amacrine cell in the mammalian retina, the physiology of fewer than ten subtypes has been studied in detail. Nevertheless, it is clear the amacrine cells show an extraordinary variety of response properties. In general, amacrine cells can be divided into transient and sustained categories on the basis of the temporal properties of their responses. In lower vertebrates, the transient cells usually show on–off responses, although transient 'on cells' and transient 'off cells' have been described in a number of species. Most sustained amacrine cells show the center-surround antagonistic receptive field organization displayed by bipolar cells and most ganglion cells. In classic experiments, amacrine cells were shown to be robustly excited by a windmill stimulus, indicating their sensitivity to moving objects.

In addition to the relatively simple center-surround receptive field of most amacrine cells, certain subtypes display complex physiology similar to that of some ganglion cells. These include amacrine cell subtypes with orientation selectivity derived from the spatial organization of excitatory and inhibitory inputs or an asymmetry in their dendritic field architecture (**Figure 4**). Amacrine cells with a response preference for the direction of stimulus motion have been reported. This includes the starburst amacrine cell, which shows a response preference for stimuli moving centrifugally away from the soma (see the section titled 'Starburst amacrine cells').

In contrast to other retinal interneurons, many amacrine cells express voltage-gated sodium channels and display action potentials (**Figure 5**). Spiking has been reported in both transient and sustained amacrine cells and is often seen superimposed atop large excitatory potentials. In addition to large-amplitude somatic spikes, some amacrine cells display all-or-none sodium-mediated spikes similar to microspikes thought to be generated within the dendritic

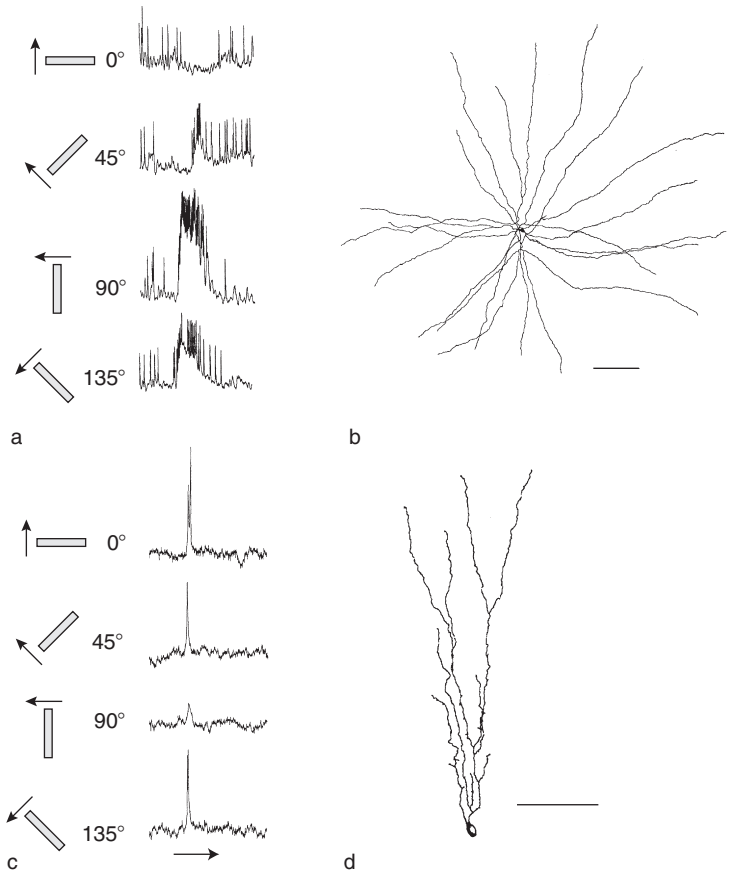

Figure 4 Orientation-sensitive amacrine cells in the rabbit retina. (a) Responses of an orientation-selective amacrine cell to a rectangular slit of light swept across the retina at four orientations. The cell excited when the stimulus was oriented at 90°, but was inhibited when the stimulus was rotated to 0°. (b) Flat-mount view of orientation-selective cell injected with horseradish peroxidase. Same cell whose responses are shown in Panel. (c) Responses of an orientation-biased amacrine cell to a rectangular slit of light swept across the retina at four orientations. A stimulus oriented at 0° evoked the largest depolarization, whereas a stimulus at the orthogonal orientation of 90° produces little response. Unlike orientation-selective amacrine cells, orientation-biased amacrine cells showed no inhibitory responses. (d) Orientation-biased amacrine cell whose responses are shown in panel (c). Note the asymmetry in the dendritic arbor with the major axis oriented at 0° matching the preferred orientation physiology. Scale bar = 150 (b), 50 μm (c). From figures 1, 5, 8, and 10 in Bloomfield SA (1994) Orientation-sensitive amacrine and ganglion cells in the rabbit retina. *Journal of Neurophysiology* 71: 1672–1691, with permission.

membrane (Figure 5(a)). In wide-field amacrine cells, including the polyaxonal cells, spikes are actively propagated centrifugally to terminal branches. In this scheme, somatic spikes are believed to rapidly invade the entire arbor to produce global signaling and release of neurotransmitter. In contrast, dendritic spikes provide only regional activation resulting in local signaling to postsynaptic cells. In this way, the multifocal impulse-generating capability of certain wide-field amacrine cells provides for a complexity in receptive field properties and integration of synaptic inputs.

The dopaminergic amacrine cells show a pacemaker activity of rhythmic, spontaneous bursts of action potentials (Figure 5(b)). This pacemaker activity is generated intrinsically by voltage-gated sodium channels that slowly depolarize the membrane to threshold levels. Since the polyaxonal dopaminergic amacrine cells act on distant cellular targets, the pacemaker activity likely ensures the tonic release of the neuromodulator. In addition to suprathreshold spiking, many amacrine cells show rhythmic subthreshold oscillations. This idea is supported by the finding that bipolar cell axon terminals display calcium-dependent spontaneous membrane oscillations, which may lead to pulsatile transmitter release and rhythmic activity of postsynaptic amacrine cells. In contrast, the oscillatory activity of some amacrine cell subtypes survives cell isolation in culture and thus must be generated intrinsically. The subthreshold oscillatory activity of bipolar and amacrine cells can produce periodic release of neurotransmitter that, in turn, will produce oscillatory activity in postsynaptic cells, particularly ganglion cells. As at other CNS loci,

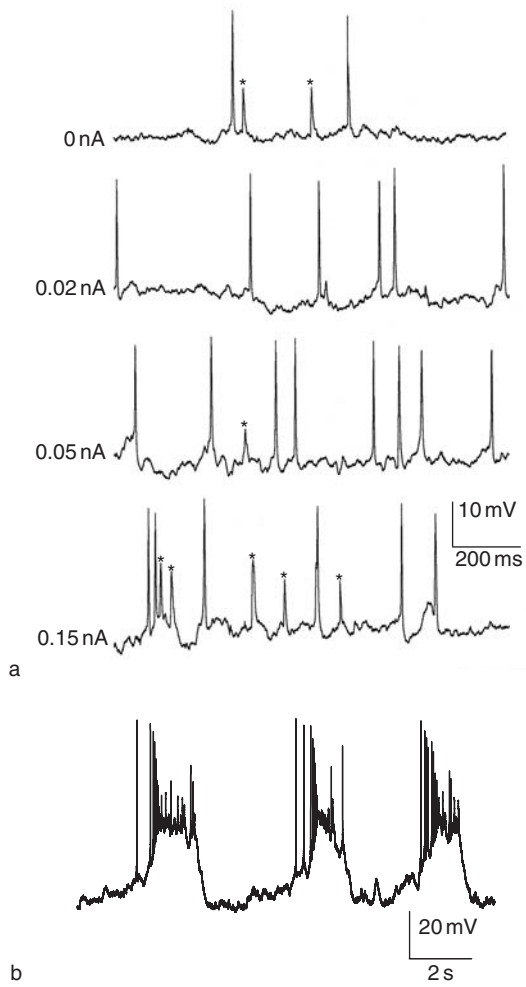

Figure 5 Spontaneous large somatic and small dendritic spikes displayed by large amacrine cells. (a) Application of extrinsic current via the recording electrode differentially evokes the small- and large-amplitude spikes in a wide-field amacrine cell in the rabbit retina. Relatively low-amplitude current effectively evokes the large somatic spikes, whereas the smaller dendritic spikes are evoked only with larger currents, suggesting more-distal initiating sites. Asterisks indicate small spikes. (b) Rhythmic, spontaneous bursts of action potentials displayed by a dopaminergic polyaxonal amacrine cell in the dark-adapted mouse retina. Both large and small amplitude are seen, presumably generated at somatic and dendritic sites, respectively.

subthreshold oscillations distributed among a network of cells can result in light-evoked synchronous activity as cells will tend to reach spike threshold together. The spontaneous oscillatory activity of amacrine cells may thereby serve to organize ensembles by coordinating the activity when activated by appropriate visual stimuli. This synchronous activity may serve to increase stimulus efficacy, encode additional information, and/or bind information about local visual features.

Examples of Specific Amacrine Cell Functional Roles

While it is clear that amacrine cells must play a wide variety of roles in visual processing, only a few of the many subtypes have been studied in detail. The physiology and morphology of two subtypes, the AII and starburst amacrine cells, have been extensively studied, and we now have a good understanding of their specific roles in retinal processing of visual signals. The results of these studies are summarized below.

AII Amacrine Cells

In the mammalian retina, rod and cone photoreceptors synapse onto largely different bipolar cells, thereby segregating their signals into different vertical streams. Whereas up to 11 different morphological types of cone bipolar cells have been reported, showing both on- and off-center physiology, only a single type of rod bipolar cell exists. It is interesting that the axons of rod bipolar cells do not directly contact ganglion cells but instead contact mainly the small-field, bistratified AII amacrine cell. In turn, AII cells form sign-conserving electrical synapses with the axon terminals of on-center cone bipolar cells and sign-inverting glycinergic chemical synapses with the axon terminals of off-center cone bipolar cells. In this way, both on- and off-center scotopic signals utilize the cone pathways before reaching the ganglion cells and ultimately higher brain centers.

The gap junctions formed between AII cells and the on-center cone bipolar cells form nonrectifying electrical synapses across which the direction of signal flow changes with stimulus intensity. As mentioned above, rod signals generated under dim, scotopic light conditions move from the AII cells to the cone bipolar cells to be distributed to the ganglion cells. In contrast, under bright, photopic conditions, cone signals move in the opposite direction, from cone bipolar cells to the AII amacrine cells. The interconnecting gap junctions thus do double duty as conduits for both rod and cone signals.

The AII amacrine cells also form gap junctions between one another, forming an extensive electrical syncytium. Computational models suggest that this coupling increases the signal-to-noise ratio of AII cell responses by summing synchronous responses and deceasing asynchronous noise. It is interesting that the conductance of these homologous AII cell–AII cell junctions is affected by light via modulation of dopamine release by changes in dark-light adaptation (Figures 6(a)–6(c)). Under dim, rod-mediated light conditions, the relationship between coupling and ambient light intensity has two phases: AII cells are relatively uncoupled in the dark-adapted retina but

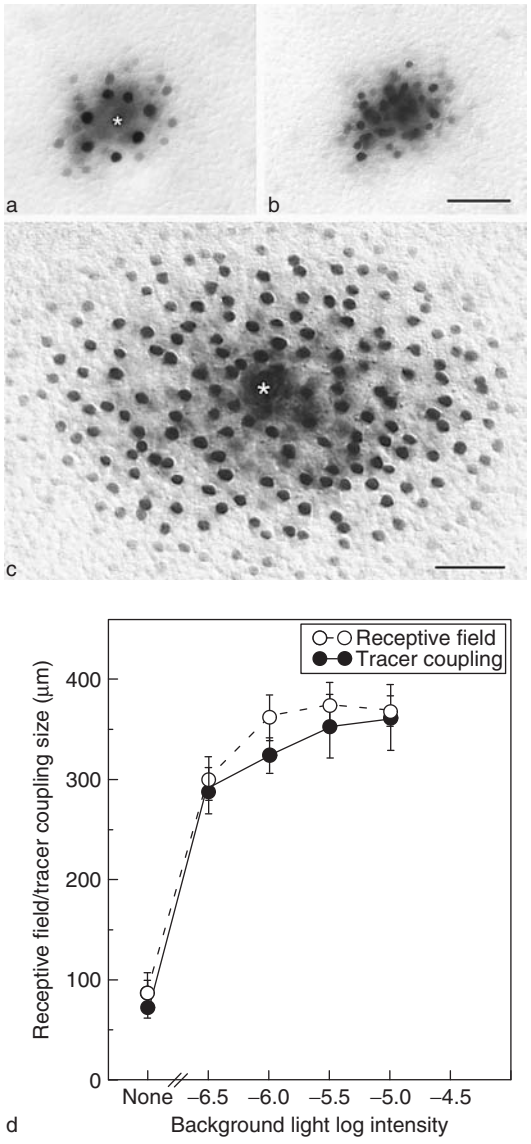

Figure 6 Light-induced changes in the coupling between AII amacrine cells. (a) Tracer coupling pattern of AII cells after injection of one AII cell (asterisk) with Neurobiotin in the dark-adapted rabbit retina. Plane of focus on a group of eight darkly labeled AII cell somata and an outer ring of 10–15 lightly labeled AII cell somata are visible. (b) Plane of focus on tracer-coupled somata of ON cone bipolar cells that lie more distal in the INL. (c) Tracer coupling pattern of AII cells after injection of one cell with Neurobiotin in retina exposed to dim background light. Size of tracer-coupled group has increased dramatically with light exposure. (d) Scatterplot illustrating the similar modulation of tracer coupling and receptive field size of AII cells across a range of background light intensities corresponding to the scotopic and mesopic levels (rod operating range). Each data point illustrates the average and standard error of multiple injections. Well dark-adapted retinas are represented by data point corresponding to 'none' background light intensity. Note that coupling increases dramatically with dim light exposure and then reaches a plateau level over the operating range of rods. Scale bars = 25 μm (a–c). (a, b, d) Adapted from figures 2 and 9 in Bloomfield SA, Xin D, and Osborne T (1997) Light-induced modulation of coupling between AII amacrine cells in the rabbit retina. *Visual Neuroscience* 14: 565–576, with permission. Reproduced (c) from figure 1 in Xin D and Bloomfield SA (1999) Comparison of the responses of AII amacrine cells in the dark- and light-adapted rabbit retina. *Visual Neuroscience* 16: 653–665, with permission.

show a dramatic increase in coupling when dim background stimuli are presented (Figure 6(d)).

What is the function of the light-induced changes in AII cell–AII cell coupling? One idea is that these changes reflect the need for AII cells, as vital elements in the rod pathway, to remain responsive throughout the scotopic-mesopic range. In this scheme, dark adaptation is analogous to starlight conditions, under which rods will only sporadically absorb photons of light. The need, then, is for AII cells to preserve these isolated signals above the background noise. Accordingly, the AII cells are relatively uncoupled in that there are few correlated signals to sum, and so extensive coupling would serve to dissipate and thereby attenuate the few isolated responses rather than enhance them. As dawn approaches, more photons become available. In turn, AII amacrine cells show an increase in coupling, which provides for summation of synchronous activity over a wider area, thus preserving the fidelity of these rod-driven, correlated signals at the expense of spatial acuity. This transition in coupling between dark-adapted retinas and those illuminated with dim background light suggests two basic operating states for AII cells under scotopic/mesopic light conditions: (1) responding to single photon events and (2) summing signals over a relatively large area to sum synchronized events above the background noise. Overall, the AII cell coupling ensures that the most sensitive rod signals are maintained at the ganglion cell level and transmitted to higher brain centers.

Starburst Amacrine Cells

A unique subtype of ganglion cell found in the retinas of many species is the direction-selective (DS) unit. DS cells respond vigorously to stimulus movement in the preferred direction yet show little or no activity following movement in the opposite or null direction. The ON–OFF variety of DS ganglion cell can be divided into four subtypes, each preferring a direction of movement roughly corresponding to the attachment points of the extraocular muscles. The mechanism underlying the computation of direction of motion has been the subject of investigation for more than 40 years. Pharmacologic studies indicate that acetylcholine and GABA can modify the selectivity of DS ganglion cells. In particular, application of GABA antagonists abolishes the direction selectivity of DS ganglion cells, suggesting that asymmetric inhibition resulting from null direction stimulus

movement plays a crucial role. Taken together, these data suggest a role for the starburst amacrine cells, which release both acetylcholine and GABA. As mentioned earlier, starburst amacrine cells are of two mirror-image subtypes. The starburst-*a* cell has a perikaryon in the INL and monostratifies within sublamina-*a* of the distal IPL, whereas the starburst-*b* cell is displaced to the GCL and monostratifies in sublamina-*b*. It is now clear that the dendrites of both the starburst-*a* and -*b* subtypes costratify and are presynaptic to the bistratified processes of ON–OFF DS ganglion cells.

Recent studies have provided compelling evidence that starburst cells indeed play a key role in the generation of direction selectivity. Specific ablation of starburst cells, either through pharmacological manipulation or genetic targeting, abolishes the selectivity of DS cells. Electrophysiological studies have shown that starburst cells provide direct inhibition to DS ganglion cells, suggesting that they provide the GABAergic null inhibition crucial for the direction selectivity of DS cells. Furthermore, the null inhibition from starburst cells is itself direction selective, being stronger for stimulus movement in the null direction.

Taken together, these data suggest that starburst amacrine cells manufacture the direction-selective responses in the retina. In particular, two intrinsic properties of starburst amacrine cells are believed crucial to the generation of DS ganglion cell responses. First, computational and calcium-imaging studies have indicated that the dendritic branches of starburst cells are electrically isolated from one another. This means that single starburst cell dendrites can perform computations independently and simultaneously. Second, light stimulation in the centrifugal direction (away from the soma) produces a greater voltage response than movement in the centripetal direction. It is unclear what establishes these two properties of starburst cells, but potassium channels, calcium channels, chloride ion transporters, and membrane cable properties have all been implicated.

As schematized in Figure 7, centrifugal movement of light along one dendritic branch of a starburst amacrine cell will result in a robust depolarization and the release of GABA onto postsynaptic DS ganglion cells. This provides the null inhibition to the DS cells. Stimulus movement in the opposite, or centripetal, direction will not produce a large response in the starburst cell, and thus no GABAergic inhibition of the postsynaptic DS ganglion cells will be evoked. Thus, the preferred and null directions for a DS

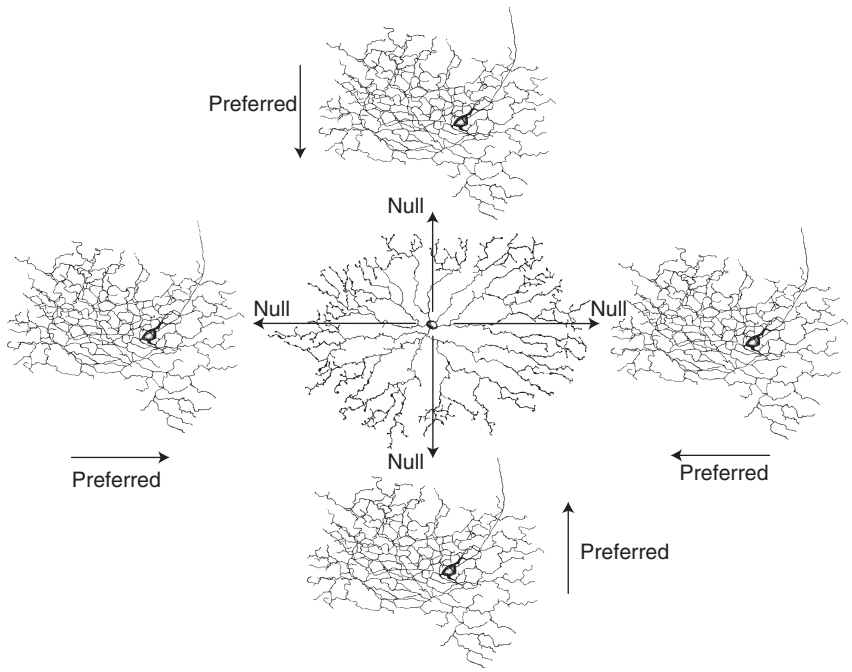

Figure 7 Schematic showing how a starburst amacrine cell generates the direction selectivity of the four ON–OFF direction-selective (DS) ganglion cell subtypes that show different preferred directions: temporal, nasal, superior, and inferior. The dendritic arbor of the starburst amacrine cell (center) can be divided into four parts that are electrically isolated and thus function autonomously. Stimulus movement in the centripetal direction evokes a depolarization of the starburst cell dendritic branch and a release of γ-aminobutyric acid. This produces the null inhibition of the postsynaptic DS ganglion cell. The preferred direction of the DS cell is opposite to the null or centrifugal in the starburst cell dendritic branch. Each starburst cell dendritic branch synapses with a different subtype of ON–OFF DS ganglion cell and thereby produces its preferred direction responses of upward, downward, leftward, and rightward.

ganglion cell are established by the direction of stimulus movement along a starburst cell dendritic branch. Since the dendritic branches of the starburst cell are electrically isolated, each can provide a different directional preference expressed by the four subtypes of ON–OFF DS ganglion cells. In this scheme, each subtype of DS ganglion cell must be postsynaptic to different dendritic branches of a given starburst amacrine cell. Indeed, recordings from DS ganglion cells indicate that while starburst cells lying on the null side of a DS cell provide inhibition, those on the preferred side do not. Likewise, morphological data indicate that DS ganglion cells selectively synapse only with particular starburst amacrine cell processes within their dendritic field.

Conclusions

Amacrine cells form an extensive, heterogeneous group of retinal interneurons that are positioned to modify the output signals carried by the postsynaptic ganglion cells. Despite the large diversity of amacrine cell morphology found in the retina, no one subtype dominates, suggesting each plays a distinct important role in visual processing. This idea is supported by the extraordinary variety of physiological response properties found for amacrine cells, despite the fact that only a few subtypes have been studied so far. These properties include complex receptive fields, including orientation and direction selectivity, somatic and dendritic spiking, pacemaker activity, rhythmic subthreshold oscillations, and electrical isolation of dendritic segments. Further, amacrine cells partake in a variety of complex circuitry that includes both chemical and electrical synapses. In fact, nearly every neurotransmitter, neuromodulator, and neuropeptide found in the CNS has been identified in specific subtypes of retinal amacrine cells. Taken together, these data indicate that the amacrine cells play diverse yet specific roles in the integration of visual signals passing through the retina.

See also: Activity in Visual Development; Fovea: Primate; Retina: An Overview; Retinal Development: An Overview; Vision: Light and Dark Adaptation.

Further Reading

Ackert JM, Wu SH, Lee JC, et al. (2006) Light-induced changes in spike synchronization between coupled ON direction selective ganglion cells in the mammalian retina. *Journal of Neuroscience* 26: 4206–4215.

Bloomfield SA (1994) Orientation-sensitive amacrine and ganglion cells in the rabbit retina. *Journal of Neurophysiology* 71: 1672–1691.

Boycott BB and Dowling JE (1969) Organization of the primate retina: Light microscopy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 255: 109–184.

Bloomfield SA and Dacheux RF (2001) Rod vision: Pathways and processing in the mammalian retina. *Progress in Retinal and Eye Research* 20: 351–384.

Bloomfield SA, Xin D, and Osborne T (1997) Light-induced modulation of coupling between AII amacrine cells in the rabbit retina. *Visual Neuroscience* 14: 565–576.

Dacey DM (1989) Axon-bearing amacrine cells of the macaque monkey retina. *Journal Comparative Neurology* 284: 275–293.

Deans MR, Völgyi B, Goodenough DA, Bloomfield SA, and Paul DL (2002) Connexin36 is essential for transmission of rod-mediated visual signals in the mammalian retina. *Neuron* 36: 703–712.

Dowling JE (1987) *The Retina: An Approachable Part of the Brain*. Cambridge, MA: Harvard University Press.

Euler T, Detwiler PB, and Denk W (2002) Directionally selective calcium signals in dendrites of starburst amacrine cells. *Nature* 418: 845–852.

Feigenspan A, Gustincich S, Bean BP, and Raviola E (1998) Spontaneous activity of solitary dopaminergic cells of the retina. *Journal of Neuroscience* 18: 6776–6789.

Fried SI, Munch TA, and Werblin FS (2002) Mechanisms and circuitry underlying directional selectivity in the retina. *Nature* 420: 411–414.

MacNeil MA, Heussy JK, Dacheux RF, Raviola E, and Masland RH (1999) The shapes and numbers of amacrine cells: Matching of photofilled with Golgi-stained cells in the rabbit retina and comparison with other mammalian species. *Journal of Comparative Neurology* 413: 305–326.

Mills SL and Massey SC (1995) Differential properties of two gap junctional pathways made by AII amacrine cells. *Nature* 377: 734–737.

Tauchi M and Masland RH (1984) The shape and arrangement of the cholinergic neurons in the rabbit retina. *Proceedings of the Royal Society London B: Biological Sciences* 223: 101–119.

Vaney DI (1990) The mosaic of amacrine cells in the mammalian retina. *Progress in Retinal Research* 9: 1–28.

Völgyi B, Xin D, Amarillo Y, and Bloomfield SA (2001) Morphology and physiology of the polyaxonal amacrine cells in the rabbit retina. *Journal of Comparative Neurology* 440: 109–125.

Xin D and Bloomfield SA (1999) Comparison of the responses of AII amacrine cells in the dark- and light-adapted rabbit retina. *Visual Neuroscience* 16: 653–665.

Retinal Bipolar Cells

S M Wu, Baylor College of Medicine, Houston, TX, USA

© 2009 Published by Elsevier Ltd.

Bipolar cells (BCs) are the central neurons of the retina which carry light-elicited signals from photoreceptors and horizontal cells (HCs) in the outer retina to amacrine cells (ACs) and ganglion cells (GCs) in the inner retina. They are also the first postphotoreceptor neurons in the visual pathway exhibiting the center-surround antagonistic receptive field (CSARF) organization, the basic alphabet for encoding spatial information in the visual system. BCs may be on-center with off-surround (named on-center BCs, or depolarizing BCs (DBCs)), or off-center with on-surround (named off-center BCs, or hyperpolarizing BCs (HBCs)). The center input of BCs is mediated by rod and cone photoreceptors which make chemical synapses on BC dendrites. Moreover, BC receptive field centers are often found to be larger than their dendritic fields, suggesting that these cells may be electrically coupled. The surround input to BCs is mediated by interneurons that carry signals laterally from the surround region to the central region of the BCs' receptive field. These interneurons include HCs in the outer retina and ACs in the inner retina. HCs mediate BC surround responses through HC-cone–BC feedback synaptic pathways and/or HC–BC feed-forward (electrical or chemical) synapses, whereas ACs mediate BC surrounds through synapses made on BC axon terminals in the inner plexiform layer (IPL).

In addition to dividing into DBCs and HBCs, BCs are further classified according to their rod/cone inputs. In mammals, rods and cones make synaptic connections with rod and cone BCs separately, and one type of rod BC and 9–10 types of cone BCs have been identified (Figure 1). In fish, small BCs make synaptic contacts with cones and large BCs make contacts with both rods and cones, and in turtle, 11 morphological types of BCs have been found, with some contacting only cones and others contacting only rods. BCs in amphibian retinas contact both rods and cones, but some are clearly rod-dominated whereas others are cone-dominated. Despite these differences between species, BCs in most vertebrate retinas have many common functional and morphological features. It has been established in many vertebrate species, for example, that axon terminals of DBCs ramify in the proximal half (sublamina B) of the IPL, whereas those of the HBCs ramify in the distal half (sublamina A) of the IPL. Axon terminals of rod BCs in mammalian retinas end near the proximal margin of the IPL, and axon terminals of cone BCs ramify in more central regions of the IPL. This agrees with the patterns of BC axon terminal stratification in the salamander: axon terminals of rod-dominated BCs ramify at the two margins of the IPL, whereas those of cone-dominated BCs ramify at central regions of the IPL. Additionally, despite the fact that mammalian rod and cone BCs make segregated synaptic contacts with rods and cones, their light responses are somewhat mixed with rod/cone signals like most BCs in lower vertebrates, partially because rods and cones are electrically coupled and partially because rod BC signals can be transmitted to cone BCs by the AII ACs. Furthermore, although rod HBCs have not been identified anatomically in mammalian retinas, recent physiological evidence suggests that they may exist at least in some mammals. Based on these and other common properties of BCs across vertebrate species, it is reasonable to propose that most vertebrate retinas have six major functional types of BCs: the rod (or rod-dominated), cone (or cone-dominated), and mixed (rod/cone) depolarizing and hyperpolarizing bipolar cells (DBC_R, DBC_C, DBC_M, HBC_R, HBC_C, and HBC_M). Each carries a characteristic set of light response attributes and projects them to the inner retina through axons that terminate at segregated regions (strata) of the IPL. Such stratum-by-stratum projection of light response attributes is exemplified by a large-scale patch clamp study of the salamander BC responses and morphology (Figure 2). This study reveals several rules for the function–morphology relationships of retinal bipolar cells: (1) Cells with axon terminals in strata 1–5 (sublamina A) are HBCs (with outward excitatory cation current (ΔI_C)), and those in strata 6–10 (sublamina B) are DBCs (with inward ΔI_C). This agrees with the sublamina A/B rule observed in many vertebrate species. (2) Cells with axon terminals in strata 1, 2, and 10 are rod-dominated, those in strata 4–8 are cone-dominated, and those in strata 3 and 9 exhibit mixed rod/cone dominance. (3) HBCs exhibit spontaneous excitatory postsynaptic currents (sEPSCs) but DBCs do not. sEPSCs in rod-dominated HBCs are of larger amplitude and higher frequency than those of the cone-dominated HBCs, probably reflecting the fact that the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor kinetics of these two populations of cells are different. (4) Light-evoked ΔI_C at light onset in rod-dominated HBCs and DBCs is sustained, that of the cone-dominated HBCs exhibited a smaller sustained outward current followed by a transient inward current at the light offset, and that of the cone-dominated DBCs exhibited a sustained inward

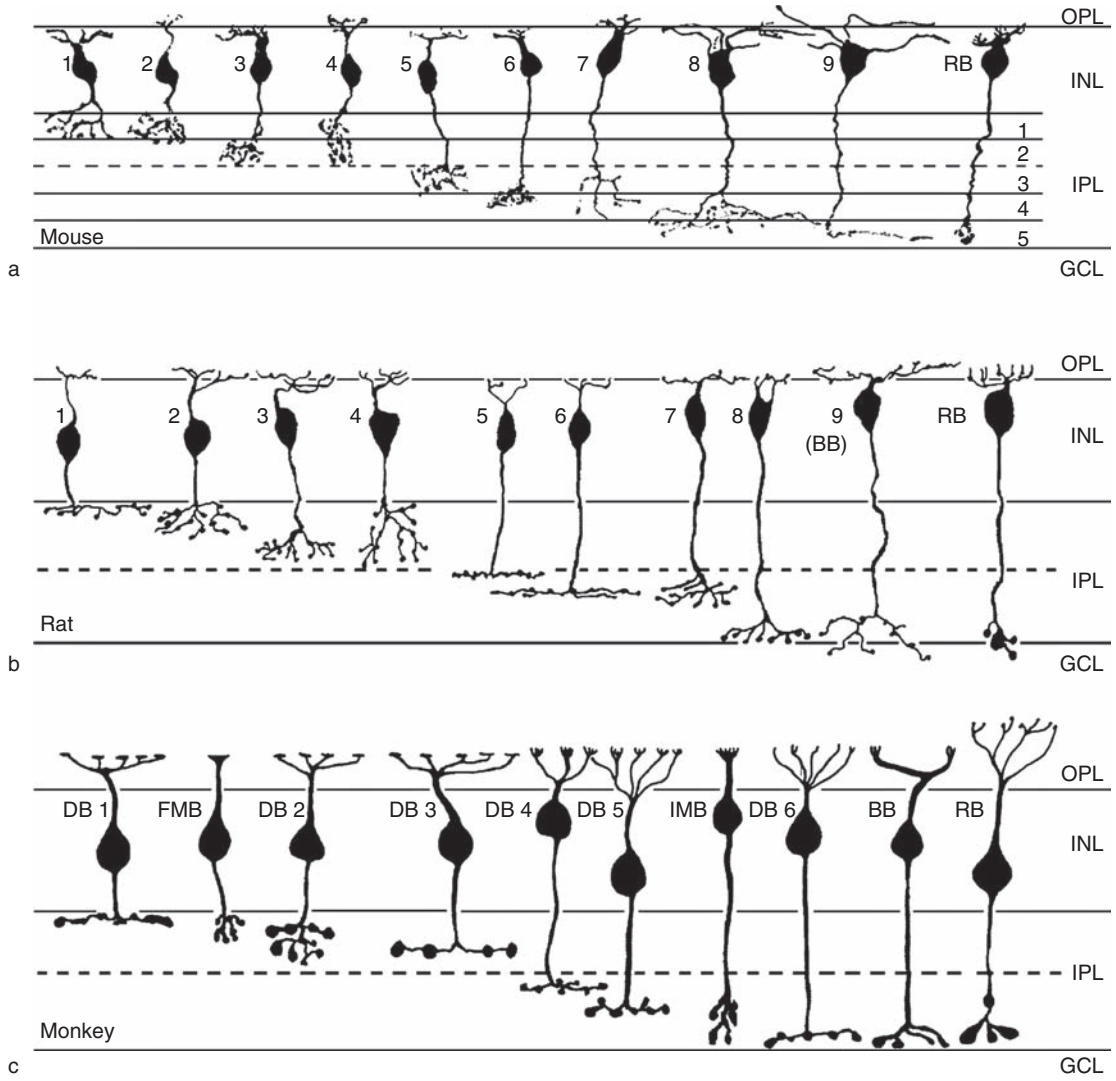

Figure 1 Schematic diagram of various types of bipolar cells in the (a) mouse, (b) rat, and (c) monkey, retinas. Bipolar cell images are derived from confocal images (a) or drawings (b) of Lucifer yellow/neurobiotin-injected cells, or drawings of Golgi-stained retinas (c). OPL, outer plexiform layer; INL, inner nuclear layer; IPL, inner pleiform layer; GCL, ganglion cell layer. Reprinted from and Wiley Liss, Inc. Types of bipolar cells in the mouse retina, Vol. 469, 2004, pp. 70–82. Copyright (2004). Reprinted with permission of Wiley-Liss,Inc., a subsidiary of John Wiley & Sons, Inc.

current followed by a small transient off outward current. (5) Light-evoked inhibitory chloride current (ΔI_{Cl}) in rod-dominated bipolar cells is sustained ON and that in cone-dominated bipolar cells is transient ON–OFF. ΔI_{Cl} in all BCs is outward, and thus is synergistic to ΔI_C in HBCs and antagonistic to ΔI_C in DBCs. (6) Bipolar cells with axon terminals stratified in multiple strata exhibited combined light response properties of the narrowly monostratified cells in the same strata. (7) Bipolar cells with pyramidally branching or globular axons exhibit light response properties very similar to those of narrowly monostratified cells whose axon terminals stratified in the same stratum as the axon terminal endings of the pyramidally branching or globular cells.

In addition to projecting signals to various strata of the IPL, where ACs and GCs gather their inputs, BCs with various light response attributes have different CSARF organizations. A summary diagram of CSARF organization of the DBC_R, DBC_C, HBC_R, and HBC_C is given in Figure 3. The figure illustrates the general synaptic circuitry of the center-surround inputs to BCs and GCs. As discussed later, many new synaptic features, often cell type specific, are not included in this general circuitry scheme. For example, a recent study indicates that in the salamander retina, BCs of the same type are dye coupled, and the receptive field center size is positively correlated with the degree of dye coupling, which varies from one BC type to another. It has also been shown that rabbit cone BCs exhibit CSARFs,

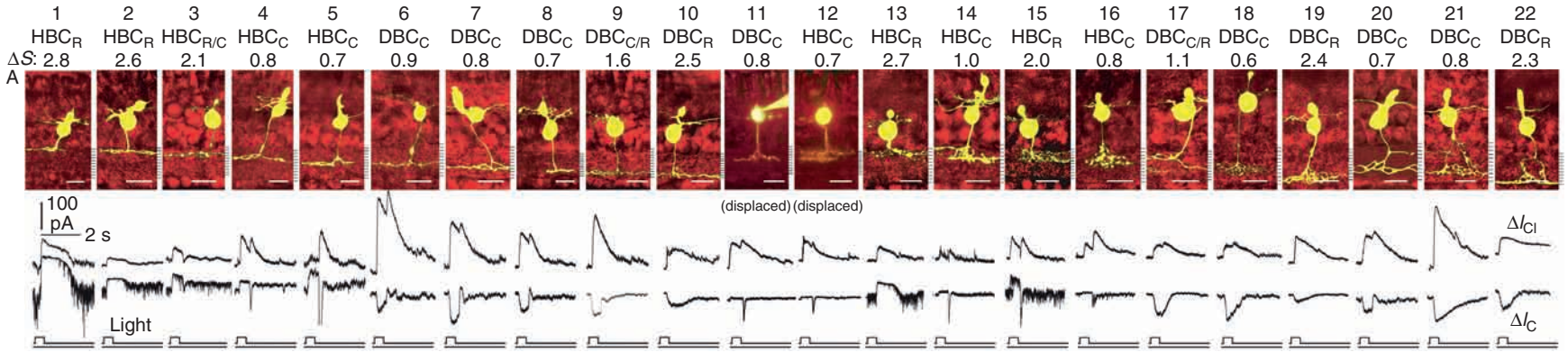

Figure 2 Twenty-two morphologically identified (by Lucifer yellow filling) bipolar cells (BCs) and their light-evoked excitatory cation current (ΔI_C) and inhibitory chloride current (ΔI_{Cl}) recorded from dark-adapted salamander retinal slices. Each cell is named according to their spectral difference (ΔS) and ΔI_C polarity as rod-dominated, cone-dominated, mixed rod/cone hyperpolarizing or depolarizing bipolar cells (HBC_R, HBC_C, $HBC_{R/C}$, DBC_R, DBC_C, or $DBC_{R/C}$). BCs with inward ΔI_C are depolarizing BCs (DBCs) and those with outward ΔI_C are hyperpolarizing (HBCs). The spectral difference, ΔS, is defined as $S_{700} - S_{500}$ (where S_{700} and S_{500} are intensities (in log units) of 700-nm and 500-nm light eliciting responses of the same amplitude). Since ΔS for the salamander rods is about 3.4 and that for the cones is about 0.1, BCs with $\Delta S > 2.0$ are rod-dominated (HBC_R or DBC_R), those with $\Delta S < 1.0$ are cone-dominated (HBC_C or DBC_C), and those with $1.0 < \Delta S < 2.0$ are mixed rod/cone cells ($HBC_{R/C}$ or $DBC_{R/C}$). HBC_C displaced: HBC_Cs with somas displaced in the outer nuclear layer. Modified from Pang JJ, Ga20 F, and Wu SM (2004) Stratum-by-stratum projection of light response attributes by retinal bipolar cells of Ambystoma. *Journal of Physiology* 558: 249–262, and Maple BR, Zhang J, Pang JJ, Gao F, and Wu SM (2005) Characterization of displaced bipolar cells in the tiger salamander retina. *Vision Research* 45: 697–705.

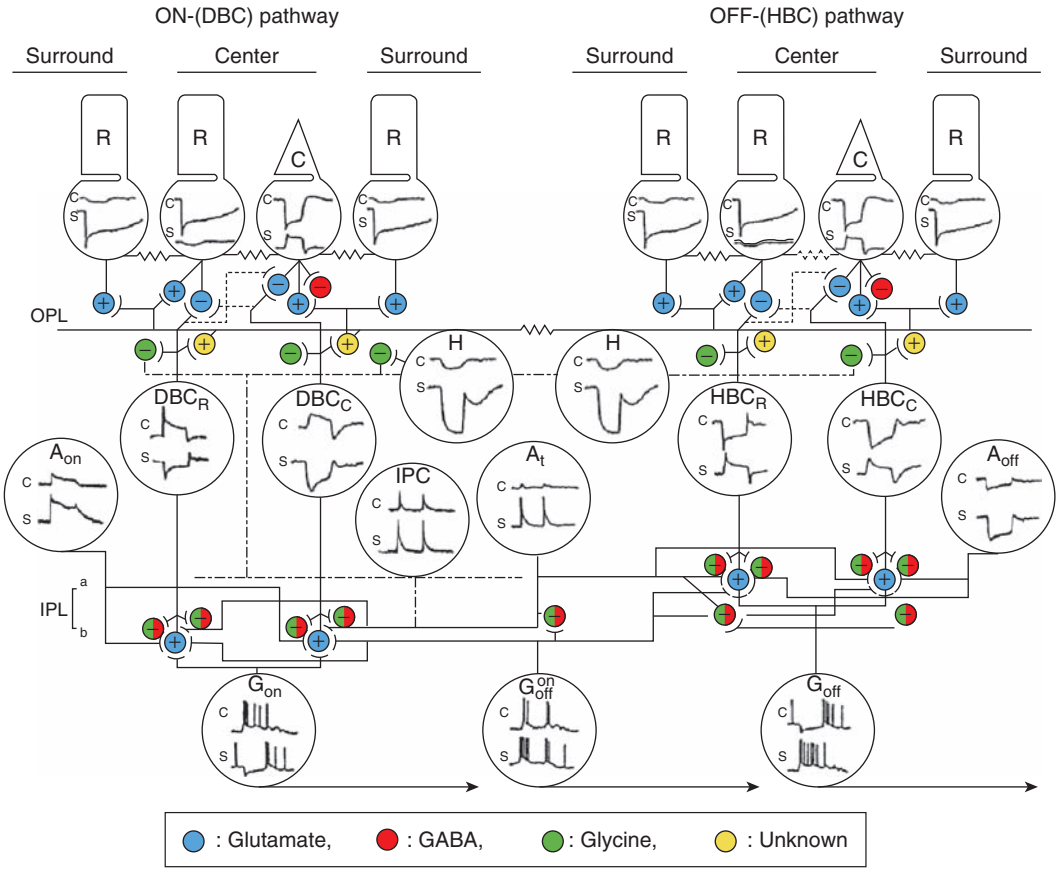

Figure 3 Summary schematic diagram of major synapses that mediate center-surround antagonistic receptive field (CSARF) of bipolar cells and ganglion cells in the retina. The left portion shows the on or depolarizing bipolar cell (DBC) pathway and the right portion shows the off or hyperpolarizing bipolar cell (HBC) pathway. The upper trace in each cell is the voltage response to center illumination, and the lower trace is the response to surround illumination. Plus signs (+) are sign-preserving chemical synapses, minus signs (–) are sign-inverting chemical synapses, and jagged lines (WW) are electric synapses. Neurotransmitter color code: blue, glutamatergic; red, GABAergic; green, glycinergic; yellow, unknown. R, rod; C, cone; H, horizontal cell; HBC_R, HBC_C, DBC_R, and DBC_C, rod- and cone-dominated bipolar cells; A_{ON}, sustained on amacrine cell; A_T, transient on-off amacrine cell; A_{OFF}, sustained off amacrine cell; IPC, interplexiform cell; G_{ON}, sustained ON ganglion cell; $G_{ON\text{-}OFF}$, ON–OFF ganglion cell; G_{OFF}, sustained OFF ganglion cell; OPL, outer plexiform layer; IPL, inner plexiform layer. From Wu SM (2003) Intracellular light responses and synaptic organization of the vertebrate retina. In: Kaufman PL and Alm A (eds.) *Adler's Physiology of the Eye*, ch. 15. St. Louis, MO: Mosby.

whereas rod BCs do not have antagonistic surround responses. Additionally, double color-opponent BCs (red off-center/on-surround and green on-center/off-surround) have been reported in the fish retina. Therefore, different sets of synaptic circuitry must be involved in mediating the center and surround inputs of different functional types of BCs.

Synapses mediating BC center inputs are glutamatergic. Glutamate is released from photoreceptors in darkness and causes a conductance decrease in DBCs and a conductance increase in HBCs. Postsynaptic responses of DBCs are mediated by the metabotropic L-AP4 receptors coupled with a second messenger cascade, and postsynaptic responses of HBCs are mediated by ionotropic receptors. Recent studies have shown that different subtypes of ionotropic glutamate receptors may be used by different types of HBCs in mediating rod versus cone and transient versus sustained signals. Glutamate release from photoreceptors and BCs is mediated by calcium-dependent vesicular processes. Calcium entry into photoreceptor and BC synaptic terminals triggers exocytosis of glutamatergic vesicles and releases packages of glutamate, each of which activates a number of glutamate receptors and initiates discrete transient EPSCs. EPSCs occur spontaneously (sEPSCs) at high rates in darkness, and the rate decreases in the presence of light. sEPSCs are seen in most HBCs, but analogous (but sign-inverted) sEPSCs are not observed in DBCs. This probably reflects a difference in the kinetics of the glutamate receptors in DBCs and HBCs, as metabotropic glutamate receptors are significantly slower than the ionotropic receptors and thus the fast sEPSCs are largely 'filtered' in the DBCs.

BC surround responses are mediated by HCs in the outer retina, through the feedback synaptic pathway (HC→cone→BC) and/or the feed-forward synapses (HC→BC). Three synaptic mechanisms have been proposed for the HC feedback actions on cones. The first is that HCs release an inhibitory neurotransmitter (γ-aminobutyric acid (GABA) in several species) in darkness that opens chloride channels in cones, and surround light hyperpolarizes the HCs, suppresses feedback transmitter release, depolarizes the cones, and results in depolarization in HBCs and hyperpolarization in DBCs. The second mechanism is that surround light hyperpolarizes HCs, resulting in an outward current through hemichannels in their dendrites near the cones, charging the cone membrane and modulating calcium currents in cones, and increases their calcium-dependent glutamate release, which depolarizes the HBCs and hyperpolarizes the DBCs. The third mechanism is that surround-induced HC hyperpolarization elevates the pH in the HC-cone synaptic cleft, leading to an increase of calcium current in cones and a higher rate of glutamate release, which depolarizes the HBCs and hyperpolarizes the DBCs. It is possible that different species under different conditions favor different feedback synaptic mechanisms, and/or different types of HC-cone synapses in the same animal may use one or more of these three HC-cone feedback mechanisms. This idea is supported by a recent study demonstrating that the responses of salamander GCs to dim surround stimuli are sensitive to GABA blockers and those to bright surround stimuli are sensitive to carbenoxolone, a gap junction/hemichannel blocker.

Properties of the feed-forward synapses are less well understood. Results from salamander retina suggest that HC→DBC feed-forward synapses may be functional. However, application of GABA on BC dendrites in Co^{2+} Ringer's solution does not elicit any response, indicating that if feed-forward synapses are chemical, the neurotransmitter is not GABA. Histochemical evidence suggests that only about 50–60% of the HCs in the salamander retina are GABAergic, and the identity of neurotransmitter(s) in the rest of the HCs is unknown (but unlikely to be glycine). Recent studies on the rabbit and salamander retinas have shown that HCs and subpopulations of BC are dye coupled. This raises the possibility that HC-BC electrical coupling may be involved in mediating the HC-BC feed-forward synapses for DBC surround responses (which does not work for HBC surround because HC-HBC coupling gives hyperpolarizing surround signal in HBCs). A systematic investigation on this issue is needed.

In addition to HC inputs in the outer retina, BC surround responses are also mediated by ACs through their feedback synapses made onto BC axon terminals in the inner retina. The majority of ACs use GABA or glycine as their neurotransmitter. GABA receptors on ACs and GCs are largely $GABA_A$, and those on BC axon terminals are largely $GABA_C$. Glycine receptors have been localized in ACs, GCs, and BC dendrites and BC axon terminals. Receptors in the IPL are postsynaptic to glycinergic ACs, and those in the OPL are postsynaptic to glycinergic interplexiform cells. Voltage clamp experiments reveal that ΔI_{Cl} in all BCs is outward. This suggests that GABA/glycine-mediated ΔI_{Cl} may be responsible for DBC hyperpolarizing surround responses, but not the HBC surround responses which are depolarizing (inward current).

In the *Xenopus* retina, GABA suppresses the surround responses of the DBCs but only slightly reduces the surround of the HBCs, and glycine suppresses the surround responses of both DBCs and HBCs. In the tiger salamander, one study shows that GABA reduces the surround responses of a subpopulation of HBCs, but another report reveals that application of picrotoxin and strychnine does not affect the surround responses of either DBCs or HBCs. Recent studies in the primate retina indicate that the HC feedback signal to cones as well as the surround responses of GCs are not sensitive to GABAergic or glycinergic agents, but are sensitive to carbenoxolone, suggesting that the surround responses in GCs are mainly mediated by HC actions on BCs in the outer retina, not by GABAergic or glycinergic AC actions in the inner retina. The reasons for these differences in GABA/glycine actions on surround responses are unclear. It is evident, however, that the surround responses of different functional types of BCs are mediated by different combinations of synaptic circuitries: HC→cone (GABA/hemichannel/pH) →BC, HC→BC (chemical/gap junction), and AC→BC (GABA/glycine). It is conceivable that the surround responses of different BCs/GCs from different animals under different conditions are sensitive to different synaptic blockers. The wide variation of synaptic circuitries underlying surround responses of various functional types of BCs allows the flexibility for function-specific modulation of BC-receptive fields and contrast sensitivity by various lighting conditions. Systematic studies on detailed synaptic pathways/mechanisms underlying surround responses for each type of BCs are needed.

See also: Fovea: Primate; Retina: An Overview; Retinal Ganglion Cells: Anatomy; Retinal Ganglion Cells: Receptive Fields; Retinal Horizontal Cells; Retinal Color Mechanisms; Retinal Development: An Overview; Retinal Development: Cell Type Specification.

Further Reading

Boycott B and Wassle H (1999) Parallel processing in the mammalian retina: The Proctor Lecture. *Investigative Ophthalmology and Visual Science* 40: 1313–1327.

Dowling JE (1987) *The Retina: An Approachable Part of the Brain.* Boston: Harvard University Press.

Ghosh KK, Bujan S, Haverkamp S, Feigenspan A, and Wassle H (2004) Types of bipolar cells in the mouse retina. *Journal of Comparative Neurology* 469: 70–82.

Kolb H (1994) The architecture of functional neural circuits in the vertebrate retina. The Proctor Lecture. *Investigative Ophthalmology and Visual Science* 35: 2385–2404.

Kolb H, Ripps H, and Wu SM (eds.) (2001) *Concepts and Challenges in Retinal Biology: A Tribute to John E. Dowling. Progress in Brain Research,* vol. 131. Amsterdam: Elsevier.

Maple BR, Zhang J, Pang JJ, Gao F, and Wu SM (2005) Characterization of displaced bipolar cells in the tiger salamander retina. *Vision Research* 45: 697–705.

Pang JJ, Gao F, and Wu SM (2004) Stratum-by-stratum projection of light response attributes by retinal bipolar cells of Ambystoma. *Journal of Physiology* 558: 249–262.

Sterling P and Demb JB (2004) Retina. In: Shepherd GM (ed.) *The Synaptic Organization of the Brain,* 5th edn., pp. 217–269. Oxford: Oxford University Press.

Wu SM (1994) Synaptic transmission in the outer retina. *Annual Review of Physiology* 56: 141–168.

Wu SM (2003) Intracellular light responses and synaptic organization of the vertebrate retina. In: Kaufman PL and Alm A (eds.) *Adler's Physiology of the Eye,* ch. 15. St. Louis, MO: Mosby.

Retinal Color Mechanisms

N W Daw, Yale University School of Medicine, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

In order to see color, one needs at least two classes of photoreceptors with different spectral sensitivities, and differential connections to cells higher in the system. Usually, this means two classes of cone photoreceptors, feeding into color-coded horizontal, bipolar, and ganglion cells in daylight at a level where the rod photoreceptors are not active. However, humans see color in circumstances where only the rods and red-absorbing cones are active. Moreover, frogs have two types of rods with different spectral sensitivities, and could presumably use them to see color, although the point has never specifically been tested. But these are exceptions. In this review, we will follow the conventional terminology, and classify color vision into categories according to the number of cone types that a species has.

Color Vision Varies with Species

Humans have three types of cones: red-absorbing, called L for long-wavelength cones; green-absorbing, called M for medium-wavelength cones; and blue-absorbing, called S for short-wavelength cones. Normal humans are thus known as trichromatic, and can match the color of any object with a mixture of three well-chosen red, green, and blue wavelengths. Old World monkeys such as macaque and chimpanzee also tend to be trichromatic, while New World monkeys such as squirrel and owl monkey tend to be either dichromatic, so that they only need two wavelengths to match the color of an object, or to have somewhat deficient trichromatic vision. Cats, rabbits, dogs, and ground squirrels are dichromatic, goldfish are tetrachromatic, and pigeons are at least trichromatic, and possibly tetrachromatic. Many other species have been shown to have color vision, but the nature has not been fully pinned down.

The photoreceptor pigments of most animals are based on vitamin A_1, with a rod pigment that is called rhodopsin, which has a spectral sensitivity peaking around 500 nm. These animals include mammals, birds, reptiles, amphibians, and marine fish. Some freshwater fish such as carp and goldfish have pigments based on vitamin A_2, with a rod pigment called porphyropsin, which has a spectral sensitivity peaking around 522 nm. The spectral sensitivity of the cone pigments is correspondingly shifted toward longer wavelengths, as a result of which goldfish can make discriminations at higher wavelengths than humans.

Some species, particularly turtles and pigeons, have oil droplets through which the light passes between the lens and the pigment in the photoreceptor. The droplets appear to the human to be either red, orange, or yellow, and cutoff short wavelengths from reaching the photoreceptor pigment. Thus, in theory, these species, which tend to have three classes of photoreceptors, could be more than trichromatic, by using a photoreceptor in combination with different classes of oil droplets. However, tests have not been completed to test whether this is the case or not.

Opponent Color and Double Opponent Responses in Fish

Thomas Young, elaborated by Herman von Helmholtz, was the first to suggest that human color vision is trichromatic, with three classes of cone photoreceptors. Ewald Hering in the nineteenth century came out with an alternative formulation of color vision, emphasizing that red and green are mutually exclusive colors and so are yellow and blue. Hering implied that these opponent color processes exist in the photoreceptors. A number of stage theorists in the early twentieth century pointed out that these alternative theories can be reconciled, if the three classes of photoreceptors feed into a second stage, some with excitatory connections and others with inhibitory connections, to produce opponent color cells. This was shown to be true in the 1960s by Russell de Valois in the macaque monkey, and by Ted MacNichol and Gunnar Svaetichin in a variety of fish.

MacNichol and Svaetichin showed further that the first cells in the fish retina after the photoreceptors, the horizontal cells, may be opponent color cells (Figure 1). Some horizontal cells hyperpolarize to light of all wavelengths: this is a luminance response. Some hyperpolarize to short wavelengths and depolarize to long wavelengths with a neutral point near 550 nm: these were named yellow–blue (Y–B) chromaticity or opponent color cells. Others hyperpolarize to short wavelengths and depolarize to long wavelengths with a neutral point near 600 nm: these were named red–green (R–G) chromaticity or opponent color cells. Fish in deep waters tend to have one type of opponent color horizontal cell, while fish in shallow waters tend to have both. In particular, goldfish and carp have both types of opponent color cells.

There is further processing of color in goldfish and carp retinas. Most ganglion cells are double opponent, that is, opponent for both color and space (Figure 2).

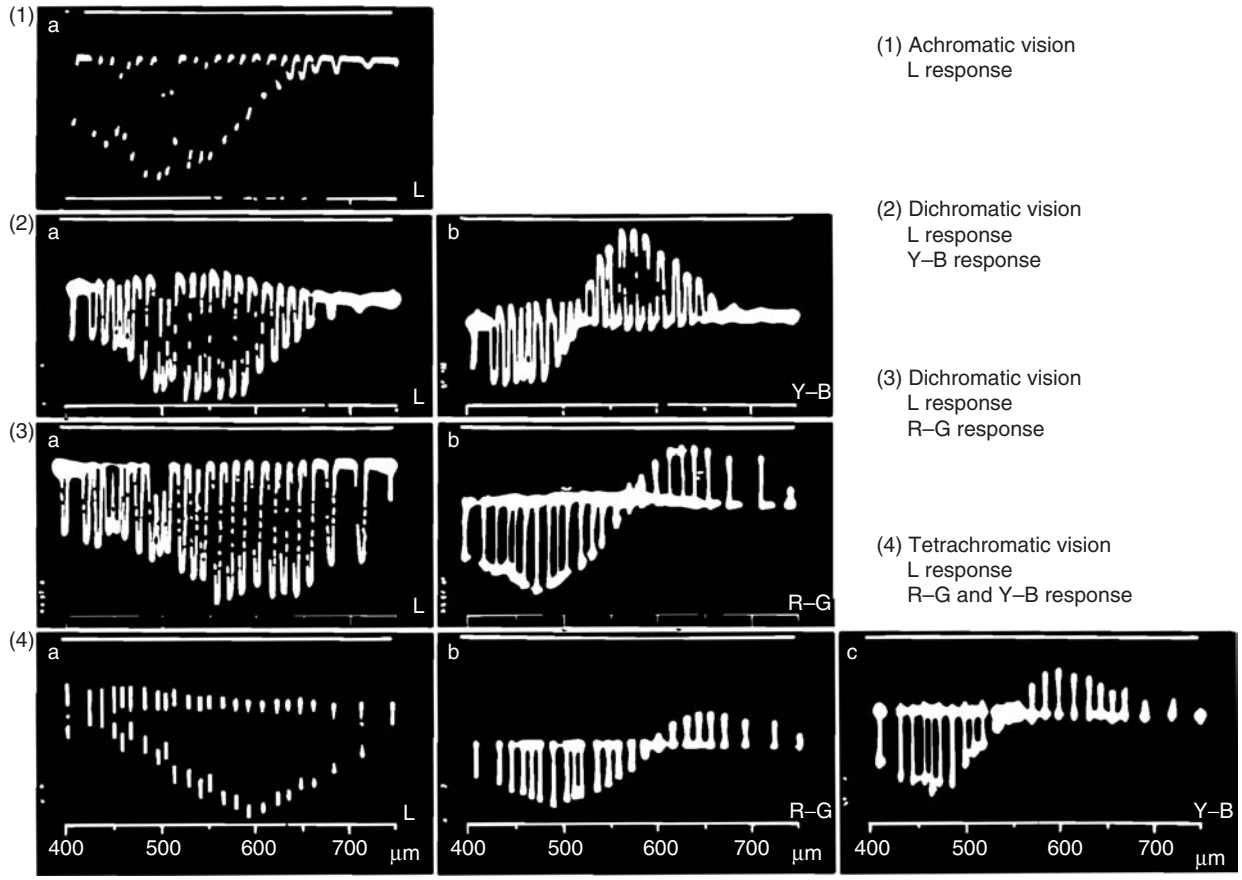

Figure 1 Responses from horizontal cells from various species of fish. Amplitude of response is plotted against wavelength, with hyperpolarization plotted downward and depolarization plotted upward. (1) Lutianidae gave only luminance responses. (2) Serranidae gave both luminance and Y–B responses. (3) Centropomidae gave luminance and R–G responses. (4) Mugilidae gave all three types of response. Reprinted from MacNichol EF and Svaetichin G (1958) Electric responses from the isolated retinas of fishes. *American Journal of Ophthalmology* 46: 26–46.

In the example given, the cell was excited by green light in the center of its receptive field, inhibited by red light in the center, excited by red light in the surround, and inhibited by green light in the surround. The reverse is also found. Excitation and inhibition tend to cancel each other out, so that these cells give little response to red or green light covering the whole receptive field and little response to white light in either center or surround. They respond maximally to color contrast. Moreover, the response to a gray spot in a red surround is similar to the response to a green spot in a gray surround, so that the response of the cells corresponds to simultaneous color contrast. Daw first showed that most ganglion cells are double opponent: subsequently Kaneko and Tachibana also found some double opponent responses in bipolar cells.

The connections for the double opponent surround of the bipolar and ganglion cells are provided by the horizontal cells in the outer plexiform layer, and the amacrine cells in the inner plexiform layer. The horizontal cells have terminals that are inserted into the cone pedicles, so that a feedback response can be recorded in individual cones. This tends to have a broad spectral sensitivity. Signals from cones and horizontal cells then feed forward onto the bipolar cells. It is not clear how far the responses of the bipolar cells would be double opponent if the inner retina was removed. However, amacrine cells contact the axon terminals of the bipolar cells, where contacts with ganglion cells are also made, and chloride channels sensitive to γ-aminobutyric acid (GABA) are found. Thus, it is not surprising that GABA antagonists affect the surround of double opponent responses.

These results show that considerable processing of information for color takes place in the retina in goldfish and carp. This is probably true for other vertebrate species, although the details have not been worked out as they have in goldfish and carp, but is not true in mammals. Double opponent cells are first found at the level of the lateral geniculate nucleus in ground squirrel and cat, and at the level of

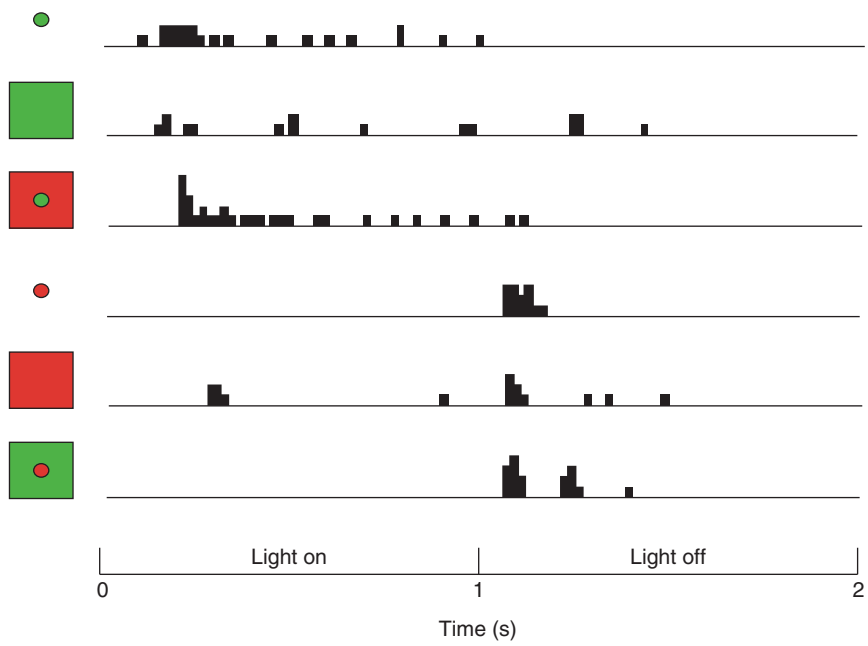

Figure 2 Response of a green on center double opponent ganglion cell from goldfish retina. Note that the size of the center of the receptive field for red is larger than that for green, consequently some responses were ON–OFF rather than pure ON or pure OFF.

the visual cortex in the macaque. However, other complications arise in the retinas of these species due to the presence of a fovea or visual streak.

Color Vision in Primates

We will now turn to the retinas of the human, and of the macaque, since vision in this species is close to that of humans. Both are trichromatic, with pigments peaking at close to the same wavelengths. In the human, there are several L cone pigments, peaking between 550 and 560 nm. There are also several M cone pigments, peaking between 530 and 536 nm. People with normal color vision will have one of the L cone pigments and one of the M cone pigments. There is less variation in the S cone pigment, which peaks around 415 nm.

The L and M cones are not arranged in a regular array (Figure 3). In some retinas, the arrangement is random. In others, there appears to be a clumping of cones with the same pigment, more than one would expect from a random arrangement. The ratio of L cones to M cones also varies considerably from retina to retina, between 1.1:1 and 16.5:1. The ratio can also vary from one location to another in the same retina. The ratio is correlated with their contribution to the electroretinogram, which is an electrical potential recorded from the whole retina that varies from individual to individual. Interestingly, in spite of these variations, the perception of the wavelength that produces a unique yellow does not vary. The clumping of L and M cones explains why subjects routinely misjudge the color appearance of tiny flashes of monochromatic light. The S cones, which are a small percentage (5–6%) of the total, form a more regular mosaic interspersed between the L and M cones.

Thirteen types of ganglion cell have been identified in the macaque retina that project to the lateral geniculate nucleus. Most types come in pairs, with two varieties, one of which arborizes in the inner part of the inner plexiform layer, and responds when light is turned on, the other of which arborizes in the outer part of the inner plexiform layer, and responds when the light is turned off. The midget ganglion cells are known to be color coded. Near the fovea they connect to a single bipolar cell which connects to a single cone. The bistratified cells are also known to be color coded, and so are the giant sparse monostratified cells. The parasol cells are not color coded. The large sparse, broad thorny, and narrow thorny cells have not yet been characterized.

Yellow/Blue Color Vision

The most clearly defined color-coded pathway within the primate retina is the blue ON yellow OFF pathway, represented by the small bistratified ganglion cell. To stimulate these cells, Dacey and Lee employed a red + blue stimulus alternating sinusoidally with a green stimulus, which produced no response in red/green color-coded cells, or cells that were not

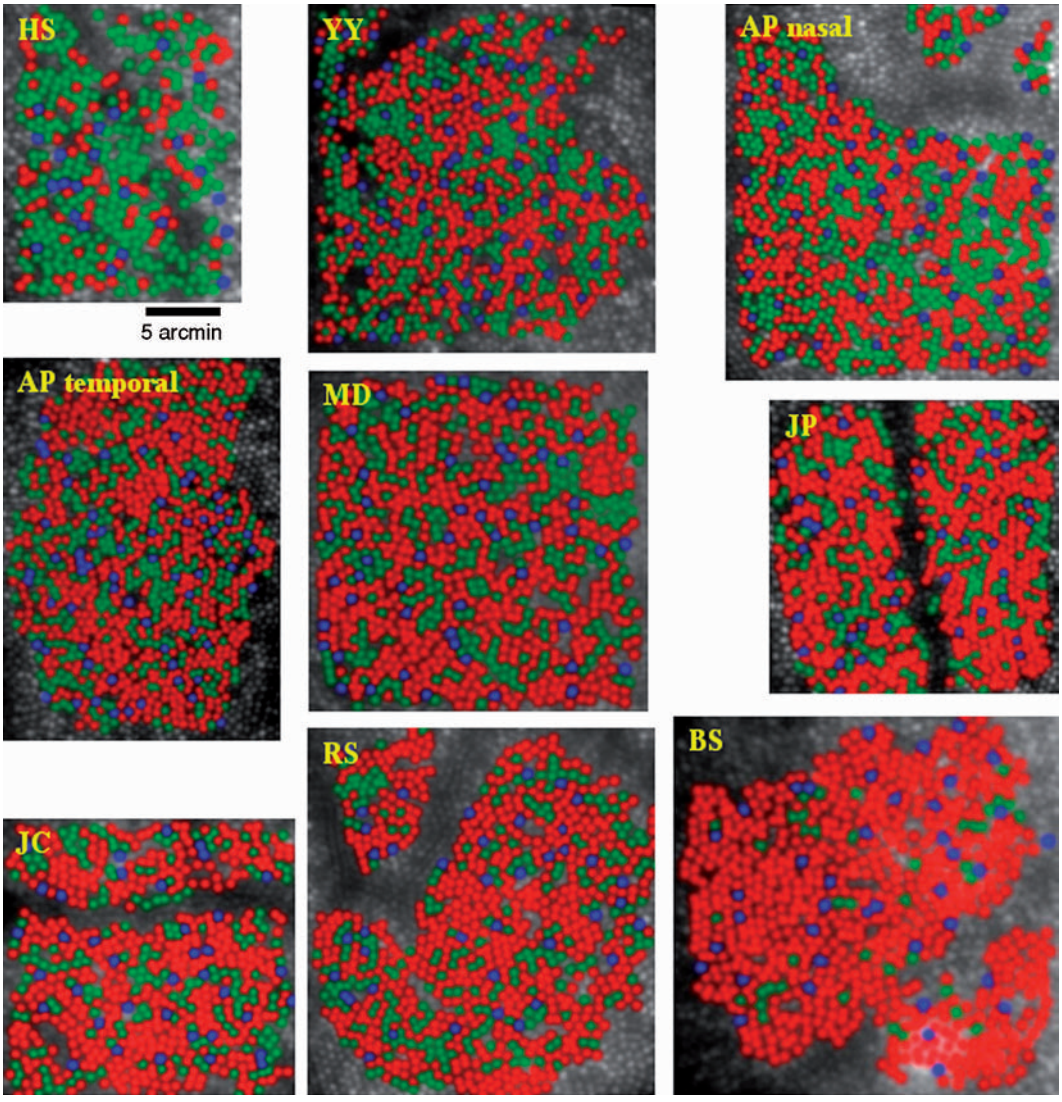

Figure 3 False color images showing the arrangement of L (red), M (green), and S (blue) cones in the retinas of different human subjects with normal color vision. Reprinted from Hofer H, Carroll J, Neitz J, Neitz M, and Williams DR (2005) Organization of the human trichromatic cone mosaic. *Journal of Neuroscience* 25: 9669–9679. Copyright 2005 by the Society for Neuroscience.

color coded. This stimulus produced a vigorous response in small bistratified cells. A flickering blue stimulus gave a response when the light was on, and a flickering yellow stimulus gave a response when the light turned off (Figure 4(c)).

This cell receives input from a cone bipolar cell that contacts the S cones in the outer plexiform layer, and arborizes in the inner part of the inner plexiform layer, explaining the blue ON response (Figure 5). It also receives input from two types of diffuse cone bipolar cells that contact L and M cones in the outer plexiform layer and arborize in the outer part of the inner plexiform layer, explaining the yellow OFF response. The spatial extent of the blue ON and yellow OFF responses is similar. However, this is not the whole story: if one applies AP4, which is a drug that interrupts transmission between the S cones and the S cone bipolar, much of the L + M cone response in the ganglion cell is abolished, as well as the S cone response. Thus, there must be a blue ON/yellow OFF response in the S cone bipolar. The OFF response probably comes from the H2 horizontal cell, which gets input from L, M, and S cones and makes lateral connections in the outer plexiform layer. Its receptive field is small – about the right size for the yellow OFF response of the small bistratified cell. Thus, there is one set of connections for the blue ON response and two sets of connections for the yellow-OFF response.

The large sparse bistratified ganglion cell also has a blue ON/yellow OFF response (Figure 4(b)). Its receptive field is larger than that of the small bistratified cell, corresponding to its larger dendritic arborization,

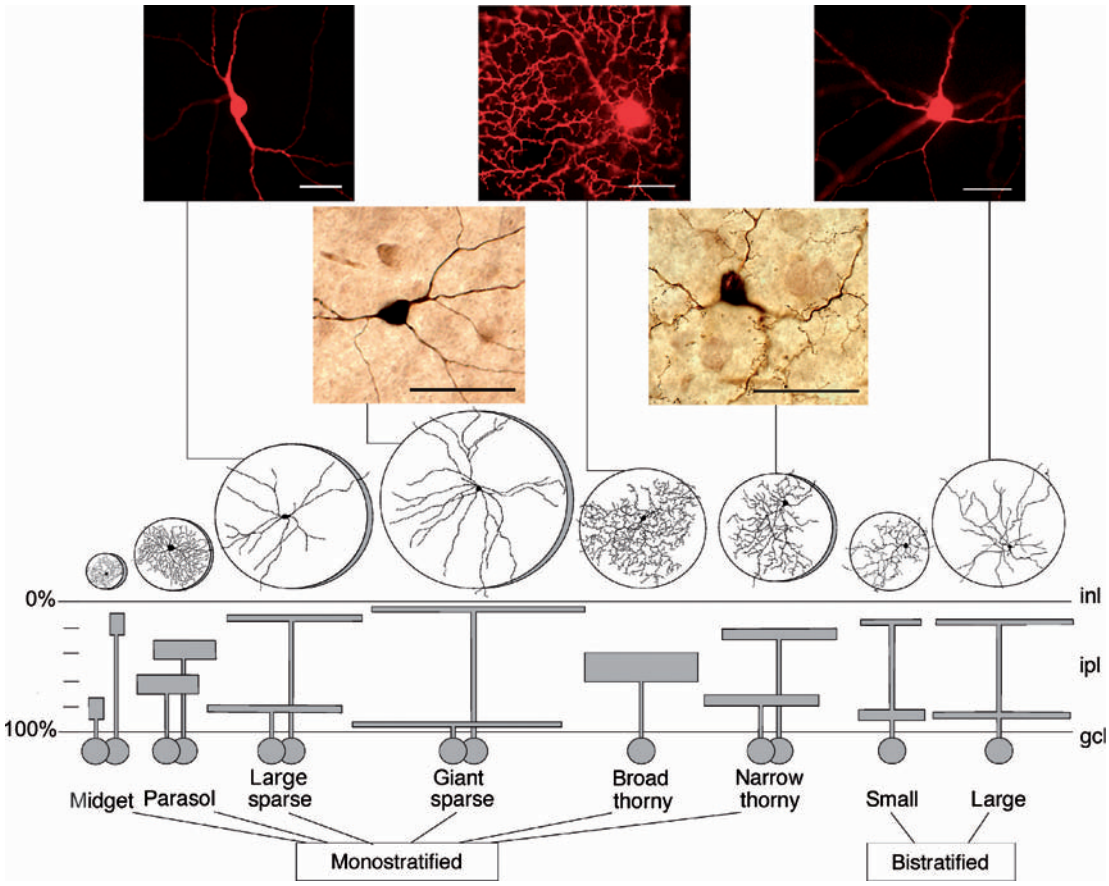

Figure 4 Responses from three types of blue/yellow opponent color ganglion cell in macaque retina. The large sparse monostratified cell giving an ON response to yellow and an OFF response to blue. The large sparse bistratified cell giving an ON response to blue and an OFF response to yellow. The small bistratified cell giving an ON response to blue and an OFF response to yellow. Scale bar = 50 μm. Reproduced from Dacey DM and Packer OS (2003) Colour coding in the primate retina: Diverse cell types and cone-specific circuitry. *Current Opinion in Neurobiology* 13: 421–427, with permission from Elsevier.

but otherwise the response is similar. The connectivity has not been worked out in detail, as has the connectivity of the small bistratified cell, but it could well be similar.

The blue/yellow system is phylogenetically older than the red/green system. Mammals that are dichromatic – that is, most mammals – have S cones and either M or L cones. It is only in recent evolution that both L and M cones have been developed. There is a long history of recordings in the mammalian retina showing blue ON responses but few or no blue OFF responses. All the blue/yellow color-coded cells recorded in cat and rabbit retinas are blue ON. Some early recordings of macaque ganglion cells showed some blue OFF cells, but most blue/yellow cells recorded there have also been blue ON. Thus, there seems to be an asymmetry in the system – either the blue ON cells are larger, so that they are recorded more frequently, or they are more numerous.

The macaque blue OFF cells have now been identified, with a variety of receptive field structures, including the large and giant sparse monostratified cells (Figure 4(a)).These cells also come in two types, ramifying at the extreme inner and outer borders of the inner plexiform layer, but do not have the usual dichotomy into ON and OFF, since they are all blue OFF. They have large receptive fields corresponding to their dendritic arborization. Their connectivity has not yet been established, since no blue OFF bipolar cells have been identified. The giant cells code for luminance over both night and daylight vision, also contain melanopsin, and give a slow response when all input from photoreceptors is abolished. Thus, one might expect that they have to do with some subcortical function such as the pupillary response or circadian rhythms, but they also project to the lateral geniculate nucleus.

Red/Green Color Vision

The red/green system is represented by the midget ganglion cells. As pointed out above, these are contacted by a single bipolar cell and a single cone in the

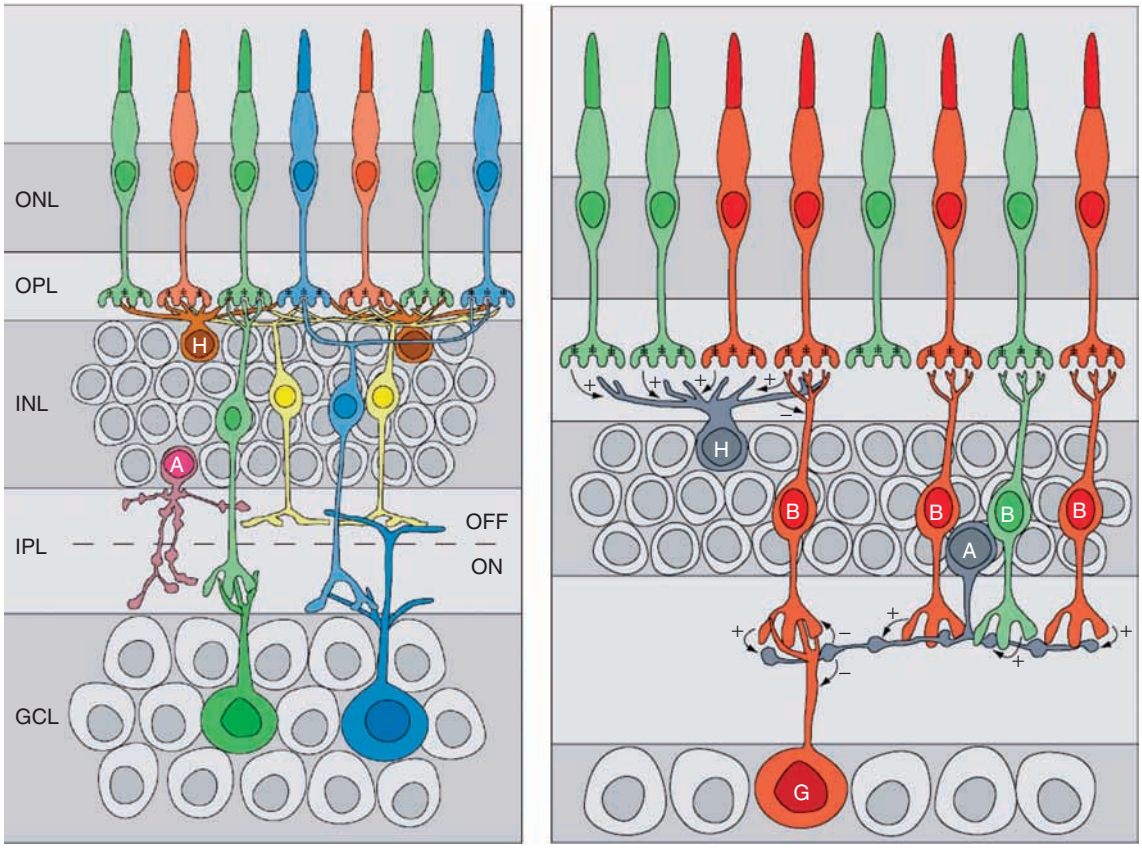

Figure 5 Wiring for the blue ON/yellow OFF small bistratified ganglion cell, and for the green ON, red OFF midget ganglion cell (left). Lateral input to the midget system from both horizontal and amacrine cells (right). Reprinted from Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, 2 vols, figures 64.4, 64.12, and 63.1, © 2003 Massachusetts Institute of Technology, by permission of the MIT Press.

fovea (Figure 5, left). There are four types – with responses in the center giving red ON, green ON, red OFF, and green OFF. Following the usual rule, the ON types ramify in the inner part of the inner plexiform layer, and the OFF types in the outer part. For each type there is a corresponding midget bipolar cell. The surround response of the midget ganglion cells is larger than the center, of opposite color and opposite polarity (e.g., red ON center and green OFF surround). The surround response could be brought in by horizontal cells, or by amacrine cells (Figure 5, right). The horizontal cells that make lateral connections to the midget system are the H1 cells, which receive input from both L and M cones, and if they are the sole pathway for the surround response, then it should not be specific for red or green alone. The amacrine cells that make lateral connections to the midget system also appear not to discriminate between L and M systems. The S cone input to this system is comparatively minor.

The midget ganglion cells project to the lateral geniculate nucleus where again there is a 1:1 connection between ganglion cell and geniculate cell. In the geniculate, measurements of the surround response suggest that it is specific to either the L or M cones. There is therefore a mystery as to how this system gets a surround response that is specific to a single type of cone, which has not yet been resolved. Some possibilities are (1) that foveal H1 horizontal cells receive a more specific input than those further out in the retina (not many foveal H1 cells have been recorded); (2) that some amacrine cells receive an input from the L or M cone systems alone (the study suggesting that amacrine cells associated with the midget system get input from both L and M cones was an electron microscope study, with a limited number of examples); or (3) that there are lateral connections within the lateral geniculate nucleus bringing in a specific surround response. Further experiments will be needed to resolve this conundrum.

To summarize, there is a red/green opponent color system with a center responding to one color, and a surround responding to the other, represented by the midget system. There is a blue/yellow opponent

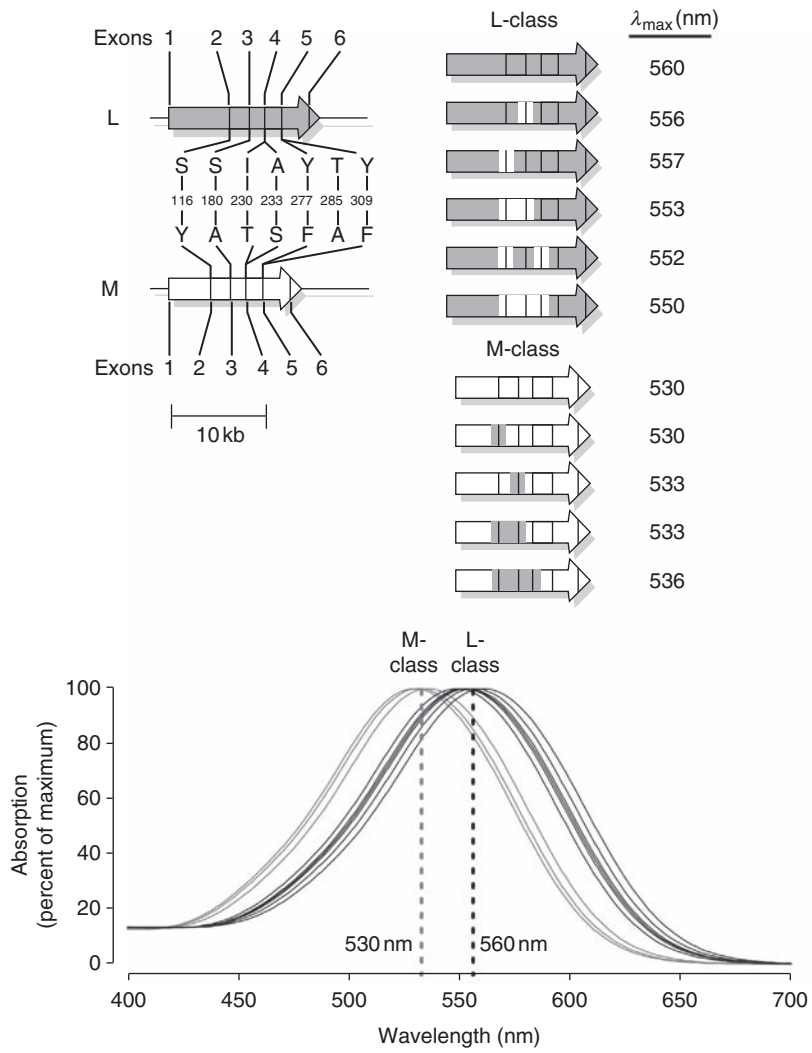

Figure 6 L and M cone pigments. Genes shown at top. Shaded and unshaded arrows represent L- and M-photopigment genes, respectively. Spectral sensitivities are shown at bottom. Reprinted from Chalupa LM and John SW (eds.) *The Visual Neurosciences*, 2 vols, figures 64.4, 64.12, and 63.1, © 2003 Massachusetts Institute of Technology, by permission of the MIT Press.

color system, with the spatial extent of the two color responses being the same, represented by the bistratified cells and the sparse cells. Horizontal cells are not opponent color, as they are in the fish. Ganglion cells and bipolar cells are opponent color, but not double opponent, as they are in goldfish and carp.

Color Vision in the Periphery of the Retina

So far the discussion has concentrated on the area near the fovea, where color vision is most acute. However, there are considerable variations in color vision across the retina in primates that have a fovea. There is a small area (about 1/8°) in the center of the fovea in humans where there are no S cones, so that blue is not perceived. Vision is trichromatic for 20–30° out from this. Red/green vision becomes deficient further out, requiring larger stimuli to see color, to 70–80° from the fovea, where the system is color blind. As one brings a red object in from the far periphery, it is seen first as salmon pink or yellow, then eventually as red.

The red/green deficiency beyond 20–30° from the fovea is probably due to a different connectivity of the midget ganglion cells. In this region there are midget bipolar cells that are connected to a single L or M cone, but the midget ganglion cells receive input from several of both types of bipolar cell. Interestingly, the L and M input to the midget ganglion cells varies in the peripheral retina, but this is believed to be due to the clumping of L and M cones. The L and M input to parasol ganglion cells and H1 horizontal cells also varies, and is similar for the three types of

cell in the same part of the retina, from experiments by Dacey and his colleagues. These results suggest that the midget ganglion cells are no more color selective than the parasol cells or the H1 cells in the same part of the retina. Martin and his colleagues, on the other hand, have recorded opponent color responses in midget ganglion cells from peripheral retina, and suggest that the color capability of these cells may even exceed that seen in psychophysical experiments. However, more recent work from this group shows that there is a variety of responses from midget ganglion cells in the periphery, with some showing no color opponency at all. Further experiments will be needed to clarify the situation.

Color Blindness in the Retina

Most color blindness is a retinal defect. There is one of six different L cone pigments in L cones, and one of five different M cone pigments in M cones (Figure 6). People with normal color vision will have one of the L cone pigments, and one of the M cone pigments. People with defective protan vision have none of the L cone pigments, and may have only one of the M cone pigments (protanopes) or two of them in separate cones (protanomalous). People with defective deutan vision have none of the M cone pigments, and may have only one of the L cone pigments (deuteranopes) or two of them in separate cones (deuteranomalous). Thus, protanopes and deuteranopes can match any color with a mixture of two wavelengths, while protanomalous and deuteranomalous people require three. The genes for the L and M cone pigments are on the X chromosome, and this form of color blindness is quite common in males, adding up to about 8% of the population. People missing the S cone pigment (tritanopes) are much more rare, and it is not X-linked. What difference these deficits make to the connections of the cones to horizontal, bipolar, and ganglion cells is not known, but there seems to be little reason why there should be any change.

Acknowledgments

The author would like to thank David Calkins for comments and **Figure 5**, Dennis Dacey for **Figure 4**, Heidi Hofer for **Figure 3**, and Maureen Neitz for **Figure 6**.

See also: Color Vision; Fovea: Primate; Representation of Color; Retinal Amacrine Cells; Retinal Bipolar Cells; Retinal Pharmacology: Inner Retinal Layers; Retinal Glia; Retinitis Pigmentosa.

Further Reading

Calkins DJ (2004) Linking retinal circuits to colour opponency. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 989–1002. Cambridge, MA: MIT Press.

Dacey DM and Lee BB (1994) The 'blue-on' opponent pathway in primate retina originates from a distinct bistratified ganglion cell type. *Nature* 367: 731–735.

Dacey DM and Packer OS (2003) Colour coding in the primate retina: Diverse cell types and cone-specific circuitry. *Current Opinion in Neurobiology* 13: 421–427.

Daw NW (1967) Goldfish retina: Organization for simultaneous color contrast. *Science* 158: 942–944.

Diller L, Packer OS, Verweij J, McMahon MJ, Williams DR, and Dacey DM (2004) L and M cone contributions to the midget and parasol ganglion cell receptive fields of macaque monkey retina. *Journal of Neuroscience* 24: 1079–1088.

Hofer H, Carroll J, Neitz J, Neitz M, and Williams DR (2005) Organization of the human trichromatic cone mosaic. *Journal of Neuroscience* 25: 9669–9679.

Jacobs GH (2004) Comparative color vision. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 962–973. Cambridge, MA: MIT Press.

MacNichol EF and Svaetichin G (1958) Electric responses from the isolated retinas of fishes. *American Journal of Ophthalmology* 46: 26–46.

Martin PR, Lee BB, White AR, Solomon SG, and Ruttiger L (2001) Chromatic sensitivity of ganglion cells in the peripheral primate retina. *Nature* 410: 933–936.

Neitz M and Neitz J (2004) Molecular genetics of human color vision and color vision deficits. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 974–988. Cambridge, MA: MIT Press.

Solomon SG, Lee BB, White AR, Ruttiger L, and Martin PR (2005) Chromatic organization of ganglion cell receptive fields in the peripheral retina. *Journal of Neuroscience* 24: 4527–4539.

Retinal Development: An Overview

H R L Napier and B A Link, Medical College of Wisconsin, Milwaukee, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Ontogenesis and Organization of the Vertebrate Retina

Development of the vertebrate eye involves the paired bulging of the forebrain to form optic vesicles. Each optic vesicle extends laterally where it contacts the surface ectoderm. Reciprocal inductive signals between the surface ectoderm and optic vesicle result in invagination of the optic vesicle to form the optic cup. The inner layer of the optic cup is then patterned to form the neural retina distally and the ciliary body and iris more proximally (Figure 1(a)).

Cell Types of the Retina

The vast majority of cells in the mature neural retina develop from neuroepithelial progenitor cells of the optic cup, the exceptions being vascular cells and some types of immune cells. When fully differentiated, the neural retina is a highly ordered, laminar structure comprising six major neuronal cell types (retinal ganglion cells, amacrine cells, bipolar cells, horizontal cells, and rod and cone photoreceptors) and one intrinsic glial cell type (Müller glial cells) (Figure 1(b)). Laminar organization of the retina facilitates the initial steps in processing visual information. Achieving this exquisite cytoarchitecture requires intricate signaling during development. Coordination of proliferation, cell cycle exit, cell type determination, directed cellular migration, and cellular morphogenesis enables the appropriate numbers, correct laminar organization, and proper synaptic connections within the mature retina.

Proliferation of Retinal Progenitor Cells

Interkinetic Nuclear Migration

During development of the vertebrate retina, the optic cup neuroepithelial cells undergo rapid, proliferative expansion. This proliferative phase is marked by several intriguing cellular behaviors that have been hypothesized to provide cell fate information to the multipotent progenitor cells. One remarkable proliferative cell behavior is known as interkinetic nuclear migration (Figure 2(a)). This is the process in which the nuclei of neuroepithelial cells migrate in an apical–basal manner and in phase with the cell cycle. M phase and cytokinesis are confined to the apical surface near the retinal pigment epithelium (RPE), whereas G_1, S, and G_2 phases occur at more basal locations. This process is heterogeneous among neuroepithelial cells. Variation has been observed in the time required to progress through one round of interkinetic nuclear migration (which equals the cell cycle period), as well as in the distance of nuclear migration. Studies suggest interkinetic nuclear migration aids in the selection of progenitors that will produce neurons in the next mitosis. This is achieved by establishing asymmetry between neuroepithelia with regard to cell body position during critical 'neurogenic' windows of the cell cycle.

Oriented Cell Divisions

A second heterogeneous cell behavior associated with progenitor cell proliferation is regulated mitotic division plane orientation. Although neuroepithelial cell divisions are confined to the apical ventricular zone near the RPE, the orientation of cleavage axis can vary among progenitors (Figure 2(b)). Typically, retinal progenitor cells divide in either an apicobasal or a horizontal manner. Genetic and imaging studies in invertebrates have indicated that mitotic division plane in neural progenitors, and other proliferate cells, is important for symmetric versus asymmetric cell fate outcomes. Symmetric cell divisions create daughter cells that assume the same fates, whereas asymmetric cell divisions produce daughters of different fates. Symmetry in cell fates is subject to definition. For example, symmetry may relate to cell cycle exit and neurogenesis. In this case, symmetric cell divisions produce two daughter cells where both remain proliferative or both leave the cell cycle and become neurons. Alternatively, symmetry can also relate to cell type fate; for example, divisions that produce two ganglion cells or equivalent daughters of other cell types. Studies on cell division orientation in the developing vertebrate retina indicate that this cell behavior influences cell type choice. Mechanistically, it appears that the plane of cell division can equally (symmetrically) or differentially (asymmetrically) partition fate determinants to daughter cells. In the vertebrate retina, a modulator of the Notch signaling pathway, the Numb protein, is the flagship example of this phenomenon. Numb associates with the apical surface of retinal progenitor cells and immunocytochemistry has shown that this protein can differentially segregate between daughter cells. Interestingly, and in contrast to invertebrates, no other examples of vertebrate retinal cell fate determinants have been shown to partition asymmetrically between daughter cells.

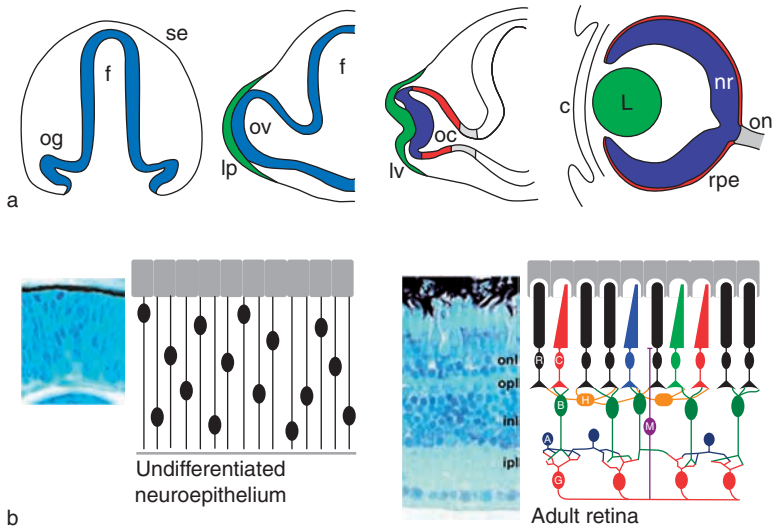

Figure 1 Development of the eye and neural retina. (a) Vertebrate eye development initiates with the paired extension of the forebrain to form optic grooves (og). Each groove extends laterally to form an optic vesicle (ov), which contacts the surface ectoderm and induces the formation of a lens placode (lp). Reciprocal inductive events lead to the invagination of the optic vesicle and lens placode to form the optic cup (oc) and lens vesicle (lv), respectively. c, cornea; L, lens; nr, neural retina; rpe, retinal pigment epithelium; on, optic nerve. (b) Histological cross sections of 24 hpf and adult zebra fish neural retinas with schematic representations of the retina at these stages. At 24 hpf, the retina is a pseudostratified neuroepithelium. The adult, laminated retina can be divided into the outer nuclear layer (onl), which includes rod (R) and cone (C) cell bodies; the inner nuclear layer (inl), which includes horizontal (H), bipolar (B), and amacrine (A) cell bodies; and two plexiform or synaptic layers – the outer plexiform layer (opl), which separates the nuclear layers, and the inner plexiform layer (ipl). M, Müller glia.

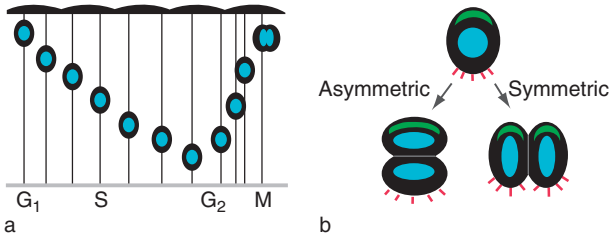

Figure 2 Cell behaviors of the proliferative neuroepithelium. (a) Interkinetic nuclear migration (blue nucleus) is schematized in one retinal neuroepithelial cell over the course of the mitotic cycle. M phase occurs at the apical region adjacent to RPE cells, whereas G_1, S, and G_2 phases take place at more basal locations. (b) Regulated mitotic division plane orientation is modeled for either an asymmetric or a symmetric division. Note that polarized fate determinants associated with either the cell surface (red molecules) or localized intracellularly (green molecules) can be equally or differentially segregated based on the plane of cytokinesis.

Molecular Regulation of Retinal Cell Proliferation

The factors that promote and regulate cell proliferation in the vertebrate retina are beginning to be discovered. Secreted growth factors such as Fibroblast growth factor (FGF), Hedgehog, and Wnt family members, as well as transmembrane proteins such as Notch function in part to control cellular proliferation, but they control other cellular processes as well. Surprisingly, secreted amino acids such as glutamate or nucleotides such as ATP also promote cell proliferation in retinal progenitors cells. In addition to extrinsic influences, specific transcription factors expressed in retinal neuroepithelia regulate parameters of the mitotic cycle. Proliferative roles for both homeodomain- and basic helix–loop–helix (bHLH)-type transcription factors have been demonstrated for retinal progenitor cells. Thus, both extrinsic and intrinsic influences converge to exert an effect on core cell cycle machinery such as the cyclins, cyclin kinases, and cyclin inhibitors. This is a common theme in development: the balance and interactions between intrinsic cell programs and extrinsic influences. This concept of complex signal integration is discussed in more detail later in the context of the competency model for retinogenesis. Specifically for retinal proliferation, however, regulatory molecules influence both

the kinetics of cell cycle progression (how fast cells pass through the various phases of the cell cycle) and neurogenic decisions (whether to remain in the cell cycle or exit mitosis and initiate migration and differentiation).

Spatial Control of Retinogenesis

The overall process of retinogenesis, from cell cycle exit to cellular differentiation, is coordinated in space and time. The first cells to exit the mitotic cycle do so at the center of the optic cup, although dorsal or ventral biases in where this process begins can vary between species. The initiation of retinogenesis, however, appears to always depend on FGF expression in the neuroepithelium. The patch of FGF expression is regulated in part by adjacent midline mesoderm-derived signals. From this central FGF organizing center, cell cycle exit proceeds in a wavelike manner toward the periphery of the optic cup. At any point within the wave, however, only a subset of progenitor cells are selected to produce postmitotic progeny. The remaining cells stay in the cell cycle to produce late-born cell types. Why some cells remain mitotic and others begin neurogenesis is an active area of research.

Lineage and Birth Date Relationships

Experiments *in vivo* and in cell culture from all vertebrate species studied have demonstrated the multipotency of retinal neuroepithelial cells. Lineage-tracing experiments using either tagged replication incompetent retroviruses or microinjection of nontransferable markers have shown that single neuroepithelial cells can divide to produce all six neuron types and Müller glia. Another evolutionarily conserved feature of retinogenesis is the link between birth date and cell type (Figure 3). Although subtle differences exist between species in the exact order of cell types generated over developmental time, retinal ganglion cells genesis is normally followed by the differentiation of horizontal cells, cone photoreceptors, and amacrine cells, followed by bipolar cells, rod photoreceptors, and Müller glia. This birth order is also reflected by clonal analyses from lineage studies. Clonal descendents from neuroepithelia labeled at early stages of development ('early' progenitors) are biased to produce ganglion cells, amacrine, horizontal cells, and cone photoreceptor cells. Progeny from 'late' progenitors are biased to produce bipolar cells, rods, photoreceptor cells, and Müller glia. Despite the trend of birth order, different cell types can and do differentiate simultaneously (Figures 3 and 4). Finally, lineage studies and cell culture experiments have suggested that the majority of retinal neuroepithelia remain multipotent until the last cell cycle prior to terminal mitosis or, for some cell types, just following cell cycle exit. Overall, observations from lineage and birth order studies, as well as data from other experiments, suggest that the timing of cell cycle exit, intrinsic lineage bias, and influences from adjacent cells all function in cell fate determination.

Competency Model for Retinogenesis

The competency model has emerged as a general paradigm for retinogenesis and posits that retinal neuroepithelia progress through temporal states in which the progenitor is competent to generate only a subset of retinal cell types (Figure 4). This bias in cell type competency is influenced by both determinant intrinsic and modulatory extrinsic factors, each of which are dynamic in nature. In other words, within the developing retina, the type, levels, and expression domain of secreted factors change with time and exert differential influences depending on the changing intrinsic state of the neuroepithelial cell. The compentency model is applicable to two interrelated cell fate decisions that neuroepithelia make during their maturation process: (1) when to exit the cell cycle and (2) what cell type to become. Although

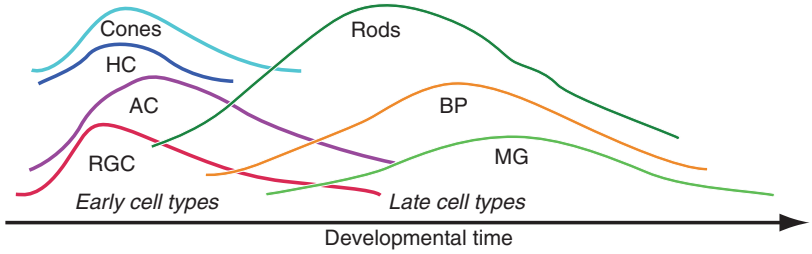

Figure 3 Birth orders for cell types of the mammalian retina; schematic showing the timing of cell cycle exit (birth date) for classes of retinal cells. The areas under the curves represent the relative proportions of cell types typically found in nocturnal (rod-dominated) rodents. Subtle differences exist in the exact birth orders between various species. Note the overlapping timing for genesis of multiple cell types and the relative clustering of 'early' and 'late' cell types.

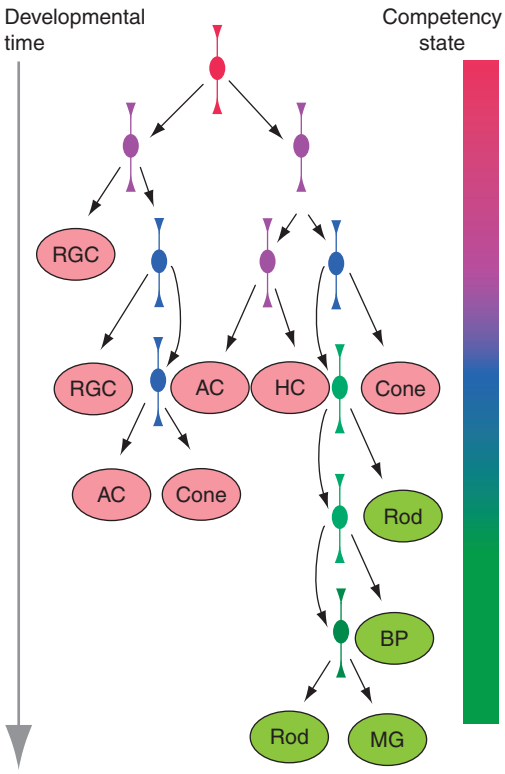

Figure 4 Competency model for retinal cell fates. As retinal progenitor cells divide and give rise to other proliferative progenitor cells (colored, elongated neuroepithelial cells), the intrinsic cellular competency to differentiate to particular classes of postmitotic cells (labeled ovals) changes. The changing colors represent changes in cellular competence. Postmitotic cells have been grouped into early cell types (pink ovals) or late cell types (green ovals). Note that multiple cell types can be generated at the same time of development and that single cells give rise to complex combinations of differentiated cell types.

general in nature, some of the molecules and signaling pathways underlying intrinsic and extrinsic components of the competency model have been discovered.

Atoh7 and Regulation of Cellular Competence

The bHLH transcription factor Atoh7 (previously known as Ath5) serves as a specific example of the molecular underpinnings of cellular competence for cell cycle exit and cell type fate determination. Atoh7 was first discovered based on its striking expression pattern and loss-of-function phenotype. The expression of *atoh7* is specific to the neural retina and commences in neuroepithelial cells in their last cell cycle preceding a neurogenic cell division (where one or both daughter cells exit the cell cycle and differentiate as neurons). Disruption of *atoh7* in mouse, frog, and fish results in a dramatic loss of retinal ganglion cells. Careful analysis of *atoh7* mutants showed that in addition to the loss of retinal ganglion cells, there was a marked decrease in the proportion of progenitor cells leaving the cell cycle. Importantly, *atoh7* linage-marking studies demonstrated that *atoh7* is expressed transiently, just before cell cycle exit in multiple cell types, but not all cell types. These data suggest that Atoh7 is required for cell cycle exit and biases cells toward particular cell type fates. The exact cell type depends on the co-expression of other transcription factors. Interestingly, ectopic expression of *atho7* alone does not drive cell cycle exit. This is consistent with Atoh7 as important for neurogenic competency, while also providing influence to eventual cell type fate. The initial expression of *atoh7* appears to be established at a time much earlier than its actual expression. This 'clock-like' mechanism is probably set by signals provided at the earliest stages of optic vesicle formation by cells associated with the embryonic midline axis. However, termination of *atoh7* expression is regulated in part by extrinsic proteins. GDF11, a secreted growth factor, has been shown to extrinsically influence the expression of *atoh7*. Mice with deletions in *gdf11* show an extended expression of *atoh7* and a concomitant increase in early born retinal cell types. Therefore, GDF11 regulates the temporal window during which retinal progenitors retain competence to produce distinct neural progeny. Other factors, both autonomous and non-autonomous, interact to provide the complete regulatory network that enable the appropriate numbers and cell types to be generated during retinal development.

Postmitotic Cell Migration

Modes of Migration

Following terminal mitosis at the apical surface, newborn retinal cells migrate to their appropriate laminar position within the retina. Classic histological and more recent time-lapse experiments have demonstrated that the various cell types of the retina employ multiple strategies during postmitotic cell migration. The modes of cell migration can be classified as guided cell migration, somal translocation, and unconstrained migration (Figure 5). These are not necessarily mutually exclusive forms of motility, and some cells may use multiple strategies while migrating. Direct evidence exists for the latter two modes of cell positioning. Within the retina, somal translocation has been demonstrated for retinal ganglion cell and bipolar cell precursors. Some newborn amacrine cells have been shown to move using unconstrained migration, in which dynamic and nonpolarized processes appear to sample the local environment

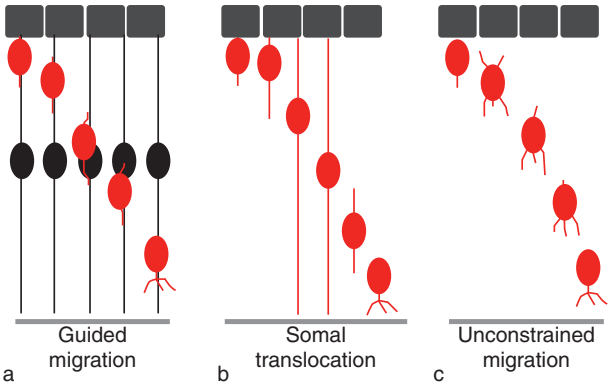

Figure 5 Modes of cell migration. Postmitotic cells (red) can migrate to their appropriate laminar position in the retina in one of three ways. (a) In guided migration, postmitotic cells use adjacent neuroepithelial cells as a scaffold to move from the apical surface toward their final destination. (b) Somal translocation occurs independently of adjacent cells. After its terminal mitosis a newly formed neuron extends processes to both the apical and the basal cell surfaces. Subsequent to this, the cell body translocates along this process until reaching its appropriate position within the retina, where it then retracts its processes. (c) In unconstrained migration, newly formed neurons migrate independently of other cells or cell processes, often extending neuritis in an 'exploratory' manner.

en route to the inner nuclear layer. It is likely that neuroepithelial-guided cell migration is utilized, to some extent, for both nuclear translocation and free migration. This form of guided cell migration is similar to that in the developing cerebral cortex, where glial cells, instead of adjacent neuroepithelia, provide the substrate and cues for cell positioning.

Within the retina, all forms of migration are restricted primarily to the radial direction along the apical–basal axis. However, clonal analysis using various progenitor cell marking techniques has demonstrated subtle lateral (or tangential) displacements of ganglion cells, horizontal cells, cone photoreceptors, and amacrine cells. This tangential displacement is thought to ensure an even 'planar' distribution of specific classes of cell subtypes and facilitate appropriate synaptic partnering. Although migration is largely restricted to the radial direction, migration is not restricted to one direction because some postmitotic cells show bidirectional movements. For example, some horizontal cells migrate first to the basal surface and then return to their appropriate outer retinal position.

Importance of Cell Polarity and Basement Membranes

Although much has yet to be discovered regarding the mechanisms and regulation of cellular migration in the retina, some important molecules and signaling pathways have been identified. One important class of molecules are those associated with apical cell junctions of neuroepithelia. The apical junctional complex includes the proteins Pard3, Pard6, Prkci, Mpp5, and Crb. Mutations in genes encoding these proteins result in cell positioning defects. Interestingly, studies *in vivo* have demonstrated that the apical junctional complex functions noncell autonomously in the retina for cell migration. This suggests that this signaling complex either regulates the adhesion between cells or facilitates polarized secretion of guidance molecules. In addition, some of these mutants result in disruptions to the RPE, exposing Bruch's membrane, which is the basement membrane of RPE cells. Basement membranes are distinct, thin sheets of extracellular matrix, rich in specific isoforms of collagens and laminins. In these mutants, ganglion cells become misdirected and migrate apically rather than basally. Guidance cues associated with the exposed Bruch's membrane may lead to these migration errors. Direct experiments have demonstrated the importance of the retinal basement membrane in directed cell migration. For example, transient disruption of the basement membrane in chick embryos results in lamination defects. Similarly, mutations in laminin α_1 in zebra fish cause migration defects for cone photoreceptors. Detailed characterizations for how the basement membrane translates into intracellular responses are limited. However, it is likely, based on cell culture work, that membrane-bound receptors such as dystroglycan and integrins, as well as their intracellular associated signaling molecules such as integrin-linked kinases and focal adhesion kinases, mediate directed cell migration.

Regulation of Cellular Differentiation

Coordination of Differentiation

Once a retinal cell reaches its appropriate laminar position, differentiation and morphogenesis begins. However, overt cellular differentiation does not always commence immediately. In addition to the spatial–temporal waves of cell cycle exit, cellular morphogenesis also progresses from central locations to the periphery in a wavelike manner. This wavelike progression of cellular morphogenesis is not a simple consequence of birth date. For example, in mice, expression of rhodopsin and other markers of rod photoreceptor differentiation are independent of birth date. Because rod genesis is very protracted over the course of retinogenesis, adjacent photoreceptor cells can have birth dates that differ by days. Despite differences in the time of terminal mitosis, differentiation and morphogenesis still occurs at approximately

the same time for adjacent cells. Two studies suggest signaling pathways important for coordinated differentiation. The SOCS3 protein (suppressor of cytokine signaling 3) has been shown to regulate and coordinate the timing of photoreceptor signaling by modulating the STAT signaling pathway. More generally, downregulation of the Notch pathway is important for multiple retinal cell types to begin morphogenesis.

Chromatin Remodeling

Chromatin remodeling factors are a diverse class of molecules that may be regulated by differentiation transducers such as the STAT and Notch pathways to coordinate differentiation programs within cells. Genetic analysis of the Brahma chromatin remodeling complex has revealed an essential role in cellular differentiation for all retinal neurons. The Brahma complex is part of the SWI/SNF group of chromatin remodeling complexes. These are large macromolecular complexes that use the energy of ATP to slide nucleosomes and create local alterations in chromatin. In addition to opening chromatin, Brahma subunits can also bind to gene-specific transcription factors and coordinate ocular differentiation programs. Other chromatin remodeling factors with demonstrated roles in global retinal cell differentiation are Dnmt1 (a DNA methyl transferase), Suv39h1 (a histone methyl transferase), and Hdac1 (a histone deacetylase).

Cell Type-Specific Differentiation

In addition to these global coordinators of retinal cell differentiation, a multitude of cell type-specific transcription factors has been shown to regulate particular differentiation programs. Many of these 'cell type-specific' transcription factors, however, are dynamic in expression and function in combination with other factors to control multiple steps of retinogenesis. The homeodomain transcription factor Ceh10 (also known as Chx10) serves as a prime example of the pleiotropy found in retinal differentiation factors. At the optic cup stages, *ceh10* is expressed in all proliferative neuroepithelia. This protein is required for normal proliferation. Following cell cycle exit, *ceh10* is downregulated in most cells but upregulated in differentiating bipolar cells. When *ceh10* is deleted, progression through the mitotic cell cycle is very slow and bipolar cells are missing. In wild-type bipolar cells, Ceh10 binds promoters of bipolar cell-specific target genes to positively regulate gene transcription. Within these same cells, Ceh10 also binds rod photoreceptor cell-specific promoters to inhibit their expression. Genetic complementation using Ceh10 pathway mutants demonstrated that the functions of Ceh10 in proliferation and differentiation are distinct. Interestingly, ectopic expression of *ceh10* in retinal progenitors promotes bipolar cell fates at the expense of rod photoreceptor cells. Cumulatively, these studies demonstrate that Ceh10 functions not only in cellular differentiation but also in cell proliferation and cell type fate decisions. Other factors that regulate specific cell differentiation pathways show similar complexity in function.

Protein Trafficking and Cellular Differentiation

In addition to transcriptional regulation during cell differentiation, protein complexes that drive cell differentiation and morphogenesis are beginning to be resolved. One of the most dramatic examples of cellular morphogenesis is that of retinal photoreceptors. Newly postmitotic photoreceptors are nearly spherical in shape and reside at the apical surface of the retina. However, as photoreceptors begin to differentiate, significant shape changes take place, including the dramatic growth of outer segments. Outer segments are elongated stacks of membranous disks and folds that contain components for phototransduction. Among the first features of photoreceptor morphogenesis is the establishment of the connecting cilium. The connecting cilium is a primary cilium composed of microtubules arranged in a 9 + 0 array. Because no protein synthesis occurs in the outer segment, an elaborate protein/membrane transport mechanism exists to both build and maintain photoreceptor outer segments. Trafficking of outer segment components occurs through the connecting cilium and along the backbone microtubules. The transport process is shared with other cilia-based cell appendages and is known as intraflagellar transport (IFT). IFT was discovered through genetic and biochemical studies of *Chlamydomonas* (green alga) and found to be the mechanism responsible for establishing and maintaining flagella. In photoreceptors, IFT particles composed of two macro-multiprotein complexes nucleate at the base of the connecting cilium where the basal body resides. IFT particles docked at the basal body and loaded with outer segment cargo are assembled onto kinesins, plus end-directed microtubule-based motors, to enable outward movement of the building blocks for the outer segment. Dyneins, minus end-directed microtubule-based motors, return the cargoless IFT particles, as well as the kinesin motors, to the basal body. Many questions remain regarding the details of IFT in photoreceptors, but it is certain that this process is essential for photoreceptor outer segment differentiation.

Neuronal Process Outgrowth and Synaptogenesis

Another obvious consequence of cell differentiation in the retina is the formation and elaboration of neuronal processes. Much of our knowledge of axon and dendrite development in the retina has been derived from serial electron microscopy analyses. Now, time-lapse experiments in mammalian slice culture and in zebra fish are revealing much of the complexity of process outgrowth and synaptogenesis. Multiple approaches have revealed that even before postmitotic neurons reach their final laminar position, process outgrowth initiates. Like retinogenesis as a whole, elaboration of axons and dendrites is dependent on both intrinsic and extrinsic cues. The initial orientation of ganglion cell axons, for example, relies on positional information provided by the retinal basal lamina. Retinal ganglion cells also depend on intrinsic factors such as Brn3b, a POU-domain transcription factor, which is essential for retinal ganglion cells to extend axons. Deletion of *brn3b* in mice results in retinal ganglion cells that prefer to generate dendrites at the expense of axons. Although analysis of process outgrowth and differentiation of individual cell types is ongoing, a general model is supported in which retinal neurons first extend processes in fields more extensive than their final arbors (Figure 6). In the synapse-refinement phase of process differentiation, cell–cell and cell–matrix communication appears to be very important. Similar to organization of neuronal cell bodies into lamina, final axon and dendrite arbors are stereotyped and highly specific. Retinal plexiform layers are organized so that specific subtypes of neurons synapse with each other in specific 'sublamina' or zones. For example, cells comprising the 'on-center' or 'off-center' pathways to light stimulus show sublaminar synaptic organization.

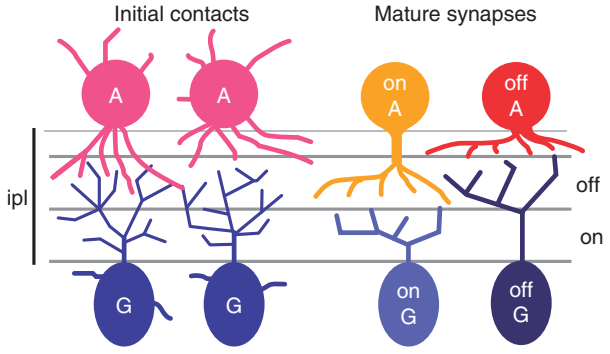

Figure 6 Sublamination of the inner plexiform layer. Early in development, as amacrine cell (A) and ganglion cell (G) dendrites first interact, process outgrowth is broad and relatively unstratified (left synaptic partners). Amacrine cells appear to guide 'on' and 'off' sublamination of the inner plexiform layer (ipl), which becomes strikingly polarized and stratified in the mature retina (right synaptic partners).

The on-center and off-center pathways can be distinguished by bipolar cells, which depolarize or hyperpolarize, respectively, when the photoreceptors from which they receive input are stimulated. On and off bipolar cells synapse with on and off ganglion cells in different substrata of the inner plexiform layer. Modulatory amacrine cells also specifically target their dendrites to either the on or off sublamina. In fact, amacrine cells play a central role in directing sublamination of the on and off circuits. Initially in development, the contacts between on and off cell processes are distributed throughout the inner plexiform layer. Over time, these synapses segregate so that off cells interact in a very discrete region apical to the synapses of on cells. This remodeling of ganglion cell synapses requires synaptic activity from amacrine cells. Interestingly, amacrine cell sublamination does not require ganglion cells or their dendrites at all and instead relies in part on cues from adjacent amacrine cells. Other sublamina, in addition to the on and off zones, exist in both the inner and the outer plexiform layers.

Cell Adhesion and Synapse Partnering

In addition to the activity-based mechanisms of synapse refinement, initial targeting of axons and dendrites to the appropriate lamina also depends on the expression of specific cell-associated guidance molecules. These molecules mediate both attractive and repulsive activities to potential synaptic partners. Based on gene expression and results from the spinal cord, various forms of the Protocadherin and Sidekick families of cell adhesion molecules have been proposed to facilitate initial axon/dendrite targeting within the retina. As with other processes of retinal development, our understanding of neuronal process differentiation and synaptogenesis is rudimentary, and much more research is required for a deeper understanding.

Conclusions

Normal retinal development occurs through a series of continuous refinements in which a simple pseudostratified epithelium is specified and patterned to form a highly organized sensory neuroepithelium. The underlying genetics and cellular interactions that drive retinogenesis are necessarily intricate and subjected to complicated feedback regulatory mechanisms. Current models of retinal development are still general, but our understanding of the gene networks and signaling factors that facilitate retinogenesis is increasing. Many important questions remain to be solved. For example, how is the cell cycle precisely controlled in progenitors, and how are cell type fate determinants integrated with regulation of

cell cycle exit? What are the cues that guide postmitotic cell migration in the retina, and what cues inform cells to stop migrating and begin morphogenesis? What are the factors that mediate the necessary cytoskeletal rearrangements required for both cell migration and morphogenesis? How are the spatial waves of neurogenesis and cell differentiation coordinated? What mechanisms work to allow neurons to find each other and form functional circuits? These and other questions are motivating hard work, driving technological advances, and pushing the creativity of many inspired retinal biologists.

See also: Fovea: Primate; Retina: An Overview; Retinal Pharmacology: Inner Retinal Layers; Retinal Color Mechanisms; Retinal Development: Cell Type Specification; Retinal Models; Retinomotor Movements; Vertebrate Eyes: Evolution; Visual Development.

Further Reading

Cayouette M, Poggi L, and Harris WA (2006) Lineage in the vertebrate retina. *Trends in Neurosciences* 29: 563–570.

Chalupa LM and Werner JS (eds.) (2004) *The Visual Neurosciences: Developmental Processes*. Cambridge, MA: MIT Press.

Chow RL and Lang RA (2001) Early eye development in vertebrates. *Annual Reviews: Cell & Developmental Biology* 17: 255–296.

Dowling JE (1987) *The Retina: An Approachable Part of the Brain*. Cambridge, MA: Belknap.

Dyer MA and Cepko CL (2001) Regulating proliferation during retinal development. *Nature Reviews Neuroscience* 2: 333–342.

Frade JM (2002) Interkinetic nuclear movement in the vertebrate neuroepithelium: Encounters with an old acquaintance. *Progress in Brain Research* 136: 67–71.

Gotz M and Huttner WB (2005) The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology* 6: 777–788.

Hatakeyama J and Kageyama R (2004) Retinal cell fate determination and bHLH factors. *Seminars in Cell and Developmental Biology* 15: 83–89.

Levine EM and Green ES (2004) Cell-intrinsic regulators of proliferation in vertebrate retinal progenitors. *Seminars in Cell and Developmental Biology* 15: 63–74.

Livesey FJ and Cepko CL (2001) Vertebrate neural cell-fate determination: Lessons from the retina. *Nature Reviews Neuroscience* 2: 109–118.

Marquardt T and Gruss P (2002) Generating neuronal diversity in the retina: One for nearly all. *Trends in Neurosciences* 25: 32–38.

Mumm JS, Godinho L, Morgan JL, Oakley DM, Schroeter EH, and Wong RO (2005) Laminar circuit formation in the vertebrate retina. *Progress in Brain Research* 147: 155–169.

Novelli E, Resta V, and Galli-Resta L (2005) Mechanisms controlling the formation of retinal mosaics. *Progress in Brain Research* 147: 141–153.

Pazour GJ, Baker SA, Deane JA, et al. (2002) The intraflagellar transport protein, IFT88, is essential for vertebrate photoreceptor assembly and maintenance. *Journal of Cell Biology* 157: 103–113.

Sernagor E, Eglen SJ, Harris WA, and Wong RO (eds.) (2006) *Retinal Development*. Cambridge: Cambridge University Press.

Tsujikawa M and Malicki J (2004) Genetics of photoreceptor development and function in zebrafish. *International Journal of Development Biology* 48: 925–934.

Retinal Development: Cell Type Specification

F L A F Gomes and M Cayouette, Institut de Recherches Cliniques de Montréal, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

The vertebrate retina is a thin layer of light-sensitive neural tissue lining the back of the eye. It comprises seven major cell classes harmoniously organized in three cellular layers (Figure 1). The outer nuclear layer (ONL) contains the cell bodies of rod and cone photoreceptors, whereas the inner nuclear layer (INL) contains the cell bodies of bipolar cells, amacrine cells, horizontal cells, and Müller glial cells. Finally, the ganglion cell layer (GCL) contains the cell bodies of retinal ganglion cells (RGCs), the projection neurons of the retina, as well as those of displaced amacrine cells. Finally, the inner and outer plexiform layers are synaptic layers located between the GCL and the INL, and between the INL and ONL respectively.

The seven major classes of retinal cells are generated in an evolutionary conserved, overlapping birth order. Generally, in vertebrates, the RGCs, cones, and horizontal cells are among the first cells to be generated, followed closely by amacrine cells, and finally by rods, Müller glia, and bipolar cells, which are among the last cell types to be generated (Figure 2).

Multipotency of Retinal Progenitor Cells

Pioneering lineage studies used retroviral vectors or injection of tracers to label retinal progenitor cells (RPCs), which made possible the identification of the entire progeny of individual RPCs, often referred to as their 'clone' (Figure 3). Some of these clones were composed of all the major retinal cell types, when the ancestor cell was labeled early during retinogenesis, indicating that at least some RPCs are multipotent at some point during retinal development. However, this does not preclude the possibility that a subset of progenitor cells may have restricted fate choices. Consistent with this idea is the finding that certain clones were composed of only one cell type, even when the RPC was labeled early in retinogenesis, suggesting that some RPCs might be constrained to generate one cell type. Whether such restricted progenitors have a characteristic 'molecular fingerprint' remains unknown, but recent gene expression profiling of RPCs should help address this issue. Interestingly, the homeodomain transcription factor Pax6 was recently shown to be required to confer the multipotent state of RPCs. Indeed, specific inactivation of Pax6 in RPCs, at the beginning of retinogenesis, prevented the generation of all retinal cell types except amacrine cells. Thus, it appears that Pax6 function in RPCs, at least early in development, is required to confer the multipotent character to RPCs.

To try to reconcile the finding that at least some RPCs are multipotent, together with the finding that different retinal cell types are generated in a strict chronological order, it has been proposed that RPCs undergo changes in their competence to generate particular cell types as development proceeds. Thus, early embryonic RPCs are competent to generate early-born retinal cell types, whereas more mature, postnatal RPCs have lost the competence to generate early-born cell types and gained the competence to generate late-born cell types. These changes in competence appear to be largely intrinsically driven. This was first suggested by the fact that the proliferative capacity and the timing of differentiation of early embryonic RPCs cultured *in vitro* do not change, even in the presence of a 50-fold excess of older postnatal retinal cells. In these mixed-age cultures, postnatal RPCs proliferated less and gave rise to rods much sooner than did embryonic RPCs, suggesting that early and late RPCs are intrinsically different. In addition, when early embryonic RPCs are cultured *in vitro* they can generate RGCs, whereas older postnatal RPCs, placed in the same *in vitro* conditions, generate mostly rods, even when mixed with an excess of early retinal cells. Moreover, isolated late RPCs, cultured at clonal density in serum-free and extract-free medium, can give rise to clones of the same general birth order, size, and composition as clones generated *in vivo*. The presence of a particular retinal cell type within such isolated clones did not appear to be required for the generation of any other retinal cell type, suggesting that instructive environmental signals from previously differentiated cells are not necessary for normal cell fate decisions, at least in late RPCs. Nonetheless, negative environmental signals most likely operate in such clonal density cultures, as well as *in vivo* (see below).

Coordinating Cell Fate and Proliferation

Of particular importance during retinal development is the coordination of proliferation and cell fate specification. Indeed, if RPCs exit the cell cycle too early during retinal development, the number of early-born cells increases at the expense of late-born cells. Thus, the decision to withdraw from the cell cycle must be coordinated with cell fate specification to guarantee that the correct proportion of each neuron and glia

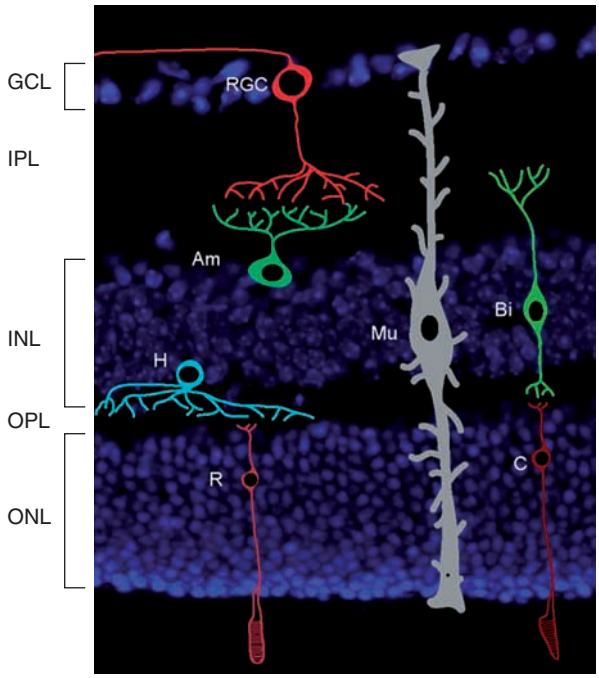

Figure 1 Organization of the vertebrate retina. The seven major retinal cell types are drawn on top of an actual histological section of rat retina (blue nuclei). Retinal ganglion cells (RGC) are found in the ganglion cell layer (GCL). Amacrine (Am), Müller (Mu), horizontal (H), and bipolar (Bi) cells are located in the inner nuclear layer (INL). Rod (R) and cone (C) photoreceptors are located in the outer nuclear layer (ONL). The inner plexiform layer (IPL) and outer plexiform layer (OPL) form the synaptic layers.

cell type is generated. But how does a cell know when to exit the cell cycle? Recent evidence suggests that there are mechanisms that connect cell cycle to fate determination. For example, subsets of RPCs express different types of cyclin-dependent kinase (CDK) inhibitors, suggesting that subpopulations of RPCs might use different components of the cell cycle machinery to control cell cycle exit. Such factors that direct cell cycle arrest may also specifically regulate cell fate determination pathways. Indeed, recent work has shown that the cyclin kinase inhibitor p27Xic1 promotes both cell cycle exit and Müller cell fate through distinct domains of the p27Xic1 protein.

Environmental Factors in Retinal Cell Fate Decisions

As retinal development proceeds, the microenvironment changes as a result of various retinal cell types being generated, raising the possibility that signals secreted by differentiated cell types act on progenitor cells to influence their fate. While the evidence for environmental signals directly instructing RPCs to produce particular retinal cell types *in vivo* is weak, the evidence for feedback inhibition signals that act to prevent the production of certain cell types is convincing. Perhaps the best-studied case of feedback inhibition concerns the RGCs. Young embryonic chick progenitors are inhibited in their ability to produce RGCs when cultured adjacent to old retinal cells, and depletion of the ganglion cells from the old retinal cell population abolishes this inhibition, indicating that signals from RGCs prevent the generation of too many RGCs. Since RGCs are generated first, one possibility is that they might also secrete signals that are required for progenitors to generate the other retinal cell types (instructive signals). This question was recently addressed using a transgenic mouse that expresses diphtheria toxin under the regulatory elements of Brn3b, a key transcription factor required for generation of RGCs. In this mouse, the toxin rapidly kills RGCs as they are born, thereby producing an RGC-depleted retina. Although the retinas of these animals showed reduced proliferation, there were no differences in the relative proportions of the various retinal cell types generated. Taken together, these results suggest that although environmental signals secreted from RGCs are important for normal RPC proliferation and to prevent RPCs from generating too many RGCs, they do not appear to be necessary to positively instruct the generation of other retinal cell types.

To determine how negative feedback signals could influence the production of RGCs in the retina, progenitors expressing green fluorescent protein (GFP) under the regulatory elements of ath5, a transcription factor required for the generation of RGCs, were imaged. In a wild-type retina, the ath5-positive cells systematically divided once to generate a RGC and another cell type (see below). However, when transplanted into a mutant retina, which lacks RGCs, these same progenitors often generated two RGCs, directly demonstrating that the lineage of RPCs can be modified by feedback inhibitory environmental signals.

What might these feedback signals be? Recent studies show that the conditional ablation of Sonic Hedgehog (Shh), which is expressed by RGCs in the mouse retina, causes a rapid depletion of the progenitor pool due to precocious cell cycle exit and neuronal differentiation. Moreover, progenitors are apparently biased to generate RGCs in the Shh knockout mouse, which contains increased number of RGCs. These results suggest that Shh provides feedback inhibition signal in the developing mouse retina. In addition, there is evidence that Shh signaling from RGCs is required for the normal laminar organization in the vertebrate retina, most likely by promoting normal Müller glial cell development. Negative feedback

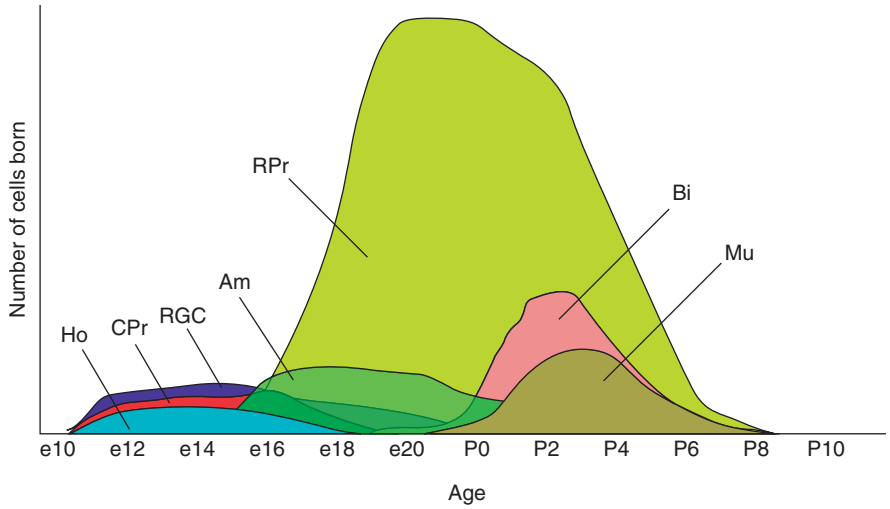

Figure 2 Chronological order of cell birth in the rat retina. The different retinal cell types are born in a strict but overlapping chronological order from around embryonic (e) day 10 to postnatal (P) day 10. Ho, horizontal cells; CPr, cone photoreceptors; RGC, retinal ganglion cells; Am, amacrine cells; RPr, rod photoreceptors; Bi, bipolar cells; Mu, Müller cells. This diagram is based on published results from Rapaport DH, Wong LL, Wood ED, Yasumura D, and Lavail MM (2004) Timing and topography of cell genesis in the rat retina. *Journal of Comparative Neurology* 474: 304–324.

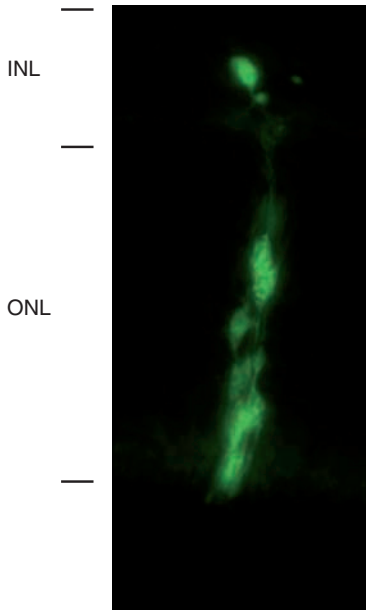

Figure 3 An example of a clone obtained by labeling a retinal progenitor cell at P0 with a GFP-expressing retroviral vector. The clone contains several photoreceptors, which are located in the outer nuclear layer (ONL), and a bipolar cell, which is located in the inner nuclear layer (INL).

signals may also affect the timing of differentiation. In the mouse retina, the growth differentiation factor 11 (GDF11) is a key regulator of RGC genesis. Evidence suggests that GDF11 acts by regulating the temporal window during which progenitors are competent to produce RGCs.

Another interesting group of environmental cues are the ligands of Notch. Constitutive stimulation of the Notch signaling pathway in rat RPCs biases the development of Müller glia at the expense of neurons, probably by interfering with the function of proneural bHLH transcription factors. At the same time, Notch clearly affects cell cycle exit. Experiments in *Xenopus* retina indicated that the Notch signaling pathway controls when a progenitor can differentiate. When cells are released from the inhibition mediated by the Notch pathway, basic helix–loop–helix (bHLH) transcription factors act as intrinsic factors that bias neuroblasts toward particular fates and histogenically appropriate cell cycle exit (see below).

Intrinsic Factors in Retinal Cell Fate Decisions

bHLH and Homeodomain Transcription Factors

As in other regions of the central nervous system, there are a number of transcription factors that intrinsically influence several facets of cell fate specification in the retina. The two major classes of transcription factors involved in cell fate identity comprise the bHLH and the homeodomain (HD). A bHLH motif consists of a short α helix connected by a loop to a second, longer α helix, whereas HD transcription factors constitute a special family of helix-turn-helix DNA-binding proteins, constructed from two α helices connected by a short extended chain of amino acids, which constitute the turn. These two classes of transcriptions factors bind to regulatory DNA sequence

either alone or via a complex with other transcriptions factors. In some cases binding triggers or enhances transcription, while in other cases it reduces or blocks transcription of developmental genes.

There are two functionally distinct groups of bHLH genes: repressors and activators. bHLH repressors (e.g., Hes1 and Hes5) are expressed in progenitors and inhibit neuronal differentiation. These repressors help maintain the undifferentiated state in the embryonic phase of retinogenesis, whereas they appear to promote gliogenesis in the postnatal retina. Misexpression of Hes1 in embryonic retina prevents differentiation and maintains the RPC fate. On the other hand, in Hes1 knockouts, progenitors show decreased proliferation, resulting in small-eye phenotype. In mice, the lack of Hes1 upregulates the expression of Mash1 (bHLH activator) resulting in increase of early-born types of neurons at the expenses of late-born cell types. Double mutation experiments have shown that another member of the Hes family, Hes5, acts cooperatively with Hes1 to regulate the maintenance of RPCs. Upon stimulation by Notch ligand from adjacent cells, Hes1 and Hes5 expression is increased, leading to the repression of bHLH activators genes, which in turns inhibits neuronal differentiation.

The role of bHLH and homeodomain transcription factors has been extensively studied in the retina. Here, we summarize the general results for the different retinal cell types, but it is beyond the scope of this article to describe in detail the function of these genes in retinal cell specification.

Müller glial cells In the postnatal retina, Hes1 and Hes5 are expressed in differentiating Müller glial cells and in RPCs. Misexpression of either Hes1 or Hes5 in postnatal RPCs increases Müller glia cell fate. Because both Hes1 and Hes5 are Notch effectors, these data indicate that activation of Notch signaling induces glial fate determination at the expense of other fates. Consistent with these results, constitutive activation of Notch signaling in postnatal RPCs increases the generation of Müller glial cells.

Interestingly, several lines of evidences suggest that progenitors and Müller glia cells are more closely related to each other than was previously thought. First, both cell types express the same transcriptions factors (Hes1, Hes5, and rax). Second, they are morphologically very similar, and lastly, Müller glia cells have been shown to have progenitor-like potency such as the capacity to self-renew and the ability to give rise to neurons and glia cells after retinal injury.

Bipolar cells It has been demonstrated that for bipolar cell fate specification two bHLH activators genes, Mash1 and Math3, and the HD factor Chx10 are required. Misexpression in RPCs of either Mash1 or Math3 alone mostly increases photoreceptor generation, whereas misexpression of Chx10 alone increases generation of Müller glia or undifferentiated cells, but not mature bipolar cells. In contrast, co-expression of Mash1 or Math3 with Chx10 in RPCs predominantly forces the generation of bipolar cells. Knockout mouse experiments have shown that Mash1/Math3 double mutants have no bipolar cells, whereas the number of Müller glial cells is increased. Loss of Chx10 function also prevents bipolar cell genesis but, in contrast to the Mash1/Math3 double mutant, is not accompanied by an increase of Müller glia cells, suggesting that Chx10 and Mash1/Math3 have distinct roles in specification of the bipolar cell fate.

Amacrine cells Two bHLH genes, NeuroD and Math3, are transiently expressed in differentiating amacrine cells. A mutation in either NeuroD or Math3 alone does not significantly affect amacrine cell development, but mutation in both NeuroD and Math3 severely decreases amacrine cell genesis, while increasing ganglion cell production.

Although NeuroD and Math3 are expressed during amacrine cell development, misexpression of each gene alone does not increase amacrine cell genesis, but significantly increases photoreceptor cell production. These results indicate that other genes may be required for amacrine cell development. Two homeodomain genes, Pax6 and Six3, are also expressed in amacrine cells. Although misexpression of each gene alone does not lead to increased amacrine cell production, concomitant expression of Pax6 or Six3 with NeuroD and Math3 significantly increases amacrine cell numbers. Thus, as in bipolar cell genesis, the co-expression of the bHLH and homeodomain genes influences the fate of amacrine cells.

Ganglion cells Math5, another member of the atonal bHLH activator gene, plays an important role in ganglion cell genesis. Math5 appears to directly activate the expression of the POU domain transcription factor Brn3b, which in turn is required for terminal differentiation of RGCs. In mice, it has been shown that a Math5 null mutation causes a dramatic decrease in RGC production and an increase of amacrine cells. These results, as stated earlier, are contrary to those of NeuroD/Math3 double mutants, which favored ganglion cell genesis at the expenses of amacrine cell production. It is plausible that NeuroD/Math3 and Math5 antagonistically control each other for amacrine versus ganglion cell fate specification. However, the mechanism for this regulation remains unknown.

Photoreceptor cells The homeodomain genes Crx and Otx2 have been shown to be required for photoreceptor cell development. In humans, mutations in Crx are associated with retinal diseases leading to photoreceptor degeneration and vision loss, such as cone–rod dystrophy-2, retinitis pigmentosa, and Leber's congenital amaurosis. Crx is expressed specifically in photoreceptor cells, where it is thought to regulate the expression of photoreceptor-specific genes such as rhodopsin and cone opsins. Overexpression of Crx in RPCs results in an increased generation of photoreceptor cells, and Crx knockout mouse photoreceptor cells do not elaborate outer segments and lack normal electrophysiological activity.

While Crx activity regulates the differentiation of both rods and cones, other factors specifically control the production of rod photoreceptors. The protein neural retina leucine zipper (Nrl) is a basic motif–leucine zipper transcription factor that is preferentially expressed in rod photoreceptors, where it is thought to act synergistically with Crx to regulate rhodopsin expression. Mutations in human NRL have been associated with autosomal dominant retinitis pigmentosa, which is characterized by rod photoreceptor degeneration. Interestingly, deletion of Nrl in mice results in the complete loss of rod function and an increase in cone function. Histological analysis of the Nrl knockout retina revealed that rod photoreceptors do not develop, whereas cone photoreceptors are overproduced. Thus, it appears that Nrl acts as a binary molecular switch during rod photoreceptor development by modulating rod-specific genes and inhibiting the cone differentiation pathway.

Another important regulator of rod photoreceptor development is the retinoblastoma (Rb) gene. During postnatal retinal development, Rb is expressed in proliferating RPCs and in differentiating rod photoreceptors. In the absence of Rb, RPCs continue to divide, and rods do not differentiate normally, as judged by abnormal inner and outer segments and disrupted pedicles. Based on these results, it has been proposed that Rb functions to regulate cell cycle exit of RPCs and rod differentiation.

Horizontal cell It was recently shown that coexpression of the homeodomain genes Pax6 or Six3 with the bHLH gene Math3 promotes primarily horizontal cell formation, in addition to amacrine cells. However, Math3 null mice develop horizontal cells normally, indicating that other genes may be compensating for Math3 function in horizontal cells.

The HD Prox1 is another essential gene for horizontal cell development. During retina development, Prox1 is expressed in differentiating horizontal cells, bipolar cells, and a subtype of amacrine cells. RPCs lacking Prox1 are less likely to stop dividing, and ectopic expression of Prox1 forces RPCs out of the cell cycle. In Prox1 knockout mice, horizontal cells are absent and misexpression of Prox1 in postnatal RPCs promotes horizontal cell production. Thus, it appears that Prox1 is both necessary and sufficient for RPC proliferation and horizontal cell fate determination in the vertebrate retina.

Asymmetric Cell Divisions in Retinal Cell Fate Decisions

Increasing evidence suggests that the asymmetric inheritance of cell fate determinants during an RPC division is an important mechanism regulating cell fate specification. In vertebrates, a number of studies have now provided evidence that the orientation of cell division is correlated to symmetric and asymmetric outcomes in various organs in different animals (Figure 4). Specifically, in the developing retina, the process of symmetric and asymmetric division has been most studied in terminal divisions that generate two postmitotic daughter cells. Upon terminal division, most RPCs divide with their mitotic spindle oriented parallel (horizontal) to the plane of the neuroepithelium, but a substantial minority divides with their spindle oriented perpendicular (vertical) to the plane of the neuroepithelium, along the apico-basal axis. A mammalian homolog of Numb (m-Numb), a natural antagonist of Notch signaling, is asymmetrically localized at the apical pole of the dividing RPCs and thus would be presumably asymmetrically inherited by the apical daughter cell of vertical divisions, whereas it would be symmetrically inherited by both daughter cells of horizontal divisions. Live imaging experiments of retrovirally labeled rat RPCs demonstrated that the two daughter cells of terminal divisions with a horizontal spindle tend to become the same cell type, whereas the two daughter cells of divisions with a vertical spindle tend to produce daughters that become different cell types. Moreover, overexpression of m-Numb in RPCs resulted in more daughter cell pairs of the same cell type and fewer daughter cell pairs of different cell types. Thus, the plane of cell division can apparently influence cell fate choice in the developing rat retina, presumably by controlling the asymmetric segregation of cell fate determinants such as m-Numb.

In the zebra fish retina, the orientation of cell division also appears to be correlated with cell fate. Although no apical–basal divisions are found in the zebra fish retina, some RPCs reorient their mitotic spindle within the plane of the neuroepithelium and divide either along the central to peripheral axis or along the circumferential axis. As development proceeds, the orientation of RPC divisions switches from

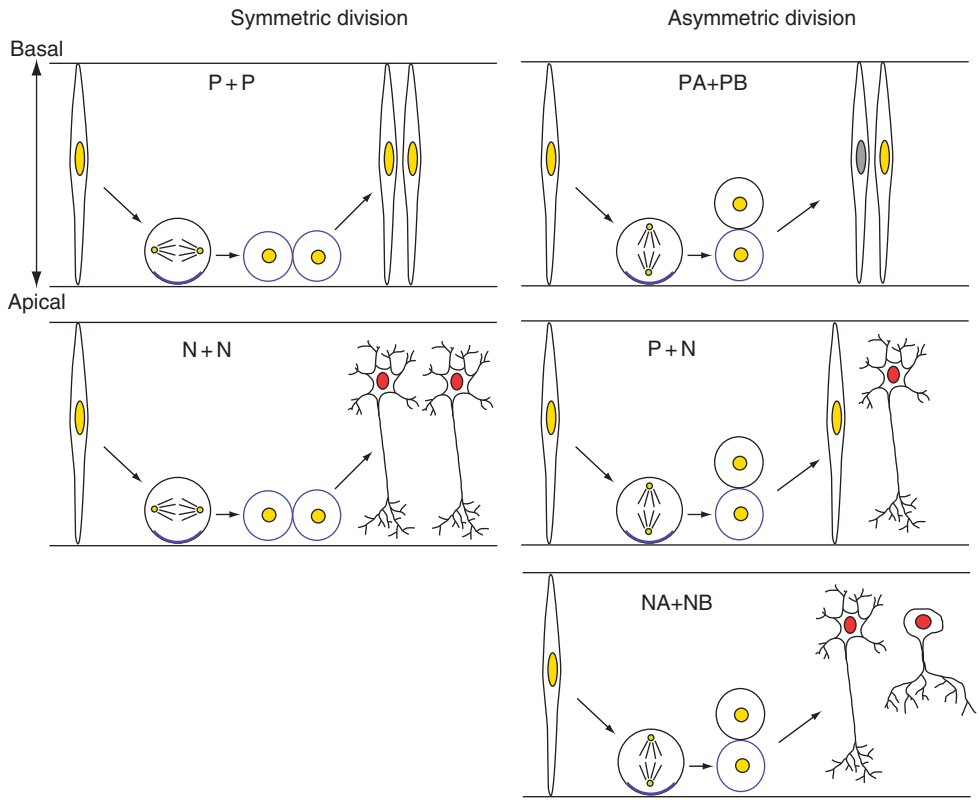

Figure 4 Different possible outcomes of symmetric and asymmetric cell divisions. RPCs undergo mitosis at the apical side of the neuroepithelium and segregate cell fate determinants (blue) asymmetrically at the apical pole of the cell cortex. In a symmetric division (left panels), the RPC divide parallel to the plane of the neuroepithelium, and cell fate determinants are inherited symmetrically by both daughter cells. The outcome of the division is two daughter cells of the same type, which can be two progenitors (P + P) or two neurons (N + N). In an asymmetric cell division (right panels), the RPC reorients its mitotic spindle to divide perpendicular to the asymmetric concentration of cell fate determinants at the apical pole of the cell cortex. This leads to asymmetric inheritance of cell fate determinants by the daughter cells. The outcome of the division is two daughter cells of different types, which can be two progenitors with different intrinsic potential (PA + PB), a progenitor and a neuron (P + N), or two neurons of different types (NA + NB).

being mostly central-peripheral to being mostly circumferential, suggesting that changing the orientation of cell division may contribute to particular cell fate decisions. Recent live imaging studies in the zebra fish retina support this possibility. In these experiments, it was shown that ath5:GFP RPCs in a wild-type retina divide along the circumferential axis to give rise to a RGC and another cell type, most likely a photoreceptor, whereas when these ath5:GFP RPCs are grafted in a RGC-depleted environment, they tend to divide along the central-peripheral axis to give rise to two RGCs, indicating that the orientation of cell division can predict the outcome in terms of the fate adopted by the daughter cells. It remains to be determined whether cell fate determining proteins are asymmetrically inherited by the daughter cells of circumferential divisions in the fish retina, but proteins involved in the establishment of planar cell polarity are certainly good candidates.

See also: Drosophila Apterous Neurons: from Stem Cell to Unique Neuron; Helix–Loop–Helix (bHLH) Proteins: Proneural; Helix–Loop–Helix (bHLH) Proteins: Hes Family; Retinal Amacrine Cells; Retinal Bipolar Cells; Retinal Ganglion Cells: Anatomy; Retinitis Pigmentosa; Müller Cells.

Further Reading

Cayouette M, Barres BA, and Raff M (2003) Importance of intrinsic mechanisms in cell fate decision in the developing rat retina. *Neuron* 40: 897–904.

Cayouette M, Poggi L, and Harris WA (2006) Lineage in the vertebrate retina. *Trends in Neurosciences* 29: 563–570.

Cayouette M and Raff M (2002) Asymmetric segregation of Numb: A mechanism for neural specification from Drosophila to mammals. *Nature Neuroscience* 5: 1265–1269.

Cepko CL, Austin CP, Yang X, Alexiades M, and Ezzeddine D (1996) Cell fate determination in the vertebrate retina. *Proceedings of the National Academy of Sciences of the United States of America* 93: 589–595.

Dyer MA and Cepko CL (2001) Regulating proliferation during retinal development. *Nature Reviews Neuroscience* 2: 333–342.

Hatakeyama J and Kageyama R (2004) Retinal cell fate determination and bHLH factors. *Seminars in Cell and Developmental Biology* 15: 83–89.

Kim J, Wu HH, Lander AD, Lyons KM, Matzuk MM, and Calof AL (2005) GDF11 controls the timing of progenitor cell competence in developing retina. *Science* 308: 1927–1930.

Livesey FJ and Cepko CL (2001) Vertebrate neural cell-fate determination: Lessons from the retina. *Nature Reviews Neuroscience* 2: 109–118.

Marquardt T (2003) Transcriptional control of neuronal diversification in the retina. *Progress in Retinal and Eye Research* 22: 567–577.

Marquardt T, Ashery-Padan R, Andrejewski N, Scardigli R, Guillemot F, and Gruss P (2001) Pax6 is required for the multipotent state of retinal progenitor cells. *Cell* 105: 43–55.

Mu X, Fu X, Sun H, et al. (2005) Ganglion cells are required for normal progenitor cell proliferation but not cell-fate determination or patterning in the developing mouse retina. *Current Biology* 15: 525–530.

Ohnuma S, Philpott A, Wang K, Holt CE, and Harris WA (1999) p27Xic1, a Cdk inhibitor, promotes the determination of glial cells in Xenopus retina. *Cell* 99: 499–510.

Perron M and Harris WA (2000) Determination of vertebrate retinal progenitor cell fate by the Notch pathway and basic helix-loop-helix transcription factors. *Cellular and Molecular Life Sciences* 57: 215–223.

Poggi L, Zolessi FR, and Harris WA (2005) Time-lapse analysis of retinal differentiation. *Current Opinion in Cell Biology* 17: 676–681.

Rapaport DH, Wong LL, Wood ED, Yasumura D, and Lavail MM (2004) Timing and topography of cell genesis in the rat retina. *Journal of Comparative Neurology* 474: 304–324.

Turner DL and Cepko CL (1987) A common progenitor for neurons and glia persists in rat retina late in development. *Nature* 328: 131–136.

Turner DL, Snyder EY, and Cepko CL (1990) Lineage-independent determination of cell type in the embryonic mouse retina. *Neuron* 4: 833–845.

Wang JC-C and Harris WA (2005) The role of combinational coding by homeodomain and bHLH transcription factors in retinal cell fate specification. *Developmental Biology* 285: 101–115.

Zaghloul NA, Yan Bo, and Moody SA (2005) Step-wise specification of retinal stem cells during normal embryogenesis. *Biology of the Cell* 97: 231–337.

Zhong W (2003) Diversifying neural cells through order of birth and asymmetry of division. *Neuron* 37: 11–14.

Retinal Ganglian Cells: Anatomy

D W Marshak, University of Texas Medical School, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Ganglion cells are the projection neurons of the vertebrate retina, conveying information from other retinal neurons to the rest of the brain. Their perikarya are the largest of any retinal neurons and are located along the inner margin of the retina, in the ganglion cell layer. Their axons run in a separate layer on the inner surface of the retina, collect at the optic disk, and then exit the eye as the optic nerve. Ganglion cells are the first type of neuron to differentiate, appearing earliest in the central retina and later in the periphery.

Ganglion cell dendrites extend into the inner plexiform layer (IPL), a neuropil located on the outer side of the ganglion cell layer. The ultrastructure of ganglion cell dendrites is typical of dendrites seen elsewhere in the central nervous system, and they usually do not contain synaptic vesicles or any other presynaptic specializations. Ganglion cells receive inputs from bipolar cells, which convey signals from photoreceptors to the IPL, and from amacrine cells, which branch in the IPL (**Figure 1**).

The presynaptic bipolar cell axon terminals contain specialized synaptic structures called ribbons that are surrounded by synaptic vesicles. There are two postsynaptic processes opposite each ribbon, and these synapses are called dyads. One postsynaptic process is usually a ganglion cell dendrite, and the second postsynaptic process is typically from an amacrine cell. At dyads, the amacrine cell process can make a reciprocal synapse back onto the bipolar cell, a feed-forward synapse onto the ganglion cell dendrite, or both. Amacrine cells also make chemical and electrical synapses onto ganglion cell dendrites at other sites, and they influence ganglion cells indirectly via synapses onto other bipolar cells and onto other amacrine cells.

The classical receptive field of a ganglion cell is defined as the area of the retina where stimulation with a small spot of light produces a change in ganglion cell firing rate. Ganglion cells are classified into three basic types based on their responses to light stimuli presented in the centers of their receptive fields. ON ganglion cells depolarize and fire action potentials to increments in light intensity, and OFF cells depolarize and fire action potentials to decrements in light intensity. Illumination presented in the region surrounding the receptive field center typically generates a response opposite in sign to that generated by central illumination. ON–OFF cells depolarize and fire action potentials to both increments and decrements in light intensity; they may or may not have antagonistic surrounds.

There are two common types of ganglion cells in typical vertebrate retinas, one type with relatively small dendritic arbors and sustained light responses and another type with larger dendritic arbors and more transient light responses. In addition, there are a number of other, less common types with more complex light responses. In all, vertebrate retinas have 10–20 types of ganglion cells, distinguishable by their morphology, synaptic connections, and light responses. The perikarya of each type are distributed across the retina in a regular array.

The light responses of ganglion cells are generated by direct, excitatory inputs from bipolar cells mediated by the neurotransmitter L-glutamate (**Figure 2(a)**). Axon terminals of ON-type cone bipolar cells contact the dendrites of the ON retinal ganglion cells in the inner half of the IPL. OFF-type cone bipolar cells terminate in the outer half of the IPL, where they contact the dendrites of OFF ganglion cells. ON–OFF ganglion cells receive input from both types of bipolar cells. For the most part, the chemical synapses made by amacrine cells are inhibitory, using either γ-aminobutyric acid (GABA) or glycine as neurotransmitters, but one type of amacrine cell, the starburst cell, releases acetylcholine and excites ganglion cells via nicotinic receptors. Some types of ganglion cells also make electrical synapses with amacrine cells or with other ganglion cells of the same type.

In mammals, there are rod bipolar cells with dendrites that receive input exclusively from rod photoreceptors; they depolarize in response to increments in light intensity (**Figure 2(b)**). The axon terminals of these rod bipolar cells make excitatory synapses onto two types of amacrine cells, A17 and AII cells. The A17 cells make reciprocal, inhibitory synapses back onto the rod bipolar cell axons. Signals from the AII cells are conveyed to the cone bipolar cells via two different pathways. The AII cells excite ON cone bipolar cells via gap junctions, and they also make glycinergic inhibitory synapses onto OFF cone bipolar cells. Signals also flow directly from rods to cones through gap junctions; however, this second rod pathway is less sensitive to light. In some mammals, there is a third rod pathway in which the dendrites of one type of OFF cone bipolar cell receive input directly from rods. Thus, signals originating in both rods and cones reach the ganglion cells through the same bipolar cells.

The focus of this article is on the anatomy of retinal ganglion cells in primates. The three most common

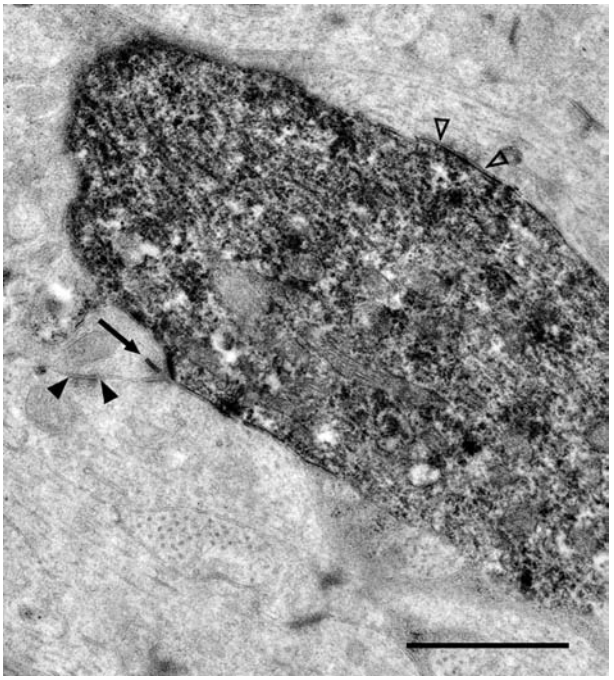

Figure 1 A large OFF parasol ganglion cell dendrite receives synapses from an amacrine cell process at a conventional synapse (open arrowheads) and a bipolar cell axon at a dyad synapse (arrow). The bipolar cell axon receives a reciprocal synapse (solid arrowheads) from an electron-lucent amacrine cell process. The parasol cell was from a macaque retina and was labeled by intracellular injection of neurobiotin. Scale bar = 1 μm.

types of ganglion cells in primates are called midget cells, parasol cells, and small bistratified cells, and together they comprise approximately 70% of all ganglion cells. The remaining ganglion cells are quite heterogeneous, but because they all have larger dendritic arbors, compared to the three common types, they are called wide-field cells. The wide-field ganglion cells of primates closely resemble the ganglion cells described in other mammalian retinas in their size and in morphology. Because primates have many more ganglion cells than other mammals have, the number of wide-field ganglion cells in primates is roughly the same as the total number of ganglion cells in typical mammalian retinas. In those instances where the light responses or synaptic connections have been studied, the wide-field ganglion cells of primates are similar to their homologs in other mammals in these respects, as well.

Midget Cells

Midget cells are the most common type of ganglion cells in primate retinas. In the central retina, approximately 70% of ganglion cells are midget cells, but in the peripheral retina the proportion decreases to 50%. Midget cells are also known as P cells because they project to the parvocellular layers of the lateral geniculate nucleus (LGN). In the central 6–7° of visual angle, they generally receive excitatory synapses from a single midget bipolar cell, which, in turn, receives input from a single long-wavelength (L) or middle-wavelength (M) sensitive cone. The dendritic fields of midget ganglion cells become larger with increasing eccentricity, and these peripheral midget ganglion cells receive input from two or more midget bipolar cells. Another difference in the peripheral retina is that rod signals reach midget ganglion cells via contacts between AII amacrine cells and midget bipolar cells, whereas the midget ganglion cells in the central retina receive only cone input.

Midget ganglion cells also receive synaptic input from amacrine cells, this comprises approximately half of their input in the central retina and a larger proportion in the peripheral retina. GABAergic amacrine cells receiving input from midget bipolar cells at dyads with midget ganglion cells often make reciprocal synapses onto the bipolar cell, feed-forward synapses onto the ganglion cell, or both. Amacrine cells containing glycine contact both midget bipolar cells and midget ganglion cells. Another amacrine cell type synapses only onto the midget ganglion cells. Interactions between amacrine cells in the neural circuit providing input to midget ganglion cells are seen very frequently in studies of serial sections using electron microscopy.

Midget ganglion cells with dendrites in the inner half of the IPL have ON center responses; that is, they are excited by increments in the intensity of white light in their receptive field centers and inhibited by that stimulus in the receptive field surround. Midget ganglion cells branching in the outer half of the IPL have the opposite, or OFF center, responses, and they have smaller dendritic fields than do the ON subtype at a given eccentricity. Because their dendritic fields are small and there is no overlap between dendrites of the same type, midget ganglion cells are well suited for high-acuity vision.

Most central and midperipheral midget ganglion cells have responses of opposite polarity to stimulation of the L and M cones, and thus they also mediate red–green color vision. When either L or M cones are stimulated selectively, midget ganglion cells respond as though their receptive field centers are driven by a single subtype of cone, and their surrounds are driven by the opposite type. Cone-selective centers and surrounds such as these are optimal to detect differences in color, although this is achieved at the expense of sensitivity to differences in brightness. Despite 40 years of research, it is unclear how the neural circuit providing input to midget ganglion cells generates these cone-selective responses.

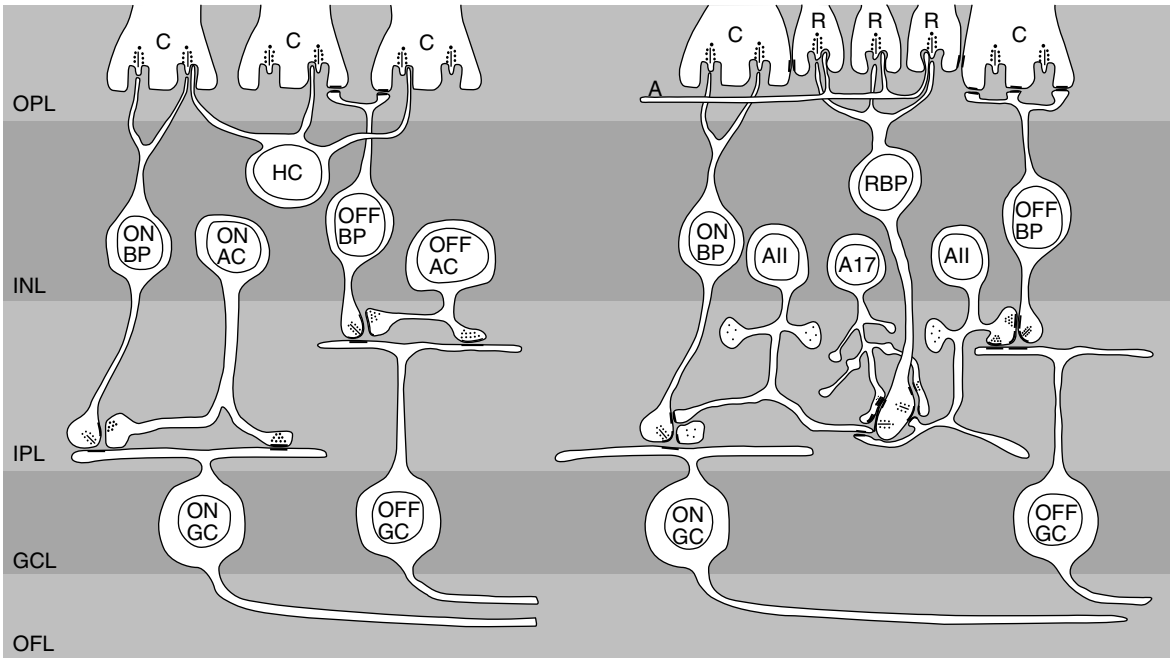

Figure 2 In mammals, the rods and cones differ in their connections with horizontal cells and contact different sets of bipolar cells. (a) Cone pathways: In the outer plexiform layer (OPL), cones (C) provide input to ON and OFF cone bipolar cells (BP) and horizontal cells (HC) in the inner nuclear layer (INL). In the inner plexiform layer (IPL), bipolar cells make excitatory, ribbon synapses onto ON and OFF ganglion cells (GC) (GCL, ganglion cell layer; OFL, optic fiber layer). The bipolar cells also synapse on amacrine cells (AC), which make inhibitory synapses back onto bipolar cells and forward onto amacrine cells and ganglion cells. (b) Rod pathways: Rods (R) provide input to rod bipolar cells (RBP), which synapse onto bistratified AII amacrine cells and A17 amacrine cells. A17 cells make inhibitory, feedback synapses onto rod bipolar cells. AII amacrine cells relay the rod signal to OFF ganglion cells via inhibitory synapses onto OFF cone bipolar cells and to ON ganglion cells via gap junctions with ON cone bipolar cells. Rods also make gap junctions with cones.

In the central 6–7°, the input from a single midget bipolar cell might account for the cone specificity of the excitatory input. However, there are small gap junctions between neighboring L and M cones, and the signal from one cone subtype has a contribution from the other. Cone-selective receptive field centers might also arise by chance in the midperipheral retina. Because L and M cones are randomly arranged and tend to form clusters of a single type, the midget bipolar cells presynaptic to a peripheral midget ganglion cell might receive input from the same subtype of cone. This would not account for the large proportion of color opponent peripheral midget ganglion cells, however.

There is no obvious anatomical explanation for the cone-selective inhibitory input in any part of the retina. Horizontal cells, neurons that branch in the outer plexiform layer (OPL), are unselective in their contacts with L and M cones, and their light responses indicate that the input from both cones is of the same polarity. Amacrine cells that contact midget ganglion cells receive input from all types of bipolar cells within their dendritic fields. Thus, the anatomical studies suggest that the light responses of midget ganglion cells simply reflect the distribution of L and M cones in their receptive fields. The apparent contradiction between the anatomical and physiological findings can be reconciled if interactions between local circuit neurons generate cone-specific responses.

All other mammalian retinas have ganglion cells that mediate high-acuity vision. These ganglion cells have brisk, sustained light responses and linear summation of light responses within their receptive field centers. None are as highly specialized as the midget ganglion cells in the central primate retina, but they resemble peripheral midget ganglion cells in many respects. These ganglion cells are not electrically coupled, and they have a high spatial density. Their dendritic arbors are small in diameter, relatively dense, and broadly stratified within the IPL.

Parasol and Other Cells with Brisk Transient Responses

Parasol cells are another major type of ganglion cell in primates, comprising approximately 5% of all ganglion cells in the central retina and 15% in the peripheral retina. Parasol cells are also known as M cells because they project to the magnocellular layers of the LGN. Parasol cells contribute to many aspects of

vision, particularly to the perception of motion. They are more sensitive to luminance contrast than midget cells are, and they also respond more transiently to light stimuli. They have a greater absolute sensitivity to light than midget cells have, because they collect from a larger area of the retina and receive more input from the rod pathway. Unlike midget cells, parasol cells are also responsive to certain types of stimuli outside the classical receptive field, and thus they convey information about global as well as local stimulus properties.

At any given eccentricity, parasol cells have much larger dendritic fields than do midget cells; their dendrites are also more narrowly stratified and ramify closer to the center of the IPL. There are two types, ON cells with branches in the inner half of the IPL, and OFF cells with branches in the outer half. As with midget cells, each type of parasol cell forms a regular array across the retina, and the dendritic field size increases with eccentricity. An important difference between midget and parasol cells is that the dendrites of adjacent parasol cells of the same type overlap extensively. The spacing between perikarya of neighboring parasol cells is roughly half of their dendritic field diameters, and thus, the dendritic coverage is three. As a result, a small spot stimulus would excite three neighboring parasol cells of the same subtype.

Many other differences between parasol cells and midget cells are attributable to the neurons that provide their input. Parasol cells receive excitatory synapses from diffuse bipolar cells, which, in turn, receive input from several cones. Two different types of diffuse bipolar cells provide input to OFF parasol cells. Diffuse bipolar cells are also presynaptic to ON parasol cells, but they have not been identified. The classical surround responses of parasol cells originate from the summed surround responses of the diffuse bipolar cells, which are generated in the OPL by horizontal cells.

Parasol cells also receive input from multiple types of amacrine cells, and the proportion of input from amacrine cells increases with eccentricity. Unlike amacrine cells in the pathway providing input to midget ganglion cells, the amacrine cells presynaptic to parasol cells rarely make reciprocal synapses with the bipolar cells providing input to the parasol cells, and they do not contact other amacrine cells in the vicinity of the parasol cell. Nevertheless, it is clear that interactions between amacrine cells are important in the pathway providing input to parasol cells. All the amacrine cells known to be presynaptic to parasol cells receive inputs from other amacrine cells. In addition, the GABA antagonist picrotoxin attenuates the light responses of parasol cells, a finding suggesting that GABAergic amacrine cells inhibit other amacrine cells in the pathway.

These interactions between amacrine cells take place on a spatial scale too large to detect in serial electron microscopic reconstruction studies of parasol cells, however. For example, some amacrine cells receive input from other amacrine cells at their dendrites and convey the signals to parasol cells via their axons, which extend for several millimeters. These inputs from axon-bearing amacrine cells presumably enable parasol cells to respond to stimuli outside their classical receptive fields. Another reason why interactions between amacrine cells in the neural circuit providing input to parasol cells are difficult to unravel is that some presynaptic amacrine cells receive synapses in one part of their dendritic trees and make synapses in another part. The starburst amacrine cell is one such amacrine cell. Its dendrites costratify with those ON and OFF parasol cells and synapse onto both types.

Dendrites of parasol ganglion cells make gap junctions with large, axon-bearing amacrine cells, and they are coupled indirectly to other parasol cells of the same subtype (**Figure 3**). As a result, neighboring parasol cells of the same subtype tend to fire synchronously, and when the signals from neighboring parasol cells converge in the brain, they are particularly effective in depolarizing their target cells. Midget ganglion cells, on the other hand, do not make gap junctions or fire synchronously, and they act independently when their signals converge in the brain. The gap junctions with amacrine cells might also enhance the sensitivity of parasol cells to luminance contrast. Because the amacrine cells coupled to parasol cells also make inhibitory synapses onto other parasol cells, neighboring parasol cells indirectly inhibit one another. At a border, parasol cells on the unstimulated side are inhibited, and this relieves parasol cells on the stimulated side from tonic inhibition. The gap junctions do not enlarge the receptive fields of parasol cells, however; the diameters of parasol cell receptive field centers are the same as their dendritic field diameters.

Primates also have wide-field ganglion cells very similar to ganglion cells known as alpha cells in other mammals. That is, they are large, monostratified ganglion cells with brisk, transient responses to light. The alpha-like cells comprise 2–3% of all ganglion cells, and they project to both the LGN and the superior colliculus. Like parasol cells, the alpha-like cells are electrically coupled to amacrine cells. In addition, their dendrites have a characteristic, radiate branching pattern and are narrowly stratified on either side of the center of the IPL. In some other

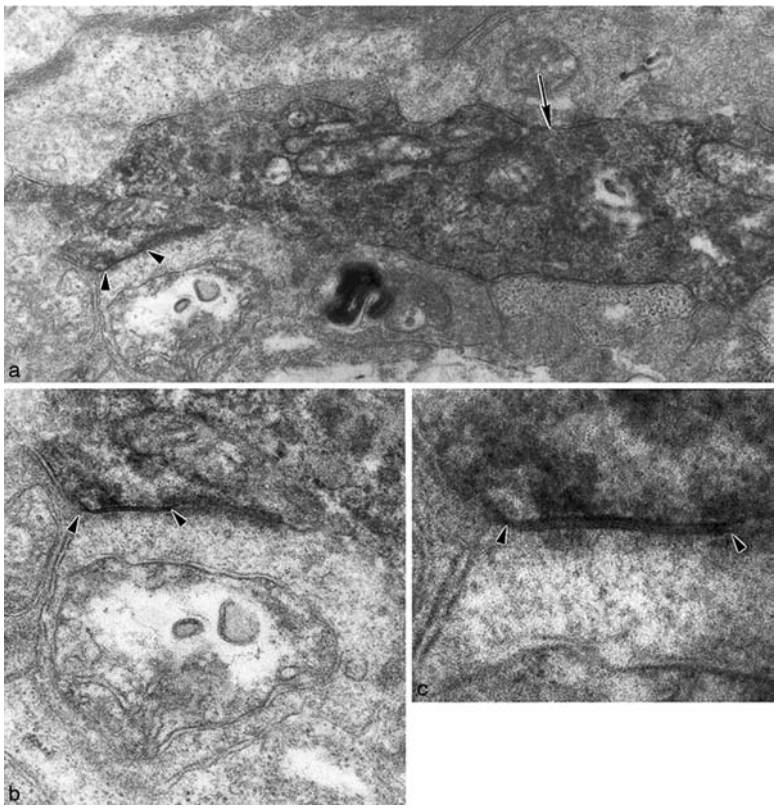

Figure 3 A large ON parasol cell dendrite receives a conventional synapse from an amacrine cell (arrow) and makes a gap junction with another amacrine cell (arrowheads). (a) Both synapses are found on a large dendrite, ~1.2 μm in diameter (30 000×). (b) The amacrine cell making the gap junction is cut in cross section (50 000×). (c) The junction is shown at higher magnification (150 000×). The parasol cell was injected with neurobiotin in a baboon retina.

mammals, there are also two types of alpha-like ganglion cells, although the differences between the two types are not as pronounced as those between alpha-like cells and parasol cells of primates.

Cells Responding Selectively to Short Wavelengths

Small bistratified cells make an important contribution to blue–yellow color vision. They depolarize and fire action potentials when short-wavelength (S) cones are stimulated and are inhibited when either L or M cones are stimulated; with 500 nm light stimulation, they do not respond at all. Small bistratified cells also depolarize and fire action potentials in response to decrements in long- and middle-wavelength light; this is called an OFF response. They have two sets of dendrites, one near the inner border of the IPL and another near the outer border. Their dendritic arbors increase in size with eccentricity and are roughly the same size as those of neighboring parasol cells. The small bistratified cells are arranged in a regular array, and their dendritic coverage is approximately two. Small bistratified cells are the third most common type of primate ganglion cells, comprising approximately 5% of all ganglion cells. They project to koniocellular layers 3 and 4 in the LGN.

The bipolar cells that provide the major excitatory input to small bistratified ganglion cells are called blue cone bipolar cells because they receive input exclusively from S cones. The horizontal cells in this pathway, the H2 type, contact all three cone types and particularly S cones. There is a second, much smaller excitatory input from diffuse bipolar cells in the outer half of the IPL that also hyperpolarize when L and M cones are stimulated. However, most of the OFF response originates in the H2 horizontal cells and is conveyed to the small bistratified ganglion cells by the blue cone bipolar cells.

Small bistratifed ganglion cells are electrically coupled to narrow-field amacrine cells and, indirectly, to neighboring small bistratified cells. In the central retina, small bistratified cells receive 30% of their input from amacrine cells, and the proportion increases to 80% in the peripheral retina. In this respect, the small bistratifed ganglion cells are typical of primate

ganglion cells; all three major types receive more input from amacrine cells in the peripheral retina. This may explain the greater responsiveness of primate ganglion cells to moving and other rapidly changing stimuli in the peripheral retina.

Several types of wide-field cells also contribute to blue–yellow color vision. Large bistratified cells comprise approximately 3% of all ganglion cells and project to the LGN. They have two sets of dendrites near the borders of the IPL, and they are excited by increments in short-wavelength light and decrements in long and middle wavelengths. However, they have considerably larger dendritic arbors than do the small bistratified type. There are two types of sparse, monostratified ganglion cells, one with dendrites near the outer border of the IPL and the other with dendrites near the inner border. Together, the two subtypes of sparse monostratified cells comprise approximately 3% of all ganglion cells, and they project to the LGN. The cells branching near the inner border have OFF responses to short wavelengths and ON responses to long and middle wavelengths.

The giant cells that contain the photopigment melanopsin also have OFF responses to short wavelengths and ON responses to long and middle wavelengths with bright stimuli. The light responses of giant ganglion cells are atypical in two other respects, however. Despite the fact that many of the giant cells have dendrites branching in the outer IPL, all of them have ON responses to dim lights that stimulate only the rods. A slow, sustained component of their light responses mediated by melanopsin enables the giant cells to transmit information about the absolute intensity of light stimuli to their targets in the brain. They comprise approximately 0.2% of all ganglion cells, and they project to the olivary nucleus of the pretectum, the suprachiasmatic nucleus of the hypothalamus, and the LGN.

Most mammals have only two cone types, S and M, and therefore have dichromatic color vision. All species studied to date have a distinctive type of ON bipolar cell that contacts S cones exclusively and ramifies in the innermost stratum of the IPL. Some mammals also have ganglion cells that costratify with the axon terminals of the blue cone bipolar and ganglion cells with color opponent responses. However, the color opponent ganglion cells have only been identified morphologically in primates. All mammals also have ganglion cells with extremely large dendritic arbors that express the photopigment melanopsin and are intrinsically sensitive to light. It is uncertain whether they also have color opponent light responses like the giant cells of primates.

Other Wide-Field Cells

A number of other types of wide-field ganglion cells have been described using anatomical techniques in primates (Figure 4). Only the most abundant types are described here. Two types of wide-field ganglion cells resemble direction-selective ganglion cells of other mammals, cells that respond vigorously to stimuli moving in one direction but not to stimuli moving in the opposite direction. They project to the superior colliculus and, together, comprise 6% of all ganglion cells. Thorny ganglion cells have fine, tortuous branches with many small branches and spines, and they resemble local edge detectors of other mammals. They project to the LGN and comprise 4% of all ganglion cells. To determine how these wide-field cells contribute to vision in primates, it will be essential to characterize each type by intracellular recording and dye injection.

Conclusions

An important goal for future research is to determine how amacrine cells contribute to the light responses of each type of ganglion cell. The classical receptive field centers and surrounds of ganglion cells can be explained in large part by synaptic interactions in the OPL. However, most of the synapses in the IPL are made by amacrine cells, and all types of ganglion cells receive a major input from amacrine cells. Ganglion cell types differ in the patterns of excitation and inhibition that they receive during the response to light, and many of these differences are attributable to amacrine cells. However, it is uncertain which types of amacrine cells provide input to most types of ganglion cells. The amacrine cells are quite diverse in their morphology and synaptic connections, but aside from the starburst cells and the two cells in the rod pathway, very little is known about their roles in visual information processing.

In primates, the different morphological types of retinal ganglion cells develop in a proscribed-sequence, and this may provide insight into their evolutionary history. Wide-field cells, which closely resemble ganglion cells in other mammals, are the first types to develop. In the central nervous system, as a rule, phylogenetically old structures develop early. Taken together, these findings suggest that wide-field cells were the retinal ganglion cells in the common ancestors of mammals, nocturnal animals having only S and M cones. It will be a challenge for the future to determine the precise sequence of developmental events that generates the morphological

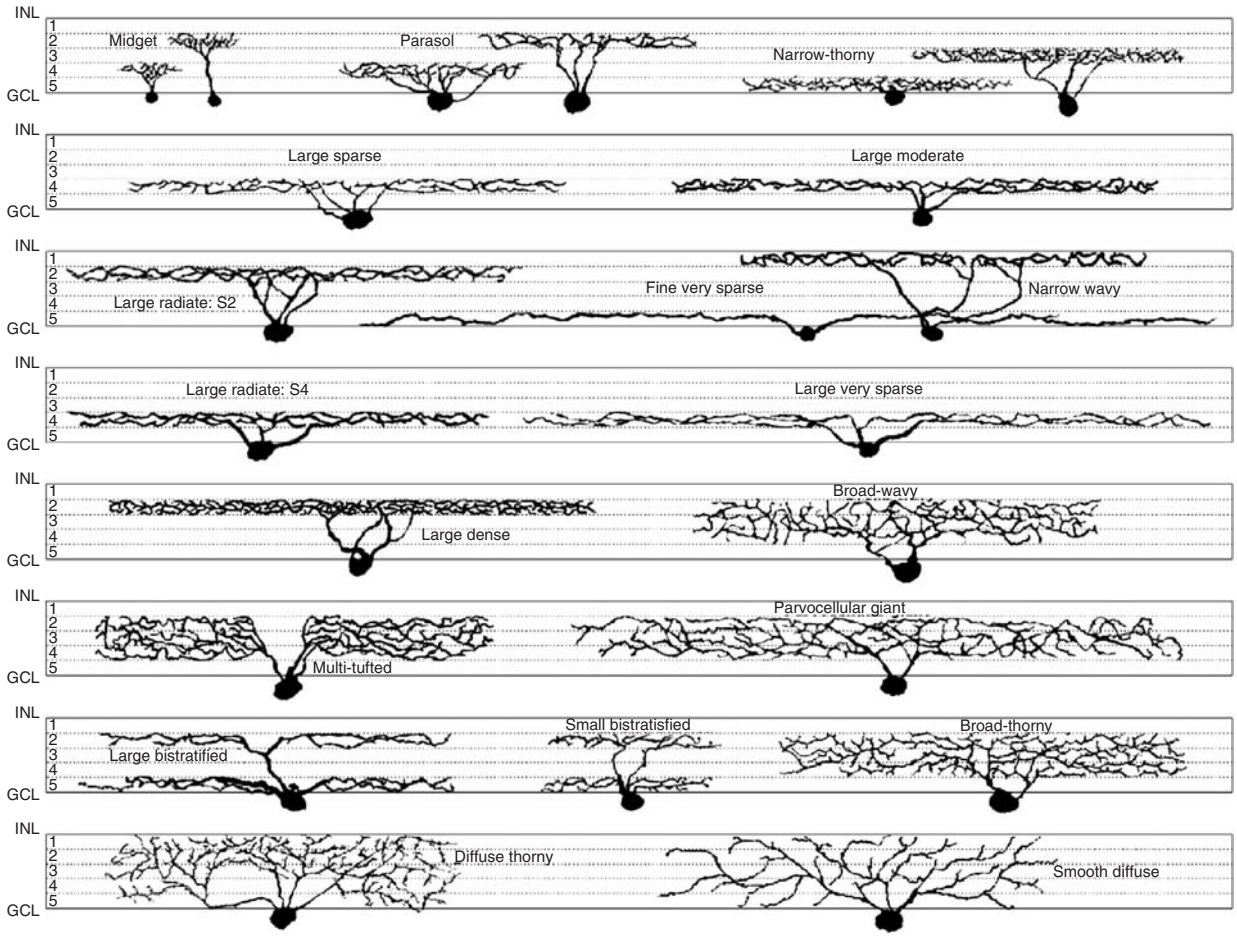

Figure 4 An artist's conception of the ganglion cell types described to date in macaque retinas, as they would appear in vertical sections. The five strata of the inner plexiform layer (IPL) are indicated by shading, S1 being closest to the inner nuclear layer (INL) and S5 being closest to the ganglion cell layer (GCL).

differences between the various types of wide-field cells, but these processes are likely to be similar in all mammalian retinas. Midget ganglion cells are the next to develop, followed by the parasol cells. These later developing cells presumably evolved more recently and, among mammals, may be unique to primates.

See also: Color Vision; Retina: An Overview; Retinal Amacrine Cells; Retinal Bipolar Cells; Retinal Ganglion Cells: Receptive Fields; Retinal Pharmacology: Inner Retinal Layers; Ribbon Synapses.

Further Reading

Bordt AS, Hoshi H, Yamada ES, et al. (2006) Synaptic input to OFF parasol ganglion cells in macaque retina. *Journal of Comparative Neurology* 498: 46–57.

Callaway EM (2005) Structure and function of parallel pathways in the primate early visual system. *Journal of Physiology* 566: 13–19.

Dacey DM (2004) Origins of perception: Retinal ganglion cell diversity and the creation of parallel pathways. In: Gazzaniga M (ed.) *The Cognitive Neurosciences,* 3rd edn., pp. 281–301. Cambridge, MA: MIT Press.

Dacey DM, Liao HW, Peterson BB, et al. (2005) Melanopsin-expressing ganglion cells in primate retina signal colour and irradiance and project to the LGN. *Nature* 433: 749–754.

Famiglietti EV (2004) Class I and class II ganglion cells of rabbit retina: A structural basis for X and Y (brisk) cells. *Journal of Comparative Neurology* 478: 323–346.

Ghosh KK and Grünert U (1999) Synaptic input to small bistratified (blue-ON) ganglion cells in the retina of a New World monkey, the marmoset *Callithrix jaccus*. *Journal of Comparative Neurology* 413: 417–428.

Jusuf PR, Martin PR, and Grünert U (2006) Random wiring in the midget pathway of primate retina. *Journal of Neuroscience* 26: 3908–3917.

Kolb H and Marshak D (2003) The midget pathways of the primate retina. *Documenta Ophthalmologica* 106: 67–81.

McMahon MJ, Packer OS, and Dacey DM (2004) The classical receptive field surround of primate parasol ganglion cells is mediated primarily by a non-GABAergic pathway. *Journal of Neuroscience* 24: 3736–3745.

Rapaport DH, Fletcher JT, LaVail MM, et al. (1992) Genesis of neurons in the retinal ganglion cell layer of the monkey. *Journal of Comparative Neurology* 322: 7–588.

Rockhill RL, Daly FJ, MacNeil MA, et al. (2002) The diversity of ganglion cells in a mammalian retina. *Journal of Neuroscience* 22: 3831–3843.

Roska B, Molnar A, and Werblin FS (2006) Parallel processing in retinal ganglion cells: How integration of space-time patterns of excitation and inhibition form the spiking output. *Journal of Neurophysiology* 95: 3810–3822.

Schein S, Sterling P, Ngo IT, et al. (2004) Evidence that each S cone in macaque fovea drives one narrow-field and several wide-field blue–yellow ganglion cells. *Journal of Neuroscience* 24: 8366–8378.

Schlens J, Field GD, Gauthier JL, et al. (2006) The structure of multi-neuron firing patterns in primate retina. *Journal of Neuroscience* 26: 8254–8266.

Solomon SG, Lee BB, and Sun H (2006) Suppressive surrounds and contrast gain in magnocellular-pathway retinal ganglion cells of macaque. *Journal of Neuroscience* 26: 8715–8726.

Solomon SG, Lee BB, White AJR, et al. (2005) Chromatic organization of ganglion cell receptive fields in the peripheral retina. *Journal of Neuroscience* 25: 4527–4539.

Wässle H (2004) Parallel processing in the mammalian retina. *Nature Reviews Neuroscience* 5: 1–11.

Yamada ES, Bordt AS, and Marshak DW (2005) Wide field ganglion cells in macaque retinas. *Visual Neuroscience* 22: 383–393.

Retinal Ganglion Cells: Receptive Fields

J B Troy, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

One can trace the term receptive field to Sherrington, who used it to name the region of a sensory surface from which a stimulus can evoke a neural response. Sherrington applied the term to a group of neurons, but today the term receptive field more commonly means the sensory region where a stimulus evokes a response from a single cell. In the visual system, one may consider the retina to be the sensory surface where receptive fields are located, but descriptions of receptive fields are more often presented in terms of the optical projection of the retina out into the world, the construct we call the visual field. Hence, for the retinal ganglion cell, the receptive field could be either a small region of retina or a small region of the visual field. Since descriptions of receptive field in the latter case incorporate an additional component due to optical blurring, which increases with decreasing light level as the pupil dilates, the two usages are not completely equivalent.

Throughout the years, the receptive field concept has taken on a meaning greater than just a spatial region over which sensory stimuli can influence neural activity. Usage of the term receptive field has come to include also a description of the configuration of the optimal stimulus for the field. For example, ever since Kuffler provided the classical description of receptive fields for mammalian retinal ganglion cells, we have come to speak of center-surround receptive fields, which are divided into two subtypes: those with off-centers and on-surrounds and those with on-centers and off-surrounds. In subsequent years, many varieties of retinal ganglion cell have been discovered in vertebrate retina, and the classification into physiological types is predicated on differences in receptive field properties. There are ganglion cells with different varieties of center-surround receptive field, some of which can be differentiated by size. In a particular retinal location, there will be some retinal ganglion cells with on-centers that are markedly larger than those of neighboring retinal ganglion cells with on-center/off-surround receptive fields. A difference in the time course of response to visual stimulation has also been used to differentiate ganglion cell types. Although one might argue that the term receptive field should be reserved for descriptions in the spatial domain, this is not followed, and perhaps with good reason. Individual neurons can represent just one spatial location since all the information they convey is carried by a single spike train. A receptive field description incorporating how the spatiotemporal distribution of light and dark and, in some cases, wavelength modulates the distribution of spikes in the spike train is richer from a functional perspective. Thus, the term receptive field encompasses both a spatial area and how the sensory stimulus should be configured over that area spatially, temporally, and spectrally to most effectively drive the cell's discharge.

By taking on this added dimension, the concept of receptive field has become one of the cornerstones of sensory physiology. We have come to view change in receptive field structure as one moves up the chain of neurons in a sensory pathway as an indicator of the processing that occurs at each stage. The spatiotemporal and spectral structure of the receptive field tells us the operation performed on the visual input by the optical and neural elements that precede spike initiation and thus what controls the neuron's discharge.

Classification of Retinal Ganglion Cell Types

No consensus exists on how to relate ganglion cell types across species and, even within a species, different terminologies coexist. Hence, the classification of mammalian ganglion cells is confusing, particularly to the nonexpert. One might hope that genetic fingerprinting will lead to a clearer classification, but the extent to which ganglion cells can be separated on the basis of gene expression remains unclear. At best, it seems that differentiation into types will require analysis of genetic expression in multidimensions. With few exceptions (perhaps the melanopsin ganglion cell), it does not seem that single gene markers could be used to identify a type of ganglion cell.

Physiological cell typing is something of an art, in that the characterization of receptive field properties depends on the skill and insight of the experimenter. An approach that some have used is to employ a spot of light (or dark) as the visual stimulus. By adjusting the position where one flashes the spot within the receptive field, one might assume that one can faithfully map the receptive field's response to the spot and thus reveal the receptive field structure. By adjusting the spectral properties of the spot, one could reveal chromatic receptive field properties and the temporal nature of the response should inform the observer

about the temporal properties of the receptive field region probed. There are many reasons why this approach to receptive field mapping is problematic. First, the sensitivity of the retina to light is not constant but, rather, varies as a function of its space- and time-averaged illuminance. The structure of the receptive field measured from darkness is not the same as the structure measured against a background of higher luminance. Second, some visual adaptation in the retina occurs locally. Thus, the light in a stimulus spot might adapt the retinal region being stimulated and thus affect the response of the cell to that stimulus. This is particularly likely when stimulating regions of a receptive field with low sensitivity. In such circumstances, one has the choice between applying a weak stimulus and obtaining a weak response or applying a stronger stimulus (a more intense light) to obtain a bigger response. The latter will produce more adaptive effect and so one might want to avoid this choice. The former runs afoul of the third difficulty faced in receptive field mapping, namely discharge noise. Repeated presentations of the same stimulus do not result in the same response from a ganglion cell; responses are noisy. Hence, in order to determine the response in the presence of this noise, one must collect and average responses to repeated presentations of the stimulus to improve signal-to-noise. The smaller the response, the more averaging is required and the longer the data collection period. Given that physiological experiments are of limited duration, the experimenter must weigh expediency against rigor. A fourth problem with the use of spots as stimuli is that the time courses of light and dark adaptation are different, so spots of significant but equal light and dark contrast will not produce exactly inverse responses. Another pitfall of reliance on single spot stimuli is that there are ganglion cell receptive fields responsive to the direction of movement of the visual stimulus. This property cannot be revealed with a flashing spot stimulus.

Cathode ray tube technology permitted the mapping of visual receptive fields with more complex stimuli. The pioneering work was done by Enroth-Cugell and Robson. They used sinusoidal luminance gratings and other two-dimensional patterns (e.g., edges) in which luminance is modulated along one of the two dimensions to map ganglion cell receptive fields. The cat was the animal subject and, in this species, they discovered two varieties of Kufflerian center-surround receptive field, one of which they ascribed to X cells and the other to Y cells. Both X and Y cells occur with on-center/off-surround and off-center/on-surround subtypes. What differentiated X cells from Y cells in Enroth-Cugell and Robson's view was the fact that the responses of X cells could be largely explained in terms of the linear spatial summation of light within their receptive fields, whereas the responses of Y cells could not. Subsequent work has shown that Y cells have significantly larger receptive field centers than X cells of the same retinal locale and that their morphologies differ. When viewed in whole mounts, Y cells have α cell morphology and X cells have β cell morphology. This difference in morphology is also related to the difference in receptive field center size, in that the spread of the dendritic tree is highly correlated with center size.

Receptive Field Modeling

The application of sinusoidal luminance gratings to the study of visual neural responses led naturally to application of the Fourier transform to the analysis and modeling of receptive fields. Essentially coincident, but using different approaches to studying ganglion cell receptive fields, Rodieck and Enroth-Cugell and Robson developed the first quantitative model of ganglion cell receptive fields. The model is the difference of Gaussians receptive field, which provides a very good quantitative description of center-surround properties. In the model, both center and surround mechanisms are assumed to have Gaussian weighting for their sensitivity to light (for on-center mechanisms) or dark (for off-center mechanisms), with the spread of the surround Gaussian being significantly broader. The central points of the two Gaussians are coincident, and it is the difference between the signals generated by the two mechanisms that accounts for the cell's response. Rodieck's model was based on measurements of ganglion cell responses for stimuli in the spatial domain, whereas Enroth-Cugell and Robson's derivation of the model relied on ganglion cell responses made for stimuli in the spatial frequency domain. The Fourier transform can be used to equate models developed in the two domains and the fact that the results were consistent demonstrated the validity of the model. This model has dominated modeling of center-surround receptive fields since, although some modifications of the original model have proven necessary. It was assumed originally that the center and surround mechanisms had circular symmetry, but later work indicated a small degree of ellipticity to the center mechanism at least. It has also been found that the midpoints of center and surround may be slightly offset and, most important, that the time course of response for center and surround mechanisms is not the same. The effect of the last observation is that the spatial structure of the receptive field is dependent on the temporal properties of the stimulus, with less surround antagonism as

temporal frequency increases. A consequence of this is that the spatial frequency tuning of the ganglion cell shifts from band-pass at low temporal frequencies to low-pass at high temporal frequencies. A similar shift from band-pass to low-pass tuning with increasing temporal frequency has long been known to occur for contrast sensitivity versus spatial frequency measurements made on human subjects psychophysically.

The difference of Gaussians receptive field model cannot explain nonlinear receptive field properties. It is a linear model. The first nonlinear receptive field model was probably that proposed by Barlow and Levick to account for the directional selective receptive field of a variety of rabbit retinal ganglion cell. Similar directionally selective receptive fields have been found in many mammalian retinas subsequently and are believed to be present in all mammals. Hochstein and Shapley developed a nonlinear receptive field model for the cat Y cell. It was built on the classical difference of Gaussians receptive field model to account for the center-surround receptive field of these cells, but with an added mechanism to account for the Y cell's nonlinear responses. The new mechanism was an extensive field of subunits, each of which sums light over a small retinal region and then feeds a rectified version of this signal as its input to the Y cell. The Y cell discharge is therefore governed by a signal that is a sum of the signal coming from the nonlinear subunits and the signal coming from the linear difference of Gaussians field. This model was found to provide a good description of Y cell responses to sinusoidal luminance gratings, whether drifted across the receptive field or held stationary, with contrast modulated in time.

Varieties of Retinal Ganglion Cells

As noted previously, the grouping of retinal ganglion cells into different types is still something of an art form and disagreements over classification have become heated at times. The current dogma holds that each type of retinal ganglion cell will have

1. unique receptive field properties;
2. a distinct dendritic tree morphology, when viewed in a retinal whole mount; and
3. a unique pattern of dendritic stratification in the retina's inner plexiform layer.

The sites of axonal projection in the brain may be another factor that helps in ganglion cell differentiation. It also seems likely that, in time, differences in gene expression will be found to correlate with receptive field phenotypes and, as the art of ganglion cell typing develops, it will come to be another basis on which to group ganglion cells. If this transpires, classification of ganglion cell types will come to be considered more scientific and less artful. Correspondences between ganglion cell types in different species will then become a less contentious issue, and one may presume that the confusing proliferation of names that currently exists will cease and there will be retrenchment with more general agreement on nomenclature.

The physiological classification of ganglion cell-types has seen its greatest advance in the cat and, hence, the naming of ganglion cell types in this species has been largely responsible for the names given to ganglion cell types in other mammalian species. Overall, the naming of cell types has fallen into two camps. There are those who prefer names that carry no implication about the physiological properties of the receptive field or other ganglion cell properties and those who wish to impart with the name some sense of what distinguishes one ganglion cell type from others. From the former camp came the terminology X and Y cells. From the latter came the terms sustained and transient cells, which were intended to reflect their temporal response properties. Both classification schemes applied to cells with Kufflerian center-surround receptive fields. The latter scheme was soon found to have problems. It was discovered that there were different types of ganglion cells with temporal response properties that could be separated into the limited binary division sustained and transient. Hence, there was a further division made along an also binary dimension into brisk and sluggish types, chosen to indicate a somewhat qualitative sense of the nature of the cell's discharge. Thus, the names brisk-sustained, brisk-transient, sluggish-sustained, and sluggish-transient cells were created. It is now generally accepted that the brisk-sustained cell and the brisk-transient cell correspond to the X and Y cell, respectively.

The use of names for ganglion cell types that carry some sense of presumed function has been more widely followed than naming on the basis of letters of the alphabet, and the brisk-sustained, etc. terminology has been carried over to other mammals more frequently than the X and Y terminology. In fact, the use of the alphabetic naming scheme for cat ganglion cells has fallen out of favor. This was probably precipitated by the unfortunate introduction of the term W cell. The axons of Y cells are larger than those of X cells, with the result that electrical stimulation of the optic nerve results in a compound action potential showing an early peak reflecting the summed response of Y cell action potentials and a later peak reflecting the summed response of X cell action potentials. A third, slower and more diffuse, peak is also observed, reflecting the summed response of

action potentials discharged by ganglion cells with smaller axons still than those of X cells. Thus emerged the term W cell, which corresponds to all ganglion cell types but the X and Y cells. That W was the letter preceding X in our alphabet made the letter sequence consistent with increasing spike conduction velocity. Unfortunately, the term W cell has limited utility since we now know that it includes many varieties of ganglion cells which have very different functional properties sharing just one, namely a slower rate of action potential conduction. With the introduction of the W cell, the advantage of an alphabetic naming scheme – 26 letter choices and no implication of function – was dealt a major blow. A letter had been used to signify a group of cell types and, at the same time, it carried meaning. To my knowledge, the use of alphabetic names has been employed only once since for a physiological type of cat ganglion cell. The name Q cell was used by Enroth-Cugell et al. for what is almost certainly the cat sluggish-sustained cell.

The classification of ganglion cell types of primate retina, although influenced by work in the cat, has occurred somewhat independently. Ganglion cells with center-surround receptive fields were divided first into tonic and phasic varieties, corresponding roughly to the sustained and transient divisions of cat ganglion cells. A later division was made between ganglion cells whose axons project to the magnocellular layers of the primate lateral geniculate nucleus and those whose axons project to its parvocellular layers. These ganglion cells were given alphabetic names – M cell for the magnocellular projecting and P cell for the parvocellular projecting ganglion cell. Linking a functional aspect with the letter had the same downside as the use of W cell for cat ganglion cells. It is now believed that multiple ganglion cell types provide input to the magnocellular and parvocellular layers of the primate lateral geniculate nucleus, so linking nomenclature with a nonunique cellular property in this way leads to confusion. However, the terms M and P cell have obtained a status that transcends their original meaning. The M cell is a physiological cell type whose receptive field is largely achromatic and has a center-surround organization. These cells have both off-center and on-center subtypes and are linked to a distinct dendritic morphology, the parasol cell. Their responses are also phasic, and thus these cells likely correspond to the classical phasic ganglion cells of primate. The P cell is a physiological cell type whose center-surround receptive field shows spectral opponency. Old World primates have three varieties of cone: the long-wavelength sensitive L cone, the middle-wavelength sensitive M cone, and the short-wavelength sensitive S cone. There are P cells with L cone-dominated on-centers and M cone-dominated off-surrounds, P cells with L cone-dominated off-centers and M cone-dominated on-surrounds, P cells with M cone-dominated on-centers and L cone-dominated off-surrounds, and P cells with M cone-dominated off-centers and L cone-dominated on-surrounds. All four physiological cell types exhibit the midget cell morphology. The responses of these cells are tonic; thus, the P cell and the classical tonic ganglion cell of primate are probably one and the same. The term K cell has been used by some to refer to the ganglion cells whose axons terminate in the koniocellular layers of the primate lateral geniculate nucleus, but again this was not a good choice because it includes a number of ganglion cell varieties. For the most part, the naming of non-M and P ganglion cell types of primate and other ganglion cells of other mammalian species has been made on the basis of presumed receptive field function.

At approximately the same time that X and Y cells were discovered in cat, it became clear that other ganglion cell types lacking center-surround receptive fields existed. Barlow and Levick had described the directional selective receptive field of a ganglion cell of rabbit retina. We now know of two types of directionally selective ganglion cell of mammalian retina. One is the on–off directionally selective cell and the other is the on directionally selective cell. The latter cell type responds with an on response to light shone over the center of its receptive field. The former type gives an on–off response to this form of stimulation. Both exhibit a pronounced preference for the direction of stimulus movement across the receptive field, and the on–off directionally selective cell has a preference for spatially restricted stimuli. Movement of large field stimuli across its receptive field, even in the preferred direction, evokes little or no response from the cell. Rodieck reported another ganglion cell receptive field whose discharge was interrupted when either light or dark contrast was introduced to its receptive field. This cell type was colorfully named the suppressed-by-contrast cell. It has been suggested that there may be more than one type of such receptive field, and others have called these uniformity detectors and edge-inhibitory off-center cells. Another ganglion cell type that was discovered early was the on–off cell. As its name implies, this cell responds to both the onset and the offset of a stimulus within the receptive field. These cells have also been called excited-by-contrast cells, phasic on–off cells, local-edge detectors, and impressed-by-contrast cells. There is also a cell with a receptive field showing spectral opponency between S cone and the L/M cones and another ganglion cell which is inherently light sensitive. This has come to be known as the

melanopsin cell. Ganglion cells with many of these receptive field properties have been found in the retinas of primate, rabbit, ground squirrel, guinea pig, mouse, and rat. It is suspected that all are present in all mammalian retinas.

Conclusion

The receptive field has become a most useful concept in sensory physiology, indicating the operation performed on the sensory stimulus by the processing stages preceding spike generation. It informs us what sensory information is transmitted and what is discarded. The presence of multiple receptive field types for retinal ganglion cells indicates that the visual image is broken into packets for transmission to the brain. Since the axonal destinations of different classes of ganglion cell differ, one can presume that the function of this packaging in part at least reflects its usefulness for the brain centers where it is destined. In the case of some simple functions (e.g., control of pupil diameter as a function of mean light level), we have a good sense of how and why the message is packaged in the way that it is. For most ganglion cell types, the nature of packaging is less obvious. Segregation of sensory messages into packets with different cell types carrying separate components seems to be a common design feature. Hence, what we learn about parallel processing in the visual system will help us to understand parallel processing in sensory systems generally. Thus, characterization of receptive field properties for ganglion cells should lead to a better overall understanding of sensory processing. We have entered a period in which it is reasonable to suppose that all ganglion cell receptive field types of mammals will be known and their properties characterized. When this has been accomplished, we will be in a better position to understand how visual information is encoded in the retina and, thus, what information reaches the next stages in the visual pathways. Obviously, we need this understanding to determine what processing takes place at each of these higher visual centers. It has long been appreciated that the characterization of ganglion cell receptive fields is foundational for an understanding of visual information processing. As our knowledge of ganglion cell properties has grown, so has an appreciation of the complex nature of the visual message sent from the retina to the brain. We have moved beyond the simplistic view of ganglion cell receptive fields with few varieties to one in which we accept that there are many varieties. In the not too distant future, we can expect that all varieties will be known along with their receptive field properties.

See also: Neural Coding of Spatial Representations; Retinal Ganglion Cells: Anatomy; Retinal Pharmacology: Inner Retinal Layers; Retinal Development: An Overview; Sensory Ganglia.

Further Reading

Barlow HB and Levick WR (1965) The mechanism of directionally selective units in the rabbit's retina. *Journal of Physiology* 178: 477–504.

Boycott BB and Wässle H (1974) The morphological types of ganglion cells of the domestic cat's retina. *Journal of Physiology* 240: 397–419.

Cleland BG, Dubin MW, and Levick WR (1971) Sustained and transient neurones in the cat's retina and lateral geniculate nucleus. *Journal of Physiology* 217: 473–496.

Cleland BG and Levick WR (1974) Brisk and sluggish concentrically organized ganglion cells in the cat's retina. *Journal of Physiology* 240: 421–456.

Dacey DM and Lee BB (1994) The blue-ON opponent pathway in primate retina originates from a distinct bistratified ganglion cell type. *Nature* 367: 731–735.

Enroth-Cugell Ch and Robson JG (1966) The contrast sensitivity of retinal ganglion cells of cat. *Journal of Physiology* 187: 517–552.

Enroth-Cugell Ch, Robson JG, Schweitzer-Tong DE, and Watson AB (1983) Spatiotemporal interactions in cat retinal ganglion cells showing linear spatial summation. *Journal of Physiology* 341: 279–307.

Gouras P (1968) Identification of cone mechanisms in monkey ganglion cells. *Journal of Physiology* 199: 533–547.

Hattar S, Liao HW, Takao M, Berson BM, and Yau KW (2002) Melanopsin-containing retinal ganglion cells: Architecture, projections and intrinsic photosensitivity. *Science* 295: 1065–1070.

Hochstein S and Shapley RM (1976) Linear and nonlinear spatial subunits in Y cat retinal ganglion cells. *Journal of Physiology* 262: 265–284.

Hoffmann KP, Stone J, and Sherman SM (1972) Relay of receptive-field properties in dorsal lateral geniculate nucleus of the cat. *Journal of Neurophysiology* 35: 518–531.

Kuffler SW (1953) Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology* 16: 37–68.

Marc RE and Jones BW (2002) Molecular phenotyping of retinal ganglion cells. *Journal of Neuroscience* 22: 413–427.

Robson JG (1966) Spatial and temporal contrast sensitivity functions of the visual system. *Journal of the Optical Society of America* 56: 1141–1142.

Rodieck RW (1965) Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research* 5: 583–601.

Rodieck RW (1967) Receptive fields in cat retina: A new type. *Science* 157: 90–92.

Shapley R and Perry VH (1986) Cat and monkey retinal ganglion cells and their visual functional roles. *Trends in Neurosciences* 9: 229–235.

Sherrington C (1906) *The Integrative Action of the Nervous System*. New Haven, CT: Yale University Press.

Stone J and Fabian M (1966) Specialized receptive fields of the cat's retina. *Science* 152: 1277–1279.

Troy JB and Shou T (2002) The receptive fields of cat retinal ganglion cells in physiological and pathological states: Where we are after half a century of research. *Progress in Retinal and Eye Research* 21: 263–302.

Retinal Glia

E A Newman, University of Minnesota, Minneapolis, MN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Glial cells outnumber neurons in the central nervous system (CNS) by ~10 to 1. Traditionally, glia were believed to provide only passive structural and metabolic support for neurons. Recent work has demonstrated, however, that glial cells in the retina as well as in the brain interact actively with neurons and have many essential functions.

There are three principal types of glial cells in the mammalian retina: Müller cells, astrocytes, and microglial cells. Müller cells are the most prominent retinal glial cell. They are a specialized form of radial glia which span nearly the entire depth of the retina. Astrocytes, the second type of retinal macroglial cell, are present only in species having a retinal circulation and, in these species, are restricted largely to the nerve fiber layer at the inner boundary of the retina. Microglia, the third type of retinal glial cell, are present in the nerve fiber layer and the inner and outer plexiform layers of the retina. Oligodendrocytes, the glial cells that form the insulating sheath of myelinated axons in the CNS, are completely absent from the retina, except in those species, including rabbit and guinea pig, possessing myelinated axons in the nerve fiber layer.

Retinal Glial Cell Development and Morphology

Müller Cells

Müller cells differentiate from progenitor cells at the outer margin of the retinal neuroepithelium during the second phase of cell differentiation, following the birth of ganglion cells, horizontal cells, and cones. The developing Müller cells migrate toward the vitreous humor and extend processes towards the inner and outer retinal margins. Their radial processes serve to guide migrating neurons during subsequent retinal development.

In the mature retina, Müller cell somata lie in the middle region of the inner nuclear layer (Figure 1). Single or multiple radial processes extend inwards toward the vitreous and end in enlarged structures termed endfeet. A second cell process extends outward to the outer limiting membrane, where apical microvilli project into the subretinal space. Secondary Müller cell processes project from the main trunk of the cell, surrounding neuronal somata and processes in all retinal layers. Müller cell endfeet terminate at the vitreal border of the retina as well as onto blood vessels, both at the vitreal surface and within the retina.

Müller cells contain numerous glycogen granules which represent the main retinal glucose store. They also contain numerous mitochondria and intermediate filaments, composed both of vimentin and glial fibrillary acidic protein (GFAP). GFAP distribution is normally sparse in Müller cells but is dramatically upregulated following retinal injury.

Astrocytes

Astrocytes migrate into the developing retina from the optic nerve and advance across the retinal surface as the retina matures. The distribution of astrocytes over the surface of the retina is determined by a 'contact-spacing' interaction between cells.

In the mature retina, astrocytes are restricted largely to the nerve fiber layer (Figure 2). They are closely associated with blood vessels and are only present in species having a retinal circulation. Their processes are confined largely to the plane parallel to the retinal surface, although processes sometimes follow blood vessels into deeper retinal layers. Astrocyte processes are terminated by endfeet which contact superficial retinal blood vessels. Astrocytes contain glycogen granules and dense bundles of intermediate filaments composed of GFAP.

Microglia

Microglia are the resident macrophages of the retina. They are derived from blood monocytes and enter the retina along with blood vessels during development. Microglia are normally present in the retina in their dormant state and are found principally in the nerve fiber layer and the inner and outer plexiform layers. They have small somata and short, irregular processes. The cells are normally flat in appearance and have slender hair-like extensions protruding from their processes.

Microglial cells proliferate following retinal injury and differentiate into macrophage-like cells that remove degenerating retinal neurons by phagocytosis. Microglia also phagocytose neurons that die during the course of normal retinal development.

Glial Induction and Guidance of Retinal Blood Vessels

Glial cells induce the formation of blood vessels and guide their growth into the retina. During development, astrocytes migrate into the retina from the

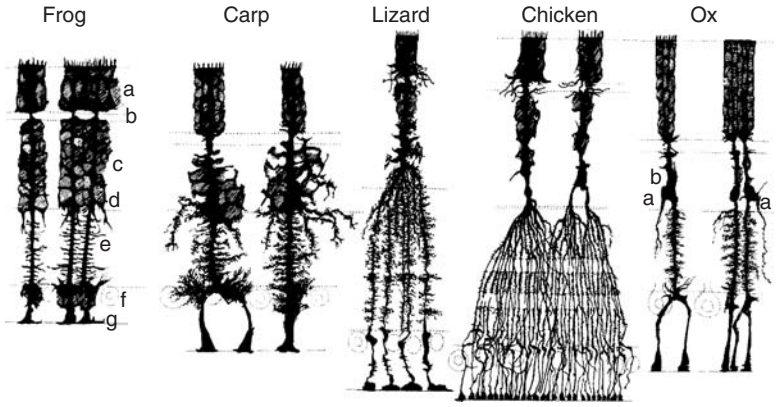

Figure 1 Drawings of Golgi-stained Müller cells from the retinas of several species. Retinal layers, indicated to the right of the frog Müller cells, are (a) outer nuclear layer; (b) outer plexiform layer; (c) inner nuclear layer; (e) inner plexiform layer; (f) ganglion cell layer; (g) nerve fiber layer. The basal endfeet of the Müller cells are seen as enlargements of the processes in the nerve fiber layer. From Ramon y Cajal S (1972) *The Structure of the Retina*. Springfield, IL: Thomas, C.C.

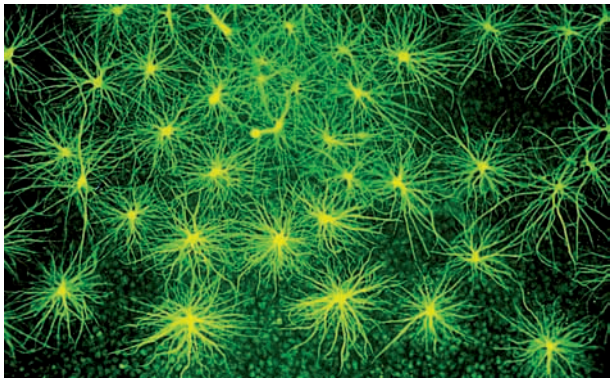

Figure 2 Fluorescence micrograph of astrocytes within the nerve fiber layer of the cat retina. The tissue was labeled with fluorescently tagged antibodies against glial fibrillary acidic protein (GFAP) which forms the intermediate filaments expressed in astrocytes. From Karschin A, Wassle H, and Schnitzer J (1986) Shape and distribution of astrocyte in the cat retina. *Investigative Ophthalmology and Visual Science* 27: 828–831.

optic nerve, first appearing at the optic disk and then advancing outwards across the retinal surface. Retinal blood vessels, which also originate from the optic nerve, grow along newly formed astrocyte processes, which function as templates for angiogenesis. Vessel growth is stimulated by the secretion of vascular endothelial growth factor (VEGF), which is released by glial cells in response to hypoxic conditions generated by neuronal activity. Astrocytes serve as guides for developing vessels as they grow across the inner surface of the retina while Müller cells serve as templates for vessel growth into the retina and across the inner nuclear layer.

Retinal Glial Cell Membrane Properties

Retinal glial cells possess many of the same voltage-gated ion channels and neurotransmitter receptors that are expressed in neurons. These glial channels and receptors play a key role in mediating glia–neuron interactions in the retina. Most research on retinal glial cell properties has been conducted on Müller cells and the following discussion will focus on the properties of these specialized glial cells.

Ion Channels

Müller cells have a high membrane permeability to K^+, which results in a negative resting membrane potential of –80 to –90 mV, near the K^+ equilibrium potential of the cell. Expression of Kir4.1, an inwardly rectifying K^+ channel, is largely responsible for this K^+ membrane permeability. In Kir4.1 knockout animals, which lack Kir4.1 channels, Müller cell input resistance increases from a wild-type level of 25 to 310 MΩ, demonstrating that Kir4.1 channels account for over 90% of the cell membrane conductance. Kir4.1 channels are not distributed uniformly over the cell surface but rather are localized to cell endfeet, which terminate at the vitreal surface of the retina and on retinal blood vessels. The dystrophin–dystroglycan complex plays a critical role in the clustering of K^+ channels to cell endfeet. As discussed below, clustering of K^+ channels contributes to Müller cell clearance of excess K^+ from the retina.

In addition to Kir4.1, Müller cells express a number of other voltage-gated ion channels, including Ca^{2+}-activated K^+ channels, fast inactivating K^+ channels, Ca^{2+} channels (which share some of the properties of L-type Ca^{2+} channels), and Na^+ channels. Müller cells also express aquaporin-4 (AQP4). These water-permeable channels are co-localized with Kir4.1 at Müller cell endfeet. Together, Kir4.1 and AQP4 make up a macromolecular complex that can be visualized in freeze-fracture electron

micrographs of endfoot membranes as orthogonal arrays of particles (OAPs).

Müller cells and astrocytes are extensively coupled to each other by gap junctions, composed of connexin43 and 45 in Müller cells and connexin30 and 43 in astrocytes. Gap junctional coupling imparts both electrical and chemical tracer coupling to these networks of glial cells. Coupling is stronger between astrocytes than between astrocytes and Müller cells. Coupling is very weak (or may not exist) between Müller cells, except in amphibians. Interestingly, coupling between astrocytes and Müller cells, as assessed by tracer spread, is asymmetric, with tracer passing from astrocytes to Müller cells more freely than from Müller cells to astrocytes. In rabbit, astrocytes are coupled to oligodendrocytes as well as to Müller cells.

Neurotransmitter Receptors

Retinal Müller cells and astrocytes express a number of neurotransmitter and neuromodulator receptors. The most prominent among these are P2Y purinergic receptors. Amphibian Müller cells express $P2Y_1$, $P2Y_2$, $P2Y_6$, and $P2Y_{11}$ purinergic receptors. Müller cells of different species also express a number of glutamate receptors, including AMPA (GluR4), NMDA, and metabotropic types. $GABA_A$, acetylcholine, dopamine, noradrenaline, adenosine (A_{2B}), thrombin, and lysophosphatidic acid receptors as well as several neuroactive peptides receptors are expressed. Activation of many of these receptors evokes glial Ca^{2+} increases (see below). Interestingly, in the mammalian retina, glutamate is largely ineffective in evoking Ca^{2+} increases in Müller cells or astrocytes. In contrast, glutamate evokes large Ca^{2+} increases in astrocytes in brain slices and in culture. The insensitivity of retinal glial cells to glutamate may be a specialization to conditions in the retina, where glutamate is released continuously from neurons.

Receptors to a number of growth factors, including basic fibroblast growth factor (bFGF), epidermal growth factor (EGF), platelet-derived growth factor (PDGF), and nerve growth factor (NGF) are expressed. The effects of these factors are varied, and include the stimulation of DNA synthesis, mitosis, and cell proliferation, expression of cytoskeletal filaments, and the modulation of ion channels. $Endothelin_B$ receptors are expressed in Müller cells and are upregulated following retinal injury.

Transporter systems

The high-affinity glutamate/aspartate transporter (GLAST, also named EAAT1) is present in Müller cells and serves to maintain a low concentration ($<1\ \mu M$) of glutamate in extracellular space. It utilizes the gradients of Na^+, K^+, and H^+ across the plasma membrane to transport glutamate into cells. A high-affinity GABA transporter of the GAT-3 type is also present in Müller cells.

Several acid/base transport systems are expressed in Müller cells. The transporters regulate intracellular pH (pH_i) and can influence extracellular pH. An electrogenic Na^+/HCO_3^- co-transporter is preferentially localized to the basal endfoot of Müller cells. A Na^+/H^+ exchanger and a Cl^-/HCO_3^- anion exchanger are also present in Müller cells. The anion exchanger is of the AE3 type and is localized to the cell endfoot. Intracellular as well as membrane-bound forms of the enzyme carbonic anhydrase (CA) are also expressed in Müller cells at high activity levels. The enzyme contributes to pH and CO_2 regulation in the retina (see below).

Calcium Signaling in Retinal Glial Cells

Calcium Waves

Activation of P2Y and other metabotropic receptors on Müller cells and astrocytes results in the release of Ca^{2+} from internal stores and to increases in intracellular Ca^{2+}. Calcium increases can also be evoked by mechanical or electrical stimulation of glial cells. Stimulation of a single glial cell often leads to the generation of propagated Ca^{2+} waves that travel through many astrocytes and Müller cells (Figure 3). Propagated Ca^{2+} waves travel at a velocity of approximately $23\ \mu m\ s^{-1}$ and can extend outwards for distances as great as $180\ \mu m$.

Glial cell-to-glial cell signaling during Ca^{2+} wave propagation is mediated by two mechanisms: by movement of the intracellular messenger IP_3, which diffuses through gap junctions coupling glial cells together; and by release of ATP, which functions as an extracellular messenger. Calcium wave propagation between astrocytes is mediated by both IP_3 diffusion and by ATP release while signaling between astrocytes and Müller cells and between Müller cells occurs solely by ATP release. The component of glial Ca^{2+} waves mediated by ATP release is blocked by purinergic antagonists.

Light-Evoked Ca^{2+} Signaling

In addition to agonist-evoked Ca^{2+} signals, Müller cells generate spontaneous, transient increases in Ca^{2+} in the absence of any applied stimulus. These endogenously generated Ca^{2+} increases occur at a frequency of 4.6 transients per cell per 1000 s in mammalian Müller cells. The transients range from 2.5 to 6 s in duration and are similar to those observed in astrocytes in the brain.

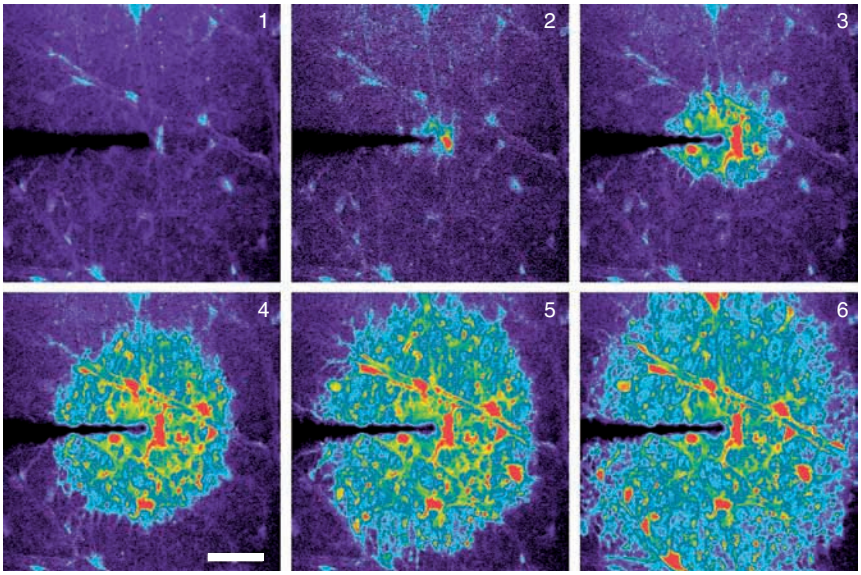

Figure 3 A Ca^{2+} wave propagated through retinal glial cells. Pseudocolor images of Ca^{2+} fluorescence within astrocytes (larger cells) and Müller cells (smaller spots) at the vitreal surface of the retina are shown. The Ca^{2+} wave is evoked by a mechanical stimulus applied to a single astrocyte. The wave is initiated at the stimulated cell (panel 2) and propagates outwards through neighboring astrocytes and Müller cells. Elapsed times following stimulation in panels 1–6 are 0, 0.2, 1.5, 3.5, 5.5, and 9.5 s, respectively. Scale bar = 50 μm. From Newman EA, unpublished.

Neuronal activity, evoked by light flashes, increases the frequency of Ca^{2+} transient generation in Müller cells by 28% (Figure 4). Signaling from activated neurons to Müller cells is mediated by release of ATP from the neurons and activation of glial P2Y purinergic receptors. Light-evoked Ca^{2+} increases are blocked by suramin, a purinergic antagonist, and by apyrase, which hydrolyzes ATP. Light-evoked Müller cell Ca^{2+} increases are also blocked by TTX, indicating that only retinal neurons that generate action potentials, amacrine and ganglion cells, are signaling to glial cells. Light-evoked increases in Müller cell Ca^{2+} transients are greatly potentiated by adenosine, which is a degradation product of ATP. Adenosine potentiation of neuron-glia signaling is of clinical interest as adenosine levels increase in certain pathological states.

Functions of Retinal Glial Cells

Regulation of the Extracellular Microenvironment

One of the principal functions of retinal glial cells is to regulate the composition of the fluid in which neurons are bathed. Neuronal activity results in changes in neurotransmitter, K^+, and H^+ levels in the extracellular space, all of which can, in turn, influence subsequent neuronal excitability. Müller cells play a major role in regulating these extracellular constituents.

Regulation of neurotransmitters Glutamate released from the synaptic terminals of photoreceptors and bipolar cells is removed from extracellular space primarily by uptake into Müller cells, mediated by GLAST, the high-affinity glutamate transporter. Glutamate, once taken up by the glial cells, is converted to glutamine and recycled back to the neuronal terminals where it serves as a substrate for the synthesis of additional neurotransmitter. Glutamate uptake by Müller cells influences the efficacy of synaptic transmission. When the Müller cell transporter GLAST is blocked, the magnitude of bipolar-to-ganglion cell synaptic transmission increases dramatically. Müller cells also express a high-affinity GABA transporter that removes this inhibitory transmitter from extracellular space following release from horizontal cell and amacrine cell synaptic terminals.

Regulation of pH Müller cells contribute to the regulation of extracellular pH (pH_o) in the retina. Neuronal activity results in the alkalinization of the extracellular space, which can dramatically alter the efficacy of synaptic transmission and neuronal excitability. The magnitude of this alkalinization may be reduced by the action of the Müller cell Na^+/HCO_3^- co-transporter. Müller cell depolarization, induced by neuronal activity, generates an acid efflux through the Na^+/HCO_3^- co-transporter, countering the alkalinization generated by the neurons. Müller cells may also regulate pH_o through the action of carbonic anhydrase, which is expressed by Müller cells in both cytoplasmic and membrane-bound forms. Carbonic anhydrase contributes to pH buffering by catalyzing the conversion of H^+ and HCO_3^- to CO_2

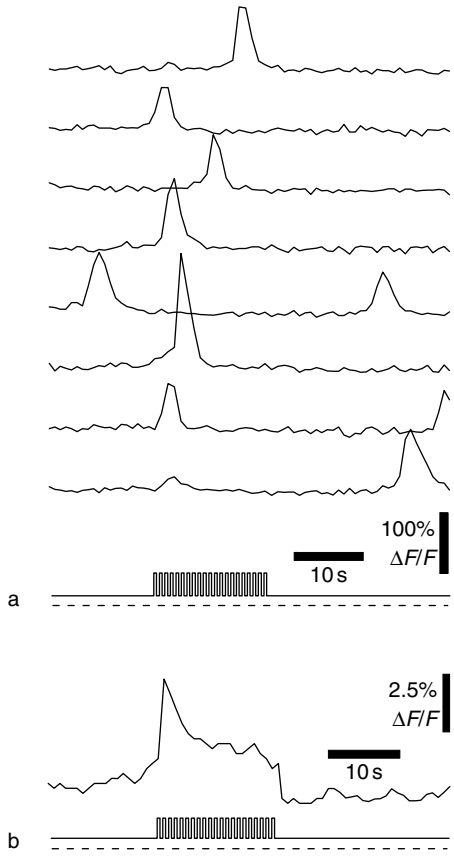

Figure 4 Light-evoked Ca^{2+} signaling in Müller cells. (a) Calcium fluorescence measured simultaneously in eight Müller cells. The retina was exposed sequentially to a uniform dim light, a bright flickering light, and a dim light (the light stimulus protocol is shown at the bottom in (a) and (b)). Calcium transients are more likely to be generated during the flickering light stimulus than during periods of constant illumination. (b) Mean Ca^{2+} fluorescence averaged over 84 trials. The flickering light evokes both a transient and a sustained increase in Ca^{2+} in Müller cells. From Newman EA (2005) Calcium increases in retinal glial cells evoked by light-induced neuronal activity. *Journal of Neuroscience* 25: 5502–5510.

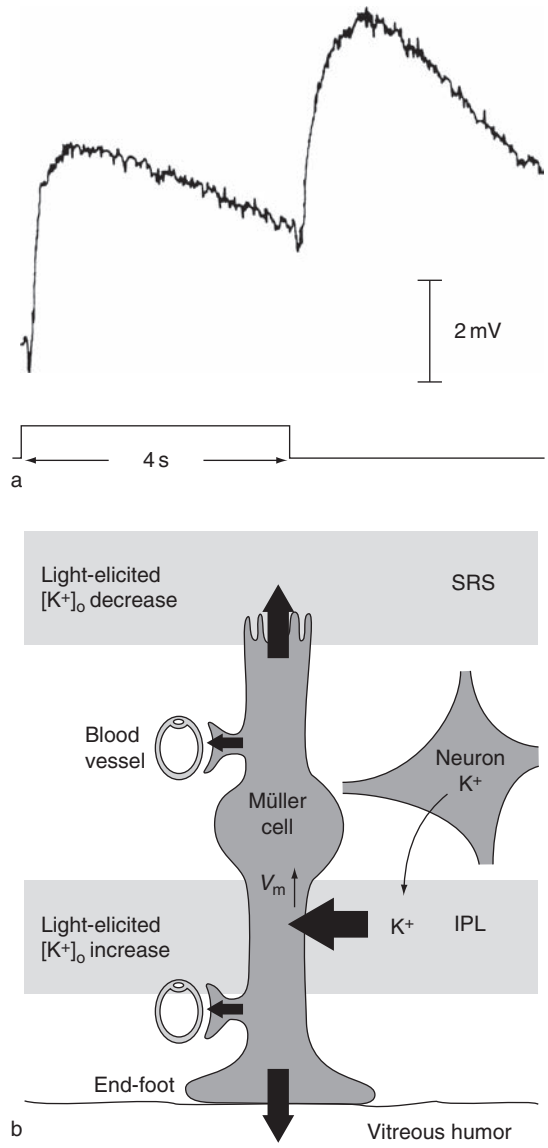

Figure 5 Müller cell regulation of K^+. (a) Light-evoked depolarizations recorded from a Müller cell of the mudpuppy. A 4 s light flash (bottom trace) evokes increases in extracellular K^+ concentration at light ON and OFF, resulting in K^+ influx into Müller cells and to cell depolarization. (b) Potassium siphoning in Müller cells. Active neurons release K^+ into the inner plexiform layer (IPL). The increase in extracellular K^+ concentration evokes a K^+ influx into Müller cells. The resulting cell depolarization generates an equal efflux of K^+ from other cell regions, into the vitreous humor, the subretinal space (SRS) and onto blood vessels. This K^+ current flowing through Müller cells reduces the initial K^+ increase in the IPL. (a) From Karwoski and Proenza (1977) Relationship between Müller cell responses, a local transretinal potential and potassium flux. *Journal of Neurophysiology* 40: 244–259. (b) From Newman (1966) *Neuroscientist* 2: 110–119.

and H_2O. Blocking the action of the enzyme leads to increases in the light-evoked pH_o alkalinization.

Regulation of potassium Müller cells play a key role in regulating extracellular K^+ ($[K^+]_o$) within the retina. Light-evoked neuronal activity results in increases in $[K^+]_o$ in the inner and outer plexiform layers and to a $[K^+]_o$ decrease in the distal retina. These $[K^+]_o$ variations will change the membrane potential of retinal neurons, altering their excitability. The $[K^+]_o$ variations are buffered by K^+ currents flowing through Müller cells, a process termed 'K^+ siphoning,' which is a specialized form of K^+ spatial buffering. Potassium released from active neurons results in an influx of K^+ into Müller cells. The K^+ influx depolarizes Müller cells (Figure 5(a)) and drives out an equal quantity of K^+ from other cell regions, primarily from cell endfeet, which have a high K^+ conductance. Potassium is released from basal endfeet into the vitreous humor, which acts as a large K^+ sink, and from apical processes into the subretinal space, where it counters the light-evoked $[K^+]_o$ decrease generated in this region (Figure 5(b)).

The influx and efflux of K^+ in Müller cells during the K^+ siphoning process occurs largely through Kir4.1 inwardly rectifying K^+ channels. The voltage- and K^+-dependent properties of these channels enhance the K^+ siphoning process; channel conductance increases with increased $[K^+]_o$, resulting in greater siphoning currents and a more rapid transfer of K^+ to the vitreous humor and the subretinal space. When Müller cell Kir4.1 channels are blocked by Ba^{2+}, the siphoning current is interrupted and $[K^+]_o$ regulation in the retina is compromised.

Generation of the electroretinogram A consequence of K^+ current flow through Müller cells is the generation of extracellular field potentials within the retina. These light-evoked potentials can be recorded with an electrode on the cornea as components of the electroretinogram (ERG). For many years, the b-wave, the most prominent component of the ERG, was believed to be generated by K^+ current flow through Müller cells. However, if K^+ siphoning is prevented by blocking Müller cell Kir channels with Ba^{2+}, or if the channels are eliminated by employing Kir4.1 knockout animals, the b-wave is not reduced. These experiments demonstrate conclusively that Müller cells do not generate the ERG b-wave. The b-wave is generated, instead, by bipolar cells, as originally suggested by Tomita.

Other components of the ERG are generated by light-evoked K^+ current flow through Müller cells, however. The slow PIII response, the retinal component of the ERG c-wave, is generated by Müller cell K^+ current flow established by a decrease in $[K^+]_o$ in the distal retina. Müller cells also generate the scotopic threshold response, a rod-driven response generated at the ON of light in mammals, and the M-wave, a negative response with prominent ON and OFF components. Both responses are generated by K^+ current flow through Müller cells established by $[K^+]_o$ increases in the inner plexiform layer.

Modulation of Neuronal Activity

In recent years, it has become clear that glial cells in the CNS modulate neuronal activity by releasing gliotransmitters. In the retina, stimulation of Müller cells results in changes in light-evoked spiking of ganglion cells. Spike activity can either be potentiated or depressed following glial cell stimulation and the initiation of Ca^{2+} waves in retinal glial cells. It is likely that Müller cells, in particular, modulate neuronal activity by several different mechanisms, two of which are described below.

Inhibition of ganglion cells Selective stimulation of Müller cells by agonist ejection evokes large Ca^{2+} increases in these cells and subsequent hyperpolarization of adjacent ganglion cells (Figure 6(a)). The hyperpolarization can be sufficiently strong to completely block the generation of action potentials (Figure 6(b)). The ganglion cell hyperpolarization is mediated by ATP release from the Müller cells. Released ATP is rapidly converted to adenosine by ecto-ATPases and ecto-nucleotidases. Adenosine, in turn, activates A_1 adenosine receptors on the ganglion cells, leading to the opening of K^+ channels and to cell hyperpolarization.

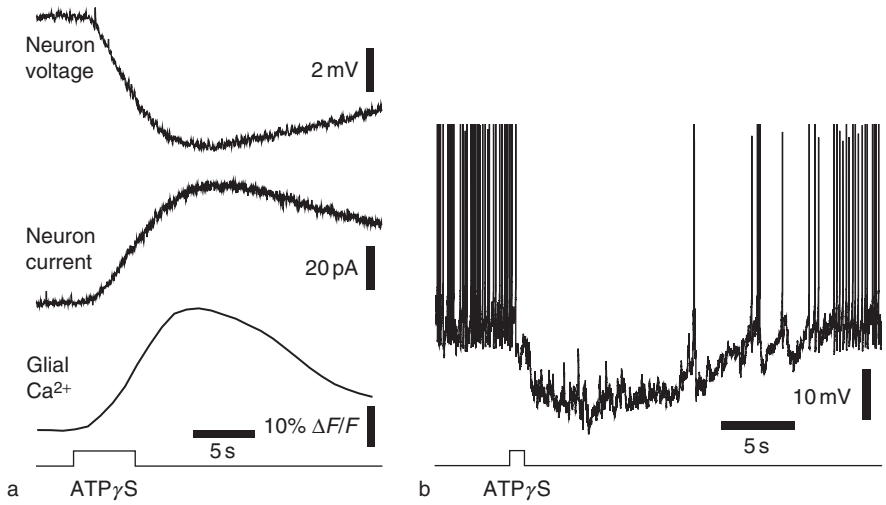

Figure 6 Glial cell inhibition of action potential generation in ganglion cells. (a) Ejection of ATPγS onto the retinal surface evokes a Ca^{2+} increase in glial cells (glial Ca^{2+}). This Ca^{2+} increase is associated with a hyperpolarization (neuron voltage) and an outward current (neuron current) in a neighboring ganglion cell. (b) ATPγS stimulation of glial cells hyperpolarizes a ganglion cell and blocks the generation of action potentials in the cell. From Newman EA (2003) Glial cell inhibition of neurons by release of ATP. *Journal of Neuroscience* 23: 1659–1666.

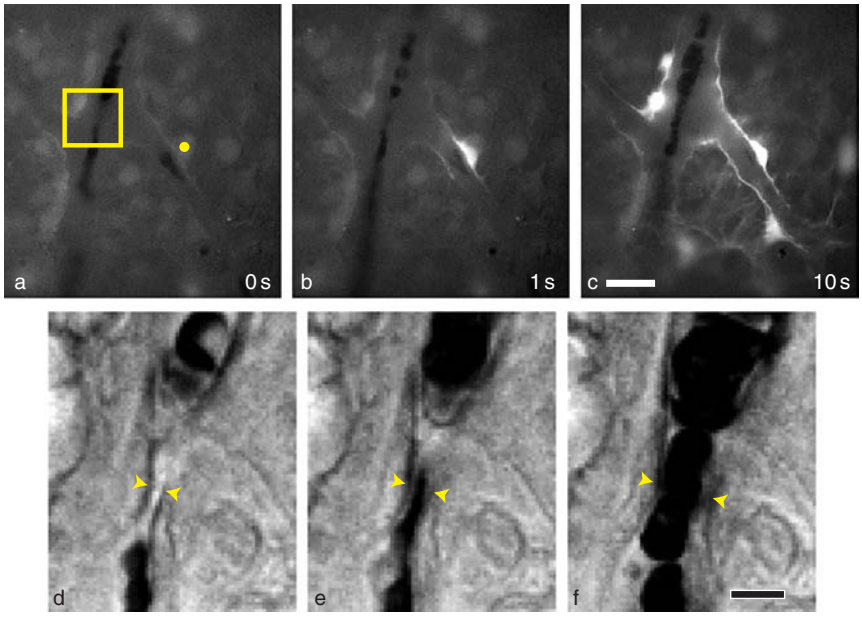

Figure 7 Glial cell-evoked dilation of a neighboring blood vessel. (a–c) Fluorescence images showing Ca^{2+} concentration within retinal glial cells. (d–f) Infrared differential interference contrast (IR-DIC) images showing the arteriole within the region indicated in (a). Photolysis of caged-Ca^{2+} in the glial cell indicated by the circle in (a) evokes a Ca^{2+} increase that propagates from the stimulated cell into neighboring glial cells surrounding the blood vessel (a–c). Glial cell stimulation results in the dilation of the adjacent arteriole (d–f). From Metea MR and Newman EA (2006) Glial cells both dilate and constrict blood vessels. A mechanism of neurovascular coupling. *Journal of Neuroscience* 26: 2862–2870.

D-serine regulation of NMDA receptors Müller cell stimulation may also lead to the release of D-serine and to potentiation of NMDA receptor neurotransmission in the retina. Müller cells contain D-serine, an endogenous NMDA receptor co-agonist, and serine racemase, the synthetic enzyme for D-serine. Stimulation of brain astrocytes results in D-serine release and a similar release of D-serine may occur in retinal Müller cells.

Regulation of Blood Flow

Light stimulation results in an increase in blood flow in retinal arteries, bringing increased oxygen and nutrients to activated neurons. Müller cells mediate neuron to blood vessel signaling, a process termed neurovascular coupling. When Müller cells are selectively stimulated, neighboring arterioles either dilate or constrict (Figure 7). Vasodilation is mediated by production of EETs while vasoconstriction is mediated by production of 20-HETE. Both vasoactive agents are metabolites of arachidonic acid. Light stimulation also evokes either vasodilation or constriction by generation of the same two arachidonic acid metabolites. When neuron to Müller cell signaling is blocked by a purinergic antagonist, light-evoked vasomotor responses are also blocked, indicating that Müller cells mediate neurovascular coupling.

Metabolic and Trophic Support of Neurons

Müller cells play an active role in supporting neuronal metabolism. Glucose is metabolized in Müller cells to lactate, which is then transferred to neurons where it serves as a primary energy source for oxidative metabolism. Müller cells also contain large reserves of glycogen, which serve as a source of glucose during times of high metabolic activity.

Retinal glial cells secrete a number of growth factors which have trophic effects on retinal neurons and blood vessels. Basic fibroblast growth factor (bFGF), which promotes photoreceptor survival, is synthesized by Müller cells and retinal astrocytes. Müller cells also secrete a number of other growth factors including brain-derived neurotrophic factor, NGF, neurotrophin-3, neurotrophin-4, and glial cell line-derived neurotrophic factor. Both Müller cells and astrocytes synthesize VEGF, which plays an essential role in the development of the retinal vasculature.

See also: Astrocyte: Calcium Signaling; Fovea: Primate; Retina: An Overview; Retinal Color Mechanisms; Retinal Development: An Overview; Retinal Development: Cell Type Specification; Retinal Pharmacology: Inner Retinal Layers.

Further Reading

Brew H and Attwell D (1987) Electrogenic glutamate uptake is a major current carrier in the membrane of axolotl retinal glial cells. *Nature* 327: 707–709.

Brew H, Gray PTA, Mobbs P, and Attwell D (1986) Endfeet of retinal glial cells have higher densities of ion channels that mediate K^+ buffering. *Nature* 324: 466–468.

Karschin A, Wassle H, and Schnitzer J (1986) Shape and distribution of astrocyte in the cat retina. *Investigative Ophthalmology and Visual Science* 27: 828–831.

Karwoski CJ, Lu H-K, and Newman EA (1989) Spatial buffering of light-evoked potassium increases by retinal Muller (glial) cells. *Science* 244: 578–580.

Kofuji P, Ceelen PW, Zahs KR, Surbeck LW, Lester HA, and Newman EA (2000) Genetic inactivation of an inwardly rectifying potassium channel (Kir4.1 subunit) in mice: Phenotypic impact in retina. *Journal of Neuroscience* 20: 5733–5740.

Metea MR and Newman EA (2006) Glial cells both dilate and constrict blood vessels: A mechanism of neurovascular coupling. *Journal of Neuroscience* 26: 2862–2870.

Newman EA (1984) Regional specialization of retinal glial cell membrane. *Nature* 309: 155–157.

Newman EA (2001) Glia of the retina. In: Ryan SJ (ed.) *Retina*, pp. 89–103. St. Louis: Mosby.

Newman EA (2003) Glial cell inhibition of neurons by release of ATP. *Journal of Neuroscience* 23: 1659–1666.

Newman EA (2005) Calcium increases in retinal glial cells evoked by light-induced neuronal activity. *Journal of Neuroscience* 25: 5502–5510.

Newman EA and Zahs KR (1997) Calcium waves in retinal glial cells. *Science* 275: 844–847.

Newman EA and Zahs KR (1998) Modulation of neuronal activity by glial cells in the retina. *Journal of Neuroscience* 18: 4022–4028.

Poitry-Yamate CL, Poitry S, and Tsacopoulos M (1995) Lactate released by Muller glial cells is metabolized by photoreceptors from mammalian retina. *Journal of Neuroscience* 15: 5179–5191.

Ramon y Cajal S (1972) *The Structure of the Retina*. Springfield, IL: Thomas, C.C.

Sarthy V and Ripps H (2001) *The Retinal Muller Cell*. Kluwer Academic/Plenum Publishers.

Stevens ER, Esquerra M, Kim P, et al. (2003) d-serine and serine racemase are present in the vertebrate retina and contribute to the functional expression of NMDA receptors. *Proceedings of the National Academy of Sciences of United States of America* 100: 6789–6794.

Retinal Horizontal Cells

I Perlman and E Weiner, Technion – Israel Institute of Technology, Haifa, Israel
H Kolb, University of Utah, Salt Lake City, UT, USA

© 2009 Elsevier Ltd. All rights reserved.

Morphological Characteristics of Horizontal Cells

Horizontal cells are second-order neurons interconnecting photoreceptors laterally across the retina at the first synaptic layer. They are wide-spreading, horizontally oriented neurons with dendrites that connect with the synaptic endings of photoreceptors. The morphological forms of horizontal cells found in vertebrate retinas are shown in Figure 1. Through Golgi staining, intracellular marking, and immunostaining techniques we have learned of the various morphologies horizontal cells can adopt in different vertebrate species (Figures 1–3).

There are at least two types of horizontal cells in the majority of vertebrate retinas, B-types, which have axons, and A-types, which are axonless (Figures 1 and 3(c)). In general, the cells bodies and dendritic ends of horizontal cells contact cone pedicles of cone photoreceptors (Figures 2(a) and 2(b)), while the axonal terminals (when present) contact rods. In nonmammalian species, horizontal cells come in several axonless subtypes with color-specific cone connections (Figure 1). Rod-dominated, nonmammalian retinas (such as those found in fish) also have a horizontal cell type devoted solely to the rods, while in rod-dominated, mammalian retinas the axonal terminal portions of the B-type cells that connect with rods act essentially independently from the cone-connected cell bodies.

B-Type horizontal cells are only concerned with light intensity changes and shaping adaptation and spatial responses of vertical pathway neurons. They are the luminosity-type horizontal cells that will be discussed later. B-Type cells can often be identified with antibodies against calcium-binding proteins (Figure 2(a)) and γ-aminobutyric acid (GABA) neurotransmitter. In turtle retina, acetylcholine has been located to H1-type horizontal cell dendrites where they invaginate the cone pedicles (Figures 2(b) and 3(a)). A-Type horizontal cells in mammals are purely cone connecting (most species have only green and blue cones). They have no connections with rod photoreceptors at all. Even in the case of the H2 cells of the trichromatic primate retina, where one or more dendrites are elongated and axonlike, their connections are with cones, and interestingly enough with blue cones (Figure 1). This may be a general theme for mammalian axonless horizontal cells that is only now being realized. In certain big cats (not yet confirmed in domestic cats), the elongated dendritic tips of A-type cells contact blue cones, while the dendrites closer to the cell body are in contact with both green cones and blue cones. In horses the A-type cell is apparently only connected to blue cones (Figure 1).

In turtles, birds, and fish, species with excellent color vision, type H2 and type H3 horizontal cells are large, stellate cells lacking axons, and are concerned solely with cone pathways (Figure 1). They are the chromaticity (C-type) horizontal cells that will be described later. Thus, in turtle (pentachromats) and fish (trichromats or tetrachromats), H2 horizontal cells connect to green and blue cones, and H3 cells connect to blue cones (Figure 1). The C-type horizontal cells of the turtle can be labeled with antibodies to nitric oxide synthase and calcium-binding proteins, suggesting a role for NO in their functioning (see later section).

Horizontal cells always interact with photoreceptor terminals at their knoblike dendritic terminals (Figure 2(b)) that invaginate the photoreceptor synaptic complex as lateral elements at a presynaptic ribbon (Figure 3(a)). The postsynaptic receptors at the lateral elements are ionotropic glutamate receptors (iGluRs). Retinal horizontal cells in all species are also connected to their homologous neighbors, by areas of gap junctions between their dendrites (Figure 3(b)) and their axon terminals, and occasionally between their dendrites and cell bodies. These areas of gap junction can be extremely large, appearing as plaques of electrical junction (Figure 3(b)).

Gap junctions are formed at closely applied plasma membranes (2–4 nm gaps) of the two horizontal cell structures (Figure 3(b)). Each half of the cell supplies connexons, or hemichannels, to complete the gap junction channel. These channels can consist of homomeric or heteromeric connexons, thus allowing a huge variety of gap junctions with slightly different properties to exist between various neurons of the nervous system. Each connexon channel contains six connexins (Cx) surrounding the channel pore. In fish and turtle retinas Cx43 and Cx26 have been demonstrated to comprise the horizontal cell gap junction connexons. Procion Yellow dye can diffuse through the gap junctions and label the horizontal cell syncytium, as seen in Figure 3(c). In this figure, the central horizontal cell is brightly stained, and the neighboring horizontal cells are faintly stained, by dye diffusion through the gap junctions. Not only dye molecules but also ion and cytoplasmic molecules of

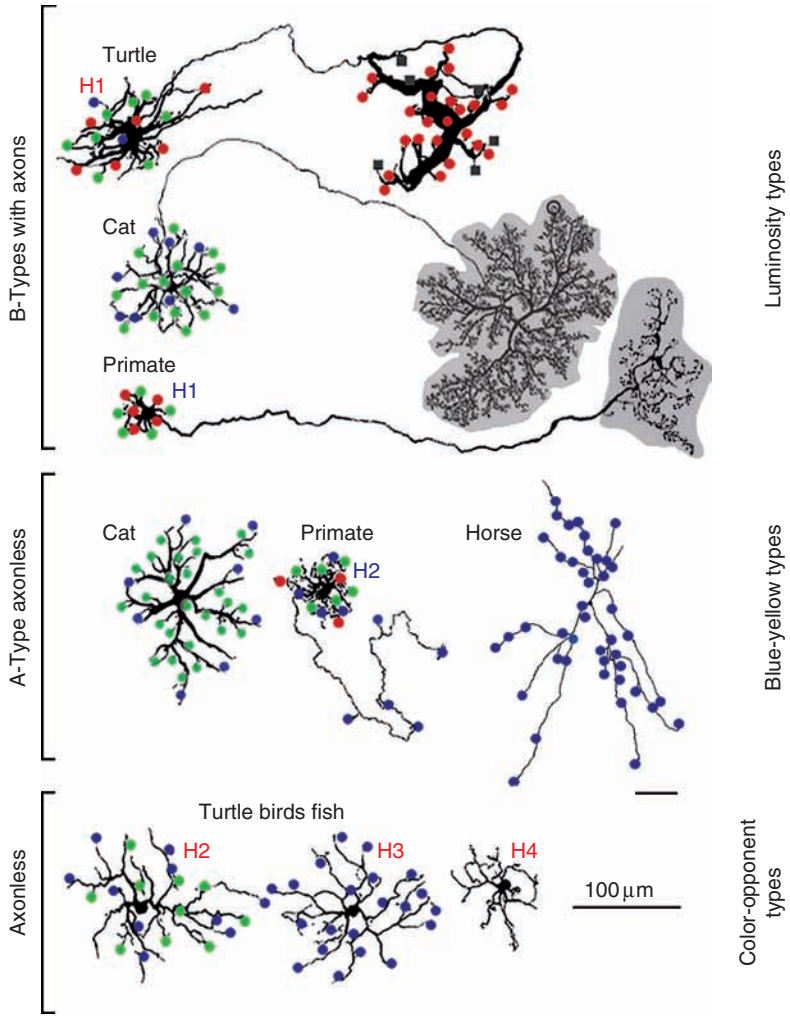

Figure 1 The morphologies of different types of horizontal cells and their known connectivities with rod and cone photoreceptors. The cells, Golgi stained or visualized using neurobiotin-filled staining techniques, are viewed in flat, whole-mounted preparations. B-Type cells with axons and axonless cell types exist in most vertebrate retinas. The physiological types that have been proved to be associated with the morphological types are indicated (i.e., luminosity, blue-yellow, and color-opponent types). Red dots represent red cones; green dots represent green cones; blue dots represent blue cones; black squares or shaded gray areas represent rod inputs.

small size can pass through the connexon pores of the gap junction, changing horizontal cell activities and ionic properties.

Horizontal cells are among the earliest neurons to develop in the vertebrate retina. They develop at about the same time as amacrine cells and ganglion cells. The horizontal cells emit processes and are at first oriented vertically like all the primordial neurons of the retina (Figure 4(a); P3). At first, horizontal cell dendrites stretch in many directions, as they seek a point of contact with cone terminals. But once they have found their cone contacts they become flattened and assume the more adult form of a horizontally running neuron with dendrites on one plane in the outer plexiform layer (Figure 4(a); P12). Horizontal cell axonal processes assume their contacts with rod photoreceptors a few days later in development.

Like during development, horizontal cells undergo bizarre morphological and structural changes after retinal detachment or retinal degenerations whereby the photoreceptors are lost. When the rods and cones are degenerating they retract their axons and become autolysed. The horizontal cells degenerate with strange transformation of their dendritic and axonal processes (Figure 4(b)). Some horizontal cell outgrowths are fairly thick processes while others are thin, often beaded cylindrical processes (Figure 4(b); red processes). These ascending ribbonlike processes occupy the subretinal space along with massive numbers of hypertrophied Müller cell processes (Figure 4(b); green processes). It is as though loss of

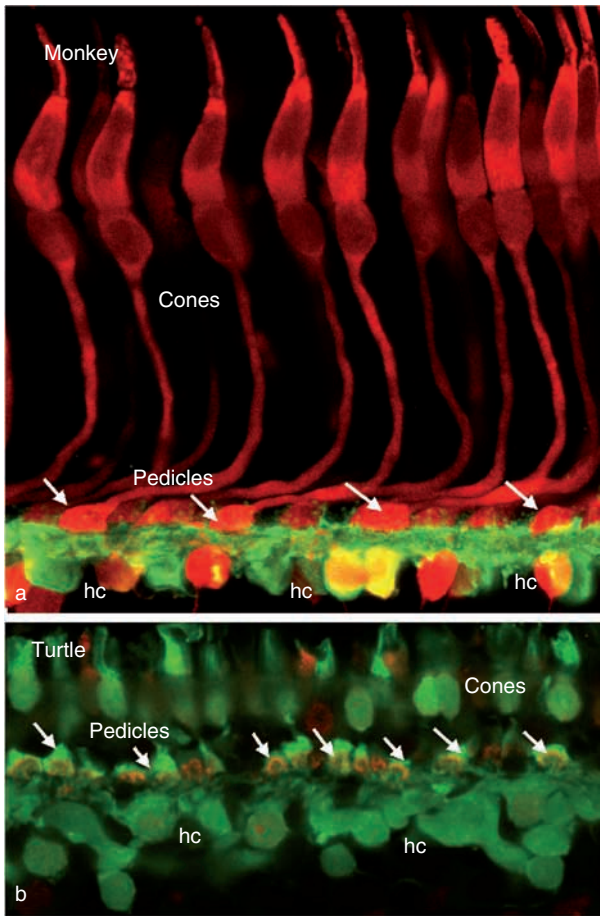

Figure 2 (a) Section of monkey retina immunostained with anticalbindin (red) and antiparvalbumin (green). The cone photoreceptors stain in their entirety and their pedicles contact the dendrites of horizontal cells (hc; green). Some horizontal cells co-localize parvalbumin and calbindin (yellow). (b) Section of turtle retina immunostained with anticalbindin (green) and anticholine acetyltransferase (red). The cones and horizontal cell structures (hc) are making synaptic contacts at the horizontal cell dendrites. The tips of the horizontal cell dendrites contain acetylcholine (red semicircles, arrows). Micrographs courtesy of Nicolas Cuenca.

photoreceptor contacts in retinal degenerations is the reverse of the morphological changes that occur in development. Thus, horizontal cells of the vertebrate retina can become highly plastic under abnormal retinal conditions.

Physiological Types of Horizontal Cells

The first light-evoked electrical responses that were recorded with intracellular microelectrodes from nerve cells in the vertebrate retina were called S-potentials. At that time, neurons were believed to be depolarized (inside becoming more positive relative to outside) by excitatory synaptic inputs. If the depolarization was of sufficient amplitude, action potentials, or nerve spikes, were generated to transfer signals down the length of the axon. The S-units, however, responded to light with graded hyperpolarization (Figure 5). Intracellular marking techniques, in which dyes were injected from the electrode tips into the cytoplasm of the recorded neurons, revealed the horizontal cells to be the source of the S-potentials. Since first described in fish retinas, S-potentials have been recorded from retinal horizontal cells in all vertebrate classes studied (Figure 5).

Horizontal cells are separated into two major physiological types based upon their responses to chromatic light stimuli, luminosity-type and chromaticity-type horizontal cells.

Luminosity-Type Horizontal Cells

Luminosity-type (L-type) horizontal cells are found in every retina that has been studied. Figure 5 shows the photoresponses of L-type horizontal cells from turtle, rabbit, and monkey retinas. These cells respond with graded hyperpolarization of amplitude and duration, depending upon the intensity and duration of the light stimuli. The relationship between peak response amplitude and stimulus intensity fits a generalized hyperbolic function,

$$V = V_{\max}[I^n/(I^n + \sigma^n)] \qquad [1]$$

where V and $V_{\max}$ are the amplitudes of the responses to a stimulus of intensity I and to a stimulus of supersaturating intensity, respectively. The parameter σ is the semisaturation constant and equals the stimulus intensity needed for response amplitude of half-maximum, and n is an exponent that depends upon the steepness of the curve. In L-type horizontal cells the exponent n ranges from 0.7 to 1.2, depending upon species and upon state of adaptation. The response–intensity relationship of retinal neurons is not linear and saturates at bright intensities.

In some species, two classes of L-type horizontal cells can be identified according to their anatomical structure and physiological properties. In the turtle retina, the L1- and L2-types are similar in their spectral sensitivity, but differ in receptive field size and in the kinetics of their photoresponses to red and green light. Anatomically, these two L-type horizontal cells are, respectively, the soma and axon terminal of the H1 horizontal cell (Figure 1). In the mammalian retina, type A and B luminosity horizontal cells differ in morphology and in their major excitatory input, but respond similarly with graded hyperpolarization to light stimuli.

In primate retina, horizontal cells also only occur as luminosity types. Some of these cells receive synergistic signals from long- and medium-wavelength

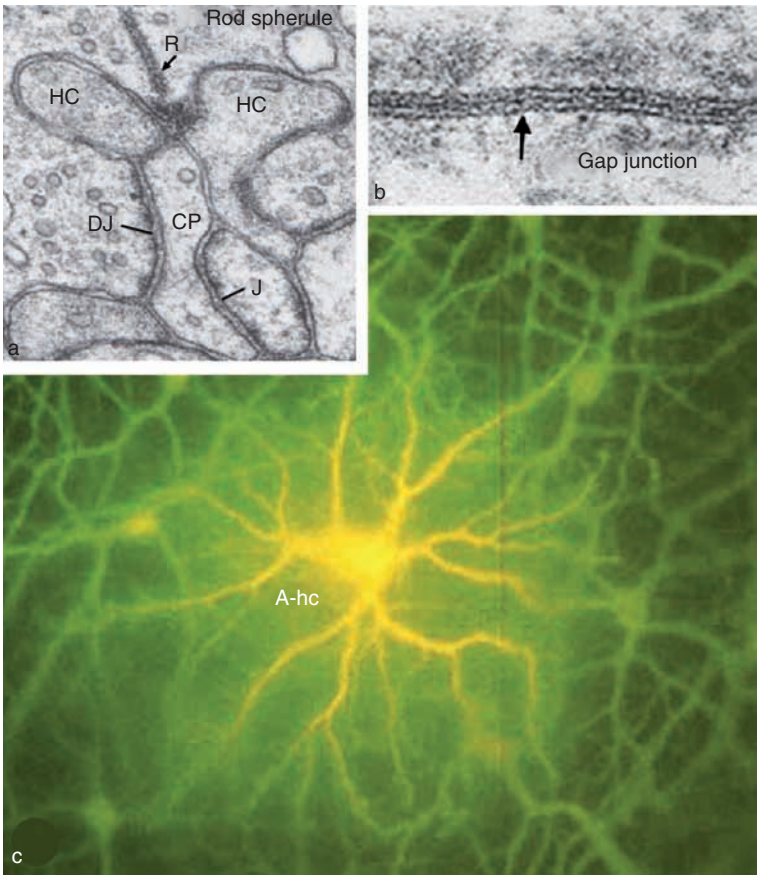

Figure 3 (a) Electron micrograph of a triad synapse in a rod photoreceptor spherule in the monkey retina. The horizontal cell processes (HC) form lateral elements of the triad of processes at the single synaptic ribbon (R). The central process (CP) is a bipolar cell dendrite (DJ, distal junction; J, junction). (b) High-magnification electron micrograph of a gap junction between two horizontal cell dendrites in the cat retina. (c) An A-type horizontal cell (A-hc) in the rabbit retina has been intracellularly stained with neurobiotin (yellow). All of the surrounding A-type horizontal cells have also become more faintly yellow stained because the stain has passed through gap junctions linking neighboring horizontal cell dendrites. (a) From Lasansky A (1972) Cell junctions at the outer synaptic layer of the retina. *Investigative Ophthalmology* 11: 265–275. (c) From Mills SK and Massey SC (1994) Distribution and coverage of A-type and B-type horizontal cells stained with Neurobiotin in the rabbit retina. *Visual Neuroscience* 11: 549–560.

cones (red and green cones), while others receive synergistic input from all three spectral types of cones: long-, medium-, and short-wavelength (red, green, and blue) cones. Anatomically, H1-type horizontal cells tend to avoid the pedicles of blue cones, and are therefore not responsive to blue light stimuli physiologically. On the other hand, many dendrites of H2-type horizontal cells make synaptic contacts with blue cone pedicles, and they are, therefore, very sensitive physiologically to the blue end of the spectrum physiologically.

Chromaticity-Type Horizontal Cells

In 1958, Svaetichin and MacNichol first reported on a class of S-units, responding with graded depolarization to long-wavelength stimuli and with graded hyperpolarization to short-wavelength stimuli. These S-units were called chromaticity-type (C-type) horizontal cells. C-Type horizontal cells have been extensively studied in cold-blooded vertebrates (e.g., fish, turtle, toad, salamander), but were not found in mammals and primates. C-Type horizontal cells are identified by the number of wavelengths at which response polarity reverses (one reversal, biphasic cells; two reversals, triphasic cells), and by the wavelength of response reversal. In the bowfin retina, response reversal for the biphasic cells is around 640 nm, while for the triphasic cells one response reversal occurs in the region of 500–530 nm and the other in the region of 650–670 nm. The biphasic and triphasic C-type horizontal cells of the fish retina are identified morphologically as types H2 and H3, respectively (Figure 1).

In the turtle retina, two classes of biphasic C-type horizontal cells have been described. For the turtle *Mauremys caspica*, typical photoresponses of C-type horizontal cells to light stimuli of different wavelength and intensity are compared in Figure 6 to

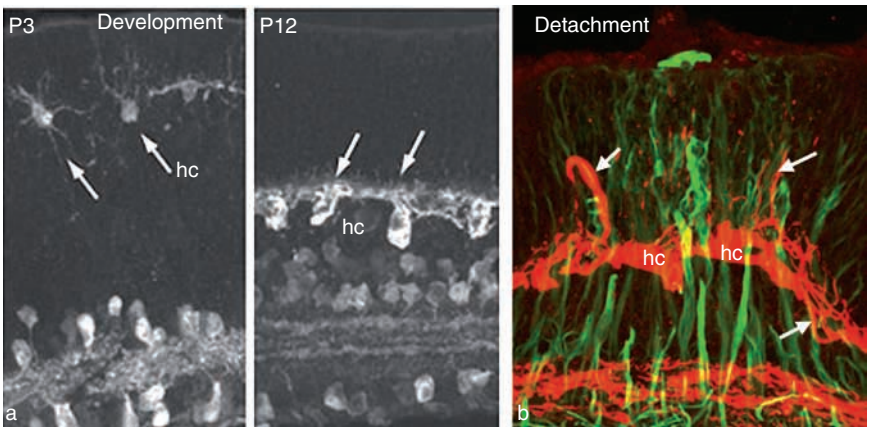

Figure 4 (a) The appearance of horizontal cells during the course of development in the neonatal mouse retina. At postnatal day 3 (P3) the horizontal cells (hc; arrows) are irregular in shape, with dendrites pointing in all directions, and have not reached the plane of the outer plexiform layer. By P12 the cells are regularly arranged and form a horizontally organized layer of cells at the outer plexiform layer. (b) After a retinal detachment whereby the photoreceptors are no longer in contact with the pigment epithelium, the retina undergoes significant remodeling. The horizontal cells (hc; red) become dysplasic and their dendrites develop strange growth modes. They appear to be searching for contacts, with dying photoreceptor terminals. Müller cells (green) also become filled with intermediate filaments and vimentin (green) and grow abnormally extensive. (a) From Morgan J and Wong R (2005) Development of cell types and synaptic connections in the retina. (b) From Fisher SK, Lewis GP, Linberg KA, et al. (2005) Cellular remodeling in mammalian retina induced by retinal detachment.

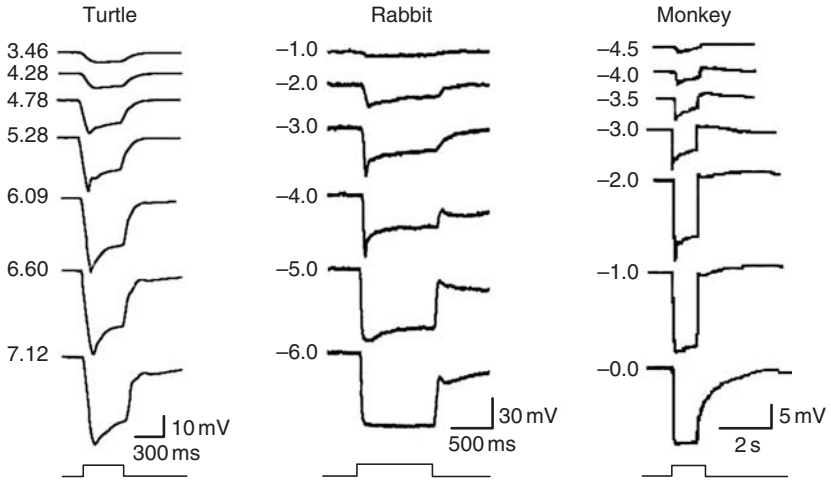

Figure 5 Photoresponses recorded with intracellular microelectrodes from luminosity horizontal cells in turtle, rabbit, and monkey. The photoresponses were elicited by light stimuli of different intensities (given in log relative units to the left of each response). Rabbit recording from Bloomfield SA and Miller RF. A physiological and morphological study of the horizontal cell types of the rabbit retina. *Journal of Comparative Neurology*, 208: 288–303. Copyright (1982), John Wiley & Sons, Inc; reprinted with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc. Monkey recording from Dacheux RF and Raviola E (1990) Physiology of HI horizontal cells in the primate retina. Proceedings of the *Royal Society of London, Series B, Biological Sciences* 239: 203–221.

those of an L1-horizontal cell. The L1-horizontal cells (Figure 6, top row) respond with graded hyperpolarization to light stimuli of any wavelength and are characterized by action spectra with peak sensitivity in the red range of the visible spectrum (640–650 nm). Red/green C-type horizontal cells (Figure 6, middle row) respond with graded depolarization to red light stimuli and with graded hyperpolarization to yellow, green, and blue stimuli. Response reversal occurs at wavelengths from 600 to 620 nm. Yellow/blue horizontal cells (Figure 6, bottom row) respond with graded depolarization to red and yellow light stimuli and with graded hyperpolarization to blue light stimuli. In these cells, response reversal occurs in the range 540–560 nm. The red/green and yellow/blue C-type horizontal cells of the turtle retina are identified morphologically as types H2 and H3, respectively (Figure 1).

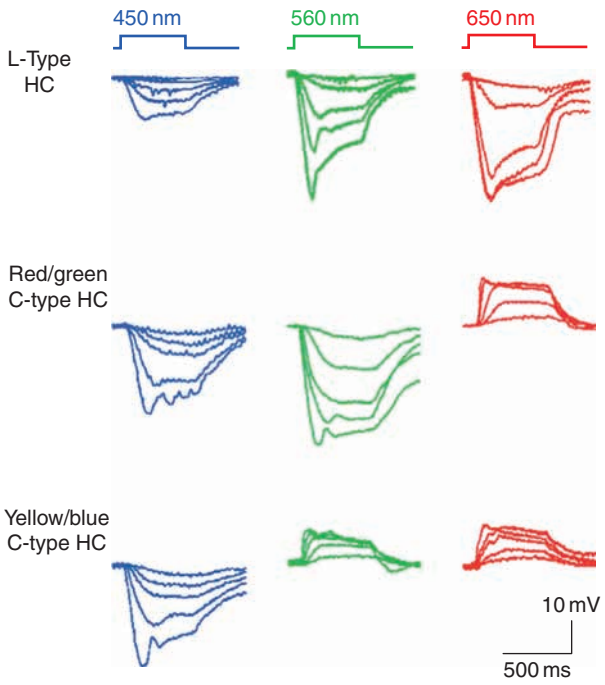

Figure 6 Photoresponses of L-type and C-type horizontal cells in the turtle retina were elicited by light stimuli of different wavelength and intensities From Perlman I, Kolb H, and Nelson R (2003) Anatomy, circuitry, and physiology of vertebrate horizontal cells. In: Chalupa LE and Werner JS (eds.) *The Visual Neurosciences*, pp. 369–394. Cambridge, MA: MIT Press.

Synaptic Transmission between Photoreceptors and Horizontal Cells

Feed-Forward Excitatory Synapse

Photoreceptors make excitatory (sign-conserving) synapses onto horizontal cells. In darkness, the photoreceptors continuously release, in a calcium-dependent manner, an excitatory neurotransmitter that opens cationic channels with a reversal potential close to 0 mV. Therefore, the horizontal cells are maintained at a relatively depolarized potential in darkness. When the photoreceptors are hyperpolarized by a light stimulus, transmitter release is reduced. The concentration of the neurotransmitter in the synapse reduces, the postsynaptic ligand-gated channels close, and the horizontal cells hyperpolarize toward the potassium equilibrium potential.

L-Glutamate is the excitatory neurotransmitter mediating light-evoked signals from the photoreceptors to second-order neurons (horizontal cells and bipolar cells), as supported by pharmacological experiments. Exposing the retina to L-glutamate causes depolarization of the horizontal cell and loss of its photoresponses. This is consistent with the notion that exogenous application of the neurotransmitter saturates the receptor sites in the horizontal cells causing further depolarization and preventing the effects of endogenous light-induced changes in neurotransmitter release from affecting the horizontal cell. This conclusion is supported using kainic acid, an agonist of L-glutamate. In contrast, kynurenic acid, a nonselective antagonist of excitatory amino acids receptors, induces the opposite effects; the horizontal cell hyperpolarizes and loses its photoresponses.

Immunostaining techniques and pharmacological analysis of the properties of the glutamate-gated channels in horizontal cells have revealed the cells to be of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid/kainic acid (AMPA/KA) type. They are specifically activated by AMPA or KA and are antagonized by 6-cyano-7-nitroquinoxaline (CNQX) or *cis*-2,3-piperidinedicarboxylic acid (PDA).

Feedback Inhibitory Synapse

Horizontal cells are postsynaptic to the photoreceptors, but are also presynaptic to them. The effects of these pathways can be revealed with intracellular recordings from cone photoreceptors, while adjusting the spatial properties of the light stimuli. In turtle retina, a stimulus composed of large spot of red light can elicit a depolarizing response in the medium-wavelength cone, while a green stimulus of the same dimensions elicits a hyperpolarizing response. Reducing the size of the illuminated retinal area reduces the differences between the responses to red and green light stimuli.

The negative feedback pathway from horizontal cells to cones is thought to take place through the horizontal cell lateral elements invaginating the photoreceptor terminal at the triad ribbon synapse. The exact mechanism(s) underlying the negative feedback pathways from horizontal cells to cone photoreceptors is still under debate. GABA was suggested as the candidate for the inhibitory feed back neurotransmitter. This notion was supported by experiments demonstrating GABA-induced currents in the terminals of red-sensitive and green-sensitive cones isolated from the turtle retina. Furthermore, color opponency in fish C-type horizontal cells could be eliminated upon exposure to GABA or its antagonists. However, other studies in other species failed to show GABA effects that were supportive for a role of GABA as the inhibitory neurotransmitter of the horizontal cell. A recent theory suggested that large extracellular currents flowing into the horizontal cells through hemi-gap-junction channels, at their lateral elements in the triad synapse, change the potential difference across the plasma membrane of the photoreceptors at the terminal region. Any change in this current will change the potential difference and thus the activation of voltage-dependent calcium channels in the

photoreceptors' terminals. According to this theory, neurotransmitter release from the photoreceptors can be altered by the biophysical properties of the horizontal cells.

Receptive Field Properties of Horizontal Cells

Gap junctions between neighboring horizontal cells are permeable to dyes that can be used to assess the size of the horizontal cell syncytium. Dye coupling, using Neurobiotin, for turtle H2 cells and for rabbit A-type horizontal cells is shown in Figure 7(a). It is clear that the entire field of view represents one large syncytium that is composed of cells with similar morphology. The gap junctions between neighboring horizontal cells are also highly permeable to small ions, allowing lateral spread of electrical signals within the horizontal cell layer. This leads to the genesis of large homogeneous receptive fields that spread to retinal regions outside the range of the immediate dendritic fields of the individual horizontal cells.

Different procedures have been used to characterize the receptive fields of horizontal cells. The most common experimental procedures have been the application of a series of concentric spots of increasing diameter but fixed intensity, or moving a narrow slit of light across the retina. Figure 7(b) shows an

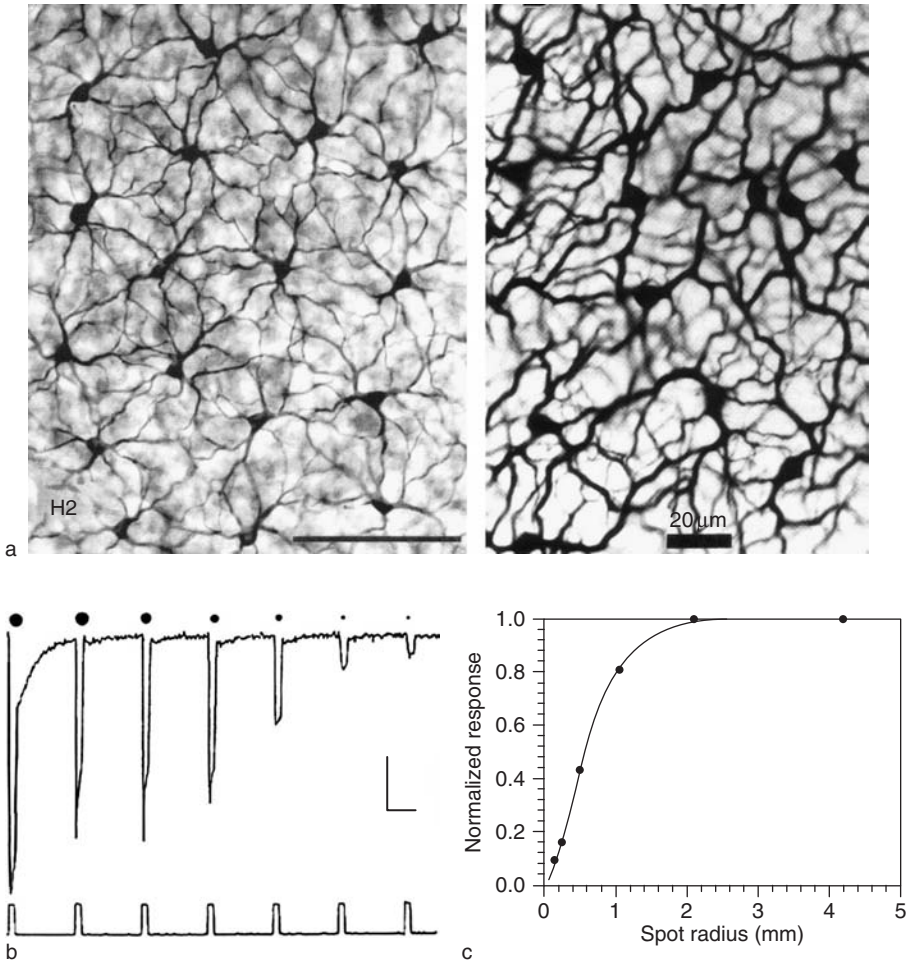

Figure 7 (a) Neurobiotin-coupling of H2 horizontal cells in the turtle (left) and of A-type horizontal cells in the rabbit retina (right). (b) Receptive field properties of L-type horizontal cell in the turtle retina. Photoresponses to light stimuli of constant intensity but different spatial pattern, as seen above each response (dots), were recorded with an intracellular microelectrode. (c) The relationship between peak response amplitude and radius of the illuminated retinal region reveals the extent of the receptive field. Scale bar = 50 μm (a). (a) Turtle micrograph from Ammermüller J and Kolb H (1996) Functional architecture of the turtle retina. *Progress in Retinal and Eye Research* 15: 393–433; rabbit micrograph from Mills SK and Massey SC (1994) Distribution and coverage of A-type and B-type horizontal cells stained with Neurobiotin in the rabbit retina. *Visual Neuroscience* 11: 549–560. (b, c) Reproduced from Perlman I and Ammermüller J (1994) Receptive-field size of L1 horizontal cells in the turtle retina: Effects of dopamine and background light. *Journal of Neurophysiology* 72: 2786–2795, used with permission from the American Physiological Society.

experiment in which the receptive field of turtle L1-type horizontal cell was assessed using a series of concentric spots of different radii. The photoresponses, elicited by light stimuli of fixed intensity and wavelength (650 nm) but different diameters, are characterized by a gradual reduction as the size of the illuminated retinal area is reduced (Figure 7(b)). This occurs because nonilluminated peripheral horizontal cells shunt the response of the central cell. With full-field illumination, all of the horizontal cells are evenly illuminated and are equipotential, and therefore no current flows between them.

The dimension of the horizontal cell receptive field can be quantified by the diameter of the spot that elicits a response that does not differ from that elicited by a full-field illumination or by the two-dimensional length constant (λ). The latter parameter can be determined by applying a mathematical model to the relationship between response amplitude (normalized to the response to full-field stimulation) and spot radius (Figure 7(c)). The horizontal cell syncytium cannot be described by a simple array of resistors interconnected by coupling resistors. The resistances of the horizontal cell layer, reflecting glutamate-gated conductances, voltage-dependent conductances, leak channels, and gap junctions, constantly change with time due to changes in ambient illumination and in the chemical composition of the extracellular space. Therefore, the size of the horizontal cell receptive field (magnitude of length constant) changes with time and with changes in environmental conditions.

Modulation of Horizontal Cell Physiology and Morphology

The properties of horizontal cells are modulated by a variety of chemicals that are released by retinal cells during changing conditions of illumination. These chemicals, generally referred to as neuromodulators, are believed to alter the modes of information processing in order to adjust retinal function to any new state of adaptation. To date, three candidates have been studied for their effects on vertebrate horizontal cells: dopamine, nitric oxide, and retinoic acid.

Dopamine

Dopamine-containing neurons reside in the amacrine cell layer of the retina in most species. In some species (e.g., fish), dopaminergic cells are interplexiform cells (IPCs), which send long processes to the outer plexiform layer to make synaptic contact with horizontal cells. In other species (e.g., turtle, mammals), the dopaminergic cells are amacrine cells, and therefore dopamine has to reach the horizontal cells by diffusion through the inner nuclear layer.

Dopamine (DA) has several influences on horizontal cells – some are general and some species dependent.

1. The size of the receptive field is reduced, as shown by dye coupling and by electrophysiological determination of receptive field size (Figure 8(a)). Here, the length constant, λ, was calculated for L-type horizontal cells using light stimuli of different intensities under control conditions and in the presence of dopamine. For each light intensity, the receptive field is narrowed (small λ) by dopamine. This effect reflects a reduction of gap junction conductance induced by activation of adenylate cyclase (AC) by dopamine at D1-type receptors (Figure 8(d)).
2. In fish horizontal cells, dopamine induces the formation of spinules in a manner similar to that seen with bright light background illumination.
3. In some species (e.g., fish), the interaction between L-glutamate and its receptors on the horizontal cell membrane is modulated by dopamine so that the amplitude of the photoresponse is reduced to any given light stimulus. This does not happen in the turtle retina (Figure 9(a)).
4. Dopamine changes the balance between rod and cone inputs to horizontal cells in *Xenopus* retina, in order to augment the rod contribution in the dark-adapted state and the cone contribution in the light-adapted state.

Nitric Oxide

Cells containing the enzyme nitric oxide synthase (NOS) have been demonstrated in all retinas studied, indicating potential synthesis of nitric oxide (NO). In all species NOS occurs in amacrine cells, but reports of NOS in photoreceptors, horizontal cells, bipolar cells, and Müller cells have been published. In the turtle retina, NOS was reported in horizontal cell processes at the photoreceptor ribbon synapses, in the inner segments of the photoreceptors, in H2 and H3 cells, and in Müller cells. In the rabbit retina, nicotinamide adenine dinucleotide phosphate (NADPH) diaphorase activity in horizontal cells was shown to depend upon the state of adaptation and the activity of glutamatergic pathways. It is increased in the dark-adapted state or by the activation of AMPA/KA-type receptors. Thus, NO can be synthesized in the distal retina to modulate the physiology of distal neurons including horizontal cells.

NO reduces the size of the horizontal cell receptive field (Figure 8(b)), reflecting direct action of NO upon guanylate cyclase (GC; Figure 8(d)). NO also modulates the horizontal cells' photoresponses. Raising the retinal level of NO by adding

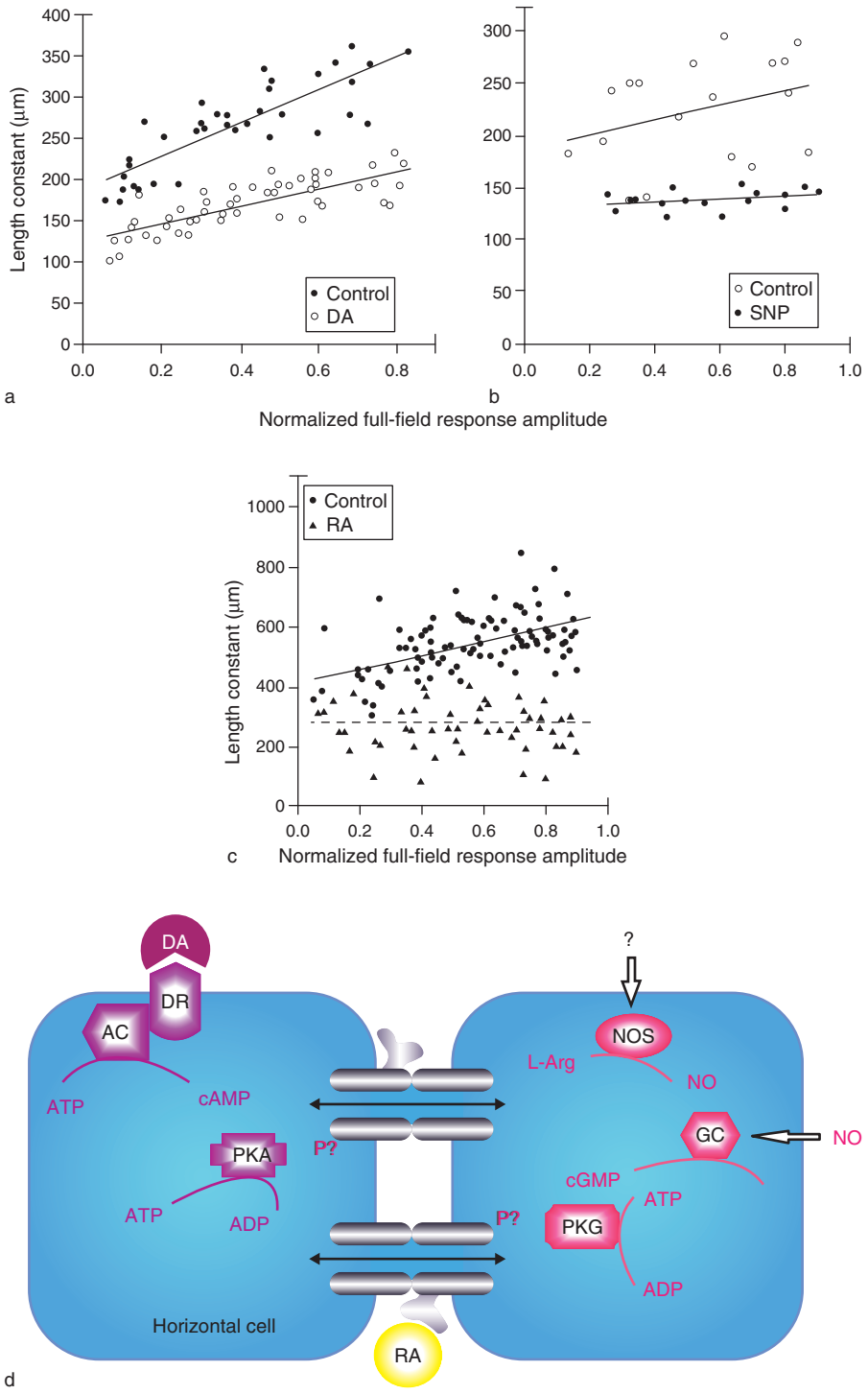

Figure 8 Modulation of horizontal cell receptive field by (a) dopamine (DA), (b) nitric oxide (from donor sodium nitroprusside (SNP)), and (c) retinoic acid (RA). For each drug, the length constant was measured using light stimuli of different intensities under control conditions and during superfusion with the experimental drug. Each drug lowered the dimensions of the horizontal cell respective field, probably by different mechanisms, as illustrated schematically in (d). (d) Dopamine acts on a dopamine receptor (DR) to activate the cyclic adenosine monophosphate (cAMP) pathway. Nitric oxide (NO) diffuses into the cell to activate the cyclic guanosine monophosphate (cGMP) pathway and retinoic acid probably acts directly on receptor sites on the gap junction proteins (AC, adenylate cyclase; PKA, protein kinase A; GC, guanylate cyclase; PKG, protein kinase G). (a) From Perlman I and Ammermüller J (1994) Receptive-field size of L1 horizontal cells in the turtle retina: Effects of dopamine and background light. *Journal of Neurophysiology* 72: 2786–2795. (b) From Pottek M, Schultz K, and Weiler R (1997) Effects of nitric oxide on the horizontal cell network and dopamine release in the carp retina. *Vision Research* 37: 1091–1102. (c) From Pottek M and Weiler R (2000) Light-adaptive effects of retinoic acid on receptive field properties of retinal horizontal cells. *European Journal of Neuroscience* 12: 437–445.

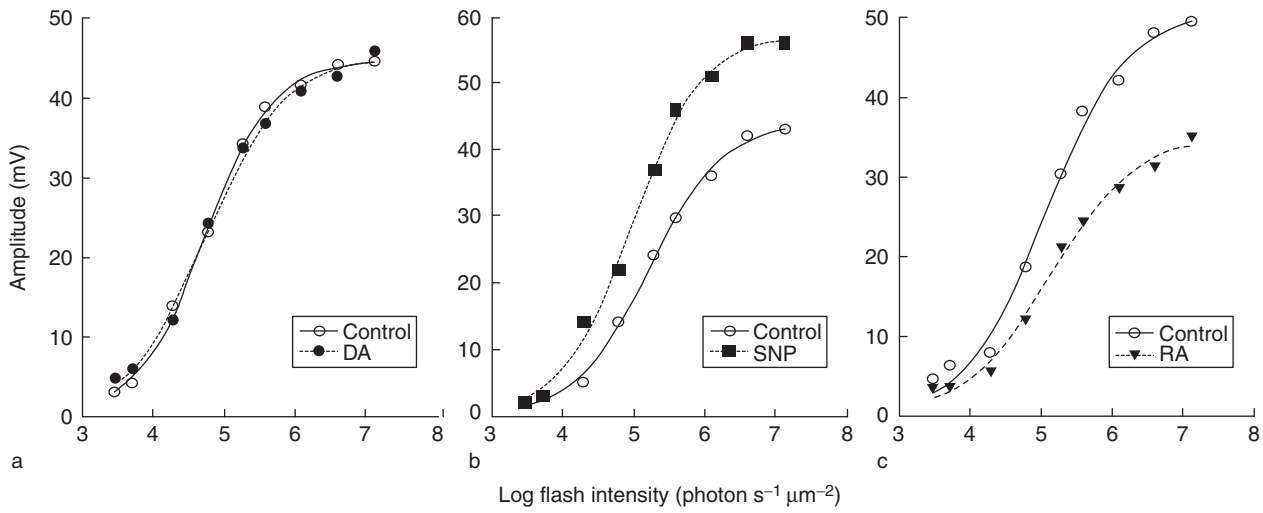

Figure 9 The effects of (a) dopamine (DA), (b) nitric oxide (from donor sodium nitroprusside (SNP)), and (c) retinoic acid (RA) upon the responsiveness to light of L-type horizontal cells in the turtle retina. The photoresponses were elicited in the dark-adapted states with light stimuli of different intensities under control conditions and during superfusion with the experimental drug. The response–intensity relationships were derived and fitted to the generalized hyperbolic function (eqn [1]) (continuous curves). It is clear that dopamine did not affect the photoresponses, that nitric oxide caused response augmentation, and that retinoic acid caused response diminution.

its precursor, L-arginine, or adding an NO donor such as sodium nitroprusside (SNP) leads to response augmentation (Figure 9(b)), while lowering the NO level leads to response diminution. The actions of NO on processing of visual signals in the distal retina of the turtle mimic the effects of background light. In darkness or in low NO, synaptic transmission from cone photoreceptors to L1-horizontal cells is nonlinear, with high gain for small signals, while in the presence of background light or high NO level, synaptic transmission becomes linear and of low gain.

Retinoic Acid

Retinoids are of great importance in the eye, having roles in photopigment cycle and in development. Retinoic acid (RA) is produced in the pigment epithelium as a side effect of the rhodopsin phototransduction process. Retinoic acid induces synaptic plasticity at the terminal dendrites of horizontal cells and promotes the growth of spinules in fish in a pattern similar to that induced by background lights or by dopamine. It exerts a pronounced effect on the electrical coupling between horizontal cells in mammalian and nonmammalian retinas, and reduces the receptive fields of the horizontal cells (Figure 8(c)), probably by acting directly on the gap junctions (Figure 8(d)). Intracellular recordings demonstrate retinoic acid-induced changes in horizontal cells' responsiveness to light. Retinoic acid augments the amplitudes of the horizontal cell photoresponses to full-field stimuli in the fish retina, but reduces them in the turtle retina (Figure 9(c)).

Functional Roles of Horizontal Cells

Spatial Information Processing

The major functional role of horizontal cells in the vertebrate retina is creating spatially opponent receptive field organization in postreceptoral retinal cells (bipolar, amacrine, and ganglion cells). Bipolar cells are the first visual neurons exhibiting spatial organization of their receptive fields. They respond with opposite polarity to light stimuli illuminating the center or the surround portions of their receptive field. The simplest model explaining the spatial properties of bipolar cells includes direct input from photoreceptors contributing to the central response and opponent inputs from horizontal cells forming the surround response. The spatial organization of the bipolar cell's receptive field is transmitted to ganglion cells, and then to higher visual centers in the central nervous system (CNS) for further processing, eventually leading to pattern recognition. In experiments in which horizontal cell membrane potentials are directly polarized by injections of current through an intracellular microelectrode, ganglion cells respond in a manner characteristic for 'surround responses.' This current-evoked response can reflect the negative feedback pathway from horizontal cells to cone photoreceptors and/or direct inputs from horizontal cells to bipolar cells. Evidence for both possibilities has been reported in different species.

Color Information Processing

Horizontal cell feedback probably serves to generate color opponency in bipolar cells and, therefore, in

ganglion cells in species having color vision, and where color opponent pathways exist in the retina. In the retina of cold-blooded vertebrates, color opponency in C-type horizontal cells is most likely generated by excitatory feed-forward and negative feedback inputs between horizontal cells and cone photoreceptors. Direct excitatory inputs from long-, medium-, and short-wavelength-sensitive cones to L-type horizontal cells, from medium- and short-wavelength-sensitive cones to red/green C-type horizontal cells, and from short-wavelength-sensitive cones to yellow/blue C-type horizontal cells determine the hyperpolarizing responses of the horizontal cells. A negative input from L-type horizontal cells onto medium-wavelength-sensitive cones provides the pathway for long-wavelength depolarizing photoresponses in turtle red/green C-type horizontal cells. Negative feedback pathways from both L-type horizontal cells and red/green C-type horizontal cells onto short-wavelength-sensitive cones underlie the depolarizing responses in turtle yellow/blue C-type horizontal cells. This is the best example of how a simple neuronal network between three spectral types of cone photoreceptors and three types of horizontal cells gives rise to color opponency channels. In mammals and primates, color opponency is observed only in bipolar cells or ganglion cells, and the role of the horizontal cells is unclear.

Photoreceptor Modulation under Different Lighting Conditions

The negative feedback pathways from horizontal cells to cone photoreceptors also provide modulation of photoreceptor activity during changing conditions of illumination, thus affecting signal transmission from photoreceptors to bipolar cells and eventually to higher visual centers. In the simple model of horizontal cell feedback onto cone photoreceptors, horizontal cells modulate the transmembrane potential of the cone pedicle, thus affecting the rate of transmitter release. This effect serves, at an early stage of the visual process, to reset the activity of visual neurons during changing conditions of illumination in a manner that will increase sensitivity and will shift the dynamic range of responsiveness to higher levels of illumination. The end result is that the visual system is insensitive to the level of constant illumination but is very sensitive to changes in illumination.

See also: Gap Junctions and Neuronal Oscillations; Photoreceptors: Physiology; Retinal Color Mechanisms; Retinal Development: An Overview; Retinal Ganglion Cells: Anatomy; Retinal Ganglion Cells: Receptive Fields; Retinal Models; Retinal Pharmacology: Inner Retinal Layers; Vertebrate Eyes: Evolution.

Further Reading

Ammermüller J and Kolb H (1996) Functional architecture of the turtle retina. *Progress in Retinal and Eye Research* 15: 393–433.

Bloomfield SA and Miller RF (1982) A physiological and morphological study of the horizontal cell types of the rabbit retina. *Journal of Comparative Neurology* 208: 288–303.

Dacey DM, Lee BB, Stafford DK, et al. (1996) Horizontal cells of the primate retina: Cone specificity without spectral opponency. *Science* 271: 656–659.

Dacheux RF and Raviola E (1990) Physiology of HI horizontal cells in the primate retina. *Proceedings of the Royal Society of London Series B, Biological Sciences* 239: 203–221.

Fahrenfort I, Klooster J, Sjoerdsma T, et al. (2005) The involvement of glutamate-gated channels in negative feedback from horizontal cells to cones. *Progress in Brain Research* 147: 219–229.

Fisher SK, Lewis GP, Linberg KA, et al. (2005) Cellular remodeling in mammalian retina induced by retinal detachment. Available at Webvision: http://www.webvision.med.utah.edu.

Kamermans M and Spekreijse H (1995) Spectral behavior of cone-driven horizontal cells in teleost retina. *Progress in Retinal and Eye Research* 14: 313–360.

Lasansky A (1972) Cell junctions at the outer synaptic layer of the retina. *Investigative Ophthalmology* 11: 265–275.

Lasater EM (1991) Membrane properties of distal retinal neurons. *Progress in Retinal Research* 11: 215–246.

Lasater EM and Dowling JE (1982) Carp horizontal cells in culture respond selectively to L-glutamate and its agonists. *Proceedings of the National Academy of Sciences of the United States of America* 79: 936–940.

Levy H, Twig G, and Perlman I (2004) Nitric oxide modulates the transfer function between cones and horizontal cells during changing conditions of ambient illumination. *European Journal of Neuroscience* 20: 2963–2974.

McMahon DG, Zhang DQ, Ponomareva L, et al. (2001) Synaptic mechanisms of network adaptation in horizontal cell. In: Kolb H, Ripps H, and Wu S (eds.) *Concepts and Challenges in Retinal Biology, a tribute to John E. Dowling*, pp. 419–436. Amsterdam: Elsevier Science.

Mills SK and Massey SC (1994) Distribution and coverage of A-type and B-type horizontal cells stained with Neurobiotin in the rabbit retina. *Visual Neuroscience* 11: 549–560.

Morgan J and Wong R (2005) Development of cell types and synaptic connections in the retina. Available at Webvision: http://www.webvision.med.utah.edu.

Perlman I and Ammermüller J (1994) Receptive-field size of L1 horizontal cells in the turtle retina: Effects of dopamine and background light. *Journal of Neurophysiology* 72: 2786–2795.

Perlman I, Kolb H, and Nelson R (2003) Anatomy, circuitry, and physiology of vertebrate horizontal cells. In: Chalupa LE and Werner JS (eds.) *The Visual Neurosciences*, pp. 369–394. Cambridge, MA: MIT Press.

Perlman I, Kolb H, and Nelson R (2005) S-potentials and Horizontal cells. Webvision. http://www.webvision.med.utah.edu.

Perlman I and Normann RA (1998) Light adaptation and sensitivity controlling mechanisms in vertebrate photoreceptors. *Progress in Retinal and Eye Research* 17: 523–563.

Pottek M, Schultz K, and Weiler R (1997) Effects of nitric oxide on the horizontal cell network and dopamine release in the carp retina. *Vision Research* 37: 1091–1102.

Pottek M and Weiler R (2000) Light-adaptive effects of retinoic acid on receptive field properties of retinal horizontal cells. *European Journal of Neuroscience* 12: 437–445.

Twig G, Levy H, and Perlman I (2003) Color opponency in horizontal cells of the vertebrate retina. *Progress in Retinal and Eye Research* 22: 31–68.

Retinal Models

G T Kenyon, Los Alamos National Laboratory, Los Alamos, NM, USA

Published by Elsevier Ltd.

Introduction

Computational models of the retina have targeted a range of biological scales, from mathematical formulations of the phototransduction cascade to biophysical models of reconstructed ganglion cells and large-scale simulations of heterogeneous networks. As with computational studies of other central nervous system (CNS) structures, retinal models have been used to investigate emergent phenomena in systems of interacting elements in order to better understand the contribution of each component to the collective behavior. Quantitative models have also been used to describe robust experimental phenomena using elegant mathematical relationships characterized by a small number of parameters. The retina has also provided a test bed for the application of information theoretic techniques to the analysis of simultaneously recorded spike trains. Finally, the retina represents an extraordinary example of biological adaptation over a vast range of operating conditions. The retina may be the first CNS structure to be completely described by biophysically detailed computational models able to account for functionally significant aspects of the system. Potential applications of such models include the computer-aided design of retinal prosthetics, the improved interpretation of the electroretinogram (ERG), and even the development of next-generation computer vision systems.

Unique Advantages of Retinal Models

As a target system for the development of computational models, the retina offers a number of distinct advantages relative to other CNS structures. Sections of intact retina can be removed and studied *in vitro*, permitting external stimuli to be controlled very precisely while recording simultaneously from multiple retinal output neurons, or ganglion cells, under effectively normal physiological conditions. With the exception of birds, the vertebrate retina receives only diffuse modulatory feedback from higher brain areas and thus to a good approximation can be studied as a stand-alone circuit. The input/output function of the retina can therefore be characterized to a much greater resolution than is possible for other CNS structures and with a minimum of confounding factors related to experimental manipulations or state of arousal. Over short distances, retinal circuitry is roughly translationally invariant and coverage of the visual space – with the exception of certain direction-selective neurons – is approximately isotropic. The connections to and from a representative retinal neuron of a given type can thus serve as a template for all such cells of the same type (to within biological variation), thereby greatly reducing the number of free model parameters. In addition, the gross anatomical structure of the vertebrate retina is largely conserved, allowing model developers to use general circuit motifs where data from a particular species may be lacking. Multiple independent evolutions of retinal-like structures may even allow some aspects of model development to be guided by universal visual processing principles. Despite these distinct advantages from the standpoint of model construction, the retina is nonetheless a highly complex structure for which a complete specification of the underlying anatomy and physiology is not available. However, emerging technologies for automating the reconstruction of neural processes from thin electron microscope sections may soon unveil retinal circuitry in unprecedented detail. Moreover, once a nearly complete specification of the retinal input/output function has been obtained – a prospect that seems increasingly plausible for several vertebrate species – it may be possible for model developers to uniquely infer the remaining unknown physiological parameters necessary to fully account for the retinal code.

Differences with Respect to Classical Models

Probably no structure in the CNS is purely classical in the sense that all of its dendrites and axons have strictly input or output functions and all of its synapses are activated solely by action potential-mediated transmitter release. The retina, however, deviates substantially from such classical neural circuits and includes several specializations that must be taken into account in the development of biophysically detailed models. Most prominently, photoreceptors possess unique transduction mechanisms and light/dark adaptation properties that drive the light responses of all retinal neurons (except for a class of photosensitive ganglion cells not discussed here). Fortunately, both the phototransduction cascade and Ca^{2+}-mediated negative feedback are relatively well understood, and detailed models of the underlying biochemical processes are actively being developed. A relatively complete mathematical characterization of the photocurrents in rods and cones is thus increasingly feasible, potentially

allowing photoreceptor responses to light flashes of varying duration and intensity and under a range of adaptation conditions to be modeled with a high degree of accuracy (Figure 1).

Retinal circuitry also involves several unconventional, or graded, synaptic contacts that do not require the generation of action potentials to mediate functional interactions, including feed-forward excitatory ribbon synapses at photoreceptor and bipolar cell terminals, dendrodendritic and dendroaxonic synapses – particularly between amacrine cells and their feedback projections onto bipolar cells – extensive gap junction coupling, as well as evidence for nonvesicular neurotransmitter release from some retinal neuron types. Indeed, conventional synapses that require action potentials for the release of chemical neurotransmitters are more the exception in the retina than the rule. Vesicular release at graded synapses can be modeled as a stochastic process whose rate depends on the presynaptic membrane potential or, if greater accuracy is required, on the concentration of intracellular calcium. Additional terms are necessary to account for reductions in gain during sustained light responses and for saturation at high intensities. Nonvesicular release, which uses voltage-dependent transporter mechanisms, can be approximated by empirically determined functions giving the change in the postsynaptic conductance produced by shifts in the presynaptic membrane potential. Mathematically, gap junctions can be effectively described as simple ohmic resistances linking coupled cells. Unlike conventional synapses, however, graded synapses and gap junctions must be accounted for – and the associated currents updated – on every time step, not just when an action potential is generated, which may significantly increase the computational load. Retinal interneurons also release a number of neuromodulators, such as dopamine, which in turn affect retinal connectivity and excitability, creating feedback loops that can operate over much longer timescales than typical light responses.

Survey of Retinal Models

Quantitative Models

Linear systems In the late 1950s, Hartline and Ratliff extended the application of linear models to the analysis of delayed, lateral inhibition between individual ommatidia in the compound eye of *Limulus*, accounting for Mach bands (enhancement of light/dark boundaries at a high contrast edge) and for amplified responses to a sinusoidally varying light intensity, the latter occurring near a resonance

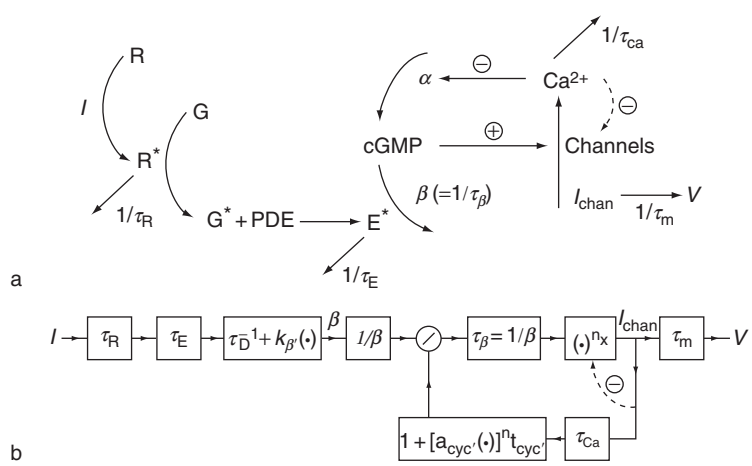

Figure 1 Reaction scheme and model for the generation of photocurrent in human cones. (a) Photopigment R is converted by light I into activated photopigment R*, which is inactivated with rate constant $1/\tau_R$. R* activates a G-protein (transducin), which subsequently forms an activated complex with PDE. The activated transducin–PDE complex is inactivated with rate constant $1/\tau_E$; it hydrolyses cGMP with rate constant β. Cyclic nucleotide-gated (CNG) channels in the membrane, opened under control of cGMP, admit an ionic current I_{chan}, part of which is carried by Ca^{2+} ions. Ca^{2+}, which is removed with rate constant $1/\tau_{Ca}$, regulates the production rate α of cGMP via guanylyl cyclase (GC). Ca^{2+} also exerts a direct inhibitory influence on the CNG channels. Part of the ionic current I_{chan} charges the membrane of the cone outer segment (with membrane time constant τ_m), yielding the output photovoltage, V. (b) The boxes with $\tau_{(\cdot)}$ represent first-order low-pass filters. The rate constant of cGMP hydrolysis consists of a dark rate constant, $1/\tau_D$, plus a light-dependent component with scaling determined by the gain factor k_β. The nonlinear differential equation describing cGMP hydrolysis can be understood as a static nonlinearity $1/\beta$ followed by a low-pass filter with time constant $\tau_\beta = 1/\beta$. The calcium feedback loop is governed by n_X, the apparent Hill coefficient of cGMP activation of the CNG channels (including the effect of a local calcium feedback, indicated by the dashed line), and by n_{cyc}, the Hill coefficient of GC deactivation; a_{cyc} is a scaling constant. Reproduced from Van Hateren JH and Lamb TD (2006) The photocurrent response of human cones is fast and monophasic. *BMC Neuroscience* 7: 34, with permission.

frequency at which the recurrent inhibition lags the excitatory center response by approximately one-half of a stimulus cycle. In 1965, Rodieck used linear systems theory, exploiting the principles of superposition and separability, to predict the responses of cat ganglion cells to arbitrary stimuli (Figure 2). The center-surround receptive field organization was modeled as a difference of Gaussians and the time dependence of the step response treated as a separate multiplicative factor with both transient and sustained components. Generalizing the Rodieck model so that the surround could be activated with a slower time course than the center, Enroth-Cugell and colleagues were able to account for the responses of cat ganglion cells to drifting sinusoidal gratings across a range of spatial and temporal frequencies. Such models continue to be widely employed as quantitative descriptions of the linear receptive field properties of retinal neurons.

Nonlinear effects In the late 1970s, computational models were developed to account for nonlinear aspects of ganglion cell responses. In a series of papers, Shapley and Victor employed a cascade of low- and band-pass temporal filters, the latter representing negative feedback, to quantify the increased selectivity of cat ganglion cells to rapidly varying stimuli as a function of luminance contrast – a phenomenon they termed contrast gain control. Such models were able to express the primary action of contrast gain control in terms of a single parameter characterizing the product of the gain and natural frequency of the negative feedback loop. Descriptions based on a cascade of temporal filters have become a standard technique for quantifying nonlinear changes in receptive field dynamics as a function of center and/or surround stimulation.

Linear–nonlinear models A linear–nonlinear model refers to a technique most commonly used to account for the responses of retinal ganglion cells to naturalistic stimuli. Unlike ganglion cell responses to very simple inputs, such as stationary spots or drifting gratings, the spike trains evoked by video clips of natural scenes are often highly reproducible from trial to trial, with temporal spreads as narrow as a few milliseconds. Linear–nonlinear models use reverse correlation methods to reconstruct spatiotemporal receptive fields via spike-triggered averaging, with the firing activity in response to a random checkerboard used to build up the average stimulus at each time point preceding a spike. The reconstructed spatiotemporal receptive field is then treated as a linear kernel that can be convolved with an arbitrary input pattern to produce a subthreshold generator potential. To account for the firing threshold, the generator potential is passed through an expansive nonlinearity. The output of the nonlinear filter is used to modulate the instantaneous rate of a stochastic

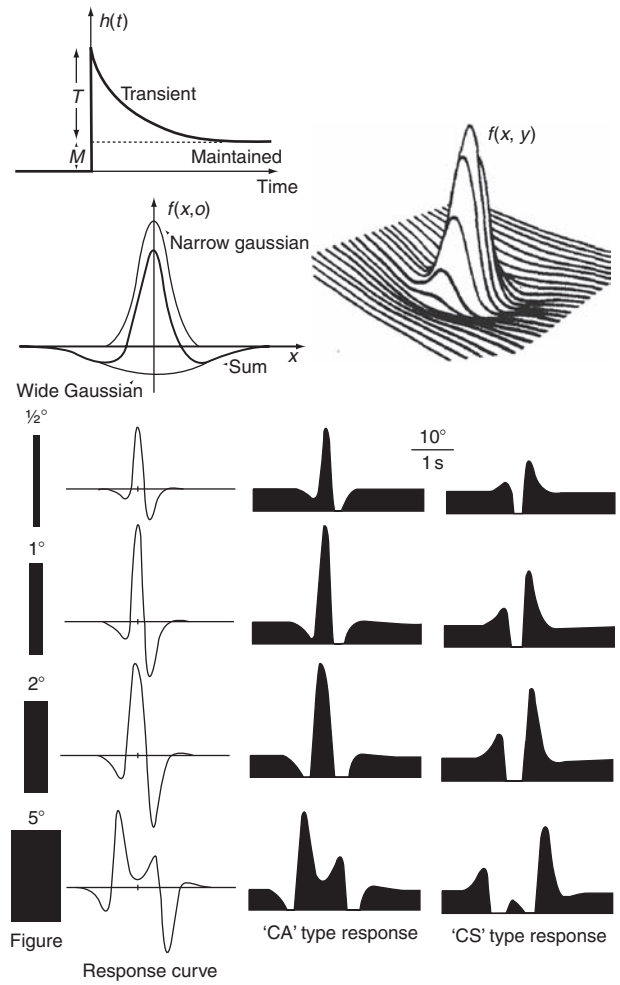

Figure 2 Analytic representation of the receptive field of a cat ganglion cell. Assuming separability, the response to a step change in illumination over a small retinal patch can be expressed as the product of separate spatial and temporal factors: $f(x, y, t) = f(x, y) \cdot h(t)$. The temporal factor is characterized by a maintained change in firing rate of amplitude M and a transient, exponentially decaying component of amplitude T. The spatial factor can be approximated as the sum of two Gaussian functions, one of which is wider and reversed in sign. The total response to any uniform stimulus is given by the integral of $f(x, y, t)$ over the bounding contour. For a uniform bar of arbitrary width moving with constant velocity perpendicular to either axis, the solution can be written analytically in terms of error functions. The response curve, which estimates the change in firing rate, is triphasic. The predicted responses of both center-activated (CA) and center-suppressed (CS) cells are shown after adding a nonzero baseline and truncating negative firing rates to zero. For these, the width of the center Gaussian = 0.59°, the ratio of the total center to surround strengths = 0.8, the ratio of surround to center widths = 3, $T/M = 4$, exponential decay constant = 40 ms, and bar velocity = 10° s^{-1}. Reproduced from Rodieck RW (1965) Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research* 5: 583–601, with permission from Elsevier.

Poisson or renewal processes, yielding a predicted output spike train in response to any given input. Model parameters, particularly those characterizing the threshold nonlinearity, are adjusted to produce the best fit to experimental data. For testing purposes, a novel naturalistic stimulus – one not used for optimizing model parameters – is presented multiple times, allowing the experimentally measured peristimulus time histogram to be compared with the prediction of the linear–nonlinear model. Such linear–nonlinear models have been used to quantify changes in the intrinsic gain of retinal ganglion cells in response to contrast increments or decrements (contrast adaptation) and can be generalized to incorporate functional interactions based on pairwise correlations between ganglion cell spike trains.

Biophysical Models

Voltage-gated channels Computational models of single retinal neurons have been used to synthesize diverse sets of experimental data, providing insight into how different mixtures and distributions of voltage-gated ion channels, in combination with realistic dendritic and axonal morphology, act collectively to shape measured input/output properties. In the salamander retina, Miller and colleagues used multicompartmental reconstructions to explore how the distribution of voltage-gated conductances in ganglion cell dendrites modulates somatic responses to injected currents. In the absence of voltage-gated currents, such models predict that capacitive charge stored in the dendritic membrane would trigger somatic spikes much more rapidly than is actually observed. Computational models of nonspiking retinal neurons further suggest that voltage-gated currents may help to maintain the membrane potential within an optimal dynamic range for modulating graded release. There can be little doubt that biophysically detailed models will continue to reveal important functional roles for the variety of voltage-gated currents present in all retinal neuron types.

Nonlinear subunits In the cat retina, α ganglion cells contain nonlinear subunits that respond asymmetrically to step changes in light intensity – a feature that increases the overall sensitivity of these neurons to moving textures, for example, by ensuring that responses to a leading edge are not canceled by an opposite response to a trailing edge. Computational models have shown that the rectification of transient inputs from both cones and bipolar cells can quantitatively account for the formation of nonlinear subunits. Because nonlinear subunits are present in a variety of mammalian retina, although not in the majority of primate ganglion cells, computational models will continue to play an important role in determining to what extent a common mechanism for nonlinear subunits has been conserved across vertebrate species.

Direction selectivity Computational models have been used extensively to investigate the synaptic mechanisms underlying the direction-selective responses of certain ganglion cells. A direct consequence of the principle of superposition is that the total integrated response of any linear system must be independent of the direction of stimulus motion, mandating that direction-selective responses must involve some essential nonlinearity. Correlation-based motion detectors were used by Reichardt in the late 1950s to account for direction-selective responses in insect visual systems, and biologically plausible circuits that could explain the main properties of directional-selective ganglion cells in the rabbit retina were proposed by Barlow and Levick in 1965. Using computational models, in 1983, Koch and Poggio investigated the role of nonlinear shunting inhibition on the dendrites of direction-selective ganglion cells, exploiting the fact that an excitatory synaptic input would be more strongly attenuated by a current shunt placed more proximally relative to the site of action potential initiation. Subsequently, in 1991, Borg-Graham and Grzywacz developed a computational model to examine the generation of direction-selective responses at the distal ends of starburst amacrine cell dendrites, which extend radially from the cell body and provide a natural axis for differentiating null and preferred directions. Computational models continue to be used to test theories of direction selectivity at both pre- and postsynaptic sites.

Network Models

Ultimately, the most interesting consequences of retinal processing are likely to arise from network-level interactions. Carver-Mead pioneered the use of analog VLSI technology (aVLSI) to physically emulate a retina-like circuit based on feedback inhibition of photosensitive elements and feed-forward inhibition of second-order neurons reminiscent of horizontal cell interactions with cones and bipolar cells. Similar to recurrent lateral inhibition in *Limulus*, the silicon retina exhibited both enhanced spatial and temporal contrast and thus was able to reproduce an essential aspect of retinal processing. By the late 1980s, computational resources had advanced to the point that network-level models were feasible. Smith and colleagues constructed some of the first biophysically detailed computer simulations of large-scale retinal networks, including models of the bipolar cell input to a reconstructed cat α ganglion cell and of the

inhibitory feedback from horizontal cells onto cone terminals. Subsequently, Werbin and coworkers developed a comprehensive circuit model based on the salamander retina that included contributions from all of the major cell classes. Implemented on a specialized parallel architecture called the PIPE, individual neurons were implemented as nonlinear filters that summed weighted activity across locally defined neighborhoods consisting of distinct presynaptic cell populations. The integrated synaptic input was then combined with a blur function to account for lateral electrical coupling. The model included explicit descriptions of various retinal interneurons, including sustained and transient bipolar cells of both ON and OFF types and of narrow and wide-field amacrine cells, providing one of the first experimentally based simulations of retinal processing stretching from cones to ganglion cell membrane potentials. Boahen and colleagues have since developed a functionally complete model of the mammalian retina configured for real-time implementation in aVLSI that accounts for contrast gain control and the spatial and temporal frequency tuning curves measured from cat ganglion cells. Models of functionally complete retinal networks are fast becoming a reality, with the incorporation of additional biophysical detail likely to proceed as the underlying anatomy and physiology become better understood.

Retinal Code

On the order of 7 million cones and more than 100 million rods converge down to 1 million fibers in the optic nerve, including both ON and OFF types. The human retina thus contains far fewer total output pixels than do modern digital cameras. Trial-to-trial variability in the responses of individual retinal ganglion cells further reduces the precision of the local visual signals conveyed by each optic nerve fiber, in part resulting from evolutionary pressure to reduce the number of impulses transmitted to the brain and thus the total metabolic cost. The resulting information bottleneck suggests that far from serving as a passive relay, the retina employs a sophisticated spatiotemporal code in order to support the perceived acuity of visual experience. Understanding the nature and origins of the retinal code, and how it maintains operational fidelity over a range of ambient light intensities covering nearly 12 orders of magnitude, represents an exciting goal of computational modeling studies.

Ensemble Activity

Population codes As the ability to record simultaneously from multiple retinal neurons becomes more widely available, computational models will be increasingly used to investigate how stimulus properties are represented by the ensemble of evoked activity. Chichilnisky and coworkers developed techniques for recording simultaneously from a nearly complete mosaic of parasol ganglion cells in the primate retina. To analyze such enormous data sets, they employed computational models to quantify how precisely a population code could be used to estimate the velocity of a moving bar. Spike trains from ganglion cells at different retinal locations were time shifted by an amount proportional to the delay in their peak activation by the moving stimulus. Estimates of bar velocity based on a population code greatly exceeded that available from any given pair of ganglion cells, demonstrating the importance of averaging over ensemble activity in order to obtain precise estimates of stimulus parameters. In the future, computational models are likely to remain essential for interpreting population-coded signals.

Correlations The spike trains generated by pairs of retinal ganglion cells are often correlated, especially when their receptive fields partially overlap. Pairwise correlations are categorized into three functional classes: (1) correlations due to stimulus coordination, which are all but inevitable when both ganglion cells respond to the same stimuli; (2) noise correlations, defined as the residual correlations remaining after the effects due to stimulus coordination have been subtracted; and (3) background correlations, which persist regardless of the stimulus condition. Of particular interest is whether the noise correlations are synergistic, providing additional information about the stimulus beyond that conveyed by the cells individually, or redundant, such that the total information transmitted is less than would have been the case if the two spike trains were statistically independent. From a computational standpoint, such issues are commonly addressed using information theory. A small time bin, usually on the order of 1 ms, is used to digitize the recorded spike train segments, typically as binary patterns in which a '1' and a '0' denote the presence and absence of an event in the corresponding bin, respectively. The information conveyed by a suitably digitized pattern of spikes in response to a particular stimulus, regardless of whether the pattern includes one or more cells, is proportional to the logarithm of the conditional probability for that pattern to occur (measured as a fraction of all patterns observed in response to the same stimulus), divided by the logarithm of the *a priori* probability for the same pattern to occur regardless of what stimulus is present. By averaging over all stimuli and over all observed patterns, a measure of the mutual information, which

quantifies how much information the digitized spike train transmits about a given set of input stimuli, is obtained. Elegant techniques have been developed for applying information theory to spike train data, along with strategies for handling finite data artifacts, for eliminating stimulus coordination effects, and for extrapolating results to the limit of infinite temporal precision and/or infinite record lengths.

Nirenberg, Latham, and coworkers used information theory to analyze noise correlations between pairs of ganglion cells in the mouse retina responding to video clips of natural scenes. Formally, they used the Kullback–Leibler distance to estimate the information lost, on average, when pairwise correlations were ignored. Under the conditions of their protocol, the information conveyed by correlations was determined to be less than 10% of the total, leading to the conclusion that retinal ganglion cells are largely independent encoders. On the other hand, the existence of high-frequency oscillatory potentials in the ERGs of humans and other primates, as well as direct measurements of oscillatory responses between 60 and 120 Hz in a variety of vertebrate species, suggests a potential mechanism for coordinating the activity of retinal ganglion cells over long distances with millisecond precision. The main properties of coherent, high-frequency oscillations were reproduced in a computational model developed by Kenyon, Marshak, and coworkers, which employed long-range inhibitory feedback from a class of axon-bearing amacrine cells that were electrically coupled to neighboring ganglion cells. Berry, Bialek, and colleagues used Ising models to estimate the consequences of realistic pairwise correlations among large numbers of retinal ganglion cells. They reported that even weak pairwise interactions can have a profound collective effect on the overall state of the network, analogous to the spontaneous alignment of local domains in a ferromagnetic material below the Currie temperature. Although the nature and origins of the retinal code remain to be elucidated, computational models will provide an indispensable complement to experimental investigations.

Models of Light/Dark Adaptation

Quantitative Models

The retina evolved to maintain contrast sensitivity across visual environments ranging from diffuse starlight to direct sunlight. Under laboratory conditions, responses can be recorded to test flashes superimposed on steady backgrounds whose combined luminance spans nearly 12 orders of magnitude. Quantitatively, many essential aspects of light adaptation are described by the Naka–Rushton equation, which in simplified form can be written as follows:

$$V/V_{\mathrm{T}} = I/(I+\sigma)$$

where $V = V_{\mathrm{T}} + V_{\mathrm{B}}$, $I = I_{\mathrm{T}} + I_{\mathrm{B}}$, such that V_{T} is the response, measured relative to a stable baseline V_{B}, to a test flash of intensity I_{T}, presented on a mean background of luminance I_{B}, and the maximum response is given by V_{m}. When the absolute intensity of the test flash, either positive or negative, is much less than the half saturation value, σ, the Naka–Rushton equation predicts that responses will be linearly proportional to the stimulus amplitude, $V_{\mathrm{T}} \sim I_{\mathrm{T}}$, or

$$V_{\mathrm{T}}/V_{\mathrm{m}} \approx I_{\mathrm{T}}/(I_{\mathrm{B}}+\sigma)$$

For very large increments in light intensity, however, the Naka–Rushton equation predicts that responses will saturate at $V \approx V_{\mathrm{m}}$, or $V_{\mathrm{T}} \approx V_{\mathrm{m}} - V_{\mathrm{B}}$.

The Naka–Rushton equation can be written more generally as

$$V/V_{\mathrm{m}} = I^n/(I^n+\sigma^n)$$

with the Hill coefficient, n, typically falling between 0.7 and 1.0. Within these limits, the generalized Naka–Rushton equation predicts that at a fixed background luminance, the dynamic range of retinal light responses, including both positive and negative flashes, covers approximately four orders of magnitude, corresponding to changes in contrast of $\pm 100\%$. The dynamic range of retinal light responses is thus consistent with the range of contrasts typically encountered in natural scenes. However, four orders of magnitude is not nearly enough to account for the shifts in mean background luminance produced, for example, by the transition from moonlight to sunlight. Adaptation mechanisms are thus required to maintain contrast sensitivity across a wide range of background light levels.

Within the framework of the Naka–Rushton equation, adaptation is modeled by allowing σ to vary as a function of the mean background intensity, with the average luminance taken over time intervals ranging from fractions of a second to tens of minutes, depending on the underlying physiological process. Generally, light adaptation is faster than dark adaptation since the latter may require additional time for bleached rod and/or cone photoreceptors to recover. In practice, the half saturation value measured under light-adapted conditions, σ_{L}, can most easily be evaluated in terms of the experimentally measurable quantities, $V_{(+)}$ and $V_{(-)}$, representing the maximum positive and negative responses to light increments and decrements, respectively, with the total response range given by $V_{\mathrm{m}} = V_{(+)} + |V_{(-)}|$. Assuming that the generalized Naka–Rushton equation remains valid,

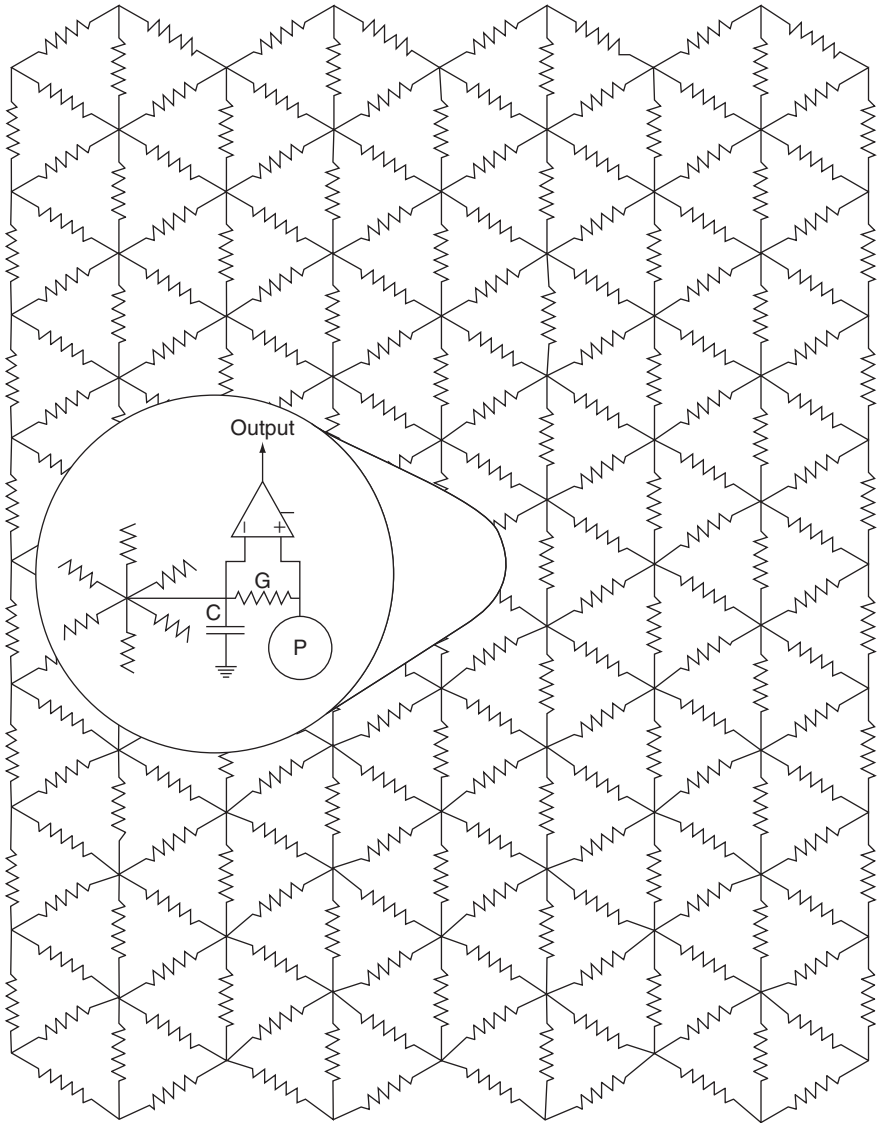

Figure 3 Silicon retina. Each pixel (vertex) represents a triad synapse (circular window) in the outer vertebrate retina, consisting of a conductance (G) that links the photoreceptor output (P) to the resistive network, a parasitic capacitance (C) arising from the interface with the silicon substrate and that provides temporal integration analogous to biological membranes, and an amplifier that takes the difference between P and the local voltage on the resistive network, implemented as a hexagonal array. All resistors are made of transistors, allowing the effective resistance to be controlled by an external input and which forces the current–voltage relation to saturate outside the linear operating range, thus limiting the potential damage from failed circuit components. The primary phototransducer is a vertical bipolar transistor whose output is logarithmic over four or five orders of magnitude, covering a similar range of light intensities to vertebrate cones. The output of the silicon retina, in analogy to bipolar cells, is sensitive to contrast, or the difference between the local receptor output and the space–time average of the surrounding elements. Reproduced from Mead CA and Mahowald MA (1988) A silicon model of early visual processing. *Neural Networks* 1: 91–97, with permission from Elsevier.

it is straightforward to solve for the value of σ_L that reproduces the measured responses range for any given mean luminance:

$$\sigma_L = I_B[\gamma/(1-\gamma)]^{1/n}$$

where $\gamma = V_{(+)}/V_m$ is the range compression factor, which equals 1 in the fully dark-adapted state and approaches 0 as light responses saturate at sufficiently large background levels.

The previous expression predicts that as a result of light adaptation, the half saturation value remains proportional to the mean background intensity, $\sigma_L \sim I_B$. For small test flashes, it follows that $V/V_m \sim I_T/I_B$, and thus that the responses to weak stimuli will be proportional to contrast. This property of light adaptation, which ensures that a 1% increment in light intensity always produces the same fractional responses regardless of the background luminance, is known as Webber's law.

The generalized Naka–Rushton equation has also been used to predict the results of threshold versus intensity (TVI) experiments. As a function of the mean background luminance, a TVI curve gives the amplitude of a test flash that is just sufficient to produce a given response, which may be measured either electrophysiologically or psychophysically, with the latter corresponding to a behavioral criteria specifying the fraction of trails on which a test flash can be successfully detected against an adapting background.

Defining ΔI_L as the minimum flash intensity required to produce a criterion response, δ, at a mean background luminance level, I_B, relative to the flash intensity required to produce the same criterion response in the dark-adapted state, ΔI_D, the generalized Naka–Rushton equation predicts that

$$\Delta I_L/\Delta I_D = I_B(1/\gamma)[1/(1-\gamma)]^{1/n}(\delta/V_B)^{n-1/n}$$

where it is assumed that $\delta \ll V_m$, V_B. As γ goes to zero as a function of increasing background luminance, the associate TVI curves diverge as $1/\gamma$, a result that is independent of the value of the Hill coefficient, n.

Remarkably, the generalized Naka–Rushton equation can accurately describe the responses of retinal neurons to test flashes presented against steady backgrounds as well the changes in both electrophysiological and psychophysical thresholds over essentially the entire range of natural visual stimuli, despite the enormous variety of biochemical and physiological processes involved. An example of the application of quantitative models to the analysis of light/dark adaptation in primate cones is presented in Figures 4 and 5.

Biophysical Models

A number of existing biophysical retinal models exhibit some form of light/dark adaptation, encompassing a wide range of underlying mechanisms. Detailed models of the phototransduction process (Figure 1) have successfully accounted for light/dark adaptation in both rod and cone photoreceptors across the entire span of naturally occurring light levels. Because light flashes produce hyperpolarizing responses by turning off a persistent depolarizing dark current, the saturation of photoreceptor responses arises naturally from the fact that the fraction of light-sensitive channels that can remain open, regardless of the flash intensity, cannot fall below zero. Adaptation in such models is governed primarily by a calcium-mediated negative feedback loop, which acts to reopen the same light-sensitive ion channels, thereby desaturating the response to large step changes in background light levels and thus partially restoring overall sensitivity. At the network level, following a step change in light intensity, both center-surround antagonism (Figure 2) and inhibitory feedback (Figure 3) contribute to light/dark adaptation by restoring the membrane potentials of retinal neurons nearer to their baseline resting levels. Electrophysiological experiments indicate that under both scotopic and photopic conditions, light/dark adaptation involves a combination of cellular processes intrinsic to individual photoreceptors as well as network-level mechanisms in both the inner and the outer retina. Typically, network-mediated mechanisms are more important when the response of individual photoreceptors is weak and thus relatively noisy, conditions under which signals from many rods and

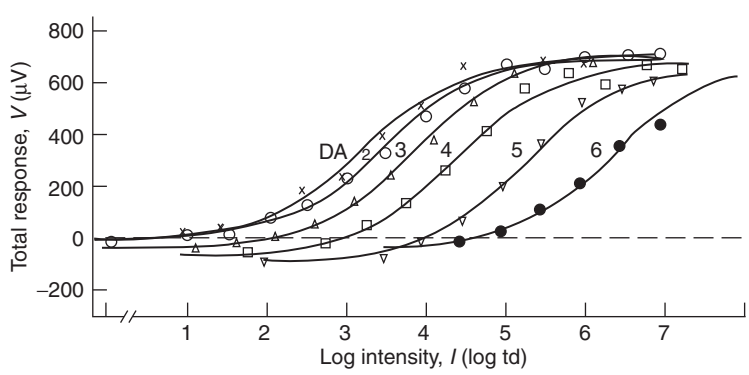

Figure 4 Quantitative model of light/dark adaptation in primate cones. Extracellular potential drop measured along cone outer segments in the primate fovea plotted as a function of flash intensity. The adapting background luminance is indicated by the number to the left of each curve (log td). The small horizontal lines indicate the background response level, V_B, at each background luminance. Solid curves are solutions to the generalized Naka–Rushton equation with $n = 0.74$, with the half-saturation value evaluated from the maximal positive and negative responses, measured relative to V_B. DA, dark adapted. Reproduced from Valeton JM and Van Norren D (1983) Light adaptation of primate cones: An analysis based on extracellular data. *Vision Research* 23: 1539–1547, with permission from Elsevier.

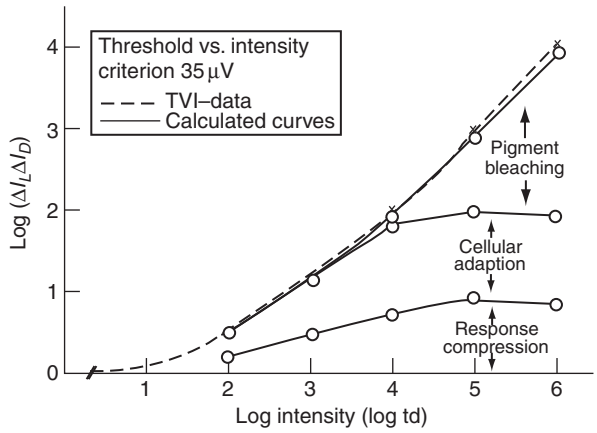

Figure 5 Threshold versus intensity (TVI) of primate cones. Thresholds at each background (ΔI_L) are measured relative to the dark-adapted threshold (ΔI_D). Data points (crosses) were obtained by determining for each background the amount of light necessary to elicit a criterion response ($\delta = 35\,\mu V$). The light-adapted half-saturation level, σ_L, was expressed as a product of three factors – σ_D, σ_{photo}, and σ_{bleach} – corresponding to the dark-adapted value, σ_D, multiplied by a factor due to adaptation in the phototransduction cascade, σ_{photo}, and a separate factor due to pigment bleaching, σ_{bleach}. Both σ_D and σ_{bleach} can be measured independently, allowing the various contributions to the light-adapted threshold to be identified. Reproduced from Valeton JM and Van Norren D (1983) Light adaptation of primate cones: An analysis based on extracellular data. *Vision Research* 23: 1539–1547, with permission from Elsevier.

cones are more likely to be pooled, whereas mechanisms intrinsic to the photoreceptors become more prominent as light levels increase.

See also: Photoreceptor Adaptation; Photoreceptors: Physiology; Phototransduction; Population Codes: Theoretic Aspects; Retinal Amacrine Cells; Retinal Bipolar Cells; Retinal Ganglion Cells: Anatomy; Retinal Ganglion Cells: Receptive Fields; Retinal Horizontal Cells; Vertebrate Eyes: Evolution.

Further Reading

Borg-Graham LJ and Grzywacz NM (1992) A model of the directional selectivity circuit in retina: Transformations by neurons singly and in concert. In: McKenna T, Davis J, and Zornetzer SF (eds.) *Single Neuron Computation*. New York: Academic Press.

Fohlmeister JF and Miller RF (1997) Mechanisms by which cell geometry controls repetitive impulse firing in retinal ganglion cells. *Journal of Neurophysiology* 78: 1948–1964.

Frechette ES, Sher A, Grivich MI, Petrusca D, Litke AM, and Chichilnisky EJ (2005) Fidelity of the ensemble code for visual motion in primate retina. *Journal of Neurophysiology* 94: 119–135.

Freed MA, Smith RG, and Sterling P (1992) Computational model of the on-alpha ganglion cell receptive field based on bipolar cell circuitry. *Proceedings of the National Academy of Sciences of the United States of America* 89: 236–240.

Frishman LJ, Freeman AW, Troy JB, Schweitzer-Tong DE, and Enroth-Cugell C (1987) Spatiotemporal frequency responses of cat retinal ganglion cells. *Journal of General Physiology* 89: 599–628.

Keat J, Reinagel P, Reid RC, and Meister M (2001) Predicting every spike: A model for the responses of visual neurons. *Neuron* 30: 803–817.

Kenyon GT, Moore B, Jeffs J, et al. (2003) A model of high-frequency oscillatory potentials in retinal ganglion cells. *Visual Neuroscience* 20: 465–480.

Koch C, Poggio T, and Torre V (1983) Nonlinear interactions in a dendritic tree: Localization, timing, and role in information processing. *Proceedings of the National Academy of Sciences of the United States of America* 80: 2799–2802.

Mead CA and Mahowald MA (1988) A silicon model of early visual processing. *Neural Networks* 1: 91–97.

Nirenberg S, Carcieri SM, Jacobs AL, and Latham PE (2001) Retinal ganglion cells act largely as independent encoders. *Nature* 411: 698–701.

Ratliff F and Hartline HK (1959) The responses of *Limulus* optic nerve fibers to patterns of illumination on the receptor mosaic. *Journal of General Physiology* 42: 1241–1255.

Rodieck RW (1965) Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research* 5: 583–601.

Schneidman E, Berry MJ 2nd, Segev R, and Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440: 1007–1012.

Shapley RM and Victor JD (1981) How the contrast gain control modifies the frequency responses of cat retinal ganglion cells. *Journal of Physiology* 318: 161–179.

Smith RG (1995) Simulation of an anatomically defined local circuit: The cone-horizontal cell network in cat retina. *Visual Neuroscience* 12: 545–561.

Teeters J, Jacobs A, and Werblin F (1997) How neural interactions form neural responses in the salamander retina. *Journal of Computational Neuroscience* 4: 5–27.

Valeton JM and Van Norren D (1983) Light adaptation of primate cones: An analysis based on extracellular data. *Vision Research* 23: 1539–1547.

Van Hateren JH and Lamb TD (2006) The photocurrent response of human cones is fast and monophasic. *BMC Neuroscience* 7: 34.

Zaghloul KA and Boahen K (2004) Optic nerve signals in a neuromorphic chip I: Outer and inner retina models. *IEEE Transactions on Biomedical Engineering* 51: 657–666.

Retinal Pharmacology: Inner Retinal Layers

P D Lukasiewicz, Washington University, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

The inner retina is composed of bipolar, horizontal, and amacrine cell interneurons, which process visual information that is received from photoreceptors, and ganglion cells that form the outputs of the retina (Figure 1). The optic nerve, which is comprised of ganglion cell axons, relays visual information to the brain. Because, on average, each ganglion cell receives input from 300 photoreceptors, the optic nerve is a potential transmission bottleneck for visual information. The inner retina processes and compresses the information received from hundreds of photoreceptors into a more compact form that can be sent to the brain by the ganglion cell axon, avoiding the bottleneck dilemma. The inner retina is involved in the complex spatial and temporal processing of visual information that is accomplished by synaptic interactions in the inner plexiform layer (IPL), the second synaptic layer of the retina. A diverse array of neurotransmitters, receptors, and transporters participates in visual processing in the IPL. The retina is an excellent model for the study of CNS synaptic pharmacology because it can be isolated intact and it can be physiologically activated with natural visual stimuli.

To understand inner retinal pharmacology, a basic knowledge of retinal circuitry is necessary (Figure 1). The retina is a layered structure that is composed of six classes of neurons. The fundamental classes of neurons and basic circuitry of the retina are well characterized. Rod and cone photoreceptors form the inputs to the retina, their somata are found in the outer nuclear layer (ONL) and they are the only photosensitive neurons involved in visual information processing in the retina. The retinal outputs are formed by ganglion cells, whose somata are found in the ganglion cell layer (GCL). Recently, a few percent of ganglion cells were found to be intrinsically photosensitive. These ganglion cells are not involved in visual processing. Instead, they provide inputs to parts of the brain that control circadian rhythms and pupillary reflexes. The retinal interneurons that process the visual signals as they flow from photoreceptors to ganglion cells are found in the inner nuclear layer (INL). Bipolar cells are important relay neurons that convey information through the vertical signaling pathway comprised of photoreceptors, bipolar cells, and ganglion cells (Figure 1, yellow neurons). Photoreceptors synapse onto bipolar cells in the first synaptic layer, called the outer plexiform layer (OPL) and bipolar cells synapse onto ganglion cells in the second synaptic layer, called the IPL. At least 10 types of bipolar cells have been described that convey different aspects of the visual signal from photoreceptors to ganglion cells, and form parallel signaling pathways. The somata of horizontal cell and amacrine cell interneurons are also located in the INL. At the OPL, horizontal cell processes make synaptic contacts with photoreceptors and bipolar cells. Horizontal cells modulate signaling from photoreceptors to bipolar cells. At the IPL, amacrine cell processes make synaptic contacts with bipolar cells and ganglion cells to modulate signaling between these two elements of the vertical signaling pathway.

Synaptic communication in the retina is mainly mediated by three neurotransmitters. Glutamate mediates fast excitation, whereas glycine and γ-aminobutyric acid (GABA) mediate fast inhibition. When bound to their receptors, all three neurotransmitters gate ion channels. Other substances released by retinal neurons act as neuromodulators. These substances do not directly excite or inhibit neurons. Instead, they modulate the effects of fast transmitters and voltage-gated channels, to fine-tune neuronal signaling. Amines and neuropeptides, and also glutamate and GABA, can act as neuromodulators.

Glutamate

Distinct Glutamate Receptors Mediate Excitatory Light Responses in ON and OFF Bipolar Cells

In the vertical signaling pathway, glutamate is the neurotransmitter utilized by both photoreceptors and bipolar cells. ON and OFF visual signals are formed at the first retinal synapse between cone photoreceptors and bipolar cells. Applied glutamate depolarizes or excites OFF bipolar cells, and hyperpolarizes or inhibits ON bipolar cells. In darkness, photoreceptors tonically release glutamate, which depolarizes OFF bipolar cells via a sign preserving synapse, and hyperpolarizes ON bipolar cells via a sign reversing synapse. Light decreases the glutamate release from photoreceptors, causing OFF bipolar cells to hyperpolarize and ON bipolar cells to depolarize. These opposite synaptic responses in OFF and ON bipolar cells are attributed to different glutamate receptor types on the dendrites of these two cell types (Figure 2). The excitatory glutamate responses of OFF bipolar cell are produced by the activation of ionotropic α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and

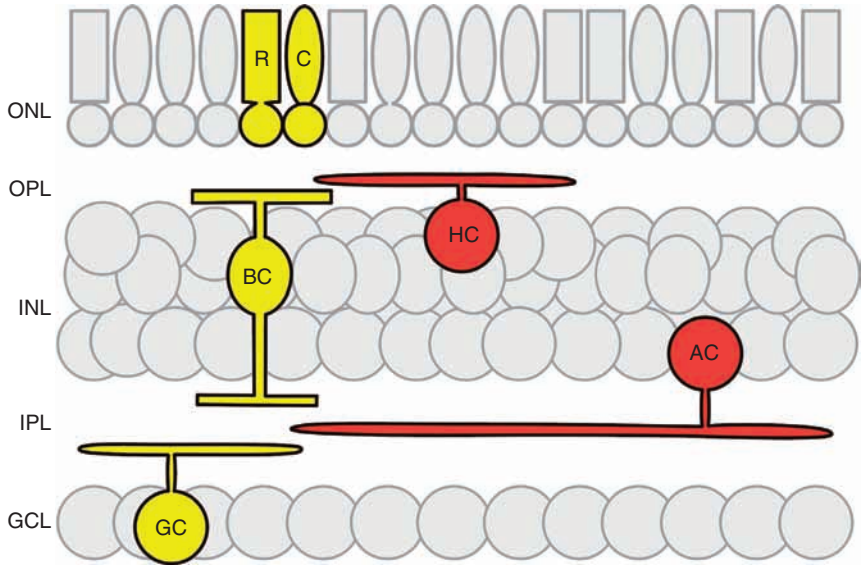

Figure 1 This schematic illustrates a cross section of the retina and the positions of the major neuronal cell types. The retina is a layered structure composed of three nuclear layers. The outer nuclear layer (ONL) contains the somata of rod (R) and cone (C) photoreceptors. The inner nuclear layer (INL) contains the somata of the retina interneurons, horizontal cells (HC), bipolar cells (BC) and amacrine cells (AC). The ganglion cell layer (GCL) contains the somata of ganglion cells, the retinal output neurons. The retina has two synaptic or plexiform layers, where most synaptic connections occur. The outer plexiform layer (OPL) is the first synaptic layer and it is where photoreceptors make synaptic contacts with horizontal cells and bipolar cells. The inner plexiform layer (IPL) is the second synaptic layer and it is where bipolar cells make synapses with amacrine and ganglion cells. The vertical signaling pathway (yellow neurons) consists of photoreceptors, bipolar cells and ganglion cells. This is the most direct signaling pathway through the retina. Horizontal cells synapse onto photoreceptors and bipolar cells and modulate signaling along the vertical pathway in the OPL. Amacrine cells synapse onto bipolar cells and ganglion cells and modulate signaling along the vertical pathway in the IPL.

kainate (KA) subtypes of glutamate receptors. Different subtypes of OFF bipolar cells have distinct synaptic response timecourses that are attributed to the predominance of either AMPA or KA receptors. The inhibitory glutamate responses of ON bipolar cells are produced by the activation of metabotropic, type 6 glutamate receptors (mGluR6), which stimulate a second messenger that closes cation channels. The mGluR6 receptors are members of group III metabotropic glutamate receptors (mGluRs) and are activated by the group-selective agonist L-amino-phosphonobutyric acid (l-AP4). These ON and OFF signaling pathways that are generated in bipolar cells are preserved throughout the visual system; they are then relayed to ON and OFF ganglion cells and then to higher visual processing centers in the brain. The locations of the major retinal glutamate receptors are summarized in Table 1 and illustrated in Figure 2.

Glutamate is the Main Excitatory Transmitter of the Inner Plexiform Layer

ON and OFF bipolar cells both release glutamate in the IPL to communicate with amacrine and ganglion cells. Bipolar cells are slow potential neurons that do not fire action potentials and tonically release glutamate when depolarized. ON bipolar cells release glutamate in response to increments of illumination and OFF bipolar cells release glutamate in response to decrements of illumination. Bipolar cell inputs to ganglion and amacrine cells activate postsynaptic AMPA and *N*-methyl-D-aspartate (NMDA) subtypes of glutamate receptors (see Figure 2 and Table 1). The AMPA receptor-mediated component of ganglion cell responses has a fast onset and rapid decay, while the NMDA receptor-mediated component has a slower onset and decay. NMDA receptor-mediated currents are also voltage dependent (unlike AMPA receptors) because at negative potentials their cation channels are blocked by magnesium, but at more positive potentials this block is relieved. NMDA receptors also differ from AMPA receptors because they are more sensitive to glutamate and require the binding of both glutamate and the coagonist glycine for activation. Glycine and D-serine, which also act as a coagonist, are abundant in the retina. Pharmacological studies suggest that extracellular levels of glycine and D-serine may be modulated to regulate NMDA receptor activation and the excitability of amacrine and ganglion cells.

The combination of AMPA and NMDA receptors may enhance the range of signaling between bipolar and ganglion cells. This is because NMDA receptor activation is voltage dependent and contributes more

to excitation at positive potentials. In contrast, AMPA responses are reduced at positive potentials. Signaling is enhanced at positive potentials because increased NMDA receptor contributions balance the reduced AMPA receptor contribution. In addition, the kinetics of excitatory synaptic signaling also depends on the relative proportions of AMPA and NMDA receptors on the postsynaptic amacrine or ganglion cell. Some amacrine cell synapses possess only AMPA receptors, while other amacrine cells and all ganglion cell synapses possess both AMPA and NMDA receptors. Thus, a single bipolar cell that contacts two postsynaptic neurons may elicit different postsynaptic responses that depend on the complement of NMDA and AMPA receptors.

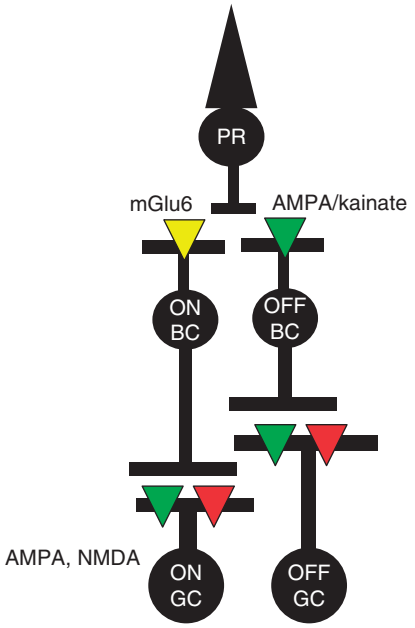

Figure 2 The locations of the major glutamate receptors in the vertical pathway are shown. In the outer plexiform layer, metabotropic, mGluR6 receptors are located on the dendrites of ON bipolar cells, and AMPA and/or kainate ionotropic glutamate receptors are located on the dendrites of OFF bipolar cells. These two distinct classes of receptors respond differently to photoreceptor synaptic input; ON bipolar cells depolarize to light increments and OFF bipolar cells depolarize to light decrements (see text). In the inner plexiform layer, bipolar cells make excitatory synapses onto ganglion cells and amacrine cells (not shown here). The glutamate receptors on the dendrites of both ON and OFF ganglion cells are similar, AMPA and NMDA ionotropic glutamate receptors.

As noted above, ON and OFF retinal pathway signaling is determined by the presence of different postsynaptic glutamate receptors on the dendrites of ON and OFF bipolar cells. Since ON and OFF bipolar cells both release glutamate that activates NMDA and AMPA receptors, ON and OFF ganglion cell synaptic responses to increments and decrement of illumination, respectively, are generally considered to be equivalent. Recent work suggests that glutamate signaling at ON and OFF ganglion cells synapses may be asymmetric. This has been attributed to differences in presynaptic inhibition of ON and OFF bipolar cell terminals (see below) and to differences in the locations of AMPA and NMDA receptors at ON and OFF ganglion cell synapses. In ON ganglion cells, AMPA receptors are located synaptically near the glutamate release sites. In contrast, NMDA receptors are located perisynaptically away from the glutamate release sites, and are activated only when sufficient glutamate is released and diffuses to perisynaptic NMDA receptor locations, a process called spillover. In OFF ganglion cells, AMPA and NMDA receptors are both located synaptically and are readily activated by even small amounts of released glutamate. The distinct locations of AMPA and NMDA receptors at ON and OFF ganglion cell synapses produce asymmetric ON and OFF visual signaling in the IPL. This asymmetry produces a larger dynamic response range in ON ganglion cells, compared to OFF cells.

Synaptic Mechanisms Shape Excitatory Signaling in the IPL

Since glutamate release from bipolar cells is tonic, prolonged increases in synaptic glutamate levels present a number of signaling challenges not found in other parts of the CNS. In other parts of the CNS glutamate is released in brief pulses in response to action potentials exciting synaptic terminals. Also, elsewhere in the CNS, glutamate responses are unaffected by receptor desensitization, are limited by agonist unbinding (deactivation), and are terminated

Table 1 Locations of major retinal glutamate receptors

Neuron type	*Synaptic layer*	*AMPA receptors*	*Kainate receptors*	*NMDA receptors*	*mGluR6*
OFF bipolar cells	OPL	Yes	Yes	No	No
ON bipolar cells	OPL	No	No	No	Yes
Rod bipolar cell	OPL	No	No	No	Yes
Amacrine cells	IPL	Yes	No	Yes	No
Ganglion cells	IPL	Yes	No	Yes	No

by rapid diffusion from the synapse. But when glutamate release is tonic, postsynaptic receptor desensitization may occur and clearance of glutamate from the synapse may be slowed. Physiological and pharmacological studies show that the clearance of glutamate from ganglion cell synapses depends upon glutamate uptake by transporters. Pharmacological blockade of glutamate transporters, located on neighboring glial cell and bipolar cell terminals, enhances and prolongs excitatory light-evoked responses in ganglion cells, suggesting that transporters normally limit the amplitude and duration of excitatory synaptic responses.

AMPA receptors rapidly desensitize to maintained activation by glutamate. Because glutamate release from bipolar cells is sustained and glutamate clearance by uptake is slow, elevated synaptic glutamate levels may desensitize AMPA receptors. Pharmacological studies suggest that AMPA receptor desensitization shapes light-evoked responses in ganglion cells. Ganglion cell responses were enhanced when AMPA receptor desensitization was pharmacologically blocked by cyclothiazide or reduced by depolarization, suggesting that AMPA receptor desensitization normally limits ganglion cell excitatory responses.

Autoinhibitory mechanisms can also modulate excitatory transmission in the IPL by limiting glutamate release from bipolar cells. Glutamate released from bipolar cells may limit release by activating mGluRs located on bipolar cell terminals. In addition to glutamate, protons are also released from vesicles. Increased extracellular proton levels also limit glutamate release, but they do so by inhibiting calcium influx into bipolar cell terminals.

mGluRs are also found on the dendrites of amacrine and ganglion cells, where they are thought to modulate synaptic signaling in the IPL. Group I, II, and III mGluRs are found on amacrine and ganglion cell processes, where they act postsynaptically to modulate synaptic transmission. Often mGluRs are found only on one of the two processes postsynaptic to the bipolar cell, suggesting that these receptors selectively modulate transmission. The group I receptor member, mGluR1, is located on amacrine cell processes and modulates the release of GABA to control the strength of feedback inhibition to bipolar cell terminals. However, the physiological roles of most of the mGluRs in the IPL remain unknown.

GABA

Inhibitory signaling in the IPL is mediated by amacrine cells, which comprise the most diverse class of retinal interneurons. Amacrine cells are morphologically and functionally distinct, possessing different sets of neurotransmitters and receptors.

Table 2 Locations of inner retinal GABA and glycine receptors

Neuron type	*$GABA_A$ receptor*	*$GABA_B$ receptor*	*$GABA_C$ receptor*	*Glycine receptor*
OFF bipolar cells	Yes	?	Yes	Yes
ON bipolar cells	Yes	?	Yes	No
Rod bipolar cell	Yes	?	Yes	Yes
Amacrine cells	Yes	Yes	No	Yes
Ganglion cells	Yes	Yes	No	Yes

Approximately 50% of all amacrine cells are GABAergic and are a major source of fast inhibitory signaling in the IPL. The population of GABAergic amacrine cells is composed of multiple subtypes that share the common feature of possessing wide field processes that extend laterally but not vertically within the IPL. Thus, GABAergic amacrine cells signal within specific layers of the IPL, but they do not signal between layers. Wide field, GABAergic amacrine cells mediate lateral inhibition between distant excitatory signaling pathways, modulating bipolar to ganglion cell transmission. GABAergic amacrine cells also mediate local feedback inhibition that controls glutamate release from bipolar cells, as described below.

GABA receptors are found throughout the IPL and are located presynaptically on bipolar cell terminals and postsynaptically on amacrine cell and ganglion cell dendrites. There is a diversity of GABA receptors in IPL that includes ionotropic $GABA_A$ and $GABA_C$ receptors and metabotropic $GABA_B$ receptors. Table 2 summarizes the locations of these GABA receptor subtypes.

$GABA_A$ and $GABA_C$ Receptors

The retina is atypical in the CNS because it possesses two distinct types of ionotropic GABA receptors, $GABA_A$ and $GABA_C$. Both receptor subtypes gate chloride channels, but they are composed of distinct molecular subunits and have different biophysical properties. $GABA_C$ receptors are more sensitive to GABA than $GABA_A$ receptors, and $GABA_C$ receptors mediate more sustained responses.

The pharmacology of $GABA_A$ and $GABA_C$ receptors is distinct. $GABA_A$ receptors are antagonized by bicuculline and gabazine and potentiated by barbiturates and benzodiazepines. $GABA_C$ receptors are unaffected by these $GABA_A$ receptor-specific agents, but their function is blocked by the selective antagonist (1,2,5,6-tetrahydropyridin-4-yl)methylphosphinic

acid (TPMPA). $GABA_A$ receptors are located throughout the IPL on amacrine, bipolar, and ganglion cells and have both pre- and postsynaptic locations (Figure 3). Anatomical and molecular studies suggest that multiple subtypes of $GABA_A$ receptors exist in the IPL. However, subunit-specific antagonists are not yet available to confirm the functional significance of this diversity of $GABA_A$ receptors. In contrast, $GABA_C$ receptors have very specific locations in the retina; they are only found on bipolar cell terminals, where they mediate presynaptic inhibition (Figure 3).

Ionotropic GABA receptors in the IPL control communication between bipolar cells and their postsynaptic targets. GABA released from amacrine cells activates $GABA_A$ receptors on the dendrites of amacrine and ganglion cells, mediating postsynaptic inhibition that limits excitation. GABA release also activates $GABA_A$ and $GABA_C$ receptors on bipolar cell terminals, limiting glutamate release. Both local and lateral forms of GABAergic inhibition occur. Lateral inhibition contributes to the antagonistic receptive field surround in ganglion cells. Local inhibition occurs at pre- and postsynaptic synapses, where excitatory input from bipolar cells causes local GABA release from amacrine cells that activates presynaptic $GABA_A$ and $GABA_C$ receptors on bipolar cell terminals or postsynaptic $GABA_A$ receptors on amacrine and ganglion cell dendrites.

Presynaptic inhibition contributes to the temporal processing of visual information by controlling the time course of ganglion cell excitation and by limiting glutamate release from bipolar cells. Distinct combinations of fast $GABA_A$ and slow $GABA_C$ receptors are found on different bipolar cells, suggesting that presynpatic inhibition is temporally filtered for specific bipolar cell pathways. Pre- and postsynaptic lateral GABAergic inhibition affects the spatial processing of visual information and contributes to the receptive field organization of ganglion cell types.

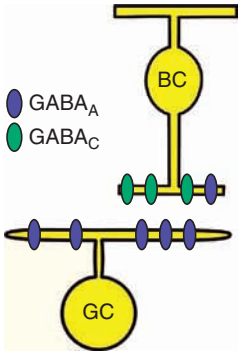

Figure 3 The presynaptic and postsynaptic locations of ionotropic GABA receptors in the IPL are illustrated. Presynaptic $GABA_A$ (blue) and $GABA_C$ (green) receptors are located on the terminals of all classes of bipolar cell. The relative proportion of each receptor class varies with bipolar cell type. Postsynaptic $GABA_A$ (blue) receptors are found on the dendrites of ganglion cells and the processes of amacrine cells (not shown here). As noted in the text, $GABA_A$ and $GABA_C$ receptors have different functional properties and are thought to mediate temporally distinct forms of inhibition. $GABA_A$ receptors mediate fast inhibition, while $GABA_C$ receptors mediate a slower form of inhibition.

$GABA_B$ Receptors

The G-protein-coupled $GABA_B$ receptors are found on amacrine and ganglion cells, where they have been shown to modulate voltage-gated potassium and calcium channels. There is no anatomical evidence for $GABA_B$ receptors on bipolar cell terminals, but physiological and pharmacological evidence from cold blooded animals suggests that they limit glutamate release from bipolar cells. The $GABA_B$ receptor agonist baclofen has been reported to enhance transient excitatory light responses, but suppress sustained, excitatory light responses in ganglion cells. However, the effects of $GABA_B$ receptor antagonists upon ganglion cells are relatively weak. Thus, there is some evidence for the presence of $GABA_B$ receptor in the inner retina, but it is unclear what the precise roles of this class of receptor are in retinal information processing.

Glycine

Glycine is the other major fast inhibitory neurotransmitter of the inner retina, and approximately one-half of all amacrine cells are glycinergic. Glycinergic amacrine cells are distinct from GABAergic amacrine cells because their processes have narrow fields that do not extend laterally. Unlike GABAergic amacrine cell processes, glycinergic cell processes extend vertically within the IPL and mediate signaling between different IPL layers. These distinct morphological features suggest that these two classes of inhibitory amacrine cells have different functional roles. The glycinergic, AII or rod amacrine cell is the best example of this principle; it receives input in the ON layers of the IPL and sends output in the OFF layer of the IPL. Glycine receptors are found presynaptically on the terminals of OFF cone and rod bipolar cells, and postsynaptically on the dendrites of amacrine and ganglion cells. Immunocytochemical labeling of receptor subunits suggests that there is a diversity of glycine receptors expressed in the inner retina, which may mediate distinct responses. However, additional physiological and pharmacological studies are needed to show that distinct glycine receptors play different roles in retinal signal processing.

There is strong evidence indicating that glycinergic inhibition plays several critical signaling roles in the retina. Glycine-mediated presynaptic inhibition from AII amacrine cells of OFF cone bipolar cells plays a critical role in rod-mediated signaling in the mammalian retina. Blocking this pathway with the specific antagonist strychnine eliminates rod-mediated OFF signals. Glycine also mediates change-sensitive surround inhibition in the retinas of lower vertebrates, which is distinct from the stationary surround inhibition, mediated by GABAergic amacrine cells.

Acetylcholine

Starburst amacrine cells are the main cholinergic neuron in the retina and they play a role in directionally sensitive (DS) responses, recorded in some ganglion cells. These cholinergic amacrine cells make synaptic contacts with directionally sensitive ganglion cells, other amacrine cells and some bipolar cell terminals via nicotinic acetylcholine (ACh) receptors. The inhibitory neurotransmitter GABA is co-localized with ACh in starburst amacrine cells. The genetic ablation of starburst amacrine cells eliminates DS responses. However DS responses are relatively unaffected when Ach receptors are blocked, suggesting that GABA signals from starburst amacrine cells may be more important for DS responses. Muscarinic ACh receptors are also found on various inner retinal neurons, but the function of these receptors is unclear.

Dopamine

Dopamine is the best-studied neuromodulator in the vertebrate retina. It is found in a small population of amacrine cells, which stain positively for the dopamine synthetic enzyme, tyrosine hydroxylase. Although the population of dopaminergic amacrine cells is small in number, their processes are wide ranging and tile or cover the entire retina. The fast inhibitory transmitter GABA is co-localized with dopamine in this class of amacrine cells. Dopamine receptors are found on most retinal neurons. In many species, dopamine acts as a paracrine substance that is released into the extracellular space to affect both local and distant targets. Dopamine is a chemical mediator of light adaptation in the retina. It acts in the outer retina to reduce signaling through rod pathways and enhance signaling through cone pathways. Dopamine also affects a number of inner retinal processes. Increased dopamine levels uncouple the electrical synapses between AII amacrine cells, reducing rod-mediated signaling. Dopamine may also modulate fast synaptic signaling by increasing the $GABA_A$ receptor mediated responses in amacrine cells and enhancing Ach release from cultured amacrine cells. In salamander retina, dopamine reduces $GABA_C$ receptor-mediated presynaptic inhibition of bipolar cell terminals, enhancing the excitatory drive to ganglion cells. Finally, dopamine affects the balance of center and surrounds receptive field inputs through its actions in the outer retina, where it reduces electrical coupling between horizontal cells.

Neuropeptides

A number of neuroactive peptides have been localized to specific populations of amacrine cells with immunocytochemistry. In other parts of the CNS, neuropeptides have modulatory roles and affect the actions of fast excitatory and inhibitory transmitters or the activity of voltage-gated channels to alter the excitability or transmitter release properties of neurons. Different populations of amacrine cells contain the following neuroactive peptides: neuropeptide Y, somatostatin, vasoactive intestinal polypeptide, and enkephalins. Calcium imaging studies suggest that both neuropeptide Y and somatostatin inhibit calcium currents in rod bipolar cell terminals. Presumably, these actions will limit glutamate release and reduce excitation in the rod signaling pathway. The precise mechanisms by which neuropeptides modify visual information processing in the retina remain unknown.

In summary, the inner retina is critical for visual processing. Information processing occurs in the second synaptic layer, the IPL. Glutamatergic, bipolar cells provide the excitatory input to the IPL and ganglion cells form the output. Excitatory signaling is shaped by postsynaptic AMPA and NMDA receptor properties, glutamate uptake and presynaptic inhibition. GABAergic and glycinergic inhibitory amacrine cells also shape the synaptic signaling between bipolar and ganglion cells. Together, these complex excitatory and inhibitory synaptic interactions mediate the temporal and spatial processing of visual information in the inner retina. Neuromodulators are thought to fine-tune fast excitatory and inhibitory signals. However, with the exception of dopamine, we still know relatively little about how these neuromodulators affect visual processing.

See also: Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Amacrine Cells; Retinal Bipolar Cells; Retinal Ganglion Cells: Anatomy; Retinal Horizontal Cells; Retinal Color Mechanisms; Retinal Development: An Overview; Retinal Glia.

Further Reading

Dowling JE (1987) *The Retina: An Approachable Part of the Brain*. Cambridge, MA: The Belknap Press of Harvard University Press.

Lukasiewicz PD (1996) GABAc Receptors in the vertebrate retina. *Molecular Neurobiology* 12: 181–194.

Lukasiewicz PD (2005) Synaptic mechanisms that shape visual signaling at the inner retina. *Progress in Brain Research* 147: 205–218.

MacNeil MA and Masland RH (1998) Extreme diversity among amacrine cells: Implications for functions. *Neuron* 20: 971–982.

Mittman S, Taylor WR, and Copenhagen DR (1990) Concomitant activation of two types of glutamate receptor mediates excitation of salamander retinal ganglion cells. *Journal of Physiology* 428: 175–197.

Roska B and Werblin F (2001) Vertical interactions across ten parallel, stacked representations in the mammalian retina. *Nature* 410: 583–587.

Slaughter MM (1995) $GABA_B$ receptors in the vertebrate retina. *Progress in Retinal and Eye Research* 14: 293–312.

Slaughter MM and Miller RF (1981) 2-Amino-4-phosphonobutyric acid: A new pharmacological tool for retina research. *Science* 211: 182–185.

Sterling P and Matthews G (2005) Structure and function of ribbon synapses. *Trends in Neurosciences* 28: 20–29.

Taylor WR and Vaney DI (2003) New directions in retinal research. *Trends in Neuroscience* 26: 379–385.

Wassle H (2004) Parallel processing in the mammalian retina. *Nature Reviews Neuroscience* 5: 747–757.

Wassle H and Boycott BB (1991) Functional architecture of the mammalian retina. *Physiological Reviews* 71: 447–480.

Wassle H, Koulen P, Brandstatter JH, Fletcher EL, and Becker C (1998) Glycine and GABA receptors in the mammalian retina. *Vision Research* 38: 1411–1430.

Werblin FS and Dowling JE (1969) Organization of the retina of mudpuppy, *Necuturus maculosus*. II. Intracellular recording. *Journal of Neurophysiology* 32: 339–355.

Witkovsky P (2004) Dopamine and retinal function. *Documenta Ophthalmologica* 108: 17–40.

Retinitis Pigmentosa

E L Berson, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Retinitis pigmentosa has a prevalence of about 1 in 4000 worldwide. About 50 000–100 000 people are affected in the United States. Affected patients typically report night blindness in adolescence, loss of mid-peripheral vision in young adulthood, and loss of far peripheral and then central vision in middle life. Clinical findings on ophthalmoscopy include intraretinal pigment around the mid-peripheral retina for which the condition is named, attenuated retinal vessels, and waxy pallor of the optic disks. Histologic examinations of autopsy eyes have shown that loss of vision is due to degeneration of both rod and cone photoreceptor cells across the retina.

This group of diseases can be inherited by an autosomal dominant, autosomal recessive, X-linked, digenic, or mitochondrial mode of transmission; uniparental isodisomy has also been rarely reported. Assuming most isolate cases (i.e., cases with no relatives known to be affected) are inherited by an autosomal recessive mode, it is estimated that some 30–40% of cases are inherited as an autosomal dominant trait, 50–60% as an autosomal recessive trait, and 10–15% as an X-linked trait. Some 15% of all cases have associated partial hearing loss (Usher syndrome, type II) and another 2–6% have associated profound congenital deafness and vestibular ataxia (Usher syndrome, type I). About 1% of patients with retinitis pigmentosa have adult-onset hearing loss with or without ataxia (Usher syndrome, type III).

Retinitis pigmentosa can be detected in early life by electroretinographic testing. Electroretinograms (ERGs) are recorded in response to flashes of light with a contact lens electrode placed on the topically anesthetized cornea; responses are electronically amplified and displayed on an oscilloscope. Patients with early retinitis pigmentosa typically have reduced rod and cone ERG responses which are delayed in their peak implicit times (i.e., time interval between stimulus flash onset and the peak of the corresponding response). Abnormal ERGs have been detected in asymptomatic children years to a decade before diagnostic changes are seen on routine ocular examination. ERG amplitudes become smaller as the disease progresses (Figure 1). Most patients with remaining vision have detectable responses when ERGs are recorded with computer averaging and narrow bandpass filtering. Patients aged 6 and older with normal ERGs in families with retinitis pigmentosa have not been observed to develop this condition at a later time.

More than 100 gene loci that cause retinitis pigmentosa have been identified or mapped; an updated list of genes causing this condition is maintained on the RetNet website. Genes so far identified as causes of nonsyndromic retinitis pigmentosa as well as Usher syndrome are given in Table 1. Molecular genetic analyses have revealed defects in genes encoding proteins involved in the phototransduction cascade (see Figure 2), the retinoid cycle, maintenance of photoreceptor cell structure, cell–cell interactions, RNA intron-splicing factors, intracellular transport of proteins, regulation of carbon dioxide/bicarbonate balance, phagocystosis, and in yet-to-be-defined functions in the photoreceptors or retinal pigment epithelium. About 50–60% of cases of retinitis pigmentosa have known molecular genetic abnormalities.

Variability of clinical expression has been observed both within and between families among patients with the same gene defect, leading to the hypothesis that some factor or factors other than the gene defect itself contributes to the pathogenesis of disease. Light exposure, nutritional factors, and/or modifying genes have been proposed as possible risk factors that might contribute to severity of disease at a given age. More than 100 mutations have been found in the rhodopsin gene alone and severity of disease has been correlated with the region of the opsin molecule affected by a given mutation; for example, patients with a mutation near the C-terminus (e.g., rhodopsin, Pro347-Leu) have more severe disease at a given age than patients with a mutation near the N-terminus (e.g., rhodopsin, Pro23His). Mutations in the peripherin/RDS gene have been associated with retinitis pigmentosa, retinitis punctata albescens, and retinal degenerations primarily involving the macula, showing that abnormalities in the same gene can be associated with different phenotypes.

A severe form of retinitis pigmentosa is called Leber congenital amaurosis. Mutations in at least ten genes have been implicated as causes of this recessively inherited condition; these genes are GUCY2D, CRB1, IMPDH1, RPE65, AIPL1, RPGRIP1, RDH12, CRX, TULP1, and LRAT. Affected children are typically born with little, if any, vision, nystagmus, and hyperopia. Optical coherence tomography has revealed that patients with little vision may still have some remaining photoreceptor nuclei that might be amenable to treatment. Promising results in restoring retinal

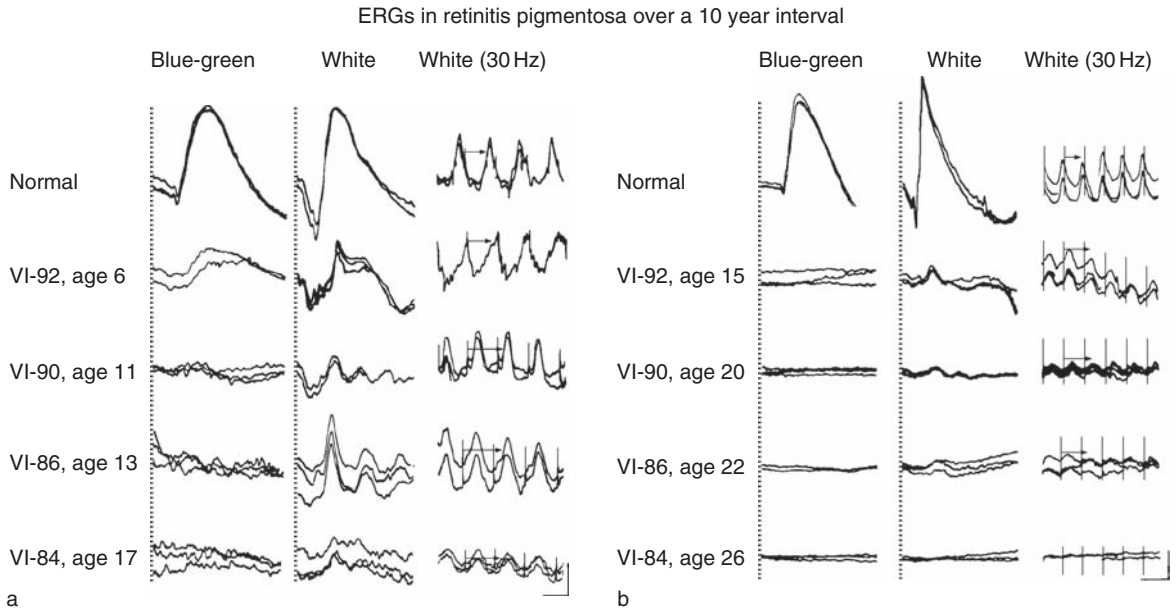

Figure 1 ERG responses recorded in 1967 (a) and 1977 (b) from a normal subject and from four affected members of a family with a dominant form of retinitis pigmentosa transmitted over six generations. Pedigree number and age at time of testing are indicated for each patient. One to three responses to same stimulus are represented. For (a) and (b), rod-isolated responses to single flashes of blue-green light are illustrated in the left column; mixed cone–rod responses to single flashes of white light are illustrated in the middle column; and cone-isolated responses to 30 Hz white flickering light are illustrated in the right column. Calibration symbol for 1967 responses designates 50 ms horizontally for columns 1 and 2, 25 ms for column 3, 50 μV vertically for column 1, and 100 μV vertically for columns 2 and 3. Calibration symbol for 1977 responses designates 60 ms horizontally and 100 μV vertically for all tracings. Reproduced from Berson EL and Simonoff EA (1979) Dominant retinitis pigmentosa with reduced penetrance: Further studies of the electroretinogram. *Archives of Ophthalmology* 97: 1286–1291, with permission from American Medical Association. Copyright © (1979) American Medical Association.

function have been achieved with a viral-mediated subretinal injection of the RPE65 gene in a canine model and the RPGRIP gene in a murine model of this group of diseases when treatment is initiated at a time when these models have remaining photoreceptors. It remains to be established whether viral-mediated gene therapy is safe and effective in humans with the early stages of Leber congenital amaurosis.

Retinitis pigmentosa has been associated not only with hearing loss but also with polydactyly, mental retardation, truncal obesity, and hypogonadism (Laurence–Moon–Bardet–Biedl syndrome). Some clinicians have subdivided this syndrome into two disorders; polydactyly is found primarily in patients with the Bardet–Biedl form, and neurological abnormalities are found usually in patients with the Laurence–Moon form. Variability of clinical expression is well known and some patients may have this condition without mental retardation or polydactyly. Mutations in at least nine genes have so far been identified (i.e., BBS1, BBS2, ARL6, BBS4, BBS5, MKKS, BBS7, TTC8, and PHTB1). Renal disease occurs in some cases; therefore, patients should have their blood pressure and urine checked periodically. No treatment is known for the retinal degeneration observed in some 90% of patients with this syndrome.

Retinitis pigmentosa is also found in association with anosmia, ataxia, dry skin, and EKG abnormalities with elevated serum phytanic acid (Refsum disease); malabsorption, acanthocytosis, ataxia, and abetalipoproteinemia (Bassen–Kornzweig syndrome); and adult-onset ataxia, dysarthria, hyporeflexia, decreased proprioception, decreased vibration sense, and low serum vitamin E levels (Friedreich-like ataxia with retinitis pigmentosa). These uncommon recessively inherited conditions are treatable as follows: a low phytol/low phytanic acid diet (excluding animal fats, milk products, and dark green leafy vegetables) in the case of Refsum disease; a low-fat diet plus supplementation with vitamin A, vitamin E, and vitamin K in the case of Bassen–Kornzweig syndrome, and vitamin E supplementation in the case of Friedreich-like ataxia associated with retinitis pigmentosa.

The common forms of retinitis pigmentosa have also yielded to treatment. In a randomized, controlled double-masked trial among 601 patients aged 18–49, the course of retinal degeneration as monitored by the ERG was slower on average among patients taking a daily supplement of vitamin A palmitate, 15 000 IU day^{-1}, than among those not on this dose. Furthermore, the course appeared to be faster on average among patients taking a daily supplement of 400 IU day^{-1} of vitamin

Table 1 Estimated proportions of cases of retinitis pigmentosa (RP) caused by Identified genes (excluding syndromic RP except for Usher syndrome)

Inheritance pattern/gene	*Percent of all RP (including Usher syndrome)*
Autosomal dominant (~40% of all cases of retinitis pigmentosa)	
1. RHO (rhodopsin) (3q)	10%
2. RP1 (8q)	2.2%
3. PRPF31 (19q)	2.0%
4. PRPF3 (1q)	1.6%
5. peripherin/RDS (6p)	~1%
6. FSCN2 (fascin) (17q)	~1%
7. PRPF8 (17p)	0.8%
8. IMPDH1 (7q)	0.8%
9. NRL (14q)	0.4%
10. CRX (19q)	0.4%
11. CA4 (carbonic anhydrase) (17q)	? frequency in USA (~2% in UK)
12. PAP1 (7p)	<1%
13. SEMA4A (semaphorin) (1q)	<1%
Autosomal recessive (~50% of all cases of retinitis pigmentosa, including isolates)	
14. usherin (1q)	~10% (Usher IIA)
15. rod PDEβ (4p)	2%
16. rod PDEα (5q)	1.5%
17. CNGA1 (rod channel α (4p)	1%
18. RPE65 (1p)	1%
19. myosin VIIa (11q)	~1% (Usher IB)
20. LRAT (4q)	<1%
21. NR2E3 (15q)	<1%
22. MERTK (2q)	<1%
23. harmonin (11p)	<1% (Usher IC)
24. CDH23 (10q)	<1% (Usher ID)
25. PCDH15 (10q)	<1% (Usher IF)
26. MASS1 (5q)	<1% (Usher IIC)
27. clarin-1 (3q)	<1% (Usher IIIA)
28. SAG (arrestin) (2q)	~0.5%
29. CRALBP (15q)	~0.5%
30. RHO (rhodopsin) (3q)	~0.5%
31. TULP1 (6q)	~0.5%
32. ABCA4 (i.e., ABCR) (1p)	<1%
33. RGR (10q)	<1%
34. CRB1 (1q)	<1%
35. CERKL (2q)	<1%
36. CNGB1 (rod channel β) (16q)	<1%
37. SANS (17q)	<1% (Usher 1G)
38. RP1 (8q)	<1%
X-linked (~10% of all cases of retinitis pigmentosa)	
39. RPGR (Xp21.1)	8% (RPGR and RP2 combined account for 1%
40. RP2 (Xp11.3)	~90% of XLRP)
Digenic (only a few families)	
41. ROM1 (11q) and peripherin/RDS (6p)	<1%
Mitochondrial (only one family)	
42. MTTS2	<1%

Percentages are approximate and are based on the breakdown of RP cases according to genetic type reported by Fishman (*Archives of Ophthalmology* 96: 822–826, 1978), Macrae (*Birth Defects: Original Article Series* 18: 175–185, 1982), and Bunker et al. (*American Journal of Ophthalmology* 97: 357–365, 1984) and on the assumptions that all isolate cases are autosomal recessive and that about one-third of cases of Usher syndrome type I (approximately 6% of all cases of RP) are caused by defects in myosin VIIa. The values are calculated based on frequency of cases in a published series multiplied by the proportion of RP with that inheritance pattern (e.g., dominant rhodopsin mutations were found in 90/363 cases of ADRP; ADRP accounts for about 40% of all cases of RP; accordingly, the percentage of cases caused by dominant rhodopsin mutations is $90/363 = 24.8\% \times 0.4 = 9.92 \approx 10\%$).
A current referenced list of genes causing retinitis pigmentosa and allied retinal diseases can be accessed on the world wide web at: http://www.sph.uth.tmc.edu/Retnet/sum-dis.htm.

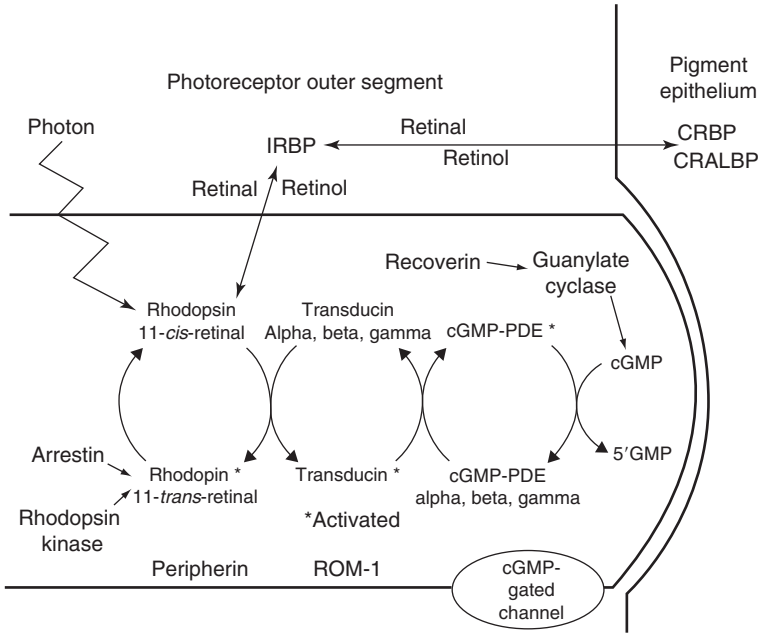

Figure 2 Schematic representation of some proteins normally found in the rod photoreceptor outer segment, interphotoreceptor matrix, and retinal pigment epithelium. Genes encoding these proteins are considered candidate genes that could cause retinitis pigmentosa because mutations in them could result in compromise of the phototransduction cascade or the mechanism by which vitamin A is transported between the photoreceptors and the pigment epithelium with consequent photoreceptor cell degeneration. IRBP, interphotoreceptor retinoid binding protein; CRBP, cellular retinol-binding protein; CRALBP, cellular retinal-binding protein; PDE, phosphodiesterase; ROM1, rod outer segment membrane protein 1; cGMP, cyclic guanosine monophosphate. Reproduced from Berson EL (1993) Retinitis pigmentosa: The Friedenwald lecture. *Investigative Ophthalmology & Visual Science* 34: 1659–1676, with permission from Association for Research in Vision and Ophthalmology.

E than those not on this dose. This has led to the recommendation that most adult patients with the common forms of retinitis pigmentosa should take a daily supplement of 15 000 IU of vitamin A in the palmitate form under the supervision of their physician and avoid high-dose supplementation with vitamin E. Since doses of $>25\,000\,\mathrm{IU\,day^{-1}}$ of preformed vitamin A can be associated with liver disease when taken over the long term, patients should have a fasting serum vitamin A and liver function profile before commencing this treatment and measures of these parameters annually thereafter while on this supplement. Beta-carotene is not predictably converted to preformed vitamin A and therefore should not be considered as a substitute for treatment with vitamin A palmitate. Sources of the appropriate dose of vitamin A palmitate are maintained on the website of the Foundation Fighting Blindness.

No toxic effects have been observed in adults with retinitis pigmentosa in good general health on this supplement of vitamin A who have been followed for up to 12 years. Because of the increased risk of birth defects associated with high-dose vitamin A supplementation, women who are pregnant or planning to become pregnant should not take this supplement. Long-term vitamin A supplementation has been associated with a decrease in bone density and up to a 1% increase in risk of hip fracture; therefore, postmenopausal women and men aged 49 and above with retinitis pigmentosa taking vitamin A should have their bone health monitored periodically by their family doctor. Since patients under age 18 were not included in this trial, no formal recommendation can be made for patients with retinitis pigmentosa under this age.

A second randomized, controlled, double-masked trial for retinitis pigmentosa revealed that docosahexaenoic acid (DHA) supplementation in a dose of $1200\,\mathrm{mg\,day^{-1}}$ did not, on average, provide benefit for adults also taking vitamin A $15\,000\,\mathrm{IU\,day^{-1}}$. However, a subgroup analysis revealed that those taking vitamin A who ate two or more 3-ounce servings of oily fish per week (i.e., salmon, tuna, herring, mackerel, or sardines) of which DHA is a major constituent showed a 40–50% slowing of annual decline in visual field sensitivity. It has been estimated that the combination of vitamin A palmitate $15\,000\,\mathrm{IU\,day^{-1}}$ and a diet rich in oily fish could result in almost 20 years of additional visual preservation. The average patient who starts this regimen in his or her thirties would be expected to retain some useful vision until about age 80 in contrast to an

untreated patient who would be expected to be virtually blind by age 60.

Night vision devices are available that amplify light and thereby allow patients with retinitis pigmentosa and night blindness to use their remaining cones to function under dim daylight, moonlight, or even starlight conditions. The best candidates for these devices are patients with best corrected vision of better than 20/200 in at least one eye and a central visual field diameter of >20° in that eye. For patients with reduced visual acuity, a closed circuit television for reading is helpful in some instances. Patients with retinitis pigmentosa should avoid exposure to bright sunlight outdoors until more is known about whether exposure to sunlight affects the course of this condition. Since no particular type of sunglass has been shown to alter the course of retinal degeneration, patients should select sunglasses for outdoor use that provide maximal protection from sunlight without compromising vision.

The precise cellular mechanisms that lead to photoreceptor cell death are under study. Among the questions remaining to be answered are why rod-specific gene defects lead to cone photoreceptor cell death and whether the process of apoptosis (or genetically programmed cell death) can help to explain pathogenesis. Efforts to replace degenerated photoreceptor cells through transplantation of healthy photoreceptor cells in the subretinal space have so far proved unsuccessful as the implanted photoreceptors have not formed functional connections with the proximal retina necessary for the transmission of visual information to the brain. Light-sensitive microchips implanted on or under the retina are under study but their value remains to be established. Studies are in progress to rescue remaining photoreceptor cells with ciliary neurotrophic factor (CNTF), gene therapy, or other nutritional supplements. A multidisciplined approach combining biochemistry, electrophysiology, molecular and cell biology, and ophthalmology would appear to have the greatest chance of making new discoveries that will benefit patients with this group of diseases.

See also: Animal Models of Inherited Retinal Degenerations; Cone Photopigment Evolution; Electroretinography; Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Color Mechanisms; Retinal Development: An Overview; Vision: Light and Dark Adaptation.

Further Reading

Acland GM, Aguirre GD, Ray J, et al. (2001) Gene therapy restores vision in a canine model of childhood blindness. *Nature Genetics* 28: 92–95.

Andréasson SOL, Sandberg MA, and Berson EL (1987) Narrow-band filtering for monitoring low-amplitude cone electroretinograms in retinitis pigmentosa. *American Journal of Ophthalmology* 105: 500–503.

Apfelstedt-Sylla E, Theischen M, Ruther K, Wedemann H, Gal A, and Zrenner E (1995) Extensive intrafamilial and interfamilial phenotypic variation among patients with autosomal dominant retinal dystropy and mutations in the human RDS/peripherin gene. *British Journal of Ophthalmology* 79: 28–34.

Ben-Arie-Weintrob Y, Berson EL, and Dryja TP (2005) Histopathologic–genotypic correlations in retinitis pigmentosa and allied diseases. *Ophthalmic Genetics* 26: 91–100.

Berson EL (1993) Retinitis pigmentosa: The Friedenwald lecture. *Invest Ophthalmology & Visual Science* 34: 1659–1676.

Berson EL (1998) Treatment of retinitis pigmentosa with vitamin A. Proceedings of the Fernström Symposium on Tapetoretinal Degenerations, Lund, Sweden. *Digital Journal of Ophthalmology* 4(7). http://www.djo.harvard.edu (accessed by August 2008).

Berson EL (2007) Long-term visual prognoses in patients with retinitis pigmentosa: The Ludwig von Sallmann Lecture. *Experimental Eye Research* 85: 7–14.

Berson EL, Rosner B, Sandberg MA, et al. (1993) A randomized trial of vitamin A and vitamin E supplementation for retinitis pigmentosa. *Archives of Ophthalmology* 111: 761–772.

Berson EL, Rosner B, Sandberg MA, et al. (2004) Clinical trial of docosahexaenoic acid in patients with retinitis pigmentosa receiving vitamin A treatment. *Archives of Ophthalmology* 122: 1297–1305.

Berson EL, Rosner B, Sandberg MA, et al. (2004) Further evaluation of docosahexaenoic acid in patients with retinitis pigmentosa receiving vitamin A treatment: Subgroup analyses. *Archives of Ophthalmology* 122: 1306–1314.

Berson EL and Simonoff EA (1979) Dominant retinitis pigmentosa with reduced penetrance: Further studies of the electroretinogram. *Archives of Ophthalmology* 97: 1286–1291.

Bishara S, Merin S, Cooper M, Azizi E, Delpre G, and Deckelbaum RJ (1982) Combined vitamin A and E therapy prevents retinal electrophysiological deterioration in abetalipoproteinemia. *British Journal of Ophthalmology* 66: 767–770.

Feskanich D, Singh V, Willett WC, and Colditz GA (2002) Vitamin A intake and hip fractures among postmenopausal women. *Journal of the American Medical Association* 287: 47–54.

Gouras P, Carr RE, and Gunkel RD (1971) Retinitis pigmentosa in abetalipoproteinemia: Effects of vitamin A. *Investigative Ophthalmology & Visual Science* 10: 784–793.

Hartong DT, Berson EL, and Dryja TP (2006) Retinitis Pigmentosa. *Lancet* 368: 1795–1809.

Michaelsson K, Lithell H, Vessby B, and Melhus H (2003) Serum retinol levels and the risk of fracture. *New England Journal of Medicine* 348: 287–294.

Pawlyk BS, Smith AJ, Buch PK, et al. (2005) Gene replacement therapy rescues photoreceptor degeneration in a murine model of Leber congenital amaurosis lacking RPGRIP. *Investigative Ophthalmology & Visual Science* 46: 3039–3045.

Refsum S (1981) Heredopathia atactica polyneuritiformis, phytanic acid storage disease Refsum's disease: A biochemically well-defined disease with a specific dietary treatment. *Archives of Neurology* 38: 605–606.

Sibulesky L, Hayes KC, Pronczuk A, Weigel-Di Franco C, Rosner B, and Berson EL (1999) Safety of <7500 RE (<25 000 IU) vitamin A daily in adults with retinitis pigmentosa. *American Journal of Clinical Nutrition* 69: 656–663.

Sieving P, Caruso RC, Tao W, et al. (2006) Ciliary neurotrophic factor (CNTF) for human retinal degeneration: Phase I trial of CNTF delivered by encapsulated cell intraocular implants. *Proceedings of the National Academy of Sciences of the United States of America* 103: 3896–3901.

Yokota T, Shiohiri T, Gotoda T, et al. (1997) Friedreich-like ataxia with retinitis pigmentosa caused by the His101Gln mutation of the α-tocopherol transfer protein gene. *Annals of Neurology* 41: 826–832.

Relevant Websites

http://www.sph.uth.tmc.edu – RetNet: Summaries of Genes and Loci Causing Retinal Diseases, The University of Texas School of Public Health (for a current referenced list of genes causing retinitis pigmentosa).

http://www.blindness.org – The Foundation Fighting Blindness (for a list of sources of the appropriate dose of vitamin A palmitate).

Retinoic Acid Signaling and Neural Patterning

G Duester, Burnham Institute for Medical Research, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Retinoic Acid Signaling Plays a Role in Patterning of the Posterior Central Nervous System

During gastrulation, the rostral–caudal axis of the mammalian central nervous system (CNS) is progressively laid down in an anterior to posterior fashion as primitive ectoderm in the epiblast differentiates into neuroectoderm. Various signaling pathways direct specific regions of the neuroectoderm to adopt forebrain, midbrain, hindbrain, or spinal cord fates. One of these signals is retinoic acid (RA), a lipid derived from metabolism of vitamin A (retinol). RA is not synthesized during early gastrulation, when the forebrain and midbrain fields are established (embryonic day E6.5–E7.5 in mouse embryos). However, at E7.5, RA synthesis begins specifically in the posterior paraxial mesoderm (presomitic mesoderm). From E7.5 to E8.5, mesoderm provides RA in a paracrine fashion to nearby neuroectoderm fated to become hindbrain and spinal cord, but RA does not reach the forebrain or midbrain. RA regulation of gene expression in the hindbrain helps define its patterning along the anteroposterior axis into discrete units called rhombomeres. In addition, RA control of gene expression in spinal cord neuroectoderm is required for dorsoventral patterning events leading to formation of motor neurons ventrally. Thus, RA is required for patterning of the posterior CNS (hindbrain and spinal cord), but not the anterior CNS (forebrain and midbrain).

Spatiotemporal Aspects of RA Signaling in Early Mouse Embryos

The establishment of when and where RA signaling occurs in mouse embryos has been instrumental in defining the role of RA in neural patterning. One approach has been to establish the identity and expression patterns of RA receptors and enzymes that may synthesize RA. RA serves as a ligand for three nuclear RA receptors (RARs) that bind DNA as heterodimers with retinoid X receptors (RXR). Under physiological conditions (normal dietary sources of vitamin A), RA is found only in the form known as all-*trans*-RA, which binds to RAR but not RXR, but under pharmacological conditions (vitamin A excess), the isomer 9-*cis*-RA is formed and can bind to both RAR and RXR. However, several studies have demonstrated that physiological RA signaling requires only binding of all-*trans*-RA to the RAR component of RAR/RXR heterodimers. During early stages of mouse development, expression of RARα and RARγ is widespread throughout the developing CNS, but RARβ expression is limited to hindbrain and spinal cord. Thus, RA signaling could potentially occur throughout the CNS because of the presence of RARα and RARγ. However, the location of RA signaling will also depend on the location of RA synthesis, which is not as widespread as RAR expression.

RA synthesis is a two-step process in which retinol is reversibly oxidized to retinaldehyde, followed by irreversible oxidation of retinaldehyde to RA. Several members of the alcohol dehydrogenase and short-chain dehydrogenase/reductase enzyme families catalyze oxidation of retinol to retinaldehyde in many tissues in an overlapping fashion such that it is essentially ubiquitous. In contrast, oxidation of retinaldehyde to RA is carried out by three nonoverlapping, tissue-specific retinaldehyde dehydrogenases encoded by *Raldh1*, *Raldh2*, and *Raldh3* (*Aldh1a1*, *Aldh1a2*, and *Aldh1a3* in the Mouse Genome Informatics database). These three *Raldh* genes are conserved in human, mouse, and several lower vertebrates. It has been demonstrated that *Raldh* genes determine where RA signaling occurs as mouse embryos carrying the *RARE–lacZ* RA-reporter transgene plus a null mutation of *Raldh1*, *Raldh2*, or *Raldh3* selectively lose *RARE–lacZ* expression in the location where that *Raldh* gene is normally expressed. *RARE–lacZ* expression is first observed at E7.5 in the posterior mesoderm (where *Raldh2* is expressed), as well as the adjacent posterior neuroectoderm and endoderm. As *Raldh1* and *Raldh3* are not expressed until E8.5 and E9.5, respectively, only *Raldh2* is involved in generating RA for early CNS patterning. Also, studies of null mutant embryos indicate that *Raldh1* and *Raldh3* generate RA for eye morphogenetic movements but not for CNS patterning.

Raldh2 is the only source of RA for CNS patterning, but not the only determinant of where RA signaling can occur. Selective RA degradation by P450 enzymes encoded by the *Cyp26* family has been found to limit how far anteriorly and posteriorly RA can travel from its site of synthesis by *Raldh2* in the posterior mesoderm (presomitic and somitic). Thus, RA signaling does not occur throughout the CNS during early neural patterning but only in the hindbrain and spinal cord (**Figure 1**).

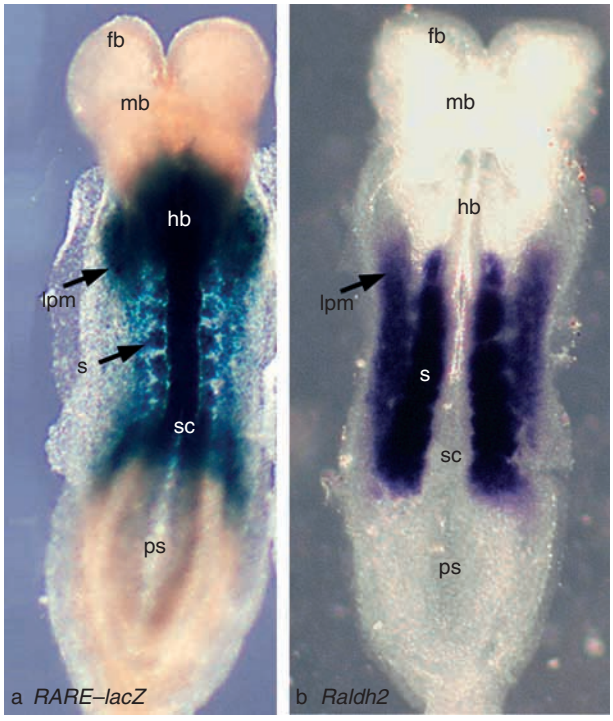

Figure 1 Retinoic acid (RA) signaling in early mouse embryos. (a) The *RARE–lacZ* RA-reporter transgene demonstrates that RA signaling activity in the central nervous system (CNS) at the six-somite stage is limited to the posterior hindbrain and spinal cord; no activity is observed in the forebrain or midbrain. (b) The only source of RA for the CNS at the six-somite stage is *Raldh2* expressed in the somitic and lateral plate mesoderm; *Raldh2* functions in a paracrine fashion by synthesizing RA in mesoderm which diffuses to the hindbrain and spinal cord neuroectoderm. fb, forebrain; hb, hindbrain; lpm, lateral plate mesoderm; mb, midbrain; ps, primitive streak; s, somite; sc, spinal cord.

RA Is Not Required for Neural Induction

In amniote vertebrate embryos, the primitive ectoderm (epiblast) consists of a pluripotent embryonic stem cell population which differentiates during primitive streak formation (gastrulation) to produce the three primary germ layers (embryonic ectoderm, mesoderm, and endoderm). As development proceeds, the primitive streak stem cell zone regresses posteriorly, and ectodermal cells emerging from the primitive streak differentiate into either neuroectoderm or epidermis in the process of neural induction. Neural induction has been found to require bone morphogenetic protein (BMP) antagonists produced in underlying mesoderm generated in the primitive streak as well as fibroblast growth factor (FGF) produced in the primitive streak. As the body axis extends, anterior neuroectoderm (forebrain/midbrain) becomes differentiated from posterior neuroectoderm (hindbrain/spinal cord) at an early stage. Posterior expansion of the hindbrain and loss of the forebrain on administration of exogenous RA to amphibian, avian, and mammalian embryos suggested that RA may be a factor normally needed to define the posterior CNS. However, as RA treatment of embryonic stem cells stimulates differentiation to neuroectoderm, it was unclear for some time whether RA is normally needed for neural induction. As ectoderm emerges from the primitive streak, *Sox1* and *Sox2*, encoding high-mobility group transcription factors, are expressed in ectoderm that has undergone neural induction to form neuroectoderm; expression occurs all along the CNS, from the forebrain to the spinal cord. *Sox1* and *Sox2* expression differentiates neuroectoderm from nonneural ectoderm, which will form epidermis. Studies on *Raldh2*$^{-/-}$ embryos has demonstrated that expression of *Sox1* and *Sox2* is not changed in the absence of RA signaling. Thus, although exogenous RA treatment has been reported to increase *Sox1* expression in mouse embryonic stem cells, endogenous RA is not required for normal induction of *Sox1* in the mouse embryo neural plate. These findings indicate that RA is not required to generate neuroectoderm. RA does, however, play a role in patterning the neuroectoderm as discussed in the section titled 'Role of RA in spinal cord dorsoventral patterning.'

Role of RA in Hindbrain Anteroposterior Patterning

During early CNS development, the hindbrain is patterned along the anteroposterior axis into segments known as rhombomeres. Studies on vitamin A-deficient quail embryos reported that a loss of the precursor for RA results in loss of rhombomeres 4–8 (r4–r8) and that the remaining hindbrain consisted of an enlarged r1–r3 joined to the anterior spinal cord. A similar phenotype was also observed in *Raldh2*$^{-/-}$ mouse embryos, vitamin A deficient rat embryos, and chick embryos treated with an RA receptor antagonist. Markers of r4 were also lost, including expression of *Hoxb1* encoding a homeobox gene and *Cyp26c1* encoding an RA-degrading enzyme.

Neural expression of *Hox* genes is limited to the posterior CNS, where these genes play essential roles in rhombomere formation. Several members of the vertebrate *Hox* gene family are direct targets of RA signaling during hindbrain development, including *Hoxb1*, whose anterior-most expression domain is in r4, located in the middle of the hindbrain. The mouse *Hoxb1* gene is regulated by an RA response element (RARE) located 3′ to the promoter that is required for early widespread induction in the posterior hindbrain up to the presumptive r3–r4 boundary, as well as another RARE located 5′ to the promoter

that has been demonstrated to be required for repression of *Hoxb1* in r3 and r5 to provide restricted expression in r4. However, *Hoxb1* r4 expression is also repressed by the homeodomain protein encoded by *vHnf1* (*Tcf2*), which is expressed in the spinal cord and posterior hindbrain up to the r4–r5 boundary. It is interesting that *vHnf1* is also inducible by RA. Studies using the *RARE–lacZ* RA reporter transgene have shown that rhombomere 3 is the anterior-most location of RA signaling during early CNS patterning but that RA activity is very transient in r3 and r4, lasting only a few hours. Transient RA activity in r3 and r4 has been found to be due to expression of RA-degrading P450s in the hindbrain. Three RA-degrading P450s encoded by *Cyp26a1*, *Cyp26b1*, and *Cyp26c1* are expressed in dynamic patterns during hindbrain development. It is now clear that the RA signal initially travels from the presomitic mesoderm (where *Raldh2* is expressed) to r3, forming a boundary next to the r2 expression domain of *Cyp26a1*. After *Hoxb1* induction, the RA boundary quickly shifts to r4–r5, coincident with induction of *Cyp26c1* in r4. Thus, during a brief period in development, RA is present in r4 to induce *Hoxb1*. Analysis of *Raldh2*$^{-/-}$ embryos has provided support for a direct role of endogenous RA signaling in *Hoxb1* induction up to r4 and repression in r3–r5 through earlier described 3′ and 5′ RAREs. Also, *Raldh2*$^{-/-}$ embryos have demonstrated that endogenous RA induces *Cyp26c1* in r4 as well as *vHnf1* posterior to the r4–r5 boundary. As *vHnf1*$^{-/-}$ embryos exhibit ectopic expression of *Hoxb1* in r5, this has provided evidence for an indirect role of RA in *Hoxb1* repression via RA induction of *vHnf1*.

It is important to note that these studies have demonstrated the existence of shifting RA boundaries in the hindbrain, as opposed to an RA gradient. *Raldh2* and *Cyp26* generate at least two distinct boundaries of RA activity along the anteroposterior axis of the hindbrain, first at r2–r3 and then at r4–r5, such that r3–r4 receives a short pulse of RA and r5–r8 receives a long pulse of RA. These two pulses of RA activity provide patterning by establishing expression of *Hoxb1* and *vHnf1* on opposite sides of the r4–r5 boundary.

Role of RA in Spinal Cord Dorsoventral Patterning

RA is required for early dorsoventral patterning of the spinal cord to generate motor neuron progenitors and later for specifying motor neuron subtype identity. RA also plays a role in assignment of positional identity of motor neurons along the anteroposterior axis of the spinal cord to generate unique neurons that innervate forelimbs, trunk, and hindlimbs. Anteroposterior effects of RA along the spinal cord are mediated primarily by regulation of *Hox* genes, similar to the role of RA in the hindbrain. In contrast, dorsoventral patterning of the spinal cord involves several different transcription factors, including the homeobox transcription factors *Pax6* and *Nkx6.1*, which are expressed dorsally and ventrally, respectively. In the region where overlapping expression of *Pax6* and *Nkx6.1* occurs (future ventral horns), the basic helix–loop–helix transcription factor *Olig2* is expressed. Genetic studies in the mouse have demonstrated that *Olig2* is required for spinal cord motor neuron differentiation. Secreted signals important for establishing spinal cord dorsoventral patterning include *Shh* expressed ventrally, *Bmp* expressed dorsally, and RA synthesized in the adjacent somitic mesoderm. A lack of RA synthesis in *Raldh2*$^{-/-}$ mouse embryos results in a loss of *Pax6* and *Olig2* expression in the spinal cord, whereas *Nkx6.1* expression remains, suggesting that *Nkx6.1* does not require RA for expression. This requirement for RA during spinal cord patterning is conserved between mammalian and avian embryos, and studies on avian embryos suggest that *Pax6* and *Olig2* are both direct targets of RA action. RA is thus needed as an additional trigger, along with *Pax6* and *Nkx6.1*, to allow induction of *Olig2*, which then stimulates undifferentiated spinal cord neuroectoderm to acquire a ventral motor neuron cell fate.

Early studies in avian embryos suggested that *Raldh2* may be responsible for synthesizing RA needed for spinal cord motor neuron differentiation as *Raldh2* is expressed in somitic mesoderm adjacent to the neural tube. Genetic studies have now shown that *Raldh2* is responsible for all RA detected posteriorly in mouse embryos at E8.5 and that *Raldh2* is necessary for *Olig2* expression in the spinal cord. The extent of RA distribution from sites of *Raldh2* expression during spinal cord development has been examined in mouse embryos. *RARE–lacZ* RA–reporter expression is observed in trunk mesoderm and throughout the dorsoventral axis of the spinal cord, demonstrating that RA synthesized by *Raldh2* in the somites can travel to all portions of the spinal cord. These findings provide evidence that RA functions exclusively in a paracrine fashion during early spinal cord patterning. As RA is not localized to any particular region of the spinal cord, this may explain its ability to induce not only *Olig2*, which is limited to the developing motor neuron domain located ventrally, but also *Pax6*, which is expressed more widely across the dorsoventral axis of the neural tube.

Development of *Raldh2*$^{-/-}$ embryos can be substantially rescued by a low-dose maternal dietary RA

supplement that returns embryonic RA levels to approximately the same level observed in wild-type embryos. Such treatment can rescue *Olig2* expression in *Raldh2*$^{-/-}$ embryos. It is interesting that *RARE–lacZ* expression in rescued *Raldh2*$^{-/-}$ embryos occurs throughout the spinal cord and posterior hindbrain, similar to wild-type, but unlike wild-type, there was no detection of *RARE–lacZ* in the somitic mesoderm, which is normally the source of RA. This phenomenon is clearly dose dependent as large doses of exogenous RA will induce *RARE–lacZ* in all cells of the embryo. These findings suggest that the low dose of exogenous RA entering the embryo was not distributed evenly. RA may be preferentially sequestered in the posterior CNS as it is known that cellular RA binding protein II (*Crabp2*) is expressed highly in the spinal cord and hindbrain but not in mesoderm. As the normal site of RA synthesis in the somitic mesoderm did not exhibit RA signaling activity in *Raldh2*$^{-/-}$ embryos under these rescue conditions, it may be that RA does not need to perform an additional function in the mesoderm during spinal cord dorsoventral patterning; the somitic mesoderm is simply acting as a source of RA. The rationale for such RA signaling may be to coordinate motor neuron differentiation with mesodermal differentiation along the anteroposterior axis as cells emerge from the primitive streak.

Many of the above observations were made possible through the use of mouse embryos carrying the *RARE–lacZ* transgene, which enables one to follow where endogenous RA signaling activity occurs in wild-type embryos and where RA signaling occurs in *Raldh2*$^{-/-}$ embryos rescued by maternal RA treatment. Such experiments have not been possible in avian, amphibian, or fish embryos because of lack of a transgene that can serve as an RA reporter during the early stages of neural development. This highlights the importance of the mouse in studies designed to understand the mechanism of RA action in the CNS.

RA Antagonism of *Fgf8* Expression in the Primitive Streak Controls Posterior Patterning

FGF signaling controlled by *Fgf8* is particularly important in the mouse as it is necessary for neural induction in the posterior CNS as well as generation of mesoderm during gastrulation. Early studies suggested that presomitic mesoderm emerging from the primitive streak provides another signal needed for differentiation of the posterior CNS, that is, a signal that antagonizes posterior FGF signaling mediated by *Fgf8*. Recent studies in avian, mouse, and zebra fish embryos indicate that RA generated by *Raldh2* is the presomitic mesodermal factor that antagonizes *Fgf8* action. The concept of embryonic tissue differentiation occurring in regions of opposing RA and FGF signals was originally proposed in studies of proximodistal outgrowth of chick limb buds where *Raldh2* and *Fgf8* also function to generate opposing RA and FGF signals.

Fgf8 is normally expressed in the primitive streak, but its expression is extinguished anteriorly as cells exit the streak. Studies in avian, mouse, and zebra fish embryos indicate that RA functions as an antagonist of posterior *Fgf8* expression. In particular, RA functions to limit the anteroposterior extent of *Fgf8* expression and confines it to the primitive streak or tailbud. Whether this reflects a direct effect of RA on the *Fgf8* gene remains to be determined, but a conserved RARE is found in the promoter regions of human, mouse, and rat *Fgf8* genes. This function is consistent with the observation that *RARE–lacZ* expression in mouse embryos is high in tissues anterior to the primitive streak, then progressively much weaker in the primitive streak itself. Primitive neuroectodermal cells thus exist in a zone of high FGF8 and low RA signaling, but on exiting the primitive streak anteriorly, these cells enter a zone of high RA emanating from the directly adjacent somitic mesoderm, which downregulates *Fgf8* and stimulates neuronal differentiation.

RA also controls left–right patterning of somite pair formation, and this is proposed to function via antagonism of *Fgf8* expression in neuroectoderm. RA synthesized in presomitic mesoderm by *Raldh2* is required to maintain bilateral symmetry between the left and right somite columns; for instance, loss of RA can result in embryos displaying 11 somites on the left side but only nine somites on the right side. Presomitic mesoderm in mouse *Raldh2*$^{-/-}$ embryos displays left–right asymmetric expression of *Hes7* and *Lfng*, required for periodic mesodermal segmentation, suggesting that a loss of RA allows left–right asymmetry to occur in presomitic mesoderm, where it normally does not occur. On the other hand, lateral plate mesoderm (where left–right asymmetry normally occurs) still maintains left–right asymmetry of key genes, including *Nodal* and *Pitx2* in *Raldh2*$^{-/-}$ embryos. RA thus acts as a buffer to prevent left–right asymmetry from occurring in presomitic mesoderm. In the 'clock and wavefront' model of somitogenesis, a moving wavefront of *Fgf8* gene expression in the primitive streak regresses posteriorly as the body axis extends, and mesodermal segmentation occurs just anterior to the *Fgf8* expression domain. FGF8 signaling also plays a required role in the node, a structure located anteromedially in the primitive streak that produces

left–right asymmetry in the lateral plate mesoderm. As *Fgf8* mRNA is not detected in the node, nearby cells in the epiblast ectoderm that do express *Fgf8* mRNA evidently secrete FGF8 that travels to the node. A hypothesis for how RA functions to prevent somite left–right asymmetry states that by limiting the anterior extent of *Fgf8* expression in the ectoderm, RA ensures that FGF8 signaling in the node is not excessive.

RA signaling monitored with *RARE–lacZ* exhibits no left–right difference in presomitic or lateral plate mesoderm at early somite stages, and RA activity extends from the posterior neural plate into node ectoderm, where it meets the *Fgf8* expression domain in the epiblast (primitive ectoderm) and then fades out. Although RA activity exists in presomitic mesoderm, it is absent in node mesoderm. Thus, it appears that RA functions in node ectoderm during somite left–right patterning. *Raldh2*$^{-/-}$ embryos exhibit a bilateral anterior shift of *Fgf8* expression into node ectoderm and neural plate, where it normally is not expressed. It is important to note that a loss of RA does not result in an anterior advance of *Fgf8* expression in presomitic mesoderm (which normally expresses *Fgf8* at a lower level than that seen in primitive ectoderm), and no ectopic *Fgf8* expression is observed in node mesoderm. These findings have provided evidence that the site of RA action during control of left–right asymmetry is the node ectoderm and adjacent neural plate.

Somite left–right patterning defects in *Raldh2*$^{-/-}$ embryos can be rescued with low-dose maternal dietary RA supplementation. It is interesting that RA provided to only the six-somite stage rescues somite patterning defects until at least the 25-somite stage. Thus, RA is needed only during a short time frame early in somitogenesis. Studies on rescued *Raldh2*$^{-/-}$ embryos carrying *RARE–lacZ* have provided further evidence that the site of RA action during control of left–right asymmetry is the node ectoderm and adjacent neural plate. Examination of such embryos at somite stages 2–10 revealed no *RARE–lacZ* expression in somitic or presomitic mesoderm (normally the source of RA), but *RARE–lacZ* was expressed in posterior neural plate and node ectoderm. As mentioned earlier, selective action of RA in neuroectodermal tissues may be due to expression of *Crabp2* that functions to sequester RA.

The limited time frame needed for RA action to correct somite left–right asymmetry is likely due to a transient RA requirement in the node as the node regresses after the ten-somite stage. At the two-somite stage, *Raldh2* is expressed in posterior presomitic mesoderm lying lateral to the node, whereas at the ten-somite stage, *Raldh2* is no longer expressed in presomitic mesoderm but remains in somites. Consistent with this observation, *RARE–lacZ* expression is present in the posterior neuroectoderm and node ectoderm during early stages but later retracts anteriorly such that by the ten-somite stage, there is no longer RA activity in the posterior neuroectoderm adjacent to the presomitic mesoderm; RA activity remains in neuroectoderm adjacent to the somites that have formed. These findings provide evidence that after the node regresses, RA is no longer needed to limit *Fgf8* expression posteriorly.

RA Acts in a Paracrine Fashion to Regulate Posterior CNS Patterning

As *Raldh2* and *RARE–lacZ* are normally expressed in presomitic and somitic mesoderm, it was initially presumed that RA functions locally in these mesodermal tissues during somite formation as well as in the adjacent neuroectoderm, where *RARE–lacZ* expression is also observed. However, studies on RA-rescued *Raldh2*$^{-/-}$ embryos demonstrate that RA needs to function only in the adjacent node ectoderm and neural plate to regulate early somite patterning, plus in the neuroectoderm of the hindbrain and spinal cord to regulate posterior CNS patterning. Thus, RA generated by *Raldh2* in mesoderm may function solely as a paracrine signal as no autocrine function has been uncovered.

See also: Floor Plate Patterning of Ventral Cell Types: Ventral Patterning; Hox Gene Expression; Morphogens: History; Motor Neuron Specification in Vertebrates; Segmentation: Segmental Boundaries, Establishment and Morphogenesis (EPH); Segmentation: Spinal Cord Segmentation and A–P Somite Patterning.

Further Reading

DelCorral RD, Olivera-Martinez I, Goriely A, Gale E, Maden M, and Storey K (2003) Opposing FGF and retinoid pathways control ventral neural pattern, neuronal differentiation, and segmentation during body axis extension. *Neuron* 40: 65–79.

Dupé V and Lumsden A (2001) Hindbrain patterning involves graded responses to retinoic acid signalling. *Development* 128: 2199–2208.

Gavalas A and Krumlauf R (2000) Retinoid signalling and hindbrain patterning. *Current Opinions in Genetics and Development* 10: 380–386.

Liu JP, Laufer E, and Jessell TM (2001) Assigning the positional identity of spinal motor neurons: Rostrocaudal patterning of Hox-c expression by FGFs, Gdf11, and retinoids. *Neuron* 32: 997–1012.

Maden M (2002) Retinoid signalling in the development of the central nervous system. *Nature Reviews Neuroscience* 3: 843–853.

Maden M, Gale E, Kostetskii I, and Zile MH (1996) Vitamin A-deficient quail embryos have half a hindbrain and other neural defects. *Current Biology* 6: 417–426.

Mic FA, Molotkov A, Benbrook DM, and Duester G (2003) Retinoid activation of retinoic acid receptor but not retinoid X receptor is sufficient to rescue lethal defect in retinoic acid synthesis. *Proceedings of the National Academy of Sciences of the United States of America* 100: 7135–7140.

Molotkov A, Molotkova N, and Duester G (2006) Retinoic acid guides eye morphogenetic movements via paracrine signaling but is unnecessary for retinal dorsoventral patterning. *Development* 133: 1901–1910.

Molotkova N, Molotkov A, Sirbu IO, and Duester G (2005) Requirement of mesodermal retinoic acid generated by *Raldh2* for posterior neural transformation. *Mechanisms of Development* 122: 145–155.

Niederreither K, Vermot J, Schuhbaur B, Chambon P, and Dollé P (2000) Retinoic acid synthesis and hindbrain patterning in the mouse embryo. *Development* 127: 75–85.

Novitch BG, Wichterle H, Jessell TM, and Sockanathan S (2003) A requirement for retinoic acid-mediated transcriptional activation in ventral neural patterning and motor neuron specification. *Neuron* 40: 81–95.

Sirbu IO and Duester G (2006) Retinoic acid signaling in node ectoderm and posterior neural plate directs left-right patterning of somitic mesoderm. *Nature Cell Biology* 8: 271–277.

Sirbu IO, Gresh L, Barra J, and Duester G (2005) Shifting boundaries of retinoic acid activity control hindbrain segmental gene expression. *Development* 132: 2611–2622.

Sockanathan S and Jessell TM (1998) Motor neuron-derived retinoid signaling specifies the subtype identity of spinal motor neurons. *Cell* 94: 503–514.

Sockanathan S, Perlmann T, and Jessell TM (2003) Retinoid receptor signaling in postmitotic motor neurons regulates rostrocaudal positional identity and axonal projection pattern. *Neuron* 40: 97–111.

Vermot J, Schuhbaur B, and LeMouellic H (2005) Retinaldehyde dehydrogenase 2 and Hoxc8 are required in the murine brachial spinal cord for the specification of Lim1+ motoneurons and the correct distribution of Islet1+ motoneurons. *Development* 132: 1611–1621.

Wilson L and Maden M (2005) The mechanisms of dorsoventral patterning in the vertebrate neural tube. *Developmental Biology* 282: 1–13.

Retinomotor Movements

B Burnside, University of California at Berkeley, Berkeley, CA, USA
C King-Smith, Joseph's University, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Nature and Occurrence

In many species, photoreceptors and retinal pigment epithelium (RPE) undergo positional rearrangements, called retinomotor movements, in response to changes in ambient light conditions and to circadian signals. Retinomotor movements are achieved by elongation and contraction of rod and cone photoreceptors and by migration of melanin pigment granules (melanosomes) within the apical projections of the RPE (Figure 1). In the light, cones shorten, rods elongate, and pigment granules disperse into the long apical projections of the RPE cells, thus positioning the cones closest to incoming light and burying the rod outer segments in the shielding pigment of the RPE. In the dark, these movements are reversed: RPE pigment granules aggregate to expose the rod outer segments and rods shorten to move their outer segments in front of those of the elongating cones. Thus in the dark (night) the retina is morphologically optimized for dim light reception with rods contracted to position their outer segments first in line for detection of incoming light, while in the light (day) the retina is optimized for bright light vision with cone outer segments first in line and rod outer segments protected from bleach by RPE pigment.

Though the cones of some fish may elongate and contract as much as 80 μm, most cone length changes range from 15 to 40 μm. Rod length changes are usually more modest. Since rates of elongation and contraction are in the 1–4 $\mu m\ min^{-1}$ range, the transition from light-adapted to dark-adapted positions and vice versa are relatively time consuming, requiring as much as 60 min in species with extensive movements. This slowness may explain why mammals deploy the more agile strategy of pupillary movement for adapting the eye to differing light intensities. Though faster, the pupillary movement strategy forfeits the use of the entire light capture area of the retina for both photopic (cone) and scotopic (rod) vision.

Because these dramatic morphological changes can be easily detected by light microscopy of sectioned retinas, many of the earliest studies of retinal morphology concerned the properties and species distributions of retinomotor movements, the first by Kühne in 1856. The late eighteenth and early nineteenth century produced an astonishingly large literature describing retinomotor movements among the vertebrate groups. Modest or nonexistent in mammals, urodeles, and reptiles, retinomotor movements are more robust and common in anurans and birds, and most elaborate and widely distributed in fishes with duplex retinas. In the pure cone retinas of early fish larvae and in the pure rod retinas of benthic fishes, retinomotor movements are absent.

Retinomotor movements are not only regulated by light but also by an endogenous circadian rhythm. When animals are maintained in constant darkness, retinomotor movements persist at expected dawn and dusk. In constant illumination, the cycles of retinomotor movements are abolished. In most species circadian cone movements mimic those of the diurnal cycle at reduced amplitude, but in some species they are more variable. Circadian retinomotor movements of cones have been observed in most fish species studied; circadian retinomotor movements of rods and RPE are considerably less common. In all species so far studied, cones contract to assume their light-adaptive retinomotor positions in anticipation of expected dawn; thus for the fish in their natural environment, circadian signals appear to be more important regulators of cone contraction than light.

Mechanisms of Force Production for Retinomotor Movements

Many of the early studies of retinomotor movement grappled with whether retinomotor movements required energy or were passive accommodations of other forces. Subsequent studies with isolated cells have made it clear that rod, cone, and RPE retinomotor movements can each occur in the absence of other cell types and indeed in the absence of the rest of the retina; thus, force production for cell elongation, contraction, and for pigment migration appears to be autonomous for each cell type.

The Photoreceptor Cytoskeleton and Force Production for Elongation and Contraction

Contraction and elongation take place in the inner segment, that is, the part of the photoreceptor between the photopigment-bearing outer segment and the junctional complexes of the outer limiting membrane, where the photoreceptors are attached to the supportive Müller cells of the retina. Thus, the motile portions of the photoreceptors protrude into the subretinal space between the retina and the RPE,

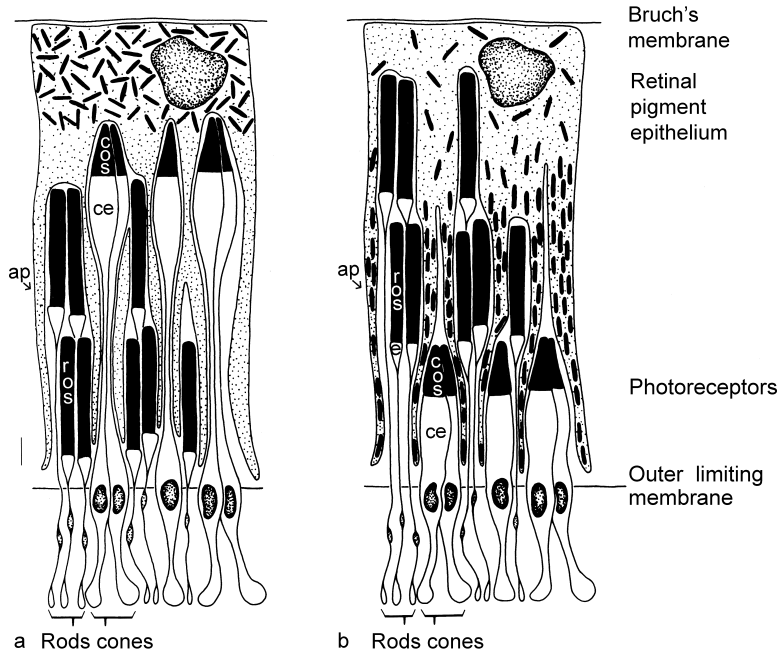

Figure 1 Retinomotor positions of rods, cones, and RPE in light- and dark-adapted fish retinas (drawn from the green sunfish *Lepomis cyanellus*). (a) In dark-adapted retinas, cone myoids are long, rod myoids are short, and pigment granules are aggregated to the base of the RPE cell near Bruch's membrane. (b) In light-adapted retinas, cone myoids are short, rod myoids are long, and pigment granules have migrated into the long apical projections (ap) of the RPE. cos, cone outer segment; ce, cone ellipsoid; ros, rod outer segment; e, rod ellipsoid. Scale bar = 10 μm. Reproduced from Burnside B and Nagle B (1983) Retinomotor movements of photoreceptors and retinal pigment epithelium: Mechanisms and regulation. In: Osborne N and Chader GJ (eds.) *Progress in Retinal Research,* vol. 2, pp. 67–109. Oxford: Pergamon, with permission from Elsevier.

where they interdigitate with the long apical projections of the RPE cells (Figures 1 and 2). The interphotoreceptor matrix filling the subretinal space thus lubricates, or at least permits, the shear produced by rods and cones moving in opposite directions. The photoreceptor inner segment is composed of the distal ellipsoid, containing a mass of aggregated mitochondria, and the more proximal myoid, so named because of its ability to contract in lower vertebrates. Almost all of the retinomotor shape change is mediated by changes in length of the myoid. In fish retinas, cone photoreceptors are generally considerably larger than rod photoreceptors and exhibit greater excursions of retinomotor movement.

Rod and cone photoreceptors have a relatively consistent cytoskeletal organization characterized by axially aligned actin filaments and microtubules (Figure 2). From the distal end of the inner segment, microvillus-like calycal processes cup the base of the outer segment. Paraxial bundles of cross-linked actin filaments form the cores of the calycal processes, continue through the peripheral cytoplasm of the ellipsoid and then splay out in the myoid into a less tightly cross-linked, but still longitudinally oriented actin filament array. Thus, the contractile myoid region is rich in paraxially aligned actin filaments. These polarized actin filaments of the ellipsoid and myoid are oriented with fast-growing (plus) ends toward the outer segment. The polarity of actin filaments plays a critical role in actin-based motility, since each type of myosin motor protein is constrained to walk in only one direction along the actin filament. Thus a plus-end directed myosin motor would walk along the photoreceptor inner segment actin bundles toward the outer segment, while a minus-end directed motor would walk toward the nucleus. Actin filaments are also highly concentrated at the outer limiting membrane, where they form the submembrane actin densities of the adherens junctions found there.

The microtubule cytoskeleton of the photoreceptor consists of two distinct microtubule populations with different properties: the highly stable $(9+0)$ ciliary axoneme of the outer segment, and the more labile paraxially aligned cytoplasmic microtubules of the photoreceptor cell body. Relatively parallel microtubules populate the sleeve of cytoplasm surrounding the mitochondrial mass in the ellipsoid and extend throughout the myoid and axon. As with actin, microtubule polarity and the orientation of the

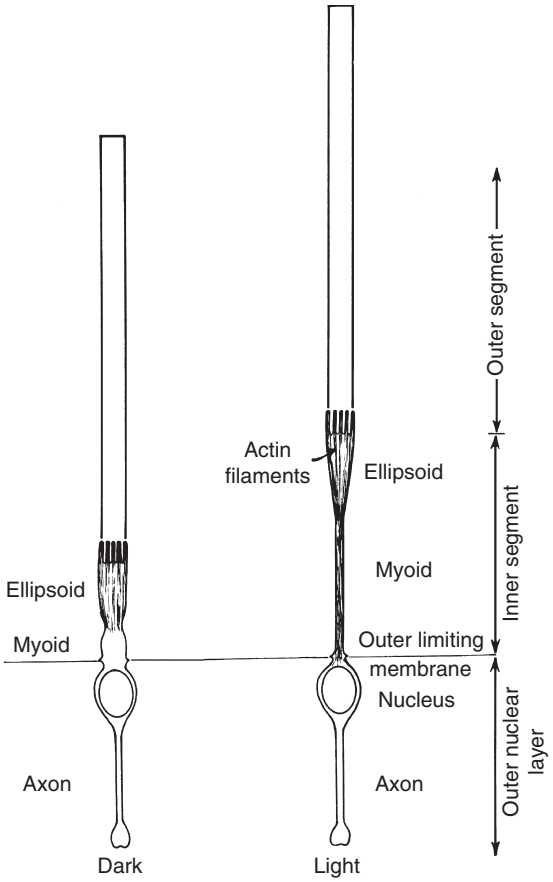

Figure 2 Schematic illustration of cell shape and thin filament distribution in dark- and light-adapted teleost rods. Not shown are ellipsoid and myoid microtubules that are also present in both states. Scale bar = 10 μm. Reproduced from *Journal of Cell Biology*, 1981, 89, pp. 517–524, by copyright permission of The Rockefeller University Press.

microtubules within a cell play critical roles in microtubule-based motility, since microtubule-based motors are also constrained to walk one direction along a microtubule. The polarity of the inner segment and axonal microtubules is quite uniform, with virtually all microtubules oriented with their slow-growing (minus) ends toward the outer segment. Thus in a photoreceptor, a minus-end directed microtubule motor like dynein would walk toward the outer segment, while a plus-end directed microtubule motor like kinesin would walk toward the synapse.

Cytoskeletal inhibitors have been used to ascertain the roles of microtubules and actin filaments in teleost rod and cone retinomotor movements. In cones, dark-adaptive myoid elongation is microtubule dependent while light-adaptive cone myoid contraction is actin dependent. Myosin II, a cytoplasmic myosin capable of forming bipolar filaments, is highly concentrated in cone myoids, suggesting that a muscle-like sliding filament model of actin–myosin interaction might mediate cone contraction. In rods, myoid elongation and myoid contraction are both actin dependent; neither elongation nor contraction requires intact microtubules. Actin inhibitor studies of rod elongation suggest that force production for rod elongation is provided by actin assembly and/or cross-linking.

The RPE Cytoskeleton and Force Production for Pigment Granule Dispersion and Aggregation

The advancement and retraction of shielding pigment in the RPE results from the translocation of individual melanin pigment granules within the long apical projections of the RPE cells. The membrane-bound melanosomes migrate out into the long apical projections in the light and aggregate back into the RPE cell body in the dark. The apical projections extend from the RPE cell apical surface into the subretinal space and interdigitate with the photoreceptor inner and outer segments, thus allowing the dispersed melanosomes to shield and protect rod outer segments from bleach in bright light. RPE apical projections may be very long (up to 100 μm in some fish species), and they remain extended throughout the phases of melanosome migration.

Apical projections of RPE cells contain both actin filaments and microtubules. Microtubule disruption by intraocular injection of inhibitors *in vivo* blocks motility of melanosomes both in the cell body and in the apical projections, suggesting that melanosome movement is microtubule dependent. However, when inhibitor studies are carried out using dissociated RPE cells in culture, microtubule disruption has no effect on melanosome aggregation or dispersion. Presumably, this discrepancy results from the need for microtubules to provide structural support *in vivo* for the maintenance of the apical projections, or to maintain patent tracks or pathways through which the melanosomes can move. In the isolated cells on the other hand, the attachment of the projections to the substrate likely provides the structural support necessary to keep apical projections intact enough to permit pigment granule translocation. The isolated cell studies therefore demonstrate that microtubules contribute to the structural integrity of RPE apical projections, but are not required for force production for translocation.

In contrast, disruption of the actin cytoskeleton in dissociated RPE cells reversibly blocks both aggregation and dispersion of melanosomes, and prevents maintenance of full aggregation or full dispersion both *in vivo* and in isolated cells. These studies show that actin filaments are necessary and sufficient

for both aggregation and dispersion, while microtubules are not required for force production for either aggregation or dispersion. The actin dependence of melanosome translocation in RPE cells suggests that myosin motor proteins are likely to play a role in force production. Myosin VIIa has been immunolocalized to RPE of human, rodent, and fish RPE. Using *in vitro* motility assays it has been shown that teleost RPE melanosomes do support plus-end directed, actin-dependent motility, and that myosin VIIa is present and enriched on melanosomal membranes.

While the *in vitro* assays demonstrate that teleost melanosomes possess myosin motors capable of moving them along actin filaments, it is not clear whether these motors translocate melanosomes during aggregation, dispersion, or both. Rates of motility observed in assays *in vitro* spanned a wide range, raising the possibility that more than one class of myosin motor is present on melanosomes. Myosin Va, which has been shown to mediate melanosome translocation in melanocytes and amphibian dermal chromatophores, has also been identified in the melanosome fraction from porcine and rodent RPE.

Intracellular Regulation of Retinomotor Movements: The Role of cAMP

Numerous observations suggest that intracellular cyclic adenosine monophosphate (cAMP) plays a critical role in the intracellular regulation of retinomotor movements in both photoreceptors and RPE cells. Experimental studies based on intraocular injections and cultured retinas show that a variety of treatments that elevate cAMP (but not cyclic guanosine monephosphute (cGMP)) produce retinomotor positions characteristic of darkness (night) in teleost rods, cones, and RPE cells and in frog cones. Similar effects were obtained using isolated RPE cells, permeabilized cones, and isolated fragments (inner/outer segments) of rods and cones, suggesting that increased cytoplasmic cAMP levels have a local and direct effect in each cell type. These findings are consistent with other reports suggesting that elevated cAMP levels in the dark are a general feature of cyclic diurnal and circadian signaling in vertebrate photoreceptors. For example, light and circadian control of retinal melatonin release is closely correlated with changes in intracellular cAMP levels. cAMP regulates melatonin synthesis by influencing the activity of the rate-limiting enzyme in melatonin production, which is located in photoreceptors. Retinal melatonin release is elevated at night in both cyclic light and constant dark conditions, consistent with higher dark levels of cAMP in photoreceptors. Together these observations suggest that diurnal and circadian changes in cAMP levels in the inner segment mediate intracellular regulation of retinomotor movements *in vivo*.

Since inhibitors of cAMP-dependent kinase (protein kinase A (PKA)) block dark-adaptive retinomotor movements in both cone and rod isolated inner/outer segments preparations, it seems likely that cAMP mobilizes the photoreceptor cytoskeleton by protein phosphorylation. The observation that cAMP can trigger elongation in detergent-permeabilized cone models further suggests that critical targets of PKA are detergent insoluble, and thus likely to be associated with the cytoskeleton itself. Pharmacological studies of rod inner–outer segment fragment motility suggest that phosphorylation of mediator proteins by PKA is enhanced in darkness and that dephosphorylation is required for light-induced elongation. In rod inner–outer segment fragments, the two primary targets of PKA phosphorylation are phosducin and phosducin-like protein. Phosducin has many roles in the photoreceptor, including binding to $\beta\gamma$ subunits of transducin in its unphosphorylated state. It is not yet clear whether there is any causal relationship between phosducin phosphorylation and retinomotor movements.

Regulation of Photoreceptor Retinomotor Movements by Paracrine Messengers

A role for paracrine signaling in regulation of retinomotor movements is implicated by spectral sensitivity studies of cone retinomotor movements in fish and frog retinas *in vivo*. The action spectrum for triggering light adaptive cone and RPE retinomotor movements most closely fits the absorption curve of the rod photopigment, suggesting that in cones and RPE cells, these movements are not directly triggered by light but instead are activated indirectly via a rod-mediated pathway. The neuromodulator dopamine participates in this rod-mediated pathway as suggested by the ability of intraocularly injected dopamine to block light-induced cone contraction and RPE pigment granule dispersion.

Observations from studies using intraocular injection *in vivo*, isolated retinas in culture, and isolated cone inner–outer segment fragments, further suggest that dopamine plays a central role in paracrine regulation of cone retinomotor movements of fish and frogs. In all these preparations, light and/or dopamine promote cone contraction, whereas darkness or the adenylate cyclase stimulator, forskolin, promotes cone elongation. Pharmacological profiles of this dopamine effect suggest that dopamine is acting through a D4-like or other D2 family dopamine receptor.

This class of dopamine receptors inhibits adenylate cyclase and thus would be expected to lower cone cytoplasmic cAMP levels, consistent with observations that experimentally increasing cAMP produces dark-adaptive movements. In isolated cone fragments, dopamine blocks cone elongation induced by the adenylate cycle stimulator forskolin but not elongation induced by the membrane permeable cAMP analog, dibutyrl cAMP, further confirming that dopamine regulates cone movement by inhibiting adenylate cyclase.

The ability of light and dopamine to regulate retinomotor movement in isolated cone inner–outer segment fragments indicates that these agents can act directly on cones. In teleost retinas, the source of dopamine is the interplexiform cell whose projections extend into the outer plexiform layer. If the D4-like receptors are on the inner (and/or outer) segments as suggested by the cone fragment findings, released dopamine would have to diffuse over tens of microns to reach the inner segments *in vivo*. The very high affinity of D4 receptors (K_d in the nM range) may facilitate this paracrine mode of signaling.

Studies with isolated photoreceptor inner–outer segment fragments indicate that dopamine is much more effective than light in producing myoid contraction in cones, whereas light is much more effective than dopamine in triggering myoid elongation in rods. Furthermore, regulation of fish rod movement by light and circadian signals was unaffected by 6-hydroxydopamine lesion of dopaminergic interplexiform cells. These observations suggest that dopamine does not play a significant role in regulating rod retinomotor movements *in vivo*.

Light-activation of myoid elongation in teleost rod inner out segment fragments *in vitro* requires relatively high light intensities (20% bleach). The photoreceptive mechanism responsible for this activation can apparently count quanta accurately for light pulse durations up to at least 10 min, suggesting that the critical factor in light activation of rod elongation is quantum catch, rather than duration of the light stimulus. Similar high intensity thresholds and photon counting have been reported to mediate entrainment of circadian rhythms in several species. High-intensity threshold and the ability to count quanta over long durations may be common denominators of photoreceptive functions that attune the organism to the daily light-dark cycle rather than mediate vision.

That dopamine also plays a role in circadian regulation of cone retinomotor movement is suggested by results obtained by intraocular injection of dopamine agonists and antagonists in animals maintained in constant darkness (DD). Cone myoid contraction can be induced at midnight by intraocular injection of dopamine or D2 receptor agonists. Partially contracted cones of DD animals at subjective midday can be induced to contract fully by intraocular injection of dopamine or D2 agonists, or to elongate by injection of the D2 antagonist sulpiride. The predawn cone contraction observed in DD animals in response to circadian signals can be completely eliminated by intraocular injection of the D2 antagonist shortly before the expected time of light onset. These observations suggest that circadian regulation of cone myoid length is mediated by endogenous dopamine, acting via D2 family receptors. Dopamine is unlikely, however, to be the sole circadian regulator, since both light induced and circadian cone movements persist (at reduced amplitude) after lesion of interplexiform cells by 6-hydroxydopamine.

The neuromodulator adenosine can also influence cone retinomotor movements in teleosts. In isolated cone inner–outer segment fragments, myoid elongation can be induced by activation of adenosine A2 receptors and inhibited by A2 receptor antagonists. Since A2 receptors are positively coupled to adenylate cyclase, adenosine effects on cone fragments are consistent with previous findings that elevation of cAMP triggers dark-adaptive retinomotor movements. The observation that cone fragments respond to adenosine indicates that A2 receptors must be present on cone inner (and/or outer) segments. Since cones elongate at subjective dusk in fish maintained in constant darkness, and since adenosine stimulates cone elongation, adenosine could provide an endogenous circadian signal for expected dark onset. Alternatively, adenosine effects on cone movements could be a local effect, reflecting photoreceptor metabolic activity. Since an adenosine transporter is present in photoreceptors, increases in adenosine in photoreceptor inner segments might be expected to be accompanied by an increase in adenosine level in the subretinal space. Darkness is associated with increased Na^+/K^+-ATPase activity in photoreceptor inner segments. Increased Na^+/K^+-ATPase activity is likely to be accompanied by an increase in intracellular adenosine levels, and consequently, increased release of adenosine into the subretinal space. The binding of released adenosine to photoreceptor A2 receptors might then enhance dark adaptation of the photoreceptors by stimulating adenylate cyclase, thereby reinforcing the dark signal.

Regulation of Retinomotor Movements in RPE Cells by Paracrine Messengers

Spectral sensitivity studies have shown that RPE pigment migration in the frog has an action spectrum

most closely resembling that of the rod photopigment. These observations suggest that for RPE, as for cones, a rod-mediated pathway triggers light-adaptive retinomotor movement. The observation that pigment position is not affected by light in isolated sheets of teleost RPE further argues that light acts indirectly via a photoreceptor mediated, paracrine pathway.

Pigment migration can be triggered experimentally in isolated sheets of teleost RPE in culture. Dopamine and D2 family agonists induce pigment granule dispersion (the light-adaptive movement) while treatments that elevate cAMP induce pigment granule aggregation (the dark-adaptive movement). In isolated RPE cells aggregated in cAMP, pigment dispersion is induced by microinjection of PKA inhibitors, suggesting that continuous phosphorylation of PKA targets is required to maintain the aggregated state. Since aggregation can also be induced by the phosphatase inhibitor okadaic acid, it seems likely that PKA and phosphatases are simultaneously active in RPE cells, and that their relative activities are altered by light and dark signals from the retina.

Surprisingly, underivatized cAMP is just as effective as membrane-permeant cAMP analogs at activating pigment granule aggregation in isolated RPE sheets. Washout of exogenous cAMP induces dispersion. Thus, cAMP appears either to act on cell surface purinergic receptors or to enter the cell via organic ion transporters. Since ATP and adenosine are ineffective and since organic ion transport inhibitors block cAMP-but not forskolin-induced pigment aggregation, it seems likely that exogenous cAMP can actually enter the cells via organic ion channels to activate pigment aggregation. Thus, in RPE cells, cAMP might be acting not only as a second (intracellular) messenger, but as a first (paracrine) messenger to trigger retinomotor movements associated with dark adaptation. These findings indicate that an increase in cAMP level in the subretinal space in the dark would activate pigment granule aggregation in RPE cells. Cyclic AMP efflux has been shown to be associated with increased intracellular cAMP accumulation in many cell types, including pinealocytes. Thus, diurnal and circadian cycles of cAMP levels in retinal photoreceptors might be expected to produce increased cAMP efflux into the subretinal space at night. This extracellular cAMP in the subretinal space could then serve as a retina to RPE signal for darkness.

Cyclic AMP and intracellular calcium often interact in cellular signaling pathways, either antagonistically or synergistically. Melanosome migration is regulated by calcium in several types of dermal chromatophores. However, increasing extracellular or intracellular calcium has no effect on either pigment granule aggregation or dispersion in dissociated RPE cells. Furthermore, intracellular calcium levels are unaffected when RPE pigment granule motility is triggered by cAMP or cAMP washout. These studies suggest that RPE pigment granule movements are not regulated by calcium but are regulated solely by cAMP, presumably by phosphorylation and dephosphorylation of protein kinase A targets.

Functions and Significance of Retinomotor Movements

Just what role retinomotor movements play in retinal function has been a topic of inquiry and speculation for researchers for more than a century. Direct experimental demonstration of effects of retinomotor movements on vision has been difficult to achieve, primarily because it is extremely difficult to interfere with retinomotor movements without affecting other aspects of retinal function. Nevertheless, several likely functions for retinomotor movements have been suggested. Occurring generally in species that lack pupillary movements, rod and RPE retinomotor movements provide an alternative mechanism for shielding rods in bright light and fully exposing them for dim light vision. Repositioning the photoreceptors likewise provides an ingenious mechanism for optimizing space utilization by positioning the rods first in line to detect light at the focal plane across the entire retinal surface under dim light conditions and then moving cones to this position for bright light vision. These movements permit the entire retinal surface to be used alternately for rod and cone vision. Some researchers have also noted that each photoreceptor type is elongated when it is expected to be silent and have suggested that the cable properties of the elongated myoid might contribute to attenuation of the signal from outer segment to synapse. The great morphological variation in retinomotor movements perhaps reflects the selective optimization of one of these functions in different species.

Summary

Though retinomotor movements are most pronounced in lower vertebrates, their study contributes to our understanding of diurnal and circadian regulation of photoreceptor and RPE physiology in all species. Retinomotor movements provide excellent models for investigating the roles of cytoskeletal elements in cell shape change and intracellular transport. Since photoreceptors are light detectors, their

metabolism and physiology is particularly attuned to the daily cycles of light and darkness. Studies of retinomotor movements have called attention to the importance of cyclic changes in intracellular cAMP levels in the diurnal and circadian regulation of other aspects of photoreceptor physiology and metabolism. The recognition that light induction of cone and RPE movements depends on a rod-mediated paracrine pathway has contributed to recognition that dopamine plays an important role in light and circadian signaling in the retina. Since retinomotor movements exhibit properties of light- and circadian-regulation characteristic of many other aspects of retinal physiology, they provide excellent experimental models for understanding light and circadian-regulatory processes.

See also: Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Ganglion Cells: Anatomy; Retinal Pharmacology: Inner Retinal Layers; Retinitis Pigmentosa.

Further Reading

Ali MA (ed.) (1975) Retinomotor responses. In: *Vision in Fishes*, pp. 313–355. New York: Plenum.

Arey LB (1915) The occurrence and significance of photomechanical changes in the vertebrate retina – A historical survey. *The Journal of Comparative Neurology* 25: 535–554.

Back I, Donner KO, and Reuter T (1965) The screening effect of the pigment epithelium on the retinal rods in the frog. *Vision Research* 5: 101–111.

Blaxter JHS (1975) The eyes of larval fish. In: Ali MA (ed.) *Vision in Fishes*, pp. 427–443. New York: Plenum.

Burnside B (1988) Photoreceptor contraction and extension: Calcium and cAMP regulation of microtubule- and actin-dependent changes in cell shape. In: Lasek RJ (ed.) *Intrinsic Determinant of Neuronal Cell Form*, pp. 323–359. New York: Alan R. Liss.

Burnside B (2001) Light and circadian regulation of retinomotor movement. In: Kolb H, Ripps H, and Wu S (eds.) *Progress in Brain Research*, pp. 478–485. New York: Elsevier Science BV.

Burnside B and Dearry A (1986) Cell motility in the retina. In: Adler R and Farber DB (eds.) *The Retina*, part 1, pp. 151–206. Orlando, FL: Academic Press.

Burnside B and Nagle B (1983) Retinomotor movements of photoreceptors and retinal pigment epithelium: Mechanisms And regulation. In: Osborne N and Chader GJ (eds.) *Progress in Retinal Research*, vol. 2, pp. 67–109. Oxford: Pergamon.

Dearry A and Burnside B (1989) Regulation of cell motility in teleost retinal photoreceptors and pigment epithelium by dopaminergic D2 receptors. In: Redburn D and Pasantes Morales H (eds.) *Extracellular and Intracellular Messengers in the Vertebrate Retina*, pp. 229–256. New York: Alan R. Liss.

Douglas RH (1982) The function of photomechanical movements in the retina of the rainbow trout (*Salmo gairdneri*). *Journal of Experimental Biology* 96: 389–403.

Khune W (1887) Fortgesetzte Untersuchungen uber die Retina und die Pigmente des Auges. *Untersuchungen Physiologie Institut Universitat Heidelberg* 2: 89–109.

McNeil EL, Tacelosky D, Basciano P, et al. (2004) Actin-dependent motility of melanosomes from fish retinal pigment epithelial (RPE) cells investigated using *in vitro* motility assays. *Cell Motility and the Cytoskeleton* 58: 71–82.

O'Connor and Burnside B (1981) Actin-dependent cell elongation in teleost retinal rods: Requirement for actin filament assembly. *Journal of Cell Biology* 89: 517–524.

Parker GH (1932) The movements of the retinal pigment. *Ergebnisseder Biologie* 9: 239–292.

Wagner HJ, Kirsch M, and Douglas RH (1992) Light dependent and endogenous circadian control of adaptation in teleost retinae. In: Ali MA (ed.) *Rhythms in Fishes Plenum*. New York: Plenum.

Retrograde Neurotrophic Signaling

C Wu and W C Mobley, Neuroscience Institute, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The survival and differentiation of neurotrophin (NT)-responsive neurons are critically regulated by the targets that these neurons innervate. However, the exact mechanism or mechanisms by which the target-derived NT signals produced at the axonal terminals are delivered to the cell body over a significant distance (>1 m in some neurons) remain unclear. These underlying mechanisms have been one of the most intriguing enigmas in modern neurobiology.

The discovery of nerve growth factor (NGF) by Levi-Montalcini and Hamberger provided an important clue to the mechanism: A soluble NGF produced in the target dramatically influenced specific populations of neurons. It is now well established that NGF produced in and released from target tissues activates specific receptors at the distal axons of innervating neurons. The signals thus elicited are subsequently retrogradely transported to the cell bodies to regulate nuclear and cytosolic events that are known to be important for the survival and maintenance of healthy neurons. However, the generation, intracellular sorting, and maintenance of the signal remain the subjects of intensive investigation.

Compartmentalized Nerve Cell Chambers Are Used to Study Retrograde Axonal Traffic

Campenot Chamber

In vitro experimental studies to support the need for retrograde NT signaling have benefited markedly from the use of compartmentalized neuronal cultures pioneered by Campenot and colleagues. In this experimental paradigm, the neuronal cell bodies and proximal axons are separated from distal axons by a Teflon divider sealed with silicone grease onto a collagen-coated surface. Radiolabeled NGF (^{125}I-NGF) can be added to the distal chamber and its retrograde transport from the axonal terminal to the proximal axon or the cell body can be quantitatively measured. Retrogradely transported signaling molecules can also be analyzed by comparing, by means of immunoblotting and immunostaining, the changes in distal axons to those in the proximal axons or cell bodies. The use of this device and its variants has allowed study of the mechanisms for retrograde NGF and NT signaling.

Microfluidic Chambers

A microfluidic device has been developed to serve as a platform for long-term cultures of primary neurons from the central nervous system *in vitro* to study axonal injury and regeneration. Though only recently introduced, it has already proven extremely useful for examining axonal transport. These miniature devices, fabricated with poly(dimethylsiloxane) (PDMS), afford the ability to precisely control, monitor, and manipulate cellular microenvironments. The microfluidic nerve cell chamber has been modified for study of the transport kinetics of retrograde axonal transport of NGF in rat embryonic dorsal root ganglion (DRG) neuronal cultures. The device is depicted in Figure 1. Dissociated DRG neurons are loaded into the left well, and both wells are supplied with standard growth medium (Figures 1(a) and 1(b)). Axons grow across the first microgroove and into the middle chamber in 2–3 days; many then traverse the second microgroove and emerge into the distal axon chamber in an additional 4–5 days.

Among the many advantages of this device is that axons can be confined to the two straight microgrooves whose cross-sectional area (3 μm × 3 μm) and length (in this case, 100 μm) are well defined. An example of rat embryonic DRG culture (7 days *in vitro*, i.e., DIV7) is shown in Figure 1(c). Furthermore, neurons can be loaded at a density at which only a single axon crosses the microgroove. This makes it possible, using microscopy, to more readily track labeled NGF and its signaling molecules. As demonstrated in Figure 2, distal axons from postnatal mouse DRG neurons that were cultured in the device for 7 days (Figure 1) were incubated with biotinylated NGF conjugated to streptavidin-Quantum dot 605, and the retrograde movement of labeled NGF was imaged in the proximal axons (Figure 2). Furthermore, the presence of a middle chamber makes it possible to modify the content of the environment (e.g., to administer various kinase inhibitors) through which the axon transmits neurotrophic signals.

Studying Retrograde NT Signaling *In Vivo*

In vivo studies have most often used ligation to examine transport. In the ligature technique, after anesthesia has been administered, a nerve (in most cases the sciatic nerve) is surgically exposed, and one or

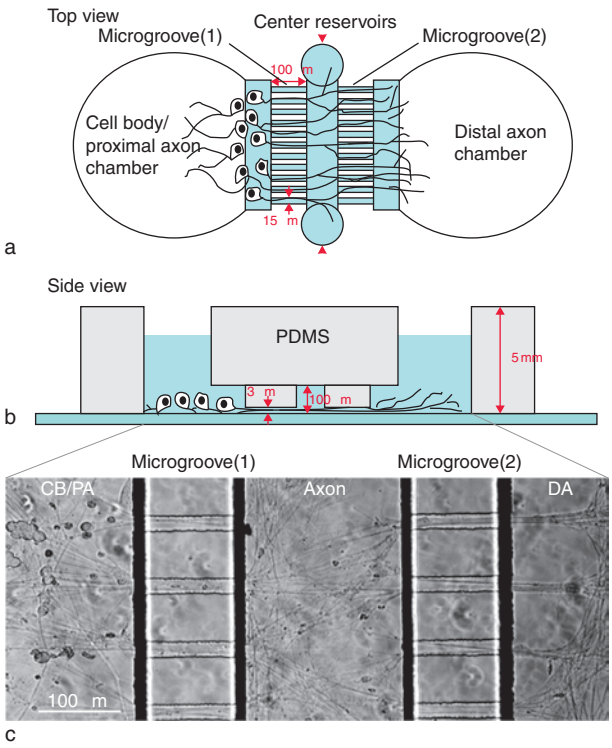

Figure 1 Schematic representation of a modified version of the microfluidic nerve cell chamber showing the top view (a) and side view (b). In (c), Rat E16 DRG neurons were cultured in the device for 7 days, and DRG axons crossed robustly the two microgrooves into the distal chamber. CB, cell body; PA, proximal axon; DA; distal axon; PDMS, poly(dimethylsiloxane).

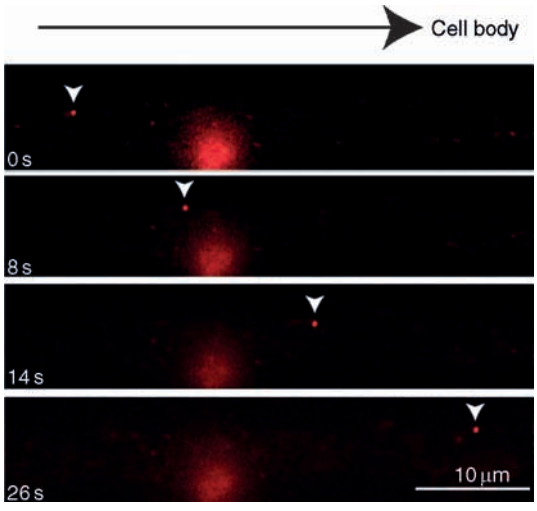

Figure 2 Postnatal (P0) mouse DRG neurons were cultured in the microfluidic nerve cell chamber for 7 days. Biotinylated NGF that was conjugated to streptavidin-Quantum-dot 605 was added to the distal chamber. Movement of labeled NGF within the proximal axons was imaged in real time by an inverted microscope (100x oil). A series of time-sequenced frames shows the retrograde NGF transport.

more ligatures, separated by a desired distance, are placed. The animals are allowed to recover. Then, after a desired interval, they are sacrificed, and the nerve is taken for immunohistocytochemical, electron microscopic, and biochemical studies. The undamaged contralateral sciatic nerve may serve as control in some experiments. To define the extent of transport of a molecule or organelle, one typically measures the extent to which it accumulates on the most distal ligature (for retrograde transport) or the most proximal ligature (for anterograde transport). The use of a double ligature is useful in controlling for the effect of ligation-mediated injury on the possible local production of the marker of interest. In an alternative procedure, a segment of nerve can be surgically removed and placed on a slab of gelled agarose with both ends placed in a collecting chamber containing a physiological medium. The eluted axoplasm can be analyzed by one or more of the methods specified above.

Another established technique that has been used to study retrograde NT signaling places the molecule of interest in the target area of afferent axons. For example, the hipposeptal cholinergic pathway has been used to examine NGF retrograde transport from the hippocampus to basal forebrain cholinergic neurons. In this experimental paradigm, either radiolabeled or fluorescently labeled NGF is stereotaxically injected into the frontal cortex or hippocampus of anesthetized animals. The animals are allowed to recover for a desired period before sacrifice and analysis. In these experiments, one can determine the extent of transport by measuring the accumulation of the labeled species in the septum using light microscopy for fluorescently tagged molecules or, in the case of radiolabeled molecules, *in situ* autoradiography or gamma counting in the septum.

Retrograde Neurotrophic Signaling: Proposed Mechanisms

Signaling Endosome Hypothesis

This hypothesis posits that the NGF signal is transmitted via endocytosis of complexes containing NGF bound to its activated tyrosine receptor kinase A (TrkA) receptor (i.e., pTrkA), followed by retrograde transport of the 'signaling endosome' thus formed (Figure 3). The hypothesis points to signaling endosomes as an important source of retrogradely transmitted signals. It is consistent with the emerging view that cellular signaling pathways are highly organized and compartmentalized, features that are used to confer specificity and sustainability of signal transduction.

The signaling endosome hypothesis is supported by an increasing body of experimental evidence from a number of laboratories. Studies *in vitro* have shown that (1) NGF at axon terminals must be internalized

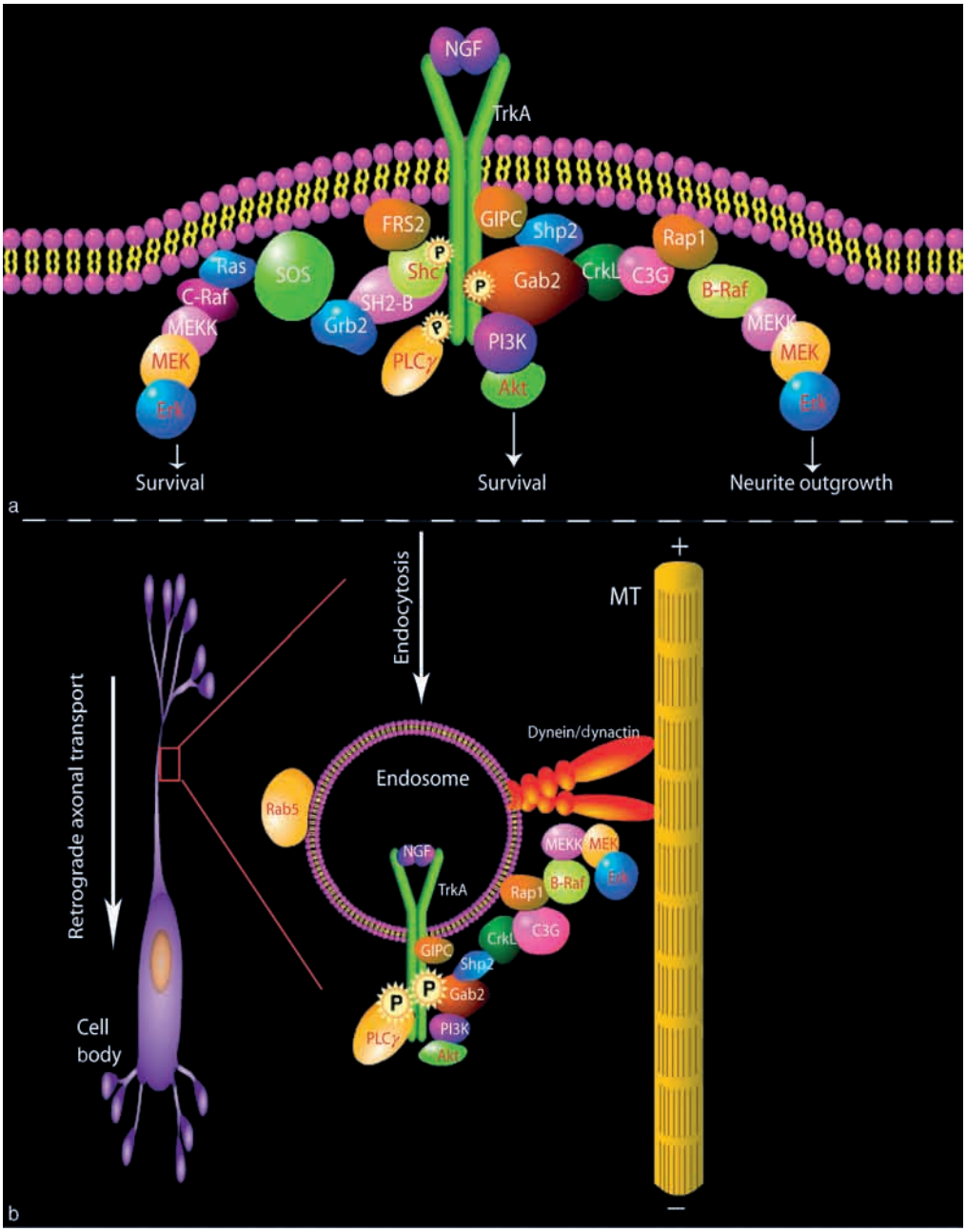

Figure 3 NGF–TrkA signaling pathways and retrograde axonal transport of signaling endosomes. (a) NGF activates its high-affinity tyrosine receptor kinase, TrkA, and elicits divergent signaling pathways. The four major pathways are (1) the transient SOS/Ras/C-Raf/MEKK/MEK/Erk pathway; (2) the persistent C3G/Rap1/B-Raf/MEKK/MEK/Erk pathway; (3) the PI3K/Akt pathway; and (4) the PLCγ pathway. Activation of these pathways effects neuronal functions of survival and differentiation. (b) Following endocytosis of NGF and activated TrkA signaling complex, the signaling endosome engages the dynein–dynactin complex to undergo retrograde axonal transport along the microtubule network (MT). NGF, nerve growth factor; TrkA, tyrosine receptor kinase A; FRS2, fibroblast growth factor receptor substrate-2; Shc, SH2-containing protein; SH2-B, SH2-containing protein B; Grb2, growth factor receptor bound protein 2; SOS, son of sevenless; Ras, a small GTPase; C-Raf, Raf kinase C; MEKK, MAPK/Erk kinase kinase; kinase; MEK, MAPK/Erk kinase kinase; Erk (or MAPK), extracellular signal-regulated kinase (or mitogen activated protein kinase); Gab2, Grb2-associated binding protein-2; GIPC, GAIP interacting protein; Shp2, SH2-containing protein phosphatase-2; CrkL, v-crk avian sarcoma virus CT10 oncogene homolog-like; C3G, Crk SH3 binding guanine-nucleotide releasing factor; Rap1, Ras-related protein 1; B-Raf, Raf kinase B; PI3K, phosphatidylinoditiol-3-kinase; Akt, protein kinase B; PLCγ, phospholipase-C-gamma; GAIP, Gα-interacting protein; Rab5, a member of the RAS oncogene family; P, phosphate.

and retrogradely transported to neuron cell bodies for the NGF signal to induce phosphorylation of cyclic adenosine monophosphate response element-binding protein (CREB) and to increase survival of immature neurons; (2) NGF-induced activation of TrkA in axon terminals is also required for survival; (3) NGF induces endocytosis of TrkA and is bound to pTrkA in endosomes; (4) retrograde NGF signals

are decreased by inhibiting endocytosis; (5) NGF–pTrkA complexes are found in clathrin-coated vesicles together with activated components of the Ras/extracellular signal-regulated kinase (Erk)1/2 pathway as well as phosphatidylinoditiol-3-kinase (PI3K) and phospholipase-C-γ; (6) NGF signaling that is present in clathrin-coated vesicles is able in *in vitro* kinase assays to phosphorylate Elk-1, a downstream target of Erk1/2; (7) in what appears to be the next step in the maturation of the signaling endosome within the endosomal pathway, and a step that requires both dynein and intact microtubules, NGF also induces the formation of complexes containing pTrkA and activated components of the Ras-related protein 1 (Rap1)/Erk1/2 pathway in early endosomes; (8) NGF–pTrkA complexes are moved retrogradely in axons to cell bodies, as are the complexes of brain-derived neurotrophic factor (BDNF) and its phosphorylated tyrosine receptor kinase B (pTrkB); (9) retrograde transport of these complexes is blocked by disrupting microtubules; and (10) retrogradely transported pTrk is required for CREB phosphorylation, c-fos induction, protein kinase B (Akt) phosphorylation, and survival of immature neurons. *In vivo* and *ex vivo* observations are consistent in showing that (1) early endosomes containing NGF–pTrkA, together with activated components of the Rap1/Erk1/2 and PI3K pathways, are retrogradely transported in sciatic nerve; (2) increasing NGF levels in the target of innervation increases retrograde transport of these membranes and increases pErk1/2 in DRG neuron cell bodies; conversely; (3) decreasing NGF in the target of innervation by sequestering it with specific antibodies decreases the retrograde transport of endosomes containing pTrk and pErk1/2; and (4) it has been demonstrated recently *in vivo* that NGF was localized within early endosomes in cholinergic terminals, and after hippocampal injection, ^{125}I-NGF was coimmunoprecipitated with antibodies against TrkA.

NGF-Independent Retrograde Signaling

There is evidence that NGF retrograde signaling may not always be coupled to retrograde transport of NGF. Senger and Campenot found that while it took at least 30 min for ^{125}I-NGF to arrive at the cell body/proximal chamber at a detectable level following addition of ^{125}I-NGF to the distal chamber in the compartmentalized culture of sympathetic neurons, TrkA and other signaling molecules were more rapidly activated in the cell body/proximal axons. Utilizing NGF that was covalently cross-linked to beads, thus rendering it incapable of internalization, Campenot's group further demonstrated that this source of NGF was able to trigger activation and axonal transport of TrkA and PI3K/Akt signaling proteins at levels adequate to support survival of sympathetic neurons for up to 30 h. Taken together, the findings suggest that signaling events initiated by NGF at the axonal terminals can be transported retrogradely without NGF. More recently, Mok and Campenot used Go6976 a very potent drug in inhibiting TrkA phosphorylation, to block TrkA phosphorylation in the cell body/proximal axons in compartmentalized sympathetic neurons. Go6976 itself exerts no survival effects on sympathetic neurons deprived of NGF. After applying Go6976 to locally block TrkA phosphorylation in the cell body/proximal axons, NGF added to distal axons still resulted in activation of survival signaling pathways (e.g., pAkt, pCREB) in the cell body/proximal axons. These investigators concluded that retrograde transport of signaling molecules downstream of phosphorylated TrkA, but not phosphorylated TrkA itself, is necessary to support neuronal survival.

Other Hypotheses

Other possible mechanisms for retrograde signaling can be envisioned. In a general sense, they conform either to the 'wave model' or to the 'signaling effector model.' The former suggests that NGF binds to and activates TrkA at the distal axonal terminals. Activated TrkA and/or signaling molecules downstream of activated TrkA are subsequently propagated and transported back to the cell body. The latter model postulates that second messengers such as ionic fluxes (e.g., cytosolic Ca^{2+}), instead of NGF–pTrkA, are retrogradely transported. There is little experimental data to support these alternatives, but given the paucity of knowledge for retrograde signaling mechanisms, continued attempts should be made to discover and characterize alternative models.

The Nature of Intracellular Vesicles That Carry Retrograde NGF Signaling

There is no consensus about the type of endocytic vesicle or vesicles that transport retrograde NT signaling in axons. An early study suggested that lysomes/multivesicular bodies (MVBs) contained the majority of ^{125}I-NGF that was retrogradely transported in cultured sympathetic neurons. Studies from Hendry laboratory showed that ^{125}I-NGF transported from the eye to the sympathetic and sensory ganglia was concentrated in MVBs. These findings led to the proposal that MVBs are the major vesiclar carriers within axons for retrograde NT signaling. Depending on the experimental conditions, investigators from a numbers of laboratories discovered that

NGF or its signaling components such as pTrk resided in pincher-positive complex structures, in 'uncoated' vesicles, in the *trans*-Golgi network, in late endosomes that are marked by Rab7, or in Rab5-positive early endosomes. It is possible that many different endocytic vesicle types are engaged in transporting NGF signals and that different neurons use a different complement of such organelles. Resolving this issue is an important next step in clarifying retrograde signaling pathways.

Role of MT/Dynein in Mediating Retrograde Transport of NT Signaling

It is not surprising that retrograde transport of NT signaling endosomes in axons, like many other subcellular organelles (mitochondria, RNA particles, etc.), requires the microtubule network (MT) for movement. In support of this, infusion of nocodazole, a drug that causes the collapse of MT, into isolated sciatic nerves blocked retrograde NGF transport. Similarly, another MT collapse-causing drug, colchicine, suppressed Trk phosphorylation of cell bodies in response to neurite stimulation in DRG neurons. Colchicine pretreatment also blocked CREB phosphorylation in response to stimulation of the cell bodies, as well as the neurites.

The dynein motor protein complex is primarily responsible for powering retrograde movement within a cell. It comprises many different subunits (the heavy chains, the intermediate chains, and the light chains). Evidence from Chao laboratory suggests that Trk associates with both the 14 kDa light and the 74 kDa intermediate chain of dynein. Furthermore, Segal and colleagues found that pTrk was found to co-localize with dynein, but not with the anterograde motor kinesin, in rat sciatic nerve axons. In addition, the 74 kDa subunit of dynein was found in retrograde transport vesicles that contain NGF–pTrkA isolated from rat sciatic nerve axons. More recently, disruption of the dynein function by introduction of the 50 kDa subunit (dynamitin) of dynactin, an established method to dissociate the dynein–dynactin complex, blocked retrograde transport of NGF and BDNF and led to atrophy and cell death in neurons. These results demonstrate that retrograde transport of NT signaling is driven by the dynein–dynactin complex.

Disrupted Retrograde Transport NT Signaling in Neurodegenerative Diseases

Given the role that active retrograde NT signaling plays in the development of the nervous system, and evidence that continued production in targets of innervation are needed to maintain mature neurons, it is noteworthy that an increasing body of data points to a link between defects in retrograde NT signaling and dysfunction and degeneration of neurons in neurodegenerative diseases. Salehi et al. recently demonstrated that in a mouse model of Down syndrome (Ts65Dn), the presence of one extra copy of the amyloid precursor protein selectively disrupted retrograde NGF transport and resulted in degeneration of basal forebrain cholinergic neurons (BFCNs), a pathological hallmark in Alzheimer's disease. In an earlier study, Cooper et al. showed that NGF was retrogradely transported via the hipposeptal pathway. The age-related degeneration of BFCNs (decrease in cell numbers and profile area of cell bodies) in Ts65Dn mice could be reversed by direct intracerebroventricular NGF infusion. Similarly, a severe defect in retrograde NGF transport has also been reported by Yang and colleagues in cultured postnatal DRG neurons of a mouse model of giant axon neuropathy, a disorder that affects a number of neuronal populations in the peripheral and central nervous system.

Defective retrograde NT transport has also been implicated in Huntington's disease (HD), a condition that results from an expansion of polyglutamine tract (>37 glutamines) in huntingtin (htt). In a recent study, Rong et al. observed that htt-associated protein-1, a contributing factor to HD pathology, interacts with the dynein motor in sympathetic neurons. Mutant htt appears to influence this interaction and in so doing reduces the stability of intracellular TrkA. As a result, retrograde NGF signaling may well be compromised in the diseased neurons.

Given that retrograde neurotrophic signaling may share mechanisms and protein machinery with other retrograde processes, it is conceivable that disruption or alteration in any components of the machinery will also inevitably affect retrograde NT signaling. Disruption of retrograde axonal transport has been suggested to contribute to the pathogenesis of a number of progressive disorders of motor neurons. In one study, transgenic overexpression of p50 was shown to result in inhibition of axonal transport and cause late-onset progressive degeneration in motor neurons. In addition, a family with chronic motor neuron disease has been shown to harbor a mutation in p150 glued, a protein that plays a critical role in the function of the dynein complex.

Genetic mutations that affect the stability of the microtubules have been suggested to cause progressive motor neuronopathy in mice. A Trp524Gly substitution at the last residue of the tubulin-specific chaperone e protein (gene locus: chromosome 13) leads to a reduced number of microtubules in the sciatic and phrenic nerves. The affected mice develop progressive caudocranial degeneration of their motor

axons from the age of 2 weeks and die 4–6 weeks after birth. Axonal degeneration apparently starts at the endplates and is prominent in the sciatic nerve and its branches and the phrenic nerve. Since intact microtubules are required for retrograde NGF transport, it is conceivable that retrograde NGF–pTrkA signaling is compromised in these mice as well, which may well in turn contribute to the disease.

In another example, alsin, a putative guanine nucleotide exchange factor for Rab5, plays a critical role in regulating early endosomal fusion. Loss of alsin-2 function is believed to be the cause for a subset of amyotrophic lateral sclerosis (ALS) in humans, juvenile-onset primary lateral sclerosis. In Als2(−/−) mice, endosomal transport of insulin-like growth factor 1 and BDNF receptors was selectively disturbed. In another mouse model of ALS, a missense mutation L967Q in Vps54 has been identified in the spontaneous autosomal recessive wobbler mutation. Axonal transport is impaired in wobbler mice, and Vps54 might be critical for retrograde vesicular transport and in particular for axonal transport in motor neurons. Since internalized NGF was found to localize in early endosomes in cholinergic neurons, retrograde NGF signaling is likely affected in these neurons as well.

Recent studies have revealed that late endosomes that are marked by Rab7 may also play a role in modulating retrograde NGF–pTrkA signaling. In genetic studies, the locus for Rab7 has been implicated as the disease gene for Charcot-Marie-Tooth disease type 2B (CMT2B). CMT2B is a rare autosomal dominant genetic disorder that leads to peripheral sensory neuropathy with prominent axonal degeneration. Missense mutations in the highly conserved C-terminus of Rab7 (Leu129Phe, Val162Met, Lys157Asn) have been identified in CMT2B patients. It will be of great interest to examine whether retrograde NT trafficking or signaling is affected in neurons that harbor Rab7 mutations.

Summary

Retrograde NT signaling plays a critical role in regulating survival and differentiation of specific populations of neurons in both the central nervous system and the peripheral nervous system. There is increasing evidence that NGF and other NTs bind to and activate their receptors at the axonal terminus to elicit the recruitment of an array of signaling molecules. Endocytosis of the NT–pTrk signaling complex is then sorted into a subpopulation of early endosomes or other cellular organelles to give rise to signaling endosomes. The signaling endosome driven by the dynein motor complex is retrogradely transported via the MT to the cell body to effect gene expression and other somal events. Retrograde NT signaling thus appears to be a very complex process that involves numerous cellular components and events. Disruption of retrograde NGF and NT signaling may contribute or lead to many types of neurodegenerative diseases.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Transport and ALS; Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Axonal Transport and Neurodegenerative Diseases; Enteric Nervous System: Neurotrophic Factors; Huntington's Disease: Neurodegeneration; Neurotrophic Factor Therapy: GDNF and CNTF; Neurotrophic Factor Therapy: NGF, BDNF and NT3; Neurotrophins: Physiology and Pharmacology; Vascular Issues in Neurodegeneration and Injury.

Further Reading

Campenot RB and MacInnis BL (2004) Retrograde transport of neurotrophins: Fact and function. *Journal of Neurobiology* 58: 217–229.

Chao MV (2003) Neurotrophins and their receptors: A convergence point for many signalling pathways. *Nature Reviews Neuroscience* 4: 299–309.

Cooper JD, Salehi A, Delcroix JD, et al. (2001) Failed retrograde transport of NGF in a mouse model of Down's syndrome: Reversal of cholinergic neurodegenerative phenotypes following NGF infusion. *Proceedings of the National Academy of Sciences of the United States of America* 98: 10439–10444.

Cui B, Wu C, Chen L, et al. (2007) One at a time, live tracking of NGF axonal transport using quantum dots. *Proceedings of the National Academy of Sciences of the United States of America* 104(34): 13666–13671.

Deinhardt K, Salinas S, Verastegui C, et al. (2006) Rab5 and Rab7 control endocytic sorting along the axonal retrograde transport pathway. *Neuron* 52: 293–305.

Delcroix JD, Valletta JS, Wu C, Hunt SJ, Kowal AS, and Mobley WC (2003) NGF signaling in sensory neurons: Evidence that early endosomes carry NGF retrograde signals. *Neuron* 39: 69–84.

Heerssen HM, Pazyra MF, and Segal RA (2004) Dynein motors transport activated Trks to promote survival of target-dependent neurons. *Nature Neuroscience* 7: 596–604.

Howe CL and Mobley WC (2005) Long-distance retrograde neurotrophic signaling. *Current Opinion in Neurobiology* 15: 40.

Huang EJ and Reichardt LF (2001) Neurotrophins: Roles in neuronal development and function. *Annual Review of Neuroscience* 24: 677–736.

LaMonte BH, Wallace KE, Holloway BA, et al. (2002) Disruption of dynein/dynactin inhibits axonal transport in motor neurons causing late-onset progressive degeneration. *Neuron* 34: 715–727.

Levi-Montalcini R (1987) The nerve growth factor 35 years later. *Science* 237: 1154–1162.

Salehi A, Delcroix JD, Belichenko PV, et al. (2006) Increased App expression in a mouse model of Down's syndrome disrupts NGF

transport and causes cholinergic neuron degeneration. *Neuron* 51: 29–42.

Taylor AM, Blurton-Jones M, Rhee SW, Cribbs DH, Cotman CW, and Jeon NL (2005) A microfluidic culture platform for CNS axonal injury, regeneration and transport. *Nature Methods* 2: 599–605.

Wu C, Lai CF, and Mobley WC (2001) Nerve growth factor activates persistent Rap1 signaling in endosomes. *Journal of Neuroscience* 21: 5406–5416.

Zweifel LS, Kuruvilla R, and Ginty DD (2005) Functions and mechanisms of retrograde neurotrophin signalling. *Nature Reviews Neuroscience* 6: 615–625.

Retrograde Transsynaptic Influences

G W Davis, University of California at San Francisco, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

More than 50 years ago, work by Victor Hamburger and his contemporaries identified target-derived factors that influence neuronal growth. Since that time it has been demonstrated that target-derived molecules are essential for many aspects of neural development, including dendrite growth, synaptic growth, synaptic stabilization, and the choice of neurotransmitter. In general, these studies have highlighted the action of growth factors, neurotrophins, and morphogen signaling proteins. These signaling proteins and their receptors are potent intercellular signaling molecules utilized repeatedly throughout development. Although studied extensively, new functions for these signaling systems continue to be identified in the context of axon guidance, stem cell maturation in the brain, and neurodegenerative disease.

Other retrograde, transsynaptic signaling systems also exist that are essential for the proper development and function of neural circuitry. In many cases, the molecular bases of these retrograde influences are not known. This article covers three fundamental phenomena that seem to require retrograde, transsynaptic signaling. First, retrograde signaling during synapse formation is considered. Retrograde signaling is well established in this context and several molecular players are known, yet the full repertoire of retrograde signaling molecules remains to be defined. This is followed by a discussion of two phenomena that require retrograde signaling of unknown identity. These include the specification of presynaptic release probability during synapse development, and the homeostatic modulation of presynaptic release probability.

Retrograde Signaling during Synapse Formation

Synapse formation involves the transformation of a motile growth cone into a stable synapse capable of mediating high-fidelity neurotransmission. This process requires the assembly and precise apposition of pre- and postsynaptic protein assemblies. These protein assemblies are known as the pre- and postsynaptic densities due to their electron-dense appearance when examined using transmission electron microscopy. The presynaptic density includes proteins that are necessary for the calcium-dependent release and recycling of synaptic vesicles. The postsynaptic density includes neurotransmitter receptors, signaling molecules, and scaffolding proteins that control the detection of neurotransmitter. Bidirectional communication between the pre- and postsynaptic cells occurs during synapse formation in order to precisely assemble the pre- and postsynaptic element. Insight into the nature of this bidirectional signaling has been achieved at both the neuromuscular junction (NMJ) and at synapses in the central nervous system.

Synaptogenesis at the Neuromuscular Junction

At the vertebrate neuromuscular junction, prior to the arrival of the growth cone, cellular specializations exist both pre- and postsynaptically; these set the stage for ensuing synapse formation. Presynaptically, the growth cone is able to release neurotransmitter and other signaling molecules that are required for synapse formation. Postsynaptically, neurotransmitter receptors (acetylcholine receptors; AChRs) are prepatterned into AChR clusters in a loose array that prefigures the site of eventual synapse formation. It remains unknown how many signaling molecules in addition to AChRs are also prepatterned to the site of eventual synapse formation.

Upon arrival at a muscle target, the presynaptic growth cone releases agrin, a heparan sulfate proteoglycan, which is deposited into the synaptic basal lamina. Agrin represents an anterograde signal that is necessary for the induction of the postsynaptic density (Figure 1). Agrin signals to the muscle via two highly conserved muscle proteins, muscle-specific kinase (MuSK) and an associated adaptor protein named rapsyn. Genetic knockout experiments demonstrate that each of these proteins is necessary for synapse formation to proceed, identifying agrin–MuSK–rapsyn signaling as an essential anterograde signaling cascade that is required for synaptogenesis. The release of the neurotransmitter acetylcholine (ACh) also participates in the early process of synapse formation. Genetic experiments in mice combining mutations in the ACh synthetic enzyme choline acetyltransferase (ChAT) with mutations in MuSK and agrin support a model in which ACh released onto the muscle actively disperses prepatterned AChRs (mechanism unknown), while agrin stabilizes those receptors that reside directly at the site of synapse formation.

The next stage of synapse formation requires a retrograde signal from muscle to nerve that directs the maturation of the presynaptic nerve terminal

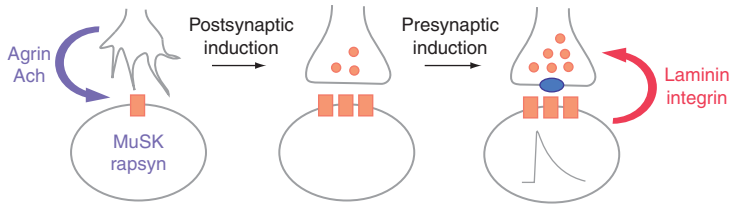

Figure 1 Reciprocal signaling during synapse formation at the neuromuscular junction. At left, the motile growth cone releases both agrin and acetylcholine (ACh). ACh disperses prepatterned ACh receptor (AChR) clusters (orange box) that reside at the site of innervation. Agrin, deposited in the basal lamina, counteracts the effects of ACh and stabilizes AChRs by signaling through muscle-specific kinase (MuSK) and rapsyn in muscle. This results in AChR clustering and stabilization at the neuromuscular junction (middle diagram). Following the formation of the postsynaptic density, a retrograde signal specifies presynaptic maturation, including the formation of presynaptic dense bodies and clustered presynaptic calcium channels (blue oval) directly opposite the postsynaptic receptors. Molecules implicated in this retrograde signal include laminins and integrins.

(Figure 1). The release of this retrograde signal follows the induction of the postsynaptic specialization. In support of this conclusion, the transformation of a motile growth cone into a fully differentiated presynaptic terminal will not proceed in animals that lack MuSK or rapsyn. Indeed, muscle transplant experiments provide evidence that presynaptic maturation may never occur in the absence of postsynaptic differentiation. Thus, it appears that the agrin-dependent induction of the postsynaptic specialization leads to the production of a retrograde signal that is subsequently required to transform a motile motoneuron growth cone into a stable presynaptic terminal (Figure 1).

The retrograde signal that induces the maturation of the presynaptic terminal remains unknown, though several candidates are strongly implicated. Laminins present in the synaptic basal lamina have a well-established function in presynaptic maturation. Genetic deletion of laminin-β2 leads to a poorly differentiated presynaptic nerve terminal and impairs the association of the Schwann cell with the presynaptic terminal. Knockout of laminin-α4 leads to misalignment of the presynaptic active zone with postsynaptic specializations, though presynaptic maturation proceeds. Although a role for laminin in presynaptic induction seems clear, the means by which the laminins induce the transformation of a growth cone into a presynaptic nerve terminal are less obvious. One interesting possibility is that laminins interact directly with key presynaptic proteins. Laminin-α4 has been suggested to be a 'ligand' for presynaptic calcium channel 'receptors.' Presynaptic calcium channels may well constitute an organizing center for the construction of a presynaptic nerve terminal. Other retrograde influences also exist. Recently, integrin-mediated signaling has also been implicated in presynaptic induction. Deletion of integrin-β1 in muscle results in failure of the presynaptic nerve terminal to fully differentiate, a phenotype resembling the MuSK knockout.

Systems amenable to forward genetic analysis of synapse formation hold promise for identification of new genes that participate in the bidirectional signaling necessary for synapse formation. Genes that influence the specificity and speed of synapse formation have been found by forward genetic approaches, several of which are highlighted here. At the *Drosophila* NMJ, a four-transmembrane-spanning protein (tetraspanin) encoded by a gene termed *late bloomer* (*lbl*) is required presynaptically, for the efficient transformation of a motile growth cone into a stable presynaptic nerve terminal. However, as is the case for many synaptogenic proteins (see later), synapse formation ultimately proceeds in the absence of the *lbl*-encoded protein.

In the nematode *Caenorhabditis elegans*, forward genetic approaches have identified an intercellular signaling system that is necessary for proper synapse formation. These studies demonstrate that two presynaptic scaffolding molecules, SYD-1 and SYD-2, are necessary for the assembly of the presynaptic active zone. In mutants that lack the genes for these proteins, synaptic vesicles and other presynaptic proteins fail to accumulate at the developing active zone. These scaffolding proteins function, genetically, downstream of the intercellular signaling proteins SYG-1 and SYG-2, which are required for the proper placement of motoneuron synapses (synapse specificity). The SYG-1 protein is an immunoglobulin (Ig) domain-containing protein with similarity to vertebrate NEPH1. SYG-2 is present on an epithelial guidepost cell that prefigures synapse placement. SYG-2 binds to SYG-1, which is expressed by the motoneuron. Together, these studies suggest a model in which intercellular signaling mediated by SYG-2 and SYG-1 organizes presynaptic SYD-1 and SYD-2, which, in turn, direct the assembly of the presynaptic active

zone. Although synapse formation proceeds at inappropriate sites in a *syg-1* mutant, this transsynaptic signaling system represents a potent intercellular signaling system capable of directing appropriate synapse formation.

Synaptogenesis in the Central Nervous System

In the vertebrate central nervous system, the Wnt family of morphogen signaling molecules are involved in both axon guidance and synapse formation. Data support a model in which Wnt, released from the target cells, causes growth cone stalling and characteristic cytoskeletal rearrangements that are observed during normal synaptogenesis Although synapse formation proceeds *in vivo* in a Wnt receptor knockout, the process is delayed, suggesting that Wnts can function to promote synapse formation. A similar function has also been attributed to target-derived fibroblast growth factor 22 (FGF-22) during synapse formation in the vertebrate central nervous system.

Other proteins have been identified that are sufficient to induce the transformation of a motile growth cone into a presynaptic nerve terminal *in vitro*. These signaling proteins include neuroligin and the Ig domain-containing protein SynCam. In both cases, expression of these proteins on the surface of a heterologous cell is sufficient to transform a growth cone into a structure that resembles a presynaptic terminal. The presynaptic terminals that form on these heterologous cells share ultrastructural and molecular hallmarks of presynaptic nerve terminals and are capable of vesicular neurotransmitter release. Neuroligin can signal to the presynaptic nerve terminal by binding presynaptic neurexin, which, in turn, can be linked biochemically to an extensive array of presynaptic proteins including presynaptic calcium channels. Thus, a model has emerged that is similar to that proposed in the nematode. Transsynaptic signaling via a receptor–ligand pair induces the formation of pre- and postsynaptic protein scaffolds that serve as a template for the eventual assembly of pre- and postsynaptic densities.

Target-Dependent Control of Presynaptic Release Probability

There is now widespread agreement that target-derived retrograde signals participate in the activity-dependent modulation of presynaptic function. Retrograde signals include the endocannabinoids, neurotrophins, and growth factor signaling molecules. Throughout the nervous systems of both vertebrates and invertebrates, evidence has also accumulated demonstrating that target cells specify the steady-state release properties of impinging presynaptic nerve terminals. The basis for this phenomenon is the demonstration that a single presynaptic neuron can establish synaptic connections with very different release properties, simultaneously, on different target cells (Figure 2(a)). A common finding is that a presynaptic neuron can simultaneously form low-release probability, facilitating synapses with one target, and high-release probability, depression-prone synapses with a second target (Figure 2). In each example, it is clearly established that differences in release probabilities and short-term dynamics are a consequence of presynaptic differences, rather than reflecting differences in the detection of neurotransmitter by the postsynaptic targets. As a result, it seems likely that the target cell, or a target-associated glial cell, determines the steady-state release properties of impinging presynaptic nerve terminals. If so, this type of retrograde influence may be a general feature of synapse development, even in circumstances wherein presynaptic terminals are formed at only a single target type.

The ability of a single presynaptic neuron to form synaptic connections expressing a different release probability at separate targets has profound importance

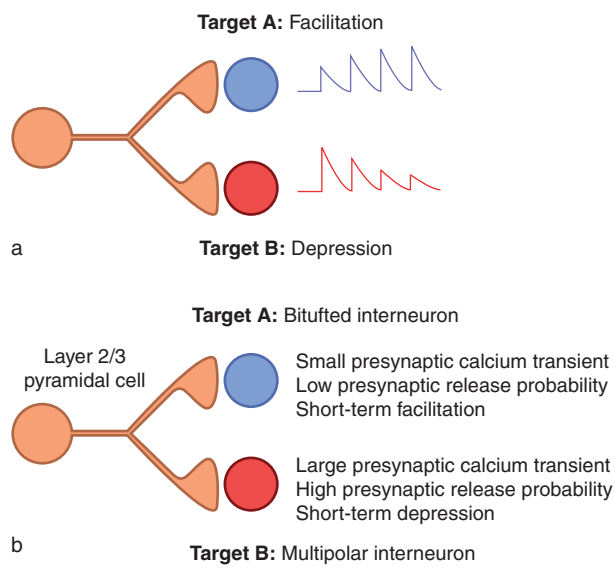

Figure 2 Retrograde specification of presynaptic release probability. (a) A single neuron is able to make synapses of different types, simultaneously, at two different target neurons. Synaptic terminals at target A have a low probability of release and facilitate during a train of action potentials. Synaptic terminals at target B have a high probability of release and depress during a train of action potentials. (b) In the rat cortex, layer 2/3 pyramidal cells are able to make synaptic connections simultaneously with bitufted and multipolar interneurons. Synapses contacting bitufted interneurons have a small presynaptic calcium transient and low-release probability and express facilitation. Synapses with multipolar interneurons have a large presynaptic calcium transient and high-release probability and express short-term depression.

for neural circuit function and animal behavior. For example, in the leech nervous system, a single sensory neuron makes low-probability, facilitatory synapses with one motoneuron (MN1) and high-probability, depression-prone synapses with a second motoneuron (MN2). In response to noxious stimulation, the sensory neuron first recruits MN2, leading to a contraction of the skin. After a short delay, synaptic facilitation brings MN1 to threshold, leading to a body bend away from the noxious stimulus. In this manner, the sequential emergence of two distinct behaviors is achieved by a single sensory neuron. As another example, in the arthropod, distinct classes of external sensory organs encode information about wind velocity versus acceleration. Acceleration detectors contact one postsynaptic target with low-probability, facilitatory synapses while velocity sensors contact a second target with high-probability, depression-prone synapses. A small number of sensory organs encode both velocity and acceleration and contact both postsynaptic targets, simultaneously making synapses appropriate for each target cell type. Again, it appears that the target cell specifies the release properties of impinging presynaptic nerve terminals, and this is directly relevant to the function of the surrounding neural circuit and animal behavior.

Nearly identical phenomena have been observed in the mammalian central nervous system (CNS). For example, layer 2/3 pyramidal neurons in the cortex contact bitufted interneurons with low-probability, facilitating synapses, and simultaneously contact multipolar interneurons with a higher probability, depression-prone synapse. Similar phenomena have been observed in other regions of the mammalian CNS. In each example, target-dependent specification of presynaptic release properties is predicted. Interestingly, in the cortex, target specification of presynaptic release probability changes over the course of development. It is interesting to speculate that stable, steady-state release properties of the presynaptic terminal could be modulated by retrograde signaling systems over time. Ultimately, the retrograde signal that is responsible for specifying presynaptic release probability and associated short-term plasticity is unknown.

Retrograde Signaling and the Homeostatic Modulation of Presynaptic Function

Homeostasis refers to the ability of a cell or a system of cells to respond to a perturbation and maintain a constant physiology. Experimental evidence now supports the existence throughout the nervous system of homeostatic signaling systems that control the firing properties of individual cells. Based upon these data, it is hypothesized that homeostatic signaling systems interface with the mechanisms of neural plasticity in order to maintain stable neural function over time.

Experimental support for homeostatic control of neural function has been documented throughout the central and peripheral nervous systems. In each case, (1) homeostatic regulation was observed following an experimental perturbation that altered the steady-state electrical properties of nerve or muscle cells, (2) the cells responded to the experimental perturbation by altering ion channel density, neurotransmitter receptor density, or the probability of neurotransmitter release at impinging presynaptic nerve terminals, and (3) the modulation of these parameters restored cellular activity to baseline levels observed prior to the experimental perturbation. This precise restoration of baseline (set-point) cellular activity is a fundamental property of homeostatic signaling.

There are several examples wherein homeostatic regulation of nerve or muscle activity is achieved by a retrograde signal that modulates synaptic transmission. For example, central neurons, both *in vitro* and *in vivo*, that experience chronic activity blockade undergo a wide range of compensatory, homeostatic changes (**Figure 3**). Chronic activity blockade leads to altered ion channel density at the cell surface as well as altered surface expression of both excitatory and inhibitory neurotransmitter receptors. In several cases, the impinging presynaptic nerve terminals are also modified in a homeostatic manner. Presynaptic changes include altered neurotransmitter release, altered active zone size, and altered synaptic vesicle pool size. Some of these effects may be a direct consequence of altered postsynaptic neurotransmitter receptor abundance. However, other changes, such as presynaptic release probability, may require active, homeostatic signaling from the postsynaptic cell to the presynaptic element. The molecular basis of such a transsynaptic, homeostatic signaling system is unknown. Interestingly, a glial-derived intercellular signal mediated by tumor necrosis factor-α (TNF-α) is both necessary and sufficient for the homeostatic modulation of glutamate receptor abundance in the CNS. Whether this signal could mediate altered presynaptic release remains to be determined.

The neuromuscular junction of organisms ranging from *Drosophila* to rodents and humans shows clear evidence of a homeostatic signaling system that requires a retrograde signal from muscle to nerve. For example, at the *Drosophila* NMJ, chronic impairment of postsynaptic neurotransmitter receptor sensitivity leads to a compensatory increase in presynaptic transmitter release that precisely counteracts the change

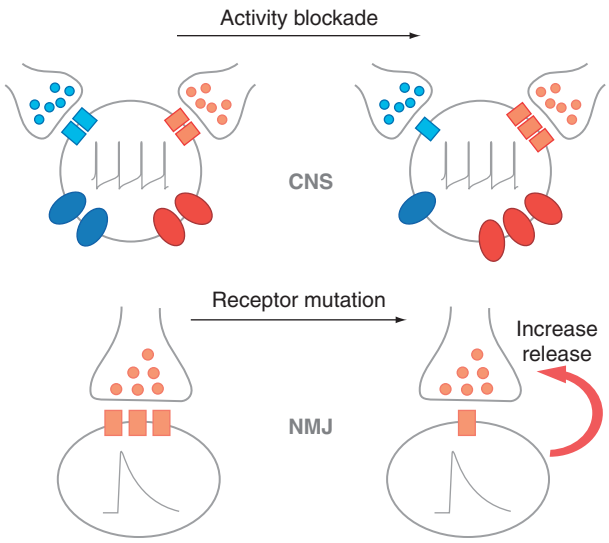

Figure 3 Homeostatic control of neural function. Top: The activity of a neuron is determined by a balance of synaptic excitation (synapses with red vesicles and red receptors), synaptic inhibition (blue vesicles and blue receptors), and the abundance of ion channels that depolarize the cell (red ovals) or hyperpolarize the cell (blue ovals) in the central nervous system (CNS). In response to pharmacological activity blockade (black arrow), central neurons can change the abundance of ion channels to reestablish baseline neural activity. Chronic activity blockade can also cause homeostatic changes in the surface expression of neurotransmitter receptors, including excitatory (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid)-, inhibitory (γ-aminobutyric acid)-, and *N*-methyl-D-aspartate-type postsynaptic neurotransmitter receptors. Bottom: At the neuromuscular junction (NMJ), impairment of postsynaptic neurotransmitter receptor sensitivity leads to a compensatory increase in presynaptic neurotransmitter release that precisely offsets the change in receptor function, to achieve normal synaptic depolarization of the muscle. Adapted from Marder E and Prinz AA (2002) Modelling stability in neuron and network function: The role of activity in homeostasis. *Bioessays* 24: 1145–1154; and Davis GW (2006) Homeostatic control of neural activity: From phenomenology to molecular design. *Annual Review of Neuroscience* 29: 307–323.

in receptor function to achieve normal muscle depolarization (Figure 3). Several variations on this experiment have been performed. For example, impairing postsynaptic membrane excitability directly, through overexpression of an inward rectifying potassium channel (Kir2.1), initiates a similar retrograde, homeostatic increase in presynaptic transmitter release. Conversely, increasing muscle excitation by superinnervating the muscle leads to a compensatory decrease in presynaptic release probability. Nearly identical experimental results have been obtained at the rodent and human NMJ, demonstrating the existence of a highly conserved, homeostatic, retrograde signaling system.

Although the nature of the retrograde signaling systems capable of increasing and decreasing presynaptic release remains unknown, a recent series of experiments has defined several parameters that specify how this retrograde transsynaptic signaling system may work. First, it was demonstrated that a homeostatic increase in presynaptic release could be achieved in 10 min following application of sub-blocking concentrations of a postsynaptic glutamate receptor antagonist (philanthotoxin). In these experiments, evidence of altered presynaptic release could be observed in as little as 2 min, suggesting that the homeostatic signaling system was able to detect altered receptor function and initiate a change in presynaptic release in a time frame of seconds to minutes, far more rapidly than previously thought. This study also demonstrated that the induction of homeostatic signaling is independent of new protein synthesis and is independent of neural activity. The experimenters suggested that the spontaneous release of neurotransmitter that persists in the absence of activity-dependent neurotransmitter release is sufficient to induce the retrograde, homeostatic modulation of presynaptic release following postsynaptic receptor inhibition.

Another interesting feature of retrograde, homeostatic signaling is that presynaptic transmitter release is precisely (quantitatively) tuned by the retrograde signaling system to correctly offset the magnitude of postsynaptic receptor perturbation. For each synapse analyzed, the decrease in postsynaptic receptor sensitivity, estimated by average spontaneous miniature excitatory postsynaptic potential (mEPSP) amplitude, is precisely offset by a compensatory increase in presynaptic vesicle release such that muscle depolarization (EPSP amplitude) is constant. Thus, the retrograde homeostatic signaling system is not only rapid, but is also sensitive enough to detect small changes in postsynaptic receptor sensitivity. Furthermore, the retrograde signal is able to convert the magnitude of the postsynaptic perturbation into a precise change in presynaptic release. Such a precise modulation of presynaptic release is quite remarkable when compared to other forms of synaptic plasticity.

The molecular mechanisms of rapid, precise, homeostatic modulation at the *Drosophila* NMJ have yet to be clearly defined. However, forward genetic screens are beginning to identify some of the molecular players. It was recently shown that mutations that perturb the functionality of a presynaptic $Ca_V2.1$ calcium channel homolog abolish the rapid induction and sustained expression of synaptic homeostasis at the *Drosophila* NMJ. The involvement of a presynaptic calcium channel fits with the rapid, precise modulation of presynaptic release in this system. The requirement of a presynaptic calcium channel for synaptic homeostasis induced by postsynaptic receptor inhibition also supports the existence of a retrograde, transsynaptic signaling system.

The nature of the retrograde signal at this synapse remains to be identified. Recent experiments have demonstrated that bone morphogenetic protein (BMP) signaling strongly influences the growth and stability of the *Drosophila* NMJ. These experiments provided evidence that the BMP ligand, named glass-bottom-boat, can be released from muscle and can signal to BMP receptors on the presynaptic motoneuron. The presynaptic receptors and the downstream canonical signaling pathway to the motoneuron nucleus are necessary for normal synaptic growth. Recent experiments demonstrate that the BMPs are necessary for the homeostatic modulation of presynaptic release in response to postsynaptic glutamate receptor blockade. However, because the BMPs are also necessary for both synaptic growth and synapse stability, it remains unclear whether the BMPs represent a homeostatic signal, or whether the NMJ is simply crippled following the loss of BMP signaling. Thus, the precise nature of the retrograde signal at the NMJ remains to be clearly defined.

Conclusions

Retrograde signaling during synapse formation and modulation is a common process. Retrograde signals are diverse and include direct transsynaptic linkages, growth factors, morphogens, and endocannabinoids. Despite remarkable advances, many phenomena exist that seem to require retrograde signaling systems of unknown identity. Forward genetic and proteomic approaches to neural development may hold the key to identification of these elusive transsynaptic signaling systems.

See also: Axonal Pathfinding: Extracellular Matrix Role; Bone Morphogenetic Protein (BMP) Signaling in the Neuroectoderm; Cell Adhesion Molecules at Synapses; Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin; Neural Induction in Chicks; Neuroproteomics; Postsynaptic Development: Neuronal Molecular Scaffolds; Receptor Trafficking; Retrograde Neurotrophic Signaling; Spine Plasticity; Synapse Formation: Competition and the Role of Activity; Synaptic Precursors: Filopodia; Voltage-Gated Calcium Channels.

Further Reading

Davis GW (2006) Homeostatic control of neural activity: From phenomenology to molecular design. *Annual Review of Neuroscience* 29: 307–323.

Goda Y and Davis GW (2003) Mechanisms of synapse assembly and disassembly. *Neuron* 40: 243–264.

Koester HJ and Johnston D (2005) Target cell-dependent normalization of transmitter release at neocortical synapses. *Science* 308: 863–865.

Kummer TT, Misgeld T, and Sanes JR (2006) Assembly of the postsynaptic membrane at the neuromuscular junction: Paradigm lost. *Current Opinion in Neurobiology* 16: 74–82.

Marder E and Prinz AA (2002) Modelling stability in neuron and network function: The role of activity in homeostasis. *Bioessays* 24: 1145–1154.

Marder E and Goaillard JM (2006) Variability, compensation and homeostasis in neuron and network function. *Nature Review Neuroscience* 7: 563–574.

Marder E and Prinz AA (2003) Current compensation in neuronal homeostasis. *Neuron* 37(1): 2–4.

Perez-Otano I and Ehlers MD (2005) Homeostatic plasticity and NMDA receptor trafficking. *Trends in Neuroscience* 28: 229–238.

Pun S, Santos AF, Saxena S, et al. (2006) Selective vulnerability and pruning of phasic motoneuron axons in motoneuron disease alleviated by CNTF. *Nature Neuroscience* 9: 408–419.

Reyes A, Lujan R, Rozov A, et al. (1998) Target-cell-specific facilitation and depression in neocortical circuits. *Nature Neuroscience* 1: 279–285.

Sanes JR and Lichtman JW (2001) Induction, assembly, maturation and maintenance of a postsynaptic apparatus. *Nature Reviews Neuroscience* 2: 791–780.

Scheiffele P (2003) Cell–cell signaling during synapse formation in the CNS. *Annual Review of Neuroscience* 26: 485–508.

Shen K, Fetter RD, and Bargmann CI (2004) Synaptic specificity is generated by the synaptic guidepost protein SYG-2 and its receptor, SYG-1. *Cell* 116: 869–881.

Stellwagen D and Malenka RC (2006) Synaptic scaling mediated by glial TNF-alpha. *Nature* 440: 1054–1059.

Turrigiano GG and Nelson SB (2004) Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience* 5: 97–107.

Retronasal Olfaction

B P Halpern, Cornell University, Ithaca, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Definitions of Retronasal Smelling and Retronasal Olfaction

Preliminary definition A literal meaning of retronasal is behind the nose. Olfaction can mean smelling. Consequently, retronasal olfaction can mean smelling from behind the nose. That definition immediately raises two questions: Where is behind the nose, and what is there to be smelled? In this context, nose denotes not only the protruding part of the face that is equipped with the nostrils but also the space (nasal cavity) and nasal structures that immediately follow each nostril (Figure 1). During respiration with the lips closed, air moves through the nasal cavities and across their structures, as well as through the pharynx. When exhalation occurs with the lips closed, air from the respiratory system travels through the pharynx, then through the nasal cavity, and exits from the nostrils.

Final definitions The region of the pharynx closest to the nasal cavity is called the nasopharynx. Farther away from the nose, the pharynx forms the back of the mouth (also called the oral cavity). For a typical mammal that moves around on four legs, the region behind the nose includes the oral cavity. In humans and other bipedal mammals, the upright posture changes behind to below for the oral cavity. Nonetheless, the oral cavity is after, or back from, the nose, if one thinks of inhaling air through the nose and then follows the inward airflow path. This inward airflow pathway, beginning at or near the nostrils and passing through the nasal cavities into the pharynx, is called the orthonasal pathway (Figure 1).

The exhalation airflow pathway, along which air flows in a direction opposite to orthonasal movement, is known as the retronasal airflow path (Figure 1). During an exhalation, air will always move through the pharynx. Because the pharynx is the back of the oral cavity, a respiratory exhalation will include some air from the oral cavity. From this oropharyngeal region, expired air may primarily exit from the mouth if the lips are parted, or retronasal airflow may move through the nasal cavities if the lips are closed (airflow from the oropharyngeal region can be excluded from the nasal cavities by elevation of the velum (Figure 1) or by closure of the nostrils). If there are substances in the oral cavity air that could be smelled (all such substances are known as odorants), and if those odorants, at a sufficient concentration, reach appropriate sensory receptors, then retronasal smelling will occur. Therefore, retronasal smelling denotes smelling odorants that originate from the oral cavity (and odorants coming from the oral cavity are called retronasal odorants). When retronasal smelling occurs with air exiting entirely from the mouth, with none reaching the nasal cavities, the relevant sensory receptors must be in the oropharyngeal region (Figure 1).

Retronasal smelling becomes retronasal olfaction if the smelling of odorants that originate from the oral cavity occurs only after the odorants reach the nasal cavity, and stimulates there only the olfactory receptors (see the section titled 'Nasal cavity receptors'). Stated differently, all smelling of odorants from the oral cavity can be called retronasal smelling. Smelling from the oral cavity is probably retronasal olfaction if the odorants must reach a nasal cavity to be smelled, and is unequivocally retronasal olfaction if the odorants require an intact olfactory system to be smelled (see the section titled 'Nasal cavity sensory systems').

Odorants from the Oral Cavity

What may be smelled from the oral cavity largely depends on what one chooses to put there. Under normal circumstances, all food or drink enters through the oral cavity. Once in the mouth, ingested foods change as they mix with saliva and are chewed and manipulated. This modifies both the concentrations and the characteristics of the food-derived odorants available for retronasal smelling. Physical measurements verify that odorants released from foods in the oral cavity (i.e., retronasal odorants) can reach a nasal cavity and exit via a nostril (Figure 2). In addition, separate from eating or drinking, pathological conditions of gums, tongue, or other oral cavity tissues can result in oral cavity odorants.

Sensory Systems for Retronasal Smelling and Retronasal Olfaction

Oral cavity receptors There are a number of sensory receptors in the oral cavity that respond to chemical, thermal, or mechanical properties of substances in the mouth. Some of these receptors can be stimulated by certain odorants. Those oral cavity sensory receptors that respond to vapor phase (dissolved or suspended in air) odorants are part of the trigeminal sensory system (Figure 1). The odorants that can stimulate the receptors of the trigeminal sensory system,

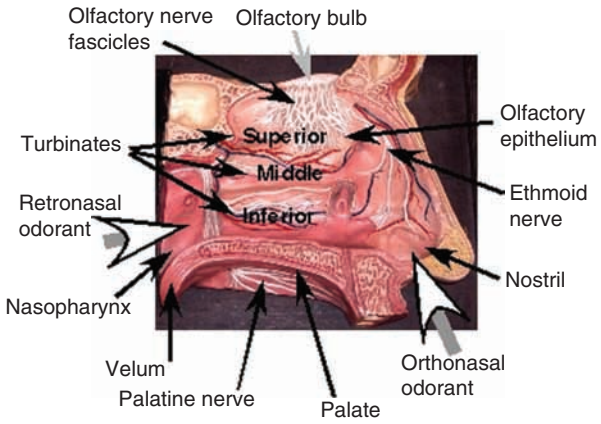

Figure 1 Model of the human left nasal cavity. The nasal valve (not shown) is a region of deformable, elastic cartilage not very far (1–2 cm) from the nostril. Much of the nasal cavity is innervated by branches of the trigeminal nerve (see **Figure 3**); the ethmoid and palatine branches are shown. Within the nasal cavity, the olfactory epithelium, where the olfactory receptor neurons (ORNs) are located, is depicted on the upper part of the middle turbinate and on the superior turbinate. The area labeled 'Olfactory epithelium' also represents the narrow cavity called the olfactory cleft. The axons of the ORNs, depicted as fascicles (nerve bundles) of the olfactory nerve, are shown originating in the olfactory epithelium, passing through the skull, and terminating in the olfactory bulb of the central nervous system. The orthonasal smelling pathway, with odorant entering through a nostril, is shown to the right of the figure. On the left, the retronasal smelling pathway, with odorant entering from the nasopharynx, is depicted. When the velum (far left portion of the soft palate) is down, as shown, a connection between the oral and nasal cavities is available. Below the nasopharynx, the pharynx continues, forming the back of the oral cavity (not shown). The palate is the roof of the oral cavity.

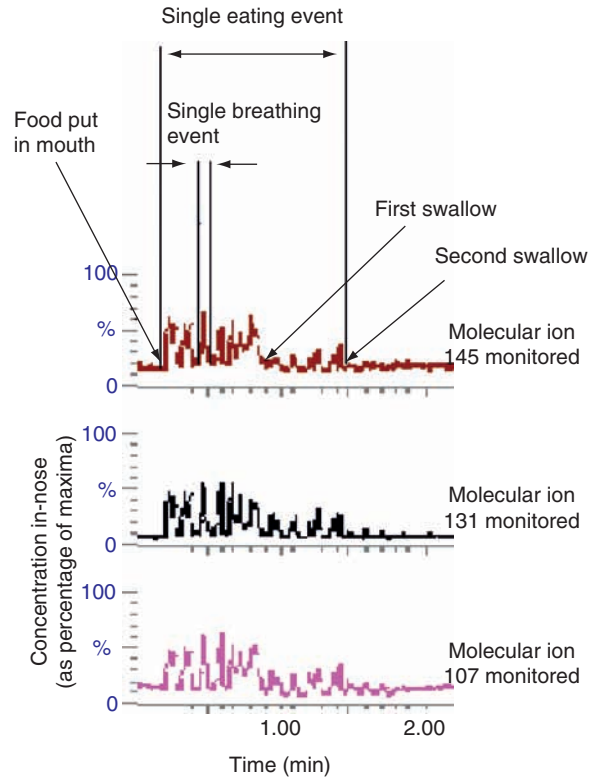

Figure 2 Mass spectrometer (MS-Nose) analyses of ions in the breath of a human while eating an imitation cheese. Input to the MS-Nose was from a nostril via an open tube and a perpendicular capillary. The cheese was in bite-size pieces and was eaten naturally, without constraints. Molecular ion 145 was ethyl hexanoate; 131 was isoamyl acetate; and 107, benzaldehyde. All three molecules had been incorporated into the imitation cheese. The known molecular ions appeared with each exhalation and disappeared after one or two swallows. Modified from Deibler KD (2001) *Measuring The Effect of Food Composition on Flavor Release Using the Retronasal Aroma Simulator and Solid Phase Microextraction*. PhD Thesis, Cornell University, with permission.

whether in the oral cavity or elsewhere (e.g., the nasal cavities), are called trigeminal odorants.

Each time a terrestrial mammal exhales, some of the air in the oral cavity, and the odorants in it, moves as part of the exhaled air. As already noted, with the lips closed, airflow during exhalation will reach and pass through the nasal cavities. Nonetheless, during an early phase of such retronasal airflow, if trigeminal odorants are present, stimulation by trigeminal odorants of trigeminal receptors in the oropharyngeal region may occur before these retronasal odorants reach the nasal cavities.

Nasal cavity receptors If one exhales with the lips closed, the air sampled from the oral cavity travels through the nasal cavities. If there are vapor phase chemicals in the oral cavity, they will reach the nasal cavities and outflow through the nostrils (**Figure 2**). As is the case for the oral cavity, there are many sensory receptors in the nasal cavities that can respond to chemical, thermal, and mechanical events. Among these nasal cavity sensory receptors are trigeminal receptors that respond to trigeminal odorants (see **Figures 1 and 3**). These nasal cavity trigeminal receptors may be stimulated not only during retronasal airflow (exhalation) but also during orthonasal airflow (inhalation).

In vertebrates the receptors of another sensory system that responds to odorants occur only in the nasal cavities. They are the receptors of the olfactory system, that is, olfactory receptor neurons (ORNs; see the section titled 'Nasal cavity sensory systems'). As a group, ORNs are responsive to very low concentrations of many different odorants, with each ORN responsive to only some odorants. ORNs can be stimulated during exhalations (retronasal airflow) or inhalations (orthonasal airflow).

There is a nonsymmetrical relationship between ORN responses to odorants and trigeminal receptor

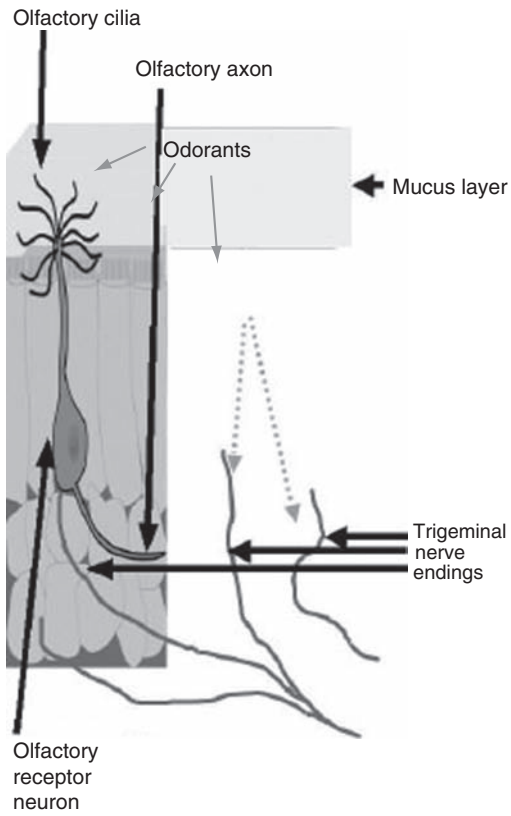

Figure 3 Schematic representation of the sensory innervation of the nasal cavity. An olfactory receptor neuron (ORN) of the olfactory epithelium is depicted surrounded by supporting cells, with the most distal part of its dendrite, and many of its cilia, extending beyond the cellular olfactory epithelium into the noncellular, continuously secreted mucus layer that overlies the epithelium. The axon of the ORN extends toward the olfactory bulb (not shown). The trigeminal nerve endings (lower portion of figure; also called trigeminal receptors) are depicted within the epithelia, both in the olfactory epithelium (left portion of figure) and in the nonolfactory respiratory epithelium (right portion of figure). In the mucus layer, odorants which had partitioned from respiratory air are shown. Once in the mucus layer, odorants can reach and stimulate the cilia of the ORN that are within it (indicated by short, solid arrow), but odorants must travel a longer distance, penetrating into the cellular epithelia (long, dotted arrows), before trigeminal stimulation is possible. Modified from a schematic drawing provided by Charles J Wysocki, Monell Chemical Senses Center, Philadelphia, PA.

responses: because of the high sensitivity and broad responsiveness of the population of ORNs, some of the ORNs will probably respond to any trigeminal odorant, but trigeminal receptors will not respond to some olfactory odorants and can respond to certain others only at high concentrations. The odorants that appear to stimulate only ORNs are known as purely olfactory odorants. In human they include vanillin, phenylethyl alcohol, and octanoic acid. Conversely, very effective human trigeminal odorants (which will also, of course, stimulate ORNs) include heptyl alcohol, DL-menthol, and valeric acid.

Air and Odorant Movement and Access to Sensory Receptors

Fluid flow Air is a fluid. It is less dense than liquid fluids such as liquid water, but nonetheless, movement of air has many similarities to movement of water. Because we usually cannot see the movement of air, it can be difficult to think about what happens when moving air abuts surfaces. With reference to the nasal and perhaps oral cavities, the effects on the flow of air of interactions between air and the surfaces or objects that air contacts as it moves can be important.

Considering liquid water instead, imagine a relatively small stream. Now imagine a rock or similar object near one bank of the stream, not blocking the stream but protruding into it. Far from the rock, both before it and after it, the water may move smoothly, with no obvious effects of interactions between the rock in the moving fluid and the fluid. However, in the immediate vicinity of the rock, turbulence develops, eddies form in some places, still water in others. The result is that contact between the water (and whatever is dissolved in the water) and the bank and bed of the stream is changed in the immediate vicinity of the object that protrudes into the flowing fluid.

Suppose that at one point the stream widens, so that one side continues at the usual depth but the other side is shallower. Most of the water will continue to flow in the deeper region, but a small proportion will reach the wider area. The small volume of water flowing through the wider region may move more slowly and is likely to contain fewer of the particles that are suspended in the rapidly flowing water. Because of this, the banks and bed of the shallower area are protected from mechanical damage that may occur in the rest of the stream. The concentrations of substances dissolved in the water are probably equal in the large flow volume and the small flow volume regions, but the total amount of the substances present in the small flow volume region must be less.

Nasal cavities The hypothetical effects of a rock in a stream of water have similarities to moving respiratory air and the odorants dissolved in it. As the air passes through the nasal cavities, small protuberances in the narrow and complex spaces of a nasal cavity can have substantial local effects on the pattern of air movement. The consequence is that odorants can more easily reach some parts of the nasal cavities than others. Because of this, some of the sensory receptors of the nasal cavities can receive greater amounts of odorant stimulation than other nearby

regions that are separated by a small bump or a sharp turn. In addition, the characteristics of some nasal cavity structures result in differences between the flow patterns of inhaled versus exhaled air. These differences mean that odorant access during retronasal airflow can be different from that during orthonasal airflow (see the section titled 'Functional anatomy of retronasal olfaction and retronasal smelling').

A similarity also exists between airflow in the nasal cavities and the hypothetical widened region of the stream, described above, that received only a small proportion of the total fluid flow. In the nasal cavities, most of the respiratory air flows across structures that have only trigeminal sensory receptors (Figure 1). Only a small proportion of the flowing air arrives at the ORNs of the olfactory epithelium, which is located on nasal cavity structures (turbinates) away from the main airflow pathway. Because of this, the amounts of odorants that can have access to the ORNs are much less than the total amount in respiratory air.

Oral cavity It is possible that airflow in the oral cavity may be complex in some locations. Unfortunately, adequate data are not available.

Functional Anatomy of Retronasal Olfaction and Retronasal Smelling

Nasal Cavity Structure

A region of high resistance to airflow near each nostril of the nasal cavities can modulate movement of respiratory air. This region, the nasal valve (there may be more than one in humans; Figure 1), functions differently during exhalation and during inhalation. One result is that the pattern of retronasal odorant movement in the nasal cavities differs from that of orthonasal odorants (orthonasal odorants are odorants originating at or near the nostrils). After the valve(s), narrow, twisting spaces through which respiratory air must flow can be created by epithelium-covered curved bones (the turbinates). The size, shape, and number of turbinates vary considerably among mammals. As the air flows, some of the odorants it carries adhere to the mucus that covers the turbinates, with the physical and chemical properties of the odorants allowing differential partitioning into the mucus (Figure 3). This permits odorants to reach the sensory receptors that protrude into (cilia of ORNs) or lie beneath (trigeminal receptors) the mucus. In addition, transfer of odorants into the mucus reduces the concentrations of odorants available for subsequent receptors. Therefore, because overall respiratory flow within the nasal cavities is in different directions for exhalation and inhalation, the locations of maximum and minimum odorant concentrations may be quite different during retronasal and orthonasal airflow. The general concept of selective transfer of odorants from the respiratory air onto and into the mucus that overlies the epithelium-covered turbinates has been termed the gas chromatographic model of olfaction.

Nasal Cavity Sensory Systems

Two different sensory systems supply receptors in a nasal cavity (Figures 1 and 3). Trigeminal receptors occur on all its surfaces, but ORNs are located only in an area away from the main path of respiratory air. In humans, the olfactory epithelium occurs on the superior turbinate and the upper part of the middle turbinate (Figure 1). This area, the olfactory cleft, receives air that is at body temperature, is humidified, and is free of large suspended particles because of the twisted airflow paths produced by the turbinates. A consequence is that only the chemicals in respiratory air are available to stimulate the ORNs. In some mammals, perhaps including humans, ORNs that respond well to one type of odorant are grouped together, somewhat separate from ORNs that respond well to another type of odorant. Because of this local grouping of ORNs, the opposite flow directions of retronasal and orthonasal olfaction may produce opposite patterns of ORNs responses.

Olfactory receptors are not permanent Individual ORNs are not permanent structures. In vertebrates, including humans, each is replaced in normal adults on a timescale of several months. Every new ORN must extend to and make appropriate connections in the central nervous system (CNS; see the section titled 'CNS connections') in order to maintain a functional olfactory system.

Trigeminal receptors are not replaced In contrast, trigeminal sensory neurons, in common with all other nonolfactory peripheral sensory neurons, are not replaced in mammals. Recovery is possible if the receptor end is damaged, but loss of an entire trigeminal neuron is permanent.

CNS Connections

Olfactory bulb Each ORN has an axon that can carry into the CNS action potentials generated at the receptor end (Figure 3). All these axons end in a CNS structure, the olfactory bulb (Figure 1), where thousands synapse with one of the principal neurons of the olfactory bulb (called mitral and tufted cells). These synaptic regions are known as glomeruli.

In some mammals, and perhaps in humans, the axons of ORNs that respond better to one type of odorant than to other types of odorants (related to the molecular features of the odorants) converge into a few glomeruli. Such odorant-dependent (molecular feature-related) convergence can produce glomerular regions that are functionally distinct from other regions. This may yield, in the olfactory bulb, clusters of neural activity that are related to the odorants present at that time at the cilia of the various ORNs. For some fixed sets of odorants, retronasal olfaction may give bulbar neural activity patterns that are different from those for orthonasal olfaction. This can occur because odorants in retronasal airflow may reach, and stimulate, some ORN regions earlier than would odorants in orthonasal airflow. In addition, the opposite concentration profiles resulting from opposite flow directions across the mucus will contribute to different active olfactory bulb regions across time.

The olfactory bulb is a complex CNS structure. Its mitral and tufted cells have connections to each other through small neurons within the olfactory bulb (interneurons) which transform the input from the ORNs. The mitral and tufted cells themselves have axons which leave the olfactory bulb and reach other parts of the CNS. In addition, axons from other CNS areas enter the olfactory bulb and affect its neural activity.

Trigeminal CNS nuclei The trigeminal sensory receptors, which are the sensory ends of neurons, have axons which enter the CNS quite separately from the axons of ORNs and synapse in very different locations. The trigeminal axons from the oral and nasal cavities go to one or more hindbrain nuclei, including the principal (i.e., sensory) nucleus and the spinal nucleus of the trigeminal nerve. The synaptic positions in these nuclei reflect the locations of trigeminal receptors in epithelia of the nasal or oral cavities, that is, a somatotopic mapping. Because of this spatial correspondence between trigeminally innervated regions in the nose or mouth and areas of termination in the brain, the patterns of activity in these CNS nuclei may reflect the movement of trigeminal odorants through the oral or nasal cavities. Consequently, the pattern of activity in trigeminal regions of the CNS in response to retronasal smelling may differ from that in response to orthonasal smelling.

Human Retronasal Responses

Psychophysical Judgments

Airflow patterns and odorant distributions in the nose and mouth during inhalation and exhalation are known to differ. These distinctions between orthonasal and retronasal airflow are related to the structural characteristics of the oral and nasal cavities. The olfactory or trigeminal sensory systems innervate these cavities. Their molecular-feature or topographic-related terminations in the CNS may have functional importance for retronasal smelling. However, interest in these anatomical, chemotopic, and fluid-dynamical data must be motivated by observations that psychophysical measures of retronasal and orthonasal smelling show noteworthy differences or important interactions. In addition, reports that retronasal smelling can show dysfunction in humans would also be important. Both the psychophysical differences and selective retronasal pathology are described in the next section and the section titled 'Clinical relevance.'

Thresholds For humans to detect them, odorants from a retronasal location, that is, normally originating from the oral cavity, must be at a higher concentration than the same odorants from an orthonasal location (at or near the nostrils). Stated more succinctly, detection thresholds are higher for odorants from a retronasal location than from an orthonasal location. This difference could be the result of disparate airflow and odorant delivery from the two loci, or it might reflect higher sensitivity of the sensory systems if odorants enter the nasal cavity from the nostrils. The latter possibility, greater sensitivity for odorants from a orthonasal location, was confirmed by delivering odorants at constant flow rates through tubes that were placed just after the nasal valve for a orthonasal locus or just beyond the nasal cavity (in the nasopharynx) for a retronasal location (see Figure 1). A physical measurement, based on odorants arriving in the olfactory cleft, demonstrated equal concentrations at the olfactory epithelium.

Identification Identifications of odorants from a retronasal location are less accurate than identifications when the odorants arrive from a orthonasal location if relatively low concentrations are used or if only quiet normal breathing is permitted. However, differences between orthonasal and retronasal identification accuracy diminish or disappear if higher concentrations are employed, if deep breathing is encouraged, or if the concentration are selected to provide equal retronasal and orthonasal intensity (i.e., isointense concentrations). Thus, rather than a difference in identification ability, a retronasal sensitivity that is less than orthonasal may be the cause. Nonetheless, a need for higher concentrations may not be the only factor, because odorants, when in a retronasal location, may be described differently than

when in an orthonasal location. This implies that, in addition to diminished sensitivity, qualitative processing of retronasal stimulation may differ from orthonasal processing.

Intensity There is a lack of consistency with regard to effects of odorant location on intensity. For the odorants lavender, butanol, farnesol, and hydrogen sulfide (possibly a trigeminal odorant), delivered either near a nostril (orthonasal location) or in the nasopharynx (retronasal location), retronasal intensity is less than orthonasal. However, applying the same procedure using phenylethyl alcohol (purely olfactory odorant) and chocolate, no difference occurs between retronasal and orthonasal intensity. It is not obvious why there are intensity differences for some odorants but not for others. With a more conventional odorant delivery procedure, odorants that had first been selected to be isointense retronasally and orthonasally are less intense retronasally when judged over time (time–intensity tracking). In this instance the psychophysical procedures may explain the difference.

Adaptation In most sensory systems, prolonged exposure to an unchanging stimulus results in a reduced response magnitude. This phenomenon is known as adaptation. The rate of response decline, the presence or absence of a brief, initial high magnitude response (phasic or dynamic component), and the existence of a sustained response to continuing stimulation (tonic or static component) all differ between sensory systems and sometimes between types of receptor within a sensory system (e.g., slowly versus rapidly adapting cutaneous mechanoreceptors). Olfactory adaptation to orthonasal odorants has been extensively described in the literature, but few data are available for retronasal adaptation. Direct comparisons using anise, orange, coffee, peppermint, and strawberry odorants found a general similarity between retronasal and orthonasal adaptation over 60 s. However, for odorants in the orthonasal location, but not in the retronasal, intensity rose before declining. This result (a phasic component only to odorants from an orthonasal location) suggests that there is a CNS processing distinction between retronasal and orthonasal smelling. Replication and further exploration of these observations will be informative.

Interactions Under typical eating conditions for humans and other terrestrial mammals, food-derived odorants originating from the oral cavity and via the nostrils may occur successively. Once in the mouth, ingested foods change as they mix with saliva and are chewed and manipulated. This modifies both the concentrations and the characteristics of the food-derived odorants available for retronasal smelling, compared with the unmodified odorants delivered via the orthonasal route. With natural breathing, a brief interval (1–2 s in humans) would separate stimulation via the oral cavity from orthonasal input during eating because there is typically a pause between inhalation and exhalation. This temporal separation might prevent sensory interactions between retronasal and orthonasal smelling or might limit interactions to an overall reduction in identifications. Surprisingly, the accuracy of identifications of odorants smelled by humans in this cyclical manner (without sniffing) was lower only for orthonasal odorants. This outcome could reflect processing effects or attention differences that favor retronasal smelling under these circumstances. However, the measurements were made after the odorants were removed. It follows that odorant memory was an important factor, perhaps with unequal consequences for orthonasal and retronasal odorants. In addition, the odorant smelled second (i.e., last) was generally reported as having been smelled first, irrespective of its oral (retronasal) or nasal (orthonasal) origin. This recency effect is commensurate with prior observations that odorant-based judgments of serial order are often unreliable. Overall, these data reveal interesting and potentially important differences between retronasal and orthonasal smelling. However, their relevance to normal eating is unclear because under natural conditions, memory would be a less significant factor.

CNS Responses in Humans

Odorant-evoked potentials Electroencephalogram-type recording made in synchrony with a brief sensory event may yield a CNS event-related potential (ERP). Generally, stimuli must be repeated many times and averaged to obtain a measurable ERP. ERP is a direct measure of CNS bioelectric activity, albeit of relatively large populations of neurons. ERP studies with auditory or visual stimuli can use large numbers of very brief stimuli, with short time intervals between the presentations. This is the case because the stimuli are rapidly transduced (converted from environmental energy to cellular biological activity), with the mechanical or electromagnetic energy fully dissipated. When the stimuli are odorants, both access to vertebrate receptors and transduction are slower. Perhaps more important, the odorant chemicals remain within the mucus and the epithelium, in the immediate vicinity of the receptors,

for an extended time. Because of these factors, studies of olfactory ERPs (OERPs) generally must employ stimulus durations of hundreds of milliseconds and must allow long intervals between presentations. A result is that about 30 min may be required for a sufficient number of odorant presentations, compared with 1–2 min for auditory or visual stimuli. An additional complexity is that odorant stimulation of terrestrial vertebrates normally requires the movement of air, thus adding a potential mechanical stimulation to the chemical event of interest. The possibility of responses to air movement itself may not be troublesome when investigators are recording directly from olfactory receptors or their axons, but it must always be controlled for in OERP studies.

Human retronasal and orthonasal OERPs were compared by means of a presentation of 200 ms of phenylethyl alcohol or H_2S, separated by 40 s and delivered 40 times each. Presentations were either orthonasal, delivered just beyond the nasal valve, or retronasal, delivered to the nasopharynx (see Figure 1). All OERP components of interest had mean latencies greater than 400 ms and would be considered long-latency cerebral cortical components, perhaps with substantial cognitive aspects. The amplitude of N1 (first negative-going component) did not differ as a function of odorants originating from retronasal or orthonasal loci, although N1 retronasal latency was longer for phenylethyl alcohol. However, retronasal P2 (second positive-going component) amplitudes to phenylethyl alcohol were significantly smaller. For both odorants, retronasal P2 latencies were longer. The longer retronasal latency and lower amplitude to retronasal odorant delivery suggest smaller CNS response magnitudes. Paradoxically, psychophysical measures found reduced retronasal intensity for H_2S but not for phenylethyl alcohol (see the earlier section titled 'Intensity'). Thus, differences between orthonasal and retronasal OERPs can be observed, but the interpretation is unclear.

Functional magnetic resonance imaging Within the CNS, the more active neural regions will have a higher metabolic rate and will use more oxygen. Therefore, an indirect indication of increased regional neural activity is a local alteration in the oxygen level of the blood, specifically in the hemoglobin of the oxygen-carrying cells, the erythrocytes. The resultant status of vascular hemoglobin is called blood oxygen level demand (BOLD). For the intact human brain, BOLD can be measured by magnetic resonance imaging (MRI). MRI uses imposed electromagnetic fields to determine the state of the hydrogen atoms in hemoglobin, thereby indicating the extent to which oxygenated hemoglobin has changed in regions of the brain. Computer-based comparisons of MRI signals in several areas of the brain is called functional MRI (fMRI). If responses to odorants are studied with fMRI, odorless airflow control conditions must be used, as is required for OERPs. Another similarity to OERP stimulus protocols is that to yield useful fMRI data, odorant presentations must be repeated many times.

In one study, odorants were delivered in 1 s odorant pulses every 3 s for 30 s in constantly flowing air to the nasopharynx (retronasal location) or just beyond the nasal valve (orthonasal location), with the velum elevated to isolate the nasal cavities from the oral cavity. Isolation of the nasal cavities using this velopharyngeal closure technique permits delivery of odorants to be independent of respiration and precludes possible oral cavity sensory stimulation. Differential cortical activation was observed, with larger responses to odorants (especially lavender) delivered to the retronasal locus found at the base of the left central sulcus and in the ventral insula. The left central sulcus region corresponds to the area previously found for somatosensory stimulation of the oral cavity. This cortical location for responses to retronasal smelling suggests a convergence of oral cavity-related sensory inputs. Unfortunately, the basis for the differential activation is unclear, because the participants judged the lavender odorant to be more intense orthonasally. However, another odorant, chocolate, which was judged to be isointense, also produced differential activation. Chocolate, delivered to the retronasal location, gave greater responses in the medial orbitofrontal cortex as well as in subcortical regions including the hippocampus and thalamus than when it was delivered orthonasally. Differences between areas of cortical activation for odorants delivered from retronasal versus orthonasal loci support the idea that retronasal and orthonasal smelling are processed differently.

Use of an elevated velum together with odorant delivery through intranasal tubes permits control of where and when odorants are delivered. This is a desirable outcome. To secure this control, the temporal pattern of odorant presentations is intentionally separated from the breathing rhythm and from sniffing, and the possibility of oral cavity stimulation by odorants is intentionally prevented. However, there may be accompanying drawbacks. First, smelling is decoupled from its normal sensorimotor context. This could be problematic because of suggestions that a lack of sniffing, or of breathing-patterned smelling, degrades olfaction. Second, preventing a potential sequence of oral cavity retronasal odorant stimulation followed by

nasal cavity stimulation might yield CNS responses to odorants from retronasal loci that differ from those to an unimpeded sequence.

Clinical Relevance

Because retronasal smelling is important in responses to food and beverages, quality of life may be reduced if retronasal smelling is impaired. Changes in nasal cavity architecture between the olfactory cleft and the nasopharynx (**Figure 1**) may alter retronasal smelling while having little effect on orthonasal smelling. A consequence of this nasal cavity pathology could be individuals who appear to have no olfactory deficit (i.e., they are normosmic) when tested orthonasally but who nonetheless complain of changes in appreciation of food. Retronasal dysfunctions of this sort have been verified: patients have been found with reduced ability to smell odorants from a retronasal location (i.e., retronasal hyposmia) but with orthonasal normosmia. This observation suggests that if complaints about food are accompanied by orthonasal normosmia, retronasal testing should be considered.

The opposite circumstance, retronasal normosmia but orthonasal hyposmia, has been found in patients exhibiting chronic swelling of nasal and sinus membranes (chronic rhinosinusitis) and numerous swollen tissue masses that arise from the epithelium of the nasal cavities or sinuses (multiple polyposis). In this case, the dysfunction of orthonasal olfaction might be due to mechanical obstructions in the anterior portion of the olfactory cleft. However, for this symptom and the previously noted retronasal hyposmia, more-complex functional factors may be involved. This is suggested because severe deficits in the ability to smell orthonasal odorants, together with retronasal normosmia, but unaccompanied by macroscopic pathology of the nasal cavities, has been reported. Specifically, abnormal smelling of odorants from an orthonasal location (orthonasal dysosmia) or even a total absence of ability to smell orthonasal odorants (orthonasal anosmia), together with absent OERPs to orthonasal odorants, but retronasal normosmia, has been observed in patients who showed no clinical abnormalities with computed tomographic scans and nasal endoscopic examinations.

See also: Olfactory Receptors; Olfactory Receptors: Microarray Analysis; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Sensory Aging: Chemical Senses; Trigeminal Motor System.

Further Reading

Brand G (2006) Olfactory/trigeminal interactions in nasal chemoreception. *Neuroscience and Biobehavioral Reviews* 30(7): 908–917.

Deibler KD and Delwiche J (eds.) (2004) *Handbook of Flavor Characterization: Sensory Analysis, Chemistry, and Physiology.* New York: Dekker.

Doetsch F and Hen R (2005) Young and excitable: The function of new neurons in the adult mammalian brain. *Current Opinion in Neurobiology* 15: 121–128.

Doty RL (ed.) (2003) *Handbook of Olfaction and Gustation,* 2nd edn. New York: Dekker.

Halpern BP (2004) Retronasal and orthonasal smelling. *ChemoSense* 6: 1–7.

Heilmann S and Hummel T (2004) A new method for comparing orthonasal and retronasal olfaction. *Behavioral Neuroscience* 118: 412–419.

Kelly JT, Prasad AK, and Wexler AS (2000) Detailed flow patterns in the nasal cavity. *Journal of Applied Physiology* 89: 323–337.

Mozell MM (1970) Evidence for a chromatographic model of olfaction. *Journal of General Physiology* 56: 46–63.

Negus V (1958) *The Comparative Anatomy and Physiology of the Nose and Paranasal Sinuses.* London: E & S Livingstone.

Pashler H and Yantis S (eds.) (2002) *Stevens' Handbook of Experimental Psychology, Vol. 1: Sensation and Perception,* 3rd edn. New York: Wiley.

Rawson NE (2000) Human olfaction. In: Finger TF, Silver WL, and Restrepo D (eds.) *The Neurobiology of Taste and Smell,* 2nd edn., pp. 257–284. New York: Wiley-Liss.

Small DM, Gerber JC, Mak YE, and Hummel T (2005) Differential neural responses evoked by orthonasal versus retronasal odorant perception in humans. *Neuron* 47: 593–605.

Sun BC and Halpern BP (2005) Identification of air-phase retronasal and orthonasal odorant pairs. *Chemical Senses* 30: 1–14.

Wrobel BB and Leopold DA (2005) Olfactory and sensory attributes of the nose. *Otolaryngologic Clinics of North America* 38: 1163–1170.

Zhao K, Scherer PW, Hajiloo SA, and Dalton P (2004) Effect of anatomy on human nasal air flow and odorant transport patterns: Implications for olfaction. *Chemical Senses* 29: 365–379.

Relevant Websites

http://www.chemosensory.com – *ChemoSense* (2006) current and back issues.

http://155.37.7.13/dentalhtml – Halpern BP (2001) Retronasal/orthonasal olfaction. *Dental continuing education video, University of Connecticut Health Center.* Requires Real Player (version 8) and Internet Explorer.

Rett Syndrome

N Sharma, Massachusetts General Hospital, Charlestown, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Gene Expression

Genomic DNA contains the core genetic information of the cell; this information is contained in the sequence of the four DNA bases: adenine (A), cytosine (C), guanine (G), and thymine (T). During development, somatic cells display distinct patterns of gene expression that are stably inherited through cell division. Thus, in addition to the need for each somatic cell to contain the same primary DNA sequence as contained in every other somatic cell, there must be an additional layer of information that specifies which genes are to be used at which stage of development. The term 'epigenetics' was coined to describe variations in gene expression that are not due to changes in the DNA sequence.

Imprinting

Genomic imprinting is an epigenetic mechanism that regulates gene expression during mammalian development. Imprinting is a normal process by which certain genes are marked as coming from either the mother or the father. Based on the imprint, a subset of genes will be expressed or repressed, depending on parental origin, during development. It is thought that the differential modification of the maternal and paternal alleles is likely to affect the overall conformation of the chromatin, containing each allele, thus influencing interactions between the transcription machinery and the genomic template.

Some imprinted genes are expressed from a maternally inherited chromosome and are silenced on the paternal chromosome; while other imprinted genes are only expressed from a paternally inherited chromosome. Thus, imprints can act as a silencer or an activator for imprinted genes. Genomic imprinting plays a role in embryogenesis. Large numbers of imprinted genes identified in humans are involved in pre- and postnatal development. Multiple lines of evidence suggest that imprinted mammalian genes influence complex neuropsychological behavior and neuropathological disorders.

DNA methylation plays a critical role in imprinting. Differential methylation of specific DNA sequences, based on whether they are maternal or paternal in origin, is one method by which genes are identified and their expression is regulated. DNA methylation occurs in almost all living organisms. DNA methylation occurs primarily at the fifth position of cytosine (5mC) in the dinucleotide CpG (in which the 'p' represents a phosphate diester). CpG methylation can affect gene expression either directly or indirectly. By the direct route, CpG methylation alters the DNA sequence so that it is no longer recognized by transcription factors. By the indirect route, CpG methylation produces a ligand for chromatin-binding proteins such as MeCP2. These proteins bind to methylated CpG sequences and then target repressor complexes that contain histone deacetylases (HDACs), leading to the formation of a deacetylated and repressed chromatin structure that is refractory to the transcription machinery. Overall, DNA methylation results in gene silencing via the repression of transcription, by both direct and indirect routes.

The list of human diseases caused by or involving DNA methylation abnormalities has become quite long. Rett syndrome is an example of a disorder due to abnormalities in DNA methylation.

Rett Syndrome

Rett syndrome is a progressive disorder of brain development that is characterized by the loss of acquired motor and language skills, stereotypic movements, autistic features, and both sleep and respiratory abnormalities. Rett syndrome is caused by mutations in the X-linked gene (*MECP2*) encoding the methyl-CpG-binding protein 2 (MeCP2). The spectrum of *MECP2* mutations that cause Rett syndrome includes nonsense, missense, and frameshift mutations as well as intragenic deletions. Different mutations are associated with distinct clinical phenotypes. Rett syndrome is an autosomal dominant disorder with a penetrance of almost 100%.

Clinical Features

Females Classic Rett syndrome is seen in girls who are normal at birth and achieve expected developmental milestones until reaching 6–18 months of age. These affected females then develop deceleration of head growth, leading to microcephaly, with a loss of acquired language, cognitive, social, and motor skills. A distinctive feature is repetitive hand wringing that replaces purposeful hand movement. Additional stereotypies include clapping and mouthing.

In addition, there is development of an array of motor abnormalities, including gait apraxia, ataxia, spasticity, tremor, and dystonia. A temporary 'autistic-like' phase often occurs at the onset of the disorder,

Table 1 Phenotypes of Rett syndrome

Characteristic	*Classic Rett*	*Severe Rett*	*Atypical Rett*
Karyotype	46,XX 47,XXY 46,XY; somatic mosaic	46,XY	46,XY
Clinical features	Normal until 6–18 months Loss of milestones Microcephaly, hand wringing	Severe encephalopathy Death during infancy	Mild-to-moderate mental retardation

and older children are known for their social engagement through intense eye gaze. In addition, seizures and breathing irregularities, such as hyperventilation and apnea, develop. After an initial regression of attained milestones, the condition stabilizes and children typically survive until reaching an age between 15 and 20 years. The incidence of Rett syndrome is between 1/10 000 and 1/15 000 live births in females.

Males Prior to the identification of the *MECP2* gene and its mutations, it was thought that Rett syndrome in a male was lethal at the embryonic stage. However, it is now known that males are born with *MECP2* mutations and fall into one of three main phenotypic categories: (1) males with classic Rett syndrome, (2) males with severe encephalopathy and infantile death, and (3) males with less severe neurologic and/or psychiatric manifestations that include mental retardation.

Males with classic Rett syndrome have either a 47, XXY karyotype or are somatic mosaics with a 46,XY karyotype. Individuals who are somatic mosaics display varying levels of expression of a specific gene in different cells and tissues within the body. Thus, males who are somatic mosaics for a *MECP2* mutation may express the mutant MeCP2 protein in only specific regions of the brain, rather than throughout the brain. In other regions of the brain, and in other regions of the body, the mutant MeCP2 may be expressed at low levels, if at all. In theory, this could result in the phenotype of classic Rett syndrome, rather than a more severe phenotype. Males who fall into the phenotype of having classic Rett syndrome express the same phenotype as found in girls with classic Rett syndrome.

Males with severe Rett syndrome exhibit severe encephalopathy, die during infancy, and have a 46, XY karyotype. Males with either classic or severe Rett syndrome have the same MeCP2 mutations that cause classic Rett syndrome in girls. Those boys who have been identified with less severe neurological and/or psychiatric manifestations carry mutations in the *MECP2* gene that are not found in girls with any identifiable phenotypic abnormality of the central nervous system. Presumably, this is because any girl carrying such a mutation is a heterozygote (i.e., also carries a normal *MECP2* gene) and therefore the effect, if any, on phenotype is mild (see Table 1).

Neuropathology

Pathologic studies of the Rett syndrome brain were first described in 1986. The Rett patient brain is small for the age of the patient. There is no evidence of atrophy, degeneration, or demyelination or an inflammatory response. There is a significant poverty of dendrites in the frontal, motor, and inferior temporal lobes, with a simplified branching pattern in the dendrites that do exist. This finding is of significance because there is a temporal correlation between the age at which clinical regression is seen in classic Rett syndrome (6–18 months) and the time during which the projection and circuits of cortical layers III and V are organized (12–36 months). Taken together, these data support the hypothesis that Rett syndrome is a disorder of brain development.

Treatment

To date, there is no treatment for Rett syndrome. Prior to the discovery of the genetic defect that causes Rett syndrome and the development of animal models, only two compounds were subjected to well-designed clinical trials. Based on two reports indicating that L-carnitine may be helpful in Rett syndrome, a randomized, placebo-controlled, double-blind study was performed. Study participants were evaluated using a standardized scale that assessed both motor and behavioral skills. The trial revealed that there was no significant benefit to treatment with L-carnitine.

Naltrexone was also used in a clinical trial. The rationale for this study was that several studies reported increases in β-endorphin levels in the cerebrospinal fluid of girls with Rett syndrome. Thus, it was hypothesized that treatment with the opiate antagonist naltrexone may be beneficial. However, a controlled, randomized, double-blind clinical trial did not find any significant improvement in motor or behavioral skills with naltrexone therapy. It is

hoped that, with the availability of animal models of Rett syndrome, the evaluation of potential treatments for Rett syndrome will be enhanced and occur at a more rapid pace.

Methyl-CpG-Binding Protein 2

Mutations in the MeCP2 result in the neurodevelopmental disorder known as Rett syndrome. MeCP2 is a chromosome-binding protein that selectively binds 5-methylcytosine residues in symmetrically positioned CpG dinucleotides. These residues are preferentially located in the promoter regions of genes that are subject to transcriptional silencing after DNA methylation. MeCP2 contains two functional domains, an 85-amino-acid methyl-CpG-binding domain (MBD), essential for binding to 5-methylcytosine and a 104-amino-acid transcriptional repressor domain (TRD) that interacts with a variety of corepressor complexes, including histone deacetylase and the transcriptional corepressor Sin3A. Thus, a functional MeCP2 recognizes and binds to 5-methylcytosine residues in CpG islands. In turn, MeCP2 interacts with HDAC and Sin3A, resulting in a DNA–protein complex that renders the promoter region of the target gene inaccessible for the initiation of transcription. In this manner, MeCP2 is thought to play a critical role in gene repression.

Mutations in MeCP2 may result in the dysregulation of expression of hundreds or thousands of target genes. It is possible that mutations resulting in the loss of function of MeCP2 could result in overexpression of genes that may be detrimental to the development of the nervous system. One hypothesis regarding the pathophysiology of Rett syndrome is that the stability of brain function, or altered brain development, is due to the misregulation of genes in the absence of functional MeCP2 protein. However, the fact that Rett patients do not suffer severe abnormalities during fetal and early postnatal development implies that no specific programs of developmental gene expression are disrupted. It is conceivable that loss of function of this protein in some cells, especially differentiated and postmitotic neurons, would lead to the overexpression of some genes that may be detrimental during nervous system maturation.

Animal Models of Rett Syndrome

In order to understand the mechanism by which MeCP2 mutations result in Rett syndrome, three distinct mouse models have been generated. One model did not express MeCP2 in any tissue of the body. Another model did not express MeCP2 in neuronal tissue. In both models, developed in differential laboratories, the resultant mice exhibited a phenotype similar to that seen in children with classic Rett syndrome. The mice displayed normal development until approximately 6 weeks of age, followed by progressive neurological dysfunction and death by 10 weeks of age. A third mouse model, in which a truncated version of the MeCP2 protein was expressed, displayed a milder phenotype, with forepaw stereotypies, motor dysfunction, behavioral impairments, and learning deficits. These mouse models recapitulate relevant features of the human disease and will serve as useful tools in the study of the pathophysiology of Rett syndrome. Specifically, by studying the relationships between the MeCP2 protein and the genes affected by it, investigators hope to develop treatments specific for the abnormalities in gene expression that are responsible for Rett syndrome.

See also: Gene Expression Regulation: Chromatin Modification in the CNS; Transcriptional Silencing.

Further Reading

Amir RE, Van den Veyver IB, Wan M, et al. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature* 23: 185–188.

Armstrong DD (2002) Neuropathology of Rett syndrome. *Mental Retardation and Developmental Disabilities* 8: 726–776.

Bienvenu T and Chelly J (2006) Molecular genetics of Rett syndrome: When DNA methylation goes unrecognized. *Nature Reviews Genetics* 7: 415–426.

Bird A (2002) DNA Methylation patterns and epigenetic memory. *Genes & Development* 16: 6–21.

Chen RZ, Akbarian S, Tudor M, et al. (2001) Deficiency of methyl-CpG binding protein-2 in CNS neurons results in a Rett-like phenotype in mice. *Nature Genetics* 27: 327–331.

Guy J, Hendrich B, Holmes M, et al. (2001) A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. *Nature Genetics* 27: 322–326.

Jellinger K and Seitelberger F (1986) Neuropathology of Rett syndrome. *American Journal of Medical Genetics* 24: 259–288.

Meehan RR (2002) DNA methylation in animal development. *Seminars in Cell & Developmental Biology* 14: 53–65.

Moretti P and Zoghbi HY (2006) MeCP2 dysfunction in Rett syndrome and related disorders. *Current Opinion in Genetics & Development* 16: 276–281.

Nomura Y and Segawa M (2005) Natural history of Rett syndrome. *Journal of Child Neurology* 20: 764–768.

Percy AK (2002) Clinical trials and treatment prospects. *Mental Retardation and Developmental Disabilities* 8: 106–111.

Scarano MI, Strazzullo M, Matarazzo MR, et al. (2005) DNA methylation 40 years later: Its role in human health and disease. *Journal of Cellular Physiology* 204: 21–35.

Shahbazian M, Young J, Yuva-Paylor L, et al. (2002) Mice with truncated MeCP2 recapitulate many Rett syndrome features and display hyperacetylation of histone H3. *Neuron* 35: 243–254.

Reward and Learning

T A Stalnaker, Y Takahashi, and G Schoenbaum, University of Maryland School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

A reward can be defined most broadly as any condition or object that an organism seeks to attain or obtain access to. Thus, rewards for humans can be of tremendous variety, including food, water, money, and social approval. Rather than being identified by their sensory characteristics, rewards can only be inferred through the behavior that they elicit, and then only relative to the current motivational state of the organism. For example, water is considered a reward because it will elicit approach and consumption in an animal that has been deprived of water, but it will fail to do so for an animal that is not thirsty. Often, a distinction is drawn between primary rewards, such as water, naturally occurring rewards to which responses are assumed to be hard-wired, and secondary rewards, such as money, to which responses are learned through experience.

When an animal detects a reward in its environment, seeking and consummatory behaviors are by definition elicited. However, rewards also trigger learning. Broadly speaking, reward-associated learning seeks to identify the conditions in which rewards are likely to be experienced in order to promote the more accurate prediction and discovery of reward in the environment in the future. Behavioral studies have dissociated a number of different aspects of reward learning, supported by different brain circuits, that occur separately or in concert when an organism experiences a reward. The sections that follow briefly describe some of these different forms of reward-associated learning.

Reward and Pavlovian Learning

One of the simplest forms of learning about rewards is assessed in a procedure called Pavlovian conditioning, in which a neutral cue, or stimulus, is repeatedly presented with a primary reward. Learning is inferred from changes in the response of an animal to the presentation of the previously neutral stimulus, called the conditioned stimulus (CS). The resultant learned, or conditioned responses (CRs) include the following: (1) conditioned approach, in which the animal approaches the area where the reward has been delivered; (2) autoshaping, in which the animal approaches the location of the CS itself; (3) conditioned orienting, in which the animal orients to the CS; and (4) many conditioned autonomic changes, such as a slowing of heart rate. In each of these cases, the neutral CS gains control over a CR by virtue of its relationship to the reward.

The striatum, a part of the basal ganglia, is thought to be particularly critical to associative learning that directly links CSs with motor responses, including CRs such as conditioned approach. Such stimulus–response, or S–R, associations, are also critical to operant responding (see the section titled 'Reward and operant learning) The central nucleus of the amygdala (CeA) is required for the acquisition and/or expression of many of the other CRs previously described. In addition, the CeA is necessary for a CS to activate nonspecific motivational representations that were originally linked to the primary reward. These motivational representations can then elicit and/or enhance certain categories of relevant behavior. The effect of associations between the CS and these motivational representations may be demonstrated by training an animal to associate a CS with a reward and then training it to make an operant response to receive the same reward, or in some cases to receive a different reward. Subsequently, if the animal is allowed to perform the operant response in the presence of the reward-predictive CS, it increases its rate of operant responding. This phenomenon is called Pavlovian-instrumental transfer (PIT), and is thought to reflect a learned association between the stimulus and affect or feeling that accompanies receipt of the reward. Damage to the CeA disrupts PIT, and further experiments have supported the hypothesis that a circuit from the CeA to midbrain dopamine neurons and onward to the nucleus accumbens is essential to the expression of PIT.

Unlike the CeA, the neighboring basolateral nucleus of the amygdala (ABL) is not directly critical for the acquisition or expression of many appetitive CRs. However, the ABL is necessary for CSs to become associated with certain intermediary representations of the primary reward. These representations become critical in situations in which expression of a previously learned CR must be modified, for example, when the incentive value of a reward changes. In so-called devaluation paradigms, the value of a reward that has been previously associated with a CS is modified by pairing the food with illness or by satiating the animal with that particular food. Either of these manipulations will cause a normal animal to reduce its conditioned approach behavior to the CS that predicts the devalued reward. This reduced response indicates that the CS is able to evoke a representation of the current value of the food. This representation differs from the one that supports PIT, because it

contains information about the specific reward that triggered the original learning. Furthermore, it depends on different neural circuits than does PIT; animals with damage to the CeA retain the ability to form these associations, while animals with damage to the ABL fail to do so. Another region that is closely connected with the ABL, the orbitofrontal cortex (OFC), is also necessary for the control of CRs by the expected value of a CS. Thus, lesions of the OFC done either before or after acquisition of the CS–food association result in rats that do not reduce their responding to a CS associated with devalued food. These and other experiments have led to the conception of the OFC and ABL as forming a circuit that encodes an 'expectancy' that reflects a predicted value under given environmental conditions and goals of the animal. In the case of the devaluation paradigm, the expectancy encoded by the OFC allows the animal to use the CS to predict the devalued food, and hence to reduce responding to that particular CS.

Reward and Operant Learning

Learning to predict and prepare for rewards based on environmental cues takes an animal only so far toward actually receiving rewards. In a natural environment, an animal must often learn to make an action or series of actions in order to obtain a reward. This kind of learning is measured through an operant-learning paradigm, in which rewards are delivered contingent on the performance of a specific arbitrary action or series of actions. After repeated experience, animals will perform the trained actions even in the absence of reward delivery, demonstrating that learning has occurred. As in Pavlovian conditioning, several different kinds of learning can support operant behavior. In some cases, operant behavior may consist of goal-directed behavior, which is defined as that which depends on (1) knowledge of action–outcome contingencies, and (2) a representation of the outcome itself. Thus goal-directed behavior depends on learned associations between representations of actions and rewards. Animals typically exhibit goal-directed operant responding after limited, but not extended, training. Conversely, after overtraining, operant responding becomes less goal directed, and is often described as being habitual. Habitual responding is thought to depend on associations between antecedent stimuli and operant responses (S–R associations); receipt of the reward encourages formation of these associations, but reward is not explicitly represented in the associations themselves. Consequently, any changes to the incentive value of the outcome or in response–outcome contingencies have little effect, at least in the short term, on established habitual responding. Various aspects of operant responding are thought to depend on different parts of the basal ganglia. Recent studies have suggested that there may be two distinct functional systems within the adjacent regions in the dorsal striatum. The dorsolateral striatum may be critical for the learning of S–R associations, whereas the dorsomedial striatum may involve in the learning of action–outcome associations. For example, in interval schedules of instrumental conditioning, lesions of the dorsolateral striatum in rats became insensitive to the posttraining devaluation of the outcome. However, both pre- and posttraining lesions of the posterior dorsomedial striatum as well as local inactivation of this area reduced the sensitivity of rats' instrumental performance both to changes in the action–outcome contingency and to posttraining outcome devaluation.

The idea that learned operant behavior can be supported either by a goal-directed system or by a habitual system begs the question of how control is apportioned between these two systems during learning. In other words, what mechanism causes an extensively practiced action to become a stronger habit and to lose the characteristic of being goal directed? Although the answer to this question is not known in detail, one hypothesis supposes that parts of the medial prefrontal cortex (mPFC) control the balance between these processes. In support of this hypothesis, lesions of the infralimbic region of the mPFC prevent an action from losing its goal-directed nature, even when it has been extensively trained. Conversely, lesions of prelimbic region of the mPFC cause even an extensively trained action to remain goal directed. These results are in accord with general theories of PFC cognitive control that suggest that the PFC controls mappings between inputs and actions.

Dopamine and Reward Signaling

As previously described, it is clear that in both Pavlovian and operant paradigms a variety of associations among stimuli, rewards, and responses underlie learned changes in behavior after exposure to reward. While one of the contributions of neuroscience has been to tease apart brain circuits that may be specialized for encoding and/or using particular aspects of these reward-related associations, it is also important to realize that many of the processes described occur in parallel as organisms learn and respond to rewards in their environment. One of the

unifying factors, thus far not discussed, that may contribute to much reward-related learning across the circuits discussed is the effects of the neurotransmitter dopamine. Dopamine receptor stimulation has been shown to effect cellular processes that may contribute to learning, and dopaminergic neurons, which project to all brain regions associated with reward learning, fire during learning in patterns consistent with a role in reinforcement learning. Specifically, dopamine neuronal firing is proposed to signal when reward or reward-predictive situations differ from expectations, based on prior associative learning. Such a signal is often called a 'teaching' signal and could facilitate new learning across multiple brain regions about stimuli and actions that predict reward.

See also: Appetitive Systems: Amygdala and Striatum; Conditioning: Theories; Cortical Processing of the Reward Value of Food; Dopamine; Goal-Directed Behavior Theories; Prefrontal Contributions to Reward Encoding; Procedural Learning: Classical Conditioning; Representation of Reward; Reward Systems: Human; Reward Decision-Making; Reward Processing: Human Imaging.

Further Reading

Baxter MG and Murray EA (2002) The amygdala and reward. *Nature Reviews Neuroscience* 3: 563–573.

Daw ND, Niv Y, and Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8: 1704–1711.

Dickinson A (1985) Actions and habits: The development of behavioral autonomy. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 308: 67–78.

Everitt BJ, Cardinal RN, Parkinson JA, and Robbins TW (2003) Appetitive behavior: Impact of amygdala-dependent mechanisms of emotional learning. *Annals of the New York Academy of Science* 985: 233–250.

Killcross S and Coutureau E (2003) Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex* 13: 400–408.

Schoenbaum G, Chiba AA, and Gallagher M (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience* 1: 155–159.

Schoenbaum G and Roesch M (2005) Orbitofrontal cortex, associative learning, and expectancies. *Neuron* 47: 633–636.

Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36: 241–263.

Schultz W, Dayan P, and Montague PR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.

Reward Decision-Making

P Shizgal, Concordia University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A Northwestern crow (*Corvus caurinus*) swoops down to a beach from its perch atop a nearby cliff. Gliding a few feet above the intertidal zone, it scans attentively before alighting. The crow's beak probes the sand repeatedly and emerges holding a small whelk. After a brief pause, the bird drops the whelk and searches for another, eventually finding a larger one that meets its fancy. The crow then flies along the beach to a rocky area, ascends almost vertically, drops the mollusk, and dives to inspect the result. The spire has broken off, and the crow easily extracts the body, which it swallows in a single gulp. Returning to the water's edge, the crow again searches the sand, rejecting small whelks and eventually flying off with a large one. This time, the shell remains intact after the first drop and several subsequent ones. Rather than abandoning the obstinate prey, the crow persists until the shell has broken and then harvests its reward.

At multiple steps in this vignette, the crow is faced with consequential choices. Rather than setting off to search for new prey on its own, the bird may have the option of scrounging food from another crow that has procured a prey item too large to consume quickly, such as a fish. While gliding over the intertidal zone, the crow selects a propitious landing spot among multiple candidates. It seems particularly choosy in accepting or rejecting whelks it has dug out of the sand and very persistent in continuing to drop a whelk that has not yet broken open.

Experiments have revealed the adaptive quality of such choices. For example, the smaller whelks do not offer a net average energy gain; they have a low probability of breaking when dropped, and their energy content fails to offset the high cost of multiple ascents to the drop point. In contrast, the larger whelks are profitable because they break more readily and contain more food energy. The probability of breaking a whelk shell seems independent of the number of times it has already been dropped; if so, abandoning a whelk that has failed to open is inefficient, because after having paid the cost of procuring another, the crow, on average, would be no closer to mealtime.

The foraging behavior of the Northwestern crow highlights several variables critical to decisions about reward seeking: reward magnitude (as indexed by the size of the whelk), procurement and handling costs, and risk (a function of the probability that the whelk shell will survive a drop) all contribute to the choices made by the crow. Experimental paradigms have been developed to manipulate such variables under simplified, highly controlled, laboratory conditions. These paradigms have been used in conjunction with neurobiological methods, such as electrophysiological recording, neurochemical measurements, pharmacological treatments, brain stimulation, and lesions, to study the neural circuitry involved in adaptive decision making.

Many experiments that apply neurobiological methods to study decision making under laboratory conditions have been carried out using rodents working at operant tasks to obtain rewards. Responses, such as pressing a lever, serve as simplified analogs to procurement behaviors seen in the wild. Rewards such as food pellets are harvested after the operant requirement has been satisfied. Cost is manipulated by varying the amount of work (e.g., the number of lever presses) required to deliver a reward, the effort required to execute each response, or the time that must be invested. The work or time requirement for earning each reward may be either fixed or variable. Additional dimensions of reward that are commonly manipulated include amount (e.g., size of a food pellet or volume of liquid), intensity (e.g., concentration of sucrose), delay (time between meeting the work or time requirement and delivery of the reward), and kind (e.g., food vs. water).

Intensity, Rate, and Cost

Brain Stimulation Reward

An experimental model that has seen particularly heavy use in studies of the neural basis of decision making in laboratory rodents entails delivery of electrical stimulation via chronically implanted electrodes. Rats will work vigorously and insatiably for such stimulation when the electrode is appropriately positioned within the brain. The behavior performed to trigger the stimulation is called self-stimulation, and the effect that draws the rat back to the lever is called brain stimulation reward (BSR). Electrical stimulation of the medial forebrain bundle can compete with, summate with, or substitute for goal objects such as sucrose solutions, water, or food pellets. It has been proposed that a neural signal arising from the stimulation mimics those that encode the intensity of a natural reward.

Computing the Payoff from BSR

Gallistel and co-workers have carried out experiments employing rewarding brain stimulation to model the decision-making processes of foragers. They reasoned that to allocate its behavior adaptively, a successful forager must multiplicatively combine information about the intensity of the reward and the rate at which it is encountered. Imagine a bird foraging for berries that vary in sugar concentration and abundance. One bush offers sweet, ripe berries at a low density; while foraging there, the bird will encounter intense rewards at a relatively low rate. Another bush offers a higher density of less ripe berries. Under such circumstances, the product of intensity and rate should provide a reasonable estimate of the energy return from each bush. In the brain-stimulation experiments, reward intensity was controlled by varying the number of pulses in each stimulation train (and thereby the induced impulse flow in the stimulated neurons), and reward rate was controlled by varying the average delay to rearming of the lever after a reward had been harvested. The experimenters measured the preferences of the rats for stimulation trains that varied in intensity and rate. The rats behaved as if they had computed the payoff from the trains of stimulation pulses by multiplying the intensity and rate of reward.

Conover and Shizgal trained rats on a task that provides control over the opportunity cost of rewarding brain stimulation. In order to earn a reward, the rats had to be holding down a lever at the time that an unpredictable, unsignaled interval timed out; cost was controlled by varying the average duration of the interval. The longer the rat held down the lever, the less it could engage in alternate activities such as grooming, exploring, or resting. Thus, cost in this paradigm is roughly analogous to the average time required for the crow to harvest a whelk; during this time, the bird must forgo the benefits of alternate activities, such as nest building. The behavior of the self-stimulating rats is well described by a model that equates the payoff from the stimulation to the ratio of reward intensity and cost. This result is consistent with the finding of Gallistel's group in that opportunity cost is the inverse of the reward rate that a subject in the Gallistel experiments would achieve if it spent all of its time working at a particular lever.

The Role of Dopamine in BSR

In models of operant performance, such as the matching law, behavioral allocation is tied to the net payoffs from the competing options; the multiple components of payoff (such as the intensity and cost of reward) act in concert in determining choice (e.g., by scalar combination, as illustrated above). This makes functional sense, but it poses a challenge for experiments on the neurobiological underpinnings of decision making: if a given manipulation shifts the allocation of behavior to reward seeking, through which component(s) of payoff did the manipulation act? This problem is illustrated by the controversy over the role of dopaminergic neurotransmission in the evaluation, selection, and procurement of rewards.

The first experiments linking dopaminergic neurons to reward were carried out on self-stimulating rats, and it has been shown repeatedly that manipulation of dopaminergic neurotransmission profoundly alters the effectiveness of BSR. Reward effectiveness can be estimated from the strength of the stimulation required to sustain a given level of performance; if low-stimulation currents or frequencies suffice, then the reward produced per unit of stimulation must be high, and thus the stimulation is deemed very effective at inducing reward. The effectiveness of BSR is increased by drugs that boost dopaminergic signaling, such as those that block the dopamine transporter, and decreased by drugs that attenuate dopaminergic signaling, such as those that block postsynaptic receptors.

Either explicitly or implicitly, most of the reports detailing the modulation of BSR following perturbations in dopaminergic neurotransmission attribute the observed change in performance to alteration of the intensity of the induced reward signal. However, an alternative account is suggested by work conducted by John Salamone's group on the effects of dopamine-specific lesions on performance for natural rewards. They show that there is little change in performance when work requirements are low and a substantial change when requirements are large. The nature of the reward costs determines the effect of the dopaminergic modulation. Performance on tasks requiring intense physical effort, such as scaling a barrier or repeatedly pressing a lever at a high rate, is very sensitive to decreases in dopaminergic neurotransmission, whereas performance appears to be little affected on low-intensity tasks that entail few responses or merely waiting for a reward. Such results could be due to dopaminergic influence on the proclivity of the subject to invest effort in the pursuit of reward. Conversely, changes in reward effectiveness could have differential effects on performance for low- and high-cost rewards due to the nonlinear relationship between performance and reward strength. The degree to which changes in reward intensity and proclivity to pay high effort costs contribute to the effects of dopaminergic manipulations on performance for either rewarding brain stimulation or natural rewards has yet to be worked out definitively.

Different Roles for Phasic and Tonic Signaling by Dopaminergic Neurons?

Should dopaminergic neurotransmission prove to influence reward intensity, questions remain concerning the underlying mechanism. Two intertwined issues must be resolved: (1) Do dopaminergic neurons lie in series with the neural signals representing reward intensity, or do they modulate transmission in the reward-signaling pathway? (2) What are the relative roles of the phasic and tonic components of dopaminergic signaling? The neural signal that gives rise to BSR is short-lived: it builds up quickly, particularly when the stimulation is strong, and it decays quickly as well. One mode of dopaminergic signaling consists of brief bursts of action potentials lasting a few hundred milliseconds. Could this phasic activity underlie the phasic signal that gives rise to BSR? This seems unlikely. The phasic release of dopamine can be monitored *in vivo* by electrochemical means. In rats trained to self-administer brief trains of pulses via electrodes positioned near the cell bodies of midbrain dopamine neurons, phasic dopamine release in the ventral or dorsal striatum was rarely seen during self-stimulation; in the few cases in which it was observed at the start of a test session, the phasic dopamine signal declined quickly and was no longer detectable after a minute or so; in contrast, the rats continued to work vigorously for the stimulation. These results are difficult indeed to reconcile with the notion that the reward signal is encoded in the phasic release of striatal dopamine.

A second mode of dopaminergic signaling entails the steady release of the neurotransmitter over longer timescales. The existing data on the influence of dopaminergic manipulations on BSR are consistent with the notion that such sustained release gates transmission of the reward signal. In this view, the dopamine neurons do not compose part of the direct pathway that translates the neural activity induced by the electrode into the rewarding effect. Instead, these neurons regulate the strength of the reward signal by modulating transmission of the phasic signal induced by the electrode. In turn, dopamine tone is increased by the rewarding stimulation. Due to positive feedback between the rewarding effect and dopamine tone, performance is remarkably persistent when stimulation strength is high.

A distinction between a phasic reward signal and the phasic firing of dopaminergic neurons has also emerged from extensive research on dopaminergic responses to conditioned stimuli and natural goal objects. Schultz's group has shown that the firing of dopaminergic neurons resembles the reward-prediction error in reinforcement-learning models. The reward is a drop of juice, and the simian subjects are thirsty during testing. Delivery of the reward is preceded by an initially neutral stimulus. The initial presentation of the juice reward elicits a vigorous burst of firing from the dopamine neurons. However, as learning proceeds and the onset of the conditioned stimulus comes to predict the drop of juice that follows, these neurons no longer respond to delivery of the reward; if the now-expected reward is omitted, firing is suppressed below its baseline rate. Thus, the phasic activity of the dopamine neurons appears to encode the difference between the reward the monkey was expecting and the reward it receives. In this view, a nondopaminergic 'primary' reward signal is relayed to the dopamine cells and compared with a reward prediction to generate an error signal, which is encoded in the phasic firing of the dopamine neurons. This error signal serves to train the neural network that generates the reward prediction and to adjust action patterns to maximize the harvest of rewards. Perhaps the neural signal that gives rise to BSR corresponds to the primary reward signal in this model. The identity of the neurons that produce this signal has yet to be established, but there is accumulating evidence implicating the extended-amygdala network.

Risk

The introductory vignette illustrates a form of risk sensitivity in foraging behavior. Every time the crow makes the arduous ascent to the drop point, it runs a risk that its efforts will be for naught. Prey selection appears to take this risk into account, in the form of a preference for larger whelks that are more likely to break when dropped.

In laboratory rats, there is evidence that circuitry in a particular brain region plays a role in determining risk sensitivity. Lesions of the core region of the nucleus accumbens, a dopaminergic terminal field, shifted choices toward small certain rewards and away from larger probabilistic ones. The lesioned rats behaved disadvantageously, choosing rewards with lower expected values (e.g., 1 food pellet delivered with $p = 1$) in preference to those with higher expected values (e.g., 4 food pellets delivered with $p = 0.32$ or 0.7). Thus, the circuitry disrupted by the lesion, which is modulated by dopaminergic input, appears to contribute to the rats' ability to assign adaptive weights to low-probability rewards. In this light, is it interesting to note case studies in which pharmacological treatments that increase dopamine tone appear to boost the incidence of pathological gambling. These findings have been interpreted to suggest that sustained activity of dopaminergic

neurons influences risk appetite in humans. Patients with lesions of the ventromedial region of the prefrontal cortex, another dopaminergic terminal field, show increased risk tolerance.

Delay

The effectiveness of rewards depends not only on their probability but also on their imminence. Imagine a cheetah lurking in the tall grass. Scanning the savanna, it spies two similar gazelles trotting towards its hiding place from different locations. Clearly, the gazelle that will first enter the cheetah's striking range is the better prospect. Thus, the values of the two gazelles can be said to be discounted according to their projected arrival times, and the value of the nearer one is less diminished due to the smaller delay. However, if the more distant gazelle is larger than the nearer one, the cheetah is faced with a dilemma: pursue the smaller, sooner reward or wait for the larger, later one? If there is a constant probability over time that either gazelle will change direction or will be pursued by another cheetah before it comes into range, then the present value of each gazelle should decline exponentially as a function of its predicted time of arrival. The cheetah will forgo the small gazelle and wait for the larger one if the size difference outweighs the difference in the temporal discounts, regardless of whether the expected arrival is imminent or far off.

Experiments carried out on many different animals paint a more complex and intriguing picture of the discounting problem. The discount function deviates from a simple exponential and is approximated by a rectangular hyperbola. This widespread finding implies that imminent rewards are attributed even greater relative weight than a simple exponential discounting model predicts. Moreover, the preference of a decision maker who discounts hyperbolically can reverse as a function of its temporal vantage point. A cheetah behaving in this way would fix its sights on the larger, later reward if both predicted arrival times are relatively far off in the future; as the gazelles approached, its preference would shift to the smaller, sooner reward. The latter behavior is said to be 'impulsive.' The steeper the discount curve, the more dramatic the preference reversal, and the more shortsighted the behavior of the cheetah will appear. Thus, it is not surprising that neural mechanisms involved in temporal discounting have been implicated in impulse-control disorders, which are believed to entail excessive discounting of negative long-term consequences in the pursuit of modest short-term gains.

Cardinal and colleagues have shown that lesions of the core region of the nucleus accumbens not only steepen probability discounting but also steepen temporal discounting, shifting the preference of laboratory rats toward smaller, sooner rewards. In the inferotemporal cortex, another dopaminergic terminal field implicated in decision making, dopamine release increases during performance of a temporal discounting task but release of serotonin does not; in contrast, serotonin release is boosted in the medial prefrontal cortex during the discounting task whereas dopamine release is decreased. It has been argued that there is a division of labor between cortical territories, with the temporal and reward-magnitude dimensions of the task processed preferentially in the medial prefrontal cortex and the orbitofrontal cortex, respectively; neurochemically distinct diffuse projection systems originating in the brain stem appear to differentially modulate the operation of the distributed cortical circuitry.

Kind

Animals must procure diverse resources in order to ensure survival and successful reproduction. These resources are sorted into differentially fungible categories (kinds) by physical, chemical, physiological, and ecological realities. An excess of stored food cannot offset a deficit in hydration or compensate for a lack of nest-building materials. In contrast, pathways of metabolism provide for partial convertibility of nutrients; although all macronutrients can be used as sources of cellular energy, carbohydrates provide a superior means of meeting short-term needs whereas fats provide a superior input to long-term stores. Thus, animals face what economists call a substitution problem. Their task is not merely to procure the most valuable single resource from a set of possible options but rather to assemble, at the lowest cost and risk, the basket of goods that best meets a diverse set of needs. Due to the variable degree to which items can be substituted for each other to meet particular needs, the contribution of each item to the value of the basket cannot be computed merely by adding the individual values. The computational task is rendered even more challenging by the fact that a given goal object may contain resources of many different kinds. For example, a prey item typically contains macronutrients from all three classes, water, minerals, and vitamins.

How does the nervous system go about solving the substitution problem? It has been suggested that functional subdivision of brain reward circuitry may contribute to this important task. Perhaps dedicated circuitry processes the reward value of resources of a given kind. Combining the outputs of such specialized circuits on the basis of the substitutability of the rewards they evaluate could then help distribute

the choices of the decision maker so as to achieve an appropriate basket of goods. Evidence for functional specialization of brain reward circuitry is provided by studies of the effects of energy balance on BSR. When the tip of the stimulating electrode lies in a particular region of the lateral hypothalamus, dorsolateral to the fornix, depletion of the long-term energy stores boosts the rewarding effect; leptin, a hormone produced by adipocytes in proportion to their fat content, attenuates the rewarding effect of stimulation delivered to this region. In contrast, when nearby portions of the medial forebrain bundle are stimulated, the rewarding effects are unaltered by depletion of the fat stores and either unchanged or increased by administration of leptin. Such evidence suggests that physiological feedback signals differentially modulate the rewarding effects produced by stimulation of different brain sites.

The Origin of Reward Signals

The neural signals representing the intensity, amount, and kind of rewards interact with signals representing general information about rate, cost, risk, and delay. In addition to the evaluation of rewards, many fundamental psychological processes, such as interval timing and memory, are essential to such computations. Thus, it is not surprising that a wide array of brain structures has been implicated in reward-related decision making, including multiple territories in the prefrontal cortex, the striatum and ventral pallidum, the amygdaloid complex and extended amygdala, the hippocampal formation, the medial forebrain bundle, and the midbrain nuclei that give rise to ascending dopaminergic, noradrenergic, serotonergic, and cholinergic projections. Using the gustatory system as an example, the following section addresses the origin of reward signals and their processing prior to combination with other information required for decision making.

Taste plays a crucial and obvious role in determining the rewarding properties of food items, and a considerable body of research has been devoted to the origin of gustatory reward signals. A bifurcation of information flow is implicit in any sensory pathway that gives rise to reward signals. One function of sensation is to determine the identity and quality of a stimulus; another is to indicate what that stimulus is worth at the present moment. In contrast to the circuitry dedicated to determining what the stimulus is, the circuitry charged with computing stimulus value functions in an inherently subjective manner. No taste stimulus is inherently good: a food that seems irresistible when we are hungry may become disgusting after we have overindulged. Thus, the computation of gustatory reward by the evaluative channel implies a weighting of sensory information by internal signals reflecting the current physiological state, a process that Cabanac has labeled 'alliesthesia.' Further modulation of stimulus value occurs as a result of learning, short-term exposure, and modulation by hormonal, circadian, and photoperiodic factors.

In rodents, there is a bifurcation in the gustatory projection. Gustatory information from the nucleus of the solitary tract is relayed to the parabrachial region, which, in turn, projects in parallel to the gustatory thalamus and to basal forebrain targets in the lateral hypothalamus and extended amygdala. Norgren and Grill proposed that the informational function is subserved by the thalamocortical branch, whereas the evaluative function depends on the basal-forebrain branch. Consistent with this view is the finding that sucrose, a rewarding gustatory stimulus, fails to increase dopamine release in the nucleus accumbens following interruption of the basal-forebrain projection but continues to do so following interruption of the thalamocortical branch.

The parabrachial relay for corticopetal gustatory information appears to be absent in the primate, and there is abundant evidence for the processing of gustatory rewards by cortical circuitry. At early stages of processing in the primate gustatory system, up to and including the primary gustatory cortex, food deprivation does not appear to modulate the responses of taste-sensitive neurons. However, such modulation is seen in the secondary gustatory cortex, within the orbitofrontal region as well as in the lateral hypothalamus. Rolls has proposed that neurons in the latter regions are involved in evaluative processing, whereas the early stages of the pathway and the primary gustatory cortex are involved in informational processing. Within the orbitofrontal cortex, his group has demonstrated separate representations of multiple sensory characteristics of foods, such as sweetness, viscosity, and graininess, as well as neurons tuned to combinations of several different sensory qualities. Also represented there are rewards processed in different sensory modalities, such as somesthesis. The elaborate representations within the orbitofrontal cortex seem well suited to tackle the substitution problem by means of a weighted combination of signals representing the values of different goal objects and their components.

Conclusion

Animals searching for sustenance, shelter, or mates, must solve multidimensional decision problems on an ongoing basis, trading off the benefits and costs of rewards against those of territorial defense and predator avoidance. Neuroscientific research into the mechanisms underlying such decisions is at an

early stage. Nonetheless, components of the computational processes underlying reward-related decisions have been brought under experimental control and linked to neurochemically and anatomically characterized circuitry in the brain.

See also: Connectivity of Primate Reward Centers; Dopamine; Dopamine in Perspective; Goal-Directed Behavior Theories; Neuroethological Perspective; Neuropsychology of Primate Reward Processes; Prefrontal Contributions to Reward Encoding; Representation of Reward; Reward Systems: Human; Reward Neurophysiology and Primate Cerebral Cortex; Reward Neurophysiology and Orbitofrontal Cortex; Reward Processing: Human Imaging; Transcription and Reward Systems.

Further Reading

Cabanac M (1971) Physiological role of pleasure. *Science* 173: 1103–1107.

Cardinal RN and Howes NJ (2005) Effects of lesions of the nucleus accumbens core on choice between small certain rewards and large uncertain rewards in rats. *BMC Neuroscience* 6(1): 37.

Cardinal RN, Winstanley CA, Robbins TW, et al. (2004) Limbic corticostriatal systems and delayed reinforcement. *Annals of the New York Academy of Science* 1021: 33–50.

Conover KL and Shizgal P (2005) Employing labor supply theory to measure the reward value of electrical brain stimulation. *Games and Economic Behavior* 52(2): 283–304.

Gallistel CR (1994) Foraging for brain stimulation: Toward a neurobiology of computation. *Cognition* 50(1–3): 151–170.

Hajnal A and Norgren R (2005) Taste pathways that mediate accumbens dopamine release by sapid sucrose. *Physiology & Behavior* 84(3): 363–369.

Hernandez G, Hamdani S, Rajabi H, et al. (2006) Prolonged rewarding stimulation of the rat medial forebrain bundle: Neurochemical and behavioral consequences. *Behavioral Neuroscience* 120(4): 888–904.

Rolls ET (2005) Taste, olfactory, and food texture processing in the brain, and the control of food intake. *Physiology & Behavior* 85(1): 45–56.

Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36(2): 241–263.

Shizgal P, Fulton S, and Woodside B (2001) Brain reward circuitry and the regulation of energy balance. *International Journal of Obesity* 25(supplement 5): S17–S21.

Waraczynski MA (2006) The central extended amygdala network as a proposed circuit underlying reward valuation. *Neuroscience & Biobehavioral Reviews* 30(4): 472–496.

Wightman RM and Robinson DL (2002) Transient changes in mesolimbic dopamine and their association with 'reward.' *Journal of Neurochemistry* 82(4): 721–735.

Winstanley CA, Theobald DE, Dalley JW, et al. (2006) Double dissociation between serotonergic and dopaminergic modulation of medial prefrontal and orbitofrontal cortex during a test of impulsive choice. *Cerebral Cortex* 16(1): 106–114.

Wise RA (1996) Addictive drugs and brain stimulation reward. *Annual Review of Neuroscience* 19: 319–340.

Zach R (1979) Shell dropping: Decision-making and optimal foraging in Northwestern crows. *Behavior* 68: 106–117.

Reward Neurophysiology and Orbitofrontal Cortex

E T Rolls, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The (primate and human) orbitofrontal cortex receives taste, olfactory, visual, somatosensory, and auditory inputs from cortical regions which are toward the end of the unimodal cortical processing streams that build representations of what stimulus is present in each of these sensory modalities (see Figure 1). Neurophysiological studies in primates show that it is the reward value of taste, olfactory, oral texture, and visual stimuli that is represented, as shown, for example, by the reduction to zero of the neuronal responses after feeding to satiety on a particular food. The reward is partly specific to a particular food, as shown by sensory-specific satiety-related neuronal responses. Visual and olfactory stimuli become rewarding by learned association with a primary (unlearned) reinforcer such as the taste of food. This learning can be reversed (in as little as one trial for visual stimuli) and may be facilitated by error neurons that respond when an expected reward is not obtained. Complementary neuroimaging studies in humans show that abstract reinforcers such as monetary reward are also represented in the orbitofrontal cortex, and that the expected value of a reward may be represented in the orbitofrontal cortex.

Taste Reward Neurons

Many neurons in the orbitofrontal cortex respond to gustatory stimuli. Each neuron responds typically with distributed encoding to a combination of sweet, salt, bitter, sour, and umami tastants. It is the reward value that is represented, for feeding the macaque to satiety decreases the responses of the neuron to zero to the food fed to satiety. The decrease is specific to the taste that has been fed, indicating that sensory-specific satiety is being represented (see Figure 2). These effects of satiety are not found one synapse earlier, in the primary taste cortex, so that the representation of taste reward is implemented in, and is an important function of, the orbitofrontal cortex. Given that in primates the brain stem taste relay, the nucleus of the solitary tract, projects to the taste thalamus and then to the primary taste cortex without direct subcortical connections, the orbitofrontal cortex is in an important position in connection with the computation of taste reward.

Taste neurons are found in many parts of the primate orbitofrontal cortex, from laterally almost in the inferior convexity, to within a few millimeters of the midline. All five prototypical tastes are represented, so taste as a punisher, as well as taste as a reward, is represented.

Olfactory Reward Neurons

Many orbitofrontal cortex neurons respond to odors, and distributed representations are typical. For approximately 40% of the olfactory neurons, the olfactory stimuli to which the neurons respond are determined by the taste with which the odor is associated. These neurons can reverse the association, with relatively slow learning (e.g., 40 trials) compared with visual to taste association learning. The learned representation of a flavor, a taste–olfactory combination, is thus relatively stable. Some of these neurons respond to odors associated with the taste and texture of food and thus learn to represent the reward value of the odor. Other orbitofrontal cortex neurons respond to odors associated with aversive tastes such as salt (NaCl) and thus learn to represent the punishment value of the odor.

The olfactory neurons with food-related responses decrease their responses to zero when the macaque is fed to satiety with a food, and this is asensory-specific effect. Thus olfactory reward and sensory-specific satiety are represented in the orbitofrontal cortex. They are probably computed there, for a functional imaging study has shown that the subjective pleasantness of odors is represented in the human orbitofrontal cortex but not in the primary olfactory cortex. The representation of odor in the orbitofrontal cortex is distributed, that is, each neuron responds to a subset of the odors being used, and this is a useful part of the mechanism of the sensory-specific reduction in the pleasantness and reward value of a particular food eaten to satiety, for this reduction can be computed by making the synaptic connections from the neurons that convey the olfactory inputs related to the food being eaten adapt over the time course of the meal.

Experiments with sensory-specific satiety have shown that the reward value or pleasantness can be decreased in part by exposure to the odor for the same time as a food would be eaten in a meal. However, for the reward value to decrease to zero, the food must be swallowed, and a crucial signal is gastric distension. This is shown by experiments in which feeding resumes almost immediately after the stomach is drained. This evidence shows that part of the computation of reward

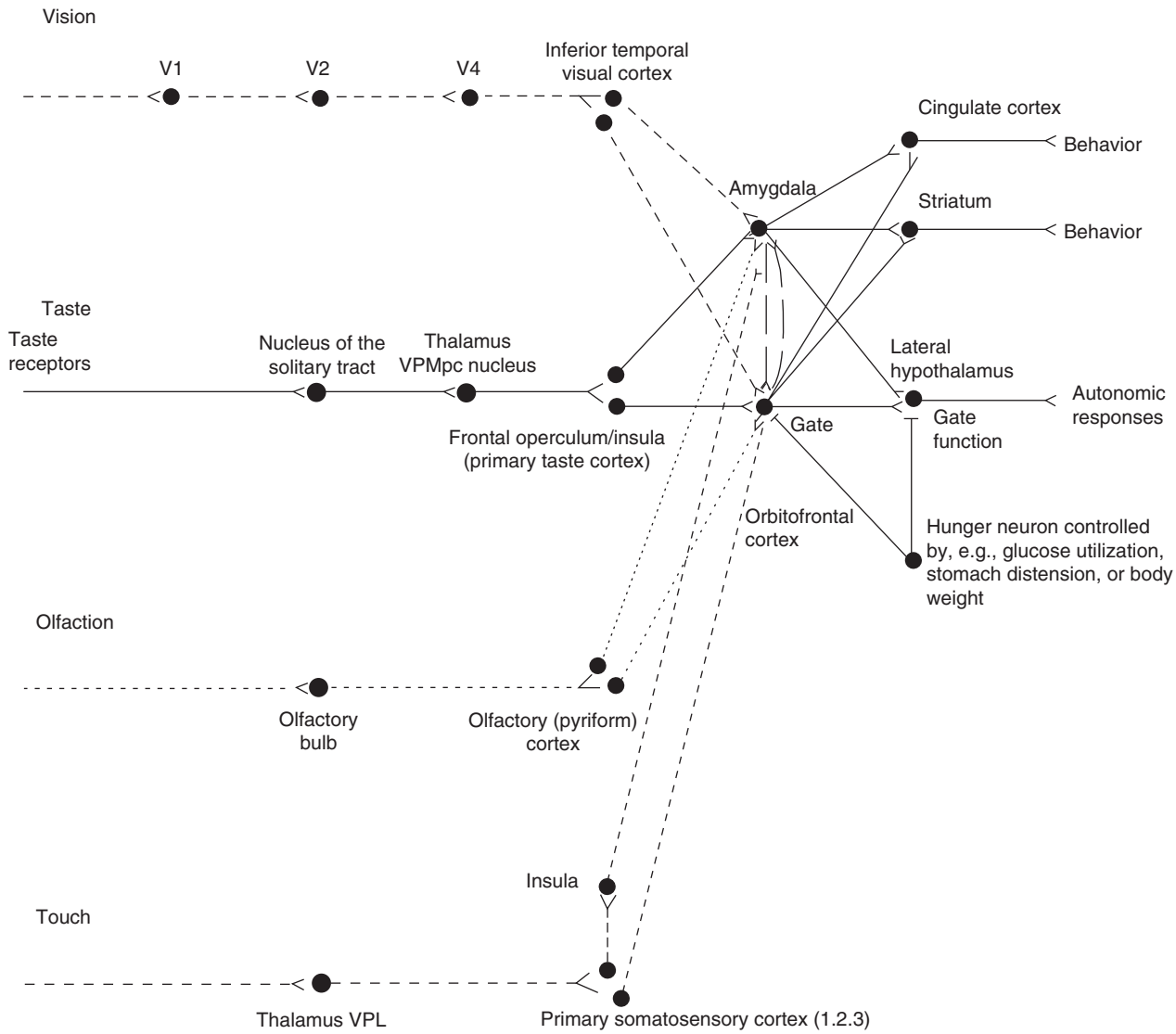

Figure 1 Schematic diagram of the connections of the orbitofrontal cortex in primates, showing how taste, olfactory, somatosensory, and visual pathways converge in the orbitofrontal cortex. The gate functions shown refer to the finding that the responses of taste, olfactory, visual, and oral somatosensory neurons in the orbitofrontal cortex and the lateral hypothalamus are modulated by hunger and thus represent reward value. VPL, ventral posterolateral; VPMpc, ventral posteromedial thalamic nucleus; V1, V2, V4, visual cortical areas.

value in the primate orbitofrontal cortex must involve feedback satiety signals from the gastrointestinal tract (and other satiety signals too), and these must reach the orbitofrontal cortex to produce this effect.

Other neurons in the orbitofrontal cortex respond to odors, but their response does not depend on the taste association of the odor. These other neurons may represent the reward and punishment value of nonfood-related olfactory stimuli, and consistent with this, the subjective pleasantness of olfactory stimuli is correlated with activations in the human medial orbitofrontal cortex, and the subjective unpleasantness of odors is correlated with activations in more-lateral parts of the human orbitofrontal cortex.

Visual Reward Neurons

Some neurons in the primate orbitofrontal cortex respond to visual stimuli, and in a visual discrimination task, some of these learn in as little as one trial to respond to the reward or punishment value of the taste associated with the visual stimuli. Some of these neurons even reverse by a rule-based rather than purely associative mechanism, in that in a well-learned go, no-go visual discrimination task, after the previously rewarded stimulus becomes associated with an aversive taste, the neuron may respond to the other stimulus as rewarding, even though it has not recently been paired with reward. A mechanism

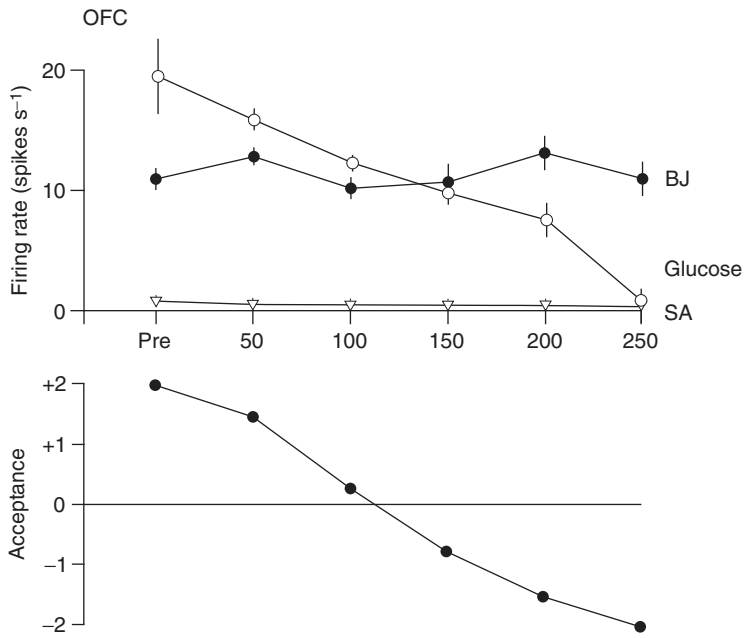

Figure 2 The effect of feeding to satiety with glucose solution on the responses of a neuron in the secondary taste cortex to the taste of glucose and of blackcurrant juice (BJ). The spontaneous firing rate is also indicated (SA). Below the neuronal response data for each experiment, the behavioral measure of the acceptance or rejection of the solution on a scale from +2 to −2 (see text) is shown. The solution used to feed to satiety was 20% glucose. The monkey was fed 50 ml of the solution at each stage of the experiment, as indicated along the abscissa, until he was satiated, as shown by whether he accepted or rejected the solution. OFC, orbitofrontal cortex; Pre, the firing rate of the neuron before the satiety experiment started. The values shown are the mean firing rate and its standard error. Adapted from Rolls ET, Sienkiewicz ZJ, and Yaxley S (1989) Hunger modulates the responses to gustatory stimuli of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. *European Journal of Neuroscience* 1: 53–60.

by which this is implemented involving quenching of a rule-based attractor network by nonreward, and the use of this network to bias the competition between neurons that respond to particular combinations of visual stimuli and whether or not they are currently associated with reward, has been proposed (see Figure 3). The quenching process on the rule attractor neurons is produced by the extra input to the orbitofrontal cortex from error neurons acting via inhibitory feedback neurons to decrease the activity in the rule attractor module. Because of some synaptic or neuronal adaptation in the rule attractor neurons, when the neuronal firing activity emerges out of the spontaneous firing rate, the neurons in the attractor that has not recently been active are those that win, and the result is that the opposite attractor state, representing the alternative rule, is expressed.

These visual neurons reflect the reward value of the stimuli, in that their responses decrease to zero to a food with which satiation has been produced. In this sense, the relative reward value of visual stimuli is represented.

Some primate orbitofrontal cortex neurons respond to face expression and to vocalization, and these are potentially reinforcers. Consistent with this, orbitofrontal cortex damage in humans can result in face and/or voice expression deficits in the absence of deficits in face or sound identification. Further, activation of the human orbitofrontal cortex can be influenced by the attractiveness of a face.

Somatosensory Reward Neurons

Some orbitofrontal cortex neurons are activated by oral texture (e.g., with viscosity tuning, or with fat responsiveness), temperature, and/or capsaicin. At least some of these neurons respond to the reward value of texture, in that the responses to fat can decrease to zero when a fatty food is eaten to satiety. Consistent with a somatosensory representation in the orbitofrontal cortex that reflects the reward value or aversiveness of somatosensory stimuli, activation of the orbitofrontal cortex in humans can be produced by pleasant touch to the hand or by pain (relative to neutral stimuli).

Auditory Reward Neurons

Some primate orbitofrontal cortex neurons respond to vocalization, different types of which are potentially

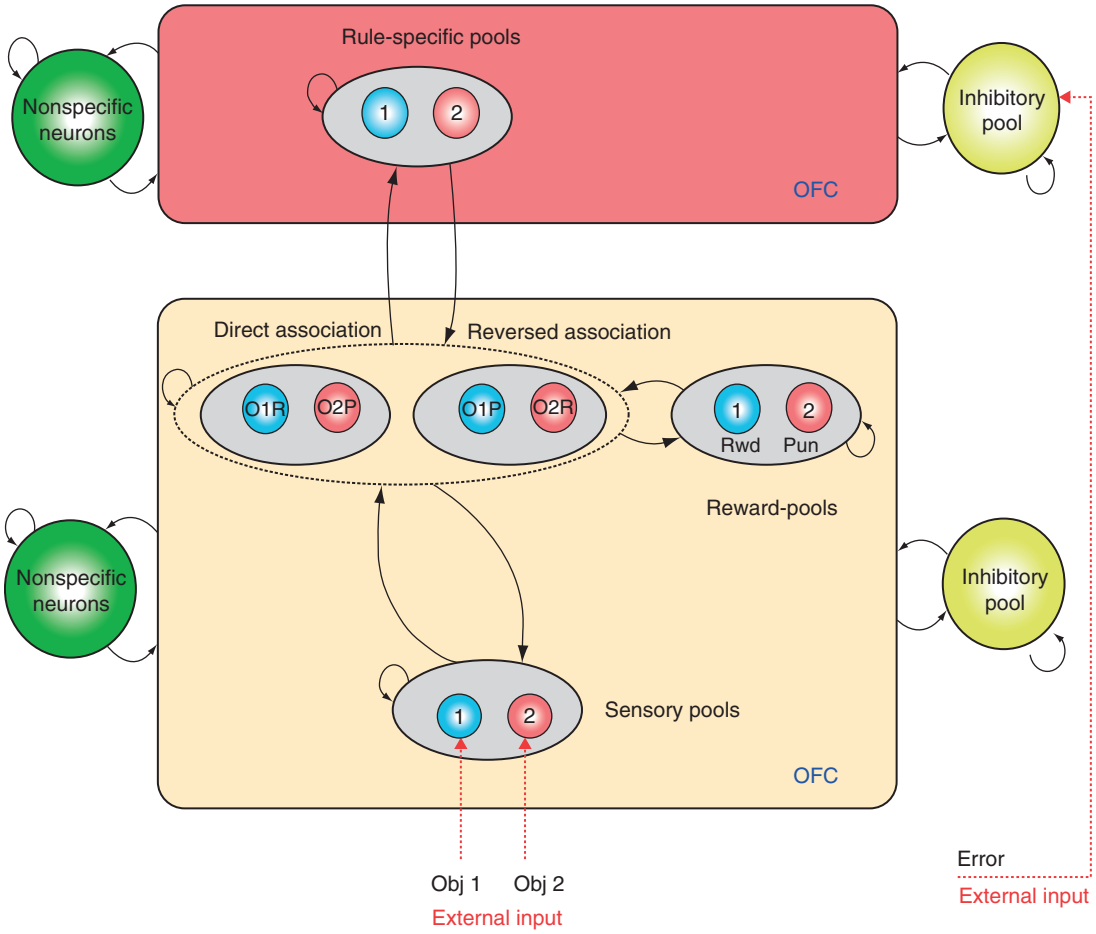

Figure 3 Cortical architecture of a reward reversal model for implementing rapid reversal of visual–taste associations. There is a rule module (top) and a sensory–intermediate neuron–reward module (below). Neurons within each module are fully connected and form attractor states. The sensory–intermediate neuron–reward module consists of three hierarchically organized levels of attractor network, with stronger synaptic connections in the forward than in the back projection direction. The intermediate level of the sensory–intermediate neuron–reward module contains neurons that respond to combinations of an object and its association with reward or punishment, such as object 1–reward (O1R, in the direct-association set of pools), and object 1–punishment (O1P in the reversed association set of pools). The rule module acts as a biasing input to bias the competition between the object–reward combination neurons at the intermediate level of the sensory–intermediate neuron–reward module. The synaptic current flows into the cells are mediated by four different families of receptors. The recurrent excitatory postsynaptic currents are given by two different types of excitatory postsynaptic potentials, respectively mediated by α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and *N*-methyl-D-aspartate receptors. These two glutamatergic excitatory synapses are on the pyramidal cells and interneurons. The external background is mediated by AMPA synapses on pyramidal cells and interneurons. Each neuron receives excitatory AMPA synaptic connections from outside the network. The visual input is also introduced by AMPA synapses on specific pyramidal cells. Inhibitory γ-aminobutyricacidergic synapses on pyramidal cells and interneurons yield corresponding inhibitory postsynaptic potentials. OFC, orbitofrontal cortex. Adapted from Deco G and Rolls ET (2005) Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. *Cerebral Cortex* 15: 15–30.

reinforcers. Consistent with this, in humans, activation of the orbitofrontal cortex can be influenced by consonant versus dissonant auditory stimuli.

Multimodal Reward Neurons

As indicated above, many neurons in the orbitofrontal cortex respond to combinations of the outputs of the unimodal cortical processing streams that generate representations of 'what' stimulus is present. For example, some orbitofrontal cortex neurons respond to the taste, odor, sight, and texture of food. These neurons reflect the reward value of the food in all these modalities, for feeding to satiety reduces the response of these neurons to each of their multimodal inputs.

Nonreward (Error) Neurons

Researchers have shown that a small proportion of neurons in the primate orbitofrontal cortex respond for many seconds when an expected reward is

not obtained. Different neurons respond to different types of nonreward, such as in extinction (when an expected reward is not obtained) or in visual discrimination reversal. These error neurons may be involved in the process by which neurons are reset to reflect new reward contingencies and, more generally, in correcting behavior. Consistent with this, activation of the human orbitofrontal cortex is produced when a face expression (e.g., anger) indicates that behavior should change in a simple model of social behavior.

Abstract Reward

More-abstract rewards are also represented in the orbitofrontal cortex. For example, it has been shown that in a probabilistic monetary reward task, the activation in the medial orbitofrontal cortex is linearly related to how much money is won on an individual trial and, in the lateral orbitofrontal cortex, to how much money is lost on each trial. Moreover, performance of the same task is impaired in humans with surgical lesions of the orbitofrontal cortex, particularly when the contingencies reverse.

However, the question arises of how the value of a stimulus is represented if the reward magnitude is high but the probability of obtaining the reward is small. Approaches used in reinforcement learning can be adopted, and we can use the terms 'reward value' (RV) for the magnitude of the reward obtained on a trial, that is, the reward outcome, and the 'expected value' (EV) as the probability of obtaining the reward multiplied by the RV. In an approach related to microeconomics, expected utility theory has provided a useful estimate of the desirability of actions and indicates that, except for very high and very low probabilities and for decisions with framing issues, expected utility, indicated by choices, approximately tracks EV. A recent investigation found that in the medial orbitofrontal cortex, the activation related to the EV period paralleled the activation to the RV. An implication is that the orbitofrontal cortex activations tracked the best estimate of RV throughout each trial. This makes the medial orbitofrontal cortex a candidate brain region where the neurons that represent RV have their firing modulated by the probability that the RV will be obtained to represent EV.

Cognitive Effects on Orbitofrontal Cortex Reward Activations

Cognition can reach down into the orbitofrontal cortex to influence affect, as shown by the effects that word descriptors (such as 'cheese' vs. 'body odor') have on the activations produced by olfactory stimuli in the medial orbitofrontal cortex. These cognitive modulations are of the pleasantness or RV of the olfactory stimuli. This modulation is suggested to be implemented by a top-down-biased competition mechanism of the type believed to be involved in attention.

Functions of the Reward Representations in the Orbitofrontal Cortex

Emotions can be defined as states elicited by rewards and punishers or, in other words, by instrumental reinforcers. Part of the importance of the orbitofrontal cortex in emotion is that it decodes and represents primary reinforcers. In other words, the orbitofrontal cortex in primates including humans is the first stage in processing of very many stimuli in which the reward or punishment value is made explicit in the representation, that is, in which the neuronal firing rate indicates the reward or punishment value of stimuli. The RV does not appear to be represented in areas that project to the orbitofrontal cortex in primates, including the primary taste cortex, the primary olfactory cortex, and the inferior temporal visual cortex.

Another part of the importance of the orbitofrontal cortex is that it learns associations between previously neutral visual stimuli, for example, visual and many olfactory stimuli, and primary reinforcers such as taste. This learning can be very fast, may be rule-based, and may be faster than that of the amygdala.

The representation in the orbitofrontal cortex of rewards and punishers is of the value of stimuli, and neuronal activity in it does not code for the behavioral responses being made. Consistent with this, it has little response-related activity. The orbitofrontal cortex is thus involved in representing the RV of stimuli and in stimulus–reinforcer association learning, in which the reinforcer is a stimulus (such as taste or touch), and the learning is thus of stimulus–stimulus associations.

Outputs of the orbitofrontal cortex reach the ventral striatum, ventral part of the head of the caudate nucleus, and the hypothalamus. It is via routes such as these that the orbitofrontal cortex may be able to influence the activity of dopamine neurons, as indicated in Figure 1.

Consistent with these concepts, orbitofrontal cortex lesions in humans have major effects on emotional behavior.

The neurophysiology of rewards and punishers in the orbitofrontal cortex described here thus gives an account of how the precise details of many different

types of reward and punisher can be represented in the primate orbitofrontal cortex and thus how the orbitofrontal cortex contributes to emotion.

See also: Connectivity of Primate Reward Centers; Cortical Processing of the Reward Value of Food; Flavor Physiology; Neuropsychology of Primate Reward Processes; Orbitofrontal Cortex: Visual Functions; Prefrontal Cortex: Structure and Anatomy; Prefrontal Contributions to Reward Encoding; Reward and Learning; Reward Decision-Making; Reward Neurophysiology and Primate Cerebral Cortex; Reward Processing: Human Imaging.

Further Reading

Deco G and Rolls ET (2005) Attention, short-term memory, and action selection: A unifying theory. *Progress in Neurobiology* 76: 236–256.

Deco G and Rolls ET (2005) Synaptic and spiking dynamics underlying reward reversal in orbitofrontal cortex. *Cerebral Cortex* 15: 15–30.

Glimcher PW (2003) The neurobiology of visual-saccadic decision making. *Annual Review of Neuroscience* 26: 133–179.

Glimcher P (2004) *Decisions, Uncertainty, and the Brain*. Cambridge, MA: MIT Press.

Hornak J, O'Doherty J, Bramham J, et al. (2004) Reward-related reversal learning after surgical excisions in orbitofrontal and dorsolateral prefrontal cortex in humans. *Journal of Cognitive Neuroscience* 16: 463–478.

Kahneman D and Tversky A (1984) Choices, values, and frames. *American Psychologist* 4: 341–350.

Kringelbach ML and Rolls ET (2003) Neural correlates of rapid reversal learning in a simple model of human social interaction. *NeuroImage* 20: 1371–1383.

Kringelbach ML and Rolls ET (2004) The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology* 72: 341–372.

O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, and Andrews C (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience* 4: 95–102.

Rolls ET (2004) The functions of the orbitofrontal cortex. *Brain and Cognition* 55: 11–29.

Rolls ET (2005) *Emotion Explained*. Oxford, UK: Oxford University Press.

Rolls ET (2006) The neurophysiology of the orbitofrontal cortex. In: Zald DH and Rauch SL (eds.) *The Orbitofrontal Cortex*, pp. 95–124. Oxford, UK: Oxford University Press.

Rolls ET and Deco G (2002) *Computational Neuroscience of Vision*. Oxford, UK: Oxford University Press.

Rolls ET, Critchley HD, Browning AS, and Inoue K (2006) Face-selective and auditory neurons in the primate orbitofrontal cortex. *Experimental Brain Research* 170: 74–87.

Rolls ET, Critchley HD, Mason R, and Wakeman EA (1996) Orbitofrontal cortex neurons: Role in olfactory and visual association learning. *Journal of Neurophysiology* 75: 1970–1981.

Rolls ET, Kringelbach ML, and de Araujo IET (2003) Different representations of pleasant and unpleasant odors in the human brain. *European Journal of Neuroscience* 18: 695–703.

Rolls ET, Sienkiewicz ZJ, and Yaxley S (1989) Hunger modulates the responses to gustatory stimuli of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. *European Journal of Neuroscience* 1: 53–60.

Schultz W, Tremblay L, and Hollerman JR (2000) Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex* 10: 272–284.

Sutton RS and Barto AG (1998) *Reinforcement Learning*. Cambridge, MA: MIT Press.

Thorpe SJ, Rolls ET, and Maddison S (1983) Neuronal activity in the orbitofrontal cortex of the behaving monkey. *Experimental Brain Research* 49: 93–115.

Reward Neurophysiology and Primate Cerebral Cortex

S Kobayashi, Cambridge University, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The primate cerebral cortex is functionally and anatomically organized such that each cortical domain processes specific information; sensory information is processed in the sensory areas that receive specific sensory input and motor information is processed in the motor areas that give output to the skeletal muscle system.

The reward system, which represents appetitive value such as pleasure and happiness, has no dedicated receptor. Reward signals have to be generated inside the brain based on sensory information and associated memories. In the example of Pavlovian conditioning, an initially neutral bell sound acquires appetitive value because it is consistently followed by food, which makes a dog salivate in response to the sound. To attach reward value to the bell sound, the neural system has to make the stimulus–reward association.

The reward system has no specific effector organ. The effects of reward may become observable as an autonomic response, such as salivation, or as voluntary actions. For instance, if animals can obtain food by pressing a lever, the neural system has to learn the action–reward association.

Reward value is not always simple to determine because many factors may influence the value. For instance, two apples are more valuable than one (magnitude), one apple for sure might be more valuable than two uncertain apples (risk), and one apple for free might be more valuable than two that require some effort (cost–benefit analysis). To make these economical considerations, the reward system might need help from some cognitive brain areas.

Mesencephalic dopamine neurons play a central role in the reward system. Reward signals carried by dopamine neurons are thought to be generated through the cortico-basal ganglia network. The generated reward signals might also influence activities in the cortex and basal ganglia via diffuse dopaminergic projections to these structures. The integrated nature of the reward system might correspond to the fact that reward-related activity is observed in surprisingly many areas of the brain. When looked at in detail, however, different brain areas process different types of reward information. This article summarizes recent findings on reward processes in the primate cerebral cortex obtained by using single-neuron recording methods with behaving monkeys.

Orbitofrontal Cortex

The orbitofrontal cortex (OFC) is a structure in which multiple sensory information and reward information converge. Anatomically, both the primary olfactory (pyriform) cortex and primary gustatory cortex send direct projections to the OFC. The OFC also receives visual, somatosensory, and auditory inputs through the uncinate fascicles and/or through the subcortical structures. The OFC is also densely connected with limbic/paralimbic structures. Dopamine neurons directly project to the OFC mostly from the ventral tegmental area and the retrorubral area. This anatomical background makes the OFC an ideal place for stimulus–reward association, which enables animals to predict reward outcome based on knowledge of the potential resources of the environment.

Monkey lesion studies provide evidence that the OFC is involved in processing stimulus–reward association. Monkeys with lesions in the OFC show deficits in object-reversal tasks, such that monkeys respond to the object which was formerly, but no longer, associated with reward. After bilateral OFC ablations, monkeys show abnormal food preference, such as eating meat, which normal monkeys rarely do.

Neurophysiological studies provide evidence that the OFC is involved in the representation of reward value. For instance, neurons in the caudolateral part of the OFC show activations to gustatory stimuli, and they are attenuated when monkeys are satiated by feeding. This suggests that the OFC reflects the value, not the identity, of the food.

The value of reward may also change by the amount of offered reward and by subjective preference for the reward. When monkeys choose between different kinds of juice of different amounts, neuronal activity in the OFC often correlates with total reward value including both magnitude and subjective preference (Figure 1). Some OFC neurons reflect the reward value of the juice that the monkey chooses to consume independently of which juice it is. Others reflect the reward value of a specific juice offered.

Unlike neurons in many other cortical areas, OFC neurons generally do not change activity by the direction of the motor response that the animal needs to make to obtain reward. Little correlation between OFC activity and motor response has also been

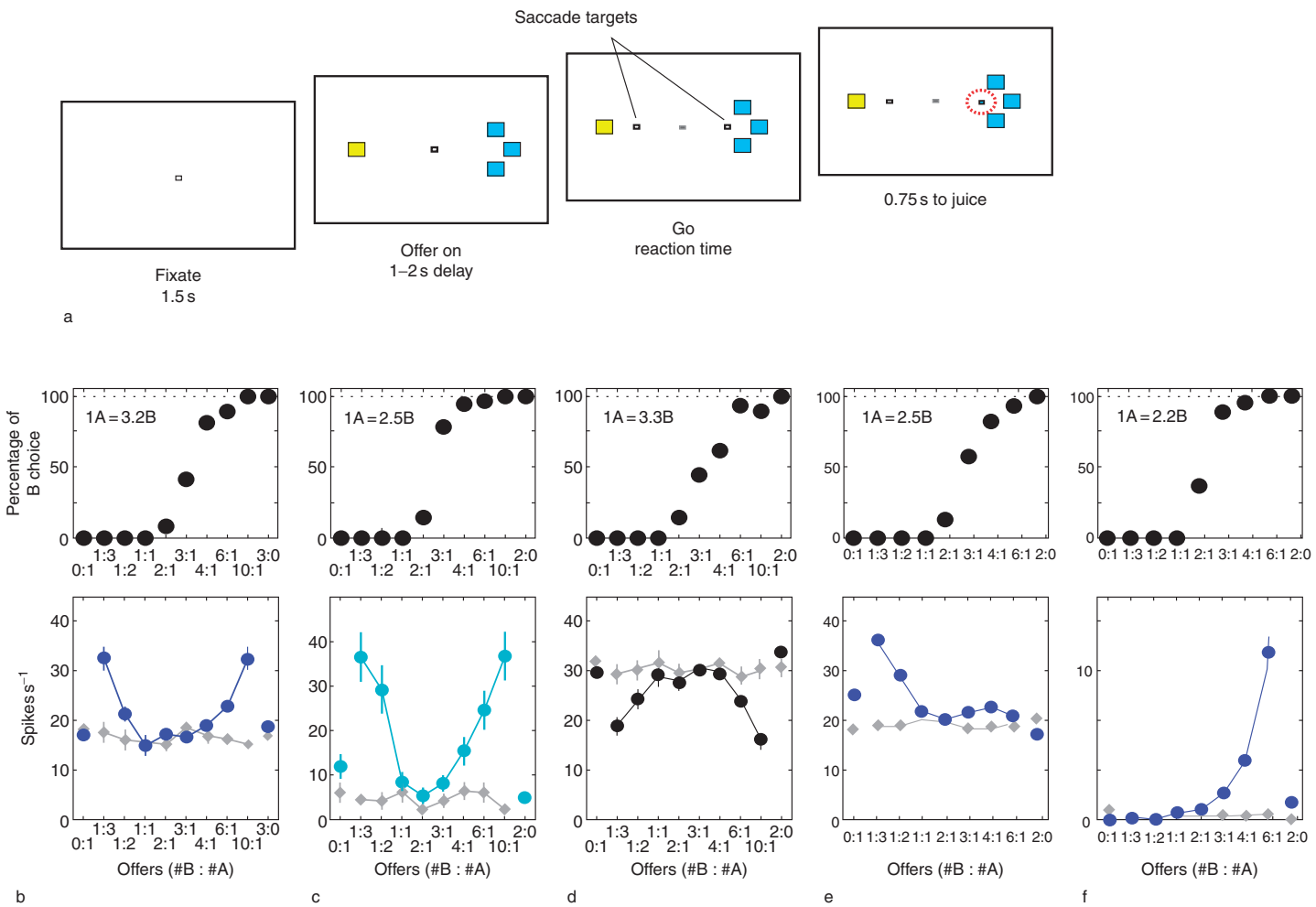

Figure 1 Orbitofrontal neurons reflect value of offered juice: (a) trial structure; (b–f) activity of six orbitofrontal neurons. As shown in (a), each trial begins with the monkey fixating the center of a computer monitor. After 1.5 s, two sets of squares appear on opposite sides of the fixation point (offer). The color of the squares indicates the juice type and the number of squares indicates the juice amount; for example, a monkey offered three blue squares versus one yellow square chooses between three drops of peppermint tea and one drop of grape juice. After a randomly variable delay (1–2 s), two saccade targets appear near the offers (go). The monkey indicates its choice with an eye movement. In (b–f), the top panels show the choice pattern for each cell. The percentage of trials in which the monkeys chose juice B (*y* axis) for various offer types ([amount of juice B] : [amount of juice A]; *x* axis) are plotted. The bottom panels show the activity for each cell. In (b–d), responses encode the chosen value; the response in (d) is negatively correlated with the chosen value (high activity for low value). In (e–f), responses encode the value of juice A offered (e) and the value of juice B offered (f). Responses were recorded in the postoffer ((b, e, f), blue), prejuice ((c), cyan), and postjuice ((d), black) time windows. For each cell, the curves in light gray show the activity in the preoffer time window. Error bars represent s.e.m. Adapted from Padoa-Schioppa C and Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441: 223–226.

demonstrated by using nonspatial types of motor response, such as go/no-go responses. When monkeys perform a go/no-go task, in which pictures instruct the behavioral reaction to be performed (movement versus nonmovement) and the reinforcer to be obtained (liquid versus sound), much of the instruction-cue-evoked activity in the OFC neurons differentiates between the associated reward outcomes. Few neurons change activity depending on the instructed behavioral reaction, irrespective of the type of reinforcer. These results indicate a greater involvement of the OFC in the stimulus–reward association than in the stimulus–response association.

Some OFC neurons reflect the learning process of stimulus–reward associations. When monkeys are presented with a new set of cue pictures that predict different reward outcomes, some OFC neurons respond exclusively during learning or, conversely, become entirely unresponsive. Such changes begin during the initial learning phase and extend into the subsequent consolidation phase.

OFC neurons appear to be sensitive to the magnitude of both reward and penalty. When monkeys perform a choice task in which they are rewarded by a large or small reward or penalized by a long or short time-out penalty, the OFC neurons generally show higher activity when the promised reward is larger and less activity when the threatened penalty is larger.

Anterior Cingulate Cortex

The anterior cingulate cortex (ACC) is a structure located in the medial wall of the frontal lobe. The ACC has reciprocal cortico-cortical connections with the dorsolateral prefrontal cortex and it sends dense projections to the motor cortex and spinal cord. The ACC is the main target area of the mesocortical dopamine system, which originates in the ventral tegmental area. This anatomical background enables the ACC to act as an interface for cognition, motivation, and action.

Neuropsychological and functional imaging studies suggest that the ACC plays a role in detecting the presence of conflicts in behavioral responses and monitoring self-performance by detecting error. Studies on event-related potentials also report a large negative polarity peak, which is thought to be generated in the ACC, occurring between 50 and 100 ms after subjects make errors in reaction-time tasks. Error-related processing in the ACC indicates its sensitivity to negative outcomes. However, evidence is accumulating that the ACC also guides actions for positive outcomes. For example, lesions or inactivation of the monkey ACC significantly impairs the ability to select actions on the basis of rewards (e.g., pull joystick upward after receipt of a pellet to obtain a second pellet and turn joystick to the side after receipt of a peanut to receive a second peanut).

Primate physiology research also provides evidence that the ACC is involved in the process of reward-based action selection. When monkeys perform a visual discrimination go/no-go task, in which different cue pictures instruct different motor responses (go/no-go) and different reward outcomes (presence or absence of liquid reward), many ACC neurons show a large response only to particular response–reward combinations (11–18% of neurons). Very few ACC neurons show activity that correlates with stimulus–reward associations.

ACC neurons are very sensitive to the proximity of the reward. When monkeys perform a task in which visual cues indicate the number of trials (one, two, or three) to be performed before receiving a reward, approximately one-third of ACC neurons increase their activity progressively in every trial as the rewarded trial approaches.

When monkeys have to choose an optimal target out of two paired stimuli associated with rewards of different magnitudes and probabilities, approximately 30% of ACC neurons change activity with the value of the game (theoretical average of reward value associated with the stimulus) independently of whether the game is probabilistic or certain. When monkeys make errors, approximately 12% of task-related ACC neurons show error-related activation, which often covaries with the amount of expected reward – the larger the missed reward, the larger the error-related activity. These findings suggest that the ACC guides behavior by encoding outcome value and deviation from a predicted goal.

Dorsolateral Prefrontal Cortex

The primate dorsolateral prefrontal cortex (DLPFC) has long been implicated in higher cognitive functions, such as switching attention, working memory, maintaining abstract rules, and inhibiting inappropriate responses. For instance, in studies of working memory, animals are typically required to remember a position of a visual cue for a brief period to make a response to the remembered position afterward. DLPFC neurons typically show sustained activity during the memory period. Also, neurons in the DLPFC often have a preferred position in the visual space, so memory-period activity is particularly high when a target is in the preferred position of the neurons (memory field). The spatially tuned activity bridging the gap between sensory and motor events has been regarded as a neural correlate of spatial working memory.

Recent research highlights the influences of reward on activity in the DLPFC. When working memory

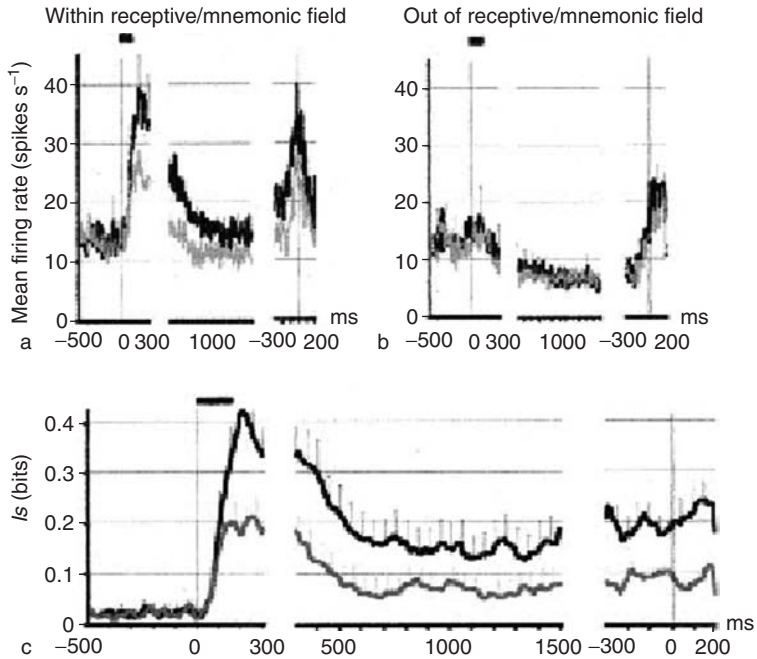

Figure 2 Modulation of spatial working memory by reward expectation in dorsolateral prefrontal neurons: (a) population activity of dorsolateral prefrontal neurons when the target was inside the memory field; (b) population activity of dorsolateral prefrontal neurons when the target was outside the memory field; (c) spatial information (*Is*) of the cue encoded in the spiking activity of dorsolateral prefrontal neurons. Monkeys performed a memory-guided saccade task in which a target appeared either inside or outside the memory field of neurons and the monkeys had to remember the location to make a saccade afterward. The color of the target indicated whether the monkeys could get reward or not after correct performance. Activity is shown in reward-present (black lines) and reward-absent (gray lines) conditions. In (a) and (b), reward expectation enhanced the memory-period activity only when the target was inside the memory field. Error bars indicate s.e.m. Adapted from Kobayashi S, Lauwereyns J, Koizumi M, Sakagami M, and Hikosaka O (2002) Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex. *Journal of Neurophysiology* 87: 1488–1498.

tasks are modified such that animals remember cued information briefly to obtain different reward outcomes, DLPFC neurons show sensitivity not only to what animals need to remember but also to what animals will obtain as a result of performing the task (Figures 2 and 3). The effects of reward are often temporally confined such that activity is enhanced during the memory period. The effects of reward are also spatially confined such that the activity is enhanced when monkeys remember the cue location that is inside the neurons' memory field. The combined representation of memory contents and reward outcome indicates that reward expectation has modulatory effects on working memory processing in the DLPFC.

DLPFC neurons are sensitive to various aspects of reward, including quantity (e.g., large versus small), quality (e.g., raisin versus cabbage), availability (present versus absent), and delay (short delay versus long delay) of rewards. The example of lateral prefrontal activity illustrated in Figure 3 shows different levels of delay-period activity to different kinds of reward expected after performing a delayed-response task.

DLPFC neurons appear to be more sensitive to appetitive than aversive processes. When monkeys perform a working memory task to obtain a reward, to avoid air-puff punishment, or to obtain only sound feedback, neurons respond preferentially to reward-predicting cues (24% of task-related neurons) than to aversively associated cues (6% of task-related neurons).

Reward information instructed by a visual cue reaches the OFC earlier than the DLPFC. This suggests that reward information in the DLPFC might come from the OFC through cortico-cortical connections. Another possible source of reward information in the DLPFC is the mesencephalic dopamine neurons. Various studies have confirmed the effects of dopamine on the functions of the DLPFC: (1) depletion of dopamine impairs the performance of working memory tasks in monkeys, (2) spatial tuning of DLPFC neurons is modulated by the stimulation of dopamine D1 receptors, and (3) extracellular dopamine concentration in the DLPFC increases while monkeys perform a working memory task.

Motor Areas

In the cerebral cortex, movements are represented in anatomically and physiologically distinguishable areas, including the primary motor cortex, premotor area (PM), supplementary motor area (SMA), and

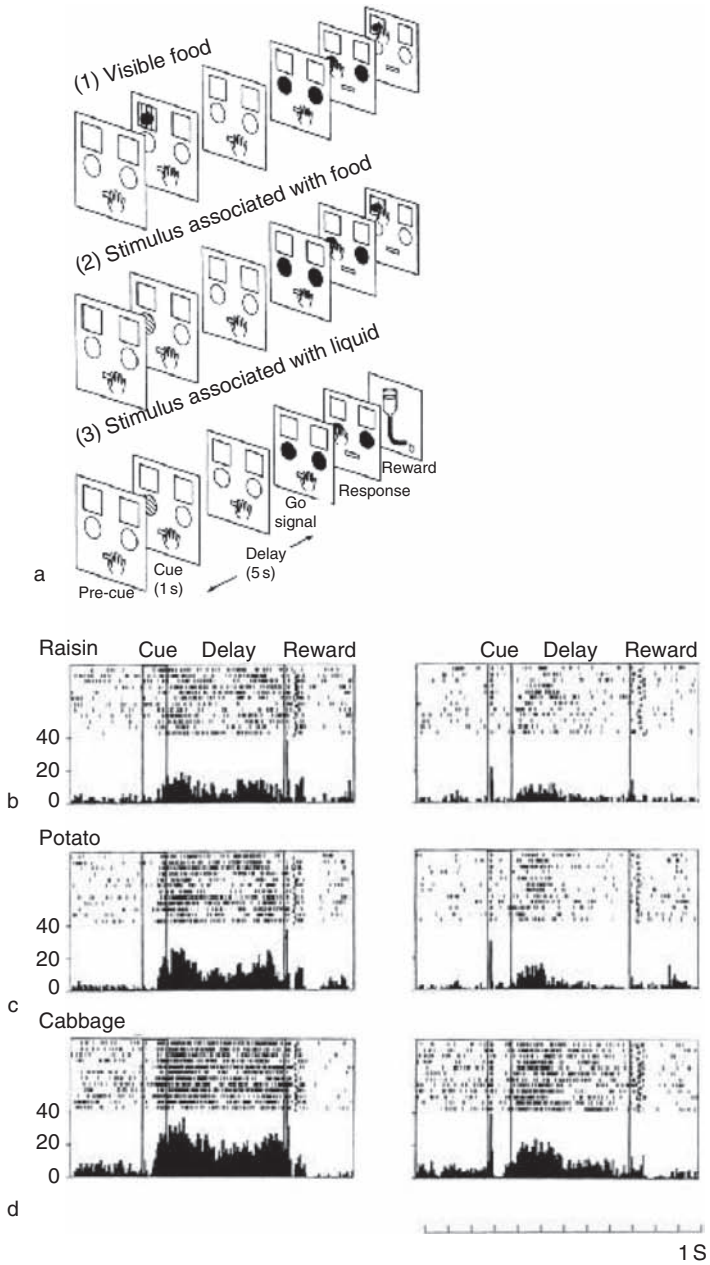

Figure 3 Effects of various food and liquid rewards on the activity of the dorsolateral prefrontal neurons: (a) sequences of events in the three different types of delayed-response tasks; (b–d) examples of a reward-dependent prefrontal delay neurons displaying spatial specificity in the three task types. In (a), the animal first depressed the lever for 6–8 s, and a cue appeared for 1 s, either (1) a visible food reward presented on the right or left window behind the transparent screen or a red light presented on the right or left key as (2) a stimulus associated with food or (3) a stimulus associated with liquid). There was then a delay period of 5 s, which was followed by a go signal of white lights on both keys, and the animal was required to respond to the cued side. Correct responses were rewarded with the food or liquid which had been presented during the cue period (1) or which had been prepared (2 and 3). As shown in (b–d), this neuron showed differential anticipatory activity changes with the different rewards in the left-side trials. Adapted from Watanabe M (1996) Reward expectancy in primate prefrontal neurons. *Nature* 382: 629–632.

cingulate motor area (CMA). There are motor areas devoted to ocular movements, including the frontal eye field and supplementary eye field (SEF). Intracortical electric stimulation in these areas evokes specific movements at low threshold.

Behavioral responses are facilitated by reward expectation in many kinds of behavioral tasks for monkeys. For example, in a saccade task, monkeys make a faster and more accurate eye movement when a higher-value reward is expected. This implies that

motivational states modulate the motor signals represented in the motor areas.

The impact of reward on neuronal activity varies across different motor areas. Neuronal activity prior to movement tends to increase when a large reward is expected, particularly in the PM and SMA. Such enhancement of movement-related activity is presumed to be attributable to motor readiness that changes with the expected reward.

Motor areas also show reward-related activity not directly linked to movements. Two types of reward-related activity were identified in SEF: one predicts the imminence of reward with activity starting before eye movement onset and culminating at the onset of reward; the other signals reward occurrence with activity silent before reward onset and bursting on reward delivery. In the CMA, one of the higher-order motor areas, neurons show more cognitive types of

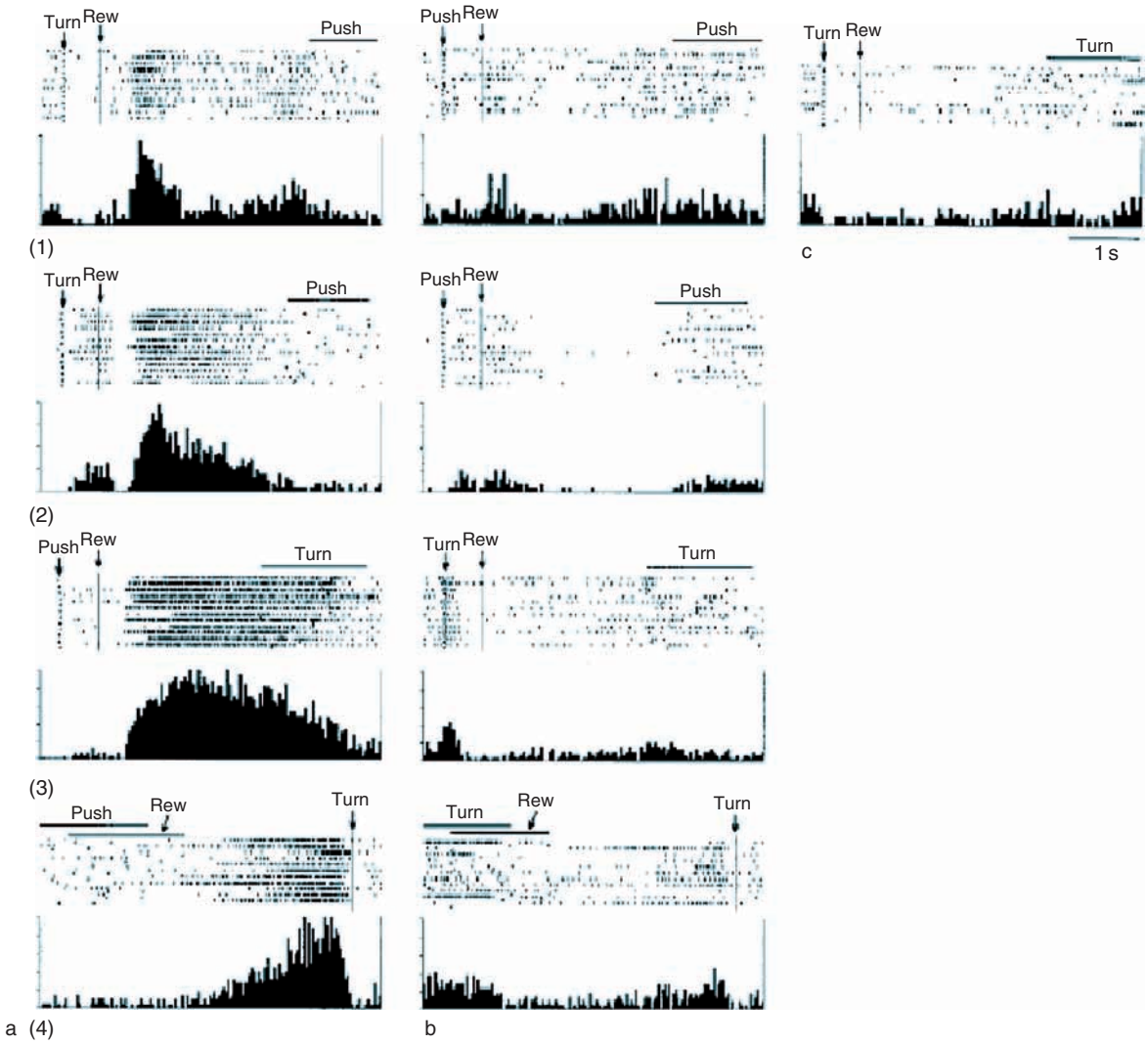

Figure 4 Activity of rostral cingulate motor area (CMAr) neurons related to voluntary movement selection based on reward: (a) activity of four types of CMAr cells (1)–(4) showing increased discharges after the reward was reduced and before initiation of a newly selected movement; (b) the four cells showing less activity under the constant reward condition, when subjects could not select an alternative movement; (c) discharges of CMAr cell (1) when the reward was reduced but the monkey did not select the alternate movement. In (a), each raster corresponds to a single trial when the animal switched from Turn to Push (1)–(2) or Push to Turn (3)–(4), with 5–12 trials in between: (1) short-lasting activity after the reward; (2) long-lasting but decaying activity; (3) continuous activity; (4) activity that increases progressively before the next movement. In (b), dots in the raster displays represent individual discharges of a single cell and small crosses denote the onset of movement. In the histograms, discharges over 11 trials are aligned at the onset of the reward and summed, except for (4), which is aligned at onset of next movement. A step in the ordinate denotes 10 spikes s^{-1}. A horizontal bar in the top right of a raster display indicates the time range of occurrences of the next movement. Adapted from Shima K and Tanji J (1998) Role for cingulate motor area cells in voluntary movement selection based on reward. *Science* 282: 1335–1338.

activity related to both reward outcome and choice of movement. In one study, monkeys selected one of two different arm movements (either pushing or turning a handle). During a series of constant-reward trials, monkeys kept selecting a particular movement. If the reward was reduced, they chose to perform the alternative movement. One-third of task-related neurons in the rostral CMA changed activity during the interval between the occurrence of the reduced reward and the initiation of a new selected movement, suggesting that the CMA plays a role in reward-based motor selection (Figure 4).

Posterior Parietal Cortex

The posterior parietal cortex plays a key role in spatial representation of objects for action planning and control. Primate neurophysiology studies suggest that the posterior parietal cortex receives multimodal sensory inputs and transforms the information from sensory-based coordinates to effector-based coordinates.

Recent studies revealed that the posterior parietal cortex not only serves for hard-wired sensory-motor conversion but also reflects endogenous factors, such as attention and reward expectation. For instance, neurons in the lateral intraparietal area (LIP) of the posterior parietal cortex change activity according to the magnitude or probability of the expected reward after making an eye-movement response.

Neurons in the LIP discharge action potentials in response to visual stimuli that lie in restricted regions of visual space (the neuron's receptive field (RF)). The activity of LIP neurons increases if a target in the RF predicts a higher reward outcome and decreases if a target outside the RF predicts a higher reward outcome (Figure 5).

Reward representation becomes critically important when animals are required to choose responses based on reward. For instance, when animals are free to choose one of the two targets and the reward outcome after each choice varies deliberately in every trial, they cannot plan actions simply by sensory-motor conversion; instead the optimal choice fluctuates in every trial based on expected reward outcome.

The LIP might contribute to guiding choice behavior by its reward-sensitive activity. According to the conventional sensory-motor conversion theory, the simultaneous presentation of two targets activates two concurrent saccade processes. One way to explain how one out of the two saccades is chosen is to assume a covert competition between the two saccade processes that are weighted by the associated reward outcomes. The saccade process associated with higher reward outcome wins the competition and guides a response to the target. The physiological properties of LIP neurons already described support this model.

Temporal Cortex

The inferior temporal (IT) cortex plays a critically important role for the visual recognition of objects. The visual recognition system in the IT cortex is distributed in multiple areas, including area TE and the rhinal cortex. Area TE is a unimodal visual-association area located at the final stage of the ventral visual pathway. The rhinal cortex, which consists of the perirhinal and entorhinal cortices, is a limbic polymodal association area located in the medial temporal lobe. Primate research has revealed that (1) lesions in these areas produce deficits in visual recognition memory, (2) TE neurons selectively respond to various complex features of objects, and (3) the perirhinal cortex is involved in the retrieval of visual long-term memory.

In addition to the functions of visual recognition and memory, IT cortex plays an important role in associating visual stimuli with reward outcome. Anatomically, area TE and the rhinal cortex are mutually interconnected. The perirhinal cortex has strong connections with reward-related sites, including the amygdala, OFC, ventral striatum, and midbrain dopamine neurons.

Lesions in the IT cortex cause behavioral changes that depend on reward schedule. In one study, monkeys learned a task in which visual cues indicated the number of successful trials (one, two, or three) required to receive a reward. Monkeys performed each trial differently depending on the reward schedule, but lesioning or suppressing the expression of dopamine D2 receptors in the perirhinal cortex abolished the reward-schedule-dependent behavior.

Neurons in the IT cortex show activity related to reward. When monkeys perform a go/no-go task in which each visual cue indicates a required action (release or keep pressing a lever) as well as reward availability, approximately one-third of perirhinal neurons and one-fifth of TE neurons show sensitivity to reward availability. In both areas, initial activity shows selectivity to visual stimuli (105 and 123 ms after stimulus onset in the TE and perirhinal cortex, respectively), followed by a reward-sensitive component (198 and 222 ms after stimulus onset in the TE and perirhinal cortex, respectively). The required actions have much less influence on neurons in both areas.

Although it is still unclear, the conversion from visual to reward information might involve the circuitries in the TE and perirhinal cortex, or the conversion might take place in further downstream areas such as the OFC and amygdala, which might in turn project the reward information back to the IT cortex.

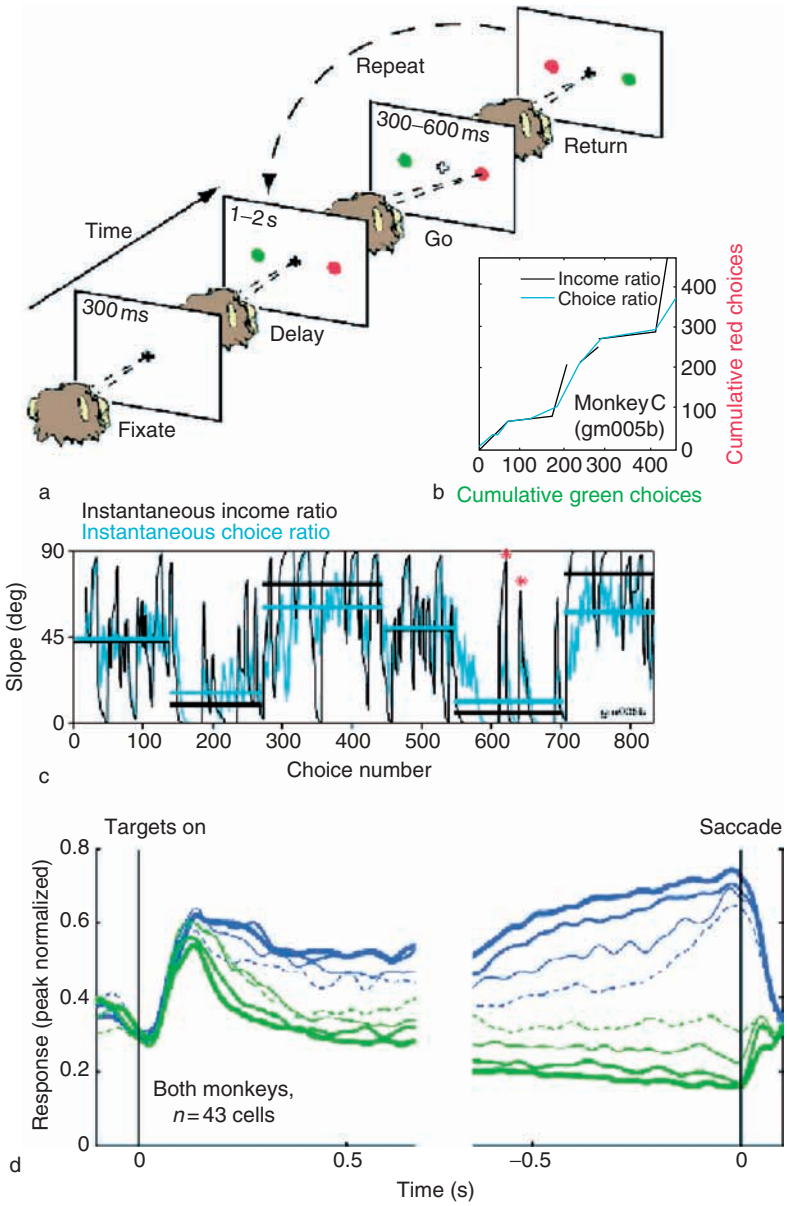

Figure 5 Representations of value in the parietal cortex: (a) the sequence of an oculomotor matching task; (b) dynamic matching behavior; (c) slope space using the same data as in (b), plotted to allow visualization of the ongoing covariation in local ratios of income and choice; (d) activity of LIP neurons dependent on reward probability. In (a), the sequence is fixate (to begin a run of trials, the animal must fixate the central cross); delay (saccade targets appear, randomized spatially by color, in opposite hemifields while the animal maintains fixation); go (dimming of the fixation cross cues a saccadic response and hold); and return (brightening of the fixation cross cues return; the target colors are then rerandomized, and the delay period of the next trial begins). The overall maximum reward rate is set at 0.15 rewards s^{-1}. Relative reward rates change in blocks (~100–200 trials) without warning; the ratios of reward rates were chosen unpredictably from the set {8 : 1, 6 : 1, 3 : 1, 1 : 1}. In (b), the representative dynamic matching behavior of a monkey during a single session is shown. The continuous blue curve shows cumulative choices of the red and green targets. Black lines show the average ratio of incomes (red : green) within each lock (here, 1 : 1, 1 : 3, 3 : 1, 1 : 1, 1 : 6, and 6 : 1). Matching predicts that the blue and black curves are parallel. In (d),the response histograms show average peak-normalized firing rates for 43 cells with activity that regresses significantly on fractional income (the income from that option divided by the total income from all available options). Before normalization and averaging, raw spike trains were smoothed with a Gaussian kernel (SD = 20 ms). Activity is aligned on both the appearance of the visual target (left) and the time of the saccade (right) and is shown separately for choices into (blue) and out of (green) the cell's RF. Trials are further subdivided into four groups according to the local fractional income of the chosen target: solid thick lines, 0.75–1.0; solid medium lines, 0.5–0.75; solid thin lines, 0.25–0.5; dotted thin lines, 0–0.25. LIP, lateral intraparietal area; RF, receptive field; SD, standard deviation. Adapted from Sugrue LP, Corrado GS, and Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. *Science* 304: 1782–1787.

See also: Cingulate Cortex; Connectivity of Primate Reward Centers; Neuropsychology of Primate Reward Processes; Orbitofrontal Cortex: Visual Functions; Prefrontal Cortex: Structure and Anatomy; Prefrontal Contributions to Reward Encoding; Prefrontal Cortex; Reward and Learning; Reward Decision-Making; Reward Neurophysiology and Orbitofrontal Cortex; Reward Processing: Human Imaging.

Further Reading

Amador N, Schlag-Rey M, and Schlag J (2000) Reward-predicting and reward-detecting neuronal activity in the primate supplementary eye field. *Journal of Neurophysiology* 84: 2166–2170.

Amiez C, Joseph JP, and Procyk E (2005) Anterior cingulate error-related activity is modulated by predicted reward. *European Journal of Neuroscience* 21: 3447–3452.

Amiez C, Joseph JP, and Procyk E (2006) Reward encoding in the monkey anterior cingulate cortex. *Cerebral Cortex* 16: 1040–1055.

Campos M, Breznen B, Bernheim K, and Andersen RA (2005) Supplementary motor area encodes reward expectancy in eye-movement tasks. *Journal of Neurophysiology* 94: 1325–1335.

Dorris MC and Glimcher PW (2004) Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* 44: 365–378.

Hadland K, Rushworth MFS, Gaffan D, and Passingham RE (2003) The anterior cingulate and reward-guided selection of actions. *Journal of Neurophysiology* 89: 1161–1164.

Kobayashi S, Lauwereyns J, Koizumi M, Sakagami M, and Hikosaka O (2002) Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex. *Journal of Neurophysiology* 87: 1488–1498.

Kobayashi S, Nomoto K, Watanabe M, Hikosaka O, Schultz W, and Sakagami M (2006) Influences of rewarding and aversive outcomes on activity in macaque lateral prefrontal cortex. *Neuron* 51: 861–870.

Liu Z and Richmond BJ (2000) Response differences in monkey TE and perirhinal cortex: Stimulus association related to reward schedules. *Journal of Neurophysiology* 83: 1677–1692.

Matsumoto K, Suzuki W, and Tanaka K (2003) Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science* 301: 229–232.

Mogami T and Tanaka K (2006) Reward association affects neuronal responses to visual stimuli in macaque TE and perirhinal cortices. *Journal of Neuroscience* 26: 6761–6770.

Padoa-Schioppa C and Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441: 223–226.

Paus T (2001) Primate anterior cingulate cortex: Where motor control, drive and cognition interface. *Nature Reviews Neuroscience* 2: 417–424.

Platt ML and Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400: 233–238.

Roesch MR and Olson CR (2003) Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. *Journal of Neurophysiology* 90: 1766–1789.

Shidara M and Richmond BJ (2002) Anterior cingulate: Single neuron signals related to degree of reward expectancy. *Science* 296: 1709–1711.

Shima K and Tanji J (1998) Role for cingulate motor area cells in voluntary movement selection based on reward. *Science* 282: 1335–1338.

Sugrue LP, Corrado GS, and Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. *Science* 304: 1782–1787.

Tremblay L and Schultz W (2000) Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. *Journal of Neurophysiology* 83: 1864–1876.

Wallis JD and Miller EK (2003) Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience* 18: 2069–2081.

Watanabe M (1996) Reward expectancy in primate prefrontal neurons. *Nature* 382: 629–632.

Reward Processing: Human Imaging

J P O'Doherty, California Institute of Technology, Pasadena, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A reward is anything an animal or human will work to attain. Some rewards are primary reinforcers in that they have innate (or unlearned) value to the organism, whereas other rewarding stimuli can be considered secondary reinforcers in that they have acquired value through being associated with primary reinforcers. Seeking out rewards (or avoiding punishers) forms the basis of most goal-directed behaviors. Therefore, understanding how rewards are processed in the brain is central to understanding the neural basis of behavior. In the last 15 years, modern neuroimaging techniques in humans such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) have complemented single-unit neurophysiology and lesion study approaches in nonhuman animals as a means of providing insight into the neural mechanisms underlying reward processing in the human brain. In this article, the contribution of human neuroimaging studies to our current understanding of neural processes underlying reward is reviewed.

Key Brain Regions Involved in Reward Processing

Although reward-related modulation of neural activity has been found to be widespread in the brain, having been reported in the parietal cortex, dorsal prefrontal cortex, and premotor and posterior cingulate cortices, among other areas, in this article we focus on the functions of a core network of brain areas that play a key role in reward processing. These areas are the amygdala in the medial temporal lobes, the orbitofrontal cortex on the ventral surface of the frontal lobes, the adjacent medial prefrontal cortex, the ventral and dorsal striatum, and the dopaminergic neuron-containing nuclei in the midbrain.

Technical Challenges Facing Human Imaging Studies of Reward

Studying reward presents a relatively unique set of challenges to the imaging neuroscientist. This arises due to an interaction between the physiological properties of the reward system and the properties of the imaging techniques themselves. First, the vast majority of functional neuroimaging studies at present make use of the blood oxygenation level-dependent contrast (BOLD) in magnetic resonance imaging (MRI). This measure depends on detecting subtle differences in local magnetic susceptibility effects caused by neural activity-induced changes in the concentration of deoxy-hemoglobin, as opposed to oxy-hemoglobin. Yet many of the brain areas that are the natural focus of reward studies – namely the orbitofrontal cortex, amygdala, and ventral striatum – happen to be at the border of areas with large susceptibility differences between brain tissue and air in the sinuses. The presence of these large extrinsic susceptibility differences can produce dropout or loss of signal as well as spatial distortion in the BOLD contrast images. A number of techniques have been developed to deal with this problem which can recover much, although not all, of the signal loss in these areas; these include tilted acquisition (typically at 30° to the anterior commissure–posterior commissure (AC–PC) line), application of a compensation gradient prepulse, and the use of a short echo time. An alternative approach is to use PET, which does not suffer from dropout, but the poor temporal and spatial resolution of that technique substantially limits its usefulness. Other MRI techniques such as perfusion MRI may also be less affected by dropout, although at the expense of sensitivity. A second challenge for the imaging of reward systems is that many aspects of reward processing may relate to the functions of specific neurotransmitter systems, including the dopamine system and opioid system, the specific effects of which cannot be isolated with standard hemodynamic imaging approaches. To study this in humans requires administration of specific targeted radioligand molecules which bind with the relevant receptors and that can then be recorded using PET, for example, using the radioligand raclopride to image the dopamine system. Alternatively, the effects of pharmacological modulation of these neurotransmitter systems (by the systemic administration of an agonist or antagonist) on neural activity can be assayed with standard hemodynamic imaging techniques.

Distinct Components of Reward Processing

Reward processing can usefully be fractionated into four distinct components: (1) neural responses pertaining to receipt of a rewarding outcome, (2) neural responses related to expectation or anticipation of a rewarding outcome, (3) neural signals underlying

learning of reward expectancies, and (4) neural responses related to action choice for reward. Here we discuss the role of the different brain regions in the reward network in mediating these distinct components.

Receipt of a Rewarding Outcome

In order to survive, an animal needs to be able to quickly evaluate the affective significance of stimuli in its environment to determine whether such stimuli are beneficial (rewarding) or harmful (punishing) so that appropriate approach or avoidance behavior can be initiated. This raises perhaps the most fundamental question about how rewards might be processed in the brain: Where in the sensory processing pathway is the reward (or punishment) value of a stimulus encoded? Rewarding or punishing stimuli can be present in each of the five sensory modalities, each of which have dedicated sensory-processing pathways. One possibility is that reward value is encoded within the sensory pathway of each modality separately, such that there are separate neural apparatuses for decoding reward value in each sensory pathway. Alternatively, the decoding of reward value may occur in multimodal brain area(s) such that a single dedicated apparatus accomplishes this for each sensory modality. Substantial evidence has now accrued from single-unit neurophysiology studies to support the latter hypothesis, which is that one or two brain areas appear to be involved in encoding the reward value of stimuli from each sensory modality. These areas are the orbitofrontal cortex and the amygdala. Both of these areas receive direct inputs from each of the sensory pathways and are therefore highly multimodal brain regions.

Role of orbitofrontal cortex in encoding stimulus reward value Single-unit neurophysiology studies have reported evidence that neural responses in the orbitofrontal cortex to a stimulus input can be modulated as a function of changes in the value of that stimulus to the animal. Brain-imaging studies in humans also support the conclusion that there is a key role for this area in encoding stimulus reward value for a variety of rewards in different modalities. In a typical study, subjects are scanned while being presented with a range of different stimuli in a specific modality that are either pleasant, affectively neutral, or aversive to the subject. For instance, to study areas encoding stimulus reward value in the olfactory modality, subjects might be presented with a range of odors, some of which are perceived to be pleasant by the subject, some of which are neutral, and some of which are considered aversive. Direct comparisons can then be computed among areas activated in response to the pleasant, aversive, and neutral odors, such that areas involved in encoding positive reward value for an odor could be revealed by testing for areas showing increased activity to the pleasant compared to the neutral or aversive odors. Such studies have revealed increased activity in parts of the orbitofrontal cortex to the receipt of pleasant as opposed to neutral or aversive odors. Other regions of the orbitofrontal cortex have been identified as showing increased responses to aversive odors. A similar methodology has been used in studies exploring the neural encoding of value in other modalities, such as touch, taste, audition, and vision. An example of rewarding stimuli in the visual modality is attractive faces, which human subjects are prepared to work for in order to view, particularly those of the opposite sex. As in the olfactory domain, attractive faces produce significant activity in the orbitofrontal cortex (**Figure 1(a)**). This is also the case for rewards in each of the other sensory modalities that have been tested. An alternative approach to comparing neural responses to stimulus categories is to correlate individual subjective ratings of pleasantness for a given stimulus with neural activity in response to that stimulus to determine the regions that correlate positively with perceived pleasantness. In the case of the attractive face example, this approach again reveals strong activity in the orbitofrontal cortex that correlates positively with subjective ratings of pleasantness for a stimulus.

Yet another approach is to make use of the phenomenon of selective satiation or sensory-specific satiety. This is the phenomenon whereby if a particular food is eaten to satiety, the pleasantness of that food can selectively decrease while the pleasantness of other foods not eaten can remain pleasant. An example of this is a person eating a main course of spaghetti in a restaurant until satiated but still having room for an additional dessert course of, say, tiramisu. The spaghetti has gone from being initially pleasant to having little or no reward value (and may even be aversive if too much is consumed), whereas the tiramisu still has high reward value until it too is consumed to satiety. Neural encoding of reward value can be studied in this way by presenting subjects with two or more food stimuli considered pleasant by a hungry subject, measuring neural activity elicited by these stimuli, and then feeding the subject to satiety on only one of the foods. If, following this feeding procedure, the subject is again scanned while being presented with the two foods, the food eaten to satiety will no longer be pleasant, whereas the food not eaten will remain pleasant. By testing for brain regions responding to the food stimuli that decrease selectively to the food eaten postsatiety, it is possible to identify regions encoding reward value. Once again, these studies reveal

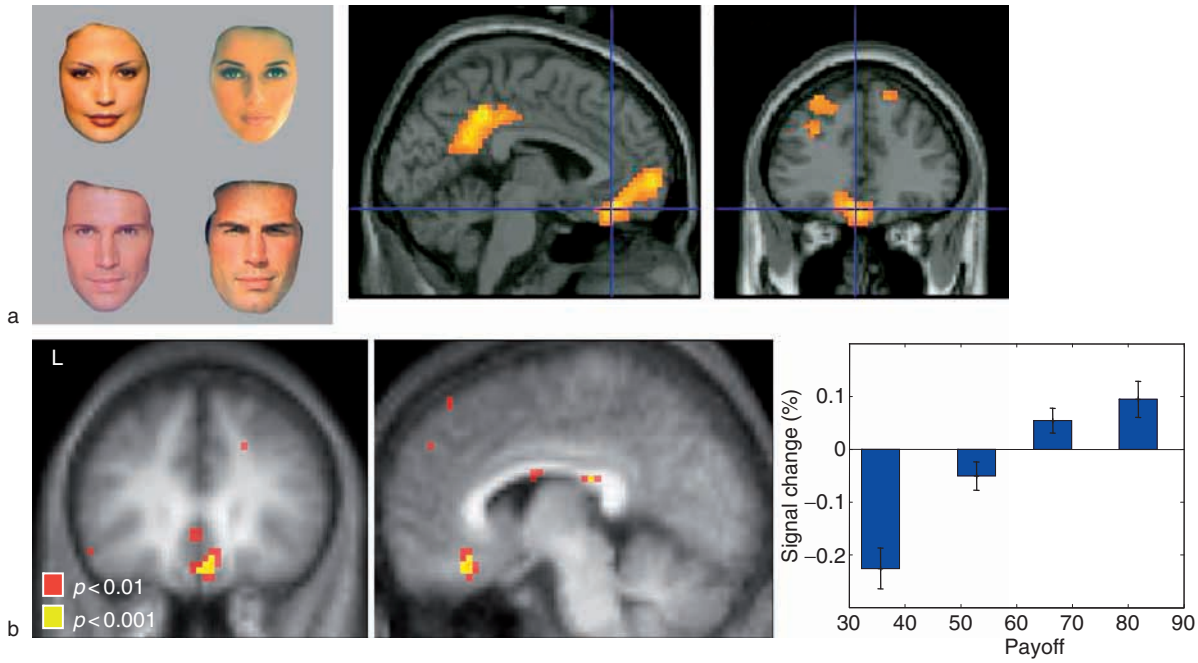

Figure 1 Representation of stimulus–reward value in human orbitofrontal cortex. (a) fMRI study in which subjects were presented with faces that were high and low in attractiveness; (b) fMRI study in which subjects made choices in order to win arbitrary points which they were informed would be converted into money after the experiment. In (a), attractive faces are a class of rewarding visual stimuli that activate the orbitofrontal cortex (example faces from the high-attractive category are shown on the left). A region of medial orbitofrontal cortex showed increased activity during presentation of high-compared to low-attractive faces (right and middle). In (b), the orbitofrontal cortex also activates in response to abstract rewards such as money or arbitrary points. The medial orbitofrontal cortex was found to respond at the time of reward receipt (left and middle). Moreover, the magnitude of activity in this region scaled according to the magnitude of the points received (right). fMRI, functional magnetic resonance imaging. (a) Adapted from O'Doherty J, Winston J, Critchley H, Perrett D, Burt DM, and Dolan RJ (2006) Beauty in a smile: The role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia* 41(2): 147–155. (b) Adapted from Daw ND, O'Doherty JP, Dayan P, Seymour B, and Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441(7095): 876–879.

activity in the orbitofrontal cortex tracking changes in the perceived reward value of stimuli (for both olfactory food rewards and for whole-food stimuli).

Some rewards are not clearly tied to specific modalities, but are more abstract in nature. A paradigmatic example of this is monetary reward, the receipt of which can be conveyed in multiple modalities. Consistent with other types of rewarding stimuli, monetary rewards also activate the orbitofrontal cortex (Figure 1(b)). Furthermore, the magnitude of activation in this area scales with the magnitude of the monetary reward received. Moreover, neural responses to abstract rewards are significantly enhanced if they are obtained in a social context, such as when delivered by a human opponent during the performance of an economic game, compared with delivery in a nonsocial context, typically induced experimentally by informing subjects that they are playing against a computer instead of an individual.

Amygdala and reward In the early neuroimaging literature, the amgydala was often implicated in aversive learning such as fear conditioning, in the perception of fearful or angry faces, or in responding to aversive tastes or odors. This led to an initial emphasis on the role of the human amygdala in negative affect. However, more recently the amgdyala has also been implicated in responding to rewarding stimuli in different modalities, including pleasant tastes, odors, and monetary gain, suggesting that, consistent with the results of single-unit neurophysiology and lesion studies in animals, the amygdala is involved in appetitive as well as aversive processing.

However, the role of the amygdala in encoding either positive or negative valence, has recently been challenged. In two separate studies, one involving gustatory reward and the other olfactory, the intensity and valence of a stimulus were varied separately such that some stimuli were rated by subjects as intense but not pleasant, whereas other stimuli were rated as pleasant but not intense. These studies found that the amygdala responded to the intensity but not the pleasantness of the stimuli. However, a further investigation showed that the amygdala cares not only about intensity but also about the interaction between intensity and valence, such that it responds

more to a highly intense but highly pleasant stimulus than it does to a less intense but still highly pleasant stimulus. This suggests that the amygdala is involved in encoding stimulus reward value, which is the combination of intensity (or magnitude) and valence rather than either valence or intensity alone.

Expectation of Reward

The ability to predict when and where rewarding or punishing stimuli will arrive confers considerable advantage to an animal because it enables appropriate behavior to be organized prospectively. For this reason, humans and other animals have developed sophisticated neural mechanisms to exploit statistical structure in the environment to learn to predict the arrival of such stimuli. Classical or Pavlovian conditioning is the simplest paradigm for studying reward prediction in the laboratory. In Pavlovian conditioning, an arbitrary neutral stimulus takes on predictive value through being repeatedly paired with the subsequent delivery of a behaviorally significant stimulus. Human imaging studies have used Pavlovian conditioning procedures to assay brain regions involved in encoding predictions of future reward. In a typical study, subjects are scanned while being presented with different visual cues (such as fractal images). One of these cues (designated the CS+) is paired repeatedly with subsequent delivery of reward (such as a pleasant odor or a pleasant taste or juice stimulus), whereas another cue (the CS−) is either paired with nothing or with an affectively neutral stimulus. Neural responses related to reward prediction can then be isolated by comparing activity to the onset of the CS+ compared to the CS−. Such studies have revealed significant responses related to reward prediction in both the orbitofrontal cortex and amygdala (Figure 2), suggesting that these regions are involved not only in responding to the receipt of a reward itself but also in anticipating the subsequent delivery of reward. Another region found to show strong expected reward responses is the ventral striatum, composed of the nucleus accumbens and adjacent ventral putamen. Expected reward signals have been found in each of these areas during anticipation of olfactory, taste, flavor, and monetary rewards.

Neuroeconomic variables and expected reward As already demonstrated, the application of paradigms borrowed from the field of behavioral psychology, such as Pavlovian conditioning, has proved useful in neuroscientific investigations of reward processing. Another fruitful avenue has been to bring to bear well-defined concepts from the field of economics as part of the emerging interdisciplinary research agenda known as neuroeconomics. Relevant concepts from economics include expected value, which formally is the product of the probability of obtaining a reward times the magnitude of the reward, and utility, which captures an individual's revealed preferences as a function of the magnitude (quantity) or type of reward. Utility can be expressed as a function of reward magnitude and can be a nonlinear function thereof. For example, the marginal (or relative difference in) utility for an individual of obtaining $102 as opposed to $100 may be much less than the marginal utility of obtaining $4 as opposed to $2, even though the difference in objective reward magnitude is the same. When the preferences for that individual are plotted as a function of magnitude, a concave utility curve is obtained showing that there is a much shallower increase in utility at the $100–102 end of the scale than at the $2–4 end. Expected utility is simply the probability of reward times its utility. This concept can be used to explain differences in preference between risky and nonrisky gambles even if the expected values for the gambles are identical. Variations of utility and expected utility theory have been proposed to account for various anomalies in human preference behavior, such as prospect theory, which accounts for asymmetries in subjects' propensity for risky decision making for gambles involving losses compared to gains.

These concepts have been applied in human neuroimaging studies to investigate neural responses related to expected value, expected magnitude, and expected probability. A number of studies have manipulated these variables by presenting visual cues that independently signal differing probabilities and magnitudes of subsequent reward. Correlations can then be computed between neural activity in response to the different cues and these variables to reveal areas responding separately to probability, magnitude, or the product of the two. Such studies revealed responses to expected magnitude, probability, and expected value in the ventral striatum and medial prefrontal cortex (Figure 3). As yet, no clear evidence has emerged of expected utility signals in the human brain, possibly because neural representations of expected utility may be difficult to dissociate experimentally from expected value because the signals may under many circumstances differ only subtly.

Risk Another variable relevant to anticipated reward is the variance in the distribution of expected reward, referred to as risk or uncertainty. Risk is a quadratic function of the probability of reward, such that risk is maximal at probability 0.5, when the delivery of reward is maximally uncertain, and minimal at 1.0 or 0 probability, when the reward is

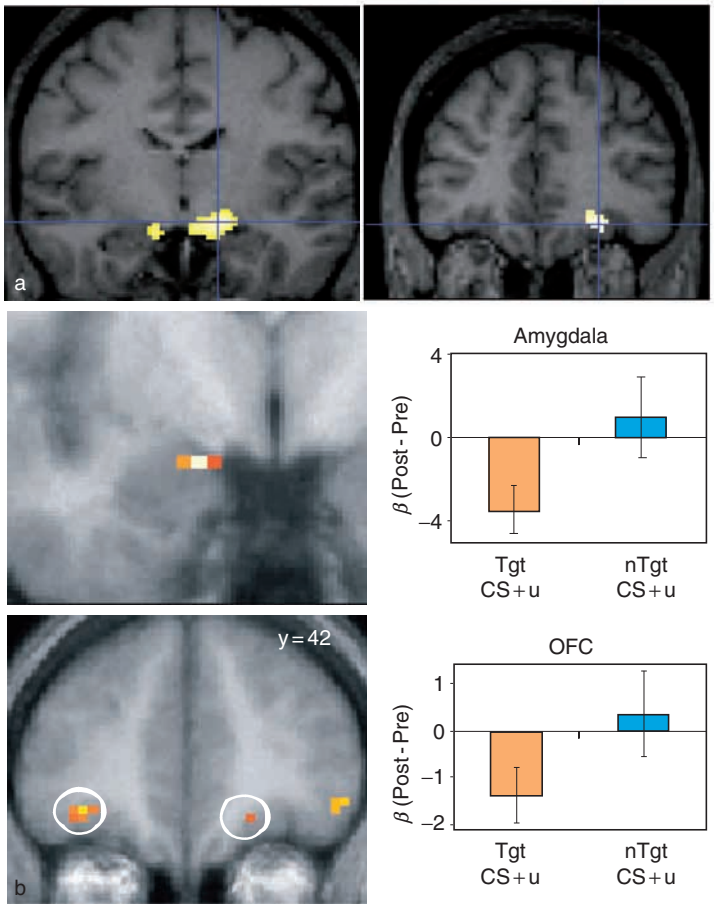

Figure 2 Neural responses during anticipation of reward: (a) neural responses during anticipation of a pleasant taste reward; (b) areas of the orbitofrontal cortex and amygdala showing anticipatory responses that track changes in the value of an associated olfactory reward induced by selective satiation. In (a), human subjects were presented with a visual cue that indicated subsequent delivery of either a pleasant sweet taste, a mildly unpleasant taste, or an affectively neutral control solution. Shown are the areas responding more to anticipation of the pleasant (left) compared to the mildly unpleasant taste (right). In (b), hungry subjects were conditioned to two cues which signaled the subsequent delivery of two food-related odors rewards (vanilla and peanut butter). Subjects were then fed to satiety on one of the foods, selectively devaluing one of the food-related odors. Subjects were then scanned again while being presented with the same two conditioned stimuli. Shown are the amygdala (top left) and orbitofrontal cortex (bottom left) responses to the cues that track changes in the value of the associated odors, indicating that these areas encode anticipatory reward value. The right plots show the difference in activity between pre- and postsatiety to conditioned cues associated with the odor of the food fed to satiety (Tgt CS + u) compared to response to cues associated with odor of food not fed to satiety (nTgt CS + u). (a) Adapted from O'Doherty J, Deichmann R, Critchley HD, and Dolan RJ (2002) Neural responses during anticipation of a primary taste reward. *Neuron* 33(5): 815–826. (b) Adapted from Gottfried JA, O'Doherty JP, and Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301(5636): 1104–1107.

either certain to occur or certain not to occur. Humans are known to vary in their sensitivity to risk; some individuals are risk averse, preferring a nonrisky gamble to a risky gamble if the gambles are equal in expected value, whereas other individuals are risk prone, favoring a risky over a nonrisky gamble with equal expected values. Other individuals are risk neutral in that they show no preference either way. Several fMRI studies have revealed increased activity in the anterior insula or lateral orbitofrontal cortex when subjects are presented with risky as opposed to nonrisky gambles. Given that these areas are often implicated in aversive processing, such as responding to pain or its anticipation, and to receipt of monetary loss, responses to risk in these areas could reflect the negative affective component of risk, at least in those subjects who are risk averse. Risk signals have also been found to be present in the tonic ramping activity of dopamine neurons in singleunit neurophysiology studies in nonhuman primates. Similar signals have been found in fMRI studies in target areas of dopamine neurons such as the ventral striatum; tonic ramping anticipatory activity in this area has been found to correlate

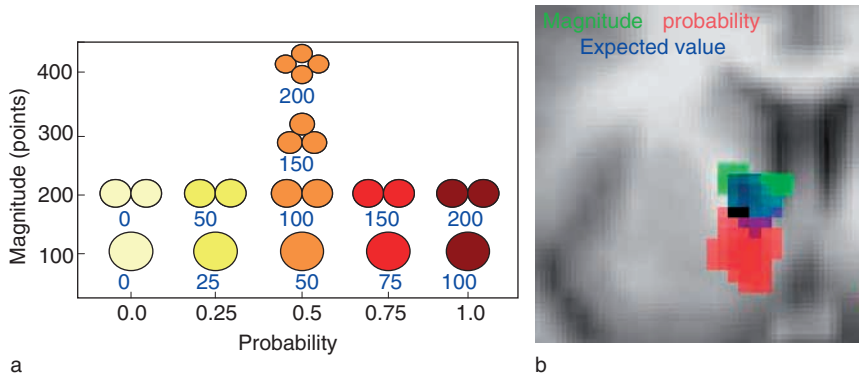

Figure 3 Neural coding of expected reward probability, expected reward magnitude, and expected value in humans: (a) outline of experimental design by Tobler and colleagues; (b) regions of the ventral striatum responding to either magnitude (green), probability (red), or expected value (blue). In (a), subjects were presented with different cues which vary in both color and shape (number of circles). The different colors signaled the probability that reward would subsequently be delivered and the number of circles of indicated the magnitude of reward that would be delivered (in arbitrary points). The blue numbers indicated the expected value (probability × magnitude) for each cue. On each trial, subjects were presented with one of these cue configurations, which signals the probability and magnitude of subsequent reward. Adapted from Tobler PN, O'Doherty JP, Dolan RJ, and Schultz W (2006) Human neural learning depends on reward prediction errors. *Journal of Neurophysiology* 95: 301–310.

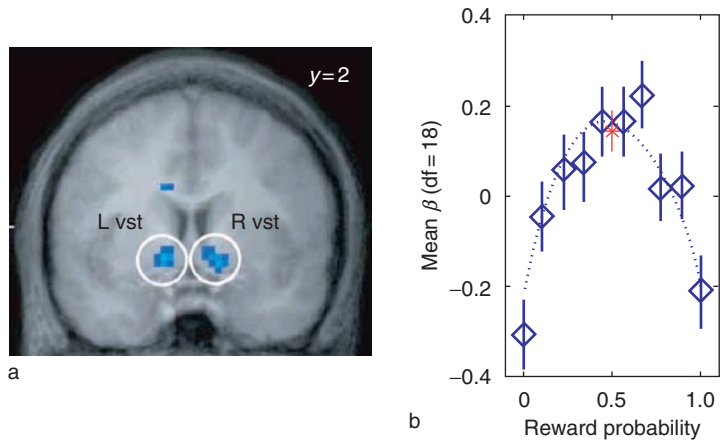

Figure 4 Anticipatory risk signals in the human brain: (a) tonic anticipatory signals in the ventral striatum related to the risk or variance in anticipated reward from an fMRI study by Preuschoff and colleagues; (b) plot of fMRI responses in this region as a function of reward probability, demonstrating a maximal response on trials when reward delivery is most unpredictable (i.e., with a probability $p = 0.5$) and a minimal response when the reward delivery is least unpredictable and is either certain to occur or not to occur ($p = 0$ and $p = 1$, respectively). fMRI, functional magnetic resonance imaging; L vst, left ventral striatum; R vst, right ventral striatum. Adapted from Preuschoff K, Bossaerts P, and Quartz S (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51(3): 381–390, with permission from Elsevier.

with risk or uncertainty (Figure 4). The presence of risk signals in addition to expected value could be important for behavioral decision making because, as discussed earlier, most individuals are sensitive not only to the expected value of a given choice but also to the variance of the expected reward outcome when making decisions.

Learning Expected Reward Signals

The finding of expected value signals in the brain raises the question of how such signals are learned in the first place. An influential theory from behavioral psychology put forward by Rescorla and Wagner suggests that learning of reward expectancy is mediated by the degree of surprise engendered when an outcome is presented or, more precisely, the difference between what is expected and what is received. Formally this is called a prediction error, which in the Rescorla–Wagner formulation can take on either positive or negative sign depending on whether an outcome is greater than expected (a positive prediction error signal) or less than expected (a negative prediction error signal). This prediction error is then used to update the expected reward associated with a particular stimulus or cue in the environment, so that, if this cue always precedes

reward (and hence is fully predictive of reward), eventually the expected value of the cue will converge to the value of the reward, at which point the prediction error is zero and no further learning will take place. The role of prediction error signals in reward learning in the brain was first uncovered through observation of the phasic activity of dopamine neurons in nonhuman primates during the performance of simple conditioning tasks. These neurons were found to demonstrate a response profile approximately captured by real-time extensions of the Rescorla–Wagner model such as temporal difference learning. fMRI studies in humans have found evidence for prediction error signals in target regions of dopamine neurons such as the ventral striatum and orbitofrontal cortex during reward learning (Figure 5). A recent study attempted to directly link prediction error signals observed in fMRI to dopamine release by administering systemic dopamine agonists and antagonists to human subjects while they performed a simple reward-learning task. Prediction error signals in the striatum were found to be significantly affected by the pharmacological manipulation, suggesting that the generation of such signals in humans may be directly related to the activity of dopamine neurons.

Action Selection for Reward

Once predictions of future reward are established, the next step is to use these predictions to guide behavior. More specifically, humans and other animals need to be able to bias their action selection so that they choose those actions which in a given context lead to the greatest probability of obtaining future reward or avoiding punishers. This phenomenon is often referred to as instrumental conditioning and involves learning of stimulus–response and/or stimulus–response–outcome associations. Human neuroimaging studies of reward-based action selection have implicated the dorsal part of the striatum, which is activated when a contingency is established between responses and reward or even when such a contingency is merely perceived. Other regions that may play a role in reward-based action selection include the anterior cingulate cortex, premotor cortex, and lateral prefrontal cortex.

A major focus of current research is to gain an understanding of how the brain can make choices or decisions between different available actions. Human imaging studies have begun to probe the brain regions involved in implementing behavioral decisions using simple choice paradigms. Perhaps the simplest paradigm for this is the *n*-armed bandit task, in which a subject is presented with a number of different options, or bandits, each of which pays out reward with differing probabilities or magnitudes. One of the bandits is generally better than the others in that the consistent selection of that bandit will lead to more cumulative reward than is available from the others. The subjects' objective is to establish which bandit

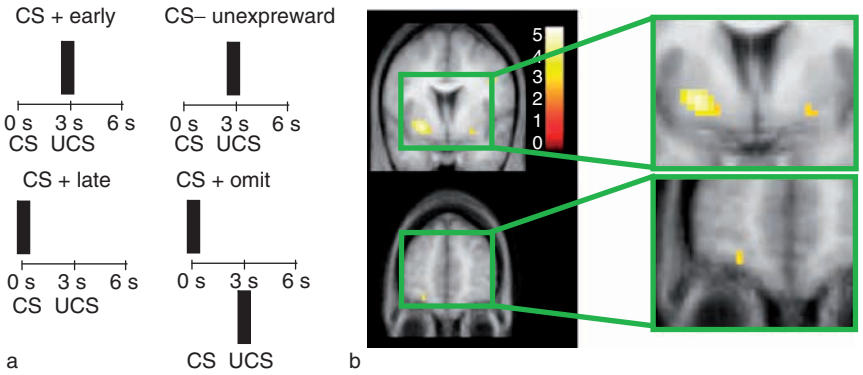

Figure 5 Prediction error (PE) signals in the human striatum (ventral putamen) and orbitofrontal cortex: (a) schematic of putative temporal difference PE signals during the experiment; (b) parts of the human ventral striatum (top) and orbitofrontal cortex (bottom) that showed a significant correlation with the temporal difference PE signal. The study was a classical conditioning paradigm in which in one trial type (CS+) an arbitrary visual cue was associated 3 s later with delivery of a taste reward (1M glucose) and in another trial type (CS–) a different cue was followed by no taste. In addition, occasional surprise trials occurred in which the CS+ was presented but the reward was omitted (CS+ omit) and in which the CS− was presented but a reward was unexpectedly delivered (CS− unexpreward). In (a) schematics of putative temporal difference PE signals during the experiment are shown. During early CS+ trials (before learning was established), the PE signal should occur at the time of delivery of the reward, whereas by the late CS+ trials (postlearning) the signal should switch to the time of presentation of the CS. During CS+ omit trials, a positive PE signal should occur at the time of presentation of the CS+, but a negative PE signal should occur at the time the reward was expected. CS− unexpreward trials should be associated with a positive signal at the time the reward was presented. CS, conditioned response; UCS, unconditioned response. Adapted from O'Doherty JP, Dayan P, Friston K, Critchley H, and Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38(2): 329–337, with permission from Elsevier.

pays the most and choose that consistently in order to maximize reward. However, the subject is not informed at the beginning which bandit is better but, instead, must work this out from trial and error by making selections in order to gain experience of the rewards available from the different bandits.

Useful insights into how humans or other animals might solve this type of problem has come from a branch of computer science known as reinforcement learning, which is concerned with building artificial algorithms for solving stochastic decision problems. The basic idea behind most reinforcement learning models is that in order to choose optimally between different actions, an agent needs to maintain internal representations of the expected reward available on each action and then subsequently choose the action with the highest expected value. Also central to these algorithms is the notion of a prediction error signal which is used to learn and update the expected values for each action through experience, just as in the Rescorla–Wagner learning model for Pavlovian conditioning described earlier. Consistent with the reinforcement learning framework, human imaging studies have found evidence for expected value signals in brain regions such as the medial and orbital prefrontal cortex and amygdala during the performance of simple variants of the *n*-armed bandit task (Figure 6). Moreover, prediction error signals have also been found in both the ventral and dorsal striatum and the prefrontal cortex during the performance of such tasks, consistent with the notion that such signals may be used to update action values as well as Pavlovian stimulus–outcome associations. These studies suggest that reinforcement learning

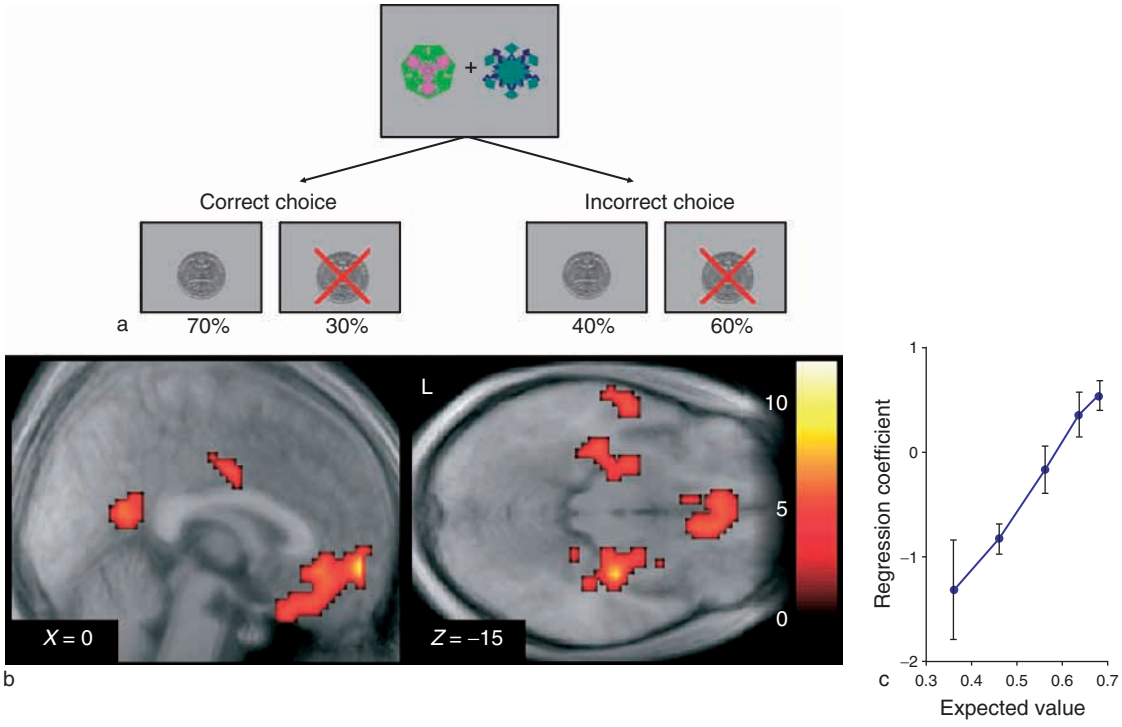

Figure 6 Expected value signals in the orbitofrontal cortex, medial prefrontal cortex, and amygdala during decision making: (a) illustration of payoffs available after the choice of the correct or incorrect stimulus; (b) areas of the medial orbitofrontal cortex, medial prefrontal cortex, and amygdala correlating with expected reward value (derived from a simple computational model) at the time of choice; (c) fMRI responses in medial prefrontal cortex plotted against expected value to illustrate the strong colinearity between fMRI signals in this area and expected value. The task is a simple variant of a two-armed bandit task called probabilistic reversal learning in which a subject chooses between two stimuli which pay out monetary gains and losses on a probabilistic basis. One stimulus is correct, in that the continual choice of that stimulus leads to an overall monetary gain, whereas the other stimulus is incorrect, in that the choice of that stimulus leads to an overall monetary loss. After a time, the contingencies reverse so that the advantageous stimulus becomes the disadvantageous and vice versa. Subjects have to keep track of which stimulus is correct and switch their responses once they deem the contingencies have reversed. In (a), a monetary gain of 25 cents is signaled by the presentation of a US quarter dollar coin, whereas the monetary loss of 25 cents is signaled by the presentation of a red cross superimposed over a quarter dollar. fMRI, functional magnetic resonance imaging. Adapted from Hampton AN, Bossaerts P, and O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience* 26(32): 8360–8367, Copyright 2006 by the Society for Neuroscience.

models may provide a useful analogy to at least some of the neural processes underlying decision making in humans.

Perhaps unsurprisingly, such models are unlikely to account for all aspects of human choice behavior. For instance, these models are unable to learn about higher-order decision rules, such as interdependencies between actions and other exogenous variables including time, whereas expected value representations in the medial prefrontal cortex have been found to encode these decision rules. Furthermore, existing reinforcement learning models may not adequately account for how humans solve the exploration/exploitation dilemma, in which a decision maker needs to balance the need to choose what is currently believed to be the best action with the need to sample other available actions just in case those actions turn out to be more lucrative. A recent study showed that regions of the anterior prefrontal cortex may become engaged when subjects switch between exploratory and exploitative action selection modes, but such a putative switching mechanism is not easily accounted for in current reinforcement learning approaches to this problem. Finally, simple reinforcement learning models may not adequately account for the behavioral distinctions between goal-directed action choice, in which actions are selected with reference to the current incentive value of the associated outcome (involving response–outcome associations), and habit learning, whereby an action is selected purely on the basis of a learned stimulus–response association without reference to the outcome which that response engenders.

See also: Appetitive Systems: Amygdala and Striatum; Cognition: An Overview of Neuroimaging Techniques; Conditioning: Theories; Connectivity of Primate Reward Centers; Cortical Processing of the Reward Value of Food; Delayed Reinforcement: Economics; Delayed Reinforcement: Neuroscience; Neuroeconomics: History; Neuroimaging; Procedural Learning: Classical Conditioning; Reinforcement Models; Reward Decision-Making; Reward Neurophysiology and Orbitofrontal Cortex; Reward Neurophysiology and Primate Cerebral Cortex; Reward Systems: Human.

Further Reading

Breiter HC and Rosen BR (1999) Functional magnetic resonance imaging of brain reward circuitry in the human. *Annals of the New York Academy of Sciences* 877: 523–547.

Daw ND, O'Doherty JP, Dayan P, Seymour B, and Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441(7095): 876–879.

Delgado MR, Nystrom LE, Fissell C, Noll DC, and Fiez JA (2000) Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology* 84(6): 3072–3077.

Gottfried JA, O'Doherty JP, and Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301(5636): 1104–1107.

Hampton AN, Bossaerts P, and O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience* 26(32): 8360–8367.

Holland PC and Gallagher M (2004) Amygdala-frontal interactions and reward expectancy. *Current Opinion in Neurobiology* 14(2): 148–155.

Knutson B and Cooper JC (2005) Functional magnetic resonance imaging of reward prediction. *Current Opinion in Neurology* 18(4): 411–417.

McClure SM, York MK, and Montague PR (2004) The neural substrates of reward processing in humans: The modern role of FMRI. *Neuroscientist* 10(3): 260–268.

Montague PR, King-Casas B, and Cohen JD (2006) Imaging valuation models in human choice. *Annual Review of Neuroscience* 29: 417–448.

O'Doherty J, Deichmann R, Critchley HD, and Dolan RJ (2002) Neural responses during anticipation of a primary taste reward. *Neuron* 33(5): 815–826.

O'Doherty JP (2004) Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology* 14(6): 769–776.

O'Doherty JP, Dayan P, Friston K, Critchley H, and Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38(2): 329–337.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, and Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442(7106): 1042–1045.

Preuschoff K, Bossaerts P, and Quartz S (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51(3): 381–390.

Rolls ET (2004) The functions of the orbitofrontal cortex. *Brain and Cognition* 55(1): 11–29.

Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36(2): 241–263.

Tobler PN, O'Doherty JP, Dolan RJ, and Schultz W (2006) Human neural learning depends on reward prediction errors. *Journal of Neurophysiology* 95: 301–310.

Zald DH and Rauch SL (2006) *The Orbitofrontal Cortex*. Oxford: Oxford University Press.

Reward Systems: Human

M R Delgado, Rutgers University, Newark, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Rewards can be broadly defined as stimuli of positive valence that can influence behavior. They can be manifested through a variety of ways, ranging from appetitive goods to a mere congratulatory pat on the back, and can shape both short-term (e.g., receiving candy after homework) and long-term (e.g., receiving a degree after years of school) goal-directed behavior. Rewards can be observed impacting basic everyday behaviors, such as searching for nutrients, but also more complex social interactions, such as developing trust. The evolutionary function of rewards, therefore, is thought to consist of three primary purposes. First, rewards can induce an overall positive emotional state. Second, the appetitive and hedonic nature of rewards can elicit exploratory or approach behavior. Third, through associative learning, rewards can also serve to reinforce a behavior, or assign a positive value to a previously neutral stimulus or action.

In order to exert an influence on behavior, however, rewards must first be detected by the brain. Information about the various properties of a stimulus, such as its valence, magnitude, probability, and time of consumption, must be integrated to form a representation of the reward value or the value of the action necessary to obtain it. Through experience and a highly adaptive neural network, the brain is then able to show preference toward specific rewards, predict when such rewards will become available, and guide behaviors toward a positive outcome. Although such learning about rewards mostly has a positive influence on life, the neural systems responsible for reward-related processing can go awry, in turn leading to abuse of behaviors that lead to rewarding feelings and contributing to many social maladies. Thus, it is imperative to understand the neurobiology underlying the human reward system, specifically how rewards come to be represented in the brain and how humans learn about them.

Neuroanatomy of Reward Processing

Much of the knowledge about reward systems of the brain come from early theories and a rich animal literature, which highlight some basic ideas about the neural basis of rewards. For example, while information about the physical and chemical attributes of a potential reward is processed by the sensory system (i.e., first through visual, gustatory, olfactory, auditory, or tactile receptors, followed by distributed sensory cortical areas), such information does not initially carry any value to the organism without knowledge of its purpose or effect on behavior. Thus, the coding of a reward representation in the brain is an integrative process that takes into account the physical, but also the subjective, value of a reward with respect to an overall goal. Animal research has been fundamental in delineating a potential neural circuit involved in reward processing and goal-directed behavior, to form the basis of understanding the human reward system. While these elegant animal models and data that derived them are not fully described herein, the neural circuit concept provides an opportunity to highlight the underlying neuroanatomy.

A simple version of the neural circuitry that mediates reward processing involves different regions that have multiple inputs and outputs, with the primary components being areas in the prefrontal cortex, basal ganglia, amygdala, and midbrain nuclei that carry the modulatory neurotransmitter dopamine (Figure 1). In the prefrontal cortex, some reward-related responses have been reported in dorsolateral areas. However, a fair amount of research implicates orbitofrontal cortex as the key prefrontal region where sensory and affective information about rewards is integrated. The orbitofrontal cortex (encompassing predominantly Brodmann's areas 10, 11, and 47) receives sensory and visceral motor information and is hypothesized to mediate many functions that involve linking stimuli classified as reward with hedonic experience or affective value. The orbitofrontal cortex can be further subdivided into a medial and lateral portion, both with distinct functionality, as will be discussed later. The orbitofrontal cortex also has strong connections with mesolimbic structures such as the amygdala, located in the medial temporal lobe, and the striatum, the primary input unit of the basal ganglia. The amygdala, a structure classically considered to be involved in aversive learning, has also been shown to participate in other aspects of emotion processing, namely appetitive or reward learning. The striatum, in turn, due to its connectivity with motor structures, has mostly been associated with motor behavior, although more recent research also links the basal ganglia with cognitive and affective processing. The striatum can be subdivided into both anatomically and functionally distinct components: dorsal (consisting of the caudate nucleus and putamen) and ventral (including nucleus accumbens and ventral putamen) striatum. All three regions, orbitofrontal cortex, amygdala, and striatum, are innervated with dopaminergic input from

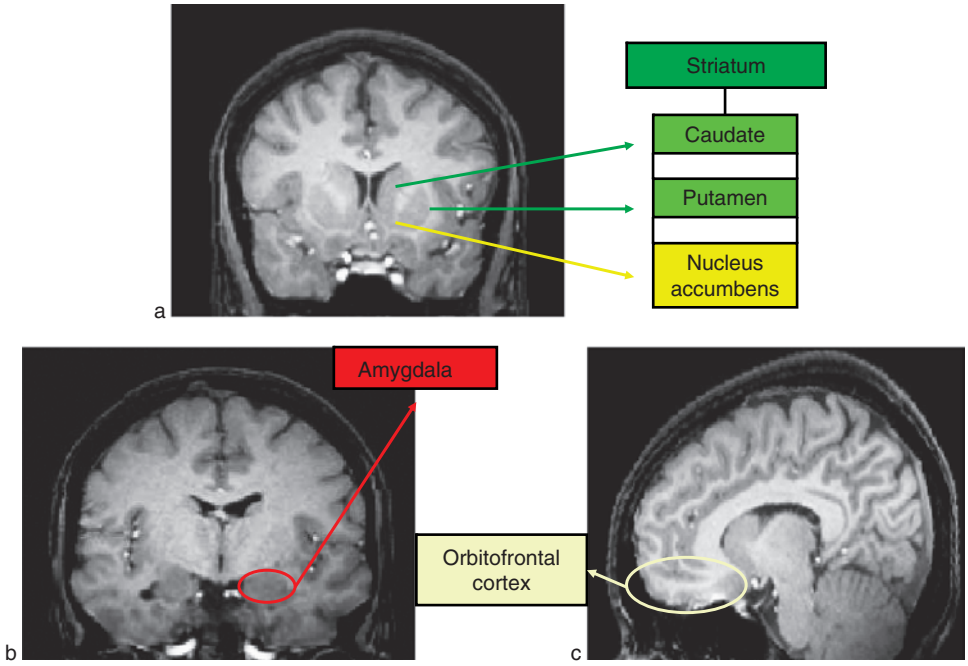

Figure 1 Brain regions involved in reward processing in the human brain. (a) Coronal view of the striatum and its components: caudate, putamen, and nucleus accumbens; (b) coronal view of the amygdala; and (c) sagittal view of the orbitofrontal cortex (approximately Brodmann's areas 10, 11, and 46) resting near the orbit of the eyes.

midbrain nuclei, such as the ventral tegmentum area and the substantia nigra. Dopamine, a neurotransmitter previously thought to be involved in representing a reward response, has more recently been linked not with the response to a reward, but instead with providing a teaching signal that helps the brain adjust expectations about the representation of reward value.

There are other neural structures that have been implicated in specific aspects of reward processing. Some of these other regions, which are not discussed at length herein, include the ventral pallidum, part of the basal ganglia complex, the thalamus, a relay station from the basal ganglia to diverse areas of the prefrontal cortex, the cingulate gyrus, and the hypothalamus. The basic circuit involved in human reward processing, therefore, involves the orbitofrontal cortex, amygdala, and the striatum, along with modulation by dopamine neurons.

Reward Representations in the Human Brain

Recent advances in technology have allowed researchers to confirm and extend basic findings from animal studies by looking at the human reward system using neuroimaging and neuropsychological methodologies. By means of similar types of rewards, such as food and liquids, this research has found a tremendous amount of overlap between species in terms of the neurocircuitry involved in reward processing. A reward such as food is known as a primary reinforcer – it is innate to an organism, and therefore is processed through similar neural pathways across species. However, due to social and cultural evolution, other types of reinforcers available for humans, chiefly money, are also thought to affect behavior through a similar circuitry. This class of reward, consisting of secondary reinforcers, has developed its reinforcing properties through associations with primary rewards (e.g., money, a secondary reinforcer, allows one to purchase food, a primary reinforcer).

Irrespective of the type of reinforcer, human neuroimaging and neuropsychological studies have suggested that regions of the brain –such as the orbitofrontal cortex, amygdala, and striatum – represent and integrate information about the value of rewards (Figure 2). Such representations are formed at time of reward delivery, when an unknown stimulus (i.e., the rewarding stimulus) is assigned value due to hedonic pleasure or satisfaction it provides. After learning an association between a reinforcer and a particular stimulus or behavior, the reward representation is then transferred to the earliest predictor of

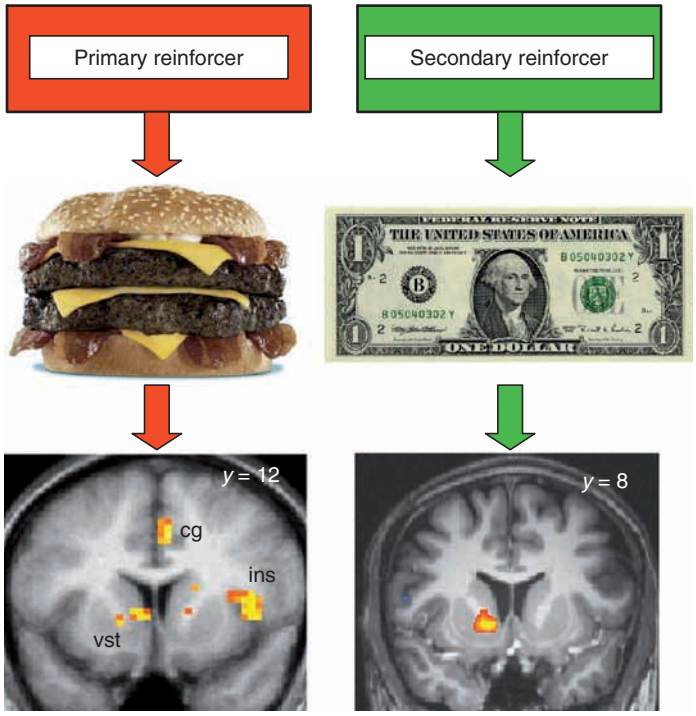

Figure 2 Example of primary and secondary reinforcers and similarities in neural systems. Primary reinforcers (e.g., food and liquids) are innate to an organism while secondary reinforcers (e.g., money) gain their reinforcing properties through learning and associations with primary rewards. Despite the differences between primary and secondary reinforcers, there is some overlap in how information about the reward is processed in the brain. For example, activation in the striatum is observed in experimental paradigms where the incentive is either a primary or secondary reinforcer. vst, ventral striatum; cg, cingulate gyrus; ins, insula; y = Talairach coordinate reflecting coronal positioning. Adapted from Gottfried JA, O'Doherty J, and Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301(5636): 1104–1107 and from Delgado MR, Frank RH, and Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience* 8(11): 1611–1618, with permission.

such reward. These new representations of expected reward value are constantly updated and, consequently, behavior is modified to maximize the availability of rewards.

Coding Reward Value

Once sensory information about a potential reward comes into the brain, a representation of the subjective value must be assigned to aid the dissociation between rewards and nonrewards. This type of representation regarding the affective value of a stimulus is thought to be coded in the orbitofrontal cortex. Activity in orbitofrontal cortex, for example, has been found during functional magnetic resonance imaging (fMRI) studies when comparing a stimulus of positive valence (e.g., a pleasant liquid) with a neutral stimulus (e.g., a tasteless solution). Similar results have been reported irrespective of sensory modality, using gustatory, olfactory (i.e., pleasant odors), and tactile (i.e., pleasant touch) stimuli. These findings also extend into more subjective, yet still sensory, reward representations, as coded by the auditory (e.g., pleasant music) and visual (e.g., beautiful faces) domains. A further dissociation within the orbitofrontal cortex has been suggested between the lateral and more medial portions. While lateral orbitofrontal has been linked with representations of negative stimuli (such as unpleasant touch, or loss of money), the medial orbitofrontal cortex has been implicated with processing stimuli with positive valence.

Neuroimaging studies have also shown that orbitofrontal cortex codes for representations of monetary rewards, concurrent with studies involving primary reinforcers. Some of these studies involve extensive activation in medial aspects of the prefrontal cortex. Further, patients with lesions in the orbitofrontal cortex demonstrate impairments in forming reward representations and updating the contingencies required to obtain such rewards, as evidenced by poor performance in probabilistic learning experiments that involve contingency reversals.

The comparison between an affective and a nonaffective stimulus, however, could also suggest that orbitofrontal activation may be due to the differences in sensory intensity between the stimuli, rather than coding for reward value. Yet, when human study

participants are presented with olfactory stimuli that differ in terms of valence (positive and negative) and intensity (high and low), activity in the orbitofrontal cortex is found to code for the valence of a stimulus, but not intensity (which is associated with amygdala activation). More support for the idea that the orbitofrontal cortex codes for representations of reward value, rather than just sensory properties of a stimulus, is observed during satiation. A hungry human, for instance, will show activation in orbitofrontal cortex when presented with two different food stimuli. After repeatedly feeding on one of the two stimuli, the activation will decrease for such food item due to satiation, but activation will persist for the other stimulus which still has rewarding value. Hence, the myriad of human data supports the existing animal models and suggests that the orbitofrontal cortex is an important structure involved in representing the value of a stimulus during reward delivery.

Expected Reward Value

There is more than one way that a reward representation can be coded by the brain. As discussed, a reward such as a food item can be processed in the orbitofrontal cortex. Yet, if an affective neutral stimulus is associated with delivery of the food item, then such stimulus comes to predict the reward. In this case, the neutral stimulus (which is no longer neutral) serves as a conditioned stimulus and provides an expected or predictive representation of reward value. This expectation of reward is created through learning (either via experience or verbal communication) and is observed in day-to-day life (e.g., pizza box predicts food reward for a hungry individual). The assignment of value to the conditioned stimulus that comes to predict reward occurs through an interaction of neural structures previously described, along with dopaminergic modulation.

Recordings from nonhuman primate dopamine neurons demonstrate that this class of neurons is sensitive to unexpected rewards, such as a drop of juice. Further, if a neutral stimulus, such as a neutral tone, is paired with the delivery of the juice reward (e.g., a pleasant liquid), dopamine neuron cell firing will cease to occur at the time of reward delivery. Instead, dopamine neurons will fire at the earliest predictor of a reward, in this example the neutral tone. This transfer of activity due to Pavlovian learning mechanisms highlights how reward representations can be transferred to any stimulus that has strong associative connections with a positive outcome or reward.

In humans, predictive reward representations can be observed in dopaminergic targets such as the orbitofrontal cortex, ventral striatum, and amygdala. Recent fMRI studies have suggested that the human orbitofrontal cortex codes not only for reward-specific representations, but also for conditioned stimuli that predict reward. Representations of expected reward value in the orbitofrontal cortex may be formed in conjunction with the ventral striatum and the amygdala (Figure 3). Activity in both of these structures, for example, has also been found during appetitive conditioning paradigms involving primary reinforcers such as pleasant liquids and odors. Further, patients with unilateral medial temporal lobe lesions (which include amygdala) are impaired in conditioned preference learning, where patterns are associated with food rewards. Unlike the orbitofrontal cortex, however, the ventral striatum and amygdala have been associated with coding the representation of reward value during the expectation or predictive phase of reward processing, rather than during actual reward delivery. This suggests an important role for these structures in learning about rewards and making predictions about potential rewards.

Representations of predicted reward can be observed in the human orbitofrontal cortex and ventral striatum with both primary (e.g., food) and secondary (e.g., money) reinforcers. Anticipation of monetary rewards, for example, leads to activation in the ventral striatum, since the stimulus inducing the anticipation carries a predictive representation of reward. Interestingly, the overlap between types of reinforcers is not observed in the amygdala, as conditioned stimuli that predict a monetary reward (i.e., a secondary reinforcer) do not usually yield amygdala activation. This potential discrepancy might be because of intensity differences between primary and secondary reinforcers, though limitations in neuroimaging methodology cannot be discounted. It is useful to examine neuropsychological investigations in conjunction with neuroimaging data, which do not imply causality. For example, one experiment found that patients with amygdala lesions did not show anticipatory responses to potential monetary wins and losses in a gambling game (i.e., the Iowa Gambling Task), although the complexities of such a task make it difficult to interpret the deficit as a failure to show predictive representations of value.

Adjusting Expected Values and Actions That Lead to Reward

The brain has a representation of reward value that, through learning, shifts to the earliest predictor of the reward. The conditioned stimulus then provides an expectation of reward. However, the brain must adjust the representation of the expected reward signal if a reward fails to occur (e.g., opening a pizza box but finding it empty). This is especially the case when a reward is contingent on a behavior, as an organism

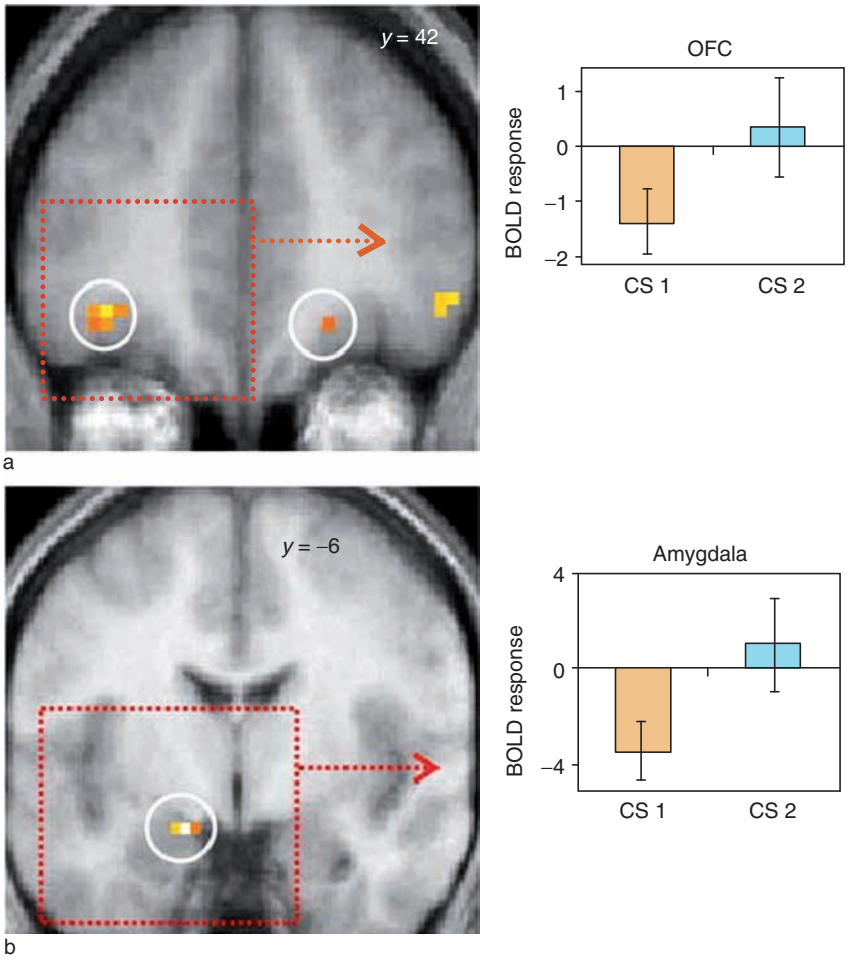

Figure 3 Representations of expected reward value in the orbitofrontal cortex and amygdala during appetitive conditioning experiment. Two predictive stimuli are presented and one stimulus (conditioned stimulus 1, CS 1; orange color) is satiated midway through experiment, while the other stimulus (CS 2; light blue color) is not. Graphs represent differential blood oxygenation level-dependent (BOLD) responses (postsatiation break minus presatiation). (a) Activation of orbitofrontal cortex (OFC) to the devalued conditioned stimulus (CS 1) was decreased compared to the nondevalued conditioned stimulus (CS 2), suggesting that representations of expected reward value are constantly changing with motivational state. (b) The same pattern of results is observed in the amygdala. BOD, blood-oxygen level dependent. Adapted from Gottfried JA, O'Doherty J, and Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301(5636): 1104–1107, with permission.

has to quickly adapt and update action representations that will lead to rewards.

Dopamine neurons provide an illustration of how the nonhuman primate brain might signal a potential error in prediction. The role of dopamine in reward-related learning is thought to be a teaching signal that relays information about the representation of rewards. As previously discussed, these cells typically show learning-related activity demonstrated by a transfer in neural firing from the unexpected reward to the predictive cue. Notably, dopamine cells will also show a depression in neuronal firing when an expected reward is omitted. By signaling the presence of an unexpected reward and the omission of a predicted reward, dopamine is able to provide a reward prediction error signal that aids in goal-directed behavior.

In humans, the prediction error signal is observed in dopaminergic targets, but primarily in the striatum. Different portions of the striatum have shown evidence of coding for errors in prediction in neuroimaging studies. Using primary reinforcers, activation in the ventral putamen has been correlated with a prediction error signal. Using secondary reinforcers, the majority of studies suggest that the nucleus accumbens may be more involved. Although this separation within the striatum between primary and secondary reinforcers has been observed in some studies, it is not uniform and it may merely reflect a difference in type of paradigms or even particular properties of the reinforcer.

A more clear dissociation within the striatum is observed between dorsal and ventral striatum. As

previously described, the ventral striatum is involved in representations of expected reward, highlighted by activation to a predictive conditioned stimulus during appetitive learning paradigms. The dorsal striatum, in contrast, is important for specifically coding the representation of a reward which is contingent on behavior. This is illustrated in conditioning paradigms, where visual cues predict the delivery of rewards either noncontingently (Pavlovian conditioning) or contingently (instrumental conditioning) on choice behavior. Activation of the caudate nucleus, part of the dorsal striatum, is observed only during instrumental conditioning. Thus, the dorsal striatum, particularly the caudate nucleus, appears to be involved in the coding of action specific representations of reward value.

Modulation by Properties of Reward

Once reward-related information is detected by the brain and a representation is formed, such representation is constantly updated to reflect changes in the current value of the reward. This updating depends on the current state of the organism and the properties of the rewarding stimulus. For example, as previously discussed, a food reward will have a representation of value for a hungry individual. This representation of reward value, however, changes as the individual becomes satiated, suggesting that the motivational state of an organism can temporarily affect representations of reward and the neural circuitry involved in reward processing. Additionally, reward representations in the brain also take into consideration the properties of a reward, such as valence, magnitude, probability, and time of consumption.

Valence

Valence refers to the positive or negative quality attributed to a stimulus. For example, glucose water is a pleasant liquid solution considered to have positive qualities, while salty water is an unpleasant solution with negative qualities. Similarly, winning money and receiving positive feedback (e.g., pat on the back) are pleasant stimuli, while losing money and receiving negative feedback (e.g., reprimand) are unpleasant stimuli. Interestingly, the interaction between the valence of a stimulus and its reward value is not as clear-cut as the preceding examples suggest, leaving open questions for future investigations. For instance, although avoidance of a negative outcome such as losing money is generally a positive outcome, can it be considered a reward? What about receiving a small reward, a positive outcome, when a preferred larger one was available (i.e., regret)?

As previously mentioned, the representation of the value of a reward during delivery of such reward is coded in the orbitofrontal cortex. However, human neuroimaging studies also suggest that the human dorsal striatum is active during reward delivery when it is contingent on a behavior. Unlike more medial portions of the orbitofrontal cortex, the dorsal striatum does not create a representation of the reward value *per se*. Instead, it codes the differential activity between stimuli of positive and negative valence during trial-and-error learning, maintaining a representation of the outcome of goal-directed actions to help guide future behavior.

Magnitude and Probability of Reward

An affective stimulus can vary in terms of valence, but also magnitude. Although magnitude could also refer to the intensity of a stimulus, the primary definition refers to the amount or quantity of a reward. Changes in reward magnitude (e.g., increases in the expectation of reward or reward delivery) evoke predictive reward representations in orbitofrontal cortex and ventral striatum that grade parametrically according to increases in value (i.e., higher magnitudes, Figure 4).

Besides magnitude, an important component of the expected reward value representation is the probability of a potential reward delivery. A stimulus that is paired with a reward elicits a prediction of reward delivery. At times, however, the probability of receiving a reward is probabilistic. That is, there is a chance a reward will not be delivered. This is observed in everyday behavior, from extremes such as playing the lottery to hoping there is one more slice of pizza in the box. The probabilistic nature of reward is an important component of behavior, since if every attainable reward was predictable there would be no evolutionary need to explore or show approach behavior. Thus, all regions of the basic reward circuit are influenced by reward probability, as evidenced by expected reward representations parametrically varying in terms of how uncertain they are in regions such as the orbitofrontal cortex and striatum. For example, the striatum is found to be more activated during unpredictable, compared to predictable, delivery of primary rewards. More recently, the medial prefrontal cortex has been implicated in integrating information about the potential magnitude and probability of a reward, forming a representation of the expected value of a reward.

Time

Finally, a representation of reward value can be influenced by the time the reward is expected to be delivered. A reward that is available immediately is thought

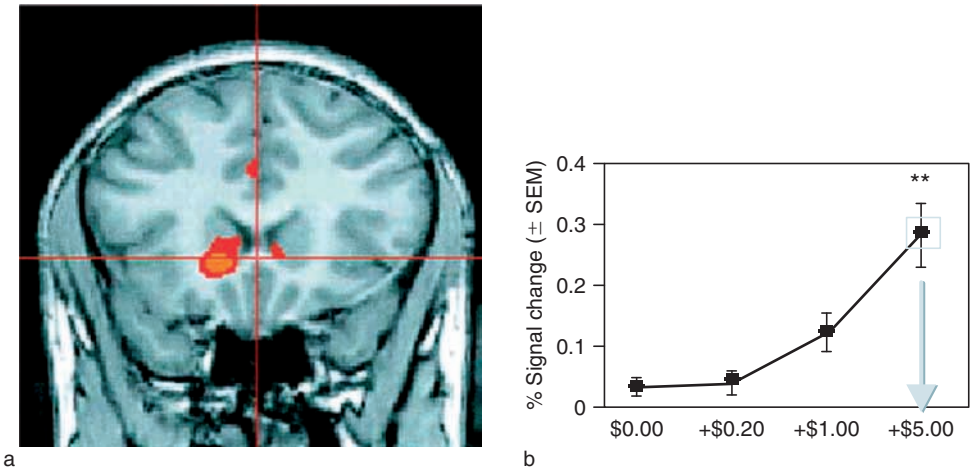

Figure 4 Reward representations in the striatum are sensitive to magnitude differences. (a) Ventral striatum activation during anticipation of different magnitudes of monetary rewards. (b) Levels of blood oxygenation level-dependent response (represented by percent signal change) in ventral striatum varied based on different cues that represented monetary rewards of different magnitudes (ranging from $0 to $5.00). Adapted from Knutson B, Adams CM, Fong G W, et al. (2001) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience* 21(16): RC159, with permission. Copyright 2001 by the Society for Neuroscience.

to be more valuable than a reward available only in the future (e.g., a slice of pizza now vs. a week later). Recent investigations have suggested that the ventral striatum is more involved with processing immediate rewards, while prefrontal cortex regions are linked to evaluation of long-term rewards. This is supported by neuropsychological work with frontal-lesioned patients, who discount future consequences of their actions in a gambling game (i.e., larger losses) in deference to immediate rewards. Yet, more research is needed to fully understand the influence of time on the neural mechanisms of human reward processing.

Modulation of Reward System by Cultural and Social Factors

Together with findings from animal studies, recent research with humans has helped delineate a basic circuit involved in processing simple rewards such as food. Yet, human beings are much more complex and form reward representations through experience with social and cultural rewards which affect behavior. New fields of study are currently being developed to probe these more complex behaviors, which necessitate the merging of disciplines, as illustrated by social neuroscience and neuroeconomics. For instance, in humans, representations of reward value are also subjective; some like Coke and others like Pepsi, while some don't like soda at all. Studies have suggested that subjective preference for cultural rewards such as sugar-based liquids and even types of cars is represented in medial orbitofrontal areas.

Interestingly, the expected representation of social rewards seems to involve primarily the striatum. It is possible that the striatum, particularly the dorsal striatum, is more involved in processing social rewards because of the link between human interaction and action. Social rewards such as love, humor, and cooperation all yield positive feelings, and presentation of a stimulus that predicts such hedonic feelings (e.g., a picture of a loved one) leads to activation in the striatum.

These findings are strengthened by studies involving social and economic interactions, where an individual can either benefit or suffer from an interaction with another person. For example, in a trust game whereby reputations are built through experience, activation of the dorsal striatum is observed coding for a representation of expected reward value when a positive, trustworthy interaction occurs. To demonstrate how complex human nature is, however, such a signal can be influenced by previous information or biases that individuals have, rather than be driven by the presence or absence of rewards.

Finally, the complexities of the human reward system are also observed during abuse of rewards, as displayed by drug addicts, for example. The lure of drugs leads to exaggerated representations of expected reward values (e.g., craving) that can 'hijack' the brain's reward system. The influence of drugs on behavior will lead an addict to bypass natural rewards in favor of drugs, eventually deteriorating quality of life for the addict and those around him.

Summary

Rewards are incentives that influence behavior. The potential rewarding properties of a stimulus and its

effect on goal-directed behavior are coded in the human brain by an intrinsic reward circuit consisting of orbitofrontal cortex, striatum, amygdala, and dopaminergic modulation. Attributes such as a stimulus valence (positive or negative) are represented in these structures. The orbitofrontal cortex is chiefly-involved in assigning value to a stimulus during reward delivery. The striatum and the amygdala are involved in forming representations of the expected reward value of a stimulus, and constantly update such representations to reflect the current motivational state or changes in reward properties (e.g., probability and time of occurrence). The striatum, particularly the dorsal components, is also involved in forming action-specific representations of reward, monitoring the outcome of actions to maximize available rewards.

The human reward circuit processes both simple primary (e.g., food) and secondary (e.g., money) rewards. Current research aims to understand how the basic reward circuitry processes more complex cultural and social rewards, as displayed by subjective preference (e.g., different types of beverages) and human interactions (e.g., trust). Explorations of the simple and complex intricacies of the neural basis underlying human reward processing afford an opportunity to further our knowledge regarding goal-directed behavior, leading to appreciation and treatment of many of societal maladies associated with abuse of the reward system.

See also: Appetitive Systems: Amygdala and Striatum; Connectivity of Primate Reward Centers; Neuropsychology of Primate Reward Processes; Orbitofrontal Cortex: Visual Functions; Prefrontal Contributions to Reward Encoding; Reinforcement Models; Reward and Learning; Reward Processing: Human Imaging; Social Emotion: Neuroimaging.

Further Reading

Anderson AK, Christoff K, Stappen I, et al. (2003) Dissociated neural representations of intensity and valence in human olfaction. *Nature Neuroscience* 6(2): 196–202.

Bechara A, Damasio H, Damasio AR, et al. (1999) Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience* 19(13): 5473–5481.

Berns GS, McClure SM, Pagnoni G, et al. (2001) Predictability modulates human brain response to reward. *Journal of Neuroscience* 21(8): 2793–2798.

Breiter HC, Aharon I, Kahneman D, et al. (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* 30(2): 619–639.

Cox SM, Andrade A, and Johnsrude IS (2005) Learning to like: A role for human orbitofrontal cortex in conditioned reward. *Journal of Neuroscience* 25(10): 2733–27740.

Delgado MR, Frank RH, and Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience* 8(11): 1611–1618.

Gottfried JA, O'Doherty J, and Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301(5636): 1104–1107.

Johnsrude IS, Owen AM, White NM, et al. (2000) Impaired preference conditioning after anterior temporal lobe resection in humans. *Journal of Neuroscience* 20(7): 2649–2656.

King-Casas B, Tomlin D, Anen C, et al. (2005) Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308(5718): 78–83.

Knutson B and Cooper JC (2005) Functional magnetic resonance imaging of reward prediction. *Current Opinion in Neurology* 18(4): 411–417.

Knutson B, Adams CM, Fong GW, et al. (2001) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience* 21(16): RC159.

Kringelbach ML (2005) The human orbitofrontal cortex: Linking reward to hedonic experience. *Nature Reviews Neuroscience* 6 (9): 691–702.

McClure SM, Laibson DI, Loewenstein G, et al. (2004) Separate neural systems value immediate & delayed monetary rewards. *Science* 306(5695): 503–507.

McClure SM, Li J, Tomlin D, et al. (2004) Neural correlates of behavioral preference for culturally familiar drinks. *Neuron* 44(2): 379–387.

Montague PR, King-Casas B, and Cohen JD (2006) Imaging valuation models in human choice. *Annual Review of Neuroscience* 29: 417–448.

O'Doherty JP (2004) Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology* 14(6): 769–776.

O'Doherty J, Dayan P, Schultz J, et al. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304(5669): 452–454.

Rolls ET (1999) *The Brain and Emotion*. New York: Oxford University Press.

Schultz W (2000) Multiple reward signals in the brain. *Nature Reviews Neuroscience* 1(3): 199–207.

Small DM, Gregory MD, Mak YE, et al. (2003) Dissociation of neural representation of intensity and affective valuation in human gustation. *Neuron* 39(4): 701–711.

Tricomi EM, Delgado MR, and Fiez JA (2004) Modulation of caudate activity by action contingency. *Neuron* 41(2): 281–292.

Reye Syndrome

D A Trauner, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In 1963, RDK Reye, an Australian pathologist, described 21 cases of acute encephalopathy with fatty infiltration of the viscera in children. Over the next decade, the clinicopathological features of Reye syndrome became increasingly well recognized. The disease is biphasic, with an antecedent viral illness followed by sudden onset of encephalopathic changes. Reye syndrome afflicts previously healthy children. There is a high mortality rate, particularly in children who have progressed to a comatose state. At the peak incidence of Reye syndrome in the late 1970s, the disorder was the second most common cause of death in the pediatric age group from virus-associated central nervous system (CNS) disease in the United States, reaching a peak incidence of 0.8 per 100 000 in children under the age of 18 years. However, in the early 1980s the incidence of Reye syndrome began to lessen, and now is extremely rare, with only a few cases per year being reported. This article discusses demographic and clinicopathological features of Reye syndrome, current incidence, conditions with similar clinicopathological features, possible etiologies, and treatment and prognosis.

Demographics

Reye syndrome is almost exclusively a disease of childhood, with only rare reports of adult cases. Although children of all ages may develop this condition, there are peak occurrences at 4–5 years of age, and again in early adolescence. Most cases of Reye syndrome occur during the winter months, December through March. This parallels the incidence of influenza virus infections. In the United States, the disease is much more prevalent in Caucasians than in other races. Reye syndrome is more likely to occur in children living in rural or suburban areas. During the peak years of Reye syndrome in the late 1970s, there were clear geographical differences in incidence. The highest numbers of cases were reported primarily from the Midwest, with Ohio, Michigan, Iowa, Nebraska, Colorado, South Dakota, and Utah each having an annual incidence of 1–2.5 cases per 100 000 children under the age of 18 years. Only Georgia, in the southeast, had a similar incidence. These geographic differences have never been explained.

During the time that Reye syndrome had a high incidence rate, an interesting demographic difference was observed relating to age and race predilection. Whereas almost 90% of children who developed Reye syndrome after 1 year of age were white, approximately 50% of infants less than 1 year of age who developed the illness were nonwhite. Furthermore, infants tended to be from more urban areas, while the older children were more likely to live in rural areas. The reason for these differences has not been established. One speculation, however, is that infants living in poorer, inner city areas were likely to become exposed to certain viruses at an earlier age as compared to children living in middle class and more rural locales.

Clinicopathological Features

The typical case of Reye syndrome is preceded by a prodromal, usually upper respiratory, viral illness. Influenza B and A viruses have been linked with most of the case clusters of Reye syndrome, while varicella is the most common viral illness associated with sporadic cases. However, large numbers of different viral agents have been associated with sporadic Reye cases. The ill child is typically recovering from the prodrome when recurrent vomiting develops, followed within 24–48 h by the onset of encephalopathic symptoms (see Table 1). The earliest symptoms of CNS involvement include disorientation and agitation, with progression in some cases to delirium, stupor, and coma. Hyperventilation is a frequent sign, even in the early stages. In the most severe cases, children develop brain stem dysfunction, massive cerebral edema, and brain stem or transtentorial herniation. Mortality rate is high (~70%) in the severe cases. Death is almost always caused by massive brain swelling.

Reye syndrome is a multisystem disease. In the early stages, however, the most consistent laboratory abnormalities are related to hepatic dysfunction. Elevated serum transaminases are uniformly found, with no or minimal increase in bilirubin levels. Prolonged prothrombin time with minimal elevation of partial thromboplastin time is common. Blood ammonia concentrations are almost invariably elevated, at least early in the course of the illness. Hyperammonemia may be transient, however, and so the absence of a clear elevation of blood ammonia concentration may not rule out Reye syndrome. Hypoglycemia is present in approximately 40% of cases, and is more common in children under the age of 4 years. Other metabolic abnormalities commonly observed include

Table 1 Staging of Reye syndrome

Stage 1: Vomiting and lethargy
Stage 2: Disorientation, delirium, combativeness; hyperventilation; appropriate responsiveness to noxious stimuli
Stage 3: Obtundation or coma; decorticate posturing; preserved pupillary reflexes
Stage 4: Coma; decerebrate posturing; dilated pupils unresponsive to light; loss of oculocephalic reflexes
Stage 5: Respiratory arrest; flaccid quadriplegia; absence of tendon reflexes

elevated levels of creatine phosphokinase, plasma lactate, uric acid, and several amino and organic acids. Arterial blood gas determinations generally demonstrate a mixed metabolic acidosis and respiratory alkalosis, with the latter usually predominating. Cerebrospinal fluid (CSF) analysis is unremarkable, with the exception of hypoglycorrachia and elevated CSF pressure.

Acute pancreatitis is a less common but potentially severe complication of Reye syndrome. Abnormal elevations in serum amylase and lipase concentrations may be found. Severe hemorrhagic pancreatitis may be fatal. Renal dysfunction has been reported late in the course of the disease process, but may in part be related to vigorous treatment with hyperosmolar agents, and is usually reversible.

The hallmark pathological feature of Reye syndrome is the presence of diffuse, microvesicular fatty accumulation in the liver, kidney, and heart. Depletion of glycogen from the liver is a common feature as well. Ultrastructural changes are found primarily in the mitochondria of liver and brain. Swollen mitochondria with distortion of the structure of the cristae are typical ultrastructural features. Dilatation of the cisternae of rough endoplasmic reticulum and proliferation of smooth endoplasmic reticulum may also be found. Pathological changes in the brain include diffuse cerebral edema and mitochondrial damage similar to that found in the liver.

Conditions with Similar Clinicopathological Features

Reye syndrome is most likely not due to a single causative factor; rather, the symptoms may be produced by a number of different mechanisms and underlying conditions. Numerous cases of acute encephalopathy with 'Reye-like' features have been documented in children with inborn errors of metabolism, including disorders of mitochondrial metabolism, urea cycle defects, fatty acid oxidation defects, and other disorders of intermediary metabolism. Large numbers of children exposed to toxins, such as aflatoxin B1 in Thailand (Udorn encephalopathy) and hypoglycin A (from eating the unripe ackee fruit) in Jamaica (Jamaican vomiting sickness), have developed acute encephalopathies and hepatopathies identical to Reye syndrome. Insecticide exposure has been implicated in outbreaks of Reye syndrome in several case studies.

Etiology/Epidemiology/Incidence

Various medications taken during the prodromal viral illness have been implicated as possible causative factors in this condition. Only salicylate exposure, however, has been repeatedly demonstrated to have a strong epidemiological association with the development of Reye syndrome. Several state studies, beginning in 1980, documented a statistical association between aspirin and Reye syndrome. Two case control studies from the United States Public Health Service (PHS) also documented a strong statistical association between salicylate use and Reye syndrome, leading the investigators to the conclusion that there might be an etiological association as well.

The results of the epidemiological studies led to the placement of a strong warning label on aspirin bottles in 1986, advising parents not to treat their children with aspirin for viral illnesses. The use of salicylate-containing compounds for children with acute illnesses has dropped dramatically since that time. The incidence of Reye syndrome in this country began to decrease in the early 1980s, prior to completion of the PHS studies and several years prior to the warning labels on aspirin bottles. The incidence continued to decrease to the point that no more than two to four cases per year are currently reported.

Similar epidemiological evidence for an association between salicylate use and Reye syndrome has come from the United Kingdom and France. In Australia and Japan, however, no similar associations have been documented. A recent cluster of cases in India was not associated with a high incidence of salicylate use. In those cases, malathion was suggested as a possible trigger for the illness. An etiologic link between salicylates and Reye syndrome has also not been proved to date. One potential mechanism by which salicylates or environmental toxins might precipitate a Reye-like illness is by reducing the activity of mitochondrial enzymes involved in fatty acid oxidation, resulting in a large increase in circulating fatty acids. High concentrations of fatty acids have been shown to be toxic to mitochondria and to produce further mitochondrial dysfunction and metabolic derangements such as hyperammonemia. It has been suggested that a virus–toxin interaction, perhaps in a

genetically susceptible host, might be required for the development of Reye syndrome.

Beginning in 1976, the US Centers for Disease Control and Prevention (CDC) maintained a voluntary reporting system for Reye syndrome in the United States. Through the late 1970s, the incidence of Reye syndrome averaged approximately 400 cases per year. A peak was reached in 1980, with the highest number of cases (555) ever reported in one year. The number of cases reported to the CDC began to diminish in 1981, and after 1987 fewer than 36 cases per year were reported. Since 1994, no more than two to four cases have been reported per year. Many of the reported cases in recent years have been of teenagers who had taken salicylates, suggesting the possibility that salicylate use among adolescents should be avoided.

The decline and virtual disappearance of Reye syndrome may be related to a decline in salicylate use among children. It has been suggested, however, that this is an apparent decline, due at least in part to the heightened recognition of inborn errors of metabolism that previously mimicked Reye syndrome. Another speculation is that the viral agents responsible for triggering Reye syndrome have changed, such that they no longer produce the same toxic changes in liver and brain that induced Reye syndrome 40 years ago. All of these explanations may be correct. Virus–toxin interactions and virus–metabolic dysregulation interactions might have precipitated the metabolic changes necessary for the development of Reye syndrome.

Although the reported incidence of Reye syndrome in the United States and many other countries is quite low now, there are 'outbreaks' of what appear to be Reye syndrome reported intermittently from different parts of the world. India, for example, has had a number of clusters of what has been reported by the news media as a 'mystery disease' associated with flulike illnesses. However, the clinical features of this illness are identical to those of Reye syndrome, and may represent isolated clusters of Reye syndrome cases.

Treatment and Prognosis

The treatment for Reye syndrome consists of intensive supportive care, use of hypertonic glucose to treat the hypoglycemia, aggressive treatment of hyperammonemia, reversal of other metabolic abnormalities, reduction of increased intracranial pressure, and neuroprotective measures such as cooling. Numerous treatment protocols have been published. One commonly used protocol is outlined in Table 2.

Table 2 General management of Reye syndrome

- Admission to pediatric intensive care unit as soon as diagnosis is entertained
- Nasotracheal intubation if child reaches stage 3 or greater
- After intubation, controlled hyperventilation to maintain pCO_2 between 25 and 30 torr
- Intravenous glucose administration with 10–15% glucose solution
- Maintain serum glucose concentration between 150 and 200 mg%
- Intravenous insulin, 1 unit per 10 g glucose initially, adjusting dose to maintain serum glucose concentration between 150 and 200 mg%
- Intravenous L-carnitine administration using 50–100 mg kg^{-1} per day in three divided doses
- Intracranial pressure monitor placement for any child who reaches stage 3
- Intracranial pressure control using a combination of hyperventilation, osmotic agents such as mannitol, and neuromuscular blocking agents such as pancuronium bromide; intravenous barbiturates may be used if other measures fail to control intracranial pressure
- Reduction of blood ammonia levels can be achieved using neomycin enemas and lactulose by nasogastric tube, or by plasmapheresis or hemodialysis

As mentioned previously, a number of conditions can mimic Reye syndrome. Infections such as varicella encephalitis with hepatitis can produce some similar clinical and biochemical features. Salicylism, poisoning with *Amanita phylloides*, and fulminant hepatic failure from valproic acid use can all cause a Reye-like illness. Numerous inborn errors of metabolism may manifest themselves as an acute metabolic encephalopathy. The most common conditions in this category include genetic disorders of mitochondrial function, medium-chain acyl-CoA dehydrogenase deficiency as well as other fatty acid oxidation defects, systemic carnitine deficiency, hereditary urea cycle defects, and carnitine palmitoyltransferase II deficiency. The immediate treatment for these conditions during periods of acute metabolic decompensation is similar to the treatment for Reye syndrome, but long-term interventions are necessary to prevent recurrent episodes of acute decompensation with life-threatening encephalopathy.

The overall mortality rate from Reye syndrome is approximately 30%. If the severity of the illness at the time of admission is taken into account (based on a severity rating scale), mortality rates are lower for the milder stages of illness, but approach 90–100% if the child is in stage 5 coma at the time of hospital admission (see Table 1). Most of the deaths are caused by massive cerebral edema with brain ischemia or herniation. Other rare causes of death include

fatal cardiac arrhythmias and acute hemorrhagic pancreatitis with shock.

Survivors of the acute illness may have subtle cognitive and behavioral sequelae from their illness, although these consequences may not be apparent for many years after recovery. One recent study found that children who had developed Reye syndrome under 1 year of age, or whose acute illness had progressed beyond stage 2, were more likely to have adverse cognitive and behavioral outcomes. Approximately one-half of the children required some remedial help in school and exhibited problems with self-esteem. Given the marked metabolic derangements and the amount of cerebral edema that can occur with Reye syndrome, it is not surprising that cognitive sequelae are common following this illness.

See also: Encephalitis; Glial Responses to Virus Infection; Mitochondrial Encephalomyopathies; Viral Vectors in the CNS.

Further Reading

Belay ED, Bresee JS, Holman RC, et al. (1999) Reye's syndrome in the United States from 1981 through 1997. *New England Journal of Medicine* 340: 1377–1382.

Hou JW, Chou SP, and Wang TR (1996) Metabolic function and liver histopathology in Reye-like illnesses. *Acta Paediatrica* 85: 1053–1057.

John T (2003) Outbreaks of killer brain disease in children: Mystery or missed diagnosis? *Indian Pediatrics* 40: 863–869.

Meekin SL, Glasgow JFT, McCusker CG, et al. (1999) A long-term follow-up of cognitive, emotional, and behavioural sequelae to Reye syndrome. *Developmental Medicine and Child Neurology* 41: 549–553.

Saudebray JM, Martin D, de Lonlay P, et al. (1999) Recognition and management of fatty acid oxidation defects: A series of 107 patients. *Journal of Inherited Metabolic Diseases* 22: 488–502.

Taylor RW, Schaefer AM, McDonnell MT, et al. (2004) Catastrophic presentation of mitochondrial disease due to a mutation in the $tRNA^{His}$ gene. *Neurology* 62: 1420–1423.

Trauner DA (1992) Reye syndrome. In: Berg B (ed.) *Neurologic Aspects of Pediatrics*, pp. 309–316. Boston: Butterworth-Heinemann Publ.

Ueki I, Koga Y, Povalko N, et al. (2006) Mitochondrial tRNA gene mutations in patients having mitochondrial disease with lactic acidosis. *Mitochondrion* 6: 39–48.

Rho GTPases and Spines

H Zhang, University of Virginia School of Medicine, Charlottesville, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Dendritic spines are tiny protrusions on neurons that receive most of the excitatory synaptic inputs in the brain. Prior to spine formation during brain development, dendrites are covered with actin-rich, filopodia-like protrusions. These protrusions are believed to be actively searching for presynaptic partners, and once a contact is formed, the protrusion matures into a mushroom-shaped spine, which is rich in actin and essentially devoid of microtubules. Thus, the formation of spines requires extensive reorganization of the actin cytoskeleton, and it is likely that regulators of actin dynamics play a central role in spine development. Indeed, evidence is emerging that points to a key function for actin regulators, particularly proteins that participate in the Rho signaling pathway, in spine morphogenesis. This article focuses on the Rho GTPase signaling pathways that lead to actin reorganization and their effects on dendritic spine morphogenesis.

Rho GTPases and Actin Dynamics

The organization of the actin cytoskeleton is regulated by members of the Rho family of small guanine triphosphate (GTP)ases. The Rho GTPases are a subfamily of proteins that belongs to the Ras superfamily of small GTP-binding proteins, which cycles between an inactive guanine diphosphate (GDP)-bound state and an active GTP-bound state. Their activity is regulated by guanine nucleotide exchange factors (GEFs), which promote GTP loading and thereby activate these proteins; and GTPase activating proteins (GAPs), which enhance the intrinsic GTPase activity of these proteins and return them to the inactive state. Inactive GTPases are removed from the plasma membrane by guanine nucleotide dissociation inhibitors (GDIs), which form a 1:1 complex with GDP-bound forms of Rho GTPases and keep them predominantly in the inactive form until they are activated in response to stimulation. Active GTPases signal to downstream effector proteins and generate a cellular response (Figure 1).

There are 20 known members of the Rho GTPase family in mammals, seven in *Drosophila melanogaster*, five in *Caenorhabditis elegans,* and 15 in *Dictyostelium discoideum*. The best-studied members of Rho GTPases include RhoA, Rac1, and Cdc42, and their effects on the actin cytoskeleton are well characterized in fibroblasts. A constitutively activated mutant of RhoA induces the formation of contractile actin and myosin filaments known as stress fibers; constitutively active Rac1 promotes actin-rich fan-like surface protrusions called lamellipodia, which are composed of a highly branched network of actin filaments. Cdc42 promotes the formation of filopodia, which are fingerlike protrusions that are made up of parallel bundles of actin filaments.

Rho GTPases do not act on the actin cytoskeleton directly. Instead, they function through their downstream effector proteins, which then act on proteins regulating actin polymerization, actin filament assembly, or myosin contractility.

RhoA

The effects of RhoA on the actin cytoskeleton are mediated by Rho kinase (also known as ROK or ROCK), which promotes myosin light chain phosphorylation. Rho kinase can achieve this through two different mechanisms. First, it directly phosphorylates myosin light chain. In addition, Rho kinase also phosphorylates and inhibits myosin light chain phosphatase, thereby stimulating myosin light chain phosphorylation indirectly. Phosphorylated myosin binds to actin, thereby promoting actin bundling and the formation of stress fibers in fibroblasts.

In addition to acting on myosin, Rho kinase also phosphorylates and activates Lin-11, Isl-1, and Mec-3 (LIM) kinase (LIMK), which in turn phosphorylates and inactivates cofilin. Cofilin and its related protein actin depolymerizing factor (ADF) increase the rate of dissociation of actin subunits from the pointed ends of actin filaments. They also have F-actin-severing activities that increase the total number of shorter actin filaments. Thus, by inactivating cofilin, the RhoA–Rho kinase pathway serves to stabilize actin filaments.

Another known downstream effector of RhoA is mDia, which is the mammalian ortholog of *Drosophila* Diaphanous. mDia is a member of the formin family of proteins that functions to stimulate actin filament assembly. It is noteworthy that mDia also functions in aligning and stabilizing microtubules, suggesting that the functions of Rho GTPases extend beyond the regulation of actin dynamics.

Rac1 and Cdc42

The pathways downstream of Rac1 and Cdc42 that lead to actin reorganization are more complex. One of the common effectors of Rac and Cdc42 is p21-activated kinase (PAK). PAKs are a family of Ste20-like serine/threonine protein kinases. They contain a conserved C-terminal kinase domain and an

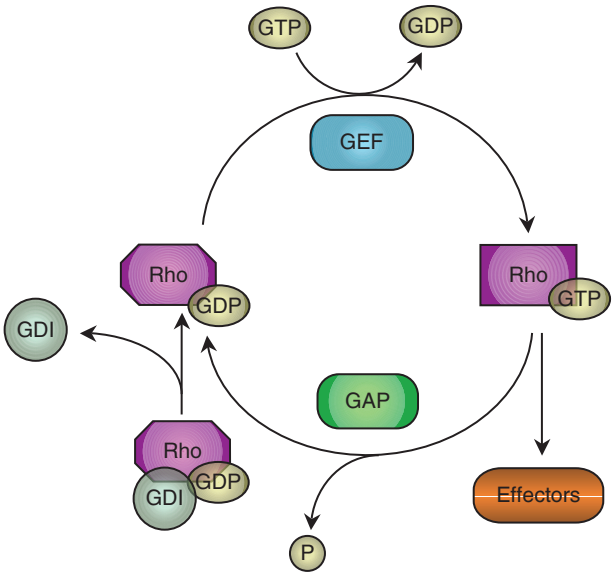

Figure 1 The Rho GTPases cycle between an inactive GDP-bound state and an active GTP-bound state. Their activity is regulated by guanine nucleotide exchange factors (GEFs), which promote GTP loading and thereby activate these proteins, and by GTPase-activating proteins (GAPs), which enhance the intrinsic GTPase activity of these proteins and return them to the inactive state. Inactive GTPases are removed from the plasma membrane by guanine dissociation inhibitors (GDIs), which form a 1:1 complex with GDP-bound forms of Rho GTPases and keep them predominantly in the inactive form. Active GTPases signal to downstream effector proteins and generate a cellular response. GDP, guanine diphosphate; GTP, guanine triphosphate; P, phosphate.

N-terminal regulatory domain. A highly conserved p21-binding domain (PBD) is localized within the N-terminal region. In its inactive form, PAK is folded onto itself. Upon binding of Rac/Cdc42 to PBD, a conformational change is induced, thereby allowing autophosphorylation within the kinase domain of PAK and the subsequent phosphorylation of its downstream targets. Activated PAK phosphorylates and inhibits LIMK1. PAK has also been shown to phosphorylate and inhibit myosin light chain kinase, thereby inhibiting myosin light chain phosphorylation. However, there is also evidence that PAK directly phosphorylates myosin light chain.

Another common downstream effector of Rac and Cdc42 is phosphatidylinositol-4-phosphate 5-kinase (PIP 5-kinase), which converts the phospholipid PI(4)P to PI(4,5)P2. PI(4,5)P2 has a variety of downstream targets that directly regulate actin dynamics. One of the effectors is capping protein, which caps actin filaments and stops their elongation. PI(4,5)P2 inhibits capping protein and facilitates filament elongation. In addition, PIP2 stimulates WASP (Wiscot–Aldrich syndrome protein), which causes the activation of the actin-related protein (Arp)2/3 complex. The Arp2/3 complex stimulates *de novo* actin polymerization and thereby promotes branch formation. Initially, PIP 5-kinase was thought to be a downstream effector of Rac. However, later studies found that PIP 5-kinase was activated downstream of Rac, Cdc42, and Rho, making it a potential convergent point downstream of Rho GTPase signaling.

In addition, both Rac1 and Cdc42 can activate insulin receptor substrate of 53 kDa (IRSp53). Activated Rac1 forms a ternary complex with IRSp53 and its downstream effector WAVE2. WAVE2 is a member of the WAVE/Scar family of proteins that activates Arp2/3. Thus, by activating IRSp53, Rac promotes actin filament branching that is necessary for lamellipodia formation.

By contrast, Cdc42 was reported to form a ternary complex with IRSp53 and Mena, a member of the Ena/vasodilator-stimulated phosphoprotein (VASP) family of proteins. The Ena/VASP proteins promote filament elongation by recruiting profilin, which is a G-actin-binding protein that accelerates the exchange of adenosine diphosphate (ADP)/adenosine triphosphate (ATP) on actin. In addition, they inhibit the activities of the Arp2/3 complex as well as capping protein. These actin-regulating activities of Ena/VASP result in longer and less-branched filaments, which fits well with the ability of Cdc42 to induce filopodia formation. However, a later study showed that IRSp53 co-localized with WAVE2 at the tips of both filopodia and lamellipodia, independently of Mena. Thus, it remains unclear whether the distinct effects of Rac and Cdc42 result from different downstream targets of IRSp53.

In addition to these common effector proteins, Rac and Cdc42 each have specific downstream targets, which may account for their differential effects on the actin cytoskeleton. Activated Rac binds the Sra-1 (specifically Rac1-associated protein 1)–cytoplasmic FMRP (fragile-X mental retardation protein)-interacting protein 1 (CYFIP1) complex, which contains Sra-1, Nck-associated protein (Nap1), Abl-interacting protein 2 (Abi2), and heat-shock protein C300 (HSPC300). The Sra-1 complex binds WAVE/Scar proteins. However, it is not clear how this interaction regulates the targeting and activation of WAVE proteins in response to active Rac.

In contrast to Rac1, the major downstream effector of Cdc42 is the WASP family of proteins. In its inactive state, WASP is folded onto itself through an intramolecular interaction between the N- and the C-termini. Cooperative binding of Cdc42 and PI(4,5)P2 relieves this interaction and activates it. Toca-1, a novel effector of Cdc42, is required for Cdc42- and PI(4,5)P2-mediated WASP activation and actin polymerization. The WASP family of proteins activates profilin, which promotes filament

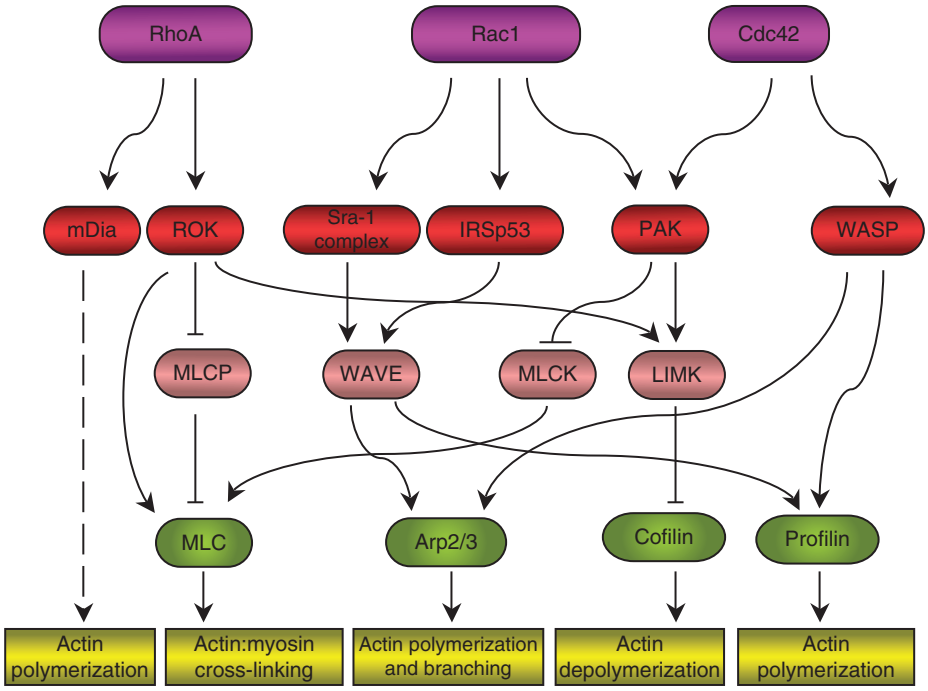

Figure 2 Rho GTPase signaling in actin dynamics, Rho GTPases (purple) signal through their effector proteins (red), which then regulate proteins that affect actin filament extension, branching, or myosin contractility (green). Dashed line indicates indirect effects. Arp2/3, actin-related protein 2/3; IRSp53, insulin receptor substrate of 53 kDa; LIMK, LIM kinase; mDia, mammalian ortholog of *Drosophila* Diaphanous; MLC, myosin light chain; MLCK, myosin light chain kinase; MLCP, myosin light chain phosphatase; PAK, p21-activated kinase; PIP2, phosphatidylinositol 4,5-biphosphate; RhoA, Rac1, Cdc42, Rho GTPases; ROK, Rho kinase; Sra-1, specifically Rac1-associated protein 1; WASP, Wiscot–Aldrich syndrome protein; WAVE, WASP family verprolin-homologous protein.

elongation. Consistent with this, neuronal WASP (N-WASP) induces filopodia formation *in vivo*. However, it remains an enigma how WASP induces the formation of parallel actin bundles characteristic of filopodia when it also activates Arp2/3, which stimulates branch formation (Figure 2).

Signaling from Adhesion and Growth Factor Receptors to Rho GTPases

The Rho GTPases reorganize the actin cytoskeleton in response to a variety of extracellular stimuli. For example, cell–cell adhesion and cell-extracellular matrix adhesion both stimulate Rho GTPases through the engagement of different adhesion receptors, including cadherins, integrins, and Ig superfamily members. A variety of different growth factors also stimulate the activity of Rho GTPases. These adhesion and growth factor receptors target different GEFs and GAPs, which regulate specific Rho GTPases. To date, over 80 different GEFs and over 70 GAPs for Rho GTPases have been found. It is interesting to speculate that these GEFs and GAPs are targeted by different extracellular stimuli. They may also function to regulate Rho GTPases in different parts of the cell so that the activity of Rho GTPases is precisely regulated both temporally and spatially.

In addition to regulation by GEFs and GAPs, Rho GTPases can also be modulated by protein degradation through the ubiquitin–proteasome pathway. In migrating cells, the PAR (partitioning defective) polarity complex recruits Smurf1, an E3 ubiquitin ligase, to filopodia and lamellipodia. Smurf1 functions by locally degrading RhoA, thereby preventing RhoA signaling in protrusive parts of the cell. This provides a novel way of regulating Rho GTPase activity through controlling local protein levels.

Rho GTPases and Spine Morphogenesis

Because the Rho GTPases are central regulators of the actin cytoskeleton, it is conceivable that they play an important role in the morphogenesis of dendritic spines, which are extremely rich in actin. Recent studies have shown that Rho GTPases are indeed key regulators of dendritic spine formation.

Signaling through Rho

Two upstream regulators of Rho have been implicated in spine morphogenesis. One is oligophrenin-1, a Rho-GAP. Oligophrenin-1 is highly expressed in the brain and is present in dendritic spines in hippocampal neurons. Knockdown of oligophrenin-1 expression

causes a reduction in spine length in rat hippocampal slices, a phenotype mimicked by activating RhoA and rescued by inhibiting Rho kinase. Oligophrenin-1 interacts with the synaptic adaptor protein Homer, which is associated with both the metabotropic glutamate receptors (mGluRs) and the *N*-methyl-D-aspartate (NMDA)-type glutamate receptor. Glutamate receptor stimulation is thereby linked to Rho signaling through this interaction.

The second upstream regulator involved is the Rho-specific GEF Lfc. Like oligophrenin-1, Lfc is highly expressed in the brain. Interestingly, Lfc normally localizes to the dendritic shaft of cultured hippocampal neurons; however, on synaptic activation, Lfc translocates to dendritic spines in an NMDA receptor-dependent manner. Overexpression of Lfc results in shorter and smaller spines. Conversely, a C-terminal fragment of Lfc, which acts as a dominant negative mutant, causes the formation of longer spines. Taken together, these studies support a role for Rho in the negative regulation of spine neck length.

The effects of RhoA itself on spine morphology have been examined in several different systems. In all cases, a constitutively active version of RhoA decreases the density of dendritic spines and shortens the spine neck, indicating that RhoA negatively regulates spine formation. Consistent with this, inhibiting Rho activity using C3 transferase was shown to increase spine density and length.

These studies all point to a function of RhoA in the negative regulation of spine density and spine neck length. However, a recent study has made the story more complex. RhoA was shown to be associated with glutamate receptors at the spine membrane. Activation of ionotropic glutamate receptors leads to the dissociation of RhoA from these receptors and the subsequent recruitment of RhoA to mGluRs. This results in a decrease in local RhoA and Rho kinase activities, which leads to spine collapse. Based on these observations, RhoA was proposed to function in stabilizing dendritic spine morphology through Rho kinase. Interestingly, when Rho kinase was locally inactivated by uncaging a caged inhibitor, Y-27632, the spines retracted; however, in another study in which hippocampal slices were treated with Y-27632, no obvious changes in spine morphology were observed. Thus, further studies, especially those involving knockdown of endogenous proteins using short interfering RNA (siRNA), are needed to clarify the roles of RhoA and Rho kinase in spine morphogenesis.

Signaling through Rac

Several Rac GEFs have been implicated in dendritic spine morphogenesis. The first one is Kalirin-7. Kalirin expression in the adult rat is restricted to the central nervous system (CNS), with Kalirin-7 being the most abundant isoform. Furthermore, Kalirin-7 is enriched in the postsynaptic density fraction, where it interacts with several PDZ (Psd-95, Dlg, and ZO1) domain-containing proteins through its extreme C-terminal binding motif. In response to EphB receptor activation, Kalirin-7 translocates to synapses, where it mediates Rac activation. Overexpression of Kalirin-7 induces the formation of numerous aberrant spinelike structures, and the induction of these structures is dependent on its GEF activity. In addition, reduced expression of Kalirin results in decreased density of spines and a concomitant loss of pre- and postsynaptic markers. These results provide a pathway by which transsynaptic ephrin-EphB receptor activation leads to Rac-mediated actin reorganization.

A second GEF involved in spine formation is β PAK-interacting exchange factor (βPIX), which is a GEF for both Rac1 and Cdc42. βPIX localizes to synapses through its interaction with G-protein-coupled receptor interacting protein 1 (GIT1) and possibly the synaptic scaffolding protein Shank. Overexpression of βPIX causes the formation of multiple long, thin dendritic protrusions, whereas a GEF-deficient mutant of βPIX decreases spine formation. In both cases, a decrease in synaptic density was observed, suggesting that the activity of βPIX needs to be precisely regulated for synapses to form.

More recently, the Rac GEF tumor invasion and metastasis 1(TIAM1) has also been shown to be involved in spine formation. TIAM1 interacts with the polarity protein partitioning defective 3 (PAR-3), the EphB2 receptor, and also the NMDA receptor. Overexpression of TIAM1 causes the formation of multiple filopodia and lamellipodia-like structures, whereas knockdown of TIAM1 expression decreases spine density. TIAM1 phosphorylation is regulated by NMDA receptor activity in a calcium-dependent manner, thereby providing a potential mechanism by which NMDA receptor stimulation induces Rac-dependent actin remodeling in dendritic spines. In addition, activation of EphBr induces tyrosine phosphosylation of TIAM1 and its recruitment to EphB complex containing NMDA receptor. This makes TIAM1 a potential convergent point of EphB and NMDA receptor signaling.

Finally, a recently identified GEF for Rho GTPases, GEFT, has been implicated in spine morphogenesis. GEFT is highly expressed in the hippocampus, granular cells, and Purkinje cells of the brain. Ectopic expression of GEFT promotes dendritic spine formation in rat dentate gyrus explant cultures. However, because GEFT seems to show guanine nucleotide exchange activity for Rac1, Cdc42, and RhoA, at

this point it is unclear which Rho GTPases mediate the effects of GEFT on spine formation.

It is interesting to note that multiple Rac GEFs have been shown to be involved in spine morphogenesis. Loss of function of an individual GEF seems to be sufficient to cause defects in spine formation. It is interesting to speculate that each Rac GEF mediates the signaling downstream of different receptors and possibly synaptic adhesion molecules. Proper signaling through each pathway is necessary for spine morphogenesis, and abnormal signaling in any one pathway can cause defects in spine formation.

What are the effects of Rac1 itself on spine morphogenesis? Studies in various different systems have shown that mutating Rac1 to either the constitutively active form or the dominant negative form causes defects in spine formation. Activated Rac1 causes the formation of long, thin protrusions as well as broad veil-like protrusions on the dendrites in dissociated hippocampal neurons and hippocampal slice cultures. In cerebellar Purkinje cells of transgenic mice expressing constitutively active Rac1, dendritic spines were much smaller in size but more numerous. A dominant negative form of Rac1 causes the elimination of dendritic spines in both dissociated hippocampal neurons and hippocampal slices. Thus, it is likely that cycling of Rac between the active and the inactive forms is necessary for proper spine morphogenesis. Interestingly, activation of Rac1 is able to recruit α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors to nascent synapses, suggesting a mechanism by which regulated activation of Rac promotes spine formation.

Which proteins that mediate spine morphogenesis are downstream of Rac? Recent data show that PAK proteins and IRSp53, both downstream effectors of Rac, are involved in spine morphogenesis. Kinase dead versions of both PAK1 and PAK3 inhibit spine formation, much like the dominant negative mutant of Rac1. Similarly, perturbing IRSp53 function, either by knocking down endogenous protein expression using siRNA or by expressing a dominant negative mutant, reduces the density, length, and width of dendritic spines. However, the activated Rac1 phenotype is not copied by these effector proteins. Unlike active Rac1, which stimulates the formation of filopodia and lamellipodia-like dendritic protrusions at the expense of mature mushroom-shaped spines, activated mutants of PAK proteins, rather, stimulate spine formation. Similarly, IRSp53 promotes spine formation when overexpressed in hippocampal neurons. These data suggest that PAKs and IRSp53 are not the only downstream effectors of Rac1 in mediating spine morphogenesis and that additional effectors remain to be found.

Rac–Rho Reciprocity

Previous studies in nonneuronal cells suggested an antagonistic effect between Rac activation and Rho activation. Rac activation was shown to downregulate Rho activity and vice versa. However, how this occurs is not entirely clear. One possible mechanism is through a pathway downstream of Rac that leads to the activation of p190RhoGAP, which in turn inactivates Rho. Similar antagonistic effects are also observed in dendritic spine morphogenesis. Activated Rac promotes the formation of dendritic protrusions that are precursors of mature spines; whereas activated Rho inhibits spine formation, a phenotype reminiscent of dominant negative Rac. However, in this context, it is not yet known whether Rho inhibition is a result of Rac activation or the two processes occur independently.

The Rac–Rho reciprocity was observed in a recent study examining the role of p120 catenin in dendritic spine and synapse development. p120 catenin (p120 ctn) is an armadillo family member that associates with the juxtamembrane region of the cadherin family of cell–cell adhesion receptors. Conditional knockout of the p120ctn gene in the dorsal forebrain results in neurons with reduced spine and synaptic density. Further analyses showed that Rac activity was downregulated in these neurons; whereas Rho activity was upregulated. Activating Rac or inhibiting Rho was able to rescue the defects in p120ctn mutant neurons. Furthermore, overexpression of p120ctn results in the formation of multiple long thin dendritic protrusions, a phenotype reminiscent of neurons expressing constitutively active Rac. Interestingly, this phenotype can be almost completely rescued by expressing a constitutively active Rho, suggesting the possibility that the active Rac phenotype is a result of inhibited Rho activity.

Signaling through Cdc42

The role of Cdc42 in spine morphogenesis is less well defined. In the *Drosophila* visual system, loss of function of Cdc42 leads to a 50% reduction in the density of spinelike structures. However, in mouse cortical and hippocampal slices, neither constitutively active nor dominant negative Cdc42 has a major effect on spine morphology. In cultured hippocampal neurons, Cdc42 is activated downstream of EphB receptor tyrosine kinase. This process is mediated by the adaptor protein Numb and the Cdc42 GEF intersectin, and N-WASP appears to function downstream of this pathway. Interestingly, dominant negative Cdc42 seems to induce filopodia formation at the expense of mature mushroom-shaped spines. Consistent with this, a dominant negative mutant of intersectin induces

the formation of filopodia-like structures; however, the overall protrusion density is significantly reduced. These studies seem to be at odds with the ability of active Cdc42 to induce filopodia in nonneuronal cells. Clearly, further studies are needed to better explain the discrepancies among different systems and to better define a role for Cdc42 in dendritic spine development.

Rnd1 and Spine Development

Although there have been more studies focusing on the roles of RhoA, Rac1, and Cdc42 in spine formation, it is important to note that there are 17 other Rho GTPase family members in mammalian cells and that they are also likely to function in regulating dendritic spine morphogenesis. One example is Rnd1, which is mainly expressed in the brain. The expression level of Rnd1 is highest during the postnatal synaptogenesis period. Overexpression of Rnd1 results in the elongation of spines without significantly affecting spine width or density. Suppression of Rnd1 expression using antisense oligonucleotide causes the formation of headless protrusions at the expense of stubby or mushroom-shaped spines. These headless protrusions are largely devoid of PSD-95, a prominent protein that localizes almost exclusively to the postsynaptic density of excitatory synapses. This shows that these protrusions are immature spines. However, in contrast to the immature filopodia-like dendritic protrusions normally present on young neurons, the headless protrusions induced by Rnd1 antisense are much shorter in length, indicating that Rnd1 may be necessary for the elongation of filopodia. Although the exact mechanism for the effect of Rnd1 on spines is not clear, previous studies suggest that Rnd1 can inactivate Rho by binding to and activating p190RhoGAP. This is consistent with the data indicating a positive effect of Rnd1 on spine length and a negative effect of Rho on spine length.

Spatial Regulation of Rho GTPase Activation

An emerging theme is that Rho GTPases are probably activated in specific locations in the cell, and that global activation or shutdown of Rho GTPase activity is likely to cause pleiotropic defects in cellular processes. This is likely to be the case in dendritic spine formation as well because both constitutively active and dominant negative Rac1 cause defects in spine formation. Therefore, the activity of Rac1 is likely to be both spatially and temporally regulated for proper spine morphogenesis. In support of this, Rac1 is locally activated in dendritic spines. This has been demonstrated using fluorescence resonance energy transfer (FRET). FRET is a nonradiative transfer of energy from a donor fluorophore to an acceptor fluorophore. To visualize Rac activity in dendritic spines, a single chain FRET probe was used. This probe has yellow fluorescent protein (YFP), Rac1, the p21-binding domain of PAK (PBD), and cyan fluorescent protein (CFP) encoded on a single cDNA. Upon activation, Rac binds PBD, which brings YFP and CFP to close proximity, thereby allowing FRET to occur. Using this probe, it was shown that Rac is locally activated in a fraction of the spines and also at the base of the dendritic protrusions. The spines containing activated Rac are contacted by presynaptic terminals, showing that they bear synapses. It is tempting to speculate that the spines with activated Rac are undergoing synaptic remodeling.

How is the spatially regulated activation of Rho GTPases achieved? Because the GTPases themselves are diffusely distributed, one possible mechanism is through the targeting of GEFs and GAPs to dendritic spines by adaptor proteins. Both GIT1 and the polarity protein PAR-3 seem to function in this capacity. In each case, proper synaptic targeting of a RacGEF (βPIX for GIT1, and TIAM1 for PAR-3) by these proteins is essential for dendritic spine morphogenesis. Mislocalized GEFs cause aberrant Rac activation and the formation of multiple dendritic protrusions at the expense of mature spines. Thus, spatially restricted Rac activation is essential for normal spine development.

Downstream Actin-Regulating Proteins Implicated in Spine Morphogenesis

As previously stated, the Rho family of small GTPases eventually targets a variety of actin-binding proteins that directly regulate polymerization, depolymerization, branching, and bundling of the actin filaments. These proteins include profilin, ADF/cofilin, the Arp2/3 complex, capping protein, and myosin. These core proteins work together with their binding partners such as cortactin, WASP and WAVE proteins, Ena/VASP, and the lipid messenger PI(4,5)P_2. Several of these actin regulators have now been implicated in spine morphogenesis.

Profilin is a small protein that allows the elongation of actin filaments. In nonneuronal cells, profilin has been shown to stabilize the actin network. In cultured hippocampal neurons, profilin II shows increased localization to dendritic spines in response to NMDA receptor activation, which coincides with actin filament stabilization in the spines and inhibition of spine motility. In addition, fear conditioning drives profilin into spines *in vivo*. Taken together, these studies suggest that profilin is involved in activity-dependent synaptic plasticity that underlies memory formation.

In contrast to profilin, ADF/cofilin functions to depolymerize and destabilize actin filaments. Cofilin is usually enriched at sites of high motility and rapid actin reorganization. Recently, cofilin has been shown to localize to the peripheral part of the dendritic spines and rarely to the spine core, supporting the observation that the spine surface is the more dynamic part that undergoes actin-dependent shape changes. LIMK1, which phosphorylates and inactivates ADF/cofilin, is important for spine morphogenesis. LIMK1-knockout mice show abnormal spines, which have a thick neck and a small head as opposed to the normal mushroom-shaped spines. In addition, these mice show impaired spatial learning. Interestingly, LIMK1 mutations are associated with William's syndrome in humans, which is characterized by visuospatial cognitive deficits.

Cortactin is an F-actin binding protein that localizes to the peripheral membrane protrusions in nonneuronal cells. This localization is dependent on Rac activation. In addition to binding actin, cortactin also binds the Arp2/3 complex and stimulates its actin nucleating activity. In hippocampal neurons, cortactin localizes to dendritic spines, where it plays a role in spine morphogenesis. Knockdown of cortactin decreases spine density, whereas overexpression of cortactin results in longer spines. Cortactin also shows activity-dependent translocation in and out of spines, indicating a role in synaptic plasticity. However, whether these effects are dependent on Rho GTPases is not yet known.

Recently, WAVE1 was shown to regulate dendritic spine morphology through phosphorylation by the cyclin-dependent kinase 5 (Cdk5). Cdk5

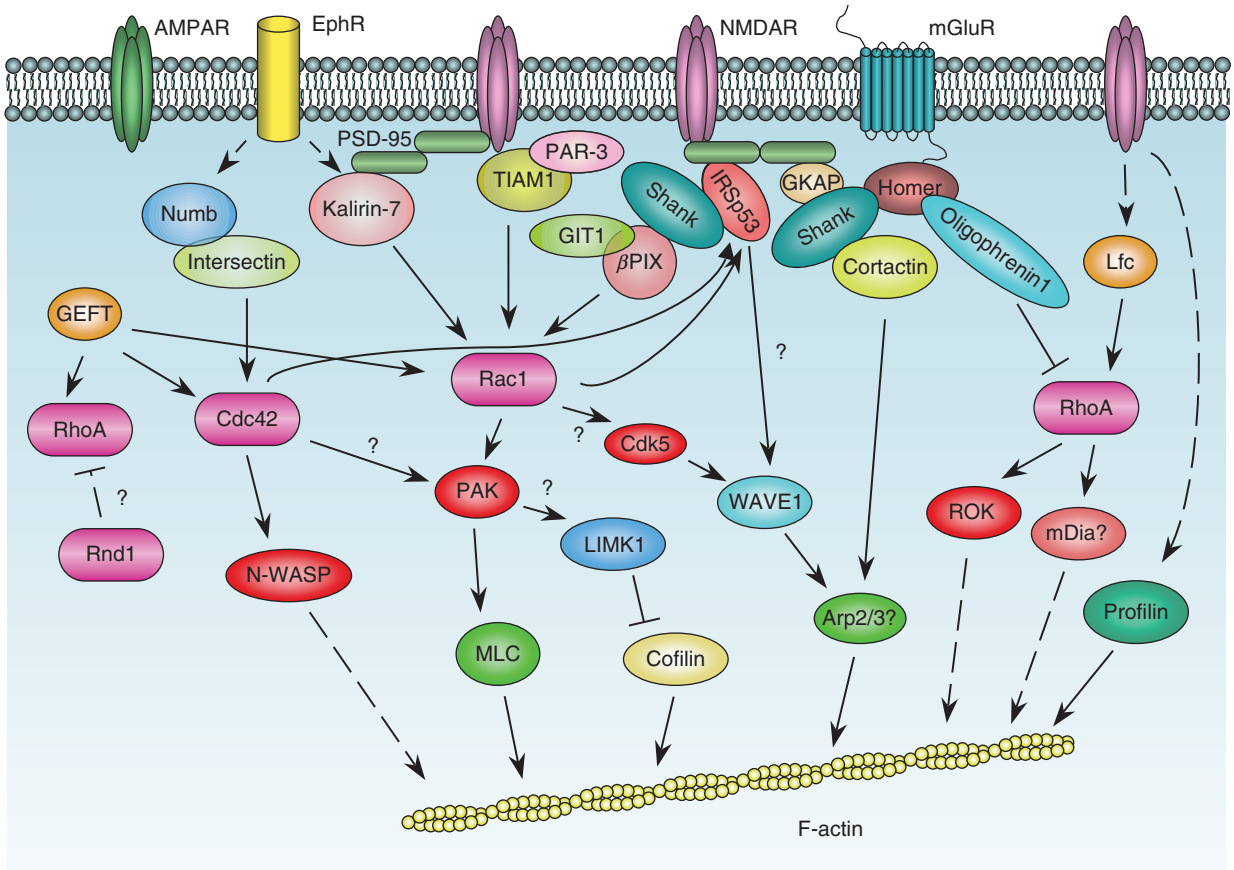

Figure 3 Rho GTPase signaling in dendritic spine morphogenesis. Known Rho signaling pathways involved in spine morphogenesis are illustrated. Dashed lines indicate indirect effects. Question marks indicate proteins or pathways that are known to be involved in regulating actin dynamics in other systems, but have not been directly demonstrated in the context of spine morphogenesis. AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; Arp2/3, actin-related protein 2/3; Cdc42, Rac1, Rnd1, RhoA, Rho GTPases; Cdk5, cyclin-dependent kinase 5; EphR, Eph receptor; GIT1, G-protein-coupled receptor kinase interacting protein 1; GKAP, guanylate kinase-associated protein; IRSp53, insulin receptor substrate of 53 kDa; Lfc, Lbc (lymphoid blast crisis)'s first cousin; LIMK1, LIM kinase 1; mDia, mammalian ortholog of *Drosophila* Diaphanous; mGluR, metabotropic glutamate receptor; MLC, myosin light chain; NMDAR, *N*-methyl-D-aspartate receptor; N-WASP, neuronal Wiscot–Aldrich syndrome protein; PAK, p21-activated kinase; PAR-3, partitioning defective 3; PIX, PAK interacting exchange factor; PSD-95, postsynaptic density protein 95; ROK, Rho kinase; Shank, SH3 and ankyrin repeat-containing protein; TIAM1, tumor invasion and metastasis 1; WAVE1, WASP family verprolin-homologous protein 1.

phosphorylation of WAVE1 inhibits its ability to activate Arp2/3-mediated actin polymerization. Loss of WAVE1 results in a decrease in mature dendritic spines and a concomitant increase in filopodia-like protrusions. This effect is rescued by a nonphosphorylatable mutant of WAVE1 but not by a pseudophosphorylated mutant. Interestingly, Cdk5 was shown before to function as a neuronal-specific effector of Rac1. This provides a potential pathway by which Rac1 activation leads to filopodia-like dendritic protrusions in neurons.

Finally, myosin II, which binds and contracts actin filaments, is important for maintaining dendritic spine morphology. Inhibiting myosin II using blebbistatin disrupts spine formation in immature neurons and induces a loss of mushroom-shaped spines and the formation of filopodia-like protrusions in mature neurons. Similarly, knockdown of endogenous myosin IIb induces the formation of filopodia-like protrusions at the expense of mushroom-shaped spines. PAK and myosin light chain kinase (MLCK), both downstream of Rac and Cdc42, have been shown to function through phosphorylating myosin light chain (MLC) in regulating spine formation. In addition, Rho kinase, a major regulator of MLC phosphorylation, may also play a role. If this is the case, it places myosin as one of the convergent points downstream of Rho GTPases signaling in spine morphogenesis (**Figure 3**).

Rho GTPases and Mental Retardation

Mental retardation (MR) is a developmental disorder that is characterized by cognitive defects. On a cellular level, MR is associated with abnormalities in dendritic spine morphology, raising the possibility that mutations in proteins regulating spine morphology could cause MR. Indeed, three out of seven recently cloned genes that cause nonsyndromic MR (i.e., MR that is not associated with gross developmental brain defects) are directly involved in Rho GTPase signaling. These genes encode oligophrenin1, a Rho GAP; αPIX, a GEF for Rac and Cdc42; and PAK3, an effector for Rac and Cdc42. The fact that these proteins participate in different aspects of Rho-family signaling points to an important role for Rho GTPases in cognitive functions. Interestingly, both oligophrenin-1 and PAK3 have been shown to be involved in spine morphogenesis. In addition, βPIX, another isoform of PIX, is also essential for spine formation.

Several other proteins on the Rho pathway have also been linked to human genetic disorders that show mental retardation. For example, the FMRP, whose gene mutation is responsible for fragile-X syndrome, binds Sra-1/CYFIP, an effector for Rac. In addition, mutations of faciogenital dysplasia protein (FGD1), a GEF for Cdc42, cause Aarskog–Scott syndrome. Future studies will no doubt bring the Rho family of GTPases to the center stage of spine morphogenesis as well as cognitive functions such as learning and memory.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Dendrite Development, Synapse Formation and Elimination; Dendritic Spine History; Eph Receptor Signaling and Spine Morphology; Glutamate Regulation of Dendritic Spine Form and Function; LIM Kinase and Actin Regulation of Spines; Postsynaptic Specialization Assembly; Spines and Mental Disorders.

Further Reading

Govek EE, Newey SE, Akerman CJ, et al. (2004) The X-linked mental retardation protein oligophrenin-1 is required for dendritic spine morphogenesis. *Nature Neuroscience* 7: 364–372.

Govek EE, Newey SE, and Van Aelst L (2005) The role of the Rho GTPases in neuronal development. *Genes and Development* 19: 1–49.

Ishikawa Y, Katoh H, and Negishi M (2003) A role of Rnd1 GTPase in dendritic spine formation in hippocampal neurons. *Journal of Neuroscience* 23: 11065–11072.

Penzes P, Johnson RC, Sattler R, et al. (2001) The neuronal Rho-GEF Kalirin-7 interacts with PDZ domain-containing proteins and regulates dendritic morphogenesis. *Neuron* 29: 229–242.

Ramakers GJ (2002) Rho proteins, mental retardation and the cellular basis of cognition. *Trends in Neurosciences* 25: 191–199.

Ryan XP, Alldritt J, Svenningsson P, et al. (2005) The Rho-specific GEF Lfc interacts with neurabin and spinophilin to regulate dendritic spine morphology. *Neuron* 47: 85–100.

Tada T and Sheng M (2006) Molecular mechanisms of dendritic spine morphogenesis. *Current Opinion in Neurobiology* 16: 95–101.

Tolias KF, Bikoff JB, Burette A, et al. (2005) The Rac1-GEF Tiam1 couples the NMDA receptor to the activity-dependent development of dendritic arbors and spines. *Neuron* 45: 525–538.

Zhang H and Macara IG (2006) The polarity protein PAR-3 and TIAM1 cooperate in dendritic spine morphogenesis. *Nature Cell Biology* 8: 227–237.

Zhang H, Webb DJ, Asmussen H, Niu S, and Horwitz AF (2005) A GIT1/PIX/Rac/PAK signaling module regulates spine morphogenesis and synapse formation through MLC. *Journal of Neuroscience* 25: 3379–3388.

Rhodopsin

T P Sakmar and T Huber, The Rockefeller University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Rhodopsin is the primary photoreceptor molecule of vision. Light striking a rhodopsin molecule in a photoreceptor cell of the retina is converted into a biochemical signal by a photochemical reaction. The signal is processed by other cells in the retina and sent to the brain. Two general types of photoreceptor cells exist in the vertebrate retina and are named according to their characteristic shapes: rods and cones. Rod cells are responsible for scotopic or dim-light vision, whereas cone cells are responsible for photopic or bright-light and color vision in vertebrates. Rod cells contain rhodopsin; cone cells contain photopsins, also known as cone pigments or color vision pigments. The photoreceptor cells of invertebrate species also contain rhodopsin, even though generally the anatomy of the invertebrate visual system varies dramatically from that of the vertebrate system.

The outer segment of the rod photoreceptor cell contains a stack of several hundred disks (Figure 1). These disks are derived from plasma membrane invaginations at the region close to the connecting cilium. The stack of disks is enclosed by the plasma membrane. The disk membranes are densely packed with rhodopsin. Cytoplasmic dynein mediates the vectorial transport of opsin-containing post-Golgi vesicles along microtubuli toward the distal region of the inner segment. The apical tip of the outer segment is constantly phagocytized by the adjacent retinal pigment epithelium (RPE), limiting the lifetime of rhodopsin to several days in the vertebrate retina.

Rhodopsin is a seven transmembrane (7TM) domain protein homologous to other members of the G-protein-coupled receptor (GPCR) family. Rhodopsin consists of the complex of a rod-type opsin with the chromophore ligand 11-*cis*-retinal acting as an inverse agonist. Photoisomerization of the chromophore converts the ligand in the all-*trans* form acting as the receptor agonist. The photoactivated receptor triggers a highly amplified signal transduction cascade. The dark-adapted rod cell is partially depolarized, causing steady release of neurotransmitters on the synaptic terminal. Light activation leads to hyperpolarization and reduction of neurotransmitter release.

The visual cycle (Figure 2) describes the retinoid pathway, including the light reaction in the photoreceptor cell and the reisomerization of vitamin A in the RPE. All visual pigments consist of an opsin protein plus a chromophore. The chromophore is a cofactor, which is linked covalently via a Schiff base bond to a specific lysine amino acid residue of the opsin protein. One of two retinoids serves as the chromophore for nearly all visual pigments. The chromophore in most vertebrate pigments is the aldehyde of vitamin A_1, 11-*cis*-retinal. The chromophore in some fishes and amphibians is the aldehyde of vitamin A_2, 11-*cis*-3,4-didehydroretinal, which contains an additional carbon–carbon double bond in the β-ionone ring. Photoisomerization of the 11-*cis* to all-*trans* form of the chromophore is the primary event in visual signal transduction, and it is the only light-dependent step. The absorption of a photon, a quantum of energy, by rhodopsin causes the photoisomerization. The wavelength of light that is maximally absorbed (λ_{max} value) is 500 nm (Figure 3). Although rhodopsin can absorb a fairly wide range of light energies, green light is most effective in activating rhodopsin due to the higher absorbance or photon capture probability. The physiological response to a single photoisomerization is independent of the wavelength of the photon causing the reaction. Upon photoisomerization of the pigment to the all-*trans*-retinylidene form, the receptor switches from the inactive receptor state (R), Meta I with a protonated Schiff base having a peak with a λ_{max} value of 480 nm, to the active receptor state (R*), Meta II with a deprotonated Schiff base and a characteristic λ_{max} value of 380 nm. The activating Meta I to Meta II transition depends on environmental factors such as temperature, proton activity, and the lipid environment. The R* is the active form of the receptor, which is able to transmit a biochemical signal as described later.

Rhodopsin is found primarily within the disk membrane of the rod cell (see Figure 1). The N-terminus of rhodopsin is in the disk lumen, or extracellular, whereas the C-terminus is intracellular, or cytoplasmic. The protein spans the membrane bilayer seven times, and the resulting 7TM helices of rhodopsin fold to form a pocket to accommodate the retinal chromophore (Figure 4). The chromophore-binding pocket resides in the membrane-embedded domain of the protein. The lysine residue that acts as the linkage site for the chromophore is conserved within the seventh TM segment in all visual pigments. In addition, a pair of highly conserved cysteine residues is found on the extracellular surface and may form a stabilizing disulfide bond. In many pigments, a carboxylic acid residue, which acts as the counter-ion to the positively charged protonated Schiff base, is conserved within the third TM segment. A cluster of serine and threonine residues is found in the C-terminal tails of most visual pigments. These sites become phosphorylated

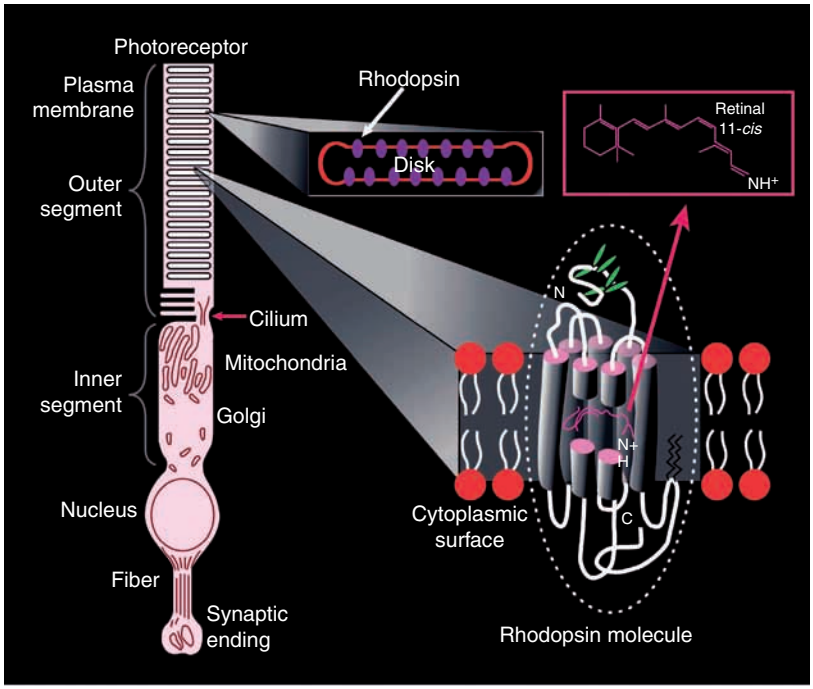

Figure 1 A schematic representation of bovine rhodopsin within the bilayer of the disk membrane. The seven transmembrane helices form a binding pocket for the retinal chromophore. The rod cell consists of an inner segment and an outer segment. The outer segment contains a stack of disk membranes, which are highly enriched in rhodopsin. Other proteins of the biochemical cascade of the vertebrate visual system, including transducin, reside in the cytoplasmic space.

in response to light and provide a mechanism to attenuate the light response because a phosphorylated rhodopsin is prevented from transmitting its biochemical signal.

Visual pigments, including rhodopsin, are members of a large family of GPCRs. Bovine rhodopsin was the first GPCR to be sequenced by amino acid sequencing, the first to be cloned, and the first for which a high-resolution crystal structure has been reported. Bovine rhodopsin is 348 amino acid residues in length and is identical to human rhodopsin at all but 23 positions. The cloning of a β-adrenergic hormone receptor led to the identification of the structural homologies that now define the superfamily of GPCRs. Molecular cloning has identified hundreds of different but related GPCRs. It has been estimated that the human genome contains 865 genes encoding GPCRs. All GPCRs reside in cell membranes and consist of seven hydrophobic membrane-spanning domains. Thus, GPCRs may also be referred to as 7TM receptors. All GPCRs are involved in the transmission of biochemical signals across cell membranes, which includes sensory processes such as vision or olfaction and responses to hormones. Some receptors responsible for synaptic transmission of a neural signal, or for modulation of synaptic signaling, are also members of the GPCR family.

GPCRs become activated either by light in the case of rhodopsin or by binding to an appropriate agonist ligand such as a neurotransmitter or hormone. The activated receptor interacts with another class of proteins, called guanine nucleotide-binding regulatory proteins, or G-proteins. The G-protein transduces the biochemical signal to some cellular effector molecule, which ultimately affects cellular function, growth, or metabolism. Examples of cellular effector enzymes include adenylate cyclase, which produces cyclic adenosine monophosphate, and phospholipase C, which produces phosphatidyl inositides. Some G-proteins can interact directly with ion channels in the cell membrane. Transducin is the G-protein that interacts with the active Meta II photoproduct of rhodopsin. When it interacts with Meta II, transducin takes up a molecule of guanosine 5′-triphosphate (GTP), which allows it to activate a cellular phosphodiesterase that lowers the levels of cyclic guanosine monophosphate (cGMP), leading to a closing of cGMP-gated cation channels. The resulting hyperpolarization of the rod cell initiates the visual signal.

The molecular cloning of opsins from a variety of species has allowed detailed comparisons of visual pigments based on their structural, spectral, and biochemical properties. Pigment genes have been classified phylogenetically and models of pigment evolution have been proposed. The homology in the opsin family of genes indicates that divergent evolution occurred from a single precursor retinal-binding protein to form long- and short-wavelength absorbing prototypes. The long-wavelength prototype diverged

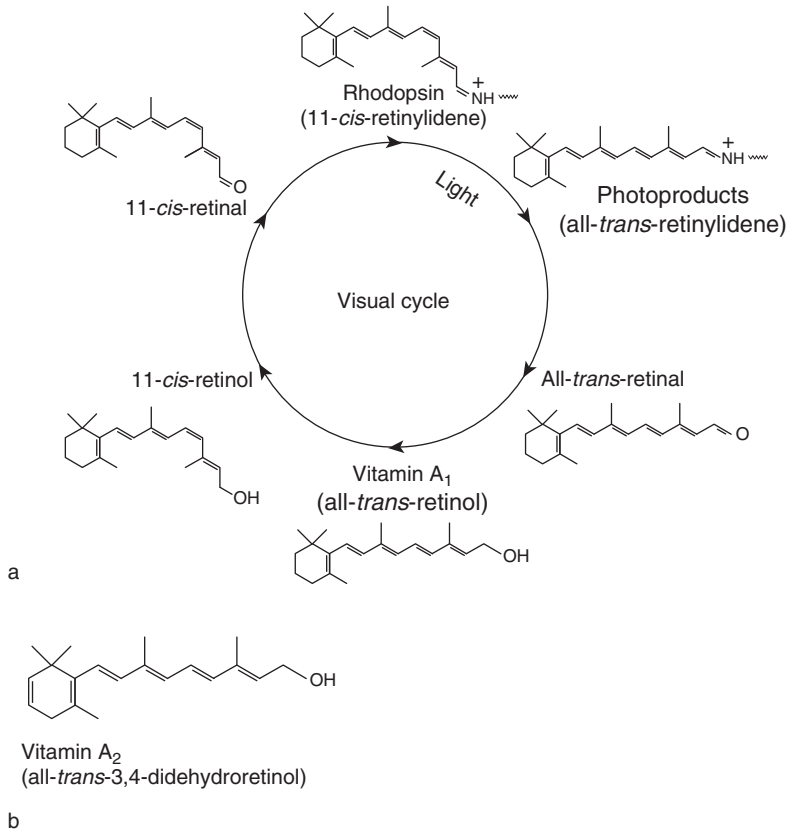

Figure 2 Photoisomerization of 11-*cis*-retinal to all-*trans*-retinal is the only light-dependent event in vision. The chromophore is covalently linked as a cofactor (11-*cis*-retinylidene) to a specific opsin lysine amino acid residue via a protonated Schiff–base bond. A salt bridge with a carboxylate counter-ion locks the receptor in the ground state. A single visible photon isomerizes the chromophore to the all-*trans* form with high quantum efficiency (~0.67). Hydrolysis of the photoproducts results in all-*trans*-retinal and the apoprotein opsin. The photopigment is regenerated from opsin and 11-*cis*-retinal, which is recycled in the dark reactions of the visual cycle (a) from all-*trans*-retinal. Vertebrate visual pigments contain one of two chromophores, 11-*cis*-retinal (vitamin A_1 aldehyde) or 11-*cis*-3,4-didehydroretinal (vitamin A_2 aldehyde), which contains an additional carbon–carbon double bond in the β-ionone ring. Typically, retinoids enter the cycle as all-*trans*-retinol (vitamin A_1) or as all-*trans*-3,4-didehydroretinol (vitamin A_2; shown in b).

to form red and green pigments. The short-wavelength prototype then diverged to form a blue pigment and the family of rhodopsins and rhodopsin-like green pigments.

Spectral Tuning of Rhodopsin and Cone Visual Pigments

Humans possess trichromatic color vision, or trichromacy. Most people can match any given reference color by combining the three primary colors. The three primary colors for additive color mixtures are red, green, and blue. In addition to rod cells, which contain rhodopsin, the retina contains three classes of cone photoreceptor cells. Each class of cone cell is sensitive to a specific wavelength of light. At the molecular level, human trichromatic color vision requires the presence of three cone pigments with broad overlapping spectral absorption. Each specific cone class contains only one type of photoreceptor molecule. The three types of cone photoreceptor molecules (red, green, and blue) are homolog of rhodopsin. The amino acid sequences of these opsins are approximately 40% identical to that of human rhodopsin. The green and red opsins are approximately 96% identical to each other and approximately 43% identical to the blue opsin. The spectral properties of human cone pigments have been studied by a variety of techniques, including psychophysical color matching and microspectrophotometry. In addition, the molecular cloning of the cone pigment genes has allowed the study of the molecular mechanism of spectral tuning and the molecular genetics of color vision.

Rhodopsin and the three cone pigments are each tuned to a characteristic λ_{max} but they all use the same chromophore, 11-*cis*-retinal. Despite the fact that retinal is the universal chromophore, the λ_{max} values of visual pigments must span the visible spectrum, which ranges from approximately 360 nm (ultraviolet) to 600 nm (deep red). Thus, a single chromophore allows all visible wavelengths of light to be detected. A spectral tuning mechanism exists so that

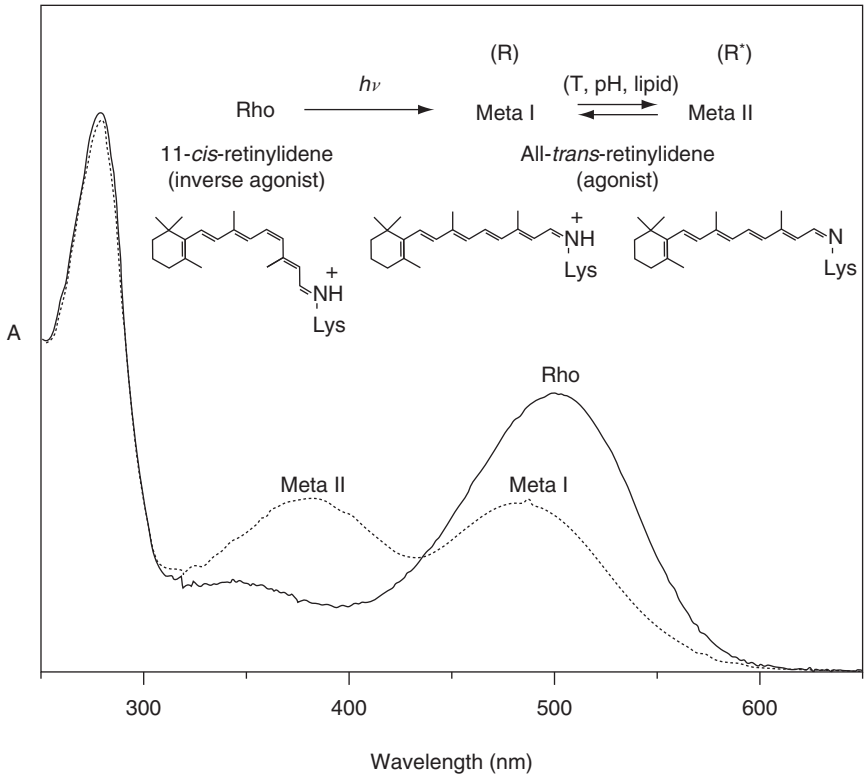

Figure 3 An ultraviolet-visible absorption spectrum of purified bovine rhodopsin shows a characteristic broad visible absorbance with a λ_{max} value of 500 nm. The 280 nm peak represents the opsin protein component. The 11-*cis*-retinylidene chromophore acts as an inverse agonist, effectively turning off the receptor. After exposure to light, *cis*-to-*trans* isomerization results in all-*trans*-retinylidene chromophore in the pigment. The all-*trans*-retinylidene acts as an agonist, triggering the transition from the inactive receptor state (R), Meta I with a protonated Schiff base having a peak with a λ_{max} value of 480 nm, to the active receptor state (R*), Meta II with a deprotonated Schiff base and a characteristic λ_{max} value of 380 nm. The R* form interacts with the rod cell heterotrimeric G-protein transducin, leading to signal transduction, and with rhodopsin kinase, leading to receptor phosphorylation and subsequent visual arrestin binding that terminates signaling.

a particular opsin protein can modulate the absorption spectrum of its retinylidene chromophore. Spectral tuning is possible because specific amino acid side chains of each opsin can interact with the chromophore and shift its λ_{max} value. For example, the human green and red opsin proteins display only 15 differences among their 364 amino acid residues. Of these 15 differences, seven amino acid changes are responsible for the observed 30 nm spectral shift in going from the λ_{max} value of the green to that of the red pigment. Most of this shift is caused by three amino acid side chains that contain hydroxyl groups: tyrosine, serine, and threonine.

Genetic variations in color vision may result from a mutation in one of the genes encoding a cone opsin. If the amino acid change affects spectral tuning in one of the cone pigments, then one of the cone types will contain a pigment with an anomalous absorption spectrum. Individuals with this genetic variation are called anomalous trichromats. The most common type of anomalous trichromacy is called red–green 'color blindness.' Red–green color vision variations predominantly affect males because the genes for the red and green opsin genes are found on the X chromosome. Severe color vision defects may result from complete deficiencies of one or more of the three cone opsin genes. Individuals who lack a functional red or green photoreceptor are called dichromats.

The Biochemical Cascade of Vertebrate Vision

Photoisomerization of retinal activates rhodopsin, allowing it to interact with a specific G-protein. The rod cell G-protein, transducin, transmits a biochemical signal from rhodopsin to a cellular effector molecule. Photoactivated rhodopsin catalyzes the exchange of guanosine 5′-diphosphate (GDP) for GTP by transducin molecules. A single activated rhodopsin can activate catalytically up to 500 transducin molecules. The activated transducin, with GTP in its nucleotide-binding pocket, then transmits a chemical signal to the effector molecule in the signaling cascade. The effector molecule in the vertebrate visual system is the enzyme cGMP phosphodiesterase. A single transducin molecule activates a single phosphodiesterase

molecule. However, each phosphodiesterase can catalyze the hydrolysis of up to 3000 cGMP molecules to GMP molecules per second. Thus, the biochemical amplification is as much as six orders of magnitude at the molecular level. The lowered levels of cGMP in the rod cell cause plasma membrane cation channels to close. A single photon causing retinal isomerization can affect the conductance of approximately 1000 plasma membrane cation channels.

Because the plasma membrane is selectively permeable to ions, which are electrically charged, an electrical potential difference exists between the inside and the outside of the rod cell. The potential increases as cation channels are closed and the influx of sodium and calcium ions, which carry positive charge, is slowed. The result is that the rod cell becomes hyperpolarized in response to light. The increase in potential varies proportionally with the strength of the light signal. The change in membrane potential is sent as an electrical signal from the plasma membrane to the synaptic terminal of the rod cell, where it is transmitted to other specialized cells of the retina.

A properly dark-adapted rod cell can detect a single photon. This extreme sensitivity is possible in part because of two important features of the visual system: (1) the stability of rhodopsin in darkness to thermal activation and (2) the high degree of signal amplification by the biochemical cascade of vision. It has been estimated that the spontaneous thermal isomerization of 11-*cis*-retinal in rhodopsin in darkness only takes place roughly once every 300–1000 years. In contrast (as presented previously), a single photon can theoretically cause as many as 1000 cation channels in a rod cell membrane to close.

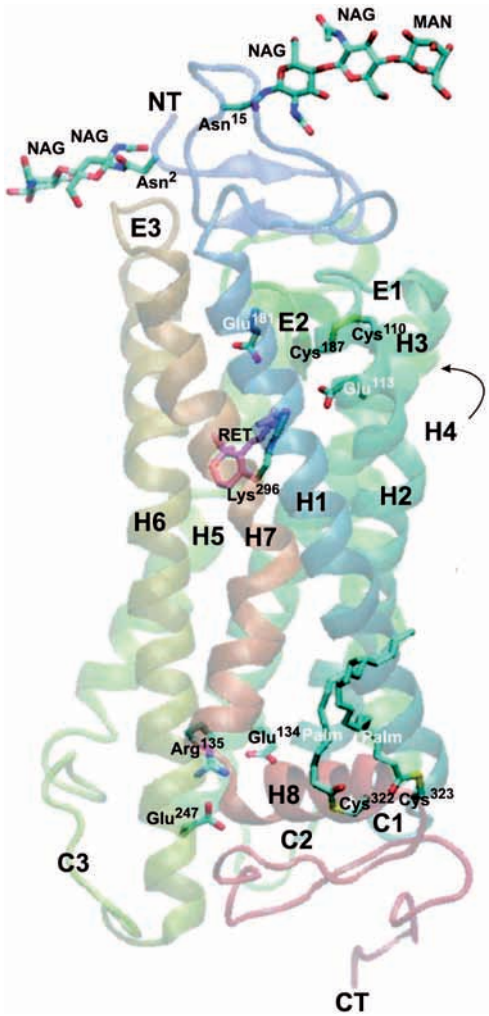

Figure 4 A molecular graphics ribbon diagram of bovine rhodopsin prepared from the 2.2 Å crystal structure coordinates (PDB access number 1U19). The N-terminus (NT) and extracellular (or intradiscal) surface is toward the top of the figure, and the C-terminus (CT) and intracellular (or cytoplasmic) surface is toward the bottom. Seven TM (H1–H7), which are characteristic of G-protein-coupled receptors (GPCRs; also known as 7TM receptors), are shown. The TM segments are tilted with respect to the plane of the membrane bilayer. They are generally α-helical, but they contain significant kinks and irregularities. Three extracellular (E1–E3) and three cytoplasmic (C1–C3) loops connect the TM helices. Posttranslational modifications are the N-glycosylations on asparagines 2 and 15 (shown as several sugar residues: NAG, *N*-acetyl-glucosaminyl; and MAN, mannosyl), the palmitoyl thioesters at cysteines 322 and 323, the disulfide bridge between cysteines 110 and 187, and the 11-*cis*-retinylidene chromophore (RET) at lysine 296. The glutamate residues 113 and 181 play a central role in the ligand-dependent activation switch after photoisomerization that is relayed to the cytoplasmic switching microdomain, between glutamate 134 and arginine 135, in a highly conserved amino acid sequence motif, and glutamic acid 247 at the cytoplasmic end of helix H6.

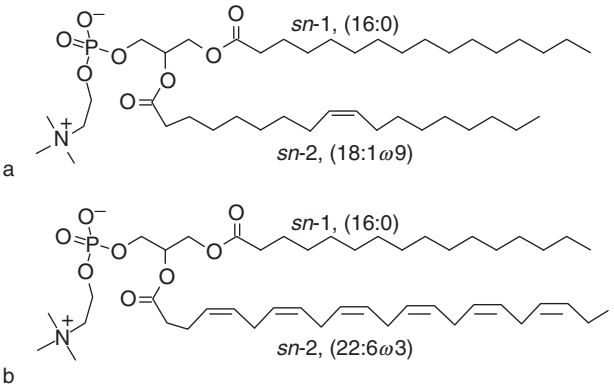

Figure 5 Photoreceptor membrane phospholipids are unusually highly enriched in ω-3 polyunsaturated fatty acids, especially docosahexaenoic acid (DHA). Structures of (a) the monounsaturated phospholipid 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC), and (b) the polyunsaturated phospholipid 1-palmitoyl-2-docosahexaenoyl-*sn*-glycero-3-phosphocholine (PDPC), which contains DHA. Lipid–protein interactions with rhodopsin provide a model to understand why essential fatty acids of the ω-3 series are important for retinal and neural development and other key biological functions.

The light signal is turned off by a number of biochemical mechanisms. Light-activated rhodopsin becomes phosphorylated by a specific rhodopsin kinase enzyme. Visual arrestin binds to the phosphorylated form of the receptor that can no longer interact with the G-protein transducin. The active GTP-bound form of transducin has an intrinsic enzymatic GTP hydrolysis activity. Over time, the bound GTP is converted to GDP, and transducin returns to its inactive state. In addition, intracellular calcium levels drop after the cation channel of the rod cell closes. A fall in intracellular calcium concentration mediates photoreceptor cell recovery and adaptation. The process of adaptation allows the sensitivity of the retina to adjust based on the level of light in a visual scene. Adaptation is extremely important for a useful visual

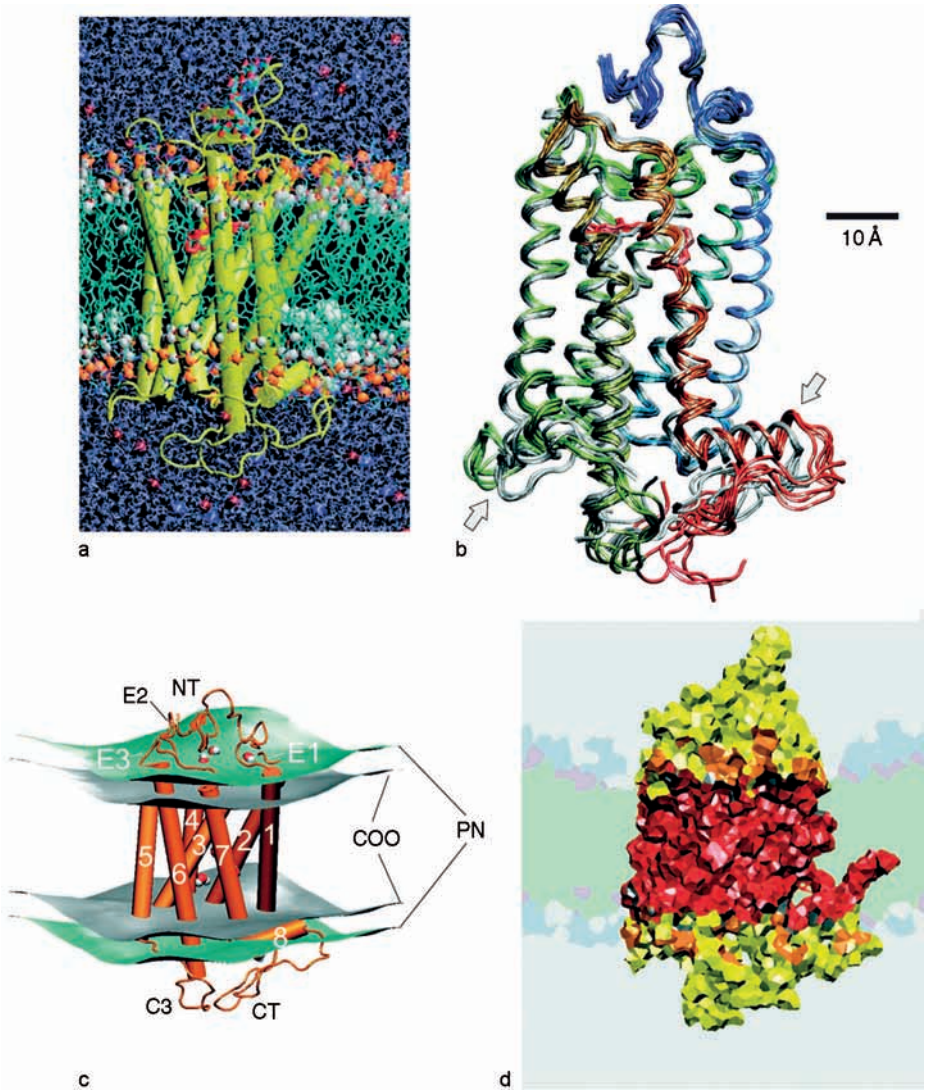

Figure 6 Biomembrane model of rhodopsin in a POPC bilayer from molecular dynamics (MD) simulations. (a) Solvation of rhodopsin by the phospholipid bilayer and aqueous salt solution. Thermal motions on the nanosecond timescale are accessible by numerical simulations on supercomputers. Disordered acyl chains (cyan) are found in the fluid membrane interior. The seven transmembrane (7TM) helices span the membrane from both polar head group regions (orange). (b) Representative protein backbone conformations from the MD simulations (colored tubes) are overlaid with several crystal structures (gray). The arrows indicate conformational changes observed in the bilayer compared with the crystal structures. The TM region exhibits very small fluctuations compared with the cytoplasmic loop regions, shown in the lower part of the structure. (c) Approximate lipid–water interfaces corresponding to the hydrophobic region (gray) and head groups (cyan) of the bilayer. All helices, except some cytoplasmic parts of helices H3, H6, and H8, are confined by the lipid head group–water interfaces. (d) Protein interfaces analyzed by a novel generalized molecular surface method based on generalized Voronoi volume regions associated with aqueous salt solution (light gray), with the glycerophosphocholine head groups (light blue), the acyl chain ester group atoms (light purple), and the acyl chain hydrocarbon atoms (light green) of the POPC bilayer. The rhodopsin interfaces with the phospholipid acyl chain hydrocarbon atoms (red), ester groups (gray), glycerophosphocholine head groups (orange), and the aqueous solvent (yellow) are color-coded.

system because the magnitude of light varies widely in the environment. For example, bright sunlight and dim starlight differ in luminance by at least several orders of magnitude.

Lipid–Protein Interactions of Rhodopsin

Photoreceptor membrane phospholipids (Figure 5) are unusually highly enriched in ω-3 polyunsaturated fatty acids, especially docosahexaenoic acid. Lipid–protein interactions with rhodopsin provide a model to understand why essential fatty acids of the ω-3 series are important for retinal and neural development and other key biological functions. Molecular dynamics simulation of a computational biomembrane model (Figure 6) gives a detailed description of the solvation of rhodopsin by the fluid phospholipid bilayer and the hydration shell. Thermally excited motions on the nanosecond timescale result in relaxation of the protein in the membrane relative to the crystal structure. The lipid–water interface adapts to the protein and vice versa. The solvation interface analysis mapped to a secondary structure diagram of rhodopsin (Figure 7) demonstrates how the helices and loops are coupled to the polar head group interface of the bilayer membrane. The results of such coupling are, for example, hydrophobic matching phenomena involved in lateral organization of biomembranes.

Rhodopsin as a Model for G-Protein-Coupled Receptors

The human genome contains nine opsin genes, one rhodopsin and three cone opsins, as well as encephalopsin, melanopsin, peropsin, retinal G-protein-coupled receptor, and neuropsin. In contrast to these opsins, different photoreceptive proteins are found in archaea, bacteria, fungi, and a green alga, *Chlamydomonas reinhardtii*. They function as light-driven ion pumps or light sensors. Despite their 7TM fold, there is no evidence that these nonanimal opsins are structurally related to rhodopsin. GPCRs provide arguably the most important drug target family for the pharmaceutical industry. However, high-resolution crystal structure data are limited to rhodopsin and computational

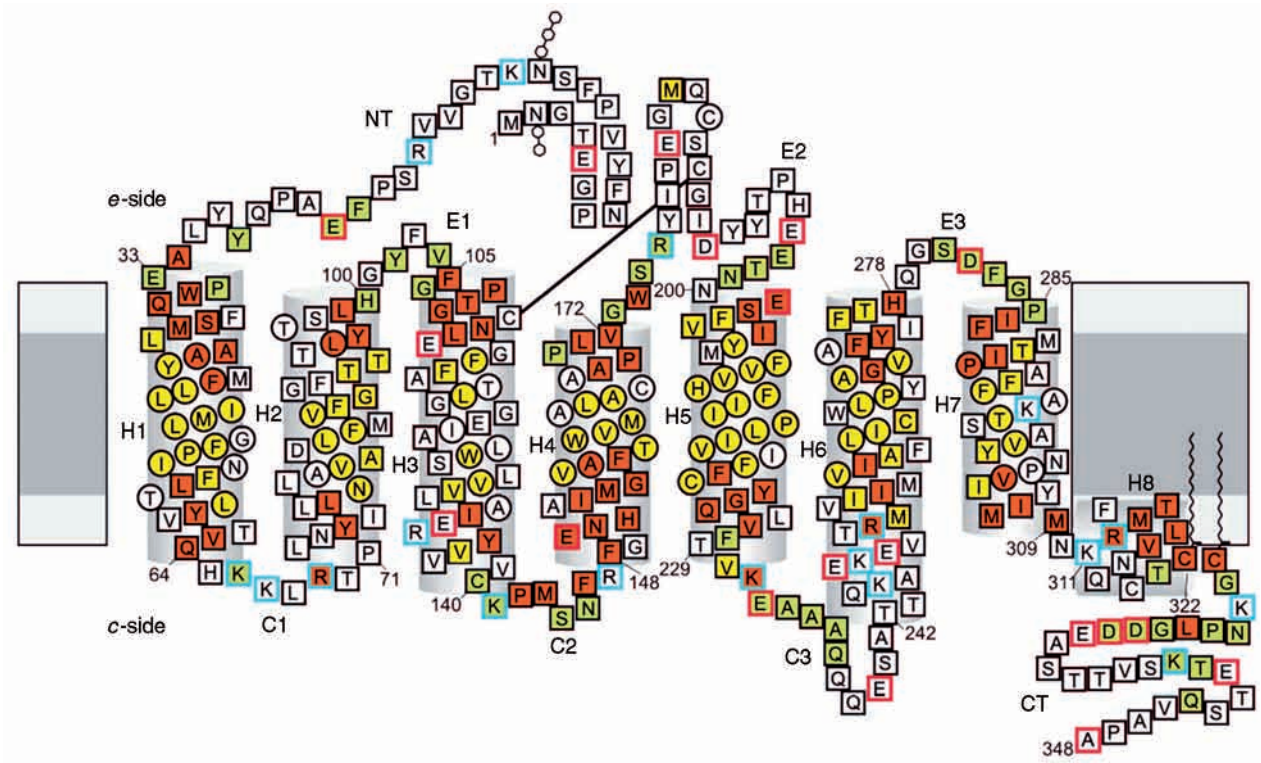

Figure 7 A secondary structure diagram of bovine rhodopsin. Amino acid residues are depicted in single-letter code. The N-terminal tail and extracellular domain is toward the top and the C-terminal tail and cytoplasmic domain is toward the bottom. TM α-helical segments (H1–H7) and the cationic amphipathic helix H8 are shown in cylinders. The topology scheme of rhodopsin illustrates the protein residues associated with the different generalized molecular surface interfaces. Residues exposed to the bilayer hydrocarbon interior (the lipid–protein hydrophobic interface; yellow), head groups (the lipid–protein polar interface; green), and both simultaneously (orange) are color-coded. Squares indicate protein residues exposed to internal or external water. The charge states of the residues in the molecular dynamics model are designated by red or blue symbol frames.

modeling of other GPCRs is mainly based on homology to rhodopsin. These models are useful in the context of computer-aided drug design methods, such as docking, in combination with site-directed mutagenesis and high-throughput screening. Despite progress in crystallization of eukaryotic membrane proteins, it is expected that computational modeling approaches will remain the main source for GPCR structures in the foreseeable future.

Mutations in the Gene for Human Rhodopsin May Cause Disease

Retinitis pigmentosa is a group of hereditary progressive blinding diseases with variable clinical presentations. One form of the disease, autosomal dominant retinitis pigmentosa (ADRP), was linked to a mutation in the gene for rhodopsin. More than 70 different rhodopsin gene mutations have been reported in patients with ADRP. The mutations reported would result in alterations in all domains of rhodopsin: extracellular, membrane-embedded, and cytoplasmic. ADRP is characterized by a progressive death of rod and, sometimes, cone cells, resulting in gradual vision loss. The molecular pathophysiology of ADRP, namely how a defect in rhodopsin leads to rod cell death, remains to be fully elucidated.

See also: Color Vision; Cone Photopigment Evolution; Photoreceptor Adaptation; Photoreceptors: Physiology; Phototransduction; Retinitis Pigmentosa; Vision: Light and Dark Adaptation.

Further Reading

Ballesteros JA, Shi L, and Javitch JA (2001) Structural mimicry in G protein-coupled receptors: Implications of the high-resolution structure of rhodopsin for structure–function analysis of rhodopsin-like receptors. *Molecular Pharmacology* 60: 1–19.

Dratz EA and Hargrave PA (1983) The structure of rhodopsin and the rod outer segment disk membrane. *Trends in Biochemical Sciences* 8: 128–131.

Dryja TP, McGee TL, Reichel E, et al. (1990) A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. *Nature* 343: 364–366.

Ebrey T and Koutalos Y (2001) Vertebrate photoreceptors. *Progress in Retinal and Eye Research* 20: 49–94.

Fanelli F and De Benedetti PG (2005) Computational modeling approaches to structure–function analysis of G protein-coupled receptors. *Chemical Review* 105: 3297–3351.

Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, and Vriend G (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Research* 31: 294–297.

Huber T, Botelho AV, Beyer K, and Brown MF (2004) Membrane model for the G-protein-coupled receptor rhodopsin: Hydrophobic interface and dynamical structure. *Biophysical Journal* 84: 2078–2100.

Kochendoerfer GG, Lin SW, Sakmar TP, and Mathies RA (1999) How color visual pigments are tuned. *Trends in Biochemical Sciences* 24: 300–305.

Kukura P, McCamant DW, Yoon S, Wandschneider DB, and Mathies RA (2005) Structural observation of the primary isomerization in vision with femtosecond-stimulated Raman. *Science* 310: 1006–1009.

Menon ST, Han M, and Sakmar TP (2001) Rhodopsin: Structural basis of molecular physiology. *Physiological Reviews* 81: 1659–1688.

Nathans J (1989) The genes for color vision. *Scientific American* 260: 42–49.

Nathans J, Thomas D, and Hogness DS (1986) Molecular genetics of human color vision: The genes encoding blue, green, and red pigments. *Science* 232: 193–202.

Palczewski K, Kumasaka T, Hori T, et al. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* 289: 739–745.

Rando RR (2001) The biochemistry of the visual cycle. *Chemistry Review* 101: 1881–1896.

Stryer L (1987) The molecules of visual excitation. *Scientific American* 257: 42–50.

Terakita A (2005) The opsins. *Genome Biology* 6: 213.1–213.9.

Wald G (1968) The molecular basis of visual excitation. *Nature* 219: 800–807.

Yan ECY, Kazmi MA, Ganim Z, et al. (2003) Retinal counterion switch in the photoactivation of the G protein-coupled receptor rhodopsin. *Proceedings of the National Academy of Sciences of the United States of America* 100: 9262–9267.

Ribbon Synapses

B Odermatt and L Lagnado, MRC Laboratory of Molecular Biology, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Ribbon synapses are sites of contact between neurons specialized for the rapid transmission of signals by the calcium-triggered secretion of neurotransmitter. Although sharing many fundamental properties with chemical synapses throughout the nervous system, ribbon synapses possess several features that set them apart as being specialized for the transmission of sensory information. The defining morphological difference is the presence of an unusual organelle, the ribbon, which holds synaptic vesicles close to the active zone, the specialized region where vesicles fuse with the surface membrane (Figure 1). Functionally, neurons that transmit through ribbon synapses are distinguished by the nature of the electrical signal controlling neurotransmitter release. Conventional synapses convey information arriving as a 'pulse code'; action potentials lasting a few milliseconds trigger a transient burst of vesicle fusion and the amplitude of the action potential is fixed. In contrast, ribbon synapses transmit 'analog signals,' such as the intensity of light, over short conduction distances; the amplitude of the electrical signal varies continuously with the stimulus and so does the rate of vesicle fusion.

The use of ribbon synapses to transfer graded voltage signals is a key feature of information transfer in the first stages of vision and hearing. Ribbons occur in photoreceptors and bipolar cells in the retina and in mechanosensitive hair cells in the cochlea. Ribbon synapses also occur in the hair cells of the vestibular organs of balance and in neurons in the pineal gland that secrete melatonin in a circadian rhythm. In fish and amphibia, electroreceptors of the lateral line organ also transmit signals through ribbon synapses.

Why are graded voltage signals and ribbon synapses employed by sensory neurons? The answer probably lies in the fact that the maintenance of a high rate of continuous vesicle fusion, and the modulation of this rate by the stimulus, allows information to be transferred more rapidly than the pulse code of neurons firing action potentials. Theoretical studies indicate that spiking neurons with conventional synapses at the input stages of senses such as vision and hearing would limit the information that could be subsequently extracted by the neural circuitry into which they feed. The way in which the ribbon contributes to the maintenance of neurotransmitter release is not clearly defined; however, given that it holds a population of vesicles close to the active zone, it seems likely that it transports vesicles to sites of fusion during ongoing activity, perhaps in the manner of a conveyor belt.

The Structure of Ribbons

The term 'ribbon' was originally used to describe unusual structures observed in the synaptic pedicle of cone photoreceptors by making transmission electron micrographs of thin sections of the retina – a long osmiophiolic band extending to the surface membrane holding synaptic vesicles close to the active zone (Figures 1(a) and 1(b)). In three dimensions, the ribbon is typically a plate approximately 30–40 nm thick with a pentalaminar structure (Figure 1(c)). The shape and size of the plate varies between different species and between different photoreceptors within the same retina, as well as varying diurnally. Typically, the plate is 200–1000 nm tall and 200–2000 nm long. In rod and cone photoreceptors, the ribbon floats just above the plasma membrane on a second electron-dense structure, the arciform density, which in turn attaches to the underlying membrane. The arciform density is only 20 nm thick, so vesicles attached to the base of the ribbon also attach to the surface membrane.

The structure of ribbons is not fixed: In some cells, they exhibit a striking structural plasticity. In photoreceptors, for instance, ribbons become smaller in the light and then larger within 10 min of returning to the dark. Similarly, in mammalian pinealocytes, the size and number of ribbons is lower during the day through a mechanism of diurnal regulation. In retinal bipolar cells and hair cells of the ear, the ribbons tend to be significantly smaller than those in photoreceptors, and the arciform density is absent. In fact, the ribbons of hair cells have a different shape: They resemble a sphere and are also known as 'dense bodies' or 'synaptic bodies' (Figure 2). The functional implications of ribbons of different sizes and/or shapes have not been studied systematically, but it seems likely that they reflect adaptations for changes in synaptic strength or speed operating under different operating conditions of sensory neurons.

At all ribbon synapses, the synaptic vesicles are clear and small (30–45 nm in diameter) and contain glutamate as the principal transmitter. They attach to the ribbon all over its surface by multiple short filaments approximately 5 nm thick and 20–40 nm long (Figures 1 and 2). The packing of vesicles is

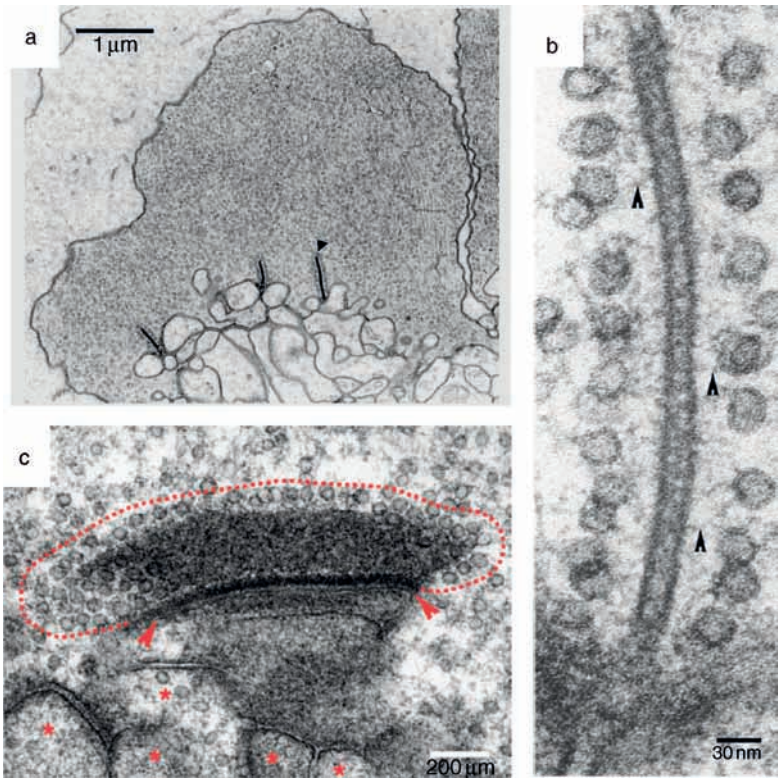

Figure 1 Structure of synaptic ribbons in the retina. (a) Electron micrograph of the synaptic terminal of a cone photoreceptor from the retina of a turtle showing three ribbons (one is marked with an arrowhead). (b) The ribbon from a bipolar cell terminal. The short filaments connecting vesicles to the ribbon are marked. (c) 'Side-on' view of a ribbon in a cone terminal showing the dense and regular arrangement of attached vesicles (within the dotted line). The dark ridge attaching the ribbon to the plasma membrane is marked by arrowheads and the postsynaptic elements are marked by asterisks. (a) Reproduced from Schaeffer SF and Raviola E (1978) Membrane recycling in the cone cell endings of the turtle retina. *Journal of Cell Biology* 79: 802–825. (b, c) Reproduced from Sterling P and Matthews G (2005) Structure and function of ribbon synapses. *Trends in Neurosciences* 28: 20–29, with permission from Elsevier.

often in a hexagonal array with a density up to 55%. A consistent feature of the synaptic terminal of these sensory neurons is that vesicles are found not just at the ribbon but also distributed through large volumes of the cytoplasm at a relatively high density. At conventional synapses, a reserve pool of vesicles can also be seen to congregate behind the active zone, but these do not number more than approximately 100 and they are not associated with a structure as large and well defined as the ribbon or dense body. It seems to be a general rule that the larger the synaptic terminal, the greater the number of ribbons at active zones, although these vary in size. For instance, rod photoreceptors in mammals typically contain just one large ribbon up to 2 µm in length with approximately 800 vesicles attached, whereas cones contain 10 or more smaller ribbons, each with less than 400 vesicles. An individual bipolar cell might contain 50 ribbons, although these will only be 200–300 nm in length with less than 100 vesicles attached. Hair cells are effectively one large synaptic terminal with small vesicles widely distributed throughout the cell and 10–20 ribbons attached to the basal end.

The Voltage Signal Controlling Transmitter Release

Rod and cone photoreceptors convert light into an electrical signal through the same basic mechanism: Absorption of light by the visual pigment in the outer segment activates a transduction cascade that results in the closure of cation-selective ion channels and hyperpolarization of the plasma membrane. At the synaptic terminal, this hyperpolarization leads to the closure of voltage-sensitive calcium channels and a reduction in the rate of vesicle fusion. Glutamate is therefore released at the maximum rate in the dark, and light causes a decrease that is graded with intensity. In rods, which are the more sensitive of the two classes of photoreceptor, the gain of this process is sufficient to signal the absorption of a single photon with a change in membrane potential of a fraction of a millivolt. The ribbon synapse can reliably transmit this small signal to postsynaptic bipolar cells and horizontal cells.

The ribbon in a rod or cone terminal is at the apex of a ridge forming a structure called a triad, in which

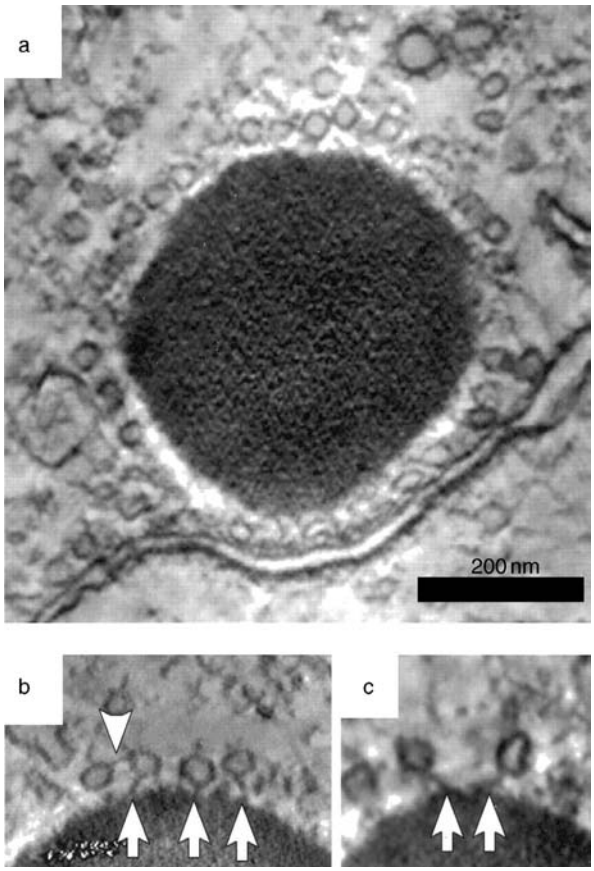

Figure 2 Structure of synaptic ribbons in hair cells. (a) A slice through the afferent synapse of a hair cell from the frog sacculus which has been reconstructed by electron tomography. Vesicles are attached to the ribbon (also known as the synaptic body) and are also docked to the plasma membrane immediately below. (b, c) Examples of the tethers linking vesicles to the ribbon. Adapted from Lenzi D, Runyeon JW, Crum J, Ellisman MH, and Roberts WM (1999) Synaptic vesicle populations in saccular hair cells reconstructed by electron tomography. *Journal of Neuroscience* 19: 119–132, with permission from Society for Neuroscience.

there are three postsynaptic elements: the processes of two horizontal cells laterally and the dendrite of a bipolar cell centrally (Figure 3(a)). The postsynaptic glutamate receptors on horizontal cells are of the AMPA type, but different classes of bipolar cells express different receptors, which confers on them electrical responses of different polarities. The bipolar cells receiving inputs from rods possess a metabotropic glutamate receptor (mGluR6) that leads to the closure of cation channels on binding glutamate. By decreasing glutamate release from rods, light causes the depolarization of these bipolar cells, which are also termed 'ON.' The polarity of the response in 'OFF' bipolar cells is inverted (i.e., they hyperpolarize to light) because these have AMPA and/or kainate receptors that pass cations when glutamate is released by photoreceptors in the dark. The organization of postsynaptic structures contacted by a bipolar cell terminal is similar, except that there are just two processes contacted by each ribbon, forming a dyad (Figure 3(b)). These two processes may be from amacrine cells (containing kainate receptors), the dendrites of ganglion cells (containing AMPA and NMDA receptors), or there may be one process from each cell type. The amacrine cells often make reciprocal synaptic connections back onto the synaptic terminal of the bipolar cell, and these release the inhibitory neurotransmitters GABA or glycine to modulate synaptic transfer from bipolar cell to ganglion cell.

In hair cells, deflection of the hair bundle toward its tall edge by a mechanical stimulus leads to the opening of cation channels and membrane depolarization. Conversely, displacement away from the kinocilium closes these channels and hyperpolarizes the membrane. Sound waves therefore cause oscillatory changes in membrane potential, the amplitude of

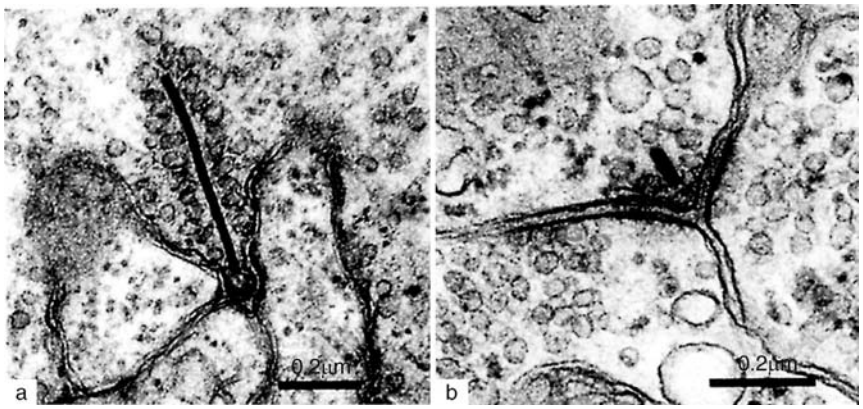

Figure 3 Electron micrographs of ribbon synapses in the frog retina. (a) A triad formed by a photoreceptor terminal and (b) a dyad formed by a bipolar cell onto two postsynaptic processes. Note that the photoreceptor ribbon is much taller than that in the bipolar cell. Adapted from Dowling (1968) Synaptic organization of the frog retina: An electron microscopic analysis comparing the retinas of frogs and primates. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 170: 205–228, with permission from Royal Society.

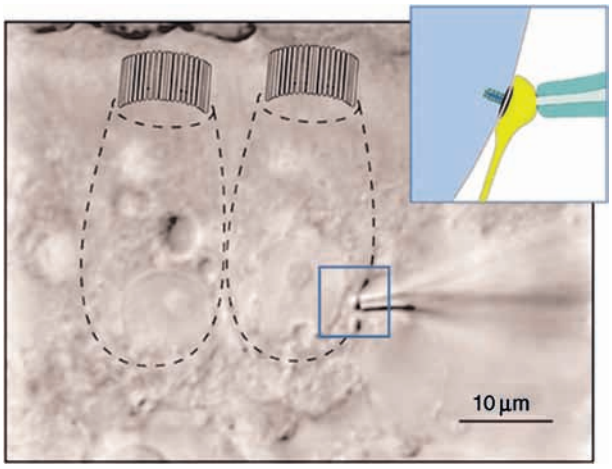

Figure 4 Rat inner ear hair cells contacted by afferent fibers. Two afferent boutons can be seen within the box at the base of the hair cell. One bouton has a glass pipette attached for electrophysiological recording. (Inset) Most inner hear hair cell synapses are served by a single synaptic ribbon. Reproduced from Glowatzki E and Fuchs P (2002) Transmitter release at the hair cell ribbon synapse. *Nature Neuroscience* 5: 147–154, with permission from Nature Publishing Group.

which varies with the loudness of the sound. In the mammalian cochlea, one inner hair cell might contact 10–30 afferent fibers that project into the central nervous system (Figure 4). In most cases, each fiber is contacted by just one synaptic ribbon, so the properties of the active zone directly determine the auditory signal that is transmitted. As in ribbon synapses in the retina, changes in membrane potential of a fraction of a millivolt cause significant modulation of the rate of vesicle fusion. A second key feature of the ribbon synapse of auditory hair cells is its ability to transmit signals at high frequency: Action potentials in the afferent fiber can be evoked at frequencies up to 300 Hz, whereas individual action potentials can remain phase-locked to a tone at frequencies up to 5 kHz.

Transient and Continuous Modes of Exocytosis

The fact that photoreceptors hyperpolarize to light, thereby closing calcium channels, suggests that they are capable of continuously releasing transmitter in the dark; otherwise, the light signal would not be transmitted. Evidence for this idea is shown in Figure 5(a), which is an intracellular recording from a horizontal cell in which the membrane potential indicates the rate of glutamate release by photoreceptors. Light flashes hyperpolarize the horizontal cell. Excess magnesium, which inhibits calcium-dependent exocytosis at the synapse, mimics the effect of light and blocks the response to a flash. The relatively depolarized potential of the horizontal cell in the dark must therefore be maintained by continuous transmitter release from photoreceptors in the dark, supported by a continuous cycle of exocytosis and endocytosis.

This vesicle cycle has been visualized in the terminals of rod-driven ('ON') bipolar cells from the retina of goldfish which have an unusually large synaptic terminal and, because of this practical advantage, have been intensively studied (Figure 5(b)). If the membrane dye FM1-43 is added in the absence of calcium, only the cell surface fluoresces because the dye cannot cross the membrane (Figure 5(b), b), but when the cell is depolarized in the presence of calcium, there is an immediate increase in fluorescence in the terminal as vesicles fuse with the surface and take up the dye. With time, there is a gradual accumulation of stained membrane retrieved from the surface by endocytosis until all the vesicles in the terminal carry the dye and the fluorescence reaches a maximum (Figure 5(b), d). Because exocytosis and endocytosis are balanced over timescales of tens of seconds, there is no net change in the surface area of the terminal.

The continuous mode of exocytosis is most obvious in ribbon synapses in the retina, and it is suited to signaling the intensity and duration of maintained stimuli. However, in some cases, ribbon synapses must signal rapid change, such as the onset of light or the frequency of a sound, and this is best done by a transient burst of neurotransmitter release similar to that which might be triggered by an action potential at a conventional synapse. Ribbon synapses can therefore also support short bursts of very rapid vesicle fusion. An electrophysiological method that allows exocytosis and endocytosis to be monitored with good time resolution is the capacitance technique, which measures the increase in surface area when vesicles fuse and the subsequent decline in area when this excess membrane is retrieved again. The vesicles docked to the plasma membrane under the ribbon can be released in a fraction of a millisecond after the calcium channels are opened, at rates up to many thousands of vesicles per second (Figure 6). Such a rapid response is expected in hair cells that transmit signals phase-locked to sounds of kilohertz frequencies, but it is also observed in retinal bipolar cells.

A key specialization enabling rapid modulation of exocytosis is found in the calcium channels that mediate calcium influx to trigger vesicle fusion. In all ribbon synapses, these channels are of the L-type, which are activated at relatively high voltages (beginning at approximately −42 mV) and which have very fast opening and closing kinetics. These calcium channels are tightly clustered in the active zone under the ribbon, very close to the docked vesicles. Over the voltage range that calcium channels are just

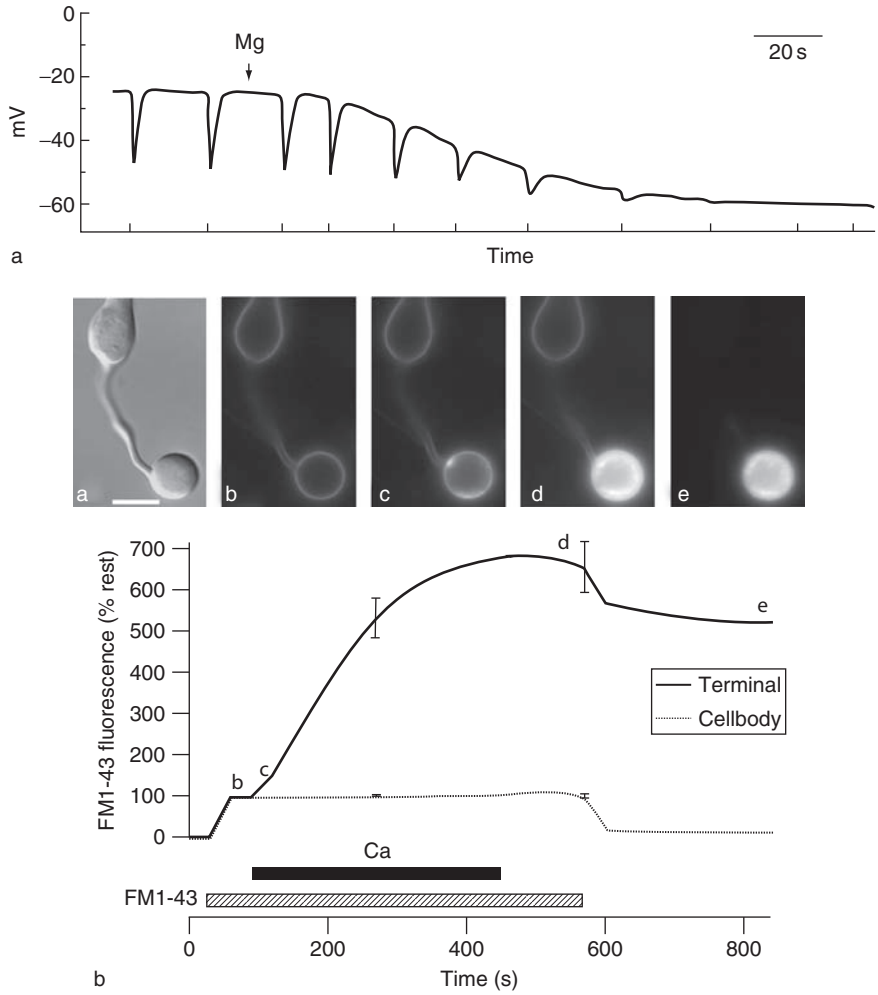

Figure 5 Continuous exocytosis. (a) The effect of magnesium on the membrane potential of a horizontal cell from the skate retina. Tick marks show the timing of flash stimuli, which cause a transient hyperpolarization. The downward arrow indicates the time at which high magnesium was applied to inhibit transmitter release, and this mimicked the effect of light to hyperpolarize the cell and gradually abolish the light response. The relatively depolarized membrane potential in the dark was therefore caused by the maintained release of neurotransmitter from photoreceptors. (b) A depolarizing bipolar cell isolated from the goldfish retina with a giant synaptic terminal (a). The fluorescence images in (b–e) show the distribution of the membrane dye FM1-43, first added in the absence of calcium (b) and then during maintained stimulation of exocytosis (c, d). When the dye is washed off the cell surface, the membrane compartments that have been retrieved by endocytosis contain the dye and can be visualized (e). The graph quantifies the increase in fluorescence in the terminal during the experiment. (a) Adapted from Dowling JE and Ripps H (1973) Effect of magnesium on horizontal cell activity in the skate retina. *Nature* 242: 101–103, with permission from Elsevier. (b) Adapted from Lagnado L, Gomis A, and Job C (1996) Continuous vesicle cycling in the synaptic terminal of retinal bipolar cells. *Neuron* 17: 957–967, with permission from Elsevier.

beginning to open, the amplitude of the calcium current is a very steep function of membrane potential so that minor changes in voltage cause major changes in the amount of glutamate released according to a power function. This voltage dependence does not appear to reflect a cooperative action of calcium in triggering release because the amount of exocytosis measured over timescales of hundreds of milliseconds has a linear dependence on calcium influx in both photoreceptors and hair cells. This fundamental property of the release apparatus is very different in conventional synapses, in which exocytosis is triggered by a highly cooperative action of calcium (Hill coefficients of 3 or 4). It has been suggested that the proportionality between vesicle release and calcium influx in hair cells contributes to the encoding of auditory stimuli by allowing the ribbon synapse to represent minor differences in sound intensity over a wide dynamic range, as well as accurately transferring variations in intensity with minimum distortion.

Closer analysis of signals in retinal bipolar cells demonstrates that there are two distinct populations of vesicles that can be released rapidly: fusion of docked vesicles is followed by release of a second group which is approximately four times the size of the docked pool and is depleted in approximately 1 s

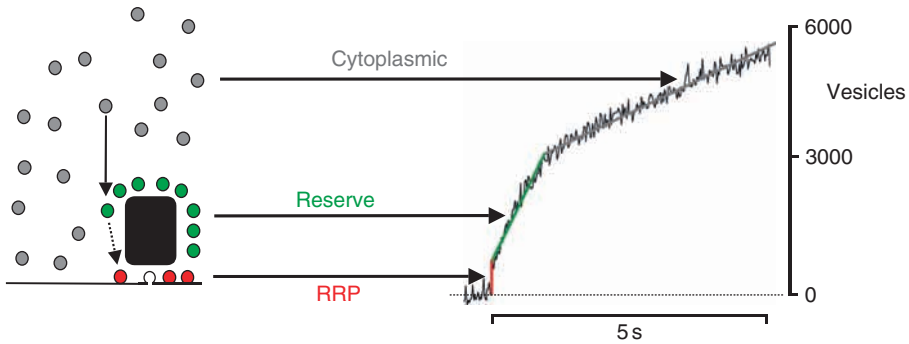

Figure 6 Vesicle pools at the ribbon synapse of bipolar cells. The right-hand side shows the time course of exocytosis assayed by using FM1-43 to stain exocytosed vesicles: Three distinct phases are observed. The first group of vesicles to be released, the rapidly released pool (RRP), is thought to correspond to vesicles docked at the active zone, which are shown in red in the schematic diagram on the left. The second group of reserve vesicles shown in green correspond numerically with the vesicles attached to the ribbon projecting into the cytoplasm. The third and slowest phase of release is continuous and is supported by a population of vesicles distributed throughout the cytoplasm. Modified from Neves G and Lagnado L (1999) The kinetics of exocytosis and endocytosis in the synaptic terminal of goldfish retinal bipolar cells. *Journal of Physiology* 515: 181–202, with permission from Blackwell Publishing.

(Figure 6). The 1:4 ratio between the first and second phases of release correlates with the observation that approximately four times as many vesicles are attached to the portion of the ribbon projecting into the cytoplasm behind the active zone as those docked at the base of the ribbon. If the ribbon is indeed a conveyor belt transporting vesicles to release sites at the active zone, then these measurements indicate that it takes an average of 250 ms for a vesicle to be transferred from the ribbon to a release site and then fuse with the surface membrane. The third and continuous phase of exocytosis that follows is supported by a much larger pool of approximately 500 000 vesicles distributed throughout the cytoplasm (Figure 6).

Efficient Recycling of Synaptic Vesicles

Although synaptic terminals containing ribbons at the active zone tend to be large and contain many vesicles, the maintenance of transmitter release over the long term requires that vesicles be recycled within the terminal. At the ribbon synapse of bipolar cells, most, if not all, the vesicles fusing with the surface membrane collapse and merge, and there is no evidence that vesicles can connect transiently with the surface through a fusion pore that allows the release of transmitter ('kiss-and-run'). These collapsed vesicles are retrieved by at least three distinct modes of endocytosis. Electron microscopy demonstrates that chronic and strong stimulation of exocytosis in cones, bipolar cells, and hair cells is followed by bulk retrieval of large chunks of surface membrane, equivalent to several vesicles' worth. This process generates cisternae within the terminal, from which vesicles subsequently bud off. Capacitance measurements of higher time resolution show that briefer stimuli in bipolar cells and hair cells are followed by two other modes of endocytosis that are kinetically distinct. Fast endocytosis occurs in 1 s or less, whereas slow endocytosis occurs on a timescale of 10–20 s.

In bipolar cells, fast and slow modes of retrieval both take in membrane chunks similar to the size of synaptic vesicles, but by mechanisms that are molecularly distinct. Slow retrieval occurs by the formation of Clathrin-coated vesicles from the cell membrane and is similar to the process that internalizes membrane receptors in all types of cells. In contrast, fast retrieval is independent of Clathrin, but the molecular mechanisms of the process remain to be identified. The relative importance of fast and slow endocytosis depends on how strongly the synapse is stimulated: In bipolar cells, weak stimuli that only release docked vesicles are followed almost exclusively by fast retrieval, but as more vesicles are released at higher rates, an increasing proportion are retrieved by the slower, Clathrin-dependent process. The two mechanisms of vesicle retrieval are also distinguished by their dependence on calcium: fast endocytosis appears to be triggered by calcium influx at the active zone, whereas slow endocytosis does not appear to be directly calcium dependent.

Once vesicles have been retrieved from the cell surface, they must somehow find their way back to the ribbon and/or the active zone. At many conventional synapses, the cluster of vesicles held in reserve behind the active zone is restricted in movement because the vesicles are tethered to the actin cytoskeleton via the vesicle-associated protein Synapsin. In hippocampal synapses, for instance, only a fraction of vesicles are mobile, and these have an effective diffusion coefficient of approximately $3 \times 10^{-3}\,\mu m^2\,s^{-1}$.

However, vesicles in the synaptic terminals of cones and bipolar cells are almost all mobile, moving in a random fashion with an effective diffusion coefficient of approximately $5 \times 10^{-2}\,\mu m^2\,s^{-1}$. In the giant synaptic terminal of bipolar cells from goldfish, the cytoplasmic reservoir contains approximately 300 000 vesicles, and this degree of mobility would result in approximately 900 vesicle collisions s^{-1} against the ribbons and 28 000 collisions s^{-1} against the surface membrane. In comparison, continuous exocytosis can be maintained at rates of 500–1000 vesicles s^{-1}. The random motion of cytoplasmic vesicles therefore seems sufficient to maintain an efficient supply to the ribbon and plasma membrane during ongoing activity.

Molecular Aspects of Ribbon Synapse Structure and Function

The synaptic vesicle cycle is regulated by a large number of specialized proteins, many of which occur in the membrane of synaptic vesicles or at the active zone. The vast majority of these proteins are common to both ribbon and conventional synapses, but there are some significant divergences, presumably reflecting different specializations for maintained or phasic release of neurotransmitter. The most obvious difference is the ribbon itself, of which the major structural component is Ribeye (which is also the only specific component identified to date). Ribeye is the product of an alternative promoter in the *CtBP2* gene and consists of an A domain specific for ribbons – presumably responsible for its multimerization – and a B domain that is identical to CtBP2, a transcriptional repressor that in turn is related to 2-hydroxyacid dehydrogenases. Other ribbon-associated proteins, such as Piccolo, CtBP1, and RIM1, are thought to be involved in tethering synaptic vesicles to the ribbon and/or priming the vesicles for fusion at the active zone (Figure 7(a)). KIF3A, a kinesin motor protein, is also attached to retinal ribbons, providing some support for the conveyor belt model, in which primed vesicles are actively transported along the ribbon to the release sites at the plasma membrane.

In photoreceptors, the arciform density is connected to the ribbon via Bassoon, a large scaffolding protein that binds Ribeye and also anchors to the cell cytoskeleton. In a mouse deficient of Bassoon, there are free-floating unanchored ribbons in photoreceptor hair cells and fast synaptic transmission is severely reduced. Surprisingly, a very similar phenotype in cone photoreceptors could be found in the zebra fish *nrc* mutant, which is deficient for synaptojanin 1, a polyphosphoinositide phosphatase important for Clathrin-mediated endocytosis and actin cytoskeletal rearrangements at conventional synapses. Bassoon is

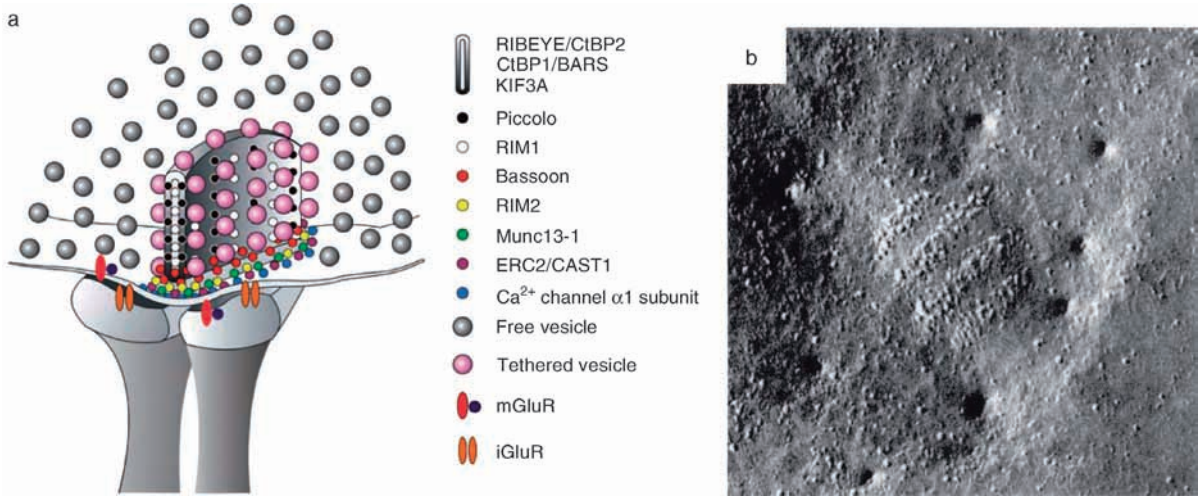

Figure 7 The distribution of different molecular components of a ribbon synapse. (a) The differential distribution of proteins defines two presynaptic compartments at a retinal ribbon synapse. The ribbon-associated compartment includes RIBEYE/CtBP2, CtBP1/BARS, KIF3A, Piccolo, and RIM1. The active zone compartment includes RIM2, Munc13-1, ERC2/CAST1, and an L-type calcium channel α_1 subunit. Bassoon localizes at the border between the two compartments. Postsynaptic glutamate receptors may be metabotropic (mGluR) or ionotropic (iGluR). (b) Freeze-fracture electron micrograph showing the array of particles in the membrane at the active zone of a hair cell from the frog sacculus. These particles are arranged in four distinct stripes and are a mixture of calcium channels (approximately 90) and calcium-activated potassium channels (approximately 40). (a) Reproduced from tom Dieck S and Brandstatter JH (2006) Ribbon synapses of the retina. *Cell Tissue Research* 326: 339–346, with permission from Springer Science and Business Media. (b) Reproduced from Roberts WM, Jacobs RA, and Hudspeth AJ (1990) Colocalization of ion channels involved in frequency selectivity and synaptic transmission at presynaptic active zones of hair cells. *Journal of Neuroscience* 10: 3664–3684, with permission from Society for Neuroscience.

not found in bipolar cell terminals, but all ribbon synapses contain Piccolo, which is similar to Bassoon in being a multidomain protein with a structural role at the active zone. Other proteins at the arciform density are RIM2, Munc13-1, and CAST1, which bind to each other and which probably play a role in docking and priming of vesicles before fusion.

The calcium signal that triggers exocytosis in ribbon synapses is generated by clusters of voltage-gated calcium channels at the active zone (**Figure 7(b)**). These channels are of the L-type ($Ca_V1.4$ in the retina and $Ca_V1.3$ in hair cells), which is an important distinction from the calcium channels at the active zone of spiking neurons, which are of the N-, P/Q-, or R-type. The activation of L-type calcium channels is inhibited by a number of neuromodulators within the retina, including adenosine, dopamine, and cannabinoids, all of which act through G-protein-coupled receptors on the surface of the terminal to alter the voltage dependence of exocytosis. A mutation in the pore-forming unit of the $Ca_V1.4$ channel produces congenital stationary night blindness, and these patients show disrupted transmission between photoreceptors and bipolar cells together with abnormal development of the synapse.

The SNARE complex, which mediates vesicle fusion with the plasma membrane, is very similar at ribbon and conventional synapses. In both cases, the v-SNARE on the synaptic vesicle is VAMP/Synaptobrevin 2, which associates with the t-SNARES on the plasma membrane – Syntaxins 1, 2, or 3 – together with SNAP-25. The SNAP receptor is modulated by Complexins; however, whereas conventional synapses contain Complexins I and II, ribbon synapses contain Complexins III and IV, implicating this molecule in the functional differences. It seems likely that the calcium sensor for fusion at all types of synapses is Synaptotagmin, a transmembrane protein in the vesicle of which there are many isoforms. Most conventional synapses contain Synaptotagmins I and II, which are also found in retinal ribbon synapses but not in hair cells, where Synaptotagmin VII is more usually found. Although it has been suggested that different Synaptotagmin isoforms with different affinities for Ca^{2+} ions may underlie variations in the functional properties of synapses, this idea remains to be tested. Calcium concentrations of 1 μM or less can trigger exocytosis at ribbon synapses of photoreceptors and bipolar cells, and this signal occurs on the spatial scale of a micrometer, being generated by the cooperative opening of many calcium channels distributed over the active zone (**Figure 7(b)**). Other proteins acting as calcium sensors for exocytosis may be the Ferlins, which are similar to Synaptotagmins. One of these, Otoferlin, has been suggested to be the major Ca^{2+} sensor triggering exocytosis in inner hair cells. Otoferlin localizes to synaptic vesicles at the ribbon, and on binding Ca^{2+} it interacts with Syntaxin 1 and SNAP-25. Mice deficient in Otoferlin are deaf, and exocytosis from inner hair cells is abolished. Otoferlin is also defective in a recessive form of human deafness.

Does the Ribbon Act as a Conveyor Belt?

Ribbon synapses possess several design features that allow the maintenance of continuous exocytosis in response to graded signals: large terminals with large numbers of vesicles and active zones, efficient retrieval of vesicles by endocytosis to maintain the cytoplasmic reservoir, and the rapid and unhindered movement of new vesicles to the ribbon and/or release sites at the active zone. The fact that the ribbon holds vesicles immediately behind docking sites at the active zone suggests that it aids this supply function by acting as a trap for vesicles moving randomly in the cytoplasm, like a piece of fly paper. Do captured vesicles move in a directed way from the ribbon to release sites on the surface membrane? The most direct evidence for the idea that the ribbon might act like a conveyor belt comes from work using total internal reflection fluorescence microscopy to image individual vesicles in the synaptic terminal of retinal bipolar cells. This approach demonstrates that active zones of exocytosis coincide with structures that hold vesicles just behind the plasma membrane until stimulation causes the transfer of these vesicles downwards. These structures are most likely to be ribbons holding vesicles 'in reserve' behind the active zone. The time required for transfer and fusion was estimated to be approximately 200 ms, which would be consistent with the kinetics of exocytosis shown in **Figure 6**. However, a definitive demonstration of the idea that the ribbon acts as a conveyor belt will require the direct visualization both of the ribbon and of vesicle movements toward the active zone in response to calcium influx. It will also be important to identify the molecules forming the filaments that tether vesicles to ribbons, which do not appear to be actin. Are synaptic vesicles attached to ribbons by a motor protein?

See also: Calcium Channels and SNARE Proteins; Exocytosis: Ca^{2+}-Sensitivity; Hair Cells: Sensory Transduction; Hair Cell Differentiation; Photoreceptor Adaptation; Photoreceptors: Physiology; Retina: An Overview; Retinal Bipolar Cells; Retinal Development: An Overview; SNAREs.

Further Reading

Roberts WM, Jacobs RA, and Hudspeth AJ (1990) Colocalization of ion channels involved in frequency selectivity and synaptic transmission at presynaptic active zones of hair cells. *Journal of Neuroscience* 10: 3664–3684.

Dowling (1968) Synaptic organization of the frog retina: An electron microscopic analysis comparing the retinas of frogs and primates. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 170: 205–228.

Dowling JE and Ripps H (1973) Effect of magnesium on horizontal cell activity in the skate retina. *Nature* 242: 101–103.

Glowatzki E and Fuchs P (2002) Transmitter release at the hair cell ribbon synapse. *Nature Neuroscience* 5: 147–154.

Griesinger CB, Richards CD, and Asmore JF (2005) Fast vesicle replenishment allows indefatigable signalling at the first auditory synapse. *Nature* 435: 212–215.

Juusola M, French AS, Uusitalo RO, and Weckstrom M (1996) Information processing by graded-potential transmission through tonically active synapses. *Trends in Neuroscience* 19: 292–297.

Keen EC and Hudspeth AJ (2006) Transfer characteristics of the hair cell's afferent synapse. *Proceedings of the National Academy of Sciences of the United States of America* 103: 5537–5542.

Lagnado L, Gomis A, and Job C (1996) Continuous vesicle cycling in the synaptic terminal of retinal bipolar cells. *Neuron* 17: 957–967.

Lenzi D and von Gersdorff H (2002) Structure suggests function: The case for synaptic ribbons as exocytic nanomachines. *BioEssays* 23: 831–840.

Lenzi D, Runyeon JW, Crum J, Ellisman MH, and Roberts WM (1999) Synaptic vesicle populations in saccular hair cells reconstructed by electron tomography. *Journal of Neuroscience* 19: 119–132.

Nouvian R, Beutner D, Parsons TD, and Moser T (2006) Structure and function of the hair cell ribbon synapse. *Journal of Membrane Biology* 209: 153–165.

Neves G and Lagnado L (1999) The kinetics of exocytosis and endocytosis in the synaptic terminal of goldfish retinal bipolar cells. *Journal of Physiology* 515: 181–202.

Prescott ED and Zenisek D (2005) Recent progress towards understanding the synaptic ribbon. *Current Opinion in Neurobiology* 15: 431–436.

Roux I, Safieddine S, Nouvian R, et al. (2006) Otoferlin, defective in a human deafness form, is essential for exocytosis at the auditory ribbon synapse. *Cell* 127: 277–289.

Schaeffer SF and Raviola E (1978) Membrane recycling in the cone cell endings of the turtle retina. *Journal of Cell Biology* 79: 802–825.

Sterling P and Matthews G (2005) Structure and function of ribbon synapses. *Trends in Neurosciences* 28: 20–29.

tom Dieck S and Brandstätter JH (2006) Ribbon synapses of the retina. *Cell Tissue Research* 326: 339–346.

Vollrath L and Spiwoks-Becker I (1996) Plasticity of retinal ribbon synapses. *Microscopy Research and Techniques* 35: 472–487.

Zenisek D, Steyer JA, and Almers W (2000) Transport, capture and exocytosis of single synaptic vesicles at active zones. *Nature* 406: 849–854.

RNA Binding Protein Methods

P T Buckley and J H Eberwine, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Gene expression is regulated by a number of mechanisms from RNA transcription to protein translation. In eukaryotes, the primary forms of regulation at the posttranscriptional level involve the interaction of messenger RNA (mRNA) with RNA-binding proteins (RBPs) (Figure 1). More than 20% of annotated genes in the human genome encode proteins with nucleic acid-binding domains. This large number of molecules suggests a great level of complexity in the ways proteins and nucleic acids interact. The complexity of regulation may in part explain why the transcriptome of a cell does not always correlate with its proteome.

As the name for this class of molecules suggests, RBPs are proteins that can associate with and bind RNA molecules facilitating stages in RNA metabolism. RBPs are implicated in a wide variety of posttranscriptional events because RBPs associate and dissociate from RNA molecules at nearly all stages of RNA processing within the cell. Particularly in the case of mRNA, RBPs are of critical importance based on their diverse functions. A better understanding of these interactions will greatly impact the way gene expression is considered.

RBPs have been shown to function in RNA metabolism at all stages: transcription, splicing/alternative splicing, capping, polyadenylation, nuclear export, subcellular localization, stabilization, RNA editing, translation (as activators or repressors), degradation (in processes such as nonsense-mediated decay and RNAi), autoregulation (fragile X mental retardation protein-1 (FMR1) and human onconeural ventral antigen-1 (NOVA1)), and micro-RNA processing (miRNA). The profile of RBPs associated with a given RNA at any stage of processing is highly dynamic and can have varied consequences depending on the cellular context in which they are found.

RBPs can be found in both nuclear and cytoplasmic regions of living cells, suggesting their involvement in aspects of RNA metabolism associated with those regions. In the nucleus, RBPs associate with mRNA precursors (pre-mRNA), heterogeneous nuclear RNAs (hnRNA), and mature mRNA prior to export to the cytoplasm. The nucleus is the site of mRNA splicing, capping, and polyadenylation. Nuclear mRNA transcripts acquire a protein coat of cap and RNA-binding proteins allowing export and controlling functional fate. Nucleoli are membraneless regions within the nuclei of some cells and their primary function is ribosome biogenesis. Ribonucleoprotein (RNP) is also one of the main components of nucleoli and this region of the nucleus is thought to be the site of assembly for most cellular RNPs.

The assembly of RBPs and RNA transcripts are often thought of as existing as membraneless granules in the cytoplasm. Granules form upon mRNA export from the nucleus for a subset of transcripts. These granules are visible by light microscopy and can contain ribosomal subunits, translation factors, decay enzymes, helicases, scaffold proteins, and other RBPs. Granules are thought to be involved in the control of mRNA localization, stability, and translation of RNA cargo. At least some of the RBPs associated with mRNA transcripts in the nucleus are retained after export, although their functional role may alter depending on the cellular compartment in which they are found.

The cytoplasm is the site of mRNA translation upon association with the ribosome. Some mRNAs are immediately translated and processed following polysome assembly. Others are translationally delayed and transported or stored until developmental or environmental cues signal translation. All of these events involve the action of a host of RBPs that associate and dissociate from transcripts, and the RBPs associated with a transcript at a given time may govern mRNA fate. In polarized cells such as neurons, where large cytoplasmic volume can exist at great distances from the nucleus, RBPs become of particular importance. Transport of mRNAs to these regions requires the coordinated action of a complex system of RBPs called a RNP complex. Once localized into subcellular compartments, RNA processing, translation, and degradation can occur in response to synaptic stimulation.

RBP Structure

At the molecular level, RBPs have traditionally been identified by the presence of domains that are capable of binding RNA. These RNA-binding domains (RBDs) are often described as ribonucleoprotein domains or RNA recognition motifs. RBDs are highly represented in the genome, suggesting that a great number of proteins are involved in posttranscriptional regulation of RNA metabolism. Known RBDs include, but are not limited to, RNP consensus sequences (RNP1 and RNP2), K homology domains, zinc fingers (CCCH

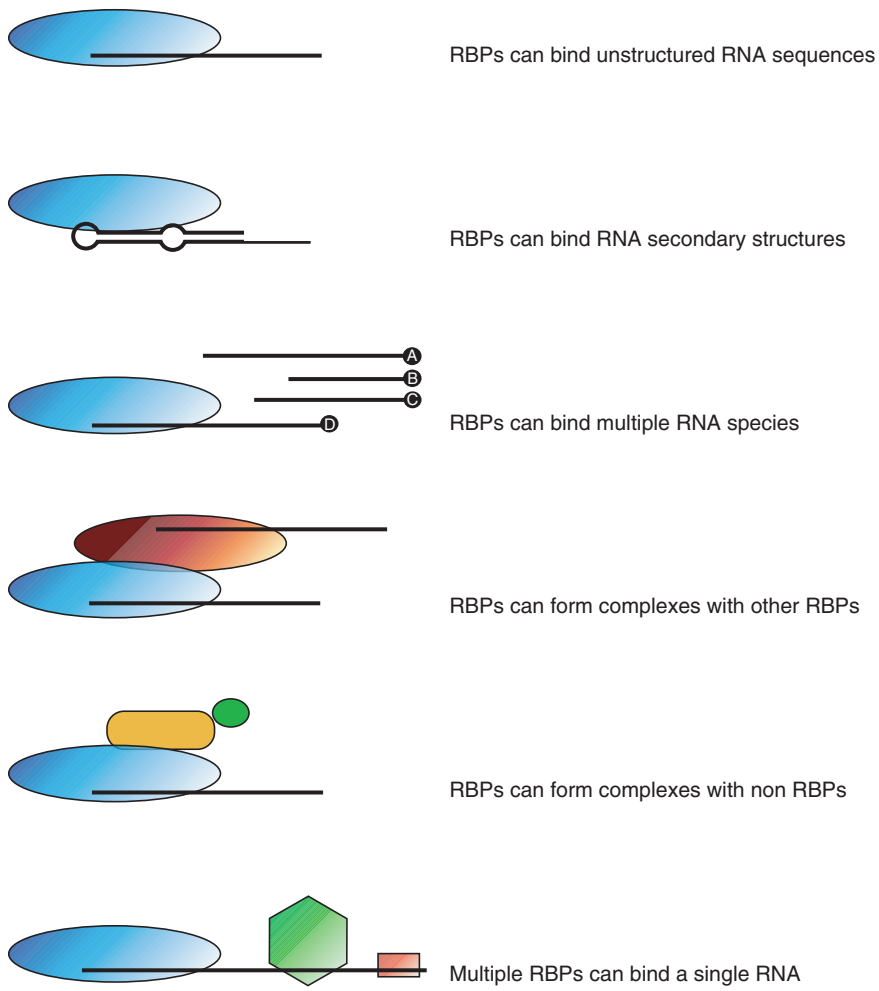

Figure 1 Diagram of reported RNA-binding protein (RBP)/RNA molecular interactions. Colored objects represent RBP molecules, and black lines represent RNA molecules.

and CCHC), RGG boxes, DEAD/DEAH boxes, Pumilio/FBF (PUF) domains, Piwi/Argonaut/Zwille (PAZ) domains, and double-stranded RNA-binding domains (dsRBD) (**Table 1**). Aromatic amino acid residues in the RNP motif result in a stacking interaction with RNA molecules that are stabilized by lysine residues in the helical portion of the molecule. Nucleic acid binding is a result of the highly positive nature of lysine side chains and the negative character of nucleic acid phosphate backbones.

RBDs with clearly defined sequences can be found throughout eukaryotic genomes. Studies have shown a surprisingly large number of potential RBPs. Presuming that these gene products do indeed act as RBPs, then this large number of RBPs suggests a complex highly regulated cellular system for controlling RNA biology. RBPs with more than one known RBD have been shown to contain the same domain represented multiple times or different domains present in the same protein.

Some of these domains are suggestive of the functional role of the RBPs in which they are found. The most well understood of these are the RNA helicase activity of DEAD/DEAH box domains, the short single-stranded RNA (RNAi and micro-RNA) binding activity of PAZ domains, and the role of SM domains in the binding of snRNA in splicing and tRNA processing. Other domains indicate RNA binding without indicating in which aspect of RNA metabolism they may participate. Homologs from other species and mutations can also be informative of the role an RBP may play in RNA metabolism.

RBPs often bind regulatory sequences in the 3′-untranslated region (UTR) of mRNA, suggesting the ability to molecularly identify binding sites in RNA molecules. It is conceivable that these binding sites may be defined by primary sequences within RNA molecules or by structural motifs formed by RNA secondary structure or, more likely, a combination of the two.

Table 1 Basic characteristics of known RNA-binding domains

	RNP1, RNP2	*K homology*	*Zinc fingers*	*DEAD/DEAH boxes*	*PUF*	*PAZ*	*dsRBD*	*R3H*
Typical structures	One or more copy of a 90-aa sequence containing four strands and two helices arranged in an alpha/beta sandwich with a third helix present during RNA binding in some cases	One or more copy of a 70-aa sequence containing a beta–alpha–alpha–beta–beta–alpha structure (type 1) or alpha–beta–beta–alpha–alpha–beta (type 2), multiple copies may function cooperatively	Each finger contains a 25- to 30-aa domain composed of a beta–beta–alpha fold stabilized by a zinc ion, contains conserved C and H residues separated by highly polar and basic residues	Occurs as a helicase domain of the aa motif DExD/H	Occurs as a tandem repeat of six to eight domains	Five stranded mix of beta-sheets surrounded by helices, tertiary structure belonging to Rnase H family	70-aa globular domain with an alpha–beta–beta–beta–alpha structure	Contain a conserved arginine and histidine residue separated by three other residues, also display a conserved pattern of hydrophobic residues, prolines, and glycines
RNA target motifs	ssRNA, U1A proteins recognize stem/loop RNA structures	ssRNA, ssDNA	mRNA 3′-UTR, short runs of guanine residues	RNA secondary structures	UGUR motif found in the 3′-UTR of target mRNAs and flanking sequences (highly selective)	dsRNA	dsRNA	ssRNA
Example proteins	hnRNPs, snRNPs	hnRNP K, FMR1, NOVA	matrin, U2AF splicing factor, many transcription factors	Involved in RNA transcription, editing, splicing, export, translation and turnover proteins across all kingdoms	Pumilio (*Drosophila*), FBF-2 (*C, elegans*), other proteins found across all eukaryotic kingdoms	Dicer, Argonate	Adenosine deaminases, PKR, Rnase III	Present in proteins from a diverse range of organisms
Proposed functions	Splicing, alternative splicing	Nucleic acid binding and protein–protein interactions	Several subtypes capable of binding both DNA and RNA C2H2 and CCCH subtypes typically associated with RNA binding	Helicase activity involved in the unwinding of dsRNA	Translational repressors affecting mRNA translation and stability	dsRNA-guided hydrolysis of ssRNA	Binding of RNA duplexes	ssRNA binding as suggested by other domains in R3H containing proteins

Data from the European Bioinformatics Institute's InterPro and Go databases.
aa, amino acid.

RNA Targets

RBPs can regulate the metabolism of multiple RNA targets, as suggested by diverse effects seen in mutational analysis leading to misregulation of many RNA targets. The RNA targets of RBPs are notoriously difficult to identify. Most RBPs have multiple targets and mutant phenotypes can be difficult to tie to an individual RNA target. Identification of RNA targets is required to fully understand RBP function. Identification of RBPs that bind to known regulatory sequences is also critical to defining the mechanism of posttranscriptional control utilized by a given RBP.

In order to function properly, RBPs must specifically differentiate and recognize target from nontarget RNA. Specific recognition of target molecules could occur by sequence recognition, structural recognition, or both. Understanding the dynamics of this recognition would help in the identification of unknown targets of RBP interaction containing similar features. However, definition of an RNA sequence motif for RBP interaction may fail to reflect physiological interactions based on concentration, subcellular localization, and the requirement of multiple proteins for proper function.

Assays (discussed later) exist to identify RNA cargoes of RBPs, and educated guesses can be made based on phenotypic information gleaned from RBP mutants. Using *in vitro* techniques to attempt to assess *in vivo* associations between molecules requires a clear recognition of the limitations involved. *In vitro* methods of defining RNA targets of RBPs may alter the *in vivo* associations of RNA–protein complexes. RNAs associated with off-target proteins may be isolated as part of complexes. Immunoprecipitation (IP) experiments have been shown capable of generating complexes that form artificially *in vitro* via reassociation of molecules after cell lysis. These methods destroy the *in vivo* context and associations of RBPs and their targets, opening these molecules to new associations with molecules they would not encounter in normal cellular regulation.

RBPs are vulnerable to being bound by stretches of repeated sequences. Therefore, this type of binding is increased when these types of sequences are expanded in disease leading to altered function of RBPs. This is particularly true with splicing, where altered splicing patterns can be created upon nucleotide expansion.

RBP–RNA Cargo Identification

Identification of RNA targets of RBPs is key to understanding how RBPs posttranscriptonally control gene expression. By virtue of RBPs having multiple RNA targets, specific features of target versus nontarget RNA can be determined once cargoes for particular RBPs have been identified. Identification of the multiple RNA targets for an RBP can also yield insight as to the type of RNA metabolism that an RBP regulates, the mechanism or mechanisms used to control metabolism for a given RBP, and how RBPs exert coordinated control of RNA metabolism. A number of assays have been developed to analyze RBPs and identify target RNAs and are described here.

Immunoprecipitation

The most common assay used to identify RNA targets of RBPs involves the use of IP to isolate RNA–protein assemblies using antibodies to the RBP of interest. Once isolated, these assemblies can be screened by microarray to identify the RNA species captured with the RBP. Prior to the development of microarray techniques, RNA targets would have been identified by subtractive hybridization, cloning, and sequencing. Microarrays allow for a tremendous increase in the power of identification of targets.

The most significant limitation of these assays is the lack of specificity of the technique for both the antibody and the capture of nontarget RNAs. RNA molecules can be particularly 'sticky' and may be identified as targets despite having no specific interaction with the RBP under scrutiny. Cellular lysis steps can significantly alter the concentrations and context of RNA–protein interactions, leading to associations that are a consequence of the isolation technique rather than the biology of the molecules. RBPs also frequently associate with other proteins as part of multiprotein complexes. These proteins could provide additional sites of RNA interaction and binding and could in fact be other RBPs. Because these proteins would be isolated in the IP technique, any RNA cargo they interact with would incorrectly be seen as interacting with the RBP target of the IP.

Cross-Linking and Immunoprecipitation

A more recently developed technique, cross-linking and immunoprecipitation (CLIP), incorporates a cross-linking step in living cells to allow for isolation of *in vivo* RNA targets. Upon cross-linking, the RNA–protein complexes are isolated from the cellular extract by IP, and the associated endogenous RNAs are identified by microarray.

The ultraviolet cross-linking step incorporated into this technique was proffered to avoid some of the issues associated with the less selective traditional IP assays. However, since protein–protein cross-linking can occur, it is possible that nontarget RBPs and their cargoes are immunoprecipitated with the target RBP. Furthermore, the sensitivity of PCR makes it difficult

to ensure that small amounts of high-abundance mRNAs are not simply 'nonspecifically' associated with the IP complex and thereafter aberrantly identified as cargoes.

Yeast Three-Hybrid Screening

The yeast three-hybrid screen is an RNA target identification technique based on the yeast two-hybrid system used to detect protein–protein interactions. Phenotypic properties of yeast are used to detect and analyze RNA–protein interactions using two hybrid proteins along with a hybrid RNA molecule. A hybrid protein composed of a known RBP (with an RBD) is generated with a DNA-binding domain that binds to the promoter region of a reporter gene. Another hybrid protein is generated that is composed of a second RBD along with a transcriptional activation domain for the reporter gene. The RBDs of these proteins can bind the RNA sequences of the hybrid RNA molecule. As binding of the RNA molecule occurs, the DNA-binding and transcription-activating hybrid proteins are brought together and RNA–protein interactions are detected by expression of the reporter gene.

Antibody-Positioned RNA Amplification

In the antibody-positioned RNA amplification (APRA) technique, an antibody to an RBP of interest is covalently coupled to a single-stranded DNA oligo composed of a sequence of nonhomologous DNA, a restriction enzyme site, and a degenerate sequence. As the antibody specifically binds to the protein of interest, the oligo can anneal to mRNA cargoes in the vicinity. Priming from the oligo allows for directed *in situ* transcription. A second complementary DNA strand can then be synthesized, and the double-stranded DNA can then be digested away from the APRA construct, isolated, and amplified for analysis by microarray.

PNA-Assisted Identification of RNA-Binding Proteins

The peptide nucleic acid (PNA)-assisted identification of RNA-binding proteins (PAIR) assay provides the opposite type of information from the previously described techniques in that it allows for a means of RBP identification for any given RNA sequence rather than identification of RNA cargoes for a given RBP. In the PAIR assay, a PNA containing photoactivatable amino acids (AAs) in association with a cell-penetrating peptide (CPP) can be synthesized that is complementary to an RNA sequence of interest. After the CPP carries the PNA into the cell, the PNA anneals with its complementary sequence *in vivo* and the PNA is then crosslinked to proteins near where it bound by activation of the photoactivatable AAs. This PNA–RBP complex can be isolated and the PNA annealed to its target sequence immobilized on a bead which serves as a way to precipitate associated protein–RNA complexes. The proteins from these complexes can be isolated from the beads and target RNA for analysis by proteomic assays for identification.

The most significant limitation to exploring post-transcriptional processing in this way is the fact that an mRNA associating with any of these RBPs does not definitively show that those mRNAs are subject to the metabolic events normally associated with those RBPs (e.g., splicing enzymes may regulate translation not just splicing). A vast number of these proteins have secondary functions that are not completely understood, and a conclusion cannot be absolutely drawn without further investigation.

Global Regulatory Role of RBPs

RBPs are essential to regulating posttranscriptional gene expression. In addition to facilitating a number of mRNA processing events in the cell, RBPs are thought to be responsible for generating proteome diversity through alternative splicing and translational control of existing mRNAs. Many cell-specific RBPs are critical for phenotypic definition and are critical to developmental regulation.

RBP dysregulation has been implicated in a number of diseases and disorders. FMR1 disregulation is the cause of fragile X syndrome, the most common form of inherited mental retardation. CELF proteins are integral to normal heart and skeletal muscle development and in the pathology of myotonic dystrophy. NOVA autoimmune antigens are neuron-specific proteins involved in the pathogenesis of the neurodegenerative syndrome paraneoplastic opsoclonus–myoclonus ataxia.

Any dysregulation or mutation of RBPs can have dramatic biological implications for living systems. The RNA cargoes associated with the RBP and the metabolic function of the RBP can both be altered by impaired protein function. The repercussions of disrupting a single RBP can affect any number of systems, leading to disease. In the same way, altering the sequence of an RNA target can alter its interaction with any number of RBPs. These changes could be localized to regulatory and noncoding sequences within a transcript and still create a physiological effect in the same manner as a change in the coding sequence. Understanding the mechanism of binding and function of RBPs is essential to treating conditions impacted by RBP disregulation.

See also: Axonal mRNA Transport and Functions; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; siRNA: Utility.

Further Reading

Anderson P and Kedersha N (2006) RNA granules. *Journal of Cell Biology* 172(6): 803–808.

Brown V, Small K, Lakkis L, et al. (1998) Purified recombinant Fmrp exhibits selective RNA binding as an intrinsic property of the fragile X mental retardation protein. *Journal of Biological Chemistry* 273(25): 15521–15527.

Job C and Eberwine J (2001) Localization and translation of mRNA in dendrites and axons. *Nature Reviews Neuroscience* 2(12): 889–898.

Kim VN and Dreyfuss G (2001) Nuclear mRNA binding proteins couple pre-mRNA splicing and post-splicing events. *Molecules and Cells* 12(1): 1–10.

Maris C, Dominguez C, and Allain FH-T (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS Journal* 272(9): 2118–2131.

Miyashiro KY, Beckel-Mitchener A, Purk T, et al. (2003) RNA cargoes associating with FMRP reveal deficits in cellular functioning in Fmr1 null mice. *Neuron* 37(3): 417–431.

Moore MJ (2005) From birth to death: The complex lives of eukaryotic mRNAs. *Science* 309(5740): 1514–1518.

Musunuru K (2003) Cell-specific RNA-binding proteins in human disease. *Trends in Cardiovascular Medicine* 13(5): 188–195.

SenGupta DJ, Zhang B, Kraemer B, Pochart P, Fields S, and Wickens M (1996) A three-hybrid system to detect RNA–protein interactions *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* 93(16): 8496–8501.

Tenenbaum SA, Lager PJ, Carson CC, and Keene JD (2002) Ribonomics: Identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. *Methods* 26(2): 191–198.

Ule J, Jensen K, Mele A, and Darnell RB (2005) CLIP: A method for identifying protein–RNA interaction sites in living cells. *Methods* 37(4): 376–386.

Venter JC, Adams MD, and Myers EW (2001) The sequence of the human genome. *Science* 291(5507): 1304–1351.

Zielinski J, Kilk K, Peritz T, et al. (2006) *In vivo* identification of ribonucleoprotein–RNA interactions. *Proceedings of the National Academy of Sciences of the United States of America* 103(5): 1557–1562.

Relevant Website

http://www.ebi.ac.uk – European Bioinformatics Institute.

RNA Granules: Functions within Presynaptic Terminals and Postsynaptic Spines

E W Khandjian, B Tournier, S Séguin, S Tremblay, and P De Koninck, Université Laval, Québec, QC, Canada
L Davidovic, Université de Nice Sophia-Antipolis, Nice, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Of all cells in nature, neurons have the most remarkable and diverse shapes. They vary from simple unipolar cells to highly complex cells, decorated with a profusion of projections that extend up to several centimeters and even a meter in the case of axons. This diversity in anatomical shapes generates myriad possibilities of cell-to-cell interactions, to form adaptable circuits through millions of synaptic connections. Far away from the cell body, synapses are functionally autonomous distinct subcellular domains; under constant remodeling and adaptation, they contribute to long-lasting changes in synaptic strength that are believed to play a role in neuronal development, learning, and memory processes. These changes rely on local rapid modulation of protein synthesis that is dependent on the presence of mRNA at the synapse. Logistic concepts arise on how to supply and functionally maintain territories and regions far away from the cell body. The majority of neuronal protein synthesis takes place in the somatic compartment and, as a result, the vast majority of the messenger RNAs (mRNAs) translated by the neuron are encountered in the cell body. However, approximately 5% of the mRNAs are also delivered either at presynaptic terminals or at postsynaptic dendritic spines, where these mRNA are translated into proteins that are required specifically in response to the local needs of the neuron. This transfer of mRNA provides a mechanism both for controlling the distribution of the corresponding protein and for regulating independently its level in individual axonal terminals or dendritic spines. The mechanisms of transport, targeting, and release of neurospecific mRNAs at the synapse are beginning to be unveiled, and involve motile macromolecular complexes that have been named RNA granules. These structures have been described as translationally silent ribonucleoparticles (RNPs), which contain mRNAs packed with RNA-binding proteins and ribosomes.

The Expanding World of Diverse Cytoplasmic RNA Granules

Originally, the term 'granule' referred to round-shaped cellular structures that, upon adequate staining, were visible using light microscopy. Granules were also subsequently seen under the transmission electron microscope as electron-dense structures. They were observed in dendrites as well as in axons, but their identities were either speculative or not documented. The possibility that some granules might contain mRNA was not considered until the 1980s, since the prevailing view was that all dendritic and axoplasmic proteins originated from translation of mRNA in the cell body, and were subsequently transported to the neuronal branches through mechanisms called 'long-range slow and rapid' dendritic or axonal transport.

The discovery of polyribosome aggregates beneath postsynaptic sites at the base of dendritic spines has changed our vision of how translation could also occur at the synapses, to regulate independently and rapidly protein synthesis in response to specific stimuli. Concomitantly, clues for the dendritic presence of extrasomatic mRNA were provided and it was established that *in vitro* protein synthesis also occurs in synaptosomes, and even in dendrites preparations, both isolated from neurons maintained in culture. The mechanism by which these mRNAs were targeted to the neuronal branches remained elusive until recently. Granules containing RNA have been documented in invertebrate neurons and also in fibroblasts, ovocytes, and oligodendrocytes. However, neuronal vertebrate RNA granules were only described in the 1990' as motile messenger ribonucleoprotein (mRNP) complexes encountered in the dendritic extensions of neurons. Although mostly studied in invertebrates systems, axonal RNA granules have also been characterized in vertebrate neurons.

RNA Granules, a Term Too General

Several types of RNA granules can be observed in all cells, including neurons. In addition to RNA detected by *in situ* hybridization, these structures can be visualized using immunofluorescent approaches with antibodies to known RNA-binding proteins. Also, they can be seen after microinjection of labeled RNA molecules, or after transfection with vectors coding for tagged-RNA-binding proteins as is the case for

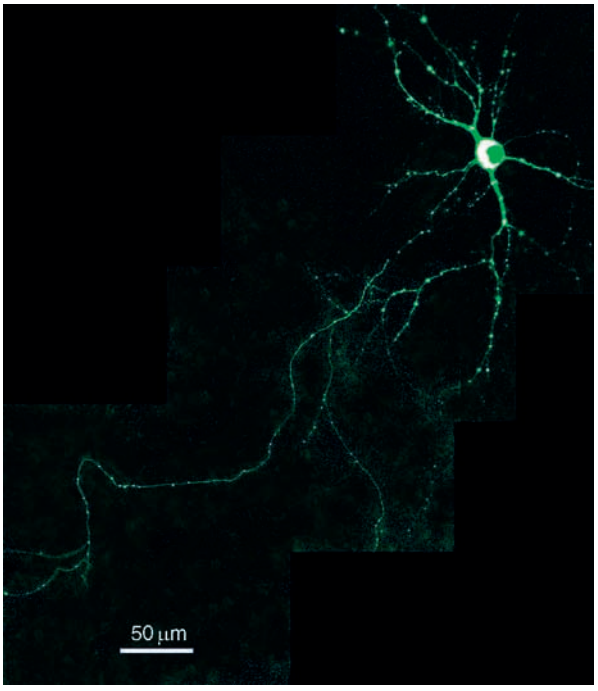

Figure 1 Fragile X mental retardation protein, an RNA-binding protein that interacts with mRNA, is transported to dendrites in granule complexes. The protein, tagged with a green fluorescent protein, was expressed in rat hippocampal neuron in culture after transfection with an expression vector coding for the fusion protein. Granules, visible by epifluorescence microscopy, are observed as far away as 500 μm from the cell body. Reproduced from Davidovic L, Tremblay S, De Koninck P, and Khandjian EW (2005) Fragile X mental retardation protein: To be or not to be a translational repressor. In: Ju Sung Y and Denman RB (eds.) *The Molecular Basis of Fragile X Syndrome*, pp. 129–144. Kerala, India: Research Signpost, with permission from Research Signpost.

FMRP (Figure 1). The list of different proteins with affinity to RNA that appear as granules is increasing daily. However, since different RNA-containing granules have been described, it is worth trying to classify them into specific types.

Polyribosomes

Polyribosome structures are made of ribosomes connected by a strand of mRNA that runs between the large and small ribosomal subunits. A number of studies have shown that nonribosomal proteins, mostly RNA-binding proteins, associate with these structures referred often as aggregates, although the term 'clusters of polyribosomes' is more appropriate. These clusters can be visualized by electron microscopy as particles, but are not individually visible by conventional light microscopy because of its limited resolution of ~0.2 μm. However, clusters of ribosomes can be visualized under certain circumstances in neurons (see below).

Stress Granules

When a cell is submitted to a stress, such as heat shock, UV irradiation, oxidative stress, or hypoxia, it responds rapidly to contend with the insult. The amount of translating polyribosomes diminishes as the translational machinery is mobilized to treat only specific mRNAs, such as the heat shock mRNAs. Meanwhile, other mRNAs are removed, stored and stabilized in stress granules (SGs) as nontranslated mRNPs. SGs are formed with an aberrant eIF2–GTP–$tRNA^{Met}$ ternary complex and by recruitment of different RNA-binding proteins that are trapped with the mRNAs. In these SGs, the initiator tRNA is not loaded onto the 40S ribosomal subunit, and recruitment of the 60S large subunit is not observed. Once the environmental stress is relieved, stored mRNAs are released from the SGs, which vanish gradually while overall translation resumes.

Stresslike granules can also be induced and detected after transient expression of different RNA-binding proteins in all kind of cells. These structures are formed when the introduced RNA-binding proteins trap mRNAs, overloading the translation machinery that becomes overcrowded and easily visible under the microscope, since granules as large as 2–3 μm can be formed.

Processing Bodies

Processing (P) bodies are structures that can be visualized as dynamic granules. They contain mRNA and components of the nonsense-mediated decay pathway, and the 5′–3′ mRNA decay machinery, including decapping enzymes and endonucleases. They also contain factors involved in small RNA-guided gene silencing. The sequestered mRNAs can be stored transiently in a repressed state, to be later released from P bodies to undergo translation in the cytoplasm. Whether the RNA-induced silencing complexes (RISCs; see later) are part of these bodies is still not clear. While P bodies might be involved in local degradation or storage of specific mRNAs in dendrites, it is not known whether the whole complex, in the form of a granule, is translocated from the soma to synapses, or whether the complexes are assembled locally and transiently in the dendrites on demand. It is important to note that P bodies do not appear to contain ribosomes and therefore cannot be associated with the translation machinery.

miRNP and RISC Complexes

Consistent with the fact that mRNAs are maintained translationally silent in granules, several small noncoding RNAs that could behave as translation repressors

have been detected in granular structures. The brain-specific cytoplasmic RNA 1 (BC1), a 152 bp-long noncoding RNA polymerase III transcript, is also targeted to dendrites, suggesting that it could control the fate of certain mRNAs.

An intriguing question is whether control of translation in granules could rely on miRNA-mediated regulation of translation. This mechanism involves a class of small, 19- to 22 nt-long RNAs, termed microRNAs (miRNA), that can specifically repress translation of homologous mRNAs. Several recent reports suggest that neuron-specific miRNAs might act in nervous system morphogenesis. These miRNAs have also been found to target specific mRNAs into P bodies involved in mRNA storage and/or degradation.

It is not clear whether RISC, the macromolecular protein complex responsible for RNAi, would appear as a granular structure in microscopic observations of dendrites or axons. However, several RISC members are found in different types of granules, including the protein Argonaute and fragile X mental retardation protein (FMRP). At the molecular level, there is evidence that the RISCs affect mRNA transport to the synapse and the synthesis of synaptic proteins associated with memory in *Drosophila*. For example, miR-134, which localizes in the dendritic compartment, affects translation of specific mRNA that control dendritic spine size. Although preliminary, these observations suggest that miRNA/RNAi could play crucial roles in granule formation and functions in neurons.

Neuronal RNA Granules

As opposed to stress granules, which are seen only under stressful conditions, neuronal RNA granules are motile complexes that transport and deliver transcripts at distant sites in the cell. Once they arrive at their destination, it is inferred that they are docked to membranous or cystoskeletal structures, awaiting a signal to initiate translation. Biochemical evidence that these granules contain mRNA and ribosomes has been obtained after isolation from primary cultured neurons of a fraction of material sedimenting faster than polyribosomes, with coefficients above 1000 S (Svedbergs: $S_{20.w}$). Electron microscopy studies reveal that RNA granules are round-shaped structures that contain electron-dense material and tightly packed ribosomes. Their reported sizes range from 150 to 1000 nm, but the large aggregates appear as if several smaller granules had coalesced.

In addition to RNA granules containing ribosomes, it appears that other species of granular structures containing mRNA, but not ribosomes, can be detected in neurites. It has been suggested that these structures carry mRNA or/and other RNA species, to be delivered at specific sites either postsynaptically or in axons. These structures are apparently the supply cargoes required to replenish and feed the translation machinery with newly transcribed mRNA, as these molecules are turning over with various half-lives.

Characterization of Neuronal RNA Granules

Neuronal RNA granules are extremely heterogeneous in size, as determined by immunofluorescent studies using different antibodies to the protein associated with these structures. Also, they are heterogeneous in composition, since each can carry different types of RNA-binding proteins. However, some RNA-binding proteins co-localize in certain subpopulations

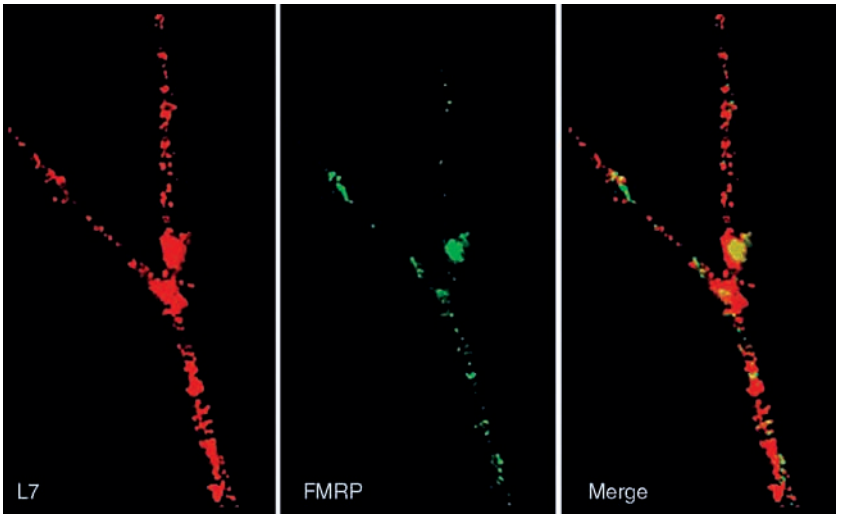

Figure 2 Axons from rat sympathetic (DRG) neurons growth culture contain RNA granules. Ribosomes were stained with anti-L7 serum (in red) and with monoclonal antibody 1C3 against fragile X mental retardation protein (FMRP), an RNA-binding protein (in green). The merged images show that FMRP is not associated with all ribosome-containing granules.

of granules (Figure 2). These observations strongly suggest that the protein composition of the granules is neither uniform nor universal. The identity and stoichiometry of the different components are subjected to high variability, and this variability is thought to determine the types of mRNAs to be transported, as well as the specific destination of the cargo.

To tentatively characterize neuronal RNA granules, biochemical approaches have been used. These are based on two main methods of separation: (1) isokinetic ultracentrifugation in sucrose density gradients to separate the subcellular components according to their sedimentation values (S) and (2) affinity chromatography, with specific antibodies or recombinant protein to isolate structures with their surrounding components.

Protein Components

Neuronal RNA granules contain RNA-binding proteins either associated with RNA transport and/or translational control, such as the members of the fragile X mental retardation protein family (FMRP, FXR1P, and FXR2P), CPEB, RNG105, Purα, Purβ, Pum2, TIAR, and TIA-1, and proteins specialized in the recognition of RNA dendritic or axonal targeting sequences, including the zipcode-binding protein-1 (ZBP1), SYNCRIP, MARTA 1, MARTA 2, and Staufen. Proteins involved in protein synthesis, such as elongation factors (e.g., EF1α, eIF2α, eIF2β and eIF2γ), are also found in RNP granules. Interestingly, they also contain proteins originally involved in nuclear posttranscriptional processes, such as RNA helicases (DDX1 and DDX3), heterogeneous nuclear ribonucleoproteins (hnRNP U, hnRNP A, and hnRNP B), and proteins involved in mRNA splicing, such as TLS, Sam68, Sc35, and polypyrimidine tract-binding protein-associated splicing factor (PSF). In addition to these specific proteins, RNP granules also contain, or at least are associated with, protein motors that are known to be involved in active transport (see below).

RNA Components

Electron microscopy analyses have revealed the presence in granules of electron-dense structures that react with the RNA stain SYTO-14. Granules also contain ribosomal RNA (rRNA) constituents of the ribosomal subunits. Due to the apparent high level of compaction of the ribosomes in the granule, it is inferred that the mRNA remains translationally incompetent. Further strengthening the hypothesis that RNA granules are repressed, it was shown that they contain little or no tRNA, a key component in translation.

Neuronal RNA granules are thought to be reservoirs of mRNAs that are transported into dendrites for subsequent site-specific utilization at synapses undergoing protein synthesis-dependent changes. To identify the mRNA content of the granules it is necessary to separate these structures from other RNPs, especially from polyribosomes. Different approaches are applied to enrich RNP granules, but these are mainly based on the different sedimentation properties between the various ribonucleoparticles present in the cell sap. It is not clear to what extent these RNP granules are pure, since they have not yet been fully identified at the biochemical level. The list of mRNA present in these RNP granules is expanding, with emphasis being put on specific mRNAs that direct the synthesis of proteins required in synaptic cytoskeleton remodeling (i.e., the microtubule-associated protein 2 (MAP2) and the activity-regulated cytoskeleton-associated protein (Arc)) and in synaptic functions (i.e.,CaMKIIα, neurogranin, voltage-gated potassium channel Kv1.1, and glutamate receptors). Other mRNAs encode proteins involved in translational control, including translation factors (i.e., elongation factor 1α, eukaryotic initiation factor 4E(eIF4E)) or RNA-binding proteins (i.e., cytoplasmic polyadenylation element-binding protein (CPEBP) or fragile X mental retardation protein). Finally, it is assumed that local translation of another class of mRNAs encoding transcription factors (i.e., the cAMP response element-binding protein, CREB) mediates short- and long-term changes in the repertoire of the genes expressed at the somatic level.

On the axonal side, several mRNAs have been identified, most of them encoding proteins involved in cytoskeleton remodeling, such as Tau and actin-depolymerizing factor (ADF), or in signal transduction, such as ephrin A2 (EpHA2) and importin β1. Other studies have inferred the presence of additional cytoskeleton-linked mRNAs (vimentin, cofilin), and several heat shock proteins mRNAs (HSP27, HSP60, HSP70), the products of which function as protein chaperones to control proper folding after protein synthesis.

Assembly and Sorting of RNA Granules: In the Nucleus or in the Cytoplasm?

To avoid ectopic translation in the cell body, the dendritic-targeted mRNAs have to be sorted from mRNA destined to somatic translation shortly after transcription and held in a translationally dormant state during transport to the appropriate destination (Figure 3). The presence of nuclear proteins such as hnRNP U, hnRNP A, and hnRNP B in some RNA granules suggests that the sorting or preassembly of the mRNA and

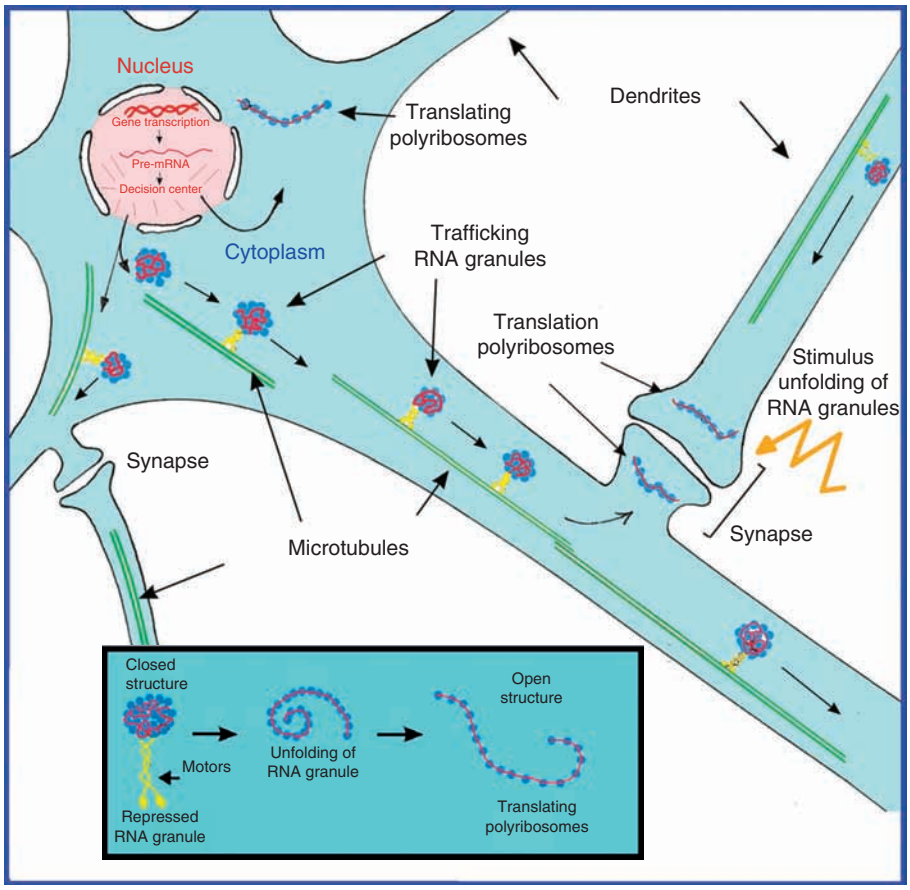

Figure 3 Model for sorting and trafficking of RNA granules. Once the 'decision' is taken in the nucleus, RNA granules are translocated onto microtubules and are transported to dendrites using motor proteins. On arrival at their destination, the granules are docked and wait for a stimulus, following which they unfold and resume protein synthesis on demand.

RNA-binding proteins in these granules may occur in the nucleus. In addition, RNP granules contain several RNA-binding proteins known to induce translation repression, such as FMRP, CPEBP, RNG105, Purα, Purβ, and p50, some of which harbor nuclear localization and nuclear export signals. These data strongly suggest that translation repression and mRNA transport within granules are tightly coupled. mRNAs could be sorted by means of the recognition of its cis-acting elements by specific RNA-binding proteins, to form mRNPs. The few known cis-acting elements are composed of a primary nucleic acid sequence and a secondary RNA structure that generate a specific RNA-binding protein recognition site. For example, the mRNA encoding β-actin is a well-known dentritically localized mRNA. Its targeting to dendritic spines or to the growth cones of developing neurites involves a specific interaction between ZBP1 and a 54-nucleotide zipcode sequence in the 3′-untranslated region (3′-UTR) of β-actin mRNA.

Another scenario can be proposed, based on new concepts about the continuous shuttling of mRNPs between hypothetical discrete cytoplasmic 'microdomains' and the translation machinery. In other words, decisions are taken in these 'centers,' whether an mRNA should be translated or be maintained as repressed mRNPs during its translocation at specific sites in the cell. Under certain physiological conditions, these structures merge, eventually to form granules large enough to be seen under the microscope.

Transport of Granules

The use of cytoskeleton-disrupting agents reveals that transport of neuronal RNA granules requires microtubules for rapid long-range axial transport and actin cytoskeleton for local dependent radial translocation locally at the synapse. Time-lapse videomicroscopy on hippocampal neurons harboring SYTO14-RNA dyed granules, green fluorescent protein (GFP)-labeled CaMKII 3′-UTR RNA, and GFP-Staufen enabled to visualize oscillatory (bidirectional) movements over short distances, or processive movements over larger distances in the direction of the growth cone and distal dendrites (anterograde). The oscillatory movement of granules has been attributed to fine local

adjustments within specific dendritic sites (i.e., clusters of synapses). RNA granules display an average velocity between 0.03 and 0.15 $\mu m\,s^{-1}$, compatible with active transport via kinesin and dynein. The kinesin I family (KIF5), and the kinesin II family (KIF3C) are thought to move a subset of granules toward distal processes, whereas dynein would move granules back to the soma. The direction of movement of RNA granules along microtubules appears to be determined by the balance of power between the conventional kinesin and cytoplasmic dynein in contact with each granule. However, due to the heterogeneity of the granules, the nature of the motors involved in their individual transports is not fully understood.

Interestingly, only a small portion of the tagged granules actually move within a neurite, and it is thought that these stationary elements might represent granules that have reached their final destination. Alternatively, it might be possible that stationary granules correspond to SGs that have been induced by high levels of the GFP-tagged proteins used, which could have trapped the mRNAs.

Derepression of Granules

When RNA granules have reached their final destination, they are docked pre- or postsynaptically until their translation is triggered by physiological and/or developmental stimuli, essentially determined by neuronal activity. Activity-dependent translation is triggered via a combination of mechanisms, including activation of the general protein-synthesizing machinery in dendrites and release from repression of specific mRNAs. It is likely that regulation also involves the selection of mRNA for translation and (maybe) competition between mRNAs for the limited available translation machinery at the synapse. Indeed, due to space limitation in the dendritic spine neck, where polyribosomes clusters are encountered, it is likely that a limited number of mRNAs is actually undergoing active translation at the synapse at the same time.

Prolonged depolarization of cortical neurons induces changes in the structure of granules, including apparent decondensation of the packed ribosomes, which can protrude outside the granules to be reshaped into polyribosomes. These changes are accompanied by postsynaptic polyribosome upregulation and a shift in the distribution of several dendritic mRNAs from the heavy-sedimenting granule fractions into actively translating polyribosomal fractions. Concomitantly, translation factors can either be immediately synthesized on the spot, or can be recruited into granules in response to synaptic activation. For example, brain-derived neurotrophic factor (BDNF) treatment of various neuronal preparations results in the redistribution of eIF4E to an RNA granule-rich cytoskeletal fraction. eIF4E linkage to cap structures mediates the recruitment of other translation factors to initiate translation. Inhibition of translation is likely to occur at this step. For example, the CPEB protein that localizes to dendrites and synapses binds to a specific cis-element in the 3′-UTR of CaMKIIα mRNA. CPEB forms a complex with the 5′-cap via interaction with Maskin, a protein that binds to eIF4E, to prevent the formation of the initiation complex and to keep the mRNA translationally dormant. When CPEB is phosphorylated, it facilitates polyadenylation and translation initiation by recruitment of polyA-binding protein (PABP) and eIF4G. Binding of eIF4G to the 5′ cap displaces Maskin from eIF4E, and enables translation of CaMKIIα mRNA. FMRP is also thought to regulate mRNA translation in dendrites by repressing translation initiation. Although the specific mechanism of FMRP action remains unknown, FMRP binds to several mRNAs via its multiple RNA-binding domains, and it has been described as a translation repressor. The fragile X syndrome (FXS), a cause of inherited mental retardation, results from the loss of FMRP. Mice and humans lacking FMRP exhibit changes in dentritic spines density and morphology, as these appear longer, thinner, and more tortuous than normal, looking immature and not fully differentiated. These observations are thought to be at the heart of behavioral and cognitive abnormalities, apparently due to altered local synaptic protein synthesis impairing proper connectivity between neurons. FXS is the only known human genetic inherited disease linked to alterations in synaptic translation and is therefore a great focus of attention in neurobiology.

Conclusion and Perspectives

Real-time fluorescence imaging has drastically changed our view and perception of the dynamics of RNA granules in neurons. It is now possible, in theory, to follow a single RNA granule, from its formation to its demise, and to monitor its complex paths of translocation. However, little is known about the signaling pathways and molecular mechanisms governing granule formation and decay. Since the discovery of polyribosomes at synaptic sites, recent studies have provided evidence linking synaptic plasticity to local protein synthesis. The presence of clusters of polyribosomes enables rapid triggering of local protein synthesis in response to exogenous stimuli. Logistically and energetically speaking, it seems that transfer of the whole translation apparatus, including ribosomes, to distal locations is

a more profitable operation than transferring bits and parts of this complex machinery. A second class of RNA granules lacking ribosomes would be necessary to replenish the translation apparatus, standing by for new rounds of translation with new mRNA transcripts. In parallel, the aging mRNAs would be addressed to P bodies, where they will be degraded. The model of the relationship between the different classes of these dynamic granules is certainly a simplistic and anthropomorphic view, and more studies are needed.

RNA-binding proteins, which are found in all neuronal RNA granules, play pivotal roles in regulating RNA metabolism. It is fascinating that among the 380 putative RNA-binding proteins coded in the mouse, 61% are detected in brain. Such high concentration of this class of proteins is certainly indicative of the high level of translation in the brain.

See also: Axonal mRNA Transport and Functions; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; Growth Cones; RNA Binding Protein Methods; siRNA: Utility.

Further Reading

Alvarez J, Giuditta A, and Koenig E (2000) Protein synthesis in axons and terminals: Significance for maintenance, plasticity and regulation of phenotype with a critique of slow transport theory. *Progress in Neurobiology* 62: 1–62.

Anderson P and Kedersha N (2006) RNA granules. *Journal of Cell Biology* 172: 803–808.

Antar LN, Dictenberg JB, Plociniak M, et al. (2005) Localization of FMRP-associated mRNA granules and requirement of microtubules for activity-dependent trafficking in hippocampal neurons. *Genes, Brain and Behaviour* 4: 350–359.

Bassell GJ and Singer RH (2001) Neuronal RNA localization and the cytoskeleton. *Results & Problems in Cell Differentiation* 34: 41–56.

Davidovic L, Jaglin XH, Lepagnol-Bestel A-M, et al. (2007) The Fragile X mental retardation protein is a molecular adaptor between the neurospecific KIF3C kinesin and dendritic RNA granules. *Human Molecular Genetic* 16: 3047–3058.

Davidovic L, Tremblay S, De Koninck P, and Khandjian EW (2005) Fragile X mental retardation protein: To be or not to be a translational repressor. In: Ju Sung Y and Denman RB (eds.) *The Molecular Basis of Fragile X Syndrome*, pp. 129–144. Kerala, India: Research Signpost.

Eberwine J, Miyashiro K, Kacharmina JE, et al. (2001) Local translation of classes of mRNAs that are targeted to neuronal dendrites. *Proceedings of the National Acadedmy of Sciences of the United States of America* 98: 7080–7085.

Giuditta A, Kaplan BB, van Minnen J, et al. (2002) Axonal and presynaptic protein synthesis: New insights into the biology of the neuron. *Trends in Neuroscience* 25: 400–404.

Hirokawa N (2006) mRNA transport in dendrites: RNA granules, motors, and tracks. *Journal of Neuroscience* 26: 7139–7142.

Kanai Y, Dohmae N, and Hirokawa N (2004) Kinesin transports RNA: Isolation and characterization of an RNA-transporting granule. *Neuron* 43: 513–525.

Knowles RB, Sabry JH, Martone ME, et al. (1996) Translocation of RNA granules in living neurons. *Journal of Neuroscience* 16: 7812–7820.

Krichevsky AM and Kosik KS (2001) Neuronal RNA granules: A link between RNA localization and stimulation-dependent translation. *Neuron* 32: 683–696.

Mazroui R, Huot M-E, Tremblay S, et al. (2002) Trapping of messenger RNA by Fragile X Mental Retardation protein into cytoplasmic granules induces translation repression. *Human Molecular Genetics* 11: 3007–3017.

Steward O and Worley P (2001) Localization of mRNAs at synaptic sites on dendrites. *Results & Problems in Cell Differentiation* 34: 1–26.

Steward O and Schuman EM (2003) Compartmentalized synthesis and degradation of proteins in neurons. *Neuron* 40: 347–359.

Tiruchinapalli DM, Oleynikov Y, Kelic S, et al. (2003) Activity-dependent trafficking and dynamic localization of zipcode binding protein 1 and beta-actin mRNA in dendrites and spines of hippocampal neurons. *Journal of Neuroscience* 23: 3251–3261.

Ule J and Darnell RB (2006) RNA binding proteins and the regulation of neuronal synaptic plasticity. *Current Opinion in Neurobiology* 16: 102–110.

Wang H and Tiedge H (2004) Translational control at the synapse. *Neuroscientist* 10: 456–466.

Wu C-WK, Zeng F, and Eberwine J (2007) mRNA transport to and translation in neuronal dendrites. *Analytical and Bioanalytical Chemistry* 387: 59–62.

Rodent Aging

B Teter, University of California, Los Angeles, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

The study of aging has incorporated such vast theoretical and experimental effort that a review of brain aging must necessarily cover a wide conceptual landscape, from evolution to biochemistry, teleology to stochasticism, and genetics to environment. Aging, determined by the entire life history and its evolutionary determinants, is influenced by aspects of fitness, disease, reproduction, mutation, genetics, and random events. Since the brain and aging are the two most complex biological phenomena, it is not surprising that our understanding of brain aging is limited. In regulating and coordinating most somatic processes, the brain is a nexus of control systems linking organ systems and the environment, and while it is rarely the organ of pathological death, its degeneration may be intimately involved in inducing, synergizing, or failing to protect the development of terminal diseases of other organs. The brain, therefore, plays an important role in aging and in life span determination.

Normal aging in the brain, defined in the absence of diseases associated with neuron loss, is not itself associated with extensive age-related neuron loss, but is recognized as sublethal, functional neurodegeneration caused by increased vulnerability to metabolic stress and trauma. Brain aging is influenced by processes and events that begin with genetic and epigenetic contributions from the parents, acting during development, and accelerate in middle age. Aging can be defined to begin when neuron birth and differentiation is surpassed by neurodegeneration that eventually affects cognitive functional ability (approximately after age 10–20 years of human life for most functions).

Animal models of brain aging address select aspects of aging to varying degrees, including animals with phenotypically accelerated aging or prolonged longevity, engineered transgenic, mutant, and knockout models that focus on a single gene's role (and *in vitro* cell and organotypic systems proximally extracted from them), and interventions and environmental manipulations that attempt to modify specific behavioral, anatomic, cellular, or biochemical targets. However, the complexities of aging are not well elucidated in the reductionist approach of animal models. The translational extrapolation to improve the human condition is where animal models receive their greatest challenge and achieve their greatest relevance. Congruencies between humans and rodents justify such extrapolation, particularly at the cellular and gene expression levels. However, evolutionarily, humans and rodents evolved with different life histories (high entropy (equilibrium) and low entropy (opportunistic), respectively) that ultimately determine their respective chronological longevities. This profoundly impacts the strength of inferences drawn from rodent models for the human condition.

Rodents as Models of Aging

Rodent strains that vary in aging processes (qualitatively and quantitatively) and life span have revealed aspects of aging that control health and longevity. Inbred strains have similar or shorter life spans than hybrids (hybrid vigor), in part because they die from strain-specific disease patterns not active in hybrids. Therefore, the more diverse genetic background and greater interanimal variation of hybrids are more relevant to human populations.

It is important to distinguish the processes of aging, disease, and life span (death). 'Normal' aging is defined as time-related change in the absence of disease, yet both can be driven by the same processes and therefore difficult to completely experimentally (or conceptually) dissociate, for example, Alzheimer's disease and normal cognitive decline. Prepathological events play important roles in aging and longevity, including lifetime accumulation of 'events' involving inflammation, infection, and nutrition. Two aspects of brain development could establish the basis for later brain aging effects: genetic antagonistic pleiotropy and cognitive reserve. Antagonistic pleiotropy occurs when genes and events that are adaptively essential for brain development become deleterious with aging, for examples, the progressive decrease in brain plasticity at early ages which serves to focus the mind on the established (and, presumably, understood) environment that continues into adulthood resulting in cognitive decline with age, and age-related changes in inflammatory cytokines which modify responses to stress and reactive oxygen species (ROS) signaling and damage. In studying postdevelopmental influences on aging in mouse transgenic models, artifacts caused by transgene expression during development can be mitigated by the use of conditional-expression transgenics, where gene expression can be experimentally turned on in adult life. The use of cell type and tissue type-specific

transgene expression and controlled genetic knockout (using the Cre-lox system) mitigates global expression issues, as well as developmental issues.

Lifetime random events including spontaneous mutation and epigenetic events can explain variation in longevity of lab rodents and may model human longevity variation. Long-term strain inbreeding within an institution can result in genetic drift that may alter strain-specific phenotypes and responses to manipulations. This random variation can be exploited to develop models of premature brain aging even within an inbred strain, using prematurely aging littermates or substrains that are 'biologically' but not chronologically older (see the section titled 'Accelerated aging models and the oxidative stress theory of aging'), illustrating that the phenotypes of aging and death do not necessarily develop chronologically.

Some lab rodent husbandry practices create selective pressure against disease-sensitive individuals by invoking the mandatory, humane euthanasia of 'suffering' animals in longevity studies. For example, age-related ulcerative dermatitis in C57Bl/6 mice, exacerbated by dominance hierarchy aggression, leads to submissive mice being culled from the population and dominant mice being 'selected.' Further, social housing strongly influences stress levels, both increasing them in male cages and decreasing them in female cages, with effects on learning and memory. The addition of an enriched environment to standard housing conditions can reverse the age-related decline in neurogenesis in the dentate gyrus. Uncontrolled variation in state of the animals at the time of (and, particularly, immediately before) sacrifice can blur effects on immediate-early responses, for example, phosphorylation cascades.

In humans, medicalization of aging (prevention or delay of lethal diseases, including that by hygiene) has contributed to increased longevity, which has also allowed expression of older age-related diseases like Alzheimer's. This raises the profound issue that 'normal' aging is clearly changing in modern humans. A link between lifelong inflammation and nutrition drove the adaptive influence of human diet changes in the distant past to increase longevity (see the section titled 'Apolipoprotein E'); however, this is drastically changing with the recent obesity epidemic in some Western cultures that will actually reduce the expected average human life span. A similar disconnect between evolved life history and current lifestyle of diet and medicalization occurs in laboratory rodents where they age in a stable caged environment at various pathogen-free levels (which limits immune system development and response) and are fed *ad libitum* low-fat, high-calorie, antioxidant diets containing soy estrogens and other compounds that are increasingly recognized for their ability to mask effects or create experimental artifacts.

The Role of the Brain in Aging and Life Span

The brain is a nexus of control systems linking organ systems and the environment that regulates and coordinates most somatic processes. It evolved as a regulator of food acquisition, energy metabolism, stress responses, and reproduction. Lethal disease processes of other organs can be induced or synergized by neurodegeneration. The brain, therefore, plays an important role in aging and in life span determination, for example, brain-specific knockdown of insulin-like growth factor 1 receptor (IGF-1R) increased life span (see the section titled 'Retarded aging and longevity extension models').

Rodent Strain Differences

Rodent strains differ in the pace of behavioral changes with age, and while ultimately phenotypically similar, might implicate fundamentally different processes. The types of strain-specific effects that would best reveal genes regulating behavioral aging are those that differ in both the rate of age-specific mortality and the rate of behavioral decline. For example, strains with shorter life spans like 129/SvJ and FVB mice seem more behaviorally vulnerable (temporal or specific vulnerability) to age-related changes than longer-lived C57Bl/6J. Further, neuronal aging at the levels of cell structure and neurochemistry, while largely studied in C57Bl/6J, show strain specific differences in vacuolar neuritic dystrophy, Ca-binding proteins, inclusion bodies, neurogenesis, neuron cell number in the hippocampal dentate gyrus, various neurotransmitters, and responses to the toxin kainic acid and ischemia.

Behavioral test performance is notoriously strain dependent. Even sublines of the C57Bl/6 strain, the J and Nnia substrains, may differ in age-related decline in Morris water maze learning, and 129/SvJ mice are poor performers in part due to their visual pathology. Further, mice and rats do not perform the Morris water maze in identical ways, particularly regarding the performance parameter that is used in probe trials. As a further testament to the value of rodent models in predicting human responses, the rodent-specific behavioral test, the Morris water maze of learning and memory, has been adapted to a virtual (computer-based) task for testing congruent age-related deficits in humans.

Genetic Contributions

Genetic variation accounts for that about one-quarter of the variance in adult human life span. Several of the genes known to influence rodent life span have important effects in the brain, including insulin-regulated genes and *SIRT-1*, *daf2*, *age1* (PI-3-kinase), *pit1/prop1* (pituitary), growth hormone-related genes, *PARP-1*, and *apoE* (see sections titled 'Retarded Aging and longevity extension models' and 'Glial influences'), to name a few, and many other candidates identified by microarray analysis. Although some senescent phenotypes can be similar between humans and rodents, they are occurring over a 30-fold difference in time; is this simply a quantitative difference in rate or are mechanisms qualitatively different? Gene expression microarray analysis of genes whose expression changes with aging in rat, mouse, and human brain (in different regions) has shown a surprising degree of agreement across these species. Among the few hundred changing genes, as many as 33% were identical between rat and human. These and other genes were in concordant categories of aging-related inflammation, oxidation, and glial processes, as well as novel (not primarily aging-related) categories like developmental regulation, structural plasticity (including *apoE*), signaling, transcription factors, lipid metabolism, and protein processing. This conservation of genomic responses across short- and long-lived species strongly validates the relevance of rodents to study age-related genetic, and by inference, physiologic, changes.

Hormonal Control

Two brain regions that interact through neuroendocrine modulation of learning and memory, the hypothalamus and hippocampus, are particularly central to aging and the control of longevity. Hypothalamic control of gonadal, adrenal, and endocrine hormones impacts functions of and aging in the hippocampus.

The hypothalamus The hypothalamus is the central regulator of endocrine system homeostasis, integrating somatic growth and development driven by pituitary and gonadal hormones, stress responses driven by glucocorticoid hormones, and nutritional and metabolic signals, all central players in the neuroendocrine regulation of life span. The two organ system axes, the hypothalamic–pituitary–ovary (HPO) axis (or more generally, hypothalamic–pituitary–gonad (HPG) and the hypothalamic–pituitary–adrenal (HPA) axis, are dramatically affected by and mediate age-related changes in the brain and fundamentally regulate aspects of somatic aging and longevity. The aging hypothalamus shows gene expression microarray changes in several gene ontogeny categories of neuronal structure, signaling, stress response, and metabolism (also see sections titled 'Accelerated aging models and the oxidative stress theory of aging,' 'Retarded aging and longevity extension models,' and 'Glial influences').

The HPO axis regulates reproductive aging. There is a well-established tradeoff between longevity and reproduction (simplistically recognized by Aristotle as each act of sex shortening life). Reproductive aging is central to life history evolution and the pleiotropic effects of gonadal hormones on brain functions of both neurons and glia. Reproductive senescence is driven by oocyte depletion, which depends on rodent strain and strain hybridization, and on random variation in embryonic oocyte number. Caloric restriction retards ovarian aging by reducing or delaying reproduction (see the section titled 'Caloric restriction'); this may be a selected response, adaptive in times of famine to preserve reproduction for later times of feast when reproduction success is greater.

Human postmenopausal loss of estrogen drives some cognitive changes. Postreproductive female rodents differ from humans most significantly in their elevated estradiol levels. Brain control of and responses to reproductive aging have been investigated using both hormone and genetic manipulation models. Modeling the human postmenopausal estrogen loss in rodents by ovariectomy (OVX) and comparing this to estrogen replacement has shown that estrogen loss accelerates reproductive senescence. Estrogen receptor knockout and transgenic mice have shown the short- and long-term effects of estrogen exposure on neurons and glia. Additional key elements identified in rodent reproductive decline are the age-related decline in follicle-stimulating hormone (FSH), which alters communication between the hypothalamus and pituitary, and the age-related change in luteinizing hormone (LH) secretion and loss of gonadotropin-releasing hormone (GnRH) neurons and systems that control them, which alter the HPG axis.

The HPA axis controls developmental growth and stress responses. This axis controls the lifelong basal and stress-induced glucocorticoid hormone levels, which are elevated with age in several rat strains, but not in C57Bl/6J mice, perhaps reflecting species-specific pathologies. Modeling adrenal hormone action by adrenalectomy (ADX) largely but incompletely eliminates glucocorticoid production. A well-studied target of glucocorticoid action is the hippocampus.

The hippocampus Aspects of age-related changes in hippocampal structure and function are partly

understood in terms of their responses to glucocorticoid and gonadal hormones and their roles in longevity. Rodent models of age-related changes in learning and memory, as assessed by hippocampal long-term potentiation (LTP) and long-term depression (LTD), are dramatically different among rat strains, where F344 show increased LTD with age and outbred Long-Evans show decreased LTD. A canonical change across all mammals is the age-related decline in the *N*-methyl-D-aspartate receptor (NMDAR) subunit NR2B, which may contribute to hippocampal decline in LTP and learning.

Neuron number in the hippocampus is relatively preserved with aging across all mammals; small changes are relatively region specific and rodent strain dependent. Historically, conflicting data are attributed to nonstandardized cell counting methods. Therefore, hippocampal atrophy with aging is attributed to the neuron cell atrophy and neurite and synapse loss. Neurogenesis of immature hippocampal granule cells shows age- and exercise-dependent changes in rodent models.

The entorhinal cortex input to the hippocampal dentate granule cells degenerates with age in humans and rats, which causes compensatory neurite sprouting in the hippocampus. This is experimentally modeled in rodents, albeit acutely, by various afferent lesions, and this recapitulates aspects of normal aging, including modulation by estrogen and impairment in aging female rat. Age-related decline in sprouting is due to defects in the aged target to which the sprouting neurite synapses, not to defects in the aged sprouting neuron, as shown in the model of implanting superior cervical ganglion tissue near the hippocampus and reciprocally varying the age of the ganglionic tissue (neurite source) or the host (hippocampal target) animal.

Estrogen stimulates neuronal function at many cellular levels, including neurogenesis, sprouting of neurites and spines, and molecular pathways. For example, aged rats do not respond to estrogen with increased spine formation in hippocampal CA1 neurons, perhaps simply due to receptor downregulation. Glucocorticoids cause neuronal and glial changes in the hippocampus of humans and rodents that contribute to cognitive changes during aging. Modeling age-related glucocorticoid exposure in rodents includes both acute and chronic treatment (e.g., hormone pellet implantation in the brain), ADX, and genetic knockout of a cortisol-regenerating enzyme. Effects and correlates of these manipulations include hippocampal atrophy, neuron pyknosis (cell body atrophy) and cell loss (pyramidal neurons), decreased neurogenesis, glial activation (hypertrophy, altered gene expression including protective changes in glial fibrillary acidic protein (GFAP) and transforming growth factor β-1 (TGFβ-1), and gliogenesis (see the section titled 'Glial influences')), and concomitant learning and memory impairments. Glucocorticoids are also the target of paradoxical regulation by caloric restriction (see the section titled 'Caloric restriction'). Rat strains differ in age-related losses of corticoid receptors in the hippocampus. The effects of developmental experience and stress on age-related hippocampal impairments are modeled in a paradigm of neonatal handling and maternal deprivation that results in reduced cortical brain-derived neurotropic factor (BDNF) in the adult rat. Therefore, uncontrolled differences in lab rodent rearing could also contribute to contradictory reports of neuronal loss with aging.

Accelerated Aging Models and the Oxidative Stress Theory of Aging

Accelerating aging models, while showing phenotypes similar to normal aging, are not necessarily driven by the same mechanisms as aging. They display only a subset of normal aging phenotypes, underscoring the greater complexity of normal aging, and show a preponderance of only one or a few disease pathologies. For example, genetically autoimmune mice, which challenge the historical doctrine of brain immune privilege by the blood–brain barrier, show age-related cognitive disability in acquisition of avoidance response that is related to the age of onset of their autoimmunity (biological age), while their retention deficit is related to chronological age.

Oxidative Stress Models and ROS

Normal brain aging is associated with decreased homeostatic reserve (allostatic load) in the metabolic triad of mitochondria-ROS-intracellular Ca. ROS is a molecular example of antagonistic pleiotropy, being controlled during normal development for its signaling effects, for example, those regulating neuroplasticity, but becoming deleterious in brain with age, due to declining resistance to its detrimental actions and its increased production (free radical theory of aging); these differ between humans and rodents. There is an optimal ROS level; therefore, drastic reductions in ROS metabolism, either genetically or chemically with high dose chronic antioxidants, can block its beneficial activities. Genetic models of oxidative stress in the aging brain include knockout of the excitatory amino acid carrier-1 gene which impairs glutathione metabolism, knockout of the p66-Shc gene (an adaptor/signaling protein regulating ROS metabolism) which extends life span 30% and prevents increased emotionality and pain sensitivity, and knockout of the hexosaminidase B gene which models

the accelerated aging in human Sandhoff disease caused by lysosomal glycosphingolipid storage dysfunction, producing premature neurodegeneration involving microglia activation and oxidation, and which also partially responds to caloric restriction.

The Senescence-Accelerated Mouse

The senescence-accelerated mouse (SAM) was established by phenotypic selection of spontaneous, prematurely senile, and short-lived littermates of the AKR/J strain; SAM-prone (SAMP) and SAM-resistant (SAMR) littermates were separately inbred. The SAMP8 and/or SAMP10 show age-related learning and memory deficits, mood disorders, other cognitive impairments with associated biochemical and synaptic changes, peripheral amyloidosis, altered blood–brain barrier, region-specific brain atrophy (mostly in cortex, but not hippocampus), and neurotransmitter-specific (cholinergic) neuron atrophy. Cognitive impairments in SAMP8 are correlated with oxidative stress and are corrected by antioxidants, implicating fundamental, age-related oxidative stress mechanisms involving defense enzymes (superoxide dismutase (SOD), glutathione, catalase), with downstream effects on neurotrophic factor expression (glial-derived neuronal factor (*GDNF*), neurotrophin-3 (*NT-3*), nerve growth factor (*NGF*)), lipid peroxidation, protein oxidation (as detected by proteomics), and defects in energy metabolism. Microarray and specific gene analyses show involvement of oxidative and stress response genes, including those regulating signaling, energy metabolism, immune response, cytoskeleton (increased tau and its hyperphosphorylation), and decreased apoE and glucocorticoid receptor subunit expression.

Retarded Aging and Longevity Extension Models

Models that increase longevity may be more informative about aging mechanisms than accelerated aging models, because increased population life span is likely to be achieved by modulating multiple age-related mechanisms, particularly in hybrid or outbred strains that die of multiple causes (see sections tilted 'Rodents as models of aging' and 'Caloric restriction'). With the definition of normal aging that excludes disease, the distinction between slowing aging and inhibiting or delaying disease must be carefully considered, since preventing one relatively common lethal disease can apparently extend population life span without modifying fundamental aging processes.

Genetic Models

Several genetic models with increased longevity involve the insulin/IGF-1/growth hormone (GH) pathways, suggesting that chronological longevity is extended by inhibiting growth pathways and by increased protection from stress. Life span extension has been achieved in mice by genetic mutations in the genes including p66, *FIRKO* (insulin receptor), and *IGF-1R* (in females only). IGF-1 control of brain cell growth, neuronal myelination and sprouting, and astrocyte activation (including GFAP induction) in development and injury-response regeneration has been shown to depend on the cellular site of its expression, by comparing neuron-specific and astrocyte-specific inducible-expression transgenics. In mice, the specific role of brain expression of IGF-1R on life span and mortality was elegantly shown by genetic knockdown of the IGF-1R gene (by brain-specific Cre-lox-mediated heterozygous knockout), producing a brain hypo-sensitive to IGF-1 that had increased life span and decreased mortality.

The pituitary is specifically implicated in growth-associated longevity effects. Various hypopituitary long-lived dwarf mice (Ames and Snell) support the model of antagonistic pleiotropy in normal brain aging and age-related decline in cognition by counteracting this effect through delayed puberty. *GH* knockout mice have preserved learning and memory even with aging, while *GH* transgenic mice age faster with earlier puberty and faster cognitive decline. Dwarf and GH-resistant mice (*GH* receptor knockout) implicate multiple complex mechanisms and abnormalities, including insulin supersensitivity (decreased insulin and hippocampal *IGF-1* expression), ROS, lipids, carbohydrates, stress, and age-related disease. Many effects are gender, diet, and strain dependent.

Caloric Restriction

Caloric restriction (CR) is a well-established method of life span extension with significant, possibly contributing effects in the brain. CR may represent a more natural diet for rodents and *ad libitum* may cause overfeeding-induced stress (see the section titled 'Rodents as models of aging'). Its intense study has informed that energy metabolism (glucose and ketones), signaling pathways of insulin, IGF-1, and GH, and ROS metabolism impact fundamental aging processes as important subsets of normal aging changes. CR slows fundamental processes of aging that affect the rate of development of multiple diseases, although other mechanisms are implicated in brain. Life span extension by CR is influenced by strain-dependent pathologic diseases and lesions and by gender, including variation in reproductive aging and the hypothalamic and pituitary control of estrus cycling.

CR extends life span in all organisms up to rodents. However, in primates and humans, fundamental

differences in life history evolution and the set points of metabolic stability and ROS metabolism have argued against predicted efficacy. A human clinical trial was initiated in 2002 by the National Institute on Aging (NIA) (called CALERIE) which has shown some neuroendocrine and pyschophysiologic benefits. CR mimetics or intermittent (hormetic) CR may be much more practical for humans, in particular for brain aging. Mimetics of CR can target glucose, insulin, oxidation, and inflammation and can be quickly identified by their induction of short-term (8-week) CR-like gene expression profile responses. The NIA tests candidate compounds for mean and maximal life extension and/or health improvement in hybrid mice with genetic heterogeneity (HET mice) and pathologic diversity, which will include cognitive testing.

Many of the beneficial effects of CR can be linked to effects on the nervous system with specific effects on age-related changes in neuroendocrine and sympathetic systems, glia, vasculature (constriction and microvessel loss), hippocampal volume (gliogenesis, but not neurogenesis), synaptic efficacy, and neurotrophic factors, particularly BDNF. CR attenuates or reverses several age-related glial responses, including the increase in hippocampal astrocyte activation and *GFAP* expression (see section titled 'Glial influences'), the region-specific modulations in *apoE* and *apoJ* mRNA in hybrid rats, and microglial activation, by several mechanisms including reducing proinflammatory inducers like advanced glycation end products and promoting anti-inflammatory effects of cortisol on cytokine gene expression. Glucocorticoids are paradoxically increased with CR in rats in response to reduced plasma cortisol binding globulin, which suggests that the neuroprotective effects of CR (including increased BDNF and attenuated microglial activation) outweigh deleterious effects of increased glucocorticoids; however, generalizations are limited because rodent species differ in the relationship between age-related cortisol and cognitive impairment. Further, while CR protects some age-related neuron loss in the periphery (enteric and retinal, and motor neurons), effects of CR on brain neuron number, particularly in relation to elevated glucocorticoids, have not been assessed. Limited studies show that CR has opposite effects on different neuronal population in the hypothalamus, protecting the age-related loss of IGF-1R neurons but inducing the loss of other populations, and in the hippocampus, CR does not impact age-related decline in neurogenesis of granule cells but does promote the survival of newborn glial cells. CR impacts aspects of oxidation and stress responses, including inducing the neuroprotective actions of heat shock protein-70 (HSP-70) and attenuating its age-related loss that involves Ca, ROS, and cell death. Proteomics of CR effects in brain have identified the striatum and hippocampus as particularly benefiting, with lowered oxidized proteins. Several neurodegenerative disease models show beneficial effects of CR, including region-specific neurotoxicity of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridire (MPTP) (substantia nigra) and kainate (hippocampus), both dependent on BDNF. The BDNF knockout (conditional) mouse shows alterations in several CR-relevant systems including stress, glucose, and insulin.

A central question is whether CR impacts the same mechanisms that drive normal brain aging. Comparison of age- and CR-related gene expression profile changes in brain showed that most of the changed neurotrophic-like genes were dissimilar, suggesting that CR achieves brain effects and perhaps longevity by mechanisms distinct from or only partially overlapping normal aging. A gene that could be central to mediating CR effects on longevity and in models of age-related neurodegeneration is sirtuin 1 (*Sirt1*), the homolog of the yeast longevity gene *sir2*, which regulates genes involved in nutrient sensing, cell cycle, and DNA repair.

Glial Influences

Astrocyte and microglial phenotypes vary in a continuum from resting to reactive, playing both protective and deleterious roles in neurodegeneration and in aging. A major controversy is whether age-related changes in glia induce neurodegeneration or are simply a response to aging-induced neurodegeneration. The inability to co-localize activated microglia with neurodegeneration argues against the latter hypothesis, in some models. Microglial number and activation state increase with age across all mammals; in one strain (C57Bl/6NNIA), glial cells increased only in females. Astrocyte activity affects the permissiveness for neuroplasticity and repair in the brain, as suggested by the age-dependent activities of astrocyte genes *GFAP* and *apoE*. They act in membrane structure (GFAP) or in lipid trafficking and signaling (apoE), contributing to the multiple pathways for regeneration in the brain and its age-related manifestations, including the suppression of reactive gliosis by estrogen regulation of their genes.

Glial Fibrillary Acidic Protein

GFAP is an intermediate filament protein integral to cytoskeletal dynamics regulating the morphology and function of astrocytes including astrocyte reactive responses during aging. GFAP is consistently upregulated with age across rodents and humans and is a surrogate marker of aging in the brain, as indicated

by unbiased microarray analysis of various brain regions across rat, mouse, and human species. The conservation of transcription control elements in the *GFAP* promoter across rat, mouse, and human species implicates this gene in canonic features of brain cell aging across both short- and long-lived species. *GFAP* also responds in caloric restriction (see the section titled 'Caloric restriction'). Human and mouse GFAP protein have similar activities, as shown by the phenotypes of mice whose *GFAP* coding region is replaced by the human *GFAP* sequence.

In modeling the age-related increase in *GFAP* expression, *GFAP* overexpressing mice show an earlier onset of the gene expression profile changes that characterize normal brain aging in mice and age-related changes that suggest a temporal relation between astrocyte and microglial activation where the former precedes the latter. The profile of expression responses shows increased stress, immune activation, oxidation, and inflammation and decreased neuronal plasticity and development genes, indicating that increased *GFAP* expression can initiate a cascade of changes that affect all other cell types in the brain. Age-related phenotypic changes in astrocyte reactivity, including increased *GFAP* mRNA, persist in *in vitro* astrocyte culture and affect neurite sprouting of co-cultured neurons. This supports the direct role of age-related astrocyte dysfunction in neurodegeneration with age.

GFAP regulates neurite and synaptic morphology, since various means of reducing its expression increases neurite sprouting. *GFAP*-knockout (ko) mice show enhanced neurite sprouting and (in combination with vimentin-ko) show increased posttraumatic regeneration and neurogenesis, an effect reproducible in an *in vitro* lesion model. *GFAP*-ko mice show abnormal, late-onset myelination, which models aspects of the human disease (Alexander disease) caused by *GFAP* mutation. Estrogen downregulation of *GFAP* controls its activity in regulating morphologic changes of astrocyte end foot processes that displace hypothalamic neuron synapses or neurites that in turn regulate endocrine hormone release during the estrus cycle. Loss of gonadal hormones with age may result in increased reactive gliosis and reduced neuroplasticity, as shown in models of stimulated neurite sprouting in both *in vitro* and *in vivo* lesion recovery models using OVX and estrogen replacement.

Apolipoprotein E

ApoE is a major lipoprotein that primarily regulates lipid metabolism but has pleiotropic activities. Of the three common human alleles, *apoE2* is rare and is associated with increased longevity, *apoE3* is the most common, and *apoE4* strongly increases risk for several diseases and disorders, including cardiovascular disease and neurodegenerative disorders, particularly Alzheimer's disease (AD). Extensive human studies show apoE4 effects on milder and mid-life aspects of cognitive impairment, which are modeled in apoE4 transgenic mice.

Human *apoE* alleles evolved as part of changing life history evolution, where the ancestral *apoE4* allele was adaptive for the lower calorie, lower fat diet of early humans, while the common allele *apoE3* became adaptive with increased dietary meat by decreasing cholesterol metabolism. This may underlie the current pandemic of obesity-induced cardiovascular disease caused by the mismatch between current diet and exercise trends and the adapted apoE3 activities and maladaptive apoE4 activities in lipid metabolism. In addition to causing a decline in human longevity, this could also alter 'normal' human brain aging.

ApoE knockout and human isotype transgenic mouse models ApoE function has been studied using the *apoE* knockout mouse (*APOE*-ko) and transgenics expressing human apoE isotypes, usually on the background of mouse *APOE*-ko. While *APOE*-ko and rare humans lacking apoE develop to adulthood with relatively minor cognitive impairments, in the aging brain, compensatory pathways may become ineffective or insufficient, increasing the reliance on apoE activity to maintain plasticity, revealing defects in apoE4. *APOE*-ko begin to show neurodegenerative changes at 5–6 months, with defective LTP and loss of dendritic cytoskeleton and synaptic markers; some effects are background strain dependent.

A major issue in understanding apoE biology in the brain is defining which cell types produce apoE, which is determined by the transcription promoter regulation. Humans express apoE in both glia and at low levels in neurons. Mouse apoE is expressed in basal conditions only in glia, but can be induced in neurons by cytotoxic challenge. There is a critical functional trafficking and cellular localization difference between receptor-mediated uptake of apoE and actual production of apoE within the neuron. The human-like neuron expression of apoE is replicated in mouse transgenics driven by the human apoE promoter (human genomic promoter, HGp transgenics). To evaluate which cellular production site is relevant to detrimental effects of apoE4, transgenics driven by cell type-specific promoters limit expression to neurons using the neuron-specific enolase (NSE) promoter (NSEp transgenics) or to astrocytes using the GFAP promoter (GFAPp transgenics). The caveat with these mice is that the cell-specific expression of

apoE is either abnormally high in NSEp transgenics or unnaturally regulated in GFAPp transgenics. This caveat is partly addressed by the newest transgenics in which the endogenous mouse apoE promoter drives expression of the human apoE isotypes (the targeted replacement transgenics, TR). These mice are ideal in that they recapitulate the human brain region-specific levels of apoE, and equivalent levels among the three isotypes, but do not show the human-like neuronal basal expression of apoE.

All human apoE isotype transgenic mice (except TR) are maintained on the, *APOE*-ko background, which has severe peripheral hypercholesterolemia with a potential influence on brain and blood–brain barrier functioning. The brain cell-specific expression transgenics (GFAPp and NSEp) maintain the peripheral hypercholesterolemia, while the genomic transgenics (HGp) correct it; although TR mice are not on the, *APOE*-ko background, they show marked isotype-specific differences in peripheral apoE levels. *APOE*-ko show disturbed blood–brain barrier permeability and increased susceptibility to microvascular disease, which contributes to the more severe responses to acute ischemia and stroke models by, *APOE*-ko and apoE4 transgenics (HGp and TR), and which is relevant to similar responses in humans and to the regional glucose utilization defects seen in human positron emission tomography (PET) studies.

Considering the definition of 'normal' brain aging and the role of apoE isotypes, there is an important distinction in transgenic models between what is normal for a mouse (with mouse apoE) and what is normal for a human (defined by the phenotype of the common allele apoE3). Comparisons of wild-type C57Bl6/J mice with isogenic apoE3 transgenics (on, *APOE*-ko) have revealed some similarities in phenotypes, but also some very important differences indicating that mouse apoE is not consistently like human apoE3 or apoE4. However, some clear congruencies between apoE transgenics and humans have been identified, for example, apoE4 in elderly humans and middle-age transgenic mice both show reduced hippocampal neuron dendritic spine density.

Compared to, *APOE*-ko mice, apoE4 shows phenotypic gain-of-negative activities in inhibiting neurite sprouting, mortality from head injury, and amyloid deposition, and loss-of-function activities in inhibiting astrogliosis, age-related dendrite loss, age-related synapse marker loss, resistance to excitotoxicity, and defective LTP. For age-related memory loss, neuronal expression (NSEp) shows a loss-of-function while astrocyte expression (GFAPp and TR) show a gain-of-negative function. This phenotypic distinction has important implications for the design and administration of drugs that target apoE activity or expression in that a gain-of-negative function of apoE4 would be exacerbated by a drug that increased apoE expression, while a loss-of-function would be rescued. Indeed, pharmacologic attempts to correct age-related disease processes including cognitive disorders, for example, lipid normalization by statins, antioxidants, hormone replacement therapy, fish oil, insulin sensitizers, and other treatments for AD, have shown apoE genotype-dependent efficacy, where apoE4 are nonresponsive.

Effects of human ApoE isotypes on aspects of brain aging ApoE has pleiotropic effects on brain and neuronal functioning, including lipid metabolism, neurite sprouting, gender-specific effects, multiple signaling cascades, LTP, inflammation, oxidation, neurotoxicity, and neurotrauma. ApoE4 accelerates the age-related change in the differential distribution of lipids between cytofacial and exofacial leaflets of the bilayer, which affects membrane fluidity and lipid raft microdomains. Rodent models of brain aging should carefully consider dietary lipid composition for its effects both developmentally (maternal diet) and in aging, particularly the age-related reductions in the essential omega-3 fatty acid docosahexaenoic acid (DHA) (see the section titled 'Rodents as models of aging'). *APOE*-ko and apoE4 show defective neurite sprouting and other aspects of neuronal plasticity, including age-related loss of synaptic markers, shown at levels from *in vitro* neuron cultures to organotypic cultures of hippocampal slices (derived from apoE isotype transgenic mice), to *in vivo* reactive neurite sprouting responses in all the promoter forms of transgenic mice. This effect models aspects of human aging in cognitively normal individuals where dentate spine loss is greater, and compensatory sprouting (in AD) is reduced or nonexistent in apoE4.

Gonadal hormone regulation of the apoE gene and its interaction with the effects of the apoE4 isotype are a paradigm of life history evolution caused by adaptive changes in diet and reproduction (somatic theory of aging). Gonadal hormones, particularly estrogen, modify apoE4 effects, both in atherosclerosis and in cognitive disorders, resulting in a greater susceptibility of females in humans and in mice to the age-exacerbated impairments in learning and exploration seen in NSEp and GFAPp mice. This occurs in brain regions that undergo estrus cycle synaptic remodeling. Conversely, estrogen's neurotrophic effect of stimulating neurite sprouting is modified by the age-dependent increased apoE expression, as shown in a rat OVX/estrogen replacement model of sprouting where the absence of apoE (*APOE*-ko mice) prevents estrogen stimulation of neurite sprouting. In addition, effects of

male androgens on learning, memory, and recognition are altered by apoE4 as shown by castration studies (NSEp) and in androgen receptor functions.

Inflammation, like lipid metabolism, is a common feature of apoE-related mechanisms in both vascular and cognitive disease with the involvement of the common cellular element, the macrophage or microglial cell, respectively; however, it is important to distinguish between adaptive versus pathological effects. ApoE is generally anti-inflammatory, its production being repressed during the cytotoxic phase of the innate immune response and promoted during the repair phase, and inflammatory challenges including age-dependent changes in the HPA axis have more severe behavioral effects (memory and anxiety) in *APOE*-ko and apoE4 mice. Inflammatory effects of apoE4 (TR) implicate activated microglial and astrocyte production of ROS, glutamate, proteases, and inflammatory immune factors like interferon gamma (INFg), NO, tumor necrosis factor-α (TNF-α), and interleukin 1 beta (IL-1b), and their feedback regulation on apoE production, including regulation by anti-inflammatory factors TGF-β and estrogen, and reversal of anti-inflammatory effects of testosterone.

Neurotoxicity of apoE4 is linked to several of its other pleiotropic activities, for examples, proteolytic fragments of apoE4 are formed and are neurotoxic only when apoE4 is expressed in neurons (NSEp) but not in astrocytes (GFAPp), hyperphosphorylation of tau (NSEp), poor antioxidant activity (where $E4 < E3 < E2$), greater microglial activation, and differential effects on cell signaling including extracellular signal-regulated kinase (ERK) and AKT. BDNF and its signaling effects are implicated in apoE4 effects on cognitive responses to environmental enrichment and head trauma in transgenic models.

Conclusions and Relevance to Human Brain Aging

Rodent models of aspects of brain aging provide valid tools to understand the roles of select components and their interactions. However, studying the intersection of the two most complex biological phenomena is not well elucidated in the reductionist approach. Rodent models are strongly validated by the multiple levels of congruence with human brain aging, largely due to their common phylogeny. However, significant differences preclude the use of lab rodents to fully understand human aging, particularly fundamental differences in dietary energy sources and utilization, reproductive aging, and temporally accumulated stochastic events that feed evolutionarily established responses of antagonistic pleiotropy in middle and later ages. Differences even between and within rodent species and strains imply that human–rodent differences are much greater than is currently understood. Genetic manipulations intended to target adult-related processes can have developmental artifacts, particularly in gene knockout and quasi-controlled transgenics. Conditional knockouts and expression-controlled transgenics eliminate such developmental artifacts and control the temporal and dose-dependent influences of gene expression.

See also: Aging of the Brain; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Gene Expression in Normal Aging Brain; Genomics of Brain Aging: Apolipoprotein E; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Hippocampus; Hippocampus: Molecular Anatomy; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Lipids and Membranes in Brain Aging; Long-Term Depression: Cerebellum; Long-Term Potentiation (LTP); Microglia Properties; Microglial Response to Injury; Synaptic Plasticity: Learning and Memory in Normal Aging; Transgenic Models of Neurodegenerative Disease.

Further Reading

Christensen K, Johnson TE, and Vaupel JW (2006) The quest for genetic determinants of human longevity: Challenges and insights. *Nature Review Genetics* 7(6): 436–448.

Conde JR and Streit WJ (2006) Microglia in the aging brain. *Journal of Neuropathology and Experimental Neurology* 65(3): 199–203.

Crimmins EM and Finch CE (2006) Infection, inflammation, height, and longevity. *Proceedings of the National Academy of Sciences of the United States of America* 103(2): 498–503.

Demetrius L (2006) Aging in mouse and human systems: A comparative study. *Annals of the New York Academy of Science* 1067: 66–82.

Finch CE (2003) Neurons, glia, and plasticity in normal brain aging. *Neurobiology of Aging* 24(supplement 1): S123–127, S131.

Holliday R (2006) Aging is no longer an unsolved problem in biology. *Annals of the New York Academy of Science* 1067: 1–9.

Ingram DK and Jucker M (1999) Developing mouse models of aging: A consideration of strain differences in age-related behavioral and neural parameters. *Neurobiology of Aging* 20(2): 137–145.

Ji LL, Gomez-Cabrera MC, and Vina J (2006) Exercise and hormesis: Activation of cellular antioxidant signaling pathway. *Annals of the New York Academy of Science* 1067: 425–435.

Kirkwood TB and Shanley DP (2005) Food restriction, evolution and ageing. *Mechanisms of Ageing and Development* 126(9): 1011–1016.

Longo VD and Finch CE (2003) Evolutionary medicine: From dwarf model systems to healthy centenarians? *Science* 299 (5611): 1342–1346.

Martin B, Mattson MP, and Maudsley S (2006) Caloric restriction and intermittent fasting: Two potential diets for successful brain aging. *Ageing Research Reviews* 5(3): 332–353.

Martin GM (2005) Genetic engineering of mice to test the oxidative damage theory of aging. *Annals of the New York Academy of Science* 1055: 26–34.

McEwen BS (2000) Allostasis, allostatic load, and the aging nervous system: Role of excitatory amino acids and excitotoxicity. *Neurochemistry Research* 25(9–10): 1219–1231.

Nadon NL (2006) Exploiting the rodent model for studies on the pharmacology of lifespan extension. *Aging Cell* 5(1): 9–15.

NIA Interventions Testing, Program to test candidate compounds for mean, maximal life extension and/or health improvement in hybrid mice. http://www.nia.nih.gov/ResearchInformation/ScientificResources/ITPapp.htmand. http://grants1.nih.gov/grants/guide/notice-files/NOT-AG-06–001.html, (accessed March 2007).

Pan F, Chiu CH, Pulapura S, et al. (2007) Gene Aging Nexus: A web database and data mining platform for microarray data on aging. *Nucleic Acids Research* 35: D756–D759.

Teter B and Finch CE (2004) Caliban's heritance and the genetics of neuronal aging. *Trends in Neuroscience* 27(10): 627–632.

Warner HR (2005) Longevity genes: From primitive organisms to humans. *Mechanisms of Ageing and Development* 126(2): 235–242.

Rodent Behavior: Approaches

L H Tecott, University of California, San Francisco, School of Medicine, San Francisco, CA, USA
E J Nestler, The University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Understanding the neural mechanisms of complex behavior, as well as the abnormalities in behavior that accompany psychiatric and neurologic disorders, represents one of the most formidable challenges in biomedical research today. Our very limited knowledge of the neural basis of behavior means that this endeavor must rely, to a unique extent, on *in vivo* studies. The human and mouse genome projects, coupled with an impressive array of molecular genetic technologies to manipulate expression of individual genes within the brain through mutagenesis and viral-mediated gene transfer, have provided unprecedented opportunities for investigating the influence of genes on behavior. In contrast, behavioral assays, and their application to large numbers of animals, have not kept pace in terms of throughput with the rapidly escalating use of genetic manipulations in mice. There is also a need for higher-throughput behavioral assays to screen increasing numbers of potential medications targeted to an increasing array of neural proteins. Similarly, the difficulty of current central nervous system (CNS) drug discovery strategies, where behavioral analysis of a potential medication comes late in the process, to develop psychiatric medications of fundamentally new mechanisms of action has led to the suggestion that behavioral testing should occur sooner. This, too, would require tests of increased throughput. Here, we briefly describe this challenge and review some of the innovative solutions that are in development to relieve the bottleneck in behavioral assessment. Such approaches may not only improve throughput but also provide new insights into the regulation of rodent behavior that are not approachable with conventional assays.

Types of Behavioral Assessments

Behavioral assays in psychopharmacology have been categorized into three classes: animal models of clinical disorders, behavioral screening tests, and behavioral bioassays. Animal models of clinical disorders are attempts to simulate symptom clusters characteristic of particular diseases. Screening tests are used to assess the impact of genetic and pharmacological manipulations on behaviors chosen to reflect particular behavioral processes of interest. For example, the forced swim test and elevated plus-maze are used to screen compounds and genetic mutations for their relevance to depression or anxiety, respectively. Alternatively, some behaviors are used as bioassays to assess the activity of particular neural pathways. Examples of such behavioral bioassays include circling behavior to test function of the brain's dopaminergic systems and head-twitch responses to serotonin receptor agonists as a measure of serotonin 5-HT2 receptor sensitivity.

Multiple behavioral assays have been used for decades to learn a great deal about the neural basis of behavior. It has become apparent, however, that there are fundamental limitations in their application to neural processes related to psychiatric and neurologic disorders and CNS drug discovery. Although *bona fide* animal models of neurologic diseases are becoming increasingly available as known disease-causing genetic factors are reproduced in mice, such models are not available for most psychiatric conditions because of the relative lack of knowledge of the specific genetic as well as nongenetic factors that cause these illnesses. In addition, it has become increasingly clear that accurate interpretation of individual behavioral tests can only be achieved in the context of a thorough set of behavioral assessments, as described in the next section.

Animal Models of Psychological Processes

Genetic and pharmacological influences on psychological processes must be inferred from their effects on observable behavior, but several factors complicate our ability to draw inferences about a rodent's psychological state from behavioral data (Table 1). The elevated plus-maze and Morris water maze are good examples of this. The plus-maze, a commonly used test of anxiety-related behavior, measures a rodent's avoidance of open or brightly lit places. In this test, animals are given a choice between narrow elevated platforms that are either walled or open. The predisposition of animals to prefer the walled arms is thought to correlate with anxiety state. The assay has some predictive validity for the identification of compounds with anxiolytic actions in humans. For example, benzodiazepines such as diazepam, which are anxiolytic in humans, increase the rodents' time in the open, brightly lit areas. Like many rodent behavioral assays, the plus-maze was originally validated in rats and more recently adapted to mice.

Table 1 Experimental approaches to behavioral assessment

Approach	*Advantages*	*Disadvantages*
Standard tests	Extensive pharmacological validation available Relatively simple and rapid to administer	Risk of data misinterpretation Limited etiologic and pathophysiologic validity Certain assays have low throughput Serial tests have very low throughput and require large numbers of animals
Batteries of standard tests	Pharmacological validation available Reduced numbers of animals required Data interpretation improved by assessing diverse behaviors	Limited etiologic and pathophysiologic validity Certain assays have low throughput Test order may confound results
Naturalistic assays	Behaviors assessed in a more ethologically valid context Reduces misinterpretations of behavioral responses Wide range of complex species-characteristic behaviors may be elicited	Low throughput May be time-consuming to establish assay conditions for certain behavioral processes Lack of widespread experience with these approaches for mice Requires extensive pharmacological validation
High-throughput screens	Capable of very high throughput Potential for detecting novel behaviors	Limited etiologic and pathophysiologic validity Some rely on limited tests, increasing incidence of false positives and false negatives Require extensive pharmacological validation
Quantitative home cage behavior analysis	Effects of handling and novelty minimized Assesses multiple behaviors over long time intervals Potential for detecting novel behaviors Relatively low labor costs	Need for chronic housing links throughput to number of available cages Requires substantial computational resources May be time-consuming to establish assay conditions for certain behavioral processes Requires extensive pharmacological validation

This is also the case for the water maze, a commonly used assay of spatial learning and memory. In a typical procedure, animals are placed within a circular pool of opaque water and must swim to a submerged platform. If the platform position remains constant, the animal's performance improves over the course of repeated trials, as indicated by reduced latencies to find the platform. Presumably, efficient performance in this task reflects a spatial memory process engaging a navigation strategy guided by visual cues. Spatial learning is verified in a subsequent 'probe trial,' in which the animal swims during a trial with the platform removed; preference for the prior platform location is viewed as a measure of spatial learning. Drugs that disrupt cognitive function impair performance in the water maze. However, the ability to demonstrate the converse – improvement of cognitive performance – is limited based on the current lack of availability of potent cognitive enhancing agents, with today's agents producing modest effects in the water maze, as they do in humans as well.

There are several sound reasons for the popularity of the elevated plus-maze and Morris water maze. Compared with other assays of similar psychological processes, these tests are simple and rapid to perform. The tests also have been pharmacologically validated with available agents, as mentioned previously. Moreover, a large literature addresses the contributions of neural circuits that regulate an animal's performance in these assays. Nevertheless, there are several limitations in the way in which such tests are sometimes interpreted. One limitation relates to a lack of accepted and standardized protocols. Interlaboratory variations often exist in both the testing apparatus and the procedures used. Often, results from one group cannot be compared to the results from another group because of such methodological differences. A second limitation is that the results of individual behavioral assays may be readily misinterpreted when they are conducted in isolation. For example, an undetected cognitive deficit interfering with the processing of contextual cues could cause abnormal performance in the elevated plus-maze. Potential perturbations of sensory perception (e.g., blindness), activity levels, or motor coordination could also lead to misleading results. Analogously, performance in the water maze can be influenced by many factors, including prior testing experience, motor impairments, abnormal stress responses, motivational state, visual problems, and perturbed thermoregulation. These considerations underscore the importance of not relying on any one assay in making a claim regarding the psychological basis of a behavioral phenotype. The evaluation of a diverse array of

behavioral domains improves the extent to which each individual assay may be correctly interpreted.

One common approach to obtaining a comprehensive behavioral assessment is to run cohorts of mice through batteries of multiple behavioral assays (Table 1). Examples include the SHIRPA protocol and PsychoScreen (PsychoGenics, Inc.; http://www.psychogenics.com), which involve an initial neurologic assessment followed by a series of standard tests of motor function, analgesia, stress responses, and cognition. Although this is a very important and useful approach, particularly given the limited numbers of mutant mice or amount of novel compounds typically available, behavioral test batteries present an additional set of caveats. First, they are time-consuming and labor-intensive, limiting their throughput, as discussed later. In addition, the extent to which prior behavioral testing influences test performance is often not examined. For example, studies have compared the performance of naive mice in the Morris water maze and elevated plus-maze with the performance of animals in which these tests were administered within the context of a behavioral battery. In one case, mice tested in the water maze within a battery did not show spatial selectivity for the platform location, whereas naive mice did so. In the elevated plus-maze, C57BL/6J mice displayed major differences in open platform time depending on whether they were tested within a battery or in the naive state. In the latter study, the test battery did not alter elevated plus-maze behavior in 129S2/Sv mice, indicating that testing order can interact with genetic background to further complicate the interpretation of behavioral phenotype data. Thus, limitations of individual assays and behavioral batteries of tests highlight the need for innovation.

Relevance of an Ethological Perspective

One way to improve the interpretation of behavioral tests is to consider them within an ethological perspective. As developed in the mid-twentieth century by investigators such as Konrad Lorenz, Niko Tinbergen, and Karl von Frisch, the discipline of ethology focuses on detailed analyses of the diverse sets of behaviors seen in animals in their natural habitats. Connotations of the term 'ethological' vary widely among investigators, and some would find it inappropriate to apply the term to studies in laboratory settings with genetically defined animals that do not exist in the wild. Regardless of whether the term may be applied to such studies, their interpretation may nevertheless be enhanced by insights that have been provided by ethology. Often, standard behavioral assays are considered in a rigid manner without regard to species-typical behaviors seen during testing. For example, mice placed directly on the open arm of the elevated plus-maze may show freezing behavior, recognized by ethologists as a characteristic response of rodents to threatening stimuli. Although this could result in increased open-arm time, it would be a mistake in this case to conclude that the animal showed reduced anxiety-like behavior. Ethological considerations can also complicate attempts to adapt rat behavioral assays for mice. As just one example, differences in the natural habitats of rats and mice have been implicated in the relatively poor performance of mice in the Morris water maze procedure. Given that rats may be better adapted to wet habitats, their superior performance in this task may relate more to task suitability than to cognitive differences.

An additional aspect of rodent behavior that is receiving increased attention relates to the fact that the behaviors of both rodents and humans are influenced by their social organization. For example, group housing of male mice can result in the formation of dominance hierarchies, which can affect many behaviors and provides an additional source of variability in behavioral studies. Attempts are underway to develop approaches to the efficient testing of interactions between same- and opposite-sex conspecifics. The value of considering social context is highlighted by work indicating that a rodent's behavioral adaptations (or maladaptations) to chronic social stress may be more relevant to stress-related disorders in humans than the various types of physical stress used in most studies.

Several groups are applying naturalistic approaches to the analysis of mouse behavior. For example, a population of mice was derived from the offspring of several laboratory strains that had been released into a 400 m^2 outdoor pen in western Russia and subjected to natural selection. Experimental subjects were captured in the enclosure, implanted with microchips for tracking, and re-released into the pen, which contained automated feeding monitors. The system enabled observations relevant to foraging strategies, social organization, and diurnal patterns of activity. A smaller modified version of this approach was used to reveal differences in foraging strategies in trkB receptor mutant mice. (TrkB is a neurotrophin receptor implicated in cognitive function and mood.) Such approaches provide opportunities for insights into mouse behavior that are not available under standard laboratory conditions.

Another example is the Mouse Defense Test Battery, used for assessment of anxiety-related behavior. In this assay, mice are placed in a test alley and approached with an anesthetized rat, a natural predator of mice that triggers unconditioned aversive reactions. A series of species-typical behaviors is assessed, including flight,

freezing, risk assessment (e.g., repeated episodes of extending and withdrawing the body in the direction of the threat), and defensive threat and attack responses. Factor analysis of these behaviors showed four factors that were differentially modified by a series of anxiogenic and anxiolytic drugs. Such approaches that more fully account for the variety of behaviors seen in mice in threatening situations hold promise for modeling several aspects of anxiety responses. However, practical challenges to implementing such naturalistic approaches for high-throughput screening have limited their widespread use.

Mouse Models of Psychiatric and Neurologic Disorders

A major determinant of the quality of an animal model of a human disorder relates to the degree to which it reproduces the disorder's pathophysiology. Mouse models have been particularly useful for studies of disorders whose underlying genetic causes have been identified in humans (e.g., neurodegenerative diseases such as Alzheimer's, Parkinson's, and Huntington's diseases, amyotrophic lateral sclerosis, hereditary ataxias, frontotemporal dementias, multiple sclerosis, and prion diseases). Mice with disease-causing mutations show some, but not all, of the neuropathological and behavioral abnormalities seen in humans, and these abnormalities have served as critical endpoints for assessing the potential clinical utility of novel therapeutic agents. These mutant mice have also been useful for genetically dissecting individual steps in the biochemical pathways responsible for disease pathogenesis. Most behavioral analyses of these models have focused on assays of motor function and cognition. It is noteworthy that many of these disorders also involve abnormalities in affect, diurnal rhythms, and impulsivity that may precede cognitive and motor symptoms. Accordingly, a study of mice with scrapie and bovine spongiform encephalopathy revealed impairments in circadian activity and open-field exploration prior to the onset of motor abnormalities. It is therefore likely that the application of comprehensive behavioral analyses to these models will improve their sensitivity and promote insights into the pathophysiology and treatment of these neurodegenerative disorders.

In striking contrast, the prospects of developing *bona fide* animal models for most psychiatric diseases are complicated by at least four major considerations. First, in comparison to neurodegenerative diseases, there is a relative paucity of definitive information regarding causal genetic factors, making the generation of mouse models difficult. Second, the precise contribution of environmental factors to the pathophysiology of these disorders remains poorly understood, precluding the production of convincing models through environmental manipulation. A third consideration relates to clinical diagnostic limitations: a great deal of heterogeneity exists among the symptom clusters of individuals sharing a common diagnosis (e.g., depression or schizophrenia), and many are not readily classifiable using standard criteria. Uncertainty regarding the clinical features of the human syndromes to be simulated obviously complicates the validation of animal models. Finally, many of the hallmark features of the human disorders pose a daunting challenge for the mouse psychiatrist. Unanticipated breakthroughs would be required to convincingly model phenomena such as guilt, religiosity, grandiosity, envy, delusions, hallucinations, grief, body image distortion, and multiple personality in the mouse.

One partial exception to these limitations is drug addiction, in which at least one of the key etiological agents – the drug – is clearly identifiable and key features of the behavioral pathology (drug self-administration) are readily measured in rodents. Progress has also been made in modeling cardinal symptoms of human addiction, such as drug craving, in rodents, although assessing whether a mouse 'wants' a drug remains challenging.

Although most behavioral models of psychiatric disorders, other than addiction, are unconvincing as disease simulations, they may nevertheless prove useful for exploring the neural basis of particular manifestations of these disorders. Thus, it is possible to model specific symptoms of a psychiatric illness (loss of appetite and disrupted sleep) even if modeling the entire syndrome is not yet feasible. In addition, many available behavioral tests have predictive value for potential medications. This can be illustrated by the forced swim test, an assay used to screen antidepressant compounds. Animals are placed in a water-containing cylinder from which they cannot escape. Initially, mice make apparent attempts at escape, but over time these give way to periods of immobility when they float at the water surface. This immobile state is proposed to reflect 'behavioral despair' – the loss of hope of escaping. Compounds that reduce immobility times have an increased likelihood of antidepressant efficacy in humans. Nevertheless, the etiologic and pathophysiologic validity of this test – that is, the extent to which the mouse resembles a depressed human, or the centrality of swim stress to human depression – is unconvincing, particularly because immobility could be viewed as an adaptive strategy in this experimental context. Therefore, if a mutation is found to result in decreased immobility time, it is risky to conclude that the mouse is 'less depressed.' A more conservative interpretation would

be warranted – that the animal has a behavioral abnormality associated with responsiveness to swim stress and antidepressants. Despite these limitations, the forced swim test has clear predictive validity, indicating that an understanding of its mechanistic basis may shed light on the function of neural pathways pertinent to the treatment of depression.

Behavioral Screening: The Problem of Throughput

The need for high-throughput behavioral screening procedures is illustrated by two endeavors: (1) the application of large-scale mutagenesis approaches for detecting genes that regulate behavior and (2) the use of behavioral testing for CNS drug discovery. Enthusiasm for chemical mutagenesis approaches has led to the establishment of international mutagenesis centers that include behavioral screening within their panels of phenotypic assays. The thousands of mutants produced yearly by these programs could provide an invaluable resource for the exploration of genetic influences on behavior. Currently, limitations in the availability of comprehensive behavioral test strategies that can be administered in a high-throughput manner create a substantial bottleneck with regard to our ability to use the valuable animal resources provided by these centers.

CNS drug discovery differs most from other therapeutic areas by its unique reliance on *in vivo* tests for target identification and validation as well as for compound selection and refinement. As stated previously, a comprehensive behavioral assessment, which is needed for accurate interpretation of behavior, is a very slow, costly, and sometimes subjective process. Consequently, at a time when the number of potential targets and the number of compounds directed toward these targets are rapidly increasing, behavioral screening has become the main bottleneck in CNS drug discovery.

Fortunately, there have been several improvements in the throughput of behavioral screening (**Table 1**). Several companies have developed semiautomated systems for measuring multiple behaviors, such as locomotor activity, fear/anxiety responses, and startle. The currently available devices, however, still must be better integrated into databases designed to coordinate the analysis of multiple behavioral responses in order to best predict clinical efficacy. Some private companies are moving in that direction. For example, Hypnion (http://www.hypnion.com) is developing a high-throughput system that monitors electroencephalograms, sleep, wakefulness, and several physiological variables in rodents. Their SCORE-2004 system uses a large database to predict efficacy and side effects of existing and experimental hypnotic and wake-producing agents. SmartCube (PsychoGenics), one of the most comprehensive approaches to date, works by compressing numerous tests into one by gathering comprehensive, detailed, and fully automated data into a large database and by implementing sophisticated data-mining algorithms to detect psychotropic drug activity. Mice are given a drug and placed in SmartCube for a designated time, during which they are presented with a sequence of stimuli. Videotracking and other automated measures are then processed by SmartBase, an intelligent database that, through sophisticated statistical analysis and data mining, identifies 'drug signatures' – for example, antidepressant, antipsychotic, or anxiolytic effects. Further research is needed to determine whether it will be possible to generate such drug signatures by a high-throughput method.

High-throughput behavioral tests could markedly increase the speed of compound screening and reduce the subjectivity inherent in many current behavioral tests. However, a key drawback must be overcome. The automated tests will probably pick up novel features of an animal's behavior that do not have the extensive predictive validity of standard behavioral assays. The elevated plus-maze, Morris water maze, and forced swim test, for example, may be labor-intensive, but they are time proven. Extensive pharmacological validation will be needed to convince a pharmaceutical company to pursue such a novel line of drug discovery. Likewise, to understand how a drug or mutation affects behavior, any finding from a high-throughput assay would require much more investigation in established and labor-intensive tests to elucidate the underlying behavioral and neurobiological mechanisms involved.

Quantitative Home Cage Behavioral Pattern Analysis

In addition to the high-throughput approaches to drug discovery described in the preceding section, efforts are also underway to apply information technologies for a comprehensive evaluation of the impact of experimental manipulations (mutations, environmental manipulations, and drugs) on mouse behavior. At the University of California at San Francisco, an initiative is under way for the collection and mining of data reflecting the diverse and spontaneous behaviors of mice in their home cages (**Table 1**). Animals are fully acclimated to housing in cages containing nest enclosures, food-intake monitors, lickometers, and activity platforms that record animal position and movement velocity with high temporal and spatial resolution.

Temporal and spatial organization of multiple behaviors are simultaneously assessed using a variety of computational approaches, revealing substantial heterogeneity in the manner in which behavioral patterns are influenced by single-gene mutations, drugs, and environmental manipulations. In association with the Jackson Laboratory Mouse Phenome Project Initiative, we have also observed that behavioral patterns vary markedly among the 16 genetically diverse inbred strains that have been examined to date. Such information will be useful as a basis to search for genes conferring strain differences and for the selection of optimal strains for the study of particular behavioral processes.

Several advantages may be associated with the use of quantitative home cage behavioral analysis. First, it helps eliminate the confounding effects of environmental novelty, as well as the impact of handling on stress and experimental variability. It also enables investigators to monitor behavioral processes regulated over longer timescales, such as circadian rhythms, patterns of ingestion, estrus cycle influences, and the gradual onset of pharmacological and genetic effects. In addition, it provides an opportunity to monitor the impact of experimental manipulations in the context of the integrated expression of multiple behavioral domains. Furthermore, hypothesis-independent data mining may reveal novel insights into the regulation of behavioral patterns. Once these behavioral patterns are established, environmental modifications may be introduced to explore additional domains of behavior (e.g., introduction of novel objects, predator stimuli, and running wheels).

On the other hand, several challenges may complicate the use of such approaches. For example, the establishment of stable behavioral patterns and the collection of sufficient data for modeling may require several weeks of monitoring. Large studies requiring analysis of terabytes of generated data require a substantial investment in computational resources. Furthermore, accurate interpretation of the functional relevance of some behavioral effects may initially require validation using standard assays.

Conclusions

The examples provided in this article reflect a growing appreciation within academic and industrial settings of the intricacies and importance of behavioral analysis. This is driving a proliferation of data acquisition and data-mining innovations that will bring comprehensive mouse neurobehavioral assessment into the information age and markedly enhance the value of rodent models for the study of psychiatric and neurologic disorders. Future research is needed to increase the throughput and consistency of such comprehensive behavioral analyses, as well as to incorporate ethological perspectives for an increasingly accurate view of rodent behavior. Such advances will greatly enhance efforts to relate genes to complex behavior and to optimally mine the genomic revolution for improved treatments of brain diseases.

See also: Amyotrophic Lateral Sclerosis (ALS); Animal Models of Alzheimer's Disease; Animal Models of Huntington's Disease; Animal Models of Parkinson's Disease; Huntington's Disease; VelociGene and VelociMouse: High-Throughput Approaches for Generating Targeted Mutations in Mice on a Genome-Wide Scale.

Further Reading

Ahmad-Annuar A, Tabrizi SJ, and Fisher EM (2003) Mouse models as a tool for understanding neurodegenerative diseases. *Current Opinion in Neurology* 16: 451–458.

Berton O, McClung CA, DiLeone RJ, et al. (2006) Essential role of BDNF in the mesolimbic dopamine pathway in social defeat stress. *Science* 311: 864–868.

Berton O and Nestler EJ (2006) New approaches to antidepressant drug discovery: Beyond monoamines. *Nature Reviews Neuroscience* 7: 137–151.

Blanchard DC, Griebel G, and Blanchard RJ (2003) The Mouse Defense Test Battery: Pharmacological and behavioral assays for anxiety and panic. *European Journal of Pharmacology* 463: 97–116.

Blanchard RJ and Blanchard DC (2003) Bringing natural behaviors into the laboratory: A tribute to Paul MacLean. *Physiology & Behavior* 79: 515–524.

Brunner D, Nestler E, and Leahy E (2002) In need of high-throughput behavioral systems. *Drug Discovery Today* 7: S107–S112.

Cryan JF, Markou A, and Lucki I (2002) Assessing antidepressant activity in rodents: Recent developments and future needs. *Trends in Pharmacological Sciences* 23: 238–245.

Dell'Omo G, Ricceri L, Wolfer DP, Poletaeva II, and Lipp H (2000) Temporal and spatial adaptation to food restriction in mice under naturalistic conditions. *Behavioural Brain Research* 115: 1–8.

Dell'Omo G, Vannoni E, Vyssotski AL, et al. (2002) Early behavioural changes in mice infected with BSE and scrapie: Automated home cage monitoring reveals prion strain differences. *European Journal of Neuroscience* 16: 735–742.

Ferkin MH and Li HZ (2005) A battery of olfactory-based screens for phenotyping the social and sexual behaviors of mice. *Physiology & Behavior* 85: 489–499.

Gerlai R (2002) Phenomics: Fiction or the future? *Trends in Neurosciences* 25: 506–509.

Gerlai R and Clayton NS (1999) Analysing hippocampal function in transgenic mice: An ethological perspective. *Trends in Neurosciences* 22: 47–51.

Goulding EH, Juneja P, Wade J, and Tecott LH (2002) Quantitative analysis of ingestive behavior pattern in 5-HT2c receptor, obese, and agouti mutant mice. *Appetite* 39: 78.

Green AR and Heal DJ (1985) The effects of drugs on serotonin-mediated behavioural models. In: Green AR (ed.) *Neuropharmacology of Serotonin*, pp. 326–365. Oxford: Oxford University Press.

Hrabe de Angelis MH, Flaswinkel H, Fuchs H, et al. (2000) Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nature Genetics* 25: 444–447.

Hunter AJ, Nolan PM, and Brown SD (2000) Towards new models of disease and physiology in the neurosciences: The role of induced and naturally occurring mutations. *Human Molecular Genetics* 9: 893–900.

Insel TR and Fernald RD (2004) How the brain processes social information: Searching for the social brain. *Annual Review of Neuroscience* 27: 697–722.

Kaakkola S and Teravainen H (1990) Animal models of parkinsonism. *Pharmacology & Toxicology* 67: 95–100.

Koob GF, Sanna PP, and Bloom FE (1998) Neuroscience of addiction. *Neuron* 21: 467–476.

Lister RG (1990) Ethologically-based animal models of anxiety disorders. *Pharmacology and Therapeutics* 46: 321–340.

Masuya H, Inoue M, Wada Y, et al. (2005) Implementation of the modified-SHIRPA protocol for screening of dominant phenotypes in a large-scale ENU mutagenesis program. *Mammalian Genome* 16: 829–837.

McIlwain KL, Merriweather MY, Yuva-Paylor LA, and Paylor R (2001) The use of behavioral test batteries: Effects of training history. *Physiology & Behavior* 73: 705–717.

Morris RGM, Garrud P, Rawlins JNP, and O'Keefe J (1982) Place navigation impaired in rats with hippocampal lesions. *Nature* 297: 681–683.

Nestler EJ (2001) Molecular basis of long-term plasticity underlying addiction. *Nature Reviews Neuroscience* 2: 119–128.

Owens T, Wekerle H, and Antel J (2001) Genetic models for CNS inflammation. *Nature Medicine* 7: 161–166.

Paigen K and Eppig JT (2000) A mouse phenome project. *Mammalian Genome* 11: 715–717.

Rogers DC, Fisher EM, Brown SD, Peters J, Hunter AJ, and Martin JE (1997) Behavioral and functional analysis of mouse phenotype: SHIRPA, a proposed protocol for comprehensive phenotype assessment. *Mammalian Genome* 8: 711–713.

Scott MR, Supattapone S, Nguyen HO, DeArmond SJ, and Prusiner SB (2000) Transgenic models of prion disease. *Archives of Virology* 16(supplement): 113–124.

Shaham Y, Erb S, and Stewart J (2000) Stress-induced relapse to heroin and cocaine seeking in rats: A review. *Brain Research: Brain Research Reviews* 33: 13–33.

Storm EE and Tecott LH (2005) Social circuits: Peptidergic regulation of mammalian social behavior. *Neuron* 47: 483–486.

Sutton MA, Schmidt EF, Choi K-H, et al. (2003) Extinction-induced upregulation in AMPA receptors reduces cocaine-seeking behaviour. *Nature* 421: 70–75.

Tecott LH (2003) The genes and brains of mice and men. *American Journal of Psychiatry* 160: 646–656.

van der Staay FJ and Steckler T (2002) The fallacy of behavioral phenotyping without standardisation. *Genes, Brain and Behavior* 1: 9–13.

Voikar V, Vasar E, and Rauvala H (2004) Behavioral alterations induced by repeated testing in C57BL/6J and 129S2/Sv mice: Implications for phenotyping screens. *Genes, Brain and Behavior* 3: 27–38.

Vyssotski AL, Dell'Omo G, Poletaeva II, et al. (2002) Long-term monitoring of hippocampus-dependent behavior in naturalistic settings: Mutant mice lacking neurotrophin receptor TrkB in the forebrain show spatial learning but impaired behavioral flexibility. *Hippocampus* 12: 27–38.

Wahlsten D, Rustay NR, Metten P, and Crabbe JC (2003) In search of a better mouse test. *Trends in Neurosciences* 26: 132–136.

Willner P (1991) Behavioral models in psychopharmacology. In: Willner P (ed.) *Behavioral Models in Psychopharmacology: Theoretical, Industrial and Clinical Perspectives*, pp. 3–18. Cambridge, UK: Cambridge University Press.

Wise RA (1998) Drug-activation of brain reward pathways. *Drug and Alcohol Dependence* 51: 13–22.

Relevant Websites

http://www.hypnion.com – Hypnion.
http://www.psychogenics.com – PsychoGenics.

Role of NO in Neurodegeneration

Y-I Lee, T M Dawson, and V L Dawson, The Johns Hopkins University School of Medicine, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Nitric oxide (NO) was first identified as endothelium-derived relaxing factor. NO is produced by many different cells in multicellular organisms, where it acts as an unprecedented biological messenger molecule. NO is synthesized by the enzymes NO synthases (NOSs), which catalyze the essential amino acid L-arginine into NO and L-citrulline. In the central nervous system (CNS), NO is an unusual neuronal messenger molecule in that conventional neurotransmitters are released by exocytosis from the nerve terminal, but because of its small size and ability to move through lipids, NO reaches its targets through diffusion. NO forms covalent and noncovalent linkages with protein and nonprotein targets. Conventional neurotransmitters undergo reversible interactions with cell surface receptors. NO is inactivated by diffusion away from its targets, by forming covalent linkages to the superoxide anion, or by scavenger proteins. In contrast, conventional neurotransmitters are terminated by presynaptic reuptake or enzymatic degradation. Because NO does not use conventional means for control of its biological actions, NO depends upon its small size, reactivity, and diffusibility more than any other biological molecule to exert its biological effects. NO is involved in a wide range of physiological processes, and under conditions of excess formation in the presence of superoxide anion it can mediate a series of pathophysiological actions. The multifaceted actions of NO in brain physiology include mechanisms related to development, synaptic function and neural plasticity, and neuronal survival and neuronal cell death.

NO in the Nervous System

NO is produced by a group of enzymes designated NOS. There are three members of the NOS family: neuronal NOS (nNOS), endothelial NOS (eNOS), and inducible NOS (iNOS). A fourth family member was proposed, mitochondrial NOS (mtNOS), but studies indicate that mtNOS is nNOSα – it is coded by the nNOS gene, and nNOS knockout mice do not have mtNOS. All the NOSs share between 50% and 60% sequence homology. In the nervous system, nNOS is expressed in neurons, eNOS is predominantly expressed in the endothelium of blood vessels, and iNOS is expressed in glial cells. nNOS and eNOS are Ca^{2+}-calmodulin-dependent enzymes that are constitutively expressed, although both isoforms can be induced following a variety of physiological and pathological stimuli. In contrast, iNOS does not require Ca^{2+}-calmodulin for activation and its regulation depends on *de novo* synthesis. Despite its highly diffusible nature, NO can exert very specific effects within the CNS and peripheral nervous system (PNS). Under physiological conditions, NO facilitates neurotransmitter release and uptake via neuron-glial communication, modulating the release of neurotransmitters, including glutamate, γ-aminobutyric acid (GABA), substance P, and acetylcholine. nNOS and eNOS are regulated via posttranslational modification, including phosphorylation, and nNOS levels are dynamically regulated through transcription. Cerebral blood flow is regulated through NO released from endothelial cells as well as autonomic nerves within the adventitia. NO may also regulate cerebral blood flow through activity-dependent activation of NOS in neurons that influence small cerebral arterioles. In the brain, nNOS is linked to the postsynaptic membrane near *N*-methyl-D-aspartate (NMDA) receptors via the postsynaptic density protein-95, and nNOS function is regulated through its interactions with the protein inhibitor of NOS (PIN) and carboxyl-terminal PDZ ligand of NOS (CAPON) (Figure 1). In the gut, NO functions as the nonadrenergic, noncholinergic (NANC) neurotransmitter where it mediates relaxation of smooth muscles associated with peristalsis. Penile erections are mediated, in part, through the neurotransmitter action of NO.

The diffusion properties and the short half-life of NO make it an ideal candidate for playing a role in nervous system morphogenesis and synaptic plasticity. In the nervous system, NO was first recognized as a messenger molecule that mediates increases in cyclic GMP (cGMP) levels that occur after activation of glutamate receptors, particularly those of the NMDA subtype. In this capacity, NO modulates long-term potentiation (LTP), which is essential for learning and memory. LTP is inhibited by administration of NOS blockers or genetic deletion of NOS isoforms, confirming the role of NO in synaptic plasticity. Nerve growth factor (NGF)-induced differentiation of pheochromacytoma cells (PC12) has been shown to be NO dependent with NO acting as a cytostasis factor. NO may play a role in the developmental processes of the nervous system, in that it is transiently expressed in the cerebral cortical plate during critical periods of neuronal development. NO is also critical for the neurite outgrowth and differentiation in murine cortical

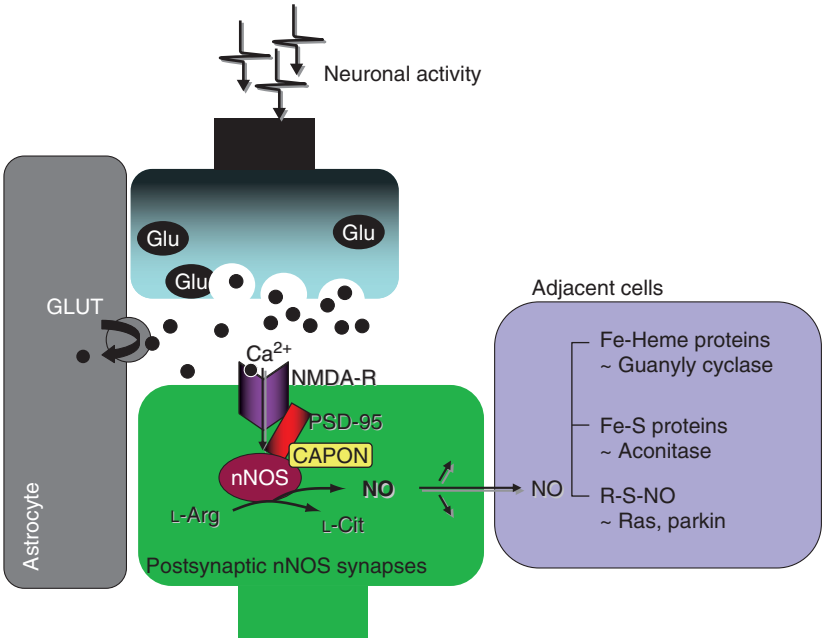

Figure 1 NO signaling in the CNS. Glutamate activation of NMDA receptor results in stimulation of neuronal NO synthase (nNOS) which is functionally coupled to the NMDA receptor by PSD-95, a scaffolding protein. The tethering of nNOS near the NMDA receptor facilitates its activation by Ca^{2+} resulting in the generation of NO. CAPON binds nNOS and regulates its interaction with PSD-95 and thus regulates the generation of NO. NO diffuses between cells and can activate a number of cell signaling pathways in adjacent cells. NO can react with the heme center of soluble guanylyl cyclase (GC) activating GC to produce cyclic GMP (cGMP). NO can react with iron–sulfur (Fe–S)-containing proteins such as aconitase to regulate their activity. NO may S-nitrosylate (R-S-NO) proteins such as Ras or parkin, thus regulating complex cell signal cascades.

neurons. NO may play a role in the molecular maturation of adult spinal cord motor neurons and also in the initial pattern of connections in ocular dominance columns in the developing visual system. It has also been suggested to play a prominent role in the activity-dependent establishment of connections in both the developing and regenerating olfactory neuronal epithelium. Thus, NO may play an important role in synaptic refinement in the developing nervous system.

In the immune system, NO generated by iNOS is released in bursts of high concentrations and is a key cytotoxic weapon for the destruction of bacteria. The CNS immune cell counterparts, microglia and astrocytes, also have iNOS, which generates a burst of NO in response to injury and can contribute to neural injury. The expression of iNOS occurs following trauma, following stroke, or during neurodegenerative diseases and is thought to contribute to neural injury.

Although NOS is the main NO source, in some special situations this molecule can be synthesized by other mechanisms. NO can be produced by the xanthine oxidase pathway or by H_2O_2 and L-arginine in a nonenzymatic reaction or by the reduction of nitrites in acid and reducing conditions, such as occurs in ischemic tissue.

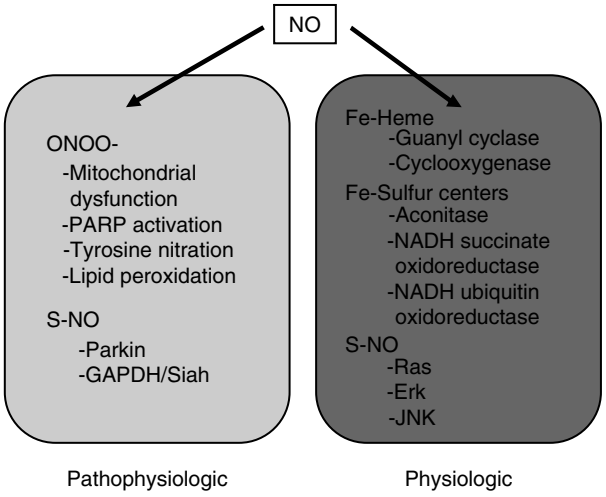

Figure 2 Physiological and pathophysiological NO signaling events.

NO Molecular Targets

There are several molecular mechanisms by which NO elicits cellular signaling (Figure 2). The unique chemical properties of NO facilitate its role as an important signaling molecule in the biological system by reacting directly with biological molecules or indirectly

following reaction with other reactive oxygen species. NO reacts with the ferrous ion in the heme complex in enzymes such as guanylate cyclase (GC), P450, and hemoglobin to form an iron nitrosyl complex. The binding of NO induces a conformational change in the protein structure activating these proteins. The prototype reaction is that of NO on GC leading to the production of cyclic guanosine-3′,5′-monophosphate (cGMP). cGMP is a second messenger that activates protein kinases. In the nervous system, NO was first recognized as a messenger molecule that regulates cGMP accumulation through activation of glutamate receptors, particularly those of the NMDA subtype. cGMP, in turn, binds to and activates a cGMP-dependent protein kinase (G kinase), which phosphorylates specific proteins on serine or threonine residues and also activates phosphodiesterase and ion channels. NO increases prostaglandin production by activation of cyclo-oxygenase, another heme-containing enzyme. Numerous enzymes (including *cis*-aconitase, NADH succinate oxidoreductase, and NADH ubiquinone oxidoreductase) that contain nonheme iron in iron sulfur clusters are also regulated by NO. NO also stimulates the RNA-binding function of the iron-responsive element binding protein in brain slices while diminishing its aconitase activity.

In recent years, the importance of protein S-nitrosylation of cysteine thiols by NO in physiological and pathophysiological actions has been appreciated. S-nitrosylation is a posttranslational protein modification that increases or decreases the activity of the target protein. Some known biological targets for S-nitrosylation include glyceraldehyde-3-phosphate dehydrogenase (GAPDH), caspases, transglutaminase, aromatases, mitochondrial aldehyde dehydrogenase, cyclooxygenase-2, G-proteins, p21-Ras, RAC1, cdc42, Dexras1, ERK, JNK, p38, and PKBa/AKT1. S-nitrosylation of the monomeric GTPase Ras activates this enzyme, triggering the MAPK and PI3K pathways in an NO-dependent and cGMP-independent manner. A recent study reports that NO donors promote HIF-1α accumulation via an increase in transcription and translation. It is not yet known how this occurs, but it is possible that these effects are due to activation of Ras and PI3K through NO-mediated nitrosylation of Ras. NO has inhibitory effects on cell cycle progression and promotes cell differentiation, in part due to activation of the retinoblastoma gene product (pRb) by S-nitrosylation. Nuclear factor kappa beta (NF-$\kappa\beta$), a transcription factor that induces iNOS expression, is inhibited by S-nitrosylation. Thus, NO inhibition of NF-kβ could be an autoregulatory feedback mechanism to regulate NO production. Another autoregulatory mechanism is the prevention of eNOS dimerization by S-nitrosylation, resulting in decreased eNOS activity. S-nitrosylation is also involved in the pathological activation of matrix metalloproteases in stroke and neurodegenerative diseases. NO regulates NMDA receptor activity through nitrosylation, thus reducing calcium permeability. One of the important features of S-nitrosylation is that the modification is reversible, and a specific protein sequence of acidic–basic motifs seems to be necessary. As usual, the modification is also selective, since only a specific cysteine is generally nitrosylated in a protein.

Oxygen and other free radicals can also react with NO to form reactive nitrogen oxide species and exertits effects on different biological molecules. Superoxide anion can react with NO to form peroxynitrite, a potent oxidant with many cellular targets. Peroxynitrite damages proteins by reacting with the tyrosine residues to form 3-nitrotyrosine. It can also induce lipid peroxidation and cause DNA damage. It is thought that many of the acute toxic effects of NO are mediated by peroxynitrite.

Neurotoxic Effects of NO

In the CNS, glutamate excitotoxicity due to activation of the NMDA receptor is mediated in large part by NO. Under physiological conditions, the binding of glutamate activates NMDA receptors in the postsynaptic neuron that result in Ca^{2+} influx, which in turn activates nNOS. Then NO either via cGMP or by S-nitrosylation suppresses NMDA receptor activity, thereby limiting further Ca^{2+} influx and thus protecting neurons. However, under pathophysiological conditions such as occur during stroke, trauma, or neurodegenerative stress, NO participates in neuronal damage through disruption of this system and generation of peroxynitrite that activates subsequent cell death cascades. NO regulation of the release of glutamate contributes to the initial toxic events. When NO concentrations are low, there is a decrease in glutamate release despite elevated cGMP levels. However, under increased NO production and excessive increases in cGMP levels, the inhibitory effect on glutamate release is reversed, and glutamate release is increased. Under the pathological conditions of excess glutamate release and NMDA receptor activation, activation of NOS leads to cell death. Excitotoxicity can be blocked by several classes of NOS inhibitors or by nNOS gene deletion. nNOS knockout mice are resistant to focal and transient global ischemia, suggesting a role for nNOS in CNS injury. The neuronal toxicity of NO is mediated by peroxynitrite, which inhibits mitochondrial respiration and manganese superoxide dismutage (MnSOD) activity,

thus establishing a feed-forward loop for peroxynitrite generation. Peroxynitrite damages DNA, which activates the nuclear enzyme poly(ADP-ribose) polymerase (PARP) that leads to the release of the cell death effector apoptosis-inducing factor (AIF) from mitochondria. AIF can induce phosphatidyl serine exposure on the plasma membrane, and in the nucleus it facilitates DNA cleavage and nuclear condensation (Figure 3). In addition, glutamate receptor-mediated excitotoxicity activates parallel cellular responses, resulting in the activation of proteolytic enzymes, lipid peroxidation, ROS and RNS formation, and activation of programmed cell death cascades. Under some conditions, NO can activate the ceramide–Jun kinase pathway and the NF-κβ–Jun kinase pathway, which regulate the expression of a large number of genes involved in both neuroprotection and neurotoxicity. The biological significance of NO stimulation of these pathways depends on the cell type and the environment. NO can also play a protective role via the blocking of caspases (cysteine–aspartate proteases), which form a complex family of proteolytic enzymes. It is not yet clear whether NO acts as a master regulator, determining which cell death pathway will be activated in neuronal injury, but these data are suggestive that NO production could direct neurons away from a classic apoptotic type of cell death and toward an excitotoxic form of cell death acutely and perhaps activate other apoptotic cascades in the late remodeling phase of neuronal injury and recovery.

Oxidative and nitrosative stress plays a pivotal role in ischemic-reperfusion injury during stroke. The increase of NO production in the presence of increased superoxide anion leads to brain damage. Brain ischemia triggers a cascade of events, including membrane depolarization, increase in intracellular Ca^{2+} concentrations via increase of extracellular glutamate concentration, and overstimulation of NMDA receptors, leading to activation of calcium-dependent nNOS. Released NO rapidly reacts with superoxide produced in excess during reperfusion resulting in peroxynitrite formation, and nitrosylation or nitration of proteins. Furthermore, ischemia or reperfusion eventually induces the expression of iNOS. This isoform is not normally present in the CNS, but it can be detected after inflammatory, infectious, or ischemic damage, as well as in the normal aging brain. Activation of nNOS or induction of iNOS mediates ischemic brain damage through mechanisms previously discussed. On the other hand, eNOS is thought to act as a neuroprotective agent by enhancing blood supply to the injured tissue. Therefore, NO generated by nNOS or iNOS contributes to injury following stroke, whereas NO generated by eNOS is protective by maintaining blood flow (Figure 4).

NO and Neurodegenerative Diseases

Abnormal activation of glutamate neurotransmission may contribute to neurodegenerative processes and

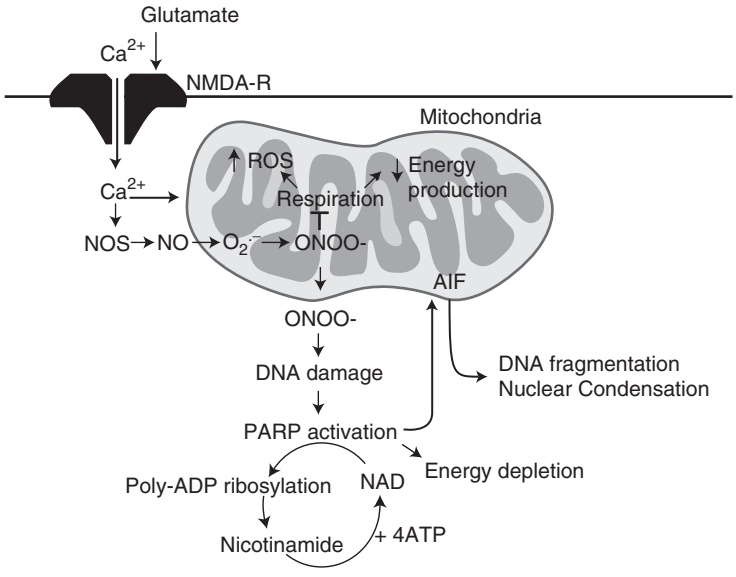

Figure 3 Mechanism of NO-mediated neurotoxicity. NMDA receptor (NMDA-R) activation causes an increase in intracellular calcium levels, which activates NOS, as well as other potential free radical-generating enzymes. These free radicals then damage DNA and also inhibit mitochondrial function, leading to more free radical formation. The damaged DNA activates PARP-1, which transfers ADP-ribose groups to nuclear proteins consuming NAD. NAD is resynthesized from nicotinamide (NAm), a reaction that consumes four high-energy equivalents of ATP. PARP activation also elicits the release and translocation of AIF from the mitochondria into the nucleus, initiating a caspase-independent cascade of cell death.

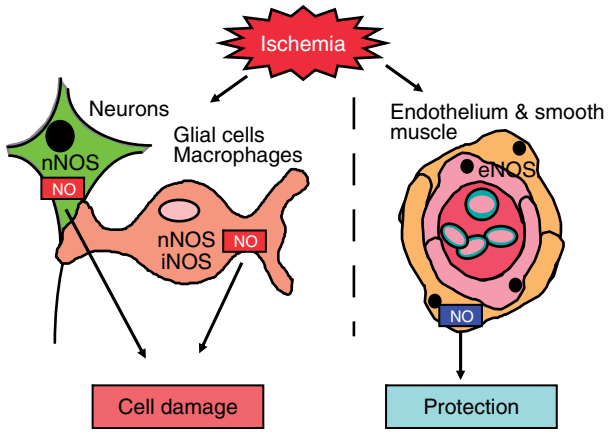

Figure 4 Dual roles for NO in brain ischemia. NO is involved in the mechanisms of neurotoxicity after cerebral ischemia. Following a stroke, the generation of NO can be either beneficial or detrimental to the brain depending on the which NOS isoform is activated and in which cellular compartment. NO generated from eNOS in endothelial cells provides protection by inducing vasodilation. Inhibition of eNOS can drastically increase stroke injury. NO produced by nNOS in neurons or iNOS in glia results in the formation of toxic peroxynitrite and cellular damage. Inhibition of nNOS or iNOS greatly reduces brain injury following stroke.

diseases such as Parkinson's disease (PD), Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS), and multiple sclerosis (MS). Neurodegenerative disorders are usually marked by a progressive loss of a selective group of neurons and are associated with aging. One of the common features of neurodegeneration is the presence of protein aggregates and increased indices of oxidative stress.

PD is a common neurodegenerative disorder characterized by impairment in motor function. It is another common age-related neurodegenerative condition associated with selective loss of dopaminergic neurons in the substantia nigra of the midbrain region. In PD, postmortem pathological studies have revealed mitochondria complex I dysfunction, reduced glutathione and ferritin levels, increased lipid peroxidation, nitrotyrosine immunoreactivity, and increased levels of iron. Genetically engineered animals reveal that NO generated by both iNOS and nNOS contributes to pathology. In the experimental 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) intoxication model of PD, NOS inhibition in wild-type mice or gene deletion of the iNOS or nNOS gene protects dopaminergic neurons. In PD patients, polymorphonuclear cells have increased expression of nNOS and increased NO production and accumulation of nitrotyrosine-containing proteins, reflecting an increase in nitrosative stress in these patients. Glial-derived neuronal factor (GDNF) is protective in animal models of PD, although its effect on patients remains controversial. GDNF action leads to the inhibition of nNOS activity and neurotoxicity. *R*-(–)-deprenyl (deprenyl) slows the progression of disability in early PD. Deprenyl prevents S-nitrosylation of GAPDH, which blocks the binding of GAPDH to Siah, preventing the GADH/Siah death cascade. In MPTP-treated mice, low doses of deprenyl prevent binding of GAPDH and Siah1 in the dopamine-enriched corpus striatum and it is neuroprotective.

PD is mostly sporadic, but rare familial cases are also found. Mutations in α-synuclein, parkin, DJ-1, PINK1, and LRRK2 have been linked to rare familial forms of PD. Point mutations or replication of α-synuclein results in autosomal dominant PD, which prompted intense investigation into the actions and pathobiology of this protein. The hallmark lesions of PD, the Lewy body, contain α-synuclein that is modified by nitration of tyrosine residues. It is possible that dityrosine cross-linking can generate stable oligomers that lead to aggregates. Both wild-type and disease-causing α-synuclein mutants can be nitrated, and the nitration induces the formation of α-synuclein inclusions with similar biochemical characteristics to protein extracted from human PD tissue. In addition to nitration, nitrosylation may also play a key role in both sporadic and familial PD, as there is a greater than twofold increase in S-nitrosylated proteins in PD. Parkin is an E3 ligase that adds ubiquitin on specific protein substrates by both K48 and K63 linkages. Disease-causing mutations and loss-of-function mutants lead to the accumulation of toxic parkin substrates in both mouse and humans. Parkin can be S-nitrosylated. This modification results in loss of function as an E3 ligase and inhibition of the protective function of parkin. In sporadic cases of PD, parkin is S-nitrosylated, suggesting that in the more common form of PD, parkin function may be compromised. While parkin and DJ-1 do not interact physiologically, under conditions of oxidative and nitrosative stress there is a gain of function and these two proteins interact. It is possible that this interaction results in the sequestering of DJ-1 into insoluble aggregates terminating its normal biological actions. The pathophysiological significance of this interaction between DJ-1 and parkin is not yet known but is an area of active scientific investigation. Actions of oxidative and nitrosative stress on PINK1 or LRRK2 await investigation.

AD is the most common chronic progressive neurodegenerative condition; it was originally described by Alzheimer about a century ago. Brains from AD patients are characterized by extracellular aggregates of amyloid β-peptide (Aβ) that form the neuritic plaques and intracellular neurofibrillary tangles due to the hyperphosphorylation of tau protein. NO can contribute to AD neurodegeneration in several

ways. Aβ fibrils are toxic, in part, by inducing ROS formation, which can produce peroxynitrite by reacting with NO. Increased expression of iNOS and eNOS in astrocytes has been associated with the presence of neuritic plaques indicating increased nitrosative stress in AD. Additional studies indicate that Aβ is critically involved in the induction of an inflammatory response through activation of astrocytes and microglia, eliciting subsequent production of inflammatory cytokines and neurotoxic factors, including NO, superoxide, and prostaglandins (via COX-2). Aβ deposition in brain vessels correlates with a decrease in the expression of eNOS in endothelial cells that can lead to decreased regulation of cerebral blood flow. In cell culture and experimental animal models, Aβ elicits NO-mediated excitotoxicity, neuroinflammation, and oxidative stress. Similar stress events are observed in patient tissue, implicating a role for NO in the progression of AD.

ALS is a neurodegenerative disease characterized by selective death of motor neurons in the cerebral cortex, brain stem, and spinal cord. The role of NO in ALS is controversial. There is evidence that nitrotyrosination induces motor neuron death and that there is irreversible inhibition of the mitochondrial respiratory chain in these cells, suggesting a role for NO. However, mice transgenic for the human disease-causing mutation in SOD-1 when crossed with the nNOS knockout mice did not derive any benefit and progressed to disease and death. On the other hand, treatment with a nonselective NOS inhibitor slightly decreased motor neuron degeneration in an ALS mouse model. It is possible that NO from different isoforms could contribute to progressive neurodegeneration in ALS and exacerbate disease progression. In postmortem tissue, there are reactive astrocytes that contain protein inclusions, express iNOS and COX-2, display nitrotyrosine immunoreactivity, and have a loss of the glutamate transporter, EAAT2. The combination of excitotoxicity and nitrosative stress could contribute to the development and acceleration of the disease.

Conclusion

NO is a multifunctional molecule that is needed for physiological functions, especially in the brain. NO has clearly revolutionized our thought about aspects of neuronal transmission. It is capable of regulating various body functions directly, via a cGMP-mediated mechanism or a cGMP-independent mechanism. NO is an important mediator of a variety of physiological effects in the nervous system; however, excessive NO production under pathological conditions could be destructive to the brain. Understanding the role of NO in these processes as well as the disposition and activation of NO will lead to a better understanding of the basic processes underlying normal and pathological neuronal functions, and hopefully to the development of selective protective therapeutic agents against neurodegenerative diseases.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Excitotoxicity in Neurodegenerative Disease; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Neurotoxins and their Neurotoxicology; Nitric Oxide; NMDA Receptors and Development; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Ahern GP, Klyachko VA, and Jackson MB (2002) cGMP and S-nitrosylation: Two routes for modulation of neuronal excitability by NO. *Trends Neuroscience* 25: 510–517.

Chung KK, Dawson TM, and Dawson VL (2005) Nitric oxide, S-nitrosylation and neurodegeneration. *Cellular and Molecular Biology* 51: 247–254.

Contestabile A and Ciani E (2004) Role of nitric oxide in the regulation of neuronal proliferation, survival and differentiation. *Neurochemistry International* 45: 903–914.

Dawson VL and Dawson TM (2004) Deadly conversations: Nuclear-mitochondrial cross-talk. *Journal of Bioenergetics and Biomembranes* 36: 287–294.

Emerit J, Edeas M, and Bricaire F (2004) Neurodegenerative diseases and oxidative stress. *Biomedicine and Pharmacotherapy* 58: 39–46.

Esplugues JV (2002) NO as a signaling molecule in the nervous system. *British Journal of Pharmacology* 135: 1079–1095.

Guix FX, Uribesalgo I, Coma M, and Munoz FJ (2005) The physiology and pathophysiology of nitric oxide in the brain. *Progress in Neurobiology* 76: 126–152.

Moro MA, Cardenas A, Hurtado O, Leza JC, and Lizasoain I (2004) Role of nitric oxide after brain ischemia. *Cell Calcium* 36: 265–275.

Thippeswamy T, Mckay JS, Quinn JP, and Morris R (2006) Nitric oxide, biological double-faced janus – Is this good or bad? *Histology and Histopathology* 21: 445–458.

Saccade–Pursuit Interactions

J-J Orban de Xivry and P Lefèvre, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Saccades are typically evoked by target steps, whereas pursuit eye movements are initiated by a moving target of interest. These two types of eye movements have been studied separately for years because it was thought that pursuit and saccadic systems were independent and were controlled by distinct cortical and subcortical networks (see Figure 1). Initially, scientists developed paradigms to elicit one of the two types of eye movements without any disruption by the other; that is, saccades were directed to stationary targets and the Rashbass paradigm was used to elicit pure smooth eye movements. Indeed, if a visual target first steps in one direction before moving in the opposite direction (this typical motion of the target is called the Rashbass paradigm), the probability of occurrence of saccades is strongly reduced during pursuit initiation. In contrast, if a target suddenly moves from the center in one direction, a catch-up saccade (saccade made during pursuit of a target to reduce the position error between the eye and the moving target) is necessary to foveate the target.

The Rashbass paradigm is essentially used in species, such as primates, that have a pursuit gain close to unity (the ratio between eye and target velocity). This high gain allows primates to pursue a moving target smoothly. In contrast, the smooth eye velocity of cats saturates at lower velocities (around $25°s^{-1}$), and their pursuit gain is much more variable. As a consequence, cats usually rely on a combination of smooth and saccadic eye movements to pursue a moving target. In addition, cats have difficulties to make large saccades but, instead, usually perform series of small saccades to achieve large gaze shifts. During the intersaccadic intervals, postsaccadic slow eye movements directed toward the stationary target help reduce the residual error at the end of saccades. Thus, cats need a combination of both saccadic and smooth eye movements either to shift their gaze or to pursue a moving target.

After some preliminary observations of saccade–pursuit interaction in primates, major progress was achieved with cat studies. Indeed, the low gain of eye movements in this species forces the two subsystems to cooperate in order to foveate the target. Thereafter, scientists began to design new paradigms to test the hypothesis of cooperation between the two oculomotor subsystems in primates. We now focus on these paradigms and the evidence of interactions they provided. We first describe evidence for shared velocity and position inputs to the saccadic and pursuit systems. We then examine the different mechanisms that are common to both oculomotor subsystems. Finally, we review the brain areas that were proven to contain signals related to the planning and execution of both pursuit and saccades.

Velocity Input to the Saccadic System

In contrast with the classical view of independent saccadic and pursuit systems (see Figure 1), lesions and behavioral studies demonstrated that there is a velocity input to the saccadic system. Lesions in brain areas (the middle temporal area (MT) or dorsolateral pontine nuclei (DLPN)) that are known to be prominent for motion processing disrupt the accuracy of saccades to moving targets without impairing saccades to stationary targets. Furthermore, the accuracy of catch-up saccades during pursuit initiation and maintenance implies that target motion is taken into account in their programming. Finally, the Rashbass paradigm itself shows that a prediction about future target motion (based on target velocity) is taken into account by the saccadic system. Indeed, after an initial step, the target moves toward the center of the visual field and the oculomotor system predicts that no saccade is necessary to catch the target but, instead, that a purely smooth movement is sufficient. Indeed, if the saccadic system did not have access to motion signals, it would trigger a saccade to the target in the direction opposite to target motion. All these experiments contributed to the accumulation of evidence showing that motion signals play a role in saccade programming and trigger but did not discriminate which parameter related to target motion (target velocity or velocity error) is taken into account. Indeed, these two parameters are highly correlated during pursuit initiation because eye velocity is very small in comparison with target velocity.

The first evidence for a prominent role of the retinal slip (RS, velocity error) came from cat studies and was confirmed later in humans. Cat experiments took advantage of the large variability of pursuit gain in this species, which yields a large range of RS for the same target velocity. Because the pursuit gain of primates is much larger and less variable, an original paradigm involving both position and velocity steps of the moving target was used to elicit catch-up

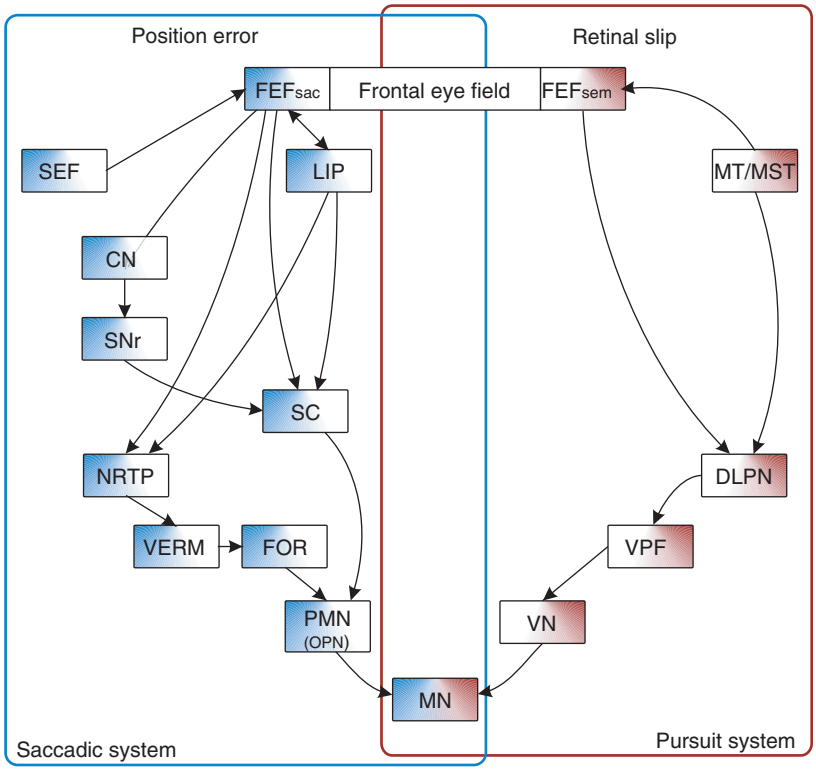

Figure 1 Diagram of the traditional view of saccadic and pursuit systems. Blue boxes are part of the saccadic system. Red boxes belong to the pursuit system. Two-color boxes are part of both systems. CN, caudate nucleus; DLPN, dorsolateral pontine nuclei, FEFsac, saccadic region of the frontal eye field; FEFsem, pursuit region of the frontal eye field; FOR, fastigial oculomotor region; LIP, lateral intraparietal area; MN, motor neurons; MST, medial superior temporal area; MT, middle temporal area; NRTP, nucleus reticularis tegmenti pontis; OPN, omnipause neurons; PMN, brain stem premotor nuclei; SC, superior colliculus; SEF, supplementary eye field; SNr, substantia nigra pars reticulata; VERM, vermis; VN, vestibular nuclei; VPF, ventral paraflocculus.

saccades during pursuit of a moving target. This paradigm (ramp–step–ramp paradigm) consisted in a first ramp that was followed by a step of the target and a second ramp. During the step, both target position and velocity varied instantaneously and randomly before the second target ramp. The position and velocity steps created both position error (PE) and RS. To compensate for these errors, the oculomotor system combined catch-up saccades and pursuit responses.

In cats and primates, the amplitude of catch-up saccades results from the addition of two components: one related to PE that represents the proportion of PE that is taken into account by the catch-up saccade (on average, 90%) and another proportional to RS that compensates for target motion during the saccade. Some electrophysiological recordings suggest that the contribution of PE might be conveyed by the superior colliculus (SC), whereas the contribution of RS originates from the motion-processing pathway (through the MT).

The processing of the two components by different networks was confirmed by the asynchrony of signals related to PE and RS. This asynchrony has been demonstrated separately with the ramp–step–ramp paradigm in two dimensions, on one hand, and by the shape of catch-up saccade velocity profiles, on the other hand. The two-dimensional ramp–step–ramp paradigm yielded curved catch-up saccades (large differences between the initial and final orientation of saccades). The asynchrony between PE and RS could be inferred from the difference between initial and final orientations of saccades that were aligned with the PE and RS vectors, respectively (see Figure 2). Similarly, the asymmetry of the shape of the velocity profile of catch-up saccades highlighted the asynchrony between PE and RS. The acceleration phase was related to PE and the deceleration phase to RS. These two results show that PE is available to the saccadic system earlier than RS because in both cases saccades take first PE into account and subsequently RS.

In addition, it has been demonstrated that both PE and RS influence the mechanism leading to the initiation of a catch-up saccade (trigger system) during ongoing smooth pursuit (Figure 3). The combination of PE and RS determines the time-to-contact between eye and target (eye-crossing time: $T_{XE} = -\mathrm{PE}/\mathrm{RS}$),

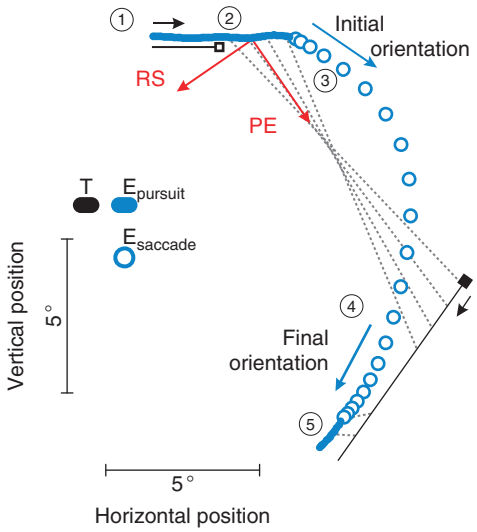

Figure 2 Velocity input to the saccadic system. A curved saccade (blue circles) occurring after both position and velocity steps of the target (black solid lines) is presented in two dimensions. The target position just before and after the target step is represented by the open and closed squares. Before and after the saccade, smooth pursuit eye movements (solid blue lines) are present to track the target. The gray dotted lines represent isochronic lines that connect eye and target at the same moment in time. Red arrows give the position error (PE) and retinal slip (RS) orientation before the saccade; blue arrows give the initial and final orientation of the eye at the start and end of the saccade, respectively, and black arrows indicate the direction of motion of the pursuit target. (1) Start of the trial: The eye smoothly pursues the rightward moving target. (2) Target step: From target step to the onset of the saccade, PE and RS are evaluated. (3) The saccade is elicited with an initial orientation close to the orientation of PE. (4) The saccade ends with a final orientation close to the orientation of RS. (5) After the saccade, pursuit is engaged to the target and the eye is very close to the target, as shown by the isochronic lines after the saccade. T, target; E, eye.

that is, the time that the eye trajectory would need to cross the target at constant velocity, which is used as a criterion to switch from pursuit mode to saccade mode. As long as the value of this sensory parameter remains in the smooth zone (between 40 and 180 ms), no saccade is needed to track the target. However, when the eye-crossing time takes values outside the smooth zone (i.e., inside the saccade zone), the decision to trigger a saccade is taken because the smooth pursuit system cannot compensate anymore for the position mismatch between eye and target. Therefore, a saccade takes place, on average, 125 ms later.

Finally, it has been shown that the saccadic and smooth pursuit systems cooperate despite the absence of retinal signals. For example, when directing gaze toward the remembered position of a flashed target, the saccadic system is able to take into account the smooth eye displacement (the integral of smooth eye velocity). This has been demonstrated during ongoing pursuit of a moving target but also in the dark during anticipatory smooth eye movements. In this case, the estimation of the smooth eye displacement is based on extraretinal information about the smooth eye movement. Furthermore, it has also been shown that during the temporary occlusion of a moving target, the saccadic system takes into account and can compensate for the modulation of the smooth response in complete darkness. These observations are evidence for the role of extraretinal signals and the existence of a velocity input to the saccadic system.

In conclusion, the saccadic system has access to retinal information about relative eye and target velocity (RS), as well as to extraretinal information about smooth eye velocity. This velocity input is available later than the position input, and these signals are combined to estimate the eye-crossing time, which determines whether a catch-up saccade should be triggered or not. The velocity input used by the saccadic system originates from the motion-processing pathway.

Position Input to the Smooth Pursuit System

The first reports of a position input to the smooth pursuit system in primates were based on indirect evidence because stationary targets do not evoke smooth eye movements in this species. In contrast, cats use a combination of small saccades and slow correcting eye movements to shift their gaze to an eccentric stationary target. Thus, studies in this species provided direct evidence of a position input to the smooth pursuit system. In primates, new paradigms needed to be developed to address this question specifically.

PE, either induced by target stabilization on to the retina or by target steps during ongoing pursuit, gave the first evidence of a position input to the primate smooth pursuit system. Target stabilization consists in making the target move at the same velocity as the eye, which stabilizes the projection of the target on to the retina. The stabilization of the target with a small offset with respect to eye position evokes slow eye movements directed toward the target either during fixation or ongoing pursuit. Similarly, target steps provided evidence of slow eye movements in response to a position error. If during smooth pursuit the target steps but continues to move at the same speed, slow eye movements are modulated by the amplitude of the step (i.e., the amplitude of the position error).

However, in these studies, the influence of PE was not direct or could be perturbed by other factors, such as concurrent RS. The first direct evidence of a

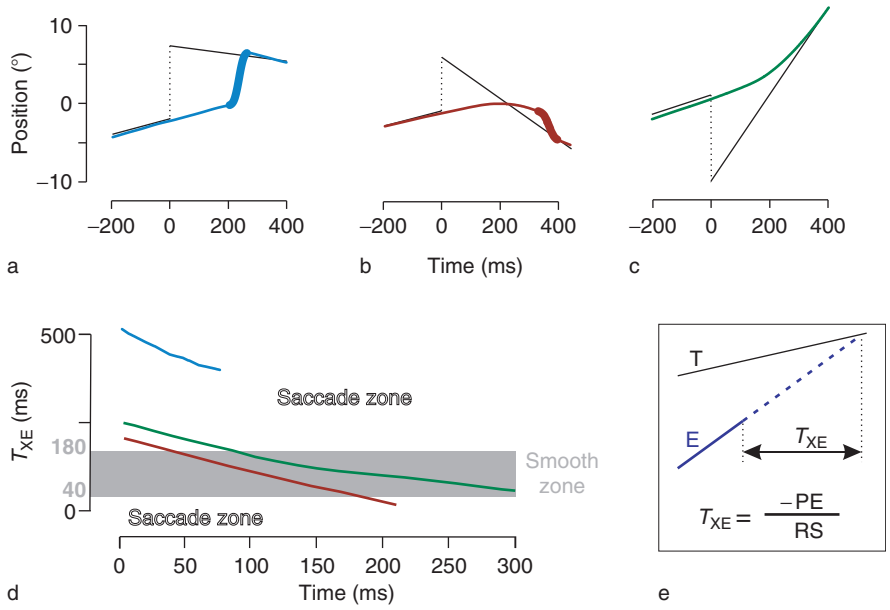

Figure 3 Saccade trigger: (a) early-saccade trial; (b) late-saccade trial; (c) smooth trial (no saccade); (d) the evolution of the eye-crossing time (T_{XE}) versus time for the three trials; (e) T_{XE}, the time that the eye trajectory would need to cross the target at constant velocity. Panels (a)–(c) show the position versus time representation of eye and target (black traces) during three different trials of a ramp–step–ramp paradigm along the horizontal axis. This paradigm consists of a moving target (first ramp) followed by a target jump (position step, represented by the dotted black lines) and a change in target velocity (velocity step) occurring at the same time (time zero on figure). The target then continues to move at constant velocity (second ramp). In panel (a), the latency of the saccade is around 200 ms; in panel (b), the latency of the saccade is around 350 ms. Panel (d) shows how the decision to trigger a catch-up saccade is taken based on T_{XE}. The T_{XE} is represented until 125 ms before the saccade. The smooth zone corresponds to the values of T_{XE} that do not trigger catch-up saccades because the pursuit system can compensate for position error (PE). In contrast, the saccade zone contains values of T_{XE} that trigger a catch-up saccade. For the early-saccade trial, T_{XE} lies far from the smooth zone and the decision to trigger a saccade is taken quickly. In contrast, for the late-saccade trial, T_{XE} lies outside the smooth zone at target step but soon enters the smooth zone for 150 ms before leaving it. As soon as the parameter leaves the smooth zone, the decision to trigger a saccade is taken. Finally, for the smooth trial, T_{XE} quickly enters the smooth zone but never leaves it; therefore, no saccade is triggered. E, eye; RS, retinal slip; T, target.

position input to the pursuit system was provided by flashing a target aside from the pursuit target path (see Figure 4). In this paradigm, subjects pursued a spot moving along a straight line. During pursuit maintenance, a second target was briefly flashed aside from the target path. This flash created a position error with a component orthogonal to the target path. This component of position error generated by the flash evoked smooth eye movements orthogonal to the direction of pursuit, directed toward the flash and scaled to the position error created by the flashed target.

Apart from all the behavioral experiments conducted to assess the influence of a position input on the smooth pursuit system, it has been shown that the stimulation of the cat SC, an area classically involved in the saccadic pathway, produces such postsaccadic slow eye movements. In addition, it has been demonstrated that the saccadic and smooth pursuit systems share a PE signal at the level of the SC in primates. In this species, recordings in the rostral SC showed that it codes small PEs during both smooth pursuit and saccades. Apart from well-established role of the SC in saccades, this shows evidence for the implication of the SC in saccade pursuit interaction.

In conclusion, evidence from several behavioral studies shows that PE influences the smooth pursuit system. Moreover, SC was proven to be the place where both saccade and pursuit systems share the PE signal, highlighting the role of SC in saccade–pursuit interaction.

Common Mechanisms for Initiation and Cancellation

Up to now, we have shown that both systems share common inputs. Indeed, PE and velocity error are used by both oculomotor subsystems. These two subsystems may either process their inputs independently or share some processes in transforming visual inputs into motor commands. We now review which processes are likely to be shared by both systems.

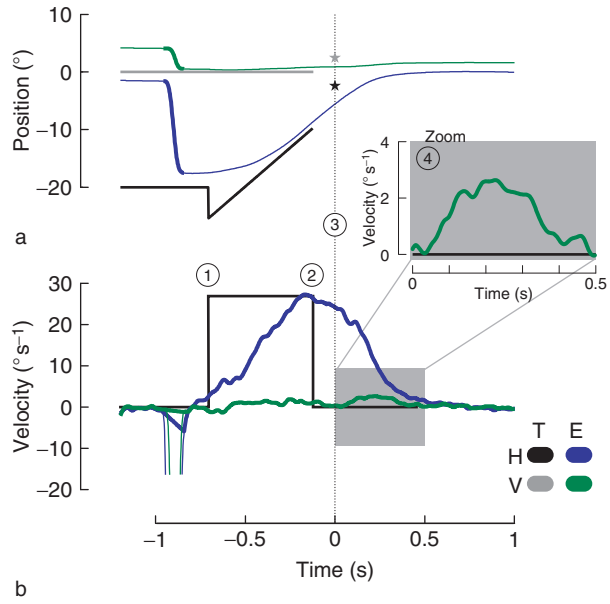

Figure 4 Position input to the pursuit system: (a) horizontal and vertical position of the eye and the target are presented as a function of time; (b) velocity of the eye and the target are presented as a function of time. The inset represents a zoom of the vertical eye velocity profile. (1) The pursuit target starts moving rightward. (2) The pursuit target disappears. (3) A target is flashed for 10 ms above the target path (slight vertical offset, see the gray (black) star that represents vertical (horizontal) flashed target position). (4) The vertical offset of the flash elicits a vertical smooth eye movement, which is orthogonal to the pursuit path. E, eye; H, horizontal; T, target; V, vertical.

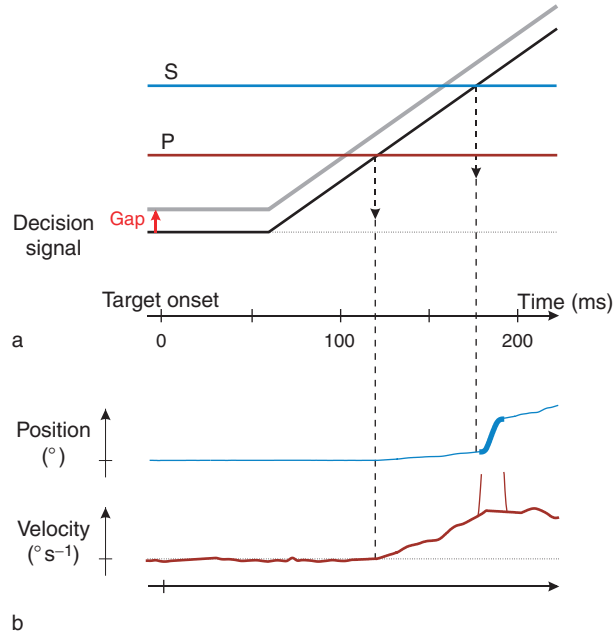

Figure 5 (a) Initiation mechanism; (b) corresponding eye position and velocity as a function of time. In panel (a), the initiation mechanism is represented by a linear rise to threshold process. After a given delay following target onset, the decision signal (black curve in no-gap condition) rises to two different thresholds (pursuit (P) in red and saccade (S) in blue). After the decision signal has reached a specific threshold, the corresponding eye movement is triggered (panel (b)). In the gap condition, the initial level of the decision signal is increased at target onset (gray curve).

Initiation

The latency of saccadic and pursuit eye movements have always been shown to differ. On average, the latency of saccades to stationary targets is around 200 ms, whereas the latency of smooth pursuit eye movements is around 100 ms. This discrepancy between smooth pursuit eye movements and saccades suggests that the initiation mechanisms are different. However, several paradigms have shown similar effects on the latency of both saccades and smooth pursuit. Their latency is shortened when the fixation point is extinguished prior to the appearance of the target (gap paradigm). The reduction in latency caused by the gap paradigm is very similar (around 50 ms) for both oculomotor subsystems. Neuronal correlates of the gap effect have been found in the SC for both types of eye movements when the buildup activity that occurs before saccade or pursuit is increased by the presence of a gap. Similar to the gap effect, the presence of a second moving target (a distractor) also modulates the latency of both pursuit and saccades in the same way (an increase in latency). Finally, when a first target is extinguished after the appearance of a new target (overlap paradigm), latencies of pursuit and saccades are comparable and highly correlated. These similar changes in latency suggest a common mechanism for the initiation of pursuit and saccades.

A linear rise to threshold model, in which a decision signal evolves toward a threshold, makes good prediction of saccade latency. As soon as this threshold is reached by the decision signal (see Figure 5), the intended action is initiated. In the case of saccade and pursuit initiation, a similar model can be used. A common decision signal linearly rises to two different thresholds (one for pursuit and a second higher one for saccades) to elicit eye movements. The crossing of one of the thresholds implies the trigger of the corresponding eye movement after some motor delay (different for pursuit and saccades). This model can account for the gap effect if it is assumed that the presence of a gap increases by the same amount the initial level of the decision signal and, hence, reduces the time needed to reach the threshold for both pursuit and saccades.

Cancellation

In a complex environment, it is often necessary to inhibit an action as circumstances that led to its

planning can change rapidly. In the oculomotor domain, a stop signal could be issued before the initiation of a movement, requiring the cancellation of the prepared eye movement. For both the saccadic and smooth pursuit systems, the probability of cancelation of the planned eye movement depends on the delay between the go signal (stimulus onset) and the stop signal. If the stop signal arises just after the go signal, the eye movement can easily be canceled; however, if it arises 200 ms later, it cannot. On average, the stop signal must be issued 60 ms before pursuit initiation for it to be canceled. Similarly, the interval between the stop signal and saccade initiation must be 80 ms to be efficient. In sum, it was found that it takes an additional 20 ms to cancel a saccade compared with smooth pursuit cancellation, suggesting that cancellation mechanisms are different for saccades and smooth pursuit.

At the neurophysiological level, cancellation of both saccadic and pursuit eye movements is mediated through omnipause neurons (OPNs, located in the brain stem). Indeed, OPNs exert an inhibitory control on the excitatory burst neurons that generate saccades, and their stimulation interrupts ongoing saccades. OPNs are also modulated during smooth pursuit eye movements, and their electrical stimulation reduces ongoing smooth eye velocity. Therefore, these results suggest that a common inhibitory group of neurons could regulate the cancellation of both saccades and pursuit. The common inhibitory neuronal population makes us think that there is a common cancellation mechanism, but it does not explain the additional delay required to cancel a saccade compared with canceling a pursuit.

Thus, the discrepancy between the cancellation delay for both saccades and smooth pursuit is probably not due to different cancellation mechanisms. Indeed, saccadic latency is measured with an eye acceleration or velocity threshold. However, 20 ms before saccade initiation, there is a point of no return for saccades. Indeed, around 20 ms before saccade onset, the OPNs are inhibited and excitatory burst neurons are activated. Therefore, under natural conditions, it is not possible to cancel saccades during this 20 ms interval. Such point of no return does not exist for pursuit. Therefore, when we take the point of no return into account instead of saccadic latency, the time needed to cancel any eye movement (either saccade or pursuit) is comparable (around 60 ms). In sum, the same inhibitory neurons could be involved in canceling saccades and pursuit eye movements. However, the details of the inhibitory process might differ between oculomotor subsystems.

Common Processes

In addition to shared inputs, saccade and pursuit systems share common processes. The initiation mechanism was proven to be common for saccades and pursuit on the basis of the similar modulation of saccade and pursuit latencies in different paradigms and on the basis of neurophysiological recordings in the SC. The proof of a common cancellation mechanism was yielded by a behavioral study on action inhibition and by a neurophysiological study in the brain stem. A third mechanism is also shared by saccadic and pursuit systems – the target-selection mechanism. This was demonstrated on the basis of target-selection paradigms and recordings in the SC (and not described here).

Other Brain Areas Subserving Both Saccades and Pursuit

In contradiction with the classical view of totally segregated systems (see Figure 1), the cortical networks of pursuit and saccades are anatomically overlapping. Among the cortical overlapping areas, some contain distinct neuronal populations for pursuit and saccades, such as the frontal eye fields (FEFs); others contain a neuronal population conveying both saccadic and pursuit signals, such as the supplementary eye fields (SEFs).

Subcortical areas generally contain a population of neurons subserving both saccades and pursuit. A subset of these subcortical areas that contain shared neuronal populations takes part in the descending pathway from the cerebral cortex toward the motor nuclei: the substantia nigra (SNr), the DLPN, the SC, and other brain stem regions (vestibular nuclei, nucleus of the optic tract, and interstitial nucleus of Cajal). Other areas containing neurons that convey both position and velocity signals take part in the ascending pathway that relays signals from premotor/motor nuclei toward cortical areas and/or form an internal feedback pathway of the ongoing eye movement: the cerebellum (vermis and fastigial oculomotor region) and the thalamus. On the other hand, the caudate nucleus (CN) contains two mainly segregated populations that convey pursuit and saccade signals separately.

Brain areas subserving saccades and pursuit contain either distinct neuronal populations for pursuit and saccades (the FEFs and CN) or a single population of neurons that conveys both pursuit and saccade signals (e.g., SNr, dorsal pontine nuclei, cerebellum, and thalamus), corresponding either to parallel or

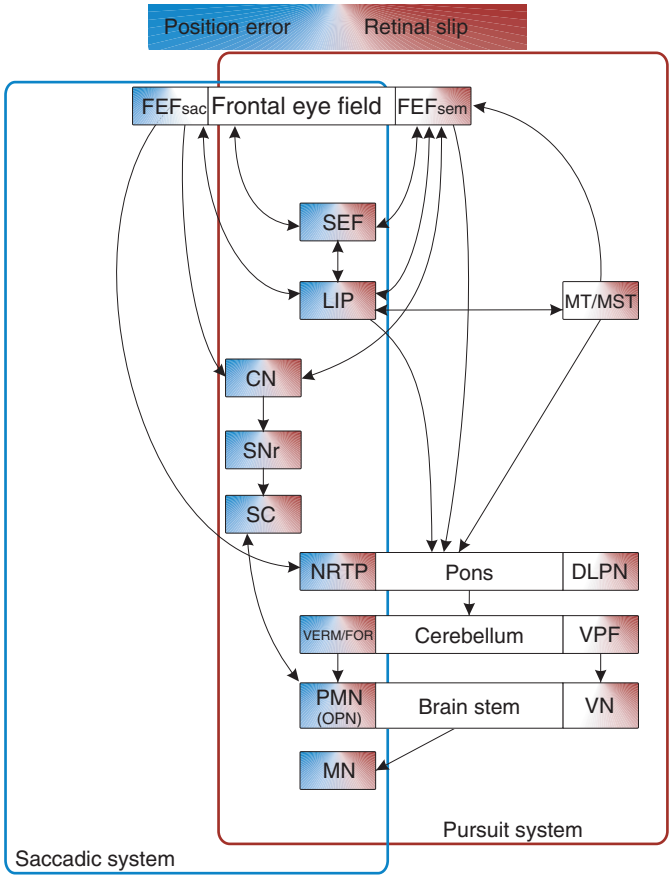

Figure 6 Diagram of the current view of saccadic and pursuit systems. Blue boxes are part of the saccadic system. Red boxes belong to the pursuit system. Two-color boxes are part of both systems. CN, caudate nucleus; DLPN, dorsolateral pontine nuclei, FEFsac, saccadic region of the frontal eye field; FEFsem, pursuit region of the frontal eye field; FOR, fastigial oculomotor region; LIP, lateral intraparietal area; MN, motor neurons; MST, medial superior temporal area; MT, middle temporal area; NRTP, nucleus reticularis tegmenti pontis; OPN, omnipause neurons; PMN, brain stem premotor nuclei; SC, superior colliculus; SEF, supplementary eye field; SNr, substantia nigra pars reticulata; VERM, vermis; VN, vestibular nuclei; VPF, ventral paraflocculus.

shared brain structures. The current view of the saccade and pursuit networks has thus evolved toward a more integrated network, subserving both oculomotor systems (see Figure 6).

Conclusion

In recent years, our view of the interaction between the smooth and saccadic systems has completely changed. From the two separate networks, a more unified scheme of the oculomotor system has emerged step by step. In this article, we highlight all the inputs, processes, and brain areas that are shared by the pursuit and saccadic systems. We emphasize that, even though they are computed by different networks, PE and RS influence both oculomotor subsystems. Finally, we underline the integration of common initiation and cancellation mechanisms and the numerous brain areas involved in both subsystems, even if, for the majority of them, their precise role in the saccade–pursuit interaction needs to be further investigated.

See also: Attention and Eye Movements; Oculomotor Control: Anatomical Pathways; Oculomotor System: Models; Pursuit Eye Movements; Saccades and Visual Search; Saccadic Eye Movements; Superior Colliculus; Target Selection for Pursuit and Saccades; Vergence Eye Movements.

Further Reading

Blohm G, Missal M, and Lefevre P (2005) Direct evidence for a position input to the smooth pursuit system. *Journal of Neurophysiology* 94: 712–721.

Blohm G, Missal M, and Lefevre P (2005) Processing of retinal and extraretinal signals for memory-guided saccades during smooth pursuit. *Journal of Neurophysiology* 93: 1510–1522.

Carpenter RH and Williams ML (1995) Neural computation of log likelihood in control of saccadic eye movements. *Nature* 377: 59–62.

de Brouwer S, Missal M, Barnes G, and Lefevre P (2002) Quantitative analysis of catch-up saccades during sustained pursuit. *Journal of Neurophysiology* 87: 1772–1780.

de Brouwer S, Missal M, and Lefevre P (2001) Role of retinal slip in the prediction of target motion during smooth and saccadic pursuit. *Journal of Neurophysiology* 86: 550–558.

de Brouwer S, Yuksel D, Blohm G, Missal M, and Lefevre P (2002) What triggers catch-up saccades during visual tracking? *Journal of Neurophysiology* 87: 1646–1650.

Kornylo K, Dill N, Saenz M, and Krauzlis RJ (2003) Cancelling of pursuit and saccadic eye movements in humans and monkeys. *Journal of Neurophysiology* 89: 2984–2999.

Krauzlis RJ (2003) Neuronal activity in the rostral superior colliculus related to the initiation of pursuit and saccadic eye movements. *Journal of Neuroscience* 23: 4333–4344.

Krauzlis RJ (2004) Recasting the smooth pursuit eye movement system. *Journal of Neurophysiology* 91: 591–603.

Krauzlis RJ (2005) The control of voluntary eye movements: new perspectives. *Neuroscientist* 11: 124–137.

Krauzlis RJ, Basso MA, and Wurtz RH (1997) Shared motor error for multiple eye movements. *Science* 276: 1693–1695.

Liston D and Krauzlis RJ (2005) Shared decision signal explains performance and timing of pursuit and saccadic eye movements. *Journal of Vision* 5: 678–689.

Missal M and Keller EL (2002) Common inhibitory mechanism for saccades and smooth-pursuit eye movements. *Journal of Neurophysiology* 88: 1880–1892.

Orban de Xivry JJ, Bennett SJ, Lefevre P, and Barnes GR (2006) Evidence for synergy between saccades and smooth pursuit during transient target disappearance. *Journal of Neurophysiology* 95: 418–427.

Rashbass C (1961) The relationship between saccadic and smooth tracking eye movements. *Journal of Physiology* 159: 326–338.

Schreiber C, Missal M, and Lefevre P (2006) Asynchrony between position and motion signals in the saccadic system. *Journal of Neurophysiology* 95: 960–969.

Tanaka M (2005) Involvement of the central thalamus in the control of smooth pursuit eye movements. *Journal of Neuroscience* 25: 5866–5876.

Thier P and Ilg UJ (2005) The neural basis of smooth-pursuit eye movements. *Current Opinion in Neurobiology* 15: 645–652.

Saccades and Visual Search

J M Findlay, University of Durham, Durham, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Visual search is a common visual activity that we engage in on a daily basis. For example, we spend time looking for a friend in the airport crowd, looking for our car in the parking lot, or looking for the tomatoes in the vegetable aisle at the supermarket. Indeed, it seems likely that visual search, and particularly the requirement to forage for food amid many nonfood items, has been a common form of behavior throughout evolutionary history. Exploration of the way in which visual search is conducted might therefore be expected to offer considerable insight into the workings of the visual system.

The extraordinary mobility of the eyes is well adapted to searching. The eyes move a few times each second, and these fast movements, known as saccades, direct the high-acuity foveal region to a succession of new visual locations for detailed inspection. Search is generally successful when the fovea is directed to the search target, since the high-resolution of foveal vision allows recognition that the target is located. The present article considers how the saccades that bring the eyes to the target are generated. Every saccade is a miniature behavioral act for which the brain makes a decision about the exact location in space to which the eyes should be redirected and a decision about the instant in time when the eyes should move. Saccades are sufficiently rapid that they cannot be guided by any visual information taken in while the movement is in progress, and so in general the saccade destination and trajectory are decided at the outset. The emphasis in this article is on human visual search, although the eye scans of monkeys trained in search tasks show considerable similarity to those of humans.

The major part of this article is devoted to consideration of eye scanning in rather artificial search displays consisting of a set of discrete items viewed against a plain homogeneous background. Usually one item is the search target and the remaining ones are nontargets, alternatively named distractors. Reference is first made to an important earlier tradition in which visual search was investigated with no explicit reference to eye scans. The final section discusses recent attempts to grapple with the additional factors that come into play when visual search occurs in more naturalistic visual scenes.

Search and Peripheral Vision

Consider the two tasks shown in Figure 1, each requiring search for a red cross. Figure 1(a) shows a display in which, for individuals with normal color vision, the target can be found with trivial ease. The target can be said to pop out of the display. In Figure 1(b), the search is clearly slower and more effortful. Around 1980, Anne Treisman suggested that, in the first case, the search was in some way carried out in parallel over the entire visual field, whereas in the second case, some serial attention-demanding process dealt with each item in turn. In support of the proposal she showed that, for displays of the type shown in Figure 1(a), the time to report the presence of a target was independent of the number of distractors present in the display, whereas for those of the type shown in Figure 1(b), the time increased linearly with the number of distractors present. For a number of years the distinction between parallel and serial processes formed the starting point for theories of visual search. It was proposed that search in the case of Figure 1(a) could be carried out 'preattentively' whereas searches such as in the case of Figure 1(b) required attention to focus on each item in sequence. It was also often believed that this attention scan could be achieved by a mental spotlight, and, for some years, these ideas were developed with little or no explicit reference to eye movements.

As empirical data accumulated, it became clear that, rather than a dichotomy, there is a continuum of visual search difficulty. Moreover, in the case of demanding searches such as those of Figure 1(b), several saccadic eye movements are normally required before the target can be found, whereas in the pop-out displays, a single saccade usually suffices to bring the eyes to the target. A critical factor is the ease with which the search target can be detected in peripheral vision. If a display is viewed briefly so that no eye movements can be made, the ease with which a target can be detected is dependent on the retinal location on which the target is projected. Detection becomes progressively more difficult with distance from the fovea (retinal eccentricity) and, in many cases, becomes impossible if the target is in the far periphery of vision. It is possible to define a functional field of view (sometimes termed the visual lobe) which describes the region where the target can be detected. Characteristically, this is shaped as a broad ellipse, centered on the fovea and extended along the horizontal axes. It is composed of two subregions. Within a central zone, the target can always be detected. This is surrounded by a transitional zone within which the

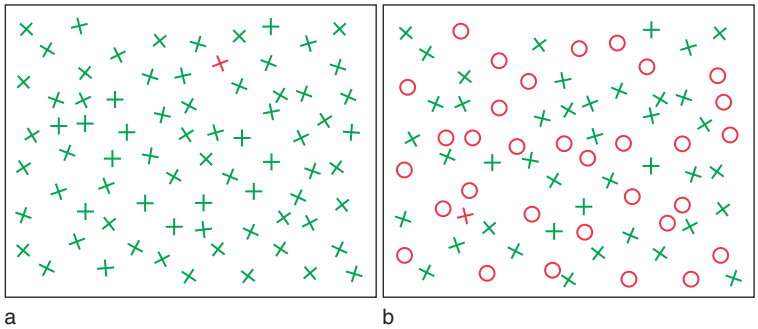

Figure 1 Demonstration of the distinction between preattentive and attentive search introduced by Anne Treisman. In both search displays the target to be searched for is a red cross. In (a) the target can be found easily and is said to pop out of the display. In (b) the search is more effortful and generally takes considerably longer.

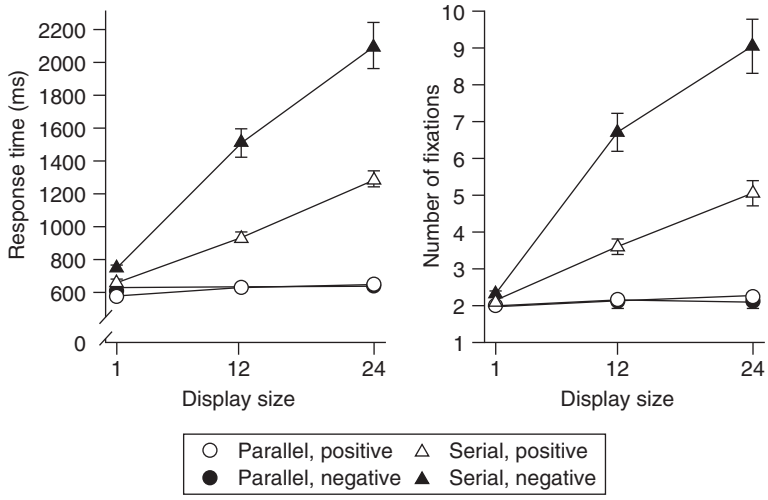

Figure 2 Results obtained from a visual search experiment with displays similar to those of **Figure 1(a)** for parallel search and **Figure 1(b)** for serial search. The task was to make a two-choice response indicating whether or not a target was present, and the searcher's eye movements were recorded. The left-hand plot shows a typical search function, giving the response times to make a target present response (positive) or a target absent response (negative). The right-hand plot shows the average number of saccades made during the searches in the different conditions. Reproduced from Williams DE, Reingold EM, Moscovitch M, et al. (1997) Patterns of eye movements during parallel and serial visual search tasks. *Canadian Journal of Experimental Psychology – Revue Canadienne De Psychologie Experimentale* 51: 151–164, with permission from Canadian Psychological Association.

target may be detected sometimes, but not always. Further out in the periphery is a region within which the target is never detected.

For displays that are large enough to extend beyond the functional field of view, it is evident that saccadic eye movements may be necessary to find the target. Study of such cases also shows that a major factor contributing to the time taken to detect the presence of a target is the number of eye saccades that are made during the search. This relationship has been shown both in humans and in nonhuman primates. Figure 2 shows data from a study in which the task was to make a two-choice Yes/No response indicating the presence or absence of a target in displays similar to those of Figure 1. The size of the display varied from trial to trial. The left-hand plot in Figure 2 shows the search function, in which the mean search time is plotted against the display size. The right-hand plot shows the mean number of fixations made. The similarity of the two plots is very apparent.

Target Selection for Saccades

Figure 3 shows a plot of the saccades made in one trial of a search task in which an observer, initially fixating at the center of the display, was asked to move the eyes to locate a red cross. The display was visible for 1 s, during which time the observer made five saccades, the third being a small correction-like movement. Against each fixation is noted its duration in milliseconds. Analysis of saccades in this task

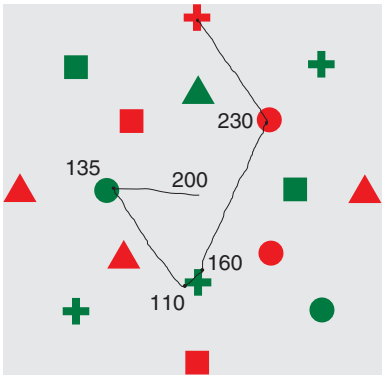

Figure 3 Example of an eye scan in a task in which the observer was asked to search for a red cross target. The observer initially fixated the center of the display, and the figures show the time in milliseconds for each fixation during the scan. Data from the study described in Findlay (1997); reproduced from Findlay JM (1997) Saccade target selection in visual search. *Vision Research* 37: 617–631, with permission.

revealed three important findings. First, if the target was in the inner ring, then it was frequently located with a single saccade. Second, saccades directed to the search target were made following a latency (time before making the movement) that was equal to, or slightly longer than, saccades made to distractors. Third, saccades made to distractors were more likely to land on elements sharing a common feature with the target than on elements with no feature in common (true for the second and fourth saccades in Figure 3, although not, in this case, for the first).

Similar findings emerged from several studies in the 1990s and were consistent with a view of brain function in relation to visual attention that was also developed in this decade. Visual processing in the brain occurs in a number of distinct cortical areas, in each of which the visual field (strictly hemifield) is represented in a retinotopic map. There are multiple cross-connections, often bidirectional, between these areas, but in each case the retinotopic mapping is maintained. This architecture promotes the operation of biased competition, a term introduced by Robert Desimone and John Duncan, who describe it as follows:

> Some kind of short-term description of the information currently needed must be used to control competitive bias in the visual system, such that inputs matching that description are favored in the visual cortex. (Desimone and Duncan, 1995: 199)

Visual cortical areas also connect in a retinotopically mapped way to three brain areas principally involved in saccadic eye movement generation, the frontal eye fields (FEFs), the lateral intraparietal area of posterior parietal cortex (LIP), and the superior colliculus (SC). The superior colliculus appears to be the brain center in which an irrevocable commitment to generate a saccade is normally made through the firing of a subpopulation of cells termed burst cells. Hence a plausible picture of saccade target selection in visual search is that the operation of biased competition results in a retinotopic representation in which neural activity codes similarity to the search target. The term salience map is often used to denote such a representation. A saccade is triggered when activity at some location on the salience map reaches a critical level. Because central regions of the visual field are exaggerated in each retinotopic map, proximity to the fovea is a further factor that combines with similarity to the target to determine the point of highest activity.

This general approach can account for each of the three findings from studies of saccades in visual search discussed at the beginning of this section. The approach maintains that visual information during visual search is always processed in parallel up to the point at which the decision is made to move the eyes. The antithesis to the approach is the alternative suggestion that some internal mental pointer (mental spotlight) is scanned around a number of locations during each fixation. The strongest evidence in favor of a fast-moving mental spotlight is the linear search functions of the type shown in Figure 2(a). As discussed previously, in the case of searches where eye movements are occurring, the linear increase in search time corresponds to an increasing number of saccades. Linear functions have also been reported in situations where eye movements are excluded, but it has been argued that these also can be accounted for in terms of the parallel processing approach. The argument runs along the lines that the target–distractor discrimination is a process of signal detection against noise. As the number of distractor (noise) elements increases, the task becomes more difficult and more time is needed to achieve a particular performance criterion.

While the general approach is widely accepted as an account of how saccade selection occurs when searching displays of the type shown in Figure 1, there is considerable debate concerning how the details of the biased competition process might be implemented. One issue concerns the degree of resolution possible. It is known that much of the visual information reaching the saccadic system is coded in a distributed form (large receptive fields; coarse or population coding). This form of coding does not preclude accurate localization in the case where a single isolated target is present. However, when two targets, or a target and a high-similarity distractor, occur in neighboring locations, the use of distributed coding can lead to the center-of-gravity phenomenon whereby an inaccurate first saccade lands between the two potential targets and is followed by further

corrective saccades. Another issue concerns how different features, which can potentially be selected for in different cortical maps, are reintegrated into an overall salience map. Thus, for example, one suggestion, that the salience map should be based on the squared Euclidian distance from the target in feature space, predicts that search for feature conjunctions should be more efficient than occurs in practice.

The salience map approach gives a clear description suggesting how bottom-up factors from the immediate stimulus display can be processed in a manner consistent with physiological findings. It is also straightforward in principle to see that top-down factors could be melded with these bottom-up factors through differential weighting in the final salience map. The influence of top-down factors will be discussed following a section considering the timing of the eye movements.

The Decision about When the Eyes Should Move

Our active vision system is capable of exquisite control of the duration of each eye pause, or fixation, between successive saccadic movements. This has been shown particularly clearly in studies of eye control during reading, where it is found that fixation durations are sensitively adapted to the difficulty of words encountered in the text. Similar effects occur in studies of visual search. Fixation durations in visual search tasks (typically 150–300 ms) are similar or slightly longer than those in reading and also increase when search is more difficult – for example, when the similarity between target and distractors increases.

In studies of reading, the relation between word difficulty and fixation duration has been found to occur both on a moment-to-moment basis, with longer fixations occurring on unusual or contextually improbable words, and on a longer term basis, with reading speed adapting to general text difficulty. An effect similar to the long-term adjustment has been noted in visual search for an oddity target (for example, a red target surrounded by a set of green distractors). If the color of the target to be found is known in advance, it is much easier to get the eyes on to the target than if the only information guiding search is that the target is different. Correspondingly, the latency to get the eyes under way is shorter in blocks of an oddity search task where all targets are of the same color than those when the target color changes unpredictably from trial to trial. Furthermore, if, in the latter case, a run of several trials occurs in which the target color remains constant, the latency is found to decrease during this sequence. This automatic priming effect of prior trials (termed priming of pop-out) shows how the system controlling when the eyes are moved can be adaptable in a flexible and efficient way.

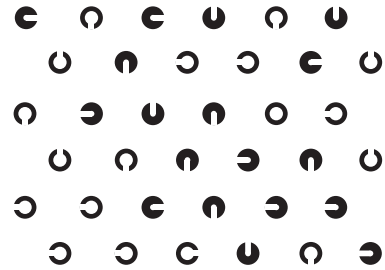

Figure 4 Example of a search display used to study the influence of various factors on eye fixation duration during search. The search target was a complete ring with no gap. From Hooge ITC and Erkelens CJ (1998) Adjustment of fixation duration in visual search. *Vision Research* 38: 1295–1302.

How do factors relating to the immediate search display affect fixation duration? This question was thoroughly investigated by Ignace Hooge and Casper Erkelens using ingeniously designed search displays of the type shown in Figure 4. The search elements were Landolt C discrimination targets and the task was to search for an element that was a complete ring without a gap. The rings were of two thicknesses, fat and thin, and the observers knew that the target would always be a thin ring. Thus the fatter rings were always distractors, and could in principle be identified as such in peripheral vision. The task allowed independent control of the difficulty of the discrimination between fat and thin rings, available in peripheral vision, and the difficulty of discrimination between thin ring distractors and the target, for which foveal or near-foveal vision was required.

It was found that the fixation duration on distractors varied in a systematic way. Fixation durations were always shortest on fat ring distractors, and those on thin ring distractors became longer with shorter gaps and thus with the difficulty of the foveal discrimination needed to distinguish the distractor from the target. The width of the fat rings was varied between blocks of trials, thus varying the difficulty of the peripheral discrimination. This factor did not affect the fixation duration, although it did affect the proportion of times that fat ring distractors were fixated. The conclusion, supported by other studies, is that fixation duration depends on the information processing required in central vision, but is unaffected by what processing is carried out in the parafovea or periphery.

Pipelined Saccade Planning

The picture developed so far has shown how individual saccades may be programmed to be responsive to the

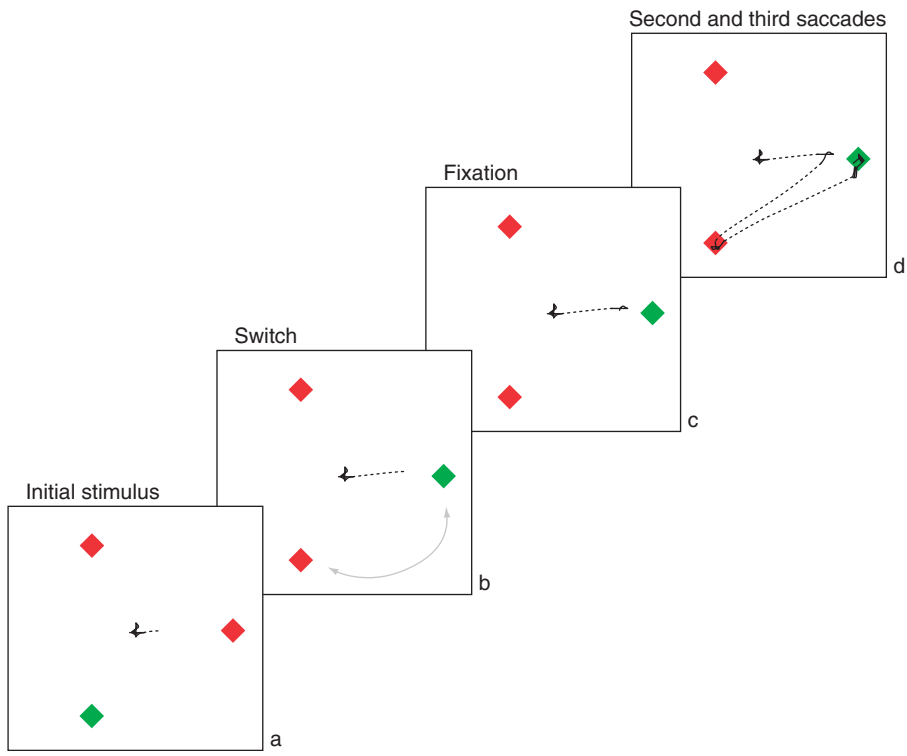

Figure 5 Sequence of displays used in a study by McPeek, Skavenski, and Nakayama, with eye scan traces superimposed. Displays always consisted of three elements and the task was to move the eyes to the odd-colored element, which could be either red or green. Immediately after the first saccade was initiated (switch frame) the display altered (two elements were interchanged). The traces show eye movements in a typical trial in which, although the first saccade landed close to the target location in the changed display, nevertheless a further saccade was made to the original target location. From McPeek RM, Skavenski AA, and Nakayama K (2000) Concurrent processing of saccades in visual search. *Vision Research* 40: 2499–2516.

immediate visual information. This basic pattern must now be supplemented in a variety of ways that allow more long-term influences to affect the eye scanning.

The first consideration is that more than one saccade can be planned at once. A particularly clear demonstration, made by Rob McPeek and collaborators, and termed by them pipelined planning, is shown in Figure 5. The task was an oddity search. On each trial three colored shapes arranged in a triangular array were presented, two red and one green, or one red and two green. The task was to move the eyes to the oddball target, a task that is in fact surprisingly difficult. On some trials, immediately after the eyes started to move, two of the elements were interchanged so the position of the oddball changed. In the case shown in Figure 5 the switch resulted in the target appearing at the destination of the initial saccade (panel b). However the eyes did not stop at this location but made a further saccade to the previous location of the target before a third and final saccade brought the eye back to the new target location. The implication is inescapable that the second saccade had been planned during the fixation before the first saccade. This was also demonstrated by the finding that the interval between first and second saccades was often extremely short. In this and similar demonstrations, the two saccades of interest are generated immediately following display occurrence. There is accumulating evidence that in this situation, one saccade is generated as a reflexive orienting movement responding to the rapid transitory visual stimulation from the display onset, while the other is a more volitional (intentional) movement. Saccades with short latency tend to be of the reflexive type and those with longer latencies are more likely to be volitional. It remains to be established whether such double planning is characteristic of search scanning generally. Fixations of short duration are often observed but these are most frequent before a small corrective saccade, without the selection of a new target.

How Are Saccades Distributed Over a Search Array?

In this section, processes are considered that contribute to search efficiency by distributing scanning to ensure that all parts of the search array are covered. It has been argued in the previous section that saccades are guided on the basis of the two factors derived from the immediate visual stimulation:

proximity to the current fixation location and similarity to the search target. A system that used these factors solely would encounter a problem if two items that were similar to the target were present in close proximity and at some distance from other potential targets. When item A was fixated, item B would be the most salient target, and when a saccade was made to B, then A would become the most salient. The system would be trapped into a salience loop. In the next two sections we consider how such loops are avoided. Obviously, what is needed is some form of memory for items already examined.

The phenomenon of inhibition of return (IOR) appears to be one attractive candidate for such a memory. IOR is a process whereby an item or location that has received attention is tagged in some inhibitory way so that return responses are slowed or rendered less likely. Ray Klein has proposed that it acts as a foraging facilitator, causing return saccades to be avoided. However, this has been disputed, since in fact, saccades returning to the immediately prior scanned item are quite commonly found in search scanpaths (3–4% of saccades is a typical figure), although return saccades to items further back in the sequence are rarer. A resolution of the conflict might be possible if IOR operated in a flexible way. It has been already shown that the time course of IOR is affected by the attentional demands of the material. Thus, rather than being generated automatically at a fixed time after the attention shift, IOR might operate in a manner sufficiently flexible that immediate returns to previously processed items were not precluded. Return saccades would then simply be a reflection of the fact that an item had been incompletely processed on first scan.

Other processes undoubtedly supplement IOR. It is evident that voluntary strategies are potentially available. For example, a regular array based on a rectangular grid could be searched with a raster-scan pattern, as in reading. Adoption of such a strategy ensures that the array is fully sampled and does not require a process such as IOR to avoid backtracking, but this is at the expense of abandoning any general guidance from peripheral vision. With less regular arrays, a similar regular scan could operate if the searcher were somehow able to use an imagined grid. In a recent study, a search task was given requiring search through a small- to medium-sized set of items (up to 12), all perceptually similar. Participants were free to adopt whatever search strategy they wished. Half the participants tested opted for a raster-like strategy. The other half generated a search which was almost as efficient, but they searched much more heuristically, making some use of the external overall contour of the array.

Learning of Contingencies

In the studies so far described, the search target might occur with equal likelihood at all locations within the display. This is, of course, not at all characteristic of more everyday search situations, where some parts of the scene would be very unlikely locations for a target. It might therefore be expected that, when a structured scene is being searched, such contingencies would influence search scanning.

Unsurprisingly, in situations where the target is likely to occur in one region, people are indeed found to concentrate their scanning in this region. However, an important discovery was that more complex contingencies can also be learned readily. These can be demonstrated even in unstructured displays and can occur without conscious awareness. In the late 1990s, Marvin Chun and Yuhong Jiang asked individuals to carry out a search task where the target was a T shape which could be in two orientations (rotated 90° clockwise or counterclockwise) and the distractors were L shapes in eight possible orientations. Target and distractors appeared at locations randomly chosen from a large set and their color was also varied. The task was to find the T shape and report its orientation. During the session, certain displays were regularly repeated. In comparison with displays not repeated, the search time on repeated displays became significantly shorter. When questioned following the experiments, participants showed no awareness that some displays had been repeated. This demonstration shows a learning process at work whereby a successful search for a target at a location in some way renders search more efficient for a target at this location when the display is encountered again. Furthermore, this learning is of an implicit nature, operating below the level of the searcher's conscious awareness.

A subsequent study by Matthew Peterson and Arthur Kramer repeated a similar experiment but in addition recorded the searchers' eye movements. They confirmed that searches were faster in repeated displays and showed that even the very first saccade made after a repeated display was presented was slightly more likely to land on the target than was the case for a novel display. The manner in which the implicit learning guided the eyes appeared to be precise. Over the first few search fixations, saccades were more likely to land on the target in the repeated displays than in the novel displays, but there was not any increased tendency to land on nearby distractors.

Searching in Coherent Scenes

In recent years several groups have developed eye recording technologies that allow gaze recordings to

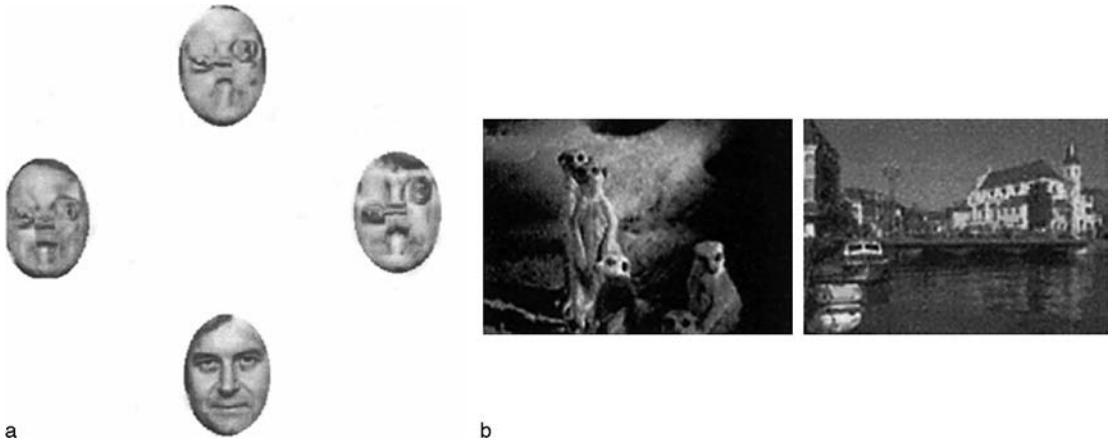

Figure 6 Two examples of search tasks designed to study the ability to use high-level perceptual features to direct saccades. In each case the observers initially fixated centrally. (a) The task was to make a saccade to the intact face. (b) The task was to make a saccade to the frame that contained an animal. (a) Reproduced from Brown V, Huey D, and Findlay JM (1997) Face detection in peripheral vision. Do faces pop out? *Perception* 26: 1555–1570, with permission. (b) Reproduced from Kirchner H and Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research* 46: 1762–1776, with permission.

be made while individuals carry out typical real world tasks (e.g., tea making, food preparation). These show an enormously impressive ability to direct gaze sequentially at the appropriate items needed to complete the task, with fixations on extraneous distractor items being very rare. Although these tasks are not primarily search tasks, they contain search requirements and the findings show the very powerful extent to which stored knowledge about the environment can direct eye fixations. In contrast to the contingency learning discussed in the previous section, memory for coherent scenes is available for conscious recall, with the capacity for storage and retrieval of thousands of different scenes being readily demonstrated. It is beyond the scope of the current article to discuss how this form of scene context might guide individual saccades, but an issue that relates more closely to the theme concerns whether complex semantic information can be extracted from peripheral vision rapidly enough to guide the subsequent saccade. Figure 6 shows material from two studies that tested this possibility and led to different outcomes.

In both studies the observer fixated centrally and displays of the type shown appeared. The task was to make a direct saccade to the designated search target. In the first study, the target was a face and the distractors were scrambled faces. Naive participants initiated a saccade rapidly but these saccades did not land on the target more often than on a distractor. Training resulted in greatly improved performance, with subjects withholding the immediate saccade. In the second study, a scene containing animals was paired with a distractor scene without animals. The task was to move the eyes to the animal scene and careful controls were used to prevent the discrimination being possible on any simpler visual factors. In this task, high accuracy (~90%) was obtained with short saccade latency (~200 ms).

These studies give the flavor of current research in the area, in which increasingly realistic visual material is being used. There is now a challenge for neuroscientists to develop an understanding of the underlying brain processes that operate in these tasks so as to extend the accounts that have been developed successfully for simpler search tasks.

See also: Attention and Eye Movements; Basal Ganglia and Oculomotor Control; Eye and Head Movements; Eye Movement Disorders; Frontal Eye Fields; Saccade–Pursuit Interactions; Saccadic Eye Movements; Superior Colliculus; Target Selection for Pursuit and Saccades; Vision for Action and Perception.

Further Reading

Bichot N, Rossi AF, and Desimone R (2005) Parallel and serial neural mechanisms for visual search in macaque area V4. *Science* 308: 529–534.

Brown V, Huey D, and Findlay JM (1997) Face detection in peripheral vision. Do faces pop out? *Perception* 26: 1555–1570.

Chun MM and Jiang Y (1998) Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology* 36: 28–71.

Desimone R and Duncan J (1995) Neural mechanisms of selective attention. *Annual Review of Neuroscience* 18: 193–222.

Fecteau JH and Munoz DP (2006) Salience, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences* 10: 382–390.

Findlay JM (1997) Saccade target selection in visual search. *Vision Research* 37: 617–631.

Findlay JM and Brown V (2006) Eye scanning of multi-element displays. I. Scanpath planning. *Vision Research* 46: 179–195.

Findlay JM and Gilchrist ID (2003) *Active Vision: The Psychology of Looking and Seeing*. Oxford: Oxford University Press.

Hooge ITC and Erkelens CJ (1998) Adjustment of fixation duration in visual search. *Vision Research* 38: 1295–1302.

Kirchner H and Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research* 46: 1762–1776.

Klein RM and MacInnes WJ (1999) Inhibition of return is a foraging facilitator in visual search. *Psychological Science* 10: 346–352.

McPeek RM, Skavenski AA, and Nakayama K (2000) Concurrent processing of saccades in visual search. *Vision Research* 40: 2499–2516.

Motter BC and Belky EJ (1998) The guidance of eye movements during active visual search. *Vision Research* 38: 1805–1815.

Navalpakkam V and Itti L (2005) Modeling the influence of task on attention. *Vision Research* 45: 205–231.

Peterson MS and Kramer AF (2001) Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception and Psychophysics* 63: 1239–1249.

Treisman AM and Gelade G (1980) A feature integration theory of attention. *Cognitive Psychology* 12: 97–136.

Williams DE, Reingold EM, Moscovitch M, et al. (1997) Patterns of eye movements during parallel and serial visual search tasks. *Canadian Journal of Experimental Psychology – Revue Canadienne De Psychologie Experimentale* 51: 151–164.

Wolfe JM (1998) What can 1 million trials tell us about visual search? *Psychological Science* 9: 33–40.

Saccadic Eye Movements

L Goffart, CNRS, Marseille, France

© 2009 Elsevier Ltd. All rights reserved.

The sudden appearance of an object in the visual field triggers an orienting shift of the line of sight (gaze) toward its location. This movement is an extremely rapid, steplike rotation of the two eyes – the saccade – which can sometimes be accompanied by a head movement. By positioning the object's image on the retinal region of highest acuity – the fovea – the saccade permits a more refined visual analysis of the object. But saccades are not exclusively aimed at objects that suddenly appear; scanning saccades are also generated when we explore the environment. In this case, saccade goals must be extracted from the visual scene based on their saliency or relevance for subsequent actions (e.g., grasping an object, heading toward a location, avoiding an obstacle, or reading the next word in a sentence).

Studying the neural control of saccades is a choice research theme for investigating how the central nervous system controls goal directed action, not only because the saccade is an essential movement used by vertebrates to spatially organize their sensory experience but also because it is a relatively simple model of the transformations that must be applied to sensory signals in order to generate motor commands for reaching a goal. This research is facilitated by the fact that eye movements can nowadays be recorded with very good spatial and temporal resolutions. Figure 1 illustrates the trajectory and the time course of a saccade aimed at a small target appearing in the peripheral visual field. The trajectory is shown by plotting the vertical versus the horizontal angular deviation (also commonly called position) of the eye (Figure 1(a)). The time course shows that, after a latency (reaction time) of approximately 150 ms, the change in eye orientation is extremely rapid (velocity profiles in Figure 1(c)) and ends with the maintenance of the eye in a new orientation (Figure 1(b)). It also shows that a secondary correction saccade sometimes follows the primary saccade.

This apparently very simple movement is the result of complex muscular and neural activities with a functional organization that is now fairly well understood. In this article instead of describing the path leading from the target-evoked retinal activity to the changes in muscle tension that rotate the eyes, we take the reverse path. Starting from the muscle contractions, we proceed upstream and present the neural mechanisms involved in generating the commands for saccadic eye movements.

Extraocular Muscles and Rotation of the Eyes

Saccades are elicited by the cooperative action of six muscles attached to the eyeball. These muscles are called extraocular muscles, to distinguish them from those located inside the eyeball which influence the diameter of the pupil or the lens curvature. Two pairs of muscles are involved during horizontal rotations of the eyes (i.e., rotations around a vertical axis through the eyeball): the lateral rectus (LR) and medial rectus (MR) muscles. Orienting the right eye toward the right is accomplished by the contraction of the LR and the relaxation of the MR (Figure 2(a)) while in the left eye the MR is contracted and the LR relaxed. Conversely, shifting the gaze toward the left is produced by MR contraction and LR relaxation in the right eye (Figure 2(b)) concomitantly with LR contraction and MR relaxation in the left eye. The muscles that contract are called the agonist muscles; those that relax are called the antagonist muscles.

The muscle contraction generates a torque which is delivered to the eyeball through tendinous attachments. Figure 2(b) shows the tension developed by the LR and MR muscles during a horizontal saccade. The agonist muscle tension waveform first shows a rapid rise in tension, which is followed by a quasi-exponential decay until a steady-state tension is reached that maintains the eye deviated in the orbit. For the antagonist muscle, the tension waveform shows a symmetrical pattern, with a sudden drop in tension followed by an exponential increase until the steady-state tension is also reached. These muscle tension recordings reveal that, after every saccade, the orientation of the eye is settled well before the agonist and antagonist muscle forces reach their final equilibrium. They also illustrate the three components of the innervation signal supplied to the extraocular muscles: a pulse, a slide, and a step component. The pulse and the slide quickly rotate the eyes and stop them abruptly against the viscosity of the muscles and orbital tissues, respectively. The step generates the force required to hold the eyes deviated in the orbit, overcoming the elastic forces that tend to drag the eyes toward their resting position.

Vertical movements of the eyes (rotations about a horizontal axis through the eyeball) are slightly more complex than horizontal ones because they are the result of a synergy between four other extraocular muscles. The joint contraction of the superior rectus (SR) and inferior oblique (IO) muscles of both eyes, combined with the relaxation of the inferior rectus (IR) and superior oblique (SO) muscles, rotates the

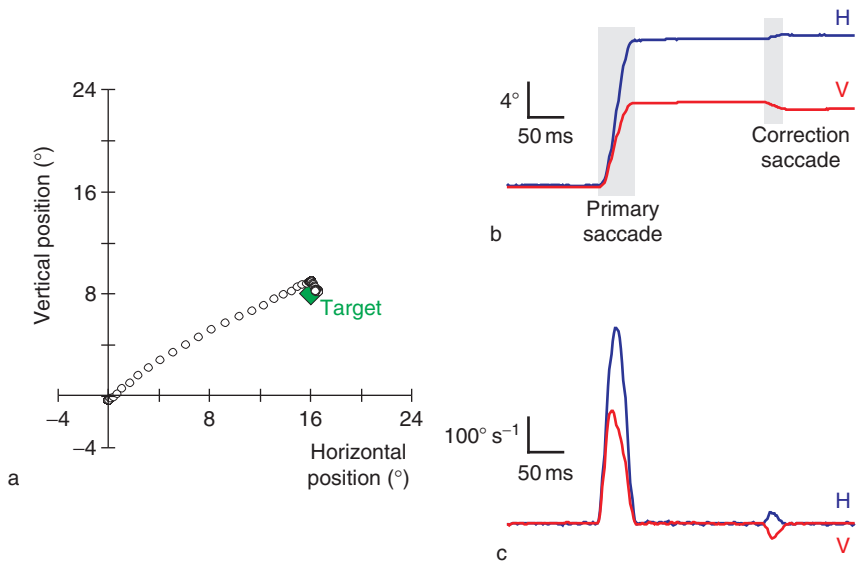

Figure 1 Oblique saccade: (a) spatial trajectory; (b) time course of eye position; (c) time course of eye velocity. The oblique saccade is initiated from a fixation stimulus located straight ahead (intersection between the axes in (a)) and aimed at peripheral target located 16° to the right and 8° upward (green diamond). Positive values of horizontal (H) and vertical (V) positions correspond to rightward and upward eye deviations in the orbit.

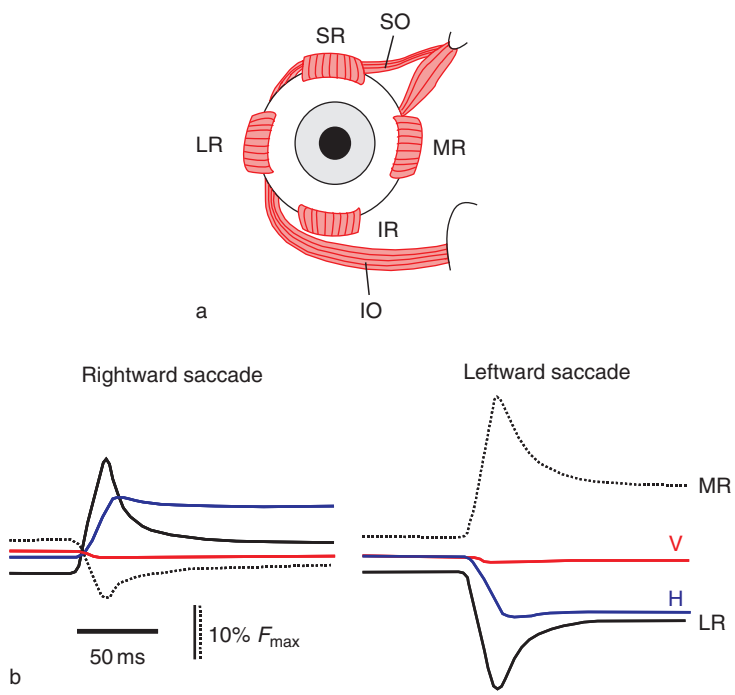

Figure 2 (a) Extraocular muscles of the right eye; (b) changes in LR and MR muscles tensions during a 16° horizontal saccade. Forces (black traces) are given as percentage of the maximum force (F_{max}). Upward deflection corresponds to muscle contraction; downward deflection corresponds to muscle relaxation. H, horizontal position; IO, inferior oblique; IR, inferior rectus; LR, lateral rectus; MR, medial rectus; SO, superior oblique; SR, superior rectus; V, vertical position. (b) Reproduced from Miller JM and Robins D (1992) Extraocular muscle forces in alert monkey. *Vision Research* 32: 1099–1113, with permission from Elsevier and the authors.

eyes upward. The bilateral contraction of the IR and SO, combined with the relaxation of SR and IO, rotates them downward. To date, there is no recording of the forces developed by these muscles during vertical or oblique saccades. Oblique saccades are generated by tensions simultaneously developed by the muscles rotating the eye horizontally and those rotating the eye vertically.

The Saccade Generators

The force generated by extraocular muscles is driven by the activity of motor neurons (MNs) located in cranial nerve nuclei III (oculomotor), IV (trochlear), and VI (abducens). The schema in Figure 3 shows the current view of the network of motor and premotor neurons which are involved in generating a rightward horizontal saccade (leftward saccades involve a similar neuronal network but located in contralateral territories of the reticular formation). The recruitment of MNs driving the agonist muscles is ensured by a population of excitatory burst neurons (EBNs) located in the ipsilateral pontine reticular formation (path p1). These EBNs drive the activity of abducens MNs innervating the LR of the ipsilateral eye. For contracting the MR in the contralateral eye, the EBN command is relayed to MNs in the oculomotor nucleus by internuclear neurons (INs), also located in the abducens nucleus. In parallel, EBNs activate a population of inhibitory burst neurons (IBNs) located in the ipsilateral medullary reticular formation and which inhibit the activity of EBNs and MNs in the contralateral side (path p2). The burst of ipsilateral IBNs thereby participates in the inhibition of MNs innervating the antagonist muscles (the LR of the left eye and MR of the right eye). In this way, the antagonist muscles relax while the agonist muscles contract. Some neurons in the IBN area also burst during contralateral saccades. If these neurons also inhibit contralateral MNs, the motor command sent to the MNs innervating the agonist muscles must be viewed as the combination of two inputs: an excitatory input from ipsilateral EBNs (path p1) and an inhibitory input from the contralateral IBNs (path p3).

The activity of EBNs and IBNs located in the left and right pontomedullary reticular formations is silenced between every saccade (e.g., when the eyes fixate) through the activity of another class of inhibitory neurons called omnipause neurons (OPNs), which are located in the nuclei raphe interpositus. These OPNs display a sustained activity which is interrupted every time a saccade is generated, irrespective of the amplitude and direction. Released from this sustained inhibition, EBNs and IBNs drive the MNs, which results in a saccadic eye movement.

During upward and downward saccades, different neuronal populations are involved. They are located in the rostral interstitial nucleus of the medial longitudinal fasciculus (riMLF), and their organization has not yet been completely elucidated. One major difference between the neural control of horizontal and vertical saccades is that the former are mostly driven by a unilateral excitatory input (from the ipsilateral pontine reticular region), whereas the latter are driven by a bilateral excitatory input from the left and right riMLF. The direction of oblique saccades is determined by the activation ratios between a unilateral horizontal saccade generator (left or right) and a bilateral vertical saccade generator (up or down). In normal healthy subjects, the examination of

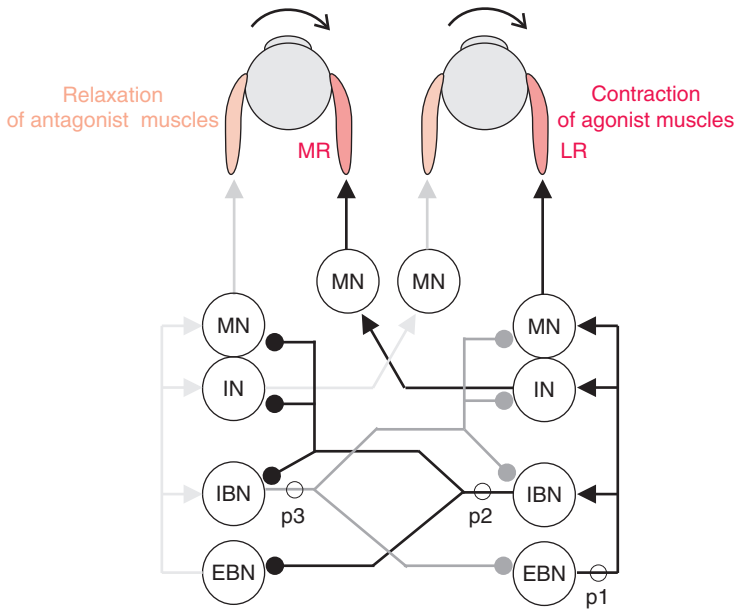

Figure 3 Neuronal organization involved during a rightward saccade. Arrows indicate excitatory connections; connecting lines ended by a circle indicate inhibitory connections. The connecting lines have different colors for clarity. EBN, excitatory burst neuron; IBN, inhibitory burst neuron; IN, interneuron; LR, lateral rectus; MN, motor neuron; MR, medial rectus; p1, path 1; p2, path 2; p3, path 3.

oblique saccades shows that both the horizontal and vertical rotations are initiated simultaneously, not sequentially (see Figure 1). Moreover, some observations show that the horizontal and vertical components are coupled – when the velocity of one component is reduced, the other component is also slowed down, thereby reducing the delay between the completion of horizontal and vertical components. Despite this component-stretching mechanism, the horizontal and vertical components may end at slightly different times. The analysis of saccades generated by neurology patients is very instructive in this matter. For example, in patients suffering from Niemann–Pick type C disease, the horizontal component precedes the vertical component, resulting in the completion of the vertical component well after the end of the horizontal component.

In Figure 3, the same network is involved for a horizontal saccade aimed at an object located 4° or 16° to the right. Additional mechanisms must be added to this network to generate saccades of different amplitudes and durations. In the next section, we describe what is known about how the central nervous system controls the amplitude and duration of saccades.

The Control of Saccade Execution

The Duration and Velocity of Saccades

In almost all studies, the duration of a saccade is defined as the time elapsed from when the eye velocity exceeds a certain threshold (e.g., $15° s^{-1}$) to when it returns below it. The duration of saccades increases linearly with saccade amplitude, ranging from 1° to 50° at a rate of approximately 2–3 ms deg^{-1} in humans (1–2 ms deg^{-1} in monkeys). For saccades less than approximately 30°, the velocity profile is bell-shaped (see Figure 1(c)), with a rapid rise to a peak value (acceleration phase) followed by a symmetrical decline in velocity (deceleration phase). For larger saccades, the velocity profile becomes skewed, with the duration of deceleration phase being longer than the acceleration phase. The peak velocity increases linearly with saccade amplitude until a value of approximately 15°, beyond which the increase in velocity begins to saturate to a maximum value as large as $700° s^{-1}$ ($900° s^{-1}$ in monkeys). These relationships between the amplitude and the duration or peak velocity of saccades are called the main sequences. Because their graphical shapes are relatively similar across healthy subjects, the main sequences have become a tool for studying the oculomotor deficits in brain-lesioned patients. Some authors have also suggested that common physiological systems are responsible for the generation of reactive saccades and voluntary saccades, as well as for the fast phases of optokinetic and rotatory nystagmi.

However, these duration–amplitude and peak velocity–amplitude relationships show considerable variability between subjects and even within the same subject. They are sensitive to alertness, motivation, and several other parameters such as saccade direction, initial deviation of the eyes in the orbit, and the testing conditions (i.e., whether the saccade is memory-guided or immediate). Saccade velocity also depends on the sensory nature of the target. Saccades toward somatosensory and auditory targets are much slower than saccades of comparable amplitude and direction aimed at visual targets. Saccades to moving targets are also slower than saccades of matched amplitude toward stationary targets and are characterized by a lower peak velocity and a prolonged deceleration phase. In some cases, a second peak eye velocity appears during the deceleration phase, presumably reflecting the late influence of a mechanism that compensates for the target displacement occurring before saccade end. Similar late reacceleration peaks are observed in saccades that accompany head movements during very large gaze shifts.

The Spatiotemporal Transformation

On the motor output side, the duration and velocity of saccades are defined by the temporal properties of MN activity. For saccades initiated from the straight ahead position in the orbit, different saccade amplitudes do not involve different groups of MNs; all MNs are recruited. Saccades of increasing amplitude are caused by increases in the duration of MN activity, the instantaneous firing rate dictating instantaneous saccade velocity (rate code). On the other side, at the sensory level, the code is topographic; targets at different locations in the visual field activate different zones in the retina. A topographical organization is found in several regions in the brain, including the intermediate and deep layers of the superior colliculi, the neurons of which are located two to three synapses from the motor nuclei. From a macroscopic point of view, the two deep superior colliculi (dSC) form a motor map, which appears to be in agreement with the retinal map, with the anterior (rostral) regions of both colliculi representing the foveal and perifoveal visual fields, posterior (caudal) regions representing the peripheral visual field, medial regions representing the upper visual field, and the lateral regions representing the lower visual field. In the dSC, neurons discharge a burst of action potentials prior to every contralateral saccade within their movement field (i.e., the range of saccade amplitude and direction that is preceded by a burst). Neurons in the rostral region discharge

maximally prior to small contralateral saccades, whereas neurons in more caudal regions burst before saccades of larger amplitude. Neurons located in the medial border burst prior to upward oblique saccades, and those located on the lateral border burst before downward oblique saccades. Interestingly, the electrical microstimulation of a particular site in the dSC evokes a saccade with amplitude and direction corresponding to those for which the stimulated neurons would maximally discharge. Thus, the amplitude and direction of impending saccades are almost fully determined by the location of active neurons in the dSC, although exceptions exist for some particular saccades.

One fundamental problem is to understand how the locus of collicular activity is converted into the MN firing rate. This transformation was initially considered to be performed by a putative spatiotemporal translator (STT) located in the cerebellum. This STT was viewed as a look-up table mechanism which associated a specific temporal pattern of motor activity with each locus of collicular activity. But, this explanation was not satisfactory because saccades generated toward the same target and from the same starting position can be slower in velocity (longer in duration), yet still have the same amplitude. A solution to this problem was proposed by David A Robinson in 1975 – the duration of saccades is determined by a negative feedback loop. According to this hypothesis, the saccade generators are driven by a motor error signal which results from the comparison between a reference signal, specifying the goal or desired state (e.g., desired eye orientation or desired change in eye orientation), and a feedback signal estimating the current state (current eye orientation or actual change in eye orientation). As long as there is a mismatch between these two internal signals, the motor error drives the saccade generators until the estimated current state matches the desired one. In this framework, the motor error is a dynamic command that changes as the direction of gaze is approaching that of the target. There is no need for an explicit STT; the duration of saccades is simply determined by the time taken to zero-out the motor error command. To date, the negative feedback loop has been the only hypothesis proposed for the spatiotemporal transformation problem. More recently, neuromimetic modeling taught us that the neuronal solutions to the computations involved in saccade generation (e.g., spatiotemporal transformation and component stretching) may not be explicitly performed by separate groups of neurons that neurophysiologists ought to identify. Instead, these computations are implemented implicitly, emerging from the activities of populations of neurons that are massively interconnected and distributed over several neural territories.

The Feedback Control of Saccades

One of the strengths of the negative feedback control is that it endows the oculomotor system with a high level of flexibility for the execution of saccades. If for some reason – pathological (e.g., spinocerebellar degeneration), experimental (functional perturbation by brief electrical microstimulation or by systemic/local injection of pharmacological agents), or natural (drop in vigilance or blink) – the saccade is executed with a velocity which is much slower than normal, its accuracy should be preserved as long as the feedback signal conveys information about the change in velocity.

The feedback signal can originate from several sources. The intrasaccadic visual signals (optic flow and/or target-elicited retinal trace) could participate in estimating the direction of the ongoing saccade. In patients affected by spinocerebellar degeneration, saccades are so dramatically slowed that saccades can even be redirected in flight toward a new target presented during the saccade. But in normal healthy subjects, most saccades are too short in duration for visual signals to influence their ongoing trajectory. Inflow signals originating from sensory receptors located in the orbital tissues or from extraocular muscle proprioception could also be involved in the feedback. But their elimination by sectioning of the ophthalmic branch of the trigeminal nerve does not affect the accuracy and velocity of saccades, nor does it affect the ability to compensate for unexpected changes in eye position prior to saccades toward flashed targets. The estimation of the ongoing oculomotor state could also be based on copies of commands sent to the MNs (efference copy). These outflow signals could originate from collaterals of the burst neurons that immediately drive the MNs. Interestingly, partial pharmacological inactivation of the region involved in the generation of the horizontal component of saccades barely affects the accuracy of the horizontal component of saccades, despite a dramatic reduction in velocity. This result, which demonstrates the feedback control of saccade accuracy also suggests that the number of active neurons influences the velocity of saccades.

The hypothesis of saccade accuracy being controlled by a negative feedback loop also predicts that a defect in estimating the current state should lead to dysmetric saccades. For example, an underestimation of the displacement performed by the eyes from saccade onset is expected to lead to hypermetric saccades (saccades larger in amplitude than required) because the motor error is canceled later than normally (i.e., until the time when the defective estimated displacement matches the desired displacement). Conversely, an overestimation of the actual displacement is expected to lead to

hypometric saccades (i.e., saccades that are smaller in amplitude than required). From this perspective, an optimal control of saccade accuracy requires a calibration of feedback signals such that they accurately represent the actual oculomotor output. A calibration is also required because the neural encoding of feedback signals is not necessarily commensurable with the signals encoding the desired state. Interestingly, one of the major structures involved in the visuomotor calibration of saccades, the medial-posterior cerebellum (composed of the vermal lobules VIc–VII and caudal fastigial nuclei), receives visual motion-related signals from pontine nuclei, proprioceptive signals from extraocular muscles, and eye position/displacement information from the nucleus prepositus hypoglossi/medial vestibular nucleus complex. In return, the output nuclei of the medial-posterior cerebellum, the caudal fastigial nuclei, send distributed projections to the territories housing the premotor neurons of the saccade generators. When the normal function of fastigial oculomotor region (FOR) is impaired, saccades become severely dysmetric.

The Adaptive Control of Saccade Accuracy

Saccade accuracy is also maintained by adaptive mechanisms that assess the error at the end of each saccade and modify the visuomotor transformation to reduce the error of the following saccade. These adaptive mechanisms have been studied in human patients and in monkeys with weakened muscles in one eye. When the subject monocularly views a visual target, saccades are accurate if the normal eye is used but are too small in amplitude when the weakened eye is the viewing eye. If the normal eye is patched and the subject is forced to use the weakened eye to look at visual targets, the amplitude of saccades gradually increases until they are nearly accurate. The fact that the amplitude changes are observed in both the weakened eye and the normal patched eye (which now makes hypermetric saccades) indicates that the adaptive changes affect a command that is issued for conjugate movements of both eyes. This flexibility of the oculomotor system to adjust saccade accuracy has also been studied in experiments in which the target is displaced to a new location during the ongoing saccade. Initially, saccades land near the initial target location, but gradually they land closer to the new one. Despite the plethora of studies characterizing the adaptive properties of the saccadic system, very few studies have tried to decipher the underlying neural mechanisms.

A mismatch between saccade amplitude and target eccentricity seems to also characterize saccades generated by newborns. When orienting toward a peripheral target, infants as young as 1–2 months of age frequently execute a series of consecutive saccades of approximately equal size, whether accompanied by the head or not. Although hypometric, the amplitude of saccades increases with the eccentricity of the target.

The Specification of the Desired State

Saccades are not exclusively aimed at visual targets. They can also be generated in the dark, toward sounds, and to stimuli applied to the skin surface. Localizing such stimuli is a much more complicated task. In the case of a somatosensory stimulus, estimating its location requires the knowledge of the stimulated body region (e.g., the index fingertip of the right hand) and of the posture of the limb on which the stimulus is applied. For auditory stimuli, they are localized on the basis of the interaural differences in intensity and timing of acoustic signals. The dSC seems to be the interface through which these different sensory signals have access to the saccadic motor system. Several saccade-related cells in the dSC are active prior to saccades to visual, somatosensory, or auditory stimuli, suggesting that the converging sensory signals are translated into oculocentric coordinates by the level of the dSC. But the differences in velocity observed among these different groups of saccades suggest that an error signal specifying the desired displacement is not the sole output signal that the dSC sends to the saccade generators. These differences could be due to different levels of firing rate in the population of active neurons or to different population sizes. It is indeed worth mentioning that regular visual saccades also have a reduced velocity and increased duration (accuracy relatively spared) during pharmacological inactivation either of small portions within the dSC or the source of one of its major cortical afferents, the frontal eye fields.

The locus of active neurons in the dSC is not sufficient for determining the amplitude and direction of impending saccades. Some studies show a mismatch between the amplitude and direction of saccades that are actually executed and those represented by the topography of active neurons in the dSC. Indeed, for the same target location, the amplitude and direction are very different between saccades to a permanent target and saccades to a memorized target location. The latter systematically end above the target location. Yet the same population of collicular neurons is involved in both saccade types, suggesting that the change in saccade accuracy is due to different neural activities below the level of the dSC. The same dissociation holds when we compare saccades toward stationary versus moving targets. The latter do not land

on the location where the target was initially detected. Some mechanism tries to compensate for the target displacement that occurred during the period preceding the end of the catch-up saccade. Chemical lesions of middle temporal (MT) cortical area or of its pontine target, the dorsolateral pontine nucleus (DLPN), two structures known to be involved in the processing of visual motion, impair the accuracy of saccades to moving targets without altering that of saccades to stationary targets. Because the DLPN is one of the major sources of motion-related visual signals to the cerebellum, the compensatory mechanism probably involves cerebellar oculomotor regions. Thus, the specification of the saccade goal presumably involves input from the dSC and the cerebellar oculomotor nuclei. Interestingly, for saccades to memorized targets and to moving targets, relationships exist between the accuracy and latency of saccades which may untangle the dynamics of the saccade goal specification.

A dependence of saccade endpoints on the latency has also been documented in studies testing saccades aimed at two stimuli that are simultaneously presented at close locations in the visual field. Under these conditions, the saccade endpoints are scattered along an axis joining the two target locations, characterizing an averaging process. This averaging depends on the relative properties of visual targets (size and energy) – the saccade landing position will be closer to the larger or the brighter target. Moreover, it mostly occurs for short latency saccades and not when the latency exceeds 300 ms. The time window of the averaging process can be delimited by presenting the two targets in a successive manner; short-latency saccades end at the location of the first target, whereas long-latency saccades end at the second target location. The amplitude transition function defines the averaging window that accounts for endpoints that proportionately change with saccade latency.

The Control of Saccade Initiation

Saccades are planned on the basis of information gathered during the time interval that precedes saccade onset. This reaction time is far longer than would be expected by adding the synaptic delays and nerve conduction times between the excitation of sensors and the contraction of extraocular muscles. It reflects the time taken to extract relevant signals from sensors; to make decisions about whether, and how, to respond to the stimulus; and to transmit a goal to the saccadic system. In response to the incoming information about a stimulus, a decision signal is supposed to rise from an initial level until it reaches a threshold level, at which point the saccade is initiated. A rate of rise that varies randomly between trials would explain the wide distribution of saccade latencies.

Several factors influence the latency of saccades. On the retinal side, latency depends, for instance, on the eccentricity of the target, on its energy, and on its contrast relative to the background. Saccade latency also depends on the presence and location of distractors and cues, as well as on more cognitive factors such as the probability of the target location, the amount of reward associated with the target, and the target valence.

Under certain experimental conditions, the distribution of saccade latencies is bimodal. The first mode, which can peak as low as 90 ms, designates the express saccades, whereas the second mode corresponds to regular saccades. The observation of express saccades is facilitated when a temporal gap separates the fixation target offset and second target appearance. Express saccades are suppressed when the deep superior colliculus is lesioned.

Conclusion

Saccadic eye movements are an important component of the animal's behavior for exploring and spatially organizing its sensory environment. They provide a useful tool for understanding how the central nervous system relates the sensorium with the motor repertoire. The neurophysiological approach's aim is to reveal the underlying neuronal organization. It helps us understand why and how the two eyes move together and at the same time and why and how it is possible to look in any direction and orient our gaze exactly to the location where a visual object has appeared. Further empirical research is required to determine the extent to which these basic neural mechanisms participate in the construction of spatial cognition and whether a similar organization underlies the control of other types of goal-directed movements.

See also: Basal Ganglia and Oculomotor Control; Cerebellum and Oculomotor Control; Cortical Control of Eye Movements; Eye and Head Movements; Eye Movement Disorders; Frontal Eye Fields; Oculomotor Control: Anatomical Pathways; Saccade–Pursuit Interactions; Saccades and Visual Search; Superior Colliculus; Supplementary Eye Fields; Target Selection for Pursuit and Saccades.

Further Reading

Barton EJ, Nelson JS, Gandhi NJ, and Sparks DL (2003) Effects of partial lidocaine inactivation of the paramedian pontine reticular formation on saccades of macaques. *Journal of Neurophysiology* 90: 372–386.

Becker W (1989) Saccades. In: Carpenter RHS (ed.) *Eye Movements: Vision and Visual Dysfunction*, pp. 95–137. London: Macmillan.

Becker W and Jürgens R (1990) Human oblique saccades: Quantitative analysis of the relation between horizontal and vertical components. *Vision Research* 30: 893–920.

Carpenter RHS (1988) *Movements of the Eyes*. London: Pion.

Cullen KE and Guitton D (1997) Analysis of primate IBN spike trains using system identification techniques, I: Relationship to eye movement dynamics during head-fixed saccades. *Journal of Neurophysiology* 78: 3259–3282.

Goffart L, Chen LL, and Sparks DL (2004) Deficits in saccades and fixation during muscimol inactivation of the caudal fastigial nucleus in the rhesus monkey. *Journal of Neurophysiology* 92: 3351–3367.

Harris CM, Jacobs M, Shawkat F, and Taylor D (1993) The development of saccadic accuracy in the first seven months. *Clinical Vision Sciences* 8: 85–96.

Hopp JJ and Fuchs AF (2004) The characteristics and neuronal substrate of saccadic eye movement plasticity. *Progress in Neurobiology* 72: 27–53.

Jürgens R, Becker W, and Kornhuber H (1981) Natural and drug-induced variations of velocity and duration of human saccadic eye movements: Evidence for a control of the neural pulse generator by local feedback. *Biological Cybernetics* 39: 87–96.

Keller EL, Gandhi NJ, and Weir PT (1996) Discharge of superior collicular neurons during saccades made to moving targets. *Journal of Neurophysiology* 76: 3573–3577.

Leigh RJ (2006) The saccadic system. In: Leigh RJ and Zee DS (eds.) *The Neurology of Eye Movements*, pp. 108–187. London: Oxford University Press.

Miller JM and Robins D (1992) Extraocular muscle forces in alert monkey. *Vision Research* 32: 1099–1113.

Moschovakis AK, Scudder CA, and Highstein SM (1996) The microscopic anatomy and physiology of the mammalian saccadic system. *Progress in Neurobiology* 50: 133–254.

Optican LM and Quaia C (2002) Distributed model of collicular and cerebellar function during saccades. *Annals of the New York Academy of Sciences* 956: 164–177.

Robinson DA (1973) Models of the saccadic eye movement control system. *Biological Cybernetics* 14: 71–83.

Robinson DA (1975) Oculomotor control signals. In: Lennerstrand G and Bach-Y-Rita P (eds.) *Basic Mechanisms of Ocular Motility and Their Clinical Implications*, pp. 337–378. Oxford: Pergamon Press.

Scudder CA, Kaneko CRS, and Fuchs AF (2002) The brainstem burst generator for saccadic eye movements: A modern synthesis. *Experimental Brain Research* 142: 439–462.

Sparks DL (2002) The brainstem control of saccadic eye movements. *Nature Reviews Neuroscience* 3: 952–964.

Stanford TR and Sparks DL (1994) Systematic errors for saccades to remembered targets: Evidence for a dissociation between saccade metrics and activity in the superior colliculus. *Vision Research* 34: 93–106.

van Gisbergen JAM, Robinson DA, and Gielen S (1981) A quantitative analysis of generation of saccadic eye movements by burst neurons. *Journal of Neurophysiology* 45: 417–442.

Salt Appetite

R S Weisinger and D P Begg, La Trobe University, Bundoora, VIC, Australia
M L Mathai, University of Melbourne, Parkville, VIC, Australia
H S Weisinger, St. Vincent's Hospital Melbourne, Fitzroy, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sodium is vital for the existence of mammals and our appetite for sodium is undeniable. A fine example is that of the elephants and the caves of Mount Elgon, above the jungles of Kenya. Although it is common to see caves in a volcanic mountain, the way in which one of them, Kitum cave, has developed is unusual. For thousands of years, elephants, with little sodium (Na^+) in their natural diet, have supplemented their intake by 'mining' Kitum cave with their tusks. Antelope and buffaloes also supplement their diets by licking Mount Elgon's cave walls. Wild rabbits in salt-poor alpine areas of Australia have been observed to consume salt from pegs impregnated with sodium salt (NaCl) in preference to pegs impregnated with other non-sodium-containing salts or water. The attraction of wild gorillas and chimpanzees to natural salt sources has also been reported.

Sodium is crucial to numerous physiological functions, from maintenance of extracellular fluid volume, blood circulation, and neuronal function to reproductive success. Thus, sodium homeostasis is fundamental to mammalian life. Mechanisms have evolved to minimize sodium loss, and when sodium is lost it must be replaced. Sodium appetite is an innate drive or motivation to seek out and ingest sodium-containing substances.

The mechanisms involved in the control of sodium appetite are complex and rely heavily on receptor-mediated events. Because sodium ions and water are intimately connected within physiological systems, neuroendocrine mechanisms are needed to coordinate information regarding body fluid status and sodium concentration with the appropriate physiological and behavioral responses. The system includes the central nervous system, cardiovascular system, skin, adrenals, kidneys, and gastrointestinal tract. The biological relevance of sodium appetite emerges with much of the planet's land and, hence, vegetation is impoverished of sodium.

Increased sodium appetite and preference for sodium occurs in humans subsequent to salt loss. A demonstration of this is in the development of highly sophisticated salt trades in African desert dwellers and the medical literature abounds with reports of salt craving in conditions in which the function of the adrenal glands has been disrupted. Both congenital adrenal hypoplasia and Addison's disease (autoimmune destruction of the adrenals) lead to rapid demise, due to salt wasting, if untreated.

Most laboratory research has involved the study of rodents, ruminants, and primates in which sodium appetite in sodium deficiency can manifest as intake of an otherwise aversive high-sodium-containing solution.

This article discusses sodium appetite caused by sodium depletion primarily, but it also covers sodium appetite caused by stress and that observed during reproduction (Figure 1).

Sodium Appetite Induced by Depletion of Body Sodium

Depletion of body sodium, as would occur with diarrhea, profuse sweating, or dietary deficiency, results in a number of important physiological changes. Renal mechanisms are activated to minimize loss of body sodium. Activation of the renin–angiotensin–aldosterone system causes increased secretion of the adrenal mineralocorticoid hormone aldosterone that acts on the kidney to maximize reabsorption of sodium. A powerful, specific sodium appetite then develops. This appetite can be caused by a decrease in brain intracellular and cerebrospinal fluid sodium concentration, a decrease in plasma volume, or an increase in blood and brain levels of aldosterone and/or the peptide angiotensin II (Ang II). The appetite is a classic example of an innate neural organization generating a behavioral drive in the instance of a nutritional deficiency. The mechanisms responsible for the generation of this appetite for sodium are complex, and although much is now known, numerous gaps remain.

Factors Involved in Sodium Depletion-Induced Sodium Appetite

Brain Sodium Concentration Evidence to support the hypothesis that changes in the intracellular fluid sodium concentration within cells that subserve sodium appetite contribute to the sodium appetite following sodium depletion was based on experiments using slow intracerebroventricular (ICV) infusion of substances that raise or lower sodium and/or osmolality of the cerebrospinal fluid and surrounding brain nuclei. Sodium intake of sodium-deplete sheep was enhanced by decreasing and reduced by increasing the sodium concentration of the cerebrospinal fluid and

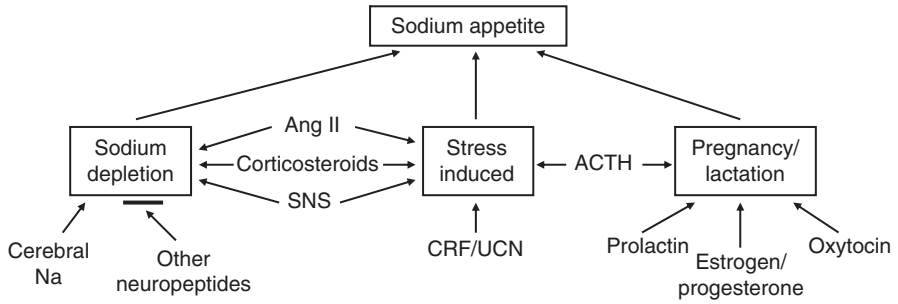

Figure 1 Factors influencing sodium appetite.

brain extracellular fluid, independently of peripheral sodium (i.e., the animal remained sodium deficient). Changes in osmolality did not influence sodium intake. Presumably, the changes in cerebral sodium concentration act on the neural mechanisms involved in the initiation and/or satiation of sodium appetite. Evidence obtained from sheep and cows suggests that cerebral sodium sensors subserving sodium appetite are some distance from the brain's ventricular system and are spatially different from those subserving thirst. A sodium channel (Na_X) thought to be the sodium sensor has been identified in rodents and is preferentially expressed in the circumventricular organs (i.e., circumventricular organs (CVOs), which are brain structures lacking a blood–brain barrier, such as subfornical organ (SFO), organum vasculosum of the lamina terminalis (OVLT), and area postrema (AP)). Increased sodium appetite is observed in mice in which the channel is absent.

Angiotensin II Ang II, an eight-amino-acid peptide, is formed in the peripheral circulation. The renal enzyme, renin, acts on its substrate, angiotensinogen, to form angiotensin I (Ang I), which is then converted to the biologically active Ang II by the action of angiotensin converting enzyme (ACE). Ang II is also formed in the brain, possibly from a precursor such as proangiotensin or preproangiotensin. Several receptor subtypes mediate the activities of Ang II, the best studied of which are type 1 (AT1) and type 2 (AT2). Administration of antagonists of the renin–angiotensin system, peripherally or centrally, has clearly shown that Ang II has a role in sodium appetite induced by sodium depletion. However, controversy exists regarding the relative contribution of Ang II generated in brain and that generated peripherally.

Peripheral administration of ACE inhibitors, such as captopril or enalapril, reduces sodium intake of sodium-deplete rats, sheep, cows, rabbits, mice, and baboons. Furthermore, the sodium appetite of the systemic captopril-treated sodium-deplete animal is restored to the appropriate level by intravenous, but not ICV, administration of Ang II. Importantly, it has been shown that the decrease in blood pressure caused by the administration of the ACE inhibitors is not the cause of the decreased appetite because the appetite is not restored by peripheral administration of phenylephrine at a dose that restores blood pressure. Since Ang II does not readily cross the blood–brain barrier, it is clear that one element in the genesis of the sodium intake of the sodium-deplete animal is mediated by peripherally generated Ang II acting in brain areas without a blood–brain barrier.

In addition to the role of peripheral Ang II, evidence suggests that central angiotensinergic mechanisms (i.e., brain Ang II) are involved in the control of sodium appetite of sodium-deplete animals. For example, administration of ACE inhibitors or Ang II receptor antagonists into the brain decreases the sodium intake of sodium-deplete rats. Whether or not Ang II has a direct influence on neurons subserving sodium appetite has still to be resolved.

Mineralocorticoid hormones The steroid hormones of both the mineralocorticoid (MC) and the glucocorticoid (GC; e.g., cortisol or corticosterone) classes appear to be involved in sodium appetite following sodium depletion, stress, or reproduction. Research has demonstrated that administration of aldosterone or deoxycorticosterone (DOC), another MC hormone, causes sodium appetite in rats, mice, rabbits, hamsters, and sheep. Curiously, the increase in sodium appetite caused by DOC appears greater than that caused by aldosterone, even though aldosterone is a more potent mineralocorticoid receptor (MR) agonist. Due to their lipophilic nature, steroid hormones cross the blood–brain barrier and increased brain levels reflect an increase in peripheral levels. One possible explanation for the difference between aldosterone and DOC in regard sodium appetite, however, is the observation that aldosterone does not penetrate the blood–brain barrier as well as DOC.

In rats, evidence suggests that the amygdala is important in steroid-induced sodium appetite. The amygdala has both MR and Ang II terminals and is the terminus for major gustatory projections from

the nucleus of the solitary tract (NTS) and parabrachial nucleus (PBN). Administration into the amygdala of aldosterone or DOC increases sodium intake in rats. Interference with MR, but not glucocorticoid receptor, formation in the amygdala (by antisense treatment) interferes with the expression of DOC-induced sodium appetite.

However, it is not clear how endogenous aldosterone can stimulate sodium appetite. That is, GCs have an affinity for MRs similar to that of aldosterone, but since the blood concentration of the GCs is 100–1000 times that of aldosterone, MRs are predominantly occupied by GCs. The answer seems to involve 11-β-hydroxysteroid dehydrogenase type 2 (11β-HSD2), an enzyme that confers MR aldosterone specificity by converting active GC to an inactive form (e.g., cortisone). Interestingly, the amygdala does not have 11β-HSD2. Conceivably, aldosterone, a hormone produced in small amounts in the brain, is generated in the amygdala. Work has shown that some neurons in the NTS contain both MRs and 11β-HSD2. These neurons are sensitive to aldosterone; however, they become activated by sodium depletion even in the absence of increased aldosterone secretion. Although research is ongoing, it seems clear that determining a biological basis of sodium appetite will involve a better understanding of these cells.

In addition, sodium intake in rats is increased by 3β,5β-tetrahydroaldosterone (or 5α-tetrahydro-deoxycorticosterone), an A-ring reduced form of aldosterone (or DOC), a steroid that binds to GABAergic membrane receptors but not to intracellular MRs. Thus, mineralocorticoids can act on brain nuclei, particularly in the amygdala, to influence sodium intake through both genomic mechanisms (involving cytosolic MR) and nongenomic mechanisms via $GABA_A$ receptors.

It has been suggested that sodium appetite caused by sodium depletion is mediated by the synergy between aldosterone and Ang II. The synergy hypothesis is tested by concurrent administration of Ang II centrally and DOC or aldosterone systemically at doses that, individually, do not stimulate NaCl intake. Using this model, increased sodium appetite has been reported in rats and pigeons, but it could not be demonstrated in sheep or rabbits. Experiments have demonstrated synergy between Ang II and aldosterone in a primate.

Finally, adrenocorticotrophin hormone (ACTH), a peptide hormone associated with stress and reproduction, causes sodium appetite, which is partially (rabbits) or entirely (sheep and rats) dependent on adrenal hormones (i.e., it is partially or entirely eliminated in adrenalectomized (ADX) animals – animals without intrinsic MCs). However, sodium depletion-induced sodium appetite still occurs in ADX animals. Administration of GC hormones can cause and/or enhance MC-induced sodium appetite.

Central factors

Excitatory mechanisms. As noted previously, Ang II does not readily cross the blood–brain barrier. Thus, the role of circulating Ang II in salt appetite is most likely mediated by its influence on CVOs. In the sodium appetite of sodium-deplete animals, the role of peripheral Ang II acting in the lamina terminalis is suggested by the fact that sodium appetite is decreased or abolished in rats with lesion of the SFO, OVLT, or the anteroventral third ventricle (AV3V; a brain area including the ventral median preoptic nucleus (vMnPO) and OVLT).

Central Ang II-driven mechanisms are involved in neural pathways thought to subserve both physiological sodium appetite and thirst. Ang II and Ang II receptors have been identified in a number of different brain structures, including the hypothalamus (e.g., paraventricular nucleus (PVN) and supraoptic nucleus (SON)), the amygdala and bed nucleus of the stria terminalis (BNST), and the brain areas located on the midline of the front wall of the third brain ventricle (i.e., SFO, MnPO, and OVLT). There is an angiotensinergic pathway between the SFO and the PVN. Lesion experiments have implicated several brain areas with a blood–brain barrier in sodium appetite, such as MnPO, PBN, anterior and lateral hypothalamus, gustatory nucleus of the thalamus, nucleus accumbens, and amygdala.

Mineralocorticoids act on brain nuclei, particularly in the amygdala, to influence sodium intake through both genomic (involving cytosolic receptors) and nongenomic mechanisms. The synergy between the two major hormones of sodium deficiency (Ang II and aldosterone) might occur by (1) interaction between brain sites that are separately sensitive to Ang II (e.g., lamina terminalis) or aldosterone (e.g., amygdala); or, more directly, (2) interaction at the same brain site, for example, by aldosterone upregulating Ang II receptors, increasing Ang II binding to receptors and enhancing responses to receptor activation (e.g., increases in Ang II binding at AT1 receptors occur with administration (at the SFO and AP) of glucocorticoid and mineralocorticoid); or (3) the inhibition of central inhibitory mechanisms.

MRs are located in SFO, MnPO, OVLT, AP, NTS, periventricular brain regions, and some amygdaloid nuclei, but not in PVN or SON. Sodium appetite induced by treatment with DOC is enhanced, rather than depressed, in rats with lesions of the OVLT and is unchanged in rats with lesions of the SFO

or AV3V. Rats with lesions of the medial amygdala have impaired aldosterone- or DOC-induced sodium appetite, and the corticosterone potentiation of aldosterone-induced salt appetite is eliminated. Rats with lesions of the central amygdala or BNST show deficits in sodium appetite induced by DOC.

Inhibitory mechanisms. Brain activity in the OVLT and, to a lesser extent, the SFO is decreased subsequent to sodium intake, whereas sodium intake after sodium depletion has been shown to increase cellular activity (as measured by Fos expression) in the PVN or NTS. These observations suggest that the SFO, OVLT, PVN, and NTS may be involved in the inhibition or satiation processes related to termination of sodium appetite.

A number of neuropeptides have been shown to interfere with sodium appetite subsequent to sodium depletion. In some instances, Ang II, presumably in the neural system subserving thirst, stimulates the release of a neuropeptide that inhibits sodium appetite and Ang II may even inhibit sodium appetite through factors such as increased blood pressure.

Because Ang II is involved in both sodium appetite and thirst (i.e., for water), specificity of intake is achieved by the interaction of angiotensinergic mechanisms with other inhibitory systems. For example, in situations requiring the ingestion of water, stimulation of angiotensinergic thirst mechanisms causes the release of neuropeptides that inhibit or delay sodium intake. In situations requiring the ingestion of sodium, stimulation of angiotensinergic mechanisms involved in sodium appetite could cause the release of neuropeptides that inhibit or delay water intake.

Much evidence suggests that oxytocinergic cells in the PVN are involved in this inhibition of sodium intake. For example, treatments that stimulate sodium intake (e.g., sodium depletion and DOC) decrease OT levels, whereas treatments that decrease sodium intake increase OT levels. Interestingly, central administration of Ang II causes a small increase in sodium intake even though it increases pituitary release of OT. However, sodium intake is greatly enhanced by prior treatment with the central administration of an OT antagonist. Increased sodium intake after central administration of Ang II is also observed with DOC treatment. Administration of DOC also decreases Ang II-induced Fos expression in the PVN and SON.

Evidence suggests that other systems may also be involved in the inhibition of sodium appetite. Decreased serotonin in the raphe nucleus following ingestion of sodium after sodium depletion suggests that serotonergic cells in the raphe nucleus may be involved in inhibition of sodium intake. The administration of fenfluramine or dexfenfluramine, drugs that enhance serotonergic transmission, decreases sodium intake of sodium-depleted rats. Pharmacological blockade of the inhibitory serotonergic system or lesion of the dorsal raphe nucleus increases spontaneous sodium intake as well as the sodium intake of sodium-depleted rats. Injection into the MnPO of L-NAME, a nitric oxide synthesis inhibitor, stimulates sodium intake, suggesting an NO inhibitory system in the MnPO. Somatostatin, atrial natriuretic peptide (ANP), and a number of the tachykinins have also been shown to have an inhibitory effect on sodium appetite. These inhibitory peptides and/or their receptors have been identified in the CVOs as well as other areas implicated in sodium appetite.

Sodium Appetite Induced by Stress and Stress Hormones

The perception of a threat (stress) causes increased release of corticotrophin-releasing factor (CRF)/urocortin (UCN), which in turn causes release of ACTH from the pituitary. ACTH acts on receptors in the adrenal gland to cause the release of adrenocortical hormones. Stress or the hormonal consequences of stress (primarily release of ACTH and the adrenocorticoids) elicits sodium appetite (Figure 2). The

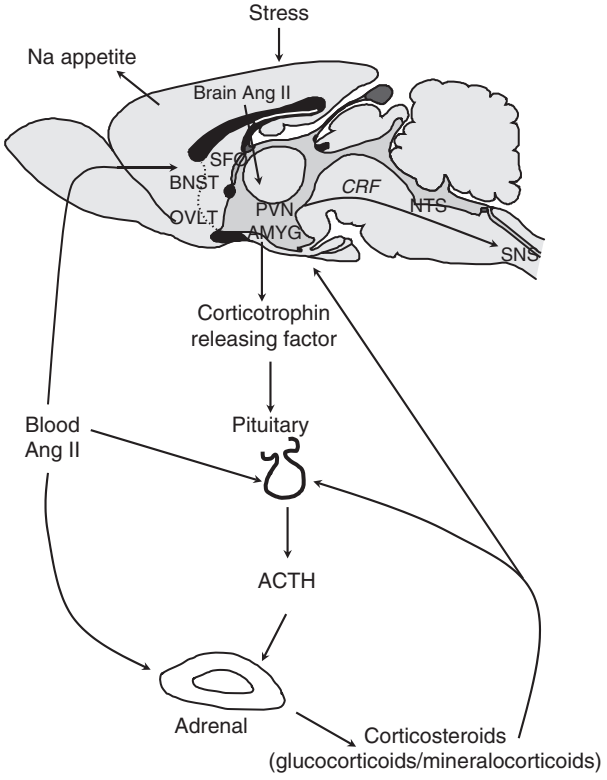

Figure 2 Stress and sodium appetite. Reproduced from Weisinger RS, Blair-West JR, Burns P, Chen N, and Weisinger HS (2004) Neurobiology of sodium appetite. In: Strickter E and Woods S (eds.) *Handbook of Behavioral Neurobiology: Neurobiology of Food and Fluid Intake*, 2nd edn., vol. 14, pp. 547–587, with kind permission of Springer Science and Business Media.

influence of the adrenocorticoids has already been discussed, and the influence of ACTH and CRF/UCN is discussed later.

One of the first demonstrations of stress-induced sodium appetite was made while attempting to fit wild rabbits with jackets designed for carrying small infusion pumps. This restraint of movement caused the rabbits to become agitated and over the course of several days a spectacular increase in sodium appetite was observed. Immobilized or restrained mice also manifest an increase in sodium appetite. 'Intruder' stress causes increased sodium intake in several rat strains, and crowding, a type of imposed social stress, has been shown to cause increased intake of hypertonic NaCl solution. Clonidine or reserpine (centrally acting sympathetic antagonists) blocks the increase in sodium intake, suggesting that the increase in intake is due to stimulation of the sympathetic nervous system. The mechanism by which stimulation of the sympathetic nervous system stimulates sodium appetite is not known.

Adrenocorticotrophic Hormone

Peripheral administration of ACTH causes a large increase in sodium intake in several species, including sheep, rabbits, rats, and mice. Sodium appetite following administration of ACTH can be eliminated by adrenalectomy and, therefore, probably involves the activity of adrenal hormones. In all of the animals tested with a cafeteria of hypertonic solutions of NaCl, KCl, $CaCl_2$, and $MgCl_2$, the appetite generated by ACTH administration is predominantly that of NaCl solution. In intact sheep, a large increase in sodium intake occurs after 2 or 3 days of ACTH administration. In ADX sheep, administration of a cocktail of hormones that are secreted from the adrenal gland during stimulation with ACTH (i.e., aldosterone, corticosterone, cortisol, 11-deoxycortisol, and deoxycorticosterone) causes an increase in sodium appetite which is similar to that induced by ACTH in intact animals. In intact rabbits, concurrent administration of cortisol and corticosterone, in doses that increase plasma levels to those observed in ACTH-treated animals, increases appetite by approximately half as much as that caused by ACTH.

ACTH-induced sodium appetite has been clearly demonstrated in ADX rabbits, maintained in good health with combined MC–GC treatment. The MC treatment was such that urinary sodium loss was minimized, whereas the GC treatment was such that blood ACTH was maintained at a basal level; blood glucocorticoids have a negative feedback control of ACTH secretion. When GC treatment was reduced such that blood ACTH was no longer maintained at basal level, sodium intake was increased. This evidence is consistent with the ability of ACTH to stimulate sodium appetite independent of steroid hormones and not due to sodium loss.

In rats, sodium intake (and output) is approximately equal to the rat's total body sodium by the fifth day of ACTH administration. ACTH also increased the sodium intake of both the SHR and WKY rats. In SHR rats, the increase was in addition to their already high-sodium intake. ACTH did not affect sodium appetite of baboons or pigs.

Corticotrophin-Releasing Factor/Urocortin

CRF, a 41-amino acid peptide, is expressed in neurons or terminals throughout the limbic system (e.g., hypothalamus and amygdala) as well as in systems associated with autonomic function (e.g., PBN and locus coeruleus). CRF is involved in the initiation of the behavioral and physiological responses to stress. UCN, also a 41-amino acid peptide, has a 45% homology with CRF and binds with high affinity to CRF receptors. UCN is expressed in many of the same brain areas as CRF, including those involved in body fluid and electrolyte homeostasis, and may contribute to responses previously attributed to CRF.

Administration of either CRF or UCN would be expected to increase sodium intake due to stimulation of sympathetic nervous system activity and ACTH secretion. Indeed, ICV administration of CRF increases sodium intake in rabbits while also causing a sustained increase in both cortisol and corticosterone. In experiments in sheep, however, prolonged ICV or intravenous administration of CRF or UCN failed to stimulate sodium appetite, despite increased secretion of ACTH. Furthermore, prolonged ICV infusions of CRF and UCN decreased need-free salt (and food) intake in sheep, mice, and baboons. The decrease in sodium intake in baboons was transient, but the decrease in food intake was sustained throughout the infusion period. Hence, there appear to be interspecies differences and the mechanisms responsible for the inhibition of sodium appetite are unclear.

Presumably, the inhibition of sodium appetite is caused by the influence of CRF on systems other than those influencing ACTH secretion or sympathetic nervous activity, such as oxytocin secretion. Clearly, the inhibitory actions of CRF and UCN need further exploration.

Angiotensin II

The increased sodium intake of restrained mice is prevented by treatment with captopril. This finding suggests that the appetite involves the renin–angiotensin system, which is known to be elevated under stress. It has been suggested that both peripherally generated and centrally generated Ang II can stimulate the

secretion of ACTH, the latter via its influence on CRF release from the PVN. Increased sodium intake in mice observed during cold stress has also been attributed to increased activity of the renin–angiotensin system. In mice, the increase in sodium intake caused by ACTH is not altered by treatment with captopril, and it appears that stress-induced Ang II precedes the increase in ACTH because stress-induced, but not ACTH-induced, sodium intake is blocked by captopril.

Brain Structures Activated by Stress or Stress Hormones

In crowded rats, Fos is increased in a number of brain areas, including the SFO, PVN, and amygdala. Increased Fos has also been reported in a number of brain areas (e.g., PVN, SON, locus coeruleus, BNST, midbrain raphe, NTS, amygdala, cortex, lateral septum, lateral preoptic area, anterior hypothalamus, LH, Barrington's nucleus, BNST, and striatum) in rats subjected to stress.

Sodium Appetite during Pregnancy and Lactation

Sodium intake is essential for reproductive success. In both rats and humans, plasma [Na^+] is decreased during pregnancy due to the actions of the hormone relaxin on the osmotic threshold for thirst and secretion of antidiuretic hormone. Despite the decreased plasma sodium concentration, retention of sodium is increased during pregnancy, with more than 55% of the sodium found in the products of conception. The reproductive process is a powerful stimulus of specific sodium appetite in several species. Increased sodium intake occurs in pregnancy and is then further augmented during lactation. ACTH seems to be part of the hormone complement necessary to induce this appetite. Treating rabbits initially with estradiol and progesterone and then with prolactin, ACTH, and oxytocin causes an increase in sodium intake similar to that seen during lactation. Thus, the behavior is primarily hormone induced, as distinct from it being determined by sodium deficiency caused by sequestration of sodium in the products of conception or loss of sodium in milk.

Ang II does not appear to have a role in the increase in water and sodium intake observed in pregnant rats because peripheral administration of losartan from day 2 to day 19 of gestation does not alter sodium intake.

Conclusions

The control of sodium appetite is as complex as it is vital to survival, involving neuroendocrine, behavioral, and multiple organ integration. As animals evolved, the challenges faced by each species varied, as did the mechanisms adapted to deal with those challenges. At the same time, presumably, the challenges imposed by sodium deficiency, stress, pregnancy and lactation, and the subsequent need to ingest sodium were integrated into the mechanisms required for survival in general. Thus, although the need for sodium and the ingestion of sodium may be common to most, if not all, terrestrial animals, it appears that in some instances different mechanisms have evolved.

Ang II has an important role in the control of sodium appetite. The relative contributions of peripheral and central Ang II to sodium depletion-induced sodium appetite are not completely understood. As sodium depletion increases, both peripheral and, presumably, central levels of Ang II increase. Peripheral Ang II, acting on AT1 receptors in CVOs, and brain Ang II, acting at AT1 receptors in brain structures with a blood–brain barrier such as the MnPO and amygdala, influence sodium appetite.

Increased blood levels of Ang II, caused by sodium depletion, stimulate the secretion of aldosterone. When the formation or action of Ang II is blocked, sodium appetite is decreased. Sodium appetite may be due to the increase in Ang II formation and concentration in the brain, possibly in synergy with the elevated blood levels of aldosterone. In addition, decreased cerebral sodium concentration may increase sodium intake by acting on sodium sensors or by increasing aldosterone secretion. Ang II, as a neurotransmitter, may have a role in these effects as well. Adrenocorticoid hormones may also contribute to sodium appetite by enhancing the central actions of Ang II or by acting directly on receptors located in the amygdala, BNST, and special cells in the NTS to cause sodium appetite. However, a role for adrenocortical hormones acting on receptors located in the lamina terminalis cannot be ruled out. Elucidation of these pathways and mechanisms will be challenging.

Inhibitory factors are important to the control of sodium appetite. Oxytocin and somatostatin are neuropeptides with some specificity with regard to body fluid homeostatic mechanisms (i.e., they inhibit sodium but not water intake). ANP, on the other hand, appears to have a universal inhibitory role in both sodium appetite and thirst. Brain serotonergic mechanisms appear to have an inhibitory input to sodium appetite.

Several peptides have been shown to influence sodium appetite, although their precise physiological role remains undefined. Although extensively explored, the pharmacoanatomical organization of the neural systems subserving sodium appetite is unclear. Regarding arousal of appetite, striking species differences have been demonstrated and are believed to reflect differing evolutionary selection pressures contingent on different

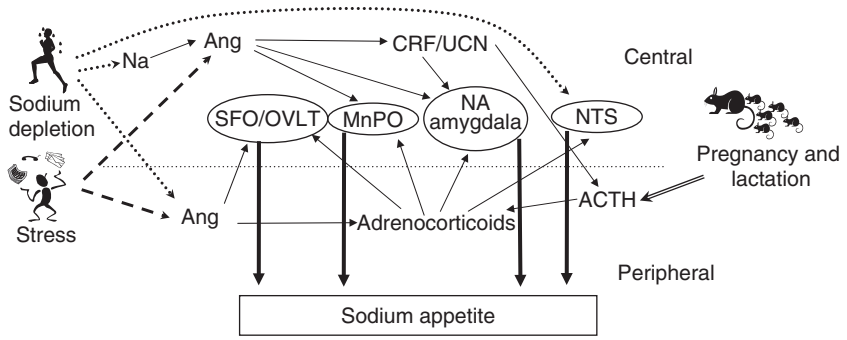

Figure 3 Central and peripheral factors involved in sodium appetite. Reproduced from Weisinger RS, Blair-West JR, Burns P, Chen N, and Weisinger HS (2004) Neurobiology of sodium appetite. In: Strickter E and Woods S (eds.) *Handbook of Behavioral Neurobiology: Neurobiology of Food and Fluid Intake*, 2nd edn., vol. 14, pp. 547–587, with kind permission of Springer Science and Business Media.

feeding habits. It is also clear that different stimuli arouse appetite in different situations, with complex neurochemical organization within the brain underlying such behavior.

Stress and hormones associated with the stress response (e.g., ACTH and adrenocortical hormones) can cause sodium appetite. Although there is evidence in the rabbit for a direct effect of ACTH on sodium appetite, in rats and sheep the increased intake caused by ACTH appears to be mediated by its influence of the secretion of adrenocortical hormones. An increase in Ang II or activity of the sympathetic nervous system caused by stress also may stimulate sodium intake. The failure of CRF/UCN to stimulate sodium appetite is paradoxical and the mechanisms responsible are still to be discovered.

Increased sodium intake occurs during pregnancy and lactation. In rats, like humans, during pregnancy the plasma volume is maintained above normal level and plasma sodium concentration is maintained below normal level. Increased intake of sodium, which is required for the products of conception and reproductive success, is observed. The increase in sodium intake has been studied in the rabbit and appears to depend on the hormonal consequences of pregnancy and lactation, including ACTH. Thus, the sodium appetite could be generated, at least in part, by adrenocortical hormones utilizing mechanisms thought to contribute to the sodium appetite observed during stress or sodium depletion. Thus, notwithstanding the exceptions noted previously, in the main similar brain mechanisms (Figure 3) may be involved in the control of sodium appetite observed with sodium depletion, stress, pregnancy, and lactation.

See also: Angiotensin II; Atrial Natriuretic Peptide: Fluid/Mineral Balance; Blood Pressure: Baroreceptors; Circumventricular Organs; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Fluid and Electrolyte Homeostasis: Clinical Disease; Neurohypophyseal System; Osmoregulation; Thirst.

Further Reading

Aumann GD and Emlen JT (1965) Relation of population density to sodium availability and sodium selection by microtine rodents. *Nature* 208(6): 198–199.

Blair-West JR, Carey KD, Denton DA, Weisinger RS, and Shade RE (1998) Evidence that brain angiotensin II is involved in both thirst and sodium appetite in baboons. *American Journal of Physiology* 275(5 Pt. 2): R1639–R1646.

Castren E and Saavedra JM (1988) Repeated stress increases the density of angiotensin II binding sites in rat paraventricular nucleus and subfornical organ. *Endocrinology* 122(1): 370–372.

Charron G, Laforest S, Gagnon C, Drolet G, and Mouginot D (2002) Acute sodium deficit triggers plasticity of the brain angiotensin type 1 receptors. *FASEB Journal* 16(6): 610–612.

Coghlan JP, Considine PJ, Denton DA, et al. (1981) Sodium appetite in sheep induced by cerebral ventricular infusion of angiotensin: Comparison with sodium deficiency. *Science* 214(4517): 195–197.

Geerling JC, Engeland WC, Kawata M, and Loewy AD (2006) Aldosterone target neurons in the nucleus tractus solitarius drive sodium appetite. *Journal of Neuroscience* 26 (2): 411–417.

Hiyama TY, Watanabe E, Ono K, et al. (2002) Na(x) channel involved in CNS sodium-level sensing. *Nature Neuroscience* 5(6): 511–512.

Johnson AK and Thunhorst RL (1997) The neuroendocrinology of thirst and salt appetite: Visceral sensory signals and mechanisms of central integration. *Frontiers in Neuroendocrinology* 18(3): 292–353.

Lucas LR, Pompei P, and McEwen BS (2000) Salt appetite in salt-replete rats: Involvement of mesolimbic structures in deoxycorticosterone-induced salt craving behavior. *Neuroendocrinology* 71(6): 386–395.

Schreihofer AM, Anderson BK, Schiltz JC, Xu L, Sved AF, and Stricker EM (1999) Thirst and salt appetite elicited by hypovolemia in rats with chronic lesions of the nucleus of the solitary tract. *American Journal of Physiology* 276(1 Pt. 2): R251–R258.

Stricker EM and Verbalis JG (1996) Central inhibition of salt appetite by oxytocin in rats. *Regulatory Peptides* 66: 83–85.

Thunhorst RL, Beltz TG, and Johnson AK (1999) Effects of subfornical organ lesions on acutely induced thirst and salt appetite. *American Journal of Physiology* 277(1 Pt. 2): R56–R65.

Vivas L, Pastuskovas CV, and Tonelli L (1995) Sodium depletion induces Fos immunoreactivity in circumventricular organs of the lamina terminalis. *Brain Research* 679(1): 34–41.

Weisinger RS, Blair-West JR, Burns P, Chen N, and Weisinger HS (2004) Neurobiology of sodium appetite. In: Strickter E and Woods S (eds.) *Handbook of Behavioral Neurobiology: Neurobiology of Food and Fluid Intake*, 2nd edn., vol. 14, pp. 547–587.

Weisinger RS, Burns P, Colvill LM, et al. (2000) Fos immunoreactivity in the lamina terminalis of adrenalectomized rats and effects of angiotensin II type 1 receptor blockade or deoxycorticosterone. *Neuroscience* 98(1): 167–180.

Weisinger RS, Considine P, Denton DA, and McKinley MJ (1979) Rapid effect of change in cerebrospinal fluid sodium concentration on salt appetite. *Nature* 280(5722): 490–491.

Schizophrenia: Epidemiology, Clinical Features, Course and Outcome

W G Frankle, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Schizophrenia represents a brain disorder (or cluster of disorders) characterized by hallucinations, delusions, impairments in cognitive functioning, and blunted emotional relatedness. It ranks among the top diseases as a cause of disability in industrialized nations. With its onset in adolescence or early adulthood and the lack of fully effective treatments, it contributes greatly to the global burden of disease. In the United States alone, the direct and indirect costs of schizophrenia were estimated at $32.5 billion in 1990. In addition, it is associated with increased rates of general medical illnesses, including heart disease and diabetes, as well as frequent comorbid psychiatric illnesses such as substance abuse. Although significant advances have been made in the treatment of this disorder during the past 50 years, much remains unknown regarding the transmission, clinical course, and optimum therapy for this illness.

Epidemiology

Studies exploring the epidemiology of schizophrenia have been performed since the work of Koller in 1895 highlighted the heritability of this disorder. However, in the intervening century, the determination of epidemiologic factors relating to schizophrenia, such as incidence, population prevalence, and risk factors, has been disadvantaged by methodological problems not encountered in nonpsychiatric complex disease processes. First and foremost has been the issue of case identification. To date, as with all psychiatric illnesses, there are no biological tests to determine the disease state, and therefore clinical interviews, clinical observations, and self-reports form the basis for diagnosis. Moreover, shifting definitions have been utilized as diagnostic criteria for schizophrenia, confounding comparisons across studies and over time. With the development of standardized criteria for diagnosis in the 1980s, the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III) and the tenth revision of the *International Statistical Classification of Diseases and Related Health Problems* allowed for structured interviews and agreed-on criteria for case identification, providing improved reliability in the diagnosis.

Prevalence and Incidence

The prevalence of a disease represents the morbidity of the disease in the population and is a combination of the incidence and the length of the illness. Although certain small populations exist in which the prevalence of schizophrenia has been found to be significantly increased or decreased compared with larger samples, the point prevalence rates across populations are, in general, similar, in the range of four per thousand, with lifetime morbidity rates in the range of seven to eight per thousand.

The incidence rate of a disease represents the risk of that disease's occurring at a point in time in a particular population. As the onset of schizophrenia is a relatively rare event, studies examining the incidence rate of schizophrenia generally use the point of first contact with mental health workers as the time of onset of the illness. This potentially represents a confound since the actual onset of symptoms frequently precedes contact with the mental health system by several years, a problem which some studies address by defining the time of onset retrospectively, although this is not ideal either. Incidence rates of schizophrenia across studies range from 0.07 to 7.1 per thousand per year. The WHO Determinants of Outcome study, published in 1992, was a ten-country study of schizophrenia which provided insights into the variability of this disease across cultures. Using a restrictive definition of schizophrenia, the WHO study found that there were no differences across countries in the annual rates of schizophrenia, with rates averaging around 0.10 per thousand annually. Given that the WHO incidence rates were similar across cultures, it would seem that factors such as climate and economic development of a country play little or no role in the risk for schizophrenia.

Moreover, there appears to be no difference between men and women in lifetime risk for schizophrenia. The lifetime risk for diagnosis of schizophrenia was determined to be approximately 0.13 per thousand in a study of first-admission cases by Häfner and colleagues. However, gender differences are apparent when comparing the age of onset of the illness. The age range of maximum risk for schizophrenia extends from 15 to 30 years old; for men the onset tends to occur about 4–5 years earlier than for women. As noted previously, first contact with mental health services is frequently delayed from the onset of the illness

and occurs between 25 and 35 years of age for both men and women. In spite of the age difference in onset between men and women, the core features of the illness are similar in both genders.

There are several other findings relating to the epidemiology of schizophrenia that remain interesting but are of uncertain significance. First is the finding of increased rates of schizophrenia in urban centers. At times this has been attributed to a 'drift' phenomenon; in other terms, individuals with schizophrenia may tend to migrate to urban areas. However, although Marcelis and colleagues in 1998 found a significant, positive correlation between the size of urban centers and the incidence of schizophrenia, they were not able to detect a difference between urban birth and urban residence. Another interesting epidemiological factor that remains unexplained is the highly replicated finding relating to season of birth. In the northern hemisphere, there is a 5–8% increase in schizophrenia births in the winter and spring compared with the other seasons. Finally, reports have repeatedly shown an increased prevalence of schizophrenia in immigrants compared with native-born individuals; however, methodological issues such as access to care, diagnostic difficulties due to language, and cultural issues confound the interpretation of these reports.

Risk Factors

Family history of the illness represents the single greatest risk factor for developing schizophrenia and was observed in the very first studies of this disease by individuals such as Koller and Kraepelin in the early twentieth century. The lifetime risk of developing schizophrenia for an individual with a first-degree relative who had the illness is 1–16%, compared with less than 1% in the general population. Individuals with a schizophrenic parent have a 15-fold increased risk of developing the illness; for siblings of an individual with schizophrenia, the risk of developing the illness is about tenfold greater than the general population. Second-degree relatives (nephews, nieces, grandchildren, and half-siblings) have approximate 3 times the rate of schizophrenia in the general population. The concordance rate for monozygotic twins ranges from 35% to 69%, depending on the study, while that for dizygotic twins is similar to that of siblings (0–26%). Adoption studies show that the risk of an adopted child's developing schizophrenia depends on the presence of schizophrenia in a biological parent rather than in the adoptive environment. Taken together, these findings indicate that genetic predisposition to schizophrenia is a risk for, but is not sufficient for, developing the illness and that environmental factors clearly play a role.

A variety of environmental factors have been examined as being potentially related to the development of schizophrenia; however, to date none have been demonstrated to be definitively causal. Among these factors are difficulties during pregnancy or birth, perinatal brain damage, and various early infections. Several studies have shown that obstetrical complications are associated with increased risk for developing schizophrenia, on the order of 3 times the risk of those without obstetrical complications. In one study, the most significant risk was represented by perinatal brain damage, which carried a sevenfold increase in risk of schizophrenia. Intrauterine exposure to influenza or central nervous system infections has also been associated with increased rates of schizophrenia, although more-recent work has not been able to replicate the findings related to maternal influenza infection.

Additional factors associated with the subsequent development of schizophrenia include poor premorbid adjustment, with few friends, isolative behavior, and lack of significant romantic relationships. However, as these patterns can be seen frequently in the at-risk age group of adolescence, their usefulness in predicting the development of schizophrenia is limited; furthermore, it is not clear whether these factors represent part of the illness prodrome or the risk factors.

Recent literature has implicated premorbid cannabis use as a risk factor in the development of schizophrenia. It has been evident for some time that the rates of cannabis use among individuals with schizophrenia are higher than those among the general population, but now, as reviewed by Arseneault and colleagues in 2004, a relatively strong association (odds ratio > 2) between premorbid cannabis use and schizophrenia has been observed, particularly when cannabis use occurs in adolescence. These data have led to an increased interest in the link between cannabis use or abuse and the pathophysiology and/or etiology of schizophrenia. Still other risk factors linked to the development of schizophrenia include low premorbid IQ, increasing paternal age, and epilepsy; however, no single risk factor has been shown to be causal in the development of schizophrenia. In conclusion, the transmission of schizophrenia is complex and likely involves multiple interacting genes which lead to a predisposition to the illness and that may combine with as yet incompletely identified contributing environmental factors.

Clinical Features

The current nosology of schizophrenia can be traced to the German psychiatrist Emil Kraepelin. Relying on his research, Kraepelin promoted the observation

that there was no specific clinical symptom or symptoms unique to the major mental illnesses but rather a specific pattern of symptoms and course which differentiated these disorders. Using this observation, he began to develop a classification system for mental illnesses which not only incorporated the clinical symptoms but also relied on criteria of age of onset, course, outcome, and prognosis to separate these disorders. It was with the publication of the sixth edition of his textbook in 1899 that Kraepelin separated the entity of dementia praecox from manic depressive illness. The term 'dementia praecox' (as 'demencé precocé') was first used by Benedict Augustin Morel in 1852 to describe the onset of dementia in young patients. Kraepelin defined dementia praecox as "subacute development of a peculiar simple condition of mental weakness occurring at a youthful age" and separated this entity into three forms – hebephrenic, catatonic, and paranoid (expanding on prior work by Hecker and Kahlbaum). The true advance by Kraepelin in this area came with his emphasis on etiology and outcome of the illness. He noted hereditary and premorbid personality as contributing to the appearance of dementia praecox, as well as suggesting that the presence of external factors, such as premature birth and/or obstetrical complications, might increase the likelihood of developing this illness. Most notably, he highlighted the underlying biological nature of the illness and the progressive, relapsing nature of dementia praecox.

The term 'schizophrenia' was introduced by the Swiss psychiatrist Eugene Bleuler (1908), who defined it as a group of psychoses characterized by alterations in thinking, feeling, and relatedness to the outside world. Bleuler noted four primary or core features of the illness, (1) looseness of associations, (2) affective flattening, (3) autism, and (4) ambivalence, which became known as 'the four A's' and were widely utilized as diagnostic criteria for schizophrenia. Bleuler considered these symptoms essential for the diagnosis, with the presence of additional, accessory features, including hallucinations and delusions, being secondary to the diagnosis. The four A's continued to be used in the diagnosis of schizophrenia until the work of Kurt Schneider, who in 1957 described first-rank symptoms of schizophrenia. These included audible thoughts; voices arguing, discussing, or commenting; somatic passivity experiences; thought withdrawal; thought broadcasting; and delusional perceptions.

Schneider's criteria shaped the current diagnostic definition of schizophrenia and have been largely incorporated into the DSM-IV, which uses a descriptive definition of the illness. The DSM-IV criteria define what have come to be known as positive symptoms and negative symptoms of the illness. Positive symptoms represent the psychotic features of the illness and include hallucinations, delusions, disorganized speech, and disorganized behavior. The negative symptoms of the illness are the absence of normal behaviors and include affective flattening, alogia, apathy, and avolition. In addition to clinical symptoms, the DSM-IV diagnostic criteria for schizophrenia requires impairment in function of one or more major areas, such as work, interpersonal relations, or self-care, defined as markedly below the level achieved prior to the onset of the illness. Continuous signs of the disturbance must persist for at least 6 months in order to meet the DSM-IV level of schizophrenia.

Although they are not specifically listed as a diagnostic criterion for schizophrenia, it has become evident that this disorder is associated with cognitive impairments. Of note, recent studies have demonstrated that it is the degree of cognitive impairment which is the best predictor of long-term outcome in individuals with schizophrenia. In general, the cognitive impairments seen in schizophrenia are most marked in the areas of memory and executive function, although mild-to-moderate impairment can be seen across all neurocognitive domains. The cognitive impairments tend to persist throughout the course of the illness, even during periods when the positive and negative symptoms have waned. The degree of impairment also appears to be relatively consistent early on in the illness; however, some individuals may experience a greater cognitive decline in later life. The striking lack of treatments available to address these cognitive deficits, combined with the association of these symptoms with poorer functional outcomes, has lead to the recent collaboration between academia, the National Institutes of Health, and the pharmaceutical industry on the Measurement and Treatment Research to Improve Cognition in Schizophrenia initiative to develop treatments for these symptoms.

Course and Outcome

It is now clear that, as noted by Kraepelin, the onset of the diagnostic symptoms of schizophrenia is frequently preceded by changes in functioning and behavior, termed 'prodromal symptoms.' These may include positive symptoms that do not rise to the DSM-IV level in severity or in duration, mood symptoms, cognitive impairments, or social withdrawal. Although useful, these prodromal symptoms exhibit significant overlap with the range of normal experience in the at-risk age groups, and so they cannot be used as diagnostic. Typically, the appearance of frank psychotic symptoms marks the onset of schizophrenia;

however, formal diagnosis and subsequent treatment are frequently delayed, on the average of approximately 1 year from the onset of psychosis. The first signs of the illness occur before the age of 30 in approximately 77% of individuals; nearly half the time, symptoms can be identified as having been present prior to the age of 20. Notwithstanding this fact, the majority of individuals diagnosed with schizophrenia achieve symptomatic recovery, with estimates as high as 75–85% achieving remission with treatment at 1–2 years in uncontrolled studies. Although definitions of remission vary, most studies have measured significant improvement in positive symptoms, with less improvement in negative symptoms, with treatment. Several factors have been identified as predicting elevated positive symptoms after treatment, including a longer duration of untreated psychosis, a greater degree and longer period of prodromal symptoms, and poor premorbid adjustment. Similar factors have been noted to predict higher levels of negative symptoms after 1 year of treatment.

Although most individuals experiencing a first episode of schizophrenia will achieve symptomatic recovery, nearly all will go on to experience at least one, and likely multiple, subsequent episodes. Relapse rates after the first episode range from 30% in the first year to 80% within 5 years. With each subsequent relapse, a significant proportion of individuals will not recover to the same degree as after the first episode of the illness and will experience persistent symptoms and/or cognitive impairments. This chronic, relapsing course of schizophrenia results in periods of relative remissions, rarely reaching the level of premorbid function, punctuated by acute exacerbation of the symptoms. This pattern of illness progression was noted by Kraepelin and appears to differing degrees across individuals diagnosed with schizophrenia, with more striking deterioration occurring in the early years of the illness and a relative stabilization of functioning in the later years. A number of large-scale studies have examined the long-term outcome in schizophrenia. These include studies such as the Iowa 500 Follow-Up Study, which, with a retrospective design and an average 35-year follow-up, examined 500 patients admitted to Iowa State Psychiatric Hospital between 1934 and 1944. Compared with patients with other diagnoses, those with schizophrenia had poorer long-term outcomes; however, approximately 20% of participants were symptom free at follow-up. Another example of this type of study is the prospectively designed Chicago Follow-Up Study, which has followed participants for up to 20 years. This study has also found that patients with schizophrenia had poorer outcomes than those with other psychiatric illnesses but that about 40% of those with schizophrenia demonstrated periods of recovery which could last several years. Comparisons between these and other studies of this nature are hampered by measurements used to assess outcome as well as the variability in the diagnostic criteria. Despite these factors, however, several points are clear. First, in all studies, a subset of patients with schizophrenia, in the range of 20–57%, demonstrate a good outcome, depending on the strictness of the criteria used in the diagnosis of schizophrenia (inclusion of individuals with some affective symptoms increases favorable outcome rates) and the definition of good outcome. Second, a small number, on the order of 6–8% of participants, met criteria for recovered at follow-up, meaning an absence of any symptoms. Finally, approximately 30% of patients with schizophrenia remain impaired, requiring long-term continuous treatment and experiencing few symptom-free intervals.

In sum, the clinical course of schizophrenia is highly variable, with the majority of individuals experiencing a relapsing course, which may include sustained periods of remission. In a smaller, but significant, number of individuals, the course of schizophrenia may be more chronic, with few, if any, symptom-free periods. In any discussion of long-term outcome, it is important to highlight the fact that schizophrenia is associated with a significant degree of mortality due to suicide. Suicide is the leading cause of premature death among individuals with schizophrenia, whose suicide risk is 8.5-fold that of the general population. Approximately 20–50% of individuals with schizophrenia will attempt suicide, and 4–13% have completed suicide.

Cross-Cultural Outcomes

Studies which have followed up the WHO Determinants of Outcome ten-country study samples have revealed some interesting differences in outcomes between developing and developed nations. A 2-year outcome study of this population found that in developing nations (Colombia, India, and Nigeria), 62.7% of individuals were remitted or in complete remission, with 35.7% experiencing continuous or episodic symptoms. Exactly the reverse was found in developed countries (Czech Republic, Denmark, Ireland, Japan, Russia, the United Kingdom, and the United States), with 36.8% remitted or in complete remission and 60.9% experiencing continuous or episodic symptoms. Several independent studies have also reported increased percentages of individuals with better outcomes in developing compared with developed countries. The reason for this difference remains unknown; possible explanations include differences in genetic variations, difference in social support

and social networks, and other environmental factors. Pharmacological management of the disease was used in only a small proportion of individuals in developing nations and therefore did not contribute to the improved outcome in this setting.

Treatment

Since the discovery of the antimanic and antipsychotic effects of chlorpromazine in the early 1950s, the mainstay of treatment for schizophrenia has been pharmacotherapy with neuroleptic medications (so termed by early researchers Jean Delay and Pierre Deniker for the properties of slowing motor activity and causing affective indifference and emotional neutrality). Earlier, therapies of questionable efficacy were employed, including application of tuberculin, insulin coma, and pharmacological agents such as opium, morphine, cocaine, or bromide. Since the introduction of chlorpromazine, numerous antipsychotic/neuroleptic medications have been introduced (on the order of 30 distinct compounds). The introduction of chlorpromazine was followed by haloperidol. Originally developed as an analgesic by Paul Janssen in Belgium but quickly found to have neuroleptic properties, haloperidol became one of the most widely prescribed medications after its introduction in 1959 (1969 in the United States). The work with chlorpromazine and haloperidol led to the development of multiple drugs screened to have properties *in vitro* similar to those of chlorpromazine and haloperidol. These medications have been termed the 'first-generation' antipsychotics, with about 15 different drugs being introduced in the United States from 1959 until 1975. Early preclinical research, along with subsequent brain imaging studies in humans, revealed a link between blockade of the dopamine D2 receptors and the ability of these medications to treat psychosis.

In 1975, clozapine, a drug with relatively low affinity for D2 receptors but relatively high affinity to serotonin 2a receptors (5-HT2a), was introduced in Europe but withdrawn shortly thereafter due to the risk of agranulocytosis associated with treatment. This medication was reintroduced in 1990 with hematological monitoring requirements, allowing for safer administration of the drug. Clinical studies have repeatedly demonstrated a superior efficacy for clozapine compared with the first-generation drugs, particularly in treatment of refractory patients. Moreover, clozapine is associated with fewer extrapyramidal side effects (EPS), which are prominent with the first-generation drugs due to blockade of the D2 receptors in the striatal brain regions. These two facts, increased efficacy with decreased EPS, resulted in clozapine being termed a 'second-generation' or 'atypical' antipsychotic and led to the development of additional medications (risperidone, olanzapine, quetiapine, and ziprasidone) with high 5-HT2a affinity relative to D2 affinity. At standard clinical doses, these medications appear to have lower rates of EPS than the first-generation drugs did, and they have also been classified as second generation or atypical antipsychotics. Although much work has been done to examine the efficacy of these second-generation drugs, with the exception of clozapine they demonstrate no clear superiority when compared with the first-generation medications. Moreover, despite the lower rates of EPS, additional side effects, most notably increased rates of impaired glucose metabolism and diabetes, have been observed with the second-generation antipsychotic medications.

While all currently available antipsychotic treatments have been determined superior to placebo in the treatment of the positive symptoms of schizophrenia, their effect on negative symptoms and the cognitive deficits can be described as minimal at best, with clozapine demonstrating the largest effect. This deficiency in adequate treatments for the negative and cognitive symptoms seen in this illness is a major shortcoming of currently available therapeutics for schizophrenia as the degree of cognitive impairment has been shown to be a predictor of poor functional outcome.

Although pharmacotherapy represents the main first-line treatment for individuals with schizophrenia, evidence has demonstrated that the best long-term outcomes are achieved in this illness using a multimodality approach. In support of this is the fact that many individuals with schizophrenia continue to have impairments in social, occupational, and cognitive function even after adequate treatment of the psychotic symptoms. In addition to pharmacotherapy, treatment systems may include family therapy, individual therapy, cognitive behavioral therapy, social skills training, employment training, or supported employment (albeit not all at the same time). These psychosocial interventions have been shown to enhance functional outcomes and the individual's quality of life.

Conclusion

Schizophrenia carries with it a significant morbidity, in terms of both the individual and the families affected and the population as a whole. Over the past half century, significant advances have been made in standardizing the diagnostic criteria and characterizing the course of this illness. However, it

remains clear that there is a large degree of inter-individual heterogeneity in the severity of symptoms, in the degree of between episode symptom remission, and in the response to treatment. Three main areas of focus hold potential for gaining a better understanding of this illness. First, the explosion of molecular genetic studies in the last 20–30 years, both at the population level and at the individual level, holds the promise of improving our understanding of the genes and genetic alterations which place an individual at risk for developing schizophrenia. Second, ongoing work aimed at further characterizing the contributory environmental factors linked to increased risk for schizophrenia may, in combination with genetic studies, allow better prediction and perhaps modulation of the risks for developing the disorder. Finally, the development of treatments, with targets beyond the standard dopamine D2 receptors, to address the cognitive deficits and negative symptoms of the illness could enhance the long-term outcome and improve the day-to-day quality of life for those with schizophrenia.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Antipsychotic Drugs; Cognitive Deficits in Schizophrenia; Cognitive Dysfunction in Psychiatric Disorders; Schizophrenia: Genetics.

Further Reading

Adityanjee Y, Aderibigbe A, Theodoridis D, and Vieweg VR (1999) Dementia praecox to schizophrenia: The first 100 years. *Psychiatry and Clinical Neurosciences* 53(4): 437–448.

an der Heiden W and Hafner H (2000) The epidemiology of onset and course of schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience* 250(6): 292–303.

Ban TA (2004) Neuropsychopharmacology and the genetics of schizophrenia: A history of the diagnosis of schizophrenia. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 28(5): 753–762.

Hafner H and an der Heiden W (1997) Epidemiology of schizophrenia. *Canadian Journal of Psychiatry* 42(2): 139–151.

Jablensky A (2000) Epidemiology of schizophrenia: The global burden of disease and disability. *European Archives of Psychiatry and Clinical Neuroscience* 250(6): 274–285.

Jobe TH and Harrow M (2005) Long-term outcome of patients with schizophrenia: A review. *Canadian Journal of Psychiatry* 50(14): 892–900.

Lichtermann D, Karbe E, and Maier W (2000) The genetic epidemiology of schizophrenia and of schizophrenia spectrum disorders. *European Archives of Psychiatry and Clinical Neuroscience* 250(6): 304–310.

Lieberman JA, Perkins D, Belger A, et al. (2001) The early stages of schizophrenia: Speculations on pathogenesis, pathophysiology, and therapeutic approaches. *Biological Psychiatry* 50(11): 884–897.

Lopez-Munoz F, Alamo C, Cuenca E, et al. (2005) History of the discovery and clinical introduction of chlorpromazine. *Annals of Clinical Psychiatry* 17(3): 113–135.

McGrath J, Saha S, Welham J, et al. (2004) A systematic review of the incidence of schizophrenia: The distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC Medicine* 2: 13.

Saha S, Chant D, Welham J, and McGrath J (2005) A systematic review of the prevalence of schizophrenia. *PLoS Medicine* 2(5): e141.

Schultz SK and Andreasen NC (1999) Schizophrenia. *Lancet* 353 (9162): 1425–1430.

Schizophrenia: Genetics

R M Twyman, University of York, York, UK
N Amin, University of Warwick, Coventry, UK

© 2009 Published by Elsevier Ltd.

Introduction

Schizophrenia is a common psychiatric disorder with a prevalence of 0.5–1% among the general population. Symptoms include auditory hallucinations, paranoid delusions, disorganized speech and thinking, and social withdrawal. However, the disease is divided into several subtypes in which certain symptoms are more apparent than others; for example, the paranoid type is characterized by hallucinations and delusions in the absence of disorganized thinking. This wide spectrum of possible symptoms makes the determination of underlying genetic factors more complex, as does the fact that the symptoms overlap to a certain degree with those of other conditions, such as bipolar disorder. The symptoms may begin in early childhood, but the disorder typically becomes fully apparent during the late teens or early 20 s. For affected individuals there is no cure, and although therapy may alleviate some of the symptoms, the prospects for a full social and occupational life are limited.

The causes of schizophrenia are not yet definitively established, but the familial tendency of the disease and results from twin concordance studies suggest there is a strong genetic component. Until comparatively recently there was little information about the genes involved, even though a number of susceptibility loci had been identified. Schizophrenia is a polygenic disease and almost certainly results from the combined effects of several susceptibility alleles at different loci, together with the appropriate environmental stimuli. There has been renewed interest in identifying schizophrenia susceptibility genes in the last few years due to the development and application of high-throughput functional genomics techniques such as expression profiling, large-scale mutagenesis, and protein interaction analysis.

Genetic Epidemiology

During almost a century of epidemiological research into schizophrenia, a surfeit of data has built up to demonstrate that the disease is heritable. The disease has a far greater incidence among the close relatives of an afflicted individual than among the general population. Furthermore, the likelihood of developing the disease increases with increasing genetic similarity to an affected individual. Among monozygotic twins, the chances of one twin getting the disease when the other has already been diagnosed have been estimated to be as low as 11% in one study and as high as 28% in another, while these rates fall to 2–5% in dizygotic twins. The incidence of schizophrenia is also consistently higher among the biological relatives of adopted schizophrenics than among the adoptive relatives. Such studies present a strong argument for a genetic basis to the disease, although many environmental triggers have also been identified, including prenatal and developmental effects, social disadvantage, drug use, neural processes, and psychological mechanisms.

The Complexity of Schizophrenia

Schizophrenia is a complex disorder, both genetically and phenotypically. Genetically, schizophrenia is unlikely to be caused by single gene mutations. No gene has yet been discovered which alone, when mutated, is sufficient to cause schizophrenia. Furthermore, the pattern of inheritance of schizophrenia is not consistent with a disease caused by a single Mendelian trait. Although the risk of acquiring schizophrenia decreases in line with genetic dissimilarity to an affected individual, the decrease is rapid and nonlinear. The observed trend is consistent with multiple susceptibility genes, each having a small individual effect but a large combined effect. To complicate matters further, schizophrenia is a genetically heterogeneous disease. There may be a pool of susceptibility genes, from which a number of different combinations may lead to different subtypes of schizophrenia.

Clinically, schizophrenia manifests itself in a wide variety of symptoms. It is still not clear if schizophrenia, as currently diagnosed, represents one disorder or several related and overlapping disorders. A further complication is that there may be genetic overlap with other psychiatric disorders. That is, genes predisposing to schizophrenia may also be involved in other psychotic disorders that display similar symptoms. As with most complex psychiatric disorders, uncertainty over the clinical, phenotypic definition hinders the study of the genetic basis of the disease.

Techniques to Identify the Genes

Linkage Studies

Linkage studies are used to identify regions of a chromosome that may contain genes responsible for or contributing to a disease or that confer susceptibility

to a disease. Depending on the genetic background, a susceptibility locus in one family may appear to be a clear, disease-causing locus in another. Linkage studies identify genes that are closely physically linked to a particular marker or set of markers found on the same chromosome. The principle is that genes on separate chromosomes will undergo random independent assortment, while those on the same chromosome will tend to be inherited together depending on the physical distance between them and hence the likelihood of a crossover occurring during meiosis. Linkage analysis relies on the availability of heterozygous markers to track normal and disease alleles through a family pedigree and to identify potential recombination events caused by crossovers during meiosis. If a candidate gene is so tightly linked to a marker that recombination never happens or is a very rare occurrence, that marker is likely to be found very close to the disease gene or even within it and can be used as the basis for map-based cloning and eventual gene isolation.

Association Studies

Association studies are used to identify alleles that are present with greater or lower frequency in a group of individuals with a disease, as compared to the general population. Ideally entire genomes would be compared, but this is not feasible with the current state of the art. Instead, association studies for schizophrenia have focused on candidate genes, that is, those already implicated in the disease. Candidate genes may be implicated in two ways, functionally or positionally. Functional candidate genes are suspected, on the basis of their known or proposed function, of having a biological role in the development of the disease. Positional candidate genes are identified by linkage analysis or more occasionally through the analysis of chromosome aberrations in which breakage events affect a schizophrenia susceptibility locus.

Major Genes Implicated in Schizophrenia

The list of candidate genes implicated in susceptibility to or the progression of schizophrenia grows longer every month, particularly with the advent of large-scale functional studies based on messenger RNA expression profiling and proteomics. However, only a small proportion of these genes have strong experimental evidence linking them functionally to the disease; those with the greatest weight of evidence behind them are summarized in Table 1 and described in more detail below.

COMT

Catechol-*O*-methyltransferase (COMT) is an enzyme involved in dopamine metabolism, specifically the breakdown of the catecholamine neurotransmitters (dopamine, epinephrine, and norepinephrine). The *COMT* gene was first implicated in schizophrenia when it was found in a region of chromosome 22 (22q11) which is deleted in velo-cardio-facial syndrome (VCFS). Individuals with VCFS have approximately a 25% probability of developing schizophrenia. There is also evidence from association studies that the *COMT* gene is involved in schizophrenia, increasing the likelihood of the condition occurring spontaneously or following adolescent cannabis use.

The pathogenic mechanism is not precisely understood, but it may involve the effects of the gene on dopaminergic transmission in the prefrontal cortex. An SNP in *COMT* has been shown to affect cognitive tasks broadly related to executive function (abstract thought, response inhibition, rule acquisition). The resulting amino acid substitution (valine for methionine) produces an enzyme variant that metabolizes dopamine and other neurotransmitters four times faster than normal, thus reducing dopaminergic stimulation of the postsynaptic neuron. Consequently, neurons expressing this version of COMT show higher levels of activation during certain cognitive tasks since more action potentials are required to achieve the same level of postsynaptic stimulation.

DAO and *G72*

DAO encodes D-amino acid oxidase, which metabolizes the oxidative deamination of neutral D-amino acids, including D-serine. Another gene, named *G72*,

Table 1 Major susceptibility genes with functional links to schizophrenia

Gene	*Description*	*Chromosomal location*
COMT	Catechol-*O*-methyltransferase	22q11
DAO/DAOA	D-amino acid oxidase and D-amino acid oxidase activator	13q22
DISC1	Disrupted in schizophrenia 1	1q42
DTNBP1	Dystrobrevin binding protein 1	6p22
RGS4	Regulator of G-protein signaling 4	1q21–22
GRM-3	Metabotropic glutamate receptor 3	7q21–22
NRG1	Neuregulin 1	8p12–21

encodes D-amino acid oxidase activator (DAOA), which physically interacts with DAO. Through linkage and association studies, both of these genes have been shown to be associated with an increased susceptibility to schizophrenia. D-serine is predominantly confined to the forebrain, where *N*-methyl-D-aspartate (NMDA) receptors are enriched, and is thought to be an endogenous ligand for these receptors. It has been shown to potentiate NMDA receptor-mediated transmission by selective stimulation of the strychnine-insensitive glycine site of the NMDA receptor with an affinity similar to glycine, but no affinity for the inhibitory glycine receptor. Mutations affecting the activity of DAO and DAOA would modulate the level of available D-serine and therefore impact signaling through NMDA receptors in the forebrain. Despite this, no specific single-nucleotide polymorphisms (SNPs) have been found in either gene, which are also associated with bipolar disorder.

DISC1

Disrupted in schizophrenia 1 (*DISC1*) was first identified at the breakpoint in a chromosomal translocation involving chromosome 1. *DISC1* has since been implicated in schizophrenia through multiple association studies. The gene has also been implicated in several other psychological disorders, including schizoaffective disorder, bipolar affective disorder, and major depression. The gene product is known to bind phosphodiesterase 4B, which regulates cyclic adenosine monophosphate signaling and is thought to be involved in learning and memory processes.

The yeast two-hybrid system for identifying interacting proteins has revealed several proteins that bind DISC1. The first was fasciculation and elongation protein zeta-1, which plays a role in the control of axonal outgrowth and fasciculation. The two proteins were subsequently shown to be colocalized in the growth cones of hippocampal axons, and evidence suggests that the interaction is required for neurite outgrowth. Further studies identified additional binding partners Kendrin and DISC1-binding zinc-finger protein, the latter of which was shown to be expressed strongly in the hippocampus, olfactory tubercle, cerebral cortex, and striatum. This pattern was similar to that of pituitary adenylate cyclase, suggesting a role for the pituitary adenylate cyclase-activating polypeptide pathway in providing further insight into the basis of psychiatric disorders.

DTNBP1

The gene encoding dystrobrevin-binding protein 1 (DTNBP1), also known as dysbindin, was first identified as a candidate schizophrenia gene through linkage studies and then through multiple independent association studies although there were also negative results from such studies in certain population groups. Many SNPs have been identified, although all in noncoding regions, so their impact, if anything at all, is likely to be on expression rather than structure. The biological function of DTNB1 is to bind α- and β-dystrobrevin, which are components of the dystrophin $\bar{N}$ glycoprotein complex. Dysbindin also forms part of the biogenesis of lysozyme-related organelles complex-1. This complex controls cellular trafficking to lysozyme-related organelles and may be involved in the regulation of vesicular trafficking and dendrite branching. There is evidence from postmortem studies that dysbindin expression is decreased in the brains of patients with schizophrenia, particularly in the synapses of certain glutamatergic neurons within the hippocampus.

RGS4

Regulator of G-protein signaling 4 (*RGS4*) has been associated with schizophrenia through both linkage and association studies. Expression levels of RGS4 are significantly reduced in the prefrontal cortex of individuals with schizophrenia. Several SNPs have been implicated in the disease. The product of this gene functions to regulate effects at G-protein-coupled receptors, including dopamine and serotonin receptors.

GRM-3

The product of the metabotropic glutamate receptor 3 (*GRM-3*) gene is involved in the regulation of synaptic glutamate levels. It accomplishes this by controlling expression of the glial glutamate transporter, which functions to remove glutamate from synapses. *GRM-3* has been linked to schizophrenia through a number of association studies, and a number of SNPs have been identified which may confer susceptibility to schizophrenia. Studies have shown that inheritance of particular alleles of *GRM-3* can result in reduced expression of the glial glutamate transporter.

NRG1

NRG1 encodes neuregulin 1, a family of protein ligands produced by alternative splicing that act on the epidermal growth factor receptor family. Six different isoforms of neuregulin 1 have been identified thus far, with a diverse range of functions. Neuregulin 1–ErbB4 interactions are thought to play a significant role in the pathological mechanism of schizophrenia, and this may reflect increased availability of the ligand rather than any differences in structure. NRG1 has

been linked to schizophrenia through a number of independent linkage and association studies, including a study of the Icelandic deCODE database which yielded a high-risk haplotype at the 5′ end of the gene (SNP8NRG243177) that enhances the expression of Type IV NRG1 in the brains of schizophrenic patients.

Is Schizophrenia a Genetic Disorder of the Synapse?

Glutamate is a major neurotransmitter of the brain that functions through the activation of several types of receptor. Crucially, one type of glutamate receptor, the NMDA receptor, has been implicated in schizophrenia. Pharmaceutical inhibition of NMDA receptors, through use of the recreational drug PCP, for example, can induce many of the symptoms associated with schizophrenia. Drugs that enhance NMDA function, on the other hand, can relieve many of the symptoms of schizophrenia. Several of the genes implicated in schizophrenia, including *DTNBP1*, *NRG1*, *GRM-3*, *DAO/DAOA*, and *RGS4*, encode proteins that have the potential to influence the function of glutamatergic synapses. Dysbindin, for example, may affect the uptake of glutamate into synaptic vesicles. GRM-3 is known to regulate synaptic glutamate levels. NRG1 is released from glutamate terminals and can regulate NMDA receptors. DAO and DAOA may also regulate NMDA receptors indirectly, through their action on D-serine.

Although the evidence pointing toward a role for glutamatergic synapses in schizophrenia is promising, there are reasons to be cautious. The genes with the strongest evidence for a role in schizophrenia, particularly *NRG1* and *DTNBP1*, encode proteins with multiple functions, many of which could be involved in schizophrenia without directly involving glutamatergic synapses. On the other hand, the genes with the most direct biological roles in affecting synapses, such as *DAO*, *DAOA*, and *RGS4*, seem to be those with the weakest associations with the disease.

Summary

Several genes have been implicated in the development of schizophrenia. However, there remain concerns about the validity of the findings since no mutation predisposing to schizophrenia has yet been discovered in any of these genes. Among the variant alleles implicated in the disease, the genetic evidence is not definitive and is frequently inconsistent, with some studies reporting positive associations with the disease and others reporting negative associations for the same gene. In several cases, the susceptibility genes appear to have quantitative rather than qualitative effects, with mutations in regulatory regions either increasing or decreasing the amount of product available. This would help explain the absence of definitive structural mutations in such genes and would also explain why association studies in different geographical and ethnic groups sometimes come up with contradictory results, the implication being that compensatory mechanisms might be present to ameliorate such quantitative effects. For example, there is no association evidence between schizophrenia and DTNBP1 in Taiwanese families, perhaps because these families have an allele at a different locus which suppresses the effect of DTNBP1. However, these inconsistencies may also be a feature of the innate limitations of the techniques used (linkage and association studies), which are generally more suited to the identification of genes with large individual effects on a disease.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Cognitive Deficits in Schizophrenia; Genomic Disorder and Gene Expression in the Developing CNS; Psychiatric Genomics and Expression Profiling; Psychiatric Disorders: Functional Genomics; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome.

Further Reading

Bray NJ, Preece A, Williams NM, et al. (2005) Haplotypes at the dystrobrevin binding protein 1 (DTNBP1) gene locus mediate risk for schizophrenia through reduced DTNBP1 expression. *Human Molecular Genetics* 14: 1947–1954.

Craddock M, O'Donovan MC, and Owen MJ (2007) Phenotypic and genetic complexity of psychosis. *British Journal of Psychiatry* 190: 200–203.

Fukumaki Y and Shibata H (2003) Glutamate receptor genes as candidates for schizophrenia susceptibility. *Drug Development Research* 60: 137–151.

Harrison PJ and Weinberger DR (2005) Schizophrenia genes, gene expression, and neuropathology: On the matter of their convergence. *Molecular Psychiatry* 10: 40–68.

McClellan JM, Susser E, and King M-C (2007) Schizophrenia: A common disease caused by multiple rare alleles. *British Journal of Psychiatry* 190: 194–199.

McGuffin P, Owen MJ, and Gottesman II (2002) *Psychiatric Genetics and Genomics*. Oxford, UK: Oxford University Press.

Owen MJ, Craddock N, and Donovan MC (2005) Schizophrenia: Genes at last? *Trends in Genetics* 21: 518–525.

Schwann Cell Development

R Mirsky and K R Jessen, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Schwann cells, found in association with axons in peripheral nerves, are the main glial cell type in the peripheral nervous system (PNS). They exist in two quite distinct forms, either making myelin sheaths round large axons to speed impulse conduction or enclosing groups of smaller axons, holding them together in tight bundles. A mature peripheral nerve therefore consists of a mixture of myelinated and unmyelinated axon–Schwann cell units, usually referred to as nerve fibers. This article provides an overview of the development of Schwann cells from their origins in the neural crest. The generation of PNS glia parallels the formation of glial cells in the central nervous system (CNS). Nevertheless, many of the key molecules that control gliogenesis differ between the two systems.

The Schwann Cell Lineage

The neural crest is a transient group of cells that segregates from the dorsal neural tube and gives rise to a variety of derivatives, including autonomic and sensory neurons, chromaffin cells, glial cells including Schwann cells and satellite cells of the autonomic and sensory ganglia, fibroblastic cells, smooth muscle cells, and melanocytes (Figure 1). The majority of Schwann cells are derived from the group of neural crest cells that migrate ventrally through the anterior part of the somites, although it has been shown that Schwann cells in the dorsal roots are derived from a specialized group of cells called boundary cap cells that also give rise to some neurons and satellite glial cells in dorsal root sensory ganglia (DRG).

Mature myelinating and nonmyelinating Schwann cells arise from the neural crest via two intermediate stages, the Schwann cell precursor and the immature Schwann cell (Figure 2). Schwann cell precursors are the glial cells found in embryonic day (E) 14 or 15 rat nerves (mouse E12 or E13). Immature Schwann cells are generated from Schwann cell precursors and populate rat nerves from E17 or E18 (mouse E15 or E16) to approximately the time of birth, when myelination starts. At the immature Schwann cell stage, Schwann cells that by chance come to be associated with large axons receive axon-associated signals that instruct them to make myelin sheaths, and they adopt the specific gene expression profile appropriate for myelination. Mature nonmyelinating Schwann cells only appear approximately 2 weeks after myelination starts. These cells also express genes that differentiate them from myelinating cells and typically hold several small-diameter axons individually in troughs that run along the cell surface, forming unmyelinated (Remak) fibers.

Three major developmental steps define the lineage – that is, the transition from migrating neural crest cells to axon-associated Schwann cell precursors, the transition from precursors to immature Schwann cells, and finally the divergence of this population to form the two mature Schwann cell types found in adult nerves. At all stages, there is a continuous close association between these cells and axons. Also, it is a striking feature of all the cell types that they are highly dependent on survival factors, mitogens, and differentiation signals from axons. Another notable feature is plasticity, because much of the developmental sequence is readily reversible. For example, mature myelinating and nonmyelinating Schwann cells respond to nerve injury by reverting to a phenotype similar to that of immature Schwann cells, and Schwann cell precursors can be diverted, at least *in vitro*, to other neural crest derivatives. Only the transition from precursors to immature Schwann cells appears to be irreversible.

In vivo, rapid proliferation is a characteristic of all early stages of the lineage – namely, neural crest cells, Schwann cell precursors, and immature Schwann cells. In contrast, the onset of myelination is clearly linked with cell cycle exit, and both myelinating and nonmyelinating cells are quiescent in normal adult nerves. However, the cells retain the potential to proliferate because they reenter the cell cycle when they dedifferentiate in response to nerve injury. Apoptotic cell death is also a feature of developing cells in early nerves, whereas myelinating and mature nonmyelinating cells are resistant to apoptosis.

Markers of Schwann Cell Development

Each cell stage in the embryonic phase of the lineage can be defined by a distinct combination of differentiation markers (Figure 3). Additional criteria, such as morphology, relationships to other cells and tissues, and response to extrinsic signals, can also be used to define the cell types and to study the signals that are important in controlling progression from one stage to the next.

The markers can be divided into five different categories depending on expression patterns (Figure 3).

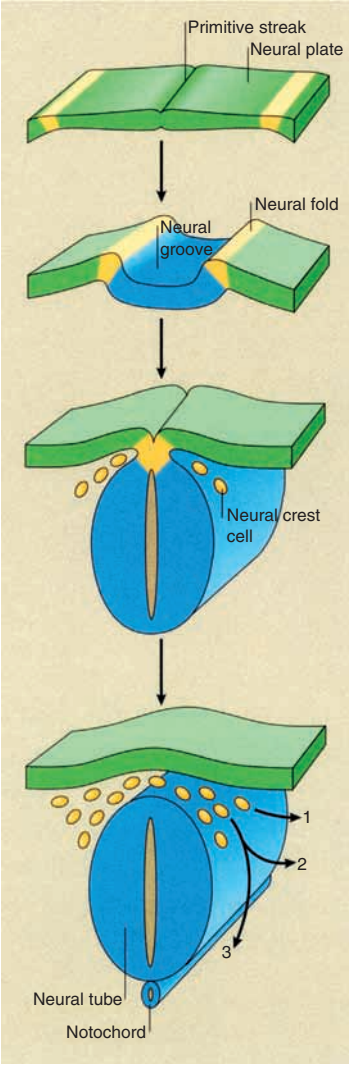

Figure 1 The neural crest. In a process known as neurulation, the neural plate, which is found along the dorsal surface of an embryo, gradually folds in on itself to form the neural groove. As the neural folds fuse to form the neural tube, the neural crest cells segregate from the tips of the folds. After taking up an initial position at the dorsal surface of the tube, the crest cells in the trunk region soon migrate along one of two major streams: in a lateral direction (1) to give rise to melanocytes in the skin, and in a ventral direction (2 and 3) to give rise to neurons in dorsal root sensory ganglia and glia (2) or glia, autonomic neurons, and chromaffin cells (3). Neural crest cells in the most anterior part of the trunk, the cardiac crest, also generate fibroblasts and smooth muscle cells, and the cephalic crest in the head region also forms the cells of cartilage and bone. The mechanisms that allow the apparently homogeneous population of crest cells to generate such diversity have been intensively studied. It is now considered likely that some neural crest cells are already committed to certain fates, whereas others are multipotent. Although some cells may enter lineages in a stochastic and undirected manner, a combination of positive and negative instructive signals probably plays an important part in directing neural crest cell differentiation. How migrating neural crest cells, which initially move through immature connective tissue on each side of the neural tube, end up as Schwann cell precursors in tight association with axons in early embryonic nerves is not clear, either in terms of their detailed migratory route or in terms of the signals that cause these cells to adopt an early glial phenotype. Reproduced from Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

The first category, exemplified by the transcription factor SRY (sex determining region Y) box 10 (SOX10), includes molecules that are present at all developmental stages; the second, exemplified by the transcription factor activator protein 2α (AP2α), includes genes/proteins present at high levels in neural crest cells and Schwann cell precursors but which are downregulated in immature Schwann cells; the third category, of which cadherin 19 is the only known example, represents genes expressed only in Schwann cell precursors; and the fourth category represents genes/proteins expressed by Schwann cell precursors and immature Schwann cells but not by neural crest cells, such as brain fatty acid-binding protein, myelin protein zero (P_0), and connexin 29. The last category represents molecules such as S100 or glial fibrillary acidic protein (GFAP) that are present in high levels on immature Schwann cells but are low or absent from neural crest cells or Schwann cell precursors. It should be emphasized that the use of technologies such as Affymatrix that allow comparative analysis across the genome reveals larger sets of different genes, but which of these will be useful in the study of mechanisms that control development of the lineage remains to be determined.

Several additional criteria provide crucial information in analyzing Schwann cell development. Schwann cell precursors and immature Schwann cells, but not neural crest cells, share a feature characteristic of glial cells in both the PNS and the CNS, namely their close physical apposition with axons (neurons). Schwann cell precursors also differ from migrating crest cells in their response to survival factors and in being relatively insensitive to the neurogenic actions of bone morphogenetic protein (BMP) 2. They are also more sensitive to the actions of Notch and are strongly biased toward the generation of Schwann cells rather than other neural crest derivatives.

A striking difference between Schwann cells and Schwann cell precursors is the ability of Schwann cells to support their own survival in the absence of axons using autocrine survival circuits. The cytoarchitecture of the nerves at the stage when they contain Schwann cell precursors (E14 or E15) is also distinctly different from that of nerves containing immature Schwann cells (E17 or E18).

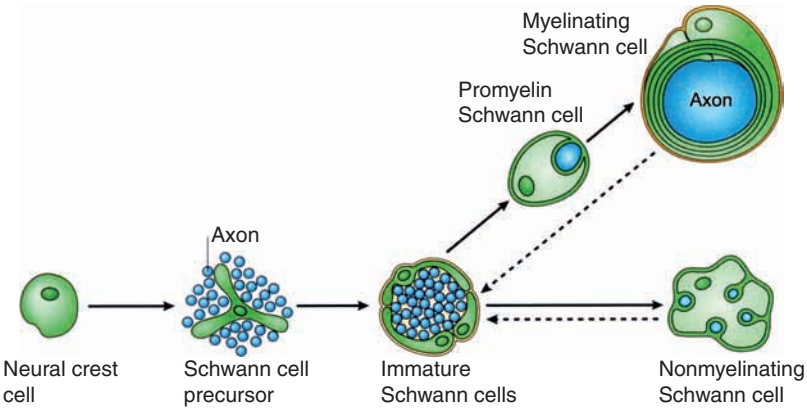

Figure 2 The Schwann cell lineage. Schematic illustration of the main cell types and developmental transitions involved in Schwann cell development. Dashed arrows indicate the reversibility of the final, largely postnatal transition during which mature myelinating and nonmyelinating Schwann cells are generated. The embryonic phase of Schwann cell development involves three transient cell populations. First are migrating neural crest cells, which are discussed further in the legend to **Figure 1**. Second are Schwann cell precursors. These cells express various differentiation markers that are not found in neural crest cells, including brain fatty acid-binding protein (BFABP), protein zero (P_0), and desert hedgehog (DHH) (**Figure 3**). At any one time, a rapidly developing population of cells, such as the glia of embryonic nerves, will contain some cells that are rather more advanced than others. Third are immature Schwann cells. All immature Schwann cells are considered to have the same developmental potential, and their fate is determined by axons with which they associate. Myelination occurs only in Schwann cells that by chance envelop large-diameter axons; Schwann cells that ensheathe small-diameter axons progress to become mature nonmyelinating cells. Reproduced from Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

Gliogenesis from the Neural Crest

The generation of more differentiated cells from stem cells or cells that share some stem cell properties, including the neural crest, is a topic that continues to generate widespread interest. The consensus is that cell specification involves interplay between cell-autonomous intracellular signaling and extracellular cues arising from the niche in which the stem cell resides. The best evidence is that generation of Schwann cells and other glia from the neural crest involves both of these mechanisms.

The Role of SOX10 in PNS Gliogenesis

In the generation of glial cells from the neural crest, the transcription factor SOX10 is the most important player known so far, because it is the only gene known to be essential for this process. Initially expressed by all migrating neural crest cells, its expression persists in the developing satellite glial cells of the DRG, in Schwann cell precursors, and in Schwann cells of peripheral nerves. However, SOX10 is downregulated very early in neurogenesis. In line with this, neurons are initially generated in normal numbers in mice in which SOX10 has been inactivated. In contrast, satellite cells and Schwann cells fail to develop in these animals. In place of satellite cells (i.e., glia), the DRG contain a population of neural crest-like cells and nerve trunks also contain a few neural crest-like cells but no Schwann cell precursors, indicating that in the absence of SOX10 glial specification is blocked. Experiments using cultured cells also suggest that SOX10 plays a role in specifying and maintaining the glial phenotype. It may act in part by upregulating levels of the neuregulin receptor ErbB3, thus increasing the responsiveness of neural crest cells to the growth factor β-neuregulin-1. There is evidence that later in development, it acts in conjunction with other transcription factors such as KROX20 (EGR2) to regulate the promoters of genes that are crucial in myelination, such as P_0 and connexin 32.

β-Neuregulin-1 and the Neural Crest

β-Neuregulin-1, signaling through its receptors ErbB2 and ErbB3, is a growth factor that has multiple important functions throughout Schwann cell development. In neural crest cell cultures, it inhibits the development of neurons, a function that might lead indirectly to increased development of glia. Overproduction of neurons, however, has not been noted in mutants that lack components of the neuregulin signaling pathway, so it is not clear that β-neureglin-1 regulates neuronal numbers *in vivo*. When β-neuregulin-1 is added to migrating neural crest cell cultures, it increases the proportion of Schwann cells generated. Nevertheless, these cultures generate Schwann cells even in the absence of added neuregulin, and the same is true when the appearance of Schwann cell

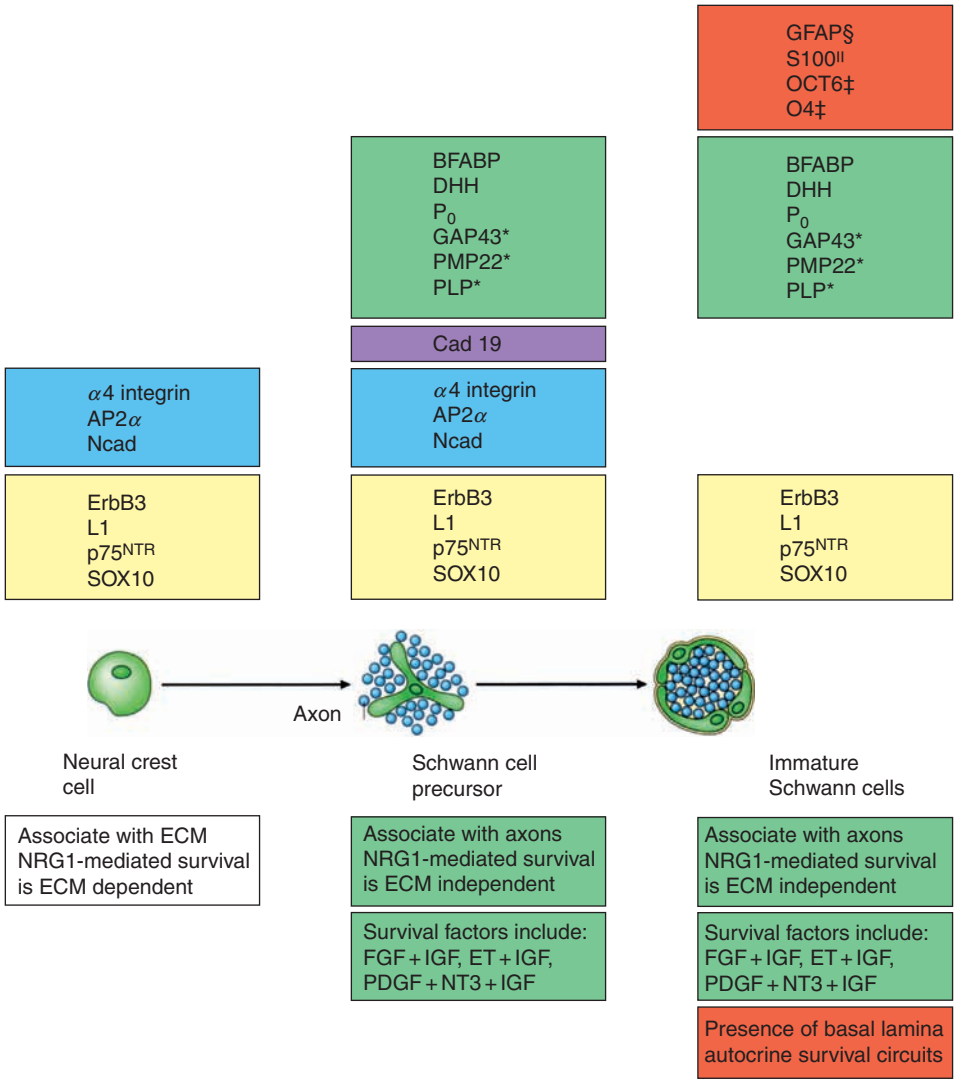

Figure 3 Changes in phenotypic profile as cells progress through the embryonic Schwann cell lineage. Shared profiles are indicated by distinct colors. The boxes above indicate the changes in gene expression that take place during embryonic Schwann cell development. The gene expression shown here is based on observations of endogenous genes rather than on observations of reporter genes in transgenic animals. Note that Cadherin 19 (Cad 19) is exclusively expressed in Schwann cell precursors. Each developmental stage also involves characteristic relationships with surrounding tissues and distinctive cell signaling properties (boxes below lineage drawing). For instance, neural crest cells migrate through extracellular matrix. By contrast, Schwann cell precursors and Schwann cells are embedded among axons with minimal extracellular spaces separating them from nerve cell membranes, a characteristic of glial cells in the CNS and PNS. Basal lamina is absent from migrating crest cells and Schwann cell precursors but appears on Schwann cells. *In vitro*, β-neuregulin-1 (NRG1) only supports neural crest survival in the presence of extracellular matrix (ECM), although this is not required for the β-neuregulin-1-mediated survival of Schwann cell precursors and Schwann cells. Migrating neural crest cells also fail to survive in the presence of several factors that support the survival of Schwann cell precursors and Schwann cells, including combinations such as fibroblast growth factor (FGF) plus insulin-like growth factor (IGF), endothelin (ET) plus IGF, and platelet-derived growth factor (PDGF) plus neurotrophin (NT)-3 and IGF. Schwann cells also have autocrine survival circuits that are absent from Schwann cell precursors. *Proteins that also appear on neuroblasts/early neurons. ‡Markers that are acutely dependent on axons for expression. §Glial fibrillary acidic protein (GFAP) is a late marker of Schwann cell generation since significant expression is not seen until approximately the time of birth. GFAP is reversibly suppressed in myelinating cells. The early expression of GFAP has not been examined carefully in mice. ||Schwann cell precursors have been shown to be S100 calcium-binding protein (S100)-negative and Schwann cells S100-positive using routine immunohistochemical methods; however, low levels of S100 are detectable in many mouse Schwann cell precursors when the sensitivity of the assay is significantly increased. AP2α, activator protein 2α; BFABP, brain fatty acid-binding protein; DHH, desert hedgehog; ErbB3, neuregulin receptor; GAP43, growth-associated protein 43; L1, L1 adhesion molecule; N-cad, N-cadherin; OCT6, octamer-binding transcription factor 6; 04, lipid antigen; PLP, proteolipid protein; PMP22, peripheral myelin protein 22 kDa; P_0, protein zero; $p75^{NTR}$, p75 neurotrophin receptor; SOX10, SRY (sex determining region Y) box 10. Reproduced from Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

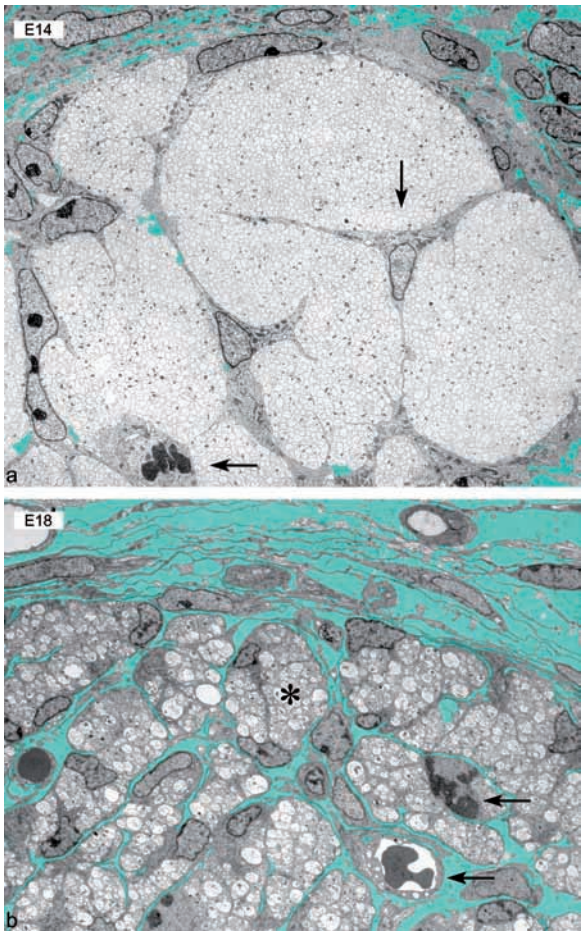

Figure 4 The appearance of early cells in the Schwann cell lineage. (a) An electron microscopic image of a transverse section of a nerve in the hindlimb of a rat embryo at embryonic day (E) 14. Schwann cell precursors branch among the axons inside the nerve (large arrow) and are also found in close apposition to axons at the nerve surface. One precursor cell is undergoing mitosis (small arrow). Extracellular connective tissue space (turquoise), which contains mesenchymal cells, surrounds the nerve but is essentially absent from the nerve. These nerves are also free of blood vessels, and the axons are of smaller and more uniform diameter than those seen in mature nerves. Magnification, ×2000. (b) Schwann cells in a transverse section of the sciatic nerve of a rat embryo at E18, shown at the same magnification. In marked contrast to the nerve at E14, connective tissue spaces now branch through the nerve among compact bundles of immature Schwann cells and their associated axons (Schwann cell families; for example, see asterisk). Blood vessels (small arrow) and fibroblasts (e.g., directly above the vessel) have also appeared inside the nerve. One Schwann cell is undergoing mitosis (large arrow). Outside the nerve (in the uppermost part of the figure), connective tissue, which contains flattened fibroblasts of the early developing perineurium and two blood vessels, can be seen. Reproduced from Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

precursors from neural crest cell cultures is monitored. Furthermore, in mouse mutants in which neuregulin signaling has been abrogated *in vivo*, satellite glia in the DRG are generated, although Schwann cell precursors are severely depleted. In these mutants, most neural crest cells fail to migrate ventrally below the level of the DRG to the sites where sympathetic ganglia are formed, resulting in underdeveloped (hypoplastic) ganglia. In summary, these data indicate that, in principle, β-neuregulin-1 is not required for the generation of glia from the neural crest, although it modulates this process by suppressing neurogenesis and by promoting migration of neural crest cells or early crest derivatives, particularly those that will form sympathetic ganglia. There is also evidence that β-neuregulin-1 accelerates the transition of Schwann cell precursors to Schwann cells.

Schwann Cell Precursors

In peripheral nerve trunks, Schwann cell precursors represent the first clearly defined stage of glial differentiation and are the cells from which the immature Schwann cells are derived. They represent the large majority of cells in the limb nerves of E14/E15 rats (E12/E13 mice). They are initially seen at the edge of early embryonic nerves but are later seen inside them as well. They communally surround large groups of axons with their sheetlike processes and divide the nerves into territories. At this stage, the nerves are compact structures that contain essentially no connective tissue and are not vascularized (Figure 4).

β-Neuregulin-1 and Schwann Cell Precursors

A vital role for β-neuregulin-1 in the Schwann cell lineage is indicated by the observation, mentioned previously, that Schwann cell precursors, and later Schwann cells, are absent or severely depleted in the peripheral nerves of mouse mutants that lack neuregulin signaling. This probably reflects the major role of β-neuregulin-1 as an essential survival factor and mitogen for Schwann cell precursors, although impaired migration might also have a role. Axonally derived β-neuregulin-1 is an essential survival factor and mitogen for Schwann cell precursors *in vitro* and it has been shown that the most important isoform expressed by axons is the membrane-bound β-neuregulin-1 type III (Figure 5). It is the major isoform present in DRG and motor neurons *in vivo*. In mice in which this isoform has been selectively eliminated, although peripheral nerves are initially populated by Schwann cell precursors, few remain after E14. *In vivo*, Schwann cell precursors die after nerve injury and this can be prevented by externally applied β-neuregulin-1. Taken together, this indicates that axonal β-neuregulin-1 type III is a key survival signal in embryonic Schwann cell development, being an essential survival factor for Schwann cell precursors.

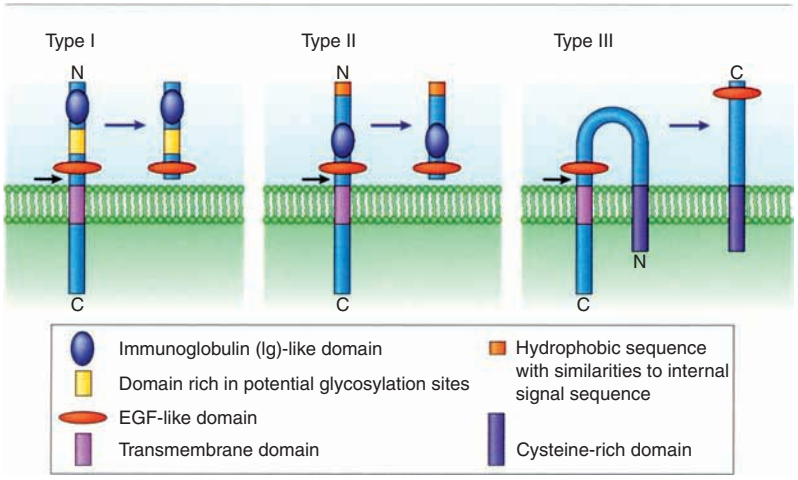

Figure 5 Neuregulin-1. β-Neuregulin-1 seems to have exceptionally numerous and varied functions in Schwann cell biology. It is involved in neural crest migration and has been implicated in the specification of neural crest stem cells and shown to be essential for the survival of Schwann cell precursors. It is also involved in Schwann cell generation, proliferation, and survival. In postnatal nerves, β-neuregulin-1 is a positive regulator of myelin sheath thickness, but paradoxically, at least in experiments, it appears to drive the dedifferentiation of Schwann cells in injured nerve fibers. No other molecule has been proposed to be so comprehensively involved in the control of Schwann cell development. There are a surprising number (>15) of neuregulin-1 α and β protein isoforms. The schematic structures of the main isoforms found in the nervous system are shown here. Although splice variants without the transmembrane domain exist for all these isoforms, transmembrane isoforms (as shown here together with the products of a proteolytic cleavage in the juxtamembrane area) predominate in the nervous system. The epidermal growth factor (EGF) domain is found in all bioactive forms of neuregulin-1 and is sufficient for activation of ErbB receptor kinase. The type III β-neuregulin-1 isoform is expressed in axons and is the main regulator of survival of Schwann cell precursors and myelin sheath thickness. It is thought to have two membrane-spanning domains and to undergo proteolytic cleavage that generates a membrane-attached protein carrying the EGF domain. β-Neuregulin-1 shows high-affinity binding to two receptors, ErbB3 and ErbB4, whereas a related protein, ErbB2, acts as a coreceptor in ErbB3–ErbB2 and ErbB4–ErbB2 complexes. The former is the main receptor in peripheral glial cells. The action of axonal β-neuregulin-1 type III on ErbB3–ErbB2 in developing Schwann cells is probably the best established molecular signaling pathway between neurons and glia in the PNS. Reproduced from Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

In vitro experiments show that in addition to its survival function, β-neuregulin-1 accelerates the transition from Schwann cell precursors to immature Schwann cells. This transition is also controlled by another signaling system present in embryonic nerves – endothelin. In contrast to neuregulin, endothelins negatively regulate Schwann cell generation, as shown by the observation that Schwann cells are generated prematurely in rats in which the endothelin B receptor is inactivated.

Notch and Schwann Cell Precursors

Notch transmembrane receptors and their ligands delta, jagged, and serrate are known to influence glial cell fate choices in the developing nervous system. In the CNS, the classical view of Notch signaling is that it acts to maintain neural stem cells in an undifferentiated state, but evidence suggests that it can promote the appearance of radial glia, astrocytes, and Muller cells. It also stimulates the formation of oligodendrocyte precursors while inhibiting their progression to myelinating cells.

The question of whether Notch instructively promotes gliogenesis from neural crest cells is still controversial. Notch activation in neural crest cells inhibits the generation of neurons *in vivo* and *in vitro*, whereas in cultures derived from a subpopulation of cells from E14 sciatic nerves it increases the number of GFAP-positive Schwann cells and the rate at which they are generated. There is also evidence that Notch, like β-neuregulin-1, acts on early glial cells, stimulating Schwann cell precursors to generate Schwann cells. Notch also promotes the proliferation of immature Schwann cells *in vitro* and *in vivo*. Although it is not settled whether the only actions of Notch are to accelerate lineage progression and stimulate proliferation, or whether Notch also directs crest cells to the glial lineage, there are strong parallels between the actions of Notch and β-neuregulin-1 in this system. Both signals suppress neurogenesis in

crest cells and stimulate Schwann cell generation from precursors and Schwann cell proliferation. It will be interesting to explore the significance of this in future studies.

Immature Schwann Cells

The switch from Schwann cell precursor to immature Schwann cell between E15 and E17 in rat (mouse E13–E15) involves a coordinated change in molecular expression and in the response of the cells to survival signals and mitogens. This is accompanied by architectural reorganization that involves the appearance of blood vessels and connective tissue in the nerve, eventually leading to the formation of basal laminae on the surface of individual Schwann cells. A distinct layer of perineurium also starts to appear at the nerve surface (Figure 4).

The mechanisms of cell survival also change with this transition. Schwann cell precursors are acutely dependent on β-neuregulin-1 for survival, whereas immature Schwann cells acquire autocrine survival mechanisms. This enables Schwann cells to support their own survival in the absence of axons, a mechanism that is likely to be important for axonal regeneration after injury. The autocrine survival factors that have been identified include a cocktail of insulin growth factor (IGF)-2, platelet-derived growth factor-BB, and neurotrophin (NT)-3, potentiated by laminin, and lysophosphatidic acid and leukemia inhibitory factor. This switch in survival strategy makes biological sense. Precursor survival is completely dependent on neuronally derived β-neuregulin-1, a property that may serve to match axon and Schwann cell precursor numbers and to keep precursors confined to developing nerves as they grow through body tissues to reach their targets. In contrast, the fact that Schwann cells can survive in the absence of axons ensures that if axonal injury occurs the surviving Schwann cells can provide essential factors to support axonal regrowth.

Little is known about the transcription factors that control the precursor to Schwann cell transition. Levels of activator protein 2α (AP2α) are high in precursors and decrease sharply in immature Schwann cells, and enforced expression of AP2α *in vitro* retards the transition. Endothelin also regulates the transition in a negative fashion, as mentioned previously.

As the cells transit from Schwann cell precursors to Schwann cells, developmental options narrow. Dedifferentiation of immature Schwann cells leading to the emergence of Schwann cell precursors has not been observed definitively, and Schwann cells in culture are resistant to signals such as BMPs and fibroblast growth factor 2 that can induce the generation of other neural crest derivatives from Schwann cell precursors.

Boundary Cap Cells Give Rise to Schwann Cells in Spinal Roots

Schwann cells in spinal roots have a different origin from the majority of Schwann cells in peripheral nerves. They are derived from boundary cap cells, a specialized set of cells that originate from the neural crest. These cells are found during embryonic development clustered in groups where dorsal and ventral roots enter and exit the spinal cord. They can be identified by the expression of the transcription factor KROX20 (EGR2) long before this gene appears in myelinating Schwann cells. Lineage tracing studies of the fate of these cells *in vivo* have revealed that they give rise not only to the Schwann cells of the dorsal and ventral roots but also to a small subset of nociceptive neurons and some satellite cells within the DRG. Few boundary cap cell-derived glia were detected in spinal nerves. Therefore, these findings do not affect the classical view that Schwann cells of limb nerves originate from migrating neural crest cells.

Functions of Schwann Cell Precursors and Schwann Cells

Trophic Support of Neuronal Survival

A major function of glial cells is to provide trophic support for developing neurons. In the case of Schwann cell precursors and immature Schwann cells, persuasive support for this view comes from experiments in which these cells have been deleted from peripheral nerves. In mouse mutants that lack SOX10, β-neuregulin-1 isoform III, or the neuregulin receptors ErbB2 or ErbB3, Schwann cell precursors and Schwann cells are depleted or absent owing to the importance of these molecules in gliogenesis and glial survival. Importantly, in these mutants most of the DRG and motor neurons that project into limb nerves die by E14 and E18, respectively, although they are initially generated in normal numbers. This suggests that an important function of precursors and immature Schwann cells is to provide essential survival signals for developing neurons. Impaired axon-target contacts may also contribute to sensory and motor death in the β-neuregulin-1 isoform III mutants.

Together with the evidence that axons provide essential β-neuregulin-1-mediated survival support for Schwann cell precursors that was discussed previously, these studies indicate that there is a discrete early phase of nerve development in which neurons and glia depend on each other for survival.

Morphogenesis: Maintaining the Normal Structure of the Spinal Cord, Nerve Trunks, and Neuromasts

An important morphogenetic role for the early glia of ventral roots has been revealed by studies using several mouse mutants, all of which have in common the absence of glial cells from nerve roots. In all of these mice, the cell bodies of motor neurons are displaced into the ventral roots. Comparable observations have been made in the chick. This indicates that an important function of boundary cap cells, or the Schwann cell precursors that are derived from them, is to maintain the normal location of motor neurons within the spinal cord.

Another morphogenetic function for immature Schwann cells and their precursors is that of holding peripheral nerve trunks together in a unified structure since in mouse mutants that lack these cells, large nerves separate into a number of smaller bundles. Remarkably, however, even without accompanying glial cells, the nerves initially grow out into limbs more or less correctly and find their way to their target areas. A morphogenetic function for neural crest-derived glia is also seen in the lateral line nerve of zebra fish, where developing glial cells (immature Schwann cells or their precursors) control nerve fasciculation and the formation of secondary neuromasts – organs that are specialized to detect water movements.

The Generation of Immature Schwann Cells and Endoneurial Fibroblasts

Lineage tracing studies have revealed an unexpected role for Schwann cell precursors. They show that the relatively small population of fibroblasts found within the nerve at birth (5–10% of total cells) arises from cells in the nerve that express desert hedgehog and p75 neurotrophin receptor ($p75^{NTR}$) and are therefore presumably Schwann cell precursors. This accords with the observation that not only Schwann cells but also fibroblasts appear in the nerve at the precursor–Schwann cell transition. It is also consistent with earlier experiments that demonstrated that, in principle, early PNS glia from rodents and bird nerves have the potential to generate cells other than Schwann cells since they can be induced to generate melanocytes in cell culture. A finding that provides an additional example of the unexpected developmental potential of early PNS glia concerns the boundary cap cells mentioned previously. These cells are tightly associated with axons and express glial-associated genes such as P_0 and KROX20, and they can therefore be regarded as having the phenotype of early glial cells. Nevertheless, these cells give rise not only to glia but also to some DRG neurons during normal development. This is reminiscent of the observation that early CNS glia, namely radial glia, generate both astrocytic glial cells and neurons, and it is in accord with the emerging concept that early glia can act as multipotent progenitors in both the developing PNS and the CNS.

The Transition from Immature Schwann Cells to Myelination

The transition from Schwann cell precursors to immature Schwann cells is essentially complete by E18, and at that time nerves have also acquired the basic tissue architecture known from postnatal nerves since they now contain relatively large extracellular spaces, connective tissue, fibroblasts, and blood vessels. At this stage, the Schwann cells, which have started to assemble a basal lamina, still ensheathe large groups of axons communally (Figure 4). Myelination starts approximately 3 days later, at approximately birth. This requires radial sorting, a process of radical change in the configuration of Schwann cells and axons that allows the larger diameter axons that will become myelinated to acquire a 1:1 relationship with individual Schwann cells, a configuration that is a prerequisite for myelin formation. At the same time as radial sorting is taking place, Schwann cell numbers are adjusted to the number of axons by controlling proliferation and survival. During this period, a number of signaling systems probably act together to prevent premature myelination.

Radial Sorting

This process, which allows individual Schwann cells to become associated with single large-diameter axons, is rapidly being dissected at the molecular level, although many aspects remain unclear. Defects in radial sorting are seen in the absence of laminin isoforms and in the absence of β_1 integrin. Radial sorting also fails in the clawpaw mutant mouse, which has a mutation in Lgi4, a secreted molecule of unknown function that belongs to a family of molecules, another member of which, Lgi1, complexes with Kv potassium channels and causes certain forms of epilepsy. Lastly, lack of a disintegrin and metalloprotease protein (ADAM)22 also causes sorting defects and severe hypomyelination. In the absence of laminin, or the laminin receptor

dystroglycan, Schwann cell proliferation is also impaired and apoptosis increased. Interestingly, these functions are unaffected by the absence of β_1 integrin.

Cell movement is a crucial component of radial sorting, and it is possible that factors that control Schwann cell migration in culture could also affect radial sorting. They include the growth factors β-neuregulin-1, IGFs, and NT-3, all of which promote migration, and brain-derived neurotrophic factor, which inhibits it. Levels of activation of the small GTPase Rac also appear to regulate Schwann cell–axon interactions, and in other cell types low levels of rac activation promote directionally persistent cell migration, whereas high levels promote random migration. There is also evidence that the p38 mitogen-activated protein kinase pathway is required just prior to myelination, perhaps for the correct alignment of axons and Schwann cells.

Control of Schwann Cell Numbers

During the process of sorting and myelination that occurs at approximately birth, it becomes important to match the numbers of Schwann cells and axons. By then, the period of neuronal cell death is largely over. This process is therefore controlled by a balance between rates of Schwann cell proliferation and death.

Experiments using cultured cells suggest that axons are the major stimulators of Schwann cell proliferation, an idea supported by the observation that *in vivo* Schwann cell proliferation decreases as Schwann cells lose contact with axons in transected nerves of newborn animals. In co-cultures of Schwann cells and DRG neurons, β-neuregulin-1 acts as a major axonal mitogen, but this has not been confirmed *in vivo*. Transforming growth factor (TGF)-βs are also potential mitogens that are present in developing nerves. Studies show that mouse Schwann cells that lack TGF-β type II receptor proliferate more slowly *in vivo* than normal Schwann cells, demonstrating that TGF-β is normally involved in regulating Schwann cell proliferation during nerve development.

Schwann cell survival in developing nerves is likely controlled by a balance between factors that support survival and those that promote cell death. Axonally derived β-neuregulin-1 and autocrine circuits provide survival support, together with laminin acting through dystroglycan receptors. Two signals that promote cell death have been identified *in vivo*, including TGF-β. Deletion of TGF-β type II receptors suppresses developmental death in E18 to newborn nerves, and cell death after newborn nerve transection is also lower in these mutant nerves. Signals acting through $p75^{NTR}$, possibly nerve growth factor, are required for the cell death that occurs after neonatal nerve transection but not for normally occurring developmental death.

The Onset of Myelination

Schwann cell myelination is a remarkable example of cell–cell interaction in which the association of an immature Schwann cell with a large-diameter axon induces a radical change in morphological and molecular phenotype. This leads to the formation of the myelin sheath and reciprocal changes in axonal membrane proteins, ion channels, and cytoskeleton that enable saltatory conduction along large nerves. The role of the basal lamina in promoting myelination is well established, although the molecular mechanisms involved are just beginning to be revealed.

The Role of β-Neuregulin-1 in Myelination

It has been shown that β-neuregulin-1 type III is one of the crucial axonal signals involved in controlling Schwann cell myelination. Mutant mice heterozygous for this isoform have thinner myelin sheaths, whereas overexpression results in thicker myelin sheaths and induces myelination of axons that would not normally be myelinated. β-Neuregulin-1 activates signaling pathways that positively regulate myelination, including the phosphatidyl inositol-3 kinase and protein kinase A pathways.

Myelin-Related Transcription Factors

The most important transcription factor involved in myelination is KROX20, which acts together with its partners, the NGF1-A-binding proteins 1 and 2 (NAB1/2), to upregulate a large number of myelin genes and proteins. The absence of KROX20 or, alternatively, the combined absence of NAB1/2, both of which are upregulated by KROX20, results in Schwann cell arrest at the 1:1 promyelin stage and failure to myelinate, indicating the central importance of the KROX20/NAB1/2 complex for gene activation and myelination. In humans, KROX20 (EGR2) mutations are associated with hereditary sensory and motor neuropathies. The octamer-binding transcription factor 6 (OCT6) and, to a lesser extent, brain 2 class III POU-domain protein (BRN2) control the timing of myelination. In mice that lack OCT6, myelination is severely delayed, probably because OCT6 is required for upregulation of KROX20 at the appropriate time in development, whereas BRN2 can partially compensate for loss of OCT6. *In vitro*, the transcription factor NF-κB and the Sloan–Kettering Institute proto-oncogene (SKI) are expressed prior to myelination, and in Schwann cell–neuron co-cultures lack of NF-κB or SKI in Schwann cells results in failure to myelinate.

Negative Regulation of Myelination

Although the onset of myelination is characterized by activation of promyelin pathways, studies indicate that signaling pathways that inhibit myelination also play a role in the timing of myelination. One of these inhibitory signals is the c-Jun N-terminal kinase (JNK) pathway. This pathway is active in Schwann cells of E10 to newborn nerves, where it is required for β-neuregulin-1 and TGF-β-induced signaling, at least in cultured cells. The pathway is inactivated as individual cells start to myelinate using a mechanism that depends in part on KROX20. In mice in which KROX20 has been inactivated, the JNK pathway is still active, proliferation remains high, and the cells arrest at the promyelin stage of development. When expression of JNK pathway constituents is enforced in cultured Schwann cells, myelination in neuron–Schwann cell cultures is blocked and myelin gene expression that would normally result from promyelin signals such as KROX20 is also blocked. The transcription factor SOX2 also promotes proliferation, inhibits myelination when overexpressed in Schwann cells in neuron–Schwann cell co-cultures, and remains high in mutant mice expressing low levels of KROX20 (EGR2). Similarly, Notch signaling, which, as mentioned previously, promotes proliferation in immature Schwann cells, is downregulated as cells start to myelinate *in vivo*, and in neuron–Schwann cell co-cultures enforced expression of Notch inhibits myelination. It is unclear how these apparently diverse inhibitory pathways are interlinked, and the details of how they interact with promyelin signals are just starting to be revealed.

See also: Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission; Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells; Oligodendrocyte and Schwann Cell Identification Methods; Schwann Cells and Axon Relationship; Schwann Cells and Plasticity of the Neuromuscular Junction; Schwann Cell Morphology.

Further Reading

Bermingham JR Jr., Shearin H, Pennington J, et al. (2006) The claw paw mutation reveals a role for Lgi4 in peripheral nerve development. *Nature Neuroscience* 9: 76–84.

Dong Z, Brennan A, Liu N, et al. (1995) NDF is a neuron–glia signal and regulates survival, proliferation, and maturation of rat Schwann cell precursors. *Neuron* 15: 585–596.

Feltri ML and Wrabetz LJ (2005) Laminins and their receptors in Schwann cells and hereditary neuropathies. *Journal of the Peripheral Nervous System* 10: 128–143.

Garratt AN, Britsch S, and Birchmeier C (2000) Neuregulin, a factor with many functions in the life of a Schwann cell. *BioEssays* 22: 987–996.

Jaegle M, Ghazvini M, Mandemakers W, et al. (2003) The POU proteins Brn-2 and Oct-6 share important functions in Schwann cell development. *Genes & Development* 17: 1380–1391.

Jessen KR and Mirsky R (2004) Schwann cell development. In: Lazzarini RA (ed.) *Myelin Biology and Disorders*, pp. 329–359. San Diego: Elsevier.

Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

Joseph NM (2004) Neural crest stem cells undergo multilineage differentiation in developing peripheral nerves to generate endoneurial fibroblasts in addition to Schwann cells. *Development* 131: 5599–5612.

Kubu CJ, Orimoto K, Morrison SJ, et al. (2002) Developmental changes in Notch1 and numb expression mediated by local cell–cell interactions underlie progressively increasing delta sensitivity in neural crest stem cells. *Developmental Biology* 244: 199–214.

Le N, Nagarajan R, Wang JY, et al. (2005) Analysis of congenital hypomyelinating Egr2Lo/Lo nerves identifies Sox2 as an inhibitor of Schwann cell differentiation and myelination. *Proceedings of the National Academy of Sciences of the United States of America* 102: 2596–2601.

Lobsiger CS, Taylor V, and Suter U (2002) The early life of a Schwann cell. *Biological Chemistry* 383: 245–253.

Maro GS, Vermeren M, Voiculescu O, et al. (2004) Neural crest boundary cap cells constitute a source of neuronal and glial cells of the PNS. *Nature Neuroscience* 7: 930–938.

Michailov GV, Sereda MW, Brinkmann BG, et al. (2004) Axonal neuregulin-1 regulates myelin sheath thickness. *Science* 304: 688–689.

Mirsky R and Jessen KR (2005) Molecular signaling in Schwann cell development. In: Dyck PJ and Thomas PK (eds.) *Peripheral Neuropathy*, 4th edn., pp. 341–376. Philadelphia: Saunders.

Morris JK, Lin W, Hauser C, et al. (1999) Rescue of the cardiac defect in ErbB2 mutant mice reveals essential roles of ErbB2 in peripheral nervous system development. *Neuron* 23: 273–283.

Morrison SJ, White PM, Zock C, and Anderson DJ (1999) Prospective identification, isolation by flow cytometry, and *in vivo* self-renewal of multipotent mammalian neural crest stem cells. *Cell* 96: 737–749.

Nagarajan R, Svaren J, Le N, et al. (2001) EGR2 mutations in inherited neuropathies dominant-negatively inhibit myelin gene expression. *Neuron* 30: 355–368.

Paratore C, Goerich DE, Suter U, Wegner M, and Sommer L (2001) Survival and fate acquisition of neural crest cells are regulated by an interplay between the transcription factor Sox10 and extrinsic combinatorial signaling. *Development* 128: 3949–3961.

Parkinson DB, Bhaskaran A, Droggiti A, et al. (2004) Krox-20 inhibits Jun-NH_2-terminal kinase/c-Jun to control Schwann cell proliferation and death. *Journal of Cell Biology* 164: 385–394.

Sagane K, Hayakawa K, Kai J, et al. (2005) Ataxia and peripheral nerve hypomyelination in ADAM22-deficient mice. *BMC Neuroscience* 6: 33.

Taveggia C, Zanazzi G, Petrylak A, et al. (2005) Neuregulin-1 type III determines the ensheathment fate of axons. *Neuron* 47: 681–694.

Topilko P and Meijer DM (2001) Transcription factors that control Schwann cell development and myelination. In: Jessen KR and Richardson WD (eds.) *Glial Cell Development,* 2nd edn., pp. 223–244. Oxford: Oxford University Press.

Vermeren M, Maro GS, Bron R, et al. (2003) Integrity of developing spinal motor columns is regulated by neural crest derivatives at motor exit points. *Neuron* 37: 403–415.

Wolpowitz D, Mason TB, Dietrich P, et al. (2000) Cysteine-rich domain isoforms of the neuregulin-1 gene are required for maintenance of peripheral synapses. *Neuron* 25: 79–91.

Yu W-M, Feltri L, Wrabetz L, Strickland S, and Chen Z-L (2005) Schwann cell-specific ablation of laminin $\gamma 1$ causes apoptosis and prevents proliferation. *Journal of Neuroscience* 25: 4463–4472.

Schwann Cell Morphology

A A Lavdas and R Matsas, Hellenic Pasteur Institute, Athens, Greece

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Along the entire length of mammalian peripheral nerves, axons of motor, sensory, and autonomic neurons are in close contact with Schwann cells (SCs), the main neuroglial cells of the peripheral nervous system (PNS). They were named after the German physiologist Theodor Schwann (1810–82), who founded modern histology by defining the cell as the basic unit of animal tissue structure, when he extended his cell theory of plant structure in animals with his 1839 publication, *Microscopic Researches into Accordance in the Structure and Growth of Animals and Plants*.

SCs are in intimate association and dynamic communication with neuronal axons thus influencing and regulating the development, maintenance, and function of peripheral nerves. They also play essential roles in the regeneration of peripheral nerves after injury and are capable of sustaining re-myelination of central nervous system (CNS) axons. They are therefore promising candidates for the development of transplantation therapies for repair of CNS lesions. Two major SC types are encountered in the adult nerves: the myelinating and the nonmyelinating SCs. Peripheral glia include more cell types than myelinating and nonmyelinating SCs. Perisynaptic glial cells or teloglia, found in the neuromuscular junction, as well as satellite cells of the peripheral ganglia are classified by some investigators as SCs. However, the consensus is that these cells are classified as non-SC peripheral glia and therefore they are not discussed here.

Myelinating SCs provide the myelin sheath of peripheral axons which is necessary for proper nerve function. The nervous system, both central and peripheral, relies on the myelin sheath for insulation and for decreasing membrane capacitance in the axon, thus allowing for rapid saltatory conduction and increased impulse speed to occur. Unlike oligodendrocytes, the myelinating cells of the CNS, each SC forms myelin around one axon only. For reasons that are not understood, only the larger diameter (typically >1 μm) axons of peripheral nerves are myelinated. A myelinated fiber is characterized by myelin internodes, which comprise the length of the myelin sheath generated by one SC, and short myelin-free areas, the nodes of Ranvier (Figure 1(a)). Myelinating SCs (Figure 2) are ultrastructurally distinct from their CNS analogs. The nucleus stains intensely, is flattened, and is oriented longitudinally along the nerve fiber. The cytoplasm of the SC is rich in organelles. A Golgi apparatus is located near the nucleus, and cisternae of rough endoplasmic reticulum as well as round mitochondria occur throughout the cell. Lysosomes, glycogen granules, multivesicular bodies, filaments, microtubules, and lipid granules can also be seen. Aggregates of dense heterochromatin are arranged peripherally.

Smaller axons (<1 μm) lie in troughs in the surface of nonmyelinating SCs (Figure 1(b)). The function of these cells is less well understood, but studies indicate that they are involved in the maintenance and function of unmyelinated axons and possibly nociception in the adult. The two types of SCs differ not only in their function, the type of axons with which they associate and their morphology, but also in molecular composition. Notably, myelinating SCs contain a set of myelin-associated proteins that are not expressed by nonmyelinating SCs.

Origins and Development of Schwann Cells

The two types of SCs, myelinating and nonmyelinating, found in adult nerves are derived from a single precursor cell type originating from the neural crest. This cell type matures in a stepwise process, giving rise to all SCs. First, SC precursors are generated by neural crest cells (at embryonic day (E) 14 or E15 in the rat) and associate with growing peripheral axon bundles, guiding their growth (Figure 3). In the limbs, these precursors are connected by adherens junctions and form elaborate scaffolds that serve as substrates for the developing growth cones as they advance to their targets. Expression of the adhesion molecule N-cadherin on SC precursors appears to be essential for their role as an axon outgrowth substrate.

SC precursors convert to immature SCs between E15 and E17 in the rat (Figure 3). This irreversible transition entails a phenotypic change in molecular expression and an altered response to survival signals and mitogens. It is also accompanied by a marked change in the cellular architecture of peripheral nerves. At E14 in rats, the nerves are compact structures consisting of axons and SC precursors (Figure 4). There is no significant connective tissue space and blood vessels are absent. SC precursors are found at the outer margin of the nerves, and inside them as well,

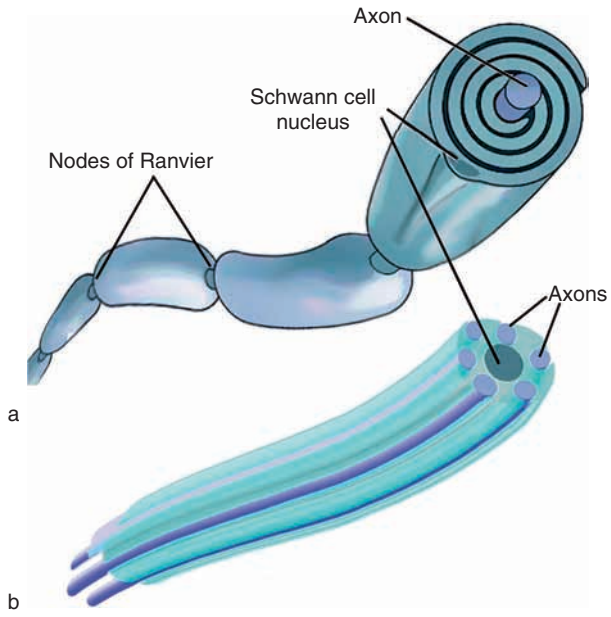

Figure 1 Schematic representation of peripheral myelinated (a) and unmyelinated (b) axons.

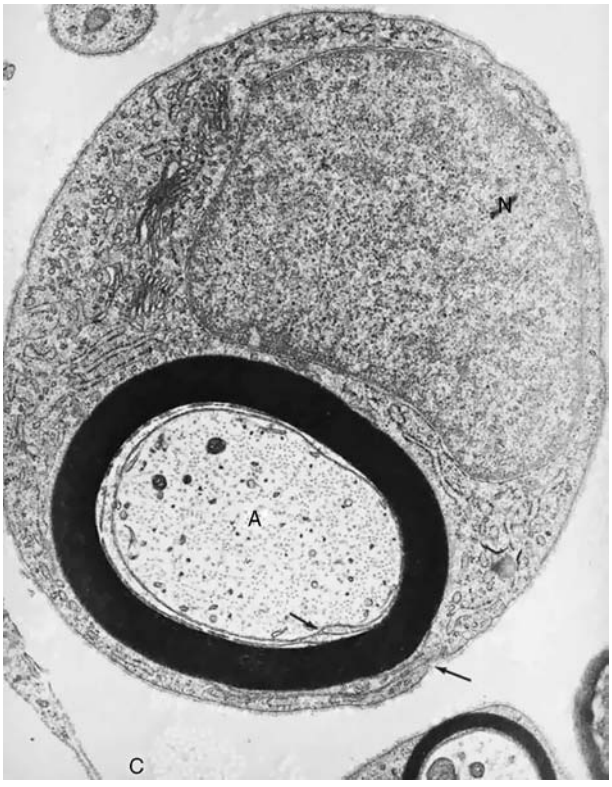

Figure 2 A large PNS axon (A) is surrounded by a myelinating Schwann cell. The SC nucleus is marked (N). Note the fuzzy basal lamina around the cell, the rich cytoplasm, the inner and outer mesaxons (arrows), the close proximity of the cell to its myelin sheath, and the 1:1 axon:Schwann cell relationship. Unstained collagen (C) lies in the endoneurial space. Reproduced from *Basic Neurochemistry*, 6th edn., with permission.

connecting to each other through sheet-like processes that communally envelope large numbers of axons. By E18, nerves are made up of irregular axon-immature SC bundles (similar to the 'families' found in newborn nerves) surrounded by connective tissue containing endoneurial fibroblasts and blood vessels (Figure 4). This basic architecture – comprising nervous tissue, connective tissue, and blood vessels – is retained in adult nerves.

On E18 in the rat, immature Schwann cells still ensheathe communally large groups of axons (Figure 3). However, for myelination to start soon after birth (approximately 3 days later), a process of radial sorting is required. During this process, which continues postnatally, the SC families of late embryonic nerves gradually become segregated into individual SCs, each associated with a single large-diameter axon. At the same time, SCs undergo a process of lateral elongation along the axon and secrete a basal lamina made up of mucopolysaccharides at their abaxonal surfaces. This radical change in cellular relationships, which results in a 1:1 relationship between axons and SCs, marks the promyelinating stage, is accompanied by changes in the molecular phenotype of the SCs, and renders them poised for myelination. Specific transcription factors, notably Krox 20, are upregulated and act to control myelination. Axon–SC contact is necessary for maintenance of the myelinating phenotype. Upon loss of axonal contact, as in the case of nerve injury, myelinating SCs revert to a nonmyelinating phenotype similar to that of immature SCs (Figure 3).

During embryonic development, as well as postnatally, numerous factors mediate specific aspects of the interactions between peripheral axons and SCs. Compelling evidence suggests that neuregulin-1 and its Erb receptors, belonging to the epidermal growth factor family of protein kinases, are key regulators of axon–SC interactions at every stage and for all SC types. The neuregulin/Erb signaling system is thus critical for SC biology and influences SC survival, proliferation, migration, and differentiation.

Myelinating Schwann Cells and the Myelin Sheath

At the promyelinating stage, as SCs start to wrap around the axon, the distance between two consecutive spirals (the extracellular gap) is 12–14 nm. At the myelinating stage, exclusion of the cytoplasm and compaction of the cytoplasmic surfaces results in the formation of compact myelin (Figure 2). Then the extracellular space is reduced from 12–14 to 2 nm, which is characteristic of compact myelin. Each SC is surrounded by the basal lamina it has secreted, which

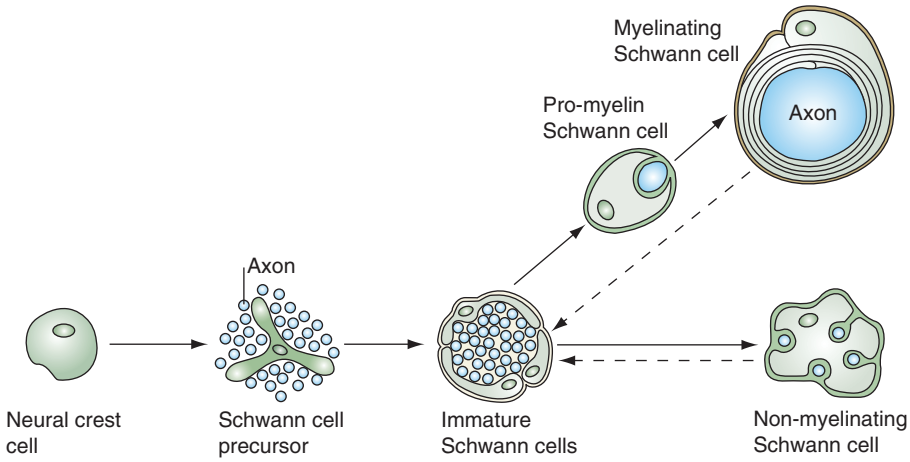

Figure 3 Schematic representation of the major stages of Schwann cell development. Dashed arrows represent reversible processes. Adapted from Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682, with permission.

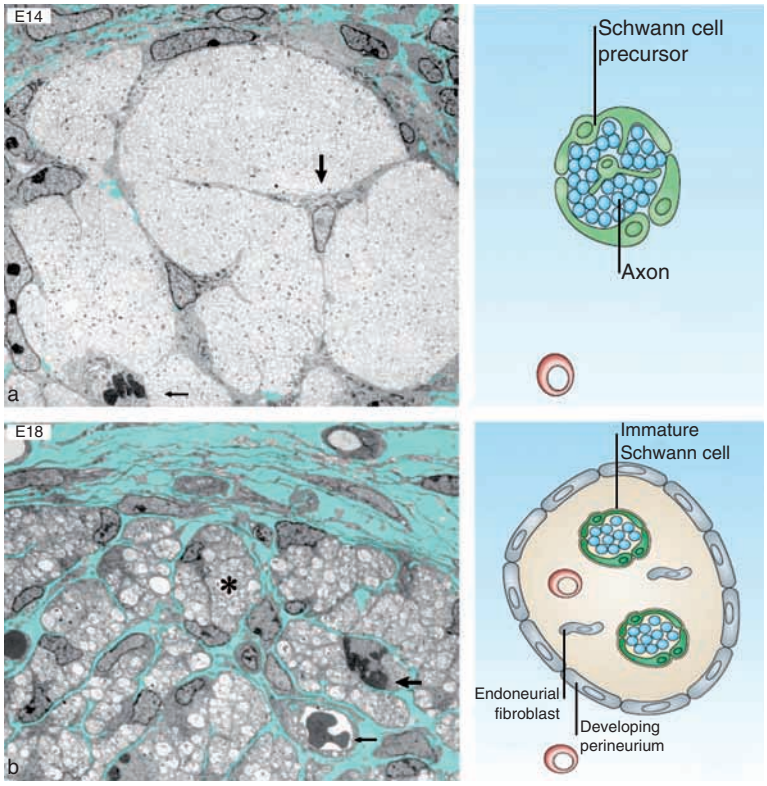

Figure 4 Electron microscopic images (pseudo-color enhanced) and corresponding schematic illustrations of rat embryo peripheral nerves at E14 and E18, constituting hallmarks in the cellular architecture of peripheral nerves. At E14 (top left) the nerve is compact and is composed only of axons and SC precursors surrounding or interdigitating (large arrow) the axons. Division of an SC precursor is visible (small arrow). Extracellular connective tissue (in turquoise) and blood vessels are absent from the nerve. In sharp contrast, by E18 the nerve has been infiltrated by connective tissue and is now segregated into compact bundles of immature SCs and their associated axons ('SC families'). One such 'family' is depicted by the asterisk. Blood vessels are now also present (small arrow), whereas SC divisions can still be seen (large arrow). Reproduced from Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682, with permission.

is approximately 20–30 nm thick and does not extend into the mesaxon. Myelination takes place during the first 2 or 3 weeks after birth in the rat. The mechanism by which SCs ensheathe the axon has been a subject of intense speculation in the past. In the PNS, the myelin spiral around a single axon may consist of up to 100 layers, a fact that has rendered the old 'jelly-roll' hypothesis, proposing that myelin

is laid down by a simple rotation of the Schwann cell nucleus around the axon, rather improbable (Figure 5(a)). In the CNS, such a hypothesis is precluded by the mere fact that a single glial cell can invest several axons with myelin. The prevailing hypothesis today is that ensheathment of the axon occurs by active progression of the inner (axonal) lip of the myelin lamella as it forms the myelin spiral (Figure 5(b)). As this movement progresses, the SC membrane on one side of the groove (in which the axon lies) comes into contact with the cell membrane on the other side of the groove. These parts of the membrane approximate and stay together and thus are seen, as the cell membrane lamella continues to wind around the axon, as a spiral of rings each made of double lines. Between adjacent double rings there is at first a layer of cytoplasm, but as the winding continues, the cytoplasm is squeezed back into the cell body.

The wrapping of the SCs around a short segment of the axon produces the myelin sheath, a lipid-rich layer formed by a greatly extended and modified plasma membrane (Figure 2). This membrane, since it is highly enriched in nonconductive ingredients and poor in conductive water-based media (cytoplasm), serves ideally as an effective insulant for the nerve fiber. The lipid component of the myelin sheath gives nerves a vacuolated appearance in histological stains. In myelinated axons, only the axon hillock and the terminal arborizations where the axon synapses with its target cells are completely devoid of the myelin sheath. Myelin and many of its morphological specializations such as the nodes of Ranvier and the Schmidt–Lantermann incisures, can be detected readily under the light microscope; however, much of our understanding of its organization is derived from studies carried out by other techniques offering higher resolution: polarized light, X-ray diffraction, and electron microscopy.

Along the myelinated fibers of the PNS, each internode of myelin is invested by one Schwann cell and each Schwann cell surrounds one internode (Figure 1). As discussed previously, this ratio of one internode of myelin to one Schwann cell represents a fundamental distinction between the SC and the oligodendrocyte, which is able to produce internodes at a ratio of 1:30 or greater. Another distinction is that the Schwann cell body always remains in close contact with its myelin internode, whereas the oligodendrocyte is not on an internode but extends processes toward them.

Under the electron microscope, myelin is visualized as a series of alternating dark and less-dark lines (protein layers) separated by unstained zones (the lipid hydrocarbon chains). This distinctive pattern of staining results from the way the myelin sheath is generated from the cell plasma membrane. The lighter (intraperiod) line represents the closely apposed protein coats of the outer aspect of the cell membrane; the darker (major dense) line corresponds to the fused inner protein coats of the membrane, formed as the cytoplasm has been squeezed out

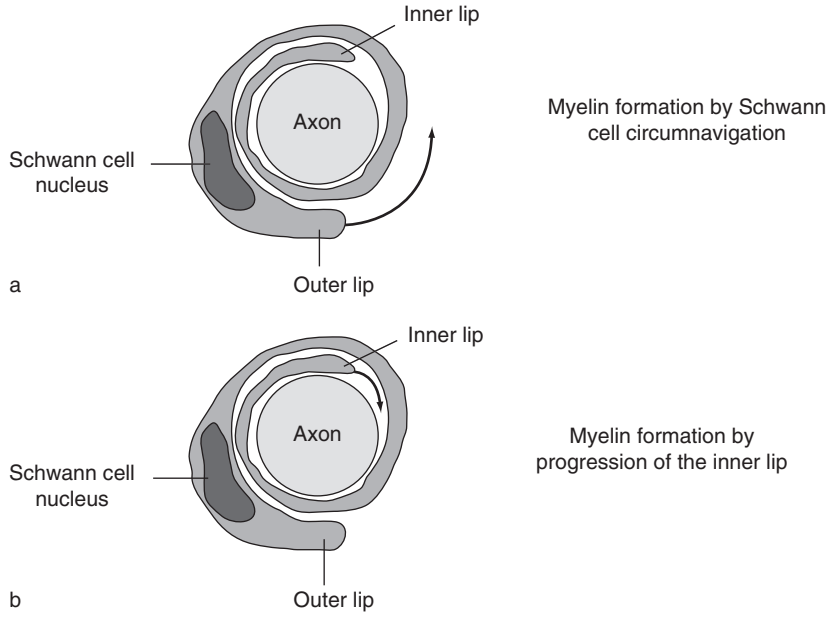

Figure 5 Myelin formation by the Schwann cell according to older (a) and current (b) hypotheses. The arrow indicates the direction of movement of the growing membrane. Reproduced from Garbay B, Heape AM, Sargueil F, and Cassagne C (2000) Myelin synthesis in the peripheral nervous system. *Progress in Neurobiology* 61(3): 267–304, with permission from Elsevier.

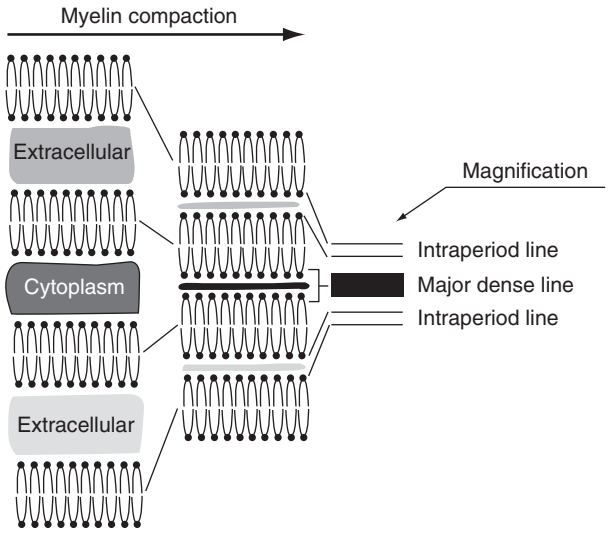

Figure 6 Schematic representation of a longitudinal section of a myelinated PNS axon. The left part is highly magnified on the right to illustrate the formation of the major dense and intraperiod lines. Reproduced from Garbay B, Heape AM, Sargueil F, and Cassagne C (2000) Myelin synthesis in the peripheral nervous system. *Progress in Neurobiology* 61(3): 267–304, with permission from Elsevier.

(Figure 6). Examination of PNS myelin swollen in hypotonic solutions confirms this interpretation. Electron microscopy observation of such preparations reveals splitting at the intraperiod line, showing the continuity of its membrane junction with the extracellular space. With good perfusion fixation and tissue staining the intraperiod line is always visualized as a double line, indicating that the extracellular sides of the unit membranes are closely apposed but not fused, as was thought in the past. When studied in fixed preparations, PNS myelin has an average repeat distance of 11.9 nm (approximately 30% less than in the unfixed state), in contrast to 10.6 nm for central myelin. External to the myelin lies the greater part of the SC cytoplasm, bounded at its outer surface by the cell membrane and basal lamina of the Schwann cell (Figure 2).

The myelin sheath can be divided, by virtue of its composition and morphology, into two distinct domains – compact myelin and noncompact myelin – each containing distinct proteins. Compact myelin contains P0, the 22 kDa peripheral myelin protein PMP-22, and myelin basic protein (MBP). MBP which is positively charged is considered important in linking the two cytoplasmic surfaces, whereas P0 is crucial for stabilizing the intraperiod line by keeping a 2-nm distance between the apposing extracellular faces of two adjacent compacted myelin membranes. Noncompact myelin, found at the paranodal regions flanking the nodes of Ranvier and in periodic interruptions of the compact myelin called Schmidt–Lanterman incisures, contains myelin-associated glycoprotein (MAG), the gap junction protein connexin32 (Cx32), and E-cadherin. E-cadherin is localized to adherens junctions, whereas the localization of Cx32 coincides with the location of gap junction-like strctures. Adherens junctions and gap junctions are usually found between neighboring cells and their presence in the myelin sheath is atypical, as in this case they join adjacent layers of the same cell. Specific neuron–SC interactions that play an important role in the organization and function of peripheral nerves take place in the regions of noncompact myelin.

Noncompact Myelin

Nodes of Ranvier

Myelin is not a continuous structure but is interrupted at regular intervals by the nodes of Ranvier, constrictions named after the French pathologist and anatomist Louis-Antoine Ranvier (1835–1922), who first described them (Figure 7). These short periodical interruptions in the myelin sheath are regularly spaced along the axon at intervals of ~100 times the axonal diameter. The area of the nerve between two consecutive nodes of Ranvier, the internode, is ensheathed by myelin produced by a single SC. The nodal region, composed of both SCs and the axon, is a highly organized structure containing several distinct domains: the nodes, paranodes, and juxtaparanodal region, each characterized by an array of interacting cell adhesion molecules (CAMs), cytoskeletal proteins, and voltage-gated ion channels (Figures 8 and 9). This highly specialized structure, in terms of both morphology and molecular composition, is essential for saltatory conduction.

At the node there is no myelin. The nodal axolemma is characterized by a high density (>1200 mm^{-2}) of voltage-gated Na^+ channels (Nav 1.2 during development or Nav 1.6 in the adult) essential for the generation of the action potential during salutatory conduction (Figure 9). Na^+ channels co-localize with KCNQ2 K^+ channels, cell adhesion molecules such as NrCAM and neurofascin 186 (NF186), and cytoskeletal proteins, including the cytoskeletal-linking protein ankyring G and the actin-binding protein βIV-spectrin (Figures 8(b) and 9). These molecules as well as gliomedin, the glial ligand for NrCAM and NF186, interact to regulate channel clustering at the node. The nodal axolemma is in contact with microvilli emanating from the outermost regions of myelinating SCs. Several other morphogenetic proteins are present, contributing to the formation of the microvilli and the particular architecture of this specialized structure (Figure 9).

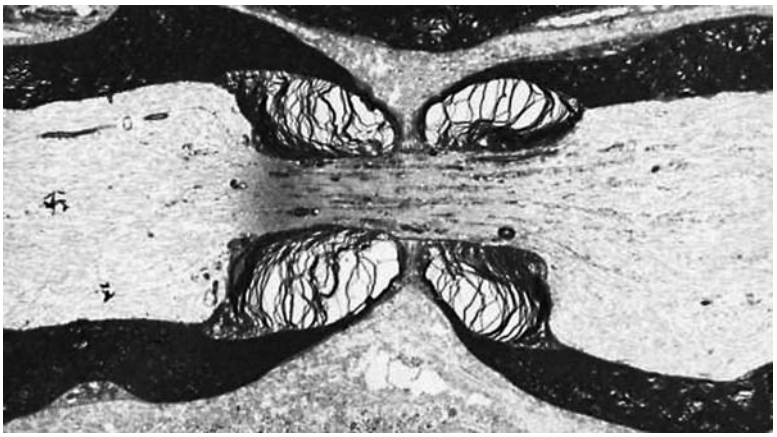

Figure 7 Low-power electron micrograph of a node of Ranvier in longitudinal section. Note the abrupt decrease in axon diameter and the attendant condensation of axoplasmic constituents in the paranodal and nodal regions of the axon. Paranodal myelin is distorted artifactually, a common phenomenon in large-diameter fibers. Reproduced from *Basic Neurochemistry*, 6th edn., with permission.

Just outside the center of the node lies the area of the paranode, in which the compact myelin lamellae open up into a series of cytoplasmic loops that spiral around the axon (**Figure 8(a)**). The length of the paranode is proportional to the axon diameter, ranging from 10–20 μm in smaller fibers (4 or 5 μm in diameter) to 80 μm in large fibers more than 15 μm in diameter. There, the cell adhesion molecules contactin (F3), paranodin (contactin-associated protein; Caspr) on the neuronal side and Neurofascin-155 (NF-155) on the glial side interact to stabilize the structure (**Figure 9**). The paranodal loops of SC are tightly attached to the axolemma, leaving an extracellular space of approximately 3 nm, in which the two membranes are joined by electron-dense structures – the paranodal or septate-like junctions. During development, these junctions are first generated next to the node and are formed gradually, as additional loops are attached to the axon in a spiral fashion. The paranodal junctions are thought to provide attachment of the myelin sheath to the axon, to electrically insulate the node from the internodal area, and to act as boundaries that limit the lateral diffusion of membrane components. Next to the paranode and distal to the node lies the juxtaparanodal region, characterized by the presence of K^+ channels Kv1.1 and Kv1.2 and their Kvβ2 subunit (**Figures 8(a) and 8(b)**). The adhesion molecule TAG1, which is expressed by both the axon and the SC at the juxtaparanode, as well as Caspr2, are crucial for the sequestration of K^+ channels (**Figure 9**).

There is one SC between every two nodes and fiber branching always occurs at nodes. The basal laminae of myelinating SCs at the nodes of Ranvier are continuous, and SC processes interdigitate so that the PNS myelinated axon never comes in direct contact with the extracellular space.

Schmidt–Lantermann Incisures, Lateral Loops, and Transverse Bands

The Schmidt–Lantermann incisures or clefts, named after the American pathologist Henry D. Schmidt (1823–88) and American anatomist AJ Lantermann (1855–1910), are recurring structures in peripheral myelinated axons. They represent regions where the cytoplasmic surfaces of the myelin sheath have not compacted to form the major dense line and thus contain SC cytoplasm (**Figures 8(a) and 8(c)**). These inclusions are present in each layer running spirally within the myelin, and they would therefore appear as tubes of cytoplasm in an unrolled myelin sheet. They transect the sheath at an angle of approximately 9°. Ultrastructurally, there is considerable variation in their number and width, which are related to the axon thickness; the density of the cytoplasm they contain, which is inversely related to their width; and the morphology of the rings that form their boundaries. Cx32 has been localized at the incisures, and the presence of gap junctions, along with adherens junctions and tight junctions, has been observed by freeze-fracture electron microscopy.

Similar tubes, in which the major dense line opens up to enclose cytoplasm, are located at the paranodal edges of the sheet and make up the lateral loops (**Figure 8(c)**). The lateral loops also contain Cx32. The presence of this gap junction protein in all uncompacted myelin structures and the electron microscopic observation of gap junction channels suggest that they act as routes of communication between turns of myelin of an individual SC for the transportation of electrolytes and metabolites, providing a path from the periaxonal cytoplasm all the way to the outermost part of the SC. The lateral loops form membrane complexes with the axonal

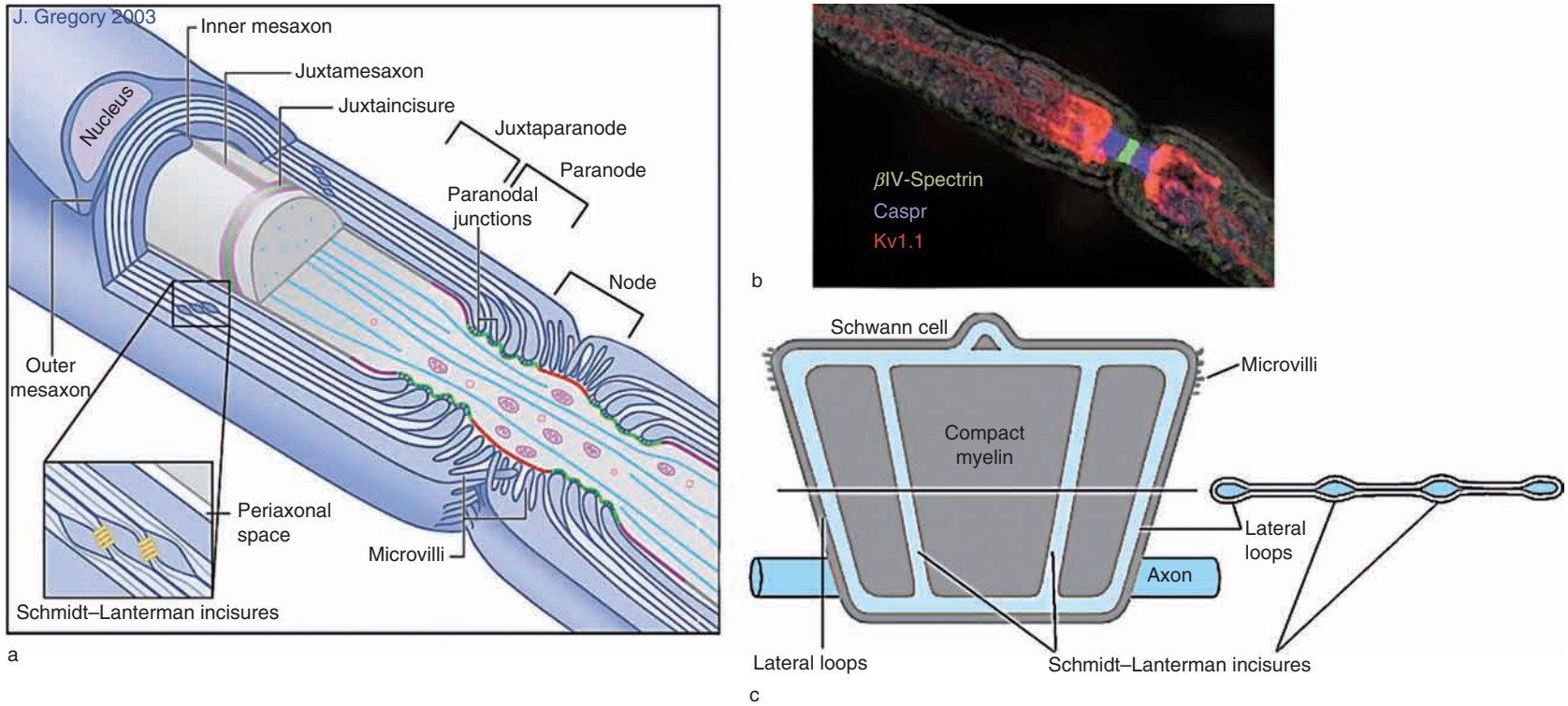

Figure 8 (a) Schematic illustration of a myelinated axon sectioned at the area of the node of Ranvier. Shown are the SC microvilli at the node as well as the SC lateral loops at the paranode. The lateral loops are kept in contact with the axon via specific protein–protein interactions (see also **Figure 9**). (b) Confocal immunofluorescence image of a myelinated sciatic nerve fiber illustrating the three distinct axonal domains in the nodal area: βIV-spectrin is localized at the node (green), Caspr at the paranode (blue), and Kv1.1 K^+ channels at the juxtaparanode (red). (c) 'Open book' view of the myelin sheath of an internode. (a, b) Reproduced from Salzer (2003), with permission. (c) Adapted from *Basic Neurochemistry*, 6th edn.

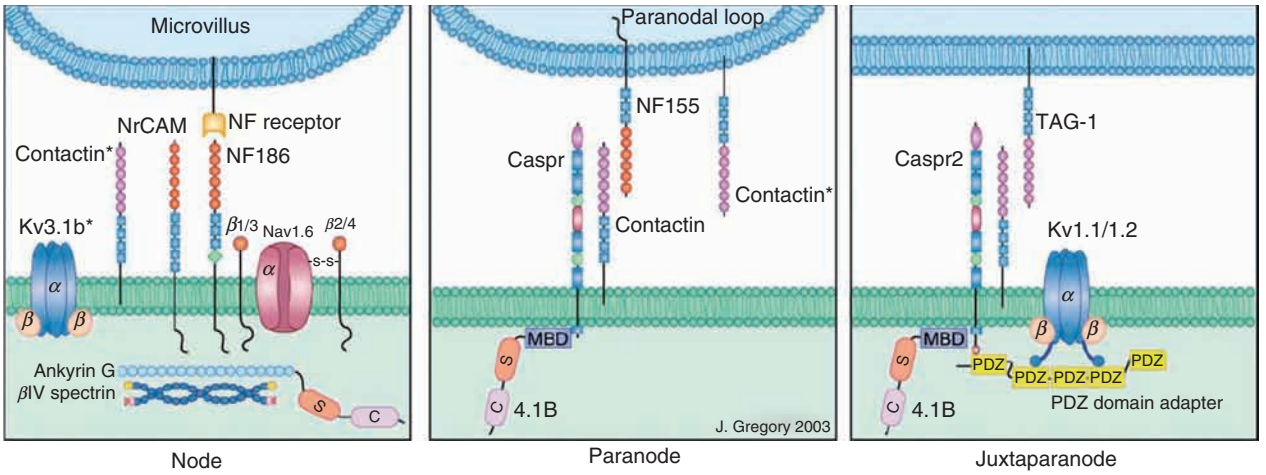

Figure 9 Protein–protein interactions between the axon and the Schwann cell at the node, paranode, and juxtaparanode of a myelinated fiber. Morphogenetic proteins in the axonal domains and their glial receptors are illustrated. Reproduced from Salzer (2003), with permission.

membrane called transverse bands. There are six to eight transverse bands per paranodal loop, spaced 25–30 nm apart. The transverse bands seal the myelin to the axolemma but provide, by the spaces between them, a path from the extracellular space to the periaxonal space, which separates myelin from the axon in the internodal area.

Nonmyelinating Schwann Cells

Some afferent and autonomic nerve fibers lack myelin sheath. These are the unmyelinated fibers and are smaller, since typically the diameter of an axon has to be larger than 1 μm for myelination to commence. After birth in the rodent, SCs diverge in myelinating and nonmyelinating cells. In contrast to oligodendrocytes, SCs surround both myelinated and unmyelinated axons. Although they are not ensheathed in a myelin structure, unmyelinated axons are nevertheless protected by SCs which are arranged in bundles (Remak bundles) so that they form a groove through which the fibers extend. In this manner, approximately 5–20 unmyelinated fibers lie protected by an SC at any given level. The protective cellular sheath enclosing a group of unmyelinated fibers thus comprises a series of interdigitating SCs arranged end to end along the bundle. Surrounding the external surface of SCs and covering the grooves where they open to the exterior is a basal lamina. Between the axolemma of each unmyelinated fiber and the SC membrane lining the housing groove is an intercellular space 10–15 nm wide. No nodal structures are present, and voltage-gated ion channels are considerably less ordered in the unmyelinated axon, spreading over the entire membrane surface. Nonmyelinating SCs (as well as SCs in culture) express a distinct and characteristic set of marker proteins, including the cytoskeletal glial fibrillary acidic protein (GFAP), the calcium-binding protein S-100, and p75 neurotrophin receptor (Figure 10).

There used to be a misconception that nonmyelinating SCs are there simply to take over myelination whenever the need arises. However, it has been shown that these cells play critical roles in neuronal survival, function, and axon maintenance while the axons support SC survival. As in the case of myelinating SCs, these reciprocal interactions are at least in part mediated via the neuregulin-1/ErbB signaling system.

Schwann Cells in Axonal Regeneration

In normal circumstances, the neuronal cell body maintains the axon through the axoplasmic flow. When even a small segment of an axon dies, the part of the axon distal to the injury undergoes a series of changes leading to its complete structural disintegration and chemical degradation. When myelinated axons distal to a site of injury die, their myelin sheaths also degenerate. This process is termed Wallerian degeneration because it was first described by neurophysiologist Augustus Volney Waller (1816–70). SCs carry out the first step in the removal of myelin by segmenting and then incorporating the degraded myelin. Recruited macrophages then join in. In the neuronal body, the rough endoplasmic reticulum disaggregates and the neuronal body swells. The cytoplasm becomes smooth as the organization of polyribosomes into Nissl bodies is temporarily disrupted and the nucleus is displaced toward the periphery of the cell. This process is called central

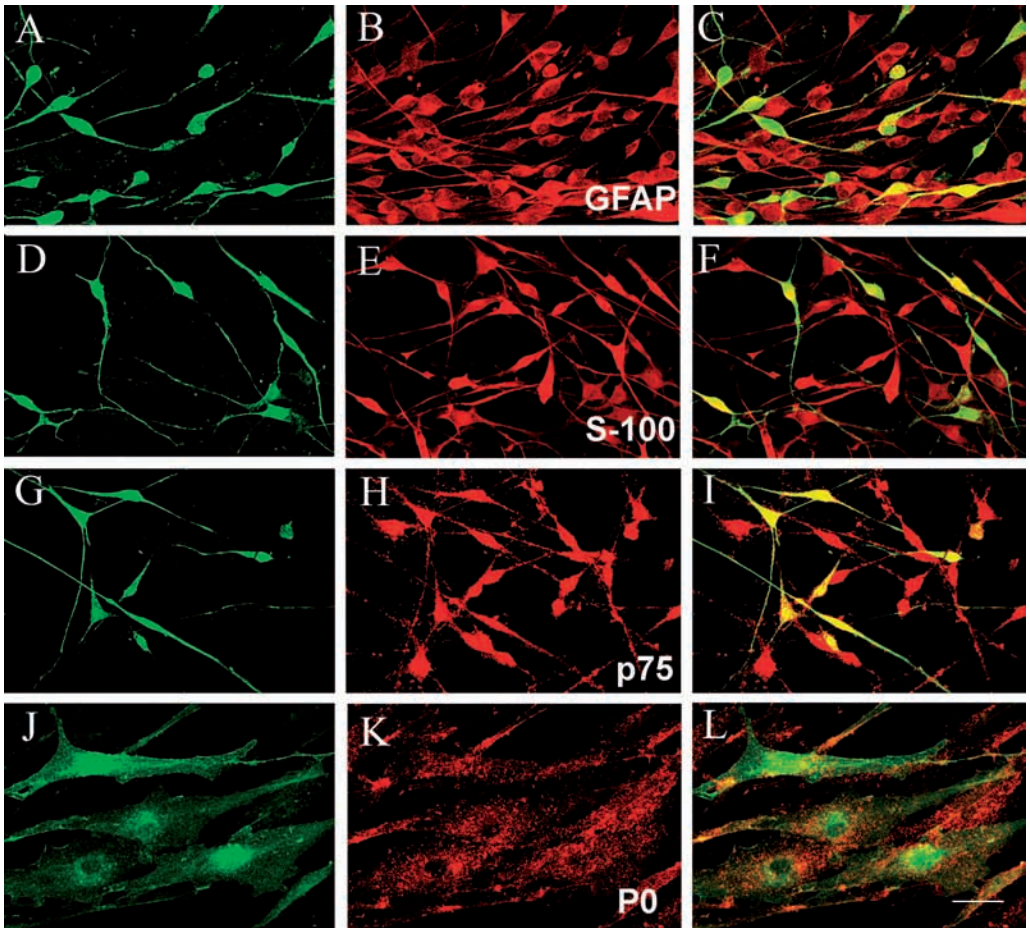

Figure 10 Confocal microscopy of SCs cultured from neonatal rat sciatic nerve. Cells were virally transduced to express, in this case, the green fluorescent protein, GFP (green, left column) and were immunostained for the characteristic SC markers *in vitro* (red, middle column): GFAP (A–C), S-100 protein (D–F), or p75 nerve growth factor receptor (G–I). Merged GFP/immunofluorescence images are shown in the right column. A fraction of SCs identified by the expression of marker proteins were transduced to express GFP (yellow, right column). Upon appropriate stimuli, these cells were induced toward a myelinating phenotype (J–L), as illustrated by their flattened appearance, production of new membrane expansions, and the expression of P0 peripheral myelin protein (red, K). Reproduced from Kenoutis et al. (2006), with permission.

chromatolysis and it reflects cytoskeletal reorganization and activation of protein synthesis in order to regenerate the axon. Cytoskeletal proteins and other materials flow down the axon and the proximal stump elongates at a rate of 1–3 mm per day. These changes begin within 24–48 h and reach a peak at approximately 2 weeks. SCs after myelin removal transiently revert from a myelinating to a nonmyelinating phenotype, and they start to divide and migrate forming longitudinally continuous columns (Bungner bands) between the severed ends. In this manner, they form tubes containing SC cell bodies and processes surrounded by a single basal lamina, thereby providing channels for growth cone navigation and axonal elongation. The presence and integrity of the SC basal lamina is essential for reinnervation. Wallerian degeneration was initially described in experimental axotomy. Neuropathies characterized by Wallerian degeneration include those that are caused by trauma, infarction of peripheral nerve (diabetic mononeuropathy and vasculitis), and neoplastic infiltration. SCs can myelinate and support the regeneration of not only peripheral but also central axons, both *in vitro* and *in vivo*. This beneficial SC property is being explored for the development of transplantation therapies for CNS repair after injury or in demyelinating diseases.

Schwann Cells *In Vitro*

Schwann cells can be cultured from neonatal peripheral nerves, after mechanical dissociation and enzymatic digestion followed by elimination of rapidly dividing residual fibroblasts with an antimitotic

agent (Figure 10). They can be maintained for weeks using proliferating media containing neuregulin-1 and a cAMP-elevating agent such as forskolin. In these culture conditions, SCs are nonmyelinating and appear mostly bipolar with a spindle-like shape, but they can also be tripolar. Cultured SCs have a molecular phenotype similar to that of nonmyelinating SCs (Figures 10(B), 10(E), and 10(H)). SCs can also be cultured from adult nerves, including human.

When culture conditions are switched from proliferating to myelinating by withdrawal of mitogens and exposure to higher concentrations of agents that elevate or mimic intracellular cAMP (forskolin, cholera toxin, and cAMP analogs), the morphological and biochemical phenotype of SCs changes dramatically, as it does *in vivo* at the onset of myelination. GFAP expression is downregulated and myelin protein expression is upregulated. SCs lose their bipolar appearance and become flattened as they produce large expansions of membrane in preparation for axonal ensheathment (Figures 10(J)–10(L)).

When in culture, SCs are amenable to genetic manipulation through viral transduction for the expression of exogenous proteins (Figure 10) with potential beneficial effects for nervous tissue regeneration. Such genetically modified SCs may be transplanted in the lesioned CNS to assist remyelination of central axons. This combination of gene and cell therapy approaches is currently attracting significant attention for CNS repair in demyelinating diseases or trauma.

See also: Axon Guidance by Glia; Glial Responses to Injury; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation; Neurotransmitter and Hormone Receptors on Oligodendrocytes and Schwann Cells; Oligodendrocyte and Schwann Cell Identification Methods; Peripheral Nerve Regeneration: An Overview; Schwann Cells and Axon Relationship; Schwann Cell Development; Schwann Cells and Plasticity of the Neuromuscular Junction; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Transplantation of Myelin Forming Cells; Wallerian Degeneration.

Further Reading

Bunge RP, Bunge MB, and Bates M (1989) Movements of the Schwann cell nucleus implicate progression of the inner (axon-related) Schwann cell process during myelination. *Journal of Cell Biology* 109(1): 273–284.

Chen S, Rio C, Ji R, et al. (2003) Loss of erbB receptor signaling in adult non-myelinating Schwann cells results in a progressive sensory neuropathy with loss of temperature sensitivity. *Nature Neuroscience* 6: 1186–1193.

Corfas G, Velardez MO, Ko CP, Ratner N, and Peles E (2004) Mechanisms and roles of axon–Schwann cell interactions. *Journal of Neuroscience* 24(42): 9250–9260.

Garbay B, Heape AM, Sargueil F, and Cassagne C (2000) Myelin synthesis in the peripheral nervous system. *Progress in Neurobiology* 61(3): 267–304.

Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

Komada M and Soriano P (2001) βIV-spectrin regulates sodium channel clustering through ankyrin-G at axon initial segments and nodes of Ranvier. *Journal of Cell Biology* 156(2): 337–348.

McIlwain DL and Hoke VB (2005) The role of the cytoskeleton in cell body enlargement, increased nuclear eccentricity and chromatolysis in axotomized spinal motor neurons. *BMC Neuroscience* 6(1): 19.

Mirsky R and Jessen KR (1999) The neurobiology of Schwann cells. *Brain Pathology* 9: 293–311.

Mirsky R and Jessen KR (2001) Embryonic and early postnatal development of Schwann cells. In: Jessen KR and Richardson WD (eds.) *Glial Cell Development: Basic Principles and Clinical Relevance*, 2nd edn., pp. 1–20. London: Oxford University Press.

Morell P and Quarles RH (1999) Myelin formation, structure and biochemistry. In: Siegel GJ (ed.) *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*. Philadelphia: Lippincott Williams & Wilkins.

Palay SL, Peters A, and Webster H (1991) *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*. New York: Oxford University Press.

Quarles RH (2002) Myelin sheaths: Glycoproteins involved in their formation, maintenance and degeneration. *Cellular and Molecular Life Sciences* 59(11): 1851–1871.

Raine CS (1982) Differences in the nodes of Ranvier of large and small diameter fibres in the PNS. *Journal of Neurocytology* 11: 935–947.

Raine SC (1999) Neurocellular anatomy. In: Siegel GJ (ed.) *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*. Philadelphia: Lippincott Williams & Wilkins.

Wanner IB, Mahoney J, Jessen KR, Wood PM, Bates M, and Bunge MB (2006) Invariant mantling of growth cones by Schwann cell precursors characterize growing peripheral nerve fronts. *Glia* 54(5): 424–438.

Schwann Cells and Axon Relationship

R D Fields, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

Vertebrates owe their evolutionary success to myelin. This multilayered wrap of electrical insulation speeds the conduction of impulses through slender axons. Without myelin, axons would need to be much thicker to carry impulses at the same rate. Evolution of myelin allowed miniaturization of nerve fibers and thus compaction of the nervous system into a brain of incomparable complexity – accelerating information processing in the nervous system of animals with backbones by magnitudes beyond the cerebral capabilities of the most evolutionarily advanced invertebrates. On a cellular level, myelin is perhaps the most intricate of all interactions between cells. Myelination requires cell–cell recognition, intercellular signaling, adhesion, motility, coordinated regulation of cell proliferation, differentiation, and synthesis of enormous sheets of insulating membrane, compacted into thin spiral layers in a manner unlike any other and with fine structural intricacies at the nodes of Ranvier to a molecular, crystalline level of precision, discernable only with the electron microscope. Myelin is unlike any other cell–cell interaction.

The cells that form myelin in the nerves of our limbs and trunk (peripheral nerves) are Schwann cells. In the brain, an entirely different type of cell, oligodendrocyte, with a distinct embryological origin, forms myelin. Schwann cells were first described in 1838 by the same pioneering biologist who conceived the concept of the cell as the basic building block of all life, Theodore Schwann. However, the purpose and origin of these peculiar cells were a mystery to him. Schwann cells coat axons like flattened pearls on a string, inspiring Schwann to speculate that they might form the nerve axon during fetal development by aligning and coalescing into a tube. In retrospect, this quaint idea appears naive, but the intimate interdependence of both axon and Schwann cell is now deeply appreciated.

Communication between Axons and Schwann Cells

Signals from axons are essential in regulating most Schwann cell functions, including proliferation, adhesion, migration, differentiation, and myelination. After an axon is damaged, Schwann cells respond to the loss of axon contact by de-differentiating. They lose their myelin and begin to proliferate and secrete growth factors to help guide the regenerating axon back to its target. The axon, in turn, is dependent on Schwann cells for its own integrity, structural development, and survival. In addition to providing structural support and electrical insulation, Schwann cells provide growth factors to axons, and they organize the arrangement of macromolecular ion channels into proper spatial domains along the axon membrane to support saltatatory conduction. Saltatory conduction describes the way an electrical impulse skips from node to node down the full length of an axon, speeding the arrival of the impulse at the nerve terminal in comparison with the slower continuous progression of depolarization spreading down an unmyelinated axon. Each node, situated between adjacent Schwann cells along the axon, is a bare segment of axon with specialized membrane properties for generating action potentials.

Embryological Origin of Schwann Cells

Schwann cells derive embryologically from the neural crest, which comprises multipotent cells migrating away from the dorsal neural tube. Neural crest cells differentiate into Schwann cell precursors, which migrate and proliferate along tracts of axons that have already extended into the periphery. The Schwann cell precursors then embark on a process of cellular differentiation into an immature Schwann cell to eventually become either a myelinating or a nonmyelinating Schwann cell. Only large-diameter axons, which conduct impulses at the highest speed, become myelinated. The slender, slow conducting fibers become bundled together and engulfed by massive, globular nonmyelinating Schwann cells.

Axon Signals Sustaining Schwann Cell Precursors

Early Schwann cell precursors are critically dependent on axon-derived signals for maintenance, proliferation, and differentiation. Axon contact stimulates proliferation of Schwann cell precursors, and neuregulin-1 (NRG-1) is recognized as one of the most potent axon-derived mitogens. Neuregulins are a family of growth factors that act as mitogens by binding the ErbB family of receptor kinases. The primary receptor for NRG-1 in Schwann cells is a heterodimer composed of ErbB2 and ErbB3, and mice genetically engineered to lack these receptors lose all Schwann cells on their peripheral axons. ErbB signaling is also

required to maintain the directed migration of Schwann cell precursors along axons. Genetic deletion experiments show that premigratory Schwann cells near the dorsal root ganglia fail to migrate out along the axons in the absence of neuregulin signaling.

The progenitors must proliferate as they migrate out along axons to produce sufficient numbers of cells to string the entire length of the axon, but precisely the correct number of cells is essential because this determines the number of nodes of Ranvier along the fiber. This is one of the important elements determining the conduction velocity and safety factor that will ensure the nodes are spaced closely enough for the wave of electrical depolarization to sustain a sufficient distance along the axon to activate the next node in sequence. As a limiting survival factor for Schwann cell precursors, NRG-1 helps establish the appropriate total number of Schwann cells on each axon. In a similar manner, transforming growth factor-β (TGF-β) also acts as a mitogen on Schwann cells at this stage to control the total number of cells.

Secreted factors inhibiting development are equally important to coordinate the timing of different development steps. Endothelins, AP2α, and bone morphogenic protein (BMP)-2 and -4 are examples of factors inhibiting the formation of Schwann cell precursors. Schwann cell precursors can also give rise to other types of cells, including neurons and fibroblasts *in vitro*. The remarkable plasticity of Schwann cells, recapitulated after nerve injury, aids in recovery from axon injury.

Differentiation to Immature Schwann Cells

Formation of immature Schwann cells from Schwann cell precursors is promoted by insulin-derived growth factor-2 (IGF-2), neurotrophin-3 (NT-3), (platelet-derived growth factor-β (PDGF-β), fibroblast growth factor-2 (FGF-2), Notch, and leukemia inhibitory factor (LIF) (**Table 1**). These factors can be produced by the Schwann cells and act in a autocrine manner when the axon signals are lost following injury.

TBF-β and the p75 neurotrophic factor induce programmed death of immature Schwann cells, and extracellular adenosine 5-triphosphate (ATP) and adenosine, acting on two different types of purinergic receptors (P2 and P1, respectively), inhibit Schwann cell proliferation and differentiation beyond the immature stage. ATP is released by axons firing action potentials, and adenosine is the end product of ATP degradation. This activity-dependent axon–Schwann cell signaling provides a mechanism for regulating the development of Schwann cells according to the state of functional activity in developing neural circuits.

The Schwann cell precursors are vitally dependent on axon survival signals, but mature Schwann cells can self-generate many of the essential survival factors in the absence of axons. Thus, these cells survive after the axon dies from injury. Indeed, the axon depends on the chain of surviving Schwann cells and the growth factors they secrete to grow back to its proper target. As discussed later, however, prolonged separation of Schwann cells from axons leads to detrimental changes in Schwann cells.

Myelination

Myelination begins after Schwann cells exit the cell cycle. In the promyelinating stage, just before beginning to form myelin, Schwann cells begin to increase expression of several myelin proteins, including myelin-associated glycoprotein (MAG) and myelin protein zero. The latter is the most abundant protein in peripheral nerve, and it is responsible for maintaining the compact layers of myelin together. A different protein, proteolipid protein, performs this function in central nervous system myelin.

The process of myelination begins by a Schwann cell first engulfing several axon segments at once (**Figure 1(a)**). These are sorted out until the cell selects only one axon segment to wrap with myelin (**Figure 1 (b)**). Formation of extracellular matrix is an essential early step in the process, and in cell culture, myelination will not commence until ascorbic acid is added to

Table 1 Growth and transcription factors affecting Schwann cell differentiation

	Schwann cell precursor	*Immature Schwann cell*	*Myelinating Schwann cell*
Promotes differentiation to this stage	NRG-1, Notch, FGF-2	Adenosine, ATP, β_1 integrin, c-Jun, Claw paw, Ets transcription factors, IGF, laminin, NRG-1, NT-3, SOX-2, p38, p75, TGF-β	BDNF, BRN-2, cAMP, GDNF, IGF, Krox20/NAB1/2, Erg2, laminin, NF-κB, NRG-1, Oct6, progesterone, PI3K-Akt, SKI
Inhibits differentiation to this stage	Endothelin, BMP2/4, AP2α		ATP, c-Jun, Notch, NT-3, PAX-3, SOX-2, TGF-β

Data from Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

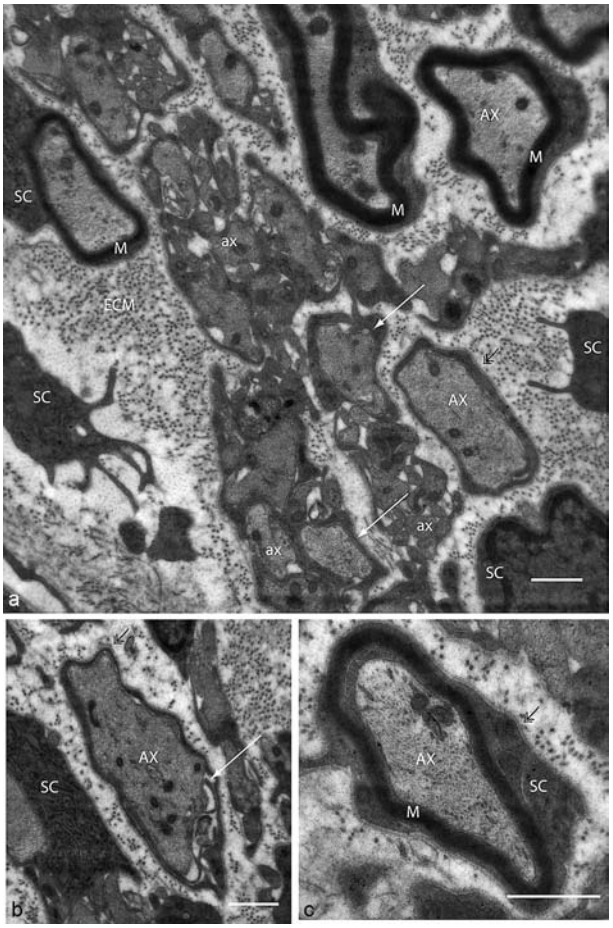

Figure 1 Schwann cell–axon relations during remyelination of regenerated axons in rat sciatic nerve. After axotomy, Schwann cells (SC) provide trophic factors and extracellular matrix components (ECM) to guide regenerating axons to form appropriate connections. (a) Early events in remyelination by SCs can be seen in sciatic nerve regenerating after being severed. Cellular processes extending from SCs (white arrows) sort out unmyelinated axons (ax) from larger diameter axons (AX). The small-diameter axons will become engulfed together in bundles inside nonmyelinating SCs and remain unmyelinated. A single SC will become associated with one segment of a large-diameter axon and begin synthesizing multiple layers of myelin membrane around a segment of the axon (M). This will provide rapid, saltatory impulse conduction by insulating the axon and organizing the fiber into internodal and nodal domains having distinct types of ion channels. Basal lamina, synthesized by SCs (double black arrows), is necessary for sorting out large-diameter axons and initiating myelination. (b) A large-diameter axon in the earliest stages of myelination is shown ensheathed by a single wrap of SC membrane. (c) Another myelinated axon shows a later stage in which compact myelin is formed by wrapping multiple layers of membrane around the axon and squeezing the cytoplasm out from between the stacks of insulating membrane. Scale bar = 500 nm.

the medium to promote formation of the basal lamina (Figure 1, black arrows). Laminin-2 is a major component of basement membrane, and mutations in α_2 laminin cause dysmyelination. β_1 integrin is one of the laminin receptors on Schwann cells, and animals lacking this receptor fail to establish the one-to-one relation with axons. Dystroglycan, another laminin receptor on Schwann cells, is not vital for early stages of myelination, but it participates in organizing the Schwann cell structure at the node.

The factors initiating myelination are not fully understood, nor is it clear why Schwann cells only myelinate large-diameter axons. Just before Schwann cells begin to myelinate, nuclear factor-kappa B (NF-κB) is upregulated, and its activation and translocation to the nucleus is necessary to regulate transcription of genes essential for myelination. A number of transcription factors are involved in initiating myelination, including Krox20/EGR2, NGFI-1-binding proteins, Oct6, and BRN2. As might be expected, cell adhesion molecules have an important role in axon–Schwann cell adhesion and myelination. The cell adhesion molecules L1 and NCAM are expressed on nonmyelinated axons, and they are downregulated during early myelination. MAG, a cell adhesion molecule that promotes association between axon and myelinating Schwann cells, is upregulated immediately after the initial layer of myelin is wrapped around the axon (Figure 1(b)). Synthesis of myelin proteins increases, and the mRNAs for myelin proteins, such as myelin basic protein (MBP),

become concentrated at distal sites in Schwann cells to provide local synthesis of myelin membrane.

Wrapping the layers of myelin around axons and squeezing the cytoplasm out from between the leaflets requires extensive cell extension and motility, which are regulated by actin–myosin cytoskeletal dynamics (Figure 1(c)). The Rho kinase (ROCK) regulates actin–myosin mechanical transduction by activating Rho, and in the absence of ROCK activity, the single myelinating process of a Schwann cell splits abnormally to form many smaller internodes, with resulting deleterious consequences for nerve conduction.

Myelin is promoted by many growth factors, including glial-derived neurotrophic factor (GDNF), NRG-1, myelin basic protein (IGFs), brain-derived neurotrophic factor (BDNF), as well as the sex hormone progesterone and the extracellular matrix molecule laminin. Multiple intracellular signaling pathways are involved, notably PI3K-Akt and cAMP. Myelination by Schwann cells is blocked by Notch activation, the neurotrophin NT-3, as well as by extracellular ATP. The latter is released by axons firing bursts of action potentials. TGF-β also inhibits myelination. Neurotrophins such as nerve growth factor (NGF) promote neuronal survival, but NGF stimulates myelination by Schwann cells through an indirect effect on axons. Indirect signals from the axons must be involved because experiments have shown that Schwann cell myelination of types of axons that are not depend on NGF for survival is unaffected by experimental manipulation of NGF.

The thickness of the myelin layer is precisely regulated in proportion to the axon diameter. Axons of larger caliber have thicker myelin, but the ratio of axon diameter to total diameter of axons including the myelin sheath falls within a range of 0.6 to 0.7 (g-ratio), regardless of the diameter of a myelinated axon. How the Schwann cell maintains this strict proportionate myelin thickness to varying axon caliber is unknown, but the thickness can be altered experimentally. BDNF, p75 neurotrophin receptor, or neuregulin overexpression results in abnormally thick myelin. Conversely, if ErbB receptors in Schwann cells are eliminated after myelination has begun, the sheath does not develop to its normal thickness. Myelin on axons regenerated after injury is typically thinner than normal, however.

Node of Ranvier

Formation of the node of Ranvier involves forming adhesive junctions between the axon and the extreme edges of the Schwann cell membrane abutting the node. At this junction, the multiple layers of myelin wrapped around the axon each seal against the axon membrane stacking in a staggered sequence at the edge of the node, thus forming the paranodal domain. In dividing the axon into nodal, paranodal, and internodal domains, the Schwann cell also affects the precise localization of ion channels and adhesion molecules along the axon, which are essential for salutatory conduction.

An important adhesion complex in forming and maintaining the paranode involves the axonal protein Caspr (contactin-associated protein, also known as paranodin), together with another protein contactin, binding to the glial partner neurofascicn 155 (Nfasc-155) at the paranodal junction. Cross-links between membrane proteins and the cytoskeleton maintain the appropriate ion channels in the proper axonal domains. Sodium channel localization at the node depends on association with Nfasc-186/NrCAM together with AnkyrinG/betaIV spectrin to anchor the channels to the actin cytoskeleton at the node. The node forms by nucleation of macromolecules in the nodal membrane, which then act as anchors for clustering additional sodium channels. Potassium channels become localized lateral to the paranode (the juxtaparanodal region), anchored by Caspr2, an axonal protein similar to Caspr, which localizes sodium channels at the node.

Myelinated Schwann cells are postmitotic, so as the axon grows, the Schwann cells must elongate at the same time, and the distance between nodes increases as the animal grows. Cytoplasmic channels beneath the outer layer of the myelin sheath, called Cajal bands, are an essential route for mRNA translocation to distal regions of the Schwann cell. When these bands are disrupted by disrupting a complex of molecules linking basement membrane to the Schwann cell plasma membrane (the L-periaxin-DRP2–dystroglycan complex), Schwann cells are impaired in their ability to elongate during nerve growth. The internodal length is reduced and conduction velocity slows.

Axon Dependence on Schwann Cells

Most of the axons die in limbs of mice genetically modified to lack Schwann cells, indicating definitively that axons and immature Schwann cells depend on each other for survival at crucial points in development. Long-term disabilities in patients with demyelinating diseases, such as Charcot–Marie–Tooth disease, can lead to axon degeneration. Myelinating Schwann cells have been shown to affect the extent of neurofilament phosphorylation in axons. Through this regulation, myelinating glia reduce the rate of slow axonal transport, in turn shrinking the caliber of the axon.

Regeneration

Axon damage results in Schwann cell responses that reflect the close interdependence of these two cells. After axotomy, the portion of the axon distal to the damage degenerates, but the Schwann cells in the distal stump begin to proliferate and synthesize growth factors that sustain the survival of the Schwann cells and the axons. Survival of neurons and elongation of sprouting axons from the proximal stump of severed axons is sustained by factors secreted by the Schwann cells and the extracellular matrix that they form. These factors include NGF, BDNF, and LIF. These beneficial properties of Schwann cells have been exploited to promote regeneration of axons in the central nervous system in Schwann cell transplantation experiments.

Schwann cells that are chronically isolated from their axons, however, begin to lose the capacity to support axon survival and outgrowth. This is one of the major reasons for permanent paralysis and failure to restore function in patients with damage to long nerves. Axon regeneration proceeds relatively slowly, at a rate of approximately 1 mm per day, and the Schwann cells in the extreme regions of limbs lose their ability to sustain axon regeneration before the regenerating axons can reach them.

Terminal Schwann Cells

At the distal end of motor neuron axons, a specialized form of Schwann cell ensheaths the neuromuscular junction. Functionally, these perisynaptic Schwann cells resemble astrocytes in the brain. Terminal Schwann cells help maintain the physical integrity of the synaptic junction, but they also have receptors for neurotransmitters that allow them to respond to neurotransmission. In turn, they release neuroactive substances that regulate the strength of synaptic transmission. These Schwann cells are also intimately associated with the axon and respond dynamically to axon injury. After axotomy, the terminal Schwann cells become highly branched and provide a route for axon sprouts from surviving junctions to reinnervate the vacated synapse.

Conclusion

Myelin appeared suddenly in the earliest vertebrates, fully formed and equally developed in both the central and the peripheral nervous system, even though the axon sheathing is made by completely different types of cells in the peripheral and central nervous system. This sudden evolutionary advance is puzzling because comparative anatomy reveals no intermediate forms. Considering the intricate choreography of cellular differentiation, recognition, mutual dependence on survival, motility, and membrane synthesis involved in myelination, this situation is especially surprising. What is clear is that the evolutionary advantage of myelin catapults nervous system function in vertebrates far beyond the limits of the invertebrate nervous system, which shares with vertebrates very similar neuronal structure and function. This success of vertebrate organisms derives from the highly regulated and complicated cellular interdependence of axon and glia.

See also: Demyelinating Diseases; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission; Schwann Cell Development; Schwann Cells and Plasticity of the Neuromuscular Junction; Schwann Cell Morphology; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Transplantation of Myelin Forming Cells.

Further Reading

Auld DS and Robitaille R (2003) Glial cells and neurotransmission: An inclusive view of synaptic function. *Neuron* 40: 389–400.

Bjartmar C, Wujek JR, and Trapp BD (2003) Axonal loss in the pathology of MS: Consequences for understanding the progressive phase of the disease. *Journal of Neurological Sciences* 206: 165–171.

Bullock TH, Moore J, and Fields RD (1984) The evolution of myelin sheaths: Both lamprey and hagfish lack myelin. *Neuroscience Letters* 48: 145–148.

Fields RD and Stevens-Graham B (2002) New views of neuron–glia communication. *Science* 298: 483–690.

Jessen KR and Mirsky R (2005) The origin and development of glial cells in peripheral nerves. *Nature Reviews Neuroscience* 6: 671–682.

Ko CP and Thompson WJ (2003) Preface to the special issue on the neuromuscular junction. *Journal of Neurocytology* 32: 423–424.

Pollak S and Peles E (2003) The local differentiation of myelinated axons at nodes of Ranvier. *Nature Reviews Neuroscience* 4: 968–980.

Salzer JL (2003) Polarized domains of myelinated axons. *Neuron* 40: 297–318.

Sherman DL and Brophy PJ (2005) Mechanisms of axon ensheathment and myelin growth. *Nature Reviews Neuroscience* 6: 683–690.

Stevens B and Fields RD (2000) Action potentials regulate Schwann cell proliferation and development. *Science* 287: 2267–2271.

Schwann Cells and Plasticity of the Neuromuscular Junction

Z Feng and C-P Ko, University of Southern California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The neuromuscular junction (NMJ) has served as a prototype synapse due to its relatively large size, simple organization, and easy accessibility. Similar to the chemical synapse in the central nervous system, the vertebrate NMJ is composed of three adjacent cellular components: the presynaptic nerve terminal, the postsynaptic specialization, and the synapse-associated glial cells. These synapse-associated cells are called perisynaptic Schwann cells (PSCs) or terminal Schwann cells at the NMJ (Figure 1). Previous studies on the NMJ have primarily focused on the roles of nerve terminals and postsynaptic components, while the role of PSCs has been largely overlooked. However, from the 1990s, a barrage of studies have revealed multiple roles of glial cells in synaptic function, formation, and plasticity at NMJs, as well as at central synapses. Emerging from these studies on synapse–glia interactions is the concept of the 'tripartite' synapse, which asserts that, along with the presynaptic nerve terminal and the postsynaptic specialization, synapse-associated glial cells should be considered as an active and integral partner of the vertebrate chemical synapse as well. In this section, we describe the key findings of synapse–glia interactions and the involvement of PSCs in the plasticity of the vertebrate NMJ and focus on the multiple roles PSCs play in synaptogenesis, maintenance, remodeling, and regeneration of the NMJ.

The General Characteristics of Perisynaptic Schwann Cells at the Neuromuscular Junction

In the vertebrate neuromuscular system, motor axons are wrapped around by myelinating Schwann cells, which are responsible for the saltatory conduction of nerve impulses; however, motor nerve terminals are capped by nonmyelinating PSCs instead. The existence of Schwann cells at the vertebrate NMJ was suspected long before the days of electron microscopy. Soon after the initial descriptions of vertebrate NMJs, Louis-Antoine Ranvier was the first to report clusters of 'arborization nuclei' in 1878. These nuclei accompany nerve terminal arbors at NMJs, and are distinct from muscle fiber nuclei. Subsequent histological studies identified these arborization nuclei as nuclei of teloglia, or terminal Schwann cells. The survival of these teloglia after nerve degeneration further suggests that they are non-neuronal cells. Electron microscopy studies in the 1950s and 1960s provided much detailed information about NMJ morphology, confirming the intimate association between terminal Schwann cells, or PSCs, and nerve-muscle contacts.

At the mammalian NMJ, PSCs do not invade the synaptic cleft between the nerve terminal and the muscle surface, and the exclusion may be attributed to laminin 11 in the synaptic cleft. At the frog NMJ, however, PSCs do extend finger-like processes looping around the nerve terminal in between active zones, and are colocalized with F-actin concentrated at the nonrelease domains of the frog nerve terminal. PSC fingers contain L-type calcium channels, but the function of PSC fingers, if any, is unknown. There are approximately three to four PSC somata per NMJ in both mammalian and amphibian muscles, and the PSC number correlates with the junctional size. PSCs are capped with basal lamina, which contain some molecules that are distinct from those in the synaptic cleft and in the extrasynaptic basal lamina of the muscle fiber. It has been suggested that the extracellular matrix associated with PSCs may play a role in guiding nerve terminal sprouts at the frog NMJ, although the molecular mechanisms remain to be studied (see below).

The advance in understanding the role of PSCs has been aided by the availability of several probes that can label them, although most of these probes are not specifically for PSCs as they can also label axonal Schwann cells. The most commonly used probe to label mammalian PSCs is an antibody to S-100, a calcium-binding protein. Recently, transgenic mice expressing green fluorescent protein (GFP) under the control of human S100B gene have been generated to reveal GFP expression in axonal Schwann cells and PSCs at living NMJs. Other probes – such as antibodies to glial fibrillary acidic protein (GFAP), neural cell adhesion molecule (NCAM), L1, LNX-1, growth-associated protein-43 (GAP-43), and the low-affinity nerve growth factor (NGF) receptor p75 – can also be used to label PSCs, although the latter two probes label PSCs only after nerve injury or blockade of nerve activity. Surprisingly, antibodies to myelinating glial markers – such as protein zero (P_0), myelin-associated glycoprotein (MAG), galactocerebroside, and 2′,3′-cyclic nucleotide 3′-phosphodiesterase – can also label PSCs even though they are nonmyelinating. The role, if any, of these myelinating markers in the differentiation of PSCs is not known.

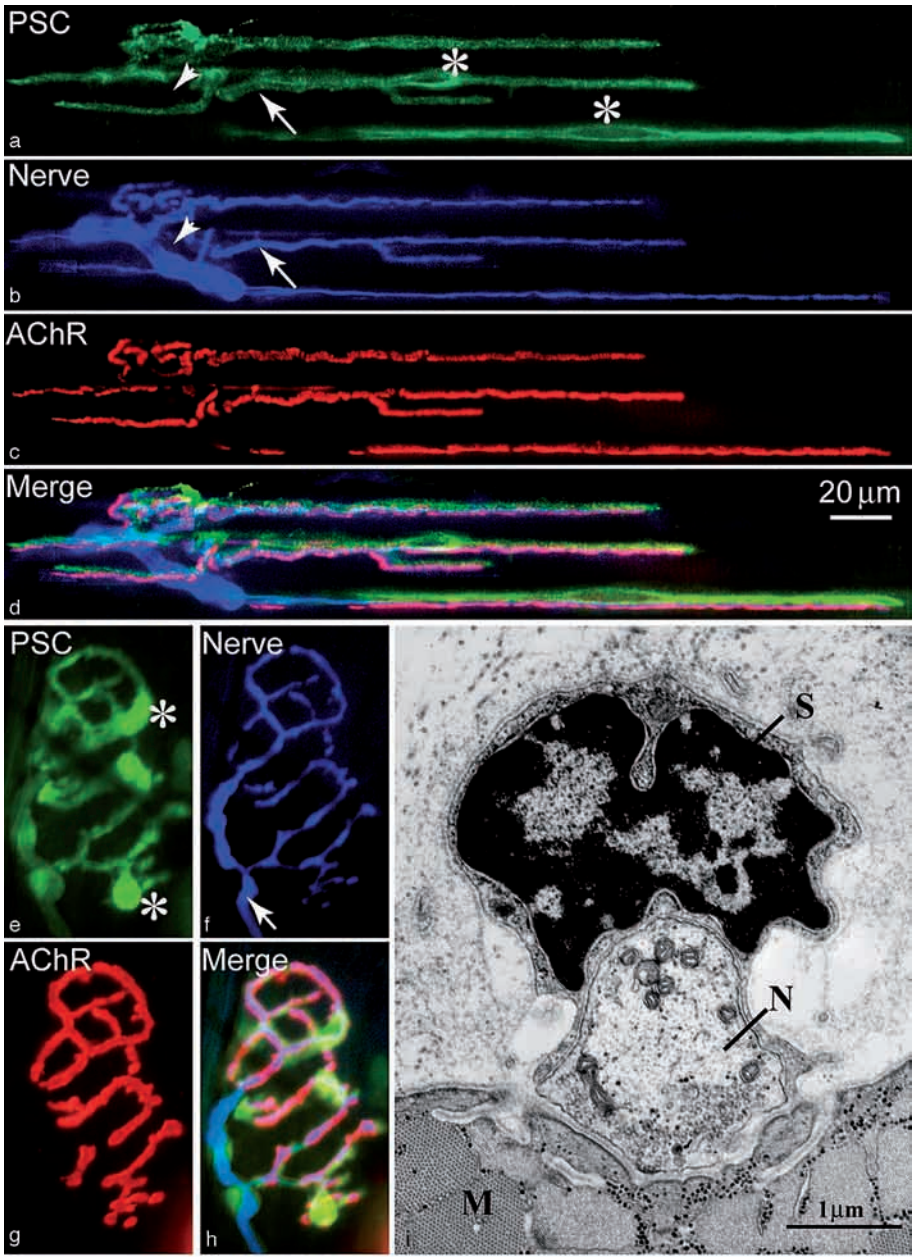

Figure 1 The tripartite organization of vertebrate neuromuscular junctions (NMJs). (a–d) A frog NMJ fluorescently labeled with a monoclonal antibody, mAb 2A12, for perisynaptic Schwann cells (PSCs) (a, green); antineurofilament and antisynapsin I antibodies for axons and presynaptic nerve terminals (b, blue); and α-bungarotoxin (α-BTX) for postsynaptic acetylcholine receptors (AChRs) (c, red). Note that mAb 2A12 labels PSC somata (asterisk in (a)) and processes (arrow in (a)), but not Schwann cells along the axon (arrowheads in (a) and (b)). As seen in the merged image (d), the three components of the NMJ are closely aligned with one another. Scale bar in (d) applies to (a)–(c). (e–h) An NMJ in a transgenic mouse that expresses green fluorescence protein (GFP) in Schwann cells (e, green) and cyan fluorescence protein (CFP) in nerve terminal (f, blue). The NMJ is also labeled with α-BTX for AChRs (g, red). Note that GFP is expressed not only in PSC somata (asterisks in (e)) and processes, but also in Schwann cells along the preterminal axon (arrow in (f)). The tripartite arrangement of the NMJ is further demonstrated in the merged image (h). (i) Electron micrograph of a frog NMJ in cross section shows the PSC (S) with its electro-dense nucleus capping the nerve terminal (N), which is apposed to postjunctional folds on the muscle fiber (M). Scale bar = 1 μm. (a–h) Reproduced from *Journal of Neurocytology*, vol. 32, 2003, 423–1037, Special issue – the neuromuscular junction, Ko CP and Thompson W, Cover picture, copyright Spring with kind permission of Springer Science and Business Media. (i) Reprinted from Reddy LV, Koirala S, Sugiura Y, Herrera AA, and Ko CP (2003) Glial cells maintain synaptic structure and function and promote development of the neuromuscular junction *in vivo*. *Neuron* 40: 563–580, with permission from Elsevier.

In addition to the above probes, there are two probes available for selectively labeling frog PSCs: peanut agglutinin (PNA), which recognizes the extracellular matrix molecules associated with PSCs; and a monoclonal antibody (mAb) 2A12, which labels the external surface membrane of PSCs. One major advantage of these two probes is that they do not interfere with synaptic function and can thus be used as vital probes for PSCs in living frog muscles. Furthermore, mAb 2A12 combined with a complement treatment can be used to selectively ablate PSCs from living frog NMJs without destroying axonal Schwann cells. The application of this complement-mediated lysis provides a unique approach to test the necessary role of PSCs *in vivo* (see below). Unfortunately, neither PNA nor mAb 2A12 labels mammalian PSCs.

With the availability of these probes and the development of new tools such as calcium imaging, it is now possible to investigate synapse–glia interactions at the NMJ. For example, it has been shown that PSCs not only can sense synaptic activity by increasing their intracellular calcium levels but also are capable of modulating transmitter release when G-proteins or intracellular calcium levels are manipulated pharmacologically. Studies on the reciprocal interactions between PSCs and nerve terminals have also unraveled the multiple roles of PSCs in synaptic development, maintenance, and repair at the NMJ, as detailed below.

Role of PSCs in Synaptogenesis

The intimate association between PSCs and nerve terminals seen at adult NMJs has begged a question as to whether synapse–glia interactions also play an essential role in the development of NMJs. It has been shown that, during development, neuronal supply of neuregulin-1 (NRG1) is required for the survival of axonal Schwann cells as well as PSCs. Disruption of the neuregulin pathway by knocking out NRG1, or its receptor ErbB2 or ErbB3, in transgenic mice results in the absence of Schwann cells both in the peripheral nerves and NMJs. Despite the lack of Schwann cells in these transgenic mice, motor axons still project to their target muscles and form transient nerve-muscle contacts. These studies suggest that, during development, while axons are required for the survival of Schwann cells, Schwann cells are required neither for axonal navigation to target muscles nor for the initial formation of nerve-muscle contacts. Consistent with this suggestion is the finding that functional nerve-muscle contacts can be formed in *Xenopus* nerve-muscle cultures in the absence of Schwann cells. However, several lines of evidence suggest that Schwann cells are essential for the subsequent growth, maturation, and maintenance of developing NMJs. First, Schwann cells are present at normal NMJs soon after initial nerve-muscle contacts are formed. Second, the number of PSCs at the NMJ increases and correlates with the size of end plates during the period of rapid synaptogenesis. Third, in the transgenic mice that are deprived of Schwann cells, although initial nerve-muscle contacts are formed, the newly formed synaptic contacts cannot be maintained without the support of Schwann cells. This suggests a possible maintenance role of Schwann cells (see below). However, the specific role of PSCs at developing mammalian NMJs cannot be confirmed since axonal Schwann cells too are absent in these transgenic mice.

To examine the direct role of PSCs in synaptogenesis, repeated *in vivo* observations of tadpole NMJs have been conducted. They revealed that PSCs extend processes beyond nerve terminals, and the subsequent growth of nerve terminals, as well as the addition of synapses, occurs preferably along the preceding PSC sprouts. The necessity of PSCs at developing frog NMJs is confirmed using the PSC ablation technique. Repeated *in vivo* observations demonstrate that less synaptic growth and more synaptic retraction, including a complete loss of synapses in some cases, occur in tadpole muscles several days after PSC ablation. Therefore, PSCs and their sprouts play an important role in guiding extending nerve terminals and promoting synaptic growth, as well as maintaining newly formed NMJs in tadpole muscles.

The necessary role of PSCs in synaptic growth has also been implicated at mammalian NMJs. For example, although partial denervation in adult muscles results in sprouting of both nerve terminals and PSCs (see below), partial denervation in neonatal rat muscles does not induce nerve terminal sprouting. This result could be attributed to the apoptosis of PSCs following nerve injury in neonatal muscles. The withdrawal of nerve terminals seen in mutant mice lacking ErbB2 and ErbB3 is also consistent with a potential role of PSCs in the maintenance of developing synapses (although the involvement of axonal Schwann cells cannot be excluded). Furthermore, both exogenous application of NRG1 to developing rat muscles and induction of constitutively active ErbB2 in neonatal mouse Schwann cells activate the sprouting of nerve terminals and PSCs. Thus, NRG1-ErbB signaling provides an essential mechanism of how axons influence the development of PSCs.

However, the molecular mechanisms of how PSCs promote the growth and maintenance of nerve terminals at developing NMJs are not understood. It has been shown that Schwann cell-conditioned medium

promotes synaptogenesis in *Xenopus* nerve-muscle cultures. One candidate for mediating Schwann cell-promoted synaptogenesis in tissue cultures may be transforming growth factor (TGF)-β1. Whether TGF-β1 and other not-yet-identified factors also promote synaptogenesis *in vivo* remains to be explored. In addition to synaptogenesis, Schwann cell-derived molecules can enhance synaptic transmission in *Xenopus* nerve-muscle cultures. Application of Schwann cell-conditioned medium causes an acute increase in the frequency, but not the amplitude, of spontaneous synaptic currents by more than 100-fold. The identity of the Schwann cell-derived molecules responsible for enhancing transmitter release is currently unknown. Whether Schwann cells also enhance transmitter release at developing NMJs *in vivo* remains to be investigated.

Role of PSCs in Synaptic Maintenance at the Adult Neuromuscular Junction

Do PSCs also play a pivotal role in maintaining adult NMJs similar to developing NMJs? This question has been examined at frog NMJs using complement-mediated cell lysis, as described above for tadpole NMJs, to selectively ablate adult PSCs *in vivo*. Shortly after PSC ablation (within 5 h), neither the ultrastructure of nerve-muscle contacts nor synaptic transmission is affected. Thus, PSCs do not acutely modulate synaptic function and are dispensable for the short-term maintenance of frog NMJs, although PSCs are capable of modulating transmitter release by pharmacological treatments. In contrast to the lack of acute effect, partial or total retraction of nerve terminals is observed at frog NMJs 1 week after PSC ablation. In addition, transmitter release is decreased approximately by half. The overall muscle function as seen in the nerve-evoked muscle twitch tension is also reduced. These observations demonstrate that PSCs are essential for the long-term maintenance of NMJs. The mechanisms by which PSCs maintain adult NMJs are not known. The retraction of nerve terminals does not seem to be attributed to nerve degeneration as there are no signs of Wallerian degeneration – such as electron dense cytoplasm and swollen mitochondria – observed in retracting nerve terminals. The absence of global nerve terminal detachment after PSC ablation suggests that PSCs do not simply serve as mechanical 'glue' to attach nerve terminals with muscle fibers. Schwann cells have been shown to express a variety of trophic factors, including TGF-β1, and extracellular matrix (ECM) molecules. The potential involvement of these trophic factors and ECM molecules in maintaining nerve terminal structure and function remains to be determined.

Recent studies suggest that PSCs also play a similar maintenance role at adult mammalian NMJs. In a transgenic mice model, in which constitutively active ErbB2 is conditionally expressed in Schwann cells, PSCs sprout massively and nerve terminals also extend processes following PSC sprouts in adult muscles. Furthermore, when the expression of active ErbB2 is turned off, both PSC sprouts and nerve terminal sprouts retract. These findings suggest that PSCs can support nerve terminal growth and that activation of the neuregulin signaling pathway in Schwann cells is sufficient to induce nerve-terminal sprouting. However, it is important to investigate the necessary role of PSCs at adult mammalian NMJs using loss-of-function approaches, such as blocking the neuregulin-signaling pathway in PSCs or deleting PSCs *in vivo*. Unfortunately, the fact that the transgenic mice lacking Schwann cells die at birth precludes the study of PSCs' role in synaptic maintenance at adult mammalian NMJs. Since mAb 2A12 does not recognize mammalian PSCs, it cannot be used to ablate PSCs with the complement-mediated cell lysis approach in mammalian muscles. Recently, it has been shown that a subtype of anti-disialosyl antibodies that is found in Miller Fisher syndrome (MFS) preferentially binds and kills PSCs via membrane attack complexes at mouse NMJs. Despite the damage of PSCs following this antibody and complement treatment, both NMJ morphology and synaptic transmission in mammalian muscles are not acutely affected. This result is similar to the finding from frog NMJs following acute PSC ablation. However, no study on the long-term effect after PSC ablation has been performed in adult mammalian muscles. It would also be interesting to examine whether defects in PSCs may contribute to MFS or other neuropathies.

Role of PSCs in Synaptic Remodeling, Degeneration, and Regeneration

Remodeling

Mature NMJs are not static, but undergo synaptic remodeling throughout adult life. For example, extension and retraction of nerve terminals can be seen with repeated *in vivo* observations at frog NMJs in normal intact muscles. Similar to nerve terminals, PSCs are also very dynamic in frog muscles. Using PNA as a vital probe for PSC-associated extracellular matrix, it has been shown that PSCs and associated ECM often extend processes beyond nerve terminals, which grow along the preceding PSC sprouts. Similar dynamic behavior of PSCs and nerve terminals can occur even within minutes, as observed at adult toad NMJs *in vivo*. These results suggest that PSCs

guide nerve terminal growth during synaptic remodeling at adult amphibian NMJs. Although mammalian NMJs are relatively less plastic than amphibian NMJs, extension of short processes of PSCs and minor nerve terminal filopodia and lamellipodia has been seen at adult mammalian NMJs as well. The remodeling plasticity of mammalian NMJs is much more pronounced in androgen-sensitive muscles, and PSC number correlates with end plate size, which can enlarge or shrink in response to androgen treatment or castration, respectively.

Degeneration and Regeneration

After nerve injury, PSCs become phagocytic as they engulf the debris of degenerating nerve terminals and occupy the denervated end plate. Surprisingly, miniature end plate potentials can be recorded even at denervated end plates, albeit at a much lower frequency than those recorded from intact end plates. These spontaneous potentials at denervated end plates are thought to be caused by acetylcholine released from PSCs, but the functional significance of these Schwann cell miniature potentials remains unknown.

Besides being phagocytic, PSCs become very dynamic after nerve degeneration as PSCs sprout their processes profusely beyond the original end plate sites. PSC sprouting can also be induced by blockade of nerve activity. Regenerating axons not only grow along axonal Schwann cells but also extend nerve terminal processes along PSC sprouts. This indicates the importance of Schwann cells in promoting synaptic restoration. Furthermore, regenerating nerve terminals can follow PSC 'bridges,' which are formed by PSC processes between neighboring end plates, to innervate adjacent denervated junctional sites. Using mice that express GFP in Schwann cells and cyan fluorescent protein (CFP) in axons, it has been demonstrated that PSC sprouts induced by denervation guide regenerating nerve terminals. Similar results using vital probes to label PSCs and nerve terminals have also been found in frog muscles after nerve injury and regeneration. However, in contrast to mice, frog PSC sprouting is not induced immediately following nerve injury, but is triggered only upon the arrival of regenerating axons. Taken together, these findings from mammalian and frog muscles strongly suggest that PSC sprouts guide nerve terminal extension and play a central role in the reinnervation of the original junctional sites after nerve injury.

The importance of PSC sprouting in synaptic repair has also been shown in muscles after partial denervation. PSC bridges are formed as PSC sprouts extend from denervated end plates to contact innervated end plates following partial denervation. Consequently, nerve terminals from the innervated end plates grow along these PSC bridges to innervate the neighboring denervated end plates. Thus, similar to PSC sprouts induced by the total axotomy, PSC bridges can guide regenerating nerve terminal sprouts and help synaptic repair following partial damage of axons in adult muscles. The necessity of PSC sprouts in inducing nerve terminal sprouting has also been implicated. As discussed above on the role of PSCs in synaptogenesis, partial denervation of neonatal muscles does not trigger nerve terminal sprouting, which could be attributed to the apoptosis of developing PSCs following nerve injury. Furthermore, poor reinnervation has been reported in aging muscle as well as in *mdx* mice (a model for Duchenne muscular dystrophy), in which PSC bridge formation is impaired. Thus, PSC bridges likely are required for inducing nerve terminal sprouting.

The mechanisms by which PSC sprouting is induced following nerve injury are not fully understood. Activating neuregulin signaling pathway by expressing constitutively active ErbB2 receptors in PSCs can induce PSC sprouting, suggesting that the neuregulin-ErbB signaling pathway may play a role in PSC sprouting and nerve terminal sprouting at adult NMJs, similar to developing NMJs. It has been shown that nerve terminal sprouting can be induced by ciliary neurotrophic factor (CNTF) and insulin-like growth factor 1 (IGF-1). Whether these trophic factors, and other molecules such as GFAP, GAP-43, and p75, which are upregulated in PSCs after nerve injury or inactivity, mediate PSC-induced nerve sprouting is unknown. Nerve activity has also been shown to influence PSC sprouting. For example, although blocking synaptic activity by botulinum toxin or α-bungarotoxin induces sprouting of both nerve terminals and PSC processes, inactivity prevents the formation of PSC bridges in partially denervated muscles. On the other hand, direct stimulation or exercise of partially denervated muscles also impairs PSC bridge formation and nerve terminal sprouting. It is not clear how activity and inactivity can both inhibit PSC bridge formation following partial denervation.

In addition to their guidance role for regenerating presynaptic nerve terminals after nerve injury, PSCs may also modulate postsynaptic specializations by expressing neuronal isoforms of agrin. PSC-derived agrin may explain the appearance of AChR clusters underneath PSC sprouts in frog muscles following nerve injury and regeneration. In addition to agrin, PSCs also express neuregulin-2, which may promote AChR synthesis and synaptic differentiation at mammalian NMJs. Whether and how the postsynaptic role of PSCs may contribute to synaptic repair remains to be further examined.

Conclusions

Despite the traditional view of glial cells as a passive supporting player at chemical synapses, growing evidence has demonstrated that glial cells are an integral and essential component of synapses in both central and peripheral nervous systems. In this section, we have described the roles of PSCs during the formation, maintenance, and regeneration of the NMJ. Axonal navigation to target muscles and the formation of initial nerve-muscle contacts do not require the presence of Schwann cells; however, PSCs are essential for guiding extending nerve terminals and promoting synaptic growth, as well as maintaining developing NMJs. In adult muscles, while PSCs are dispensable for the short-term maintenance of NMJs, long-term absence of PSCs causes retraction of nerve terminals and reduction in synaptic transmission. This is possibly due to the lack of trophic support provided by PSCs. PSC sprouts guide nerve terminal sprouts during synaptic remodeling in intact adult muscles, as well as guide regenerating nerve terminals after nerve injury. NRG1-ErbB signaling is thought to play an essential role in the survival and sprouting of developing PSCs, as well as in inducing sprouting of PSCs and nerve terminals in adult denervated muscles. Similar to PSCs, astrocytes also play multiple roles at CNS synapses; therefore, synapses should be viewed as a tripartite structure, taking into account the contribution of glial cells in the formation, function, and maintenance of synapses. The molecular mechanisms by which synapse-associated glial cells contribute to multiple aspects of the synapse are not well understood. Identification of these mechanisms in the future will lead to better understanding of how synapses form, function, maintain, and repair.

See also: Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction (NMJ): Aging; Neuromuscular Junction (NMJ): Inherited and Acquired Disorders; Neuromuscular Junction (NMJ): Presynaptic Stretch Effects on Neuromuscular Transmission; Neuromuscular Junction (NMJ): Presynaptic Schwann Cells and Modulation of Neuromuscular Transmission; Schwann Cells and Axon Relationship; Schwann Cell Morphology.

Further Reading

Auld DS, Colomar A, Belair EL, et al. (2003) Modulation of neurotransmission by reciprocal synapse–glial interactions at the neuromuscular junction. *Journal of Neurocytology* 32: 1003–1015.

Halstead SK, Morrison I, O'Hanlon GM, et al. (2005) Anti-disialosyl antibodies mediate selective neuronal or Schwann cell injury at mouse neuromuscular junctions. *Glia* 52: 177–189.

Hayworth CR, Moody SE, Chodosh LA, et al. (2006) Induction of neuregulin signaling in mouse Schwann cells *in vivo* mimics responses to denervation. *Journal of Neuroscience* 26: 6873–6884.

Kang H, Tian L, and Thompson W (2003) Terminal Schwann cells guide the reinnervation of muscle after nerve injury. *Journal of Neurocytology* 32: 975–985.

Ko CP, Sugiura Y, and Feng Z (2007) The biology of perisynaptic (terminal) Schwann cells. In: Armati P (ed.) *The Biology of Schwann Cells: Development, Differentiation and Immunomodulation*, pp. 72–99. Cambridge University Press.

Ko CP and Thompson W (2003) Special issue – the neuromuscular junction. *Journal of Neurocytology* 32: 423–1037.

Koirala S, Reddy LV, and Ko CP (2003) Roles of glial cells in the formation, function, and maintenance of the neuromuscular junction. *Journal of Neurocytology* 32: 987–1002.

Reddy LV, Koirala S, Sugiura Y, Herrera AA, and Ko CP (2003) Glial cells maintain synaptic structure and function and promote development of the neuromuscular junction *in vivo*. *Neuron* 40: 563–580.

Son YJ and Thompson WJ (1995) Schwann cell processes guide regeneration of peripheral axons. *Neuron* 14: 125–132.

Son YJ and Thompson WJ (1995) Nerve sprouting in muscle is induced and guided by processes extended by Schwann cells. *Neuron* 14: 133–141.

Volterra A, Magistretti PJ, and Haydon PG (2002) *The Tripartite Synapse: Glia in Synaptic Transmission*. Oxford: Oxford University Press.

Zuo Y, Lubischer JL, Kang H, et al. (2004) Fluorescent proteins expressed in mouse transgenic lines mark subsets of glia, neurons, macrophages, and dendritic cells for vital examination. *Journal of Neuroscience* 24: 10999–11009.

Seasonal Changes in Night-Length and Impact on Human Sleep

K P Wright Jr., University of Colorado, Boulder, CO, USA

© 2009 Published by Elsevier Ltd.

Sleep Regulation and Seasonal Changes

Sleep in humans is primarily regulated by two factors, sleep homeostasis and internal circadian timing. In general, homeostatic sleep drive increases the longer we are awake and is dissipated during sleep. Most humans require 7–9 h of sleep each night to fulfill the homeostatic drive for sleep. An internal biological clock, located in the brain's hypothalamus, regulates sleep across the 24 h day. In humans and other diurnal animals, the biological or 'circadian clock' promotes sleep at night and wakefulness during the day. Specific brain regions are responsible for promoting wakefulness and others for promoting sleep, and these brain regions are primarily regulated by the sleep homeostatic and circadian systems. However, environmental factors can also influence sleep-wakefulness states. In humans, the timing, duration, and structure of sleep are reported to change across seasons. Specifically, bedtimes and wake times are reported to be later and sleep duration is reported to be longer in the winter compared to the summer. In addition, the total amount of rapid eye movement (REM) or dream sleep has been reported to be greater in the winter than in the summer. Why do these seasonal changes in sleep occur? The most common environmental factors that change with the season and influence the sleep and circadian systems are light and temperature. Seasonal changes in the hours of daily sunlight (photoperiod) and environmental temperatures occur at latitudes away from the equator. When the northern hemisphere is tilted toward the sun during summer, the photoperiod is longer and environmental temperatures are warmer, while at the same time the southern hemisphere is experiencing winter with a shorter photoperiod and colder temperatures. Near the equator, the photoperiod is relatively constant across the year, with approximately 12 h of sunlight and 12 h of darkness. At latitudes away from the equator, there are larger changes in the photoperiod across the seasons. For example, the city of Seattle in Washington State receives up to 16 h of sunlight each day in the summer and less than 9 h of sunlight each day in the winter. At the polar extremes, there are times when the sun never sets in the summer and never rises in the winter.

Influence of Environmental Factors on Sleep

How do environmental factors influence sleep? When we sleep, our senses are still active. Whether remembered or not, the sleeping brain responds to sounds, smells, pain, temperature, and even visual stimuli (e.g., light). The finding that sensory information can be processed by the brain during sleep provides a mechanism by which environmental factors disrupt sleep.

Temperature and Humidity

It is well documented that hot and cold environmental temperatures disrupt sleep. When it is too hot or too cold, the time it takes to fall asleep is longer, sleep is lighter, awakenings from sleep are more frequent, and REM sleep is reduced. REM sleep appears to be particularly sensitive to being disturbed by ambient temperatures that are too hot or too cold because brain thermoregulatory processes that maintain body temperature are altered during REM sleep.

Humans have constructed environments and developed technologies (e.g., heated homes) that reduce the influence of seasonal changes in environmental temperature on sleep. Sleep is best in a relatively cool environment. Thus, individuals who like warmer house temperatures in the winter should wear warmer cloths to bed; a house temperature that is too high may disrupt the sleep of others in the house. Assuming that heating in homes is adequate, the influence of cold temperatures on sleep is most commonly experienced when sleeping outside, such as in a tent during the fall or winter. In contrast, the influence of hot summer temperatures on sleep affects many more people, and as a result, there is widespread use of air-conditioning and fans to improve sleep at night. A high humidity environment has also been reported to increase the disruptive effects of hot ambient temperature on sleep.

Light Exposure

Exposure to light can acutely alter sleep. For example, exposure to light in the morning increases the ease of awakening, and exposure to light in the evening prior to bedtime increases the time it takes to fall asleep.

Influence of Seasonal Changes in Photoperiod on the Circadian Timekeeping System and Sleep

Seasonal changes in photoperiod alter the timing of the internal circadian clock, which in turn may alter

the timing of sleep. Light is the strongest environmental time cue for the circadian clock in humans. Exposure to light at night delays the timing of the internal circadian clock in a westward direction, and as a result, the circadian clock promotes sleep later the next day; whereas exposure to light in the morning advances the internal circadian clock in an eastward direction, and as a result the circadian clock promotes sleep earlier the next day.

The circadian clock regulates daily rhythms in many physiological functions such as alertness, performance, cardiovascular function, immune function, endocrine function, and sleep. One particular hormone that is strongly regulated by the circadian clock is the sleep-promoting hormone melatonin. Melatonin levels are typically high at night and low during the day. On average, humans go to bed approximately 2 h after melatonin levels begin to rise at night. The clock hour when melatonin levels rise or peak has been reported to be later during the winter than during the summer, and thus the time when the internal circadian clock promotes sleep is later in the winter than in the summer (Figure 1). This seasonal change in circadian timing is most likely related to the later time of sunrise in the winter. Changes in the timing of sleep and of the melatonin rhythm have also been reported to occur at research stations in Antarctica and in cities at latitudes away from the equator such as in northern Norway. In general, findings from most studies indicate a delay in the timing of the internal circadian clock during the winter. However, findings from some studies indicate a lack of synchronization between the internal circadian clock and the 24 h day. Lack of circadian synchronization occurs either when individuals are required to be awake when the levels of the sleep-promoting hormone are high or when individuals go to sleep later and awaken later each day and thus their sleep schedule cycles around the clock.

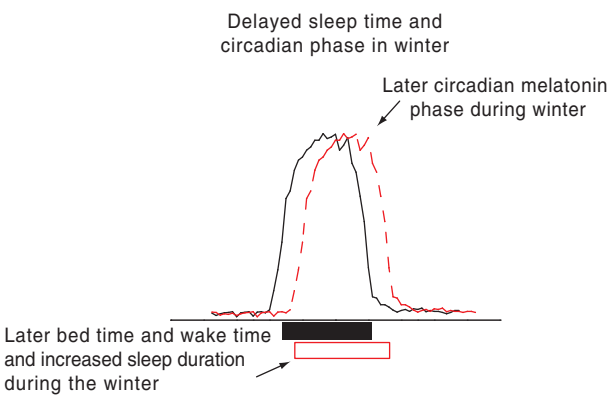

Figure 1 The daily pattern of melatonin is primarily regulated by the internal circadian timekeeping system. Melatonin levels are low during the day, increase on average ~2 h prior to habitual bedtime, and are high at night. Sleep quality is best at internal biological times when melatonin levels are high. The desire to go to bed later and to sleep in later in the morning during the winter is related to a shift in internal biological time, as represented by the delay in the timing of the circadian melatonin rhythm (dashed line). The phase delay in circadian timing may be related to the later sunrise in the winter. The solid line represents melatonin curve in the summer and the dashed line represents melatonin curve in the winter. The closed box represents sleep time in the summer, and the open box represents sleep time in the winter.

Many seasonal changes in physiology and behavior that are observed in nonhumans (e.g., seasonal reproduction, winter vs. summer coat color, and weight gain in hamsters) are primarily regulated by changes in the duration of high melatonin levels at night, which itself is regulated by changes in photoperiod. Melatonin is the primary biological signal regulated by the circadian clock to indicate to the rest of the body information about environmental photoperiod. In seasonal animals, the duration of high melatonin levels is longer during long winter nights than during short summer nights. The human circadian system is similar to that of other animals in that the human circadian system has the capacity to respond to seasonal changes in photoperiod with respect to the duration of high melatonin levels. In experimental studies, when participants slept in a sleep laboratory under simulated long winter nights (14 h of darkness) for 4 weeks, the duration of high melatonin levels was longer than during simulated short summer nights (8 h of darkness). Sleep duration was also longer during the simulated winter than during the simulated summer nights. However, outside of these experimental conditions, seasonal changes in the duration of the melatonin rhythm and of sleep in humans appear to be small. This is likely due to the widespread use of artificial lights before sunrise and after sunset. The use of artificial light may indicate to our circadian clock that we live in a near constant summer photoperiod year round.

Other Seasonal Changes That Influence Sleep

Other factors that may influence seasonal changes in human sleep include seasonal changes in health/sickness, environmental noise, and schedule-induced changes in sleep times. Seasonal exposure to foreign pathogens such as a flu virus in the winter or allergens in the spring can alter sleep. In response to being exposed to pathogens and allergens, cells of the immune system release proteins such as cytokines that alter sleep regulatory processes in the brain to promote sleep.

Physical activity levels are reported to be lower in the winter than in the summer. A reduction in physical activity levels can disrupt the quality of sleep. Environmental noise is more likely to disrupt sleep in the summer when sleeping with the window open (e.g., traffic noise or birdsong).

Last, seasonal changes in sleep timing may also be schedule induced. For example, the circadian clock in adolescents and young adults is reported to promote late bed and wake times. Specifically, levels of the sleep-promoting hormone melatonin are reported to rise later at night in adolescents and young adults as compared to children and older adults. During summer vacation, later bed and wake times can be accommodated; however, when school starts in the fall, the students are required to awaken earlier to accommodate the early school start times.

Winter Depression

Some humans appear to be particularly sensitive to seasonal changes in the environment. These individuals are at risk for developing winter depression, also known as seasonal affective disorder. Patients with winter depression report symptoms of depressed mood, irritability, loss of interest in performing activities, fatigue and sleep problems during the winter months, and lack of these symptoms in the spring and summer months. Sleep duration has been reported to be increased in the winter, although the increase in sleep time seen in patients with winter depression was not different than the increase in sleep time seen in nondepressed subjects. However, the amount of deep slow-wave sleep was reported to be lower and the number of rapid eye movements during REM sleep was reported to be higher in patients with winter depression during the winter months. Patients with winter depression are thought to be particularly sensitive to seasonal changes in light exposure. One hypothesis suggests that the phase delay of the internal circadian clock that occurs in the winter is a risk factor and/or underlying mechanism for winter depression. Another hypothesis suggests that a reduction in the amount of light exposure can increase the risk of winter depression. Why some people, but not others, experience winter depression in response to the circadian phase delay and/or changes in light exposure in the winter is unknown. Regardless of the underlying mechanism, the most common treatment for winter depression is exposure to bright light in the morning, although light exposure at other times of day and melatonin administration in the evening have also been reported to be effective.

See also: Circadian Genes and the Sleep–Wake Cycle; Circadian Function and Therapeutic Potential of Melatonin in Humans; Circadian Rhythms: Influence of Light in Humans; Circadian Rhythms in Sleepiness, Alertness, and Performance; Entrainment of Circadian Rhythms by Light; Peripheral Circadian Oscillators; Photoperiodic Regulation of Reproductive Cycles; Photoreceptors and Circadian Clocks; Seasonal Timing: Neural Mechanisms; Shift Work and Circadian Rhythms; Sleep: Development and Circadian Control; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Binkley S, Tome MB, Crawford D, and Mosher K (1990) Human daily rhythms measured for one year. *Physiology and Behavior* 48: 293–298.

Carskadon MA, Wolfson AR, Acebo C, Tzischinsky O, and Seifer R (1998) Adolescent sleep patterns, circadian timing, and sleepiness at a transition to early school days. *Sleep* 21: 871–881.

Czeisler CA and Wright KP Jr. (1999) Influence of light on circadian rhythmicity in humans. In: Turek FW and Zee PC (eds.) *Regulation of Sleep and Circadian Rhythms*, pp. 149–180. New York: Marcel Dekker.

Honma KI, Honma S, Kohsaka M, and Fukuda N (1992) Seasonal-variation in the human circadian-rhythm—Dissociation between sleep and temperature rhythm. *American Journal of Physiology* 262: R885–R891.

Kleitman N (1963) *Sleep and Wakefulness*. Chicago: University of Chicago Press.

Kohsaka M, Fukuda N, Honma K, Honma S, and Morita N (1992) Seasonality in human sleep. *Experientia* 48: 231–233.

Lewy AJ, Lefler BJ, Emens JS, and Bauer VK (2006) The circadian basis of winter depression. *Proceedings of the National Academy of Sciences of the United States of America* 103: 7414–7419.

Stuck BA, Czajkowski J, Hagner AE, et al. (2004) Changes in daytime sleepiness, quality of life, and objective sleep patterns in seasonal allergic rhinitis: A controlled clinical trial. *Journal of Allergy and Clinical Immunology* 113: 663–668.

Usui A, Obinata I, Ishizuka Y, Okado T, Fukuzawa H, and Kanba S (2000) Seasonal changes in human sleep-wake rhythm in Antarctica and Japan. *Psychiatry and Clinical Neurosciences* 54: 361–362.

Wehr TA (1991) The durations of human melatonin secretion and sleep respond to changes in daylength (photoperiod). *Journal of Clinical Endocrinology and Metabolism* 73: 1276–1280.

Wehr TA, Giesen HA, Moul DE, Turner EH, and Schwartz PJ (1995) Suppression of men's responses to seasonal changes in day length by modern artificial lighting. *American Journal of Physiology* 269: R173–R178.

Wright KP Jr. (2005) Environmental influences on sleep and sleep deprivation. In: Kushida CA (ed.) *Sleep Deprivation: Basic Science, Physiology, and Behavior*, pp. 121–155. New York: Marcel Dekker.

Seasonal Hormonal Changes and Behavior

B D Goldman, University of Connecticut, Storrs, CT, USA
C K Song and T J Bartness, Georgia State University, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Mechanisms of Seasonal Timing

It is adaptive for organisms to prepare for future changes in seasons so that biologically important physiological and behavioral responses occur at the optimal time of year. A number of environmental cues are used by various species to achieve seasonal timing. Of these, photoperiod is the most widely used cue, likely because it is readily detectable and is the most noise-free environmental signal for establishing seasonal time. Thus, although photoperiod may be of much less direct importance than factors such as food availability and temperature, in many environments it is the best single predictor of seasonal change. Most photoperiodic organisms employ circadian timing mechanisms to measure photoperiod (photoperiod time measurement (PTM)). This phenomenon has been most widely studied in birds, mammals, insects, and plants. In many organisms, including all mammals and birds studied to date, a biological circadian oscillator is involved in PTM.

Retinal Ganglion Cells Are Ultimately Connected to Pinealocytes via the Sympathetic Nervous System

In mammals, the photoperiod cue is received by the retinal ganglion cells and is transmitted through a multisynaptic pathway to the pineal gland and their primary cell type, pinealocytes. This has been most elegantly demonstrated using a retrograde transneuronal viral tract tracer, pseudorabies virus (PRV). PRV infects chains of functionally connected neurons labeled from the site of injection (end of the circuit) to its beginning. In brief, the circuit begins with retinal ganglion cells that ultimately project to the pinealocytes through the following neuronal chain: suprachiasmatic nucleus, hypothalamic paraventricular nucleus, intermediolateral horn of the spinal cord, and superior cervical ganglia (Figure 1).

The Pineal Gland Is Part of the Mammalian Photoperiodic Mechanism

In addition to the involvement of a circadian clock in PTM mammals, the pineal gland and its primary hormone, melatonin (MEL), are also key features of PTM. Once the neural signal that codes day length information reaches the pineal via the circuit described previously, it is transduced into a neuroendocrine signal in the form of the rhythmic secretion of MEL into the circulation. Pineal and plasma MEL concentrations are at their nadir during the light phase of the photocycle and at their peak during the dark phase. This pattern of MEL synthesis and secretion results from the regulation of the pineal gland by an endogenous circadian oscillator that is entrained to the light:dark cycle. Thus, the duration of the night is faithfully coded into the duration of MEL secretion and serves to trigger seasonal responses: Long MEL secretion durations signal 'fall/winter,' whereas short MEL secretion durations signal 'spring/summer' (Figure 2).

Photoperiodic mammals can be divided into two categories: Some species, such as the long day-breeding Syrian and Siberian hamsters, and several other rodent species, do not exhibit indications of seasonal change when held under a constant long day length in the laboratory. Exposure to the decreasing day lengths indicative of late summer is a necessary stimulus to induce a shift to the winter physiology/behavior that includes a complete inhibition of reproductive activity. Although an increase in day length can now stimulate a return to reproductive activity and other responses that are typical of spring/summer, this is not a necessary cue. Hamsters will 'spontaneously' return to a spring/summer physiological state after being held in short days for approximately 4 or 5 months, with the exact time depending on the species. The exact time of return to the spring/summer phase under constant conditions is apparently determined by the operation of an endogenous interval timer, the anatomical identity and underlying mechanisms of which are unknown. Regardless, the return to spring condition even under conditions of continuous short photoperiod is termed 'short-day photorefractoriness,' and this part of the seasonal timing mechanism appears to be a general characteristic of species with this type of annual cycle.

A second category of photoperiodic mammals includes those termed 'circannual.' These species exhibit repeated cycles of seasonal changes even when held under seasonally constant conditions (fixed day length and unchanging temperature). Under constant conditions, however, the circannual cycles generally have period lengths that are substantially shorter than 365 days. Under field conditions, day length cues are used to adjust the timing of the circannual mechanism so as to yield a true annual periodicity. The domestic

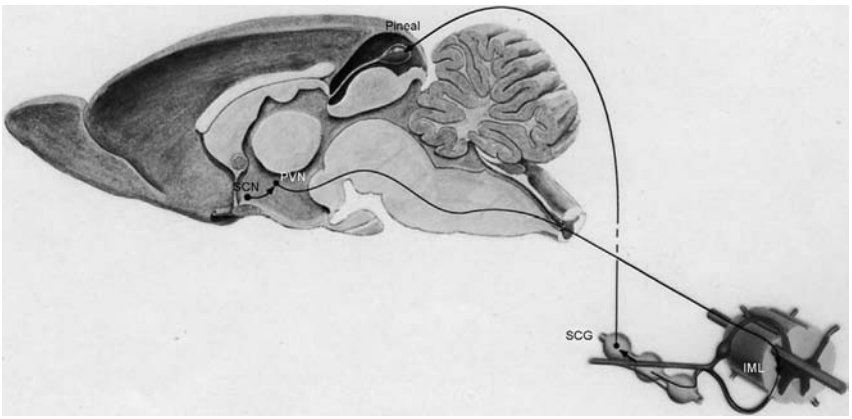

Figure 1 Using the pseudorabies virus, a transneuronal retrograde tract tracer, the multisynaptic pathway connecting the retina to the pineal gland was defined for the first time within an animal. Retinal projections via the retinohypothalamic tract (not shown) synapse on neurons in the suprachiasmatic nucleus (SCN). In turn, these neurons project to the hypothalamic paraventricular hypothalamic nucleus (PVN), then the preganglionic sympathetic neurons in the intermediolateral cell column (IML), then the postganglionic sympathetic neurons in the superior cervical ganglion (SCG), finally projecting to pinealocytes. Reproduced from Larsen PJ, Enquist LW, and Card JP (1998) Characterization of the multisynaptic neuronal control of the rat pineal gland using viral transneuronal tracing. *European Journal of Neuroscience* 10: 128–145, with permission.

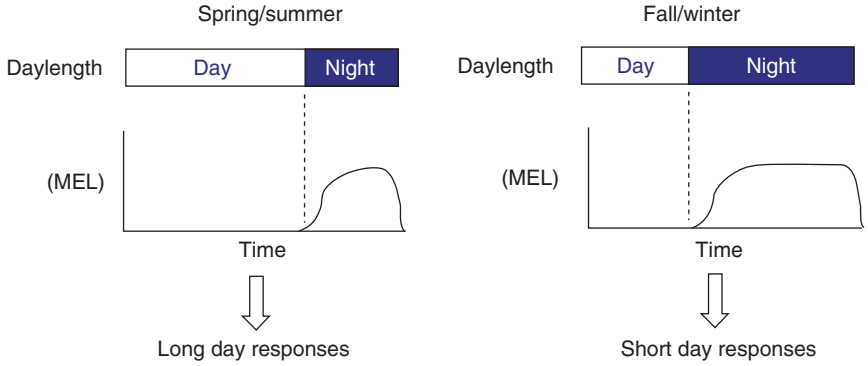

Figure 2 Schematic of photoperiod transduction into durational melatonin signals by the pineal gland. Long days and short nights generate short nocturnal melatonin (MEL) durations of synthesis/secretion (left), whereas short days and long nights generate long nocturnal MEL durations of synthesis/secretion (right).

sheep has been the most thoroughly studied of photoperiodic, circannual mammals with respect to how photoperiod cues interact with the endogenous circannual timing mechanism. Sheep are short-day breeders and hence show reproductive activation and mating during late summer, with lambs being born the following spring. As with hamsters, sheep become refractory to continuous short-day exposure; ewes held under constant short days stop showing ovulatory cycles after several months, hence returning to the spring condition. Unlike hamsters, sheep also become refractory to long days. Ewes held under continuous long days – which are initially inhibitory to reproduction – eventually return to ovulatory cyclicity (hamsters held under continuous long day lengths remain reproductively active indefinitely). Thus, the circannual cycle of the sheep and the seasonal cycle of the hamster both include a role for photoperiodic responsiveness and an endogenous timing mechanism. Whereas both hamsters and sheep can show a spontaneous shift from winter to summer phase (short-day photorefractoriness), sheep show spontaneous shifts from summer to winter phases as well (long-day photorefractoriness). The presence of two distinct phases of 'spontaneous' change in the sheep (compared to a single spontaneous change in hamsters) may help to account for the expression of circannual cycles when this species is held under seasonally constant conditions.

The process of PTM appears to be essentially the same in sheep and hamsters, despite the fact that sheep are short-day breeders with true circannual

rhythmicity, whereas hamsters are long-day breeders that are not circannual. In both species, and in most mammals that have been studied, the duration of the nocturnal elevation of MEL secretion is roughly proportional to the length of the dark phase (or inversely proportional to day length). Thus, the duration of nocturnal MEL secretion contains information about the day length, and this information apparently is decoded and then triggers seasonally appropriate changes in target tissues either by direct effects (e.g., pars tuberalis) or via neural activation (e.g., white adipose tissue).

MEL Receptors Occur in the Brain and Periphery, but Stimulation of Specific Central Sites Appears Necessary for Seasonal Changes in Reproductive Status and Adiposity

Brain MEL receptors were first suggested by the binding of a radioactive MEL analogue (2-[^{125}I]-iodomelatonin; IMEL) to neural tissue. Only a few IMEL binding sites in the brain exist, although MEL binds extensively to many peripheral tissues. IMEL binds in the brains of a wide range of mammals (e.g., sheep; deer; goats; rabbits; laboratory rats; Syrian, Siberian, and European hamsters; white-footed mice; golden-mantled ground squirrels; guinea pigs; nine-banded armadillos; and humans) as well as nonmammalian vertebrates (Atlantic hagfish, lamprey, hedgehog skates, rainbow trout, and clawed toads). Across many mammalian species, one brain and one peripheral site consistently show IMEL binding – the suprachiasmatic nucleus of the hypothalamus (SCN) and the pars tuberalis (PT) of the pituitary gland, respectively. Many of these species show binding in both the SCN and the PT, although there are a few seasonally breeding species in which IMEL binding only is found in the PT (mink and skunks). Gene expression for one of the MEL receptor subtypes, the MEL 1_a receptor (mt_1 receptor) is present in the SCN of Siberian hamsters, laboratory rats, and humans, as well as the PT of the first two. Despite extensive studies of MEL binding sites, the specific functional roles of most of these sites have not been established.

Several Model Systems Are Used to Study Seasonal Changes

Studies of seasonal changes in endocrine-mediated events have revealed a number of interesting model systems that can be exploited to yield information of general application to both seasonal and nonseasonal species. Several of these systems are described next.

There Is a Similarity between the Mechanisms Involved in the Alternation between Breeding and Nonbreeding Seasons Compared with Mechanisms Leading to Puberty Onset

There is considerable similarity between the endocrine mechanisms that lead to a shift from the seasonal anestrous condition to breeding status and mechanisms involved in puberty onset. Two types of mechanisms have been found to account for the decreased secretion of the pituitary gonadotropic hormones (GTH), leuteinizing hormone (LH), and follicle-stimulating hormone (FSH), during seasons of anestrus. In some species, the hypothalamic–pituitary axis shows considerably increased sensitivity to the negative feedback effects of gonadal steroid hormones on secretion of pituitary GTH during seasonal anestrus. The consequence is that blood GTH concentrations are held at very low levels even in the presence of the decreased amounts of steroid hormones produced by the unstimulated gonads. In other species, GTH levels decrease during anestrus even in gonadectomized individuals, indicating that a steroid-independent or direct drive mechanism is responsible for the inhibition of GTH during anestrus. The same two types of mechanisms also have been described as operative in prepubertal mammals, with changes occurring at puberty that lead to an increase in GTH secretion much as similar changes lead to the shift from seasonal anestrus to reproductive activation with the onset of the breeding season. Interestingly, GTH secretion always seems to be largely steroid dependent during the breeding season; that is, during reproductive activation steroid negative feedback always exerts a powerful influence on GTH secretion. This negative feedback system is important for regulating the level of reproductive activity during the breeding season, as has been most clearly illustrated in studies demonstrating the important role of negative feedback to regulate the number of ovarian follicles that develop during an ovulatory cycle and, hence, the number of ova that will be released. During prepubertal development and during the season of anestrus, such precise titration of GTH activity is not required; rather, in these states it is presumably only necessary to achieve an essentially total inhibition of reproductive activity. This can be achieved either via greatly increased sensitivity of the negative feedback regulatory system or by a neural overriding of the system.

Pubertal mechanisms may be triggered at a particular age or in relation to attainment of a sufficient amount of utilizable metabolic fuels, usually lipid

stores. Seasonal activation of the reproductive system generally depends on environmental cues and/or the operation of endogenous timing mechanisms that are relatively independent of age or fat stores. Some cases of puberty incorporate both age-related and environmentally cued components. For example, Siberian hamsters will achieve active reproductive status (puberty) when 40–50 days old if raised in long days; however, if day lengths are short from birth, as in young born at the end of the summer breeding season, puberty will not occur until approximately 5 or 6 months of age, coinciding with the beginning of the next breeding season. The roles of steroid negative feedback and steroid-independent regulation of GTH during the juvenile and pubertal phases are similar to what is seen in seasonal reproductive cycles.

Neuroendocrine Regulation of the Preovulatory Surge of GTH

Considerable research has centered on determining the neuroendocrine pathways that regulate the timing of release of the preovulatory LH surge that is accompanied by an increase in FSH secretion of lesser magnitude. The amount of LH released is in excess of the amount required to lead to the ovulation of all mature follicles in the ovaries. The timing of preovulatory LH release is important in that it determines the time of actual ovulation – approximately 10 h after the LH surge in laboratory rodents – and ovulation time should most adaptively be bracketed by the time period when the female is most likely to mate. In mammals that are coitus-induced (or reflex) ovulators, sensory stimulation associated with mating triggers the release of LH so that ovulation will occur several hours later, when viable sperm are still present in the reproductive tract of the female. In spontaneous (or cyclic) ovulating species, a sharp increase in circulating estrogen, originating from newly matured ovarian follicles, is the endocrine stimulus that leads to release of the preovulatory surge of LH. Less emphasis has been placed on the ability of progesterone to delay the LH surge by one or more days; perhaps this is because in some species progesterone can actually advance the LH surge by a short period of time if administered early on the proestrous day, whereas progesterone administered earlier in the cycle may result in a 1- or 2-day delay of LH release. In rats and other cycling rodents studied in the laboratory, the LH surge occurs approximately 1 day after the initial large increase in circulating estrogen. In some species, the timing of the surge is further refined by the action of a circadian clock in the suprachiasmatic nuclei of the hypothalamus. For example, in laboratory rats and Syrian hamsters, the LH surge occurs during the early afternoon of the day after increased estrogen, with ovulation occurring at approximately midnight or slightly thereafter. This is presumably a time when these nocturnal and relatively nonsocial species are most likely to be successfully inseminated during an encounter with a male. The involvement of a circadian clock in the precise timing of LH release has been clearly demonstrated in studies showing that ovariectomized laboratory rats subjected to continuous administration of estrogen exhibit daily LH surges, occurring at the same time of day as the preovulatory surge that normally occurs only every fourth or fifth day.

Studies in the photoperiodic Syrian hamster have revealed an interesting variation on the theme of triggering LH surges by increased levels of estrogen. Syrian hamsters are long-day breeders and ovulatory cycles cease after 4–7 weeks of exposure to short day lengths (<12.5 h light/day). During the anovulatory phase of the annual cycle, female hamsters exhibit daily surges of LH. These surges occur during the early afternoon at about the same time as the preovulatory surges that occur during the breeding season, and the daily surges are of comparable magnitude to the preovulatory surges. These daily surges are expressed, however, even though estrogen levels are continuously very low. Indeed, the daily LH surges continue unabated in anestrous (short-day) females even following ovariectomy or combined ovariectomy and adrenalectomy. Curiously, daily LH surges are also observed in this species during lactation, another situation in which ovulatory cycles are suspended. It is not known whether the daily LH surges that occur in anestrous and lactating hamsters serve a physiological function. Syrian hamsters have extraordinarily regular, 4-day ovulatory cycles and have been well studied with respect to patterns of secretion of reproductive hormones, so the species may provide an interesting model in which to further examine factors that lead to the triggering of the preovulatory LH surge.

Seasonal Changes in Pelage Are Also Governed by Changes in Hormones Triggered by Alterations in the Photoperiod

Many temperate zone mammals undergo seasonal changes in pelage, molting their summer fur and developing a more insulative coat in the fall, with reversal of this process as spring approaches. In some species, the winter fur not only is more insulative but also may be different in color from the summer coat. One such species is the Siberian hamster,

which molts from a grayish-brown summer coat to a white winter pelage that no doubt provides camouflage during winters of heavy snowfall. Several studies in Siberian hamsters have demonstrated that circulating levels of the pituitary hormone prolactin determine what type of fur will be grown following a molt (Figure 3). Baseline prolactin levels are high during the summer and several fold lower during the winter. These seasonal changes in prolactin levels are determined by day length and by the action of the pineal hormone, MEL, as described later. When prolactin levels are artificially manipulated either by administering drugs that stimulate (e.g., pimozide, the dopamine antagonist) or inhibit secretion of the hormone (e.g., bromocryptine, the dopamine agonist) or by direct administration of prolactin or hypophysectomy to remove its source, the type of fur that is produced is always governed by the levels of prolactin. Thus, summer fur grows when prolactin levels are elevated, and winter fur grows when prolactin is decreased or absent. Many photoperiodic mammals show seasonal changes in prolactin that parallel those observed in Siberian hamsters – with high levels of the hormone under long days and low levels under short photoperiods. Seasonal changes in blood prolactin concentrations also have been implicated in seasonal pelage changes in mink, arctic blue fox, and red deer. Thus, prolactin may prove to be the hormone that usually regulates seasonal changes in fur quality.

It might seem strange that prolactin, which is known for its role in supporting mammary gland function (lactogenesis) in mammals, also should be the hormone that regulates the type of hair produced by hair follicles. Hair follicles, however, are one of a few types of integumentary glands in mammals, including sebaceous glands and sweat glands. Mammary glands are thought to have evolved as specializations of sweat glands so that there is an evolutionary connection between these two types of integumentary glands. Prolactin has also been reported to synergize with testosterone in the regulation of sebaceous gland activity in laboratory rats, suggesting a pattern of prolactin involvement in regulation of different types of integumentary glands and their derivatives, the mammary glands.

It may be that for some species prolactin has actions on hair follicles that are not limited to seasonal pelage changes. Laboratory rats develop a more bristly type of hair at the time of puberty compared to the softer fur of juveniles. Furthermore, hypophysectomy of adult rats results in a return to this prepubertal fine/soft fur. An increase in prolactin associated with achievement of puberty has been implicated in this change in pelage quality as well.

Finally, although prolactin has some documented effects on body fat, its role in the photoperiodic control of body fat appears minimal at best because manipulations of the hormone to produce long-day circulating physiological concentrations of prolactin in short days, and vice versa, do not in and of themselves create short-day- or long-day-like changes in adiposity, respectively.

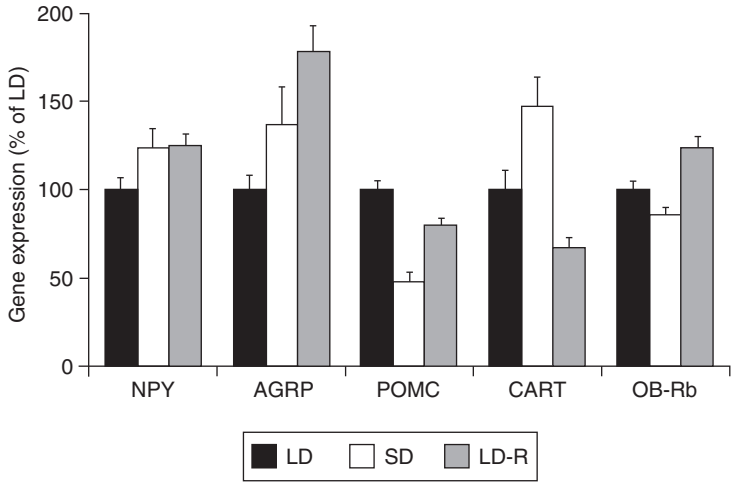

Figure 3 Neuropeptide and leptin receptor gene expression in the hypothalamic arcuate nucleus of adult male Siberian hamsters fed *ad libitum* (ADLIB) in long days (LD) or short days (SD) or held in long days with restricted food to mimic short day body mass changes (LD-R). Means ± SEM as percentages of values in LD-ADLIB hamsters. NPY, neuropeptide Y; AGRP, agouti-related protein; POMC, pro-opiomelanocortin; CART, cocaine and amphetamine-regulated transcript; OB-Rb, leptin receptor long form. Reproduced from Mercer JG and Tups A (2003) Neuropeptides and anticipatory changes in behaviour and physiology: Seasonal body weight regulation in the Siberian hamster. *European Journal of Pharmacology* 480(1–3): 43–50, with permission.

Seasonal Changes in Adiposity and Food Intake Are Also Often Regulated by the Photoperiod and Melatonin

Many animals exhibit seasonal changes in body mass, usually almost exclusively as changes in body fat (cf., collard lemmings). The patterns of seasonal changes in body fat fall into four major categories: (1) species in which body fat/mass increases in the short 'winter-like' days of the fall (e.g., Syrian hamsters and prairie voles) and then decrease, (2) species in which body fat/mass increases from spring to the beginning of fall and then decreases (e.g., Siberian hamsters, meadow and bank voles, deer, and domestic sheep, goats, and cattle), and (3) species that evidence photoperiod-controlled reproductive responses but show no body mass/fat changes (e.g., Turkish hamsters). Because the photoperiod induces a functional gonadectomy of these animals and because gonadal steroids have profound effects on body mass/fat, it may seem that the seasonal changes in body fat are simply secondary to the changes in gonadal steroids. This notion, however, is not supported because the majority of the photoperiod-induced changes in body fat occur independent of the gonads (i.e., they still occur in previously gonadectomized animals subjected to the photoperiod change).

There does not appear to be a hormonal intermediary of MEL for the photoperiod-mediated seasonal changes in adiposity because white adipose tissue lipolysis (fat mobilization) or lipogenesis (fat synthesis) are not triggered in isolated adipocytes (fat cells) by physiological doses of MEL in several species. Instead, at least in Siberian hamsters, it appears that MEL acts through a neural intermediary to decrease body fat in short days. Specifically, injections of PRV, the transneuronal viral tract tracer, into white fat combined with identification of MEL_{1a} receptor mRNA by *in situ* hybridization, reveal PRV + MEL_{1a} co-localization in neurons in brain sites that are part of the sympathetic outflow to white fat (e.g., suprachiasmatic anterior, arcuate, and paraventricular nuclei of the hypothalamus, reuniens/xiphoid, and paraventricular thalamic nuclei). These neuroanatomical data are supported functionally by the increased sympathetic drive to white fat in short-versus long-day-housed Siberian hamsters. Moreover, surgical or chemical denervation of the sympathetic nerves innervating white fat blocks short-day-induced decreases in body fat by Siberian hamsters.

In addition to seasonal body fat changes, often there are also corresponding changes in feeding; however, the changes in food intake do not completely account for the changes in body/lipid mass. In these cases, alterations in energy expenditure complete the accounting of the mechanisms underlying seasonal changes in body/lipid mass. The seasonal changes in food intake have most extensively been studied in Siberian hamsters, in which body/lipid mass most rapidly decrease during the first 5 or 6 weeks of short day exposure, but food intake only significantly decreases later, after approximately 7–12 weeks of short day exposure (**Figure 4**).

The mechanisms underlying these ultimate decreases in food intake by Siberian hamsters are not precisely known, but in laboratory rodents, changes in food intake are often triggered by alterations in orexigenic (increased feeding) and anorexigenic (decreased feeding) peptides. In Siberian hamsters, however, the well-established roles of these peptides in laboratory rats and mice often do not, at least superficially, hold. For example, in laboratory rodents, decreases in body/lipid mass are associated with increases in gene expression of the orexigenic

Figure 4 Siberian hamster (*Phodopus sungorus*) littermates treated identically except the animal on the left has been housed in long 'summer-like' days, whereas the animal on the right has been housed in short 'winter-like' days. Note the change to a white winter pelage in the short day-housed hamster.

peptides neuropeptide Y (NPY) in the arcuate nucleus and melanin concentrating hormone (MCH) and orexin A in the lateral hypothalamus. In addition, there are decreases in two anorexigenic peptides – leptin, mostly synthesized by adipocytes and secreted into the circulation, and arcuate nucleus cocaine- and amphetamine-regulated transcript (CART). In contrast, when short days trigger decreases in body/lipid mass in Siberian hamsters, these relations do not hold in that arcuate nucleus NPY gene expression and lateral hypothalamic MCH and orexin A do not change and arcuate nucleus CART decreases (Figure 5), but circulating leptin concentrations are decreased. In terms of the latter, when leptin is decreased in laboratory rodents, food intake is increased (as is arcuate nucleus NPY gene expression), but in Siberian hamsters food intake is decreased. Some of the disparity between the photoperiodic control of these peptides appears to be reconciled by embracing two views. The first is that unlike the fasting-induced body mass/fat-reduced laboratory rats and mice, the decreases in body mass/fat of short day-exposed Siberian hamsters are seasonally appropriate; thus, because these peptides serve to counteract deviations in body mass/fat, they might not be expected to be engaged because the short day decreases are not true 'deviations' from normal. Indeed, if Siberian hamsters are fasted to promote decreased body/fat mass in either long or short days, they show the same changes in these peptides as do laboratory rats and mice. Second, the unusual inability of decreased circulating leptin concentrations to stimulate food intake in short days and vice versa appears to be due to changes in leptin sensitivity. That is, long day-housed hamsters are leptin insensitive when circulating leptin concentrations are high, whereas short day-housed hamsters are leptin

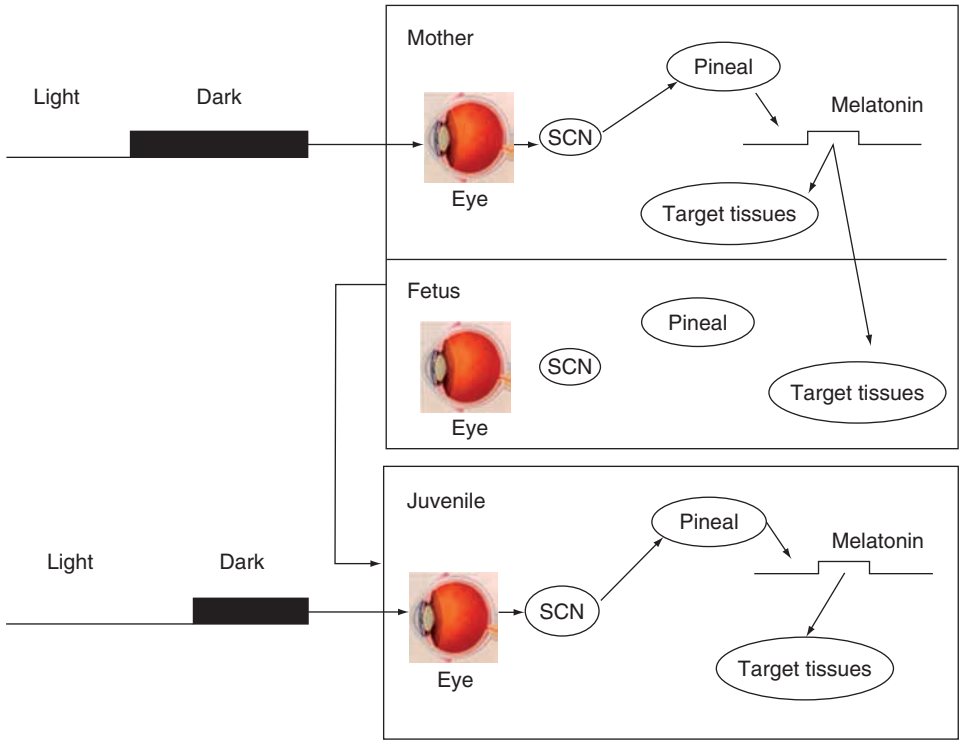

Figure 5 Representation of mechanism by which mother hamster is thought to transmit photoperiod information to her fetuses. Photoperiod information received by the mother's eye is conveyed by a neural pathway (the retinohypothalamic tract) to the circadian pacemaker in the suprachiasmatic nuclei (SCN). Neural signals from SCN traverse a multisynaptic pathway to the maternal pineal gland to regulate the circadian rhythm of melatonin secretion. Maternal melatonin not only acts on target tissues in the mother but also crosses the placenta to act on similar targets in fetus. (The fetal SCN may be one of these targets.) The fetus does not yet have neural connections between the eye and SCN or between SCN and pineal. Hence, the fetal SCN/pineal unit does not generate a melatonin rhythm of its own. At approximately 2 weeks after birth, all appropriate neural pathways have become established and juvenile hamster (bottom) is able to generate its own melatonin rhythm, shaped by the prevailing photoperiod. The day lengths to which the juvenile is exposed will be either longer or shorter than those that were in effect during gestation, and both absolute day length and direction of change influence the trajectory of early reproductive development. Photoperiod information provided from the mother to the fetus may allow the juvenile to more rapidly assess the direction of change in photoperiod and, hence, to 'determine' the time of year. Reproduced from Goldman BD (2003) Pattern of melatonin secretion mediates transfer of photoperiod information from mother to fetus in mammals. *Science STKE* 2003(192): pe29 (DOI: 10.1126/stke.2003.192.pe29), with permission.

sensitive when circulating leptin concentrations are naturally low.

Photoperiod Information Is Transmitted from Mother to Fetus

It has been demonstrated that the photoperiod experienced by a female rodent during late pregnancy can influence the photoperiodic responses of her offspring. MEL passes from the maternal to the fetal circulation so that the rhythm of circulating MEL is similar in mother and fetus. In Siberian hamsters, there is strong evidence that the mother's MEL rhythm acts on fetal MEL target tissues and that this allows the fetus to establish a photoperiodic history that can influence its responses to the photoperiod experienced during forthcoming juvenile life. This early photoperiod history is a particularly important influence on juvenile testis development in male Siberian hamster pups that are born into 'intermediate day lengths' (~14 h of light (14L)). Males born to dams that experienced shorter day lengths (e.g., light:dark cycle 8:16) show relatively rapid testis growth when exposed to 14L postnatally, whereas males born to dams that experienced a longer gestation photoperiod (e.g., light:dark cycle 16:8) show slow testis development in a postnatal 14L day length. This mechanism may allow for rapid and appropriate adjustments with respect to reproductive development. Testis growth is rapid when the photoperiod is increasing, as it would be at the beginning of the breeding season, but is inhibited when the photoperiod is decreasing.

See also: Circadian Function and Therapeutic Potential of Melatonin in Humans; Gonadotropin-Releasing Hormone: GnRH-1 System; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Neuroendocrine Aging: Pituitary–Gonadal Axis in Males; Neuroendocrine Aging: Hypothalamic–Pituitary–Gonadal Axis in Women; Photoperiodic Regulation of Reproductive Cycles; Pineal Gland and Melatonin; Seasonal Changes in Night-Length and Impact on Human Sleep; Seasonal Timing: Neural Mechanisms.

Further Reading

Bartness TJ, Demas GE, and Song CK (2002) Seasonal changes in adiposity: The roles of the photoperiod, melatonin and other hormones and the sympathetic nervous system. *Experimental Biology and Medicine* 227: 363–376.

Bartness TJ and Goldman BD (1989) Mammalian pineal melatonin: A clock for all seasons. *Experientia* 45: 939–945.

Bartness TJ, Powers JB, Hastings MH, Bittman EL, and Goldman BD (1993) The timed infusion paradigm for melatonin delivery: What has it taught us about the melatonin signal, its reception, and the photoperiodic control of seasonal responses? *Journal of Pineal Research* 15: 161–190.

Bittman EL (1993) The sites and consequences of melatonin binding in mammals. *American Zoologist* 33: 200–211.

Elliott JA and Goldman BD (1981) Seasonal reproduction: Photoperiodism and biological clocks. In: Adler NT (ed.) *Neuroendocrinology of Reproduction*, pp. 377–423. New York: Plenum.

Goldman BD (2003) Pattern of melatonin secretion mediates transfer of photoperiod information from mother to fetus in mammals. *Science STKE* 2003(192): pe29 (DOI: 10.1126/stke.2003.192.pe29).

Hastings MH and Herzog ED (2004) Clock genes, oscillators, and cellular networks in the suprachiasmatic nuclei. *Journal of Biological Rhythms* 19: 400–413.

Karsch FJ, Malpaux B, Wayne NL, and Robinson JE (1988) Characteristics of the melatonin signal that provide the photoperiodic code for timing seasonal reproduction in the ewe. *Reproduction, Nutrition and Development* 28: 459–472.

Larsen PJ, Enquist LW, and Card JP (1998) Characterization of the multisynaptic neuronal control of the rat pineal gland using viral transneuronal tracing. *European Journal of Neuroscience* 10: 128–145.

Malpaux B, Viguie C, Skinner DC, Thiery JC, and Chemineau P (1997) Control of the circannual rhythm of reproduction by melatonin in the ewe. *Brain Research Bulletin* 44: 431–438.

Mercer JG and Tups A (2003) Neuropeptides and anticipatory changes in behaviour and physiology: Seasonal body weight regulation in the Siberian hamster. *European Journal of Pharmacology* 480(1–3): 43–50.

Morgan PJ, Ross AW, Mercer JG, and Barrett P (2003) Photoperiodic programming of body weight through the neuroendocrine hypothalamus. *Journal of Endocrinology* 177: 27–34.

Nelson RJ, Demas GE, Klein SL, and Kriegsfeld LJ (1995) The influence of season, photoperiod, and pineal melatonin on immune function. *Journal of Pineal Research* 19: 149–165.

Reiter RJ (1993) The melatonin rhythm: Both a clock and a calendar. *Experientia* 49: 654–664.

Reppert SM (1997) Melatonin receptors: Molecular biology of a new family of G protein-coupled receptors. *Journal of Biological Rhythms* 12: 528–531.

Seasonal Timing: Neural Mechanisms

M A Hofman, Netherlands Institute for Neuroscience, Amsterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The rotation of the Earth and the Earth's orbit around the sun result in daily and seasonal fluctuations in illumination, temperature, and humidity. It would be appropriate for an organism to have its internal organization and behavioral activity synchronized with this cyclic environment to provide an adaptive optimum. Periodic fluctuations, in fact, are found to be a fundamental feature of living systems and can be observed throughout the plant and animal kingdoms. Although a specific and constant phase relationship between the solar day and these rhythms is maintained, endogenous biological rhythms are not directly driven by the environmental light–dark cycle or by any other zeitgebers. Numerous chronobiological studies, in which environmental conditions were rigorously controlled, showed that organisms not only follow the rhythms of the diurnal and seasonal cycle but also possess their own intrinsic rhythm. This article focuses on the cellular and molecular mechanisms underlying the complex interaction between the circadian timing system and life in a photoperiodic environment.

The Biological Clock as Calendar

Suprachiasmatic Nucleus as Circadian Pacemaker

In mammals, the suprachiasmatic nucleus (SCN) of the hypothalamus is considered to be a critical component of a neural system implicated in the timing of circadian processes. The pacemaker of the SCN oscillates with a near 24-h period and is entrained to the diurnal light–dark cycle. The SCN is located on either side of the midline of the brain just dorsal to the optic chiasm. Since the discovery of the SCN as the site of the master circadian pacemaker, many attempts have been undertaken to unravel the mechanisms underlying its endogenous circadian rhythmicity. In particular, lesion and transplantation experiments of the SCN and *in vitro* slice studies have provided firm evidence for its biological clock characteristics. Furthermore, consistent with its role in circadian timing, investigations in rodents and nonhuman primates suggest that the SCN is the locus of the brain's endogenous calendar, enabling organisms to anticipate seasonal environmental changes (Figure 1).

Recent studies indicate that the biological clock is a multioscillator system, consisting of independently operating subdivisions located in the core and shell of the SCN. The core is located in the ventral region of the SCN above the optic chiasm, receives primary and secondary visual afferents, and predominantly contains neurons producing vasoactive intestinal polypeptide (VIP). The shell is located in the dorsal region of the SCN, largely surrounds the core, receives input mainly from nonphotic sources, and is composed predominantly of neurons containing arginine vasopressin (AVP). A similar topographic organization of SCN efferent projections was found, indicating that the nature of the signal conveyed to areas receiving core or shell projections varies as a function of the subdivision from which the innervation is derived.

Melatonin and Photoperiodic Time Measurement

In mammals, the SCN–pineal complex is the principal neural substrate involved in the regulation of annual cycles. In fact, many of the functions that exhibit seasonal cycles in mammals, such as sexual behavior, energy metabolism, food intake, and hibernation, are regulated by this timing system in the brain. Day length (photoperiod) has been shown to be the strongest synchronizer of seasonal functions in most species.

The periodic synthesis and secretion of the hormone melatonin exhibited by the pineal gland is generated by oscillators in the SCN and is communicated to the pineal via a complex multisynaptic pathway. Disruption of any portion of this pathway from the SCN to the pineal gland abolishes melatonin rhythmicity. These findings strongly suggest that the endocrine activity of the mammalian pineal is under neural control and receives a major input from the SCN. This means that, in addition to its role as a circadian pacemaker, the SCN may also be involved in the seasonal timing of a number of physiological and behavioral processes by regulation of the photoperiod-dependent changes in melatonin secretion (Figure 2).

Melatonin is secreted at night, and the duration of secretion is inversely related to day length; thus, photoperiodic information is encoded reciprocally in the duration of nocturnal melatonin secretion, which becomes longer in winter and shorter in summer. Several studies in hamster and sheep, employing a paradigm of timed daily infusions of melatonin in pinealectomized animals, indicate that photoperiodic information is indeed encoded in the melatonin signal. Thus, in both short-day breeding sheep and

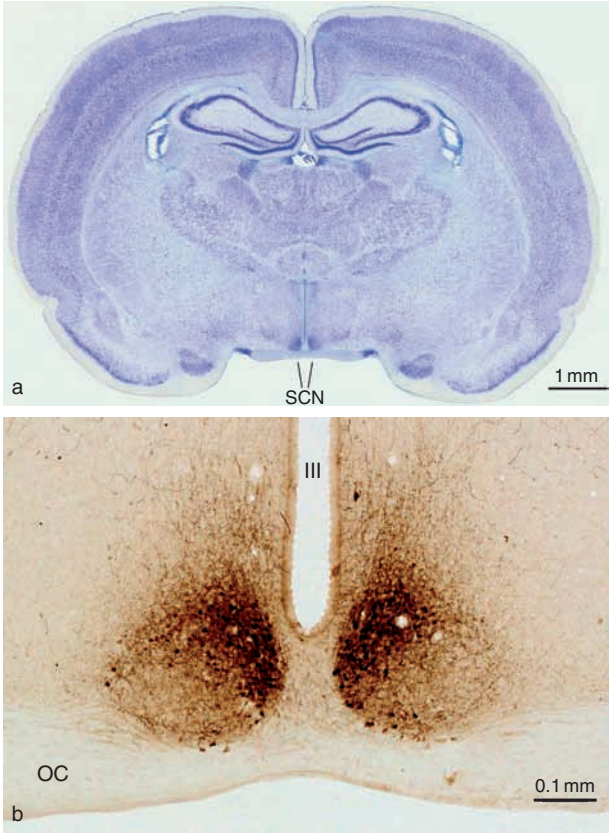

Figure 1 Photomicrograph of the rat brain: (a) frontal section through the brain at the level of the suprachiasmatic nucleus (SCN), the site of the brain's circadian clock and calendar; (b) frontal section through the midportion of the rat SCN stained with an antibody against vasopressin (AVP). The principal mass of AVP-immunoreactive neurons lies in the dorsal region of the SCN, from which strands of loosely packed cells descend on all sides, surrounding a central unlabeled area. OC, optic chiasm; III, third ventricle. Courtesy of CG van Eden, Netherlands Institute for Neuroscience.

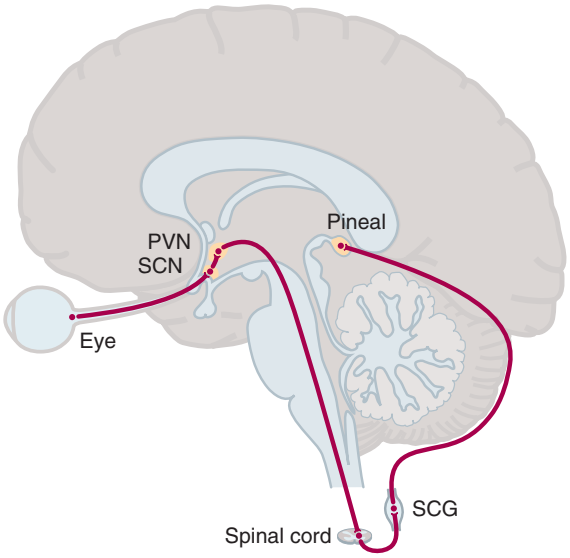

Figure 2 Diagram of the human brain (midsagittal section) showing the neural pathways of the circadian timing system by which photic information reaches the SCN and the pineal gland. PVN, paraventricular nucleus; SCG, superior cervical ganglion; SCN, suprachiasmatic nucleus.

long-day breeding hamsters, short-duration melatonin infusions led to the long-day responses typical for the given species, whereas long-duration infusions evoked responses that are associated with short days. These results suggest that day-length signals are encoded in the duration of nocturnal melatonin secretion. In other words, the circadian system regulates the pattern of melatonin secretion, and the melatonin pattern serves as a humoral signal that conveys day-length information.

It has been suggested that two light-entrainable circadian oscillators are responsible for regulating the duration of nocturnal melatonin secretion. Thus, a dusk oscillator may trigger the onset of melatonin secretion and a dawn oscillator might determine the termination of pineal activity. If the duration of melatonin secretion does indeed depend on the phase relationship between dawn and dusk oscillators, then there is a direct link between this mutual phase relationship and the endocrine coding of photoperiod via the duration of nocturnal melatonin secretion. It means that the circadian clock should exhibit sensitivity to phase resetting by melatonin during the crepuscular domains, times of the day which are most affected by changes in photoperiod. Indeed, the circadian clock in the SCN is most sensitive to melatonin levels at dusk and dawn.

The cellular and molecular mechanisms through which the melatonin signal is decoded to drive or synchronize photoperiodic responses remain unclear. Potential target sites in the mammalian brain, where the effects of melatonin are mediated via high-affinity melatonin receptors, are located in the pars tuberalis of the pituitary and in the SCN. The presence of melatonin receptors in the SCN suggests that melatonin feeds back to regulate circadian pacemaker function.

Seasonal Cycles in Humans

Most of the anatomical and molecular substrates of the system that encodes changes in photoperiod in the duration of melatonin secretion, and the receptor molecules that read this signal, have been shown to be conserved in monkeys and humans, and the functions of this system appear to be intact from the level of the retina to the level of the seasonal-dependent duration of the melatonin signal. Although photoperiodic seasonal breeding has been shown to occur in monkeys, it remains unclear whether photoperiod

and mediation of its effects by melatonin influence human reproduction.

Historical and experimental evidence indicates that human responses to seasonal changes in the natural photoperiod may have been more robust prior to the Industrial Revolution and that subsequently they have been increasingly suppressed by alterations of the physical environment. In fact, some observations in humans suggest that the seasonal fluctuations in pineal activity in modern urbanized human beings, who are living in an artificial photic environment, resemble that of mammals living under fixed light–dark conditions. It seems as if the extension of the natural photoperiod with artificial light to maintain long summer-type photoperiods throughout the year has significantly altered the temporal organization of many physiological and behavioral functions in humans.

During the past decade, several studies were conducted to determine whether variations in photoperiod and light intensity affect the morphology and neuronal activity of the human SCN by studying brains obtained at autopsy. The AVP- and VIP-producing neurons in the SCN, for example, were found to exhibit distinct seasonal oscillations with an asymmetrical bimodal waveform. The AVP cycle had a maximum in early autumn, a lower plateau in winter, a second peak in early spring, and a decline beginning in late spring, leading to a nadir in the period from May through July. The population of VIP-expressing neurons also reached a maximum in early autumn, a lower plateau in winter, a second peak in early spring, a short nadir in April to May, and another plateau in summer (Figure 3).

There is also convincing evidence for seasonal variations in monoamine concentrations in the human brain, based on postmortem studies and single-photon emission computed tomography (SPECT). Among the monoamines investigated, the level of serotonin (5-HT) in the human hypothalamus showed the most striking seasonal rhythm, with a maximum in autumn and a minimum in early winter. No corresponding seasonal fluctuations in the concentration of hypothalamic 5-hydroxyindoleacetic acid (5-HIAA) were found. Furthermore, the hypothalamic dopamine level showed a more complex, but less distinct,

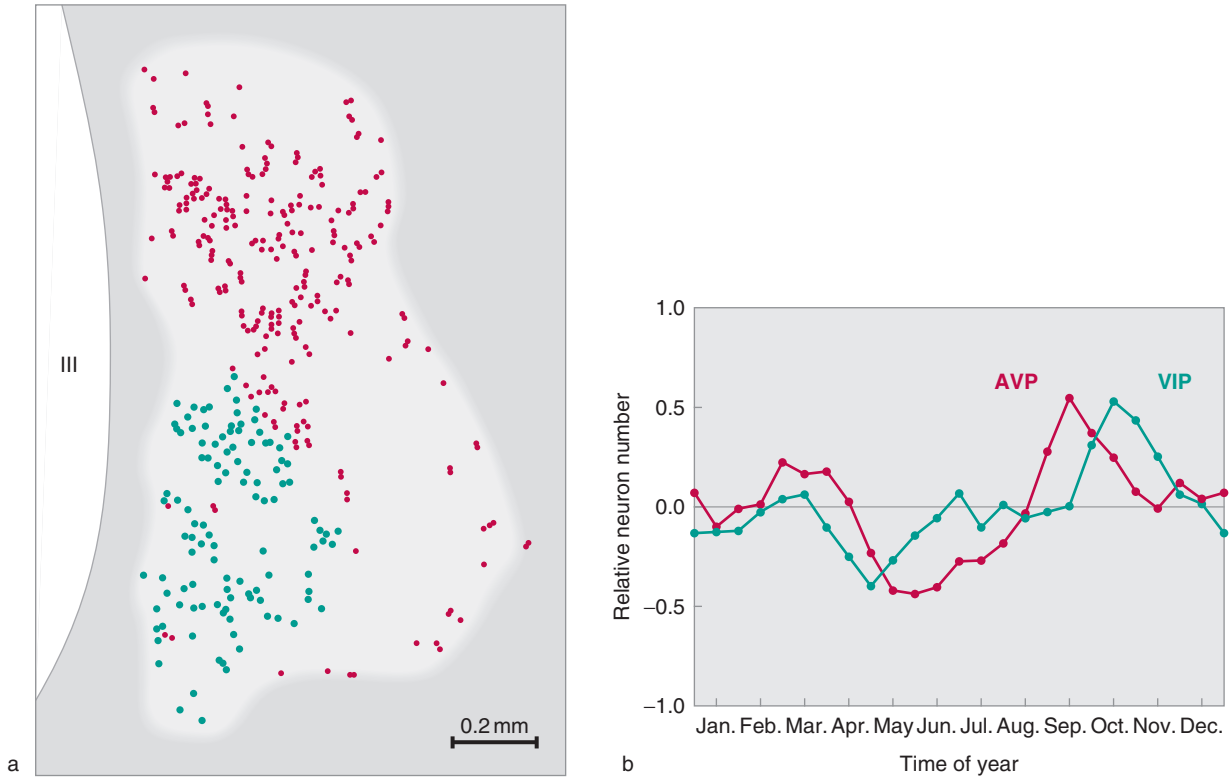

Figure 3 Arginine vasopressin (AVP) neurons (red dots) and vasoactive intestinal polypeptide (VIP) neurons (green dots) in the human suprachiasmatic nucleus (SCN): (a) distribution; (b) normalized values of the seasonal cycles. Notice the bimodal waveform of the cycles. III, third ventricle. Modified from Hofman MA and Swaab DF (2002) A brain for all seasons: Cellular and molecular mechanisms of photoperiodic plasticity. In: Hofman MA, Boer GJ, Holtmaat AJGD, Van Someren EJW, Verhaagen J, and Swaab DF (eds.) *Progress in Brain Research*, *Vol. 138: Plasticity in the Adult Brain: From Genes to Neurotherapy*, pp. 255–280. Amsterdam: Elsevier, with permission from Elsevier.

seasonal profile, with two peaks and two nadirs, like the seasonal profiles found in the synthesis of AVP and VIP in the human SCN. How these periodic fluctuations in 5-HT and dopamine – neurotransmitters that innervate the SCN – are functionally related to the seasonal oscillations in the biological clock is not known.

Seasonal Affective Disorder

Seasonal affective disorder (SAD) is a condition of regularly occurring depressions in winter with a remission the following spring or summer. In addition to depressed mood, the patients tend to experience increased appetite and an increased duration of sleep during the winter. SAD is a relatively common condition, affecting 1–3% of adults in temperate climates, and it is more prevalent in women. The pathological mechanisms underlying SAD are incompletely understood. Certain neurotransmitters have been implicated; a dysfunction of the serotonin system in particular has been demonstrated by a variety of approaches. Furthermore, epidemiological evidence suggests that SAD and seasonal fluctuations in human reproduction are governed by a common biological process.

Although the role of the circadian timing system in SAD is not clarified yet, there are strong indications that the circadian pacemaker in SAD patients is delayed relative to the sleep–wake cycle. According to this phase-delay hypothesis, the symptoms of SAD should improve if the circadian rhythms can be phase advanced. There is some experimental support for this idea. SAD can be treated successfully with light therapy, in which the patient is exposed to 2000–10 000 lux for a period of 30–120 min daily during the late autumn or early winter. Other forms of light treatment and pharmacotherapy are currently being tested.

A Clock for All Seasons

Models of Seasonal Timing

Studies of the formal properties of photoperiodic mechanisms have revealed that most organisms use circadian oscillators to measure day length. Two types of paradigms – external coincidence (EC) and internal coincidence (IC) models – have been proposed to account for photoperiodic time measurements by a circadian mechanism. Both models postulate that the timing of light exposure, rather than the total amount of light (as in the classical model), is critical to the organism's perception of day length.

In the EC model for photoperiodic time measurement, light has two distinct and separate roles. First, light is the major diurnal zeitgeber for the entrainment of the circadian system; second, light is the photoperiodic zeitgeber for the entrainment of the biological clock to changes in day length. The model postulates that the photoinductive phase of the circadian cycle is subdivided in a relatively continuous fashion, enabling organisms to discriminate between small variations in day length. Light falling at one time point during the photoinductive phase might be interpreted by the circadian system as representative of a 12 h photoperiod light falling at a different time point might be interpreted as a 14 h photoperiod, and so forth. It is difficult to obtain a complete and precise role of the photoinductive phase, but it is clear that it is largely included within the portion of the circadian cycle where light is able to evoke phase shifts (i.e., the subjective night). For an EC system to discriminate unerringly among various day lengths, it is crucial that the organism be appropriately entrained to the 24 h light–dark cycle. Thus, a change in the phase angle between the circadian oscillator and the light–dark cycle might be expected to result in an alteration in the interpretation of day length.

In the IC model for photoperiodic time measurement, the phase relationship between two circadian oscillators varies as a function of day length. In the IC model, it is the degree of the phase difference between these oscillators that determines the type of photoperiodic response. Light is the zeitgeber for the entrainment of both oscillators. However, the two oscillators may be differentially phase shifted by a given light pulse, so that the phase relationship between them can vary depending on the parameters of the photic stimuli. The phase of each oscillator is also influenced via mutual coupling between the two oscillators, and the strength of the coupling relation may vary in response to a number of factors, including photoperiod.

The EC and IC models have frequently been treated as if they were distinct alternatives. The models are clearly different in that internal coincidence requires the interaction of two oscillators, whereas external coincidence requires only one. However, an IC system can behave in a manner fully consistent with the EC model. Thus, it could be that light falling during the photoinductive phase (EC) evokes long-day responses because that is the phase at which light produces differential phase shifts of two oscillators (IC), causing them to assume a long-day phase relationship to one another.

The Dual Oscillator Model

A hypothetical principle based on two coupled oscillators was proposed more than 30 years ago by Pittendrigh and Daan, who suggested that the mammalian circadian system consists of a morning oscillator (M) and an evening oscillator (E). One of the central

propositions of the original M–E oscillator model was that with M locking on the dawn and E on the dusk, the phase relationship of M and E in the coupled system reflects the day length to which the animal has been exposed. This 'clock for all seasons' could simultaneously be exploited to measure day length, and this information could be used to adjust adaptively to seasonal timing.

From a functional perspective the main attraction of the dual oscillator model, based on IC, remains that it accounts both for the adjustment of circadian organization to season and latitude and for the measurement of day length in photoperiodic responses. The mechanism of this neural calendar has remained elusive, however. Two main sets of facts, emerging from electrophysiological and molecular studies, now simultaneously suggest how the mammalian circadian clock may indeed be built from M and E oscillators, and how the phase relationship of these oscillators reflects the photoperiod, and thus time of year, to which an animal has been exposed.

In addition to the involvement of circadian oscillators in photoperiodic time measurement, there are other types of biological timing mechanisms that are generally associated with photoperiodism. Some photoperiodic species exhibit endogenous circannual rhythms when held for long periods of time under seasonally constant conditions, that is, constant photoperiod and temperature. For circannual rhythms, changes in day length may be viewed as a zeitgeber for entrainment of the putative circannual oscillator. However, the anatomical substrate for a circannual timing system has not been unequivocally identified in any mammalian species. Therefore, the concept of a circannual oscillator remains a hypothetical construct to explain the expression of circannual rhythms. It remains possible, however, that overt circannual rhythms result from interactions among multiple neural and endocrine systems rather than being a product of one discrete timing system, as for circadian oscillations.

Molecular Genetics of Biological Timing

Circadian Timing

As internal systems, circadian clocks arise from the expression and temporally regulated activity of specific genes and gene products. Over the past decade there have been significant advances in the understanding of the molecular basis of circadian clocks. A number of clock genes have been identified that play a key role in the generation of circadian rhythms. These clock genes are expressed in a circadian manner, with peak values at different times. The clock's core consists of autoregulatory feedback loops with oscillating levels of nuclear proteins that negatively regulate the transcription of their own genes and a positive feedback loop involving the rhythmic regulation of CLOCK–BMAL1 heterodimers.

Based on the oscillation patterns of the *Period* genes within the SCN (high during the day, low during the night), the mammalian clock was predicted to be a day-active clock (in which the oscillating components of the clock are at higher levels during daytime than during the night). However, analysis of the phase response curve (PRC) to light shows that the circadian clock is more greatly affected at nighttime than during the day (daytime is within a dead zone, during which light exposure has little or no effect on the phase of the clock). Thus, during the dead zone of the PRC, the oscillating clock molecules are abundant, and during the active phase of the PRC, these molecules are scarce. What follows from this is the prediction that this clock is phase shifted (and thus entrained) not by degradation of clock molecules following zeitgeber exposure but rather by the induced expression of clock molecules.

Seasonal Timing

Recently, the hypothesis was advanced that the circadian pacemaker in the mammalian SCN is composed at the molecular level of a nonredundant double complex of circadian genes (*Per1–Cry1* and *Per2–Cry2*). Each one of these sets is sufficient for the maintenance of endogenous rhythmicity and thus constitutes an oscillator, and each of these oscillators has slightly different temporal dynamics and light responses. The *Per1–Cry1* (or morning (M)) oscillator is accelerated by light and decelerated by darkness, and the *Per2–Cry2* (or evening (E)) oscillator is decelerated by light and accelerated by darkness. Changes in the activity of these M and E oscillators give rise to the adaptive tuning of the endogenous behavioral program to day length. In other words, the phase relationship of M and E in the coupled system reflects the day length to which the animal has been exposed. The phase angle between oscillator and zeitgeber tells us something about the oscillator's intrinsic period relative to the external 24 h period. This suggests that changes in day length involve changes in the intrinsic period of the two component cycles, whereas the external period remains at 24 h (**Figure 4**).

There must be some kind of coupling for the two components of the oscillating system in the SCN to maintain their mutual phase relationship under free-running conditions. In nature, the coupling force opposes the differential control of each oscillator by light because it acts to revert to this particular phase relationship from which the light responses pull

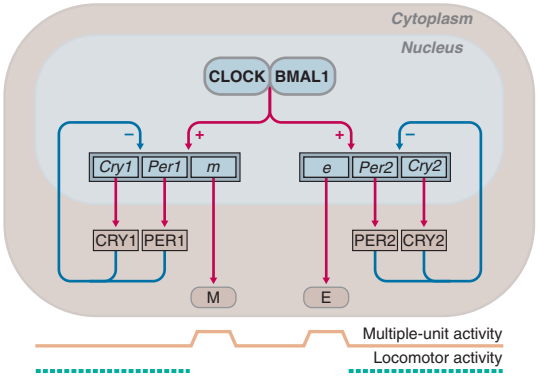

Figure 4 Dual oscillator model of photoperiodic timing proposed by Daan et al. In the cell nucleus, the transcription of the clock genes *Per* and *Cry* is promoted by the CLOCK–BMAL1 complex and suppressed by PER and CRY. The two hypothetical main pathways in the complex model are indicated as separate oscillatory loops, although they interact both via CLOCK–BMAL1 and via multimerization of PER and CRY between the loops. The two parts have slightly different temporal dynamics, leading to differences in the timing of expression of putative clock-controlled genes, indicated here as *m* and *e*. This, in turn, leads to two bouts of multiple-unit activity, M (morning) and E (evening), together inducing or suppressing behavioral activity during part of the 24-h period (lower part of the panel). The phase relationship of the M oscillator and E oscillator is thought to reflect the photoperiod to which an organism has been exposed.

the system away. There are various ways in which the coupling of these oscillators may be envisaged, depending on whether the oscillators are restricted to different populations of neurons in the SCN or whether they are co-localized in the same neuron. In the coupling interaction, it is likely that there is some degree of asymmetry in the strengths of the two oscillators.

At the moment, however, there is no clear evidence that one oscillator is more dominant in synchronizing the circadian clock with the 24-h day than the other. In nocturnal animals, however, under various photoperiods both pineal and SCN rhythms are locked to the morning light onset rather than to the evening light offset. It seems that in these animals, at least for the entrainment of circadian rhythms, the M oscillator is the dominant oscillator and the E oscillator the less flexible slave oscillator. Recent findings in fruit flies support this hypothesis, showing that the oscillators that control morning and evening peaks in locomotor activity reside in distinct neuronal cell populations and that the M oscillator resets the E oscillator once a day. In between these resets, the E oscillator free-runs with its own faster intrinsic period.

The data on human sleep–wake cycles under different photoperiods, furthermore, demonstrate that sleep onset and offset lock on to dusk and dawn along with the melatonin profiles, whereas both rhythms may become bimodal in their nocturnal distribution. These data, along with observations on the bimodal circadian profile of neuronal activity in the human SCN, lend support to the idea that eventually the M–E model will be applicable also to our own species.

Modeling the Molecular Calendar

The M–E model for photoperiodic timing understandably focuses very much on the novel molecular discoveries that support the dual oscillator view of the clockwork, and, as such, the notion of a molecular calendar in the SCN based on morning and evening oscillators is an attractive hypothesis. The M–E model, however, may be too simplistic in view of the complexity of the generation of circadian oscillations. The genetic details of the molecular calendar may be different from the models proposed so far, and other clock genes might be involved in the encoding of season by the circadian clockwork.

A complicating factor could be that there are species differences in the way photoperiodic modulations are encoded within the circadian clockwork. Recent findings, for example, support the conclusion that in sheep, unlike hamsters or rats, an EC model best explains the way photoperiod affects the temporal organization of clock gene expression in the circadian pacemaker, whereas an IC model best explains the way melatonin affects the phasing of clock gene expression in the pars tuberalis of the pituitary to mediate the photoperiodic control of physiological functions and behavior.

Although at present the details of how season is encoded by the circadian clockwork are not known, it is clear that there are properties in the behaviors of the clock genes that could be interpreted as M and E oscillators. The phase relationship of these oscillators or the temporal organization of the expression of their genes reflects the photoperiod, and thus the time of year, to which an animal has been exposed.

The dual oscillator model has obvious implications for the evolution of circadian timing. The rotation of the Earth provides two precise timing signals, sunrise and sunset, rather than a single one. It might well be that life that evolved on its surface has developed ways to exploit both signals, for the fine-tuning of its temporal organization over the 24 h as well as for adapting its physiology and behavior to the seasonal fluctuations in day length. Experimental data consistent with the model have formed the cornerstone of our understanding of how the duration of the nocturnal secretion of melatonin by the pineal gland encodes day length and how, at some level, the circadian waveform generated by the SCN is responsible for driving the pineal.

Finally, there is the intriguing question of whether humans have a seasonal timing system like their

circadian one. The distinct seasonal oscillations in neuropeptide expression in the human SCN, together with the seasonal rhythms in monoamine and hormone concentrations in the hypothalamus and other parts of the human brain, strongly suggest that humans, like other mammals, have a biological calendar.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Function and Therapeutic Potential of Melatonin in Humans; Genetics of Circadian Disorders in Humans; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Photoperiodic Regulation of Reproductive Cycles; Photoreceptors and Circadian Clocks; Seasonal Hormonal Changes and Behavior; Seasonal Changes in Night-Length and Impact on Human Sleep.

Further Reading

Daan S, Albrecht U, Van der Horst GTJ, et al. (2001) Assembling a clock for all seasons: Are there M and E oscillators in the genes. *Journal of Biological Rhythms* 16: 105–116.

Goldman BD (2001) Mammalian photoperiodic system: Formal properties and neuroendocrine mechanisms of photoperiodic time measurement. *Journal of Biological Rhythms* 16: 283–301.

Gorman MR, Goldman BD, and Zucker I (2001) Mammalian photoperiodism. In: Takahashi JS, Turek FW, and Moore RY (eds.) *Handbook of Behavioral Neurobiology, Vol. 12: Circadian Clocks*, pp. 481–508. New York: Plenum-Kluwer.

Hastings MH and Follett BK (2001) Toward a molecular biological calendar. *Journal of Biological Rhythms* 16: 424–430.

Hastings MH and Herzog ED (2004) Clock genes, oscillators, and cellular networks in the suprachiasmatic nucleus. *Journal of Biological Rhythms* 19: 400–413.

Hofman MA (2004) The brain's calendar: Neural mechanisms of seasonal timing. *Biological Reviews of the Cambridge Philosophical Society* 79: 61–77.

Hofman MA and Swaab DF (2002) A brain for all seasons: Cellular and molecular mechanisms of photoperiodic plasticity. In: Hofman MA, Boer GJ, Holtmaat AJGD, Van Someren EJW, Verhaagen J, and Swaab DF (eds.) *Progress in Brain Research, Vol. 138: Plasticity in the Adult Brain: From Genes to Neurotherapy*, pp. 255–280. Amsterdam: Elsevier.

Magnusson A and Boivin D (2003) Seasonal affective disorder: An overview. *Chronobiology International* 20: 189–207.

Schwartz WJ, De la Iglesia HO, Zlomanczuk P, and Illnerová H (2001) Encoding Le Quattro Stagioni with the mammalian brain: Photoperiodic orchestration through the suprachiasmatic nucleus. *Journal of Biological Rhythms* 16: 302–311.

Stoleru D, Peng Y, Nawathean P, and Rosbash M (2005) A resetting signal between *Drosophila* pacemakers synchronizes morning and evening activity. *Nature* 438: 238–242.

Sumová A, Bendova Z, Sladek M, Kovacikova Z, and Illnerová H (2004) Seasonal molecular timekeeping with the rat circadian clock. *Physiological Research* 53: S167–S176.

Tramontin AD and Brenowitz EA (2000) Seasonal plasticity in the adult brain. *Trends in Neurosciences* 23: 251–258.

Wehr TA (1996) A clock for all seasons in the human brain. In: Buijs RM, Kalsbeek A, Romijn HJ, Pennartz CMA, and Mirmiran M (eds.) *Progress in Brain Research, Vol. 111: Hypothalamic Integration of Circadian Rhythms*, pp. 321–342. Amsterdam: Elsevier.

Wehr TA, Aeschbach D, and Duncan WC (2001) Evidence for a biological dawn and dusk in the human circadian timing system. *Journal of Physiology* 535: 937–951.

Zucker I (2001) Circannual rhythms: Mammals. In: Takahashi JS, Turek FW, and Moore RY (eds.) *Handbook of Behavioral Neurobiology, Vol. 12: Circadian Clocks*, pp. 509–528. New York: Plenum–Kluwer.

Second Language Acquisition

P Indefrey and D J Davidson, Max Planck Institute for Psycholinguistics and F. C. Donders Centre for Cognitive Neuroimaging, Nijmegen, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction: Second Language Acquisition and the Brain

Speaking a second language (L2) means having acquired a new inventory of phonemes (speech sounds), new words, and new inflectional and syntactic rules. Because this new knowledge must be stored somewhere in the brain, it does not require neuroimaging experiments to conclude that a bilingual speaker's brain must be in some way different from that of a monolingual speaker. The question is rather: Can the difference between the brain of a bilingual speaker and a monolingual speaker tell us something about the functional organization of a bilingual language system? Quite different answers are possible, ranging from 'more of the same,' meaning that at all processing levels the neural substrates for first language (L1) and L2 processing completely overlap, to 'radically different,' as suggested by early hypotheses that an L2 relies on the right rather than the left hemisphere of the brain. In between, there are hypotheses suggesting that L2s rely predominantly on a subset of L1 processes, such as storage and declarative processing, but not rule or procedural processing. Such hypotheses predict that a subset of the brain regions involved in L1 processing, namely those subserving rule or procedural processing, should not be recruited for L2 processing.

Neuroimaging studies on bilingual speakers are thus potentially of high relevance for our understanding of the functional architecture of a bilingual language system. At the same time, they face considerable methodological problems. In experiments involving a task, differences in neural activity patterns may be explained by general differences in task performance rather than the components of the task that the experiment was designed to assess. Given that most speakers are less proficient in their L2 than in their L1, however, equal task performance is rarely achieved. A second problem is that experimental variables such as 'age of L2 acquisition' are almost inevitably confounded with other variables such as 'L2 proficiency' or vice versa. Often, it becomes a matter of interpretation which of the observed neuronal activation differences are due to experimental variables and which are due to possible confounds. In this situation, comparisons across studies can be helpful because they may indicate which activation patterns or electrophysiological signatures are reliably found in certain tasks when performed in the L2, irrespective of particular L2 speaker characteristics. They may also show brain activity that is reliably found for subsets of studies testing particular types of L2 speakers.

This article therefore focuses on experimental tasks that have been used in a number of hemodynamic (i.e., positron emission tomography (PET) and functional magnetic resonance imaging (fMRI)) and electrophysiological studies of L2 processing, such that an assessment of the reliability of findings is possible. Where possible, we attempt to identify L2 speaker characteristics that seem to influence observed brain activation patterns.

Hemodynamic Correlates of L2 Acquisition

Hemodynamic studies on the neural correlates of L2 processing have been conducted on all levels of linguistic structure from phonology to discourse. At higher linguistic levels, most studies have investigated the processing of an L2 after years of learning rather than L2 acquisition as such. These studies provide evidence of changes in functional neuroanatomy that must have occurred at some time point during L2 acquisition rather than during the study. Experiments with true learning (longitudinal) designs are relatively rare and mostly found at the phonemic level.

Phonemic Learning

A number of studies have investigated the neural consequences of learning a novel phonemic contrast using within-subject training designs. These studies are very instructive because they illustrate a methodological problem of neurocognitive research on bilingualism and possible ways to overcome the problem. To detect neural correlates of phonemic learning, participants performing phoneme classification or recognition tasks were scanned with fMRI before and after a training phase (several hours spread over 1–4 weeks), during which they learned nonnative phonemic contrasts (e.g., Japanese speakers learning the English *r/l* contrast). All studies tested for differences between pre- and posttraining fMRI measurements. The number and anatomical location of regional changes in activation varied considerably across studies, but there were five regions that were

replicated across studies: the left and right posterior inferior frontal gyri, the left mid superior temporal gyrus (auditory cortex and planum temporale), the left superior parietal lobule, and the left caudate nucleus. According to the logic of the design, it is tempting to conclude that these regions are involved in the processing of the novel phonemic contrast. Note, however, that by definition successful training resulted in improved performance such that the two fMRI measurements were no longer comparable in this respect. Therefore, it cannot be excluded that the observed regions simply reflect unspecific effects, such as reduced attention or effort, lower error rates, and so forth. To be able to assess which of the observed activation differences were due to phonemic processing and which were unspecific effects of higher performance or even simple repetition effects, some studies included a control condition by training a second, easy phonemic contrast that was similar to a native language contrast. Because performance in this condition is sometimes even higher than that on the critical nonnative contrast, it is possible to interpret activation changes that only occurred in the critical condition as being specific for learning a nonnative contrast. When taking into account these additional data, commonalities between the findings of the different studies are further reduced.

Remaining regions that might be specifically recruited for learning a nonnative phonemic contrast are the left inferior frontal gyrus and the left caudate nucleus, but other data suggest that posttraining learner brain activation may be very much like native speaker brain activation. One possible reason for such seemingly discrepant findings is a potentially nonlinear relationship between learning and hemodynamic activation. The activation level of regions that are recruited for a difficult task (e.g., recognizing a nonnative contrast) may be initially high but fall off again after some level of performance is reached. This issue is further discussed later.

Word-Level Processing

The two most frequently used tasks for the study of word production are picture naming and word generation (typically generation of words that start with a given letter or fit semantially to a given stimulus word). Both tasks involve the core components of word production (lemma retrieval, word form retrieval, syllabification, and phonetic encoding) but differ with respect to the processes that are employed to come up with a lexical concept to be produced. No study comparing picture naming or word generation in the L1 and L2 of bilingual speakers has reported stronger activation when the task was performed in the native language. Stronger activation for the L2 has been reported in a small but reliable number of studies in the left (for word generation also the right) posterior inferior frontal lobe (**Figure 1**). Studies reporting stronger activation of this region in L2 picture naming had participants with later L2 onset and lower L2 proficiency compared to studies reporting no differences between L2 and L1 picture naming. Stronger activation of the left posterior inferior frontal gyrus was also found in a verb-generation study with quadri-lingual participants for languages that were learned later and spoken less proficiently. These results show that at a group level, L1 and L2 word production seem to engage the same cortical areas, and that L2 speakers with late onset or lower proficiency may recruit one or more of the areas more strongly. Note, however, that hemodynamic studies show all brain areas that are activated. They do not provide evidence of which brain areas are actually necessary for word production in individual speakers. Evidence for necessary areas comes from a study in which different cortical areas in fluent bilingual epilepsy patients were electrically stimulated prior to surgery. The patients performed an object-naming task in their first and their L2, and it was recorded whether electrical stimulation interfered with object naming or not. For every individual, this procedure resulted in a map of sites that were necessary for L1 picture naming, L2 picture naming, or both. Shared sites were found in all left peri-Sylvian regions. L1-specific sites were also found in both posterior frontal and temporal regions but more so in the frontal cortex. L2-specific sites were exclusively found in mid- to posterior temporal cortex and adjacent inferior parietal sites. These data show that although there may not be any L1- or L2-specific regions at a group level, there seem to be cortical sites in many individuals that are only necessary for word production in one of the languages. Moreover, the left inferior frontal cortex, which in hemodynamic studies has been found to be more strongly activated for L2 word production, does not contain L2-specific sites but, rather, L1-specific sites possibly subserving L1-specific processes. A possible interpretation might be that L2 speakers attempt to make use of processes that are in some way tailored to L1 word production (to the extent that L1, but not L2, word production becomes impossible without them). A good candidate for such a process seems to be postlexical syllabification, which is subject to language-specific constraints and engages Broca's area in L1 word production.

Most studies on bilingual word-level comprehension have used written stimuli and asked their subjects to perform some kind of semantic decision. Typically, control conditions involved nonlinguistic stimuli so that the observed activations may reflect

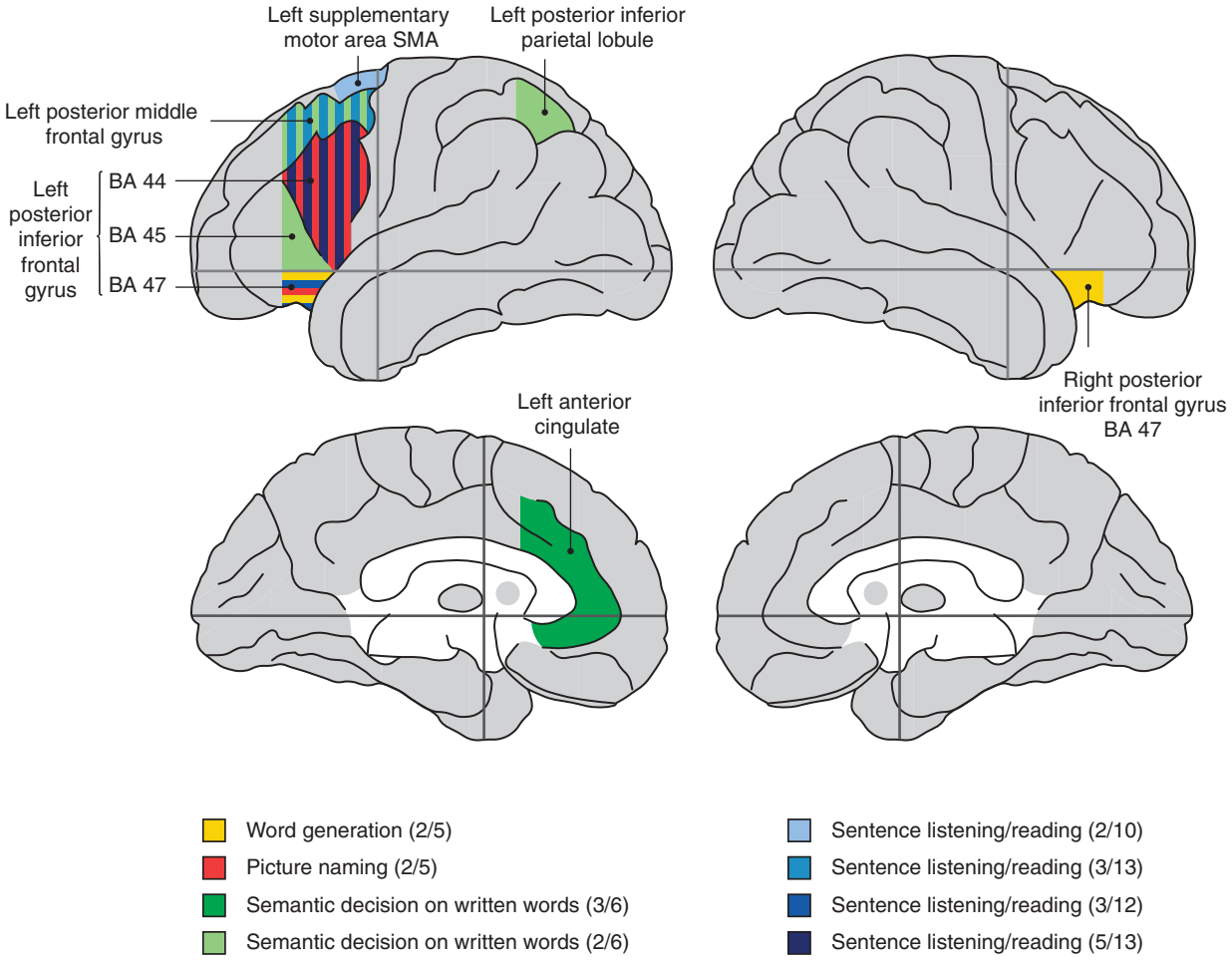

Figure 1 Schematic representation of brain areas reliably reported to be more strongly activated in L2 compared to L1 (based on a meta-analysis by Indefrey). Colors indicate the task in which stronger L2 hemodynamic activation has been found. Numbers in brackets refer to the number of studies reporting a brain area and the number of studies that examined this area. SMA, supplementary motor area.

all processing components involved in reading as well as semantic decision. Stronger activation for semantic decisions on L1 compared to L2 words has been reported for the anterior left middle temporal gyrus in late L2 learners with moderate to high proficiency. Stronger activation for semantic decisions on L2 compared to L1 words has been reported for the left posterior middle frontal gyrus, the left posterior inferior frontal gyrus (Broca's area (BA) 45), the left posterior inferior parietal lobule, and the anterior cingulate gyrus (Figure 1). The areas tend to be found in bilingual speakers with lower L2 proficiency, whereas there is no obvious influence of L2 onset or use. As in word production, areas of potentially stronger activation in L2 word comprehension are not L2 specific but are also observed in the native language. They are part of a modality-independent semantic system found to be activated during semantic decisions on both words and pictures. Anterior left middle temporal lobe activation, by contrast, has been found for semantic decisions on written words only, not on pictures, suggesting that a stronger activation of this area for L1 compared to L2 words may be due to L1-specific word representations at the graphemic or lemma level.

The processing of inflectional morphology in a L2 has been investigated with fMRI. Japanese learners of English were visually presented with English verb stems and had to choose between correct ('caught' and 'smelled') and incorrect past tense forms ('catched' and 'smold'). After a 2-month training involving explicit instruction on verb inflection, 13-year-old beginners of English showed activation of the dorsal left posterior inferior frontal gyrus (BA 44/45) when performing the task. The activation was stronger the better participants performed on the task. The area overlapped with the activation area observed when the same task was conducted with Japanese verbs. In a follow-up study, the same task was used on 19-year-old learners of English, who by this age

had 6 years of English instruction. Again, increased activation of left posterior inferior frontal gyrus (together with parieto-occipital and motor areas) was found when comparing the past tense task to a control task. A subgroup of high-performing advanced learners showed much weaker activation in the inferior frontal gyrus area, and the activation was also weaker in older compared to younger participants when performing the task with Japanese verbs. Taken together, the two studies suggest a change in the linkage between performance on a certain linguistic task and accompanying hemodynamic responses. The activation level of an area may initially be positively correlated with performance but show a negative correlation after years of practice. The positive correlation may simply reflect neural and behavioral consequences of the same underlying factor – the degree of effort put into the task. By contrast, a negative correlation may be explained by assuming that neural structures support the linguistic process involved in the task more effectively, which means that they must have in some way adapted to the process. Note, however, that 'adaptation' is not the only option to interpret reduced activation with better performance. Alternatively, a different but more effective brain region might have taken over or such a negative correlation might occur because the task is performed in a different, more effective way that no longer involves the original process.

Sentence-Level Processing

A number of studies have presented proficient bilingual speakers with spoken or written sentences. Although some regions have been reported as being more strongly activated in L1 compared to L2 sentence or story comprehension, no region has been reported in more than one study. Regions that have been reliably replicated as being more active in L2 than in L1 sentence comprehension are mainly found in the left posterior frontal gyrus (middle frontal gyrus, BA 44 and BA 47 of the inferior frontal gyrus, and the supplementary motor area; Figure 1). Stronger L2 activation of these regions has typically not been observed when subjects simply read or listened to sentences or stories, but it seems to depend on the use of additional judgment tasks. In particular, stronger L2 activation of BA 44 has been found consistently when bilinguals were asked to distinguish between correct sentences and syntactic violations. Another shared, but not distinctive, feature of the studies reporting stronger L2 activations seems to be late L2 onset (after age 10 years). In summary, the available hemodynamic studies suggest that proficient L2 speakers recruit the same cortical areas (but possibly stronger) for L2 as for L1 sentence-level comprehension.

We investigated changes in Broca's area activation in beginning Chinese learners of Dutch as an L2 for more than 1 year. Significantly stronger activation of Broca's area for listening to sentences compared to word sequences first emerged after 6 months of learning the new language, when the participants still scored in the low to moderate range of a standardized Dutch proficiency test. These findings suggest that brain regions involved in L1 syntactic processing are relatively soon also recruited for the processing of a new language.

Recruitment of similar brain areas for L1 and L2 sentence processing has also been observed in sentence production. Comparable activation levels for the two languages in frontal and temporal regions were found when bilingual speakers were asked to covertly report events of the previous day in their first and their L2s or when they generated sentences in a much better controlled task. Note, however, that in these studies differences between L1 and L2 in the exact localization of activation peaks in the frontal cortex (not in temporal cortex) were found, and these localization differences covaried with the age of L2 acquisition.

Electrophysiological Correlates of L2 Acquisition

The general picture that emerged from this overview of hemodynamic studies is that L2 processing involves the same brain structures as L1 processing but may be associated with greater activity. Complementary evidence for possible underlying processing differences may be provided in studies using electrophysiological measures such as the event-related potential (ERP) which are sensitive to the time course of neuronal activity. Past L2 research has mainly characterized L2 responses of speakers with preexisting L2 knowledge, whereas recent work has examined L2 learning.

Sublexical Processing

Studies of sublexical processing in bilinguals and learners have focused on the mismatch negativity response (MMN) observed in a passive listening task in which a sequence of standard sounds (e.g., syllables) are presented with an occasional mismatching sound. An MMN is taken as evidence that the brain has detected a difference between the two sounds. This technique has been used to examine whether native speakers and L2 learners show different electrophysiological responses to features of phonemic sounds that are distinctive in an L2.

Several studies have shown that an MMN response can be observed in both younger and older L2 learners, but that L2 learners who start relatively late show reduced amplitude MMN for L2 phonological contrasts.

Word-Level Processing

Several studies have compared L1 and L2 N400 responses to single words. The N400 is a negative brain potential deflection, peaking approximately 400 ms after stimulus onset, that is modulated by associative, semantic, or meaning-identity relations between a word and its context. Similar patterns of both associative and semantic priming have been shown in L2 N400 responses compared to L1 responses, but the pattern is modulated by L2 proficiency, with lower proficiency being associated with reduced conceptual but not associative priming. Current studies suggest that priming effects operate similarly in L1 and L2 single-word recognition, with some modulation by overall L2 proficiency.

Sentence-Level Processing

Although the priming effects that have been observed in L2 comprehenders suggest similar word-level semantic processing in L1 and L2, this does not necessarily mean that the semantic integration of words into a sentence context works equally well. Studies of L2 responses to semantic violations in sentence contexts have observed the N400 response to L2 violations, and most have shown a similar topographical distribution of this effect for L2 violations compared to L1 violations, sometimes with a delayed peak in the N400 response. Studies that have directly contrasted age of acquisition and proficiency have found that both age of acquisition and vocabulary proficiency make independent contributions to the longer latency of the N400 response in L2 readers. There is some variability in L2 response patterns, however, because no difference in the N400 violation response or a greater N400 violation response have both been observed. In most studies, however, the L2 N400 response in sentence contexts has been observed with a similar topography.

Two other electrophysiological violation responses are elicited by grammatical rather than semantic violations in native speakers. One is a late positive deflection (P600, or syntactic positive shift), and the other is a left anterior negativity (LAN) that is observed between 200 and 500 ms after a grammatical violation. L2 ERP studies have usually compared grammatical violation responses versus controls. Studies examining responses to phrase structure violations have demonstrated the P600 in early starting and/or proficient L2 learners, but the LAN has only been shown for relative high proficient L2 users, and some reports have also shown an atypical P600 in late L2 learners. Other studies have observed the P600 response for certain combinations of L1 and L2 pairs but not others. The existing patterns across studies are not easily reconciled, but it appears that L2 P600 effects are more likely to be observed if the L1 and L2 belong to the same language family (e.g., Western Germanic for German and Dutch), if one takes into account the proficiency of the speakers tested. The limited number of magnetoencephalographic L2 studies have revealed similar variability in responses to grammatical violations. Whereas some studies have shown relatively early (in time) L2 violation responses, others have not. Finally, comparisons of L1 and L2 responses to open- and closed-class words (without grammatical violations) have shown L1/L2 response differences for closed-class words in later-starting L2 learners even though similar responses to L1 and L2 open-class words have been shown.

Studies have examined violation responses in L2 learners as they acquire the L2. Similar to the results in learners with existing knowledge, P600 violation responses have been observed when learners have acquired L2 proficiency. Although no early (in time) components have been observed in learners, some studies have observed an N400 response to grammatical violations in L2 learners.

These results show that responses to grammatical violations in an L2 can be observed using electrophysiological techniques, but the exact form of the violation response varies greatly, more so than semantic effects. Two factors, age of acquisition and global language proficiency in the L2, appear to play an important role in the appearance of all of these responses.

Conclusions

Neurocognitive studies of bilingual language processing have indeed provided some answers to the question of how an L2 is implemented in the brain. First, one of the three hypothesis mentioned in the introduction, the 'radically different' hypothesis, can clearly be rejected. L2 processing does not use any new brain areas nor does it give rise to new electrophysiological effects that are not found in L1 processing. With respect to the other two hypotheses, the situation is more complex. Whether L2 processing uses the same processes and brain structures as L1 processing or just a subset seems to depend on the processing level, and hemodynamic and electrophysiological data are not always in agreement.

The available evidence for sublexical processing suggests no fundamental differences. Some nonnative phonemic contrasts can be acquired to a level at

which MMN responses and hemodynamic responses are indistinguishable from those observed for native phonemic contrasts. Insofar as nonnative contrasts elicited differential responses, these seem to be of a quantitative rather than a qualitative nature, such as a reduced MMN or a stronger hemodynamic activation of the left inferior frontal gyrus, which is also active for phonemic decisions in the native language.

Similarly, at the lexical level, electrophysiological and hemodynamic data are in agreement that there are no L1/L2 differences with respect to the type or location of responses. Brain responses may be modulated by proficiency in the L2, with lower proficiency being correlated with reduced electrophysiological semantic priming effects and stronger recruitment of primarily left posterior frontal regions. Note, however, that there is hemodynamic evidence for L1-specific processing of written words. It is also important to note that the overall observation of similar recruitment of brain areas for L1 and L2 lexical processing may only hold at the level of larger areas, but within those areas, L1 and L2 may be separated to some extent in every individual bilingual.

Although below the sentence level the available data generally support the 'more of the same' hypothesis, electrophysiological and hemodynamic data give different answers with regard to sentence processing. Whereas hemodynamic studies report similar L1 and L2 processing areas with age of acquisition-dependent modulations (stronger recruitment of Broca's area in comprehension and location of activation peak differences within Broca's area in production), electrophysiological studies suggest that one typical native language response to grammatical violations, the LAN, is less commonly observed in the L2. Also, it cannot be said that speakers who acquired an L2 early across the board show the P600 in the same way as native speakers (as might be expected from similar L1 and L2 hemodynamic activation patterns in early L2 learners). Whether or not a P600 is observed seems to depend on factors such as distance between languages and the exact type of violation in ways that remain to be more closely examined in the future. These electrophysiological data are the main source of evidence supporting the hypothesis that L2 processing may only use a subset of the processes that are engaged for the L1. Insofar as this holds true, L1/L2 differences seem to be restricted to the sentence level, most likely syntactic processing.

How, then, can it be that such differences do not seem to affect the set of brain regions that are hemodynamically activated in sentence processing? First, there are design differences that reduce the comparability of hemodynamic and electrophysiological studies. Hemodynamic studies used mainly correct sentences and within-subject L1/L2 comparisons, whereas ERP studies used grammatically violating sentences and L1/L2 differences were observed in group comparisons. Second, both the L1/L2 differences observed in ERP studies (scarcity of LAN or P600 in the L2) and the L1/L2 similarities in hemodynamic studies are negative findings, so it remains possible that evidence to the contrary will be found in the future. Third, and probably most important, the presence or absence of ERP components does not hinge only on certain neural populations being active or not but, rather, on their synchronous activity. It could be that what appears as a qualitative difference in the data is, in fact, a quantitative difference in the timing of neuronal activation. If L2 speakers were more variable with respect to the exact timing of the cognitive operations underlying, for example, the LAN response in native speakers, this ERP response might not be observable anymore, whereas the corresponding hemodynamic effect of neuronal activation might still be observable due to the much coarser temporal resolution of fMRI or PET.

In summary, the most parsimonious interpretation of the available neurocognitive data on L2 processing is a 'more of the same' account assuming that L2s are implemented in the same brain structures as the native language, but with quantitative differences in timing and strength of activation. Alternative views assuming qualitative L1/L2 differences, such as the 'subset' hypothesis, have some, although not conclusive, support from electrophysiological data but are not supported by hemodynamic data.

See also: Bilingualism; Language Development; Language: Cortical Processes; Language: Aphasia; Psycholinguistics; Sentence Production; Sentence Comprehension; Statistical Learning of Language; Word Production; Word Recognition; Word Learning.

Further Reading

Golestani N, Alario F-X, Meriaux S, LeBihan D, Dehaene S, and Pallier C (2006) Syntax production in bilinguals. *Neuropsychologia* 44: 1029–1040.

Gullberg M and Indefrey P (2006) *The Cognitive Neuroscience of Second Language Acquisition.* Malden, MA: Blackwell.

Hahne A (2001) What's different in second-language processing? Evidence from event-related brain potentials. *Journal of Psycholinguistic Research* 30: 251–266.

Indefrey P (2006) A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean? *Language Learning* 56: 279–304.

Lucas TH, McKhann GM, and Ojemann GA (2004) Functional separation of languages in the bilingual brain: A comparison of electrical stimulation language mapping in 25 bilingual patients

and 117 monolingual control patients. *Journal of Neurosurgery* 101: 449–457.

Osterhout L, McLaughlin J, Kim A, Greenwald R, and Inoue K (2004) Sentences in the brain: Event-related potentials as real-time reflections of sentence comprehension and language learning. In: Carreiras M and Clifton C Jr. (eds.) *The On-line Study of Sentence Comprehension: Eyetracking, ERP, and Beyond*, pp. 271–308. New York: Psychology Press.

Paradis M (2004) *A Neurolinguistic Theory of Bilingualism*. Amsterdam: Benjamins.

Perani D and Abutalebi J (2005) The neural basis of first and second language processing. *Current Opinion in Neurobiology* 15: 202–206.

Rossi S, Gugler MF, Friederici AD, and Hahne A (2007) The impact of proficiency on syntactic second-language processing of German and Italian: Evidence from event-related potentials. *Journal of Cognitive Neuroscience* 18: 2030–2048.

Stowe LA and Sabourin LL (2006) Imaging the processing of a second language: Effects of maturation and proficiency on the neural processes involved. *International Review of Applied Linguistics* 43: 327–351.

Tatsuno Y and Sakai KL (2005) Language-related activations in the left prefrontal regions are differentially modulated by age, proficiency, and task demands. *Journal of Neuroscience* 25: 1637–1644.

Weber-Fox CM and Neville HJ (1996) Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience* 8: 231–256.

Segawa Dopa Responsive Dystonia

M Segawa, Segawa Neurological Clinic for Children, Tokyo, Japan

© 2009 Elsevier Ltd. All rights reserved.

Definition and Classification

Autosomal dominant (AD) GTP cyclohydrolase I (GCH-I) deficiency or Segawa disease is a dopa responsive dystonia (DRD) caused by heterozygous mutation of GCH-I gene, located on chromosome 14q22.1–q22.2. This disease was initially reported by Segawa as hereditary progressive dystonia with marked diurnal fluctuation (HPD). Prior to identification of the causative gene of HPD, this disease was called as Segawa syndrome with recessive type, which is clarified later as recessive tyrosine hydroxylase (TH) deficiency. In the 1990s, this was called as DRD. Recently, it is classified in DYT5 with recessive TH deficiency.

AD GCH-I deficiency is an autosomal dominantly inherited, generalized, postural dystonia with female predominance. The hallmarks of this disease included onset in childhood, diurnal fluctuation of symptoms, and marked and sustained response to levodopa. But some patients show action dystonia, ballistic dystonic movements, associated with postural dystonia from late childhood. AD GCH-I deficiency is characterized by the age dependency of the symptoms and clinical courses, and adult onset patients may start with tremor but without generalized dystonia or diurnal fluctuation. Other neurological conditions include psychiatric disorders, autism, depression, and migraine; furthermore, subjects with compound heterozygotes show hypotonia, failure in locomotion, and delay in mental and motor development in infancy.

Neuropathological studies revealed no degenerative changes. Neurohistochemistry shows decrease of the TH protein at the terminal of the nigrostriatal (NS)–dopamine (DA) neuron with predominance in the ventral area, which causes postural dystonia through the descending pathways of the basal ganglia. However, clinical, neurophysiological, and neuroimaging studies suggest the involvement of the NS–DA neurons projecting to the subthalamic nucleus (STN) with D_1 receptor for tremor and action dystonia. With lesions at the terminal, this disease shows diurnal fluctuation and age-related variation of symptoms.

Clinical Signs and Symptoms

Clinical symptoms are characterized by age dependency both in the initial signs and the clinical course. The ages of onset are childhood from 1 to 11 years, mostly ~6 years. The tremor disappears by passive stretching of the muscle, that is, it does not appear as cogwheel rigidity. However, there are patients who have onset in adulthood and some in ages older than fifties.

Childhood onset cases start with postural dystonia, dystonic posture of one of the lower extremities, mostly with pes equinovarus. There are patients who start with dystonia of one upper extremity at ages a little bit later. However, some patients show action dystonia, rigorous dystonic movements, besides postural dystonia. These appear later than postural dystonia from ~8 years. These are observed in neck and shoulder commonly as action retrocollis. Occurogyric crisis may associate with this. The occurrence of action dystonia is depending on the family. Postural tremor appears later in upper extremities with asymmetry, mostly after 10 years.

Adult onset patients start with hand tremor. They may show gait disturbance due to generalized mild rigidity. Some adult onset patients start with writer's cramp. Patients with onset after 50 years show tremor and generalized rigidity with bradykinesia similar to parkinsonism. These adult onset patients do not show the postural dystonia expanding to generalized dystonia. Asymmetry is a characteristic feature and is observed in dystonia, rigid hypertonus, and tremor irrespective of ages at onset.

Dystonia shows marked diurnal fluctuation, that is, aggravates toward the evening and recovers markedly or nearly completely in the morning after sleep. The fluctuation is also observed in tremor. But these fluctuations are mild or not apparent in adult onset patients.

In patients with onset in early childhood, the body length fails to grow normally with the onset of dystonia and becomes short stature in the late teen ages. This is not observed in patients with onset after adolescence. Locomotion or interlimb coordination is preserved and psychomental activities are not affected even in advanced stages.

Clinical courses are also characterized by age dependency. The postural dystonia of the lower extremity occurring in childhood expands to all limbs and develops to be generalized dystonia by the middle of the teen age and its grade progressively aggravates toward the early twenties. However, the

progression attenuates with age, and once the age has crossed 30 years, symptoms become stational. With the attenuation of the progression, diurnal fluctuation decreases its grade and becomes not apparent in the stational stage. However, tremor, with onset in early teen, expands all limbs and trunk with age and this progression is observed until 30 years. Writer's cramp or torticollis, that is, focal or segmental dystonia, may appear in adulthood.

Neurological Examinations

Muscle stretch reflexes demonstrate a rigid hypertonus, but there is no plastic rigidity, and repeated testing will produce fluctuations in the tonus. The tremor is a high-frequency postural tremor (8–10 Hz), but a parkinsonian, resting tremor is not observed. However, adult onset patients may show resting tremor of lower frequency. These clinical signs show asymmetry, but the pattern of involvement of the sternocleidomustoideus (SCM) differs between rigidity and tremor. That is, the side predominantly affected in the SCM is contralational to that of extremities in rigidity, while it is ipsilateral in tremor. However, in adult onset cases, the side of predominance of the rigid hypertonus is ipsilateral between the SCM and the muscles of extremities. Bradykinesia or postural instability appears with advancing symptoms of dystonia. However, freezing phenomena or the *marche a petit pas* of Parkinson's disease (PD) are not seen, and locomotion is preserved throughout the course of illness. The tendon reflexes are brisk and ankle clonus may be observed, but the plantar reflexes are flexor. Although some patients exhibit sustained dorsiflexion of the toe, this is 'striatal toe sign', not elicited by plantar stimulation, and is associated with basal ganglia involvement. There are neither cerebellar signs nor sensory disturbances. Psychomental activities are preserved normally.

Clinical Variation

Besides the classical type with generalized postural dystonia, there is a type associated with action dystonia. This variation is depending on the family and suggests existence of two types, that is, postural and action dystonia type. Ages at onset also show another phenotypical variation, that is, adult onset patients show tremor, gait disturbance, focal dystonia, or parkinsonism but without generalized dystonia and diurnal fluctuation. There are patients with paroxysmal dystonia, dystonic cramp, or oculogyric crisis. These symptoms are tended to be observed in patients with action dystonia. There are also intrafamilial variations, that is, there are families with anticipation in the ages at onset, while others show identical features or marked phenotypical variation irrelevant to the generation. Some patients show autistic features, depressive state, or migraine, and others, particularly in patients with compound heterozygotes, show hypotonia, failure in locomotion, and delay in mental and motor development. These symptoms are caused by serotonergic deficiency. These association are expected because GCH-I deficiency affects the 5-hydroxytryptophan (5-HTP), the coenzyme for synthesizing serotonin (5-HT) and observed in patients with marked deficiency of GCH-I. Thus, the heterogeneity of symptoms in this disorder appears to be related to age at onset, pathophysiological differences among families, and the grade of the involvement of the serotonergic neurons.

Treatment and Prognosis

In most cases, a dose of 20 mg kg^{-1} day^{-1} of plain levodopa without decarboxylase inhibitor alleviates the symptoms completely. Some patients starting treatment with plain levodopa before 10 years tend to decrease the response after ~13 years. Older onset subjects do not always respond to plain levodopa. To these patients, levodopa with decarboxylase inhibitor with doses of 4–5 mg kg^{-1} day^{-1} alleviates the symptoms completely. In a few patients, choreic movements develop by a rapid increase of dosage or by administration of a higher dose of levodopa in the initial stage of treatment. In patients with action dystonia, action retrocollis and oculogyric crisis may be aggregated by initial doses. In patients with compound heterozygote, aggravation of dystonia by initial dosage is prominent. In these patients, the unfavorable symptoms disappear with decrease of the doses. After titration to an optimal dosage by starting with smaller doses and by slowly increasing it, levodopa shows sustained and favorable effects without side effects. Levodopa is effective in almost all patients without any relation to the age of onset and the longevity of the clinical courses, and improves short stature, if administered before puberty.

Anticholinergic drugs may have a marked and prolonged effect, but do not afford a complete relief, either clinically or polysomnographically. It does not improve the tremor. Amantadine has proven beneficial for levodopa-related chorea. Tetrahydrobiopterin (BH_4) monotherapy is not favorable, but there are a few patients who show complete remission after administration of BH_4 in addition to levodopa. In a patient with compound heterozygote, administration of BH_4 is necessary for complete recovery.

Investigations

Biochemical Studies

Cerebrospinal fluid (CSF) examinations reveal low levels of homovanillic acid (HVA). But characteristic features are marked decrease (20% of normal levels) of both biopterin and neopterin levels. Moderate reduction of these pteridine metabolites is also observed (about 30–50% of normal levels) in CSF of asymptomatic carriers.

The activity of GCH-I in the mononuclear blood cells of patients is less than 20% of those in normal individuals, while asymptomatic carriers reached 30–40% of normal levels. Phenylalanine loading tests in both child and adult patients reveal a 6-hour increase in phenylalanine levels. In addition, the phenylalanine-to-tyrosine ratios remain at elevated levels during the postloading period, while biopterin levels decline. However, this test tends to show false negative results.

Neuroimaging Studies

Magnetic resonance imaging (MRI) and computed tomography (CT) scans of the brain show no abnormalities, while positron emission tomography (PET) scanning demonstrates normal or low normal [18F] dopa uptake levels. [11C] raclopride PET shows normal activity in symptomatic subjects. [11C] *N*-spiperone PET reveals mild increase in receptor binding. There is no increase in receptor binding in follow-up PET analysis after 7 months of levodopa therapy. Childhood onset patients with 30-year untreated courses and a 59-year-old patient with onset at 58 years showed no abnormalities either in [18F] dopa or [11C] *N*-spiperone PET scan. [123I] β-CIT SPECT scanning is normal in this disease.

Neurophysiological Studies

Polysomnographies (PSGs) reveal abnormalities in the phasic components of sleep. These changes include a decrease in the number of gross movements (GMs) and twitch movements (TMs) and abnormality in the pattern of occurrence of GMs against sleep stages.

These phasic components of sleep are modulated by the basal ganglia and the NS–DA neuron; of them, the numbers of TMs during rapid eye movement (REM) sleep reflect NS–DA neuronal activity. Sleep structure, percent sleep stages, and other parameters modulated by the brain stem aminergic neurons are preserved normally.

Normally, the number of REM-associated TMs decline with age and show decremental nocturnal variation with sleep cycle. In AD GCH-I deficiency, these ages and the nocturnal variations of the TMs are preserved, but the amount of TM declines to approximately 20% of normal values. Abnormalities of the patterns of GMs differ between postural dystonia and action dystonia type and that of the latter suggests DA-D_2 receptor supersensitivity.

Evaluation of saccadic eye movements reveals abnormalities in both visually guided and memory guided saccades, and implicates involvement of the both the direct and the indirect pathways. Besides hyperactive nigrocollicular inhibition associated with slowing in both memory and visually guided saccades, disinhibition of the superior colliculi has been postulated by failure in suppression of unnecessary saccade in memory-guided task.

Supracranial magnetic stimulation was normal, showing preservation of the corticospinal tract. Paired pulse magnetic stimulation showed normal short-interval intracortical inhibition of the motor cortex in postural type AD GCH-I deficiency. This suggests that reduction of GABAergic inhibition of the thalamocortical pathway may not contribute to generation of dystonia in postural-type AD GCH-I deficiency.

Brain Pathology and Histochemistry

Neuropathological examination reveals no demonstrable changes in the substantia nigra (SN) except decrease in the melanin, particularly in the ventral tier of the pars compacta. Histochemically, DA content is reduced in the pars compacta in the SN and striatum. Similar to PD, the reduction is greater in the putamen than in the caudate nucleus, and subregionally, more in the rostral caudate and the caudal putamen. In contrast to PD, AD GCH-I deficiency shows a greater DA loss in the ventral subdivision of the rostral caudate than its dorsal counterpart, and the activity and protein content of TH is decreased only in the striatum, while it is within the normal range in the SN.

There are marked reductions of total biopterin (84%) and neopterin (62%) in the putamen, despite normal concentration of aromatic acid decarboxylase, DA transporter, and vesicular monoamine transporter. A postmortem study on an asymptomatic carrier shows modest reduction of TH protein (52%) and DA (44%), despite marked reduction of striatal biopterin (by 82%).

Molecular Biological Studies

The causative gene of AD GCH-I deficiency is the GCH-I gene located on 14q22.1–q22.2. Although more than 100 independent mutations have now been identified in the coding region of GCH-I, the loci of mutation differ across families but are identical in one family. The rate of mutant GCH-I

mRNA production against normal RNA was 28% in a patient but it was 8.3% in the asymptomatic carrier.

Molecular analysis remains unable to determine mutations of the coding region of the gene in approximately 40% of subjects with AD GCH-I deficiency. In some of these subjects, abnormalities in intron genomic deletion, a large gene deletion, an intragenic duplication or inversion of GCH-I, or mutation in as yet undefined regulatory gene modifying enzyme function are suspected.

Pathophysiology

Although the pathogenetic mechanisms for dominant inheritance are unknown, a classic dominant negative effect and destabilizing effect have been considered. The ratio of mutant/wild-type GCH-I mRNA in lymphocytes is higher in an affected individual than an unaffected heterozygote, and varies depending on the locus of the mutation. Furthermore, the ratio differed among affected individuals in some families, depending on the locus of the mutation. Thus, the degree and the pattern of inactivation of normal enzyme by mutant gene differ among the locus of mutation and may cause inter- and intrafamilial variation of the phenotype as well as the rate of penetrance.

In AD GCH-I deficiency, the TH appears to be preferentially affected when compared to tryptophan hydroxylase. This could be explained by the difference in distribution of GCH-I mRNA in DA and 5-HT neurons, the destabilization of the molecule of TH, or impairment of axonal transport. However, the difference of K_m value for TH and tryptophan hydroxylase is most probable. With heterozygotic mutant gene, the BH_4 decreases partially in AD GCH-I deficiency. Thus, TH with higher affinity to BH_4 is affected rather selectively. In molecular conditions with marked decrease of BH_4, both tryptophan hydroxylase and TH are affected, producing symptoms induced by deficiencies of the 5-HT neurons.

Complete and sustained response to levodopa, especially given the absence of morphological changes, further suggests that the lesion in AD GCH-I deficiency is restricted to the NS–DA neurons. TH activity of the NS–DA neurons shows age-related decrement and circadian oscillation in the terminals, but these age- and state-dependent variations are not observed in the SN or the perikaryon of the NS–DA neuron. The onset of symptoms in the first decade of life with diurnal fluctuation and age related the clinical course correlates to the activities of TH in the synaptic terminals of the NS–DA neurons in the caudate. Neurohistochemical studies confirm the decrease of the TH protein and its activities only at the terminal, and the PSGs suggested that the TH activities at the terminal follow the decremental age and nocturnal variation of normal individuals with low levels but without progressive decrement of the activities. These features and the results of PET scan suggest that in AD GCH-I deficiency are not a progressive or degenerative disorder and the NS–DA neurons preserve their fundamental functions.

Study of GCH-I activity in stimulated mononuclear blood cells shows age-dependent, decreasing activity in the first three decades of life. Putaminal biopterin levels increase in postnatal period, reaching a plateau at 1–13 years of age, before declining in adulthood. These results imply that pteridine metabolism has a critical period beginning early in infancy and extending to early childhood and important roles of GCH-I and BH_4 for neuronal development in the first and the second decades of life.

As for the loss of striatal TH protein with normal preservation of it in the SN, following processes are considered. BH_4 may control protein stability rather than expression. Animal experiments revealed stabilization of TH protein by co-expression of GCH-I and loss of TH protein but not of TH mRNA in the brains of BH_4-deficient mice.

In the rostral caudate in particular, the medial/ventral portions, the striosomes/patches or D_1 direct pathways are more numerous, whereas in the dorsal/lateral portions the matrix compartment is more homogenous. Thus, histochemical findings suggest that the DA loss in AD GCH-I deficiency is more prominent in the striosomes/patches compartment, the terminal for the D_1 receptor. Clinically it is suggested that the D_1 direct pathways mature earlier than the D_2 indirect pathways. Dopa-responsive growth arrest seen in children with AD GCH-I deficiency is a reflection of the tuberoinfundibular D_4 receptor involvement. The D_4 receptor belongs to the D_2 receptor family, which, however, matures early among D_2 families. Thus, the DA neuron in which the DA synthesis is modulated by pteridine metabolism might regulate DA receptors that mature early in the developmental course.

Tremor is levodopa responsive but develops independently, from symptoms of dystonia. The side predominance of tremor is identical to dystonic hypertonus in the extremities while it is contralateral in the SCM. These suggest a different pathophysiology of tremor from that of dystonia and postulate the responsible lesion in the downstream of the striatum, the DA neuron innervating to the STN with D_1 receptor. This confirms unresponsiveness of tremor to anticholinergies. For generation of the rhythmic discharge of tremor, the circuits consisted of the STN

and two globus palliduses are considered. In addition, response of the tremor to stereotactic ventrolateral (VL) thalamic nucleus thalamotomy performed in the era before levodopa suggests involvement of the ascending pathway to the VL nucleus of the thalamus. Given that ascending pathways to the thalamus develop later than the descending pathways, increasing age may be a factor for development of tremor. Further support for the D_1 receptor-mediated hypothesis stems from the PET findings revealing preservation of function of D_2 receptors, and suggesting that the striatal indirect pathway does not play a role in the generation of symptoms. The PSG findings observed in patients with action dystonia could be explained by the hypofunction of the STN.

The 5-HT neuron involves in modulation of locomotion and postural tone as well as of behavioral function. The hypotonia and failure in locomotion observed in patients with compound heterozygotes are associated with deficiency of the 5-HT-regulated activities. Preservation of interlimb coordination or locomotion in AD GCH-I deficiency without symptoms of 5-HT deficiency may depend on the preservation of the descending output of the basal ganglia to the pedunculo pontine nucleus (PPN).

Two studies on penetrance estimating the ratio of symptomatic carries revealed just identical results, that is, for females 87% and 87% and for males 38% and 35%. Gender difference of the base levels of GCH-I in the mononuclear blood cells is yet to be confirmed. Thus, marked female predominance might depend on a genetically determined gender difference of the DA neuron.

The pathophysiology of AD GCH-I deficiency is postulated as follows. The decrease of the TH protein or DA in the ventral area of the striatum causes disfacilitation of the D_1 striatal direct pathway, and disinhibits the output projection of the internal segment of the globus pallidus and pars reticulata of the SN, which suppress the reticulospinal tract and the superior colliculus. These may cause postural dystonia with exaggeration of the tendon reflexes without extensor plantar reflexes and abnormalities in voluntary saccades. Whereas, with involvement of the DA neuron innervating to the STN with D_1 receptor, the ascending outputs to the thalamus are disfacilitated and develop tremor and action dystonia. However, action dystonia is observed in infants with recessive sepiapterin reductase deficiency; the projection of the descending output to the brain stem reticular formation could get involved in this process. This process also disinhibits of the superior colliculus and involves the failure in suppression of unnecessary saccade in memory-guided task. Focal and segmental dystonia observed in patients with action dystonia may be caused by dysfunction of the motor cortex due to disinhibition of the thalamocortical pathway. Considering the ipsilateral finging of the predominant side of involvement in SCM and extremities, this pathway may be preferentially involved in adult onset patients. These pathophysiologies are shown in Figures 1(a)–1(c).

Diagnosis

Confirming the diagnosis of AD GCH-I deficiency is usually not difficult in the setting of characteristic clinical symptoms. The estimation of GCH-I activity in peripheral nucleated cells is accurate, but technically complicated. Thus, estimation of neopterin and biopterin levels in CSF is most reliable for diagnosis.

Differential Diagnosis

All children with gait disturbance and limb dystonia with asymmetry should be evaluated for AD GCH-I deficiency. Other conditions with this clinical presentation may include Wilson's disease, Hallervorden–Spatz disease, hereditary spastic paraplegia, and cerebral palsy. AD GCH-I deficiency is often misdiagnosed as hereditary spastic paraplegia. The differentiation of AD GCH-I deficiency from these disorders is usually not difficult with careful clinical examination.

Cases with axial torsion dystonia, including early onset autosomal dominant torsion dystonia (DYT1), can be differentiated clinically from AD GCH-I deficiency by evaluating the side that is predominantly involved between the SCM and muscles of the extremities; the side of the SCM predominantly affected is contralateral to that of the limb muscles in AD GCH-I deficiency, while it is ipsilateral in dystonias with axial torsion. This difference in the side preference reflects involvement of the NS–DA neurons in AD GCH-I deficiency while involvement of the striatum or the pallidum in torsion dystonia. However, it should be taken care that AD GCH-I deficiency and also DYT1 may appear with focal or segmental dystonia without general dystonia in adolescence or adulthood. Besides biochemical and molecular biological studies, levodopa loading test is recommended.

Dopa Responsive Dystonia Other Than AD GCH-I Deficiency

This group includes recessive disorders of pteridine metabolism and recessive TH deficiencies (recessive DYT5). All of the inherited disorders of pteridine

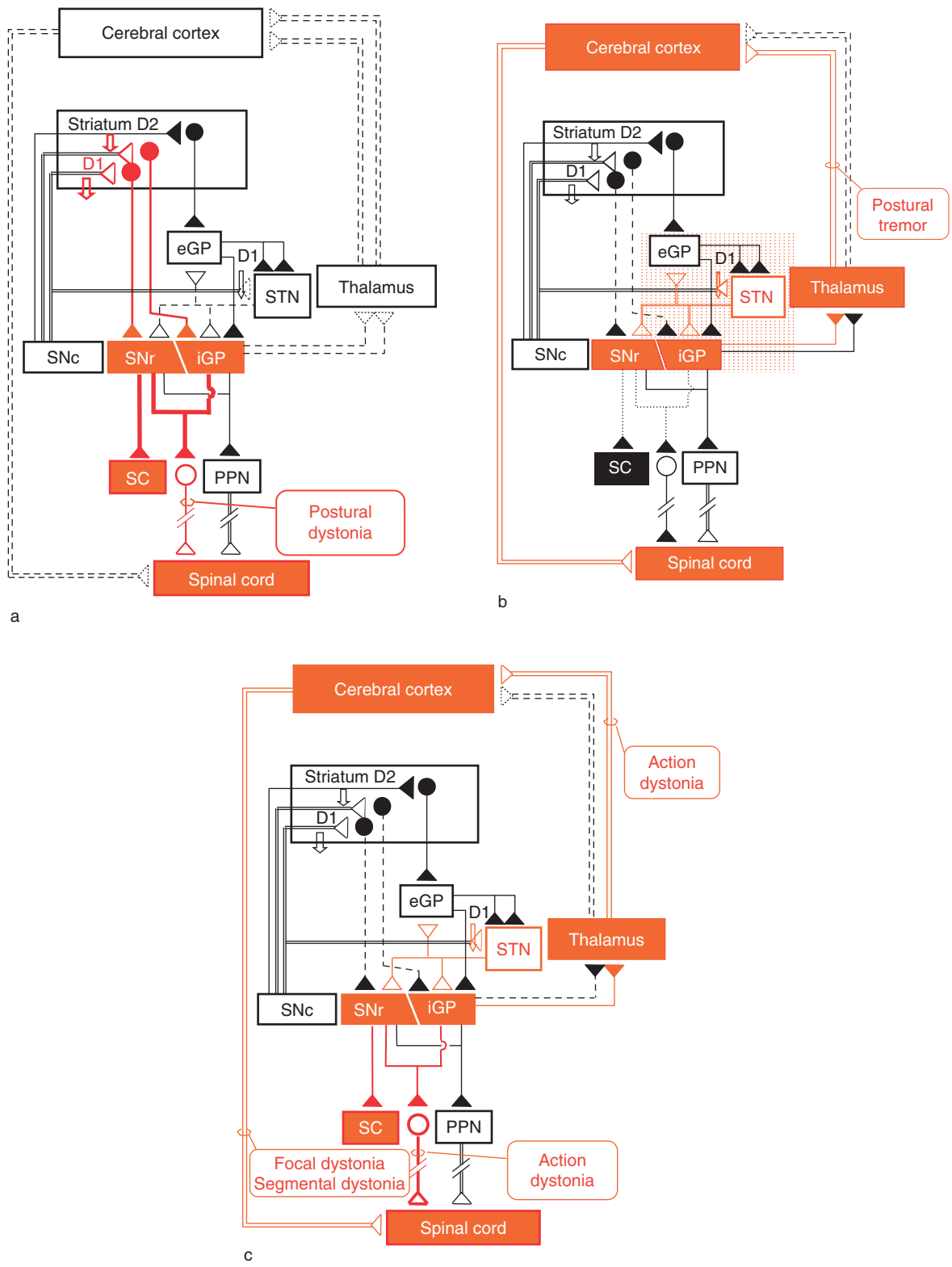

Figure 1 Pathophysiology of SDRD. (a) Neuronal pathways involving in postural dystonia. (b) Neuronal pathways involving in tremor. (c) Neuronal pathways involving in action dystonia and focal or segmental dystonia. eGP, external segment of globus pallidus; iGP, internal segment of globus pallidus; STN, subthalamic nucleus; SNc, substantia nigra pars compacta; SNr, substantia nigra pars reticulata; SC, superior colliculus; PPN, pedunculopontine nucleus. Red lines, pathways involved in pathophysiology; black lines, pathways not involved in pathophysiology; solid line, inhibitory neuron; dashed line, excitatory neuron; closed triangle, inhibitory neuron; open triangle, excitatory neuron; shaded region, the area of the circuit for postural tremor. White letters with red back ground, targeted neurons or neuronal system. Dotted lines and triangles show pathways and terminals involved in SDRD but not involved in the symptoms shown in the figure.

metabolism develop levodopa responsive dystonia caused by decrease of BH_4 in infancy and early childhood as in AD GCH-I deficiency. However, they show postural hypotonia and psychological disturbances caused by deficiency of the 5-HT activities. Recessive TH deficiency shows ptosis, hyperperspiration, and psychomotor disturbances due to noradrenaline (NA) deficiency. However, some show almost identical features as AD GCH-I deficiency and respond well to levodopa. However, the diurnal fluctuation is not marked as AD GCH-I deficiency. For definite diagnosis, estimation of pteridine metabolites and catecholamine metabolites in CSF is necessary.

Dopa Responsive Dystonia Parkinsonism

All of juvenile parkinsonism (JP) appears as dystonia when it occurs in childhood to early teens. Although dystonia of JP responds briskly to levodopa, dyskinesia develops soon after levodopa is started. Among JPs, that caused by parkin gene (PARK2) is a particularly important disease to differentiate from AD GCH-I deficiency when it occurs in early ages.

In addition, some patients with AD GCH-I deficiency develop symptoms later in life (e.g., fifties and sixties). In this population, tremor and gait disturbance are the primary signs, and dystonia is absent or not prominent. In these patients, diurnal fluctuation is not observed and often misdiagnosed as PD. However, the tremor in these cases is mainly postural and their clinical features are milder with minimal progression. For definite diagnosis of PARK2, gene analysis is necessary.

See also: Dopaminergic Agonists and L-DOPA; Dystonia: Classification, Genetics and Therapeutics; Parkinsonian Syndromes; Torsion Dystonia.

Further Reading

Furukawa Y, Kapatos G, Haycock JW, et al. (2002) Brain biopterin and tyrosine hydroxylase in asymptomatic dopa-responsive dystonia. *Annals of Neurology* 51: 637–641.

Furukawa Y and Kish SJ (1998) Influence of development and aging on brain biopterin: Implications for dopa-responsive dystonia onset. *Neurology* 51: 632–634.

Furukawa Y, Kish SJ, Bebin EM, et al. (1998) Dystonia with motor delay in compound heterozygotes for GTP-cyclohydrolase I gene mutations. *Annals of Neurology* 44: 10–16.

Furukawa Y, Lang AE, Trugman JM, et al. (1998) Gender-related penetrance and *de novo* GTP-cyclohydrolase I gene mutations in dopa-responsive dystonia. *Neurology* 50: 1015–1020.

Furukawa Y, Nygaard TG, Gutlich M, et al. (1999) Striatal biopterin and tyrosine hydroxylase protein reduction in dopa-responsive dystonia. *Neurology* 53: 1032–1041.

Hibiya M, Ichinose H, Ozaki N, et al. (2000) Normal values and age-dependent changes in GTP cyclohydrolase I activity in stimulated mononuclear blood cells measured by high-performance liquid chromatography. *Journal of Chromatography. B, Biomedical Sciences and Applications* 740: 35–42.

Hirano M, Tamaru Y, Ito H, et al. (1996) Mutant GTP cyclohydrolase I mRNA levels contribute to dopa-responsive dystonia onset. *Annals of Neurology* 40: 796–798.

Hornykiewicz O (1995) Striatal dopamine in dopa-responsive dystonia: Comparison with idiopathic Parkinson's disease and other dopamine-dependent disorders. In: Segawa M and Nomura Y (eds.) *Age-Related Dopamine-Dependent Disorders,* Monographs in Neural Sciences, vol. 14, pp. 101–108. Basel: Karger.

Hyland K, Fryburg JS, and Wilson WG (1997) Oral phenylalanine loading in dopa responsive dystonia; a possible diagnostic test. *Neurology* 48: 1290–1297.

Ichinose H, Ohye T, Takahashi E, et al. (1994) Hereditary progressive dystonia with marked diurnal fluctuation caused by mutations in the GTP cyclohydrolase I gene. *Nature Genetics* 8: 236–242.

Kreiss DS, Anderson LA, and Waters JR (1996) Apomorphine and dopamine D receptor agonists increase the firing rates of subthalamic nucleus neurons. *Neuroscience* 72: 863–876.

Kreiss DS, Mastropietro CW, Rawji SS, and Walters JR (1997) The response of subthalamic nucleus neurons to dopamine receptor stimulation in a rodent model of Parkinson's disease. *The Journal of Neuroscience* 17: 6807–6819.

McGeer EG and McGeer PL (1973) Some characteristics of brain tyrosine hydroxylase. In: Mandel J (ed.) *New Concepts in Neurotransmitter Regulation*, pp. 53–68. New York: Plenum.

Nygaard TG, Wilhelmsen KC, Risch NJ, et al. (1993) Linkage mapping of dopa-responsive dystonia (DRD) to chromosome 14q. *Nature Genetics* 5: 386–391.

Rajput AH, Gibb WR, Zhong XH, et al. (1994) Dopa-responsive dystonia: Pathological and biochemical observations in one case. *Annals of Neurology* 35: 396–402.

Segawa M, Hosaka A, Miyagawa F, Nomura Y, and Imai H (1976) Hereditary progressive dystonia with marked diurnal fluctuation. *Advances in Neurology* 14: 215–233.

Segawa M, Nomura Y, and Nishiyama N (2003) Autosomal dominant guanosine triphosphate cyclohydrolase I deficiency (Segawa disease). *Annals of Neurology* 54: 32–45.

Segawa M, Nomura Y, Yamashita S, et al. (1990) Long term effects of levodopa on hereditary progressive dystonia with marked diurnal fluctuation. In: Berardelli A, Benecke R, Manfredi M, and Marsden CD (eds.) *Motor Disturbance II*, pp. 305–318. New York: Academic Press.

Segawa M, Nomura Y, Yukishita S, Nishiyama N, and Yokochi M (2004) Is phenotypic variation of hereditary progressive dystonia with marked diurnal fluctuation/dopa-responsive dystonia (HPD/DRD) caused by the difference of the locus of mutation on the GTP cyclohydrolase I (GCH-I) gene? *Advances in Neurology* 94: 217–223.

Segawa M, Ohmi K, Itoh S, Aoyama M, and Hayakawa H (1971) Childhood basal ganglia disease with remarkable response to levodopa, hereditary basal ganglia disease with marked diurnal fluctuation. *Shinryo* 24: 667–672.

Segmentation: Segmental Boundaries, Establishment and Morphogenesis (EPH)

D G Wilkinson, National Institute for Medical Research, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Significance of Boundary Formation

The subdivision of the neural epithelium into domains with distinct regional identity is a crucial phase in establishing the highly complex organization of the central nervous system. This regionalization is initiated by localized signals that induce the spatially restricted expression of specific transcription factors, which in turn regulate which types of differentiated cells are generated within the tissue domain. At early stages the boundaries of regulatory gene expression can be fuzzy, but subsequently become very sharp. This sharpening is significant for two reasons: First, because it establishes a strict demarcation of regional domains that presumably contributes to the precision of subsequent tissue organization. Second, because some boundaries form a signaling center that further refines pattern by inducing new cell types. Since the distribution of the boundary-derived signal affects the location of the induced cell types, it is important that the signaling cells are precisely positioned by formation of a sharp interface.

Cell lineage studies have shown that there can be extensive intermingling within the neural epithelium due to the intercalation of proliferating cells. This raises the question of why such intermingling does not disrupt the formation and maintenance of precisely organized interfaces and signaling centers. One general mechanism to achieve this is the specific inhibition of cell intermingling between adjacent domains. Such territories within tissues that do not mix with their neighbors, termed compartments, were first recognized in studies of fruit fly development and shown to correspond to the domains of action of 'selector' genes that control regional identity. Clonal analyses and time lapse imaging have revealed a number of compartment boundaries in the developing central nervous system of vertebrates: between the cortex and lateral ganglionic eminence, the interface of the diencephalon and mesencephalon, between the zona limitans intrathalamica (zli) and adjacent dorsal and ventral thalamus, at the midbrain–hindbrain junction, and at the interfaces of hindbrain segments (rhombomeres). At some of these boundaries – the zli and midbrain–hindbrain junction – a signaling center forms that regulates patterning and differentiation of the adjacent territory.

Role of Cell Affinity in Boundary Sharpening

Important insights into mechanisms that inhibit cell intermingling came from studies showing that cell populations with different adhesive properties will segregate from each other after they are intermingled *in vitro*. Such segregation can be driven by quantitative or qualitative differences in homophilic cell-to-cell adhesion mediated by cadherins. These studies led to the concept that segregation is established and maintained by cells maximizing their contact with others for which they have the strongest affinity. Studies of the roles of cadherin family members with spatially restricted expression during development have confirmed the major role of these adhesion molecules in maintaining tissue organization. For example, the inhibition of cell intermingling between the cortex and lateral ganglionic eminence is mediated by R-cadherin and cadherin-6 that are expressed in these respective domains.

This article focuses on a second type of mechanism that restricts cell intermingling mediated by cell responses to signal transduction by Eph receptor tyrosine kinases and ephrins.

Interactions and Activation of Eph Receptors and Ephrins

Eph receptors are transmembrane receptor tyrosine kinases that bind to ephrin ligands. In vertebrates there are 14–16 Eph receptor tyrosine kinases and 8–9 ephrins that largely fall into two binding specificity classes: EphA receptors interact with ephrinA ligands that are anchored in the cell surface by a glycosylphosphatidylinositol moiety, and EphB receptors interact with ephrinB ligands that have a transmembrane and cytoplasmic domain. There are a few exceptions to these binding classes, for example, EphA4 that binds to ephrinB1 and ephrinB2 as well as all ephrinA ligands. Soluble ephrins do not activate Eph receptors unless they are artificially clustered, whereas Eph receptors are clustered and activated upon binding to membrane-bound ephrins. The clustering of Eph receptors triggers intracellular signal transduction, in part via tyrosine phosphorylation and activation of the receptor kinase domain, and also by kinase-independent mechanisms.

Remarkably, ephrins themselves have a receptor function and transduce signals upon being clustered by binding to Eph receptors. Consequently, Eph–ephrin

interactions mediate cell contact-dependent signaling that can lead to bidirectional cell responses. The mechanisms of ephrinA signal transduction are unclear, but for ephrinB proteins they involve phosphorylation of conserved tyrosine residues in the cytoplasmic domain by src family kinases.

Diverse Cell Responses to Eph–Ephrin Signaling

Eph receptor and ephrin signaling regulates a number of cellular responses, including proliferation, death, and migration. Of particular relevance to boundary formation are their roles in controlling cell migration by modulation of the state of assembly of the actin cytoskeleton and integrin-mediated adhesion. This involves multiple adaptor proteins that bind Eph receptors or ephrinB proteins on specific phosphotyrosine motifs (e.g., Nck, Grb4, RasGAP, Src), on a C-terminal PDZ interaction domain (e.g., Dishevelled), or on other motifs not requiring tyrosine phosphorylation (e.g., Ephexin). These adaptor proteins modulate the activity of a number of regulators of cell adhesion and the actin cytoskeleton, including Rho family GTPases, the Ras/mitogen-activated protein kinase (MAPK) pathway, and the planar cell polarity pathway.

One major role of EphA, EphB, or ephrinB activation is in cell contact-dependent repulsion in which there is disassembly of the actin cytoskeleton and collapse of cell membrane processes. Concomitantly, cell-to-cell attachment mediated by binding of Eph receptor to ephrin is disrupted by the endocytosis of intact Eph–ephrin complexes or the proteolytic cleavage of ephrinA protein, thus enabling the interacting cells to disengage. The repulsion response of migrating cells or neuronal growth cones to Eph or ephrinB activation prevents them from entering ligand-expressing territory, thus confining them to the correct pathway. Examples of such repulsive guidance in the vertebrate nervous system include the restriction of spinal cord motor axons expressing EphB receptors from the posterior half of somites that express ephrinB protein, and the inhibition of corticospinal axons expressing EphA4 from inappropriately crossing the ventral midline that expresses ephrinB3. A corollary of this role in guidance by contact repulsion is that Eph receptors and interacting ephrins are commonly expressed in complementary regions.

In contrast to such roles in repulsion, in other contexts Eph receptor or ephrin signaling can increase cell adhesion and promote the migration of cells or outgrowth of axons. For example, EphA–ephrinA signaling increases cell adhesion required for neural tube closure, and signal transduction through ephrinA5 promotes the targeting of vomeronasal axons to territory expressing high levels of EphA6. Furthermore, in some cells Eph receptor activation can lead to either repulsion or invasion responses. Switching between these alternative cellular responses contributes to the correct migration of axons during topographic mapping in the nervous system.

The nature of the cell response to Eph receptor activation depends at least in part upon the amount of clustering and/or the number of activated receptors. Unlike other classes of receptor tyrosine kinases, which are activated by dimerization, Eph receptors form higher order multimers upon binding to ephrin. The activation of some intracellular signaling components depends upon the amount of multimerization. Low levels of Eph receptor activation or clustering increase assembly of the actin cytoskeleton and integrin-mediated adhesion, whereas high levels of activation lead to disassembly of polymerized actin and decreased adhesion. This has been demonstrated most clearly for retinal axons that form a topographic map of connections by responding to a gradient of ephrinA ligands: low amounts of ephrinA stimulate axon outgrowth, whereas high amounts of ephrinA inhibit outgrowth. The nature of the cell response may thus depend upon the amount of ephrin ligand encountered. Alternatively, cell responses can be affected by mechanisms that modulate how much Eph receptor activation occurs in response to a given amount of ephrin. Examples of such modulatory mechanisms are the endogenous expression of truncated Eph receptors that inhibit Eph receptor phosphorylation, and the co-expression of ephrins and Eph receptors in the same cells, which leads to *cis* interactions and formation of inactive complexes. Cell responses may also be influenced by coactivation of Eph and ephrin signaling pathways within a cell, and by cross-activation of Eph receptor or ephrinB by fibroblast growth factor (FGF) receptor.

Roles of Eph–Ephrin Signaling in Boundary Sharpening

Repulsion at Boundaries

The clearest example of roles of Eph–ephrin interactions in the sharpening of boundaries in the nervous system has been uncovered in studies of the hindbrain. Expression of EphB and EphA4 receptor and their high-affinity ephrinB binding partners occurs in complementary sets of hindbrain segments, such that interactions occur at segment interfaces (Figure 1). The results of loss- and gain-of-function experiments

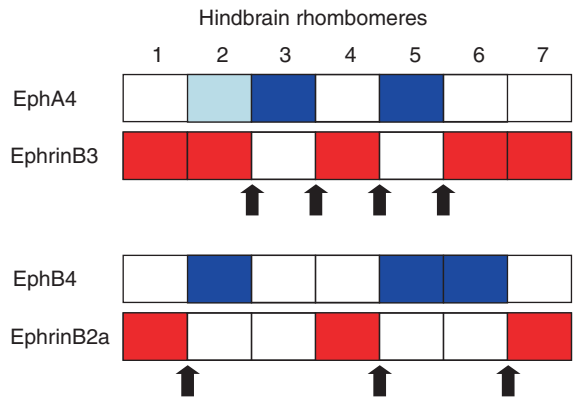

Figure 1 Expression of Eph receptors and ephrins in the zebra fish hindbrain. There is a largely complementary segmental expression of Eph receptors and ephrins that have high affinity for each other, for example, EphA4 and ephrinB3, and EphB4 and ephrinB2. This leads to interactions at hindbrain boundaries (indicated with arrows). There are some overlaps in expression, such as between ephrinB3 and the low level expression domain of EphA4 in rhombomere 2. Since this does not account for the role of EphA4 in regulating cell affinity within rhombomeres 3 and 5, other ligands may overlap with EphA4 in these segments.

in zebra fish embryos reveal that Eph–ephrin signaling is required for formation of sharp and straight interfaces between segments and that cell responses are regulated by signal transduction through Eph receptors and ephrinB proteins. Direct evidence for a requirement for bidirectional signaling comes from *in vitro* experiments in which Eph receptor and ephrinB-expressing cells are juxtaposed. Cell intermingling is restricted by signaling through full-length Eph receptor and ephrinB that mediate bidirectional activation at the interface. In contrast, cells expressing truncated Eph receptor – which acts as a ligand but cannot itself transduce signals – invade ephrinB territory, but ephrinB cells do not invade Eph receptor territory. Analogous roles of Eph–ephrin signaling in the segregation of distinct cell populations have been uncovered in mesoderm segmentation, the vasculature, limbs, and intestinal crypts. These findings, together with the complementary expression of Eph receptors and ephrins in many vertebrate tissues, suggest that Eph–ephrin interactions are widely used to stabilize tissue organization by mediating signaling across boundaries.

Activation within Domains

A role of Eph–ephrin activation at interfaces in restricting the mixing of cells predicts that Eph receptor function is required only at boundaries. For example, the expression of EphA4 in hindbrain segments r3 and r5 enables interaction with ephrins at the borders with adjacent even-numbered segments (Figure 1). However, mosaic knockdown of EphA4 leads to segregation of cells lacking EphA4 protein away from the central, nonboundary regions of these segments. Thus, EphA4 function is required throughout r3 and r5 to regulate cell affinity. These data suggest that EphA4 is activated by a co-expressed ligand(s). This is consistent with the findings, discussed previously, that overlapping Eph-ephrin expression causes low-level receptor activation that promotes cell adhesion.

Eph–ephrin activation thus contributes to the stabilization of boundaries by two mechanisms: by mediating bidirectional repulsion across the interfaces of cell populations and by regulating cell affinity throughout a cell population. This dual role could involve different levels of activation of the same receptor(s). Alternatively, there may be a different role of the two binding specificity classes, since although EphA and EphB receptors are both capable of triggering cell repulsion, in some tissues they mediate distinct cell responses.

Restriction of Communication across Boundaries

The formation of gap junctions between cells allows the passage of small molecules that may regulate cell phenotype, and the loss of such communication often correlates with the acquisition of a distinct cell fate. It is therefore intriguing that whereas cells within hindbrain segments are connected by gap junctions, there is a restriction to such communication across rhombomere boundaries. *In vitro* studies show that ephrinB and EphB receptor signaling each block the sharing of small molecules by gap junctions. Eph-ephrin signaling may thus regulate both the restriction of cell mixing and the inhibition of cell communication across the boundaries of hindbrain segments.

Relationships with Other Regulators of Boundary Formation

Eph–ephrin interactions have a major role in the restriction of cell intermingling in the hindbrain, but it is currently not known whether they contribute to the stabilization of other boundaries in the nervous system. Roles at other boundaries may be obscured due to overlapping expression and functions of Eph-ephrin family members. However, it is likely that other regulators of cell affinity are responsible for boundary sharpening or act in parallel with Eph receptors. This raises the question of the relationships between different signaling and adhesion systems implicated in boundary formation.

Notch and Boundary Formation

The inhibition of cell mixing between the zli and thalamus is regulated by lunatic fringe, a glycosyl transferase that by glycosylating the Notch receptor modulates the strength of its binding and activation by ligands. This situation has a parallel in the fruit fly wing imaginal disc, in which the restriction of cell mixing between the dorsal and ventral compartments requires differential activation of Notch in boundary cells and adjacent cells. The restriction of cell mixing in the wing disc involves transcriptional targets of the canonical Notch pathway and modulation of the actin cytoskeleton by a noncanonical pathway. How Notch receptors regulate boundaries in the developing nervous system is not known, but interestingly in somitogenesis Notch regulates the formation of morphological boundaries via the upregulation of EphA4 expression.

Cadherins and Cell Sorting

In view of evidence that Eph-ephrin signaling and cadherin-mediated adhesion can each restrict the intermingling of cell populations, it is possible that one or the other mechanism acts at specific boundaries, for example, at hindbrain boundaries and the lateral ganglionic eminence–cortex boundary, respectively. Alternatively, these mediators may synergize by acting in parallel at boundaries and/or modulate each other's activity. Indeed, E-cadherin is required for the correct cell surface localization and activation of EphA2. Furthermore, strong cadherin-mediated adhesion requires interaction with the actin cytoskeleton, which could indirectly be modulated by the regulation of cytoskeletal assembly by Eph-ephrin signaling.

Integrins and Boundary Stabilization

Eph receptor activation can lead to the promotion of the integrin-mediated adhesion of cells to extracellular matrix. A potential connection with boundary formation has been revealed by genetic studies of the segmentation of mesenchymal paraxial mesoderm to form epithelial somites. The formation of somite boundaries requires EphA4–ephrinB2 interactions at the interface of presumptive somites that induce the generation of polarized epithelial cells. The subsequent maintenance of sharp interfaces between somites requires integrin $\alpha 5$ and the accumulation of fibronectin at boundaries. Thus, Eph–ephrin interactions and integrin-mediated adhesion are both required and potentially have connected roles in the stabilization of boundaries. This situation may have a parallel at hindbrain boundaries where there is an accumulation of extracellular matrix in the spaces between cells.

Perspectives

Eph receptors and ephrins have versatile roles in the establishment and maintenance of tissue organization by regulating cell repulsion and adhesion responses. They can act at boundaries or within tissue domains to maintain sharp interfaces or in gradients to position cells or axonal connections at appropriate locations. Elucidation of how the level of Eph receptor clustering and other mechanisms control the restriction or invasion of cells remains a key challenge. There is increasing evidence for cross-talk with other classes of receptors that may modulate cell migration, as well as cell proliferation and death, and it is important to understand how intracellular signal transduction by these multiple receptors is correctly integrated.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axonal Pathfinding: Extracellular Matrix Role; Eph Receptor Signaling and Spine Morphology; Midbrain Patterning; Neural Patterning: Midbrain–Hindbrain Boundary; Notch Signal Transduction: Molecular and Cellular Mechanisms; Notch Pathway: Lateral Inhibition; Segmentation: Spinal Cord Segmentation and A–P Somite Patterning.

Further Reading

Cooke JE, Kemp HA, and Moens CB (2005) EphA4 is required for cell adhesion and rhombomere-boundary formation in the zebrafish. *Current Biology* 15: 536–542.

Holder N and Klein R (1999) Eph receptors and ephrins: Effectors of morphogenesis. *Development* 126: 2033–2044.

Kiecker C and Lumsden A (2005) Compartments and their boundaries in vertebrate brain development. *Nature Reviews Neuroscience* 6: 553–564.

Mellitzer G, Xu Q, and Wilkinson DG (1999) Eph receptors and ephrins restrict cell intermingling and communication. *Nature* 400: 77–81.

Pasini A and Wilkinson DG (2002) Stabilizing the regionalisation of the developing vertebrate central nervous system. *Bioessays* 24: 427–438.

Pasquale EB (2005) Eph receptor signalling casts a wide net on cell behaviour. *Nature Reviews Molecular and Cell Biology* 6: 462–475.

Poliakov A, Cotrina M, and Wilkinson DG (2004) Diverse roles of Eph receptors and ephrins in the regulation of cell migration and tissue assembly. *Developmental Cell* 7: 465–480.

Steinberg MS and Takeichi M (1994) Experimental specification of cell sorting, tissue spreading, and specific spatial patterning by quantitative differences in cadherin expression. *Proceedings of the National Academy of Sciences of the United States of America* 91: 206–209.

Segmentation: Spinal Cord Segmentation and A–P Somite Patterning

G M W Cook, K E Lewis, and R J Keynes, University of Cambridge, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Segmentation, the process of tissue subdivision into a periodic pattern of repeated elements, is a key strategy in animal development. It can expand body size, by increasing segment size and/or number, and it promotes tissue diversification, by enabling variation and differentiation of individual segments independently of others in the series. Accordingly, it appears to have evolved convergently several times in different animal phyla. The nervous system is no exception to this rule because the anatomy of the adult vertebrate nervous system shows clear evidence of its segmental origins, most clearly in the hindbrain and peripheral nervous system (PNS) of the trunk.

Intrinsic Segmentation of the Spinal Cord

As discussed elsewhere, the early neural epithelium of the developing hindbrain is subdivided into a series of repeat units, or rhombomeres, causing both motor neurons and interneurons of the hindbrain to be arranged segmentally. The PNS of the trunk region is also manifestly segmented, with aggregates of primary sensory neurons, or dorsal root ganglia, and associated spinal nerves (mixed motor and sensory) placed at regular, repeated intervertebral positions in relation to the segmented vertebral column. In most vertebrates, however, there is no direct evidence for any intrinsic segmental patterning within the spinal cord, unlike the hindbrain. Distinct subpopulations of motor neurons that occupy specific motor columns are specified at particular anterior–posterior (A–P) axial levels (e.g., lateral motor column neurons at limb levels and visceral/sympathetic motor neurons at thoracic levels). However, these differences reflect a subdivision of the A–P axis into broad regions rather than a more fine-grained, regular periodicity of different neuronal types, and similar subdivisions of interneurons have not been described.

Segmentation of Motor Neurons in the Zebra fish Spinal Cord

An important exception to this lack of segmentation in the vertebrate spinal cord is the zebra fish embryo, in which the first spinal motor neurons to form, primary motor neurons (PMNs), are located in a segmentally reiterated pattern along the A–P axis. Zebra fish have three different PMN subtypes: rostral primary (RoP), middle primary (MiP), and caudal primary (CaP) (Figure 1). One PMN of each subtype innervates the myotome (developing musculature) derived from each successive somite (mesoderm segment alongside the developing spinal cord), with the exception that approximately half the hemisegments (one hemisegment = one somite's length of one side of the spinal cord) initially have two CaP-like PMNs, one of which is called variably present (VaP) and usually dies. Each PMN subtype is uniquely identifiable by its soma (cell body) position relative to the overlying somites and by its axon trajectory. MiP and RoP somata, for example, are positioned adjacent to the somite boundaries, whereas CaP somata lie adjacent to the middle of each somite. CaP axons project into ventral myotome, MiP axons project into dorsal myotome, and RoP axons project into medial myotome. Such a fine-grained, reiterated pattern of distinct motor neuron subtypes has not been described in other vertebrates, although this may be because there are many more motor neurons in the species that have been examined to date, and distinct types of individual cells have not yet been recognized.

The tight spatial correspondence between the reiterated pattern of zebra fish PMNs and the overlying somites has long suggested that signals from paraxial mesoderm might specify different PMN subtypes. Consistent with this, transplantation experiments have shown that environmental signals can specify zebra fish PMN subtypes. For example, when a MiP is transplanted 2 or 3 h before axogenesis to a position where a CaP normally develops, the transplanted cell forms a CaP-like axon and initiates expression of a CaP-specific gene marker *islet2*. Studies of CaP and MiP specification have shown that zebra fish mutants which either lack somites or have very narrow somites form PMNs with hybrid identities, again suggesting that somite-derived signals specify the different PMN subtypes. This has been confirmed using cell transplantation to create genetically mosaic embryos: CaPs and MiPs were correctly specified in embryos where wild-type somites surrounded mutant neural tube but not in embryos with wild-type neural tube but lacking somites.

In addition, either morphological somite boundaries or some aspects of A–P somite patterning are required for the correct segmental spatial organization of PMNs. Zebra fish mutants that lack morphological

somite boundaries and aspects of A–P somite patterning form normal numbers of CaPs and MiPs, but the precise alternation and spacing of these cell types is lost and PMNs on the two sides of the embryo form out of register with each other. This shows that PMN subtype specification and spatial organization are separable aspects of PMN patterning, and it suggests that these somite segmentation mutants are lacking signals that normally fine-tune or maintain the precise segmental organization of different PMN subtypes. However, none of these somite signals have been identified at the molecular level.

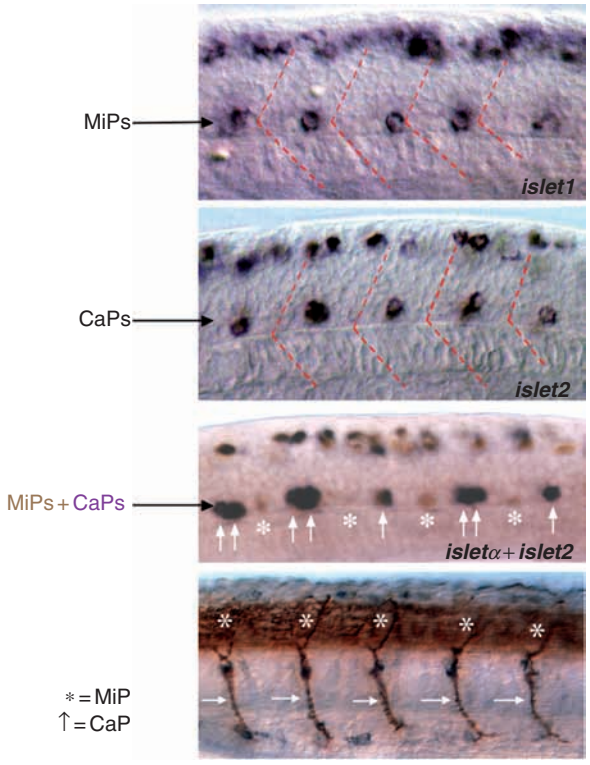

Figure 1 Zebra fish MiPs and CaPs form in different positions, express different genes, and have different axon trajectories. Lateral views of a five-somite length of zebra fish trunk 18–20 h postfertilization (hpf) (top three pictures) and 26–30 hpf (bottom picture). In all cases, dorsal is up and anterior is right. MiPs express *islet1* and line up with overlying somite boundaries (top), whereas CaPs and VaPs express *islet2* and are found adjacent to the middles of the overlying somites. Red dotted lines indicate somite boundaries. The alternation of these different PMN fates can be seen in the third picture, in which MiPs are only labeled by nuclear Islet antibody staining (small brown circles indicated by asterisks) and CaPs and VaPs are labeled with both nuclear Islet antibody staining and *in situ* hybridization to *islet2* RNA (blue cytoplasmic staining indicated by white arrows). In all three pictures dorsal Rohon–Beard sensory neurons also express *islet* genes. The bottom picture shows ZNP1 antibody staining, which labels axons. MiP axons extend dorsally (white asterisks) and CaP axons extend ventrally (white arrows).

Segmentation of Interneurons in the Zebra Fish Spinal Cord

As with motor neurons, in most vertebrates there is no evidence for a segmental pattern of spinal cord interneurons. The main exception is again the early zebra fish embryo. At least one morphologically identified interneuron class is segmentally organized in the early embryonic zebra fish spinal cord, and several transcription factors that are good candidates for being expressed by particular classes of spinal cord interneuron are initially expressed by irregularly spaced cells along the A–P axis of the spinal cord, suggesting that there may be an imprecise segmental arrangement of other interneuron classes (Figure 2). For example, ventral lateral descending (VeLD) interneurons are, at least initially, segmentally arranged: at 24 h of development, the cell bodies of one or two VeLDs can be found just dorsal to PMNs on each side of the spinal cord. In contrast, Rohon–Beard sensory neurons and commissural primary ascending interneurons do not seem to be segmentally arranged, even at this early stage. Circumferential ascending interneurons express the transcription factor Engrailed, and expression occurs initially in a roughly

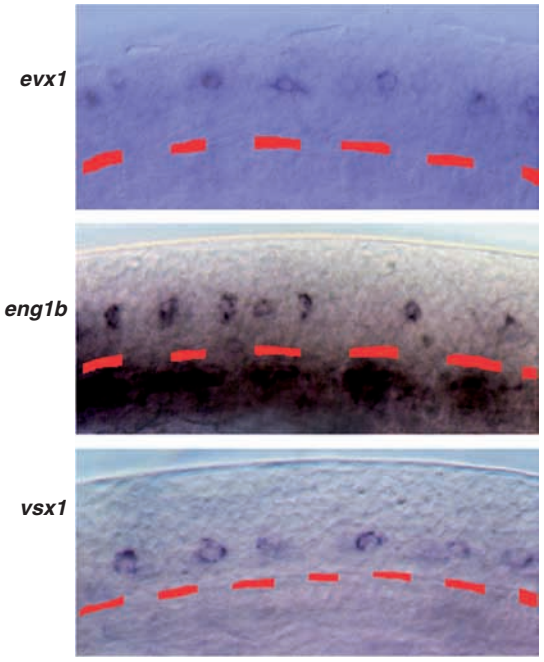

Figure 2 Transcription factors that are good candidates for labeling different interneuron classes in the zebra fish are expressed in an imprecise segmental pattern early in development. Lateral views of a five- to six-somite length of zebra fish trunk at approximately 20 hpf showing RNA expression of *evx1*, *eng1b* (*engrailed1b*), and *vsx1*. Each blue circle is an individual neuron showing cytoplasmic staining. In all cases, dorsal is up and anterior is right; dashed lines demarcate the ventral edge of the spinal cord.

segmented pattern in the spinal cord, although at later stages it becomes more continuous. This is reminiscent of *islet1* and *islet2*, which are initially expressed by PMNs but are then also expressed by later-forming secondary motor neurons, at which point expression along the A–P axis occurs in a more or less continuous row of cells and is no longer obviously segmental. A similar result is also observed for *chx10*-expressing interneurons (at least some of which are circumferential descending interneurons). Interestingly, these imprecise segmental patterns resemble the imprecise spacing of PMNs seen in A–P somite mutants, suggesting that early forming neurons in the zebra fish spinal cord may have a stochastic/imprecise intrinsic periodicity, and then some neurons are more precisely aligned with the overlying somites by signals from the somites. One possibility is that these somite signals set up segmental differences in cell adhesion along the spinal cord. This would explain why some neurons have such precise segmental distribution (they respond to differences in cell adhesion) and others do not (they are refractory to these differences). If this hypothesis is correct, then the CaP/MiP results discussed previously suggest that these cell adhesion signals should be disturbed or lacking in somite segmentation mutants.

Segmentation of Spinal Cord Neurons in Chordates and Protochordates – An Ancient Phenomenon in Chordate Evolution?

There is evidence from anatomical studies that other teleosts, lampreys, and *Amphioxus* have spinal cord motor neurons and/or interneurons that are arranged segmentally with a period equal to that of the myotome/somite series. This suggests that the segmentation seen in the zebra fish spinal cord reflects an evolutionarily ancient segmentation of the primitive spinal cord that has either been lost in tetrapods or is obscured by the larger numbers of cells (in amniotes – reptiles, birds, and mammals) or by the displacement of the somites relative to the spinal cord (e.g., amphibia such as *Xenopus*). In this context, it is notable that even in zebra fish, the spinal cord appears less obviously segmented as the embryo develops; the only clearly segmental neurons that have been observed are those that form early in development.

Peripheral Nerve Segmentation in Amniote Vertebrates

As mentioned previously, there is no direct evidence for any intrinsic segmental patterning process within the spinal cord of amniote vertebrates, yet their peripheral spinal nerves are manifestly segmented. How does this arise? In particular, since the amniote spinal cord is encased by a flexible, weight-bearing vertebral column that gives it necessary mechanical protection, how do the spinal nerves, which must make essential interconnections between the central nervous system and the tissues and organs of the body, navigate their way through this developing column? How do they avoid being buried in bone and cartilage, and how do they achieve their final position between adjacent vertebrae in the intervertebral foramina, unperturbed by bending movements of the vertebral column throughout the life of the animal? The answer lies in the relative positioning of their earliest precursors – migrating neural crest cells and the growth cones of outgrowing motor and sensory axons – and the segmental pattern of the mesodermal somites that develop alongside the developing spinal cord and construct the vertebral column.

Somite Segmentation

Somites bud off sequentially from the mesenchymal 'paraxial' mesoderm, lying on either side of the neural tube, in a rhythmical sequence, arising as pairs of epithelial spheres that flank the neural tube and accumulate in a progressive A–P (craniocaudal) direction. The segmentation process is driven by a clock mechanism, intrinsic to paraxial mesoderm cells, that involves *Notch*, *Wnt*, and *Fgf* signaling, the details of which are beyond the scope of this article. Each somite then generates a diverse set of tissues (Figure 3(a)). The mesenchymal sclerotome arises from the ventromedial somite cells, triggered by signaling proteins such as sonic hedgehog and noggin secreted from the midline notochord and floor plate and characterized by expression of transcriptional regulators such as *Pax1*, *Pax9*, and *cFkh-1*. Together with cells from the midline notochord, the sclerotome differentiates into the vertebral column. The dermomyotome derives from the remaining dorsolateral somite, differentiating into axial dermis and skeletal muscle, whereas a further group of cells between the myotome and the sclerotome, the syndetome, provides the tendons that attach segmental muscles to vertebrae.

Peripheral nerve segmentation is critically dependent on the sclerotomes. When somites first appear in the embryos of birds and mammals, neural crest cells simultaneously migrate from the dorsal neural tube and navigate the somites in two broad streams. One is dorsal, between the ectoderm and somites, and provides skin melanocytes. The second stream penetrates ventrally, its earliest cells migrating between adjacent somites to a final position alongside the dorsal aorta, where they contribute to the sympathetic ganglia. Once the sclerotome appears, however,

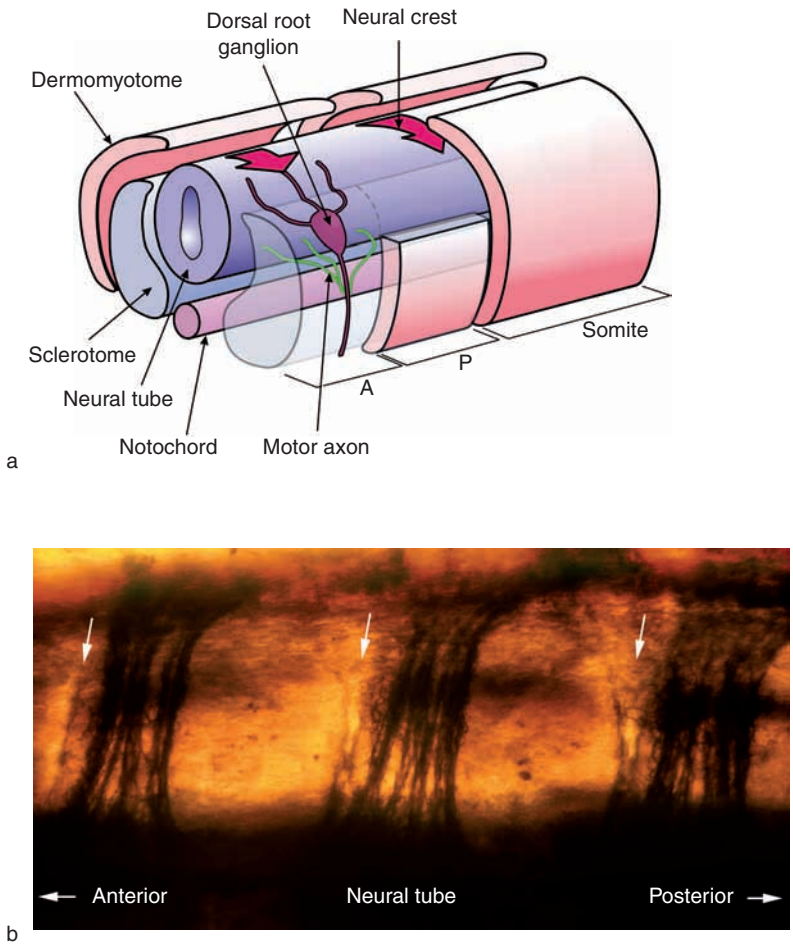

Figure 3 Development of spinal nerve segmentation in amniote vertebrates. (a) Schematic diagram showing the early components of the developing PNS (the neural crest, sensory axons from the dorsal root ganglion, and motor axons from the ventral neural tube/spinal cord) and the derivatives of the somite (dermomyotome and sclerotome). The dermomyotome is largely removed on one side to show the subdivision of the sclerotome into anterior (A) and posterior (P) halves, and the dashed line shows the A–P boundary within the sclerotome. Migration of neural crest cells (large red arrows), development of the crest-derived dorsal root ganglion, and growth of spinal motor and sensory axons are restricted to the anterior half-sclerotome. (b) Whole mount of a chick embryo at 3.5 days of incubation, stained with zinc iodide and osmium tetroxide. Motor axons extend from the neural tube on one side of the embryo and grow exclusively in the anterior (left) half of each sclerotome of the somite series. Arrows mark the boundaries between neighboring somites. From Kuan C-YK, Tannahill D, Cook GMW, and Keynes RJ (2004) Somite polarity and segmental patterning of the peripheral nervous system. *Mechanisms of Development* 121: 1055–1068.

crest cells selectively invade the anterior half of each sclerotome while avoiding the posterior half-sclerotomes (Figure 3(a)). Many of these cells will aggregate alongside the spinal cord as the segmentally arranged dorsal root (primary sensory) ganglia and alongside the aorta as the sympathetic ganglia, whereas others differentiate into peripheral nerve Schwann cells. Soon after crest cells begin to penetrate the anterior half-sclerotomes, motor axons project from the motor pools of the ventral neural tube through the sclerotomes and, like the crest, their growth cones show the same exclusive preference for anterior versus posterior half-sclerotome (Figure 3).

Peripheral nerve segmentation is not an intrinsic property of the neural tube. Instead, experimental analysis in the chick embryo has shown that it is dictated by the A–P polarity of the sclerotomes. In particular, whereas A–P reversal of a short length of neural tube does not reverse the position of axon outgrowth within the sclerotome, when the presomite mesoderm is similarly reversed, motor axons grow through the original anterior (now posterior) half-sclerotomes. Supporting this conclusion, both axon outgrowth and sensory ganglia are unsegmented after surgical construction of multiple anterior half-sclerotomes *in ovo*.

Besides PNS segmentation, the A–P subdivision of the somite is critical for the formation and/or maintenance of sclerotome boundaries. Transplantation experiments in chick embryos have shown that anterior cells are nonmiscible with posterior cells, and that

confrontations between them generate boundaries. The A–P somite polarity is also important in the segmental patterning of the vertebral column since each half-sclerotome contributes to positionally distinct elements of each vertebra. Placing axons in the anterior half-sclerotomes ensures that the spinal nerves and dorsal root ganglia will be sited in the intervertebral foramina, where they are isolated from movements of the adjacent vertebral elements.

Spinal nerve segmentation appears to be determined by the A–P sclerotome polarity in all reptiles, birds, and mammals. These terrestrial vertebrates require a weight-bearing vertebral column to be constructed early in development, and the polarized sclerotome is, accordingly, a prominent component of the somite. In fishes, the sclerotome is only a minor somite constituent during the establishment of PNS segmental patterning, and the mechanism of PNS segmentation is discussed above. Classical experiments have also shown the dependence of amphibian PNS segmentation on somite segmentation but have not resolved which somite component may be critical in these vertebrates.

Molecular Determinants of the A–P Subdivision of the Somite

Studies have shown that the generation of somite A–P polarity is linked to the clock mechanism that underlies somite segmentation. According to one model, polarized expression of the transcriptional regulator, Mesp2, in the future anterior half-somite causes arrest of Notch activity in this somite domain and subsequent generation of anterior identity; Notch activity continues in the future posterior half-somite where Mesp2 is not expressed, resulting in posterior identity and a boundary between A–P somite domains. Consistent with this, loss of the Notch ligand, Delta1, causes loss of A–P somite polarity and a single, anterior somite identity, whereas loss of Mesp2 likewise generates a single posterior identity. It remains unknown, however, what determines the anterior restriction of Mesp2 expression. Several other transcriptional regulators have been identified that show polarized expression domains in the somite, such as the T-box transcription factor Tbx6, the homeobox transcription factors Meox1 and Meox2, and the bHLH transcription factor paraxis. Identifying how these interact with other transcriptional regulators involved in establishing A–P polarity, as well as with Wnt and Fgf signaling, remain an important challenge for future research.

Molecular Basis of Peripheral Nervous System Segmental Patterning

In principle, the preference of crest cells and growth cones for anterior half-sclerotome might reflect attraction by molecules from anterior half-sclerotome and/or repulsion from posterior half-sclerotome. Several extracellular matrix proteins, including laminin, fibronectin, collagens, thrombospondin and tenascin-C, are expressed by neural crest cells and/or anterior half-sclerotome cells, and these are likely to promote migration of the crest in the anterior half-sclerotome. However, there is no evidence that these proteins determine the preference of the crest or axon growth cones for anterior half-sclerotome. Work has concentrated largely on the repulsive activity of posterior half-sclerotome cells for migrating crest cells and growth cones, and this system was among the first to provide strong evidence that cell–cell repulsion plays a major role in axon guidance and the construction of the nervous system.

Using the growth cone collapse assay to detect contact repulsion of axon growth, detergent extracts of chick somites have been shown to cause collapse of primary sensory axons and motor axons growing on laminin *in vitro*. The main repulsive mechanism is likely to be contact repulsion rather than long-distance (chemo-) repulsion. Detailed observations of interactions *in vitro* between motor axon growth cones and posterior half-sclerotome cells are consistent with contact repulsion. Moreover, chemorepulsion is not seen in collagen gel co-cultures of posterior half-sclerotomes with primary sensory axons, whereas chemorepulsion of these axons by the notochord and dermomyotome is readily visible under identical conditions. The posterior half-somite has been shown to express a variety of candidate repulsive systems that might mediate spinal nerve segmentation, and these are now considered in turn.

Eph/ephrin signaling Several members of the Eph receptor tyrosine kinase family and their ephrin ligands, initially implicated in the repulsive guidance of retinal axons projecting to the optic tectum, show polarized expression patterns in the sclerotome. Ephrins are subdivided, according to their mode of attachment to the plasma membrane, into ephrin-A ligands (attached via a GPI anchor) or ephrin-B ligands (with a transmembrane domain and cytoplasmic tail). Their cognate receptors (EphA and EphB) bind ephrin-A and ephrin-B ligands, respectively. Within each subclass, a receptor can bind to more than one ligand and vice versa, and signaling can be bidirectional; with some exceptions, there is little cross-talk between the two subclasses.

Eph/ephrin signaling has been implicated in the contact repulsion of neural crest cells by posterior sclerotome cells. EphB3 is expressed by migrating neural crest cells in chick anterior half-sclerotome, whereas the ligand ephrin-B1 is expressed reciprocally

in posterior sclerotome; in the rat, EphB2 and ephrin B2 are expressed similarly. Moreover, these studies have shown that addition of soluble ephrin-B to chick trunk explants disrupts segmental migration of the crest, and that avoidance by crest cells and motor axons of immobilized ephrin-B *in vitro* is also blocked by soluble ephrin-B. Surprisingly, however, defective spinal nerve segmentation has been reported neither in an EphB2/EphB3 double knockout mouse nor in an ephrin-B2 null mutant, perhaps because of functional redundancy in this signaling system or because additional repulsive systems operate in the sclerotome. There is also evidence that EphA4/ephrin-A5 interaction promotes crest and axon migration in anterior half-sclerotome, showing the 'bifunctional' nature of this signaling system.

F-spondin Another molecule with bifunctional properties in PNS segmentation is F-spondin, which belongs to the thrombospondin type 1 repeat (TSR) superfamily and acts as a floor plate attractant for chick spinal commissural axons. F-spondin is expressed throughout the sclerotome with the exception of the region corresponding to the early migratory pathway of neural crest, suggesting a repulsive role in crest migration. Consistent with this, crest cells migrate into posterior half-sclerotome after *in ovo* microinjection of anti-F-spondin antibodies into the embryo trunk, leading to expansion of dorsal root ganglia. F-spondin also inhibits outgrowth of cultured motor axons, although paradoxically it is reported to promote the outgrowth of rat primary sensory axons. Another member of the TSR superfamily, thrombospondin-1, acts as a permissive rather than repulsive molecule for neural crest migration. It is expressed on the myotome basal lamina, adjacent to the stream of ventrally migrating crest cells, and it promotes neural crest migration and adhesion *in vitro*.

Semaphorins Sema3A was originally identified as a secreted repellent protein causing collapse of chick primary sensory axon growth cones, and it is expressed selectively in the posterior half-sclerotome, albeit at low levels after the initial phase of neural crest migration. Its coreceptor, neuropilin-1, is expressed reciprocally on migrating crest cells in the anterior half-sclerotome, and crest cells migrating *in vitro* avoid stripes of Sema3A-Fc. For several reasons, however, it is unlikely to be a critical determinant of PNS segmentation. These include the finding that explanted posterior half-sclerotomes show no chemorepulsive activity for cultured sensory axons, and that there are no abnormalities of spinal nerve segmentation in mouse null mutants for Sema3A or neuropilin-1. On the other hand, zebra fish SemaZ1B, which is closely related to Sema3A, has been implicated in the segmental outgrowth of motor axons in this species.

Expression of another semaphorin, Sema3F, is also restricted to posterior half-sclerotome, and migrating crest cells express its coreceptor neuropilin 2 (Npn2). Strikingly, in both Sema3F and Npn2 mutant mice, anterior-restricted crest migration is replaced by uniform migration in the sclerotome, whereas somite polarity is otherwise unchanged. This system is therefore a good candidate for mediating segmental crest migration in the somites, but it is also clear that other mechanisms must be critical for generating the final segmented pattern of spinal nerves; despite the loss of segmental crest migration, segmented dorsal root ganglia and spinal nerves still develop in these mutant mice.

Peanut lectin-binding glycoproteins The search for molecular differences between anterior and posterior half-sclerotome was pioneered by the use of plant agglutinins: Among a panel of lectins, only peanut agglutinin (PNA) showed a preferential binding to posterior half-sclerotome cells in the chick embryo. PNA has a specificity for Galβ1–3 GalNAc residues, but binding is inhibited upon sialylation. Subsequent use of jacalin, a lectin whose binding to Galβ1–3 GalNAc is not blocked by sialylation, also showed preferential binding to posterior half-sclerotome cells, indicating that differential sialylation does not account for the differences in glycosylation between anterior and posterior half-sclerotome cells detected by PNA. The biological significance of this differential glycosylation is not established, but experiments conducted with detergent extracts of somites have shown that all the growth cone collapse-inducing activity present in these extracts is removed by absorption to immobilized PNA. In addition, two PNA-binding glycoproteins of molecular weight 48 and 55 kDa are localized in the posterior half-sclerotome, and antibodies raised against these proteins can also be used to remove all of the growth cone collapse-inducing activity present in detergent extracts of somites. The addition of soluble PNA to explant cultures of chick embryo trunks permits crest cells to enter both halves of the sclerotome, confirming a role for PNA-binding glycoproteins in establishing PNS segmentation. Molecular cloning of PNA-binding proteins will enable a more detailed understanding of the role of these proteins in spinal nerve segmentation.

Other molecules Several other molecules are expressed selectively in the posterior half-sclerotome

and are candidate contact repellents. T-cadherin is a GPI-linked cadherin that inhibits motor axon outgrowth *in vitro*, but its role in mediating axon segmentation *in vivo* remains unclear. Chondroitin sulfate proteoglycans (CSPGs) implicated in PNS segmentation include collagen IX and versican, and the study of versican illustrates well the potential difficulties in interpreting molecular function in this system. Versican is initially expressed throughout the sclerotome, becoming progressively confined to the posterior half as the dorsal root ganglion compacts in the anterior half. However, this redistribution may be a consequence rather than a cause of segmental crest migration since expression remains uniform following ablation of the crest. Also, when versican-coated micromembranes are implanted into the crest migration paths *in ovo*, crest cells are attracted to them rather than repelled. Like the ephrin/Eph system discussed previously, versican may regulate crest migration in a bifunctional manner, depending on which isoforms are expressed where and when.

Surround Repulsion

In addition to their A–P segmental pattern, spinal axons also adopt a particular dorsoventral trajectory. Each differentiating primary sensory neuron initially emits short extensions in random directions before producing definitive growth cones at its dorsal and ventral poles, which then grow in a bipolar fashion (Figure 4). Again, the chick embryo has provided a valuable model system for investigating how this trajectory is sustained. The influence of the tissues surrounding the dorsal root ganglia can be studied by placing tissue fragments adjacent to a dorsal root ganglion embedded in a collagen gel, the latter acting as a medium for stabilizing gradients of any diffusible molecules released from the tissues. Using this system, ectoderm, dermomyotome, and notochord have been shown to secrete potent sensory axon growth repellents. When a dorsal root ganglia, which would normally extend axons radially in a collagen gel, is sandwiched between a fragment of dermomyotome and notochord, axons extend in a polarized trajectory reminiscent of the bipolar orientation seen *in vivo*. Such experiments suggest that opposing gradients of chemorepellents cause axons to extend along the path of least repulsion, a phenomenon termed 'surround repulsion.'

CSPGs and Sema3A are expressed by the notochord, and Sema3A is also expressed by the dermomyotome. Neutralizing antibodies have shown that Sema3A and neuropilin-1 mediate dermomyotome and notochord repulsion. Studies with Sema3A- or neuropilin-1-deficient mice have indicated, moreover,

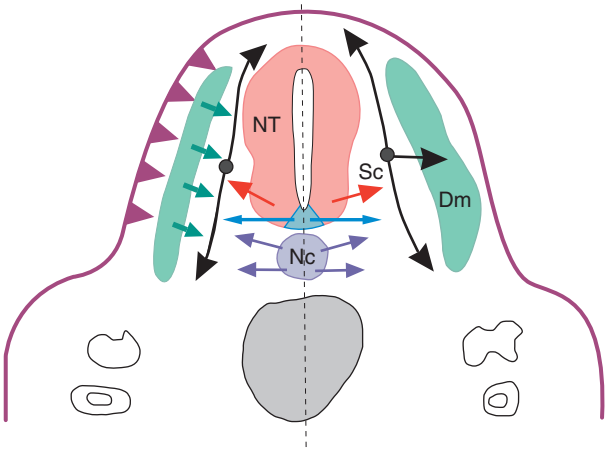

Figure 4 Surround repulsion of sensory axons. The schematic diagram shows a transverse section through the chick and mouse embryo trunk at the level of an anterior half-sclerotome. Diffusible repellents are secreted from surrounding tissues and determine the dorsoventral growth trajectory of primary sensory axons (black arrows) extending from dorsal root ganglia neurons (black circles) in the sclerotome (Sc): ectoderm (magenta arrowheads), dermomyotome (Dm; green arrows), floor plate (blue arrows) basal plate of neural tube (NT; red arrows), and notochord (Nc; purple arrows). The left half of the diagram shows normal development. The right half shows the consequence of loss of function of sema3A/neuropilin-1; repulsion from the midline is now unopposed, causing some sensory axons to project aberrantly toward the dermomyotome (black arrow to the right).

that dermomyotome repulsion is substantially due to this molecular system since sensory axons extend toward the dermomyotome in these mice (Figure 4). Evidence suggests that Sema3A acts as one component of a surround repulsion mechanism in this system. The lack of aberrant axonal projections toward the notochord in these animals (Figure 4) is not surprising because sensory axon repulsion by the notochord is only partially blocked by soluble neuropilin-Fc, and retinal axons, which are insensitive to Sema3A, are still repelled by the notochord. This suggests that the notochord secretes a number of different repellents that contribute to surround repulsion in this system.

Conclusions

During vertebrate evolution, segmentation of the trunk nervous system has transformed from an early type in which the positioning of motor and sensory neuronal cell bodies was determined by somite-derived signals and in which sensory and motor axons extend independently in the space between neural tube and somites. In birds and mammals, the polarized sclerotome forces motor and sensory spinal axons to segment together in register with the developing series of vertebrae. It is likely, then, that

during the amphibian–reptile transition a new mechanism evolved, alongside the expansion of the sclerotome, to align spinal motor and sensory axons together along the A–P axis. This transformation may have incorporated substantial molecular redundancy into the process of segmental patterning of spinal axons. It is possible, for example, that the cell types involved – neural crest cell, anterior or posterior sclerotome cell, motor axon, or sensory axon – express independent cues that mediate their segmentation. The search for further molecular mediators of spinal nerve segmentation should help to distinguish between this possibility and the alternative – that one molecular system, still to be identified, orchestrates segmentation *in vivo*.

See also: Anterior-Posterior Spinal Cord Patterning of the Motor Pool; Midbrain Patterning; Motor Neuron Specification in Vertebrates; Segmentation: Segmental Boundaries, Establishment and Morphogenesis (EPH); Transcriptional Networks and the Spinal Cord.

Further Reading

Brown M, Keynes R, and Lumsden A (2001) *The Developing Brain*. Oxford: Oxford University Press.

Davies JA, Cook GMW, Stern CD, and Keynes RJ (1990) Isolation from chick somites of a glycoprotein fraction that causes collapse of dorsal root ganglion growth cones. *Neuron* 4: 11–20.

Dubrulle J and Pourquié O (2004) Coupling segmentation to axis formation. *Development* 131: 5783–5793.

Eisen JS and Pike SH (1991) The spt-1 mutation alters the segmental arrangement and axonal development of identified neurons in the spinal cord of the embryonic zebra fish. *Neuron* 6: 767–776.

Gammill LS, Gonzalez C, Gu C, and Bronner-Fraser M (2006) Guidance of trunk neural crest migration requires neuropilin 2/semaphorin 3F signaling. *Development* 133: 99–106.

Goldstein RS and Kalcheim C (1991) Normal segmentation and size of the primary sympathetic ganglia depend upon the alternation of rostrocaudal properties of the somites. *Development* 112: 327–334.

Keynes R and Lumsden A (1990) Segmentation and the origin of regional diversity in the vertebrate central nervous system. *Neuron* 4: 1–9.

Keynes R, Tannahill D, Morgenstern DA, et al. (1997) Surround repulsion of spinal sensory axons in higher vertebrate embryos. *Neuron* 18: 889–897.

Keynes RJ and Stern CD (1984) Segmentation in the vertebrate nervous system. *Nature* 310: 786–789.

Kimmel CB, Sepich DS, and Trevarrow B (1988) Development of segmentation in zebrafish. *Development Supplement* 104: 197–207.

Kuan C-YK, Tannahill D, Cook GMW, and Keynes RJ (2004) Somite polarity and segmental patterning of the peripheral nervous system. *Mechanisms of Development* 121: 1055–1068.

Le Douarin NM and Kalcheim C (1999) *The Neural Crest*. Cambridge, UK: Cambridge University Press.

Lewis KE and Eisen JS (2003) From cells to circuits: Development of the zebrafish spinal cord. *Progress in Neurobiology* 69: 419–449.

Lewis KE and Eisen JS (2004) Paraxial mesoderm specifies zebrafish primary motoneuron subtype identity. *Development* 131: 891–902.

Morimoto M, Takahashi Y, Endo M, and Saga Y (2005) The Mesp2 transcription factor establishes segmental borders by suppressing Notch activity. *Nature* 435: 354–359.

Relevant Website

http://www.pdn.cam.ac.uk – University of Cambridge, Department of Physiology, Development and Neuroscience.

Seismic and Vibrational Signals in Animals

P M Narins and N Losin, University of California, Los Angeles, CA, USA
C E O'Connell-Rodwell, Stanford University, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Substrate-borne vibrations are a potentially important but largely unexplored signal class for communication in terrestrial animals. The detection of seismic stimuli is a phylogenetically widespread phenomenon, with known examples in every terrestrial vertebrate class. Many of these cases involve animals using seismic cues to evade predators, to locate prey, or to detect environmental changes. Although this is not considered true communication, these cases are worth further investigation since sensitivity to substrate-borne vibrations is often the scaffolding preceding the evolution of intra- or interspecific seismic communication.

The phylogenetic distribution of seismic communication *sensu stricto* is less clear. Well-documented examples of intra- and interspecific seismic communication occur among the mammals, with a few amphibian examples as well. It seems probable, however, that future discoveries will substantially broaden the range of taxa in which seismic communication is known to occur. The two most common methods of creating these signals are percussion and direct coupling of vocalizations with the ground. In the sections below, we will describe recent advances in our understanding of both seismic sensitivity and seismic communication proper in several vertebrate taxa.

Seismic Signal Generation

Percussion can produce both short- and long-distance seismic cues. A spider jumping on a web, foot-drumming in the kangaroo rat and Cape mole rat, head banging in the termite and blind mole rat, and chela drumming of male fiddler crab are examples of close-range percussion-induced signals.

The white-lipped frog (*Leptodatylus albilabris*), a terrestrial amphibian endemic to the neotropics, produces seismic vibrations concomitant with its air borne acoustic signals. During calling, the posterior portion of the frog's body is often buried in the muddy substrate, while the head and forelimbs are above ground. The seismic signal component is created while vocalizing when the frog's rapidly expanding vocal sac strikes ('thumps') the substrate impulsively. Such substrate-borne vibrations might simply be a natural consequence of the sound production mechanism in this species, but in fact males of *L. albilabris* can detect and respond to these vibrations. The sacculus in the inner ear of *L. albilabris* exhibits the highest sensitivity to substrate-borne vibrations of any terrestrial vertebrate studied. The frequency corresponding to the maximum saccular sensitivity in *L. albilabris* and the frequency spectrum of the seismic component of its vocalization match closely, suggesting that the thumps may be important for communication. Moreover, in response to either audio playback of the conspecific call or seismic playback of the vibrational call components, males of *L. albilabris* reliably modify the timing of their own vocalizations. This is the clearest demonstration of seismic communication in any anuran amphibian.

Seismic detection is not limited to adult anurans. Embryos of the red-eyed tree frog (*Agalychnis callidryas*) respond to particular vegetation-borne vibrations by hatching. The eggs of this species are laid on vegetation overhanging water so that on hatching, tadpoles drop directly into the water below. Aquatic predators may await in abundance, so it pays to delay hatching until the embryos are as mature as possible. Egg mass strikes by oophagous snakes trigger hatching *en masse*, up to 30% earlier than undisturbed clutches; other intense vibrational disturbances, such as rainstorms, do not induce hatching. Vibrations from snake attacks typically exhibit a slower temporal pattern than do rainstorm vibrations; vibration playback experiments have shown that snake recordings induced significantly more hatching than rain recordings. Stimuli that mixed snakelike and rainlike characteristics elicited intermediate levels of hatching. The *A. callidryas* embryos' sophisticated seismic detection and discrimination capabilities, although not an example of true communication, suggest that the use of substrate-borne vibrations in this group may be much more widespread than is currently recognized.

Many mammal species are known to thump their feet against the ground, and in some species these simple thumps may be rapidly repeated in complex 'drumming' sequences. These signals are best known in the Rodentia such as *Dipodomys* and *Georhycus*, but they are likely more widespread. Many kangaroo species are known to thump their feet; eastern gray kangaroos (*Macropus giganteus*) foot thump when confronted by predators. It is not clear for whom these signals are intended; two possibilities are predators or conspecifics.

Among the rodents, foot drumming is widespread, and these signals are known to propagate seismically between individual burrows. For example, individually distinctive drumming sequences have been found in the banner-tailed kangaroo rat (*Dipodomys spectabilis*). Evidence suggests that the animals may use these signals for individual recognition; drumming sequences are learned early in life and are modified with experience so that they differ from those of close neighbors. When an individual moves to a new territory, it may adjust its drumming sequence to be more recognizable among its new neighbors.

Some species, including kangaroo rats, also drum interspecifically at potential predators, particularly snakes. Snakes are highly sensitive to substrate-borne vibrations, so this antipredator behavior is often an effective signal, causing the snake to retreat. While foot-drumming signals are known to propagate through the soil, it is unknown whether conspecific recipients detect the substrate-borne vibrations directly (via somatosensory or auditory mechanisms) or via air borne vibrations created at the soil–air interface of their burrow walls. In either case, the animals are exploiting properties of seismic transmission, but more work needs to be done to understand the sensory capabilities of these organisms.

Massive terrestrial mammals may generate high-2 amplitude percussive signals; examples include foot thumping in kangaroos, trunk banging in the Asian elephant, and body slamming of male elephant seals. Although the propagation of these particular signals has not been measured through the ground, a man jumping generates a seismic disturbance that is measurable at a distance of 1 km. Locomotion of large mammals also produces ground-borne vibrations; examples include a stampede of bison or an elephant mock charge which ends in foot-stomping behavior. A running herd of zebra or giraffe generates propagated seismic waves which are distinctive at the species level and which could provide information to a predator at a distance.

The coupling of vocalizations with a substrate and the separate propagation of signals along the surface of a particular substrate, be it a plant stem, the forest floor, or open savannah, have the potential for short and long-range transmission. Distance traveled and propagation velocities of the signals depend on signal frequency and energy level as well as the properties of the substrate.

Seismic Signal Propagation

Maximum transmission efficiency of seismic energy occurs in the frequency band between 10 and 40 Hz, precisely the range in which many animals communicate seismically. In areas with little or no anthropogenic seismic noise, the 20 Hz channel is relatively noise free, save for vibrations associated with thunder or earth tremors, making it a reasonably quiet communication band. In general, for species that generate percussive substrate vibrations or low-frequency vocalizations that couple to the ground, communication signals are produced that often exhibit a spectral peak near 20 Hz and propagate along the surface of the earth as Rayleigh waves. The velocity of these seismic signals depends on the propagation velocity in soil and may be either slower or faster than the velocity of their acoustic signal counterpart. Moreover, depending on physical and environmental limitations, the seismic component may propagate within or potentially beyond the range of the airborne component.

Both airborne and substrate-borne (seismic) waves are subject to interference and alteration due to environmental factors. Variables such as wind and temperature influence airborne sound propagation, whereas the soil type and heterogeneity are among the environmental variables affecting the propagation of a seismic signal. Airborne sound waves spread spherically rather than cylindrically, attenuating more rapidly (losing 6 dB for every doubling of distance) than ground surface waves such as Rayleigh waves (losing 3 dB for every doubling of distance), which all else being equal, favors the long-distance propagation of seismic waves.

Seismic Signal Detection

The signal source intensity and attenuation during transmission, together with the sensitivity and depth of receptors in the receiver and the threshold at which the receptor will be stimulated relative to the frequency and strength of the stimulus define the active space of seismic communication signals. Detection of these signals depends on the expectation, motivation, sensitivity, and decision-making ability of the receiver, as well as the *in situ* condition – for example, the environmental backgound noise level and the signal trajectory.

There are at least two pathways for detecting seismic signals, either through bone-conducted acoustic detection or through somatosensory reception. Mammals that employ bone-conducted seismic detection have enlarged ossicles, the malleus being particularly hypertrophied, which facilitates independent oscillations of the middle ear bones relative to the skull due to inertia. This is true for species such as the blind mole rat as well as the elephant, but nowhere is this more evident than in members of the family Chrysochloridae (golden moles), which have the

largest malleus relative to body size of any animal. Moreover, recent molecular data place the Proboscidia (elephants) and the golden moles in the same clade – the Afrotheria – which also includes the Sirenia (manatees and dugongs), Hyracoidea (hyraxes), Tubulidentata (aardvarks), and Tenrecomorpha (tenrecs and otter shrews).

In elephants, vibrations travel from the substrate through the feet, up the legs and shoulders, and into the middle ear cavity, or in the case of some marine mammals and small rodents, straight through the skull. Another pathway for seismic signals in small terrestrial mammals is directly from the substrate to the skull. This is facilitated, for example, in the golden mole *Eremitalpa* by head-dipping behavior (in which the animal stops its forward motion and inserts its head in the sand) and in the blind mole rat *Nannospalax* by placing the head against the burrow wall; both behaviors appear to improve substrate–skull coupling.

Vibrations of the middle ear ossicles lag behind the skull motion due to the inertia of the hypertrophied ossicles as well as the displacement of the ossicular center of mass from the rotatory axis, resulting in inner ear stimulation. This appears to be the case for *Eremitalpa* and perhaps *Nannospalax*. In the Cape golden mole (*Chrysochloris asiatica*), there is evidence that the axis of rotation of the middle ear ossicles is frequency dependent. Moreover, low-frequency seismic stimuli result in rotation of the ossicles about the short process of the incus (roughly perpendicular to the long axis of the malleus), whereas high-frequency airborne sound causes a rotation about the long axis of the malleus. A middle ear in which the rotatory axis changes with acoustic modality may provide increased seismic acuity without compromising sensitivity to higher-frequency airborne sounds.

In addition to the enlarged ear bones, the dense fat within the foot pad of the elephant may facilitate a bone-conducted pathway and is similar to the 'acoustic fat' found in some marine mammals, particularly in the dolphin melon. This structure functions as an acoustic waveguide for the reception of acoustic communication signals ('jaw hearing'). An analogous structure, the mandibular fat body, may be used for the detection of vibrational signals by bone conduction in blind mole rats. For reception, the mandibular fat body provides a twofold increase in sound intensity by more efficiently coupling acoustic energy from the vibrating burrow to the bone and thus serving as an impedance-matching mechanism. Likewise, the elephant foot pad may perform an impedance-matching function across the air–ground interface for more-efficient seismic signal detection.

The Manatee's cartilaginous, fat-filled jaw lacunae and its skull's aerated sinuses and fatty deposits are thought to play a similar role in coupling sound to the manatee's ear. Analogously, the elephant's fatty deposits and aerated skull sinuses may facilitate low-frequency acoustic and seismic detection.

The circumference of the elephant foot just above the toenails increases by as much as 10% when the foot bears weight. The fatty digital cushion is much larger in the front than in the back foot, which may explain why elephants appear to lean forward on their front feet during periods of seismic detection, orienting their feet directly in line with the ear due to the unique graviportal structure of their forelimbs, thus facilitating efficient transmission of vibrations from foot to ear.

Anuran amphibians have in their inner ears two auditory organs sensitive to airborne stimulation: the amphibian papilla (AP), sensitive to low and middle frequencies, and the basilar papilla, sensitive to high frequencies. In addition, the sacculus responds to low-level vibrational stimuli and high-level airborne sounds. Recordings from single fibers in the eighth nerve of *L. albilabris* in response to controlled seismic stimuli revealed the greatest vibration sensitivity of any vertebrate. AP fibers exhibited clear phase-locking to a 130 Hz vibration with a peak acceleration of $0.002\ \mathrm{cm\,s^{-2}}$. In contrast, using firing rate increase as the threshold criterion, AP thresholds in both northern leopard frogs (*Rana pipiens pipiens*) and white-lipped frogs (*L. albilabris*) reached $0.01\ \mathrm{cm\,s^{-2}}$.

Seismic communication is not known definitively in any reptile, though chameleons probably possess some form of inter- or intraspecific signaling behavior via vegetation-borne vibrations. Crocodilians generate thoracic vibrations causing a showering of water at the surface prior to their high-amplitude, low-frequency vocalizations and appear to be likely candidates for further investigation of substrate-borne signal communication. In addition, dense pressure receptors located on the body and particularly around the snout of crocodilians are sensitive to water surface vibrations and are thought to be important in prey detection. These same receptors might also facilitate the detection of vibrations generated by conspecifics. Snakes are known to detect both airborne and substrate-borne vibrations through their body surface and can respond behaviorally to both airborne and substrate-borne stimuli. There is, however, no evidence of either airborne or substrate-borne acoustic communication in snakes.

One species of semifossorial snake (the Saharan sand viper, *Cerastes cerastes*) uses substrate-borne vibrations to detect its prey. The snake often strikes prey from a completely submerged position in the

loose sand of the Sahara desert. In the laboratory, *C. cerastes* can strike with accuracy even when its olfactory and vomeronasal organs are denervated. Occluding the eyes reduces striking accuracy somewhat, but the animals are still able to capture live mouse prey. Even with chemosensory-neutral, temperature-neutral targets, these snakes strike with a fair degree of accuracy, providing circumstantial evidence for the importance of vibration sensitivity. Sand vipers, with their sinuous bauplan, are suited to a fossorial lifestyle, and it might be expected that other snake species also utilize vibrational cues in prey location, and perhaps in communication as well. Further work is needed to elucidate the mechanism of seismic detection in snakes.

A second pathway for seismic detection that terrestrial species might employ in addition to or instead of bone conduction is somatosensory reception. Pacinian corpuscles, or pressure receptors, the largest cutaneous mechanoreceptors in mammals (0.5 mm in diameter) are deeply placed whereas Meissner's corpuscles, or touch receptors, are superficially located and smaller (40–60 μm in diameter). Pacinian corpuscles in humans respond over a frequency range of 65–400 Hz, with peak sensitivity at approximately 250 Hz, whereas Meissner's corpuscles exhibit best sensitivity between 10 and 65 Hz.

Many small and large mammals are equipped to detect seismic cues via somatosensory reception. One notable example is the star-nosed mole (*Condylura cristata*), with an elaborate snout surrounded by 22 fleshy, mobile appendages covered with thousands of mechanoreceptive Eimer's organs, which act like a 'tactile eye.' Pacinian corpuscles found in the paws and knees of cats, lamellated corpuscles in the legs of kangaroos, and vibrationally sensitive Herbst corpuscles in the legs of birds are all thought to facilitate prey detection and predator avoidance. Both Pacinian and Meissner's corpuscles can also be found around the tip of the Asian elephant trunk and are as sensitive as the lips of a primate.

Enhancing Signal Detection Range

Temporal summation can be responsible for significant auditory threshold reduction; thus repeated signals can facilitate signal detection. Moreover, longer signals can increase the signal-to-noise ratio, improving signal detection as well. Examples of this can be found in animals communicating in both the acoustic and the seismic realms.

Weddell seals (*Leptonychotes weddelli*) repeat calls to enhance signal detection over long distances and during times of high background noise, as do some whales. Killer whales increase call duration in high background noise from boats. It follows, therefore, that species that communicate seismically would also employ similar mechanisms to enhance the propagation and detection of signals in noisy environments or over long distances.

To this end, African elephant family groups vocalize within interactive bouts more often than while calling individually. These bouts create an enhanced signal that is effectively three times longer than one produced by a single individual, serving to increase the signal-to-noise ratio and facilitating signal detection and processing at great distances. Furthermore, during departure from a resource, calling bouts are repeated at a greater rate. Elephants listening at a distance might use multiple bouts of longer signals to adjust their physical position to optimize signal resolution. Observations suggest they do just this by freezing for long periods and shifting positions, aligning themselves in the direction of the acoustic or seismic signal source.

Seismic Frequency Discrimination

Elephants can discriminate very subtle frequency differences in the seismic environment. If the pathway for detecting these signals is via bone conduction to the ear, then the seismic frequency discrimination ability should be correlated with the airborne frequency discrimination ability of this species. The acoustic critical bandwidth for the elephant is calculated at 0.75–0.95 Hz across a range of 15–19 Hz, allowing them to detect very small changes in frequency modulation across their low-frequency vocalizations in both the acoustic and the seismic realms.

If the seismic detection pathway is via vibration-sensitive corpuscles, then elephants should still be able to discriminate fine frequency differences, assuming they are at least as sensitive as primates at this task. The tactile frequency discrimination difference limen of adult owl monkeys, for example, is 2 Hz in the region of 20 Hz, which is similar to that of macaques and humans. Since elephants often lay their trunks on the ground during seismic events, they may make use of vibration-sensitive corpuscles described from the trunk, as well as those in their feet, to distinguish frequency modulation between different calls.

Summary

Seismic stimuli are used to navigate within the environment, enhance localization, and detect predators or prey; for intraspecific communication; and possibly as a cue to initiate migration. For percussive signals, large size is often associated with greater source amplitudes, leading to a greater propagation range. Moreover, the use of low-frequency, high-amplitude signals favors more-efficient coupling of airborne

vocalizations to the substrate and increased propagation range.

White-lipped frogs, kangaroo rats, blind mole rats, Cape mole rats, and golden moles appear particularly adept at generating and detecting seismic signals in the face of high levels of background noise. Moreover, they respond to playbacks of seismic signals by changing their signaling behavior or by moving. The seismic components of the foot-drumming signals of the Cape mole rat (*Georychus*) are capable of propagating in the sandy soil between natural burrows of these solitary rodents with little attenuation, whereas their auditory components attenuate into the background noise within a meter of the source. Thus, seismic signaling is strongly implicated in this species as well.

Elephants appear to use long-distance seismic communication by producing low-frequency, high-amplitude vocalizations that efficiently couple to the substrate, where they propagate over long distances and attenuate slowly. Moreover, they are equipped to detect seismic signals. Since of all extant terrestrial animal species, Asian elephants have the largest volume of cerebral cortex available for cognitive processing, we suggest they are likely well-equipped to integrate multimodal signals.

In theory, if a seismic eavesdropper, such as a lion on the African plain, were able to identify a potential prey item such as an antelope or a zebra by its seismic signature, it could be more selective in its hunting forays for preferred food. Likewise, elephant eavesdroppers might learn of the availability of a resource at great distances, which would facilitate intergroup spacing.

The ability to tap into the seismic environment to discriminate biologically relevant information from background noise and to discriminate subtle frequency differences in signals indicates that vertebrates may be using the ground as a communication channel for obtaining much more subtle cues than previously imagined.

See also: Auditory/Somatosensory Interactions; Communication in Frogs and Toads; Communication in Terrestrial Animals; Electrical Perception and Communication; Kinship Signals in Animals; Localizing Signal Sources; Multimodal Signaling in Animals; Signal Identification: Peripheral and Central Mechanisms; Signal Transmission in Natural Environments; Somatosensory Perception.

Further Reading

Barnett KE, Cocroft RB, and Fleishman LJ (1999) Possible communication by substrate vibration in a chameleon. *Copeia* 1999(1): 225–228.

Borg E (1997) Cutaneous senses for detection and localization of environmental sound sources: A review and tutorial. *Scandinavian Audiology* 26(4): 195–206.

Catania KC and Remple FE (2004) Tactile foveation in the star-nosed mole. *Brain, Behavior and Evolution* 63(1): 1–12.

Christensen-Dalsgaard J and Narins PM (1993) Sound and vibration sensitivity of VIIIth nerve fibers in the frogs *Leptodactylus albilabris* and *Rana pipiens pipiens*. *Journal of Comparative Neurology* 172: 653–662.

Hill PSM (2001) Vibration as a communication channel: A symposium. *American Zoologist* 41: 1133–1240.

Ketten DR (2000) Cetacean ears. In: Au WWL, Popper AN, and Fay RR (eds.) *Hearing by Whales and Dolphins*, pp. 43–108. New York: Springer.

Lewis ER and Narins PM (1985) Do frogs communicate with seismic signals? *Science* 227: 187–189.

Lewis ER, Narins PM, Jarvis JUM, Bronner G, and Mason MJ (2006) Preliminary evidence for the use of microseismic cues for navigation by the Namib golden mole. *Journal of the Acoustical Society of America* 119: 1260–1268.

Markl H (1983) Vibrational communication. In: Huber F and Markl H (eds.) *Neuroethology and Behavioral Physiology*, pp. 332–353. Berlin, Germany: Springer.

Mason MJ (2004) Functional morphology of the middle ear in *Chlorotalpa* golden moles (Mammalia, Chrysochloridae): Predictions from three models. *Journal of Morphology* 261: 162–174.

Mason MJ and Narins PM (2002) Seismic sensitivity in the desert golden mole (*Eremitalpa granti*): A review. *Journal of Comparative Psychology* 116: 158–163.

Narins PM (1990) Neuroethology of seismic communication in anuran amphibians. *Bioscience* 40: 268–274.

Narins PM, Lewis ER, Jarvis JJUM, et al. (1997) The use of seismic signals by fossorial southern African mammals: A neuroethological gold mine. *Brain Research Bulletin* 44: 641–646.

O'Connell-Rodwell CE, Arnason BT, and Hart LA (2000) Seismic properties of Asian elephant (*Elephas maximus)* vocalizations and locomotion. *Journal of the Acoustical Society of America* 108: 3066–3072.

O'Connell-Rodwell CE, Wood JD, Rodwell TC, et al. (2006) Wild elephant (*Loxodonta africana*) breeding herds respond to artificially transmitted seismic stimuli. *Behavioural Ecology and Sociobiology* 59(6): 842–850.

Rado R, Terkel J, and Wollberg Z (1998) Seismic communication signals in the blind mole rat (*Spalax ehrenbergi*): Electrophysiological and behavioral evidence for their processing by the auditory system. *Journal of Comparative Physiology* 183: 503–511.

Randall JA and Lewis ER (1997) Seismic communication between the burrows of kangaroo rats, *Dipodomys spectabilis*. *Journal of Comparative Physiology* 181: 525–531.

Saxod R (1996) Ontogeny of the cutaneous sensory organs. *Microscopy Research and Technology* 34(4): 313–333.

Shoshani J (1996) Skeletal and other basic anatomical features of elephants. In: Shoshani J and Tassy P (eds.) *The Proboscidea: Evolution and Palaeoecology of Elephants and Their Relatives*, pp. 9–20. Oxford, UK: Oxford University Press.

Soares D (2002) Neurology: An ancient sensory organ in crocodilians. *Nature* 417: 241–242.

Vliet KA (1989) Social displays of the American alligator *(Alligator mississippiensis)*. *American Zoologist* 29: 1019–1031.

Warkentin KM (2005) How do embryos assess risk? Vibrational cues in predator-induced hatching of red-eyed treefrogs. *Animal Behaviour* 70: 59–71.

Willi UB, Bronner GN, and Narins PM (2006) Ossicular differentiation of airborne and seismic stimuli in the Cape golden mole (*Chrysochloris asiatica*). *Journal of Comparative Physiology* 192: 267–277.

Young BA (2003) Snake bioacoustics: Toward a richer understanding of the behavioral ecology of snakes. *Quarterly Review of Biology* 78: 303–325.

Self-Organizing Maps

R Martin and K Obermayer, Bernstein Center for Computational Neuroscience and Berlin University of Technology, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Self-organization is found at various levels of the brain, ranging from global areal organization and modularization to regularities in the fine structure of neural layers. Equally variable are explanations for this organization. Genetic regulation, chemical markers, and cell death account for much of the larger-scale specialization of areas, for anatomical differences observed in different regions and layers, and for differential afferentiation.

One ubiquitous property of the brain not fully accounted for by these mechanisms is the organization into topologically ordered maps. Smooth representations of features processed by the brain range from representations of visual space (Figure 1) and orientation in early visual areas to representations of motor output and of semantic content in higher brain areas. Activity-driven self-organization appears to play a major role in their development. The attempt to understand this process has inspired a number of models, one of which is the self-organizing map (SOM; sometimes called the self-organizing feature map).

Based on Hebbian learning and lateral inhibition, the SOM organizes its units into topographically ordered maps. The original SOM provided an abstraction of a class of processes underlying neuronal pattern formation, where units represented local populations of neurons and the algorithm was inspired by, rather than derived from, an underlying neuronal dynamics. However, later models also implemented SOM-like algorithms with units explicitly representing cortical columns, and there are also models which implement the SOM algorithm based on spiking neurons that employ a temporal code. Independent of these neuroscience aspects providing insight into the workings of the brain, the SOM has also developed into an important machine learning method.

History

Biologically plausible candidate principles for activity-driven self-organization of brain areas were identified early on, both in experimental and in theoretical neuroscience: the principle of Hebbian learning and that of lateral inhibition. Early studies combining these principles in neural network models to explain the spatial organization of brain function include the spatially ordered line detectors of von der Malsburg and the neural field model of Amari. In these studies, lateral inhibition resulted in the creation of localized activity blobs, and Hebbian reinforcement led first to the localization of features in the output map and then to a dependence of the projection zone size on frequency of stimulation. Kohonen suggested considering network activity and the synaptic modifications as two separate processes. The lateral inhibition essentially implements a winner-takes-all (WTA) mechanism, which confines activity in response to a stimulus to one subregion; a plasticity-control system restricts Hebbian learning to the winner and its neighborhood. Maps developing according to these principles are referred to as SOMs.

Consider using these rules on a representation of input features (e.g., position and orientation) in a map of model neurons. The WTA principle results in the grouping of input properties because different feature combinations can be assigned to the same unit. On the other hand, Hebbian learning restricted to the region around the winning unit creates a smooth representation. Similar input patterns are represented by nearby units of the map. Thus, the SOM combines two complementary principles of unsupervised learning: clustering and smoothness. Therefore, it is similar to many of the biological maps found in the brain.

Alternative models suggested for development of feature maps include elastic nets, wire length minimization models, and correlation-based learning. These are not reviewed in detail in this article.

Method

Low-Dimensional SOMs

In its simplest form, the algorithm operates on input which already contains an abstract representation of the features to be encoded. The variant that uses a rich high-dimensional input is discussed later.

A system such as the primary visual cortex of cat can be described as a two-dimensional neural sheet. Disregarding cortical depth, which by virtue of columnar organization leaves some response properties invariant, each unit is assigned a vector of response properties, giving the vector field or map $\bar{w}(\bar{r})$, where $\bar{r}$ denotes cortical location and $\bar{w}$ quantifies the response properties of the unit (i.e., the neurons at this location or in this cortical column). Maps are not necessarily restricted to two dimensions in $\bar{r}$. For example, a description of the functional organization of cat lateral geniculate nucleus could be based

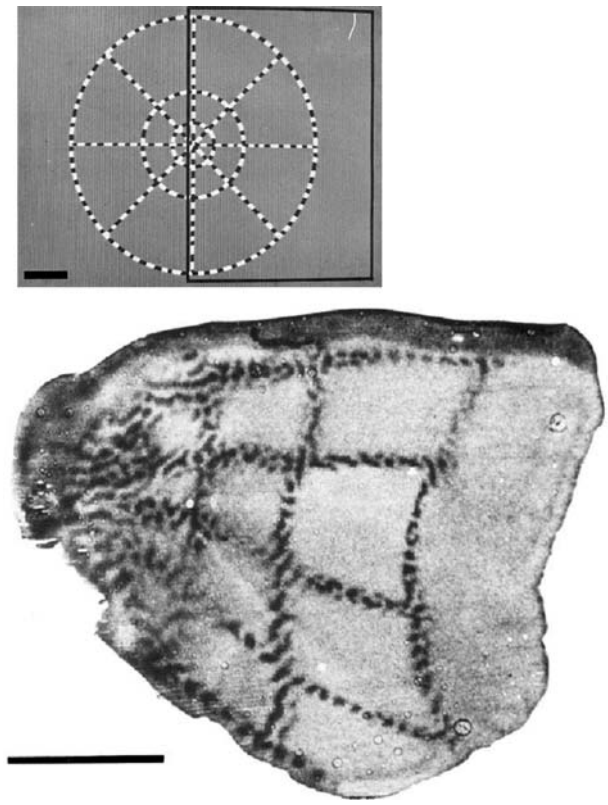

Figure 1 The retinotopic projections from the retina onto the visual cortex of macaque monkey. The top shows the image the animal was exposed to and the bottom the activity in visual cortex during exposure revealed by processing deoxyglucose uptake. Scale bars indicate 2° of visual angle (top) and 1 cm (bottom). Reproduced from Tootell RBH, Switkes E, Silverman MS, and Hamilton SL (1988) Functional anatomy of macaque striate cortex: II. Retinotopic organization. *Journal of Neuroscience* 8: 1531–1568. Copyright 1988 by the Society for Neuroscience.

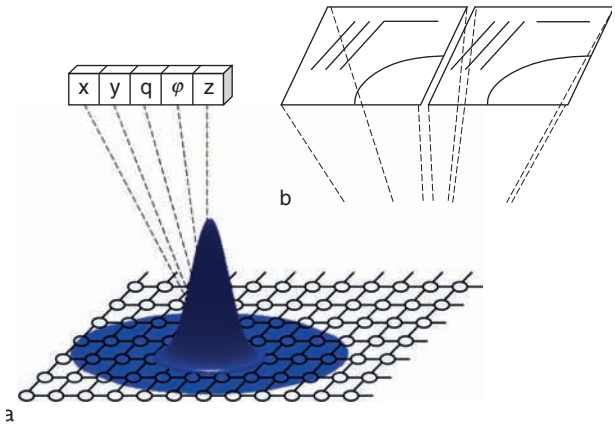

Figure 2 A representation of the network layout of a SOM. (a) A low-dimensional SOM receiving feature vector input coding for position *x* and *y*, orientation specificity *q* and orientation preference φ, and ocular dominance *z*. Connections to just one of the network units $\bar{r}$ are shown as well as the neighborhood function for this unit. (b) A high-dimensional SOM input layer corresponding to the low-dimensional SOM on the left. Two image patches provide information about features (position, orientation, and, possibly, intensity) for each of the two eyes. Topographic representations for such input take longer to arise. The representations can be compared to the corresponding low-dimensional representation through analysis of the connectivity matrix (dominant eigenvectors of left and right eye image determine orientation preference, etc.).

on a three-dimensional map $\bar{w}(\bar{r})$, where visual space varies along two dimensions of $\bar{r}$ and ocular dominance along the third (Figure 2(a)).

The SOM is an iterative algorithm which operates on such a map. Each stimulus is described by a vector $\bar{x}$ of input patterns or features, where the elements of $\bar{x}$ correspond to elements of the vector $\bar{x}$ of response properties. If $\bar{w}(\bar{r})$ describes the localization of receptive fields in visual space, the vector $\bar{x}$ describes the retinotopic location of a stimulus. Given a set of input patterns occurring with probability $P(\bar{x})$, the following algorithm describes self-organized map formation:

- Present stimulus $\bar{x}$ with probability $P(\bar{x})$.
- Determine the location $\bar{r}_0$ in the map with response properties most similar to stimulus $\bar{x}$ (WTA rule). Often, the Euclidean distance is used – that is,

$$\bar{r}_0 = \text{argmin}_{\bar{r}}(\bar{x} - \bar{w}(\bar{r}))^2 \qquad [1]$$

- Change the vectors $\bar{w}(\bar{r})$ of response properties according to

$$\Delta\bar{w}(\bar{r}) = \eta h(\bar{r}, \bar{r}_0)(\bar{x} - \bar{w}(\bar{r})) \qquad [2]$$

where η is the learning step. A common choice for the neighborhood function is

$$h(\bar{r}, \bar{r}_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} \exp\left(\frac{(\bar{r} - \bar{r}_0)^2}{2\sigma^2}\right) \qquad [3]$$

Both learning step η and width σ of the neighborhood function are normally changed during the iteration of the previous algorithm from initially large to small (for σ) or zero (for η) final values. This annealing procedure helps obtain good convergence in the face of the frequent local minima in the model space.

Convergence Behavior

The neighborhood function ensures that neighboring locations in the map perform similar learning steps, leading to a smooth progression of response properties across the final map. However, the precise functional organization of the resulting map depends on a number of parameters. The overall representation is determined by the probability distribution of input patterns and by the final width of the neighborhood function. The fine spatial structure depends more on

the initial and boundary conditions and on the precise change in η and σ during iterations.

The convergence of the SOM has not been demonstrated in general; however, results have been obtained for specific geometric configurations. If the Euclidean distance $(\bar{x} - \bar{w}(\bar{r}))^2$ in the assignment step (eqn [1]) is replaced by its local average, $\sum_{\bar{r}'} h(\bar{r}, \bar{r}')(\bar{x} - \bar{w}(\bar{r}'))^2$, one can demonstrate that the SOM algorithm exactly minimizes the weighted average squared distance

$$E = \frac{1}{2} \sum_{\bar{x}} \sum_{\bar{r}} h(\bar{r}, \bar{p}(\bar{x}))(\bar{x} - \bar{w}(\bar{r}))^2 \qquad [4]$$

between input patterns and the feature vectors. Here, $\bar{p}(\bar{x})$ is the nearest unit according to the WTA rule. Both the closest unit and its neighbors contribute to the average, but squared distances are weighted by the neighborhood function. Feature vectors converge to the center of mass of the data assigned to their unit as well as to neighboring units, all weighted by the neighborhood function.

In the limit of a zero-range neighborhood function, eqns [1]–[3] are identical to the K-means clustering algorithm, which minimizes the average squared distance between the input data and the feature vectors of the units to which the input data are assigned (vector quantization). Feature vectors converge to the center of mass of their own data cluster. However, in this limit, the organizing power of the algorithm is completely lost.

Other Variants

The algorithm described previously can be referred to as the low-dimensional SOM since it provides an abstract representation of the input in the form of a feature vector with intensities corresponding to the features to be encoded in the map. This requires advance insight into the problem to be modeled. The most suited parameterization is not always obvious; also, different parameterizations of the same problem may lead to different results.

The high-dimensional alternative has an input layer, which provides a rich representation of the features as would be received by the system to be modeled (e.g., images contain a rich representation of local features such as color, orientation, and spatial structure; **Figure 2(b)**). The winning region for particular inputs, as determined by the WTA rule, no longer lends itself to interpretation as easy as that for the low-dimensional variant. However, through an analysis of the connectivity matrix for the winning unit, one can determine which features in the input (e.g., which pixel combinations) are responsible for a high response. Naturally, given the higher dimensionality, the high-dimensional SOM converges more slowly and its convergence depends more critically on the annealing procedure.

The use and interpretation of the parameters can also vary in different SOM applications. In modeling, the learning step η sometimes retains nonzero values to reflect continued plasticity of the network. The remaining competition between the different stimulus classes can result in a reorganization of the map if the input statistics change.

Soft topographic vector quantization contains a probabilistic version of the SOM as a special case. It offers a soft nonlinearity using a temperature hyperparameter with a deterministic annealing scheme to control the convergence and avoid local minima. Thus, σ annealing is not necessary for controlling good convergence, and σ can be used solely to determine the desired neighborhood structure of the map.

Over time, a number of variants of the standard SOM have been developed to reflect different goals. Soundly based on information-theoretic principles yet biologically plausible, van Hulle introduced a self-organized map in which the area representing a part of feature space is proportional to the density of the data in that part of space, thus providing load balancing for local calculations to be carried out on the map. Other variants include a time-adaptive SOM with adaptive learning rates and neighborhood functions for individual neurons, a parameterless SOM that removes the problem of appropriate choice of local neighborhood and learning rate by proving an internal state variable to achieve stable convergence, a batch-learning version of the SOM that improves convergence properties and provides a tool for stochastic approximation, and kernel-based variants of the SOM. Thus, SOMs remain an area of active research, although much of the focus has shifted to the domain of machine learning problems.

Properties

Clustering

As explained previously, grouping is achieved according to similarity because stimuli are assigned to units (clusters) minimizing a distance measure (eqn [1]). Algorithms have been provided to achieve efficient clustering for similarity measures other than the standard Euclidean ones (general topographic proximity mapping).

Clustering behavior is also well documented in feature maps in the brain. Similar input features are clustered in similar regions of the map, and the size of these regions depends on the probability of encountering these features.

Mapping

If the neighborhood function (eqn [3]) has finite range, neighboring units in the map perform similar learning steps according to eqn [2]. Thus, weight vectors of neighboring units converge to similar values, resulting in a smooth or neighborhood-preserving map of input space. In contrast to the clustering property, however, where a quality measure for grouping is directly optimized by the SOM, the mapping property is an implicit result of the neighborhood function (eqn [3]). Measures for smoothness and for the degree of neighborhood preservation have been designed, and it has been shown that the SOM in general improves representations according to these measures, but there is no evidence that an optimum is actually achieved. One of the most surprising features of the SOM is that without actually explicitly minimizing any smoothness measure, such near-optimal maps are nonetheless achieved.

Many of the maps found in different brain model systems are also neighborhood preserving. This may apply on different scales for different features represented in the map, as will become clear in the discussion of dimension reducing mappings.

Topographic Maps

If the dimension of the input data and the dimension of the map are equal, the representation is called a topographic map. Utilizing a formal relationship between clustering methods and mixture models, it can be shown that SOMs tend to generate representations where the number of input data per cluster and the size (variance) of the clusters are constant. If certain classes of stimuli appear more often than others, however, the constraints are incompatible, and the algorithm converges to a compromise solution. Equal size requires the number of units with feature vectors close to the input feature to be proportional to the occurrence probability of this feature; equal variance requires the number of units to be independent of this probability. The compromise is a representation in which the number of units increases sublinearly with occurrence probability.

Many of the maps found in the brain can be called topographic. Most prominently, the maps of visual space in early visual areas are organized retinotopically (**Figure 1**). However, maps in other modalities (e.g., tonotopic and somatotopic maps), maps representing the exterior environment, and maps producing motor output are also topographic. Properties organized in topographical fashion locally (local orientation, spatial frequency, or ocular dominance in V1) are considered later.

Magnification Factors

If the dimensions of input data and map are equal, one can define magnification factors based on volume (the ratio between a certain volume on the map and the volume in input space mapped onto this target region) or length changes (the ratio between distance along one direction in the map and corresponding distance in input space). Magnification factors are larger for regions in data space with a higher input density, and they are influenced by boundary conditions. If there is a mismatch between the shape of the data region in input space and the shape of the map, anisotropic magnification factors emerge or, in extreme cases, there are discontinuities in the mapping.

Magnification factors are also encountered in brain maps. Well-known is the magnification of the fovea in retinotopic V1, where the V1 area representing the fovea is larger than the already high-resolution representations found in retina and lateral geniculate nucleus. This may well be a consequence of the fact that input received via the fovea tends to show higher variability due to eye movements directed to features with more complex local structure. Other examples of magnification factors are found in the representation of the body surface. Areas which receive much distinct input (fingers, lips, and genitals) or input with high behavioral relevance (the index finger of brail readers) have larger representations in somatosensory cortex. Similarly, in the natural environment, more frequent vertical and horizontal orientations are represented by a larger number of neurons in V1 than other orientations, and the frequency range used by bats for navigation has an enlarged representation in the tonotopic map of bat auditory cortex.

Dimension Reducing Mappings

If the dimension of the input data is larger than that of the map, the representation is dimension reducing; its structure depends on the variance of the data distribution in input space. For example, when mapping two-dimensional input data with one highly variable and one not so variable dimension onto a one-dimensional map, the small variance dimension is discarded and only the large one will be represented. This automatic selection of feature dimensions generalizes to higher dimensions and to curved data manifolds. The suppression of the less prominent dimensions occurs up to a critical level of variance. Above this value, a hierarchical representation emerges, in which the direction of largest variance determines the coarse structure of the map and less variable directions in input space are represented on a smaller scale. The critical value increases with increasing range of the neighborhood function.

As long as the neighborhood function has a finite range, dimension reducing maps are also smooth in the sense that response properties of neighboring neurons are similar. However, there is a trade-off between the rates of change along different input dimensions. Generally, if one response property changes rapidly, the other response property changes slowly.

The emergence of such hierarchical representations of multiple stimulus feature dimensions is best illustrated with the well-studied feature maps of primary visual cortex as found in cat and macaque. Superimposed on the retinotopic presentation of visual space, orientation of local stimuli is represented around pinwheel centers; ocular dominance and possibly also spatial frequency selectivity are represented in stripes along the cortical surface, and color information is represented mostly in the cytochrome oxidase blobs (Figure 3a). Of these features, orientation represents a noteworthy special case because its representation around pinwheels also demonstrates a topological defect: The cyclical property orientation is mapped onto a flat surface, resulting in the creation of a singularity at the pinwheel center.

Coverage and Continuity

It has been proposed that cortical maps in the brain are often organized according to the principles of coverage and continuity. Coverage requires all combinations of response properties to occur. For example, in macaque V1, there should be groups of cells responsive to stimuli for every combination (or small intervals) of spatial location, orientation, ocularity, and spatial frequency. If this requirement is not met, 'perceptual holes' emerge and cortex is less responsive to one sort of localized stimuli than to others. Continuity requires response properties to vary smoothly along most of the cortical surface and discontinuities to be rare.

The SOM model generates maps according to these principles, even though measures of coverage and continuity are not optimized directly. Coverage is achieved because of the clustering property. Continuity emerges from the fact that neighboring units in the map perform similar learning steps and, therefore, converge to similar feature vectors.

Analysis of optical recording data from visual cortical maps has shown that coverage and continuity are indeed better for the experimentally observed arrangements of response properties than for random superpositions of visual feature maps.

Applications

The previous section highlighted many formal similarities between properties of SOMs and maps found in brains. Modeling studies of SOMs actually go beyond this and produce explicit parallels between maps generated in computational models and maps encountered in various *in vivo* model systems.

Visual Feature Maps

The vertebrate visual system is arguably the system in which SOMs have had the greatest impact. Early qualitative studies of self-organization in this system as pioneered by von der Malsburg were taken to a new level through optical recoding. The maps that optical recording unveils and the richness in visual features represented in the areas of visual cortex have

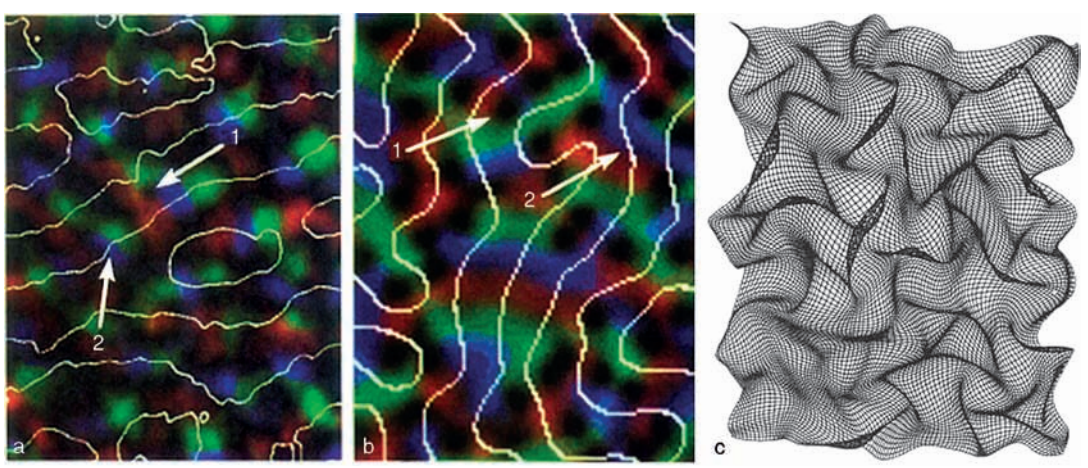

Figure 3 Orientation and ocular dominance in visual cortex. Color and brightness code for orientation tuning and strength; superimposed white lines show the borders of ocular dominance stripes. (a) Experimental results from adult macaque. (b) Stationary result of a SOM simulation. (c) Spatial structure of the topographic projection. Reproduced with permission from Obermayer K, Blasdel GG, and Schulten K (1992) Statistical–mechanical analysis of self-organization and pattern formation during the development of visual maps. *Physical Review A* 45: 7568–7589. Copyright 1992 by the American Physical Society.

allowed the demonstration of quantitative matches between the maps observed *in vivo* and those generated in model studies. Obermayer and co-workers first applied Kohonen's principle of self-organization to primary visual cortex and demonstrated that it can account for co-development of retinotopic organization and orientation columns (Figure 3).

Subsequent successes include the extension of this to other feature maps, such as ocular dominance and spatial frequency selectivity. Indeed, SOMs account for the interaction between the different feature maps: The higher density of pinwheel formations in the center of ocular dominance columns and the linear change of orientation preference along the border of the ocular dominance slabs could be attributed to the development of hierarchical representations of the different features represented in the map combined with the tendency of the SOM to keep the variance within the clusters constant. In fact, Sur and co-workers demonstrated that *in vivo*, high-gradient changes of different features in maps tend to occur perpendicular to one another, as SOM models predict, and that *in vivo* reorganization of visual cortex following the removal of one of the input features is also well described by an SOM.

Detailed analysis of the conditions under which ocular dominance and orientation selectivity in retinotopic maps develop shows that for stable emergence of the columnar structure, the neighborhood function (*in vivo*, the lateral connectivity) needs to be below a critical value. The fact that lateral interaction cannot be completely absent, however, was demonstrated by Piepenbrock and co-workers. They provided a generalized framework of soft competition with varying degrees of interaction for the development of topographic maps. It contains SOMs as well as correlation-based learning as limit cases and establishes that at least for Gaussian-shaped input patterns, the model must have a certain degree of lateral interactions (and, thus, SOM-like makeup) in order to develop topographic receptive fields. Interestingly, there was a sharp phase transition to maps without topographic structure when lateral interaction was decreased too much.

Boundary Conditions

Further applications indicating similarities in the development of local structure in SOMs and brain areas derive from simulations using elongated target maps. The simulations predict that, in this case, island-like representations will appear, thus violating the overall topography of the map while maintaining a locally topographic representation. These results are in agreement with experimental data indicating that visual areas with elongated shape (as found in owl monkey, cat, and mouse) will display an island-like structure (Figure 4).

Experimental observations of feature map development in visual cortex also allow the assessment of the influence of initial conditions on map formation. Wolf and co-workers analytically studied map formation for general self-organized map models. Assuming translational, rotational, and orientation shift invariance of the orientation maps found in V1, they demonstrated that early during development, maps have a certain minimum pinwheel density. Experimental data for pinwheel densities show that in some species (e.g., cat, squirrel monkey, and tree shrew), this initial orientation map structure undergoes a process of further refinement, during which a number of pinwheels are annihilated by merging with nearby pinwheels with opposite chirality, thus decreasing pinwheel density. Maps with strong ocular dominance structure, however, are more resistant to such map reorganization, consequently showing a larger density of pinwheels (e.g., macaque).

The co-development of lateralized feature map representations has also been shown to be compatible with activity-dependent synaptic changes. Provided that a crude initial mirror symmetric map is present and that the distribution of the sensory afferents is sufficiently divergent, mirror symmetric maps as frequently observed in adjacent topographic maps can develop.

Plasticity

Both the structure of maps observed in the brain and also the changes these maps undergo depending on their input resemble those of maps created using SOMs. The representation in the map encodes the most variable directions in the manifold of feature space and is thus adjusted to the statistics of the stimuli the system has to analyze; a change in these statistics results in reorganization of model maps to a new equilibrium. Similar behavior can be observed in maps *in vivo* when altering the input to certain brain areas. Kittens with eyes sewn shut do not develop ocular dominance and orientation columns in primary visual cortex, whereas rearing kittens with just one eye shut results in the absence of corresponding ocular dominance stripes and rearing them in an environment containing only horizontal or vertical structures will result in an impairment of their ability to see other orientations.

These effects tend to be most pronounced or even restricted to a critical period in the early life of the animal, typically a few weeks. Kittens without visual input for a sufficiently long period will remain blind, despite normally organized eyes and lateral geniculate nucleus. However, there is also evidence for a certain degree of reorganizing capacity remaining

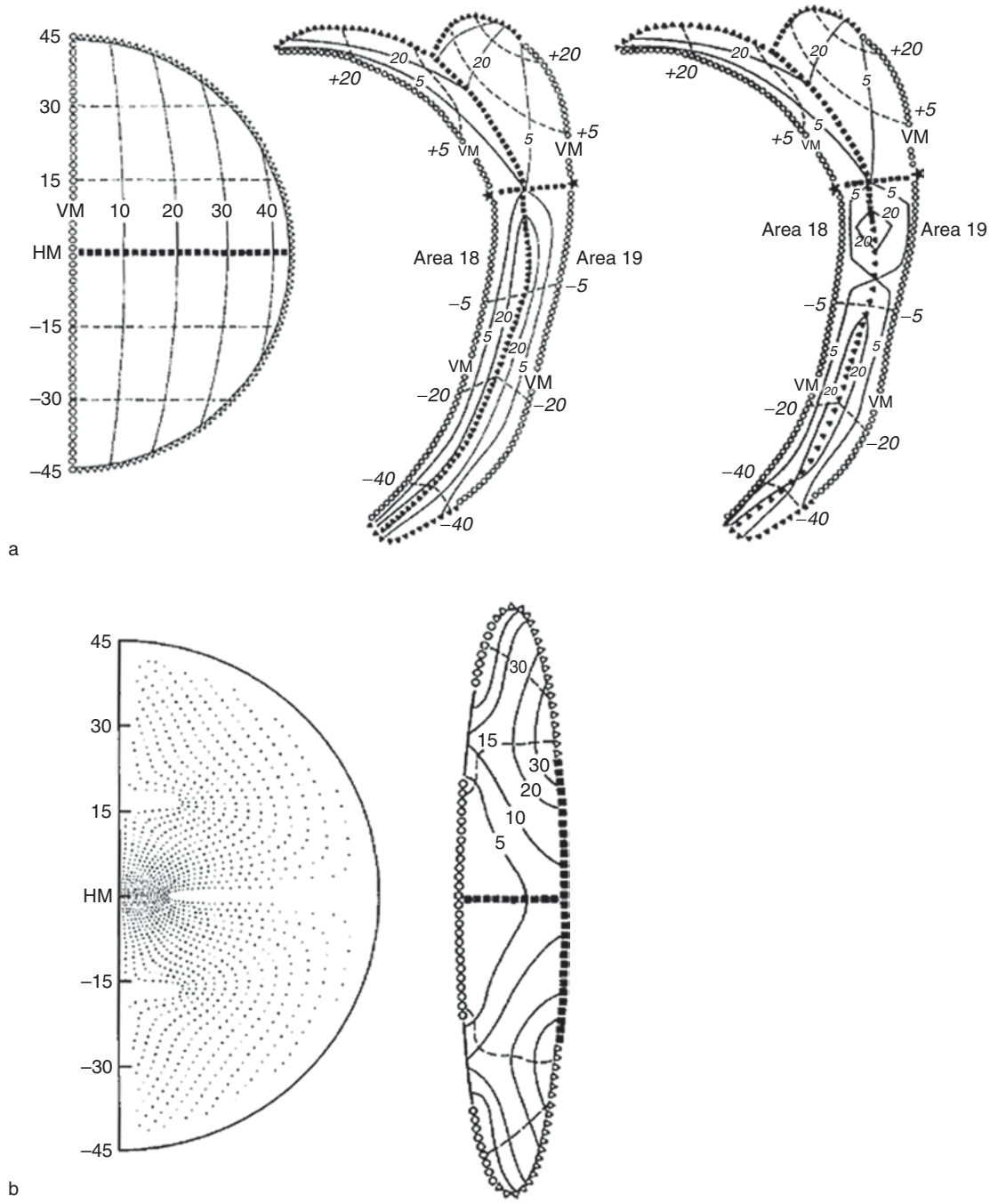

Figure 4 Islands in retinotopic representations. (a) Retinotopic organization of cat area 18. The figures in the middle and on the right show how the visual field on the left is mapped onto area 18 for two different individuals. Both show a discontinuity in the representation of the horizontal meridian at the outer boundary of area 18 and the one on the right shows an island representation of a part of the periphery in the lower visual field. (b) Result of self-organization simulation showing two islands and one field discontinuity. Reproduced from Wolf F, Bauer H-U, and Geisel T (1994) Formation of field discontinuities and islands in visual cortical maps. *Biological Cybernetics* 70: 525–531. Copyright 1994 Springer-Verlag Berlin-Heidelberg. With kind permission of Springer Science and Business Media.

beyond the critical period. Lesions that remove a section of visual input lead to changes in the network structure; loss of limbs results in a reorganization of somatosensory cortex, where the area representing the lost limb is taken over by nearby areas. Replacing lost or damaged afferent fibers from the sensory periphery to primary auditory cortex with electrical stimulation can result in a reorganization of mature brains that allows them to make use of this representation of the sensory environment.

A possible anatomical correlate of plasticity changes throughout development is the development of lateral

connections. It coincides with the development of orientation and ocular dominance columns in the early visual areas. Moreover, lateral connections are initially uniformly distributed, but long-range connections become patchy during development as a result of neural activity. These longer range lateral connections preferentially connect neurons encoding similar properties (e.g., orientation preference) and are thought to be important for functionality such as perceptual grouping. The short-range lateral connections, however, remain approximately uniformly distributed, irrespective of a neuron's position in the orientation map. This is similar to a decrease in the neighborhood parameter, which improves convergence of SOMs and assists the formation of hierarchical representations of multiple dimensions in one map.

Semantic Maps

Whereas sensory feature maps can be studied *in vivo* with relative ease, cortical representations of higher faculties such as language are not as readily accessible. However, the pervasiveness of the columnar design throughout neocortex suggests that the SOM principle may also be relevant for other faculties.

The SOM has been applied to phonemes of different languages. Providing a representation of the phoneme's waveform as input patterns, a map emerges that represents the phonemes grouped according to their similarity in sound.

Modeling studies of more abstract representations such as semantic meaning are only suggestive in character. However, they do demonstrate that topologically correct map representations can also be found for these abstract features. For instance, words can be organized into a map representing their semantic relatedness from semantic feature vectors describing their meaning or even providing only the context in which they occur as input (Figure 5). Further examples can also be found in the many nonneuroscience applications that employ SOMs for organizing complex data in an intuitive way, as exemplified by the WEBSOM application for text mining.

An argument supporting such topological organization of these more abstract concepts *in vivo* derives from the degradation pattern observed in human patients after strokes or brain lesions, where failures in the use of specific concepts or categories have been observed.

Functional Considerations

Besides the indications that SOM principles may aid the self-organization of the brain, one can also consider functional advantages they offer over other approaches. However, this area is still an open topic.

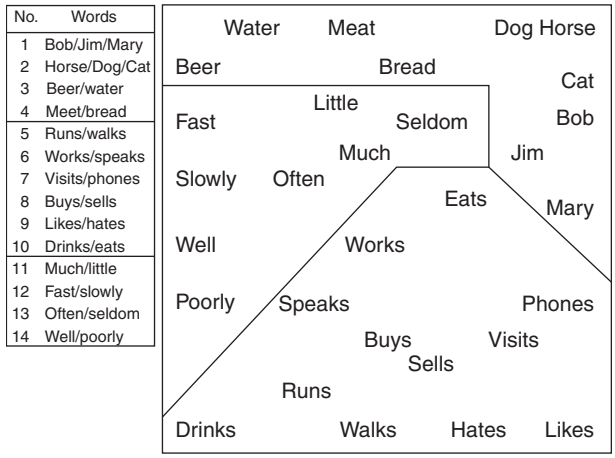

Figure 5 Semantic map. (Left) The word list is used to generate simple three-word sentences (noun–verb–adverb, such as 1–5–12, 1–8–2, or 2–10–3). Sentences which make sense and have the correct structure are produced at random and concatenated to generate a text. Words in the text are encoded as random high-dimensional vectors of unit length and are presented, together with their predecessors and successors, as sensory input for a SOM. (Right) Representation of the words on the network layer. Words are grouped according to their grammatical categories as well as their meaning. Adapted from Obermayer K (1995) Mathematical theories of neural development. *B.I.F. FUTURA* 10(2): 72–87.

It has been argued that coverage and continuity are desirable features for map representations since they ensure a gap-free representation of the feature space. Indeed, feature maps found in the brain appear to at least locally maximize coverage and continuity.

Furthermore, studies have suggested that SOM-like models minimize wiring length for local calculations. In general, one can argue that SOMs provide a trade-off: on the one hand, they reduce the communications overhead through clustering and near-optimal wiring; on the other hand, they also provide a certain load balancing through adjustment of the representation to the input distribution (but, unlike equiprobabilistic topographic maps, also to input variance). However, it is difficult to define the optimum for this trade-off.

Finally, the cost function (eqn [4]) also provides a clue to the advantages of the SOM algorithm. The squared error term ensures the map representation is effective for signals affected by Gaussian noise.

Discussion

The overwhelming body of evidence for activity-driven plastic reorganization of cortex leaves no doubt that self-organization of the network is important for its overall structure. In this context, the SOM has been a very useful model of the result of this self-organization and, to some extent, of the process

generating this result. The extent to which map structure is determined through self-organization, however, is still debated. In particular, the formation of topographic mappings between different areas remains contended. Kohonen argued that such maps tend to signal co-occurring events (e.g., the orientation occurring in neighboring image positions) and provide no overall 'winner,' as required by the SOM algorithm. The soft competition approach can defuse these arguments somewhat: the restricted range of the neighborhood function effectively limits the number of winner foci locally. Point-to-point maps can thus form from restricted activity (like retinal waves) early during development, when the neighborhood function may still be large. The lateral interaction becomes more localized later in order to restrict the range of competition. However, evidence suggests that this discussion may be futile since the point-to-point-like maps may be determined chemically to a larger extent than previously thought, thus requiring only the local ordering power provided by the traditional SOM.

A trend in the study of map formation is a greater emphasis on the importance of lateral connectivity, which at greater distances is patchier. Miikkulainen and co-workers argued that the applicability of the general SOM principles can be broadened if the lateral (longer range) connectivity is also modified according to Hebbian rules. They extensively explored how this class of models (LISSOM) can account in detail for a range of phenomena observed in visual cortex.

However, to what extent should the connectivity between different areas be predetermined using genetic tools? This is an example of the bias–variance dilemma: predetermination of map structure effectively introduces a bias into the representation a subsequent SOM process will build. It has the advantage that the feature representation can be learned from fewer examples and is thus more quickly available. However, a bias that is too strong will result in inferior feature representations if the bias does not reflect stimulus statistics. Thus, in order to empower feature representations to be efficient in a wide variety of environments, the bias needs to be kept sufficiently low and the training period sufficiently long. Based on this, one could argue that the unusual and, in comparison to other animals, seemingly excessive length of the maturation period in humans reflects our extraordinary capacity for adjusting to different (sensory and otherwise) environments.

It must be cautioned that although the simple nature of the SOM combined with its proven explanatory power for relatively simple feature map models is intriguing, it should not be regarded as the universal key to understanding the brain. First, for some features there may not actually be a need to achieve selectivity through some organizational principle since even haphazard wiring with some basic order in connectivity may suffice to explain the observed distribution of selectivities. Second, a simplistic view such as considering the brain as something akin to a stack of hierarchically organized SOMs which represent increasingly complex features disregards much of the highly complex pattern of interareal connectivity. Finally, examining the detailed mechanisms of how self-organization in the brain occurs demonstrates that mechanisms such as spike time-dependent synaptic plasticity add further complexity to the picture. Thus, SOMs in their simplicity will not be able to cover all behavior of the brain.

In summary, the SOM provides major insights into the organizational principles used in the brain. Particularly noteworthy is the fact that such a simple principle can quantitatively account for experimental results in so much detail. It remains an active area of research that has yet to achieve all its potential.

See also: Hebbian Plasticity; Notch Pathway: Lateral Inhibition; Topographic Maps: Molecular Mechanisms; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging; Visual Cortical Models of Orientation Tuning.

Further Reading

Erwin E, Obermayer K, and Schulten KJ (1995) Models of orientation and ocular dominance columns in the visual cortex: A critical comparison. *Neural Computation 7:* 425–468.

Farley BJ, Yu H, Jin DZ, and Sur M (2007) Alteration of visual input results in a coordinated reorganization of multiple visual cortex maps. *Journal of Neuroscience* 27: 10299–10310.

Kaas JH (1997) Topographic maps are fundamental to sensory processing. *Brain Research Bulletin* 44: 107–112.

Kaski S, Kangas J, and Kohonen T (1998) Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys* 1: 102–350.

Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43: 59–69.

Kohonen T (1982) Analysis of a simple self-organizing process. *Biological Cybernetics* 44: 135–140.

Kohonen T (2001) *Self-Organizing Maps*, 3rd edn. New York: Springer.

Miikkulainen R, Bednar JA, Choe Y, and Sirosh J (2005) *Computational Maps in the Visual Cortex*. New York: Springer.

Miller KD (1995) Receptive fields and maps in the visual cortex: Models of ocular dominance and orientation columns. In: Domany E, van Hemmen JL, and Schulten K (eds.) *Models of Neural Networks III*, pp. 55–78. New York: Springer.

Oja M, Kaski S, and Kohonen T (2003) Bibliography of self-organizing map (SOM) papers: 1998–2001 addendum. *Neural Computing Surveys* 3: 1–156.

Obermayer K (1995) Mathematical theories of neural development. *B.I.F. FUTURA* 10(2): 72–87.

Obermayer K, Blasdel GG, and Schulten K (1992) Statistical–mechanical analysis of self-organization and pattern formation during the development of visual maps. *Physical Review A* 45: 7568–7589.

Obermayer K, Ritter H, and Schulten KJ (1990) A principle for the formation of the spatial structure of cortical feature maps. *Proceedings of the National Academy of Sciences of the United States of America* 87: 8345–8349.

Obermayer K and Sejnowski TJ (eds.) (2001) *Self-Organizing Map Formation*. Cambridge, MA: MIT Press.

Swindale NV (1996) The development of topography in the visual cortex: A review of models. *Network: Computation in Neural Systems* 7: 161–247.

Tootell RBH, Switkes E, Silverman MS, and Hamilton SL (1988) Functional anatomy of macaque striate cortex: II. Retinotopic organization. *Journal of Neuroscience* 8: 1531–1568.

van Hulle MM (2000) *Faithful Representations and Topographic Maps*. New York: Wiley.

Wolf F, Bauer H-U, and Geisel T (1994) Formation of field discontinuities and islands in visual cortical maps. *Biological Cybernetics* 70: 525–531.

Semantic Memory

A Martin, National Institute of Mental Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

'Semantic memory' refers to a major division of long-term memory that includes knowledge of facts, events, ideas, and concepts. As such, it covers a vast terrain, ranging from information about historical and scientific facts, details of public events, and mathematical equations to the information that allows us to identify objects and understand the meaning of words. Semantic memory can be distinguished from episodic/autobiographical memory by an absence of temporal and spatial details about the context of learning. In relation to episodic memory, semantic memory is considered to be both a phylogenically and an ontologically older system. Although many animals, especially mammals and birds, acquire information about the world, they are assumed to lack the neural machinery to consciously recollect detailed episodes of their past. Finally, although retrieval of semantic memory often requires explicit, conscious mediation, the organization of semantic memory can also be revealed via implicit tasks such as semantic priming.

Semantic Memory and the Medial Temporal Lobe Memory System

Studies of patients with amnesia due to damage to the medial temporal lobes have established three broadly agreed on facts about the functional neuroanatomy of semantic memory. First, like episodic memory, acquisition of semantic memories is dependent on medial temporal lobe structures, including the hippocampal region (CA fields, dentate gyrus, and subiculum) and surrounding neocortex (parahippocampal, entorhinal, and perirhinal cortices). Damage to these structures results in deficient acquisition of new facts and public events and the extent of this deficit is roughly equivalent to the deficit for acquiring personal information about day-to-day occurrences. However, despite broad agreement that acquiring semantic memories requires medial temporal lobe structures, there is disagreement concerning the role of the hippocampal region. One position, championed by Larry Squire and colleagues, holds that the hippocampus is necessary for acquiring semantic information. In contrast, others have argued that acquisition of semantic memories can be accomplished by the surrounding neocortical structures alone; participation of the hippocampus is not necessary. Recent studies have favored the hippocampal position by showing that carefully selected patients with damage limited to the hippocampus are impaired in learning semantic information, and that the impairment is equivalent to their episodic memory deficit. One potentially important caveat to this claim comes from studies of individuals who have sustained damage to the hippocampus at birth or during early childhood. These cases of developmental amnesia have disproportionately better semantic than episodic memories, suggesting that the hippocampal region may not be necessary for acquiring semantic information. Reconciliation of this issue will depend on direct comparison of adult onset and developmental amnesias with regard to extent of medial temporal lobe damage and its behavioral consequences.

The second major finding established by studies of amnesic patients is that the medial temporal lobe structures have a time-limited role in the retrieval of semantic memories. Retrieval of information about public events shows a temporally graded pattern with increasing accuracy for events further in time from the onset of the amnesia. Conceptual information about the meaning of objects and words acquired many years prior to the amnesia onset remains intact as assessed by both explicit and implicit tasks. Third, semantic memories of all types are stored in the cerebral cortex. In amnesic patients, impairments in semantic memory for information acquired prior to amnesia onset are directly related to the extent of damage to cortex outside the medial temporal lobes.

Cortical Lesions and the Breakdown of Semantic Memory

Studies of semantic memory in amnesia have concentrated largely on measures of public event knowledge. The reason for this is that these tasks allow memory performance to be assessed for events known to have occurred either prior to, or after, amnesia onset. These measures also allow performance to be evaluated for events that occurred at different times prior to amnesia onset to determine if the memory impairment shows a temporal gradient – a critical issue for evaluating theories of memory consolidation. Because these patients have either no or, more commonly, limited damage to regions outside the medial temporal lobes, they are not informative about how semantic information is organized in the cerebral cortex. To address this issue, investigators have turned to patients with relatively focal lesions compromising

different cortical areas. In contrast to the studies of amnesic patients, these studies have focused predominantly on measures designed to probe knowledge of object concepts.

Object Concepts

An object concept refers to the representation (i.e., information stored in memory) of an object category (a class of objects in the external world). Concepts are central to all aspects of cognition; they are the glue that holds cognition together. The primary function of a concept is to allow us to quickly draw inferences about an object's properties. That is, identifying an object as, for example, a 'hammer' means that we know that this is an object used to pound nails, so that we do not have to rediscover this property each time the object is encountered. Object concepts are hierarchically organized, with the broadest knowledge represented at the superordinate level, more specific knowledge at an intermediary level referred to as the 'basic level,' and the most specific information at the subordinate level. For example, 'dog' is a basic-level category that belongs to the superordinate categories 'animals' and 'living things,' and has subordinate categories such as 'poodle' and 'collie.' As established by Eleanor Rosch and others in the 1970s and 1980s, the basic level has a privileged status. It is the level used nearly exclusively to name objects (e.g., 'dog' rather than 'poodle'). It is also the level at which we are fastest to verify category membership (i.e., we are faster to verify that a picture is a 'dog' rather than an 'animal' or a 'poodle'). It is also the level at which subordinate category members share the most properties (e.g., collies and poodles have similar shapes and patterns of movement). Finally, the basic level is the easiest level at which to form a mental image (you can easily imagine an elephant but not an 'animal'). Studies of patients with cortical damage have documented the neurobiological reality of this hierarchical scheme and the central role of the basic level for representing objects in the human brain.

Semantic Dementia and the General Disorders of Semantic Memory

Several neurological conditions can result in a relatively global or general disorder of conceptual knowledge. These disorders are considered general in the sense that they cut across multiple category boundaries; they are not category specific. Many of these patients suffer from a progressive neurological disorder of unknown etiology referred to as semantic dementia (SD). General disorders of semantic memory are also prominent in patients with Alzheimer's disease (who, compared to SD patients, typically have a greater episodic memory impairment), and can also occur following left hemisphere stroke, prominently involving the left temporal lobe. The defining characteristics of this disorder, initially described by Elizabeth Warrington and colleagues in the mid-1970s, are deficits on measures designed to probe knowledge of objects and their associated properties. These deficits include impaired object naming (with errors typically consisting of semantic errors – retrieving the name of another basic-level object from the same category, or retrieval of a superordinate category name), impaired generation of the names of objects within a superordinate category, and an inability to retrieve information about object properties – including sensory-based information (shape, color) and functional information (motor-based properties related to the object's customary use, but this may include other kinds of information not directly related to sensory or motor properties). The impairment is not limited to stimuli presented in a single modality, like vision, but rather extends to all tasks probing object knowledge regardless of stimulus presentation modality (visual, auditory, tactile) or format (words, pictures). The semantic deficit is hierarchical in the sense that broad levels of knowledge are often preserved, while specific information is impaired. Thus, these patients can sort objects into superordinate categories, having, for example, no difficulty indicating which are animals, which are tools, which are foods, and the like. The difficulty is manifest as a problem distinguishing among the basic-level objects as revealed by impaired performance on measures of naming and object property knowledge.

Recent studies have expanded our understanding of SD in two important ways: one is related to location of neuropathology, the other to functional characteristics of the disorder. The initial neuropathological and imaging studies of SD indicated prominent atrophy of the temporal lobes, especially to the anterolateral sector of the left temporal lobe, including the temporal polar cortex inferior and middle temporal gyri, and the most anterior extent of the fusiform gyrus. However, recent advances in neuroimaging that allow for direct and detailed comparison of brain morphology in SD patients relative to healthy controls have shown that the atrophy extends more posteriorally along the temporal lobe than previously appreciated. In fact, the amount of atrophy in ventral occipitotemporal cortex, including the posterior portion of the fusiform gyrus, has been reported to be as strongly related to the semantic impairment in SD as is atrophy in the most anterior regions of the temporal lobes.

The other major advance in our understanding of SD is that it is not as global a conceptual disorder as

initially thought. Rather, certain domains of knowledge may be preserved, and the pattern of impaired and preserved knowledge appears to be related to the locus of pathology. Specifically, left-sided atrophy seems to impair information about all object categories except person-specific knowledge (i.e., information about famous people), which, in turn, is associated with involvement of the right anterior temporal lobes. Also relatively spared is knowledge of number and mathematical concepts, and information about motor actions needed to use familiar tools. Both of these knowledge domains are associated with left posterior parietal cortex.

Category-Specific Disorders of Semantic Memory

Although reports of cases of relatively circumscribed knowledge disorders date back over 100 years, the modern era of the study of category-specific disorders began in the early 1980s with the seminal reports of Warrington and colleagues. Category-specific disorders have the same functional characteristics as does SD, except that the impairment is largely limited to members of a single superordinate object category. For example, a patient with a category-specific disorder for 'animals' will have greater difficulty naming and retrieving information about members of this superordinate category relative to members of other superordinate categories (e.g., tools, furniture, flowers). Similar to patients with SD, patients with category-specific disorders have difficulty distinguishing among basic-level objects (e.g., between dog, cat, horse), thereby suggesting a loss or degradation of information that uniquely distinguishes among members of the superordinate category (e.g., four-legged animals).

Although a variety of different types of category-specific disorders have been reported (e.g., for fruits and vegetables), most common have been reports of patients with relatively greater knowledge deficits for animate entities (especially animals) than for a variety of inanimate object categories. While less common, other patients show the opposite dissociation of a greater impairment for inanimate manmade objects (including common tools) than for animals and other living things.

Models of category-specific disorders Two major theoretical positions have been advanced to explain these disorders. Following the explanation posited by Warrington for her initial cases, most current investigators assume that category-specific deficits are a direct consequence of an object property-based organization of conceptual knowledge, an idea that was prominent in the writings of Karl Wernicke, Sigmund Freud, and other behavioral neurologists during the late nineteenth and early twentieth centuries. The central idea is that object knowledge is organized in the brain by sensory (e.g., form, motion, color, smell, taste) and motor properties associated with the object's use, and in some models, by other functional/verbally mediated properties, such as where an object is typically found. In this view, category-specific semantic disorders occur when a lesion disrupts information about a particular property or set of properties critical for defining, and for distinguishing among, category members. Thus damage to regions that store information about object form, and form-related properties like color and texture, will produce a disorder for animals. This is because visual appearance is assumed to be a critical property for defining animals, and because the distinction between different animals is assumed to be heavily dependent on knowing about subtle differences in their visual forms. A critical prediction of sensory/motor-based models is that the lesion should affect knowledge of all object categories with this characteristic, not only animals. In a similar fashion, damage to regions that store information about how an object is used should produce a category-specific disorder for tools, and all other categories of objects defined by how they are manipulated.

The alternative to these property-based theories is the domain-specific view championed most recently by Alfonso Caramazza and colleagues. On this account, our evolutionary history provides the major constraint on the organization of conceptual knowledge in the brain. Specifically, the theory proposes that selection pressures have resulted in dedicated neural machinery for solving, quickly and efficiently, computationally complex survival problems. Likely candidate domains offered are animals, conspecifics, plant life, and possibly tools. Property-based and category-based accounts are not mutually exclusive. For example, it is certainly possible that concepts are organized by domains of knowledge, implemented in the brain by large-scale property-based systems. Much of the functional neuroimaging evidence (to be discussed later) is consistent with this view.

Functional neuroanatomy of category-specific disorders There is considerable variability in the location of lesions associated with category-specific disorders for animate and inanimate entities. Nevertheless, some general tendencies can be observed. In particular, category-specific knowledge disorders for animals are disproportionately associated with damage to the temporal lobes. The most common etiology is herpes

simplex encephalitis, a viral condition with a predilection for attacking anteromedial and inferior (ventral) temporal cortices. Category-specific knowledge disorders for animals also have been reported following focal, ischemic lesions to the more posterior regions of ventral temporal cortex, including the fusiform gyrus. In contrast, category-specific knowledge disorders for tools and other manmade objects have been most commonly associated with focal damage to lateral frontal and parietal cortices of the left hemisphere. However, it is important to stress that the lesions in patients presenting with category-specific knowledge disorders are often large and show considerable variability in their location from one patient to another. As a result, these cases have been relatively uninformative for questions concerning the organization of object memories in cerebral cortex. In contrast, recent functional neuroimaging studies of the intact human brain have begun to shed some light on this thorny issue.

Organization of Conceptual Knowledge: Neuroimaging Evidence

The functional neuroanatomy of semantic memory has been explored using a wide variety of experimental approaches and paradigms. The overarching goal of these studies has been to identify brain regions that show a heightened response during a conceptual or semantic processing task (e.g., is the object a living thing?) versus performing a nonsemantic, but equally difficult, task (e.g., does the object's name contain the letter 'p'?) with the same set of stimuli. Assigning a particular brain region a role in semantic processing is further strengthened when the activity is associated with semantic task performance, regardless of physical differences in the stimuli used to denote the objects (i.e., visual or auditory, pictures or words). As a result, activity in these brain regions can be linked to conceptual processes and not features of stimulus input modality or format.

Studies based on this type of approach have consistently isolated two regions, the left ventrolateral prefrontal cortex (VLPFC) and the ventral and lateral regions of the temporal lobes (typically stronger in the left than in the right hemisphere). Detailed evaluations of left VLPFC have shown that this region is critically involved in the top-down control of semantic memory. Specifically, left VLPFC is responsible for guiding retrieval and postretrieval selection of conceptual information stored in posterior temporal and perhaps other cortical regions. This neuroimaging finding is consistent with studies of patients with left inferior frontal lesions that have word-retrieval difficulties, but retain concept knowledge of the words they have difficulty retrieving. In contrast, a large body of evidence has linked activity in posterior regions of the temporal lobes to memory representations, especially concerning conceptual representations of specific properties associated with concrete objects.

Object Knowledge Is Organized by Sensory and Motor-Based Properties

One source of evidence concerning the organization of object property information comes from property production tasks. In these tasks, individuals are presented with an object, denoted either by a picture or the object's name, and are required to generate a word denoting a specific property associated with that object. Several studies have shown that generating an action associate (e.g., saying 'pull' in response to a child's wagon) elicits heightened activity in premotor cortex as well as in a posterior region of the left lateral temporal cortex, centered on the middle temporal gyrus (pMTG) just anterior to primary visual motion-sensitive cortex. In contrast, relative to action word generation, color word generation (e.g., saying 'red' for the child's wagon) activates the fusiform gyrus, anterior to regions associated with passive perception of color. Subsequent studies have shown that tasks requiring finer perceptual discrimination of colors, or perceiving more complex patterns of object motion, elicit activity in the same temporal lobe regions active when color or action knowledge is retrieved. The findings from these and similar studies suggest two important points. First, that information about different types of object-associated properties (e.g., color, action) is represented in different anatomical regions. Second, these regions directly overlap with sites that mediate perception of these object properties. A number of studies have replicated these two basic points using property verification tasks requiring subjects to answer yes/no questions about the relation between a particular object and property. These studies have linked retrieval of information about visual, auditory, and tactile properties to regions associated with processing within each of these sensory modalities. Similarly, retrieval of object use-associated information has been linked to motor regions active when objects are actually manipulated. These studies suggest that the same neural systems are involved, at least in part, in perceiving, acting on, and knowing about specific object properties.

Neural Networks for Animate Entities and Tools

A large number of studies have addressed questions concerning how information about different object

categories is represented in the brain. Many of these studies were motivated by the neuropsychological evidence for category-specific disorders discussed in the preceding sections. As a result, these studies have concentrated on the neural systems for perceiving and knowing about two broad domains: animate agents (living things that move on their own), and tools (manmade objects with a close association between their function and motor movements associated with their use). These studies have provided evidence for four major points about the organization of conceptual knowledge. First, information about a specific object category is not represented in a single cortical region, but rather is represented by a network of discrete regions that may be widely distributed throughout the brain. Second, the informational contents of these regions are related to specific properties associated with the object. Third, some of these property-based regions are automatically active when objects are identified, thus suggesting that object perception is associated with the automatic retrieval of a limited set of associated properties that may be necessary and sufficient to identify that object. These properties prominently include information about what the object looks like, how it moves, and, for a tool, its use. Fourth, this object property-based information is stored in sensory and motor systems active when that information was acquired.

Conceptual representations of animate entities have been most strongly linked to two regions of temporal lobe cortex, the lateral region of the fusiform gyrus, located on the ventral surface of the temporal lobe (including the fusiform face area; FFA), and the posterior region of the superior temporal sulcus (pSTS), located on the lateral surface of the temporal lobe (Figure 1). Activation of these regions is typically stronger in the right, rather than in the left, hemisphere. In contrast, conceptual representations of tools have been linked to four regions. Included in this network are the medial portion of the fusiform gyrus, pMTG (as discussed previously in relation to generating action words), and posterior parietal and ventral premotor cortices. Activity in these regions is typically stronger in the left, rather than in the right, hemisphere (Figure 1).

Activity in these networks transcends stimulus features, thus strengthening the claim that these regions are involved in conceptual processing. For example, category-related activity in the lateral region of the fusiform gyrus has been shown not only in response to pictures of people and animals, but also to the written and heard names of animals, human voices, animal-associated sounds (like the moo of a cow), and when simply imagining faces. In addition, category-related activity in the lateral fusiform gyrus has been found in response to degraded and abstract visual

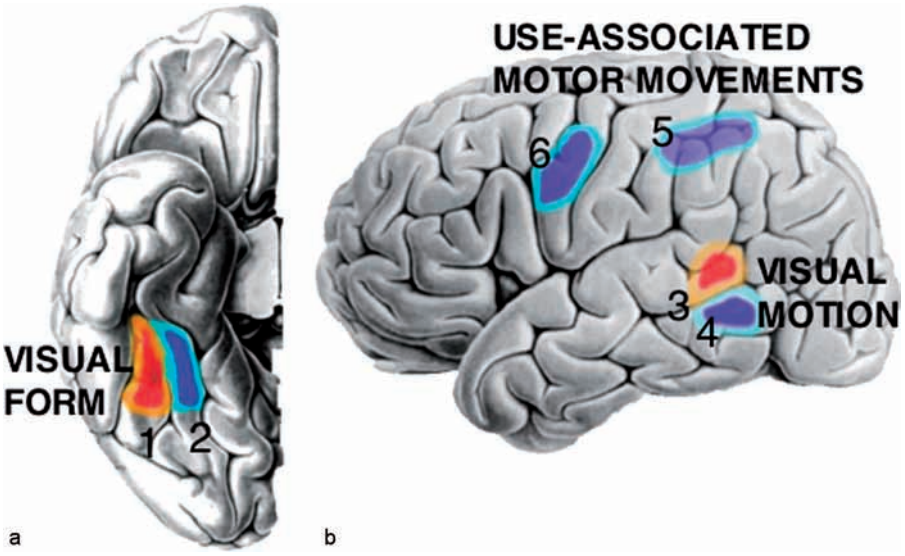

Figure 1 Schematic illustration of location of regions showing category-related activity for animate entities (red) and tools (blue). (a) Ventral view of the right hemisphere, showing relative location of regions assumed to represent visual form and form-related properties such as color and texture of animate entities (1, lateral region of the fusiform gyrus, including, but not limited to the fusiform face area) and tools (2, medial region of the fusiform gyrus). (b) Lateral view of the left hemisphere, showing relative location of regions assumed to represent biological motion (3, posterior region of the superior temporal sulcus) and rigid motion vectors typical of tools (4, posterior region of the middle temporal gyrus). Also shown are the relative locations of the posterior parietal (5, typically centered on the intraparietal sulcus) and ventral premotor (6) regions of the left hemisphere assumed to represent information about the motor movements associated with using tools.

stimuli such as humanlike stick figures, point light displays interpreted as human figures in motion, degraded pictures of objects when misinterpreted as faces, and when simple geometric shapes in motion are interpreted as depicting social interactions. Similarly, category-related activity in the more medial region of the fusiform gyrus has been shown not only in response to pictures of tools, but also to the written and heard names of tools and to tool-associated sounds (like the banging of a hammer). In addition, category-related activity in the medial fusiform gyrus has been observed in response to degraded and abstract visual stimuli such as point light displays interpreted as depicting tools in motion, and when simple geometric shapes in motion are interpreted as depicting mechanical interactions. Taken together, these findings indicate that the critical determinant of differential activity in these regions is the conceptual interpretation assigned to a stimulus (e.g., as a person, animal, or tool), not the physical characteristics of the stimulus, *per se*.

Linking Category-Related Representations to Sensory and Motor Properties

Neuroimaging studies also have provided evidence about the functional role of regions showing category-related activity. The two regions in the posterior region of the temporal lobe that show a stronger response to animate rather than inanimate objects – the lateral region of the fusiform gyrus and pSTS – have been linked to representing the visual form and the flexible, fully articulated, patterns of motion associated with animate objects, respectively. These findings are consistent with a large body of neurophysiological data from studies of monkeys showing that neurons in inferior temporal cortex and STS are differentially tuned to visual form and biological motion, respectively. In contrast, the two regions in the posterior region of the human temporal lobe that show a stronger response to tools than to animate objects – the medial region of the fusiform gyrus and pMTG – have been linked to representing the visual form and rigid, unarticulated patterns of motion typically associated with manipulable man-made objects.

Other studies have provided evidence linking the representation of tools to two regions of the dorsal action-processing stream – the left posterior parietal cortex and left ventral premotor cortex. These findings are consistent with data from monkey neurophysiology showing that neurons in ventral premotor and parietal cortices respond both when monkeys grasp objects and when they see objects that they have had experience manipulating. Along with pMTG, functional brain imaging studies in humans have shown that these dorsal regions are active when individuals retrieve information about the functional properties of objects, suggesting that information about how objects are used is stored in these regions. Support for this claim comes from recent neuropsychological investigations linking impaired knowledge of tools and their associated actions to damage to either left pMTG, posterior parietal, or premotor cortices.

Finally, recent studies have shown that the category-related, property-based neural systems discussed here are not only active when objects are encoded into memory, but also that this activity is reinstated prior to recalling that information at a later time. These findings underscore the central role that these systems play in memory encoding, storage, and retrieval.

See also: Aging and Memory in Humans; Cognition: An Overview of Neuroimaging Techniques; Declarative Memory System: Anatomy; Dementia; Hippocampus and Neural Representations; Memory Representation; Memory Disorders.

Further Reading

Caramazza A and Shelton JR (1998) Domain-specific knowledge systems in the brain: The animate–inanimate distinction. *Journal of Cognitive Neuroscience* 10: 1–34.

Damasio H, Tranel D, Grabowski T, et al. (2004) Neural systems behind word and concept retrieval. *Cognition* 92: 179–229.

Forde EME and Humphreys GW (2002) *Category Specificity in the Brain and Mind*. Hove and New York: Psychology Press.

Garrard P and Hodges JR (1999) Semantic dementia: Implications for the neural basis of language and meaning. *Aphasiology* 13: 609–623.

Manns JR, Hopkins RO, and Squire LR (2003) Semantic memory and the human hippocampus. *Neuron* 38: 127–133.

Martin A (2007) The representation of object concepts in the brain. *Annual Review of Psychology* 58: 25–45.

Martin A and Caramazza A (eds.) (2003) *The Organization of Conceptual Knowledge in the Brain: Neuropsychological and Neuroimaging Perspectives*. Hove and New York: Psychology Press.

Murphy GL (2002) *The Big Book of Concepts*. Cambridge, MA: The MIT Press.

Polyn SM, Natu VS, Cohen JD, et al. (2005) Category-specific cortical activity precedes retrieval during memory search. *Science* 310: 1963–1966.

Vargha-Khadem F, Gadian DG, Watkins KE, et al. (1997) Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* 277: 376–380.

Warrington EK (1975) The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology* 27: 635–657.

Warrington EK and Shallice T (1984) Category specific semantic impairments. *Brain* 107: 829–854.

Semaphorins

J Verhaagen, Netherlands Institute for Neuroscience, Amsterdam, The Netherlands
R J Pasterkamp, University Medical Center Utrecht, Utrecht, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

In the nineteenth century, a 'semaphore line' was a signaling system composed of rods or flags that was widely used in the maritime world to navigate ships, in rail transportation, and to direct movements of army battalions. In the early 1990s, a family of proteins was discovered that functions as signals involved in growth cone guidance and navigation. The term 'semaphore' was adopted to refer to this family of proteins as semaphorins. The first semaphorin, fasciclin IV, was identified in the grasshopper and was later renamed semaphorin-1a. The first vertebrate semaphorin was originally designated collapsin-1, a name referring to its collapsing effect on growth cones of chick sensory ganglia. Collapsin-1 was later renamed semaphorin 3A. In 1999, the Semaphorin Nomenclature Committee published a unified nomenclature for the semaphorins/collapsins based on structural features and phylogenetic tree analysis. To date, more then 25 semaphorin genes have been identified that encode both membrane-associated and secreted molecules subdivided into eight classes. Most semaphorins contain approximately 750 amino acids and share a conserved 'sema' domain of approximately 500-amino-acid residues (Figure 1). This article reviews the current knowledge on how semaphorins function in neuronal network formation during development. In addition, it discusses progress in understanding the role of semaphorins during adulthood in synaptic transmission and plasticity, in neural trauma, and in certain neurological and psychiatric diseases. The article does not touch upon the role these proteins play in organogenesis, vascularization, the immune system, and cancer.

The Concept of Chemorepulsion

Approximately 20 years ago, it was recognized that developing axons bend away from or can be inhibited by (at that time) hypothetical molecular cues they encounter on their path of growth. The first molecules shown to be able to mediate such effects were subsequently identified in genetic or antibody perturbation screens or by combining biochemical purification with well-defined bioassays as read-out systems for chemorepulsive activity (Figure 2). To date, five families of cellular guidance molecules that can function as chemorepulsive proteins are known, including (in order of their discovery) netrins, ephrins, semaphorins, slits, and repulsive guidance molecule (RGM). Interestingly, some members of these protein families also influence axon guidance decisions by functioning as neurite attractants. A large body of work has shown that the ability of growth cones to respond to chemorepulsive proteins is critical for the formation of stereotypical neural pathways and maps. Chemorepulsion is therefore an essential element of the total arsenal of coordinated morphogenetic processes that shape the brain's circuitry.

In the mid-1980s, observations on the differential substrate properties of peripheral and central nervous system (CNS) tissue led to the hypothesis that central nervous system myelin contains inhibitory factors for regenerating nerve fibers. Myelin-associated proteins, including Nogo, myelin-associated glycoprotein (MAG), and oligodendrocyte myelin glycoprotein (OMgp), were subsequently identified as inhibitors of regenerative neurite outgrowth in the CNS. Thus, around 1985 two independent lines of thinking led to the notion that the inhibitory or repulsive regulation of neurite outgrowth during development and the inhibition of regenerative neurite outgrowth are regulated by distinct molecular entities. Recently, however, developmental chemorepulsive guidance cues, including semaphorins, ephrins, and RGMs, have also been linked to regenerative failure. Therefore, the concept of chemorepulsion appears to have implications for our understanding of the molecular mechanisms that underlie the failure of damaged nerve tracts to regenerate as well.

Semaphorins Signal through Multimeric Receptor Complexes

To detect semaphorins in the extracellular space, growth cones at the leading tip of extending axons express specialized receptor complexes. A unifying feature of the neuronal semaphorin receptor complexes identified to date is the presence of a member of the plexin family within each complex (Figure 1). The founding member of this homogeneous family of transmembrane proteins, plexinA1 (plexin), was found in a screen for antigens involved in axon guidance in the developing tadpole retinotectal system. Since the identification of plexinA1, ten more plexins have been discovered (two plexins can be found in invertebrates and nine in vertebrates) and categorized into four subclasses on the basis of sequence similarities (Figures 1(a)–1(d)). Like semaphorins, extracellular sema domains are characteristic of plexins and it

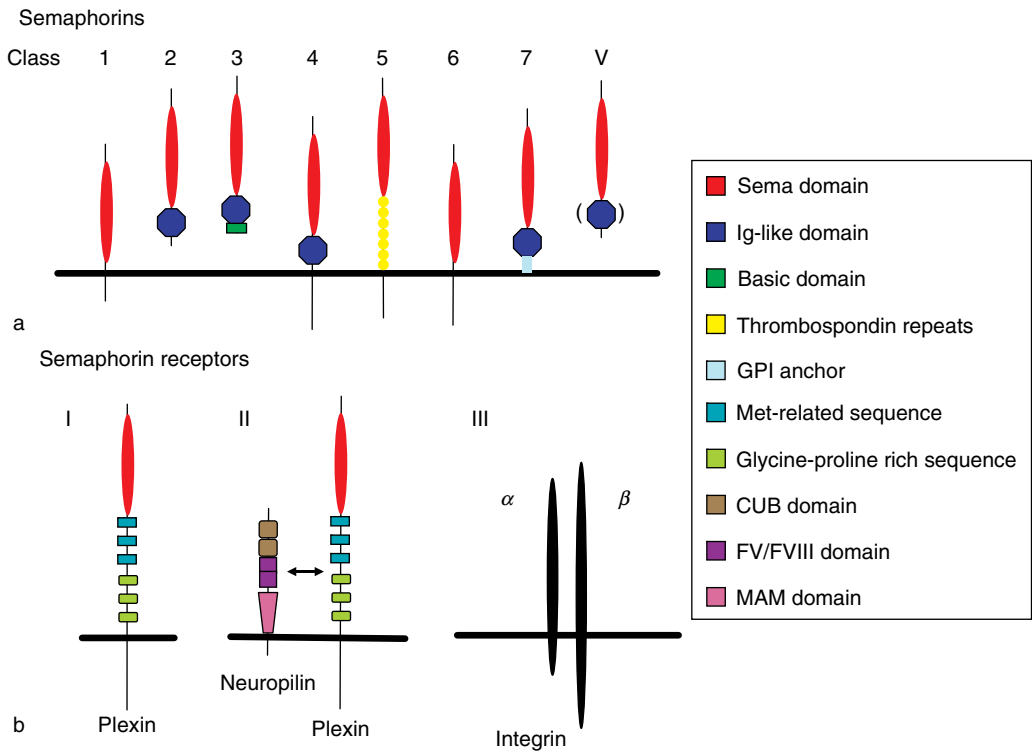

Figure 1 Semaphorins and their neuronal receptors. The semaphorin family is subdivided into eight classes. Classes 1 and 2 contain semaphorins found in invertebrate species, classes 3–7 vertebrate semaphorins, and class V viral semaphorins. Semaphorins exist as secreted (classes 2, 3, and V) and membrane-associated molecules (classes 1 and 4–7). All neuronal semaphorin receptors identified to date contain a member of the plexin family. Semaphorins can either bind directly to plexins (a; classes 1, 2, 4, V, and Sema3E) or require neuropilins as ligand binding subunits (b; class 3 semaphorins). β_1 subunit-containing integrins also serve as receptors for class 7 semaphorins in a plexin-independent fashion (c). Modulatory receptor subunits such as L1 and Off-track have also been identified.

is believed that an intramolecular interaction between the plexin sema domain and the rest of the plexin extracellular domain inhibits receptor activity in the absence of semaphorin ligand. The plexin intracellular region shares homology with the GAP domain of p120 RasGAP and is highly conserved between plexins. Dominant-negative forms of plexins lacking their cytoplasmic region fail to mediate semaphorin-mediated repulsion which shows that plexins function as signal-transducing receptor subunits.

Although most semaphorins exert their functions by directly interacting with plexins, class 3 semaphorins (Sema3s), except for Sema3E, fail to bind any of the known plexin family members. Instead, Sema3s require neuropilins as essential semaphorin-binding coreceptors to signal through class A plexins (plexinAs) (Figure 1). Neuropilins are transmembrane proteins with short intracellular domains that lack intrinsic enzymatic activity and that are indispensable for Sema3 function. The initial identification of neuropilin-1 (A5, neuropilin (Npn)) in the *Xenopus laevis* nervous system using hybridoma techniques was followed by the characterization of the only other neuropilin, neuropilin-2, in rodents. Cell culture studies suggest that all four plexinAs (plexinA1–A4) can form receptor complexes with both neuropilins. The specific responses of different classes of neurons to individual Sema3s can be explained in part by restricted and unique 'neuropilin' distribution patterns and by the preferential binding of individual Sema3s to Npn-1 and/or Npn-2. PlexinAs also contribute to the specificity of axonal responses to Sema3s. A large body of work supports the idea that *in vivo* Npn-1/plexinA4 complexes act as principal Sema3A receptors, whereas Npn-2/plexinA3 complexes mediate responses to Sema3F. The Npn/plexinA complexes that mediate the *in vivo* responses of other Sema3s remain to be identified. In addition to Sema3s, neuropilins serve as (co)receptors for structurally unrelated ligands such as VEGFs and PDGFs.

Although less well-characterized, semaphorin receptor (components) unrelated to plexins and neuropilins do exist in neurons: (1) off-track, a protein similar to receptor tyrosine kinases but with a catalytically inactive kinase domain, associates with the *Drosophila* PlexA receptor to mediate semaphorin repulsion; (2) cell adhesion molecules of the immunoglobulin superfamily, including L1 and NrCAM, play a role in

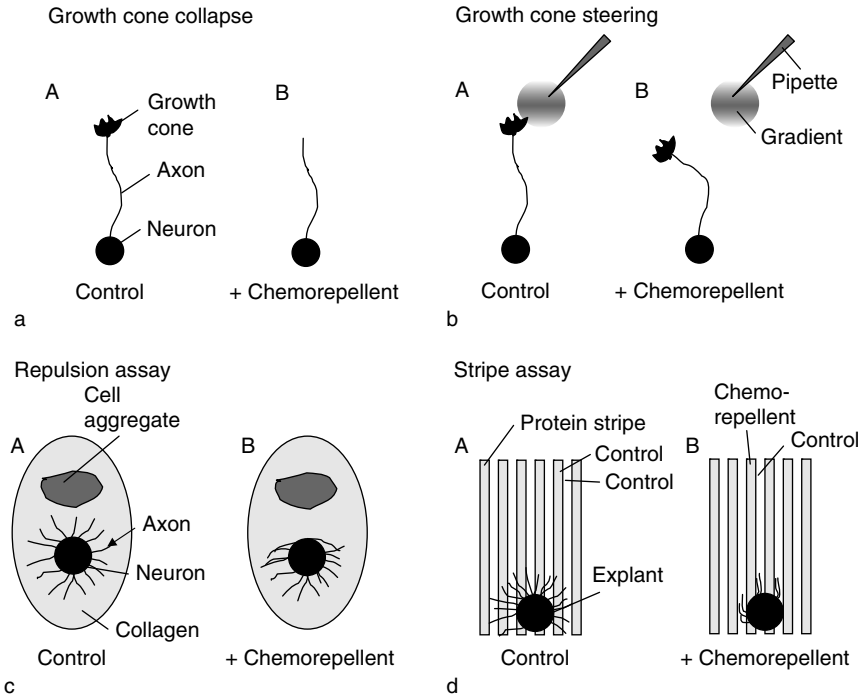

Figure 2 *In vitro* assays for studying the effects of chemorepellents on growth cone behavior and axon steering. (a) In the growth cone collapse assay, dissociated neurons are grown and a putative chemorepulsive protein is added to the tissue culture medium to assess whether it can cause growth cone collapse. (b) Similarly, in the growth cone steering assay, dissociated neurons are grown but now the chemorepulsive protein is supplied in a small gradient by a micropipette. Chemorepulsive proteins will cause the growth cone to steer away from the gradient. (c) In the repulsion assay, a neuronal explant is cultured next to an aggregate of nonneuronal cells that secretes control or chemorepulsive proteins in a collagen gel. The chemorepulsive gradient that emerges from the aggregate will cause the axons to grow away from the aggregate. (d) Whereas the first three assays (a–c) are normally used to study secreted chemorepellents, the stripe assay is used to study membrane-associated proteins. Alternating stripes of protein or cell membranes are generated and explants are grown on top. If the axons can choose between lanes of control and chemorepulsive proteins, they will grow preferentially on control lanes.

the responsiveness of axons to Sema3A and are thought to be part of the neuropilin–plexinA receptor complex; and (3) β_1 subunit-containing integrins act as Sema7A receptors, or components thereof (Figure 1). Whether nonneuronal semaphorin receptors such as CD72 or Tim2 identified in the immune and other organ systems also function in neuronal semaphorin signaling remains to be determined.

Intracellular Signaling

On activation by semaphorins, plexins induce a series of intracellular signaling events that regulate the (dis) assembly of the neuronal cytoskeleton to control growth cone morphology and motility. The signaling mechanisms that are utilized by repulsive Sema3s have been studied in most detail and are summarized here. One of the first cytosolic proteins shown to link Sema3 receptor complexes to the neuronal cytoskeleton is collapsin response mediator protein (CRMP)-2, a member of a small family of membrane-associated phosphoproteins. Five different CRMPs have been identified, which in addition to plexinAs bind a plethora of other intracellular proteins, including kinases and cytoskeleton-associated proteins. Interestingly, sequential phosphorylation of CRMPs by different kinases is essential for Sema3-induced growth cone collapse as well as for semaphorin-independent functions of CRMPs during axonogenesis. In addition to CRMPs, members of the Rho family of monomeric GTPases play a prominent role in repulsive Sema3 signaling. RhoGTPases are critical regulators of cytoskeletal dynamics. Both RhoGTPases and the proteins that regulate their activity (i.e., GEFs and GAPs) associate with the plexin cytoplasmic region. In addition, the plexin cytoplasmic domain contains a RasGAP domain which in the case of plexinAs and plexinBs acts to inactivate R-Ras. This is considered as one of several signaling events by which plexins slow down integrin signaling to reduce growth cone adhesion and allow for changes in growth cone morphology. Cytoplasmic kinases also play a prominent role in Sema3s repulsion. After the initial observation of tyrosine phosphorylation of plexins in culture, several cytoplasmic kinases have been implicated in Sema3-mediated growth cone collapse and steering (e.g., Fyn, Cdk5, and GSK-3). However, their functional roles have not been fully explored. This also

holds true for cyclic nucleotide signaling. cGMP and cAMP have strong modulatory effects on growth cone turning responses induced by semaphorins. However, our understanding of how cyclic nucleotide levels are regulated under physiological conditions by semaphorins is still rudimentary. In addition to the four classes of signaling molecules discussed previously (CRMPs, RhoGTPases, kinases, and cyclic nucleotides), an ever increasing number of intracellular proteins and signaling mechanisms are implicated in semaphorin function. For example, recent findings support roles for redox reactions, novel protein synthesis, and microdomain signaling in axon responses induced by semaphorins. In addition, a high degree of cross-talk between semaphorin and unrelated signaling cascades (e.g., those mediating integrin or neurotrophic factor functions) exists that integrates different cellular processes during neuronal network formation. Furthermore, signaling cascades downstream of semaphorins other than Sema3s are being unveiled at a rapid pace, revealing both shared and unique signaling mechanisms between different classes of semaphorins. This, together with evidence that different classes of plexins may interact and the possibility for reverse signaling (i.e., plexin acting as a ligand and semaphorins as receptors), indicates that we are just beginning to understand the complexity of the semaphorin signaling network.

Semaphorin Function during Neurodevelopment

Loss-of-function studies in mice and invertebrate species such as *Drosophila* and *Caenorhabditis elegans* have provided compelling insights into the role of semaphorins during development. Several general themes have emerged. First, deletion of semaphorin or semaphorin receptor genes often results in severe defasciculation of peripheral nerves. Second, the effects on CNS development are usually more subtle. For example, errors in CNS axon guidance as a result of semaphorin or semaphorin receptor deletion in mice are most pronounced in the olfactory system, hippocampus, anterior commissure, and thalamocortical projections. Third, many of the null mutant mouse lines display severe abnormalities in the development of nonneural tissues, particularly the cardiovascular system. This may be an important reason for the perinatal death of many of these mutants. Fourth, the phenotypes of a ligand mutant and its corresponding receptor mutant are usually very similar. For instance, the phenotype of Sema3A and Npn-1 knockouts and of Sema3F and Npn-2 are remarkably similar. This indicates that these two secreted semaphorins indeed signal through Npn-1 and Npn-2, respectively, *in vivo*. Knockouts for plexinA3 and plexinA4 also display phenotypes that suggest that these plexins function as receptor components for Sema3F and Sema3A, respectively. Similarly, the neural phenotypes observed in fruit flies lacking Sema-1a resemble those of PlexA-deficient flies. Studies on genetically modified mice have also revealed additional novel functions for semaphorins in the developing nervous system, including their involvement in pruning of axons, synapse development and maturation, and the migration of neural crest and cerebellar granual cells.

Semaphorins in the Adult Intact Nervous System

The expression of several secreted and transmembrane semaphorins persists in the adult nervous system. This intriguing observation indicates that these guidance molecules also have functions in the fully wired adult brain. Unlike their role in neural development, research on a function for semaphorins in the adult nervous system is only just beginning.

So far, biochemical and cell culture studies have shown that several class 4 semaphorins are localized at synapses and associate with PDZ domains in postsynaptic density protein-95 (PSD-95), suggesting a role in as yet unknown aspects of synaptic functioning. Two Sema3s, Sema3F and Sema3A, have been studied in the context of synaptic transmission. Sema3F modulates synaptic transmission in acute slices of the adult rat hippocampus and Sema3A application to fully differentiated cultured hippocampal neurons decreases the efficacy of synaptic transmission. The expression of some secreted and membrane-associated semaphorins is sensitive to electrical activity, as has been shown in temporal lobe epilepsy and kainate acid-induced status epilepticus. Downregulation of Sema3s in the entorhinal–hippocampal complex of epileptic rats has been associated with pathological mossy fiber sprouting into the molecular layer of the hippocampus. Sema3A induces clustering of PSD-95 and NP-1 at synaptic sites in cultured neurons. In the cortex of adult Sema3A knockout mice, spine density appears to be reduced.

Semaphorins and Neurodegenerative and Psychiatric Disorders

Traumatic Injury

Regenerating axons stop growing when they reach the border of the glial–fibrotic scar because they encounter a potent molecular barrier that inhibits growth cone motility. Semaphorins have been implicated as factors

that contribute to the failure of axon regeneration. Theories on semaphorin function in traumatic nervous system injury are based on the cellular context in which semaphorins, their receptors, and intracellular signaling molecules are (re)expressed after damage. Four possible functions have been postulated. First, semaphorins may act in concert with the classical myelin-associated inhibitors of regeneration: multiple membrane-associated semaphorins are expressed by oligodendrocytes and Sema4D is upregulated in oligodendrocytes located in the vicinity of CNS lesions. Second, Sema3s are expressed by fibroblasts in the neural scar and may act as repulsive signals for injured axons approaching the scar. Evidence has shown that semaphorins can bind to chondroitin sulfate proteoglycans; these molecules are prominent components of the extracellular matrix of the neural scar and influence the activity of axonal guidance cues such as semaphorins. Third, neovascularization of the scar may be affected by semaphorins: blood vessels in the scar express Npn-1 and are surrounded by meningeal fibroblasts that secrete Sema3s. Sema3s may also directly influence the formation of the scar by affecting the migration of cells in the scar. Finally, in the peripheral nervous system semaphorins are regulated following injury in motor neurons, in epineural fibroblasts, and in a subpopulation of terminal Schwann cells at neuromuscular junctions where they may be involved in limiting plasticity of motor axons. The first causal evidence for an inhibitory role of Sema3s in axon regeneration comes from a study in which application of a Sema3A-blocking agent isolated from *Penicillum* resulted in enhanced regeneration. Additional evidence for a role of Sema3A in injury-associated changes in neurite outgrowth comes from observations on viral vector-mediated expression of Sema3A in the spinal cord following a dorsal root lesion. Enhanced ectopic Sema3A expression counteracts the nerve growth factor-induced sprouting response of injured sensory neurons, thereby preventing neuropathic pain.

Neurodegenerative Disorders: Alzheimer's and Parkinson's Disease

Alzheimer's and Parkinson's disease have in common the fact that specific populations of neurons and synaptic contacts atrophy and are then gradually lost. Alterations in neuronal connectivity in the mature nervous system appear to depend on a delicate balance between growth-promoting and growth-inhibiting influences. When this balance is disturbed, neurons may lose synaptic contacts and could eventually degenerate. It has been postulated that Sema3A may be involved in the pathogenesis of Alzheimer's disease. The expression and distribution of Sema3A are altered in the subiculum and hippocampus during the early stages of Alzheimer's disease. As described previously, Sema3A can stimulate Cdk-5 and GSK-3beta, which can lead to (hyper)phosphorylation of CRMP-2 and microtubule-associated protein-1B (MAP1B). During the clinical stages of Alzheimer's disease, a protein complex is formed in neurons that are affected by the disease consisting of Sema3A, plexinA1, plexinA2, and hyperphosphorylated CRMP-2 and MAP1B. It is not known whether Sema3A contributes directly or indirectly to the pathogenesis of Alzheimer's disease, but it is important to note that Sema3A can act as an apoptotic factor and as an axon retraction inducer during development. Both processes are part of the neuropathological repertoire in Alzheimer's disease. The possible involvement of Sema3A in dopamine-induced oxidative stress (one possible mechanism responsible for dopaminergic (DA) neuron death in Parkinson's disease) is so far only based on *in vitro* evidence. Application of Sema3A to E14 murine ventral mesencephalic neurons induces apoptosis of DA neurons. Antibodies that block Sema3A function can protect neurons from DA-induced apoptosis.

Neurodevelopmental Disorder: Schizophrenia

Two observations have linked possible changes in semaphorin signaling to schizophrenia. First, the expression of Sema3A is increased in the cerebellum of schizophrenic patients and elevated expression of Sema3A correlates with a decline in expression of synaptophysin and reelin, proteins involved in synapse formation and synaptic transmission. Second, variants of plexinA2 have been associated with schizophrenia in a genomewide association study. These findings are in line with an early developmental origin of this psychiatric disorder and may point to a disturbed balance between trophic and chemorepulsive influences in this disease.

Neuromuscular Disease: Amyotrophic Lateral Sclerosis

In amyotrophic lateral sclerosis (ALS), selective synaptic weakening and denervation of muscle fibers starts long before the actual motor neuron degeneration occurs and takes place prior to the clinical manifestation of the disease. Furthermore, fast-fatiguable muscle fibers are selectively denervated during the earliest phases of the disease. The chemorepellent Sema3A is selectively upregulated in terminal Schwann cell that occupy the neuromuscular junctions of fast-fatiguable muscle fibers in a mouse model of ALS. This observation has led to the hypothesis that Sema3A may act as a motor axon retraction inducer at the neuromuscular synapse in ALS, thereby contributing to the neuropathology of ALS. Alternatively,

elevated levels of Sema3A at the neuromuscular junction may affect retrograde signaling in motor neurons, a process that is disturbed in ALS.

Future Challenges

The current literature underscores that a more rigorous dissection of the function(s) of semaphorins in the adult intact and dysfunctional nervous system is essential. The hypotheses that predict a role for these chemorepulsive proteins in the normal physiology and plasticity of the adult brain and in neuropathological processes are predominantly based on descriptive findings. Many of the null mutant mice that have been so useful to study the role of semaphorins during development die soon after birth. Mouse mutants that do survive into adulthood may not be good models to study semaphorin function during adulthood because compensatory changes may mask functionally relevant phenotypes. Therefore, mice with conditional and cell type-specific semaphorin or semaphorin receptor gene ablation or transgenic animals with targeted and conditional overexpression will have to be generated to study the role of semaphorins in adult neurons. Furthermore, other experimental strategies (genetic modification based on direct viral vector-mediated expression in adult rodent brain and pharmacological intervention) will be instrumental to advance our knowledge of the role of semaphorins in the normal intact adult nervous system and under neuropathological conditions.

During the past century, our view of the nervous system as a fixed and immutable structure has changed radically. Most neuroscientists will agree that the nervous system is a highly plastic organ in which the structure and function of connections and synapses are continually modified. Like the formation of neuronal connections during development, the ongoing plasticity in the mature nervous system is regulated by a myriad of molecular interactions. It will be fascinating to witness how future neurobiological research will define the precise function(s) of semaphorins in plasticity of the mature nervous system. Moreover, whether or not the chemorepulsive activity of these molecules contributes to the limited capacity of the nervous system to repair itself after damage is another pivotal issue that will be resolved in the future.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Axon Guidance: Guidance Cues and Guidepost Cells; Axonal Pathfinding: Netrins; Axonal Pathfinding: Guidance Activities of Sonic Hedgehog (Shh); Axonal Pathfinding: Extracellular Matrix Role; Axon Guidance: Morphogens as Chemoattractants and Chemorepellants; Growth Cones; Neurodegeneration in Psychiatric Illness; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Adams RH, Lohrum M, Klostermann A, Betz H, and Puschel AW (2006) The chemorepulsive activity of secreted semaphorins is regulated by furin-dependent proteolytic processing. *EMBO Journal* 20: 6077–6086.

Bagri A, Cheng HJ, Yaron A, Pleasure SJ, and Tessier-Lavigna M (2003) Stereotyped pruning of long hippocampal axon branches triggered by retraction inducers of the semaphoring family. *Cell* 113: 285–299.

Barzilai A, Zilkha-Falb R, Daily D, et al. (2000) The molecular mechanism of dopamine-induced apoptosis: Identification and characterization of genes that mediate dopamine toxicity. *Journal of Neural Transmission Supplement* 60: 59–76.

Castellani V, Chédotal A, Schachner M, Faivre-Sarrailh C, and Rougon G (2000) Analysis of the L1-deficient mouse phenotype reveals cross-talk between sema3A and L1 signaling pathways in axonal guidance. *Neuron* 27: 237–249.

DeWinter F, Vo T, Stam FJ, et al. (2006) The expression of the chemorepellent Semaphorin 3A is selectively induced in terminal Schwann cells of a subset of neuromuscular synapses that display limited anatomical plasticity and enhanced vulnerability in motor neuron disease. *Molecular and Cellular Neuroscience* 32: 102–117.

Eastwood SL, Law AJ, Everall IP, and Harrison PJ (2003) The axonal chemorepellant semaphorin 3A is increased in the cerebellum in schizophrenia and may contribute to its synaptic pathology. *Molecular Psychiatry* 8: 1248–1255.

Giger RJ, Pasterkamp RJ, Heijnen S, Holtmaat AJGD, and Verhaagen J (1998) Anatomical distribution of the chemorepellent semaphorin III/collapsin-1 in the adult rat and human brain: Predominant expression in structures of the olfactory–hippocampal pathway and the motor system. *Journal of Neuroscience Research* 52: 27–42.

Good PF, Alapat D, Hsu A, et al. (2004) A role for semaphorin 3A signaling in the degeneration of hippocampal neurons during Alzheimer's disease. *Journal of Neurochemistry* 91: 716–736.

Goshima Y, Kawakami T, Hori H, et al. (1997) A novel action of collapsin: Collapsin-1 increases antero- and retrograde axoplasmic transport independently of growth cone collapse. *Journal of Neurobiology* 33: 316–328.

Kantor DB, Chivatakarn O, Peer KL, et al. (2004) Semaphorin 5A is a bifunctional axon guidance cue regulated by heparin and chondroitin sulfate proteoglycans. *Neuron* 44: 961–975.

Kapfhammer JP, Grunewald BE, and Raper JA (1986) The selective inhibition of growth cone extension by specific neurites in culture. *Journal of Neuroscience* 6: 2527–2534.

Kerjan G, Dolan J, Haumaitre C, et al. (2005) The transmembrane semaphorin Sema6A controls cerebellar granule cell migration. *Nature Neuroscience* 8: 1516–1524 (doi:10.1038/nn1555).

Kikuchi K, Kishino A, Konishi O, et al. (2003) *In vitro* and *in vivo* characterization of a novel semaphorin 3A inhibitor, SM-216289 or xanthofulvin. *Journal of Biological Chemistry* 278: 42985–42991.

Kolodkin AL, Levengood DV, Rowe EG, Tai YT, Giger RJ, and Ginty DD (1997) Neuropilin is a semaphorin III receptor. *Cell* 90: 753–762.

Kolodkin AL, Matthes DJ, and Goodman CS (1993) The semaphorin genes encode a family of transmembrane and secreted growth cone guidance molecules. *Cell* 75: 1389–1399.

Luo Y, Raible D, and Raper JA (1993) Collapsin: A protein in brain that induces the collapse and paralysis of neuronal growth cones. *Cell* 75: 217–227.

Pasterkamp RJ, Giger RJ, Ruitenberg MJ, et al. (1999) Expression of the gene encoding the chemorepellent semaphorin III is induced in the fibroblast component of neural scar tissues formed following injuries of adult but not neonatal CNS. *Molecular and Cellular Neuroscience* 13: 143–166.

Pasterkamp RJ, Peschon JJ, Spriggs MK, and Kolodkin AL (2003) Semaphorin 7A promotes axon outgrowth through integrins and MAPks. *Nature* 424: 398–405.

Sahay A, Kim C-H, Sepkuty JP, et al. (2005) Secreted semaphorins modulate synaptic transmission in the adult hippocampus. *Journal of Neuroscience* 25: 3613–3620.

Schwab ME and Thoenen H (1985) Dissociated neurons regenerate into sciatic but not optic nerve explants in culture irrespective of neurotrophic factors. *Journal of Neuroscience* 5: 2415–2423.

Takahashi T, Fournier A, Nakamura F, et al. (1999) Plexin–neuropilin-1 complexes form functional semaphorin-3A receptors. *Cell* 99: 59–69.

Tamagnone L, Artigiani S, Chen H, et al. (1999) Plexins are a large family of receptors for transmembrane, secreted, and GPI-anchored semaphorins in vertebrates. *Cell* 99(1): 71–80.

Taniguchi M, Yuasa S, Fujisawa H, et al. (1997) Disruption of semaphorin III/D gene causes severe abnormality in peripheral nerve projection. *Neuron* 19: 519–530.

Unified, Nomenclature for the Semaphorins/Collapsins(1999) Letter to the editor. *Cell* 97: 551–552.

Yang J, Houk B, Shah J, et al. (2005) Genetic background regulates semaphorin gene expression and epileptogenesis in mouse brain after kainic acid status epilepticus. *Neuroscience* 131: 853–869.

Yaron A, Huang P-H, Cheng H-J, and Tessier-Lavigne M (2005) Differential requirement for plexin-A3 and-A4 in mediating responses of sensory and sympathetic neurons to distinct class 3 semaphorins. *Neuron* 45: 513–523.

Sensation from the Face

B J Sessle, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Orofacial Touch, Thermal, Pain, and Related Sensations

The face and mouth have special biological, emotional, and psychological meaning. Humans, for example, use the orofacial region for ingestion of foods and liquids vital for life and for communicating with others by verbal and emotional expressions; furthermore, facial appearance is very important for most humans, and dysfunction or pain in the face and mouth can severely affect quality of life. In addition, some tissues in the face and mouth are unique in the body (e.g., teeth, cornea), and a wide range of sensory functions occurs in the orofacial region. While this article is limited to the somatosensory functions of touch, warmth, cool, and pain, it should be noted that the special senses of taste and smell are unique to this region, and sensory information from the mouth provides us with information about the texture, hardness, brittleness, and viscosity of foodstuffs.

As a consequence of the dense and diverse innervation of the face and mouth, the orofacial perceptions of tactile, thermal, and nociceptive events are unsurpassed in detail. With the exception of sensitivity to high-frequency mechanical stimuli, the tactile sensitivity of some orofacial tissues is as high, or higher, than any other region of the body. The same applies to the thermal sensations of cool and warmth as well as to pain sensibility. Nerve signals from the orofacial region provide us with sensory information on the temperature of some foodstuffs or liquids that we consume, and thereby not only add to our enjoyment of what we eat and drink but also help protect us from ingesting food or fluids that are too hot or too cold. Furthermore, the face and mouth represent sites of some of the most common pains in the body. Although pain arising from orofacial tissues shares many features in common with the pain originating in other body regions, there are several pain conditions that are unique to this region, for example, toothache, trigeminal neuralgia, various types of headaches, and temporomandibular disorders (TMD).

The following sections provide an overview of the neural mechanisms underlying these somatosensory functions. Peripheral processes are considered first, and then attention is given to the mechanisms in the central nervous system (CNS) that underlie these functions.

Peripheral Mechanisms

Like peripheral tissues elsewhere in the body, those in the orofacial region are supplied by primary afferent nerve fibers that terminate in sense organs (receptors) that respond to peripheral stimulation of the tissues; nearly all these fibers are carried in the trigeminal (V) nerve, the fifth cranial nerve. Most of the small-diameter, slowly conducting primary afferents terminate as nonspecialized free nerve endings which are activated by noxious stimuli, and so function as nociceptors, although some terminate in endings sensitive to cooling or warming stimuli and some may respond to nonnoxious mechanical stimuli. The larger-diameter, faster-conducting primary afferents end as specialized low-threshold receptors (e.g., Ruffini-like; Pacinian-like) that respond to nonnoxious mechanical stimuli or movements. In most muscles, some of these large-diameter afferents are associated with muscle spindles and Golgi tendon organs that respond, as in comparable tissues in other parts of the body, to muscle stretch and contractile tension respectively. Nearly all the afferents that supply these specialized and nonspecialized receptors are less than 10 μm in diameter, and have conduction velocities in the A–β, A–δ, and C-fiber ranges. However, not all the properties of the peripheral components of the V sensory system are equivalent to analogous components of the spinal system. There are some notable differences, and unique features, as outlined in Table 1.

Mechanoreceptors

The orofacial region has a dense and diverse supply of mechanoreceptors that provide the CNS with sensory input that is used for fine tactile detection and discrimination and for the regulation of the many behaviors that characterize orofacial function. The mechanoreceptors in the facial skin and oral mucosa may also be activated during lip and jaw movements associated with chewing, swallowing, speech, etc., and so may also provide proprioceptive information of importance to the control of orofacial motor functions. Like those in other parts of the body, facial and intraoral mechanoreceptors include slowly adapting type I and type II receptors, rapidly adapting type I receptors, and hair follicle receptors. Mechanoreceptors also occur in the periodontal tissues that surround the root of the teeth. These periodontal mechanoreceptors can accurately encode information about which teeth are loaded and the intensity, velocity, or direction of forces applied to one or more teeth. Their input to the CNS is crucial for the detection and

Table 1 Trigeminal sensory systems: differences from spinal sensory systems

Peripheral tissues and innervation
1. Tissues unique to the craniofacial region (e.g., tooth pulp; cornea)
2. Higher innervation density in many craniofacial tissues than in most spinally innervated tissues
3. Shorter conduction distances of peripheral nerve pathways
4. Slower conduction velocities of peripheral nerve fibers
5. Higher ratio of myelinated: unmyelinated fibers
6. Lower proportion of sympathetic efferents
7. Certain craniofacial receptors (e.g., some periodontal mechanoreceptors; jaw muscle spindles) have their primary afferent cell bodies within CNS
Central nervous system
1. Face and mouth represented completely at most rostrocaudal levels of VBSNC, and a dual representation of some tissues occurs in Vc
2. Distinctive brain stem termination patterns of some nociceptive afferents
3. Transitional regions between Vc and Vi, and between Vc and CDH, with distinctive properties (e.g., bilateral afferent inputs to Vc/Vi)
4. 'Deep bundle' fiber system especially prominent in Vc (connects caudal and rostral levels of VBSNC), whereas Lissauer's tract absent
5. Significant ipsilateral as well as contralateral projections from VBSNC to thalamus
Pain conditions specific to craniofacial region
1. Headaches (e.g., migraine; cluster headache)
2. Toothaches (e.g., pulpitis pain)
3. Trigeminal neuralgia
4. Miscellaneous (e.g., atypical facial pain; burning mouth syndrome; atypical odontalgia)

VBSNC: Trigeminal brain stem sensory nuclear complex; CDH: upper cervical dorsal horn; Vc: subnucleus caudalis; Vi: subnucleus interpolaris.

Reproduced from Sessle BJ (2005) Orofacial pain. In: Merskey H, Loeser JD, and Dubner R (eds.) *The Paths of Pain 1975–2005*, pp. 131–150. Seattle: IASP Press, with permission from IASP Press.

discrimation of occlusal forces applied to the teeth, and are also used in the fine motor control of masticatory function associated with the intraoral manipulation of foodstuffs.

The temporomandibular joint (TMJ) is also supplied by low-threshold nonnociceptive afferents with either slowly adapting or rapidly adapting responses to joint movement or to a change in jaw position. These responses have been implicated in both mandibular movement and position sense (kinesthesia). Nonetheless, other nonarticular primary afferents may show activity during jaw movements; these include not only skin and intraoral afferents as noted above, but also afferents supplying the muscle spindles that occur in some masticatory muscles. By their movement-related activity, these nonarticular afferents may thereby contribute to kinesthetic sense and motor control. The contribution of joint receptors, relative to muscular and cutaneous receptors, to joint position sense has been the subject of considerable investigation and debate in both humans and experimental animals, and it is currently thought that their role is quite limited and only becomes significant at extreme joint positions.

Thermoreceptors

As for sensations evoked in other parts of the body, thermal sensations in the face and mouth are subserved by thermoreceptors that respond specifically to cold or warm stimulation. The nasal mucosa and anterior tongue are especially sensitive to small changes in temperature, and the exquisite sensitivity of cold or warm thermoreceptors in these regions have been well documented. These warm and cold thermoreceptors respond to the resting temperatures of the tissues as well as to changes in temperature and so underlie the human capacity to distinguish between thermal stimuli at very low to very high temperatures. Some slowly adapting mechanoreceptors may also respond to cooling stimuli, but their role in cold sensation seems to be minor.

Nociceptors

Like nociceptors elsewhere in the body, those in the orofacial region are associated with a diverse population of mechanical, thermal, and chemically responsive A–δ and C-fiber afferents that detect and encode intense physical, thermal, and chemical events associated with actual or near-tissue damage. As in other parts of the body, tissue damage or inflammation can produce an increased excitability of the nociceptors at the site of an orofacial injury. The increased excitability of the nociceptors (so-called peripheral sensitization) can be reflected in their spontaneous activity, lowered activation thresholds, and increased responsiveness to subsequent noxious stimuli which appear to contribute, respectively, to the spontaneous pain, allodynia (pain produced by a

stimulus that is normally nonnoxious), and hyperalgesia (increased sensitivity to noxious stimuli) that are features of many chronic or persistent orofacial pain conditions.

The properties of those nociceptive afferents supplying facial skin or oral mucosa appear to be generally much the same as those of spinal nociceptive afferents, and include high-threshold mechanosensitive afferents (that are mostly A–δ), mechanothermal nociceptive afferents (also A–δ), and A–δ or C-polymodal nociceptive afferents excited by intense mechanical, thermal, or chemical stimuli. The cornea is a unique tissue found only in the orofacial region and is supplied by A–δ and C-fibers that respond to a wide range of mechanical, thermal, and chemical stimuli. Despite their low activation thresholds, their input to the CNS can evoke pain. The masticatory muscles and TMJ also are supplied by A–δ and C-fiber afferents that may respond to a wide range of peripheral stimuli, including heavy pressure and algesic chemicals. Small-diameter afferents also supply the cerebrovasculature and dura. Many of these afferents can be activated by noxious stimuli, but their modulation by peripheral chemical mediators (e.g., serotonin (5-HT), calatonin gene-related peptide (CGRP)) appears to be important in the initiation and control of certain types of headaches (e.g., migraine).

The tooth pulp of course is a tissue unique to the orofacial region, and its nerve supply warrants some particular emphasis. It is richly innervated by A–δ and C-fiber afferents, but it also receives some A-beta afferents and sympathetic efferents. The A–β, A–δ afferents are sensitive to a wide range of dental stimuli that produce pain in humans. Since A–β afferents are not normally nociceptive elsewhere in the body, it is unclear if these intradental afferents represent a notable exception or, rather, that their sensitivity has been altered by the experimental procedures required for the recording of their activity. Recent electrophysiological recording studies have suggested that many A-fibers terminate in the tubules within dentine and are responsible for the sensitivity of dentine, probably through a hydrodynamic mechanism, whereas C-fibers predominate in the pulp and may be activated only if the dental stimulus reaches the pulp. Injury or inflammation of the pulp can lead to neurochemical changes and nerve-sprouting as well as induce peripheral sensitization of intradental afferents. Since inflammation of the pulp occurs in a noncompliant environment (it is encased by dentine) with a high extracellular tissue pressure, these features may account for the exquisite sensitivity of pulp afferents and explain why toothache is such an intense pain.

Neurochemical Features

In vivo and *in vitro* recordings and molecular and immunocytochemical approaches have shown that the V primary afferent cell bodies in the V ganglion (see below) synthesize the vast array of chemicals that help define the role that the primary afferent neurons play in orofacial sensation. These include CGRP, substance P, somatostatin, and nerve growth factor, and the afferents may express several different types of receptors or ion channels (e.g., cholinergic, purinergic, bradykinin, histamine, anandamide, prostaglandin, transient receptor potential vanilloid receptor (TRPV1)). Each may be involved in the activation, sensitization, or modulation of orofacial afferent inputs into the CNS. In addition, receptors also exist for chemical mediators long thought to be involved in nociceptive transmission or modulation within the CNS, for example, glutamate, opiate, and γ-aminobityric acid (GABA) receptors, and sex differences occur in the peripheral actions of glutamate and morphine suggesting that peripherally based physiological mechanisms may contribute to the sex differences documented in the prevalence of many chronic pain conditions.

Central Trigeminal Processes

Brain Stem Nociceptive and Nonnociceptive Transmission

In the spinal somatosensory system, all spinal afferents supplying deep as well as superficial tissues have their primary afferent cell bodies in the dorsal root ganglia. In the V system, nearly all the primary afferents supplying orofacial tissues have their peripheral cell bodies in the V (Gasserian or semilunar) ganglion. The central projections of these primary afferent cell bodies enter the brain stem and may ascend or descend in the V spinal tract from which they give off collaterals that terminate in one or more subdivisions of the V brain stem sensory nuclear complex (VBSNC). Notable differences from the spinal system are that muscle spindles occur only in some muscles (e.g., jaw-closing musculature) and that their afferents, and some mechanosensitive afferents supplying periodontal tissues, have their primary afferent cell bodies within the CNS, in the V mesencephalic nucleus. From here they project to the V motor nucleus, or to adjacent nuclei where they excite interneurons involved in orofacial reflex function. In addition to its predominant input from V nerve afferents, the VBSNC (especially its subnucleus caudalis) may also receive afferent inputs from other cranial nerves (e.g., VII, IX, X, and XII) as well as from upper cervical nerves.

The VBSNC consists of the main sensory nucleus and the spinal tract nucleus, and the latter is subdivided into three subnuclei: oralis, interpolaris, and caudalis. Subnucleus caudalis is a laminated structure with many morphological and functional similarities with the spinal dorsal horn (although see below) to the extent that caudalis is now often termed the medullary dorsal horn. The structure of the other three components of the VBSNC is more uniform, although subdivisions of each can be distinguished in some species. Each component contains neurons, glia, and numerous nerve fibers that are the collaterals of the incoming primary afferents, as well as fibers intrinsic to the VBSNC (e.g., the 'deep bundles') and fibers derived from other brain stem or central neural structures that contribute to the substrate for modulatory influences on VBSNC neurons (see below). The V spinal tract lateral to each nucleus or subnucleus is composed largely of nerve fibers, but islands of neuronal elements are present in or adjacent to the tract lateral to subnuclei caudalis and interpolaris (the so-called paratrigeminal nucleus or interstitial islands).

The low-threshold mechanosensitive primary afferents terminate mainly in the more rostral components of the VBSNC and in laminae III–VI of subnucleus caudalis. Subnucleus caudalis also receives thermosensitive afferent inputs. Some cutaneous nociceptive afferents and dental pulp afferents terminate in some of the rostral components but most of the small-diameter V nociceptive afferents terminate in laminae I, II, V, and VI of caudalis. Some of the small-diameter afferents stain positive for CGRP, substance P, and neurotrophins, whereas others stain negative for the neuropeptides but positive for the cell surface marker, isolectin B4. It is noteworthy that these IB4 afferents have a different distribution in caudalis compared with the spinal dorsal horn.

Many neurons in the four components of the VBSNC project to CNS regions involved in somatosensory function or modulation, such as ventrobasal thalamus, hypothalamus, pontine parabrachial area, periaqueductal gray, or brain stem reticular formation. Some of the connections to the reticular formation and other parts of the brain stem are utilized in autonomic reflex responses to orofacial stimuli. Some neurons in the VBSNC and adjacent to it (e.g., supratrigeminal nucleus and reticular formation) also serve as interneurons in orofacial and cervical muscle reflex pathways, and there are also intrinsic connections between neurons in the different components of the VBSNC. One example is the caudalis axonal projections via the deep bundles to subnucleus oralis that underlie modulatory influences between rostral and caudal VBSNC neurons. The lamina II of caudalis, the so-called substantia gelatinosa (SG), also represents an important interneuronal system contributing to segmental and descending modulation of somatosensory transmission in subnucleus caudalis and more rostral components of the VBSNC (see below).

The VBSNC is somatotopically organized: the neurons receiving ophthalmic afferent input occur in the ventral part and those with mandibular afferent input are located in the dorsal part of each nucleus or subnucleus. The maxillary region is represented in between, and oral and perioral structures are represented medially. This medially facing, inverted somatotopic pattern of the face and mouth may shift in subnucleus caudalis to an 'onion-skin'-like arrangement, with perioral regions represented in the rostral part of the subnucleus and more lateral regions of the face more caudally. Investigations specifically of the somatotopic pattern of the vibrissae (facial whiskers) in the rodent VBSNC have revealed that the brain stem projections of single vibrissal afferents are organized into columns or cylinders that run rostrocaudally and are somatotopically parcellated in the transverse plane in the main sensory nucleus and in subnuclei interpolaris and caudalis; however, subnucleus oralis does not have such a clear-cut spatial patterning. This patterning, so-called barreletes, reflects the pattern of vibrissae on the ipsilateral face, and the vibrissal afferents and most low-threshold mechanoreceptive (LTM) neurons associated with each barrelete are optimally responsive to tactile stimulation of a single vibrissa.

Structure–function relationships have been well-documented in the spinal cord and dorsal column nuclei but are usually less clear-cut and often variable between types of afferents in the four different components of VBSNC. Nonetheless, investigations combining anatomical intra-axonal labeling with electrophysiological identification of the functional properties of labeled single primary afferents have shown that the spatial distribution of an afferent's collateral endings is influenced by the location in the orofacial region of the afferent's mechanoreceptive field (RF), thus accounting for the somatotopic pattern mentioned above. As for the V primary afferent endings, the morphological and physiological features of neurons in the VBSNC reveal also some fundamental differences between V and spinal systems in the extent of structure–function correlations, for example, correlations have been shown to be a strong feature for vibrissa-sensitive neurons in some parts of the VBSNC but not in other parts.

Properties of LTM Neurons

The rostral components of the VBSNC have long been considered to be the essential brain stem regions that relay orofacial mechanosensory information related to so-called 'fine touch' and thereby to be analogous to the dorsal column nuclei of the spinal somatosensory system. Such a view has been reinforced by findings that the RF properties of some of the neurons that relay this mechanosensory information (LTM neurons) may allow for a precise coding of orofacial mechanical stimuli, for example, the single-vibrissa receptive field of most LTM neurons in the V main sensory nucleus versus the multi-vibrissa RF of the majority of more caudally located neurons in the rat. Nonetheless, LTM neurons are abundant in all four nuclei/subnuclei of the VBSNC, including subnucleus caudalis where they predominate in laminae III/IV. At least some LTM neurons at all levels of the VBSNC appear to be capable of receiving and faithfully transmitting detailed somatosensory information about light tactile stimuli delivered to localized regions of the face and mouth, in keeping with the evidence mentioned above that V low-threshold mechanosensitive afferents project to all levels of the VBSNC.

Like analogous neurons in the dorsal column nuclei and spinal dorsal horn, LTM neurons in the VBSNC are activated by light tactile stimuli applied to a localized RF, show graded responsiveness as stimulus intensity or area of RF stimulation is gradually increased, and receive low-threshold A-fiber afferent inputs that may release glutamate as the principal excitatory neurotransmitter. In general, as noted above, they have functional properties ensuring a secure transmission through the VBSNC of tactile information from orofacial mechanoreceptors. The LTM neurons thereby provide the higher levels of the CNS with detailed information of the modality and spatiotemporal features (e.g., location, intensity) of an orofacial tactile stimulus.

Properties of Other Nonnociceptive Somatosensory Neurons

A small proportion of neurons in the VBSNC can be excited by low-threshold afferent inputs other than those derived from orofacial mechanoreceptors in skin, mucosa, or periodontal tissues. Some neurons, especially in the rostral components, can be activated by jaw movements (in particular, jaw opening), presumably through inputs from low-threshold mechanoreceptors in the TMJ or stretch-sensitive receptors in masticatory muscles. These neurons may contribute to ascending sensory pathways (e.g., to thalamus and cortex) or to brain stem circuits underlying the reflex regulation of masticatory muscle activity.

Some LTM neurons also respond to cooling of their orofacial RF, but other neurons, especially in V subnucleus caudalis, respond exclusively and with much greater sensitivity to innocuous cooling or warming. These are specific thermosensitive neurons which receive their afferent input from the specific cold or warm thermoreceptors mentioned earlier, and relay detailed information about orofacial thermal stimuli (e.g., intensity, duration, location) to the higher brain centers.

Properties of Nociceptive Neurons

Several lines of evidence indicate that subnucleus caudalis is critical in V brain stem nociceptive processing. These include its nociceptive afferent inputs, its laminated morphology and cell types resembling those of the spinal dorsal horn, and findings that the disruption of caudalis interferes with many behavioral responses as well as higher center and reflex responses evoked by noxious stimuli. In addition, two classes of nociceptive neurons, based on their cutaneous RF and response properties, occur in the superficial and deep laminae of caudalis: high-threshold or nociceptive-specific (NS) neurons and wide dynamic range (WDR) neurons. Like their spinal dorsal horn counterparts, NS neurons receive small-diameter nociceptive afferent inputs and so respond only to noxious stimulation whereas WDR neurons receive both large- and small-diameter afferent inputs and respond to both tactile and noxious stimuli. Additional features of caudalis that are consistent with its nociceptive role include the several receptor types associated with nociceptive processing that are present in the afferent endings or neurons in caudalis (e.g., neurokinin, opiate, GABA, purinergic, TRPV1, estrogen, and *N*-methyl-D-aspartate (NMDA) and non-NMDA receptors), and the presence of a SG which, like the SG of the spinal dorsal horn, has neurons with axons that terminate locally within the VBSNC and release neurochemical substances such as enkephalin or GABA that modulate somatosensory transmission in caudalis (also see below).

The caudalis NS and WDR neurons with a cutaneous RF have been shown to play an important role in our ability to localize, detect, and discriminate cutaneous noxious stimuli. The majority of the caudalis NS and WDR neurons can nonetheless be excited by afferents from deep tissues (e.g., tooth pulp, cerebrovasculature, TMJ, jaw muscle) as well as by cutaneous (or mucosal) afferent inputs, and some can be

even activated by other cranial nerve afferents or cervical afferents. Such features are thought to contribute to central sensitization (see below) and to the referred pain patterns that are common in many pain conditions in the head and neck, and also underscore the inadvisability of using 'convergent' neurons as an alternative term for WDR neurons, since NS neurons typically also reveal extensive afferent convergence.

The similarities between caudalis and spinal dorsal horn have led to the long-held concept that these two structures are homologous, but this useful homology may need some revision as recent evidence indicates that portions of caudalis are organized differently from the spinal system (also see Table 1). A unique feature of caudalis is its dual representation of several different orofacial tissues in its rostral border with subnucleus interpolaris as well as in the transitional region between the caudal part of caudalis and the upper cervical spinal cord. Immunohistochemical and electrophysiological studies have shown that the caudalis/interpolaris transition region has a ventral transition area, which is especially clear in rodents and includes bilateral orofacial afferent inputs, and a dorsomedial transition area. Each of these two areas of the rostral caudalis, plus the caudal transitional part of caudalis, has its own unique morphological and functional features that may be differentially involved in contributing to perceptual, autonomic, endocrine, and muscle reflex responses to orofacial stimuli.

In addition, subnucleus caudalis is not the only VBSNC component with a nociceptive role. Some orofacial nociceptive behaviors may persist after caudalis lesions, lesions of rostral components (e.g., oralis) of the VBSNC may disrupt some orofacial nociceptive responses, and the rostral components have substantial numbers of NS and WDR neurons with an intraoral or perioral nociceptive RF (including tooth pulp). Furthermore, the rostral components of the VBSNC may also contribute to ascending and reflex nociceptive pathways and manifest some neurochemical markers for nociceptive processes. Nonetheless, these rostral regions receive only sparse direct input from small-diameter primary afferents and have a limited neurochemical substrate and lack the modulatory SG circuitry that are characteristic of caudalis (and spinal dorsal horn). Indeed, their nociceptive functions may be dependent on the functional integrity of caudalis since recent investigations have documented that caudalis, especially via its SG, exerts a modulatory influence on oralis through processes that include NMDA, purinergic, and opioid receptor mechanisms within caudalis itself (see also below). These various findings collectively suggest that the cutaneous RF and response properties of rostral nociceptive neurons, particularly in subnucleus oralis, may be dependent on caudal regions such as subnucleus caudalis for the relay of nociceptive signals from primary afferents supplying superficial craniofacial tissues. However, the responses of some neurons in the rostral VBSNC to tooth pulp stimulation may be dependent on pulp afferent relays in both rostral and caudal components of the VBSNC.

Thalamocortical Processes

Neurons in the VBSNC project to higher levels of the CNS, in particular to the posterior thalamus, and from there to the cerebral cortex. The thalamus receives direct contralateral input from each component of the VBSNC: an ipsilateral projection also occurs from some components. Although, in general, the majority of the V-thalamic projection neurons occur in the trigeminal main sensory nucleus, and are primarily involved in tactile sensation, the major thalamic projection related to pain and temperature perception arise mainly from neurons in subnucleus caudalis; there are, however, species differences in some of these projections.

The posterior thalamic regions that are especially involved in receiving and relaying this orofacial sensory information are the ventrobasal complex (VB) (or ventroposterior nucleus in humans), the posterior group of nuclei (PO), and parts of the medial thalamus. Those VB neurons that receive and relay somatosensory information from the face and mouth are concentrated in the medial half of the VB (the nucleus ventralis posteromedialis, VPM), whereas neurons in the lateral portion of the VB relay somatosensory spinal information derived from the limbs, trunk, and neck. The VPM neurons are somatotopically organized, and most are LTM neurons with a localized RF and other properties that allow for retention of the functional specificity that is characteristic of the LTM primary afferents and second-order LTM neurons. Many nociceptive neurons also occur in the VPM and generally have properties similar to those described for NS and WDR neurons in the VBSNC. The most medial aspect of the VB contains neurons that relay orofacial thermosensitive or gustatory information. The neurons in the PO and medial thalamus in general are much less specific in the information they relay. Nociceptive neurons may also occur in other areas such as hypothalamus and basal ganglia.

The RF and response properties of the LTM and nociceptive (NS and WDR) neurons in VPM and their connections with the overlying somatosensory cerebral cortex indicate that most may be involved in the

spatiotemporal coding of touch and pain, respectively. The somatosensory cortex also contains neurons that receive thalamic relays of gustatory or thermosensitive information, but the latter, in particular, have not been extensively studied. The primary somatosensory cortex (SI) is critical for fine tactile discrimination and face SI contains many LTM neurons which in general retain the features of the primary afferents, VBSNC, and VPM neurons that provide for detection and discrimination of the spatiotemporal characteristics of an orofacial tactile stimulus. SI (and VPM) neurons responding to tactile stimulation of the vibrissae aggregate in a barrel-like arrangement and their properties have been extensively studied. In comparison, only a limited number of studies have tested for nociceptive neurons in face SI. Nonetheless, NS and WDR neurons do occur in the face SI and respond to noxious facial or tooth pulp stimuli in a manner suggesting a role in the sensory-discriminative dimension of pain. Neurons responding to orofacial tactile or noxious stimuli also occur in other cortical regions such as the anterior cingulate cortex, the function of which includes the affective or motivational dimension of pain.

Trigeminal Somatosensory Modulation

Orofacial somatosensory transmission in the brain stem, thalamus, and somatosensory cortex can be modulated at each of these levels, although the modulation in the thalamus and cortex in large part reflects modifications of the sensory signals occurring first in the VBSNC. The intricate organization of each subdivision of the VBSNC and the variety of inputs to each from peripheral orofacial tissues or from different parts of the CNS provide important substrates for numerous interactions between the various inputs. For example, the responses of brain stem LTM neurons to orofacial tactile stimuli can be modified by anesthesia, by behavioral state, and by segmental (afferent) influences and modulatory influences intrinsic to the CNS that are derived from structures within the VBSNC itself as well as from other parts of the brain stem and higher centers (e.g., periaqueductal gray and somatosensory cortex).

Inhibitory Influences

Some of the first studies detailing segmental (so-called afferent) inhibition of tactile-transmission neurons were demonstrated in LTM neurons at various levels of the VBSNC. Subsequent studies also documented such inhibitory influences on thalamic and cortical LTM neurons. As in the spinal somatosensory system, these modulatory effects are thought to represent mechanisms for 'sharpening' the coding properties of LTM neurons and so contribute to, for example, spatial acuity in orofacial touch sensation. Afferent influences have also been shown to modulate V nociceptive transmission, especially through the interneuronal circuitry and neurochemical substrate existing within caudalis. These influences include inputs into the CNS from nociceptive and nonnociceptive afferents as well as visceral afferents (e.g., baroreceptors). The modulation by nociceptive inputs underlies the powerful diffuse noxious inhibitory controls (DNIC) effects that can be exerted on V nociceptive neurons, for example, by noxious stimulation of spially innervated tissues. These effects of noxious and nonnoxious afferent inputs may contribute to the analgesia reported for acupuncture and transcutaneous electrical nerve stimulation.

Electrical or chemical stimulation of the periaqueductal gray matter or rostroventral medial medulla activates descending pathways that project to the VBSNC and can inhibit V brain stem neuronal and related reflex and behavioral responses to noxious as well as nonnoxious orofacial stimulation in experimental animals. Other powerful modulatory effects on V somatosensory transmission include pathways emanating from the locus coeruleus, reticular formation, pontine parabrachial area, anterior pretectal nucleus, amygdala, red nucleus, and thalamic nucleus submedius, as well as the cerebral cortex (e.g., somatosensory and motor areas). Some of these influences are also expressed on somatosensory transmission in the thalamus and cortex. The descending pathways to the VBSNC exert their effects by the release from their endings within the VBSNC of certain neurochemicals (e.g., 5-HT) or by their causing other chemicals (e.g., enkephalins, GABA) to be released from the endings of interneurons intrinsic to the VBSNC (e.g., in the SG of caudalis). These pathways and chemical processes are involved in the modulation of somatosensation that occurs in a diverse set of conditions that include different states of consciousness and attention, emotion, exaggerated pain states, and the analgesia associated with deep brain stimulation and placebo effects.

Facilitatory Influences

As in the spinal somatosensory system, afferent inputs and intrinsic CNS processes may produce a prolonged facilitation of somatosensory transmission in certain conditions. In the V system, extensive investigations of the organization and neuronal properties of the whisker barrelettes in the VBSNC and analogous structures in VB thalamus and somatosensory cortex have revealed the susceptibility of the LTM neuronal properties to peripheral manipulations such as deafferentation in neonatal animals and even in adults. The effects of such deafferentation include facilitatory influences that have been viewed as evidence

of the neuroplasticity of the vibrissal somatosensory system. Interestingly, studies of the effects of the neonatal depletion of C-fibers reveal that C-fiber afferents appear to play an important role in shaping the adult RF and response properties of the LTM neurons. Furthermore, deafferentation of other orofacial tissues such as the tooth pulp in adult animals can also lead to profound changes in the properties of LTM neurons at the different levels of the VBSNC.

Nociceptive transmission in the V system can also be facilitated by alterations to the peripheral afferent inputs to the CNS that may result from inflammation or trauma of peripheral tissues and nerves, thus emphasizing the plasticity of the neural circuitry underlying nociceptive transmission, that is, that it is not 'hard-wired.' Afferent inputs from deep musculoskeletal tissues (e.g., TMJ, masticatory muscle) and tooth pulp are especially effective in inducing spontaneous activity, lowering of activation threshold, RF expansion, and enhancement of peripherally evoked responses of caudalis NS and WDR neurons. These features reflect an increased neuronal excitability analogous to the 'central sensitization' documented in the spinal nociceptive system and implicated in allodynia, hyperalgesia, and pain spread and referral following tissue injury or inflammation. Maintenance of these changes is thought to underlie chronic pain states. Studies in the V system have also shown that purinergic as well as glutamatergic (NMDA and non-NMDA) receptor mechanisms are involved in this central sensitization process, and that central depressive influences (e.g., opioid-related) can normally be 'triggered' by noxious orofacial stimulation and may serve to modulate, even limit, the sensitization. It is also noteworthy that V central sensitization is not restricted to subnucleus caudalis. It has also been shown in nociceptive neurons in subnucleus oralis and in higher brain regions such as VPM thalamus, although caudalis is responsible for their expression of central sensitization by way of its projections to both structures.

In addition to central sensitization, several other factors have been implicated in acute and chronic orofacial pain mechanisms related to peripheral tissue trauma or inflammation. These include ectopic impulses generated in damaged afferent fibers; peripheral sensitization, enhanced sympathetic modulation, and central sprouting of afferent fibers; phenotypic changes in afferent fibers and central nociceptive neurons; and changes in segmental and descending inhibitory and facilitatory influences on nociceptive transmission. Recent findings suggest that nonneural cells (e.g., glia) may also play an important role in some of these processes.

See also: Cingulate Cortex; Mechanoreceptors; Multisensory Convergence and Integration; Neural Circuitry in the Somatosensory System; Nociceptor Responses; Pleasant Touch; Somatosensory Cortex: Functional Architecture; Somatosensory Receptive Fields; Somatosensory Cortex; Tactile Texture; Temperature Sensation; Trigeminal Motor System.

Further Reading

Bereiter DA, Hiraba H, and Hu JW (2000) Trigeminal subnucleus caudalis beyond homologies with the spinal dorsal horn. *Pain* 88: 221–224.

Byers MR and Närhi MVO (1999) Dental injury models: Experimental tools for understanding neuroinflammatory interactions and polymodal nociceptor functions. *Critical Reviews in Oral Biology and Medicine* 10: 4–39.

Campbell JN and Meyer RA (2005) Neuropathic pain: From the nociceptor to the patient. In: Merskey H, Loeser JD, and Dubner R (eds.) *The Paths of Pain 1975–2005*, pp. 229–242. Seattle: IASP Press.

Chiang CY, Hu B, Park SJ, et al. (2003) Purinergic and NMDA-receptor mechanisms underlying tooth pulp stimulation-induced central sensitization in trigeminal nociceptive neurons. In: Dostrovsky JO, Carr DB, and Koltzenburg M (eds.) *Proceedings of the 10th World Congress on Pain, Progress in Pain Research and Management*, vol. 24, pp. 345–354. Seattle: IASP Press.

Darian-Smith I (1966) Neural mechanisms of facial sensation. *International Review of Neurobiology* 9: 301–395.

Dostrovsky JO and Craig AD (2006) Ascending projection systems. In: McMahon SB and Koltzenburg M (eds.) *Textbook of Pain*, 5th edn., pp. 187–203. London: Elsevier.

Dostrovsky JO and Sessle BJ (2006) Nociceptive processing in the brainstem. In: Schmidt RF and Willis WD (eds.) *Encyclopedic Reference of Pain*. Heidelberg: Springer.

Dray A (2004) Future pharmacologic management of neuropathic pain. *Journal of Orofacial Pain* 18: 381–385.

Dubner R and Ren K (2004) Brainstem mechanisms of persistent pain following injury. *Journal of Orofacial Pain* 18: 299–305.

Dubner R, Sessle BJ, and Storey AT (1978) *The Neural Basis of Oral and Facial Function*. New York: Plenum Press.

Lam DK, Sessle BJ, Cairns BE, and Hu JW (2005) Neural mechanisms of temporomandibular joint and masticatory muscle pain: A possible role for peripheral glutamate receptor mechanisms. *Pain Research and Management* 10: 145–152.

Sessle BJ (2000) Acute and chronic craniofacial pain: Brainstem mechanisms of nociceptive transmission and neuroplasticity, and their clinical correlates. *Critical Reviews in Oral Biology and Medicine* 11: 57–91.

Sessle BJ (2005) Orofacial pain. In: Merskey H, Loeser JD, and Dubner R (eds.) *The Paths of Pain 1975–2005*, pp. 131–150. Seattle: IASP Press.

Svensson P and Graven-Nielsen T (2001) Craniofacial muscle pain: Review of mechanisms and clinical manifestations. *Journal of Orofacial Pain* 15: 117–145.

Sensorimotor Control of Manipulation

R S Johansson, Umeå University, Umea, Sweden
J R Flanagan, Queen's University, Kingston, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Contact Events and Action Goals in Manipulation Tasks

Most natural object manipulation tasks involve a series of actions or phases that accomplish specific goals (or task subgoals) typically associated with mechanical contact events. For example, consider the task of lifting, holding, and replacing a box on a tabletop. This task involves a series of action phases separated by contact events involving either the making and breaking of contact (Figure 1(a)). Thus, the goal of the initial reach phase is marked by the digits contacting the box, and the goal of the subsequent load phase (during which increasing vertical load forces and horizontal grip forces are applied under isometric conditions) is marked by the breaking of contact between the object in hand and the support surface. These and subsequent contact events give rise to discrete sensory signals from one or more sensory modalities. For example, when the box is replaced on the table, the contact between the box and surface gives rise to discrete tactile and auditory signals and, if the box is in the field of view, visual signals as well. It is likely that such contact events also give rise to discrete signals in proprioceptive systems. Thus, in general, contact events can be associated with distinct signals in multiple sensory modalities, each providing a neural signature of the event. These signals provide information about not only the timing of the event but also the characteristics of the mechanical interaction (e.g., direction and amplitude of contact forces). Thus, a given object-manipulation task can be represented as a sensory plan wherein a sequence of sensory goals is specified. The implementation of such a plan requires the selection and execution of a corresponding sequence of basic actions, or action-phase controllers, to achieve the sensory goals. To be accurate, each of these controllers requires information about the initial configuration of the motor system and objects in the environment, as well as their mechanical properties. This state information (multimodal) is used for feed-forward or predictive adaptation of motor output of the controllers with reference to attainment of their sensory goals.

The implemented action-phase controllers also generate predictions about the sensory consequences of the motor output, which are compared with the actual sensory signals (Figure 1a)). Contact events, which denote completion of action goals, represent critical sensorimotor control points in this respect because they give rise to discrete and distinct sensory signals in one or more modalities. By comparing predicted and actual sensory feedback, the brain can monitor the progression of the task and, if a mismatch occurs because of misleading initial state information or a faulty action-phase controller, launch corrective actions (Figure 1(a)). This use of sensory information as well as the characteristics of the corrective actions are highly task- and phase-specific and are presumably learned with the learning of the underlying action-phase controller. Moreover, mismatches can lead to an updating of the memory representations of object properties that can be subsequently used for predictive control later in the task and in other tasks with the same object, thus reducing future prediction errors.

Although we have emphasized that manipulation tasks can be divided into a sequence of action phases, it is important to note that sensory predictions generated in one action phase are used in the next. Specifically, sensory predictions about the terminal state of the motor system and environment of one action phase also provide initial state information for the subsequent action-phase controller. This information can be used to parameterize the next action phase in advance, which allows a smooth transition between the component actions of the task. In the absence of such anticipatory control, delays between phases would occur because the brain would have to rely on peripheral afferent signals to obtain this state information. Such a process would be time consuming due to long time delays in sensorimotor control loops associated with receptor transduction and encoding, neural conduction, central processing, and muscle activation. For example, it takes approximately 100 ms before signals from tactile sensors in the digits can bring about a significant change in fingertip actions. Even longer delays, in excess of 200 ms, are usually required to transform visual events into purposeful fingertip actions. Because of these long time delays in sensorimotor control loops operating on the hand, anticipatory control policies govern the swift and smooth transitions characteristic of dexterous manipulation.

Vision provides critical information for predictive control of task kinematics. In addition to locating objects in the environment, vision is used for identifying contact sites that are both stable and advantageous for various actions that we want to perform with the grasped object. For example, when we pick

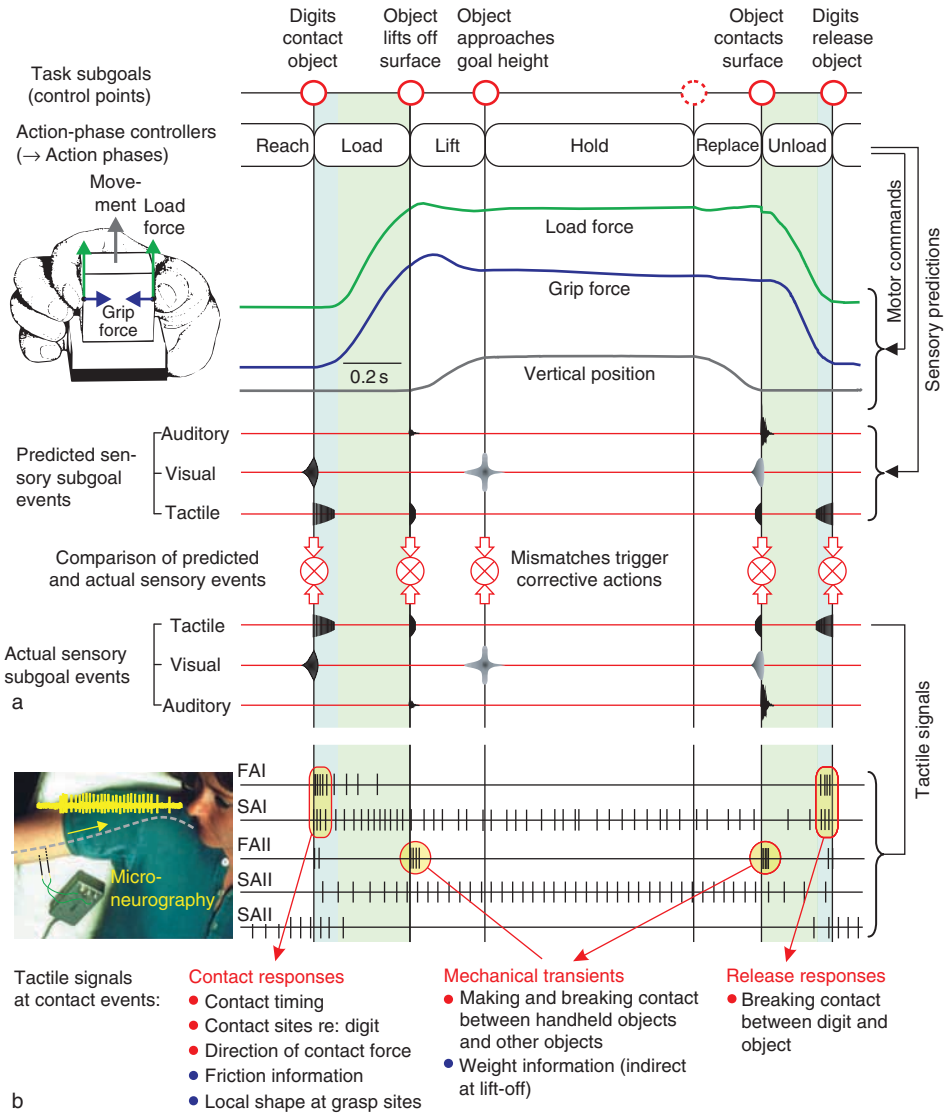

Figure 1 A person grasps and lifts an instrumented test object from a table, holds it in the air, and then replaces it, using the precision grip: (a) contact events and subgoals; (b) signals in the four types of tactile afferents from the glabrous skin of the human hand. In (a), the contact events shown at the top define the subgoals of the task (i.e., goals of each action phase). Sequentially implemented action-phase controllers generate motor commands that bring about the required action phases. After digits contact the object, the grip force increases in parallel with the tangential (load) force applied during isometric conditions to the object to overcome gravity. When the load force overcomes the force of gravity, the object lifts off. After the object is replaced such that it contacts the support surface, the load and grip forces decline in parallel until the object is released. In conjunction with generating motor command, the action-phase controllers predict internally the sensory consequences of these commands in one or more modalities (predicted sensory subgoal events), whereas the actual sensory consequences are signaled in one or more modalities (actual sensory subgoal events). For example, when the object is replaced on the surface, the contact event gives rise to tactile, visual, and auditory sensory events. By comparing predicted and actual sensory events, the sensorimotor system can monitor task progression and detect mismatches used to bring about corrective actions tailored to the action phase (see also **Figures 5, 6(b),** and **6(c)**). The predicted sensory events from each action phase can be used to provide initial state estimates for the next phase (not illustrated). In (b), signals in the four types of tactile afferents are recorded with the technique of microneurography during the lifting trial. At four points corresponding to subgoal events, tactile afferents show distinct burst discharges: (1) contact responses preferentially in FAI and SAI afferents when the object is first contacted, (2) burst responses in the FAII afferents related to the mechanical transients at lift-off and (3) when the object contacts the support surface, and (4) responses primarily in FAI afferents when the object is released. In addition to these event-related responses, SAII afferents and many SAI afferents show ongoing impulse activity when forces are applied to the object. Some spontaneously active SAII units are unloaded during the lift and cease firing. FA, fast-adapting; SA, slow-adapting. Data from Johansson RS and Westling G (1991) Afferent signals during manipulative tasks in man. In: Franzen O and Westman J (eds.) *Somatosensory Mechanisms*, pp. 25–48. London: Macmillan.

up a hammer to drive in a nail, we probably use different grasp sites than when we pick it up to give it to another person. Furthermore, visual cues about the identity, size, and shape of an object provide cues about its mechanical properties that are useful for predicting the forces required for successful manipulation. Visual cues related to object weight and mass distribution can be used to predict magnitudes of required fingertip forces and visual cues about the shape of grasp surfaces can be used to predict stable fingertip force directions. However, vision is of limited utility when objects are out of sight or partially occluded.

The use of vision in manipulation relies on learned associations between visual cues and their mechanical meaning. Such associations are grounded in movement–effect relationships evaluated through signals in sensors that transmit veridical information about mechanical interactions between the body and objects in the environment. The tactile modality directly provides information about mechanical interactions between our hands and objects and plays a pivotal role in the learning, planning, and control of dexterous object manipulation tasks. Relatively little is known about the contribution of proprioceptive signals in manipulation. Like vision, proprioception can provide only indirect information about mechanics. For example, signals related to muscle length, joint angle, and muscle force do not directly code the contact state between the hands and objects.

The importance of tactile afferent signals in manipulations is most apparent in people with impaired digital sensibility. Such individuals have great difficulty with routine tasks even under visual guidance. For example, they often drop objects, may easily crush fragile objects, and have tremendous difficulties with everyday activities such as buttoning a shirt or picking up a match. In humans, the density of tactile innervation is highest in body surface areas that typically contact objects including the palmar surfaces of the hands, the soles of the feet, and the tongue and lips. Approximately 2000 tactile afferents innervate each fingertip and some 10 000 afferent neurons innervate the remaining glabrous skin on the palmar surface of the hand. Microneurography studies in humans have shown that there are four types of tactile afferents that encode different aspects of the deformations of the soft tissues when the hands interact with objects (Figure 2). Overall, these sensors have evolved for extracting, rapidly and with high spatiotemporal fidelity, features of dynamic mechanical events that occur on top of the low-frequency and often large forces typically present when holding and transporting handheld objects. These dynamic tactile signals code the contact events that mark the attainment of subgoals in manipulation tasks (Figure 1(b)).

In the remainder of this article, we describe a series of experimental results from manipulation tasks that illustrate and further develop the general principles just described.

Predictions and Control Points in the Tactile Modality

The first action phase in most manipulation tasks is to bring the hand to the object in order to grasp it. The goal of this reach phase it to place the fingertips on the object in locations that will allow the development of a stable grasp in the context of the actions that will be performed with the object. In many manipulation tasks, it is important that the fingertips contact the object at around the same time and that the fingertip force vectors sum to zero. This is particularly important when grasping a light object that might otherwise be displaced or rotated when initially contacted. The contacts between the fingertips and object represent control points for goal completion of the reach phase. In some cases, these contact events may be visible and can be encoded visually. However, in many instances, the contact sites of at least some of the digits are out of view (as when digits contact the back side of an object). In contrast, tactile signals encode reliably and robustly all contact events (Figure 1(b)).

For each digit, contact responses in ensembles of afferents convey information about contact timing, the contact location on the digit, and the direction of contact force. Ensembles of tactile afferents also encode the frictional status of the contact and the shape of the contact surface, which are critical parameters in the control of object manipulation. All of these contact parameters are represented by signals in ensembles of tactile afferents that encode the patterns of stresses and strains distributed throughout the viscoelastic fingertip. Fast-adapting type I (FAI) afferents provide the most well-defined contact responses; however, slow-adapting type I (SAI) and fast-adapting type II (FAII) tactile afferents may also respond dynamically at the moment of contact. Tactile information associated with discrete mechanical fingertip events (including contact events) can lead to appropriate changes in motor output (i.e., fingertip forces) within 100 ms or less. Traditionally, it has been posited that tactile information is coded by the firing rates of tactile afferents. However, such a mechanism would be too slow to account for the rapid use of tactile information in manipulation. That is, to estimate firing rates at least two impulses in a given neuron are

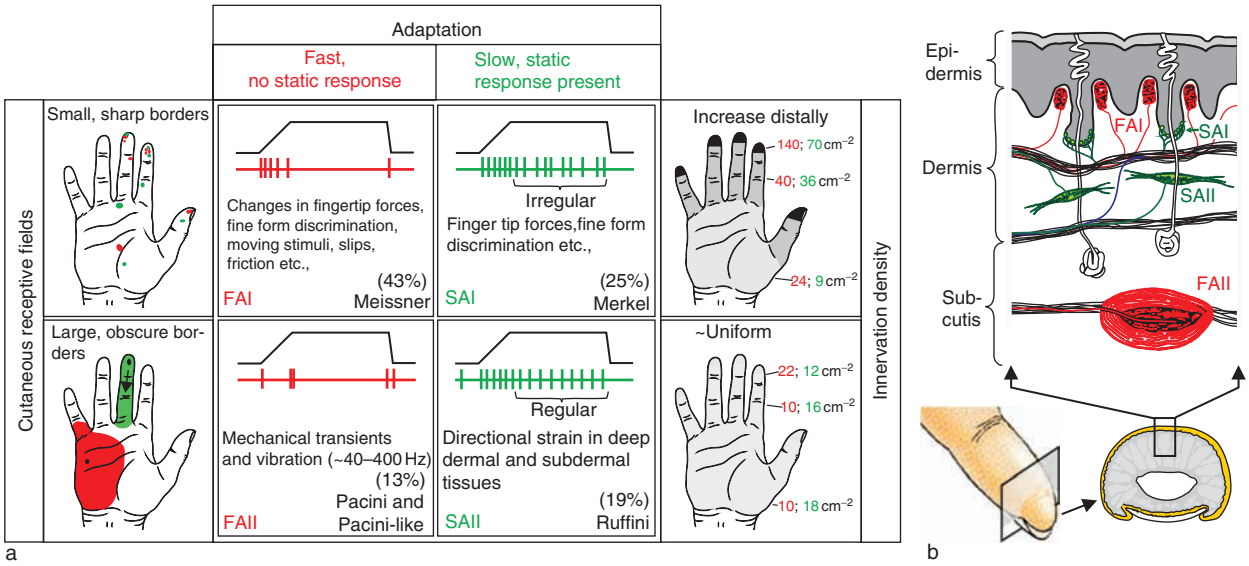

Figure 2 Types of tactile (mechanoreceptive) sensors in the glabrous skin of the inside of the human hand: (a) functional properties; (b) nerve endings in the fingertip. In (a), the middle graphs schematically show the impulse discharge (lower traces) to perpendicular ramp indentations of the skin (upper traces) for each sensor type. Two types (FA) show fast adaptation to maintained tissue deformation (i.e., they only respond to deformation changes). Two types adapt slowly (SA); that is, in addition to being dynamically sensitive (particularly the SAIs) they show an ongoing response related to the strength of maintained tissue deformation. The type I sensors (FAI and SAI) have small and well-defined coetaneous receptive fields (typically 10 mm^2) when defined by light, pointed stimuli (patches in the left drawing of the hand represent the fields of 15 different sensors). The density of these sensors increase in the distal direction of the hand and is especially high in the very tips of the digits (see the right drawing of the hand; the numbers indicate the innervation density as the number of afferents per square centimeter of skin area for the fingertips, fingers, and palm). In contrast, the FAII and SAII sensors show lower and approximately uniform densities over the hand and their receptive fields are larger and less well defined (left drawing of the hand). The FAIIs are especially responsive to mechanical transients that propagate though the tissues, whereas the SAIIs sense strain in the dermal and subdermal fibrous tissues, often with a remote origin. Sensors with encoding properties akin to the type II sensors of the hand are found in virtually all fibrous tissue in the body. The relative frequency of occurrence in the glabrous skin and the probable morphological correlate are indicated for each type of sensor. In (b), the organized nerve terminals corresponding to the four types of tactile sensors are shown for the fingertip. FA, fast-adapting; SA, slow-adapting. Data from Johansson RS and Vallbo ÅB (1983) Tactile sensory coding in the glabrous skin of the human hand. *Trends in Neuroscience* 6: 27–31.

required and reliable estimates generally require that neural responses are averaged over a substantial time window and over several trials. Recently it has been proposed that the relative timing of first spikes in ensembles of tactile afferents (temporal codes) encode tactile information for use in manipulation tasks. Neural recordings in humans indicate that such a mechanism would provide information about tactile events more rapidly than rate coding and fast enough to account for the use of tactile signals in natural manipulation.

Sensory Predictions That Support Grasp Stability

After contact is established between the object and the hand, practically all manipulation tasks require the application of forces tangential to the contacted surfaces (load forces). For example, to lift an object with the digits at the side, vertical load forces must be applied to overcome the weight of the object (Figure 1(a)). In many cases, twist forces (torques) tangential to the grasped surfaces are also applied. For example, if we lift a bar from one end, in addition to the vertical load force we need to apply tangential torque to prevent the bar from rotating as we lift it. These tangential loads destabilize the grasp, and to prevent the object from slipping (either linearly or rotationally) we need to apply forces normal to the surface (grip forces) to create stabilizing frictional forces.

When manipulating objects, grip forces are controlled to increase and decrease in phase with increases and decreases in the tangential loads with no time delay (Figures 1(a) and 3). This anticipatory modulation of grip force with movement-dependent changes in load force generates grip forces that are greater than the minimum required to prevent slip while at the same time not excessive. This coordination of grip and load forces is not innate but develops during ontogeny and is gradually refined until approximately 8 years of age. The coupling between grip force and load force supports grasp stability in a wide variety of maneuvers that we perform with objects. For example, it is observed at the level of

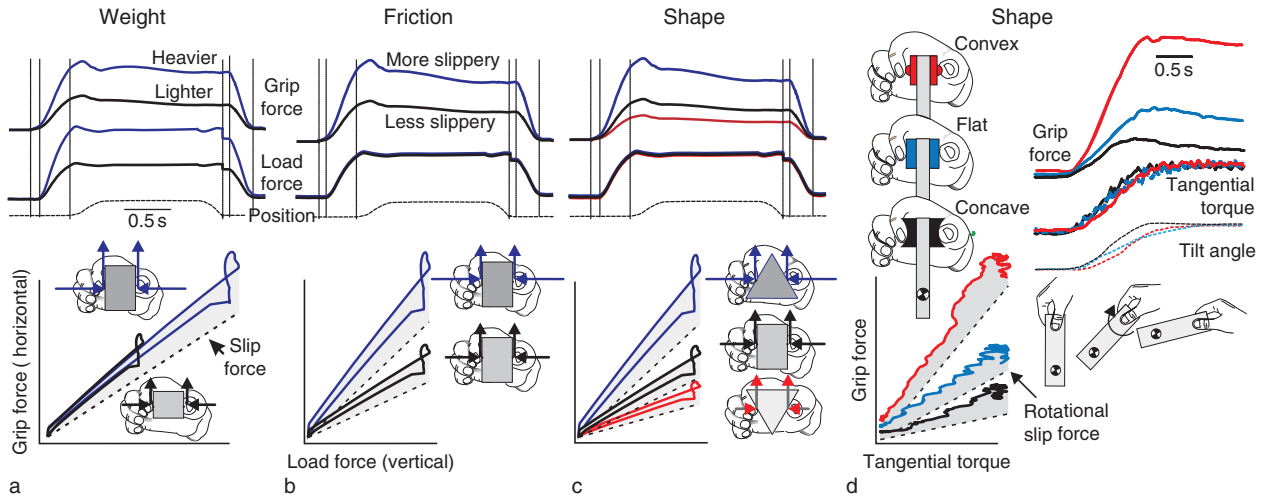

Figure 3 Parametric adjustments of motor output: (a) to change in object weight; (b) to change in friction between the object and skin; (c) to change in shape of the contact surface during object lifting; (d) to change in shape of the contact surface during object tilting change in tilt. In (a–c), the subject lifts an instrumented test object from a table, holds it in the air, and then replaces it, using a precision grip. Upper graphs show the horizontally oriented grip force, the vertically oriented load force (lift force), and the object's vertical position as a function of time for superimposed trials, indicated by differently colored curves. The lower graphs show the grip force as a function of the load for the same trials. The dashed lines indicate the minimum grip force to prevent slips and the safety margin against slips is indicated by hatching. In (a), with weight variations, the parallel change in grip and lift forces coordinated by the action-phase controllers ensures grasp stability when lifting objects of different weights. In (b), the balance between grip force and load force is adjusted to deal with changes in friction. In (c), a similar scaling of the grip-to-load force ratio is observed when object shape is varied. In either instance, the coordination of grip and load force ensures an adequate safety margin against slips. In (d), an already lifted object is tilted by 65° around the grip axis, which causes tangential torques at the grasp surfaces. Three superimposed trials are shown, one for each of the following surface curvatures: spherically convex with a radius of 5 mm, flat, and spherically concave with a radius of 20 mm. The top traces show the grip force, tangential torque, and tilt angle against time. Note the parallel change in grip force and the tangential load and the higher grip force with the greater surface convexity. The bottom graph emphasizes this adjustment by plotting the grip force against tangential torque load for the corresponding data. The shading indicates a similar safety margin against rotational slips irrespective of curvature. (a) Based on data from Johansson RS and Westling G (1988) Coordinated isometric muscle commands adequately and erroneously programmed for the weight during lifting task with precision grip. *Experimental Brain Research* 71: 59–71. (b) Data from Johansson RS and Westling G (1984) Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. *Experimental Brain Research* 56: 550–564. (c) Data from Jenmalm P and Johansson RS (1997) Visual and somatosensory information about object shape control manipulative fingertip forces. *Journal of Neuroscience* 17: 4486–4499. (d) Data from Goodwin AW, Jenmalm P, and Johansson RS (1998) Control of grip force when tilting objects: Effect of curvature of grasped surfaces and applied tangential torque. *Journal of Neuroscience* 18: 10724–10734.

individual digits when people use two or more digits of the same or both hands to manipulate objects, when grasping and moving objects held between the palms, and when load forces on a handheld object are generated by jumping up and down. Likewise, it applies to bimanual coordination when the hands have different and complementary roles, such as when we use one hand to add things into, or remove things from, a receptacle held by the other hand. Coupling between grip and load forces also operates when we move objects with different complex dynamics (combinations of inertial, elastic, and viscous loads), even though this involves altering the mapping between motor commands that generate load forces and those that generate grip forces.

The control of grasp stability requires that the balance between the grip and load forces be adapted to the properties of contacted surfaces. The friction between digit and object surface determines the minimum ratio between grip force and load force required to prevent slip, and people parametrically adapt the balance between grip and load forces for different frictional conditions, using greater grip forces with more slippery surfaces. This is achieved by varying the ratio between the grip force change and load force change (Figure 3(b)). In the same vein, people parametrically scale the balance between the grip and load forces to the shape of the contacted surface. For example, the greater the curvature of a spherically curved grasp surface, the larger the grip force required to generate a given tangential torque (Figure 3(d)). Similarly, when people lift tapered objects, a greater grip-to-load force ratio is required when the grip surfaces are tapered upward as compared to downward (Figure 3(c)). These parametric adaptations to contact surface friction and shape typically result in grip forces that exceed the minimum required to prevent slips by a safety margin of 10–40% of the applied grip force (gray areas in the force coordination plots of Figure 3).

When manipulating objects, people control their fingertip forces based on predictions of surface properties. Vision can, under favorable conditions, provide accurate information about object shape that can be used for predictive parameterization of fingertip forces. In contrast, it appears that vision is unhelpful in predicting the friction of an object, in part because friction depends not only on the object surface but also on sweating rate and the greasiness and wetness of the skin (Figure 4(a)). Instead, predictions of frictional conditions are based on the memory of previous haptic experiences with the same or similar objects, whereas a default prediction, perhaps based on some average, appears to be used in novel situations. The critical role of tactile afferents in the control of grasp stability is clearly illustrated by the fact that people with digital sensory impairments fail to adapt grip-to-load force ratios to object surface properties and, instead, consistently use large grip forces (cf. Figures 4(b) and 4(c)).

Information about surface friction and shape is rapidly extracted during initial contact between the fingertip and object surface by signals in ensembles of tactile afferents (Figure 4(d)). Thus, the initial contact event

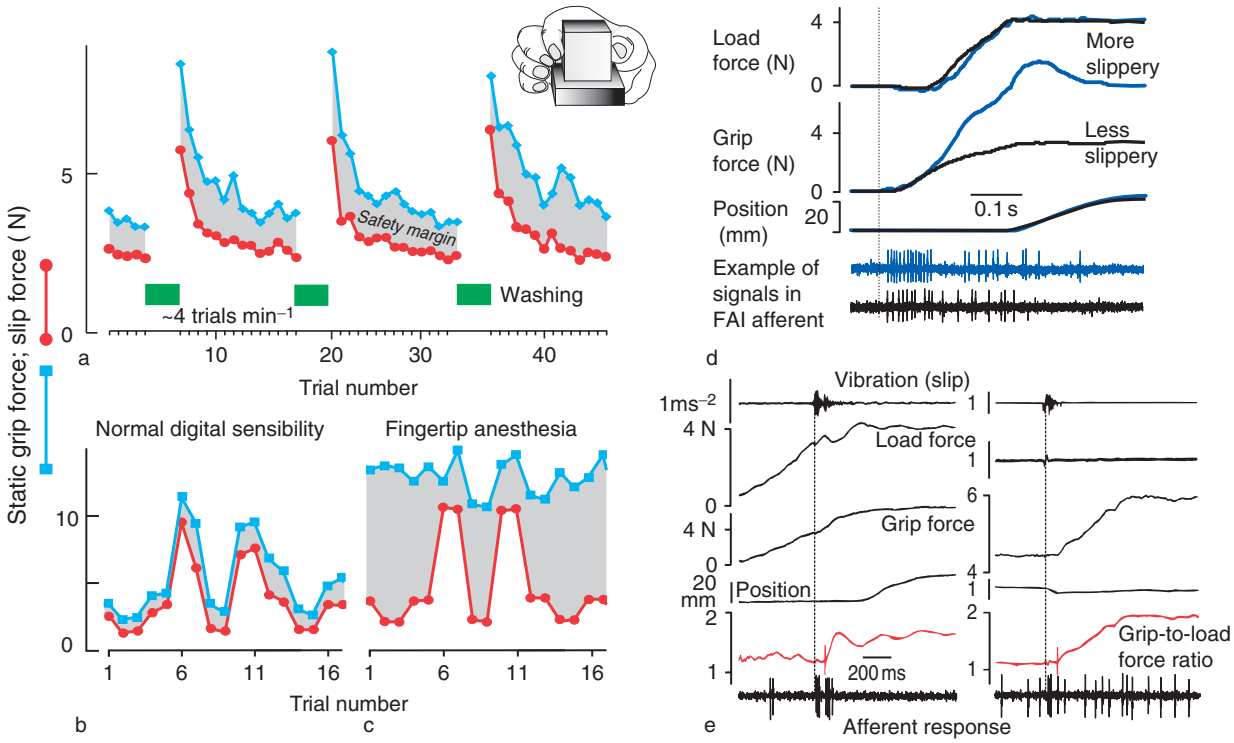

Figure 4 Adjustment of hand grip force to frictional changes in everyday situations: (a) static grip force before and after washing the hand; (b) with changes in object surface structure with normal digital sensibility; (c) with changes in object surface structure with anesthetized fingertips; (d) with friction; (e) in response to accidental slippage. In (a), the static grip force when holding an object in air using a precision grip (diamonds) and the corresponding minimum grip forces to prevent slips (slip force (dots)) are shown before and after the subject's hand was washed (with soap and water) and dried (towel); there were 48 consecutive lifts with suede as surface structure (single subject, weight constant at 400 g). In (b–c), the corresponding grip force data is shown during frictional changes between lifts caused by changing the surface structure among fine-grain sandpaper, suede, and smooth rayon. In (d), the influence of friction on force output and initial contact responses in a fast-adapting type I (FAI) afferent are shown. Two trials are superimposed, one with less slippery sandpaper (black lines) and a subsequent trial with more slippery silk (blue lines). The sandpaper trial was preceded by a trial with sandpaper, and therefore the force coordination is initially set for the higher friction. The vertical line indicates initial touch. In (e), examples are shown of afferent slip responses and the upgrading of the grip-to-load force ratio elicited by small slip events occurring at only one of the digits engaged. Vertical dotted lines indicate the onset of the slips as revealed by vibrations in the object (acceleration signal); short vertical lines indicate the onset of the slip-triggered upgrading of the force ratio. The new higher and stable ratio restores the safety margin, preventing future slips. On the left, slip during the load phase is shown; the middle burst is the afferent slip response, whereas the third burst represents dynamic responses to the force changes following the slips. On the right, slip during the static phase is shown. (a) Based on data from Johansson RS and Westling G (1984) Influences of cutaneous sensory input on the motor coordination during precision manipulation. In: von Euler C, Franzen O, Lindblom U, and Ottoson D (eds.) *Somatosensory Mechanisms,* pp. 249–260. London: Macmillan. (b–c) From Johansson RS and Westling G (1991) Afferent signals during manipulative tasks in man. In: Franzen O and Westman J (eds.) *Somatosensory Mechanisms,* pp. 25–48. London: Macmillan. (d–e) Based on data from Johansson RS and Westling G (1987) Signals in tactile afferents from the fingers eliciting adaptive motor responses during precision grip. *Experimental Brain Research* 66: 141–154.

represents a critical control point in the manipulation task at which predicted and actual sensory feedback related to surface properties can be compared. Mismatches between predicted and actual contact responses – primarily provided by ensembles of FAI afferents – lead to an updating of the grip–load force coordination approximately 100 ms after contact. Figure 5 illustrates such updating when people repeatedly lift objects with tapered grasp surfaces without vision of the object. The tapering was changed between trials in a pseudorandom order. In all trials, the initial increase in grip force was determined by the tapering (and hence force requirements) in the previous trial. This indicates that knowledge about the status of the object, obtained in previous trials, specifies the force coordination. When the tapering was changed, the grip force output is modified approximately 100 ms after the digits contacted the object and tuned for the actual object properties (Figure 5(a)). By the second trial after the change, the force coordination is appropriately adapted right from the onset of force application. Thus, when a prediction error occurs, tactile information obtained at initial contact with the object rapidly initiates a corrective action accompanied with an updating of the representation of the object used to control forces in future interactions with the object. As previously indicated, people with digital sensory impairments fail to adapt grip-to-load force ratios to object surface properties (cf. Figure 4(c)).

Under favorable conditions, visual cues about object shape can influence the initial contact forces used in lifting in a feed-forward manner. For example, in sequential lifting trials with tapered contact surfaces, a change from upward to downward tapering results in a slower increase in grip force right from the beginning of the force application (Figure 5(b)). Similarly, with a change from downward to upward tapering surfaces, the grip force is adjusted from the very start of the lift in anticipation of the higher grip force required to lift the object. Hence, visual cues about object geometry can provide state information for anticipatory parametric adaptation of component actions of the manipulation task. Once the object is contacted, tactile signals also provides state information about object shape that can override visual predictions if necessary.

Occasionally, the updating of frictional and shape representations that typically occurs at initial contact is

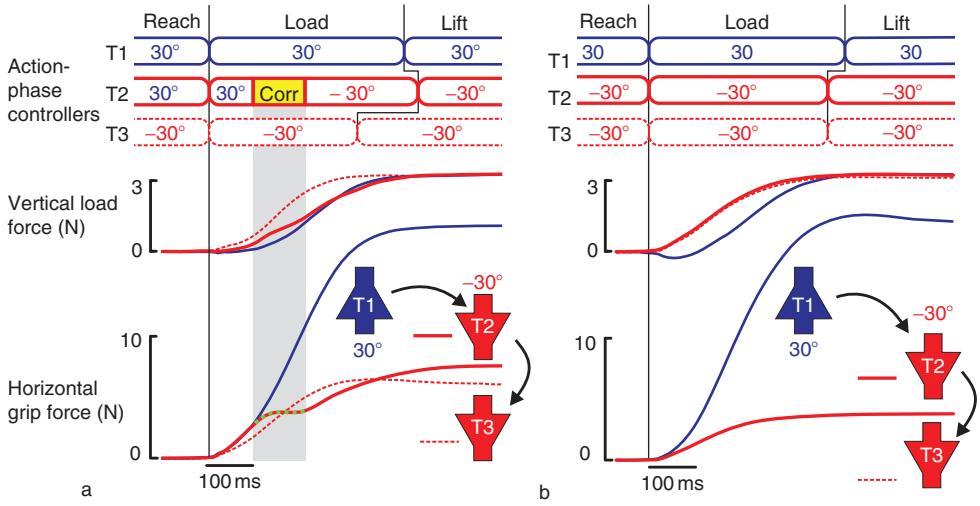

Figure 5 Adaptation of fingertip forces to changes in object shape; vertical load force and horizontal grip force shown as a function of time for trials from a lift series in which surface angle was unpredictably varied between lifts: (a) without vision of the contacted surfaces; (b) with vision of the contacted surfaces. The blue curves refer to the last trial with the 30° object before the switch (T1; upward tapered grasp surfaces). (This trial was preceded by a trial with a 30° object.) The solid red curves show the next trial (T2) performed with the –30° object (downward tapered grasp surfaces). These curves thus illustrate adjustments to the smaller angle. The yellow dashed segment in the grip force curve in (a) indicates the epoch of the corrective action elicited by the new surface angle. The thin red dashed curves show the following trial, again with the –30° object. The top diagram represents the status of the sequentially implemented action-phase controllers. In T1, they are parameterized for the 30° object throughout. In T2, without vision of the contacted surfaces (a), a corrective action (Corr) is triggered approximately 100 ms after contact based on a mismatch between predicted and actual tactile information obtained at contact related to object shape. This action is interspersed during the loading phase and involves a change in the ratio between the grip force change and load force change, which, in turn, change the balance between grip and load forces to better suit the –30° object for the remainder of the trial (see also Figure 3(c)). In T3, the controllers remain updated to the –30° object. With vision of the contacted surfaces (b), visual cues about object geometry provide state information for anticipatory parametric adaptation of the motor output after the change in tapering (T2), and no corrective action is triggered. Data from Jenmalm P and Johansson RS (1997) Visual and somatosensory information about object shape control manipulative fingertip forces. *Journal of Neuroscience* 17: 4486–4499.

inadequate and may result in an accidental slip later in the task. For example, when an object is held in air after being lifted, such a slip usually results in a transitory and partial unloading at one digit, and this increases the loads on the other digits engaged (Figure 4(e)). These transient load shifts are signaled by tactile afferents and trigger a corrective action that increases the grip-to-load force ratio at all digits for the remainder of the task. The nature of the corrective action triggered by slips depends on the phase of the task. Slips during the load phase in lifting lead to a slowing down of the increase in load force, whereas slips during the hold phase lead to an increase in grip force. Hence, different action-phase controllers are associated with different smart reflexes that support grasp stability and help achieve the current subgoal of the task.

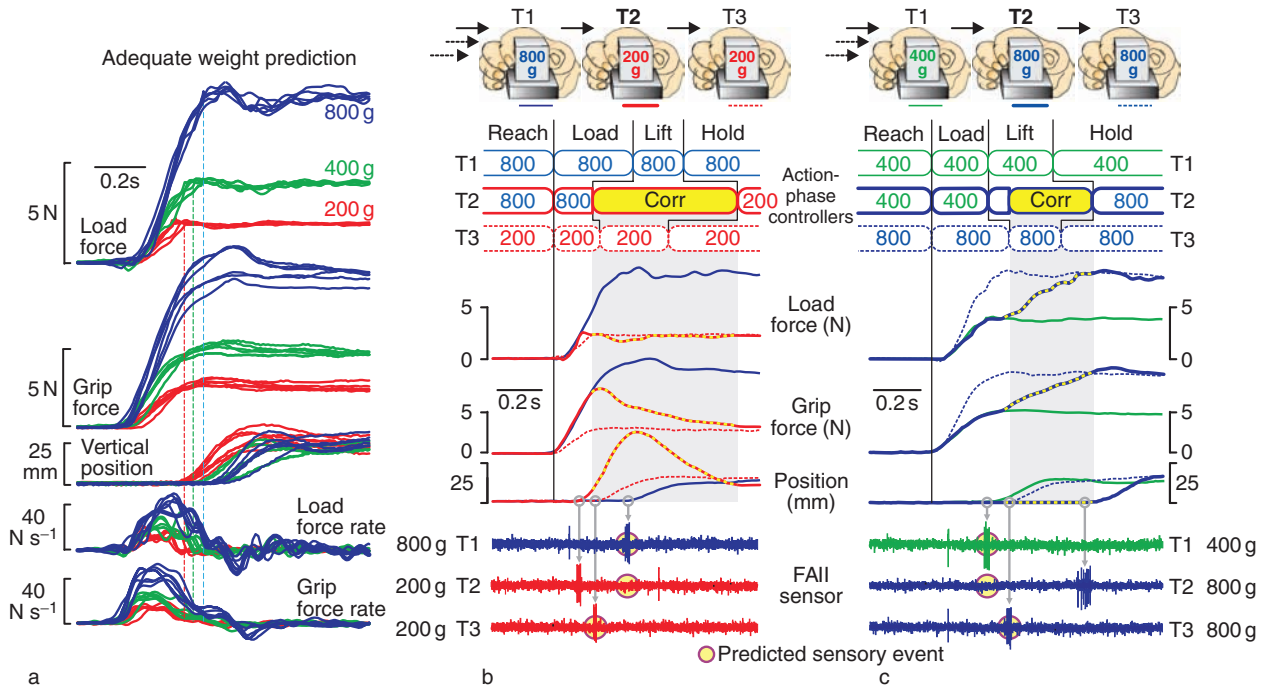

Figure 6 Adaptation of motor output to object weight: (a) fingertip forces and object movement during the initial part of adequately programmed lifts with three objects of different weights (data from 24 single trials superimposed; single subject); (b) single-unit tactile afferent responses and adjustments in force to unexpected decrease in object weight (trials between lifts from 800 to 200 g); (c) second experiment showing single-unit tactile afferent responses and adjustments in force to unexpected increase in object weight (trials from 400 to 800 g). In (b–c), the data from the single-lift trials are aligned on initial touch (vertical line). Gray circles and vertical lines indicate the moment of lift-off for each trial and the arrowheads point to the actual sensory events generated by the lift-off in a fast-adapting type II (FAII) afferent. The circles behind the nerve traces indicate the corresponding predicted sensory events. In (b), three successive trials (T1–T3) are shown in which the subject lifted an 800 g object (blue curves), a 200 g object (red solid curves), and then the 200 g object again (red dashed curves). The forces exerted in the first lift are adequately programmed because the subject had previously lifted the 800 g object. The forces are erroneously programmed in the first lift of the 200 g object (T2) because they are tailored for the heavier 800 g object lifted in the previous trial. The sensory information about the start of movement occurs earlier than expected for the erroneously programmed 200 g object trial (cf. the actual and predicted sensory events) and initiates a corrective action (yellow dashed red curves). The strong force drive of the ongoing load phase is terminated but still results in an overshoot in position due to the reflex delay, and corrective motor command is launched that brings the object back to the intended position. In (c), the participant performed an adequately programmed lift with a 400 g weight (T1, green curves), followed by a lift with an 800 g object erroneously programmed for the lighter 400 g weight lifted in the previous trial (T2, blue solid curves) and then a lift with the 800 g object again (T3, blue dashed curves). The absence of burst responses in FAII afferents at the predicted point for the erroneously programmed 800 g object trial (cf. the actual and predicted sensory events) elicits a corrective action (yellow dashed blue curves). This involves abortion of the lift-phase command followed by triggering of a second load-phase command that involves a slow, discontinuous, and parallel increase in grip and load forces until terminated by sensory input signaling lift-off. The top diagrams in (b) and (c) represent the status of the sequentially implemented action-phase controllers. In T1, they are parameterized for the 800 g (b) and 400 g (c) weight throughout. In T2, a corrective action (Corr) is triggered approximately 100 ms after the occurrence of the mismatch between predicted and actual sensory information related to object lift-off. This action involves abortion of the operation of the current action-phase controller and the implementation of action patterns that allow the task to continue. The corrective action is linked to an updating of the subsequently implemented controllers for the new weight. In T3, the controllers remain updated to this weight. Elements compiled from Johansson RS (1996) Sensory control of dexterous manipulation in humans. In Wing AM, Haggard P, and Flanagan JR (eds.) *Hand and Brain: The Neurophysiology and Psychology of Hand Movements*, pp. 381–414. San Diego: Academic Press.

Sensory Predictions Related to Object Motion

The successful lifting of an object requires that the forces developed during the load phase be predictively tuned to the weight of the object. People tend to lift objects of varying weight in about the same amount of time. To accomplish this, they scale the rate of increase of vertical load force, prior to lift off, to the expected weight of the object – increasing load force more rapidly for objects they expect to be heavy (Figure 6(a)). In addition, people predict the load force required for lift-off and, when lifting an object just off the surface, reduce the increase in load force rate so that it approaches zero at the expected lift-off time. When people lift to different heights – requiring different amount of object acceleration – load forces are likewise scaled predictively to achieve the desired acceleration and lift height. This ensures a smooth and critically damped lifting motion. Because no sensory information is available about weight until lift-off, this tuning of force output relies on estimates of weight. Such estimates are based on sensorimotor memory of the object derived from previous interactions with the object. Familiar objects can be identified visually (or by haptic exploration) for retrieval of weight estimates. Moreover, memory representations of familiar objects are organized into families or categories of objects in which the objects within a family can vary in size (e.g., cups, books, and loafs of bread). In such situations, people exploit size–weight associations when predicting the motor commands required to lift objects.

The fact that people rely on predictions related to object properties is clearly illustrated when people make erroneous predictions that result in pronounced performance errors. For example, when we lift up an object that is lighter than we expect, we typically make a jerky movement and lift the object higher than intended. A jerky movement occurs because the load phase is programmed for a heavier weight such that lift-off happens earlier than predicted and excessive lifting force is generated (Figure 6(b), T2). Conversely, an unexpected increase in weight results in a slow and hesitant load phase because the force drive is targeted for a lighter weight and additional increases in force are required to bring the object aloft (Figure 6(c), T2).

The expected lift-off represents a control point related to the goal of the load phase. Ensembles of FAII afferents, supplied by Pacinian corpuscles located in subcutaneous tissues, most quickly and reliably signal the outcome of the load phase by encoding both the timing and the dynamics of the lift-off event. Such afferents, which are distributed throughout fibrous tissues of the body, are particularly sensitive to transient mechanical events that propagate through handheld objects into the hand and arm. The transient events take place when contacts between handheld objects and surfaces are formed or broken and effectively excite these highly sensitive afferents. When lifting an object that is lighter than expected, the sensory events elicited by the lift-off occur before the predicted events that are part of the sensory plan (cf. predicted and actual neural events in Figure 6(b)). The mismatch triggers a learned corrective action that involves termination of the load phase force followed by corrective motor commands that bring the object back to the intended position. However, due to delays in sensorimotor control loops, this corrective action pattern kicks in too late to avoid an overshoot in the lifting movement (see position signal in Figure 6(b)). When lifting an object that is heavier than expected, the sensory event elicited by lift-off occurs neither before nor at the point predicted by the sensory plan (cf. predicted and actual neural events in Figure 6(c)). Hence, again there is a mismatch between actual and predicted sensory events. In this case, the corrective action triggered by the mismatch results in slow, probing increases in finger forces until terminated reactively by sensory events signaling lift-off. Thus, the sensorimotor system reacts to both the presence of an unpredicted event and the absence of a predicted sensory event. Significantly, in addition to triggering corrective actions, these sensory mismatches lead to an updating of memory representations related to object weight, which in turn improves predictive control in subsequent action phases and tasks that engage the same object. In natural situations, this updating generally occurs in a single trial (see Figures 6(b) and 6(c)). However, in the presence of misleading or confounding visual and haptic cues, or when people deal with objects with unusual dynamics (relating applied force to motion), updating may require repeated lifts or movements of the object.

When people transport and hold an object, knowledge about object weight, mass distribution, and surface properties remains critical for controlling action and maintaining grasp stability. When people replace the object on a surface, the sensory goal is to produce an afferent signature signifying contact. This contact event, which represents a sensorimotor control point, is signaled by FAII afferents that encode the timing and nature of the event (Figure 1(b)). The contact event is followed by an unloading phase where grip and load forces decrease in parallel. The sensory events related to the digits breaking contact with the surface of the object represent the goal of this phase. The coupled decrease in the load and grip forces ensures the maintenance of a grip-to-load force ratio that provides grasp stability even in this phase. The outcome of the unloading phase is distinctly signaled by release

responses in ensembles of tactile afferents, especially FAI afferents (Figure 1(b)).

Predictions and Control Points in the Visual Modality

Because visual resolution is far better in and around the fovea, we use saccadic eye movements (fast jumps in the line of sight) to shift gaze to targets of interest. In addition to providing high-resolution visual information about target features, gaze fixations provide information about target location in space based on knowledge about the orientation of the eyes with reference to the head and to other body segments as well. Different types of visually supported tasks have characteristic patterns of eye movements that accompany them. In object manipulation, people direct their gaze to successive contact locations as they gain salience in the evolving task according to the demands of the implemented action plan. For example, when people pick up a bar, move the bar in hand to contact a target switch, and then replace the bar, gaze is successively directed to the grasp site on the bar, the target, and the landing surface where the bar is replaced (Figures 7(a) and 7(b)). Furthermore, people may direct fixations to points where contact must be avoided, including obstacles that must be circumnavigated with the hand (or object in hand). In familiar manipulation tasks, subjects almost never fixate their hand or objects being moved by the hand. An important role of foveal vision is to monitor contact events representing subgoals of the task.

The spatiotemporal coordination of gaze and hand movements emphasizes the segmentation of manipulation tasks into distinct action phases (Figure 7(c)). At the start of most action phases, congruent hand and eye movements are launched concurrently to the contact location representing the goal. Thus, both hand and eye movements are specified based on peripheral vision about the contact location (or on memorized landmark locations). Because eye movements are quick, gaze reaches the contact location well before the hand (or object in hand). This permits the use of foveal and parafoveal vision in monitoring the final approach of the hand to the contact location. Gaze typically remains at the contact location until around the time of goal completion (e.g., until the grasp is established, the target switch is released, or the bar is replaced in the target contact task) or remains at a location where a contact should be avoided until the time of the potential contact (e.g., when the tip of the bar passes closest to the obstacle in the target contact task). Thus, the gaze shift to the contact location associated with the next action phase occurs around

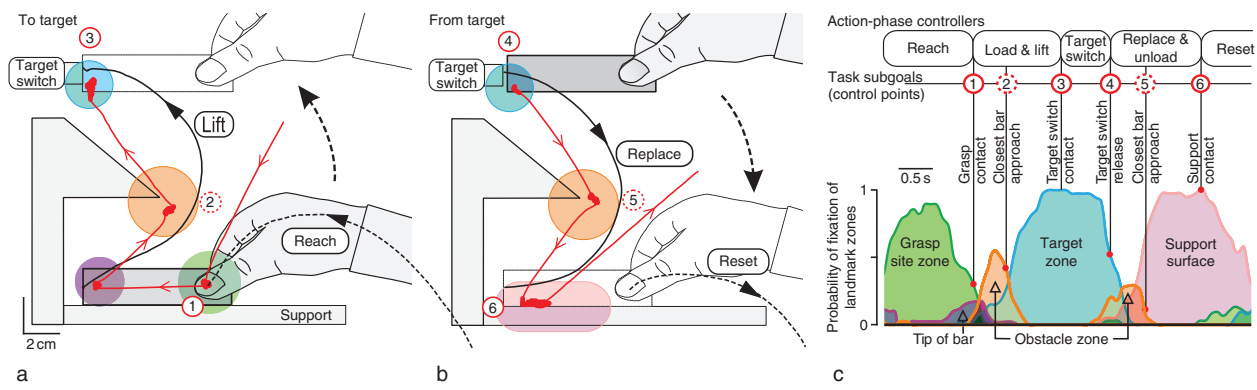

Figure 7 Predictions and control points in the visual modality, gaze, and hand movements for a target contact task: (a) a single trial in which the participant reached for and grasped a bar and moved it to press a target switch; (b) and then moved it from target contact and replaced it on the support surface; (c) spatiotemporal coordination of gaze and manipulatory actions shown as time-varying instantaneous probability of fixations within the landmark zones indicated in (a) and (b). In the task in (a–b), a triangular obstacle was located between the bar and target. Dashed black lines in (a–b) represent the path of the tip of the index finger during the reach for the bar and during the reset phase when the hand was transported away from the bar after it was replaced and released. The solid black lines represent the path of the tip of the bar. The red lines indicate the position of gaze; the thin segments with the arrowheads represent saccadic eye movements and the thick patches represent gaze fixations. The colored zones represent landmark zones that captured 90% of the fixations recorded during several trials by 10 participants; these zones are centered on the grasp site (green), tip of the bar (purple), protruding part of the obstacle (orange), target (blue), and support surface (pink). The plot in (c) is derived from data pooled across 10 participants, each performing four trials with the triangular obstacle. The red circles and vertical lines mark contact events demarcating phase transitions in the task, and the spatial locations of these events are schematically indicated by the location of the corresponding numbered circles in (a) and (b). The common time base has been normalized such that each phase of each trial has been scaled to the median duration of that phase. Data from Johansson RS, Westling G, Bäckström A, and Flanagan JR (2001) Eye-hand coordination during learning of a novel visuomotor task. *Journal of Neuroscience* 21: 6917–6932.

the predicted time of goal completion. In fact, in most cases, both gaze and hand movement commands are initiated in anticipation of goal completion and are not delayed until sensory feedback is obtained verifying goal completion. With the latter strategy, a smooth transition between successive phases of the manipulation task would not be possible because of the substantial time delays in sensorimotor control loops. In summary, contact events that demarcate action phases can be predicted and monitored in both the tactile modality and the visual modality.

Predictions in Action Observation

Although there is no question that tactile feedback related to control points is essential for skilled object manipulation, there are control points that do not give rise to tactile events. For example, when we drop a ball onto a surface, we typically direct our gaze to the predicted contact point between the ball and surface. Here, sensory feedback related to contact is available only through vision (and possibly audition). Thus, vision and the tactile system can play complementary roles in monitoring contact events. We have recently argued that, when observing an actor perform an object manipulation task, the observers – like the actor – predict and monitor contact events linked to the subgoals of the task. Although the observers cannot use tactile feedback to monitor contact events, they are free to move their gaze to monitor contact events via vision.

When people observe an actor performing a predictable block-stacking task, their gaze behavior is similar to the actor's gaze behavior. Specifically, the gaze of both the actor and observer predicts forthcoming contact sites (where blocks are grasped and replaced), and gaze is maintained at each contact site until around the time of goal completion (grasp contact and block landing). Even when observers do not know, in advance, which of two possible blocks the actor will grasp, they shift their gaze proactively to the correct target as soon as possible based on vision information. Taken together, these results indicate that observers implement a dynamic sensory representation of the task they are observing. Furthermore, these findings provide support for the hypothesis that observing and understanding the actions of another person involves a mechanism that maps the observed action on to sensorimotor representations in the observers' brain. In action observation, as in action, we argue that gaze is controlled, in part, for monitoring the consequences of action in central vision. By comparing actual and predicted visual feedback related to contact events, both observers and actors may be able to obtain valuable information about the outcomes of action and action plans that can be exploited by the sensorimotor system when learning, planning, and controlling future actions.

Sensory Control Policies during Learning

Although skilled behaviors involve learning to execute motor commands that fulfill sensory goals, little is known about possible changes in sensory control policies across learning stages. We have recently studied the learning of a novel visually guided object manipulation task in which participants had to discover how to move a cursor to successive targets presented on a screen by applying forces and torques to a rigid tool held between the two hands. As when people discover how to ride a bicycle, the learning proceeds in three stages. In the initial exploratory stage, participants attempt to discover the basic mapping rule relating motor and sensory signals by busily generating apparently uncontrolled cursor movements. In the skill acquisition stage that follows, control begins to emerge and performance improves rapidly. In the final skill refinement stage, performance continues to improve but gradually. During the exploratory stage, gaze reactively pursues the cursor via saccades directed to successive cursor locations, possibly allowing the learner to build a map between hand actions and gaze-related signals. As the learner starts to gain control over the cursor during the subsequent skill acquisition stage, gaze begins to predict forthcoming cursor positions. In the subsequent skill refinement phase, learners launch spatially and temporally congruent gaze and cursor movements directed to the target as in well-learned natural manipulations. These findings indicate that people adopt different sensorimotor schemas depending on the stage of learning and that learning to predict sensory consequences of motor commands precedes accurate control. It is not yet known how the use of other sensory modalities in manipulation may also depend on learning stage.

Conclusion

Dexterity in object manipulation depends on anticipatory control policies that rely on knowledge about movement–effect relationships when interacting with environmental objects. A sensory plan specifies the sequence of task subgoals in terms of sensory signals linked to contact events. This sensory representation of the task provides a scaffold for the selection and shaping of the action-phase controllers required for achieving sensory subgoals. Prediction of sensory states pertaining to the outcome of executed action phases

allows for smooth phase transitions, in contrast to the stuttering transitions that would result if the brain relied on peripheral afferent information to confirm subgoal completion and to update state information before launching the next action phase. Tactile sensors signal information about both the timing and the physical nature of the discrete mechanical events that represent the outcomes of the action phases. Contact events in manipulation are also represented in the visual and auditory modalities and, presumably, the proprioceptive modality. When the brain successfully predicts the actual sensory signals at a given contact event, no corrective action is required and the task runs in a feed-forward fashion. When a prediction error arises, a learned corrective action, the nature of which depends on the task and its action phase, is generated and representations of object properties are updated. We suggest that contact events that represent sensorimotor control points are encoded in multiple sensory modalities and serve at least two critical functions. First, by comparing actual and predicted sensory events in multiple sensory modalities, the sensorimotor system can simultaneously monitor multiple aspects of task performance and, if prediction errors arise, respond to the pattern of errors observed in different modalities. Second, because contact events give rise to salient sensory signals from multiple modalities that are linked in time and space, they provide an opportunity for sensorimotor integration and intermodal alignment helpful for learning and upholding multimodal sensorimotor correlations that support prediction of purposeful motor commands.

See also: Cross-Modal Interactions Between Vision and Touch; Finger Movements: Control; Human Haptics; Posterior Parietal Cortex and Tool Usage and Hand Shape; Reaching and Grasping; Sensorimotor Integration: Models; Spatial Transformations for Eye–Hand Coordination; Tactile Coding in Peripheral Neural Populations; Tactile Texture.

Further Reading

Birznieks I, Jenmalm P, Goodwin AW, and Johansson RS (2001) Encoding of direction of fingertip forces by human tactile afferents. *Journal of Neuroscience* 21: 8222–8237.

Flanagan JR, Bowman MC, and Johansson RS (2006) Control strategies in object manipulation tasks. *Current Opinion in Neurobiology* 16: 650–659.

Flanagan JR and Johansson RS (2003) Action plans used in action observation. *Nature* 424: 769–771.

Goodwin AW, Jenmalm P, and Johansson RS (1998) Control of grip force when tilting objects: Effect of curvature of grasped surfaces and applied tangential torque. *Journal of Neuroscience* 18: 10724–10734.

Jenmalm P and Johansson RS (1997) Visual and somatosensory information about object shape control manipulative fingertip forces. *Journal of Neuroscience* 17: 4486–4499.

Johansson RS (1996) Sensory control of dexterous manipulation in humans. In: Wing AM, Haggard P, and Flanagan JR (eds.) *Hand and Brain: The Neurophysiology and Psychology of Hand Movements*, pp. 381–414. San Diego: Academic Press.

Johansson RS and Birznieks I (2004) First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nature Neuroscience* 7: 170–177.

Johansson RS and Westling G (1984) Influences of cutaneous sensory input on the motor coordination during precision manipulation. In: von Euler C, Franzen O, Lindblom U, and Ottoson D (eds.) *Somatosensory Mechanisms,* pp. 249–260. London: Macmillan.

Johansson RS and Westling G (1984) Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. *Experimental Brain Research* 56: 550–564.

Johansson RS and Westling G (1987) Signals in tactile afferents from the fingers eliciting adaptive motor responses during precision grip. *Experimental Brain Research* 66: 141–154.

Johansson RS and Westling G (1988) Coordinated isometric muscle commands adequately and erroneously programmed for the weight during lifting task with precision grip. *Experimental Brain Research* 71: 59–71.

Johansson RS and Westling G (1991) Afferent signals during manipulative tasks in man. In: Franzen O and Westman J (eds.) Somatosensory Mechanisms, pp. 25–48. London: Macmillan.

Johansson RS, Westling G, Bäckström A, and Flanagan JR (2001) Eye-hand coordination in object manipulation. *Journal of Neuroscience* 21: 6917–6932.

Johansson RS and Vallbo ÅB (1983) Tactile sensory coding in the glabrous skin of the human hand *Trends in Neuroscience* 6:27–31.

Land MF (2006) Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research* 25: 296–324.

Sailer U, Flanagan JR, and Johansson RS (2005) Eye–hand coordination during learning of a novel visuomotor task. *Journal of Neuroscience* 25: 8833–8842.

Wing AM, Haggard P, and Flanagan JR (eds.) (1996) *Hand and Brain: Neurophysiology and Psychology of Hand Movement*. San Diego, CA: Academic Press.

Wolpert DM and Flanagan JR (2001) Motor prediction. *Current Biology* 11: R729–R732.

Sensorimotor Integration: Attention and the Premotor Theory

T Moore, B Noudoost, and K M Armstrong, Stanford University School of Medicine, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

The Neural Basis of Attention before the Neuron Doctrine

Our ability to selectively process important sensory information at the expense of extraneous or potentially distracting information is afforded us by the process of attention. Without attention, as William James (*c*. 1890) stated, "experience is an utter chaos." In the visual domain, attention allows us to sequentially process individual items among the multitude stimulating the retina, thus permitting us to, for example, read this text. Much work, and indeed much progress, has been made in recent years in establishing a neurobiological basis of visual attention. However, a significant fraction of the experimental groundwork was laid during the late nineteenth century. In addition to William James, the American psychologist who fractionated attention into two types, voluntary and involuntary, other noteworthy thinkers and experimentalists set the stage for what has only recently been explored further. Having carried out an extensive set of cortical lesions studies, primarily in dogs and monkeys, David Ferrier (*c*. 1876) concluded that the faculty of attention was "intimately related to the volitional movements of the head and eyes." Ferrier had observed that large lesions of the prefrontal cortex of monkeys resulted not only in paralysis of head and eye movements but also in deficits that would today be best described as 'neglect' – that is, a tendency to ignore sensory events in the affected half of space. Although Ferrier understood that visual attention did not require eye or head movements, this fact was not clearly established experimentally until Hermann von Helmholtz demonstrated it in 1894. Helmholtz set up a screen on which letters were scattered systematically and fixated the center of the display. Although there were too many letters, and the display was too large to view without moving the eyes, Helmholtz found that he could nonetheless selectively perceive different portions of the display while fixating. Using an electric spark, Helmholtz could briefly illuminate the entirety of an otherwise darkened display while still holding his gaze at its center. He found that he could choose a portion of the display to attend to while still fixating centrally, and that by doing this he could identify letters in that portion during the brief period of illumination, whereas he could not identify letters in other portions. Helmholtz thus observed experimentally that attention can be deployed covertly – that is, in the absence of eye movements. Helmholtz concluded,

> Our attention is quite independent of the position and accommodation of the eyes, and of any known alteration in these organs; and free to direct itself by a conscious and voluntary effort upon any selected portion of a dark and undifferentiated field of view. This is one of the most important observations for a future theory of attention.

As discussed later, the ability of primates, most notably humans and monkeys, to covertly attend has facilitated the study of the neural basis of visual attention. However, Ferrier's view regarding independence of the focus of attention from the direction of gaze during covert attention was more complicated and suggested a possible mechanistic basis for Helmholtz's later observation: "During (attention) we suppress actual movements, but keep up in a state of greater or less tension the centres of the movement or movements with which the various sensory factors of ideation cohere."

Ferrier thus argued that the apparent independence of movements from the direction of visual attention resulted from an active suppressing of it. Interestingly, noted physiologist, psychologist, and color theory nemesis of Helmholtz, Ewald Hering (*c*. 1894), further suggested that

> whilst attending to the marginal object we must always attend at the same time to the object directly fixated. If even for a single instant we let the latter slip out of our mind, our eye moves towards the former.

Thus, even before the neuron doctrine was established in the late nineteenth century, testable hypotheses regarding one of the brain's most complex mental functions had already been posed. In the many years to follow, much of the progress in pinpointing the neural mechanism of visual attention and in understanding its relationship with motor control came from neuropsychological studies of human subjects and monkeys with brain damage.

Parietal and Frontal Neglect

Neglect, or the failure in spatially directing visual attention to a region of space, is typically found in patients with unilateral lesions of parietal or frontal cortical regions. Typically, human patients with neglect omit elements on their contralateral side when copying simple drawings. They also show a bias

toward the intact side when asked to bisect a horizontal line. It is generally believed that the more 'perceptual' deficits tend to be associated with parietal lesions, whereas the more 'motor' tasks tend to be associated with frontal lesions. However, lesions of parietal cortex also lead to impairments on motor tasks. Gross inaccuracy in reaching for objects and marked clumsiness in picking up objects are also common following posterior parietal damage in both human patients and monkeys. Parietal lobe lesions also disrupt oculomotor behavior in monkey and man. Monkeys with parietal lesions are unable to track quickly moving visual targets with eye movements. In addition, the slow phase of vertical optokinetic nystagmus is disrupted following bilateral parietal lesions. Cooling of parietal area 7 causes a slowing of eye movements toward the contralateral side. Indeed, there is clear involvement of both parietal and frontal cortices in both perceptual and motor tasks, and this suggests a strong coupling of attention and motor programming. In the framework of a 'premotor theory of attention,' neglect has been interpreted as an imbalance of spatial representations that control motor programs, thus producing an attentional deficit.

Consistent with the initial observations of David Ferrier, subsequent lesion studies involving monkeys carried out by numerous investigators during the past century have repeatedly implicated the frontal eye field (FEF) in the control of visual spatial attention. Modern neurophysiological studies have established that FEF neurons play a key role in the programming and triggering of saccadic eye movements. However, studies both new and old indicate that the FEF is crucial to attention as well. Similar to parietal lesions, lesions of the FEF in monkeys have long been known to result in marked neglect. Like Ferrier, Bianchi (*c.* 1895) observed clear perceptual defects after excision of the frontal lobe in monkeys and dogs. Later, Kennard (1939) found that visual impairments following restricted lesions of the frontal lobe were indistinguishable from a hemianopia following occipital lobectomy. Other frontal lesions did not result in neglect. Subsequently, Clark and Lashley (1947), Denny-Brown (1951), Welch and Stuteville (1958), and Brucher (1966) all observed contralateral neglect after unilateral FEF lesions. Mapping visual fields before and after FEF lesions, Latto and Cowey were the first to quantify visual field deficits by training monkeys to make manual responses in response to a flash of light presented at different locations throughout the visual field. They found that unilateral FEF removal transiently impaired performance in the contralateral visual field. The animals were not blind, however, because they could still respond correctly on a fraction of trials. Yet, monkeys were far more likely to fail to respond to stimuli presented in the affected hemifield than elsewhere. Latto and Cowey suggested that the FEF might normally modulate the visual system according to eye position, and that this could explain the fact that such small lesions were able to produce such pronounced deficits.

More recently, studies have used reversible pharmacological inactivation to temporarily disrupt neural activity in a number of structures with known roles in the control of saccades, including the FEF, the superior colliculus (SC), and the lateral interparietal area. Generally, inactivation of these structures increases the time it takes an animal to initiate saccades to visual targets appearing in the affected part of space. In addition, the inactivation also impairs performance on attention-demanding tasks in the affected part of space when other distracting stimuli are present. Attention performance on tasks with isolated stimuli is not affected. Thus, there is good evidence that brain structures involved in controlling visually guided saccades also play a causal role in directing visuospatial attention.

Neurophysiology of Covert and Overt Visual Attention

It is well established that the responses of visual cortical neurons to covertly attended stimuli are enhanced above that of unattended stimuli. In a typical experiment, a monkey is trained to attend to peripheral stimuli appearing at a particular location or to a particular stimulus, regardless of location, while maintaining central fixation. The responses of extrastriate cortex neurons to physically identical stimuli presented to their receptive fields (RFs) are then compared between the different conditions. During covert spatial attention, responses to attended RF stimuli are enhanced, and this enhancement has been found to result in changes in orientation tuning, direction of motion tuning, contrast sensitivity, color tuning, and the representations of multiple, competing RF stimuli. Thus, the perceptual benefits of covert attention appear to be supported by neuronal sensitivity changes within visual cortex. However, the mechanism driving these correlates of attention is unknown.

Before it had been established that covert attention effects were widespread within visual cortex, it had been demonstrated that visual-evoked neuronal responses in a number of structures are enhanced when an animal targets an RF stimulus with a saccadic eye movement. This effect was first observed by Wurtz and Goldberg among neurons within the superficial layers of the SC and subsequently by Mountcastle and colleagues in posterior parietal cortex. Later studies demonstrated that the presaccadic

visual enhancement is also observed among neurons in area V4 and in inferior temporal cortex. Although there is a striking parallel apparent between the presaccadic enhancement of visual responses throughout the visuo-oculomotor axis and the attention-related modulation in visual cortex, a causal connection between the two was difficult to draw for many years since attempts to find covert attention effects in the SC and in the FEF were unsuccessful. However, recent studies have indeed found modulation of SC and FEF neuronal activity during covert attention. Thus, there is now good evidence that visual responses, whether exhibited by neurons more closely related to saccades or more closely related to coding the parameters of visual stimuli, are modulated by both covert and overt attention.

Testing a 'Premotor Theory' of Attention

A number of psychophysical studies have found evidence linking covert attention and the preparation of saccades. Generally, this evidence suggests one or both of two possibilities: that visual spatial attention influences saccade preparation and that saccade preparation influences the deployment of visual spatial attention. The most classic demonstration of the influence of directed attention on saccades was provided by Rizzolatti and colleagues. In their study, the authors examined how covert attention perturbed saccade trajectories. Subjects were instructed to initiate saccades to a location in one half of the visual field according to cues presented in the other half. The cues could be presented in one of several locations in the cued half of the visual field. The authors found that saccade trajectories were systematically deviated according to the location of the covertly attended (i.e., cued) location. Thus, the deployment of covert attention can perturb saccade programming. The previous observation as well as similar findings from other investigators are frequently cited as evidence for what is commonly referred to as the premotor theory of attention, a more modern incarnation of the ideas of Ferrier and others.

To address the role of saccade preparation in the deployment of attention, Hoffman and Subramaniam had subjects make saccades to a location while also detecting visual targets presented prior to the eye movement. These authors found that target detection was typically best at the location of planned saccades, and that subjects were unable to completely dissociate the attended and saccade locations. These results suggest that the preparation of saccades to a location deploys attention to that location. Deubel and Schneider obtained similar results and concluded that a single mechanism must drive the selection of objects for perceptual processing and the information needed to drive the appropriate saccadic response.

Although these psychophysical studies provide a sound rationale for the hypothesis that oculomotor and attentional circuits have a common neural substrate, they cannot rule out the alternative view that these are two distinct and independent systems. In recent years, neurophysiological studies have taken a more direct approach to this problem by perturbing neural signals within oculomotor structures and examining the effects of those perturbations on spatial attention. These structures include the FEF and the SC, both having a known role in the programming and triggering of saccades. The approach of directly manipulating signals in these areas addresses the causal relationship between neural activity in oculomotor circuits and shifts of spatial attention.

The results from recording and lesions studies suggest that the locus of spatial attention is driven by neural activity from saccade-related structures. If this is the case, then it should be possible to direct attention by perturbing neural activity within these structures. Studies have used electrical microstimulation to test the hypothesis that saccade-related circuits play a causal role in driving attention to particular locations in the visual field. Although it had long been known that microstimulation of the FEF and the SC can elicit saccades, Moore and Fallah were the first to examine the effect of microstimulation on visual attention (Figure 1). They found that when they stimulated sites within the FEF using currents that were too low to evoke a saccade, they could nonetheless enhance monkeys' ability to detect changes in a visual target. Importantly, the improvements observed with microstimulation were dependent on the target of attention being positioned at the location to which suprathreshold microstimulation would shift the monkey's gaze. Thus, it appeared that by increasing the likelihood that the monkey would foveate a location in visual space, the experiments also enhanced the animal's ability to process visual events there. Subsequently, two other studies reported similar enhancements in visual spatial attention following subthreshold microstimulation of the SC. In both cases, the performance-enhancing effects of microstimulation were spatially dependent as in the FEF studies.

The effect of electrical microstimulation of the FEF and the SC on visual spatial attention performance in monkeys suggests that during stimulation, visual representations of particular stimuli are enhanced above those of others. As mentioned previously, a wealth of evidence has established that correlates of visual attention can be observed throughout much of macaque visual cortex; visual responses to RF stimuli are typically enhanced when a monkey is trained

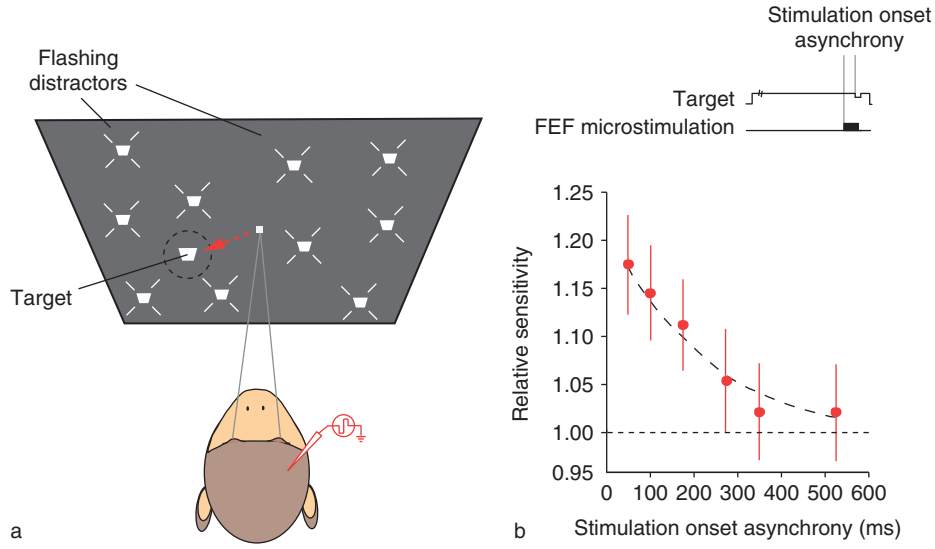

Figure 1 Enhancement of covert spatial attention with microstimulation of the FEF. (a) Monkeys were trained to maintain central fixation and covertly attend to detect the transient dimming of a peripheral target while ignoring a flashing distracter. Before engaging in the attention task, the movement field of the FEF site was determined by mapping the saccades (red arrow) evoked by suprathreshold microstimulation of a site within the FEF. The attention target was then placed at the endpoint of the suprathreshold saccade vector (dotted circle) and the microstimulation was decreased to 50% of the threshold current (i.e., to a level that did not evoke a saccade). Next, the luminance change sensitivity of the monkey at the target location was determined during control trials or trials in which subthreshold microstimulation was delivered to the FEF site. (b) Microstimulation increased the monkey's sensitivity to the target change, relative to control trials. Average relative sensitivity values (microstimulation/control) are shown for each stimulation onset asynchrony tested (red circles), with red lines indicating the standard error of the mean. Monkeys were able to detect smaller changes in target luminance when the target was in the part of space represented at the microstimulation site. The increase in sensitivity with microstimulation was greatest when the target change and the stimulation occurred at the same time.

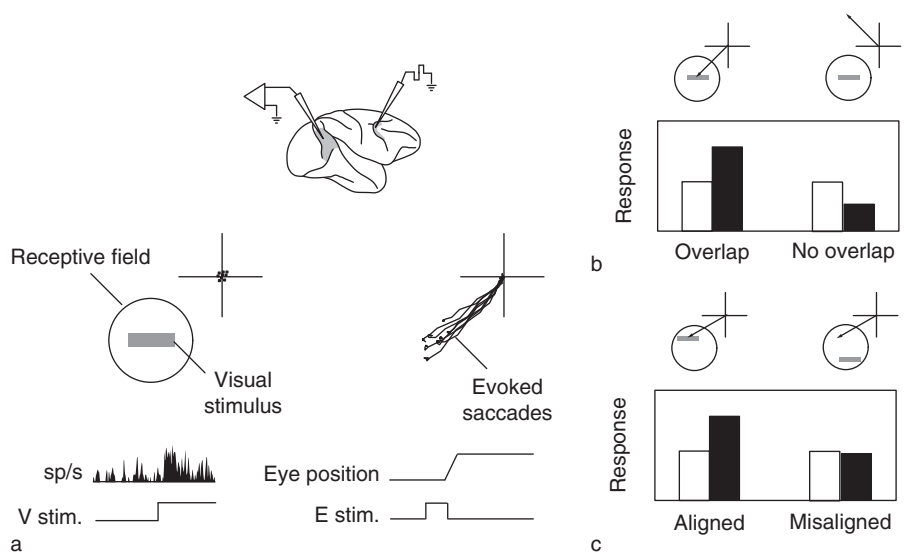

Figure 2 Modulation of visually driven responses of area V4 neurons with microstimulation of the FEF. Sites within the FEF were electrically stimulated while recording from neurons in area V4. (a) A side view of the macaque brain. (Top) The locations of the FEF in the anterior bank of the arcuate sulcus and of area V4 in the prelunate gyrus are shown (shaded). (Bottom left) The response of a V4 neuron when a visual stimulus (V stim) appears in its receptive field (RF) while the monkey maintains central fixation (dots at origin). (Bottom right) Eight saccades evoked to a location overlapping the V4 RF following microstimulation of an FEF site (E stim). (b) Subthreshold microstimulation of the FEF increases the visual activity ('Response') of V4 neurons with RFs (circles) that overlapped with the saccade represented at the stimulated FEF site (arrow from crosshairs), but it decreases visual activity when the receptive fields and the saccades did not overlap. Gray bars represent the visual stimulus used to drive the V4 response. Responses for control and stimulation conditions are shown in white and black, respectively. (c) The enhancement of V4 visual responses depends on the spatial alignment of the RF stimulus and the FEF saccade vector.

to covertly attend to them. Thus, one might expect that microstimulation of the FEF or the SC will likewise result in the enhancement of the responses of visual cortical neurons to RF stimuli. Indeed, this was the hypothesis posed by Moore and Armstrong, who paired FEF microstimulation with single-neuron recording in extrastriate area V4. Specifically, they studied the impact of subthreshold microstimulation of FEF sites on the visual responses of V4 neurons with RFs that were either overlapping or non-overlapping with the spatial location represented at the FEF site (Figure 2). They found that following very brief trains of FEF stimulation (20–50 ms), visual responses were enhanced at overlapping locations in the visual field, whereas they were suppressed at non-overlapping locations. Furthermore, this attention-like modulation was greatest when there was another 'distracter' stimulus presented outside of the V4 neuron's RF. A subsequent study also found that FEF microstimulation can alter the structure of V4 RFs and can bias V4 responses in favor of one of two RF stimuli. Importantly, these changes in V4 responses are very similar to what has been observed in the same area during voluntary attention. Together, the previous results suggest that microstimulation of oculomotor structures drives covert attention and its correlates in visual cortex, and that indeed, as speculated by Latto and Cowey, visual spatial attention seems to be the result of feedback from the oculomotor system onto visual cortical areas, and this feedback influence may exist in order to modulate processing according to eye movement plans.

Conclusion

Neurophysiological work has directly addressed and found much evidence of a role of saccade mechanisms in directing covert visual spatial attention. This evidence appears to support the conclusion that the neural mechanisms of covert and overt spatial attention are identical with respect to the influence of the saccade plan on visual representations. Although it should be assumed that a different set of mechanisms are in effect when attention is directed overtly (i.e., when planned movements are triggered and shifts in gaze are actually carried out) than when it is directed covertly, results suggest that these differences are less important than the similarities.

See also: Attentional Networks in the Parietal Cortex; Cingulate Cortex; Frontal Eye Fields; Multisensory Convergence and Integration; Neglect Syndrome and the Spatial Attention Network; Neural Coding of Spatial Representations; Parietal Cortex and Spatial Attention; Sensorimotor Integration: Barrels, Vibrissae and Topographic Representations; Sensorimotor Integration: Models; Spatial Cognition; Spatial Cognitive Maps; Supplementary Eye Fields; Vision for Action and Perception; Visual Attention; Visual–Vestibular Interactions.

Further Reading

Armstrong KM, Fitzgerald JK, and Moore T (2006) Changes in visual receptive fields with microstimulation of frontal cortex. *Neuron* 50: 791–798.

Awh E, Armstrong KM, and Moore T (2006) Visual and oculomotor selection: Links, causes and implications for spatial attention. *Trends in Cognitive Sciences* 10: 124–130.

James W (1890) *Principles of Psychology.* New York: Holt.

Kolb B and Whishaw IQ (2003) *Fundamentals of Human Neuropsychology.* New York: Worth.

Latto R and Cowey A (1971) Visual field defects after frontal eye-field lesion in monkeys. *Brain Research* 30: 1–24.

Moore T and Armstrong KM (2003) Selective gating of visual signals by microstimulation of frontal cortex. *Nature* 421: 370–373.

Moore T, Armstrong KM, and Fallah M (2003) Visuomotor origins of covert spatial attention. *Neuron* 40: 671–683.

Moore T and Fallah M (2004) Microstimulation of the frontal eye field and its effects on covert spatial attention. *Journal of Neurophysiology* 91: 152–162.

Mountcastle VB, Lynch JC, Georgopoulos A, Sakata H, and Acuna C (1975) Posterior parietal association cortex of the monkey: Command functions for operations within extrapersonal space. *Journal of Neurophysiology* 38: 871–908.

Reynolds JH and Chelazzi L (2004) Attentional modulation of visual processing. *Annual Review of Neuroscience* 27: 611–47.

Wurtz RH, Goldberg ME, and Robinson DL (1982) Brain mechanisms of visual attention. *Scientific American* 246: 124–135.

Sensorimotor Integration: Barrels, Vibrissae and Topographic Representations

T A Woolsey, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Barrels

A barrel is a cylindrical group of neurons located in the middle cortical layer (layer IV) of the somatosensory cortex. Barrels receive segregated inputs from the thalamus that are related to spatially separate parts of the somatic sensory periphery. In this respect they are similar to ocular dominance columns in the primate visual cortex that receive segregated inputs related to each eye. Many rodents, rabbits, and other small mammals have barrels. Barrels are most clearly related to whiskers (some of which are termed 'vibrissae'), which are tactile organs located principally on the face (see Figure 1). Barrels are also related to other peripheral structures such as the toes and footpads of the paws. Rodent barrel cortex is planar, compact, and accessible. Barrels can be seen with the naked eye in unstained sections of the rat brain.

The Barrel Field

The barrel field is the architectonic field defined by the barrels. It is homomorphic (similar) to the whisker pattern on the face and on parts of the paws on the limbs. There is a precise topographic anatomical correlation between the sensory periphery and the barrel cortex, first understood by recording surface potentials evoked by stimulation of the contralateral face and body. Each whisker barrel has a unique position corresponding to the vibrissa projecting to it (Figure 1). This relationship has been detailed in numerous independent ways, including (1) via microelectrode recordings, because most cells in a particular barrel are activated by the appropriate contralateral whisker, (2) by metabolic markers, which measure short- and long-term changes in activity and are altered in corresponding barrels with stimulation of selected parts of the body, and (3) by selective surgical removal of different vibrissae or limbs in early postnatal life, after which the corresponding barrels are altered.

Vibrissae

A whisker or vibrissa is a large tactile hair in a mechanically isolated hair follicle. Each whisker on the face receives a substantial, segregated sensory innervation from the trigeminal (fifth) nerve. The facial vibrissae are arranged in species-specific, stereotypical, gridlike patterns. The arrangement of the barrels is also species specific, stereotypical, and gridlike. Each vibrissa has a unique position to which a barrel in the homologous grid position corresponds. Single afferent fibers are related to one of several receptor specializations in the hair follicle that collectively encode direction, velocity, frequency, magnitude, and duration of whisker movements. Sensory inputs to the vibrissae are modulated through coordinated motor activity. Striated muscles innervated by the facial (seventh) nerve are attached to and actively move or 'vibrate' the larger whiskers (vibrissae) back and forth during exploratory whisking behavior at frequencies averaging 7 Hz in rats and 15 Hz in mice. Each vibrissa follicle contains erectile tissue (sinus hair is the general name), under opponent autonomic control, to dynamically alter tactile thresholds.

Topographic Representations

A topographic representation is a coherent, if distorted, map in the brain. In rats, mice, and some other animals, histological markers outline this map in layer IV in single sections parallel to the cortical surface (Figure 2). This visible representation of the body – head, body, limbs, tail – in the somatosensory cortex is a concrete example of representation of sensory and motor functions in the brain deduced from clinical observations, electrophysiological recording, and experimental neuroanatomy. This illustrates a general principle: sensory inputs from the body (skin), eyes (retina), and ears (cochlea) and outputs for voluntary movements (muscle groups) have systematic representations in many central nervous system structures that are topographically similar to the order of the parts represented.

Barrel Components

The larger barrels in mice are ovoid, with a major diameter of about 300 μm; in rats they have a major diameter of about 600 μm. Cells are arranged around dense clusters of afferent endings from the thalamus (Figure 3). In mice, there are about 2500 neurons in the larger mouse barrels and approximately 10 000 neurons in a functional barrel column extending throughout the full cortical thickness. The number of neurons in a barrel is directly proportional to, but 17-fold greater than, the number of fibers innervating its contralateral whisker. Segregated thalamic projections for single whiskers from

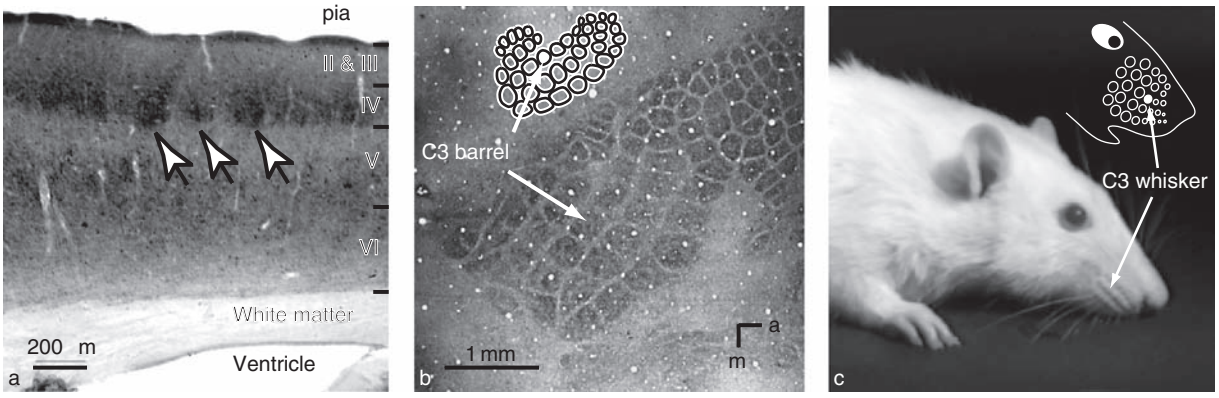

Figure 1 Whisker barrels in rat somatic cortex. (a) Barrels stain intensely for activity of an enzyme (mitochondrial cytochrome oxidase) involved in metabolism to produce energy. Each barrel (arrowheads) in layer IV is part of a functional column that extends through the full cortical thickness from the pia to the white matter (50-μm perpendicular section). (b) A section in the plane of layer IV of the left somatic cortex shows a pattern that is similar to that of the whiskers. The white holes are penetrating vessels. The marker in the lower right-hand corner indicates anterior a and medial m orientations. (c) The C3 whisker is identified on the right face of a rat.

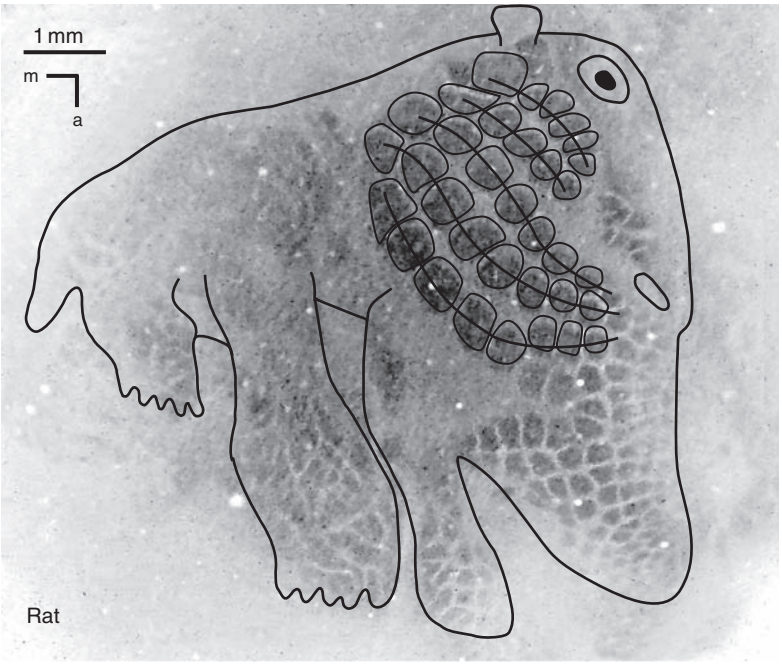

Figure 2 Barrels are demonstrable in the appropriate part of the physiologically defined cortical map of the contralateral body surface in histological sections from the flattened cortex. The distortions of the form of the animal in the map are directly related to peripheral innervation density (orientation: a, anterior; m, medial).

the ventroposterior medial (VPM) nucleus of the thalamus end largely on neurons in layer IV, and also in upper layer VI. The stellate neurons in layer IV are of two principal types: spiny and smooth, in a ratio of about 10:1. Most smooth cells are positive for glutamic acid decarboxylase (GAD) and are inhibitory interneurons. The axons and dendrites of layer IV neurons cells are largely confined in layer IV to the barrel in which their soma is located. About 93% of the synapses in a barrel are asymmetrical (excitatory) and of these about 20% of these are from the somatosensory thalamus. Both types of stellate cells receive thalamic synapses. Those on the smooth cells are distributed on the soma and proximal dendrites, while those on the spiny cells terminate on spines at regular intervals. Each barrel has a rich capillary plexus that is supplied by several penetrating arterioles.

Barrel Activation

Functional organization of the barrel cortex is predominantly vertical, or 'columnar.' Neurons in layer IV, and to a lesser extent those in layer VI, respond principally

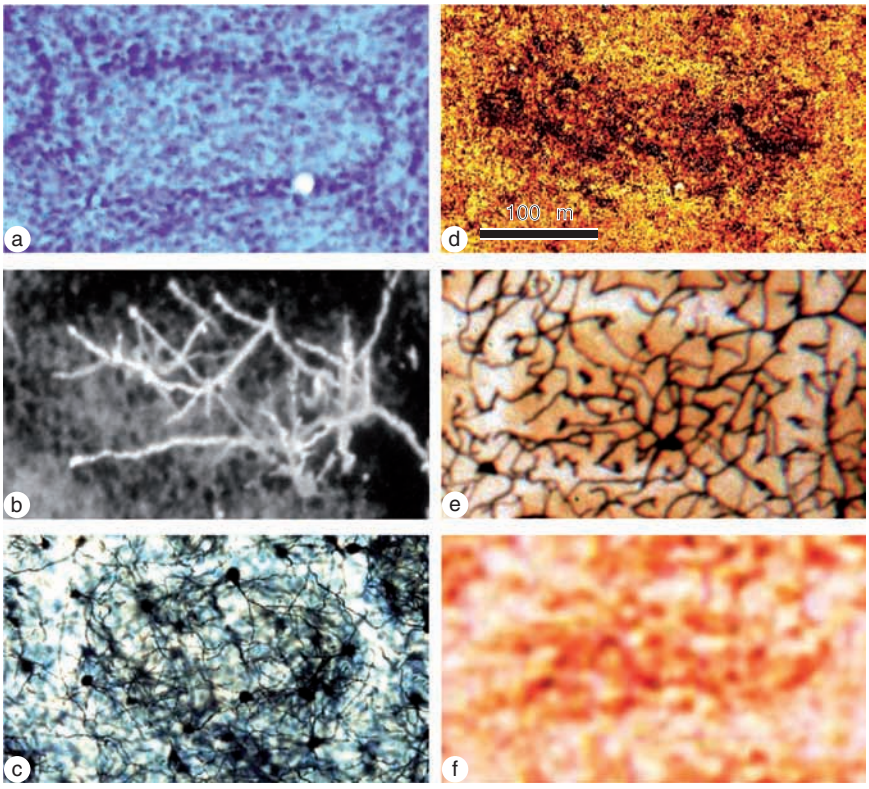

Figure 3 Barrel components from tangential sections of mouse cortex. (a) Barrel neurons in a ring define a barrel with a cell body (Nissl) stain. Each barrel is separated from its neighbors by relatively cell sparse regions. (b) Several thalamocortical axons projecting to a barrel labeled with a brightly fluorescent dye. Throughout the whisker barrel pathway, whisker-related afferents have dense, segregated ramifications confined to the appropriate whisker patch (prepared by Dr. Stephen Senft). (c) Single neurons stained by a Golgi method with Nissl counterstain showing barrel neuron processes (dendrites and axons) that mainly arborize in the barrel where their cell bodies are located. (d) Radioactive 2-deoxy-D-glucose labels a single whisker barrel after stimulation of the homotypical single contralateral whisker (autoradiogram prepared by Dr. Tony Mosconi). (e) Dense capillaries (filled with a black marker) in a cytochrome oxidase-stained barrel. (f) Serotonin-containing axons from the brain stem raphe are confined to barrels in young animals and mediate aspects of the development of the anatomic and functional organization of the cerebral cortex (prepared by Dr. Carol Bennett-Clarke).

to deflections of a single vibrissa. They indicate direction of whisker movement, have sustained or transient discharges to whisker deflections, can code for amplitude and rate of displacement, and may follow sinusoidal stimuli to frequencies exceeding 50 Hz. In these respects, they closely resemble those of the primary whisker afferents. Metabolic (2 deoxy-D-glucose) uptake after stimulation of a single whisker labels a single column of cells and neuropil that coincides precisely in layer IV with the relevant barrel and extends throughout the full cortical thickness. Prompt, focal changes in blood flow to barrels accurately reflect the location, timing, and magnitude of whisker-evoked neuronal activity.

Cells in all layers of barrel cortex are activated by highly specific stimuli that vary by layer. With whisker stimulation they exhibit precise directional sensitivities, are often inhibited by certain forms of stimulation, and integrate information from several different neighboring vibrissae. Interactions are through asymmetrical patterns of intracortical connections within and between different whisker columns. For instance, cells in layer V are responsive to particular spatial and temporal sequences of movement of adjacent whiskers as if the neurons were detectors for objects of finite sizes touched by whisker hairs. Models suggest that convergence from different thalamocortical fibers and intracortical inhibition improve the signal-to-noise ratio during whisker activation.

Whisker-to-Barrel Pathways

Activity is conveyed from the whiskers to the barrel cortex over two direct, parallel pathways, each crossing three synapses (see Figure 4). Afferent cell bodies are in the trigeminal ganglion where, as a rule, one ganglion cell innervates only one whisker. Central processes of these neurons terminate mainly in two separate ipsilateral brain stem nuclei: the principal

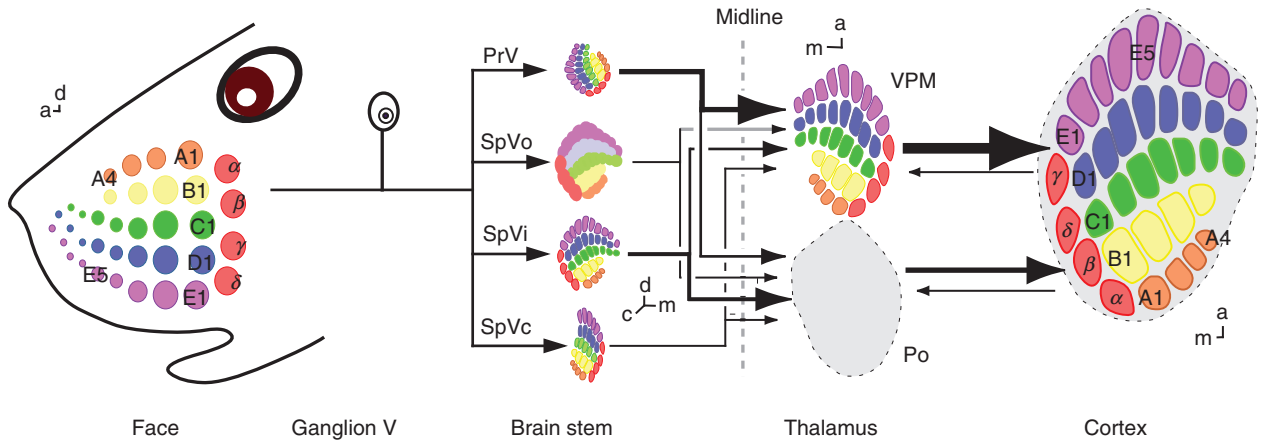

Figure 4 Parallel rodent trigeminal pathways. Sharply somatotopic neuronal patterns and projections (lemniscal), identical to the pattern of the whiskers on the face, can be seen from standard histology of the brain stem, with the thalamus just as in layer IV barrels of the cortex. The hallmark of this system is the predominance of single whisker representations that are dominated by activation of the appropriate whisker. A parallel pathway (paralemniscal) is activated at slightly slower latencies by multiple whiskers. Primary afferents converge on neurons in brain stem trigeminal nuclei, which project mainly to the posterior nucleus of the thalamus and to the cortex between and above and below the barrels in the cortex. Rows of whiskers and barrels are labeled with letters (A–E), whiskers and barrels within the rows are numbered (e.g., A1–A4), and whiskers and barrels between the ends of the rows are labeled with Greek letters ($\alpha-\gamma$). PrV, principal sensory nucleus of the trigeminal nerve; SpV, spinal nucleus of the trigeminal nerve (SpVo, subnucleus oralis; SpVi, subnucleus interpolaris; SpVc, subnucleus caudalis); VPM, ventroposterior nucleus of the thalamus; Po, posterior nucleus of the thalamus. Lines and lower case letters indicate the principal orientations in the animal (d, dorsal; a, anterior; m, medial; c, caudal).

trigeminal nucleus (PrV) and the spinal trigeminal nucleus (SpV), which is subdivided into three subnuclei where whiskerlike patterns of afferents and neurons are termed 'barrelettes.' Cells from these two nuclei project to two contralateral somatosensory thalamic nuclei: the ventroposterior medial nucleus, where whiskerlike patterns of afferents and neurons are termed 'barreloids,' and the architecturally homogeneous posterior nucleus (Po). Neurons in the thalamus project to the somatosensory cortex of the same side.

The 'lemniscal' pathway is faster and is activated by single whiskers. Afferent terminals in each locus on this path are grouped in whiskerlike patterns and the target cells typically have processes confined to a whisker patch and are strongly activated by single whiskers. The PrV is the main brain stem relay and the main thalamic relay is the VPM nucleus. Thalamocortical axons from VPM nuclei end in clusters in barrels in layer IV and in upper layer VI. Representations can be demonstrated throughout the entire lemniscal pathway, in a single individual, in sections cut in specific planes, with simple, routine staining methods.

In contrast, a 'paralemniscal' pathway originates mainly from large cells in the SpVi (subnucleus interpolaris), on which afferents from many whiskers converge. These project to the Po in the thalamus, which then projects between the barrels in layer IV and also to layers III and V of the barrel cortex. This pathway is slower and its cells are activated by multiple adjacent whiskers (Figure 4). Information from the two systems is integrated to provide accurate information about spatial locations of objects encountered as the animal moves its whiskers back and fourth during exploration.

Development of the Whisker Barrel System

The whisker barrel system offers special advantages for studies of brain development. Rodents such as hamsters, mice, and rats are very immature (altricial) at birth. Although the neurons of somatosensory pathway are generated *in utero*, at birth, young neurons are still migrating into the cortex. Further, the segregation of the central axonal projections along the pathway is not yet complete, synapses are forming, and most of the target neurons in the brain have only begun to differentiate. The vibrissae are recognizable in the fetus a week before birth. Temporal gradients in the development of whisker representations follow the same sequences as found for earlier neurogenesis along the neuraxis and within a particular part of the brain. The pattern of central projections is evidently instructed from the periphery and transmitted from one central locus to the next. Cortical transplant experiments demonstrate that visual cortex placed in the somatosensory area can express barrels, and, conversely, putative barrel cortex transplanted to the visual area takes on visual cortical architecture. The size of representations in the whisker pattern is related to relative, rather than absolute, peripheral innervation density.

Developmental and Adult Plasticity

Surgical removal of whiskers changes the architecture and function of the central whisker system in predictable, time-dependent ways. Repeatedly cutting off whisker hairs in neonates or in adults reduces activity in the whisker pathway and deprives it of patterned input. Chronic whisker clipping changes a number of biochemical parameters related to energy metabolism and transmitter synthesis in appropriate whisker barrels in adults. Deprivation can be tested in the short term, or reversed in the longer term, by letting the whisker hairs grow back. Neurons in barrel columns deprived by whisker clipping from birth exhibit higher spontaneous activity and less specific spatial tuning, and are more strongly activated by adjacent unclipped whiskers than is seen in normal animals. Sensory deprivation or pharmacological blockade of neuronal activity or synaptic transmission in the postnatal period does not apparently alter the patterns of long axonal projections. These manipulations, however, change patterns of intracortical connections in ways that are consistent with a strong role for activity in establishing circuits for functional integration. Novel synapses form in the appropriate barrel of adults after only 16 h of continuous whisker stimulation.

Genetic Analyses

Mice are the mammalian species of choice for genetic analysis. Animals created through selective breeding and homologous recombination permit better understanding of the molecular basis of central nervous system development, peripheral pattern specification, and brain function. Numerous strains of mice with abnormalities of the central nervous system and vibrissal patterns have been well documented. For example, mice have been selectively bred without barrels (barrelless) and a gene uniquely expressed in somatosensory cortex has been described. The *N*-methyl-D-aspartate (NMDA) receptor expression is necessary for development of the whisker pattern in the trigeminal brain stem, but patterns are preserved in the thalamus and cortex. Pattern formation is not (NMDA) subunit specific. These and a host of other findings have opened the way to profound insights from the detailed genetic analysis of the mouse central nervous system. The whiskers, rather like fruit fly bristles, and the barrel pathway provide practical markers for interpreting genetic control of brain development, function, and pathology.

Behavior and Ethology

The large whisker barrels constitute 20% of the total somatic representation that constitutes 30% of the dorsolateral surface of the rat cerebral cortex. A neural system as prominent as the whisker-to-barrel system would seem to have considerable importance for the creatures that possess it. Rodents spend significant time in active exploration in the dark, in burrows or other close quarters, and in contact with other individuals. In these species, the whisker barrel system appears to have developed for analysis and recognition of objects which can be touched (this is how other species, such as carnivores, use their whiskers), and also for navigation (rather as the blind use canes to scan the environment as they walk). Rats can be trained to discriminate different tactile surfaces and they can detect textures that are at least as fine as those that can be distinguished by humans with their fingertips. Rats can be taught to use their whiskers to determine whether to jump across an abyss from one platform to another in the dark. Interestingly, it has been shown that animals can be guided remotely, via paired stimulation of indwelling electrodes in the barrel cortex and other structures, through complex three-dimensional environments. These studies add to classical work showing that rats use their whiskers to quickly traverse mazes for food. Simultaneous recording from different sites in the whisker pathway of active animals shows that the system monitors coordinated whisking and whisker inputs. The whisker barrel pathway is not used solely for tactile processing; mice establish social hierarchies by chewing off the whiskers of subordinates.

The whisker barrel pathway has many features – behavioral importance, accessibility, ease of stimulation, certainty of localization, and genetic and developmental modifiability – that make it a good model system for the study of the mammalian brain and brain disorders in common laboratory species.

See also: Barrel Cortex Circuits; Cortical Plasticity and Learning: Mechanisms and Models; Neural Circuitry in the Somatosensory System; Neuronal Plasticity after Cortical Damage; Somatosensory Cortex: Functional Architecture; Somatosensory Receptive Fields; Somatosensory Plasticity; Somatosensory Cortex; Vibrissa Movement, Sensation and Sensorimotor Control; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging.

Further Reading

Celikel T and Sakmann B (2007) Sensory integration across space and in time for decision making in the somatosensory system of rodents. *Proceedings of the National Academy of Sciences of the United States of America* 104: 1395–1400.

Fox KD (2007) *Barrel Cortex*. Cambridge: Cambridge University Press.

Fukuchi-Shimogori T and Grove EA (2001) Neocortex patterning by the secreted signaling molecule FGF8. *Science* 294: 1011–1012.

Knott GW, Quairiaux C, Genoud C, et al. (2002) Formation of dendritic spines with GABAergic synapses induced by whisker stimulation in adult mice. *Neuron* 34: 265–273.

Lu HC, Gonzalez E, and Crair MC (2001) Barrel cortex critical period plasticity is independent of changes in NMDA receptor subunit composition. *Neuron* 32: 619–634.

Miller B, Blake NMJ, and Woolsey TA (2001) Barrel cortex: Structural organization, development and plasticity. In: Kossut M (ed.) *Plasticity of Adult Barrel Cortex*. Johnson City, TN: FP Graham.

Simons DJ and Land PW (1987) Early experience of tactile stimulation influences organization of somatic sensory cortex. *Nature* 326: 694–697.

White EL and Keller A (1989) *Cortical Circuits: Synaptic Organization of the Cerebral Cortex*. Boston: Birkhäuser.

White EL, Weinfeld L, and Lev DL (1997) A survey of morphogenesis during the early postnatal period in PMBSF barrels of mouse SmI cortex with emphasis on barrel D4. *Somatosensory & Motor Research* 14: 34–55.

Woolsey CN (1952) Patterns of localization in sensory and motor areas of the cerebral cortex. In: *Milbank Symposium, The Biology of Mental Health and Disease*, pp. 193–206. New York: Hoeber.

Woolsey TA (1967) Somatosensory, auditory and visual cortical areas in the mouse. *Johns Hopkins Medical Journal* 121: 91–112.

Woolsey TA, Rovainen CM, Cox SB, et al. (1996) Neuronal units linked to microvascular modules in cerebral cortex: Response elements for imaging the brain. *Cerebral Cortex* 6: 647–660.

Woolsey TA and Van der Loos H (1970) The structural organization of layer IV in the somatosensory region (SI) of mouse cerebral cortex: The description of a cortical field composed of discrete cytoarchitectonic units. *Brain Research* 17: 205–242.

Relevant Websites

http://www.brain-map.org – Allen Institute for Brain Science: Allen Brain Atlas.

http://www.ibro.info – International Brain Research Organization: Science Issues (Thomas A. Woolsey, Barrel cortex).

Sensorimotor Integration: Models

F A Mussa-Ivaldi and A Fishbach, Northwestern University, Chicago, IL, USA

© 2009 Published by Elsevier Ltd.

Introduction

With the advent of robotics, scientists have become increasingly aware of the complex problems that the brain must solve to carry out even the common task of bringing a glass of water to one's mouth. In this respect, our nervous system is an expert in mechanics, far more competent than the most advanced physics student. Two approaches have been considered as a means to simplify the problems associated with movement control. One is based on reflexes and on the concept of feedback. The other is based on the idea of memory-based computation.

The study of reflexes, in combination with the engineering concept of feedback control, led to early models of voluntary movements. In a feedback control model, the nervous system generates forces that tend to eliminate the sensed error between desired and actual state of motion (position and velocity). This concept is attractive because it does not require any knowledge of dynamics. However, feedback control systems tend to exhibit severe instability problems in the presence of large delays.

An alternative approach to the control of movements was based on the computational idea of storing motor programs into memory and retrieving them when needed. This is analogous to the concept of a 'lookup table.' In a lookup table, the control signals that are needed to generate a movement are not computed explicitly from knowledge of the limb dynamics. Instead, patterns of neural activations are stored as they are produced, in conjunction with the movement that they cause, in an associative memory bank. Subsequently, when a movement is required, the desired movement is used as an address for retrieving the corresponding neural activations. David Marr and James Albus proposed this as a computational model of cerebellar motor control. Subsequently, Marr criticized this approach based on its inability to capture and express the wide variety of movements that characterize natural behaviors. The failure of lookup table models is due to a combinatorial explosion arising from the structure of multijoint kinematics: The number of movement patterns that need to be stored grows exponentially with the number of degrees of freedom.

This article discusses how the nervous system learns to control body-environment dynamics by representing and learning the transformations that relate movements and forces to the neural activities that drive the muscles. These transformations are maps that are accessed in different ways. For example, one way to access a motor map is by deriving a prediction of expected movement given the activities of the muscles. Conversely, another way to access the same map is by deriving the muscle activities that are needed to generate a desired movement. From a mathematical standpoint, the input and output entities are exchanged in these two situations. The transformation from muscle activities to movement is called a 'forward' model, and the transformation from desired movement to muscle activities is called an 'inverse' model. There is ample evidence suggesting that the brain represents both types of models.

The use of the term 'model' may engender some misunderstanding because we are used to thinking of mathematical models as representations involving symbolic operations, such as trigonometric functions, logarithms, integrals, and derivatives. Here, it is not suggested that the brain carries out these symbolic operations. The concept of internal model in motor neuroscience refers to the existence of transformations that are equivalent to the computation of dynamics' equations but are based on a radically different form of 'calculus.' This calculus is based on elementary units, or primitives, that execute simple operations and that can be combined to generate a broader repertoire of behaviors. This article reviews some of the experimental evidence and theoretical arguments supporting the existence of such primitives and their role in the learning and generation of motor behavior.

A Geometrical Framework

A variety of computations associated with the generation and control of movements can be formulated as coordinate transformations. Sensory information about the state of motion of the body comes from a variety of sources, each concerned with particular mechanical variables. For example, skeletal muscles are endowed with sensors that measure the muscle strain and its rate of change. Muscles are also endowed, at their junction with tendons, with Golgi tendon organs that sense variations in muscle force. Signals from other sensory modalities, such as the eyes and the vestibular organs, provide information about the position of the body and its parts with respect to the environment.

The quantitative description of motor and sensory signals implies the choice of a coordinate system,

which falls quite naturally into three classes: actuator coordinates, generalized coordinates, and endpoint coordinates.

Endpoint Coordinates

Endpoint coordinates reflect the properties of the environment in which we operate. Perhaps the most striking of these is the Euclidean symmetry: the invariance of distance between points under rotations and translations. The dimensionality of endpoint space is generally rather low. Since the position of a moving rigid body is completely described by six coordinates (three translations and three angles), the force applied to the body is a six-dimensional force/torque vector. Human hand movements tend to be spontaneously organized in endpoint coordinates: The kinematics of reaching is generally a smooth trajectory (i.e., a straight line with bell-shaped velocity profile) when described in the coordinates of the hand. This is remarkable because the kinematics of a multijointed limb defines a curved manifold that does not obey the symmetry of Euclidean space. For instance, a given angular excursion of the shoulder and elbow causes a net displacement of the hand that varies in amplitude and direction depending on its starting position; however, the length of a segment in Euclidean geometry does not depend on the location or the orientation of the segment. Similar observations have been made across species.

Actuator Coordinates

Information about the position of the arm in space can also be conveyed by specifying a collection of muscle lengths, $l = (l_1, l_2, \ldots, l_m)$. In this coordinate system, a force is a collection of muscle tensions, $f = (f_1, f_2, \ldots, f_m)$. Actuator coordinates afford the most direct representation for the motor output of the central nervous system. The final output stage (or 'final common path') of the motor system is provided by the motor neurons in the spinal cord, whose activities determine the state of contraction of the muscles. The latter is measured by the receptors in the muscle spindles, which deliver this information to the nervous system via the dorsal root ganglia.

Although each muscle is operated by a distinct set of neural signals, actuator coordinates do not constitute a system of mechanically independent variables: one cannot set arbitrary values to muscle lengths without eventually violating a kinematic constraint. Different approaches to the study of motor control may employ different levels of detail in the description of physiological processes, leading to actuator models with different dimensionality. For example, some approaches focus on simplified models in which a single joint is operated by a pair of muscles acting as reciprocal (agonist/antagonist) actuators. At the most detailed level of analysis, individual motor units may be considered as actuator elements. In this case, the dimensionality of the actuator space can reach hundreds of thousands.

Generalized Coordinates

A different description of body motions is afforded by the values of the joint angles that define the orientation of each skeletal segment either with respect to fixed axes in space or with respect to the neighboring segments. Joint angles are a particular instance of generalized coordinates. Generalized coordinates are independent variables suitable for describing the dynamics of a mechanical system. The term dynamics refers to a description of how the state of motion of a limb evolves under the influence of the muscle forces, of the passive mechanical properties of the moving limb, and of the forces generated by the environment with which the limb comes in contact. Mathematically, dynamics are expressed by differential equations that relate these forces and the current state of motion of a limb (its position and velocity) to the rate of change of velocity (i.e., the limb's acceleration). A typical problem in engineering is to find the forces that must be applied to the mechanical joints of a robotic arm so as to generate a desired trajectory. The solution to this problem – called 'inverse dynamics' – requires representing the dynamics equation of the arm and calculating the applied force from the desired position, velocity, and acceleration of each joint.

The concept of state is critically important in a dynamical formulation. The state of a mechanical system is the smallest set of variables that suffices to describe the future evolution of a system given its current value and the applied forces.

The idea that the state of a system is specified by its position and velocity is generally applicable to all mechanical systems, but specific applications require a careful definition of what is meant by 'position' and 'force.' This extension to generalized coordinates and generalized forces was introduced soon after Newton's early definition of force for simple point mass systems, and it involves the incorporation of constraints between the constituents of a composite system.

General Formulation of Limb Dynamics

In vector notation, the dynamics equations for a multijoint limb can be written as

$$I(q)\ddot{q} + G(q, \dot{q}) + E(q, \dot{q}, \ddot{q}, t) = C(q, \dot{q}, u(t)) \quad [1]$$

where $q = (q_1, q_2, \ldots, q_n)$ is the generalized coordinate description for the position – or 'configuration' – of the limb though its n joint angle coordinates; $\dot{q}$ and $\ddot{q}$ are respectively the first (velocity) and second (acceleration) time derivatives of q, I is an $n \times n$ matrix of inertia (that depends on the configuration); $G(q, \dot{q})$ is a vector of centripetal and Coriolis torques; and $E(q, \dot{q}, \ddot{q}, t)$ is a vector of external torques, which depends, in general, on the state of motion of the limb, its acceleration, and also, possibly, time. In a more compact notation, Coriolis and inertial forces are lumped in a single function $D(q, \dot{q}, \ddot{q})$ representing the passive dynamics of the limb, and eqn [1] becomes

$$D(q, \dot{q}, \ddot{q}) + E(q, \dot{q}, \ddot{q}, t) = C(q, \dot{q}, u(t)) \qquad [2]$$

The left side of eqns [1] and [2] represents the torque due to inertial properties and to the action of the environment (part of which may be considered as 'noise'). The term $C(\cdot)$ on the right side represents the net torque generated by the muscles. The time-dependent function $u(t)$ is a vector of control variables representing, for example, a set of neural signals directed to the motor neurons or the desired limb position at time t. An additional term, to be added to the right side of eqns [1] and [2], could represent the noise associated with $u(t)$.

Costs and Benefits of Motor Redundancy

A distinctive feature of biological control systems is what appears to be an overabundance of controlled components and of degrees of freedom. The term 'kinematic redundancy' describes this feature. Advanced robotics technologies have recognized the value of design that allows for 'extra' degrees of freedom in favor of greater versatility and dexterity. Ultimately, the term 'redundancy' could be considered as a misnomer because it does not acknowledge that the dimension associated with a task may well exceed the dimension of the kinematics space. For example, Hogan pointed out the important role played by what appear to be extra degrees of freedom that modulate the impedance of the endpoint of a limb in the face of an expected impact.

The computational problem associated with kinematic redundancy is illustrated in Figure 1, which represents the maps of force and position in a serial limb, such as the arm. On the right side of the figure are the representations of state and force variables in endpoint coordinates. These variables are observed at the interface with the environment – for example, the position of the fingertip and the contact force at this point. On the left side are descriptions in actuator coordinates. Note that the force generated by the muscular system results from a function that maps the state of motion of the muscles – collectively indicated by the two vectors $l, \dot{l}$ of muscle lengths and their rates of change – and the muscle control signals $u(t)$ into a vector f of muscle tensions. This map of muscle tensions is in general nonlinear and noninvertible. The generalized coordinates, typically joint angle variables, are an intermediate representation between actuator and endpoint coordinates. The directions of the arrows in Figure 1 correspond to the directions in which kinematic and force transformations are well posed. For example, joint angles map into endpoint position via a direct kinematics function L and into muscle length via a function M. Both functions are generally noninvertible, although for different reasons. The map L is noninvertible because a given endpoint position can be obtained through a large number of joint configurations that define a so-called 'null space.' The term null space refers to directions of variability in the space of generalized coordinates that result in no change of the endpoint location. In more formal terms, if A is a matrix acting on a vector space x, the null space of A is the set of all vectors x that satisfy $Ax = 0$. The map M is noninvertible because an arbitrary set of muscle lengths may not be realizable because it may have no corresponding image in configuration space. Although the kinematic maps L and M refer to coordinate transformations, forces transform in a dual, reciprocal way through the Jacobian derivatives ∂L and ∂M of the corresponding kinematic functions. Note the reversed direction of the corresponding arrows.

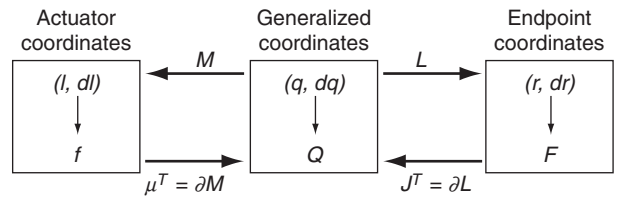

Figure 1 Coordinate systems and coordinate transformations for planning and control.

The Representation of Space in Motor Planning

The remarkable ability of the visual system to capture the Euclidean invariance of object size and shape with respect to rigid translations and rotations has been extensively studied. For example, it has been proposed that the Euclidean metric of object space is implicitly learned and represented internally. In contrast with the long-standing interest in the visual perception and representation of space, little is known about how the fundamental geometrical properties of space are represented in the motor system, in which

neither sensory nor muscle command signals are characterized by Euclidean invariance. Our ability to generate straight motions of our hand in different directions and from different starting points, without apparent effort, is a simple demonstration of our motor system's de facto ability to capture the Euclidean properties of the space in which we move.

Mosier et al. investigated how movement representations are reorganized by the central nervous system to capture the Euclidean properties of the space within which actions take place. In these experiments, subjects wore a data glove that allowed for the measurement of 22 signals related to hand and finger postures. A highly redundant linear mapping was used to transform these 22 signals into two coordinates defining the location of a cursor on a computer monitor (Figure 2). The results suggested that subjects not only learned to produce novel coordinated movements

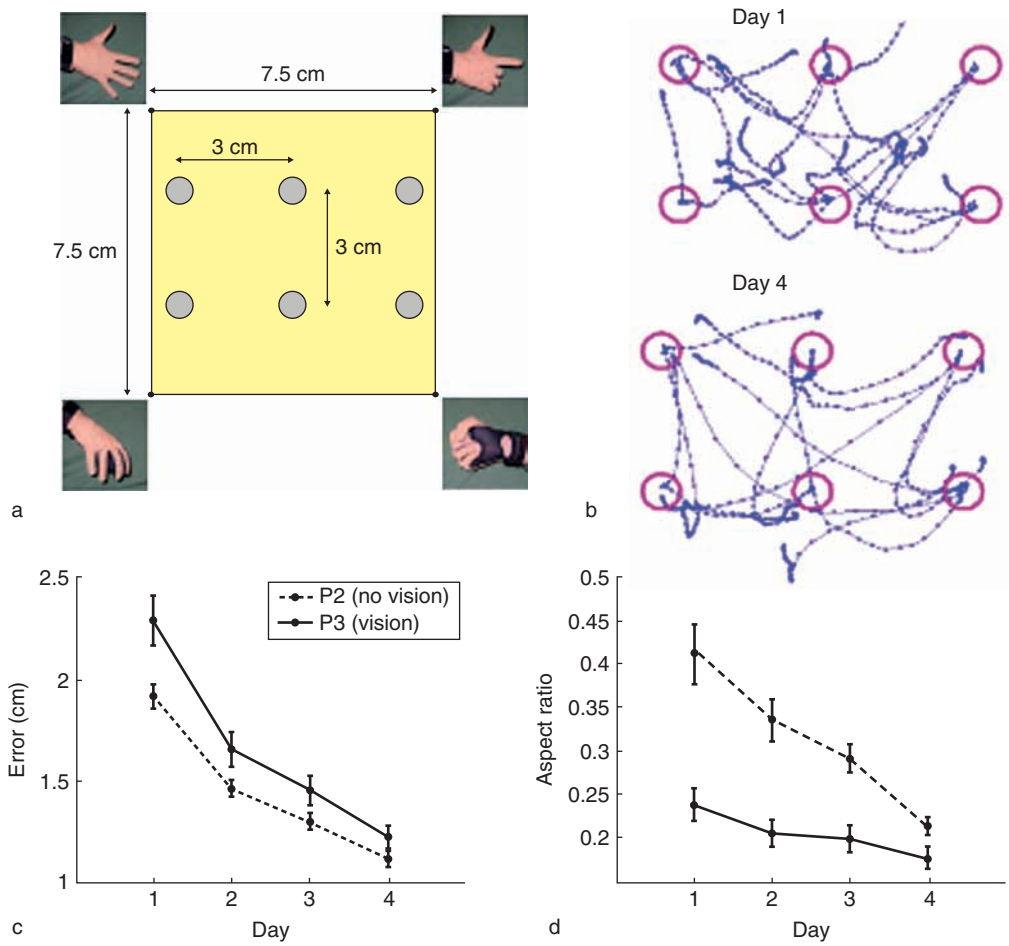

Figure 2 Reorganization of hand movements. Subjects were asked to control movement of a cursor by changing the configuration of hand and fingers. The hand was instrumented with a data glove. Hand configuration controlled the position of a cursor on the screen via a smooth and linear map from the 22-dimensional vector of glove signals to the two coordinates of the monitor. Circular targets were presented and subjects were asked to make a rapid movement to the target, stop, and then correct. Movements were executed under one of two training protocols. In the no-vision (NV) protocol, the cursor was suppressed during the reaching movements. The cursor was only presented at the end of reaching to allow for a correction (not analyzed). In the vision (V) protocol, the cursor was always visible. Subjects in the P2 group were trained and tested under the NV protocol. Subjects in the P3 group were trained under the V protocol and tested under the NV protocol. Data analysis was carried out only for test trials allowing comparison of how subjects executed reaching movements without visual guidance after having trained without (P2) and with (P3) visual guidance. (a) The linear map from glove signals to screen coordinates was set up by asking the subjects to hold their hand in four 'standard gestures' corresponding to the corners of a rectangular region on the monitor. (b) Two sets of cursor trajectories obtained from a subject on day 1 and on day 4 of training. (c) Subjects in both P2 and P3 groups learned the reaching task with similar results. The two plots show the average reaching error measured in each of four sessions on consecutive days. The vertical bars are 99% confidence intervals. (d) Under both training conditions, subjects tend to produce straighter cursor motions as evidenced by the decreasing trend in 'aspect ratio,' the ratio of lateral deviation to movement extent. Subjects trained with vision generated significantly straighter trajectories than subjects trained without vision of cursor motion. Modified from Mosier KM, Sceidt RA, Acosta S, and Mussa-Ivaldi FA (2005) Remapping hand movements in a novel geometrical environment. *Journal of Neurophysiology* 94: 4362–4372, with permission.

to control the placement of the cursor but also developed a representation of the two-dimensional space (the plane of the monitor) onto which hand movements were mapped. The emergence of this representation is evidenced by the practice-induced tendency to generate controlled trajectories that were both shorter and less variable. The construction of this representation appears to involve the subjects' ability to partition the degrees of freedom of the hand into combinations that contributed to cursor movements and null space combinations that did not generate any cursor movement.

From Planning to Control

Although the transformation from a desired motion to a set of control signals $u(t)$ may be ill-defined, a plan of action may be implemented as an approximation of a desired force field by an appropriate tuning of viscoelastic actuators. The force field f generated by the combined viscoelastic behavior of the muscles under a pattern of control signals $u(t)$ is given by a function $f(l, \dot{l}, u(t))$. This force field can be expressed in terms of generalized coordinates through the Jacobian derivative of the map M, resulting in a computational structure that does not involve any ill-posed inverse mapping:

$$Q = \phi_f(q, \dot{q}, u(t)) = \partial M(q)^T f\Big(l(q), \dot{l}(q), u(t)\Big) \quad [3]$$

The planning of a desired behavior can in turn be expressed as a force field F that maps a state of the endpoint (e.g., a heading direction) into a corrective force $F(x, \dot{x}, t)$. This is a way to represent what some researchers would call a 'policy' – a prescribed action in response to an observed state. This policy can be represented as a force field in terms of generalized coordinates through the Jacobian derivative of the map L, resulting also in this case in a computational structure that does not involve any ill-posed inverse mapping:

$$Q = \psi_F(q, \dot{q}, t) = \partial L(q)^T F(x(q), \dot{x}(q), t) \quad [4]$$

Therefore, the biological implementation of a planned policy can be seen as the approximation

$$\varphi_f(q, \dot{q}, u(t)) \approx \psi_F(q, \dot{q}, t) \quad [5]$$

to be achieved through the appropriate choice of $u(t)$.

Motor Primitives and Field Approximation

In a series of experiments, the activity induced by chemical and electrical stimulation of the spinal interneurons of the frog was found to spread to several groups of motor neurons. This distribution of activity was not random but, rather, imposed a specific balance of muscle contractions. The activation of a group of muscles generated a force that was recorded by a sensor at the endpoint of the limb. This force vector changed in amplitude and direction depending on the position of the limb. The resulting force field converged toward a location in the reachable space of the limb – a stable equilibrium point.

Perhaps the most interesting aspect of these investigations was the discovery that the fields induced by the focal activation of the cord follow a principle of vectorial summation: when two separate sites in the spinal cord were simultaneously active, the resulting force field was the vector sum of the force fields induced by the separate activation of each site.

Consistent with the finding of vector summation, the net force field C induced by a pattern of K motor commands may be represented as a linear combination over the corresponding K primitive force fields, labeled by an index i running from 1 to K:

$$\sum_{i-=1}^{K} c_i \varphi_i(q, \dot{q}, t) \quad [6]$$

In this expression, each spinal force is a field that depends on the state of motion of the limb, $(q, \dot{q})$ and on time t. The descending commands $(c_1, c_2, \ldots, c_K)$ modulate the degree with which each spinal field $(\varphi_1, \varphi_2, \ldots, \varphi_K)$ participates in the combination. These commands select the modules by determining how much each one contributes to the net control policy. The linear combination of eqn [6] generates the torque, the generalized force that drives the limb inertia. Substituting it for the right-hand side of eqn [2] results in

$$D(q, \dot{q}, \ddot{q}) + E(q, \dot{q}, \ddot{q}, t) = \sum_{i=1}^{K} c_i \varphi_i(q, \dot{q}, t) \quad [7]$$

For simplicity, let us start by analyzing the implications of eqn [7] in the case in which the controlled limb is not affected by external forces – that is, $E = 0$. In this case, a least squares approximation (**Figure 3**) can efficiently determine the optimal set of coefficients $(c_1, c_2, \ldots, c_K)$ for a desired trajectory $q_D(t)$:

$$c_i = \sum_{i=1}^{K} [\Phi]_{i,j}^{-1} \Lambda_j \quad [8]$$

with

$$\begin{cases} \Phi_{l,m} = \int \varphi_l(q_D(t), \dot{q}_D(t), t) \bullet \varphi_m(q_D(t), \dot{q}_D(t), t)\, \mathrm{d}t \\ \Lambda_j = \int \varphi_j(q_D(t), \dot{q}_D(t), t) \bullet D(q_D(t), \dot{q}_D(t), \ddot{q}_D(t))\, \mathrm{d}t \end{cases} \quad [9]$$

where the symbol $\bullet$ indicates the ordinary inner product.

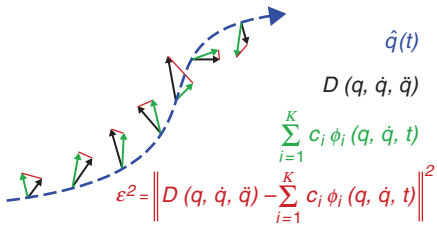

Figure 3 Control by field approximation. The desired trajectory is represented schematically in blue. The black arrows are the dynamic forces generated by D along the desired trajectory. The green vectors are the forces generated by a linear combination of force fields. The red bars are the errors to be minimized by least squares.

Although spinal force fields offer a practical way to generate movement, they also provide the central nervous system with a mechanism for movement representation. The possible movements of a limb can be considered as 'points' in an abstract geometrical space in which the fields $\{\varphi_i\}$ provide a basis such as that provided by Cartesian axes, and the selection parameters $\{c_i\}$ that generate a particular movement may be regarded as generalized projections of this movement along the $\{\varphi_i\}$ axes.

The Presence of Noise and Uncertainty Leads to a Segmentation of Motor Control

Primitive force field combinations do not have to be restricted to simultaneous activation, as in eqn [6]; they can be combined over time:

$$\sum_{i=1}^{N} c_i \varphi_{n_i}(q, \dot{q}, t - \Delta t_i)$$

where $n_i \in \{1, 2, \ldots, K\}$. Such a temporal summation of primitive force fields can arise when executing a preplanned complex movement or when the executed movement needs to be updated. Updating an ongoing movement may be required when the goal of the movement has changed during movement execution, or when the executed movement deviates from the planned one. Deviations between the planned and actual movement may be caused by inaccurate inverse/forward models or because of noise in the planning and execution stages of the movement. Indeed, many studies have demonstrated that rapid reaching movements of human and nonhuman primates are often characterized by irregular multipeaked velocity profiles, which can be interpreted as a temporal summation of movement primitives with bell-shaped velocity profiles. Many movement segmentation methods have been proposed, most of which assumed that all movement primitives share the same basic velocity profile but could be scaled in time and magnitude to best match the combined movement. It is important to note that demonstrating that a velocity trace can be represented by a linear superposition of a set of discrete primitives is not sufficient evidence for the existence of these primitives. Indeed, studies have suggested that a continuous control process can produce kinematic irregularities that do not reflect superposition of discrete movement primitives. However, Fishbach and collaborators have reported several experimental findings that are difficult to reconcile with a purely continuous controller. In their experiments, monkeys were trained to perform a one-dimensional reaching task, in which the target was occasionally perturbed at movement onset. Analysis of the movement corrections shows a wide distribution of the time of the first correction and a strong correlation between the correction timing and the amplitude of the initial portion of the movement. These results suggest a model for movement update in which a forward model is used to continuously estimate both the expected error of the ongoing movement and the variance of the predicted error. Once the probability distribution of the predicted endpoint is statistically different from the target's location, a discrete movement primitive is triggered to compensate for the discrepancy. This control strategy is similar to the sliding variable control which uses a threshold nonlinearity on the predicted error, except that the former uses a threshold nonlinearity on the predicted z score of the error. Fishbach and coworkers used their model to successfully predict the timing of the observed corrections. Whereas Fishbach et al.'s study provides strong support for the view of temporal superposition of discrete movement primitives, the interpretation of multipeaked velocity profiles of reaching movements remains a strongly debated subject.

Sensory Motor Adaptation

If the arm dynamics change while the modules remain unchanged, then the representation of the movement must change accordingly. Suppose that a desired trajectory $q_D(t)$ is represented by a selection vector $c = (c_1, c_2, \ldots, c_K)$ for a limb not affected by external forces (i.e., $E = 0$). If the limb dynamics are suddenly modified by an external force, $E(q, \dot{q}, \ddot{q}, t)$, then the full form of eqn [7] must be considered:

$$D(q, \dot{q}, \ddot{q}) + E(q, \dot{q}, \ddot{q}, t) = \sum_{i=1}^{K} c_i \varphi_i(q, \dot{q}, t)$$

The solution to this equation will be a trajectory $\tilde{q}(t)$, generally different from the original $q_D(t)$ for $E = 0$. To recover the desired trajectory, the selection vector needs to be changed from c to a new vector, $c' = c + e$, where the correction vector e satisfies

$$E(q_D(t),\dot{q}_D(t),\ddot{q}_D(t),t) = \sum_{i=1}^{K} e_i\varphi_i(q_D(t),\dot{q}_D(t),t) \quad [10]$$

The modified coefficients $c' = c + e$ offer a new representation of the desired trajectory $q_D(t)$; the change reflects an alteration of the dynamic conditions within which the trajectory is to be executed. This computational procedure for forming a new representation and for recovering the original trajectory is consistent with the empirical observation of aftereffects in force field adaptation. If the force field is removed after the new representation is formed, the dynamics becomes

$$D(q,\dot{q},\ddot{q}) = \sum_{i=1}^{K}(c_i + e_i)\varphi_i(q,\dot{q},t) \quad [11]$$

which can be rewritten as

$$D(q,\dot{q},\ddot{q}) - \sum_{i=1}^{K} e_i\varphi_i(q,\dot{q},t) = \sum_{i=1}^{K} c_i\varphi_i(q,\dot{q},t) \quad [12]$$

Therefore, removing the load while maintaining the new representation will result in a trajectory that corresponds approximately to applying the opposite load with the old representation, in agreement with observed aftereffects.

The modification of motion representation through the selection coefficients thus provides a useful mechanism for rapidly adjusting to changes in limb dynamics. It is of interest to ask whether these representations need to be recalculated in response to every dynamical change or whether previously existing representations can be restored. From a computational standpoint, whenever a dynamical change becomes permanent, such as when due to growth or damage, it would be convenient for the central nervous system to have the ability to restore the previously learned motor skills (i.e., the previously learned movement representations) without need to relearn them. It is possible for the adaptive system described here to do this by modifying the force fields expressed by the individual modules. A specific mechanism follows from expressing the coefficients $e = (e_1, e_2, \ldots, e_K)$ as a linear transformation of the original coefficients $c = (c_1, c_2, \ldots, c_K)$:

$$e = Wc \quad [13]$$

This is a simple coordinate transformation applied to the change e in selection vector; a linear associative network can implement it. A few steps of algebra lead to

$$\begin{aligned} D(q,\dot{q},\ddot{q}) + E(q,\dot{q},\ddot{q}) &= \sum_{i=1}^{K} c_i'\varphi_i(q,\dot{q},t) \\ &= \sum_{i=1}^{K}(c_i + e_i)\varphi_i(q,\dot{q},t) \\ &= \sum_{i=1}^{K} c_i\bar{\varphi}_i(q,\dot{q},t) \quad [14] \end{aligned}$$

where the old fields ϕ_i have been replaced by the new fields

$$\bar{\varphi}_i = \sum_{i=1}^{K}(\delta_{li} + W_{li})\varphi_l \qquad \delta_{li} = \begin{cases} 1 & \text{if } l = i \\ 0 & \text{otherwise} \end{cases} \quad [15]$$

This coordinate transformation is then made permanent through a change in the modules that represent individual force fields. By means of such coordinate transformation, one obtains the important result that the movement representation (i.e., the selection vector c) can be maintained invariant after a change in limb dynamics.

Forward and Inverse Models as Basis for Adaptive Control

The ability to generate a variety of complex behaviors cannot be attained by just storing the vector c of control signals for each action and recalling these selection coefficients when subsequently needed. Simple considerations about the geometrical space of meaningful behaviors are sufficient to establish that this approach would be inadequate. To achieve its typical competence, the motor system must take advantage of experience for going beyond experience by constructing internal representations of the controlled dynamics, such as the projections of D onto the fields φ_i, as in eqn [7] for $E = 0$. These representations need to allow the nervous system to generate new behaviors and to handle situations that have not yet been encountered. The term 'internal model,' used to indicate the collection of available internal representations, implies the ability to perform two distinct mathematical transformations: (1) the transformation from a motor command to the consequent behavior and (2) the transformation from a desired behavior to the corresponding motor command. A model of the first kind is called a 'forward model.' Forward models provide the controller with the means not only to predict the expected outcome of a command but also to estimate the current state of the limb in the presence of feedback delays. Conversely, a representation of the transformation from desired actions to motor commands (as in eqns [8] and [9])

is called an 'inverse model.' Studies by Kawato, Wolpert, and Miall have led to the proposal that neural structures within the cerebellum perform sensorimotor operations equivalent to a combination of multiple forward and inverse models. Strong evidence for the biological and behavioral relevance of internal models has been provided by numerous experiments. Some of these experiments involved the adaptation of arm movements to a perturbing force field generated by a manipulandum. The major findings of these studies are as follows (**Figure 4**):

1. When exposed to a complex but deterministic field of velocity-dependent forces, arm movements are at first distorted, but the initial kinematics are recovered after repeated practice.
2. If, after adaptation, the field is suddenly removed, aftereffects are clearly visible as mirror images of the initial perturbations, as described by eqn [12].

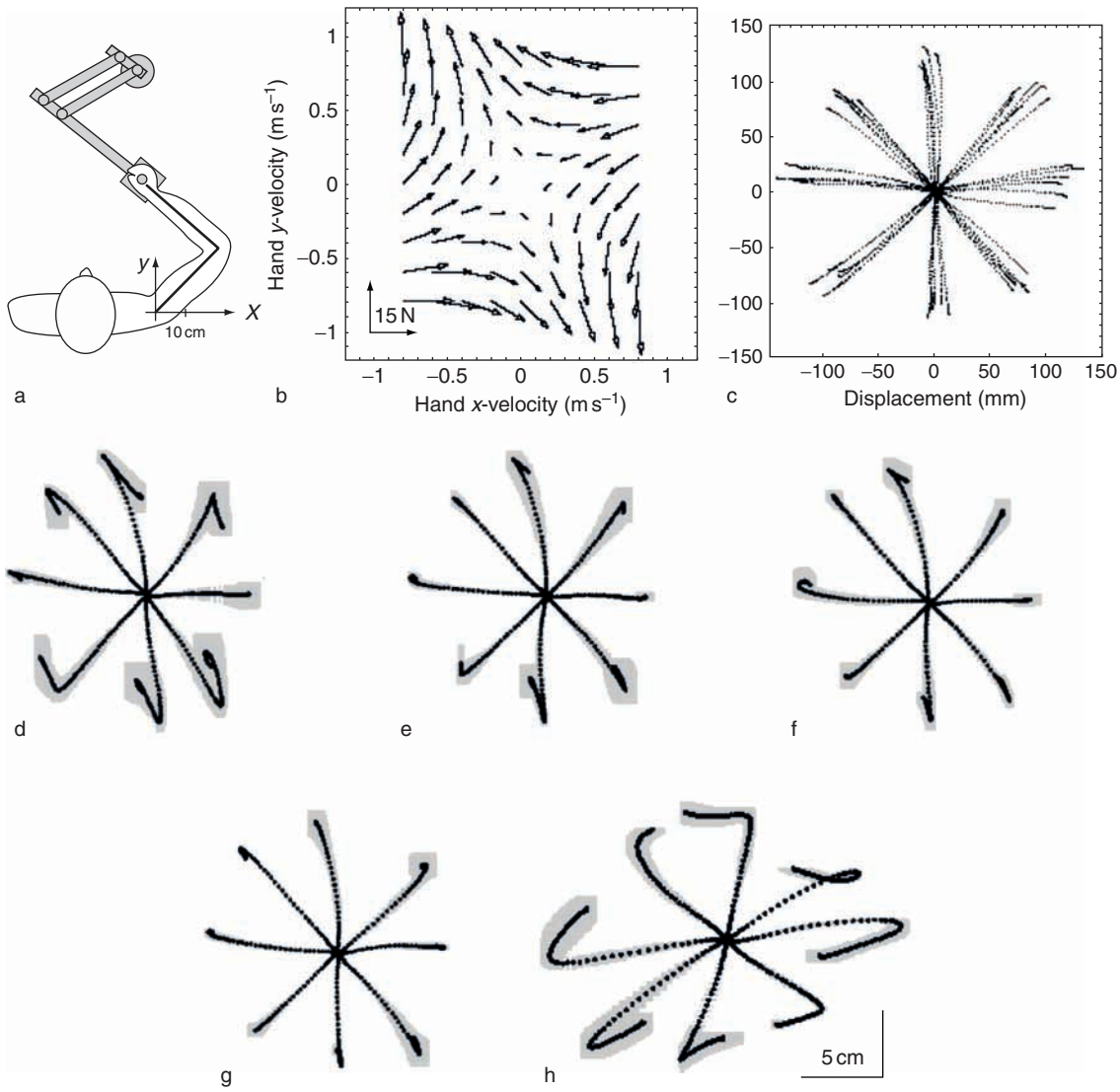

Figure 4 Force field adaptation experiment. (a) Subjects hold the end effector of a two-joint planar robot and execute reaching movements to a set of targets. (b) The robot can generate a force that depends on the measured velocity of the end effector. (c) Typical trajectories generated by a subject when the robot is not generating any force field. (d–g) Average ± standard deviation of hand trajectories during the training period in the force field (b). The four panels represent the performance during four consecutive sets of 250 targets. (h) Aftereffects of adaptation. These trajectories were obtained at the end of the training period by exposing the subject to random trials in which the field was unexpectedly removed. Note that the shape of these trajectories is approximately the mirror image of the trajectories in (d). Modified from Shadmehr R and Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience* 14(5): 3208–3224, with permission.

3. Adaptation is achieved by the central nervous system through the formation of a local map that associates the states (positions and velocities) visited during the training period with the corresponding forces.
4. This map, which is the internal model of the field, undergoes a process of consolidation after adaptation.

In order to generate the appropriate command for a desired movement of the arm, an inverse model must take into account that multijoint inertia depends on limb position and velocity. Therefore, an inverse model must be informed about the current state of motion of the limb. This information may come from (1) the input to the inverse model that specifies where the limb should be or (2) a prediction of the current state based on delayed sensory feedback and on the past history of motor commands. Bhushan and Shadmehr found compelling evidence for the second way. Their experimental results are consistent with the hypothesis that we learn to compensate changes in limb dynamics by a process that involves the combined adaptation of a forward and an inverse model of the limb.

State Space Models of Motor Learning

The process of learning requires practice. In this respect, the formation of an internal model can be described as the temporal evolution of a dynamical system. This is the point of view taken by Thoroughman and Shadmehr and others. Thoroughman and Shadmehr proposed to describe the adaptation to a force field as the temporal evolution of a system whose internal state is characterized by a hidden, unobservable variable z. The system produces an observable output to which a 'performance error' y can be associated. An input signal p describes the applied perturbation through binary values that indicate the presence ($p = -1$) or absence ($p = +1$) of the applied field. The proposed state space model is a quite simple because it involves only linear first-order dynamics over discrete time steps counted by the index n:

$$\begin{aligned} z_{n+1} &= az_n + bp_n \\ y_n &= z_n + dp_n \end{aligned} \qquad [16]$$

Despite its simplicity, the model proved adequate to capture a number of observed features of the learning process. One such feature is the time course of adaptation, as captured by the deviation of each movement from a straight line; this deviation was taken to be the observable error y. Thoroughman and Shadmehr noticed that following each 'catch trial' – a trial in which the force field is unexpectedly removed – the experimental data indicated a brief but distinguishable unlearning period. The time course of this effect was

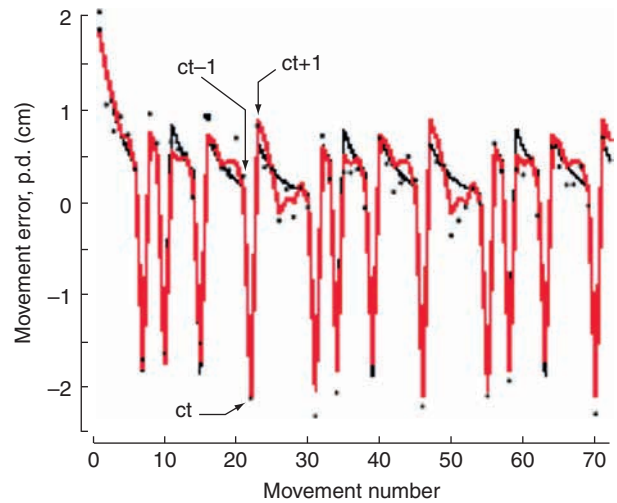

Figure 5 Unlearning effects captured by a state-space model. The plot depicts the time course of error in subject reaches in a particular direction. The circles are the errors (maximum lateral deviations from straight line) measured in each movement during training in the field. The field was suppressed at random trials during this experiment. When the field was suppressed in a 'catch trial', there was an aftereffect, indicated by a large negative error spike. Following a catch trial, the error returned at a level higher than the level immediately preceding the catch trial. Thus, catch trials cause unlearning. This effect is captured by state-space models of learning both in a scalar (black line) and in a vector (red line) form. Reproduced from Thoroughman KA and Shadmehr R (2000) Learning of action through adaptive combination of motor primitives. *Nature* 407: 742–747.

adequately captured by the state space model (Figure 5). Another important feature of this model is expressed by the value of the parameter b, which represents the effect on the current state (and therefore on the current error) of having experienced a force in the previous state. By estimating from the empirical error data the dependence of b on the direction of the previous movement, Thoroughman and Shadmehr were able to derive 'tuning curves' which demonstrate a destructive interference from previously experienced forces when there is a major difference between the current and the previous direction of movement.

An intriguing feature of this state space model is its low order as a dynamical system. All that seems to matter is the experience that the learner has acquired in the previous repetition of the task. This conclusion is supported by the observations of Scheidt and collaborators, who exposed a group of subjects to randomly varying force fields. In each trial, subjects were exposed to velocity-dependent forces perpendicular to the direction of motion. The amplitude of the force was related to movement speed by a gain factor B. The actual value of this factor was randomly drawn on each trial from a Gaussian distribution. With practice, the best performance of the subjects (zero error) was achieved for the mean value of the perturbing gain factor (Figures 6(a) and 6(b)).

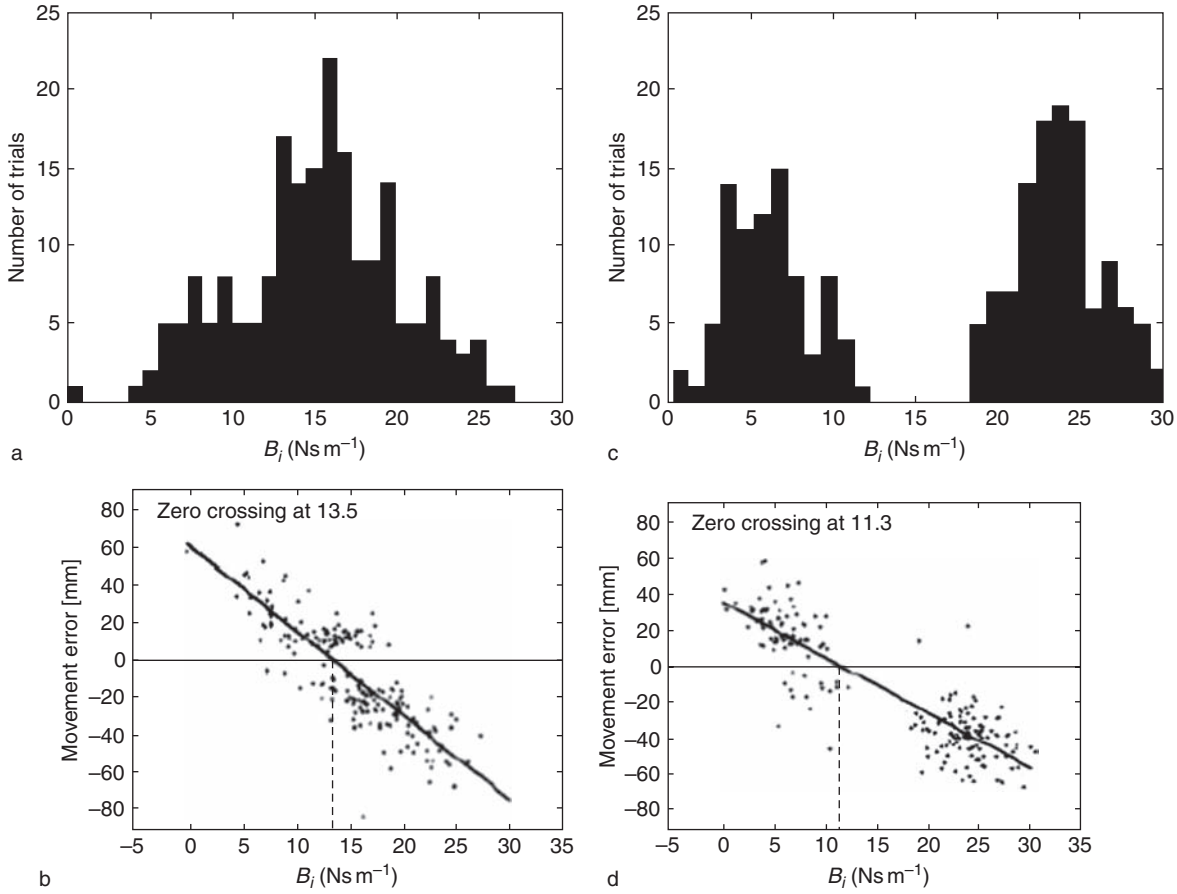

Figure 6 Adaptation to random force fields. Subjects were asked to move the hand against a velocity-dependent force field, whose gain varied randomly from trial to trial. For one group of subjects, the gain was drawn from a Gaussian probability distribution (a). The scatter plot of movement error versus field gain (b) exhibited a nearly linear relation ($r = 0.82$). Note that the zero crossing is near the mean value of field gain: subjects learned the mean field. A second group of subjects was exposed to a random perturbation that followed a bimodal distribution (c). The scatter plot of movement error versus field gain showed a linear trend ($r = 0.85$) and the zero crossing was again in the vicinity of the mean value of the gain (d). This is remarkable because subjects learned to compensate the mean perturbation, which in this case was never encountered during learning.

Remarkably, the same result was observed when the random gain factor was drawn from a bimodal distribution (Figures 6(c) and 6(d)). This is apparently paradoxical because in the bimodal case the mean value of the gain factor B is never experienced. The situation is reminiscent of the Buridan ass, which died of starvation because he could not chose between two identical stacks of hay placed in opposite directions. However, paradoxical as it may seem, a strategy based on the estimated mean value of the gain factor B is optimal in the sense of minimizing the mean square error.

These experimental observations indicate the existence of an underlying linear mechanism that guides the learning process. This suggests the possibility of using a simple linear time series – an autoregressive model with external input – to describe the learning behavior:

$$y_i = \sum_{j=1}^{L} a_j y_{i-j} + \sum_{k=0}^{M} b_k B_{i-k} \qquad [17]$$

In this model, y_n and B_n are, respectively, the error (deviation from linearity) and the input (field gain) at the nth trial. An estimation of the time windows L and M of the autoregressive process showed that the error at the current trial is only affected by the current input and by the immediately preceding input and error. The previous equation can then be rewritten as

$$y_i = a_1 y_{i-1} + b_0 B_i + b_1 B_{i-1} \qquad [18]$$

A simple transformation establishes the equivalence between this $L = M = 1$ model and the formulation of Thoroughman and Shadmehr shown in eqn [16].

See also: Bayesian Models of Motor Control; Computational Approaches to Motor Control; Kinematics and Dynamics; Motor Skill Learning; Motor Primitives; Multisensory Convergence and Integration; Neural Integrator Models; Pattern Generation; Sensorimotor Control of Manipulation; Spinal Motor Neurons: Properties.

Further Reading

Bernstein N (1967) *The Coordination and Regulation of Movement*. Oxford: Pergamon.

Bizzi E, Mussa-Ivaldi F, and Giszter S (1991) Computations underlying the execution of movement: A biological perspective. *Science* 253: 287–291.

Flanagan J and Wing A (1997) The role of internal models in motion planning and control: Evidence from grip force adjustments during movements of hand-held loads. *Journal of Neuroscience* 17: 1519–1528.

Flash T and Hogan N (1985) The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience 5*: 1688–1703.

Goldstein H (1980) *Classical Mechanics*. 2nd edn. Reading, MA: Addison-Wesley.

Kawato M and Wolpert D (1998) Internal models for motor control. *Novartis Foundation Symposium* 218: 291–304.

McIntyre J, Berthoz A, and Lacquaniti F (1998) Reference frames and internal models. *Brain Research: Brain Research Reviews* 28(1–2): 143–154.

Morasso P (1981) Spatial control of arm movements. *Experimental Brain Research* 42: 223–227.

Mosier KM, Sceidt RA, Acosta S, and Mussa-Ivaldi FA (2005) Remapping hand movements in a novel geometrical environment. *Journal of Neurophysiology* 94: 4362–4372.

Mussa-Ivaldi FA and Bizzi E (2000) Motor learning through the combination of primitives. *Philosophical Transactions of the Royal Society of London, Series B* 355: 1755–1769.

Mussa-Ivaldi FA and Hogan N (1991) Integrable solutions of kinematic redundancy via impedance control. *International Journal of Robotics Research* 10: 481–491.

Sciavicco L and Siciliano B (2000) *Modeling and Control of Robot Manipulators*. New York: Springer.

Shadmehr R and Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience* 14(5): 3208–3224.

Sutton R and Barto A (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Thoroughman KA and Shadmehr R (2000) Learning of action through adaptive combination of motor primitives. *Nature* 407: 742–747.

Sensorimotor Plasticity and Control of Movement Following Spinal Cord Injury +D18

V R Edgerton and R R Roy, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Published by Elsevier Ltd.

Comparison of the Control of Movement in the Intact Neuromuscular System and after a Spinal Cord Injury with and without Sensory Input

Normally, movement is controlled with a high level of precision as a result of a combination of highly integrated input between the brain and the spinal cord and the peripheral sensory information associated with the movement. All these components are important elements in the control of normal movement. However, perhaps due to the extensive redundancy and the intrinsic plasticity within the neuromuscular system, some of these components can be lost and reasonably effective movements still can be performed.

It is clear that locomotor-like activity can be generated by the spinal cord in the absence of any input from the brain when the peripheral sensory information that is generated during the movement is available to the spinal cord. Just as impressive is the fact that the spinal cord can generate coordinated motor outputs that share many of the features of normal locomotion, even in the absence of supraspinal or peripheral sensory oscillatory input. The isolated spinal cord can generate repetitive steplike cycles for hours with an impressive level of coordination of the flexor and extensor muscles that normally control the hindlimb joints. This is particularly evident when the spinal circuitry is at a high tonic and excitable state, a state usually induced in the laboratory pharmacologically, by tonic brain stem stimulation, by tonic spinal cord stimulation, and/or by tonic peripheral nerve or dorsal root stimulation. This oscillatory and coordinated flexor–extensor output of the spinal cord in the absence of oscillatory supraspinal or peripheral input is referred to as fictive locomotion. It should be noted, however, that the functional capability of the spinal cord in the absence of supraspinal input is even greater when sensory input is present. For example, when the spinal cord has access to the sensory inflow associated with weight-bearing stepping, it can adjust to varying loads and speeds placed on the hindlimbs, can perform complicated motor responses to unpredictable events such as stepping over obstacles, and can learn to perform specific motor tasks when trained appropriately over a period of weeks.

With respect to spinal cord injury and the practical application of the ability of the spinal cord isolated from the brain to perform useful movements, the spinal cord must have access to the cutaneous and proprioceptive input that provides the information that enables the spinal cord to generate functionally useful weight-bearing stepping. The role of sensory input is emphasized because of its critical importance in rehabilitative strategies designed to recover motor function following a spinal cord injury. The spinal cord circuitry cannot generate effective locomotion in the absence of sensory input, and it cannot learn to step or to stand in response to training. Fortunately, the pathways for normal sensory input to the spinal cord are preserved following most spinal cord injuries. Equally fortunate is the fact that the spinal circuitry and neurons necessary to interpret and generate the appropriate motor patterns are located within the lumbosacral spinal cord, a region that is not usually damaged when a spinal cord injury occurs.

An exception to the preservation of the lumbosacral segments following a spinal cord injury is when there is direct damage to the dorsal and ventral roots, as can occur from an impact on the sacrum as in falling, often resulting in damage to the cauda equina. In this case, both the sensory and motor output to the muscles of the hindlimb usually can be critically damaged. The major barrier to motor recovery in these cases is the magnitude of the damage, that is, the number of motor neurons and sensory fibers damaged. The critical questions are: (1) are there enough intact motor neurons remaining to generate sufficient force to accomplish weight-bearing stepping or standing? and (2) is there a sufficient number of sensory fibers that can provide adequate information to contribute to the control of these movements? The functional prognosis when there is an injury to the cauda equina, in some cases, can be better than for injuries at a higher level because the supraspinal control of the remaining spinal circuitry can effectively compensate for a significant loss of neurons in the damaged sacral area. In contrast, the loss of function following injuries within the thoracic or cervical regions of the spinal cord is usually not due to the loss of neurons associated with motor output or with sensory input but, instead, is related to the lack of supraspinal control of those neurons persisting below the lesion that continue to have the potential to generate effective standing and stepping.

Why Is the Sensory Input to the Spinal Cord So Crucial When the Spinal Circuitry Alone Can Generate a Coordinated Output?

As noted previously, in order to be of practical significance in the rehabilitation setting, a critical feature of the spinal circuitry is its ability to receive and interpret online and very accurately the ensemble of sensory information that it is provided continuously. For example, the interneuronal components that are necessary to activate the flexor muscles at the appropriate level and time during a step cycle must recognize the sensory input that is associated with each phase of the step cycle. In this case, the initiation of extension of the plantarflexors of the foot and the sensory input associated with the first phase of foot contact during stance must know that the end of the swing phase has just occurred and, furthermore, must know which muscles to activate next. That is, those neurons associated with a given phase of a motor event recognize the ensemble of sensory input that is normally associated with that motor event (e.g., the specific phase of a step cycle). But it is important to recognize that these sensorimotor circuits are not acting as simple reflexes (i.e., stimulus-response during weight-bearing stepping). In essence, the sensory enables the spinal circuitry to interpret and anticipate the motor event that will be initiated next during a given phase of the step cycle.

In addition to being able to generate routine repetitive stepping, the spinal circuitry is highly capable of rapid adaptability. For example, it can adapt the stepping mechanics in spinal cats when there is an attempt to prevent the hindlimb from moving forward during the swing phase of a step. More specifically, when a stimulus is provided to the dorsum of the paw of the spinal cat during the flexion phase of a step cycle, ipsilateral flexion is enhanced, thus enabling the cat to step above an object in its path. In contrast, when the same stimulus is applied during the stance phase of a step cycle, the duration and magnitude of extension is enhanced. These experiments demonstrate that the motor output generated depends on the functional state of the spinal cord and, more specifically, on the phase of the step cycle. This is an obvious example of decision making or interpretation made by the spinal circuitry based on temporal sensory input, pointing out that the neural processing that occurs within the spinal circuitry reflects much more sophisticated circuitry than is typically associated with spinal reflexes. There are similar examples of the importance of sensory information derived from the limbs during stepping in human subjects. For example, the level of activation of extensor muscles during stepping is modulated directly with the level of loading. Even more interesting is the observation that this modulation of level of activation of extensor muscles with load changes is similar whether the subject is nondisabled (Figure 1(a)), has a severe spinal cord injury (Figure 1(b)), or even has a complete spinal cord injury (Figure 1(c)).

As already stated, it is apparent that the motor responses generated by the spinal cord are a function of the ensemble of sensory information and the physiological state of the spinal circuitry. The degree to which

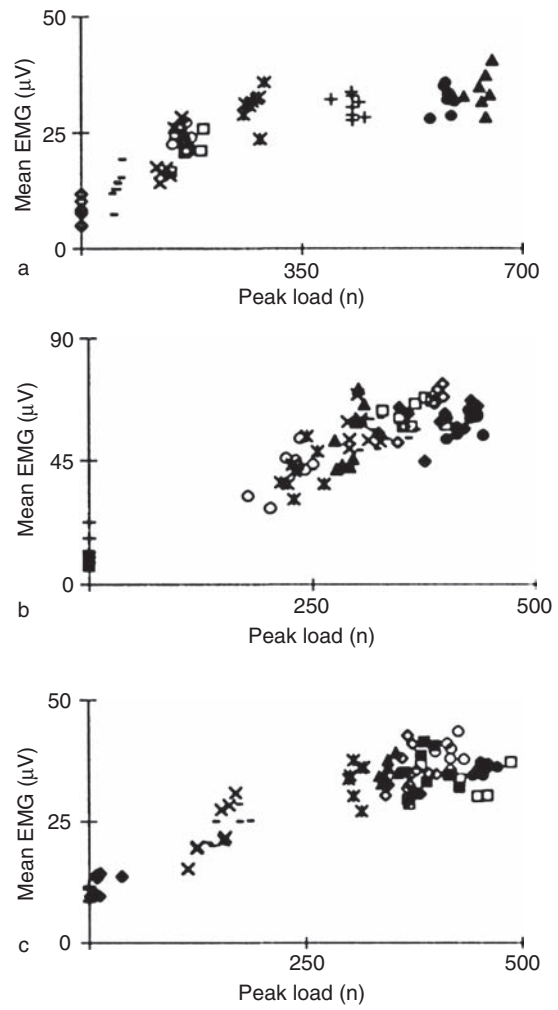

Figure 1 The effects of changes in levels of loading on the lower limbs on the level of activation of an extensor muscle of the ankle (i.e., the soleus) when stepping on a treadmill: (a) nondisabled subject; (b) subject with a severe but incomplete spinal cord injury, classified as ASIA C; (c) subject with a complete spinal cord injury, classified as ASIA A. The loading is controlled through a suspension apparatus, with the subject wearing a vest similar to that used during mountain climbing. The amplitude of the EMG increases similarly with increasing loading for all subjects. EMG, electromyogram. Modified from Harkema SJ, Hurley SL, Patel UK, et al. (1997) Human lumbosacral spinal cord interprets loading during stepping. *Journal of Neurophysiology* 77: 797–811.

this response depends on an ensemble of information versus a specific response of a specific type of receptor, however, remains to be determined. Although there is increasing evidence that the spinal circuitry, as well as the brain circuitry, responds to profiles of sensory information rather than to a specific receptor inducing a specific response, the critical experiments that confirm this hypothesis have not been performed. An analogy to this ensemble concept might be that the individual letters within this paragraph have no meaning in themselves, but when the letters are combined in certain patterns (words, phrases, sentences, and paragraphs) complex meanings or states can be derived by the brain. A similar analogy can be conceptualized in processing visual information. For example, an individual pixel of a high-density photograph has little or no meaning, but when all the pixels are in their normal spatial positions complex meanings and responses are generated.

Physiological State of the Spinal Cord

Pharmacological Modulation

After a complete thoracic spinal cord injury in cats, rats, and mice, the spinal circuitry necessary for generating full weight-bearing stepping remains anatomically intact. For example, several months after a complete spinal cord transection in an adult cat that had not recovered any stepping ability, weight-bearing steps over a range of speeds on a treadmill could be generated within 30 min after the administration of strychnine, a glycinergic receptor blocker that results in less inhibition, and therefore a greater level of excitability, of the spinal circuitry. Although it is not entirely known which neurotransmitter receptors are actually affected by this pharmacological intervention, spinal inhibitory circuits are mediated by glycinergic and γ-aminobutyric acid (GABA)ergic neurotransmitters. These neurotransmitters usually are co-expressed at the same terminals and play an important role in the coordination of agonist–antagonist motor pools. Our interpretation of experiments in which spinal cats were administered strychnine is that it has a facilitatory effect on the sensorimotor pathways involved in stepping. In general, the hindlimbs were flaccid prior to the administration of the inhibitory antagonist. We suggest that the spinal circuitry needed to generate weight-bearing stepping was present but that this circuitry was unable to interpret the sensory input until its general level of excitability was elevated. The physiological state of the circuitry was modified significantly by this pharmacological treatment. Similar conclusions can be drawn with the administration of other pharmacological agents. For example, Goldberger and colleagues reported that bicuculline, an inhibitory antagonist for GABAergic neurotransmission, had an effect similar to strychnine in facilitating locomotion in spinal cats. More recently, similar effects were observed in complete spinal adult mice given the serotonergic agonist quipazine. The general concept illustrated by these observations is that the physiological state of the spinal cord circuitry can be readily modulated. A very important feature of this concept that evolved from the most recent experiments, however, is that these pharmacological interventions need not, and actually preferably do not, induce the locomotion themselves but, instead, facilitate locomotion by enabling the spinal circuitry to interpret the cutaneous and proprioceptive information associated with specific phases of stepping and standing. This enabling phenomenon potentially can allow an individual with a spinal cord injury, particularly paraplegia, to manipulate the position of the legs and hip joints for generating weight-bearing stepping and sensory input.

Modulation of the Physiological State of the Spinal Cord with Epidural Stimulation

The physiological state of the spinal cord also can be readily modulated by tonic stimulation of the lumbosacral spinal cord via epidurally located electrodes. As with pharmacological interventions, the stimulation itself can be administered at an intensity level and frequency that does not induce stepping unless sensory signals associated with weight-bearing stepping are provided, as occurs when the hindlimbs are placed on a treadmill belt. Thus, these experiments also demonstrate the changing of the physiological state of the spinal cord so that the sensory information can reach those spinal circuits that have the capability to accurately interpret the sensory ensemble associated with specific phases of the weight-bearing state of the hindlimbs and then generate the appropriate motor response. Based on the success of prior experiments using epidural stimulation over the lumbosacral spinal cord in mice, rats, and cats, this approach represents an exciting potential for enhancing the recovery of stepping and standing in human subjects following a complete spinal cord injury. We have recently demonstrated that epidural stimulation also has been successfully applied to facilitate rhythmic activity and improve the training effect in chronic spinal rats.

In humans, spinal cord electrical stimulation has been successfully used for the suppression of pain and spasticity. To control severe spasticity, an optimal location of the electrodes was reported to be between L1 and L3, with a stimulus frequency of 50–100 Hz. It appears that the optimal stimulation parameters for

the control of spasticity and for initiating steplike electromyography (EMG) differ. A nonpatterned tonic stimulation at a frequency of 25–50 Hz elicited rhythmic, alternating EMG patterns in individuals with chronic complete spinal cord injury. Recent studies have suggested improved overground ambulation following a combined treatment of body-weight-support treadmill therapy and epidural stimulation in a chronic incomplete spinal-cord-injury subject.

Changing the Physiological State of the Spinal Cord by Combining Pharmacological Agents, Epidural Stimulation, and Motor Training

Other, more long-term modulation techniques can be used to change the physiological state of the spinal cord. Step and stand training represent two examples in which the repetitive activation of specific sensorimotor pathways results in a reinforcement of the order of activation of these pathways, resulting in a higher probability of generating successful weight-bearing standing or stepping.

In developing more effective rehabilitative strategies following spinal cord injury, it appears that a productive direction will be to identify how we can modulate the physiological state of the spinal cord in a highly predictable manner using a combination of electrical stimulation techniques, which can be rather noninvasive, and modulation via pharmacological cocktails targeting specific inhibitory and excitatory pathways. Both electrical stimulation and pharmacological modulation have been shown to be effective ways to enhance concomitant rehabilitative procedures in achieving an improved ability to step. Similar procedures are likely to be effective in improving standing.

Improved Coordination Is an Important Component of Learning to Step as a Result of Step Training Following a Complete Spinal Cord Transection

After a complete mid-thoracic spinal cord transection in an adult animal, the motor pools projecting to the muscles of the hindlimb remain active although at a lower than normal level. In addition, the activation of these motor pools is not well coordinated. This is shown by the higher incidence of co-contractions of the motor pools that normally function as agonists and antagonists after a spinal cord injury. Experiments in animals and humans suggest that step training can improve the level of coordination of agonistic and antagonistic muscles to a level more typical of normal than when there is no training. Improved coordination between the two hindlimbs in spinal cats is an important component of successful stepping. For example, we observed that one of the primary factors contributing to an unsuccessful step (i.e., continuous stepping fails when the hindquarters can no longer be supported), was a gradual shift in the coordination of the two hindlimbs, resulting in a shorter duration of double-weight support to the point at which both legs were in a non-weight-supporting position and the hindquarters collapsed. The results of a number of other studies examining the coordination within and across limbs in spinal animals and in humans with a spinal cord injury are consistent with the interpretation that an important aspect of step training is the improved coordination of the motor pools that generate stepping. This lack of precise timing of the normal pattern of the activation of the motor pools after a spinal cord injury is manifested as limb movements that are highly variable and, therefore, less predictable (Figure 2). Step training has been shown to reduce this variability by improving the preciseness of the motor pool coordination.

The Role of Muscle Plasticity in the Recovery of Locomotion Following a Complete Spinal Cord Injury

After a complete spinal cord transection, the muscles below the level of the lesion atrophy and generally shift in their functional and metabolic properties toward those of a faster muscle. These adaptations are observed to some extent in most muscles below the lesion, but the magnitude of the changes appears to be greater in predominantly slow extensor muscles (vastus intermedius, adductor longus, and soleus) than in predominantly fast extensor muscles (vastus lateralis, gastrocnemius, and plantaris) and greater in the extensor (vastus lateralis, gastrocnemius, plantaris, and soleus) than in the flexor (semitendinosus and tibialis anterior) muscles.

Motor training has some beneficial effects in restoring these properties toward normal levels in some muscles, but full recovery of muscle size and function does not seem to be an essential factor for improved motor performance. For example, the relative muscle mass and maximum tetanic tension of the soleus muscle of complete adult spinal cats that were weight-supported step trained for approximately 5 months were greater than in nontrained spinal cats but still lower than in normal control cats. Furthermore, these properties for the medial gastrocnemius muscle were similar in step-trained and nontrained spinal cats and were similar to normal control values for the tibialis anterior in both groups. Because the step-trained cats showed markedly better locomotor ability than the nontrained cats, the combined results

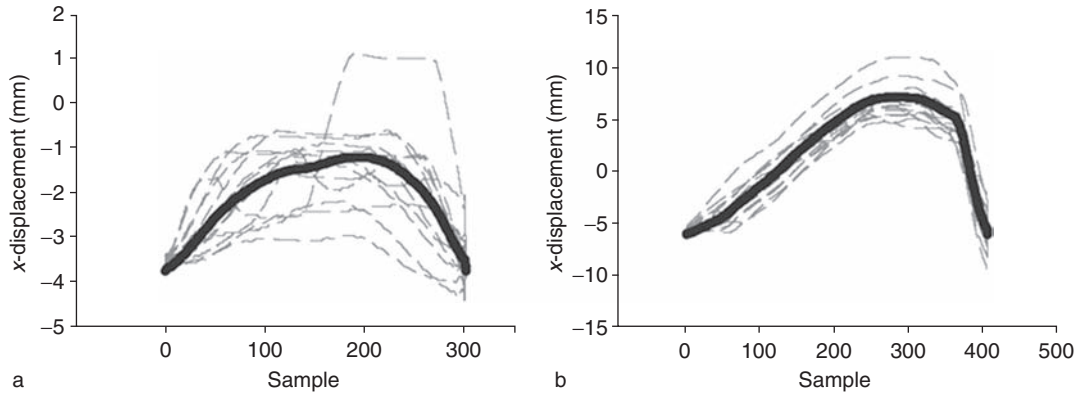

Figure 2 Plots of several *x* trajectories (dashed lines) and their corresponding first principal component (PC; solid lines): (a) for a quipazine-treated nontrained spinal (mid-thoracic, complete) mouse; (b) for a quipazine-treated robotically trained spinal (mid-thoracic, complete) mouse. The first PC from a principal component analysis (PCA) identifies the most representative step pattern executed by a mouse. A mouse with a high PCA percentage for its first PC stepped with a more consistent step shape than a mouse with a lower PCA percentage score. The PCA percentages, 67.9% (a) and 97.6% (b), respectively, indicate that quipazine facilitated a greater improvement in spatial stepping consistency in the trained mouse than the nontrained mouse. Note also from the different scales of the *x*-displacement (ordinate) axis that the trained mouse generally took much longer steps than the nontrained mouse. From Fong AJ, Cai LL, Otoshi, et al. (2005) Spinal cord-transected mice learn to step in response to quipazine treatment and robotic training. *Journal of Neuroscience* 25: 11738–11747.

indicate that adaptations in muscle properties alone cannot account for the improved locomotor ability of step-trained spinal cats. This is particularly evident when we consider that a decrease in ankle flexion (foot drop) during the swing phase of the step cycle is considered to be a primary deficit in generating a successful step, yet the mass and maximum tension capability of the tibialis anterior, the primary ankle flexor, is unaffected by a complete spinal cord transection. Thus, it appears that a combination of a critical mass and functional capacity of the musculature and adequate excitability of the appropriate spinal circuitry can contribute to an improvement in motor performance in complete spinal animals.

The size, metabolic, and functional properties of a skeletal muscle are highly dependent on the amount and pattern of neuromuscular activity imposed on that muscle. The actual amount of activation and/or loading required to maintain normal muscle properties, however, is undefined. Recent evidence from our and other laboratories indicates that these requirements may be minimal. For example, we have shown that as little as 1 min of high-load contractions (30 contractions, 1 s duration, performed twice per day) is effective in maintaining the size and force-generating properties of the otherwise inactive rat medial gastrocnemius muscle.

Dramatic Differences in the Mechanisms of Recovery Following Complete versus Incomplete Spinal Cord Injuries

Improved posture and locomotion can occur in complete spinal animals following training, as noted previously. As many as 10–15 consecutive weight-bearing steps have been generated successfully when human subjects with a complete mid-thoracic spinal cord injury were trained to step on a treadmill with some body-weight support. In human subjects, however, the success of independently regaining stepping capability with full weight bearing has been insufficient in individuals with a complete spinal cord injury. Thus, to date some weight-bearing assistance has been necessary for the steps to be generated consecutively by individuals with a clinically complete injury. It also has been reported, however, that an individual with a complete spinal cord injury learned to take some steps using a walker, through which considerable weight-bearing assistance and balance was provided by the subject's upper body.

On the other hand, human subjects with incomplete, even severe, spinal cord injuries can dramatically improve their locomotor ability. In a recent clinical trial, 90% of the subjects with incomplete injuries, 95% of whom could not generate any steps approximately 6 weeks following their injury, were able to step at nearly normal speeds following rehabilitation interventions that included weight-bearing step training. Similar recovery levels were observed in individuals who were trained using only overground step training compared to those who were trained on a treadmill with weight-bearing assistance. The important point is that the probability of recovering effective posture and locomotor function following a very severe, but incomplete, spinal cord injury is extremely high, assuming that appropriate rehabilitative and postinjury procedures are followed.

The results on motor recovery following a complete mid-thoracic spinal cord transection in mice and rats are more similar to that observed in humans than in cats with complete spinal cord injuries; that is, there is a more modest level of motor recovery unless the interventions involve pharmacological facilitation in combination with training. In contrast, partial lesions of the spinal cord in mice, rats, cats, and nonhuman primates result in near-normal recovery of posture and locomotion without any extensive training and rehabilitation. Thus, the specific mechanisms by which recovery occurs following a complete versus an incomplete lesion are fundamentally different. If the injury is incomplete, there is a possibility of reorganization of the descending control from the brain and of the spinal circuitry. With even a small amount of supraspinal input, the reorganization of the spinal circuitry allows for the remarkable levels of recovery seen in humans and laboratory animals. This phenomenon probably reflects not only the redundancy in the sensorimotor system, as noted earlier, but it also appears that novel pathways for sensorimotor control evolve, particularly with training. The important point in terms of rehabilitative strategies and of strategies to achieve novel reconnectivity of the appropriate neural circuits for generating posture and locomotion is that as long as some continuity (even very indirect) remains, the reorganization of existing circuits can be the basis for the extensive motor recovery. Thus, with incomplete injuries, reorganization can occur and reconnectivity can be achieved indirectly around the lesion by the expansion of existing noninjured neural circuits. This is not possible with a complete lesion because no noninjured descending pathways remain. After a complete lesion, on the other hand, (so far) to achieve full weight-bearing stepping new pathways by previously injured axons must grow through a major physical and chemical barrier. After an incomplete lesion, the existing circuitries can be enhanced, and these circuits can go around rather than though the lesion site. The reorganization of connectivity after a complete injury must involve regeneration through the lesion site.

Therefore, although remarkable and impressive levels of recovery are being reported regularly as a result of various types of surgical and pharmacological interventions, almost all of these experimental models involve incomplete lesions. The real challenge is to achieve improved function in humans with complete spinal cord lesions. Furthermore, in individuals with severe incomplete lesions, the degree to which we can take advantage of the minimal connectivity remaining above and below the lesion by using activity-dependent interventions to promote reorganization of this functional connectivity needs to be studied carefully.

Potential Advances on the Horizon

An exciting and growing list of widely varying strategies are being tested experimentally for their efficacy in facilitating the recovery of sensorimotor function after a spinal cord injury. There is a general consensus among scientists working on spinal cord injury that greater improvement can be achieved with multiple interventions. Theoretically, multiple strategies could be additive, particularly if the mechanisms involved with each intervention are different. Some of the strategies that have been reported to have positive effects, although very modest in most cases, are (1) modifying the properties of endogenous cells (i.e., altered gene expression); (2) transplanting cells that will replace those that are lost or transplanting cells that can generate molecules that have growth-promoting properties of the remaining axons (e.g., adult olfactory ensheathing cells; embryonic or adult stem cells); (3) transplanting segments of peripheral nerves to bridge the gap across the lesion; (4) transplanting embryonic tissue; (5) modifying the concentration of cyclic adenosine monophosphate (cAMP); and (6) modifying the extracellular matrix of the neuronal networks.

One of the most consistent observations has been that supraspinal and spinal plasticity is highly activity-dependent. Consistent improvement in motor function has been observed in different models of spinal cord injury, and it is likely that the effectiveness of each of the other interventions noted here will depend, in part, on whether there is an optimal level of neuromotor activity for the tissues that are involved in that intervention.

It cannot, however, be assumed that any given combination of interventions when used in concert with training will have a favorable outcome. Each intervention, including the activity-based therapies, ideally will be optimized with respect to dosage, time of administration, route of administration, and so on. But it will be even more difficult to identify the optimal activity-related behaviors when there are interactive effects of multiple interventions being used concomitantly. The challenge remains, however, to learn to use every mechanism of recovery so that their actions are complementary rather than competitive.

With continuing progress toward a better understanding of the basic biology of neural repair and of the mechanisms of the spinal control of locomotion and how it can reorganize in concert with supraspinal descending systems, it is likely that at least modest success in facilitating axonal growth across a complete spinal cord lesion can be achieved eventually. Even when this has been demonstrated, however, we still will not understand whether the connections will be of a sufficient magnitude or whether they will be

beneficial or disruptive with respect to pain and the control of movement. Based on the results obtained from incomplete lesions, it appears that the brain and spinal cord have a rather remarkable ability to reorganize in a functionally advantageous manner. The remarkable level of locomotor recovery that occurs after a lateral hemisection in mice, rats, and monkeys is consistent with this interpretation.

See also: Neuronal Plasticity after Cortical Damage; Reticulospinal System; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Regeneration and Functional Recovery: Strategies; Spinal Cord Injuries; Spinal Motor Neurons: Properties; Spine Plasticity.

Further Reading

Cai LL, Fong AJ, Otoshi CK, et al. (2006) Implications of assist-as-needed robotic step training after a complete spinal cord injury on intrinsic strategies of motor learning. *Journal of Neuroscience* 26: 10564–10568.

Edgerton VR, de Leon RD, Harkema SJ, et al. (2001) Topical review: Retraining the injured spinal cord. *Journal of Physiology* 533: 15–22.

Edgerton VR, Kim SJ, Ichiyama RM, Gerasimenko YP, and Roy RR (2006) Rehabilitative therapies after spinal cord injury. *Journal of Neurotrauma* 23: 560–570.

Edgerton VR and Roy RR (2002) Paralysis recovery in humans and model systems. *Current Opinion in Neurobiology* 12: 658–667.

Edgerton VR, Tillakaratne NJT, Bigbee AJ, de Leon RD, and Roy RR (2004) Plasticity of the spinal circuitry after injury. *Annual Review of Neuroscience* 27: 145–167.

Fong AJ, Cai LL, Otoshi, et al. (2005) Spinal cord-transected mice learn to step in response to quipazine treatment and robotic training. *Journal of Neuroscience* 25: 11738–11747.

Grillner S (2006) Biological pattern generation: The cellular and computational logic of networks in motion. *Neuron* 52: 751–766.

Harkema SJ, Hurley SL, Patel UK, et al. (1997) Human lumbosacral spinal cord interprets loading during stepping. *Journal of Neurophysiology* 77: 797–811.

Murray M and Fischer I (2001) Transplantation and gene therapy: Combined approaches for repair of spinal cord injury. *Neuroscientist* 7: 28–41.

Rossignol S, Dubuc R, and Gossard JP (2006) Dynamic sensorimotor interactions in locomotion. *Physiological Review* 86: 89–154.

Rossignol S, Giroux N, Chau C, et al. (2001) Pharmacological aids to locomotor training after spinal injury in the cat. *Journal of Physiology* 533: 65–74.

Roy RR, Baldwin KM, and Edgerton VR (1991) The plasticity of skeletal muscle: Effects of neuromuscular activity. In: Holloszy J (ed.) *Exercise and Sports Sciences Reviews,* vol. 19, pp. 269–312. Baltimore, MD: Williams and Wilkins.

Thuret S, Moon LD, and Gage FH (2006) Therapeutic interventions after spinal cord injury. *Nature Reviews Neuroscience* 7: 628–643.

Timoszyk WK, de Leon RD, London N, et al. (2002) The rat lumbosacral spinal cord adapts to robotic loading applied during stance. *Journal of Neurophysiology* 88: 3108–3117.

Wernig A, Muller S, Nanassy A, and Cagol E (1995) Laufband therapy based on "rules of spinal locomotion" is effective in spinal cord injured persons. *European Journal of Neuroscience* 7: 823–829.

Relevant Websites

http://www.christopherreeve.org – Christopher Reeve Foundation.
http://www.chnfoundation.org – Craig H. Neilsen Foundation.

Sensory Aging: Chemical Senses

R L Doty, University of Pennsylvania, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The senses of taste and smell are critical for esthetics, discerning the flavor of foods and beverages, and detecting environmental toxins, including spoiled foods, leaking natural gas, polluted air, and smoke. Although age-related losses are present in all sensory modalities, the chemical senses – particularly olfaction – seem to be unduly influenced by aging. While approximately 2% of the American population has demonstrable olfactory loss before the age of 65 years, this percentage rises to 50% for those between the ages of 65 and 80 years, and to 75% for those above this age range. This likely explains why a disproportionate number of the elderly die from accidental gas poisoning, and why many elderly report that their food lacks flavor. The latter problem can lead to decreased motivation to eat, nutritional deficiencies, depression, and, in rare instances, death.

The age-related declines in taste and smell function are influenced by a number of factors, including gender. For example, in the case of odor identification ability, men exhibit a greater age-related decline in function than do women, a sex difference that increases in magnitude in the later years of life. The reasons why women maintain better function than men into older age are poorly understood, although estrogen may be involved. If this is the case, estrogen's organization effects on the brain may be largely responsible, since some chemosensory-related sex differences are present before puberty.

Age-Related Alterations in Olfaction

The decline in olfactory function observed in the later years of life is, on average, quite large. This decrement is apparent on a wide variety of olfactory tests, including tests of odor detection, identification, discrimination, memory, and suprathreshold odor intensity perception. In addition, age-related declines have been noted in tests sensitive to chemically mediated somatosensory sensations derived from intranasal stimulation of the trigeminal cranial nerve (CN V), such as irritation, sharpness, coolness, warmth, and fullness. The changes in odor identification ability, as measured by the standardized University of Pennsylvania Smell Identification Test, and odor memory, as measured by a standardized Odor Memory Test, are shown in Figures 1 and 2, respectively. Such age-related changes are accompanied by alterations in brain activation, as indexed by functional magnetic resonance imaging (fMRI; Figure 3). The losses have been demonstrated in a number of different cultural groups, are much more pronounced than the influences of gender or smoking, and are present in the prepubescent period (Figure 4).

In general, the age-related decrease in sensitivity occurs for a spectrum of odorants, although some odors are more influenced than others, depending upon such factors as an individual's threshold to the odorant, whether multiple chemicals are involved in producing the odor, the number of receptor types activated by the odorant or odorant combination, and the nature of the psychophysical functions relating odorant concentration to perceived intensity. The tendency for threshold values of different odorants to be correlated suggests that a 'general olfactory acuity' factor exists, perhaps analogous to the general intelligence factor derived from items of intelligence tests. Interestingly, thresholds of a number of sensory modalities tend to be correlated with one another, as well as with several cognitive measures (e.g., verbal memory), implying that a 'general sensory acuity/cognitive factor' may also exist that subsumes intermodal correlations.

It is noteworthy that experience with odors can enhance olfactory sensitivity, suggesting the possibility that variation in age-related changes among people could reflect differences in prior experience with odorants. Improvement in olfactory function as a result of prior exposure is greater in women than in men, or in some cases may even be confined to women, implying that somehow the female brain is more plastic in this regard. The underlying physiological mechanisms regarding this phenomenon are poorly understood.

Age-Related Alterations in Taste

Commonly, persons who complain of taste loss have, in fact, altered olfactory function. This is because flavor sensations, such as chocolate, mint, pizza, apple, strawberry, and so on, are dependent upon stimulation of the olfactory receptors via the retronasal route. Blocking the nose while eating, for example, greatly decreases or eliminates such flavor sensations, since active movement of air into the nasal pharynx from the oral cavity is significantly impeded.

The sensations mediated by the taste system are largely limited to such qualities as sweet, sour, bitter, salty, and metallic. Unlike olfaction, which is mediated by a single cranial nerve (CN I), the taste system

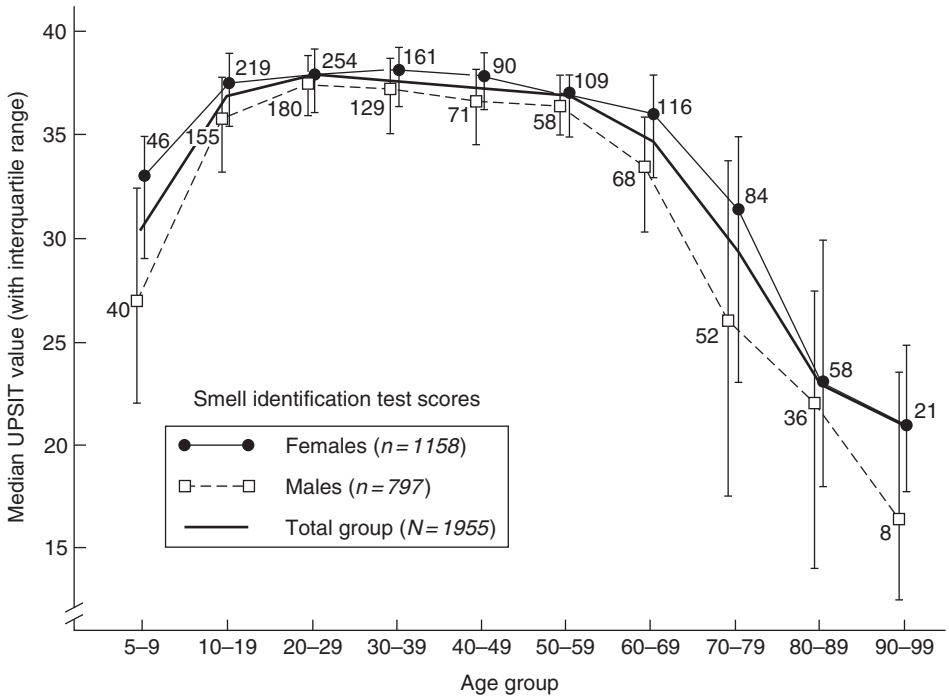

Figure 1 Scores on the University of Pennsylvania Smell Identification Test (UPSIT) as a function of age in a large heterogeneous group of subjects. Numbers by data points indicate sample sizes. From Doty RL, Shaman P, Applebaum SL, et al. (1984). Smell identification ability: Changes with age. *Science* 226: 1441–1443. Copyright 1984, American Association for the Advancement of Science.

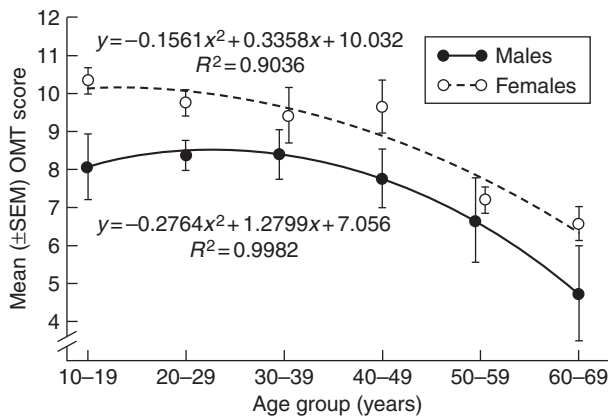

Figure 2 Scores on the Odor Memory Test (OMT) as a function of sex and age. Data averaged across 10, 30, and 60 s delay intervals. From Choudhury ES, Moberg P, and Doty RL (2003) Influences of age and sex on a microencapsulated odor memory test. *Chemical Senses* 28: 799–805. Copyright 2003, Oxford University Press.

is subserved by a number of cranial nerves: the facial nerve (CN VII), the glossopharyngeal nerve (CN IX), and the vagus nerve (CN X). Like olfaction, age-related decreases in gustatory function are well documented. However, when the whole mouth is evaluated (e.g., by "sip and spit" testing), the magnitude of the age-related loss is not as marked as that seen for the sense of smell. This is not true when small regions of the tongue are evaluated, where marked age-related declines are observed.

Decreases in taste threshold sensitivity have been reported for such tastants as caffeine, citric acid, hydrochloric acid, magnesium sulfate, propylthiourea, quinine, sodium chloride, sucrose, tartaric acid, and a large number of amino acids. As with odorants, not all taste qualities appear to exhibit the same degree of age-related loss. Thus, sweet taste sensation appears to be less ravaged by age than are salty, bitter, and sour sensations. Figure 5 shows the marked age-related decrement to NaCl seen in small regions of the tongue.

Changes in suprathreshold taste perception to chemicals as a result of aging have also been observed. For example, flatter functions relating tastant concentrations to perceived intensity have been reported in elderly persons relative to young adults, at least for some stimuli. However, the slope of such functions can be markedly altered by changes in intensity that occur at the extremes of the stimulus continuum, and may not reflect decreased sensitivity across the entire suprathreshold stimulus range. In a manner analogous to what occurs in olfaction, sex differences in taste function become more apparent in the later years.

Although the genetic influences on smell function are poorly established, a number for taste have been worked out. For example, the sensitivity to bitter-tasting agents is a heritable trait, presumably having

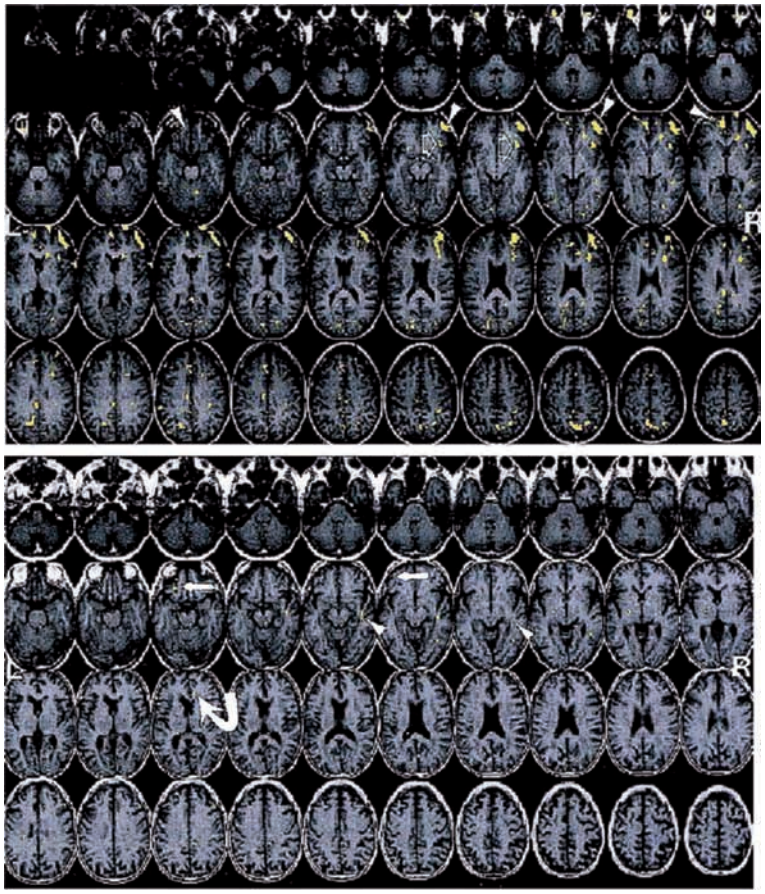

Figure 3 Odorant-induced fMRI activation of brain regions in a young (top) and an old (bottom) individual. In the image of the young person's brain, note the strong activation in both inferior frontal lobes (arrowheads), on the right more than left. Perisylvian activation inferiorly on the right (open arrows) is also noteworthy. In the bottom image(older person), note minimal activation in the left inferior frontal (straight arrows), right perisylvian (arrowheads), and right cingulate (curved arrow) areas. From Yousem DM, Maldjian JA, Hummel T, et al. (2003) The effect of age on odor-stimulated function MR imaging. *American Journal of Neuroradiology* 20: 600–608. Copyright 2003, American Society of Neuroradiology.

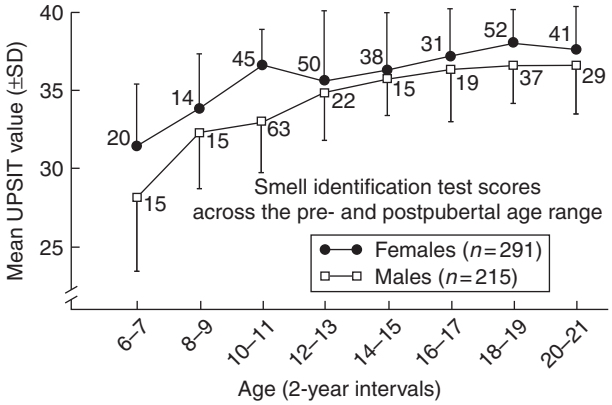

Figure 4 Scores on the University of Pennsylvania Smell Identification Test (UPSIT) as a function of gender across the prepubertal, adolescent, and young adult years. Note the consistently better performance of women than of men. From Doty RL (1986) Gender and endocrine-related influences on human olfactory perception. In: Meiselman HL and Rivlin RS (eds.) *Clinical Measurement of Taste and Smell*, pp. 377–413. New York: Macmillan Publishing Co. Copyright 1986, Macmillan Publishing Company.

evolved to protect against ingestion of poisonous toxins, such as a number of bitter-tasting plant-derived alkaloids and glycosides. The tastants that have received the most experimental attention in this regard are phenylthiocarbamide (PTC) and 6-*n*-propylthiouracil (PROP). About three times more Europeans and North Americans of European descent are PTC 'nontasters' than are Asians, Africans, and Native Americans. Tasters and nontasters exhibit different food preferences, as well as different preferences for hot spices. Research from the 1930s suggested that sensitivity to PTC followed a simple Mendelian pattern, with tasters being homozygous or heterozygous for the dominant gene, and nontasters having two recessive genes. In the 1980s, alternative genetic models proved to be a better fit to the data (e.g., a two-locus model in which one locus controls PTC taste perception and the other controls more general taste acuity). Recently, genes have been identified that account, to a large degree, for bitter taste perception.

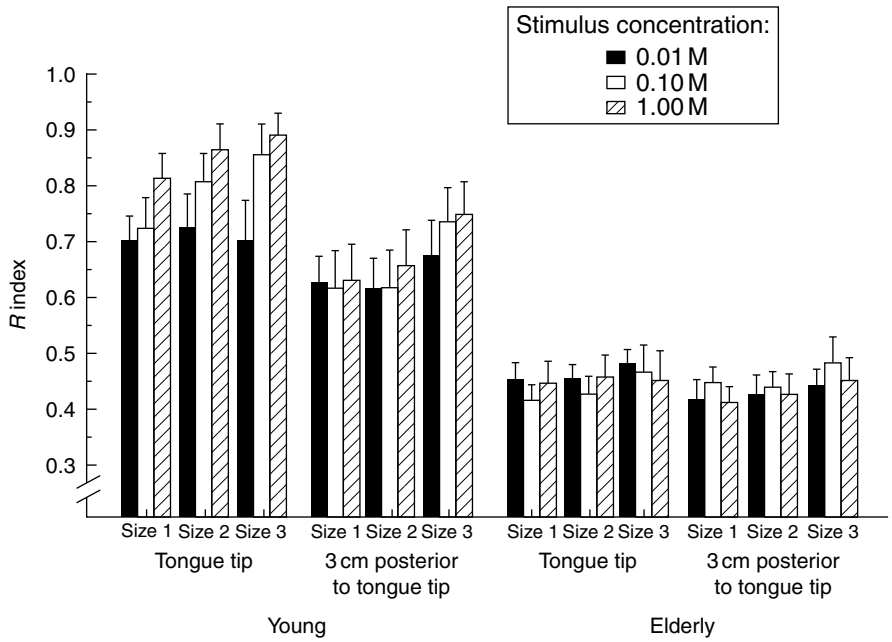

Figure 5 Detection sensitivity to NaCl in two age groups for three stimulus concentrations at two tongue loci for three different sizes of tongue regions. Size 1 = 12.5 mm^2; size 2 = 25 mm^2; size 3 = 50 mm^2. Means reflect scores on a nonparametric signal detection measure of olfactory sensitivity. Error bars represent +1 SEM. From Matsuda T and Doty RL (1995) Regional taste sensitivity to NaCl: Relationship to subject age, tongue locus and area of stimulation. *Chemical Senses* 20: 283–290. Copyright 1995, Oxford University Press.

Despite evidence for strong heritability, sensitivity to PTC and PROP declines with age. In one study, for example, the percentage of a group of 378 women who perceived PROP as being strong dropped from over 70% between the ages of 18 and 29 years to approximately 50% between the ages of 60 and 70 years. Corresponding values for men in the same age ranges were 53% and 30%. The complexity of the inheritance of the PTC taster gene is illustrated by a community living in an isolated mountain region of Ecuador. In this group, PTC nontasters eat maize-derived bread which contains a goitrogen, and consistently develop large unsightly goiters. PTC tasters, on the other hand, avoid such bread and do not develop such goiters. However, the PTC nontasters exhibited greater resistance to malaria, providing an explanation as to why the nontasting genes remain in the gene pool.

Alterations in Olfactory and Gustatory Function in Alzheimer's Disease

A number of age-related neurodegenerative diseases, most notably Alzheimer's disease (AD), exhibit marked olfactory loss relative to age-matched controls. Taste loss, while found in some studies, is much less marked. As with the influences of age, the olfactory decrements of AD are observed on numerous types of tests, and can be seen in functional imaging studies (Figure 6). A recent meta-analysis of 11 studies examining olfaction in AD found effect sizes ranging from 0.98 to 12.15 (median = 2.17), suggesting this is a robust phenomenon.

It has been argued that olfactory test scores are inversely related to the severity of dementia. However, such associations are weak. Moreover, they are enigmatic, given the influences of dementia on nonolfactory components of the tests. The AD-related olfactory dysfunction may reflect, to some degree, physiological changes associated with normal aging, although very early stage AD patients with mild dementia consistently score more poorly on olfactory tests than do age-matched controls. Recent studies suggest that smell loss may be occurring in the so-called preclinical stage of the disease process, and that it is more likely to occur in relatives than in nonrelatives of AD patients. In one study of 1604 non-demented community-dwelling senior citizens 65 years of age or older, a test of odor identification was a better predictor of future cognitive decline than was a global cognitive test. Persons who were anosmic and possessed at least one $\varepsilon 4$ allele of apolipoprotein E, a genetic risk factor for AD, had 4.9 times the risk of exhibiting cognitive decline than did persons with normal smell not possessing this allele. This is in contrast to the 1.23 times greater risk for cognitive decline in persons with normal smell who possessed this allele.

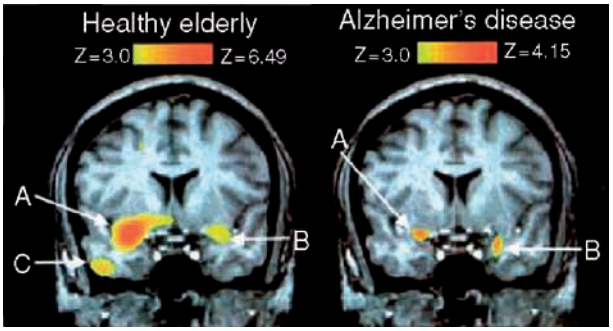

Figure 6 Left: Whole brain subtraction of olfactory stimulation minus control stimulation in healthy elderly control participants, co-registered on a T1-weighted reference magnetic resonance image ($N=7$). Right and left are reversed. Regions of interest are as follows: (A) right frontotemporal junction (piriform area), (B) left piriform area, and (C) right anterior–ventral temporal lobe (Brodmann 20). Right: Same whole brain subtraction in patients with Alzheimer's disease. Regions of interest are as follows: (A) 7 mm anterior to the frontotemporal junction (piriform) region in the control participants' image and (B) left amygdala–uncus. From Kareken DA, Doty RL, Moberg PJ, et al. (2001) Olfactory-evoked regional cerebral blood flow in Alzheimer's disease. *Neuropsychology* 15: 18–29. Copyright © 2001, American Psychological Association.

Causes of Olfactory Deficits in Aging and Alzheimer's Disease

The most common cause of age-related alterations in the ability to smell appears to be damage to the olfactory neuroepithelium from environmental insults (e.g., from viruses, bacteria, toxins, and pollutants). The dendritic knobs and the olfactory receptor-laden cilia are exposed rather directly to the external milieu, thereby being susceptible to nosogenic attack. In general, the olfactory neuroepithelium undergoes cumulative damage throughout life, exhibiting increasing numbers of islands of respiratory-like metaplasia soon after birth, decreased epithelial thickness, and decreased numbers of olfactory receptor cells. Some environments, such as cities or workplace environments polluted with airborne toxins, accelerate such pathology. Older persons may be more susceptible to epithelial damage than are younger ones, having a less resilient olfactory membrane due to such factors as reduced protein synthesis (as in hypothyroidism), loss of neurotrophic factors, altered vascularity and decreased intramucosal blood flow, altered nasal mucus viscosity, atrophy of secretory glands and lymphatics, and decreases in enzymes that deactivate xenobiotics within the olfactory mucosa. Some age-related deficits reflect subclinical neuropathology of the development of plaques and neurofibrillary tangles (NFTs) related to AD. Additionally, the number and cross-sectional area of the foramina of the cribriform plate are decreased in the elderly as a result of appositional bone growth, in effect pinching off the bundles of olfactory nerve axons as they traverse the foramina. In one study of this phenomenon, the mean area of the foramina was 47.3% less in men over the age of 50 years compared to those under the age of 50 years. The corresponding figure for women was 28.8%.

Despite the fact that olfactory receptor cells have some capacity to reconstitute themselves, this plasticity is compromised by age-related processes. Animal studies suggest that the ratio of dead or dying cells to the number of receptor cells increases with age, implying age-related decreases in mitotic activity. In older animals, epithelial repair is slower or nonexistent after damage by chemical agents such as zinc sulfate or methyl-formimino-methylester, unmasking an altered neurogenic process. Damage to the olfactory receptors leads to degenerative changes within the glomerular and mitral cell layers of the olfactory bulb. The number of mitral cells and glomeruli decreases steadily with age at an approximate rate of 10% per decade. By the ninth and tenth decades of life, less than 30% of these structures are present. *In vivo* quantitative magnetic resonance imaging studies confirm that the volume of the human olfactory bulbs and tracts decreases with age. Higher order structures associated with olfactory processing, such as the piriform cortex, show much less age-related change than do the olfactory bulb and epithelium. For example, in one study of rats, no significant age-related volumetric changes were found within cortical laminae Ia and Ib. The numerical and surface densities of the synaptic apposition zones in layer Ia, which are formed largely by mitral cell axons, were also not altered by age, although the proportion of layer Ia occupied by dendrites and spines was modestly decreased. An increase in the proportion of glial processes, but not in the proportion of axons and terminals, was also present.

The basis of the olfactory loss of patients with AD is poorly understood. No one has found within the olfactory epithelia of AD patients marker proteins or messenger RNAs that are specific to the disease. Nevertheless, tau-reactive and partially ubiquitin-reactive dystrophic neurites have been found in the lamina propria of AD patients. Both tau and ubiquitin contribute to the aggregation events associated with AD pathogenesis (e.g., NFTs). Recent behavioral studies in transgenic mice reveal that those that overexpress tau develop olfactory dysfunction, whereas those that overexpress β-amyloid, the major aberrant protein associated with neuritic plaques, do not. Interestingly, the $\varepsilon 4$ allele of apolipoprotein E is more than threefold greater in AD than in control olfactory receptor cells, and heat-shock proteins – proteins that are highly conserved and upregulated in cells

after exposure to stressors – are markedly downregulated in the olfactory epithelia of AD patients. This conceivably increases the susceptibility of the receptor cells to damage from xenobiotic agents.

There is considerable evidence that the olfactory bulbs and tracts are damaged early in the AD process, although the most salient early signs of AD neuropathology appear within transentorhinal cortical regions. The pathology then seems to spread from these cortical areas to other brain regions. Postmortem studies report up to a 40% reduction in the cross-sectional area of the olfactory tracts of individuals with AD, as well as ~50% loss of myelinated axons in these tracts. The AD-related NFTs and neuropil treads and PD-related Lewy bodies have been found in all layers of the olfactory bulb, with the exception of the olfactory nerve cell axon layer. Interestingly, the olfactory bulbs of over 40% of nondemented persons 50 years of age and older exhibit NFTs, suggesting such individuals may be in the earliest stages of developing the ultimate disease phenotype. Central limbic structures that receive olfactory bulb projections preferentially exhibit relatively high numbers of NFTs and neuritic plaques, including the hippocampus, the periamygdaloid nucleus, the prepiriform and piriform cortices, and the entorhinal cortex.

Among the more interesting and controversial theories concerning the etiology of the olfactory dysfunction in AD and some other neurodegenerative disorders (e.g., Parkinson's disease), as well as the etiology of the diseases themselves, is the 'olfactory vector hypothesis.' One form of this hypothesis suggests that the olfactory loss results from a xenobiotic agent (e.g., virus, bacterium, prion, pesticide) that enters the brain from the nasal cavity via the olfactory fila. Since the time of Galen it was known that dyes can move from the nasal cavity into the cerebrospinal fluid (CSF), and as early as 1912 it was shown that the olfactory neuroepithelium is a major route of entry of the poliovirus into the brain. The latter observation led to the discovery that cauterizing the olfactory mucosa of monkeys with picric acid and other caustic chemicals protected them against contracting polio. This led, in the late 1930s, to public health initiatives to prophylactically spray the noses of Toronto school children with zinc sulfate during polio epidemics.

The central-to-peripheral direction of the development of the neuropathology of AD need not be in opposition to the olfactory vector hypothesis. Thus, a xenobiotic could be transported to brain regions that are more vulnerable to its effects and could initiate genetically determined pathology in the susceptible region. The possibility also exists that damage to the olfactory system, for whatever reason, leads to degenerative alterations in brain structures that ultimately produce the cognitive alterations associated with AD. Limited support for this understudied concept comes from the observation that older rats that have been anosmic for a period of time have considerable difficulty in learning an active avoidance learning task, unlike older rats of the same age whose anosmia is of recent origin.

Causes of Taste Deficits in Aging and Alzheimer's Disease

Relative to olfaction, comparatively little research has been done to explain age- or AD-related alterations in gustatory function. As with other sensory systems, a relationship between the number of functioning receptor elements and the system's sensitivity is present. Thus, the perceived intensity of tastants presented to localized regions of the anterior tongue is correlated with the number of fungiform papillae and, hence, the number of taste buds within the stimulated regions. Individuals with higher densities of taste buds and fungiform papillae rate the intensity of a number of tastants as being 'more strong' than do individuals with lower densities of such structures. Age-related loss of taste buds has been found for human circumvallate papillae, although recent studies of rodent, monkey, and human tongues suggest that taste bud numbers on the fungiform papillae may be little influenced by age. For example, the average percentage of fungiform papillae-containing taste buds in Fischer 344 rats aged 4 to 6 months, 20 to 24 months, and 30 to 37 months was found in one study to be 99.6%, 99.3%, and 94.7%, respectively. A human study found no statistically meaningful relationship between age and taste bud densities on either the tip or the midregion of tongues from young adults (22–36 years, $N = 5$), middle-aged adults (50–63 years, $N = 7$), and old adults (70–90 years, $N = 6$). This work, however, is limited by its small sample sizes and marked variability in the number of taste buds on the tongue at all ages.

Although taste bud numbers appear not to be markedly decreased in old rats, electrical responsiveness of the chorda tympani nerve, which innervates the anterior tongue, is decreased to a number of salts, acids, and sugars. Among the possible explanations of this phenomena are decreased intrinsic reactivity of taste buds to taste solutions, decreased multiple neural innervation of taste buds by some taste fibers, alterations in the general structure of the epithelium (which, e.g., might impair the movement of stimulus solution into the taste bud pore), and decreased taste nerve responsiveness, *per se*. It is also possible that some taste buds, although anatomically present, are

not fully functional because of altered receptor cell turnover or related metabolic events. Since taste buds function in a complex ionic milieu and are bathed with saliva and other secretory products, changes in such products may also undergo age-related changes. There is suggestion that heightened taste threshold values and flattened suprathreshold psychophysical functions observed in many elderly persons reflect background tastes noticeable at low, but not at moderate or high, stimulus concentrations. Changes in both the neural and oral chemical milieu (e.g., salivary constituents) could contribute to this noise. Evidence that improved oral hygiene improves taste sensitivity in some elderly persons is in accord with this hypothesis.

Relatively few studies have examined the structure of the gustatory system of patients with AD. In one study, the innervation of the foliate and circumvallate taste buds was examined in five AD (three men, two women) and two control (both women) study participants, using antisera to protein gene product 9.5 (PGP 9.5), neuron-specific enolase (NSE), tyrosine hydroxylase (TH), dopamine β-hydroxylase (Dβ-H), and calcitonin gene-related peptide (CGRP). The mean number of PGP 9.5-immunoreactive intragemmal nerve fibers in taste buds of the foliate and circumvallate papillae was lower in the AD patients. Although very provocative, this study is limited by the small sample of individuals evaluated.

See also: Aging of the Brain; Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: MRI Studies; Central Gustatory System and Ingestive Behavior; Dementia; Flavor Physiology; Olfactory Receptor Genes: Human Loss During Evolution; Taste: Vertebrate Central Pathways.

Further Reading

Arvidson K and Friberg U (1980) Human taste: Response and taste bud number in fungiform papillae. *Science* 209: 807–808.

Baker H and Genter MB (2003) The olfactory system and the nasal mucosa as portals of entry of viruses, drugs and other exogenous agents into the brain. In: Doty RL (ed.) *Handbook of Olfaction and Gustation*, pp. 549–573. New York: Marcel Dekker.

Bartoshuk LM, Rifkin B, Marks LE, et al. (1986) Taste and aging. *Journal of Gerontology* 41: 51–57.

Braak H, Braak E, Yilmazer D, et al. (1996) Pattern of brain destruction in Parkinson's and Alzheimer's diseases. *Journal of Neural Transmission* 103: 455–490.

Choudhury ES, Moberg P, and Doty RL (2003) Influences of age and sex on a microencapsulated odor memory test. *Chemical Senses* 28: 799–805.

Cowart BJ (1981) Development of taste perception in humans: Sensitivity and preference throughout the life span. *Psychological Bulletin* 90: 43–73.

Curcio CA, McNelly NA, and Hinds JW (1985) Aging in the rat olfactory system: relative stability of piriform cortex contrasts with changes in olfactory bulb and olfactory epithelium. *Journal of Comparative Neurology* 235: 519–528.

Dalton P, Doolittle N, and Breslin PA (2002) Gender-specific induction of enhanced sensitivity to odors. *Nature Neuroscience* 5: 199–200.

Doty RL (1986) Gender and endocrine-related influences on human olgactory perception. In: Meiselman HL and Rivlin RS (eds.) *Clinical Measurement of Taste and Smell*, pp. 377–413. New York: Macmillan Publishing.

Doty RL (2003) Odor perception in neurodegenerative diseases. In: Doty RL (ed.) *Handbook of Olfaction and Gustation*, pp. 479–501. New York: Marcel Dekker.

Doty RL, Shaman P, Applebaum SL, et al. (1984) Smell identification ability: Changes with age. *Science* 226: 1441–1443.

Ferreyra-Moyano H and Barragan E (1989) The olfactory system and Alzheimer's disease. *International Journal of Neuroscience* 49: 157–197.

Hinds JW and McNelly NA (1977) Aging of the rat olfactory bulb: Growth and atrophy of constituent layers and changes in size and number of mitral cells. *Journal of Comparative Neurology* 72: 345–367.

Kalmey JK, Thewissen JG, and Dluzen DE (1998) Age-related size reduction of foramina in the cribriform plate. *Anatomical Record* 251: 326–329.

Kareken DA, Doty RL, Moberg PJ, et al. (2001) Olfactory-evoked regional cerebral blood flow in Alzheimer's disease. *Neuropsychology* 15: 18–29.

Kishikawa M, Iseki M, Nishimura M, et al. (1990) A histopathological study on senile changes in the human olfactory bulb. *Acta Pathologica Japonica* 40: 255–260.

Langan MJ and Yearick ES (1976) The effects of improved oral hygiene on taste perception and nutrition of the elderly. *Journal of Gerontology* 31: 413–418.

Loo AT, Youngentob SL, Kent PF, et al. (1996) The aging olfactory epithelium: Neurogenesis, response to damage, and odorant-induced activity. *International Journal of Developmental Neuroscience* 14: 881–900.

Matsuda T and Doty RL (1995) Regional taste sensitivity to NaCl: Relationship to subject age, tongue locus and area of stimulation. *Chemical Senses* 20: 283–290.

McBride MR and Mistretta CM (1986) Taste responses from the chorda tympani nerve in young and old Fischer rats. *Journal of Gerontology* 41: 306–314.

Mesholam RI, Moberg PJ, Mahr RN, et al. (1998) Olfaction in neurodegenerative disease: A meta-analysis of olfactory functioning in Alzheimer's and Parkinson's diseases. *Archives of Neurology* 55: 84–90.

Miller IJ and Reedy FE Jr. (1990) Variations in human taste bud density and taste intensity perception. *Physiology & Behavior* 47: 1213–1219.

Mistretta CM (1984) Aging effects on anatomy and neurophysiology of taste and smell. *Gerodontology* 3: 131–136.

Murphy C and Gilmore MM (1989) Quality-specific effects of aging on the human taste system. *Perception & Psychophysics* 45: 121–128.

Nakashima T, Kimmelman CP, and Snow JB Jr. (1984) Structure of human fetal and adult olfactory neuroepithelium. *Archives of Otolaryngology* 110: 641–646.

Ohm TG and Braak H (1987) Olfactory bulb changes in Alzheimer's disease. *Acta Neuropathologica* 73: 365–369.

Rosli Y, Breckenridge LJ, and Smith RA (1999) An ultrastructural study of age-related changes in mouse olfactory epithelium. *Journal of Electron Microscopy* 48: 77–84.

Schiffman SS, Hornack K, and Reilly D (1979) Increased taste thresholds of amino acids with age. *American Journal of Clinical Nutrition* 32: 1622–1627.

Schultz EW and Gebhardt LP (1937) Zinc sulfate prophylaxis in poliomyelitis. *Journal of the American Medical Association* 108: 2182–2184.

Smith CG (1942) Age incident of atrophy of olfactory nerves in man. *Journal of Comparative Neurology* 77: 589–594.

Stevens JC and Cain WS (1986) Aging and the perception of nasal irritation. *Physiology & Behavior* 37: 323–328.

Weiffenbach JM, Cowart BJ, and Baum BJ (1986) Taste intensity perception in aging. *Journal of Gerontology* 41: 460–468.

Yamagishi M, Takami S, and Getchell TV (1995) Innervation in human taste buds and its decrease in Alzheimer's disease patients. *Acta Oto-Laryngologica* 115: 678–684.

Yousem DM, Maldjian JA, Hummel T, et al. (1999) The effect of age on odor-stimulated functional MR imaging. *American Journal of Neuroradiology* 20: 600–608.

Sensory Aging: Hearing

M K Pichora-Fuller, University of Toronto, Mississauga, ON, Canada
E MacDonald, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Presbycusis

Causes

The term presbycusis, or presbyacusis, refers to age-related hearing loss. There are many causes of hearing loss in older people, including environmental factors such as exposure to noise and ototoxic drugs, genetic factors, and generalized effects of aging such as cell damage and neural degeneration. Many age-related hearing impairments are types of sensorineural hearing loss involving damage to the cochlea, or inner ear. The hallmark of this damage is the elevation of the thresholds of audibility for high-frequency pure tones.

Our knowledge of cochlear pathology, its effects on perception, and many issues pertaining to treatment and rehabilitation can be generalized across the adult age range. In addition to types of cochlear damage that are common to adults across the age range, the peripheral and central auditory systems of older adults may also be damaged in ways that are not typical in younger adults. Furthermore, older adults often exhibit perceptual deficits that are disproportionate to the problems observed in younger adults with similar hearing thresholds. Importantly, studies of several species of mammals have demonstrated that hearing declines as a function of chronological age, even when all genetic and potentially damaging environmental effects have been carefully controlled.

The classification of subtypes of presbycusis, defined according to the particular structures of the auditory system that are affected by age, has continued to be refined for over four decades. Unfortunately, there is not a straightforward correspondence between damage to particular structures and perceptual deficits. Damage at multiple sites likely contributes to the differences in auditory processing that are observed between younger and older adults with similar hearing thresholds. Although high-frequency pure-tone threshold elevation is very common in older adults, the more age-specific structural and perceptual aspects of presbycusis have been the focus of most research conducted in the past decade.

Prevalence

The prevalence of hearing loss increases with age and is greater in males than in females. The prevalence of presbycusis, based on audiometric pure-tone threshold data, ranges from almost 50% to over 80%, depending on the sample, age, and the definition of hearing loss. Some reports indicate that hearing loss is the third most common chronic disability in senior citizens. About three-quarters of those who have impaired hearing are over the age of 75 years.

Age-related hearing loss characterized by high-frequency threshold elevations can begin in the fourth decade of life. The degree of hearing loss and the range of frequencies affected continue to increase with age. The specific prevalence estimates based on pure-tone thresholds depend on the degree of loss and on which frequencies are used to determine the inclusion criteria. One explanation for the greater prevalence of hearing loss in males is that their cumulative exposure to noise, including industrial and military noise, is typically greater than that of females. Age cohorts may also differ in history of noise exposure associated with wars and changing industrial practices, as well as changes in longevity.

In industrialized countries, a century ago, fewer than 1 in 20 persons lived to age 65 years, but by 1980, there were 1 in 10, and by 2030, 1 in 5 persons will be over age 65 years. As the baby-boom generation enters retirement age and longevity increases, with the fastest growing age group now being those over the age of 85 years, because hearing loss increases with age, estimates of the prevalence of hearing loss increase for the coming decades.

In general, self-reported hearing loss yields prevalence estimates that are lower than those based on pure-tone thresholds measured objectively in the clinic. A typical finding is that by 65 years of age, over one-third of adults report hearing problems, and the prevalence of hearing loss has been reported to be as high as 80% for the frail elderly living in residential care facilities. The most obvious explanation for this apparent discrepancy in the estimates based on subjective and objective measures is that the effect of hearing loss on functioning in everyday life depends greatly on the individual's auditory ecology and not only on the degree of hearing impairment.

Research Issues

Most early research on presbycusis was confounded by the high degree of correlation between high-frequency hearing loss and age. More recent studies of the effects of age on auditory perception have

attempted to minimize this confound by matching younger and older adults as closely as possible for audiometric thresholds, by simulating the effects of hearing loss in younger listeners with normal hearing, or by amplifying sound so that it is equally audible to younger and older listeners. By controlling better for the effects of audibility it has been possible to isolate changes in auditory perception that are specific to aging.

Studies of nonhuman mammal species have enabled researchers to differentiate between the effects on the peripheral auditory system of genetic factors, environmental factors, and aging *per se*. A high degree of variability in hearing thresholds in quiet-reared gerbils and patterns of hearing loss in different strains of mice suggest that there is a strong genetic component in age-related hearing loss. A consistent finding is that heritability coefficients for age-related hearing loss in humans are strong (0.22–0.55), especially for mother/daughter and sister/sister relationships, with magnitudes similar to those reported for hypertension and hyperlipidemia. Whereas in studies of humans it is virtually impossible to eliminate the role of environmental factors that are known to cause hearing loss, it has been possible to rear animals in quiet and to control their diet and acoustical environment so that the effects of aging can be isolated.

Animal models have greatly advanced our understanding of age-related changes in peripheral auditory anatomy and physiology and how these changes relate to changes in pure-tone thresholds. Nevertheless, the auditory central nervous system, from the brain stem to the cortex, also undergoes significant age-related changes. It is more difficult to compare the nature and consequences of these changes across species, especially with regard to the perception of complex signals such as speech or music. Studies of the human central auditory system using electrophysiological techniques, such as evoked potentials and brain imaging, are beginning to address questions concerning age-related changes in the more cortical aspects of auditory perception.

Anatomy and Physiology

The main structures of the peripheral auditory system are the outer, middle, and inner ear. The effects of aging on auditory perception result mainly from changes in the inner ear and the auditory nervous system.

Outer Ear and Middle Ear

The outer ear, including the pinna, concha, and the ear canal to the ear drum, does not undergo any age-related change that significantly affects auditory perception in everyday life. Two outer ear conditions that may affect hearing testing are cerumen buildup and collapsing ear canals. Buildup of cerumen, or ear wax, may impede the perception of sound and possibly cause tinnitus, or ringing in the ears; however, cerumen can be safely removed by a qualified health professional. The cartilage of the outer ear loses rigidity, with the consequence that the ear canal may collapse if pressure is applied to the pinna. Collapsing canals occur in about a third of older adults. The occlusion produced when the ear canal collapses impedes the transmission of high-frequency sounds such that thresholds measured using circumaural earphones may over-estimate the degree of hearing loss. This testing artifact can be avoided by using insert earphones. A tester can easily identify a collapsing ear canal by pressing on the pinna with a finger and observing a change in the shape of the opening of the ear canal.

The middle ear cavity, from the ear drum to the inner ear, is normally air filled and contains a chain of three ossicles. The middle ear structures transduce the airborne acoustical sound vibrations arriving at the ear drum into mechanical vibrations that are relayed to the fluid-filled inner ear. Some older adults have hearing loss resulting from middle ear damage that developed earlier in their lifetime; however, the middle ear does not undergo any changes specifically related to age that significantly affect auditory perception in everyday life or during hearing testing.

Inner Ear

The high-frequency threshold elevations commonly observed in older adults may be partially the result of environmental factors such as exposure to noise or ototoxic drugs, rather than age *per se*. Losses due to such environmental factors are characterized by damage to and loss of the outer hair cells of the cochlea, or inner ear. In the normal cochlea, the outer hair cells have a gain-control function that enables low-intensity sounds to be heard. The outer hair cells are motile and under efferent control. Their status is evaluated clinically by testing otoacoustic emissions. Importantly, there is less change in otoacoustic emissions with age than would be expected if outer hair cell damage were the only pathology contributing to the high-frequency hearing loss typical of presbycusis.

In addition to outer hair cell damage resulting from environmental factors rather than to age *per se*, there are two subtypes of presbycusis in which cochlear pathology is attributable specifically to aging. One subtype of presbycusis, sensory loss, characterized by atrophy or degeneration of the sensory receptor

cells, the inner hair cells, and supporting cells, is now considered to be of only minor importance in the aged ear. A second subtype of presbycusis, metabolic loss, characterized by atrophy and degeneration of the lateral wall of the cochlea, especially the stria vascularis and spiral ligament, is now considered to be the predominant and most distinctive lesion of the aging ear.

The stria vascularis has a dense capillary network and an exceptionally high metabolic rate. There are significant losses of strial capillaries in aging animals. Age-related changes in the stria vascularis affect the basic mechanoelectric transduction process of the cochlea, the generation of electrochemical gradients, and the regulation of cochlear ionic homeostasis. Specifically, there is an age-related decrease in the endocochlear potential and in the gain of the cochlear amplifier. Importantly, metabolic presbycusis accounts for high-frequency threshold elevations, but is not observed in cases of noise-induced or ototoxic hearing loss. Furthermore, animal studies show that there can be at least temporary reversals of age-related threshold elevations on the order of 20–40 dB when the endocochlear potential is increased by the introduction of a direct current voltage into the scala media of the cochlea.

Auditory Nerve

A third subtype of presbycusis, neural loss, is characterized by loss of spiral ganglion neurons and pronounced degeneration of the auditory nerve. In quiet-reared gerbils and other species, spiral ganglion cell density decreases and the average volume of the surviving cells also decreases in older compared to younger ears. In aging gerbils, reduced slopes of the input–output functions are observed for the compound action potential of the auditory nerve and for the auditory brain stem response. Such reduced amplitudes of physiological responses in aging ears to moderate to high-level signals, in cases in which there are only small changes in auditory thresholds, are consistent with reduced synchronization of neural firing. Importantly, these changes have been observed in the presence of intact inner hair cells, consistent with age-related primary degeneration of the auditory nerve, as opposed to degeneration secondary to loss of sensory receptor cells, as is observed in cases of noise-induced or ototoxic hearing loss.

Perception

Auditory aging alters the perception of sound. The most commonly reported problem of older adults is difficulty understanding speech in noise. For the most part, the perceptual consequences of high-frequency hearing loss due to environmental factors are well known and will be the same for older adults as for younger adults. The perceptual consequences of the three subtypes of presbycusis are less well known, but seem to be distinctive. Importantly, older adults often find it more challenging than do their younger counterparts with similar audiograms when listening in complex acoustical environments, even though they have little or no problem listening in quiet situations. Furthermore, these problems can occur even if the audiometric thresholds of the older adults are within the normal range for speech or if sound is adequately amplified by hearing aids.

Psychoacoustic studies have documented changes in how older adults perceive the basic dimensions of sound. Studies have also documented age-related declines in speech perception and spoken language comprehension in more realistic, complex listening conditions. In humans, it has been difficult to distinguish the perceptual consequences of outer hair cell damage due to environmental factors from the perceptual consequences of the three subtypes of presbycusis, because most older adults have a combination of these pathologies. Insight into how the perceptual abilities of older adults resemble, or differ from, those of younger adults have been provided by studies in which the pure-tone thresholds of younger and older adults are matched as closely as possible, especially in the low-frequency range from 250 to 3000 Hz that is most important for speech perception.

Psychoacoustics

Absolute auditory thresholds Compared to a population of human adults with normal hearing, the perceptual decline that defines the degree of hearing loss is the increase in absolute thresholds. The absolute threshold of audibility is the lowest intensity at which a sound is detectable. Clinically, these thresholds are measured for pure-tone frequencies at octave intervals between 125 and 8000 Hz and they are plotted on an audiogram in decibels referenced to standards based on population hearing level (HL) norms, or dB HL. The typical hearing loss associated with presbycusis is characterized by normal thresholds, less than or equal to 25 dB HL, in the low-frequency range, and a sloping high-frequency hearing loss with thresholds increasing progressively with frequency. Figure 1 depicts the common progression of audiometric pure-tone threshold hearing loss with age for males and females.

An increase in high-frequency absolute thresholds, and thus a decline in sensitivity, can affect perception and function in everyday life; for example, some speech sounds, especially consonants, become inaudible. In addition, such hearing loss is usually

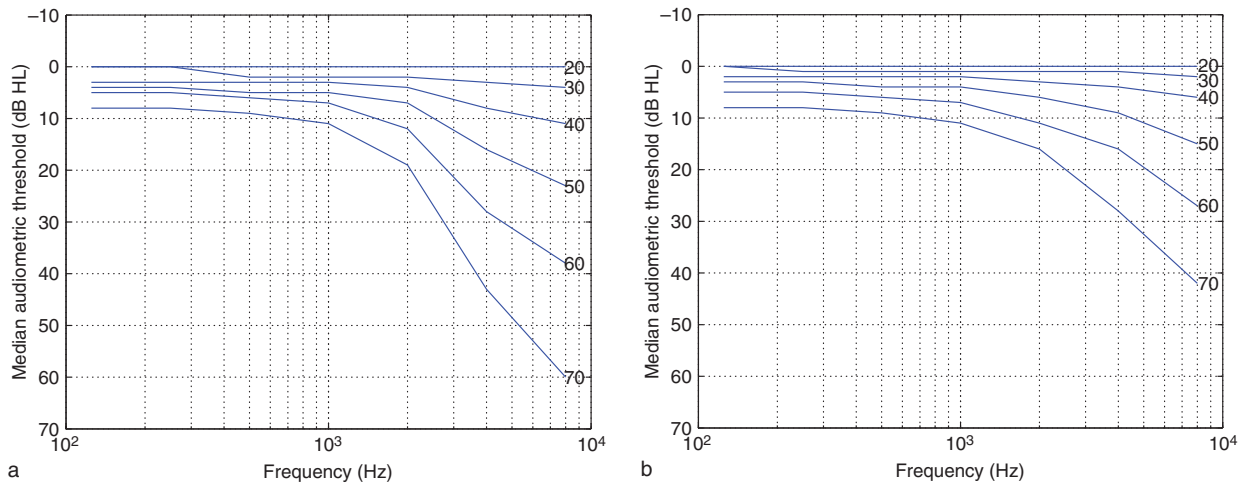

Figure 1 Median thresholds in decibel 'hearing level' (dB HL; i.e., relative to normal hearing in 18-year-olds) as a function of age for males (a) and females (b), based on population data from the International Standards Organization (ISO 7029).

associated with alterations in the perception of the physical dimensions of suprathreshold sounds. The main physical dimensions of suprathreshold sound are frequency, intensity, and duration. In general, variations in the frequency of sound are perceived as variations in pitch, while variations in the intensity of sound are perceived as loudness. Older adults often exhibit changes in suprathreshold perception that are not predictable from their audiometric thresholds.

Frequency selectivity and discrimination The cochlea is tonotopically arranged such that different regions respond maximally to different frequencies (i.e., there is place coding of frequency). The characteristic frequency, or the frequency that elicits the maximum response, is highest in the basal region of the cochlea (nearest to the middle ear) and it decreases more or less logarithmically with distance along the cochlea, with the most apical region having the lowest characteristic frequency. The tuning of the cochlear response to different frequencies has been examined using psychoacoustic experiments in humans and physiological measurements in mammals. The frequency tuning of the cochlea is often characterized in terms of auditory filter properties. The auditory filters of listeners with outer hair cell loss are broader than are those of listeners with normal hearing. The degree to which the auditory filters are broadened is correlated with degree of audiometric hearing loss.

A perceptual consequence of broadened auditory filters, or decreased frequency selectivity, is an increased susceptibility to masking of one signal by a simultaneous competing signal, especially masking signals lower in frequency than the target signal. Listeners with outer hair cell loss often have difficulty understanding speech in noise because of their increased susceptibility to masking. The deleterious effects of the upward spread of masking continue to undermine speech perception in noise, especially when the noise is the speech of other talkers, even when hearing aids amplify sound sufficiently to restore the audibility, because both the target and any competing signals are amplified.

The effect of outer hair cell loss on ability to discriminate sequentially presented tones differing in frequency is less clear. While frequency discrimination is usually adversely affected, there is significant variability among individuals, and the size of the just-noticeable frequency difference is generally not well correlated with the degree of threshold hearing loss.

In general, older adults with normal audiograms for low-frequencies do not discriminate sequentially presented tones of different frequencies as well as younger adults do. Importantly, these age-related differences in frequency discrimination do not appear to vary with presentation level and they are larger for low-frequency pure tones than for high-frequency tones. The place coding of frequency is dominant for high frequencies while temporal coding is dominant at low frequencies. Temporal coding of low frequencies is possible because the timing of the firing of primary auditory neurons is phase-locked to the signal which causes the movement of the cochlear structures. Thus, the synchronized timing of the firing of auditory neurons provides a cue that can be used to code low frequencies. At higher frequencies, the speed of firing needed for this type of coding exceeds the limits of the neural response. The pattern of age-related declines in frequency discrimination is

consistent with disruptions in temporal coding, rather than with the disruptions in place coding associated with outer hair cell damage that results in high-frequency threshold elevations.

Loudness and intensity discrimination Loudness recruitment is a phenomenon that is observed with almost all cases of outer hair cell loss. This phenomenon refers to an increased rate of growth in perceived loudness as the intensity level of a signal is increased. Consider an individual with a unilateral hearing loss – that is, one ear with normal thresholds and abnormally elevated thresholds in the other ear. While the range of audible sound levels will be different for both ears, the full range of perceived loudness will be the same (i.e., from barely audible to painfully loud). In the case of complete recruitment, once a tone becomes audible in the ear with elevated thresholds, the perceived loudness of that tone will grow faster there than in the normal ear as the level of the tone is increased, until the sound level is sufficiently high to produce the same perceived loudness in both ears. Overrecruitment or underrecruitment can also occur in some cases of cochlear hearing loss; respectively, these refer to the perceived loudness of a sound presented at a high level as being either greater or less in the impaired ear compared to the normal ear.

Typically, when tested at low intensity levels relative to their absolute thresholds, listeners with cochlear hearing loss who have recruitment discriminate smaller intensity differences compared to listeners with normal absolute thresholds. This is the basis of the behavioral tests sometimes used in the clinical evaluation of cochlear hearing loss characterized by recruitment. However, when tested at high intensity levels relative to their absolute thresholds, listeners with recruitment discriminate intensity differences that are similar, or sometimes larger, compared to young adults with normal audiograms.

Older adults with normal audiograms in the low-frequency range often exhibit larger intensity discrimination thresholds than do younger adults, with the largest age-related differences in intensity discrimination occurring for low-frequency tones. For high-frequency tones, loudness is mostly coded by the spread of activation along the length of the cochlea; that is, the overall rate of neural firing increases because activation occurs at more places in the cochlea. However, coding of loudness for low-frequency tones is influenced by temporal cues based on the number of neurons firing synchronously. The pattern of larger age-related threshold differences for low-frequency tones than for high-frequency tones is not consistent with the disruptions in place coding associated with outer hair cell damage resulting in high-frequency threshold elevations, but it is consistent with an age-related decline in neural synchrony.

Temporal processing The pattern of age-related differences in frequency and intensity discrimination for low frequencies is consistent with declines in auditory temporal processing. Age-related declines in temporal processing undermine the coding of the durational and transitional time-varying properties of the input signal and the extraction of signals from noise.

The gap detection threshold, the smallest gap between two sound markers that a listener can detect, is the most common measure of temporal resolution. Many older adults do not detect gaps until the gaps are significantly longer than those that can be detected by younger adults. Notably, age-related differences in gap detection thresholds are not associated with absolute threshold elevations. The effect of age on gap detection threshold is more pronounced when the markers surrounding the gap are shorter than 10 ms, when the gap is nearer to the onset or offset of the signal, and when the markers differ in frequency composition.

Older adults are also significantly worse than younger adults in discriminating the duration of two tones. The effect of age on duration discrimination is more pronounced when the target tone is embedded in a sequence of tones. There are also significant age-related declines in the discrimination and recall of the temporal order of short sequences of sounds, especially when the sounds are presented at fast rates.

Synchrony or periodicity coding is another aspect of auditory temporal processing that declines with age. Loss of neural synchrony is implicated by the pattern of age-related differences in frequency and intensity discrimination, as well as by the pattern of age-related differences in masking-level differences. Other evidence pointing to loss of periodicity coding in older adults is that they have more difficulty than do younger adults in detecting mistuned harmonics in complex sounds and in segregating concurrently presented sounds.

Speech Processing

Like listeners of any age who have outer hair cell loss, audiometric thresholds explain nearly all of the problems encountered by older adults when they listen to speech in relatively quiet environments. However, specifically age-related deficits seem to be most apparent in difficult or complex auditory perception tasks, such as when speech is presented rapidly or in noisy or reverberant conditions, particularly in a background of interrupted or modulated noise, or when there is competing speech, even from a single talker.

Importantly, age-related deficits are much larger for temporally complex conditions compared to spectrally complex conditions. There is growing evidence that age-related declines in auditory temporal processing account for the disproportionate difficulties in understanding speech in noise that are experienced by older adults, including those for whom the audibility of speech is not a problem because they have normal audiometric thresholds throughout the speech frequency range up to 3000 Hz.

Declines in auditory temporal processing affect the ability of older adults to use different types of speech cues. At the sentence level, prosodic cues, such as syllable stress and rhythmic patterning, impart syntactic and affective information. At the word level, gaps and durations serve phonemic contrasts that enable word recognition (e.g., the stop consonant gap of [p] distinguishes the words split and slit). At the subphonemic level, periodicity cues derived from the voice fundamental frequency and harmonic structure contribute to voice quality and talker identification. Older adults exhibit deficits at all levels of speech processing that can be related to declines in auditory temporal processing.

Interactions between Auditory and Cognitive Processing

There is a strong correlation between sensory and cognitive aging. The difficulties of older adults in understanding spoken language in complex listening conditions may be exacerbated by the effect of age on cognitive processing, including declines in working memory and attention. As the listening condition becomes more adverse, it becomes more cognitively demanding for listeners of any age to process heard information. However, older adults demonstrate preserved knowledge and better ability, compared younger adults, to use contextual cues to compensate for difficulty hearing in challenging listening conditions.

Rehabilitation

In quiet, older adults with outer hair cell loss exhibit perceptual deficits resembling those of younger adults with comparable loss. Beyond the difficulties in speech perception explained by loss of audibility, many older adults also have problems attributable to subtypes of presbycusis involving declines in temporal processing. Currently, hearing aids are unable to correct for these declines; however, other assistive listening technology which enables the segregation of sound sources can reduce the challenge of listening in complex acoustical environments. Other forms of rehabilitation focus on training the older person and their communication partners to use strategies to avoid and cope with everyday listening challenges.

See also: Auditory Scene Analysis; Auditory System: Central Pathway Plasticity; Auditory Evoked Potentials; Auditory System: Central Pathways; Cochlear Mechanics; Cochlear Development; Cochlear Prosthesis; Deafness; Hair Cell Regeneration; Otoacoustic Emissions; Speech Perception: Adult; Speech Perception: Neural Encoding; Tinnitus.

Further Reading

Cilento B, Norton S, and Gates G (2003) The effects of aging and hearing loss on distortion-product otoacoustic emissions. *Otolaryngology, Head and Neck Surgery* 129: 382–389.

Davis AC (1995) *Hearing Impairment in Adults*. London: Whurr Publishers Ltd.

Frisina DR, Frisina RD, Snell KB, et al. (2001) Auditory temporal processing during aging. In: Hof PR and Mobbs CV (eds.) *Functional Neurobiology of Aging*, pp. 565–579. New York: Academic Press.

Gallacher J (2005) Hearing, cognitive impairment and aging: A critical review. *Reviews in Clinical Gerontology* 14: 199–209.

International Organization for Standardization (2000). *ISO 7029–2000: Acoustics – Statistical Distribution of Hearing Thresholds as a Function of Age*, 2nd edn. Geneva: International Organization for Standardization.

Kiessling J, Pichora-Fuller MK, Gatehouse S, et al. (2003) Candidature for and delivery of audiological services: Special needs of older people. *International Journal of Audiology* 42(supplement 2): 2S92–2S101.

Moore BCJ (1998) *Cochlear Hearing Loss*. London: Whurr Publishers Ltd.

Pichora-Fuller MK and Carson AJ (2000) Hearing health and the listening experiences of older communicators. In: Hummert ML and Nussbaum J (eds.) *Communication, Aging, and Health: Linking Research and Practice for Successful Aging*, pp. 43–74. Mahwah, NJ: Lawrence Erlbaum Associates.

Pichora-Fuller MK and Singh G (2006) Effects of age on auditory and cognitive processing: Implications for hearing aid fitting and audiological rehabilitation. *Trends in Amplification* 10: 29–59.

Schneider BA and Pichora-Fuller MK (2000) Implications of perceptual deterioration for cognitive aging research. In: Craik FIM and Salthouse TA (eds.) *The Handbook of Aging and Cognition*, 2nd edn., pp. 155–219. Mahwah, NJ: Lawrence Erlbaum Associates.

Vaughan N and Fausti S (eds.) (2006) The aging auditory system: Considerations for rehabilitation. Proceedings from the National Center for Rehabilitative Auditory Research (NCRAR), 2005. *Seminars in Hearing* 27(4): 213–352.

Willott JF (1991) *Aging and the Auditory System: Anatomy, Physiology, and Psychophysics*. San Diego: Singular.

Sensory Aging: Vision

J C Blanks and C K Dorey, Florida Atlantic University, Boca Raton, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Aging of the eye affects all structures, and changes in visual function accompany the aging process. The elderly often exhibit a decrease in visual acuity, ocular accommodation, and dark adaptation. Diseases can interfere with the aging process and cause blindness. Blindness among white and black persons in the US was caused by age-related macular degeneration (AMD; 46% and 4% of (cases); cataract (9% and 37%); glaucoma (6% and 26%); and diabetic retinopathy (5% and 7%). Vision loss is currently a major public health problem in the United States.

Approximately 3.3 million Americans age 40 years and older are blind or have low vision. With aging of the US population, the number of Americans with eye disease will increase such that by 2020 the number of people who are blind or have low vision is projected to increase substantially, to 5.5 million. Advanced age is the strongest risk factor for AMD, which is currently estimated to affect 20–25 million people in the industrialized world, causing severe vision loss in 8 million. With the aging of society, the number of people with AMD will triple in the next 40 years. This article discusses the normal pattern of aging in the choroid, retinal pigment epithelium (RPE), retina, vitreous humor, lens, and cornea (Figure 1). Diseases of aging associated with these structures are also covered.

Outer Retina

The retina is divided into the light-detecting outer retina and the visual-processing inner retina (Figure 2). The outermost layer of the retina, composed entirely of rod and cone photoreceptors, is physiologically and functionally dependent on the neuroectoderm-derived RPE that covers the photoreceptors, and the choriocapillaris that delivers vitamin A, nutrients, and all oxygen utilized by the photoreceptors. The choriocapillaris is separated from the RPE and photoreceptors by Bruch's membrane, a thick avascular area of extracellular matrix (Figure 3). Gradual age-related changes in the RPE, Bruch's membrane, and choriocapillaris become quite marked at advanced age, and influence photoreceptor life and function.

It has been suggested that it would be productive to consider age-related degenerative diseases as a cluster of problems associated with poor aging, rather than as problems arising from disease-specific pathologies. AMD may well be an archetypical condition for exploration of this view. As the following discussion illustrates, reduced photoreceptor function in older eyes can derive from age-related changes in any or all of these functionally intertwined layers. Moreover, age-related changes observed in the RPE are known to induce pathology.

Photoreceptors

Aging Changes in Photoreceptors

Decline in visual acuity is especially prominent in low light. Human rods are lost more rapidly with age, declining ~3% per decade, than are cones, which decline in number by 1.8% per decade. With increasing age, the diameter of rod outer segments and the optical density of rhodopsin increase, whereas the scotopic sensitivity of the rods, and the time needed to recover visual sensitivity following a bleach, both decline. Accumulated oxidative damage contributes to photoreceptor loss.

Photoreceptor outer segments contain photosensitive visual pigments (e.g., rhodopsin), which mediate light damage to the photoreceptors, and a high concentration of readily oxidized polyunsaturated fatty acids. This structure and the component environment (high concentration of oxygen and constant light exposure) result in a high rate of photooxidation in outer segment membranes. Photoreceptor function is maintained by rapid turnover of outer segments: new materials are added proximally, and the RPE ingests and digests the distal oxidized tips daily (Figure 4). Photoreceptor degeneration occurs if the RPE is unable to ingest or digest the photoreceptor tips – age-related changes in the RPE affect both processes.

Photoreceptors and the RPE are protected from damaging blue light by accumulation of the carotenoids lutein and zeaxanthin in photoreceptor axons and outer segments, and in axons in the inner plexiform layer of the central retina. The high carotenoid concentration in the retina creates a visible yellow region known as macular pigment.

AMD-Related Changes in Photoreceptors

Soft drusen, a clinical hallmark of AMD, cause photoreceptor distortions (Figure 5). Large areas of soft drusen are associated with disturbances in photoreceptor physiology, including decreased sensitivity of rods and short-wavelength cones, delayed dark adaptation of rods, and, with longer duration of AMD,

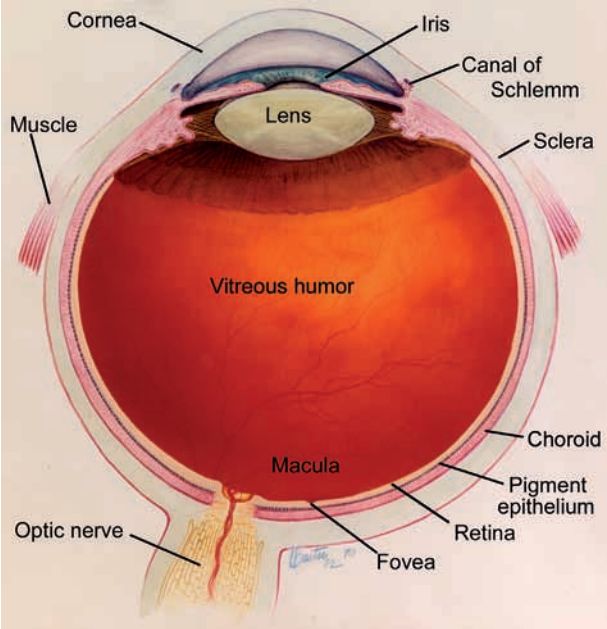

Figure 1 Diagram of a human eye, illustrating the structures discussed in the text. Data from the National Eye Institute, National Institutes of Health (NEI/NIH) website, http://www.nei.nih.gov.

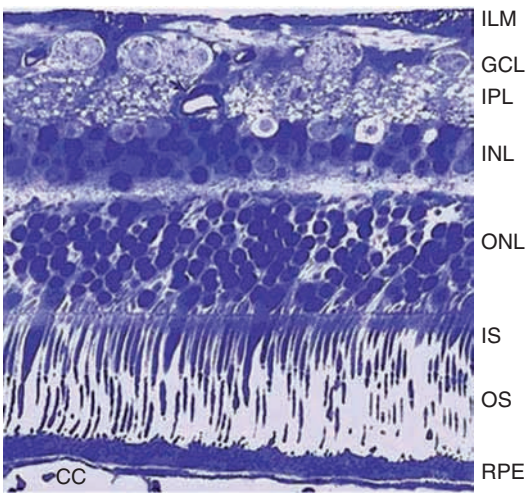

Figure 2 Histological section of a retina. The inner retina is composed of neural cells and their processes, including the inner limiting membrane (ILM), ganglion cell layer (GCL), inner plexiform layer (IPL), and inner nuclear layer (INL). The outer retina contains only photoreceptor nuclei in the outer nuclear layer (ONL) and photoreceptor inner and outer segments (IS, OS) in the subretinal space. The photoreceptors are covered and maintained by the retinal pigment epithelium (RPE) and the vascular supply (choriocapillaris; CC). Modified from Kolb H, Fernandez E, and Nelson R, *WebVision: The Organization of the Retina and Visual System*, available online at http://webvision.med.utah.edu/.

deficiencies in color vision. Rod photoreceptors are lost preferentially and at an accelerated rate, leading to measurable loss of extrafoveal acuity, while retention of cones leaves central visual acuity unaffected.

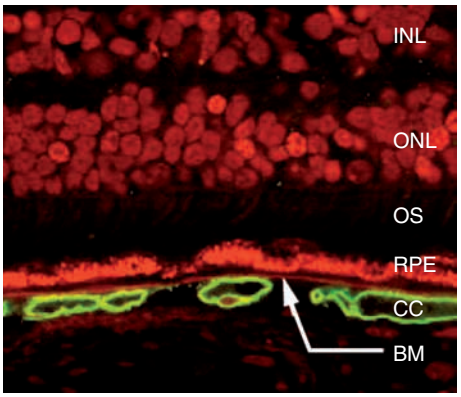

Figure 3 Histological section of retina stained to reveal nuclei (red) in the inner nuclear layer (INL) and outer nuclear layer (ONL), and choriocapillaris (CC, yellow-green). Autofluorescence from retinal pigment epithelium (RPE) lipofuscin and Bruch's membrane (BM) are also shown in red. Choriocapillaris (CC) profiles contact BM; outer segments (OS) are not stained. Adapted from Yang Z, Alvarez BV, Chakarova C, et al. (2005) Mutant carbonic anhydrase 4 impairs pH regulation and causes retinal photoreceptor degeneration. *Human Molecular Genetics* 14: 255–265, with permission from Investigative Ophthalmology Visual Science.

Eventually in 'dry AMD,' visual field defects occur in macular areas of geographic atrophy, where both photoreceptors and RPE are missing. 'Wet AMD' – characterized by growth of choroidal neovascularization into Bruch's membrane and/or the photoreceptor layer – also causes marked loss of visual acuity. Both types can ultimately cause blindness.

Retinal Pigment Epithelium

The RPE is a single layer of nondividing cells that extend processes across the subretinal space toward the photoreceptors. The basal surface of the RPE rests on the innermost layer of Bruch's membrane (see Figure 2). Age-related changes dramatically alter both structure and function of the RPE, impact photoreceptor function, and may contribute to the pathogenesis of AMD.

RPE Functions Essential for Survival and Function of Photoreceptors

Visual cycle Photoreceptor function is dependent on constant delivery of vitamin A in the proper form. Retinol in the choriocapillaris is transported through Bruch's membrane to the basal RPE, transported into the RPE, isomerized to the required retinol isomer, and transported to the photoreceptors; after light absorption, it cycles back to the RPE.

Absorption of short-wavelength light The apical RPE contains numerous eumelanin granules. Melanin absorption of ultraviolet (UV) and blue light serves

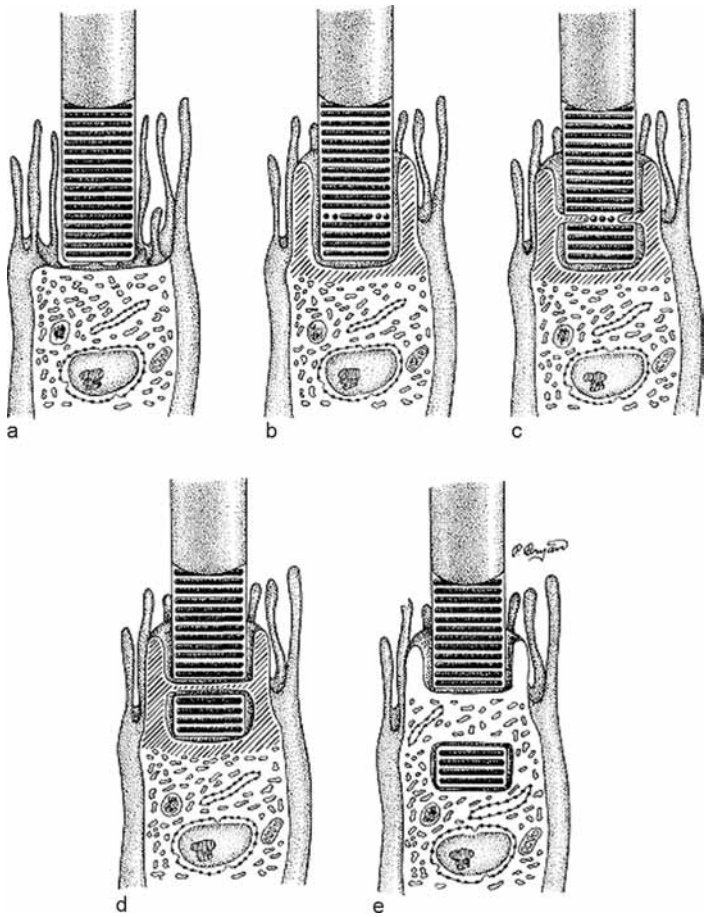

Figure 4 Illustration of retinal pigment epithelium phagocytosis of damaged rod outer segment tips. In panel (e), lysosomal enzymes will enter the phagosome to digest the tip, leaving fluorescent residue. Adapted from Matsumoto B, Defoe DM, and Besharse JC (1987) Membrane turnover in rod photoreceptors: Ensheathment and phagocytosis of outer segment distal tips by pseudopodia of the retinal pigment epithelium. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 230: 339–354, with permission from The Royal Society.

two purposes: reduced scatter of light back onto the photoreceptors improves visual acuity whereas reduced exposure to blue light reduces generation of singlet oxygen by A2E, an endogenous photosensitizer (see later).

Blood–retinal barrier The RPE serves as the interface between the photoreceptors and the rest of the body. Its tight junctions form the outer blood–retinal barrier, regulating movement of molecules and cells into the retina. Factors expressed by RPE cells modulate the activities of inflammatory cells that enter Bruch's membrane, and control their access to the retina. Risk for inflammatory damage to the retina is reduced by RPE inhibition of T cell proliferation and activation, and inhibition of complement cascades. However, when exposed to inflammatory cytokines, the RPE produces factors chemotactic for neutrophils and monocytes and expresses cellular adhesion molecules that facilitate migration of leukocytes across the RPE barrier.

Age-Related Changes in the RPE

Age-related cell loss The RPE population declines about 3% per decade, losing 30% of RPE cells over a lifetime. As in aging neurons, RPE cells experience age-related loss in number and area of mitochondria, and loss of cristae. These losses occur earlier and are exaggerated in AMD. Taken together, these findings indicate that the ability of the RPE to execute its phagocytic, transport, and visual cycle functions will decline with age.

Increased oxidative damage A major change with advancing age of the RPE is decreasing ability to face oxidant challenge. Expression of antioxidant enzymes and levels of blue-light-absorbing melanin both decline. At the same time, the RPE accumulates a blue-light-absorbing photosensitizer. Increasing oxidative stress may contribute to the age-related increase in DNA mutations, and is strongly linked to the progression of AMD; a multicenter prospective

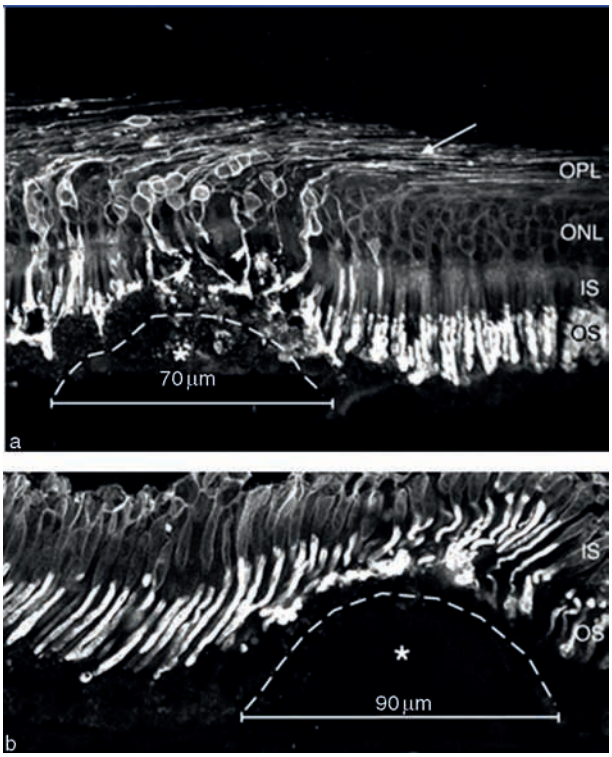

Figure 5 Immunohistochemistry of rod (a) and cone (b) opsins. Drusen deposits (*; inside dashed lines) cause shortening of outer segments (OS) and relocation of opsin to (a, b) inner segments (IS), and (a) outer nuclear layers (ONL) and rod axons (arrow) in the outer plexiform layer (OPL). Reproduced from Johnson PT, Lewis GP, Talaga KC, et al. (2003) Drusen-associated degeneration in the retina. *Investigative Ophthalmology & Visual Science* 44: 4481–4488, with permission from Investigative Ophthalmology & Visual Science.

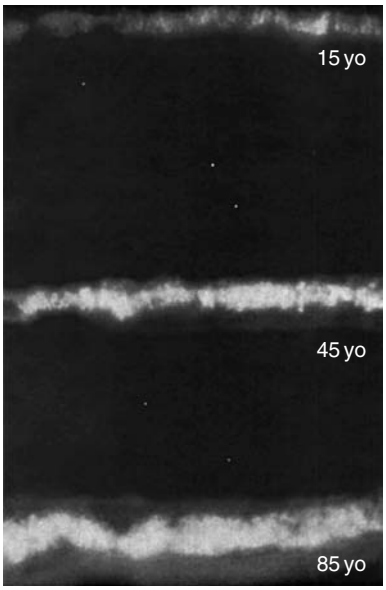

Figure 6 Increase in lipofuscin autofluorescence in eyes from 15-year-old (15 yo), 45-year-old, and 85-year-old donors. Reproduced from Okubo A, Rosa RH Jr., Bunce CV, et al. (1999) The relationships of age changes in retinal pigment epithelium and Bruch's membrane. *Investigative Ophthalmology & Visual Science* 40: 443–449, with permission form Investigative Ophthalmology & Visual Science.

clinical trial has demonstrated that high doses of antioxidants (vitamin C, vitamin E, β-carotene, and zinc) slow progression of AMD.

Ingestion of oxidized outer segments probably impairs RPE digestion, since lysosomal enzyme activity declines in RPE fed oxidized lipid, and in other cell types fed oxidized protein. Moreover, exposure to oxidized membranes may help switch the RPE to a proinflammatory state: ingestion of oxidized outer segments triples RPE production of monocyte chemotactic protein-1, which would increase macrophages along Bruch's membrane.

Lipofuscin accumulation With advancing age, RPE lysosomes become engorged with lipofuscin (Figure 6), autofluorescent residue from the daily ingestion of oxidatively damaged tips of photoreceptor outer segments. The prominent fluorophore (A2E) is of particular interest, for it can be cytotoxic. The number of lipofuscin granules increases with age: by the ninth decade they occupy almost 20% of the free cytoplasmic space in the macular RPE cell. Lipofuscin levels are low in the fovea, increase across the macula, and decrease in the periphery. Regions with higher fundus autofluorescence (of RPE lipofuscin) exhibit delayed dark adaptation and decreased scotopic sensitivity, and are regions of preferential loss of rod photoreceptors in AMD.

A2E: Fluorophore by-product of the visual cycle A2E is a prominent lipofuscin fluorophore, composed of two molecules of vitamin A (retinol) and one molecule of ethanolamine. Upon absorption of a photon, visual pigments release highly reactive retinaldehyde. Under normal conditions, the photoreceptor-specific ATP binding cassette transporter (ABCA4) rapidly transports retinaldehyde across the disk membrane to the cytoplasm for reduction. When the retinaldehyde levels are sufficiently high to permit two retinaldehydes to sequentially interact with one phosphatidylethanolamine, an A2E precursor is formed. RPE digestion releases A2E. A2E formation is accelerated under conditions of high light irradiance, low antioxidant capacity in the photoreceptor, or when the ABAC4 activity is impaired by oxidative damage or inherited polymorphisms. A2E accumulates rapidly in Stargardt's retinal dystrophy (juvenile macular degeneration), caused by polymorphisms in ABCA4. The role of A2E is strongly implicated by evidence that mice lacking this enzyme accumulate high levels of A2E before photoreceptor atrophy.

Toxicity of A2E A2E may contribute to pathogenesis of AMD. RPE cells loaded with A2E *in vitro* have impaired lysosomal acidification and, consequently, inhibited lysosomal enzymes and impaired digestion of outer segments. Noxious A2E is also a photosensitizer in blue light, generating reactive singlet oxygen and lysing lysosomes. Released into the cytoplasm, A2E targets mitochondria, releases apoptosis-inducing proteins, and causes RPE apoptosis. The blue light hazard for RPE is increased by higher A2E levels and is decreased by absorption of blue light by lens brunescence, macular pigment optical density (determined by dietary lutein and zeaxanthin), and RPE melanin (which declines with age). Higher dietary, plasma, and retinal levels of lutein and zeaxanthin are associated with decreased risk for AMD; higher retinal zeaxanthin prevents light-induced photoreceptor death in quail.

The RPE in AMD

The functionally impaired RPE is unable to support the visual cycle – consequently, patients exhibit impaired dark adaptation. RPE changes contribute to the pathology seen in Bruch's membrane and the choriocapillaris. Foci of high lipofuscin (fundus hyperfluorescence) surrounding geographic atrophy (Figure 7) identify areas of subsequent photoreceptor and RPE apoptosis. The RPE in AMD has reduced ability to suppress inflammation, as described in the following sections.

Bruch's Membrane and Choriocapillaris

The choriocapillaris is a bed of interconnected, highly fenestrated capillaries lying on the outer surface of Bruch's membrane (Figure 8). All peptide factors and waste products from the RPE and all nutrients from the choriocapillaris must pass through Bruch's membrane. Unfortunately, the aging Bruch's membrane accumulates debris that interferes with transport and induces inflammatory reactions. The RPE and choriocapillaris are functionally intertwined – death of either layer will ultimately cause attenuation or death of the other. The RPE produces angiogenic vascular endothelial growth factor (VEGF) and anti-angiogenic factors (e.g., pigment epithelial growth factor and endostatin) that maintain and regulate growth and status of the choriocapillaris.

Age-Related Changes in the Choriocapillaris

The choriocapillaris volume declines with advancing age. By age 90 years, almost half of the capillary

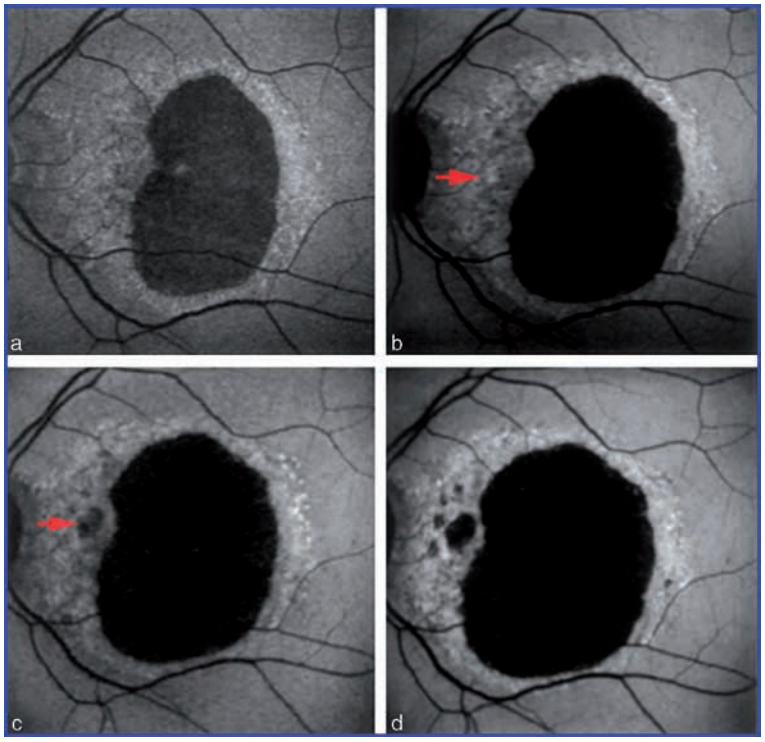

Figure 7 Clinical images of fundus fluorescence from retinal pigment epithelium lipofuscin taken at yearly intervals (a–d). The dark areas are geographic atrophy, where both retinal pigment epithelium and photoreceptors have died. Atrophy expands in areas of higher lipofuscin; the bright area in b (arrow) became an atrophic area 1 year later (arrow in c). Reproduced from Holz FG, Bellman C, Staudt S, et al. (2001) Fundus autofluorescence and development of geographic atrophy in age-related macular degeneration. *Investigative Ophthalmology & Visual Science* 42: 1051–1056, with permission form Investigative Ophthalmology & Visual Science.

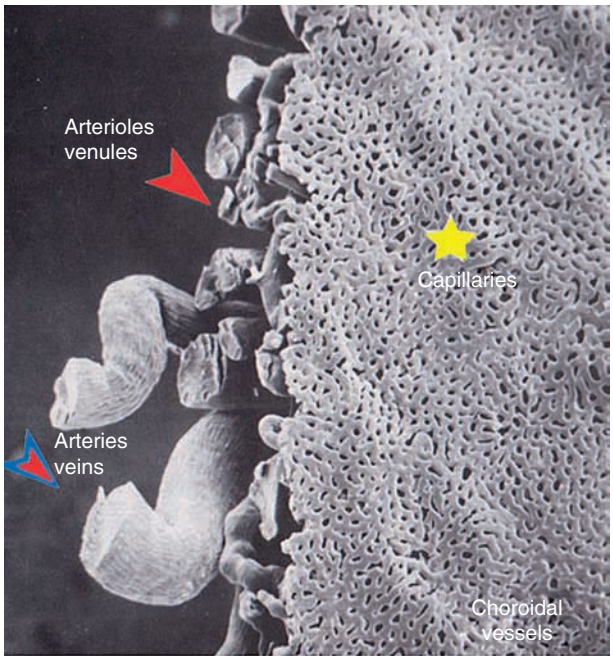

Figure 8 The meshwork of the healthy choriocapillaris (star), and larger choroidal arteries and veins that supply it, are visible in this corrosion cast of a human choroid. Reproduced from Zhang HR (1994) Scanning electron-microscopic study of corrosion casts on retinal and choroidal angioarchitecture in man and animals. *Progress in Retinal & Eye Research* 13: 243–270, with permission from Investigative Ophthalmology & Visual Science.

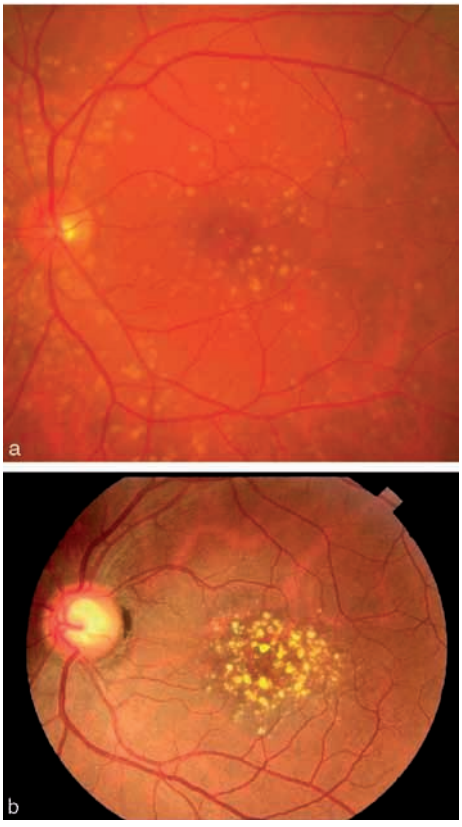

Figure 9 Drusen in clinical photographs of the human macula. Scattered hard drusen deposits with distinct boundaries are seen in most older eyes (a), whereas the presence of soft indistinct drusen (b) is associated with decline in photoreceptor function, and high risk for age-related macular degeneration. (a) Reproduced from Galan A and Tavolato M (2004) Cataract surgery as risk factor for AMD progression. *Oculista Italiano* 138, with permission from La Oculista Italiano.; (b) Reproduced from Charles Retina Institute, http://www.charles-retina.com, with permission from Dr. Charles.

profiles have been lost and their diameter has decreased by 30%. Consequent RPE hypoxia induces production of factors that can cause choroidal neovascularization in AMD. RPE production of antiangiogenic factors declines with aging and oxidative stress. At advanced age, the balance of angiogenic and antiangiogenic factors is tipped toward angiogenesis, increasing risk for choroidal neovascularization and visual loss due to retinal edema or scarring.

Aging changes in Bruch's membrane Age-related changes can be characterized as deposits which thicken Bruch's membrane and/or change in its biophysical properties. Deposits of amorphous material and long spacing collagen accrue above and below the RPE basal lamina, whereas vesicular and fibrillar deposits accumulate within Bruch's membrane. Focal exaggeration of these deposits under the basal lamina form a small lump that elevates RPE and is visible clinically as hard drusen in virtually all older eyes (Figure 9(a)). The thickening of Bruch's membrane is linearly correlated with lipofuscin accumulation. Lipids, cholesterol, phospholipids, triglycerides, and apolipoproteins accumulate throughout life. The hydraulic conductivity of Bruch's membrane declines rapidly, losing 70% by age 30 years; permeability to macromolecules declines 10% per decade. Reduced permeability to RPE factors could contribute to loss of the choriocapillaris.

Some eyes also develop larger soft drusen with indistinct borders; this feature is used to identify eyes at risk for AMD. Bruch's membrane in these eyes is thickened by ubiquitous layers of membranous debris and proteins; localized areas of thicker deposits form soft drusen where RPE displacement narrows the subretinal space, and distorts the regular array of the photoreceptors. Photoreceptors over drusen exhibit altered distribution of opsin, shortened outer segments, and altered synaptic terminals. Patients with maculopathy exhibit reduced color contrast sensitivity and decreased light sensitivity. Large numbers of confluent soft drusen deposits increase risk of progression to geographic atrophy or choroidal neovascularization.

The known components of hard and soft drusen are not different. Deposits include oxidatively

damaged proteins, and serum and RPE proteins that regularly traverse Bruch's membrane. It is noteworthy that complement C5 and the membrane attack complex (MAC) (C5–C9) are found in all types of drusen.

Inflammation in Drusen and AMD

Advanced glycation endproducts (AGEs) are cross-linked complexes of oxidized sugars, lipids, and proteins; they are considered a universal phenomenon of aging. Continuous deposits of AGEs are found in Bruch's membrane in AMD (Figure 10). Cellular responses to AGEs are mediated by cell surface receptors for advanced glycation endproducts (RAGEs) (Figure 11). Photoactivation of A2E induces RAGE expression in cultured RPE. RAGEs are expressed in the RPE over drusen, but not in RPE associated with normal extracellular matrix. RAGEs have been strongly implicated in the pathogenesis of complications of diabetes, atherosclerosis, and arthritis. Activation of RAGEs is associated with release of reactive oxygen species and increased expression of RAGEs, nuclear factor-kappa B (NF-κB)-regulated inflammatory cytokines, chemotactic factors for inflammatory cells, adhesion molecules, and members of the complement cascade.

Complement Activation

Activation of the complement cascade on cell surfaces can result in damage by the MAC, a membrane-piercing aggregate of the final five complement components (C5–C9). Activated complement components, including MAC, are routinely found in both hard and soft drusen (Figure 12). Cell responses to MAC include shedding of damaged membrane and/or production of regulators (such as fibronectin or clusterin) that impede formation of MAC. Vitronectin and clusterin are present in drusen and are produced by RPE cells (Figure 12(c)).

Despite evidence that complement activation occurs in drusen, only a fraction of people with drusen develop AMD. When complement C3b is deposited on cell surfaces, factor H prevents cell damage by binding and inactivating C3b. Recently, single polymorphisms in factor H were found in ~50% of patients with AMD; adding significant variants in other complement factors (component C2 regulator factor B) predicts up to 75% of AMD outcomes. Thus

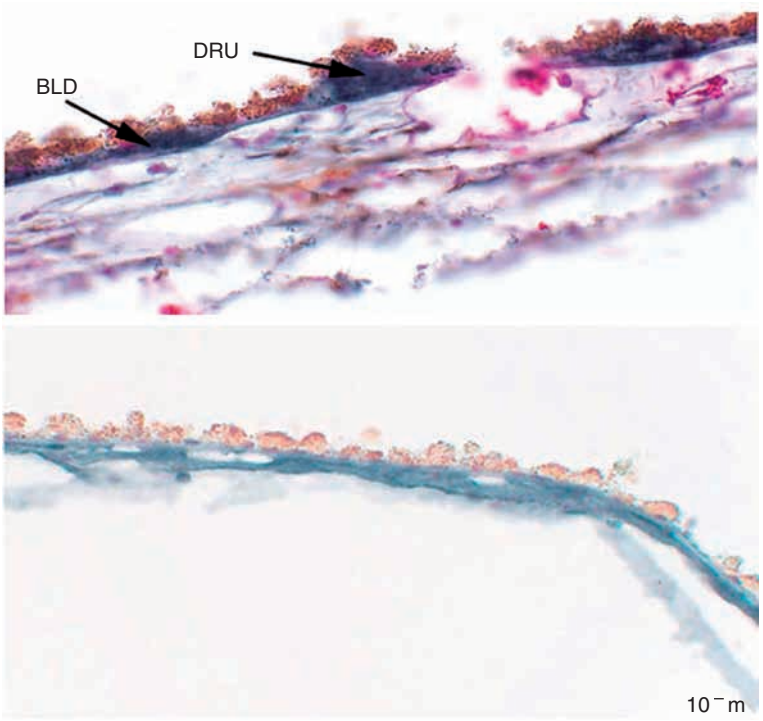

Figure 10 Immunohistochemical localization (blue) of advanced glycation endproducts in druse (DRU) and thickened Bruch's membrane under the retinal pigment epithelium (brown) of an 82-year-old donor (BLD, basal laminar deposit). Reprinted from Farboud B, Aotaki-Keen A, Miyata T, et al. (1999) Development of a polyclonal antibody with broad epitope specificity for advanced glycation endproducts and localization of these epitopes in Bruch's membrane of the aging eye. *Molecular Vision* 5: 11, with permission from Molecular Vision.

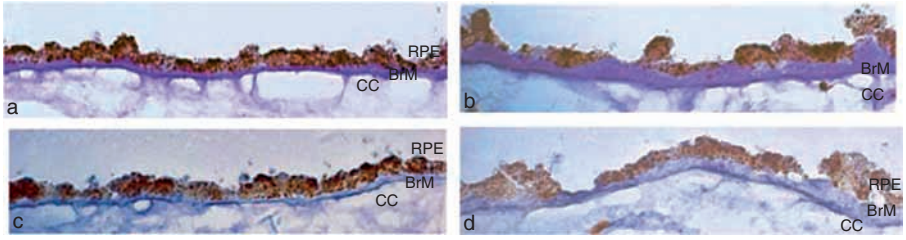

Figure 11 Blue stain reveals immunohistochemical localization of the receptors for advanced glycation endproducts in an 85-year-old donor eye. The receptors are present in (a, c) normal retinal pigment epithelium (RPE) and Bruch's membrane (BrM) and in (b, d) thick basal deposits in Bruch's membrane. CC, choriocapillaris. Reproduced from Yamada Y, Ishibashi K, Bhutto IA, et al. (2006) The expression of advanced glycation endproduct receptors in RPE cells associated with basal deposits in human maculas. *Experimental Eye Research* 82: 840–848, with permission from Elsevier.

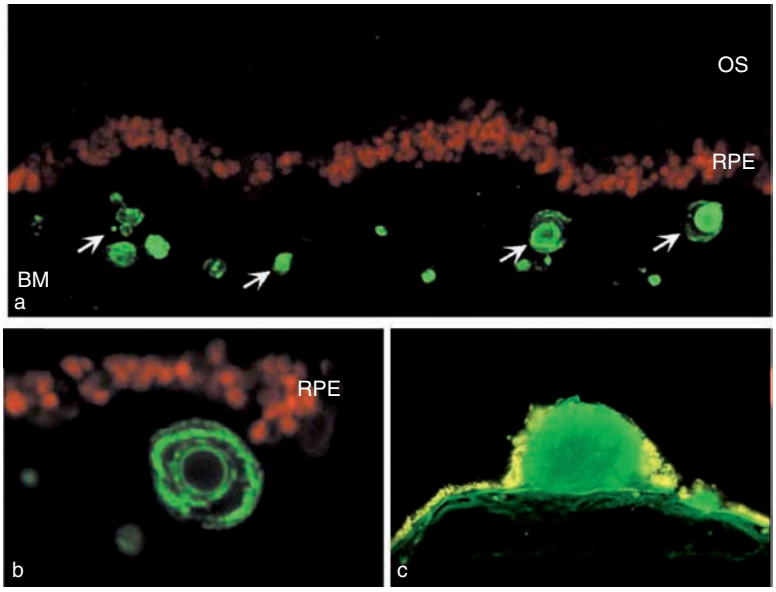

Figure 12 (a, b) Activated C3b complement (green) is present in drusen under retinal pigment epithelium (RPE) with lipofuscin autofluorescence (red) (BM, Bruch's membrane; OS, outer segment). (c) RPE cells (yellow autofluorescence) secrete inhibitors of complement activation such as vitronectin (green), which is localized throughout Bruch's membrane and drusen. (a, b) Reprinted from Johnson LV, Leitner WP, Staples MK, et al. (2001) Complement activation and inflammatory processes in Drusen formation and age related macular degeneration. *Experimental Eye Research* 73: 887–896, with permission from Elsevier Science & Technology; (c) Reprinted from Hageman GS and Mullins RF (1999) Molecular composition of drusen as related to substructural phenotype. *Molecular Vision* 5: 28, with permission from Molecular Vision.

inability to prevent complement-mediated cell damage may be a key factor determining whether the inflammation observed in aging eyes will lead to macular degeneration.

Normal Aging of the Inner Retina

Age-Related Changes of the Inner Retina

While there is a paucity of studies analyzing changes in interneurons (bipolar, horizontal, and amacrine cells) in the retina with age, a number of histological studies show an inherent loss of retinal ganglion cells (RGCs) with age. RGCs, located in the innermost layer of the retina (adjacent to the inner limiting membrane and the vitreous humor; Figure 13) have large somas with dendritic processes that project into the inner synaptic layer, where they synapse with amacrine cells and bipolar cells. A single axon projects from each RGC and runs centripetally along the vitreous surface in the nerve fiber layer (NFL). These axons (approximately 1 million in total) produce the first action potentials in the visual system and carry information already processed by the retina to targets in the brain (primarily the lateral geniculate nucleus of the thalamus) for further processing. Although a larger percentage of RGCs is lost with age in the peripheral retina than in the central retina, the number of RGCs actually lost is similar, because of the higher density of RGCs in the central region. Therefore, a relatively constant number of RGCs die with age in different retinal locations. The

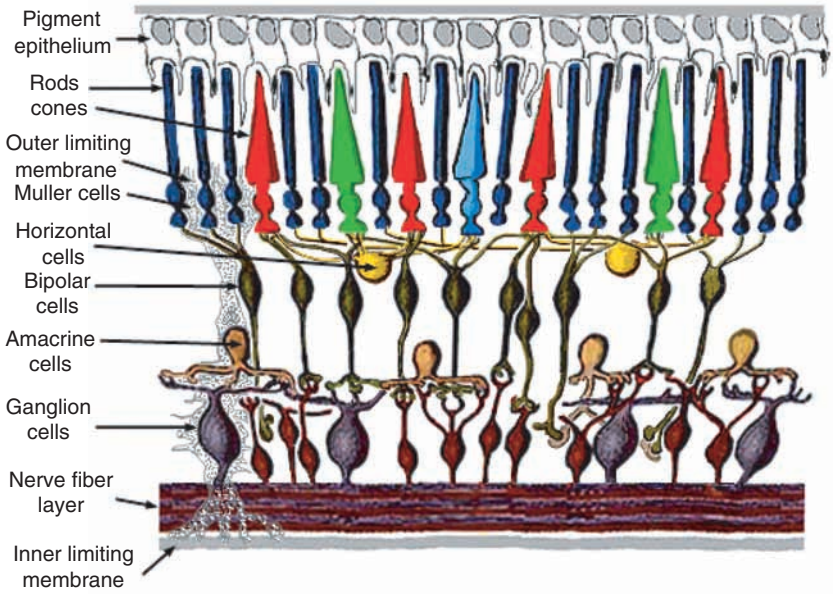

Figure 13 Schematic illustrating the synaptic connections between the photoreceptors and horizontal and bipolar cells of the inner retina, and between bipolar cells and ganglion cells and amacrine cells. Note that the retinal pigment epithelium is at the top. From Kolb H, Fernandez E, and Nelson R, *WebVision: The Organization of the Retina and Visual System*, available online at http://webvision.med.utah.edu.

naturally occurring RGC loss with age means that about 5000 RGC axons (about 0.5%) die per year per eye. Noninvasive imaging techniques that measure the topography of the optic nerve head and the thickness of the NFL show at least a 20% thinning of the NFL with age. This thinning appears to be more pronounced in the superior quadrant. The reduction in the number of visual cells (photoreceptor cells and/or RGCs) in the retina and in higher visual centers is thought to be associated with the reduced visual acuity, decreased contrast sensitivity, and higher thresholds for light detection in visual field measurements that occur with age.

Diseases of Aging Inner Retina

RGC loss may also play a role in visual dysfunction observed in neurodegenerative diseases. Parkinson's disease (PD) is characterized by a progressive loss of dopamine-containing neurons in the substantia nigra. There is also a progressive loss of dopaminergic cells in the retina and possibly within other areas of the visual system in PD. Selective spatial–temporal abnormalities in RGC function (arising from altered receptive field organization) may be due to the retinal dopamine deficiency.

Many patients with Alzheimer's disease (AD) also exhibit impairments of visual perception and eye movements. These defects can include anomalies in color vision, spatial contrast sensitivity, susceptibility to visual masks, electroretinographic and visual evoked potential abnormalities, and saccadic and eye tracking dysfunction. The precise basis for these deficits is unknown, but senile plaques, neurofibrillary tangles, and amyloid protein deposits are found in multiple regions of the brain that mediate vision, excluding the retina. Some of these visual impairments may also be due to retinal lesions, since RGC loss, reduction in the number of RGC axons in the optic nerve, and electroretinographic abnormalities have been reported in AD patients.

Glaucoma

Glaucoma is an eye disease in which the normal fluid pressure inside the eye slowly rises, leading to vision loss or blindness. The disease is characterized by changes in the morphology of the optic disk, the pattern of visual field loss, and the progressive death of RGCs. RGC damage is usually not uniformly distributed throughout the eye. The uneven regional death of RGC axons in the NFL is clearly demonstrated in stained specimens of retinal whole mounts from glaucomatous animals (Figure 14). The principal treatment for glaucoma involved lowering the intraocular pressure (IOP). Increased IOP is directly related to the flow of aqueous fluid. Aqueous fluid is produced by the ciliary body and passes through the pupil into the anterior chamber of the eye (Figure 15). Aqueous fluid nourishes the lens and cornea and generally drains through the trabecular meshwork, a complicated arrangement of cells and extracellular 'beams' found at the junction of the iris, cornea, and sclera. This

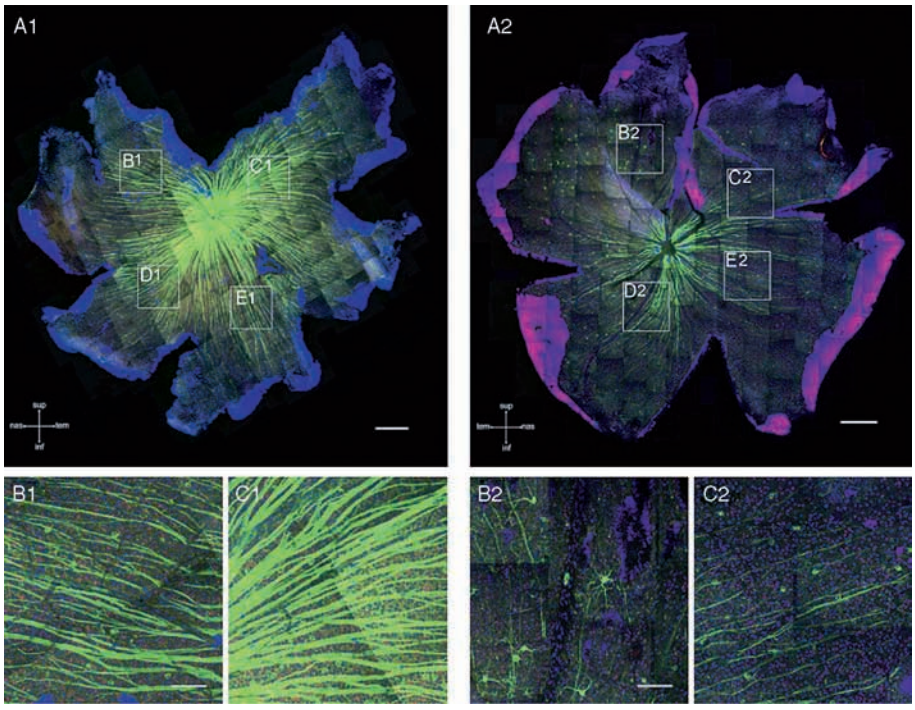

Figure 14 Comparison of ganglion cell axons entering the optic disks of normal (A1; higher power views B1, C1) and glaucomatous (A2; higher power views B2, C2) retinas. More fibers are present in the central retina (C1, C2) than in the periphery (B1, B2) but numbers in both areas are diminished in glaucoma. (Higher power views of insets D1, D2, E1, and E2 not shown.) Reproduced from Jakobs TC, Libby RT, Ben Y, et al. (2005) Retinal ganglion cell degeneration is topological but not cell type specific in DBA/2J mice. *Journal of Cell Biology* 171: 313–325, with permission from Rockefeller University Press.

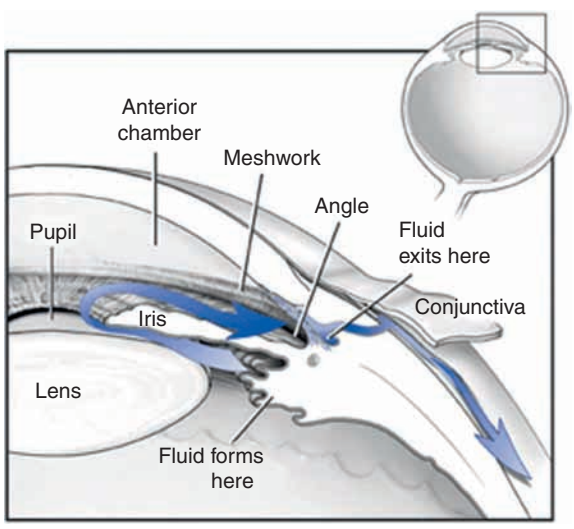

Figure 15 Illustration of fluid formation in the ciliary body and movement around the iris to the anterior chamber, where fluid exits through the canal of Schlemm (seen in **Figure 1**). Impaired drainage raises intraocular pressure, causing glaucoma. Courtsey: National Eye Institute, National Institutes of Health.

outflow pathway maintains normal levels of aqueous humor outflow. When aqueous production exceeds outflow, the pressure within the eye rises.

The most common form of glaucoma is primary open angle glaucoma (POAG). It is a late-onset disease usually accompanied by elevated IOP resulting from pathology in the aqueous outflow pathway (the trabecular meshwork and Schlemm's canal). A combination of genetic and environmental factors is likely involved in the pathology of the outflow pathway. Oxidative stress may play an important role in the pathophysiology of POAG, since the cells of the trabecular meshwork are constantly exposed to an oxidative environment. Decreased antioxidant potential, increased expression of oxidative stress markers, and increased oxidative DNA damage have all been reported in glaucoma patients.

One or more of the following events may lead to the apoptosis of RGCs observed in POAG: elevated IOP, mechanical compression, ischemic changes, glutamate neurotoxicity, astrocytic reaction, autoimmune attack, neurotrophic deprivation, intracellular calcium toxicity, changes due to reactive oxygen species, and/or nitric oxide production. Several studies suggest that the region where the axons exit the eye, the lamina cribrosa, is the primary site of mechanical damage leading to RGC death. Most likely the compromised blood flow in the optic nerve noted in glaucoma patients would cause ischemic injury to RGCs. Therefore, it is possible that excitotoxic injury (routinely seen in ischemic injury) may play a major role in the pathogenesis of glaucoma.

Figure 16 Erosion of fibers in glaucoma causes the pale area around the optic disk (yellow) and cupping, where vessels curve downward toward the optic disk. Reproduced from the Southwestern Medical Center; image available online at http://www.swmed.edu/home_pages/ophth/images/dlc-fun8.jpg, with permission from UT Southwestern Medical Center.

Clinically, the loss of RGCs is monitored by University of Texas by biomicroscopy and is seen as a progressive loss of tissue in the optic nerve head. The optic nerve head in POAG has an unusual appearance on examination with an indirect ophthalmoscope – it appears to have a halo around the rim of the optic disk due to the massive loss of optic nerve fibers (Figure 16). Visual field testing is important in documenting the location and progression of RGC loss during the progression of the disease. A scotoma is an area of relative or complete blindness in an area of the visual field. Glaucoma invariably causes isolated scotomas in the visual field, such that the patient loses the ability to see one or more peripheral regions, while vision is completely normal in other areas of the eye. As the disease progresses, the field of vision gradually narrows and blindness can result.

Risk Factors for Primary Open Angle Glaucoma

Advanced age and increased IOP are fundamental risk factors for POAG. Other predisposing factors include mutations in specific genes, toxic effects, and mechanical damage due to elevated IOP. There is broad consensus that genetic risk factors play an important role this disease. The trabecular meshwork-induced glucocorticoid response (TIGR) gene, *MYOC*, for juvenile glaucoma (J-POAG) was identified by mapping and mutational analysis. This gene codes for myocillin, a protein which plays an important role in maintaining the structural integrity of the trabecular meshwork. A mutation in this gene is a high risk factor for the onset of POAG. Other strong risk factors include positive family history of glaucoma in a first-degree relative, higher incidence in African-Americans, and central corneal thickness less than 0.5 mm.

A possible diagnostic indicator for glaucoma may be the presence of the endothelial leukocyte adhesion molecule-1 (ELAM-1). ELAM-1, a member of the cell-adhesion family of molecules (CAMs), is present in trabecular meshwork cells from glaucomatous donor eyes, compared to normal eyes. This finding suggests that the activation of a tissue-specific stress response in the aqueous outflow pathway of the eye defines the glaucoma disease phenotype.

Treatments for Glaucoma

Current therapeutic strategies for glaucoma are aimed at reducing IOP by pharmacologically decreasing aqueous humor production or by microsurgically increasing the outflow capacity. However, lowering the IOP in glaucoma may not prevent further damage.

Neuroprotection or Rescue of RGCs

Knowledge of mechanisms of neuronal death of RGCs, and prevention, delay, or reversal of the processes, are now sufficient that glaucoma therapy may be designed to preserve RGCs and their axons. Neuroprotection is a therapeutic strategy to keep RGCs alive and functional. Neuroprotective strategies proven successful in animal models of glaucoma have included blocking retinal excitotoxicity mediated by glutamate and administration of neurotrophic factors (such as brain-derived neurotrophic factor) which maintain RGC survival. To date, none of the these strategies has worked in randomized controlled clinical trails in patients with glaucoma, but the use of neuroprotective agents to treat glaucoma still remains promising.

Vitreous Humor

Eighty percent of the interior volume of the eye is made up of the vitreous humor. The primary structure of the vitreous is collagen and hyaluronic acid. In the vitreous, the predominant proteoglycans are chrondroitin and heparan sulfate.

Age-Related Changes

The vitreous is primarily a collagen gel at 2 or 3 years of age, but the gel portion decreases and the liquid portion increases steadily throughout life. The continual increase in the liquid fraction of the vitreous suggests that there is a gradual change in the vitreous collagen. Data suggest that bundles of collagen form in the vitreous with aging.

The vitreous is attached to the posterior retina, but detaches with aging in 25–30% of the population. The process of posterior vitreal detachment occurs early in

the second to third decades of life in people with high myopia, but typically occurs anywhere along the inner margin of the retina between the fourth and ninth decades of life. Detachment occurs suddenly, when fluid trapped in the middle of the vitreous humor floods into the interface between the vitreous base and the inner limiting membrane (Figure 17). Because the attachment over the macula is much firmer, the floating vitreous gel tugs on the retina. The process can cause the individual to see 'sparkles' or 'flashes' of light. When the vitreous base is firmly attached, the tugging on the retina can cause a tear in the retina, or a macular hole.

Lens

Normal Aging of the Lens

The lens is the optically clear structure located behind the iris and in front of the vitreous humor and retina (see Figure 1). The lens functions to focus light onto the retina. In order to form an image on the retina, the lens must maintain its shape, transparency, and refractive index. The lens has only one cell type, and these cells are precisely aligned to lead to the transparency of the lens. Young lens cells have organelles, but the organelles are destroyed during early development. What remains are cells made up of a dense solution of large proteins termed crystallins, specifically the α-, β-, and λ-crystallins. The α-crystallins act as molecular chaperones as well as structural proteins. As stated by Bloemendal and colleagues, "The absence of protein turnover in the lens and the ability of the lens to retain life-long transparency means that the crystallins must meet not only the requirement of solubility associated with high cellular concentration but that of longevity as well. Longevity, as it relates to proteins, is commonly assumed to be correlated with long-term retention of native structure, which in turn can be due to inherent thermodynamic stability, efficient capture and refolding of non-native protein by chaperones, or a combination of both."

The complexes formed by crystallin proteins make the cytoplasm optically homogeneous – that is, the refractive index does not change inside the cell or

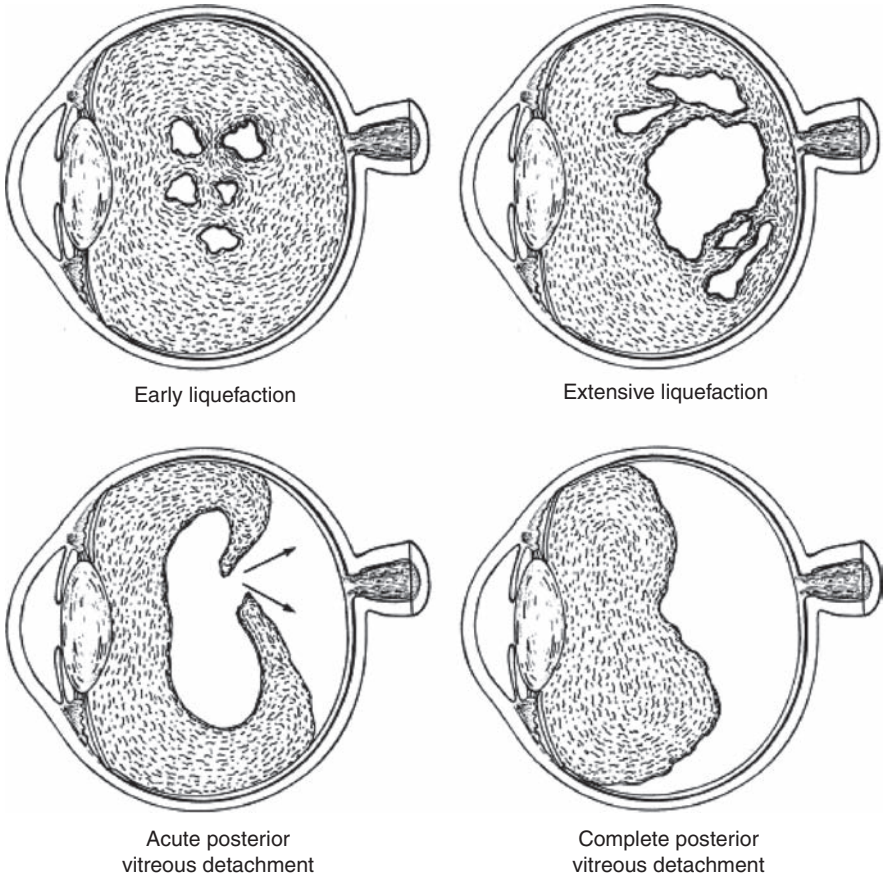

Figure 17 Liquefaction of the vitreous during aging can cause complete or partial vitreous detachment and is a risk factor for retinal detachment and macular holes. Reproduced from Bishop PN (2000) Structural macromolecules and supramolecular organisation of the vitreous gel. *Progress in Retinal & Eye Research* 19: 323–344, with permission from Elsevier Science & Technology.

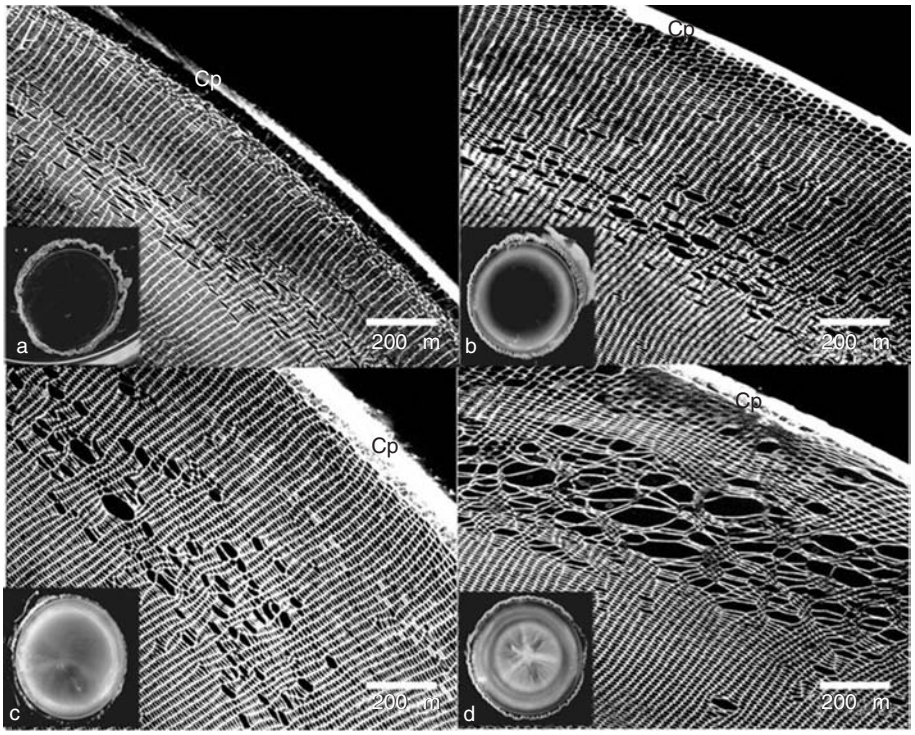

Figure 18 The regularly arrayed fiber cells in a young lens (a; Cp, capsule) result in a transparent lens (inset, a). As the fiber cells become more disorganized with damage (b, c), there is a corresponding increase in the opacity of the lens, ultimately forming a full cataract (inset, d). Reproduced from Merriman-Smith BR, Krushinsky A, Kistler J, and Donaldson PJ (2003) Expression patterns for glucose transporters GLUT1 and GLUT3 in the normal rat lens and in models of diabetic cataract. *Investigative Ophthalmology & Visual Science* 44: 3458–3466, with permission form Investigative Ophthalmology & Visual Science.

from one cell to another. As the lens ages, new layers of fibers are added to the cortex, causing the lens nucleus to become compressed, harder, and more compact. Yellowing of the lens also occurs with age. The lens increases in bulk and mass, gaining weight steadily until about age 50 years. Some lenses remain clear with advancing age; but significant protein oxidation at the center of the lens causes cataracts. The key factor in preventing oxidation in the nucleus of the lens seems to be the concentration of nuclear glutathione (L-γ-glutamyl-L-cysteinylglycine; GSH). Nuclear sclerosis progresses slowly over the years, causing only a slight change in refraction, but these changes may not significantly affect vision. Further progression of nuclear sclerosis can lead to loss of color discrimination and also loss of distance vision.

Diseases of Aging Lens

Any opacity of the lens or its capsule is called a cataract. Age-related cataracts are divided into nuclear, cortical, and posterior subcapsular types, depending on their location and clinical appearance. Lens crystallins accumulate damage with age (from ultraviolet light, oxidation, dehydration, or elevated blood sugar from diabetes), and may collapse into misfolded fibers. The misfolded fibers aggregate, producing heterogeneity in the protein concentration within the lens. This protein aggregation can lead to light scattering and to cataract formation (Figure 18). The most common treatment for cataracts is surgical removal.

It should be noted that oxidation is the hallmark of age-related nuclear (ARN) cataracts. Loss of protein sulfhydryl groups and the oxidation of methionine residues are progressive and increase as the cataract worsens. Therefore, the ARN cataract is associated with a loss of GSH in the center of the lens, and extensive modifications to the nuclear proteins, including coloration, oxidation, insolubilization, and cross-linkng.

Risk Factors and Prevention

Cataracts are the most common cause of reversible vision loss worldwide. Age is by far the major risk factor for cataract formation. Ethnic variation has been reported with different cataract types. Genetic factors may account for as much as 50% of the severity of nuclear cataracts and may also be important in the development of cortical cataracts. Many reports suggest that women are at slightly higher risk than are men to develop cataracts. Other risk factors are

of much less importance to cataractogenesis, but are ones that we may be able to control: cigarette smoking, exposure to sunlight, alcohol use, and low dietary levels of carotenoids or antioxidants. Epidemiological studies also suggest that exposure to ultraviolet B causes cortical cataract changes.

Presbyopia

Presbyopia is the loss of accommodative amplitude which occurs with age. Around age 45 years, most people are unable to hold a book at the normal reading distance and still see the text clearly (i.e., the near point of accommodation is farther away than the normal reading or working distance). Presbyopia is largely thought to be associated with age-related changes in the lens. The many changes in the lens from birth onward (increase in diameter and thickness, volume, and mass; addition of new cells; changes in curvature) occur during the decline of accommodation that occurs with age. The link between lenticular changes and presbyopia is thought to be due to the increase in bulk and mass of the lens, such that the elastic forces are less effective in changing the curvature of the lens.

Cornea

Normal Aging of the Cornea

Research in aging of the cornea has focused on the inability of the corneal endothelium to replace cells that die with age. The corneal endothelium is a single layer of metabolically active cells that lines the entire posterior surface of the cornea. The aqueous humor in the anterior chamber of the eye is in direct contact with the corneal endothelium. The loss of the corneal endothelium during a life span is substantial: about half of the endothelium is lost. As endothelial cells die, neighboring endothelial cells expand and migrate in order to maintain a confluent layer of cells. Age-related changes in the tear film that bathes the cornea cause dry eye, a condition of irritating, burning, or itching eyes found in over 10 million people in the US.

In contrast to the endothelium, the corneal epithelium heals quickly and completely. The corneal epithelium replaces itself about every 10 days. It is believed that stem cells for the corneal epithelium are located in the limbus of the eye (the region between the transparent cornea and the sclera). If the limbal epithelium is damaged, then the epithelium is unable to repair itself. Corneal transplants were developed to replace damaged tissue with healthy limbal tissue.

See also: Animal Models of Inherited Retinal Degenerations; Fovea: Primate; Optic Nerve, Optic Chiasm and Optic Tracts; Photoreceptor Adaptation; Photoreceptor Mosaic; Photoreceptors and Circadian Clocks; Photoreceptors: Physiology; Phototransduction; Retina: An Overview; Retinal Ganglion Cells: Anatomy; Retinal Ganglion Cells: Receptive Fields; Retinal Color Mechanisms; Retinal Development: An Overview; Retinal Development: Cell Type Specification; Retinitis Pigmentosa; Vision: Light and Dark Adaptation.

Further Reading

Algvere PV, Marshall J, and Seregard S (2006) Age-related maculopathy and the impact of blue light hazard. *Acta Ophthalmologica Scandanavica* 84: 4–15.

Asbell PA, Dualan I, Mindel J, et al. (2005) Age-related cataract. *Lancet* 365: 599–609.

Bloemendal H, de Jong W, Jaenicke R, et al. (2004) Ageing and vision: Structure, stability and function of lens crystallins. *Progress in Biophysics & Molecular Biology* 86: 407–485.

Bonnel S, Mohand-Said S, and Sahel JA (2003) The aging of the retina. *Experimental Gerontology* 38: 825–831.

Bourne WM and McLaren JW (2004) Clinical responses of the corneal endothelium. *Experimental Eye Research* 78: 561–572.

Budovsky A, Muradian KK, and Fraifeld VE (2006) From disease-oriented to aging/longevity-oriented studies. *Rejuvenation Research* 9: 207–210.

Congdon N, O'Colmain B, Klaver CC, et al. (2004) Causes and prevalence of visual impairment among adults in the United States. *Archives of Ophthalmology* 122: 477–485.

Gupta N and Yucel YH (2007) Glaucoma as a neurodegenerative disease. *Current Opinion in Ophthalmology* 18: 110–114.

Hageman GS, Anderson DH, Johnson LV, et al. (2005) A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proceedings of the National Academy of Sciences United States of America* 102: 7227–7232.

Harman A, Abrahams B, Moore S, and Hoskins R (2000) Neuronal density in the human retinal ganglion cell layer from 16–77 years. *Anatomical Record* 260: 124–131.

Jackson GR and Owsley C (2003) Visual dysfunction, neurodegenerative diseases, and aging. *Neurology Clinics* 21: 709–728.

Jackson GR, Owsley C, and Curcio CA (2002) Photoreceptor degeneration and dysfunction in aging and age-related maculopathy. *Ageing Research Reviews* 1: 381–396.

Quigley HA (2005) Glaucoma: Macrocosm to microcosm the Friedenwald lecture. *Investigative Ophthalmology & Visual Science* 46: 2662–2670.

Ramasamy R, Vannucci SJ, Yan SS, et al. (2005) Advanced glycation end products and RAGE: A common thread in aging, diabetes, neurodegeneration, and inflammation. *Glycobiology* 15: 16R–28R.

Ramrattan RS, van der Schaft TL, Mooy CM, et al. (1994) Morphometric analysis of Bruch's membrane, the choriocapillaris, and the choroid in aging. *Investigative Ophthalmology & Visual Science* 35: 2857–2864.

Sacca SC, Izzotti A, Rossi P, and Traverso C (2007) Glaucomatous outflow pathway and oxidative stress. *Experimental Eye Research* 84: 389–399.

Sparrow JR and Boulton M (2005) RPE lipofuscin and its role in retinal pathobiology. *Experimental Eye Research* 80: 595–606.

Stadtman ER (2004) Role of oxidant species in aging. *Current Medicinal Chemistry* 11: 1105–1112.

Truscott RJ (2005) Age-related nuclear cataract-oxidation is the key. *Experimental Eye Research* 80: 709–725.

Zarbin MA (2004) Current concepts in the pathogenesis of age-related macular degeneration. *Archives of Ophthalmology* 122: 598–614.

Relevant Websites

http://www.afb.org – American Foundation for the Blind (link to CareerConnect > For Employers > Visual Impairment and Your Workforce for information on normal changes in the aging eye).

http://webvision.med.utah.edu – John Moran Eye Center, University of Utah (WebVision: The Organization of the Retina and Visual System).

http://www.nei.nih.gov – National Eye Institute (U.S. National Institutes of Health).

http://www.stlukeseye.com – St. Lukes Cataract & Laser Institute (Florida, USA).

http://www.agingeye2.net – The Eye Digest.

Sensory Ganglia

R E Papka, Northeastern Ohio Universities College of Medicine, Rootstown, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Ganglia are enlargements along bundles of nerve cell processes derived from collections of nerve cell bodies. Several categories of ganglia are identified: motor ganglia of the autonomic nervous system, special sensory ganglia of the auditory–vestibular nerves, and general sensory ganglia associated with spinal nerves (e.g., dorsal root ganglia) and cranial nerves (e.g., nodose ganglion of cranial nerve X and trigeminal ganglion of cranial nerve V) (Figure 1). The focus of this article is general sensory ganglia.

Types and Locations of Sensory Ganglia

General sensory ganglia are the spinal (dorsal root) ganglia (DRG) and cranial nerve ganglia(CG). DRG are located along dorsal roots of the cervical, thoracic, lumbar, and sacral spinal nerves near their entry to the spinal cord (Figure 1(a)). The CG are located along the roots of cranial nerves V (trigeminal), VII (facial), IX (glossopharyngeal), and X (vagus) as they enter the brain stem. Sensory neurons (though not necessarily distinct sensory ganglia) are described in the walls of some tubular organs, particularly the gastrointestinal tract, where they function in reflexes, responses to distention, and sampling local and luminal environments. Each neuron cell body (soma) (Figures 2 and 3) has a single process (Figures 3 and 4) that divides into two: a peripherally directed process and a centrally directed process. Peripheral processes (neurites, or more specifically dendrites) respond to stimuli which are conducted toward the DRG and CG, and are relayed by the central process into the spinal cord or brain stem (Figure 1(a)). Sensory information is derived from somatic structures (e.g., cutaneous areas) and visceral structures (e.g., organs and body cavities). Sensory neurons have a spatial organization in ganglia; in general, proprioceptive neurons are located mainly in the ventral half of a ganglion whereas cutaneous neurons are broadly distributed throughout the ganglion.

Development and Neurotrophin Dependence

Sensory neurons develop embryologically from neural crest cells. Crest cells migrate from the developing neural tube to form ganglia near the future spinal cord and brain stem. The developing neurons are initially unipolar; however, this single process bifurcates with one branch growing peripherally, to reach a target structure, and one growing centrally, for connection with the central nervous system (CNS) (Figures 1). The migration, differentiation, and survival of ganglionic neurons are dependent on factors produced by the neuron and on protein growth/trophic factors produced by target tissues. Growth-associated protein of 43 kDa (GAP-43) (Figure 5) is an integral membrane protein that is such a factor and is expressed at elevated levels by developing or regenerating neurons during axon growth. Neurotrophins (NTFs) are a family of trophic factors that are survival substances essential during development of the neurons, though these factors continue to exert influences throughout the life span of neurons. Actions of NTFs are mediated through receptors expressed on sensory neurons – three high-affinity receptor tyrosine kinases (TrkA, TrkB, and TrkC) and low-affinity $p75^{NTR}$ receptors. Among the NTFs are nerve growth factor (NGF), brain-derived neurotrophic factor (BDNF) (Figure 6), and neurotrophin-3 and neurotrophin-4/5 (NT-3 and NT-4/5). NGF utilizes TrkA; BDNF and NT-4/5 utilize TrkB; NT-3 primarily activates TrkC. All NTFs may bind $p75^{NTR}$ receptors. Another family of trophic factors is glial-derived neurotrophic factor (GDNF); GDNF utilizes a receptor complex for signaling. Many other growth factors influence sensory neurons, such as transforming growth factor-β (TGF-β), fibroblast growth factors (FGFs), insulin-like growth factors (IGFs), leukemia inhibitory factor (LIF), and ciliary neurotrophic factor (CNTF). Trophic factor requirements correlate with specific sensory neuron populations (e.g., neuron size) and transmission of specific modalities (e.g., nociception, thermal information, or mechanoreception). For example, NGF supports small-sized and nociceptive neurons; though approximately one-half of nociceptors are dependent on NGF, the other half are sensitive to other NTFs such as GDNF. BDNF supports mechanoreceptive populations and NT-3 supports large proprioceptive neurons.

Interactions between trophic factors and neurons are complex and new information is emerging from gene-deficient animals. For example, lack of NGF, BDNF, and NT-3 genes results in severe neuronal deficits and early postnatal death. In BDNF knockout mice, sensory ganglion cells forming slowly adapting mechanoreceptors are less sensitive, there is a reduction in number of cells in trigeminal ganglia, and the defects are worse when NT-4 is also deficient. In NT-3 knockouts, there is loss of large ganglion cells and absence of muscle spindles and Golgi tendon organs.

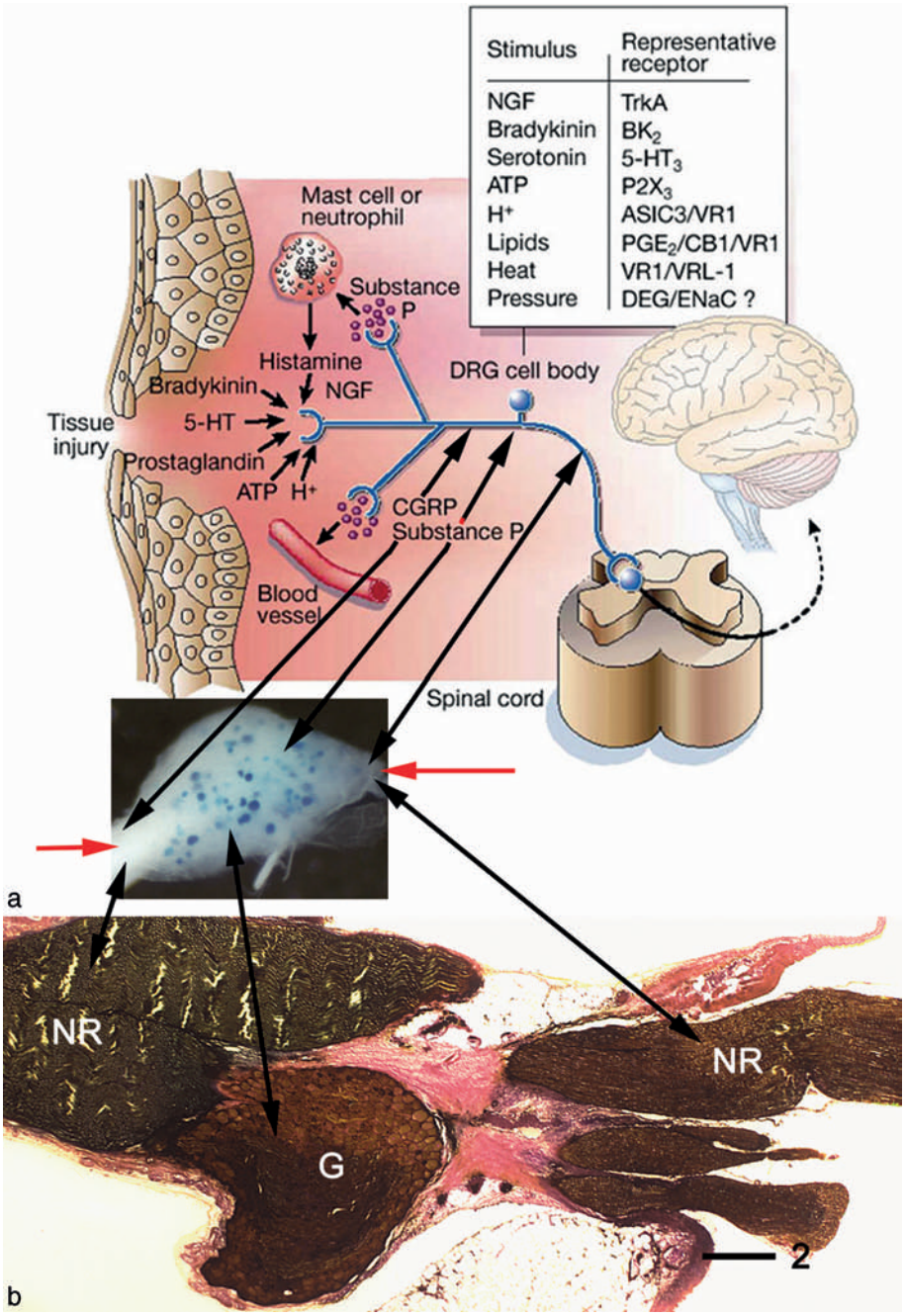

Figure 1 (a) Diagram illustrating a sensory neuron cell body in a dorsal root ganglion (DRG) and its peripheral and central processes (blue). The peripheral process (functionally a dendrite) ends as primary sensory nerve endings in tissue near an epithelium (left), and the central process (functionally an axon) terminates in the spinal cord (right). The branching of the peripheral process illustrates the basis for axon reflexes; such reflexes are the basis for neurogenic inflammatory reactions. Also shown are examples of neurotransmitters (e.g., substance P and calcitonin gene-related peptide (CGRP)) that are released from peripheral sensory nerve endings and a number of substances that can influence activity of the nerve ending. Some of the latter substances are the main components of the acidic inflammatory soup during inflammation, including peptides (bradykinin), lipids (prostaglandins), neurotransmitters (serotonin (5-HT) and ATP), and neurotrophins (nerve growth factor (NGF)). Each of these factors sensitizes (lowers the threshold) or excites the terminals of the C-fiber that acts as a nociceptor by interacting with cell surface receptors expressed by these neurons. Examples of these stimuli and representative molecular receptor targets are indicated. Activation of the nociceptor not only transmits afferent messages to the spinal cord dorsal horn (and from there to the brain), but also initiates the process of neurogenic inflammation. This is an efferent function of the nociceptor whereby release of neurotransmitters, notably substance P and CGRP, from the peripheral terminal induces vasodilation and plasma extravasation (leakage of proteins and fluid from postcapillary venules), as well as activation of many nonneuronal cells, including mast cells and neutrophils. These cells in turn contribute additional elements to the inflammatory soup. Inset: Whole mount of a dorsal root ganglion, illustrating its enlargement along the sensory dorsal root fibers (red arrows) of a spinal nerve (blue cells are the result of herpes virus-mediated gene transfer of β-galactosidase). Double-headed black arrows indicate corresponding parts of the dorsal roots and ganglion in the sensory neuron diagram

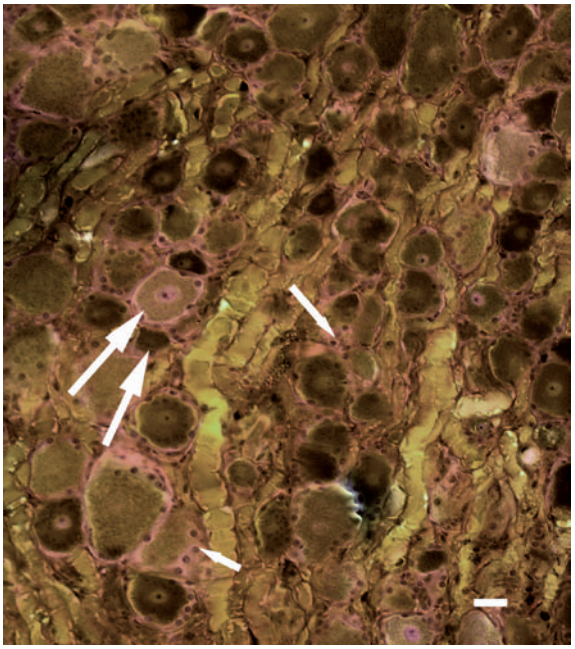

Figure 2 Medium-magnification view of a dorsal root ganglion, showing the numerous neuron cell bodies (somas; large arrows) that exhibit variations in sizes and staining. Many small nonneuronal cells are evident between neurons (small arrows). Scale bar = 25 μm.

In TrkC knockout mice, spinal proprioceptive afferents are absent in limb skeletal muscles; however, in the same animals, proprioceptive afferents from the trigeminal nerve innervate head muscles and form muscle spindle sensory endings. This implies that spinal and trigeminal proprioceptive afferents have different neurotrophin-supporting systems during differentiation and survival. In $p75^{NTR}$ receptor knockouts, most types of sensory axons are reduced by 50%. On the other hand, some neurons exhibit a switch in trophic factor dependence during development; for example, IB4-binding (isolectin from *Griffonia simplicifolia*) sensory neurons switch from NGF to GDNF dependence in early postnatal life.

Structure

Cellular Constituents

Each DRG and CG contains several thousand neurons (Figures 3 and 4); numbers vary at different levels of the spinal cord and brain stem. For example, neurons are most numerous in cervical and lumbar DRG that serve the upper and lower extremities. The trigeminal ganglia are the largest sensory ganglia

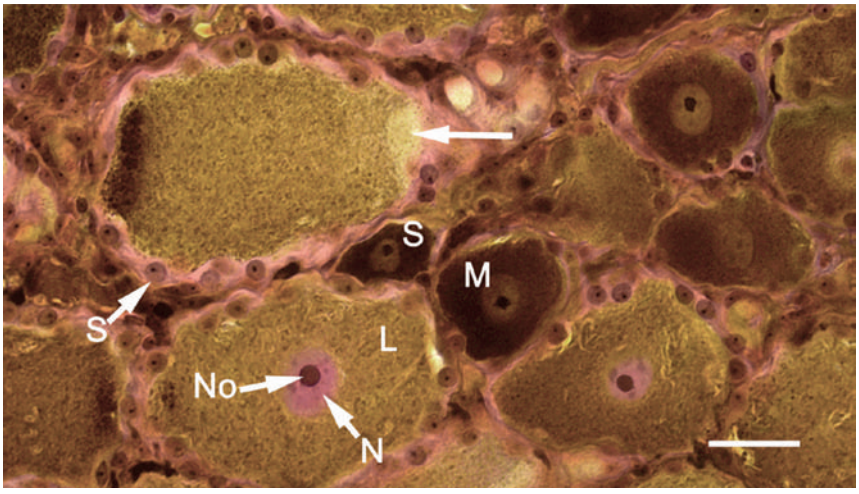

Figure 3 Higher magnification view of a dorsal root ganglion, illustrating neurons and nonneuronal cells. Neurons are classified into three size categories, large (L), medium (M), and small (S within dark brown cell). Within the neurons a centrally located nucleus (N) and nucleolus (No) are visible. Surrounding and enwrapping each neuron is a collection of nonneuronal satellite cells (S adjacent to arrow). Such cells isolate the neurons from the connective tissue extracellular space as well as contribute to the metabolism of the neuron. The granularity of the neurons results from the content of Nissl bodies (cisternae of rough endoplasmic reticulum). The lightly stained area (top horizontal arrow) of a large neuron is the axonal hillock, and is the site of origin of the single process that emanates from the cell body. Scale bar = 25 μm.

and in the adjacent dorsal root sensory ganglion micrograph (see (b)). Diagram adapted from Julius D and Basbaum AI (2001) Molecular mechanisms of nociception. *Nature* 413: 203–210. Inset reproduced courtesy of Dr. Steven P Wilson, Department of Pharmacology, Physiology and Neurobiology, University of South Carolina School of Medicine. (b) Low-magnification micrograph of a dorsal root sensory ganglion (also called spinal ganglion). The ganglion (G) is located along the nerve roots (NR) that will connect to the spinal cord and project to peripheral targets. Scale bar = 500 μm.

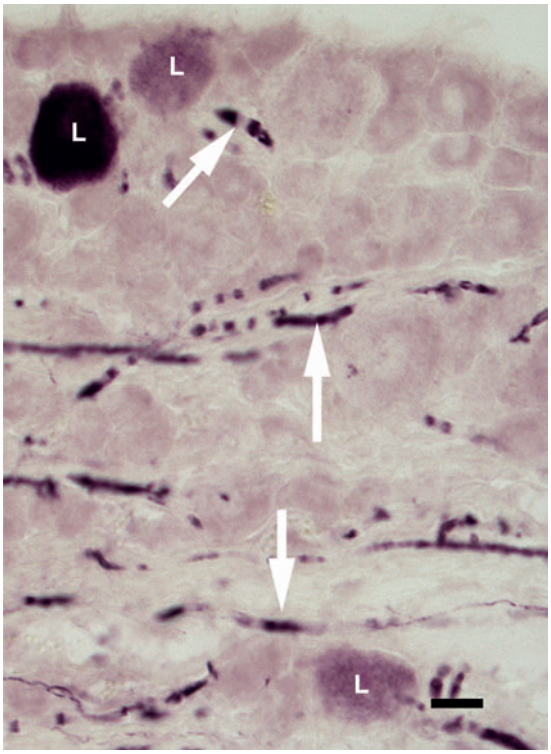

Figure 4 Dorsal root ganglion, immunostained for calretinin. There are subclassifications of sensory neurons based on size, neurochemistry, and other characteristics. In this view, a ganglion is specially immunostained by an antibody against the calcium-binding protein calretinin. A subclass of predominately large neurons (L) is stained, and the intensity of staining varies from moderate to intense, which may indicate a further subclass of neurons. A number of neurites coursing through the ganglion are also immunoreactive for calretinin (arrows). Such neurites appear to be large myelinated processes that likely originate from large neurons. Scale bar = 25 μm.

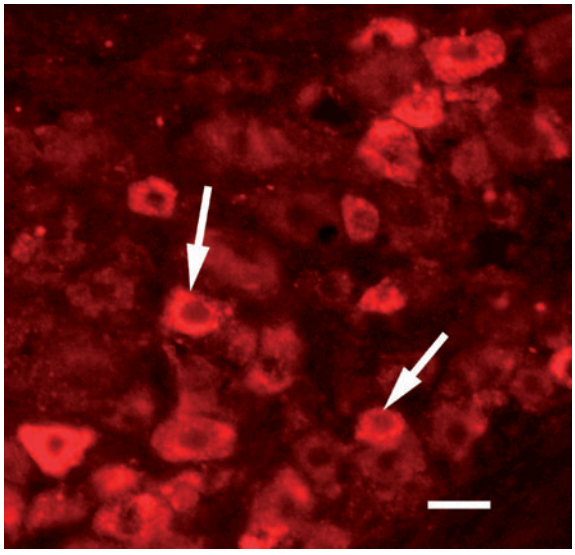

Figure 5 Dorsal root ganglion, immunostained with an antibody against the growth-associated protein GAP43. A subpopulation of mainly small and medium-sized neurons (arrows) is revealed by the immunostaining. Scale bar = 25 μm.

and innervate the face and head. Ganglionic neurons are organized as clusters (Figures 2, 7, and 8) among bundles of neurites (axons and dendrites) (Figures 2–4 and 7–10). Their somas are rounded and contain a centrally located nucleus and prominent nucleolus (Figure 3). They contain abundant Golgi complexes (Figures 10–13) and cisternae of rough endoplasmic reticulum (Figure 3) that are essential for synthesis of proteins (e.g., neuropeptide transmitters, structural

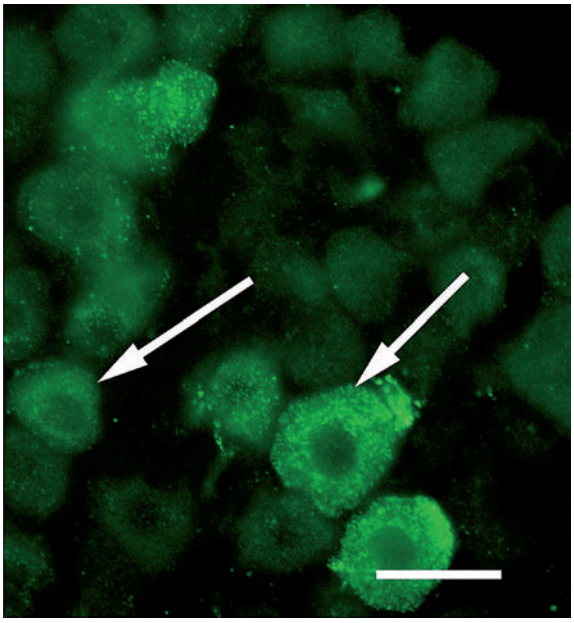

Figure 6 Dorsal root ganglion, immunostained with an antibody against brain-derived neurotrophic factor. A subpopulation of mainly small and medium-sized neurons (arrows) is revealed by the immunostaining. Scale bar = 25 μm.

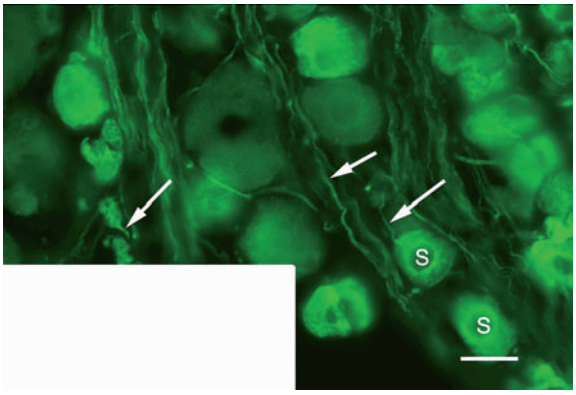

Figure 7 Dorsal root ganglion, immunostained with an antibody against the neurofilament protein peripherin. This antibody immunostains only the small C-type neurons (S) of the ganglia. In addition, it is evident that some neurites extending from these small neurons are immunopositive for peripherin (arrows). Such processes are classified as small unmyelinated C-fibers that are slow conducting and transmit nociceptive and thermal information. Note that clusters of neurons are separated by bundles of nerve fibers. Scale bar = 25 μm.

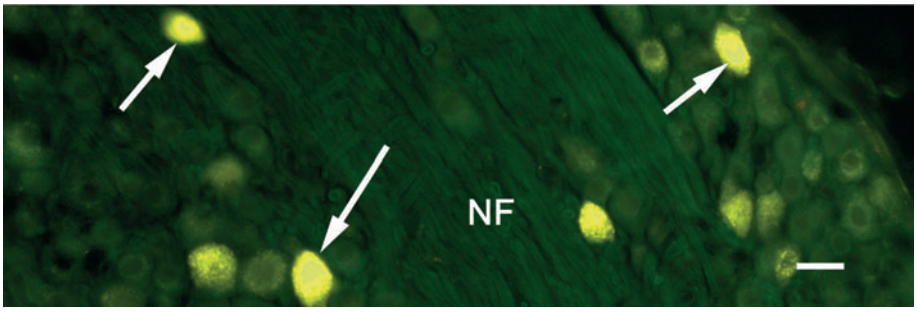

Figure 8 Dorsal root ganglion, illustrating a subpopulation of neurons; these neurons fluoresce yellow (arrows) due to their content of the retrograde tracer Fluorogold. The tracer was injected into the wall of the uterine cervix, endocytosed by the peripheral sensory nerve terminals in the cervix, and retrogradely transported by axoplasmic flow to the cell bodies of neurons in the dorsal root ganglia which innervate the cervix. This is one method to 'mark' or 'identity subgroups of neurons that innervate a specific target. Note that neurons tend to be clustered and separated by bundles of nerve fibers (NF). Scale bar = 25 μm.

proteins, and cytoplasmic enzymes). Such features reflect that sensory neurons with large somas and extensive neurites have considerable volumes of protein-rich cytoplasm to maintain. The rough endoplasmic reticulum is organized into aggregates called Nissl bodies, giving the cytoplasm a fine granular appearance (Figure 3). Additionally, the soma may stain lightly, moderately, or intensely with general histologic stains (Figures 2 and 3), depending on their organellar and cytoplasmic content.

Sensory neurons are described as pseudounipolar because a single process (neurite) emanates from the soma (Figures 3, 4, 7, 9 and 10), but then bifurcates, sending one process (axon) to the CNS where it associates synaptically with a CNS neuron (Figure 1(a)). The other process (dendrite) innervates a peripheral target, such as the skin, muscle, or visceral organ, where it terminates as a receptor ending. This ending can be a specialized (morphologically) receptor (e.g., Meissner's corpuscle) or an unspecialized free ending (Figure 8). Special membrane-associated receptor molecules are inserted in the membranes of these terminals and provide responsiveness to specific stimuli (e.g., ions, cytokines, neurotransmitters). Stimuli are transduced at terminals of sensory neurons and this information is conveyed toward the ganglia.

Since sensory neurons and their processes are excitable, they must be protected by supporting cells equivalent to neuroglia in the CNS. The somas are encapsulated by satellite cells that are small and closely apposed to the neuron somas (Figure 3). They isolate neurons from the connective tissues and extracellular milieu, and influence their metabolism. Schwann cells are similar to satellite cells except that they are associated with axons and dendrites. They produce a lipoprotein layer called myelin around neurites of large and medium-sized neurons (basically, myelin insulates the neurites and increases the conduction velocity of action potentials). Additionally,

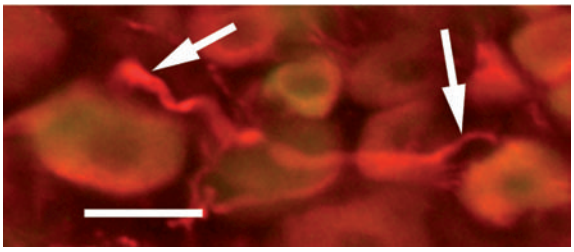

Figure 9 Higher magnification view of peripherin-immunoreactive C-fibers (arrows) in the ganglion. Scale bar = 25 μm.

Schwann cells are important in guiding and enhancing regeneration of damaged peripheral processes. Moreover, supporting cells appear to be metabolically and functionally involved with neurons in terms of transmitter metabolism and signaling.

Normally there are not synaptic structures in sensory ganglia. However, there are suggestions that neuropeptide transmitters are released in ganglia to influence adjacent sensory neurons. On the other hand, sympathetic innervation of sensory neurons is described in traumatic pathological conditions (discussed later). However, the peripheral processes of sensory neurons make contact with and influence peripheral autonomic neurons (e.g., in the pelvic and enteric ganglia).

Connections to the CNS

Activation of the peripheral nerve ending generates an action potential which conveys the sensory stimulus toward the cell body in the ganglia, and then to the CNS via the axon. Information is processed within the spinal cord dorsal horn or brain stem nuclei and then relayed to higher centers of the CNS and interpreted or perceived.

Central terminations of sensory axons enter into a complex neuronal circuitry. There is a differential pattern of distribution and termination of the centrally

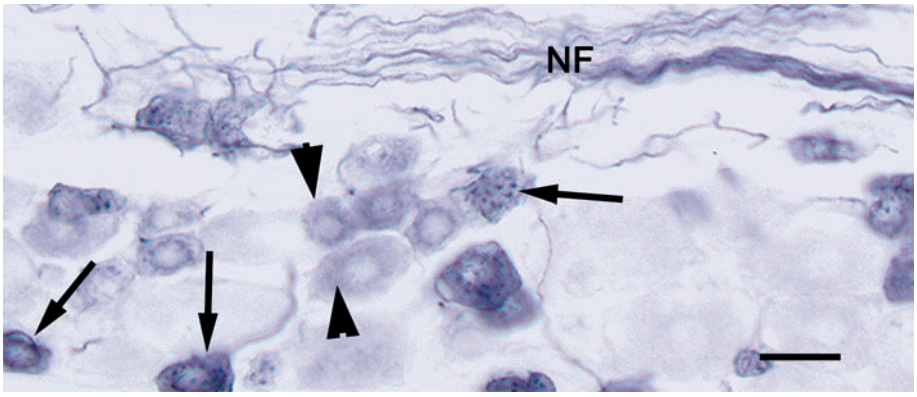

Figure 10 Dorsal root ganglion, labeled with the biotinylated-lectin IB4 (isolectin from *Griffonia simplicifolia*). This lectin reportedly binds to peptide-poor, glial-derived neurotrophic factor-responsive small neurons. There is a subpopulation of small neurons (arrows) which are IB4-positive and others that are IB4-negative (arrowheads). Many neurites (NF) are also evident and bundled, separating groups of neurons. Scale bar = 25 μm.

directed processes in the laminae of the spinal cord. Larger heavily myelinated processes emanating from larger neurons enter the cord most medially and terminate in deeper laminae. Smaller unmyelinated fibers of smaller neurons enter more laterally and terminate in more superficial laminae.

Neuron Variability

Size, Conduction Velocity, and Sensory Modality Conducted

Neurons of sensory ganglia form heterogeneous groups of cells varying in size, response to sensory modalities, signal conduction velocity, neurochemistry (neurotransmitters and receptors), development and survival dependence on neurotrophins, and tissue/organ innervated.

Sensory neurons are classified into three subgroups based on diameter: small (<20 μm), medium (21–40 μm), and large (>40 μm) (Figures 2–4, 7, 10, 12, and 13). Functional differences exist among these groups. Small neurons have thin processes (0.3–1.3 μm in diameter) that lack myelin (referred to as unmyelinated C-fibers) and contain the neurofilament protein peripherin (Figures 7 and 9). Such fibers conduct signals (e.g., nociception, temperature) relatively slowly (0.7–2.3 m s^{-1}). Medium-sized neurons have lightly to moderately myelinated processes (called A-δ fibers) and conduct signals faster (1.7–7 μm in diameter, conduct at 3–15 m s^{-1}). Such fibers are mechanoreceptive. Large neurons have moderately to heavily myelinated processes (A-β and A-α fibers) and are the fastest conducting (for A-β and A-α fibers, respectively: 7–15 μm in diameter, conduct at 20–80 m s^{-1}; 13–19 μm in diameter, conduct at 75–120 m s^{-1}). The latter fibers conduct proprioceptive information.

Small C-type neurons are receiving considerable attention due to their involvement in nociception and inflammation. Such C-neurons are subtyped into groups based on neurotrophic factor dependence, cytochemical markers (such as expression of the neurofilament protein peripherin (Figures 7 and 9), and pattern of termination in the spinal dorsal horn:

- Group I: Classified as peptide rich; express pro-inflammatory neuropeptides (substance P (SP), and calcitonin gene-related peptide (CGRP)); function in neurogenic inflammation, respond to NGF and the sensory neurotoxin capsaicin, express the TRPV1 (VR1) receptor, and terminate in the spinal laminae I and II_{outer}.
- Group II: Neuropeptide poor; respond to GDNF, bind lectin IB4, and project to spinal lamina II_{inner}.

Groups I and II most likely are not exclusive and distinct classes of neurons because there is overlap of characteristics of both groups in certain subpopulations of cells. Furthermore, adding electrophysiological current signatures of acutely dissociated DRG neurons as another characteristic allows further subtyping of small neurons to nine types.

Neurochemistry

Transmitters Sensory neurons express peptide and nonpeptide neurotransmitters – for example, the peptides CGRP (Figures 11–13), SP, neurokinin-A (NK-A), secretoneurin (SN), galanin (GAL), somatostatin (SOM), vasoactive intestinal peptide (VIP), cholecystokinin , angiotensin II, bombesin, prolactin, oxytocin, and pituitary adenylate cyclase-activating peptide (PACAP). Moreover, neurons exhibit calcium-binding proteins (e.g., calretinin) (Figure 4), the gas messenger nitric oxide, and the excitatory amino acids glutamate and aspartate. Expressions of these substances in sensory neurons have variability based on species, neuron size, target organ innervated, and

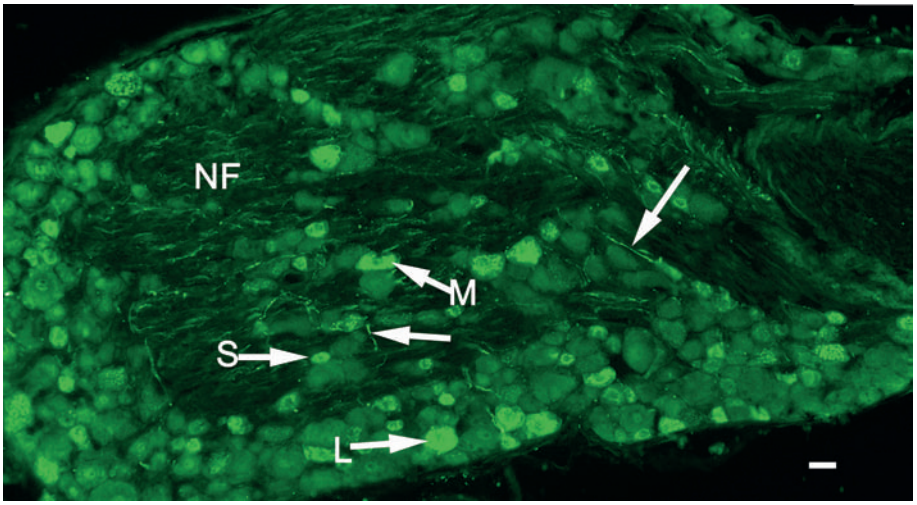

Figure 11 Low-magnification view of a dorsal root sensory ganglion, immunostained for the neuropeptide calcitonin gene-related peptide (CGRP). The sensory neurons of the ganglion are concentrated at the periphery of the ganglion (bottom of the figure) and many neurites (NF) populate the interior of the ganglion and are mixed with neurons. There are subclasses, small (S), medium (M), and large (L), of neurons that are immunoreactive to CGRP; also many neurons are not immunoreactive for CGRP. CGRP-immunoreactive neurites are also evident in the central part of the ganglion (arrows). Scale bar = 25 μm.

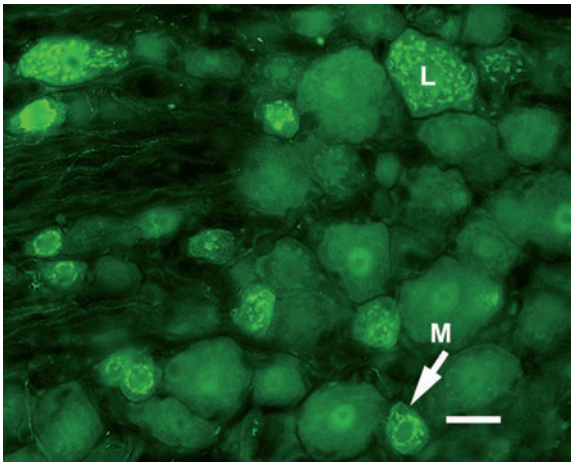

Figure 12 Higher magnification view of a dorsal root ganglion, illustrating neurons immunoreactive for calcitonin gene-related peptide (CGRP). Only a subpopulation of neurons is CGRP-positive, and some neurons are large (L) and medium (M) in size; many neurons are negative for CGRP. Scale bar = 25 μm.

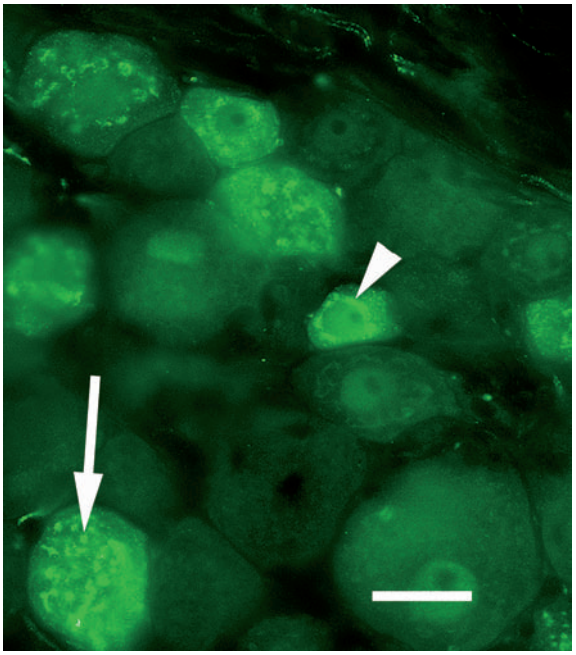

Figure 13 Subpopulations of neurons in the dorsal root ganglion are immunopositive for calcitonin gene-related peptide (CGRP). In some medium-sized neurons, multiple Golgi complexes appear immunopositive (arrow), and in small neurons (arrowhead) an immunostaining Golgi complex is evident. This staining pattern may indicate that a propeptide is staining. Note that nonneuronal satellite cells enwrapping the sensory neurons are not immunoreactive for CGRP. Scale bar = 25 μm.

pathophysiological state. Most, if not all, sensory neurons contain more than one neurotransmitter. However, even though most primary afferent neurons use glutamate as a neurotransmitter, recent evidence suggests that glutamate may not be co-expressed in peptidergic neurons.

Receptors Further classification of ganglionic-sensory neurons stems from their response to certain agents and receptor expression. For example, small C-type neurons are sensitive to capsaicin (pungent ingredient of peppers). Capsaicin-sensitive neurons are destroyed by treatment of neonates with capsaicin, whereas treatment during adulthood causes depletion of neuropeptide transmitters. Receptor types fall into two general classes: ligand-gated channels

and G-protein-linked receptors. Examples of substances operating through ligand-gated channels are capsaicin, serotonin (5-hydroxytryptamine; 5-HT), adenosine triphosphate (ATP), protons, and glutamate. Substances operating through G-protein-linked receptors are opiates, bradykinin, SOM, neuropeptide Y, adenosine, and prostaglandin E_2 (PGE_2). As indicated previously, tyrosine kinase-linked receptors respond to NTFs.

Some sensory neurons express temperature-sensitive transient receptor potential vanilloid ($TRPV_{1-6}$) receptors (e.g., receptors responding to various temperatures and to nonthermal stimuli such as acid, the vanilloid compound capsaicin, anandamide, and certain lipooxygenase products, and hypotonicity). TRPV1 acts as a polymodal transducer for nociceptive stimuli in primary sensory neurons. Purinergic receptors (P2; responsive to ATP) (Figure 14) are expressed as P2X ionotropic receptors and P2Y G-protein-linked receptors. P2X receptors are localized predominantly to small and medium-sized neurons and may have a role in pain. P2Y receptors are located mainly in large neurons and may play a role in touch sensation. Other receptors include NK-1 receptor for substance P (Figure 15), CGRP receptor, steroid hormone receptors for estrogen and androgens (Figures 15–18), opioid receptors

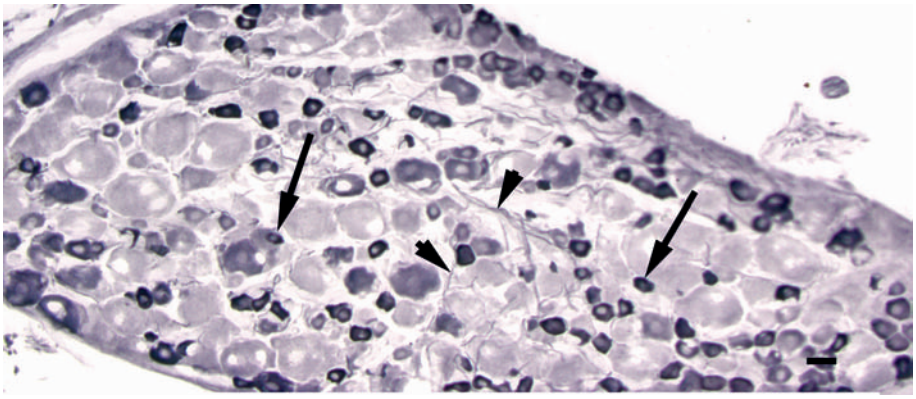

Figure 14 Low-magnification view of a dorsal root sensory ganglion, immunostained with an antibody against the purinergic receptor $P2X_3$. Note that a subpopulation of neurons is immunopositive and most are of the small size (arrows). Numerous neurites are also $P2X_3$ immunoreactive (arrowhead). Scale bar = 25 μm.

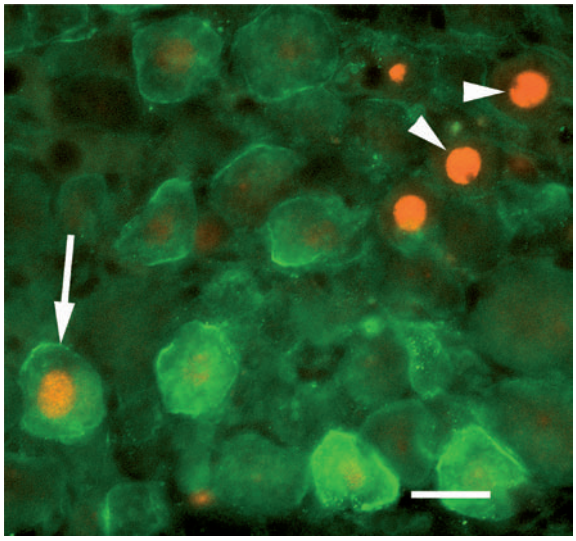

Figure 15 Dorsal root ganglion, double immunostained for estrogen receptor-α (red nuclear localization) and the neurokinin-1 receptor (green cytoplasmic and membrane associated). Subsets of neurons contain estrogen receptor-α (arrowheads) whereas others contain both estrogen receptor-α and neurokinin-1 receptor (arrow). Scale bar = 25 μm.

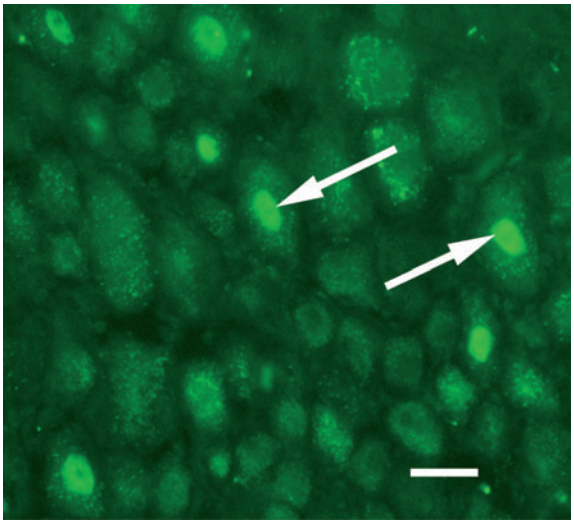

Figure 16 Male rat dorsal root ganglia, immunostained to reveal androgen receptors. A subset of sensory neurons of this ganglion expresses androgen receptors. Such receptors are nuclear in location (arrows). Scale bar = 25 μm.

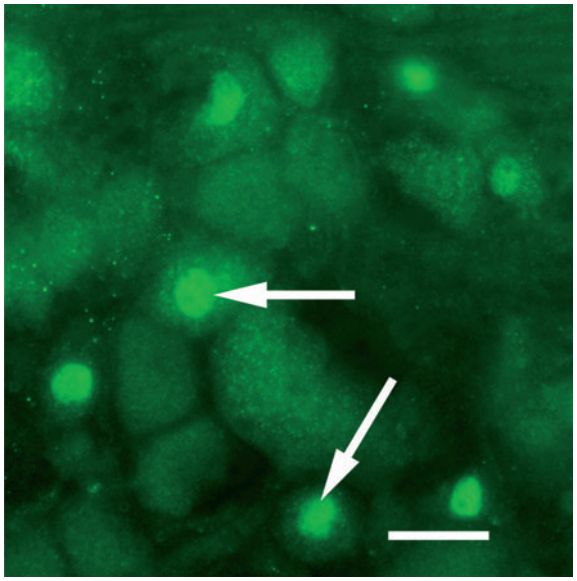

Figure 17 Female rat dorsal root ganglia, immunostained to reveal androgen receptors. A subset of sensory neurons expresses androgen receptors. Such receptors are nuclear in location (arrows). Scale bar = 25 μm.

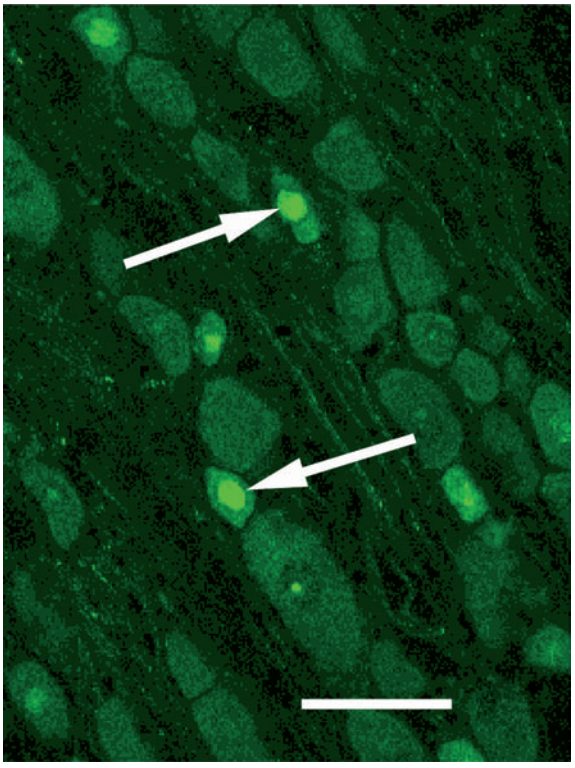

Figure 18 Female rat dorsal root ganglion, immunostained for estrogen receptor-α. A subset of neurons exhibits nuclear immunolocalization of the steroid hormone receptor (arrows). Scale bar = 50 μm.

involved in pain modulation, 5-HT receptors (especially in the gastrointestinal tract, where they influence nausea and reflexes, and on cutaneous afferents, where they modulate inflammation), and D_2 dopamine receptors.

Steroid Influences

Estrogen (E_2) and androgens bind to nuclear estrogen receptors (ERs) and androgen receptor (ARs), respectively, in DRG neurons (Figures 15–18) and initiate changes in gene expression and alterations in cellular functions. Some, but not all, of these neurons innervate reproductive organs. Sex steroids modulate neurochemical features at many sites in the CNS that involve both reproductive and nonreproductive processes. Recent data indicate that rising serum E_2 over pregnancy alters the neurochemistry of DRG neurons during gestation. E_2 stimulates synthesis of SP, CGRP, SN and, PACAP and modulates expression of $P2X_3$ receptors of sensory nerves. Since sensory nerves are influenced by E_2, they in turn could have afferent and efferent roles in normal pregnancy and parturition. Finally, sensory neurons transmit information such as pain and neuroendocrine reflexes from their target organs, and, importantly, the presence of estrogen receptors makes them a likely target of estrogen action.

Functions

Conduction of Modalities

Signal detection Peripheral sensory nerve terminals have multiple functions: reporting normal and abnormal sensations, releasing transmitter substances in response to local factors, and releasing trophic factors to maintain the structure and function of cells in their local environment (e.g., epithelial cells). Signal detection by nerve terminals varies based on location, morphology or type of ending (e.g., encapsulated vs. free nerve endings), and molecular receptor properties. Sensory terminals are populated with specialized receptors that transduce stimuli (e.g., thermal, mechanical, and chemical) into electrical potentials which generate action potentials in their parent axons. Action potentials are propagated through neurites to communicate the modality and intensity of stimuli toward neuron somas and to the CNS. The physiology of these endings is usually dependent on intracellular [Ca^{2+}]. Generally, nociceptors are free nerve endings; mechanoreceptors or touch receptors show some type of encapsulation and are located in cutaneous and subcutaneous regions, and receptors located in muscle, joints, and deep structures monitor mechanical forces as proprioceptors.

Depolarization of nociceptors (mainly polymodal C-fibers) elicits pain and inflammation. In addition to transmitting signals to the CNS, nociceptors release peptides (SP and CGRP) peripherally to produce vascular leakage and vasodilation, leading to inflammation and tenderness at the site of irritant application

(see later and Figure 1(a)). Noxious chemicals (e.g., capsaicin) mediate these effects by activating a TRPV1 receptor belonging to a family of transient receptor potential (TRP) channels that are key components of signal transduction pathways in sensory systems. TRP channels respond to temperature, touch, pain, osmolarity, and other stimuli and are activated by stimulation of G-protein-coupled receptors (GPCRs) and receptor tyrosine kinases. For example, TRPV1 is activated by ligands, including the vanilloid capsaicin, and anandamide (endogenous ligand for cannabinoid receptor 1), as well as by moderate temperatures (>42 °C) and decreased pH. TRPV1 also participates in purinergic signaling from the urinary bladder epithelium. TRPV2 may mediate high-threshold noxious heat sensation (>52 °C), perhaps through lightly myelinated A nociceptors. Increased temperature also activates TRPV3 (>31 °C) and TRPV4 (>25 °C).

Touch or mechanoreceptors are distributed in various tissues (e.g., skin and hair follicles). A variety of mechanoreceptor neurons respond to touch; most show some degree of encapsulation, which determines the nature of the stimulus to which they respond. All of these receptors work in fundamentally the same way: touch produces deformation or changes the nerve ending, which alters the ionic permeability of the receptor membrane generating a depolarizing current, ultimately triggering an action potential. The peripheral endings of such neurons use touch-transducing molecules to generate the depolarizing current. A channel superfamily – epithelial amiloride-sensitive sodium channel (ENaC) – is expressed nearly exclusively in touch receptor neurons. Mammalian members of the ENaC superfamily fall into two subfamilies that are expressed in either epithelia (ENaCs) or neurons (acid-sensing ion channels activated by extracellular protons; ASICs).

Efferent Function of Sensory Nerves

DRG neurons not only relay information centrally, they also have peripheral neuroeffector functions by releasing mediators from their peripheral terminals. As early as 1901, the physiologist Bayliss demonstrated that stimulation of dorsal root sensory nerves induced peripheral vasodilation (Figure 1(a)). This phenomenon was eliminated by destruction of the DRG or peripherally directed processes of DRG neurons, suggesting that the vasodilation was caused by antidromic impulses in sensory nerves. Subsequent experiments demonstrated that cutaneous axon reflexes occur without connections to the CNS. Such phenomena utilize primarily polymodal C-fibers, but also A-δ fibers.

Axon reflexes involve activation of a C-fiber, which conducts information to the CNS as well as to other branches of the stimulated afferent nerve ending (Figure 1(a)). In this way, action potentials spread into branches of the activated terminal and mediators are released from all activated branches. Through this mechanism, stimulation of nociceptors releases neuropeptides resulting in vasodilation (flare), release of histamine from mast cells (leading to venular permeability and edema – protein extravasation), and influx of immune cells. Taken together, these actions constitute neurogenic inflammation. Neurogenic inflammation has roles in inflammatory responses, repair, and tissue remodeling. Substances that stimulate nerve endings can originate in local tissues and cells or from external sources and include ATP, PGE_2, bradykinin, 5-HT, protons, and histamine. The C-fibers release SP, CGRP, GAL, PACAP, SN, and other tachykinins from their peripheral ends. This expanded view of the sensory neuron stems from studies of the expression and function of their neuropeptides. It is emphasized here that many (if not all) sensory peptidergic nerves contain more than one transmitter, so peripheral stimulation and axon reflexes may produce multiple and complex effects. Such functions of sensory nerves play important roles in cutaneous blood flow, renal excretory function, protection during gut inflammation, release of proinflammatory agents, rheumatoid arthritis, and airway disease.

Pathophysiology

Plasticity in Sensory Neurons

Following injury to peripheral sensory endings, long-lasting changes can occur in the somas; for example, such changes may be in soma neurochemistry, including alterations of expression of neuropeptide neuro transmitters, receptors, immediate-early genes (e.g., c-*jun*), and gene-products such as growth-associated protein 43 (GAP43; associated with reparative events) (Figure 5). Injury to small and medium-sized neurons causes, in general, downregulation of excitatory peptides SP and CGRP and upregulation of inhibitory neuropeptide tyrosine (NPY) and GAL.Such changes result in decreased neurotransmission in the spinal dorsal horn and are thought to represent adaptive responses to protect the individual and to promote survival of the sensory neuron. Moreover, GAP43 increases along with upregulation of c-*jun*, both of which are thought to have roles in reparative processes.

Levels of neuropeptides are influenced by multiple regulatory factors, including E_2, electrical activity, substances derived from the environment of the terminals, and inflammation. After tissue injury, NGF is produced by inflammatory cells and is critical for maintenance of NGF-responsive peripheral nerves.

NGF binds to TrkA and $p75^{NTR}$ receptors, and TrkA mediates NGF internalization, retrograde transport, and subsequent cellular responses. E_2 can regulate NGF influence on DRG neurons via regulation of NGF receptors (NGF receptors are increased by 17-β-estradiol). As indicated previously, many DRG neurons contain estrogen receptors and express SP, CGRP, SN, and PACAP; rising serum E_2 levels in late pregnancy promotes expression of these neuropeptides. Thus, a linkage may exist between injury, inflammation, immune cells, NGF, rising E_2, and increase of neuropeptides in DRG neurons. Consequently, E_2 positively influences the plasticity of small C-type neurons.

Diabetes Mellitus

Patients with diabetes mellitus (DM) frequently suffer polyneuropathy with nerve conduction abnormalities. Decreases in sensory amplitudes and slowing of the nerve conduction velocities indicate axonal destruction and demyelination. Factors important in the pathogenesis of diabetic neuropathy are hyperglycemia, reduced Na^+/K^+-ATPase activity in nerves, and microvascular abnormalities.

In DM, peripheral nerves suffer deficient neurotrophic support – a situation which resembles axotomy. In animal models of diabetes, NGF is depleted in nerves; along this line, experimentally induced NGF deprivation results in decreased sensation and hypoalgesia. Early in DM, depletion of NGF leads to dysfunction of small-diameter sensory fibers and loss of SP. There may be a correlation between NGF depletion and decreased organ (particularly in skin) axon reflex vasodilation (promoted by the efferent function small sensory fibers releasing SP). Decreases in endogenous target-derived NGF may influence the onset of diabetic polyneuropathy, whereas metabolic or vascular abnormalities may underlie a full neuropathy. Loss of nociception and axon reflex vasodilation contributes to diabetic foot ulceration in patients. Exogenous NGF can reverse some pathological changes in diabetic nerves; thus, early and prolonged NGF treatment may provide a rational approach to maintenance of sensory nerve function in diabetes.

Other NTFs may be affected in DM. For example, NT-3 is a neurotrophic factor for large-diameter sensory nerves and helps maintain proprioceptive and mechanoreceptive fibers. Neuropathy involving large sensory fibers is frequent in DM. In DM animal models, decreased NT-3 gene expression occurs in hind limb muscle, and treatment with insulin prevents the decrease. Decreases in axonal transport of NTFs are also reported. Moreover, expression of the NT-3 receptor, TrkC, is also downregulated in lumbar DRG. Decreases in NT-3 production and axonal transport may indicate that large sensory neurons expressing TrkC may experience suboptimal neurotrophic support in DM.

It is interesting that expression of genes that are responsive to axotomy (i.e., increase after axotomy) – genes that code for GAP43, GAL, NPY, VIP, neuronal nitric oxide synthase (nNOS), TrkA, TrkB, and TrkC– seem to be unaffected in ganglia from diabetic rats. This suggests that damages caused by DM do not necessarily replicate axonal damage by trauma.

Viral Invasion

Sensory nerves are avenues for neurotropic viruses to enter the nervous system and cause neuropathy, skin lesions, and pain. Following infection, neurotropic viruses replicate in the epithelium, invade sensory nerve endings, and are transported retrogradely to the somas in sensory ganglia (Figure 19). Viruses become latent in the somas, with the possibility of later reactivation, often triggered by a stressful stimulus. Viruses that commonly invade sensory nerves are herpes simplex virus (HSV; types 1 and 2) and herpes zoster (varicella), which produce common clinical syndromes (e.g., chicken pox, shingles, and blisters). Viral presence in an epithelium (e.g., oral cavity or genital region) provides opportunity for invasion of mucous membranes and uptake by sensory nerve endings. A short period of viral replication within sensory ganglia is concomitant with establishment of a nonproductive latent infection. Neurotropic viruses may reemerge from the nervous system, spread to cutaneous areas by anterograde transport of the virus through sensory nerves, and cause disease within the same dermatome at a location distal from the initial site of infection. Spread of the virus requires an intact nerve supply. Both larger A-δ and smaller C-fibers take up and transport neurotropic viruses, thus pain and proprioceptive pathways can be infected.

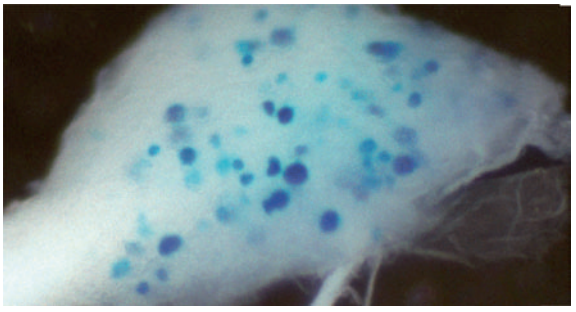

Figure 19 Sensory neurons can take up viruses in their cutaneous terminals and transport them to the cell soma in ganglia. This micrograph shows a whole dorsal root ganglion, illustrating herpes virus-mediated gene transfer of β-galactosidase (shown as blue cells). Reproduced courtesy of Dr. Steven P. Wilson, Department of Pharmacology, Physiology and Neurobiology, University of South Carolina School of Medicine.

Sympathetically Maintained Pain

There are numerous pain conditions involving sensory nerves directly or indirectly, and these have been reviewed extensively. One condition involves sensory and autonomic nerve interaction.

In humans, trauma to a peripheral nerve may be followed by chronic pain which is only relieved by blockade of sympathetic nerve activity. This has been referred to as sympathetic maintained pain (SMP). SMP is a symptom occurring in neuropathic pain syndromes of different etiologies. However, there is controversy regarding use of the term SMP in clinical practice. A movement is under way to use the phrase 'complex regional pain syndrome' (CRPS) (older terms for CRPS are 'reflex sympathetic dystrophy syndrome' and 'causalgia') to describe several types of chronic pain conditions believed to result from dysfunction in the central or peripheral nervous system. Typical features of CRPS include changes in color and temperature of the skin over affected body parts, intense burning, skin sensitivity, sweating, and swelling. Since some of the symptoms could involve sympathetic activity, abnormal activity in the sympathetic nerves may be involved in the pathogenesis of such chronic pain syndromes. For example, fibromyalgia (FM) may be a type of SMP, because patients with FM display signs of relentless sympathetic hyperactivity, and pain is modulated by sympathetic blockade and is exacerbated by noradrenalin (norepinephrine) injections. Along this line, a recent theory suggests that pain receptors in the affected parts of the body become responsive to noradrenalin released from sympathetic nerves after tissue or nerve injury, and animal experiments have shown that nociceptive afferents, after partial nerve lesions, develop adrenergic sensitivity at the site of the injury.

A sympathetic–afferent coupling reportedly takes place in the DRG, but its exact nature has been enigmatic. However, animal experiments reveal that peripheral nerve injury often induces sympathetic nerve fiber sprouting into nearby DRG. It is presumed that, after the lesion, noradrenalin released from sympathetic axons excites primary afferent neurons by activating α-adrenoceptors and generating signals that enter the pain pathways of the CNS. Such connections provide a possible explanation for sympathetic–afferent coupling and SMP.

See also: Diabetic Neuropathy; Enteric Nervous System: Sensory Pathways; Neural Crest; Neuropathic Pain; Neuropeptides: Sensory Systems; Neurotrophins: Physiology and Pharmacology; Pain and the Sympathetic Nervous System.

Further Reading

Gibbins IL, Furness JB, and Costa M (1987) Pathway-specific patterns of the co-existence of substance P, calcitonin gene-related peptide, cholecystokinin and dynorphin in neurons of the dorsal root ganglia of the guinea-pig. *Cell and Tissue Research* 248: 417–437.

Guo A, Vulchanova L, Wang J, et al. (1999) Immunocytochemical localization of the vanilloid receptor 1 (VR1): Relationship to neuropeptides, the $P2X_3$ purinoceptor and IB4 binding sites. *European Journal of Neuroscience* 11: 946–958.

Hökfelt T, Broberger C, Xu ZQ, et al. (2000) Neuropeptides – an overview. *Neuropharmacology* 39: 1337–1356.

Hökfelt T, Zhang X, and Wiesenfeld-Hallin Z (1994) Messenger plasticity in primary sensory neurons following axotomy and its functional implications. *Trends in Neuroscience* 17: 22–30.

Julius D and Basbaum AI (2001) Molecular mechanisms of nociception. *Nature* 413: 203–210.

Kirkup AJ, Brunsden AM, and Grundy D (2001) Receptors and transmission in the brain–gut axis: Potential for novel therapies. I. Receptors on visceral afferents. *American Journal of Physiology – Gastrointestinal and Liver Physiology* 280: G787–G794.

Kunze WA and Furness JB (1999) The enteric nervous system and regulation of intestinal motility. *Annual Review of Physiology* 61: 117–142.

Lauria G, Morbin M, Lombardi R, et al. (2006) Expression of capsaicin receptor immunoreactivity in human peripheral nervous system and in painful neuropathies. *Journal of the Peripheral Nervous System* 11: 262–271.

Maggi CA (1991) The pharmacology of the efferent function of sensory nerves. *Journal of Autonomic Pharmacology* 11: 173–208.

McLachlan EM, Jänig W, Devor M, et al. (1993) Peripheral nerve injury triggers noradrenergic sprouting within dorsal root ganglia. *Nature* 363: 543–546.

Mendell L (1996) Neurotrophins and sensory neurons: Role in development, maintenance and injury. A thematic summary. *Philosophical Transactions of the Royal Society of London B* 351: 463–467.

Mowa CN and Papka RE (2004) The role of sensory neurons in cervical ripening: Effects of estrogen and neuropeptides. *Journal of Histochemistry and Cytochemistry* 52: 1249–1258.

Papka RE, Storey-Workley M, Shughrue PJ, et al. (2001) Estrogen receptor-alpha and beta-immunoreactivity and mRNA in neurons of sensory and autonomic ganglia and spinal cord. *Cell and Tissue Research* 304: 193–214.

Petruska JC, Napaporn J, Johnson RD, et al. (2002) Chemical responsiveness and histochemical phenotype of electrophysiologically classified cells of the adult rat dorsal root ganglion. *Neuroscience* 115: 15–30.

Zochodne DW, Verge VM, Cheng C, et al. (2001) Does diabetes target ganglion neurones? Progressive sensory neurone involvement in long-term experimental diabetes. *Brain* 124: 2319–2334.

Sensory Re-education

G Lundborg and B Rosén, Malmö University Hospital, Malmö, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Peripheral nerve injuries still represent some of the most challenging and difficult surgical reconstructive problems. Although evolving scientific concepts during the past decades have resulted in an enormous amount of new laboratory data, this has had little impact on the treatment of such injuries. The majority of nerve injuries occurs in the upper extremity of young males, varying in severity and level from clear-cut nerve transection to severe nerve lacerations and from digital nerve injuries to brachial plexus lesions. A simple transection of a median nerve at wrist level may have disastrous consequences for hand and upper extremity function. The probability of work loss is very high and there is a substantial economic impact on the patient as well as society. Nerve injuries may substantially impair the life quality for patients and may result in lifelong hand-function impairment, pain, cold intolerance, and dysesthesia.

Following nerve repair at distal forearm level the fingers will be without sensibility up to 6 months before any reinnervation at distal finger level occurs, and 5 years following surgery there still is a substantial disability in adults.

Nerve injuries are usually surgically repaired so that both ends of the cut nerve are anastomosed by use of microsurgical techniques. However, in spite of optimal technical refinement of the repair techniques, the outcome in adult patients is often disappointing with remaining deficits in terms of impaired sensory and motor function and discomfort based on pain and cold intolerance. A crucial factor is the great extent of axonal misdirection which always takes place after nerve repair regardless of how sophisticated technique is used. A result, and a major reason for bad functional outcome, is the profound cortical functional reorganizations, which takes place as a result of the peripheral axonal misdirection. The problem is even greater when a nerve injury is treated at a later stage when a reconstruction with a nerve graft may be required. Several factors influence the outcome from nerve repair such as timing of the repair, type of nerve, level of injury, and type of injury. In adult patients it has been demonstrated that specific cognitive capacities such as verbal learning capacity and visuospatial logic capacity help to explain variations in the recovery of functional sensibility. However, among all influencing factors, the age of the patient has turned out to be the most important. Although functional restitution after nerve repair in adults is usually far from satisfactory, the situation is quite different in children, who consistently show superior functional results as compared to adults.

The Cortical Response to Nerve Injury and Repair

It was demonstrated 70 years ago by Penfield and Boldrey that the hand is projected in a large area in the cortical body map. A nerve transection represents an acute deafferentiation with immediate and long-standing influence on the corresponding representational areas in brain cortex as well as in adjacent cortical territories. In early primate experiments it was shown by the Merzenich group that transection of a median nerve at wrist level results in a silent area in the somatosensory brain cortex corresponding to the projectional areas of the fingers that had been deprived of their central connections. The result was an expansion of adjacent cortical areas that gradually occupied a substantial area of the former median nerve territory. Such changes probably occur very fast after nerve transection, within minutes, as a result of unmasking of normally occurring but inhibited synaptic connections.

Later, when reinnervation of the hand occurs there is a new phase of cortical reorganization. Regardless of how refined surgical technique is used, there is a great deal of axonal misorientation at the repair site. Schwann cell tubes in the distal nerve segment may be reinnervated by sprouts from a large number of various axons, and axons in the proximal segment may hereby reinnervate incorrect distal pathways. As a result, distal skin areas will not, to a large extent, be reinnervated by the original axons but may instead be reinnervated by axons originally meant to innervate other parts of the hand. It is known from primate experiments that the original well-organized cortical representation hereby is transformed into a mosaic like pattern and that the nerve does not recapture all of its original territory. The former well-defined individual cortical representation of separate fingers disappears and changes into disbursed discontinuous islands. Skin surfaces, which previously had well-defined cortical representations such as a finger, a segment of a finger, or a palmar pad, become instead represented across multiple small, sometimes overlapping cortical patchlike areas. Cortical regions that receive tactile signals from reinnervated skin regions will now contain multiple recording sites

with abnormally located or multiple cutaneous receptive fields. In monkeys, such representational changes after nerve transection and repair occur both immediately and progressively at the cortical as well as the subcortical level. In humans, fuctional magnetic resonance imaging (fMRI) investigation has shown that tactile stimulation of the nerve-injured hand activates larger areas in the somatosensory cortex compared with the noninjured hand. This enlarged area has been regarded as a result of a cortical reorganization, and it is believed that larger cortical areas are involved in the discriminatory tasks after the supposed derangement of the peripheral input. From a clinical perspective, the result is an incomplete sensory restitution of the hand with impaired discriminative capacity – the hand speaks a new language to the brain. The capacity to detect touch may be restored so that an adult person can regain some protective sensibility. However, the prerequisites for a discriminatory capacity to identify textures, structures, and shapes in the act of touch are very limited and a sensory relearning process is required.

The Sense of Touch

Our senses are the windows through which we perceive our environment. Hearing, vision, taste, smell, and the sense of touch together provide our mind with an inner picture of the outer world. For the sense of touch, the hand is a key organ with huge cortical representation, and the sensation has therefore a vital role in hand function. Lost or impaired hand sensation after nerve injury is a major problem, causing not only considerable body function impairment but also substantial disability and reduction in life quality.

Four major modalities of somatic sensibility can initially be defined: the pure and discriminative touch, proprioception, nociception, and temperature sense. For clinical examination of hand sensibility, a hierarchy of sensory functions can be identified: if detection of touch is present, the next level is to examine if the patient can discriminate the touch – the most basic tactile gnosis. Two-point discrimination test is the classic test here – although the validity and reliability is disputed. Localization of touch is as well an aspect of discriminative touch. Identification by active touching is the third level – a more refined tactile gnosis. Functional sensibility – or what some name haptics – is the endpoint where tactile gnosis acts in concert with muscles and joints for a useful grip function.

In the interplay between the senses, the eye remains at the surface, while hearing and the touch of the hand obtain information on the inner of an object, the microstructure, thereby playing a great role in developing a belief in the reality of the external world. More than a 100 years ago, von Frey and Weber paid attention to clinical methods to describe the capacity to detect touch and spatial acuity, respectively. Spatial acuity is the smallest discriminated distance between two points of contact. Moberg paid further attention to the importance of tactile discrimination, and established the term 'tactile gnosis.' Tactile gnosis enables recognition of qualities and characters of objects without using vision, which is of utmost importance for functional sensibility – a prerequisite for what the hand can do. Napier described the hand as "an organ of touch which feels around corners and sees in the dark." The sense of touch can provide us with information not only about shape and texture but also about temperature, consistency, elasticity, density, dryness, wetness, stickiness, and oiliness. The discriminative touch, the basis for tactile gnosis, is based on various types of sensory receptors found in the glabrous and hairy skin which are sensitive to mechanical stimulation, transducing mechanical stimuli such as pressure, vibration, and stretching. Afferent signals from cutaneous mechanoreceptors of the hand reach the somatosensory brain cortex after passing the dorsal root ganglia, the dorsal column of the spinal cord via the medial lemniscuses pathway, and intermediate relay stations situated in the cuneate nucleus in the brain stem and the ventroposterior nucleus in thalamus. Signals elicited by touch primarily reach the contralateral hemisphere, but to a lesser extent also the ipsilateral somatosensory cortex. The somatosensory cortex has a complex multidimensional functional architecture with three major divisions: the primary (S-I) and secondary (S-II) somatosensory cortices and the posterior parietal cortex. In the primary sensory cortex there are four distinct somatotopic maps corresponding to the four areas of S-I that have been identified by Brodman named 1, 2, 3a, and 3b; each map represents one subquality of sensation. Information from the skin, important for touch, is primarily represented in area 3b.

Sensory Relearning and Sensory Re-Education

In the adult, the original neural program for understanding tactile information, which was generated in early childhood, is lost after nerve injury and repair as a result of cortical reorganization and distortion of the original cortical hand map. A relearning process is required. The capacity to recover tactile gnosis after nerve repair, as related to age, has much in common with learning to speak a second language

(Figures 1 and 2). In hand rehabilitation following nerve injury and repair, such sensory relearning is based on the use of sensory re-education protocols (Figure 3). Normally, such a relearning process may in parts be gained in daily activities by conscious use of the hand in common manual tasks, using vision as a guide and a code for impaired hand sensibility. The concept is to reprogram the brain to understand the new 'language' from the hand. Specific programs for sensory re-education have been suggested and are routinely used clinically in adult patients, subjected to nerve repair in the hand forearm, for regaining tactile gnosis. Protocols for sensory re-education have been presented in numerous books and articles. First, the perception of different touch modalities and the capacity to localize touch is trained, followed by touching and exploration of items, presenting shapes and textures of varying and increasing difficulty with the eyes open or closed. In this way, an alternate sense (vision) trains and improves the deficient sense (sensation).

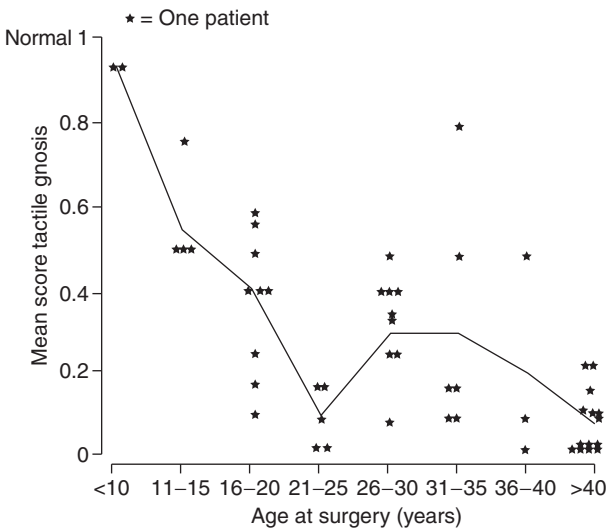

Figure 1 Mean recovery of tactile gnosis in 54 patients, as related to age at surgery, assessed at least 2 years after injury and repair of major nerves of the forearm. Reproduced from Lundborg G and Rosen B (2001) Sensory relearning after nerve repair. *Lancet* 358: 809–810.

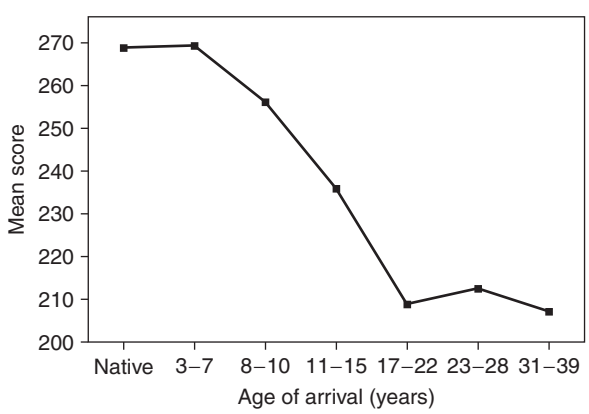

Figure 2 Curve indicating how the scores of immigrants on a grammar test decline with age at which they started to learn to speak the English language. From Newport EL (1991) Contrasting conceptions of the critical period for language. In: Carey S and Gelman R (eds.) *The Epigenesis of Mind: Essays on Biology and Cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates. Redrawn from Johnson JS and Newport EL (1989) Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21: 60–99.

The training is integrated in the overall rehabilitation program. Specific simple and repetitive sensory relearning exercises of increasing complexity and difficulty are supposed to be performed at home by the patient on a regular basis in frequent and short training sessions. Training in a quiet location for high attention is recommended.

Several factors may influence the results from the sensory re-educational programs in a positive or negative way. For instance, active and conscious use of the hand in daily activities and attention during training combined with high motivation by the patient are extremely important. From primate experiments, it is obvious that a behavioral state and the strength of reward and punishment in behavioral training have strong influence on the remodeling of cortical representations. When animals are stimulated on a routine stimulus schedule with the stimulation unattended cortical remodeling does not occur, and nonassociated meaningless inputs can drive negative representational changes. It is important that individuals are well rewarded in the training program since successful performance generates the most rapid representational changes. The molding of the brain is influenced in a positive direction in a stimulating environment with meaningful activities and much encouragement.

How Is Brain Organization Influenced by Sensory Re-Education?

Little is known about the effects of sensory re-education on brain organization. One possibility is that sensory training can reverse the disturbed cortical organization toward a normalized pattern. Another possibility is that the cortical hand map is permanently distorted and that the key is an improved processing and perceptual capacity in the sensory networks at a higher order of cortical level, facilitating interpretation of the distorted hand map. Although young individuals may have a greater capacity for normalization of the cortical hand map

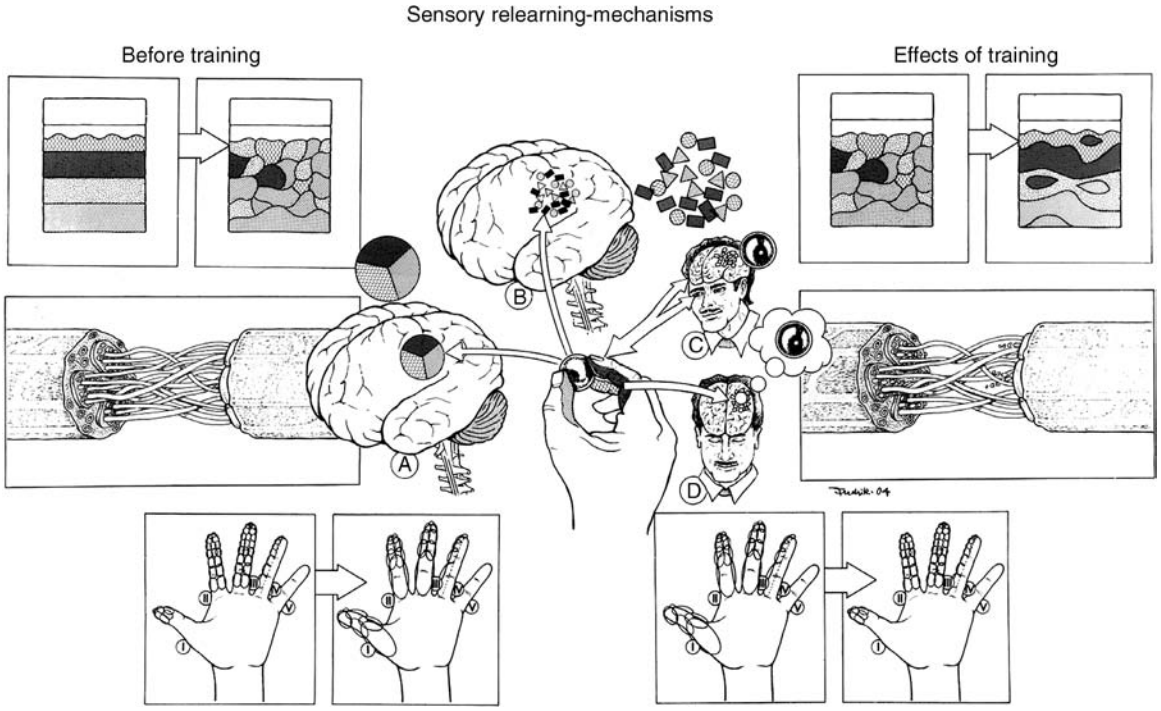

Figure 3 Basic principles for sensory re-education and underlying physiologic mechanisms. By touching items of increasing difficulty with the eyes open or closed, a relearning process is initiated in which vision is the guiding sensation to understand 'the new language spoken by the hand.' The pathophysiologic events behind the altered sensibility in the hand, induced by the injury, are illustrated to the left ('before training'). The cortical hand map is changed (top) as a result of misdirected axons at the suture site (middle). The cortical receptive fields in the fingertips are becoming large and overlapping as reflected in an impaired tactile discrimination capacity (bottom). The 'sensory relearning process' passes through several steps. The cortical representation of a ball, schematically illustrated by (A), has been changed by the nerve injury into a dispersed, nonorganized cortical pattern (B). By touching items of increasing difficulty, with alternately closed and open eyes (C, D), vision guides sensation toward a better understanding of shapes and textures. Possible physiologic changes explaining effects of sensory re-education are illustrated to the right ('effects of training'). The distorted cortical hand map is modified toward a more normal pattern (top). The misdirection of axons at the suture site is not significantly changed, although some pruning of misdirected axons has occurred (middle). The receptive fields in fingertips shift toward a normalized pattern showing a finer grain in resolution (bottom). Reproduced from Lundborg G (2004) *Nerve Injury and Repair. Regeneration, Reconstruction and Cortical Re-Modelling*, 2nd edn. Philadelphia: Elsevier, with permission from Elsevier.

following nerve section and repair compared to adults, such capacity for normalization of the cortical hand map by training probably does not exist in the mature brain. The topic was studied by Florence, who investigated the effects of enriched environmental sensory experience on cortical organization on monkeys that had previously undergone nerve section and repair of the median nerve. One group was subjected to an enriched environmental sensory experience involving presentation of food treats on an artificial grass surface and toys, while the control group was allowed only restricted mobility with no such enriched sensory experience. Using direct cortical recording techniques, it was demonstrated that the enriched sensory environment had a significant effect on receptor field size with smaller, normalized, and better-located receptive fields likely to provide better resolution. In addition, the cortical hand maps in area 3b were less distorted in the sensory-enriched animals compared with the sensory-restricted animals. Thus, also in humans, sensory relearning programs, including an enriched environment, may hypothetically explain functional improvement after rehabilitation.

Evolving Strategies for Enhanced Sensory Relearning

The sensory re-education programs that are used today have not changed much during the last decades, in spite of the amount of increased knowledge and evolving concepts in neuroscience, cognitive science, and brain imaging techniques that have been presented over recent years together with our increased understanding of mechanisms underlying brain plasticity with a focus on learning. For instance, the timing of the sensory relearning has been little considered. Routine sensory re-education starts with beginning reinnervation of the hand that is 3–5 months after a median nerve lesion at wrist level. At this point, the

reorganizing cortex presents a random pattern which can be very difficult for the patient to control and utilize for regaining functional sensibility. It may be that the timing of onset of sensory relearning is a crucial factor, since functional reorganization changes of the cortex occur early after changes in peripheral input. For instance, following amputation of the hand, touching the face – the face being close to the hand in the cortical body map – gives rise to referred sensations in the phantom as early as 24 h after the injury. Analogous phenomena have been demonstrated, also after local anesthetic blocks, that temporarily induce shifts in neuronal receptive fields with cortical reorganization.

When discussing onset and principles for sensory re-education in the future, two postoperative time windows should be defined: phase one, covering the early postoperative phase when the hand is still without sensation, and when the original cortical projectional area of the denervated body part is being occupied by expanding adjacent cortical areas, and phase two, when reinnervation of the hand is beginning and when the original cortical projection is changed into a disbursed, mosaiclike pattern.

Phase One

In phase one, maintenance of the cortical hand projection, which otherwise will be occupied by adjacent cortical areas, may be a way to facilitate the course of sensory recovery when the hand is reinnervated. In uninjured individuals premotor areas, can be activated by solely using motor imagery exercises (imaging hand movements), observing movements of hands belonging to others by activating mirror neurons in frontal cortex, or reading or listening to action words, related to hand movements. In tactile imagery, the somatosensory cortex is activated by mental imagery and tactile stimulation of the hand. Alternative ways to activate somatosensory cortex are less documented, but somatosensory neural substrates have been activated by watching touch of legs or hands. Another principle is to use another sense – hearing – to substitute for sensibility. This method allows early sensory relearning, long before any reinnervation of the hand can be identified, in order to feed the sensory cortex with 'relevant' information. The ambition here is to maintain the cortical map from the affected hand until 'real' sensibility is present. Cortical audiotactile interaction has recently, with fMRI technique, been demonstrated in persons trained with such a system.

Phase Two

In phase two, when reinnervation of the hand is beginning and when the original cortical projection is changed, sensory re-education programs focus on detection and discrimination of touch in combination with identification of specific objects. A very conscious perceptive behavior when using the hand in daily activities is also of importance. New training strategies have been suggested in the literature recently. Examples of such strategies are cutaneous anesthesia of the forearm of the nerve-injured hand and continuous tactile stimulation to enhance the sensory relearning. Such strategies, based on new knowledge in brain plasticity and mechanisms in the brain's capacity for rapid redistribution of cortical resources, may have potential for improving sensory functions in the hand after nerve injuries.

See also: Adult Cortical Plasticity; Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Injury: Neuronal Responses; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Neuronal Plasticity after Cortical Damage; Nogo-A: Its Role in Axon Regeneration; Peripheral Nerve Regeneration: An Overview; Spinal Cord Regeneration and Functional Recovery: Strategies.

Further Reading

Chen R, Cohen LG, and Hallett M (2002) Nervous system reorganization following injury. *Neuroscience* 111: 761–773.

Dellon AL (1981) *Sensibility and Re-Education of Sensation in the Hand*. Baltimore: Williams & Wilkins.

Florence S, Jain N, Pospichal MW, Beck PD, Sly DL, and Kaas JH (1996) Central reorganization of sensory pathways following peripheral nerve regeneration in fetal monkeys. *Nature* 381: 69.

Florence SL, Boydston LA, Hackett TA, Lachoff HT, Strata F, and Niblock MM (2001) Sensory enrichment after peripheral nerve injury restores cortical, not thalamic, receptive field organization. *European Journal of Neuroscience* 13: 1755–1766.

Hauk O, Johnsrude I, and Pulvermuller F (2004) Somatotopic representation of action words in human motor and premotor cortex. *Neuron* 41: 301–307.

Jeannerod M and Frak V (1999) Mental imaging of motor activity in humans. *Current Opinion in Neurobiology* 9: 735–739.

Johansson BB (2000) Brain plasticity and stroke rehabilitation. The Willis lecture. *Stroke* 31: 223–230.

Johansson RS and Vallbo AB (1979) Tactile sensibility in the human hand: Relative and absolute densities of four types of mechanoreceptive units in glabrous skin. *Journal of Physiology* 286: 283–300.

Johnson JS and Newport EL (1989) Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21: 60–99.

Kaas JH and Florence SL (1997) Mechanisms of reorganization in sensory systems of primates after peripheral nerve injury. *Advances in Neurology* 73: 147–158.

Kandel ER, Schwartz JH, and Jessel TM (eds.) (2000) *Principles of Neural Science*, 4th edn. New York, NY: McGraw-Hill.

Katz D (1989) *The World of Touch*. London: Lawrence Erlbaum Associates.

Keysers C, Wicker B, Gazzola V, Anton J-L, Fogassi L, and Gyallese V (2004) A touching sight: SII/PV activation during the observation and experience of touch. *Neuron* 42: 335–346.

Lundborg G (2000) A 25-year perspective of peripheral nerve surgery: Evolving neuroscientific concepts and clinical significance. *Journal of Hand Surgery* 25A: 391–414.

Lundborg G (2003) Richard P Bunge memorial lecture. Nerve injury and repair – A challenge to the plastic brain. *Journal of the Peripheral Nervous System* 8: 209–226.

Lundborg G (2004) *Nerve Injury and Repair. Regeneration, Reconstruction and Cortical Re-Modelling,* 2nd edn. Philadelphia: Elsevier.

Lundborg G and Rosen B (2001) Sensory relearning after nerve repair. *Lancet* 358: 809–810.

Merzenich MM and Jenkins WM (1993) Reorganization of cortical representations of the hand following alterations of skin inputs induced by nerve injury, skin island transfers, and experience. *Journal of Hand Therapy* 6: 89–104.

Merzenich MM, Kaas JH, Wall J, Nelson RJ, Sur M, and Felleman D (1983) Topographic reorganization of somatosensory cortical areas 3B and 1 in adult monkeys following restricted deafferentiation. *Neuroscience* 8: 33–55.

Newport EL (1991) Contrasting conceptions of the critical period for language. In: Carey S and Gelman R (eds.) *The Epigenesis of Mind: Essays on Biology and Cognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rizzolatti G and Craighero L (2004) The mirror–neuron system. *Annual Review of Neuroscience* 27: 169–192.

Rosén B, Balkenius C, and Lundborg G (2003) Sensory re-education today and tomorrow. Review of evolving concepts. *British Journal of Hand Therapy* 8: 48–56.

Rosén B, Lundborg G, Dahlin LB, Holmberg J, and Karlsson B (1994) Nerve repair: Correlation of restitution of functional sensibility with specific cognitive capacities. *Journal of Hand Surgery* 19B: 452–458.

Rossini PM, Martino G, Narici L, et al. (1994) Short-term brain plasticity in humans: Transient finger representation changes in sensory cortex somatotopy following ischemic anesthesia. *Brain Research* 642: 169–177.

Vallbo ÅB and Johansson RS (1984) Properties of cutaneous mechanoreceptors in the human hand related to touch sensation. *Human Neurobiology* 3: 3–14.

Wall JT, Kaas JH, Sur M, Nelson RJ, Felleman DJ, and Merzenich MM (1986) Functional reorganization in somatosensory cortical areas 3b and 1 of adult monkeys after median nerve repair: Possible relationships to sensory recovery in humans. *Journal of Neuroscience* 6: 218–233.

Wall JT, Xu J, and Wang X (2002) Human brain plasticity: An emerging view of the multiple substrates and mechanisms that cause cortical changes and related sensory dysfunctions after injuries of sensory inputs from the body. *Brain Research Reviews* 39: 181–215.

Witzel C, Rohde C, and Brushart TM (2005) Pathway sampling by regenerating peripheral axons. *Journal of Comparative Neurology* 485: 183–190.

Yoo SS, Freeman DK, McCarthy JJ, and Jolesz FA III (2003) Neural substrates of tactile imagery: A functional MRI study. *Neuroreport* 14: 581–585.

Sensory System Specializations

E C Henry and K C Catania, Vanderbilt University, Nashville, TN, USA

© 2009 Published by Elsevier Ltd.

Introduction

Studies of specialized species often reveal general principles of brain organization and function and also help guide theories of brain evolution by revealing the neuronal traits that are correlated with specific sensory abilities. Although there is great diversity in the organization of visual, auditory, and olfactory systems across mammals, in some ways specializations of the somatosensory system are most easily appreciated because this system is so closely tied to the morphology of the body. Every different mammal has a differently organized somatosensory system reflecting the unique configurations of the skin receptors, joints, and muscles. Species with prominent mechanoreceptors usually have predictably large corresponding areas of the central nervous system for processing touch information.

Mice and rats, with their large facial whiskers, are well-known examples. In 1970, Woolsey and Van der Loos made a landmark discovery in the mammalian neocortex by identifying a series of cortical modules (barrels) in somatosensory cortex (S1) that correspond to the pattern of whiskers on the contralateral side of the face in mice (Figure 1). This visible reflection of peripheral mechanoreceptors in the central cortical representation, or map, was best revealed by cutting the cortical tissue tangent to the surface so that large areas of the map in cortical layer 4 could be viewed in a single section of tissue.

Refinements to the flattened cortical preparation and new histological techniques, in conjunction with electrophysiological recordings of cortical neurons, have provided increasingly detailed understanding of brain organization in a range of different species. In the case of somatosensation, it is particularly informative to examine species that depend most heavily on touch to navigate their environments. Subterranean mammals must 'feel' their way through their underground burrows, using tactile cues for orientation and prey identification, and they have correspondingly enlarged and specialized somatosensory systems. Two different subterranean species, the star-nosed mole and the naked mole rat, are specialized for life underground, but their somatosensory systems are very different.

Star-Nosed Moles (*Condylura cristata*)

Mechanoreceptor Specializations

There are about 30 different species of moles (family Talpidae) that exhibit a variety of different specializations for touch. However, a common theme among all these species is the use of the tip of the nose, rather than long whiskers, to collect the most detailed tactile information from their environment. For this purpose, nearly every mole species has a series of domed mechanosensory organs covering the distal surface of the snout. These are called Eimer's organs (Figure 2), and each contains a series of free nerve endings that run through an epidermal cell in the center of the organ. At the base of the organ, one or more merkel cells is contacted by a neurite, and just below this, a lamellated corpuscle is situated in the underlying dermis, just below the central cell column. In addition, recent investigations have identified a second, smaller set of free nerve endings that ring the periphery of the dome and are positive for substance P – a marker for nociceptive afferents. This subset of nociceptive afferents is well positioned to detect damage to the delicate organs.

An array of these complicated sensory organs generally covers the skin surface of a mole's nose. However, the snout of the star-nosed mole is by far the most elaborate and complex. The entire star is essentially composed of Eimer's organs that cover all the 22 appendages that form the star. Despite its small size (about 1 cm diameter), the star is innervated by more than 100 000 primary afferents that serve 25 000 Eimer's organs. Using this array of mechanosensory organs, the star-nosed mole can make very rapid sensory discrimination of small objects.

The star stands out as one of the more unusual mechanosensory specializations among mammals, and the organization of this touch organ shares similarities with the organization of visual systems in other species. For example, the star is divided into a high-resolution central area (the paired 11th appendages in the center) and a larger, low-resolution area consisting of the other 20 appendages of the star. When moles explore their environment, the star is moved like an eye, with the large surface area of appendages 1–10 used to scan the environment through touch, and the central pair used as the tactile fovea for detailed investigations of objects of interest. The sudden, jerky movements of the star that shift the tactile fovea to different areas of the environment are very similar to saccadic eye movements that position

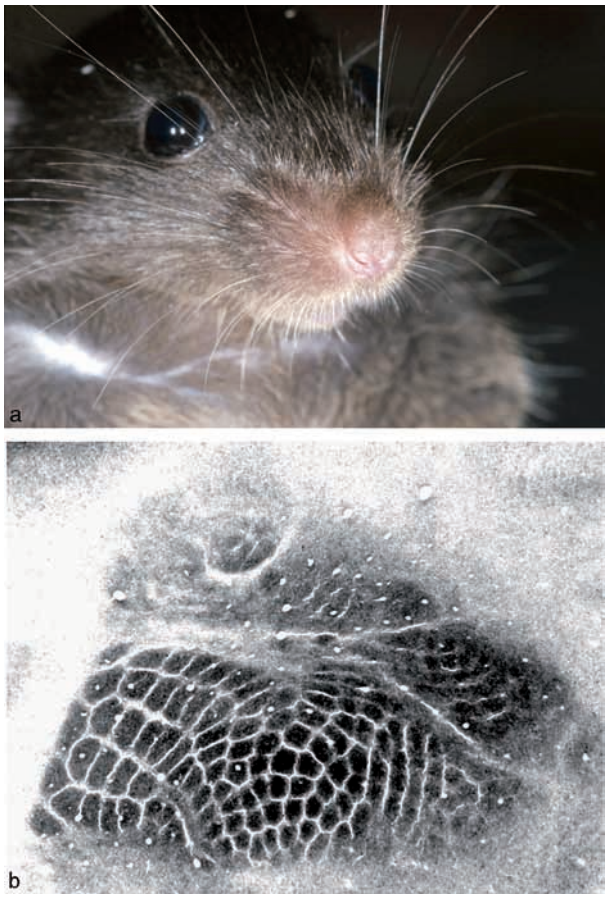

Figure 1 Cortical barrels. (a) The face of a laboratory mouse, including the prominent whiskers or 'vibrissae.' (b) A section of flattened right cortex processed for the metabolic enzyme cytochrome oxidase to reveal the prominent cortical barrels in layer IV. Rostral is to the right and medial is up. The large facial vibrissae correspond to the barrels visible in the lower left quadrant whereas the smaller facial whiskers are represented by the smaller, more rostral barrels.

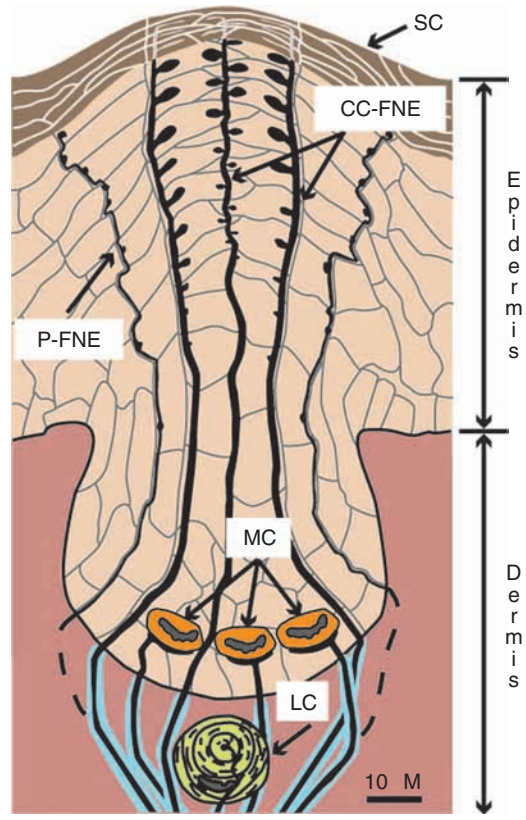

Figure 2 Schematic representations of the structural components of Eimer's organ of the mole. Eimer's organ contains a central column of large diameter intraepidermal free nerve endings (CC-FNE) that course directly to the stratum corneum (SC). A ring of smaller satellite free nerve endings surrounds the central column free nerve endings (S-FNE). The satellite fibers take a circuitous route to the skin surface, tracing the boundaries between epithelial cells one or two cell diameters removed from the central column fibers. There are Merkel cell–neurite complexes (MC) located at the base of the epidermal papilla. There are also lamellated corpuscles in the dermis below the papilla of each Eimer's organ (LC).

images on the retinal fovea in primates. This elaborate sensory system allows star-nosed moles to discriminate and consume small prey in only 200 ms.

Central Nervous System

The complexity of the star and associated behaviors suggest that star-nosed moles should have enlarged and specialized somatosensory areas in the central nervous system. This turns out to be true (Figure 3) and is most easily appreciated at the cortical level by examining flattened cortical sections from the somatosensory region – as previously described in mice and rats. Figure 3(b) shows a single section of layer 4 cortex, where touch information from half the star projects to primary somatosensory cortex (S1). As in many rodents, the distribution of mechanoreceptors on the mole's snout projects to a modular region of the cortex that reflects the morphology of the face. Simply put, the map of the star is visible in the brain.

Closer examination of the relationship between the star and the cortical map reveals a disproportionate expansion of the region representing appendage 11 (only the contralateral half of the star is represented in each cortical hemisphere; hence only 11 stripes are visible on each side of the brain). Although the 11th appendage is relatively small on the star, it has a disproportionately large representation in the brain, reflecting the behavioral importance of this region of the sensory surface. This is also true for the cortical representation of the retinal fovea in primary visual cortex of more-visual species.

Examining a larger region of the cortex reveals additional maps of the star in more lateral cortex (Figure 3(c)). These maps are densely interconnected with one another to form a network that most likely processes different aspects of touch information in each different area. Star-nosed moles have three maps

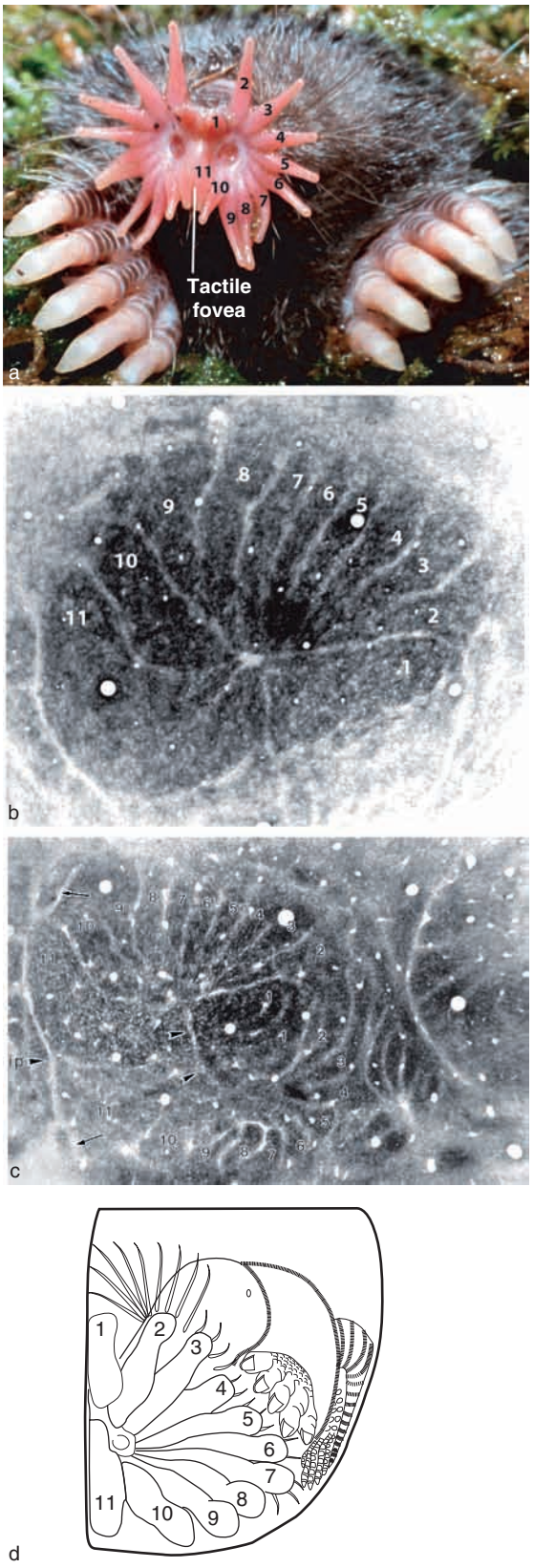

Figure 3 The unusual mechanosensory star, its corresponding cortical representation, and the organization of the superior colliculus in the star-nosed mole (*Condylura cristata*). (a) A star-nosed mole emerging from an underground tunnel shows its large forelimbs and the 22 fleshy appendages that surround each nostril. The 11th appendages on each side act as the somatosensory fovea and are used for detailed tactile investigations. (b) The pattern of cortical modules in S1 visible in flattened cortical sections processed for the metabolic enzyme cytochrome oxidase. Note the greatly expanded representation of the somatosensory fovea (appendage 11). (c) A section of flattened cortex revealing all three cortical representations of the star, visible as a series of modules, each representing an appendage from the contralateral star. (d) Schematized representation of the star-nosed mole's body proportions in the superior colliculus. The star has a greatly magnified representation and takes up approximately half of the somatosensory map.

of the star in cortex, whereas other mole species have only two maps. Star-nosed moles appear to have added a new cortical area to the processing network to handle the large volumes of somatosensory information from the star, and this may be one way they have achieved such rapid discriminatory abilities.

Finally, it is interesting to consider how the superior colliculus is organized in star-nosed moles. The superior colliculus in most mammals is dominated by visual inputs and plays a critical role in guiding orientation movements of the eyes. In star-nosed moles, the superior colliculus receives only minimal input from the eyes. Instead, somatosensory inputs dominate, and much like the organization of the somatosensory cortex, a map of the body surface in the colliculus is dominated by the representation of the star (Figure 3(c)). These findings suggest the superior colliculus in star-nosed moles plays a role in guiding the saccadic orientation movements of the nose.

Naked Mole Rats (*Heterocephalus glaber*)

The naked mole rat is a remarkable rodent with many unusual traits. Mole rats live in large subterranean colonies of 80 or more individuals but only a single reproductively active female. This eusocial colony structure includes separate castes – a queen, many workers, and occasional disperser morphs – thus closely resembling the organization of social insect colonies such as ants and termites. Naked mole rats also have extraordinarily long lives compared with other rodents, with captive animals living more than 20 years.

Mole Rat Mechanoreceptor Specializations

Naked mole rats also have a number of sensory specializations that allow them to navigate dark, underground tunnels. Much like star-nosed moles, naked mole rats have reduced eyes and sparse visual projections to the superior colliculus. They have few retinal ganglion cells compared with sighted rodents, and the architecture of their retina is disorganized such that form vision is impossible.

Not surprisingly, mole rats depend heavily on touch to navigate their environment. Although virtually hairless, they have unique 'body vibrissae' in rows along the trunk that act as touch sensors for navigation within their tunnels. These body vibrissae have follicle diameters and innervation that more closely resemble those of facial vibrissae than fur. These hairs guide orientation movements of the mole rat's head in the direction of novel stimuli. Mole rats also have facial whiskers more typical of other rodents, which they use for tactile perception (Figure 4).

However, the most obvious feature of mole rats' external anatomy is the large incisors that protrude beyond the oral cavity. The mole rat's mandibles connect at their anterior extent by a flexible mandibular symphysis, allowing the lower incisors to move independently. Mole rats use these teeth for digging extensive tunnel networks, carrying objects and young, chewing, and investigating their surroundings (by tapping their incisors against objects of interest). The teeth are also important for social interactions between colony members – 'incisor fencing' is often observed in struggles for dominance. From these observations, one might predict that the front teeth, rather than the snout, have the most important role in gathering sensory information from the environment. Investigation of mole rat cortex suggests this is the case.

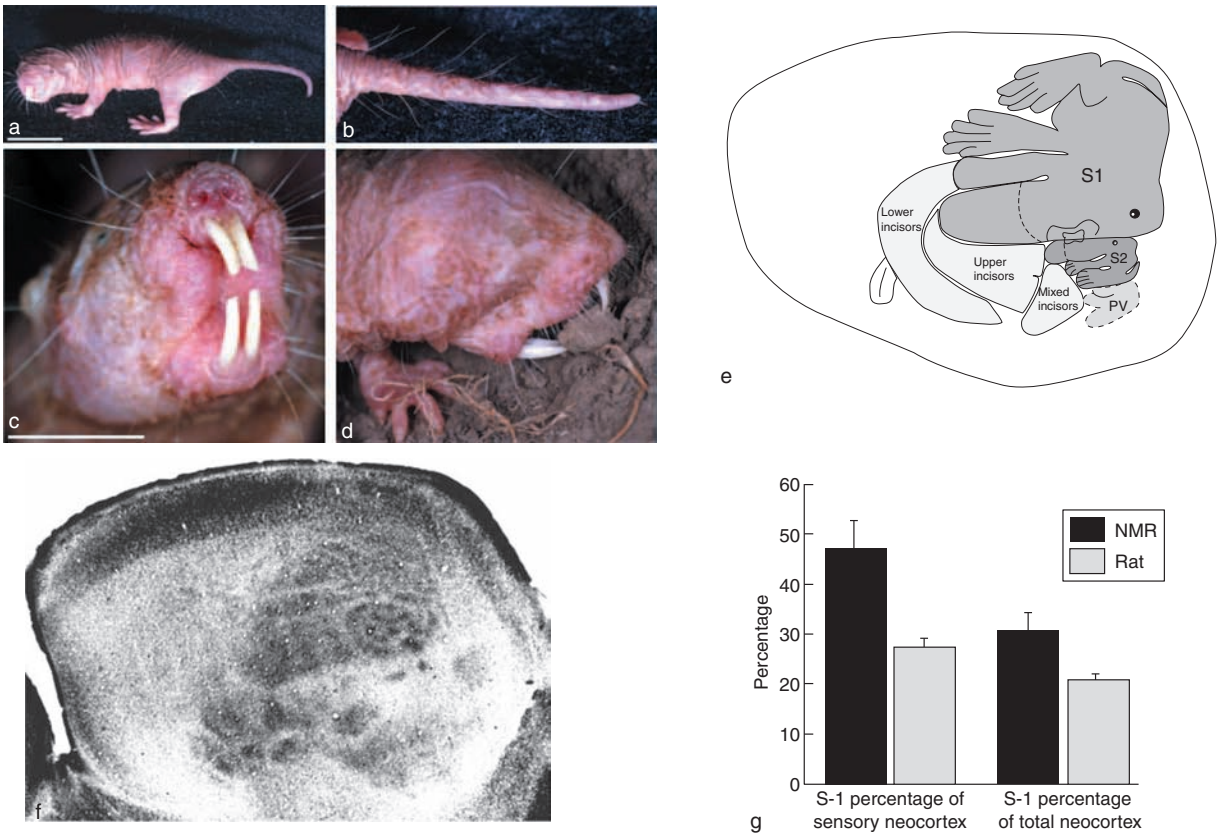

Figure 4 The unusual anatomy of the naked mole rat (*Heterocephalus glaber*). (a) Photo of an adult mole rat, demonstrating the lack of fur and general proportions of the body. (b) The sensory hairs covering the tail are used for navigation while the mole rat is traveling backward through tunnels. (c) Front view of a naked mole rat, showing the enlarged incisors that are located permanently exterior to the oral cavity. Note also the sensory whiskers covering the front part of the face and mouth region. (d) A mole rat moves a small piece of earth with its teeth. (e) Summary schematic of the somatotopic organization, location, and relative size of areas S1, S2, and a region presumed to be the parietal ventral area (PV; dotted line). The large representations of the upper and lower incisors and the rostral tongue area are all part of S1. The 'mixed Incisors' region is composed of neurons responsive to the upper incisor, lower incisor, or both. (f) A single section of cortical tissue corresponding to layer IV in the naked mole rat. The section was flattened and cut parallel to the surface before processing for the enzymatic marker for oxidative metabolism, cytochrome oxidase (Wong-Riley et al., 1978; Wong-Riley and Welt, 1980). Cytochrome oxidase processing revealed dense modules corresponding to the predicted locations of S1 body areas and orofacial structures. (g) Comparison of the overall proportion of neocortex taken up by S1 in the naked mole rat (NMR) and Long Evans laboratory rat (Rat). S1 is approximately 50% larger, as a proportion of total neocortex, in the naked mole rat (right side of graph). A second comparison is the proportion of sensory cortex taken up by S1 in the two species (left side). Total sensory cortex here is defined as cortex caudal to the rostral border of S1.

Central Nervous System

As might be expected, the sensory specializations described above are reflected in an expanded somatosensory cortex. Somatosensory cortex (S1) in naked mole rats has been found to represent 31% of total neocortical territory and at least 47% of sensory cortex. This is a much larger area of the cortex than is typically taken up by somatosensory inputs. For example, S1 in laboratory rats takes up approximately 21% of neocortex and only 27% of all possible sensory cortex. Within mole rat somatosensory cortex, S1 is dominated by an expanded representation of the behaviorally important front teeth (Figure 4). This representation is dramatically magnified compared with the size of the dentition representation in rats, which takes up only about 7% of S1.

The S1 representation of the body can be visualized by reacting sectioned cortical tissue for CO. Although the subdivisions in mole rat cortex are less distinct that those in mouse cortex (Figure 1), this procedure nevertheless provides an indication of the relative position the various body-part representations in the cortex. It also reveals a series of modules within the S1 dentition representation located in rostrolateral cortex. These modules are similar to CO modules identified in primate area 3b (the homologous cortical area to S1 in other mammals), corresponding to the oral structures. One hypothesis for this modular representation of the dentition is that different modules process different kinds of touch information from the teeth.

Conclusions

Both star-nosed moles and naked mole rats have specialized somatosensory systems, yet each has different mechanoreceptor arrays and different areas of the body magnified in the neocortex. Star-nosed moles depend most heavily on the skin of the snout to explore their environment and have developed a tactile fovea on the star. This region of the body is greatly magnified in the cortical representation, which contains a series of modules representing the appendages of the star. The superior colliculus in star-nosed moles is dominated by somatosensation rather than vision, and this structure likely plays an important role in guiding saccadic head movements. In mole rats, a series of body whiskers is used for orienting in underground tunnels. In addition, the unusually large front teeth are used for collecting tactile information, and their representations are greatly magnified in the cortical map of the body. More than 30% of somatosensory cortex is devoted to the front teeth. Both species have been useful in understanding the organization and evolution of different somatosensory specializations in mammals.

See also: Neural Circuitry in the Somatosensory System; Sensorimotor Integration: Barrels, Vibrissae and Topographic Representations; Somatosensory Perception; Somatosensory Cortex: Functional Architecture; Somatosensory Receptive Fields; Somatosensory Plasticity; Somatosensory Cortex; Visual System: Adaptive Regression and Progression in Subterranean Mammals.

Further Reading

Catania KC (2000) Epidermal sensory organs of moles, shrew-moles, and desmans: A study of the family Talpidae with comments on the function and evolution of Eimer's organ. *Brain, Behavior and Evolution* 56: 146–174.

Catania KC and Kaas JH (1997) Somatosensory fovea in the star-nosed mole: Behavioral use of the star in relation to innervation patterns and cortical representation. *Journal of Comparative Neurology* 387: 215–233.

Cooper HM, Herbin M, and Nevo E (1993) Ocular regression conceals adaptive progression of the visual system in a blind subterranean mammal. *Nature* 361: 156–159.

Crish SD, Rice FL, Park TJ, and Comer CM (2003) Somatosensory organization and behavior in naked mole-rats I: Vibrissa-like body hairs comprise a sensory array that mediates orientation to tactile stimuli. *Brain, Behavior and Evolution* 62: 141–151.

Henry EC, Remple MS, O'Riain MJ, and Catania KC (2006) Organization of somatosensory cortical areas in the naked mole-rat (*Heterocephalus glaber*). *Journal of Comparative Neurology* 495: 434–452.

Jain N, Qi H, Catania KC, and Kaas JH (2001) Anatomical correlates of the face and oral cavity representations in the somatosensory cortical area 3b of monkeys. *Journal of Comparative Neurology* 429: 455–468.

Lacey EA, Alexander RD, Braude SH, Sherman PW, and Jarvis JUM (1991) An ethogram for the naked mole-rat: Nonvocal behaviors. In: Sherman PW, Jarvis JUM, and Alexander RD (eds.) *The Biology of the Naked Mole-Rat*, pp. 209–242. Princeton, NJ: Princeton University Press.

Park TJ, Comer C, Carol A, Lu Y, Hong HS, and Rice FL (2003) Somatosensory organization and behavior in naked mole-rats: II. Peripheral structures, innervation, and selective lack of neuropeptides associated with thermoregulation and pain. *Journal of Comparative Neurology* 465: 104–120.

Woolsey TA and Van der Loos H (1970) The structural organization of layer IV in the somatosensory region (SI) of mouse cerebral cortex: The description of a cortical field composed of discrete cytoarchitectonic units. *Brain Research* 17: 205–242.

Sentence Comprehension

M C MacDonald and R S Sussman, University of Wisconsin–Madison, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When people talk to each other, they translate their ideas into a series of sounds (or in the case of signed languages, gestures), and the recipient of the message must take this string of essentially arbitrary sounds or movements and interpret it as a discrete, abstract message. As such, processes of language comprehension run the gamut, from the earliest stages of perception of a speech sound or gesture to high-level inferences concerning the implications of the messages that have been comprehended. Sentence comprehension forms a middle ground in this sequence and concerns the processes by which a language perceiver transforms a linear string of words into an understanding of the essential sentence meaning.

The meaning of a sentence can be thought of as a hierarchically organized structure connecting and relating the meanings of individual elements to create propositional meaning. For example, in the sentences in example 1, a comprehender must understand that Silvia is the one doing the giving (that is, that Silvia has the thematic role of Agent in the sentence), that Mom is receiving something (Mom has the Recipient role), and the flowers are the thing being given (the Theme role). Although the three sentences in 1 have quite different word orders, these basic role assignments remain the same.

1a. Silvia gave some flowers to Mom.
1b. Silvia gave Mom some flowers.
1c. Mom was given some flowers by Silvia.

As this example illustrates, the role that any individual element will play in the event that is being described is determined by the sort of sentence structure it occurs in, as well as the position it occupies within that structure. Moreover, words and phrases toward the ends of utterances may have an effect on how the beginning of a sentence must be interpreted. Consider example 2:

2. I saw her fly . . .
   a. the fighter jet.
   b. was open.

Before listeners have heard the end of the utterance, they cannot be sure whether the phrase 'her fly' is the beginning of an embedded clause describing the actions of a pilot or a noun phrase denoting the front closure of a woman's pants. Sentence comprehension has the property of immediacy of interpretation: people begin interpreting language as soon as they hear it; listeners do not wait to find out how an utterance will end before they assign at least a tentative interpretation to its beginning. Immediacy of interpretation is an essential feature of the comprehension process because the acoustic speech signal is fleeting, and memory for unanalyzed acoustic information is strictly limited. Thus the memory of the input signal will be lost if it is not rapidly converted to a meaningful representation. The same is essentially true in reading, despite the fact that words on a page, unlike a speech signal, are available for reprocessing. If comprehenders did not advance beyond a visual or phonological (sound-based) representation of the written input for the early part of a sentence, then processing later parts of a sentence would simply create interference in the representation of the initial sequence. Thus, sentence comprehension proceeds through the rapid development of an initial interpretation of a sentence, with enrichment and adjustment of this interpretation as the sentence unfolds. The key research questions have concerned the nature of that preliminary interpretation and the mechanisms for its enrichment and revision. These issues have typically been addressed in investigations of sentence comprehension processes by adult native speakers of a language, but the work has clear overlap with studies of child language acquisition, comprehension processes in second language acquisition, and the nature of impairments in these processes following brain injury.

The first attempt to develop a theory of the processes executed during sentence comprehension emerged in the 1960s through a collaboration between linguist Noam Chomsky and psychologist George Miller, who tried to integrate the rapidly developing field of generative linguistics with studies of cognitive processes. The derivational theory of complexity (DTC) was built on Chomsky's transformational grammar, which posited that the structure of a sentence (the surface structure) was related to a meaning-based representation (a deep structure) through a series of transformations. The DTC posited that during the process of sentence comprehension, people mentally worked through the various sentence transformations to convert a perceived sentence to its deep structure, from which sentence meaning could then be extracted. The more transformations that had to be undone to arrive at the deep structure, the longer the sentence comprehension process should take. For example,

the passive sentence above 1c was hypothesized to need at least one more transformation than the active sentence 1a, yielding the prediction that passive sentences should take longer to comprehend than their active counterparts. Whereas early studies found good support for these predictions, later work suggested that the number of posited transformations had essentially no value in predicting comprehension difficulty. As a result, not only did the DTC fall out of favor as an account of sentence comprehension processes, but many psychologists became disillusioned with the idea that formal linguistic theories of the structure of language could offer insight into accounts of language comprehension.

More currently, research in sentence comprehension has addressed the cognitive processes, assumptions, and information that underlie the mechanisms and strategies of understanding everyday language in real time. This focus on the precise timing of comprehension processes emerges from a general interest in modularity claims within language research. Applied to sentence comprehension, the key modularity question concerns whether the many levels of language representation (sound, word, sentence, discourse, etc.) interact to derive sentence meaning or whether processes operate autonomously on different levels, saving any cross-level integration for the final stage of processing. Sentence comprehension research, situated between research on lexical (word) processing and the interpretation of entire conversations or texts, can be viewed as spanning three subfields – syntactic ambiguity resolution, syntactically complex utterances, and the relating of sentences to the broader discourse or text, including the interpretation of referentially dependent expressions such as pronouns.

Ambiguity Resolution

Language is rife with ambiguities at all levels of representation, and sentence comprehension research addresses the challenges presented by lexical ambiguity and syntactic ambiguity. Lexical ambiguities arise when a word has more than one meaning, as in example 3. Because the word 'glasses' can denote either eyewear or kitchenware, the intended meaning of the sentence is not obvious. Syntactic ambiguities, on the other hand, arise when a sequence of words is at least temporarily compatible with more than one sentence structure and more than one set of thematic roles and relations among the participants of the action. In example 4 below, the prepositional phrase 'with a lollipop' could be interpreted as modifying either the main verb of the sentence, as in 4a, such that the lollipop is the instrument used to tease the boy, or as modifying the object of the sentence, as in 4b. Lexical ambiguity and syntactic ambiguity frequently interact, as in example 2, in which the lexical ambiguity of the words 'her' and 'fly' created a syntactic ambiguity.

3. Phil wiped off the glasses.
4. The surly teenager teased the boy with the lollipop.
   a. The surly teenager teased the boy by using a lollipop (as bait).
   b. The surly teenager teased the boy who was holding the lollipop.

The modularity debate in syntactic ambiguity resolution has concerned the time course over which information is brought to bear on an ambiguity. Listeners must attach at least a partial interpretation to each word of an utterance as soon as it is encountered. The question then is, given that they have not yet heard the end of the utterance, on what information are they basing their interpretations? When do they integrate the information into their understanding of the unfolding utterance? Are certain types of information more useful (or used) than others?

One approach to these questions is exemplified by Lyn Frazier and colleagues' Garden Path model. 'Garden path' refers to an initial misinterpretation of a sentence, followed by a moment of surprise when later information in the sentence reveals that the initial interpretation was incorrect. The Garden Path model is a modular, two-stage approach in which the first-stage parser (the syntactic interpretation component of the comprehension system) develops an initial syntactic structure for the input, guided by only the lexical categories of the input words (noun, verb, etc.), the syntactic rules of the language, and structure-based heuristics that direct structure building when the given lexical categories can combine in more than one way within the syntax of the language. The structure-building heuristics typically favor syntactically simpler interpretations over more-complex ones. At some later point, a second stage integrates semantic and contextual information into the representation, and if this information conflicts with the initial interpretation built by the parser, the conflict may trigger a revision and reanalysis of the input. The guiding assumption of this and other 'syntax-first' models was that the demands of immediacy of interpretation could not be satisfied if all information were used in developing the initial interpretation, because probabilistic information, such as the implications of the prior context or the likelihood of the speaker's conveying a particular meaning, was too complex and would take too long to compute to be applied rapidly. The first-stage parser, with access to minimal and rapidly available information, was

therefore seen as serving the function of getting a syntactic interpretation started, leaving additional time for the second-stage system to finish computations.

The preeminent role accorded to the syntactic parser in the Garden Path model and other syntax-first models is accompanied by a prediction for sentence comprehension behavior during the interpretation of ambiguous sentences: namely that all nonsyntactic factors such as the context in which a sentence is produced, the actual words used, the relative frequency of its constructions with respect to the words involved, the prosodic features of the utterance, and so on, will affect the course of sentence comprehension only secondarily, in the second stage of processing. The alternative, constraint-based approach argues that the preference for one interpretation over another during comprehension of an ambiguous sentence emerges from the rapid combination of many probabilistic constraints, with no privileged role for syntactic information. Proponents of this approach explicitly reject the notion that nonsyntactic information is somehow harder to compute. A key observation of this approach is that different types of information tend to be correlated. For example, a verb's meaning is strongly related to the kinds of nouns and prepositional phrases (such as 'with a lollipop') it tends to appear with in sentences. As a result, even weak cues can combine with other correlated cues and have a strong effect on interpretation preferences.

Both the constraint-based and modular accounts have primarily been assessed with online measures, that is, precise, time-locked measures of comprehension processes as the sentence progresses, in order to look for behavioral evidence consistent with either early interaction or two modular subprocesses. A significant body of work in the 1980s and 1990s used measures of reading time during comprehension of written ambiguous sentences. Some early work using eye movement monitoring (eye tracking) during reading claimed to find evidence of a modular, two-stage interpretation system, with the earliest eye fixations on a region of text guided by operations of the first-stage parser, and later fixations on the text and rereading were driven by second-stage semantic integration processes. This view was motivated in part by studies in which manipulations of semantic information in syntactically ambiguous sentences were found to affect late eye fixations but not early ones. Subsequent studies suggested that at least some delayed effects of nonsyntactic information in eye fixation patterns were attributable to weak or infelicitous contexts or other biases in the ambiguous stimuli. More-robust manipulations of context have shown clear evidence of use of nonsyntactic information very early. These results challenged the modular claims for isolated subprocesses during sentence comprehension. Additional evidence for the constraint-based approach comes from studies of comprehenders' eye fixations on visual scenes while listening to spoken language related to the scene. Patterns of eye fixations suggest that information in the visual scene affects the early stages of interpreting ambiguous sentences. These results offer a powerful challenge to claims for modularity in sentence comprehension: not only does nonsyntactic information in the linguistic signal appear to constrain early syntactic ambiguity resolution, but even nonlinguistic information (such as from the visual scene) has a powerful early effect.

Complex Sentence Interpretation

A second area of investigation in sentence comprehension is syntactically complex sentences. The sentences in example 5 show two different types of relative clause. The example in 5a is known as a subject relative because the head of the relative clause (the element modified by the relative clause, here 'the reporter') is the subject of the attacking action in the relative clause. The object relative clause (also known as a center embedded clause) in 5b has a similar word order, but here the head of the relative clause ('the reporter') is the direct object of 'attacked' rather than the one doing the attacking.

5a. Subject relative: The reporter that attacked the senator admitted the error.
5b. Object relative: The reporter that the senator attacked admitted the error.

Despite the clauses' identical words and highly similar word order, sentences containing object relatives are typically much more difficult to comprehend than those with subject relatives. The added comprehension difficulty is not thought to reflect ambiguity resolution processes. The explanation of processing difficulty for object relatives varies, but the dominant view is that comprehenders encounter two unintegrated nouns (e.g., 'The reporter that the senator') before a verb, thus creating a working memory load in which two nouns must be held in memory while waiting for verb information that relates them. By contrast, in subject relatives such as 5a, the alternating pattern of nouns and verbs permits rapid integration of these sentence elements, reducing the memory burden. A minority view holds that object relative clauses are more difficult than the subject relatives because the object relative structure is unlike other sentence structures in the language, effectively leading to ambiguity in how to interpret them. If so, then relative clause interpretation may be a form of

ambiguity resolution, similar to that in the more widely recognized syntactic ambiguities discussed above.

Because of their syntactic complexity and their presumed dependence on working memory processes, relative clause structures have been a popular choice in investigations of sentence comprehension in atypical populations, including patients with aphasia, Alzheimer's disease, or other forms of dementia. The relationship between memory processes and relative clause comprehension has also been investigated in children, for whom both the ability to interpret relative clauses and the ability to perform well on working memory tasks develops fairly late.

Referentially Dependent Expressions and Links to Discourse Representations

Referentially dependent expressions are those which do not directly refer to something in the world but typically refer instead to some previously mentioned entity. For example, in 6a, we know that 'John' refers to a person, and we can interpret 'John' in this way even if we don't know him personally. By contrast, the pronoun 'he' does not refer directly to someone in this way but instead acquires its meaning indirectly, by being linked to the expression 'John.' This linking process is a form of ambiguity resolution, as the referent of a pronoun is not always clear. For example in 6b, 'he' might refer to either 'John' or 'Javier.' Research on pronoun comprehension spans both the ambiguity resolution processes that identify the correct referent of a pronoun and the consequences of pronominal reference – that is, how the comprehender's representation of the sentence changes as a result of having identified the referent of a pronoun.

6a. John went to the store after he got off work.
6b. John called Javier after he got off work.

The processes of pronoun reference resolution have a character similar to that of the processes of syntactic ambiguity resolution in that comprehenders appear to integrate a variety of probabilistic constraints to converge on the correct referent of a pronoun, including order of mention (earlier mentioned referents are more likely to be pronoun referents than later ones, at least within a sentence), grammatical role of referents (subjects of sentences are more likely to be referents than are other roles), and parallelism in grammatical role (nouns with the same grammatical role as the pronoun – e.g., both are subjects or both are objects – are more likely to be referents). These three constraints all serve to promote 'John' rather than 'Javier' as the referent of 'he' in example 6b.

In addition, the syntactic structure surrounding a referentially dependent expression may constrain the choice of referent. In 7a, the reflexive 'himself' must refer to John, while in 7b, the pronoun 'him' must not.

7a. John likes himself.
7b. John likes him.

These patterns of permissible co-reference have led to the hypothesis that syntactic structures define possible domains of reference, that is, the regions in which a referentially dependent expression may find a referent or may be blocked from finding a referent. Specifically, reflexives must find a referent within their own syntactically defined domain; pronouns must find a referent outside that domain. In examples 7a and 7b, this domain is the complete sentence. Thus, in 7a, 'himself' refers to the only possible referent within the relevant syntactic domain: 'John.' In 7b, 'him' can refer to any male entity except John. Extensive work within the field of linguistics has been devoted to defining these syntactic domains of reference, and modular accounts of pronoun resolution have suggested that these syntactic constraints have a privileged, early effect on pronominal reference resolution. However, recent work within both linguistics and sentence comprehension suggests that the syntactic domains are not so clear as originally thought, nor do they have a privileged status in referent resolution during comprehension. Online behavioral measures of referential interpretation (e.g., using eye tracking to monitor looks to displays of possible referents) confirm this intuition. These results suggest that while syntactic structure is certainly a factor in defining the preferred referents of pronouns and reflexives, other factors – such as point-of-view information or the specific lexical biases of other words in the construction – may also influence early stages of interpretation.

When someone identifies the referent of a pronoun, the mental representation of the sentence appears to be updated to reflect the re-mention (via pronoun) of a previously mentioned entity in the discourse. For example, when a comprehender has determined that 'he' must refer to 'John' in the examples above, the representation of 'John' appears to become more prominent in the discourse representation, as evidenced by shorter reaction times to the word 'John' compared with control conditions just after the pronoun has been comprehended. Moreover, other previously mentioned entities in the discourse (such as 'Javier' in 6b) appear to be inhibited and yield slower reaction times, as the pronoun indicates a shift toward focus on the referent in the ongoing linguistic input. The work

provides time-locked data concerning the processes of memory updating during the comprehension process.

Sentence Comprehension Methodologies

Research in sentence comprehension employs both online behavioral and brain imaging methodologies. Behavioral measures in sentence comprehension typically assess processing difficulty, as indexed by assessments of response time, errors, or both. The most common behavioral method has been reading times of particular words or regions within a sentence as a measure of processing difficulty. The focus on reading rather than comprehension of spoken sentences emerged out of the necessity to have a continuous measure of comprehension processes as the linguistic signal unfolds over time. Because reading proceeds at the comprehender's own pace, continuous measures of processing difficulty can be obtained from assessments of reading time, whereas there is no listening-time equivalent for spoken sentence comprehension.

In reading comprehension studies, the immediacy assumption holds that the reader is building an interpretation of the sentence as each new word or phrase is read. When the reader encounters a word or words that do not fit with this interpretation, this new material should take longer to read than in cases in which the material easily fits into the ongoing sentence interpretation. Comparing reading times of critical regions across different constructions yields a measure of the relative difficulty of integration of new material into the comprehender's interpretation of the sentence so far, which in turn sheds light on the interpretations that must have been developed by the reader in order for this difficulty to arise. The two primary reading measures are self-paced reading, in which the comprehender reads text on a computer screen and is allowed to see only a small portion of a sentence at a time, pressing a key to see each new word or region, and eye tracking, in which the entire sentence can be viewed while a device records the location and duration of eye fixations. Both measures yield data concerning the overall length of time spent reading a region, but eye tracking also yields data concerning the patterns of eye movements through a sentence, including patterns of rereading previously read text. Some researchers have interpreted different sorts of fixations within this sequence (e.g., the first fixation to a region vs. the second fixation or looks back to a region after reading later regions) as corresponding to the integration of different sources of information (syntactic, semantic, pragmatic, etc.) into the reader's interpretation of the sentence.

Eye movements have also been used to provide a time-locked measure of spoken sentence comprehension in the visual world paradigm, in which comprehenders view an array of objects or a scene on a computer and listen to speech referring to the scene, such as instructions to act on certain objects in view. Eye fixations on elements of the scene reflect their current state of understanding of the spoken material and its relationship to the visual environment. For instance, listeners who hear a sentence containing the verb 'eat' are more likely to look at edible items in a visual display than are listeners who heard a sentence that featured the verb 'take.' By manipulating both the visual display and the language accompanying it, researchers can investigate how context and language interact during spoken sentence comprehension.

Measures of brain activity as rendered by positron-emission tomography, functional magnetic resonance imaging (fMRI), magnetoencephalography, and event-related potentials (ERPs) have also been used to assess online sentence comprehension, often accompanied by an explicit task to encourage attention to the stimuli, such as to judge the sensibility of sentences. As with studies using behavioral measures, researchers typically compare materials that are similar save for one critical manipulation, such as contrasts between sensible sentences and scrambled words or between subject and object relative clauses.

Various fMRI studies of sentence comprehension have revealed a large number of areas that appear to be crucial to the comprehension process, beyond those that are important in interpretation of isolated words. These sentence-specific areas are believed to be important for interpreting the syntax of the sentence, identifying the thematic roles, and linking the sentence representation to prior knowledge. Whereas studies of patients with brain injury have named the left hemisphere as the site of virtually all language processes in most individuals, imaging studies have tended to find bilateral activation of a number of frontal and temporal brain regions during sentence comprehension, often with somewhat greater levels of activation in the left hemisphere than in the homologous area on the right. The fMRI studies of sentence comprehension have not factored heavily into the research issues addressing the time course of processing described above, which have been the focus of many behavioral studies in sentence comprehension. The temporal resolution of the blood oxygen level-dependent (BOLD) signal in fMRI, with about a 5 s latency, is not an appropriate instrument to address the precise time course of processes which behavioral data suggest operate on the order of a few dozen milliseconds.

The shorter latency of ERPs makes them more applicable to issues concerning the time course of comprehension, and some measures are thought to map onto specific comprehension processes. For instance, an early left anterior negativity is thought to be associated with syntactic difficulty, while negativity approximately 400 ms after stimulus onset is hypothesized to be an indicator of semantic discord. A more generalized process of information integration has been tentatively linked to a positivity 600 ms after stimulus onset. It should be emphasized, however, that these connections between activation patterns and specific comprehension processes are by no means uncontroversial and are at best only loosely understood.

See also: Connectionist Models of Language Processing; Language Development; Reading; Sentence Production; Speech Perception: Adult; Speech Perception: Development; Speech Perception: Neural Encoding; Speech Perception: Cortical Processing; Statistical Learning of Language; Word Recognition; Word Learning.

Suggested Reading

Altmann G (1998) Ambiguity in sentence processing. *Trends in Cognitive Sciences* 2(4): 146–152.

Ferreira F, Bailey K, and Ferraro V (2002) Good-enough representations in language comprehension. *Current Directions in Psychological Science* 11: 11–15.

Frazier L (1987) Sentence processing: A tutorial review. In: Coltheart M (ed.) *Attention and Performance XII: The Psychology of Reading,* pp.559–586. Hove, UK: Lawrence Erlbaum.

Garnham A (2001) *Mental Models and the Interpretation of Anaphora*. Philadelphia: Psychology Press.

Gibson E (1998) Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68: 1–76.

Henderson J and Ferreira F (2004) *The Interface of Language, Vision, and Action: Eye Movements and the Visual World.* New York: Psychology Press.

Kaan E and Swaab T (2002) The brain circuitry of syntactic comprehension. *Trends in Cognitive Sciences* 6(8): 350–356.

Lewis RL, Vasishth S, and VanDyke JA (2006) Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences* 10: 447–454.

MacDonald MC, Pearlmutter NJ, and Seidenberg MS (1994) The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101: 483–506.

MacDonald MC and Seidenberg MS (2006) Constraint satisfaction approaches to lexical and sentence comprehension. In: Traxler M and Gernsbacher MA (eds.) *Handbook of Psycholinguistics*, pp. 581–612. New York: Academic Press.

Pickering M and van Gompel R (2006) Syntactic parsing. In: Traxler M and Gernsbacher MA (eds.) *Handbook of Psycholinguistics*, pp. 455–504. New York: Academic Press.

Runner JT, Sussman RS, and Tanenhaus MK (2006) Assigning referents to reflexives and pronouns in picture noun phrases: Experimental tests of binding theory. *Cognitive Science* 30: 1–49.

Tanenhaus M, Spivey-Knowlton M, and Hanna J (2000) Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In: Crocker M, Pickering M, and Clifton C Jr. (eds.) *Architectures and Mechanisms for Language Processing*, pp. 90–118. Cambridge, UK: Cambridge University Press.

Townsend D and Bever T (2001) *Sentence Comprehension: The Integration of Habits and Rules.* Cambridge, MA: MIT Press.

Trueswell JC and Tanenhaus MK (2004) *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions.* Cambridge, MA: MIT Press.

Sentence Production

V S Ferreira, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sentence production refers to the ability of speakers of natural language to assemble known words into larger structured phrases or sentences. These phrases and sentences have two critical and related properties: They are grammatically well formed, as speakers generally do not produce 'the cat were happy' or 'cat happy the were,' and they convey the relational meaning of a sentence, or 'who did what to whom,' that the words by themselves cannot. For example, there is no single word (or, more precisely, morpheme – the smallest meaningful unit of language) that conveys 'the cat was happy'; one must assemble a sentence to relate one lexically denoted meaning, happiness, to another, the cat.

This article reviews different issues in sentence production research, roughly following a historical sequence. It begins with the first insights into sentence production from speech error evidence, followed by empirical and theoretical issues that have emerged from laboratory study, and moves on to more recent investigations of naturalistic issues. The article closes with a brief discussion of the neuroscience of sentence production.

Initial Insights: Evidence from Speech Errors

Compared with its sibling fields of language acquisition and language comprehension, research on language production in its current form began later and more slowly. This is likely due primarily to methodological challenges. In language comprehension at least, it is obvious how to precisely control the input to the system. For language production, however, it was not obvious how to get speakers to produce the sentences that were necessary for theory development and testing.

This methodological roadblock was overcome in the 1970s, through the pioneering work of Vicki Fromkin and Merrill Garrett. They sidestepped this input problem by amassing and exploring relatively large corpora of naturally occurring speech errors or slips of the tongue. These researchers made two critical discoveries that allowed speech error evidence to reveal the workings of sentence production mechanisms. First, they recognized that speech errors were often the (bungled) products of cognitive mechanisms involved in core language processing. At least for a relatively large class of errors, an utterance like 'that log could use another fire' is not a 'slip of the mind,' because speakers know when they make such errors, so they are not the result of bad conceptual knowledge. It is also (ironically) not a 'slip of the tongue,' because speech errors are unaffected by very peripheral articulatory or acoustic properties of speech. Instead, such errors are 'slips of the sentence production mechanism,' revealing the basic processing and representations speakers use to put sentences together. Second, Fromkin and Garrett discovered that when analyzed *en masse*, speakers slipped into consistent patterns, and these hinted at how sentence production mechanisms encoded nonlinguistic thought into phrases and sentences.

In a series of analyses published from the mid-1970s onward, Merrill Garrett used speech error evidence to lay a foundation for a theory of sentence production, the general outline of which can be recognized in most theories to this day. Garrett's initial investigations looked primarily at movement errors – slips in which one or more elements that a speaker intends are positioned in the wrong place. 'That log could use another fire' is a movement error because though the speaker intended to use 'log' and 'fire,' each word was produced in the place of the other and so were 'moved' with respect to the speaker's original intention. The same is true of 'I met someone who had a spet pider,' but with respect to the positioning of the sound 's.' Garrett first noticed that such movement errors clustered into two types. One typically involved words, like the log/fire error above, though larger units like phrases can be involved. In these errors, the words moved about as far as speakers could plan and nearly always came from the same part of speech (so that nouns traded places with nouns, verbs with verbs, and so forth). Such errors were also largely insensitive to the properties of the sounds of words. The other type of movement error typically involved individual speech sounds, like 'spet pider,' though again larger units like morphemes can be involved. In these, sounds rarely moved more than two or three words and exchanged between words of any part of speech. Also, the movement patterns in these errors were highly sensitive to the properties of the sounds of words, such as stress level or whether they were consonants or vowels.

These and related patterns of errors led Garrett to posit that sentence production proceeds through two distinct stages, sketched in Figure 1. The first type of

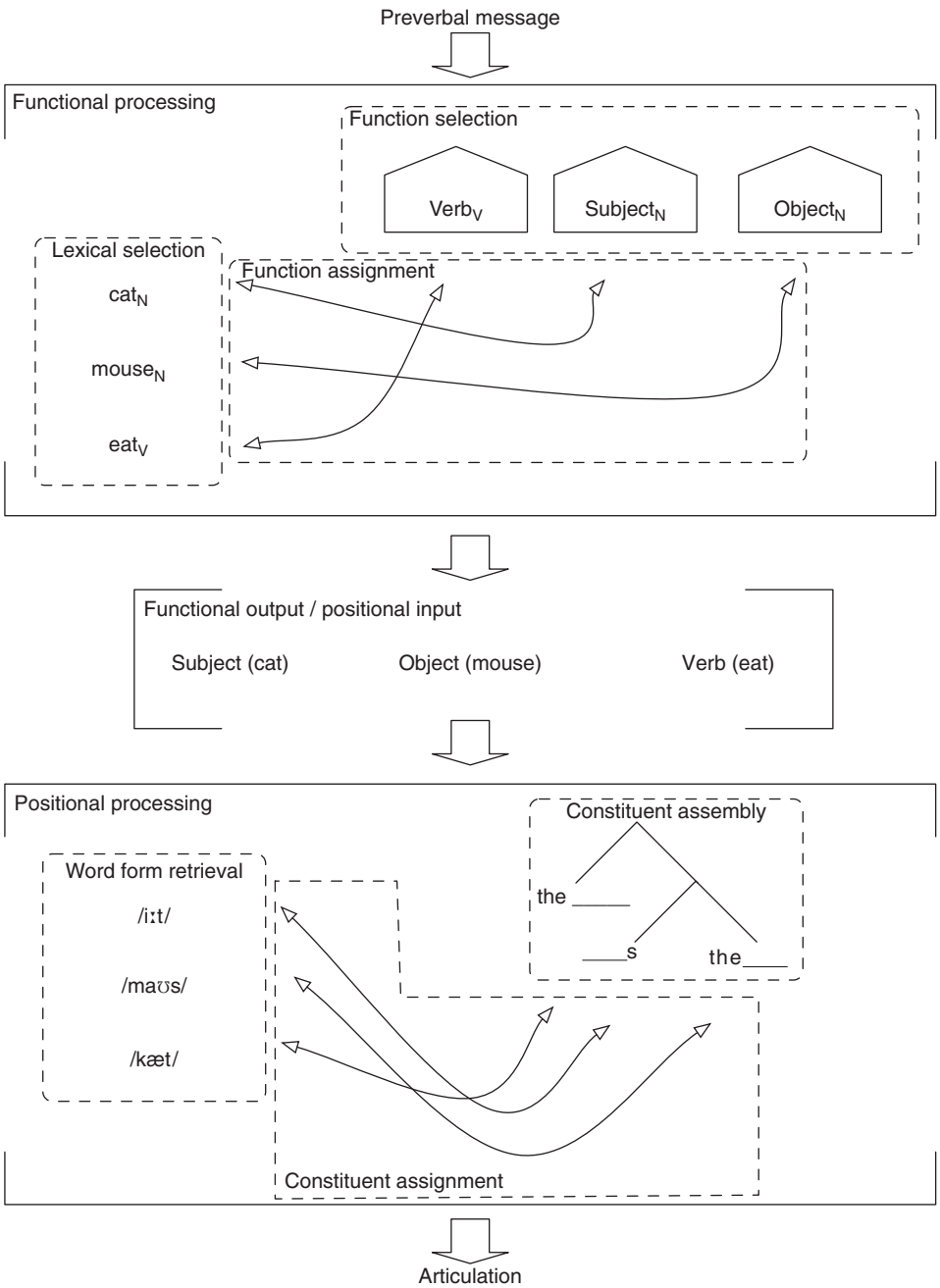

Figure 1 Sketch of the architecture of sentence production, as proposed by Garrett. Processing flow begins with the preverbal message (top), proceeds through functional processing and then positional processing, and concludes with articulation (bottom). V, verb; N, noun.

error reflects the organization of the first stage of sentence production, called functional processing, in which speakers assign words on the basis of their part of speech (noun, verb) to grammatical functions such as 'subject of the sentence.' These words are not phonologically defined, which is the reason why errors involving words are not sensitive to how they sound. According to Garrett, the processing dynamics at the functional level are independent of the spoken order of words. This explains why misplaced words can move large distances: what determines whether two units are processed at about the same time, and so might exchange, is unrelated to whether those units are adjacent in the eventual utterance. The second type of error reflects the organization of the next stage of sentence production, termed positional processing, at which speakers configure words into hierarchical constituent structures that determine their order in the spoken utterance. At positional processing, segmental content (individual speech sounds)

is specified, whereas functional information is not represented.

Though basic outlines of Garrett's proposal are still current, certain details of Garrett's proposals have been scrutinized. One prominent effort was a series of empirical and theoretical speech error studies carried out by Gary Dell. Dell first identified patterns of speech errors that revealed mutual influence or interactions among representations at different stages of processing. In particular, Dell showed that when one word replaces another that is similar in meaning, the words are more likely than chance to be similar also in sound. He also showed that when speakers erred in producing the individual sounds of words, the outcomes were more likely to constitute a known word than chance would predict. These patterns can be accommodated by assuming that processing at one level of sentence production is influenced by processing at other levels. So the phonological similarity of semantically related substituting words can be explained if representations at the analog of Garrett's positional level – phonologically defined words – are able to influence selection at the analog of Garrett's functional level – semantically selected and syntactically defined words. The tendency for sound errors to result in real words can be explained if the representations of words other than the one a speaker intends to say can influence the ongoing selection of individual sounds that a speaker does intend to say. Later, Dell developed a comprehensive connectionist model of the sentence production process that explained these observations and more. In connectionist models, knowledge structures are represented as units that become activated as a function of degree of evidence supporting those knowledge structures and are interconnected to represent the relationships among the knowledge structures. Dell's model could explain not only the speech error patterns already described but other, even less intuitive effects, such as the fact that sound errors are more likely to result in known words more as speech rate slows (because representations at different levels have more time to influence one another).

The Move to the Laboratory

The theory developed by Garrett proved seminal. Garrett's theory was important not only for investigations of speech errors; it succeeded in bootstrapping a laboratory sentence production enterprise as well. Recall that the primary impediment to developing a science of sentence production was the search for a valid way to elicit controlled, theoretically informative sentences. With Garrett's model, based as it was on face-valid evidence from speech errors, laboratory techniques further removed from natural production could confirm their own validity. This allowed laboratory investigations of sentence production to flourish until today, when diverse laboratory techniques are used to explore diverse theoretical issues. These issues are discussed next.

Scope of Planning

One of the first laboratory investigations of sentence production explored a simple question that has proven to have a rather complex answer: When speakers produce sentences, how far do they plan in advance, and what are the linguistic units they plan with? If freely speaking participants are asked to press buttons on hearing randomly presented tones, their responses are slowest at the ends of linguistic units called basic clauses, provided that they are about to produce more sentence material. A basic clause is a part of a sentence that includes a verb and all its directly associated arguments; thus, speakers evidently plan speech one basic clause at a time (fitting well with Garrett's claims regarding functional processing). Analyses of patterns of pausing in natural speech later verified this conclusion.

Subsequent work has revealed, however, that these original conclusions were likely too simple. Two factors complicate the story. First, the scope of planning seems different for different levels of sentence production. Early, during semantic and functional processing, the scope of planning is likely broader than later, during phonological and positional processing. Second, the scope of planning appears to be highly sensitive to task circumstances and speaker strategies. It is interesting that as production becomes more constrained and formulaic, speakers seem to plan ahead less. The idea is that speakers plan as little as they can get away with; when production is highly constrained and formulaic, speakers can be confident that they can come up with linguistic material to produce 'on the fly' and so can plan ahead less. With free-form production, speakers are less confident that they will be able to formulate on the fly, and so they plan ahead more. Relatedly, speakers are able to plan the same basic utterance about twice as fast (presumably because they plan ahead less) when given tight deadlines to begin speaking as they are if allowed to begin speaking whenever they like.

Accessibility Effects

Recall that Garrett assumed that functional processing is not affected by the order of words in the eventual utterance. A heavily investigated set of effects that revisits this assumption is termed accessibility effects. When speakers produce sentences, they tend

to mention easily retrieved or highly accessible material sooner, leaving harder-to-retrieve or less accessible material to be mentioned later. For example, Kay Bock discovered that speakers choose different syntactic structures to allow semantically primed words to be mentioned as the subjects of their sentences. Thus, speakers tend to say, 'The lightning struck the church' if semantically primed by 'thunder,' but 'The church was struck by lightning' when semantically primed by 'worship' – first retrieved, first spoken. This seems inconsistent with Garrett's assumption that functional processing is insensitive to the eventual orders of words in sentences.

Yet with reasonable assumptions, the core of Garrett's claim can be preserved. Specifically, one can claim that words that are accessible (e.g., because they are semantically primed) are selected sooner during functional processing (see the left box in Figure 1) and so are assigned to 'higher' grammatical functions (where 'subject' is higher than 'object,' for example). Across languages, higher grammatical functions tend to be mentioned earlier in sentences, and so these assumptions can accommodate accessibility effects without requiring that functional processing directly reference the eventual positions of words in sentences. Further support for this assumption comes from demonstrations that accessibility manipulations that are more meaning-based (e.g., differences in semantic priming or the imageability of words) more readily influence function assignment (e.g., whether something is produced as a subject or object); in contrast, manipulations that are more form-based (e.g., the lengths of words) influence word order independent of function assignment (e.g., the order of words in a conjunction like 'salt and pepper'). This fits with Garrett's original division of production into a meaning-based functional system and a form-based positional system.

Within models of sentence production, accessibility effects are generally motivated by incrementality – that sentences are produced piecemeal from beginning to end. Incrementality implies that if one part of a sentence becomes accessible before another, production is most efficient if the part that is accessible sooner is mentioned sooner and the part that is accessible later is mentioned later.

Grammatical Agreement

Another topic that can be couched within Garrett's framework is agreement, referring to when the form of one element in a sentence depends on some property of another element. In English, this arises with subject–verb agreement: One says, 'the cat was happy' and 'the cats were happy,' such that the form of the verb ('was' or 'were') depends on the plurality of the subject ('cat' or 'cats').

The core agreement-related object of study is attraction, in which some irrelevant part of a sentence influences agreement. For example, when given a subject noun phrase like 'The key to the cabinets' and asked to complete it, subjects are likely to say 'The key to the cabinets were lost,' with the verb ('were') agreeing with the local noun 'cabinets' rather than the subject noun 'key.' The empirical enterprise has primarily been to assess the properties of local nouns that do and do not cause such attraction. One factor that powerfully causes attraction is grammatical plurality: Local nouns that are grammatically plural (like 'cabinets') cause plural-agreeing verbs to be produced rather than singular-agreeing verbs. Curiously, this effect is asymmetrical, such that grammatically singular local nouns do not cause attraction (in utterances like 'The keys to the cabinet was/were . . .'). This has led to the proposal that in English, plurality is marked – it is an explicit feature that is associated with a noun argument; the singular, in contrast, is unmarked, such that nouns default to being singular when not otherwise specified.

Given the robust effect of plurality on attraction, the question arises whether properties that are correlated with plurality, such as plurals' meaning (they are usually more numerous than the singular, like 'keys' vs. 'key') or sound (the '-s' at the ends of words), cause attraction errors. Generally, such properties are much less potent (if influential at all) compared with true grammatical plurality. For example, local nouns that denote collectives but are not grammatically plural ('The general of the army . . .') do not cause attraction, nor do local nouns that sound like plurals but are not ('The petal on the rose . . .'). In contrast, local nouns that denote single entities but are grammatically plural ('The ad for the scissors . . .') cause attraction, as do nouns that are plural but do not sound that way ('The trap for the mice . . .').

Comprehensive theories of agreement are currently the subject of active debate, the details of which are beyond the scope of this article. The weight of evidence suggests that within Garrett's model, agreement processing should arise within positional processing. Some properties of attraction fit in with this, including the above-noted insensitivity to semantic correlates of plurality and the sensitivity to the hierarchical structure of sentences. However, given that positional processing has a phonological character, additional assumptions may be needed to explain why attraction is insensitive to phonological factors.

Syntactic Priming

A heavily investigated phenomenon in sentence production that does not fit as neatly into Garrett's framework is termed syntactic priming (sometimes

called structural or syntactic persistence). When a speaker produces or comprehends a sentence with a particular syntactic structure, the speaker is then more likely to use that structure rather than a reasonable alternative. For example, after hearing a prepositional dative structure ('The doctor gave the hat to the sailor'), speakers are likely to use another prepositional dative ('The ballerina showed the cake to the boxer') rather than a double-object structure ('The ballerina showed the boxer the cake').

Evidence has shown that at core, syntactic priming reflects the repetition of the syntactic structures of sentences. Properties correlated with syntactic structure either do not affect priming or affect priming independent of syntactic repetition. Such inconsequential properties include aspects of meaning (e.g., the order of agents and themes and recipients), sounds (e.g., rhythm or auditory structure), or words (especially syntactically relevant words like 'to') of sentences. Syntactic priming is enhanced by repeating meaningful words (especially verbs) in sentences but is reliable even without such repetition.

Three different theories of syntactic priming dominate current scientific discourse. One views priming as implicit learning, such that comprehending or producing particular structures (quasi-) permanently strengthens the knowledge of those structures, thereby resulting in repetition. Supporting this view are the observations that priming can be notably long-lived (surviving the comprehension and production of at least ten intervening neutral trials), is independent of explicit memory, and is evident in patients with anterograde amnesia. Another views priming as reflecting recent shorter-term activation of knowledge structures representing syntax. Supporting this is the observation that priming can be short-lived, especially when meaningful words are repeated from prime to target sentences. A third view is that priming is a force to coordinate knowledge structures among interlocutors in order to promote communicative success. Supporting this is the finding that priming is sensitive to conversational variables (such as the linguistic capabilities of the participants or whether participants are speakers, hearers, or side participants).

Earlier, syntactic priming was described as ill-fitting with respect to Garrett's framework. The reason is that priming operates similarly at both functional and positional processing. For example, the selection of an active rather than a passive sentence form – a sentence choice involving different assignments of grammatical functions – is influenced by priming in the same way as the selection of 'on the table is a book' versus 'the book is on the table' – a function-independent structure choice. Furthermore, to date, no aspect of priming has been revealed to dissociate along functional–positional lines (unlike accessibility effects, as noted earlier). Thus, though priming might affect the two separate processing stages identically, a more parsimonious proposal might be that the two stages should be collapsed into one.

The Separation of Words and Syntax

An area of active controversy in sentence production concerns the relationship between lexical and syntactic knowledge. At one extreme of a theoretical continuum, lexical knowledge could ground syntactic knowledge, such that the structure of a sentence is organized around the structures permitted by a particular verb, for example. These are termed lexically based theories. At the other extreme, lexical knowledge could be separate from syntactic knowledge, such that syntactic structures are constructed independently to obey communicative and grammatical constraints, and then words are slotted into those structures. These are termed frame-based theories. Between these are closed-class immanence accounts, whereby some words – closed-class or function words in particular, like 'the' and 'to' and 'by' – ground syntactic knowledge, with other words 'slotted in' as in the independent syntactic account.

Evidence from syntactic priming can be taken to support either extreme. Frame-based theories are supported by the fact that syntactic priming is observed even when no words are repeated between prime and target. Lexically based theories are supported by the fact that repeated words boost priming. Closed-class immanence accounts are supported especially by neuropsychological evidence showing that with breakdown of syntactic knowledge, speakers' use of function words is especially compromised.

From the Lab to the Wild

A current trend in sentence production work since the migration to the lab is toward evaluating issues of relevance to naturally produced sentences.

Corpora

The analysis of collections of natural speech, or language corpora, has a long tradition in sentence production research, with its current incarnation beginning with the publication of the famous Brown corpus in 1967. Ready access to greater computational power has allowed more-recent corpus work to take a considerable leap. Also, powerful statistical techniques such as mixed-effects modeling have allowed corpus researchers to isolate the influence of specific theoretically interesting factors while taking into account confounded uninteresting factors.

The Linguistic Data Consortium at the University of Pennsylvania is a useful source of corpus data.

Nonverbal Devices

Because of its laboratory tradition (and perhaps because of the emphasis on reading in sentence comprehension research), most investigations of sentence production have focused on issues relevant to 'idealized' utterances – sentences with an uninterrupted delivery characteristic of a prepared speech or words on a page. However, natural language is rife with disfluencies – violations of idealized delivery – including fillers like 'uh' and 'um,' backtracking, self-interruptions, and so forth. The lore was that such violations are nuisance factors, to be avoided or ignored. Recent work has challenged this assumption, instead identifying the performance factors that underlie disfluency. These performance factors seem to be metalinguistic, largely reflecting the state of mind of the speaker (e.g., an 'uh' and 'um' indicate that the speaker is having difficulty thinking of the material).

Gesture also sometimes conveys metalinguistic information. Some gesture simply accompanies the rhythm of a spoken sentence (what are sometimes called beat gestures). Other gesture conveys physical properties of the matter under discussion (e.g., 'The fish I caught was THAT big!'). Yet other gesture conveys information redundant with that in the speech signal, for example, the path along which an object is traveling. With the trend toward more naturalistic research, investigations of gesture are becoming more prominent.

Spoken sentence production has one other critical nonverbal property: because it is presented in a perceptually graded medium (sound), some of the acoustic properties of a sentence can vary independent of the lexical and syntactic properties (a situation that holds for signed languages as well). The higher-order structure of such acoustic properties is termed the prosody (e.g., rhythm and stress) of a sentence. Research suggests that these two levels of higher-order structure, syntactic and prosodic, are related but independent. Prosodic structure gets some of its characteristics from syntactic structure, but it is sensitive to its own pressures (including the facts that it must respect other acoustic properties of sentences that syntactic structure does not and that it represents a lower dimensionality of structure, without recursive properties).

Communicating Discourse Properties

Though sentences are the basic units of certain aspects of communication, they rarely occur in isolation. Rather, sentences are typically parts of discourses such as a story or conversation. Critically, devices within sentences help them cohere into discourses. This is sometimes termed the information structure of a sentence, referring to the linguistic devices that help convey the foreground and background status of referents in sentences. A systematic and universal example of a discourse device in sentence production is given–new ordering, whereby speakers tend to produce already described (given) information before newly introduced (new) information; accordingly, addressees can take early mention to indicate that they should already know a referent and later mention to indicate that they should add a referent to their discourse models. Note that such information–structure effects bear a tight relationship to the above-described accessibility effects. Indeed, it is reasonable to suppose that at least given–new ordering could emerge directly from accessibility if given items are more accessible than new.

The Neuroscience of Sentence Production

Thus far, this article has included little on the neuroscience of sentence production. Regarding traditional neuropsychology, this is because the range of work that has been done is extensive and detailed and involves its own set of phenomena and debates. A useful recent review is included in the recommended readings. An especially commendable characteristic of the primary literature on the neuropsychology of sentence production is that it has made very close contact with the core theories of normal sentence production outlined here.

Only a handful of studies have used techniques for assessing online neurological functioning (such as fMRI, PET, or EEG) to better understand mechanisms of sentence production. One major reason for this gap is (again) methodological: most techniques for assessing ongoing neural functioning are disrupted by head motion. It is difficult to investigate sentence production unless speakers actually produce sentences, and for spoken production, this requires the movement of the head and jaw. These challenges have begun to be overcome, however, and the cognitive neuroscience of sentence production promises to be an exciting avenue of research in the future.

Currently, debates in the cognitive neuroscience of sentence production have largely concerned the precise localization of general sentence production abilities. At this point, it is clear that brain regions within Brodmann areas 44 and 45 and adjacent regions are implicated in processes that combine words into grammatical sentences, as opposed to processes that just list words or simpler structures. Going forward, cognitive neuroscience investigations would (like

the traditional neuropsychological literature) gain by making greater contact with the rich and well-established range of theories on normal sentence production.

See also: Aphasia: Sudden and Progressive; Connectionist Models of Language Processing; Language Development; Language: Cortical Processes; Language: Aphasia; Psycholinguistics; Sentence Comprehension; Speech Production: Adult; Speech Production: Development; Word Production.

Further Reading

Bock JK (1982) Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review* 89: 1–47.

Bock JK (1986) Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 12: 575–586.

Bock JK (1986) Syntactic persistence in language production. *Cognitive Psychology* 18: 355–387.

Bock JK (1995) Producing agreement. *Current Directions in Psychological Science* 8: 56–61.

Clark HH (1996) *Using Language*. Cambridge, UK: Cambridge University Press.

Dell GS (1986) A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93: 283–321.

Ferreira F (1993) The creation of prosody during sentence production. *Psychological Review* 100: 233–253.

Ferreira VS and Slevc LR (2007) Grammatical encoding. In: Gaskell MG (ed.) *The Oxford Handbook of Psycholinguistics*, pp. 453–469. Oxford, UK: Oxford University Press.

Fromkin VA (ed.) (1973) *Speech Errors as Linguistic Evidence*. The Hague: Mouton.

Garrett MF (1975) The analysis of sentence production. In: Bower GH (ed.) *The Psychology of Learning and Motivation,* vol. 9, pp. 133–177. New York: Academic Press.

Griffin ZM and Ferreira VS (2006) Properties of spoken language production. In: Traxler MJ and Gernsbacher MA (eds.) *Handbook of Psycholinguistics,* 2nd edn. Amsterdam: Academic Press.

Indefrey P, Brown CM, Hellwig F, et al. (2001) A neural correlate of syntactic encoding during speech production. *Proceedings of the National Academy of Sciences of the United States of America* 98(10): 5933–5936.

Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Pickering MJ and Garrod S (2004) Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(2): 169–226.

Schütze CT and Ferreira VS (eds.) (2007) *MIT Working Papers in Linguistics 53: The State of the Art in Speech Error Research: Proceedings of the LSA Institute Workshop*. Cambridge MA: MITWPL.

Relevant Website

http://www.ldc.upenn.edu – The Linguistic Data Consortium.

Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation

A G Ramage, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Transmission

In this review the nature of 5-hydroxytryptamine (5-HT) transmission will be discussed with particular reference to the control of the cardiovascular system, micronutrition, the gut, and airways (autonomic nervous system) which most reviews fail to cover, concentrating on mood, cognition, perception, sleep, appetite, sex, and thermoregulation, to name but a few! There is direct evidence that 5-HT neurotransmission is involved in the central regulation of autonomic function as autonomic reflexes can be accurately measured and the underlying neuronal pathways have been fairly well delineated. Considering the relatively small number of 5-HT-containing neurons that are found in the brain, it is surprising that 5-HT is involved in so many functions; the total number of neurons in the mammalian brain is in the billions whereas the number of 5-HT-containing neurons is in the thousands. The majority of these 5-HT-containing neurons are located in the brain stem in the raphe nuclei, which are mainly located along the midline. These clusters of 5-HT-containing neurons have been labeled B1–B9. They send dense and divergent projections to all regions of the brain, as might be expected by their involvement in so many functions. Although some of these projections form classical synapses, many do not, indicating they may contribute to paracrine transmission, otherwise known as volume transmission (VT). VT would allow transmitter (information) to reach a variety of targets over a wide area, and thus it is more likely that 5-HT-containing neurons are involved in the modulation of a series of processes, such as rhythm generation. It should be noted that VT can also involve communication with glial cells. When 5-HT-containing neurons form classical synapses with their target neurons, which are sometimes referred to as 'private synapses,' the transmitter has only a very short distance to diffuse within such synapses, and so its action is rapid. Such fast transmission is referred to as 'wiring transmission' (WT). VT differs in that it can allow neuronal interaction via the transmitter and additional substances in the extracellular space. The ability of 5-HT to be involved in VT may explain the involvement of 5-HT neurons in such diverse functions. Data in support of this idea, that VT is occurring in a particular system, is the extrasynaptic localization of transmitter uptake transporters. In addition, the large number of receptor subtypes that 5-HT can interact with, which can have opposing actions, gives a vast repertoire of possible responses. For VT, the topographical organization of these different receptor subtypes would be important.

More recently, retrograde neural transmission has been demonstrated; however, whether this applies to 5-HT remains to be determined as presumably 5-HT-containing vesicles would be required in the postsynaptic specialization. The question arises as to whether 5-HT is the only transmitter released by 5-HT-containing neurons. In this respect, the term '5-hydroxytryptaminergic or serotoninergic' is often probably inappropriately used to indicate that these effects are mediated by such neurons using that one named transmitter. This is known as 'Dale's principle.' There is, however, much evidence that 5-HT-containing neurons release more than one transmitter and that cotransmission may be considered the norm. Normally cotransmission means the release of a peptide together with the main transmitter, which is usually associated with higher frequencies of neuronal firing. It is interesting to note that for 5-HT, there is also good evidence that glutamate may be co-released. In this respect, it has been suggested that where 5-HT-containing neurons form conventional synapses, that is, are involved in WT, glutamate only is released from these synapses, and at unconventional synapses (VT), 5-HT is released. This would in a sense be a return to Dale's principle in that each terminal (synapse) is releasing only one type of transmitter, although overall the neuron will release more than one type of transmitter depending on the synapse.

The Need for Multiple Receptors

The affinity of 5-HT for the different 5-HT receptor subtypes varies. For instance, the affinity that 5-HT has for 5-HT_1 receptor subtypes is higher than for most other subtypes, with 5-HT having the lowest affinity for the ligand-gated ion channel receptor, the 5-HT_3 receptor. This different affinity enables the target neuron or neurons to respond differentially to different amounts of 5-HT released, with lower concentrations favoring 5-HT_1 receptor activation. In addition, receptor desensitization rates vary; thus a slow and regular firing rate, a supposed characteristic of 5-HT-containing neurons, will mainly recruit 5-HT_1 and 5-HT_2 receptors. However, high firing rates would recruit 5-HT_3 and 5-HT_4 receptors.

All these receptors may undergo cycles of desensitization and recovery as firing patterns change. Thus changes in firing patterns would activate different receptors and in turn initiate temporal changes in the responsivity of the target neurons and so on to the next coding sequence. In addition, the fact that 5-HT receptors use a large repertoire of second messengers enables 5-HT neurons to have a large capacity to influence cellular events modulated by these second messengers. Thus it would be possible for neurons to have a 'memory' of the normal inputs they receive from 5-HT terminals. Overall, there are multiple ways that central 5-HT-containing neurons can modify transmission in other neurons. It should also be noted that 5-HT neurons can modulate their output of transmitter by using dendritic and somatodendritic autoreceptors.

Evidence that 5-HT Transmission Is Involved in Physiological Processes with Reference to Central Control of the Autonomic Nervous System

Central Cardiovascular Regulation

Brain stem 5-HT pathways are known to play important roles in cardiovascular regulation, although the precise role of each of the identified 5-HT receptor subtypes in this process remains to be determined. The main receptor subtypes that are known to be in this pathway are 5-HT_{1A}, 5-HT_3, and 5-HT_7. This has been determined using selective 5-HT receptor antagonists to interfere with defined reflex pathways, that is, inputs, which, when activated, give a well defined and characteristic output. In addition, studies depleting 5-HT have been carried out although it is very hard to know if all the 'functional' 5-HT has been removed in such studies. Nevertheless, data indicate that in rats with around only 6% of their total brain stem 5-HT, blood pressure is elevated and baroreceptor gain is attenuated. On the other hand, there is, at present, no evidence that 5-HT receptor antagonists can increase resting blood pressure in rats, suggesting that background tone in the resting condition is not mediated by 5-HT-containing neurons. In this respect, it has been suggested that 5-HT_7 receptors are involved in normalizing blood pressure after it has been destabilized by sensory input. This may have implications in cardiovascular changes induced by stress. The main types of receptors that have been identified in controlling sympathetic drive and thus the central maintenance of blood pressure are 5-HT_{1A} (sympathoinhibitory) and 5-HT_2 (sympathoexcitatory). Data suggest that it is the A subtype of the 5-HT_2 receptor that is primarily involved. However, blockade of either of these receptor subtypes has little effect on resting blood pressure. (It is interesting to note that it is these receptors which are also believed to be responsible for the hallucinogenic effects of LSD and related compounds.) It is somewhat surprising that 5-HT_{2A} receptor antagonists do not lower blood pressure, as activation of these receptors causes large central increases in sympathetic drive. However, this may be a problem with selectivity of the 5-HT_2 receptor antagonists used. Long-term block of 5-HT_2 receptors does prevent deoxycorticosterone acetate-salt hypertension, which probably reflects the importance of central 5-HT-containing neurons in the release of vasopressin (antidiuretic hormone). The release of vasopressin, though, involves the activation of 5-HT_{2A} receptors on a central angiotensinergic pathway. Overall, this implicates central 5-HT pathways in the control of blood volume, and in this respect, one of the characteristic unwanted effects of selective serotonin reuptake inhibitors is dilutional hyponatremia due to the release of vasopressin.

It is the role of central 5-HT neurotransmission in the control of vagal tone to the heart rather than sympathetic tone to the blood vessels that has been more thoroughly elucidated. The archetypical 5-HT_{1A} receptor agonist 8-OH-DPAT was observed to cause large increases in vagal drive to the heart. Vagal tone to the heart can also be increased by activation of a number of cardiovascular reflexes (cardiopulmonary and baroreceptor reflexes and the chemoreflex), and the afferents for these reflexes run in the IXth and Xth (vagus) cranial nerves and terminate along with other visceral afferents (e.g., gut) in the nucleus tractus solitarius (NTS). Of these reflexes, the cardiopulmonary has been identified in all species tested as involving 5-HT_{1A} receptor-mediated transmission. Those sensory nerve endings located in the pulmonary circulation and in the heart and coronary vessels have 5-HT_3 receptors located on them. The sites at which 5-HT transmission can modulate these reflexes in the central nervous system (CNS) are the NTS, the dorsal vagal motor nucleus, and the nucleus ambiguus. These latter two nuclei contain vagal (parasympathetic) preganglionic neurons. In the former, the neurons have unmyelinated axons mainly innervating parasympathetic ganglia (myenteric plexus) in the gut, while the latter have myelinated axons and mainly innervate parasympathetic ganglia in the heart and airways. Their activation could cause the heart to stop or cause severe bronchoconstriction. Central application of the archetypical 5-HT_{1A} receptor antagonist WAY-100635 attenuates the cardiopulmonary reflex vagal bradycardia, and this is believed to occur at the level of the nucleus ambiguus. As 5-HT_{1A} receptors are

usually inhibitory, this excitatory action is probably one of disinhibition (see Figure 1). It has thus been suggested that when NTS neurons activate cardiac vagal preganglionics in the ambiguus, this causes the release of 5-HT, which activates a 5-HT_{1A} receptor inhibiting a tonically active γ-aminobutyric acid (GABA)-ergic interneuron.

More recently, a 5-HT_7 receptor-mediated pathway has been shown to play a global role in the control of vagal tone by all these reflexes. 5-HT_7 receptors differ from 5-HT_{1A} receptors in that they are excitatory, increasing the levels of cyclic adenosine monophosphate. The site at which these receptors act is believed to be within the NTS as 5-HT_7 receptor blockade affects most of the components of these reflexes. It should be noted that there is overlap in the pharmacology of 5-HT_7 and 5-HT_{1A} receptors. In addition, at the level of the NTS, 5-HT_3 receptors are also involved in the vagal afferent excitation of NTS neurons. They do this indirectly by causing the release of glutamate, which has also been identified as a crucial transmitter in these reflex pathways. It is interesting to note that in dealing with the observations that blockade of 5-HT_7 receptors or glutamate receptors abolishes the baroreceptor reflex, the question arises as to why both transmitters are seemingly of equal importance. Furthermore, how 5-HT_3 receptors fit into such a system remains unclear. The importance of 5-HT_3 receptors in the control of these reflex pathways, although seemingly essential for transmission within the NTS, is at present not clear. It is thought that 5-HT_3 receptors may be located on glial cells, causing the release of glutamate from glial cells themselves. However, it is surprising that after vagotomy, there is considerable reduction in 5-HT_3 receptor density within the NTS, indicating that 5-HT_3 receptors are mainly located at the terminals of vagal afferents; see Figure 2 for a proposed functional 5-HT pathway within the NTS.

Thus 5-HT transmission plays an important role in cardiovascular afferent integration and in controlling vagal efferent output to the heart. Although 5-HT receptor activation can affect sympathetic outflow

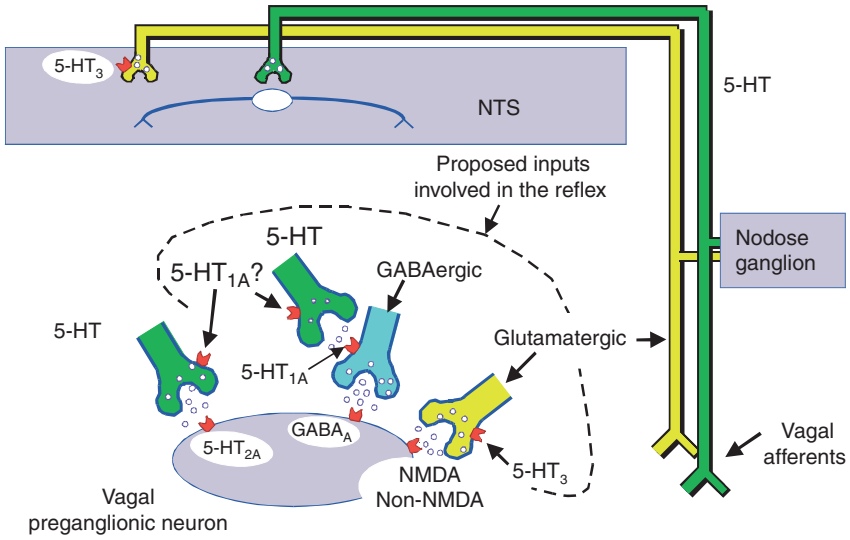

Figure 1 A diagrammatic representation of involvement of pathways containing 5-hydroxytryptamine (5-HT) and receptors in the control of the activity of cardiac and bronchoconstrictor (airway) vagal preganglionic neurons. The 5-HT-containing neurons are in green whereas glutamate-containing neurons are in yellow. The vagal preganglionic neurons can be reflexly activated via the nucleus tractus solitarius (NTS) by the cardiopulmonary afferents that run in the vagus (see Figure 2 for fuller details of the role of 5-HT_3, 5-HT_7, glutamate, and glial cells in activating NTS neurons). It is proposed that the NTS neurons activate (1) a 5-HT-containing and (2) a glutamatergic pathway. The 5-HT pathway inhibits the γ-aminobutyric acid (GABA)-mediated (blue) 'brake,' allowing the glutamatergic pathway to fully excite the preganglionic vagal neurons, resulting in a bradycardia or bronchoconstriction. It should be noted that the 5-HT_{1A} receptors are located presynaptically, not postsynaptically. In addition, the diagram shows a 5-HT-containing pathway directly innervating vagal preganglionic neurons, which activates 5-HT_{2A} receptors to cause excitation. This pathway is not believed to be involved in the reflex activation of vagal preganglionic neurons; however, it is speculated that the putative 5-HT_{1A} receptors that mediate inhibition of vagal preganglionic neuronal activity could be located on the nerve terminals of this pathway and as such would function as autoreceptors. It should be noted that these inhibitory 5-HT_{1A} receptors could also be located on the terminals of the 5-HT-containing pathway, which inhibits the putative GABAergic brake. This could apply to all bronchial and bladder vagal preganglionic neurons in which 5-HT_{1A}-mediated transmission has been identified. NMDA, *N*-methyl-D-aspartate. Reproduced from Wang Y and Ramage AG (2001) The role of central 5-HT_{1A} receptors in the control of B-fibre cardiac and bronchoconstrictor vagal preganglionic neurons in anaesthetized cats. *Journal of Physiology* 536: 753–767, with permission from Blackwell publishing.

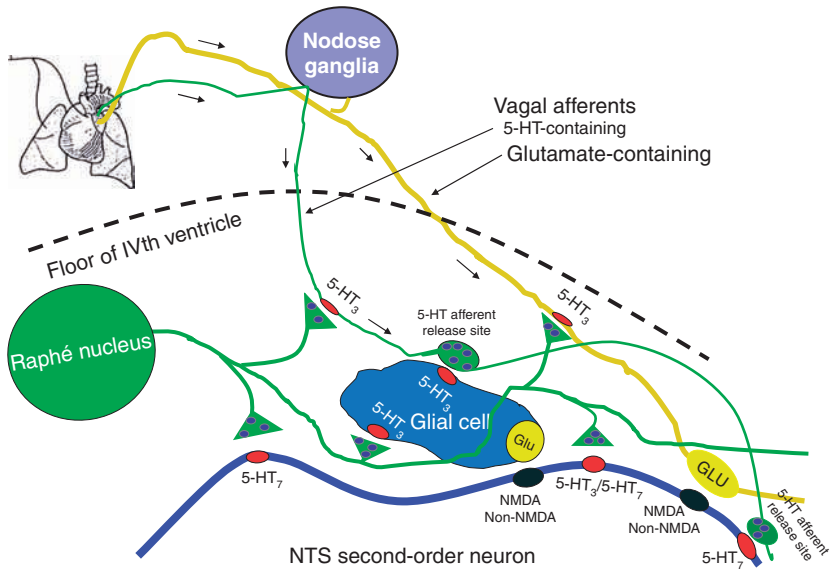

Figure 2 A diagrammatic representation of the interaction of vagal sensory afferents, involved in the cardiopulmonary reflex, and a 5-HT-containing projection from the raphe on a nucleus tractus solitarius (NTS) neuron. It is proposed that the vagal afferent can contain either 5-HT (green) or glutamate (GLU; yellow). It should be indicated that there is no reason why both transmitters cannot be contained in one afferent fiber. Further, these vagal afferent projections could just release either 5-HT or glutamate. However, for the present it is considered that they release both. In this respect, the 5-HT-containing afferents make apposition to the glial cell (the transmission here could be volume transmission (VT) or wiring transmission (WT); intuitively, VT is preferred). The released 5-HT activates 5-HT_3 receptors, which cause the release of glutamate from the glial cell and thus excitation of the NTS neuron via the activation of *N*-methyl-D-aspartate (NMDA) and non-NMDA receptors. The other afferent releases glutamate directly to activate the NTS neuron. It is proposed that the 5-HT-containing raphe projection controls the 'gain' of this system via activation of 5-HT_3 receptors on the sensory afferent input and/or glial cell and 5-HT_7 receptors on the NTS neuron (this again could be a mixture of WT and VT). It should be noted that 5-HT terminal autoreceptors would be expected to be found in this system. This proposed system could apply to other afferent processing involving 5-HT.

to blood vessels, a role for 5-HT transmission in sympathetic (blood pressure) control is unclear, but 5-HT does seem to play a role in the regulation of blood volume by the CNS.

Airways

The predominant drive to airway smooth muscle is again the vagus, and again data indicate that 5-HT neurotransmission operating through 5-HT_{1A} and probably 5-HT_7 receptors is important for their regulation, increasing vagal drive. These bronchoconstrictor preganglionic neurons have a respiratory modulation effect opposite to that of cardiac neurons' being excited on inspiration. So the effects of 5-HT are unlikely to be due to an action on respiratory drive. Sadly, very little experimentation has been carried out on such neurons despite their ability to close down the airways.

Micturition – The Bladder

Storage and elimination of urine are performed by the bladder and urethra. This involves the sympathetic nervous system, which is considered important for storage, that is, increasing the compliance of the bladder, and the parasympathetic, which causes contraction of the bladder and thus voiding. The somatic nervous system is important in the control of the external urethral sphincter for the control of continence. The CNS seems to control micturition at two levels, one in the midbrain and the other at the sacral level of the spinal cord, where the preganglionic vagal neurons innervating the bladder are found. It has long been thought that 5-HT neurotransmission is inhibitory to micturition. As central regulation of micturition is very complex, it is not surprising that the majority of 5-HT receptors have been implicated; in order of known importance, they are: 5-HT_{1A}, 5-HT_7, 5-$HT_{2A/C}$, 5-HT_3, 5-HT_4, and 5-HT_{5A}. Again, what is surprising is that both 5-HT_{1A} and 5-HT_7 receptor antagonists inhibit micturition, indicating, as above, that although both have opposite effects on adenylyl cyclase, both play an excitatory role in this process, implying that 5-HT-containing pathways can be considered excitatory rather than inhibitory. The 5-HT pathway utilizing 5-HT_7 receptors operates only at the level of the brain stem whereas the 5-HT_{1A} pathway operates at both the brain stem and the sacral spinal cord regions involved in micturition.

A pathway mainly operating through 5-HT_{2C} receptors has also been strongly implicated in the control of micturition, and this pathway is proposed to be inhibitory. As this is an excitatory receptor, again disinhibition has been proposed for these actions, that is, activation of GABAergic interneurons. Indeed, administering selective 5-HT_{2C} receptor agonists does cause the inhibition of micturition, and this occurs at both brain stem and spinal cord levels. Nevertheless, the selective 5-HT_{2C} receptor antagonist SB 242084 did not affect the micturition reflex, indicating that these receptors have no physiological role. However, preliminary experiments suggest that there may be an excitatory role for 5-HT_{2A} receptors.

Recently the discovery of a high density of 5-HT_{5A} receptors in Onuf's nucleus, the site of somatic motor neurons controlling the external urethral sphincter, implies a role for 5-HT transmission in the control of this nucleus, which is also important for fecal continence.

Thus, the above data would suggest that depletion of brain 5-HT or destruction of 5-HT neurons would interfere with micturition. This has not been observed so far, although destruction of 5-HT neurons and terminals in the spinal cord by around 75% increased micturition volume, favoring an excitatory action of 5-HT.

Another question that arises from the above observations is, where are the 5-HT-containing neurons located that are involved in these actions? Further, do these neurons send axons all the way down to the sacral spinal cord? In the brain stem/pons area, the major structures involved in the control of micturition are the pontine micturition center, periaqueductal grey, and locus coeruleus. In this area of the brain is also located the dorsal raphe nucleus. This nucleus has very large numbers of 5-HT-containing neurons, which mainly project rostrally to the cortex. It is possible that the 5-HT-containing terminals, which presumably innervate these micturition areas, originate from this nucleus, probably operating by activating 5-HT_7 receptors to initiate micturition and/or switch from filling to voiding, that is, acting as a 'flip-flop switch.' Once more, 5-HT_{1A} receptors, as for cardiac vagal neurons, appear to switch off a tonic GABAergic input, a 'brake,' in the pontine micturition center so these neurons can then activate bladder parasympathetic preganglionic neurons to cause contraction of the bladder and voiding. This system could also act at the level of parasympathetic neurons in sacral spinal cord. However, the site these sacral 5-HT terminals come from remains to be determined. Yet again the role of 5-HT neurotransmission in controlling the sympathetic and sensory afferent side of this system remains to be fully determined.

Gut

5-HT is an important neurotransmitter and paracrine signaling molecule in the gut and is released from enterochromaffin cells by local stimulation of the intestinal mucosa. The released 5-HT activates 5-HT_3 and 5-HT_4 receptors on sensory afferents, initiating the peristaltic reflex. In this respect, treatment with a 5-HT_3 receptor antagonist slows down transit time through the gut. Another use of these antagonists is the treatment of vomiting and nausea. This action is believed to occur at two levels, at the level of the sensory nerves in the gut and in the area postrema (chemical triggers zone), sometimes referred to as the vomiting center and one of the circumventricular organs (allowing factors to 'circumvent' the blood–brain barrier). It also projects to the NTS, which is found just below it on the floor of the fourth ventricle. The area also receives vagal afferents and is rich in 5-HT_3 receptors, which again are mainly on vagal afferent nerve endings, as in the NTS (see above). The data suggest that these vagal afferents again release 5-HT; whether directly or indirectly remains to be determined.

Overview

The above data indicate that 5-HT neuronal transmission plays a vital role in the central control of the autonomic nervous system. How this 5-HT innervation integrates between the use of volume transmission and so-called tight private synapses to control these processes remains to be elucidated. In addition, the relationship between the different types of receptors involved also remains to be determined, and the site from which this 5-HT innervation arises is still unknown.

See also: Autonomic Nervous System; Cotransmission; Parasympathetic Nervous System; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors; Sympathetic Nervous System.

Further Reading

Bunin MA and Wightman RM (1999) Paracrine neurotransmission ion the CNS: Involvement of 5-HT. *Trends in Neuroscience* 22: 377–382.

Crowell MD (2004) Role of serotonin in the pathophysiology of the irritable bowel syndrome. *British Journal of Pharmacology* 141: 1285–1293.

Damaso EL, Bonagamba LGH, Kellett DO, et al. (2007) Involvement of central 5-HT_7 receptors in modulation of cardiovascular reflexes in awake rats. *Brain Research* 1144: 82–90.

De Groat WC (2006) Integrative control of the lower urinary tract: Preclinical perspective. *British Journal of Pharmacology* 147: S25–S40.

Jacobs BL and Azmitia EC (1992) Structure and function of the brain serotonin system. *Physiological Reviews* 72: 165–229.

Jordan D (2005) Vagal control of the heart: Central serotonergic (5-HT) mechanisms. *Experimental Physiology* 90: 175–181.

Ramage AG (2001) Central cardiovascular regulation and 5-hydroxytryptamine receptors. *Brain Research Bulletin* 56: 425–439.

Ramage AG (2006) The role of central 5-hydroxytryptamine (5-HT, serotonin) receptors in the control of micturition. *British Journal of Pharmacology* 147: S120–S131.

Silva ALS, Cabral AM, Abreu GR, et al. (2005) Chronic treatment with mianserin prevents DOCA-salt hypertension in rats: Evidence for the involvement of central 5-HT_2 receptors. *European Journal of Pharmacology* 518: 152–157.

Syková E (2004) Extrasynaptic volume transmission and diffusion parameters of the extracellular space. *Neuroscience* 129: 861–876.

Trudeau L-E (2004) Glutamate co-transmission as an emerging concept in monoamine neuron function. *Reviews Psychiatry Neuroscience* 29: 296–310.

Uphouse L (1997) Multiple serotonin receptors: Too many, not enough, or just the right number? *Neuroscience and Behavioral Reviews* 21: 679–698.

Serotonin (5-Hydroxytryptamine; 5-HT): Receptors

T P Blackburn, Helicon Therapeutics Inc., Farmingdale, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During the past 50 years, since its discovery in 1948, studies of 5-hydroxytryptamine (5-HT; serotonin) have provided considerable insight into the metabolism and physiological functions of this monoaminergic neurotransmitter in the peripheral and central nervous system (CNS). Considerable interest in the biosynthesis, metabolism, and receptor pharmacology of 5-HT (Figure 1) has elevated this simple indolealkylamine into one of the most researched neurotransmitters of the twentieth century. The synthesis of 5-HT begins with the essential amino acid tryptophan, which can be found in foods such as bananas, turkey, and milk. The enzyme tryptophan hydroxylase hydroxylates tryptophan's benzene ring, to form 5-hydroxytryptophan. Aromatic amino acid decarboxylase then decarboxylates 5-hydroxytryptophan to produce 5-HT. 5-HT is formed in nerve terminals, where it is held in storage granules; when nerve impulses release 5-HT into the synaptic cleft, it acts on the postsynaptic neuron to affect the process of neurotransmission across the synapse. Released 5-HT is inactivated primarily by an active serotonin reuptake transporter (SERT) process that moves 5-HT back into the neuron. Once inside the 5-HT neuron, the neurotransmitter is either repackaged into storage granules or degraded metabolically by monoamine oxidase (MAO). Pre- and postsynaptic 5-HT receptors exist on the 5-HT nerve terminals and cell bodies. Activation of these receptors leads to a disruption in the synthesis and release of serotonin. The role of the presynaptic autoreceptor is to modulate the concentration of serotonin in the synaptic cleft, thereby reducing further release and synthesis of 5-HT (see the sections 'The 5-$HT_{1A(Gi/Go)}$ receptor' and 'The 5-$HT_{1B(Gi/Go)}$ receptor').

It is probable that early work on the synthesis, neurochemistry, and distribution of 5-HT stimulated the expectation that 5-HT might be important in anxiety, depression, and schizophrenia. It is quite remarkable, and a far cry from those early days, that the total sales of drugs modulating central and peripheral 5-HT activity exceeded $20 billion (US) in 2006. This is undoubtedly due to an increasing awareness of the 5-HT receptor subtypes associated with the multitude of disorders treated by drugs modulating 5-HT function; migraine, chemotherapy-induced emesis, depression, anxiety, schizophrenia, and irritable bowel syndrome are examples of such disorders. The promise of further therapeutic opportunities emerging awaits the development of more selective agents (combination agents) and the possibility of discovery of further novel 5-HT receptor subtypes.

Evidence for Multiple Serotonin Receptor Genes

The classification and nomenclature of 5-HT receptors in the past 10 years has emerged from criteria established by the International Union of Basic and Clinical Pharmacology (IUPHAR) Subcommittee for the Classification and Nomenclature of 5-HT receptors Guide to Receptors and Channels (GRAC) 2007 (Figure 2; Tables 1–5). To date, 13 distinct subtypes of human serotonin receptor are recognized on the basis of structural, transductional, and operational characteristics. A fourteenth purported receptor subtype (5-ht_{5B}) has been identified in rodents only.

The various subtypes are subdivided into seven different classes (5-$HT_{1–7}$), based on gene encoding, and in some cases only one gene product has been identified. These gene products are referred to by use of lower case (e.g., 5-ht_{1E}, 5-ht_{1F}) and will only be accepted into the 5-HT family and given 'upper-case' status when there is unequivocal evidence that the receptor exists as an operational endogenous, physiologically relevant receptor.

G-Protein-Coupled 5-$HT_{(1–7)}$ Receptor Family

5-HT_1 serotonin receptors are expressed in a variety of neurons in the peripheral and central nervous systems. The 5-HT_1 receptor family – 5-HT_{1A}, 5-HT_{1B} (formerly termed 5-$HT_{1D\beta}$), 5-HT_{1D} (formerly 5-$HT_{1D\alpha}$), 5-ht_{1E}, and 5-ht_{1F} – are all seven-transmembrane (7TM) G-protein-coupled receptors (via G_i or G_o; see nomenclature in Figure 2). They are all encoded by intronless genes (of approximately 365–421 amino acids) and are negatively coupled to adenylyl cyclase, and principally cause hyperpolarization.

The 5-$HT_{1A(Gi/Go)}$ Receptor

The most studied of the 5-HT_1 receptor family is the 5-HT_{1A} receptor, located on human chromosome HTR1/5q11.2–q13, which encodes 421 amino acids.

The receptor is found largely on neuronal soma and dendrites and epithelial cells, where it is found on both apical and basolateral membranes. Release of 5-HT in the CNS is under the control of a 5-HT_{1A} autoreceptor (and 5-HT_{1B} autoreceptors, see later). These autoreceptors fall into two categories: cell body autoreceptors and terminal autoreceptors. The former inhibit 5-HT release through inhibition of cell firing; the latter act through direct inhibition of release at the terminal. Cell body (or somatodendritic) autoreceptors belong to the 5-HT_{1A} receptor subtype in all species studied so far. *In vivo* immunocytochemical experiments have clearly demonstrated that, in striatal medium spiny neurons, the 5-HT_{1A} receptor is restricted to the somatodendritic level, while 5-HT_{1B} receptors are shipped exclusively toward axon terminals. In all systems examined to date, there appears to be a differential sorting of the 5-HT_{1A} and 5-HT_{1B} receptors. This observation is largely based on an *in vivo* transgenic system and appears to represent the only model that reconstitutes proper sorting of these receptors.

5-HT_{1A} receptors in the raphe nuclei act as somatodendritic autoreceptors and have little effect on extraneuronal 5-HT levels alone, but potentiate the increase seen after selective serotonin reuptake inhibitor (SSRI) administration. It is this effect that has been purported to be the mechanism of action of SSRIs, whereby desensitization of the 5-HT_{1A} autoreceptors may underlie the ability of chronic, but not acute, SSRI administration to increase synaptic cleft levels of 5-HT.

Evidence from *in vivo* studies indicates that activation of the postsynaptic 5-HT_{1A} receptor in the rat results in a characteristic 5-HT syndrome consisting of flat body posture, forepaw treading and head weaving, hypothermia, and adrenocorticotropic hormone (ACTH) release. Stimulation of postsynaptic 5-HT_{1A} receptors may also cause anxiogenic-like responses. In contrast, activation of presynaptic 5-HT_{1A} receptors induces both hyperphagia and anxiolytic-like effects in rats and hence may account for the clinical anxiolytic efficacy of the 5-HT_{1A} receptor agonists, buspirone and gepirone. 5-HT_{1A} receptor agonists are also active in animal models of depression, such as the rat forced swimming and mouse tail suspension tests, consistent with their purported antidepressant efficacy.

Several agonists show selectivity for the 5-HT_{1A} receptor, particularly *R*(+)-8-hydroxy-(di-*n*-propylamino) tetralin (8-OH-DPAT; Table 1), U92016A, and *R*(+)-UH301, which act as high-efficacy agonists in most systems. In contrast, the nonbenzodiazepine anxiolytics (buspirone and gepirone) and other ligands (such as MDL72832) have been shown to be partial agonists (low efficacy) at the 5-HT_{1A} receptor. Several apparent antagonists have also been characterized, such as NAN190, BMY7387, MDL 73005EF, WAY100135, *S*(−)-UH301, *S*(−)-pindolol, spiroxatrine, spiperone, SDZ216525, and NAD299 (robalzotan). However, all have demonstrated partial agonist properties in studies of somatodendritic autoreceptor function, perhaps due to the much larger receptor reserve associated with these, as opposed to postsynaptic receptors. To date, two selective high-affinity silent antagonists at this receptor have been identified; WAY100635 (pK_B 8.7; Table 1) and lecozotan (SRA-333; K_i 1.6 nM). The latter compound is currently in phase I studies for mild to moderate symptoms of Alzheimer's disease.

The 5-HT_{1A} receptor remains an enigma; however, continued great interest in this receptor persists, based on the ability of azapirone partial agonists (e.g., gepirone, isapirone, and zalospirone) to modify 5-HT function by an action on the 5-HT autoreceptor. 5-HT_{1A} agonists in this class include GR127935 and WAY163426, which are reported to

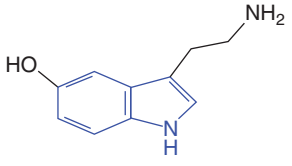

Figure 1 5-Hydroxytryptamine structure.

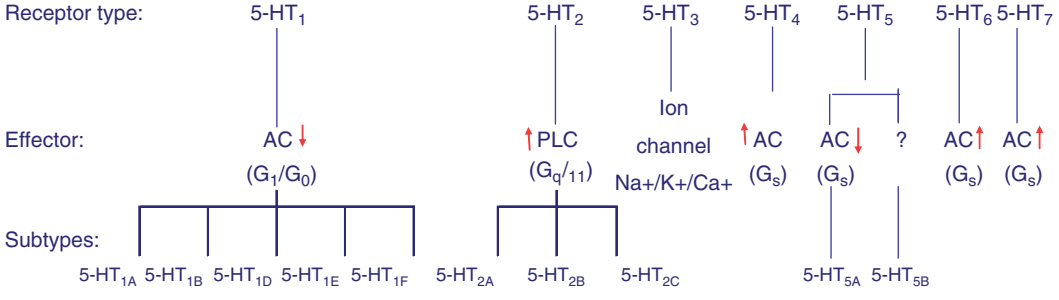

Figure 2 Classification of 5-HT receptor subtypes.

Table 1 5-HT_{1A}, 5-HT_{1B}, 5-HT_{1D}, and 5-ht_{1E}, receptors[a]

Feature	*Receptor*			
	5-HT_{1A}	*5-HT_{1B}*	*5-HT_{1D}*	*5-ht_{1E}*
Other names	–	5-$HT_{1D\beta}$	5-$HT_{1D\alpha}$	
Ensemble ID	ENSG00000178394	ENSG00000135321	ENSG00000179546	ENSG00000168830
Principal transduction	$G_{i/o}$	$G_{i/o}$	$G_{i/o}$	$G_{i/o}$
Selective agonists	8-OH-DPAT, (*R*)-UH301, U92016A	Sumatriptan, L694247, CP-93129, CP-122288	PNU10929, sumatriptan, L694247	
Selective antagonists (pK_B)	(±)WAY100635 (8.7), (*S*)-UH301, NAD299 (robalzotan, 9.2)	SB236057 (8.9), SB224289 (8.5), GR55562 (7.4)	SB714786 (pK_i 9.1), BRL15572 (7.9)	
Probes (K_D)	[^{3}H]WAY100635 (0.3 nM), [^{3}H]8-OH-DPAT, [^{11}C]WAY100635 (PET ligand)	[^{3}H]Sumatriptan, [^{125}I]GTI, [^{3}H]GR125743 (2.6 nM), [^{3}H]L694247	[^{3}H]Sumatriptan, [^{125}I]GTI, [^{3}H]GR125743 (2.8 nM), [^{3}H]L694247	[^{3}H]5-HT

[a]Overview (see also **Tables 2–5**) of 5-HT receptors. Modified nomenclature as agreed by the Nomenclature Committee of the International Union of Basic and Clinical Pharmacology (IUPHAR), subcommittee on 5-HT receptors; receptors and subsequently revised types are, with the exception of the ionotropic 5-HT_3 class, seven-transmembrane receptors, for which the endogenous agonist is 5-HT. The diversity of 5-HT receptors is increased by alternative splicing that produces isoforms of the 5-HT_{2A} (nonfunctional), 5-HT_{2C} (nonfunctional), 5-HT_4, and 5-HT_7 receptors. RNA editing produces 5-HT_{2C} receptor isoforms that differ in function, such as in efficiency and specificity of coupling to $G_{q/11}$.The 5-HT_3 receptor exists as a pentamer of 4TM subunits that form an intrinsic cation selective channel. Three 5-HT_3 receptor subunits (5-HT_{3A}, 5-HT_{3B}, and 5-HT_{3C}) have been cloned, but only homo-oligomeric assemblies of 5-HT_{3A} and hetero-oligometric assemblies of 5-HT_{3A} and 5-HT_{3B} subunits have been characterized in detail. Putative HTR3D and HTR3E genes have also been described but there is no evidence to date that they encode functional subunits. Also, a hetero-oligometric receptor that contains two copies of the 5-HT_{3A} subunit and three copies of the 5-HT_{3B} subunit has been described. The 5-HT_{3B} subunit has distinctive biophysical properties upon hetero-oligomeric versus homo-oligomeric (5-HT_{3A}) recombinant receptors but generally has little effect upon the apparent affinity of agonists and antagonists. However, homo- and hetero-oliogomeric 5-HT_3 receptors have been reported to differ in their allosteric regulation by some general anesthetic agents. The diversity of 5-HT_3 receptors is also further increased by alternative splicing of the 5-HT_{3A} subunit. Evidence to date, indicates that the 5-HT_{3A} subunit appears to be essential for 5-HT_3 operational characteristics.
Tabulated K_D values refer to binding to human 5-HT receptors, with the exception of SB207710 (piglet) and RP62203 (rat). The nomenclature of 5-HT_{1B}/5-HT_{1D} receptors has been revised. Only the nonrodent form of the receptor was previously called 5-HT_{1Db}: the human 5-HT_{1B} receptor (tabulated) displays a pharmacology different from that of the rodent forms of the receptor, due to Thr335 of the human sequence being replaced by Asn in rodent receptors. NAS-181 is a selective antagonist of the rodent 5-HT_{1B} receptor. Fananserin binds with high affinity to dopamine D4, in addition to 5-HT_{2A}receptors. The human 5-ht_{5A} receptor has been claimed to couple to several signal transduction pathways when stably expressed in C6 glioma cells. The human ortholog of the mouse 5-ht_{5B} receptor is nonfunctional due to interruption of the gene by stop codons. In addition to the receptors listed in the table, an 'orphan' receptor, unofficially termed 5-HT1P, has been described in the literature. Abbreviations here and for **Tables 2–4** (subtype-selective 5-HT compounds): **BIMU8**, (endo-*N*-8-methyl-8-azabicyclo[3.2.1]oct-3-yl)-2,3-dihydro-3-isopropyl-2-oxo-1*H*-benzimidazol-1-carboxamide hydrochloride; **BRL15572**, 3-[4-(3-chlorophenyl)piperazin-1-yl]-1,1,-diphenyl-2-propanol; **BW723C86**, 1-[5(2-thienylmethoxy)-1*H*-3-indolyl]propan-2-amine hydrochloride; **5-CT**, 5-carboxamidotryptamine; **8-OH-DPAT**, 8-hydroxy-2-(di-*n*-propylamino)tetralin; **EGIS-7625**, 1-benzyl-4-[(2-nitro-4-methyl-5-amino)-phenyl]-piperazine; **GR55562**, 3-[3-(dimethylamino)propyl]-4-hydroxy-*N*-[4-(4-pyridinyl)phenyl] benzamide; **GR113808**, [1–2[(methylsulfonyl)amino]ethyl]-4-piperidinyl]methyl-1-methyl-1*H*-indole-3-carboxylate; **GR125743**, *n*-[4-methoxy-3-(4-methyl-1-piperizinyl)phenyl]-3-methyl-4-(4-pyrindinyl)benzamide; **CP 93,129** ([3-(1,2,5,6-tetrahydropyr-id-4-yl)pyrrolo[3,2-b]pyrid-5-one]), **CP-122,288** (5-methyl-aminosulfonylmethyl-3-(N-methylpyrrolidin-2R-yl-methyl)-1H-indole), **GTI**, [^{125}I]5-hydroxytryptamine-5-O-carboxymethylglycyltyrosinamide, **GR65630**-(5-methyl-1*H*-imidazol-4-yl)-1-(1-methyl-*H*-indol-3-yl)-1-propanone, **LY278584**, 1 methyl-N-(8-methyl-8-azabicyclo[3.2.1]oct-3-yl)-1H-indazole-3-carboxamide, **TMB-8**, 8-(diethylamine)octyl-3,4,5-trimethoxybenzoate; **GTI**, 5-*O*-carboxamidomethylglycyl[^{125}I]tyrosinamide-tryptamine; **PNU109291**, (*S*)-3,4-dihydro-1-[2-[4-(4-methoxyphenyl)-1-piperazinyl] ethyl]-*N*-methyl-1*H*-2-benzopyran-6-carboximide; **Ro 60–0175**, (*S*)-2-(6-chloro-5-fluroindol-1-yl)-1-methyethylamine; **Ro 63–0563**, 4-amino-*N*-[2,6-bis(methylamino)pyridin-4-yl]benzenesulfonamide; **RP62203**, 2-[3-(4-(4-fluorophenyl)-piperazinyl)propyl]naphto[1,8-ca] isothiazole-1,1-dioxide; **RS127445**, (2-amino-4-(4-fluoronaphthyl-1-yl)-6-isopropylpyrimidine); **RS57639**, 4-amino-5-chloro-2-methoxybenzoic acid 1-(3-[2,3-dihydrobenzo[1,4]dioxin-6yl)-propyl]-piperidin-4yl methyl ester; **RS100235**, 1-(8-amino-7-chloro-1,4-benzodioxan-5-yl)-5-((3-(3,4-dimethoxyphenyl)prop-1-yl)piperidin-4-yl)propan-1-one; **RS102221**, 8-[5-(5-amino 2,4-dimethoxyphenyl) 5-oxopentyl]-1,3,8-triazaspiro[4,5]decane-2,4-dione; **SB204070**, 1-butyl-4-piperidinylmethyl-8-amino-7-chloro-1–4-benzoioxan-5-carboxylate; **SB207710**, 1-butyl-4-piperidinylmethyl-8-amino-7-iodo-1,4-benzodioxan-5-carboxylate; **SB224289**, 10-methyl-5[((20-methyl-40-)5-methyl-1,2,4-oxadiazol-3-yl)biphenyl-4-yl]carbonyl-2,3,6,-7-tetrahydrospiro[furo[2,3-f]indole-3,40-piperidine]oxalate; **SB236057**, 10-ethyl-5-(20-methyl-40-(5-methyl-1,3,4-oxadiazol-2-yl)biphenyl-4-carbonyl)-2,3,6,7-tetrahydrospiro[furo[2,3-f]indol3–3,40-piperidine; **SB242084**, 6-chloro-5-methyl-1-[2-(2-methylpyridyl-3-oxy)-pyrid-5-yl-carbamoyl] indoline; **SB258585**, 4-iodo-*N*-[4-methoxy-3-(4-methylpiperazin-1-yl)-phenyl]-benzenesulfonamide; **SB258719**, (*R*)-3,*N*-dimethyl-*N*-[1-methyl-3-(4-methylpiperidin-1-yl)propyl]benzene sulfonamide; **SB269970**, (*R*)-3-(2-(2-(4-methylpiperidin-1-yl)ethyl)pyrrolidine-1-sulfonyl)phenol; **SB271046**, 5-chloro-*N*-(4-methoxy-3-piperazin-1-yl-phenyl)-3-methyl-2-benzothiophenesulfonamide; **SB357134**, *N*-(2,5-dibromo-3-fluorophenyl)-4-methoxy-3-piperazin-1-yl-benzenesulfonamide; **SB656104**, 6-((*R*)-2-[2-[4-(4-chlorophenoxy)-piperidin-1-yl]-ethyl]-pyrrolidine-1-sulfonyl)-1*H*-indole hydrochloride; **SB699551**, 3-cyclopentyl-*N*-[2-(dimethylamino)ethyl]-*N*-[(40-{[(2-phenylethyl)amino]methyl}-4-biphenylyl)methyl]propanamide dihydrochloride; **SB714786**, 2-methyl-5-({2-[4-(8-quinolinylmethyl)-1-piperazinyl]ethyl}oxy)quinoline; **SR57227**, 4-amino-(6-chloro-2-pyridyl)-1-piperidine hydrochloride; **UH301**, 5-fluoro-8-hydroxy-2-(dipropylamino) tetralin; **U92016A**, ()-(*R*)-2-cyano-*N*,*N*-dipropyl-8-amino-6,7,8,9-tetrahydro-3*H*-benz[e]indole; **WAY100635**, *N*-(2-(4-(2-methoxyphenyl)-1-piperazinyl)ethyl)-*N*-(2-pyridyl)-cyclohexanecarboxamide trichloride.

Table 2 5-HT_{1F}, 5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C} receptors[a]

Feature	*Receptor*			
	5-HT_{1F}	*5-HT_{2A}*	*5-HT_{2B}*	*5-HT_{2C}*
Other names	5-$HT_{1E\beta}$, 5-HT_6	D, 5-HT_2	5-HT_{2F}	5-HT_{1C}
Ensemble ID	ENSG00000179097	ENSG00000102468	ENSG00000135914	ENSG00000147246
Principal transduction	$G_{i/o}$	$G_{q/11}$	$G_{q/11}$	$G_{q/11}$
Selective agonists	LY334370, LY334864	DOI	DOI, Ro 60-0175, BW723C86	DOI, Ro 60-0175
Selective antagonists (pK_B)		Ketanserin (8.5–9.5), M100907 (9.4)	RS127445 (9.5), EGIS-7625 (9.0)	SB242084 (9.0), RS102221 (8.6)
Probes (K_D)	[^{3}H]LY334370 (0.5 nM), [^{125}I] LSD	[^{3}H]Ketanserin (0.45 nM), [^{3}H]RP62203 (fananserin, 0.13 nM), [^{11}C]M100907 (PET ligand), [^{18}F]altanserin (PET ligand)	[^{3}H]5-HT	[^{3}H]Mesulergine (0.67), [^{3}H]LSD

[a]See **Table 1** footnote for overview and abbreviations.

Table 3 5-HT_3 receptors[a]

Nomenclature	5-HT_3
Other names	M
Ensemble ID	5-HT_{3A} ENSG00000166736, 5-HT_{3B} ENSG00000149305
Selective agonists	2-methyl-5-HT (5.3–5.5), 3-chlorophenyl-blguanide (5.4–5.7)
Selective antagonists (pIC_{DD})	granisetron (9.5), ondansetron (9.5), tropisetron (9.2)
Channel blockers	Diltiazem, TMB-B, picrotoxin
Radioligands (K_D)	[^{3}H]ramosetron (0.15 hM). [^{3}H]granisetron (1.2 nM), [3H]GR65630 (2.6 nM). [^{3}H]LY278584

[a]See **Table 1** footnote for overview and abbreviations.

Table 4 5-HT_4, 5-ht_{5A}, 5-ht_{5B}, and 5-HT_8, receptors[a]

Feature	*Receptor for*			
	5-HT_4	*5-ht_{5A}*	*5-ht_{5B}*	*5-HT_8*
Other names		5-$HT_{5\alpha}$		
Ensemble ID	ENSG00000164270	ENSG00000157219	ENSMUSG00000050534	ENSG00000158748
Principal transduction	G_s	G_i/G_o	None identified	G_s
Selective agonists	BIMU8, ML10302, RS67506			
Selective antagonists (pK_B)	GR113808 (9.0–9.5), SB204070 (10.8), RS100235 (11.2)	SB699551 (8.0)		SB271046 (8.7), SB357134 (7.6), Ro 63–0563 (7.9)
Radioligands (K_D)	[^{3}H]GR113808 (0.1 nM), [^{125}I] SB207710 (86 pM), [^{3}H]RS57639	[^{3}H]5-CT, [^{125}I]LSD	[^{3}H]5-CT, [^{125}I]LSD	[^{125}I]SB258585 (1.0 nM), [^{3}H]Ro 63–0563 (5 nM), [^{3}H]5-CT, [^{125}I]LSD

[a]See **Table 1** footnote for overview and abbreviations.

raise serotonin levels more rapidly than occurs with an SSRI, and which exhibit activity in chronic models of depression that are consistent with a more rapid onset. Two compounds, VPI-013 and PRX-00023, are also adding to the renewed interest in this area; although they have mixed 5-HT_{1A} receptor antagonist and sigma agonist properties, they are reported to be in phase II studies for depression.

Table 5 5-HT_7 receptor features[a]

Other names: 5-HT_X, 5-HT_1-like
Ensemble ID: ENSG00000148680
Principal transduction: G_s
Selective agonists: none
Selective antagonists (pK_B): SB656104 (8.5), SB269970 (8.5), SB258719 (7.2)
Radioligands (K_D): [^{3}H]SB269970 (1.2 nM), [^{3}H]5-CT, [^{125}I]LSD, [^{3}H]5-HT

[a]See Table 1 footnote for overview and abbreviations.

The 5-$HT_{1B(Gi/Go)}$ Receptor

5-HT_{1B} receptors are expressed throughout the mammalian central nervous system. These receptors are located on the axon terminals of both serotonergic and nonserotonergic neurons, where they act as inhibitory autoreceptors or heteroreceptors, respectively. 5-HT_{1B} receptors inhibit the release of a range of neurotransmitters, including serotonin, noradrenaline, γ-aminobutyric acid (GABA), acetylcholine, and glutamate. In rats and mice, the terminal autoreceptor is known to be a 5-HT_{1B} receptor, whereas in humans, pigs, rabbits, and guinea pigs, the terminal autoreceptor is thought to belong to the 5-HT_{1D} receptor subtype (see later). The 5-HT_{1B} receptor is located on human chromosome HTR1/6q13 and the receptor gene is largely concentrated in the basal ganglia, striatum, and frontal cortex. The receptor was originally defined by its pharmacology and also because of species differences in the binding affinity of key ligands, such as the β-adrenoceptor antagonist, cyanopindolol, which was thought to exist only in rodents. More recently, the amino acid sequence of the receptor has been characterized (humans, 390 amino acids; rodents, 386 amino acids) and found to be 93% identical overall and 96% identical with the transmembrane domains of the 5-$HT_{1D\beta}$ receptor, a close homolog found in higher species. This has a distribution and function similar to that of the 5HT_{1B} receptor. Indeed, the differences in the pharmacology of these two homologs are now attributed to the mutation of a single amino acid. Only the nonrodent form of the receptor was previously called 5-HT_{1D}: the human 5-HT_{1B} receptor displays a pharmacology different from that of the rodent forms of the receptor due to Thr335 of the human sequence being replaced by Asn in rodent receptors in the transmembrane-spanning region. In the rodent it is possible to differentiate between the pharmacology of the 5-HT_{1B} and 5-HT_{1D} receptors, as the 5-HT ortholog in these species exhibits a different pharmacology due to an Asp123/Arg123 switch in the rodent receptor. Thus, it has recently been agreed to classify the receptors as species homologs of the same receptor, termed h5-HT_{1B} (formerly 5-$HT_{1D\beta}$) and r5-HT_{1B}, with the h and r prefixes referring to the human and rat species, respectively.

Highly selective agonists and antagonists (Table 1) have now been identified to facilitate characterization of this receptor. There has been accumulating evidence, however, that 5-HT_{1B} receptors modulate drug reinforcement, stress sensitivity, mood, anxiety, and aggression. The general results of a number of studies and human genetics suggest that reduced 5-HT_{1B} heteroreceptor activity may increase impulsive behaviors, whereas reduced 5-HT_{1B} autoreceptor activity may have an antidepressant-like effect.

A large number of studies have established the 5-HT_{1B} terminal autoreceptor in controlling 5-HT release in rats, guinea pigs, and humans. The notion that a selective 5-HT_{1B} antagonist might prevent 5-HT negative feedback via this site, thereby increasing extraneuronal 5-HT and mimicking SSRI antidepressants, awaits clinical efficacy studies.

RU24969 was the first reported full agonist at the 5-HT_{1B} receptor, but, in binding studies, it is only fivefold selective over the 5-HT_{1A} and 5-HT_{1D} receptors. Since the first reports, several other compounds with dubious selectivity have emerged, at best having five-to tenfold selectivity in binding and functional assays (CGS 12066B, anpirtoline trifluoromethylphenylpiperasine (TFM), *m*-cholorophenylpiperazine (mCPP), MK 464, BW311C90, KSF99101H, and GR46611). The most potent agonist reported is L694247 (pK_D 10.0). None of these, however, differentiates between the 5-HT_{1B} and the 5-HT_{1D} receptor subtypes. More recently, selective 5-HT_{1B} receptor agonists (L694247) and antagonists (GR55562, pK_B 7.4; SB224289, pK_B 8.5; SB236657, pK_B8.9; SB236057, pK_B8.9) have been reported, generating further excitement around this receptor subtype and promising to shed light on the therapeutic utility of such agents.

The first potent and selective 5-$HT_{1B/1D}$ antagonist to be reported, GR127935 (pK_D 9.9), decreased extraneuronal 5-HT in the brain when administered systemically to guinea pigs. This may be due to its partial agonist activity at the 5-HT_{1B} receptor or to the presence of 5-HT_{1D} receptors on the raphe acting as somatodendritic autoreceptors, at which GR127935 also has agonist properties. This interpretation is supported by the failure of the recently characterized selective silent 5-HT_{1B} antagonist, SB224289 (pK_B 8.5, >60-fold selective over 5-HT_{1D} and other receptors; Table 1), to increase 5-HT release in the frontal cortex. NAS-181 has been reported to be the only selective antagonist of the rodent 5-HT_{1B} receptor and it may add to our knowledge of the operational characteristics of the rat receptor subtype.

Complementary studies in 5-HT_{1B} null mutant ('knockout') mice have proved unhelpful in understanding the pharmacology of the 5-HT_{1B} receptor, as it has been suggested that a compensatory adaptation to the constitutive loss of 5-HT_{1B} receptors becomes an important determinant of the altered response of 5-HT_{1B} knockout mice to a variety of pharmacological challenges. Altered expression and functions of serotonin 5-HT_{1A} and 5-HT_{1B} receptors in knockout mice lacking the 5-HT transporter (SERT knockout mice) have also been reported.

Although the effects on 5-HT release are unclear from 5-HT_{1B} knockout mice, *in vivo* studies have shown that stimulation of central postsynaptic 5-HT_{1B} receptors in mice, but not rats, causes hyperlocomotion, and penile erection and hypophagia in rats are also reportedly 5-HT_{1B} receptor mediated. Postsynaptic 5-HT_{1B} receptor activation in another species, the guinea pig, is reported to induce hypothermia, while the hypothermic response to 5-HT_{1B} agonists in the rat remains to be fully characterized. The putative 5-HT_{1B} receptor agonist, anpirtoline, has analgesic, cognitive, and antidepressant-like properties in rodents and it is of interest that mutant mice lacking the 5-HT_{1B} receptor are reported to be both highly aggressive and have an increased preference for alcohol.

Great interest in 5-HT_{1B} receptor agonists has been largely generated by the highly successful antimigraine drug sumatriptan, a nonselective 5-HT_{1D} and 5-HT_{1B} receptor agonist with low selectivity for other receptors in functional studies (Table 1). This compound may act either via constriction-mediating 5-HT_{1B} receptors on cerebral arteries or by blocking neurogenic inflammation and nociceptive activity with trigeminovascular afferents. This latter action has been argued to be 5-HT_{1B} receptor mediated, as protein extravasation induced by trigeminal ganglion stimulation is blocked by sumatriptan, the nonselective 5-HT_{1B} receptor agonist, by CP-93129 (>100-fold selectivity over 5-HT_{1A}, 5-HT_{1B}, and 5-HT_2 receptors; Tables 1 and 2), and by the close structural analog of sumatriptan, CP-122288, in wild-type, but not mutant, mice lacking the 5-HT_{1B} receptor. However, while 5-HT_{1B} receptor mRNA has been detected in rats, only 5-HT_{1D} mRNA has been detected in the guinea pig and human trigeminal ganglia. Thus, in humans, the antimigraine properties of sumatriptan may be either 5-HT_{1B} or 5-HT_{1D} receptor mediated. It is therefore abundantly clear that our understanding of the operational and physiological consequences of the 5-HT_{1B} activation across species is still largely in its infancy.

The 5-$HT_{1D(Gi/Go)}$ Receptor

The 5-HT_{1D} receptor has 63% overall structural homology with the 5-HT_{1B} receptor (formerly 5-$HT_{1D\beta}$) and a 77% amino acid sequence homology in the seven-transmembrane domains. The receptor is located on human chromosome HR1D/1p34.3–p36.3 and contains 377 and 374 amino acids for the human and rodent gene, respectively. Low levels of the 5-HT_{1D} receptor mRNA are found in the rat brain, predominantly in the caudate putamen, nucleus accumbens, hippocampus, and cortex, but also in the dorsal raphe and locus coeruleus.

It has been proposed that neurogenic inflammation and nociceptive activity within the trigeminovascular afferents may be 5-HT_{1D} receptor mediated due to the presence of 5-HT_{1D}, but not 5-HT_{1B}, receptor mRNA in the guinea pig and human trigeminal ganglia. Thus antagonism of plasma extravasation, induced by electrical stimulation of the trigeminal nerves, is observed with the antimigraine 5-$HT_{1B/1D}$ receptor agonists sumatriptan, naratriptan, rizatriptan, and zolmitriptan. Some advocates, however, claim that this is a 5-HT_{1D} receptor-mediated response. Several other pharmacological nonselective tool 5-$HT_{1B/1D}$ ligands have been used extensively to characterize the 5-HT_{1D} receptor. These include sumatriptan, PU109291, L694247 (pK_D 10.0), CP-122288, MK 464, and BW311C90, and the more brain penetrant agonists SKF 99101H, GR46611, and GR125743.

The location of 5-HT_{1D} receptor mRNA in the raphe suggests that, like the 5-HT_{1B} receptor, it too may function as a 5-HT autoreceptor. Indeed, there is electrophysiological, release, and voltammetric evidence to this effect. These data are further substantiated by the use of ketanserin and ritanserin, which, in addition their high affinity for the 5-HT_{2A} site, also have high affinity for the 5-HT_{1D}, but not for the 5-HT_{1B}, site (Table 1). The antagonist SB714786 (pK_D 9.1) and BRL-15572 (pK_D 7.9) 5-$HT_{1B/D}$ ligands mediate their effects at autoreceptors involved in the local inhibitory control of 5-HT release and may play a role in the pathogenesis of major depressive disorder (MDD) and in the antidepressant effects of the SSRIs in patients. A number of compounds in this class – including GR125743, SB224289, and L694247; AZD8129, a 5-HT_{1B} receptor antagonist; and CP-448187, a 5-$HT_{1B/1D}$ receptor antagonist – are in phase II studies for depression, and the outcome of these studies is eagerly awaited.

The 5-$HT_{1E(Gi/Go)}$ Receptor

The 5-HT_{1E} receptor was first characterized in humans as a [^{3}H]-5-HT binding site in the presence of 5-carboxyamidotryptamine (5-CT) that blocked binding to the 5-HT_{1A} and 5-HT_{1D} receptors. Human brain binding studies have reported that 5-HT_{1E} receptors (representing up to 60% of 5-HT_1 receptor binding) are concentrated in the caudate putamen,

with lower levels in the amygdala, frontal cortex, and globus pallidus. This is consistent with the observed distribution of 5-HT_{1E} mRNA. The receptor has been mapped to human chromosome HTR1E/6q14–q15, and the gene encodes a 365-amino-acid protein (Table 1). There are no reported selective or high-affinity ligands for this receptor (except for 5-HT) and its function is currently unknown. However, based on its distribution, one could argue that the amygdala is regarded as the 'emotional engine,' and dysfunction of the amygdala (and caudate putamen) has been associated with numerous psychiatric and neurological disorders, ranging from epilepsy to anxiety disorders, attention-deficit/hyperactivity disorder (ADHD), and social phobia, to Alzheimer's disease. Some schools of thought have even proposed that the amygdala is involved in sociopathic and criminal behaviors. Only time will tell, with the development of selective agents for the 5-HT_{1E} receptor.

The 5-$HT_{1F(Gi/Go)}$ Receptor

This receptor subtype is most closely related to the 5-HT_{1E} receptor, with 70% sequence homology across the 7TM domains. The gene is located on chromosome HTR1F/3p11, encoding 366 amino acids across human and rodent species. The mRNA coding for the receptor is concentrated in the dorsal raphe, hippocampus, and cortex of the rat and also in the striatum, thalamus, and hypothalamus of the mouse. 5-HT_{1F} receptor mRNA has been detected in human brain and is also present in the mesentery and uterus. Sumatriptan has almost equal affinity for the 5-HT_{1F} (pK_i 7.6) and 5-$HT_{1B/1D}$ receptors (pK_i 8.4 and 8.1, respectively; Table 1). Thus, it has been hypothesized that the 5-HT_{1F} receptor might be a target for drugs with antimigraine properties. 5-HT_{1F} mRNA has been detected in the trigeminal ganglia, the stimulation of which leads to plasma extravasation in the dura, a component of neurogenic inflammation that is thought to be a possible cause of migraine. The first 5-HT_{1F} receptor selective agonist, LY334370 (pK_D 9.4; see Table 2), with >100-fold separation over the 5-$HT_{1B/1D}$ receptors, has been claimed to block the effects of trigeminal nerve stimulation, as does sumatriptan, naratriptan, rizatriptan, and zolmitriptan . LY334370 has also been used successfully as a radioligand (K_D 0.5 nM) and demonstrated a reasonable correlation between the receptor protein and mRNA distribution, with the highest binding in the cortical areas, striatum, hippocampus, and olfactory bulb.

The 5-$HT_{2(Gq11)}$ Receptor Family

5-HT_2 receptors are characterized by having a relatively lower affinity for indolealkylamines, including 5-HT, and are linked to the G_q/phospholipase C pathway of signal transduction. The structural organization of the 5-HT_2 receptor has largely been determined by mutagenesis analyses that have identified a number of primary binding residues located mainly on TMs 3, 5, and 6, which are believed to constitute the purported binding pocket. Other residues identified on TMs 2, 3, 6, and 7 have been implicated in receptor activation and G-protein coupling. The 5-HT_2 receptor family mediates a large array of physiological and behavioral functions in humans via three distinct subtypes: 5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C}. While selective 5-HT_{2A} receptor antagonists have been known for some time, knowledge of the precise effects of 5-HT_{2C} and 5-HT_{2B} receptor antagonists has been hampered by the existence of only mixed receptor antagonists for 5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C}. However, selective 5-HT_{2A}, 5-HT_{2C}, and 5-HT_{2B} receptor antagonists have emerged over the past 5–10 years. Indeed, several structural classes belonging to the various pharmacophores have been reported.

The 5-HT_{2A} Receptor Subtype

The 5-HT_{2A} receptor subtype refers to the classical 'D' 5-HT_2 receptor as defined by the neuroleptic [^{3}H]spiperone in rat brain frontal cortex. Much of the early pharmacology of the 5-HT_{2A} receptor was pioneered using the selective ligand, [^{3}H]ketanserin, which made an important contribution to the characterization and localization of 5-HT_{2A} receptors. Since the early ligand binding studies, only a few compounds have shown over all selectivity for the 5-HT_{2A} site. Of these compounds, MDL100097 possesses subnanomolar affinity for the 5-HT_{2A} receptor (pK_B 9.4), with a minimum 300-fold selectivity for the 5-HT_{2A} site (pK_B 8.5–9.5) over other receptors examined, including the 5-HT_{2C} receptor. Other less selective compounds include ICI 169369, ICI 170809, RP62203 (fananserin), and SR46349B (20- to 30-fold selective). The recent introduction of more selective 5-HT_{2A} receptor antagonists (e.g., APD-125) will greatly accelerate the clinic potential of this receptor subtype. However, as yet, no selective agonists exist to provide a more precise classification of this receptor. Only α-methyl-5-HT has been shown to be a useful selective agonist for identifying 5-$HT_{2A/2C/2B}$ receptors (Table 2) Other (nonselective) compounds with agonist properties at 5-HT_2 receptors include 1-(2,5-dimethoxy-4-methylphenyl)-2-amino propane (DOM) and its 4-bromophenyl (DOB) and 4-iodophenyl (DOI) congeners. DOI is the preferred radioligand, albeit that it is relatively nonselective.

The actions elicited via 5-HT_{2A} receptors comprise mostly those previously considered to be mediated by postsynaptic 'D' receptor and include vascular

smooth muscle contraction in several preparations (e.g., rabbit aorta, rat caudal artery, dog gastrosplenic vein); contraction of bronchial, uterine, bladder, and some gastrointestinal smooth muscle; platelet aggregation; increased capillary permeability; and some behavioral syndromes in rodents, such as head twitch, forepaw treading, and coarse body tremors induced by 5-HT agonists. In humans, hallucinations and vivid dreams have been reported to be induced by 5-HT_{2A} receptor stimulation. The contribution of 5-HT neurotoxins in understanding the role of 5-HT_{2A} receptors and the mechanisms of action of hallucinogenic amphetamine analogs such as 3,4-methylenedioxymethamphetamine (MDMA) have implicated the 5-$HT_{2A/2C}$ receptors in psychiatric disorders. Several nonselective 5-HT_{2A}/5-HT_{2C} antagonists under investigation have failed to show any clinical utility for the treatment of several psychiatric disorders, including schizophrenia, anxiety, neurotic depression, migraine, sleep disorders, and drug abuse. However, a number of compounds have been shown to have therapeutic potential in cardiovascular disorders and migraine prophylaxis.

As considered earlier, the striking similarities between 5-HT_{2A} and 5-HT_{2C} receptors with respect to amino acid sequence (70% in 7TM domains) holds true for the signal transduction mechanism, as both receptors stimulate phosphatidylinositol (PI) turnover. However, the receptors have distinct chromosomal localizations (HTR2A/13q14–q21 and HTR2C/Xq24). The characterization of the 5-HT_{2A} receptor-coupled signal transduction system was first studied in human platelets. It is likely that the biochemical events mediated by the 5-HT_{2A} receptor are important for both the shape change and the aggregation of human platelets. Controversy still exists, but a few studies have reported 5-HT-induced platelet aggregation in patients with peripheral cerebrovascular disease. Peripheral arterial vasospasms subsequent to platelet activation appear to be mediated in part by 5-HT and can be effectively reduced by ketanserin. In animal studies, postthrombotic peripheral collateral circulation can be significantly restored by treatment with ketanserin (pK_D 8.5–9.5) and MDL100907 (pK_D 9.4), and 5-HT-induced reduction of blood supply through the collateral system can be effectively counteracted by pretreatment with ketanserin. In *ex vivo* human platelet studies, the 5-HT_{2A}/5-HT_{2C}/5-HT_{2B} receptor antagonist ICI 170809 significantly inhibited 5-HT-induced aggregation (minimum effective dose, 0.1 mg kg^{-1} per os), thus lending support for the potential therapeutic role of 5-HT_{2A} receptor antagonists in modulating the vasospastic effects of 5-HT released from platelets during vascular trauma. The clinical significance of selective 5-HT_{2A} (see also discussions of 5-HT_{2B} and 5-HT_{2C}) receptor antagonists in this vascular disorder and various psychiatric disorders still awaits further conclusive evidence of clinical efficacy with selective agents. In psychiatric disorder studies, two recent 5-HT_{2A} receptor antagonists have shown significant clinical efficacy in schizophrenia. Pimavanserin (ACP-103) significantly improved efficacy scores (positive and negative syndrome scale; PANSS), showing a faster onset and improved side-effect profile, following co-therapy with risperidone. Fananserin, which binds with high affinity to dopamine D_4, in addition to 5-HT_{2A} receptors, also showed good clinically efficacy in schizophrenia, with an improved side-effect profile. This therapeutic opportunity is currently been actively pursued by several pharmaceutical companies, and several compounds acting at 5-HT_{2A} and dopamine D_2 receptors, are at various stages of clinical development (e.g., asenapine, blonanserin, eplivanserin, pruvanserin, paliperidone, uoperidone, and vabicaserin). Another selective 5-HT_{2A} receptor antagonist, APD-125, is currently being assessed in the clinic for insomnia and sleep maintenance.

The 5-$HT_{2B(Gq11)}$ Receptor

The 5-HT_{2B} mRNA transcript has been now been identified in human and nonvertebrates and the receptor has been mapped to human chromosome HTR2B/2q36.3–2q37.1, with the human gene encoding 481 amino acids (Table 2). It was first identified in the stomach of the rat and in the small intestine, kidney, heart, and cerebellum in the mouse. The presence of the transcript in the rat stomach and the pharmacology of the cloned 5-HT_{2B} receptor have led ultimately to defining the rat stomach fundus receptor. In Gaddum and Picarelli's first classical attempt at subdivision of 5-HT receptors, the D subtype (5-HT_{2A}) preceded the first descriptions of the rat stomach fundus 5-HT receptor by only a few pages in the same issue of the *British Journal of Pharmacology*, whereas the 5-HT_{2C} receptor was not identified until some 28 years later.

The 5-HT_{2B} receptor represents the second member of the 5-HT_2 family, but it is still unclear how the clone relates pharmacologically to that 5-HT receptor which mediates contraction of the rat stomach fundus. Evidence indicates that the 5-HT_{2B} contractile receptor in the rat stomach fundus is coupled to calcium influx through voltage-dependent calcium channels, intracellular calcium release, and activation of protein kinase C (PKC). These actions may reflect a novel coupling mechanism unrelated to increases in PI hydrolysis, unlike the other 5-HT_2 receptors.

In the 1990s a series of potent, selective 5-HT_{2B} receptor antagonists were identified, of which

SB204741 (pK_i 7.8) has proved to be an important tool for characterizing this receptor, along with the less selective 5-HT_{2B} receptor agonists BW723C86 and α-Me-5-HT. Recent additions to the 5-HT_{2B} tool-kit have been the introduction of two highly selective 5-HT_{2B} receptor antagonists, RS127445 (pK_i 9.5) and EGIS-7625 (pK_i 9.0).

The clinical potential of compounds selective for the 5-HT_{2B} receptor site has pointed to an alternative target for initiation of migraine and may account for the continued use of methysergide, pizotifen, and cyproheptadine for migraine prophylaxis. More selective compounds have been developed for this indication, but to date no selective 5-HT_{2B} compound has shown clinical efficacy for migraine. Moreover, the only 5-HT_{2B} compound in current clinical development for the treatment of pulmonary hypertension is PRX-08066, which is surprising, considering the critical importance of the 5-HT_{2B} receptor in mammalian heart development.

The 5-$HT_{2C(Gq11)}$ Receptor

The human 5-HT_{2C} receptor gene HTR2C/Xq24 encodes 377 amino acid residues with seven hydrophobic domains of 20–30 residues (Table 2). The rat and mouse sequences suggest the presence of an eighth transmembrane region not seen in the human receptor. The rat and human sequences are otherwise very similar, with a 90% overall homology. They are also positively coupled to phosphoinositide hydrolysis, providing a common secondary messenger system for 5-HT_{2C} and 5-HT_{2A} receptors. It is well established that 5-HT_{2C} receptors display a high degree of constitutive activity, and consequently inverse agonists have a large effect when they bind to the 5-HT_{2C} receptor. Furthermore, the mRNA for the 5-HT_{2C} receptor undergoes mRNA editing, which results in the expression of multiple 5-HT_{2C} receptor isoforms. Correspondingly, INI (unedited) and VSV (a fully edited version) isoforms are abundant in rat brain. The VSV isoform lacks the high-affinity recognition site for 5-HT, which may be caused by low-efficiency coupling to G-proteins. These edited isoforms appear to have reduced constitutive activity and reduced capacity to couple to G_{q11} signaling proteins. This suggests that the INI isoform of the 5-HT_{2C} receptor is pharmacologically similar to the VSV form of the 5-HT_{2C} receptor, but that it couples more efficiently to G-proteins. Importantly, the capacity of some cells in the brain to edit the 5-HT_{2C} mRNA is reduced. It has been suggested that this may well be the underlying pathophysiology in schizophrenic patients and that this reduced capacity is evident in brains of patients who had committed suicide. Changes in the RNA-editing capacity of the brain will result in altered patterns of expression of the edited 5-HT_{2C} isoforms and perhaps in altered responses to drugs. Consequently, it is important to understand the signaling characteristics and mechanism of the edited 5-HT_{2C} receptor isoforms in various psychiatric and neurological conditions, and selective agents are allowing this work to progress.

The pharmacological characteristics of the 5-HT_{2A} and 5-HT_{2C} (formally named 5-HT_{1C}) receptors are remarkably similar, with many of the 'classical' 5-HT receptor antagonists showing high affinity for both sites. Furthermore, some compounds may have different affinities for human as opposed to rat receptors. The first 5-HT_{2C} selective antagonist SB200646A (50-fold) was reported in the early 1990s; however, few truly selective 5-HT_{2C} receptor agonists have been identified. MK 212 (30-fold) is the most selective CNS active agonist, but is also a 5-HT_{2A} agonist, while α-methyl-5-HT has equal affinity for all 5-HT_2 receptor subtypes. Of the various 5-HT_2 receptor agonists, (+)-DOB and *m*-chlorophenylpiperazine (mCPP) possess approximately tenfold selectivity over other binding sites in the rat but not in humans. However, mCPP is of particular interest because of its use as a clinical probe in the 1990s to study human 5-HT_{2C} receptor function.

mCPP is commonly observed to induce anxiety and panic attacks in humans. This mCPP-induced anxiogenesis is inhibited by the 5-HT_{2C} and 5-HT_{2A} receptor antagonist ritanserin, which is consistent with 5-HT_{2C} receptor mediation. If 5-HT_{2C} receptor activation is anxiogenic, the blockade of this receptor might induce anxiolysis, provided the receptors are tonically innervated. This argument is supported by the anxiolytic properties of SB243213, a selective 5-HT_{2C} antagonist, in rat models of anxiety with quite different motivational and aversive bases (the rat social interaction and Geller–Seifter tests). To date this hypothesis awaits clinical testing.

In support of this hypothesis, the selective 5-HT reuptake inhibitors (SSRIs) paroxetine, fluoxetine, clomipramine, sertraline, and citalopram, on chronic administration, have all been observed to desensitize behavioral responses to mCPP. The monoamine oxidase inhibitors phenelzine and nialamide, which also raise extraneuronal 5-HT levels, have a similar effect after chronic treatment. These results suggest that SSRIs and MAO inhibition antidepressant treatments may desensitize the 5-HT_{2C} receptor, although there is no direct evidence from receptor binding, mRNA, or secondary messenger systems as yet. Furthermore, unlike other SSRIs, fluoxetine and its metabolite norfluoxetine have weak affinity for the 5-HT_{2C} receptor and are therefore likely to accumulate and directly block 5-HT_{2C} sites.

Paradoxically, there is an emerging rationale and literature for the use of selective 5-HT_{2C} receptor agonists in psychiatric and eating disorders that has been driven by recently reported compounds that have shown greater selectivity for the 5-HT_{2C} receptor than has mCPP. Early studies with mCPP pointed the way to far more selective agonists in eating and psychiatric disorders. Several reports suggest that the 5-HT_{2C} receptor agonists VER-2692 and WAY16909 are potent and selective agonists, and these are currently being evaluated as antiobesity agents.

Early clinical studies with mCPP in bulimic patients induced migraine-like headaches 8–12 h later. This response was correlated with plasma levels of mCPP and was more pronounced in patients with a personal or family history of migraine. Conversely, the nonselective 5-HT_{2C}/5-HT_{2A} receptor antagonists pizotifen, cyproheptadine, and methysergide are all clinically effective as migraine prophylactics. ICI 169369 and sergolexole, which have high affinity for 5-HT_{2C} and 5-HT_{2A} receptors, but little affinity for other sites, have subsequently been found to have modest antimigraine efficacy in an early trial. As these drugs only share a high affinity for the 5-HT_{2C} and 5-HT_{2A} receptors, and as the selective 5-HT_{2A} receptor antagonist ketanserin is ineffective as a migraine prophylactic, it was proposed that blockade of 5-HT_{2C} receptors mediated this effect (although this has recently been challenged on the basis of their affinity for the 5-HT_{2B} receptor). Furthermore, downregulation of 5-HT_{2C} receptors may mediate the antimigraine properties of the antidepressants amitriptyline and fluoxetine; alternatively, the latter compounds may act by direct antagonism of vascular 5-HT_{2C} receptor sites. Other possible therapeutic targets for drugs acting at 5-HT_{2C} receptors include eating disorders, sexual dysfunction, and head trauma, based on the actions of nonselective 5-$HT_{2C/2A}$ agonists and antagonists.

It is highly likely that the 5-HT_2 family will continue to grow. There is already considerable evidence for a number of other 'orphan' receptors which may be adopted into the family. Several orphans are thought to exist, one of which is the endothelial 5-HT receptor, which is present in rabbit and rat jugular vein and pig pulmonary artery and vena cava. The rat jugular vein is now known to represent a 5-HT_{2B} receptor. More recently, a site identified in choroid plexus has been shown to bind the 5-HT_2-selective ligand RP62203 with high affinity, but does not possess identity with any of the three established 5-HT_2 receptors. However, it would appear that the endothelial receptors do not represent a pharmacologically homogeneous class. While possessing many of the characteristics of 5-HT_{2C} and 5-HT_{2B} receptors, subtle yet robust differences render any explicit affiliation impossible. These problems are confounded by the need, in many instances, to make comparisons across species. The potential for substantial species differences in pharmacology is highlighted on pharmacological comparison of mouse and rat homologs of the cloned 5-HT_{2B} receptor. Thus the eventual classification of the orphan endothelial 5-HT receptor awaits robust interspecies comparison of 5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C} receptors. In this regard, in rat jugular vein, the endothelial receptor, which was previously classified as 5-HT_{2C}-like, has now been shown to be more like 5-HT_{2B} receptors, using ligands which discriminate the rodent (rat) 5-HT_2 receptor subtypes.

Another therapeutic area that has gained prominence recently arose from observations in humans that mCPP reduces total sleep time, sleep efficiency, slow-wave sleep (SWS), and rapid eye movement sleep (REMS). The actions of 5-$HT_{2A/2C}$ receptor antagonists on sleep architecture have since been widely studied. Thus ritanserin has been reported to increase SWS, reduce sleep onset latency, and improve subjective sleep quality in volunteers, while REMS is reduced in some, but not all, studies. Ritanserin has also been observed to be of therapeutic use in insomniacs and in patients suffering from dysthymia. These effects were maintained with chronic treatment. Other nonselective 5-HT_{2A}/5-HT_{2C} receptor antagonists such as mianserin, cyproheptadine, and pizotifen have similar effects. In rats, ritanserin generally increases SWS and reduces REMS, although some studies report that the deepest phase of SWS (SWS2) is increased but total SWS is unaffected. Studies with three other 5-HT_{2A}/5-HT_{2C} receptor antagonists, ICI 169369 and ICI 170809 and SR46349B, report reduced REMS, but little effect on undifferentiated SWS. These results, therefore, suggest that a 5-HT_{2C} receptor antagonist may improve sleep quality (see also agomelatine).

Of great interest to pharmaceutical companies is the role of the 5-HT_{2C} receptor in anxiety and depression. Two 5-HT_{2C} receptor inverse agonists, SB243213 and WAY163909, have been reported to be in development for the treatment of depression. The fact that both compounds show a similar pharmacology in several animal models of depression reflects the ability of the 5-HT_{2C} receptor agonists and antagonists/inverse agonists to desensitize the 5-HT_{2C} receptor, as observed following chronic SSRI treatment. Another drug that has recently gained European regulatory approval is agomelatine, which is a melatonergic, MT1 and MT2 receptor agonist/5-HT_{2C} antagonist with

antidepressant properties and which may be acting through a similar mechanism. It is reported to improve disrupted sleep patterns in depressed patients (which also is likely, due to its 5-HT_{2C} receptor antagonist properties), without affecting daytime vigilance. The reported lack of effect on sexual function, tolerability problems, or discontinuation symptoms offers several advantages over current therapy for the treatment of depression.

It is clearly evident that therapeutic opportunities are still to be realized in the 5-HT_2 receptor family and will undoubtedly follow with clinical evaluation of selective, nonselective, and combination agents.

The Ligand-Gated 5-HT_3 Receptor

The original naming of the 'M' receptor by Gaddum and Picarelli was based on the ability of morphine to antagonize the serotonin receptor which mediates depolarization of cholinergic nerves in the guinea pig ileum. Morphine was the first pharmacological tool to help characterize this receptor, now referred to as the 5-HT_3 receptor. Since the time of these early observations, 5-HT_3 receptors have been reported to be present on postganglionic autonomic neurons in the peripheral sympathetic and parasympathetic nervous systems, on enteric neurons, and on sensory neurons in various tissues.

The 5-HT_3 receptor is a member of the Cys-Cys loop ligand-gated ion channel superfamily and as such is composed of multiple subunits, containing 478 amino acids. Five human genes have been identified on chromosome HTR3/11, which encodes the 5-HT_3 subunit (5-$HT_{3A–E}$) (Table 3). To date, conflicting reports as to the existence of central 5-HT_{3B} subunits have been reported. However, immunohistochemical studies using sections of human temporal lobe demonstrated both 5-HT_{3A} and 5-HT_{3B} immunoreactivity associated with pyramidal neurons within all CA fields of the hippocampus, as well as with large neurons within the hilus. The expression of both h5-HT_{3A} and h5-HT_{3B} subunit mRNAs in human hippocampus would therefore support the presence of both homomeric h5-HT_{3A} and heteromeric h-5-$HT_{3A/3B}$ receptors in human hippocampus.

The pathophysiological relevance of central h5-HT_{3A} and h5-HT_{3B} receptors remains to be fully evaluated with selective receptor antagonists in preclinical and clinical studies. The pioneering animal studies in the 1980s with the nonselective antagonist ondansetron clearly point to a role for 5-HT_3 receptors in anxiety, depression, and schizophrenia. However, this was not substantiated in human studies with several 5-HT_3 receptor antagonists, including, ondansetron, granisetron, BRL46470, and dolasetron.

5-HT is found in both the brain and the gut, but it is now widely understood that 95% of the serotonin in the body resides in the gut. The two leading 5-HT_3 receptor antagonists, ondansetron and granisetron, have proved to be highly successful antiemetic agents, and they are widely used in the prophylactic treatment of chemotherapy-induced nausea and vomiting. This discovery arose from carefully reasoned experimental research. It was shown in animal studies that chemotherapeutic agents such as cisplatin released 5-HT in the gut in vast quantities, activating 5-HT_3 receptors in this region. Activation of 5-HT_3 receptors resulted in the firing of a vagally mediated response resulting in emesis. The discovery of the antiemetic properties of these 5-HT_3 receptor antagonists, along with several others that followed, opened the way for this class of compounds to become the gold standard in the clinical treatment of chemotherapy-induced nausea and emesis.

A second non-CNS therapeutic target that has gained prominence in recent years has been irritable bowel syndrome (IBS) and its association with the 5-HT_3 receptor and 5-HT_4 receptor. 5-HT_3 receptor antagonists were predicted to be effective in diarrhea-predominant IBS based on animal studies. Clinical studies with ondansetron and granisetron showed a reversal in rectal hypersensitivity in IBS, and the longer acting agent alosetron showed significant clinical improvement in IBS symptoms. However, alosetron has had a checkered clinical existence following early withdrawal after reports of rare serious gastrointestinal adverse events among patients taking the drug. Similar drugs (i.e., cilansetron, ramosetron, and DDP-733) are also in clinical studies for IBS, with the hope that these side effects are not class related. The latter compound is a partial agonist that has recently been reported to be active in a phase II IBS study.

The complex symptomatology in IBS may be attributed to activation of other 5-HT receptor subtypes in the gastrointestinal tract. It was the identification of 5-HT_4 receptors that led many pharmaceutical companies to believe that agents acting as 5-HT_4 receptors would be useful in the treatment of this condition (see discussion of the 5-HT_4 receptor and Table 4).

The 5-$HT_{4(Gs)}$ Receptor

The 5-HT_4 receptor was originally identified in primary cell cultures of mouse embryo colliculli, where later it was found to have a broad tissue distribution and to be positively coupled to adenylate cyclase. In early functional assays studies the 5-HT_4 receptor was identified in the rat esophageal muscularis mucosae and guinea pig colon, where activation resulted in

relaxation and contraction, respectively. The cloned human and rodent receptor was shown to consist of 388 and 406 amino acids, respectively. Various isoforms of the 5-HT_4 receptor have now been identified (5-$HT_{4b,c}$ and 5-HT_{4d}), encoded by a gene located on chromosome HTR4/5q31–q33 (Table 4).

In the brain, 5-HT_4 receptors appear to be located on neurons, whereas in the peripheral nervous system they are reported to facilitate the release of acetylcholine in smooth muscle of the intestinal tract. It was these observations on the gastrointestinal tract that led to the notion that 5-HT_4 receptors may have a role in the pathophysiology of irritable bowel syndrome and gastroesophageal reflux disease (GERD).

The complex symptomatology and pharmacology of IBS led to the development of several highly selective 5-HT_4 receptor agonists and antagonists that are now established in clinical practice for this indication. Tegaserod (HTF-919) and purcalopride (R93877) are 5-HT_4 receptor partial agonists, and piboserod (SB207266) is an antagonist. All of these medications are reported to relieve the symptoms of constipation, the notion being that the 5-$HT_{3/4}$ agonists are intended to have prokinetic activity, like the mixed 5-HT_4 receptor agonist/5-HT_3 receptor agonist, cisapride. Such agents would therefore be beneficial for constipation-predominant IBS. Conversely, the antagonist piboserod would be indicated for diarrhea-predominant IBS. However, the therapeutic potential of piboserod was noted to be limited to a decrease in rectal sensitivity, and the compound was not advanced in further clinical studies. The disappointment with this class of drugs has been their lack of efficacy in IBS-related pain. However, other 5-HT_3 agonists (pumosetrag) and 5-HT_4 agonists (TD-2749 and TD-5108) are also being evaluated in IBS, which may address this important, unmet medical need.

The potential CNS efficacy of 5-HT_4 receptor antagonists is less well developed, though the 5-HT_4 partial agonist SL65.0155 (5-(8-amino-7-chloro-2,3dihydro-1,4-benzodioxin-5yl)-3-[1-(2-phenylethyl)-4-piperidinyl]1,3,4-oxadiazol-2(3*H*)-one-monohydro chloride) has been reported to show cognitive enhancing effects in animal models. The high density of 5-HT_4 receptors in the nucleus accumbens has suggested that these receptors may be involved in reward mechanisms and may influence self-administration. Recent studies with GR113808 and its effects on reducing alcohol intake in rats would concur with this suggestion.

The clinical benefit of selective agents for the 5-HT_4 receptors in medical practice and the promise of more developments clearly represent one of many success stories of 5-HT drug development.

The 5-ht_5 Receptor (G-Protein Coupling, None Identified)

The 5-HT_5 receptor is highly localized in the rodent CNS; its distribution in the hippocampus (CA1, CA3, dentate gyrus), cortex, cerebellum (granular layer), olfactory bulb, habenula, and spinal cord would suggest a role in the pathophysiology of psychiatric and neurological disorders. A human 5-HT_{5A}, but not a 5-ht_{5B}, receptor has been identified; the human 5-HT_{5A} receptor gene, like the rodent gene, contains two codon exons separated by a single large intron. In the case of the human and rodent 5-HT_{5A} receptor, the gene contains the same number of amino acids (357); similarly, the rodent $5ht_{5b}$ gene contains 387. To elucidate the pathophysiology of the 5-HT_5 gene, chromosomal localization studies have identified the receptor on HTR5A/7q36, whereas the $5ht_{5b}$ gene is localized on htr5b/2q11–q13. The signal transduction mechanism for the 5-HT_5 receptor is still a matter for conjecture, as some suggest it may utilize a novel (perhaps ion channel) second-messenger system (Table 4).

The physiological function of the 5-HT_{5A} receptor is poorly understood. 5-HT_{5A} null mutant mouse studies suggest a role in exploratory behavior, but the lack of selective pharmacological tools has hampered progress. Localization studies have revealed that 5-HT_{5A} receptors have widespread distribution in the CNS. It has been speculated that, on the basis of their localization, the receptors may be involved in motor control, learning and memory consolidation, and anxiety and depression.

Recently a selective 5-HT_{5A} receptor antagonist, SB699551A, has been identified. This compound has been reported to attenuate 5-carboxyamidotryptamine-induced inhibition of raphe neuronal cell firing *in vitro* and to increase 5-HT levels in prefrontal cortex *in vivo*, suggesting a role for the 5-HT_{5A} receptor in the modulation of raphe 5-HT neuronal activity. This is another pharmacological opportunity whereby 'tool' compounds will expose the receptor subtype operational and functional characteristics in native tissue, as it has been suggested that many of the operational characteristics of the 5-HT_{1D} receptor may be attributed to the 5-HT_{5A} receptor.

The 5-$HT_{6(Gs)}$ Receptor

5-HT_6 receptors have been identified in areas of the rat and human brain associated with learning and memory: hippocampus, CA1, CA3, dentate gyrus, olfactory tubercles, cerebral cortex, nucleus accumbens, and striatum. The 5-HT_6 gene is localized on chromosome HTR6/1p35–p36, and studies to

determine its linkage to several CNS disorders (memory dysfunction and schizophrenia) are actively being pursued (Table 4). However, the potential of 5-HT_6 receptor antagonists in the treatment of cognitive disorders has been an actively researched topic in recent years, and the availability of numerous tool compounds for this receptor, some of which are currently in clinical trial, will advance our understanding of the role of this receptor in cognitive process.

The notion that several atypical antipsychotic agents such as clozapine, quetiapine, and olanzapine possess high affinity for the 5-HT_6 receptor adds translational creditability to this approach. The leading pharmacophores are indole-related compounds, of which several have emerged, including the recently described aryl sulfonamide indole chemotype, SB271046 (pK_i 8.6), with excellent CNS bioavailability and *in vivo* efficacy, compared to several earlier compounds.

More recently, 5-HT_6 receptors have been demonstrated to regulate central cholinergic neurotransmission. In this regard, administration of the 5-HT_6 receptor-selective antagonist Ro 04–6790 reversed scopolamine-induced rotation in 6-hydroxydopamine-lesioned rats. Additionally, 5-HT_6 receptor antisense oligonucleotides or 5-HT_6 receptor-selective inhibitors enhanced retention by rats of the learned platform position in the Morris water maze. These data suggest that 5-HT_6 receptor antagonists might boost cholinergic neurotransmission and reduce the cognitive impairments experienced by patients with dementia or schizophrenia. Intriguingly, it has recently been determined that the 267C allele of the 5-HT_6 receptor is a significant risk factor for Alzheimer's disease. Taken together, these findings indicate that 5-HT_6 antagonists might prove useful in treating a number of common illnesses, including dementia and schizophrenia.

A major concern in drug discovery is that the cloned mouse receptor is significantly different in rat and human 5-HT_6 receptors. In addition to species differences in the binding of drugs to 5-HT_6 receptors, differences in the regional expression of 5-HT_6 receptors are also apparent. Furthermore, quantitative polymerase chain reaction studies have demonstrated that the mouse 5-HT_6 receptor mRNA is at least tenfold less abundant than are the rat and human 5-HT_6 receptor mRNAs, in every brain region examined. Surprisingly, whereas 5-HT_6 receptor mRNA and radioligand binding activity is enriched in the basal ganglia of rat and human brain, there is no such enrichment in the mouse brain.

Finally, using a combination of site-directed mutagenesis and molecular modeling studies, it has been shown that the peculiar mouse 5-HT_6 pharmacology is attributable to two amino acids – Tyr188 (in helix 5, which is Phe188 in rats and humans) and Ser290 (in helix 6, which is Asn290 in rats and humans) – and these account for the bulk of the differences in pharmacology. These findings are a cautionary note but may have important implications for drug discovery.

The 5-$HT_{7(Gs)}$ Receptor

Since the identification of the 5-HT_7 receptor in the early 1990s, evidence has accumulated supporting a role for 5-HT_7 receptors in various physiological functions, including sleep disorders, circadian rhythms, cognition, mood, and sleep (see Table 5). A number of splice variants of the 5-HT_7 receptor have been identified and exhibit overlapping mRNA distributions, and all couple to G_s, though no differences in operational characteristics have been shown, which may have led to differentiation of behavioral phenotypes. Work is still ongoing to elucidate the pathophysiology of these splice variants, and chromosomal localization studies have identified the receptor on HTR5A/7q36, encoding 445 and 488 amino acids for the human and rodent receptor, respectively (Table 5). The finding that several drugs used in clinical psychiatric practice had high affinity for 5-HT_7 receptor binding sites in the brain initiated an ongoing interest in evaluating the involvement of neuropsychiatric disorders. The distribution of 5-HT_7 receptors in areas of the brain associated with neuropsychiatric disorders – hippocampus (CA_1, CA_2), hypothalamus, thalamus, raphe nuclei, and superior colliculus – supported this notion. In concordance with its distribution in the hippocampus, the selective 5-HT_7 receptor antagonist SB269700-Z was shown to attenuate phencyclidine (PCP)-induced cognition dysfunction associated with schizophrenia.

The strongest case has been made for the importance of the 5-HT_7 receptor in depression. Inactivation or blockade of the receptor leads to an 'antidepressant' state in behavioral models of depression. Furthermore, 5-HT_7 receptor downregulation following chronic administration of antidepressants in animal studies points to a possible molecular mechanism of action. This hypothesis is supported by recent animal experiments using 5-HT_7 receptor knockout mice that clearly show an antidepressant-like profile in the rat forced swim test and mouse tail suspension test. Selective antagonist tool compounds SB258717, SB269970, and SB656104-A have all been shown to exhibit antidepressant-like properties in animal studies and to modulate REM sleep in rats, results that concur with this line of investigation. Several lines

of evidence have shown that 5-HT-induced hypothermia is mediated by the 5-HT_7 receptor, as defined with selective antagonists and in 5-HT_7 knockout mice studies. What is interesting, however, is recent work with the endogenous fatty acid oleamide that suggests it may act through an independent mechanism as well as at an allosteric 5-HT_7 receptor site to regulate body temperature and perhaps circadian rhythms. Clinical progress in this area has been limited by suboptimal clinical candidates, and clinical efficacy data for selective agents are still awaited.

The Future of 5-HT Research

It is now over 50 years since Gaddum first suggested "that a drug with a specific antagonistic action to 5-HT – might be used in therapeutics." Clearly, drugs acting at the 5-HT receptor subtypes have revolutionized therapeutics during the intervening years. However, it is only now that selective agents are appearing to differentiate many operational and physiological characteristics of 5-HT actions on the various receptor subtypes currently identified. The future therapeutic potential of drugs acting at these 5-HT subtypes and isoforms or in combination with other neurotransmitter receptors will open up next generation of novel serotonergic therapeutic agents.

See also: Cells: 5-Hydroxytryptamine Receptors; Depression and the Brain; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin-Related Psychedelic Drugs; Stress Response: Genetic Consequences.

Further Reading

Alexander SPH, Mathie A, and Peters JA (2008) Guide to receptors and channels. *British Journal of Pharmacology* 153 http://www.nature.com/bjp/journal/vgrac/ncurrent/index.html (accessed Jul. 2008).

Barnes NM and Sharpe T (1999) A review of central 5-HT receptors and their function. *Neuropharmacology* 38: 1083–1152.

Bonasera SJ and Tecott LH (2000) Mouse models of serotonin receptor function: Toward a genetic dissection of serotonin systems. *Pharmacology & Therapeutics* 88: 133–142.

Branchek TA and Blackburn TP (2000) 5-HT_6 receptors as emerging targets for drug discovery. *Annual Review of Pharmacology and Toxicology* 40: 319–334.

Brockaert J, Claeysen S, Compan V, et al. (2004) 5-HT_4 receptors. *Current Drug Targets – CNS & Neurological Disorders* 3: 39–51.

Glennon RA (2003) Higher-end serotonin receptors: 5-HT_5, 5-HT_6, and 5-HT_7. *Journal of Medical Chemistry* 46: 2795–2812.

Hartig PR, Hoyer D, Humphrey PP, et al. (1996) Alignment of receptor nomenclature with the human genome: Classification of 5-HT_{1B} and 5-HT_{1D} receptor subtypes. *Trends in Pharmacological Sciences* 17: 103–105.

Hoyer D, Clarke DE, Fozard JR, et al. (1994) International Union of Pharmacology classification of receptors for 5-hydroxytryptamine (serotonin). *Pharmacological Reviews* 46: 157–203.

Hoyer D and Martin G (1997) 5-HT receptor classification and nomenclature: Towards a harmonization with the human genome. *Neuropharmacology* 36: 419–428.

Lanfumey L and Hamon M (2004) 5-HT_1 receptors. *Current Drug Targets – CNS & Neurological Disorders* 3: 1–10.

Jones JJ and Blackburn TP (2002) The medical benefit of 5-HT. *Pharmacology, Biochemistry and Behavior* 71: 555–568.

Leysen JE (2004) 5-HT_2 receptors. *Current Drug Targets – CNS & Neurological Disorders* 3: 11–26.

Lui M and Gershon MD (2005) Slow excitatory ("5-HT1P"-like) responses of mouse myenteric neurons to 5-HT: Mediation by heterodimers of 5-HT1B/1D and Drd2 receptors. *Gastroenterology* 128(4, supplement 2): A87.

McLean PG, Borman RA, and Kee K (2007) 5-HT in the enteric nervous system: Gut function and neuropharmacology. *Trends in Neurosciences* 1: 9–13.

Nelson DL (2004) 5-HT_5 receptors. *Current Drug Targets – CNS & Neurological Disorders* 3: 53–58.

Pauwels PL (2000) Diverse signalling by 5-hydroxytryptamine (5-HT) receptors. *Biochemical Pharmacology* 60: 1743–1750.

Peters JA, Hales TG, and Lambert JJ (2005) Molecular determinants of single channel conductance and ion selectivity in the Cys-loop transmitter-gated ion channels: Insights from the 5-HT_3 receptor. *Trends Pharmacological Sciences* 26: 587–594.

Sanders-Bush E, Fentress H, and Hazelwood L (2003) Serotonin 5-HT_2 receptors: Molecular and genomic diversity. *Molecular Interventions* 3: 319–330.

Steward LJ, Ge J, Bentley KR, et al. (2007) Guide to receptors and channels (GRAC), 2nd edition (2007 revision). *British Journal of Pharmacology* 150(supplement 1): S1–S168.

Thomas DR and Hagen JJ (2004) 5-HT_7 receptors. *Current Drug Targets – CNS & Neurological Disorders* 3: 81–90.

Woolley ML, Marsden CA, and Fone KC (2004) 5-ht_6 receptors. *Current Drug Targets – CNS & Neurological Disorders* 3: 59–79.

Relevant Website

http://www.iuphar-db.org – IUPHAR Receptor Database.

Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology

G Aghajanian and R-J Liu, Yale School of Medicine, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Serotonin (5-hydroxytryptamine; 5-HT), a biogenic indoleamine derived metabolically from the dietary amino acid tryptophan, was first discovered in peripheral tissues and subsequently found in brain. Shortly afterward, in the early 1950s, a structural relationship was noted between serotonin and *d*-lysergic acid diethylamide (LSD), the most potent of all known hallucinogenic drugs. Subsequently, it was recognized that a wide range of psychedelic hallucinogens, both indoleamines such as *N*,*N*-dimethyltryptame (DMT) and phenethyamines such as mescaline, had structural similarities to LSD (Figure 1). This relationship to powerful psychotropic drugs sparked interest in the idea that serotonin might function as a central 'neurohumoral' substance with a profound influence on behavior. However, it was not obvious how to investigate this idea directly until the mid-1960s, when, through the use of a new histochemical method (the Falck–Hillarp method), it was discovered that serotonin was highly concentrated in specific sets of neurons located almost exclusively in the raphe nuclei of the brain stem. Histochemical maps showed that serotonergic projections emanating from the raphe nuclei innervate, to varying degrees, virtually all other parts of the central nervous system, ranging from the cerebral cortex to the spinal cord. Based on these maps, it became possible to study the properties of identified serotonergic neurons and their postsynaptic targets.

Although it was its link to hallucinogenic drugs that first brought it to prominence, serotonin has now been implicated in almost every conceivable physiological or behavioral function – affect, aggression, appetite, cognition, emesis, endocrine function, gastrointestinal function, motor function, neurotrophism, perception, sensory function, sex, sleep, respiration, and vascular function. Moreover, many drugs that are currently used for the treatment of psychiatric disorders (e.g., depression, mania, schizophrenia, autism, obsessive–compulsive disorder, anxiety disorders) are thought to act, at least partially, through serotonergic mechanisms.

How is it possible for serotonin to be involved in so many different processes? One answer lies in the anatomy of the serotonergic system, where small clusters of serotoninergic cell bodies located in the brain stem raphe nuclei are positioned through their vast projections to influence all regions of the neuraxis. Another answer lies in the molecular diversity and differential cellular distribution of the many serotonin receptor subtypes that are expressed differentially in brain and other tissues. For example, the effects of hallucinogenics are mediated via the 5-HT_{2A} subtype of serotonin receptor, which is expressed extensively in the cerebral cortex. In contrast, antidepressant effects are thought to involve members of the 5-HT_1 family of receptors located in the brain stem raphe as well as in certain postsynaptic regions. Given this diversity, it is useful to approach serotonin neurophysiology within the context of the neuroanatomical and cellular distribution of serotonin receptor subtypes and their associated transduction pathways.

Serotonergic Neuronal Pathways

Serotonergic neurons are clustered in a series of midline nuclei, labeled B1–B9, extending from the midbrain to the medulla oblongata. The dorsal and median raphe nuclei (B7 and B8) comprise the largest of these groupings of brain serotonergic neurons, and together they provide the major serotonergic input to the forebrain; the lower brain stem raphe nuclei supply the main serotonergic projections to the spinal cord. A single serotonergic neuron projects to large numbers of postsynaptic cells and does so in an overlapping fashion with other serotonergic neurons. While the density of this input varies widely, both regionally and between different cell types within a given region, virtually no corner of the CNS is lacking a serotonergic input despite the fact that serotonergic neurons represent far less than 1% of all the neurons in the brain. Given these anatomical considerations, the serotonergic system is well positioned to provide a broad modulatory influence upon diverse functions rather than the precise point-to-point transmission characteristic of traditional sensory, motor, or associational pathways.

In view of its widespread projections arising from a relative handful of neurons, is it possible to assign an overarching function to the serotonergic system? An answer to this question may come from the fact that serotonergic neurons display a pattern of firing that is dependent upon behavioral state: they fire most rapidly during alert waking, decelerate considerably during slow-wave sleep, and become almost entirely silent during dream – or rapid-eye-movement

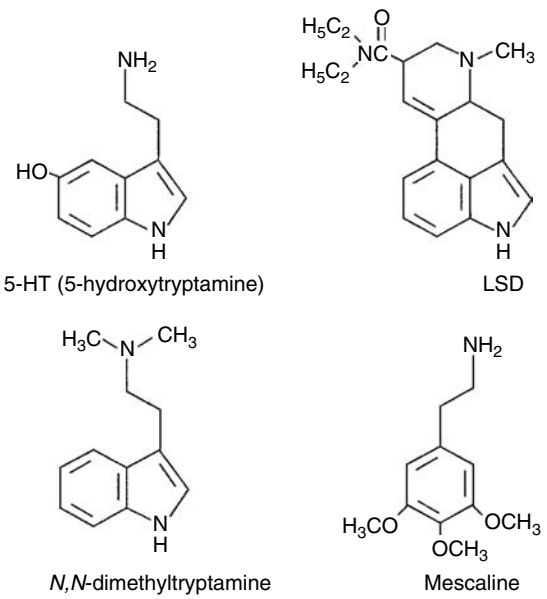

Figure 1 The structural formulas of serotonin and three related hallucinogenic drugs: LSD, DMT, and mescaline. Note that serotonin, LSD, and DMT have in common an indolethylamine moiety. While it is not an indolethylamine, mescaline shares with LSD a phenethylamine moiety. The structures are drawn so as to emphasize these shared structural features. Provided by GK Aghajanian.

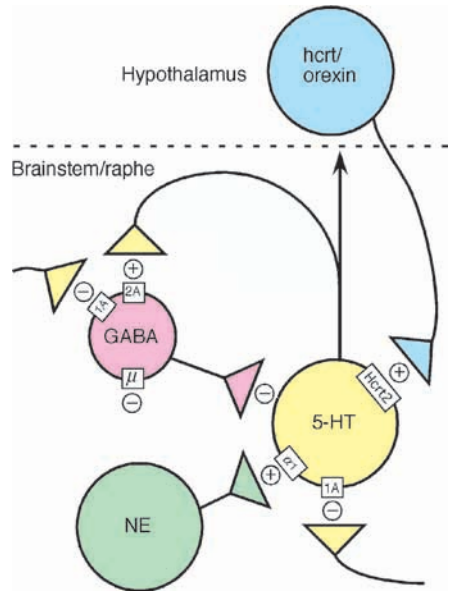

Figure 2 Schematic diagram showing norepinephrine (NE), hypocretin/orexin (hcrt/orexin), and γ-aminobutryic acid (GABA) inputs to a 5-HT cell in the brain stem raphe. The NE input originates from nearby sites in the brain stem while the hcrt/orexin input is hypothalamic in origin; both of these are excitatory (+). Also depicted is a 5-HT collateral input to the 5-HT cell; this input is inhibitory (−) via somatodendritic 5-HT_{1A} autoreceptors. Finally, an inhibitory input (−) from a local GABA interneuron is shown; the GABA cell itself receives 5-HT collaterals, which can be excitatory (via 5-HT_{2A} receptors) or inhibitory (via 5-HT_{1A}) receptors. Provided by GK Aghajanian.

(REM) – sleep. Thus, in a broad sense, one can view serotonergic neurons as constituting a 'wake-on' system. In this context, many of the electrophysiological effects of serotonin in postsynaptic regions can be seen as facilitating motor, sensory, and cognitive functions characteristic of the waking state. All of this would be reversed when serotonergic neurons become quiescent during sleep states, when there is an emergence of sleep- and dream-related systems. Such a broad, state-related function for serotonin has been likened to the peripheral sympathetic system where small clusters of noradrenergic neurons project to a myriad of organs including the heart, gut, and skin to influence their function in a concerted manner. Interestingly, a central counterpart to the sympathetic system exists in the brain stem where small clusters of noradrenergic neurons (e.g., in the locus coeruleus) parallel the broad projections of the serotonergic system and similarly have a broad influence on a wide range of wake-related functions.

Regulation of Serotonergic Neuronal Activity

Ultimately, the electrophysiological effects of serotonin depend upon the activity of the serotonergic cells of origin in the raphe nuclei. In various mammalian species, serotonergic neurons in waking states have been found to have a slow, tonic pattern of firing ($\sim$1–4 spikes s^{-1}). During slow-wave sleep, there is a deceleration of firing rate, and during REM (or dream) sleep, nearly a total cessation of activity. Thus, serotonergic neurons can be seen as being 'wake-on' cells. The maintenance of rhythmic firing during the waking state has suggested that serotonergic neurons possess regulated tonic pacemaker mechanisms. Intracellular and whole cell recordings from dorsal raphe neurons show that spikes arise from gradual depolarizing ramps (pacemaker potentials) rather than synaptic potentials. The regular pacemaker rhythm of serotonergic neurons is shaped by a complex interplay of intrinsic ionic currents; these include a voltage-dependent transient outward potassium current, a low-threshold inward calcium current, and a large calcium-activated outward potassium current. However, the drive behind the pacemaker activity mainly comes from extrinsic excitatory inputs, including brain stem noradrenergic inputs and hypothalamic hypocretin (orexin) inputs (Figure 2). The hypocretin system is essential for maintaining the normal waking state and loss of this system leads to the condition of narcolepsy, characterized by excessive somnolence. Both norepinephrine (via α_1-adrenoceptors) and the peptide hypocretin/orexin

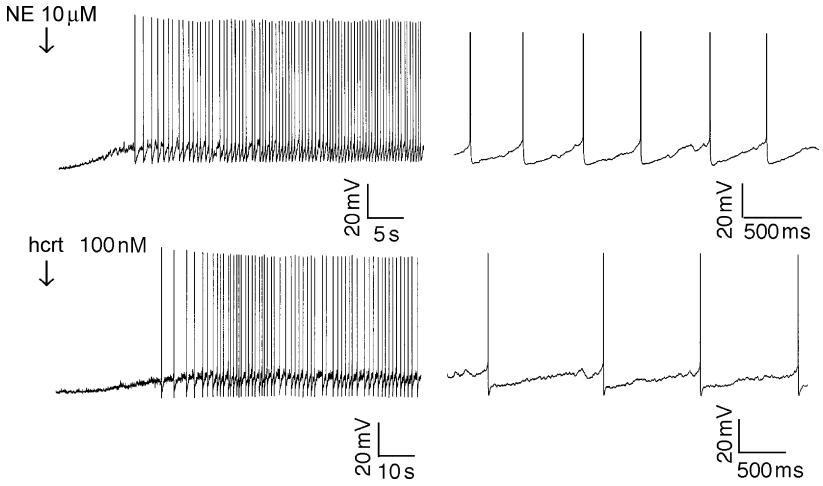

Figure 3 Activation of pacemaker activity in a serotonergic neuron by norepinephrine (NE) or hypocretin in brain slice. Whole cell recording from a serotonergic neuron located in the dorsal raphe nucleus of a midbrain slice. As is typical of serotonergic neurons in brain slice where excitatory afferents are disconnected, this cell shows no basal spiking activity. When NE (top traces) or hypocretin/orexin (hcrt; bottom traces) is applied, there is a slow depolarization leading to tonic pacemaker activity. The expanded traces to the right show the gradual depolarizing ramps leading up to the onset of spiking activity. NE and hcrt/orexin can be seen as restoring the tonic pacemaker activity that is typical of serotonergic neurons *in vivo*. Provided by R-J Liu and GK Aghajanian.

(via hypocretin receptors 1 and 2), by closing potassium channels and opening nonselective cation channels, accelerate pacemaker activity of serotonergic neurons by inducing a slow depolarization (Figure 3). In addition to neurotransmitter inputs, elevated levels of CO_2 also have a strong activating effect on serotonergic as well as noradrenergic neurons. This would provide an alerting signal in response to states of respiratory insufficiency.

Serotonergic pacemaker activity is negatively modulated by the autoinhibitory action of serotonin, acting via somatodendritic 5-HT_{1A} autoreceptors. The maintenance of tonic 5-HT_{1A} autoinhibition depends upon the rate of serotonins synthesis, which in turn depends on availability of the initial serotonin precursor L-tryptophan. The activity of tryptophan hydroxylase, the limiting step in serotonin synthesis, is highly regulated by the tryptophan hydroxylase-activating kinases calcium/calmodulin protein kinase II (CaMKII) and protein kinase A (PKA). Increased calcium influx at higher firing rates, by activating tryptophan hydroxylase via CaMKII and PKA, can work together with tryptophan to enhance negative feedback control of the output of the serotonergic system.

Electrophysiology of Serotonin Receptor Subtypes

The electrophysiological actions of serotonin are a function of the receptor subtype(s) expressed by a given cell. The G_i/G_o-coupled 5-HT_1 receptors generally mediate inhibitory effects on neuronal firing through an opening of potassium channels or a closing of calcium channels. Inhibitions mediated by 5-HT_{1A} and other 5-HT_1 receptor subtypes have been observed in postsynaptic neurons located in many regions of the central nervous system. Presynaptic 5-HT_{1A} receptors, located at the somatodendritic region of serotonergic neurons, have been termed 'autoreceptors,' mediating negative feedback inhibition via the cell's own transmitter. The 5-HT_{2A} receptors mediate slow excitatory effects through a decrease in potassium conductance or an increase in nonselective cation conductance. These are merely a few examples of how serotonin, with its widespread projections and large array of receptor subtypes/transduction pathways, is able to modulate neuronal excitability in a complex fashion throughout the neuraxis.

While the electrophysiological actions of serotonin may seem quite varied, there is considerable uniformity within each of the major receptor families. Because of their common G-protein coupling, all members of the 5-HT_1 receptor family tend to have inhibitory actions either pre- or postsynaptically. Similarly, all members of the 5-HT_2 family tend to have excitatory actions. Therefore, the discussion of serotonin electrophysiology in selected regions of brain will be organized according to receptor families and their transduction pathways (Figure 4).

5-HT_1 Receptors

Members of the 5-HT_1 receptor family are coupled via G_i/G_o G proteins and produce their effects through opening of inwardly rectifying K^+ channels or closing of Ca^{2+} channels. Particularly high levels of 5-HT_{1A} receptors are found in certain regions,

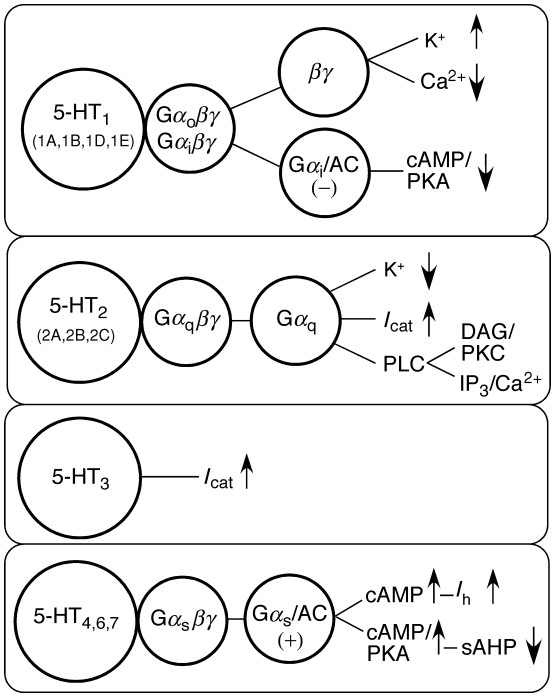

Figure 4 The four major groupings of serotonin receptor families and their transduction pathways. Three of the four groups are coupled through G-proteins: 5-HT_1-G_i/G_o, 5-HT_2-G_q, and 5-$HT_{4,6,7}$-G_s. The remaining receptor is 5-HT_3, which is a ligand-gated channel and is not coupled through a G protein. cAMP, adenosine 3′,5′-cyclic monophosphate; DAG, diacylglycerol; PKC, protein kinase C; sAHP, slow after hyperpolarization; PLC, phospholipase C. Provided by GK Aghajanian.

including the dorsal raphe nucleus, hippocampus, lateral septum, and certain layers of the cerebral cortex. Studies in these regions and other regions have been useful in delineating the physiological role of this receptor.

Raphe nuclei Serotonergic neurons of the raphe nuclei are inhibited by the local (microiontophoretic) application of serotonin to their cell body region. Thus, the receptor mediating this effect has been termed a somatodendritic autoreceptor (as opposed to the prejunctional autoreceptor). Functionally, the somatodendritic serotonin autoreceptor has been shown to mediate collateral inhibition. The ionic basis for the autoreceptor-mediated inhibition, either by serotonin or LSD, is an opening of K^+ channels to produce a hyperpolarization; these channels are characterized by their inwardly rectifying properties. Patch clamp recordings in the cell-attached and outside-out configuration from such acutely isolated dorsal raphe neurons show that the increase in K^+ current results from a greater probability of opening of unitary K^+ channel activity.

The somatodendritic autoreceptors of serotonergic neurons in both the dorsal raphe and other raphe nuclei appear to be predominantly of the 5-HT_{1A} subtype, as a variety of drugs with 5-HT_{1A} selectivity (e.g., 8-OH-DPAT) share the ability to potently inhibit raphe cell firing in a dose-dependent manner. Highly selective 5-HT_{1A} antagonists (e.g., WAY 100635) have been found which potently block the direct inhibition of dorsal raphe serotonergic neurons both by serotonin and selective 5-HT_{1A} agonists. As described above, the somatodendritic 5-HT_{1A} autoreceptor is importantly involved in the negative feedback regulation of sertonergic cell activity through the release of endogenous serotonin.

Other subcortical regions Inhibitory or hyperpolarizing responses to serotonin have been reported in a wide variety of neurons in the spinal cord, brain stem, and diencephalon. In general, such responses have been attributed to mediation by 5-HT_1 receptors. In the nucleus prepositus hypoglossi, focal electrical stimulation evokes inhibitory postsynaptic potentials (IPSPs) that are mediated by 5-HT_{1A} receptors to activate an inwardly rectifying K^+ conductance and a novel outwardly rectifying K^+ conductance. In the midbrain periaqueductal gray, a region known to be involved in pain modulation and fear responses, approximately half the cells are inhibited/hyperpolarized by 8-OH-DPAT, suggesting mediation by 5-HT_{1A} receptors. In the ventromedial hypothalamus and lateral septum, serotonin and 5-HT_{1A} agonists produce inhibitory effects also by activating a potassium conductance. In addition to these postsynaptic effects, serotonin has been shown to suppress glutamatergic synaptic transmission via presynaptic 5-HT_{1B} receptors in various regions including the hypoglossal nucleus and the nucleus accumbens.

In the rat laterodorsal tegmental nucleus (LDT), bursting cholinergic neurons are hyperpolarized by serotonin via 5-HT_1 receptors. In freely behaving rats, the direct injecton of serotonin into the LDT was found to suppress REM sleep. In unanesthetized cats, a corresponding population of neurons that are active selectively during REM states (REM-on neurons) in the LDT are inhibited by direct application of the 5-HT_{1A} agonist 8-OH-DPAT. It has been proposed that during REM sleep, the removal of a tonic inhibitory serotonin influence from these cholinergic neurons may be responsible for the emergence of an activated EEG during this behavioral state (see later).

Hippocampus Pyramidal cells of the CA1 region express high levels of 5-HT_{1A} receptor mRNA and 5-HT_{1A} receptor binding. Serotonin-induced inhibition in both CA1 and CA3 pyramidal cells is mediated

by the activation of receptors of the 5-HT_{1A} subtype coupled to an opening of K^+ channels. In addition to the above direct effects on pyramidal cells, serotonin has been shown to depress both excitatory and inhibitory synaptic potentials in the hippocampus. Relatively high concentrations of serotonin cause a reduction in electrically evoked excitatory postsynaptic potentials (EPSPs) in CA1 pyramidal cells, an effect that is mimicked by 8-OH-DPAT, suggesting mediation by 5-HT_{1A} receptors. Indirect measures indicate that serotonin acts presynaptically to reduce Ca^{2+} entry and thereby gluamatergic synaptic transmission. In addition, there is a 5-HT_{1A}-mediated inhibitory effect on putative inhibitory interneurons of the hippocampus. Functionally, the 5-HT_{1A}-mediated inhibition of GABAergic interneurons in the hippocampus leads to a disinhibition of pyramidal cells in CA1. Clearly, the effects of serotonin in the hippocampus are highly complex, involving both pre- and postsynaptic actions which may, to varying degrees, be inhibitory or disinhibitory, facilitatory or disfacilitatory.

Cerebral cortex 5-HT_{1A}-induced hyperpolarizing/inhibitory responses in pyramidal cells of the cerebral cortex have been described in a number of studies. In entorhinal cortex, where there is an especially high density of 5-HT_{1A} receptors (and a low density of 5-HT_{2A} receptors), unopposed 5-HT_{1A} receptor-mediated hyperpolarizing responses are seen. However, cortical neurons in most other regions typically display mixed inhibitory and excitatory responses to serotonin due to the expression by the same pyramidal cells of multiple serotonin receptor subtypes (e.g., 5-HT_{1A}, and 5-$HT_{2/2C}$). Hyperpolarizing responses mediated by 5-HT_{1A} receptors are often unmasked or enhanced in the presence of 5-HT_2 antagonists, consistent with the idea that there is an interaction between 5-HT_{1A} and 5-HT_{2A} receptors at an individual neuronal level. In addition to these postsynaptic effects, there are various presynaptic effects mediated by 5-HT_1 receptors in the cerebral cortex. In cingulate cortex, serotonin, acting upon presynaptic 5-HT_{1B} receptors, reduces the amplitude of electrically evoked EPSPs, including both *N*-methyl-D-aspartate (NMDA) and non-NMDA components. Similar modulations of EPSPs, mediated by 5-HT_{1A} or 5-HT_{1B} receptors, have been reported for several cortical regions including medial prefrontal and entorhinal cortex.

5-HT_2 Receptors

Members of the 5-HT_2 family of receptors are coupled through Gq-type G proteins and have excitatory effects through closing of K^+ channels or opening nonselective cation channels. High concentrations of 5-HT_2 are expressed in certain regions of the forebrain such as the neocortex, piriform cortex, claustrum, and olfactory tubercle. With few notable exceptions (e.g., motor nuclei and the nucleus tractus solitarius), relatively low concentrations of 5-HT_2 receptors or mRNA expression are found in the brain stem and spinal cord. Studies aimed at examining the physiological role of 5-HT_2 receptors in several of these regions are described in the following sections.

Motoneurons In spinal cord and brain stem motor nuclei, motoneurons have a high density of 5-HT_2 receptor binding sites. Early studies *in vivo* showed that serotonin applied microiontophoretically does not by itself induce firing in the normally quiescent facial motoneurons but does facilitate the subthreshold and threshold excitatory effects of glutamate. Intracellular recordings from facial motoneurons *in vivo* or in brain slices *in vitro* show that serotonin induces a slow, subthreshold depolarization associated with an increase in input resistance, indicating a decrease in a resting K^+ conductance. Selective 5-HT_2 antagonists are able to selectively block the excitatory effects of serotonin in facial motoneurons. The facilitation of motoneuron excitability contributes to the role of serotonin as a 'wake-on' system to promote activity during the waking state (see later).

Other subcortical regions In brain slices of the medial pontine reticular formation, serotonin induces depolarizing responses that have a 5-HT_2 pharmacology and are associated with a decrease in membrane conductance resulting from a decrease in an outward K^+ current. In brain slices of the substantia nigra pars reticulata, a majority of neurons are excited by serotonin via 5-HT_2 receptors, possibly of the 5-HT_{2C} rather than 5-HT_{2A} subtype. Neurons in the inferior olivary nucleus are excited by serotonin via 5-HT_{2A} receptors, thereby altering the oscillatory frequency of input to cerebellar Purkinje cells. In the nucleus accumbens, the great majority of neurons are depolarized by serotonin, inducing them to fire. This depolarization is associated with an increase in input resistance due to a reduction in an inward rectifier K^+ conductance. In addition, GABAergic neurons within a number of regions (e.g., dorsal raphe nucleus, medial septal nucleus, hippocampus, cerebral cortex) are also excited by serotonin via 5-HT_2 receptors, suggesting that in multiple locations within the CNS there are subpopulations of interneurons that are excited by serotonin via 5-HT_2 receptors, giving rise to indirect inhibitory effects.

Cerebral cortex The electrophysiological effects of serotonin have been studied in several cortical-regions. *In vitro* studies in the brain slice preparation have shown that pyramidal cells in various cortical regions respond to serotonin by either a small hyperpolarization, depolarization, or no change in potential. Depending on the region of cortex under study, as described below, the depolarizations appear to be mediated by 5-HT_{2A} or 5-HT_{2C} receptors. In addition to these postsynaptic effects, serotonin induces an increase in 'spontaneous' (nonelectrically evoked) postsynaptic potentials or currents (PSPs/PSCs) recorded in brain slices from various cortical regions. These may originate from the activation of intracortical pathways or through interactions with subcortical inputs such as the thalamus. Activation of 5-HT_{2A} receptors also enhances late components of evoked responses in cerebral cortex (Figure 5). These late responses (also termed UP states) are generated by sustained recurrent network activity. It has recently been shown that UP states rely on glutamate spillover onto extrasynaptic NMDA receptors. Interestingly, these UP states are characteristic of alert waking and are enhanced by psychedelic hallucinogens which are known to be 5-$HT_{2A/2C}$ partial agonists. It has been proposed that the hallucinogenic drugs produce their characteristic aberrations in perception, cognition, and affect by driving cortical UP states beyond normal limits. It is significant that the facilitation of UP states by 5-HT_{2A} receptors is opposed by 5-HT_{1A} receptors which tend to suppress UP states. This damping effect of 5-HT_{1A} receptors on UP states may explain why selective 5-HT_{2A} agonists are hallucinogenic, whereas serotonin, the natural transmitter, is not. This example further illustrates how serotonin, rather than having a monolithic action, exercises numerous checks and balances at a cellular and systems level via its diverse receptor subtypes.

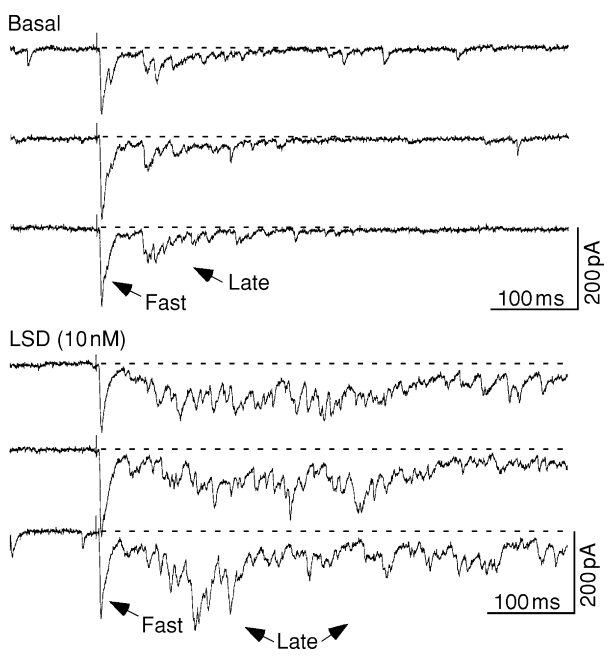

Figure 5 Enhancement of electrically evoked recurrent network activity (prolonged late excitatory postsynaptic current (EPSC) or UP state) in prefrontal brain slice by the hallucinogen LSD. Under basal conditions (top traces), note the fast EPSC, which occurs within a few milliseconds of a local electrical stimulus applied to the brain slice; following the fast EPSC, a low-amplitude late EPSC can be seen. After the application of a low concentration of LSD in the bath, the late EPSC or UP state is greatly enhanced both in amplitude and duration. Whole cell recording from a layer V pyramidal cell in rat prefrontal cortex. Provided by GK Aghajanian.

5-HT_3 Receptors

Excitatory responses to serotonin have been found in various central neurons that have a rapid onset of action and rapid desensitization, features that are typical of ligand-gated ion channels rather than G-protein-coupled receptor responses. The cloned 5-HT_3 receptor homologous with the nicotinic acetylcholine receptor and the β_1 subunit of the $GABA_A$ receptor indicate that it is a member of the ligand-gated ion channel superfamily. Typically members of this superfamily are comprised of multiple subunits; however, only one 5-HT_3 receptor subunit and an alternatively spliced variant has been cloned to date.

In hippocampal slices, serotonin has been reported to increase spontaneous GABAergic IPSPs, most likely through a 5-HT_3 receptor-mediated excitation of inhibitory interneurons; these responses also show fading with time. A similar induction of 5-HT_3 receptor-mediated induction of inhibitory postsynaptic currents (IPSCs) has been reported in the neocortex. While fast, rapidly inactivating excitation has generally become accepted as characteristic of 5-HT_3 receptors, non-desensitizing responses have also been reported. In dorsal root ganglion cells, a relatively rapid but non-inactivating depolarizing response has been described that has a 5-HT_3 pharmacological profile. In neurons of the nucleus tractus solitarius brain slices, there is a postsynaptic depolarizing response to 5-HT_3 agonists that is not rapidly desensitizing.

5-HT_4, 5-HT_6, 5-HT_7

There are three known G_s-coupled serotonin receptors: 5-HT_4, 5-HT_6, and 5-HT_7, all mediating positive coupling of serotonin responses to adenylyl cyclase. The resulting increase in cyclic AMP can produce its electrophysiological effects through interacting directly with ion channels or through the activation of the PKA signal transduction pathway, which can interact

with a myriad of ion channels, transporters, and other kinases. At this time, electrophysiological studies are available only for the 5-HT_4 and 5-HT_7 receptors, as described below.

5-HT_4 receptors Binding studies using a selective 5-HT_4 ligand indicate that 5-HT_4 receptors are present in several discrete regions of the mammalian brain including the striatum, substantia nigra, olfactory tubercle, and hippocampus. As these regions also express 5-HT_4 receptor mRNA, it appears likely that these receptors function postsynaptically to mediate certain actions of serotonin. The best studied of these regions is the hippocampus where both biochemical and electrophysiological studies have provided detailed picture of the actions of serotonin at 5-HT_4 receptors. Electrophysiological studies show that 5-HT_4 receptors mediate an inhibition of a calcium-activated potassium current, which is responsible for the generation of a slow afterhyperpolarization (AHP) in hippocampal pyramidal cells of the CA1 region. A suppression of the AHP would enhance the ability of these cells to respond to excitatory inputs with robust spike activity.

5-HT_7 receptors The 5-HT_7 receptor is positively coupled to adenylate cyclase and can mediate electrophysiological actions through the cyclic AMP pathway. One such effect is through a cyclic nucleotide enhancement of the hyperpolarizing-activated nonselective cationic current I_h. An increase in I_h tends to prevent excessive hyperpolarization and increase neuronal excitability. 5-HT_7 receptors can also influence neuronal excitability through cyclic AMP-activated PKAs, leading to suppression of slow AHPs. This occurs in the midline/intralaminar nuclei of the thalamus where the highest expression of 5-HT_7 receptors in the brain occurs. As the thalamocortical pathway projecting from the midline/intralaminar nuclei is the final link in the ascending arousal pathway, this is one way that serotonin exerts its effects as a wake-on system to promote alertness and attention (see later).

Behavioral and Clinical Correlations

The diversity of receptors and transduction pathways underlying the varied electrophysiological actions of serotonin, together with the differential expression of these receptors in different neuronal populations, helps explain how one transmitter can be involved in such a large array of behaviors, clinical conditions, and drug actions. Alterations in serotonin function have been linked to affective disorders, anxiety states, schizophrenia, obsessive–compulsive disorder, eating disorders, migraine, and sleep disorders, including sleep apnea. Drugs that modify serotonergic transmission include antidepressants, atypical antipsychotics, antiemetics, hallucinogens, antimigraine drugs, and appetite suppressants. For example, the most commonly used antidepressant drugs act by blocking serotonin reuptake, which results in increased synaptic availability of this transmitter. Antagonism of 5-HT_{2A} receptors has been proposed as contributing to the favorable clinical profile of atypical antipsychotic drugs. Conversely, an agonist action of psychedelic hallucinogens at 5-HT_{2A} receptors is thought to mediate the psychotomimetic effects of these drugs. Because of its ubiquitous distribution and diverse actions in the central nervous system, it is likely that future research will continue to uncover an involvement of serotonin in many pathological as well as normal behavioral processes.

It has been hypothesized that activation of serotonin neurons works in concert with other components of the ascending arousal system projecting to the forebrain. For example, noradrenergic neurons, which also constitute a wake-on system, directly activate serotonergic neurons. A further linkage is seen by the fact that both noradrenergic and serotonergic neurons are activated by inputs from hcrt/orexin neurons, which are essential for normal wakefulness. A concerted action is also seen upstream, by the excitatory/facilitatory effects of both serotonin and hcrt upon the relay neurons of the midline/intralaminar nuclei which project to prefrontal cortex. This prefrontal input, which represents the final synapse in the ascending arousal system, is essential for awareness and attention. Within the cortex itself, 5-HT_{2A} receptors enhance UP states, which are closely associated with the alert waking state. However, excessive stimulation of 5-HT_{2A} receptors by hallucinogenic drugs can be pathological by leading to uncontrolled UP states.

In conclusion, through its effects on neuronal excitability in diverse regions of the brain and spinal cord, the serotonergic system is positioned to coordinate complex sensory and motor patterns during different behavioral states. More specifically, since the activity of serotonergic neurons is greatest during periods of active waking, reduced during slow-wave sleep, and almost completely absent during REM (dream) sleep, the serotonin system can be seen as promoting neuronal activity in support of the waking state while concomitantly suppressing neuronal activity underlying sleep and dream states.

See also: Aversive Emotions: Genetic Mechanisms of Serotonin; Cells: 5-Hydroxytryptamine Receptors; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Serotonin

(5-Hydroxtryptamine; 5-HT): Receptors; Serotonin and the Regulation of Mammalian Circadian Rhythms; Serotonin-Related Psychedelic Drugs.

Further Reading

Aghajanian GK and Sanders-Bush E (2002) Serotonin. In: Davis KL, Charney D, Coyle JT, et al. (eds.) *Neuropsychopharmacology: The Fifth Generation of Progress*, pp. 15–46. New York: Raven Press, Ltd.

Jacobs BL and Azmitia EC (1992) Structure and function of the brain serotonin system. *Physiological Reviews* 72: 165–229.

Lambe EK and Aghajanian GK (2006) Electrophysiology of $5HT_{2A}$ receptors and relevance for hallucinogen and atypical drug actions. In: Roth BL (ed.) *The Serotonin Receptors: From Molecular Pharmacology to Human Therapeutics*. Totowa, NJ: Humana Press.

Lambe EK and Aghajanian GK (2006) Hallucinogen-induced UP states in the brain slice of rat prefrontal cortex: Role of glutamate spillover and NR2B-NMDA receptors. *Neuropsychopharmacology* 31: 1682–1689.

Liu R-J, Lambe EK, and Aghajanian GK (2005) Somatodendritic autoreceptor regulation of serotonergic neurons: Dependence on L-tryptophan and tryptophan hydroxylase-activating kinases. *European Journal of Neuroscience* 21: 945–958.

Liu R-J, van den Pol AN, and Aghajanian GK (2002) Hypocretins (orexins) regulate serotonin neurons in the dorsal raphe nucleus by excitatory direct and inhibitory indirect actions. *Journal of Neuroscience* 22: 9453–9464.

Richerson GB (2004) Serotonergic neurons as carbon dioxide sensors that maintain pH homeostasis. *Nature Reviews Neuroscience* 5: 449–461.

Whitaker-Azmitia PM (1999) The discovery of serotonin and its role in neuroscience. *Neuropsychopharmacology* 21: 2S–8S.

Serotonin and the Regulation of Mammalian Circadian Rhythms

P J Sollars and G E Pickard, Colorado State University, Fort Collins, CO, USA
J S Sprouse, Pfizer Global Research and Development, Groton, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

The Suprachiasmatic Nucleus Regulates Mammalian Circadian Rhythms

Organisms from plants to humans have developed a self-sustaining clocklike mechanism that coordinates biological functions both with one another and with specific times of the day or night. In mammals, the hypothalamic suprachiasmatic nucleus (SCN) is the primary circadian oscillator responsible for the generation of circadian rhythms in physiology, behavior, and metabolism. The SCN is an autonomous and self-sustaining circadian oscillator generating an oscillation of neural activity with a period of approximately 24 h (the term circadian is derived from the Latin *circa* 'about' and *dies* 'day'). Environmental time cues, the most potent being the day/night cycle, entrain the SCN circadian oscillation of neural activity to the 24-h day/night cycle. Entrainment is achieved by light-induced signals that are generated in the retina and conveyed to the SCN by axons of the retinohypothalamic tract (RHT) traveling via the optic nerve.

Anatomical Organization of the SCN

The SCN in mammals is located in the anterior hypothalamus adjacent to the midline third ventricle and situated just dorsal to the optic chiasm from which the RHT emerges. The SCN is composed of a heterogeneous group of about 10 000 small neurons within a complex neuropil. The dendritic arbors of SCN neurons are relatively simple, with most cells having only two or three dendrites that branch very little. Most SCN neurons are γ-aminobutyric acid (GABA) ergic and are capable of generating a circadian oscillation in their action potential firing rate. SCN GABAergic projection neurons also possess local collaterals that innervate other neurons within the nucleus, thus forming a complex intra-SCN network. The intra-SCN GABAergic network is important for the functional coordination of the unitary output of SCN circadian oscillations that are derived from a large number of oscillating clock cells.

SCN Divisions

Several different neuroactive peptides have been localized within the SCN. Based primarily on the nonhomogeneous distribution of these peptidergic neurons within the SCN, the nucleus has been subdivided into two somewhat imprecise divisions – a ventrolateral region rich in vasoactive intestinal polypeptide (VIP) neurons and a dorsomedial region containing vasopressin (VP) neurons, although as noted earlier, most, if not all, of these neurons also use GABA as a neurotransmitter. Whether these two regions defined by their peptidergic phenotype represent distinct functional units within the SCN remains to be firmly established. It has been suggested by some, based primarily on the distribution of peptidergic neurons in the rat SCN, that the VIP neuron-containing region represents a receptive core that receives both retinal and serotonergic afferents and is surrounded by a shell region containing VP neurons and receiving afferents from nonvisual cortical and subcortical regions. However, significant species variations in the core/shell organizational plan suggest that this may be an oversimplification of the complex organization of the mammalian SCN.

Afferent Fiber Inputs to the SCN

The SCN receives many afferent fiber inputs arising from neurons located in several regions of the brain. The RHT arising from retinal ganglion cells (RGCs) and afferent fibers arising from serotonergic neurons located in the midbrain median raphe nucleus have received significant attention and are perhaps the best-understood afferent fiber systems innervating the SCN.

The Retinohypothalamic Tract

The retinal input to the SCN is bilateral and retinal fibers are found distributed throughout much of the SCN except typically for a medial crescent-shaped region that is almost devoid of retinal input. The RGCs that give rise to the RHT represent 1–2% of the total RGC population, and these cells were recently discovered to be intrinsically photosensitive due to their expression of a novel mammalian opsin, melanopsin. Melanopsin-expressing intrinsically photosensitive retinal ganglion cells (ipRGCs) depolarize in response to light and send environmental irradiance information to the SCN. Melanopsin ipRGCs removed from the retina and maintained in culture retain their photosensitivity, although *in situ* the ipRGCs also receive input in the conventional manner: rod/cone photoreceptor → bipolar/amacrine → RGC. The ipRGCs use glutamate as a neurotransmitter, and these cells also synthesize pituitary adenylate cyclase-activating polypeptide (PACAP), which may

be co-released with glutamate in the SCN. In addition to the SCN, ipRGCs send axons to the thalamic intergeniculate leaflet (IGL), olivary pretectal nucleus (OPN), ventral subparaventricular zone of the hypothalamus, and ventrolateral preoptic (VLPO) nucleus, suggesting that these cells may play a broad role in nonvisual photoreception.

The primary role of the RHT input to the SCN is to convey light/dark information (i.e., level of ambient illumination), thereby entraining the intrinsic SCN circadian oscillation to the environmental 24 h day/night cycle. Entrainment is accomplished by a daily resetting of the SCN such that the light-induced phase shift is equal in magnitude to the difference between the free-running circadian period of the endogenous SCN oscillation and the 24 h period of the day/night cycle. This light-induced resetting process is described experimentally by the response of animals free running under constant dark (DD) conditions and exposed briefly to light at different phases of the circadian cycle. The response of the SCN to brief light pulses is phase dependent: light exposure early in the subjective night produces phase delays in circadian rhythms, while light exposure late in the subjective night results in phase advances. In contrast, light exposure during the subjective day has no effect on the phase of the free-running circadian rhythm. The effects of drugs on light-induced behavioral phase shifts have often been used to explore the neuropharmacology of the SCN circadian system.

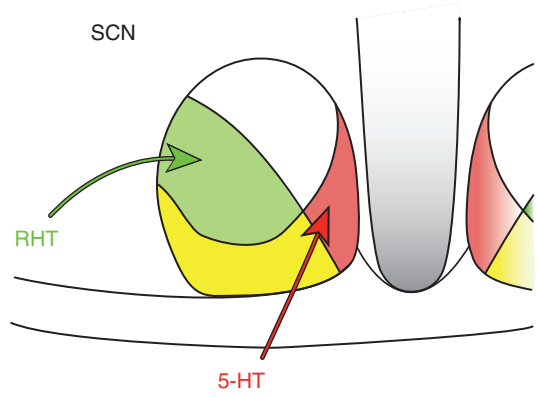

Figure 1 Representation of the retinohypothalamic tract (RHT) and serotonin (5-hydroxytryptamine; 5-HT) terminal fields in the mouse suprachiasmatic nucleus (SCN). The region of the SCN containing the heaviest concentration of retinal fibers is shown in green and the region of densest 5-HT fibers is depicted in red; the yellow region represents the area where the terminal fields are co-extensive.

Serotonergic Afferents

The SCN has long been known to receive one of the densest serotonergic innervations in the brain. This monoaminergic input to the SCN arises from ascending projections of serotonin (5-hydroxytryptamine; 5-HT) neurons in the midbrain median raphe nucleus. 5-HT fibers are distributed throughout much of the SCN, but they are most heavily concentrated in the ventral aspect of the nucleus, forming a curved, shaped terminal plexus that is co-extensive to a considerable degree with the RHT terminal field (Figure 1). In this region of overlap between the two afferent fiber systems, 5-HT fibers preferentially make axo-dendritic and axo-somatic synaptic contacts with VIP neurons that also receive retinal synapses. Ultrastructural analyses have also shown that 5-HT varicosities in the SCN are very often apposed to retinal and GABA terminals; these 5-HT axo-axonic interfaces represent additional important sites for serotonergic modulation of SCN activity.

Early on it was demonstrated that the 5-HT input to the SCN was not essential for the expression of circadian behavior. However, removal of the serotonergic input to the SCN following ablation of the median raphe nucleus does produce an increased sensitivity to photic input in the SCN and an alteration in the phase angle of entrainment of the circadian activity rhythm to the light/dark cycle. These findings are consistent with the generally accepted theme that 5-HT neurons do not play an essential role in physiological and behavioral processes but rather exert tonic modulatory influences on their targets.

Serotonin receptor subtypes in the SCN A family of seven serotonin receptor subtypes (5-HT_{1-7}) with at least 14 members mediates the diverse modulatory effects of serotonin in the central nervous system. Using a variety of techniques, including radioligand binding, *in situ* hybridization for 5-HT receptor mRNA, and immunocytochemical localization for 5-HT receptor proteins at both the light and electron microscopic level of examination, six 5-HT receptor subtypes have been identified in the SCN: 5-HT_{1A}, 5-HT_{1B}, 5-HT_{2A}, 5-HT_{2C}, 5-HT_{5A}, and 5-HT_7. *In situ* hybridization and binding studies typically lack the resolution to provide detailed information on subcellular localization. Electron microscopic immunocytochemical procedures have been used to describe the subcellular distribution for the 5-HT_{1B} and 5-HT_7 receptors in the SCN. 5-HT_{1B} receptor immunostaining is associated primarily with the membrane of axon terminals, many of which are of retinal origin. These 5-HT_{1B} presynaptic receptors are synthesized by RGCs and are transported via the RHT to the SCN, where they are then inserted into the terminal membrane, but almost never at the synaptic active

zone. 5-HT_7 receptors are synthesized by SCN neurons, and 5-HT_7 immunoreactivity is often associated with the dendrites of SCN GABAergic neurons as a postsynaptic receptor.

Effects of Serotonergic Compounds on Retinal Input to the SCN

To determine if the serotonergic innervation of the SCN modulates retinal input, several different approaches have been used. *In vivo* studies have utilized systemic administration of serotonergic drugs to evaluate their effect on light-induced responses of the SCN. Although relatively easy to perform, these studies lack the ability to determine the precise site of action of the drugs administered. For example, the dorsal raphe nucleus provides serotonergic input to the retinorecipient IGL. The IGL sends to the SCN afferent fibers that modulate SCN responses to light. Thus, compounds administered systemically may alter SCN-mediated responses indirectly. Direct microinjection of drugs into the region of the SCN via an indwelling guide cannula provides better assurance that the drugs are acting locally on the SCN, but these types of studies are more difficult to conduct and also require *post hoc* analyses to confirm the precise location of the injection site in the region of the SCN. *In vitro* studies using electrophysiological recording from SCN neurons in a hypothalamic slice preparation provide cellular detail, but the natural stimulus pathway (i.e., light-induced activation of the RHT) is unavailable and electrical stimulation of the optic nerve must be used to simulate photic input to the SCN. Taken together, these different approaches have firmly established that serotonergic afferents modulate photic responses of the SCN.

5-HT Modulation of the SCN Response to Light

Based on the intimate anatomical relationship between 5-HT afferent fibers and RHT afferent fibers in the SCN, it is not surprising that 5-HT has been shown to alter the response of the SCN to light. Single-unit activity recorded from the SCN *in vivo* was initially used to examine the effects of 5-HT agonists on photically responsive cells. Intravenous application of the nonselective 5-HT agonist quipazine was shown to inhibit light-induced responses in the SCN. With the development of more selective serotonergic compounds, it became possible to identify the involvement of specific serotonin receptor subtypes in mediating the effects of 5-HT on photic responses.

5-HT_{1B} receptor-mediated presynaptic inhibition of retinal input 5-HT modulates the sensitivity of the SCN to light by acting on 5-HT_{1B} presynaptic receptors located on retinal terminals and by inhibiting glutamate release. 5-HT_{1B} receptor agonists, including *m*-trifluoromethylphenylpiperazine (TFMPP), CGS-12066A, and CP-94253, injected peripherally inhibit light-induced phase shifts of rodent circadian activity rhythms in a dose-dependent manner. Similarly, 5-HT_{1B} receptor agonists inhibit light-induced expression of the immediate-early gene c-*fos* in the SCN as well as light-induced expression of VIP mRNA in the SCN. 5-HT_{1B} receptor agonists have also been shown to inhibit the light-induced suppression of pineal melatonin during the night. TFMPP injected directly into the SCN has been shown to inhibit light-induced behavioral phase shifts. TFMPP is a well-characterized 5-HT_{1B} receptor agonist, although it also has a relatively high affinity for 5-HT_{1A} and 5-HT_{2C} receptors. However, TFMPP has no effect on light-induced behavioral phase shifts in 5-HT_{1B} receptor knockout mice, which lack functional 5-HT_{1B} receptors, indicating that its effect in wild-type (WT) mice is mediated via 5-HT_{1B} receptors.

Although the studies described here are consistent with the interpretation that the 5-HT agonists act at 5-HT_{1B} receptors located presynaptically on retinal terminals in the SCN, the data do not exclude the possibility that they act at receptors located postsynaptically on the soma or dendrites of SCN neurons. Support for a presynaptic effect has been provided using whole-cell patch clamp recording from SCN neurons in a slice preparation (Figure 2). 5-HT_{1B} receptor agonists decreased reversibly the amplitude of the optic nerve-evoked excitatory postsynaptic current (EPSC) but had no effect on glutamate-evoked whole-cell currents, indicating that the drug effect was due to activation of receptors located presynaptically to the recorded neuron (Figure 3).

5-HT_7 receptor-mediated inhibition of retinal input The mixed 5-$HT_{1A/7}$ receptor agonist 8-hydroxy-2-(di-*n*-propylamino)tetralin (8-OH-DPAT) administered *in vivo* inhibits light-induced behavioral phase shifts and light-induced gene expression in the SCN. 8-OH-DPAT *in vitro* also inhibits optic nerve stimulation-evoked field potentials and optic nerve stimulation-evoked EPSCs in the SCN slice preparation. The effect of 8-OH-DPAT in the latter case is minimally attenuated by a highly selective 5-HT_{1A} antagonist, WAY 100635, but is reversibly and significantly reduced in the presence of ritanserin, a 5-$HT_{2/7}$ antagonist. Although the site of action of 8-OH-DPAT within the SCN has not been addressed directly, 5-HT_7 receptors are postsynaptic in the SCN and have been described on VIP cells that receive retinal synapses.

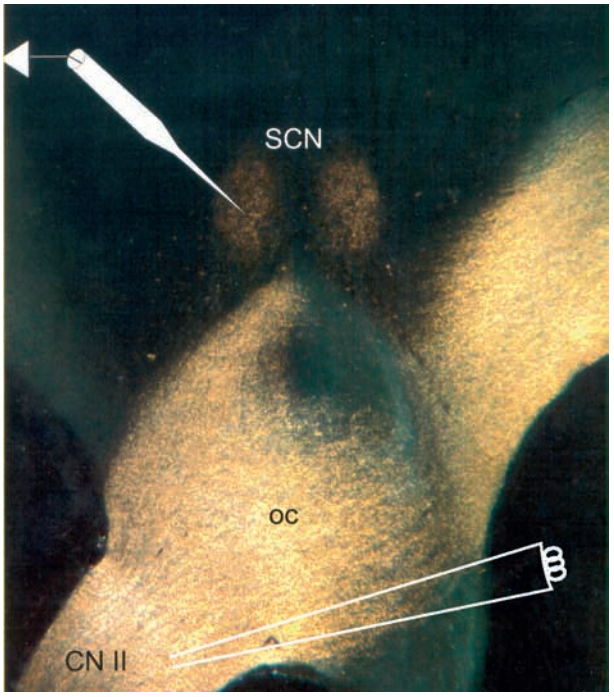

Figure 2 Composite figure of a horizontal slice of the ventral hypothalamus, illustrating recording configuration for examining retinohypothalamic input to the suprachiasmatic nucleus (SCN). To illustrate bilateral retinal innervation of the SCN, retinal fibers were labeled after monocular injection of horseradish peroxidase into the vitreous body of the eye. Horseradish peroxidase-labeled optic fibers are visible in this dark-field image in the optic nerve (CN II) of the injected eye, crossing in the optic chiasm, and terminating bilaterally in the SCN just posterior to the optic chiasm (OC). Whole-cell patch clamp recordings of evoked excitatory postsynaptic currents in SCN neurons after optic nerve stimulation were conducted to examine the effect of serotonergic compounds on retinal input. Adapted from Pickard GE, Smith BN, Belenky M, et al. (1999) 5-HT_{1B} receptor-mediated presynaptic inhibition of retinal input to the suprachiasmatic nucleus. *Journal of Neuroscience* 19: 4034–4045, copyright 1999 by the society for Neuroscience.

5-HT Alterations in Circadian Phase

Alterations in the phase of the SCN clock generated by stimuli other than light (e.g., arousal, neuropeptides, and 5-HT agonists) fall into the general category of nonphotic phase shifts. Such phase shifts typically are largest during the subjective day, when light has little effect on the SCN. 8-OH-DPAT administered systemically during the subjective night does not alter the phase of the circadian activity rhythm, whereas it produces phase advances with maximal effects observed in the midsubjective day. These effects may be due to 8-OH-DPAT acting on somtatodendritic 5-HT_{1A} autoreceptors in the raphe nucleus, resulting in a decrease in 5-HT firing and subsequent decrease in 5-HT release in the SCN. The disinhibition of SCN unit activity may be responsible for the phase shift. It is thought that the 5-HT_{1A} agonist MKC-242 potentiates the effects of light on the SCN by such an effect on median raphe activity. The inability of 8-OH-DPAT to elicit phase shifts when injected into the SCN supports a site of action outside the nucleus. However, it is unclear how much these data can be generalized, as there appears to be significant species differences in the susceptibility of the circadian system to modulation by systemically delivered serotonergics.

Effects of Serotonergic Compounds on SCN Rhythms *In Vitro*

Circadian rhythms may be monitored *in vitro* by isolating the SCN in a brain slice preparation and observing changes in the rhythms of spontaneous unit firing. These rhythms readily survive outside the animal, generating distinct peaks near the midpoint of the light phase and repeating this pattern at roughly 24 h intervals. Peaks occur at the same time of day in tissue obtained from both nocturnal and diurnal species, suggesting that the ecological preference for daytime or nighttime activity is downstream of the oscillator. Shifts in the occurrence of the peak, either in a phase advance or phase delay direction, can be accomplished by briefly incubating the slices with fixed concentrations of drug to yield both concentration-response and phase-response curves. Presumably the magnitude and direction of the shift observed as such *in vitro* form the basis of the shifts revealed in awake, behaving animals. That this is not the case, based on extensive research in the past few years, hints at the complexity of the pharmacology and our crude understanding of it.

Roles for 5-HT_{1A}, 5-HT_{5A}, and 5-HT_7 Receptors

Early *in vitro* studies with 8-OH-DPAT (at the time believed to be a selective 5-HTIA receptor agonist), when given at midday or ZT6 (zeitgeber time, with ZT0 defined as 'lights on'), demonstrated robust phase advances. Such shifts were attributed to this receptor subtype, a tentative finding in the absence of confirmatory experiments with 5-HT antagonists. When the latter approach was undertaken, a somewhat surprising interpretation surfaced with a significantly greater role for 5-HT_7 receptor activation. Specifically, pretreatment with antagonists selective for the 5-HT_{1A} subtype failed to block the 8-OH-DPAT-induced advances; pretreatment with those counting 5-HT_7 antagonism among their activities were considerably more effective. Although the density of 5-HT_7 sites within the SCN certainly gave credence to this interpretation, it left the somewhat more potent activity of 8-OH-DPAT at the 5-HT_{1A}

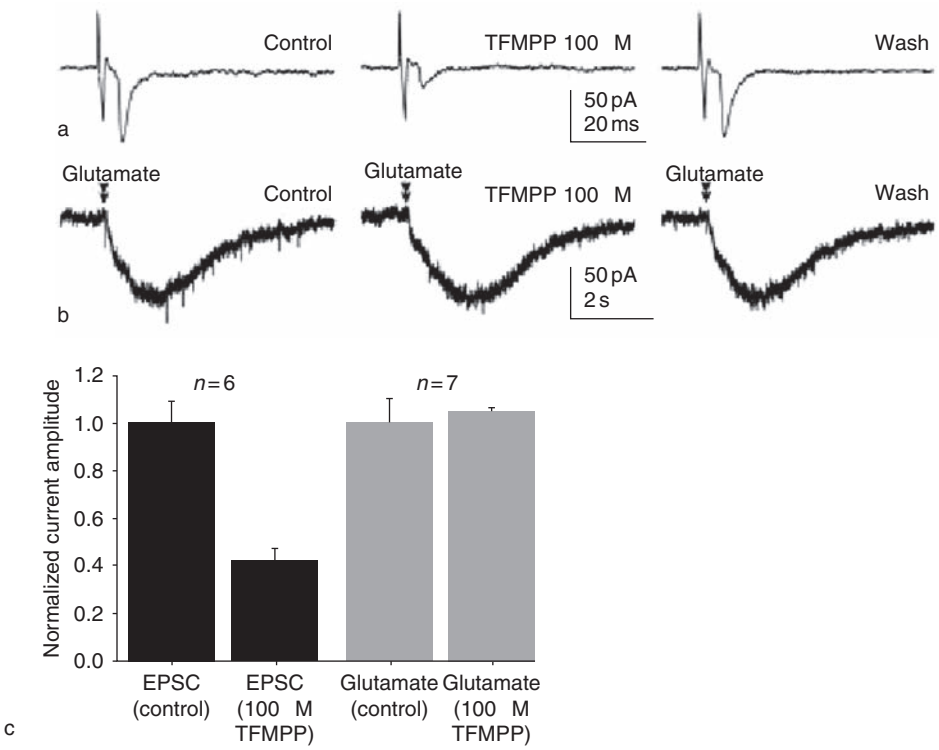

Figure 3 *m*-Trifluoromethylphenylpiperazine (TFMPP) reduces the amplitude of evoked excitatory postsynaptic currents (EPSCs) but has no effect on the inward current evoked by direct application of glutamate. (a) TFMPP reversibly reduced the amplitude of the optic nerve-evoked EPSC. The average of eight events is shown. (b) In the same neuron, direct application of glutamate (20 mM) to the surface of the slice evoked an inward current that was not reduced by TFMPP. (c) The effect of 100 μM TFMPP on the optic nerve-evoked EPSC and on the glutamate-evoked inward current in several neurons is shown. The number of neurons examined is indicated above each column set. Adapted from Pickard GE, Smith BN, Belenky M, et al. (1999) 5-HT_{1B} receptor-mediated presynaptic inhibition of retinal input to the suprachiasmatic nucleus. *Journal of Neuroscience* 19: 4034–4045, copyright 1999 by the Society for Neuroscience.

subtype (and its relative abundance) without an apparent role. At least, an immediate role was not apparent, although this was to change in due time.

What was true for 8-OH-DPAT would surely be true for 5-HT itself and initial studies did confirm very similar phase-response curves for both the time of dosing for the peak effect and the relative magnitude of the response (Figure 4). When co-incubated with 5-HT_7 antagonists, however, a surprising pattern emerged: the phase advances elicited by 5-HT at ZT6 were apparently reduced by one 5-HT_7 antagonist, ritanserin, but not by another, mesulergine. Subsequent receptor binding studies demonstrated a 25-fold greater affinity of ritanserin for 5-HT_{5A} sites compared to mesulergine, an observation suggestive of a 5-HT_{5A} mechanism for 5-HT and consistent with earlier observations of robust labeling of 5-HT_{5A} sites in the SCN. The Rp diastereomer of adenosine 3′,5′-cyclic monophosphothioate (Rp-cAMPS), a cAMP antagonist, failed to block the phase shifts elicited by 5-HT at a concentration previously shown to be effective against 8-OH-DPAT-induced phase shifts, in keeping with the proposed negative coupling of 5-HT_{5A} receptors to cAMP production and the positive coupling assigned to 5-HT_7. While these data were internally consistent, the broader picture grew considerably more complicated: similar phase advances by an exogenous agonist and an endogenous neurotransmitter that were mediated by two distinct 5-HT receptor subtypes.

Efforts to simplify this story only served to add to its complexity. The development of a 5-HT_7 knockout (KO) mouse promised a quick confirmation of earlier findings – the loss of phase advances following incubations with 8-OH-DPAT and the continuing presence of these when 5-HT was used – as a receptor mechanism functioning exclusively through 5-HT_{5A} receptors should not be affected. The results were again surprising, this time in two ways. In slices obtained from 5-HT_7 KO animals, bath application of 8-OH-DPAT produced a phase advance in neuronal firing when applied one circadian cycle earlier at ZT6, but the mean phase advance was no different in slices prepared from KO or matched WT animals. Co-application of 8-OH-DPAT with a highly selective 5-HT_{1A} antagonist significantly reduced the phase advance, both in experiments with WT and KO

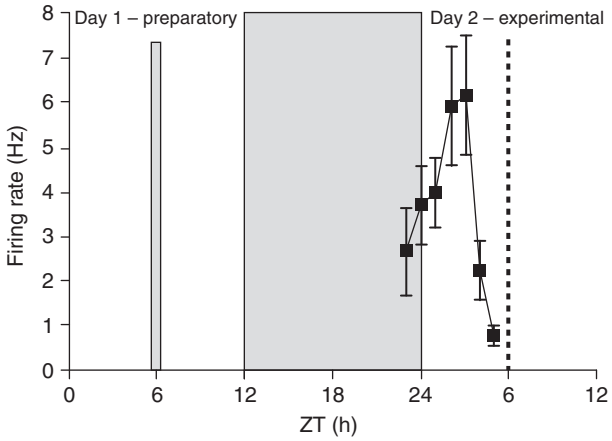

Figure 4 Phase-shifting effects of 5-hydroxytryptamine in rat suprachiasmatic nucleus slices maintained *in vitro*. 5-Hydroxytryptamine (0.5 μM) was added in a 1-h static bath application at ZT6 (zeitgeber time) nominally on the preparatory day. One circadian cycle later, on the experimental day, firing rates of single units were derived from neuronal recordings in sequential electrode passes. An impartial observer determined peak activity visually. In this example, 5-hydroxytryptamine pretreatment shifted the peak to an earlier time point, a phase advance, 3.0 h before that noted in vehicle-treated slices (mean of nine experiments at dashed line) and similar to that obtained with 8-hydroxy-2-(di-*n*-propylamino) tetralin (10 μM). (Zeitgeber time is a 24-h notation of the phase in an entrained circadian cycle; ZT0 indicates the beginning of day, or the light phase, and ZT12 is the beginning of night, or the dark phase.) Adapted from Sprouse JS, Reynolds L, Braselton J, and Schmidt A (2004) Serotonin-induced phase advances of SCN neuronal firing *in vitro*: A possible role for 5-HT5A receptors? *Synapse* 54: 111–118, copyright 2004 John Wiley & Sons, Inc.

mice, suggesting the greater importance of this serotonin subtype independent of genetic modification. 5-HT at ZT6 also yielded phase advances that were indistinguishable in slices prepared from WT and KO mice and that were also sensitive to 5-HT_{1A} antagonism. Unlike the pattern with 8-OH-DPAT, however, the 5-HT-induced phase advances were blocked by ritanserin, in this paradigm useful as a 5-HT_{5A} antagonist (in addition to its more typical role as a 5-$HT_{2A/2C/7}$ antagonist). In sum, these results highlight the importance of the species used in establishing receptor mechanism, a rather well-worn observation. More provocatively, they support the involvement of multiple 5-HT receptors (5-HT_{1A} and 5-HT_{5A}, as well as 5-HT_7) in shifting the phase of circadian rhythms at ZT6.

Actions of SSRIs

Taking a step back from the goal of role-assigning for specific receptor subtype, other efforts have investigated the utility of selective serotonin reuptake inhibitors (SSRIs) in modulating rhythmicity, and in doing so modulating affective state. Fluoxetine, perhaps the prototypical SSRI, when added to SCN slices, has no effect on the timing of phasic activity. While initially puzzling, it was reasoned that the fluoxetine-induced phase shifts would be dependent on existing 5-HT in the slices, likely at a minimum after multiple hours in culture. L-Tryptophan, the amino acid precursor of 5-HT, when applied at ZT6, was subsequently found to produce phase advances similar to those observed with 5-HT agonists; co-incubation of a subthreshold concentration of this 5-HT synthesis-boosting agent with fluoxetine yielded the expected phase advances.

What is to be made of fluoxetine-induced phase advance of SCN neuronal firing? In terms of a clinical dosing strategy, patients receiving their total daily dose of fluoxetine in the morning versus in the evening differed little in their depression scores, and there have been no reported changes in circadian rhythm measurements that might track with symptom improvement. In all of these studies, however, dosing was performed in the morning or in the evening and did not include a midday time point. This is typical strategy in which drugs that tend to sedate are given before bedtime and those that more likely would stimulate are given at the start of the day. A comparison of optimal dosing times utilizing a noontime dose, in this case using chlomipramine, did establish it as being the most effective in a single clinical study, superior to morning or evening doses or to three equally divided doses. More work needs to be done; it is hoped that a greater appreciation of what might be accomplished will spur such activities. In making the considerable leap to the clinical setting, several other factors are of concern – the magnitude of the shift, the direction of the shift, the effect of chronic dosing – and it is clear that further laboratory experimentation, complemented by its clinical counterpart, is the key to greater understanding.

Circadian Rhythms and Mood Disorders

Over the years, evidence has accumulated suggesting that mood disorders such as major depressive disorder and bipolar disorder may be related to alterations in circadian clock functioning. In recent years there has been considerable effort to correlate mood regulation with the activity of individual clock genes. Moreover, some treatments for mood disorders such as seasonal affective disorder (SAD) successfully employ bright light therapy as a means of shifting or resetting the circadian clock; the magnitude of advance phase shifts correlates with the extent of mood improvement.

5-HT_{1B} Receptors and SAD

Seasonal affective disorder is a subtype of major depression most often characterized by autumn onset

of depressive symptoms that worsen throughout the winter and then dissipate as summer approaches. Patients suffering from SAD typically have a phase-delayed circadian system. Morning light therapy, especially if timed to maximize the corrective phase advance, has proved to be an effective treatment for SAD. The mechanisms underlying the pathophysiology of this subtype of depression are poorly understood, although it has been suggested that 5-HT plays a key role.

Very recently, it has been shown that p11 (S100A10), a member of the S100 protein family, interacts with 5-HT_{1B} receptors and p11 is decreased in brain tissue from depressed patients. p11 appears to participate in the translocation of proteins to the cell surface of neurons and in particular plays a role in the localization of the 5-HT_{1B} receptor to the presynaptic terminal membrane. p11 KO mice have reduced 5-HT_{1B} receptor function and exhibit a depression-like phenotype.

Although it is clear that activation of 5-HT_{1B} presynaptic receptors on RHT terminals reduces glutamate release, paradoxically, the SCN of 5-HT_{1B} receptor KO mice is less sensitive to light. This is probably due to an increase in GABAergic tone in the SCN resulting from the concomitant loss of 5-HT_{1B} receptors on GABA terminals and a reduction in 5-HT_{1B} receptor-mediated presynaptic inhibition of GABA release. Elimination of 5-HT_{1B} receptor function in the KO mouse results in an alteration in entrainment to the day/night cycle. When 5-HT_{1B} KO mice are maintained under short day (winterlike) photoperiod conditions, their circadian activity rhythm is significantly phase delayed compared to WT mice. The phase-delayed pattern of entrainment in mice lacking functional 5-HT_{1B} receptors under short day condition is similar to patients with SAD, and thus the 5-HT_{1B} receptor KO mouse may be a valuable model to study the relationship between circadian rhythms, 5-HT, and depression.

Chronopharmacology

Chronopharmacology, the strategy of exploiting in pharmacological terms what the clock has set in motion, is well accepted, as it is tied to the cyclical nature of biological phenomena and the rationale for pacing the delivery of drugs to optimize patient benefit. Remarkable success has been achieved in the timing of chemotherapy and the treatment of cardiovascular disease, and more widespread utility may be in the offing as the importance of peripheral circadian oscillators residing in various organ systems gains greater acceptance. The results presented here focus on a simple logical extension of this argument, the notion that the administration of drugs to directly alter SCN function might be of great importance. In doing so, two highly desirable outcomes might be obtained: an improvement in general timekeeping and, as a consequence, a facilitation of psychotherapeutic efficacy.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Circadian Function and Therapeutic Potential of Melatonin in Humans; Clock Gene Regulation of Endocrine Function; Clock Genes and Metabolic Regulation; Entrainment of Circadian Rhythms by Light; Genetics of Circadian Disorders in Humans; Photoreceptors and Circadian Clocks; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Sleep: Development and Circadian Control.

Further Reading

Belenky M and Pickard GE (2001) Subcellular distribution of 5-HT_{1B} and 5-HT_7 receptors in the mouse suprachiasmatic nucleus. *Journal of Comparative Neurology* 432: 371–388.

Hedlund PB and Sutcliffe JG (2004) Functional, molecular and pharmacological advances in 5-HT_7 receptor research. *Trends in Pharmacological Sciences* 25: 481–486.

Jacobs BL and Azmitia EC (1992) Structure and function of the brain serotonin system. *Physiological Reviews* 72: 165–229.

Klein DC, Moore RY, and Reppert SM (1991) *Suprachiasmatic Nucleus: The Mind's Clock*. New York: Oxford University Press.

McClung CA (2007) Circadian genes, rhythms and the biology of mood disorders. *Pharmacology & Therapeutics* 114: 222–232.

Morin LP (1999) Serotonin and the regulation of mammalian circadian rhythmicity. *Annals of Medicine* 31: 12–33.

Morin LP and Allen CN (2006) The circadian visual system, 2005. *Brain Research Reviews* 51: 1–60.

Nagayama H, Nagano K, Ikezaki A, et al. (1991) Double-blind study of the chronopharmacotherapy of depression. *Chronobiology International* 8: 203–209.

Pickard GE, Smith BN, Belenky M, et al. (1999) 5-HT_{1B} receptor-mediated presynaptic inhibition of retinal input to the suprachiasmatic nucleus. *Journal of Neuroscience* 19: 4034–4045.

Sollars PJ, Ogilvie MD, Simpson AM, et al. (2006) Photic entrainment is altered in the 5-HT_{1B} receptor knockout mouse. *Journal of Biological Rhythms* 21: 21–32.

Sprouse JS (2004) Pharmacological modulation of circadian rhythms: A new drug target in psychotherapeutics. *Expert Opinion in Therapeutic Targets* 8: 25–38.

Sprouse JS, Braselton J, and Reynolds L (2006) Fluoxetine modulates the circadian biological clock via phase advances of suprachiasmatic nucleus neuronal firing. *Biological Psychiatry* 60: 896–899.

Sprouse JS, Reynolds L, Braselton J, and Schmidt A (2004) Serotonin-induced phase advances of SCN neuronal firing *in vitro*: A possible role for 5-HT5A receptors? *Synapse* 54: 111–118.

Svenningsson P, Chergui K, Rachleff I, et al. (2006) Alterations in 5-HT_{1B} receptor function by p11 in depression-like states. *Science* 311: 77–80.

Terman M and Terman JS (2005) Light therapy. In: Kryger MN, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, 4th edn., pp. 1424–1442. Philadelphia: Elsevier.

Thomas DR (2006) 5-ht5A receptors as a therapeutic target. *Pharmacology & Therapeutics* 111: 707–714.

Serotonin-Related Psychedelic Drugs

M A Geyer, University of California, San Diego, CA, USA
D E Nichols, Purdue University, West Lafayette, IN, USA
F X Vollenweider, University of Zurich, Zurich, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Historical Introduction

Hallucinogen use by humans dates back at least two millennia and continues in virtually all cultures today. The vast majority of human experience with these compounds derives from the ingestion of plant materials; only recently have the chemical constituents of these plants been identified, synthesized, and ultimately modified and elaborated. Traditional uses of hallucinogens were typically characterized by sacramental ritual uses appropriate to the particular culture. With the advent of modern chemistry, potent hallucinogenic compounds are more widely available and used in more diverse ways. Although several chemical classes are now understood to be represented among the large family of hallucinogenic agents, the present focus is on those whose primary mechanism of action is believed to be mediated by actions on serotonergic neuronal systems.

Defining Phenomenology

Fundamentally, the category of hallucinogens is defined only by the subjective reports by humans of various distortions of perception and affect. Specifically, serotonergic hallucinogens such as lysergic acid *N*, *N*-diethylamide (LSD), mescaline, ayahuasca, and psilocybin increase the intensity and lability of affective responses and produce profound distortions of perceptual processes encompassing visual, auditory, somesthetic, and olfactory modalities. Hallucinogens produce profound changes in mood, thought, intuition, and sensory perception, including changes in the experience of time, space, and self that are otherwise rarely experienced except in dreams. Depending upon person and circumstance, the same hallucinogen may, on the one hand, produce a loss of ego boundaries including heightened mood ranging from pleasure to bliss, and feelings of oneness with the universe, or may, on the other hand, lead to more psychotic ego dissolution, including fear and paranoid ideation associated with experiences of split ego. In these states, perceptual hypersensitivity, illusions, and elementary hallucinations are common. In general, the intensity of these alterations of consciousness is dose dependent, so that hallucinations involving cognitive impairment or disorientation in person, place, and time rarely, if at all, occur with low to medium doses.

Altered States of Consciousness

According to the work of Dittrich, the common nucleus of hallucinogen-induced altered states of consciousness (ASCs) can be described and assessed by three primary dimensions: (1) 'oceanic boundlessness' (OB), referring to positively experienced ego dissolution; (2) 'anxious ego dissolution' (AED), including thought disorder and loss of self-control; and (3) 'visionary restructuralization' (VR), referring to perceptual alterations and altered meaning of percepts. Recent brain imaging studies demonstrate that the degree to which each of these key dimensions of ASCs is expressed across individuals correlates with functional alterations in segregated neuronal networks, including prefrontal, temporal, and parietal cortical regions as well as the basal ganglia and thalamus.

Chemical Structures

For the classical serotonergic hallucinogens, which presumably exert their psychedelic actions predominantly through a 5-HT_{2A} receptor mechanism, molecules are comprised of a basic amino group, separated by two carbon atoms from an aromatic system. This motif reflects the fact that serotonin and the other monoamine neurotransmitters arise through biosynthesis from tryptophan and tyrosine. Thus, the hallucinogens embody the same carbon framework as the natural transmitters. There are two major classes of serotonergic hallucinogens: the phenethylamines and the tryptamines. The tryptamines can further be divided into the simple tryptamines and the tetracyclic ergolines that include LSD and related lysergic acid amides.

Phenethylamines

The simplest phenethylamine hallucinogen is mescaline, which occurs naturally in the peyote cactus, *Lophophora williamsii*. Apparently the only naturally occurring phenethylamine hallucinogen, the potency of mescaline is low, requiring a dose on the order of 250–500 mg for a full effect lasting 8–10 h. Despite its low potency, however, mescaline has served as a molecular template for the chemical synthesis of hundreds of structural modifications that have been made and tested.

Early modifications involved adding an alkyl group to the side chain, to provide TMA, which is about

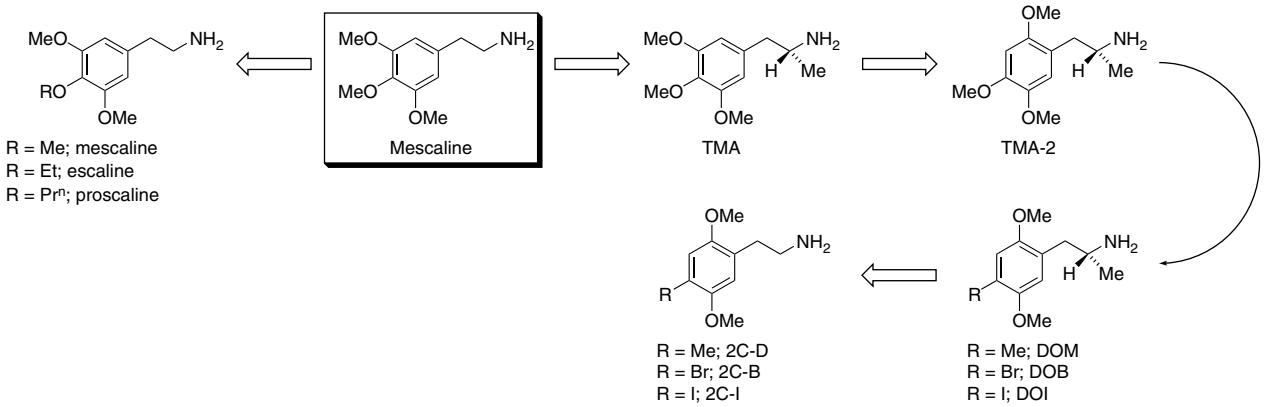

Figure 1 Mescaline and related phenethylamine hallucinogens.

twice as potent as mescaline. Transposing one of the distal methoxy groups to an *ortho*-position gave TMA-2, which was an order of magnitude more potent than mescaline. These alpha-methyl-substituted compounds are often referred to as 'hallucinogenic amphetamines.' Finally, replacing the 4-alkoxy of TMA-2 with alkyl, alkylthio, or halogen groups gives the most potent phenethylamine-type compounds, some of which are two orders of magnitude more potent than mescaline itself. The introduction of the alpha-methyl group to provide the hallucinogenic 'amphetamines' also introduces a center of chirality. It is known that the *R*-(–) isomer, as shown in Figure 1, is the more potent one. This *R* stereochemistry is opposite to psychostimulant amphetamines, where the *S*-(+) 'dextro' isomers are more potent. In general, substituted amphetamines that are substrates for and inhibitors of monoamine transporters have stereochemistry like *S*-(+)-amphetamine. Because the target of the hallucinogenic amphetamines is a postsynaptic receptor, the different stereochemistry is not surprising.

A variety of 'designer' phenethylamines that lack the alpha-methyl group have appeared on the illicit market. Except for the 2,4,5-trimethoxy compound itself, virtually everything that has a more hydrophobic group at the 4-position has been found to possess oral activity. For example, as shown in Figure 2, several 2C-T compounds have been reported, where the 4-position is an *S*-alkylthio compound. Another example that made frequent appearance on the illicit market was the 4-bromo analog, known as 2C-B (Figure 1). Generally, these phenethylamines are somewhat less potent than the corresponding amphetamine, and of shorter duration of action. They do not, however, produce qualitatively the same type of pharmacology as their amphetamine counterparts.

The most extensive studies of structure–activity relationships of the phenethylamine-type hallucinogens have involved variations in aromatic ring substituents.

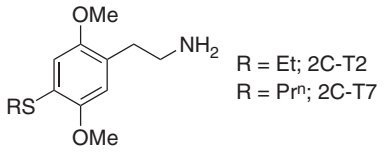

Figure 2 Phenethylamine hallucinogens that lack the alpha-methyl group.

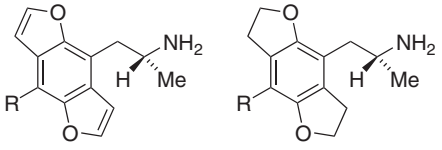

Figure 3 Phenethylamine hallucinogens of the benzofurano type.

Methoxy groups at the 2- and 5-positions of the aromatic ring are optimal. Extension of either methoxy group by one carbon, to an ethoxy group, markedly reduces or abolishes activity. The one caveat to this conclusion is that the methoxy groups can be incorporated into heterocyclic oxygen rings. Indeed, the benzofurano-type compounds shown in Figure 3 are the most potent phenethylamine hallucinogens reported to date, with the fully aromatic compounds having K_i values at cloned human 5-HT_{2A} receptors of <0.10 nM. Human data for many phenethylamines have been reported by the Shulgins and also on some Internet sites. In contrast to the relatively low potency of mescaline, many phenethylamines have effective dosages between 1 and 10 mg, with most being orally active and having durations of action ranging from a few hours to as much as 24 h or more.

As illustrated in Figure 1, a large variety of different substituents can be placed at the R position. Of the simple halogens, bromine (DOB) and iodine (DOI) are the most potent, whereas the highly electronegative

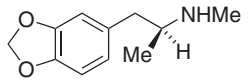

MDMA

Figure 4 3,4-Methylenedioxymethamphetamine (MDMA; 'ecstasy').

group CF_3 is even more potent. Alkylthio groups also have given very potent and interesting compounds.

Although 3,4-methylenedioxymethamphetamine (MDMA; ecstasy) is not properly classified as a hallucinogen, it is relevant here because it is a substituted phenethylamine derivative and is in the same legal class as the hallucinogens (Figure 4). In contrast to the hallucinogenic amphetamines, however, which have a direct receptor agonist action, MDMA is an indirect-acting agent that causes release of endogenous serotonin, dopamine, and norepinephrine from neuron terminals. The mechanism appears to be similar to amphetamine itself, and is believed to be related to a reverse transport of the monoamine uptake transporter, leading to drug-induced efflux of monoamine from the neuron. Although the primary behavioral effects of MDMA are attributable to the release of presynaptic serotonin and consequent activation of 5-HT_{1B} receptors, it is clear that the release of dopamine also contributes. There has been very little study of the role of norepinephrine, but it is also robustly released by MDMA.

The mechanism of action for MDMA clearly differs from that of the hallucinogenic amphetamines. That conclusion is evident initially from a chemical perspective. First of all, in the hallucinogenic amphetamines, the isomer with the *R* configuration in the side chain is most active. By contrast, it is the *S* enantiomer (shown) of MDMA that is more active, the same stereochemistry that is most active in the psychostimulants amphetamine and methamphetamine. Second, when an *N*-methyl is attached to the amine of a hallucinogenic amphetamine, there is a dramatic and near-total loss of activity. Yet MDMA possesses this feature, and retains about half the potency of the primary amine analog MDA. Finally, if the alpha-methyl group in the side chain of the hallucinogenic amphetamines is lengthened by one carbon atom to an ethyl group, hallucinogenic activity is completely abolished. If the alpha-methyl of MDMA is changed to an ethyl group, however, about 70% of the potency of MDMA is retained.

Tryptamines

The tryptamines bear the closest resemblance to the neurotransmitter serotonin, as seen in Figure 5.

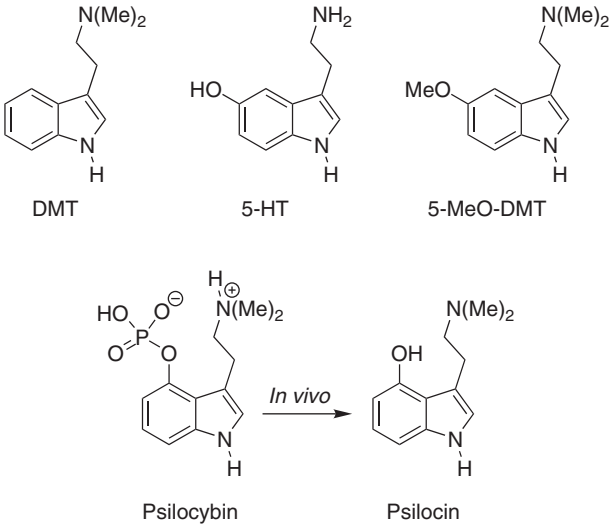

Figure 5 Serotonin and the major tryptamine-based hallucinogens.

A number of these substances are found in nature. The simplest is *N,N*-dimethyltryptamine (DMT), which is found in a number of plant species, and also has been detected in human cerebrospinal fluid. It is not active orally, being degraded by monoamine oxidase (MAO) in the liver, and is usually smoked or injected, where doses of 50–75 mg would be typical, with transient effects. There is an orally active preparation of DMT known as ayahuasca, or yagé, which has been used by Indians of the Amazonian basin in Brazil. Ayahuasca is a decoction made from two different plants, *Banisteriopsis capii*, a source of beta-carboline alkaloids, and DMT-containing species such as *Psychotria viridis*. The beta-carbolines include harmine, harmaline, and tetrahydroharmine, and are reversible MAO-A inhibitors. It is thought therefore, that ayahuasca has pharmacological activity because the harmala alkaloids inhibit the MAO in the liver that would normally deaminate the DMT following oral ingestion.

Another tryptamine that occurs in nature is 5-methoxy-*N,N*-dimethyltryptamine (5-MeO-DMT). Like DMT, it is not orally active. It occurs naturally in a number of species of *Virola* that grow in central and South America. The red exudate from the inner surface of bark off these trees is dried and finely ground to produce a powerful snuff called epená. The effects of both DMT and 5-MeO-DMT generally last less than 30 min.

In contrast to DMT and 5-MeO-DMT, psilocybin is orally active in the 6–20 mg range, with a duration of action of 4–6 h. Psilocybin itself serves as a prodrug, as serum phosphatases cleave the *O*-phosphate to give psilocin, the active moiety. Although there are

a few other substituted tryptamines that are known to be hallucinogenic, for example α-methyltryptamine (AMT), in general there is little information available on their human psychopharmacological properties, metabolism, etc. A number of anecdotal human reports have been presented by Shulgin and Shulgin.

LSD and Other Ergolines

The tetracyclic ergoline *N,N*-diethyllysergamide (LSD; LSD-25) is the most potent of the known hallucinogens and is shown in Figure 6. A typical clinically effective oral dose of the tartrate salt is 0.10 mg, with effects lasting for 8–10 h. LSD is a semisynthetic material, prepared by the condensation of (+)-lysergic acid with diethylamine. Lysergic acid is obtained by basic hydrolysis of ergot alkaloids (e.g., ergotamine), which are obtained from any of a variety of *Claviceps* (ergot) species.

The stereochemistry illustrated here is critical for activity, and if the chirality at either the 5- or the 8-position is reversed, activity is lost. In contrast to the phenethylamine hallucinogens, the structural requirements for activity in the ergolines are narrowly restricted. No other amide group, other than diethyl, has the potency of LSD. Isosteric groups such as methyl-isopropyl, or methyl-propyl, or heterocyclic rings such as pyrrolidine, all fail to provide compounds with more than about 10% of the potency of LSD, at least in humans. An isomeric series of azetidines, four-membered nitrogen-containing heterocyclic rings with two methyl groups attached, has been used to show that the diethyl group must adopt a very specific stereochemical orientation when LSD binds to its receptors. This knowledge, coupled with examination of homology models for the 5-HT_{2A} receptor, suggests that there is a very specific amino acid motif in the area where the diethylamide binds. This topographically defined region apparently has steric and spatial requirements that are uniquely complementary to the diethylamide of LSD.

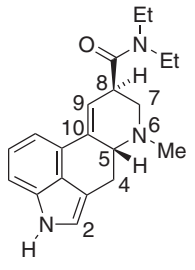

LSD

Figure 6 The ergoline-derived hallucinogen lysergic acid *N,N*-diethylamide (LSD).

Psychopharmacology in Animals

Serotonin Receptor Mechanisms

Studies in the 1950s indicated that hallucinogens had profound effects on serotonin in the brain. The recognition that both LSD and serotonin were built on the same tryptamine scaffold led to early hypotheses that LSD was affecting brain serotonin neurotransmission. It was only after radioligand binding techniques were developed in the 1970s that it was discovered that hallucinogens bound with high affinity to specific serotonin receptor subtypes and subsequently that all the chemical classes of hallucinogens activate serotonin 5-HT_{2A} receptors, a member of the Gαq-coupled family of G-protein-coupled receptors.

The earliest mechanistic evidence for a serotonergic action of psychedelics came from George Aghajanian's laboratory, where it was found that LSD and tryptamine hallucinogens suppressed firing of serotonergic cells in the dorsal raphe nucleus. The correlation between raphe cell firing and general level of vigilance, and the cessation of firing during rapid-eye-movement (REM) sleep, were intriguing phenomena that seemed to identify raphe cells as likely targets for hallucinogens. Subsequent studies indicating that some ligands that suppressed raphe firing were not hallucinogenic led to the abandonment of this hypothesis. When subtype-selective serotonin receptor ligands were finally developed, it was found that suppression of raphe cell firing was mediated by somatodendritic serotonin 5-HT_{1A} receptors. Ideas then moved to a focus on other serotonin receptor subtypes and other brain areas. Of greatest interest currently are the 5-HT_{2A} receptors in multiple regions such as the thalamus and frontal cortex. Nevertheless, several experimental studies in animals indicate that LSD and other tryptamines may have some subtle psychopharmacological properties that could be attributed to their agonist actions on 5-HT_{1A} receptors. Unfortunately, the well-controlled clinical studies employing specific 5-HT_{1A} receptor ligands necessary to elucidate these issues in humans have not been conducted.

Behavioral Effects in Animals

Given the complexity, variety, and variability of the effects of psychedelics in humans, it has been difficult to define animal behavior tests for hallucinogenic activity. In animals, psychedelics (1) potentiate neophobia and other anxiety-related behaviors, (2) increase responsiveness to sensory stimulation, (3) retard habituation in a variety of input modalities and

response output systems, and (4) induce a variety of abnormal behaviors such as ear scratches and head twitches. Using various behavioral paradigms in rodents, researchers have identified the 5-HT_{2A} subtype of serotonin receptors as being responsible for the major psychological effects of the classical hallucinogens. These behavioral paradigms included measures based on abnormal behaviors induced by hallucinogens, alterations within the normal range of unconditioned behaviors such as locomotor activity or habituation to sensory stimuli, changes in simple forms of classical or instrumental learning, and assessments of the discriminative stimulus properties of psychoactive drugs. Despite their limitations, animal models of the effects of hallucinogens enabled the evaluation of several hypotheses regarding their neurochemical substrates of action. Indeed, the major contributions of $5HT_{2A}$ receptor agonism to the behavioral effects of hallucinogens and the contrasting importance of serotonin release to MDMA effects were first demonstrated using animal behavioral models and more recently confirmed by studies in humans. Thus, clinical studies demonstrated that – as was originally discovered in animal studies – the selective 5-HT_{2A} receptor antagonist ketanserin, but not serotonin reuptake inhibitors, blocks the psychopharmacological effects of the hallucinogen psilocybin in human volunteers. Similarly, confirming the earlier studies of locomotor activity in rats and mice, serotonin reuptake blockers such as citalopram block the psychological effects of MDMA in humans, whereas 5-HT_{2A} receptor antagonists have minimal effects. These results are the most compelling and conclusive proof of 5-HT_{2A} receptor mediation of the effects of hallucinogens.

Interactions with Other Systems

Although 5-HT_{2A} receptors appear to be the most important, the activation of 5-HT_{1A} receptors also has been shown to influence many of the same behavioral end points and often contribute to and/or interact with the impact of hallucinogens on 5-HT_{2A} receptors. The contributions of activation of 5-HT_{2C} receptors have not been assessed adequately in the absence of a hallucinogen that has selective effects at the 5-HT_{2A} receptor, relative to the 5-HT_{2C} receptor. Possible contributions of the 5-HT_4, 5-HT_5, 5-HT_6, or 5-HT_7 receptors also await additional study. Nevertheless, one could make the argument that activation of the 5-HT_{2A} receptor is a necessary, but not sufficient, condition for producing hallucinogenic effects, especially with regard to the mechanism of action of phenethylamine and simple tryptamine hallucinogens. It should be noted, however, that the pharmacology of LSD may be more complex. Although LSD is known to have affinity for the majority of serotonin receptor subtypes, the receptor affinities and intrinsic activity of LSD are unremarkable. In particular, the receptor affinity of LSD for the 5-HT_{2A} and 5-HT_{2C} receptor is essentially identical to that of simple phenethylamines such as DOB or DOI. Furthermore, the intrinsic activity of these phenethylamines is on the order of about 80%, whereas that of LSD is less than 50%. Yet, LSD is more than an order of magnitude more potent than these phenethylamines. Therefore, the question "why is LSD so potent?" has often been asked. There has been no definitive answer as yet, but it should be noted that LSD has relatively high affinities and intrinsic activities for the dopamine D_2-like family of receptors.

Psychopharmacology in Humans

In addition to the defining aspects of the effects of serotonergic hallucinogens discussed above, experimental neurophysiological and neuropsychological studies confirm that hallucinogens produce dose-dependent cognitive impairments in healthy subjects, including deficits in attention, working memory, and associative learning, while leaving executive functions almost unaffected. Serotonergic hallucinogens also impair high-level but not low-level motion perception and reduce binocular rivalry rate and rhythmicity in a manner reflecting subjective changes in conscious state alterations that also have been reported in bipolar disorder and schizophrenia.

Relevance of Psychedelics to Clinical Pharmacotherapeutics

Studies of the mechanisms responsible for the psychotomimetic aspects of serotonergic hallucinogens have provided important insights into the neurobiology of schizophrenia and related psychotic disorders. Indeed, recent research into hallucinogenic drug action has revitalized alternative hypotheses to the classic dopamine hyperactivity model of schizophrenia. Historically, the serotonin hypothesis of schizophrenia derived from the early studies of the pharmacology of LSD and its structural relationship to serotonin. This hypothesis has received renewed support because indoleamine hallucinogens produce their psychological effects predominantly via activation of 5-HT_{2A} receptors, 5-HT_{2A} receptors are abnormal in the brains of schizophrenic patients, and 5-HT_{2A} receptor antagonism appears to contribute substantially to the therapeutic effects of atypical antipsychotics such as clozapine.

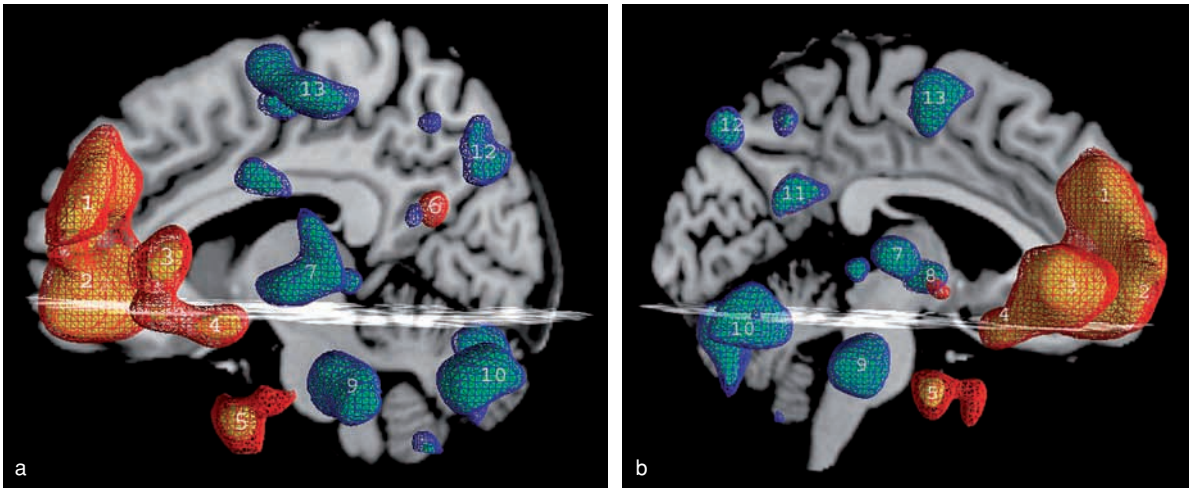

Figure 7 (a,b) The effects of a 5-HT_{2A} agonist, psilocybin, in healthy human volunteers as measured by changes of cerebral blood flow (CBF) using H_2O-PET (positron emission tomography). Red is a relative increase as indexed by H_2O-PET (CBF); blue is a relative decrease in brain activity. Psilocybin alters neuronal activity in key components of limbic cortico-striato-thalamic pathways (CSTC loops), which are important for gating or filtering of extrinsic and intrinsic information flow to the cortex. Disruption of the functional integrity of CSTC loops and subsequently reduced thalamic gating may explain at least in part the sensory flooding and psychotic symptom formation reported by individuals during psilocybin-induced altered states. Marked increases in regional activity were found in areas that are important for both cognitive and affective processes such as the frontomedial cortex extending into the anterior cingulate (1,2), the dorsolateral (3), insula (4), and temporal cortices (5), as well as the left posterior cingulate (6). In comparison, decreased cerebral blood flow was found in brain areas that are important for gating or integrating cortical information processing such as bilateral thalamus (7), right globus pallidus (8), bilateral pons (9), and in cerebellum (10). Finally, psilocybin also reduced neuronal activity in components responsible for higher-order visuospatial processing such as precuneus (12) and angular gyrus (11), as well as in supplementary eye fields of the premotor area (13). In combination, psilocybin alters brain activation in such a way as to increase cognitive and affective processing in the context of reduced gating and reduced focus on external stimulus processing. Unpublished data from FX Vollenweider.

Several authors have argued that hallucinogen-induced ASCs share multiple phenomenological features with early acute stages of schizophrenia and may provide useful models to elucidate the neuronal basis of productive symptoms and cognitive deficits of schizophrenic pathophysiology. Recent theories about the neurobiology of ASCs suggest that deficits in basic information processing may be elaborated into the diverse psychotic symptoms and cognitive disturbances seen in both drug-induced model psychoses and naturally occurring psychoses such as schizophrenia. For example, one fundamental feature of information processing dysfunction in both hallucinogen-induced states and schizophrenia-spectrum disorders is the inability of these subjects to inhibit, filter, or gate extraneous stimuli and to attend selectively to important features of the environment. Extensive neurobiological studies have identified cortico-striato-thalamic pathways (CSTC loops) that subserve major aspects of gating functions in mammals. As shown in Figure 7, acute treatment with psilocybin alters cerebral blood flow within the CSTC gating circuits in healthy human volunteers. Gating deficits may cause both psychedelic-treated and schizophrenia subjects to become overloaded with excessive exteroceptive and interoceptive stimuli and subsequently disrupt normal cognitive integrity. Impaired gating and loosening of ego boundaries could lead to positive symptoms such as delusions, hallucinations, thought disturbances, persecution, and the loss of a coherent ego experience. In addition, various negative symptoms, such as emotional and social withdrawal, could reflect efforts to protect the individual from input overload. Nevertheless, despite the number of similarities between hallucinogen-induced models of psychoses, substantial differences also are apparent in the limited human studies. One of the most striking of these differences involves the essentially lifelong nature of schizophrenia disorders, which contrasts with the rapidity with which tolerance develops to the psychological effects of serotonergic hallucinogens.

Possible Therapeutic Applications of Psychedelics

In the 1950s and early 1960s, LSD was hailed as an important new technology that had the potential to revolutionize psychiatry. Parallels were often drawn between the discovery of LSD and the discovery of atomic energy. Thus, there were more than a thousand published clinical studies of over 40 000 subjects. Unfortunately, the unbridled enthusiasm for the value of LSD most often meant that clinical studies were

poorly designed, poorly controlled, and poorly analyzed, with inadequate posttreatment follow-up. As a result, the widespread impression remains that LSD was found to lack any therapeutic value.

In the absence of well-designed studies, the only conclusion that can safely be drawn is that we really do not know what medical value LSD or other hallucinogens might have. Like any new technology, and here we might usefully employ the analogy to atomic energy, it may take some time to discover how best to use it. Certainly, the use of radium to produce luminescent watch and clock dials was a trivial one and did not reflect the great potential that the atom represented. Similarly, we might speculate that LSD might have important medical uses if the time and effort were spent to investigate its potential.

What leaves us curious are the many anecdotal reports of individuals who used LSD and experienced profound sorts of unexpected healing, from a variety of disorders. Many alcoholics claim to have stopped drinking, and individuals with substance abuse problems have reported losing their desire to continue their substance use. These reports fail to meet the test as science, because they were case reports, or personal essays, and either were not replicated or could not be replicated.

The most well-documented value of hallucinogens as therapeutics was in the treatment of dying cancer patients. Starting with the work of Chicago internist Eric Kast and ending in several studies at the Baltimore State Hospital, LSD treatment in a program that included only a modest amount of counseling produced some degree of benefit in 60–70% of patients, measured as a reduced need for pain medication and improved mood. The patients who had the most profound LSD experiences gained the most therapeutic benefit. The overall conclusion is that the LSD treatment program appeared to have led to a reduced fear of death, and thus a reduction in the anxiety normally caused by impending death. Recently, a small number of new clinical studies have begun to reexamine the possible clinical benefits of serotonergic hallucinogens. With modern clinical designs and appropriate controls, there is a need to assess whether or not hallucinogens do represent useful tools for psychiatry. This chapter of history has not been closed, and hopefully future clinicians will make the efforts necessary to gain the answers to this question once and for all.

In another context, explored primarily in Europe, hallucinogens have been suggested to be useful tools in so-called psycholytic or hallucinogen-aided psychotherapy, given that they have the potential to improve the patient–therapist relationship, to free the patient from autistic fixation and isolation, and to call up forgotten and repressed unconscious material. Under the effects of hallucinogens, repressed memory traces, even those going back to early childhood and which may often contribute to psychiatric disorders, can enter vividly into consciousness. In the context of psychodynamic concepts, the performance of psycholytic therapy is similar to the imaginative techniques, such as guided affective imagery, in terms of setting, the interaction between the patient and therapist, and the clinically relevant phenomena. Nevertheless, without clinical supervision, patients and nonpatients should not experiment with substances having such deep and profound effects on mind and body, for many of the effects, their intensity, duration, and potential after effects, can in no way be foreseen. From a therapeutic perspective, such uncontrolled use of hallucinogens is dangerous.

Conclusions

Serotonergic hallucinogens, the classical psychedelic compounds, have profound effects on human consciousness. These substances have been used ceremonially for many centuries in many cultures. Neurochemists and behavioral neuroscientists have identified agonist actions at specific subtypes of serotonin receptors, primarily the 5-HT_{2A} receptors, as being the primary mechanism of action for the behavioral effects of psychedelics. More recent clinical studies of human volunteers have confirmed that the psychedelic effects of hallucinogens such as psilocybin are largely attributable to the activation of 5-HT_{2A} receptors. The study of serotonergic hallucinogens has been informative as to the neurobiological substrates of ASCs and likely mechanisms contributing to some psychiatric disorders such as schizophrenia. Further research in this area could lead to the identification of novel pharmacotherapeutics, potentially including therapeutic effects of the hallucinogens themselves in certain conditions.

See also: 3,4-Methylenedioxymethamphetamine (MDMA, "Ecstasy"); Cells: 5-Hydroxytryptamine Receptors; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Serotonin (5-Hydroxtryptamine; 5-HT): Neurotransmission and Neuromodulation; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Serotonin (5-Hydroxtryptamine; 5-HT): Receptors.

Further Reading

Aghajanian GK and Marek GJ (1999) Serotonin and hallucinogens. *Neuropsychopharmacology* 21: 16S–23S.

Carter OL, Burr DC, Pettigrew JD, et al. (2005) Using psilocybin to investigate the relationship between attention, working

memory, and the serotonin 1A and 2A receptors. *Journal of Cognitive Neuroscience* 17: 1497–1508.

Dittrich A (1998) The standardized psychometric assessment of altered states of consciousness (ASCs) in humans. *Pharmacopsychiatry* 31: 80–84.

Fischer R (1971) A cartography of the ecstatic and meditative states. *Science* 174: 897–904.

Geyer MA (1998) Behavioral studies of hallucinogenic drugs in animals: Implications for schizophrenia research. *Pharmacopsychiatry* 31(supplement 2): 73–79.

Geyer MA and Callaway CW (1994) Behavioral pharmacology of ring-substituted amphetamine analogs. In: Cho AK and Segal DS (eds.) *Amphetamine and Its Analogs: Neuropsychopharmacology Toxicology and Abuse*, pp. 177–208. New York: Academic Press.

Geyer MA and Krebs KM (1994) Serotonin receptor involvement in an animal model of the acute effects of hallucinogens. In: Lin G and Glennon RA (eds.) *NIDA Research Monograph Series: Hallucinogens: An Update*, pp. 124–156. Rockville, MD: US Department of Health and Human Services.

Glennon RA (2003) Higher-end serotonin receptors: 5-HT(5), 5-HT(6), and 5-HT(7). *Journal of Medicinal Chemistry* 46: 2795–2812.

Gouzoulis-Mayfrank E, Habermeyer E, Hermle L, et al. (1998) Hallucinogenic drug induced states resemble acute endogenous psychoses: Results of an empirical study. *European Psychiatry* 13: 399–406.

Hofmann A (1968) Psychotomimetic agents. In: Burger A (ed.) *Chemical Constitution and Pharmacodynamic Actions,* 2nd edn., pp. 169–235. New York: Marcel Dekker.

Leuner H (1994) Hallucinogens as an aid in psychotherapy: Basic principles and results. In: Pletscher A and Ladewig D (eds.) *50 Years of LSD: Current Status and Perspectives of Hallucinogen Research*, 1st edn., pp. 175–189. New York: The Parthenon Publishing Group.

Mandell AJ and Geyer MA (1980) The euphorohallucinogens – a study in hemispheric relating. In: Freedman AM, Kaplan HI, and Sadock BJ (eds.) *The Comprehensive Textbook of Psychiatry III*, pp. 586–599. Baltimore: Williams & Wilkins.

Nichols DE (1997) Role of serotoninergic neurons and 5-HT receptors in the action of hallucinogens. In: Baumgarten HG and Gothert M (eds.) *Serotoninergic Neurons and 5-HT Receptors in the CNS*, pp. 563–585. Berlin: Springer.

Nichols DE (2004) Hallucinogens. *Pharmacology & Therapeutics* 101: 131–181.

Shulgin A and Shulgin A (1991) *PIHKAL. A Chemical Love Story.* Berkeley: Transform Press.

Shulgin AT and Shulgin A (1997) *TIHKAL. The Continuation.* Berkeley: Transform Press.

Strassman RJ, Qualls CR, Uhlenhuth EH, and Kellner R (1994) Dose-response study of *N,N*-dimethyltryptamine in humans. II: Subjective effects and preliminary results of a new rating scale. *Archives of General Psychiatry* 51: 98–108.

Vollenweider FX and Geyer MA (2001) A systems model of altered consciousness: Integrating natural and drug-induced psychoses. *Brain Research Bulletin* 56: 495–507.

Vollenweider FX, Remensberger S, Hell D, and Geyer MA (1999) Opposite effects of 3,4-methylenedioxymethamphetamine (MDMA) on sensorimotor gating in rats versus healthy humans. *Psychopharmacology* 143: 365–372.

Vollenweider FX, Vollenweider-Scherpenhuyzen MFI, Bäbler A, Vogel H, and Hell D (1998) Psilocybin induces schizophrenia-like psychosis in humans via a serotonin-2 agonist action. *NeuroReport* 9: 3897–3902.

Relevant Website

http://www.erowid.org – Erowid: Documenting the Complex Relationship between Humans and Psychoactives.

Sexual Behavior: Neuroendocrine Control

D Pfaff, N Vasudevan, and L-M Kow, The Rockefeller University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Female Behavior

The most prominent, biologically crucial, strongly sexually differentiated behavior in mammals is lordosis behavior (Figure 1), the vertebral dorsiflexion response by which female quadrupeds permit reproduction. Lordosis is governed by a limbic–hypothalamic system of estrogen-binding neurons (Figure 2), a system proven to be universal among mammals. Lordosis depends on a spino–midbrain–spinal neural circuit (Figure 3), which explains the physiology of the behaviorally adequate sensory stimulation and the production of the motor response of lordosis.

Later immunocytochemical analyses identified the contributions of two different gene products: those due to the estrogen receptor (ER)-α gene and those due to the ER-β gene. In the ventromedial hypothalamus, a cell group containing neurons crucial for lordosis behavior, the number of ER-α-expressing neurons far exceeds that expressing ER-β.

Because estrogen receptors are ligand-activated transcription factors, estrogen actions in the hypothalamic neurons that govern lordosis could be used to generate a list of genes that have two properties: Their transcript levels are elevated by estrogens, and their gene products foster lordosis behavior (Figure 4). It can be seen from Figure 4 that these genes encode many different kinds of biochemicals: growth-related products, another transcription factor (the progesterone receptor), neurotransmitter receptors, neuropeptides, and their receptors. Nevertheless, we have argued that the genes identified so far encode products that participate in functional modules ensuring the biological adaptiveness of reproductive behaviors.

Overall, the lordosis behavior neural circuit – the first worked out for a mammalian behavior – and the corresponding molecular mechanisms prove that specific biochemical reactions in specific parts of the mammalian brain produce a specific, biologically important behavior.

Membrane-Initiated Hormone Actions Potentiate Genomic Actions

Historically, biochemical and physiological actions of steroid hormones were laid exclusively to effects in the cell nucleus on gene transcription. However, it has become clear that some steroid hormone effects on electrical activity of neurons occur too rapidly to be dependent on genomic actions. In the past, scientists argued that either the rapid, membrane-initiated actions or the slow nuclear actions were most important. Vasudevan resolved this argument by showing that rapid, membrane-initiated effects of estrogens actually potentiate the cell nuclear actions of estrogens on gene transcription. Thus, there now exists a unified theory of steroid hormone actions on neurons.

What might this mean for the facilitation of lordosis behavior? In the ventromedial nucleus of the hypothalamus (Figure 2), which is crucial for facilitating lordosis, estrogens have at least three ways to rapidly (within seconds or minutes) increase electrical activity: increasing inward, excitatory ionic currents; decreasing outward, inhibitory ionic currents; and amplifying excitatory responses to arousal-related neurotransmitters such as histamine and glutamate. The signal transduction pathways potentially involved in these various rapid actions of estrogens and important for lordosis behavior are partially known (Table 1).

In summary, estrogens rapidly heighten electrical activity in nerve cells crucial for lordosis and amplify gene transcription actions of estrogens. Therefore, it must be concluded that rapid, membrane-initiated effects of estrogens cooperate with slower genomic effects in order to facilitate lordosis behavior.

Reverse Engineering the Lordosis Behavior Circuit

Having worked out the major mechanisms that produce lordosis behavior (hormonal, neuronal, and genomic), we can now look back on these mechanisms and ask what engineering design principles they manifest. 'Reverse engineering' asks the following question: Why was that particular set of neuronal connections (or, for example, that particular action of a hormone on a gene's transcription) evolved that way, with the eventual effect of producing biologically adaptive, well-regulated mating behaviors?

In the small laboratory animals in which lordosis mechanisms have been worked out, mating behaviors (which may involve exposure to predators as well as the expenditure of metabolic energy) should only occur when the female is about to ovulate so that the mating will produce the young animals of the next generation. Years ago, we found that the neuropeptide gonadotropic-releasing hormone (GnRH, or leutinizing hormone-releasing hormone), whose main job is to trigger ovulation, also facilitates lordosis behavior; thus the 'unity of the body' – the physiology of peripheral reproductive organs coordinated with the brain's control of mating behavior. Furthermore,

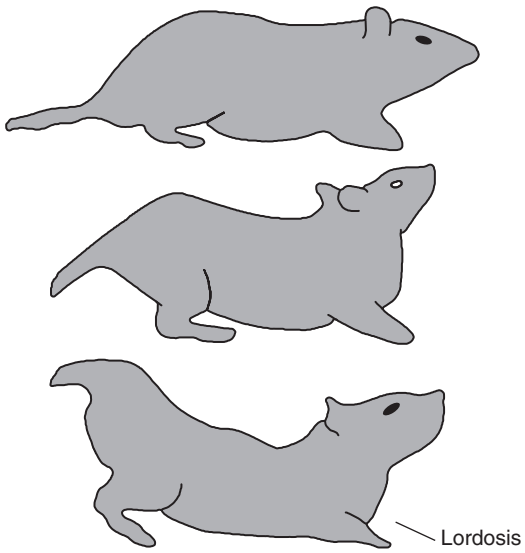

Figure 1 Lordosis behavior (bottom) as performed by female rats. In all female quadrupeds, the vertebral dorsiflexion by the female permits fertilization by the male and thus controls reproduction. Adapted from Becker et al. (1992).

the sex steroid hormones emanating from the ovaries and circulating in the blood to the brain signal that the ovaries are doing their job and are ready to ovulate. From a reverse engineering standpoint, the genomic actions of estrogens in hypothalamic neurons (Figure 4) provide the hormone dependence required for whole-body management of reproduction and its associated behaviors.

Ventromedial hypothalamic neurons that govern lordosis project to the midbrain central gray (Figure 3), where a transformation takes place in the temporal dynamics of neuronal controls. Hypothalamic neurons are slow. Genomic actions of hormones there take hours, neuronal responses are sluggish, and resting discharge rates are low. Yet, in the real world of mating behavior, the female's courtship and lordosis behaviors are fast. Nerve cells in the central gray of the midbrain take the gradually fluctuating inputs from the hypothalamus and translate them into response biases that are important for the postural, behavioral, and autonomic controls managed by the hindbrain. These central gray neurons project to the medullary and pontine reticular formation.

The reticulospinal cells projecting from the medullary and pontine reticular formation, as well as the vestibulospinal neurons projecting from the lateral vestibular nucleus (also important for lordosis behavior; Figure 3), are responsible for managing the coordinated actions of all the segments of the spinal cord that receive the somatosensory inputs from the mounting by the male and that execute the vertebral dorsiflexion (Figure 1) of lordosis behavior. Finally, the descending signals from reticulospinal and vestibulospinal neurons are integrated, at each individual spinal level, with the relevant somatosensory inputs from the male's mounting so that local spinal reflexes executed by motor neurons in the ventromedial corner of the ventral horns bilaterally can stimulate the necessary activation of the deep back muscles, whose contractions up and down the spinal column constitute lordosis.

Usually when we think of design principles during engineering exercises, we are searching for negative and positive feedback. In the controls over mating behaviors, some of these are most prominent in the endocrine dynamics involved. The hypothalamus tells the pituitary gland to send protein hormones, 'gonadotropic hormones,' to the ovaries and testes so that those gonadal organs will produce and release steroidal hormones such as estradiol and testosterone. When those steroidal hormones reach the brain and pituitary, they reduce the hypothalamic/pituitary action and thereby keep hormone levels within a prescribed physiological range. Thus, this process results in negative feedback. However, on the day of ovulation in the female, a temporary but powerful positive feedback mechanism kicks in. For a brief period of time in the late afternoon, higher estrogen levels, supplemented by progesterone, increase gonadotropic hormone release such that a massive amount of gonadotropic hormones in the blood circulates from the pituitary to the ovary and causes ovulation. As noted previously, we believe that the coordinated actions by GnRH, the neuropeptide in charge of the ovulatory surge of hormones causing both behavioral and endocrine changes, provide the neurochemical mechanism for behavior to be rendered consonant with the physiology of the rest of the body.

Male Behavior

Analyses of male mating behaviors have not proceeded as quickly as those of female mating behavior for at least two reasons. First, the steroid chemistry involved is more complicated. Rather than simply depend on testosterone, the peripheral and central physiological changes required for normal male sexual performance depend on testosterone plus its metabolites estradiol and dihydrotestosterone. Second, the behavior is more complicated. Compared to the simple vertebral dorsiflexion that constitutes lordosis behavior (Figure 1), male mating behaviors in laboratory animals involve locomotion and highly coordinated balancing actions during mounting. Nevertheless, in 2002 Hull et al. summarized an impressive body of data on the neuroanatomy and neurochemistry that underlie male behaviors.

A male must chase the female, mount the female, thrust his hindquarters against her so as to achieve

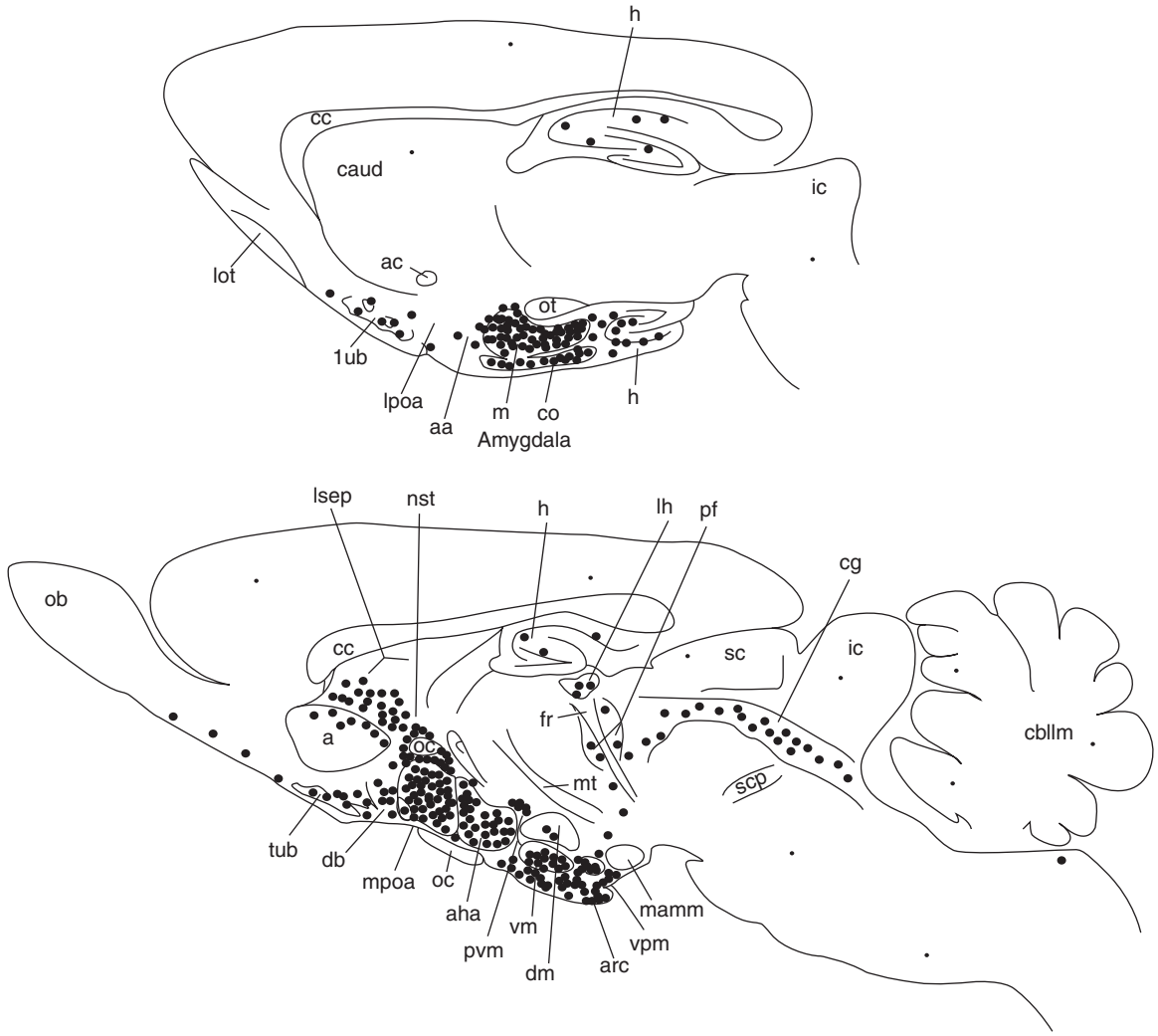

Figure 2 Two sagittal sections through the rat brain looking at it from the left side. Black dots indicate locations of cells expressing estrogen receptors as proven by the binding and retention of tritiated estradiol. The limbic–hypothalamic system revealed in this experiment with rat brain was found to be universal among vertebrate brains. Adapted from Pfaff D and Keiner M (1973) Atlas of estradiol-concentrating cells in the central nervous system of the female rat. *Journal of Comparative Neurology* 151: 121–157.

intravaginal penile insertion ('intromission'), and, after multiple intromissions, ejaculate. Preoptic area neurons, especially those in the medial preoptic area, are essential for the performance of these male sexual behaviors in every vertebrate animal studied. Axonal projections from these neurons reach, among other places, the central gray of the midbrain. The most important axonal termination sites in the central gray are among those neurons that project in turn to the nucleus paragigantocellularis in the medullary reticular formation. These neurons send reticulospinal axons down to the spinal cord, where they excite motor neurons responsible for penile erection and ejaculation.

Although the preoptic area is the brain region most obviously controlling male behaviors, several other forebrain regions contribute. For example, lesions of the corticomedial nuclei of the amygdala reduce male copulatory behaviors, perhaps because they interrupt olfactory signaling from the olfactory and accessory olfactory bulbs to the preoptic area.

Among neurotransmitter inputs to the preoptic area, dopamine acting through D1 receptors seems to be the most involved. In contrast, D2 receptor agonists tend to reduce male copulatory behavior. There is a similar split among receptor subtypes for serotonin. 5-HT-1a agonists increase male sex behaviors, whereas 5-HT-1b agonists decrease them. The inhibitory transmitter GABA working in the preoptic area through $GABA_A$ receptors, as expected, inhibits male mating. A gaseous transmitter, nitric oxide, supports erection but not necessarily through mechanisms located in the forebrain. For example, systemic

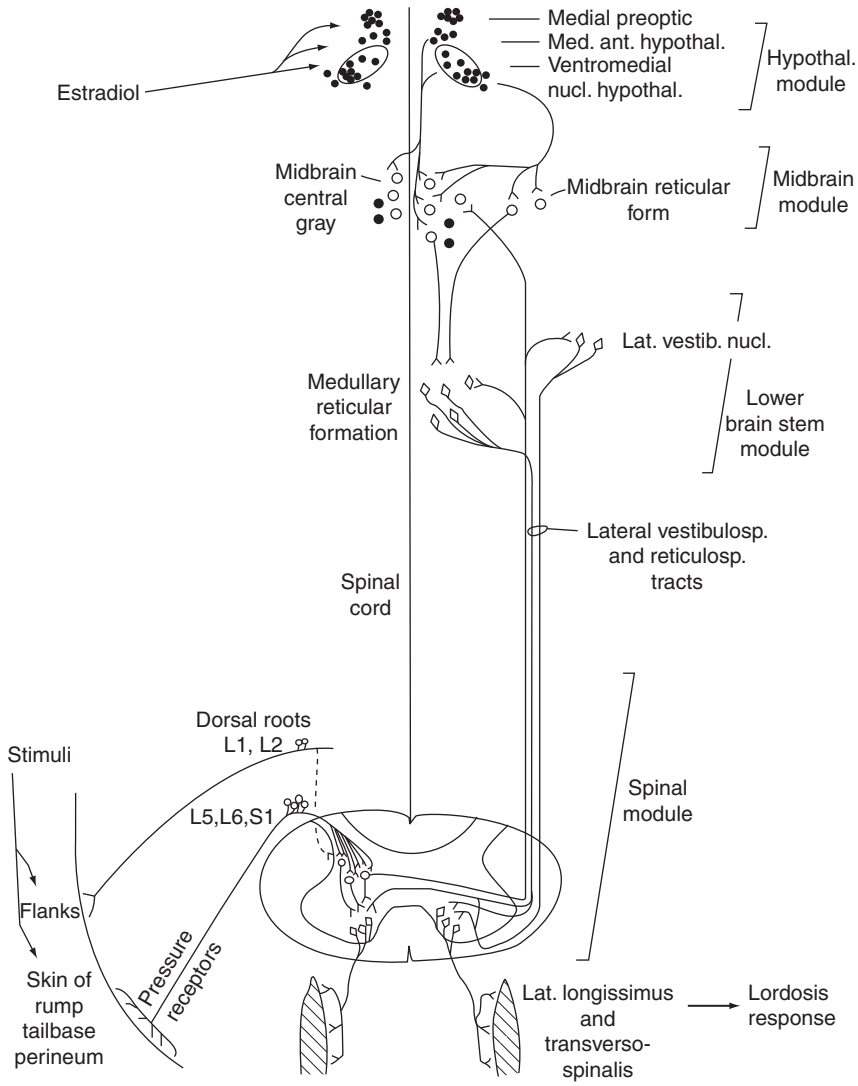

Figure 3 Neural circuit for the production of lordosis behavior. A spinal–midbrain–spinal circuit is regulated by estrogen-modulated neurons in the ventromedial nucleus of the hypothalamus. Adapted from Pfaff D (1980) *Estrogens and Brain Function*. New York: Springer.

inhibition of nitric oxide synthase enzymatic activities will reduce the capacity for erection and, thus, normal male mating behavior.

Neuropeptides emanating from forebrain neurons are also important. For example, oxytocin, a nine-amino-acid peptide associated with affiliative, friendly behaviors and with parental behaviors, can increase spontaneous erections when administered through the cerebral ventricles. Furthermore, GnRH tends to decrease latencies and increase frequencies of male mating responses.

Kevin McKenna and colleagues reported that temporally coordinated sympathetic and parasympathetic autonomic nervous system influences must be integrated with somatosensory inputs from the penile surface to manage erection and then ejaculation. William deGroat charted the sympathetic nervous system connections through the pelvic and pudendal nerves and the parasympathetic nervous system connections through the superior hypogastric and pelvic plexuses. How sympathetic, parasympathetic, and somatosensory influences interact at various stages of the male's mating behaviors is still debated. Nevertheless, the old stereotype of sympathetic 'versus' parasympathetic influence is not correct. Instead, the two autonomic influences must be viewed as a team, not identical or correlated with each other but, rather, coordinated in a biologically adaptive manner.

Generalized Central Nervous System Arousal

Underlying all motivated behaviors in animals and in humans is a set of functions identified as 'generalized

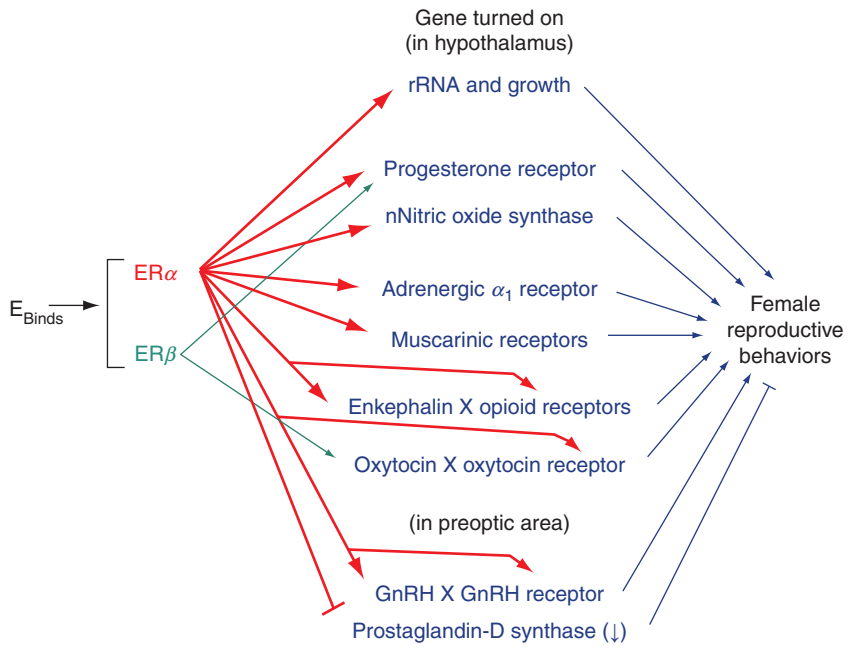

Figure 4 Genes that have two properties: Their transcript levels are elevated by estrogens (E) following binding to an estrogen receptor (ER) and their products foster female reproductive behaviors. The exception is the expression of prostaglandin D synthase in the preoptic area, which works through an estrogen-triggered disinhibition. Adapted from Pfaff D (1999) *Drive.* Cambridge, MA: MIT Press.

Table 1 Comparisons of effects of various neurotransmitters and neuropeptides on lordosis behavior, neuronal electrical activity, and second messengers

Agents (type of receptor)	*Effects on*		
	Lordosis behavior	*Neuronal activity*	*Membrane enzymes*
Lordosis-facilitating agents			
Acetylcholine (M_3?)	↑	↑	↑ on PLC
Norepinephrine (α_1)	↑	↑	↑ on PLC, PLA_2, PLD
Serotonin (5-HT_2)	↑	↑	↑ on PLC
LHRH	↑	∼	↑ on PLC, PLA_2
PRL	↑	∼	↑ on PLC
Oxytocin	↑	↑	↑ on PLC
Substance P (NK_1)	↑	↑	↑ on PLC
$GABA_A$	↑	↓	
Lordosis-inhibiting agents			
Serotonin (5-HT_{1A})	↓	↓	↓ on AC
Dopamine (D_1?D_2?)	↓	↑ or ↓	↑ (D_1) or ↓ (D_2) on AC
CRF	↓	?	↑ on AC
β-Endorphin	↓	↓	↓ on AC
Neuropeptide Y	↓	↓?	↓ on AC
Cholecystokinin	↓	↑	↑ on AC, PLC
Glutamate	↓	↑	

CRF, corticotropin-releasing factor; LHRH, leutinizing hormone-releasing hormone; PLC, phospholipase C; PLD, phospholipase D; PRL, prolactin.
Adapted from Kow L-M, Mobbs CV, and Pfaff DW (1994) Roles of second-messenger systems and neuronal activity in the regulation of lordosis by neurotransmitters, neuropeptides and estrogen: A review. *Neuroscience & Biobehavioral Reviews* 18: 251–268.

central nervous system (CNS) arousal.' Its operational definition states that a more aroused animal or human (1) is more responsive to sensory stimuli in all sensory modalities, (2) emits more voluntary motor activity, and (3) is more reactive emotionally. In our laboratory, we use an automated assay of all three components to examine the functional genomics of generalized arousal in mice. At least five major ascending systems in the CNS and at least four major descending systems participate in heightening

Table 2 Relations among arousal mechanisms and sex behaviors (male and female)

Function	*Requires*			*CNS Features*		*Mechanisms Require*	
	Ag	*As*	*SB*	*Bilateral*	*VMN/POA neuronal excitation*	*Sex hormones*	*Genes for non-specific ascend systems*[*]
Generalized CNS arousal (*Ag*)		No	No	Yes	No	No	Yes
Sexual arousal[a] (*As*)	Yes		No	Yes	Yes	Yes	Yes
Sex behavior (Sb)	Yes	Yes		Yes	Yes	Yes	Yes

[a]See text for explanation.

generalized CNS arousal. More than 120 genes are involved with the receptors, the transporters, and the chemical synthesis and regulated breakdown of the relevant neurotransmitters and neuropeptides. However, what does this mean for mating behaviors of males and females?

Generalized Arousal Fuels Sexual Arousal, Which Fuels Sex Behaviors

In neuroanatomical terms, ascending systems that signal generalized arousal send axons that terminate in the ventromedial nucleus of the hypothalamus (necessary for female sex behavior) and the preoptic area (necessary for male sex behavior). These are so-called 'nonspecific' arousal systems. Among the most important of these are noradrenergic signals (from the locus coeruleus), cholinergic signals (likely from the dorsomedial nucleus of the tegmentum), dopaminergic signals (from the ventral tegmental area), serotonergic signals (from the raphe nuclei of the midbrain), and histaminergic signals (from the tuberomammillary nucleus of the hypothalamus). These arousal-related transmitters have the capacity to potentiate a wide variety of behavioral responses, including mating behaviors (Table 2).

In electrophysiological terms, noradrenergic, cholinergic, and histaminergic transmitters turn on the electrical activity of ventromedial hypothalamic neurons crucial for lordosis behavior in the female. Correspondingly, dopamine, acting through D1 receptors in the preoptic area, facilitates male sex behavior. Thus, heightened electrical activity in these hypothalamic and preoptic neurons is the neurophysiologic equivalent of sexual arousal in the female and male, respectively.

In behavioral terms, sexual arousal is occasioned by heightened muscular tone throughout the body and by courtship behaviors designed to bring reproductively competent males and females into contact. In the female rat, extreme axial muscle tension and rapid darting movements that attract the male are easily noticed. In the male, searching and following the female are obvious. These are accomplished not only by heightened electrical activity in the ventromedial hypothalamic and medial preoptic neurons but also by the generalized CNS arousal mechanisms referred to previously (Table 2).

Logically, therefore, it must be concluded that generalized arousal is required for heightened sexual arousal, which in turn is required for the normal performance of both male and female sexual behaviors. This field of work offers the possibility of linking well-understood, concrete mechanisms for specific biologically important behaviors to a generalized CNS function crucial for the potentiation of all mammalian behaviors.

See also: Gonadal Steroid Actions on Brain; Neurosteroids; Sexual Differentiation of the Brain; Sexual Differentiation of the Central Nervous System; Stress Response: Sex Differences.

Further Reading

Hull E, Meisel R, and Sachs B (2002) Male sex behavior. In: Pfaff D, Arnold A, Etgen A, Fahrbach S, and Rubin R (eds.) *Hormones, Brain and Behavior*, pp. 3–137. San Diego: Elsevier.

Pfaff D (1980) *Estrogens and Brain Function*. New York: Springer.

Pfaff D (1999) *Drive*. Cambridge, MA: MIT Press.

Pfaff D (2006) *Brain Arousal and Information Theory*. Cambridge, MA: Harvard University Press.

Sexual Differentiation of the Brain

D G Zuloaga, C L Jordan, and S M Breedlove, Michigan State University, East Lansing, MI, USA

© 2009 Published by Elsevier Ltd.

Men and women differ not only in their anatomy and reproductive functions but also in cognitive and affective behaviors. On an average, men are generally better at tasks involving the mental rotation of objects, whereas females are superior at memory tasks in which they are asked to recall objects in a spatial array. There are also sex differences in the prevalence of mental disorders, with males more commonly diagnosed with schizophrenia, autism, and attention deficit hyperactivity disorder. Females, on the other hand, are more likely to suffer from disorders of depression and anxiety. During the past several decades, scientists have uncovered sexual dimorphisms in the central nervous system (CNS) of mammals that may contribute to the etiology of these sex differences. Studies on rodents suggest that many sexual dimorphisms in the CNS result from exposure to hormones at approximately the time of birth, lending credence to the idea that the masculinizing or feminizing seed is planted early in life. However, some areas of the brain remain vulnerable to hormones after this critical period and can incur morphological changes during adolescence and in adulthood. This article discusses instances of each in the mammalian CNS, the mechanisms that underlie these sexual dimorphisms, and examples of sexual differentiation of the human brain and behavior.

The sex of a mammal is determined at the time of conception when sperm from the male penetrates the egg and either contributes an X or a Y chromosome to the genetic makeup. If the sperm donates a Y chromosome, the offspring will be a male, and if it donates an X chromosome, the offspring will be a female. The sex-determining region Y (SRY) gene on the Y chromosome of the male is responsible for the formation of the testes, and in the absence of this gene, ovaries form. In males, it is the formation of the testes and the subsequent secretion of testicular hormones that masculinize the body. One of these hormones, anti-Müllerian hormone (AMH), suppresses the development of the female reproductive organs (e.g., the fallopian tubes and uterus), whereas another, the steroidal androgen testosterone (T), induces the formation of the male reproductive system (including the seminal vesicles and prostate) and external genitalia. In females, T and AMH are nearly absent, so male genitalia and internal sex organs fail to form and sex organs develop a feminine phenotype. Gonadal hormones can only influence genital phenotype during specific developmental periods; T in adulthood does not result in further development of male genitalia.

Testicular androgens such as T are primarily responsible for masculinization of the brain. In 1959, Phoenix and colleagues discovered that prenatal administration of T to female guinea pigs resulted in a defeminization of female sexual behavior. As adults, these early T-treated females (when administered estrogen and progesterone (E + P) to stimulate female sexual behavior) displayed little or no lordosis, the female rodent posture of sexual receptivity. However, these same females responded to testosterone in T, showing increased pelvic thrusting (typical of male mounting behavior). Similarly, adult males that were castrated immediately after birth showed increased lordosis after E + P administration as adults and decreased pelvic thrusting after T treatment. This outcome suggested that perinatal androgens are necessary for masculinization (promoting male-typical behaviors such as pelvic thrusting) and defeminization (interfering with female-typical behaviors such as lordosis) in male rodents. Females treated with T only in adulthood show little or no difference in sexual behavior, implying that rodents need 'androgen priming' in early development in order for adult testosterone to 'activate' behavior. This series of experiments laid the foundation for the organizational hypothesis, which postulates that behavioral sex differences result from (1) exposure to different hormones that act in early development to organize the neural machinery underlying behaviors and (2) exposure to different sex hormones in adulthood that activate the previously organized neural machinery. This conceptual framework has been amply confirmed in studies of nonhuman mammals.

Sexual Differentiation of the CNS

Exposure to steroid hormones at approximately the time of birth can also result in permanent morphological changes in the CNS, which may explain the lasting effects of early hormone exposure on adult behavior. In rodents, there are structural sex differences in several areas of the brain, one of the most robust of which is the sexually dimorphic nucleus of the preoptic area (SDN-POA). The volume of the SDN-POA is several times greater in males than in females, a difference which results from T exposure at approximately the time of birth. Castration of

newborn male rats decreases the size of the SDN-POA in adulthood, whereas T treatment of female rats just before and after birth increases adult SDN-POA volume. In the SDN-POA, masculinization results not from the direct activation of androgen receptors (ARs) by T or its metabolite dihydrotestosterone, as it does in the genitalia. Rather, T is converted to estrogens such as estradiol by the enzyme aromatase within the brain, and this brain-produced estrogen activates estrogen receptors (ERs) to masculinize the SDN-POA. The effect of activating ERs is to keep neurons in the SDN-POA from dying during the course of neuronal development. In turn, this rescuing of neurons from death by estradiol results in a greater SDN-POA volume in males than in females (**Figures 1(a)** and **1(b)**).

In another sexually dimorphic brain region, the anteroventral periventricular nucleus (AVPV), which is involved in regulating ovulatory cycles by influencing secretion of the luteinizing hormone, regional volume is greater in female rats and mice than in males. This sex difference is also due to the binding of aromatized metabolites of T to ERs. However, in the AVPV, ER activation promotes cell death, contributing to a larger female nucleus.

The spinal nucleus of the bulbocavernosus (SNB) provides another example of a sexually differentiated structure in the CNS in which gonadal hormones play a decisive role. The SNB is a sexually dimorphic group of motor neurons located in the lower lumbar spinal cord which innervates muscles involved in male copulatory behavior: the bulbocavernosus and

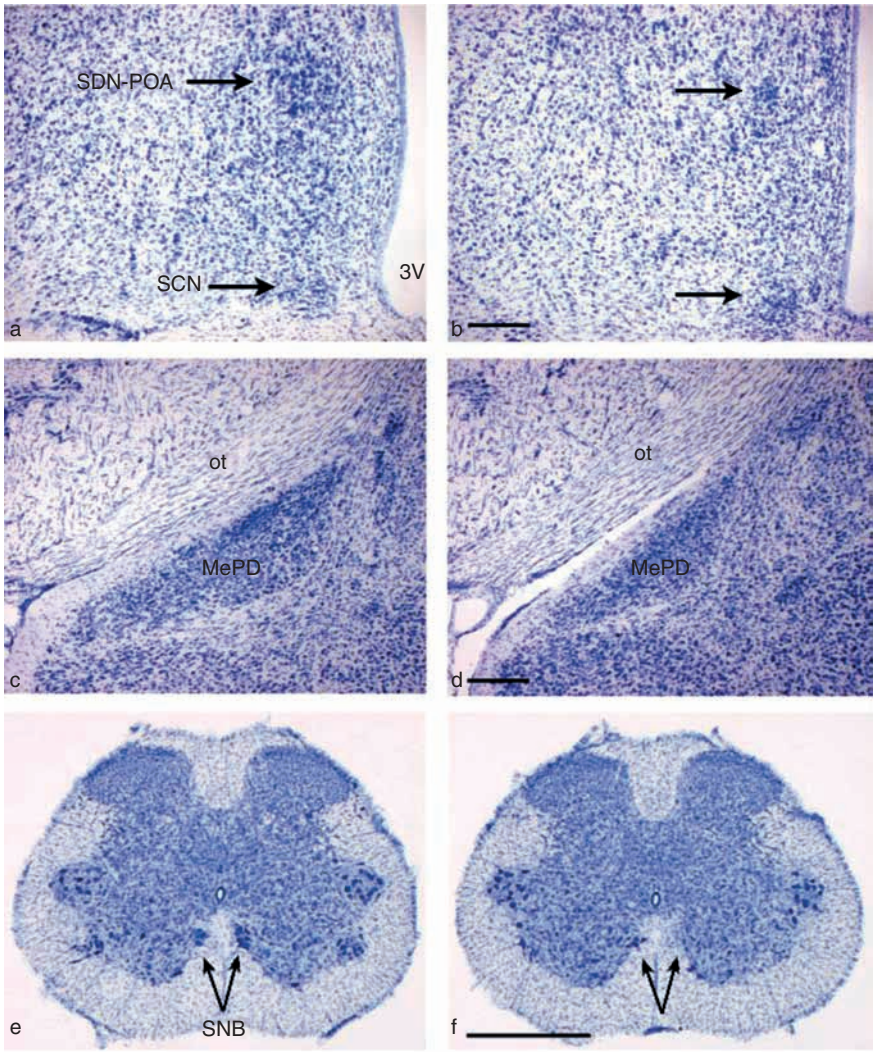

Figure 1 The sexually dimorphic nucleus of the preoptic area (SDN-POA) and the posterodorsal portion of the medial amygdala (MePD) are larger in males (left) than in females (right). There are also more motor neurons in the spinal nucleus of the bulbocavernosus (SNB) of males (left) than of females (right). SCN, suprachiasmatic nucleus.

levator ani perineal muscles. SNB motor neurons and the muscles they innervate are present at approximately the time of birth in both sexes, but they die in females postnatally while surviving in males. The SNB system is spared from death by early exposure to T in males (Figures 1(e) and 1(f)). Neonatal castration of males causes the SNB system to die, whereas T treatment of females spares the system. However, unlike the SDN-POA, SNB cells are spared from apoptosis (cell death) by the action of androgens upon ARs and not ERs. Genetic male rats with a defective AR fail to develop a masculine SNB. It is also known that androgens do not act directly on SNB motor neurons to spare them from death. Rather, androgens first act in the periphery to spare the muscles that are innervated by the SNB, and this action secondarily spares SNB motor neurons from death.

Sexual Differentiation during Adolescence

The maturation of the reproductive system during adolescence is marked by a period of elevated gonadal hormones. It is likely that these gonadal hormones also play a role in molding both the male and the female adolescent brain, and emerging evidence in rodents suggests that pubertal hormones indeed organize brain morphology and behavior. For example, some sex differences in the brain develop during the course of adolescent development, as in the locus coeruleus (LC) and AVPV, where volume is greater in females, and in the primary visual cortex, where volume is greater in male rats. Both the LC and the AVPV enlarge in females during adolescence, although the mechanisms that underlie these increases in volume are not known (neurogenesis, increased migration, or cell differentiation?). There is evidence that sex differences in the primary visual cortex are the result of ovarian hormones that induce cell death in females during adolescence, suggesting that female gonadal hormones can also drive brain organization. Ovariectomy of female rats prior to puberty results in an increased number of neurons in the primary visual cortex compared to females with intact ovaries. The visual cortex of ovariectomized females becomes similar to males in both neuron number and volume.

Changes in synaptic organization also occur during adolescence, a process that has been best studied in the medial amygdala (MeA). The MeA is a brain region that receives input from olfactory and pheromonal systems by way of the olfactory bulbs, and it has been implicated in male sexual behavior. The MeA is approximately 150% larger in adult male rats than in females. A significant amount of remodeling, involving the pruning of dendrites as well as changes in dendritic spine density in the MeA, accompanies adolescent development in hamsters and may serve to organize aspects of cognition and behavior.

Studies of the Syrian hamster further indicate that adolescence is a second critical period for behavioral development. T during puberty is necessary for the full masculinization of male hamster sexual behavior since prepubertal castration decreases mounts and intromissions when the hamster is exposed to a receptive female in adulthood. Similarly, pubertal hormones appear to masculinize male agonistic behavior since castration prior to puberty decreases aggressive attacks and increases submissive behavior compared to animals that received T during puberty.

Sexual Differentiation in Adulthood

In certain areas of the CNS, sexual differentiation occurs in adulthood and is similarly dependent on gonadal hormones. In rats, portions of the MeA remain sensitive to circulating androgens into adulthood because castration of adult males causes some areas within the MeA to shrink to a size typical of females, whereas T treatment of adult females causes the same area to grow to a male-typical volume. The size of the posterodorsal medial amygdala (MePD; Figures 1(c) and 1(d), which receives olfactory and pheromonal information and is important for some aspects of male sexual behavior, appears to be solely dependent on adult hormones. MePD volume is 1.5 times greater in male compared to female rats and mice, but this sex difference can be abolished by castration of males or administration of T to females. Structural plasticity in the MePD is mediated in adulthood through activation of both ARs and ERs, and the overall amygdala also appears to grow at puberty in boys, but not girls, suggesting that it is sensitive to androgens in humans as well as in rodents.

Another brain sexual dimorphism, arginine vasopressin (AVP) fiber innervation of the septal area in rodents, depends on T in both early development and adulthood for full masculinization. Septal AVP innervation is greater in males and plays a role in pair bonding, parental, and aggressive behavior. Castration of adult males decreases septal vasopressin innervation, although AVP depletion following neonatal castration is more extensive than the reduction following adult castration. The sex difference in septal vasopressin in adulthood is primarily dependent on aromatized metabolites of T that act on ERs, although the AR also contributes to some degree to the perinatal organization of this system.

The SNB system, like septal AVP innervation, is influenced by hormones in both development

and adulthood. Castration of adult males causes widespread regression throughout the system: The muscles, the motor neurons, and their synapses all regress after castration in adulthood. These morphological changes are accompanied by a reduction in spinal reflexes of the penis that are necessary for male sexual behavior. Interestingly, although both SNB somas and dendrites shrink after castration, each appear to be regulated by distinct mechanisms. In rats, androgens act directly through ARs in SNB motor neurons to regulate soma size while acting indirectly, through AR in the target muscles, to regulate the length of SNB dendrites.

Hormonal Site of Action and Downstream Regulation

Although the particular hormones and receptors involved in sexual differentiation of brain and spinal cord structures are becoming clear, little is known about the location and cell types that these hormones act upon and the downstream genes regulated to masculinize the CNS. In the SDN-POA, T acts to prevent apoptosis; however, it is not known whether it is acting directly on neurons in the SDN-POA or in some other area that indirectly spares these neurons. In the SNB system, motor neuron survival results, secondarily, from the sparing of target musculature that provides SNB motor neurons with a signal that keeps them alive. Thus, it is possible that sexual differentiation of the SDN-POA and AVPV results from T acting on distant target cells. However, although it is known that T acts on the target musculature to spare SNB motor neurons, it is not known what cell types in the muscle (muscle fibers, fibroblasts, or Schwann cells) are involved. Similarly, in the SDN-POA and the AVPV, it is not known which cell type (neurons, astroctyes, or oligodendrocytes) provides the site(s) of action for T. The only morphological sex difference in rodents for which the cellular target of T is known is in the regulation of SNB soma size. In female rats heterozygous for the testicular feminization mutation, a mutation that renders the AR protein nonfunctional, testosterone treatment in adulthood increased soma size only in SNB neurons with functional ARs. This suggests that T acts directly via the AR in SNB motor neurons to regulate soma size.

Researchers are also just beginning to uncover the genes acting downstream from T. Evidence suggests that T may modulate the Bcl2 family of genes that are involved in regulating apoptosis in the SNB, AVPV, and the central division of the medial preoptic nucleus (which comprises a significant portion of the SDN-POA). Deletion of a member of the Bcl2 family, the Bax gene, which signals downstream regulators of cell death, decreases sex differences in neuron number in both the SNB and the AVPV by decreasing the amount of cell death. Likewise, sex differences are reduced in the SNB and AVPV in transgenic mice that overexpress the anti-apoptotic Bcl2 gene. Interestingly, T promotes cell death in the AVPV while preventing it in SNB and SDN-POA. In both cases, T appears to affect expression of Bcl2 family genes in such a way that it can spare cells in some areas and condemn them in others.

Neurotrophic factors, as mediators of hormone effects, also appear to play a role in the sexual differentiation of the CNS. In the SNB system, T-dependent signaling through the ciliary neurotrophic factor receptor α (CNTFRα) is necessary for the development of a sex difference. CNTFRα is highly expressed in SNB motor neurons and administration of CNTF to female rats rescues SNB motor neurons from death. Similarly, male mice lacking CNTFRα fail to develop a masculine SNB system. Evidence from our laboratory suggests that another neurotrophic factor, brain-derived neurotrophic factor (BDNF), plays a role in the arborization of SNB dendrites in adulthood. Castration of adult rats, which results in decreased SNB dendritic length, is paralleled by a decrease in BDNF signal in SNB motor neurons and their dendrites, whereas BDNF signal is unaltered in androgen-insensitive motor neurons of the nearby retrodorsolateral nucleus.

In the POA, T induces the formation of prostaglandin E_2 (PGE_2), which is involved in dendritic spine development. The proliferation of dendritic spines in the POA appears to be important in the organization of male sexual behavior. Pharmacological blockade of PGE_2 in male newborns inhibits the formation of dendritic spines, whereas administration of PGE_2 to newborn females increases spine formation. Neither manipulation of PGE_2 has an effect on SDN-POA volume, suggesting that T may regulate the expression of different genes to promote dendritic spine formation in the POA, on the one hand, and increase volume of the SDN-POA on the other hand. For other areas of the CNS where sex differences have been located, little is known about the genes acting downstream from T, primarily because not enough is known about the target cells that respond directly to T or its metabolites.

Sexual Differentiation of the Human Brain and Behavior

The role of gonadal steroid hormones in the masculinization of the human brain and behavior is not as clear as in rodents, partially because human

behavior is very complex. Humans are so dramatically influenced by social cues that it is difficult to decipher whether a particular behavior is the result of T exposure or a byproduct of the milieu in which they were raised. Many sex differences in behavior may simply reflect the fact that boys have a penis and girls have a vagina, because from the moment of birth something as simple as genital development will radically shape the child's social environment.

Much of what we have learned about the role of hormones in the development of human behavior has come from studies of people who have been exposed to unique cascades of hormonal development. One such example is found in genetic males (XY) with complete androgen insensitivity syndrome (CAIS), a disorder characterized by a mutation in the AR gene. In people with CAIS, the testes develop and release T, but in the absence of functional AR, T fails to masculinize the periphery, resulting in a female phenotype. These individuals are generally raised as girls, and as adults they self-identify as women. They do not differ from genetic (XX) females in sexual orientation, are just as likely to be married, and as children were involved in female-typical play. Since people with CAIS have ample T and functional ER, this suggests either that, unlike rodents, the aromatization of T to estrogen and subsequent activation of ERs is not responsible for the masculinization of behavior (and thus AR may be necessary), or that the social context in which these individuals are raised and continue to live outweighs any influence of gonadal hormones.

Studies of women with congenital adrenal hyperplasia (CAH), who are exposed to T in fetal development because their adrenal glands overproduce androgens, suggest that androgens may play a role in masculinizing the female brain. Females with CAH are more likely to engage in male-typical activities and are less interested in female-typical activities throughout life. In addition, although most women with CAH are self-reported heterosexuals, they are more likely to be homosexual than are other women. Women with CAH also show male-typical spatial performance, suggesting that androgens play a role in the masculinization of behavior in humans. However, the genitalia of these women are also slightly masculinized, which could affect the way these women are perceived either by their parents or by themselves.

Another strategy that has been used to study the effects of early androgen exposure on human behavior is through the examination of somatic markers that are affected by prenatal androgen exposure. These somatic markers of androgen exposure, including finger length measurements (ratio between second and fourth digit lengths (2D:4D)), otoacoustic emissions, eyeblink patterns, and limb measurements, can be co-varied across measures of sexual orientation, cognitive abilities, and sex-biased affective and mood disorders. Using these somatic markers, links have been made between an increased exposure to prenatal T and homosexuality in women. A relationship between somatic markers of T in male homosexuality is less clear. Examination of finger length ratios in males suggests that levels of prenatal testosterone may be related to some affective and mood disorders. Boys with autism and attention deficit hyperactivity disorder, which are more common in males, tend to have hypermasculine finger length ratios (smaller 2D:4D; increased prenatal testosterone), whereas boys with anxiety disorders have more feminine ratios (larger 2D:4D; decreased prenatal testosterone).

Differences in hormone exposure may contribute to structural sex differences in the human CNS, including the corpus callosum, anterior commissure, third interstitial nucleus of the anterior thalamus (INAH3), bed nucleus of the stria terminalis, Onuf's nucleus, and suprachiasmatic nucleus (SCN). There is also a sex difference in overall brain weight favoring men, although women have a greater cortical surface area. These sexual dimorphisms in brain morphology may contribute to gender differences in behavior and the prevalence of some mental illnesses. Very little is known about how sex differences in the human brain may be related to differences in behavior, although the sizes of two hypothalamic nuclei have been correlated to sexual orientation in men. Inspired by the dramatic sex difference in the SDN-POA of rats, in 1991 Simon LeVay compared nuclei within the preoptic area of heterosexual men and women and discovered that one of these nuclei, INAH3, was larger in heterosexual men than in women, and that homosexual men had a feminine-sized nucleus. On the other hand, the vasopressinergic subnucleus of the SCN was found to be larger in volume and contained twice as many cells in homosexual men compared to heterosexual men. In this instance, the nucleus of homosexual men was not feminized. Together, these findings indicate that the brains of homosexual men may be feminized in some areas but not in others. However, it is important to note that it is unclear whether these morphological differences in the brains of homosexual men are the result of androgen exposure or arise as a result of differences in experience and behavior.

In some areas of the brain, sex differences in morphology may exist as a means to minimize sex differences in behavior. Differences in gonadal hormone levels between men and women could potentially contribute to maladaptive sex differences in behavior,

with alterations in brain structure compensating for these differences. An example of such a phenomenon has been suggested in the prairie vole, in which the male and female both contribute to the care of their offspring. In females, changes in hormones that accompany pregnancy and childbirth prime maternal behavior. Males, on the other hand, are not exposed to these hormones and their parental behavior appears to rely on sexually differentiated vasopressin innervation. This more extensive vasopressin innervation of the septal area in males may then compensate for the lack of maternal hormone exposure and make male prairie voles more effective fathers. It is also possible that hormones *per se* are not the sole contributor to the sexual differentiation of brain and behavior. Evidence suggests that in some songbirds and perhaps in mice, the expression of sex chromosome genes within cells may exert some influence on sexual differentiation of the brain and behavior independently of gonadal steroids. However, gonadal hormones are the primary instigators of sex differences in neural structure.

In many ways, hormone exposure during critical periods of development affects us, partially because these hormones have made us look like a male or a female, and also because these hormones have influenced the development of our brain and subsequently our behavior. Hormones may also continue to act throughout our lives to influence behavior; decreases in gonadal hormones with aging are often associated with depression, anxiety, debilitated memory, and general unhappiness. Lifelong plasticity in areas such as the MePD and the SNB demonstrates how these changes in behavior may also be accompanied by changes in structure. It is likely that there are many additional sex differences in the CNS waiting to be revealed. As we discover more of these dimorphisms and uncover how these sex differences emerge, we gain a better understanding about the influence of hormones and why, in many ways, men and women are really not the same.

See also: Gonadal Steroid Actions on Brain; Neurosteroids; Sexual Differentiation of the Central Nervous System; Stress Response: Sex Differences; Vasopressin/Oxytocin and Receptors.

Further Reading

Breedlove SM and Arnold AP (1980) Hormone accumulation in a sexually dimorphic motor nucleus of the rat spinal cord. *Science* 210: 564–566.

Cohen-Bendahan CC, van de Beek C, and Berenbaum SA (2005) Prenatal sex hormone effects on child and adult sex-typed behavior: Methods and findings. *Neuroscience and Biobehavioral Reviews* 29: 353–384.

De Vries GJ (2004) Minireview: Sex differences in adult and developing brains: Compensation, compensation, compensation. *Endocrinology* 145: 1063–1068.

Forger NG (2006) Cell death and sexual differentiation of the nervous system. *Neuroscience* 138: 929–938.

LeVay S (1991) A difference in hypothalamic structure between heterosexual and homosexual men. *Science* 253: 1034–1037.

Morris JA, Jordan CL, and Breedlove SM (2004) Sexual differentiation of the vertebrate nervous system. *Nature Neuroscience* 7: 1034–1039.

Nelson RJ (2005) *An Introduction to Behavioral Endocrinology*, 3rd edn. Sunderland, MA: Sinauer.

Phoenix CH, Goy RW, Gerall AA, and Young WC (1959) Organizing action of prenatally administered testosterone propionate on the tissues mediating mating behavior in the female guinea pig. *Endocrinology* 65: 369–382.

Sisk CL and Zehr JL (2005) Pubertal hormones organize the adolescent brain and behavior. *Frontiers in Neuroendocrinology* 26: 163–174.

Swaab DF and Hofman MA (1990) An enlarged suprachiasmatic nucleus in homosexual men. *Brain Research* 537: 141–148.

Sexual Differentiation of the Central Nervous System

R B Simerly, Childrens Hospital Los Angeles, Los Angeles, CA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Despite a wide variety of behaviors that are shared between male and female members of vertebrate species, certain traits are strongly associated with a single sex. Such behaviors are said to exist in two forms and are therefore dimorphic. Behaviors that are characteristic of males are designated as masculine and those that are characteristic of females as feminine. Sexually dimorphic behaviors are widespread among vertebrate species because nearly all reproduce sexually. Accordingly, dimorphic phenotypic traits have developed to support these sex-specific behaviors and the physiological processes that accompany them. Thus, males and females display distinct copulatory behaviors and neuroendocrine physiological responses. The importance of these sex-specific functions to vertebrate species is reflected in their widespread occurrence and in the reliability of developmental mechanisms that have evolved to ensure their unambiguous display. Sexually dimorphic behavior extends beyond reproduction and maternal behavior, although such differences can be more subtle and at times less reproducible.

Chromosomal Sex and the Developing Gonad

Masculinization of the body begins with a genetically controlled process termed sex determination that leads to the sex-specific differentiation of the gonad. The genetic sex of mammals is specified by the presence of either paired X chromosomes (females) or single X and Y chromosomes (males). The paternally derived Y chromosome contains the Sry (sex-determining region of the Y) gene, which encodes a transcription factor that causes the undifferentiated gonads to develop into testes; the absence of Sry expression leads to development of ovaries. Hormones secreted by the testes act on various tissues to masculinize the body, and it is testosterone secreted by Leydig cells in the testes that is responsible for masculinization of the central nervous system (CNS). In males, levels of testosterone are elevated significantly at approximately the time of birth (perinatal) and pass through the blood to the CNS, where they have free access to developing neurons and glia. Although all of the evidence to date supports a defining role for testosterone in sexual differentiation of the brain, it would be surprising if genes located on the Y chromosome, and therefore only present in males, did not influence masculinization of the CNS in some manner. Genetic studies in mice suggest that chromosomal sex does appear to impact certain behaviors, but these effects are relatively subtle compared with the robust changes caused by sex steroid hormones.

Major Principles and Model Systems

The developing brain is considered to be bipotential with respect to sexual differentiation, and in the absence of testosterone it develops phenotypic traits that are characteristic of females. For example, the ability of males to mount receptive females and the ability of females to display lordosis are dependent on the perinatal hormone environment. Similarly, whether the brain can induce a surge in secretion of gonadotropins from the pituitary gland, and thereby cause ovulation, is also determined by perinatal exposure to testosterone. Sex steroids also regulate expression of reproductive behavior and associated neuroendocrine events throughout life in response to what are described as activational influences of steroid hormones. The more permanent, developmental actions of testosterone are viewed as having an organizational action on the CNS and tend to be restricted to discrete periods of development. Thus, treating a female rat with testosterone during the first few days of life will permanently masculinize copulatory behavior and eliminate ovulation, but applying a similar treatment in adulthood does not. However, in adulthood, sex steroids regulate sexually dimorphic behavior and these activational actions are required for full expression of behavior (Figure 1).

The ability of perinatal testosterone to permanently alter neural structure and function is primarily due its action on brain development. The neural circuits controlling sexually dimorphic functions are thought to be specified during a developmental critical period occurring at or near the time of birth. For example, injecting female rats with testosterone during the first week of life generally increases the size of the sexually dimorphic nuclei, but treatments later on do not. However, sex steroid hormones continue to affect the structure and function of the CNS throughout life, but to a more limited degree than during the perinatal period. Gonadal hormones can alter patterns of connectivity, change expression of neurotransmitters and their receptors, and influence electrophysiological properties of neurons in adults. These effects can be very rapid (seconds to minutes) or can occur over extended periods (days to weeks). Thus, it is the combined action of sex steroid

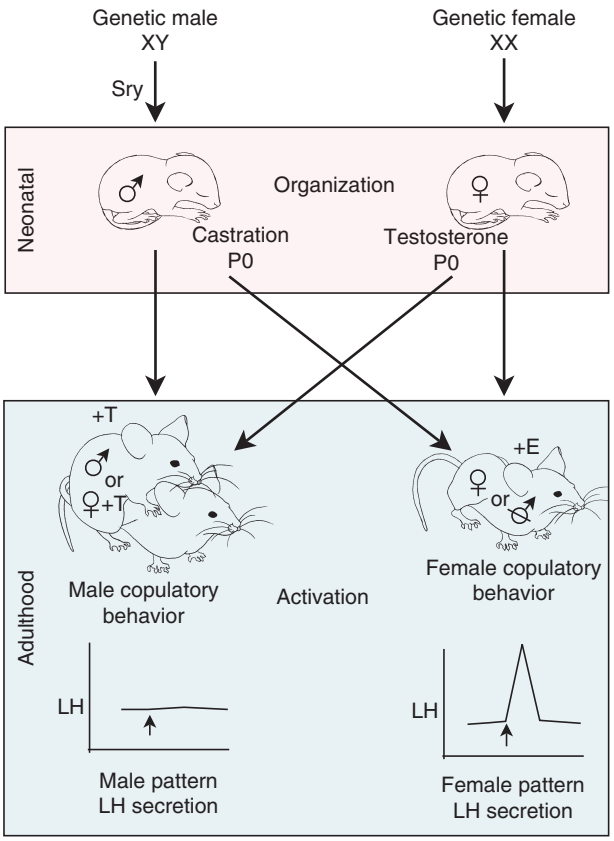

Figure 1 Sexual differentiation and hormonal activation. Sex determination is specified by the complement of sex chromosomes (X and Y). Expression of the Sry gene on the Y chromosome in males causes formation of the testes, which in rodents secrete testosterone at high levels during the neonatal period (the first few days of life). During this period, testosterone acts on the brain to 'organize' the structure of regions that mediate sexually dimorphic functions. During adulthood, sex steroid hormones such as testosterone (T) and estrogen (E) act on the brain to 'activate' expression of copulatory behavior, mounting in males and lordosis or arching of the back in females, as well as sex-specific patterns of luteinizing hormone (LH) secretion from the pituitary gland, which causes ovulation. Castration of male newborn mice, or treating females with testosterone, on the day of birth (P0) reverses expression of these sexually dimorphic functions.

hormones, acting during discrete periods of sensitivity during development, together with the more short-term activational actions of these hormones later in life, that determines display of sexually dimorphic behaviors and physiological responses. However, it is important to remember that levels of circulating hormones in adult animals are different in males and females, which may influence expression of phenotypic traits independent of the developmental actions of testosterone. Surprisingly, it is estrogen that effects many of the masculinizing developmental actions of testosterone in the brain through intracellular conversion of testosterone to estradiol (a form of estrogen) by the enzyme aromatase. Once formed in the cell, the estradiol binds to estrogen receptors, which function as ligand-activated transcription factors to control gene expression. By regulating expression of genes involved in multiple aspects of cell function and development, estrogen directs the formation of sexually dimorphic neural circuits that are presumably responsible for expression of sexually dimorphic behaviors and physiological responses.

Because of its prominent role in regulating reproductive function, the hypothalamus has been the focus of many studies on sexual dimorphism in the brain, and most sexually dimorphic nuclei have been found within the hypothalamus or in limbic regions that innervate it. Nuclei in the preoptic region of the hypothalamus are known to control copulatory behavior and ovulation, and generally these regions contain high densities of neurons that express receptors for testosterone and estrogen. In a variety of mammalian species, preoptic nuclei involved in these functions contain different numbers of neurons in males and females, as well as having sexually dimorphic patterns of connectivity. The medial preoptic nucleus of the hypothalamus (MPN) contains distinct components that are larger in males and contain more neurons. These cellular components were included within a region designated as the sexually dimorphic nucleus of the preoptic area (SDN-POA) by Roger Gorski and co-workers. This region is so dimorphic as to be identifiable in stained tissue sections without magnification. Treatment of females with either testosterone or estrogen increases the size of the SDN-POA to that characteristic of males, and removal of the gonads from newborn male rats reduces the size of the region. Most morphological sex differences follow this pattern of hormonally specified male dominance, but the pattern is reversed in another preoptic nucleus, the anteroventral periventricular nucleus (AVPV). The AVPV plays a critical role in the control of ovulation and is larger in females. Exposure to testosterone during perinatal life reduces the size of the AVPV and permanently alters numbers of distinct subpopulations of neurons in a cell-type-specific manner. For example, treatment of female rats with testosterone decreases the number of dopaminergic neurons in the AVPV but increases numbers of neurons that express the opioid peptide dynorphin in the same nucleus. In humans, the rostral hypothalamus contains several distinct cell groups, termed the interstitial nuclei of the anterior hypothalamus, and at least two of these nuclei have been reported to be sexually dimorphic. Although these findings are controversial, and present difficulties for functional interpretation, it seems clear that structural sex differences are not restricted to nonhuman species.

The MPN and AVPV receive inputs conveying a variety of sensory modalities, but the most direct are those than transmit olfactory signals. The medial nucleus of the amygdala and principal nucleus of the bed nuclei of the stria terminalis (BSTp) are sexually dimorphic nuclei (larger in males) that are key components of the pathway from the accessory olfactory bulb to the hypothalamus. Androgen receptors may play a more prominent role in regulating the development of these telencephalic nuclei than they do in the hypothalamus, and these regions appear to remain relatively sensitive to the effects of testosterone manipulations in adulthood. Breedlove and colleagues showed that castration of adult males decreased, and treatment of adult females increased, the size of the posterodorsal part of the medial amygdala. Concurrent changes in the hormone-treated animals in their responses to airborne cues from receptive females suggest that hormonal regulation of nuclear volume may underlie changes in the function of this sexually dimorphic olfactory pathway.

The most thoroughly studied sexually dimorphic motor pathway is the spinal nucleus of the bulbocavernosus (SNB), which innervates striated muscles of the perineum (Figure 2). The motor neurons of the SNB are present before birth in both sexes and innervate the muscles, but the muscles regress and the neurons die soon after birth in females unless they are exposed to testosterone. In contrast to the sexual differentiation of the preoptic region, which is largely mediated by estrogen receptors, the action of testosterone on the SNB relies solely on androgen receptors.

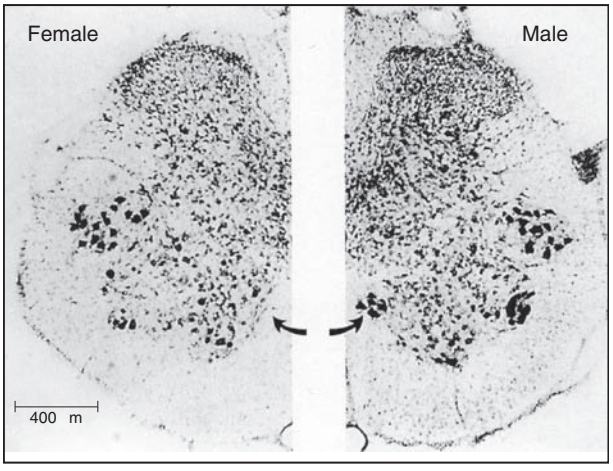

Figure 2 The spinal nucleus of the bulbocavernus (SNB). The SNB is found in the lumbar region of the spinal cord in rats. There are 3 times as many motor neurons in the SNB (arrows) of male rats as there are in females. Adapted from Breedlove SM, Jordan CL, and Kelley DB (2002) What neuromuscular systems tell us about hormones and behavior. In: Pfaff D, Arnold A, Etgen A, Fahrbach S, and Rubin R (eds.) *Hormones, Brain and Behavior*, pp. 192–222. San Diego, CA: Academic Press.

However, testosterone does not act directly on the SNB to influence neuron number but, rather, acts on the perineal muscles and then secondarily affects the motor neurons of the SNB. Such target-dependent control of development is a common feature of neuromuscular systems. However, hormones can also act at multiple levels and at multiple times during development of neuromuscular systems. For example, in the frog vocal control neuromuscular system studied by Darcy Kelley and colleagues, both muscle cells and motor neurons express androgen receptors and appear to represent direct targets for hormone action on sexual differentiation of this system.

The avian song control system presents yet another pattern of steroid action. Male zebra finches sing more than females, and the system of nuclei responsible for song production is sexually dimorphic. Treatment of newly hatched female finches with estrogen masculinizes both the pattern of singing and the size of sexually dimorphic song system nuclei. Surprisingly, neither castration of young birds nor blockade of steroid receptors prevents sexual differentiation of these brain nuclei. *In vitro* studies of isolated brain slices provide a possible explanation: the active hormone (estrogen) is produced by forebrain neurons, which then directs sexual differentiation locally. However, even in identical culture conditions the slices derived from the brains of male birds produced more estrogen than those derived from females, suggesting that the male genotype may cause a higher level of hormone production in the male brain, independent of hormone levels produced by the gonad. Evidence from mammalian systems also supports a possible role for genetic signals in sexual differentiation of the CNS. Dispersed cultures of midbrain dopaminergic neurons derived from male and female mice placed *in vitro* on embryonic day 14, a time that precedes the normal increase in testosterone levels, nevertheless develop differently: the cultures derived from males end up with significantly more neurons. Moreover, similar embryonic cultures derived from mice lacking the Sry gene, and therefore lacking testicular testosterone secretion, have a greater number of dopaminergic neurons when derived from male embryos compared to those derived from females. The laboratory of Arthur Arnold used transgenic mice, engineered to display gonadal sex independent of chromosomal sex, to study correlations between Y chromosomal genes and sex differences in behavior and brain chemistry. Together, these findings suggest that multiple cell types may be masculinized by genetic signals. However, it should be noted that the overwhelming majority of sexually dimorphic traits evaluated in this transgenic paradigm support a dominant role for sex steroids in the control of sexual differentiation.

Neurobiology of Sexual Differentiation

Few studies have provided definitive proof that individual sexual dimorphisms are responsible for specific aspects of functional sex differences. In large part this is due to the fact that the study of sexual dimorphisms proceeded well ahead of our understanding of the functional neural systems that mediate sexually dimorphic behavior and physiology. However, an emerging perspective is that sexually dimorphic aspects of neural circuits conveying or decoding sensory information, or that control motor output systems, function differently in males and females because of structural differences in their organization. Alternatively, there may be differences in the way these pathways respond to changes in hormone levels acutely. Brain function is dependent on the numbers, organization, and patterns of connections that constitute functional neural systems, so it follows that developmental events that alter these parameters of neural circuits will impact their function. Numerous regions in the mammalian forebrain have been identified as sexually dimorphic, and their normal development is dependent on testosterone exposure during the perinatal period. For hypothalamic regions, conversion of testosterone to estradiol and activation of either or both estrogen receptors (ER-α and ER-β) are initial events in the process. These hormone receptors impact development through changes in gene expression that ultimately cause changes in cell number, morphology, neurochemistry, and connectivity among a wide variety of sexually dimorphic neural circuits.

Sex Steroids Regulate Programmed Cell Death

Assembly of neural circuits involves generation and differentiation of component populations of neurons, migration of neurons to resident sites, and establishment of functional connections with other parts of the nervous system. Hormonal regulation of neurogenesis does not appear to play a major role in specifying sexually dimorphic numbers of neurons in the CNS; however, estrogen appears to influence neurogenesis in mature animals. Nor does formation of sexually dimorphic nuclei appear to be influenced significantly by hormonal regulation of neuronal migration patterns; however, it is possible that experimental limitations are largely responsible for a lack of definitive information on migration. The cumulative evidence indicates that the major way in which sex steroid hormones alter neuronal number in sexually dimorphic regions of the CNS is by influencing cell death. Testosterone decreases cell death in the MPN, BSTp, and SNB, all of which have more neurons in males. In contrast, testosterone exposure increases cell death in the AVPV, where cell number is greater in females. In each of these regions, dying neurons display the morphological features of an apoptotic cell death mechanism, including DNA fragmentation. Moreover, cell loss in the AVPV appears to depend on activation of caspases, enzymes that mediate key aspects of the apoptotic cell death cascade. Although the evidence for cell death as a mechanism underlying sexual differentiation of neuronal number is most thoroughly documented for these nuclei, there is compelling support for cell death as a major determinant in several sexually dimorphic nuclei, in a variety of vertebrate model systems, indicating that hormonally induced control of cell number is a widespread neurobiological mechanism. Less clear is how sex steroid hormones bring about cell death. One possibility is that testosterone regulates expression of antiapoptotic genes such as members of the Bcl2 family, and overexpression of Bcl2 appears to function in a neuroprotective way. The laboratory of Nancy Forger showed that overexpression of Bcl2 in mice reduces the magnitude of sex differences in the SNB and AVPV, suggesting that testosterone regulates cell death in these nuclei through a Bcl2-dependent mechanism. Sex differences were eliminated in mice in which the proapoptotic gene Bax was deleted, demonstrating that Bax is required for sexual differentiation of these nuclei. However, it is a mistake to assume that all sexually dimorphic populations of neurons are regulated through differential control of such proapoptotic or antiapoptotic genes. For example, dopamine neurons in the AVPV are resistant to the developmental effects of manipulating Bcl2 expression or deletion of Bax, despite the overall effects on cell number in the AVPV, indicating that the reduction in dopaminergic neurons seen following exposure to testosterone is caused by a cell death program that is independent of Bcl2 family proteins. In addition to regulation of cell death, testosterone acting early in development may cause lasting changes in neurotransmitter phenotype. In the BST of the rat, perinatal steroids appear to specify a subset of galanin-containing neurons to co-express vasopressin, consistent with the notion that hormones acting perinatally may induce marked and lasting changes in gene expression that lead to sex-specific patterns of neuronal differentiation.

Sex Steroids Specify Patterns of Connectivity

Axon guidance The assembly of functional neural circuits requires neurons to establish polarity, undergo dendritic morphogenesis, and extend axons that traverse significant distances in the CNS guided to their

targets by a complex array of molecular guidance cues. Estrogen is capable of exerting a direct trophic action on neural tissue, especially if the neurons express estrogen receptors. For example, hypothalamic explants extend a profusion of neurites in response to estrogen treatment, suggesting that sex steroids can promote axonal growth. Two sexually dimorphic nuclei in the zebra finch song control system, designated the higher vocal center (HVC) and the robust nucleus of the archistriatum (RA), play important roles in the motor control of song production. A massive projection from the HVC to the RA is required for expression of song in adulthood, and this pathway forms in male but not female finches. Formation of this pathway can be induced in female finches by exposure to exogenous estrogen, which *in vivo* appears to be derived locally rather than from the gonad. In contrast, neurons in the preoptic region of the quail hypothalamus provide sexually dimorphic projections to the midbrain periaqueductal gray. These neurons contain aromatase and convert gonadally derived testosterone to estrogen during development. The homologous pathway in rats shows a similar sex difference, with more preoptic neurons innervating the periaqueductal gray in males than in females. The most dimorphic pathway in the mammalian forebrain appears to be a projection from the BSTp to the AVPV (Figure 3). The intensity of this limbic–hypothalamic pathway is at least an order of magnitude greater in males relative to the homologous pathway in females, and it appears to develop through a directed mechanism rather than forming in both sexes and then regressing in females. The molecular mechanisms directing development of the BSTp-to-AVPV pathway remain unknown, but it is clear that testosterone acts on the AVPV target cells to promote innervation by BSTp neurons. Because the BSTp is larger in males, it is not surprising that it provides a more robust projection to its targets in males. However, the density of projections from the BSTp to the AVPV in male rats is much greater than would be predicted simply on the basis of the greater number of neurons in the BSTp. Numbers of projection neurons are often influenced by the size of their targets, but the AVPV contains fewer neurons in males relative to females, which would argue for a weaker input in males if target size were controlling the number of afferents from the BSTp. Thus, exposure to sex steroids during the perinatal period decreases the number of cells in the AVPV while increasing the innervation of the remaining cells, resulting in a hormonally directed increase in the convergence of BSTp inputs onto AVPV neurons in males. Because the BSTp conveys olfactory information to the hypothalamus, the hormonally directed differentiation of the BSTp-to-AVPV pathway means that sensory information will impact certain hypothalamic circuits differently in males and females. Innervation of perineal muscles by the SNB is also regulated by target-derived factors. In this system, testosterone acts on the muscles to induce endogenous factors, which then act on SNB motor neurons to mediate

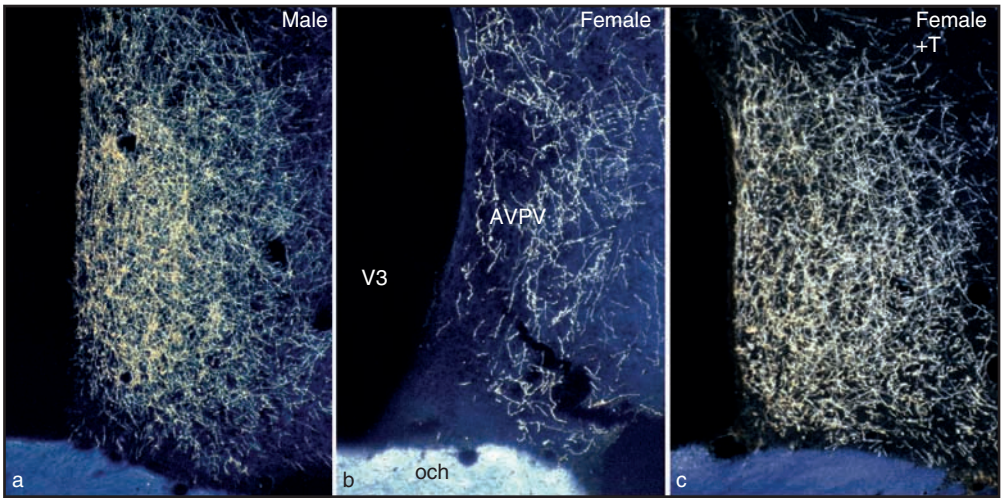

Figure 3 Sexually dimorphic limbic–hypothalamic pathway. The principal nucleus of the bed nuclei of the stria terminalis (BSTp) innervates the anteroventral periventricular nucleus (AVPV) of the hypothalamus. The axonal tracer PHAL was used to label neural projections from the BSTp to the AVPV in male (a), female (b), and testosterone-treated female (c) rats. Treatment of newborn females with testosterone induces innervation of the AVPV by BSTp neurons, thereby changing this pathway from the female to the male pattern. och, optic chiasm; V3, third ventricle. Adapted from Gu G, Cornea A, and Simerly RB (2003) Sexual differentiation of projections from the principal nucleus of the bed nuclei of the stria terminalis. *Journal of Comparative Neurology* 460: 542–562.

sexual differentiation of the SNB. One such factor has been identified as the ciliary neurotrophic factor, but other factors that influence growth and maintenance of neuromuscular junctions may also play a role in specifying innervation of perineal musculature by SNB neurons. The brain-derived neurotrophic factor (BDNF) is important for development of the avian song control nuclei, but in contrast to the SNB system, BDNF appears to act through both retrograde and nonretrograde signaling mechanisms.

Synaptogenesis Sex steroid hormones also appear to be involved in regulation of dendritic structure and synapse formation in sexually dimorphic regions. *In vitro* experiments document the ability of estrogen to promote dendritic length and branching in embryonic limbic and hypothalamic neurons. Estrogen specifies sexually dimorphic patterns of dendritic morphology in the hypothalamus of postnatal rats, and glial cells appear to play a determinant role in this process. Sexually dimorphic patterns of synapses on hypothalamic neurons have been identified that are determined by exposure to testosterone during perinatal life. In the arcuate nucleus of the hypothalamus, synapses on both dendrites and cell bodies of neurons are more abundant in females. In contrast, there is a different synaptic pattern in the ventromedial hypothalamic nucleus, in which the density of synapses is greater in males, indicating that there is considerable regional variation in sexually dimorphic patterns of synaptogenesis. Whether these different patterns of synaptology are due to differential innervation by regions with varying numbers of afferent neurons or are locally specified by postsynaptic mechanisms is unknown. The arcuate nucleus is innervated by the AVPV, which contains more neurons in females and provides more robust innervation of the arcuate nucleus. Similarly, several regions that have greater numbers of neurons in males innervate the ventromedial nucleus. However, inputs to the ventromedial nucleus from the BSTp are similar in males and females, even though the BSTp provides a sexually dimorphic projection to the AVPV, indicating that at least for the BSTp, hypothalamic targets play an important role in specifying patterns of innervation. Thus, in addition to regulating the numbers of neurons that reside in sexually dimorphic nuclei, perinatal testosterone specifies patterns of connectivity in multiple neural networks. The functional consequence of these sex differences is that the input–output relationships of neurons comprising key components of both sensory and motor pathways are different in males and females, which means that information is processed in unique ways, leading to expression of sex-specific behaviors.

The Role of Receptors for Estrogen and Testosterone

A key developmental factor in the regulation of sexual differentiation of the brain is expression of receptors for estrogen and testosterone by neurons in sexually dimorphic nuclei (**Figure 4**). All of the major sexually dimorphic nuclei in the mammalian brain contain high densities of neurons that express ER-α, AR, or ER-β during the perinatal period, when testosterone exerts its greatest effect on brain development. These receptor proteins function primarily as ligand-activated transcription factors that regulate cell development through regulation of gene expression. However, sex steroids can also influence the activity of neurons through what appear to be membrane-associated receptors. Because neural activity has a profound influence on neuronal differentiation and formation of connectivity, such membrane receptor-mediated effects of sex steroids may alter development of regions that do not express the nuclear steroid receptors in abundance. In the rodent, estrogen receptors are responsible for regulation of sexual differentiation of the hypothalamus, whereas the AR appears to mediate most of the sex-specific developmental events in telencephalic regions. This regional specification of receptor function may underlie the enhanced role of the AR in sexual differentiation of behavior in primate species, in which telencephalic regions play a more dominant role in controlling behavior.

The signaling events underlying the developmental actions of sex steroid hormones appear to interact at several levels with neurotrophin signal transduction pathways. For example, estrogen can activate the mitogen-activated protein kinase cascade in postnatal neurons as well as regulate the activity of transcription factors such as c-fos and the cAMP-regulated enhancer binding protein (CREB). These cellular signals in turn play an important role in coupling neurotrophin signaling to the nucleus. Work from the laboratory of Margaret McCarthy has directly implicated CREB in sexual differentiation of copulatory behavior. Few downstream targets of these hormonally activated signaling pathways have been identified, but similarities between some of the morphological effects of estrogen and molecules such as BDNF suggest there is a convergence of neurotrophin and sex steroid signaling on common pathways impacting development. These signaling pathways are commonly activated by a variety of environmental factors and are sensitive to the overall effects of experience. It is therefore intriguing to speculate that the observed interactions between developmental

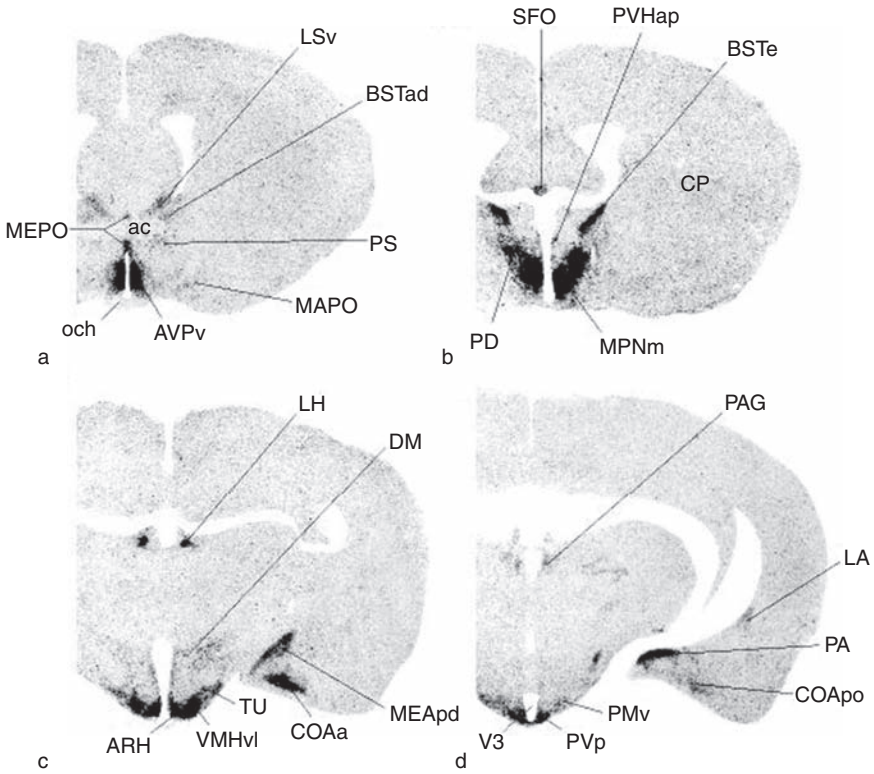

Figure 4 The distribution of forebrain neurons that express estrogen receptor (α form). Estrogen receptors are expressed in distinct regions of the forebrain in adulthood, as illustrated here by expression of ER-α mRNA, and during development. Sexually dimorphic regions of the forebrain tend to contain high densities of neurons that express ER-α. ac, anterior commissure; ARH, arcuate nucleus of the hypothalamus; AVPv, anteroventral periventricular nucleus; BSTad, anterodorsal nucleus of bed nuclei of the stria terminalis; BSTe, encapsulated (principal) nucleus of bed nuclei of the stria terminalis; COa, cortical nucleus of the amygdala (anterior part); COApo, cortical nucleus of the amygdala (posterior part); CP, caudoputamen; DM, dorsomedial hypothalamic nucleus; LA, lateral nucleus of the amygdala; LH, lateral habenula; LSv, lateral septal nucleus (ventral part); MAPO, magnocellular preoptic nucleus; MEPO, median preoptic nucleus; MeApd, medial nucleus of the amygdala (posterodorsal part); MPN, medial preoptic nucleus (m, medial part); PA, posterior nucleus of the amygdala; PAG, periaqueductal gray; PD, posterodorsal preoptic nucleus; PMv, ventral premammillary nucleus; PS, parastrial nucleus; PVHap, paraventricular nucleus of the hypothalamus (anterior parvicellular part); PVp, posterior periventricular nucleus; och, optic chiasm; SFO, subfornical organ; TU, tuberal nucleus; V3, third ventricle; VMHvl, ventromedial hypothalamic nucleus (ventrolateral part). Reproduced from Simerly RB (1995) Hormonal regulation of limbic and hypothalamic pathways. In: Micevych PE and Hammer RP Jr. (eds.) *Neurobiological Effects of Sex Steroid Hormones*, pp. 85–116. New York: Cambridge University Press.

events specified by hormones and the impact of external environmental influences on behavior may be linked through common intracellular signaling pathways.

See also: Gonadal Steroid Actions on Brain; Neurosteroids; Sexual Differentiation of the Brain; Stress Response: Sex Differences; Vasopressin/Oxytocin and Receptors.

Further Reading

Arnold AP (2002) Concepts of genetic and hormonal induction of vertebrate sexual differentiation in the twentieth century, with special reference to the brain. In: Pfaff D, Arnold A, Etgen A, Fahrbach S, and Rubin R (eds.) *Hormones, Brain and Behavior*, pp. 105–135. San Diego, CA: Academic Press.

Arnold AP, Xu J, Grisham W, Chen X, Kim Y-H, and Itoh Y (2004) Minireview: Sex chromosomes and brain sexual differentiation. *Endocrinology* 145: 1057–1062.

Balthazart J and Adkins-Regan E (2002) Sexual differentiation of brain and behavior in birds. In: Pfaff D, Arnold A, Etgen A, Fahrbach S, and Rubin R (eds.) *Hormones, Brain and Behavior*, pp. 223–301. San Diego, CA: Academic Press.

de Vries GJ and Simerly RB (2002) Anatomy, development, and function of sexually dimorphic neural circuits in the mammalian brain. In: Pfaff D, Arnold A, Etgen A, Fahrbach S, and Rubin R (eds.) *Hormones, Brain and Behavior*, pp. 137–191. San Diego, CA: Academic Press.

Forger NG (2006) Cell death and sexual differentiation of the nervous system. *Neuroscience* 138: 929–938.

Forlano PM, Schlinger BA, and Bass AH (2006) Brain aromatase: New lessons from non-mammalian model systems. *Frontiers in Neuroendocrinology* 27: 247–274.

McCarthy MM, Auger AP, and Perrot-Sinal TS (2002) Getting excited about GABA and sex differences in the brain. *Trends in Neurosciences* 25: 307–312.

McCarthy MM and Konkle AT (2005) When is a sex difference not a sex difference? *Frontiers in Neuroendocrinology* 26: 85–102.

Morris JA, Jordan CL, and Breedlove SM (2004) Sexual differentiation of the vertebrate nervous system. *Nature Neuroscience* 7: 1034–1039.

Simerly RB (2002) Wired for reproduction: Organization and development of sexually dimorphic circuits in the mammalian forebrain. *Annual Review of Neuroscience* 25: 507–536.

Swaab DF, Chung WC, Kruijver FP, Hofman MA, and Ishunina TA (2002) Sexual differentiation of the human hypothalamus. *Advances in Experimental Medicine and Biology* 511: 75–105.

Wagner CK (2006) The many faces of progesterone: A role in adult and developing male brain. *Frontiers in Neuroendocrinology* 27: 340–359.

Wallen K (2005) Hormonal influences on sexually differentiated behavior in nonhuman primates. *Frontiers in Neuroendocrinology* 26: 7–26.

Sexual Selection and the Evolution of Animal Signals

W A Searcy, University of Miami, Coral Gables, FL, USA
S Nowicki, Duke University, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Darwin defined sexual selection in *On the Origin of Species* as a mechanism that "depends, not on a struggle for existence, but on a struggle between the males for possession of the females; the result is not death to the unsuccessful competitor, but few or no offspring." Thus sexual selection is a form of selection in which variation in fitness is caused by variation in mating success, and which favors traits associated with enhanced mating success. Darwin's initial description of sexual selection – "a struggle between the males for possession of the females" – implies that he had in mind only aggressive competition among males, but the paragraphs that followed make it clear that Darwin recognized that the struggle "is often of a more peaceful character," with males competing through displays to attract females for mating. Darwin later expanded on these themes in *The Descent of Man, and Selection in Relation to Sex*, treating at length both of the major components of sexual selection: intrasexual selection through aggressive competition and intersexual selection through mate choice. Here we concentrate on the intersexual, mate choice component and its effects on animal communication.

Darwin thought that it was almost universally true that males are more eager to mate than are females and that as a consequence, females are the sex that exercises mate choice. This pattern is still considered the rule, though exceptions are now recognized, both cases in which males and females are almost equally selective and cases in which male choice of mates predominates. Theory successfully explains both the rule and the exceptions, either on the basis of the relative parental investment made by the two sexes or, almost equivalently, on the basis of maximal rates of reproduction. In brief, the sex that makes the greater investment in producing and rearing offspring, or that has the lower maximal rate of reproduction, limits reproduction in the other sex; consequently, the limited sex competes to mate, while the limiting sex exercises choice. Females typically provide greater parental investment than males do, whether in taxa in which both sexes invest only in gametes or in groups such as mammals and birds, with extensive parental care. Largely as a result of these parental investment patterns, females typically have lower maximal rates of reproduction. Thus sexual biases in both parental investment and maximal rates of reproduction predict the predominance of female choice.

Many of the greats of evolutionary biology who followed Darwin were skeptical of the importance of female choice as a selective mechanism. Alfred Russel Wallace, for example, thought "it almost inconceivable that female preference could have been effective in the way suggested," that is, in bringing about the evolution of male displays. Such skepticism may be attributed in part to unwary references by Darwin to an "aesthetic capacity" in female animals, allowing them "an appreciation of the beautiful in sound, colour or form. . . ." Attribution of an esthetic sense to animals, or even of simple choice, implied a higher level of cognition than many scientists were willing to accede to birds of paradise and peafowl, let alone to crickets and moths. Nowadays, however, evolutionary biologists realize that the kinds of mating preferences observed in many animals do not actually call for much in the way of cognitive ability. The words 'choice' and 'preference' currently are used to describe biases in the probability that females will mate with males having certain traits, without regard to whether these biases require cognitive processing. From an evolutionary perspective, what matters most is whether female mating behavior is biased in a way that affects male mating success, not the mechanism – cognitive or otherwise – that underlies the bias. Moreover, it is now realized that, though we may be tempted to label the preferences of nonhuman animals as esthetic, these preferences do not necessarily mesh with our own sense of what is beautiful, as should be conceded by anyone who observes such preferences as that of female common grackles for the songs of their own species.

Another reason for Wallace's skepticism about female choice was the lack of empirical evidence for the hypothesized preferences. One could in theory explain the evolution of the extravagant song of a nightingale or the brilliant coloration of a coral reef fish by positing the appropriate female preferences, but there was no evidence that the required preferences actually existed. Since Wallace's day, however, and in particular during the past 30 years, a great deal of evidence has supported the existence of female preferences in various animal groups. More often than

not, the female preference is for a male trait that we would label as a signal, and which is therefore an element of animal communication.

Evidence of Female Preferences Based on Male Signals

An excellent illustration of female preferences based on male signal characteristics is provided by the work of Michael Greenfield and colleagues on female choice in the lesser wax moth (*Achroia grisella*). Lesser wax moths are obligate symbionts of honeybees (*Apis mellifera*), feeding as larvae on honeycomb and other materials found in honeybee colonies. As adults, male lesser wax moths aggregate in the vicinity of a honeybee colony and produce sexual displays to attract females for mating. The primary display is an ultrasonic acoustic signal produced by fanning the wings. The sound-producing organs are a pair of tymbals located at the base of the forewings; these are hardened, platelike structures that produce sound when they buckle either in or out. A single stroke of the wings produces a pair of sound pulses, one from the left tymbal and one from the right, slightly offset in time (**Figure 1(a)**). Pairs of pulses are produced by both the downstroke and the upstroke of the wings. Males also release pheromones in conjunction with wing fanning, but playback of the sound from a loudspeaker is as effective in attracting females as is a live signaling male, indicating that the ultrasonic signal is the only cue essential for mate attraction.

Female lesser wax moths have been tested in the laboratory for preferences based on acoustic properties of male sounds, using phonotaxis to a loudspeaker playing a signal as a measure of preference for that signal. In each test, a female was given a choice between a standard stimulus, one having the average properties of a male signal, and a stimulus artificially modified along one dimension. Tested this way, females show a preference based on call amplitude, preferring louder calls across the entire range of amplitudes presented (**Figure 1(b)**). This preference should generate directional selection on males to produce signals of higher amplitude. Females also show a preference based on call rates, but here the shape of the preference function is different: instead of always preferring more-extreme signals, females most prefer signals of an intermediate rate, with response falling away for higher as well as lower rates (**Figure 1(c)**). The most preferred rate is higher than the mean rate in the male population (**Figure 1(c)**), however, so this preference should also impose directional selection, in this case for higher signal rates. Finally, females show a preference based on 'asynchrony interval,' the time interval between the onsets of the two pulses in a pair. This preference function shows yet another shape, with females discriminating against signals with very short asynchrony intervals but treating all signals with higher intervals equivalently (**Figure 1(d)**).

Further experiments confirmed these preferences under more naturalistic conditions. Series of females were allowed to choose between sets of four males confined in separate cages. Approach was again assumed to indicate choice. Partial regression coefficients indicate that the number of females attracted by a male is significantly positively related to each of the three signal characteristics mentioned above: call amplitude, call rate, and asynchrony interval.

These results from lesser wax moths are typical of a large body of evidence from insects and anurans showing female preferences for male displays that have greater acoustic energy. In a review of results on several dozen species of frogs and insects, Gerhardt and Huber found that females prefer higher male call rates over lower in the great majority of the studied species. Similarly, females typically prefer male calls of longer duration to shorter ones. By contrast, preferences based on the rate of pulses within a call or on carrier frequency are most often stabilizing, with calls of average properties preferred over either extreme. Exceptions to all these generalizations have been found; in particular, females in several species of frogs show preferences for male calls of lower frequency than the population mean.

In birds, female preferences based on vocalization rates are again observed quite commonly, with females consistently preferring males that sing at higher rates. A second common preference in female birds is for local songs over foreign ones. An example is provided by our work on geographic preferences in female song sparrows (*Melospiza melodia*). We tested for preferences using the solicitation display assay, in which captive females are first treated with estradiol and then exposed to playback of male songs. Subjects respond to conspecific songs with a species-typical precopulation display, and the magnitude of this response is used to infer preferences. We tested females from a site in Pennsylvania for discrimination between a sample of local songs and songs recorded at various points along a transect leading away from the local site. Females showed no discrimination between local songs and songs from the first foreign site (at 18 km) but gave significantly more displays for local songs than for songs from the second foreign site, located 34 km from the start of the transect (**Figure 2(a)**). Females preferred local songs to songs from all sites

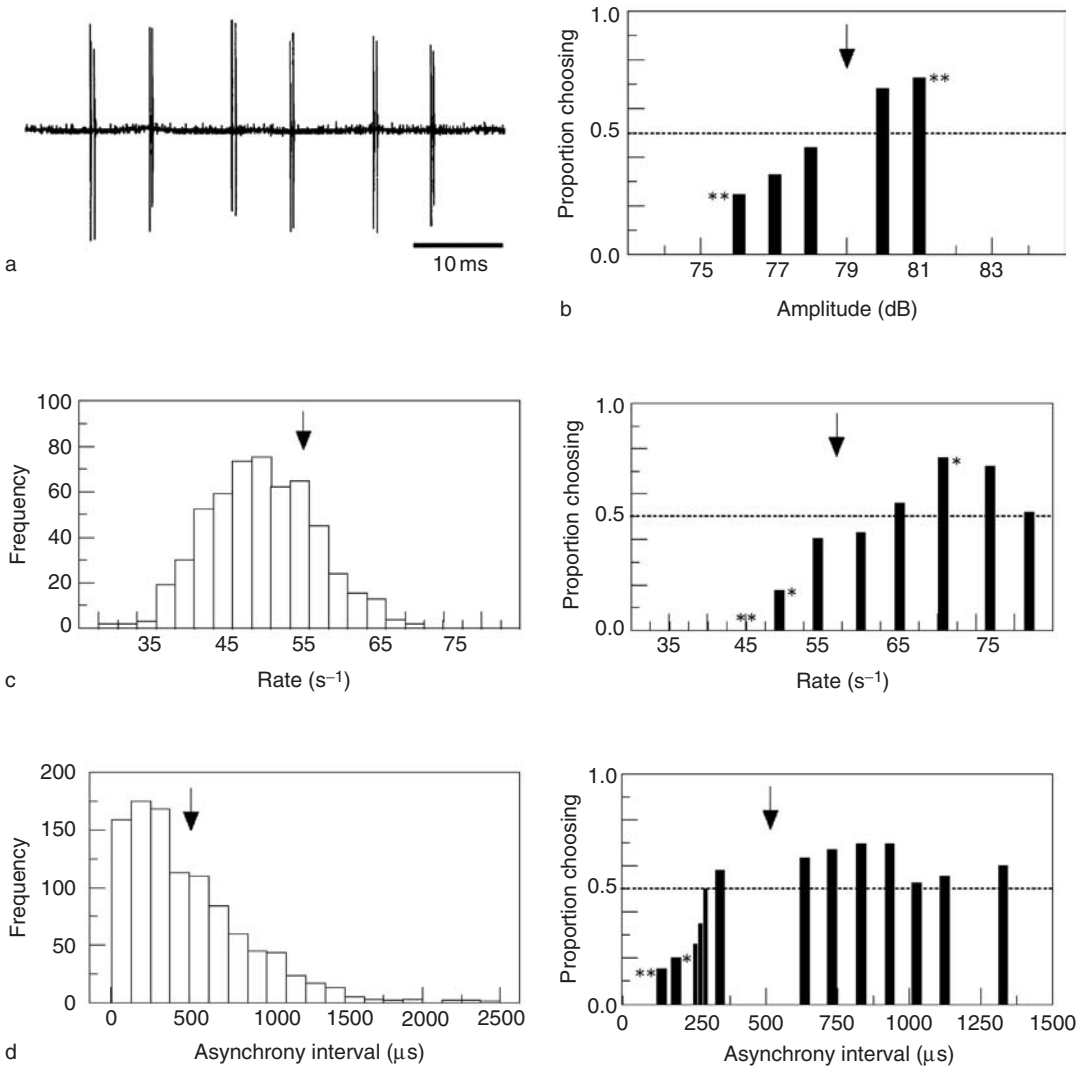

Figure 1 The shape of female preference functions for male ultrasonic signals in lesser wax moths (*Achroia grisella*), as measured by female orientation in two-speaker playback tests. (a) Oscillogram showing the temporal sequence of pulses produced by male wax moths. One pair of pulses is produced by each downstroke of the wings and the subsequent pair by the following upstroke. (b) Preferences of females for male calls differing in amplitude. Bars indicate the proportion of subjects choosing calls of specific amplitudes over a standard call in two speaker choice tests. The arrow indicates the amplitude of the standard call. (c) On the left is the frequency distribution of call rates in 529 calls of 33 males. Rate is the reciprocal of the time between one downstroke and the next. On the right are the preferences of females for calls differing in rate. Bars indicate the proportion of subjects choosing calls of specific rates over a standard rate. The arrow indicates the standard rate. Note that the most preferred call rate is well above the population mean. (d) On the left is the frequency distribution of asynchrony intervals in 1057 measurements of calls from 33 males. Asynchrony interval is the time between the beginning of the first pulse in a pair and the beginning of the second. On the right are the preferences of females for calls differing in asynchrony intervals. Bars indicate the proportion of subjects choosing specific intervals over a standard interval. The arrow indicates the standard interval. * indicates a significant preference ($p < 0.25$) for or against the indicated signal without correction for multiple tests. ** indicates a significant preference ($p < 0.05$) for or against the indicated signal after correction for multiple tests. Note that the only significant preferences are against especially short asynchrony intervals. Reprinted from Jang Y and Greenfield MD (1996) Ultrasonic communication and sexual selection in wax moths: Female choice based on energy and synchrony of male signals. *Animal Behaviour* 51: 1095–1106, with permission from Elsevier.

farther away than 34 km, and in general the strength of their preferences increased with increasing distance to the foreign site.

A third preference common in female birds is for more-complex song or more-complex song repertoires. Song sparrows again provide an illustration. Male song sparrows possess repertoires of 5–13 song types, each of which is a discretely different version of the species song. Female song sparrows tested in captivity perform more courtship for 32 songs of four types than for

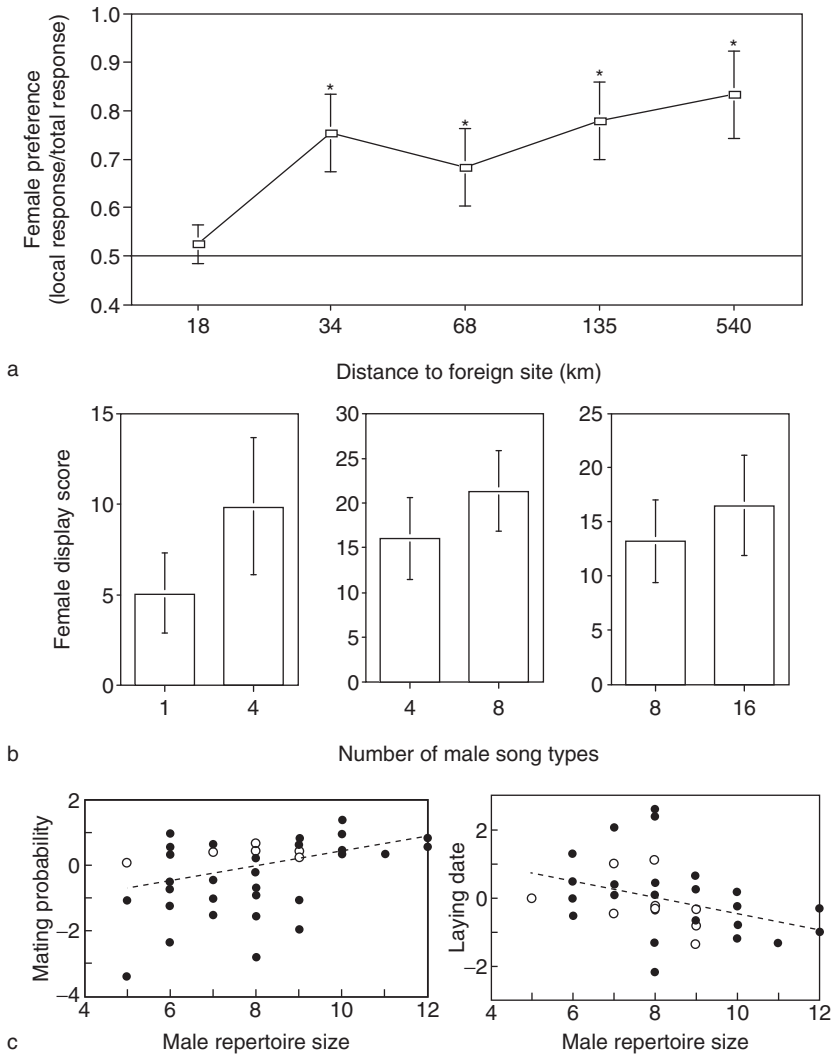

Figure 2 Female preferences based on song in song sparrows (*Melospiza melodia*). (a) Preferences based on geographic origin of song. Preference is measured as the number of courtship displays given by females in response to songs from their local site, as a proportion of the total given to songs from the local site plus the foreign site. An asterisk denotes a significant preference for the local songs over the foreign. Error bars indicate standard errors. Data are from Searcy WA, Nowicki S, Hughes M, and Peters S (2002) Geographic song discrimination in relation to dispersal distances in song sparrows. *American Naturalist* 159: 221–230. (b) Female preferences based on numbers of song types. Preferences are measured using display scores that combine the number and intensity of displays. Females were tested in single-speaker playbacks with (from left to right) 32 songs of one or four types; 64 songs of four or eight types; and 64 songs of eight or 16 types. Females in each test showed significantly higher response to the greater number of song types. Data are from Searcy WA (1984) Song repertoire size and female preferences in song sparrows. *Behavioral Ecology and Sociobiology* 14: 281–286; and Searcy WA and Marler P (1981) A test for responsiveness to song structure and programming in female sparrows. *Science* 213: 926–928. (c) Field measures of female preferences for males with larger repertoires in song sparrows. On the left, a male's residual probability of mating in his first year, after controlling for territory size and population sex ratio, increases with his repertoire size. On the right, a first-year male's residual laying date, after controlling for between-year variation in the timing of breeding, decreases as his repertoire size increases. Thus first-year males with large repertoires are more likely to mate and mate earlier than males with small repertoires. (c) Reproduced from Reid JM, Arcese P, Cassidy ALEV, et al., (2004) Song repertoire size predicts initial mating success in male song sparrows, *Melospiza melodia. Animal Behaviour* 68: 1055–1063, with permission from Elsevier.

32 songs of one type. Similarly, holding the number of songs constant, females perform more courtship for eight song types than for four, and more for 16 types than for eight (Figure 2(b)). In the field, Jane Reid and colleagues have shown that for male song sparrows holding territories for the first time, the probability of attracting a mate increases with repertoire size, holding the yearly sex ratio constant (Figure 2(c)). Among the males that do breed, laying date decreases as repertoire size increases, indicating that males with large repertoires are chosen before males with smaller repertoires (Figure 2(c)).

Female preferences can, of course, be based on signals other than acoustic ones. Preferences based

on color are common in both birds and fishes. In sticklebacks (*Gasterosteus aculeatus*), females prefer to lay eggs in the nest of males with red ventral coloration over males lacking red. Females show the same preference when the red is added experimentally as when the color is natural. Similarly, male house finches (*Carpodacus mexicanus*) are more likely to find a mate the redder their natural coloration, and their chances of being chosen by a female can be altered by manipulating their red coloration in either direction. Other visual ornaments known to be important to female preferences in birds include the exaggerated trains or tails of male common peafowl (*Pavo cristatus*), long-tailed widowbirds (*Euplectes progne*), and barn swallows (*Hirundo rustica*). In satin bowerbirds (*Ptilonorhynchus violaceus*), males decorate individual mating arenas ('bowers') with objects such as blue and yellow flowers, and the number of these external ornaments affects the likelihood that females will mate with a bower's owner.

In some animals, female preferences are based on olfactory signals rather than on visual or acoustic displays. In house mice (*Mus musculus*), females show preferences for the odors of healthy males over those of males experimentally infected with viral or protozoan parasites. Female house mice also show preferences for males whose major histocompatibility complex genotypes are dissimilar to their own, and these preferences are again thought to be based on odor. In Iberian rock lizards (*Lacerta monticola*), females prefer to associate with the odors of older males over those of younger males, and with the odors of males in good body condition over those of males in poor condition. In field crickets (*Gryllus integer*), females exhibit preferences based on male acoustic signals, but they also show preferences based on male odors, preferring odors of dominant males to odors of subordinates. Conversely, female cockroaches prefer the mix of pheromones produced by subordinate males to the mix produced by dominants.

Evolution of Female Preferences: Theory

If we accept the evidence for female preferences favoring such traits as the exaggerated tails of peacocks and the ultrasonic displays of male wax moths, we can explain the evolution of these traits in males; however, we have solved one difficulty only at the cost of creating another – explaining the evolution of female preferences. RA Fisher was the first to grapple with this problem, remarking that "the tastes of organisms, like their organs and faculties, must be regarded as the product of evolutionary change, governed by the relative advantage which such tastes may confer. . . ." In other words, a female preference is a trait that must have itself evolved, presumably in most cases in response to selection in one form or another. Understanding how selection works in such cases turns out to present considerable difficulties. Fisher stated his own solution concisely, perhaps too concisely given the subtlety of his proposal. Suppose females express a preference based on a male trait in a species in which female preferences have a considerable effect on the reproductive success of males. If the male trait has an initial selective advantage, for example, in promoting male survival, then the preference will begin to spread among the females because of the effect it has on the fitness of their sons. The increase in the preference among females then gives an additional advantage to the preferred trait in males, through increased mating success. As long as the sons of females exercising the preference have a net advantage, due to the combination of the initial natural selective advantage and the subsequent sexual selective advantage, the intensity of the preference will continue to increase. In this way, the male trait can be exaggerated past the point at which its initial natural selective advantage has been lost. The core idea expressed by Fisher – that the exaggeration of a female preference for a male trait leads to stronger selection on both the trait and the preference, leading to a further exaggeration of the preference and the preferred trait – is now termed the 'Fisher' or 'runaway' mechanism.

The Fisher mechanism may seem hopelessly circular, but its logic has been confirmed repeatedly during the past quarter century by mathematical models of various types. The gist of these models can be summarized as follows. We start with a female preference already existing at some low frequency, perhaps due to an initial advantage from securing mates of high viability. Females with the preference mate with males with the preferred trait, and their offspring therefore inherit both the genes for the preference and the genes for the male ornament. This creates a genetic correlation between trait and preference. Consequently, when sexual selection favors the male trait because of its advantage in mating, the preference is favored indirectly, because of the genetic correlation. As the preference increases, sexual selection in favor of the ornament strengthens, which in turn strengthens selection for the preference, and so on. Under certain conditions, such as weak genetic covariance between trait and preference, evolution proceeds toward a line of equilibria, at which the degree of exaggeration of the male trait balances the degree of strength of the preference. Under other conditions, such as stronger genetic covariance, the process can run away toward ever increasing levels of trait and preference. Even in the latter case, the process is

expected to stabilize eventually as the preferred trait becomes increasingly costly under natural selection. The benefit of increased mating success conferred by a bird's long tail, for example, may be outweighed at some point by the cost of poorer flight capabilities associated with having a long tail or the physiological cost of producing a long tail in the first place.

The Fisher mechanism is classified as an indirect selection hypothesis because selection acts on the preference only indirectly, through its genetic correlation with the preferred trait. A second indirect selection hypothesis is provided by the 'good genes' mechanism. Here the male ornamental trait is thought to serve as an 'indicator' of male viability because, for example, only males in good condition are able to develop the ornament fully. Mating between females with the preference and males with the preferred trait creates a genetic correlation between preference, preferred trait, and viability. Natural selection always favors high viability, and the preference and ornamental trait are consequently favored indirectly because of their genetic correlation with viability. A general problem with the good genes hypothesis is that it requires viability to be heritable; this is a problem because natural selection always favors high viability and thus tends to exhaust additive genetic variance for traits contributing to viability. Among several hypotheses for escaping this dilemma, the best known is that of William Hamilton and Marlene Zuk, who suggested that coevolutionary cycles between parasites and hosts ought to maintain additive genetic variance for resistance in hosts and thus the heritability of viability. A preference for males with good genes would then often manifest itself as a preference for males resistant to parasites.

Some theoreticians have considered the good genes mechanism to be a distinct alternative to the Fisher process – and a more parsimonious alternative by virtue of being both less circular and less arbitrary. Those taking this view have assumed that the fitness of offspring of preferred males is a test of these alternatives: under the good genes hypothesis, preferred males should have sons and daughters of superior viability, whereas under the Fisher hypothesis, only sons of preferred males would have an advantage, and the advantage would stem only from superior mating success. Recently, however, others have argued that Fisher effects – those in which sexual selection acts indirectly on the preference – are inevitable in a good genes process and that therefore the two mechanisms cannot be disentangled.

It is also possible for selection to act directly on a preference, because the preference affects the fitness of the female expressing the preference rather than the fitness of her offspring. Direct selection will occur if the preference affects the quality of material benefits provided by the male; such benefits might include nuptial gifts of food, a breeding territory, or resources passed to the offspring in parental care. If a female can observe something about a male, such as his size or vigor, that allows her to assess directly his potential for providing material benefits, then communication does not need to be invoked. If instead a female assesses a male through a signal that correlates with his ability to provide material benefits, then communication is involved. The signal can again be considered an 'indicator' of male quality, albeit of direct material benefits instead of indirect genetic benefits.

Another possibility is for the preference to affect the costs of mate choice to the female; for example, if searching for a mate exposes the female to predation risk, an appropriate preference shortens the search and thus reduces risk. Yet another possibility is that mating with preferred males may be harmful to females, as for example when male seminal fluids contain substances toxic to females, as occurs in *Drosophila*. In this case, females may evolve resistance to the preferred male display, in the sense of an increased threshold for response, which paradoxically may lead to selection for exaggeration of the display in order to overcome resistance.

A final possibility is for a preference to be the consequence of properties of the sensory or neural systems of the female that have evolved in contexts having nothing to do with mating. The preference is then said to represent a 'sensory bias,' and the evolution of a male signal in response to the preference is said to represent 'sensory exploitation.' Sensory bias is sometimes considered to fall in the category of direct selection hypotheses, as natural selection has presumably acted directly on the relevant sensory or neural properties in some other context, such as predator avoidance or food finding. Alternatively, sensory bias can be considered to be a case of no selection, because the mating preference *per se* has not affected either the female's fitness or that of her offspring and thus has not been acted on by selection.

Evolution of Female Preferences: Evidence

To a layperson, the simplest explanation of female preferences might seem to be the good genes hypothesis: a female prefers a male whose displays advertise his health and vigor, and she benefits when he passes those traits to her offspring. To a population geneticist, however, this explanation may seem the least parsimonious of the possibilities, depending as it does on a series of genetic correlations and heritabilities that are all likely to be low. Nevertheless, there is

evidence in some instances that females benefit from their preferences in ways predicted by the good genes hypothesis. In great reed warblers *(Acrocephalus arundinaceus*), for example, females prefer males with larger syllable repertoires, and such males sire offspring that have superior postfledging survival. Similarly, peafowl females prefer males with larger trains, and those males sire offspring that experience higher survival. In gray treefrogs (*Hyla versicolor*), females prefer calls of long duration, and males that give such calls sire offspring that grow at a faster rate and complete development in a shorter interval than the offspring of males giving shorter calls. This advantage, however, is seen only at low densities; at high densities, offspring of males of short call duration complete development faster. Similarly, in the lesser wax moths described above, the offspring of males with attractive ultrasonic signals develop faster under favorable environmental conditions but slower under unfavorable conditions. Thus support for the good genes mechanism must be considered equivocal for gray treefrogs and lesser wax moths.

A crucial prediction of the Fisher mechanism for the evolution of female preferences is that a genetic correlation ought to exist between the preference and the preferred trait. This prediction has been verified in several cases using two methods. In one, artificial selection is imposed on the preferred trait only; a genetic correlation between the two is demonstrated if the preference changes along with the preferred trait. In stalked-eyed flies (*Cyrtodiopsis dalmanni*), for example, females normally prefer males with long eye spans. When males were selected for shorter eye spans for 13 generations, eye stalk length responded, and female preferences also changed, such that females in the selected population came to prefer males with short rather than long eye stalks. Similarly, when artificial selection was imposed on male guppies for greater or lesser area of orange, females in the selected lines also diverged in the strength of their preference for orange males. The second way to demonstrate a genetic correlation is by measuring the covariance of the trait and preference in genetic relatives, such as full sibs, half sibs, or parent and offspring. Using this method, a genetic correlation has been demonstrated in the field cricket *Gryllus texensis* between pulse rate in male calls and female preference for pulse rates, and in sticklebacks between area of red in males and female preference for red in their mates.

A second prediction of the Fisher mechanism is that the traits that make males attractive are heritable, so that attractive fathers produce attractive, or 'sexy,' sons. The predicted heritability has been demonstrated for pulse rate of male calls in the field cricket *G. texensis* mentioned above and for all three of the male call traits contributing to attractiveness in lesser wax moths. The prediction has been verified more directly in the lekking sandfly *Lutzomyia longipalpis*, through evidence of a positive relationship between mating success of fathers and mating success of sons. Pheromones are thought to mediate male attractiveness in this case. The most direct evidence confirming the prediction comes from guppies, in which it has been shown (through the resemblance of half-sibs) that male color pattern is heritable and that there is a genetic correlation between male color pattern and male attractiveness to females. Thus, a female choosing a male based on his attractive color pattern will acquire genes for her sons that enhance their attractiveness.

Although the good genes and Fisher mechanisms receive support in some cases, it is thought that the most parsimonious explanations for the evolution of female preferences are ones in which selection acts directly on a preference. In variable field crickets (*Gryllus lineaticeps*), for example, females prefer males producing calling songs with higher chirp rates and longer chirp durations. Among females maintained on a restricted diet, longevity was positively associated with the chirp duration of the male to which they were mated, while the number of fertile eggs produced was positively associated with male chirp rate. Both effects seem likely to be due to nutrients passed in the seminal fluid, since neither effect occurred in females maintained on a better diet. Both aspects of the preference, for call rate and call duration, affect the female's own fitness, and thus selection acts directly on these preferences. Even in guppies, in which such a good case can be made that females benefit indirectly from mating with preferred males through the enhanced mating success of sons, it may be more important that the preferred males – those with a greater area of orange – have larger sperm stores and deliver more sperm to the female.

Sensory bias proposes that mating preferences are a consequence of properties of a female's sensory or neural systems that have evolved due to selection in some context other than mating. The hypothesis thus predicts that the preference existed before the preferred trait evolved, a prediction not made by other hypotheses for the evolution of preferences. Evidence supporting this prediction comes from the neotropical frog *Physalaemus pustulosus*. Here males produce calls consisting of a low-frequency whine plus one or more high-frequency chucks, and females prefer calls with chucks to calls without them. Females of a related species, *P. coloradorum*, also prefer conspecific calls to which chucks are added, even though males of this species never produce chucks. The preference can be explained by the organization of the anuran ear, which has two papillae, one tuned to low frequencies

and thus stimulated by a whine, and the other tuned to higher frequencies and thus stimulated by chucks. Phylogenetic analysis indicates that both the tuning properties of the ear and the preference for the chuck probably existed in an ancestor that lacked the preference; thus the preference seems to have evolved before the preferred trait. The same pattern of auditory tuning, however, also exists in another *Physalaemus* species that lacks the preference, arguing that the preference is not solely determined by the tuning pattern. Support for a sensory bias explanation has also been found in analyses of female preferences for swords in swordtail fish and for water vibrations in water mites.

Conclusions

Initial skepticism about the plausibility of female choice as a mechanism of sexual selection has given way in the face of overwhelming evidence that preferences for male traits do bias female mating patterns. The male traits on which female preferences act are most often male signal characteristics, and thus sexual selection plays a large role in shaping animal communication. Understanding why sexual selection shapes communication the way it does requires understanding the evolution of female preferences. Various theories have been elaborated on how and why female preferences evolve, including the Fisher mechanism, the good genes hypothesis, sensory bias, and various direct selection explanations. These hypotheses are not mutually exclusive, so it is not surprising that empirical support exists for multiple hypotheses, not only with respect to separate preferences in different species, but even with respect to a single preference in a single species.

See also: Kinship Signals in Animals; Multimodal Signaling in Animals; Pheromones in Mammals; Pheromones and other Chemical Communication in Animals; Signal Identification: Peripheral and Central Mechanisms; Visual Signaling in Animals.

Further Reading

Andersson M (1994) *Sexual Selection*. Princeton, NJ: Princeton University Press.

Clutton-Brock TH and Vincent ACJ (1991) Sexual selection and the potential reproductive rates of males and females. *Nature* 351: 58–60.

Darwin C (1859) *On the Origin of Species*. London: John Murray.

Darwin C (1871) *The Descent of Man, and Selection in Relation to Sex*. London: John Murray.

Fisher RA (1930) *The Genetical Theory of Natural Selection*. Oxford, UK: Clarendon Press.

Gerhardt HC and Huber F (2002) *Acoustic Communication in Insects and Anurans*. Chicago: University of Chicago Press.

Greenfield MD (2002) *Signalers and Receivers: Mechanisms and Evolution of Arthropod Communication*. New York: Oxford University Press.

Hamilton WD and Zuk M (1982) Heritable true fitness and bright birds: A role for parasites? *Science* 218: 384–387.

Jang Y and Greenfield MD (1996) Ultrasonic communication and sexual selection in wax moths: Female choice based on energy and asynchrony of male signals. *Animal Behaviour* 51: 1095–1106.

Jang Y and Greenfield MD (1998) Absolute versus relative measurements of sexual selection: Assessing the contributions of ultrasonic signal characters to mate attraction in lesser wax moths. *Achroia grisella (Lepidoptera: Pyralidae)*. *Evolution* 52: 1383–1393.

Kirkpatrick M and Ryan MJ (1991) The evolution of mating preferences and the paradox of the lek. *Nature* 350: 33–38.

Kokko H, Brooks R, McNamara JM, and Houston AI (2002) The sexual selection continuum. *Proceedings. Biological Sciences/The Royal Society B* 269: 1331–1340.

Reid JM, Arcese P, Cassidy ALEV, et al. (2004) Song repertoire size predicts initial mating success in male song sparrows. *Melospiza melodia*. *Animal Behaviour* 68: 1055–1063.

Ryan MJ and Rand AS (1993) Sexual selection and signal evolution: The ghost of biases past. *Philosophical Transactions of the Royal Society B: Biological Sciences* 340: 187–195.

Searcy WA (1984) Song repertoire size and female preferences in song sparrows. *Behavioral Ecology and Sociobiology* 14: 281–286.

Searcy WA and Marler P (1981) A test for responsiveness to song structure and programming in female sparrows. *Science* 213: 926–928.

Searcy WA and Nowicki S (2005) *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton, NJ: Princeton University Press.

Searcy WA, Nowicki S, Hughes M, and Peters S (2002) Geographic song discrimination in relation to dispersal distances in song sparrows. *American Naturalist* 159: 221–230.

Wallace AR (1889) *Darwinism*. London: Macmillan.

Shape Representation in Inferotemporal Cortex

L Woloszyn and D L Sheinberg, Brown University, Providence, RI, USA

© 2009 Elsevier Ltd. All rights reserved.

The macaque visual system is commonly divided into two parallel processing streams. The dorsal pathway travels from the occipital lobe into the parietal lobe and is thought to be responsible for analyzing the spatial layout and movement of objects. The ventral pathway, which follows the inferior longitudinal fasciculus, extends from the occipital lobe into the anterior and ventral portions of the temporal lobe; unlike the dorsal pathway, it is thought to be primarily responsible for representing those aspects of visual information critical for encoding object identity. Anatomical and physiological evidence indicates that the flow of visual information in the ventral pathway proceeds in a hierarchical fashion through several distinct functional areas. In nonhuman primates, a collection of visual areas that comprises the most anterior and ventral portions of this pathway are generally referred to as the inferotemporal (IT) cortex. As the name implies, this large region of cortex is located on the ventrolateral surface of the central and anterior temporal lobe. A major source of feed-forward input into the IT cortex is area V4, a retinotopically organized visual area that responds to visual information of intermediate complexity, such as oriented contours. Individual neurons within IT, because of their large receptive fields and lack of retinotopic arrangement, integrate this incoming information into even more elaborate representations. Many experiments have shown that single IT neurons respond selectively to the presentation of faces, animals, cars, fruits, and a myriad of other complex images. Coupled with the IT cortex's position in the visual hierarchy, these observations have led researchers to speculate that the IT cortex underlies the perception and representation of complete visual objects.

Selectivity, Invariance, and Tuning Functions

The tuning profiles of individual IT neurons reveal considerable individual variation, ranging from broadly tuned neurons that respond to the presentation of many objects to narrowly tuned neurons that respond extremely selectively to only one test stimulus. The breadth of observed response selectivities is consistent with the remarkable ability of the visual system to simultaneously mediate two seemingly antagonistic processes required for successful visual object recognition. On one hand, the visual system must be able to identify the same visual object regardless of its size, position, angle, and other factors that affect its appearance but not what it is. On the other hand, the visual system must also be capable of making fine discriminations between very similar yet different objects. This interplay between invariance and selectivity is illustrated in Figure 1. Ever since the pioneering studies of Gross and colleagues, in which it was shown that single IT neurons can display extraordinary selectivity for individual visual objects, a major focus of research has been to identify the particular attributes that determine neuronal selectivity.

Shape Selectivity

Implicit in the idea that visual recognition capacities are directly linked to the activity of cells in IT cortex is the notion that those features critical for judging identity will be reflected in individual cell response profiles. Consistent with this suggestion is the observation that IT neurons respond to a given shape independent of the cue used to construct that shape. For example, a bright square patch against a dark background or a square patch of upward-moving dots against a background of downward-moving dots can both indicate the shape of a square. An IT neuron selective for a square will give a vigorous response to both the luminance- and the motion-defined shapes. This finding has led some to propose that shape is the fundamental visual attribute that the IT cortex represents. Support for the shape hypothesis, however, is complicated by the wide array of complex visual images that are used to explore the selectivity of IT neurons. Whereas some labs use images of real objects, other labs decompose these objects into their constituent parts or employ completely arbitrary shapes. Because the concept of shape has been so operationalized, results from one study are difficult, if not impossible, to interpret in the context of another study. Notwithstanding these difficulties, numerous efforts have been undertaken to systematically analyze shape representation in IT cortex. One such approach has been to record from IT neurons while monkeys view a parameterized set of shapes.

What is a parameterized set of shapes? A shape's quantitative description lies in several dimensions. Some simple and concrete examples of these dimensions are the length, position, orientation, and curvature of the line segments that compose a particular shape. Different values along the dimensions lead to

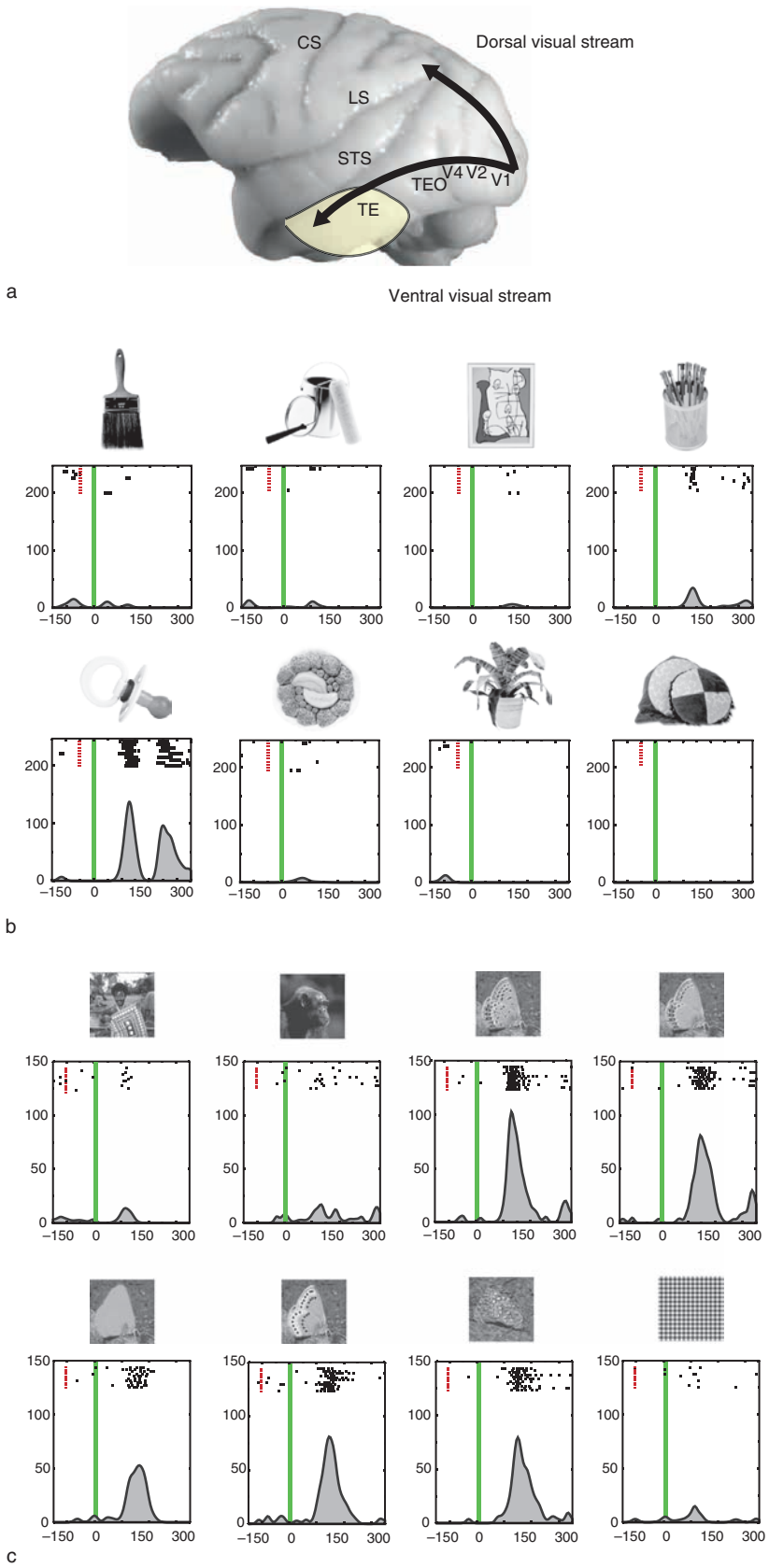

Figure 1 Location of inferotemporal (IT) cortex and the response properties of individual IT neurons. (a) Lateral view of the macaque cortex with the anatomical location of IT shaded in yellow. IT is part of the ventral visual stream, which descends from the posterior pole

different shapes. It is possible that IT neurons explicitly represent these dimensions, but using real-world objects to assess this is difficult because two different real-world objects will differ along several dimensions at a time. This makes it hard to interpret to which dimension a neuron is sensitive. Parameterized shapes overcome this problem because their distinguishing characteristics vary along a known number and identity of dimensions. Hence, a set of parameterized shapes can hold constant all but a few dimensions, allowing for a better understanding of what it is that an individual IT neuron represents. This sort of approach has shown that IT neurons have the capability to encode simple dimensions of shapes. For instance, an IT neuron selective for a triangle's height will systematically increase or decrease its firing rate as a function of the triangle's height. A separate neuron selective for how curved the triangle is will similarly increase or decrease its firing rate as a function of the triangle's curvature. It is interesting that single neurons can be sensitive to changes along more than one dimension. When both dimensions change simultaneously, the neuron's response can be predicted on the basis of its response to the individual changes, implying that the dimensions can be encoded independently.

In most situations, the dimensions that define a shape are numerous and complex. It would seem to be an ineffective and inefficient strategy to represent each of these dimensions separately and independently. So how can such shapes be represented systematically at the neural level? Given a set of complex shapes, some of them are naturally judged to be more similar than others. If the stimuli have been parameterized properly, the perceptual similarities can be plotted in an arbitrary low-dimensional space (e.g., a two-dimensional plane). The larger the distance between a pair of points in this space, the less perceptually similar the two shapes are. It has been shown that IT neurons have tuning functions consistent with the representation of such low-dimensional arrangements of parameterized shapes. Specifically, an IT neuron that has a preferred shape will decrease its response to the other shapes, and the magnitude of the decrease corresponds to the magnitude of the perceived dissimilarity. This coding scheme, depicted in Figure 2, suggests that a clever neural code for representing stimuli is one that constructs arbitrary dimensions which capture most of the variations in a set of shapes.

Parameter Invariance

The ability of IT neurons to detect subtle differences between shapes supports the claim that this population of cells allows observers to make fine discriminations between visual objects. This selectivity, however, must be balanced by an equally important aspect of shape representation: invariance. Any two particular instances of a shape are unlikely to be exactly the same. A shape's size, position, or orientation on the retina will never be exactly the same. The visual system faces the difficult task of identifying shapes independent of these (contextually) irrelevant variations. To couch this in the adopted terminology, IT neurons, in addition to their sensitivity to one set of parameters, must also be insensitive to another set of parameters. What are some of the neural correlates of invariance?

At its anterior location in the ventral visual processing pathway, IT cells have receptive field sizes that are significantly larger than those of cells in more posterior visual areas. Receptive fields as large as 30–50° in diameter and spanning both the contra- and ipsilateral visual fields are encountered routinely during recordings. This size allows individual IT neurons to integrate sensory information from a substantial portion of the visual scene. As such, IT neurons can maintain their relative selectivity for a given shape over large changes in its location. Even though the absolute responses to shapes may well be affected by changes in eccentricity (almost always being reduced with increasing distance from the fovea), the shape selectivity is generally preserved. Likewise, this relative selectivity is preserved across variations in the shape's size. It is intriguing to note that there does appear to be a trade-off in the ability of IT neurons to be very selective for one or a small number of shapes and the ability of IT neurons to demonstrate invariance in the face of image transformations. That is, the responses of narrowly tuned neurons are more likely

of the occipital lobe into the anterior portions of the temporal lobe. The main visual areas comprising the ventral visual stream are indicated: V1, V2, V4, TEO, and TE. CS, central sulcus; LS, lateral sulcus; STS, superior temporal sulcus. (b) The response profile of a narrowly tuned IT neuron, showing high selectivity. This neuron responds selectively to the presentation of the pacifier. None of the other images in the set produce a reliable response. (c) The response profile of a different, broadly tuned IT neuron. The presentation of any kind of butterfly wing elicits a consistent response from this neuron. This neuron's response is somewhat invariant to different instances of a butterfly wing. For each panel in (b) and (c), the visual stimulus is placed above its corresponding raster and spike density function. Each row in the raster (top) is one trial; the black bars indicate individual action potentials for that trial (each trial is marked by the aligned horizontal red ticks). The spike density function (below) is the averaged data for all trials with a particular image. The *x*-axis is time in milliseconds. The green line in the middle of each panel designates time of stimulus onset. Mruczek and Sheinberg (unpublished data).

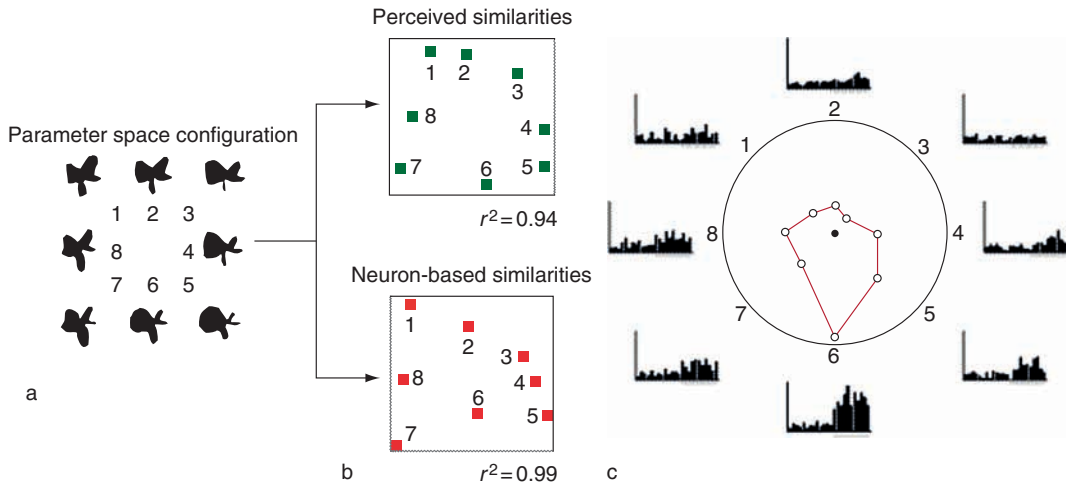

Figure 2 Low-dimensional configurations of parameterized shapes. (a) A set of eight parameterized shapes used to explore the representational scheme in inferotemporal (IT) cortex. Parameter space configuration: The eight shapes varied systematically along two radial frequency components. The shape in the top left has low amplitude values for both components. The shapes in the top right and bottom left have high amplitude values for one or the other component. The shape in the bottom right has high amplitude values for both components. (b) Perceived similarities: monkeys judged a pair of these stimuli (for all possible pairs) as same or different. By means of a technique known as multidimensional scaling, the perceived similarities of the monkey were plotted in a low number of arbitrary dimensions. The low-dimensional configuration of perceived similarities has a squarelike arrangement very similar to the actual parameter space configuration. Neuron-based similarities: multidimensional scaling can also be applied to the neuronal responses obtained while the monkey passively views the shapes. The correspondence between the low-dimensional configuration of the neuron-based similarities and the perceived similarities is readily apparent. (c) The response profile of a single IT neuron to the eight shapes. The inside part of the figure is a polar plot which shows the normalized response of the neuron to each of the eight shapes. On the outside of the polar plot are the spike density functions used to generate the polar plot. This neuron is most responsive to the presentation of stimulus 6. Shapes perceived by the monkey to be the least similar to shape 6 caused the largest decrease in the response of this neuron. From Op de Beeck H, Wagemans J, and Vogels R (2001) Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience* 4: 1244–1252.

to be affected by changes in size and position than are those of broadly tuned neurons, possibly suggesting that distinct populations of neurons within IT cortex mediate selectivity and invariance.

Bell-Shaped versus Monotonic Response Functions

From studies using parameterized shape stimuli has emerged the hypothesis that IT neurons have monotonic tuning functions. In order to explain this coding scheme, it is easiest to contrast it with an alternative, bell-shaped tuning. Bell-shaped tuning refers to tuning functions whereby each neuron has some optimal stimulus to which it will respond, and any deviations from this stimulus will cause a decrease in the response. The most obvious example of this kind of tuning is seen in V1 with respect to orientation. A V1 neuron maximally responsive to a 180°-oriented bar will respond less to 150°- and 210°-oriented bars and even less than that to 120°- and 240°-oriented bars. It has always been implicitly assumed (and largely observed) that IT neural tuning would be similar to these bell-shaped curves of V1. An IT neuron sensitive to variations in a shape dimension would have a preferred value along this dimension (e.g., some particular height of the triangle which it liked the most), and any departures from this value would elicit a smaller response. Different neurons would have different preferred values, and the collection of preferred values would span the whole range of values that the dimension could take.

Unlike cells in V1, studies using parameterized shapes have suggested that IT neurons often respond maximally to those stimuli which attain the least or most extreme value of the dimension under investigation, such as the shortest or tallest triangle. Hence, there exist only two preferred values along the whole span of the dimension, and individual IT neurons are required to adopt one or the other extreme. This observation is inconsistent with bell-shaped tuning because the bell-shaped tuning account would predict a comparable number of neurons maximally responsive to the intermediate values. It is worth noting that the concept of 'most' extreme is operationally defined because dimensions can attain values higher than those tested in a particular experiment. IT neurons, however, cannot increase their response magnitude forever. Is there a way around this dilemma?

Two general solutions are possible. First, it could be the case that at the more extreme values, the neural response does start to decrease. This would support somewhat broad bell-shaped tuning functions. Second, at some point the higher-dimension value could

cause a sharp and distinct transition in the shape itself. To provide a concrete example, imagine a trapezoid whose two parallel sides are of very different length. As the lengths of these two parallel sides are made more and more equal, eventually this trapezoid will become a rectangle (if the endpoints of the parallel side are the same). It has been shown that an IT neuron selective for the trapezoid will maintain its response to the trapezoid over a rather wide range of the relative lengths of the two parallel sides but that the response will abruptly cease if the trapezoid is converted into a rectangle. This observation suggests that IT neurons show enhanced sensitivity to qualitative, as opposed to quantitative, differences in shapes. Figure 3 depicts how qualitative shape differences are able to modulate the responses of individual IT neurons more significantly than are equally large quantitative differences.

Prototype Shapes and Combination Encoding

A large but finite set of such qualitatively different 'prototype' shapes could be used to construct an almost infinite number of new shapes. Any new and arbitrary shape would simply be some unique combination of the already existing prototype shapes. In essence, the prototype shapes can be thought of as parts that are used to make a whole, complete object. It is plausible that IT cortex contains cells with sensitivities spanning a complete set of prototype shapes. Such a representational scheme demands that for any object to be accurately represented, it has to be decomposed into its individual parts.

Binding Problem

This decomposition raises the question of how the parts are eventually bound together to form a coherent percept of one complete visual object. When there is only one object in the environment, all the parts are assumed to belong to that one object (though their configuration is not specified). In most circumstances, however, biological vision systems have to deal with multiple objects simultaneously. Suppose there are two multipart objects present, a triangle on top of a square and a circle on top of a pentagon. A simple decomposition would result in several separate parts: triangle, square, circle, and pentagon. How does the visual system know to put the triangle with the square and not the pentagon or the circle?

Most current theories on how this binding process is carried out posit a substantial role for attention. Attention refers to enhanced sensory processing at those locations in the visual field that are attended to. How might attentional processes be related to binding? Prior to the allocation of attention to particular spatial locations, one is aware only of individual parts; at this point, misbinding is likely to occur at the unattended locations. However, subsequent to the deployment of attention, the parts at the attended locations get bound together and, as a result, complete object representations materialize. The specific mechanism through which attentional effects are manifested at the neural level remains a topic of debate. One interesting possibility is that attention synchronizes the neurons that are representing different parts of the same object. To be precise, each object is represented through an ensemble of neurons, and each ensemble contains neurons that are firing in concert. Synchronization prevents parts belonging to different objects from being falsely associated because these parts are embedded in different ensembles of neurons.

Spatial Configurations

Another implicit assumption in the prototype-based scheme is the ability of IT neurons to encode the spatial relationships between the individual parts. For example, it must be made explicit that the triangle goes on top of the circle, not off to the side or underneath it. A study that used a large battery of two-dimensional silhouette shapes showed that IT neurons can represent not only the particular parts comprising a shape but also the spatial arrangement of the object components. A separate experiment found that keeping the identities of the parts constant but varying their spatial arrangement causes a considerable decrease in the response of a subset of IT neurons. It seems then that the IT cortex is able to integrate simpler information from more posterior visual areas into something more closely resembling a complete object representation.

Dynamics of Shape Encoding

Despite the intuitive appeal and biological plausibility of prototype shapes and the electrophysiological evidence supporting such a scheme, the existence of a fixed set of structural primitives is incongruent with the phenomenon of visual experience. That is, prototype shapes are building blocks that can be arranged in any arbitrary fashion to describe something completely new and previously unseen. As such, the difference between the representation of novel and familiar shapes should be minimal as all the shapes are constructed in essentially the same manner. Yet we have all had the experience of encountering a new environment and being astounded by the novelty and complexity of the visual objects therein. Over time, as we become more and more familiar with our surroundings, we are able to more efficiently and accurately identify all

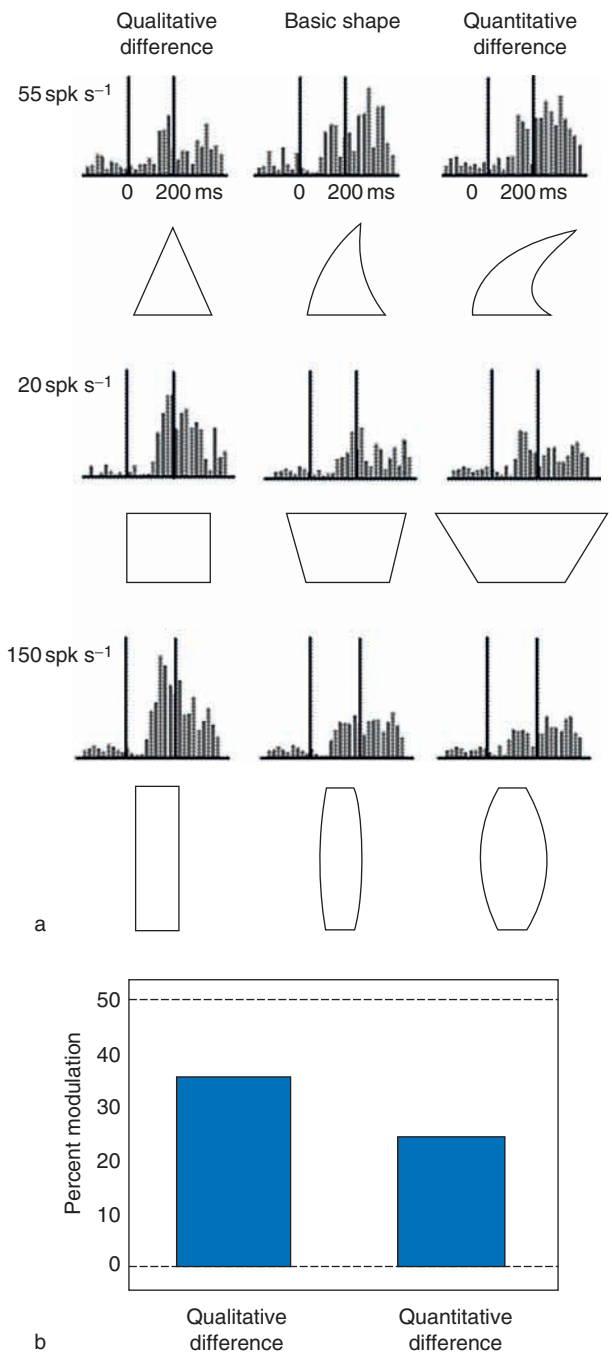

Figure 3 Qualitative vs. quantitative shape changes. (a) Responses of three example neurons to a basic shape (middle column), a qualitatively different shape (left column) and a quantitatively different shape (right column). The neuron in the top row maintained its response to the curved triangle as the curvature of the main axis was increased. However, when the triangle was made upright, the response was markedly reduced. For the neuron in the middle row, neither of the trapezoids elicited a reliable response, but the rectangle did. For the neuron in the bottom row, a change from a slightly oval rectangle to a more oval rectangle caused no significant change in the firing rate whereas a change to a rigid rectangle produced a substantial increase in the response. The changes between the basic shape and the qualitatively different shape were equal in magnitude (as measured through Euclidian distance) to the changes between the basic shape and the quantitatively different shape. (b) Across the whole population of neurons, qualitatively different shapes caused more modulation in the responses of inferotemporal neurons than did quantitatively different shapes. Spk, spikes. From Kayaert G, Biederman I, and Vogels R (2003) Shape tuning in macaque inferior temporal cortex. *Journal of Neuroscience* 23: 3016–3027.

the objects and shapes that are present in this setting. This is presumably the result of some changes that are taking place at the neural level. The concept of prototype shapes, therefore, needs to be complemented by a neural representational scheme that is better equipped to take into account visual experience.

Long-Term Plasticity

In contrast to primary visual cortex, plasticity in IT cortex appears to be prevalent throughout adulthood. Studies have shown that repeated visual experience may lead to an increased proportion of neurons responsive to the previously seen stimuli. More interesting, individual IT neurons show greater response differences to well-known stimuli. The existence of neurons that respond selectively to familiar objects presumably aids the identification and discrimination process. Recent observations suggest that changes caused by visual experience are also present in the local field potential, which is thought to reflect the integrated synaptic activity of thousands of neurons. These modifications alone, however, seem insufficient to underlie all the different forms of visual experience.

In real-world situations, different objects (especially of the same class) often share the same overall global shape. The proper representation of these shapes requires that the observer makes use of those local features that are informative while disregarding the irrelevant ones. Bird experts, for example, learn that in some situations the shape of the wings can identify a bird whereas in other situations the shape of the beak may be more critical. A plausible neural mechanism underlying this perceptual phenomenon might be for individual IT neurons to be sensitive to changes in the diagnostic features but insensitive to changes in the uninformative features. This hypothesis has in fact received supporting evidence. In a recent study, monkeys were trained to categorize stimuli as belonging to one of two arbitrary classes. Even though the stimuli had four different local features (e.g., dorsal fin, tail, ventral fin, mouth), only two local features (e.g., dorsal fin and tail) were informative with regard to performing the categorization. Physiological recordings indicated that only changes in the relevant feature dimensions modulated the responses

of individual IT neurons. Thus behavioral experience can bias to which aspects of a shape individual IT neurons devote their processing and selectivity.

For certain classes of shapes, individual local features are inadequate for correct identification. A classic example is that of faces; all faces have two eyes, a nose, a mouth, and so forth. While any of the features alone is enough to signal the presence of a face, it provides little information as to the identity of the face. To recognize a face requires a more holistic approach, a process whereby the spatial configurations of the features are given more weight. How does one go about exploring the possible neural mechanisms underlying holistic processing?

Monkeys can be trained to discriminate between objects composed of the same parts presented in unique combinations of parts (Figure 4). As in the classic XOR problem, this is achieved by creating two-part objects in such a way that the identity of any one part is ambiguous with respect to the identity of the object; both parts must be seen together in order to perform this task correctly. IT neurons demonstrate several changes in their tuning profiles following extensive experience with this task. Notable among these is nonlinear integration: neural responses to wholes that cannot be predicted from the responses to the individual parts. Such holistic responses allow for a more efficient representation of well-known, multipart stimuli.

The presence of neurons that develop such extraordinary selectivity for whole and complete shapes indicates the existence of a sparse representational code within the IT cortex, a code in which relatively few neurons are required to encode and represent an object. A well-known example used to illustrate this concept is the 'grandmother' cell. Does IT cortex have neurons that are selective for just one shape (such as a grandmother's face), and is the firing of this small subset of neurons sufficient to signal the presence of that shape? The idea does not seem far-fetched, considering the studies discussed. However, given the flexibility of the visual system to perceive and identify novel shapes, a sparse representational code has to be complemented by a more distributed and population-based code. A population code uses many neurons to encode and represent an object, each neuron contributing slightly different information about the same object. This sort of code is not only better equipped to handle the myriad of new objects experienced throughout adult life but is also able to identify well-known objects in unfamiliar contexts.

Short-Term Dynamics

It is interesting that the neural response reflecting the nonlinear integration of parts into wholes appears to take more time to develop than does the response to the individual parts alone. This suggests that the informational content of a single neuron's response varies as a function of time. Specifically, the early and late responses have the ability to indicate different facets of the same visual stimulus. Studies that have examined the temporal structure of the informational

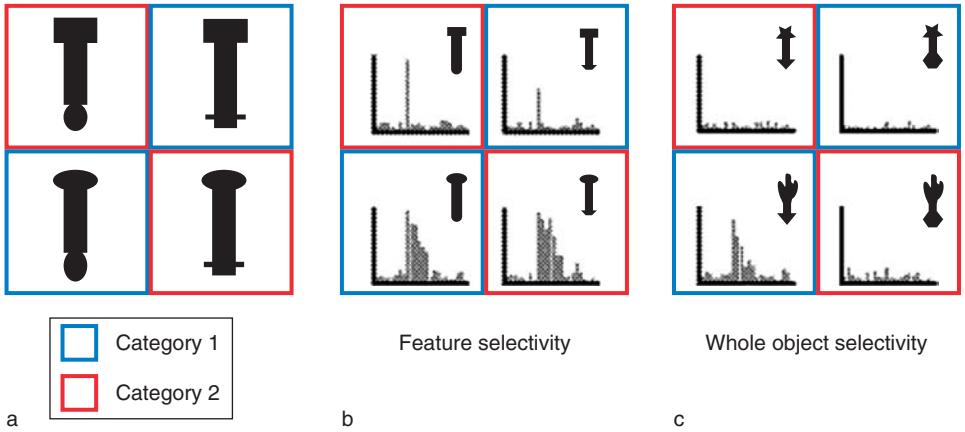

Figure 4 Parts- and whole-based shape coding. (a) Monkeys were trained to classify 'baton' stimuli as belonging to one of two classes. The stimuli from different categories shared the same parts, so the monkeys had to encode the upper and lower parts of the stimuli simultaneously to perform the classification correctly. (b) Compared to stimuli that were not learned, inferotemporal (IT) neurons showed enhanced selectivity for individual parts. (c) More interesting, some cells responded to learned stimuli as wholes. The preferred object here shared its parts with two other images; however, neither of the images had both parts together and, as such, did not elicit a response from the neuron. The response is thus 'nonlinear' because the activation to a whole object cannot be predicted on the basis of its response to the individual parts composing that object. Neuronal responses of this sort imply the presence of a sparse representational code within IT. From Baker CI, Behrmann M, and Olson CR (2002) Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience* 5: 1210–1216.

content have shown that a neuron's early response transmits coarser information about the stimulus, such as its category, whereas the late response carries information about the finer aspects of the stimulus, such as its identity. This is congruent with the intuitive notion that not all information about a visual stimulus is immediately available for conscious access. For instance, when presented with a face, we are much more likely to first identify it as a face, as opposed to a car or a house, than we are to make judgments about its identity or emotional state.

Noninvasive Imaging: Three-Dimensional Shape Representation, Human IT Homologs, and Future Directions

Recently, functional magnetic resonance imaging (fMRI) has been extensively used in studies exploring shape representation in the visual system, allowing for a more broadscale investigation of all the brain regions involved in a specific function. As such, it has been particularly useful in investigating the areas of cortex mediating the representation of complex, three-dimensional shapes. The importance of real-world three-dimensional shape representation extends beyond perception into cognitive domains. As expected, studies employing fMRI have shown that a distributed network of brain regions, including the occipital, temporal, parietal, and frontal lobes, mediates three-dimensional shape representation. These observations are consistent, not only with three-dimensional shapes playing an important role in perception, action, and cognitive functions, but also with three-dimensional shapes being defined through a variety of cues – similar to two-dimensional shapes. It is intriguing to note that within IT cortex itself, there seems to be a specialized area that responds only to three-dimensional shapes.

The use of fMRI in both monkeys and humans has also provided an opportunity to compare and contrast the brain regions important for shape representation in the two species. It has emerged from such comparisons that a likely human homolog of the monkey IT cortex is most likely the lateral occipital complex (LOC). This is a region in the human ventral occipitotemporal cortex that responds preferentially to images of intact objects. It shares with IT cortex the property of cue invariance: LOC will respond to shapes regardless of whether the shapes are luminance-, texture-, or motion-defined. Furthermore, LOC exhibits relative invariance to image transformations involving size and position.

This article has focused on what most would agree to be the critical role of the IT cortex: the representation of visual shape. Although it has examined several pieces of converging evidence that all point toward individual IT neurons being sensitive to some aspect of a shape, the representation of shapes is unlikely to be confined to a small, localized region of cortex. Rather, given the large role that the perception and recognition of shapes plays in daily life, the neural code underlying shape representation is likely to be distributed across several different interconnected areas of cortex, with IT cortex being an integral part of this circuit. To gain a more thorough understanding of this complex circuit, recent research has focused on delineating the interactions between the bottom-up hierarchical processes required for the initial representation of shapes and the top-down attentional and cognitive modulations that influence exactly how the sensory information is perceived and utilized. With the advent of fMRI, it has become possible to observe the functional connectivity between disparate areas of cortex during various perceptual and cognitive tasks. Furthermore, fMRI has enabled electrophysiologists to localize their recording sites and will also guide single-cell recordings made from different areas of cortex. Only recently have studies begun to record the simultaneous activity of multiple neurons from the same cortical area. This will potentially allow for a better understanding of the computations that populations of neurons withinan area carry out. While there is little doubt that IT cortex plays a critical role in shape representation, many details about how this area contributes to shape recognition in the real world remain to be worked out.

See also: Neural Coding of Spatial Representations; Perception of Surfaces and Forms; Vision: Surface Segmentation; Visual System: Multiple Visual Areas in Monkeys.

Further Reading

Baker CI, Behrmann M, and Olson CR (2002) Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience* 5: 1210–1216.

Brincat SL and Connor CE (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience* 7: 880–886.

De Baene W, Premereur E, and Vogels R (2007) Properties of shape tuning of macaque inferior temporal neurons examined using rapid serial visual presentation. *Journal of Neurophysiology* 97: 2900–2916.

Denys K, Vanduffel W, Fize D, et al. (2004) The processing of visual shape in the cerebral cortex of human and nonhuman primates: A functional magnetic resonance imaging study. *Journal of Neuroscience* 24: 551–2565.

Freedman DJ, Riesenhuber M, Poggio T, and Miller EK (2006) Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cerebral Cortex* 16: 1631–1644.

Kayaert G, Biederman I, and Vogels R (2003) Shape tuning in macaque inferior temporal cortex. *Journal of Neuroscience* 23: 3016–3027.

Kiani R, Esteky H, Mirpour K, and Tanaka K (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology* 97(6): 4296–4309.

Kourtzi Z and DiCarlo JJ (2006) Learning and neural plasticity in visual object recognition. *Current Opinion in Neurobiology* 16: 152–158.

Logothetis NK and Sheinberg DL (1996) Visual object recognition. *Annual Review of Neuroscience* 19: 577–621.

Op de Beeck H, Wagemans J, and Vogels R (2001) Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience* 4: 1244–1252.

Riesenhuber M and Poggio T (1999) Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2: 1019–1025.

Sary G, Vogels R, and Orban GA (1993) Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science* 260: 995–997.

Sigala N and Logothetis NK (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415: 318–320.

Sugase Y, Yamane S, Ueno S, and Kawano K (1999) Global and fine information coded by single neurons in the temporal visual cortex. *Nature* 400: 869–873.

Yamane Y, Tsunoda K, Matsumoto M, Phillips AN, and Tanifuji M (2006) Representation of the spatial relationship among object parts by neurons in macaque inferotemporal cortex. *Journal of Neuroscience* 96: 3147–3156.

Shift Work and Circadian Rhythms

T H Monk, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Biological Underpinnings

Homo sapiens is without doubt a diurnal creature, designed to sleep at night and to be up and about during daylight (except, perhaps, for a short nap after lunch). The endogenous circadian system facilitates this process by the generation and synchronization of various circadian rhythms which are appropriate to a diurnal (i.e., day-oriented) lifestyle. Thus, for example, higher values of endogenous melatonin in its circadian rhythm (which peaks at night) are associated with sleepiness in the diurnal human but activity in the nocturnal hamster. Therefore, having to work at night (as many shift workers do) is definitely an unnatural act for a human being. Humans, of course, do many things which are unnatural; some of them are not harmful at all. A key question (which this article seeks to address) is whether shift work is not only unnatural but also harmful. This article is divided into sections devoted to alertness and safety at work, sleep at home, social and domestic factors, and physical and mental health.

The Significance of Shift Work

In most industrialized countries, approximately one-fifth of all employees are engaged in some form of paid occupation that requires their presence at work outside of the 'daylight' 7 a.m. to 6 p.m. working day on a regular basis and who can thus be regarded as 'shift workers' using a broad definition of the term. This figure is expected to rise as second jobbing and mandatory overtime increase. The fastest growing sector of most Western economies is the service sector, and people are increasingly demanding and receiving around-the-clock availability of such services. Even in the production sector, plant machinery has become so expensive and so quickly obsolete that it has to be run 24 h per day, 7 days per week in order for it to be profitable. Also, the United States has adopted health care provision and business evaluation strategies (e.g., in assessing profitability) that encourage employers to squeeze as many work hours per year as possible from their existing employees, rather than hiring new ones, as the volume of business increases. This leads to extended work weeks and fewer different work teams covering each 24-h day.

Some people cope well with shift work, others poorly. At the extreme, one can define a 'shift work maladaptation syndrome' in those failing to cope, and thus having problems sleeping at home and being alert and productive at work. Several international classifications also refer to a sleep disorder associated with abnormal work hours. The International Classification of Sleep Disorders formally lists shift work sleep disorder as one of the circadian rhythm sleep disorders. The *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition, lists shift work as a subtype of the circadian rhythm sleep disorder (No. 307.45). Thus, there is increasing acceptance that the difficulties some people experience with shift work should properly be regarded as a disorder worthy of medical diagnosis and treatment.

In many countries, shift work is not simply restricted to a small group of people who can be carefully selected or self-selected to experience minimal problems. Neither is it restricted to the youngest and fittest of workers who can bid their way out of abnormal hours when they advance into their middle age. Employment trends, particularly in the manufacturing sector, now dictate that the 'bidding out of shift work by seniority' option often no longer applies. For many middle-aged and late middle-aged workers, the only option is between shift work and no work. Thus, people are 'doing shift work' typically for many years at a time, perhaps even almost until retirement. This is particularly important as the 'baby boom' generation reaches their late fifties because several studies have indicated that late middle age is a time of particular vulnerability to shift work coping problems.

What Does Night Work Involve?

As has been documented by several sleep diary and time budget studies, most night workers like to have their evenings free to be with family and friends and thus take their sleep immediately after work rather than after an interval of recreation. Night workers therefore have to phase shift (delay) their sleep by approximately 9 or 10 h from day-working society and from their own routine when on a diurnal ('day-oriented') schedule. The latter point is important because even permanent night workers seldom keep to a nocturnal schedule on their days off and are thus, *de facto*, rotating between nocturnal and diurnal schedules. This enables them to meet family and social commitments such as taking their children to soccer games on a Saturday morning or going to church on a Sunday. Monk and Wagner suggested that social and religious commitments on a Sunday could

be invoked to explain an observed Sunday night increase in industrial accidents (relative to other nights) in 10-year data from open cast mining operations. There is also empirical evidence for a reversion to a diurnal orientation after an off-duty break. In a study of permanent night nurses, it was shown that after an off-duty (weekend-type) break of at least 2 days, nurses' circadian temperature rhythms were identical in phase to those of day workers. Thus, those who work at night (even on fixed shifts) are continually challenging their circadian system with phase adjustments, perhaps for many years at a time.

What Do Evening Shifts Involve?

From the standpoint of sleep and circadian rhythms, evening shift work is relatively benign. Such workers typically finish work between 10 p.m. and 2 a.m. and can thus take most of their sleep during the night hours. Should social and domestic constraints allow it (which they often do not), the evening shift worker can sleep in late, and typically such workers on average obtain much more sleep than night workers and slightly more sleep than day workers. As discussed later, however, evening shifts are often devastating from a social and domestic standpoint, particularly for those with school-age children.

What Do Rotating Shifts Involve?

The term 'rotating shifts' describes a shift roster in which a given individual, over time, works all three shifts (day, evening, and night). The time that the individual spends on each of the given shifts is termed the 'speed of shift rotation.' Many shift workers rotate on a weekly or 2-weekly basis. A very common shift system in the United States is the 'southern swing' in which the individual works 1 week of nights, followed by 1 week of evening shifts, followed by 1 week of days. Parenthetically, most experts would agree that such a system is a disaster from a chronobiological standpoint. In Europe, a 'rapid rotation' of shifts is often seen, with all three shift timings done within a given week. In the 2-2-2 system, for example, individuals work 2 day shifts, 2 evening shifts, and 2 night shifts and then have 2 days off – all within one 8-day period.

As one might expect, no rotating shift system is easy to cope with. Many rotating shift workers report being in a state of permanent circadian desynchrony, in which they find themselves moving to a new timing just as they are starting to get used to the old one. Typically, most rotating shift workers sleep at night whenever they possibly can, thus further complicating the phase adjustment picture.

Consequences of Shift Work

Alertness and Safety at Work

One major function of the human circadian system is to 'shut down' various processes so that a restful 7 or 8 h of uninterrupted sleep can be obtained. *H. sapiens* is a social and inquisitive creature, and thus in addition to shutting down hunger, thirst, and the need to void during the night hours, the circadian system also renders the individual sleepy and inattentive during these night hours. The trough in these measures typically occurs at or around 4 a.m., leading to night shift workers often describing a '4 a.m. slump.' This slump is beneficial for the day worker who is asleep at night, but it is potentially harmful for the night worker who must perform his or her job during the night hours, often with complex critical processes and/or dangerous machinery.

Many studies have shown that there is are circadian rhythms in alertness and performance which are endogenously generated by the circadian pacemaker, and are of nontrivial magnitude. These performance effects are equivalent in magnitude to that seen when the 'legal limit' of alcohol is ingested. The human circadian system typically requires 1 day for each 90 min of phase delay adjustment, and thus for at least the first several nights of a run of night duty, individuals will be performing at work during the down phase of the circadian performance rhythms with resultant impairments in alertness, productivity, and safety.

In addition to night shift alertness and performance suffering from being in the down phase of the circadian performance cycle, there are also the effects of partial sleep loss. If the circadian system is at a diurnal (day-oriented) circadian phase, then it will not be doing its job of preparing the individual for a restful sleep during the daylight hours at home. Thus, the day sleep of the night worker is likely to be, from biological processes alone, shorter and less refreshing than the night sleep of the day worker. However, as noted later, to this must be added environmental and psychosocial factors such as daytime noise and domestic responsibilities which will further worsen the night worker's day sleep .

There are several things that a tired night shift worker might do, all of which could potentially compromise safety. First is the obvious one of falling asleep. This is most prevalent in the long-haul trucking industry, where every year several thousand lives are lost by truckers falling asleep and running off roads. However, this has also happened in the airline industry, where the entire cockpit crew was asleep as one airplane full of passengers failed to make its

approach to a west coast US city and headed out over the Pacific Ocean, and in the nuclear power industry, where regulators made a surprise visit to Peach Bottom Nuclear Power Plant in Pennsylvania and found everyone in the control room asleep. Safety can also be compromised by night workers making a simple error in response, even though they are not asleep. The classic example of this is the 4 a.m. accident at Three Mile Island, where the operators were very much awake but were not processing information appropriately. A nocturnal increase in the errors of reading industrial gas meters was also reported in a classic study in the field. A third way in which safety can be compromised is by the night shift worker being angry. This is comparatively neglected as an area of scientific enquiry but undoubtedly contributes to poor safety on the night shift given that affect dysregulation is a major consequence of sleep loss and circadian desynchrony, and that the social and domestic lives of shift workers are often in a state of turmoil.

As one review of shift work and safety notes, acquiring good quality data to support the hypothesis that safety on the night shift is impaired is not easy. Often, there are very different numbers of people present at the facility at different times of the day and night, and the work is often of a different nature (e.g., saving long production runs for the night shift and saving new additions to the process for the day shift). Thus, a simple comparison of accident rates on the night and day shifts will often fail to reveal a night shift increase. One study obtained accident data from a UK automobile manufacturing plant where conditions were exactly the same over day, evening, and night shifts. An analysis of 4645 injuries revealed that the relative risk of an accident was approximately 20% higher (relative risk, 1.2) on the night shift compared to the day shift. Importantly, when only serious injuries occurring away from the production line were considered, the relative risk for the night shift increased to 1.8.

Sleep at Home

Sleep is the major preoccupation of most night shift workers. In both Europe and the United States, surveys have indicated that night workers get approximately 5–8 h less sleep per week than their day-working counterparts. There are both biological and psychosocial reasons for this decrement. To the shift worker, the most salient disruptions are the psychosocial ones of traffic noise, household noise, and family commitments. The day sleep of the night worker is not protected by society's taboos as is the night sleep of the day worker. Few people would make a nonurgent telephone call to a day worker at 1 a.m., whereas few would think twice about making a 1 p.m. call to a day-sleeping night worker. Also, as discussed later, the family will often make demands on the shift worker (particularly if she is a mother of young children) which are unreasonable and which will severely limit her sleep.

To these psychosocial disruptions must be added the biological ones. As noted previously, the biological reasons for day sleep disruption stem from the fact that *H. sapiens* is a diurnal species, hard-wired to sleep at night rather than during the day. The circadian system is slow to make the approximately 9-h phase delay adjustment required by night work, typically taking at least 6 days to make the full 9-h change, and often failing to achieve a full nocturnal orientation. Unlike the transmeridianal traveler, the night shift worker experiences photic cues stemming from daylight and darkness which are resolutely opposed to any circadian adjustment. Studies have shown that the commute home in daylight following a night shift can be particularly problematic since it involves bright light exposure at the point in the phase response curve which inhibits the circadian phase delay that the night shift worker is trying to obtain.

Because the circadian system is not anticipating a day sleep, it thus fails to do its job of shutting down processes so that a restful 7 or 8 h of sleep during the day can be obtained. Thus, even when traffic noise, domestic noise, and family commitments are removed from the picture, the day sleep of night workers will be disrupted. This was demonstrated clearly in a study in which shift workers were removed from their family and domestic constraints and placed in a sound-attenuated, electrically shielded bedroom for their sleep periods, with the subjects commuting to their work from the laboratory rather than from home. Even in this closely protected environment, there was a highly significant difference between the duration of the day sleep of night workers and the night sleep of day workers (5.1 vs. 6.7 h). In addition, there were reliable differences between the polysomnographic characteristics of the sleep, with a smaller amount of rapid eye movement (REM) sleep and a greater proportion of slow wave sleep for the night workers.

In Akerstedt's general surveys of field and laboratory shift work sleep studies, it was concluded that the shortening in a night worker's day sleep comes primarily from a reduction in stage 2 and REM sleep, with slow wave sleep relatively unaffected. Not surprisingly, given the prolonged levels of partial sleep deprivation involved, sleep latency can be somewhat reduced in night workers, and some studies have found shorter REM latencies to occur. Essentially,

the problem is usually one of sleep maintenance insomnia rather than sleep onset insomnia. A shift worker's sleep loss is sometimes partially recouped on days off but does represent a chronic state of partial sleep deprivation which undoubtedly affects the mood and performance abilities of the individual. Laboratory studies have shown that 1 week of sleeping only 5 h per 24 h (as is typical of some individuals performing night work) shows significant cumulative effects on subjective alertness and objective performance abilities. There are several well-controlled studies which document the pathological sleepiness levels exhibited by many shift workers, both at work and on the drive home after work.

Social and Domestic Interactions

Human beings are essentially social creatures, and one could argue that the social and domestic factors are at least as important in shift work as the biological ones. Certainly, if a shift worker's domestic and social life is unsatisfactory, then the individual will not be coping well, however well adjusted the sleep and circadian rhythm factors may be. Usually, however, poor domestic adjustment adversely affects the ability of the shift worker to do the 'right thing' with regard to sleep and circadian rhythms. A common example concerns the child care and household management tasks that can be expected of a female shift worker. Unlike her male colleagues, she is often expected by her spouse to continue to run the household and can thus find herself completely unable to comply with the routine that good sleep hygiene and circadian adjustment might require.

Another aspect of domestic disruption concerns the role of the male shift worker as husband and parent. With regard to the former, there are three major spouse roles that are affected: sexual partner, social companion, and protector–caregiver. All three roles are compromised. Perhaps as a consequence, shift work has been shown in a longitudinal follow-up study to increase the risk of divorce by 57%. Although some of the problems are sexual in origin, many arise from the spouse's inability to be there when needed. Feelings of loneliness and insecurity in a wife left alone at home every night, for example, can represent a much more chronic and insoluble problem than that associated with the timing of lovemaking. Similarly, the evening shift, which has minimal impact on the sleep and circadian factors, can have a crushing impact on the role of the shift worker as social companion. With regard to the family role of parent, the evening shift is again the most disruptive. Often during the school week, the shift worker may only get to see his or her children when they are asleep in bed. In addition, both spouse and parent roles are heavily disrupted when the shift worker is required to work on weekends.

In addition to disrupted family roles, the shift worker often suffers from social isolation from day-working friends and from religious and community organizations that work under the expectation that evenings or weekends will be free for meetings and activities. One might advance the view that perhaps a shift worker who is denied access to community meetings and social and political associations is as much disadvantaged as a handicapped person who is denied wheelchair access to a museum.

Health

A number of reviews have documented the health consequences of shift work. Apart from the obvious problems related to sleep and alertness, there appear to be both physical health and mental health consequences of shift work exposure. In terms of physical health, the main problems implicated are gastrointestinal disorders, cardiovascular disorders, and cancer. All are nontrivial and cannot be attributed simply to lifestyle factors attributed to shift work, such as diet and smoking. For example, in one study, having more than 10 years of shift work exposure doubled the risk of cardiovascular disease. Other studies have confirmed this relation between shift work and cardiovascular disease, even when other risk factors are controlled for. In a prospective study of almost 80 000 nurses, extensive (>30 years) rotating night work exposure increased the risk of breast cancer by 36%. Again, this increase has been confirmed in other studies. In terms of mental health, neuroticism, substance abuse, and alcohol abuse increases have all been implicated in rotating shift workers. Also, a telephone survey of 98 current and former shift workers suggested a more than fivefold increase in the lifetime prevalence of formally diagnosed major depression in those with 6–20 years of shift work exposure compared to those with less than 5 years of exposure.

Clearly, any consideration of the effects of chronic shift work must also be concerned with the weight (or, rather, the body mass index (BMI)) of the worker. From work at the University of Chicago and elsewhere, it is now well established that partial sleep deprivation affects glucose tolerance, and it would thus not be surprising were such effects, compounded with the poor dietary choices that shift workers often make (or have to make), to result in increased BMI and an increased likelihood of sleep apnea and type II diabetes as years of shift work exposure increase. Supportive evidence for this comes from a study

from The Netherlands which concludes that BMI increases with shift work experience, a study from Sweden which concludes that there is an association between shift work and the metabolic syndrome, and a study from Japan that found that shift work is associated with insulin resistance in shift workers younger than 50 years of age.

Conclusions

Shift work is a multifaceted challenge resulting not only from our biology as a diurnally oriented species but also from the social structures which we have built around us. These structures make assumptions regarding sleep and recreation times that can be unhelpful for the shift worker. Thus, shift workers are likely to have problems concerning alertness and safety at work, sleep at home, social and domestic interactions, and the maintenance of good health. It is unrealistic to assume that the prevalence of shift work will do anything other than increase, and it thus behooves the scientific community to provide information to employees, employers, and governments so that the negative effects of shift work can be ameliorated to the greatest extent possible.

See also: Circadian Function and Therapeutic Potential of Melatonin in Humans; Circadian Rhythms: Influence of Light in Humans; Circadian Rhythms in Sleepiness, Alertness, and Performance; Endocrine Function During Sleep and Sleep Deprivation; Genetics of Circadian Disorders in Humans; Immune Function During Sleep and Sleep Deprivation; Psychiatric Disorders Associated with Disturbed Sleep and Circadian Rhythms; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function; Sleep and Circadian Rhythm Disorders in Human Aging and Dementia; Sleep: Development and Circadian Control.

Further Reading

Akerstedt T (1985) Adjustment of physiological circadian rhythms and the sleep–wake cycle to shiftwork. In: Folkard S and Monk TH (eds.) *Hours of Work: Temporal Factors in Work Scheduling*, pp. 185–197. New York: Wiley.

Akerstedt T (2003) Shift work and disturbed sleep/wakefulness. *Occupational Medicine (London)* 53: 89–94.

Belenky G, Wesensten NJ, Thorne DR, *et al.* (2003) Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose–response study. *Journal of Sleep Research* 12: 1–12.

Carrier J and Monk TH (2000) Circadian rhythms of performance: New trends. *Chronobiology International* 17: 719–732.

Costa G (2003) Shift work and occupational medicine: An overview. *Occupational Medicine (London)* 53: 83–88.

Dawson D and Reid K (1997) Fatigue, alcohol and performance impairment [letter]. *Nature* 388: 235.

Drake CL, Roehrs T, Richardson G, *et al.* (2004) Shift work sleep disorder: Prevalence and consequences beyond that of symptomatic day workers. *Sleep* 27: 1453–1462.

Eastman CI, Stewart KT, Mahoney MP, *et al.* (1994) Dark goggles and bright light improve circadian rhythm adaptation to night-shift work. *Sleep* 17: 535–543.

Folkard S and Tucker P (2003) Shift work, safety and productivity. *Occupational Medicine (London)* 53: 95–101.

Knauth P and Hornberger S (2003) Preventive and compensatory measures for shift workers. *Occupational Medicine (London)* 53: 109–116.

Knutsson A (2003) Health disorders of shift workers. Occuptional Medicine (London) 53: 103-108.

Megdal SP, Kroenke CH, Laden F, *et al.* (2005) Night work and breast cancer risk: A systematic review and meta-analysis. *European Journal of Cancer* 41: 2023–2032.

Monk TH (2000) What can the chronobiologist do to help the shift worker? *Journal of Biological Rhythms* 15: 86–94.

Office of Technology Assessment (1991) *U.S. Congress, Office of Technology Assessment, Biological Rhythms: Implications for the Worker.* (OTA-BA-463). Washington, DC: U.S. Government Printing Office.

Sack RL, Blood ML, and Lewy AJ (1992) Melatonin rhythms in night shift workers. *Sleep* 15: 434–441.

Scott AJ (2000) Shift work and health. *Primary Care* 27: 1057–1079.

Scott AJ, Monk TH, and Brink L (1997) Shiftwork as a risk factor for depression: A pilot study. *International Journal of Occupational and Environmental Health* 3: S2–S9.

Smith L, Folkard S, and Poole CJM (1994) Increased injuries on night shift. *Lancet* 344: 1137–1139.

Turek FW (1986) Circadian principles and design of rotating shift work schedules. *American Journal of Physiology* 251: R636–R638.

White L and Keith B (1990) The effect of shift work on the quality and stability of marital relations. *Journal of Marriage and the Family* 52: 453–462.

Short Term and Working Memory

B R Postle, University of Wisconsin – Madison, Madison, WI, USA
T Pasternak, University of Rochester, Rochester, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Definition

Short-term memory (STM) refers to the active retention (for humans, the 'keeping in mind') of information when it is not accessible from the environment. Working memory can be thought of as 'STM+,' the '+' referring to the ability to manipulate or otherwise transform this information, to protect it in the face of interference, and to use it in the service of such high-level behaviors as planning, reasoning, and problem solving. STM and working memory are of central importance to the study of high-level cognition because they are believed to be critical contributors to such essential cognitive functions and properties as language comprehension, learning, planning, reasoning, and general fluid intelligence.

Historical Backdrop

The modern study of working memory began with the work of Jacobsen, who, in the 1930s, demonstrated that large bilateral lesions of the prefrontal cortex (PFC) in the monkey produced profound deficits in spatial working memory, and thus Jacobsen ascribed to the PFC a function he termed "immediate memory." A subsequent study, however, revealed that this deficit could be erased by darkening the testing cage and thus reducing interference during the memory delay. These results highlighted a critical question, still the subject of an active debate, which is whether PFC plays a critical role in storage in addition to its role in behavioral control.

The early 1970s witnessed two developments that were seminal in shaping contemporary conceptions of working memory. The first was the observation made by Joaquin Fuster that individual neurons in PFC of monkeys, trained to compare visual stimuli separated by a memory delay, exhibit sustained activity throughout that delay. This observation suggested a neural correlate of two potent ideas from physiological psychology – that of a PFC-dependent immediate memory and that of a reverberatory mechanism for "a transient 'memory'" proposed by psychologist Donald Hebb. The second development, which occurred in the field of human cognitive psychology, was the multiple component model of working memory proposed by Alan Baddeley and Graham Hitch. In its initial instantiation this model comprised two independent buffers for the storage of verbal and of visuospatial information, and a central executive to control attention and to manage information in the buffers. Prompted by these two developments, the neuroscientific and the psychological study of working memory each proceeded along parallel, but largely independent, paths until the late 1980s, when a third important advance occurred.

The third advance was the conceptual integration of the neuroscientific and psychological traditions of working memory research, proposed by Patricia Goldman-Rakic, that the sustained delay-period neural activity in PFC and the storage buffers of the multiple-component model of Baddeley and Hitch were cross-species manifestations of the same fundamental mental phenomenon. This association between prefrontal cortex and working memory has been very influential in systems and cognitive neuroscience.

Current State of Working Memory Research

A growing body of evidence provided by behavioral, physiological, and neuroimaging studies indicates that information about sensory stimuli may be stored in a segregated, feature-selective manner, and that the relevant cortical regions include relatively early stages of sensory cortical processing. The principle emerging from this work is that the same brain regions that are responsible for the precise sensory encoding of information also contribute to its short-term retention. In the remaining portions of this article we briefly describe these recent advances, with the emphasis on working memory for fundamental dimensions of sensory stimuli.

Visual Working Memory

In the laboratory, the ability to briefly retain visual information can be measured with delayed discrimination tasks. In the simplest version of such tasks, individuals discriminate between two stimuli, the sample and the test, separated by a temporal delay of various durations, and report whether and how the test differs from the previously seen sample. The results of such experiments have revealed that fundamental stimulus features, such as orientation, contrast, size, or speed, can be faithfully preserved for many seconds, although the duration of this preservation often differs for different features. For example, stimulus size or orientation can be retained accurately longer than can luminance contrast. This

difference suggests that different stimulus attributes may be retained by separate, feature-selective mechanisms. More direct support for feature-selective storage mechanisms comes from studies that have used interference consisting of an irrelevant stimulus (a 'memory mask') introduced during the delay separating S1 and S2. With this procedure, determination of the parameters that maximize the interfering effects of the mask can provide insights into the nature of the remembered stimulus and thus into the mechanisms involved in its short-term retention. For example, the memory for spatial frequency of gratings (size) can be disrupted only if the masking stimulus is of a different spatial frequency, irrespective of its orientation. This selective interference suggests that information about spatial frequencies may be preserved by mechanisms that are narrowly tuned for that stimulus attribute, and closely associated with its processing, but distinct from mechanisms concerned with stimulus orientation. Similar specialization is seen within the domain of visual motion, in that stimulus speed can be preserved by the mechanisms that are relatively narrowly tuned for speed, independently of stimulus direction.

Representation of visual motion in memory appears to be localized in space in a manner consistent with properties of neurons in cortical area MT, a region specialized in processing of visual motion, thereby supporting the idea that neurons processing visual motion may also be involved in its short-term storage. Overall, behavioral studies of mechanisms that preserve basic attributes of visual stimuli can be characterized as narrowly tuned, spatially localized filters, supporting a model of working memory that involves the contribution of sensory cortical areas.

A close link between visual processing and storage is also suggested by the consequences of cortical damage. For example, damage to motion processing area MT in monkeys results in deficits in remembering motion direction, but only when the remembered stimuli require area MT for their encoding (**Figure 1**). Similarly, damage to inferotemporal cortex, associated with processing of complex shapes, results in deficits in remembering such shapes. Humans with focal lesions in the occipitotemporal cortical area, implicated in processing of specific visual features, also have difficulties remembering these features.

The activity of visual cortical neurons during working memory tasks also attests to their role in sensory maintenance. Such tasks similar to the delayed discrimination tasks described earlier, require the comparison of two stimuli separated by a memory delay. When such stimuli consist of moving random-dot stimuli and the monkeys are required to identify and remember the dot direction, the activity of MT neurons during the memory delay shows selectivity reflecting the remembered direction. This activity is pronounced early in memory delay but weakens toward its end and does not correlate with the animals' decision, suggesting that additional neural mechanisms are needed to account for the maintenance of motion signals. However, at the time of the comparison and decision, these neurons are strongly affected by the remembered direction, suggesting that throughout the task they have access to the remembered direction, and thus are likely to participate in the circuitry subserving storage. Similar behavior has been observed in inferotemporal (IT) cortex, associated with the processing of complex shapes, in that stimulus-selective activity occurs not only in response to specific shapes but also during the memory delay. During the comparison phase of the task, these neurons, like neurons in area MT, reflect the remembered shape. This type of activity is consistent with the possibility that visual cortical neurons actively participate the circuitry subserving storage of signals they process, although the nature of this participation awaits further study.

Functional imaging studies of visual working memory indicate that regions of the visual system are differentially recruited by working memory tasks, depending on the stimulus dimension that must be remembered. For example, working memory for the location versus the identity of visual stimuli, such as color patches and geometric shapes, recruits domain-specific memory-related activity in posterior cortical regions associated with the 'where' and 'what' visual streams, respectively. For location memory this includes dorsal occipital cortex, the intraparietal sulcus, and superior parietal lobule, whereas for object identity it includes ventral temporal and occipital cortex. Location memory tasks also typically recruit frontal regions associated with oculumotor control – the frontal eye fields and the supplementary eye fields.

Within the ventral streams, delay-period activity is segregated in a category-specific manner. For example, multiple images of faces and of naturalistic scenes can be presented as samples, with a postsample cue indicating which is the to-be-remembered category. When individuals are cued to remember the faces, delay-period activity is selectively elevated in the fusiform face area (FFA). When they are cued to remember scenes, delay-period activity is selectively elevated in the parahippocampal place area (PPA). Another study of working memory for faces featured three delay periods interposed between the presentation of the first and second, second and third, and third and fourth stimuli. The logic was that the multiple distracting events in this task might serve to 'weed out' from the first delay period activity that was not involved directly in storage, because only regions

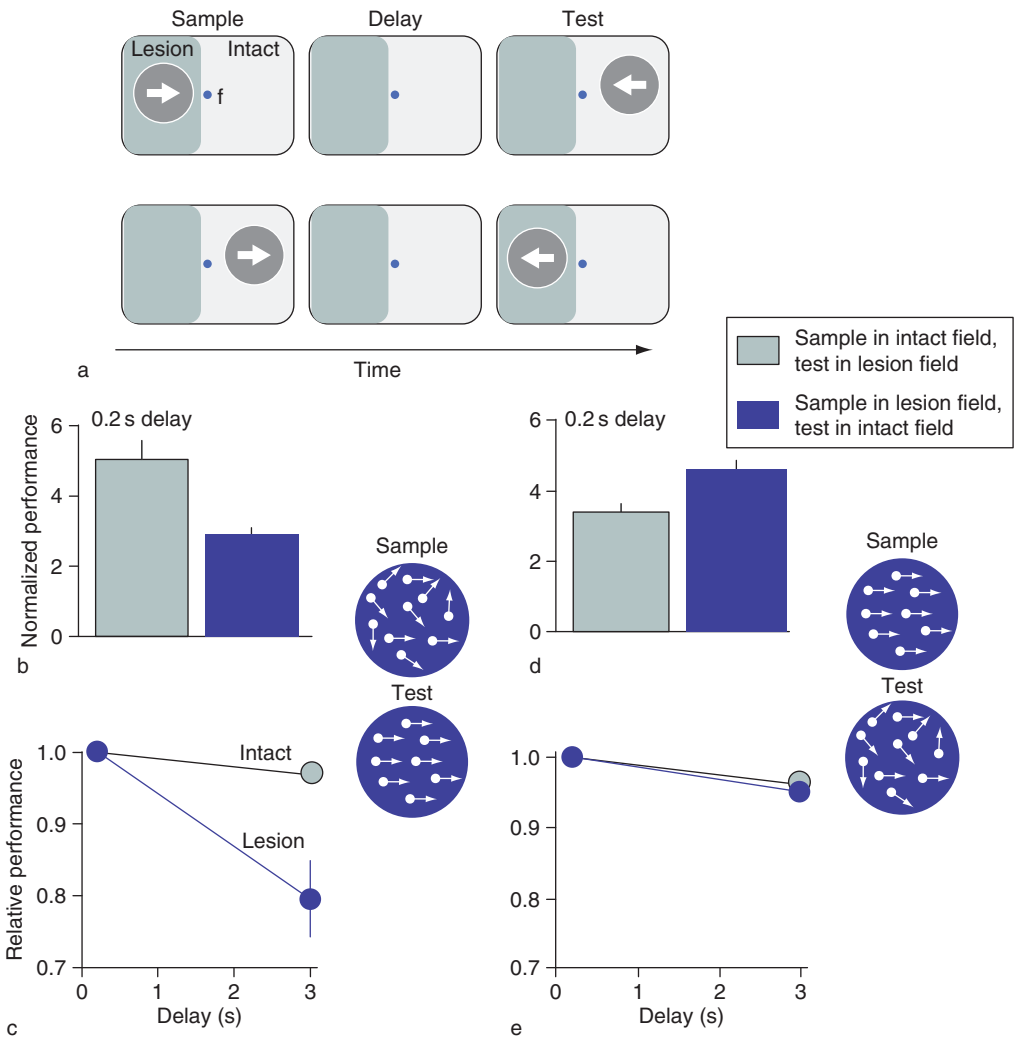

Figure 1 The role of cortical areas processing visual motion in remembering motion direction. (a) Monkeys with unilateral lesions of motion processing areas MT/MST performed a task requiring integration of complex motion and remembering its direction. In this task, sample and test stimuli were separated by a delay and positioned in opposite hemifields, so that one was placed in the lesioned and the other in the corresponding location in the intact hemifield. The monkeys reported whether sample and test moved in the same or in different directions by pressing one of two response buttons. During each trial the monkeys fixated a small target at the center of the display (blue dot) while attending to moving random dots (indicated by arrows) presented in the periphery. (b, d) Normalized direction range thresholds (1/(360-range threshold)/360) measured when the delay between sample and test was minimal (0.2 s). The sample or the test stimulus was composed of dots moving in a range of directions, while the other stimulus contained only coherent motion (stimulus configurations shown to the right of each set of plots). Performance was decreased whenever the stimulus containing noncoherent motion, and thus requiring integration, was placed in the lesioned field, demonstrating the importance of areas MT/MST for motion integration. (c, e) Effect of memory delay on performance for two direction range tasks. Performance was measured with both stimuli placed in the intact (gray circles) or in the lesioned hemifields (blue circles), either by varying the range of directions in the sample, while the test moved coherently (left plots), or by varying the range of directions in the test, while the sample moved coherently (right plots). Thresholds were normalized to the data measured at a 0.2 s delay. Error bars are SEM. A delay-specific deficit was present only when the remembered stimulus (sample) contained a broad range of directions and required integration. This result demonstrates that stimulus conditions requiring motion integration depend on intact areas MT/MST. However, coherently moving random dots can be discriminated and remembered at a normal level even in the absence of areas MT/MST. This shows that a cortical area involved in processing of sensory signals is also involved in their storage. Adapted from Bisley JW and Pasternak T (2000) The multiple roles of visual cortical areas MT/MST in remembering the direction of visual motion. *Cerebral Cortex* 10: 1053–1065.

with activity necessary for retaining the information to the end of the trial would maintain their activity across distracting stimuli. The results revealed that the posterior fusiform gyrus was the only region that retained the relevant signal during the last delay, immediately preceding the decision. Still other studies of working memory for faces have varied memory load (i.e., the number of items to be remembered),

on the logic that storage-related activity is sensitive to variations in load. These studies confirmed the importance of posterior fusiform gyrus in short-term storage.

Studies of working memory for spatial location have also addressed the mechanisms that support storage. One of these is attention-based rehearsal, a mechanism hypothesized to contribute to the short-term retention of locations via covert shifts of attention to the to-be-remembered location. These studies presented sample stimuli in one or the other visual hemifield while individuals fixated a central spot. Delay-period activity was greater in the hemisphere contralateral to the target location, an effect comparable to what one sees in studies of attention. This lateralized delay-period bias was strongest in extrastriate regions, decreased in magnitude across the parietal cortex, and was no longer reliable in frontal cortex. A second mechanism for the short-term retention of location information is prospective motor coding – the formulation, and then retention, of a motor plan for the acquisition of a target with, say, a saccade or a grasp. Electrophysiology studies in monkeys and neuroimaging studies in humans that encourage a prospective strategy localize this activity to frontal oculomotor regions, to prefrontal cortex, and to the caudate nucleus.

Tactile Working Memory

As with visual stimuli, tactile stimuli can be faithfully represented in working memory. Delayed discrimination of vibration stimuli can be performed with delays of many seconds, although the accuracy of this discrimination is maximal at short delays and decreases rapidly during the first 5 s of the delay. At longer delays, however, performance does not continue to deteriorate, suggesting that a two-stage memory process might be involved. Delayed discrimination of vibration stimuli can be disrupted by the application of transcranial magnetic stimulation (TMS) to human primary somatosensory cortex (area S1) during the initial portion of the delay period. This effect not only implicates area S1 in storage of vibration information, but also demonstrates the vulnerability of the storage mechanism early in the delay, an effect also observed in studies of working memory for visual motion.

In the awake behaving monkey, electrophysiological evidence for delay-period activity in S1 is mixed, with some laboratories reporting evidence in favor of it and others failing to find it. In area S2, however, there is no such equivocality, with strong evidence for stimulus frequency-specific activity persisting far into the delay period. Furthermore, responses of S2 neurons to the test stimulus contain information about the remembered stimulus, thereby reflecting the relationship between the two stimuli.

The ability to recognize and remember objects based on tactile input has also been studied with human neuroimaging. For example, positron emission tomography (PET) studies have revealed activation in the parietal operculum (area S2), an associative somatosensory area, during working memory tasks involving vibratory stimuli or palpated wire forms.

Auditory Working Memory

Although humans discriminate both pitch and loudness of sounds with high accuracy and remember these dimensions for many seconds, memory for pitch and loudness decline at different rates, suggesting that the two dimensions may be processed separately in auditory memory. The observed differences in the precision of memory for intensity and pitch parallel the findings in vision for contrast and spatial frequency (see earlier). As with vision, the use of interfering stimuli during the memory delay has revealed the nature of representation of auditory stimuli in memory. For example, memory for pitch can be disturbed by distractor tones, but only if these tones are within a narrow range of frequencies, relative to the frequency of the remembered stimulus, supporting the existence of separate pitch memory modules. Furthermore, sound frequency is likely to be stored separately from its location, suggesting that auditory memory obeys the same patterns as does auditory perception for physical parameters of the remembered stimuli.

In the auditory cortex, neuronal activity during the delay period of a delayed two-tone discrimination carries information about the frequency of the sample tone. In addition, the response to the second tone depends on whether its frequency matches that of the first, implicating neurons in auditory cortex in the circuitry subserving working memory for tone frequency. Participation of auditory cortex in auditory working memory is also supported by studies of the human using magnetoencephalography (MEG).

Memory-related activity has also been seen in associative auditory cortex of humans during tasks requiring memory for differences in pitch and tonal sequences; the right superior temporal lobe was associated with tonal sequences, and pitch judgments were associated with increased right frontal lobe activation. Interestingly, the auditory cortex appears to be maximally active early in the delay, whereas regions in the supramarginal gyrus and parts of the cerebellum are activated later, suggesting that for audition, too, working memory may be accomplished in multiple stages.

The involvement of regions processing auditory information in working memory is also supported by selective deficits in patients with damage to auditory associative cortex in the right temporal lobe. These patients are impaired primarily on tone discrimination with long, distracted delays.

Multiple Encoding in Working Memory

It seems unlikely that the representational bases of working memory are limited to the domain in which stimuli are perceived. In addition to the multiple serially ordered stages in working memory for visual, tactile, and auditory domains, sensory information can also be represented in parallel in multiple formats. Thus, as reviewed earlier, working memory for location can be accomplished via retrospective sensory memory-based mechanism, as well as a prospective motor planning mechanism. Analogously, there is growing evidence that working memory for visual identity recruits verbal (and, perhaps, semantic) representations in addition to the ventral stream visual mechanisms already discussed.

Working Memory and Prefrontal Cortex

For many years PFC has been strongly associated with the storage of sensory information in working memory. Consistent with this idea are the facts that PFC is directly interconnected with cortical areas processing fundamental sensory dimensions, and that PFC neurons often display stimulus-related activity during the memory delay. However, a growing body of evidence suggests that PFC is more likely to play a key role in directing attention to behaviorally relevant sensory signals and in making decisions concerning these signals, than in directly supporting their retention in short-term and working memory. Consistent with this view is the fact that extensive damage to prefrontal cortex does not eliminate the ability to perform simple short-term memory tasks, such as the digit span and spatial span variants of immediate serial recall. The limited effect of such lesions is apparent on delay tasks if during the memory period there is no interference and no requirement to keep track of multiple items. Also consistent with this view is the finding that although the delay activity of PFC neurons recorded in monkeys performing a delayed direction discrimination task (see earlier) was direction-specific, it did not predict decisions about stimulus direction that they made at the end of the trial. The contribution of these neurons to the task became apparent only after the end of the memory delay, when the monkeys compared test direction with that of the remembered sample. Furthermore, direction selectivity in PFC neurons, prevalent during the task, was greatly reduced when the monkeys were not required to use the information about stimulus direction. This work showed that PFC neurons are capable of gating sensory signals according to their behavioral relevance, it but does not support the key role for PFC neurons in the maintenance of these signals. Other studies, designed to isolate the contributions of attention and response selection from those of sensory storage, indicate a stronger role for the former in PFC.

In the human neuroimaging literature there is some debate about stimulus domain- or category-specific segregation of delay-period activity in PFC – for example, comparable patterns of delay-period activity during tasks involving memory for location versus memory for objects or memory for faces versus memory for scenes. The same is true for tasks that involve haptic object encoding versus those that involve visual object encoding. Similarly, in tasks requiring cross-modal integration, individual PFC neurons can be active during the memory delay following the presentation of different stimulus modalities. These examples suggest that PFC assists early-level sensory processing regions to form supramodal mental representations of objects and/or of task contingencies when the relevant stimuli belong to more than one sensory domain.

Another strategy for evaluating the neural basis of sensory storage is to require the retention of information across multiple delay periods interrupted with intervening distracting stimuli. This approach has revealed that neurons in the monkey PFC, but not IT cortex, could sustain a representation of the sample stimulus across multiple delays. Robust sample-specific activity across multiple delay periods is not limited to PFC, however, and has also been observed in other regions, including the temporal pole and the entorhinal cortex of the medial temporal lobe. Functional magnetic resonance imaging (fMRI) studies in humans remembering faces showed that only posterior fusiform gyrus retains sustained activity across all three delay periods, although many regions displayed activity during the first delay period. An analogous result was seen in the individuals remembering spatial locations, with distraction-spanning delay-period activity distributed across multiple areas, including intraparietal sulcus, superior parietal lobule, frontal eye fields, and supplementary eye fields in frontal cortex (Figure 2). Finally, repetitive TMS (rTMS) applied to PFC has failed to disrupt working memory performance, but has altered performance when applied to parietal cortex.

Although the weight of evidence is inconsistent with a storage role for PFC, virtually every electro-

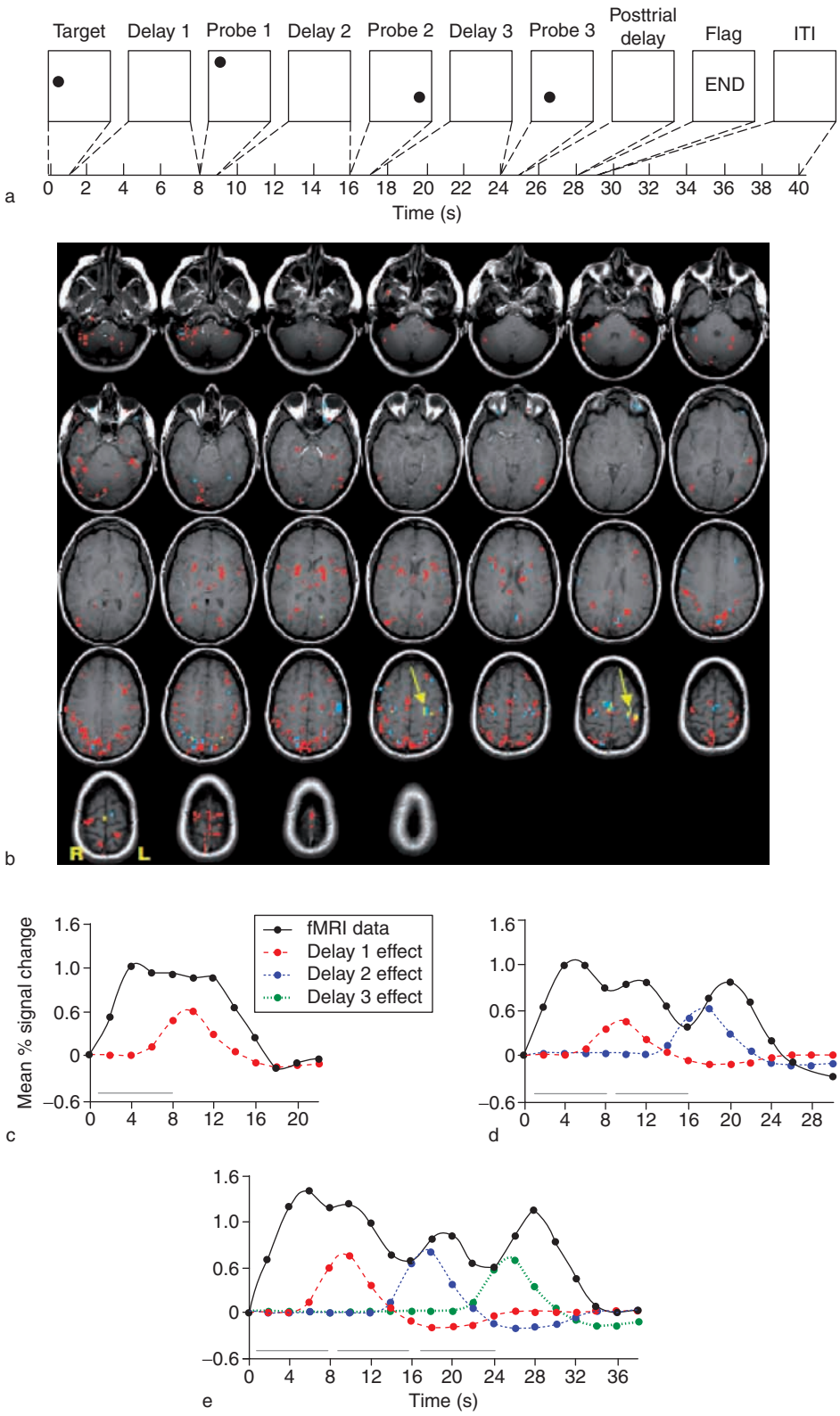

Figure 2 A functional magnetic resonance imaging experiment demonstrating sustained memory-related activity for location across multiple delay periods. (a) Behavioral task. Individuals view and encode the target location, then, after a 7 s delay, indicate with a button press whether the probe does or does not appear in the same location. One-third of the trials end after probe 1. On two- and three-delay trials, the offset of probe 1 is followed by another delay period, after which individuals evaluate the location of probe 2 with respect to the target. On one- and two-delay trials, the 'END' message appears at times 12 and 20 s, respectively. (b) Results from a single representative individual. Voxels in red showed sustained delay-period activity for delay 1 only. Voxels in blue are the subset of voxels

physiological and neuroimaging study of primate working memory finds delay-period activity in this region of the brain. Thus, it is important to address the role of this activity. If this activity does not represent the storage of information, what are alternative explanations of its function? One possibility is that PFC plays an important role in the control of potentially disrupting effects of distraction and interference during working memory tasks. Lesion, electrophysiology, and fMRI studies have all provided evidence that dorsolateral PFC can accomplish this by controlling the gain of activity in sensory processing regions, such that the delay-period processing of potentially distracting stimuli is suppressed. Analogously, a region of ventrolateral PFC has been implicated in the control of proactive interference, the deleterious effect of previous mental activity on current task performance. In addition, an important function of delay activity in PFC may be the attentional selection of task-relevant information, as well as in planning for the response. This is consistent with the well-characterized role for PFC in the biasing of stimulus–response circuits so that novel or less salient behaviors can be favored over well-learned associations and behavioral routines. Finally, the potential role of PFC in the integration of cognitive and motivational factors is supported by the modulation of delay-period activity of its neurons during spatial memory tasks by the type of the anticipated reward.

Conclusion

The emerging picture of the neural basis of working memory is of a class of behaviors that does not depend on one or more functionally specialized regions. Rather, working memory is supported by the coordinated activity of circuits responsible for the sensory processing of the critical information, and, to varying degrees, those that control the flexible allocation of attention and the selection of task-relevant behavior.

See also: Attentional Functions in Learning and Memory; Cognition: An Overview of Neuroimaging Techniques; Event-Related Potentials (ERPs); Hippocampus and Neural Representations; Prefrontal Cortex: Structure and Anatomy; Somatosensory Perception; Synaptic Plasticity and Place Cell Formation; Working Memory: Capacity Limitations.

Further Reading

Baddeley A (1992) Working memory. *Science* 255: 556–559.

Bisley JW and Pasternak T (2000) The multiple roles of visual cortical areas MT/MST in remembering the direction of visual motion. *Cerebral Cortex* 10: 1053–1065.

Cowan N (2005) *Working Memory Capacity.* Hove, UK: Psychology Press.

Fuster J (1995) *Memory in the Cerebral Cortex*. Cambridge, MA: MIT Press.

Goldman-Rakic PS (1995) Cellular basis of working memory. *Neuron* 14: 477–485.

Hebb D (1949) *Organization of Behavior.* New York: Wiley.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24: 167–202.

Pasternak T and Greenlee M (2005) Working memory in primate sensory systems. *Nature Reviews Neuroscience* 6: 97–107.

Petrides M (2000) Dissociable roles of mid-dorsolateral prefrontal and anterior inferotemporal cortex in visual working memory. *Journal of Neuroscience* 20: 7496–7503.

Postle BR (2006) Distraction-spanning sustained activity during delayed recognition of locations. *NeuroImage* 30: 950–962.

Postle BR (2006) Working memory as an emergent property of the mind and brain. *Neuroscience* 139: 23–38.

Postle BR and Hamidi M (in press) Nonvisual codes and nonvisual brain areas support visual working memory. *Cerebral Cortex-Advance Access*. published online on December 5, 2006.

Romo R and Salinas E (2003) Flutter discrimination: Neural codes, perception, memory and decision making. *Nature Reviews Neuroscience* 4: 203–218.

Todd J and Marois R (2005) Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective & Behavioral Neuroscience* 5(2): 144–155.

Zaksas D and Pasternak T (2006) Directional signals in the prefrontal cortex and area MT during a working memory for visual motion task. *Journal of Neuroscience* 26: 11726–11742.

with delay-period activity sustained across delay 1 and delay 2. Voxels in yellow are the still smaller subset with delay-period activity sustained across all three delays. Note that although many regions, including the prefrontal cortex, show delay-period activity during delay 1, only dorsal-stream parietal and frontal oculomotor regions sustain this delay-period activity across all three delays. Arrows highlight the voxels from left frontal eye field (the activity of which is shown in panels c, d, and e). (c) Activity from three-delay voxels in the left frontal eye field, averaged across one-delay trials. 'Delay 1 effect' reflects the estimated magnitude of delay 1 activity. Gray bar along the horizontal axis indicates the duration of the delay period. (d) Activity from these same frontal eye field voxels averaged across two-delay trials. Graphical conventions are the same as in panel b. (e) Activity from these same frontal eye field voxels averaged across three-delay trials. Graphical conventions are the same as in panel c. Adapted from Postle BR (2006) Distraction-spanning sustained activity during delayed recognition of locations. *NeuroImage* 30: 950–962.

Signal Design Rules in Animal Communication

H C Gerhardt, University of Missouri, Columbia, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Communication occurs when signals produced by one animal affect the behavior of another animal; here the focus is on exchanges between members of the same species. A widely held assumption is that communication leads to behaviors that ultimately increase the fitness of both senders and receivers. Signal design is the result of natural and sexual selection on signal structure that contributes to this increase in fitness. The rationale is that if receivers use signal information to make behavioral decisions, those decisions are likely to affect both their own fitness and that of the signaler and henc e, directly or indirectly, exert selection on the form of the signal. Such reciprocal selection by senders and receivers is a hallmark of communication and generates the expectation that signal form and receiver selectivity will coevolve to some degree. One obvious outcome is elaboration or exaggeration of one or more components of a signal, and such changes may also result in more efficient communication by increasing signal detectability against background noise. Receivers can also promote innovations in signal design if, for example, rare mutations result in signals that tap sensory biases that conventional signals do not.

Understanding signal design depends on analyzing the physical properties of signals and assessing variation in these properties that correlates with the behavior of both the sender and the receiver under a given set of circumstances. We also need to know about the capabilities of the species in question to produce and perceive signals and the characteristics of the environment that can affect the signal during transmission. The signal's modality immediately delimits the set of relevant physical properties needed to describe signals as well as the likely constraints on signal production, reception, and the environmental variables that are likely to affect its transmission. The need to objectively describe signals is reinforced by the fact that many animals use signals (many chemicals, ultraviolet and infrared wavelengths of light, infrasonic and ultrasonic sounds, and weak electric currents) that humans cannot even perceive.

The three signal modalities that are most widely used for animal communication will be the main organizing feature of this article. Within each modality, signals that operate over relatively long distances often have two functions in a reproductive context: attraction of mates and repulsion of potential competitors. Other long-range signals serve to recruit members of a social group to exploit a food resource or for defense. Short-range signals are most frequently used for further assessments of potential mates or rivals. Whereas the opportunistic nature of evolution assures that signals with the same function will show a diversity of forms, it is possible to describe some generalizations about the properties of signals within each modality that make them suitable for long- or short-range communication.

Olfactory Signals

The first signals to evolve were almost certainly chemical in nature. Chemical messengers are used extensively within and between cells, so modifications for their use in intraspecific communication would include mechanisms for senders to expel them into the environment and for receivers to acquire, detect, and recognize them. Chemical signals that are transmitted from individual to individual are called pheromones.

One of the first formal approaches to signal design considered relationships between the chemical composition of pheromones, the capabilities of senders and receivers to produce and detect these signals, problems of transmission, and the kinds of information to be communicated. Pheromones should generally have a minimum size of about five carbon atoms plus functional groups (alcohol, ketones, etc.) because smaller and less complex molecules can be too volatile in air and may not be sufficiently specific for coding information. The latter limitation can be offset by mixtures of simple pheromones, which are commonly found in invertebrates and mammals. The upper limit of about 20 carbons may be set by production and storage costs and by low diffusion rates in air. Even though production costs have seldom been determined, pheromones are still probably less energetically expensive than signals of other modalities because conversion from chemical to some other form of energy is unnecessary.

The active space of a pheromone is the area in which its concentration is above the threshold of detection by receivers; the active space corresponds to the maximum range or distance over which communication is possible. The fade-out time refers to the time after emission when the concentration falls below threshold everywhere. In still air or water, animals can increase the active space either by increasing

Table 1 The predicted spread and fade-out (active space) of a signal comprising a substance of low molecular weight

Q/K[a]	*Diffusion coefficient*		*Maximum radius of active space (cm)*		*Time required to reach maximum radius (s)*		*Fade-out time (s)*	
	Air	*Water*	*Air*	*Water*	*Air*	*Water*	*Air*	*Water*
1	10^{-1}	10^{-5}	0.6	0.6	0.4	4×10^3	1	10^4
10^2	10^{-1}	10^{-5}	2	2	8	8×10^4	20	2×10^5
10^4	10^{-1}	10^{-5}	10	10	1.5×10^2	1.5×10^5	5×10^2	5×10^6
10^6	10^{-1}	10^{-5}	60	60	4×10^4	4×10^7	10^4	10^8

Diffusion coefficient arbitrarily set at 0.1 cm S^{-1}.
[a]Q, number of pheromone molecules released at one time; K, detection threshold.
From Wilson EO (1970) Chemical communication within species. In: Sondheimer E and Simeone JB (eds.) *Chemical Ecology*, pp. 135–155. New York: Academic Press.

the number of pheromone molecules released at one time (Q) or by increasing their sensitivity (K = detection threshold). As shown in Table 1, the active space is limited and the time to reach its maximum range is long for a pheromone with a relatively high diffusion rate, even when its concentration is high initially and receivers are sensitive. Although the maximum transmission distance can be achieved by releasing pheromones into a wind or current, transmission is uncertain because the location of the receiver is unknown and the direction of flow is unidirectional and may change unpredictably. Moreover, if the wind or current is too strong, many molecules are dispersed laterally, thus reducing the concentration downrange. These attributes make olfactory communication the least reliable modality for long-range communication and source localization.

The fade-out time in still air or water is also dependent on the ratio of Q to K and the diffusion rate. For example, the use of small molecules with a relatively high diffusion rate will result in the rapid dispersal of the message to its maximum range and decrease the fade-out time. Rapid dispersal is a desirable characteristic of alarm, threat, and recruitment signals used for guiding members of a social group to a new food source or to defend against a predator. Short fade-out times mean that alarm and recruitment signals do not persist after the food source is depleted or the predator is no longer present. Pheromones with short fade-out times are also useful for short-range courtship because they do not alert rivals in the neighborhood of the presence of a receptive female, and they also allow senders to change messages more rapidly. In contrast, large molecules with slow diffusion rates can have long-duration effects that are useful for marking of territories or a trail to a long-lasting food source or shelter.

The disadvantages of pheromones with respect to reliable transmission and localization are offset by the fact that they can be used day and night in a variety of habitats. Another advantage is that the extreme specificity of some olfactory systems facilitates the identification not only of species and populations but also of colonies and individuals. Human olfaction is sufficiently poor that we cannot appreciate the subtle information in pheromones, and special instrumentation is necessary to show the effects of selection on the evolution of pheromones. For example, gas chromatography has been used to assess individual and population variation in cuticular hydrocarbon pheromones in *Drosophila*. These data made it possible to study geographic variation and to conduct genetic experiments that demonstrated the shift in signals that sometimes occurs when two formerly isolated species come into contact (reproductive character displacement).

Visual Signals

In contrast to olfactory communication, visual communication is extremely rapid and has the potential to operate over long distances. Moreover, because nearly all communication by light is line-of-sight, detection and localization of visual signals are usually simultaneous. The line-of-sight requirement does, however, limit communication range whenever there are opaque objects between senders and receivers, and the fact that most visual signals depend on reflected light means that communication at night, in caves, and in murky or deep water is difficult if not impossible. Animals that produce their own light source (bioluminescence) to produce signals may be limited by energetic costs.

The design of visual signals used for long-distance intraspecific communication hinges on making signals conspicuous against the visual background. Such signals make use of color, contrast, and movement, to which visual systems of receivers are especially sensitive. At the same time, predators may take advantage of these same features, imposing a fitness cost to senders that may exceed the production cost.

Thus, many species have evolved mechanisms by which conspicuous body parts can be concealed until potential mates or rivals are present. Indeed, the visual design of whole animals often serves as camouflage unless they are protected by toxins or venom (or mimic such creatures), in which case, conspicuous patterns and colors are common. One fascinating innovation for increasing detection by prospective mates without incurring significant risks of predation is the use of decorations as signals in Australian bowerbirds (Figure 1). The items used to adorn their bowers show greater visual contrast in their respective habitats than do their plumage characteristics, even in males that are more brightly colored and patterned than females.

Because long-range mate attraction often involves large repetitive movements that are energetically costly and risky, prospective mates (usually females) often use the rate and duration of such displays to assess the physical condition and perhaps the genetic quality of the signaler. The same is probably true for some rivals. These qualities can be assessed by the color and intensity of nuptial pigmentation in animals such as fish, lizards, and birds. At closer distances, the ability to rapidly change display elements makes visual signals especially suitable for courtship, in which interactions between individuals of both genders depend on rapid exchanges of signals. Here subtle differences and quick reactions to the behavior of the prospective mate are more important than conspicuousness, which would also have the disadvantage of attracting rivals or predators. Visual signals used in short-range threats often involve postures that make the animal seem larger and that display body parts that can serve as weapons.

Figure 1 Photograph of a male satin bowerbird (*Ptilonorhynchus violaceous*) in its bower, which is decorated (foreground) with natural objects (feathers) and artificial objects (e.g., straws) that are light blue. In terms of the bird visual system, the dark blue, iridescent plumage of the male bird shows far less contrast (= less conspicuousness) against the visual background than do the bower ornaments. See Endler JA, Westcott DA, Madden JR, and Robson T (2005) Animal visual signals and the evolution of color patterns: Sensory processing Illuminates signal evolution. *Evolution* 59: 1795–1818 for details. Photo by H.C.G.

Specificity in visual signals resides in traits at two ends of a continuum. Some visual patterns are permanent (badges) that are specific to the species, population, or even individuals. Other such patterns may be temporary: appearing at the onset of the breeding season and gradually disappearing afterwards. Displays, which involve postures and movements, may also contain some highly stereotyped components (static properties) that vary little from one repetition to the next. Such elements are sufficiently stereotyped to identify the sender's species and sometimes even its individual identity. Other elements of displays show great variability (dynamic properties) and are often correlated with the sender's motivation, the social context (proximity of a prospective mate or rival), or the sender's physical or genetic fitness. The degree of stereotypy in individual signals can be quantified statistically and also applies to auditory signals.

Auditory Signals

Auditory signals are rapidly transmitted and can have a large active space. Depending on their physical properties, sounds can be relatively easy or hard to localize. Auditory signals can be used at any time of day, and sound waves readily bend around objects, such as trees, and hence can propagate well in densely vegetated habitats. The energetic costs of producing auditory signals can be high, and some species of insects and frogs exhaust nearly all their stored reserves during a breeding season. The risk of predation may be somewhat lower than in visual communication because sounds are harder to localize. Moreover, in vertebrates and other animals that power signals with their respiratory system, the sender can signal during ongoing behavior or from a concealed, stationary position without making large movements that could attract predators. Animals can also use private channels for auditory communication by using sounds that have frequencies below the hearing range of other animals (infrasonic signals of elephants), above the hearing range of predators (small birds), or even in the ultrasonic range (bats, whales, and dolphins). Most ultrasonic production occurs in the context of echo-location, however, and the social signals of most bats and marine mammals are much lower in frequencies that lie within the hearing range of ordinary mammals. Short-range signals, commonly produced during courtship or as predator alarms, are inconspicuous by virtue of low intensity. In addition, alarm signals of small birds

typically consist of high single-frequency notes with gradual onsets and offsets. These features make them difficult to detect and localize by predators. By contrast, sounds that serve to rally individuals of the same and other species for mobbing predators have properties (intense, broadband, with sharp onsets and offsets) that increase their conspicuousness and ease of localization.

The design of long-distance auditory signals used in intraspecific communication is constrained by mechanisms of production and environmental transmission. On the one hand, low-frequency sounds travel farther with less attenuation by the environment than do high-frequency sounds. On the other hand, the most efficient production of auditory signals requires that the dimensions of the sound producing structure be not very much smaller than the wavelength of the frequency with the most energy. In other words, small animals, such as insects, frogs, and small birds, will usually be limited to producing signals with relatively high frequencies. Some species ameliorate this problem to some extent by using leaves as acoustic baffles (to direct the sound energy) or take advantage of the resonant properties of burrows or hollows. Another tactic is to produce sounds from elevated positions because ground attenuation of high-frequency sounds is especially severe. Some studies of birds confirm that small species with relatively high-frequency songs tend to sing at higher elevations than do large species with low-frequency signals. Still another, counterintuitive tactic is for small animals such as frogs or insects to call on the ground at night when temperature inversions create a channel for lower-than-expected attenuation. Some birds and mammals may not be as constrained as insects and frogs, and correlations between the context ('motivation' or function) and signal structure are widespread. Aggressive vocalizations tend to be low in frequency and broad in bandwidth, whereas vocalizations of submissive or fearful animals are high in frequency and more tonal in nature.

The effective range of communication is also limited by the fact that features of the signal that convey information will be affected by the environment in various ways. Two common forms of signal degradation mainly affect the temporal structure of signals: reverberations and amplitude fluctuations. Reverberations arise because at a distance, sound waves reach the receiver both directly and indirectly after being reflected from the ground or vegetation. The indirect waves arrive with a slight delay and can interfere and partially cancel the direct waves or, more important, obscure the silent intervals between repeated short notes. Reverberations are especially problematic in forests when animals use signals that are rapidly modulated or trilled. Because degradation by reverberation is somewhat predictable in a given habitat, some birds have been found to take advantage of this phenomenon to estimate the distance of the singer. An additional cue for species with broadband signals is that the excess attenuation mentioned above will result in a greater loss of high-frequency components relative to low-frequency ones as distance increases. Amplitude fluctuations occur when pockets of warm air rise from the ground, especially during the day in open habitats. In contrast to reverberation, the periodicity of amplitude fluctuations is unrelated to any features of the signal but typically occur irregularly at rates of about 50 Hz or less. Some evidence for signal design (including signaling behavior) is that birds in forests tend to produce relatively long tonal signals that lack rapid modulations (**Figure 2(a)**). Birds in open habitats often have relatively long, rapidly trilled signals; low-frequency, short amplitude fluctuations will only partially degrade these trills, with enough consecutive pulses getting through to preserve information used by receivers. Numerous studies, however, have discovered only weak correlations between environmental variation and signal structure, and there are many counterexamples in which the auditory signals of a species propagate farther and with less degradation in nonpreferred habitats than in preferred ones.

Perhaps the most serious problem with auditory communication is masking interference from the sounds produced by rushing water, wind, and other animals, both conspecific and heterospecific. Masking by rushing water may result in the use of relatively high frequencies (higher than the dominant frequencies of this noise), the augmentation or substitution of auditory with visual signals, or both. For example, some neotropical frogs in such habitats have elaborate hand-waving signals. Because animals such as orthopteran insects, cicadas, frogs, and some birds often signal from aggregations, the most severe masking problems are biological. Some amelioration occurs because some simple animals such as insects and frogs have auditory systems that are roughly tuned to the frequencies emphasized in their advertisement calls. Hence the masking effect of signals of other species with different frequencies will be somewhat reduced. However, members of the same species, whose signals will be especially effective maskers because of their similarity in both frequency and temporal structure, will always be an individual's most serious competitor for clear 'airtime.' Indeed, the production of very loud advertisement calls, typical of temperate-zone anurans, probably represents an evolutionary response to masking interference in

choruses rather than a tactic for attracting mates from a distance. Masking interference has resulted in the evolution of diverse tactics involving signal timing among nearby individuals. At one end of the continuum, individuals vie to time their calls to lead and overlap those of rivals; at the other end, individuals seemingly attempt to maintain a completely nonoverlapping, alternating form of etiquette. In

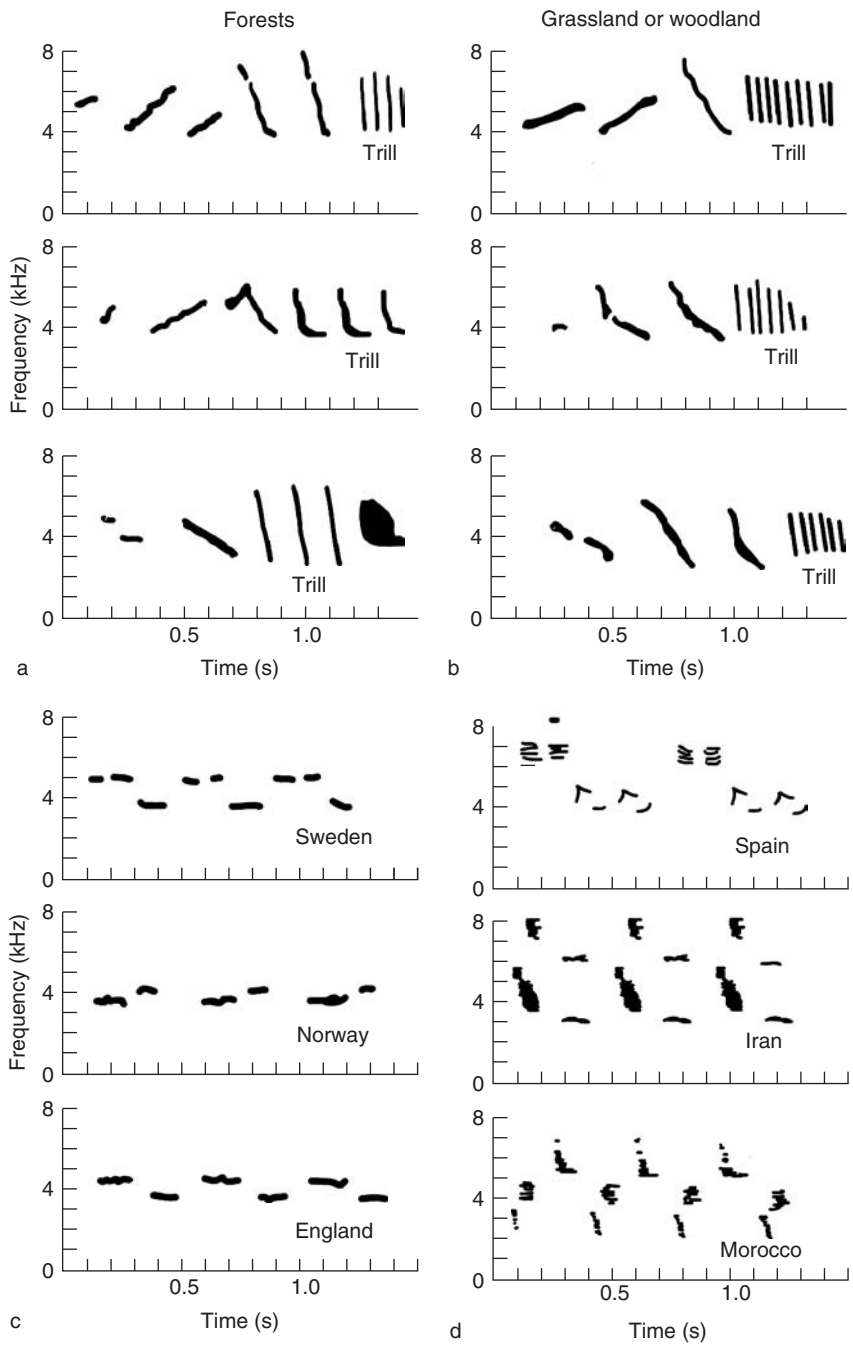

Figure 2 Habitat effects on signal structure. Sonograms (frequency vs. time) of bird songs of (a) and (b) the chingolo sparrow (*Zonotricia capensis*) and (c) and (d) great tit (*Parus major*) in different habitats are shown. Notice that the songs of both species in forests lack long rapid trills that would be subject to degradation by reverberations. Long rapid trills are typical in chingolo sparrows living in grasslands, and extensive frequency modulation is prevalent in great tits in open woodland where amplitude fluctuations rather than reverberations are the major source of signal degradation. From Bradbury JW and Vehrencamp SL (1998) *Principles of Animal Communication.* Sunderland, MA: Sinauer.

still other species, calls may be produced in a highly synchronous fashion to make the localization of individuals difficult for acoustically orienting predators.

As in the visual modality, auditory signals have the potential to encode many messages and to change them rapidly. Some properties, however, show limited variation within the signals of an individual (static properties), whereas other properties may change dramatically from one signal to the next (dynamic properties). The degree of variability may be quantified for comparison across signal properties and species (**Figure 3(a)**). Although variability within the signals of individuals is continuous, comparative studies usually find a distinctly bimodal distribution in which static and dynamic properties occupy opposite ends of the continuum. Static properties such as frequency are potentially useful in insects, anurans, and birds for identifying individuals or assessing their body size because they are consistent within the signals of individuals but typically vary significantly among individuals in the same population. Other static properties, such as fine-scale temporal patterning, are species specific and show little variation within and among individuals in any population and limited geographical variation as well. These properties are analogous to the static features used in visual communication, such as breeding plumage. By contrast, dynamic properties such as the rate and duration of signaling correlate with energetic costs and may be useful indicators of male fitness.

Both physical constraints on signal production, such as body size and energetic reserves, and the preferences of prospective mates contribute to patterns of variation in signal properties. For example, preference functions show a strong trend for stabilizing selection for static properties and directional selection (preference for one end of the range of values) for dynamic properties (**Figure 3(b)**). Directional selection on dynamic properties is probably an especially effective process for assessing the physical and genetic status of prospective mates because high values of these properties are costly both in terms of energy and predation risks. Of course, receivers respond to the entire signal, which has both static and dynamic properties, and the relative influence of such criteria on signal selection may vary according to the fitness consequences of mate choice. For example, females in areas of overlap with other species having static properties with values similar to those of conspecific males weigh static properties more heavily than dynamic properties that may indicate differences in quality among conspecific males. This can result in reproductive character displacement of the mean values of static properties because mating with a male of the wrong species will result in a much greater fitness loss than will mating with a low-quality conspecific male. This and other examples serve to emphasize that traits providing information that enhances receiver fitness may become subject to selection in the context of communication even if those traits did not originally evolve as signals.

Multimodal Communication

Animals are not restricted to using signals of a single modality, and many displays that are especially fascinating to humans combine elaborate visual and auditory elements. Sequential use of different kinds of signals is also common. Loud long-range auditory signals produced in a dense forest, for example, are likely to be replaced by short-range auditory, visual, and olfactory signals at close range. Chemical and tactile signals are more common at even closer distances. Signals of different modalities can also be used to transmit different kinds of messages and can be directed to different kinds of receivers. The diversity of animal communication systems and sensory capabilities makes it difficult to find general correlates between signal modality and function.

Summary

Visual and auditory signals have the greatest potential for rapid long-range communication. The properties of these signals or the kind of signal can also be changed quickly, thus allowing for a multitude of messages to be delivered in a short time. Visual and auditory signals can also be designed or modified to make the localization of their source easy or difficult. Visual signals have three disadvantages in comparison with auditory signals: (1) They usually depend on reflected light, and hence are they not as useful at night or in other dark environments; (2) they are not useful in densely vegetated or other habitats where line-of-sight vision is restricted; and (3) in many animals, visual displays require the commitment of body parts (wings, legs) that are also needed for flight or locomotion. Chemical signals are almost certainly less expensive to produce than are other kinds of signals, and they can be used at night and in all sorts of habitats. But the range of action is limited, and the localization of long-range pheromones is uncertain; the persistence of chemical signals can also limit the ability to change messages rapidly. Signals of all three modalities have physical properties whose values vary more or less at the individual and among-individual levels, thus providing the potential for most signals to convey multiple messages. One of the great challenges

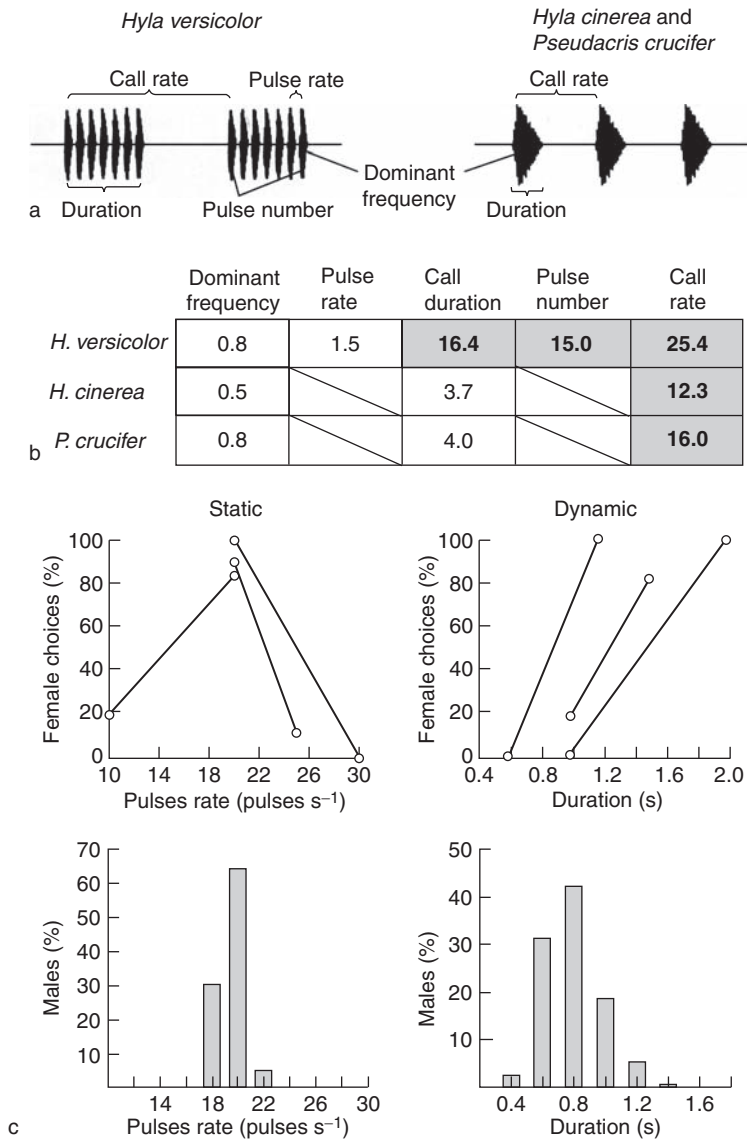

	Dominant frequency	Pulse rate	Call duration	Pulse number	Call rate
H. versicolor	0.8	1.5	**16.4**	**15.0**	**25.4**
H. cinerea	0.5		3.7		**12.3**
P. crucifer	0.8		4.0		**16.0**

Figure 3 Static and dynamic acoustic properties in treefrogs. (a) Gray treefrogs (*Hyla versicolor*), green treefrogs (*Hyla cinerea*), and spring peepers (*Pseudacris crucifer*) produce long-range advertisement calls that vary from call to call in dominant frequency, call duration, and call rate; the repetition rate of the pulses making up the calls (pulse rate) of gray treefrogs also varies, and the number of such pulses largely determines call duration; (b) Variability in these properties is quantified by the coefficient of variation (CV), which is the standard deviation expressed as a percentage of the mean, thus allowing comparisons of variability across properties and across species. Static properties (unshaded) generally have a CV of 5% or less whereas dynamic properties (shaded) have a CV of 10% or more. (c) Female gray treefrogs prefer values near the mean (= stabilizing preferences) for static properties such as pulse rate, and values with high values (= highly directional preferences) for dynamic properties such as call duration. The percentages of females choosing each of two alternatives with different values are shown by circles connected by lines (e.g., in the left panel, 100% of the females chose a call with a pulse rate of 20 pulses s^{-1}, the mean for the population, and none chose the call with a pulse rate of 30 pulses s^{-1}). Histograms show the distribution of values of pulse rate (left panel) and duration (right panel) in a population of gray treefrogs. From Bradbury JW and Vehrencamp SL (1998) *Principles of Animal Communication.* Sunderland, MA: Sinauer.

of the study of animal communication is to learn how receivers assess multiple messages, especially conflicting ones, in order to make fitness-enhancing decisions.

See also: Animal Communication: Honesty and Deception; Communication Networks and Eavesdropping in Animals; Communication in Terrestrial Animals; Endocrinology of Animal Communication: Behavioral; Facial Expression in Primate Communication; Game Theory and the Economics of Animal Communication; Multimodal Signaling in Animals; Pheromones in Mammals; Pheromones in Humans and Social Chemosignals; Primate Communication: Evolution; Referentiality and Concepts in Animal Cognition; Signal Production and Amplification in Birds; Visual Signaling in Animals; Vocal Communication in Birds.

Further Reading

Bradbury JW and Vehrencamp SL (1998) *Principles of Animal Communication*. Sunderland, MA: Sinauer.

Catchpole CK and Slater PJB (1995) *Bird Song: Biological Themes and Variations*. Cambridge, UK: Cambridge University Press.

Endler JA, Westcott DA, Madden JR, and Robson T (2005) Animal visual signals and the evolution of color patterns: Sensory processing illuminates signal evolution. *Evolution* 59: 1795–1818.

Gerhardt HC (1991) Female mate choice in treefrogs: Static and dynamic acoustic criteria. *Animal Behaviour* 42: 581–595.

Gerhardt HC and Huber F (2002) *Acoustic Communication in Insects and Anurans: Common Problems and Diverse Solutions*. Chicago: University of Chicago Press.

Hauser MD (1996) *The Evolution of Communication*. Cambridge, MA: MIT Press.

Hödl W and Amézquita A (2001) Visual signaling in anuran amphibians. In: Ryan MJ (ed.) *Anuran Communication*. Washington, DC: Smithsonian Institution.

Marler PR and Slabberkoorn H (eds.) (2004) *Nature's Music: The Science of Birdsong*. San Diego: Elsevier Academic Press.

Morton ES (1977) On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist* 111: 855–869.

Wilson EO (1970) Chemical communication within species. In: Sondheimer E and Simeone JB (eds.) *Chemical Ecology*, pp. 135–155. New York: Academic Press.

Wilson EO and Bossert WH (1963) Chemical communication among animals. *Recent Progress Hormone Research* 19: 673–716.

Signal Identification: Peripheral and Central Mechanisms

G Pollack and Krahe R, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Communication signals may contain various sorts of information about the sender, including its identity (whether of species, gender, or individual), sexual receptivity, aggressiveness, and fitness. This information is encoded into physical features of the signal, such as its chemical composition, sound frequency, or rhythmical structure. The receiver must detect the signal, discriminate it from other sensory stimuli, and decode the signal's contents (its 'message'). Signal processing begins with accessory sensory structures (e.g., pinnae and olfactory mucosa), which may constitute a first filtering stage. Further, more detailed processing is accomplished by sensory receptor neurons and within the central nervous system. This article focuses on the neural mechanisms underlying the processing of communication signals. The emphasis is on sensory systems that are highly specialized for communication, namely olfaction, electroreception, and audition.

Peripheral Processing

Matched Peripheral Filters

Communication signals are often narrow in bandwidth, consisting of a restricted range of sound frequencies, chemical compounds, or light wavelengths. A common peripheral mechanism for enhancing signal reception is selective responsiveness, by the receiver, to the same bandwidth used by the signaler. Selective 'tuning in' to the signal bandwidth attenuates sensory stimuli falling outside of this bandwidth, thus increasing the signal-to-noise ratio of signal reception. It should be noted, however, that selective tuning of the receiver may also be a mechanism by which the receiver evaluates how well the sender is able to produce the signal, which may be an indicator of fitness.

Matched filters for sound frequency In several insects that communicate acoustically, the frequency sensitivity of the receiver is matched to the sound frequencies used in acoustic signals. For example, the acoustic signals (songs) of crickets have most of their power in restricted frequency ranges, and the majority of their auditory receptor neurons are most sensitive to frequencies in this same range (Figure 1). The ears of mosquitoes are their antennae, which vibrate in response to near-field sound, thereby exciting a sensory organ in the antenna's basal segment. Vibrational amplitude of the antenna of male *Toxorhynchites brevipalpis* is greatest for sound frequencies near 420 Hz, which matches precisely the wingbeat frequency of females. The acoustical properties of the antennae are set in part by 'standard' physical parameters such as length, mass, and stiffness. However, active mechanical input to the system, analogous to that provided by outer hair cells of the organ of Corti of vertebrates, also contributes to the frequency selectivity and remarkable (nanometer-level displacement) sensitivity of the system.

Matched filtering also occurs in acoustically signaling vertebrates; indeed, the concept of the matched filter was developed in the context of acoustic communication among frogs. Communication sounds of the bullfrog, *Rana catesbeiana*, have a bimodal spectrum. Power is divided into low- and high-frequency bands, centered on approximately 200 and 1500 Hz, respectively, with a pronounced absence of power at an intermediate range centered on 500 Hz. Upon hearing the calls of conspecifics, males respond with their own calls. Studies using evoked calling as a behavioral assay showed that for maximal effectiveness, a signal must contain energy in both the low- and the high-frequency bands and, in addition, must lack energy in the intermediate band. These spectral requirements have clear parallels in the frog's peripheral auditory system. There are two distinct sensory structures in the inner ear: the amphibian papilla, which responds best to low-frequency sounds, and the basilar papilla, which is tuned to higher frequencies. Neurons of the amphibian papilla have the interesting property that their responses to a low sound frequency are suppressed by simultaneous stimulation with a slightly higher frequency. Thus, neurons from the basilar papilla are stimulated by high frequencies, and those from the amphibian papilla are driven by low frequencies, with the additional requirement of an absence of midfrequencies. The combined input from both papillae thus represents the structure of the natural signal and matches the spectral requirements for behavioral effectiveness.

The degree of matching between sender and receiver may vary through time. Male midshipman fish court females using acoustic signals. The tuning at the level of the inner ear of females is modified seasonally by steroid hormones and more closely matches the spectral properties of male calls during the breeding season than when the females are in nonreproductive state.

Electroreception Weakly electric fish produce electrical discharges and monitor the resulting electrical

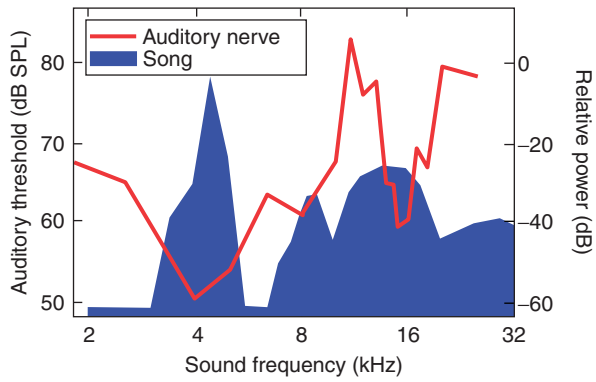

Figure 1 A matched frequency filter. The red curve shows sensitivity of the auditory nerve of a cricket (*Gryllus campestris*) as a function of sound frequency. The blue area represents the frequency content of the species' calling song, a signal that is produced by males and serves to attract females and adjust intermale spacing. Auditory sensitivity is greatest (threshold sound level required to produce a neural response is lowest) at frequencies near 4 and 16 kHz, the same frequencies that are most prominent in the song. Adapted from Nocke H (1972) Physiological aspects of sound communication in crickets (*Gryllus campestris* L.). *Journal of Comparative Physiology* 80: 141–162.

fields with electroreceptors in the skin. In wave-type fish, the discharges are periodic and quasi-sinusoidal. The frequency of the electric organ discharge can range from 50 to 2200 Hz (depending on species, gender, and individual identity), but it is quite constant within an individual. The fields of conspecifics cause modulations in the fish's self-generated field; thus, detection of communication signals relies on the tuning of the electroreceptor organs in the skin to the fish's own electric organ discharge frequency. Accordingly, the best frequency of the primary afferent fibers is matched to the discharge rate of the electric organ of the individual fish, and their spiking activity shows enormously steep filter curves with slopes of up to 60 dB per octave. The mechanism underlying this sharp tuning seems to be based on electrical resonance of the receptor cells, as determined by the kinetic properties of ion channels. In a manner similar to that described previously for midshipman auditory receptors, the tuning of the electroreceptors follows changes in the fish's discharge frequency that occur in connection with changes in reproductive status. Interestingly, discharge rate and electroreceptor tuning are both under the control of steroid hormones, although apparently via independent pathways.

Another group of electric fish produces discrete, pulse-like electrical discharges rather than waves. Their electroreceptors are more broadly tuned than those of wave-type fish. Broad tuning presumably supports the exquisite sensitivity of the receptors to the timing of the onset and offset of electric organ discharges of conspecifics. This would allow the fish to measure pulse duration, an indicator of species and sex. In this case, broad, rather than narrow, receptor tuning may contribute to the processing of communication signals.

Vision A rare example of matched filters in a system using visual signals is found in fireflies, in which the absorption spectra of visual and screening pigments result in color sensitivity that is tuned to the wavelengths of conspecific bioluminescent emissions.

Matched temporal filters A typical feature of acoustic communication signals is that much of the information that is transmitted is contained in modulations of signal amplitude. Auditory receptor cells of grasshoppers as well as primary auditory afferents of bull frogs show enhanced transmission of information about sound stimuli if the distribution of amplitudes and the frequency characteristics of amplitude modulations match the characteristics of the natural species-specific songs. The optimal stimuli driving grasshopper auditory receptor cells, however, do not match the full song structure but, rather, resemble a subset of features of the songs, such as the sharp increases in sound level following the gaps between pulses. For particularly salient features of the songs, the variability of spike timing of receptor cells can be as low as 0.15 ms, an order of magnitude smaller than the fastest stimulus timescale.

Chemoreception Precise matching between signaler and receiver is perhaps most spectacular in pheromonal communication systems, in which the signal consists of a mixture of chemical compounds, the precise makeup of which may code for the species, gender, and/or individual identity of the signaler, as well as the signal's meaning. The complex social systems of ants and other hymenopteran insects are regulated in large part by chemical signals, which are used in reproduction, to lead nest mates to food sources, to alert them to possible threats, and for a host of other purposes. Male moths locate sexually receptive females by tracking the pheromone plumes that they release into the air. Mammalian pheromones mediate mother–infant communication as well as recognition of the gender, strain, reproductive state, and unique identity of conspecific individuals.

The neuronal basis of pheromonal communication has been studied most thoroughly in insects and mammals. In both groups, pheromones are detected by peripheral chemoreceptor neurons that combine high sensitivity with exquisite selectivity. It has been estimated that capture of as few as five pheromone molecules by the moth *Spodoptera littoralis* is

sufficient to elicit a physiological response (change in heart rate). Pheromone-sensitive olfactory receptor neurons of mice may respond to concentrations of their specific ligand as low as 10^{-11}M yet remain unresponsive to other pheromone components at much higher (100 000-fold) concentrations. This remarkable specificity of the receptor neurons reflects the properties of the receptor molecules that they express, and it differs from the broader response profiles of olfactory receptors that respond to nonpheromonal stimuli. Thus, pheromone detection appears to be accomplished by a series of 'labeled lines': distinct receptor neurons, expressing distinct receptor molecules, report the presence and concentration of specific, semiotically important, chemicals to central processing areas where inputs from the separate lines are integrated.

Central Processing

Signals, once detected peripherally, are processed further in the central nervous system. The tasks of central processing include further sharpening of neural responses to specific stimulus features; generating new forms of selectivity, for example, for the combined presence of stimulus components that are detected by separate groups of receptor neurons, and performing more 'abstract' functions, such as decoding the information that is encoded into receptor–neuron spike trains, discrimination among signals, and evaluation of the signal's (and thus the signaler's) quality.

Inhibition Enhances Selectivity

In some systems, the selectivity for communication signals that is established by peripheral tuning is sharpened in the central nervous system through inhibition. Some of the most selective neurons yet described occur in brain nuclei of songbirds. These neurons may respond exclusively to the song of the same individual bird in which the neuron resides. This exquisite selectivity is believed to be important for song learning and is due in part to inhibition. When an ionic current that mediates inhibition is blocked pharmacologically, the neurons respond to a broader range of stimuli. In the midbrain of frogs, selectivity for frog song-like sound frequencies is sharpened by 'lateral' inhibition – that is, inhibition elicited by sound frequencies above and/or below that to which the neuron is most sensitive (Figure 2). A remarkable example of the role of inhibition in sharpening spectral selectivity occurs in an identified interneuron of tettigoniids (long-horned grasshoppers) that is tuned, in a species-specific manner, to the major spectral component of conspecific song. The species specificity of tuning is 'sculpted' from a broadly tuned response curve, which is identical across species, by synaptic inhibition originating in neurons (the identities of which are unknown) that are tuned to higher and lower frequencies. The species-specific differences in the tuning of this neuron are eliminated when inhibition is blocked pharmacologically.

An inhibitory mechanism is responsible for the dedication of an entire class of electroreceptors to communication signals in mormyrid weakly electric fish. Primary afferents conveying information from Knollenorgan electroreceptors in the skin of these fish to a hindbrain nucleus respond to both the fish's own electric organ discharge and the discharge of conspecifics. At the level of the hindbrain, a corollary discharge mechanism effectively gates information about conspecific signals by inhibiting all responses

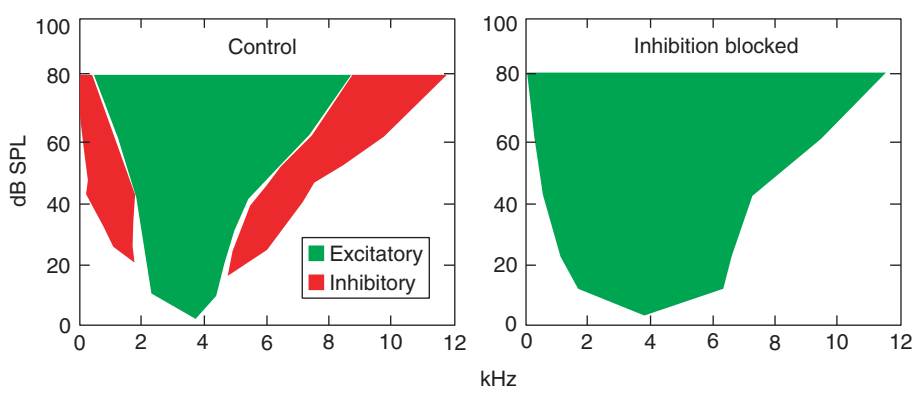

Figure 2 Sharpening of frequency tuning by lateral inhibition. The graph on the left (control) indicates the frequency/intensity combinations that excite (green) or inhibit (red) an auditory neuron in the brain of a frog. The range of sound frequencies that excites the neuron is restricted by inhibition produced by lower and higher sound frequencies. When this inhibition is blocked using the drug bicuculline (right), the formerly inhibitory regions become excitatory, broadening the range of frequencies that can excite the neuron. Adapted from Hall JC (1999) GABAergic inhibition shapes frequency tuning and modifies response properties in the auditory midbrain of the leopard frog. *Journal of Comparative Physiology A* 185: 479–491.

related to the fish's own electric activity such that only the electric signals generated by conspecifics are processed.

Combination Sensitivity

Another common theme of central processing is combination sensitivity – that is, selective responsiveness to the simultaneous occurrence of separate components of a communication signal. In the auditory midbrain of frogs, most neurons at early stages of processing have frequency-tuning properties similar to those of auditory receptors, with best sensitivity to frequencies corresponding to receptors of either the amphibian or the basilar papilla. At higher processing levels (midbrain and thalamus), however, neurons are found that can be driven only by simultaneous presentation of both the low- and high-frequency components of conspecific acoustic signals, paralleling the requirement for both spectral components for maximum behavioral effectiveness. Pheromonal signals of moths typically consist of blends of several components. Closely related species may use the same chemical components but with species-specific differences in their relative concentrations. Thus, an important task of central processing is to gauge the level of activity across component-specific peripheral channels. Inputs from these channels converge in the central nervous system. In the antennal lobes of moths, different neuropil areas (glomeruli) receive input from receptor neurons that are selective for different pheromone components. Some interneurons within the lobe have dendrites projecting to only a single glomerulus and accordingly are excited by only one of the pheromone components (and in fact may be inhibited by the other component via local circuitry), whereas other interneurons, with dendritic branches in multiple glomeruli, are excited by multiple components. Further convergence of excitation induced by multiple pheromone components occurs at still higher processing stations in the brain.

Central Processing of Temporal Pattern

Communication signals are often strongly rhythmic, with distinct components occurring at specific times, and a signal's 'message' may be carried by its temporal pattern. For example, the temporal structure of insect and frog songs may inform listeners of the species identity of the signaler and may also carry information about its sex, size, robustness, and other fitness-related attributes. Neural specializations for behaviorally relevant features of signals may be expressed in two ways: (1) increased responsiveness to meaningful temporal features (i.e., higher firing rate) and (2) more accurate representation of the rhythmical structure by the timing of action potentials (Figure 3).

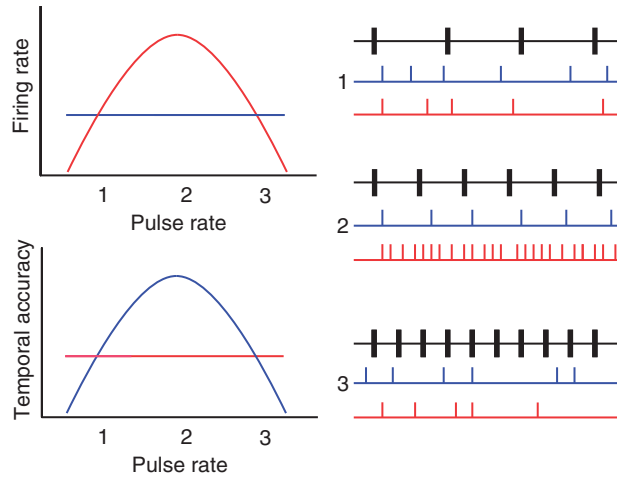

Figure 3 Firing-rate and time codes for temporal pattern. Each graph depicts the responses of two neurons that are 'tuned' to the intermediate pulse rate. The traces on the right represent the spike trains of the two neurons to three stimulus pulse rates: one slower than optimal (1), the optimal rate (2), and a rate that is faster than optimal (3). The red neuron uses a firing rate code; firing rate is highest for the optimal pulse rate, but the precise time of each action potential is unrelated to the times of the individual stimulus pulses. The blue neuron uses a time code; firing rate does not vary with stimulus pulse rate, but action potentials are precisely time-locked to the stimulus pulses only for the optimum pulse rate. These are idealized examples. In real neurons, both firing rate and precision of spike timing typically vary with stimulus pattern, although one or the other type of coding strategy may predominate.

Sensitivity to pulse rate In crickets, tettigoniids, and frogs, communication signals consist of brief sound pulses that are repeated at species-specific rates, and neurons in the central nervous systems of these animals respond most strongly and/or with greatest temporal accuracy to stimuli with species-typical pulse rates. Several central mechanisms have been implicated in pulse rate sensitivity. 'Filter' neurons are found in the brains of both crickets and frogs, which may be low-pass (i.e., respond most strongly to low pulse rates), high-pass, or bandpass. An intriguing, although unconfirmed, idea is that bandpass selectivity might be an example of a neural implementation of a logical 'AND' operation, whereby bandpass neurons would acquire their selectivity for a restricted range of pulse rates by requiring simultaneous input from low-pass and high-pass neurons, the pass bands of which would overlap for species-typical signals (Figure 4). If true, this would be yet another example of combination sensitivity as a mechanism for signal recognition.

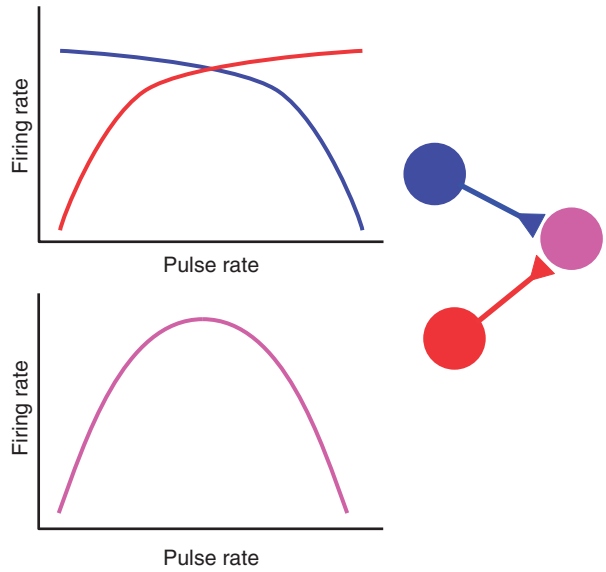

Figure 4 Pulse rate-selective filters. The top graph represents the responses of two neurons, one of which has 'low-pass' filtering properties (i.e., responds best to low pulse rates) (blue) and the other of which has high-pass properties (red). The bottom graph represents a bandpass neuron, which responds best to the species-typical range of pulse rates. The diagram on the right illustrates a neural circuit that could produce bandpass selectivity. The blue and red neurons correspond to the low-pass and high-pass curves shown in the top graph. They provide input to the magenta neuron, which responds best when receiving input simultaneously from both presynaptic cells. This occurs when the pulse pattern falls within the high-response ranges of both the low- and the high-pass neurons. Thus, the magenta neuron has bandpass selectivity.

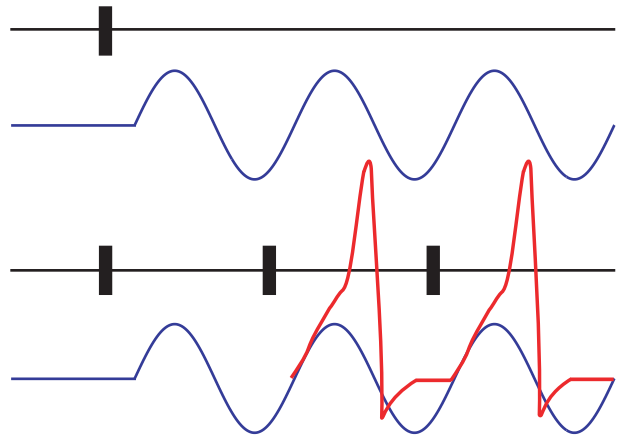

Figure 5 Pulse rate selectivity through neuronal resonance. The top traces illustrate the membrane potential of a neuron in response to a single stimulus pulse. The neuron experiences repeated waves of excitation, occurring with a specific period. The bottom traces show that if the stimulus pulses arrive at the neuron's resonant frequency, succeeding pulses will coincide with the intrinsic excitatory wave resulting from earlier stimuli. The responses to stimuli with this pulse rate will thus be enhanced, resulting, in this illustration, in production of action potentials.

The cellular mechanisms underlying pulse rate filtering are unknown in these examples, although studies on the sensitivity of electric fish to amplitude-modulation rates show that basic features of neuronal physiology can be exploited to build rate filters. For example, rate-dependent synaptic depression and/or electrotonic filtering of synaptic inputs can result in low-pass selectivity. Summation and/or facilitation of synaptic inputs, as well as intrinsic subthreshold ion conductances, might contribute to high-pass filtering.

Another mechanism for pulse rate selectivity, which does not require low-pass or high-pass filters, is based on the concept of neuronal resonance – that is, the endogenous generation of regular, periodic activity by single neurons or neural circuits. Oscillatory activity has been well studied in neural circuits underlying rhythmic motor behaviors such as walking, swimming, and flying. The initial sensory input (i.e., from the first stimulus pulse of a series) would serve to initiate resonance, and inputs from succeeding pulses would, if their period matches that of the resonator, reinforce the endogenous activity, resulting in response enhancement (Figure 5). A prediction of this model is that inputs arriving at integer fractions of the resonant frequency would also result in partial response enhancement because they would coincide with every second, third, *n*th cycle of the endogenous oscillation. Behavioral experiments on an insect, *Tettigonia cantans*, have provided tantalizing evidence for a resonant mechanism. Phonotaxis responses of females are strongest for stimuli with the species-typical pulse rate, 25 pulses s^{-1}, but secondary response peaks occur at 12.5 and at 8 s^{-1} (i.e., at one-half and one-third the optimum rate).

Central processing of bird song Bird song is characterized by complex modulations of amplitude and frequency of sound. Neurons at the level of the auditory midbrain show enhanced temporal precision when stimulated with species-specific sounds compared to noise stimuli that were modulated in amplitude and frequency over the same ranges as the natural signals. The enhanced precision of single neuron firing may support the high fidelity of encoding of natural song patterns by the neuronal population.

Although matched filters are of obvious benefit for the detection of signals from background, apparent 'mismatch' may play a role in enhancing discrimination between different natural signals. This relates to one hypothesis of optimal coding – that the responsiveness of a sensory neuron to particular stimulus components should be inversely related to the probability with which those components occur. Increased sensitivity to stimulus features other than the most common

ones maximizes a neuron's ability to discriminate among different naturally occurring stimuli. Such a relationship between stimulus statistics and neural responses has been observed in a number of systems, although without specific reference to its role in communication. With respect to bird song, neurons in the auditory midbrain and forebrain of zebra finches show a similar relationship between the temporal coding properties of neuronal ensembles and the power distributions of temporal modulations of natural songs. As a consequence, differences in the responses to similar, but nevertheless distinct, song segments are enhanced.

Signal rejection The emphasis thus far has been on mechanisms for recognition of a 'correct' signal. A related concept is the active rejection of 'incorrect' signals. Songs of male grasshoppers, *Chorthippus biguttulus*, consist of a sequence of syllables, each produced by a series of up-and-down stridulatory movements of the hindlegs. The legs are slightly out of phase so that the brief silent periods that occur at the transitions between upward and downward movements of one leg are obscured by sound produced by the other leg. Syllables produced by males that have lost one leg (presumably in an encounter with a predator), however, contain distinct gaps. These songs are rejected by females, suggesting that male sound production is under intense sexual selection. Peripherally, gap detection is accounted for by the synchronous, precisely timed action potentials of groups of receptor neurons in response to a sharp sound onset following silence. Centrally, an interneuron has been identified that is a good candidate for detecting, and reporting to the brain, the 'gappiness' of syllables. The neuron responds well throughout the duration of gap-free syllables, but its response falls off with increasingly wide gaps, paralleling precisely the preference functions of females. The mechanism responsible for the neuron's selectivity is synaptic inhibition, which is evoked by the onset of sound following silence. For gap-free syllables, only a single inhibitory synaptic potential occurs at the onset of each syllable, but for syllables that contain gaps, inhibition is renewed with each gap. The summation of these successive inhibitory potentials suppresses the neuron's response to syllables that contain gaps.

Summary

Both peripheral and central mechanisms participate in the identification of communication signals. For sensory systems that are dedicated to communication, peripheral processing serves mainly to restrict activity to behaviorally relevant signals through matched filtering. Peripheral tuning is further enhanced by central processing. Other central functions include analysis of temporal patterns, integration of separate signal components, discrimination among different signals or components, and evaluation of signal quality.

See also: Agonistic and Affiliative Signals: Resolutions of Conflict; Localizing Signal Sources; Pheromones and other Chemical Communication in Animals; Signal Production and Amplification in Birds; Signal Transmission in Natural Environments; Signal Design Rules in Animal Communication; Vocal Communication in Birds.

Further Reading

Bradbury JW and Vehrencamp SL (1998) *Principles of Animal Communication*. Sunderland, MA: Sinauer.

Bullock TH, Hopkins CD, Popper AN, and Fay RR (eds.) (2005) *Springer Handbook of Auditory Research: Electroreception*. Berlin: Springer.

Capranica RR (1966) Vocal response of the bullfrog to natural and synthetic mating calls. *Journal of the Acoustical Society of America* 40: 1131–1139.

Gerhardt HC and Huber F (2002) *Acoustic Communication in Insects and Anurans: Common Problems and Diverse Solutions*. Chicago: University of Chicago Press.

Hennig RM, Franz A, and Stumpner A (2004) Processing of auditory information in insects. *Microscopic Research and Technique* 63: 351–374.

Hall JC (1999) GABAergic inhibition shapes frequency tuning and modifies response properties in the auditory midbrain of the leopard frog. *Journal of Comparative Physiology A*. 185: 479–491.

Luo M and Katz LC (2004) Encoding pheromonal signals in the mammalian vomeronasal system. *Current Opinion in Neurobiology* 14: 428–434.

Nocke H (1972) Physiological aspects of sound communication in crickets (*Gryllus campestris* L.). *Journal of Comparative Physiology* 80: 141–162.

Pollack GS (2001) Analysis of temporal patterns of communication signals. *Current Opinion in Neurobiology* 11: 734–738.

Simoncelli EP and Olshausen BA (2001) Natural image statistics and neural representation. *Annual Review of Neuroscience* 24: 1193–1216.

Simmons A, Popper AN, and Fay RR (eds.) (2003) *Springer Handbook of Auditory Research: Acoustic Communication*. Berlin: Springer.

VanderMeer RK, Breed MD, Espelie KE, and Winston ML (1998) *Pheromone Communication in Social Insects: Ants, Wasps, Bees, and Termites*. Boulder, CO: Westview.

Wyatt TD (2003) *Pheromones and Animal Behaviour*. Cambridge, UK: Cambridge University Press.

Relevant Websites

http://www.birds.cornell.edu – Cornell Lab of Ornithology, Macaulay library of sounds.

http://buzz.ifas.ufl.edu – Singing Insects of North America.

http://acp.eugraph.com – The Animal Communication Project.

Signal Production and Amplification in Birds

R A Suthers, Indiana University, Bloomington, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

Avian Vocal Signals

Calls and Songs

Birds are a highly mobile group of vertebrates that depend on vocal signals for communication over long distances, through dense vegetation, and in dim light. Their vocalizations include calls and songs. Calls are typically short, acoustically simple sounds that have a social or survival function, such as maintaining contact between individuals, establishing dominance, or warning of the approach of a predator. They are generally used over short distances and are produced by both sexes throughout the year. Nearly all calls are thought to be innate. Songs, on the other hand, usually have a more complex, often musical, acoustic structure than calls. They are usually, but not always, sung only by the male and primarily during the breeding season, when they are broadcast over a large area to attract a mate or defend a territory.

Vocal Learners and Nonlearners

Oscine songbirds (i.e., members of the oscine suborder of the Passeriformes), parrots, and some hummingbirds learn their species song with the aid of sensory feedback. Only a few vertebrate groups besides humans are known to be capable of vocal learning. Birds which possess this ability thus provide a valuable animal model in which to study this phenomenon. Just as human children must be exposed to speech in order to acquire normal language skills as adults, avian vocal learners must hear conspecific song as juveniles and require auditory feedback of their own 'practice' vocalizations in order to develop normal adult song. In some birds, vocal learning is restricted to critical periods during the first year of life; other species undergo seasonal periods of vocal motor plasticity throughout their lives when they add or replace song syllables. The brain of the oscine songbirds contains a well-developed network of song control nuclei devoted to vocal learning and song production. Analogous areas are present in the brains of parrots and hummingbirds, but no similar system has been found in other avian taxa. The songs of these nonvocal learners appear to be innate and do not depend on auditory experience.

Structure of the Avian Vocal Organ

Songbirds

Birds have evolved a unique vocal organ, the syrinx, which is located in the interclavicular air sac at or near the point where the two primary bronchi converge to form the trachea. The syrinx consists of a framework of modified tracheal and/or bronchial cartilages that support the syringeal membranes or fleshy labia, depending on the species, which vibrate in respiratory airflow to generate sound.

Vocal organ anatomy varies considerably between different avian taxa. In the oscine songbirds, the syrinx is formed from the caudal end of the trachea and the cranial end of each bronchus (Figure 1). Each bronchial portion contains a thin medial tympaniform membrane adjacent to a pair of fleshy labia. It was long assumed that the tympaniform membranes were the vibratory sound source, but ablation of the membranes and endoscopic observations of syringeal biomechanics by Goller and Larsen in anesthetized birds during sounds elicited by forced airflow or brain stimulation indicate that sound is generated by oscillation of the labia at the end of each bronchus. Songbirds thus have a separate sound source on each side of their syrinx (Figure 2). This bipartite vocal organ is controlled by several bilaterally arranged pairs of syringeal muscles innervated by the ipsilateral tracheosyringeal branch of the hypoglossal nerve and is under primary control of the song control nuclei on the same side of the brain.

Non-Songbirds

Many non-songbirds also have a similarly constructed bipartite syrinx at the confluence of the bronchi and trachea, but they have fewer syringeal muscles and the tympaniform membranes may replace the labia as the vibratory sound source. In some species, the syrinx is controlled only by extrinsic muscles in which only one end is attached to the syrinx.

In other non-songbirds, the syrinx is located either in the trachea or in the bronchi. The syrinx of doves and parrots, for example, is in the caudal portion of the trachea (Figure 3(a)). Sound is generated by the vibration of a pair of lateral tympaniform membranes situated on the wall of the trachea. A still different anatomical arrangement is present in many nocturnal birds, including owls, nightjars, the echolocating oilbird (*Steatornis caripensis*), and some cuckoos. In these birds, half of the syrinx is located in each

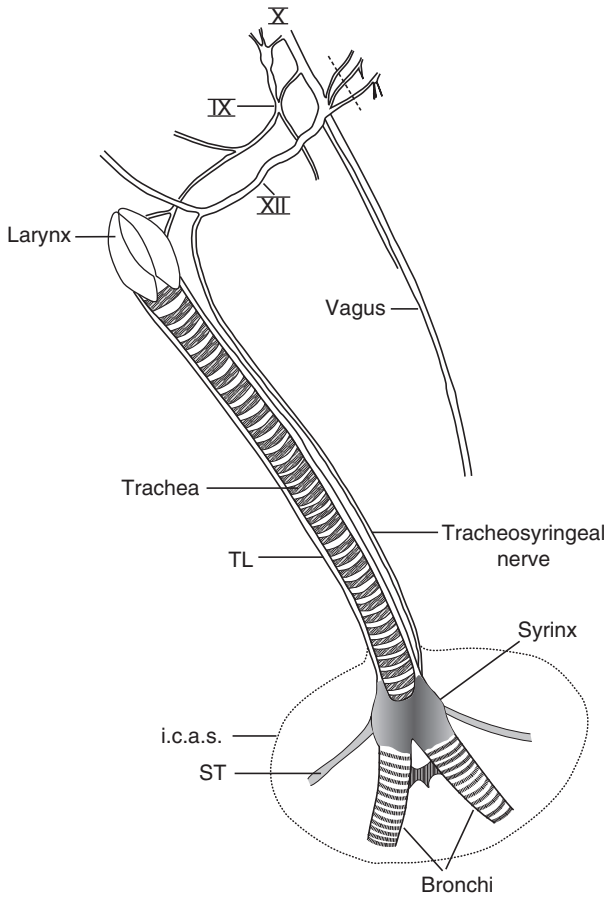

Figure 1 The songbird's vocal tract showing innervation for one side. i.c.a.s., interclavicular air sac, which is the only unpaired air sac and is located in the midline surrounding syrinx; ST, sternotrachealis muscle; TL, tracheolateralis muscle; X, vagus nerve; XI, accessory nerve; XII, hypoglossal nerve. Adapted from Nottebohm F (1971) Neural lateralization of vocal control in a passerine bird: I. Song. *Journal of Experimental Zoology* 177: 229–262, with permission.

primary bronchus and sound is produced by oscillation of a tympaniform membrane on the medial side of each bronchus (Figure 3(b)).

Respiratory and Syringeal Dynamics of Song Production

Birdsong, like human speech, requires the precisely coordinated activity of many muscles in both the respiratory and vocal systems. Together, these muscle groups determine the temporal and spectral properties of the sound that enters the suprasyringeal vocal tract, on its way to the beak and radiation into the external environment as a vocal signal.

Vocalization begins with the contraction of respiratory muscles that causes air to flow through the syrinx. As air flows across the labia (or syringeal membranes in some species) it causes them to oscillate, generating sound. Birds do not have a muscular diaphragm like mammals do, so respiratory airflow is determined by the activity of thoracic and abdominal muscles that expand or compress several pairs of thin-walled air sacs in the thorax and abdomen. These air sacs function as bellows connected to the lungs and bronchi, through which they pump air.

A singing bird must adapt both the amplitude and the timing of its respiratory pattern to meet its requirements for vocal communication, as well as those for pulmonary gas exchange. During song, respiratory pressures are increased approximately ten-fold compared to those of quiet, silent respiration. Phonation is accompanied by high-amplitude bursts of electrical activity in expiratory muscles that alternate with similar bursts in inspiratory muscles (Figure 4(a)). The temporal pattern of respiration is also modified and determines the tempo of the song.

Song Tempo and Duration

Nearly all song is produced during expiratory airflow. The overall temporal pattern of song is thus determined by the timing and duration of each expiration. Within the expiratory phase, the timing of phonation can be further fine-tuned by the muscles that adduct the sound oscillators into the syringeal lumen. In addition to generating sound, these structures function as a pneumatic valve that gates phonation and can prevent or interrupt sound production during expiration by closing the syringeal lumen, thus blocking airflow despite a positive subsyringeal pressure.

Birds face a trade-off between achieving a high syllable repetition rate and the length of time they can sustain the phrase or song. At relatively low repetition rates (<30 or <16 syllables per second for canaries and cardinals, respectively), the volume of air required to produce each syllable is replaced by a brief 'minibreath' immediately after the syllable. As repetition rate increases, it reaches a point where there is not enough time between syllables for a minibreath. Syllable repetition rates higher than this are produced by pulsatile expiration (Figure 4(b)). This motor pattern maintains a positive subsyringeal pressure while keeping one side of the syrinx closed to conserve respiratory air for phonation. The trill is generated by rapid repetitive opening of the other side. Each escaping puff of air generates a syllable, but the duration of the trill is limited since air is not being replaced.

In songbirds this syringeal valve is formed by the medial and lateral labia at the cranial end of each bronchus (Figure 2(a)). Its aperture is controlled by the dorsal syringeal muscles on the same side of the syrinx (Figure 2(b)). When the dorsal muscles relax and other antagonistic abductor muscles contract, the

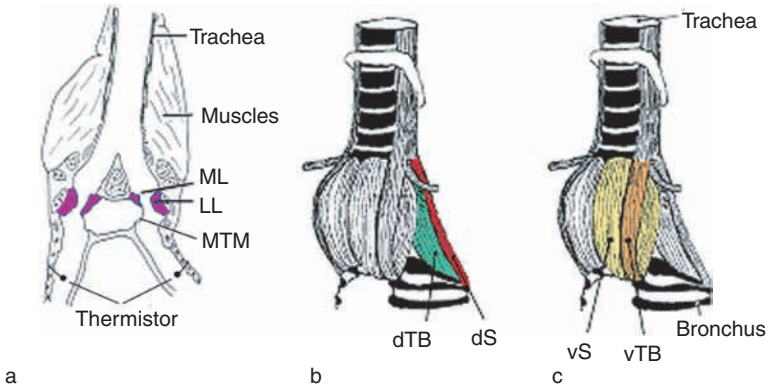

Figure 2 Songbird syrinx showing its bipartite structure and multiple muscles. (a) Frontal section through syrinx of brown thrasher showing separate labial oscillators at the cranial end of each bronchus. A heated microbead thermistor in each bronchus monitored the rate of airflow through each side of the syrinx during spontaneous song. (b) External ventrolateral view of syrinx indicating the dorsal syringeal muscles primarily responsible for moving the labia into the syringeal lumen where they vibrate in the airflow or, if fully adducted, silence the ipsilateral side of the syrinx by blocking airflow. (c) Ventral syringeal muscles that are primarily responsible for varying fundamental frequency by changing tension on the labia. dS, dorsal syringeal muscle; dTB, dorsal tracheobronchial muscle; LL, lateral labium; ML, medial labium; MTM, medial tympaniform membrane; vS, ventral syringeal muscle; vTB, ventral tracheobroncial muscle. (a) Adapted from Suthers RA (1990) Contributions to birdsong from the left and right sides of the intact syrinx. *Nature* 347: 473–477, with permission. (b, c) Adapted from Goller F and Suthers RA (1996) Role of syringeal muscles in controlling the phonology of bird song. *Journal of Neurophysiology* 76: 287–300, with permission.

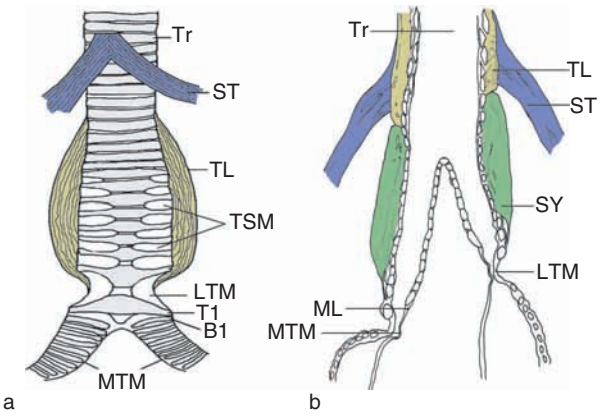

Figure 3 Syringeal diversity. (a) Tracheal syrinx of collared dove (ventral view). (b) Bronchial syrinx of oilbird (frontal section). B1, first bronchial semiring; LTM, lateral tympaniform membrane; ML, medial labium; MTM, medial tympaniform membrane; ST, sternotrachealis muscle; SY, syringealis muscle; TL, tracheolateralis muscle; TSM, tracheosyrngeal membrane. (a) Adapted from Ballintijn MR and ten Cate C (1997) Sex differences in the vocalizations and syrinx of the collard dove (*Streptopelia decaocto*). *Auk* 114: 22–39, with permission. (b) Adapted from Suthers RA and Hector DH (1985) The physiology of vocalization by the echolocating oilbird, *Steatornis caripensis*. *Journal of Comparative Physiology A* 156: 243–266, with permission.

labia are fully abducted out of the air stream to provide a low-resistance pathway during inspiration and silent respiration. During phonation, the dorsal muscles are active, causing the labia to move toward each other into the syringeal lumen, where they form a slit and vibrate in airflow, generating sound. Fine gradations of dorsal muscle action during phonation can modulate the amplitude and spectral properties of the sound. Further contraction of the dorsal syringeal muscles fully adducts the labia, which meet in the middle of the lumen and prevent phonation on that side of the syrinx by blocking airflow.

In many non-songbirds, including doves, the syringeal valve is controlled by tracheal muscles. If the song contains trills with high-note repetition rates, the muscles controlling the syringeal valve must be able to operate very rapidly. The coo of the ring dove includes a trilled portion with note repetition rates up to $30\,s^{-1}$. Dove tracheal muscles controlling the syringeal valve are specialized for high-speed contractions and have half-twitch times of approximately 9 or 10 ms. It is likely that syringeal muscles serving a similar function in songbirds are also superfast.

The production of stereotyped vocal signals requires precise coordination between respiratory and syringeal motor patterns. By opening and closing the syrinx, motor commands to the dorsal syringeal muscles can determine which side(s) is producing sound at any instant and whether one or several notes are produced within a single expiration. A neural pathway believed to be important in coordinating respiratory and syringeal motor patterns during learned song involves the robust nucleus of the arcopallium, a song control nucleus in the songbird's telencephalon, which projects to respiratory premotor nuclei as well as to syringeal motor neurons.

Control of Fundamental Frequency

The ventral syringeal muscles of songbirds have an important role in determining the fundamental frequency

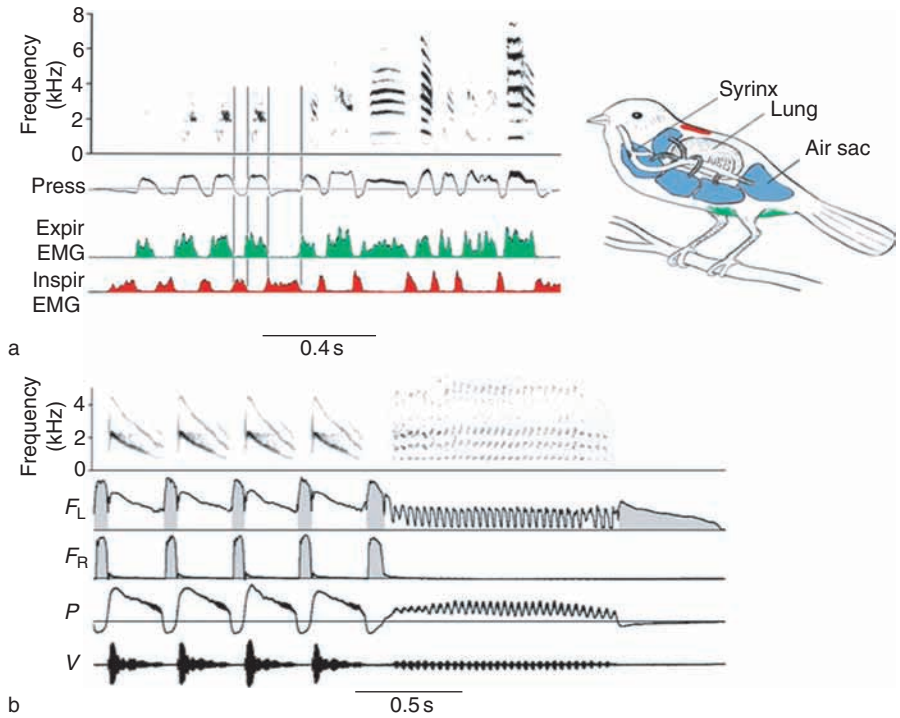

Figure 4 Respiratory patterns during song. (a) Segment of zebra finch song showing air sac pressure, together with alternating EMG patterns of the envelope of rectified abdominal expiratory muscle (green) and thoracic inspiratory scalenus muscle (red). Schematic drawing of bird indicates approximate location of air sacs (blue). (b) Portion of a song by the northern cardinal. Envelope of rectified EMG is shown. The first four syllables are produced by airflow through the left side of syrinx. Despite positive air sac pressure during the syllables, there is no airflow through the right side, indicating it is closed by full adduction of labia. Shaded airflow between syllables indicates minibreaths. The fourth syllable is followed by a pulsatile trill in which a high syllable repetition rate is achieved by maintaining a continuous positive air sac pressure and releasing repetitive puffs of air through the left side of the syrinx while keeping the right side closed. Top shows the sound spectrogram of song. F_L and F_R, rate of airflow through the left and right side of syrinx, respectively. Shaded flow is inspiratory. Horizontal line equals zero airflow. *P*, air sac pressure driving airflow. Horizontal line is ambient external pressure. Negative pressure accompanies inspiratory flow; *V*, time waveform of vocalizations. (a) Reproduced from Wild J, Goller F, Suthers R (1998) Inspiratory muscle activity during bird song. *Journal of Neurobiology* 336: 441–453, with permission from Blackwell. (b) and bird adapted from Suthers RA (2004) How birds sing and why it matters. In: Marler P and Slabbekoorn H (eds.) *Nature*'s *Music. The Science of Birdsong*, pp. 272–295. New York: Elsevier, with permission.

of vocalizations (Figure 2(c)). Unlike the human vocal fold, the avian labia do not contain intrinsic muscles. The fundamental frequency at which they oscillate can be varied by the ventral syringeal muscles which increase the tension on the labia when they contract. The electrical activity of these muscles is positively correlated with changes in the song's fundamental frequency.

The complexity of song in many oscine songbirds is at least in part related to the fact that they have multiple pairs of syringeal muscles. This increases the potential for particular muscles to be functionally specialized and facilitates the ability to independently control various acoustic and aerodynamic parameters. Non-songbirds with few syringeal muscles may face a challenge in controlling fundamental frequency independent of subsyringeal pressure, for example.

Ring doves (*Streptopelia risoria*) provide an interesting case in which perisyringeal pressure controls sound frequency independent of subsyringeal pressure. The tracheal syrinx of doves has only two muscles. Tension in these muscles may influence the range of frequencies over which the syringeal membranes can oscillate, but the fine control of fundamental frequency during the course of the coo is precisely correlated with pressure fluctuations in the interclavicular air sac that surrounds the syrinx. This interclavicular pressure impinges directly on the outer surface of the syringeal membranes, presumably controlling their frequency of oscillation by varying their tension as they bulge into the tracheal lumen. Interestingly, the subsyringeal pressure measured in the thoracic air sac varies independent of that in the interclavicular sac and is not correlated with sound frequency.

Sensory Feedback in Adult Song

Auditory Feedback

These complicated vocal motor patterns of songbirds are dependent on sensory feedback, even in adult,

crystallized song. Auditory feedback, in addition to its essential role in song learning, continues to be important in the maintenance of adult song. The song of deafened adults gradually deteriorates over a time course that varies from days to months, depending on the species. One of the more dramatic demonstrations of the importance of auditory feedback after song crystallization comes from the work of Leonardo and Konishi, who placed a microphone and speaker in the cages of adult zebra finches (*Taeniopygia guttata*). Whenever the bird sang, it heard its song reproduced through the speaker after a delay of 100 ms, disrupting the normal sensorimotor integration. The normal crystallized songs of these zebra finches began to deteriorate after a few weeks. Abnormal features of this 'decrystallized' song included stuttering, distortion or deletion of syllables, or the addition of new syllables. If the delayed auditory feedback was discontinued, the song gradually returned to its original form over a few months, indicating that memory of the original song persists.

Somatosensory Feedback

Respiratory and syringeal motor patterns of adult cardinals are also modulated by somatosensory feedback. If the subsyringeal pressure is briefly increased by injecting a small amount of air into an air sac during a syllable, there is a real-time (latency of approximately 50 ms) compensatory decrease in the contraction of expiratory muscles (Figure 5). There are also compensatory responses by syringeal muscles. These adjustments in muscle activity minimize the increase in syringeal airflow that would otherwise accompany the increased air sac pressure and distort the vocalization.

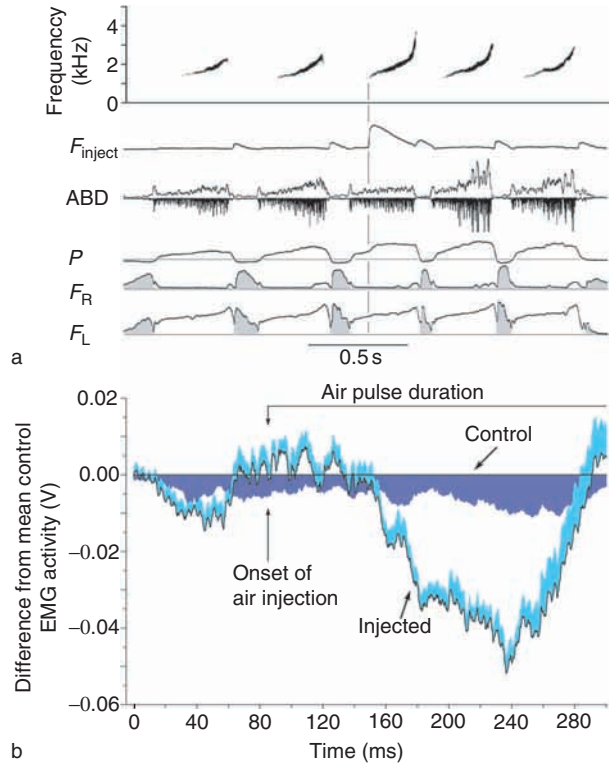

Figure 5 Real-time modulation of song respiratory motor pattern by somatosensory feedback. (a) Northern cardinal song. A fraction of a milliliter of air was injected into the thoracic air sac early in the third syllable (vertical line). (b) Relative mean amplitudes of abdominal expiratory muscle EMG during syllables with air injection compared to the same syllable without air injection in deafened bird. Blue shading indicates one SE for control (dark blue) and experimental (light blue). Reduction in amplitude of expiratory muscle EMG following air injection stabilizes syringeal airflow and minimizes the acoustic effect of unpredictable pressure change. F_{inject}, thermistor output showing rate of airflow through cannula into air sac. ABD, EMG of abdominal expiratory muscle (rectified below and integrated envelope above); F_L and F_R, rate of airflow through the left and right side of syrinx, respectively; *P*, air sac pressure driving airflow. Adapted from Suthers RA, Goller F, and Wild JM (2002) Somatosensory feedback modulates the respiratory motor program of crystallized birdsong. *Proceeding of the National Academy of Sciences of the United States of America* 99: 5680–5685, with permission.

Song Lateralization

Lateralized neural control of behavior was first described by Broca in 1861 with regard to human speech. Birdsong was one of the first nonhuman examples of functional lateralization in the brain. Early studies on songbirds by Nottebohm and others assessed lateralization by comparing song deficits following unilateral denervation of the syrinx. With rare exceptions, species tested in these experiments all showed a larger song deficit when the left side of the syrinx was denervated, limiting song production to the right side, than when the right side was denervated. This left dominance is most pronounced in a Wasserslager strain of Belgium canaries (*Serinus canaria*). Paralysis of the left side of the Wasserslager's syrinx eliminates approximately 90% of the syllables in the bird's song repertoire, whereas only approximately 10% is lost after paralysis of the right side. The extreme left side dominance of song production by the Wasserslager canary is unique among the birds that have been studied, including even a conspecific domestic strain of canary.

Insight into the neural basis of song lateralization was obtained from recordings of syringeal muscle electromyograms in brown thrashers. These experiments showed that lateralization is based on the motor commands to the dorsal syringeal muscles responsible for gating airflow through each side of the syrinx. The action of these muscles determines which side of the syrinx produces sound. The ventral syringeal muscles that primarily control the phonetic structure of

syllables are active on both sides of the syrinx regardless of which side is phonating. The song control nuclei above the hypoglossal motor nucleus do not exhibit anatomical or functional laterality. These findings argue against the hypothesis that song lateralization evolved in order to free brain space for other tasks such as memory storage of songs or in order to provide a single command center for both sides of the song control system. Respiratory motor patterns are not lateralized during song.

Patterns of Song Lateralization

The most important contribution that the bipartite syrinx makes to avian vocal communication is not the possibility of unilateral dominance in song production but, rather, the additional degrees of freedom for signal diversity and complexity provided by two independently controlled sound sources. Songbirds have exploited this opportunity by evolving different acoustic specializations for each side of the syrinx and using the two sides in particular ways that appear optimized for producing the species-specific features of their song.

Independent Voices from the Bipartite Syrinx

By singing with both sides of the syrinx simultaneously, syllables containing two harmonically unrelated fundamental frequencies can be generated. Such two-voice syllables are common, for example, in the repertoires of brown thrashers (*Toxostoma rufum*) and catbirds (*Dumetella carolinensis*) (family Mimidae) (Figure 6(a)). The separate 'voices' in such syllables may be frequency modulated in different directions and often have a distinctive dissonance to our ears. Since the sounds from the two sides of the syrinx are combined in the trachea they can, in certain circumstances, generate a vocalization containing rapid amplitude modulation at a modulation frequency higher than can be produced by cyclical contractions of the syringeal or respiratory muscles.

Surprisingly little is known about the behavioral significance in vocal communication of this two-voice phenomenon. Ironically, the exception to this statement comes from studies on a non-songbird, the penguin, by Aubin and colleagues. Emperor and king penguins (*Aptenodytes forsteri* and *A. patagonicus*) are faced with the daunting task of finding and identifying mates or chicks in a noisy, tightly packed colony of thousands of birds. Since these two species carry the egg on their feet, there is no nest site to which to return. Emperor and king penguins produce display calls in which each side of the syrinx simultaneously generates a different fundamental frequency (Figure 7). The fundamentals of these two-voice calls interact to produce individually identifiable amplitude-modulated beat notes that propagate well through the crowded colony.

The bipartite syrinx of songbirds also facilitates increased temporal complexity in song. This is illustrated in the songs of the brown-headed cowbird (*Molothrus ater*). Each song begins with two or three clusters of notes produced in rapid succession at different frequencies and with little or no silent interval between them (Figure 6(b)). Recordings of respiratory airflow, pressure, and bronchial sounds show that within each note cluster the first note, which also has the lowest fundamental frequency, is produced on the left side of the syrinx and the following notes originate alternately from opposite sides of the syrinx. This motor pattern allows abrupt frequency steps between successive notes despite the absence of a silent interval between them. This is possible because while one note is being produced, a motor program to the contralateral side of the syrinx is presetting muscle tensions needed for the next note. Each cowbird song ends with a high-frequency, loud 'final whistle' that is produced on the right side of the syrinx.

Lateralized Functional Specialization of the Syrinx

In all songbirds studied, each side of the syrinx has a partially different frequency range. Frequencies below approximately 3.5–4 kHz are generally produced on the left side and higher frequencies are produced on the right, although there is typically an extensive band of midrange frequencies that can be generated on either side. This lateralized specialization for different frequencies is particularly pronounced in the northern cardinal. Much northern cardinal (*Cardinalis cardinalis*) song is characterized by FM sweeps that often extend over two or more octaves. To the human listener, these whistled sweeps sound continuous, but recordings from the syrinx during song show that fundamental frequencies below approximately 3 or 4 kHz are sung by the left syrinx, whereas higher frequencies are sung by the right syrinx (Figure 6(c)). When the frequency of a syllable extends across this frequency divide, the bird switches phonation in midnote from one side to the other, depending on the direction of the FM. Usually, this lateral switch is so expertly executed that there is no perceptible interruption in the sound or perturbation of the frequency.

Acoustic Role of the Vocal Tract

A Variable Resonance Filter

The sound that is produced by the syrinx is modified by resonances of the suprasyringeal vocal tract which

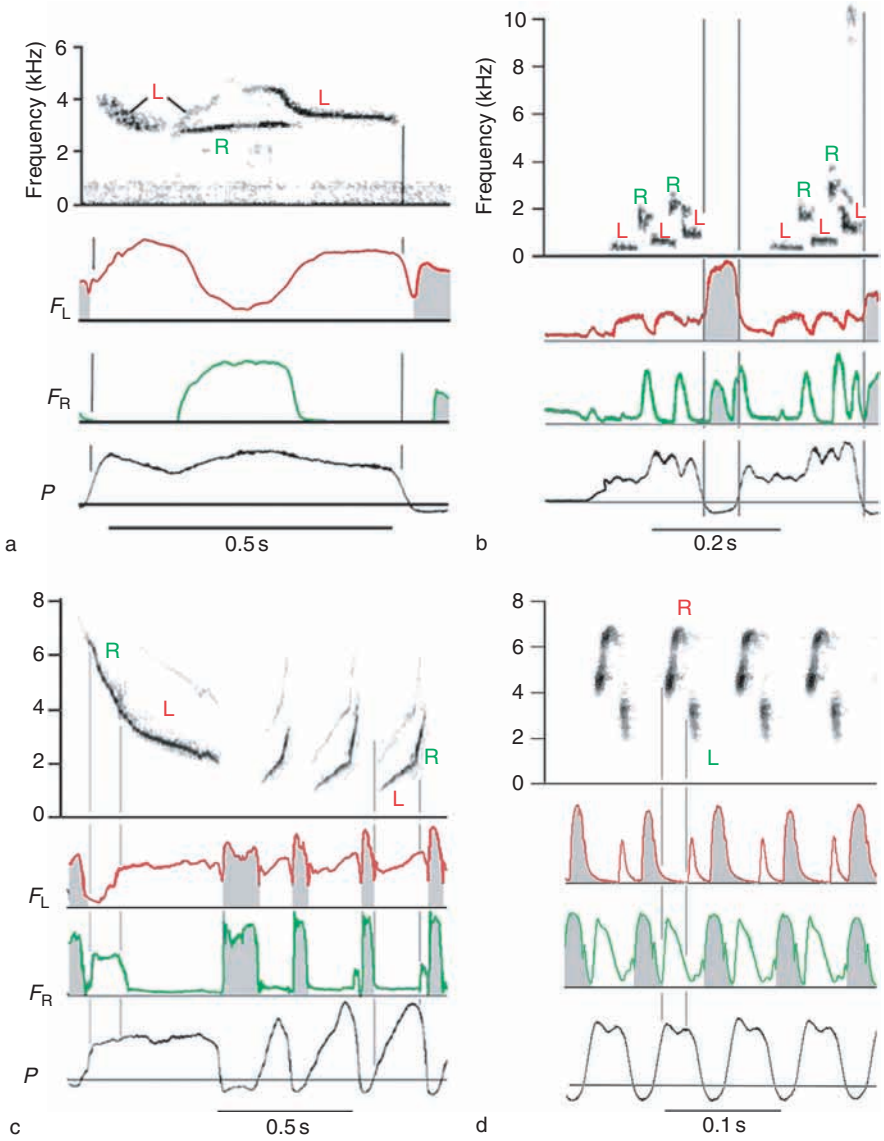

Figure 6 Species differences in the pattern of syringeal lateralization. (a) Independently modulated two-voice syllable of brown thrasher. (b) Two note clusters from song of brown-headed cowbird. Abrupt frequency steps without a silent interval between notes are achieved by alternating note production between opposite sides of the syrinx. Final whistle is not shown. (c) Frequency modulated sweeps from northern cardinal song with smooth transition of phonation from one side of syrinx to the other in the middle of the sweep. (d) Part of a 'sexy' phrase from a domestic canary, showing the high repetition rate, wide bandwidth, and two-note structure of the syllables. In all species studied, the right side of syrinx tends to produce higher frequencies than the left. F_L and F_R, rate of airflow through the left and right side of syrinx, respectively; *P*, air sac pressure driving airflow. (a–c) Adapted from Suthers RA and Zollinger SA (2004) Producing song: The vocal apparatus. *Annals of the New York Academy of Sciences* 1016: 109–129, with permission.

allow some frequencies to pass through the vocal tract and attenuate other frequencies. The resonance frequencies are expressed in the emitted vocalization as bands of relatively high acoustic energy called formants.

Evidence that the vocal tract acts as a resonance filter of syringeal sound was obtained through experiments with songbirds singing in a light gas mixture consisting of 80% helium and 20% oxygen. These experiments indicated that the vocal tract resonance in air is tuned to the fundamental frequency of the song. Because the speed of sound is faster in heliox than in air, the resonant frequency in heliox shifts upward nearly an octave, increasing energy in the normally suppressed second harmonic and reducing the energy in the fundamental of the emitted vocalization. Light gas experiments also suggested that songbirds adjust the tuning of their vocal tract resonance to match changes in the fundamental frequency, thus increasing the tonality, or whistlelike, quality of the vocalization by reducing the relative amplitude of the higher harmonics generated by the syrinx.

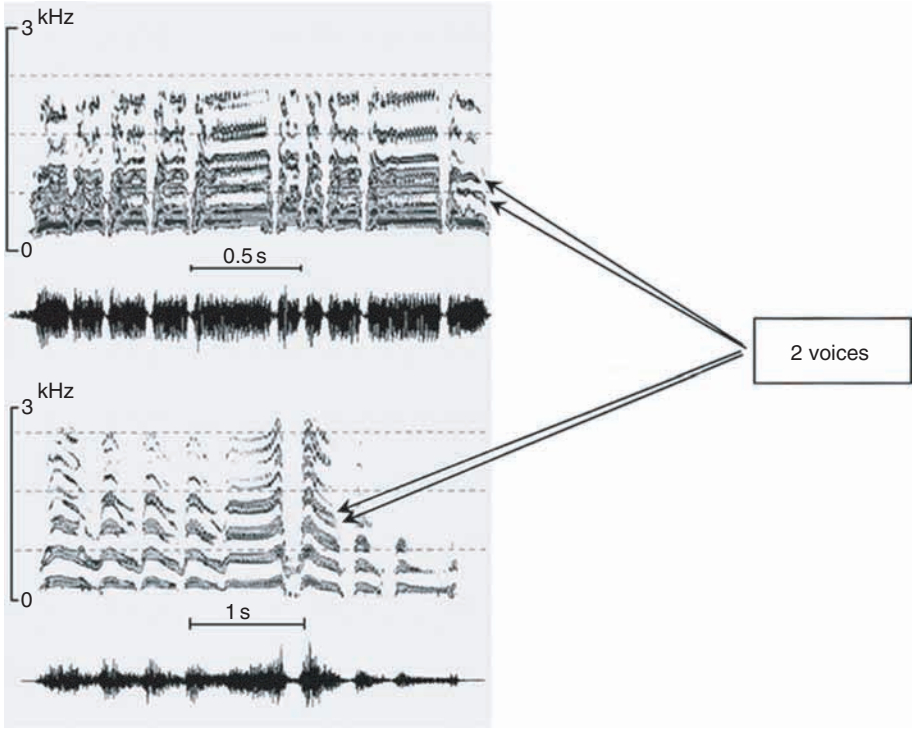

Figure 7 Spectrograms of two-voice phenomenon represented as simultaneous harmonically unrelated frequencies in calls of the emperor penguin (top) and king penguin (bottom). Reproduced from Aubin T (2004) Penguins and their noisy world. *Anais da Academia Brasileira de Ciencias* 76: 279–283, with permission from Anais da Academia Brasileira de Ciências.

In order to vary the vocal tract's resonance frequency, the bird must be able to alter its shape. Which part of the vocal tract is responsible for this, and the mechanism by which shape is regulated, is a topic of controversy and current research. Primary candidates include the trachea, the oropharyngeal cavity including the tongue, and the beak.

The Trachea

The tracheal resonance can be approximated by modeling it acoustically as a tube closed at the syrinx and open at the glottis. Songbirds can only make very small changes in the length of their trachea, however, making it an unlikely candidate for the tracking filter. Nevertheless, the trachea appears to have evolved a special role in vocal communication by some non-songbirds, such as cranes. The trachea of these species is coiled inside the body, greatly increasing its length (Figure 8). Tecumseh Fitch showed that the amount of tracheal elongation is inversely correlated with the distance between formants in the call. He hypothesized that birds use formant spacing to assess the sender's body size and that tracheal elongation may have evolved as a means of exaggerating the sender's size.

The Oropharynx and Esophagus

The acoustic role of the oropharyngeal cavity, which lies between the glottis and the beak, has been studied using x-ray cinematography. Data from northern cardinals show that the oropharyngeal cavity undergoes large cyclical, song-related changes in its shape and volume due to movement of the hyoid apparatus and larynx. The volume of this cavity varies inversely with the fundamental frequency of the syringeal sound. At low fundamental frequencies the oropharyngeal cavity is further enlarged by expansion of the cranial end of the esophagus (Figure 9). A computational acoustic model of this cavity predicts that the primary resonance frequency of the oropharyngeal–esophageal cavity closely tracks the song's fundamental frequency, reinforcing the fundamental while attenuating the higher harmonics. It appears to be an important part of the songbird's tracking filter.

The Tongue

The tongue lies within the oropharyngeal cavity but little is known about its acoustic role in avian vocalizations. The base of the tongue is attached to the hyoid apparatus and moves with it and the larynx. Parrots have a thick muscular tongue, in contrast to the relatively much thinner, smaller tongue of songbirds. In the monk parakeet (*Myiopsitta monachus*), small changes in the position of the tongue can alter the formant frequencies, and tongue movements appear to contribute to the articulation of vowel sounds when gray parrots mimic speech. Although

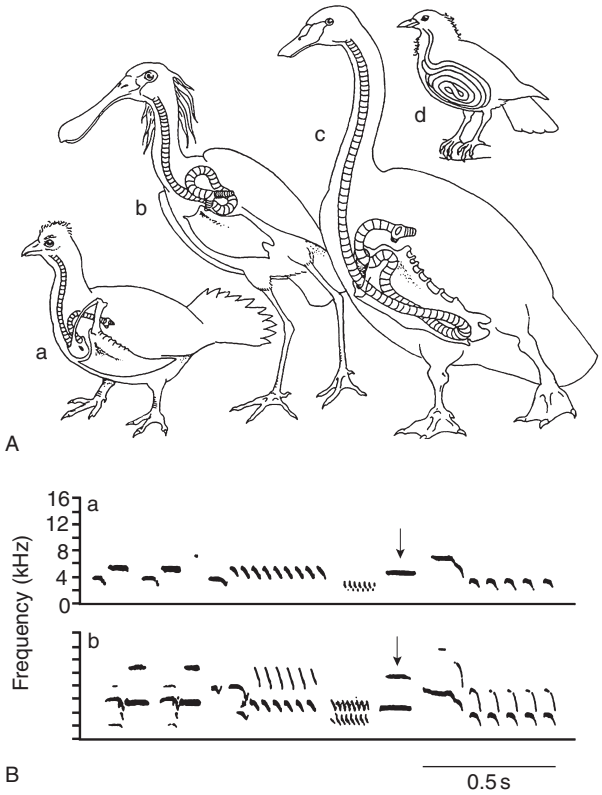

Figure 8 (a) Example of four types of tracheal elongation that may have evolved as a way of exaggerating body size by altering the formant frequency dispersion. Species shown are (a) intraclavicular coiling, crested guinea fowl *Guttera edouardii*; (b) intrathoracic coiling, European spoonbill *Platalea leucorodia*; (c) intrasternal coiling, trumpeter swan *Cygnus buccinators*; and (d) subdermal coiling, trumpet manucode *Manucodia keraudrenii*. (b) Segment of a song sparrow song sung in air (top) and in heliox (bottom). Second harmonic is attenuated compared to fundamental in air but is relatively prominent in heliox due to an upward shift in the resonant frequency in light gas. The arrows indicate a syllable in which the second harmonic, which is suppressed in air but prominent in helium, is at the same frequency as the fundamental of the following syllable, suggesting a change in the tuning of the vocal tract filter. (A) From Fitch WT (1999) Acoustic exaggeration of size in birds by tracheal elongation: Comparative and theoretical analyses. *Journal of Zoology (London)* 248: 31–48, with permission. (B) Adapted from Nowicki S (1987) Vocal tract resonances in oscine bird sound production: Evidence from bird songs in a helium atmosphere. *Nature* 325: 53–55, with permission.

some songbirds, such as the hill mynah (*Gracula religiosa*) and starling (*Sturnus vulgaris*), can also mimic speech sounds, the role of the passerine tongue in vocal communication is unknown.

The Beak

The beak is the only part of the vocal tract that can be readily observed during song. A number of experiments have shown that the beak opening, or gape, also exhibits song-related changes, being large when the fundamental frequency is high and relatively small when the frequency is low. It has been widely assumed that these song-related beak movements provide the mechanism by which birds adjust their vocal tract resonance to support the fundamental frequency. According to this hypothesis, changes in beak gape alter the resonance frequency of the vocal tract by varying its end correction or effective length. The emission of high frequencies is facilitated by a large beak gape. Conversely, by nearly closing its beak, a bird can suppress the higher harmonics of a low-frequency fundamental. The relative contribution of the oropharyngeal–esophageal cavity versus beak gape to the vocal tract's resonance filter is controversial. Although there is general agreement that beak gape is correlated with changes in the filter properties of the vocal tract, the physical basis of this relationship is unclear. Some data suggest that the beak *per se* has a limited acoustic effect, raising the possibility that related changes in the upper vocal tract that are linked to beak gape may be acoustically important.

Production Constraints

Two of the important functions of birdsong are territorial defense and mate attraction. Exactly how song contributes toward accomplishing these essential behaviors is a subject of current research and beyond the scope of this article. Advances in understanding the mechanisms of singing underline the need for precise and often very rapid coordination between the two sides of the syrinx, the respiratory rhythm and movements of the suprasyringeal vocal tract. To what extent do the vocal gymnastics that are required to produce the distinctive acoustic features of much species-specific song impose performance limits? The fact that females of some species prefer males that sing in certain ways raises the possibility that by attending to how well a prospective mate sings difficult song elements, a female may be able to assess the relative quality of competing males.

Female Preference and Song

A growing body of evidence suggests that at least in some species, males that excel in rapid, precise coordination of vocal motor patterns have an advantage in attracting a mate. Behavioral experiments on brown-headed cowbirds by West and King suggest that the temporal precision of high- and low-frequency notes, which alternate between the two sides of the syrinx in the introductory note clusters (Figure 6 (b)), is an important factor in eliciting copulation solicitation displays from females. Researchers at the University of Paris showed that female domestic canaries are most likely to solicit copulation from

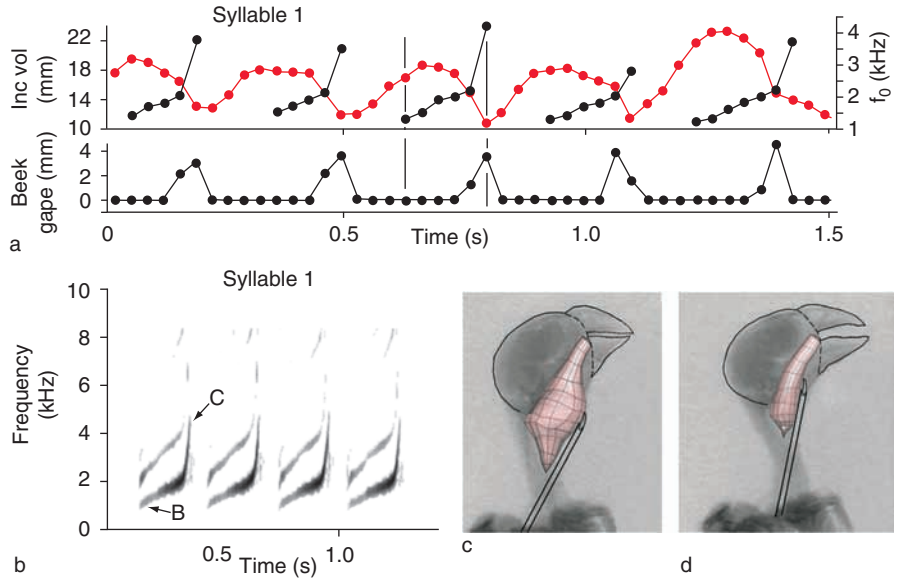

Figure 9 The northern cardinal tunes the major resonance of its vocal tract to the changing fundamental frequency of its syllables. (a) The volume of the oropharyngeal cavity (red curve) increases at low fundamental frequency of upward sweeping song syllable (black lines). At high fundamental frequency, cavity volume decreases and the beak opens. Each dot is a measurement from successive frames of an X-ray movie. Red dots indicate movement of larynx relative to anatomical reference. Upward movement of the red line corresponds to increased volume of the oropharyngeal cavity. (b) Four song syllables in which the fundamental sweeps upward from approximately 1 to approximately 5 kHz. (c, d) X-ray images at the beginning (1 kHz) and end (5 kHz) of syllable 1. (c) The pink polygon shows the expanded oropharyngeal cavity and cranial end of the esophagus at the beginning of syllable. (d) At the end of syllable, the esophagus has collapsed and the volume of the oropharynx is much reduced. A computational acoustic model based on the volume of the upper vocal tract estimated from a three-dimensional model indicates the primary resonance is approximately 1 kHz in (c) and 5 kHz in (d). The trachea forms the tube slanting upward from the syrinx (not in image) to the larynx which opens into the mid-ventral region of the oropharynx. Adapted from Riede T, Suthers RA, Fletcher NH, and Blevins W (2006) Songbirds turn their vocal tract to fundamental frequency of their song. *Proceedings of the National Academy of Sciences of the United States America* 103: 5543–5548, with permission.

males that include special 'sexy syllables' in their songs. Important features of these FM syllables include their high repetition rate, a relatively large frequency bandwidth, and the fact that each syllable is composed of two notes sung in rapid succession. In subsequent studies on song production in these canaries, it was found that the two successive notes of these sexy syllables are generated on opposite sides of the syrinx, requiring extremely rapid switching of phonation from side to side (Figure 6(d)). The ability to maintain a wide frequency bandwidth at high syllable repetition rates is also behaviorally important in swamp sparrows (*Zonotrichia georgiana*). Scientists at Duke University found that males who are able to maintain exceptionally wide frequency bandwidth at high trill repetition rates are more successful than other males in eliciting sexual displays from females.

Performance Constraints in Vocal Mimicry

Further insight into the likely influence of performance constraints on the evolution of birdsong comes from experiments on how the northern mockingbird, a vocal mimic, produces heterospecific songs. If there are multiple ways a songbird can achieve specific acoustic effects, one might expect that, by trial-and-error motor learning, the mockingbird would sometimes produce heterospecific song using different motor strategies than those used by the tutor species. Alternatively, if the mockingbird uses the same production mechanism as the tutor, it suggests there may be only one way to achieve the particular acoustic properties of the tutor song.

Juvenile northern mockingbirds (*Mimus polyglottus*) were tutored with the natural or computer synthesized songs of other species. When the mockingbirds sang these heterospecific songs during their first breeding season as adults, the motor pattern they used to produce the song was similar to that of the tutor (Figure 10). When a mockingbird's motor pattern deviated from that of the tutor, the acoustic accuracy of the copy also decreased. The fact that the vocal mimic uses the same pattern of lateralization as the tutor suggests that production constraints have played an important role in the evolution of song diversity.

See also: Agonistic and Affiliative Signals: Resolutions of Conflict; Animal Communication: Honesty and Deception;

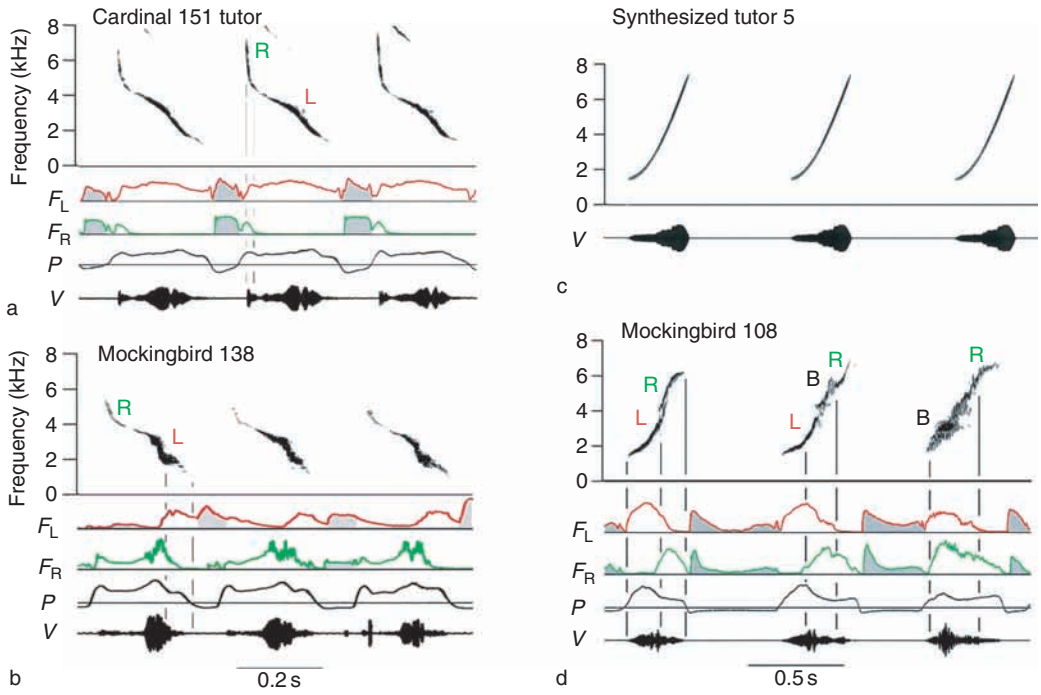

Figure 10 Mockingbird copies of extended frequency modulated sweeps. (a) Cardinal tutor sings fundamental above approximately 4 kHz with the right syrinx and switches to the left side of the syrinx for lower frequencies. (b) Mockingbird uses cardinal-like motor pattern to sing the same syllable, switching sides cleanly like the cardinal tutor. (c) Computer synthesized cardinal-like upward frequency sweep. (d) Mockingbird copies of (c). The first syllable is sung like a cardinal with a clean switch from the left to the right side of the syrinx near the middle of the sweep. The second and third syllables include portions when both sides of the syrinx are phonating simultaneously. As the mockingbird's motor pattern deviates from that of cardinal, the sound also becomes less cardinal-like. F_L and F_R, rate of airflow through the left and right side of syrinx, respectively. *P*, air sac pressure driving airflow; *V*, time waveform of vocalizations. From Zollinger SA and Suthers RA (2004) Motor mechanisms of a vocal mimic: implications for birdsong production. *Proceedings of the Royal Society of London, Series B* 271: 483–491, with permission.

Animals and the Biology of Music; Bird Song Systems: Evolution; Birdsong Learning: Evolutionary, Behavioral, and Hormonal Issues; Birdsong Learning; Birdsong: The Neurobiology of Avian Vocal Learning; Communication Networks and Eavesdropping in Animals; Endocrinology of Animal Communication: Behavioral; Localizing Signal Sources; Signal Identification: Peripheral and Central Mechanisms; Signal Transmission in Natural Environments; Vocal Communication in Birds.

Further Reading

Aubin T and Jouventin P (2002) How to vocally identify kin in a crowd: The penguin model. *Advances in the Study of Behavior* 31: 243–277.

Beckers GJL, Nelson BS, and Suthers RA (2004) Vocal-tract filtering by lingual articulation in a parrot. *Current Biology* 14: 1592–1597.

Beckers GJL, Suthers RA, and ten Cate C (2003) Pure-tone birdsong by resonance filtering of harmonic overtones. *Proceedings of the National Academy of Sciences of the United States of America* 100: 7372–7376.

Doupe AJ and Kuhl PK (1999) Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience* 22: 567–631.

Elemans CPH, Spierts ILY, Hendriks M, et al. (2006) Syringeal muscles fit the trill in ring doves (*Streptopelia risoria* L.). *Journal of Experimental Biology* 209: 965–977.

Goller F and Larsen ON (1997) A new mechanism of sound generations in songbirds. *Proceedings of the National Academy of Sciences of the United States of America* 94: 14787–14791.

Hoese WJ, Podos J, Boetticher NC, and Nowicki S (2000) Vocal tract function in birdsong production: Experimental manipulation of beak movements. *Journal of Experimental Biology* 203: 1845–1855.

Larsen ON and Goller F (2000) Direct observation of syringeal muscle function in songbirds and a parrot. *Journal of Experimental Biology* 205: 25–35.

Mindlin GB and Laje R (2005) *The Physics of Birdsong*. Berlin: Springer.

Riede T, Suthers RA, Fletcher NH, and Blevins W (2006) Songbirds turn their vocal tract to fundamental frequency of their song. *Proceedings of the National Academy of Sciences of the United States of America* 103: 5543–5548.

Suthers RA (1990) Contributions to birdsong from the left and right sides of the intact syrinx. *Nature* 347: 473–477.

Suthers RA (1999) The motor basis of vocal performance in songbirds. In: Hauser M and Konishi M (eds.) *The Design of Animal Communication*, pp. 37–62. Cambridge, MA: MIT Press.

Suthers RA (2004) How birds sing and why it matters. In: Marler P and Slabbekoorn H (eds.) *Nature's Music. The Science of Birdsong*, pp. 272–295. New York: Elsevier.

Signal Transmission in Natural Environments

R H Wiley, University of North Carolina, Chapel Hill, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

All communicatory signals consist of energy or matter transferred from a signaler to a receiver, but not all transmitted energy is a signal. The fundamental property of a signal is its ability to elicit a response from a receiver without providing all the power for the response. Coercion is not communication. Instead, a signal produces a response that is mediated by the receptors, decision mechanisms, and effectors of the receiver. Even tactile signals are separated from coercion by this property of insufficient power and thus rely on the receiver for their effectiveness.

Insufficient power to produce a response assures that transmission is a fundamental property of all signals. Transmission of a signal to the receiver in turn introduces the possibility of alterations in the structure of the signal and of masking by irrelevant energy in the environment. These possibilities result in errors in the receiver's responses, and the possibility of errors has in turn some fundamental consequences for the evolution of optimal behavior by both receivers and signalers. Understanding signal transmission is thus fundamental to understanding communication in general.

The alterations and masking of signals during transmission are most easily studied and most thoroughly understood in the case of sounds in the atmosphere.

Acoustic Signals in Air

Spherical Spreading

Sound radiating from a point in a homogeneous medium spreads spherically, so that the density of energy in a unit of area (the intensity) decreases in inverse proportion to the square of the distance from the source. The energy emitted by a source at any instant spreads over an imaginary spherical surface that depends on how far the sound has traveled from the source. Once beyond about one wavelength and one diameter from the source (in the far field), sound propagates as a wave of pressure as a result of minute movements of the molecules in the medium. Closer than this distance (the near field), sound has somewhat different properties as a result of the larger movements of molecules in the medium. Although no real source emits sound exactly from a point and no natural environment is exactly homogeneous, nevertheless measurements have shown that, at least in the far field, sounds produced by animals in relatively homogeneous conditions spread approximately spherically. As a result, the intensity of sound in general attenuates at least in proportion to the inverse square of distance from the source. This attenuation equals approximately 6 dB for every doubling of distance from the source, by the definition of decibels as -10 times the log of a ratio, in this case $-10\log((1/2)^2)$.

In the presence of gradients of wind speed or temperature, sound no longer spreads spherically. These conditions result in sound that attenuates from spreading either more or less than the 6 dB for a doubling of distance expected from spherical spreading. Sound travels faster in warmer than in cooler air, so a temperature inversion (warmer air near the ground, because of sunlight heating the surface, and cooler air above) results in sound that refracts upward as it propagates over the ground. This bending reduces sound spreading horizontally. As a result, attenuation in a horizontal direction is greater than 6 dB for a doubling of distance. Probably for this reason, few birds in open environments sing near the ground on sunny days. Instead, many sing in flight well above ground, behavior that increases the horizontal distance at which their songs can be heard.

In forests during early morning, the air above the canopy warms faster than the air below the canopy. In this circumstance sounds produced in the canopy are refracted downward from the air above, so the intensity of the sound in the canopy attenuates less than expected by spherical spreading. Primates of tropical forests often vocalize in the canopy at this time and thus get the advantage of this focusing effect to boost the propagation of their vocalizations.

Gradients of wind speed have somewhat similar effects on the propagation of sound. Wind speed is slowest next to the ground (or other surfaces) and increases with height, sometimes for 20 m or so. In these conditions, sound propagating upwind refracts upward, and sound propagating downwind refracts downward. As a consequence, a shadow zone develops upwind beyond which little sound penetrates, whereas downwind, sound is channeled horizontally. Thus under windy conditions, sound attenuates more in the horizontal direction upwind but less downwind than expected by spherical spreading. Although this effect is prominent, there are no documented cases of animals that selectively communicate downwind. For instance, we might expect in windy conditions

that birds would sing closer to the upwind sides of their territories so as to equalize the reception in all neighboring territories.

Frequency-Dependent Attenuation

Absorption of sound by molecular vibrations in the medium results in additional attenuation. Because absorption occurs in addition to the universal attenuation from spreading, it is a form of excess attenuation. Absorption of sound by molecules in the atmosphere is complex, but the practical consequences for communication with frequencies in the range of many animals' hearing (100–10 000 Hz) are relatively simple. In this range, absorption is not significantly affected by temperature or humidity of the air. However, absorption increases monotonically with frequency in all natural environments. As a result, the lowest frequencies produced by animals propagate farthest in any natural environment.

There are two other important sources of excess attenuation. First, near a boundary layer, such as the ground, sound propagating in the atmosphere interacts with energy propagating near the boundary or reflected from the boundary. This effect is also strongly frequency-dependent in accordance with the properties of the boundary. Highly porous soil can result in a nearly 180° phase shift of the low frequencies in reflected sound, so there is nearly complete destructive interference between direct and reflected sound at these frequencies during transmission near the ground. As a result of this ground effect, many birds that forage on the ground move to higher perches when they sing.

Second, sound is absorbed and reflected by objects in the path of transmission. Large objects simply reflect sound and create a sound shadow beyond. Smaller objects absorb sound in their vibrations or refract sound away from the path of transmission in complex ways. Very small objects, those well below a wavelength in diameter, have little effect on the transmission of sound. The leaves in forests have particularly interesting effects on reverberation. Many broad leaves should refract or reflect sounds with wavelengths smaller than roughly 10 cm, corresponding to frequencies greater than 3400 Hz in air. In contrast, coniferous needles should only affect sounds with much higher frequencies (1 cm diameter would correspond to a frequency of 34 000 Hz). As expected, birds in coniferous forests use higher frequencies on average in their songs that do birds of similar size in broad-leaved forests.

Provided the air and vegetation are homogeneous in composition along the path of propagation, excess attenuation is a constant fraction of the intensity of sound for each unit of distance. Because atmospheric absorption results in constant decreases in intensity of sound with distance but spherical spreading results in exponential decreases, spherical spreading dominates attenuation at short ranges, and atmospheric absorption dominates at long ranges. As a result, we should expect to find adaptations to minimize excess attenuation in communication only at relatively long ranges. Human speech, for instance, has notably poor transmission at a distance. Presumably it has evolved for communicating lots of information rapidly at short range rather than for efficacy of communicating a few crucial messages at long range.

Reverberation and Amplitude Fluctuations

Frequency-dependent attenuation alters the spectrum of sound as it propagates by disproportionately reducing the higher frequencies in a sound. Reverberation and amplitude fluctuations also have pronounced effects on the temporal properties of sound, sometimes called degradation rather than attenuation.

Reverberation results from multiple refractions and reflections of sound from objects in the path of transmission. As a result, sound no longer arrives at the receiver along a single, direct path. Instead, it arrives from a wide angle along many paths of different lengths. Some of the sound is thus delayed, in relation to the rest, in reaching the receiver. The result is a smearing of any temporal patterns in the sound. Distinct notes at the source arrive as blurred sounds. Staccato becomes sostenuto. To the extent that shapes or timing of notes is important for a receiver's response, reverberation degrades signals. Reverberation is more severe in environments with greater density of objects in the path of transmission. Forests, for instance, produce more reverberation than open air above grasslands. As expected, birds in forests, especially broad-leaved forests, are less likely to incorporate rapid repetitions of notes at any one frequency in their songs than are birds of grasslands. Because broad leaves reflect mostly sounds with higher frequencies in the range used by many songbirds, reverberation is greater for these frequencies. Greater reverberation of higher frequencies, as well as greater attenuation, can explain why birds in forests use lower frequencies on average than do birds of open habitats.

Amplitude fluctuations also degrade the temporal structure of propagating sounds. These fluctuations arise from turbulence in the flow of air. Even a gentle breeze, particularly near a rough surface such as the ground or the canopy of a forest, includes irregular eddies. The dimensions of these eddies (or their periodicities as they pass any one point) have an exponential spectrum with progressively decreasing representation of smaller dimensions (or smaller

periodicities). These eddies refract sound away from a direct path of transmission and thus impose fluctuations in the intensity of the sound for a receiver. Because the periods of these fluctuations are mostly long in relation to the time constants for animals' ears, they produce dropouts in sounds (periods of irregular length with lower intensity). Amplitude fluctuations affect sounds primarily in open environments, where wind is rarely absent, as opposed to, for instance, the interior of a forest. As a result, birds in open country (such as sandpipers, larks, pipits, and longspurs) often have prolonged tinkling songs (several seconds of extremely rapid notes in complex patterns). These birds appear to produce songs with high specificity in any brief part so that unpredictable dropouts in the song would not affect the possibility for recognition. Conspecific listeners might identify these songs by hearing any brief part them.

Background Sound

All natural environments include background sound, sound not relevant to communication by any one species or particular individuals. This background sound arises from physical features of the environment – wind, vegetation, and water. It also includes sounds made by other organisms – often not only different species but also other individuals of the same species. In recent centuries, especially in recent decades, human-generated noise, especially from airplanes and automobile traffic, pervades natural environments even at surprising distances from the source. These sources of irrelevant sound mask to varying degrees the sounds produced by animals for communication.

In tropical forests, insects (mostly Tettigoniidae, Gryllidae, and Cicadidae) create a nearly constant and ubiquitous band of noise with frequencies of 4–10 kHz. The tendency of birds in the interior of tropical forests to use lower frequencies in their songs than do those in open habitats might result primarily from the ceiling imposed by noise from insects and the floor imposed by the dimensions of the birds themselves. In addition, as described above, reverberation and excess attenuation of sound increase more rapidly with frequency in forests (particularly broad-leaved forests) than in the open. The combination of these features often creates a sound window for communication in these forests.

For any level of background sound, attenuation sets a distance beyond which the signal can no longer be detected in the noise. To determine the active space of a signal, we need measurements of attenuation and background energy and also experiments to find the signal–noise ratio necessary for recognition by conspecific listeners. Such investigations have shown that the signal space for birds' songs, primates' loud calls, or insects' stridulations equals approximately twice the average spacing of individuals, as expected if individuals moving around in their territories or home ranges stay in touch with all their neighbors at all times.

Minimizing or Compensating for Attenuation and Degradation

Investigation of the transmission of animals' sounds in natural environments has led to the environmental adaptation hypothesis: animals' sounds should evolve features that transmit well in the relevant habitat. Although sometimes not explicitly acknowledged, this hypothesis should apply primarily to signals used for relatively long-range communication. At short range, transmission should be dominated by spherical spreading. Only at long range should transmission be dominated by excess attenuation, reverberation, amplitude fluctuations, or background energy. Recent analyses of independent phylogenetic contrasts have often confirmed that birds in forests (especially broad-leaved forests) use lower frequencies on average and avoid rapid repetitions of notes in comparison with species in open habitats, as predicted by this hypothesis. Comparisons of different populations of the same or closely related species have often had similar results. Nevertheless, not all studies have confirmed the hypothesis.

Insects and anurans (frogs and toads) in particular have provided little support for the environmental adaptation hypothesis, perhaps because communication in these animals often occurs over relatively short distances. One study of cricket frogs *Acris crepitans* in Texas, however, did reveal two populations that differed as predicted in the frequencies of sound in their calls. The population that inhabited denser vegetation had lower-frequency calls, which propagated with less distortion in this habitat. These calls also propagated better in the more open habitat, although the difference was much less there. The tuning of the females' auditory neurons nearly matched the differences in calls between the populations. As a consequence, the auditory neurons of the population with high-frequency calls did a better job of rejecting the predominantly low-frequency noise in both habitats.

Adjustments to pervasive human influences on environments have also come under investigation recently. Birds in areas subject to intense and continual noise from traffic produce louder songs than conspecifics in quieter environments (the Lombard effect). They also produce songs with higher frequencies, which avoid overlap with most of the predominantly low-frequency noise from traffic.

Almost no research has focused on adjustments of signal perception (as opposed to signal production) to attenuation, degradation, and masking of acoustic

signals. A suggestion of the possibilities comes from one study that determined thresholds for hearing in the great tit *Parus major*, both in quiet conditions, as such measurements are usually conducted, and in the presence of noise matching natural background sound. In quiet conditions, the best frequencies for hearing were considerably lower than those in the species' songs, but in the presence of noise, there was a close match between the best frequencies for hearing and the frequencies in songs. Because the masking bandwidth for hearing decreased steadily with frequency, the optimal frequency for hearing shifted upward in the presence of noise.

Using Attenuation and Degradation to Localize Signals

Attenuation and degradation of signals could be put to good use by listeners to determine the distance to a signaler, an ability called ranging. Any change during transmission of a sound might be used for ranging. This capability is studied experimentally by playing back signals (such as birds' songs) in order to compare responses to relatively clean signals and to artificially or naturally attenuated and degraded signals. If the subjects decide the signals come from a distance, they might respond as if they came from a neighboring territory rather than from their own. A responding bird can also reveal a judgment of distance simply by overshooting or undershooting its approach to the speaker.

Overall intensity of sound (as affected, for instance, by spherical spreading) might not be a good cue for distance, at least in competitive interactions, because intensity can be manipulated by the sender. Nevertheless, experiments have shown that overall intensity is used by birds for ranging conspecific songs. However, they also use frequency-dependent attenuation and degradation by reverberations to judge distance, even when overall intensity is held constant. Furthermore, ranging can occur, at least approximately, even when the birds lack prior experience with a particular song pattern. Experience does play a role, however, because birds adjust their responses to reverberated songs depending on the season. During spring and summer in temperate forests, when more leaves are present and reverberation accumulates more rapidly with distance, Carolina wrens *Thryothorus ludovicianus* give territorial responses only to songs with more reverberation than during winter.

Acoustic Signals in Water

Sound propagates in water about 5 times faster than in air, in most natural conditions, and thus wavelengths for any frequency of sound are 5 times longer. As a consequence, reverberation in water results only from objects that are 5 times the size of objects that produce reverberation at similar frequencies in air. No studies have documented effects of reverberation on animals' signals in water. Excess attenuation as a result of molecular absorption also occurs much less rapidly with distance than in air. As a result, signal space is much greater in water. In addition, background sounds from farther away can effectively mask signals. One consequence is that human-generated sound, principally from ships' engines and propellers, pervades underwater habitats much more than it does terrestrial habitats.

A further complexity of sound in water, as a consequence of the longer wavelengths of sounds at audible frequencies, is the likely importance of wave guides. Any body of water less than a few tens of meters in depth might constitute a wave guide for sound. The practical consequences are reduced attenuation because of nonspherical spreading and dispersive transmission as a result of frequency-dependent speed of sound in wave guides. Dispersion alters the basic frequency–time structure of sound because a listener hears different frequencies in the source as they arrive at different times.

Although it is clear that sound is a major modality for communication by marine mammals, fish, and invertebrates, we know little about how transmission affects the structure of signals or responses to them. Because of the lower attenuation and longer wavelengths of sound in water, more communication occurs at ranges where spherical spreading dominates or even in the near field, where particle movement rather than pressure dominates. Nevertheless, there is evidence that some whales maintain communication over great distances in the oceans, distances at which frequency-dependent attenuation, reverberation, and dispersion during transmission could affect the structure of signals.

Visual and Olfactory Signals

Visual communication differs from acoustic communication because the source of the transmitted energy is not the signaler (except in the special cases of bioluminescence). Instead the signaler provides a surface that reflects light from the sun. Because sunlight can arrive either directly from the sun or indirectly from the atmosphere (the blue sky) and because it is also reflected and transmitted by any nearby objects, including vegetation, the light incident on a visual signal varies with the exact location of the signaler. Furthermore, the ability of a conspecific to detect a visual signal depends on its contrast with light reflected from the background as well as with diffuse

light scattered from particles in the atmosphere. In some cases at least, the colors of animals' signals, the places they choose to display these signals, and even their modifications of their immediate surroundings all serve to increase the contrast of signals with the background light as viewed by conspecific observers.

Because the wavelengths of visible light are so small in comparison with those of audible sound, reverberation and attenuation of visual signals over relatively short distances in the atmosphere are minor. Instead, the problem in environments with dense vegetation is assuring direct line-of-sight paths to the potential receiver. Indeed, animals communicating over large distances in the interiors of forests make little use of visual signals and correspondingly prominent use of acoustic signals.

In water, light attenuates much more rapidly than in air, and the frequency dependence within the visual spectrum is much more prominent. Long wavelengths of light attenuate more rapidly than short wavelengths, so that even at modest depths in water, red light is largely absent. Red objects near the surface appear black at depths of a few meters. Although many fish and other animals that live in relatively shallow water have conspicuous visual signals for communication, we know little about how these signals are perceived by conspecifics.

Olfactory communication often takes one of two extreme forms. Either a chemical, often one with low diffusion in the atmosphere, is deposited on the substrate to await an occasion when a receiver might pass nearby, or a chemical is released into the atmosphere for propagation by wind and diffusion to a distant receiver. Communication can thus occur over long distances in relatively short times, as in acoustic or visual communication, but also over long times but short distances. In both cases, environmental conditions affect the duration or distance of potential communication.

The dispersal of an odor in the atmosphere depends on both diffusion in the air and movement of the air. Because movement of air is turbulent, the process is complex. As a result of eddies in the atmosphere, the concentrations of an odorant at even a few meters from a source can vary substantially over distances of centimeters, a situation that complicates any calculation of the distance at which an odor can be detected by a receiver.

Errors and the Evolution of Communication

The attenuation and degradation of signals during transmission and the possibilities for masking by background energy make it clear that a signal rarely if ever arrives at its intended receiver in the form that it left the signaler. As much of the attenuation and degradation of signals and the generation of background energy depends on random turbulence and heterogeneities in the environment, every signal arrives at a receiver with some unpredictable properties. This situation raises the issue of errors in communication.

Signal detection and decision theory provide a way to understand the consequences of errors in communication. In these conceptualizations of communication, every time a receiver checks its receptors for a signal, there are four possible outcomes: a correct detection (if the receiver performs an appropriate response to the signal, one that in evolutionary terms tends to increase the spread of its genes in the population by increasing its survival or reproduction), a correct rejection (if the receiver fails to respond when there is in fact no signal present), a missed detection (if the receiver fails to respond when a signal is in fact present), and a false alarm (if the receiver responds when there is in fact no signal present). These four possibilities are the logically exhaustive set of two possibilities for response (present or absent) for each of two possibilities for the occurrence of a signal (present or absent).

Signal detection theory has had only minor application to animal communication so far, but its implications suggest that it should receive more attention. A fundamental insight is that, in realistic conditions, a receiver cannot simultaneously minimize its probability of false alarms and its probability of missed detections. Raising its threshold for response, for instance, reduces false alarms but also increases missed detections. Lowering its threshold for response has the opposite effects. Complex criteria for response likewise cannot avoid an inevitable trade-off between false alarms and missed detections. In the real world, a receiver cannot usually avoid errors. It can only adjust its criterion for response (either by evolution or by learning) to optimize its expected utility. The expected utility is a function of the probability that a signal will occur and the probabilities and the consequences of each of the four possible outcomes listed above.

An optimal threshold for response thus depends critically on the probabilities and consequences for errors. For instance, if false alarms are more costly than missed detections, then with all else equal, a high threshold would be optimal. Choosing a suitable mate might fit these conditions. In this case, an optimal receiver would behave as if it were choosy, often passing up suitable mates in order to avoid choosing an unsuitable one. This case can be called adaptive fastidiousness. As a consequence, signalers should evolve more exaggerated signals in order to surpass the high thresholds of optimal receivers. From the

perspective of signal detection theory, exaggeration and redundancy of signals evolve to increase the detectability of signals by fastidious receivers.

Alternatively, if missed detections are more costly than false alarms, then with all else equal, a low threshold would be optimal. Responding to alarm calls for predators might fit these conditions. In this case an optimal receiver would behave as if it were error-prone, often responding to false alarms in order to avoid missing an actual signal. This case can be called adaptive gullibility. As a consequence, some signalers might evolve deceptive signals to take advantage of the low thresholds of optimal receivers. There are several cases among primates and birds of such deceptive alarm calls that, for instance, occasionally allow a subordinate to scare a dominant away from food or a potential mate.

Conclusion

An understanding of the alteration and masking of signals as they pass from the signaler to the receiver is necessary for an understanding of evolutionary adaptations in communication. It is the possibility of error arising from the attenuation, degradation, and masking of signals that leads to predictions for the optimal behavior of receivers and thus for the optimal behavior of signalers.

See also: Animal Communication: Honesty and Deception; Audiovocal Communication in Bats; Communication in the Honeybee; Facial Expression in Primate Communication; Localizing Signal Sources; Multimodal Signaling in Animals; Pheromones and other Chemical Communication in Animals; Primate Communication: Evolution; Sexual Selection and the Evolution of Animal Signals; Signal Design Rules in Animal Communication; Social Communication in Whales and Dolphins; Visual Signaling in Animals; Vocal Communication in Birds.

Further Reading

Brenowitz EA (1982) The active space of red-winged blackbird song. *Journal of Comparative Physiology A* 147: 511–522.

Brown CH (1989) The active space of the blue monkey and grey-cheeked mangabey vocalizations. *Animal Behaviour* 37: 1023–1034.

Brumm H and Slabbekoorn H (2005) Acoustic communication in noise. *Advances in the Study of Behavior* 35: 151–209.

Endler JA (1993) The color of light in forests and its implications. *Ecological Monographs* 63: 1–27.

Gerhardt HC and Huber F (2002) *Acoustic Communication in Insects and Anurans*. (pp. 370–379, 431–438). Chicago, IL: University of Chicago Press.

Gerhardt HC and Klump GM (1988) Masking of acoustic signals by the chorus background noise in the green tree frog: A limitation on mate choice. *Animal Behaviour* 36: 1247–1249.

Kime NM, Turner WR, and Ryan MJ (2000) The transmission of advertisement calls in Central American frogs. *Behavioral Ecology* 11: 71–83.

Klump GM (1996) Bird communication in the noisy world. In: Kroodsma DE and Miller EH (eds.) *Ecology and Evolution of Acoustic Communication in Birds*, pp. 321–338. Ithaca, NY: Cornell University Press.

Langemann U and Klump GM (2001) Signal detection in amplitude-modulated maskers: I. Behavioral auditory threshold in a songbird. *European Journal of Neuroscience* 13: 1025–1032.

Naguib M and Wiley RH (2001) Estimating the distance to a source of sound: Mechanisms and adaptations for long-range communication. *Animal Behaviour* 62: 825–837.

Slabbekoorn H (2004) Singing in the wild: The ecology of birdsong. In: Marler P and Slabbekoorn H (eds.) *Nature's Music: The Science of Birdsong*, pp. 178–205. Amsterdam: Elsevier.

Warren PS, Katti M, Ermann M, and Brazel A (2006) Urban bioacoustics: It's not just noise. *Animal Behaviour* 71: 491–502.

Waser PM and Brown CH (1986) Habitat acoustics and primate communication. *American Journal of Primatology* 10: 135–154.

Wiley RH (1991) Associations of song properties with habitats for territorial oscine birds of eastern North America. *American Naturalist* 38: 973–993.

Wiley RH (1994) Errors, exaggeration, and deception in animal communication. In: Real L (ed.) *Behavioral Mechanisms in Ecology*, pp. 157–189. Chicago, IL: University of Chicago Press.

Wiley RH (2006) Signal detection and animal communication. In: Brockman HJ (ed.) *Advances in the Study of Behavior*, vol. 36, pp. 217–248. San Diego, CA: Elsevier.

Wiley RH and Richards DG (1978) Physical constraints on acoustic communication in the atmosphere: Implications for the evolution of animal vocalizations. *Behavioral Ecology and Sociobiology* 3: 69–94.

Wiley RH and Richards DG (1982) Adaptations for acoustic communication in birds: Sound propagation and signal detection. In: Kroodsma DE and Miller EH (eds.) *Acoustic Communication in Birds*, vol. 1, pp. 131–181. New York: Academic Press.

Witte K, Farris HE, Ryan MJ, and Wilczynski W (2005) How cricket frog females deal with a noisy world: Habitat-related differences in auditory tuning. *Behavioral Ecology* 16: 571–579.

Wollerman L (1999) Acoustic interference limits call detection in a Neotropical frog *Hyla ebraccata*. *Animal Behaviour* 57: 529–536.

Wollerman L and Wiley RH (2002) Background noise from a natural chorus alters female discrimination of male calls in a neotropical frog. *Animal Behaviour* 63: 15–22.

Single Cell Electroporation

D Dunfield and K Haas, University of British Columbia, Vancouver, BC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Single-cell electroporation (SCE) is a method for delivering DNA or other charged compounds to individual cells in dissociated culture or within intact tissues, including *in vivo* preparations. SCE can be used to simultaneously control expression of multiple genes within the same cell while leaving surrounding tissue unaltered, allowing the testing of the cell-autonomous effects of gene expression. SCE is particularly useful as a method to label neurons by introducing fluorescent dyes or DNA encoding fluorescent proteins, enabling morphological imaging.

SCE is a specialization of the general electroporation technique commonly used to transfect dissociated cells. Electroporation refers to the ability to permeabilize cell membranes by the application of an external electric field, and it is typically used in conjunction with the electrophoretic transfer of charged molecules, primarily DNA, into cells. Traditional electroporation employs an electric field generated between two flat plate electrodes on opposing sides of a solution containing both dissociated cells and the exogenous molecules to be delivered. An applied voltage between the electrodes induces movement of ions within the cells to membrane regions adjacent to the electrodes, creating localized increases in transmembrane potential. When the transmembrane potential reaches a critical point of approximately 0.5 V, the electrostatic forces maintaining the lipid bilayer integrity break down to form water-permeable pores. Charged molecules in the external medium are moved by the applied field and enter cells through these open pores. On termination of the field, pores remain open for tens to hundreds of milliseconds before collapsing to reestablish a contiguous lipid bilayer, trapping exogenous molecules within the cells. Electroporation is a versatile technique applicable to the delivery of a wide range of charged molecules relevant to research and medicine, including DNA, RNA, oligonucleotides for antisense knockdown, peptides, proteins, antibodies, dyes, and drugs. Furthermore, electroporation lacks cell type or maturational stage restrictions, has no residual transfer agents, and can be used to simultaneously co-deliver multiple compounds or genes.

More recently, electroporation has been applied to the delivery of DNA and other compounds to cells within intact tissues, including the skin, muscle, testis, spinal cord, retina, and brain. Efforts have been made to control the extent of electroporation by focally applying the delivery molecules or using smaller electrodes. The ability to restrict electroporation to an individual cell within an intact tissue represents the culmination of this effort. The SCE technique uses a pulled-glass pipette electrode with tip area approximately two to three orders of magnitude smaller than the surface area of a cell body to target electroporation to an individual cell by restricting the electric field and concentrating molecular delivery. The pipette is filled with a solution containing the delivery compound, and electrical stimulation is generated between a wire inserted into the pipette in contact with the delivery solution and a second external electrode placed in electrotonic contact with the preparation. The pipette and preparation electrodes are connected to a standard voltage source capable of generating stimulation parameters appropriate for SCE tranfection or molecular loading. Once the pipette tip is positioned in contact with a target cell, electrical stimulation induces membrane permeation near the tip and causes the delivery of charged compounds from the pipette solution through these pores and into the cell. Following the termination of the electric stimulation, the pores reseal, the pipette is retracted, and the delivery compounds are sequestered within the target cell.

Single-Cell Electroporation Pipettes – Electrodynamics

The physical geometry of the SCE pipette directly affects the amplitude and spatial distribution of the generated electric field, dramatically influencing the success of SCE. The field strength at the tip increases steeply with decreased tip diameter and increased taper angle (**Figures 1(c)** and **2(a)**). Pipette tips that have been found to be most effective for SCE have tip diameters in the range of 0.5–1 μm and taper angles greater than 10°. **Figures 1(a)** and **1(b)** show electron micrograph pictures of an SCE pipette tip of the type used for transfecting neurons within the intact *Xenopus laevis* tadpole brain, as shown in **Figures 3** and **4**. We have modeled the electric field a few micrometers within and outside of these pipettes using finite element analysis (**Figures 1(d)** and **1(e)**). These models predict that the field produced is nonuniform, with the field strength gradually increasing within the pipette toward the tip. External to the tip, the field

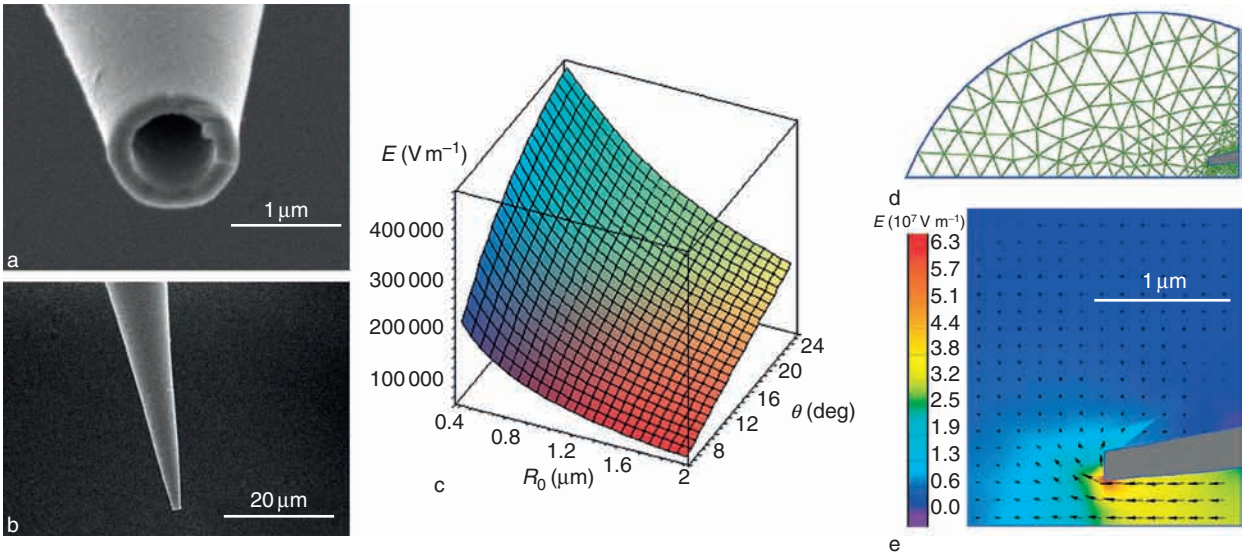

Figure 1 SCE micropipette: (a, b) scanning electron microscopy images at different scales, tip diameter $<1\,\mu m$ and tip cone angle $>10°$; (c) electric field strength (E) 500 nm from the pipette aperture with a 1 V applied voltage; (d) the field generated within and outside a SCE pipette modeled using axisymmetric FEA; (e) electric field diagram for a pipette tip with a 0.8 μm diameter aperture and 13° cone angle, with a 100 V applied voltage. In (c), R_0 is the radius of the tip opening and θ is the cone angle of the inner wall of the pipette. In (d), the diagram represents a cross section through one-half of the pipette tip, which is represented in gray on the lower right-hand side of the figure. In (e), the color code represents the amplitude of field strength and the arrows indicate the direction of the field. FEA, finite element analysis; SCE, single-cell electroporation.

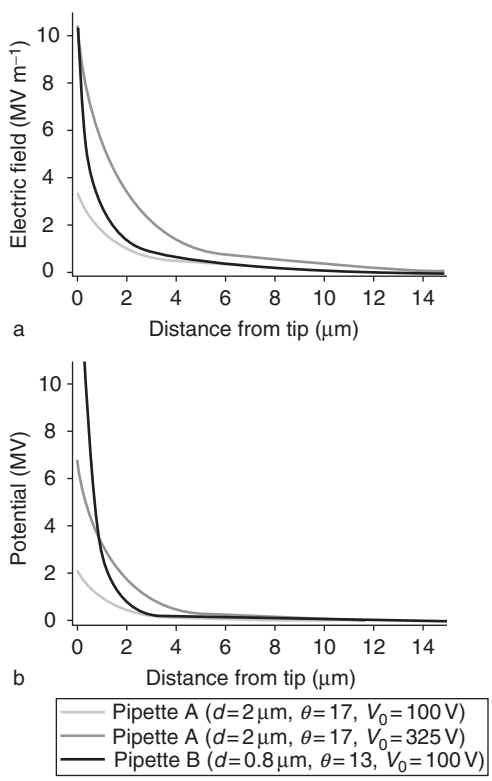

Figure 2 Electric fields and voltage potentials for various pipettes: (a) electric field along the pipette axis out of the pipette away from the tip; (b) voltage potential along the pipette axis out of the pipette away from the tip. As shown in (a), the electric fields drop off as second-order exponential decay curves. Pipette B is a typical SCE micropipette. SCE, single-cell electroporation.

strength peaks directly outside but then rapidly decays as a second-order exponential, which explains the ability of SCE to restrict transfection to individual cells within intact tissues.

The strong effect of tip aperture on SCE success is due to the varying strengths and spatial distributions of the electric fields and voltages generated by pipettes of different sizes. For a given applied voltage, as tip aperture diameter increases, the electric field strength at the tip gets weaker and broader. Although the function is not perfectly linear, doubling the aperture diameter requires approximately tripling the applied voltage to achieve the same field strength just outside the tip opening (Figure 2(a)). The decay constant for the exponential fall-off of the electric field outside the pipette tip increases as the aperture diameter increases. Because the voltage potential of an exponentially decaying electric field is the product of the electric field and the reciprocal of its decay constant (Gauss's law), the potential drop in response to the same applied voltage becomes smaller as tip aperture increases (Figure 2(b)). Therefore, as the pipette diameter increases, higher voltages are required to induce the threshold transmembrane potentials required for poration. Nonetheless, matching the electric field strength at the tip openings of two pipettes with differing aperture sizes by increasing the voltage applied to the larger pipette will not result in the same electroporation effects (Figure 2(b)). The

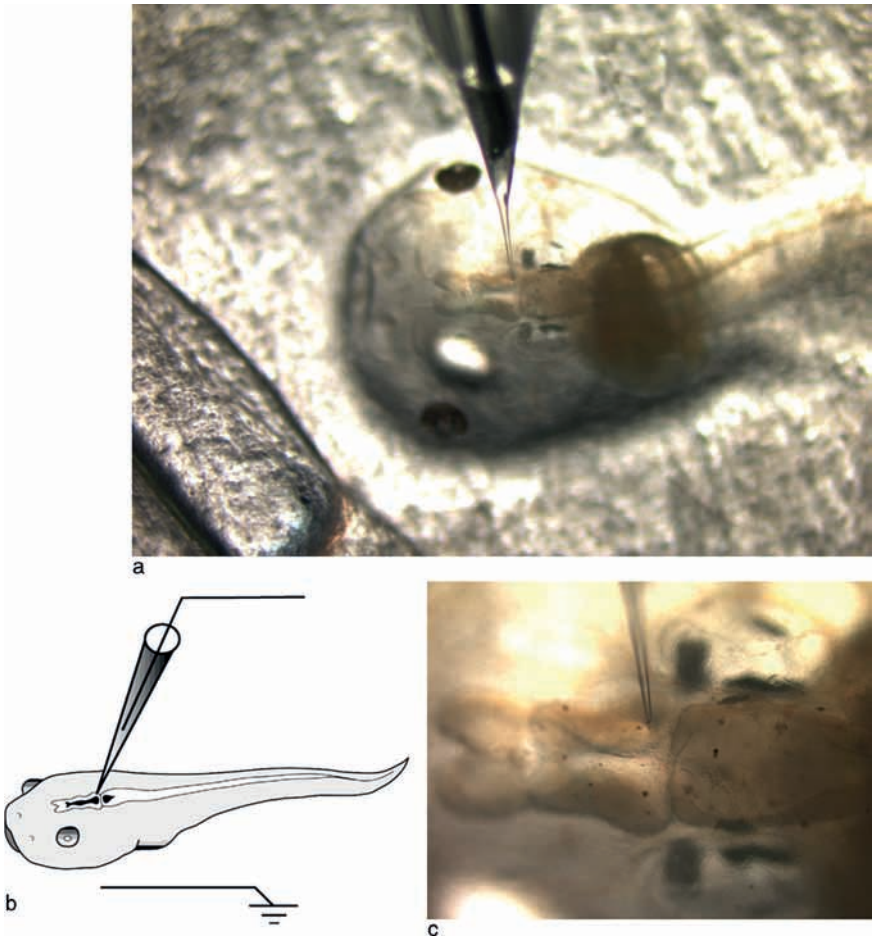

Figure 3 SCE pipette positioning in the albino *Xenopus laevis* tadpole brain: (a) positioning for *in vivo* electroporation of single neurons; (b) diagram of (a); (c) enlargement showing the insertion of the SCE pipette. Here, the SCE pipette tip is positioned blind in a region of the target neurons using low-magnification long-working-distance stereomicroscopy. SCE, single-cell electroporation.

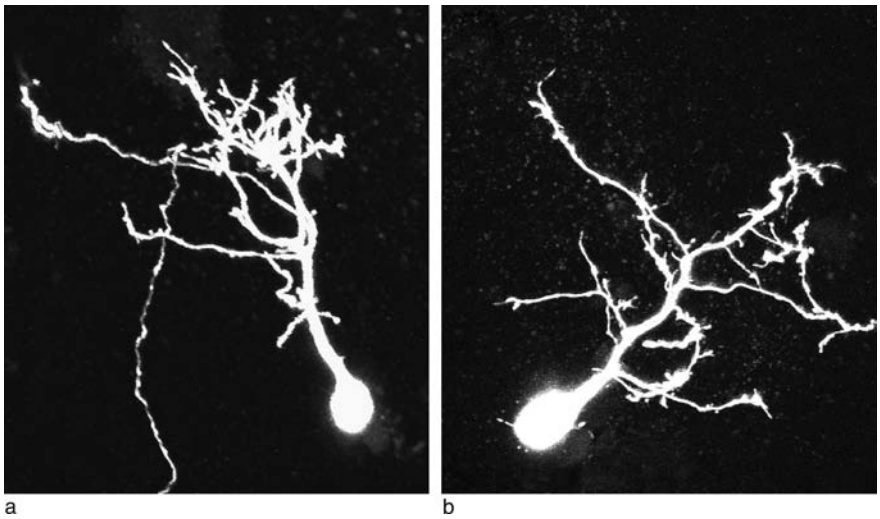

Figure 4 *In vivo* two-photon imaging of developing tadpole brain neurons labeled with SCE: (a) SCE with a plasmid encoding enhanced green fluorescent protein (eGFP); (b) SCE with dextran-conjugated Alexa Fluor 488. SCE, single-cell electroporation.

distance over which the potential drop occurs is greater for larger tips and higher voltages, producing a broader spatial distribution for cells to be affected. The spatial distribution of cells affected by a pipette tip with twice the diameter of a standard SCE pipette will be 2–3 times larger. Therefore, small tips are better able to generate the tip voltage amplitudes required to induce pore formation while restricting poration to a target cell. Larger tips either generate a low-amplitude voltage insufficient for poration or electroporate a cluster of cells.

The choice of pipette tip shapes must also take into account preparation-specific considerations involving interactions with tissue, clogging, and tip strength. Although tips with wide-tapered angles and small tip apertures are optimal for SCE, wider-tapered angles produce more displacement when inserted into intact tissues. Clogging is more common as the tip diameter and taper angle decrease and can arise from particles within the pipette solution or the accumulation of debris from the preparation. To reduce internal clogging, pipette solutions should be centrifuged at high speeds to remove contaminants. In especially sticky preparations, such as organotypic slice cultures, the application of positive pressure to the pipette can help counter clogging from external tissue. If significant layers of tissue, such as skin, must be penetrated, then increasing the tip strength by increasing the taper angle will help prevent breaking.

Single-Cell Electroporation Currents

Another critical component of SCE is the behavior of the specific charged molecules to be delivered when they are exposed to the electric fields in SCE pipette tips. Three major forces influence the movements of ions within the SCE pipette: electroosmotic flow, electrophoresis, and dielectrophoresis. Electroosmotic flow refers to the movement of liquids when a voltage is applied to an ionic fluid within a narrow glass tube. In the case of SCE, which uses common borosilicate glass, negative charges are adsorbed onto the glass surface, leaving behind a layer of positive ions (the Debeye layer), which is pulled toward the negative electrode with the application of the electric field. The motion of the Debye layer is transmitted to the rest of the fluid by friction. Electrophoretic flow is caused by the attraction or repulsion of charged ions within solutions by the oppositely charged electrodes. Dielectrophoresis occurs when particles within a solution are subjected to a nonuniform electric field and is dependent on the product of the scalar electric field and its gradient. The dielectrophoretic force affects all particles, independent of their charge, but has a strong dependence on the particles' size, shape, and chemical makeup. In the case of an SCE pipette, dielectrophoresis results in a nonlinear force from the pipette regardless of the polarity of the applied voltage. In most SCE experiments, it is expected that electrophoretic flow will be larger than the electroosmotic flow because the delivery compound will consist of a high salt concentration at a neutral of physiological pH at which the electroosmotic flow is known to be low.

Single-Cell Electroporation Pore Formation

Our current understanding of the process of SCE-mediated poration is incomplete. The creation and evolution of pores in plasma membranes induced by electric fields have been extensively modeled. However, these models, largely based on the Smoluchowski equation, may have only partial relevance to SCE because they typically describe entire cells within a relatively large and uniform electric field. During SCE, only a discrete area of the cell membrane is exposed to the applied field because the micropipette tip is relatively small (0.5–1 μm in diameter) compared to the size of neuronal somata (approximately 10–50 μm in diameter) and the applied potential is negligible, less than 1 μm outside the tip opening (**Figure 2(b)**).

Therefore, recent attempts to mathematically describe the effects of a locally applied field to a patch of membrane may be more applicable to SCE. These models describe electroporation as a multistep process initiated when exposure to an external electric field induces a sufficient transmembrane potential to overcome the electrostatic forces maintaining the lipid bilayer, inducing membrane breakdown and the formation of water wires that penetrate the lipid bilayer. The number of pores initially formed is controlled by the membrane voltage achieved, with the number of pores increasing exponentially with increasing voltage amplitude. The majority of the minute hydrophobic breaks in the lipid bilayer are rapidly destroyed by lipid fluctuations, but a small proportion continues to expand. Once growing pores exceed a radius of approximately 0.5 nm, the fatty acids of the lipid bilayer reconfigure, forming a more stable configuration with acyl groups lining the pore to create a hydrophilic channel. Membrane potential and membrane tension initially promote pore creation and expansion, but as relative pore area increases, membrane potential and tension decrease,

eventually leading to the cessation of pore growth. Therefore, initial pore density contributes to the size to which pores expand. Thus, high voltages produce numerous pores that remain small, and low voltages produce few pores with greater expansion. Because of the addition of effective tension into these membrane patch models, these findings are in contrast with previous theoretic results, which predict that high voltages create larger pores which expand rapidly, sometimes beyond the point where electroporation becomes irreversible. On termination of the voltage pulses, the pores shrink and collapse to reform the lipid bilayer by lipid fluctuations. However, if pore growth is unchecked, the pores reach a diameter at which they are unable to reverse expansion, leading to cell lysis and death. Further modeling studies predict that a second pulse delivered at short intervals changes the pores' parameters generated by the first pulse and that, by modulating the amplitude of the second pulse and interpulse interval, pore size and stability can be controlled.

Single-Cell Electroporation for DNA and Dyes

Here, we review the design of SCE stimulation parameters for the delivery of two commonly used, but disparate, compounds: plasmid DNA and dextran conjugates of fluorescent dyes. As expected from the theoretical models, different electrical stimuli are required for SCE of compounds of different charges and sizes. SCE-mediated DNA transfection is particularly challenging due to the large size of typical supercoiled plasmid expression vectors, which are often greater than 4 kb and estimated to have effective radii of 5–9 nm. Furthermore, successful transfection probably requires that DNA be delivered to the nucleus for transcription. Much debate has centered on the precise mechanisms by which DNA translocates across the plasma and nuclear membranes given the relatively large size of DNA and the small pores produced by electroporation. Using models designed for focal membrane poration, Smith et al. suggest that the formation of pores sufficiently large for DNA yet small enough to reseal can occur within a limited voltage range and that repeated pulses improve the stabilization of pores appropriate for DNA transfer. Given that DNA is strongly negatively charged, the pipette wire is typically connected to the negative pole of the voltage source and the external preparation electrode is connected to the positive pole. The application of voltage pulses induces DNA to be repelled by the pipette electrode and attracted to the external electrode. In this case, electroosmosis is negligible, and electrophoresis and dielectrophoresis drive DNA out of the pipette. However, DNA expulsion from micropipettes and DNA transfection have also been demonstrated by reversing the polarity of the voltage field to induce a phenomenon termed trapping. In this case, voltage pulses induce the electroosmostic and electrophorestic flow of DNA away from the pipette tip and back into the shaft, whereas dielectric flow remains directed out of the pipette. Near the tip, the dielectrophoretic force is slightly greater than the opposing forces, causing DNA to concentrate and become trapped. When the voltage pulse ends and the pipette field collapses, the dielectrophoretic force overwhelms electroosmosis and electrophoresis, resulting in brief ejection of a high concentration of DNA from the tip. Trapping from micropipettes has been employed for controlled DNA deposition, as well as for loading cells. The use of SCE for DNA transfection has been applied to cells in dissociated cultures, neurons in organotypic brain slice cultures, and neurons within the intact brain (**Figures 3 and 4(a)**). One of the advantages of SCE for DNA is the ability to co-deliver multiple plasmids expressing different genes to the same cell. Optimal tip diameters for SCE-mediated DNA transfection range from 0.5 to 1 μm. Reports vary, however, on the concentration of DNA required and whether additional salts in the pipette solution contribute to SCE success of cell viability. Some groups use low DNA concentrations (0.2–33 ng μl^{-1}) dissolved in saline solutions; we use DNA concentrations of 1–2 μg μl^{-1} diluted in dH_2O because no effect of additional salts was observed. Groups agree on the general stimulation parameters that yield the highest SCE transfection efficiencies: brief high-frequency trains of short pulses. For *in vivo* SCE of neurons in the *Xenopus laevis* tadpole brain (see **Figures 3 and 4(a)**), using pipettes with tip diameters of 0.5–1 μm and resistances of 20–40 MΩ, typical stimulation parameters include a DNA concentration of 1–2 μg μl^{-1} dissolved in dH_2O; 500-ms-long trains of 1-ms square pulses delivered at 200 Hz and 25–50 V; and a current of 1–2 μA.

A second type of commonly electroporated compounds are dyes (**Figure 4**), such as dextran-conjugated fluorescent dyes (Molecular Probes, Eugene, OR). The same SCE pipettes effective for DNA loading are effective for the SCE of dyes. In contrast to DNA, these dyes are small, readily pass through minute pores, and need to cross only the plasma membrane. The ease of passing through minute pores and the rapid movement of charged dyes in electric fields make it challenging to load an individual cell with

sufficient dye for complete filling of neuronal processes without also labeling neighboring cells. A solution to this problem is to use a high concentration of dye and brief high-voltage stimuli. High voltages induce many small pores and brief high voltage delivers a discrete concentrated bolus of dye only to the cell at the SCE pipette tip. The great advantage of loading neurons with dyes over SCE-mediated DNA transfection, which has a 4- to 12-h lag in the detection of the exogenous fluorescent protein, is that dye loading allows direct and immediate feedback on SCE success. Direct visualization of SCE using fluorescent microscopy allows the researcher to rapidly test a range of stimulation parameters for optimal SCE. We recommend the use of either single square pulses or brief trains of short (1-ms) pulses lasting 10–50 ms. By adjusting both the voltage amplitude and pulse or train duration while examining cell loading, the researcher can rapidly determine the parameters that brightly label but restrict labeling to single neurons. Both positive and negative polarity should be tested for any new dye because the effective charge and contribution of dielectrophoresis and electroosmosis forces may counter expected movement. It is best to fine-tune the stimulation parameters for each pipette by direct fluorescence observation of loading. However, given the phototoxic effects of fluorescent excitation, once the optimal parameters are determined, subsequent loading should be performed without fluorescent illumination. Neurons can be checked for complete loading of neuronal processes 10–30 min following SCE. SCE-mediated dye loading is a rapid and effective method for dye loading neurons for morphological imaging and for testing the quality of SCE pipettes to be used for DNA transfection. Although different stimulation parameters are required, pipette tips that function well for SCE-mediated dye loading work well for DNA transfection.

There are two approaches to carrying out SCE: blind and direct visualization of the target cells. The blind approach is useful in preparations such as the hippocampal brain slice and the *in vivo* tadpole brain (Figure 3) in which regions containing target neurons can be readily determined using low-magnification long-working-distance microscopy. In these conditions, the SCE pipette can be inserted into these regions and the stimuli applied without directly observing tip–cell contact. Although incorrect tip–cell interactions probably contribute significantly to SCE failure and lower efficiency, the long working distance and ability to reuse the same tip allow the rapid repositioning of the pipette tip to multiple sites within the same preparation. An alternative approach is to use high-magnification microscopy to directly visualize SCE tip–cell contact. Although the shorter working distances inherent in such systems restrict the rapid targeting of many sites, visualization has been shown to greatly improve SCE efficiency.

Long-Term Effects of Single-Cell Electroporation on Neurons

An important consideration for applying electroporation to neural tissue is the long-term effects induced by transient membrane breakdown. A number of studies using morphological and electrophysiological methods have detected no lasting effects on neuronal growth or neuronal function. Whole-cell patch clamp recordings of neurons 24 h following SCE transfection found no alterations in resting membrane potential, characteristics of action potential discharge, or evoked Ca^{2+} transients. Neurons in culture transfected using a similar electroporation method (microporation) also found no effects on electrophysiological measures, including synaptic transmission. *In vivo* time-lapse imaging of the growth of developing neurons within the *Xenopus laevis* tadpole brain, considered a sensitive measure of altered neuronal function, found no differences in dynamic growth between neurons transfected by SCE and those labeled with DiI. Furthermore, the expression of fluorescent proteins can be seen for up to 2 weeks in slice cultures and 3 weeks *in vivo* following SCE, the longest we have examined, suggesting continued neuronal health.

Summary

SCE is a timely advance in the technology for controlled compound delivery and gene expression for determining the roles of specific proteins in neuronal growth and function. Recent advances in genomic sequencing and gene discovery have created a bottleneck of orphan genes whose functions are poorly understood or unknown. At the same time, there have been tremendous advances in the development of dyes and molecular biology tools, such as RNAi, peptide inhibitors, and genetically encoded constructs, including fluorescent-tagged proteins and fluorescence resonance energy transfer (FRET) compounds. The effective application of these tools is limited by our ability to control their delivery to appropriate cells. SCE offers the unprecedented ability to direct these tools to individual neurons within intact tissues. Significantly, such targeted manipulation leaves the remaining tissue unaltered, so any effects observed are not caused by the secondary alteration of the target cell's environment.

See also: Gene Therapy: Genetically Modified Cells; Gene Therapy: Direct Viral Delivery; Imaging Studies Using Reporter-Gene Transgenic Rats; Single Cell Genomic DNA Analysis; Single Cell Molecular Analysis Procedures; Single Cell PCR Coupled with Electrophysiology.

Further Reading

Haas K, Jensen K, Sin WC, Foa L, and Cline HT (2002) Targeted electroporation in *Xenopus* tadpoles *in vivo* – from single cells to the entire brain. *Differentiation* 70: 148–154.

Haas K, Sin WC, Javaherian A, Li Z, and Cline HT (2001) Single-cell electroporation for gene transfer *in vivo*. *Neuron* 29: 583–591.

Ho SY and Mittal GS (1996) Electroporation of cell membranes: A review. *Critical Reviews in Biotechnology* 16: 349–362.

Inoue T and Krumlauf R (2001) An impulse to the brain – using *in vivo* electroporation. *Nature Neuroscience* 4: 1156–1158.

Itasaki N, Bel-Vialar S, and Krumlauf R (1999) 'Shocking' developments in chick embryology: Electroporation and in ovo gene expression. *Nature Cell Biology* 1: E203–E207.

Lovell P, Jezzini SH, and Moroz L (2006) Electroporation of neurons and growth cones in *Aplysia californica*. *Journal of Neuroscience Methods* 151: 114–120.

Miyasaka N, Arimatsu Y, and Takiguchi-Hayashi K (1999) Foreign gene expression in an organotypic culture of cortical anlage after *in vivo* electroporation. *NeuroReport* 10: 2319–2323.

Muramatsu T, Mizutani Y, Ohmori Y, and Okumura JI (1997) Comparison of three nonviral transfection methods for foreign gene expression in early chicken embryos in ovo. *Biochemical and Biophysical Research Communications* 230: 376–380.

Neu JC and Krassowska W (1999) Asymptotic model of electroporation. *Physical Review E* 59: 3471–3482.

Neumann E, Kakorin S, Tsoneva I, Nikolova B, and Tomov T (1996) Calcium-mediated DNA adsorption to yeast cells and kinetics of cell transformation by electroporation. *Biophysical Journal* 71: 868–877.

Neumann E, Kakorin S, and Toensing K (1999) Fundamentals of electroporative delivery of drugs and genes. *Bioelectrochemistry and Bioenergetics* 48: 3–16.

Olofsson J, Nolkrantz K, Ryttsen F, Lambies B, Weber S, and Orwar O (2003) Single-cell electroporation. *Current Opinion in Biotechnology* 14: 29–34.

Potter H (1988) Electroporation in biology: Methods, applications, and instrumentation. *Analytical Biochemistry* 174: 361–373.

Rae JL and Levis RA (2002) Single-cell electroporation. *Pflugers Archiv* 443: 664–670.

Rathenberg J, Nevian T, and Witzemann V (2003) High-efficiency transfection of individual neurons using modified electrophysiology techniques. *Journal of Neuroscience Methods* 126: 91–98.

Sin WC, Haas K, Ruthazer ES, and Cline HT (2002) Dendrite growth increased by visual activity requires NMDA receptor and Rho GTPases. *Nature* 419: 475–480.

Smith KC, Neu JC, and Krassowska W (2004) Model of creation and evolution of stable electropores for DNA delivery. *Biophysical Journal* 86: 2813–2826.

Sukharev SI, Klenchin VA, Serov SM, Chernomordik LV, and Yu AC (1992) Electroporation and electrophoretic DNA transfer into cells: The effect of DNA interaction with electropores. *Biophysical Journal* 63: 1320–1327.

Swartz M, Eberhart J, Mastick G, and Krull CE (2001) Sparking new frontiers: Using *in vivo* electroporation for genetic manipulations. *Developmental Biology* 233: 13–21.

Washbourne P and McAllister AK (2002) Techniques for gene transfer into neurons. *Current Opinion in Neurobiology* 12: 566–573.

Ying L, White SS, Bruckbauer A, Meadows L, Korchev YE, and Klenerman D (2004) Frequency and voltage dependence of the dielectrophoretic trapping of short lengths of DNA and dCTP in a nanopipette. *Biophysical Journal* 86: 1018–1027.

Single Cell Genomic DNA Analysis

P B Crino, University of Pennsylvania Medical Center, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

The use of human brain tissue for basic and disease-based research has rapidly expanded over the past decade. Tissue resected during tumor or epilepsy surgery or obtained postmortem from patients with a variety of neurological or psychiatric diseases can provide a new window for viewing the cellular processes that are involved in disease pathogenesis, and our understanding of many human brain disorders has been greatly facilitated by human tissue studies. In addition, many fundamental neurobiological processes that are well defined in rodent systems can also be studied in frozen or fixed human brain tissue. Perhaps most exciting, however, is that new experimental methodologies to extract genomic DNA from archival tissue samples now permit an in-depth analysis of the molecular mechanisms and genetic causes of virtually any inherited or sporadic neuropsychiatric syndrome. With the advent of tissue microdissection, an even finer experimental resolution can be gained, so gene sequence alterations can be defined in select cell populations. There are new protocols and optimization procedures to assay DNA sequences in human brain tissue that are now widely available.

Why Evaluate DNA Sequences in Human Brain Tissue?

For many neurological and psychiatric disorders, there is little, if any, evidence for genetic or Mendelian inheritance patterns. For example, epilepsy, schizophrenia, autism, sporadic brain tumors, developmental malformations, and vascular abnormalities often occur as sporadic disorders without a clear family history. Thus, the only plausible genetic or molecular mechanism to account for these disorders is a *de novo,* rather than inherited, gene mutation. From an experimental perspective, these disorders provide a significant challenge because large family pedigrees or even monozygotic twins are unlikely to yield insights. Perhaps even more challenging is that brain structure in many sporadic neuropsychiatric disorders is intact and thus morphological analysis may not be helpful.

One interesting hypothesis is that sporadic neurological disorders may in some cases result from somatic mutations occurring in brain progenitor cells but that are not present in the germline DNA. In this scenario, a deleterious gene mutation might occur during the many rounds of cellular mitosis occurring throughout the course of brain development or perhaps within a progenitor cell in the mature brain. There is extensive evidence in the brain tumor literature that somatic mutational events are commonly detected in genes such as EGFR, KIT, and PDGFR in brain tumor tissues. Furthermore, large heterozygous gene deletions in chromosomes 1 and 19 have been detected in tumors and often correlate with clinical variables such as survival. Thus, studies sequencing genomic DNA from resected or postmortem brain tissue specimens may aid in identifying somatic mutations that are confined to small regions of the brain but cannot be identified in peripheral blood (Figure 1). Furthermore, because the gene mutation may affect only a subset of cells within a given brain region, the ability to sequence DNA from phenotypically defined cell populations is a powerful approach.

Cataloging Tissue

Most human brain tissue used in research is obtained either directly from a neurosurgeon in the operating room, through a pathologist, or from an existing tissue bank. In view of the Health Insurance Portability and Accountability Act (HIPAA) regulations on the security of electronic healthcare information, it is of the utmost importance to maintain patient confidentiality in human research. Unless specific phenotypic parameters are under investigation, investigators must avoid including patient-identifying information in their research publications. It should be noted that birth dates, social security numbers, medical record numbers, and pathology/autopsy numbers are considered patient identifiers and thus cannot be used to label tissue in the laboratory. Nonetheless, it remains imperative for researchers to have some ability to track samples for a variety of reasons such as collaboration, phenotypic analysis, and requests for more tissue, and thus a system should be designed for tissue identification and storage of necessary patient information. One strategy is to assign each tissue sample a lab-specific number or label so that no member of the lab can identify the source of the sample. A separate strictly confidential database (for example, in Excel) of information for each specimen is maintained (in duplicate) in a password-protected limited-access file. Thus, an investigator can maintain tissue identity for research purposes but protect patient confidentiality. Prior to investigators' embarking on human tissue research, we strongly suggest that investigators meet with

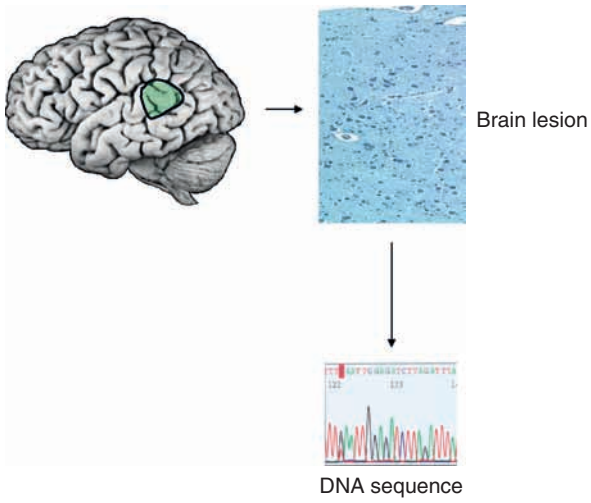

Figure 1 Schematic representation of how a somatic mutation could affect a focal brain region. In this scenario, a small brain area (depicted in green) contains cells that have sustained a mutation during brain development (e.g., a malformation) or after maturity (e.g., a brain tumor). The altered morphology (brain lesion) can harbor a causative gene mutation (DNA sequence) that can be detected only by an analysis of DNA extracted from that brain area. Note that the mutation is not detectable in cells outside the brain lesion.

appropriate protocol personnel at their home institutions to assure complete compliance with the standards for human tissue research.

Tissue Fixation and Storage

A common myth is that resected tissue used for molecular analysis must be kept frozen or used immediately after removal. Although it is true that frozen tissue yields high-quality mRNA or DNA, paraformaldehyde- or formalin-fixed paraffin-embedded (FFPE) tissue can be used for both DNA and mRNA analysis. As a general rule, frozen tissue is exquisitely sensitive to transient rises in temperature. Even 10–20° changes can lead to mRNA and protein degradation over time. Frozen tissue should be maintained at –70 °C in a frost-controlled freezer. The freezer should have a backup power source with an alarm system should the internal temperature rise above –50 °C. Contact people and appropriate information should be displayed prominently near the freezer in case of emergencies. A backup freezer should be identified in a nearby lab to prevent loss of tissue in case the power should fail. Perhaps the most optimal situation for tissue analysis is to obtain both fresh-frozen and fixed material, although this may pose logistical procurement problems because specimens must typically be analyzed by a neuropathologist prior to experimental analysis. The use of fixed tissue has particular appeal because molecular analysis can be combined with tissue immunohistochemistry to define, for example, protein expression or to use protein expression as a phenotypic marker for cell-specific DNA analysis. In addition, it is often much easier to obtain archival FFPE pathological material stored than fresh-frozen tissue.

Investigators may have little control over the selection of tissue fixatives by pathologists. In most cases, fixation for pathological preparation uses paraformaldehyde or 10% neutral buffered formalin. These protein cross-linking fixatives afford excellent protein fixation, but paraformaldehyde can compromise the successful extraction of nucleic acids. Boiun's fixative is occasionally used; it combines picric acid with formalin and is very stringent. In general, most labs avoid using Bouin's fixed tissue for nucleic acid work. Ethanols are denaturing fixatives and thus do not pose significant problems for nucleic acid extraction. Carnoy's fixative and 10% neutral buffered formalin do not appear to compromise DNA extraction. Some recent studies have shown that the fixative methacarn yields highly successful amplification of long DNA stretches extracted from FFPE.

In general, for small blocks of tissue (less than 2 cm^3), a fixation period of 2–3 weeks may be effective, whereas for larger blocks or for whole brains or hemispheres several more weeks are necessary. One question that arises is whether there is a fixation time after which DNA or mRNA cannot be extracted from fixed tissue. There is no definite answer to this question; in general, an initial failure to recover DNA can often be overcome by the manipulation of various protocols. There is reasonable evidence that DNA extracted from tissue stored for long periods may fragment, leading to small or short sequences that can prove problematic for sequencing studies. DNA fragmentation becomes an issue when longer sequences are needed to define gene mutations or polymorphisms or when DNA is used to probe array platforms such as single nucleotide polymorphism (SNP) arrays. Paraffin blocks should be kept in cool, dark, dust-free boxes. However, DNA is a stable biomolecule, and thus its breakdown is unlikely even in fixed, paraffin-embedded tissue blocks. Our lab has sequenced through several genes including *PTEN*, *TSC1*, and *TSC2* in genomic DNA extracted from paraffin-embedded tissue blocks stored for up to 5 years, and reports in the literature cite successful polymerase chain reaction (PCR) amplification of DNA extracted from brain tissue stored for up to 30 years.

Work Environment

In contrast to mRNA analysis, DNA degradation is uncommon due to lab practices. However, all glassware should be autoclave sterilized and fresh

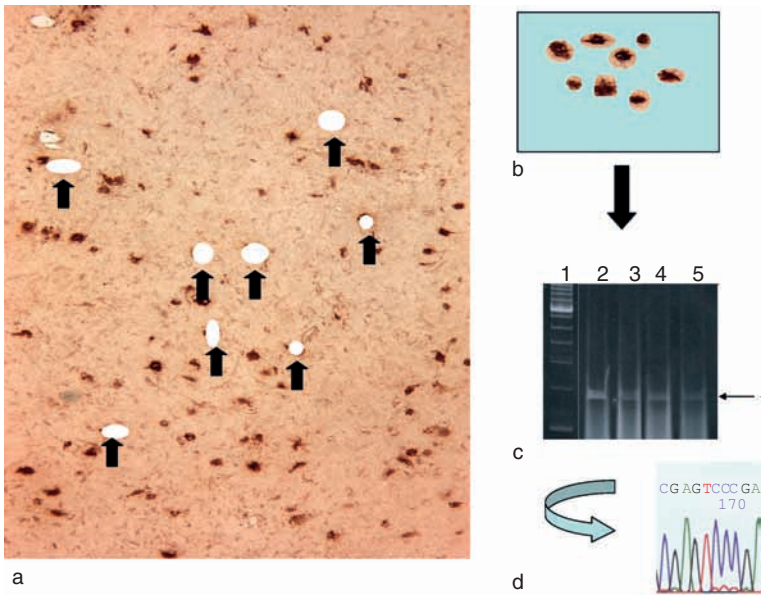

Figure 2 Microdissection of cells and DNA extraction and sequencing: (a) microdissection of multiple single cells (arrows); (b) pooling of microdissected cells; (c) DNA extraction (agarose gel; arrow depicts PCR amplicon); (d) sequencing of the band for mutation or polymorphism detection using specific DNA primers. PCR, polymerase chain reaction.

plasticware should be used to avoid contamination. Fresh tissue and frozen unfixed tissues should be treated as potentially infectious, and thus universal precautions (gloves, eye protection, and lab coat) should be implemented. Although fixation of human tissue largely inactivates most pathogens (e.g., HIV and hepatitis B and C), universal precautions for everyday work with fixed human tissue should be implemented. It is useful to know or ascertain the HIV and hepatitis serology status in the patient samples, and these data are usually available if the tissue is procured from existing tissue banks. For those working on surgical specimens associated with known human pathogens, communication with institutional safety offices is requisite to ensure safe practices for all lab staff. Transmissible spongiform encephalopathies are a particularly serious risk when working with human brain tissue. Even though it is believed that formalin fixation at least diminishes the risk of lab-related transmission of spongiform encephalopathy, utmost care must be taken to protect lab personnel. As a general practice, gloves must be worn for every instance of contact with human brain tissue, whether it is during microtome sectioning, transfer of slides, or immunohistochemistry. Investigators should review their own institutional policies on working with human tissues and pathogens prior to embarking on human tissue research.

Microdissection of Single Cells

Single cells (e.g., neurons, astrocytes, and endothelial cells) can be microdissected from either frozen or FFPE brain tissue with good success. It is up to the investigator's discretion to determine how cells will be grouped (e.g., morphology, age, laminar location, or protein expression). Although individual cells can be recognized by phase-contrast microscopy or by simple cresyl violet labeling, perhaps a more elegant method for visualizing cells is with immunohistochemistry. In this way, individual cells can be phenotyped on the basis of protein expression. For both frozen and fixed tissue, immunohistochemistry provides a reliable strategy for identifying classes of cells, either by fluorescence or by diaminobenzidine visualization.

The technical feat of microdissection can be accomplished by a single joystick-driven micromanipulator system or by a laser-capture microdissection system. In the former approach, a stainless steel microscalpel can be used to remove single cells under light microscopy; microdissected cells can then be pooled for DNA extraction. In the latter approach, multiple cells can be removed by laser capture and thereby provide a more rapid means of attaining adequate concentrations of DNA for analysis (Figure 2).

Extraction of Genomic DNA

There are several commercially available kits for DNA extraction that use proprietary reagents and buffer systems. These can be applied to whole tissue blocks and to single sections. For example, the Trizol reagent provides a means of simultaneously extracting DNA, mRNA, and protein from a single fixed or

frozen tissue block, but does not work well for single microdissected cells from FFPE tissue.

One protocol that provides high-quality DNA from FFPE tissue sections or blocks is the following. The sections are first rehydated through xylenes and graded ethanols. The cells are digested in a digestion buffer (NaCl 0.1 M, Tris–Hcl 0.01 M, EDTA 0.025 M, SDS 0.5%) overnight at 55 °C. Then, cells are treated with Proteinase K ($20\,mg\,ml^{-1}$) to reverse cross-linking fixation and permit more effective DNA extraction. Following Proteinase K, the sample is extracted in phenol-chloroform-isopropyl alcohol. The DNA is then extracted with ammonium acetate (7.5 M)/glycogen. The extracted DNA is then assessed by an optical density (OD) measurement to determine its concentration and purity and assessed by agarose gel analysis to determine the size of extracted products.

Single-cell DNA extraction requires a bit more technical expertise and specific protocols. In particular, companies (e.g., Arcturus) that make laser-capture microdissection systems also provide proprietary kits for extracting genomic DNA from single pooled cells. These generally work well for most applications. It is implied that for most single-cell DNA work, an amplification procedure such as PCR will be necessary to screen for mutations. Alternatively, several new whole-genome amplification kits exist that can adequately enhance the copy number of large numbers of genes. DNA extracted from paraffin-embedded tissue can be used for DNA sequencing and as a probe for SNP arrays. As already discussed, immunolabeling aids in identifying cells for dissection. In our lab, we use immunohistochemistry and visualize antibody labeling with the avidin-biotin-diaminobenzidine approach. It is recommended that multiple cells be dissected and pooled for analysis. For most applications, the isolation of DNA from 50 to 200 cells provides adequate template DNA for PCR analysis, although in high-quality tissue samples 20–50 cells may be adequate.

Applications of Genomic DNA Analysis in the Brain

Analysis of genomic DNA provides a unique opportunity to study gene sequences in human tissue. For example, in monogenic disorders, if the genotype of a particular patient is unknown, it can be determined by direct sequencing of the candidate gene in genomic DNA from a brain tissue specimen. Gene sequencing in brain DNA can define mutations or polymorphisms as easily as in peripheral blood DNA. Indeed, reliable and reproducible results can be determined with published DNA sequence primers for most genes. Sequencing of DNA extracted from brain tissue also has a unique application – the detection of somatic gene mutations that cannot be otherwise detected in the blood. For example, there is good evidence that a variety of brain malformations (e.g., vascular malformations) and tumors result from somatic, not germline, gene mutations. In these disorders, the putative molecular mechanism is that mutations occurred not as inherited or meiotic events but rather in progenitor cells during brain development. For some glial-derived tumors, there is evidence that an accumulation of mutations in multiple sites occurs as the lesions progress. Thus, screening DNA extracted from brain tissue provides a new and high-yield strategy for studying the molecular pathogenesis of these disorders. Note that labs must meet Clinical Laboratory Improvement Amendments (CLIA)-directed regulatory requirements; if not, the genetic screening results cannot be communicated to the patient or family member, nor can it be used for clinical practice, decision making, or counseling.

Although there have been a growing number of studies that assay DNA sequence alterations in human brain tissue, especially in brain tumor specimens, there have been only a few recent studies that have used single-cell DNA analysis in human brain tissue. In an initial analysis, sequencing of the TSC1 and TSC2 genes was accomplished in human tuberous sclerosis complex. A compelling study demonstrated the feasibility of performing loss of heterozygosity (LOH) and microsatellite instability assays in human brain malformations. *In silico* high-throughput sequencing was accomplished in microdissected cells in ganglioglioma, a congenital brain tumor. An analysis of genomic DNA from microdissected single cells demonstrated an intact gene sequence and several potentially significant polymorphisms in the PTEN gene in focal brain malformations associated with epilepsy. Similarly, the β-catenin gene was sequenced in human hemimegalencephaly, a brain malformation characterized by unilateral hemispheric enlargement and associated with severe seizures.

If adequate template DNA is extracted, there is no limit to the potential technical applications. For example, in addition to gene sequencing, analysis of SNP arrays to search for heterozygous or homozygous deletions can be accomplished. Alternative techniques include LOH and muliplex ligation product amplification (MLPA) assays. Recent studies have shown that single-strand conformation polymorphism analysis (SSPA) can be used in microdissected tissue specimens to identify gene mutations. Array-based comparative genome

hybridization (aCGH) is a high-throughput strategy for detecting deletions in candidate gene regions. This approach has been successfully implemented in FFPE brain tumor specimens.

See also: Single Cell Electroporation; Single Cell Molecular Analysis Procedures; Single Cell PCR Coupled with Electrophysiology.

Further Reading

Becker AJ, Klein H, Baden T, et al. (2002) Mutational and expression analysis of the reelin pathway components CDK5 and doublecortin in gangliogliomas. *Acta Neuropathologica* (Berlin) 104: 403–408.

Becker AJ, Urbach H, Scheffler B, et al. (2002) Focal cortical dysplasia of Taylor's balloon cell type: Mutational analysis of the TSC1 gene indicates a pathogenic relationship to tuberous sclerosis. *Annals of Neurology* 52: 29–37.

Crino PB (2002) Gene expression analysis as a strategy to understand the molecular pathogenesis of infantile spasms. *International Review of Neurobiology* 49: 367–389.

Crino PB and Becker AJ (2006) Gene profiling in temporal lobe epilepsy tissue and dysplastic lesions. *Epilepsia* 47: 1608–1616.

Crino PB, Dichter MA, Trojanowski JQ, and Eberwine J (1996) Embryonic neuronal markers in tuberous sclerosis: single cell molecular pathology. *Proceedings of the National Academy of Sciences of the United States of America* 93: 14152–14157.

Crino PB and Eberwine JH (2001) Analyzing mRNA in single neurons. *Current Protocols in Neuroscience* 1: 5.3.1–5.3.15 [online].

Fassunke J, Blumcke I, Lahl R, et al. (2004) Analysis of chromosomal instability in focal cortical dysplasia of Taylor's balloon cell type. *Acta Neuropathologica* (Berlin) 108: 129–134.

Schick V, Majores M, Engels G, et al. (2006) Activation of Akt independent of PTEN and CTMP tumor-suppressor gene mutations in epilepsy-associated Taylor-type focal cortical dysplasias. *Acta Neuropathologica* (Berlin) 112: 715–725.

Yu J, Baybis M, Lee A, et al. (2005) Targeted gene expression analysis in hemimegalencephaly: Activation of β-catenin signaling. *Brain Pathology* 15: 179–186.

Single Cell Molecular Analysis Procedures

T J Bell, R Sánchez-Alvarez, and J Eberwine, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Single Cell mRNA Analysis Using *In Situ* Hybridization

In situ hybridization was the first technique developed to detect target mRNAs in individual cells. It simply involves the hybridization of a labeled RNA or single-stranded DNA probe with a specific target nucleic acid (RNA or genomic DNA) within a cell. The initial *in situ* hybridization experiments were performed to detect DNA and ribosomal RNA in individual cells. This technique was later used to map the locations of particular genes on chromosomes. This procedure has been successfully used to analyze mRNA expression at cellular and subcellular levels in several different sample preparations – tissue sections, cultured cells, or intact whole mounted tissues. Studies in the nervous system frequently enhanced this procedure by performing *in situ* hybridization in combination with immunocytochemistry. This approach can be used to detect mRNAs in immunocytochemically defined cell types and also determine subcellular localization of the mRNAs and their corresponding protein products. The main limitations of the *in situ* hybridization technique are that only a small number of mRNAs within the same cell can be analyzed, it lacks the sensitivity to detect low-abundance mRNA transcripts and correlating functional properties, and the mRNA expression profiles within the same neuron are experimentally challenging. Therefore, this approach is not always the best option for studies using neuronal tissues. The main advantages of using *in situ* hybridization are that technically it is the simplest single cell mRNA analysis procedure to perform and this technique can be used to quantitatively analyze mRNA levels in single cells.

General Considerations for mRNA Analysis by *In Situ* Hybridization

The primary precautions that need to be taken when analyzing mRNA expression by *in situ* hybridization are essentially high-quality general lab practices. Gloves need to be worn when handling the samples and reagents to avoid contamination by RNA-degrading enzymes. All solutions, reagents, glassware, and dissecting instruments need to be prepared RNAase-free and preferably designated for RNA or *in situ* hybridization work only. As an extra precaution, all the final working solutions can be filtered through 0.2-μm filters before use and a high-grade fixative, such as electron microscopy grade paraformaldehyde, should be used for cell fixation. Fundamentally, successful *in situ* hybridization experiments depend on designing high-quality gene-specific probes and stringent hybridization conditions. The probe sequence should be checked in BLAST search to confirm specificity to target mRNA. Depending on the target mRNA, the probe length will most likely range between 100 and 1000 bp. This probe size range has three important advantages: (1) It allows efficient access to the cellular mRNAs; (2) it is sufficient in length for highly specific probe target mRNA hybridization; and (3) it is an adequate length for visualization detection. A stringent hybridization temperature ranges between ~37 and 50 °C. Finally, depending on the abundance of target mRNA, the probe concentration is usually between 0.1 and 10 ng μl^{-1}. One potential pitfall to consider when using the *in situ* hybridization procedure is the obscuring of the mRNA target sequence by a secondary RNA structure or association with RNA-binding proteins. Generating multiple probes to the same target mRNA is a simple approach to solving this problem.

Subcellular mRNA Analysis Using *In Situ* Hybridization

In the nervous system, cell-specific gene expression provides a potent mechanism for altering the properties of individual neurons. Neurons can be regulated to modify the expression profile and ultimately their functional properties by limiting specific mRNAs to different cellular compartments of a neuron. Cell-specific mRNA expression coupled with mRNA localization provides a neuron with a highly sophisticated focal mechanism to regulate the functional properties of neurons and regulate synaptic transmission. *In situ* hybridization is the best procedure to analyze subcellular mRNA expression patterns. For example, the Arg 3.1 gene encodes activity-regulated cytoskeleton-associated protein and the mRNA for Arg 3.1 is rapidly transported to the dendrites of cortical neurons in an activity-dependent manner. The *in situ* hybridization procedure was used to demonstrate the selective transport and localization of Arg 3.1 to active synapses in dendrites required for the activation of the glutamatergic NMDA receptors. *In situ* hybridization has also been used in combination with transgenic mice to correlate the disruption of subcellular mRNA expression patterns in the absence of the

RNA-binding protein, FMRP (fragile X mental retardation-1). *In situ* hybridization analysis of neurons from *Fmr1* knockout mice showed changes in the subcellular distribution of several known FMRP cargo mRNAs.

Single Cell Reverse-Transcriptase Polymerase Chain Reaction

Several studies have successfully used the polymerase chain reaction to analyze mRNA from single cells. This technique greatly increases sensitivity and the number of RNAs that can be analyzed in a single cell. The first studies to use single cell reverse-transcriptase polymerase chain reaction (scRT-PCR) demonstrated mRNAs encoding different glutamate receptor subunits, and these were present at various abundances in single neurons. Because of the sensitivity and versatility of the procedure, scRT-PCR was the first method to correlate molecular analysis at the single cell level with electrophysiological responsiveness in neurons. For splice variant analysis at the single cell level, scRT-PCR is usually the best approach compared to amplified antisense RNA amplification (aRNA) and *in situ* hybridization. The population of RNA products generated by aRNA amplification limits its use for splice variant analysis to the 3′ region of mRNAs. *In situ* hybridization lacks the sensitivity to detect low-abundance splice variants and probe design could be problematic when creating splice variant-specific probes. All of these limitations are overcome by using the scRT-PCR approach. Primers can be designed to amplify any region of the mRNA, and splice variants that differ in only a single based can be separated on high-resolution gels. However, unlike aRNA amplification and *in situ* hybridization, scRT-PCR is an exponential amplification so quantification of mRNA levels is challenging. Another potential limitation is designing high-efficiency PCR primers for mRNA or a specific region of the mRNA of interest. The sequence of the target mRNA must be known and the region of interest must not be overly G–C rich to permit the placement of high-quality PCR primers in the appropriate region of the target mRNA.

General Considerations for mRNA Analysis by Single Cell RT-PCR

Different precautions need to be taken when analyzing genes at the cellular level compared to whole tissue RT-PCR analysis. When making cDNA from single cells, one of the main concerns is preventing the small amount of mRNAs from degrading. Therefore, some type of RNase inhibitor must be used during the cDNA synthesis reaction. To optimize the single cell cDNA synthesis reaction and to make the protocol more cost-efficient, the reverse transcription reaction should be downscaled to levels more appropriate to single cells. To ensure primer specificity, the primer sequences should be checked by BLAST search in GenBank to confirm specificity for the target gene. The PCR primers should be placed on different exons to avoid amplification of genomic DNA. The reliability and efficiency of the PCR primers can be determined by testing the primers on cDNA derived from whole tissue. The whole tissue cDNA samples can be diluted to levels of total RNA corresponding to approximately 10 cells (1 ng) and a single cell (100 pg). If the primers perform well under these conditions, they should be suitable for scRT-PCR. Lastly, the PCR products should be cloned and sequenced to confirm the identity of the sequence of the PCR product from both whole tissues and single cells. To reduce variability from the reagents, RT-PCR reagents and enzymes should always be divided into aliquots to screen single cells in large groups (~50 neurons) with identical stocks of RT-PCR reagents and enzymes. Negative controls lacking cDNA have to be performed and processed in parallel with single cell RT-PCR reactions.

Splice Variant Analysis by Single Cell RT-PCR

The nervous system contains thousands of functionally distinct subsets of neurons. These neuronal populations express specific subsets of genes that underlie their unique properties. Cell-specific alternative splicing offers a finer level of control within specific regions of the nervous system. Several studies have used scRT-PCR to analyze splice variant expression patterns at the cellular level. For example, the KCNMA1 gene encodes the pore-forming α subunit of BK channels and is subjected to cell-specific alternative splicing. scRT-PCR was used to show select KCNMA1 splice variants and show differential expression along the tonotopic map in individual hair cells. Functionally, the splicing of KCNMA1 plays a role in altering the repolarization rate of the hair cell, thus setting the electrical tuning frequency in individual auditory hair cells. scRT-PCR has also been used in combination with whole cell patch recordings to correlate to functionally analysis with cell-specific splicing in nociceptive neurons of dorsal root ganglia. The Cav2.2 gene encodes the pore-forming α subunit of N-type calcium channels, and a splice variant of this gene was identified by scRT-PCR to be preferentially expressed in subsets of nociceptive neurons. Functionally, scRT-PCR has been used to link this splice variant to changes in the N-type calcium current densities in these neurons.

Antisense RNA Amplification

Some of the limitations of *in situ* hybridization and scRT-PCR have been overcome by a technique called amplified antisense RNA amplification. This procedure is useful to amplify RNA from a small amount of material and is employed particularly for amplifying RNA from single cells in which the level of the RNA is approximately 0.1 pg. As shown in Figure 1, a cell is isolated by using a patch pipette or by laser capture microdissection, and then it is aspirated and transferred to a microcentrifuge tube. The RNA of a cell is reverse transcribed into cDNA by using an oligo-dT primer that contains a T7 RNA polymerase promoter site. Then, the obtained cDNA contains the T7 promoter sequence at the 5′ end. After cDNA is double stranded, T7 RNA polymerase is used to transcribe RNA copies. The aRNA product can be used as a template for cDNA synthesis and conversion into double-stranded cDNA containing a T7 RNA polymerase promoter site so that subsequent rounds of RNA amplification can be made. Only two rounds of RNA amplification are required to obtain a sufficient amount (approximately 106-fold) of RNA from the original pool. This procedure can be used in fixed cells following a similar protocol. Figure 2 provides a brief summary of the protocol. The aRNA protocol is a linear amplification so the mRNA levels can be quantified. This technique is predominantly used to screen microarrays and macroarrays. Approximately, 70% of published microarray studies have used aRNA amplification to generate their probes.

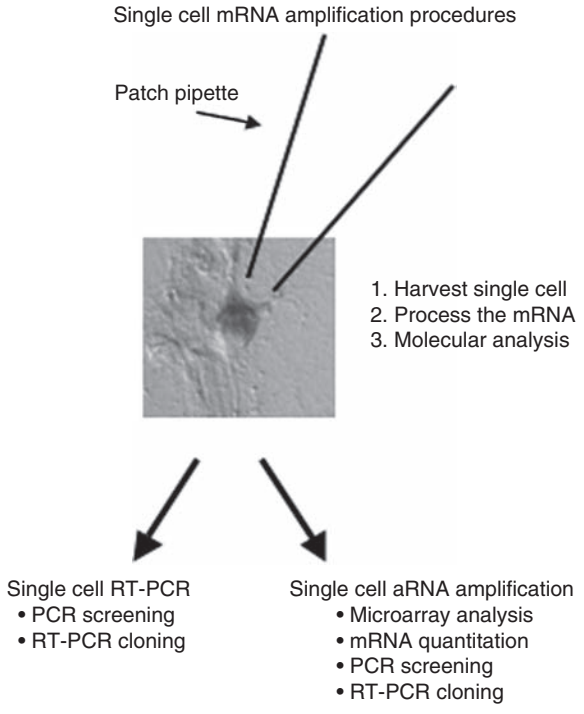

Figure 1 Schematic of single cell mRNA procedures. Individual cells can be collected from a culture or tissue by aspirating a single isolated cell into a recording micropipette. Once the single cell is collected, the mRNAs expressed by the cell can be analyzed by either single cell RT-PCR or aRNA amplification. Another single cell methodology, *in situ* hybridization (not shown), is used predominately to analyze mRNA localization in cells.

Application of aRNA Amplification

Single cell aRNA amplification, along with other techniques, can be used for expression profile studies and other approaches. Thus, the RNA amplification technique has been employed for demonstrating the presence of neuronal mRNA species in individual senile plaques in Alzheimer's disease and the presence of mRNA in dendrites and dendrite growth cones. In combination with array platforms, expression profiling demonstrated a decrease in G-protein transcripts, GluR3, GluR5, synaptic-related markers, and signaling molecules from the affected brain area in schizophrenia. In addition, there was senescence-associated cell type-specific downregulation in dopamine receptor mRNA in brain regions related to learning and memory. The combination of patch clamp recording and aRNA amplification methods in single dentate granule cells from an animal model of temporal lobe epilepsy allows the demonstration of a different expression profile in $GABA_A$ receptor subunits correlating with alterations in GABAergic function that precede the epileptic episode. Gene expression profile or any other approach using the single aRNA amplification technique can be performed in other neurological disorders in order to gain insight into the molecular basis of the diseases and then help develop new strategies to prevent or cure them. On the other hand, these techniques have been used in tissues different from the brain with positive results.

Future Directions

The regulation of gene expression eukaryotes can occur at several different levels: transcriptional, posttranscriptional (alternative splicing, mRNA localization, and mRNA stability), translational, and posttranslational. All of these processes can be regulated in a cell-specific manner. In neuronal tissues, studying gene expression profiles at single cell levels is of particular importance because the nervous system contains thousands of functionally distinct subsets of neurons. These neuronal populations express specific subsets of genes that underlie their unique functional properties. Studying gene expression at the cellular level with single cell molecular analysis procedures will lead to a better understanding of the molecular mechanisms

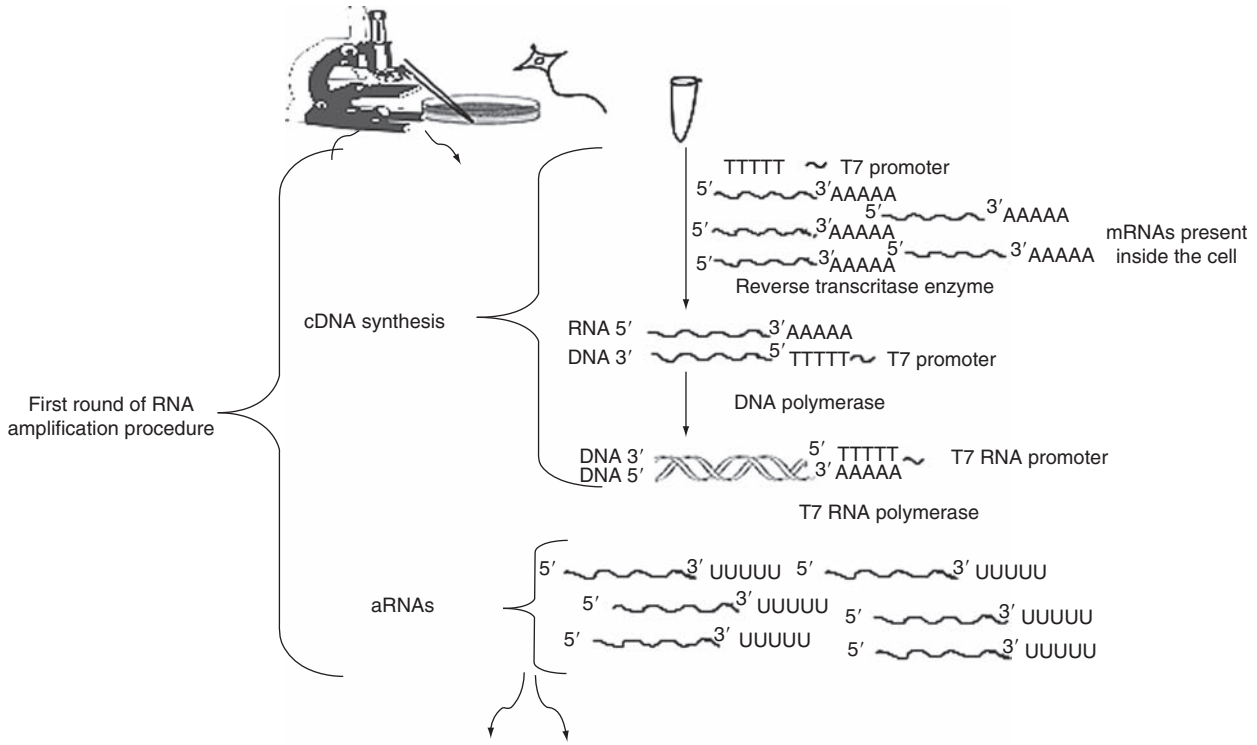

Figure 2 RNA amplification procedure in single cells. A single cell is isolated from primary culture using a patch pipette or by laser capture microdissection and then transferred into a tube. An oligo-dT primer containing a T7 RNA polymerase promoter site is hybridized into poly A tail of mRNAs present in the cell and RNA is reverse transcribed into cDNA. As a result, a double-stranded mRNA–DNA hybrid is obtained. Then, a second strand of DNA is synthesized by using DNA polymerase. Subsequently, T7 RNA polymerase mediates transcription from DNA that contains a T7 RNA polymerase promoter site, creating amplified antisense RNA copies. At this point, further rounds of amplification can be performed.

that underlie functional diversity in neuronal cell types. In addition to a better understanding of the molecular basis of neuronal function, the identification of cell-specific mRNAs and splice variants may be clinically relevant to human disease states, such as ion channelopathies. Cell-specific mRNAs and splice variants in functionally distinct sets of neurons may be useful as potential therapeutic targets for neurological disorders.

See also: Axonal mRNA Transport and Functions; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; Single Cell Genomic DNA Analysis; Single Cell Electroporation; Single Cell PCR Coupled with Electrophysiology.

Further Reading

Bell TJ, Thaler C, Castiglioni AJ, Helton TD, and Lipscombe D (2004) Cell-specific alternative splicing increases calcium channel current density in the pain pathway. *Neuron* 41: 127–138.

Brooks-Kayal AR, Shumate MD, Jin H, Rikhter TY, and Coulter DA (1998) Selective changes in single cell GABA(A) receptor subunit expression and function in temporal lobe epilepsy. *Nature Medicine* 4: 1166–1172.

Buongiorno-Nardelli M and Amaldi F (1970) Autoradiographic detection of molecular hybrids between RNA and DNA in tissue sections. *Nature* 225: 946–948.

Crino P, Khodakhah K, Becker K, Ginsberg S, Hemby S, and Eberwine J (1998) Presence and phosphorylation of transcription factors in developing dendrites. *Proceedings of the National Academy of Sciences of the United States of America* 95: 2313–2318.

Crino PB, Trojanowski JQ, Dichter MA, and Eberwine J (1996) Embryonic neuronal markers in tuberous sclerosis: Single-cell molecular pathology. *Proceedings of the National Academy of Sciences of the United States of America* 93: 14152–14157.

Dixon AK, Richardson PJ, Pinnock RD, and Lee K (2000) Gene-expression analysis at the single-cell level. *Trends in Pharmacological Sciences* 21: 65–70.

Eberwine J, Belt B, Kacharmina JE, and Miyashiro K (2002) Analysis of subcellularly localized mRNAs using *in situ* hybridization, mRNA amplification, and expression profiling. *Neurochemical Research* 27: 1065–1077.

Eberwine J, Yeh H, Miyashiro K, et al. (1992) Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America* 89: 3010–3014.

Eberwine JH, Spencer C, Newell D, and Hoffman AR (1993) mRNA structure, *in situ*, as assessed by microscopic techniques. *Microscopy Research and Technique* 25: 19–28.

Fettiplace R and Fuchs PA (1999) Mechanisms of hair cell tuning. *Annual Review of Physiology* 61: 809–834.

Fostel J, Narayanswami S, Hamkalo B, Clarkson SG, and Pardue ML (1984) Chromosomal location of a major tRNA gene cluster of *Xenopus laevis*. *Chromosoma* 90: 254–260.

Gall JG and Pardue ML (1969) Formation and detection of RNA–DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America* 63: 378–383.

Ginsberg SD, Crino PB, Hemby SE, et al. (1999) Predominance of neuronal mRNAs in individual Alzheimer's disease senile plaques. *Annals of Neurology* 45: 174–181.

Hemby SE, Ginsberg SD, Brunk B, Arnold SE, Trojanowski JQ, and Eberwine JH (2002) Gene expression profile for schizophrenia: Discrete neuron transcription patterns in the entorhinal cortex. *Archives of General Psychiatry* 59: 631–640.

Hemby SE, Trojanowski JQ, and Ginsberg SD (2003) Neuron-specific age-related decreases in dopamine receptor subtype mRNAs. *Journal of Comparative Neurology* 456: 176–183.

Hudspeth AJ (1989) How the ear's works work. *Nature* 341: 397–404.

Jones EM, Gray-Keller M, Art JJ, and Fettiplace R (1999) The functional role of alternative splicing of Ca(2+)-activated K^+ channels in auditory hair cells. *Annals of the New York Academy of Sciences* 868: 379–385.

Lambolez B, Audinat E, Bochet P, Crepel F, and Rossier J (1992) AMPA receptor subunits expressed by single Purkinje cells. *Neuron* 9: 247–258.

Mackler SA and Eberwine JH (1993) Diversity of glutamate receptor subunit mRNA expression within live hippocampal CA1 neurons. *Molecular Pharmacology* 44: 308–315.

Middleton FA, Mirnics K, Pierri JN, Lewis DA, and Levitt P (2002) Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *Journal of Neuroscience* 22: 2718–2729.

Mirnics K, Middleton FA, Marquez A, Lewis DA, and Levitt P (2000) Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron* 28: 53–67.

Miyashiro KY, Beckel-Mitchener A, Purk TP, et al. (2003) RNA cargoes associating with FMRP reveal deficits in cellular functioning in Fmr1 null mice. *Neuron* 37: 417–431.

Navaratnam DS, Bell TJ, Tu TD, Cohen EL, and Oberholtzer JC (1997) Differential distribution of Ca^{2+}-activated K^+ channel splice variants among hair cells along the tonotopic axis of the chick cochlea. *Neuron* 19: 1077–1085.

O'Dell DM, Raghupathi R, Crino PB, Morrison B III, Eberwine JH, and McIntosh TK (1998) Amplification of mRNAs from single, fixed, TUNEL-positive cells. *Biotechniques* 25: 566–568, 570.

Pardue ML and Dawid IB (1981) Chromosomal locations of two DNA segments that flank ribosomal insertion-like sequences in *Drosophila*: Flanking sequences are mobile elements. *Chromosoma* 83: 29–43.

Rosenblatt KP, Sun ZP, Heller S, and Hudspeth AJ (1997) Distribution of Ca^{2+}-activated K^+ channel isoforms along the tonotopic gradient of the chicken's cochlea. *Neuron* 19: 1061–1075.

Steward O and Worley PF (2001) Selective targeting of newly synthesized Arc mRNA to active synapses requires NMDA receptor activation. *Neuron* 30: 227–240.

Todd R and Margolin DH (2002) Challenges of single-cell diagnostics: Analysis of gene expression. *Trends in Molecular Medicine* 8: 254–257.

VanGelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, and Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences of the United States of America* 87: 1663–1667.

Zola-Morgan S and Squire LR (1993) Neuroanatomy of memory. *Annual Review of Neuroscience* 16: 547–563.

Zola-Morgan S, Squire LR, and Ramus SJ (1994) Severity of memory impairment in monkeys as a function of locus and extent of damage within the medial temporal lobe memory system. *Hippocampus* 4: 483–495.

Single Cell Neuronal Circadian Clocks

S Honma and K-i Honma, Hokkaido University Graduate School of Medicine, Sapporo, Japan

© 2009 Elsevier Ltd. All rights reserved.

Neuronal Activity Rhythms of the Suprachiasmatic Nucleus

The suprachiasmatic nucleus (SCN) of the hypothalamus is the site of the mammalian master clock, where individual neurons exhibit robust circadian firing rhythms. Electrical activity of SCN neurons has been measured *in vitro* using acutely prepared SCN slices from adult animals as well as *in vivo* by monitoring multiunit activity (MUA) using chronically implanted electrodes. Irrespective of nocturnality or diurnality of behavioral rhythms, the firing rate of SCN neurons is higher in the subjective day than in the subjective night; however, in most of the brain areas outside the SCN, neuronal activity rhythms of nocturnal and diurnal animals are 180° out of phase with each other.

Measuring circadian periodicity of SCN neurons requires analysis of the order of days or even weeks. Nevertheless, in traditional slice physiology, duration of a single-unit activity recording was limited to the order of minutes, which is hardly sufficient for circadian rhythm analysis. Therefore, until recently, it had been difficult to demonstrate whether the circadian rhythm was driven by single pacemaker neurons or emerged from neural networks within the SCN. This problem was solved by the use of the multielectrode array dish (MED) system. In addition, bioluminescent and fluorescent reporters enabled us to monitor gene expression rhythms in single SCN cells for days and weeks.

Circadian Periods of Single-SCN Neuronal Rhythm

Most SCN neurons in nocturnal rodents exhibit robust circadian rhythms in their spontaneous discharges (Figure 1), and their periods were specific to each neuron. The percentage of SCN neurons that exhibit significant circadian periodicity depends on the culture methods. In wild-type rats, about 75% of SCN neurons exhibit significant circadian periodicity in their spontaneous discharges in the dispersed cell culture, while nearly 85% exhibit in the organotypic slice culture. The percentage of rhythmic neurons is significantly higher in the slice culture than in the dispersed cell culture. In the dispersed cell culture, many SCN neurons exhibit circadian firing rhythms with individually different periods, indicating that the SCN is comprised of multiple autonomous single-neuronal oscillators. However, some of them seem to exhibit circadian firing rhythms driven by other neurons. The range of circadian period distribution of single-SCN neurons also depends on culture methods. In rats, the periods of dispersed neurons vary between 20.0 and 30.9 h with an average period of 24.1 ± 1.4 h (mean $\pm$ SD), while those from slice culture vary between 22.4 and 26.7 h, with an average of 24.2 ± 0.7 h. In both cultures, circadian periods follow a Gaussian distribution, and their mean circadian periods are not significantly different from each other. However, the distribution range of circadian period is significantly different between the two culture systems (Figure 2). Circadian periods of single-SCN neuronal rhythms are widely varied in dispersed neurons even within a single culture dish, whereas they are basically synchronized in a single-SCN slice culture. Therefore, the SCN cell assemblage for cell-to-cell interactions is critical for coherent circadian rhythm expression from the SCN. Not only the intact SCN cell assemblage but also the intrinsic circadian period is suggested to play an important role in synchronizing the constitutional cellular oscillators in the SCN. Compared to the cultured neurons, the free-running periods of behavioral rhythms in rats are distributed in a strikingly narrow range of 24.0–24.8 h with an average of 24.4 ± 0.2 h (Figure 2). The entire SCN structure with intact afferents and efferents seems to be necessary for the SCN neurons to exhibit rhythms with physiological periodicity.

Regional Specificity and Neuronal Properties

The SCN is functionally and anatomically organized into two subdivisions: ventrolateral (core) and dorsomedial subdivisions (shell). The former lies adjacent to the optic chiasm and comprises predominantly neurons producing vasoactive intestinal peptide (VIP) and gastrin-releasing peptide (GRP). The core receives major afferents to the SCN from the retina and midbrain, whereas the shell surrounds the core where arginine vasopressin (AVP)-containing neurons are predominated. In addition, basically all SCN neurons produce γ-aminobutyric acid (GABA), a major inhibitory neurotransmitter in the central nervous system. The percentage of neurons exhibiting significant circadian firing rhythm is higher in the dorsomedial subdivision than in the ventrolateral. Dissociation of

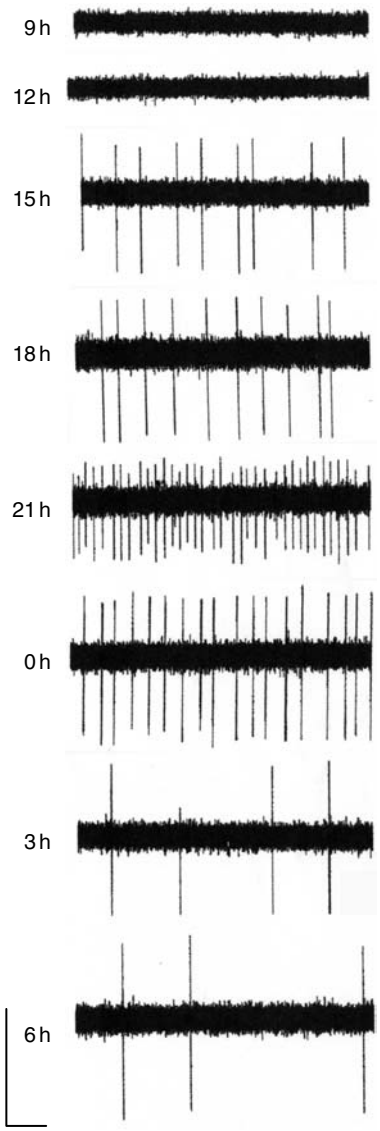

Figure 1 Circadian changes in spontaneous discharges of a rat SCN neuron. Local time of recording was indicated in the left of each record. Vertical and horizontal bars represent 50 mV and 1 s, respectively. Adapted from Honma S, Namakura W, Shirakawa T, et al. (2005) Application of a multielectrode array dish to chronobiology: Monitoring the circadian firing rhythm of single SCN neurons. In: Honma K and Honma S (eds.) *Biological Rhythms*, pp. 91–104. Sapporo: Hokkaido University Press.

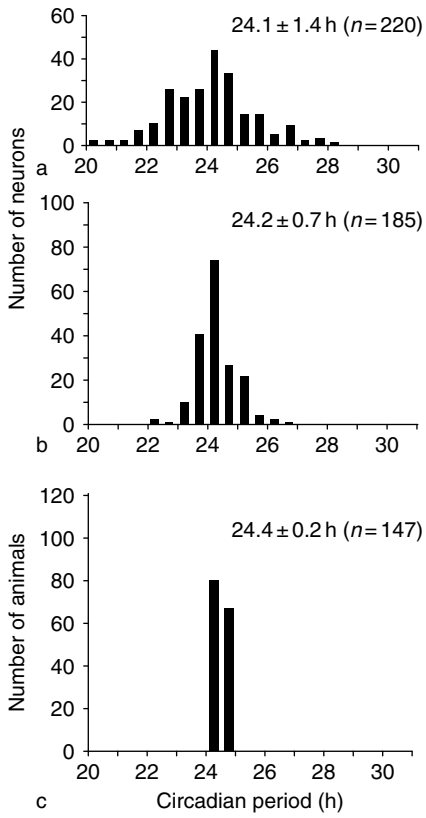

Figure 2 Distribution of circadian period of single SCN neurons. Numbers of neurons in the dispersed cell culture (a); and slice culture (b). Circadian periods of spontaneous locomotor activity in individual rats (c) are expressed in histograms. Numbers are mean ± SD. Reprinted from *Neuroscience Letters,* 358, Honma S, Nakamura W, Shirakawa T, et al., Diversity in the circadian periods of single neurons of the rat suprachiasmatic nucleus depends on nuclear structure and intrinsic period, 173–176, Copyright 2004, with permission from Elsevier.

the circadian rhythms in the SCN evidentially supports the existence of two subdivisional pacemakers in the SCN. Such dissociations have been observed in the rhythms in AVP and VIP releases from cultured SCN slices and clock gene expression rhythms *in vivo* after the abrupt shift of light–dark (LD) cycles. Interestingly, dissociation of clock gene rhythms following LD shift does not accompany similar dissociation in the MUA rhythms. The neural networks within the SCN seem to connect the two subdivisional rhythms, because surgical dissection dissociates them. Therefore, with intact neuronal connections, single neuronal rhythms we can measure are the integrated outputs of their own intracellular clocks and extracellular rhythmic signals.

The two oscillatory components within the SCN are also observed in relation to the photoperiodic responses in circadian rhythms. In the coronal SCN slices, the MUA always exhibits unimodal circadian rhythm with a peak in the subjective day. However, two distinct peaks are detected at the subjective dawn and dusk in the MUA from horizontal SCN slices of hamsters, the phases of which depend on day length. A long-standing hypothesis for photoperiodic responses of the behavioral rhythms in nocturnal rodents is mutually coupled two oscillators, evening (E) and morning (M) ones, which separately track dawn and dusk. The horizontal SCN slices have entire rostrocaudal interactions yet lack the dorsal SCN. One of the E and M oscillators may be lacking in coronal slices, while both of them still exhibit

each peak in horizontal slices. Even under steady entrainment to an LD cycle, cellular clocks seem to be differentially phased in different subdivisions of the SCN.

Mechanisms Generating Circadian Rhythms in Spontaneous Action Potentials

In addition to the spontaneous firing, SCN neurons exhibit circadian rhythms in various functions such as neuropeptide release, metabolic rate, and intracellular Ca^{2+} levels. Current understanding of the intracellular molecular machinery for circadian rhythm generation consists of interlocked transcriptional and translational feedback loops involving several clock genes, as positive elements, *Clock* and *Bmal1*, and as negative ones, *Per1*, *Per2*, *Cry1*, and *Cry2*. Various kinds of clock gene-deficient as well as mutant mice have been developed for the study of molecular clock mechanisms, and the results support these molecular feedback loops. However, the intracellular signal transduction mechanisms from the molecular core to cellular functions are basically unknown. Although SCN neurons possess various kinds of ion channels, the target channels of the cellular clock and the mechanisms of circadian regulation remain to be studied. Substantial data support the hypothesis that voltage-dependent calcium channels contribute to the clock function. Diurnal modulation in the calcium currents may suggest that the channels are the targets of intracellular circadian signal output. Circadian modulation is also observed in Na^+ channels, and Ca^{2+} gated and voltage-gated K^+ channels. Almost all SCN neurons contain GABA, and $GABA_A$ receptors are also found in SCN neurons. However, controversial data have been reported on the effects of GABA in the SCN, whether it is an inhibitory transmitter throughout the day or as an excitatory and inhibitory transmitter depending on the circadian phase. These conflicts may be partly due to the dose of the drug application. In addition, different types of SCN neurons may respond to GABAergic transmission differently. Neuronal activities are often measured in a limited number of neurons without identifying their localization within the SCN and types of neurotransmitters, channels, or receptors. SCN neurons are functionally as well as anatomically heterogenous. NMDA and 5-HT receptors are localized in the ventrolateral subdivision. The dorsal neurons lack K^+–Cl^- cotransporter 2, which results in elevated intracellular Cl^- concentration. Although it is apparent that the SCN is composed of multiple autonomous single-neuronal oscillators, they are not homogenous, especially in their response to the external signals.

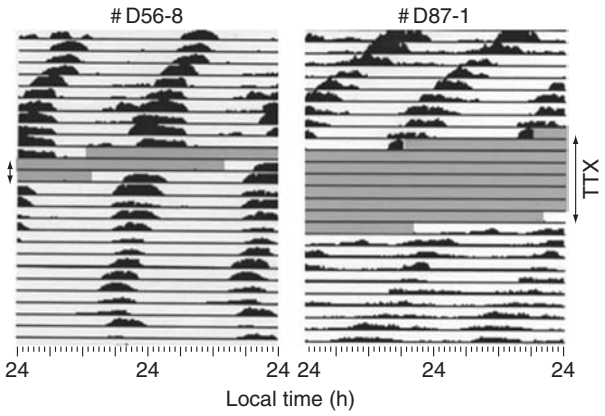

Figure 3 Effects of TTX on the firing rhythms of single SCN neurons. Double plotted firing rhythms of single SCN neurons treated with 200 nM TTX for 1 (#D56-8, full scale 3 Hz) and 7 days (#D87-1, full scale 1.5 Hz). Arrows and shaded areas indicate days in TTX treatment.

Administration of tetrodotoxin (TTX), a sodium channel blocker, into the SCN suppresses the spontaneous firing of SCN neurons *in vitro*, and behavioral rhythms *in vivo*. However, TTX does not terminate the intracellular circadian oscillation (Figure 3). Circadian firing rhythms reappear after the TTX washout from the phase predicted by the pretreatment rhythm. Furthermore, neither clock gene expression rhythm nor intracellular Ca^{2+} rhythm in each SCN neuron is affected by TTX. TTX does block synaptic interaction among SCN neurons, which eventually results in the dissociation of the SCN neuronal rhythms. On the other hand, membrane hyperpolarization reversibly abolishes the clock gene expression rhythms. Therefore, membrane phenomena such as electrical impulses and ionic currents of the SCN cell membrane are not only output or input for the intracellular clock machinery but may also participate in the rhythm generation mechanisms.

Circadian Signal Communication between SCN Neurons

The SCN is composed of a number of neurons with slightly but significantly different circadian periods. Therefore, it is critical for them to communicate with each other in order to synchronize individual rhythms for the circadian output from the SCN. The SCN *in vivo* receives daily photic inputs via the retinohypothalamic tract, which reset the period of individual cellular oscillators to 24 h. Even without daily entraining cues, individual neuronal rhythms synchronize with each other in the SCN. At least three modes of communication patterns have been reported for SCN neurons to convey circadian information – chemical synapses, gap junctions, and diffusible

factors. Cross-correlation analyses of spike timings is useful to determine the existence of synaptic coupling and, if so, the mode of transmission (Figure 4). Both excitatory and inhibitory connections have been observed between neurons exhibiting synchronized firing rhythms in dispersed cell culture. On the other hand, no functional connections are detected in neuron pairs with different circadian periods. Sodium-dependent action potential exerts its influence as one of the synchronization signals. Recently, VIP, a predominant neurotransmitter in the core, has been shown to exert as a synchronization signal among cellular oscillations in the SCN. Type 2 VIP receptor (VPAC2) of the SCN neurons receives these synchronizing signals, since VPAC2 knockout mice have no circadian output due to the desynchronization among multiple cellular oscillators. In addition to chemical synapses, gap junctions are found in SCN neurons as well as in astrocytes. Presence of connexin 32 and dye coupling, especially during daytime in the SCN, also suggests gap-junctional communication among SCN neurons.

Imaging of Single-SCN Cellular Rhythms and Clock Gene Dysfunction

Transcriptional and translational feedback loops involving clock genes and their protein products are the molecular machinery for intracellular circadian rhythm generation. Bioluminescent reporters are successfully introduced to long-term monitoring of clock gene expression or their protein products in the cultured SCN. Especially, clock functions in individual cells are assessed using real-time imaging of circadian gene expression in cultured SCN of transgenic mice carrying a *Per1* promoter-driven luciferase reporter gene. Most SCN cells exhibit robust and synchronized circadian rhythms in their clock gene expression. TTX treatment desynchronizes cellular circadian gene expression rhythms. Cycloheximide, a protein synthesis inhibitor, stops the intracellular circadian oscillation. Clock gene dysfunctions can also be examined in single SCN cells. Individual cellular rhythms of *Cry1*- and *Cry2*-deficient mice exhibit no circadian fluctuation in *Per1* expression, whereas desynchronized is observed among SCN cells of VPAC2-deficient mice, although both of them are behaviorally arrhythmic in constant darkness.

Mouse mutant in the gene *Clock* are known to exhibit altered circadian rhythms. Under constant darkness (DD), heterozygotes exhibit circadian behavioral rhythms with periods longer than those in wild-type mice, and homozygotes eventually become arrhythmic (Figure 5). However, in the organotypic SCN slice culture, robust circadian firing rhythms are observed in 77% neurons of the homozygotes examined. The percentage is still smaller than that in the wild-type SCN neurons of 95%. Notable is a dramatic decrease of neurons exhibiting circadian firing rhythm in dispersed cell culture (only 15% and 46% in homozygote and wild-type mice, respectively). In addition, the circadian firing rhythms in the dispersed cell culture were much unstable in *Clock* mutants than wild-type mice, indicating that

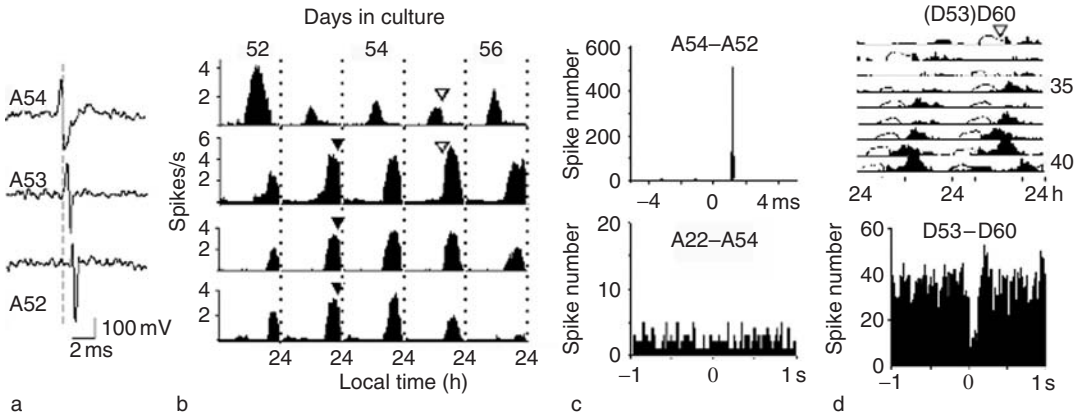

Figure 4 Cross-correlation analyses of single SCN neuronal rhythms. (a) Spike discharges from three electrodes. Synchronized circadian firing rhythms among three neurons (A53, A53, and A54) and one desynchronized (A22) exhibiting a different phase and a period from the other three. (b) Open triangles indicate the time of cross-correlation analysis. Numbers on the top indicate days in culture. (c) Cross-correlograms showing an excitatory communication between A54 and A52 (upper graph), and a lack of functional communication between A22 and A54 (lower graph). (d) Double-plotted circadian firing rhythms of D53 (outline) and D60 (filled) with the same period of 23.6 h and phase differences of about 6 h (upper graph), and a cross-correlogram between them (lower graph). Numbers in the right margin indicate days in culture. An open triangle indicates the time of cross-correlation analysis. Adapted from Shirakawa T, Honma S, Katsuno Y, et al. (2000) Synchronization of circadian firing rhythms in cultured rat suprachiasmatic neurons. *European Journal of Neuroscience* 12: 2833–2838.

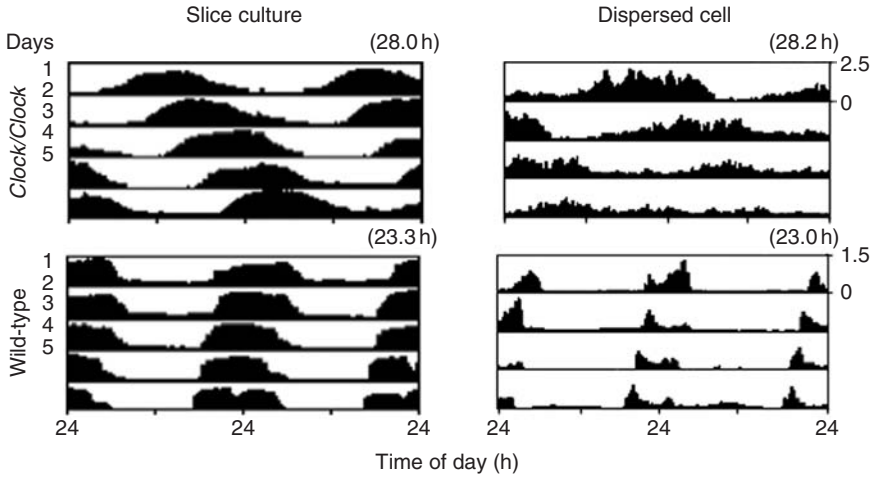

Figure 5 Spontaneous firing rhythms of the SCN neurons in *Clock* mutant mice. Double-plotted firing rhythms of the SCN neurons from *Clock* mutant mice (*Clock*/*Clock*) and wild-type mice in slice culture and dispersed cell culture. Circadian period of each neuron is indicated on the top right of each rhythm. Adapted from Nakamura W, Honma S, Shirakawa T, et al. (2002) *Clock* mutation lengthens the circadian period without damping rhythms in individual SCN neurons. *Nature Neuroscience* 5: 399–400.

circadian firing rhythms in single-SCN neurons are affected by genotypes as well as culture system. Similar results are obtained in luciferase reporter activity rhythms in the slice and dispersed cultures of the SCN from *Cry*- or *Per*-deficient mice. These results suggest that intense cell-to-cell communication in the SCN increases the stability of single-neuronal oscillators to exhibit robust circadian rhythms to orchestrate peripheral clocks throughout the body. The intercellular interaction is especially important for mice having genetic dysfunctions in clock system.

See also: Circadian Organization; Circadian Rhythm Models; Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Rhythms: Influence of Light in Humans; Sleep: Development and Circadian Control.

Further Reading

Albus H, Vansteensel MJ, Michel S, et al. (2005) A GABAergic mechanism is necessary dissociable ventral and dorsal regional oscillators within the circadian clock. *Current Biology* 15: 886–893.

Aujard F, Herzog ED, and Block GD (2001) Circadian rhythms in firing rate of individual suprachiasmatic nucleus neurons from adult and middle-aged mice. *Neuroscience* 106: 255–261.

Herzog ED, Aton SJ, Numano R, et al. (2004) Temporal precision in the mammalian circadian system: A reliable clock from less reliable neurons. *Journal of Biological Rhythms* 19: 35–46.

Herzog ED, Takahashi JS, and Block GD (1998) Clock controls circadian period in isolated suprachiasmatic nucleus neurons. *Nature Neuroscience* 1: 708–713.

Honma S, Namakura W, Shirakawa T, et al. (2003) Cellular oscillators in the duprachiasmatic nucleus: Studies of cultured SCN on a multi-electrode dish. In: Honma K and Honma S (eds.) *Circadian Clock as Multi-Oscillation System*, pp. 155–166. Sapporo: Hokkaido University Press.

Honma S, Namakura W, Shirakawa T, et al. (2006) Monitoring the clock neuron's tick: Circadian rhythm analysis using a multi-electrode dish. In: Taketani M and Baudry M (eds.) *Advances in Network Electrophysiology, Using Multi-Electrode Arrays*, pp. 409–424. New York: Springer.

Honma S, Shirakawa T, Katsuno Y, et al. (1998) Circadian periods of single suprachiasmatic neurons in rats. *Neuroscience Letters* 250: 157–160.

Liu C, Weaver DR, Strogatz SH, et al. (1997) Cellular construction of a circadian clock: Period determination in the suprachiasmatic nuclei. *Cell* 91: 855–860.

Maywood E, Reddy AB, Wong GKY, et al. (2006) Synchronization and maintenance of timekeeping in suprachiasmatic circadian cock cells by neuropeptidergic signaling. *Current Biology* 16: 599–605.

Nakamura W, Honma S, Shirakawa T, et al. (2001) Regional pacemakers composed of multiple oscillator neurons in the rat suprachiasmatic nucleus. *European Journal of Neuroscience* 14: 666–674.

Nakamura W, Honma S, Shirakawa T, et al. (2002) *Clock* mutation lengthens the circadian period without damping rhythms in individual SCN neurons. *Nature Neuroscience* 5: 399–400.

Shirakawa S, Honma S, Katsuno Y, et al. (2000) Synchronization of circadian firing rhythms in cultured rat suprachiasmatic neurons. *European Journal of Neuroscience* 12: 2833–2838.

Welsh DK, Logothetis DE, Meister M, et al. (1995) Individual neurons dissociated from rat suprachiasmatic nucleus express independently phased circadian firing rhythms. *Neuron* 14: 697–706.

Yamaguchi S, Isejima H, Matsuo T, et al. (2003) Synchronization of cellular clocks in the suprachiasmatic nucleus. *Science* 302: 1408–1412.

Single Cell PCR Coupled with Electrophysiology

T Tkatch and D J Surmeier, Northwestern University, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Neurons Are Not All the Same

The enormous computational capacity of the brain is in large measure a reflection of its cellular heterogeneity. Most brain nuclei or regions are composed of complex mixtures of neurons that differ in many ways. A powerful array of tools has been developed during the past century to explore this diversity in an attempt to understand how it shapes brain function in health and disease. Most of these approaches, however, are limited to one aspect of the cellular phenotype. For example, traditional anatomical techniques that utilize light or electron microscopy are limited to revelations about neuronal structure. *In situ* hybridization and immunocytochemical techniques provide information about gene and protein expression but little about structure or function.

A major leap forward occurred approximately 30 years ago when electrophysiological techniques were combined with intracellular labeling that allowed the function of a neuron to be correlated with its morphology. Although this approach has provided important clues about how structure influences function, its principal value is in cellular identification; that is, it allowed a dynamic, physiological signature or phenotype to be connected to a more static morphological phenotype.

With the genomics 'revolution' has come the opportunity to determine how differences in gene expression are translated into functional differences. Conventional *in situ* hybridization and immunocytochemical approaches only allow the co-localization of one or two genes in any one particular cell. Moreover, both approaches are difficult to make quantitative. An alternative approach combines patch clamp recording with reverse transcription (RT) and amplification of the complementary DNA (cDNA), allowing gene expression and physiology to be linked. There are two ways in which cDNA from single cells is typically amplified. One uses a T7 RNA polymerase strategy. The other uses the polymerase chain reaction (PCR). The strengths and weaknesses of this so-called single cell RT-PCR (scRT-PCR) approach are discussed here. This technique has been used extensively during the past decade to generate expression profiles from electrophysiologically characterized individual neurons.

Harvesting Cellular mRNA

Cellular mRNA can be harvested from patch clamped neurons that have either been acutely isolated or been accessed *in situ*, usually in tissue slices (although *in vivo* patch clamping is becoming more common). The patch electrode, with its relatively large diameter tip (~0.5 μm), provides an excellent means of accessing cytoplasmic mRNA.

Neurons can be acutely dissociated from tissue slices by a variety of methods. Dissociation generates a suspension of individual neurons that is plated into a sterile recording chamber and allowed to settle (Figure 1). These isolated neurons have short dendritic and axonal processes, making them excellent subjects of voltage clamp studies aimed at characterizing ion channel gating and modulation. Following electrophysiological study, cells can be harvested for scRT-PCR profiling. When harvesting acutely isolated neurons, the entire cell (cytoplasm and nucleus) is sucked up into the patch pipette by applying negative pressure to the back of the pipette/electrode. It is critical in these experiments to approach cells for patching with positive pressure on the electrode to prevent harvesting cellular debris that might contain mRNA. It is also important during the harvesting phase that the cell be lifted off the bottom of the chamber where debris has settled; the best strategy is to lift the cell with the patch electrode into a stream of sterile saline solution, aspirate the cell into the pipette, and then quickly remove the electrode from the chamber solution. All of these steps must be visually monitored.

Neurons in tissue slices also have been subjected to scRT-PCR analysis following patch clamp recording. This preparation is better suited to questions about the synaptic and integrative properties of neurons (for example, how synaptic currents are translated into spike trains). However, there are several obstacles to reliable harvesting of mRNA from neurons in tissue slices. The first is that the recording/harvesting electrode can easily be contaminated with mRNA from surrounding cells as it is driven through the tissue to the desired cell. Positive pressure must be maintained on the electrode during the descent into the tissue. The second obstacle is related to the first. During the harvesting procedure, it is critical that the high-resistance seal that forms between the electrode and the cell be maintained to prevent tissue surrounding the cell from being harvested. This is accomplished by monitoring seal resistance while applying modest negative pressure to the back of the electrode for 5–10 min at the conclusion of the recording phase,

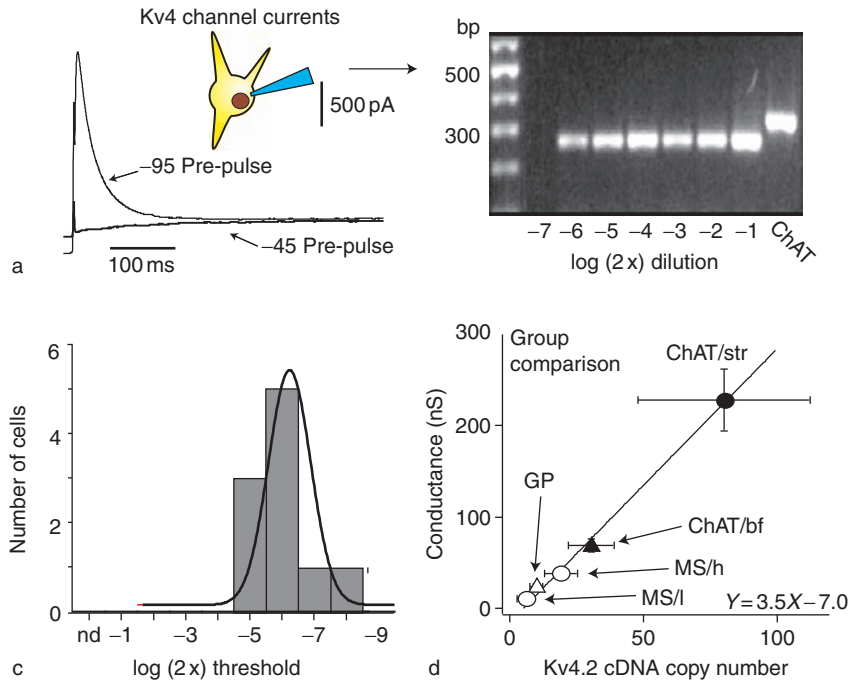

Figure 1 Somatodendritic A-type conductance in striatal cholinergic (expressing choline acetyltransferase (ChAT) mRNA) neurons is directly correlated with estimates of Kv4.2 mRNA abundance assessed using scRT-PCR. (a) Low-threshold, TEA-resistant A-type current was recorded before aspirating cell cytoplasm. (b) After serial dilution of RT product, PCR targeting Kv4.2 was performed. (c) The threshold distribution from a sample of neurons was fit with a Gaussian function. (d) Average maximal conductance of A-type channels was plotted against the mean estimated number of copies of Kv4.2 cDNA for five cell groups: striatal cholinergic interneurons (ChAT/str), basal forebrain cholinergic neurons (ChAT/bf), high and low expressing GABAergic striatal medium spiny neurons (MS/h and MS/l, respectively), and GABAergic globus pallidus neurons (GP).

bringing cytoplasm (and mRNA) into the pipette. Next, the electrode must be slowly withdrawn to form an outside-out patch, capping the harvesting electrode. Unlike the situation with acutely isolated cells in which the entire cytoplasm and nucleus are harvested, only a limited amount of cytoplasm is obtainable from neurons in slices without compromising cellular specificity. Another issue with this approach is that larger bore pipettes that do a better job of harvesting often compromise the electrophysiological recording because the cytoplasm is too well dialyzed; this is particularly problematic for small neurons. In this case, harvesting can be performed with a second pipette at the conclusion of the recording phase.

The amount of mRNA within a single vertebrate neuron is estimated to be between 0.1 and 1 pg, making the detection of mRNA transcripts a technical challenge, even with the most efficient harvesting. This makes it critical that as near an RNase-free environment be maintained during harvesting as possible. Several steps are helpful. RNase-free gloves should be worn during preparation and handling of electrodes, recording solutions, and harvested material. Electrodes should be baked to denature RNases. Silanes can be used to minimize binding of mRNAs to the interior wall of the harvesting pipette. The solutions, particularly the harvesting solution placed in the pipette, should be free of RNase contamination; RNase inhibitors can be included in the harvesting solution. Electrode holders and wires need to be decontaminated between harvesting neurons to avoid cross-contamination. Although degradation of mRNAs can be reduced during harvesting, it cannot be completely stopped, leading to diminished profiling fidelity with prolonged electrophysiological recording.

The efficiency of the RT step, leading to the conversion of mRNA to more stable cDNA, is critical to the fidelity of profiling. There are several steps that need to be taken to maximize the quality of profiling. First, the volume of the solution in the harvesting pipette needs to be carefully controlled (~10 μl). The ionic composition of this solution, particularly the Mg^{2+} concentration, needs to be taken into account (and compensated for) in the RT step because the enzymes involved are very sensitive to divalent ion concentration. Efficient priming of RT is important. Three types of primers can be used. Most commonly, oligo-dT primers are used that bind to the poly A tail of mature mRNA. This prevents reverse transcription of immature heteronuclear mRNA and ribosomal RNA, which can be a major fraction of the harvested RNA. The disadvantage of this priming strategy is

that the reverse transcriptase must work through the 3′ untranslated region (UTR) of the mRNA before it gets to the coding sequences. The 3′ UTR can be very long for some mRNAs and possess secondary structure that can lead to premature termination of the RT reaction. An alternative strategy is to use random hexamers to prime the RT reaction. This does not limit reverse transcription to mature mRNA (or to mRNA). A third RT priming strategy is to use gene-specific antisense primers that target specific mRNAs. This strategy is less common because it limits subsequent amplifications to these mRNAs.

Positive controls should be run for diagnostic purposes in every experiment. The simplest control is to spike an RT sample with a highly diluted RNA from whole brain or a tissue known to express a gene of interest. Because of potential contaminants in mRNA harvesting, solutions, and instruments (e.g., pipettors), it is crucial to also run negative controls with each experiment. Two types of negative controls should be performed. First, a sample of the reaction solutions that ostensibly lacks cellular mRNA should be processed to ensure that reaction solutions are not contaminated with PCR products from previous reactions or some other undesirable RNA. Second, cells should be harvested without the RT enzyme to verify that genomic DNA is not amplified in subsequent PCR steps. To reduce the chances of contamination, several precautions should be taken. Separate areas of the lab should be devoted to procedures in which high concentrations of potential contaminants are handled (e.g., tissue RNA preparations and electrophoresis of PCR products). Aerosol-resistant pipette tips are an absolute prerequisite. Pipettors should be dedicated to one task and frequently cleaned.

Characterization of Cellular mRNAs

Following reverse transcription of cellular mRNAs to cDNA, PCR can be used to amplify specific transcripts. There are two ways to approach such a task. The simplest one is to divide total cellular cDNA and then use each aliquot of cDNA for separate PCR reaction. Although simple, this approach significantly limits the number of cDNAs that can be tested for because each reaction requires a substantial fraction of the total cDNA. An alternative approach is to perform 'multiplex PCR,' in which a combination of primers (one pair for each transcript under investigation) is used in the same PCR reaction. After several cycles of amplification, aliquots of resulting PCR product are used as a template for a second round of PCR amplification with a single pair of transcript-specific primers. Often, the primers used in this second stage of amplification are nested; that is, they target sequences inside the stretch of cDNA targeted by the primary primer set. If targeted transcripts belong to the same gene family, 'degenerate' primers can be used in the first round of PCR. Although appealing because it broadens the single cell profile that can be generated, designing multiplex primers that do not dimerize and yield spurious amplicons is often very difficult.

A variety of factors influence the efficiency of transcript amplification with PCR. One of the most important determinants of success with scRT-PCR is primer design. There are a variety of excellent computer programs available to make this task easier. The length of primers, their homology with target nucleic acids sequences, and guanosine–cytosine content all influence amplification efficiency. In addition, primers should not have any significant amount of homology with other sequences in the targeted transcript or other genes. The primers should not have complementary 5′ and 3′ ends to prevent the formation of hairpin structures or primer dimers, since these can lower amplification efficiency. When possible, primers should span splicing sites to reduce the chances of amplifying genomic DNA.

Correlating mRNA Expression and Cellular Function

In many situations, the goal of scRT-PCR following patch clamp recording is simply cellular identification, much like intracellular labeling. For this type of analysis, the goal is to determine whether a particular mRNA is detectable or not. Because inefficiencies in harvesting, reverse transcription, or amplification can lead to a failure to detect an mRNA that is in fact present, it is important to draw a distinction between detection and actual expression of mRNAs. It is also possible to miss low-abundance transcripts when the cellular cDNA is divided into separate aliquots for amplification. As a consequence, with low-abundance mRNAs, it is important to replicate profiling results to yield estimates of detection probability; without this kind of replication, it is impossible to interpret negative results. Positive controls that target ubiquitously expressed high-abundance mRNAs (e.g., actin) are always a good idea but they do not exclude the possibility that undetected mRNAs are expressed.

Quantitative analysis of mRNA expression in individual neurons can be combined with electrophysiological studies to establish a more compelling association between mRNA expression and physiological function. This approach also helps address the uncertainty about the relationship between mRNA and protein abundance. Posttranscriptional control of mRNA (e.g., stability) and posttranslational processing of protein can influence the RNA/protein

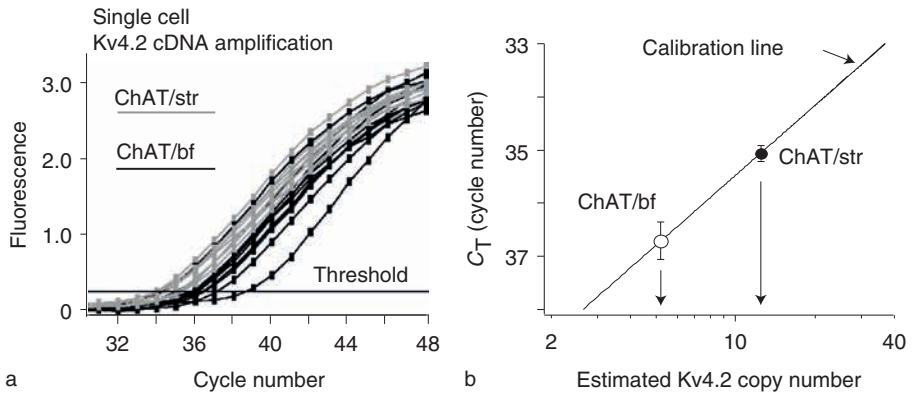

Figure 2 (a) Kv4.2 cDNA was quantified in striatal and basal forebrain ChAT expressing striatal and basal forebrain neurons using real-time PCR detection. Plot of fluorescence as a function of PCR cycle number is shown. (b) Calibration plot generated with known concentrations of plasmid Kv4.2 was used to estimate Kv4.2 cDNA copy numbers.

ratio. Combining quantitative electrophysiological assays of protein abundance and quantitative analysis of mRNA expression provides a direct assessment of this relationship.

A number of strategies have been developed for quantifying cDNA generated from reverse transcription of cellular mRNAs. In competitive or end point analysis PCR, mRNA abundance can be estimated by using carefully designed internal standards. The base sequence of these standards needs to be very close to that of the targeted, comparison transcript to ensure similar amplification efficiency. A single base can be changed to introduce a restriction site that will allow the internal standard to be distinguished after amplification. Alternatively, the standard might contain a small deletion in the region between the two primers, allowing it to be distinguished on the basis of size. A key assumption of this approach is that amplification remains in a linear range. In practice, this is often difficult to ensure. Moreover, this requires that the number of amplification cycles be kept reasonably small (<35 cycles), making detection of amplicons difficult when starting from single cells.

Another rather simple quantitative approach is the serial dilution technique. With this semiquantitative method, a sample is serially diluted until the detection threshold is reached. The more abundant a transcript, the greater the number of dilutions possible before the detection threshold is crossed. Thresholds determined in this way typically have a near normal distribution when sampling is done from a homogeneous population of neurons. Different cell types can then be compared for their relative expression of a particular mRNA. These levels of expression can be compared with a functional measure. For example, the relative abundance of the K^+ channel subunit Kv4.2 was strongly correlated with the amplitude of A-type K^+ currents in a wide variety of brain neurons (Figure 1).

Real-time fluorescence-based PCR has gained in popularity. In this approach, the accumulation of PCR product is monitored by its fluorescence (Figure 2). The abundance of a particular mRNA is estimated from the number of amplification cycles required to distinguish sample fluorescence from background fluorescence. Cloned cDNA can be used to establish calibration curves for absolute quantification. Probes that hybridize to the amplicon (TaqMan) give the most reliable quantitation and also allow for multiplexing. SYBR green (intercalating DNA dye) can also be used if the PCR reaction is optimized to yield only the product to be quantified.

Note that all methods described previously only allow quantification of cDNA molecules, not mRNA. It is often assumed that the efficiency of the RT reaction is perfect, but this is unlikely. RT efficiency depends on the secondary structure of mRNA, priming strategy, and length of the 3′ UTR when using oligo-dT priming. In the absence of appropriate RNA standards, absolute quantitation of mRNA is not possible. Standards used in competitive PCR usually do not have full 3′ UTRs and the efficiency of their transcription is likely to be different from that of native mRNA.

See also: Axonal mRNA Transport and Functions; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; RNA Binding Protein Methods; Single Cell Genomic DNA Analysis; Single Cell Electroporation; Single Cell Molecular Analysis Procedures.

Further Reading

Lambolez B, Audinat E, Bochet P, Crepel F, and Rossier J (1992) AMPA receptor subunits expressed by single Purkinje cells. *Neuron* 9: 247–258.

Liss B and Roeper J (2004) Correlating function and gene expression of individual basal ganglia neurons. *Trends in Neurosciences* 27: 475–481.

Monyer H and Lambolez B (1995) Molecular biology and physiology at the single-cell level. *Current Opinion in Neurobiology* 5: 382–387.

Sucher NJ, Deitcher DL, Baro DJ, Warrick RM, and Guenther E (2000) Genes and channels: Patch/voltage-clamp analysis and single-cell RT-PCR. *Cell and Tissue Research* 302: 295–307.

Tkatch T, Baranauskas G, and Surmeier DJ (2000) Kv4.2 mRNA abundance and A-type current amplitude are linearly correlated in basal ganglia and basal forebrain neurons. *Journal of Neuroscience* 20: 579–588.

Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, and Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences of the United States of America* 87: 1663–1667.

Single Photon Emission Computed Tomography (SPECT)

J K Staley and F Bois, Yale University School of Medicine and VA Connecticut Healthcare System, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Single photon emission computed tomography (SPECT) imaging offers the unprecedented opportunity to visualize specific brain chemicals, including sites of drug action as well as distinct functional states of the living human brain, using radioactive drugs (also known as radiotracers or radiopharmaceuticals). Over the past three decades, significant advances have been made both in the technology and in the development of novel radiopharmaceuticals for imaging of specific neural receptors and transporters in the living human brain using SPECT. These advances have allowed researchers for the first time to visualize the chemical state of the brain; this is important in persons suffering from neuropsychiatric disorders that result from an imbalance of chemicals in the brain, such as alcoholism, Alzheimer's disease, bipolar disorder, cocaine dependence, major depression, Parkinson's disease, schizophrenia, and tobacco dependence. Furthermore, SPECT imaging has permitted the visualization of sites of drug action as well as the quantitation of drug levels in the brain and the relationship with treatment response, thus facilitating the design of drug therapies for the treatment of these devastating brain disorders.

SPECT Radiochemistry

SPECT cameras (Figure 1) measure gamma rays emitted from radioactive atoms, called radionuclides. The SPECT radionuclides, iodine and technetium, ^{123}I and ^{99m}Tc, respectively, emit a single photon at 159 and 140 keV, and have half-lives of 13.2 and 6 h. The longer half-life of ^{123}I allows it to be generated by a central cyclotron facility and then shipped to imaging centers within a 3000-mile radius. ^{99m}Tc is easily produced on site in the local nuclear pharmacy using a molybdenum generator. For SPECT imaging of specific neural receptors, ^{123}I or ^{99m}Tc is attached to a drug (specific for a receptor in the body or brain), resulting in a radioactive drug, commonly called a radiopharmaceutical.

Iodine-123 forms strong covalent bonds with carbon and conveniently can take different chemical forms (metal, anion, cation). Molecular I_2 is treated with an oxidant such as hydrogen peroxide or chloramine T (Bolton–Hunter method) to generate the cation (I^+). I^+ is the reactive form of iodide that is incorporated into the chemical compound by 'exchange' or by 'electrophilic substitution' (Figure 2). The sensitivity of the molecule to oxidation and the labeling kinetics are the parameters that usually define the labeling method used. Since iodine is a fairly large atom, when it is introduced into a compound, the affinity and/or specificity of binding may change dramatically compared to the parent drug. For this reason, preexisting drugs that already have iodine in their structure are obvious prime candidates for development of novel radiotracers.

There is considerable ongoing effort to develop SPECT radioligands that incorporate ^{99m}Tc because of its significantly lower cost and more convenient availability (generated on site by inexpensive molybdenum generators). Several SPECT brain radiotracers have been produced incorporating ^{99m}Tc in their structure. However, its metallic properties (chelation required) and its incorporation generally result in a drastic loss of affinity, and the large size of the ^{99m}Tc moiety significantly reduces penetration of the blood–brain barrier, thus far limiting its use to image neural receptors in brain imaging studies.

Since large interindividual differences in metabolism occur due to genetics, environment, nutrition, and age, the metabolic fate of the radiopharmaceutical must be evaluated after administration to a human. Catabolism (breakdown) of the radiopharmaceutical into metabolites or by-products decreases the concentration in the blood and thus the amount available to enter the brain. To evaluate radiopharmaceutical metabolism, blood samples are obtained at various time points before and after radiotracer administration, and are analyzed to determine the amount of remaining total parent radiotracer, free parent radiotracer, plasma protein binding of the radiotracer, and the lipophilic and polar metabolites. These measurements are used to normalize brain radioactivity levels, to correct for individual differences in metabolism and plasma protein binding.

SPECT Physics

The radiopharmaceutical is most commonly introduced to the individual's body by intravenous injection. The radiopharmaceutical is administered in trace (at least 500-fold lower than the EC_{50}) quantities that do not disturb the biochemical process of interest.

The SPECT image is created by measuring the gamma photons that are released when the radionuclide

Figure 1 Picture of Picker Prism 3000XP SPECT camera.

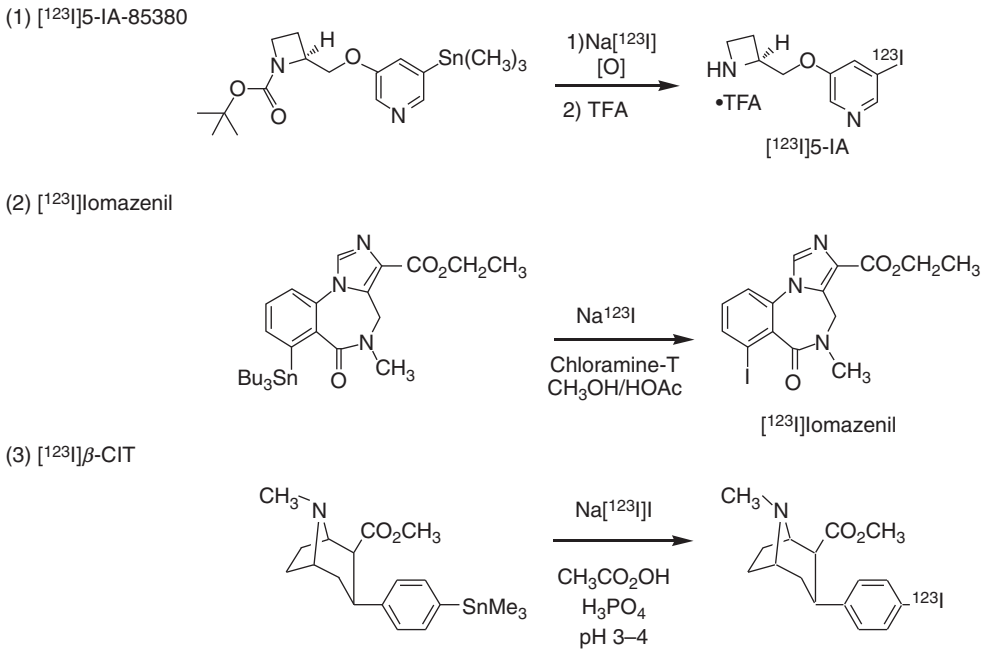

Figure 2 SPECT radiosyntheses.

decays. The high-energy gamma photons pass through the body tissue and bone, and are measured by the SPECT camera. In the SPECT camera, the gamma photon passes through a channel in a lead block called a collimator. Collimators are made from pieces of lead approximately 4–5 cm thick and measuring 20 × 40 cm on the side that contain thousands of channels through which the gamma rays pass. Behind the collimator is a crystal (commonly made of sodium iodide) that releases light in the visible range when a photon interacts with it. Photomultiplier (light-sensitive) tubes adjacent to the crystal collect the emitted light. The holes in the collimator limit the possible origins of the emitted photon, because the lead walls (septae) block all photons that do not enter at a straight angle. SPECT images are produced by computers linked to the SPECT camera (Figure 1); the software processes the light signals and reconstructs the original site in which the activity was emitted, in a slice (or a tomograph), using the principle of back-projection.

Physical Constraints of SPECT

Attenuation

Between the site of origin and the detector in the camera, the gamma photon may travel through many different mediums, including bone, air, tissue, and fluid. Passage through these mediums may reduce the energy intensity of the photon such that it is not detected by the camera. Thus the activity is 'attenuated' by the tissue.

Measurements of absolute tissue activity levels are not accurate without a correction for the attenuated signal. Typically SPECT images have been corrected for attenuation using a first-order approximation in which an ellipse is drawn around the edge of the skull and attenuation is assumed to be uniform and equal to that of water. In general, SPECT activity is attenuated by about 15% per centimeter of path length of the photon. Thus the greater the distance the photon travels to the detector in the camera, the greater the attenuation. This type of attenuation correction, commonly referred to as uniform attenuation correction, does not take into account differences in attenuation that occur between different mediums (bone, air, tissue, fluid), or between individuals with heads that vary in size and shape, for example. Individual differences in attenuation may be corrected by obtaining a transmission scan. The transmission scan, which is similar to a computed tomography (CT) scan, provides a measurement of attenuation specific for each individual head. Although not yet common practice in the clinical setting, this type of attenuation correction, called nonuniform attenuation correction, is commonly used for research studies.

Scatter

Scattered activity is when the high-energy gamma photons deviate from the straight path; scattered activity is measured in a location different from that of its site of origin. Compton scattering occurs when the interaction of gamma photons with tissue results in the transfer of electrons from the photon to the tissue. Because of the lower energy, scattered photons are typically measured at a photopeak value less than the primary photopeak (e.g., 159 keV for ^{123}I). Scatter alters the accuracy of quantification of SPECT images throughout the brain, although the effects will vary by tracer and by brain region. For cerebral blood flow studies, scatter alters quantitation in brain areas including the precentral, temporal, posterior hippocampus, and cerebellum, with minimal effects in the parietal and central areas. Importantly, the cerebellum, which is often used as a reference region for both cerebral blood flow and neuroreceptor imaging studies, appears to be the most vulnerable to scatter.

Several methods for scatter correction have been developed over the past few years, including (1) the triple energy window (TEW) method, which estimates scatter based on photons detected in adjacent energy windows, (2) image-based scatter correction (IBSC), in which scatter is estimated based on the reconstructed image and Chang's attenuation correction factor determined from the transmission scan, and (3) a Monte Carlo (MC)-based scatter correction, which applies a simulation code of all image-degrading effects, including attenuation, fanbeam collimator response, and scattering for nonhomogeneous voxelized objects, into the reconstruction algorithm. While not yet routinely used, incorporation of a scatter correction will increase the accuracy of quantitative measurements of SPECT images.

Spatial Resolution

Scattered activity causes the SPECT image to appear blurred. The blurred image limits the ability to distinguish two distinct points or the spatial resolution. The spatial resolution is commonly measured by the determination of the full width at half-maximum (FWHM) of the line-spread function. If the activity emitted from a point source is measured, the result is a Gaussian curve in which the amount of activity in the peak is significantly less than the actual activity in the point source, and the residual activity is measured in adjacent areas of no activity. The resolution of the image may be defined as the width of the Gaussian curve at 50% of the maximum value. To visualize two point sources of equal activity, they must be separated by a distance equal to the FWHM.

SPECT is often criticized for having poorer spatial resolution as compared to positron emission tomography (PET) imaging. However, while this is currently true, in that the spatial resolution of the most commonly used SPECT cameras is around 8–10 mm, compared to the most commonly used PET cameras, 3–6 mm. It is important to note that, theoretically, SPECT is capable of having greater spatial resolution than PET has because the gamma photons that are measured with the SPECT camera are back-projected to the site of original emission, in contrast to PET, wherein dual gamma photons are measured using coincidence detection, a site a short distance away from the site of original emission. As SPECT technology advances, spatial resolution will continue to improve and will soon be comparable to PET.

Partial Volume Effects

The limited spatial resolution of SPECT alters the accuracy of the quantitation of the regional activities. Activity from a point is spread out or blurred in three

dimensions. Thus, the measured activity in any one area reflects the actual activity and also scattered activity from the surrounding areas. This 'partial volume' effect causes errors in quantitative measurements that increase as resolution decreases. While currently a limitation , as SPECT technology advances and resolution improves, partial volume effects will be minimized.

Quality Control

Quality control is crucial to optimize image quality and quantification of neural receptor numbers using SPECT. Artifacts may be visualized in the image by errors in the ability of the detector to measure counts uniformly or by the center of rotation being off. Quality control procedures should be performed as recommended by the manufacturer and typically include (1) photomultiplier (PM) tube gain calibration, (2) linearity calibration, (3) sliding energy calibration, (4) center of rotation calibration, and (5) crystal uniformity calibration. PM gain is performed to measure the energy gain near the centers of the photomultiplier tubes; this calculates a correction factor to ensure that the photons emitted are within the correct energy window (159 keV for ^{123}I). The linearity calibration measures and corrects for the residual spatial distortion. The sliding energy calibration measures the response of the detector at each location on the crystal and calculates a correction factor for the peak energy response at each location. The center of rotation calibration measures the angle of the center of each collimator segment relative to a sensor on the camera, and the crystal uniformity calibration measures the inhomogeneity of detector response to a uniform distribution of radioactivity.

Perfusion Imaging

SPECT imaging has been used to measure blood flow patterns in brain, a procedure also known as regional cerebral blood flow (rCBF). Brain tissue regions with a lot of blood flow are imaged using brain perfusion agents that are lipophilic (or 'fat-liking'), so that they cross the blood–brain barrier in sufficient quantities to create a good visual image. These agents, which enter the brain quickly (90–120 s), are rapidly metabolized into a lipophobic (or 'fat-disliking') compound that is trapped in the brain neuron. The slow washout restricts the localization to brain areas that are active and have a lot of blood flow. Brain tissue known as gray matter, which contains neuron cell bodies and terminals, has higher rates of blood flow compared to white matter, which contains neuronal axons. Regional cerebral blood flow has been measured using ^{99m}Tc-labeled hexamethylpropyleneamine oxime (HMPAO), and bicisate ethyl cysteinate dimer (ECD). In a perfusion or blood flow SPECT images, the radiotracer is typically symmetrically distributed throughout the cerebral cortex, cerebellum, thalamus, and basal ganglia.

Brain perfusion SPECT is used clinically for the diagnosis and clinical maintenance of a variety of neurological conditions, including cerebrovascular disease, Huntington's disease, and dementia disorders such as Alzheimer's disease. Brain perfusion SPECT is superior to anatomic imaging (CT or magnetic resonance imaging (MRI)) for the detection of acute stroke in the first few hours following the event. The sensitivity and specificity of brain SPECT for stroke localization are 85.5% and 97.6%, respectively. The sensitivity of perfusion SPECT for detection of transient ischemic attacks is 60% in the first 24 h and 40% in the first week.

Brain perfusion SPECT can be used in identifying the location of brain tumors. Thalium-201 (^{201}Tl) and ^{99m}Tc are used to evaluate recurrences of brain tumors after therapy; the differential diagnosis between tumor and radiation changes is of importance for the consequent therapy. SPECT permits to distinguish metabolically active and nonactive regions – a process difficult to obtain with technologies that provide only images of structures in the body, such as MRI. [^{123}I]Iodomethyltyrosine (IMT) has been used to diagnose primary or recurrent cerebral gliomas and has been found to be superior to 2-[^{18}F]fluoro-2-deoxy-D-glucose (FDG) PET in delineating the extension of gliomas. In epilepsy, SPECT is used to localize the seizure focus for surgical therapy. Presently SPECT is the only imaging modality that can capture the rCBF changes that occur with seizures. Brain perfusion SPECT is also useful in the diagnosis of different dementia disorders. Each disorder has its own distinct rCBF pattern. Alzheimer's disease is diagnosed with SPECT with a diagnostic confidence of 98%.

SPECT Neuroreceptor Imaging

SPECT neuroreceptor imaging studies have provided insight into the neurochemical status of numerous neural receptors and transporters throughout the spectrum of neuropsychiatric brain disorders (Table 1; Figure 3). While there are millions of chemical sites in brain tissue, only a handful of neural receptors and transporters have been imaged using SPECT to date; these include the dopamine transporter, dopamine D2 receptors, serotonin transporters, serotonin (5-hydroxytryptamine 2A, or 5-HT_{2A}) receptors, cannabinoid (CB_1) receptors, nicotinic acetylcholine receptors (containing the β_2 subunit), γ-aminobutyric

Table 1 SPECT neuroreceptor radiopharmaceuticals

Receptor site	*Radiopharmaceutical*
Cannabinoid	
CB1 receptor	[^{123}I]AM281
Cholinergic	
Nicotinic acetylcholine eceptor	[^{123}I]Iodonicotine
β^2-subunit, nAChR	[^{123}I]5-IA-85380
α7-nAChR muscarinic receptors	[^{123}I]QNB, [^{123}I]iododexetimide
Dopaminergic	
Dopamine transporters	[^{123}I]β-CIT, [^{123}I]FP-CIT, [^{123}I]altropane, [^{123}I]PE21, [^{99m}Tc]TRODAT
D2 receptors	[^{123}I]3-Iodo-6-methoxybenzamide, [^{123}I]epidepride
GABAergic	
Benzodiazepine–GABA$_A$ receptor	[^{123}I]Iomazenil
Glutamatergic	
NMDA receptor (PCP site)	[^{123}I]CNS-1261
Serotonergic	
5-HT$_2$ receptors	[^{123}I]R93274 or R91150, [^{123}I]IBSP, [^{123}I]iodoketanserin
5-HT transporters	[^{123}I]INQUIP, [^{123}I]β-CIT, [^{123}I]nor-β-CIT, [^{123}I]ADAM

[^{123}I]AM281, *N*-(morpholin-4-yl)-1-(2,4-dichlorophenyl)-5-(4-[^{123}I]iodophenyl)-4-methyl-1*H*-pyrazole-3-carboxamide; nAChR, nicotinic acetylcholine receptor; [^{123}I]5-IA-85380, 5-[^{123}I]iodo-3-(2(*S*)-azetidinylmethoxy)pyridine; [^{123}I]QNB, [^{123}I]iodoquinuclidinyl benzilate; [^{123}I]β-CIT, 2β-carbomethoxy-3β(4-[^{123}I]iodophenyl)tropane; [^{123}I]FP-CIT, 2β-carbometoxy-3β-(4-[^{123}I]iodophenyl)-*N*-(3-fluoropropyl)nortropane; [^{123}I]PE21, *N*-(3-[^{123}I]iodoprop-2-enyl)-2β-carbomethoxy-3β-(4'-methylphenyl)nortropane; [^{99m}Tc]TRODAT, [^{99m}Tc][2[[2-[[[3-(4-chlorophenyl)-8-methyl-8-azabicyclo[3.2.1]oct-2-yl]-methyl](2-mercaptoethyl)amino]ethyl] amino]ethane-thiolato(3-)-*N*2,*N*2',*S*2,*S*2']oxo-[1*R*-(exo-exo)]; GABA, γ-aminobutyric acid; NMDA, *N*-methyl-D-aspartate; [^{123}I]CNS-1261, *N*-(1-naphthyl)-*N'*-(3-[^{125}I]iodophenyl)-*N'*-methylguanidine; 5-HT, 5-hydroxytryptamine (serotonin); [^{123}I]R93274 or R91150, 4-amino-*N*-[1-[3-(4-fluorophenoxy)propyl]-4-methyl-4-piperidinyl]-5-[^{123}I]iodo-2-methoxybenzamide; [^{123}I]IBSP, 8-[3-(4-fluorobenzoyl)propyl]1-(4-[^{123}I]iodobenzoyl)-1,3,8-triazaspiro[4,5]decan-4-one; [^{123}I]INQUIP, 5-[^{123}I]iodo-6-nitroquipazine; [^{123}I]ADAM, 2-([2-([dimethylamino]methyl)phenyl]thio)-5-[^{123}I]iodophenylamine.

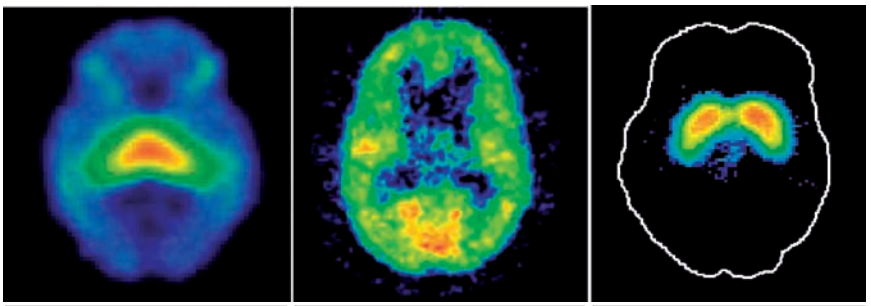

[^{123}I]5-IA-85380 Nicotinic acetylcholine receptor	[^{123}I]Iomazenil Benzodiazepine-GABA-A receptor	[^{123}I]β-CIT Dopamine and serotonin transporters

Figure 3 SPECT neuroreceptor images.

acid (GABA)$_A$–benzodiazepine receptors, *N*-methyl-D-aspartate (NMDA receptors), and the norepinephrine transporter.

[^{123}I]2β-Carbomethoxy-3β(4-iodophenyl)tropane ([^{123}I]β-CIT) SPECT imaging has demonstrated decreased expression of the serotonin (5-HT) transporter in the diencephalon in major depressive disorder, which appears to be limited to women under the age of 50 years, suggesting a sex-specific decrease in 5-HT transporter availability. Several studies have suggested that 5-HT transporter availability does not predict or differentiate responder and nonresponders to antidepressant treatment; however, since these studies failed to evaluate sex-specific differences, it remains to be determined if this would be useful for young women suffering from major depressive disorder. In contrast, occupancy of 5-HT transporter by citalopram is correlated with a better clinical response in patients suffering from obsessive–compulsive disorder.

[^{123}I]Iomazenil SPECT imaging of the benzodiazepine site on the GABA$_A$ receptor is decreased in patients with probable Alzheimer's disease. The

same site has been shown to upregulate over the first week of abstinence in alcohol-dependent nonsmokers but not in alcohol-dependent smokers and to correlate with the severity of alcohol withdrawal, suggesting that smoking suppresses the upregulation of the receptor and also decreases the severity of alcohol withdrawal symptoms. [^{123}I]Iomazenil SPECT imaging of the benzodiazepine site on the $GABA_A$ receptor in alcohol-dependent patients abstinent for 1–6 months suggests that the receptor normalizes and/or decreases to below control levels.

[^{123}I]β-CIT SPECT imaging in cocaine-dependent patients has demonstrated higher striatal dopamine transporters that normalize after 6 months of abstinence. SPECT [^{123}I]β-CIT imaging of the loss of striatal dopamine transporter expression is a highly sensitive marker for the presence and severity of Parkinson's disease. SPECT imaging of the dopamine D2 receptor before and after amphetamine administration has demonstrated greater amphetamine-induced dopamine release in schizophrenic patients compared to healthy individuals. SPECT *N*-(1-naphthyl)-*N*′-(3-[^{125}I]iodophenyl)-*N*′-methylguanidine ([^{123}I]CNS-1261) imaging of the NMDA receptor demonstrated a significant decline in receptors throughout the brain of clozapine-treated schizophrenic patients but did not demonstrate this decline in the brain of drug-free schizophrenic patients.

While these are just a few illustrations of the knowledge that neuroreceptor imaging has provided on the chemical abnormalities in neuropsychiatric disorders, the utility of SPECT for understanding the neuropathological bases of these highly complex brain disorders is clear. Additional information about SPECT imaging research studies may be sought in the literature listed in the section 'Further reading.'

Clinical Utility of and Future of SPECT

Over the past two decades, SPECT has gained merit as a tool to image brain function and neurochemistry in living humans. Perfusion SPECT is routinely used for the clinical diagnosis and assessment of several neurological disorders. Neuroreceptor SPECT imaging has been useful in research to begin to identify the chemical state of different neuropsychiatric disorders that result from abnormal expression of neuroreceptors and transporters. While in its youth, neuroreceptor SPECT imaging holds tremendous potential in the clinical setting for the diagnosis of a myriad of psychiatric disorders for which currently there are no biological or chemical diagnostic tools. SPECT is highly amenable to being used as a tool in a majority of clinical settings. Compared to positron emission tomography, SPECT is less costly, does not require an on-site cyclotron, and, because of major technological advances, is rapidly improving in sensitivity and resolution. With continued progress in the development of radiopharmaceuticals and in the technology for the acquisition and image processing of SPECT data, SPECT imaging has the potential to radically change the way that neuropsychiatric disorders are diagnosed and treated.

See also: Cognition: An Overview of Neuroimaging Techniques; Executive Function and Higher-Order Cognition: Neuroimaging; Neuroimaging; Single Photon Emission Computed Tomography (SPECT): Technique.

Further Reading

Camargo E (2001) Brain SPECT in neurology and psychiatry. *Journal of Nuclear Medicine* 42: 611–623.

Devous MD (2002) Functional brain imaging in the dementias: Role in early detection, differential diagnosis, and longitudinal studies. *European Journal of Nuclear Medicine* 29: 1685–1687.

Erritzoe D, Talbot D, Frankle WG, et al. (2003) Positron emission tomography and single photon emission CT molecular imaging in schizophrenia. *Neuroimaging Clinics of North America* 13: 817–832.

Malison RT, Laruelle M, and Innis RB (1995) Positron and single photon emission tomography. In: Bloom FE and Kupfer DJ (eds.) *Psychopharmacology, the Fourth Generation of Progress*, pp. 865–879. New York: Raven Press.

Morano GN and Seibyl JP (2003) Technical overview of brain SPECT imaging: Improving acquisition and processing of data. *Journal of Nuclear Medicine Technology* 31: 191–195.

Pimlott SL (2005) Radiotracer development in psychiatry. *Nuclear Medicine Communications* 26: 183–188.

Ravina B, Eidelberg D, Ahlskog JE, et al. (2005) The role of radiotracer imaging in Parkinson disease. *Neurology* 64: 208–215.

Staley JK, Malison RT, and Innis RB (1998) Imaging of the serotonergic system: Interactions of neuroanatomical and functional abnormalities of depression. *Biological Psychiatry* 44: 534–549.

Volkow ND, Fowler JS, and Wang G-J (2003) Positron emission tomography and single photon emission computed tomography in substance abuse research. *Seminars in Nuclear Medicine* XXXIII: 114–128.

Warwick JM (2004) Imaging of brain function using SPECT. *Metabolic Brain Disease* 19: 113–123.

Single Photon Emission Computed Tomography (SPECT): Technique

A B Newberg and A Alavi, University of Pennsylvania, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Single photon emission computed tomography (SPECT) was developed in the 1960s and was used to obtain the first transaxial reconstructions of radionuclide distribution in the brain. This technique was used to study a number of neurological and psychiatric disorders and the impact of these disorders on radionuclide distribution in the central nervous system (CNS). SPECT was the first functional imaging modality that was used in the study of the CNS, andthese early studies met with much success. The SPECT studies employed single-photon emitters such as iodine-123 (^{123}I) or technetium-99m (^{99m}Tc) as the radionuclide that was attached to the substrate. The most commonly used clinical radioligands for SPECT are ^{99m}Tc hexamethyl propylene amine oxime (HMPAO) and technetium-99m bicisate to measure blood flow. However, a number of radiopharmaceuticals (see Table 1) for SPECT imaging have been developed to study various aspects of cerebral function, including a wide array of neurotransmitters. The radiopharmaceuticals used with SPECT imaging has been widely used in the evaluation of patients with various neurological and psychiatric disorders as well as various activation states of the brain.

SPECT imaging is highly advantageous for a number of reasons, comparing favorably with other imaging modalities such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET). Several noteworthy points about SPECT imaging include:

- *In vivo* imaging techniques allow the study of humans over extended periods of time. These longitudinal studies provide valuable information and may be particularly important in studies of the progression of disease and the response to treatment.
- The SPECT system, using the fan beam collimator, has good sensitivity and excellent spatial resolution (although not typically as good as PET or fMRI).
- The single-photon tracers, especially ^{99m}Tc- and ^{123}I-labeled agents, are readily available at a relatively low cost.
- The tracers can be readily prepared from premade kits without the need for an on-site cyclotron.
- There is little or no possibility of saturating cerebral binding sites with the SPECT tracers because the specific activity of these tracers is very high, $>100\,000$ Ci mmol^{-1}. The chemical amount to be injected in humans is too small to affect the number of available binding sites in the brain. Such tracers do not result in any measurable pharmacological effects.
- The localization and density of specific binding sites or biological processes in their native states can be investigated. The kinetic information determined by SPECT imaging can only be obtained by such techniques.

Normal Variants in SPECT Imaging

The ability to detect abnormalities on clinical brain SPECT scans or in research studies initially requires a determination of the normal variations that might be observed on such scans. One of the primary issues related to the definition of normal that arises with brain imaging is the determination of the baseline brain state (Figure 1). There has been considerable attention in the literature regarding the best conditions for obtaining a baseline functional imaging study. What the patient is doing and experiencing may have a profound influence on brain function. This has provided a foundation for activation studies that are typically performed on healthy individuals involved in various sensorimotor, cognitive, or affective tasks. For example, the brain function with eyes open is markedly different than the brain with the eyes closed. Once the eyes are open, there is a dramatic increase in the primary visual areas. This activity increases with the complexity of the scene being presented to the brain and may also include other association areas. Movement in the external environment also alters function in the visual centers of the brain as well as in the medial prefrontal cortex, temporoparietal junction, basal temporal regions, and extrastriate cortex. Sounds also elicit changes in cerebral function and, therefore, whether the ears are occluded may influence brain function.

Technical factors must also be considered when evaluating brain function because they may introduce artifacts and findings that are, in reality, variants of normal brain perfusion. These include the time between injection and scanning, the dose administered, filtering and processing steps, test–retest variability, and the type of scanner. As the duration of time from the injection increases, it is possible that

Table 1 Partial list of existing SPECT tracers[a]

Compound	*Application*
HMPAO, bicisate	Cerebral blood flow
IQNB	Muscarinic cholinergic receptor
Iodopride, IBZM, iodospiperone	Dopamine receptor
AMIK, DOI	Serotonin receptor
TRODAT	Dopamine transporter
ADAM	Serotonin transporter
β-CIT	Dopamine or serotonin transporter
Iomazenil	Benzodiazepine
2-Iodomorphine	Opioid receptor
^{123}I-d(CH_2)5[Tyr(Me)$_2$, Tyr(NH_2)$_9$]AV	Vasopressin receptor

[a]β-CIT, [^{123}I]-2β-carbomethoxy-3β-(4-idophenyl)-tropane; ADAM, 2-((2-((dimethylamino)methy) phenyl)thio)-5-iodophenylamine; AMIK, 7-amino-8-iodo-ketanserin; DOI, 1-(2,5-dimethoxy-4-iodophenyl)-2-aminopropane; HMPAO, technetium-99m hexamethyl propylene amine oxime; IBZM, 3-iodo-N-[(1-ethyl-2-pyrrolidinyl)] methyl-2-hydroxy-6-methoxybenzamide; IMP, iodine-123-N-N′, N,-trimethyl-N′-[2-hydroxyl-3-methyl-5-iodo-benzyl]-1, 3 propane diamine; IQNB, 3-quinuclidinyl benzilate; TRODAT, Tc-99m [2-[[2-[[[3-(4-chlorophenyl)-8-methyl-8-azabicyclo[3.2.1]oct-2-yl]-methyl](2-mercaptoethyl) amino]ethyl]amino]ethane-thiolato(3-)-N2, N2′, S2, S2′]oxo-[1R-(exo-exo)].

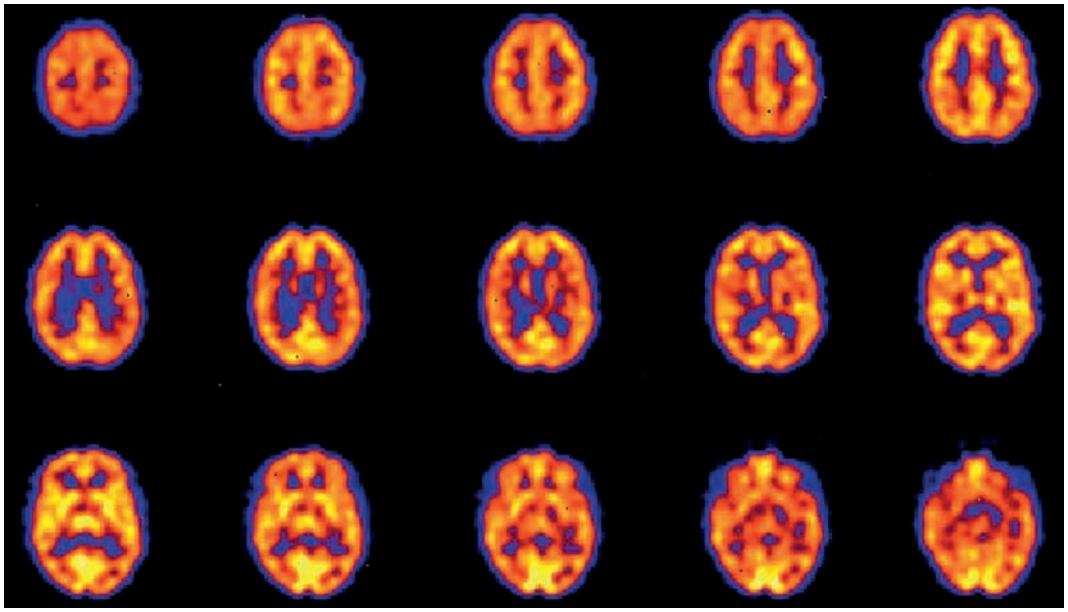

Figure 1 SPECT scan using ^{99m}Tc bicisate showing uniform distribution of blood flow throughout the cortical and subcortical structures in a healthy subject. The images are axial slices from the top of the head to the cerebellum.

certain tracers may have an altered cerebral distribution. Obviously, the ideal tracer would remain completely fixed within the brain at the time of extraction from the blood. However, most tracers do experience some degree of washout and some even undergo redistribution. ^{99m}Tc bicisate, for example, begins to wash out of the cortical regions faster than the subcortical areas. The result is that a scan obtained several hours after injection might appear to have pathologically low activity in the cortex compared to a scan of the same person obtained within the first hour of the injection. Other tracers, especially neurotransmitters, may have altered biodistribution over time, and this is a factor that must be considered for any clinical or research study.

Filtering and processing of the images is also crucial to determining cerebral activity. SPECT imaging with fan beam collimators and the appropriate application of filters is necessary to adequately visualize subcortical structures and central cortical areas (Figure 2). Attenuation correction is also necessary for adequately visualizing and comparing activity in various structures (Figure 3). Patient motion can always complicate interpretation of SPECT scans. Methods for

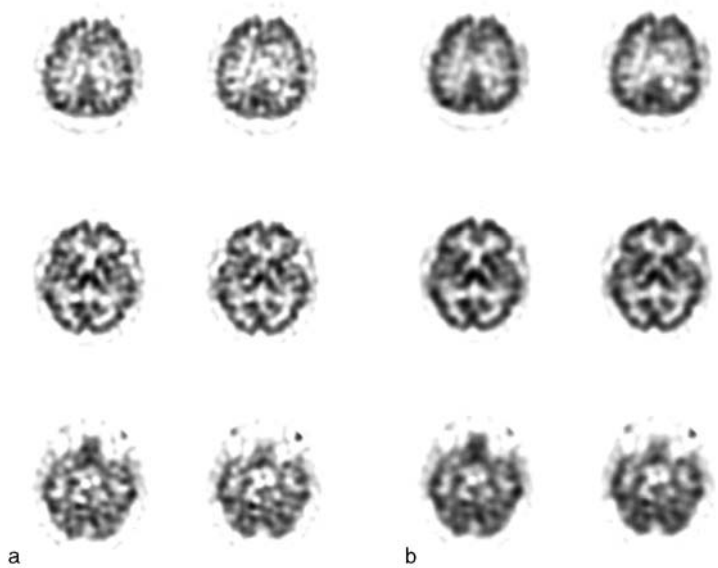

Figure 2 Transaxial SPECT images of ^{99m}Tc bicisate showing the differences in scans when processed via different Weiner filter multipliers of: (a) 0.3; (b) 0.6. In (a), the scan appears very grainy, making some structures difficult to observe. In (b), the scan appears to be much smoother, but there is less delineation between structures.

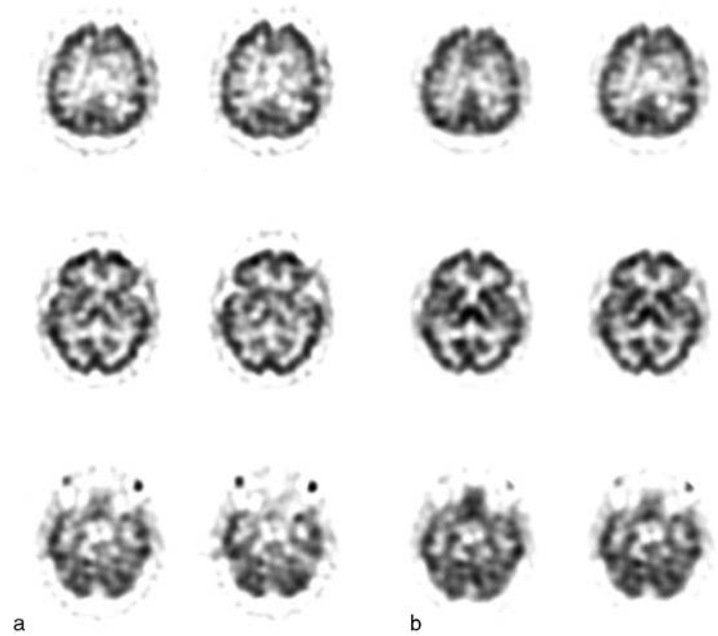

Figure 3 Transaxial SPECT images of ^{99m}Tc bicisate: (a) before correction for photon attenuation; (b) after correction for photon attenuation. The ability to restore counts to central structures results in much better observation of the subcortical structures, such as the thalamus, after attenuation correction is applied.

preventing patient motion from interfering with scanning include firm head holders and the use of tape or some other method to hold the head in place. Also, shortening image acquisition makes the session easier on the patient. Alternatively, breaking up the acquisition time into multiple shorter scans can be useful because scans with motion can be excluded without losing the entire study. Another processing factor that can affect the visual inspection of brain scans is head tilt. Because a patient's head is rarely in complete alignment with the camera, slight and especially significant head tilt might result in certain areas appearing asymmetric in comparison to the contralateral structure. For example, the thalami can frequently

appear to have asymmetric activity if there is head tilt. However, this can usually be identified by examining all the slices containing the structure. As the slices proceed through the structure, asymmetries associated with head tilt should flip from one side to the other. This can usually be corrected by computer programs that realign brain images.

In spite of many of these issues, most studies of cerebral blood flow, metabolism, or even neurotransmitter systems have demonstrated good test–retest results with small variability within structures. Most studies have shown that in healthy controls repeat scans typically demonstrate regional activity and absolute activity to be within 5–10%.

SPECT Imaging from Infancy to Adulthood

The normal aging process is associated with a number of biochemical changes in the brain. Although postmortem studies may help to elucidate the nature of these changes, the advent of functional neuroimaging techniques allowed these biochemical changes in the brain to be measured *in vivo* across the life span of the person.

During the first year of normal development, studies have shown that the pattern of activity observed on functional brain scans generally corresponds to the phylogenetic order of development. Thus, the functional maturation (as measured by increased metabolism) of developmentally older structures precedes that of structures that develop later. For example, as the visuospatial and visuosensorimotor function and primitive reflexes develop, increased cerebral function is observed in the parietal, temporal, cerebellar, and primary visual cortices. Increases have also been observed in the basal ganglia as movement and sensorimotor function becomes more integrated. Function in the frontal lobe remains low during the first 4 months of life and increases as the infant begins to develop higher cortical and cognitive capabilities. Thus, as the infant develops more complex social interactions as well as improves his or her abilities in various neuropsychological tests that specifically involve frontal lobe function, these changes are reflected by increases in frontal lobe activity. By 1 year of age, the overall pattern of activity is qualitatively similar to that of a young adult.

The results of studies in adults that have measured cerebral blood flow in normal aging have generally noted a global decrease in cerebral blood flow with age. This typically affects the gray matter more than the white matter. Furthermore, the decrease appears to be nonlinear, with a sharper decrease prior to the age of approximately 36 years followed by a slower decline. Several areas appear to be affected, but the most commonly observed decrease is in the frontal lobes bilaterally (Figure 4). However, age-related decline in tracer uptake also occurs in the anterior cingulate gyrus, bilateral basal ganglia, left prefrontal, left lateral frontal, and left superior temporal and insular cortex.

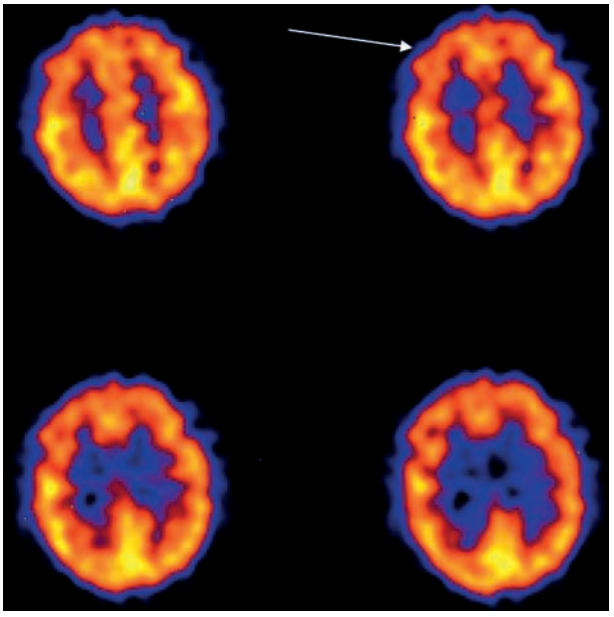

Figure 4 Transaxial SPECT images of ^{99m}Tc bicisate in normal aging with decreased bilateral frontal lobe perfusion (arrow).

Neurotransmitter Function in Aging

In addition to measuring changes in cerebral blood flow with age, SPECT can measure neurotransmitter activity, which may also change with age. A large number of neurotransmitter systems can be studied using SPECT (see Table 1). Many of these have been used to study the effects of aging.

The nigrostriatal dopaminergic pathways have been studied with SPECT imaging. This pathway is important with regard to extrapyramidal symptoms that occur in the elderly. This system also has particular significance in the study of Parkinson's disease (PD). *In vitro* studies have generally shown decreases in both pre- and postsynaptic receptor levels with age. Several SPECT studies have been performed with ligands that bind to postsynaptic receptors such as D1 and D2. Radioligands that can measure the dopamine transporter system, which removes dopamine from the neuronal synapse into the terminal for storage, have also been developed. One study, using the cocaine analogs such as ^{99m}Tc TRODAT-1 and 2β-carbomethoxy-3β-(4-iodophenyl)tropane (β-CIT), indicated an age-related decrease in dopamine transporters. ^{99m}Tc TRODAT-1, has demonstrated a nonlinear decrease

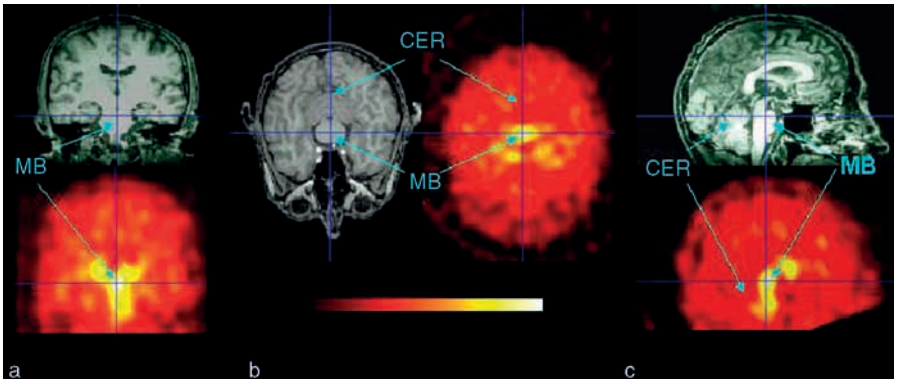

Figure 5 Multiple projections of ^{123}I ADAM SPECT and corresponding MRIs showing areas of substantial serotonin transporter binding in the midbrain with relatively little uptake in the cortex and cerebellum: (a) coronal; (b) transaxial; and (c) sagittal. CER, cerebellum; MB, midbrain.

in uptake with age. With this tracer, most of the decline appears to occur before the age of 40 years, with a slower decline from 40 to 75 years of age.

The serotonergic system is critical for a number of functions in the brain and is most notably involved in the regulation of mood. Several studies have examined the changes with age of the serotonin (5-HT) system using ^{123}I β-CIT or ^{123}I ADAM (Figure 5). Generally, there is a decline in binding to the 5-HT transporter of approximately 10% per decade in the thalamus and midbrain. An age-related decreased activity for the 5-HT1A receptor has also been reported, with a decline of approximately 10% per decade in a number of cortical areas except for the medial temporal cortex. Similar decreases have been reported for the 5-HT2A receptor.

Other neurotransmitter systems have not been widely studied with regard to the effects of aging. However, given the list of radioligands available for SPECT imaging, it seems that these techniques may have vast applications in the study of the neurotransmitter effects that result from normal aging as well as neurological and psychiatric disorders. Several studies have been mentioned, but there remains a significant amount of neurotransmitter systems which have yet to be thoroughly explored with regard to the effects of aging.

SPECT Imaging in Neuropsychiatric Disorders

Alzheimer's Disease

Since 1980, a large number of studies have used SPECT in the assessment of patients with Alzheimer's disease (AD). Most studies show that the bilateral parietal and temporal lobes are particularly affected in AD. This parietal hypoperfusion (Figure 6) is often referred to as representing the typical pattern of AD

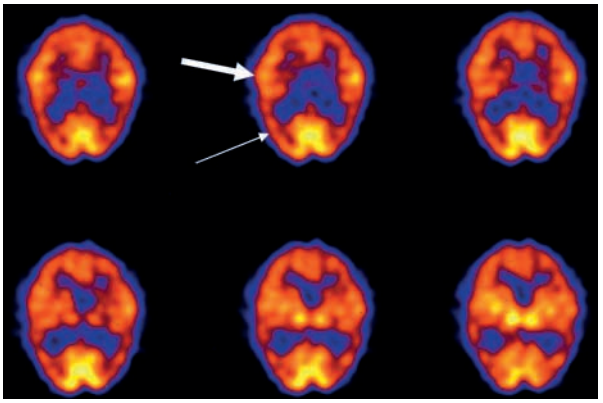

Figure 6 Transaxial SPECT images of ^{99m}Tc bicisate in a patient with late Alzheimer's disease (AD). There is moderately decreased perfusion in both temporoparietal regions (thin arrow) in comparison to the control studies shown in previous figures. In this later stage of AD, there is also decreased perfusion in the frontal lobes with preservation of perfusion only in the sensorimotor areas (thick arrow), visual cortex, and subcortical structures.

and may be particularly pronounced in patients younger than 65 years. However, it should be noted that, although the bilateral parietal pattern is highly predictive of AD, the pattern is not pathonomic for AD and may be seen in patients with PD, bilateral parietal subdural hematomas, bilateral parietal stroke, and bilateral parietal radiation therapy ports. In spite of this limitation, until recently, SPECT has been widely used for assisting in the diagnosis of AD. More recently, fluoro-2-deoxyglucose (FDG) PET has been used for the diagnosis of AD and frontal lobe dementia. Given the superior image quality of FDG PET, this may ultimately replace the use of SPECT in the workup of patients with memory problems.

In patients with AD of varying severity, the magnitude and extent of hypoperfusion correlate with the severity of the dementia symptoms. Other areas (the sensorimotor and visual cortices, subcortical nuclei,

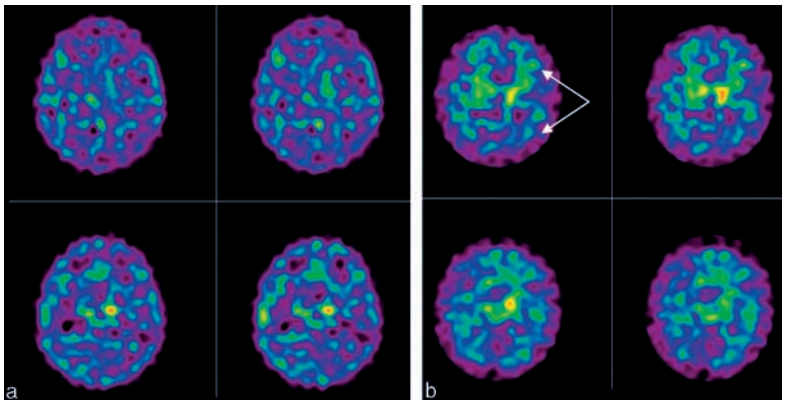

Figure 7 Transaxial SPECT images of ^{123}I IMPY: (a) in a elderly control subject; (b) in a patient with Alzheimer's disease (AD). There is substantially greater binding in the cortex in the AD patient than in the control (arrows).

brain stem, and cerebellum) have relatively preserved perfusion except in patients with specific neuropsychological deficits.

Newer SPECT radiopharmaceuticals are being developed to specifically target the pathophysiological process of AD. For example, ^{123}I IMPY, which binds to the amyloid protein, has recently been developed and tested in human subjects (Figure 7). It is hoped that such compounds will be useful for early diagnosis as well as monitoring responses to interventions designed to attenuate the pathophysiological process.

Cerebrovascular Disease

Cerebrovascular disease is the third leading cause of death in the United States and affects approximately half a million people. However, stroke is often associated with a poor outcome, in part due to the lack of understanding of the mechanisms that underlie stroke and the process by which recovery may take place. SPECT, in addition to PET, imaging has been of great benefit in advancing the understanding of the pathophysiology of cerebrovascular disorders. SPECT imaging allows for the detection of stroke earlier and with higher sensitivity than anatomic imaging with either MRI or computed tomography (CT). Further, SPECT imaging has been useful in evaluating the extent of the functional damage because areas not immediately affected by the infarct may show hypoperfusion (Figure 8). However, SPECT studies have not become prominent in the management of stroke patients for several reasons. One of the most important initial issues is whether a stroke is hemorrhagic, and SPECT cannot make this determination. Further, SPECT has not been shown to substantially alter management in the acute setting. In the chronic setting, SPECT may help with prognosis and to predict functional recovery, but this is not often useful in the clinical setting.

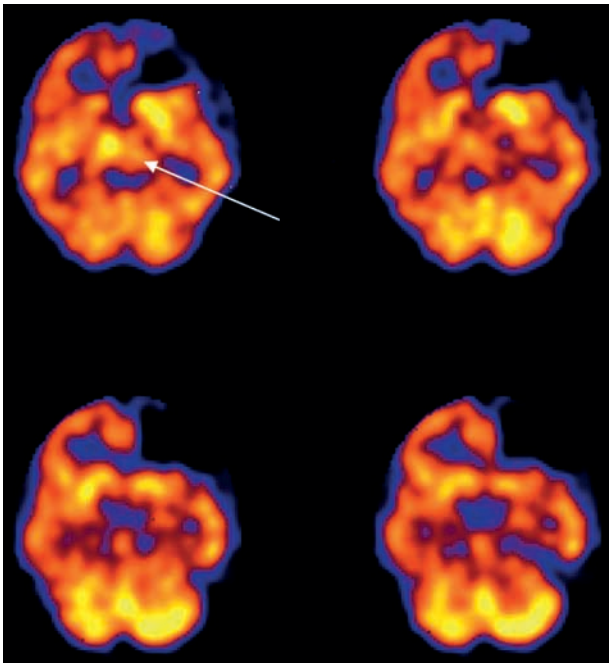

Figure 8 Transaxial SPECT images of ^{99m}Tc bicisate in a patient with recent left frontal lobe stroke. The scan shows near-absent activity in the left frontal region. There is also decreased perfusion in the ipsilateral thalamus (arrow), which may be associated with deafferentation from the affected regions in the left frontal lobe.

Where SPECT has been helpful is in the evaluation of cerebrovascular reserve, patients are given an initial SPECT scan followed by a second scan in which they have been pretreated with acetazolamide. Acetazolamide causes cerebral vasodilation, which unmasks areas that are at risk for future ischemic events. This helps to guide possible surgical interventions such as carotid endarterectomy.

Head Trauma

There have been a limited number of studies using SPECT in the evaluation of patients with head trauma. One of the problems with the use of SPECT in these cases is that SPECT cannot distinguish between structural damage and cerebral dysfunction because these may all result in areas of decreased perfusion. Thus, it is helpful to compare SPECT to anatomic images such as those obtained by MRI or CT, especially because cerebral dysfunction can extend beyond the boundary of anatomic lesions and may appear even in locations remote from the trauma.

Lesions such as cortical contusions, intracranial hematoma, and resultant encephalomalacia have perfusion defects that are confined primarily to the site of injury . However, subdural and epidural hematomas often cause widespread hypoperfusion and may affect even the contralateral hemisphere. Another entity, diffuse axonal injury, has been found to cause diffuse cortical hypoperfusion, but there is a particularly marked decrease in metabolism in the thalamus which appears to correlate with the severity of cognitive dysfunction. Further, crossed cerebellar diaschisis, as well as ipsilateral cerebellar hypoperfusion, has been found in head-injury patients with supratentorial lesions.

Seizures

Epilepsy affects 0.5–1.0% of the population, can cause focal or generalized seizures, and usually begins in childhood. In general, during an epileptic seizure, cerebral metabolism and blood flow are markedly increased. The focus of partial seizures can be identified using SPECT imaging because these areas have increased blood flow during the seizure and decreased blood flow in the interictal period (Figure 9). It has been shown that single hypoperfused regions can be identified in up to 80% of patients with focal electroencephalogram (EEG) abnormalities. The degree of asymmetry in the region of the seizure focus appears greater with increasing duration of the seizure disorder. The type of seizure preceding the SPECT study may also affect the cerebral blood flow landscape such that hypoperfusion is limited to the epileptogenic zone if the preceding seizure is focal limbic,

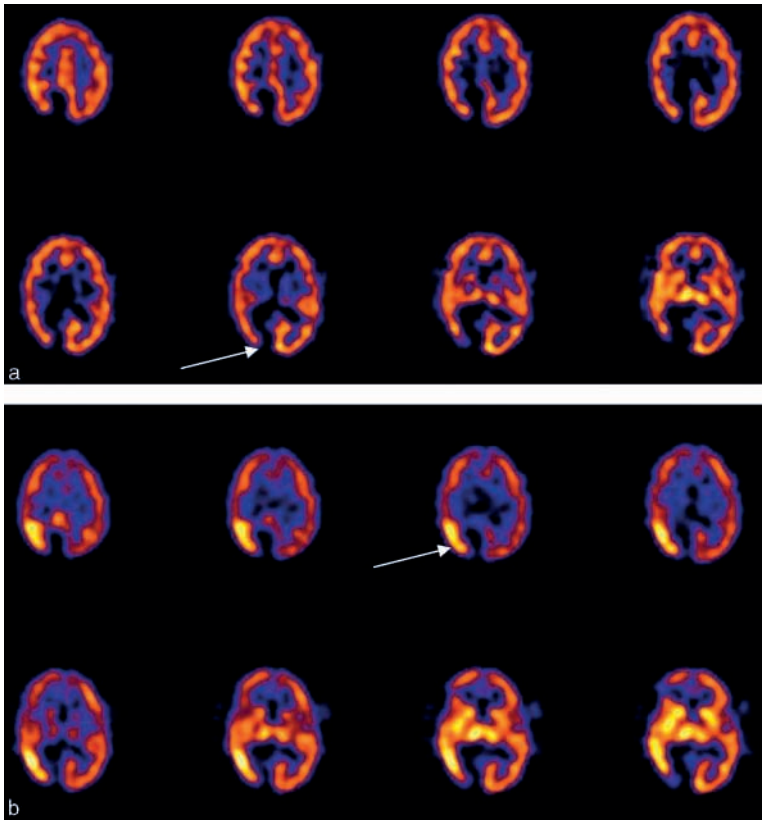

Figure 9 Transaxial SPECT images of ^{99m}Tc bicisate in a patient with seizures: (a) interictal scan; (b) ictal scan. The interictal scan in (a) shows decreased activity in the region of prior surgery with mildly decreased activity in the adjacent cortex (thin arrow). The ictal scan in (b) shows the surgical area as decreased perfusion, but there is an intense focus of activity in the adjacent cortex (thin arrow), clearly delineating the seizure focus.

whereas patients with widespread limbic seizures have hypoperfusion that includes one or several additional areas of the limbic cortex.

The temporal lobe is the most common focus of partial epilepsy (Figure 9). Studies show that the sensitivity of interictal SPECT in detecting temporal lobe epilepsy (TLE) foci is over 70%. Ictal SPECT studies provide a more sensitive evaluation of seizure foci and can detect areas of hyperperfusion in up to 80% of patients. When statistical parametric mapping or coregistration with MRI is used to enhance detection, the sensitivity can be improved up to 90%. When additional cerebral areas of abnormality are detected, the implication is that the seizures have a more complex origin. This also makes surgical excision of the seizure focus, an important intervention in patients refractory to medications, less likely to be successful.

Depression

Depression is generally characterized by decreased cortical perfusion, particularly affecting the frontal lobes (Figure 10). However, the exact pattern has been somewhat controversial, with different studies reporting slightly discrepant findings. Furthermore, generalized decreases in cortical perfusion have also been reported with a number of psychotropic medications such as benzodiazepines, sedatives, antipsychotic medications, and anticonvulsant medications. Thus, SPECT cerebral blood flow imaging of depression has generally not been very useful in a clinical or research setting.

A number of functional neuroimaging studies have focused on the role of the 5-HT system in depression, including both pre- and postsynaptic receptors. 5-HT2A postsynaptic receptors are generally reduced in the frontal cortex. SPECT studies of the 5-HT transporter using ^{123}I β-CIT and ^{123}I 2-((2-((dimethylamino) methy) phenyl)thio)-5-iodophenylamine (ADAM) revealed reductions in binding in the brain stem in drug-free depressed patients compared to controls. Several initial SPECT studies of treatment effects suggested that antidepressant medication and also psychotherapy improve 5-HT function in patients who demonstrate a good response.

Movement Disorders

Parkinsonian symptoms are associated with a number of neurodegenerative disorders, such as PD, multiple system atrophy (MSA), and progressive supranuclear palsy (PSP). These disorders are frequently associated with abnormalities in a number neurophysiological processes. In particular, pathological evidence has shown that these disorders are associated with a loss of neurons, particularly in the nigrostriatal dopaminergic pathway.

SPECT and PET have become important tools in the differential diagnosis of these diseases and may have sufficient sensitivity to detect neuronal changes before the onset of clinical symptoms. Imaging is now being used to elucidate the genetic contribution to PD and, in longitudinal studies, to assess the efficacy and mode of action of therapeutic interventions such as neuroprotective drugs and surgery.

SPECT studies of radiotracer binding to postsynaptic dopamine receptors and presynaptic dopamine

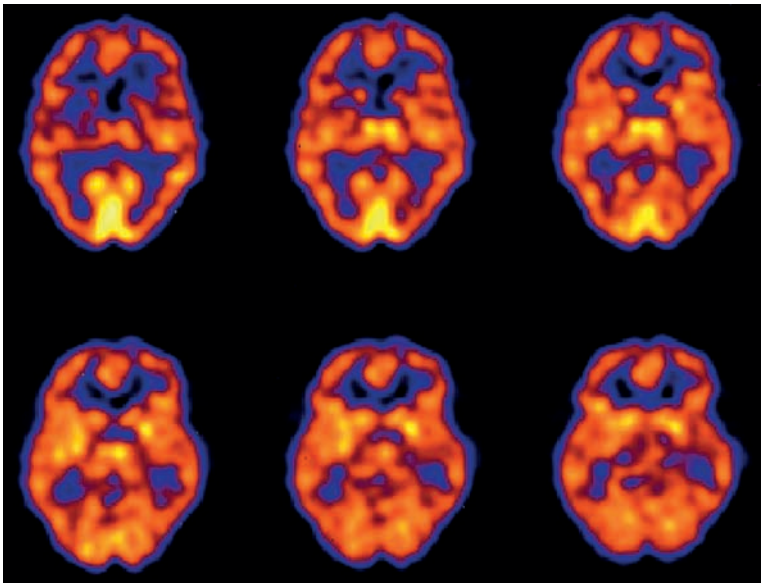

Figure 10 Transaxial SPECT images of ^{99m}Tc bicisate in a patient with depression show overall decreased perfusion in the cortex, with the frontal lobes most prominently affected. There is preserved perfusion in the visual cortex and thalamus.

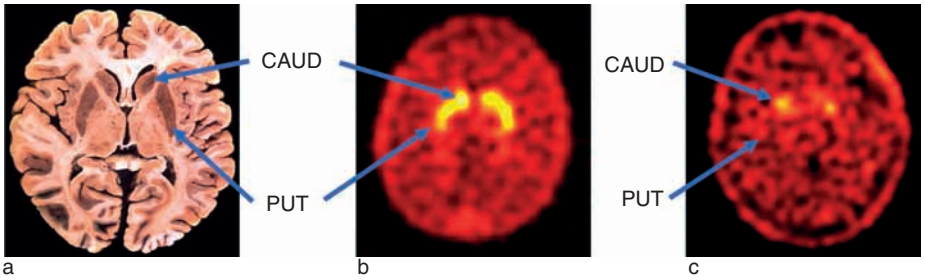

Figure 11 Transverse SPECT images of [^{99m}Tc]TRODAT-1 binding to dopamine transporters in human subjects: (a) histology slice through the brain at the level of the striatum showing the head of the caudate nucleus (CAUD) and the putamen (PUT); (b) image at the same level from a healthy subject showing high concentrations of [^{99m}Tc]TRODAT-1 binding to dopamine transporters in the caudate and putamen; (c) image at the same level in a patient with bilateral Parkinson's disease showing significant reductions of [^{99m}Tc]TRODAT-1 binding in the putamen but with the caudate relatively spared.

transporters and neurons have proved to be powerful techniques for quantifying the loss of dopaminergic neurons in normal aging and in PD and other neurodegenerative disorders. SPECT studies have indicated a consistent pattern of dopaminergic neuronal loss in PD, usually with a more pronounced decrease in the putamen rather than in the caudate (Figure 11). In addition, there is often a marked asymmetry of uptake in the striatum, particularly in the early stages of the disease. This asymmetry usually has a good correlation with symptom severity and illness duration. Most important, SPECT imaging studies may be sensitive enough to detect very early PD, perhaps even before clinical symptoms become apparent.

Although most of the SPECT studies have shown highly significant differences between groups of PD patients and age-matched normal controls, the statistically significant differential diagnosis of an individual subject is more problematic. Patients with severe PD are easily separated from healthy controls just from a simple visual inspection of striatal images, which can be quantified using some form of discriminant analysis which has a sensitivity and specificity from 90% to 100% in the proper clinical setting. However, patients presenting much earlier in the course of the disease are more difficult to detect because there can be a substantial overlap with an age-matched control group and, consequently, a loss of diagnostic accuracy. The situation may be further complicated if the early differential diagnosis among several different neurodegenerative disorders is required. Many of the symptoms associated with parkinsonian disorders are nonspecific, which is why the accurate clinical diagnosis of these diseases is difficult. Indeed, some histopathalogical studies have shown that as many as 25% of all patients who were diagnosed with PD before death had been misdiagnosed. Studies measuring dopamine transporter activity have had more difficulty in separating PD from MSA or PSP. Essential tremor (ET) is another possible diagnosis to consider when evaluating a patient for PD. ET may be easier to distinguish from PD because the former does not typically involve the dopaminergic pathways to the same degree.

Conclusion

Functional imaging using SPECT has provided detailed *in vivo* measurements of cerebral biochemical activity that occur in the normal brain and various neuropsychiatric conditions. The ability to interpret findings depends both on technical issues that may affect the normal pattern of activity and on normal age-related changes. Changes with aging and various neuropsychiatric conditions can be reflected in cerebral blood flow and neurotransmitter concentrations. Overall, SPECT imaging will probably continue to play an important role in the evaluation of brain function in both the clinical and research setting.

See also: Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: MRI Studies; Cognition: An Overview of Neuroimaging Techniques; Depression and the Brain; Map Plasticity and Recovery from Stroke; Neuroimaging; Single Photon Emission Computed Tomography (SPECT); Stroke; Stroke: Neonate vs. Adult; Two-Photon Imaging.

Further Reading

Alavi A and Hirsch LJ (1991) Studies of central nervous system disorders with single photon emission computed tomography and positron emission tomography: Evolution over the past 2 decades. *Seminars in Nuclear Medicine* 21: 58–81.

Booij J, Habraken JB, Bergmans P, et al. (1998) Imaging of dopamine transporters with iodine-123-FP-CIT SPECT in healthy controls and patients with Parkinson's disease. *Journal of Nuclear Medicine* 39: 1879–1884.

Camargo EE (2001) Brain SPECT in neurology and psychiatry. *Journal of Nuclear Medicine* 42: 611–623.

Chugani HT (1992) Functional brain imaging in pediatrics. *Pediatric Clinics of North America* 39: 777–799.

Diksic M and Reba RC (eds.) (1991) *Radiopharmaceuticals and Brain Pathology Studied with PET and SPECT.* Boca Raton, FL: CRC Press.

Goodwin GM (1996) Functional imaging, affective disorder and dementia. *British Medical Bulletin* 52: 495–512.

Kung HF (1991) Overview of radiopharmaceuticals for diagnosis of central nervous disorders. *Critical Review in Clinical Laboratory Science* 28: 269–286.

Masdeu JC, Zubieta JL, and Arbizu J (2005) Neuroimaging as a marker of the onset and progression of Alzheimer's disease. *Journal of Neurological Sciences* 236: 55–64.

Masterman DL, Mendez MF, Fairbanks LA, and Cummings JL (1997) Sensitivity, specificity, and positive predictive value of technetium 99-HMPAO SPECT in discriminating Alzheimer's disease from other dementias. *Journal of Geriatric Psychiatry and Neurology* 10: 15–21.

Mozley PD, Acton PD, Barraclough ED, et al. (1999) Effects of age on dopamine transporters in healthy humans. *Journal of Nuclear Medicine* 40: 1812–1817.

Mozley PD, Sadek AM, Alavi A, et al. (1997) Effects of aging on the cerebral distribution of technetium-99m hexamethylpropylene amine oxime in healthy humans. *European Journal of Nuclear Medicine* 24: 754–761.

Newberg AB, Alavi A, and Payer F (1995) Single photon emission computed tomography in Alzheimer's disease and related disorders. *Neuroimaging Clinics of North America* 5: 103–123.

Pagani M, Salmaso D, Jonsson C, et al. (2002) Regional cerebral blood flow as assessed by principal component analysis and (99m)Tc-HMPAO SPECT in healthy subjects at rest: Normal distribution and effect of age and gender. *European Journal of Nuclear Medicine and Molecular Imaging* 29: 67–75.

Pirker W, Asenbaum S, Bencsits G, et al. (2000) [^{123}I] β-CIT SPECT in multiple system atrophy, progressive supranuclear palsy, and corticobasal degeneration. *Movement Disorders* 15: 1158–1167.

Van Heertum RL and Tikofsky RS (2000) *Functional Cerebral SPECT and PET Imaging.* Philadelphia: Lippincott Williams & Wilkins.

Van Paesschen W (2004) Ictal SPECT. *Epilepsia* 45(supplement 4): 35–40.

Wang J, Jiang YP, Liu XD, et al. (2005) 99mTc-TRODAT-1 SPECT study in early Parkinson's disease and essential tremor. *Acta Neurologica Scandinavia* 112: 380–385.

Weng YH, Yen TC, Chen MC, et al. (2004) Sensitivity and specificity of ^{99m}Tc-TRODAT-1 SPECT imaging in differentiating patients with idiopathic Parkinson's disease from healthy subjects. *Journal of Nuclear Medicine* 45: 393–401.

Single-Nucleotide Polymorphism (SNP) Analysis

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

Single-nucleotide polymorphisms (SNPs) are individual nucleotide positions in the genome that vary at polymorphic levels in human populations. Most SNPs are diallelic (i.e., the SNP site may be occupied by either of two different bases), and the less common base must have a frequency of at least 1%, else it is classed as a rare mutation rather than a polymorphism. SNPs represent the most abundant form of genetic variation in humans, accounting for more than 90% of all differences between unrelated people. SNPs influence many central nervous system phenotypes, including susceptibility to neurological disorders. For example, an important genetic determinant in Alzheimer's disease is the *APOE* gene, which encodes apolipoprotein E (ApoE). The presence of two SNPs in this gene yields three common variants – *APOE* ε2, ε3, and ε4, which produce the proteins ApoE2, E3, and E4. Each of these variants is associated with a different level of disease susceptibility, *APOE* ε4 conferring the greatest risk and *APOE* ε2 the lowest. *APOE* ε4 homozygotes are the highest-risk group for Alzheimer's disease.

As well as affecting disease susceptibility, SNPs also affect our responses to drugs and other xenobiotics. Extending the example presented above, SNPs in the *APOE* gene contribute not only to the risk of Alzheimer's disease but also to drug responses, with the high-risk *APOE* ε4 homozygotes showing the poorest response to tacrine, an acetylcholinesterase inhibitor. SNPs therefore play a major role in all stages of the drug development process, from target identification through to clinical trials. The analysis of SNPs may help scientists tailor drugs and drug regimes to particular genotypes, which is the underlying principle of pharmacogenomics. In this context, it has been necessary to develop techniques for the rapid, high-throughput genotyping of SNPs in large samples.

Principles of SNP Genotyping

All SNP genotyping methods combine two distinct steps – one to discriminate between alternative alleles at the SNP locus and one to report the nature of the allele by generating a signal that can be detected, in an automated fashion if possible.

There are three general allele discrimination methods: hybridization/annealing (sometimes with a downstream enzymatic discrimination step), primer extension, and enzyme cleavage. In each case, the discrimination assay may take place in solution (so-called homogeneous assays) or in a solid phase immersed or coated in liquid, such as a microarray (so-called heterogeneous assays). In many of these assays, it is necessary to amplify the genomic DNA prior to detection of the allele, and amplification may indeed be integrated into the detection step. In other cases, the detection assay is sensitive enough to work directly on nonamplified genomic DNA or complementary DNA.

While allele discrimination methods come in three broad varieties, there are many different ways to generate and detect the signal, some based on fluorescence, others based on enzymatic reactions, and a few based on direct quantification. Most follow the fate of a label, either in real time or at the assay end point. Some methods based on mass spectrometry (MS) measure the masses of allele-specific oligonucleotides, so a label is not required.

Allele Discrimination

Allele-Specific Hybridization

The simplest method for discriminating between alleles at a SNP locus is hybridization using allele-specific oligonucleotide (ASO) probes. Two probes are required, one specific for each allele, and stringency conditions are employed such that a single-base mismatch is sufficient to prevent hybridization of the nonmatching probe (Figure 1). The two probes can be combined in labeled/unlabeled pairs, in which case two separate assays are required for each SNP (one in which probe 1 is labeled and probe 2 is unlabeled, and one in which these roles are reversed). Alternatively, and preferably in the case of high-throughput assays, distinct labels can be used for each probe (e.g., different fluorophores or mass tags) so that each SNP can be genotyped in a single reaction.

Allele-specific polymerase chain reaction (PCR) is an extension of the allele-specific hybridization principle, in which discrimination is achieved by allele-specific primer annealing followed by PCR amplification rather than the direct detection of a hybridized probe (Figure 2). Three primers are required, two to match alternative alleles at the SNP site and one to anneal to the opposite strand. Two reactions are set up, each containing one of the SNP primers together with the reverse primer. Although the variable base position can be sited somewhere in the middle of the SNP primers, in a manner directly analogous to the ASO strategy, this method tends to allow the extension of a mismatched primer–template because the

stringency of annealing cannot be optimized (annealing has to favor both primers in the reaction and is therefore difficult to optimize for the allele-specific primer alone). The preferred approach is to place the 3′ end of one of the primers at the variable base position since extension is dependent on perfect complementarity at the 3′ end of the primer.

A further extension of the allele-specific hybridization approach is allele-specific ligation, in which two oligonucleotides are annealed at adjacent sites and DNA ligase is used to join them together. One of the oligonucleotides has two allele-specific variants, and the discriminatory base is at the most 3′ position. If there is perfect complementarity between the allele-specific oligonucleotide and the template, it will be successfully ligated to the common oligonucleotide. If there is a mismatch, then ligation will fail, and the two oligonucleotides will remain separate. The assay benefits from the inclusion of additional steps such as the ligase chain reaction (in which the ligation reaction is reiterated with further oligonucleotides complementary to the first ligation product) or rolling circle amplification technology, in which two allele-specific oligonucleotides about 80 nt in length are designed to form a circle on the template, generating a closed loop (or padlock probe) when ligated. Primers annealing to this circle can be extended with a strand-displacing DNA polymerase so that when the nascent strand completes the circle and encounters itself, it is continually displaced, generating a long concatemer that is easy to detect using fluorescence methods.

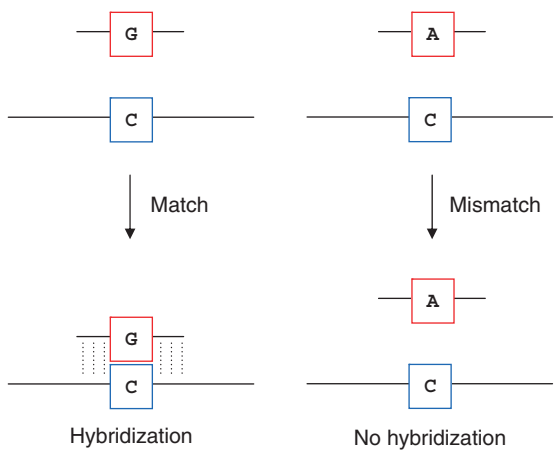

Figure1 Principles of allele-specific hybridization as an allele discrimination method.

Minisequencing

Minisequencing is also known as allele-specific primer extension. Like allele-specific PCR, it involves the extension of a primer annealing to the template DNA, but the primer in minisequencing is nondiscriminatory. The primer anneals one base upstream of the polymorphic site, and discrimination is achieved by detecting which base is incorporated in the first polymerization step (Figure 3). Allele-specific primer extension methods are more adaptable than hybridization/annealing assays because a much greater diversity of labeling strategies can be used. The free nucleotides in solution can be labeled with four different fluorescent tags, mass tags, or haptens, allowing the same mix to be used in the detection of many SNPs in a highly multiplexed fashion, such as on a microarray. Primer extension methods for genotyping have been developed commercially, for example as Genetic Bit Analysis technology by Orchid Biosciences, and they form the basis of many popular commercial genotyping systems, such as SNaPshot (Applied Biosystems).

Allele-Specific Enzymatic Cleavage

The use of enzymes to distinguish between alternative alleles is the basis of one of the earliest genotyping

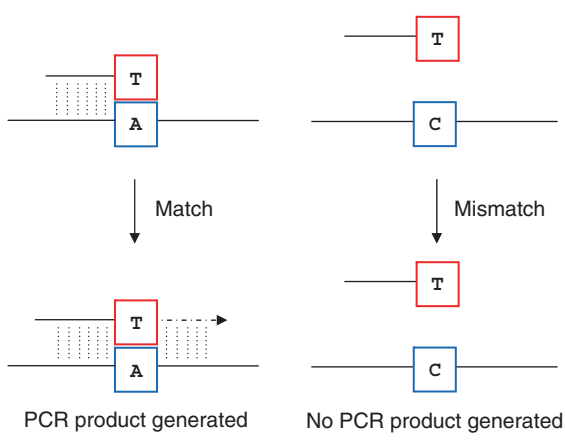

Figure 2 Principles of allele-specific polymerase chain reaction (PCR) as an allele discrimination method.

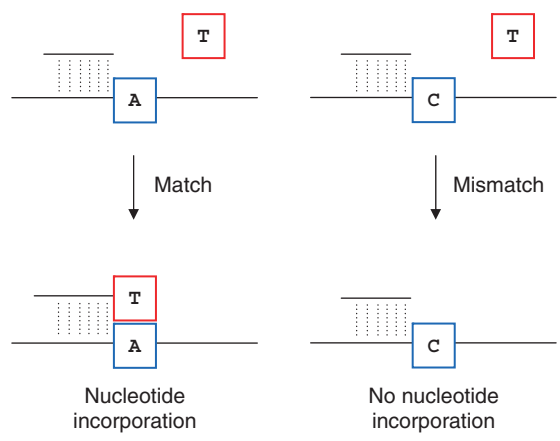

Figure 3 Principles of allele-specific primer extension (minisequencing) as an allele discrimination method.

methods – the analysis of restriction fragment length polymorphisms (RFLPs). An RFLP is generated when a single nucleotide polymorphism occurs at a restriction endonuclease recognition sequence, and one allele preserves the sequence while the other destroys it. In a sequence of genomic DNA with three adjacent restriction sites, the middle one containing a SNP, then digestion of amplified genomic DNA with the appropriate restriction endonuclease will produce either a single large fragment (if the central restriction site is absent) or two smaller fragments (if the central restriction site is present and cleavage occurs).

Traditional RFLP analysis cannot be applied on a large scale because it relies on gel electrophoresis and Southern blot hybridization, which are cumbersome to use in high-throughput experiments. Modern takes on RFLF analysis which combine restriction enzymes and PCR amplification can accelerate this process and facilitate moderate multiplexing (e.g., amplified fragment length polymorphisms and cleaved amplified polymorphic sites). However, these have been used more for linkage analysis than for genotyping. One enzymatic assay developed for genotyping is the Invader assay, which employs a unique method of allelic discrimination involving overlapping probes and an enzyme that specifically recognizes the resulting 'flap' (Figure 4). An advantage of the Invader assay is that it requires no PCR amplification. Two signal probes are used, one recognizing each allele, plus an invader probe. The signal and invader probes hybridize in tandem, and the signal probe overlaps the invader probe, generating a flap that is recognized by an enzyme called cleavase, a modified form of the thermostable FEN-1 (for flap endonuclease) enzyme. A flap with the appropriate structure is generated only if the signal probe completely matches the template and there is a one-base invasion. The cleaved flap can then be used in a second-round invader assay to generate a fluorescent signal, or the cleaved flap can be detected by MS.

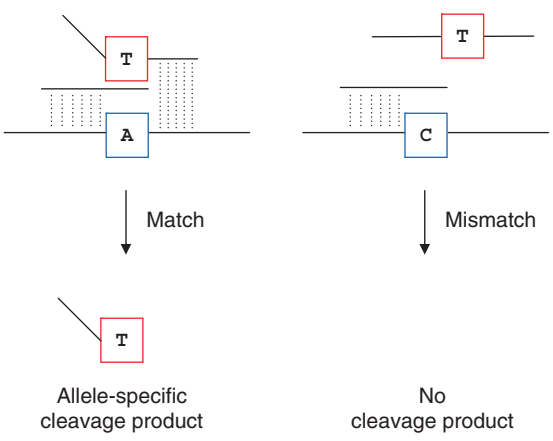

Figure 4 Principles of allele-specific invasive cleavage as an allele discrimination method.

Allele Detection

Fluorescence Methods

Direct detection of incorporated fluorescent labels The most straightforward way to detect an allele-specific product is to incorporate one or more nucleotides conjugated to a fluorescent dye. Direct fluorescence detection is generally used with heterogeneous assay formats (microarrays, bead arrays) and is particularly suitable for minisequencing methods since single-base extension can be carried out with four nucleotides bearing different fluorophores, which allows the immediate detection of alleles on the basis of the emission wavelength of the extension product. Direct fluorescence is also often used when gel or capillary electrophoresis is used to separate allele-specific products.

Direct detection of intercalated fluorescent labels Intercalating fluorescent dyes such as ethidium bromide fluoresce much more brightly in the presence of double-stranded DNA than single-stranded DNA. This is the basis of a unique detection method known as dynamic allele-specific hybridization, which involves a single (unlabeled) allele-specific oligonucleotide probe that is added to a PCR product containing an SNP site and allowed to anneal when ethidium bromide is present. The mixture is then heated, and fluorescence is measured in real time as the temperature rises. The fluorescence decreases as denaturation (melting) occurs, and the thermal melting profile for each allele will be different because only one is perfectly complementary to the probe and the other contains a mismatch.

Fluorescence resonance energy transfer Fluorescence resonance energy transfer (FRET) occurs when two fluorophores are in proximity and one of them (the donor) has an emission spectrum that overlaps the excitation spectrum of the other (the acceptor). When a lone donor fluorophore is excited, light is produced with a characteristic emission spectrum. However, when the donor fluorophore is excited in proximity to the acceptor, energy is transferred to the acceptor fluorophore with the result that the intensity of emission from the donor is reduced (quenched) while that of the acceptor is enhanced.

Although FRET enhancement can be used for SNP genotyping, two of the most popular homogeneous assays in current use employ FRET quenching. The *Taq*Man assay exploits the intrinsic 5′ nuclease activity of *Taq* DNA polymerase to generate a fluorescent

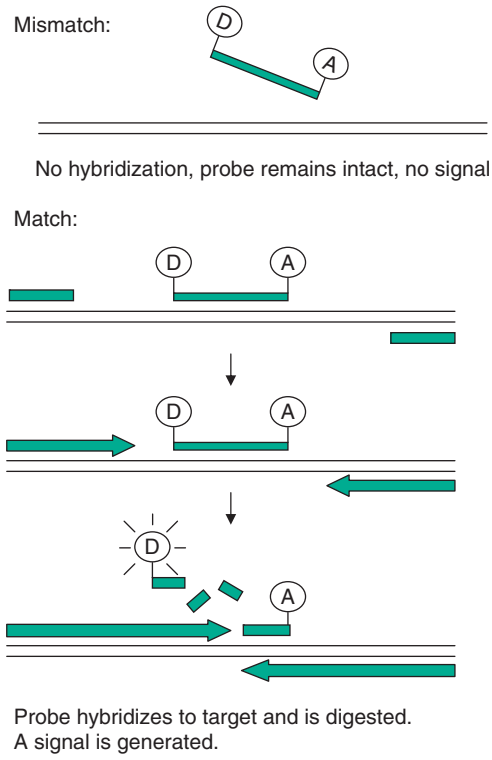

Figure 5 Principles of the *Taq*Man assay; A, acceptor fluorophore; D, donor fluorophore.

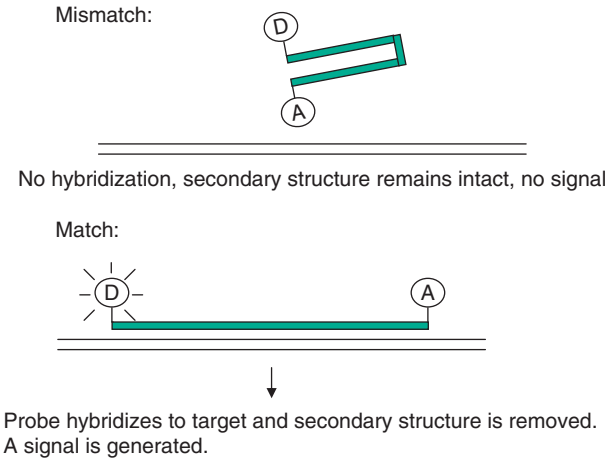

Figure 6 Principles of the molecular beacon assay; A, acceptor fluorophore; D, donor fluorophore.

signal from a short allele-specific oligonucleotide probe. Two probes are required, one specific for each allele. Each probe contains a unique donor fluorophore and a common acceptor fluorophore and is short enough for the donor to be quenched when the probe is intact (either hybridized or in solution). When the probe hybridizes to the PCR template, however, the 5′ exonuclease activity of *Taq* DNA polymerase digests it, thus releasing the two fluorophores into solution and eliminating the quenching effect (Figure 5). The molecular beacon assay involves the use of longer probes that have self-complementary ends labeled with the donor and acceptor fluorophores. Nonhybridized probes will self-anneal, bringing the donor and acceptor fluorophores together, resulting in quenching. Where hybridization occurs, the donor and acceptor fluorophores are separated such that the quenching effect is eliminated and a fluorescent signal is produced. As with the *Taq*Man assay, two probes labeled with different donor fluorophores are required, one specific for each allele (Figure 6).

Fluorescence polarization When a fluorophore is excited by plane polarized light, the fluorescence emitted by the dye is also polarized, a phenomenon termed fluorescence polarization (FP). Complete FP occurs only when a dye molecule is stationary. The degree of FP thus depends on the movement of molecules in solution, which in turn depends on various factors, including the molecular mass. SNP genotyping assays can change the molecular mass of a target oligonucleotide, for example by primer extension, probe hydrolysis, or invasive cleavage, and this is the basis of commercial genotyping systems such as SNaPshot. Such changes can be recorded as changes in FP as long as the temperature and viscosity of the solution remain the same.

Pyrosequencing

Pyrosequencing is a novel method for sequencing short stretches of DNA on the basis of the detection of pyrophosphate, a normal by-product of DNA synthesis. Although similar in principle to primer extension allele discrimination methods, pyrosequencing is suitable not only for typing SNPs but also for scoring entire haplotypes (groups of linked SNPs). Pyrosequencing works on the basis that pyrophosphate can be used to generate adenosine triphosphate (ATP), which then stimulates luciferase activity, causing the emission of a chemiluminescent signal. To achieve this, the primer extension reaction must include APS and the enzyme ATP sulfurylase, which converts adenosine 5′ phosphosulfate into ATP in the presence of pyrophosphate. Also present are luciferin (the substrate of luciferase) and the enzyme apyrase, which continuously degrades unincorporated deoxynucleotide triphosphates (dNTPs) and excess ATP. The dNTPs are added to the reaction one by one. Only if the incoming dNTP is complementary to the template will it extend the primer and release pyrophosphate, resulting in the production of an equimolar amount of ATP. Visible light is generated in proportion to the amount of ATP and is detected as a peak on a pyrogram. When the degradation of the present dNTP is completed by apyrase, the next is added. A thioderivative of dATP must be used to avoid constant stimulation of luciferase activity.

Mass Spectrometry

Genotyping assays involving MS are distinct from those discussed above in that signal detection is direct, based on differing molecular weights of small DNA fragments rather than the behavior of a label. The analysis of DNA by MS requires soft ionization (i.e., without fragmentation) and is usually achieved by matrix-assisted laser desorption/ionization (MALDI) followed by time of flight (TOF) analysis. The MALDI procedure involves mixing the allele-specific products of the discrimination assay with a matrix compound on a metal plate. The mixture is then heated with a short laser pulse, causing it to expand into the gas phase, where ionization is achieved by application of a strong potential difference. Ions are accelerated toward the detector, and the TOF (the time taken to reach the detector) is measured, allowing the mass–charge ratio to be calculated.

High-resolution mass spectrophotometers can accurately discriminate between alternative alleles in DNA fragments of three to 20 nt in length. The smallest difference in mass between alternative allelic products is about 9 Da, representing an adenosine–thymidine polymorphism. Advantages of MS-based methods include accuracy, as long as the sample is very pure, and the fact that each genotyping reaction takes a fraction of a second, which allows thousands of reactions to be carried out serially in a single day.

The majority of current MS genotyping platforms involve the use of DNA chips because the chip can be used directly as a MALDI plate. Thus far, the technique has been used to determine the masses of allele-specific hybridization probes, allele-specific primer extension products, and the products of invasive cleavage reactions. Commercial systems utilizing MS include MassArray (Sequenom) and PinPoint (PerSeptive Biosystems).

Another promising MS genotyping technology involves the use of cleavable mass tags (chemical moieties of known molecular mass) as labels to replace fluorophores. The advantage of mass tags is that, because of the accuracy of MS-based mass determination, the range of labels that can be used is virtually limitless, so ultrahigh-throughput genotyping is possible. The MassCode system (Qiagen Genomics) uses this technology.

Summary

High-throughout SNP genotyping has helped facilitate association studies aiming to find patterns among the 0.1% of variation in the human genome (mostly represented by SNPs) corresponding to differences in complex neurological phenotypes such as susceptibility to neuropathophysiological disorders, psychiatric disorders, and responses to drugs. Many of the technologies described above facilitate highly multiplexed analysis, although MS stands out as a platform in which increased throughput is realized by ultrarapid serial processing.

The sensitivity of most genotyping techniques reflects some form of amplification step either before, during, or after the actual allele discrimination assay. In the majority of cases, this amplification step relies on the PRC, but novel formats such as the invader assay and rolling circle amplification technology are also becoming popular. While homogeneous assay formats are versatile, array-based technology platforms have the advantage of higher throughput. Recent developments include the use of electrostringent hybridization arrays, in which electric currents are used to increase the speed and sensitivity of hybridization, and the detection of allele-specific products on the basis of their electrochemical properties when hybridized to oligonucleotide probes attached to minute electrodes. Another emerging technology, bead arrays, combines the flexibility of the homogeneous assay format with the convenience and throughput of chip-based detection systems. Bead array systems that employ fluorescence-based detection and sorting are commercially available from Lynx Therapeutics and Illumina.

See also: Alzheimer's Disease: Molecular Genetics; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Fluorescence Microscopy in the Neurosciences; Genomics of Brain Aging: Apolipoprotein E; Single Cell Genomic DNA Analysis.

Further Reading

Brookes AJ (1999) The essence of SNPs. *Gene* 234: 177–186.

Campbell DA, Valdes A, and Spur N (2000) Making drug discovery a SN(i)P. *Drug Discovery Today* 5: 388–396.

Griffin TJ and Smith LM (2000) Single nucleotide polymorphism analysis by MALDI-TOF mass spectrometry. *Trends in Biotechnology* 18: 77–84.

Kwok P-Y (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics & Human Genetics* 2: 235–258.

Syvänen A-C (1999) From gels to chips: 'Minisequencing' primer extension for analysis of point mutations and single nucleotide polymorphisms. *Human Mutation* 13: 1–10.

Syvänen AC (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* 2: 930–942.

Taylor JG, Choi E-H, Foster CB, and Chanock SJ (2001) Using genetic variation to study human disease. *Trends in Molecular Medicine* 7: 507–512.

Twyman RM (2004) SNP discovery and typing technologies for pharmacogenomics. *Current Topics in Medicinal Chemistry* 4: 1421–1429.

Twyman RM and Primrose SB (2003) Techniques patents for SNP genotyping. *Pharmacogenomics* 4: 67–79.

siRNA: Utility

D M Dykxhoorn, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

RNA interference (RNAi) is a highly conserved, naturally occurring cellular process that mediates the sequence-specific silencing of gene expression. The discovery of this process has challenged the long-held idea that the flow of genetic information occurs in a unidirectional manner from DNA to RNA to protein. In fact, it is now clear that RNA can act bidirectionally, with certain types of RNA being translated into proteins, whereas other forms can direct the regulation of gene expression. Regulatory RNAs have been shown to regulate a variety of processes, including transcription, splicing, and translation. In addition, the discovery of these regulatory RNAs has helped to begin to explain why the genomes of complex organisms retain large amounts of DNA that lack protein coding genes, referred to as intergenic regions. Analysis of the total RNA content, the so-called transcriptome, of a variety of organisms has shown that the transcriptional output of many organisms' genomes is much greater than the number of protein coding genes and that the intergenic regions of the genome are rich sources for noncoding RNAs, the majority of which are probably involved in various regulatory processes.

Small Regulatory RNAs

The small regulatory RNAs involved in RNAi have garnered a great deal of attention, both for their biological role as endogenous regulators of gene expression and as research tools for the selective suppression of gene expression. These small RNA molecules, ranging in size from approximately 19 to 26 nucleotides (nt), are highly conserved across diverse organisms from plants to worms and mammals and have been found to be involved in the regulation of a variety of cellular processes, including development and differentiation, metabolism, and the suppression of mobile genetic elements (e.g., transposons and retrotransposons) and viruses. Although micro (mi)RNAs were the first class of small regulatory RNA molecules identified, with the discovery that let-7 and lin-4 were noncoding RNAs involved in the regulation of genes responsible for proper developmental timing in the nematode *Caenorhabditis elegans*, the impact of this discovery was not immediately appreciated. It was the demonstration that the introduction of long double-stranded RNA (dsRNA) molecules into *C. elegans* led to the sequence-specific suppression of gene expression that signaled the beginning of a new era in molecular biology. It was for this discovery, termed 'RNA interference,' that Drs. Andrew Fire and Craig Mello were awarded the 2006 Nobel Prize in Physiology or Medicine. In rapid succession, it was shown that these long dsRNAs were rapidly cleaved into small duplexed RNA molecules of 19–21 nt, referred to as small interfering (si)RNAs, and that these siRNAs were the effector molecules that guided the degradation of the homologous mRNA. It soon became clear that RNAi was related to a variety of silencing phenomena, including quelling in fungus and posttranscriptional gene silencing in plants. Since these initial discoveries, an increasing number of small regulatory RNAs have been identified using biochemical, genetic, and computational approaches. These small RNA species include miRNAs, siRNAs, repeat-associated siRNAs (rasiRNAs), and piwi-interacting RNAs.

Both miRNAs and siRNAs are processed from dsRNA precursor molecules – long hairpin RNAs for miRNAs and long dsRNAs for siRNAs – by RNase III-type endonucleases (**Figure 1**). miRNAs are expressed as long transcripts, termed primary miRNAs (pri-miRNAs), from RNA polymerase II or III promoters. These highly structured pri-miRNAs are recognized by the nuclear localized RNase III-type enzyme Drosha that cleaves the pri-miRNA into an approximately 60–70 nt precursor miRNA (pre-miRNA) stemloop containing both the mature miRNA and the accompanying passenger strand separated by a noncomplementary loop region. This truncated stemloop is recognized by exportin-5 and translocated from the nucleus to the cytoplasm. In the cytoplasm, it is cleaved by the RNase III-type endonuclease Dicer into a short-lived small (~21 nt) duplexed RNA containing 2 or 3 nt 3′ overhangs characteristic of an RNase III-type cleavage reaction. In addition to producing miRNAs, Dicer is also able to cleave long dsRNAs into siRNAs. In fact, Dicer was originally identified based on its ability to process long dsRNAs. Additional dsRNA-binding proteins help Dicer to distinguish between these two precursor molecules. Interestingly, in plants, the pre-miRNA processing and long dsRNA cleavage activities of Dicer have been segregated into two RNase III family members, Dicer-like 1 (DCL-1) and DCL-2, respectively. In each case, the small duplexed RNAs, either the siRNAs or the miRNA, are loaded into an effector complex whereupon the passenger strand is dissociated from the active silencing strand (guide strand). Both siRNAs and miRNAs have an asymmetric nature that helps to distinguish the guide strand from the passenger strand. This asymmetric nature is determined by the relative thermodynamic

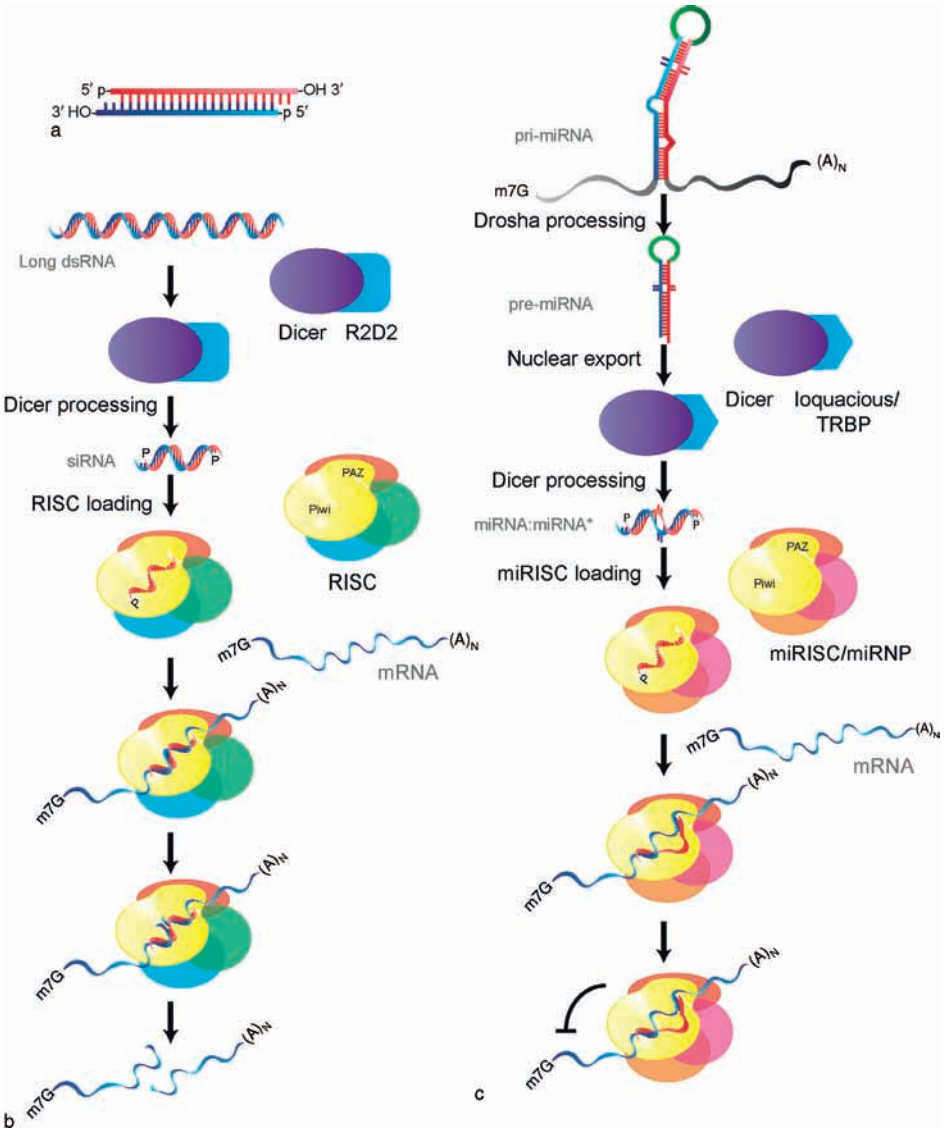

Figure 1 RNA interference pathways. (a) Biochemical structure of siRNAs. siRNAs have the characteristic structure of an RNase III cleavage product with 5′ phosphorylated ends, a 19 nt duplexed region, and 2–3 nt overhangs at the 3′ termini. (b) The 'classic' RNAi pathway. Long dsRNAs were the first 'trigger' molecules associated with RNAi. These long dsRNAs were processed by the RNase III family member Dicer into siRNAs that serve as the effector molecules that guide the sequence-specific silencing of gene expression. Although Dicer can cleave the long dsRNAs in the absence of additional factors, the dsRNA-binding protein R2D2 helps to identify the end of the siRNA that has the lowest thermodynamic stability, a mechanism used to differentiate the active guide strand from the passenger strand. The duplexed siRNA enters a multiprotein complex, termed the RNA-induced silencing complex, where the passenger strand is cleaved and released, leaving the guide strand to identify the target mRNA sequences. Although the structure of RISC is not well understood, the key component of this complex is the Argonaute protein, Ago2, which is responsible for holding the siRNA (the 5′ phosphorylated termini of the siRNA binding to a cleft in the Piwi domain and the 3′ termini interacting with the PAZ domain) in the proper conformation to facilitate its interaction with target mRNA. In addition, Ago2 contains the endonuclease activity, located within the Piwi domain, responsible for cleaving the target mRNA. This pathway can be harnessed for the experimental silencing of gene expression by providing either the long dsRNA trigger molecules, a technique commonly used in *C. elegans* and *Drosophila*, or by directly transducing cells with the siRNA effector molecules. Although the RNAi machinery is highly conserved from plants to mammals, endogenous siRNAs, which are present in lower eukaryotes, have not been identified in mammalian cells. (c) The microRNA pathway. miRNAs, the most sequence-diverse class of small RNA regulatory molecules, are encoded on long, highly structured primary transcripts (pri-miRNA). These transcripts can be produced by RNA polymerase II or III and can encode a single miRNA or a cluster of miRNAs. These pri-miRNAs are processed by the nuclear localized RNase III family member Drosha, which recognizes the structure of the miRNA-encoding RNA hairpin and not the specific miRNA sequence. The Drosha-mediated cleavage reaction releases a short hairpin RNA, termed the precursor (pre)miRNA. This approximately 60–70 nt pre-miRNA is translocated to the cytoplasm by Exportin 5, where it is cleaved by Dicer into a dsRNA duplex containing the mature miRNA and the usually short-lived passenger strand (miRNA*). Similar to the long dsRNA-directed RNAi pathway, Dicer is associated with a dsRNA-binding protein called loquacious in *Drosophila* or TRBP (HIV-1 TAR RNA-binding protein) in mammals. The duplexed miRNA molecule is taken up by the effector complex, variously termed the

properties of each of the 5′ termini, with the end having the lower stability being preferentially maintained in the effector complex and used to guide the silencing reaction. Although this effector complex, variously termed the RNA-induced silencing complex for siRNAs or the miRNA-containing RNA-induced silencing complex (miRISC) for miRNAs, is poorly defined, it contains at its core a member of the Argonaute (Ago) family of proteins. The guide strand of the siRNA or the mature miRNA provides the specificity determinant for the effector complex by binding to complementary sites on the target mRNA. Gene silencing can take place by one of two mechanisms, either the inhibition of translation or the endonucleolytic cleavage of the mRNA. The method of silencing used for any target depends largely, but not entirely, on the degree of sequence complementarity between the siRNA or miRNA and the cognate mRNA. Those interactions with complete or nearly complete homology lead to mRNA cleavage, whereas interactions with a lower level of homology induce translation inhibition.

The analysis of these silencing mechanisms has not only allowed for a better understanding of the role that small RNA molecules play in the regulation of many important cellular processes but also facilitated the development of technologies for the targeted silencing of gene expression. Although long dsRNAs had been used for silencing gene expression in *C. elegans*, this technology was not amenable for use in higher eukaryotes due to innate immune responses that have evolved in mammalian cells to recognize long dsRNA species and induce the activation of cell death pathways. In normal circumstances, the interferon response serves to protect cells from viral pathogens, many of which produce long dsRNAs as intermediates in their life cycle. Since the interferon response is efficiently triggered by dsRNAs more than 30 nt in length, it was believed that siRNAs, which fall below this size threshold, would be unable to activate this response. This was borne out by experiments that showed that the treatment of mammalian cells with chemically synthesized siRNAs could effectively silence gene expression without any untoward toxicity. This discovery led to the development of RNAi-based technologies allowing reverse genetic experiments to be performed in organisms that were intractable to more traditional gene-silencing approaches and laid the groundwork for the application of RNAi-based silencing approaches as novel therapeutic modalities. In fact, the use of siRNAs for the localized silencing of gene expression has already entered clinical trials for the treatment of age-related macular degeneration and Rous sarcoma virus infections.

siRNAs as a Gene-Silencing Tool

There are several features that make the use of siRNAs an attractive approach for the targeted silencing of gene expression. Foremost among these features is the fact that these silencing technologies take advantage of an endogenous gene-silencing mechanism that is conserved in organisms as diverse as plants, *C. elegans*, *Drosophila*, mice, and humans. In addition, unlike antisense approaches, potential target sequences can be readily identified. Initial studies showed that siRNA-mediated gene silencing was highly specific, being able to distinguish between genes that differed by as little as a single nucleotide, with no apparent toxicity.

siRNA Sequence

A better understanding of the characteristics that distinguish highly potent siRNAs from nonfunctional siRNAs will facilitate the optimization of siRNA design to maximize silencing potential and minimize off-target effects. There are a number of traits associated with highly active siRNAs, including low internal stability, moderate to low GC content (optimally between 30% and 52%), lack of secondary structures within the siRNA sequence, and low thermodynamic stability at the 5′ end of the guide strand. In addition, siRNA sequences that have homology with unintended targets or that contain immunostimulatory sequence motifs should be avoided to reduce the potential off-target effects. Many of these criteria have been incorporated into computer algorithms designed to aid in the identification of effective siRNA sequences targeting a particular gene of interest. Although these algorithms can improve the selection of potential siRNA sequences, experimental testing is necessary to determine the optimal siRNA sequence for any given gene.

siRNA Specificity

Early studies using siRNAs to mediate gene silencing suggested that siRNAs, by virtue of their action,

miRNA-containing RISC (miRISC) or the miRNA nucleoprotein (miRNP) complex; the passenger strand (miRNA*) is dissociated; and the mature miRNA guides the complex to the target mRNA. The binding of the miRNA to target sites that have partial complementarity leads to the induction of translational repression. Although these pathways represent the classic views of the RNA interference pathways that induce posttranscriptional gene silencing, in actuality, siRNAs and miRNAs have interchangeable effector functions and either molecule can induce translational repression or mRNA cleavage depending on the degree of target site complementarity.

required full complementarity and that a single nucleotide difference between the siRNA and the target binding site would be sufficient to abrogate silencing. In fact, it appears that mismatches between the siRNA and the target site can be accommodated without the complete loss of silencing. The degree to which mismatches can be tolerated depends on a number of factors, including the position and nucleotide composition of the mismatch. For example, the central nucleotides of the siRNA, around the site where cleavage of the mRNA occurs, are more sensitive to mismatches than either the 5′ or 3′ ends of the molecule. This is presumably due to steric constraints imposed by the active site of the Ago2 endonuclease, the core enzymatic component of RISC. It is not simply the presence of mismatched nucleotides that lends specificity but, more important, it is the nature of the mismatched nucleotides. That is, the juxtaposition of two purines within the central region of the siRNA:mRNA binding site will interfere with silencing more readily than the juxtaposition of two pyrmidines or a non-base-pairing purine:pyrmidine pair. However, by judiciously choosing the site and type of mismatch, it is possible to selectively silence genes that differ by as little as a single nucleotide. This has permitted the development of silencing tools that can distinguish between disease-causing alleles of a gene that differ by a single nucleotide from the 'normal' allele. This has been applied to disease causing single nucleotide polymorphisms (SNPs) involved in the development of cancer (e.g., RasV12), sickle cell anemia, and several dominantly inherited neurodegenerative disorders including superoxide dismutase 1 in amyotrophic lateral sclerosis and tau in frontotemporal dementia. Even for diseases whose phenotype is not directly attributable to an SNP, selective targeting of the disease can be accomplished by targeting SNPs that are linked to the disease allele. This has been used for the preclinical development of tools for the silencing of MJD1 (ataxin-3) in Machado–Joseph disease by targeting an SNP that tightly segregates with the pathogenic allele.

Off-Target Effects

Although the initial studies using siRNAs for the targeted silencing of gene expression showed a high level of specificity and a lack of toxicity, as the technology has become more widely applied, the potential for off-target effects has become more apparent. The off-target effects of siRNAs are classified into two categories: (1) the silencing of unintended targets that have partial complementarity to the siRNA and (2) the induction of cellular responses that nonspecifically silence gene expression. The analysis of mRNA levels in cells treated with specific siRNAs has shown changes to a large number of genes. Although the majority of these changes are small, a few genes can be severely affected. These nonspecific effects have been found for targets that share as few as 15 complementary nucleotides with as few as 11 contiguous nucleotides. These nonspecific effects can be generated from either strand of the duplexed siRNA. Microarray analysis can effectively measure changes in the amount of mRNA but fails to represent the changes that may occur due to translational inhibition, the common form of RNAi-mediated gene silencing induced by miRNAs in animal cells. Since miRNAs require a much lower level of complementarity, the potential off-targets could be significantly higher. It is difficult to evaluate these changes in protein levels since high-throughput proteomic approaches are much more difficult. Since effective translational repression mediated by miRNAs requires the concerted action of several miRNAs, the overall impact of these miRNA-style off-target effects may be minimal.

As part of the innate antiviral response, mammalian cells have mechanisms for the recognition of dsRNA species, a common intermediate found in the life cycle of many viruses. This response induces the global inhibition of gene expression by impairing the cellular translation machinery. Although initial studies using siRNA for gene silencing found no induction of the interferon response with dsRNA smaller than 30 nt, subsequent studies using highly sensitive microarray analysis found that in certain circumstances expression of a subset of interferon-responsive genes was upregulated. This induction of the interferon response was found for only certain siRNA sequences and only when cells were treated with high concentrations of siRNAs. The treatment of cells with different siRNA sequences led to an increase in the expression of different subsets of interferon genes. Since no cytotoxicty was associated with any of these siRNA treatments, it appears that only a partial interferon response was induced. Experiments have suggested that the induction of the interferon genes may not be through the dsRNA-activated protein kinase; instead, interferon gene expression may be indirectly induced by the activation of Toll-like receptors (TLRs), TLR7 and TLR8. These TLRs can recognize specific immunostimulatory sequence motifs, particularly GU-rich sequences. The mode of administration of the siRNA affected the ability of the siRNA to stimulate an interferon response, suggesting that the method of cellular uptake plays a role in determining the immunostimulatory potential of an siRNA. For example, transduction of the cell by lipid-based transfection reagents resulted in a pronounced enhancement of off-target gene expression compared to delivery by electroporation.

Since these off-target effects are concentration dependent, many of them can be avoided by minimizing the siRNA dosage. This can be accomplished by identifying highly potent siRNAs that can efficiently silence target gene expression at doses below the threshold necessary to induce off-target effect. Potential off-target effects can be minimized by careful siRNA sequence selection. Important criteria to reduce the potential for off-target effects include the avoidance of immunostimulatory sequences and designing the siRNA to minimize uptake of the passenger strand into RISC. In addition, all the potential siRNA sequences should be compared to the mRNA in the human genome to minimize the potential off-target binding sites. Although it may be impossible to avoid short stretches of homology, siRNAs that have extensive homology with unintended targets should not be used. In addition, the incorporation of chemical modifications within the siRNA can mitigate the immunostimulatory effects of the siRNA while retaining its silencing efficiency.

RNAi-Based Silencing Technologies

Two basic approaches have been used for the targeted silencing of gene expression by siRNAs: (1) the introduction of chemically synthesized siRNAs into cells and (2) the introduction of DNA vectors expressing siRNA precursors that are processed intracellularly into active siRNAs (Figure 2). Each of these methods has distinct advantages and disadvantages.

Chemically Synthesized siRNAs

The demonstration that chemically synthesized siRNAs could induce the silencing of a target gene of interest without activating an innate immune response has revolutionized biological research, providing a powerful tool for gene function analysis and opening up the possibility for the therapeutic application of these potent technologies. The transfection of synthetic siRNAs into tissue culture cells remains the most widely used method for gene silencing because of its specificity, ease of use, and adaptability to a variety of experimental conditions and cell types. An increasing number of lipid-based transfection reagents have been developed to effectively deliver the siRNAs into the cytoplasm of target cells. There are a number of factors that have limited the use of this technology, principal among these are the transient nature of the siRNA-based silencing response in actively dividing cells and the inability of traditional transfection reagents to transduce many important cell types. In most dividing cell lines, the silencing phenotype peaks at approximately day 3 posttransfection and subsides approximately 5–7 days posttransfection. This loss of silencing over time appears to be due to the diluting out of the siRNAs during cell division until the intracellular concentration of the siRNA falls below a critical threshold needed to maintain the silencing phenotype. Long-lasting silencing (up to 3 weeks) has been achieved in cells that are nondividing, as demonstrated by *in vitro* experiments in postmitotic, terminal differentiated primary macrophage cells and *in vivo* delivery to hepatocytes. The degree to which a specific protein can be silenced depends on the half-life of the protein, the effectiveness of the siRNA, and the rate of cell doubling. Since the silencing phenotype is mediated by the targeted degradation of mRNA and not by degrading the protein, it is possible to effectively deplete the cell of a specific mRNA without altering the levels of the protein. This is particularly the case for proteins that are long-lived. Although the transfection of a wide variety of tissue culture cells is possible with the judicious choice of transfection reagents, there are a large number of cells that are refractory to lipid-based transfection protocols, including many clinically important primary cell types. In addition, the inherent toxicity of many of these lipid-based transfection reagents limits their usefulness for *in vivo* applications. Therefore, a variety of alternative transduction mechanisms have been developed (Figure 2(a)). One approach for the *in vitro* delivery of siRNAs to cell types that are difficult to transfect has been the combination of electroporation technologies with lipid-based transfection reagents, termed nucleofection. Although this approach can have some untoward toxicity, it has been effectively used for the silencing of gene expression in a wide range of cell types, including many primary cells. Another approach that has been proved to be effective for the delivery of siRNAs both *in vitro* (in tissue culture cell experiments) and *in vivo* has been the conjugation of the siRNA with cholesterol. Not only did this lead to the efficient uptake of the siRNA while retaining the siRNA's silencing activity but also it improved the pharmacological properties of the siRNA by increasing the retention of the siRNA within the circulatory system. Since cholesterol receptors are relatively ubiquitously expressed, the cholesterol-conjugated siRNAs can enter a wide variety of cells. The broad spectrum of cells that can be effectively transduced by cholesterol-conjugated siRNAs makes this technology widely applicable. However, when cholesterol-conjugated siRNAs are given *in vivo*, they will be delivered to a wide variety of cells and tissues, only some of which are the intended target. This increases the risk of silencing gene expression in unintended cell types and could increase off-target effects. In addition, this nonspecific uptake increases the amount of the

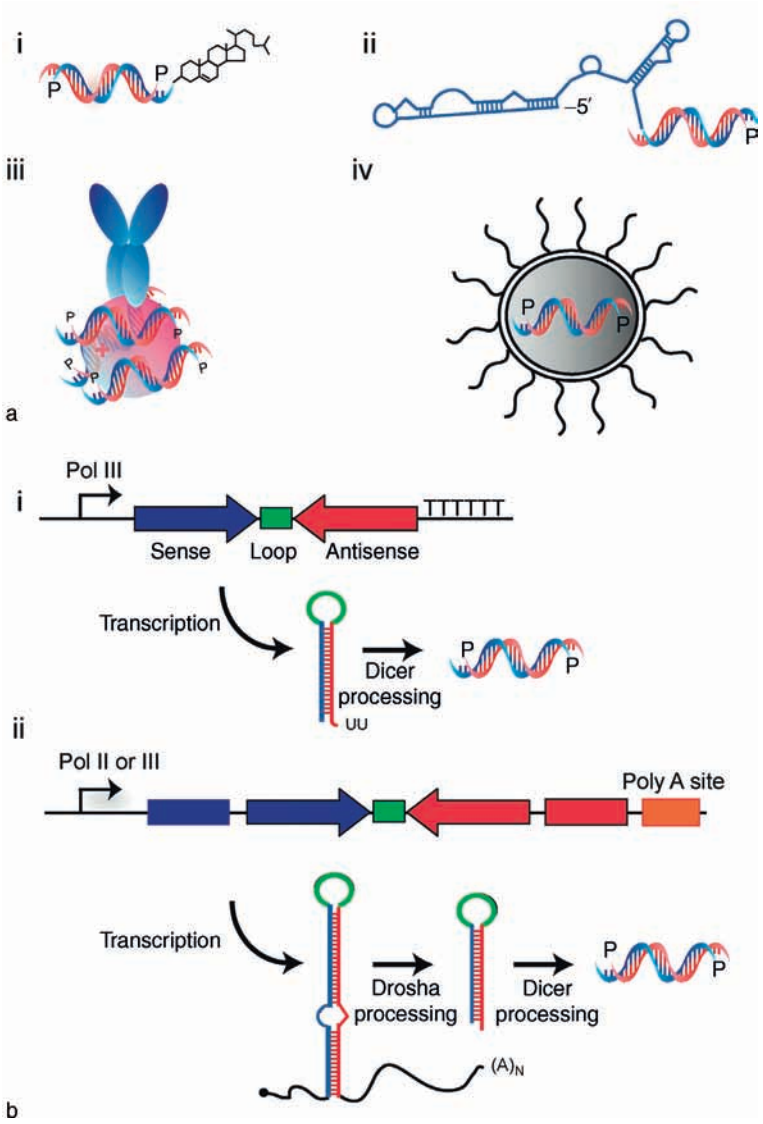

Figure 2 Methods for siRNA delivery. (a) Methods for the delivery of chemically synthesized siRNAs into cells. Although the use of lipid-based transfection reagents remains the most common method for the transduction of siRNAs into tissue culture cells, several methods have been developed that allow for the efficient delivery of siRNAs into cells for both *in vitro* and *in vivo* gene-silencing experiments. The covalent coupling of the siRNA, through the passenger strand, with either a cholesterol moiety (i) or an RNA aptamer (ii) has been shown to increase the pharmacological properties of the siRNAs (e.g., increasing the retention time of the siRNA within the circulatory system) while facilitating the uptake of the siRNA and maintaining its silencing capacity. Cholesterol permits the binding of the siRNA to cellular low-density lipoprotein receptors, which are expressed in a wide variety of cell types, and the uptake of the siRNA. Alternatively, the passenger strand of the siRNA can be incorporated into the termini of an aptamer, a structured RNA molecule that has been selected *in vitro* for specific properties, such as high-affinity binding to a distinct molecule. Using this approach, one can achieve cell type-specific delivery by choosing an aptamer that has a high affinity for a particular cell surface molecule. Similarly, cell type-specific delivery has been achieved using a recombinant protein containing the recognition domain of a single chain antibody or a fragment antibody fused to a nucleic acid binding motif (e.g., protamine) (iii). In this case, the siRNA interacts with the nucleic acid binding motif and is directed to the cell type of interest by the affinity of the antibody for a specific cell surface molecule. Although the majority of lipid-based transfection reagents have associated toxicities that prevent their use for *in vivo* applications, there are some lipid formulations that can be used for efficient delivery *in vivo* with negligible toxicity (iv). This usually involves the encapsulation of the siRNA into a lipid nanoparticle composed of several different types of lipids to provide it with the appropriate stability while allowing for the uptake and release of the siRNA into the cytoplasm of target cells. These lipid particles can be made to target specific cell types by the incorporation of specific moieties (e.g., antibody fragments) on their surface. (b) Stable silencing using DNA-based siRNA expression systems. Long-term silencing of gene expression can be achieved by expressing siRNA precursors, either (i) short hairpin (sh)RNAs or (ii) long hairpin RNAs based on the structure of endogenous miRNAs. These hairpin RNAs are processed by the endogenous RNAi machinery producing active siRNAs. These siRNA expression cassettes have been incorporated into a variety of DNA-based expression systems, including viral systems such as retroviral (oncoretroviral or lentiviral) and adeno-associated viral vectors. Typically, shRNAs have been expressed from RNA polymerase III (pol III) promoters that are uniquely suited for the production of short RNA species, whereas longer hairpin RNAs can be expressed from either Pol III or RNA polymerase II (pol II) promoters. Unlike pol III promoters that are ubiquitously expressed, pol II promoter-driven siRNA expression systems can be developed that permit the tissue-specific or inducible expression of the siRNA.

siRNA needed to effectively silence expression of the target gene in the intended tissue or cell type. Therefore, delivery vehicles that could deliver siRNAs to specific cell types would be beneficial, particularly for therapeutic siRNA-based gene-silencing technologies. A number of technologies have been developed for the cell type-specific delivery of siRNAs. One such approach uses a recombinant bifunctional protein that contains a peptide that can recognize a specific cell-surface antigen (e.g., a single chain antibody or a fragment antibody against a cellular receptor or a ligand for a specific receptor) fused to a nucleic acid binding domain. Unlike the cholesterol-conjugated siRNAs, the siRNAs bind noncovalently to the recombinant protein and can be delivered to the cytoplasm of only those cells that express the specific protein on the cell surface. Therefore, the same deliver vehicle can be used to deliver any siRNA of interest for delivery. Cell type-specific delivery has also been achieved by conjugating the siRNA to an aptamer. Aptamers are structured nucleic acids that have been selected *in vitro* for specific characteristics, such as the ability to bind to specific proteins or to have a specific catalytic activity. The conjugation of an RNA aptamer that recognizes a specific cell surface protein to an siRNA can effectively deliver the siRNA to the cell type or tissue of interest. In order to retain the activity of the siRNA, the protein or RNA moiety is conjugated to the passenger strand of the siRNA, leaving the guide strand to perform its silencing function unimpaired. The mechanism by which the siRNAs are taken up and delivered to the cytoplasm to silence gene expression remains an area of active research. It remains to be determined how broadly applicable these technologies will be for the delivery of siRNAs. Despite these unknowns, these approaches have taken the therapeutic application of siRNAs one step closer to becoming a reality.

Vector-Mediated RNA Interference

To overcome some of the intrinsic limitations of siRNA transfections, the transient nature of the silencing phenotype in actively dividing cells, and the difficulty of efficiently transducing certain cell types that are refractory to traditional lipid-based transfection protocols, siRNA-expressing DNA vectors have been developed (Figure 2(b)). Although Dicer was originally identified as an endonuclease that is capable of cleaving long dsRNA into siRNAs, it was soon realized that Dicer was also capable of processing the hairpin precursor molecule of miRNAs. By mimicking the structure of endogenously produced miRNA precursors, short hairpin RNAs were designed that could be efficiently expressed from DNA vectors and processed into active siRNAs. The majority of these expression systems were based on expression from RNA polymerase III (pol III) promoters, either the U6 or the H1 promoter, which are uniquely suited for the expression of small RNA species since all of the necessary sites for activation of transcription are present upstream of the transcription start site and a run of uridine residues is sufficient to terminate transcription. The initial attempts to develop RNA pol II-based siRNA expression systems were unsuccessful. Although the transcripts were expressed, they were not processed by Dicer into functional siRNAs. It was only after the mapping of several miRNA transcripts that it was demonstrated that miRNA was encoded on primary transcripts that were highly structured hairpin RNAs containing frequent bulges and mismatched base-paired interactions. Based on these longer hairpin RNA structures, successful expression and processing of siRNAs was achieved. Unlike the ubiquitously expressing pol III promoters, pol II promoters that allow inducible, tissue- and cell type-specific transcription can be used for the regulated expression of siRNAs. In all these cases, the promoter drives transcription of a hairpin RNA containing the siRNA sequence and the accompanying passenger strand separated by a short noncomplementary loop sequence. These hairpin RNAs are processed by the endogenous RNAi machinery producing active siRNAs that enters RISC and direct the silencing of gene expression. The siRNA expression cassettes have been incorporated into an assortment of plasmid and viral vectors for the transduction of cells. Unlike synthetic siRNAs, plasmid vectors can be used for the transient delivery of siRNAs and, if the vector contains a suitable selectable marker, the stable delivery of siRNAs into many cell lines. However, their use is limited, in the same way as synthetic siRNAs, to cells that can be readily transfected. In addition, the low efficiency of generating stable silencing clones further limits their use in primary cells to transient siRNA expression. On the other hand, recombinant viral vectors have been developed which incorporate the siRNA expression cassette into the viral genome that is packaged into viral particles allowing for the infection of susceptible cell types. A variety of viral vectors have been adapted for use as siRNA delivery vehicles, including retroviral- and adeno-associated viral-based vectors. Since these viruses possess the necessary factors to facilitate uptake into cells, they are not limited to cells that can be readily transfected. Each of these systems has unique cell tropisms and life cycles, allowing them to productively deliver the siRNA expression cassette to specific subsets of cells. However, by changing the characteristic of the viral surface proteins involved in cellular recognition and uptake, it is possible to broaden the variety of cell types that can be infected. For example, HIV-1-based lentiviral constructs that utilize the HIV-1 viral

envelope glycoprotein will infect only human cells that express the CD4 molecule on their surface. However, if the viral envelope glycoprotein is replaced by the glycoprotein from the vesicular stomatitis virus, which has a broad cell tropism, the lentiviral construct can be used to infect a variety of cell types from many different organisms, including human, mouse, rat, and even avian cells. Retroviral-based systems, derived from either lentiviruses or oncoretroviruses, have been used extensively for siRNA delivery. In these recombinant retroviral vectors, the genes necessary for productive infection have been deleted to ensure that the virus is unable to replicate in the target cells. Retroviruses are particularly well suited for the stable, long-term expression of siRNAs since they become integrated into the host genome and are replicated along with the host DNA. In addition, vectors based on lentiviral genomes, a class of retroviruses that include HIV-1, feline immunodeficiency virus, and equine infectious anemia virus, can infect postmitotic, nondividing cell types (e.g., terminal differentiated macrophage cells and neuron). Despite these advantages, retroviral vectors have the potential to cause insertional mutagenesis, leading to oncogenic transformation and clonal expansion of affected cells. Since these effects were seen in cells that were already actively replicating, it may be that this insertional mutagenesis will not be a significant barrier to the use of retroviral vectors in quiescent cells, such as those found in most tissues including the brain. Adeno-associated viral vectors have also been used successfully for *in vivo* siRNA delivery. Unlike retroviral vectors, adeno-associated vectors do not integrate into the host genome but can readily mediate long-term siRNA expression, particularly in nondividing tissues. In addition, adeno-associated viral vectors lack expression of viral proteins and as a result induce a limited immune response. Although these viral-based siRNA delivery systems have been shown to be effective both *in vitro* and *in vivo*, the packaging and purification of the recombinant viral particles can be a laborious process. For *in vivo* applications, siRNA expression from DNA-based vector systems is usually limited to the injection site and delivery requires a local regional administration (injection) or surgical procedure.

Applications of RNAi-Based Technologies

RNAi-mediated gene silencing has been widely applied as a research tool for reverse genetic experiments, which determine the function of a gene based on its disruption. The majority of these experiments have been done in a focused manner, examining the function of a single gene of interest at a time as well as other genomes, coupled with the ubiquitous nature of the RNAi machinery, this has made it possible to silence any gene in the genomes of humans and other organisms. These approaches have been extensively applied in *C. elegans* and *Drosophila* cells to identify unknown genes involved in a wide array of cellular functions. Although large-scale screens in mammalian cells have presented a greater technical challenge, the refinement of RNAi-based technologies (siRNA transfection protocols and shRNA-based expression systems) has made these screens increasingly popular.

It took very little time after RNAi was discovered in mammalian cells for researchers to recognize the potential of harnessing these pathways for therapeutic gene silencing. RNAi-based approaches have been extensively used for the silencing of pathogenic states ranging from viral infections to tumor development and metabolic and neurological disorders. The ease of design and synthesis, the low cost of production relative to other therapeutic agents such as antibodies and recombinant growth factors, the specificity of silencing, and the potency of silencing make siRNAs an attractive candidate for therapeutic applications. Delivery remains the major challenge for the successful therapeutic application of RNAi-based gene-silencing approaches. This is particularly the case for the systemic administration of siRNAs since unmodified siRNAs are sensitive to serum RNases and their small size results in their rapid clearance from the circulatory system. Despite these challenges, localized delivery of siRNAs has entered clinical trials for the treatment of age-related macular degeneration and respiratory syncytial virus infection. In a very short period of time, RNAi has gone from a novel silencing phenomenon in *C. elegans* to a highly conserved and ubiquitous gene regulatory network involved in many important cellular processes and, along the way, has been harnessed, providing an indispensable research tool and potential novel therapeutic modality.

See also: Axonal mRNA Transport and Functions; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; RNA Binding Protein Methods; Rodent Behavior: Approaches.

Further Reading

Alisky JM and Davidson BL (2004) Towards therapy using RNA interference. *American Journal of Pharmacogenomics* 4: 45–51.

Aravin A, Gaidatzis D, Pfeffer S, et al. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442: 203–207.

Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* 116: 281–297.

Boudreau RL and Davidson BL (2006) RNAi therapy for neurodegenerative diseases. *Current Topics in Developmental Biology* 75: 73–92.

Chu TC, Twu KY, Ellington AD, et al. (2006) Aptamer mediated siRNA delivery. *Nucleic Acids Research* 34: e73.

Davidson BL and Harper SQ (2005) Viral delivery of recombinant short hairpin RNAs. *Methods in Enzymology* 392: 145–173.

Dorsett Y and Tuschl T (2004) siRNAs: Applications in functional genomics and potential as therapeutics. *Nature Reviews Drug Discovery* 3: 318–329.

Dykxhoorn DM and Lieberman J (2005) The silent revolution: RNA interference as basic biology, research tool, and therapeutic. *Annual Review of Medicine* 56: 401–423.

Dykxhoorn DM, Palliser D, and Lieberman J (2006) The silent treatment: siRNAs as small molecule drugs. *Gene Therapy* 13: 541–552.

Elbashir SM, Harborth J, Lendeckel W, et al. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* 411: 494–498.

Elbashir SM, Lendeckel W, and Tuschl T (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Development* 15: 188–200.

Fire A, Xu S, Montgomery MK, et al. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391: 806–811.

Girard A, Sachidanandam R, Hannon GJ, et al. (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442: 199–202.

Jackson AL and Linsley PS (2004) Noise amidst the silence: Off-target effects of siRNAs? *Trends in Genetics* 20: 521–524.

Judge AD, Sood V, Shaw JR, et al. (2005) Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. *Nature Biotechnology* 23: 457–462.

Khvorova A, Reynolds A, and Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115: 209–216.

Kim DH and Rossi JJ (2007) Strategies for silencing human disease using RNA interference. *Nature Reviews Genetics* 8: 173–184.

Kim VN (2005) MicroRNA biogenesis: Coordinated cropping and dicing. *Nature Reviews Molecular Cell Biology* 6: 376–385.

Liu J, Carmell MA, Rivas FV, et al. (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science* 305: 1437–1441.

Lu C, Tej SS, Luo S, et al. (2005) Elucidation of the small RNA component of the transcriptome. *Science* 309: 1567–1569.

McNamara JO 2nd, Andrechek ER, Wang Y, et al. (2006) Cell type-specific delivery of siRNAs with aptamer-siRNA chimeras. *Nature Biotechnology* 24: 1005–1015.

Morrissey DV, Lockridge JA, Shaw L, et al. (2005) Potent and persistent *in vivo* anti-HBV activity of chemically modified siRNAs. *Nature Biotechnology* 23: 1002–1007.

Pasquinelli AE, Hunter S, and Bracht J (2005) MicroRNAs: A developing story. *Current Opinion in Genetics & Development* 15: 200–205.

Reynolds A, Leake D, Boese Q, et al. (2004) Rational siRNA design for RNA interference. *Nature Biotechnology* 22: 326–330.

Schwarz DS, Hutvagner G, Du T, et al. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 115: 199–208.

Song E, Zhu P, Lee SK, et al. (2005) Antibody mediated *in vivo* delivery of small interfering RNAs via cell-surface receptors. *Nature Biotechnology* 23: 709–717.

Soutschek J, Akinc A, Bramlage B, et al. (2004) Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. *Nature* 432: 173–178.

Stewart SA, Dykxhoorn DM, Palliser D, et al. (2003) Lentivirus-delivered stable gene silencing by RNAi in primary cells. *RNA* 9: 493–501.

Tomari Y and Zamore PD (2005) MicroRNA biogenesis: Drosha can't cut it without a partner. *Current Biology* 15: R61–R64.

Zamore PD, Tuschl T, Sharp PA, et al. (2000) RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* 101: 25–33.

Zaratiegui M, Irvine DV, and Martienssen RA (2007) Noncoding RNAs and gene silencing. *Cell* 128: 763–776.

Sleep and Circadian Rhythm Disorders in Human Aging and Dementia

M V Vitiello, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

General Introduction

In 2007, approximately 12% of the United States population consisted of people 65 years of age or older. By 2030, this proportion will rise to 20%. Because the older portion of the population is increasing twice as fast as other age groups, the number of people 65 year of age or older in the United States will effectively double, to 72 million, by 2030.

Alzheimer's disease (AD) is a progressive neurodegenerative disorder, common in older adults, that accounts for approximately two-thirds of all dementias worldwide. In the United States, it has recently been estimated that between 2 and 4 million older adults have AD. Further estimates suggest that this number is likely to quadruple over the next 50 years.

Among older adults, one of the major changes that commonly accompany the aging process is an often profound disruption of an individual's daily sleep–wake cycle. This disruption is often further exacerbated in older individuals suffering from neurodegenerative disorders, such as AD. Here we review what is known about sleep and circadian rhythms in normal aging and in AD.

Normal Aging

As many as 50% of older individuals complain about sleep problems, including disturbed or light sleep, frequent awakenings, early morning awakenings, and undesired daytime sleepiness. Such disturbances are often associated with impaired daytime function and compromised quality of life. The most striking change in the sleep in older adults is the frequent interruption of nighttime sleep by periods of wakefulness, possibly the result of an age-dependent intrinsic lightening of sleep homeostatic processes. Further, older adults are more easily aroused from nighttime sleep by auditory stimuli, suggesting that their sleep may be more easily disrupted by exogenous stimuli. Both of these changes are indicative of impaired sleep maintenance and depth and contribute to the characterization of the sleep of older adults as lighter, or more fragile, than that of younger adults.

These age-associated increases of nighttime wakefulness are mirrored by increases in daytime fatigue, excessive daytime sleepiness, and increased likelihood of napping or falling asleep during the day. Aging is also associated with a tendency to both fall asleep and awaken earlier. That is, a tendency for older individuals to be larks rather than owls. Older individuals also tend to be less tolerant of phase shifts in time of the sleep–wake schedule such as those produced by jet lag and shift work. These changes suggest an age-related breakdown of the normal adult circadian sleep/wake rhythm.

Polysomnographic (PSG) studies of sleep architecture have consistently demonstrated four age-related changes in nighttime sleep quality: reductions in total sleep time, reductions in sleep efficiency, reductions in slow wave sleep, and increases in waking after sleep onset. Nevertheless, age-related changes in numerous other PSG sleep measures remain unclear. Sleep onset latency, stages 1 and 2, and rapid-eye movement (REM) have all been reported to change with age, whereas other studies have observed no such age-related changes.

It is important to note that even carefully screened older adults who do not complain of sleep disturbance and who have minimal medical burdens show the changes described when compared to younger adults. This suggests that at least some of the sleep disturbance seen in older adults is part of the aging process *per se*, apparently independent of any medical or psychiatric illnesses or primary sleep disorders and often referred to as age-related sleep change. As to whether this age-related decline in the ability to generate sleep equates with a decreased need for sleep in the later years of the human life span remains unclear. Nevertheless, the available scientific evidence suggests that it is important to remember that as individuals age, they make need to modify their expectations about the duration and quality of their nighttime sleep.

Sleep in Normal Aging

Clearly sleep changes with advancing age, but an important question remains: exactly when during the adult life span do these changes occur? It has typically been assumed that the age-related sleep changes that characterize the sleep of older adults begin to appear in early adulthood and progress steadily across the full continuum of adult human life span, including the older adult years, so that, for example, the sleep of the typical 75-year-old was assumed to be worse than that of the typical 65-year-old.

However, the results of a recent and comprehensive meta-analysis of objective sleep measures across the

human life span demonstrated that the bulk of the age-related changes seen in adult sleep patterns do not occur continuously across the human adult life span but, rather, occur between early adulthood, beginning at age 19, and that after age 60 changes in sleep macroarchitecture becomes effectively asymptote, declining only minimally from age 60 to age 102. These sleep changes across adult life-span are summarized in Table 1. When the full adult life span is examined (2391 adults, ages 19–102 years), such analyses confirmed the four most consistently reported age-related changes in PSG studies of sleep macroarchitecture: decreases in total sleep time (TST), decreases in sleep efficiency (SE), decreases in slow wave sleep (SWS), and increases in waking after sleep onset (WASO). Also confirmed were less consistently reported age-related sleep changes; increases in stage 1 (S1) and stage 2 (S2) sleep and decreases in REM sleep were all observed in their meta-analyses. Conversely, age-related changes in both sleep latency (SLAT) and REM latency (REMLAT) were minimal.

It is crucial to note that all these significant age changes in objectively assessed sleep architecture were found only when the full adult life span of 19–102 years was examined. When the sleep of only older adults (1142 adults, ages 60–102 years) was examined, the meta-analyses demonstrated that no sleep measures showed any significant age-related change within the older adult portion of the study sample, with the single exception of SE, which declined significantly but at a modest rate of ~3% per decade.

Table 1 Summary of significant findings from meta-analyses examining the association between various sleep measures and age

	Adults 19–102 years[a]	*Older adults 60–102 years*[b]
Total sleep time	⇓	⇔
Sleep latency	⇔	⇔
WASO	⇑	⇔
Sleep efficiency	⇓	⇓
Stage 1 (%)	⇑	⇔
Stage 2 (%)	⇑	⇔
SWS (%)	⇓	⇔
REM (%)	⇓	⇔
REM latency	⇔	⇔

[a] $n = 2391$ adults.
[b] $n = 1142$ adults.
From Ohayon MM, Carskadon M, Guilleminault C, and Vitiello MV (2004) Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan. *Sleep* 27(7): 1255–1273.
⇓, decreased; ⇑, increased; ⇔, unchanged; REM, rapid-eye movement sleep; SWS, slow wave sleep; WASO, waking after sleep onset.

These findings were comparable for both older men and older women.

These findings seem counterintuitive and fly against our commonly held concepts of sleep changes with aging. However, it is important to remember that Ohayon and colleagues used very rigorous selection criteria, such that the subjects involved in the studies analyzed were not representative of the older population *per se* but, rather, were in excellent health and more likely represented individuals who were optimally or successfully aging. That is, these findings represent normative age-related changes in sleep architecture and not average sleep changes of the entire aging population. When various medical and psychiatric comorbidities that typically accompany the aging process were controlled for and optimal aging was examined, the bulk of age-related sleep changes occurred in early and middle adulthood (ages 19–60 years); after age 60, assuming the individual remained in good health, further age-related sleep changes were likely to be, at most, modest. Conversely, if such comorbidities were present, these minimal age-related sleep changes might well be exacerbated, consistent with the level of sleep-related complaints seen in the general aging population.

Circadian Rhythms in Normal Aging

Not only does the quality of sleep change across the human life span, but the timing of when that sleep occurs also apparently changes with aging. Circadian rhythms are those that occur within a period of 24 h (from the Greek about (*circa*) a day (*dies*)), such as the adult human sleep–wake cycle. Interestingly, as with sleep, there is a considerable disparity in the conventional wisdom concerning circadian rhythms and aging and which evidence is supports or does not support these conventionally held beliefs. The conventional wisdom regarding what happens to human circadian processes with aging holds that (1) the circadian amplitude is reduced; (2) the circadian phase is advanced (i.e., the circadian rhythm moves earlier relative to the environment); (3) the circadian free-running period (tau) is shortened; and (4) the ability to tolerate rapid phase shifts, such as those experienced in shift work or rapid transmeridian travel (jet lag), declines.

However, the available evidence convincingly supports only two of these beliefs: that older people tend to have earlier circadian phases, with a corresponding tendency to go to bed and arise from bed earlier than younger adults; and that they have more trouble than younger adults adjusting to the rapid phase shifts due to shift work and jet lag, at least in terms of sleep quality, subjective complaint, and performance measures. Interestingly, the data in support of

diminished circadian amplitudes and shortened circadian taus in healthy older adults are, at best, equivocal. Again, it is important to note that these circadian changes are those seen in normal, or more accurately normative, aging, in which age-related health comorbidities have been carefully screened out.

Excessive Daytime Sleepiness and Napping in Normal Aging

Two other commonly held assumptions about sleep and aging are that older adults typically report more excessive daytime sleepiness (EDS) and nap more than younger adults. Few community-based epidemiological studies have reported the prevalence of regular napping and its association with sleep complaints and other mental and physical health problems, especially in relation EDS. Somewhat counterintuitively, although epidemiological studies typically show a significant increase in the prevalence of regular napping with advancing age, EDS does not demonstrate a similar increase in prevalence among older adults. Consistent with the findings already described for nighttime sleep measures, very recent findings indicate that the presence of comorbidities (medical illness, depression, etc.) is highly associated with the likelihood of an older adult reporting regular napping or EDS. That is, healthy older adults, even those complaining of significant nighttime sleep disturbance, are much less likely to report regular napping or EDS than their more health-burdened cohorts.

There is considerable debate as to whether regular napping among older adults, particularly those in good health, may be beneficial to daytime wakefulness or perhaps detrimental to their nighttime sleep propensity. Studies of naps in healthy noncomplaining older adults demonstrated that napping had only a mild to moderate impact on nighttime sleep quality and may even result in improved postnap cognitive performance. However, these results should be interpreted with caution because, again, it is important to emphasize that the subjects involved were carefully screened healthy older adults. It is unclear if similar results would be obtained, for example, with a sample of older insomniacs.

Causes of Sleep and Circadian Disturbances in Older Adults

It is often repeated that epidemiological studies report that as much as 50% of older adults complain of significant, chronic sleep disturbance. However, it is important to keep in mind that the other 50% of older adults do not so complain. As previously reviewed, the sleep of very healthy older adults changes only very slowly across the later human life span. Conversely it has also been shown that older adults who do not complain of any sleep problems nevertheless have objective sleep quality that is markedly compromised (e.g., less TST, less SE, less SWS, and more WASO) compared to that of healthy, non-sleep-complaining, younger adults. Nevertheless, there are many factors in addition to the age-related sleep and circadian changes already discussed that contribute to the significant sleep disturbance reported by half of all older adults. These include (1) medical and psychiatric comorbidities and their treatments, such as cardiovascular disease, arthritis, depression, or AD and many of the drugs used to treat them; (2) primary sleep disorders, many of which, such as sleep apnea, restless legs syndrome, and REM behavior disorder, tend to occur with increasing frequency in older adults; (3) the many behavioral, environmental, and social factors, often collectively referred to as sleep hygiene, that have the potential to maximize or compromise an individual's sleep quality; and (4) some combination of these factors. A more comprehensive list of such contributing factors with specific examples appears in Table 2.

Table 2 Causes of poor sleep in older adults

Cause/problem	*Examples*
Physiological	Age-related sleep change
	Age-related circadian rhythm change (phase advance)
Medical illnesses	Arthritis or other conditions causing chronic or intermittent pain
	Chronic cardiac or pulmonary disease
	Gastro-esophageal reflux disorder
Psychiatric illnesses	Depression
	Alzheimer's disease
Medications	Diuretics (leading to nocturnal awakenings)
	Inappropriate use of OTC medications
Primary sleep disorders	Sleep disordered breathing (sleep apnea)
	Restless legs syndrome
	REM behavior disorder
Behavioral/social	Retirement/lifestyle change reducing need for regular bed and rise times
	Death of a family member or friend
	Inappropriate use of social drugs
	Transmeridian air travel
	Napping
Environmental	Bedroom environment (e.g., ambient noise, temperature, light, bedding)
	Moving to a new home or downsizing to a smaller space or a retirement community or related facility
	Institutionalization

OTC, over-the-counter; REM, rapid-eye movement.
Reproduced from Vitiello MV (2006) Sleep in normal aging. *Sleep Medicine Clinics* 1(2): 171–176, with permission from Elsevier.

Alzheimer's Dementia

Both normal aging and AD are associated with disturbances in the daily sleep–wake cycle, although the disturbances in AD are typically much more severe. Within the AD population itself, significant sleep disturbances are quite common, affecting as much as half of community- and clinic-based samples.

For AD patients, sleep disturbance adds an additional burden to the compromised function and quality of life directly attributable to dementia. For AD caregivers, disturbances in patients' sleep and nighttime behavior, particularly the reduced nighttime sleep time, increased nighttime wakefulness, and wandering that commonly require considerable caregiver attention, with a subsequent chronic sleep loss for the caregivers themselves, are a significant source of physical and psychological burden and are often cited as one of the principal reasons for a family's decision to institutionalize a demented person.

The Biological Bases of Sleep Disturbances in AD

The sleep disturbances that accompany early-stage or mild dementia are remarkable in that they appear as exacerbations of the sleep changes found in normal aging rather than as unique disease-related phenomena. The sleep of AD patients is marked by an increased duration and frequency of awakenings, decreased SWS and REM sleep, and more daytime napping. Damage to neuronal pathways that initiate and maintain sleep is the most likely cause of the acceleration of these age-related sleep changes in AD patients. These compromised neural structures, including the suprachiasmatic nucleus (SCN) of the hypothalamus and neuronal pathways originating in subcortical regions, also regulate arousal and the sleep–wake cycle. Recent work suggested that at least some of the sleep disruption seen in AD many be an inherent trait, probably linked to apolipoprotein E (APO E) status, such that AD patients positive for the APO E $\varepsilon 4$ allele show greater sleep disruption over time.

There is considerable evidence that sleep disturbance grows more severe with increasing severity of AD. However, the moderate, or intermediate, stages of the disease are typically when most other behavioral disturbances occur with peak frequency. The breakdown of the basic circadian sleep–wake rhythm of dementia patients can be severe and in extreme cases may lead to an extremely fragmented sleep–wake pattern or nearly complete day–night sleep pattern reversals. In end-stage AD, patients may appear to sleep throughout most of the day and night, awakening only for brief periods. However, as of this time there have been no prospective longitudinal studies of sleep in AD that would inform an accurate understanding of both the course and the individual differences of these sleep disturbances. This remains an important gap in both the biology and therapeutics of sleep disturbances in demented patients.

The current standard of care for treating cognitive disorders in AD is the use of oral acetylcholinesterase inhibitors. These drugs may improve sleep patterns in some AD patients, particularly by increasing REM sleep. Because they act by enhancing cholinergic transmission in the brain, the involvement of both forebrain and brain stem cholinergic nuclei in regulating sleep–wake cycles and arousal forms the rationale for expecting both positive and negative sleep effects with use of these agents.

Circadian Rhythm Disorder in AD

AD patients commonly manifest a breakdown of their circadian sleep–wake rhythm. Termed irregular sleep–wake rhythm disorder (ISWRD), this perturbation of circadian rhythmicity is commonly associated with neurological impairment in older adults, such as the deficits experienced by AD patients. ISWRD is characterized by the relative absence of a circadian pattern in the sleep–wake cycle. Total sleep time can well be comparatively normal, but, instead of being consolidated into a single long relatively continuous major sleep period, sleep times are shortened and, in extreme cases, relatively brief sleep periods are distributed almost randomly throughout the day and night. The cause, or more likely the causes, of this association are unknown, but AD-related damage to the SCN, locus of the circadian pacemaker, is clearly implicated as an important, if not a major, etiological factor. Experimental ablation of the SCN in animals produces a loss of circadian rhythmicity that strongly resembles the sleep–wake pattern typically seen in older adults with dementing disorders, particularly in the later stages of the dementia. However, it is important to note that clinical studies, the bulk of which have been carried out in older adults with dementia (particularly AD), have rarely used formal sleep diagnostic criteria for ISWRD.

Although complaints of nighttime sleep fragmentation and daytime napping have been consistently reported to increase with age, the weight of the available evidence strongly indicates that, although the prevalence of ISWRD increases with age, this increase is secondary to the increased prevalence of associated medical disorders with advancing age. Age is apparently not an independent risk factor for ISWRD; rather, it is probably the age-related medical comorbidities which constitute the main risk factors

for sleep pathology, including both increasing disturbances in sleep architecture and the increase in prevalence of ISWR with advancing age. ISWRD is particularly associated with neurological impairment in older adults, most significantly AD.

Although there is no direct evidence for a genetic basis for ISWRD, there are several lines of evidence that suggest that the sleep disturbance seen in AD is at least partially based on genetic factors. Actigraphic studies of AD patients have demonstrated longitudinal deterioration of sleep quality, and most of this longitudinal variance in sleep appears to be related to an inherent trait of the individual patient. This suggests that genetic factors may help determine the ultimate course and level of sleep deterioration seen in a given AD patient, a hypothesis consistent with research indicating that much circadian variation in physiological systems is controlled by a limited number of similar genes across species. Regardless, the tentative relationship of genetic factors as a risk for ISWRD remains to be fully explored.

Older adults are exposed to reduced light levels in their daily lives relative to younger individuals. This reduction may be exacerbated by disorders of vision, which are common in older adults and which many further attenuate the impact of ambient light on the SCN. Finally, the SCN itself is adversely impacted by age. The impact of each of these factors is magnified in patients with AD, including those who are community dwelling, who have been shown to be exposed to less light than age-matched healthy normal controls. Furthermore, the retina and the optic nerve can both be compromised in AD. Older adults, especially those who are institutionalized, are exposed to less intense light than younger people, but questions remains as to the importance of reduced light exposure as an etiological factor in ISWRD.

Other Causes of Sleep and Circadian Disturbances in AD

Population-based studies examining the causes, incidence, and persistence of sleep disturbances in AD patients are lacking; consequently, little is known about the risk factors for their development. Therefore, we must look to the literature concerning sleep disturbance in the nondemented elderly and factor this with clinical assessments of individual patients to make informed inferences about the AD population. Certainly the same neurodegenerative mechanisms that result in the evolving cognitive deficits seen in AD are a potential primary cause of the sleep disruption seen in this population. However, it is important to recognize that sleep can be disrupted for many other reasons as well. The likely major causes of sleep disruption in AD, as with normal aging, include (1) physiological changes that arise as part of normal nonpathological aging; (2) sleep problems due to comorbid physical or mental health conditions, including AD, and possibly their treatments; (3) primary sleep disorders; (4) poor sleep hygiene or sleep-related habits; or (5) commonly, some combination of these factors.

Studies of patients living in institutional settings provide much of the available information about sleep and circadian disturbance in cognitively impaired patients. Environmental factors that promote circadian dysregulation in dementia patients living in such settings include light, noise, activity schedules, and the needs of staff. Ambient light levels are typically too low in many congregate care facilities to support natural light-dependent internal rhythms, and noisy conditions, especially during the night, are both common and inimical to sleep. The staffing schedules and timing of specific activities in many facilities that care for demented people may be driven less by the needs of the patients than by compliance with federal and state requirements governing nursing-home operations. Regulatory requirements in general fail to incorporate many of the positive evidence-based practices found to be beneficial for demented residents, focusing more attention on feeding and bathing schedules, injury prevention, and detection of medical problems than on sleep and other issues related to quality of life.

Conclusion

Previously it was assumed that sleep and circadian rhythms decline inexorably with advancing age. However, recent research shows that, at least for health older adults, most of the age changes in sleep that have typically been thought to progress steadily across the adult human life span actually occur before age 60 and that adults over age 60 who remain healthy can expect their sleep quality to remain relatively stable as they age. Similarly age-related changes in circadian changes, again in older adults free of comorbid illnesses, are not as marked as was previously assumed. However, even healthy older adults tend to have earlier circadian phases, with a corresponding tendency to go to bed and arise from bed earlier than younger adults and to experience more trouble than younger adults adjusting to the rapid phase shifts such as those experienced in shift work and jet lag, at least in terms of sleep quality, subjective complaint, and performance measures.

Although sleep quality and circadian rhythms are apparently relatively well preserved in healthy older

adults, the negative impact of age-related comorbid illnesses on sleep and circadian rhythms is also clear. AD is a classic example of this negative impact, with AD patients typically demonstrating progressively fragmented sleep and irregular circadian rhythms with increasing severity of the disease.

Nevertheless, there is a growing body of evidence that the sleep quality and circadian rhythmicity of AD patients can, like those of other less-than-healthy older adults, be improved through nonpharmacological treatment approaches, whether the AD patients are community dwelling or institutionalized. Such groundbreaking work offers a firm foundation for future efforts to improve the sleep and related circadian rhythms of older adults, regardless of their health status.

See also: Aging of the Brain and Alzheimer's Disease; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: Transgenic Mouse Models; Animal Models of Alzheimer's Disease; Brain Glucose Metabolism: Age, Alzheimer's Disease and ApoE Allele Effects; Genetics of Circadian Disorders in Humans; Psychiatric Disorders Associated with Disturbed Sleep and Circadian Rhythms; Shift Work and Circadian Rhythms.

Further Reading

Alessi CA, Martin JL, Webber AP, Cynthia Kim E, Harker JO, and Josephson KR (2005) Randomized, controlled trial of a nonpharmacological intervention to improve abnormal sleep/wake patterns in nursing home residents. *Journal of the American Geriatrics Society* 53: 803–810.

Alessi CA and Schnelle JF (2000) Approches to sleep disorders in the nursing home setting. *Sleep Medicine Reviews* 4(1): 45–56.

Ancoli-Israel S, Gehrman P, Martin JL, et al. (2003) Increased light exposure consolidates sleep and strengthens circadian rhythms in severe Alzheimer's disease patients. *Behavioral Sleep Medicine* 1: 22.

Bonnefond A, Harma M, Hakola T, Sallinen M, Kandolin I, and Virkkala J (2006) Interaction of age with shift-related sleep-wakefulness, sleepiness, performance, and social life. *Experimental Aging Research* 32(2): 185–208.

Campbell SS, Murphy PJ, and Stauble TN (2005) Effects of a nap on nighttime sleep and waking function in older subjects. *Journal of the American Geriatrics Society* 53: 48–53.

Cajochen C, Munch M, Knoblauch V, Blatter K, and Wirz-Justice A (2006) Age-related changes in the circadian and homeostatic regulation of human sleep. *Chronobiology International* 23(1–2): 461–474.

Dijk DJ, Duffy JF, and Czeisler CA (2001) Age-related increase in awakenings: Impaired consolidation of nonREM sleep at all circadian phases. *Sleep* 24(5): 565–577.

Dijk DJ, Duffy JF, Riel E, Shanahan TL, and Czeisler CA (1999) Ageing and the circadian and homeostatic regulation of human sleep during forced desynchrony of rest, melatonin and temperature rhythms. *Journal of Physiology* 516: 611–627.

Duffy JF, Zeitzer JM, Rimmer DW, Klerman EB, Dijk DJ, and Czeisler CA (2002) Peak of circadian melatonin rhythm occurs later within the sleep of older subjects. *American Journal of Physiology, Endocrinology and Metabolism* 282(2): E297–E303.

Foley DJ, Ancoli-Israel S, Britz P, and Walsh J (2004) Sleep disturbances and chronic disease in older adults: Results of the 2003 National Sleep Foundation Sleep in America Survey. *Journal of Psychosomatic Research* 56: 497–502.

Foley DJ, Monjan AA, Brown SL, Simonsick EM, Wallace RB, and Blazer DG (1995) Sleep complaints among older persons: An epidemiological study of three communities. *Sleep* 18: 425–432.

Foley DJ, Monjan AA, Simonsick EM, Wallace RB, and Blazer DG (1999) Incidence and remission of insomnia among elderly adults: An epidemiologic study of 6,800 persons over three years. *Sleep* 22(supplement 2): S366–S372.

Foley DJ, Vitiello MV, Bliwise DL, Ancoli-Israel S, Monjan AA, and Walsh JK (2007) Frequent napping is associated with excessive daytime sleepiness, depression, pain and nocturia in older adults: Findings from the national sleep foundation "2003 sleep in America" poll. *American Journal of Geriatric Psychiatry* 15: 344–350.

Hoffman MA and Swaab DF (2006) Living by the clock: The circadian pacemaker in older people. *Ageing Research Review* 5(1): 33–51.

Martin JL, Webber AP, Alam T, Harker JO, Josephson KR, and Alessi CA (2006) Daytime sleeping, sleep disturbance, and circadian rhythms in the nursing home. *American Journal of Geriatric Psychiatry* 4: 121–129.

McCurry SM, Gibbons LE, Logsdon RG, Vitiello MV, and Teri L (2005) Nighttime insomnia treatment and education for Alzheimer's disease: A randomized, controlled trial. *Journal of the American Geriatrics Society* 53: 793–802.

McCurry SM, Logsdon RG, Teri L, and Vitiello MV (2007) Evidence-based psychological treatments for older adult sleep disorders. *Psychology and Aging* 22(1): 18–27.

Monk TH (2005) Aging human circadian rhythms: Conventional wisdom may not always be right. *Journal of Biological Rhythm* 20(4): 366–374.

Monk TH, Buysse DJ, Carrier J, Billy BD, and Rose LR (2001) Effects of afternoon "siesta" naps on sleep, alertness, performance, and circadian rhythms in the elderly. *Sleep* 24(6): 680–687.

Ohayon MM, Carskadon MA, Guilleminault C, and Vitiello MV (2004) Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan. *Sleep* 27(7): 1255–1273.

Rybarczyk B, Stepanski E, Fogg L, Lopez M, Barry P, and Davis A (2005) A placebo-controlled test of cognitive behavioral therapy for comorbid insomnia in older adults. *Journal of Consulting Clinical Psychology* 73(6): 1164–1174.

Van Someren EJ (2000) Circadian and sleep disturbances in the elderly. *Experimental Gerontology* 35(9–10): 1229–1237.

Van Someren EJ, Riemersma RF, and Swaab DF (2002) Functional plasticity of the circadian timing system in old age: Light exposure. *Progress in Brain Research* 138: 205–231.

Vitiello MV (2000) Effective treatment of sleep disturbances in older adults. *Clinical Cornerstone* 2(5): 16–27.

Vitiello MV (2006) Sleep in normal aging. Sleep Medicine Clinics 1(2): 171–176.

Vitiello MV and Borson S (2001) Sleep disturbances in patients with Alzheimer's disease: Epidemiology, pathophysiology and management. *CNS Drugs* 15(10): 777–796.

Vitiello MV, Larsen LH, and Moe KE (2004) Age-related sleep change: Gender and estrogen effects on the subjective-objective

sleep quality relationships of healthy, non-complaining older men and women. *Journal of Psychosomatic Research* 56(5): 503–510.

Vitiello MV, Moe KE, and Prinz PN (2002) Sleep complaints co-segregate with illness older adults: Clinical research informed by and informing epidemiological studies of sleep. *Journal of Psychosomatic Research* 53: 555–559.

Vitiello MV, Prinz PN, Williams DE, Frommlet MS, and Ries RK (1990) Sleep disturbances in patients with mild-stage Alzheimer's disease. *Journal of Gerontology* 45(4): M131–M138.

Wu YH and Swaab DF (2005) The human pineal gland and melatonin in aging and Alzheimer's disease. *Journal of Pineal Research* 38: 145–152.

Yesavage JA, Friedman L, Kraemer H, et al. (2004) Sleep/wake disruption in Alzheimer's disease: APOE status and longitudinal course. *Journal of Geriatric Psychiatry and Neurology* 17(1): 20–24.

Yesavage JA, Taylor JL, Kraemer H, Noda A, Friedman L, and Tinklenberg JR (2002) Sleep/wake cycle disturbance in Alzheimer's disease: How much is due to an inherent trait? *International Psychogeriatrics* 14(1): 73–81.

Sleep and Sleep States: Cytokines and Neuromodulation

J M Krueger, L Churchill, and D M Rector,
Washington State University, Pullman, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sleep is of central importance to neurobiology, yet our understanding of the biochemical and neurophysiological mechanisms of sleep is rudimentary and sleep function has not been experimentally demonstrated. Our quality of life, performance, and mental well-being are all adversely affected by even a single night's loss of sleep. Chronic sleep loss is associated with pathologies such as the metabolic syndrome. Primary sleep disorders increase morbidity and mortality. Sleep plays a central role in many mental and physiological functions – for example, memory improves during sleep. It is thus important that we understand sleep regulatory mechanisms. This article focuses on the biochemical regulation of sleep and discusses how recent knowledge has led to new insights into how the brain is organized to produce sleep.

There are two somewhat independent literatures concerning the fundamental mechanisms of sleep regulation. One is based on neurophysiological methods, and this literature has led to the identification of circuits involved in non-rapid eye movement sleep (NREMS) regulation, such as corticothalamic projections and the hypothalamic ventrolateral preoptic and median preoptic circuits, and those involved in rapid eye movement sleep (REMS) regulation, such as the laterodorsal tegmental nucleus. Satisfactory explanations of how these circuits impose sleep on the brain and how they keep track of past sleep–wake activity likely will involve the biochemical mechanisms that interact with these circuits. Such interactions are particularly important for sleep homeostasis, the phenomenon exemplified by sleep rebound after prolonged wakefulness. The second sleep regulatory literature is based on biochemical methods. This work has its basis in the homeostatic nature of sleep and the nearly 100-year-old finding that has been replicated many times showing that the transfer of cerebrospinal fluid from sleep-deprived, but not control, animals enhances sleep in the recipients.

During the past 25 years, several sleep regulatory substances have been identified and extensively tested. However, only a handful are strongly implicated in sleep regulation. The list includes tumor necrosis factor-α (TNF-α), interleukin-1β, growth hormone releasing hormone (GHRH), prostaglandin D_2, and adenosine for NREMS regulation and vasoactive intestinal peptide, prolactin (PRL), and nitric oxide for REMS. When injected into animals, each of these substances promotes sleep; if they are inhibited, spontaneous sleep is reduced, their levels in brain vary with sleep propensity, and they act on sleep regulatory circuits to promote sleep. Here, we focus on TNF-α in NREMS regulation because it is the only substance identified whose circulating levels in humans correlate with sleepiness and sleep. We also discuss PRL because it has been implicated in REMS regulation by several laboratories and is a cytokine. In addition to these humoral agents, all neurotransmitters have been implicated in sleep–wake regulation; this literature is not discussed here. There is yet another literature demonstrating changes in gene expression, usually determined by changes in mRNA levels using gene chips, with sleep or sleep loss. That literature is distinguished from the sleep regulatory literature because every physiological function changes during sleep; thus, it is not possible from gene expression data alone to know if sleep *per se* is causative of the change. It is important to recognize that those humoral agents implicated in NREMS and REMS regulation affect each other's production and seem to act in concert with each other to affect sleep (Figure 1).

TNF-α in Sleep Regulation

The ability of TNF-α to promote NREMS was first described in 1987. TNF-α given directly into the brain or systemically enhances the duration of NREMS. For instance, mice receiving 3 μg TNF-α intraperitoneally spend approximately 90 min extra in NREMS during the first 9 h postinjection. In addition, NREMS after TNF-α treatment is associated with supranormal electroencephalogram (EEG) delta (0.5–4 Hz) waves, which are indicative of a greater intensity of NREMS. TNF-α is somnogenic in all species thus far tested: rabbits, mice, rats, and sheep. TNF-α has little effect on REMS if low NREMS-promoting doses are used; however, higher doses inhibit REMS. Sleep following TNF-α treatment appears to be normal in that sleep architecture remains normal, sleep remains easily reversible, postures remain normal, and animals remain responsive to handling. Changes in sleep-coupled autonomic functions, such as the decreases in brain temperature upon entry into NREMS, also persist after TNF-α treatment.

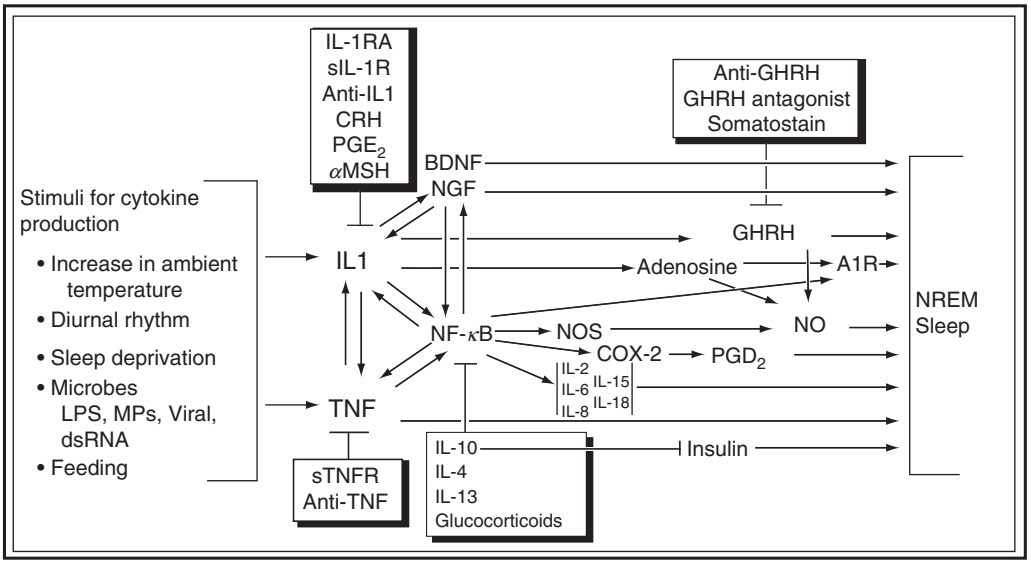

Figure 1 Molecular networks are involved in sleep regulation. Substances in boxes inhibit sleep and inhibit the production or actions of sleep-promoting substances illustrated via feedback mechanisms. Inhibition of one step does not completely block sleep since parallel sleep-promoting pathways exist. These redundant pathways provide stability to sleep regulation. Our knowledge of the biochemical events involved in sleep regulation is more extensive than that illustrated. The molecular network shown here possesses many of the characteristics of biological networks and engineered systems. Thus, the network is modular in that several proteins (cytokines) are working in 'overlapping coregulated groups' in this pathway. Second, the molecular network is robust in that removal of one of the components does not result in complete sleep loss. Third, the network operates as a recurring circuit element with multiple feedback loops affecting other pathways to the extent that similar networks involving many of the same substances and component network parts are used to regulate body temperature, inflammatory responses, microcirculation, memory, food intake, etc., and these systems, to a limited degree, coregulate. Specificity for any one physiological process, such as sleep, results from multiple interacting molecular and cellular circuits, each possessing different, but similar, reactivity. A1R, adenosine A1 receptor; αMSH, α-melanocyte-stimulating hormone; anti-IL1, anti-IL-1 antibodies; anti-TNF, anti-TNF antibodies; BDNF, brain-derived neurotropic factor; COX2, cyclooxygenase 2; CRH, corticotrophin-releasing hormone; GHRH, growth hormone releasing hormone; IGF1, insulin-like growth factor 1; IL-1RA, IL-1 receptor antagonist; LPS, lipopolysaccharide; MPs, muramylpeptides; NF-κB, nuclear factor-kappa B; NGF, nerve growth factor; NOS, nitric oxide synthase; PGD_2, prostaglandin D_2; sIL-1R, soluble IL-1 receptor; sTNFR, soluble TNF receptor; TGFβ, transforming growth factor-β.

Inhibition of TNF-α inhibits spontaneous NREMS. Thus, treatment with anti-TNF-α antibodies, the full-length soluble TNF receptor, or TNF soluble receptor fragments containing the TNF recognition site all inhibit spontaneous NREMS in rabbits and rats. Furthermore, pretreatment of animals with TNF inhibitors prior to sleep deprivation reduces the expected sleep rebound that normally occurs after sleep loss. Substances that inhibit TNF-α action or production also inhibit spontaneous sleep (e.g., interleukin-4, -10, and -13) (Figure 1). In addition, these substances also inhibit the production of certain other cytokines, such as interleukin-1β; therefore, their action on sleep may not be specific to TNF. However, they do form part of the negative feedback loops that help to regulate these nuclear factor-kappa B (NF-κB)-sensitive cytokines. Furthermore, inhibition of TNF also blocks the increases in NREMS observed in response to an acute mild increase in ambient temperature. Mice lacking the TNF 55-kDa receptor fail to exhibit NREMS responses if given TNF-α, thereby implicating this receptor in TNF-α-enhanced sleep. These mice also sleep less than corresponding control strains; the reduced NREMS occurs mostly during daylight hours.

Brain levels of TNF-α and the TNF-α mRNA vary diurnally and are influenced by sleep deprivation. The highest levels in rats occur at daybreak. The amplitude of the day–night changes in TNF protein is approximately tenfold and that in mRNA approximately twofold; this reflects the predominate posttranscriptional regulation of TNF. After sleep loss, hypothalamic TNF-α mRNA also increases. Sleep deprivation also increases the expression in brain of the 55 kDa TNF receptor mRNA. TNF-α serum levels increase in mice after sleep deprivation but not after stress. In normal humans, blood levels of TNF-α correlate with EEG delta wave activity. After sleep loss, circulating levels of TNF-α and the 55 kDa soluble TNF receptor, but not the 75 kDa TNF soluble receptor, increase.

Clinical conditions associated with sleepiness correlate with higher blood levels of TNF. Thus, patients with sleep apnea, chronic fatigue, chronic insomnia,

myocardial infarct, excessive daytime sleepiness, and preeclampsia all have elevated TNF plasma levels. AIDS patients have disrupted TNF and sleep rhythms. Postdialysis fatigue is associated with higher TNF levels, and cancer patients receiving TNF report fatigue. Furthermore, the 308A TNF-α polymorphic variant is associated with the metabolic syndrome and sleep apnea. Rheumatoid arthritis patients receiving the soluble TNF 75 kDa receptor report reduced fatigue. Sleep apnea patients treated with the soluble TNF receptor have reduced sleepiness. If obstructive sleep apnea patients are surgically treated, their elevated TNF-α plasma levels return to normal.

Systemic TNF, like several other cytokines, signals the brain via multiple mechanisms; one involves vagal afferents since vagotomy attenuates intraperitoneal TNF-α-induced NREMS responses. The effects of systemic bacterial products such as endotoxin, a component of gram-negative bacterial cell walls, may also involve TNF. For instance, in humans, endotoxin doses that induce transient increases in sleep also induce concomitant increases in circulating TNF-α. In addition, the soluble TNF receptor fragment attenuates bacterial cell wall peptidoglycan-enhanced NREMS in rabbits.

The site(s) of action of TNF-α-induced NREMS includes the preoptic area of the anterior hypothalamus, the locus coeruleus, and the somatosensory cortex. Thus, microinjection of TNF-α into the preoptic area enhances NREMS in rats. In contrast, injection of a soluble TNF receptor fragment into this site inhibits spontaneous NREMS. Microinjection of TNF-α into the locus coeruleus, after a brief period of excitation, induces prolonged increases in sleep and EEG synchronization. These effects are antagonized by anti-TNF-α antibodies. Microinfusion of TNF-α into the subarachnoid space just beneath the basal forebrain in rats enhances NREMS and reduces REMS. Finally, unilateral application of TNF-α onto the surface of the somatosensory cortex induces unilateral state-dependent increases in EEG delta wave power. Conversely, the soluble TNF receptor unilaterally reduces EEG power during the NREMS occurring after sleep deprivation. These latter two findings are germane to the brain organization of sleep and are discussed further later.

TNF Cell Biology

TNF signaling is an intense area of research due to its clinical importance and its inherent interest. TNF-α is synthesized as a 26 kDa membrane-associated protein. Soluble TNF-α, a 17 kDa protein, is cleaved from the 26 kDa membrane-associated protein by TNF-α converting enzyme. TNF-α production is tightly regulated in a tissue-specific manner, with transcription, translation, and secretion all controlled at multiple points. For instance, the 3′ UTR of TNF contains 24 miRNA complementary sites, suggesting translational regulatory sites. TNF-α induces its own expression, in part, via an NF-κB regulatory site (Figure 1). TNF-α can exist as a transmembrane protein that has biological activity. This membrane integrated form of TNF-α also receives signals and thus acts as a receptor, as well as a ligand.

There are two TNF cell surface receptors, the 55 kDa TNF receptor and the 75 kDa TNF receptor. The intracellular domains of these receptors lack intrinsic enzymatic activity. Rather, both receptors signal by recruitment of cytosolic proteins via protein–protein interaction domains. The diversity of these adaptor proteins and their ability to interact with other members of the TNF receptor superfamily help explain the pleiotropic actions of TNF-α. The TNF receptor family is divided into two large groups based on the adaptor proteins recruited in response to ligand binding. One major TNF-activated signaling pathway activates new gene transcription, whereas the other leads to cell death. Activation of the cell death pathway via the caspases 8 and 3 appears to be a rare physiological event because TNF-responsive gene products function to prevent cell death.

Both TNF receptors are cell surface receptors; they form trimeric complexes with TNF. The spatial distribution of the receptor and associated intracellular adaptor molecules is likely an important determinant of cellular specificity of action. Consequently, TNF activates signaling at multiple subcellular compartments, including the plasma membrane, mitochondria, and the nucleus. The ectodomains of both receptors are shed to form soluble receptors. The 55 kDa soluble TNF receptor is a constituent of normal cerebrospinal fluid, although its physiological role is unknown. However, given its effects on sleep and fatigue, it likely plays a role in sleep regulation.

TNF-α is expressed by microglia, astrocytes, and neurons and has a variety of biological actions in the brain, including a role in mediating both brain damage and neuroprotection. Whether TNF-α is protective or damaging may depend on the receptor type present – either the TNF 55 kDa or the TNF 75 kDa receptor – as well as the stimulus context and the presence or absence of substances that modify TNF-α activity. In addition to sleep, TNF-α has a role in thermoregulation, food intake, brain development, and neuronal connectivity. For instance, TNF-α potentiates AMPA-induced postsynaptic potentials,

AMPA-induced cytosolic Ca^{2+} increases, as well as several voltage-dependent calcium channels. TNF has a role in synaptic scaling, a process involved in maintaining synaptic diversity and plasticity.

Other NREMS Cytokine and Peptide Regulatory Substances

There is extensive evidence implicating both interleukin-1β (IL-1β) and GHRH in NREMS regulation that parallels that described previously for TNF-α. Thus, for example, inhibition of either GHRH or IL-1β inhibits spontaneous sleep. Mutant mice lacking functional GHRH receptors or mice lacking the IL-1β type I receptor have substantially less NREMS than respective strain controls. IL-1β induces growth hormone release via a hypothalamic mechanism that includes GHRH; anti-GHRH antibodies block IL-1β-induced growth hormone release. These antibodies also block IL-1β-induced NREMS. In fact, hypothalamic GABAergic neurons are receptive to both IL-1β and GHRH; both substances enhance intracellular calcium levels in these neurons. It is speculated that these neurons are the sleep-active hypothalamic neurons that increase their firing rates in response to IL-1β. In contrast, wake-active hypothalamic neurons decrease their firing rates in response to IL-1β. IL-1β and GHRH are related to TNF via the pathways outlined in Figure 1; collectively, it seems likely that these three substances form part of an NREMS regulatory network. Many other cytokines and members of the somatotropic axis can also affect sleep, but extensive knowledge concerning their roles in sleep regulation is lacking. They could in their own right be sleep regulatory substances or they may affect sleep via TNF-α, IL-1β, or GHRH.

Brain Organization of Sleep and TNF-α

Many investigations support the idea that sleep is regulated by specific brain circuits, such as the anterior hypothalamus. Perhaps the best evidence for this paradigm with regard to NREMS is that electrical, chemical, or thermal stimulation of the anterior hypothalamus induces NREMS. However, this paradigm inadequately addresses the issue of how the brain keeps track of its prior sleep–wake history over long periods of time. The transfer of cerebrospinal fluid from a sleep-deprived animal, but not a control animal, into a fully rested recipient enhances sleep in the recipient. This finding suggests that the homeostatic processes for sleep involve sleep regulatory substances.

There are also numerous results suggesting that sleep can be targeted to brain areas disproportionately activated during prior wakefulness. For instance, unilateral stimulation of the somatosensory cortex in awake humans produces asymmetry in EEG delta wave activity during subsequent NREMS. Similar results were obtained from rats and mice. Furthermore, dolphins only show EEG slow wave sleep on one side of the brain at a time. After a specific learning task, the subsequent EEG delta wave activity during sleep is enhanced in the area activated by the learning task. This finding reinforced previous work showing that either EEG activity or brain metabolism during sleep is regionally localized depending on prior use during wakefulness.

Our view of brain organization of sleep posits that sleep is initiated within neural assemblies as a function of prior cellular activity and this provides the mechanism for the localization and targeting of sleep, as well as for sleep homeostasis. In this view, for example, TNF-α is produced and released in response to neural and glial activity. These substances in turn are posited to be made and act locally in autocrine, paracrine, and juxtacrine ways to alter the input–output relationships of neural assemblies (e.g., cortical columns, barrels, or a cortical–thalamic–cortical loop) and thereby induce a functionally different state in these assemblies. Indeed, cortical TNF-α levels increase during prolonged wakefulness and during seizures – two conditions thought to be associated with enhanced cellular activity – and activity-dependent increases in TNF-α appear to be important for the developing nervous system. Furthermore, whisker stimulation upregulates TNF-α expression in neurons in the stimulated whisker's corresponding somatosensory barrel, but not in adjacent barrels, thereby directly demonstrating activity-dependent production of TNF-α.

Individual auditory and somatosensory cortical columns oscillate between functional states as defined by the amplitude of auditory- or whisker stimulation-induced surface evoked potentials. One of the functional states usually occurs simultaneously with whole animal sleep. This sleeplike state is dependent on its prior history in the sense that the probability of its occurrence is higher if, for the 15 min period prior to the measurement, it had been in the other functionally defined state. Furthermore, if two whiskers are stimulated, one at twice the rate of the other, the probability of the cortical column being in the sleep state is greater in the column receiving afferent input from the more rapidly stimulated whisker. If such changes in the functional state are dependent

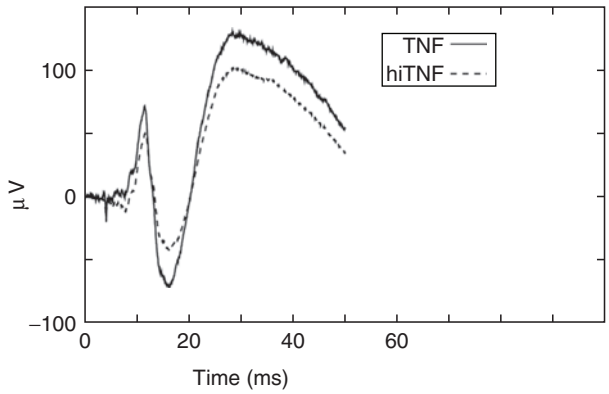

Figure 2 TNF-α enhances the amplitude of surface evoked potentials induced by whisker stimulation. A rat with two-bilateral 12 electrode arrays plus a guide cannula was trained to accept the restraint associated with the recording apparatus and slept during recordings. If the rat was pretreated with TNF-α applied to the surface of the somatosensory cortex, surface evoked potentials were enhanced during quiet sleep compared to those occurring after heat-inactivated TNF-α (hiTNF). We propose that cytokines are driving neurons within cortical columns into a down state and thereby amplify surface evoked potentials. We also propose that the sleep functional state of the column is promoted by cytokines produced in response to cellular activity and is a function of the number of neurons in the down state.

on activity-dependent sleep regulatory substances such as TNF-α, then one would predict that direct application of TNF-α onto the surface of the cortex would drive the affected neural assemblies to simultaneously express the same functional state. Indeed, when TNF-α is applied in this manner, EEG delta power is enhanced unilaterally. Furthermore, if TNF-α is applied to the surface of the cortex and individual cortical columns are probed by whisker twitches, the probability of the column being in the sleeplike state increases (Figure 2). Such experiments strongly support the notion that sleep begins as a local process, being induced by the production of activity-dependent sleep regulatory substances such as TNF-α. Such a view has profound implications for our understanding of sleep regulation, sleep function, and unconsciousness.

Regulation of REMS by Prolactin and Related Molecules

The first anecdotal account implicating PRL in REMS regulation was from an experiment in which REMS was enhanced after systemic administration of PRL in hypophysectomized pontine cats. Subsequently it was shown that systemic PRL causes dose-dependent enhancements of REMS in rabbits. Thereafter, several lines of evidence implicated PRL in the regulation of REMS. The following were the major observations:

1. Injection of exogenous PRL stimulates REMS in cats, rabbits, and rats.
2. The increases in REMS develop slowly, over 2 or 3 h postinjection, and are maintained for several additional hours.
3. PRL-induced changes in sleep are selective for REMS.
4. PRL enhances REMS only during the light period in the rat.
5. In contrast, REMS is suppressed during the light period and is enhanced at night in mutant, PRL-deficient rats and PRL knockout mice.
6. Rats rendered chronically hyperprolactinemic by means of grafting pituitaries under the kidney capsule exhibit increases in REMS.
7. Both REMS and NREMS are enhanced in hyperprolactinemic pseudopregnant rats.
8. Restraint stress and ether stress are associated with enhanced circulating PRL and enhanced REMS. Restraint stress also enhances hypothalamic release of PRL, although stressor-induced release of pituitary PRL is thought to be responsible for the increases in REMS after stress.

Intracerebral administration of PRL or antibodies to PRL stimulates or inhibits REMS, respectively. The modulation of REMS is therefore a central action of PRL. PRL may come from two sources: it is produced by neurons in the hypothalamus and brain stem, and it is also released into the blood from the anterior pituitary. Circulating PRL is transported into the brain via specific receptor-mediated transport mechanisms residing in the choroid plexus. The REMS-promoting activity of blood-borne PRL is supported by the observation that REMS is enhanced in response to systemically injected PRL, excess PRL released from pituitary grafts, or stimulation of endogenous PRL secretion from the pituitary. Nevertheless, only slight decreases in REMS are observed when basal, nonstimulated blood PRL is suppressed by means of immunoneutralization. Although plasma PRL levels in humans are sleep related, being highest in the early morning hours, patients with PRL-producing adenoma fail to display changes in REMS; instead, their NREMS is increased. In summary, increases in blood PRL stimulate REMS in animal models, but circulating PRL at normal concentrations in the blood has only a minor effect on the regulation of REMS. The impact of the increased blood PRL concentrations on REMS is, in part, attributed to stimulation of the expression of PRL

receptors in the choroid plexus, whereby PRL enhances its own transport into the brain. It is thus posited that intracerebral PRL modulates REMS under normal conditions, whereas pituitary PRL provides additional stimulating influence when PRL secretion is high (e.g., during stress). PRL entering the brain from the systemic circulation acts on the same structures normally under the influence of intracerebral prolactinergic neurons.

Intracerebroventricular injections of hypothalamic peptides that cause pituitary release of PRL – vasoactive intestinal peptide (VIP) and pituitary adenyl cyclase activating peptide, both of which promote REMS – elicit expression of PRL mRNA in the hypothalamus and anti-PRL antibodies can block VIP-enhanced REMS responses. Another hypothalamic releasing factor for pituitary PRL, prolactin releasing peptide, also enhances REMS in rats if injected centrally.

The mechanism by which PRL increases REMS is unclear. PRL may modulate the circadian regulation of REMS at the level of the hypothalamus, or it may act in the brain stem structures generating REMS. The long (2 or 3 h) latency of the REMS response to exogenous PRL suggests that some metabolic actions mediate the sleep effects. Finally, growth hormone, which is structurally related to PRL, also stimulates REMS. Human growth hormone has a high affinity to PRL receptors and promotes REMS in both humans and rats.

The neuronal circuits responsible for the generation of REMS reside in the brain stem. The brain stem control of REMS involves the interactions between REMS-on neurons, which promote REMS, and REMS-off neurons, which inhibit REMS. Acetylcholine (ACh) is the best characterized neurotransmitter that promotes REMS. Cholinergic neurons localized in the laterodorsal tegmental nucleus (LDT) and the pedunculopontine tegmental nucleus (PPT) are involved in the generation of REMS signs such as EEG desynchronization, ponto-geniculo-occipital waves, and muscle atonia. Cholinergic neurons increase their firing rates approximately 20 s before the EEG signs of REMS and maintain higher activity levels during REMS than during NREMS, suggesting that these neurons participate in the initiation and maintenance of REMS. Numerous microinjection and microinfusion studies show that REMS is enhanced by cholinergic agonists and inhibited by cholinergic antagonists. Furthermore, kainate lesions of the PPT result in prolonged decreases in REMS. Anatomical data indicate that single cholinergic neurons in the LDT or PPT project to multiple brain regions that are responsible for the generation of individual signs of REMS. Therefore, these cholinergic neurons might be involved in the orchestration of different REMS signs to generate REMS.

A Proposed Mechanism for PRL-Enhanced REMS

REMS is influenced by metabolism. The length of the REMS cycles correlates with brain size and general metabolic activity. The brain primarily utilizes glucose as an energy source. Glucose is transported from the blood and into astrocytes through an active transport system and metabolized into pyruvate, the main fuel for neurons. In neurons, pyruvate may be metabolized into lactate by the enzyme lactate dehydrogenase in the absence of oxygen, or it may be metabolized into CO_2 and H_2O through the oxidative pathway which involves many enzyme actions, including pyruvate dehydrogenase (PDH). REMS is selectively suppressed during hypoxic hypoxia, whereas NREMS or waking may be enhanced. In contrast, an increase in oxygen availability induces a significant increase in REMS in chronic pontine cats kept at a fixed temperature. The high glucose and oxygen consumption levels during REMS compared to those during NREMS suggest that PDH activity is enhanced during REMS. It is suggested that glucose may be metabolized through the anaerobic pathways during wakefulness, converted to glycogen during NREMS, and metabolized through the aerobic pathways during REMS.

Stimulation of PDH in the mesencephalic/pontine cholinergic cells may serve as a final common pathway for various experimental interventions that promote REMS, such as REMS deprivation, immobilization stress, or administration of PRL. The connection between PRL and PDH was suggested by observations that PRL stimulated PDH activity in the prostate and the mammary gland. Increases in brain stem PDH activity occur in response to systemic injection of PRL.

The synthesis of ACh is critically dependent on the supply of acetyl-CoA, which in turn relies on the oxidation of pyruvate by the PDH complex during high glucose and oxygen consumption. PDH is particularly enriched in brain cholinergic neurons. Enhanced PDH activity results in increases in the concentration of acetyl-CoA and thereby stimulates the synthesis of ACh. Furthermore, brain cholinergic neurons are vulnerable to the deficiency of PDH. ACh levels are decreased by the PDH inhibitor 3-bromopyruvate

in brain homogenates or in the brain after local injection of this compound into the basal forebrain. Decreases in PDH activity elicited by experimental thiamine deficiency are also associated with the suppression of REMS.

Based on these observations, the proposed mechanism of the effects of PRL on sleep is as follows: PRL in the brain stimulates PDH and thereby energy production in neurons. This stimulation of PDH results in enhancement of ACh synthesis in cholinergic neurons because of an increased supply of acetyl-CoA. In our opinion, increased ACh production in itself does not trigger REMS. Due to the increased ACh pool, however, the quantity of ACh released increases whenever REMS spontaneously occurs. Therefore, PRL does not trigger REMS; instead, it facilitates spontaneous REMS. Furthermore, there is no evidence to suppose that PRL is specific to REMS-promoting cholinergic neurons. Cholinergic neurons also exist in the basal forebrain. These neurons are implicated in the mechanism of arousal, and their activities are associated with wakefulness. The importance of normal PDH activity has been demonstrated in the function of cholinergic neurons of the basal forebrain in experiments modeling the pathology of Alzheimer's disease. Stimulation of PDH in these neurons may explain the promotion of wakefulness by PRL during the dark period in rats. If further experiments substantiate that PRL acts via an increased synthesis of ACh due to enhanced activity of PDH, then PRL might be the first metabolic hormone whose effects on sleep can be linked to a distinct biochemical mechanism.

Conclusion

Our understanding of the biochemical regulation of sleep has progressed rapidly during the past 20 years. We now have a firm understanding of at least some of the humoral substances involved in sleep regulation. No doubt, other sleep regulatory substances will be described. Perhaps more important, these substances have already given insight into how the brain is organized to produce sleep. These mechanisms include a dynamic oscillation of states within neural assemblies coordinated into coherent whole animal states by sleep regulatory circuits and homeostatically driven by the accumulation of sleep regulatory substances.

See also: Dopamine Control of Arousal; Sleep and Sleep States: Thalamic Regulation; Sleep and Sleep States: Hypothalamic Regulation; Sleep and Sleep States: Histamine Role; Sleep–Wake State Regulation by Noradrenaline and Serotonin; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Eissner G, Kolch W, and Scheurich P (2004) Ligands working as receptors: Reverse signaling by members of the TNF superfamily enhance the plasticity of the immune system. *Cytokine & Growth Factor Reviews* 15: 353–366.

Krueger JM and Obal F Jr. (1993) A neuronal group theory of sleep function. *Journal of Sleep Research* 2: 63–69.

Krueger JM and Obal F Jr. (2003) Sleep function. *Frontiers in Bioscience* 8: 511–519.

Krueger JM, Obál F Jr., Opp M, Toth L, Johannsen L, and Cady AB (1990) Somnogenic cytokines and models concerning their effects on sleep. *Yale Journal of Biology and Medicine* 63: 157–172.

Majde JA and Krueger JM (2005) Links between the innate immune system and sleep. *Journal of Allergy and Clinical Immunology* 116: 1188–1198.

Maquet P, Peigneux P, Laureys S, et al. (2003) Memory processing during human sleep as assessed by functional neuroimaging. *Revue Neurologique (Paris)* 159: 6S27–6S29.

Marshall L and Born J (2002) Brain-immune interactions in sleep. *International Review of Neurobiology* 52: 93–131.

Obal F Jr. and Krueger JM (2003) Biochemical regulation of sleep. *Frontiers in Bioscience* 8: 520–550.

Opp MR (2005) Cytokines and sleep. *Sleep Medicine Reviews* 9: 355–364.

Pollmacher T, Schuld A, Kraus T, Haack M, Hinze-Selch D, and Mullington J (2000) Experimental immunomodulation, sleep, and sleepiness in humans. *Annals of the New York Academy of Sciences of the United States of America* 917: 488–499.

Rector DM, Topchiy IA, Carter KM, and Rojas MJ (2005) Local functional state differences between rat cortical columns. *Brain Research* 1047: 45–55.

Roky R, Obal F Jr., Valatx JL, et al. (1995) Prolactin and rapid eye movement sleep. *Sleep* 18: 536–542.

Shoham S, Davenne D, Cady AB, Dinarello CA, and Krueger JM (1987) Recombinant tumor necrosis factor and interleukin 1 enhance slow-wave sleep. *American Journal of Physiology* 253: R142–R149.

Steriade M (2003) The corticothalamic system in sleep. *Frontiers in Bioscience* 8: d878–d899.

Sleep and Sleep States: Gene Expression

C Cirelli and G Tononi, University of Wisconsin–Madison, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

The Molecular Correlates of Sleep May Offer Clues to Its Functions

All animal species studied to date sleep, and sleep may serve multiple functions, which may differ in different species. It is possible, however, that sleep also has a core function that is conserved from invertebrates to mammals. If this is the case, that function is most likely a cellular one because flies and humans share most pathways for intercellular and intracellular signaling, from membrane receptors and ion channels to nuclear transcription factors, but differ significantly in the number, anatomy, and complexity of brain circuits. The idea that the functions of sleep may ultimately relate to cellular and molecular aspects of neural function is not new. Giuseppe Moruzzi argued that "sleep concerns primarily not the whole cerebrum, nor even the entire neocortex, but only those neurons or synapses, and possibly glial cells, which during wakefulness are responsible for the brain functions concerned with conscious behavior." Moruzzi suggested that neural cells or synapses supporting waking conscious activity undergo plastic changes that make sleep necessary, although he did not speculate about the mechanisms. Others have suggested that sleep may help restore brain energy metabolism, may be needed to maintain the synaptic efficacy of the neural circuits not frequently used during wakefulness, or may favor a generalized synaptic downscaling following waking-induced synaptic potentiation. It has also been suggested that the rich spontaneous activity of neocortical neurons during slow-wave sleep may consolidate memory traces acquired during wakefulness. Several of these hypotheses are not mutually exclusive, and sleep may actually favor different cellular processes.

The analysis of the molecular correlates of sleep and wakefulness may help to understand the benefits that sleep brings at the cellular level. Specifically, the identification of the genes whose expression changes in the brain between sleep and wakefulness may clarify if, and why, brain cells need sleep and why their functions are impaired if they are prevented from doing so during sleep deprivation. In the past 15 years, subtractive hybridization, mRNA differential display, and microarrays have emerged as new powerful methods in transcriptomics, the study of the complete set of RNA transcripts produced by the genome at any one time. Microarrays are currently the most comprehensive and sensitive method to detect genome-wide expression changes, and they have allowed the identification of hundreds of transcripts whose expression is modulated by behavioral state. In addition, transcriptomics analysis has also identified functional categories of genes whose expression varies during sleep, wakefulness, and after different periods of sleep deprivation. At least some of these gene categories are conserved from invertebrates to mammals.

Table 1 lists the studies that have used transcriptomics approaches to identify genes whose transcript (mRNA) levels vary as a function of sleep, wakefulness, and/or sleep deprivation. It is clear from this table that the main focus has been on the cerebral cortex, although other brain regions have been studied, including cerebellum, hypothalamus, and brain stem. The cerebral cortex is of major interest because it is responsible for the cognitive defects observed after sleep deprivation, and it is at the center of most hypotheses concerning the functions of sleep.

The following general conclusions can be drawn from the studies listed in Table 1, and they are discussed further in this article:

- Hundreds of genes in the brain change their expression due to a change in behavioral state.
- Some genes increase their expression during sleep, whereas others do so during waking and sleep deprivation.
- Sleep and waking genes belong to different functional categories.
- Three functional categories of waking genes are conserved from flies to mammals.
- The noradrenergic system plays a role in the induction of some waking genes.
- The expression of several genes changes after chronic, but not acute, sleep loss.

Hundreds of Genes in the Brain Change Their Expression between Sleep and Waking

A first important finding of transcriptomics studies is that gene expression in the brain changes extensively as a function of behavioral states, sleep, spontaneous wakefulness, and short-term (a few hours) or long-term (several days) sleep deprivation.

Table 1 Studies that use transcriptomics approaches to identify molecular correlates of sleep (S), wakefulness (W), and sleep deprivation (SD)

Study[a]	*Method used/no. of tested transcripts*	*Experimental details*	*Major findings*
1	Subtractive hybridization (~4000 transcripts)	24 h SD (forced locomotion) vs. time-matched controls; rat forebrain	Four transcripts ↓ by SD, six transcripts ↑ by SD, all unknown
2	mRNA differential display (~9000 transcripts)	3 h S, W, SD (novel objects); rat cerebral cortex	One unknown transcript ↑ in S; 11 transcripts ↑ in W-SD, including *c-fos*, *NGFI-A*, and several mitochondrial genes
3	mRNA differential display (~9000 transcripts) + cDNA arrays (Clontech, 1176 transcripts)	8 h S, W, SD (novel objects); rat cerebral cortex	12 transcripts ↑ in S (all unknown except membrane protein E25); 44 transcripts ↑ in W-SD, including IEGs and heat shock proteins
4, 5	cDNA arrays (Clontech, 1176 + 588 +140 transcripts)	6 h SD (gentle handling), 6 h SD + 4 h recovery S vs. time-matched controls; mouse cerebral cortex (changes must be >100%); four array replicas	IEGs and heat shock proteins ↑ during SD Grp94 ↑ in recovery sleep
6	Affymetrix GeneChips, >24 000 probe sets, >15 000 reliably detected	8 h S, W, SD (novel objects); rat cerebral cortex (changes must be >20%); three to five array replicas; two statistical methods; qPCR; estimated false positives ~20%, based on qPCR confirmation of 52 transcripts (using independent groups of animals)	106 known transcripts and 155 ESTs ↑ in S; 95 known transcripts and 395 ESTs ↑ in W-SD
7	mRNA differential display (~5000 transcripts) + Affymetrix GeneChips, ~8000 probe sets, ~5000 reliably detected	8 h SD with or without prior LC lesion (DSP-4); rat cerebral cortex (changes must be >20%); three array replicas; two statistical methods; estimated false positives <20%, based on qPCR confirmation of 32 transcripts (using independent groups of animals)	20% (16/95) of the known transcripts ↑ in W-SD downregulated after cortical depletion of norepinephrine
8	Affymetrix GeneChips, 18 955 probe sets, ~9400 reliably detected	9 h S, W, SD (forced locomotion); fly head (changes must be >50% in two fly lines); four array replicas; one statistical method; estimated false positives <20%, based on qPCR confirmation of 19 transcripts (using independent groups of animals)	12 transcripts ↑ in S and 121 transcripts ↑ in W-SD
9	Affymetrix GeneChips, 1322 probe sets, ~600 reliably detected	6 h SD (gentle handling), 6 h SD + 2 h recovery S vs. time-matched controls; rat cerebral cortex, hypothalamus, basal forebrain (changes must be >50%); two array replicas	
10	Affymetrix GeneChips, 26 261 probe sets, ~17 000 reliably detected	1-week SD (disk-over-water method) compared to yoked controls, 9 h W, and 9 h SD; rat cerebral cortex (changes must be >20%); three to eight array replicas; one statistical method; estimated false positives <5%, based on qPCR confirmation of 15 transcripts (using independent groups of animals)	75 transcripts ↑ and 28 transcripts ↓ after long-term SD

[a]Studies: 1, Rhyner TA, Borbely AA, and Mallet J (1990) Molecular cloning of forebrain mRNAs which are modulated by sleep deprivation. *European Journal of Neuroscience* 2: 1063–1073; 2, Cirelli C and Tononi G (1998) Differences in gene expression between sleep and waking as revealed by mRNA differential display. *Brain Research: Molecular Brain Research* 56: 293–305; 3, Cirelli C and Tononi G (2000) Gene expression in the brain across the sleep–wakefulness cycle. *Brain Research* 885: 303–321; 4, Terao A, Steininger TL, Hyder K, et al. (2003) Differential increase in the expression of heat shock protein family members during sleep deprivation and during sleep. *Neuroscience* 116: 187–200; 5, Terao A, Greco MA, Davis RW, Heller HC, and Kilduff TS (2003) Region-specific changes in immediate early gene expression in response to sleep deprivation and recovery sleep in the mouse brain. *Neuroscience* 120: 1115–1124; 6, Cirelli C, Gutierrez CM, and Tononi G (2004) Extensive and divergent effects of sleep and wakefulness on brain gene expression. *Neuron* 41: 35–43; 7, Cirelli C and Tononi G (2004) Locus coeruleus control of state-dependent gene expression. *Journal of Neuroscience* 24: 5410–5419; 8, Cirelli C, LaVaute TM, and Tononi G (2005) Sleep and wakefulness modulate gene expression in *Drosophila*. *Journal of Neurochemistry* 94: 1411–1419; 9, Terao A, Wisor JP, Peyron C, et al. (2006) Gene expression in the rat brain during sleep deprivation and recovery sleep: An Affymetrix GeneChip study. *Neuroscience* 137: 593–605; 10, Cirelli C, Faraguna U, and Tononi G (2006) Changes in brain gene expression after long-term sleep deprivation. *Journal of Neurochemistry* 98(5): 1632–1645.
ESTs, Expressed sequence tags; IEGs, immediate early genes.

Five percent or more of the transcripts tested in the brain show changes in gene expression according to behavioral state. Interestingly, a similar percentage of genes (1–10%) change because of differences in circadian time. In the cerebral cortex of rats, out of the 15 459 transcripts we tested, 808 (5.2%) were affected by time of day independently of behavioral state, and 752 (4.9%) were affected by sleep and wakefulness independently of time of day. These data indicate that, at least in the rat, day/nighttime and sleep/wakefulness influence gene expression in the cerebral cortex to a similar extent. Peripheral tissues such as liver and heart also show significant changes in gene expression as a function of circadian time. Whether this is also true for sleep-dependent and wakefulness-dependent genes is unknown because none of the array studies published so far has focused on peripheral tissues.

Some Genes Increase Their Expression during Sleep, Whereas Others Do So during Waking and Sleep Deprivation

We found that in the rat cerebral cortex, 95 known genes and 395 expressed sequence tags (ESTs; i.e., short transcribed sequences whose function is unknown) show higher expression after 8 h of spontaneous wakefulness or sleep deprivation, whereas 106 known genes and 155 ESTs show increased expression after 8 h of sleep. In the fly head, for which we used the same experimental protocol previously applied to rats (three experimental groups, including 8 h of sleep, 8 h of spontaneous wakefulness, and 8 h of total sleep deprivation), we identified 121 waking transcripts, including both known genes and ESTs, and 12 sleep transcripts, mostly unknown. Finally, we compared gene expression in the cerebral cortex of long-term sleep-deprived rats and sleeping controls and found that 226 transcripts are upregulated after long-term sleep deprivation and 113 after sleep. Thus, it seems that if all transcribed sequences are considered (including ESTs), there are more genes whose expression is upregulated during waking than during sleep. However, hundreds of genes are also upregulated during sleep, confirming the results of intracellular recording studies that show that sleep is far from being a quiescent state of global inactivity. In general, the largest fold changes are observed for waking genes. However, this is mainly due to the fact that waking increases the expression of transcription factors (e.g., c-fos) and effector genes (e.g., Arc) that behave as 'immediate early genes,' whose mRNA and protein levels may increase severalfold relative to sleep. Most waking and sleep genes, instead, change their expression by 20–50%, a fold change similar to that seen for most genes whose expression varies because of circadian time.

Sleep and Waking Genes Belong to Different Functional Categories

Sleep and waking transcripts belong to different functional categories and thus may favor different cellular processes. Most genes whose expression increases during wakefulness (or short-term sleep deprivation) code for proteins involved in energy metabolism, synaptic excitatory transmission, transcriptional activity, acquisition of new information, and the response to cellular stress. Sleep genes, instead, code for proteins involved in protein synthesis, synaptic downscaling and/or memory consolidation, lipid metabolism, and in membrane trafficking and maintenance.

Three Functional Categories of Waking Genes Are Conserved from Flies to Mammals

Three functional categories of genes are consistently induced in several animal species (fruit flies, mice, rats, and hamsters) during waking and short-term sleep deprivation relative to sleep. The first one includes genes involved in energy metabolism, including mitochondrial genes such as cytochrome oxidase (CO) I and IV. One study in rats also found an increase in CO enzymatic activity after 3 h of wakefulness. In mice, 6 h of sleep deprivation also affects the mRNA levels of genes involved in glycogen metabolism (glycogen synthase and glycogen phosphorylase) and increases the activity of glycogen synthase, possibly to counterbalance the rapid depletion, at the transition from sleep to wakefulness, of astrocytic glycogen stores. Eight hours of spontaneous waking or sleep deprivation also induces the expression in the rat cerebral cortex of *Glut1*, the gene coding for one of the more widely expressed glucose transporters in the central nervous system. The upregulation during waking of genes involved in energy metabolism is perhaps not surprising because glucose metabolism in many brain regions is higher in wakefulness than in non-rapid eye movement (NREM) sleep, and the brain relies almost exclusively on glucose as its energy substrate. These changes, however, do not seem to persist when sleep deprivation is protracted for 24 h or several days. This is consistent with imaging studies in animals, normal human subjects, and fatal familial insomnia patients showing no change or a decrease in cerebral metabolic rate after prolonged sleep loss relative to normal wakefulness.

Thus, the rapid upregulation of mitochondrial genes, *Glut1*, and glycogen-related genes may represent a mechanism by which the brain responds to the immediate increase in energy requirement during spontaneous waking or after a few hours of sleep deprivation, whereas prolonged sleep loss is associated with reduced energy demands.

A second group of genes consistently upregulated during wakefulness and downregulated during sleep includes heat shock proteins, chaperones, and, in general, genes involved in the response to cellular stress. Cellular stress is the cell's reaction to any adverse environmental condition that perturbs cellular homeostasis, with potential macromolecular damage (i.e., damage to proteins, DNA, RNA, and lipids). Wakefulness and short-term deprivation are associated with higher mRNA levels of small heat shock proteins with actin-stabilizing properties, such as HSP27 and αB crystallin; molecular chaperones (HSP70 and HSP60); and chaperones in the endoplasmic reticulum, such as BiP and Grp94. BiP (immunoglobulin heavy chain-binding protein) shows the largest and most consistent upregulation in the brain of awake animals, including mice, rats, hamsters, and sparrows, as well as in the head of fruit flies. BiP assists in the folding and assembly of newly synthesized glycoproteins and secretory polypeptides and increases in *BiP* expression have been classically described after heat shock and ischemia, when unfolded proteins accumulate in the endoplasmic reticulum. This unfolded protein response is also induced by other conditions that adversely affect the function of the endoplasmic reticulum, such as anoxia, glucose deprivation (BiP is also called glucose-regulated protein 78), and depletion of Ca^{2+} stores in the endoplasmic reticulum. Importantly, the unfolded protein response also triggers a slowing down of protein synthesis. In the rat cerebellum, 8 h of spontaneous wakefulness or sleep deprivation is associated with an increase in the mRNA levels of *PEK*. *PEK* codes for the EIF2a kinase that mediates the inhibition of protein synthesis during the unfolded protein response. Moreover, in the mouse cerebral cortex, 6 h of sleep deprivation is also associated with signs of slowing down of protein synthesis, as indicated by the phosphorylation of key components of the translational machinery and by changes in the ribosomal profile.

A third group of genes consistently upregulated during wakefulness and downregulated during sleep includes transcripts sensitive to membrane depolarization and synaptic activity. In the rat cerebral cortex, they include *RGS2*, *Homer/Vesl*, tissue-type plasminogen activator, casein kinase 2, cyclooxygenase 2, *cpg2*, and connexin 30. Moreover, several genes involved in activity-dependent neural plasticity and long-term potentiation, such as *Arc*, *BDNF*, *Homer/Vesl*, and *NGFI-A*, have been consistently found to be upregulated in the cerebral cortex of awake or short-term sleep-deprived rats. At least some of the same genes are induced during wakefulness and/or short-term sleep deprivation in mice, hamsters, and sparrows. In the fly head, many of the depolarization-sensitive genes involved in synaptic plasticity as identified by Guan et al. in 2005 are also waking-related genes, including *Astray*, *Kek2*, *stripe*, and *Hr38*. The fly wakefulness-related gene *stripe* is a member of the Egr family of transcription factors. In the rat brain, several members of this family, including *NGFI-A* (*Egr-1*), *krox 20* (*Egr-2*), and *NGFI-C* (*Egr-4*), are induced during wakefulness. *Hr38*, another fly wakefulness-related gene, is the *Drosophila* ortholog of the mammalian NGFI-B family of orphan nuclear receptors. *NGFI-B* is a wakefulness-related gene in the rat cerebral cortex. Thus, it seems that in both mammals and flies several plasticity-related genes involved in synaptic growth and potentiation are specifically induced during wakefulness.

The observation that genes involved in long-term potentiation and the acquisition of new memories are specifically induced during wakefulness should not be too surprising. After all, we learn when we are awake and able to adapt to the environment, and not when we are asleep. However, if sleep is not involved in learning, is there a role for it in other aspects of synaptic plasticity? Well-controlled experiments show that sleep may favor the performance of some, but not all, tasks. The strongest and more consistent data relate to procedural tasks and suggest an important role for NREM sleep or a combination of NREM and REM sleep. What remains unknown, however, are the mechanisms by which sleep may enhance performance. Animal and human studies have shown that brain circuits activated during the execution or the acquisition of certain tasks are 'reactivated' during the subsequent sleep period. The presence of this memory trace, as shown by electrophysiological recordings and imaging studies, could reflect the fact that sleep replays the waking experience, and by doing so it somehow allows the specific consolidation of the learned tasks. However, it is also possible that this trace is the inevitable consequence of the synaptic potentiation that occurs during learning, and that sleep enhances performance by actually allowing a generalized downscaling of synapses. The first and most important effect of this downscaling is thought to be the decrease in the metabolic cost of synaptic activity, which increases during waking due to synaptic potentiation. A second effect of the generalized but proportional decrease in synaptic weight is an increase in the signal-to-noise

ratio, which could improve the performance of some learned tasks.

If sleep is associated with a generalized depression of synaptic strength, molecular markers of synaptic depression should be increased during sleep relative to waking. Studies have identified and validated several specific and direct markers of synaptic strength, with the best established among them being the density, composition, and phosphorylation state of the GluR1 subunit of the glutamatergic AMPA receptors. We are currently measuring the number and phosphorylation state of synaptic GluR1 in the rat cerebral cortex after 6 h of sleep or sleep deprivation. Preliminary results show that in several cortical areas, sleep is associated with a relative dephosphorylation of GluR1 at serine 845, a finding consistent with the hypothesis that sleep is associated with synaptic downscaling.

The Noradrenergic System Plays a Role in the Induction of Some Waking Genes

One of the mechanisms that underlie the widespread changes in cortical gene expression between sleep and wakefulness is the activity of the noradrenergic system of the locus coeruleus (LC), whose neurons project diffusely over the entire brain. LC cells are tonically active during wakefulness, reduce their firing rate during NREM sleep, and cease firing during REM sleep. Moreover, LC activity increases phasically in response to salient events and in relation to the decision to act. In a series of experiments, rats were subjected to unilateral or bilateral LC lesions to deplete one or both sides of the brain of norepinephrine. In these animals, the raw electroencephalogram (EEG) and its power density spectrum were not significantly affected by the lesion. However, all cortical areas depleted of norepinephrine showed a marked decrease, during wakefulness, of the expression of plasticity-related genes such as *Arc*, *BDNF*, *NGFI-A*, and *P-CREB* as well as of stress response genes such as heat shock proteins and *BiP*. By contrast, the transcript for the translation elongation factor 2 was the only known sleep-related transcript whose expression increased after LC lesion (Figure 1). In a complementary experiment in mice, the activity of the LC of one side was increased using a conditional transgenic approach. This manipulation resulted in an increased ipsilateral expression of *NGFI-A* in cortical and subcortical target areas. Thus, LC activity during wakefulness modulates neuronal transcription to favor synaptic potentiation and memory acquisition and to counteract cellular stress, whereas LC inactivity during sleep may play a permissive role to enhance brain protein synthesis.

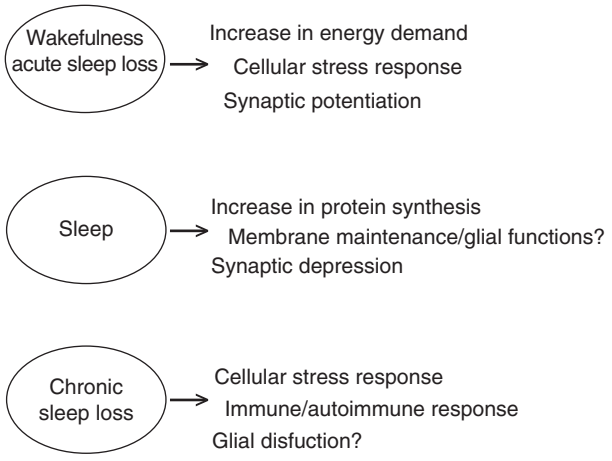

Figure 1 Schematic representation of the major functional categories of genes upregulated in the rat cerebral cortex after 3–8 h of spontaneous waking or sleep deprivation (acute sleep loss), 1 week of total sleep deprivation (chronic sleep loss), and 8 h of spontaneous sleep.

The Expression of Several Genes Changes after Chronic, but Not Acute, Sleep Loss

In an array study, we analyzed the expression of more than 26 000 transcripts in the cerebral cortex of long-term sleep-deprived rats. Animals were kept awake for 7 days using the disk-over-water method. This method uses minimal stimulation to enforce chronic sleep deprivation in the sleep-deprived rat, while it simultaneously applies to the control rat the same stimulation but without severely limiting its sleep. The sleep-deprived and the control rats are each housed on one side of a divided horizontal disk suspended over a shallow tray of 2 or 3 cm deep water. EEG and electromyogram are continuously recorded to detect sleep states. When the experimental rat falls asleep, the disk is automatically rotated at low speed, awakening the rat and forcing it to walk opposite to disk rotation to avoid being carried into the water. The yoked control rat receives the same physical stimulation because it is on the same disk. However, whereas sleep is severely reduced (by 70–80%) in the sleep-deprivation rat, the control rat can sleep whenever the sleep-deprived rat is spontaneously awake and eating, drinking, or grooming, and thus its sleep is reduced by only 25–40%.

We compared long-term sleep-deprived rats with rats spontaneously awake or sleep deprived for only 8 h and identified 75 transcripts upregulated and 28 transcripts downregulated after chronic sleep loss. Most transcripts specifically increased or decreased their expression due to the long-term sleep deprivation and not because of the disk stimulation, because they were not affected in the yoked controls. A large group

of genes upregulated after prolonged sleep loss codes for immunoglobulins, including two autoantibodies. Another group includes stress response genes such as the macrophage inhibitor factor-related- protein MRP14, expressed in microglia. Arylsulfotransferase, an enzyme involved in the catabolism of catecholamine, is also upregulated after chronic sleep deprivation. These genes are either not induced (antibodies and MRP14) or induced to a lesser extent (arylsulfotransferase) after acute sleep loss, suggesting that short-term and long-term sleep deprivation differ to some extent at the molecular level. We also compared long-term and short-term sleep deprivation to sleep and found that only a minority of genes were similarly up- or downregulated in the two sleep- deprivation conditions. Among the genes whose expression was most significantly downregulated in long-term sleep-deprived animals relative to sleeping animals were myelin constituents, including the most abundant structural protein component of myelin – the proteolipid protein. These findings suggest that sustained sleep loss may trigger a generalized inflammatory and stress response in the brain. Some glial cells (e.g., microglia) may help to protect neurons against this cellular insult, but others (e.g., oligodendrocytes) may also suffer some of its negative consequences. It will be important in the future to test whether sleep loss may be detrimental to the maintenance of cellular membranes and specifically to myelin.

Genes Upregulated during Sleep

Figure 1 shows the sleep-related transcripts identified so far in the rat cerebral cortex. Unfortunately, only one study in mammals was designed to identify genes upregulated after several (8) hours of sleep relative to both spontaneous waking and sleep deprivation to rule out effects of stress and circadian time. Another study in flies used the same experimental conditions used in rats, but for technical reasons the entire fly head, rather than the brain only, was used to hybridize the arrays. This most likely decreased the ability to detect state-dependent genes. Only 12 genes were identified as sleep-related in flies, and the function of most of them is unknown. Thus, the extent to which sleep genes are conserved across species is yet to be determined.

Two sleep genes in the rat cerebral cortex code for the translation elongation factor 2 and the initiation factor 4AII, and they are involved in protein synthesis. The mRNA levels of the elongation factor EF2 are also increased during sleep in the brain of Djungarian hamsters. A positive correlation between sleep and protein synthesis has been suggested by previous studies, and evidence suggests that waking and sleep deprivation are less conducive to protein synthesis than sleep. Whether sleep favors protein synthesis globally or whether it enhances the synthesis of specific classes of proteins is unclear.

Other sleep genes identified in the rat cerebral cortex are involved in the consolidation of long-term memory as well as in synaptic depression and depotentiation. Among them are calmodulin-dependent protein kinase IV, calcineurin, FK506 binding protein 12, inositol 1,4,5-trisphosphate receptor, and amphiphysin II. As mentioned previously, the mechanism by which sleep enhances memory and performance is still debated, but these molecular data are compatible with the idea that sleep benefits the brain by producing a global synaptic downscaling.

Another large group of sleep-related transcripts in the rat cerebral cortex is involved in membrane trafficking and maintenance. Some of these transcripts are important for the synthesis/maintenance of membranes in general and of myelin in particular, including oligodendrocytic genes coding for myelin structural proteins (e.g., *MOBP*, *MAG*, *plasmolipin*, and *CD9*), myelin-related receptors, and enzymes. Transcripts with higher expression in sleep also code for enzymes involved in the synthesis and transport of cholesterol, a major constituent of myelin and other membranes and an important factor in regulating synaptic efficacy. As mentioned previously, long-term sleep deprivation decreases the expression of the gene coding for proteolipid protein, a major myelin component. Intriguingly, among the few known sleep genes in *Drosophila* are the gene *anachronism* (*ana*) and three genes involved in lipid metabolism (*CG8756*, *CG9009*, and *CG11407*). Ana is a glycoprotein secreted by glial cells, and it controls the timing of neuroblast development. Whether *ana* plays any role in membrane trafficking and synaptic plasticity remains to be determined.

Conclusions

As it is apparent from the small number of studies in Table 1, transcriptomics analysis is still a work in progress, and it would be inappropriate to draw firm conclusions from what has been published to date. Continuous improvement in the array technology will certainly favor more progress in this area. For instance, the current arrays can assess the expression of 20 000–30 000 transcripts, but this is still far from a true 'whole genome' analysis. Low expression genes are often difficult to detect, especially in large and heterogeneous tissue samples such as most brain areas. Moreover, current arrays focus on the coding portion of the genome, but new 'tiling' arrays, in which probe sets are laid down at regular intervals

across the entire genome, are becoming commercially available. This will permit the identification of new transcriptionally active regions of the genome that have not been tested in more conventional arrays. Moreover, proteomics, the study of the complete repertoire of proteins that contribute to a cell's physiological phenotype, is a natural next step in the study of molecular correlates of behavioral states.

See also: Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Architecture; Sleep Oscillations and PGO Waves; Sleep and Sleep States: PET Activation Patterns; Sleep and Sleep States: Hippocampus–Neocortex Dialog; Sleep and Sleep States: Network Reactivation; Sleep Deprivation and Brain Function; Sleep and Waking in Drosophila.

Further Reading

Cirelli C, Faraguna U, and Tononi G (2006) Changes in brain gene expression after long-term sleep deprivation. *Journal of Neurochemistry* 98: 1632–1645.

Cirelli C, Gutierrez CM, and Tononi G (2004) Extensive and divergent effects of sleep and wakefulness on brain gene expression. *Neuron* 41: 35–43.

Cirelli C, LaVaute TM, and Tononi G (2005) Sleep and wakefulness modulate gene expression in *Drosophila*. *Journal of Neurochemistry* 94: 1411–1419.

Cirelli C, Pompeiano M, and Tononi G (1996) Neuronal gene expression in the waking state: A role for the locus coeruleus. *Science* 274: 1211–1215.

Cirelli C and Tononi G (1998) Differences in gene expression between sleep and waking as revealed by mRNA differential display. *Brain Research: Molecular Brain Research* 56: 293–305.

Cirelli C and Tononi G (2000) Gene expression in the brain across the sleep–wakefulness cycle. *Brain Research* 885: 303–321.

Cirelli C and Tononi G (2004) Locus coeruleus control of state-dependent gene expression. *Journal of Neuroscience* 24: 5410–5419.

Guan Z, Saraswati S, Adolfsen B, and Littleton JT (2005) Genome-wide transcriptional changes associated with enhanced activity in the *Drosophila* nervous system. *Neuron* 48: 91–107.

Moruzzi G (1972) The sleep–waking cycle. *Ergebnisse der Physiologie* 64: 1–165.

Naidoo N, Giang W, Galante RJ, and Pack AI (2005) Sleep deprivation induces the unfolded protein response in mouse cerebral cortex. *Journal of Neurochemistry* 92: 1150–1157.

Nikonova EV, Vijayasarathy C, Zhang L, et al. (2005) Differences in activity of cytochrome C oxidase in brain between sleep and wakefulness. *Sleep* 28: 21–27.

Petit JM, Tobler I, Allaman I, Borbely AA, and Magistretti PJ (2002) Sleep deprivation modulates brain mRNAs encoding genes of glycogen metabolism. *European Journal of Neuroscience* 16: 1163–1167.

Rhyner TA, Borbely AA, and Mallet J (1990) Molecular cloning of forebrain mRNAs which are modulated by sleep deprivation. *European Journal of Neuroscience* 2: 1063–1073.

Taishi P, Sanchez C, Wang Y, Fang J, Harding JW, and Krueger JM (2001) Conditions that affect sleep alter the expression of molecules associated with synaptic plasticity. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 281: R839–R845.

Terao A, Greco MA, Davis RW, Heller HC, and Kilduff TS (2003) Region-specific changes in immediate early gene expression in response to sleep deprivation and recovery sleep in the mouse brain. *Neuroscience* 120: 1115–1124.

Terao A, Steininger TL, Hyder K, et al. (2003) Differential increase in the expression of heat shock protein family members during sleep deprivation and during sleep. *Neuroscience* 116: 187–200.

Terao A, Wisor JP, Peyron C, et al. (2006) Gene expression in the rat brain during sleep deprivation and recovery sleep: An Affymetrix GeneChip study. *Neuroscience* 137: 593–605.

Tononi G and Cirelli C (2006) Sleep function and synaptic homeostasis. *Sleep Medicine Reviews* 10: 49–62.

Sleep and Sleep States: Hippocampus–Neocortex Dialog

J J Chrobak, University of Connecticut, Storrs, CT, USA
A Sirota and G Buzsáki, Rutgers University, Newark, NJ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Understanding how large groups of neurons interact across brain structures is important for understanding brain function. During both awake and asleep brain states there are structured interactions between groups of neurons (ensembles) in the hippocampus and groups of neurons in the neocortex. Ensembles fire together in predictable patterns during hippocampal theta and sharp waves in both the awake and sleeping brain, as well as during neocortical sleep spindles, delta waves, and slow oscillations. How are the ensemble patterns observed in the hippocampus related to ensemble patterns observed in the neocortex, and do they differ in the awake versus asleep brain?

Two prominent hippocampal activity patterns (theta and sharp waves) occur during specific behavioral states (Figure 1). The theta pattern occurs during active exploration of the environment, with a number of neurons throughout the hippocampal system discharging collectively every 50–100 ms. Theta is considered an 'online' state in which information about the 'world' is channeled from primary sensory pathways through the neocortical associative cortices (e.g., perirhinal and parahippocampal cortices) and then fed into the circuitry of the hippocampus. Theta pattern also occurs during rapid eye movement (REM) sleep.

During moments of immobility and quiet resting, sharp waves (SPWs) dominate the hippocampus, with large numbers of hippocampal neurons discharging in a fast-frequency (200 Hz) 'ripple' volley in relation to the slow (~100 ms) SPWs. Sharp waves are considered an output state, when the hippocampus reengages neocortical circuits, perhaps providing a comment, an index, or simply a replay about the content of prior theta-related input. This pattern has been suggested to serve an 'offline' consolidation process. Similar SPW patterns also dominate the hippocampus during slow-wave sleep (SWS). Do the ensemble firing patterns during SPWs in awake and sleep states support different functions? Defining the similarities and differences across awake and sleep states may shed light on their significance as well as on the function of distinct sleep states. More importantly, it is necessary to uncover the mechanism by which these 'replay' patterns are coordinated with the ongoing neocortical activity, to ensure information transfer between the two structures.

It has been often suggested that sleep and particular sleep stages (e.g., REM) play an important role in memory, perhaps consolidating important information or erasing trivial information gathered by the waking brain. According to several models, memory involves an initial encoding stage followed by a longer period of consolidation, with some models suggesting that ensemble interactions during REM and SWS contribute to memory consolidation. Several research groups have demonstrated that neurons within the hippocampus that fire together during exploration fire together and in temporally related sequences during both REM and SWS. Could these structured interactions during sleep support memory consolidation processes? What are the mechanisms that give rise to these interactions?

Information about the spatial extent and temporal synchronization of hippocampal and neocortical networks during both asleep and awake brain states may offer insight into the mechanism and function of large-scale integration and segregation of ensemble activity. Do the coordinated discharges of hippocampal neurons during SPWs actually influence neocortical neurons? What influence does the concerted action of neocortical neurons during sleep rhythms (spindles, delta waves, slow oscillation) have on hippocampal ensemble activity? Sleep spindles and delta waves reflect a concerted discharge of thalamic, thalamocortical, and neocortical neurons, and several findings now link the occurrence of SPWs to the synchronization of neocortical neurons during sleep spindles and delta waves. Thus, fairly large networks of hippocampal, thalamic, and neocortical neurons have the potential to synchronize their activity during SPWs. Do these coordinated interactions support the specific transfer of information between hippocampal and neocortical cell assemblies?

We suggest that structured interactions among neurons during awake and sleep serve as building blocks for systems-level functions. The dialogs between the hippocampus and neocortex during SPWs, during both awake and sleep states, may play an important role in consolidating hippocampal-dependent memory traces in neocortical stores. A protracted consolidation process, taking place over days/months, could then support the observation that hippocampal insults do not affect memories acquired at some point prior to insult (equally days/months/years). The logic and evidence in

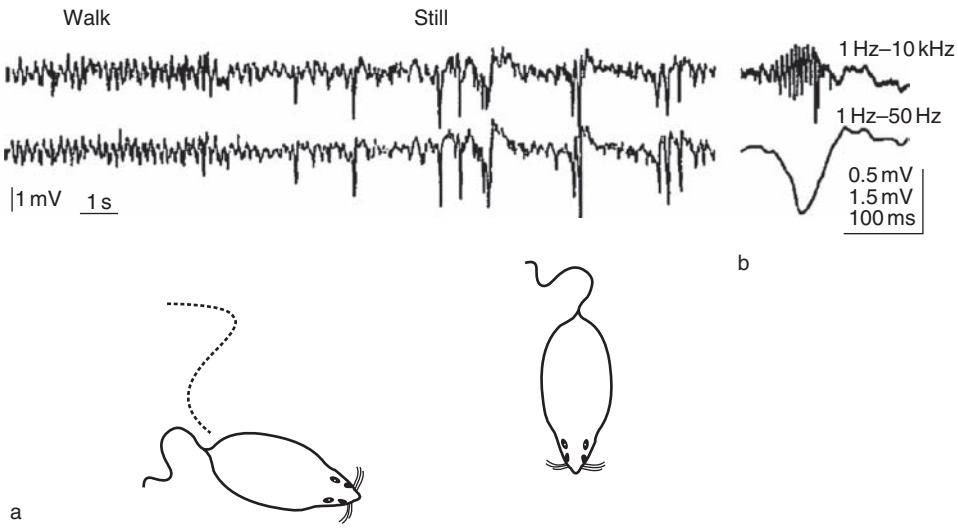

Figure 1 Behavior-dependent macroscopic (electroencephalogram) states in the hippocampus. (a) Extracellular recordings from the CA1 stratum radiatum of the left (top) and right (bottom) hippocampus during the transition from exploration walking to being still. Note regular theta waves during walking, and large, negative sharp waves during immobility. Note also that sharp waves are bilaterally synchronous. (b) A single sharp wave with simultaneously recorded fast-field oscillation from the CA1 pyramidal layer at a faster time scale. Adapted from Buzsaki G (1989) Two-stage model of memory formation: A role for 'noisy' brain states. *Neuroscience* 31: 551–570, with permission from Elsevier.

support of a role for SPWs and neocortical oscillations in hippocampal-dependent memory, and their role in hippocampal–neocortical dialogs during SWS, are overviewed in the following sections.

The Basics: A Two-Stage Model of Memory Trace Formation

The neuronal activity of the hippocampal formation (hippocampus, subiculum, and entorhinal cortex) is organized around two population patterns, theta and SPW. Theta and SPW patterns seem to serve companion processes. The theta pattern occurs during exploratory behavior and REM sleep. During theta, hippocampal circuits receive rhythmic input from neurons within the entorhinal cortex. The entorhinal cortex transmits information from neocortical association areas (e.g., perirhinal, parahippocampal) to the hippocampus. Thus, theta synchronizes the input pathways into the hippocampus from the neocortex.

In contrast, SPWs occur during consummatory behavior and SWS. During SPWs, the output neurons of the hippocampal formation participate in organized population bursts. During each burst, SPW neurons in the CA1, subiculum, and deep layers of the entorhinal cortex fire fast-frequency population volleys at ~200 Hz. The fast-frequency discharge is accompanied by a fast-frequency (~200 Hz) field potential referred to as a 'ripple' near the perisomatic region. The ripple field potential reflects synchronized inhibitory input to the soma from local γ-aminobutyric acid (GABA)ergic neurons (e.g., basket cells).

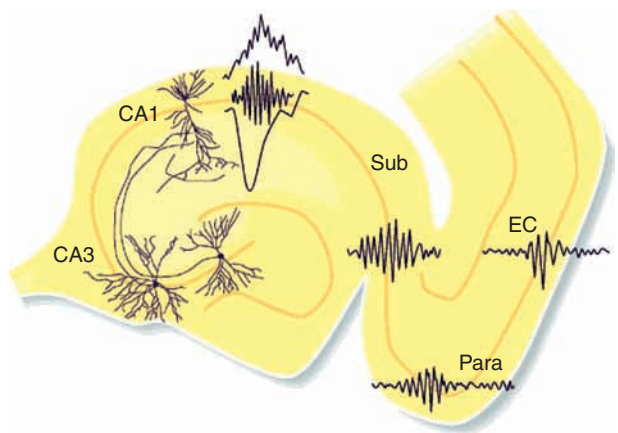

Figure 2 Self-organized burst of activity in the hippocampal CA3 region produces a field potential in the dendritic layer of CA1 and a short-lived fast-frequency field oscillation (200-Hz ripple) within stratum pyramidale, as well as a phase-related discharge of the neurons. The hippocampal output, in turn, produces similar sharp wave–ripple complexes in the subiculum (Sub), parasubiculum (Para), and deep layers of the entorhinal cortex (EC). Adapted from Chrobak JJ and Buzsaki G (2005) Synaptic plasticity and self-organization in the hippocampus. *Nature Neuroscience* 11: 1560–1567, with permission.

Each SPW is a transient (40–100 ms) population volley that involves the discharge of roughly 40 000 to 80 000 neurons in the CA3–CA1–subicular complex–entorhinal cortex axis within a 50- to 200-ms time frame. Both SPWs and their associated fast-frequency ripples (Figure 2) are present in the hippocampus of

all mammals investigated, including humans. Several features of this fast-frequency population volley make this network pattern an excellent candidate for altering synaptic connectivity. First, the time window of the SPW population burst roughly corresponds to the time constant of the *N*-methyl-D-aspartate (NMDA) receptor, through which Ca^{2+}, the key ion for the induction of synaptic plasticity, enters the neuron. Second, the fast-frequency network output (~200 Hz) is the ideal frequency range for initiating long-term potentiation (LTP). Third, during the SPW burst there is a three- to fivefold gain of network excitability, creating an ideal condition for synaptic potentiation. Last, the SPW engages the neurons that make up the output network of the hippocampus (CA1, subiculum, and entorhinal cortex). This network logically would transfer information back to neocortical structures, and SPW-related discharge has been observed in a number of hippocampal/entorhinal target structures, including the perirhinal cortex, amygdala, ventral striatum, and prefrontal cortices.

Hippocampal Replay of Waking Patterns during Sleep

Neuronal activity during sleep may promote memory-related synaptic plasticity more effectively than does neuronal activity during awake states. Why? First, neurons, free from the influence of sensory-driven input, can discharge independently and perhaps randomly, thereby erasing synaptic modifications brought about by the awake brain. The expected result could be to refresh or wipe clean the synaptic modifications associated with previous experience. This would provide a fresh 'tabula rasa' every morning, ready for information from the new day. Alternatively, neuronal pathways used and modified in the awake brain could be replayed repeatedly, and the repeated activation could be used to enhance important/relevant synaptic modifications. The molecular processes supporting synaptic potentiation are quite protracted (hours/days), thus perhaps necessitating repeated instantiation of the patterns supporting potentiation.

Because SPWs occur during immobility, during consummatory behaviors, and, mostly, during SWS, it was inferred that SPWs during SWS could play an important role in memory consolidation. The repeated replay of ensemble patterns (representations) related to recent ensemble patterns could provide a dynamic network mechanism for the slow, offline, transfer of hippocampal computations to the neocortex. Some specific predictions of the model are that (1) neurons do not randomly participate in SPW volleys; (2) synaptic connectivity, firing rates, and population patterns during SWS remain similar unless specifically altered by awake experience; and (3) SPWs induce synaptic modifications in the output networks of the hippocampal formation.

Several physiological experiments support the conjectures and predictions of the two-stage model of memory consolidation. Ensembles of neurons discharging in spatial specific (place cells) environments continue to fire together at a high rate during subsequent SWS and REM sleep episodes (Figure 3). This replay of neuronal patterns has been observed in both rats and monkeys and has also been observed during SPWs during awake immobility as well as during SWS. Neurons 'tied' together by sensory experience(s) in the awake world tend to discharge together whenever the probability of their discharge is increased. In most of these studies, no specific 'learning' experience took place. Place (spatial)-specific neurons fired as a rat navigated on a linear track or circular environment, and subsequently discharged together some time later, during both awake states and during subsequent sleep (both REM sleep and SWS). Typically, the probability of two or more neurons discharging together is increased during sleep directly following an experience, as compared to sleep before spatial exploration. In addition to showing that neurons fire together, these experiments demonstrate that the specific sequence of place cells that discharge as a rat navigates an environment replay later in the same or related sequences. The degree to which these sequential patterns are maintained over time (hours to days)

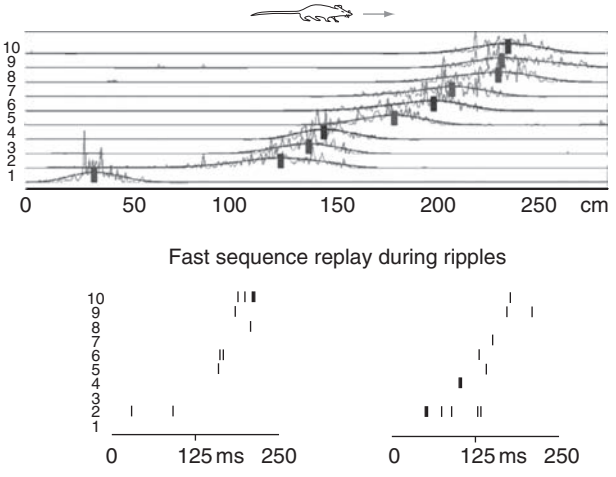

Figure 3 Temporally compressed replay of learned neuronal sequences. Top: Spatiotemporal firing rate changes of 10 CA1 place units on an elevated track. Bottom: Representative examples of firing patterns of the same units during two ripple events of nonrapid eye movement sleep. Note the similarity between ripple-related spike sequences and the order of place-cell activation on the track. Adapted from Lee AK and Wilson MA (2002) Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* 36: 1183–1194, with permission from Elsevier.

has been somewhat equivocal and may vary as a function of experience and/or novelty.

The latter can be examined by comparing firing rates and coactivation of neuronal groups in sleep episodes interrupted by performance in a well-trained task, such as food foraging in an open field and running in a wheel. The discharge frequency of concurrently recorded pyramidal neurons remains remarkably stable across several sleep/wake cycles. Neurons with high firing rates during wheel running (associated with theta) continue to fire at a high rate during drinking or awake immobility. Similarly, fast-firing neurons in SWS sustain their relatively high rates during REM sleep. When firing rates of neuronal populations are compared in two sleep episodes, interrupted by a wheel-running session or foraging in an open field, discharge frequencies of individual cells are robustly correlated. Thus, coactivation patterns during sleep episodes preceding and following an awake session are very strongly correlated. Several studies suggest that an underlying anatomical or anatomical/physiological matrix governs the propensity of neurons to fire, and the probability that they fire together is simply altered by awake experience. Any recent alterations in the probability of firing are maintained across subsequent sleep episodes.

All told, these findings demonstrate that sensory-related firing patterns during awake states are correlated with similar patterns during subsequent periods of immobility and during sleep states. Do these correlated patterns demonstrate a unique role for sleep patterns in maintaining ensemble representations of prior events? No, they do not. The correlated patterns demonstrate that neurons that fire together in a relatively stable network generally fire together when any member of the ensemble discharges. Important caveat in these studies is that causality is derived from the difference between sleep sessions before and after exploration, while many studies demonstrate considerable nonspecific difference of network activity patterns between these seemingly equivalent states. While subtle modifications, small perturbations, in the stability of a relatively hard-wired network could support mnemonic representations and be used for recognition and retrieval of prior events, the described studies do not demonstrate that they do. One possible way to attribute causality to sleep–wake correlated patterns is to manipulate or perturb the system.

Novel Experience Alters Firing Patterns in Subsequent Sleep

Can synaptic change associated with a specific experience persist in such a way that changes could be detected in subsequent sleep? Few studies have examined the role of learning on ensemble firing patterns. Exposure to a novel situation can result in long-lasting changes in the firing patterns of hippocampal pyramidal neurons. The correlated firing rates of individual neurons during exploration of a novel environment and subsequent sleep episode are stronger than are the correlated rates between exploration and the preceding sleep episode. Importantly, firing rates observed in the sleep episodes before and after exposure to novelty are poorly correlated (**Figure 4**). Exploration in a novel environment has a similar effect on correlated activity among neurons, with neuronal pairs exhibiting high awake correlations continuing to display high correlation during subsequent sleep episodes.

A more direct support for the 'replay' of learned neuronal patterns comes from experiments in which, instead of two neurons firing together, three or more neurons (up to six) exhibit sequential temporal relationships during awake and subsequent sleep. Typically neurons involved in the sequences were both identical and replayed much faster during sleep than in the waking animal. The dynamics of each ripple indicate that tens of thousands of hippocampal output neurons discharge as a population volley at 200 Hz. In summary, specific neuronal sequences during SWS are determined by prior sensory experience and by the dynamics of the SPW ripple process.

Interactions between Neocortex and Hippocampus during Sleep

Though the hippocampus is critical to the encoding of new memories, it is fairly well understood that the hippocampus plays a time-limited role in memory consolidation. After an ill-defined time period (days, weeks, and months), most memories can be retrieved in the absence of the hippocampus. Thus, a gradual translation or translocation of neural patterns to neocortical targets is presumed to mediate 'long-term' memory storage. A number of models posit that the hippocampus plays a role in bringing ensembles of cortical neurons together to form a linked 'chain' of information related to a particular memory. The role of the hippocampus is to activate each link in the chain, allowing for memory retrieval, and/or to forge the links such that the entire chain can become hippocampal independent. In the absence of hippocampal input, as occurs in human temporal lobe amnesia or experimental animals, the links in the chain cannot readily be forged or retrieved. With an intact hippocampus, the repeated replay of hippocampal patterns during offline states, during REM or SPWs in SWS, may forge the neocortical links such that memories can eventually be retrieved without the hippocampus.

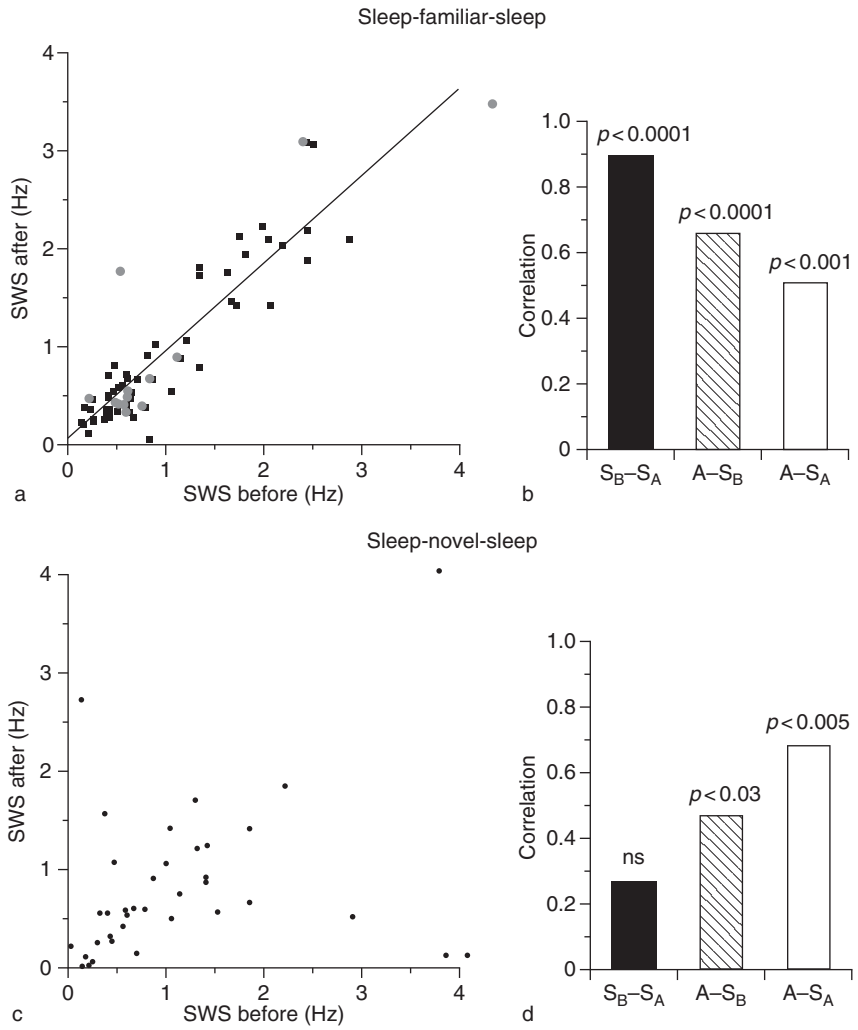

Figure 4 Comparison of discharge frequency of pyramidal cells in two successive slow-wave sleep (SWS) episodes, interrupted either (a) by a wheel running session (familiar) or (c) by novel exploration. Mean correlation values for 'sleep before' vs. 'sleep after' (S_B–S_A), 'awake' vs. 'sleep before' (A–S_B), and 'awake' vs. 'sleep after' (A–S_A) sessions in the well-trained running wheel task (b) and in the novel environment (d). Note high and low correlations between successive sleep sessions in the familiar and novel environments, respectively. Adapted from Hirase H, Leinekugel X, Czurko A, et al. (2001) Firing rates of hippocampal neurons are preserved during subsequent sleep episodes and modified by novel awake experiences. *Proceedings of the National Academy of Sciences of the United States of America* 98: 9386–9390, with permission.

The SPWs seem like an ideal candidate mechanism for the gradual offline strengthening of neocortical patterns related to memory, regardless of how the process is conceptualized.

Deep slow-wave sleep is characterized by widespread synchronized oscillatory patterns, defined primarily by spatially coherent delta waves during which spiking activity of nearly all principal cells and interneurons in the neocortex is suspended. The intervals between the silent periods define slow oscillations (0.3–1.5 Hz). The relatively abrupt shifts in population firing and silence states are reflected by the bimodal (depolarized and hyperpolarized or UP and DOWN) distribution of the membrane potential in neocortical, entorhinal, and subicular neurons. The UP state is often associated with K-complexes and thalamocortical sleep spindles (12–18 Hz).

Several studies have described a concurrence of hippocampal SPWs in association with neocortical sleep spindles and delta waves. The excitatory front of the UP cortical states can spread to the hippocampus via the entorhinal cortex, as reflected by the active sinks in the dentate molecular layer and stratum lacunosum-moleculare of the CA1 region, the main targets of the entorhinal cortex. As a result of the strong neocortical drive, dentate granule cells and other neurons of the dentate area fire coherently with the input from the entorhinal neurons during the slow oscillation. The rest of the hippocampus is less affected, mainly due to the strong feed-forward

inhibitory circuits between granule cells and CA3 pyramidal cells. Nevertheless, the majority of the output of CA1 pyramidal cells also increases their discharge rate in phase with the cortical UP state, likely driven by direct perforant path input. Important in the present context, a majority of SPW/ripple complexes are also associated with the enhanced neocortical activity during slow-wave sleep: the probability of occurrence of SPWs/ripples is highest at and after the onset of the cortical UP state. Furthermore, SPWs are modulated by single cycles by neocortical spindles (Figure 5), suggesting a more direct role for neocortical inputs in triggering SPWs.

In summary, cortical slow oscillations temporally coordinate the occurrence of K-complexes, thalamocortical spindles, and hippocampal SPWs. Such periodic 'rebooting' of neuronal activity in widespread neocortical, paleocortical, and hippocampal networks should have important functional consequences. Could specific subtle patterns of neocortical inputs 'select' the ensemble participants in each specific SPW? Is the neocortex querying the hippocampus about the relationship between ensemble members as instantiated in the neocortical input? The possible scenario of such a mechanism is as follows: After each DOWN state, the neocortex self-organizes its global activity from locally generated patterns. The DOWN–UP transition, perhaps involving thalamocortical spindle, triggers organized firing patterns of cortical neurons, which, in turn, can lead to activation of specific subpopulations of hippocampal neurons. As a result of the neocortex-biased selection, the activated hippocampal neurons can generate SPW-related synchronous output. Because the SPW is a punctuate event, whereas the UP state and sleep spindle are temporally protracted, the hippocampal output can be directed to the still-active neocortical assemblies. We hypothesize that the temporal coordination of neocortical and hippocampal strong bursting patterns favors the conditions in which the hippocampal output may assist in modifying neocortical synaptic connectivity.

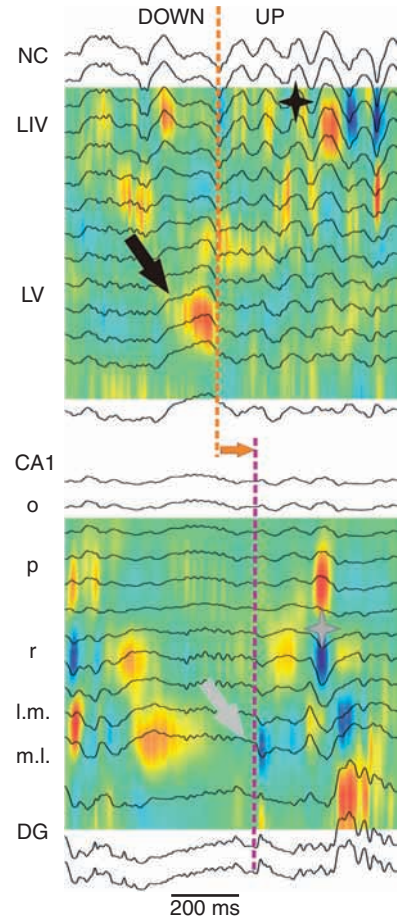

Figure 5 Spread of neocortical activity into the hippocampus during slow-wave sleep. Current-source density and superimposed local field potential traces of simultaneously recorded events in the neocortex (NC; top) and the hippocampus (bottom). A delta wave, seen as the red source in layer 5 (LV; black arrow), is followed by a sleep spindle (black star). DOWN–UP state transition is indicated by the orange dashed line. In the hippocampus, the entorhinal cortex-mediated DOWN–UP transition is reflected by a phase reversal of the local field potential and a sink (blue) in the molecular layer (white arrow). The horizontal arrow denotes the neocortico-dentate time lag (~100 ms); gray star denotes sharp wave in CA1 stratum radiatum; o, stratum oriens; p, pyramidal layer; r, radiatum; l.m., lacunosum-moleculare; m.l., molecular layer; DG, dentate gyrus. Adapted from Isomura Y, Sirota A, Ozen S, et al. (2006) Integration and segregation of activity in entorhinal-hippocampal subregions by neocortical slow oscillations. *Neuron* 52: 871–882, with permission from Elsevier.

Conclusions

The collective behavior of neurons (e.g., theta and gamma oscillations, SPWs and related 200 Hz ripples, and sleep spindles) offers a link between complex cognitive functions and behaviors governed by the brain and the biophysical, synaptic properties of individual neurons. Many of these collective behaviors can be observed in both the awake and sleeping brain. Understanding the differences between these collective patterns in the awake and sleeping brain should provide further insight into function. The current assumption about this function is that experience-initiated changes of synaptic connectivity are consolidated during sleep by way of transient, reciprocal, and strongly synchronous interactions between the neocortex and hippocampus.

See also: Hippocampus; Neocortex: Origins; Sleep Architecture; Sleep Oscillations and PGO Waves; Sleep and Sleep States: PET Activation Patterns; Sleep-Dependent Memory Processing.

Further Reading

Born J, Rasch B, and Gais S (2006) Sleep to remember. *Neuroscientist* 12(5): 410–424.

Buzsaki G (1989) Two-stage model of memory formation: a role for 'noisy' brain states. *Neuroscience* 31: 551–570.

Chrobak JJ and Buzsaki G (2005) Synaptic plasticity and self-organization in the hippocampus. *Nature Neuroscience* 11: 1560–1567.

Foster DJ and Wlson MA (2006) . *Nature* 440: 680–683.

Hirase H, Leinekugel X, Czurko A, et al. (2001) Firing rates of hippocampal neurons are preserved during subsequent sleep episodes and modified by novel awake experiences. *Proceedings of the National Academy of Sciences of the United States of America* 98: 9386–9390.

Isomura Y, Sirota A, Ozen S, et al. (2006) Integration and segregation of activity in entorhinal–hippocampal subregions by neocortical slow oscillations. *Neuron* 52: 871–882.

Kudrimoti HS, Barnes CA, and McNaughton BL (1999) Reactivation of hippocampal cell assemblies: Effects of behavioral state, experience, and EEG dynamics. *Journal of Neuroscience* 19: 4090–4101.

Lee AK and Wilson MA (2002) Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* 36: 1183–1194.

Nadasdy Z, Hirase H, Czurko A, et al. (1999) Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience* 19: 9497–9507.

Peigneux P, Laureys S, Fuchs S, et al. (2004) Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron* 44: 535–545.

Sejnowski TJ and Destexhe A (2000) Why do we sleep? *Brain Research* 886(1–2): 208–223.

Siapas AG and Wilson MA (1998) Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron* 21: 1123–1128.

Sirota A, Csicsvari J, Buhl D, et al. (2002) Communication between neocortex and hippocampus during sleep in rodents. *Proceedings of the National Academy of Sciences of the United States of America* 100: 2065–2069.

Steriade M and Timofeev I (2003) Neuronal plasticity in thalamocortical networks during sleep and waking oscillations. *Neuron* 37: 563–576.

Tononi G and Cirelli C (2006) Sleep function and synaptic homeostasis. *Sleep Medicine Reviews* 10(1): 49–62.

Wilson MA and McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* 265(5172): 676–679.

Sleep and Sleep States: Histamine Role

H L Haas, O Selbach, and O A Sergeeva, Heinrich-Heine-Universität, Düsseldorf, Germany

© 2009 Elsevier Ltd. All rights reserved.

Overview

Histamine is a signaling molecule in the immune system, the skin, the stomach, and the brain of vertebrates. Histaminergic neurons are located exclusively in the posterior hypothalamus and extend their axons throughout the central nervous system (CNS) (Figure 1). They are active during the awake state but not during sleep. Three of the four known histamine receptors are widely expressed in the brain. Close mutual interactions between histaminergic and other aminergic (acetylcholine, norepinephrine, dopamine, serotonin) and peptidergic (orexins/hypocretins) systems form a network with basic homeostatic functions that are closely connected with the sleep–wake regulation. Because the fluorescence histochemistry techniques that have delivered complete pictures of the other aminergic systems for decades fail to make the histaminergic system visible, histamine has been a relatively neglected transmitter. Only after the discovery of antibodies against histamine and histidine decarboxylase did the existence of histamine gain general acceptance. Histaminergic neurons are now recognized as the major regulator of the waking center of the brain.

Histamine Metabolism

Histidine is taken up by histaminergic neurons through the L-amino acid transporter and is decarboxylated via the specific histidine decarboxylase to form histamine. This synthesis can be inhibited by α-fluoromethylhistidine (α-FMH) – a useful tool to study the histaminergic system functions. Histamine is carried by the vesicular monoamine transporter VMAT-2 into vesicles, from which it can be released upon arrival of action potentials. Inactivation in the extracellular space is achieved by methylation via histamine methyltransferase. Tele-methylhistamine undergoes oxidative deamination to tele-methylimidazoleacetic acid. Diamine oxidase leads directly to the production of imidazoleacetic acid, mainly in invertebrates; this effective γ-aminobutyric acid ($GABA_A$) receptor agonist does not seem to occur to a significant degree in mammalian tissues. The level of histamine in brain tissue is somewhat lower than that of other biogenic amines but its turnover is considerably faster (on the order of minutes) and varies with functional state. The synthesis of histamine is controlled by feedback through H3 autoreceptors.

Morphology and Location of Histaminergic Neurons

The histaminergic neurons, consisting of five subgroups, are located in the tuberomamillary nucleus. Many of the neurons lie close to the outer and inner (third ventricle) surfaces of the posterior hypothalamus. They are relatively large cells, with typically two or three thick major dendrites and an axon emanating from each cell at considerable distance from the soma. The dendrites contain vesicles with histamine, which is released to act as a feedback signal, reducing the firing of action potentials and the synthesis of histamine at the soma level. Histamine synthesis in and release from histaminergic endings, varicosit0ies on the multifold branching axons, is also reduced. These varicosities have been seen to form synaptic contacts with postsynaptic membranes, but a vast majority of the endings are located at some distance from the target membranes.

Several species have variable numbers of brain mast cells which contain and release histamine and can contribute substantially to the histamine content. A massive invasion of mast cells into the brain has been reported in doves during courting.

Physiology of Histaminergic Neurons

The histaminergic neurons display a rather regular firing pattern below 4 Hz and their membrane potential dwells at about −50 mV during waking (Figure 2). Their action potential displays a clear Ca^{2+} component, and two major opposing currents give the responses to current injection a very typical appearance, with a strong hyperpolarization-dependent inward rectification and marked outward rectification upon return to the membrane potential from more negative voltages. These conductances carry currents of the I_h- and the A-type, respectively. Ca^{2+}-dependent prepotentials in the dendrites and a tetrodotoxin-sensitive, noninactivating Na^+ current, together with a postspike K^+ current, create a pacemaker cycle. The actual firing rate is variable during waking and is dependent on behavioral state. A minimal polarization suffices to stop the firing as it occurs during sleep. This inhibition is attributed to the (GABA)ergic input mainly from the ventrolateral preoptic (VLPO) area, a known sleep active nucleus.

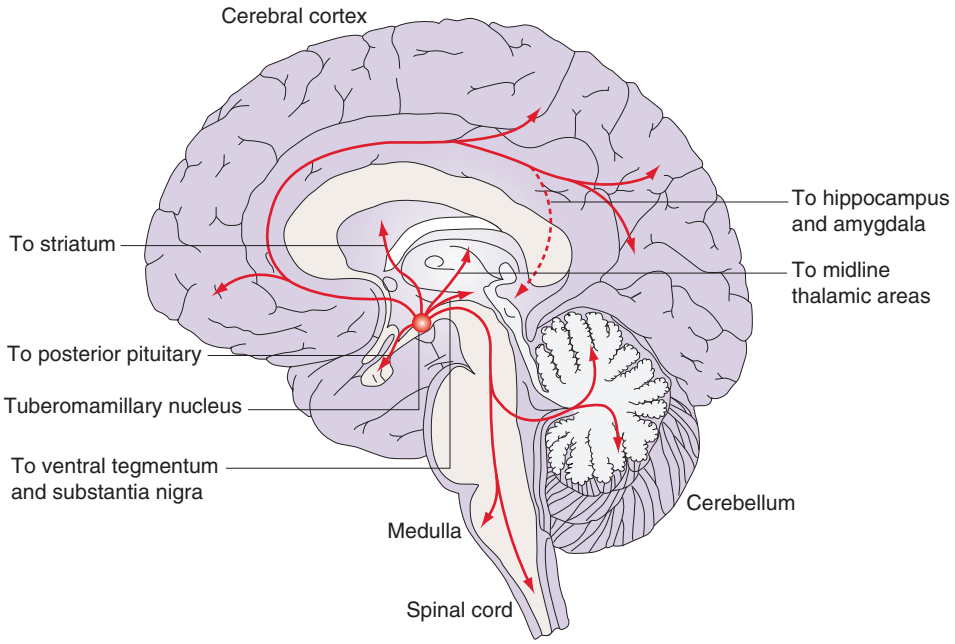

Figure 1 The histaminergic system in the human brain: origin and projections. Reproduced with permission from *Nature Reviews Neuroscience* 4: 121–130, Haas HL and Panula P, The role of histamine and the tuberomamillary nucleus in the nervous system, copyright 2003, Macmillan Magazines Ltd.

Local injections of GABA agonists in the tuberomamillary nucleus (TMN) cause sedation while GABA antagonists have the opposite effect.

The histaminergic neurons contain further transmitters with unknown specific functions. GABA and glutamic acid decarboxylase (GAD) are expressed in most TMN neurons, but so far no evidence for GABA release from the somata or the axon varicosities has been presented. Should GABA be released from a substantial fraction of such a projecting system the physiological impact would be expected to be significant and possibly opposite to the normal histamine release. Similarly, the inhibitory neuropeptide galanin is found in the TMN neurons and in the GABAergic afferents to them. Furthermore, enkephalins, thyrotropin-releasing hormone, and substance P are present in populations of the TMN neurons. The histamine neurons display a high level of expression of adenosine deaminase, which may be involved in a number of basic cellular functions.

Inputs to Histaminergic Neurons

Amino Acids

Sources for the inputs to histaminergic neurons are widespread: for example, glutamatergic fibers from the cortex and the hypothalamus are a source of glutamate, which excites histaminergic neurons, and the fibers carry both α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and *N*-methyl-D-aspartate (NMDA) receptors. Glycine inhibits a subpopulation of histaminergic neurons, but the role of glycinergic fibers is uncertain. Taurine, an osmolyte that can reach relevant concentrations in the extracellular space, gates strychnine-sensitive glycine receptors and $GABA_A$ receptors.

GABAergic inputs come from several mostly hypothalamic locations; functionally most prominent with respect to sleep–waking regulation is the innervation from the VLPO area, which fires high during sleep and thus suppresses the firing of histamine neurons. $GABA_A$ receptors are quite heterogeneous among histamine neurons; three groups with different GABA sensitivity have been identified, depending on the expression of the γ subunit of the ionotropic GABA receptor. The sedative component of general anesthetic effects (e.g., propofol) is attributed to an action on the TMN, with a key to this action being the low expression of the $GABA_A$ receptor ε subunit. Cessation of histaminergic neuron firing is associated with the loss of consciousness. The GABAergic inputs to the TMN are under feedback control of $GABA_B$ receptors.

Amines

All aminergic nuclei give rise to projections and they are all functionally excitatory, but use a variety of mechanisms. A nicotinic fast-desensitizing action occurs through α7-type receptors, but muscarinic actions have not been detected. Noradrenaline does not affect histaminergic neurons directly but

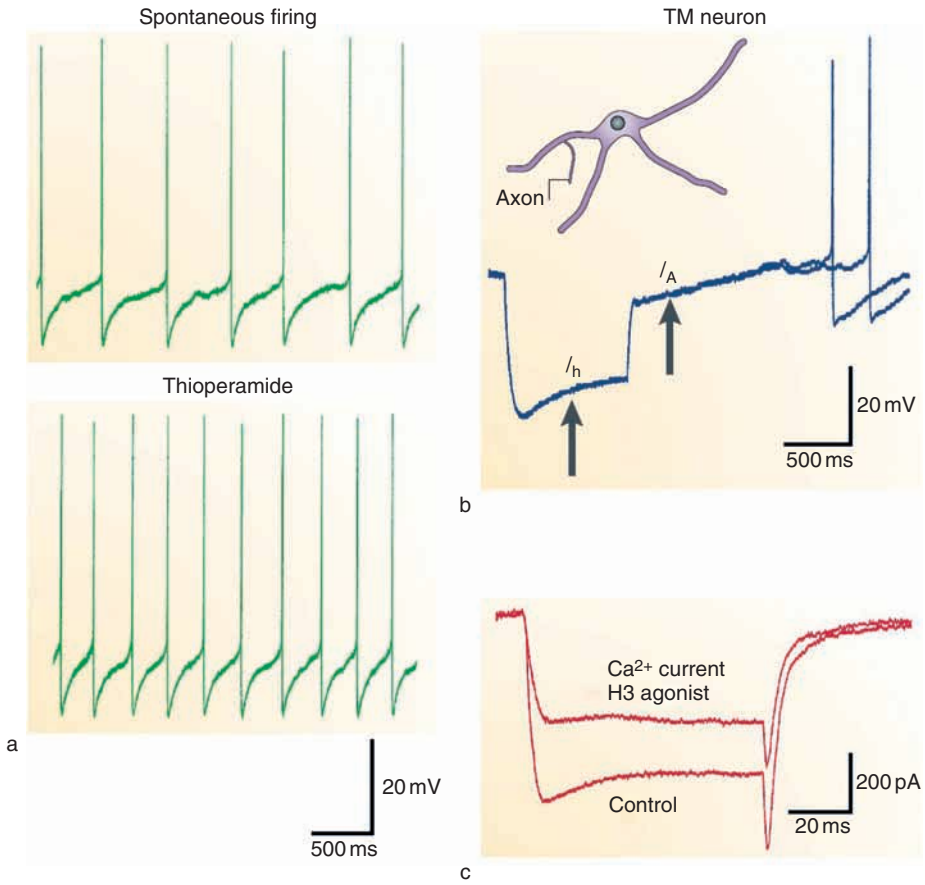

Figure 2 Properties of histaminergic neurons. (a) Spontaneous firing of a tuberomamillary neuron recorded *in vitro* at a membrane potential of −50 mV. Under superfusion with a histamine H3-receptor inverse agonist (thioperamide) the firing is markedly enhanced, indicating a tonic inhibition of the neuron through its autoreceptors. (b) Camera lucida drawing and recordings from a TM neuron. (Top) Response to negative current injection: a hyperpolarization-activated inward rectification (due to the current Ih, HCN1, and HCN3) and an outward rectification that delays the return to the membrane potential (due to A-type currents). (c) Voltage clamp recording (switch clamp) showing an inward Ca^{2+} current that is markedly reduced by a histamine H3-receptor agonist (R-alpha-methylhistamine). Reproduced with permission from *Nature Reviews Neuroscience* 4: 121–130, Haas HL and Panula P, The role of histamine and the tuberomamillary nucleus in the nervous system, copyright 2003, Macmillan Magazines Ltd.

effectively controls GABAergic input through $\alpha 2$ receptors, mediating an inhibition of inhibitory postsynaptic currents (IPSCs). Dopamine excites histamine neurons through a yet undefined indirect action, whereas serotonin excites the histaminergic neurons through activation of Na^+–Ca^{2+} exchange (NCX). This electrogenic transporter has to let three Na^+ ions enter the cell to expel one Ca^{2+} ion; resulting in a depolarization and excitation in the absence of any conductance change.

Peptides

Galanin is expressed in histaminergic neurons and in the GABAergic inputs to them and it inhibits their firing. It is thus possible that galanin participates in both the autogenic (feedback) inhibition and in the extrinsic inhibition from the VLPO area. Neurons containing orexins (also called hypocretins) are neighbors of the histamine neurons; the nuclei intermingle partially. Both types of orexins (hypocretin-1 and -2) excite histamine neurons through the Hcrt2 receptor and largely via activation of NCX. This action is secondary to a rise in intracellular Ca^{2+} that probably depends on both extra- and intracellular sources of the Ca^{2+}. Hypocretin neurons also express dynorphin, which can contribute to the excitation of histaminergic neurons, as it suppresses GABAergic input. Leptin, the hormone from fat that controls food intake, has no obvious effect, but the appetite-stimulating ghrelin inhibits a potassium channel (Kir3). Nociceptin strongly inhibits TMN neurons at the postsynaptic level while morphine (μ receptors) excites them, probably through disinhibition.

Purines

The tuberomamillary nucleus displays a very strong expression of adenosine deaminase, which has led to the suggestion that it may also use adenosine as a

transmitter. So far such a role of adenosine is elusive, and there is no evidence for synaptic release of this nucleoside but it is sedative through A1 receptors. A2A adenosine receptors have also been implicated in sleep regulation, through enhancement of the inhibition of histamine neurons. Adenosine accumulates during wakefulness and is likely an endogenous sleep pressure substance. Interestingly, adenosine, which strongly inhibits many neurons in the brain, has no such effect on histamine neurons. In contrast, ATP excites histamine neurons through P_2Y_1 and P_2Y_4 receptors.

Histaminergic Pathways

Histaminergic fibers innervate virtually all parts of the CNS, from the spinal cord to the retina. A dense innervation is found in the hypothalamus, cortex, amygdala, and basal ganglia and more variably in the thalamus and hippocampus. The mutual innervation of aminergic systems forms a network that acts in concert to ensure vegetative, endocrine, metabolic, osmotic, and energetic homeostasis.

The waking action of histamine is attributed to projections to thalamus and cortex, as well as to other aminergic nuclei and their emanating axons. Relay cells in the thalamus are depolarized and thus are shifted to the regular firing mode, which represents an 'open door' for perception. In the absence of histaminergic tone these cells return to the burst firing mode, which closes the door to the cortex for sensory input. Most targets are depolarized and excited through a number of different mechanisms; it is only in the hippocampus that pyramidal cells are hyperpolarized and inhibited through intracellular Ca^{2+} increase, which activates a K^+ conductance. On the other hand, histamine can block Ca^{2+}-dependent K^+ conductance, and thus accommodate firing through another receptor (H2). Some hippocampal interneurons are rather strongly excited by H2 receptor activation, making the overall histamine action in this structure an inhibitory one.

Histamine Receptors

Histamine can activate Cl^- conductances in thalamus and hypothalamus: in the supraoptic nucleus this effect is blocked by picrotoxin and H2 receptor antagonists. In thalamic interneurons Cl^- conductance blocking is also H2 receptor mediated but is not picrotoxin sensitive. Histaminergic ionotropic receptors may play only a minor role, if any, in the vertebrate brain. In contrast, arthropod sensory neurons, the insect eye photoreceptor, and lobster olfactory neurons display synaptic transmission through histamine-gated Cl^- channels. Four metabotropic histamine receptors have been cloned; H1, H2, and H3 occur in the brain and H_4 receptors are mainly expressed in peripheral tissues such as blood, spleen, lung, and liver.

H1 Receptors

Histamine H1 receptors occur throughout the central nervous system, with a particularly high density in regions involved in arousal and waking, such as the thalamus and cortex and the cholinergic, noradrenergic, dopaminergic, and serotonergic nuclei (Figure 3). H1 receptor activation causes excitation in most brain regions (brain stem, thalamus, hypothalamus, cortex, amygdala, striatum) through G_{q11} protein and direct block of leak K^+ conductance or phospholipase C, inositol trisphosphate (IP_3), and diacylglycerol (DAG) mediation. IP_3 releases Ca^{2+} from internal stores and activates a number of Ca^{2+}-dependent processes, including the opening of a cation channel of the transient receptor potential canonical (TRPC) type or the stimulation of a Na^+–Ca^{2+}-exchanger. Furthermore, the elevated intracellular Ca^{2+} can stimulate NO synthase and consequently guanylate cyclase. On the other hand, Ca^{2+}-dependent K^+ channels can be opened, leading to hyperpolarization and inhibition, for instance, in hippocampal pyramidal neurons (Figures 4).

H1 receptor antagonists are classic antihistaminics and are widely prescribed for allergies. Their well-known sedative actions have prompted early suggestions for an involvement of endogenous

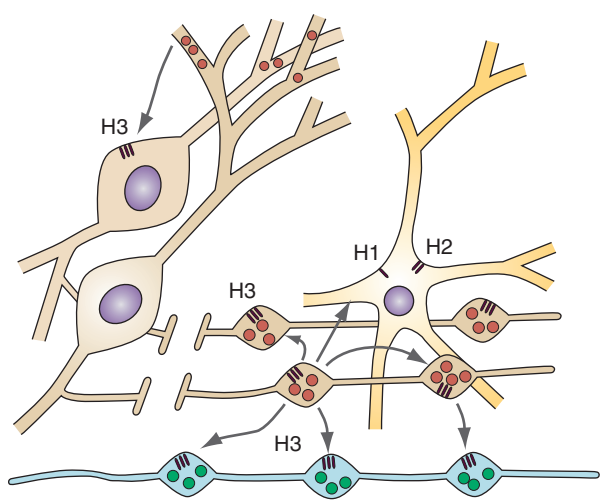

Figure 3 Cellular location of histamine receptors in the brain. TM neurons (left) with H3 receptors on their somata and their axons, as well as on the axons of other cells (containing glutamate or other amines). H1 and H2 receptors on a target cell body. Reproduced with permission from *Nature Reviews Neuroscience* 4: 121–130, Haas HL and Panula P, The role of histamine and the tuberomamillary nucleus in the nervous system, copyright 2003, Macmillan Magazines Ltd.

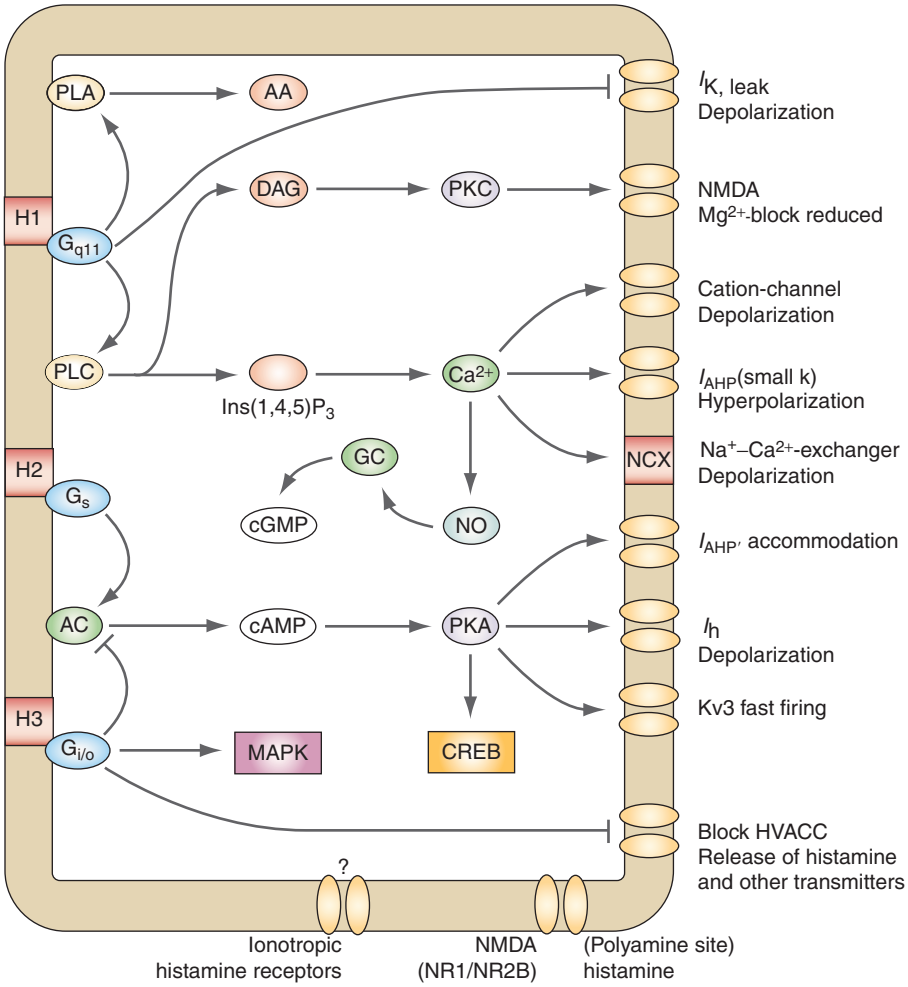

Figure 4 Signaling pathways activated by histamine receptors. cGMP, cyclic guanosine monophosphate; HVACC, high-voltage activated Ca^{2+} current; MAPK, mitogen-activated protein kinase; PKA, protein kinase A; PKC, protein kinase C. Reproduced with permission from *Nature Reviews Neuroscience* 4: 121–130, Haas HL and Panula P, The role of histamine and the tuberomamillary nucleus in the nervous system, copyright 2003, Macmillan Magazines Ltd.

histamine in sleep–waking regulation. There is strong evidence for H1 receptors being the main mediators of histamine waking action. In H1 receptor knockout (KO) mice, however, the sleep–waking pattern is hardly changed, but the waking response to H3 receptor antagonists, which relieve the autoinhibition of histamine release, is abolished. Most antipsychotic and antidepressive pharmaceuticals occupy H1 receptors, but this is not part of the therapeutic effect. Genetic and molecular studies have recently attributed an important role in peripheral and central immunity to histamine receptors. The H1 receptor is identical to an autoimmune locus receptor (*Bordetella pertussis*-induced histamine sensitization; Bphs), with relevance to inflammatory brain diseases (e.g., multiple sclerosis). Through an action on astrocytes, histamine regulates the permeability of the blood–brain barrier and the immigration of immune cells in the otherwise privileged CNS. Histidine decarboxylase (HDC) KO mice, compared to wild type, produce more proinflammatory signals and display a particularly marked experimental autoimmune encephalitis.

H2 Receptors

Histamine from enterochromaffin-like cells in the stomach stimulates acid secretion by parietal cells, an action that is insensitive to the classic antihistamines (H1). Development of H2 receptor antagonists by Jim Black revolutionized the therapy of stomach ulcers. H2 receptors are also widespread in the brain, in particular in the basal ganglia, the hippocampus, and the amygdala. Activation of histamine (H2) as well as some β-adreno-, dopamine, serotonin, corticotropin-releasing hormone (CRH), and vasoactive intestinal peptide (VIP) receptors has similar actions on neuronal excitability: blockage of a Ca^{2+}-activated K^+ conductance (small K), which is responsible for the accommodation of

firing and a long-lasting (seconds) afterhyperpolarization following action potentials (Figure 5). The accommodation of firing determines the number of action potentials in response to a longer (~100 ms) depolarizing stimulus – that is, the number of action potentials in a burst, and the amount of Ca^{2+} inflow through high-voltage activated Ca^{2+} channels (HVACCs) or NMDA channels. This sets up a neuron state of readiness to respond. These actions are mediated by cyclic AMP and protein kinase A (PKA) and lead to phosphorylation of the transcription factor cAMP response element-binding protein (CREB; Figure 4). Cyclic AMP can also interact directly with another current, the hyperpolarization-activated cation current (I_h), by shifting its activation curve to more positive levels, leading to depolarization in thalamic and hippocampal neurons. Brief exposure of hippocampal slices to histamine, noradrenaline, or cyclic AMP causes long-term enhancements of CA1 population spikes and pyramidal neuron firing in the absence of high-frequency stimulation. Adenosine A1 receptors are negatively coupled to adenylyl cyclase and have effects opposite to those of histamine H2 receptor activation (i.e., increasing the long-lasting afterhyperpolarization and accommodation of firing). Adenosine has a sedative action (in contrast to the waking action of histamine) and is a major factor in sleep induction.

H3 Receptors

The histamine autoreceptor was termed H3 by the Paris group of J-C Schwartz, who first described it in 1983. H3 receptors are present on TMN somata, dendrites, and axons (Figure 3). They are coupled directly through a $G_{i/o}$ protein to Ca^{2+} channels, causing inhibition of TMN firing, histamine release from varicosities, and histamine synthesis. Adenylyl cyclase is depressed by H3 receptor activation. Importantly H3 receptors are found on many nonhistaminergic varicosities, such as those containing glutamate, acetylcholine (ACh), dopamine, noradrenaline, serotonin, GABA, and various peptides. Excitatory transmission in the glutamatergic perforant pathway from the entorhinal cortex to the dentate area is depressed by about 30% by histamine in rats, but not in mice. The glutamatergic corticostriatal pathway and the dopaminergic nigrostriatal pathway are also suppressed. In the substantia nigra, GABAergic potentials are reduced.

In contrast to the H1 and H2 receptor genes the H3 receptor gene features two or three introns. Alternative splicing yields a large number of H3 receptor isoforms with differential distribution and constitutive activity. The H3 receptors are thus highly promising targets for the development of selective

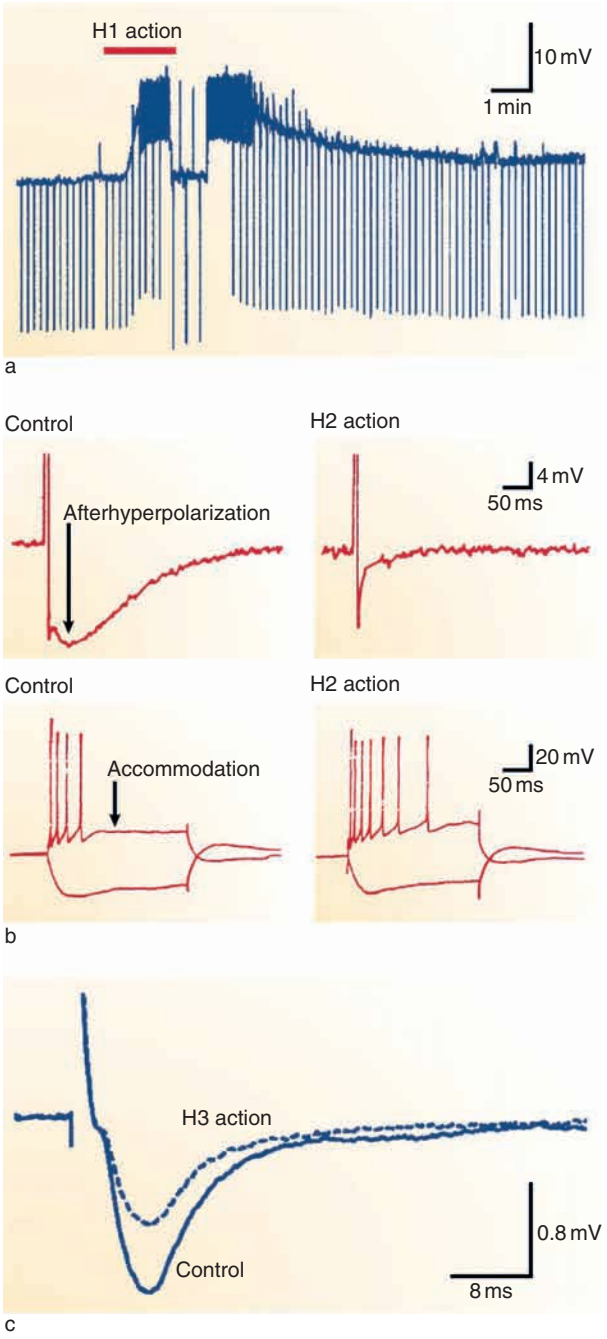

Figure 5 Electrophysiology of histamine actions. (a) H1-receptor: a depolarization of a pontine reticular formation neuron through block of a potassium conductance. (b) H2-receptor: block of the long-lasting afterhyperpolarization after Ca^{2+} inflow (Ca^{2+}-dependent potassium conductance, sK-channel) and block of the accommodation of action potential firing in a human hippocampal pyramidal cell *in vitro*. (c) H3-receptor: decrease of field excitatory postsynaptic potential (EPSP) in the dentate gyrus of the hippocampus evoked by perforant path stimulation. Histamine reduced release of the excitatory transmitter glutamate in these two pathways. Reproduced with permission from *Nature Reviews Neuroscience* 4: 121–130, Haas HL and Panula P, The role of histamine and the tuberomamillary nucleus in the nervous system, copyright 2003, Macmillan Magazines Ltd.

pharmaceuticals to treat neuropsychiatric disorders affecting sleep–waking regulation, feeding, mood, learning, and memory, as well as epilepsy, migraine, and inflammatory disorders. H3 receptor antagonists have, indeed, facilitatory actions on learning and memory, presumably in part by removal of a tonic inhibition of the release of other transmitters, such as glutamate and acetylcholine. In the light of the many possible actions on H3 heteroreceptors, the systemic effects of interference are difficult to predict.

Histamine and Synaptic Plasticity

Histamine (and hence the activity of the histaminergic neurons) can cause long-lasting changes in synaptic transmission in several structures by itself, and it can prime or modulate other forms of synaptic plasticity such as long-term potentiation (LTP) and long-term depression (LTD) (Figure 6). These phenomena represent coincidence detection and are associated with learning and memory. An important signal transduction pathway is the transient rise of intracellular Ca^{2+} and its binding to a Ca^{2+}/calmodulin kinase (CaMK).

Interference of the polyamine spermidine with early attempts to stain histamine neurons came to mind when direct interaction of the diamine histamine with the NMDA receptor on hippocampal pyramidal cells was first detected: this effect is similar to the potentiation by spermidine of the NMDA receptor (NR2B subunit) and opposite to the inhibitory action of protons. It is facilitated in slightly acidic conditions, such as occur in the CA3 region during synchronous discharges, with 'sharp waves' that evoke specifically distributed LTP in CA1 and that are likely a decisive element in memory trace formation. Histamine strongly promotes burst firing in the CA3 area, frequently leading to 200 Hz oscillations (ripples) during quiet waking or sleep. The replay of ripples during sleep has been proposed as critical for memory consolidation.

H1 receptor activation increases intracellular Ca^{2+} and potentiates the H2 effects that are proved proplasticity signals. H2 receptor activation stimulates adenylyl cyclase and controls a Ca^{2+}-dependent potassium current through CaM kinase II and the extent of firing in response to an excitatory stimulus. These mechanisms are deeply involved with triggering long-term potentiation and the cellular basis of learning and memory. Brief exposure of hippocampal slices to histamine causes LTP in the absence of any electrical stimulation. Adenosine A1 receptors are negatively coupled to adenylyl cyclase and have effects opposite to those of histamine on LTP. The histamine H3 receptor is an auto- and heteroreceptor with inhibitory action. The glutamatergic corticostriatal pathway undergoes long-term depression after H3 receptor stimulation.

Memory trace formation occurs in the hippocampus where a selected population of CA3 pyramidal cells produces a synchronous burst that is relayed to the CA1 area and acts as the natural trigger for long-term potentiation. Whether, and at which intensity, this burst is fired depends on the subcortical control by largely aminergic systems. Histamine interacts at all levels of this process.

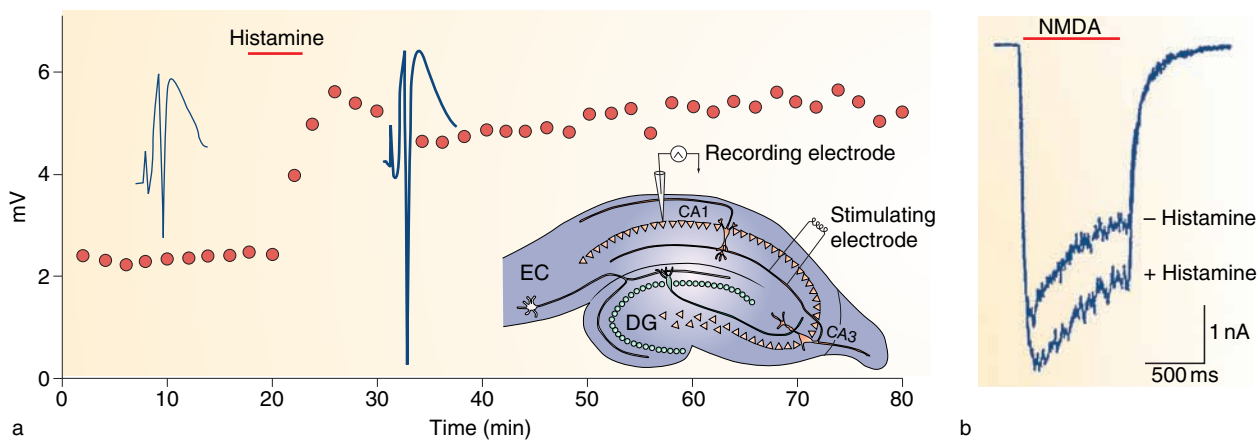

Figure 6 Histamine and synaptic plasticity. (a) Histamine (H2-receptor activation) induces LTP of population spike in CA1 (Schaffer-collateral stimulation) in hippocampal slice. Inserted are population spikes at the times where they are shown and a slice. EC, entorhinal cortex; CA, cornu ammonis; DG, dentate gyrus. (b) Whole cell recording from hippocampal pyramidal cell. Neuron exposed to NMDA for 1 s displays inward current, which is larger in the presence of histamine. Reproduced with permission from *Nature Reviews Neuroscience* 4: 121–130, Haas HL and Panula P, The role of histamine and the tuberomamillary nucleus in the nervous system, copyright 2003, Macmillan Magazines Ltd.

Histamine and Behavior

Histaminergic systems are phylogenetically old and are concerned with basic body functions of organisms in both the invertebrate and the vertebrate world. They are found in vertebrates in the hypothalamus, where feeding, drinking, osmoregulation, temperature control, energy administration, and the endocrinium are integrated with the sleep–waking and circadian rhythms. Histamine neurons fire exclusively during waking and particularly during attention. Excitation of the large cells in the supraoptic nucleus leads to antidiuresis. Injection of histamine into other hypothalamic sites induces drinking, and dehydration increases synthesis and release of histamine. Stimulation of oxytocin-containing cells and those releasing the prolactin-releasing hormone have far-reaching influences on mammalian reproduction. The suprachiasmatic nucleus, our circadian zeitgeber, receives strong histaminergic excitation. Locomotion is maintained by histaminergic activation and serves the basic function of feeding. $HDC^{-/-}$ mice display reduced exploratory activity, but normal habituation to novel environments, increased fear of heights, and superior motor coordination. H3 $receptor^{-/-}$ mice are lazy walkers, but H3 antagonists can cause hyperactivity. H1 $receptor^{-/-}$ mice display reduced ambulatory and exploratory activity and a reduced sensitivity to pain. Histamine activity on H2 receptors is an essential link in opioid receptor-mediated analgesia. In the periphery, histamine participates in the stimulation of nociceptors and itching, and the central histamine system is involved through both H1 and H2 receptors in pain perception.

Leptin-induced anorexia is mediated by the histaminergic system. Glucose-responsive neurons in the ventromedial nucleus are excited by histamine. In the anterior hypothalamus, H1 receptor activation lowers the set point of the hypothalamic thermostat; in the posterior hypothalamus, H2 receptors are involved with initiating heat loss mechanisms. The vegetative nervous system is under the control of the hypothalamus, with a significant contribution by histamine; this also includes cardiovascular control.

Control of Waking and Arousal

Sleep and awake states are closely correlated with essential body functions. Histaminergic neurons interact with the neighboring orexin/hypocretin-containing neurons to determine the right time for sleeping and waking. Loss-of-function of orexin/hypocretin neurons results in narcolepsy. Cataplexy, the major symptom of narcolepsy, is characterized by loss of muscle tone in the presence of consciousness. In contrast to other aminergic systems, histamine neurons are fully active during cataplexy. Thus histamine is mainly responsible for consciousness and noradrenaline is mainly responsible for muscle tone.

Destruction or inhibition of the tuberomamillary region causes somnolence, whereas activation evokes arousal. During the devastating influenza epidemic of 1918–1919, a significant number of patients suffered from sleep disturbances. Von Economo in Vienna investigated the brains of these patients and found posterior hypothalamic lesions in those who had suffered from hypersomnia (encephalitis lethargica) and anterior hypothalamic lesions in those who had suffered from insomnia. It is now clear that the first group must have lost their histamine neurons and probably the orexin/hypocretin neurons, whereas the insomnia patients displayed lesions in an area now known to be exceptionally active during slow-wave sleep (the VLPO area). The VLPO houses GABAergic neurons that project to wake-active aminergic neurons, specifically the histaminergic neurons, the firing of which is suppressed by the input. $GABA_A$ receptors display a large variability due to differential assembly of subunits in a receptor-channel complex. Depending on γ subunit expression, populations with different sensitivities to GABA have been identified – presumably reflecting an involvement in different aspects of the waking function.

The TMN is a major target of general anesthesia because it mediates the sedative component. $GABA_A$ receptor stimulation (local muscimol) is sedative whereas the local injection of a $GABA_A$ antagonist causes arousal. Systemic administration of general anesthetics that potentiate $GABA_A$ action increases activity of the sleep-active preoptic area and reduces activity of the TMN. Low-level expression of the $GABA_A$ receptor ε subunit in most TMN neurons (in contrast to the neighboring orexin/hypocretin neurons) may be key to the high sensitivity of the TMN to anesthetics.

Much of the more recent knowledge in this field comes from the rare, technically difficult recordings from identified histaminergic neurons over several hours in freely behaving animals, but results from experiments using local excitation, inhibition, or destruction all point in the same direction. In combination, measures of histamine release by microdialysis or c-fos activation and the selective, complete or reversible removal of the histamine system or its receptors have also given congruent evidence.

Hibernation, an adaptation to low external temperature and the lack of food, is a state of radically reduced physiological functions – brain activity, body temperature, heart rate, and metabolism – that highlights the link between feeding and sleeping. Interestingly, the turnover of histamine, the waking transmitter, is considerably enhanced during

hibernation. It seems to play a major role in regulating maintenance of and arousal from the hibernation state.

Learning and Memory

Correlation of synaptic plasticity with memory mechanisms is often claimed but the evidence is rarely convincing. The effects of histamine H1 and H2 receptor activation on principal cells in hippocampus, cortex, amygdala, or striatum are quite compatible with a positive influence on synaptic plasticity and memory. The excitatory action and the intracellular increase of Ca^{2+} ions (H1) are important mechanisms for maintaining wakefulness and attention and are a prerequisite for effective learning. Block of the accommodation of firing (H2) shifts target neurons toward a state of quiet readiness, characterized by little difference in membrane potential and spontaneous action potential firing, but a dramatically changed response to longer lasting (<100 ms) depolarizing signals. This is a response to compartmentalized Ca^{2+} signals. Classical conditioning (eye-blink reflex in rabbits) can reduce the amplitude and duration of the Ca^{2+}-dependent afterhyperpolarization in hippocampal pyramidal cells. These mechanisms are also of interest with respect to pathophysiological events leading to exaggerated activity and epilepsy. Contrary to expectation, histamine H1 antagonists (the classical antihistamines) are slightly epileptogenic. Two clues can be offered to explain this: (1) H1 receptor activation leads to hyperpolarization and inhibition of principal neuron firing in the (rodent) hippocampus through activation of Ca^{2+}-dependent K^+ channel opening; (2) interneurons in the hippocampus are strongly excited by H2 receptor activation. The same interneurons suffer a cutoff at high-frequency firing. Thus, the final action on a whole structure and its output cannot be predicted by studying principal cells alone and *in vitro*. H2 and H3 receptor blocking has revealed ambiguous results on learning paradigms. The histaminergic system can affect memory through at least two independent mechanisms, directly on the synaptic plasticity in hippocampus, striatum, and amygdala, and by an indirect inhibitory effect through the brain's reinforcement system. H1 receptor antagonists can act as reinforcers, alone or together with opiates, cocaine, or amphetamine.

H1 antagonists improve water maze performance but reduce radial maze performance. Locomotor activity and exploratory behavior are reduced, but learning and memory seem unaffected in H1 receptor knockout mice. H3 receptor$^{-/-}$ mice display enhanced spatial learning and reduced anxiety. Lesions of the histaminergic nucleus can lead to facilitation of learning, and pharmacological inactivation by posttrial injection of lidocaine leads to memory improvement. Such lesions or stimulations of the tuberomamillary nucleus are very difficult or impossible to perform without affecting neighboring structures or bypassing fibers.

Histidine decarboxylase KO mice cannot synthesize histamine. JS Lin has described an important defect in HDC$^{-/-}$ mice that is quite relevant for memory formation: their response to novelty is strongly impaired and they fall asleep rather than explore a new environment. JP Huston's group revealed a number of rather specific changes in HDC$^{-/-}$ mice: an improvement in negatively reinforced water maze performance but a defect in (nonreinforced) episodic object memory.

Histamine Pathophysiology

Hepatic encephalopathy (HE) develops as a result of high ammonia levels. Histidine and some other amino acids are significantly increased in the brains of HE patients. This leads to a higher synthesis of histamine, which does not cause alertness in these patients; instead, they suffer from sleepiness. Histamine turnover is not increased. Model rats with a portacaval anastomosis display defects in histamine-induced plasticity and are sleepy and uninterested in their environment; their histamine receptors may be desensitized or incorporated in the cells. In contrast to low levels of histamine, high levels inactivate NMDA receptors – in keeping with cognitive deficits.

Within the hypothalamus only the TMN neurons show tangles and degeneration in Alzheimer's disease and Down syndrome. In Parkinson's disease, histaminergic endings seem to take over the empty places left by degenerated dopaminergic axon varicosities. Histamine turnover is increased and H1 receptor binding is low in the frontal cortex of schizophrenic patients. H2 receptor antagonism can reduce negative symptoms in schizophrenia. Histamine receptor expression is altered in rats with ethanol preference. Finally, the histaminergic system has an antiepileptic influence; antihistamines (H1) can cause seizures and H3 receptor antagonists are anticonvulsive.

Conclusion

The brain histaminergic system is our major waking center and is deeply involved in many basic functions. It keeps the CNS alert and ready to react and is the gatekeeper of consciousness. H1 and H2 receptors mediate excitation and (long-term) potentiation of excitation, while the H3 autoreceptors provide a feedback control of synthesis, release, and electrical activity. As heteroreceptors, these receptors also control the varicosities of most other transmitter systems,

making them a prime target for pharmaceutical research and development.

See also: Circadian Genes and the Sleep–Wake Cycle; Circadian Rhythms in Sleepiness, Alertness, and Performance; Histamine Receptors and their Ligands: Mechanisms and Applications; Sleep: Development and Circadian Control; Synaptic Plasticity: Short-Term Mechanisms.

Further Reading

Black JW, Duncan WA, Durant CJ, et al. (1972) Definition and antagonism of histamine H2-receptors. *Nature* 236: 385–390.

Brown RE, Stevens DR, and Haas HL (2001) The physiology of brain histamine. *Progress in Neurobiology* 63: 637–672.

Haas H and Panula P (2003) The role of histamine and the tuberomamillary nucleus in the nervous system. *Nature Reviews Neuroscience* 4: 121–130.

Haas HL and Reiner PB (1988) Membrane properties of histaminergic tuberomamillary neurones of the rat hypothalamus *in vitro*. *Journal of Physiology (London)* 399: 633–646.

Hill SJ, Ganellin CR, Timmerman H, et al. (1997) International Union of Pharmacology. XIII. Classification of histamine receptors. *Pharmacological Reviews* 49(3): 253–278.

Hough LB, Nalwalk JW, Li BY, et al. (1997) Novel qualitative structure–activity relationships for the antinociceptive actions of H2 antagonists, H3 antagonists and derivatives. *Journal of Pharmacology & Experimental Therapeutics* 283(3): 1534–1543.

John J, Wu MF, Boehmer LN, et al. (2004) Cataplexy-active neurons in the hypothalamus: Implications for the role of histamine in sleep and waking behavior. *Neuron* 42: 619–634.

Leurs R, Bakker RA, Timmerman H, et al. (2005) The histamine H3 receptor: From gene cloning to H3 receptor drugs. *Nature Reviews Drug Discovery* 4: 107–120.

Lin JS, Sakai K, and Jouvet M (1988) Evidence for histaminergic arousal mechanisms in the hypothalamus of cat. *Neuropharmacology* 27(2): 111–122.

Nelson LE, Guo TZ, Lu J, et al. (2002) The sedative component of anesthesia is mediated by GABA(A) receptors in an endogenous sleep pathway. *Nature Neuroscience* 5: 979–984.

Panula P, Yang HY, and Costa E (1984) Histamine-containing neurons in the rat hypothalamus. *Proceedings of the National Academy of Sciences of the United States of America* 81(8): 2572–2576.

Passani MB, Lin JS, Hancock A, et al. (2004) The histamine H3 receptor as a novel therapeutic target for cognitive and sleep disorders. *Trends in Pharmacological Sciences* 25: 618–625.

Schwartz JC, Arrang JM, Garbarg M, et al. (1991) Histaminergic transmission in the mammalian brain. *Physiological Reviews* 71(1): 1–51.

Vorobjev VS, Sharonova IN, Walsh IB, et al. (1993) Histamine potentiates *N*-methyl-D-aspartate responses in acutely isolated hippocampal neurons. *Neuron* 11: 837–844.

Wada H, Inagaki N, Yamatodani A, et al. (1991) Is the histaminergic neuron system a regulatory center for whole-brain activity? *Trends in Neurosciences* 14(9): 415–418.

Sleep and Sleep States: Hypothalamic Regulation

P M Fuller and J Lu, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The fact that the hypothalamus plays a central role in the regulation of sleep–wake behavior was first suggested in 1916 when the Viennese neurologist and psychiatrist, Baron Constantine von Economo, began to study patients with a viral encephalitis that profoundly affected sleep–wake regulation (i.e., encephalitis lethargica or von Economo's sleeping sickness). The preponderance of patients with von Economo's encephalitis lethargica slept excessively (some more than 20 h per day). Postmortem brain analysis revealed that these individuals had lesions at the junction of the midbrain and posterior hypothalamus, suggesting that this area of the brain contained wake-promoting circuitry. Interestingly, a small percentage of individuals afflicted with encephalitis lethargica became insomniacs, often sleeping only a few hours each day. These insomniacs had lesions involving not the midbrain–diencephalon junction but, rather, the basal forebrain (BF) and anterior hypothalamus, suggesting that these areas of the brain contained sleep-promoting circuitry. von Economo's observations in the lethargica victims were later recapitulated experimentally by Ranson (1939) through his demonstration of hypersomnolence in rhesus monkeys sustaining hypothalamic lesions and by Nauta (1946) in his findings of profound changes in sleep–wake behavior in rats following surgical ablation of preoptic and BF tissue. Remarkably, however, the key role of the hypothalamus in the regulation of sleep–wakefulness was obscured for almost a half century by the discovery of a brain stem 'ascending arousal system' in the late 1940s.

Electrocortical and Behavioral Arousal: The Ascending Reticular Activating System

During cortical arousal (i.e., the neurologically defined vigilance state of wakefulness), the cortical electroencephalogram (EEG) is desynchronized (i.e., high frequency/low amplitude), presumably reflecting differences in the timing of processing of cognitive, motor, and perceptual functions. The EEG directly reflects the collective synaptic potentials of inputs largely to pyramidal cells within the neocortex and hippocampus. In 1935, Bremer uncovered evidence of an ascending arousal system necessary for forebrain and cortical arousal when he demonstrated that transection of the brain stem at the pontomesencephalic level (but not the spinomedullary junction) produced coma in anesthetized cats. More than a decade later, Morruzi and Magoun provided additional support for the concept of an ascending arousal system when they demonstrated that electrical stimulation of the rostral pontine reticular formation produced a desynchronized EEG (an electrophysiological correlate of the conscious state) in anesthetized cats. It was subsequently shown that transections, electrolytic lesions, and infarcts of this same region (i.e., the putative ascending arousal system) produced coma, although it remained unclear whether this was a result of damage to fibers of passage or specific cell groups. Because these ascending arousal pathways traversed (or appeared to originate in) the midbrain reticular formation, the term 'ascending reticular activating system' (ARAS) was coined and the functional integrity of the ARAS was henceforth considered the *sine qua non* for cortical arousal and consciousness. What remained unresolved, however, was the neuroanatomical pathways and population of neurons by which the ARAS might control cortical activity. In the 1970s and 1980s, experiments revealed that, contrary to popular conception, the origin of the ARAS was not a functionally homogeneous mass of reticular tissue in the brain stem tegmentum but, rather, composed of discrete neuronal populations (and neurotransmitters) which can be parsed into two distinct anatomical branches. The first branch of the ARAS consists of cholinergic neurons originating in the pedunculopontine (PPT) and laterodorsal (LDT) tegmental nuclei of the mesopontine tegmentum. The PPT and LDT send excitatory cholinergic projections to the midline and intralaminar nuclei of the thalamus and play a critical gating role in thalamocortical transmission by preventing relay neurons from being hyperpolarized and entering into burst mode, thus clearing the way for thalamocortical sensory transmission.

Regarding thalamic oscillations during wake and sleep, thalamocortical neurons fire in one of two stereotypic modes. When their membrane potential is near threshold (as during wake), they respond to incoming stimuli by firing single spikes. However, when they are deeply hyperpolarized, a low-threshold calcium channel is de-inactivated. The resultant excitatory postsynaptic potentials now produce calcium spikes which are prolonged and produce a depolarized

plateau, from which the neuron fires a series or burst of action potentials (as during sleep). The increase in burst sequences inhibits thalamocortical sensory transmission, thus leading to deprivation of sensory signals from the environment. The state of these thalamocortical neurons is determined by inputs from the thalamic reticular nucleus, which surrounds the lateral surface of the thalamus. The reticular nucleus sends an inhibitory GABAergic projection to $GABA_B$ receptors on the thalamocortical neurons. These $GABA_B$ receptors are potassium channels that can deeply hyperpolarize the neuron sufficiently to de-inactivate the low-threshold calcium channels. Therefore, it is the state of the reticular nucleus that in turn determines the firing state of thalamocortical neurons.

Predictably then, neurons of the PPT and LDT fire most rapidly during wakefulness and rapid eye movement (REM) sleep, which is the stage of sleep associated with cortical activation, muscle atonia, and dreams. Historically, the thalamocortical system has been viewed as a critical component of the ARAS. In this view, the thalamus is thought to function as a major relay to the cortex for the ARAS, with the overall activity of the thalamocortical system forming the origin of the cortical EEG. Indeed, thalamic relay neurons fire in patterns that correlate with the cortical EEG. In turn, the overall activity in the thalamocortical system is thought to be regulated by the ARAS. Nevertheless, a critical role of the thalamocortical system in the ARAS has been challenged by studies demonstrating persistent cortical activation (i.e., desynchronized EEG) following thalamic lesions.

The second branch of the ARAS, which bypasses the thalamus and originates in the upper brain stem and caudal hypothalamus, consists of a series of wake-promoting monoaminergic cell groups that project to and activate neurons in the lateral hypothalamic area (LHA), BF, and cerebral cortex. The monoaminergic systems include the noradrenergic ventrolateral medulla and locus coeruleus, the dopaminergic neurons of the ventral periaqueductal gray matter, the serotoninergic dorsal and median raphe nuclei, and histaminergic neurons in the tuberomammillary nucleus. In general, neurons in all of these cell groups fire more during wakefulness than during non-rapid eye movement (NREM) sleep and show virtually no activity during REM sleep. Furthermore, of these monoaminergic cell groups, the dopaminergic neurons of the ventral periaqueductal gray appear to most potently influence arousal. More rostrally, the LHA, in addition to receiving input from these ascending monoaminergic systems, projects to the BF, cerebral cortex, and, in a reciprocal manner, components of the brain stem arousal systems. The LHA projection to the cortex is thought to augment the arousal system as lesions of the LHA produce profound sleepiness, stupor, or even coma in monkeys and rats. In fact, the LHA contains at least two distinct populations of neurons that are thought to contribute to the regulation of wakefulness. LHA neurons that contain orexin (also called hypocretin) are active during wakefulness and contribute to cortical arousal through ascending projections to the cortex and descending projections to neurons in the tuberomammillary nucleus, locus coeruleus, and dorsal raphe. Mice lacking orexin demonstrate narcolepsy-like symptoms, including profound behavioral state instability (e.g., frequent state transitions and cataplexy). putative glutamatergic LHA neurons that contain melanin-concentrating hormone have similar projections to the orexin neurons but are mostly active during REM sleep, during which time they are thought to inhibit the ascending monoaminergic systems. A third group of putative glutamatergic LHA neurons is also thought to be involved in arousal; however, this group remains unresolved.

Finally, the BF (including the nucleus basalis and magnocellular preoptic nucleus in the substantia innominata, and the medial septal nucleus and nucleus of the diagonal band of Broca) consists of cholinergic neurons intermixed with noncholinergic (largely GABAergic) neurons which project to the cortex, hippocampus, and, to a lesser extent, the thalamus and are implicated in waking and EEG desynchronization.

New Perspective on the ARAS

As previously described, the traditional model for the ARAS posits an origin in the upper brain stem monoaminergic and cholinergic cell groups and reticular formation, and it emphasizes the role of the thalamus as an obligate relay in cortical activation. However, cell-specific lesions using excitoxins (ibotenic acid and kainic acid) in these cell groups have caused relatively limited alteration of wakefulness in rats and cats. Hence, the origin of the arousing influence (i.e., the ARAS) has never been fully explained, and the source has been presumed to be the collective input from all of these structures so that lesions of any one of them will not cause loss of consciousness.

Experimental evidence has strongly suggested that the parabrachial nucleus (PB) and precoeruleus region (PC) of the dorsolateral pontine region, located caudal to the cholinergic PPT/LDT, are the likely neuroanatomical structures of origin for the ARAS, henceforth referred to as the 'ascending arousal system.' In general support of this concept are the findings that stimulation of the PB (in cats) produces marked electrophysiological activation of

the cortex and cell-specific lesions of the PB (in rats) produce profound coma with persistent sub-delta EEG. It has also been shown that the PB sends excitatory projections to the BF that, in turn, project to the cortex, presumably to maintain cortical arousal. In parallel with the PB–BF projection, the PC projects to the medial septum that, in turn, projects to the hippocampus and is presumably important in ascending control of hippocampal function. This hypothesis is supported by the observation that lesions of the PC selectively eliminate hippocampal theta activity during REM sleep. Thus, in this more contemporary view, the PB/PC, like the cholinergic PPT/LDT, is active during all states of cortical activation, including both wakefulness and REM sleep. In the presence of concomitant activation of the monoaminergic arousal systems, the cortical response to PB/PC–BF activation would be wakefulness; in the absence of monoaminergic stimulation, a state of REM sleep would prevail. This ascending arousal model differs fundamentally from the prevailing one in which the wake–sleep state is determined by the level of activation only of the monoaminergic and cholinergic arousal systems. In this alternate view, monoaminergic and cholinergic systems may modulate the activity of the PB/PC and can determine (via their cortical and thalamic projections) whether the PB/PC induces wakefulness versus REM sleep, but they cannot sustain a waking (or REM) state without the PB/PC. In essence, it appears that the loss of wakefulness caused by electrolytic or mechanical lesions in the midbrain reticular formation may be due mainly to disruption of the ascending projections to the BF (but not thalamus) from the PB/PC. Although many other cell groups, including the monoaminergic and cholinergic components of the arousal system, may play specific roles in modulating arousal, particularly under specific physiological circumstances, no other area has this property of being absolutely necessary to maintain a waking state.

As indicated previously, the functional integrity of the thalamus has been widely considered to be necessary for cortical arousal and the maintenance of consciousness. A well-entrenched role for the thalamus in the ARAS is further indicated by the explicit representation in contemporary neuroscience textbooks of the thalamocortical system as an obligate system/relay for maintaining cortical awareness (i.e., consciousness). It is indeed likely that the thalamus, which transmits a wide range of sensory inputs to the cerebral cortex, is necessary for its normal cognitive function. However, experiments dating back to Ranson's 1939 work have repeatedly shown that rodents, nonhuman primates, and humans lacking a thalamic relay and/or thalamocortical system demonstrate relatively normal sleep–wake EEG patterns and cortical arousal. Work has provided additional data supporting a limited role for the thalamic relay as well as the (reticulo thalamocortical pathway (originating in the LDT/PPT) in the overall level of arousal or cortical EEG activation during wakefulness. Again, however, these data do not deny the thalamus a potentially critical role in defining the content of cortical function during wakefulness.

The reported evidence that the integrity of the thalamus is not necessary to maintain an aroused cortical state strongly indicates that activating influences from the brain stem may reach the neocortex via an extrathalamic route. Experimental data have indicated a critical role for the BF in cortical arousal, suggesting the BF might function as the primary structural relay in the ARAS. Indeed, lesions of the BF produce coma (i.e., persistent sub-delta <1 Hz EEG intrinsically generated by cortical neurons) in experimental animals. Work has demonstrated that BF cholinergic and noncholinergic (likely GABAergic) neurons jointly control cortical arousal because loss of either the cholinergic or the noncholinergic neuronal populations (i.e., each lesion sparing one population) has minimal effects on cortical arousal, but combined lesions produce coma.

In summary, work has suggested that (1) the reticulothalamocortical pathway may play a very limited role in the regulation of the overall level of behavioral arousal or EEG activation during wakefulness, (2) the PB/PC–BF-cortical axis is a critical system for maintaining cortical arousal, and (3) cholinergic and noncholinergic (possibly GABAergic) BF neurons jointly contribute to cortical arousal.

Sleep – Turning the Ascending Arousal System Off

As outlined previously, projections from cholinergic neurons, monoaminergic cell groups, and orexin neurons act in a coordinated manner to produce arousal. But what turns off this arousal system to produce sleep? As indicated previously, von Economo proposed a putative sleep-promoting area in the BF/anterior hypothalamic area based on postmortem brain analysis of his insomniac patients. In the 1990s, von Economo's prediction was borne out experimentally by a series of studies which demonstrated a critical role for the ventrolateral preoptic nucleus (VLPO) in inhibiting these arousal circuits during sleep. Neurons of the VLPO are sleep active, contain the inhibitory neurotransmitters galanin and GABA, and receive afferents from each of the major monoaminergic systems. Consistent with a sleep-promoting function,

cell-specific lesion of the VLPO produces profound insomnia and sleep fragmentation. The VLPO contains two populations of neurons. The first is a cluster of neurons in the 'core' of the VLPO that projects most heavily to the tuberomammillary nucleus. A more diffusely distributed second population of VLPO neurons (i.e., the extended VLPO) project mainly to the locus coeruleus and the dorsal and median raphe nuclei and are thought to be important in gating REM sleep. The interaction between the VLPO and components of the arousal systems (e.g., tuberomammillary nucleus, dorsal raphe, and locus coeruleus) has been demonstrated to be mutually inhibitory and, as such, these pathways function analogously to an electronic 'flip-flop' switch/circuit (Figure 1). By virtue of the self-reinforcing nature of these switches (i.e., when each side is firing it reduces its own inhibitory feedback), the flip-flop switch is inherently stable in either end state but avoids intermediate states. The flip-flop design thus ensures stability of behavioral states and facilitates rapid switching between behavioral states. Within the framework of the flip-flop model, the VLPO represents the 'sleep side,' whereas the monoaminergic system represents the 'arousal side.' Importantly, it has not been determined if a mutually inhibitory relationship exists between the VLPO and the PB/PC (putative origin of the arousal system).

Flip-Flop Switch Stability and the Lateral Hypothalamic Area

Flip-flop switches also possess the, at times, undesirable property of abruptly undergoing unwanted state transitions. The frequency of unwanted state transitions may increase if one side of the switch is 'weakened' as the weakened side becomes less able to inhibit the other side, thereby biasing the switch more toward a midpoint where smaller perturbations may trigger a state transition. For example, it has been suggested that cell loss in the VLPO during aging may weaken the switch, ultimately manifesting in sleep fragmentation and daytime napping, both of which are frequent complaints in the elderly. This phenomenon is seen in rats with VLPO lesions, which demonstrate a significant decrease in total sleep time that is largely attributable to shortened sleep bouts. That is, rats with VLPO lesions can initiate but not maintain sleep because the switch is no longer balanced but, rather, biased toward the arousal side.

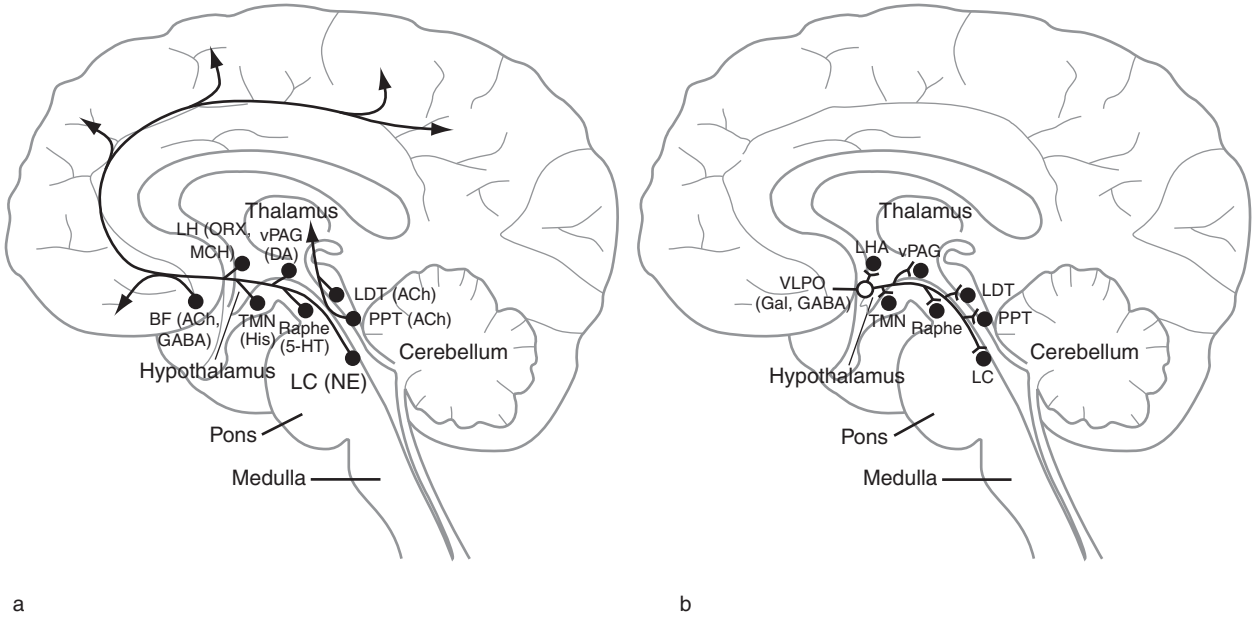

Figure 1 (a) The ascending arousal system consists of noradrenergic neurons of the ventrolateral medulla and locus coeruleus (LC), cholinergic neurons in the pedunculopontine (PPT) and laterodorsal tegmental (LDT) nuclei, serotoninergic neurons in the dorsal raphe nucleus, dopaminergic neurons of the ventral periaqueductal gray matter (vPAG), and histaminergic neurons of the tuberomammillary nucleus (TMN). These systems produce cortical arousal via two pathways: a dorsal route through the thalamus and a ventral route through the hypothalamus and basal forebrain (BF). The latter pathway receives contributions from the orexin and melanin-concentrating hormone (MCH) neurons of the lateral hypothalamic area (LHA) as well as from GABAergic or ACh neurons of the BF. (b) A schematic of the projections of the ventrolateral preoptic nucleus (VLPO; open circle) to the main components of the ascending arousal system. The VLPO neurons are primarily active during sleep and contain the inhibitory transmitters GABA and galanin.

Unwanted state transitions are a hallmark of narcolepsy. Patients with narcolepsy do not sleep more than normal people but experience profound fragmentation of sleep–wake behavior in the form of 'sleep attacks' during the day and frequent awakening during the night. In the late 1990s, two groups of investigators discovered that a lack of lateral hypothalamus (LH) orexin or their type 2 receptors can cause narcolepsy in experimental animals. It was subsequently demonstrated that humans with narcolepsy have few LH orexin neurons and low orexin levels in the cerebrospinal fluid. As indicated previously, orexin neurons are wake-active, and through their ascending projections to the cortex and descending projections to the monoaminergic and cholinergic system of the arousal system, they are thought to augment the ARAS in maintaining cortical arousal. These orexin neurons also project to the VLPO, although, curiously, the VLPO does not contain orexin receptors. This suggests that LH orexin neurons actively reinforce arousal tone as opposed to having a direct, mutually inhibitory relationship with the VLPO. Taken together, it has been hypothesized that LH orexin neurons and other, possibly glutamatergic neurons, are actually 'external' to the flip-flop switch but play an important stabilizing role for the switch. That is, the position of the orexin neurons outside the flip-flop switch permits them to stabilize behavioral state, thus reducing transitions during both sleep and wakefulness, whereas narcoleptic humans or animals lacking orexin have increased transitions in both states.

Homeostatic Regulation of Sleep

Although a teleological explanation for the function of sleep remains unresolved, it is clear that sleep has a restorative effect on the brain. This is perhaps best exemplified by the phenomenon of diminished cognitive and physical performance following sleep deprivation. Although the nature of 'sleep drive' is unknown, it has been conceptualized as a homeostatic pressure that builds during the waking period and is dissipated by sleep. This homeostatic process or 'sleep homeostat' thus represents the need for sleep (i.e., 'sleep propensity'). The cellular determinant of homeostatic sleep drive is unknown, although a putative endogenous somnogen, adenosine (AD), is thought to play a critical role. AD is a naturally occurring purine nucleoside that is hypothesized to accumulate during wake and, upon reaching sufficient concentrations, inhibits neural activity in wake-promoting circuitry of the BF and likely activates sleep-promoting VLPO neurons located adjacent to the BF. Consistent with this hypothesis are the observations that intracerebroventricular injections of AD promote sleep, extracellular concentrations of AD increase with prolonged waking and decline with sleep, administration of AD agonist near the VLPO increases sleep and induces Fos in VLPO neurons, and nonspecific AD antagonists (e.g., caffeine) potently increase waking and decrease sleep. However, data have demonstrated that accumulation of AD in the BF is not necessary for sleep drive, leaving the cellular basis of the sleep homeostat elusive.

Circadian Regulation of Sleep

Although a critical role of the circadian timing system (CTS) in providing temporal organization for behavioral states, including sleep–wakefulness, is well established, the role of the CTS in determining the duration, intensity, and propensity of sleep (and component sleep) is less clear. Nevertheless, regulation of sleep–wakefulness by the CTS is evident because the sleep–wake cycle continues on an approximately 24 h basis even when environmental conditions are constant (i.e., in the absence of environmental time cues). In addition, a clear circadian variation in sleep propensity and sleep structure has been demonstrated in humans by uncoupling the rest–activity cycle from the output of the circadian pacemaker (i.e., 'forced desynchrony protocol'). Circadian modulation of the cortical EEG during NREM sleep and REM sleep (i.e., independent of prior wake time) has also been demonstrated in humans. It has been hypothesized that this modulation of the cortical EEG, particularly REM sleep propensity, may occur through indirect projections from the suprachiasmatic nucleus (SCN) to the mesopontine tegmental nuclei involved in REM sleep generation. Rather strikingly, however, despite a demonstrated role for the circadian pacemaker in governing the timing of sleep, the SCN has only minimal monosynaptic outputs to sleep-regulatory centers, such as the VLPO and LHA orexin neurons, and none to brain stem arousal sites (**Figure 2**).

In 1982, Borbely proposed a two-process model of sleep regulation in which a homeostatic process (i.e., sleep drive) builds during the day and declines exponentially during sleep and interacts with a circadian process that is independent of sleep and waking. A further elaboration of Borbely's model for sleep–wake regulation, the 'opponent process' model, was proposed in 1993. Although conceptually similar to Borbely's two-process model, the opponent process model was the first to identify a role for the SCN in actively facilitating the initiation and maintenance of wakefulness and opposing homeostatic sleep tendency during the subjective day. Specifically, the

model predicted a wake-, but not sleep-, promoting role for the SCN in sleep–wake regulation – that is, an 'alerting' signal produced by the SCN that enhanced wakefulness and actively opposed the sleep homeostat. Human sleep studies, however, have yielded data largely consistent with both a wake- and a sleep-promoting role for the SCN in sleep–wake regulation. Dijk and Czeisler provided data supporting the concept that the primary role of the SCN in sleep–wake regulation is the generation of an alerting signal necessary for generating a consolidated bout of wakefulness. This circadian drive for wakefulness (i.e., the alerting signal) becomes gradually stronger during the waking period so as to oppose increasing homeostatic sleep drive. Similarly, consolidated sleep is achieved by having maximal circadian drive for sleep near the end of the sleep cycle when homeostatic sleep drive is minimal. The concept of a circadian 'hypnotic' signal subserving the circadian drive for sleep has also been proposed; however, no endogenous correlate of a hypnotic signal has been uncovered.

Hypothalamic Circadian Integrator for Sleep–Wake Regulation

As indicated previously, despite the widespread role of the SCN in shaping daily rhythms of behavior, physiology, and gene expression, SCN efferent projections to sleep systems (e.g., LHA orexin neurons and VLPO) are surprisingly limited (Figure 2). The densest projection from the SCN terminates dorsally and caudally in the adjacent subparaventricular zone (SPZ). Similar to the effects of SCN ablation, lesions that include the ventral SPZ abolish behavioral circadian rhythms, including rest–activity, sleep–wake cycles, and feeding. Surprisingly, lesions of the dorsal SPZ eliminate the circadian rhythms of body temperature but have little effect on the rhythms of locomotor activity or sleep. Thus, taken together, these observations suggest that neurons of the SPZ function as an obligate relay to maintain circadian rhythms of sleep and body temperature. Like the SCN, however, the SPZ has limited projections to the major components of the sleep–wake regulatory system. The SCN and ventral SPZ do however project densely to the dorsomedial hypothalamic nucleus (DMH). Lesions of the DMH also abolish circadian rhythms of locomotor activity, wake–sleep, feeding, and corticosteroid secretion, but not the rhythm of body temperature, which retains a normal circadian variation. These findings suggest that the principal neuronal output pathway that determines the timing of circadian behavior is mediated by a primary projection from the SCN to the SPZ, followed by a secondary projection to the DMH. Interestingly, the DMH sends

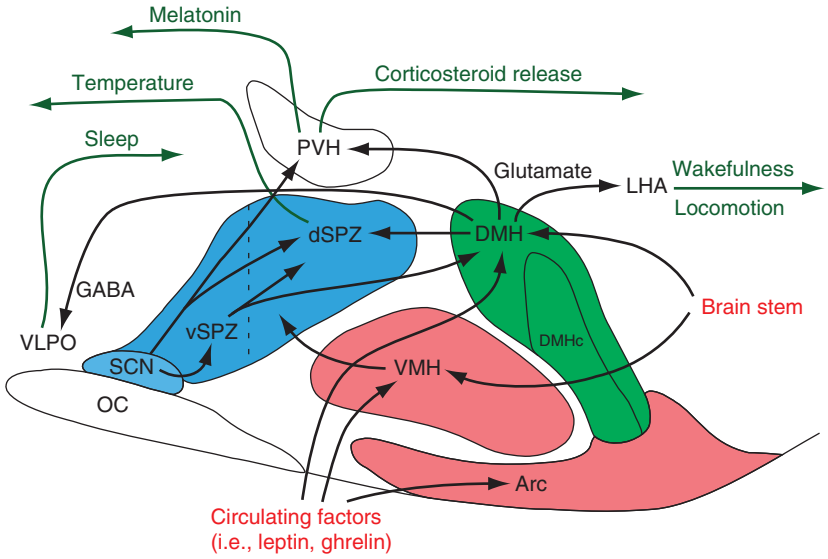

Figure 2 Hypothalamic integrator model for sleep–wake regulation. The suprachiasmatic nuclei (SCN) have relatively modest projections to sleep-regulatory networks, including the VLPO and orexin neurons. However, the bulk of its output goes to the ventral (vSPZ) and dorsal (dSPZ) subparaventricular zone and the dorsomedial hypothalamus (DMH). Lesions of the vSPZ disrupt the circadian rhythms of sleep–wakefulness and locomotor activity, whereas lesions of the dSPZ severely impair the circadian rhythms of body temperature. These findings suggest that vSPZ and dSPZ are obligate relays for the normal expression of the circadian rhythms of these functions. A major target of the SPZ is the DMH. Lesions of the DMH profoundly attenuate the circadian rhythms of sleep–wake, locomotor activity, corticosteroid secretion, and feeding. Interestingly, the DMH is one of the largest sources of input to the VLPO and lateral hypothalamic orexin neurons. This multistage regulation of circadian behavior in the hypothalamus allows for the integration of multiple time cues from the environment to shape daily patterns of sleep–wake.

a dense glutamatergic projection to the LHA orexin neurons as well as an intense GABAergic projection to the VLPO, suggesting a primary wake-promoting role for the DMH. Thus, these projections from the DMH neurons to sleep–wake centers of the hypothalamus provide a putative mechanism for the circadian regulation of sleep–wake cycles.

These findings raise the question of why circadian rhythms of behavior and physiology are regulated by a complex multisynaptic pathway in the hypothalamus rather than by a direct projection from the SCN to sleep–wake centers of the brain. In both diurnal and nocturnal species, the SCN is active during the light cycle, whereas the VLPO is always active during the sleep cycle (i.e., irrespective of circadian phenotype), suggesting an absolute requirement for additional (intervening) neural circuitry for organizing diurnal and nocturnal circadian programs despite identical clock input and sleep-control systems. In consideration of the fact that the timing of environmental pressures such as food availability and predation does not always track the solar day, and that the SCN normally remain phase-locked to the light–dark cycle, having flexibility in circadian organization may be extremely adaptive. Thus, for example, this series of multiple relays in the hypothalamus (SCN, SPZ, and DMH) could allow for the integration of light entrained circadian cues from the SCN with environmental time cues (e.g., feeding, temperature, and social cues) to sculpt patterns of rest–activity and sleep–wake cycles that are optimal for survival.

Corticolimbic Projections to the Hypothalamus – A New Frontier

In addition to the integration of external cues, these hypothalamic relays likely coordinate signals from visceral, emotional, and cognitive inputs which collectively influence arousal state. Although studies have demonstrated extensive corticolimbic (e.g., medial prefrontal cortex, infralimbic cortex, bed nucleus of the stria terminalis, and ventral subiculum) projections to the VLPO, SCN, and LHA, we still know very little about the mechanisms by which emotional or cognitive inputs interact with sleep–wake and arousal systems. Nevertheless, it has been hypothesized that during times of emergency these corticolimbic inputs (so-called 'allostatic load') may produce a state of hyperarousal that overrides the homeostatic and circadian drives for sleep, thereby enhancing cognitive and physical performance. Although the adaptive nature of this transient hyperarousal state is clear, chronic or inappropriate activity in corticolimbic sites secondary to psychological stress or depression may in fact be a primary contributing etiology to insomnia.

Summary

Approximately one century ago, von Economo proposed that sleep–wakefulness was governed by a waking influence arising from the midbrain and a sleep influence from the BF/anterior hypothalamic area. The remarkable accuracy of von Economo's predictions has been confirmed experimentally by the subsequent discovery of (1) a brain stem ascending arousal system that maintains consciousness and activates the forebrain both during wakefulness and during REM sleep and (2) the identification of a key hypothalamic switch that shuts off the arousal system to produce sleep. Work has further revealed the critical role of several additional hypothalamic nuclei in the circadian and homeostatic regulation of sleep–wakefulness. Despite this remarkable progress, significant gaps remain in our knowledge of how the arousal system and hypothalamic circuits interact to execute control of behavior state, including sleep–wake regulation.

See also: Circadian Rhythms in Sleepiness, Alertness, and Performance; Sleep and Sleep States: Thalamic Regulation; Sleep and Sleep States: Histamine Role; Sleep and Sleep States: Cytokines and Neuromodulation; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Oscillations and PGO Waves; Sleep Oscillations; Sleep: Development and Circadian Control; Sleep–Wake State Regulation by Noradrenaline and Serotonin; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Blanco-Centurion C, Xu M, Murillo-Rodriguez E, et al. (2006) Adenosine and sleep homeostasis in the basal forebrain. *Journal of Neuroscience* 26: 8092–8100.

Borbely AA and Tobler I (1985) Homeostatic and circadian principles in sleep regulation in the rat. In: McGinty D (ed.) *Brain Mechanisms of Sleep*, pp. 35–44. New York: Raven Press.

Buzsaki G, Bickford RG, Ponomareff G, Thal LJ, Mandel R, and Gage FH (1988) Nucleus basalis and thalamic control of neocortical activity in the freely moving rat. *Journal of Neuroscience* 8: 4007–4026.

Chou TC, Scammell TE, Gooley JJ, Gaus SE, Saper CB, and Lu J (2003) Critical role of dorsomedial hypothalamic nucleus in a wide range of behavioral circadian rhythms. *Journal of Neuroscience* 23: 10691–10702.

Dijk DJ and Czeisler CA (1995) Contribution of the circadian pacemaker and the sleep homeostat to sleep propensity, sleep structure, electroencephalographic slow waves, and sleep spindle activity in humans. *Journal of Neuroscience* 15: 3526–3538.

Edgar DM, Dement WC, and Fuller CA (1993) Effect of SCN lesions on sleep in squirrel monkeys: Evidence for opponent

processes in sleep–wake regulation. *Journal of Neuroscience* 13: 1065–1079.

Kinney HC, Korein J, Panigrahy A, Dikkes P, and Goode R (1994) Neuropathological findings in the brain of Karen Ann Quinlan. The role of the thalamus in the persistent vegetative state. *New England Journal of Medicine* 330(21): 1469–1475.

Lu J, Greco MA, Shiromani P, and Saper CB (2000) Effect of lesions of the ventrolateral preoptic nucleus on NREM and REM sleep. *Journal of Neuroscience* 20: 3830–3842.

Lu J, Sherman D, Devor M, and Saper CB (2006) A putative flip-flop switch for control of REM sleep. *Nature* 441: 589–594.

Lu J, Zhang YH, Chou TC, et al. (2001) Contrasting effects of ibotenate lesions of the paraventricular nucleus and subparaventricular zone on sleep–wake cycle and temperature regulation. *Journal of Neuroscience* 21: 4864–4874.

Mochizuki T, Crocker A, McCormack S, Yanagisawa M, Sakurai T, and Scammell TE (2004) Behavioral state instability in orexin knock-out mice. *Journal of Neuroscience* 24: 6291–6300.

Moruzzi G and Magoun HW (1949) Brain stem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology* 1: 455–473.

Nauta WJH (1946) Hypothalamic regulation of sleep in rats. An experimental study. *Journal of Neurophysiology* 9: 285–314.

Porkka-Heiskanen T, Alanko L, Kalinchuk A, and Stenberg D (2002) Adenosine and sleep. *Sleep Medicine Reviews* 6: 321–332.

Ranson SW (1939) Somnolence caused by hypothalamic lesions in monkeys. *Archives of Neurology and Psychiatry* 41: 1–23.

Saper CB, Lu J, Chou TC, and Gooley J (2005) The hypothalamic integrator for circadian rhythms. *Trends in Neurosciences* 28: 152–157.

Saper CB, Scammell TE, and Lu J (2005) Hypothalamic regulation of sleep and circadian rhythms. *Nature* 437: 1257–1263.

Sherin JE, Shiromani PJ, McCarley RW, and Saper CB (1996) Activation of ventrolateral preoptic neurons during sleep. *Science* 271: 216–219.

Starzl TE, Taylor CW, and Magoun HW (1951) Ascending conduction in reticular activating system, with special reference to the diencephalon. *Journal of Neurophysiology* 14: 461–477.

Steriade M, McCormick DA, and Sejnowski TJ (1993) Thalamocortical oscillations in the sleeping and aroused brain. *Science* 262: 679–685.

Vanderwolf CH and Stewart DJ (1988) Thalamic control of neocortical activation: A critical re-evaluation. *Brain Research Bulletin* 20(4): 529–538.

Villablanca J and Salinas-Zeballos ME (1972) Sleep–wakefulness, EEG and behavioral studies of chronic cats without the thalamus: The 'athalamic' cat. *Archives Italiennes de Biologie (Pisa)* 110: 383–411.

von Economo C (1930) Sleep as a problem of localization. *Journal of Nervous Disorders* 71: 249–259.

Sleep and Sleep States: Network Reactivation

S Ribeiro, Edmond and Lily Safra International Institute of Neuroscience of Natal, and Universidade Federal do Rio Grande do Norte, Natal, Brazil
M Nicolelis, Duke University Medical Center, Durham, NC, USA, and École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The persistence of neural activity following a given stimulus or behavior is a widespread phenomenon in vertebrates. Simple introspection suffices to experience the phenomenon, which occurs during waking as well as during sleep. Network reactivation occurs, for instance, when a song heard by chance on the radio lingers in one's auditory imagination for days thereafter. Such reverberation seems to reflect a particularly robust encoding of the song memory, causing an involuntary recollection that runs on top of the heavy background of sensory interference that typically characterizes waking. When occurring during sleep, network reactivation gives rise to dreams as well as nonvivid thoughts ('mentation'), depending on the specific sleep state considered.

The Canadian psychologist Donald Hebb was probably the first to point out that patterns of neuronal activity present during memory encoding remain in the brain after stimulus cessation, maintaining the freshly acquired information 'alive' until structural cellular changes, able to transform a short-lived reverberatory trace into a long-lasting memory, have time to occur. In the past 20 years, evidence has accumulated suggesting that network reactivation is a critical step for learning, being related to many of its idiosyncrasies. It has also been demonstrated that network reactivation is enhanced during sleep, offering a plausible explanation for the beneficial role of sleep in memory consolidation. In addition, network reactivation elucidates the persistence of salient waking experiences during dreaming, a phenomenon labeled 'day residue' by Sigmund Freud. To date, network reactivation has been detected with a variety of techniques from the cellular to the systems level, including the transcriptional upregulation of activity-dependent genes, single neuron activity recorded with microwires, and large-scale neural activity measured by electroencephalographic waves and cerebral blood flow. This article reviews the properties and functions of network reactivation.

Feedback and Rhythms

The mechanistic investigation of network reactivation dates back to the pioneering studies conducted by Lorente de Nó on the dynamics of electrical excitation in hardwired neuronal loops. The famous disciple of Santiago Ramón y Cajal was interested in understanding how activity induced in a given neuron would be passed on to other cells so as to eventually return to the stimulus origin by way of recurrent connections. Based on much anatomy and some physiology, Lorente de Nó proposed that feedback loops comprising multiple neurons echo activation upon stimulation of a single neuron, creating depolarization cycles that only dissipate after several laps take place.

The notion of small-circuit network reactivation enthused neuroscientists in the twentieth century because recursive processes such as this could in principle implement clocks, pacemakers, and memory building blocks. This enthusiasm led to the discovery of specialized neuronal loops that generate recurrent traveling waves within and across brain structures. The combination of recursive architecture, local inhibitory neurons, and the modulatory input of different neurotransmitters (e.g., acetylcholine) gives rise to rhythms of variable duration which are characteristic of global brain states such as waking, sleep, and dreaming. Different substates occur within the major states, such as very long episodes of sustained regional oscillation, including theta rhythm (4–9 Hz in different species), which is composed of sinusoidal oscillations that strongly engage the hippocampus during alert waking as well as during rapid eye movement (REM) sleep, the dreaming phase of sleep in humans. Other rhythms include short-lived oscillations in the range of milliseconds to seconds, such as the cortical spindles (10–14 Hz) that are particularly abundant during stage 2 sleep, constituting its defining characteristic. Stage 2 sleep is one of the substates of slow-wave sleep (SWS), the nondreaming phase of sleep in humans.

The span of brain oscillations deemed behaviorally relevant goes from very slow, whole brain waves ($<$1 Hz) to very fast ripples ($\sim$200 Hz) that engage small groups of hippocampal neurons. All these oscillations depend on the interaction of specific neural circuits, such as the septohippocampal axis for theta rhythm, the thalamocortical loop for spindles, and intrahippocampal networks for ripples. Oscillations may coexist in time and space, phase locking at specific moments so as to optimize communication

among brain areas. For instance, slow oscillations of the membrane potential in the cerebral cortex constrain neurons to 'upstates' and 'downstates.' The slow oscillatory upstate represents the depolarizing phase of the slow oscillation, leading to discrete bursts of cortical activity that are followed by hippocampal sharp wave/ripples complexes, preferentially triggered by the upstate.

Reactivation, Reverberation, and Replay

Postulated by Donald Hebb more than 50 years ago, the term 'reverberation' was replaced in the past few decades by 'reactivation' and 'replay,' but neither term fully accounts for the general phenomenon in question. Although network reactivation decreases substantially during waking, it does not disappear completely when animals are awake. This was demonstrated by electrophysiological studies in rodents and can easily be verified by introspection. For most people, it suffices to remain quiet under reduced sensory stimulation in order for mnemonic images to appear. The available data indicate that salient experience causes sustained neural reverberation rather than discrete reactivation, in the sense that memory traces are continuously detectable during subsequent periods across all behavioral states in a state-dependent manner. Reverberating patterns of neural activity associated with past experience are largely, but not completely, masked during waking by incoming sensory inputs.

Another related term that became well known in the early 1990s is memory replay, which implies a high-fidelity repetition of past patterns of neural activity. However, the accuracy of neuronal reverberation is still controversial. In the rodent hippocampus, the path run by the animal during wakefulness can be very accurately re-created based on the spatiotemporal neuronal firing patterns observed during sleep. In contrast, researchers found low values for firing rate correlations among neurons in the cerebral cortex, suggesting that cortical reactivation implements a low-fidelity emulation of memories rather than a perfect replay. This is probably related to the fact that many portions of the mammalian cerebral cortex are dedicated to the joint representation of multiple perceptions or acts. As a consequence, single cells are recruited into many different neuronal groups, multiplexing information in a way that likely blurs the detectable network reactivation of any discrete engram. A notable counterpoint is found in songbirds: The nucleus RA, single output of the telencephalic song circuitry, displays a highly accurate replay of song-related firing patterns (**Figure 1(d)**). This remarkably rare case of high-fidelity replay derives from the very specialized neural processing carried out by this motor nucleus, namely the descending activation of a single, unique, and invariant sequential engram: the bird's own song.

A Link with Memory

Despite the importance of brain rhythms to explain major differences among behavioral states, nowadays it is recognized that the kind of network reactivation that underlies the recall of fine-grain memories arises from the interaction of much more complex neural circuits, in which the reverberation of neuronal activity does not derive trivially from a preexisting anatomical loop. Beginning in the 1970s, the investigation of primates performing working memory tasks showed that neurons in the prefrontal cortex show persistent firing rate changes during delay periods that separate stimulation from behavioral response. Currently, there is a consensus that network reactivation in the prefrontal cortex and other associative areas of intricate connectivity is directly implicated in the temporary storage of stimulus-related information required for correct behavioral responses. These so-called 'memory neurons' can remain active for many seconds after stimulus removal, keeping information about the past available for future decision making. In these cases, the experimental task shaped the neuronal behavior, uncovering neurons able to reverberate information on voluntary demand. This kind of delayed goal-directed reactivation is extremely important in nature for hunters and hunted alike. However, it does not account for all the time we spend in spontaneous network reactivation, either day- or night dreaming.

Network Reactivation Increases during Sleep

A major breakthrough in the search to identify the mechanisms underlying spontaneous network reactivation resulted from an experiment on the poststimulus sleep activity of hippocampal neurons performed by Constantine Pavlides and Jonathan Winson at the Rockefeller University in the late 1980s. Their elegant study took advantage of the fact that pyramidal hippocampal neurons in the rat tend to fire very selectively according to the spatial position of the animal in a given environment, hence the name 'place cells.' Using chronic electrode implants in the hippocampus, the experimenters first identified and recorded pairs of neurons with nonoverlapping place fields – that is, places in which the neurons fired the most. Then, rats were restricted from entering the place field of either cell overnight. Finally, animals were confined for 10–15 min within the place field of one of the cells so as to strongly stimulate one neuron while suppressing the activity of the yoked one. By

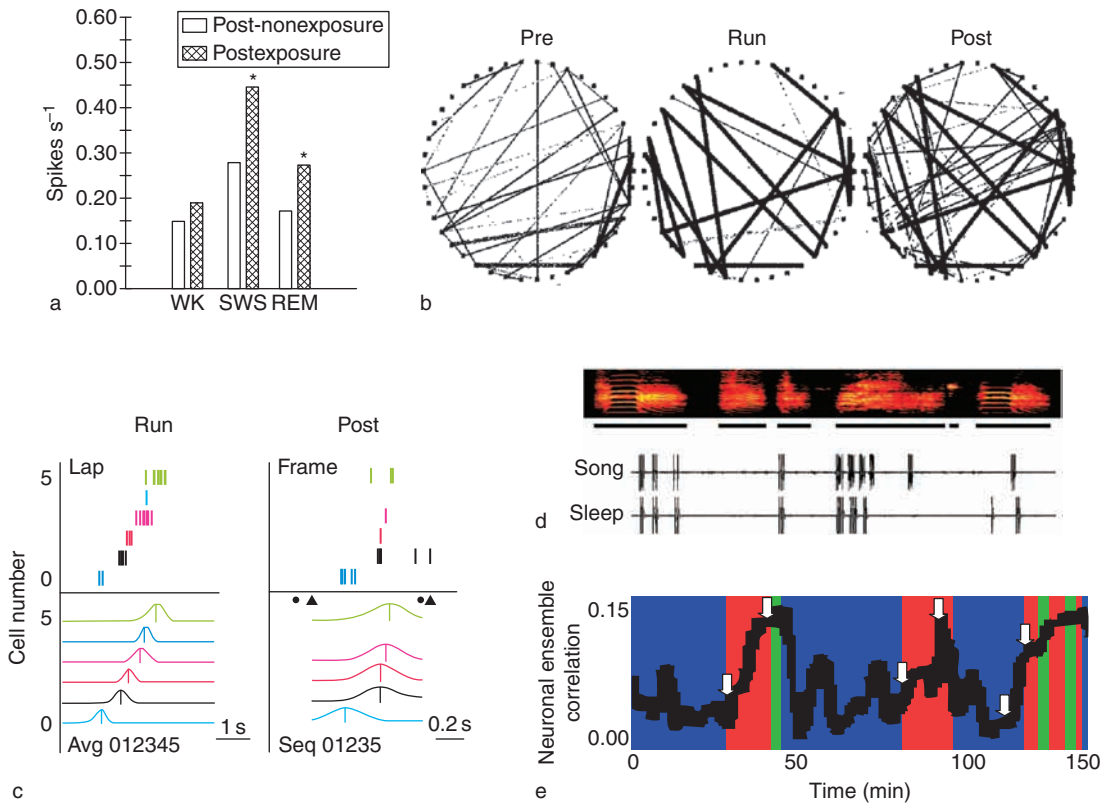

Figure 1 Neuronal reactivation in rats and birds. (a) Hippocampal place cells are reactivated during SWS and REM sleep after waking (WK) exposure to their corresponding place fields. $^*p < 0.05$. (b) For hippocampal place cells, pairwise neuronal correlations established during waking exposure to a circular maze (RUN) are preserved during subsequent SWS (POST), while being absent from prior SWS (PRE). In each circle, neurons are represented by dots over the perimeter, and correlations are represented as lines whose thickness is proportional to correlation strength. (c) SWS replays neuronal firing sequences observed during RUN in the hippocampus. Notice the temporal compression of the neuronal patterns during SWS. Shown are firing patterns during a single running lap on the left–right trajectory (left) or during an equivalent sleep 'frame' period (right). One neuron per row, one spike per tick. (d) Premotor neurons in nucleus RA of a zebra finch accurately replay singing-specific activity during sleep. Shown are raw traces of neuronal activity (900 ms) recorded during singing (premotor activity) and sleep (spontaneous activity). A color spectrograph of the song that the bird sang is shown on top, with horizontal bars indicating different song syllables. (e) To estimate similarity with past patterns of neuronal ensemble activity produced by the exploration of novel objects in rats, neuronal activity patterns were compared with extensive recordings of neuronal action potentials using a template-matching algorithm. As indicated by the white arrows, neuronal ensemble correlations increase gradually between waking (blue) and sleep (red for SWS and green for REM sleep). (a) Adapted from Pavlides C and Winson J (1989) Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of Neuroscience* 9(8): 2907–2918. (b) Adapted from Wilson MA and McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* 265(5172): 676–679. (c) Adapted from Ji D and Wilson MA (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience* 10(1): 100–107. (d) Adapted from Dave AS and Margoliash D (2000) Song replay during sleep and computational rules for sensorimotor vocal learning. *Science* 290(5492): 812–816. (e) Adapted from Ribeiro S, Gervasoni D, Soares ES, et al. (2004) Long-lasting novelty-induced neuronal reverberation during slow-wave sleep in multiple forebrain areas. *PLoS Biology* 2(1): 126–137.

quantifying the neuronal firing rates during and after spatial confinement across the wake–sleep cycle, Pavlides and Winson found that changes in firing rates recorded during waking persist during ensuing SWS and REM sleep (Figure 1(a)). In other words, place cells exposed to their place fields prior to sleep showed increased firing rates during SWS and REM sleep compared to place cells that were not exposed to their place fields.

In 1994, Matthew Wilson and Bruce McNaughton showed that pairs of neurons recorded in the hippocampus carry into subsequent SWS the temporal firing relationships observed when animals run a circular path beforehand (Figure 1(b)). Follow-up of these discoveries by different research groups established that reactivation is strongest and most stable during SWS, becomes more variable during REM sleep, and decreases during waking. The difference between sleep states probably arises from the overall low activation of most of the brain during SWS, in contrast with the elevated neural activation that characterizes REM sleep, which is so high as to be considered waking-like, hence the alias 'paradoxical sleep' preferred by most European neuroscientists. Given the

consistent increase in network reactivation during SWS, its higher variance during REM sleep, and the short duration of total REM sleep in comparison with total SWS (~1:4 in rats and humans), one must conclude that SWS plays the major role in network reactivation, although dreaming sleep (i.e., REM sleep) also participates in the process.

Reactivation and Learning

The great importance ascribed to network reactivation by neuroscientists stems from its putative link with memory consolidation. Although the early breakthroughs in the field were mostly the result of electrophysiological investigation of rodent neurons, it was the survey of human brains with electroencephalography, positron emission topography, and functional magnetic resonance imaging that established a firm link between learning and network reactivation (Figure 2(a)). Postacquisition reverberation of regional brain activity during sleep has been shown to be proportional to memory acquisition and to quantitatively predict learning. Such reverberation is reflected in local increases of blood oxygenation as well as the amplitude of low-frequency electroencephalographic waves (delta band, <4 Hz) (Figure 2(b)). If learning coincides with an increase in slow oscillatory activity, the converse is also true: The application of 0.75 Hz pulses by transcranial magnetic stimulation during SWS enhances the retention of hippocampus-dependent memories in humans, but similar stimulation at 5 Hz does not produce the effect.

Network reactivation during sleep seems to favor the anatomical reorganization of memory traces, an increasingly recognized hallmark of memory consolidation. In the case of spatial memory tasks, which depend on the hippocampus for initial encoding, sleep favors a displacement of the activation elicited by task performance, from the hippocampus to striatal networks. Indeed, different lines of evidence suggest that network reactivation during sleep is directly implicated in the temporally graded propagation of memories from the hippocampus to cortical areas, a phenomenon made famous by patient H.M., who developed anterograde amnesia for declarative memories after bilateral removal of the hippocampus. Likewise, sleep may promote engram propagation in the case of implicit memory tasks, which initially involve the cerebellum but progressively engage the basal ganglia. To summarize, network reactivation during sleep seems to play a rather important role in the strengthening and relocation of memory traces within the brain.

Characterization of Reactivation Properties

One possible explanation for the increase in network reactivation during sleep is that endogenous brain activation, freed from sensory interference during

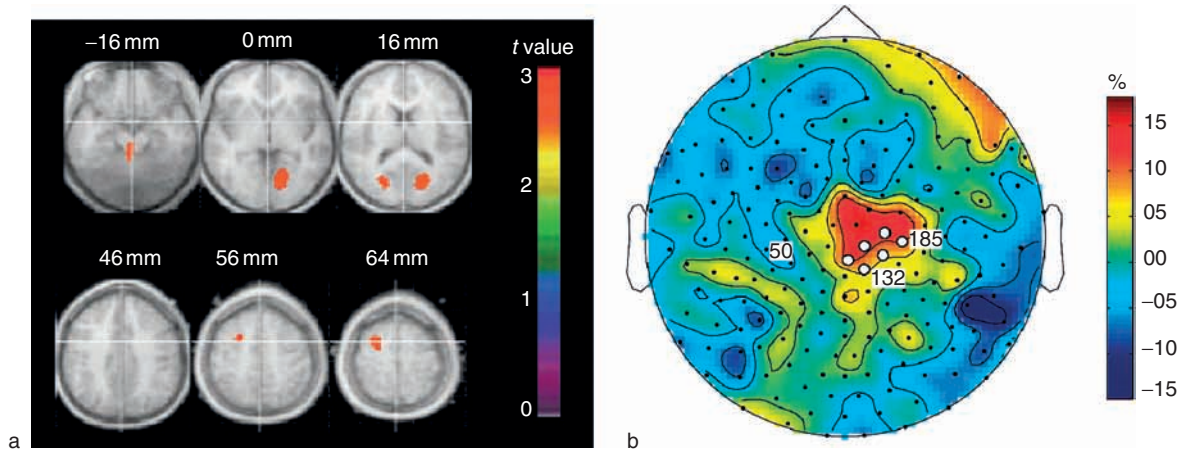

Figure 2 Regional brain reactivation in humans. (a) Brain regions reactivated during REM sleep in subjects previously trained in a visuomotor task. Shown are brain maps of regional cerebral blood flow as measured by positron emission tomography. The activated regions comprise the cuneus and adjacent striate striate cortex bilaterally, the left premotor cortex, and the mesencephalon. Maps show six different brain levels (from 16 mm below to 64 mm above the bicommissural plane), superimposed on the average MRI image of the sleeping subjects. (b) The amplitude of slow waves recorded with electroencephalography increases locally in subjects trained in a motor task. Shown is the topographic distribution of the percentage change in the amplitude of slow waves between the task condition and a control task, void of motor learning. White dots indicate the cluster of six electrodes showing increased amplitude of slow waves after motor learning. (a) Adapted from Maquet P, Laureys S, Peigneux P, et al. (2000) Experience-dependent changes in cerebral activation during human REM sleep. *Nature Neuroscience* 3(8): 831–836. (b) Adapted from Huber R, Ghilardi MF, Massimini M, and Tononi G (2004) Local sleep and learning. *Nature* 430: 78–81.

off-line states, flows preferentially through previously activated synapses. In principle, this would suffice to cause neuronal firing patterns originally produced during waking to reverberate significantly above chance levels during sleep. For instance, telencephalic reactivation may be driven by the pontine waves that are particularly abundant during REM sleep. To date, experience-dependent network reactivation during sleep has been observed in rodents, macaques, humans, and songbirds, indicating a very widespread biological phenomenon. Sleep promotes the coordinated reverberation of waking-acquired activity patterns across large groups of neurons in the hippocampus, primary sensory and associative cortices, basal ganglia, and even primary sensory thalamic relays. The synchronization of neuronal firing by slow cortical oscillations seems to generate discrete activity frames between the hippocampus and the cerebral cortex, promoting coordinating interactions across these brain regions during the membrane potential upstate (i.e., the depolarizing phase of the slow oscillation).

It is unclear whether reverberation is favored by novel stimuli, which are more salient for perception, or by well-trained stimuli/behaviors, which are deeply engrained by habit. Results favoring either option coexist in the literature. There is evidence that major changes in firing rate patterns last longer in the cerebral cortex than in the hippocampus, but more research is needed to elucidate this issue. The duration of network reactivation is also under intense debate. Whereas several studies point to fast-fading effects lasting less than 1 h, some experiments have detected signs of mnemonic reverberation for up to 48 h after stimulus removal.

It has been proposed that network reactivation during sleep occurs under different regimes of time compression or expansion, perhaps allowing for more efficient mnemonic processing. Some studies indicate that the experience-dependent reverberation of neuronal firing patterns is hastened during SWS, presumably because the activity bursts characteristic of that state (e.g., hippocampal ripples) promote the fast replay of entire neuronal firing sequences (Figure 1(c)). Similarly, there is some evidence suggesting that network reactivation is slowed during REM sleep, but a consensus regarding this issue is yet to be achieved. Another point of contention, with controversial evidence, is whether sleep also harbors antireverberation – that is, patterns of neuronal ensemble activity statistically more dissimilar from the waking reference patterns than expected by chance. In principle, the novelty-induced reverberation and antireverberation of neuronal firing patterns could play balancing roles in the delineation of specific memory traces, embossed in the synaptic landscape of the forebrain.

Reactivation at the Molecular Level

The very wide prevalence of network reactivation within the brain and across species suggests that it derives from extremely conserved elements of the vertebrate nervous system, such as the major anatomical connections and the molecular backbone shared by all neurons. As discussed previously, the overall plan of the vertebrate brain comprises several circuits prone to recursive activity, such as the thalamocortical loop. The conspicuous rhythms implemented by these circuits are ridden by finer-grain reactivation at the level of single neurons. Such neuronal reactivation, rather than large-scale rhythms, constitutes the most likely substrate for the recall of specific memories.

The acquisition of a single new memory involves activity-dependent molecular mechanisms able to store information at the synaptic level from many minutes to a few hours. Most of these mechanisms depend critically on the rise of calcium, either permeated through membrane channels (e.g., NMDA voltage-gated channels) or as a result of increased intracellular release. Elevated calcium levels activate phosphorylating enzymes (kinases) that trigger major changes in the internal biochemical milieu of the cell. Network reactivation immediately after memory acquisition seems to require, for instance, the phosphorylation of calcium/calmodulin kinase II (CaMKII), an autocatalytic enzyme that can by itself sustain local changes in synaptic strength. In this case, the substrate of reactivation is not an anatomical loop but, rather, a molecular one since CaMKII maintains memory by persistent self-phosphorylation.

Whereas the initial stabilization of recently acquired memories relies on spatially restricted synaptic mechanisms such as CaMKII activation, it is well established that the long-lasting persistence of mnemonic traces requires *de novo* protein synthesis. Indeed, calcium-dependent kinase activation triggers signaling cascades that couple membrane depolarization to genomic regulation inside the cell nucleus, leading to the transcriptional upregulation of several immediate early genes (IEGs) required for the long-term maintenance of memory. One of these genes is the activity-regulated cytoskeleton-associated protein (Arc), a calcium-dependent IEG that interacts with glutamatergic AMPA receptors, actin, and CaMKII to promote synaptic remodeling. Arc mRNA is transported to dendrites for local translation, leading to retrograde (presomatic) effects. Another activity-dependent gene regulated by calcium is zif-268 (also known as egr-1, krox-24, NGFI-A, and ZENK), which encodes a transcription factor with binding sites on the promoters of hundreds of different genes. The zif-268 protein is thought to modulate synaptic

remodeling by controlling the expression of genes directly involved in synaptic function, such as synapsins, the most abundant protein constituent of synapses, involved in the regulation of synaptic vesicle release. The transcriptional control of synapsin levels by the zif-268 protein illustrates how neuronal depolarization can prompt anterograde (postsomatic) synaptic remodeling. Such a mechanism could, in principle, disseminate synaptic changes downstream of long neuronal chains, effectively propagating memory traces across separate circuits.

The large-amplitude neural oscillations typical of SWS promote marked periodic fluctuations of calcium levels in activated synapses, as indicated by extracellular calcium measurements during sleep. As a consequence, it has been hypothesized that SWS is concomitant with the activation of multiple calcium-dependent kinases, with a role in memory formation. This would result in a pretranscriptional amplification of synaptic changes encoding novel memory traces during SWS, as suggested by recent data (Figure 1(e), white arrows). Although experimental evidence linking all the elements of these calcium-dependent metabolic cascades to sleep-dependent network reactivation is missing, results indicate that REM sleep triggers the activation/upregulation of several calcium-dependent molecules required for neuronal plasticity and memory formation, such as zif-268, arc, cyclic AMP response element-binding protein, and brain-derived nerve growth factor (Figure 3(a)). The sleep-dependent transcriptional upregulation of the zif-268 gene seems to also reveal memory trace relocation from the hippocampus to the cerebral cortex (Figure 3(b)). The activity-dependent anterograde response caused by the zif-268 gene represents a putative molecular mechanism for the propagation of synaptic changes across neuronal circuits, leading to memory relocation.

The fact that neuronal activity can be traced to molecular markers of plasticity establishes a strong parallel between the cellular and systems levels, offering support for the Hebbian notion that memory requires two consecutive and distinct steps, namely network reactivation for short-term recall and synaptic remodeling for long-term storage. The current results suggest that SWS and REM sleep play distinct and complementary roles in memory consolidation, with memory recall (network reactivation) occurring mainly during SWS and memory storage (plasticity-related gene) sparked off during REM sleep. Such a mechanism fulfills early conceptual notions of a two-step process for memory consolidation during sleep and is in line with evidence that SWS and REM sleep have synergistic effects on human procedural learning and developmental plasticity.

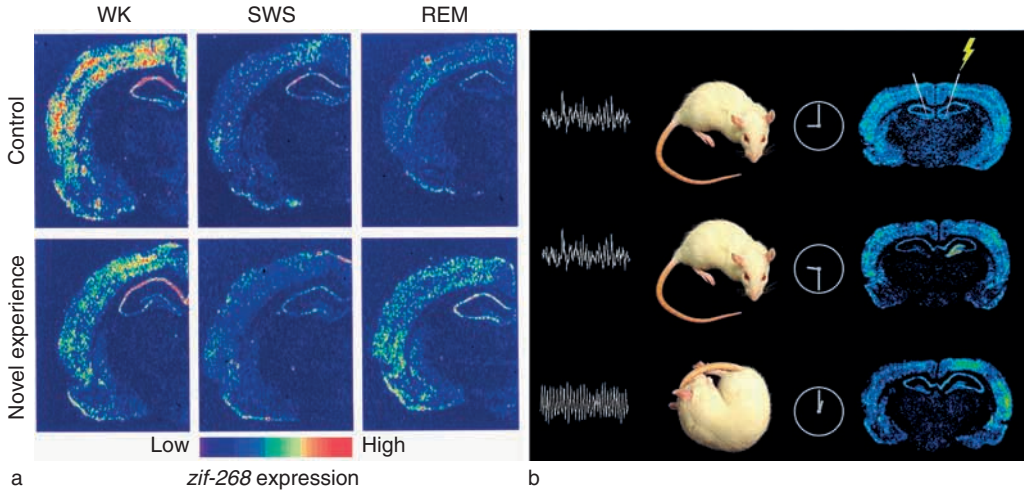

Figure 3 Molecular reactivation. (a) The mRNA levels of the activity-dependent gene *zif-268*, a transcription factor with an important role in memory consolidation, are upregulated during REM sleep in rats previously exposed to novel experience but not in control animals kept in their habitual home cages. Shown are autoradiograms of brain sections subjected to radioactive *in situ* hybridization. (b) The unilateral induction of hippocampal long-term potentiation during waking leads to marked unilateral gene upregulation in the hippocampus 30 min after high-frequency electrical stimulation (yellow arrow). Four hours later, during REM sleep, zif-268 is downregulated in the hippocampus but strongly induced in extrahippocampal regions such as the cerebral cortex and the amygdala, indicating that REM sleep disengages the hippocampus and engages extrahippocampal regions. Traces on the left represent hippocampal local field potentials typical of waking (top and middle) and REM sleep (bottom). (a) Reproduced from Ribeiro S, Goyal V, Mello CV, and Pavlides C (1999) Brain gene expression during REM sleep depends on prior waking experience. *Learning & Memory* 6(5): 500–508, with permission from Gold spring Harbor Laboratory press. (b) Adapted from Ribeiro S, Mello CV, Velho T, Gardner TJ, Jarvis ED, and Pavlides C (2002) Induction of hippocampal long-term potentiation during waking leads to increased extrahippocampal zif-268 expression during ensuing rapid-eye-movement sleep. *Journal of Neuroscience* 22(24): 10914–10923.

Dreams, Waking, and Sleep Mentation

In adult humans, REM sleep is almost always accompanied by dreaming, as can be easily demonstrated by waking experimental subjects during REM sleep and requesting dream reports. During human REM sleep, a selected set of forebrain areas are reactivated, including portions of the hypothalamus, amygdala, septum, and ventral striatum, as well as the anterior cingulate, orbitofrontal, entorhinal, and insular cortices. The dorsolateral prefrontal cortex, however, shows reduced activity. Given the role of prefrontal cortex activity in the executive control of behavior, its reduction during REM sleep has been proposed to underlie the poor logical concatenation and reduced self-awareness of dreaming in comparison with waking.

Dreams may have first evolved as a collateral effect of extended REM sleep, a characteristic mammalian trait. Pet owners know that cats and dogs seem to act out dreams during sleep. More controlled evidence of dreaming in nonhuman mammals was obtained by lesions of the brain stem nuclei that promote muscle atonia during REM sleep. Cats with such lesions sleep quietly through SWS, but upon entering REM sleep they become suddenly agitated by vigorous species-specific behaviors, such as meowing and pouncing. How are dreams generated, and what is their biological function?

As noted by Sigmund Freud, dreams often involve elements of the experience of the preceding day(s), the day residue. Although high-fidelity memory reactivation must be at the roots of the cognitive functions of sleep and dreams, it does not account entirely for the symbolic complexity that characterizes the oneiric narrative. After all, it is not common to dream about the exact repetition of waking scenes. On the contrary, most human dreams are characterized by the intrusion of illogical elements, leading to unforeseen associations. Human dreams are subjective narratives composed of familiar and unfamiliar beings, things, and places interacting around a self-representation of the dreamer that mostly observes an unfolding plot. Dreams vary in intensity, ranging from confused and faint impressions to complex time-evolving narratives with vivid imagery and surprising turns. Although dreams tend to be dominated by visual images, they can also involve combinations of auditory, olfactory, tactile, gustatory, motor, vestibular, and linguistic modalities. Dreams can sometimes be extremely pleasant or just the opposite, but they are usually characterized by a mix of emotions. Dreams are also hyperassociative, linking characters, places, and actions in bizarre ways. Dreams can also anticipate events of the coming day(s), particularly when subjects undergo extreme anxiety and expectation. A good example is provided by the dreams of students during the night before taking difficult exams; these dreams often contain detailed anticipatory simulations of the expected challenges, in content and/or context.

Freud also postulated that dream narratives fulfill wishes (or anti-wishes) of the awake subject, simulating the realization (or frustration) of specific desires. Whereas the prevalence of overt wish fulfillment in the dreams of normal adults is low, it is common in young children. Children living in war zones have a much higher incidence of nightmares, a result that has been interpreted as evidence that the function of dreams is to simulate environmental threats. Based on the existence of blissful dreams, other neuroscientists have expanded the threat-simulation theory to include the simulation of rewarded behaviors as a dream function. The notion of dreaming as behavior simulation leading to either reward or punishment is supported by the neuropsychological investigation of dreamless patients. These patients, who often conserve REM sleep, represent rare clinical cases with lesions of mesocorticolimbic pathways connecting reward centers with the thalamus, striatum, and cerebral cortex. Although REM sleep is spared in these cases, the loss of dreaming is accompanied by subjective reports of sleep disturbance, as assessed by qualitative interviews. This suggests a role for dreaming in sleep maintenance, which is another Freudian postulate.

The current evidence indicates that dreams reflect the intense and noisy network reactivation that takes place during extended REM sleep, with the cerebral cortex lightened up at high neuronal firing rates and oscillatory frequencies predominantly above 30 Hz (gamma band). The qualitative feeling of 'quasi-reality' in dreams derives from the fact that network reactivation during REM sleep occurs at an intensity and oscillatory spectrum comparable to those of waking. Such high levels of excitation cause the reverberating memories that comprise dreams to appear bright and vivid to the self-representation, and variations in neural activity levels explain the dynamic range of vividness that characterizes normal dreaming. Likewise, the simultaneous occurrence of neuronal reverberation in multiple forebrain sites during REM sleep explains the wide variation in dream 'sensory' modality. There is also a compelling relation between the higher variability of network reactivation during REM sleep and the fragmentation, condensation, and bizarreness of dreams. Far from being random, dream narratives highlight waking events according to how recent, novel, and behaviorally significant they were (i.e., dreams are directed by the anxieties and preoccupations of the dreamer). Dreams seldom occur during SWS. Instead,

the subjective experience of SWS consists of low-intensity but coherent thoughts resembling waking reasoning. The coherence of such 'mentation' likely reflects the high stability of neuronal reverberation during SWS. Similarly, the lack of intense imagery during SWS probably reflects the decreased cortical activity and slow oscillations that characterize network reactivation in this state.

Concluding Remarks

Network reactivation is a fundamental property of neuronal circuits that occurs at several levels, ranging from molecular and cellular mechanisms to the consequent systemic and behavioral manifestations. Reactivation is based on the existence of recurrent biological loops, from self-activating enzymes to large networks of correlated neurons distributed over multiple brain regions, with various degrees of duration. Although reactivation is favored by the interruption of sensory and motor interference provided by sleep, data suggest that it occurs nonstop during waking as well, spanning one's entire lifetime.

See also: Dream Function; Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep Research and Sleep Medicine in Historical Perspective; Sleep Architecture; Sleep Oscillations and PGO Waves; Sleep and Sleep States: PET Activation Patterns; Sleep and Sleep States: Hippocampus–Neocortex Dialog; Sleep and Sleep States: Gene Expression; Sleep-Dependent Memory Processing.

Further Reading

Dave AS and Margoliash D (2000) Song replay during sleep and computational rules for sensorimotor vocal learning. *Science* 290(5492): 812–816.

Hebb DO (1949) *The Organization of Behavior: A Neuropsychological Theory.* New York: Wiley.

Huber R, Ghilardi MF, Massimini M, and Tononi G (2004) Local sleep and learning. *Nature* 430: 78–81.

Ji D and Wilson MA (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience* 10(1): 100–107.

Llinas RR and Pare D (1991) Of dreaming and wakefulness. *Neuroscience* 44(3): 521–535.

Maquet P, Laureys S, Peigneux P, et al. (2000) Experience-dependent changes in cerebral activation during human REM sleep. *Nature Neuroscience* 3(8): 831–836.

Marshall L, Helgadottir H, Molle M, and Born J (2006) Boosting slow oscillations during sleep potentiates memory. *Nature* 444(7119): 610–613.

Nádasdy Z, Hirase H, Czurkó A, Csicsvari J, and Buzsáki G (1999) Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience* 19(21): 9497–9507.

Orban P, Rauchs G, Balteau E, et al. (2006) Sleep after spatial learning promotes covert reorganization of brain activity. *Proceedings of the National Academy of Sciences of the United States of America* 103(18): 7124–7129.

Pavlides M and Winson J (2006) Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of Neuroscience* 9(8): 2907–2918.

Ribeiro S, Goyal V, Mello CV, and Pavlides C (1999) Brain gene expression during REM sleep depends on prior waking experience. *Learning & Memory* 6(5): 500–508.

Ribeiro S, Gervasoni D, Soares ES, et al. (2004) Long-lasting novelty-induced neuronal reverberation during slow-wave sleep in multiple forebrain areas. *PLoS Biology* 2(1): 126–137.

Ribeiro S, Mello CV, Velho T, Gardner TJ, Jarvis ED, and Pavlides C (2002) Induction of hippocampal long-term potentiation during waking leads to increased extrahippocampal zif-268 expression during ensuing rapid-eye-movement sleep. *Journal of Neuroscience* 22(24): 10914–10923.

Ribeiro S and Nicolelis MAL (2004) Reverberation, storage and postsynaptic propagation of memories during sleep. *Learning & Memory* 11(6): 686–696.

Wang XJ (2001) Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences* 24(8): 455–463.

Wilson MA and McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* 265(5172): 676–679.

Winson J (1985) *Brain and Psyche.* New York: Anchor Press.

Sleep and Sleep States: PET Activation Patterns

T T Dang-Vu and M Desseilles, University of Liege and Centre Hospitalier Universitaire, Liege, Belgium
P Peigneux, University of Liege, Liege, Belgium
S Laureys and P Maquet, University of Liege and Centre Hospitalier Universitaire, Liege, Belgium

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Positron emission tomography (PET) shows the distribution in the body (or in the brain, in the context of neuroimaging) of compounds labeled with positron-emitting isotopes. In the field of sleep research, two compounds have mainly been used to assess the global and regional cerebral activity during the different sleep stages: [^{18}F]fluorodeoxygucose, which is a marker of glucose metabolism, and oxygen-15-labeled water ($H_2^{15}O$), which is an indirect marker of blood flow. In the framework of sleep research, PET images are acquired in combination with simultaneous recordings of brain electrical activity by electroencephalography for sleep staging. After data transformation and inter-individual standardization, statistical procedures allow the comparison of regional cerebral activity between two conditions (e.g., a specific sleep stage and wakefulness) or the examination of correlation patterns between regional activity distribution in a specific condition and measurements of relevant physiological events (e.g., sleep oscillations within a specific sleep stage). The resulting functional brain maps have contributed important information about human sleep physiology, assessed in the light of animal experimental data. Sleep functions have also been investigated with this technique. In particular, a role for sleep in learning and memory has been evidenced using PET recording during posttraining sleep associated with behavioral measurements.

Sleep Stages and Sleep Oscillations

PET studies have shown that global and regional patterns of brain activity during wakefulness are remarkably different from those obtained during sleep (i.e., rapid eye movement (REM) sleep and non-REM sleep).

Non-Rapid Eye Movement Sleep

Non-REM sleep is a state of synchronization of the electroencephalogram (EEG), along with the production of specific oscillations within thalamo-cortical networks: spindles, delta waves, and slow oscillations.

When compared to wakefulness and REM sleep, non-REM sleep is characterized by a decrease in global cerebral blood flow and regional cerebral blood flow (rCBF). The largest decreases in rCBF are observed in a set of subcortical and cortical areas, including the dorsal pons, mesencephalon, thalami, basal ganglia, basal forebrain, prefrontal cortex, anterior cingulate cortex, and precuneus (Figure 1(b)).

A lower activity in the brain stem and thalamus was expected from animal data of non-REM sleep-generation mechanisms; a decreased firing rate in the brain stem induces the sequential alternation of long hyperpolarization and short depolarization patterns in thalamic neurons, which leads to the formation of non-REM sleep rhythms (spindles, delta oscillations, and slow) among thalamocortical networks. Due to the low time resolution of the PET technique (i.e., one scan is the activity averaged over a period of time ranging from 40 to 90 s) and because the hemodynamic influences of hyperpolarization predominate over those of depolarization phases, brain areas where non-REM sleep rhythms are most expressed appear deactivated in PET studies.

At the cortical level, the pattern of deactivation is not homogeneously distributed. Indeed, the least active areas in non-REM sleep are located in associative cortices, especially the ventromedial prefrontal cortex (VMPF), which includes the orbitofrontal and anterior cingulate cortices. The VMPF is also one of the most active brain areas during the awake resting state and is involved in important cognitive processes such as action monitoring and decision making. In contrast, the primary cortices were the least deactivated cortical areas during non-REM sleep. This specific segregation of cortical activity remains poorly understood, although some hypotheses have been proposed, for instance, (1) that associative areas might be more profoundly influenced by non-REM sleep rhythms than primary cortices because they are the most active cerebral areas during wakefulness and (2) that sleep intensity is homeostatically related to prior waking activity at the regional level.

The precuneus is another cortical area that displays a reduced activity during non-REM sleep in PET studies. It is a region particularly active in wakefulness, during which it is involved in visual mental imagery processes, explicit memory retrieval, and consciousness. The precuneus is also deactivated during other states of decreased consciousness such as pharmacological sedation, hypnotic states, and vegetative states. The role of the precuneus during sleep still remains unclear. Its decreased activity during

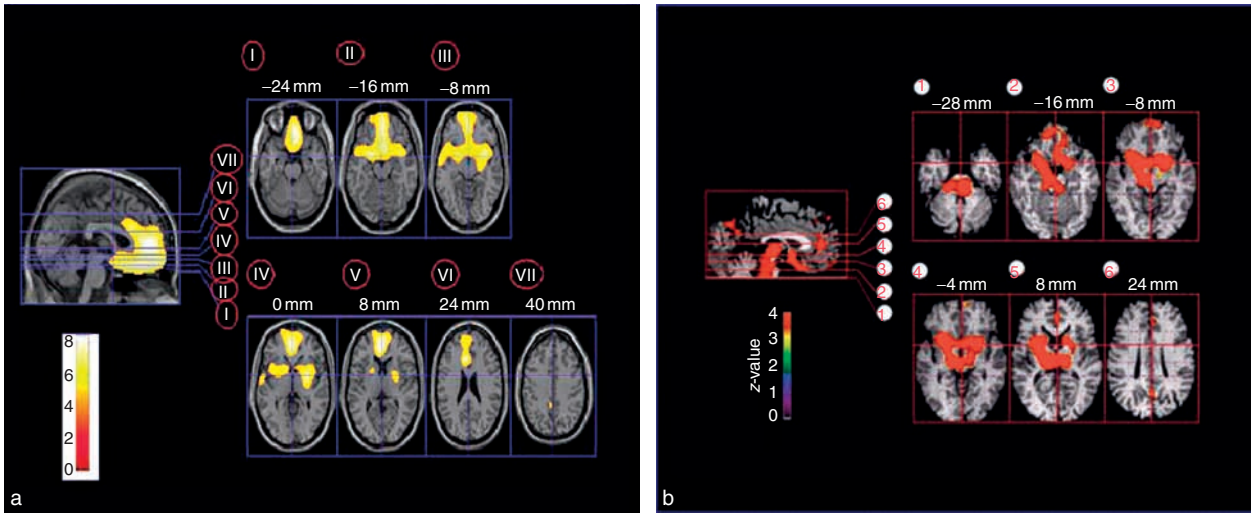

Figure 1 Functional neuroanatomy of normal human non-REM sleep, as assessed by $H_2^{15}O$ PET: (a) brain areas in which rCBF decreases as a function of delta power during non-REM sleep (stages 2–4); (b) brain areas in which rCBF decreases during non-REM sleep as compared to wakefulness and REM sleep. In (a), image sections are displayed on different levels of the *z*-axis, as indicated at the top of each picture. The color scale indicates the range of *z*-values for the activated voxels, superimposed on a canonical T1-weighted MRI image. Displayed voxels are significant at $p < 0.05$ after correction for multiple comparisons. Note that different analyses (correlation (a) vs. subtraction (b)) gave rise to strikingly similar results in terms of the regional blood flow distribution. MRI, magnetic resonance imaging; PET, positron emission tomography; rCBF, regional cerebral blood flow; REM, rapid eye movement. Adapted from Maquet P, Degueldre C, Delfiore G, et al. (1997) Functional neuroanatomy of human slow wave sleep. *Journal of Neuroscience* 17(8): 2807–2812. Copyright 1997 by the Society for Neuroscience.

non-REM sleep might reflect a homeostatic compensation of a high waking activity.

The basal forebrain and the basal ganglia (mostly the striatum) have also been found to be consistently deactivated during non-REM sleep in PET sleep studies. The basal forebrain is a functionally and structurally heterogeneous structure in which a majority of neurons is involved in cortical activation during wakefulness and REM sleep. Its deactivation during non-REM sleep may therefore reflect a lower activity of these arousal-promoting neurons. The role of the basal ganglia, and especially the striatum, in sleep regulation remains speculative, however. Two hypotheses have been put forward. First, the striatum receives massive afferent inputs from the frontal cortex and the thalamus, which are also deactivated during non-REM sleep. These structures are most likely to participate to the formation of non-REM sleep rhythms by oscillating synchronously between long phases of hyperpolarization and bursts of discharges. Due to the fronto- and thalamo-striatal connections, basal ganglia neurons may likewise oscillate following these non-REM sleep-rhythm sequential patterns and, thus, appear deactivated at the macroscopic level. According to the second proposal, the striatum may also send projections to the pedunculopontine tegmental nucleus (PPT) in the brain stem and induce the disinhibition of this activating structure, subsequently leading to cortical arousal during wakefulness. In this perspective, the decreasing activity in the striatum during non-REM sleep may be related to a reduced propensity to arousal as well.

PET studies have not merely compared the activity between non-REM sleep and other stages of sleep or wakefulness. Another way to describe brain activity during this sleep stage was to search for the neural correlates of non-REM sleep oscillations (spindles and delta waves) by looking for brain areas in which rCBF values correlate with the EEG activity of interest (i.e., power density in the sigma or delta frequency band). Using this approach, spindle activity (12–15 Hz) has been shown to correlate negatively with rCBF in the thalamus, meaning that the higher the power density within the spindle frequency range on EEG recordings, the lower the thalamic activity. This result is in line with spindle-generation mechanisms in mammals, which are dominated by the cyclic repetition of hyperpolarization and spike bursts in the thalamic neurons. Delta activity (1.5–4 Hz) correlates negatively with rCBF in the VMPF, basal forebrain, striatum, and precuneus (Figure 1(a)). The resulting map is very similar to the brain map of the regions less activated during non-REM sleep compared to REM sleep and wakefulness (Figure 1(b)), which emphasizes the notion that delta activity is a prominent feature of non-REM sleep. A major difference,

however, is the absence of significant correlation between delta and thalamus activity, whereas the thalamus is markedly deactivated during non-REM sleep compared to other sleep stages or wakefulness. This discrepancy can be explained taking into account that two types of delta activity have been described in animals: a stereotyped delta rhythm, whose generation depends on intrinsic properties of thalamocortical neurons, and a cortical polymorphous delta rhythm, which persists after extensive thalamectomy. Therefore, the delta correlation map might preferentially reflect the brain areas involved in the generation of cortical delta waves during non-REM sleep. The physiology of these cortically generated delta oscillations, and their relationship with the slow rhythm, is still poorly understood.

It should be emphasized here that deactivation patterns found with PET studies do not imply that these brain areas remain idle during non-REM sleep. As already stated, non-REM sleep oscillations are produced by the recurrent and sequential alternation of hyperpolarization and depolarization phases in the thalamic and cortical neurons. The latter are characterized by bursts of neuronal firing temporally organized by the non-REM sleep slow oscillation. PET is insensitive to these bursts because it averages brain activity over long periods, during which the effects on regional brain function of prolonged hyperpolarization periods exceed those of shorter depolarization phases. This issue should be addressed in future studies using techniques with higher spatial and temporal resolution, such as combined EEG-functional magnetic resonance imaging (fMRI), that will provide activation patterns closer to the genuine non-REM sleep physiology, dominated by synchronous and low-frequency oscillations.

Rapid Eye Movement Sleep

REM sleep is characterized by a desynchronized EEG with high-frequency, low-amplitude activity, a major muscle atonia, spontaneous rapid eye movements, and an intense oneiric mentation. In contrast to non-REM sleep, PET studies showed that the global cerebral blood flow during REM sleep is sustained to a level comparable to wakefulness. At the regional level, several brain areas even display a higher rCBF during REM sleep compared to wakefulness and/or non-REM sleep: the pontine tegmentum, thalamus, amygdala, hippocampus, anterior cingulate cortex, temporal–occipital areas, and the basal forebrain. Conversely, regional decreases in activity are also found during REM sleep, in the dorsolateral prefrontal cortex (DLPF), posterior cingulate gyrus, precuneus, and inferior parietal cortex (Figure 2).

Activation of the pontine tegmentum, thalamic nuclei, and basal forebrain is in agreement with REM sleep-generation mechanisms in animals. REM sleep is generated by cholinergic processes arising from brain stem structures, located in the pedunculopontine tegmentum and laterodorsal tegmentum, which mediate widespread cortical activation via a dorsal pathway innervating the thalamus and a ventral pathway innervating the basal forebrain.

Consistent activations during REM sleep are also found in limbic and paralimbic structures, including the amygdaloid complexes, hippocampal formation, and anterior cingulate cortex. There is evidence that the amygdala plays a role in REM sleep modulation. For instance, carbachol (cholinergic agonist) injections in the central nucleus of amygdaloid complexes enhance REM sleep duration, and a shift from non-REM sleep to REM sleep is observed in rats after serotonin injection into the amygdala. The amygdala also appears to modulate important physiological features of REM sleep. For example, recent PET data suggest that the large variability in heart rate during REM sleep could be explained by a prominent influence of the amygdaloid complexes. Both amygdala and hippocampal formation are also critical for memory systems and may thus participate in the processing of memory traces during REM sleep.

PET studies showed higher rCBF values in temporal–occipital areas during REM sleep compared to wakefulness. These areas include the inferior temporal cortex and the fusiform gyrus, which are extrastriate cortices belonging to the ventral visual stream. Functional interactions between the primary visual cortex (striate cortex) and the extrastriate cortex during REM sleep were also assessed in a PET study. This study showed that extrastriate cortex activation is significantly correlated with striate cortex deactivation during REM sleep, whereas their activities are positively correlated during wakefulness. This result has been interpreted as supporting the hypothesis that REM sleep allows internal information processing (between extrastriate areas and their paralimbic projections, both activated during REM sleep) in a closed system dissociated from interactions with the environment (via striate cortex and prefrontal cortex, both deactivated during REM sleep).

Regional deactivations during REM sleep are mostly located in the DLPF (inferior and middle frontal gyrus), precuneus, posterior cingulate cortex, and a part of the parietal cortex (temporal–parietal region and inferior parietal lobule). In contrast, the activity in the superior parietal lobe and in the superior and medial prefrontal cortex is similar to waking level. The reasons of these cortical deactivations are still unclear. Animal data show that the cortical areas less

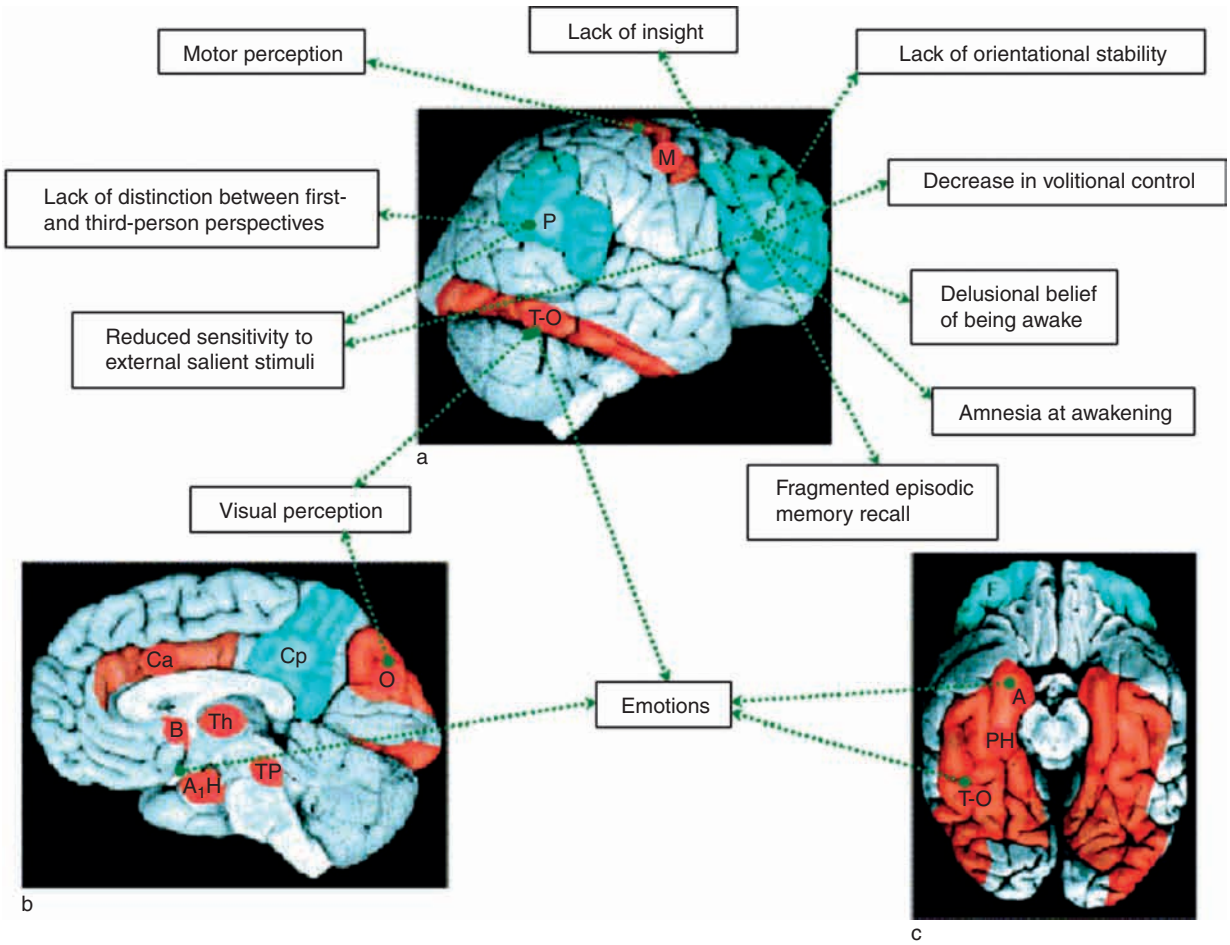

Figure 2 Schematic representation of the functional neuroanatomy of normal human REM sleep: (a) lateral view; (b) medial view; (c) ventral view. Regions in which there is a relative increase in neural activity during REM sleep are colored in red; regions in which neural activity decreases are colored in blue. Green arrows illustrate the proposed relationships between the activity of specific brain areas and several dreaming features, which may be accounted for by regional patterns of activity during REM sleep. A, amygdala; B, basal forebrain; Ca, anterior cingulate gyrus; Cp, posterior cingulate gyrus and precuneus; F, prefrontal cortex (middle, inferior, and orbito-frontal cortices); H, hypothalamus; M, motor cortex; P, parietal cortex (inferior parietal lobule); PH, parahippocampical gyrus; O, occipital-lateral cortex; REM, rapid eye movement; Th, thalamus; T-O, temporal–occipital extrastriate cortex; TP, pontine tegmentum. Reproduced from Schwartz S and Maquet P (2002) Sleep imaging and the neuro-psychological assessment of dreams. *Trends in Cognitive Science* 6(1): 23–30, with permission from Elsevier.

active during REM sleep (the inferior parietal and DLPF) receive only few inputs from the amygdala, whereas areas more active during REM sleep (the anterior cingulate and right parietal operculum) receive rich amygdalar inputs, suggesting that amygdala may modulate the pattern of cortical activity during REM sleep. This hypothesis is also supported by PET results showing that functional interactions between the amygdala and the temporal–occipital cortices are different in the context of REM sleep than in non-REM sleep or wakefulness. The amygdalo-cortical network during REM sleep might contribute, in particular, to the selective processing of emotionally relevant memories.

In animals, rapid eye movements during REM sleep are closely related to the occurrence of ponto-geniculo-occipital (PGO) waves. PGO waves are bioelectrical phasic potentials occurring during the transition from non-REM sleep to REM sleep or during REM sleep itself. They are observed in many parts of the animal brain, but they are most easily recorded in the pons, lateral geniculate bodies of the thalamus, and occipital cortex, hence their name. Animal data also suggest that PGO waves might have important functional roles, such as a facilitation of brain plasticity. In humans, a PET study found correlations during REM sleep, but not during wakefulness, between

spontaneous eye movements and rCBF in the occipital cortex and lateral geniculate bodies of the thalamus, supporting the hypothesis that PGO-like activities are present during REM sleep in humans as well as in animals.

Neuroimaging and Dreams

Dreaming is experienced every night by many humans as multisensory mental representations occurring spontaneously during sleep, often organized in a narrative manner. Dreams are more often associated with REM sleep (dream reports are present in 90% of the episodes in subjects awakened during a REM sleep period). Dream during REM sleep is classically considered the canonical dream with several characteristics: hallucinoid imagery, narrative structure, bizarreness, hyperemotionality, delusional acceptance, and deficient memory of its content. Dream reports are usually less organized, poorer, and shorter during non-REM sleep.

At the neuroimaging level, it has been proposed that the PET activation patterns during REM sleep might be interpreted in the light of dreams. Indeed, brain functional segregation during REM sleep, as already described, may account for several typical dreaming features (Figure 2). We present a few examples here.

Dream reports differentially involve the sensory modalities. Visual components are nearly always present, auditory components are present in 40–60% of dreams, movement and tactile sensations are present in 15–30%, and smell and taste components are present in less than 1%. The activation of the associative posterior (temporal–occipital) cortices may be related to these perceptual aspects of dreams, consistently dominated by visual and auditory elements. Accordingly, a cessation of visual dream imagery was reported in some patients with temporal–occipital lesions.

Dream content is also characterized by the prominence of emotions especially negative emotions such as fear and anxiety. Responses to threatening stimuli or stressful situations are known to be modulated by the amygdala during wakefulness. The high limbic, and especially amygdalar, activity during REM sleep may therefore underpin the emotional intensity occurring during dreams. Moreover, PET data have shown functional interactions between the amygdala and temporal–occipital cortices during REM sleep, and fMRI studies during wakefulness have found positive relationships between the emotional intensity of visual stimuli and both amygdalar and inferotemporal cortex activity. Together, these data suggest that emotional experience during dreams may engage specific brain networks encompassing the amygdala and temporal–occipital cortices rather than a single brain area.

The bizarreness of dream reports is another recurrent hallmark. The cortical patterns of hypoactivity during REM sleep may contribute to these aspects. Indeed, the prefrontal regions deactivated during REM sleep overlap with the areas involved in the selection of stimulus–response associations according to contextual signals, past events, and internal goals. The decreased activity of prefrontal regions consequently impairs the efficiency of this integrative system during REM sleep. This may, for instance, account for the inability of the dreamer to integrate information of a whole episode, leading to an oneiric content in which characters, times, and places are fused, incongruous, and discontinuous. It may also explain the decrease in volitional control and the dreamer's failure to organize mental representation toward a well-identified internal goal. Prefrontal inactivity during REM sleep may, then, explain why the dreamer is unable to control the flow of dream events.

These hypoactive areas also encompass frontal areas participating in the processing of episodic memory, that is, the ability to encode and recollect personally experienced events set in a particular spatiotemporal context. Prefrontal areas are involved in the monitoring of episodic memory retrieval by checking the accuracy and completeness of the processed information. The impairment of these areas during REM sleep might explain why, although 65% of dream reports contain residues of recent waking activity, only 1.4% of them are considered to represent a replay of full memory episodes. During dreams, episodic elements might be reactivated in a fragmented fashion, but the deactivation of the prefrontal cortex prevents the various details of past events from being integrated into an identifiable life episode. Impairment in these areas might also explain the well-known amnesia on awakening that prevents most people from accurately remembering the events experienced during the dream episode.

Overall, PET studies have brought interesting hypotheses about human dream organization. However, these assumptions remain largely speculative and partial, especially because combined dream and functional imaging data are still very sparse. In future studies, the use of scales to quantify and categorize the dream narrative in terms of different perceptual, emotional, or bizarre elements should provide crucial explanatory variables to model neuroimaging data. This combination of systematic neuropsychological assessments of dreams with neuroimaging data might greatly improve our insight into dreaming mechanisms by bringing genuine functional maps of the dreaming brain.

Sleep and Memory

The idea that sleep is involved in the processing of information that should be memorized dates back to the theory of memory consolidation a century ago. Consolidation here refers to the processing of memory traces during which the traces may be reactivated, analyzed, and gradually incorporated into long-term memory. In addition to anecdotic similarities between brain areas activated in canonical maps of normal human sleep and learning-related areas, other PET studies have supported a role for sleep in learning and memory. These studies suggest that neuronal activity patterns observed during a learning episode are reinstated during posttraining sleep. These reactivations allow for the adaptation of intercellular connection strengths between the elements of the network and the incorporation of the new experience into long-term memory. Both REM sleep and non-REM sleep are differentially involved in these processes.

During REM sleep, but not during non-REM sleep, the premotor and visual cortical areas engaged in the implicit learning of a procedural motor task are reactivated in subjects previously trained, compared to non-trained subjects. Such reactivations do not occur if the subjects are trained in a random version of the same task, showing that processing during sleep is not merely due to intense task exposure but rather relies on the sequential content of the material. These data speak for the reprocessing during REM sleep of procedural memory traces acquired during previous wakefulness.

During non-REM sleep, but not during REM sleep, the hippocampal areas activated during a spatial learning task are reactivated in subjects previously trained, compared to nontrained subjects. This result suggests a reprocessing during non-REM sleep of recent spatial memory traces acquired during previous wakefulness. Moreover, this study demonstrated a significant correlation between rCBF increases in the hippocampal areas during non-REM sleep and an overnight gain in behavioral performance, supporting the hypothesis that this offline reprocessing is related to plastic changes underlying a subsequent improvement in performance.

PET studies have thus contributed to a large amount of evidence supporting a role of sleep in learning and memory. These results are in agreement with behavioral data suggesting that REM sleep and non-REM sleep differentially modulate the consolidation of procedural and spatial/episodic memories, respectively, in the model called the dual process hypothesis. However, other behavioral data support, and brain-imaging results do not oppose to, the sequential hypothesis, in which the ordered succession of non-REM sleep and REM sleep are necessary for the consolidation of memory traces, whatever the memory system. Furthermore, memory systems and mechanisms are complex and heterogeneous, and it remains unclear whether sleep influences all memory systems. Behavioral data already suggest that not all memories need sleep to consolidate. More studies are needed to refine the characterization of the precise contribution of each sleep stage in the processing of the different categories of memory traces.

Conclusion and Perspectives

PET studies have contributed to the understanding of human sleep neurophysiology by describing the functional neuroanatomy of sleep stages, by proposing brain correlates of dreaming features, and by showing cerebral reactivations during sleep of regions involved in prior learning. Yet our knowledge about normal human sleep remains fragmentary. These seminal studies only pave the way for future works that will benefit from the development of multimodal techniques such as combined EEG-fMRI. This neuroimaging method is technically more challenging in sleep studies, but it also has a number of advantages, including better spatial and temporal resolutions. This should help in describing more accurately the dynamics of brain activity throughout the sleep–wake cycle and especially in relation to the sleep physiological events, such as sleep oscillations. Future studies should also continue to characterize the specific extent of sleep's contribution to the learning and memory processes, which remains a topic of intense debates. There is no doubt that functional neuroimaging studies bring us significant insights into sleep functions (e.g., the relationships between sleep and brain plasticity) and sleep physiology, which is a prerequisite to the investigation of sleep disorders.

See also: Autonomic Dysregulation During REM Sleep; Dream Function; Electroencephalography (EEG); Positron Emission Tomography (PET); Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep Research and Sleep Medicine in Historical Perspective; Sleep Architecture; Sleep Oscillations and PGO Waves; Sleep Oscillations; Sleep-Dependent Memory Processing; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Braun AR, Balkin TJ, Wesensten NJ, et al. (1998) Dissociated pattern of activity in visual cortices and their projections during human rapid eye movement sleep. *Science* 279(5347): 91–95.

Dang-Vu TT, Desseilles M, Albouy G, et al. (2005) Dreaming: A neuroimaging view. *Swiss Archives of Neurology and Psychiatry* 156(8): 415–425.

Dang-Vu TT, Desseilles M, Laureys S, et al. (2005) Cerebral correlates of delta waves during non-REM sleep revisited. *Neuroimage* 28(1): 14–21.

Dang-Vu TT, Desseilles M, Peigneux P, and Maquet P (2006) A role for sleep in brain plasticity. *Pediatric Rehabilitation* 9(2): 98–118.

Maquet P (2000) Functional neuroimaging of normal human sleep by positron emission tomography. *Journal of Sleep Research* 9(3): 207–231.

Maquet P (2001) The role of sleep in learning and memory. *Science* 294(5544): 1048–1052.

Maquet P, Degueldre C, Delfiore G, et al. (1997) Functional neuroanatomy of human slow wave sleep. *Journal of Neuroscience* 17(8): 2807–2812.

Maquet P, Laureys S, Peigneux P, et al. (2000) Experience-dependent changes in cerebral activation during human REM sleep. *Nature Neuroscience* 3(8): 831–836.

Maquet P, Peters J, Aerts J, et al. (1996) Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. *Nature* 383(6596): 163–166.

Maquet P, Ruby P, Maudoux A, et al. (2005) Human cognition during REM sleep and the activity profile within frontal and parietal cortices: A reappraisal of functional neuroimaging data. *Progress in Brain Research* 150: 219–227.

Maquet P, Smith C, and Stickgold R (2003) *Sleep and Brain Plasticity*. Oxford: Oxford University Press.

Peigneux P, Laureys S, Fuchs S, et al. (2001) Generation of rapid eye movements during paradoxical sleep in humans. *Neuroimage* 14(3): 701–708.

Peigneux P, Laureys S, Fuchs S, et al. (2004) Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron* 44(3): 535–545.

Rauchs G, Desgranges B, Foret J, and Eustache F (2005) The relationships between memory systems and sleep stages. *Journal of Sleep Research* 14(2): 123–140.

Schwartz S and Maquet P (2002) Sleep imaging and the neuropsychological assessment of dreams. *Trends in Cognitive Sciences* 6(1): 23–30.

Sleep and Sleep States: Phylogeny and Ontogeny

J A Lesku, D Martinez-Gonzalez, and N C Rattenborg, Max Planck Institute for Ornithology, Seewiesen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sleep is a prominent yet enigmatic animal behavior. Despite the apparent ubiquity of sleep and associated vulnerability that (in part) defines sleep, the functions of sleep remain elusive. The seemingly simplest approach to determining sleep's functions would be to compare animals that sleep with those that do not and then determine whether a relationship exists between various traits thought to be functionally involved in sleep and the presence or absence of sleep. However all animals adequately studied sleep in one form or another making such comparisons currently impossible. Nevertheless, of the 30 or more animal phyla, detailed sleep information is known for only two: Chordata (includes vertebrates) and Arthropoda (includes insects). Thus, there is an opportunity for truly sleepless animals to be discovered in the future. For example, sponges (phylum: Porifera) are nearest to the base of the Metazoan phylogenetic tree and lack a nervous system. Based on the generally accepted belief that a plastic nervous system is the biological target benefiting from sleep, sponges should not sleep.

Determining whether sleep evolved many independent times or only once early in the evolution of animals is of fundamental importance in contemporary comparative sleep research. With the application of genetic approaches to the study of sleep, it has become important to establish homology between sleep in mammals and sleep in invertebrates, such as the fruit fly (*Drosophila melanogaster*), where genomes generally have less redundancy and are amenable to manipulation. Establishing homology between sleep in fruit flies and sleep in mammals would reinforce the usefulness of invertebrate models for determining the molecular correlates of human sleep. Even without homology at the molecular level, comparative studies can enhance our understanding of sleep. For instance, evolutionary convergence (i.e., distantly related species that independently evolved similar sleep-related traits) can provide insight into sleep by revealing overriding principles otherwise obscured by nonessential traits specific to one lineage.

The lack of a highly resolved cladogram of sleep across Animalia has not impeded comparative analyses of sleep. Indeed, many studies have tried to determine the relationships between the time species spend asleep and various constitutive, physiological, and ecological traits. Historically, this approach has been applied only in mammals, in which the electrophysiological correlates of sleep have been identified and quantified in a large number of species; however, a recent study provides insight into the correlates of sleep in birds as well. Although the strengths of functional hypotheses derived from comparative analyses are necessarily limited given the correlational nature of the data, the evolutionary patterns gleaned from such analyses, especially those that employ modern phylogenetic statistical techniques, provide a framework for the development of experimentally testable hypotheses.

As with comparing sleep among taxonomic groups, changes in sleep occurring during early development can provide insight into the functions of sleep. For example, changes in sleep duration, intensity, or the relative proportion of the two sleep states in mammals and birds over early ontogeny might be functionally linked to the concurrent development of the central nervous system. In this article, we review current knowledge of sleep in various taxonomic groups and over early ontogeny, and we relate important patterns to existing hypotheses for the functions of sleep. By doing so, we demonstrate the contribution that comparative sleep research has made to our greater understanding of why we sleep.

Behavioral Definition of Sleep

Sleep is foremost an animal behavior broadly characterized by quiescence with reduced responsiveness to stimuli. Sleep can be distinguished from other quiescent states (e.g., hibernation) by rapid reversibility to wakefulness with sufficient stimulation and homeostatic regulation; that is, sleep shows a compensatory increase in intensity or time following sleep loss. Sleeping animals often retreat to a species-specific location and assume a characteristic posture. Although the occurrence of concurrent sleep and swimming in some aquatic mammals is an exception to the quiescence criterion, these criteria apply to the vast majority of animals studied, ranging from insects to mammals and birds.

Mammalian Sleep

Behavioral sleep is associated with specific changes in brain activity. Placental and marsupial (or therian) mammals exhibit two basic states of sleep: slow-wave sleep (SWS), also known as non-rapid eye movement

(non-REM) sleep, and REM sleep. In nonhuman animals, the term SWS usually refers to all non-REM sleep, whereas SWS refers only to stages 3 and 4 of non-REM sleep in humans. SWS is characterized by an electroencephalogram (EEG) of low-frequency, high-amplitude activity arising from the large-scale synchronous slow oscillations of neurons in the neocortex. SWS is homeostatically regulated with SWS-related slow-wave activity (SWA) (typically 0.5–4.0 Hz EEG power density) reflecting the intensity of SWS, as arousal thresholds are correlated with the amount of SWA and SWA increases as a function of prior time awake. REM sleep is characterized by an EEG of high-frequency, low-amplitude activity similar to wakefulness, with a hippocampal theta rhythm (between 4 and 9 Hz or higher depending on the species) observed in some mammals, and an atonic electromyogram. Heart and respiratory rate are irregular during REM sleep, and thermoregulatory mechanisms that rely on motor control are diminished. REM sleep does not appear to have an intensity component, although the time spent in REM sleep can increase following sleep loss. In mammals that engage in extended periods of wakefulness, the time spent in SWS is the greatest and SWS-related SWA is the highest early in the subsequent sleep bout, whereas the proportion of time devoted to REM sleep increases toward the end of the sleep bout.

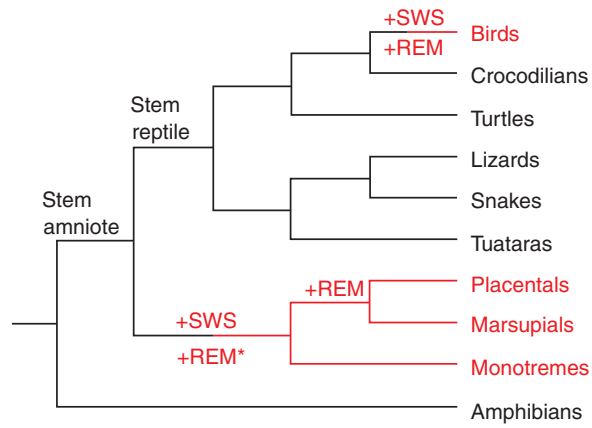

Figure 1 Cladogram for amniotes showing the evolutionary appearance of slow-wave sleep (SWS) and rapid eye movement (REM) sleep. Here, SWS and REM sleep arose independently in the most recent common ancestor to mammals and the most recent common ancestor to birds. However, in mammals, REM sleep first appeared in the common ancestor to all mammals as a heterogeneous state with neuronal activity in the brain stem indicative of REM sleep (+REM*) occurring concurrently with EEG activity indicative of SWS. In the most recent common ancestor to placental and marsupial mammals, REM sleep and SWS became segregated into two distinct states, with EEG activation occurring in conjunction with REM sleep-related brain stem activity. Evidence suggests that reptiles engage in SWS, but without the EEG slow-wave activity similar to that observed in mammals and birds.

Monotremes

The study of sleep in monotremes, an egg-laying order of mammals that are the closest extant relatives to therian mammals (Figure 1), may provide insight into sleep in the most recent common ancestor to therian mammals. An early study of the short-beaked echidna (*Tachyglossus aculeatus*) found only EEG activity indicative of SWS during sleep, thereby suggesting that REM sleep evolved after the appearance of the therian lineage. The electrophysiological correlates of sleep in the echidna were reexamined using a combination of EEG and brain stem neuronal recordings. As in the earlier study, REM sleep with cortical activation was not observed. However, during sleep with cortical slow waves, brain stem reticular neurons fired in an irregular pattern similar to that observed in placental mammals during REM sleep. This suggests that aspects of REM sleep in the brain stem and SWS in the cortex occur concurrently. Moreover, an EEG-based study of the duck-billed platypus (*Ornithorhynchus anatinus*) suggests that this mixed sleep state is typical of monotremes (Figure 1). Although neuronal activity in the brain stem was not recorded, the platypus showed frequent rapid eye movements and twitching of the head and bill, similar to that associated with REM sleep in therian mammals, whereas the cortex exhibited an EEG pattern indicative of SWS. Based on the incidence of twitches, the time spent in REM sleep was estimated at up to 8 h per day, the highest in any animal. A comparison of sleep times with therian mammals is difficult to interpret, however, given the heterogeneous nature of sleep and the absence of cortical correlates of REM sleep in monotremes.

Aquatic Mammals

Sleeping in the water poses a significant challenge for air-breathing mammals. Among Cetaceans, the electrophysiological correlates of sleep have been recorded only in the Odontocetes (dolphins and porpoises). Dolphins and porpoises exhibit unihemispheric SWS (USWS), a mixed state in which one cerebral hemisphere shows EEG activity characteristic of deep SWS while the other shows a pattern indistinguishable from wakefulness (Figure 2). Although the lighter stages of SWS can occur concurrently in both hemispheres, deep SWS occurs only unihemispherically. Evidence suggests that SWS is homeostatically regulated independently in the two hemispheres, which is an example of regional use-dependent SWS homeostasis, a phenomenon discovered only recently in terrestrial mammals including humans. Dolphins and porpoises swim and

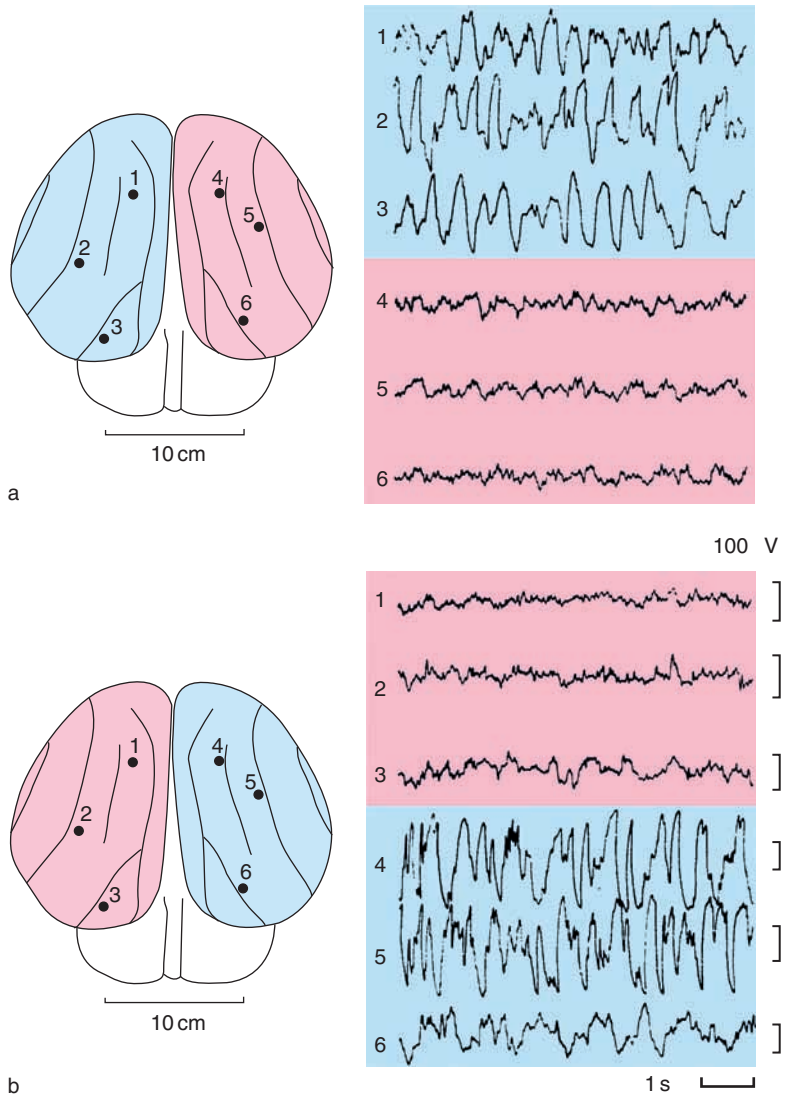

Figure 2 Examples of unihemispheric slow-wave sleep in the bottle-nosed dolphin (*Tursiops truncatus*). The EEG was recorded from the anterior, medial, and posterior neocortex. Note the high-amplitude, low-frequency activity indicative of slow-wave sleep (blue) in only the left (a) or the right (b) hemisphere concurrent with low-amplitude, high-frequency activity indicative of wakefulness (red) in the other hemisphere. Reproduced from Mukhameotv LM, Supin AY, and Polyakova IG (1977) Interhemispheric asymmetry of electroencephalographic sleep patterns in dolphins. *Brain Research* 134: 581–584, with permission from Elsevier.

surface to breath during USWS, but they can also float at the surface or rest motionless underwater. Given that the eye opposite the awake hemisphere remains open, USWS may allow dolphins and porpoises to monitor the environment for conspecifics and predators during sleep. Although behavioral signs of REM sleep (e.g., twitching and rapid eye movements) have been observed infrequently in quiescent Cetaceans, unequivocal REM sleep has not been recorded electrophysiologically, suggesting that Cetaceans have lost REM sleep secondarily. Nevertheless, REM sleep could occur in very small amounts or in a modified form in Cetaceans. Interestingly, manatees (order: Sirenia) also exhibit USWS and small amounts of REM sleep. Unlike Cetaceans and manatees, seals can sleep in the water and on the land. Seals in the family Phocidae hold their breath during periods of bilateral SWS and REM sleep underwater, whereas seals in the family Otariidae show interhemispheric asymmetries in the intensity of SWS while sleeping on the surface of the water.

Early Ontogeny of Mammalian Sleep

The time spent asleep changes greatly over early development. In mammals, newborns generally sleep the longest. This high amount of sleep declines over early ontogeny until it stabilizes at a species-specific level similar to that seen in adults. Sleep homeostasis

has been demonstrated in postnatal day 5 rats, before SWS and REM sleep appear in the EEG as differentiated states. Evidence suggests that SWS and REM sleep actually differentiate before the EEG shows sleep-related changes in cortical activity. The proportion of total sleep time devoted to REM sleep (%REM sleep) and SWS also changes greatly during the first postnatal weeks. Interestingly, the magnitude of these changes appears to be influenced by the degree of precociality at birth (Figure 3). Altricial mammals (those that are relatively more dependent on their parents at birth for food, warmth, and protection; e.g., rats and cats) show a marked reduction in %REM sleep through the first postnatal weeks (Figure 3) and show more exaggerated behavioral characteristics of REM sleep, such as rapid eye movements and twitches. Conversely, newborn precocial mammals (e.g., guinea pigs) are more developed relative to altricial mammals, and they exhibit a %REM sleep comparable with that of adults (Figure 3). These observations led (in part) to the hypothesis that REM sleep provides endogenous stimulation necessary for the early development of the central nervous system, particularly that related to the visual system. Indeed, experimental evidence implicates REM sleep in maturational processes of the visual system during early ontogeny. As discussed later, phylogenetic comparative analyses also support the idea that REM sleep facilitates early brain development.

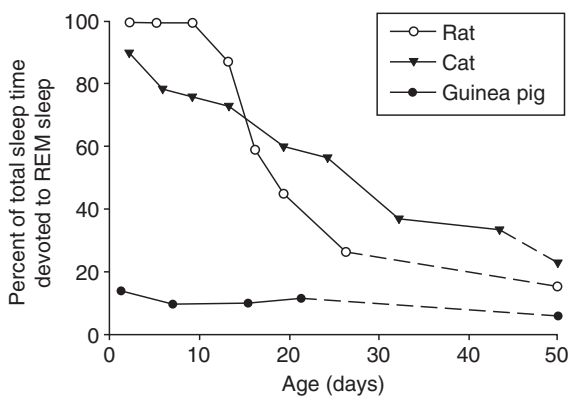

Figure 3 Changes in the percentage of total sleep time devoted to rapid eye movement (%REM) sleep across early development in the rat, cat, and guinea pig. In the altricial rat and cat, REM sleep constitutes most of total sleep time and decreases to levels close to that of adults approximately 1 month after birth. Conversely, %REM sleep in the precocial guinea pig is relatively stable over early ontogeny and into adulthood. Note, however, that even as adults, altricial species have higher %REM sleep than more precocial species. Reprinted Jouvet-Mounier D, Astic L, and Lacote D (1970) Ontogenesis of the states of sleep in rat, cat, and guinea pig during the first postnatal month. *Developmental Psychobiology* 2: 216–239, with permission from John Wiley & Sons, Inc.

Unlike newborn terrestrial mammals, newborn Cetacean calves are continuously active, swimming alongside their mothers during the first few weeks postpartum. Although electrophysiological recordings have not been obtained during this period of activity, the calves close one eye intermittently while swimming underwater and, therefore, might be engaging in USWS. However, calves are unlikely to be engaging in REM sleep if REM sleep is incompatible with swimming. The possible absence of REM sleep in newborn calves during their first weeks of life seemingly challenges theories for a functional role of REM sleep in brain development; however, given that Cetacean calves are extremely precocial, REM sleep could nonetheless play a role in brain development *in utero*.

Avian Sleep

Birds are a particularly interesting taxonomic group in which to study sleep. Birds have brains comparable in relative size to those of mammals and cognitive abilities that include vocal learning and tool making; however, much of the avian forebrain is organized in a manner markedly different from mammals. Although the dorsal two-thirds of the avian forebrain (a region formerly thought to be primarily striatal) is derived from the same embryonic neural tissue (the pallium) that gives rise to the mammalian neocortex, the avian pallium is arranged in a nuclear manner that lacks the true laminar organization of the neocortex (Figure 4). Interestingly, despite this difference in pallial organization, birds are the only nonmammalian taxonomic group to show unequivocal SWS and REM sleep (Figure 1).

Like mammalian SWS, avian SWS is characterized by an EEG of low-frequency, high-amplitude activity. Birds that sleep predominantly at night (e.g., domestic hens, *Gallus domesticus*; European blackbirds, *Turdus merula*; nonmigrating white-crowned sparrows, *Zonotrichia leucophrys gambelii*) show slow-wave activity during SWS that is greatest early in the night and gradually declines thereafter in a manner suggestive of mammalian-like SWS homeostasis. Indeed, we recently demonstrated that pigeons (*Columba livia*) show a compensatory increase in SWS-related low-frequency EEG power density following short-term sleep deprivation, suggesting that mammalian and avian sleep is regulated in a similar manner.

Avian REM sleep also shares several features in common with mammals. REM sleep is associated with closure of both eyes, rapid eye movements, occasional bill movements, and behavioral signs of reduced muscle tone such as dropping of the head. However, muscle atonia has been observed only in

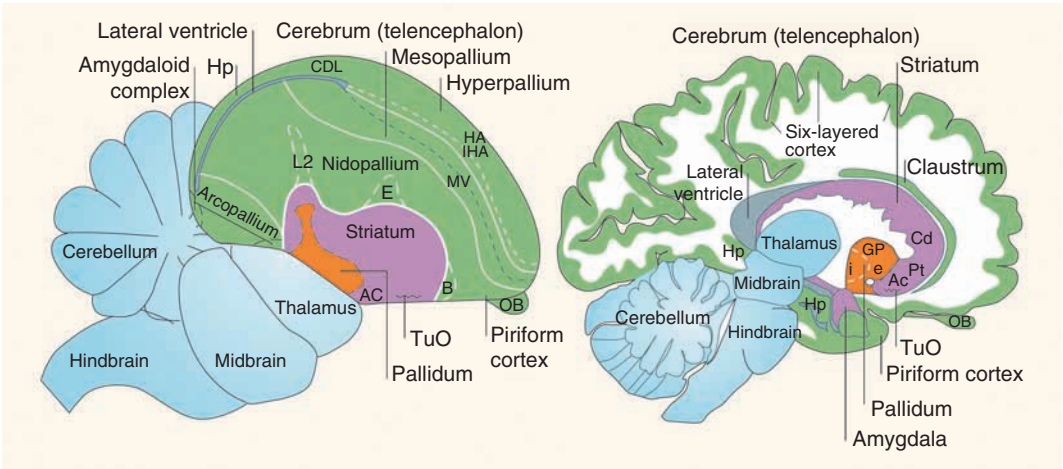

Figure 4 Modern consensus view of avian and mammalian brain relationships. Historically, much of the avian telencephalon was thought to be homologous to the mammalian striatum. However, converging lines of evidence demonstrate that the dorsal two-thirds of the avian telencephalon is actually derived from the pallium (green), the same neural tissue that gives rise to the mammalian neocortex. The new nomenclature reflecting this fundamental change in our understanding of brain evolution is depicted in the sagittal view of a zebra finch (left) and human (right) brain. Ac, accumbens; B, basorostralis; Cd, caudate nucleus; CDL, dorsal lateral corticoid area; E, entopallium; GP, globus pallidus (i, internal segment; e, external segment); HA, hyperpallium apicale; Hp, hippocampus; MV, mesopallium ventrale; IHA, interstitial hyperpallium apicale; L2, field L2; LPO, lobus parolfactorius; OB, olfactory bulb; Pt, putamen; TuO, olfactory tubercle. The several large, white regions are axon pathways (i.e., white matter) in the cerebrum. Lamina (cell-sparse zones separating brain subdivisions) are marked as solid white lines, primary sensory neuron populations are distinguished from neighboring regions by dashed white lines, and regions differing in cell density or size are demarcated by dashed gray lines. Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Neuroscience* (Jarvis E, Gunturkun O, Bruce L, et al., 2005, 6: 151–159), Copyright 2005.

birds that can securely rest their head on the back. As in mammals, thermoregulatory responses are diminished during REM sleep. During REM sleep, the EEG reverts to a pattern similar to wakefulness, but often with lower amplitude. Unlike mammalian REM sleep, however, a hippocampal theta rhythm has not been found in birds. Episodes of REM sleep typically last 2–10 s and occur in clusters. The short duration of REM sleep episodes does not appear to be related to a need to maintain balance because REM sleep episodes are equally short when birds are sitting with their heads supported on their backs. In many birds, REM sleep increases throughout the night, in a manner similar to humans. In pigeons, this reflects an increase in the incidence and the duration of REM sleep episodes across the night. The proportion of total sleep time devoted to REM sleep appears to be lower in bird species (mean, 8%) than in mammalian species (mean, 17%). As in mammals, the time spent in REM sleep increases following sleep deprivation. Finally, since REM sleep and SWS have been recorded in every avian species investigated, both states were likely present in the most recent common dinosaur ancestor to birds (Figure 1).

Like aquatic mammals, birds often sleep with one eye open, a behavioral state associated with SWS in the hemisphere opposite the closed eye and EEG activity intermediate between wakefulness and SWS in the hemisphere opposite the open eye. Birds have the ability to switch from sleeping with both eyes closed to sleeping with one eye open in response to a perceived increase in the risk of predation. Here, birds direct the open eye toward the potential threat and are able to respond to threatening stimuli presented to the open eye. In contrast to mammals, where unihemispheric SWS occurs only in aquatic mammals, such interhemispheric asymmetries in slow-wave activity are common in birds and may be an ancestral trait. As in Cetaceans that swim in a coordinated manner during unihemispheric SWS, sleeping with one eye open and half the brain awake may allow birds to sleep during flight. Although there is strong evidence showing that birds, such as common swifts (*Apus apus*) and frigatebirds (*Fregata* sp.), spend periods lasting days to weeks or longer in constant flight, sleep in flight has not been confirmed with electrophysiological recordings.

Early Ontogeny of Avian Sleep

The ontogenetic development of avian sleep has been studied in relatively few species. In precocial chickens (*Gallus gallus*), the EEG correlates of SWS and REM sleep can be distinguished 1 day before hatching, whereas in the altricial pigeon (*Columba livia*) EEG activity is absent at hatching and does not resemble that of adults until 14 days later. Although it is

unclear whether chicken chicks show more or less REM sleep than adults shortly after hatching, the proportion of sleep allocated to REM sleep seems to reach adult levels within the first week post-hatch. Among altricial birds, the ontogeny of SWS and REM sleep has only been examined in magpies (*Pica pica*), in which juveniles sleep longer and have more REM sleep than adults. Additional studies on early sleep ontogenesis in species of greater taxonomic diversity are needed to determine whether the ontogenetic changes in sleep in altricial and precocial birds parallel those observed in mammals. Interestingly, chicken chicks that imprint on a moving object on the day of hatching exhibit more REM sleep during the first 3 h following imprinting than nonimprinted control chicks, suggesting that REM sleep may play a role in this acute form of learning. Sleep also appears to play a role in song learning in juvenile zebra finches (*Taeniopygia guttata*).

Sleep in Reptiles

The presence of SWS and REM sleep in all mammalian and avian species investigated suggests that either these states were present in the most recent common ancestor to extant mammals, birds, and reptiles or they evolved independently in mammals and birds. Several studies have attempted to distinguish between these alternatives by examining the electrophysiological correlates of sleep behavior in reptiles and, to a lesser extent, amphibians and fish. Unlike mammals and birds, however, where largely similar results have been obtained across species and laboratories, the results from reptiles have been less consistent and therefore subject to more diverse interpretations. Although some controversy persists, the EEG during reptilian sleep behavior typically shows intermittent high-voltage spikes arising from a background pattern similar to or slightly reduced in amplitude compared to quiet wakefulness. Because the incidence of spikes is correlated with arousal thresholds and increases following sleep deprivation, spikes appear to reflect sleep intensity. Studies have shown that these spikes originate in the reptilian hippocampus, thereby corroborating earlier pharmacological evidence indicating that reptilian spikes are comparable to similar spikes occurring in the mammalian hippocampus during SWS. Despite this similarity at the level of the hippocampus, however, the reptilian dorsal cortex does not generate concurrent high-amplitude slow waves typical of SWS in mammals and birds.

REM sleep has been reported in reptiles and fish based on the presence of eye, head, and limb movements during sleep. However, it remains unclear whether these behaviors truly reflect REM sleep-related twitching similar to that observed in mammals and birds or partial arousals from sleep. Although the presence of brain stem neural activity suggestive of REM sleep in sleeping echidnas raised the possibility that reptiles also exhibit REM sleep at the level of the brain stem, no sign of REM sleep was detected in the brain stem of sleeping turtles, despite the presence of neural structures involved in generating REM sleep in mammals. The presence of unequivocal SWS and REM sleep in mammals and birds, but not in reptiles, amphibians, or fish, suggests that these sleep states arose independently in the mammalian and avian lineages through convergent evolution (**Figure 1**).

Corticocortical Connectivity and SWS

Historically, the absence of slow waves in the EEG of sleeping reptiles has been attributed to the lack of a thick cortex similar to that which generates slow waves in mammals. However, the presence of slow waves in sleeping birds, despite the absence of a neocortex (**Figure 4**), demonstrates that the neocortex is not essential for the genesis of EEG slow waves. Instead, the extent of connections within the mammalian neocortex, avian hyperpallium, and reptilian dorsal cortex may explain why slow waves are present during sleep in mammals and birds but not reptiles (**Figure 5**). In mammals, corticocortical connections in layers II and III play an integral role in synchronizing the slow oscillation of neurons in a manner sufficient to generate slow waves in the EEG. In accord with the absence of slow waves in sleeping reptiles, the three-layered reptilian dorsal cortex lacks layers II and III, and it shows limited corticocortical connectivity (**Figure 5**). Furthermore, although birds lack a true neocortex, the hyperpallium shows extensive interconnectivity (**Figure 5**). Thus, the occurrence of sleep-related slow-wave activity in amniotes seems to be related to the extent of interconnectivity in the neocortex, hyperpallium, and dorsal cortex, although additional factors may also play a role.

A persistent question in sleep research is whether the EEG correlates of sleep are involved in the functions of sleep or simply reflect an epiphenomenon of the state. For example, the presence of slow waves in the EEG of sleeping mammals and birds may simply be an emergent property of a heavily interconnected neocortex and hyperpallium. Alternatively, as suggested by experimental work in mammals, the corticocortical connections that give rise to slow waves may also depend on slow waves to maintain the level of connectivity at an energetically and functionally adaptive level. Experimental evidence indicates that slow waves may also be involved in sleep-dependent memory processing and plasticity. Additional studies are needed, however, to determine whether slow

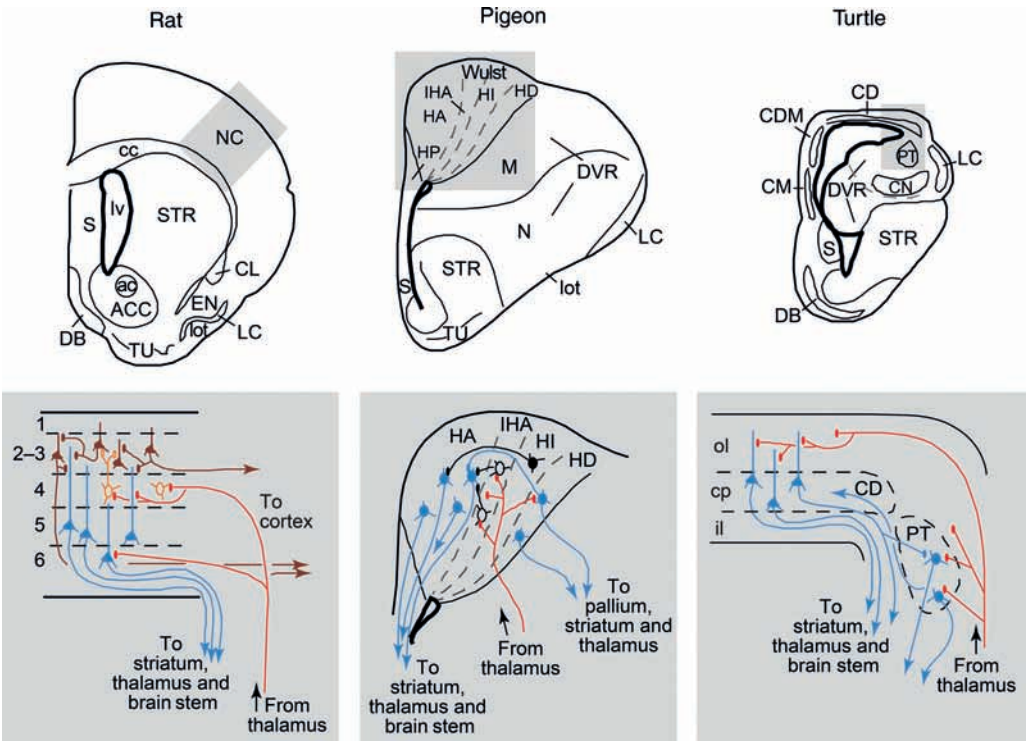

Figure 5 Comparison of the dorsal pallium in representative mammalian, avian, and reptilian species. The top row shows coronal cross sections through the brain. The area shaded in gray is expanded in the bottom row to show network connections. Note the comparatively high degree of interconnectivity in the rat neocortex and the pigeon Wulst (i.e., hyperpallium) compared to the turtle dorsal cortex. ac, anterior commissure; ACC, nucleus accumbens; cc, corpus callosum; CD, dorsal cortex; CDM, dorsomedial cortex; CL, claustrum; CM, medial cortex; CN, core nucleus of the DVR; cp, cell plate; DB, diagonal band of Broca; DVR, dorsal ventricular ridge; EN, endopiriform region; HA, hyperpallium apicale; HD, hyperpallium densocellulare; HI, hyperpallium intercalatum; HP, hippocampus; IHA, nucleus interstitialis hyperpallii apicalis; il, inner layer; LC, lateral cortex; lot, lateral olfactory tract; lv, lateral ventricle; M, mesopallium; N, nidopallium; NC, neocortex; ol, outer layer; PT, pallial thickening; S, septum; STR, striatum; TU, olfactory tubercle. Reproduced from Medina L and Reiner A (2000) Do birds possess homologues of mammalian primary visual, somatosensory and motor cortices? *Trends in Neurosciences* 23: 1–12, with permission from Elsevier.

waves evolved independently in mammals and birds (and not in reptiles) to maintain their heavily interconnected brains and associated cognitive abilities, or whether similar processes occur during reptilian sleep in a manner undetectable in the EEG. As indicated by the presence of hippocampal spikes during sleep, at least some of the neural correlates of SWS seem to be present in reptiles (Figure 1). A comparison of sleep-related gene expression among mammals, birds, and reptiles may clarify the evolutionary history of SWS and REM sleep.

Comparative Perspectives on the Functions of Sleep

More is known about sleep in mammals than in any other group of animals. Electrophysiological sleep data exist for just less than 100 mammalian species. An examination of this comparative data set reveals that the time spent in SWS and REM sleep varies greatly across Mammalia. Many studies have tried to identify the evolutionary determinants responsible for maintaining such interspecific variation in the structure of sleep. The identification of significant predictors of sleep times may help shed light on the functions of SWS and REM sleep. For example, species with a higher mass-specific basal metabolic rate (i.e., basal metabolic rate per gram of body mass) were once thought to engage in more SWS, a relationship that could support an energy conservation role for SWS, because metabolic rates are lower during SWS. Additionally, it was once thought that species with greater encephalization allocated a lower proportion of time spent sleeping to REM sleep, seemingly refuting a neurophysiological role for REM sleep, such as memory processing and plasticity. However, excluding our own analysis, all comparative studies of sleep treated each species (or, in one case, family) as a statistically independent unit. Independence is a basic assumption of all statistical analyses, but species cannot be considered independent because they are related to one another through common

ancestry. Phylogenetic comparative methods (e.g., independent contrasts) were developed by evolutionary biologists to deal with such interspecific relatedness in comparative studies; however, until recently, these procedures had not been used in comparative sleep research.

In our analysis, we controlled for shared evolutionary history among species using independent contrasts and analyzed our phylogenetically controlled data using a multivariate analysis that incorporated hypotheses for the functions of SWS and REM sleep. Many of our results were different from those in previous studies. Unlike all previous studies, we found that species with greater encephalization (i.e., brain mass controlling for body mass using regression) allocate a greater proportion of time asleep to REM sleep, thus providing the first comparative support for a neurophysiological role for REM sleep. Although no relationship was found between encephalization and the time spent in SWS, cumulative slow-wave activity may be the more accurate measure of SWS. Contrary to some comparative studies and expectations under the energy conservation hypothesis, we found that species with a higher residual basal metabolic rate (BMR) engage in less SWS and sleep less altogether. These relationships might reflect increased foraging demands associated with higher residual BMR and thus less time available for sleep. Nevertheless, mammals with higher residual BMRs may obtain functionally comparable amounts of SWS by engaging in more intense SWS, although such intensity data are largely unavailable. Alternatively, the restorative processes occurring during sleep could be achieved more quickly with higher residual BMR and thus take less time to accomplish.

Predation risk should be among the strongest selection pressures influencing how to structure sleep and how long to sleep because sleeping is dangerous. Interestingly, the vulnerability associated with sleep may depend on the sleep state. For example, due to high arousal thresholds, deep SWS and REM sleep may be particularly dangerous sleep states from an antipredator standpoint. Indeed, in the context of our multivariate models, we found that species sleeping at relatively exposed or risky locations in the wild engage in less REM sleep and allocate a lower proportion of sleep time to REM sleep in the laboratory. Although the time spent in SWS was largely independent of predation risk, once again, SWS intensity may respond more strongly to the risk of predation. As per previous studies, we found that species more precocial at birth engage in less REM sleep as adults, a relationship that was not mediated strongly by predation risk because precocial species did not generally sleep in more vulnerable locations than relatively altricial species. The degree of precociality as a predictor of REM sleep time in adults may reflect an extension of the ontogenetic changes in REM sleep shown in **Figure 3**, which supports the idea that REM sleep is particularly important in early brain development. However, why high levels of REM sleep during early ontogeny should persist in adults remains enigmatic.

Correlates of Avian SWS and REM Sleep

As discussed previously, the electrophysiological correlates of avian sleep are remarkably similar to those observed in mammals. In addition to these similarities, there is also great variation in the time avian species spend in SWS and REM sleep. Thus, an obvious question is whether birds share the same evolutionary determinants of SWS and REM sleep as mammals? We addressed this question by conducting the first electrophysiologically based comparative analysis of avian sleep architecture using the same phylogenetically controlled variables as in our mammalian analysis. Overall, we found that birds that sleep at relatively exposed sites in the wild engage in less SWS in the laboratory, but this was the only significant relationship identified. Thus, if relationships identified in mammals reflect functional aspects of sleep architecture, then the same functions may not apply broadly to birds.

Sleep in Invertebrates

More than 97% of all animal life is invertebrates, of which 80% are arthropods, the taxonomic group which includes insects. Surprisingly, however, of the 30 or more animal phyla, sleep has been studied in only 2: Chordata (includes vertebrates) and Arthropoda (includes insects), with much work having been focused on the fruit fly (*Drosophila melanogaster*) and the honeybee (*Apis mellifera*). Sleep in honeybees is characterized as a sustained period of quiescence accompanied by increased arousal thresholds and specific postures (e.g., antennal immobility). Optomotor interneurons in the optic lobes of forager honeybees show lowered sensitivity during the subjective night, when forager bees are often quiescent, than during the subjective day. Sleep in honeybees was shown to be homeostatically regulated as antennal immobility increased following 12 h of sleep deprivation. Sleep has also been demonstrated in scorpions (*Heterometrus* and *Pandinus* spp.), cockroaches (*Blaberus giganteus* and *Leucophaea maderae*), and crayfish (*Procambarus clarkii*). Also, preliminary work suggests that sleep is present in three additional animal phyla: Nematoda (*Caenorhabditis elegans*), cephalopods (*Octopus vulgaris* and *Sepia pharaonis*) in the phylum Mollusca, and box jellyfish (*Chironex fleckeri*) in the phylum Cnidaria.

Conclusions

Our understanding of how and why we sleep has been enhanced by studies examining the evolution and ontogeny of sleep. The presence of SWS and REM sleep in all avian and mammalian species studied, and their apparent absence in reptiles, suggests that these states arose independently twice: once in the ancestor to birds and once in the ancestor to mammals (Figure 1). A comparison of neurocytoarchitecture among homeotherms suggests that the degree of corticocortical (or palliopallial) connectivity is responsible for EEG slow-wave activity in mammals and birds. Interestingly, REM sleep-related cortical activation evolved independently in the ancestor to therian mammals and the ancestor to birds (Figure 1). The similarity in the electrophysiological correlates of behavioral sleep in homeotherms suggests similarities in functions, possibly related to having heavily interconnected brains and associated cognitive abilities. However, in phylogenetically controlled comparative analyses, the evolutionary determinants of sleep times in birds were markedly different from those identified in mammals. Identifying the reasons for these differences is a promising area for future work.

Our understanding of the evolution, ontogeny, and functions of sleep would benefit greatly from studies in species representing broader phylogenetic diversity. Here, the study of species nearest the base of the Metazoan phylogenetic tree would be most revealing. A systematic approach to selecting taxa for study should be employed that takes into consideration neuroanatomy, neurophysiology, and phylogenetic position in the context of current hypotheses for the functions of sleep. Future work should also expand existing genetic work on fruit flies and rats to nonmammalian vertebrate taxa and to nonarthropod invertebrate taxa. The success of these endeavors will depend on the broad collaboration of animal behaviorists, evolutionary biologists, geneticists, and neurophysiologists, but will do much to aid our greater understanding of sleep.

See also: Circadian Function and Therapeutic Potential of Melatonin in Humans; Mammalian Sleep and Circadian Rhythms: Flies; Sleep Architecture; Sleep Oscillations and PGO Waves; Sleep and Sleep States: Hippocampus–Neocortex Dialog; Sleep and Sleep States: Gene Expression; Sleep and Waking in Drosophila.

Further Reading

Amlaner CJ and Ball NJ (1994) Avian sleep. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 2nd edn., pp. 81–94. Philadelphia: Saunders.

Eiland MM, Lyamin OI, and Siegel JM (2001) State-related discharge of neurons in the brainstem of freely moving box turtles, *Terrapene carolina major. Archives Italiennes de Biologie* 139: 23–36.

Hartse KM (1994) Sleep in insects and nonmammalian vertebrates. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 2nd edn., pp. 95–104. Philadelphia: Saunders.

Jarvis E, Gunturkun O, Bruce L, et al. (2005) Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience* 6: 151–159.

Jouvet-Mounier D, Astic L, and Lacote D (1970) Ontogenesis of the states of sleep in rat, cat, and guinea pig during the first postnatal month. *Developmental Psychobiology* 2: 216–239.

Kaiser W and Steiner-Kaiser J (1983) Neuronal correlates of sleep, wakefulness and arousal in a diurnal insect. *Nature* 301: 707–709.

Lesku JA, Roth TC, Amlaner CJ, and Lima SL (2006) A phylogenetic analysis of sleep architecture in mammals: The integration of anatomy, physiology, and ecology. *American Naturalist* 168: 441–453.

Lesku JA, Roth TC, Rattenborg NC, Amlaner CJ, and Lima SL (2008) Phylogenetics and the correlates of mammalian sleep: A reappraisal. *Sleep Medicine Reviews* 12: 229–244.

Lima SL, Rattenborg NC, Lesku JA, and Amlaner CJ (2005) Sleeping under the risk of predation. *Animal Behavior* 70: 723–736.

Lyamin OI, Mukhametov LM, and Siegel JM (2004) Relationship between sleep and eye state in Cetaceans and Pinnipeds. *Archives Italiennes de Biologie* 142: 557–568.

Margoliash D (2005) Song learning and sleep. *Nature Neuroscience* 8: 546–548.

Martinez-Gonzalez D, Lesku JA, and Rattenborg NC (2008) Increased EEG spectral power density during sleep following short-term sleep deprivation in pigeons (*Columba livia*): Evidence for avian sleep homeostasis. *Journal of Sleep Research* 17: 140–153.

Medina L and Reiner A (2000) Do birds possess homologues of mammalian primary visual, somatosensory and motor cortices. *Trends in Neurosciences* 23: 1–12.

Mukhaneotv LM, Supin AY, and Polyokova IG (1977) Interhemispheric asymmetry of electroencephalographic sleep patterns in dolphins. *Brain Research* 134: 581–584.

Rattenborg NC (2006) Evolution of slow-wave sleep and palliopallial connectivity in mammals and birds: A hypothesis. *Brain Research Bulletin* 69: 20–29.

Rattenborg NC, Amlaner CJ, and Lima SL (2000) Behavioral, neurophysiological and evolutionary perspectives on unihemispheric sleep. *Neuroscience and Biobehavioral Reviews* 24: 817–842.

Roth TC, Lesku JA, Amlaner CJ, and Lima SL (2006) A phylogenetic analysis of the correlates of sleep in birds. *Journal of Sleep Research* 15: 395–402.

Siegel JM, Manger PR, Nienhuis R, Fahringer HM, and Pettigrew JD (1998) Monotremes and the evolution of rapid eye movement sleep. *Philosophical Transactions of the Royal Society Series B* 353: 1147–1157.

Szymczak JT, Kaiser W, Helb HW, and Beszczynska B (1996) A study of sleep in the European blackbird. *Physiology & Behavior* 60: 1115–1120.

Tobler I (1995) Is sleep fundamentally different between mammalian species? *Behavioural Brain Research* 69: 35–41.

Sleep and Sleep States: Thalamic Regulation

A Destexhe, Centre National de la Recherche Scientifique (CNRS), Gif-sur-Yvette, France
T J Sejnowski, Salk Institute for Biological Studies and University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During slow-wave sleep, brain activity is dominated by oscillations, such as delta oscillations and slower oscillations. The thalamus, together with the cortex, participates in generating and sustaining these oscillations. Delta waves were observed in the isolated cortex *in vivo* by Frost and colleagues in 1966, and delta-wave activity has been found in the isolated thalamus as well in thalamic slices, although these are more regular than delta waves *in vivo*. The latter studies demonstrated that this form of delta-wave activity is generated intrinsically by the interplay of I_T and I_h currents in thalamocortical (TC) neurons. Slow oscillations (<1 Hz) can be generated by the cortex; they were observed in cortical slices by Sanchez-Vives and McCormick in 2000. A similar type of oscillation has also been observed by the group of V Crunelli to be an intrinsic property of thalamic neurons when metabotropic receptors are stimulated.

Spindle oscillations, which are mostly visible in the early stages of sleep (stage II in humans), are the best-studied sleep oscillation and are generated by thalamic circuits under physiological conditions. In contrast to delta and slow oscillations, for which cortical generators also exist, spindles are generated exclusively in the thalamus. We focus here on sleep spindles, their genesis by thalamic networks, how these oscillations are distributed across the entire TC system, and how they are modulated.

In the human electroencephalogram (EEG), spindle oscillations are grouped in short 1–3 s periods of 7–14 Hz oscillations, organized within a waxing-and-waning envelope, that recur periodically every 10–20 s. In cats and rodents, spindle waves of similar characteristics appear during slow-wave sleep and are typically more prominent at sleep onset. They are enhanced by some anesthetics, such as barbiturates, which, when administered at an appropriate dose, generate an EEG dominated by spindles. In ferret thalamic slices, spindle oscillations occur spontaneously. This *in vitro* model of spindle waves has made it possible to precisely characterize the ionic mechanisms underlying spindle oscillations.

The first indication that spindles could originate outside the cerebral cortex was from Bishop in 1936, who showed that rhythmical activity was suppressed in the cerebral cortex following destruction of its connections with the thalamus. In 1938, Bremer showed that rhythmical activity is still present in the white matter following destruction of the cortical mantle. Later, Adrian in 1941 and Morison and Bassett in 1945 observed that spindle oscillations persist in the thalamus on removal of the cortex, providing firm experimental evidence that these oscillations originate in the thalamus. These experiments led to the development of the thalamic pacemaker hypothesis, according to which rhythmic activity is generated in the thalamus and communicated to the cortex, where it entrains cortical neurons and is responsible for the rhythmical activity observed in the EEG.

Thalamic Pacemakers for Sleep Spindle Oscillations

The first cellular mechanism for the genesis of spindle oscillations was proposed by Andersen and Eccles in 1962. From intracellular recordings from TC relay neurons during spindles, they reported that TC cells fired bursts of action potentials interleaved with inhibitory postsynaptic potentials (IPSPs). They suggested that TC cells fire in response to IPSPs (postinhibitory rebound), which was later demonstrated to be a characteristic electrophysiological feature of thalamic cells by Jahnsen and Llinas. Andersen and Eccles suggested that the oscillations arise from the reciprocal interactions between TC cells and inhibitory local-circuit interneurons. This mechanism was later incorporated into a computational model that provided a phenomenological description of the inhibitory rebound.

The mechanism proposed by Andersen and Eccles was seminal but not entirely correct because reciprocal connections between TC cells and thalamic interneurons have never been observed in anatomical studies. It was later found by Scheibel and Scheibel that the thalamic reticular (RE) nucleus, a sheetlike structure of inhibitory neurons surrounding the thalamus, could provide the inhibition of TC cells and that TC-RE loops could underlie the recruitment phenomena and spindle oscillations. They predicted that the output of the RE nucleus should be inhibitory and that the inhibitory feedback from RE cells onto TC cells should be critical for the genesis of thalamic rhythmicity. This hypothesis was supported by the observation that the pattern of firing of RE neurons was tightly correlated with IPSPs in TC neurons.

Several critical experiments by the group of Steriade firmly established the involvement of the RE nucleus in the generation of spindles in cats *in vivo*. The typical intracellular features of spindle oscillations in the two thalamic cell types is shown in Figure 1(a). By using intracellular or extracellular experiments, it was shown that (1) cortically projecting thalamic nuclei lose their ability to generate spindle oscillations if deprived of input from the RE nucleus and (2) the isolated RE nucleus can itself generate rhythmicity in the spindle frequency range.

Thus, thalamic rhythmicity can be explained by three different mechanisms: the TC interneuron loops of Andersen and Eccles, the TC-RE loops of Scheibel and Scheibel, and the RE pacemaker hypothesis of Steriade (although Steriade's work also demonstrated the importance of the cortex; see the section titled 'Dialog between thalamus and cortex: emergence of large-scale synchrony'). The introduction of an *in vitro* model of spindle waves in ferrets by McCormick's group (Figure 1(b)) supported the second of these mechanisms. Slices of the visual thalamus that contain the dorsal (lateral geniculate nucleus (LGN)) and reticular nuclei (perigeniculate nucleus (PGN)) as well as the interconnections between them generated spindles spontaneously, confirming earlier experimental evidence for the genesis of spindles in the thalamus.

The *in vitro* preparation allowed detailed pharmacological investigation of the ionic currents and synaptic receptors underlying the spindle oscillations. The spindle waves disappeared after the connections between TC and RE cells were severed, consistent with the mechanism based on intrathalamic TC-RE loops proposed by Scheibel and Scheibel. This *in vitro* experiment also confirmed the observation that the input from RE neurons is necessary to generate spindles, as found by Steriade's group in 1985. However, the RE nucleus maintained *in vitro* did not generate oscillations without connections with TC cells, in contrast with the observation of spindle rhythmicity in the isolated RE nucleus *in vivo* by Steriade's group in 1987.

Thus, *in vitro* experiments appear to support a mechanism by which oscillations are generated by the TC-RE loop, in contrast with the RE pacemaker hypothesis. Computational models suggested a way to

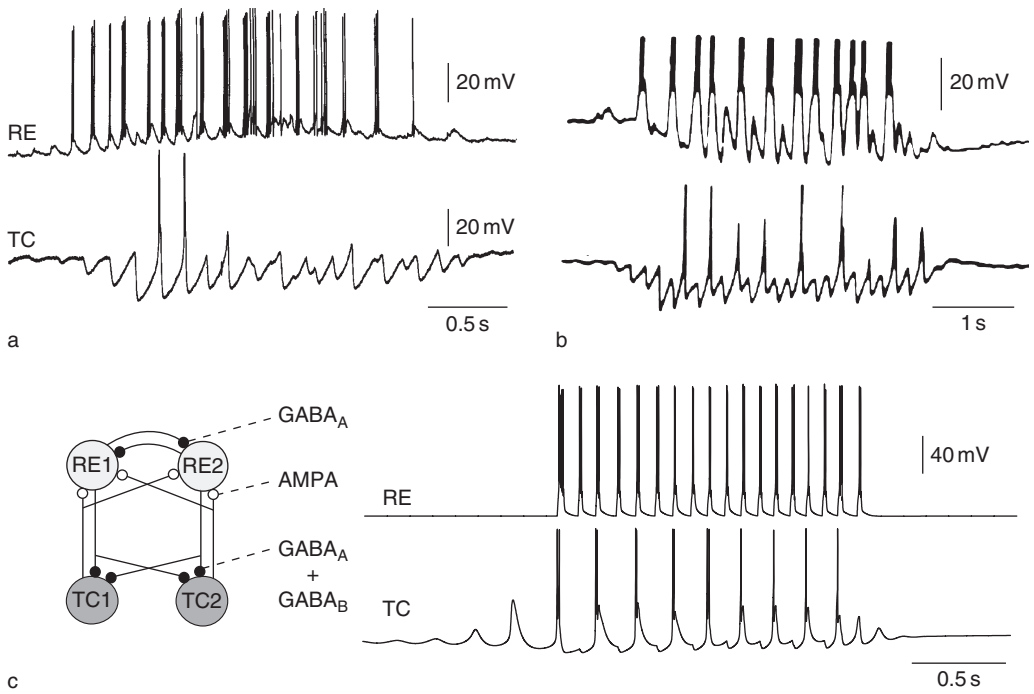

Figure 1 Spindle oscillations in thalamic circuits: (a) *in vivo* intracellular experiments in cats under barbiturate anesthesia, showing the activity of thalamocortical (TC) and thalamic reticular (RE) cells during spindle waves; (b) *in vitro* intracellular experiments in ferret visual thalamic slices, showing the activity of the same type of thalamic neurons during spindle waves; (c) computational model of spindle waves generated by TC-RE interactions. The left drawing in (c) shows a simple circuit, consisting of two TC and two RE cells interconnected with glutamatergic (AMPA) and GABAergic receptors (GABA$_A$ and GABA$_B$). The right drawing in (c) shows the activity of two model neurons during simulated spindle waves. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; GABA, γ-aminobutyric acid. (a) Modified from Steriade M and Deschênes M (1984) The thalamus as a neuronal oscillator. *Brain Research Review* 8: 1–63. (b) Modified from von Krosigk M, Bal T, and McCormick DA (1993) Cellular mechanisms of a synchronized oscillation in the thalamus. *Science* 261: 361–364. (c) Modified from Destexhe A, Bal T, McCormick DA, and Sejnowski TJ (1996) Ionic mechanisms underlying synchronized oscillations and propagating waves in a model of ferret thalamic slices. *Journal of Neurophysiology* 76: 2049–2070.

reconcile these apparently contradictory experimental observations. Models showed that both types of rhythmicities are possible and suggested ways to reconcile the experiments. First, different modeling studies found that oscillations very similar to isolated RE preparation *in vivo* could be generated by networks of RE cells. The oscillations were generated by the interaction between the rebound properties of RE cells, as provided by the T-type Ca^{2+} current I*T*, and γ-aminobutyric acid (GABA)-mediated IPSPs, consistent with experiments. Second, several modeling studies found that spindle oscillations with characteristics identical to those in the *in vitro* preparation could be generated from TC-RE loops (Figure 1(c)). In this case, the oscillations were dependent on I*T*-mediated rebound properties in the TC cells and GABA-mediated IPSPs, together with excitatory postsynaptic potential (EPSP)-evoked bursts in RE cells. Third, it has been proposed that the differences observed between spindles *in vitro* and *in vivo* could be explained by the limited connectivity between the RE neurons in the slice and/or insufficient levels of neuromodulation in the slice needed to maintain isolated RE oscillations.

Dialog between the Thalamus and Cortex: Emergence of Large-Scale Synchrony

The initiation and distribution of spindle oscillations in large circuits was investigated more recently by using multiple recordings *in vivo* and *in vitro* (Figures 2(a) and 2(c)). Spindle oscillations *in vitro* show traveling wave patterns, with the oscillation typically starting on one side of the slice and propagating

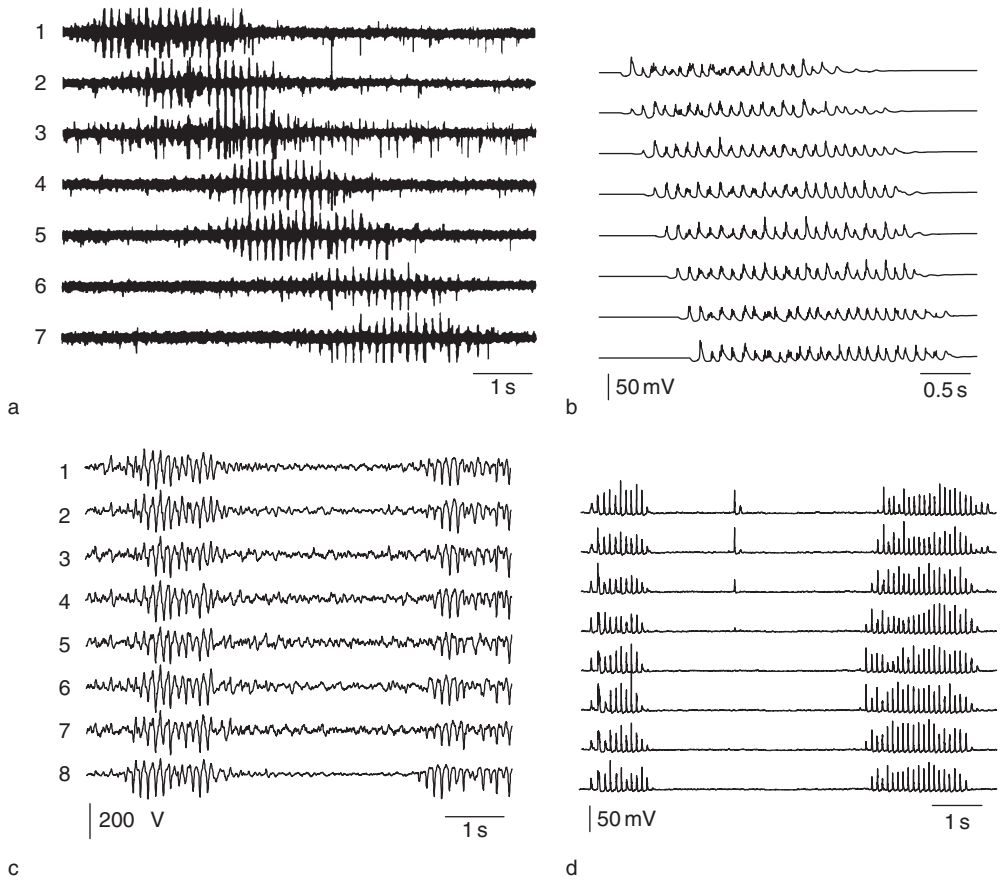

Figure 2 Control of spindle oscillations by the cerebral cortex: (a) multisite extracellular recordings *in vitro* showing the propagating activity of spindle waves in the visual thalamic slice; (b) model of propagating spindle wave activity in thalamic networks with topographic connectivity; (c) multisite extracellular recordings in the thalamus of cats *in vivo* showing the large-scale synchrony of spindle waves in the intact thalamocortical (TC) system; (d) model of the TC network showing large-scale synchrony. (a) Modified from Kim U, Bal T, and McCormick DA (1995) Spindle waves are propagating synchronized oscillations in the ferret LGNd *in vitro*. *Journal of Neurophysiology* 74: 1301–1323. (b) Modified from Destexhe A, Bal T, McCormick DA, and Sejnowski TJ (1996) Ionic mechanisms underlying synchronized oscillations and propagating waves in a model of ferret thalamic slices. *Journal of Neurophysiology* 76: 2049–2070. (c) Modified from Contreras D, Destexhe A, Sejnowski TJ, and Steriade M (1996) Control of spatiotemporal coherence of a thalamic oscillation by corticothalamic feedback. *Science* 274: 771–774. (d) Modified from Destexhe A, Contreras D, and Steriade M (1998) Mechanisms underlying the synchronizing action of corticothalamic feedback through inhibition of thalamic relay cells. *Journal of Neurophysiology* 79: 999–1016.

to the other side at a constant propagation velocity (Figure 2(a)). In contrast to thalamic slices, the intact TC system *in vivo* does not display such clear-cut propagation, but spindle oscillations are remarkably synchronized over extended thalamic regions and show very limited traveling-wave activity (Figure 2 (c)), in agreement with early observations. Moreover, a study by Contreras and colleagues in 1996 showed that large-scale synchrony was lost when the cortex was removed, suggesting that, although the oscillation is generated by the thalamus, its synchrony depends on the cortex. However, cutting intracortical connections has no effect on large-scale synchrony, so cortical connections are not responsible for organizing the synchrony of sleep spindles.

The mechanisms for large-scale synchrony were investigated by computational models by first simulating the propagating properties found in slices (Figure 2(b)). These models assumed a topographic connectivity between TC and RE layers and could generate traveling spindle waves consistent with *in vitro* data. Second, a model by Destexhe and colleagues in 1998 simulated TC networks with bidirectional interactions between the cortex and thalamus and showed that all experiments could be reproduced under the assumption that the cortex recruited TC cells primarily through inhibition. This inhibitory-dominant cortical feedback to the thalamus is consistently observed experimentally and can explain large-scale synchrony by the mutual recruitment of thalamic and cortical networks (Figure 2(d)). The same mechanism can also explain the genesis of absence-type of epileptic seizures when the excitability of cortical neurons is too high. The concept of cortical control of thalamic spindle oscillations can be generalized to other oscillation types, and it was proposed by Steriade that the slow oscillation organizes spindle oscillations, delta oscillations, and more complex patterns such as K-complexes.

Control of Sleep Oscillations by Neuromodulators

Neuromodulators such as acetylcholine (ACh), norepinephrine, (NE), serotonin, (5-HT), histamine (HA), and glutamate abolish the low-frequency rhythms in TC systems during sleep. In the thalamus, where both the relay neurons and the reticular thalamic cells are hyperpolarized during sleep, the activation of the neuromodulatory systems depolarizes the thalamic cells by 5–20 mV, inactivating the low threshold Ca^{2+} current and inhibiting bursting. The transition to tonic firing enhances the participation of sensorimotor processing, which is blocked during sleep.

Conclusion

The thalamus shifts during sleep from a tonic mode, suitable for relaying information from the periphery to the cortex, to a rhythmic mode that generates activity and produces highly spatially and temporally coherent states through interactions with the cortex. During sleep, the feedback connections from the cortex to the thalamus become highly effective in globally coordinating activity in the thalamus. Thus, the thalamus becomes a mirror during sleep, linking distant parts of the cortex. This could be important for maintaining and adjusting the overall balance of activity in the cortex.

See also: Sleep and Sleep States: Hypothalamic Regulation; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Oscillations and PGO Waves; Sleep Oscillations.

Further Reading

Contreras D, Destexhe A, Sejnowski TJ, and Steriade M (1996) Control of spatiotemporal coherence of a thalamic oscillation by corticothalamic feedback. *Science* 274: 771–774.

Destexhe A, Bal T, McCormick DA, and Sejnowski TJ (1996) Ionic mechanisms underlying synchronized oscillations and propagating waves in a model of ferret thalamic slices. *Journal of Neurophysiology* 76: 2049–2070.

Destexhe A, Contreras D, and Steriade M (1998) Mechanisms underlying the synchronizing action of corticothalamic feedback through inhibition of thalamic relay cells. *Journal of Neurophysiology* 79: 999–1016.

Destexhe A and Sejnowski TJ (2001) *Thalamocortical Assemblies*. Oxford: Oxford University Press.

Destexhe A and Sejnowski TJ (2003) Interactions between membrane conductances underlying thalamocortical slow-wave oscillations. *Physiological Reviews* 83: 1401–1453.

Kim U, Bal T, and McCormick DA (1995) Spindle waves are propagating synchronized oscillations in the ferret LGNd *in vitro*. *Journal of Neurophysiology* 74: 1301–1323.

Sanchez-Vives MV and McCormick DA (2000) Cellular and network mechanisms of rhythmic recurrent activity in neocortex. *Nature Neuroscience* 10: 1027–1034.

Steriade M (2003) *Neuronal Substrates of Sleep and Epilepsy*. Cambridge, UK: Cambridge University Press.

Steriade M (2006) Grouping of brain rhythms in corticothalamic systems. *Neuroscience* 137: 1087–1106.

Steriade M and Deschênes M (1984) The thalamus as a neuronal oscillator. *Brain Research Review* 8: 1–63.

Steriade M, Jones EG, and McCormick DA (eds.) (1997) *Thalamus*, 2 vols. Amsterdam: Elsevier.

Steriade M, McCormick DA, and Sejnowski TJ (1993) Thalamocortical oscillations in the sleeping and aroused brain. *Science* 262: 679–685.

von Krosigk M, Bal T, and McCormick DA (1993) Cellular mechanisms of a synchronized oscillation in the thalamus. *Science* 261: 361–364.

Wang XJ, Golomb D, and Rinzel J (1995) Emergent spindle oscillations and intermittent burst firing in a thalamic model: Specific neuronal mechanisms. *Proceedings of the National Academy of Sciences of the United States of America* 92: 5577–5581.

Sleep and Waking in *Drosophila*

C Cirelli and G Tononi, University of Wisconsin, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

The Molecular Correlates of Sleep May Offer Clues to Its Functions

All animal species studied so far sleep, and sleep may serve multiple functions, which may differ in different species. It is possible, however, that sleep also has a core function that is conserved from invertebrates to mammals. If this is the case, that function is most likely a cellular one, because flies and humans share most pathways for intercellular and intracellular signaling, from membrane receptors and ion channels to nuclear transcription factors, but differ significantly in the number, anatomy, and complexity of brain circuits. The idea that the functions of sleep may ultimately relate to cellular and molecular aspects of neural function is not new. Giuseppe Moruzzi argued that "... sleep concerns primarily not the whole cerebrum, nor even the entire neocortex, but only those neurons or synapses, and possibly glial cells, which during wakefulness are responsible for the brain functions concerned with conscious behavior...." (Moruzzi, 1972). Moruzzi suggested that neural cells or synapses supporting waking conscious activity undergo plastic changes that make sleep necessary, although he did not speculate about the mechanisms. Others have suggested that sleep may help in restoring brain energy metabolism, may be needed to maintain the synaptic efficacy of the neural circuits not frequently used during wakefulness, or may favor a generalized synaptic downscaling following waking-induced synaptic potentiation. It has also been suggested that the rich spontaneous activity of neocortical neurons during slow-wave sleep may consolidate memory traces acquired during wakefulness. Several of these hypotheses are not mutually exclusive, and sleep may actually favor different cellular processes.

The analysis of the molecular correlates of sleep and wakefulness may help to understand the benefits that sleep brings at the cellular level. More specifically, the identification of the genes whose expression changes in the brain between sleep and wakefulness may clarify if, and why, brain cells need sleep and why their functions are impaired if they are prevented from doing so during sleep deprivation. In the last 15 years, subtractive hybridization, mRNA differential display, and microarrays have emerged as new powerful methods in transcriptomics, the study of the complete set of RNA transcripts produced by the genome at any one time. Microarrays are currently the most comprehensive and sensitive method to detect genome-wide expression changes, and have allowed the identification of hundreds of transcripts whose expression is modulated by behavioral state. In addition, transcriptomics analysis has also identified functional categories of genes whose expression varies during sleep, wakefulness, and after different periods of sleep deprivation. At least some of these gene categories are conserved from invertebrates to mammals.

Table 1 lists the studies that used transcriptomics approaches to identify genes whose transcript (mRNA) levels vary as a function of sleep, wakefulness, and/or sleep deprivation. It is clear from the table that the main focus has been so far on the cerebral cortex, although other brain regions have recently been studied, including cerebellum, hypothalamus, and brain stem. The cerebral cortex is of major interest because it is responsible for the cognitive defects observed after sleep deprivation, and it is at the center of most hypotheses concerning the functions of sleep.

Several general conclusions can be drawn from the studies listed in Table 1. Each of them is further discussed below.

1. Hundreds of genes in the brain change their expression due to a change in behavioral state.
2. Some genes increase their expression during sleep, others during waking and sleep deprivation.
3. Sleep and waking genes belong to different functional categories.
4. Three functional categories of waking genes are conserved from flies to mammals.
5. The noradrenergic system plays a role in the induction of some waking genes.
6. The expression of several genes changes after chronic, but not acute sleep loss.

Hundreds of Genes in the Brain Change Their Expression between Sleep and Waking

A first important finding of the recent transcriptomics studies is that gene expression in the brain changes extensively as a function of behavioral states, sleep, spontaneous wakefulness, short-term (a few hours) or long-term (several days) sleep deprivation. Five percent or more of the transcripts tested in the brain show changes in gene expression according to behavioral state. Interestingly, a similar percentage of genes (from 1% to 10%) changes because of differences in

Table 1 Published studies that used transcriptomics approaches to identify molecular correlates of sleep (S), wakefulness (W), and sleep deprivation (SD)

Reference	*Method used/N of tested transcripts*	*Experimental details*	*Major findings*
Rhyner et al. *Eur J Neurosci* 2: 1063–73, 1990	Subtractive hybridization (~4000 transcripts)	24 h SD (forced locomotion) vs. time-matched controls; rat forebrain	four transcripts ↓ by SD, six transcripts ↑ by SD, all unknown
Cirelli and Tononi *Mol Brain Res* 56: 293–305, 1998	mRNA differential display (~9000 transcripts)	3 h S, W, SD (novel objects); rat cerebral cortex	one unknown transcript ↑ in S; 11 transcripts ↑ in W-SD, including *c-fos, NGFI-A,* and several mitochondrial genes
Cirelli and Tononi *Brain Res* 885: 303–321, 2000	mRNA differential display (~9000 transcripts) + cDNA arrays (Clontech, 1176 transcripts)	8 h S, W, SD (novel objects); rat cerebral cortex	12 transcripts ↑ in S (all unknown except membrane protein E25); 44 transcripts ↑ in W-SD, including IEGs and heat shock proteins
Terao et al. *Neurosci* 116: 187–200, 2003; 120: 1115–1124, 2003	cDNA arrays (Clontech, 1176 + 588 + 140 transcripts)	6 h SD (gentle handling), 6 h SD + 4 h recovery S vs. time-matched controls; mouse cerebral cortex (changes must be >100%); four array replicas	IEGs and heat shock proteins ↑ during SD Grp94 ↑ in recovery sleep
Cirelli et al. *Neuron* 41: 35–43, 2004	Affymetrix GeneChips, >24 000 probe sets, >15 000 reliably detected	8 h S, W, SD (novel objects); rat cerebral cortex (changes must be >20%); 3–5 array replicas; two statistical methods; qPCR; estimated false positives ~20%, based on qPCR confirmation of 52 transcripts (using independent groups of animals)	106 known transcripts and 155 ESTs ↑ in S; 95 known transcripts and 395 ESTs ↑ in W-SD
Cirelli and Tononi *J Neurosci* 24: 5410–5419, 2004	mRNA differential display (~5000 transcripts) + Affymetrix GeneChips, ~8000 probe sets, ~5000 reliably detected	8 h SD with or without prior LC lesion (DSP-4); rat cerebral cortex (changes must be >20%); three array replicas; two statistical methods; estimated false positives <20%, based on qPCR confirmation of 32 transcripts (using independent groups of animals)	20% (16/ 95) of the known transcripts ↑ in W-SD in Cirelli et al. (2004) are downregulated after cortical depletion of norepinephrine
Cirelli et al. *J Neurochem* 94: 1411–1419, 2005	Affymetrix GeneChips, 18 955 probe sets, ~9400 reliably detected	9 h S, W, SD (forced locomotion) fly head (changes must be >50% in two fly lines); four array replicas; one statistical method; estimated false positives <20%, based on qPCR confirmation of 19 transcripts (using independent groups of animals)	12 transcripts ↑ in S and 121 transcripts ↑ in W-SD
Terao et al. *Neurosci* 137: 593–605, 2006	Affymetrix GeneChips, 1322 probe sets, ~600 reliably detected	6 h SD (gentle handling), 6 h SD + 2 h recovery S vs. time-matched controls; rat cerebral cortex, hypothalamus, basal forebrain (changes must be >50%); two array replicas	
Cirelli et al. *J Neurochem* 98: 1632–1645, 2006	Affymetrix GeneChips, 26 261 probe sets, ~17 000 reliably detected	One week SD (disk-over-water method) compared to yoked controls, 9 h W, and 9 h SD; rat cerebral cortex (changes must be >20%); 3–8 array replicas; one statistical method; estimated false positives <5%, based on qPCR confirmation of 15 transcripts (using independent groups of animals)	75 transcripts ↑ and 28 transcripts ↓ after long-term SD

IEGs, immediate early genes; ESTs, expressed sequence tags.

circadian time. In the cerebral cortex of rats, out of the 15 459 transcripts we tested, 808 (5.2%) were affected by time of day, independently of behavioral state, and 752 (4.9%) were affected by sleep and wakefulness independently of time of day. These data indicate that, at least in the rat, day/night time and sleep/wakefulness influence gene expression in the cerebral cortex to a similar extent. Peripheral tissues such as liver and heart also show significant changes in gene expression as a function of circadian time. Whether this is also true for sleep-dependent and wakefulness-dependent genes is currently unknown, because none of the array studies published so far has focused on peripheral tissues.

Some Genes Increase Their Expression during Sleep, Others during Waking and Sleep Deprivation

We found that in the rat cerebral cortex 95 known genes and 395 expressed sequence tags (ESTs) (i.e., short transcribed sequences whose function is still unknown) show higher expression after 8 h of spontaneous wakefulness or sleep deprivation, while 106 known genes and 155 ESTs show increased expression after 8 h of sleep. In the fly head, in which we used the same experimental protocol previously applied to rats (three experimental groups, including 8 h of sleep, 8 h of spontaneous wakefulness, and 8 h of total sleep deprivation), we identified 121 waking transcripts, including both known genes and ESTs, and 12 sleep transcripts, mostly unknown. Finally, we recently compared gene expression in the cerebral cortex of long-term sleep deprived rats and sleeping controls, and found that 226 transcripts are upregulated after long-term sleep deprivation, and 113 after sleep. Thus, it seems that if all transcribed sequences are considered (including ESTs), there are more genes whose expression is upregulated during waking than during sleep. However, hundreds of genes are also upregulated during sleep, confirming the results of intracellular recording studies that show that sleep is far from being a quiescent state of global inactivity. In general, the largest fold changes are observed for waking genes. However, this is mainly due to the fact that waking increases the expression of transcription factors (e.g., c-fos) and effector genes (e.g., Arc) that behave as 'immediate early genes,' whose mRNA and protein levels may increase several folds relative to sleep. Most waking and sleep genes, instead, change their expression by 20–50%, a fold change similar to that seen for most genes whose expression varies because of circadian time.

Sleep and Waking Genes Belong to Different Functional Categories

Sleep and waking transcripts belong to different functional categories, and thus may favor different cellular processes. Most genes whose expression increases during wakefulness (or short-term sleep deprivation) code for proteins involved in energy metabolism, synaptic excitatory transmission, transcriptional activity, acquisition of new information, and the response to cellular stress. Sleep genes, instead, code for proteins involved in protein synthesis, synaptic downscaling and/or memory consolidation, lipid metabolism, and in membrane trafficking and maintenance.

Three Functional Categories of Waking Genes Are Conserved from Flies to Mammals

Three functional categories of genes are consistently induced in several animal species (fruit flies, mice, rats, hamsters) during waking and short-term sleep deprivation relative to sleep. The first one includes genes involved in energy metabolism, including mitochondrial genes such as cytochrome oxidase (CO) I and IV. One study in rats also found an increase in CO enzymatic activity after 3 h of wakefulness. In mice, 6 h of sleep deprivation also affect the mRNA levels of genes involved in glycogen metabolism (glycogen synthase and glycogen phosphorylase), and increase the activity of glycogen synthase, possibly to counterbalance the rapid depletion, at the transition from sleep to wakefulness, of astrocytic glycogen stores. Eight hours of spontaneous waking or sleep deprivation also induce the expression in the rat cerebral cortex of *Glut1*, the gene coding for one of the more widely expressed glucose transporter in the CNS. The upregulation during waking of genes involved in energy metabolism is perhaps not surprising, because glucose metabolism in many brain regions is higher in wakefulness than in non-rapid eye movement (NREM) sleep, and the brain relies almost exclusively on glucose as its energy substrate. These changes, however, do not seem to persist when sleep deprivation is protracted for 24 h or several days. This is consistent with imaging studies in animals, normal human subjects, and fatal familial insomnia patients, showing no change or a decrease in cerebral metabolic rate after prolonged sleep loss relative to normal wakefulness. Thus, the rapid upregulation of mitochondrial genes, *Glut1*, and glycogen-related genes may represent a mechanism by which the brain responds to the immediate increase in energy requirement during spontaneous waking or after a few hours of sleep deprivation,

while prolonged sleep loss is associated with reduced energy demands.

A second group of genes consistently upregulated during wakefulness and downregulated during sleep includes heat shock proteins, chaperones, and, in general, genes involved in the response to cellular stress. Cellular stress is the cell's reaction to any adverse environmental condition that perturbs cellular homeostasis, with potential macromolecular damage, that is, damage to proteins, DNA, RNA, and lipids. Wakefulness and short-term deprivation are associated with higher mRNA levels of small heat shock proteins with actin-stabilizing properties, such as HSP27 and alpha-B crystallin, molecular chaperones (HSP70, HSP60), and chaperones in the endoplasmic reticulum, such as BiP and Grp94. BiP (immunoglobulin heavy chain-binding protein) shows the largest and most consistent upregulation in the brain of awake animals, including mice, rats, hamsters and sparrows (Cirelli et al., unpublished observations), as well as in the head of fruit flies. BiP assists in the folding and assembly of newly synthesized glycoproteins and secretory polypeptides and increases in *BiP* expression have been classically described after heat shock and ischemia, when unfolded proteins accumulate in the endoplasmic reticulum. This unfolded protein response is also induced by other conditions that adversely affect the function of the endoplasmic reticulum, such as anoxia, glucose deprivation (BiP is also called glucose-regulated protein 78), and depletion of Ca^{2+} stores in the endoplasmic reticulum. Importantly, the unfolded protein response also triggers a slowing down of protein synthesis. In the rat cerebellum, 8 h of spontaneous wakefulness or sleep deprivation are associated with an increase in the mRNA levels of *PEK*. *PEK* codes for the EIF2a kinase that mediates the inhibition of protein synthesis during the unfolded protein response. Moreover, in the mouse cerebral cortex, 6 h of sleep deprivation are also associated with signs of slowing down of protein synthesis, as indicated by the phosphorylation of the key component of the translational machinery and by changes in the ribosomal profile.

A third group of genes consistently upregulated during wakefulness and downregulated during sleep include transcripts sensitive to membrane depolarization and synaptic activity. In the rat cerebral cortex, they include *RGS2*, *Homer/Vesl*, *tissue-type plasminogen activator*, *casein kinase 2*, *cyclooxygenase 2*, *cpg2*, and *connexin 30*. Moreover, several genes involved in activity-dependent neural plasticity and long term potentiation, such as *Arc*, *BDNF*, *Homer/Vesl*, and *NGFI-A*, have been consistently found to be upregulated in the cerebral cortex of awake or short-term sleep deprived rats. At least some of the same genes are induced during wakefulness and/or short-term sleep deprivation in mice, hamsters, and sparrows (Cirelli et al., unpublished results). In the fly head, many of the depolarization-sensitive genes involved in synaptic plasticity as identified in Guan et al. are also waking-related genes, including *Astray*, *Kek2*, *stripe*, and *Hr38*. The fly wakefulness-related gene *stripe* is a member of the Egr family of transcription factors. In the rat brain several members of this family, including *NGFI-A* (*Egr-1*), *krox 20* (*Egr-2*), and *NGFI-C* (*Egr-4*), are induced during wakefulness. *Hr38*, another fly wakefulness-related gene, is the *Drosophila* ortholog of the mammalian NGFI-B family of orphan nuclear receptors. *NGFI-B* is a wakefulness-related gene in the rat cerebral cortex. Thus, it seems that in both mammals and flies several plasticity-related genes involved in synaptic growth and potentiation are specifically induced during wakefulness.

The observation that genes involved in long-term potentiation and the acquisition of new memories are specifically induced during wakefulness should not be too surprising. After all, we learn when we are awake and able to adapt to the environment, and not when we are asleep. But if sleep is not involved in learning, is there a role for it in other aspects of synaptic plasticity? Well-controlled new experiments show that sleep may favor the performance of some, but not all, tasks. The strongest and more consistent data relate to procedural tasks, and suggest an important role for NREM sleep or a combination of NREM and REM sleep. What remains unknown, however, are the mechanisms by which sleep may enhance performance. Animal and human studies have shown that brain circuits activated during the execution or the acquisition of certain tasks are 'reactivated' during the subsequent sleep period. The presence of this memory trace, as shown by electrophysiological recordings and imaging studies, could reflect the fact that sleep replays the waking experience, and by doing so somehow allows the specific consolidation of the learned tasks. However, it is also possible that this trace is the inevitable consequence of the synaptic potentiation occurred during learning, and that sleep enhances performance by actually allowing a generalized downscaling of synapses. The first and most important effect of this downscaling is thought to be the decrease in the metabolic cost of synaptic activity, which increases during waking due to synaptic potentiation. A second effect of the generalized but proportional decrease in synaptic weight is an increase in the signal to noise ratio, which could improve the performance of some learned tasks.

If sleep is associated with a generalized depression of synaptic strength, molecular markers of synaptic depression should be increased during sleep relative to waking. Recent studies have identified and validated several specific and direct markers of synaptic strength, the best established among them being the density, composition, and phosphorylation state of the GluR1 subunit of the glutamatergic AMPA receptors. We are currently measuring the number and phosphorylation state of synaptic GluR1 in the rat cerebral cortex after 6 h of sleep or sleep deprivation. Preliminary results show that in several cortical areas, sleep is associated with a relative dephosphorylation of GluR1 at serine 845, a finding consistent with the hypothesis that sleep is associated with synaptic downscaling.

The Noradrenergic System Plays a Role in the Induction of Some Waking Genes

One of the mechanisms that underlie the widespread changes in cortical gene expression between sleep and wakefulness is the activity of the noradrenergic system of the locus coerulues (LC), whose neurons project diffusely over the entire brain. LC cells are tonically active during wakefulness, reduce their firing rate during NREM sleep, and cease firing during REM sleep. Moreover, LC activity increases phasically in response to salient events and in relation to the decision to act. In a series of experiments rats were subjected to unilateral or bilateral LC lesions to deplete one or both sides of the brain of norepinephrine. In these animals the raw EEG and its power density spectrum were not significantly affected by the lesion. However, all cortical areas depleted of norepinephrine showed a marked decrease, during wakefulness, of the expression of plasticity-related genes such as *Arc*, *BDNF*, *NGFI-A*, and *P-CREB*, as well as of stress response genes such as heat shock proteins and *BiP*. By contrast, the transcript for the translation elongation factor 2 was the only known sleep-related transcript whose expression increased after LC lesion (Figure 1). In a complementary experiment in mice the activity of the LC of one side was increased using a conditional transgenic approach. This manipulation resulted in an increased ipsilateral expression of *NGFI-A* in cortical and subcortical target areas. Thus, LC activity during wakefulness modulates neuronal transcription to favor synaptic potentiation and memory acquisition and to counteract cellular stress, while LC inactivity during sleep may play a permissive role to enhance brain protein synthesis.

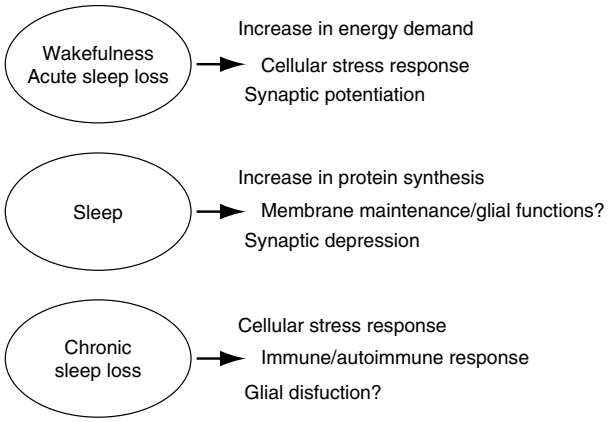

Figure 1 Schematic representation of the major functional categories of genes upregulated in the rat cerebral cortex after 3–8 h of spontaneous waking or sleep deprivation (acute sleep loss), 1 week of total sleep deprivation (chronic sleep loss), and 8 h of spontaneous sleep.

The Expression of Several Genes Changes after Chronic, but Not Acute Sleep Loss

In a recent array study we analyzed the expression of >26 000 transcripts in the cerebral cortex of long-term sleep deprived rats. Animals were kept awake for 7 days using the disk-over-water method (DOW). This method uses minimal stimulation to enforce chronic sleep deprivation in the sleep-deprived rat, while it simultaneously applies to the control rat the same stimulation, but without severely limiting its sleep. The sleep-deprived and the control rat are housed each on one side of a divided horizontal disk suspended over a shallow tray of 2–3 cm deep water. EEG and EMG are continuously recorded to detect sleep states. When the experimental rat falls asleep, the disk is automatically rotated at low speed, awakening the rat and forcing it to walk opposite to disk rotation to avoid being carried into the water. The yoked control rat receives the same physical stimulation because it is on the same disk. However, while sleep is severely reduced (by 70–80%) in the sleep deprivation rat, the control rat can sleep whenever the sleep deprived rat is spontaneously awake and eating, drinking, or grooming and thus its sleep is only reduced by 25–40%.

We compared long-term sleep deprived rats with rats spontaneously awake or sleep deprived for only 8 h, and identified 75 transcripts upregulated, and 28 transcripts downregulated after chronic sleep loss. Most transcripts specifically increased or decreased their expression due to the long-term sleep deprivation and not because of the disk stimulation, because they were not affected in the yoked controls. A large

group of genes upregulated after prolonged sleep loss codes for immunoglobulins, including two autoantibodies. Another group includes stress response genes such as the macrophage inhibitor factor-related-protein MRP14, expressed in microglia. Aryl-sulfotransferase, an enzyme involved in the catabolism of catecholamine, is also upregulated after chronic sleep deprivation. These genes are either not induced (antibodies, MRP14), or induced to a lesser extent (aryl-sulfotransferase) after acute sleep loss, suggesting that short-term and long-term sleep deprivation differ to some extent at the molecular level. We also compared long-term and short-term sleep deprivation to sleep, and found that only a minority of genes were similarly up- or downregulated in the two sleep deprivation conditions. Among the genes whose expression was most significantly downregulated in long-term sleep deprived animals relative to sleeping animals were myelin constituents, including the most abundant structural protein component of myelin, the proteolipid protein. These findings suggest that sustained sleep loss may trigger a generalized inflammatory and stress response in the brain. Some glial cells (e.g., microglia) may help to protect neurons against this cellular insult, but others (e.g., oligodendrocytes) may also suffer some of its negative consequences. It will be important in the future to test whether sleep loss may be detrimental to the maintenance of cellular membranes and more specifically to myelin.

Genes Upregulated during Sleep

Figure 1 shows the sleep-related transcripts identified so far in the rat cerebral cortex . Unfortunately, so far only one study in mammals was designed to identify genes upregulated after several (8) hours of sleep relative to both spontaneous waking and sleep deprivation, to rule out effects of stress and of circadian time. Another study in flies used the same experimental conditions used in rats, but for technical reasons the entire fly head, rather than the brain only, was used to hybridize the arrays. This most likely decreased the ability to detect state-dependent genes. Only 12 genes were identified as sleep related in flies, and the function of most of them is currently unknown. Thus, the extent to which sleep genes are conserved across species is yet to be determined.

Two sleep genes in the rat cerebral cortex code for the translation elongation factor 2 and the initiation factor 4AII, and are involved in protein synthesis. The mRNA levels of the elongation factor EF2 are also increased during sleep in the brain of Djungarian hamsters (Cirelli et al., unpublished results). A positive correlation between sleep and protein synthesis has been suggested by previous studies, and recent evidence suggests that waking and sleep deprivation are less conducive to protein synthesis than sleep. Whether sleep favors protein synthesis globally, or whether it enhances the synthesis of specific classes of proteins is still unclear.

Other sleep genes identified in the rat cerebral cortex are involved in the consolidation of long-term memory, as well as in synaptic depression and depotentiation. Among them are *calmodulin-dependent protein kinase IV*, *calcineurin*, *FK506 binding protein 12*, *inositol 1,4,5-trisphosphate receptor*, and *amphiphysin II*. As mentioned above, the mechanism by which sleep enhances memory and performance is still debated, but these molecular data are compatible with the idea that sleep benefits the brain by producing a global synaptic downscaling.

Another large group of sleep-related transcripts in the rat cerebral cortex is involved in membrane trafficking and maintenance. Some of these transcripts are important for the synthesis/maintenance of membranes in general and of myelin in particular, including oligodendrocytic genes coding for myelin structural proteins (e.g., *MOBP*, *MAG*, *plasmolipin*, *CD9*), myelin-related receptors and enzymes. Transcripts with higher expression in sleep also code for enzymes involved in the synthesis and transport of cholesterol, a major constituent of myelin and other membranes and an important factor in regulating synaptic efficacy. As mentioned above, long-term sleep deprivation decreases the expression of the gene coding for proteolipid protein, a major myelin component. Intriguingly, among the few known sleep genes in *Drosophila* is the gene *anachronism* (*ana*), as well as three genes involved in lipid metabolism (*CG8756*, *CG9009*, *CG11407*). Ana is a glycoprotein secreted by glial cells, and controls the timing of neuroblast development. Whether *ana* plays any role in membrane trafficking and synaptic plasticity remains to be determined.

Conclusions

As it is apparent from the small number of studies in Table 1, transcriptomics analysis is still a work in progress, and it would be inappropriate to draw firm conclusions from what has been published so far. Continuous improvement in the array technology will certainly favor more progress in this area. For instance, the current arrays can assess the expression of 20 000–30 000 transcripts, but this is still far from been a true 'whole-genome' analysis. Low-expression genes are often difficult to detect, especially in large and heterogeneous tissue samples such as most brain areas. Moreover, current arrays focus on the coding portion of the genome, but new 'tiling' arrays, in

which probe sets are laid down at regular intervals across the entire genome, are becoming commercially available. This will permit the identification of new transcriptionally active regions of the genome that have currently not been tested in more conventional arrays. Moreover, proteomics, the study of the complete repertoire of proteins that contribute to a cell's physiological phenotype, is a natural next step in the study of molecular correlates of behavioral states.

See also: Circadian Organization; Circadian Organization in Non-Mammalian Vertebrates; Circadian Regulation in Invertebrates; Circadian Rhythms: Influence of Light in Humans; Genetic Regulation of Circadian Rhythms in Drosophila; Mammalian Sleep and Circadian Rhythms: Flies; Photoreceptors: Physiology.

Further Reading

Benington JH and Heller HC (1995) Restoration of brain energy metabolism as the function of sleep. *Progress in Neurobiology* 45: 347–360.

Cirelli C and Tononi G (1998) Differences in gene expression between sleep and waking as revealed by mRNA differential display. *Brain Research. Molecular Brain Research* 56: 293–305.

Cirelli C and Tononi G (2000) Differential expression of plasticity-related genes in waking and sleep and their regulation by the noradrenergic system. *Journal of Neuroscience* 20: 9187–9194.

Cirelli C and Tononi G (2000) Gene expression in the brain across the sleep-waking cycle. *Brain Research* 885: 303–321.

Cirelli C and Tononi G (2004) Locus ceruleus control of state-dependent gene expression. *Journal of Neuroscience* 24: 5410–5419.

Cirelli C, Pompeiano M, and Tononi G (1996) Neuronal gene expression in the waking state: A role for the locus coeruleus. *Science* 274: 1211–1215.

Cirelli C, Gutierrez CM, and Tononi G (2004) Extensive and divergent effects of sleep and wakefulness on brain gene expression. *Neuron* 41: 35–43.

Cirelli C, LaVaute TM, and Tononi G (2005) Sleep and wakefulness modulate gene expression in *Drosophila*. *Journal of Neurochemistry* 94: 1411–1419.

Cirelli C, Faraguna U, and Tononi G (2006) Changes in brain gene expression after long-term sleep deprivation. *Journal of Neurochemistry* 98: 1632–1645.

Everson CA, Smith CB, and Sokoloff L (1994) Effects of prolonged sleep deprivation on local rates of cerebral energy metabolism in freely moving rats. *Journal of Neuroscience* 14: 6769–6778.

Guan Z, Saraswati S, Adolfsen B, and Littleton JT (2005) Genome-wide transcriptional changes associated with enhanced activity in the *Drosophila* nervous system. *Neuron* 48: 91–107.

Hairston IS, Peyron C, Denning DP, et al. (2004) Sleep deprivation effects on growth factor expression in neonatal rats: A potential role for BDNF in the mediation of delta power. *Journal of Neurophysiology* 91: 1586–1595.

Kavanau JL (1997) Memory, sleep and the evolution of mechanisms of synaptic efficacy maintenance. *Neuroscience* 79: 7–44.

Krueger JM, Obal F Jr., Kapas L, and Fang J (1995) Brain organization and sleep function. *Behavioural Brain Research* 69: 177–185.

Maquet P (1995) Sleep function(s) and cerebral metabolism. *Behavioural Brain Research* 69: 75–83.

Moruzzi G (1972) The sleep–waking cycle. *Ergebnisse der Physiologie* 64: 1–165.

Naidoo N, Giang W, Galante RJ, and Pack AI (2005) Sleep deprivation induces the unfolded protein response in mouse cerebral cortex. *Journal of Neurochemistry* 92: 1150–1157.

Nelson SE, Duricka DL, Campbell K, Churchill L, and Krueger JM (2004) Homer1a and 1bc levels in the rat somatosensory cortex vary with the time of day and sleep loss. *Neuroscience Letters* 367: 105–108.

Nikonova EV, Vijayasarathy C, Zhang L, et al. (2005) Differences in activity of cytochrome C oxidase in brain between sleep and wakefulness. *Sleep* 28: 21–27.

Petit JM, Tobler I, Allaman I, Borbely AA, and Magistretti PJ (2002) Sleep deprivation modulates brain mRNAs encoding genes of glycogen metabolism. *European Journal of Neuroscience* 16: 1163–1167.

Reich P, Driver JK, and Karnovsky ML (1967) Sleep: Effects on incorporation of inorganic phosphate into brain fractions. *Science* 157: 336–338.

Rhyner TA, Borbely AA, and Mallet J (1990) Molecular cloning of forebrain mRNAs which are modulated by sleep deprivation. *European Journal Neuroscience* 2: 1063–1073.

Salbaum JM, Cirelli C, Walcott E, Krushel LA, Edelman GM, and Tononi G (2004) Chlorotoxin-mediated disinhibition of noradrenergic locus coeruleus neurons using a conditional transgenic approach. *Brain Research* 1016: 20–32.

Stickgold R (2005) Sleep-dependent memory consolidation. *Nature* 437: 1272–1278.

Taishi P, Sanchez C, Wang Y, Fang J, Harding JW, and Krueger JM (2001) Conditions that affect sleep alter the expression of molecules associated with synaptic plasticity. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* 281: R839–845.

Terao A, Greco MA, Davis RW, Heller HC, and Kilduff TS (2003a) Region-specific changes in immediate early gene expression in response to sleep deprivation and recovery sleep in the mouse brain. *Neuroscience* 120: 1115–1124.

Terao A, Wisor JP, Peyron C, et al. (2005) Gene expression in the rat brain during sleep deprivation and recovery sleep: An Affymetrix geneChip((R)) study. *Neuroscience* 26: 26.

Terao A, Steininger TL, Hyder K, et al. (2003b) Differential increase in the expression of heat shock protein family members during sleep deprivation and during sleep. *Neuroscience* 116: 187–200.

Tononi G and Cirelli C (2006) Sleep function and synaptic homeostasis. *Sleep Medicine Reviews* 10: 49–62.

Voronka G, Demin NN, and Pevzner LZ (1971) Total protein content and quantity of basic proteins in neurons and neuroglia of rat brain supraoptic and red nuclei during natural sleep and deprivation of paradoxical sleep. *Dokland Akademii Nauk SSSR* 198: 974–977.

Sleep Apnea

S C Veasey, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Sleep Apneas Are Neurally Mediated

Distinguishing the types of sleep-disordered breathing events and sleep apnea syndromes is of utmost importance in gauging patient risk and guiding optimal therapy. The term apnea is defined in the adult as cessation in ventilatory airflow for 10 s or longer. In children, cessation of airflow for any duration qualifies as an apnea. Sleep apnea syndromes have been divided into two categories: obstructive and central sleep apnea. This terminology has led to confusion because it implies neural and mechanical processes for central and obstructive apneas, respectively. Indeed, a common misperception is that obstructive sleep apnea (OSA) has a primarily physical etiology (a fixed obstruction in the upper airway), whereas central sleep apnea (CSA) involves alterations in neural input. It is essential to understand, however, that all sleep-disordered breathing syndromes, including OSA, occur only during sleep and thus involve sleep-dependent changes in neural drive to various ventilatory and airway stabilizer muscles. During wakefulness, individuals with OSA have normal ventilation and no significant upper airway obstruction. In wakefulness, these individuals effectively use upper airway muscle to stent open the upper airway. A normal consequence of sleep is reduced neural drive to many muscle groups, including upper airway dilator muscles (e.g., muscles of the tongue and soft palate reduce muscle activity). The termination of an obstructive apnea typically requires wakefulness with resumption of upper airway muscle activity. In other words, the obstructive process in OSA is dynamic and neurally mediated. The neural mechanisms by which OSA occurs may differ for nonrapid eye movement (NREM) and rapid eye movement (REM) sleep. There are differences in neurochmeical drive to the upper airway dilator motor neurons in NREM and REM sleep, and subsets of individuals with OSA are affected differentially in the two distinct sleep states. Specifically, a fraction of patients with OSA have apneas exclusively in REM sleep, whereas others have events in both sleep states but may have more frequent events in NREM sleep than in REM sleep. In most patients, reductions in arterial oxygen levels induced by apneas are more pronounced in REM sleep because REM sleep atonia of chest wall muscles compromises lung volume and oxygenation.

Individuals with CSA do not have increased collapsibility of the upper airway and therefore do not rely on upper airway muscles to maintain patency. There are two major neural mechanisms underlying central apneas. First, in one group of individuals with CSA, ventilatory drive is markedly less in sleep than in wakefulness, and thus a transition into sleep can induce an apnea. These patients are most likely to have apneic events in stages 1 and 2 NREM sleep, when a sigh at the termination of one apnea can lower the $PaCO_2$ sufficiently to initiate the next apnea. In the second group of patients with CSA, there is a reliance on respiratory muscles other than the diaphragm for ventilation. In NREM sleep, these accessory muscles are somewhat suppressed, and in REM sleep the muscles may have atonia. Thus, in these individuals, CSA is expected to be most pronounced in REM sleep.

A second basic concept regarding sleep apnea and its terminology is that not all significant sleep-disordered breathing events are apneas. This concept is most relevant for OSA, for which epidemiological studies justify inclusion of other ventilatory disruptions. In fact, a partial reduction in ventilation occurring with a drop in the oxyhemoglobin desaturation or with an arousal may result in as much sleepiness as apnea events of the same frequency. Clinically significant OSA events occur along a spectrum from what is termed respiratory effort-related arousals (an airflow limitation with an arousal but normal ventilation) to hypopneas (commonly defined as 30–90% reductions in ventilation, measured as airflow, with a 4% decline in the oxyhemoglobin saturation) to apneas (as defined previously). These clinical definitions are relevant because respiratory effort-related arousals and hypopneas can cause significant sleepiness, whereas hypopneas may contribute to cardiovascular disease. These types of disordered breathing events are categorized and defined in Figure 1. Disease severity for OSA is defined by the frequency of apnea and hypopnea events per hour of sleep. The American Academy of Sleep Medicine defines mild, moderate, and severe sleep apnea as 5–15, 16–30, and more than 30 apneas and/or hypopneas per hour, respectively. The mathematical modeling for OSA disease severity and morbidity risk, however, is not complete, and it is anticipated that future definitions of severity of OSA will include a factor for the magnitude of oxygenation changes, the time spent hypoxic, and/or the degree to which sleep is fragmented, as described later.

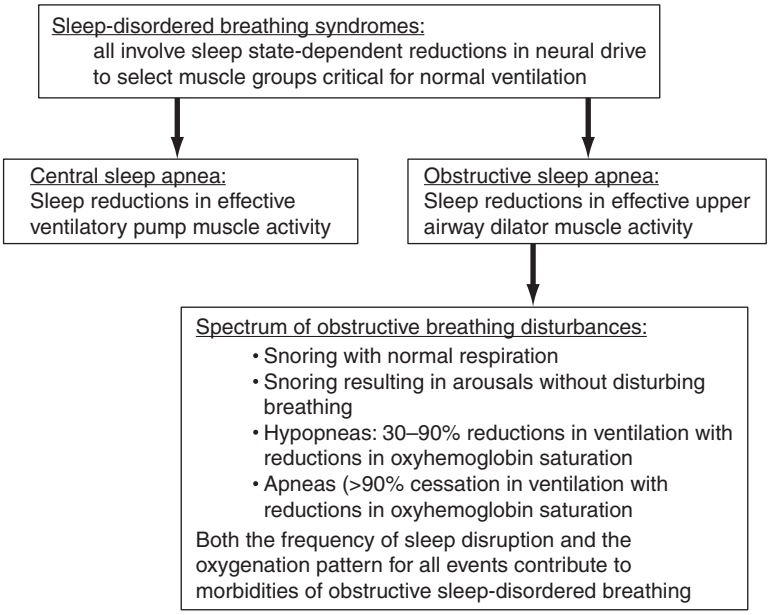

Figure 1 Delineation of sleep apnea syndromes and definitions of sleep-disordered breathing events.

Epidemiology and Genetics of Sleep Apnea

Beyond the common mechanistic theme of sleep-related reductions in neural drive to ventilatory muscles, the pathogeneses for OSA and CSA are remarkably distinct. Consequently, the various sleep apnea syndromes have minimal overlap in epidemiology and genetic bases.

Obstructive Sleep Apnea

The prevalence and/or severity of OSA vary with gender, age, family history, and ethnicity. Specifically, the prevalence of symptomatic OSA (OSA with sleepiness or fatigue) in adults in North America, Europe, and in select regions of Asia is 4–7% for males and 2% or 3% for females. In prepubertal children, there are no apparent gender differences. Prevalence increases with age, rising to more than 40% in elderly individuals. Although some studies have shown increased prevalence of OSA in females after menopause, longitudinal studies are needed to better delineate hormonal and aging effects on OSA. The prevalence of OSA in subjects in the Sleep Heart Health Study appears not to vary with race. However, it is clear that the severity of OSA is greater in age-matched African American versus Caucasian individuals, and overall, OSA develops at an earlier age in African Americans than in Caucasian adults. Familial aggregation is clear for OSA. Approximately 40% of the variance in the apnea hypopnea frequency may be explained by familial factors, and a positive family history increases the relative risk by two- to fourfold. Craniofacial anomalies that compromise upper airway space and stability provide additional risk factors for OSA. At least two endocrine conditions predispose to OSA: hypothyroidism and acromegaly. The most significant risk factor for OSA, however, is obesity, which is best characterized using the body mass index (kg m^{-2}) and/or the neck circumference. With the prevalence of obesity increasing in North America, prevalence rates determined 10 years ago will soon require reexamination. In children, the major risk factor has been adenoid and tonsil hypertrophy, although the increase in obesity in children is threatening to make it a major risk factor. A small number of patients diagnosed with OSA are thin. In this group, chronic nasal obstruction from allergies or septal deviation, retrognathia, or chronically enlarged tonsils may predispose patients to increased collapsibility of the upper airway and OSA. In summary, the most important risk factors for OSA appear to be obesity (body mass index and neck circumference), age, family history of OSA, male gender, craniofacial and/or upper airway soft tissue anomalies, hypothyroidism, and acromegaly. However, because OSA is so prevalent, it is essential that the possibility of OSA is considered in all individuals presenting with fatigue and/or unrefreshed sleep.

Although it is clear that OSA has a significant genetic basis, it is less clear whether the genetic basis rests entirely with the intermediate phenotypes (e.g., obesity and craniofacial morphology) or whether there are also genes that determine the severity of OSA and the clinical sequelae of the disease severity. It is important to acknowledge that the many risk factors for OSA and the varied clinical

presentations and comorbidities render this syndrome a complex disorder, determined not only by numerous genetic factors but also by environmental and cultural factors. The relative importance of body mass index versus craniofacial structure influences is expected to vary substantially across races with craniofacial differences. Genetics of OSA will require reexamination in large populations in India and China, where obesity may increase with economic expansion. Although there are many candidate genes for the intermediate phenotypes and for ventilatory control, very few have been explored and fewer still in large-scale clinical studies. In adults with OSA, polymorphisms of angiotensin converting enzyme, APOε4, and tumor necrosis factor-α have been identified. Whether these have a direct role in OSA risk or in the severity of comorbidities will require further investigation. Several studies on nonhuman mammals support genetic variation in apnea and hypoxia responses. These studies suggest that genetics may play a very important role in individual susceptibility to complications of OSA.

Central Sleep Apnea

It is estimated that CSA represents just 4% of the sleep apnea diagnoses made in clinical sleep laboratories. CSA occurs in many vastly different conditions. These CSA predispositions may be classified as disorders associated with (1) heightened ventilatory drive, (2) reduced ventilatory drive, and (3) insufficient and abnormal ventilatory pump function. Because of these diverse predispositions, the prevalences of CSA and for many of its subtypes are not known. However, it is estimated that 25–40% of individuals with congestive heart failure with systolic dysfunction may have CSA, and many of these individuals will have Cheyne–Stokes respiration. An additional risk factor for CSA associated with heightened respiratory drive includes periodic breathing or CSA associated with ascent to high altitude (acute hypoxia-induced increased ventilatory drive). The predisposition to CSA at altitude is associated with several angiotensin converting enzyme polymorphisms. Neuromuscular disorders with pump muscle weakness also predispose to CSA. Thus, CSA may be observed in individuals with Duchenne's or myotonic muscular dystrophy or amyotropic lateral sclerosis, myasthenia gravis, multiple sclerosis, and other disorders involving lower ventilatory pump muscle function. In this collection of neuromuscular disorders, the predisposition to sleep-disordered breathing arises from reliance on pump muscles other than the diaphragm. These accessory pump muscles normally have greater reductions in activity in both NREM and REM sleep compared to the diaphragm. Thus, when these accessory muscles are required for normal ventilation, it is expected that CSA will be present. A very rare form of CSA is congenital central hypoventilation syndrome (CCHS), with apneas occurring in 1 of every 200 000 births. Mutations in the *Phox2B* gene occur in most, if not all, individuals with CCHS. This gene is essential in the development of the autonomic nervous system. Sleep and breathing studies in mice with transgenic absence of the gene have confirmed the importance of the *Phox2B* gene in the development of autonomic ventilatory control and CSA by demonstrating impaired autonomic neuronal development CSA in mice lacking this gene. Whether specific mutations in *Phox2B* or in other gene mutations identified in children with CCHS (e.g., *Ret*, *Mash-1*, and *hash-1*) may contribute to adult forms of idiopathic CSA is not known.

Pathophysiology

Obstructive Sleep Apnea

Three decades ago, Sauerland and Harper measured the electromyographic activity of a major upper airway dilator muscle, the genioglossus, across sleep in subjects with OSA. Coincident with each apneic event, they observed reductions in genioglossus muscle activity, and upon arousal they observed resumption of genioglossus activity and resolution of the apnea. These landmark observations set the stage for further elucidation of mechanisms underlying the pathogenesis of OSA. People with OSA may have normal or reduced upper airway dimensions in wakefulness and may have increased soft tissue surrounding the upper airway, but the universal feature of OSA is increased collapsibility of the upper airway, particularly when upper airway muscle activity is reduced.

Collapse of the upper airway is a dynamic process. Inspiration creates negative pressures in the upper airway. Counterpressure must be generated to prevent collapse of the upper airway in individuals with severe OSA. Collapsibility of the upper airway in snorers is greater than collapsibility in individuals without sleep-disordered breathing but far less than in people with OSA. This increased collapsibility parallels the severity of sleep-disordered breathing events, where the most collapsible airways are present in individuals with OSA who have predominantly apneas.

With the understanding that sleep-dependent reductions in upper airway muscle activity are central to the collapse of the upper airway in OSA, a significant effort has been put forth to determine the neurochemical mechanisms underlying reduced muscle activity. It is now evident that sleep-dependent reductions in serotonin and noradrenaline and increases in GABA

delivery to motor neurons innervating the genioglossus contribute to the reduced muscle activity in sleep. It is unfortunate that serotonergic and noradrenergic receptor subtypes on upper airway motor neurons do not represent safe pharmacological targets. It is hoped that novel G-protein-coupled receptors will be identified on upper airway motor neurons and that these may be pharmacotherapeutic targets.

Central Sleep Apnea

As a condition with diverse causes, including congestive heart failure, CCHS, and many neuromuscular disorders, the pathophysiology of CSA varies with the cause of the condition. In deep NREM sleep (stages 3 and 4), ventilation is controlled almost exclusively by autonomic factors and is therefore extremely regular. In lighter NREM sleep (stages 1 and 2), there are major differences between the ventilatory response to low $PaCO_2$. A low $PaCO_2$ in wakefulness will not result in a cessation of ventilation. In contrast, a low $PaCO_2$ in light NREM sleep can suppress ventilation, causing an apnea until the $PaCO_2$ rises enough to stimulate ventilation. Typically, there is an overshoot so that upon arousal, the ventilatory response is so robust that the $PaCO_2$ falls again to a level that induces an apnea if the patient falls quickly back to sleep. This principle underlies the pathophysiology in several forms of CSA. If ventilatory drive is high in waking and subjects enter into NREM sleep with a sufficiently low $PaCO_2$, below the apneic threshold in NREM sleep, apnea will ensue until the CO_2 rises sufficiently to augment ventilatory drive in sleep or induce an arousal. Through this same sleep-dependent reduced ventilatory drive mechanism, a healthy individual may have a post-sigh apnea in NREM sleep. High-altitude periodic breathing also occurs through a similar change in ventilatory drive. At high altitude, there is some hypoxic drive to increase ventilation and sufficiently lower $PaCO_2$. Upon entering sleep, again if the $PaCO_2$ is below the apnea threshold, an apnea will occur until an increased $PaCO_2$ is detected resulting in an arousal that, in turn, greatly augments ventilation, blowing the $PaCO_2$ back down below the apnea threshold, where the process repeats itself throughout the night.

This same sleep-dependent change in ventilatory response to $PaCO_2$ underlies Cheyne–Stokes breathing. In this condition, the heightened ventilatory drive in wakefulness is, in part, a consequence of increased sympathetic drive that drives down the wake $PaCO_2$. Cheyne–Stokes respiration has been successfully treated pharmacologically with adrenergic blockade.

The remainder of conditions predisposing to CSA are those with impaired neuromuscular function. Individuals with neuromuscular weakness may have impaired diaphragm function yet can have normal waking ventilation by using accessory muscles (e.g., the sternocleidomastoid or internal intercostals and abdominal muscles). Sleep-related central apneas occur when sleep-dependent reductions in accessory muscle activity occur. In these conditions, sleep apnea will be worse in REM sleep when atonia is present in all accessory muscles of ventilation.

Clinical Features of Sleep Apnea

Obstructive Sleep Apnea

The clinical presentation of OSA varies across subsets of individuals. Adults with severe OSA, particularly obese males, present with a remarkably consistent constellation of: loud snoring, witnessed pauses in breathing at night, unrefreshed sleep, and daytime sleepiness or fatigue. Dyspepsia, rhinitis, and nocturia are also commonly associated symptoms. Many individuals with severe OSA will also report clumsiness, poor concentration, and impaired dexterity and/or memory impairments. Recognition of mild to moderate OSA in thin males, premenopausal females, and children presents a greater challenge. Patients with less severe OSA are more likely to have complaints of insomnia, fatigue, and quiet or minimal snoring. Therefore, of significant concern is the potential for underdiagnosis. Somatic complaints may predominate and may be attributed to chronic fatigue syndrome, depression, and fibromyalgia. The presentation in children differs greatly from the presentation in adults. Specifically, parents and other caregivers of children with OSA may recognize increased distractibility, labile mood, and hyperactivity in these children, without appreciating sleepiness, sleep disturbances, or even snoring.

The physical exam in many, but not all, cases will provide clues of upper airway compromise, such as a large tongue, a large neck circumference (>17 in. in adult males or >15 in. in adult females), a low-lying soft palate, a posteriorly displaced soft palate (1–4 mm anterior–posterior velopharyngeal space), large tonsils, a small mandible, and, rarely, a high arched palate. However, it is possible to have a normal head and neck exam and yet have enlarged lingual tonsils or collapse of the pharynx that is evident only in sleep. The physical exam should also always include an exploration for signs of hypothyroidism and acromegaly because these would be treated differently, as indicated later.

Central Sleep Apnea

The clinical presentation in CSA varies with the underlying cause, and as with the different

pathophysiologies, depending on the cause, the clinical presentation may be classified into two categories: individuals with lower arterial $PaCO_2$ values in wakefulness and those with normal or increased arterial $PaCO_2$. Individuals with CSA and low arterial $PaCO_2$ values may describe sleep as disturbed and unrefreshing, or they may have daytime hypersomnolence or fatigue. Increased fatigue with sedentary behaviors compared to exertional behaviors should alert physicians to examine the possibility of sleep-disordered breathing. Some patients may present with paroxysmal nocturnal dyspnea or may present with quiet snoring and witnessed apneas. However, the witnessed apneas and snoring are less obvious than events witnessed in people with severe OSA. Individuals with normal to increased $PaCO_2$ may present with clinical features more consistent with chronic hypoxemia and hypoventilation: polycythemia, cor pulmonale, or morning headaches.

Diagnosis

The gold standard for diagnosing both CSA and OSA has been overnight polysomnography. Recording sleep in parallel with ventilatory pump effort and nasal and oral airflow, the full spectrum of snoring, hypopnea, and apnea events can be examined. In addition, the effects of sleep-disordered breathing events on sleep quality and on arterial oxygenation may be determined. Together, this information distinguishes CSA and OSA and the severity of OSA can be determined. In OSA, night-to-night variability can be high, particularly in individuals with mild to moderate disease. Thus, when clinical suspicion is high and the first polysomnogram does not show OSA, it may be necessary to repeat the study. In females and in people with mild OSA, events may cluster in REM sleep. Thus, it is essential to examine sufficient REM sleep in individuals before excluding the possibility of OSA. Also in these patient populations, it is important to search for frequent arousals (>15 h) associated with increased ventilatory effort. With CSA, polysomnographic features may be instrumental in elucidating causes of sleep-disordered breathing. For example, Cheyne–Stokes presents with a very characteristic crescendo–decrescendo pattern in ventilation, with the arousal occurring at the peak of ventilation. Congenital central hypoventilation syndrome, which has rarely been identified in adults, manifests with most severe apneas occurring in slow-wave sleep. In contrast, because ventilatory motor neuron suppression is most pronounced in REM sleep, many patients with pulmonary disease or neuromuscular disorders (those that rely on accessory muscles for ventilation) will have more severe sleep-disordered breathing in REM sleep, when these muscle have atonia. Neither the sensitivity of polysomnography for the diagnosis of CSA nor the night-to-night variability in CSA has been established. Thus, if clinical suspicion is high, repeated polysomnography may be indicated in the hypersomnolent patient.

Of significance, OSA progresses with age and weight gain; thus, reconsideration of the severity of disease and its treatments is required periodically.

Therapies

Obstructive Sleep Apnea

The frontline therapy for moderate to severe OSA is the application of positive airway pressure (PAP) to the upper airway to prevent collapse. The pressure must be titrated individually to a level that prevents obstruction (apneas, hypopneas, and snores with arousals) in both NREM and REM sleep and yet is not so high as to interfere with sleep. Levels typically vary between 5 and 15 cm H_2O. Although remarkably effective for OSA, the therapy is an air pump connected to a relatively airtight mask or nasal interface, and as such this system can be cumbersome and uncomfortable. Consequently, less than half of the individuals prescribed PAP regularly use this therapy. Nonetheless, every effort should be made to encourage use of PAP regularly because this is the only therapy for OSA shown to lessen cardiovascular and neurobehavioral morbidity, even in mild to moderate OSA. Masks and nasal interfaces are more comfortable than the first equipment designed, and many patients who thought they could never tolerate a PAP mask are now able to comply with newer interfaces. Patients with claustrophobia and mask intolerance may undergo behavioral therapy for PAP desensitization. Developments in PAP therapy include machines that can self-adjust the level of PAP based on airflow patterns, gradually increasing pressure as needed across sleep stages and different sleeping positions.

Other therapies for OSA should be considered in young individuals with mild sleep apnea and in individuals with mild disease and inability to desensitize to PAP. These alternative therapies include surgical procedures to shorten the soft palate and reduce collapsibility of the pharynx (uvulopalatoplasty), reduce the tongue volume (genioglossectomy), or advance the genioglossus forward (genioglossus advancement hyoid myotomy) as initial surgical therapies and, if clinically indicated, a second phase of surgery to more dramatically increase pharyngeal space (maxillary advancement or maxillary and mandibular osteotomy). Laser-assisted uvulopalatoplasty is not as effective as surgical uvulopalatoplasty

for treating sleep apnea. Clinical indications for use of temperature-controlled radio frequency are still being determined. Some patients who do not tolerate PAP or in whom OSA is mild may benefit from oral appliances for sleep-disordered breathing. Although well tolerated, therapy takes weeks to months to reach optimal effectiveness. Factors predicting success with oral appliances include female gender, non-obese body habitus, young age, and positional OSA with more events in the supine position.

Weight loss should be recommended for all obese individuals with sleep-disordered breathing. Dietary counseling should be the first step taken, and all patients should understand that reduced caloric intake is the primary principle for most, if not all, successful weight loss programs. Exercise may help maintain weight, but in most nonathletic individuals, healthy caloric restriction should be the primary strategy for weight loss. Bariatric surgery is typically reserved for individuals with morbid obesity because there is a 10% chance of significant perioperative adverse events and a 1% chance of mortality in the postoperative period. The majority of individuals who have substantial weight loss after bariatric surgery will experience marked reductions in OSA, if not reversal of the disease. Treatment of endocrine disorders and OSA (hypothyroidism and acromegaly) should begin with PAP therapy, but across the treatment of the underlying endocrine disorder, the PAP may need adjusting. Several medical therapies for OSA may be considered as second-line therapies for mild OSA. There may be subsets of individuals who respond to supplemental oxygen, positional therapy, and, rarely, pharmacotherapies such as selective serotonin reuptake drugs in individuals with mild REM sleep-predominant apnea. Because these therapies are only rarely effective, treatment success should be determined with repeated polysomnography during treatment.

Central Sleep Apnea

Treatment of CSA varies with the cause of sleep apnea. PAP initially appeared promising for treating Cheyne–Stokes respiration in heart failure, but with newer medical therapy including treatment of the sympathetic overdrive, indications for PAP are less clear. High-altitude CSA may be treated with descent, acetazolamide, and medications reducing the hypoxia–angiotensin response or with supplemental oxygen. For other forms of CSA, ventilation must be augmented in sleep, and this may be achieved with PAP that assists ventilation rather than acts as an air splint. This bilevel PAP may be delivered with a mask as in OSA or, rarely, by tracheostomy tube if the noninvasive delivery fails.

Morbidities and Disease Interactions

One of the most significant advances during the past decade regarding OSA has been the identification of OSA as a risk factor for numerous and diverse morbidities. OSA is a major risk factor for excessive sleepiness and sleepiness-related impaired performances. Indeed, OSA is an independent risk factor for motor vehicle crashes, raising the relative risk by 2.5-fold. A direct link is supported by the reduction in car crash risk with successful treatment of sleep apnea. Several reports suggest that OSA is an independent risk factor for insulin resistance. This risk remains after controlling for obesity. As with motor vehicle accidents, further support for this disease interaction comes from studies showing improvement in glucose control with successful use of PAP therapy. It will be exciting to determine whether long-term PAP therapy reduces the occurrence of complications of diabetes. Several reports suggest that OSA may impair liver function and might contribute to nonalcoholic fatty liver disease, a major risk factor for liver failure in developed countries. Perhaps the most impressive observation in individuals with OSA has been the elucidation of OSA as an independent risk factor for several cardiovascular diseases, including hypertension, congestive heart failure, and stroke. Here again, obesity was controlled for. Importantly, the relative risk for hypertension increases even at levels of mild OSA (5–15 events per hour). The rates of significant cardiovascular events during 10-year follow-up in a large European trial also increased with mild to moderate OSA, but to a far greater extent (fourfold increase) in individuals with severe untreated OSA. This risk in individuals with severe OSA declines almost to baseline when treated effectively with PAP across the duration of follow-up. The risk of cardiovascular death is also reduced with PAP therapy in people with severe OSA. In light of the seriousness of morbidities and the disease interactions associated with OSA, every effort to improve PAP therapy acceptance and use should be made.

Future Directions

This area of research is extremely vibrant, and consequently significant progress has been made in understanding the pathophysiology, epidemiology, genetics, therapies, and disease interactions of sleep apnea. In light of the serious morbidities associated with OSA, future directions should emphasize disease prevention, improved detection of mild disease, early disease intervention, and development of novel well-tolerated effective therapies. As disorders with diverse systemic

effects, furthering of multidisciplinary approaches to patient care will be ever more important in the future.

See also: Cataplexy; Narcolepsy; Sleep in Adolescents; Sleeping Sickness.

Further Reading

Budhiraja R and Quan SF (2005) Sleep-disordered breathing and cardiovascular health. *Current Opinion in Pulmonary Medicine* 11(6): 501–506.

Culebras A (2005) Sleep apnea and stroke. *Reviews in Neurological Diseases* 2(1): 13–19.

Gozal D (2004) New concepts in abnormalities of respiratory control in children. *Current Opinion in Pediatrics* 16(3): 305–308.

Javaheri S (2005) Central sleep apnea in congestive heart failure: Prevalence, mechanisms, impact, and therapeutic options. *Seminars in Respiratory and Critical Care Medicine* 26(1): 44–55.

Lavie L (2005) Sleep-disordered breathing and cerebrovascular disease: A mechanistic approach. *Neurologic Clinics* 23(4): 1059–1067.

Narkiewicz K, Wolf J, Lopez-Jimenez F, and Somers VK (2005) Obstructive sleep apnea and hypertension. *Current Cardiology Reports* 7(6): 435–440.

Pack AI (2006) Advances in sleep-disordered breathing. *American Journal of Respiratory and Critical Care Medicine* 173(1): 7–15.

Remmers JE (2005) A century of control of breathing. *American Journal of Respiratory and Critical Care Medicine* 172(1): 6–11.

Schwab RJ (2005) Genetic determinants of upper airway structures that predispose to obstructive sleep apnea. *Respiratory Physiology & Neurobiology* 147(2–3): 289–298.

Strohl KP (2003) Periodic breathing and genetics. *Respiratory Physiology & Neurobiology* 135(2–3): 179–185.

White DP (2006) The pathogenesis of obstructive sleep apnea: Advances in the past 100 years. *American Journal of Respiratory Cell and Molecular Biology* 34(1): 1–6.

Young T, Palta M, Dempsey J, Skatrud J, Weber S, and Badr S (1993) The occurrence of sleep-disordered breathing among middle-aged adults. *New England Journal of Medicine* 328(17): 1230–1235.

Sleep Architecture

E F Pace-Schott, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Human Sleep

Human sleep is a behavioral state that alternates with waking, relative to which it is characterized by a heightened threshold to sensory input, attenuation of motor output, characteristic changes in central and peripheral physiology, and diminished conscious awareness. It is during this state of altered consciousness that dreaming occurs. Two major types of sleep, rapid eye movement (REM) and non-rapid eye movement (NREM), are characteristic of all mammals and alternate with regular, species-specific periods, indicating that they are timed by a reliable oscillator. In adult humans, this period is 90–120 min, resulting in approximately four or five cycles per night, as shown in Figure 1.

Polysomnographic Recording of Sleep

Since the late 1960s, sleep stages have been visually scored using the standardized electroencephalographic (EEG), electrooculographic (EOG), and electromyographic (EMG) criteria of the 1968 manual developed by Alan Rechtschaffen and Anthony Kales. EEG, EOG, and EMG are collectively termed polysomnography (PSG). Figure 2 shows the sites of placement of electrodes for recording these three measures, which constitute oscillating field potentials from the underlying cortex (EEG), relative distance of the electronegative retina from the electrode (EOG), and muscle tone (EMG). In recent years, these electrophysiological descriptors of sleep stage have been expanded to include results from functional neuroimaging (e.g., functional magnetic resonance imaging), quantitative EEG (e.g., spectral analysis), and magnetoencephalography (MEG). Although these new technologies are rapidly expanding our understanding of sleep neurophysiology, the PSG criteria of Rechtschaffen and Kales remain the gold standard by which the progression of sleep stages (or 'sleep architecture') is routinely characterized in clinical and research sleep labs. (Note, however, that the American Academy of Sleep Medicine published a substantive update of the Rechtschaffen and Kales manual in 2007.) The following sections first characterize sleep architecture in terms of the traditional PSG measures and then describe cellular neurophysiology relevant to the control of the sleep–wake and REM–NREM cycles and, lastly, findings on regional brain activation in sleep.

Sleep Architecture

Before commencing with a description of sleep architecture, it is important to define three terms used to characterize the EEG: frequency, amplitude, and morphology. EEG measures changes in electrical potential on the scalp that, in turn, reflect fluctuation of membrane potentials integrated across large numbers of cortical neurons beneath the recording electrodes. Such fluctuations or 'brain waves' are rhythmic, and their features reflect varying degrees of synchrony in membrane potential changes among neurons in these large aggregates. Frequency refers to the rapidity of fluctuation in electrical potential between successive peaks (or successive troughs) of measured EEG brain waves and is measured in cycles per second or Hertz (Hz). Amplitude refers to the magnitude of electrical potential change between such waves' peaks and troughs and is measured in (micro) volts. Morphology refers to the shape of a particular EEG wave or group of waves, reflecting both their frequency and amplitude, which can assume characteristic forms such as the sleep spindle or K-complex described later.

The initiation of NREM sleep is gradual and is characterized by a slowing in frequency of brain waves in the EEG from predominance of beta (14–25 Hz) and gamma (25–80 Hz) frequencies seen with eyes open (Figure 3(a)) to alpha (8–13 Hz) with eyes closed (Figure 3(b)) and finally to a predominance of theta rhythm (4–7 Hz) in stage 1 NREM sleep (Figure 3 (c)). The subsequent increasing depth of NREM sleep through stages 2–4 can be generally characterized as a progressive decrease in the predominant EEG frequency along with a progressive increase in average wave amplitude (Figures 3(d), (e) and (g)). In stage 1 NREM, characteristic slow eye movements are recorded in the EOG and muscle tone (EMG) is diminished relative to waking (Figure 4). Stage 1 is succeeded by stage 2 sleep, with an EEG characterized by further decreased mean frequency and increased amplitude of brain waves (Figure 3(d)). Stage 2 EEG is also characterized by two distinctive wave forms: sleep spindles, which are intermittent clusters of spikes at sigma frequency (12–14 Hz) with characteristic waxing and waning morphology, and the K-complex, which is a large negative (upward) deflection followed by a positive (downward) deflection (Figure 3(d)). In stage 3 sleep, sleep spindles decrease, EEG frequency further slows, and the amplitude of brain waves increases.

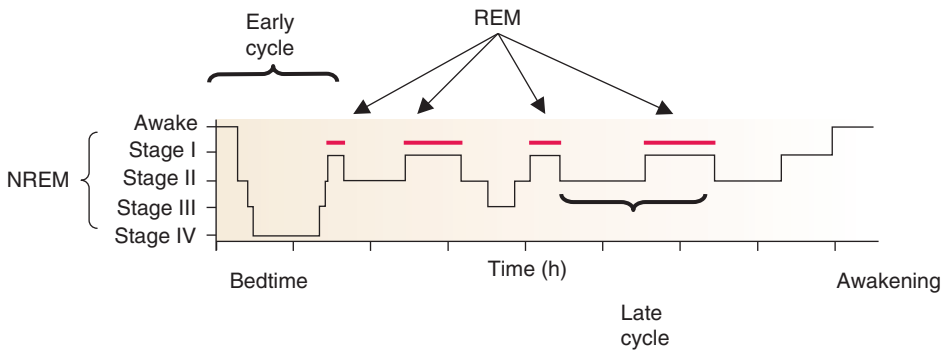

Figure 1 A night of sleep shown in the form of a 'hypnogram' in which the vertical axis shows depth of NREM sleep, red bars show REM, and the 90–120 min periodicity of each sleep cycle can be estimated as the duration from the end of one REM period to the end of the next one. Note that early in the night NREM is deeper and occupies more of the sleep cycle, whereas toward the end of the night, NREM sleep is lighter and more time is spent in REM. Adapted from Pace-Schott EF and Hobson JA (2002) The neurobiology of sleep: Genetic mechanisms, cellular neurophysiology and subcortical networks. *Nature Reviews Neuroscience* 3: 591–605, with permission.

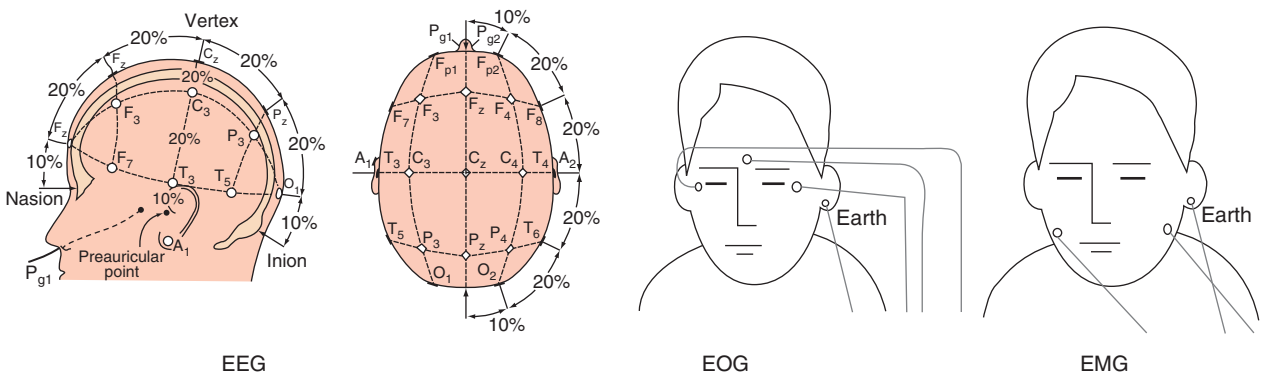

Figure 2 Placement of electrodes for polysomnography (PSG) that includes electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG).

Very high-amplitude delta waves (1–3 Hz) occur in the deepest NREM sleep, stages 3 and 4, which are also known as slow-wave sleep (SWS) or delta sleep. A slower oscillatory rhythm of less than 1 Hz, the 'slow oscillation,' has been described in cats and humans. In healthy young adults between ages 20 and 40 years, stage 1 NREM occupies approximately 2–5% of total sleep, stage 2 occupies 45–55%, stage 3 occupies 3–8%, and stage 4 occupies 10–15%.

REM sleep, discovered in 1953 by Eugene Aserinsky and Nathaniel Kleitman, has also been termed 'paradoxical' sleep because, by contrast with NREM, frequencies in the EEG resemble those of waking but muscle tone reaches its minimum level (Figure 3(f)). This EMG minimum in REM results from central, brain-stem-mediated inhibition of spinal alpha motor neurons ('REM atonia'). Because of its wake-like EEG (Figures 3(f) and 3(g)), REM has also been termed 'active' sleep. Also, because of its low-amplitude, high mixed-frequency waves, it has been termed 'desynchronized' sleep, in contrast with NREM (especially SWS), in which high-amplitude waves reflect synchronous activity of cortical neurons. As its name suggests, saccadic rapid eye movements (REMs) occur during REM sleep (Figure 4). These REMs, which occur in singlets and clusters, reflect the fact that unlike skeletal muscles, oculomotor muscles are not atonic in REM. REMs are believed to reflect underlying, central phasic (brief) neural events that can also be measured as peripheral muscle twitches (that briefly overcome REM atonia), middle-ear muscle movements, and periorbital potentials. Such events, in humans, are likely related to ponto-geniculo-occipital waves – phasic potentials recorded sequentially in the pons, thalamic lateral geniculate body, and the occipital cortex that are a characteristic electrophysiological feature of REM in the cat. Unlike REM's phasic events, its tonic (persisting), wake-like EEG is believed to reflect cortical activation by the cholinergic ascending arousal systems described later. An additional EEG feature of REM sleep, the 'sawtooth wave,' is illustrated in Figure 3(f).

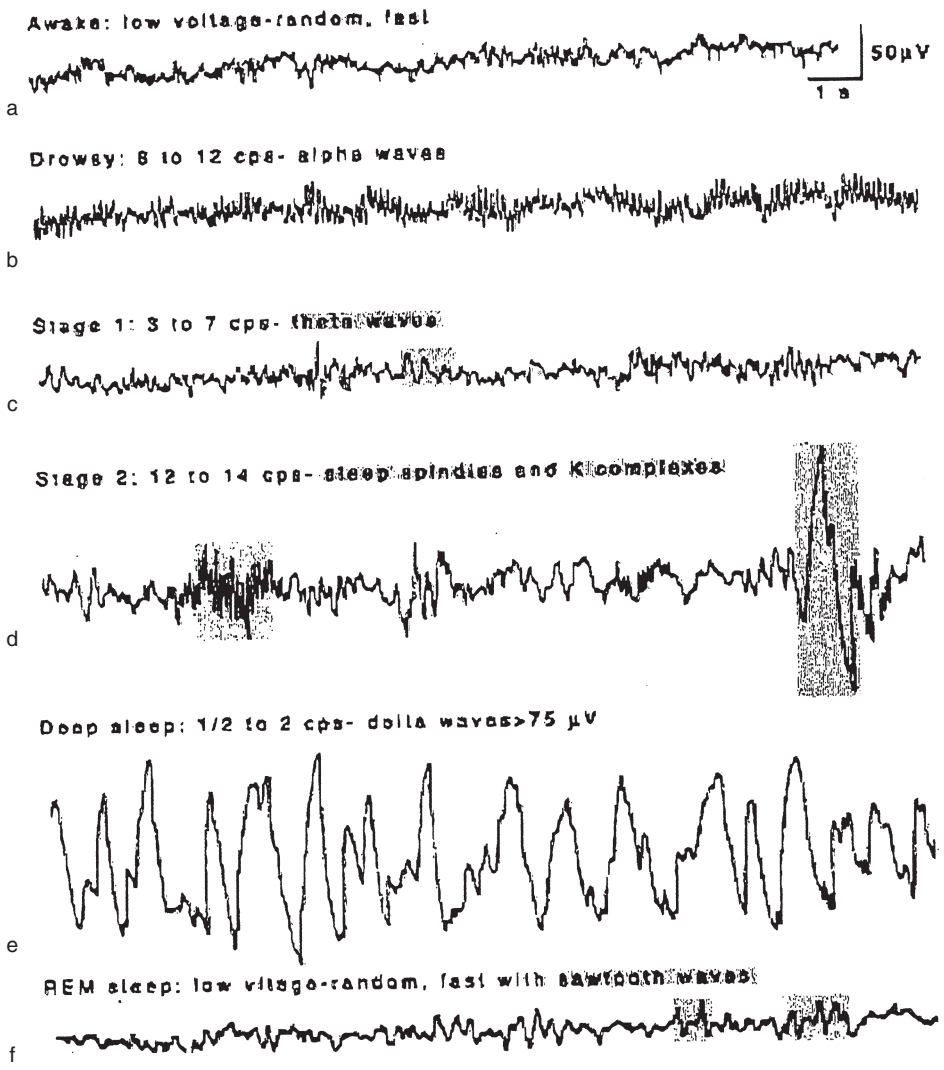

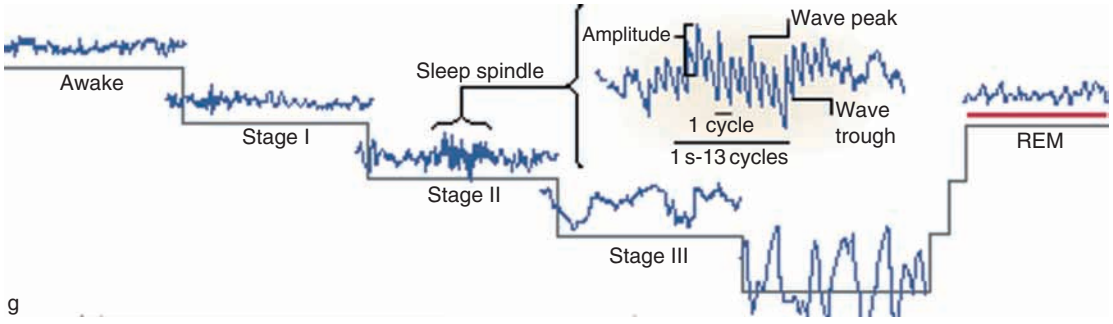

Figure 3 Depth of sleep is characterized by the frequency and amplitude of EEG waveforms with slower frequency and higher amplitude indicative of greater depth. The dominant EEG frequencies waking and sleep stages: (a) waking with eyes open (beta); (b) waking with eyes closed (alpha); (c) stage 1 NREM sleep (theta); (d) stage 2 NREM sleep (theta frequency with sleep spindle and K-complex wave forms); (e) stages 3 and 4 NREM sleep collectively termed 'slow-wave sleep' (delta); (f) REM sleep (wake-like EEG, characteristic saw-tooth waves); (g) a single sleep cycle from early in the night showing the complete series of stages from waking to NREM stages 1–4 and back through these stages to REM. (a–f) Hobson JA and Pace-Schott EF (2003) Sleep, dreaming, and wakefulness. In: Squire L, Bloom F, McConnell S, Roberts J, Spitzer N, and Zigmond M (eds.) *Fundamental Neuroscience*, 2nd edn., ch. 42. San Diego: Academic Press, with permission from Elsevier. (g) Pace-Schott EF and Hobson JA (2002) The neurobiology of sleep: Genetic mechanisms, cellular neurophysiology and subcortical networks. *Nature Reviews Neuroscience* 3: 591–605, with permission.

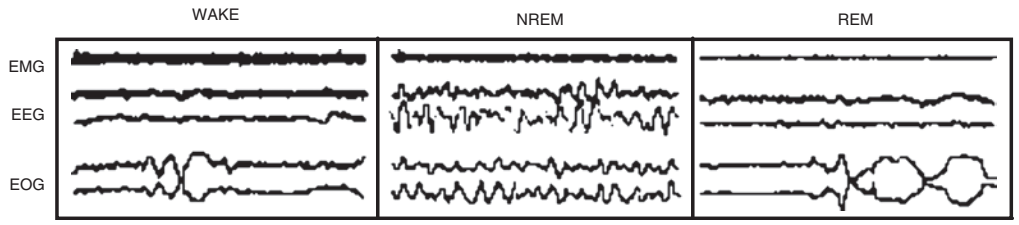

Figure 4 The combined EEG, EOG, and EMG of waking, NREM, and REM sleep. Note the decline in muscle tone from wake to NREM to REM, the similarity of waking and REM EEG, and the rapid eye movements in the EOG for which REM sleep is named. Reproduced from Hobson JA and Pace-Schott EF (2003) Sleep, dreaming, and wakefulness. In: Squire L, Bloom F, McConnell S, Roberts J, Spitzer N, and Zigmond M (eds.) *Fundamental Neuroscience*, 2nd edn., ch. 42. San Diego: Academic Press, with permission from Elsevier.

Autonomic and Peripheral Physiological Changes in Sleep

In NREM sleep relative to waking, heart rate, blood pressure, and respiration rate decline (Figure 5), sympathetic outflow of the autonomic nervous system decreases, and parasympathetic activity increases. Parasympathetic activity, as measured by heart rate variability indices, becomes closely linked to changes in the delta power band. In REM, heart rate, blood pressure, and respiration rate again rise but, as noted previously, muscle tone remains actively suppressed. In REM, homeostatic responses are attenuated. For example, respiration becomes insensitive to blood CO_2 and thermoregulation is largely absent. Additional peripheral features of REM include penile erections and clitoral engorgement. In the transition from NREM to REM sleep, sympathetic activity increases and parasympathetic activity decreases, but the parasympathetic:sympathetic ratio remains high as evidenced by pupillary miosis.

Neurophysiological Basis of Sleep Stage Changes

Control of sleep onset from the waking state is believed to be a function of the hypothalamus. Clifford Saper and colleagues have shown that there is a population of neurons in the anterior hypothalamus that, unlike much of the brain, shows increased metabolic activity during sleep (Figure 6). These neurons are concentrated in the ventrolateral preoptic area (VLPO) and produce both the inhibitory amino acid γ-aminobutyric acid (GABA) and the inhibitory neuropeptide galanin. At sleep onset, these neurons become active and inhibit ascending arousal systems of the brain stem, posterior hypothalamus, and basal forebrain (described later), thereby effecting the transition from the waking state to NREM sleep.

Once NREM is initiated, its delta and slow (<1 Hz) oscillatory rhythms, as well as its spindle and K-complex wave forms are generated by reciprocally interconnected thalamic and cortical neural circuits that have been extensively investigated and described by the late Mircea Steriade and colleagues. When the influence of subcortical ascending arousal systems decreases following sleep onset, thalamocortical circuits shift into an oscillatory mode during which transmission of sensory information to the cortex is blocked. Intrinsic oscillations of NREM represent the combined influence of two neuronal mechanisms: (1) the inhibitory influence of neurons of the thalamic reticular nucleus – a thin sheet of GABAergic neurons surrounding the thalamic periphery – on thalamic neurons projecting to the cortex, and (2) an oscillation within cortical neurons involving alternation between a prolonged hyperpolarization phase and a shorter depolarized phase with rapidly repeated spiking. This latter rhythm, generated within cortical networks, is the neuronal basis of the <1 Hz slow oscillation in the EEG and has been shown in animal models to exert an organizing or grouping influence on other intrinsic NREM sleep rhythms such as sleep spindles and the K-complex.

During waking, NREM oscillatory rhythms are prevented due to neuromodulation of thalamocortical circuits and cortex by ascending projections >from arousal-promoting systems of the brain stem, hypothalamus, and basal forebrain (Figure 6). Such neuromodulators include (1) acetylcholine from the midbrain–pons (mesopontine) junction in the brain stem, as well as from the basal forebrain (e.g., nucleus basalis); (2) the monoamines including norepinephrine (NE) and serotonin (5-HT) from the locus coeruleus and dorsal raphe, respectively, in the mesopontine brain stem dopamine from the midbrain periaqueductal gray, and histamine from the posterior hypothalamus; and (3) the neuropeptide orexin from the lateral hypothalamus. In contrast, during REM sleep, the intrinsic rhythms of NREM are prevented mainly by influence on the thalamus and basal forebrain from the cholinergic (acetylcholine) subcomponent of these systems originating in the mesopontine brain stem, including the laterodorsal tegmental and pedunculopontine nuclei.

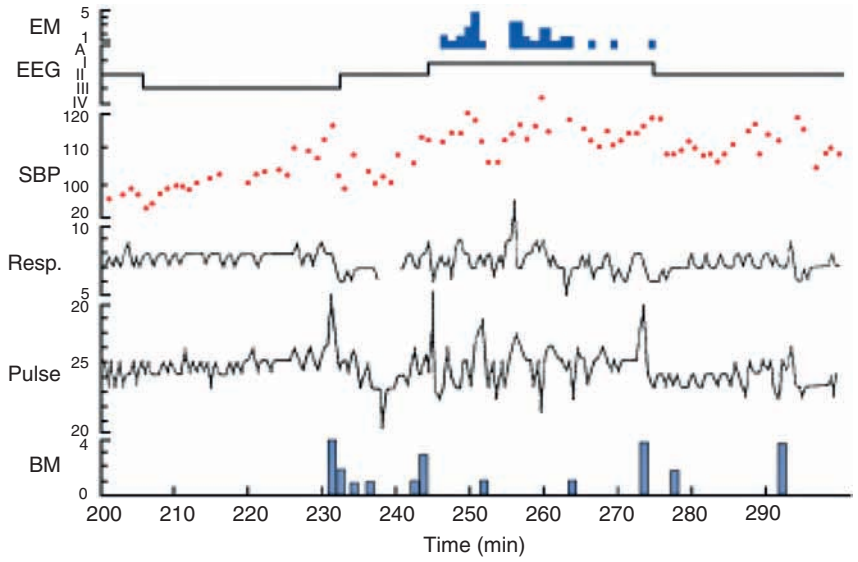

Figure 5 Peripheral physiological changes accompanying one sleep cycle. The EM trace shows the number of rapid eye movements per minute indicating the occurrence of REM sleep. The EEG trace is a hypnogram as in **Figure 1**. The SBP trace is systolic blood pressure, the Resp trace is respiration rate, the Pulse trace is heart rate, and the BM trace is the number of body movements per minute (movements increase at the beginning and end of REM). Reproduced from Hobson JA and Pace-Schott EF (2003) Sleep, dreaming, and wakefulness. In: Squire L, Bloom F, McConnell S, Roberts J, Spitzer N, and Zigmond M (eds.) *Fundamental Neuroscience*, 2nd edn. San Diego: Academic Press, with permission from Elsevier.

Homeostatic and Circadian Control of Sleep Onset

A two-process model of sleep propensity, proposed by Alexander Borbely, suggests that homeostatic mechanisms (process S) interact with circadian factors (process C) to bring about sleep–wake transitions. Process S refers to a drive to sleep that increases as a function of the duration spent awake, whereas process C refers to sleep propensity varying as a function of time of day. The physiological basis of process S is believed to be substances that build up in the brain as a function of time awake ('endogenous somnogens'), and the purine nucleoside adenosine is currently the most widely suggested candidate for such a substance, although others have also been proposed (e.g., interleukin-1, prostaglandin D1, and growth hormone releasing hormone). Robert McCarley and colleagues have suggested that the specific site at which adenosine acts to promote sleep is in the basal forebrain, although its somnogenic effects have been reported at other subcortical sites as well.

Sleep propensity due to process C is an example of a wide variety of physiological circadian rhythms whose 'master clock' resides in the suprachiasmatic nucleus (SCN) of the anterior hypothalamus. Cells of the SCN contain molecular clocks that maintain a near-perfect 24-h periodicity via interlocking positive and negative feedback control of transcription and translation of circadian genes. These SCN cells can become entrained to the ambient photoperiod by light impinging on retinal ganglion cells that contain photoreceptive pigments such as melanopsin and which communicate monosynaptically with the SCN via the retinohypothalamic tract. Clifford Saper and colleagues have shown that the SCN, in turn, projects to various hypothalamic nuclei with circuits through subparaventricular and dorsomedial nuclei able to influence sleep–wake transitions via projections to the VLPO. Secretion of the pineal sleep hormone melatonin is also controlled by the SCN via a multisynaptic pathway and provides an internal signal of circadian time, as well as feedback to the SCN via its melatonin receptors.

Recent Discoveries Regarding the Brain in Sleep

The advent of new technologies in brain imaging has resulted in a greatly increased understanding of the behavioral and electroencephalographic features of human sleep. Positron emission tomography (PET) studies indicate that global cerebral oxygen utilization (detected using $H_2^{15}O$) and glucose metabolism (detected using ^{18}fluorodeoxyglucose) is decreased during NREM sleep compared to waking. Relative to waking, glucose metabolism is decreased by approximately 11% in stage 2 NREM and approximately 40% in SWS. fMRI studies have shown similar declines with deepening NREM sleep in the cerebral blood oxygen level-dependent signal that is measured using this technique. In contrast, during

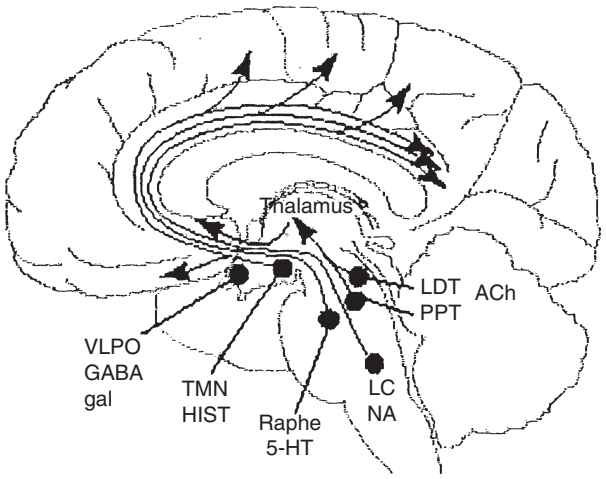

Figure 6 Major components of sleep inducing and wake-inducing neuromodulatory systems. To induce sleep, projections from the gamma-aminobutyric acid (GABA) and galanin (gal) producing cells of the ventrolateral preoptic nucleus (VLPO) of the anterior hypothalamus send inhibitory signals to each of the ascending arousal systems of the brain stem, basal forebrain, and hypothalamus. These arousal systems include: the tuberomammillary nucleus (TMN) of the posterior hypothalamus that produce histamine (HIST), serotonin (5-HT) producing cells of the dorsal raphe (Raphe), acetylcholine (ACh) producing cells of the laterodorsal tegmental (LDT) and pedunculopontine tegmental (PPT) nuclei and norepinephrine (NA) producing cells of the locus coeruleus (LC). Other ascending arousal systems not shown include the perifornical lateral hypothalamus that produces orexin, dopamine producing cells of the midbrain periaqueductal gray, and cholinergic projections from the basal forebrain (nucleus basalis, diagonal band of Broca, medial septum) to the limbic system and the cortex. Reproduced from Saper CB, Chou TC, and Scammell TE (2001) The sleep switch: Hypothalamic control of sleep and wakefulness. *Trends in Neurosciences* 24: 726–731, with permission from Elsevier.

REM, global cerebral energy metabolism increases so that it is equal to or even greater than that of waking.

Deactivation of frontal cortices is one of the first physiological signs of human sleep. This has been observed whether activation is measured by EEG, MEG, PET, or fMRI. Functional imaging studies have also shown deactivation of the thalamus to be an early event in sleep onset. As NREM sleep deepens, more posterior cortices also decrease their neuronal activity, as do subcortical areas including the hypothalamus, basal ganglia, and hippocampus. In addition to their earlier deactivation in NREM, frontal areas may also display a greater need for recovery sleep than more posterior regions. Alexander Borbely and colleagues have shown that slow-wave activity (SWA), measured as the spectral power in delta frequencies, can serve as a marker of homeostatic sleep pressure. SWA is inversely related to cortical activation, is greater in frontal than in parietal and occipital regions in the first NREM episode of the night, and spreads in a frontal to posterior direction within the cortex. In addition to being the first part of the cortex to 'go to sleep,' frontal areas also appear to be the last part to wake up because their reactivation lags behind that of the rest of the brain.

After the transition from NREM to REM, much of the lateral frontal cortex remains less active than during waking. However, PET studies have shown that blood flow and glucose metabolism increase during REM in subcortical brain regions such as the pons, midbrain, thalamus, amygdala, hypothalamus, and basal ganglia, as well as limbic-related cortices such as parahippocampal cortex, temporal pole, anterior insula, caudal orbitofrontal cortex, subcallosal anterior cingulate, and medial prefrontal cortex (e.g., Brodmann areas 24, 32, and 10). Eric Nofzinger and colleagues have termed these midline structures an 'anterior paralimbic REM activation area' and have shown that their reactivation during REM following low activity in NREM can be as great or greater than their activation during waking.

Summary

Sleep, consisting of alternating REM and NREM states, is a universal behavior of all mammals. In humans, this cycle has a periodicity of 90–120 min and consists of characteristic substages of NREM characterized by brain waves of increasing amplitude and decreasing frequency as NREM deepens. The oscillating EEG patterns of NREM reflect synchronous activity in thalamocortical circuits. REM sleep is characterized by a unique, paradoxical combination of a fast, wake-like EEG but the least muscle tone of any stage. Characteristic peripheral and autonomic changes accompany stages of the sleep cycle with higher, more variable respiration, pulse, and blood pressure during REM than during NREM. Sleep propensity is determined by the interaction of a circadian rhythm ('process C') that is controlled, like other such rhythms, by the biological clock in the suprachiasmatic nucleus and a sleep homeostatic 'process S' that increasingly drives the need for sleep the longer one is awake. Waking is maintained by the ascending arousal systems of the brain stem, hypothalamus, and basal forebrain that send projections containing wake-promoting neuromodulators such as serotonin, norepinephrine, acetylcholine, histamine, orexin, and dopamine to the thalamus, limbic system, and cerebral cortex. During sleep, inhibitory projections from sleep-active sites in the anterior hypothalamus suppress these same arousal systems. In REM, mainly one component of these wake-active systems, that of

acetylcholine, activates the forebrain. Functional neuroimaging studies have revealed that NREM sleep is associated with widespread cortical deactivation relative to waking that is particularly pronounced in frontal areas. During REM, midline limbic and paralimbic areas are selectively activated while much of the lateral cortex remains in the less activated state characteristic of NREM.

See also: Circadian Genes and the Sleep–Wake Cycle; Circadian Rhythms in Sleepiness, Alertness, and Performance; Sleep Research and Sleep Medicine in Historical Perspective; Sleep Oscillations and PGO Waves; Sleep and Sleep States: PET Activation Patterns; Sleep and Sleep States: Hippocampus–Neocortex Dialog; Sleep and Sleep States: Network Reactivation; Sleep and Sleep States: Gene Expression; Sleep: Development and Circadian Control.

Further Reading

Borbely AA and Achermann P (2005) Sleep homeostasis and models of sleep regulation. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 405–417. Philadelphia: Saunders.

Braun AR, Balkin TJ, Wesensten NJ, et al. (1997) Regional cerebral blood flow throughout the sleep–wake cycle. *Brain* 120: 1173–1197.

Carskadon MA and Dement WC (2005) Normal human sleep: An overview. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 13–23. Philadelphia: Saunders.

Hobson JA (1989) *Sleep*. New York: Scientific American Library.

Hobson JA, Pace-Schott EF, and Stickgold R (2000) Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences* 23: 793–842.

Kaufmann C, Wehrle R, Wetter TC, Holsboer F, Auer DP, and Pollmacher T (2006) Brain activation and hypothalamic functional connectivity during human non-rapid eye movement sleep: An EEG/fMRI study. *Brain* 129: 655–667.

Kryger MH, Roth T, and Dement WC (eds.) (2005) *Principles and Practice of Sleep Medicine,* 4th edn. Philadelphia: Saunders.

Lydic R and Baghdoyan HA (eds.) (1999) *Handbook of Behavioral State Control: Molecular and Cellular Mechanisms,* Boca Raton, FL: CRC Press.

Mallick BN and Inoue S (eds.) (1999) *Rapid Eye Movement Sleep*. New York: Dekker.

Maquet P (2000) Functional neuroimaging of normal human sleep by positron emission tomography. *Journal of Sleep Research* 9: 207–231.

Maquet P, Ruby P, Maudoux A, et al. (2005) Human cognition during REM sleep and the activity profile within frontal and parietal cortices: A reappraisal of functional neuroimaging data. *Progress in Brain Research* 150: 219–227.

Nofzinger EA (2005) Functional neuroimaging of sleep. *Seminars in Neurology* 25: 9–18.

Ohayon MM, Carskadon MA, Guilleminault C, and Vitiello MV (2004) Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan. *Sleep* 27: 1255–1273.

Opp MR, Kilduff TS, Marcus CL, Armitage R, and Nofzinger EA (2005) *SRS Basics of Sleep Guide*. Westchester, IL: Sleep Research Society.

Pace-Schott EF and Hobson JA (2002) The neurobiology of sleep: Genetic mechanisms, cellular neurophysiology, and subcortical networks. *Nature Reviews Neuroscience* 3: 591–605.

Saper CB, Scammell TE, and Lu J (2005) Hypothalamic regulation of sleep and circadian rhythms. *Nature* 437: 1257–1263.

Steriade M (2006) Grouping of brain rhythms in corticothalamic systems. *Neuroscience* 137: 1087–1106.

Sleep Deprivation and Brain Function

S P A Drummond, University of California at San Diego and Veterans Affairs San Diego Healthcare System, San Diego, CA, USA
B S McKenna, San Diego State University, University of California at San Diego, and Veterans Affairs San Diego Healthcare System, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The behavioral and cognitive consequences of sleep deprivation have been the subject of scientific investigation for over 100 years, since the pioneering work of Patrick and Gilbert in 1896. In general, sleep deprivation produces neurobehavioral deficits in experimental tasks of alertness, attention, memory, cognition, learning, and motor responses, as well as performance deficits in operational settings (e.g., on the highway and in the factory). Despite the wealth of knowledge concerning the behavioral effects of sleep loss, much less work has examined the neurophysiological effects of sleep loss. This article discusses what we have learned about the effects of sleep deprivation on brain function from the perspective of functional neuroimaging studies. First, the definitions of sleep deprivation and brain function are covered. Then we review the literature in which functional neuroimaging is used to study sleep deprivation. Because the widespread application of functional neuroimaging techniques is still relatively new, there are fewer than 20 such studies published. However, rather than providing a detailed review of each one, we highlight the common findings and provide details only when they are illustrative or of particular interest. Finally, we provide a synthesis of the various findings.

Definitions

Sleep deprivation in humans can be broadly classified into three categories: total sleep deprivation (TSD), partial sleep deprivation (PSD), and sleep fragmentation. TSD is the complete lack of sleep for at least one night and often longer. PSD involves restricted sleep for multiple nights, that is, individuals obtaining an inadequate amount of sleep for several consecutive nights. Sleep fragmentation is repeated awakenings from sleep throughout the night. This results in a decreased amount of sleep but a normal time spent in bed. This article focuses on the research on TSD because all published neuroimaging studies in healthy adults to date have used this paradigm.

The term 'brain function' can be used to describe activity ranging from the cellular level (e.g., production of neurotransmitters) to the systems level (e.g., interactions among several brain regions). Functional neuroimaging refers to a set of technologies that allow researchers to examine brain activity with various spatial and temporal resolutions and relate that activity to behavior. Here we focus on two aspects of brain function typically examined in functional neuroimaging studies: (1) that cognitive performance is a behavioral measure often conceptualized as the output of brain function and (2) that functional neuroimaging techniques all provide some measure of brain metabolism. The two functional neuroimaging techniques that have been used most commonly to understand the effects of sleep deprivation are positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). The specific PET technique used measures glucose metabolism over the course of 45 min (although the signal is dominated by activity in the first 20 min). The measure of brain function obtained in the fMRI studies is a complex signal called the blood oxygen-dependent (BOLD) signal. When neurons in a local area increase firing rates (e.g., in response to a cognitive demand), blood flow and blood volume in that area increase. These, in turn, increase oxygen delivery to the area. However, more extra oxygen is delivered than is needed, and the net result is an elevated level of oxyhemoglobin in the local region of activity. It is this elevated level of oxyhemoglobin that produces the BOLD signal. fMRI has a temporal resolution on the order of seconds. Both techniques allow investigators to examine the regions of the brain that are involved in processing a cognitive task being performed by a subject. Brain regions showing positive signal changes when an individual performs the task are said to be activated.

Functional Neuroimaging Studies of Sleep Deprivation

This section reviews the findings from the PET and fMRI studies examining sleep deprivation in healthy young adults. The various reports are categorized by the type of cognitive task examined. Not surprisingly, the type of task has a tremendous influence not only on the brain regions activated when the

subject is well rested but also on how brain activation changes with sleep deprivation.

Attention Tasks

There are many types of attention (e.g., sustained, selective, spatial, and divided), but functional neuroimaging studies have examined only sustained attention and short-term visual attention. Sustained attention is the ability to focus on a single task for an extended period of time. Visual attention refers to the need to process a short stream of visual stimuli and respond to a specific target. In well-rested individuals, sustained attention relies on the right middle frontal gyrus and inferior parietal lobe and, to a lesser extent, on the left inferior parietal lobe and bilateral thalamus. Shorter-term visual attention tasks rely more heavily on the extrastriate cortex, as well as the fusiform, lingual, and superior parietal regions, depending on the specific target of attention.

The first study to use neuroimaging to examine TSD was a PET study involving a sustained attention task in which individuals pressed a button every 3 s for 45 min. Along with a decrease in alertness, this study showed decreased whole-brain metabolism of glucose following approximately 32 h of TSD. In addition, the frontal and temporal lobes showed a decrease in absolute metabolic rates, whereas the parietal lobes showed an increase and the occipital lobes showed no change. The largest decreases, however, were seen in the thalamus, basal ganglia, and cerebellum. These findings are consistent with a later PET study of sustained attention in which subjects performed a serial addition task for 45 min. That study showed a significant drop in cerebral metabolic rates in most brain regions after 24 h of TSD, which continued throughout 72 h of TSD. Both these studies used very long tasks and only examined brain function averaged across the entire task.

Two other studies used fMRI to examine attention after TSD on a finer temporal scale. The first measured visual attention over 40 s and found that increased activation in the ventrolateral thalamus supports intact attention after TSD. The final study used an even finer temporal scale to examine performance on the psychomotor vigilance task (PVT), which measures sustained attention and arousal lability. In this task, an individual presses a button as soon as a millisecond counter starts, which begins randomly within a predetermined interval. The task lasts 10 min and requires sustained attention. Within a testing session, performance on the PVT has been thought to reflect state instability (the variability (waxing and waning) in performance produced by the increased sleep drive during TSD). Findings showed that the frontoparietal sustained attention network supported fast reaction times whether the subjects were well rested or sleep deprived. These brain regions did not respond during slow reaction times under normal sleep conditions (which is probably why those reaction times were slow). Interestingly, however, these same regions showed increased responses immediately following slow reaction times after TSD. This was thought to reflect individuals reengaging in the task following a slow response because they received feedback regarding this slow reaction time. In essence, then, a slow reaction time prompts compensatory effort from an individual, and this compensatory response is captured in fMRI as an increase in activation in the frontoparietal sustained attention network following the slowest reaction times. A final noteworthy finding is that the study also suggests that the mechanism underlying poor performance after TSD is either (1) the impaired ability of the brain to appropriately allocate resources to task-specific brain regions or (2) the inability to inhibit activation within brain regions that actively interfere with performance.

Overall, it appears that TSD results in decreased brain responses in regions underlying sustained attention demands, as well as diminished performance on sustained attention tasks. Nonetheless, it appears that sleep-deprived individuals can indeed engage these attention regions for brief periods while sleep deprived, but only when accompanied by compensatory activation within the thalamus. This brief engagement allows for intermittent normal levels of performance, despite the fact that average performance across the entire task is diminished with TSD. This set of findings illustrates the importance of the temporal scale at which brain function is measured during TSD.

Learning Tasks

Learning is an essential task for humans, and there are multiple types of learning and memory that rely on various brain systems. Verbal learning (i.e., memorization of language-based material) is the only type of learning studied with functional neuroimaging techniques in the context of sleep deprivation. Studies of TSD have examined brain function during the encoding of verbal material, but none has directly studied recall or recognition of learning materials. Cortical regions associated with verbal encoding are almost exclusively in the left hemisphere and include the medial temporal structures, inferior frontal gyrus, premotor areas, middle temporal gyrus, and cerebellum.

Unlike sustained attention tasks, in which brain function generally decreases with TSD, verbal encoding tasks elicit increased activation in several brain

regions after TSD. The most reliable increases are seen in the left and right inferior frontal gyri, as well as the left-hemisphere inferior and superior parietal lobes. This increased activation is typically associated with intact, or only minimally impacted, performance on the learning tasks. In fact, the increased activation within the parietal lobes is directly related to performance after TSD; individuals showing the greatest activation within the left inferior parietal regions after TSD show the best learning performance after TSD.

In addition, the difficulty level of the verbal learning task also influences brain function with TSD. This was demonstrated in a study comparing learning an easy word list versus learning a hard word list. During the memorization of easy words, brain activation did not differ between well rested and TSD states. Memorization of hard words, however, elicited strong increases in brain activation after TSD, and those increases were directly related to performance. The influence of task difficulty in verbal learning has also been examined by employing a distracter task to interfere with learning. That study showed a similar result, with the learning-plus-distraction condition showing greater cerebral responses after TSD than after a normal night of sleep. Once again, the greater the increase in activation with TSD, the better the subjects performed on the learning task. Interestingly, in all the studies that examined verbal learning, the left parietal lobes showed increased activation following TSD and individuals showing the greatest increases in this region also showed the best performance after TSD. This suggests that the left inferior parietal lobe plays a key role in maintaining verbal learning performance during TSD.

Finally, one recent study examined cerebral responses during a verbal learning task from a network perspective using functional connectivity analyses. This study showed that memorization of easy words and hard words each activated the same brain regions (the left and right inferior frontal gyrus and left inferior and superior parietal lobes) but that the interconnections among those regions differed. Following TSD, this study found that the interactions among brain regions within these networks changes. Specifically, in both cases, the interhemispheric connection within the frontal lobe weakened, whereas the intrahemispheric connection between the inferior and superior parietal lobes increased. Thus, this study suggests that changes, positive or negative, in cerebral activation with TSD may reflect alterations in communication among brain regions rather than simply increased or decreased activation within discrete brain regions.

Working Memory

Working memory, also called executive function, has long been a popular cognitive domain in TSD research. This is because (1) working memory tasks typically rely, at least in part, on brain function within the prefrontal cortex; (2) the prefrontal cortex has traditionally been thought of as the brain region most vulnerable to TSD; and (3) electroencephalography (EEG) studies have consistently demonstrated that changes in the frontal lobes following TSD contribute to cognitive performance deficits. The term 'working memory' refers to a large number of cognitive functions, however. It is likely that the specific aspect of working memory examined in any given study influences the measured effect of TSD on brain function. The types of working memory tasks used in functional neuroimaging studies of TSD include those requiring arithmetic, maintenance of information, and manipulation of information.

Arithmetic working memory relies on activation within the bilateral dorsolateral prefrontal cortex, anterior cingulate gyrus, and bilateral parietal lobes. Two studies examined arithmetic working memory following TSD. One of those is the PET study previously discussed in the section on attention tasks. This study showed widespread decreases in glucose metabolism during task performance starting after 24 h of TSD. However, activation in this study was confounded by the sustained attention demands of the task. The second study was an fMRI study examining serial subtraction after 35 h of TSD. Even though this study measured brain activation over 40 s intervals (thereby reducing the sustained attention demands), it confirmed the decreased responses within task-related brain regions following TSD. In both studies, subjects showed diminished performance with TSD.

Tasks requiring maintenance of information in working memory elicit bilateral activation in the prefrontal, precentral, and inferior parietal regions, as well as in the insula and thalamus. Tasks requiring the manipulation of information in working memory show more extensive activation in those same areas, reflecting the fact that manipulation requires more cognitive resources than simple maintenance. The studies examining brain function after TSD during tasks requiring maintenance and/or manipulation of information in working memory have reported both increased and decreased activation with TSD. Some of that inconsistency can be attributed to the task difficulty, similar to the verbal learning studies. More specifically, tasks requiring only the maintenance of information in working memory are more likely to show deceased activation, whereas those requiring the manipulation of information are more

likely to show increased activation with TSD. Furthermore, increasing the difficulty level within either a maintenance or manipulation task also results in greater activation during TSD. This is evident from a study that used a grammatical transformation task to manipulate task difficulty and found a stronger modulation of activation in task-related regions as difficulty increased following TSD, compared to the degree of modulation after normal sleep.

Overall, the effect of TSD on the cerebral response to working memory demands remains unclear. It appears that arithmetic demands (at least those not overlearned) produce decreased activation after TSD. The brain response to other working memory tasks probably depends on the specific nature of the task demands and the specific subregions with the prefrontal cortex underlying performance of those specific demands.

Visuospatial Functioning

Visuospatial functioning refers to those cognitive processes involved in visual perception of spatial relationships among objects. One example of visuospatial functioning involves the recognition of objects, such as circular line configurations, that have no verbal properties. In imaging experiments using recognition of nonverbal objects, visual association areas in the ventral visual stream are commonly activated along with the precuneus, an area associated with the dorsal visual stream. As with working memory tasks, studies using visuospatial tasks have shown both increases and decreases in cerebral activation following TSD. Most commonly, activation in the visual association cortex and precuneus decreases after TSD, and this correlates with decreased behavioral performance. Regions showing increased activation during recognition of nonverbal objects after TSD include the basal ganglia and insula. Interestingly, the one study that attempted to use a more real-world spatial navigation task (a piloting task) found exclusively increased activation after TSD, including in both visuospatial processing regions and dorsolateral executive controls regions.

Summary and Conclusion

Research using functional neuroimaging techniques to better understand the effects of sleep deprivation on brain function is still in the early stages. Most studies have taken a relatively descriptive approach; they administered various tasks to examine the differences in brain function after normal sleep and after TSD. A few studies have used more sophisticated manipulations such as parametrically manipulating task demands or functional connectivity analyses. Despite the fairly small number of studies in this area, some generalizations can be made about the results. There appear to be two distinct brain responses to cognitive challenges during TSD. The first is characterized by decreased activation relative to the well-rested state and diminished performance; this can be thought of as an impaired brain response. The second type of response is characterized by increased activation and relatively intact performance; this has been interpreted as a compensatory recruitment response.

Compensatory Recruitment Hypothesis

The compensatory recruitment hypothesis states that the brain has the ability to recruit additional cognitive resources during task performance following TSD that it typically does not recruit to perform the same task when the subject is well rested. These resources are then engaged in performing the task, and the more an individual engages these cognitive resources following TSD, the better the individual performs the task. It is this correlation between recruitment (measured as activation) and performance that makes the process compensatory. Clearly, however, such recruitment does not always occur because several studies reported decreased activation with TSD. Three broad classes of factors are thought to influence whether, and where in the brain, recruitment occurs following TSD: (1) cognitive task-related factors (What is the subject asking the brain do to?); (2) the length and type of sleep deprivation; and (3) individual difference variables (i.e., some individuals are able to recruit new resources and others are less able).

The studies reviewed here suggest that cognitive task factors certainly play a role in brain function following TSD. For example, sustained attention tasks seem to show impaired brain responses following TSD, whereas verbal learning tasks consistently show increased activation. In addition, each study that systematically manipulated the task difficulty showed that increasing the task difficulty facilitated the compensatory recruitment response. We might think that there is a functional limit to this difficulty-influenced recruitment; however, no study to date has effectively demonstrated such a limit. A similar argument could be made with respect to the length of sleep deprivation. After enough sleep loss, even the verbal learning tasks that consistently show recruitment with TSD should show decreased activation; however, only one published study took functional neuroimaging measurements at multiple time points during TSD. This was the PET study showing decreased glucose metabolism after 24, 48, and 72 h of TSD. Because

this study did not even show recruitment at the first time point, it cannot serve as a test of this aspect of the hypothesis. Likewise, no study has examined any length of PSD. The final factor thought to affect recruitment is individual differences, which has only been investigated in a few preliminary studies and, thus, also remains largely untested.

Future Directions

Clearly, much work needs to be done before we will fully understand the impact of sleep deprivation on brain function. All three factors proposed to affect cerebral activation following TSD need more research. Cognitive task-related factors are the most well-studied area. Nonetheless, we need to design cognitive tasks better so as to focus on very specific aspects of cognition rather than on general domains of cognition. For example, instead of examining working memory generically, investigators should employ event-related designs to examine the various components of working memory (e.g., attention, working memory buffer, manipulation, and encoding/retrieval). To the extent this has been done already, the results suggest that some components are more sensitive to TSD than others. Furthermore, future studies should use more real-world tasks when possible. For the second factor, models of PSD need investigation. This is a more realistic model of sleep loss in day-to-day life. Behavioral studies have suggested that a week of 4 or 6 h in bed per night produces cognitive deficits on some tasks similar to 24 or 48 h, respectively, of TSD. However, the time course of the deterioration and whether the brain can show compensatory recruitment during PSD remain unknown. Finally, behavioral data show that some individuals are relatively resilient to the effects of TSD on specific tasks, whereas others are more vulnerable. Understanding the source of those differences could lead to important advances in job selection, safety training, and effective countermeasures to sleep deprivation. Currently, we do not even know if demographic measures such as age, education, IQ, and gender play a role in the brain's response to TSD. Other variables to consider include the subject's mastery of a given task, genomic and gene expression differences, and prior experience with sleep deprivation. Work in this area is growing exponentially. The promise is that a better understanding of the consequences of sleep deprivation, and the factors that protect against the negative consequences, will help us better understand the function of sleep itself.

See also: Autonomic Dysregulation During REM Sleep; Endocrine Function During Sleep and Sleep Deprivation; fMRI: BOLD Contrast; Immune Function During Sleep and Sleep Deprivation; Metabolic Syndrome and Sleep; Positron Emission Tomography (PET); Sleep and Sleep States: PET Activation Patterns; Sleep Deprivation: Neurobehavioral Changes; Sleep-Dependent Memory Processing; Stimulant and Wake-Promoting Substances; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Chee MWL and Choo WC (2004) Functional imaging of working memory after 24 hr of total sleep deprivation. *Journal of Neuroscience* 24: 4560–4567.

Chee MWL, Chuah LYM, Venkatraman V, et al. (2006) Functional imaging of working memory following normal sleep and after 24 and 35 h of sleep deprivation: Correlations of fronto-parietal activation with performance. *NeuroImage* 31(1): 419–428.

Choo WC, Lee WW, Venkatraman V, Sheu FS, and Chee MWL (2005) Dissociation of cortical regions modulated by both working memory load and sleep deprivation and by sleep deprivation alone. *NeuroImage* 25: 579–587.

Drummond SPA, Bischoff-Grethe A, Dinges DF, et al. (2005) The neural basis of the psychomotor vigilance task. *Sleep* 28: 1059–1068.

Drummond SPA, Brown GG, Gillin JC, et al. (2000) Altered brain response to verbal learning following sleep deprivation. *Nature* 403: 655–657.

Drummond SPA, Brown GG, Stricker JL, et al. (1999) Sleep deprivation-induced reduction in cortical functional response to serial subtraction. *NeuroReport* 10: 3745–3748.

Drummond SPA, Meloy MJ, Yanagi MA, Orff HJ, and Brown GG (2005) Compensatory recruitment after sleep deprivation and the relationship with performance. *Psychiatry Research: Neuroimaging* 140: 211–223.

Drummond SPA, Smith MT, Chengazi V, and Perlis ML (2004) Functional imaging of the sleeping brain: Review of findings and implications for the study of insomnia. *Sleep Medicine Reviews* 8: 227–242.

Mu Q, Nahas Z, Johnson KA, Yamanaka K, et al. (2005) Decreased cortical response to verbal working memory following sleep deprivation. *Sleep* 28: 55–67.

Portas CM, Rees G, Howseman AM, et al. (1998) A specific role for the thalamus in mediating the interaction of attention and arousal in humans. *Journal of Neuroscience* 18: 8979–8989.

Stricker JL, Brown GG, Wetherell LA, and Drummond SPA (2006) The impact of sleep deprivation and task difficulty on networks of fMRI brain response. *Journal of the International Neuropsychological Society* 12: 591–597.

Thomas ML, Sing HC, Belenky G, et al. (2000) Neural basis of alertness and cognitive performance impairments during sleepiness. I: Effects of 24 h of sleep deprivation on waking human regional brain activity. *Journal of Sleep Research* 9: 335–352.

Wu JC, Gillin JC, Buchsbaum MS, et al. (1991) The effect of sleep deprivation on cerebral glucose metabolic rate in normal humans assessed with positron emission tomography. *Sleep* 14: 155–162.

Sleep Deprivation: Neurobehavioral Changes

J D Minkel, S Banks, and D F Dinges, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sleep restriction is common in modern society, not only as a result of medical and psychiatric disorders but also because of sociocultural influences including occupational and domestic responsibilities. Artificial lighting and around-the-clock services have facilitated lifestyles that do not allow for adequate levels of sleep. The resulting 'sleep debt' is associated with significant daytime performance deficits that carry considerable social, financial, and personal cost. The most well-documented changes in behavior resulting from sleep deprivation include microsleeps intruding into wakefulness and resulting in lapses in cognitive functions that are indicative of state instability. Sleep deprivation experiments have also found deleterious effects on mood, cognitive performance, and motor function due to an increasing sleep propensity and destabilization of the wake state. Vigilant and executive attention, working memory, and divergent higher cognitive functions are particularly vulnerable to sleep loss. These effects have been demonstrated in studies of total sleep deprivation (i.e., those that allow no sleep opportunity) and in studies of chronic partial sleep deprivation.

History of Sleep Research

Although observations on the effects of sleep loss can be found throughout recorded human history, the first systematic investigation of the effects of sleep loss in humans was conducted more than a century ago at the University of Iowa. Observations of three adults across 90 h of sustained wakefulness showed a general decrease in sensory acuity, slowing of reaction times and motor speed, and problems with memorization relative to performance after recovery sleep. This nineteenth-century study identified microsleeps, hypnagogic reverie, lapsing, circadian variation, and recovery sleep depth, all of which have withstood the test of time as behavioral consequences of sleep deprivation in humans.

The first systematic sleep studies were conducted by Dr. Nathaniel Kleitman beginning in the 1920s. He noted that subjects often seemed to increase their motivation during sleep deprivation to perform as well as they did prior to the sleep loss and that on some tasks, performance appeared normal at the beginning of the test but deteriorated as the test duration increased. These results demonstrated that sleep deprivation did not produce a 'lesion' in performance, which was the conventional wisdom at the time, but reduced neurocognitive capacity and performance.

The next major insight into the consequences of sleep deprivation occurred in the 1930s when Neil Warren and Brant Clark found that performance during sleep deprivation became more variable, despite little change in measures of central tendency of performance. This observation – namely that variability in behavioral output characterized the neurobehavioral performance of sleep-deprived healthy adults – offered a way of reconciling conflicting results among sleep deprivation studies without invoking motivation and duration of measurement. Modern experiments have confirmed the original observation that sleep deprivation increases performance variability rather than simply or solely decreasing average performance.

The physiological basis for the effects of sleep deprivation was first investigated in the late 1940s by Bo Bjerner, who demonstrated that sleep deprivation-induced lapses in performance (which contributed to greater variability) were associated with changes in the human scalp-recorded electroencephalogram (EEG). Specifically, it was observed that alpha-wave (8–12 Hz) depression and high-amplitude slow-wave activity accompanied delayed actions (lapses) on a performance task during deprivation. This EEG profile had been documented 10 years earlier to co-occur with subjective drowsiness.

Thus, by 1950 some of the most fundamental features of performance changes during sleep deprivation had been identified. However, the disparate findings had not yet been integrated into a theoretical model of how sleeplessness affected neurobehavioral functions. Moreover, the behavioral effects of sleep deprivation had not yet been understood relative to the homeostatic drive for sleep. Only in the past 10 years has the full range of consequences from sleep deprivation begun to be understood and a theory that attempts to systematically integrate findings from the neurobiology of sleep deprivation with data on cognitive and neurobehavioral performance changes been developed.

Functional Consequences of Sleep Deprivation

Sleep Deprivation and Accident Risk

Accidents involving sleep-deprived operators have been estimated to have an annual economic impact of \$43–\$56 billion. Motor vehicle crashes due to drowsy driving contribute substantially to these estimates. Drowsiness markedly increases the risk of crashing, and motor vehicle accidents related to sleep loss are particularly common. Length of time awake, nocturnal circadian phase, reduced sleep duration, prolonged driving duration, and use of sedative–hypnotic medications have all been found to contribute to the occurrence of drowsy-driving and fatigue-related motor vehicle crashes. Studies of shift workers, truck drivers, medical residents, and airline pilots show an increased risk of crashes or near misses due to sleep deprivation. Table 1 shows the results of a nationwide study of the increased motor vehicle crash risks posed by sleep deprivation from the common practice of scheduling medical residents for extended work shifts of 30 h or more. Sleepiness-related motor vehicle crashes have a fatality rate and injury severity level similar to those of alcohol-related crashes.

Sleep deprivation poses a risk to safety in all modes of transportation and to performance in many other safety-sensitive occupations. For example, studies of interns working prolonged shifts in intensive care units demonstrated that they made 35% more serious medical errors during work schedules that allowed the current practice of $\geq$30-h shifts compared to work shifts limited to 16 h. Limiting the prolonged shifts also resulted in the interns sleeping 5.8 h more per week, and they had less than half the rate of attentional failures while working during on-call nights. Thus, sleep deprivation is a preventable cause of accidents in most people.

Sleep Deprivation and the Propensity to Fall Asleep

Under normal conditions, humans experience either stable sleep or wakefulness with relatively little time spent transitioning from one state to another. When sleep is resisted, however, the stability of the sleep–wake system begins to break down. Intrusions of sleep into wakefulness and goal-directed performance become evident through increases in a variety of neurobehavioral phenomena: performance lapses, microsleeps and sleep attacks (i.e., involuntary naps), voluntary naps, shortened sleep latency, slow eyelid closures, and slow rolling eye movements. Wake state instability from inadequate sleep duration or timing can occur while people engage in virtually any type of motivated behavior. In fact, sleep deprivation has been shown to produce psychomotor impairments equivalent to those induced by alcohol consumption at or above the legal limit.

As sleep deprivation increases, the biological initiation of sleep occurs more often and sooner, as evident in sleep propensity/latency tests, which use polysomnography to measure physiological transitions from wake to sleep. Sleep deprivation results in not only a reduction in the latency to sleep onset during the Multiple Sleep Latency Test but also a reduction in latencies from lighter stages of nonrapid eye movement (REM) sleep to deeper slow-wave sleep (SWS). Such rapid transitions from waking to physiologically

Table 1 Risk of motor vehicle crashes and near-miss incidents after extended shifts

Variable	*Extended work shifts (>24 h)*	*Nonextended work shifts (<24 h)*
Crashes		
No. reported	58	73
No. of commutes	54 121	180 289
Rate (per 1000 commutes)	1.07	0.40
Odds ratio (95% CI)	2.3 (1.6–3.3)	1.0
Near-miss incidents		
No. reported	1971	1156
No. of commutes	54 121	180 289
Rate (per 1000 commutes)	36.42	6.41
Odds ratio (95% CI)	5.9 (5.4–6.3)	1.0

A within-person case-crossover analysis was used to assess the risks of motor vehicle crashes and near-miss incidents among interns during commutes after extended shifts compared with nonextended shifts. A two-by-two table was constructed for each intern who reported either a crash or a near-miss incident, consisting of the number of crashes or near-miss incidents after an extended shift, the number of crashes or near-miss incidents after a nonextended shift, the number of extended shifts that did not precede a crash or a near-miss incident, and the number of nonextended shifts that did not precede a crash or a near-miss incident. CI, confidence interval. Reproduced from Barger LK, Cade BE, Ayas N, et al. (2005) Extended work shifts and the risk of motor vehicle crashes among interns. *New England Journal of Medicine* 352: 125–134.

deeper stages of sleep reveal the increased neurobiological drive for sleep when wakefulness has been sustained too long, sleep duration has been too brief, and/or sleep consolidation has been abnormal, disrupted, or fragmented.

Whereas the Multiple Sleep Latency Test involves instructing subjects to try to fall asleep while supine in a sleep-conducive environment, the Maintenance of Wakefulness Test involves instructing subjects to remain awake rather than fall asleep in a non-sleep-conducive environment. Like the Multiple Sleep Latency Test, the Maintenance of Wakefulness Test shows reduced sleep latency in response to sleep deprivation. Thus, whether attempting to fall asleep or resist sleep, the latency from waking to sleeping is reduced by sleep deprivation. Although there is some controversy regarding the interpretation of sleep latency measures relative to actual cognitive performance measures during sleep deprivation, sleep latency tends to reflect subjective reports of daytime sleep tendency and the duration of nighttime sleep in the general population.

Microsleeps, Lapses, and Errors

Sleep deprivation not only reduces latencies to enter into normal sleep but also produces microsleeps (i.e., brief episodes of sleep that occur during waking performance). These intrusions result in behavioral lapses or errors of omission (i.e., failure to respond in a timely manner during cognitive performance demands). Lapses contribute to increased moment-to-moment variability in cognitive performance in sleep-deprived adults. Sleep deprivation can also increase errors of commission (i.e., response when no stimulus is present or to the wrong stimulus). Although lapses generally exceed the number of errors of commission on tasks in which the sleep-deprived subject controls the rate of performance, the opposite can be the case when the pace of the cognitive task permits only a very limited time for a response (i.e., tasks done under time pressure). In the latter circumstances, both types of errors are common in sleepy subjects.

Wake State Instability

The consequences of sleep deprivation on human performance were previously understood by the 'lapse hypothesis', which asserted that performance was relatively unchanged except during brief periods of low arousal. Although partially correct, this hypothesis does not explain some of the most robust cognitive changes associated with sleep deprivation, including a general slowing of cognition and reaction time, increased variability in performance, and potentiated decline in performance as time on task increases. To accommodate these effects of sleep deprivation, the wake state instability theory was developed. Rather than suggesting that sleep deprivation simply produces brief periods of impairment in otherwise normal performance, this theory posits that variability in neurocognitive performance increases as homeostatic sleep-initiating mechanisms become progressively more dysregulated with sleep loss. Thus, the theory posits that the co-occurrence of sleep and waking neurobiologies destabilizes performance, making it more erratic, less predictable, and poorly executed toward goal attainment. Not only do behavioral lapses result from sleep deprivation but also even optimal responses can be slowed or made less effective by sleep loss. Higher-order cognitive functions are adversely affected via the impact of sleep loss on basic attention and memory functions, as well as through direct effects on the areas of the cerebrum that mediate executive functions.

Wake state instability is theorized to occur when the brain's capacity to maintain alertness is hindered by the activation of sleep initiation from cortical and subcortical 'sleep switch' mechanisms. In such circumstances, behavioral state instability occurs when sleep pressure is elevated in the presence of goal-directed waking behavior. This conceptualization integrates the neurocognitive effects of sleep deprivation with what is known neurobiologically about sleep–wake regulation. When state instability is high, neurocognitive performance of the sleep-deprived individual is unpredictable and a product of interactive, reciprocally inhibiting neurobiological systems mediating sleep initiation and wake maintenance.

Sleep Deprivation and Basic Cognitive and Psychomotor Performance

Attentional networks, one of the basic building blocks of all cognitive activity, appear to be especially vulnerable to disruption by sleep deprivation. Therefore, any cognitive performance task that relies on stable vigilance and sustained attention with timely responses (e.g., driving) is at risk for deficits induced by sleep loss. Lapses of attention increase in both frequency and duration with increasing sleep deprivation, resulting in a progressive shift toward attentional lability. Individuals who are motivated to remain awake while sleep deprived can temporarily mask their impairment by engaging in compensatory behaviors (e.g., standing, walking, and talking), but the masking effects are typically short-lived. Ultimately, the compensatory effort to resist sleep cannot prevent intrusions of sleep initiation into wakefulness, as evidenced by Kleitman's observations

of sleep-deprived subjects 'semi-dreaming' while engaged in verbal cognitive tasks and by first-person reports of sleep-deprived people falling asleep while ambulating in dangerous environments.

There are hundreds of published studies on the cognitive deficits induced by sleep deprivation in humans. Sleep deprivation induces a wide range of effects on cognitive functions, although cognitive tasks vary considerably in their sensitivity to sleep loss. In general, regardless of the task, cognitive performance deteriorates more rapidly in sleep-deprived subjects than in well-rested controls; this is the classic 'fatigue' effect that is exacerbated by sleep loss. However, even performance on very brief cognitive tasks that precisely measure speed of cognitive throughput, and other aspects of attention and memory, consistently has been found to be sensitive to sleep deprivation. In this sense, sleep deprivation adversely affects the cognitive building blocks of attention and memory, on which higher-order cognition depends.

Two often overlooked confounding factors that can obscure the effects of sleep loss on many cognitive tasks are intersubject variability (e.g., aptitude on a task) and intrasubject variability (e.g., rate of learning a task). For example, one person's poorest performance during sleep deprivation may be superior to the best performance of a non-sleep-deprived person. Similarly, someone may be cognitively diminished by sleep loss but continue to improve when measured over time due to continued learning. Moreover, a failure to understand that sleep deprivation increases variability within subjects and between subjects and/or a failure to measure behavior at a frequent rate can result in the effects of sleep deprivation being missed in cognitive measures because less sensitive metrics or inappropriate data analyses are used.

Sleep Deprivation and Higher Cognitive Functions

More complex tasks involving higher cognitive functions have often been reported to be insensitive to sleep deprivation. In general, convergent rule-based deduction and critical thinking (e.g., logic-based tasks) show little change in accuracy during sleep deprivation, although they will reflect cognitive slowing. When tasks are made more divergent, such as multitasking and flexible thinking, sleep deprivation appears to have adverse effects on problem-solving performance. Divergent skills involved in decision making that appear to be affected by sleep loss include assimilation of changing information, updating strategies based on new information, lateral thinking, innovation, risk assessment, maintaining interest in outcomes, insight, communication, and temporal memory skills.

Implicit to divergent thinking abilities is a reliance on executive functions subserved by the prefrontal cortex. Executive function can be defined as the ability to plan and coordinate a willful action in the face of alternatives, monitor and update action as necessary, and suppress distracting material by focusing attention on the task at hand. Many tasks thought to engage different aspects of executive function have been used in studies of sleep-deprived individuals. These tasks also rely on working memory and executive attention, and most show some adverse effects of sleep loss.

Sleep Deprivation and Working Memory

Working memory involves the ability to hold and manipulate information and can involve multiple sensorimotor modalities. Deficits in neurocognitive performance requiring working memory result in difficulty determining the scope of a problem due to changing or distracting information, impairment in remembering temporal order of information, difficulty maintaining focus on relevant cues, taking inappropriate risks, difficulty maintaining flexible thinking, demonstrating poor insight into performance deficits, perseverating on thoughts and actions, and problems making behavioral modifications based on new information. It can therefore be concluded that sleep deprivation adversely affects prefrontal cortex-related executive attention and working memory abilities. Although both neurobehavioral and neuroimaging experiments confirm that sleep deprivation adversely affects neurocognitive functions involving the prefrontal cortex, there is also increasing evidence that subcortical systems involved in sleep initiation – especially hypothalamic, thalamic, and brain stem nuclei – have an important role in altering prefrontal cortical activity during sleep deprivation.

Sleep Deprivation, Mood, and Emotion

In comparison to the effects of sleep deprivation on cognition and performance, relatively little is known about the effects of sleep deprivation on emotional processes, with the exception of subjective mood states. Experimental data have clearly shown that virtually all forms of sleep deprivation result in increased negative mood states, especially feelings of fatigue, loss of vigor, sleepiness, and confusion. These data are limited, however, in that they were obtained without systematic mood induction procedures or other experimental probes designed to influence mood. It is not known if there are systematic differences in how specific events influence mood in sleepy versus rested subjects.

Sleep deprivation is widely believed to be associated with impaired emotional functioning as well. Field studies of sleep deprivation in military and medical personnel have found that reduced sleep is associated with increased negative and decreased positive emotional responses to specific events, but the uncontrolled nature of these studies allows for multiple interpretations of the findings. Other than simple subjective self-report measures of affect, there are few laboratory-based studies on affective responses in sleep-deprived humans. A controlled experiment in rats, however, found that sleep deprivation was associated with increases in aggression and risk-taking behavior. Although data are somewhat limited, the available evidence suggests that emotional processes are sensitive to sleep loss.

Effects of Chronic Partial Sleep Deprivation

Although acute total sleep deprivation is commonly used in research, chronic partial sleep restriction is more frequently experienced in real-world settings. Medical conditions (e.g., pain), sleep disorders (e.g., apnea), work demands (e.g., night shift work), and social and domestic responsibilities generally reduce the amount and quality of sleep that people are able to achieve but rarely eliminate sleep time entirely. Despite the pervasiveness of partial sleep restriction relative to acute total sleep deprivation, only recently has there been substantial experimental focus on its effects.

Daytime Sleep Propensity during Chronic Sleep Restriction

Basic questions concerning how much sleep is required for normal functioning, whether or not there is a cumulative cost to chronically restricted sleep, and to what extent individuals differ in sleep need have only recently been addressed experimentally. Few experiments have studied the cognitive effects of chronic sleep restriction over more than 7 days in a controlled laboratory setting to ensure participants only obtained the sleep durations they were assigned, and that they used no stimulants (e.g., caffeine and nicotine) or other techniques to reduce sleepiness. The two most extensive investigations found adverse effects of chronic sleep restriction on measures of vigilant attention, reaction time, cognitive throughput, and working memory. In one study, truck drivers were randomized to seven nights of 3, 5, 7, or 9 h time in bed for sleep per night. Only the 9-h time in bed condition showed no cognitive impairments across days. In an even larger experiment in which healthy young adults had their sleep duration restricted to 4, 6, or 8 h time in bed for 14 nights, only the 8-h time in bed condition showed no cognitive impairments across days. Three discoveries in these two experiments were seminal. First, both studies found dose–response cumulative adverse effects of sleep restriction on cognitive functions. That is, deficits in neurocognitive performance accumulated across consecutive days of sleep restriction when time in bed for sleep was chronically reduced below 7 h per night, and the accumulation was more rapid as sleep time became shorter. Second, the cognitive deficits induced by chronic sleep restriction accumulated to levels of impairment equivalent to those observed after one or two nights of acute total sleep deprivation. Figure 1 illustrates these phenomena for psychomotor vigilance task lapses from the two experiments.

Attenuated Subjective Effects of Chronic Sleep Restriction

The third major finding from the two large-scale controlled dose–response experiments on chronic sleep restriction concerned effects on subjective estimates of sleepiness. Although it is commonly assumed that people know when they are tired from inadequate sleep and can therefore correctly anticipate their ability to perform safely, these and other laboratory studies do not support this assumption. Subjective reports of sleepiness during periods of chronic sleep restriction showed an immediate modest elevation after the first one or two nights of sleep restriction, but this response reached a plateau after a few days and did not demonstrate the cumulative increases that were evident in cognitive performance deficits. Thus, as chronic sleep restriction progressed, subjective perception of alertness became increasingly dissociated from the reality of neurocognitive ability. Subjects often reported that they had adjusted to the reduced sleep durations, but objective performance measures clearly indicated that they had not. This suggests that people frequently underestimate the cognitive impact of sleep restriction and overestimate their performance readiness when sleep deprived.

Polysomnography and Waking EEG and Electrooculography Effects of Chronic Sleep Restriction

Waking EEG changes have also been observed during chronic sleep restriction. These typically involve shifts toward slower EEG frequencies in the theta (4–8 Hz) and delta range (3.75–4.5 Hz). Chronic sleep restriction to 3 or 4 h time in bed has been reported to lead to a decrease in saccadic velocity and slow eye movements (via electrooculography).

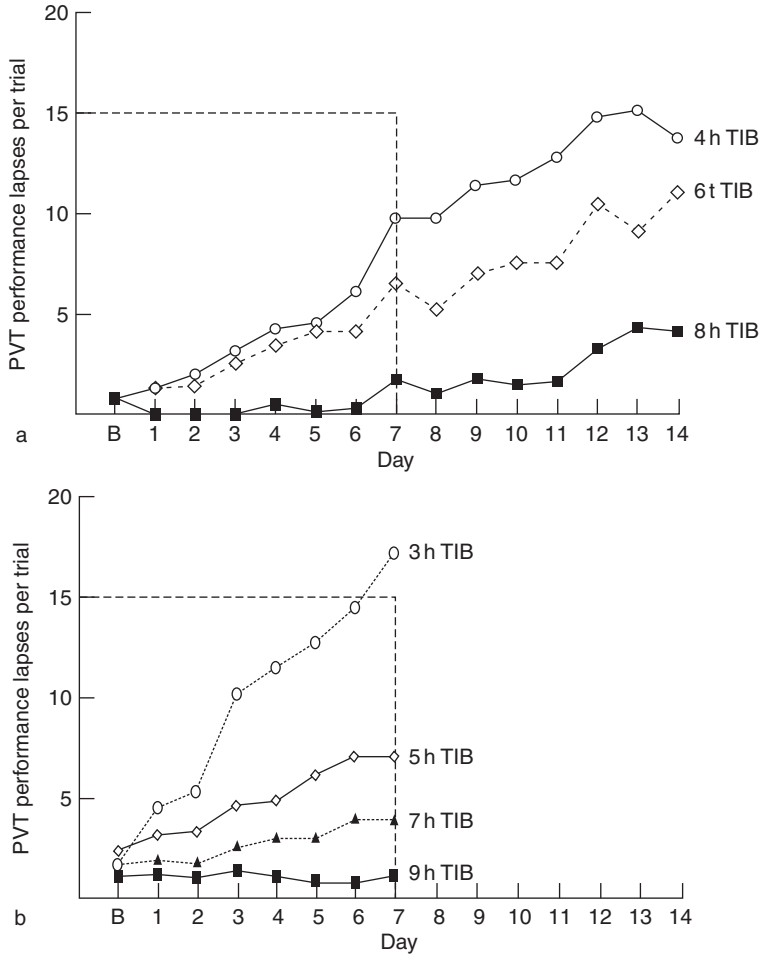

Figure 1 Lapses of attention during a psychomotor vigilance test from two experiments of chronic sleep restriction. (a) In this experiment, sleep was restricted for 14 consecutive nights in 36 healthy adults. Subjects were randomized to 4 h (n = 13), 6 h (n= 13), or 8 h (n = 9) sleep opportunities at night. Performance was assessed every 2 h from 7:30 a.m. to 11:30 p.m. The graph shows cumulative increases in lapses of attention during the psychomotor vigilance task (PVT) per test bout across days. The horizontal dotted line shows the level of lapsing found in a separate experiment when subjects had been awake continuously for 64–88 h. (b) In this experiment, sleep was restricted for seven consecutive nights in 66 healthy adults (mean age, 48 years). Subjects were randomized to 3 h (n = 13), 5 h (n = 13), 7 h (n = 13), or 9 h (n = 16) sleep opportunities at night. Performance was assessed 4 times each day from 9 a.m. to 9 p.m. As in a, the graphs show cumulative increases in PVT lapses per test bout across days within the 3 h and 5 h groups (p = 0.001). The horizontal dotted line shows the level of lapsing found in a separate experiment by Van Dongen et al. when subjects had been awake continuously for 64–88 h. (a) Adapted from Van Dongen HP, Maislin G, Mullington JM, and Dinges DF (2003) The cumulative cost of additional wakefulness: Dose–response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep* 26: 117–126. (b) Adapted from Belenky G, Wesensten NJ, Thorne DR, et al. (2003) Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose–response study. *Journal of Sleep Research* 12: 1–12. Reprinted from Opp MR (ed.) (2005) *SRS Basics of Sleep Guide*. Westchester, IL: Sleep Research Society.

Slow eyelid closures (which are associated with vigilance lapses) have also been shown to increase with sleep loss and have been found to be a sign of drowsiness while driving.

In addition to affecting Multiple Sleep Latency Test and Maintenance of Wakefulness Test sleep propensity measures, sleep restriction can affect sleep physiology. Following acute total sleep deprivation, rebound increases in non-REM SWS and non-REM EEG slow-wave activity (SWA) – the putative marker of sleep homeostasis – are commonly observed, but these EEG slow-wave rebounds are much less marked in chronic sleep restriction. When sleep duration is 4 h or more, there appears to be modest loss of SWA but marked reductions of stage 2 non-REM and REM sleep time. Consequently, it appears that deficits in SWS and non-REM SWA are not the basis for the cumulative cognitive performance deficits observed during chronic sleep restriction to 4–7 h per night, although they may be more closely related to subjective sleepiness. Other aspects of sleep must be critical to maintaining waking cognitive functions. Total sleep time, rather than non-REM

EEG SWA, therefore remained the major predictor of waking cognitive functions.

Sleep Debt Hypothesis versus Wake Extension Hypothesis

In the chronic sleep restriction experiment involving sleep limited to 4, 6, or 8 h time in bed for 14 consecutive nights, physiological sleep responses to sleep reduction did not mirror waking neurobehavioral responses, but cumulative wakefulness in excess of 15.84 h predicted performance lapses across all sleep deprivation conditions. This suggested that the concept of 'sleep debt' (or cumulative lost sleep) accounting for the cumulative waking neurobehavioral deficits was perhaps a misnomer since it was the additional wake time in excess of 15.84 h per day that appeared to accumulate and have a neurobiological 'cost' over days of sleep restriction. Thus, it is possible that the cumulative cognitive deficits during chronic sleep restriction were not caused by reduction of sleep time *per se* but, rather, by excessive wakefulness beyond a maximum period during which stable neurobehavioral functioning could be maintained. Viewed this way, cumulative wake extension (i.e., excess wakefulness), rather than cumulative loss of sleep (i.e., sleep debt), may be the primary cause of progressively reduced behavioral alertness both across days of chronic sleep restriction and across days of total sleep deprivation. This hypothesis awaits a neurobiological explanation. There has been little basic research published on possible neurobiological mechanisms for the effects of chronic sleep loss on waking cognitive performance, but it has been proposed that an increase in A_1 adenosine receptors in the basal forebrain in response to initial sleep loss may sensitize the brain to subsequent sleep loss.

Physiological Effects of Sleep Deprivation

Endocrine and Metabolic Effects of Chronic Sleep Restriction

Sleep restriction has also been found to affect metabolic responses. Restriction of sleep to 4 h time in bed per night for six nights, compared to extension of sleep to 12 h time in bed per night for six nights, has been associated with elevation in evening cortisol, increased sympathetic activation, decreased thyrotropin activity, and decreased glucose tolerance. Leptin levels have been observed to decrease following sleep restriction to 4 h time in bed. It is believed that these endocrine changes contribute to altered signaling of hunger and appetite, which may promote increased weight gain and obesity. Support for this hypothesis comes from epidemiological evidence of a U-shaped curvilinear association between sleep duration and body mass index (BMI). Those who slept less than 8 h had increased BMI proportional to decreased sleep. In addition, short sleep (5 vs. 8 h) was associated with low leptin (an adipocyte-derived hormone that suppresses appetite) and high ghrelin (predominantly a stomach-derived peptide that stimulates appetite), independent of BMI. The association of sleep duration, especially reduced sleep time, with obesity had also been observed in other epidemiological studies.

Cardiovascular Effects of Chronic Sleep Restriction

An increase in cardiovascular morbidity has been associated with sleep duration in a number of epidemiological studies. For example, there is epidemiological evidence of hypertension and increased risk of coronary events in women reporting less than 7 h of sleep or more than 9 h sleep per day. Similar findings have also been observed in studies examining cardiovascular health in shift workers, who typically experience chronic reductions in sleep duration in addition to circadian disruption. The mechanism underlying the link between chronic sleep restriction and increased cardiovascular risk is unknown. One potential mechanism, however, may be activation of inflammatory processes during sleep loss, which has been reported to occur in laboratory studies of healthy sleep-deprived adults.

Immune and Inflammatory Effects of Chronic Sleep Restriction

A single night of total or partial sleep deprivation can alter natural killer cell activity and circulating levels of leukocytes and cytokines (interleukin-6 and tumor necrosis factor-α) in healthy humans. Although there have been few controlled experiments, it appears that restriction of sleep to 4 h time in bed for six nights acutely reduced antibody titers to influenza vaccination. Attenuation in the febrile response to an endotoxin (*Escherichia coli*) challenge in subjects undergoing chronic sleep restriction to 4 h time in bed for 10 nights has also been observed. The basic mechanisms by which these acute effects of sleep deprivation on immune responses occur remain to be identified, and their clinical significance is unknown.

Differential Neurobehavioral Vulnerability to Sleep Loss

Sleep deprivation (both acute total and partial chronic) not only increases neurobehavioral variability within subjects (i.e., state instability) but also

reveals marked cognitive performance differences between subjects. As sleep pressure accumulates over time, intersubject differences also increase markedly. Although most people suffer neurobehavioral deficits when sleep deprived, there are individuals with extreme reactions at both ends of this spectrum. Some people experience very severe impairments even with modest sleep loss, whereas others show few, if any, cognitive deficits until sleep deprivation is very severe. Studies have shown that these responses to sleep deprivation are stable and reliable across subjects, suggesting that they are traitlike and likely genetic rather than the result of situational variability or 'noise'. Thus, some people are much more vulnerable to the neurocognitive effects of sleep deprivation than others. The biological basis of the differential vulnerability to neurocognitive deficits from sleep loss is not known.

Summary

For more than a century, experiments on the effects of total and chronic partial sleep deprivation in healthy human subjects have revealed that sleep cannot be deprived without immediate and often profound changes to neurobehavioral functions, especially those involving attention and memory. In addition to the brain being sensitive to sleep loss, experiments indicate that endocrine, metabolic, immunologic, and inflammatory processes are also affected by sleep deprivation. Despite evidence of the behavioral and physiological risks posed by sleep deprivation, time devoted to sleep has declined in industrialized societies, especially in the past 20 years. Instead of protecting sleep time, there is a tendency in modern cultures to spend resources on ways to safely reduce sleep time to accommodate 24-h operations. These efforts have met with modest success. There is as yet no substitute for sleep, and sleep duration remains the most reliable predictor of wake state stability and neurobehavioral functioning.

See also: Autonomic Dysregulation During REM Sleep; Circadian Rhythms in Sleepiness, Alertness, and Performance; Endocrine Function During Sleep and Sleep Deprivation; Immune Function During Sleep and Sleep Deprivation; Metabolic Syndrome and Sleep; Sleep Research and Sleep Medicine in Historical Perspective; Sleep Deprivation and Brain Function; Sleep-Dependent Memory Processing; Stimulant and Wake-Promoting Substances; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Banks S, Barnes M, Tarquinio N, et al. (2004) The maintenance of wakefulness test in normal healthy subjects. *Sleep* 27: 799–802.

Barger LK, Cade BE, Ayas N, et al. (2005) Extended work shifts and the risk of motor vehicle crashes among interns. *New England Journal of Medicine* 352: 125–134.

Belenky G, Wesensten NJ, Thorne DR, et al. (2003) Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose–response study. *Journal of Sleep Research* 12: 1–12.

Carskadon MA, Dement WC, Mitler MM, et al. (1986) Guidelines for the Multiple Sleep Latency Test (MSLT): A standard measure of sleepiness. *Sleep* 4: 519–524.

Chuah YML, Venkatraman V, Dinges DF, and Chee MWL (2006) The neural basis of inter-individual variability in inhibitory efficiency following sleep deprivation. *Journal of Neuroscience* 26: 7156–7162.

Dinges DF, Baynard M, and Rogers NL (2005) Chronic sleep restriction. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 67–78. Philadelphia: Saunders.

Dinges DF, Pack F, Williams K, et al. (1997) Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep* 20: 257–267.

Durmer JS and Dinges DF (2005) Neurocognitive consequences of sleep deprivation. *Seminars in Neurology* 25: 117–129.

Landrigan CP, Rothschild JM, Cronin JW, et al. (2004) Effect of reducing interns' work hours on serious medical errors in intensive care units. *New England Journal of Medicine* 351: 1838–1848.

Saper CB, Scammell TE, and Lu J (2005) Hypothalamic regulation of sleep and circadian rhythms. *Nature* 27: 1257–1263.

Stutts JC, Wilkins JW, Osberg JS, and Vaughn BV (2003) Driver risk factors for sleep-related crashes. *Accident Analysis & Prevention* 35: 321–331.

Taheri S, Lin L, Austin D, Young T, and Mignot E (2004) Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index. *PLoS Medicine* 1: 210–217.

Thomas M, Sing H, Belenky G, et al. (2000) Neural basis of alertness and cognitive performance impairments during sleepiness: I. Effects of 24 h of sleep deprivation on waking human regional brain activity. *Journal of Sleep Research* 9: 335–352.

Van Dongen HP, Baynard MD, Maislin G, and Dinges DF (2004) Systematic interindividual differences in neurobehavioral impairment from sleep loss: Evidence of trait-like differential vulnerability. *Sleep* 27: 423–433.

Van Dongen HP, Maislin G, Mullington JM, and Dinges DF (2003) The cumulative cost of additional wakefulness: Dose–response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep* 26: 117–126.

Sleep in Adolescents

L Tarokh, Brown University, Providence, RI, USA
M A Carskadon, Brown University, Providence, RI, USA; and E. P. Bradley Hospital, Providence, RI, USA

© 2009 Elsevier Ltd. All rights reserved.

Sleep behavior and sleep physiology manifest dramatic changes across adolescent development. Certain sleep pattern changes arise from psychological and societal sources, such as a growing sense of independence, increasing access to electronic apparatus in the bedroom, greater academic obligations, and more social opportunities. Some have noted that emergent changes in affect regulation and social affiliation, including 'first love,' can have striking influences on sleep patterns.

Certain hormonal changes that herald and accompany puberty are sleep related, such as the pulsatile sleep-dependent release of luteinizing hormone (LH), which leads to subsequent sleep-related release of sex steroids. Whether these hormonal changes have a direct effect on sleep is unknown; that sleep changes substantially in the pubertal transition is well established. As discussed later, maturation of the central nervous system (CNS) affects the phenomenology of sleep as well as the underlying intrinsic regulatory processes.

In addition to CNS changes, adolescent maturation is associated with new physical attributes and reproductive development. A common method to gauge pubertal development in humans is Tanner staging, which relies on identifying benchmarks in secondary sexual characteristics. Thus, in boys genital development and the growth and distribution of pubic hair are evaluated, and in girls breast development and pubic hair growth and distribution are evaluated. Tanner stage 1 is assigned if no signs of development are identified, Tanner stages 2–4 designate intermediate pubertal stages based on achieving developmental criteria, and Tanner stage 5 requires achievement of all the benchmarks identifying full maturation of secondary sexual characteristics. Note that central signs of puberty (e.g., nocturnal LH release and likely CNS changes) occur earlier than the external physical attributes of puberty.

Adolescent Development and the Brain

The behavioral, emotional, and cognitive changes of adolescence are accompanied by large-scale remodeling of the brain that provides a biological basis for many features of adolescent development. *In vivo* longitudinal studies of the developing human brain using magnetic resonance imaging (MRI) show that gray matter volume increases during preadolescence and declines postadolescence. This inverted U-shape change in gray matter volume corresponds on a cellular level with a marked proliferation of axons/synapses in early puberty and synaptic pruning in later adolescence. Positron emission tomography studies have shown a similar inverted U-shape increase and then decrease in waking cerebral glucose metabolic rates. These cross-sectional studies demonstrate a peak in absolute cerebral glucose metabolism at the age of 9 or 10 years followed by a steady decline until the ages of 16–18 years, at which point levels reach adult values. Not surprisingly, there is a strong correspondence between the developmental curve for blood glucose metabolism, gray matter volume, and synaptogenesis. The combined results from these studies provide strong evidence that adolescence is a time of major neural reorganization.

Adolescent Development and the Waking Electroencephalogram

The changes in gray matter volume, glucose metabolism, and synaptogenesis parallel the decline of electroencephalographic (EEG) power across adolescent development. The EEG signal is thought to arise from synchronous synaptic activity at the cortex; therefore, a reduction in the number of cortical synapses would logically result in dampened EEG power. Although gray matter volume and EEG power decline across adolescence, white matter volume – an index of myelination – increases in a linear manner. Adolescent development is also accompanied by greater functional connectivity between distant cortical regions (i.e., connections between frontal and occipital lobes), as inferred from EEG coherence. Cross-sectional studies of waking EEG coherence show a pattern of increasing inter- and intrahemispheric coherence before and during adolescence.

Along with waking EEG power and coherence changes, the frequency of the dominant posterior rhythm (alpha rhythm) increases as the brain matures. This rhythm is associated with a relaxed waking state and increases in amplitude with eye closure. The dominant posterior rhythm is seen during eye closure in infants as young as a few months old and increases in frequency from approximately 4 Hz in infancy to approximately 8 Hz in mid-childhood, reaching adult frequencies (8–12 Hz) by adolescence.

Adolescent Development and the Sleeping EEG

The EEG is a central measure of sleep, and indeed, EEG patterns are a defining characteristic of sleep in mammals. Typically measured along with electrooculogram and electromyogram, the EEG in humans contributes to defining sleep states (nonrapid eye movement (NREM) sleep and rapid eye movement (REM) sleep) and four sleep stages within NREM sleep. As outlined previously, a significant amount is known about the maturation of the waking EEG. Less is known about adolescent development of the sleeping EEG, although a number of characteristic changes have been identified. As with waking, for example, the age-related decrease in EEG power is strikingly apparent during sleep, particularly in the lower frequencies.

Figure 1 illustrates a number of changes of nocturnal sleep that accompany adolescent maturation by displaying the sleep hypnogram of a representative 10-year-old Tanner stage 1 child and a 15-year-old Tanner stage 5 adolescent. When sleep quantity is held constant at approximately 9 or 10 h, adolescent development is accompanied by a profound decline of approximately 40–50% in the quantity of NREM sleep stages 3 and 4, commonly referred to as slow-wave sleep (SWS). The dampening of EEG amplitude consequent to cortical synaptic pruning may account for the developmental reduction of SWS. At the same time, stage 2 NREM sleep, the definition of which involves EEG features unrelated to amplitude (i.e., sleep spindles and K complexes), increases. Another impressive developmental feature of sleep is the advance of REM sleep to an earlier point in sleep, from a latency of approximately 150–180 min to a latency of approximately 100 min in most studies.

Figure 2 illustrates the marked reduction of EEG amplitude for NREM sleep stages 2 and 4 and REM sleep across adolescence. The sleep EEG amplitudes are notably lower in the more mature adolescent. This EEG amplitude decline appears to be non-state-dependent because it is seen in wakefulness and NREM and REM sleep, although definitive longitudinal assessments have not been carried out. Figure 3 emphasizes this point by displaying the absolute EEG power spectra across the entire night for the pre- and postpubertal youngsters. Warmer colors (designating greater power) in the Tanner stage 1 child versus the Tanner stage 5 child are evident across frequencies and states. The differences are especially clear for low frequencies (1–4 Hz), particularly early in the night.

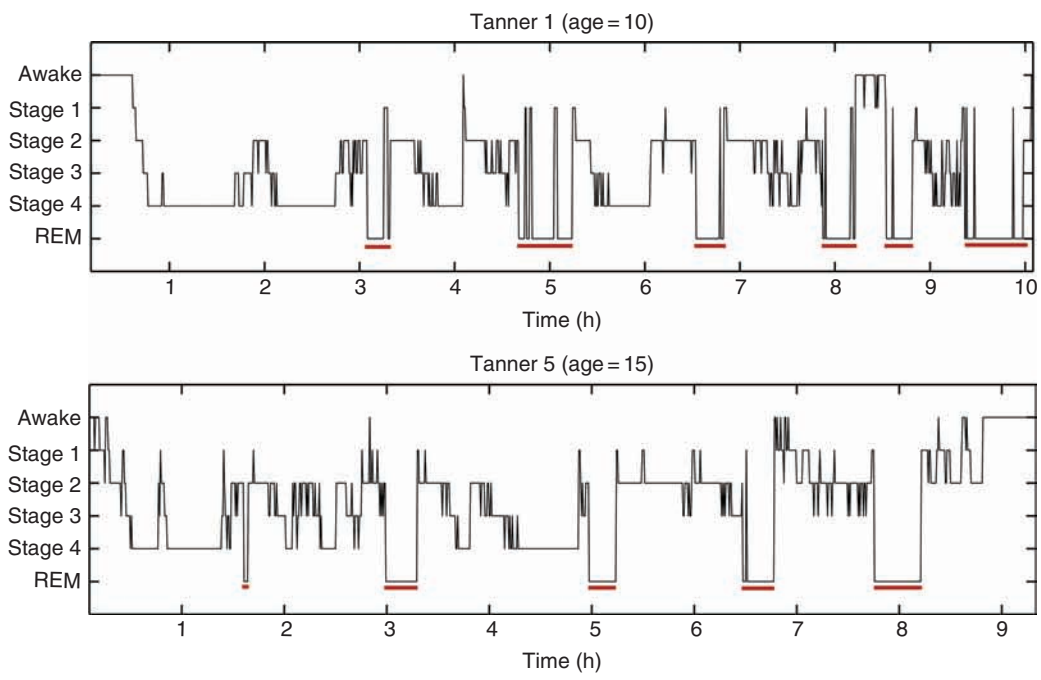

Figure 1 Sleep hypnograms of two boys, one at pubertal Tanner stage 1 (top) and the other at Tanner stage 5 (bottom). The *y*-axis designates stage/state, and the *x*-axis indicates elapsed time from lights out. REM episodes are underlined with a red bar. Stages were determined by visual scoring of 30 s epochs according to standard staging criteria. Significantly greater stage 4 sleep is typical of prepubertal adolescents compared to postpubertal adolescents. The decreased stage 4 sleep in more mature adolescents is accompanied by increased stage 2 sleep. Note also that REM sleep episodes increase in duration over the course of the night. Of note is the delayed onset of the first REM episode in the Tanner 1 preadolescent. Indeed, the first REM episode is often 'skipped' in Tanner 1 preadolescents.

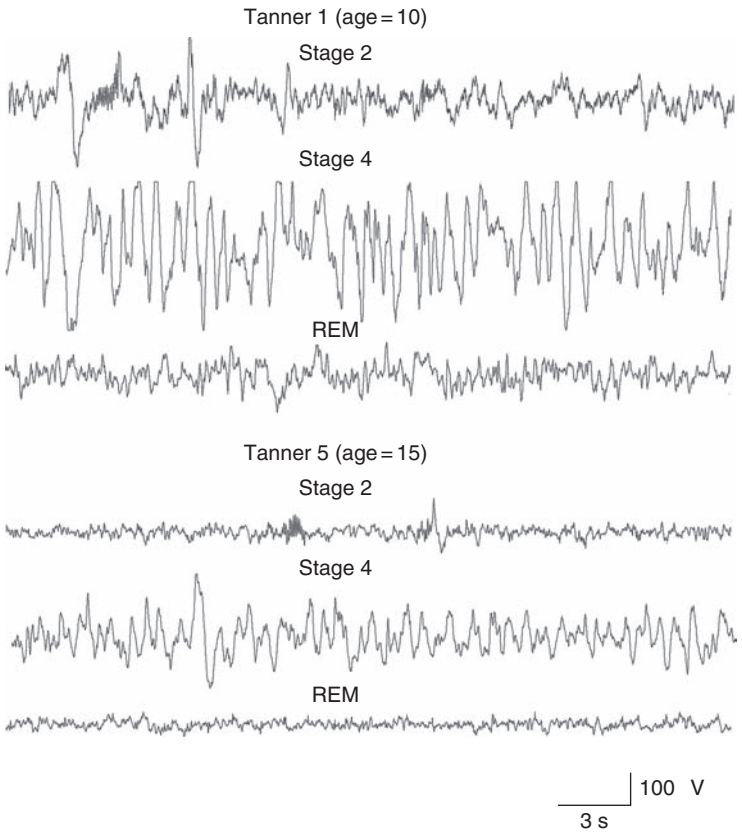

Figure 2 Comparison of Tanner stage 1 (top) and Tanner stage 5 (bottom) EEG tracings of NREM sleep stage 2, stage 4, and REM sleep from the same children and nights as in Figure 1. The calibrations designate the voltage and timescale of all tracings. The diminished EEG amplitude in the more mature adolescent is obvious from these tracings and is most evident in stage 4 sleep; however, this developmental change is not stage specific and occurs in all sleep stages and in the waking EEG. The reduction of EEG amplitude during early adolescence is believed to result from cortical synaptic pruning that occurs during this developmental stage.

This EEG balance of power toward the less mature brain is also clear for the sigma frequencies (approximately 14 Hz). Note that the sigma 'stripe' in these plots is evident only for NREM sleep and disappears during REM episodes.

Sleep Regulation and Adolescent Development

Homeostatic Process (Process S)

In the two-process model of sleep regulation, independent homeostatic (process S) and circadian (process C) processes operate interactively to regulate the amount and timing of sleep. Process S, which can be characterized colloquially as sleep pressure, has been modeled using the NREM sleep EEG slow-wave activity (0.5–4.5 Hz; SWA). As waking is extended, NREM EEG SWA power increases during subsequent sleep and is modeled using a saturating exponential function. With sleep, NREM EEG SWA spectral power manifests an exponential decay across the sleep episode.

The substantial decline in absolute SWA EEG power across adolescence raises the additional question of whether the homeostatic regulatory process is altered and perhaps secondarily accounts for certain expressed sleep behaviors in adolescents. The Brown University group examined this issue in several ways using cross-sectional data sets. First, they examined the dynamics of the decay of SWA across the sleep episode of Tanner stage 1 and Tanner stage 5 adolescents. As Figure 4 illustrates, the exponential fits were overlapping functions, and the mean decay time constant for Tanner stage 1 subjects was 143.2 min (range, 87.0–195.5) and that for Tanner stage 5 subjects was 128.2 min (range, 80.4–176.9). This finding indicates that the nocturnal recovery process is similar for adolescents across this developmental phase, with an implication that neither sleep intensity nor sleep need change with adolescence.

On the other hand, when the accumulating function of process S is examined in the context of extended wakefulness, the data indicate the presence of a maturational change. One approach examined sleep pressure by assessing the speed of falling asleep

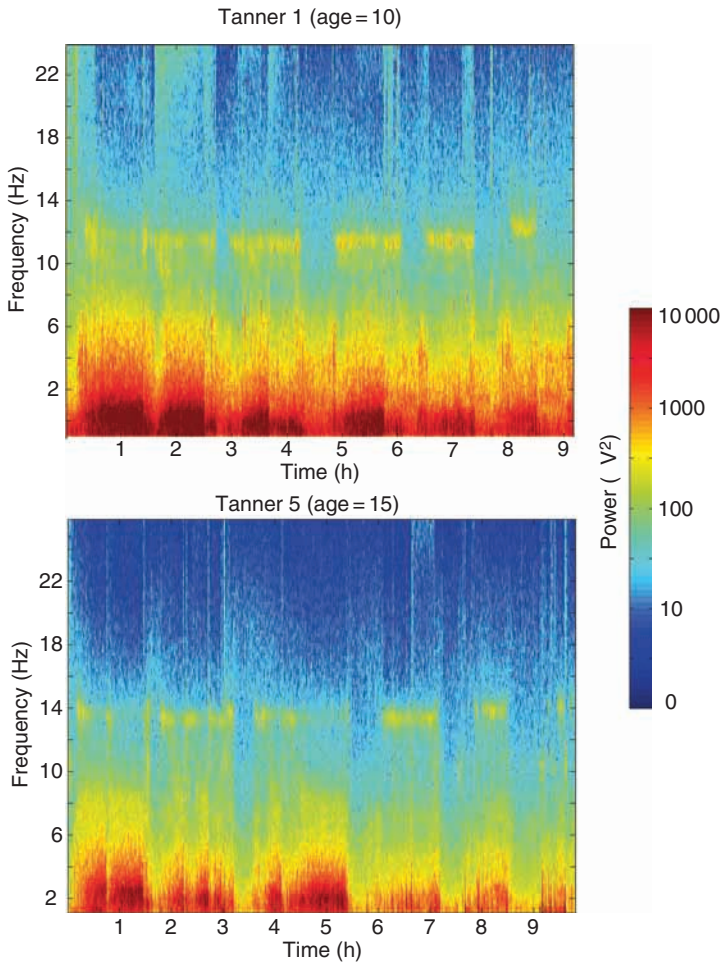

Figure 3 Time–frequency plot of the same nights and children depicted in the hypnograms displayed in **Figure 1**. The figure is derived from fast Fourier transform analysis of the C3/A2 EEG signal. The plots depict how the frequency content of the EEG signal changes over the course of the night. The *y*-axis is the frequency and the *x*-axis shows hours since lights out. EEG power is indicated by the color, in which warmer colors (reds and yellows) indicate higher power and cooler colors (aqua and blue) indicate lower power. During NREM sleep, the power of the signal at frequencies below 4 Hz is maximal, particularly early in the night, ebbing in strength across the sleep episode. Note also that the power in these slow frequencies is markedly greater for the younger boy. NREM episodes are also marked by high power in the spindle frequency range (12–16 Hz). In contrast, during REM episodes, the power in the lower frequency and spindle bins is diminished. The characteristic cycle-by-cycle nature of the sleep EEG is apparent in these plots. Overall EEG power differences between the Tanner 1 and Tanner 5 participants are readily apparent at all frequency bins.

in adolescents at Tanner stage 1 and Tanner stage 5. This study found that after approximately 14.5, 16.5, and 18.5 h awake, the less mature youngsters fell asleep significantly faster than the more mature group. A second approach modeled process S using SWA in a sleep deprivation study and found fitted functions for the exponential saturating function with a time constant of 15.4 ± 2.5 h for the mature Tanner stage 5 adolescents compared to 8.9 ± 1.2 h for a group of pre-/early pubertal (Tanner stages 1/2) children. These two findings indicate that the more mature adolescents, although they exhibit the same sleep recovery pattern, may find it easier to remain awake longer than the less mature adolescents. This developmental alteration may contribute to the later timing of sleep that emerges across adolescence.

Circadian Timing (Process C)

Maturational changes in the circadian timing system (CTS) also appear to contribute to the behavioral changes in adolescent sleep patterns. The master 'clock' in mammals has been localized to a small paired nucleus in the hypothalamus, the suprachiasmatic nuclei (SCN). A direct pathway from the retina, particularly the light-sensitive melanopsin-containing retinal ganglion cells, carries light input directly to the SCN, where the light signals affect the molecular machinery of the CTS within SCN neurons.

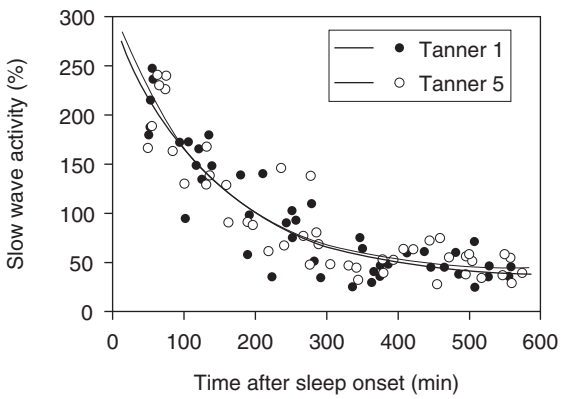

Figure 4 Dynamics of slow-wave activity (SWA; 0.6–4.6 Hz) across the sleep episode in eight prepubertal (solid circles) and eight mature (open circles) adolescents. SWA power for each NREM sleep cycle is shown as a percentage of the individual child's average overnight NREM sleep, and each symbol represents this value plotted at the midpoint of the time relative to the onset of sleep. The lines show exponential functions fitted to the Tanner 1 (thick line) and Tanner 5 (thin line) data. (Both fits achieved r^2 values of 0.83.) These data indicate that the dissipation of SWA, and inferentially of process S, does not change with pubertal development, although the power of the EEG SWA signal is markedly changed as shown in **Figure 3**. Reproduced from Jenni OG and Carskadon MA (2004) Spectral analysis of the sleep electroencephalogram during adolescence. *Sleep* 27: 774–783.

One of the first findings implicating the CTS in the adolescent delay of sleep timing was an association between self-reports of phase preference (e.g., morningness vs. eveningness) and puberty in sixth-grade girls: Girls who rated themselves more mature also rated themselves with greater evening-type phase preference. Subsequent studies examined these phenomena using better measures, such as Tanner staging and circadian phase measured using the dim light melatonin onset (DLMO) phase. DLMO is a reliable and noninvasive assessment of circadian phase acquired from serial plasma or saliva samples collected under dim light (less than approximately 40 lux) conditions. Because melatonin levels are low during the biological daytime and rise during the biological night, the circadian phase can be measured by assessing the onset phase of melatonin release. In adolescents, saliva is collected approximately every 30 min in dimly lit conditions during the participants' usual evening hours.

A major feature of the CTS is its sensitivity to environmental light. Effects of light on the CTS are phase dependent; to wit, light impinging on the clock in the early nighttime produces phase delays (later timing) and light occurring in the early morning produces phase advances (earlier timing). Because of this phase-dependent light sensitivity, mature adolescents living in their normal circumstances will inevitably express later circadian phase positions because their delayed sleep pattern exposes them to light at a later time. Therefore, to determine whether the internal process undergoes developmental change, light exposure needs to be controlled, for example, by controlling sleep (dark) schedules. In such controlled circumstances, the DLMO phase is positively correlated with Tanner stage, whereby more mature adolescents have a later DLMO phase. This circadian phase delay across adolescence may affect the ability of older adolescents to fall asleep easily in the evening since the CTS favors wakefulness in the early circadian night, sometimes called the forbidden zone for sleep. In addition to humans, a pubertal delay of sleep behavior has been reported in rhesus monkeys and in the octodon degus, a small diurnal mammal with a relatively protracted developmental course from birth to reproductive maturity.

Attempts have been made to identify the underlying mechanism for this adolescent phase delay. One hypothesis holds that the intrinsic period of the circadian timing system may slow from prepubertal to postpubertal adolescence, thus delaying circadian phase. Studies that permit assessment of internal circadian period are both difficult and time-consuming and have not established the validity of this hypothesis. One study, however, has shown that the period is longer on average in adolescents than in adults. Another theory for the developmental change in circadian phase posits that the phase-dependent sensitivity to light's phase shifting effects may favor phase delays either through strengthening the delay response to evening light or through weakening the phase advance to morning light. Attempts to demonstrate this phenomenon have not been successful to date, although none has used the most circadian-sensitive shortwave (blue) light sources to examine the possibility of a maturational change in light sensitivity.

Thus, the circadian timing system does manifest a phase delay associated with pubertal/adolescent development; however, the mechanism for the delay is unknown. Unlike the sleep system in which the adolescent decline in EEG amplitude can be attributed directly to cortical synaptic pruning, no molecular, cellular, or systems basis has been identified to explain the maturation of the CTS across adolescence.

Interaction of Process S and Process C

As noted previously, the circadian timing system favors waking in the biological evening. Indeed, the strength of the arousing signal from the circadian timing system increases from the usual time of morning arousal until about the time of usual bedtime and

decreases across the typical nighttime window. On the other hand, the homeostatic sleep–wake process builds as waking is extended, only to dissipate during sleep. Either process on its own is untenable for supporting our usual sleeping and waking patterns; however, the interaction provides an arousing circadian signal later in the day to support alertness when the homeostatic pressure would otherwise favor sleep. Conversely, the circadian arousing signal wanes during the night. As morning comes, with circadian arousal at its nadir, wakefulness is underwritten by the restorative function of sleep.

In the case of adolescent development, this interaction plays out in several important ways. When daytime alertness is evaluated in children keeping an optimizing sleep schedule (10 h per night from 10:00 p.m. to 8:00 a.m.), the interaction of processes S and C manifests as maximal alertness from rising until bedtime. With the onset of puberty, however, the pattern shifts to one that augments midday sleepiness. In the experimental paradigm that showed this phenomenon, young people were followed annually across several years with summertime assessments on the optimized sleep schedule and sleepiness measured with the Multiple Sleep Latency Test (MSLT), a standardized measure of the speed of falling asleep (sleep propensity) at 2 h intervals across the day. In prepuberty, the youngsters maintained alertness near the ceiling of this measure throughout the day; at midpuberty, the pattern changed to exhibit faster sleep onsets on tests given in the middle of the afternoon. Subsequent evaluation of these data indicated that the phase delay of pubertal development resulted in a reorganization of the phase relationship between sleep and the circadian timing system. With this reorganization, the circadian buffering of accumulating sleep pressure across the waking day did not occur until a later phase, and thus midday sleepiness emerged.

In normal circumstances, however, adolescents do not live with optimal sleep. As outlined below, sleep occurs later and is shorter in older adolescents. This pattern results in chronic insufficient sleep on school nights that is not recuperated fully by nighttime sleep. In these circumstances, morning alertness specifically is impaired since adolescents wake with elevated homeostatic pressure and also at a circadian phase that does not support alertness.

One study examined the consequences of this mismatch in students attending the 10th grade of high school in the United States. These students were required to attend school beginning at 7:20 a.m. When followed on their normal sleep schedule for 2 weeks, the students obtained approximately 7.5 h of sleep each night, going to bed on average at approximately 10:30 p.m. and waking at approximately 6:00 a.m. When their alertness was measured using the MSLT, these adolescents were sleepiest at 8:30 a.m. (average speed of falling asleep was 5 min), and alertness improved in the late morning and early afternoon, presumably due to the alerting effects of the circadian timing system. Thus, the speed of falling asleep at 2:30 p.m. was 11 min. Of greater note in this study, however, was the appearance of an abnormal pattern of sleep in nearly 50% of the young people on the MSLTs: They had at least one test (usually at 8:30 a.m.) in which they fell directly into REM sleep rather than NREM sleep (i.e., sleep-onset REM episodes). Furthermore, the speed of falling asleep in this group was even faster at 8:30 a.m. – 3.4 min on average. A distinguishing feature of this group was that the phase of DLMO was later, and the investigators concluded that they woke at a much earlier circadian phase than the other students, at a time favoring sleeping over waking and REM sleep over NREM sleep. The MSLT results in these young people looked more like those of patients with a serious sleep disorder – narcolepsy – than healthy teens.

Sleep Habits and Adolescent Development

As mentioned previously, sleep patterns undergo major alterations across adolescent development – changes that have been confirmed in many surveys of young people on every continent (except Antarctica). In each case, the timing of sleep is later overall in older adolescents. In societal circumstances requiring early morning awakening – for example, with early starting time for school as found in many communities in the United States – sleep quantity also is reduced, at least on school days. Indeed, the difference between school day and weekend schedules becomes greater during adolescence. At age 9 or 10 years, for example, children may delay weekend bedtime slightly but tend not to delay or extend sleep as much on weekend mornings. By age 14 or 15 years, on the other hand, bedtime is much later on school nights, delayed by 1–3 h on weekends, and weekend sleep can be several hours longer than on school days. These weekday–weekend differences express both the pressure for sleep to delay during adolescence and the compensatory need for sleep on weekends due to insufficient sleep on school nights in the older adolescents.

The 2006 Sleep in America poll of the National Sleep Foundation surveyed approximately 1600 adolescents and caregivers. Data from this poll provide a contemporary perspective of adolescent sleep. Although this article has emphasized the biological bases of sleep pattern alterations across adolescence,

the influence of environmental factors is undeniable. Furthermore, the interactions among sleep regulatory changes and social/environmental influences and constraints can lead to troubling outcomes in susceptible youngsters. As the Sleep Foundation poll data show, school night sleep time reported by sixth-grade students in the United States is approximately 8.4 h and it is only 6.9 h in 12th-grade students, which by general consensus is significantly below the sleep need of teens.

The starting time for school plays an important role in setting the stage for short sleep in many adolescents. Thus, although teens' bedtimes delay significantly across adolescence, rising time on school days does not delay and indeed may actually become earlier (Figure 5). Later bedtimes fit well with the intrinsic maturational changes; earlier rising times, however, do not arise from a natural tendency on the part of adolescents but in direct response to waking in time for school. The result is truncated sleep time. In cases in which school begins later in the day, as shown by a number of ongoing studies of students in Brazil, Croatia, and the United States, teens sleep later and longer. One US study that examined sleep patterns of high school students before and after imposition of a later school schedule found that bedtimes did not change significantly; however, rising times were later and, consequently, youngsters slept more.

A newly emerging consensus also credits access to contemporary technologies as contributing to adolescent sleep patterns. For example, Belgian adolescents who report using electronic media to help them fall asleep in fact sleep less than other teens and report themselves as more tired. Furthermore, the National Sleep Foundation's 2006 Sleep in America Poll found

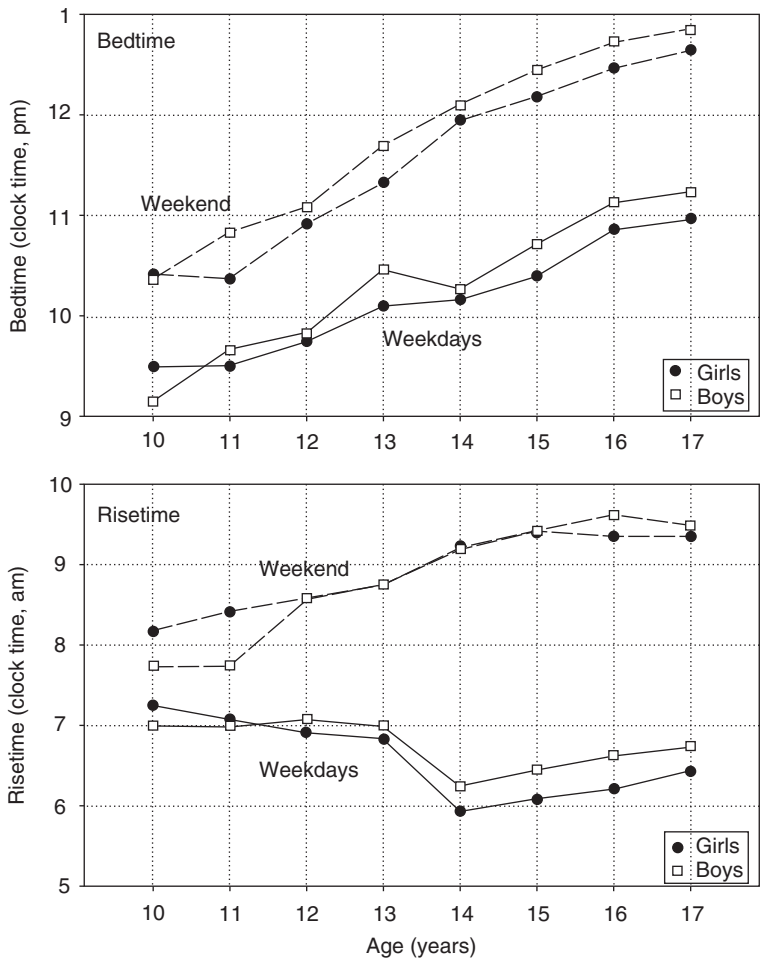

Figure 5 Adolescent self-reported sleep patterns taken from surveys performed in the 1990s with children living in the United States. Girls are represented by closed symbols and boys by solid symbols. Weekend bedtimes (top) and rise times (bottom) are shown with dotted lines; weekday (school day) schedules are indicated by solid lines. The sample included children ages 10–17 years who were enrolled in school at the time of the survey. Times on the *y*-axes are indicated using 24 h clock hours. Reproduced from Jenni OJ and Carskadon MA (2005) Infants to adolescents. In: *SRS Basics of Sleep Guide*, pp. 11–20. Westchester, IL: Sleep Research Society.

that compared to younger teens, older teens have increased numbers of technological devices (e.g., telephone, instant messaging, MP3 player, electronic games, television, computer, and Internet connection) accessible to them in their bedrooms. In addition, teens who reported having four or more such devices in the bedroom also reported sleeping 30 min less on average each night than those with fewer devices.

Consequences of Chronic Insufficient and Ill-Timed Sleep in Adolescents

The association of maturational changes of the intrinsic regulatory systems with the seductive nocturnal lifestyle options available to teenagers makes for a synergistic drive toward later timing of sleeping and waking for most teens. When confronted with the inevitability of an early school bell, a pattern emerges that comprises late, short, and irregular sleep. As noted previously, for some teens the combination results in extreme morning sleepiness and sleep-onset REM episodes. For others, the combination can be deadly. For example, in car crashes attributed to the driver having fallen asleep, the age of the driver is 16–25 years in more than 50% of cases.

Consequences of insufficient, irregular, and ill-timed sleep also include potential problems with mood regulation and depression, learning and memory, weight regulation, substance use or abuse, school tardiness and absenteeism, along with impulsivity and risk taking. Many of these behavioral tendencies are thought to be endemic problems of adolescence, yet the contributions of poor sleep may be involved. Serious consideration and scientific examination of these issues is needed. Trends point to continued deterioration of adolescent sleep patterns and at younger ages. Thus, whether adolescent sleep may affect the maturation of neurons, neural circuitry, or neurochemistry is in urgent need of study.

See also: Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Circadian Function and Therapeutic Potential of Melatonin in Humans; Circadian Rhythms: Influence of Light in Humans; Circadian Rhythms in Sleepiness, Alertness, and Performance; Endocrine Function During Sleep and Sleep Deprivation; Entrainment of Circadian Rhythms by Light; Narcolepsy; Photoreceptors and Circadian Clocks; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Architecture; Sleep and Sleep States: PET Activation Patterns; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function; Sleep: Development and Circadian Control; Sleep-Dependent Memory Processing.

Further Reading

Carskadon MA (ed.) (2002) *Adolescent Sleep Patterns: Biological, Social, and Psychological Factors*. New York: Cambridge University Press.

Carskadon MA and Acebo C (2002) Regulation of sleepiness in adolescents: Update, insights, and speculation. *Sleep* 25: 606–614.

Carskadon MA, Harvey K, Duke P, Anders TF, Litt IF, and Dement WC (1980) Pubertal changes in daytime sleepiness. *Sleep* 2: 453–460.

Carskadon MA, Labyak SE, Acebo C, and Seifer R (1999) Intrinsic circadian period of adolescent humans measured in conditions of forced desynchrony. *Neuroscience Letters* 260: 129–132.

Carskadon MA, Vieira C, and Acebo C (1993) Association between puberty and delayed phase preference. *Sleep* 16: 258–262.

Carskadon MA, Wolfson AR, Acebo C, Tzischinsky O, and Seifer R (1998) Adolescent sleep patterns, circadian timing, and sleepiness at a transition to early school days. *Sleep* 21: 871–881.

Dahl RE and Spear LP (eds.) (2004) *Annals of the New York Academy of Sciences: Adolescent Brain Development: Vulnerabilities and Opportunities*.

Feinberg I (1982) Schizophrenia: Caused by a fault in programmed synaptic elimination during adolescence? *Journal of Psychiatric Research* 17: 319–334.

Grigg-Damberger M, Gozal D, Marcus CL, et al. (2007) The visual scoring of sleep and arousal in infants and children. *Journal of Clinical Sleep Medicine* 3: 201–240.

Jenni OG, Achermann P, and Carskadon MA (2005) Homeostatic sleep regulation in adolescents. *Sleep* 28: 1446–1454.

Jenni OG and Carskadon MA (2004) Spectral analysis of the sleep electroencephalogram during adolescence. *Sleep* 27: 774–783.

Lenroot RK and Giedd JN (2006) Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging. *Neuroscience and Biobehavioral Reviews* 30: 718–729.

National Sleep Foundation (2006) *Sleep in America Poll Summary Findings*. Available at http://www.sleepfoundation.org.

Tanner JM (1962) *Growth at Adolescence*. Oxford: Blackwell.

Taylor DJ, Jenni OG, Acebo C, and Carskadon MA (2005) Sleep tendency during extended wakefulness: Insights into adolescent sleep regulation and behavior. *Journal of Sleep Research* 14: 239–244.

Sleep in Aging

M Rissling and S Ancoli-Israel, University of California at San Diego, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Background

Sleep is one of many biological processes that invariably undergoes change with age. Older adults may experience an increase in sleep disturbance as the homeostatic and circadian rhythm processes regulating the sleep cycle become more fragile. Sleep is also affected by the many complications associated with physical and mental illness that often accompany aging. It is important to understand the distinction between expected changes in sleep associated with aging and those likely to arise from pathology unrelated to aging *per se*. This article reviews both.

Changes in Sleep with Age

It is generally well accepted that aging results in an overall decrease in slow-wave sleep (SWS). The amplitude of slow-wave delta waves is attenuated, and the percentage of SWS observed is markedly decreased. This process is gradual and begins in young adulthood, at around 20 years of age, but stabilizes by about age 60, with no further reduction in SWS. There is evidence of a gender difference, with men showing greater decreases in SWS than women. Latency to rapid eye movement (REM) sleep is also decreased in the older adult and is considered to be a direct consequence of the circadian phase advancement that occurs with aging.

The largest overall change observed in older adults is an increase in stage shifts, with more time spent awake in bed following sleep onset. The disruptions are often characterized by early morning awakenings and an inability to fall back to sleep and may be due to a number of sleep disturbances (such as insomnia or other specific sleep disorders) rather than to aging. The result is marked decreases in sleep efficiency (defined by the percentage of sleep that occurs while in bed) and consequently increases in daytime sleepiness. Stage 1 sleep is slightly increased in association with such arousals; stage 2 sleep appears to be more or less constant throughout the life span.

The need for sleep, or the sleep drive, however, does not diminish with age. Rather, it is the ability to satisfy the sleep drive that becomes impaired. Therefore, although the need for sleep remains the same, nighttime sleep is curtailed, with the remaining sleep drive redistributed to the daytime. The decreased ability to maintain nighttime sleep is most often attributed to changes in circadian rhythms, insomnia comorbid with medical or psychiatric illnesses and medication use, and the presence of other specific sleep disorders, such as periodic limb movements in sleep or restless leg syndrome (RLS), REM behavior disorder, and sleep-disordered breathing.

Changes in Circadian Rhythm

Circadian rhythms, such as hormone secretion, body temperature, and the sleep–wake cycle, fluctuate every 24 h. In humans, the sleep–wake cycle is regulated by the combined effect of an endogenous pacemaker and exogenous stimuli. The hypothalamic suprachiasmatic nucleus (SCN) is considered the endogenous clock of the brain and is believed to be the primary mediator of circadian rhythms. Circadian rhythms maintained by the SCN are naturally entrained by exogenous stimuli, or zeitgebers (time cues). The most significant zeitgeber is light, but other factors, such as physical activity, the timing of meals, and social interactions, have also been shown to be important regulators.

As people age, their circadian rhythms become weaker, desynchronize, and lose amplitude. It is hypothesized that the progressive deterioration of the SCN and its subsequent weakened functioning contribute to the disruption of circadian rhythms in older adults. The most common desynchronized pattern often seen in older adults is an advancement of the sleep phase.

Older adults with advancement of the sleep phase have increased numbers of early morning awakenings and difficulty maintaining sleep. They typically experience an increased drive for sleep between 7 p.m. and 9 p.m. and a drive to wake up between 3 a.m. and 5 a.m. Despite being sleepy earlier in the evening, they will often try to stay up later because of societal pressure, resulting in fewer hours spent in bed sleeping. The decreased total sleep time at night then results in daytime sleepiness and napping. Some older adults take inadvertent evening naps in front of the television or while reading early in the evening and then have difficulty falling asleep when they go to bed later in the evening; yet they still wake up early in the morning. This pattern gets interpreted as a complaint of insomnia, that is, difficulty falling asleep, difficulty staying asleep, and daytime sleepiness.

Treatment of Circadian Rhythm Disturbances

The most common treatment for circadian rhythm disturbances is bright light therapy. An advanced circadian rhythm can be delayed and the sleep–wake

cycle strengthened with evening light exposure. As the sun is the strongest source of bright light, older adults with advanced rhythms should spend more time outdoors during the late afternoon or early evening and avoid bright light in the morning hours (as morning light would advance their rhythms even more). Studies have also shown that exposure to artificial light via a bright light box in the early evening has a similar effect and can improve sleep continuity in both healthy and institutionalized older adults. Adherence to a regular sleep schedule will promote a more robust sleep–wake cycle. Physical activity has also been shown to improve nighttime sleep quality. Such therapies are without adverse effects, and each has been shown to improve mood, performance, and quality of life and to reduce fatigue.

While melatonin has been shown to be safe, the efficacy for circadian rhythm disturbances has not been established. There is little consensus on the recommended dose or timing of administration, and melatonin purity and composition are currently not regulated in the United States by the Food and Drug Administration (FDA).

Insomnia

Insomnia is defined as difficulty falling asleep or staying asleep or nonrestorative sleep which results in daytime consequences. Insomnia is found in 30–50% of those more than 60 years old. Gender differences exist as well, with older women being more likely to complain of insomnia than older men. In older adults, sleep maintenance insomnia (an inability to maintain sleep throughout the night) and early morning or terminal insomnia (awakening early in the morning, with difficulty returning to sleep) are the most common complaints. Insomnia in the older adult is most often comorbid with medical and psychiatric illness or the medications used to treat those illnesses. Insomnia may also be a complaint in patients with specific sleep disorders, such as RLS or periodic limb movements in sleep.

Consequences of Insomnia

The consequences of disturbed sleep are well known and may include difficulty sustaining attention, slowed response time, difficulty with memory, and decreased performance. These consequences can be more pronounced in older adults and may even be misinterpreted as symptoms of dementia. Insomnia disrupts the sleep of 64% of bed partners of both healthy and demented older adults and 84% of caregivers (i.e., who may sleep in separate rooms). Nighttime restlessness secondary to insomnia is often cited as the primary reason for institutionalizing a demented relative.

Untreated insomnia is associated with a nearly 500% increase in the general risk for serious accidents and injuries, including car accidents due to daytime sleepiness and hip fractures due to falls. In the older adult, insomnia is most often comorbid with medical and psychiatric disorders and the medications used to treat them.

Medical and Psychiatric Illness and Insomnia

Older adults are the most susceptible to comorbid insomnia as they are more likely to have multiple medical conditions. Data from the National Sleep Foundation 2003 poll on sleep in older adults showed that those with three or four medical conditions, including cardiac disease, pulmonary disease, stroke, and depression, reported significantly more sleep complaints than did those with none, one, or two conditions.

Medical and psychiatric illnesses most commonly associated with insomnia include pain resulting from osteoarthritis or malignancies, chronic obstructive pulmonary disease, congestive heart failure, nocturia (e.g., due to enlarged prostate) and incontinence, neurological deficits (e.g., stroke, Parkinson's disease, and dementia), gastroesophageal reflux, depression, and anxiety disorder. Difficulty falling asleep has a prevalence rate of 31% for patients with osteoarthritis and 66% for those with chronic pain conditions. In addition, 81% of arthritis patients, 85% of chronic pain patients, and 33% of diabetes patients report difficulty staying asleep.

Two additional conditions that are associated with aging and contribute to poor sleep are nocturia and menopause. Nocturia or frequent voluntary voiding of urine during the night and nocturnal polyuria, the production of an abnormally large volume of urine during sleep, are common causes of awakenings and disruption of nighttime sleep. Although nocturia is often perceived as causing nocturnal awakenings, it has been hypothesized that the fragmented sleep experienced by the older adult leads to more awareness of the need to void. The circadian rhythm of urine excretion is also altered with age such that there is a fall of renal concentrating ability, sodium conservation, and secretion of renin–angiotensin–aldosterone, and arginine vasopressin secretion and an increase of atrial natriuretic hormone. The bladder's ability to store urine decreases with age, and age-related lower urinary tract problems, such as detrusor overactivity, can affect frequency and urgency.

Menopause is associated with hormonal, physiologic, and psychological changes that affect sleep and may play a critical role in modulating both the presence and the degree of sleep disruption. Menopausal

vasomotor symptoms or hot flashes are known to lower sleep efficiency and cause multiple arousals during the night. Insomnia is a common symptom of menopause and may be partially related to psychological factors such as depression and anxiety.

In fact, insomnia can be prognostic or indicative of depression. While depressed mood may predict insomnia, unresolved insomnia also places adults at greater risk for developing depression, anxiety, and substance abuse. Studies in younger adults have suggested that treating the insomnia may improve depression.

Medication Use and Insomnia

Alerting or stimulating drugs, such as central nervous system stimulants, beta-blockers, bronchodilators, calcium channel blockers, corticosteroids, decongestants, stimulating antidepressants, stimulating antihistamines, and thyroid hormones are known to cause difficultly falling asleep when taken late in the day. Sedating drugs, such as sedative antihistamines, antihypertensives, and sedative antidepressants, may cause excessive daytime sleepiness when taken early in the day and contribute to daytime napping behavior and consequently delay nighttime sleep onset. Adjusting the dose or the time of day of administration often improves the insomnia.

Treatment of Insomnia

Effective treatments for insomnia in older adults include cognitive behavioral therapy (CBT) and pharmacological therapy.

CBT CBT is a combination of cognitive restructuring and behavioral modification for insomnia. The cognitive portion deals with misconceptions or unrealistic expectations about sleep (i.e., absolute requirement of 8 h or more of sleep), amplification of the consequences of not getting enough sleep, incorrect thinking about the cause of insomnia, and misunderstandings about healthy sleep practices. The behavioral component of CBT involves a combination of sleep restriction therapy, stimulus control therapy, and good sleep hygiene practices. Tables 1–3 list the rules for each of these, respectively.

A recent National Institutes of Health (NIH) State-of-the-Science Conference on Insomnia concluded that CBT is as effective as prescription medications for the treatment of chronic insomnia and that there are indications that the beneficial effects of CBT may last well beyond the termination of treatment. If pharmacologic treatment is initiated, it is most effective when combined with CBT.

Table 1 Stimulus control therapy

Stimulus control is a behavioral therapy that seeks to remove the association of sleeplessness with the bed. The following steps reflect the principles of stimulus control:
Step 1. Go to bed only when sleepy.
Step 2. Use the bed only for sleeping – do not read, watch television, or eat in bed.
Step 3. If unable to fall asleep within about 20 min (don't watch the clock), move to another room. Stay up until really sleepy. The goal is to associate the bed with falling sleep quickly.
Step 4. If you are awake at night for more than about 20 min, get out of bed.
Step 5. Repeat Step 4 as often as necessary.
Step 6. Awaken at the same time every morning regardless of total sleep time.
Step 7. Avoid naps.

From Bootzin RR, Epstein D, and Wood JM (1991) Stimulus control instructions. In: Hauri PJ (ed.) *Case Studies in Insomnia*, pp. 19–28. New York: Plenum; Bootzin RR and Epstein D (2000) Stimulus control. In: Lichstein KL and Morin CM (eds.) *Treatment of Late-Life Insomnia*, pp. 167–184. Thousand Oaks, CA: Sage and Morin C (2005) Psychological and behavioral treatments for primary insomnia. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, 4th edn., pp. 726–748. Philadelphia: Elsevier/Saunders.

Table 2 Sleep restriction therapy

Sleep restriction therapy that seeks to keep patients from spending excessive time in bed. The following steps reflect the principles of sleep restriction therapy:
Step 1: Cut bedtime to actual amount patient reports sleeping but not less than 4 h per night.
Step 2: Prohibit sleep outside these hours.
Step 3: Get up at the same time each morning.
Step 4: Have patient report daily the amount of sleep obtained.
Step 5: Compute sleep efficiency (SE); based on moving average of five nights; when SE is >85%, increase bedtime by15 min.
Step 6: Compute sleep efficiency (SE); when SE is > 80%, increase time in bed by going to bed 15 min earlier. Continue to get up at same time each morning.

From Glovinsky PB and Spielman AJ (1991) Sleep restriction therapy. In: Hauri PJ (ed.) *Case Studies in Insomnia*, pp. 49–63. New York: Plenum; Morin C (2005) Psychological and behavioral treatments for primary insomnia. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, 4th edn., pp. 726–748. Philadelphia: Elsevier/Saunders; and Spielman AJ, Saskin P, and Thorpy MJ (1987) Treatment of chronic insomnia by restriction of time in bed. *Sleep* 10: 45–56.

Pharmacological therapy There are several classes of medications used to treat insomnia in older adults, in particular benzodiazepine receptor agonists (benzodiazepines and nonbenzodiazepines), melatonin receptor agonists, antidepressants, antipsychotics, and anticonvulsants. In addition, over-the-counter

Table 3 Sleep hygiene rules for older adults

1. Maintain a consistent bedtime and uptime.
2. Use the bed only for sleep and satisfying sex. Avoid watching television, eating, or discussing stressful issues while in bed.
3. Minimize noise and temperature extremes.
4. Keep the bedroom environment as dark as possible. Minimize the exposure to bright light if you get out of bed during the night. Use the dimmest night-light possible.
5. Avoid naps if you have difficulty falling asleep at night. If must nap, restrict napping to 30 min or less in the early afternoon.
6. Avoid nicotine near bedtime and during night awakenings.
7. Avoid all forms of caffeine for 4–6 h before bedtime.
8. Although alcohol may help you fall asleep, metabolism of alcohol will cause awakenings.
9. Avoid eating heavy meals and large amounts of protein close to bedtime. A light meal of carbohydrates or dairy products (particularly milk) is preferable.
10. Avoid vigorous exercise just before bedtime.
11. Avoid having pets in bed at nighttime.

From Morin C (2005) Psychological and behavioral treatments for primary insomnia. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, 4th edn., pp. 726–748. Philadelphia: Elsevier/Saunders; and Hauri PJ (1991) Sleep hygiene, relaxation therapy, and cognitive interventions. In: Hauri PJ (ed.) *Case Studies in Insomnia*, pp. 65–86. New York: Plenum.

Table 4 Does and half-life of US Food and Drug Administration-approved agents for insomnia

Generic agent (trade name)	*Dose (mg)*[a]	*Half-life (h)*
Benzodiazepine receptor agonists: benzodiazepines		
Flurazepam HCL (Dalmane)	15, 30	47–100
Quazepam (Doral)	7.5, 15	39–73
Estazolam (ProSom)	0.5, 1, 2	10–24
Temazepam (Restoril)	7.5, 15, 30	3.5–18.4
Triazolam (Halcion)	0.125, 0.25	1.5–5.5
Benzodiazepine receptor agonists: nonbenzodiazepines		
Eszopiclone (Lunesta)	1, 2, 3	6
Zaleplon (Sonata)	5, 10, 20	1
Zolpidem (Ambien)	5, 10	1.4–4.5
Zolpidem MR (Ambien CR)	6.25, 12.5	2.8–2.9
Melatonin receptor agonist		
Ramelteon (Rozerem)	8	2–5

[a]The lowest dose is generally recommended as the starting dose for older people.
Physicians' Desk Reference (1991, 1999, and 2004).
FDA (2004) Lunesta (eszopiclone) Tablets: 1 mg, 2 mg, 3 mg. Retrieved January 24, 2008, from www.fda.gov/cder/foi/label/2004/021476lbl.pdf.
From Consensus conference (1984) *JAMA* 251: 2410–2414.

Table 5 Indications for the newer nonbenzodiazepine and melatonin receptor agonist agents

Agent	*Indicated for sleep onset*	*Indicated for sleep maintenance*	*Required time in bed (h)*
Eszopiclone[a]	X	X	8
Zaleplon[b]	X		4
Zolpidem[c]	X		7–8
Zolpidem MR[d]	X	X	7–8
Ramelteon[e]	X		No minimum time

[a]From Lunesta (eszopiclone) prescribing information. Sepracor.
[b]From SONATA (zaleplon) prescribing information. King Pharmaceuticals, Inc.
[c]From Ambien (zolpidem tartrate) prescribing information. Sanofi-Synthelabo Inc.
[d]From Ambien CR (zolpidem tartrate extended-release tablets) prescribing information. Sanofi-Synthelabo Inc.
[e]From Rozerem (ramelteon) prescribing information. Takeda Pharmaceutical Company Inc.

medications such as antihistamines (diphrenhydramine) and nutritional or herbal supplements are often used. The NIH conference concluded that there is no systematic evidence for the effectiveness of antidepressants, antipsychotics, anticonvulsants, antihistamines, or herbal preparations in the treatment of insomnia. The panel also expressed significant concerns about the risks associated with the use of these medications, particularly in older adults.

The NIH panel also concluded that the benzodiazepine receptor agonists and in particular the newer nonbenzodiazepines have been shown to be the most effective and safest of the sedative hypnotics (at the time of the conference, the melatonin receptor agonist remelteon had not yet been approved by the FDA). The key to successful pharmacologic treatment of insomnia lies in choosing the sedative or hypnotic with the characteristics that best fit the patient's complaints. (It is important to note that only ten agents have been approved by the FDA for the treatment of insomnia. It is important to note that there are only 10 agents approved by the FDA for the treatment of insomnia, listed in Table 4 with the recommended dose and half-life. Table 5 lists the indications for the five newer nonbenzodiazepine receptor agonists and the melatonin receptor agonist.)

If sedatives or hypnotics are presecribed, the potential side effects must be taken into account. The administration of long-acting benzodiazepines can cause adverse daytime effects such as excessive daytime sleepiness and poor motor coordination, which may lead to injuries. In older adults, the risk of falls, cognitive impairment, and respiratory depression are of particular concern, although recent data suggest that insomnia *per se*, and not hypnotics, might increase the risk of falls. Chronic use of long-acting benzodiazepines can lead to tolerance and withdrawal symptoms if they are abruptly discontinued. In addition, the potential for exacerbating coexisting

medical conditions such as severe hepatic or renal disorders exists when some of these medications are used. The newer nonbenzodiazepine agents have been shown to be effective with a lower propensity for causing clinical residual effects, withdrawal, dependence, or tolerance.

Specific Sleep Disorders

There are also specific sleep disorders which are quite common in the older adult and which contribute to poor sleep. These would include sleep-disordered breathing (SDB), RLS and periodic limb movements in sleep, and REM behavior disorder.

SDB

SDB, also called sleep apnea, is a disorder in which the patient periodically stops breathing during sleep. Each apnea (complete cessation in respiration) or hypopnea (partial decrease in respiration) lasts a minimum of 10 s and generally ends with an arousal. This results in fragmented sleep and nocturnal hypoxemia. The total number of apneas plus hypopneas per hour of sleep is called the apnea–hypopnea index (AHI) or respiratory disturbance index (RDI). Depending on the laboratory, an AHI or RDI greater than or equal to 5–10 is required for the diagnosis of SDB.

The prevalence of SDB in older adults has been estimated to be between 30% and 60%, substantially higher than in younger adults. Cross-sectional studies have shown that there is a small increase in prevalence of SDB with increasing 10-year age groups. In the longest longitudinal study, older adults followed for 18 years showed changes in AHI only when associated with changes in body mass index.

There are two main types of SDB. The first, central sleep apnea, occurs when the respiratory centers in the brain shut down during sleep, and there is no attempt to breathe. Central sleep apnea is most common in congestive heart failure and is often associated with Cheyne–Stokes breathing. The second, obstructive sleep apnea (OSA), is the more common and occurs when the muscles of the airway collapse. The two main symptoms of OSA are loud snoring and excessive daytime sleepiness (EDS). EDS, which results from the recurrent nighttime arousals and sleep fragmentation, may manifest as unintentional napping or falling asleep at inappropriate times during the day. EDS can cause reduced vigilance and is associated with cognitive deficits, which may be particularly serious in older adults who may already have some cognitive impairment at baseline.

In addition to the cognitive deficits that may occur as a result of SDB, there is some evidence that many of the progressive dementias (e.g., Alzheimer's and Parkinson's disease) involve degeneration in areas of the brain stem responsible for the regulation of respiration and other autonomic functions relevant to sleep maintenance, thus placing the patient at increased risk for SDB.

Patients with SDB are at greater risk for hypertension, cardiac arrhythmias, congestive heart failure, myocardial infarction, and stroke. Most of the research, however, has not focused on older adults, so the relationship between SDB and some of these various morbidities remains unknown in the older population. Research that has targeted older adults has reported an association between SDB and the risk of developing cardiovascular disease, including coronary artery disease and stroke.

Treatment of SDB Treatment of SDB in older adults should be guided not only by the severity of the SDB but by the significance of the patient's symptoms. Patients with more severe SDB (AHI > 20) deserve a trial of treatment. For those with milder SDB (AHI < 20), treatment should be considered if comorbid conditions such as hypertension, cognitive dysfunction, or EDS are present. Treatment should never be withheld because of age or because of assumed noncompliance.

The gold standard for treating SDB is continuous positive airway pressure (CPAP), a device which provides continuous positive pressure via the nasal passages or oral airway, thus creating a pneumatic splint to keep the airway open during inspiration. If used appropriately, CPAP has been shown to be safe and efficacious. Alternative treatments include oral appliances, weight loss (although most older adults with SDB are not obese), and surgery; however, no other treatments have been shown to be as effective as CPAP.

Significance of SDB in development The presentation of sleep apnea in the older adult is not necessarily the same as in the younger adult. This has led to the idea that OSA in older adults might be a different disorder from OSA in younger adults. Two questions that arise are whether OSA is pathological in older adults and whether OSA in the older patient should be treated, and if so, when?

From a clinical standpoint, does it matter what is causing SDB in the older adult? Does it matter if the SDB in the older adult is the same as or different from the SDB seen in the younger adult? Probably not. What matters is whether SDB in the older adult has negative consequences. Most studies suggest that SDB does not increase the risk of mortality in the older adult, but that older adults with SDB are excessively sleepy and that SDB may contribute to decreased

quality of life, increased cognitive impairment, and greater risk of nocturia, hypertension, and cardiovascular disease. In addition, older adults with both insomnia and SDB have more functional impairment than those with just insomnia.

Any patient, no matter the age, who presents with symptoms of impairment as well as with a history of traffic accidents and repeated falls, should be evaluated for a sleep disturbance. Treatment of a symptomatic older patient should not be withheld on the basis of age, as the consequence of SDB can be just as serious in older as in younger patients. The bottom line is that if the sleep apnea is associated with clinical symptoms (e.g., hypertension, cognitive dysfunction, nocturia, high levels of sleep-disordered breathing, cardiac disease), then it should be treated, regardless of the age of the patient.

Periodic Limb Movements and RLS

The condition known as periodic limb movements in sleep (PLMS) is characterized by repetitive leg movements during sleep, occurring every 20–40 s and recurring several hundred times over the course of the night, each causing a brief awakening, which results in sleep fragmentation. Patients with PLMS may complain of EDS and/or insomnia. The number of limb movements per hour of sleep is called the periodic limb movement index, with ten or more limb movements per hour needed for a clinical diagnosis. A diagnosis of PLMS therefore must be made with an overnight sleep study that records leg movements. The etiology of PLMS is unknown. PLMS can be seen in patients with fibromyalgia and in conjunction with other primary sleep disorders, including SDB and narcolepsy.

While the prevalence of PLMS in adults is estimated to be 5–6%, the rate increases dramatically with age, with reported prevalence rates of up to 45% in community-dwelling people over the age of 65.

RLS, a condition strongly linked to PLMS, is characterized by leg dysesthesia, often described as 'a creepy crawling' or 'restless' sensation, occurring in the relaxed awake state. The uncomfortable sensations can be relieved only by movement. The diagnosis of RLS is made on the basis of history alone, often with the question, When you try to relax in the evening or sleep at night, do you ever have unpleasant, restless feelings in your legs that can be relieved by walking or movement? Patients may have no knowledge that they kick, so interviewing the patient's bed partner may be helpful in elucidating the history of both RLS and PLMS. Patients with symptoms of RLS should also be assessed for anemia, uremia, and peripheral neuropathy as each of these conditions can cause or exacerbate RLS.

Treatment of PLMS and RLS RLS and PLMS are treated primarily with dopamine agonists, which reduce the number of kicks and associated arousals. At this time, only ropinirole and pramipexole have been FDA-approved for the treatment of RLS.

REM Behavior Disorder

REM behavior disorder (RBD) is characterized by the intermittent absence of normal skeletal muscle atonia during REM sleep, resulting in patients' acting out their dreams. RBD generally occurs during the last third of the night, when REM sleep is more common. The movements during RBD may be violent and potentially harmful to patient and bed partner. Vivid dreams, consistent with the patient's aggressive and/or violent behavior, may be recalled on waking.

The prevalence of RBD is unknown, but it has been shown to be most common in older men. The etiology of RBD is also unknown, but there is a strong association between idiopathic RBD and neurodegenerative diseases, including Parkinson's disease, multiple system atrophy, and Lewy bodies dementia, with the RBD often preceding the other disorders by years. Data suggest that close to half of RBD patients develop Parkinson's disease within several years of the RBD diagnosis. The onset of acute RBD has been associated with stress disorders and the use of monoamine oxidase inhibitors and fluoxetine as well as withdrawal of REM-suppressing agents (e.g., alcohol, tricyclic antidepressants, amphetamines, and cocaine).

Treatment of RBD Treatment of RBD with clonazepam, a long-acting benzodiazepine, results in partial or complete cessation of abnormal nocturnal motor movements in 90% of patients. However, because of the long half-life, many patients report residual sleepiness in the morning and during the day. Recently, melatonin has also been shown to be an effective treatment. In addition to the pharmacologic treatment, other aspects of treatment should include good sleep hygiene and education for the patient and bed partner. The bedroom should be made safer by removing heavy or breakable or potentially injurious objects from the bed's vicinity. Heavy curtains should be placed on bedroom windows and doors, and windows should be locked at night. Finally, to avoid falling out of bed, patients may consider sleeping on a mattress on the floor.

Conclusion

The ability to have restful sleep decreases with age but is likely independent of age itself. Medical and psychiatric disorders, medication, circadian rhythm changes, and primary sleep disorders (such as SDB,

RLS or PLMS, and RBD) contribute to older adults' decreased ability to maintain a full night's sleep. Each of these conditions is treatable, and health care professionals should never withhold treatment because of age or because of assumed noncompliance.

See also: Circadian Genes and the Sleep–Wake Cycle; Endocrine Function During Sleep and Sleep Deprivation; Pharmacology of Sleep: Adenosine; Psychiatric Disorders Associated with Disturbed Sleep and Circadian Rhythms; Restless Leg Syndrome; Reticular Activating System; Sleep and Sleep States: Thalamic Regulation; Sleep and Sleep States: Hypothalamic Regulation; Sleep and Sleep States: Histamine Role; Sleep and Sleep States: Cytokines and Neuromodulation; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Architecture; Sleep and Sleep States: Hippocampus–Neocortex Dialog; Sleep and Sleep States: Network Reactivation; Sleep and Sleep States: Gene Expression; Sleep Apnea; Sleep and Circadian Rhythm Disorders in Human Aging and Dementia; Sleep: Development and Circadian Control; Sleep–Wake State Regulation by Noradrenaline and Serotonin; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Ancoli-Israel S (2005) Sleep and aging: Prevalence of disturbed sleep and treatment considerations in older adults. *Journal of Clinical Psychiatry* 66: 24–30.

Ancoli-Israel S (2007) Guest editorial: Sleep apnea in older adults: Is it real and should age be the determining factor in the treatment decision matrix? *Sleep Medicine Review* 11: 83–85.

Ancoli-Israel S, Kripke DF, Klauber MR, et al. (1991) Sleep disordered breathing in community-dwelling elderly. *Sleep* 14(6): 486–495.

Bliwise DL (1993) Sleep in normal aging and dementia. *Sleep* 16(1): 40–81.

Bootzin RR and Epstein D (2000) Stimulus control. In: Lichstein KL and Morin CM (eds.) *Treatment of Late-Life Insomnia*, pp. 167–184. Thousand Oaks, CA: Sage.

Bootzin RR, Epstein D, and Wood JM (1991) Stimulus control instructions. In: Hauri PJ (ed.) *Case Studies in Insomnia*, pp. 19–28. New York: Plenum Press.

Edinger JD, Heolscher TJ, Marsh GR, Lipper S, and Ionescu-Pioggia M (1992) A cognitive-behavioral therapy for sleep-maintenance insomnia in older adults. *Psychological Aging* 7: 282–289.

Foley DJ, Monjan A, Simonsick EM, Wallace RB, and Blazer DG (1999) Incidence and remission of insomnia among elderly adults: An epidemiologic study of 6,800 persons over three years. *Sleep* 22: S366–S372.

Glass J, Lanctot KL, Herrmann N, Sproule BA, and Busto UE (2005) Sedative hypnotics in older people with insomnia: Meta-analysis of risks and benefits. *British Medical Journal* 331: 1169.

Glovinsky PB and Spielman AJ (1991) Sleep restriction therapy. In: Hauri PJ (ed.) *Case Studies in Insomnia*, pp. 49–63. New York: Plenum.

Hauri PJ (1991) Sleep hygiene, relaxation therapy, and cognitive interventions. In: Hauri PJ (ed.) *Case Studies in Insomnia*, pp. 65–86. New York: Plenum.

Lichstein KL and Reidel BW (1994) Behavioral assessment and treatment of insomnia: A review with an emphasis on clinical application. *Behavior Therapy* 25: 659–688.

Martin JL and Ancoli-Israel S (2003) Insomnia in older adults. In: Szuba MP, Kloss JD, and Dinges DF (eds.) *Insomnia: Principles and Management*, pp. 136–154. Cambridge, UK: Cambridge University Press.

Morin C (2005) Psychological and behavioral treatments for primary insomnia. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, 4th edn., pp. 726–748. Philadelphia: Elsevier/Saunders.

Morin CM, Colecchi C, Stone J, Sood R, and Brink D (1999) Behavioral and pharmacological therapies for late life insomnia. *Journal of the American Medical Association* 281: 991–999.

National Institutes of Health (2005) State-of-the-Science conference statement on manifestations and management of chronic insomnia in adults. *Sleep* 28(9): 1049–1057.

Spielman AJ, Saskin P, and Thorpy MJ (1987) Treatment of chronic insomnia by restriction of time in bed. *Sleep* 10: 45–56.

Van Someren EJ (2000) Circadian rhythms and sleep in human aging. *Chronobiology International* 17(3): 233–243.

Vitiello MV (2000) Effective treatment of sleep disturbances in older adults. *Clinical Cornerstone* 2(5): 16–47.

Relevant Websites

http://www.aasmnet.org – American Academy of Sleep Medicine.
http://www.sleepfoundation.org – National Sleep Foundation.
http://www.sleepresearchsociety.org – Sleep Research Society.

Sleep Mentation in REM and NREM: A Neurocognitive Perspective

J S Antrobus, The City College of New York, New York, NY, USA
E J Wamsley, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Dreaming in Rapid Eye Movement Sleep

The neurocognitive study of sleep mentation began when Aserinsky, in the first all-night electroencephalograph (EEG) study of sleep, observed periodic intervals of rapid eye movements (REMs) associated with reports of long, visually vivid, and often bizarre dreams. Although Aserinsky quite plausibly assumed that sleepers were scanning their dream imagery, this assumption has not been confirmed. Relative to periods of non-REM (NREM) sleep, REMs occur during periods of low-voltage, high-frequency EEGs so similar to that of waking that European scientists called it 'paradoxical' sleep. The first REM period typically follows about 90 min of NREM sleep and lasts about 10 min. Subsequent REM periods increase in duration and make up a total of 21% of sleep time. Although REMs are most frequent at the onset of each REM period, and are associated with the most visually vivid dream reports, vivid dreaming is reported throughout the period, even after intervals of REM quiescence. Sleepers report some content from about 80% of REM awakenings, as compared to 50% of NREM awakenings. REM reports are, on average, more vivid, emotional, storylike, and bizarre than are NREM reports. Relative to these REM reports, NREM reports are generally shorter and thoughtlike, more often consisting of isolated, static images.

Although the association between REM sleep and dreaming is one of the strongest, nonpathological, neurocognitive associations ever discovered, in 1953 what was known about waking neurophysiological processes was insufficient to instruct a model of dreaming sleep. Consequently, neurological models of dreaming initially emphasized the pontine processes that produced the alternation between REM and NREM sleep, at the expense of describing how cortical structures might generate the thought and imagery of those states. REM/NREM models were also silent about the fact that 6% of NREM reports were more dreamlike than were REM reports, matched by sleeper and time of night, and about the rather dramatic increase in the dreamlike quality of both REM and NREM reports in the late morning.

The Activation–Synthesis Model

Though the study of reported sleep mentation is necessarily limited to humans, the first brain models of REM/NREM sleep were based on the cat. Hobson, McCarley, and their colleagues solved the puzzle of the paradoxical wakelike EEG pattern of REM sleep when they showed that it is produced by reticular–thalamic–cortical circuits similar to those that Moruzzi and Magoun had previously shown to produce general wakefulness. The 90-min periodic alternation between REM and NREM sleep is controlled by pontine circuits. The primary distinction between waking and REM sleep is that the pontine-initiated REM activation process strongly inhibits both afferent input, from the sensory pathways, and efferent motor output. In short, it produces an active brain, with high sensory thresholds and, except for the oculomotor system, motor paralysis. In REM sleep, the dreamer can attempt to run, but efferent commands from the motor cortex are inhibited and the afferents that normally confirm feedback of movement, both kinesthetic and visual, are not received. The dreamer may interpret this to mean that he/she is 'stuck.'

But how, in the absence of sensory input, does the brain produce imagery and thought? In their activation–synthesis model (1977), Hobson and McCarley proposed that the 'input' is oculomotor information initiated in the pons and transmitted via pontogeniculooccipital (PGO) circuits to the cortex. In the attempt to 'synthesize' this information, the cortex creates a semicoherent visual story (i.e., the dream).

The large amplitude of the PGO spikes during REM sleep in cats suggested that they might interrupt the cortical EEG and thereby the coherence of the ongoing story sequence, producing the bizarre characteristics of the dream. But this assumption has yet to receive empirical support. Because of the pontine origin of this circuit, Hobson further argued that the Freudian assumption of a cortical (i.e., higher order) origin of dream process was without merit. The synthesis portion of the model is based on the quite reasonable neurocognitive isomorphic assumption that the cortex interprets subcortical processes similarly during sleep and waking. Thus, a command from the lateral geniculate nucleus to move the eyes rightward in waking should be interpreted similarly in REM sleep. However, the geniculooculomotor left–right orientation pattern of waking is not sustained, at least consistently, in REM sleep – the isomorphism assumption does not hold in this particular case.

Synthesis and Dream 'Incorporation'

The stimulus–response experimental procedures that are so effective in waking cognitive neuroscience are almost useless in sleep, since processing a stimulus generally drives the sleeper out of the biological state being studied. Nevertheless, once it was established that REM sleep is the time when the likelihood of dreaming is high, investigators began to systematically study how the dreamer 'incorporates' external stimuli into the dream. Although a fine spray of water or a soft bell presented during sleep generally elicit a short interval of EEG alpha, sleep often returns quickly to the prior REM state. Using this procedure, Dement showed that the time interval between stimulus presentation and awakening was similar to the length of the subsequently reported dream experience – which discredited the theory that dreams occurred in the momentary transition between sleep and waking.

The content of such incorporations, while fascinating, provided little information about the processes by which they were produced. Their salient characteristic was that they rapidly made remarkably good sense out of one or more features of the external stimulus – within the apparently prior context of the ongoing dream episode. For example, Dement found that spraying individuals with water during REM sleep led one dreamer to report that his children had spilled water on him while he was reading a book, whereas another dreamer reported that while watching an actress walk on stage ahead of him, it began to rain through the ceiling, so the actress immediately opened her umbrella. These quick, apparent 'interpretations' by the sleepers, in response to their wet skin, are remarkably similar to the left-brain interpretations of right-brain input later reported by Gazzaniga in split-brain patients.

Activated Cortex as an Attractor Landscape

A neural network attractor is essentially a network with recurrent connectivity that represents a well-learned feature, object, person, or event. Attractors that share many features ('wet face,' 'wet clothes,' 'rain,' 'shower') lie closer together than do attractors that represent dissimilar objects. A large array of attractors can be represented in an attractor landscape in which each attractor is a basin within a large plane; attractors with similar features are neighbors within a common attractor valley. When information reaches the landscape, it is cycled recurrently until it 'settles' into (i.e., recognizes) one particular attractor. Cortical regions that process different classes of features or events may be represented as different, but interconnected, landscapes. Several characteristics of attractors make them particularly useful for representing the brain processes that generate dreaming and other forms of sleep mentation. Even when information is extremely 'noisy,' it can settle into the basin of an attractor. Second, the settling process can be strongly biased by prior context, whether the context is the preceding dream narrative, the events of the preceding day, or chronic life stressors. Biases may be represented as temporary deepening of attractor basins – neurally by long-term potentiation (LTP) of the synapses that form their neural networks, particularly in the hippocampal system. Finally, the patterns of regional cortical activation that distinguish REM, NREM, and waking states can be represented by leveling or deepening the basins across different regional attractor landscapes.

The pontine–reticular–cortical activation network described by Hobson and his colleagues provides a basis for amplifying a large cortical attractor network during REM sleep. Other sources of activation capable of supporting attractor networks, thereby sleep mentation, even in NREM sleep, will be described in the following sections. This attractor model is nicely supported by the 'incorporation' studies, whereby a single feature of an external stimulus is attracted to a basin that has previously been deepened by the context of the prior dream. For example, if the dream is located in the living room close to the telephone, when the experimenter, Dement, rings a buzzer, the attractor landscape 'interprets' the sound as the telephone, and the dreamer gets up to answer it.

Incorporation studies of sleep mentation imply that external input is the origin of the sleep mentation content. A neural network attractor model, however, can 'recognize' and produce well-learned sequences of activity even when the input is sheer neural noise (i.e., possesses no information). This suggests that an activated cortical region can produce the same type of output that it was trained to do in the waking state, even when there is no concurrent external stimulus input. It further implies that the higher the level of regional activation, the deeper the attractor basins, and therefore the greater the likelihood that any noisy information will settle into one of the basins of attraction (i.e., produce mentation).

Characteristics of Sleep Mentation

Hallucinatory Quality

The ability of activated cortical attractors to produce vivid, even bizarre, imaginal sequences is characteristic not only of REM sleep but also of waking mentation, under conditions of mild sensory deprivation. The truly unique characteristic of the REM dream is its hallucinatory quality. Because inhibition of sensory

input allows minimal processing of external stimuli, the constructed imaginal world is experienced as veridical perception. The dreamer imaginally interacts with the imaginal perceptions. By contrast, though waking daydreams in a dark room may be similarly vivid, the daydreamer reports them as imaginal. We assume that the hippocampal system tends to maintain a background representation of the daydreamer's external environment even when the daydreamer is not actively attending to those stimuli.

Lucid Dreaming

Opposing this hallucinatory, defining characteristic of REM dreams is the relatively infrequent phenomenon of 'lucid' dreams, whereby the 'sleeper' reports being aware that he/she is dreaming, and in some cases is able to deliberately modify the events of his/her ongoing dream. Lucid dreams are commonly reported from the final hours of morning sleep. The phenomenon suggests that the hallucinatory characteristic of dreamlike mentation may have a neurological basis different from that of other properties. Circadian-driven reticular formation/cortical circuits may not produce the deafferentation associated with the REM/NREM cycles, so that sensory, including proprioceptive, thresholds are lower than in earlier REM periods. These lower thresholds may allow the lucid dreamer to process stimuli that identify the external world and distinguish it from the internal dream experience, thereby producing the experience of 'knowing' that one is dreaming.

Personal Motives, Values, and Emotions

Night terrors are associated with the most violent emotions in sleep. These experiences occur upon sudden arousal from what is otherwise the deepest stage of NREM sleep (stage 4), and continue into a sustained hallucination of being attacked. The cry of terror and motor escape from the imagined attacker is preceded by only a few seconds of autonomic arousal, more intense than is seen in waking. The 'cause' is thought to be the loss of normal parasympathetic inhibition of the sympathetic system.

By contrast, emotions reported from REM sleep are more modest and tend not to awaken the sleeper or produce an autonomic response. We assume that the inhibitory afferent–efferent processes that protect sleep during the REM period extend to the autonomic system. Reported emotions provide us with an index of the motives that are active during sleep. This index suggests that emotions are stronger in REM than in NREM sleep, and that negative emotions are more prevalent than positive. Nevertheless, REM reports, especially in the early night, often describe without emotion events that in the waking state would elicit a strong emotional response.

The abstract nature of motives and the difficulty of determining whether a motive or emotion precedes a visual dream event, or is elicited in response to it, make the study of motivation in sleep very difficult. Fortunately, an increasing understanding of how the amygdala, medial prefrontal, and frontal cortices represent goal or motive structures in the waking state provides us with a neurological basis for exploring how these structures may participate in the production of sleep mentation.

Positron emission tomography (PET) studies show that the amygdala is unusually active in REM sleep, relative to NREM and waking, providing a neural basis for the negative emotions characteristic of that sleep stage. That the amygdala is associated with the experience of fear and pain in the waking state suggests that it plays a similar role in REM dreaming. Following the amygdala–fear model of LeDoux, we suggest that the imaginal events continuously generated by the cortex in REM sleep are also continuously evaluated by the amygdala, via their links to the medial prefrontal cortex, for their potential imagined threat to the dreamer. Positive emotions are reported less frequently than are negative emotions in REM mentation, but, with the exception of night terrors, both are reported less frequently in NREM mentation. We note that this description of affect in REM dreams is inconsistent with the long, vivid, emotional dreams that define most people's experience of their dreams. The latter tend to be produced while sleeping late on a weekend morning, and their neural generators are circadian and homeostatic processes that amplify the generation of mentation in REM and NREM sleep.

Analyzing brain-lesioned patients, Solms demonstrated that forebrain regions with limbic connections – which, in the waking state, participate in estimating the value of stimuli – are in sleep critical for dreaming in both REM and NREM sleep. Arguing against the activation–synthesis model, he finds that dreaming can persist even in the face of massive lesioning of the pontine brain stem mechanisms responsible for REM sleep. By contrast, dreaming can be entirely eliminated by forebrain lesions, particularly in the ventromesial quadrant of the frontal lobes, an area which connects frontal and limbic structures with dopaminergic cells in the ventral tegmentum. This forebrain circuit is often referred to as the 'seeking' or 'wanting' center of the human brain. Administration of L-dopa, stimulating this area, results in excessive and intense dreaming (in addition to psychotic symptoms). Conversely, inhibition of this circuit with haloperidol inhibits dreaming. Solms therefore hypothesizes that dreaming is controlled, not by the cholinergic brain

stem mechanisms described by Hobson but rather by this dopaminergic forebrain motivational system. In summary, the construction of sleep mentation is strongly influenced by neural processes that, in waking, determine the value of stimuli.

Bizarreness

There are two primary classes of bizarre mentation. The most common in laboratory reports are sudden changes in scene or location. The second, the juxtaposition of features, objects, or persons that would be improbable or impossible in the waking state, is the hallmark of the popular conception of dreaming, but is relatively infrequent in both REM and NREM sleep, especially in the early night. The neural basis for both classes is a matter of speculation. Certainly, the characteristics of attractors that make sense out of noisy input mitigate against bizarre patterns. Densely connected attractors in the same feature neighborhood tend to actively inhibit features that rarely occur together (i.e., they prevent bizarre juxtaposition). For example, in waking, a person cannot be recognized simultaneously as a boy and a girl. Yet one dreamer reported that "it was my brother," but "he was a girl" in the dream. We assume that the sense of knowing the relationship – "it was my brother" – was accomplished in an attractor field other than the field that saw his visual features as "a girl." Many dreamers do not recognize these incongruities until they awaken, at which time the normal communication between these cortical processing regions is restored. PET and transcranial magnetic stimulation studies have shown that cortical regions tend to be less effectively interconnected during sleep, as compared to wakefulness. Transcollosal communication may be especially impaired during REM sleep. We suggest that this lack of connectivity in sleep allows attractors that normally constrain one another in the waking state to produce bizarre combinations of visual objects, conversation, meanings, and responses, which would otherwise mutually inhibit one another.

The neural basis of sudden changes of topic or scene is even less well understood. We speculate that all thought, sleep or waking, is sustained by the value assigned to it by the amygdala and other limbic structures. We assume that the continuity that we, correctly or not, attribute to waking thought is due largely to the stability of our external environment. Lacking this external basis for continuity in sleep, each dream scenario simply fades away and is replaced by a new scene of significance to the dreamer.

Visual and Auditory Imagery

Visual imagery is much brighter and clearer in REM than in NREM periods (reported as being about 70% of the magnitude of brightness and clarity of waking perception). Although we assume that most of this imagery is generated in the occipital and right parietal cortex, heightened activity in extrastriate visual areas during REM sleep is accompanied by substantial decreases in blood flow to primary visual cortex. Sleepers who have been blind from birth report surprisingly normal dreams, which are apparently constructed from spatial coordinates learned in the waking state from auditory, tactile, and motor cues. Reported auditory imagery in REM sleep is also more frequent than in NREM, as are reports of conversation and the meanings of those conversations. Reports of tactile or pain imagery in sleep are extremely rare.

Total Reported Information

Given that REM reports exceed NREM on almost any variable on which they are compared, one must ask whether the cortical regions that generate mentation simply generate more information in REM sleep. With the exception of emotional intensity, after the effect of total reported information is partialled out, no other variable contributes to the qualitative difference in REM versus NREM mentation. REM reports, for example, tend to contain more bizarreness than NREM reports do, simply by virtue of their greater length. This suggests that aside from the more vivid visual imagery, which we attribute to right parietal cortex activation in REM sleep, the remaining characteristics of REM mentation may be attributed to a widespread increase in cortical activation. The active regions simply do what they were trained to do in the waking state, except that with high sensory thresholds, their processing sequences are not responsive to external stimuli, and decreased connectivity between regions results in less mutual constraint than is operative in waking perception. Many investigators have claimed, and continue to claim, that the unique pattern of cortical activation in REM must produce a unique pattern of mentation, but few have statistically corrected for the total amount of reported information, and some have even discarded reports with no content.

Circadian Influences on Sleep Mentation

Report length, and with it most characteristics of sleep mentation, tends to increase during the final hours of sleep. It is this longer, more dramatic imagery and emotion that people typically remember and identify as dreaming. Because REM periods also increase in length across the night, it has been generally assumed that these longer, more dramatic late-morning dreams are the result of a more active REM

process. However, increases in report length and dreamlike quality across the night occur in NREM as well as in REM sleep, correlated with the rising phase of a circadian cortical activation cycle, roughly correlated with the core body temperature rhythm. Propensity for REM sleep, which increases across the night, is controlled by the endogenous circadian clock, as is spindle frequency activity in stage 2 and alpha activity in REM sleep. Levels of the wake-promoting neuromodulator cortisol also rise as the sleeper moves toward morning wakefulness. Each of these activation-related physiological changes depends not on the number of hours of prior wakefulness or sleep, but rather on internal clocklike mechanisms. Although circadian-driven changes in cortical activation are difficult to measure directly, core body temperature is known to begin rising about 2 h before the time of normal waking. Inasmuch as circadian-driven cortical activation does not include the inhibition of sensory input characteristic of the REM generator, REM mentation produced during this late-morning interval may sometimes tend to be lucid in the sense described earlier. The size of the circadian effect on report length can be quite large, equaling the strength of the REM/NREM effect on propensity to report dreaming.

Joint and Unique Patterns of REM and Circadian-Based Mentation

That both REM and circadian generators are associated with increased mentation reporting and dreaming places us in a better position to ask what patterns of sleep mentation are unique to, or common to, the two generators. We note that the late-morning circadian increase in NREM (as well as REM) dreaming accounts, at least in part, for the often disputed source of dreaming in NREM sleep. Consistent with the heightened amygdala activation specific to REM sleep, a study from our lab recently found that circadian-driven activation amplified the emotional intensity of late-morning REM reports, but did not effect emotion in NREM sleep.

Bizarreness Revisited

That bizarreness increases with the active phases of both the circadian and the REM cycles suggests that bizarreness is not produced by an activation process unique to REM sleep and its pontine generator. We suggest that as cortical activation increases, more objects and events are produced, and if the regions that generate these objects and events are weakly interconnected, failing to mutually constrain one another, each will be regarded as improbable in the context of the other upon awakening.

Homeostatic-Driven Processes, Naps, and Dreaming

Circadian-driven cortical activation would be expected to increase steadily throughout the morning hours and into the afternoon. The joint effect of the circadian and REM/NREM generators should therefore produce intense dreaming at this time, compared to during the night. However, these activation effects appear to be opposed by the familiar, but poorly understood, 'sleep need' process. This homeostatic process, commonly measured by propensity for slow-wave activity in the EEG, accumulates as an exponentially saturating function of time awake, and then discharges during sleep according to an exponentially decaying function. It follows that even a short period of waking might suppress the activating effects of the rising phase of the circadian rhythm. Supporting this assumption, a recent study in our laboratory found that NREM reports from early-afternoon naps, when circadian activation is presumably high, yielded less total information than did late-morning NREM reports preceded by sustained sleep. Ultimately, the generators and suppressors of cortical activation responsible for mentation can only be isolated by further studies that employ forced desynchrony protocols in which individuals sleep at various phases of the circadian rhythm.

Sleep State Ambiguity and Mentation Features

Although the vernacular labels 'wake' and 'sleep' imply that the two states are discretely different, the father of sleep research, Nathaniel Kleitman, showed that biological characteristics of the two states can be dissociated. Inasmuch as the feature patterns of night terrors and lucid dreaming, discussed earlier, are incompatible with the defining characteristics of either waking or sleep, it would obfuscate our understanding of those phenomena if we were to force them into one category or the other. This dilemma raises the question of how much weight to give to an individual's own classification of his/her sleep/wake state when it conflicts with neurological and behavioral criteria. Sleep scientists traditionally distinguished sleep from waking by EEG criteria. However, these criteria were brought into question when it was found that many normal sleepers reported having been awake when interrupted from stage 2 NREM sleep during the first 2 h of the night. Interestingly, the latency of some of these 'awake' responses was sometimes over 30 s. One systematic study of this issue found that people use the subjective quality of their preinterruption mentation to make the judgment as to whether they were awake or asleep.

Thus, if mentation were an internal subvocal monolog common in the waking state, the stage 2 sleeper might say, "I was awake," whereas streams of vivid imagery common to REM dreaming would elicit the judgment of being asleep.

See also: Dream Function; Dreams and Nightmares in PTSD; Dreams and Dreaming: Incorporation of Waking Events; Dreams, Dreaming Theories and Correlates of Nightmares; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Research and Sleep Medicine in Historical Perspective; Sleep: Development and Circadian Control; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Antrobus J, Kondo T, and Reinsel R (1995) Dreaming in the late morning: Summation of REM and diurnal cortical activation. *Consciousness and Cognition* 4: 275–299.

Aserinsky E and Kleitman N (1953) Regularly occurring periods of eye motility and concomitant phenomenon during sleep. *Science* 188: 273–274.

Bertini M, De Gennaro L, Ferrara M, et al. (2004) Reduction of transcallosal inhibition upon awakening from REM sleep in humans as assessed by transcranial magnetic stimulation. *Sleep* 27: 875–882.

Braun AR, Balkin TJ, Wesenten NJ, et al. (1997) Regional cerebral blood flow throughout the sleep–wake cycle: An H_2O-O-15 PET study. *Brain* 120: 1173–1197.

Dement W and Wolpert EA (1958) The relation of eye movements, body motility, and external stimuli to dream content. *Journal of Experimental Psychology* 55: 543–53.

Ellman S, and Antrobus J (eds.) (1991) *The Mind in Sleep*, 2nd edn. New York: Wiley.

Hobson J and McCarley R (1977) The brain as a dream state generator: An activation-synthesis hypothesis of the dream process. *American Journal of Psychiatry* 134: 1335–1348.

Hobson JA, Pace-Schott E, and Stickgold R (2000) Dreaming and the brain: Towards a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences* 23: 793–842.

Maquet P, Ruby P, Maudoux A, et al. (2005) Human cognition during REM sleep and the activity profile within frontal and parietal cortices: A reappraisal of functional neuroimaging data. *Progress in Brain Research* 150: 219–227.

Massimini M, Ferrarelli F, Huber R, et al. (2005) Breakdown of cortical effective connectivity during sleep. *Science* 309: 2228–2232.

Moruzzi G and Magoun HW (1949) Brain stem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology* 1: 455–473.

Nielsen TA (2000) A review of mentation in REM and NREM sleep: "Covert" REM sleep as a possible reconciliation of two opposing models. *Behavioral and Brain Sciences* 23: 793–1121.

Pivik T and Foulkes D (1968) NREM mentation: Relation to personality, orientation time, and time of night. *Journal of Consulting and Clinical Psychology* 32: 144–151.

Smith M, Antrobus J, Gordon E, et al. (2004) Motivation and affect in REM sleep and the mentation reporting process. *Consciousness and Cognition* 13: 501–511.

Solms M (2000) Dreaming and REM sleep are controlled by different mechanisms. *Behavioral and Brain Sciences* 23: 793–850.

Wamsley EJ, Hirota Y, Tucker MA, et al. (2007) Circadian and ultradian influences on dreaming: A dual rhythm model. *Brain Research Bulletin* 47(4): 347–354.

Sleep Oscillations

A Destexhe, Centre National de la Recherche Scientifique, Gif-sur-Yvette, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The discovery that the brain oscillates during sleep is almost as old as the discovery of the electroencephalogram (EEG). The first human EEG recordings reported oscillations, the type, frequency, and amplitude of which highly depend on behavioral state. In an alert, awake subject, the EEG is dominated by low-amplitude fast activity ('desynchronized EEG') with high-frequency oscillations (beta, gamma), whereas during slow-wave sleep, the EEG shifts to large-amplitude, slow oscillations. The early stage of slow-wave sleep is associated with EEG spindle waves, which occur at a frequency of 7–14 Hz. As sleep deepens, waves with slower frequencies (0.3–4 Hz), including delta waves and slow oscillations, appear and progressively dominate the EEG. During paradoxical sleep, also called rapid eye movement sleep, EEG activities resemble those of wakefulness.

The cellular bases of slow-wave sleep oscillations have been investigated since the first extracellular and intracellular recordings in mammals. The major brain regions identified are the thalamus and cerebral cortex, which are intimately linked by means of reciprocal projections. The activities of thalamic and cortical neurons during sleep have been largely documented by electrophysiological studies. The cellular mechanisms underlying these oscillations depend on many factors, such as the connectivity and intrinsic properties of the different types of thalamic and cortical neurons. To determine these cellular mechanisms, it is necessary to use computational models, which are based on experimental data, and to suggest mechanisms and, if possible, predictions to test them. This type of interaction has been quite successful in the (still ongoing) exploration of the mechanisms of sleep oscillations, which this article attempts to summarize.

Sleep Spindle Oscillations

Sleep spindles are found during the early stages of sleep (sleep stage 2 in humans) and constitute an electrographic landmark for the transition from waking to sleep. Spindle oscillations consist of 7–14 Hz waxing and waning potentials, grouped in sequences lasting 1–3 s and recurring every 3–10 s. Spindle oscillations constitute an interesting and well-constrained problem to investigate by computational models for several reasons. First, these oscillations are generated in the thalamus, which is a well-known structure anatomically, with well-defined connectivity between the different cell types (see circuit in Figure 1(a)). Second, spindles are remarkably well documented experimentally and have been extensively characterized both *in vivo* and *in vitro*. Third, this oscillation is generated by an interplay of complex cellular properties (schematized in Figure 1(b), left), such as burst firing, and synaptic interactions via multiple types of postsynaptic receptors (see Figure 1(b), right). Computational models are needed to understand this complex interplay.

The typical electrophysiological features of spindle oscillations are shown in Figure 1(c). The two cell types involved, thalamocortical (TC) and thalamic reticular (RE) neurons, oscillate synchronously and display burst discharges according to a mirror image: RE cells display bursts following excitatory postsynaptic potentials while TC cells burst following inhibitory postsynaptic potentials. While RE cells tend to burst at every cycle of the oscillation, TC cells produce bursts only once every few cycles. These features are typical of spindles recorded in thalamic neurons in different mammals.

Several hypotheses for the genesis of oscillations by thalamic circuits have been proposed. These involve reciprocal synaptic interactions between TC neurons and local inhibitory interneurons, loops between TC and RE neurons, or loops within the RE nucleus. The involvement of the RE nucleus was firmly demonstrated in a series of experiments by Steriade's group. In particular, the deafferented RE nucleus *in vivo* can exhibit spindle rhythmicity in extracellular recordings. In contrast, the RE nucleus does not display autonomous oscillations *in vitro*, but spindles have been observed in thalamic slices based on TC–RE interactions. These *in vitro* spindles display the same intracellular features as *in vivo* (Figure 1(c)).

To attempt clarifying these contrasting results, the genesis of spindle oscillations was investigated with computational models. First, models investigated whether the RE nucleus is capable of displaying oscillations consistent with experiments. Models found that RE neurons interacting through γ-aminobutyricacidergic (GABAergic) synapses can generate spindle rhythmicity and suggested different mechanisms. GABAergic interactions between RE neurons can make them oscillate synchronously through mutual inhibitory rebound interactions, either with slow GABAergic synapses or with fast ($GABA_A$ receptor-mediated) GABAergic synapses with extended connectivity. Synchronized oscillations can also be generated

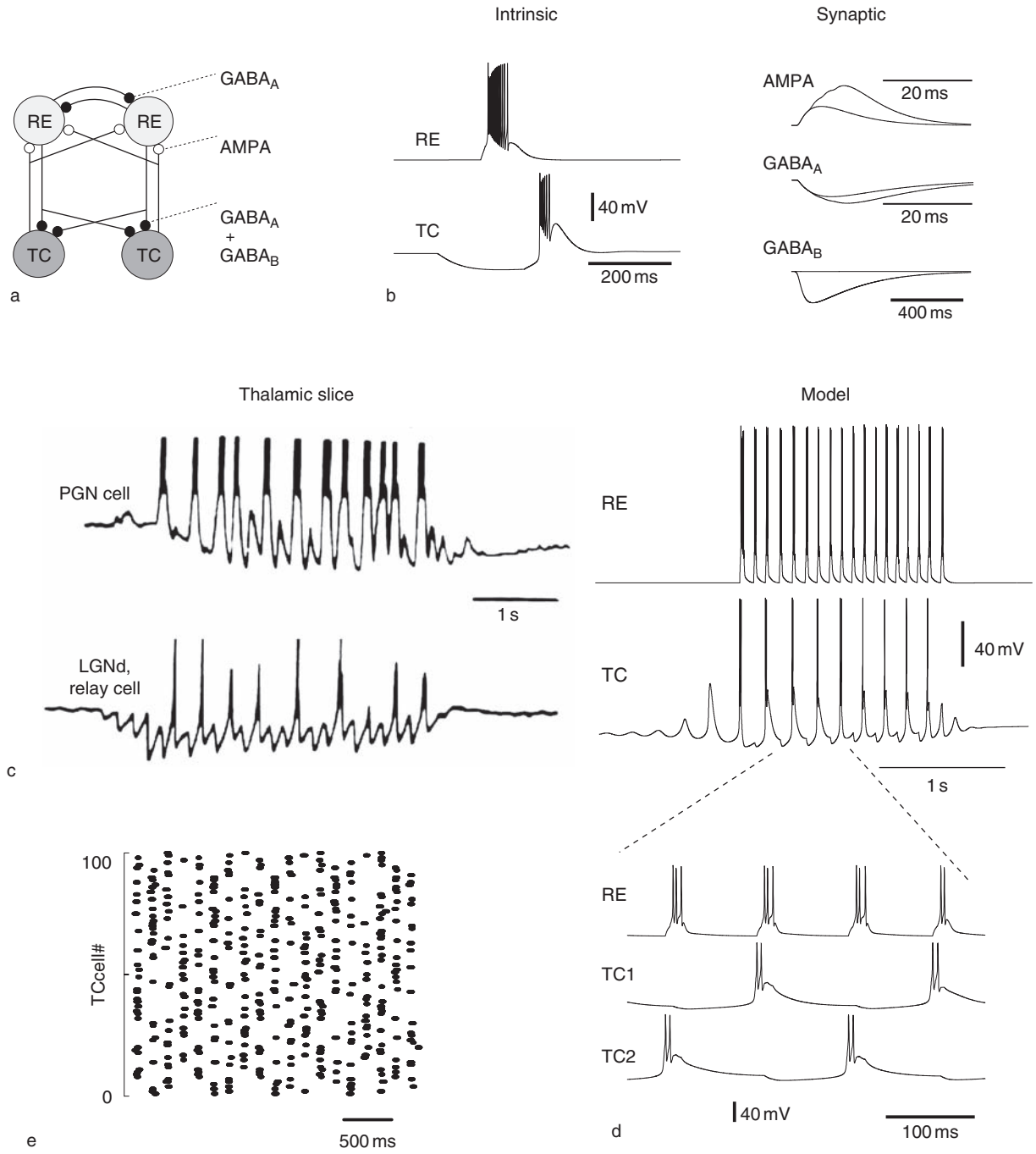

Figure 1 Modeling the interactions between intrinsic and synaptic properties to generate spindle oscillations. (a) Circuit of interconnected thalamocortical (TC) and thalamic reticular (RE) neurons with different receptor types. (b) Models of the intrinsic properties of thalamic neurons (left) and of the synaptic receptor types (right) mediating their interactions. (c) *In vitro* recordings of TC (LGNd) and RE (PGN) neurons from the visual thalamus during spindle oscillations. (d) Computational model of spindle oscillations in circuits of interconnected TC and RE cells. The expanded trace below shows the phase relations of the two cell types. (e) Phase relations of TC cells during spindle oscillations in a different computational model. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; GABA$_A$, γ-aminobutyric acid class A receptor; LGN, lateral geniculate nucleus; LGNd, dorsal lateral geniculate nucleus; PGN, perigeniculate nucleus. (a, b, d) From Destexhe A, Bal T, McCormick DA, and Sejnowski TJ (1996) Ionic mechanisms underlying synchronized oscillations and propagating waves in a model of ferret thalamic slices. *Journal of Neurophysiology* 76: 2049–2070. (c) From von Krosigk M, Bal T, and McCormick DA (1993) Cellular mechanisms of a synchronized oscillation in the thalamus. *Science* 261: 361–364. (e) From Wang XJ, Golomb D, and Rinzel J (1995) Emergent spindle oscillations and intermittent burst firing in a thalamic model: Specific neuronal mechanisms. *Proceedings of the National Academy of Sciences of the United States of America* 92: 5577–5581.

from RE neurons connected with depolarizing GABA-ergic synapses. The presence of gap junctions in the reticular nucleus was also incorporated in models and also reinforced its propensity to oscillate in the spindle frequency range. Thus, models suggest that the intrinsic properties of thalamic RE neurons, combined with their synaptic or electrical interactions, support the RE pacemaker hypothesis.

Second, models including TC and RE cells showed that spindle oscillations can also be obtained from TC–RE loops. This TC–RE loop model is shown in Figure 1(d). Neurons were modeled using Hodgkin–Huxley type representations of Na^+, K^+, and Ca^{2+} voltage-dependent currents, which were based on voltage-clamp data on thalamic neurons. These models reproduced the most salient intrinsic properties of thalamic neurons, such as the production of bursts of action potentials (Figure 1(b)). Synaptic interactions were modeled with conductance-based kinetic models, which were used to simulate the main receptor types (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), $GABA_A$, and $GABA_B$) identified in thalamic circuits (Figure 1(b), right). Under these conditions, the circuit generated 7–14 Hz spindle oscillations with the typical features identified intracellularly in the different thalamic neuron types. The model reproduced the typical mirror image between TC and RE cells during spindles, as well as the phase relations between cells (see Figure 1(d)). In particular, TC cells produced bursts once every two or three cycles, a feature consistently observed experimentally (compare with Figure 1(c)). More-irregular behavior, similar to the experiments, was obtained in larger networks (Figure 1(e)) or in the presence of the cortex (see the section titled 'Propagation or large-scale synchrony'). The oscillations also showed the typical waxing-and-waning envelope of spindles; this property was due in the model to Ca^{2+}-mediated slow regulation of the hyperpolarization-activated current, a prediction that was verified experimentally.

Thus, models show that the complex bursting properties of thalamic neurons, combined with their interactions through well-defined types of synaptic receptors, account for both RE pacemaker oscillations and spindle oscillations arising from TC–RE loops without assuming pacemaker activity in the RE nucleus. How may such an apparent inconsistency be resolved? This question was addressed by a computational model of the RE nucleus which took into account the action of neuromodulators (such as noradrenaline) in depolarizing RE cells. This model produced oscillations only when a sufficient level of neuromodulator was present. The difference between *in vivo* and *in vitro* preparations may therefore be explained by the limited connectivity between the RE neurons in the slice and/or by the fact that slices lack the necessary level of neuromodulation to maintain isolated RE oscillations. The main prediction from this model is that applying neuromodulators to slices of the RE nucleus should induce oscillations similar to those observed *in vivo*. This prediction still awaits testing.

Propagation or Large-Scale Synchrony?

How spindle oscillations are initiated and distributed in large circuits was investigated with multiple recordings *in vivo* and *in vitro*. Spindle oscillations *in vitro* showed traveling wave patterns, with the oscillation typically starting on one side of the slice and propagating to the other side, at a constant propagation velocity. Traveling spindle waves were simulated by computational models of networks of interconnected TC and RE cells (one-dimensional extensions of the circuit shown in Figure 1), in two independent modeling studies. These models were similar in spirit to the circuit shown in Figure 1(a) but assumed that there was a topographic connectivity between TC and RE layers, consistent with anatomical data. Under these conditions, the models generated traveling waves consistent with *in vitro* data.

However, in contrast to thalamic slices, the intact TC system *in vivo* does not display clear-cut propagation, but spindle oscillations are synchronized over extended thalamic regions and show little signs for traveling wave activity (Figure 2(a)). This large-scale synchrony was lost when the cortex was removed, suggesting that although the oscillation is generated by the thalamus, its synchrony depends on the cortex. To investigate the mechanism underlying large-scale synchrony, a TC network model was developed by combining the previous model of thalamic slices with a model of deep cortical layers. The principal prediction of this model was that, in order to generate large-scale coherent oscillations, the powerful action of cortex on thalamus (via corticothalamic fibers) must be included, and most important, the cortex had to recruit the thalamus primarily through the RE nucleus. Because of the powerful inhibitory action of RE cells, the action of corticothalamic input is 'inhibitory dominant' on TC cells, a property essential to maintain large-scale synchrony by synchronizing the rebound bursts of TC cells. This property of inhibitory dominance is generally observed intracellularly in thalamic neurons when the cortex is stimulated. In these conditions, the same model was capable of generating large-scale synchrony in the presence of the cortex, as well as traveling wave activity in the isolated thalamus (Figure 2(b)). Consistent with these models, propagating activity has indeed been observed in the thalamus of decorticated cats *in vivo*.

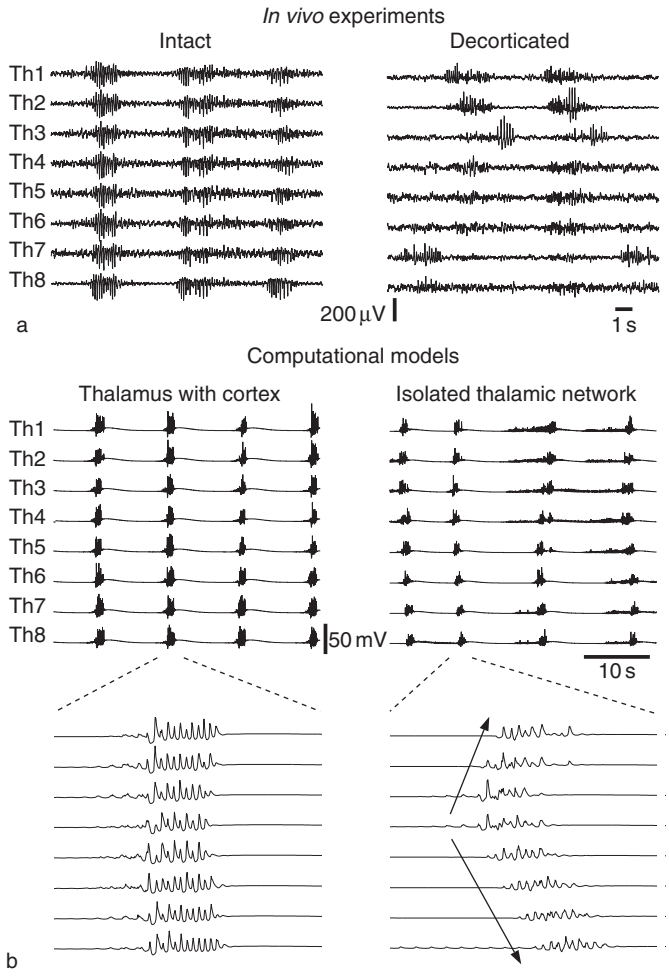

Figure 2 Modeling the large-scale synchrony of spindle oscillations. (a) Multiple extracellular field potential recordings (Th1 . . . Th8) in the thalamus of cats *in vivo*. In the intact thalamocortical system (Intact), the oscillations were synchronized over large distances (7 mm between Th1 and Th8). After removal of the cortex (Decorticated), the large-scale synchrony was largely abolished. (b) Computational models of thalamocortical networks displaying large-scale synchrony when cortex was present (left) and local synchrony with traveling-wave activity in the isolated thalamus (right). (a) Adapted from Contreras D, Destexhe A, Sejnowski TJ, and Steriade M (1996) Control of spatiotemporal coherence of a thalamic oscillation by corticothalamic feedback. *Science* 274: 771–774. (b) Adapted from Destexhe A, Contreras D, and Steriade M (1998) Mechanisms underlying the synchronizing action of corticothalamic feedback through inhibition of thalamic relay cells. *Journal of Neurophysiology* 79: 999–1016.

The cortical control of thalamic relay cells through dominant inhibitory mechanisms has important consequences, not only for explaining large-scale synchrony, but also for explaining pathological situations such as absence epileptic seizures. As a result of inhibitory dominance, a too strong corticothalamic feedback can overactivate thalamic $GABA_B$ receptors and entrain the physiologically intact thalamus into hypersynchronous rhythms at ~3 Hz. This scheme may explain the genesis of absence seizures, which are hypersynchronous ~3 Hz rhythms that appear suddenly in the TC system. These seizures can be provoked experimentally by increasing cortical excitability in a physiologically intact thalamus. The TC model accounts for those experiments and can simulate seizures based on inhibitory-dominant corticothalamic feedback. This model directly predicted that manipulating corticothalamic feedback should entrain intact thalamic circuits to generate hypersynchronous rhythms at ~3 Hz, a prediction verified by two independent studies. A similar mechanism, with a different balance between $GABA_A$ and $GABA_B$ receptors, can also generate faster hypersynchronous rhythms (around 5–10 Hz), as observed in rat or mouse experimental models of absence seizures.

Slow-Wave Oscillations

During the deepest phases of sleep (stages 3 and 4 in humans), as well as for some anesthetized states, cortical activity is dominated by delta and slow oscillations, in a frequency range of 0.3–4 Hz. The intracellular

correlate of these slow waves is the alternation between depolarized states (up states) and hyperpolarized states (down states), which occurs in perfect synchrony with the EEG (Figure 3(c)). Thus, entire cortical regions are simultaneously switching between up and down states, as also shown by multiple extracellular studies. The origin of these oscillations seems to be cortical because they survive extensive thalamic lesions, and they are also observed in cortical slices.

One of the interesting features of up states is that this activity is very similar to that during the awake state. This is supported by several observations. First, during the up state, the EEG is of low amplitude, and fast activity is similar to desynchronized EEG. Second, extracellular recordings have shown that the up states obey the same dynamics of firing, have similar local correlations, and display similar relations between EEG and unit firing as wakefulness does. Third, simulating nuclei participating in the ascending arousal system induces periods of desynchronized EEG corresponding intracellularly to prolonged up states. Fourth, conductance measurements from intracellular recordings in anesthetized or EEG-activated states show similar conductance patterns during both states. The absolute conductance is lower in activated states than in up states, but both states are characterized by similar ratios between excitatory and inhibitory conductances. Fifth, computational models of up/down states and of activated states in cortical circuits suggest that both can be generated by similar mechanisms.

Computational models have investigated the genesis of slow-wave oscillations and the associated up- or down-state patterns. These models were based on recurrent circuits of excitatory and inhibitory cortical neurons (Figure 3(a)). The two main electrophysiological types of cortical neurons were considered, as well as their synaptic interactions through glutamate (AMPA) and GABAergic ($GABA_A$) receptors (Figure 3(b)). These models showed that up states can be generated by recurrent excitatory and inhibitory connections, which self-sustain the activity (Figure 3(d)). Different exact mechanisms by which up states begin and terminate have been proposed. Up states can start either by the interaction between subthreshold Na currents (persistent Na current) and miniature excitatory synaptic potentials. Another possible mechanism is spontaneously active cells that would initiate the wave of activity in the network. In either case, none of the mechanisms has been verified experimentally. The termination of the up state is apparently due to a progressive run down of synaptic activity, as indicated by conductance measurements. What causes this run down could be either an intrinsic property, such as the progressive buildup of a slow potassium conductance, or simply depression of excitatory synapses. Both hypotheses are supported by experimental data and are also consistent with the refractoriness of the up states found in slices (this refractoriness could be due to the potassium conductance, or recovery from synaptic depression).

Another property of up/down states is that the duration of the down state is proportional to network size. Down states are typically short *in vivo* (a few hundred milliseconds) whereas they can last up to 20 s in slices. Cutting cortical slabs of different sizes *in vivo* confirmed that the down-state duration varies inversely in proportion to slab size. Here again, this property is consistent with the two mechanisms of initiation outlined above as they both depend on coincident activation of either miniature or spontaneously active cells, both of which will occur more often in large networks.

Models were also used to simulate the transition to the sustained and irregular firing activity during wakefulness. Self-sustained irregular states similar to activated states have been simulated by various models. Only a few models, however, provided the transition from up/down states to activated states. For all such models, up states and activated states are very similar and differ only by the level of excitability of the neurons (mostly by downregulating potassium conductances). Some of these models were confronted to input resistance or conductance measurements and reproduced qualitatively the values measured experimentally. The fact that up states and activated states can be simulated with few differences by the same models is another indication that those two states stem from similar network activity.

A final property of slow waves is that the up states clearly show propagating properties *in vitro*. This propagation can be reproduced by computational models (Figure 3(e)), assuming that synaptic connections are made locally in the cortical network. In contrast, there is evidence that up states are highly synchronized *in vivo*, because the local EEG is always phase locked with intracellular activity (Figure 3(a)). Multiple extracellular recordings in natural sleep also demonstrated that the up states of slow waves are highly synchronized across distances up to 7 mm in cortex. A detailed investigation of this synchrony, and the conditions under which up states can propagate *in vivo*, as well as the modeling of such phenomena, still await.

What Is the Role of Slow-Wave Sleep Oscillations?

As mentioned above, the up states during slow waves share many different features of the sustained activity during wakefulness, and thus, up states can be viewed

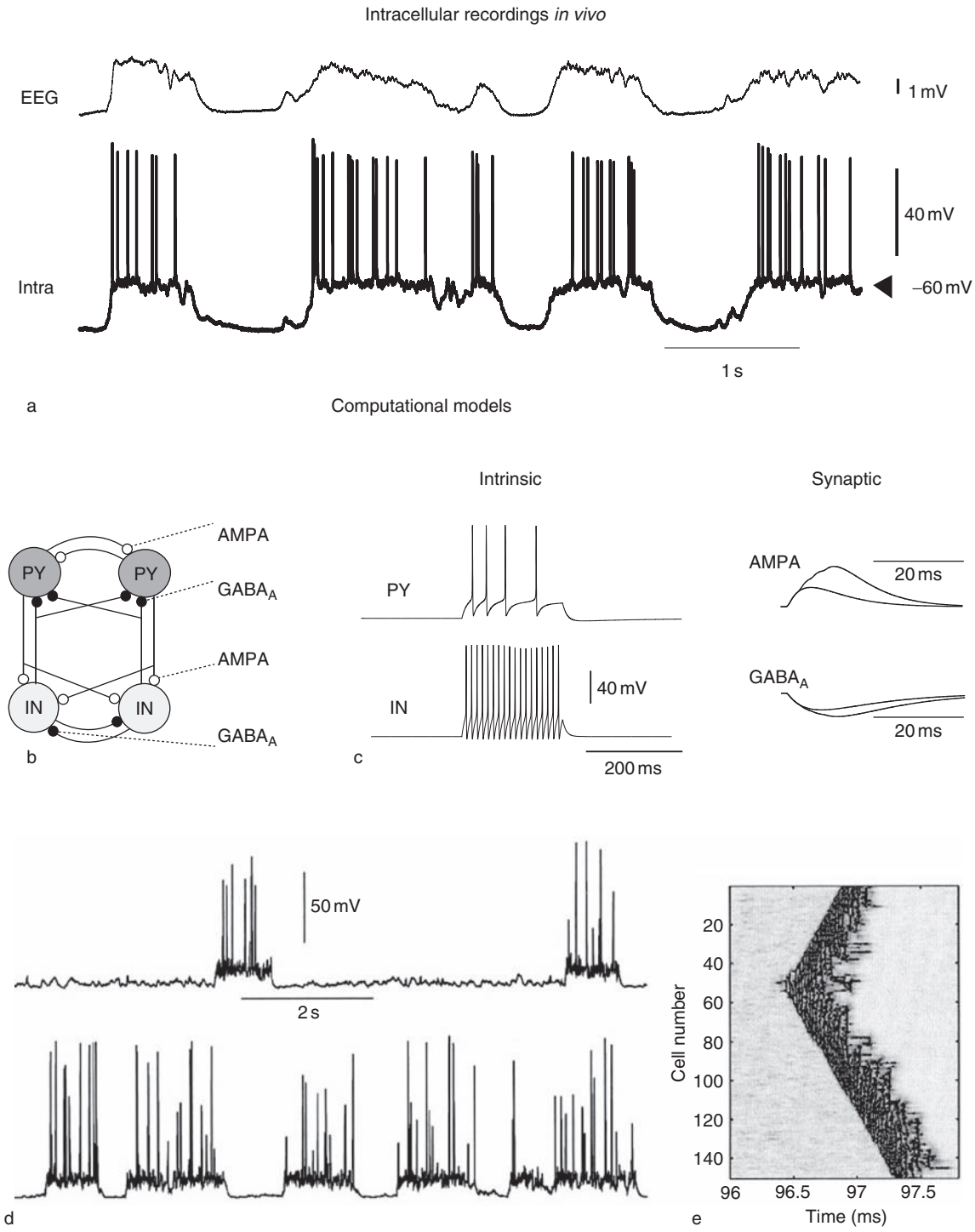

Figure 3 Computational models of slow-wave oscillations in cerebral cortex. (a) *In vivo* recordings of a morphologically identified pyramidal neuron during slow waves. (b) Schematic circuit showing the two main types of cortical neurons, pyramidal cells (PY) and inhibitory interneurons (IN). Those neurons are connected via different types of synaptic receptors, with the two main types illustrated here. (c) Models of the intrinsic properties of cortical neurons (left) and of the synaptic receptor types (right) mediating their interactions. (d) Computational model of slow-wave oscillations arising from reverberation of activity through recurrent connections in networks of cortical circuits. The network displays up and down states with different frequency of occurrence, depending on the level of spontaneous activity. (e) Snapshot of activity in the network, showing the initiation and propagation of the up state. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid; GABA$_A$, γ-aminobutyric acid class A receptor. (a) Adapted from Rudolph M, Pelletier J-G, Paré D, and Destexhe A (2005) Characterization of synaptic conductances and integrative properties during electrically-induced EEG-activated states in neocortical neurons *in vivo*. *Journal of Neurophysiology* 94: 2805–2821. (d, e) Adapted from Timofeev I, Grenier F, Bazhenov M, Sejnowski TJ, and Steriade M (2000) Origin of slow cortical oscillations in deafferented cortical slabs. *Cerebral Cortex* 10: 1185–1199.

as brief periods of activity in which network dynamics are very similar to the dynamics during wakefulness. This is consistent with the fact that up states would represent 'replayed' events that have occurred previously during the wake state. There is abundant experimental evidence, from birds to higher mammals, for such a replay during sleep.

Why would such a replay occur during sleep and not during wakefulness? There is presently no clear answer to this question, but it was proposed that another type of sleep oscillation, sleep spindles, may be implicated in gating of long-term plasticity mechanisms (for example, permanent changes that require protein synthesis). Computational models of reconstructed pyramidal neurons, combined with intracellular measurements of thalamic inputs during spindles, have shown that the pattern of synaptic bombardment during spindles in the cortex is a strong excitatory-driven depolarization in dendrites, with an equally strong inhibition around the soma, preventing the cell from excessive firing. This pattern is likely to induce massive calcium entry at around 10 Hz in the dendrites, which is an ideal signal to activate molecular gates, such as protein kinase A. Sleep spindles would therefore provide a physiological signal similar to the repeated tetanus used to induce long-term synaptic changes in slices. However, instead of inducing potentiation directly, spindles may in fact provide a 'priming' signal, opening a gate that allows permanent changes to subsequent inputs ('replayed' events mentioned earlier) following the sleep spindles.

Taking those two observations together leads to the following scenario for how these different mechanisms may contribute to memory consolidation during sleep. During conscious experience, latent memories are formed throughout the cortex, together with links to the hippocampal formation that allow top-down retrieval to occur. During the early stages of sleep, spindle oscillations would mobilize the molecular machinery needed for memory consolidation in cortex. In the deeper phases of slow-wave sleep, during the brief periods of wakelike activities (up states), the hippocampal formation would activate latent memories stored in the neocortex ('replay') and induce permanent changes in intrinsic or synaptic conductances. This hypothetical mechanism of memory consolidation during sleep is consistent with all electrophysiological characteristics of sleep oscillations, and it predicts that special correlations between hippocampal and cortical activities should occur during the up states of slow waves. Such correlations have been found recently between cortical slow waves (up states) and hippocampal sharp waves. No computational model has been proposed to date to explain these observations, and certainly they constitute one of the most exciting directions to pursue toward exploring the role of sleep oscillations.

See also: Electroencephalography (EEG); Electrophysiology: EEG and ERP Analysis; Epilepsy; Executive Function and Higher-Order Cognition: EEG Studies; Gap Junctions and Neuronal Oscillations; Sleep and Sleep States: Thalamic Regulation; Sleep Oscillations and PGO Waves; Sleep-Dependent Memory Processing.

Further Reading

Bal T, Debay D, and Destexhe A (2000) Cortical feedback controls the frequency and synchrony of oscillations in the visual thalamus. *Journal of Neuroscience* 20: 7478–7488.

Battaglia FP, Sutherland GR, and McNaughton BL (2004) Hippocampal sharp wave bursts coincide with neocortical "up-state" transitions. *Learning & Memory* 11: 697–704.

Bazhenov M, Timofeev I, Steriade M, and Sejnowski TJ (2002) Model of thalamocortical slowwave sleep oscillations and transitions to activated states. *Journal of Neuroscience* 22: 8691–8704.

Blumenfeld H and McCormick DA (2000) Corticothalamic inputs control the pattern of activity generated in thalamocortical networks. *Journal of Neuroscience* 20: 5153–5162.

Compte A, Sanchez-Vives MV, McCormick DA, and Wang XJ (2003) Cellular and network mechanisms of slow oscillatory activity (<1 Hz) and wave propagations in a cortical network model. *Journal of Neurophysiology* 89: 2707–2725.

Contreras D, Destexhe A, Sejnowski TJ, and Steriade M (1996) Control of spatiotemporal coherence of a thalamic oscillation by corticothalamic feedback. *Science* 274: 771–774.

Contreras D, Timofeev I, and Steriade M (1996) Mechanisms of long lasting hyperpolarizations underlying slow sleep oscillations in cat corticothalamic networks. *Journal of Physiology* 494: 251–264.

Destexhe A (1998) Spike-and-wave oscillations based on the properties of $GABA_B$ receptors. *Journal of Neuroscience* 18: 9099–9111.

Destexhe A, Bal T, McCormick DA, and Sejnowski TJ (1996) Ionic mechanisms underlying synchronized oscillations and propagating waves in a model of ferret thalamic slices. *Journal of Neurophysiology* 76: 2049–2070.

Destexhe A, Contreras D, Sejnowski TJ, and Steriade M (1994) A model of spindle rhythmicity in the isolated thalamic reticular nucleus. *Journal of Neurophysiology* 72: 803–818.

Destexhe A, Contreras D, Sejnowski TJ, and Steriade M (1994) Modeling the control of reticular thalamic oscillations by neuromodulators. *NeuroReport* 5: 2217–2220.

Destexhe A, Contreras D, and Steriade M (1998) Mechanisms underlying the synchronizing action of corticothalamic feedback through inhibition of thalamic relay cells. *Journal of Neurophysiology* 79: 999–1016.

Destexhe A, Contreras D, and Steriade M (1999) Spatiotemporal analysis of local field potentials and unit discharges in cat cerebral cortex during natural wake and sleep states. *Journal of Neuroscience* 19: 4595–4608.

Destexhe A and Sejnowski TJ (2001) *Thalamocortical Assemblies*. Oxford, UK: Oxford University Press.

Destexhe A and Sejnowski TJ (2003) Interactions between membrane conductances underlying thalamocortical slow-wave oscillations. *Physiological Reviews* 83: 1401–1453.

Gloor P and Fariello RG (1988) Generalized epilepsy: Some of its cellular mechanisms differ from those of focal epilepsy. *Trends in Neurosciences* 11: 63–68.

Hodgkin AL and Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* 117: 500–544.

Kim U, Bal T, and McCormick DA (1995) Spindle waves are propagating synchronized oscillations in the ferret LGNd *in vitro*. *Journal of Neurophysiology* 74: 1301–1323.

Landisman CE, Long MA, Beierlein M, Deans MR, Paul DL, and Connors BW (2002) Electrical synapses in the thalamic reticular nucleus. *Journal of Neuroscience* 22: 1002–1009.

Ribeiro S, Gervasoni D, Soares ES, et al. (2004) Long-lasting novelty-induced neuronal reverberation during slow-wave sleep in multiple forebrain areas. *PLoS Biology* 2: 126–137.

Rudolph M, Pelletier J-G, Paré D, and Destexhe A (2005) Characterization of synaptic conductances and integrative properties during electrically-induced EEG-activated states in neocortical neurons in vivo. *Journal of Neurophysiology* 94: 2805–2821.

Sanchez-Vives MV and McCormick DA (2000) Cellular and network mechanisms of rhythmic recurrent activity in neocortex. *Nature Neuroscience* 3: 1027–1034.

Sirota A, Csicsvari J, Buhl D, and Buzsaki G (2003) Communication between neocortex and hippocampus during sleep in rodents. *Proceedings of the National Academy of Sciences of the United States of America* 100: 2065–2069.

Steriade M (2003) *Neuronal Substrates of Sleep and Epilepsy*. Cambridge, UK: Cambridge University Press.

Steriade M, McCormick DA, and Sejnowski TJ (1993) Thalamocortical oscillations in the sleeping and aroused brain. *Science* 262: 679–685.

Timofeev I, Grenier F, Bazhenov M, Sejnowski TJ, and Steriade M (2000) Origin of slow cortical oscillations in deafferented cortical slabs. *Cerebral Cortex* 10: 1185–1199.

von Krosigk M, Bal T, and McCormick DA (1993) Cellular mechanisms of a synchronized oscillation in the thalamus. *Science* 261: 361–364.

Wang XJ, Golomb D, and Rinzel J (1995) Emergent spindle oscillations and intermittent burst firing in a thalamic model: Specific neuronal mechanisms. *Proceedings of the National Academy of Sciences of the United States of America* 92: 5577–5581.

Wang XJ and Rinzel J (1993) Spindle rhythmicity in the reticularis thalami nucleus: Synchronization among inhibitory neurons. *Neuroscience* 53: 899–904.

Relevant Website

http://senselab.med.yale.edu – ModelDB, database of freely available model codes, some of which simulate sleep oscillations.

Sleep Oscillations and PGO Waves

M Steriade, Université Laval, Laval, QC, Canada

© 2009 Published by Elsevier Ltd.

Behavioral and Electrographic Signs of Sleep Stages

Sleep Onset

The popular view that behavioral quiescence is the sign that heralds sleep may be valid for the full-blown state of sleep but not for the preparatory period during which many animal species display complex motor behaviors directed to find a safe home for sleep. The defining signs of the period when one falls asleep are changes in brain electrical activity (electroencephalogram (EEG)) that are associated with long periods of inhibition in neurons located in the thalamus, a deep structure in the brain and a gateway for most sensory signals in their route to the cerebral cortex. The consequence is that the incoming messages are blocked in the thalamus and the cerebral cortex is deprived of information from the outside world.

In contrast with the rapid awakening from sleep, the period of falling asleep is quite long. Thus, sleep onset may not be exclusively attributable to neuronal mechanisms, which operate on short timescales. It is conceivable that the mechanisms of falling sleep will only be revealed with concerted studies on sleep humoral factors (acting on longer timescale) and on their actions on neurons in critical brain areas. Currently, little is known about the effects of sleep-promoting chemical substances on neurons. The list of putative sleep humoral factors is quite long. Some of them are (1) adenosine, which exerts a tonic inhibitory control on both upper brain stem and basal forebrain cholinergic activating (awakening) neurons; (2) muramyl peptides, which have a chemical structure that is similar to that of serotonin, a neurotransmitter that was implicated in the generation of early sleep stages; and (3) prostaglandin D_2, whose receptors are found in the anterior hypothalamus, thought to be a critical zone for promoting sleep.

The concepts postulating that sleep is a passive phenomenon due to closure of cerebral gates leading to brain deafferentation or, alternatively, an active phenomenon promoted by inhibitory mechanisms arising in some cerebral areas have long been considered as opposing views. However, neurons with inhibitory influences (located in the anterior hypothalamus) act on neurons that exert excitatory influences on the brain (located in the posterior hypothalamus and upper brain stem reticular core), and the final outcome of these relations is the disconnection of the forebrain. Therefore, the two mechanisms (active and passive) of sleep onset are probably successive steps within a chain of events, and they are complementary rather than opposed.

Slow-Wave and Rapid Eye Movement Sleep

Sleep consists of an alternation between two distinct stages. The early stage is characterized by large-scale synchronization of low-frequency EEG activity (0.5–15 Hz), which is generated by the cerebral cortex and thalamus. Synchronization is a state in which two or more oscillators display the same frequency because of mutual influences. The notion of synchronization supposes the coactivation as well as concerted inhibitory processes in a large number of neurons. The summation of neuronal activities is sufficiently large to be recorded with gross electrodes, such as scalp electrodes when recording the human EEG. Because of the prevalent low-frequency waves, this stage is termed slow-wave sleep. This electrophysiological feature is associated with behavioral quiescence and suspension of conscious processes related to the external world because the transmission of signals is inhibited in the thalamus, even before these signals reach the cerebral cortex, the highest site of consciousness. However, not all mental processes are obliterated during slow-wave sleep because the brain oscillations that occur during this stage give rise to synaptic plasticity that underlies processes through which information is stored, such as consolidation of memory traces acquired during the state of wakefulness. Also, a peculiar form of dreaming mentation is present during slow-wave sleep.

The second stage of sleep is associated with rapid eye movements (REMs) and the sign that distinguishes it from the other two states of vigilance (slow-wave sleep and waking) is the largely suppressed motor output (muscular atonia) due to inhibition of spinal motor neurons. In addition, spiky potentials (ponto-geniculo-occipital (PGO) waves), related to ocular saccades, can be recorded from the thalamus and cortex and are regarded as the physiological correlate of dreaming. This state was also termed active because the level of brain alertness is similar to, or even higher than, the state of wakefulness. (This should not imply, however, that slow-wave sleep is inactive from the standpoint of brain functions since neuronal recordings have shown that cortical neurons fire during this stage at almost the same level as during waking.) Thus, REM sleep is a brain-active state associated with paralysis of limbs' muscles. In other words, motor commands from a highly excitable cerebral

cortex cannot be executed because of the inhibition of motor neurons in the spinal cord. Although REM sleep seems very similar to wakefulness because of intense brain activity, some differences also exist between these two states, mainly due to more effective inhibitory processes in waking and the virtual silence of monoaminergic neurons. This might explain the psychological differences between the logical thought in waking and the bizarre, illogical, and emotional content of dreaming in REM sleep.

Sleep Oscillations: Sites of Generation and Cellular Mechanisms

Three Major Oscillations Define Slow-Wave Sleep

The brain rhythms that characterize slow-wave sleep are the slow oscillation (0.5–1 Hz) that appears from the onset of full-blown sleep; spindles (7–15 Hz), whose amplitudes are highest during stage 2 of slow-wave sleep; and delta waves (1–4 Hz), which appear in stages 3 and 4, when their great amplitudes may obscure spindles.

The slow oscillation was first described using intracellular recordings from different neuronal types in anesthetized animals and, in the same article, was also detected in EEG recordings during natural slow-wave sleep in humans in which cyclic groups of delta waves at 1–4 Hz recurred with a slow periodicity of 0.4 or 0.5 Hz. The distinctness between delta and slow oscillations also came from human studies showing that the typical decline in delta activity (1–4 Hz) from the first to the second sleep episode was not present at frequencies characteristic for the slow oscillation (range, 0.55–0.95 Hz). The cortical nature of the slow oscillation was demonstrated by its survival in the cerebral cortex after large thalamic lesions and its absence in the thalamus of decorticated animals. This rhythm was recorded in all major types of neocortical neurons, including pyramidal-shaped and local circuit inhibitory neurons. The slow oscillation is composed of a prolonged depolarizing ('up') phase followed by a long-lasting hyperpolarizing ('down') phase. The hyperpolarizations of cortical neurons occur from the onset of sleep and distinguish sleep from the steady depolarization seen during waking (Figure 1). The depolarizing phase consists of excitatory postsynaptic potentials (EPSPs), some intrinsic excitatory currents, and also inhibitory postsynaptic potentials (IPSPs) that reflect the action of synaptically coupled local

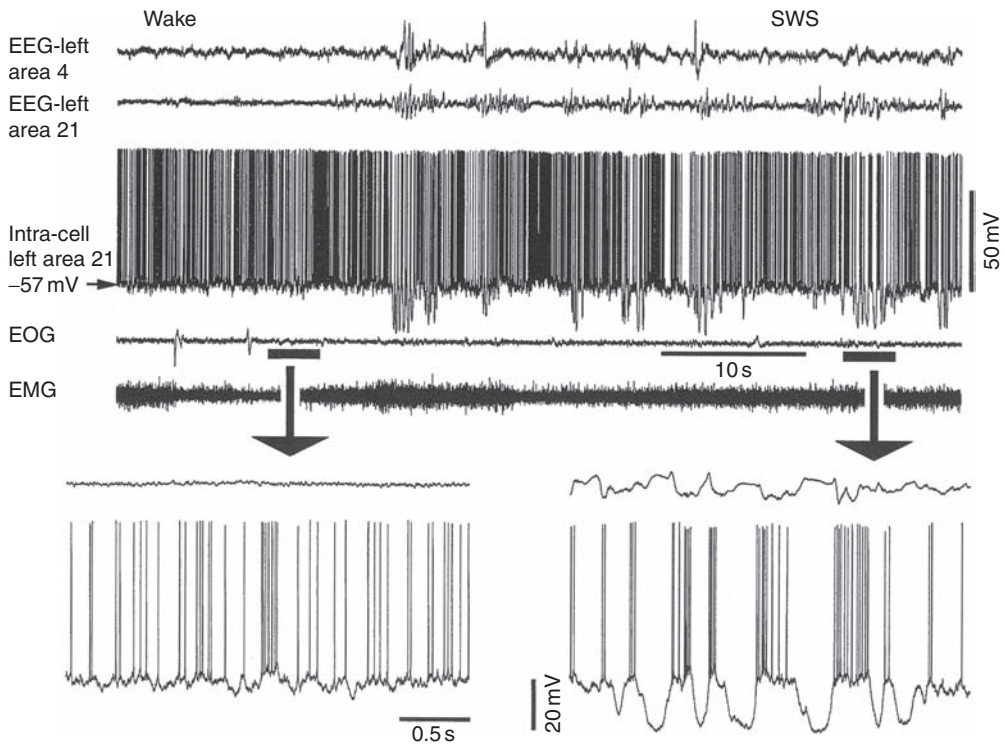

Figure 1 The slow oscillation marks the difference between waking and slow-wave sleep (SWS). Intracellular recording of regular-spiking neuron from left cortical area 21 of chronically implanted cat, together with depth EEG from left areas 4 and 21, electrooculogram (EOG), and electromyogram (EMG). Transition from waking to SWS. Two epochs, one in waking and the other in SWS, are marked by horizontal traces (below EOG) and expanded below (arrows). Note the occurrence of episodic cyclic hyperpolarizations of the slow oscillation since the very onset of SWS, as indicated by EEG.

circuit GABAergic neurons. The hyperpolarizing (silent) phase is not produced by inhibitory interneurons but is due to disfacilitation (removal of synaptic, mainly excitatory, inputs) in intracortical and thalamocortical networks and also to some K^+ currents.

Sleep spindles are waxing-and-waning waves, grouped in sequences that recur every 2–5 s. They are generated in the thalamus, even in the absence of the cerebral cortex, but corticothalamic neurons have a decisive role in the near simultaneity of spindles over widespread thalamic and cortical territories. The pacemaker of spindles is the thalamic reticular nucleus, which consists exclusively of inhibitory (GABAergic) neurons. In the absence of this nucleus, spindles are not seen in the remaining thalamus and cortex, but this oscillation is present in the deafferented reticular nucleus. The cellular mechanisms underlying spindles are as follows. At each spindle wave, thalamic reticular neurons fire spike bursts that impose IPSPs on their targets, thalamocortical neurons. Following IPSPs, thalamocortical neurons discharge postinhibitory rebound spike bursts that are transmitted to cortex, where they rhythmically excite cortical neurons and thus give rise to spindles. The rebound is an intrinsic property of thalamic neurons and depends on a Ca^{2+}-mediated conductance, which is inactive during waking and REM sleep when neurons are depolarized and is de-inactivated (uncovered) during slow-wave sleep when thalamocortical neurons are hyperpolarized.

Some delta waves arise in the cortex without thalamic participation. Another type of delta potentials is generated in thalamocortical neurons through interplay between two of their intrinsic properties.

Coalescence of low-frequency corticothalamic rhythms by the slow oscillation Unlike rhythms that appear within distinct frequency bands and are generated in restricted neuronal circuits of simplified experimental preparations (e.g., isolated thalamic or cortical slices maintained *in vitro* and decorticated or thalamectomized animals *in vivo*), the normal brain does not generally display separate oscillations during slow-wave sleep. Rather, it displays a coalescence of the slow oscillation with other sleep rhythms (spindles and delta) as well as with faster (beta and gamma) rhythms that are superimposed on the depolarizing phase of the slow oscillation (Figure 2). The circuitry and neuronal mechanisms that account for the grouping of low-frequency and fast-frequency rhythms by the slow oscillation are discussed later.

During the depolarizing phase of the slow oscillation, the synchronous firing of neocortical neurons impinges upon thalamic reticular pacemaking neurons, thus creating conditions for formation of spindles, which are transferred to thalamocortical neurons and up to cortex, at which level spindles shape the tail of the slowly oscillatory cycle (Figure 2, middle panel in 'slow + spindle'). This connectivity explains why a cycle of the slow oscillation is followed by a brief sequence of spindles in thalamocortical neurons and in the cortical EEG (Figure 2, left panel in 'slow + spindle'), as seen with intracellular recordings in animal studies as well as with EEG recordings in human slow-wave sleep. The sequence of grapho-elements consisting of an ample surface-positive transient, corresponding to the excitation in deeply lying cortical neurons, followed by a slower, surface-negative component and eventually a few spindle waves, represents the combination between the slow and spindle oscillations. It is termed the K-complex and is a reliable sign for stage 2 of human sleep, but it is apparent in all stages of slow-wave sleep. These investigations indicate that K-complexes are the expression of the spontaneously occurring, cortically generated slow oscillation, although K-complexes can also be evoked by sensory stimuli during sleep.

Why fast rhythms are also observed during slow-wave sleep The unexpected association between a slow sleep rhythm and fast oscillations that are conventionally regarded as defining the electrical activity of waking and REM sleep is explained by the voltage dependency of fast oscillations. Indeed, cortical and thalamic neurons generate beta and gamma oscillations at relatively depolarized values of the membrane potential. Thus, fast rhythms are sustained during the steady depolarization of cortical neurons during waking and REM sleep, selectively appear over the depolarizing phase of the slow sleep oscillation, and are absent during the hyperpolarizing phase of the slow oscillation (Figure 2). The association between slow oscillation and fast rhythms has also been reported in human sleep.

Fast Oscillations during REM Sleep and Wakefulness

Beta and gamma waves (20–80 Hz) occur spontaneously during REM sleep and waking and are evoked by intense attention, conditioned responses, tasks requiring fine movements, or sensory stimuli. Beta and gamma rhythms can interchangeably be termed fast because neurons may pass from beta to gamma oscillation in very short periods of time (0.5–1 s), with slight depolarization of cortical neurons. Studies in humans have also shown that there is no precise cutoff between the beta and gamma bands since these activities may fluctuate simultaneously, as indicated by increased activities within both beta and gamma frequency bands (20–40 Hz) during cognitive processes implicating memory. The synchronized fast

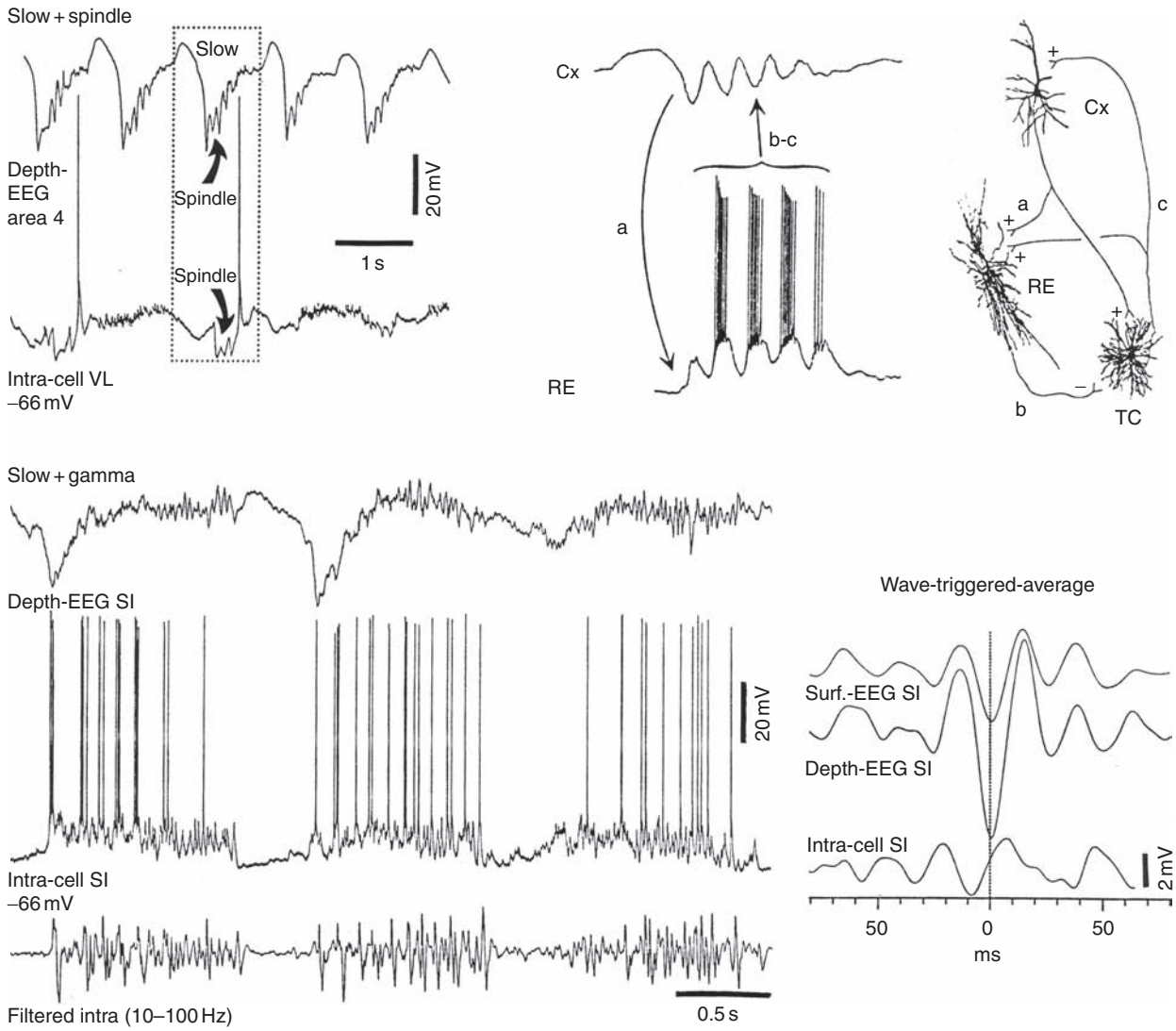

Figure 2 Coalescence of slow oscillation with spindle and gamma rhythms. Intracellular recordings from cortical and thalamic neurons in cats. Slow + spindle, combined slow oscillation and spindle. (Left) Depth EEG from cortical area 4 and intracellular recording of thalamocortical neuron from ventrolateral (VL) nucleus. The excitatory component (negative depth EEG wave, downward deflection) of the slow cortical oscillation (0.9 Hz) is followed by a sequence of spindle waves at 10 Hz (arrows). One typical cycle of these two combined rhythms is indicated by the dotted box; note the inhibitory postsynaptic potentials in the VL neuron leading to a postinhibitory rebound. (Right) Top and bottom traces represent field potential from the depth of association cortical area 5 and intracellular recording from thalamic reticular (RE) neuron. In neuronal circuits (far right), synaptic projections are indicated with small letters, corresponding to the arrows at left, which indicate the time sequence of the events. The depolarizing phase of the field slow oscillation (depth-negative, downward deflection) in the cortex (Cx) travels through the corticothalamic pathway (a) and triggers in the thalamic reticular nucleus (RE) a spindle sequence that is transferred to thalamocortical cells (TC) of the dorsal thalamus (b) and thereafter back to the cortex (c), where it shapes the tail of the slow oscillatory cycle (middle panel). Slow + gamma, fast activity (40 Hz) crowns the depolarizing phase of the slow oscillation. Three traces depict depth EEG waves from primary somatosensory cortex (S1), intracellular recording from S1 neuron, and filtered intracellular trace (between 10 and 100 Hz). Note the fast waves (40 Hz) during the depolarizing phase of the slow sleeplike oscillation and the absence of such fast waves during hyperpolarization.

rhythmic activity led to hypotheses postulating that linkages between spatially distributed oscillatory elements in the visual cortex may be the bases for pattern recognition function. However, beta and gamma activities are also present in the spontaneous activity of neurons and EEG, and these oscillations have been recorded under deep anesthesia when cognitive processes are suspended.

The best candidates for generation and synchronization of fast activities are fast rhythmic-bursting neurons. These neurons are both corticothalamic excitatory and local circuit inhibitory neurons, which may serve synchronization of fast activities within corticothalamic circuits and also be part of the role played by cortical inhibitory neurons in these oscillations. The thalamus and neocortex display coherent beta and gamma

rhythms. At variance with the long-range synchrony of the slow sleep oscillation, fast rhythms are synchronized over restricted cortical territories and within specific circuits between thalamocortical and neocortical areas.

Synaptic Plasticity Induced by Sleep Oscillations

Far from being epiphenomena of intrinsic properties and synaptic operations in cortical and thalamic neuronal networks, with little or no functional significance, sleep oscillations lead to synaptic plasticity and are implicated in memory consolidation. Earlier views that considered sleep to be associated with widespread inhibition throughout the cortex and subcortical structures, which would lead to abolition of cognitive and conscious events, have been challenged by intracellular recordings during natural slow-wave sleep demonstrating rich spontaneous firing of neocortical neurons. These data refute the assumption that cortical neurons are inactive in this state. Although external signals are blocked at the thalamic level during slow-wave sleep, mainly because thalamocortical neurons are inhibited during spindle oscillations, the intracortical dialog is maintained and the responsiveness of cortical neurons to cortical volleys is even increased during slow-wave sleep. In humans too, this cortically evoked response is stronger in slow-wave sleep than in waking. That the neocortex is active during slow-wave sleep suggests a reorganization/specification of neuronal circuits. This view is supported by studies using indicators of neuronal activities during slow-wave sleep in humans, revealing more marked changes in those neocortical areas that are implicated in memory tasks and decision making during wakefulness.

Then, spontaneously occurring brain rhythms may lead during sleep to increased responsiveness and plastic changes in the strength of connections among neurons, a mechanism through which information is stored. Next, the results of animal experiments and human studies on synaptic plasticity are briefly presented.

The experimental model of sleep spindles is the sequence of augmenting responses, defined as thalamically evoked cortical potentials that grow in size during the first stimuli at a frequency of approximately 10 Hz, which mimics the initially waxing pattern of spindle waves. In the intact brain, augmenting responses evoked by rhythmic thalamic stimulation are characterized in cortical neurons by an increase in the secondary depolarization, at the expense of the primary EPSP. Synaptic plasticity is not only seen by progressively enhanced amplitudes of responses during the pulse train but also by persistence of self-sustained potentials, with the same pattern and frequency as those of responses during the prior stimulation period (Figure 3(a)). Stimulation of the corticothalamic projection with pulse trains at 10 Hz results in evoked responses and also, after protracted stimulation, in 'spontaneously' occurring spike bursts whose form and rhythmicity are similar to those of evoked responses, as if the repetition of volleys was imprinted in the 'memory' of the corticothalamic network (Figure 3(c)). Short- and medium-term (5–30 min) neuronal plasticity can occur inside cortical circuitry, even in the absence of the thalamus (Figure 3(b)).

Thus, besides their role in cortical disconnection through inhibition of incoming messages in the thalamus, sleep spindles are also operational in important cerebral functions. During spindles, rhythmic and synchronized spike bursts of thalamic neurons depolarize the dendrites of neocortical neurons, which is associated with massive Ca^{2+} entry that may provide an effective signal to efficiently activate Ca^{2+}/calmodulin-dependent protein kinase II, implicated in synaptic plasticity of excitatory synapses.

Similar phenomena, with Ca^{2+} entry in dendrites and somata of cortical neurons, occur with rhythmic spike trains associated with oscillations in the slow (0.5–1 Hz) or delta (1–4 Hz) frequency bands, during later stages of slow-wave sleep. The hypothesis that the slow oscillation is responsible for the consolidation of memory traces acquired during the state of wakefulness is supported by data showing that slow and delta oscillations are implicated in the cortical plasticity of the developing cortex.

These experimental data are congruent with human studies demonstrating that the overnight improvement of discrimination tasks requires some steps, including those in early slow-wave sleep stages. Also, procedural memory formation may be associated with oscillations during early sleep stages. The early part of night sleep favors retention of declarative memories, whereas the late part of sleep favors retention of nondeclarative memories.

At variance with the commonly used notion of global brain processes in slow-wave sleep are two major findings reported in humans subjects: slow-wave sleep activity increases 2 h after a motor learning task, and the enhancement is expressed locally in parietal association areas that receive converging visual and proprioceptive inputs relevant to spatial attention and skilled actions.

The increased cortical activity that accounts for consolidation of memory traces during slow-wave sleep also explains the presence of dreaming mentation during this state. In slow-wave sleep, dreaming is rational and repetitive, in contrast with the vivid, illogical, and emotional perceptions during REM

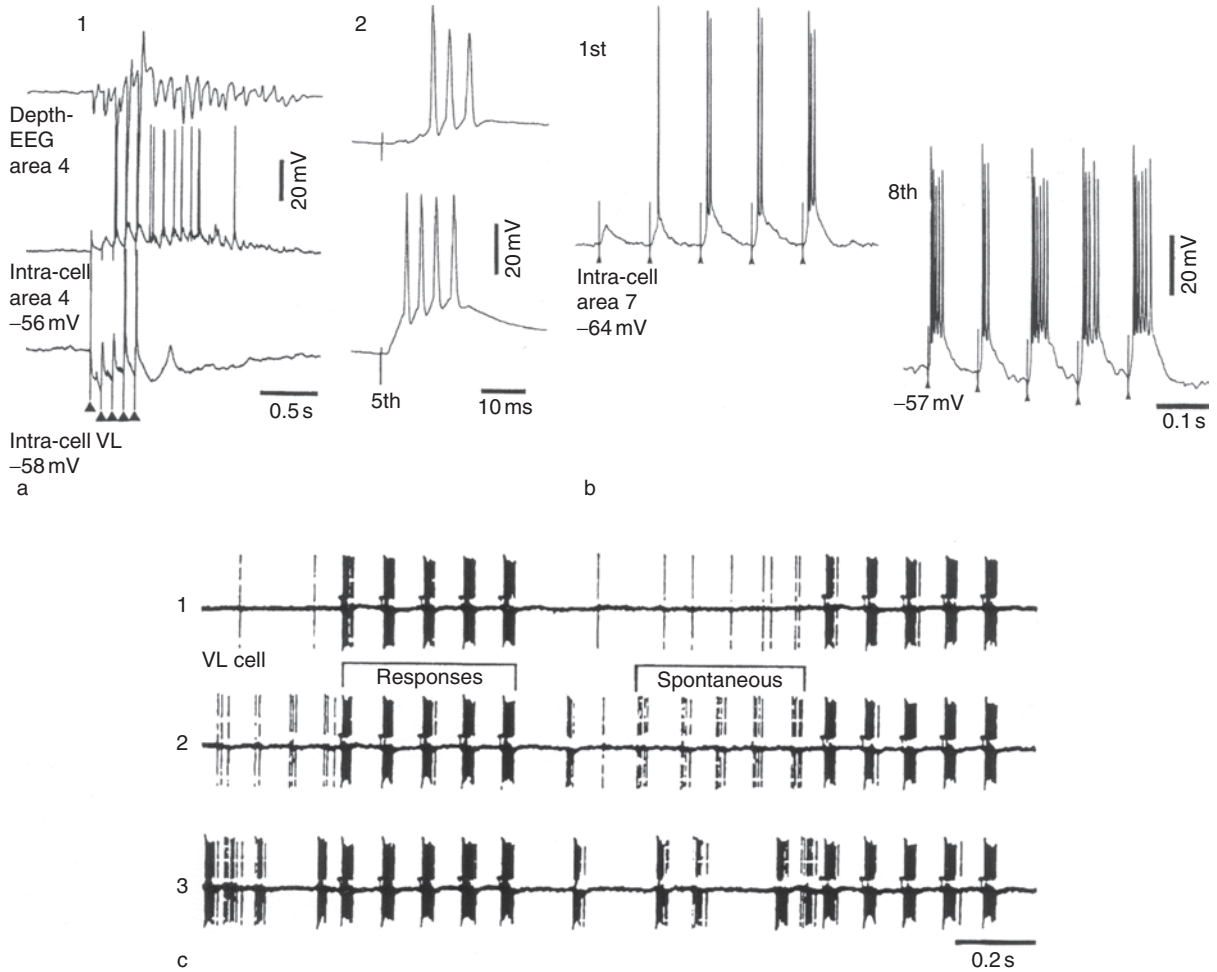

Figure 3 Synaptic plasticity in thalamocortical and intracortical systems, and 'memory' of electrical responses in corticothalamic system, induced by low-frequency stimuli mimicking sleep spindles. (a) Dual simultaneous intracellular recordings from cortical and thalamocortical (TC) neurons in cat (top trace is depth EEG from area 4). (1) Pulse train (five stimuli at 10 Hz; arrowheads) applied to the thalamic ventrolateral (VL) nucleus produced augmenting responses in the cortical neuron, whereas simultaneously recorded VL neuron displayed hyperpolarization. (2) Expanded fifth response; the augmented response in cortical neuron followed the rebound spike burst of the VL neuron. Note the self-sustained oscillatory activity at 10 Hz in the cortical neuron after cessation of thalamic stimuli, despite persistent hyperpolarization in the VL neuron. (b) Intracellular responses of cat area 7 bursting cortical neuron to repetitive callosal stimulation (10 Hz). The thalamus ipsilateral to the recorded neuron was extensively lesioned using kainic acid. The intracortical augmenting responses to the first and eighth pulse trains are illustrated. Note the depolarization by approximately 7 mV and the increased number of action potentials within bursts after repetitive stimulation. (c) Extracellular recording of the VL neuron in brain stem-transected cat. Motor cortex stimulation with pulse trains at 10 Hz (stimuli are marked by dots). (1) The pattern of responses in the thalamic VL neuron in early stages of rhythmic pulse trains. (2 and 3) Responses at later stages of stimulation. Note the appearance of spontaneous spike bursts resembling the evoked ones as a form of 'memory' in the corticothalamic circuit.

sleep. The brain is never 'empty,' and mental activity is present during all stages of normal sleep.

Ponto-Geniculo-Occipital Waves

PGO waves are sharp field potentials that are generally recorded in the visual thalamic lateral geniculate (LG) nucleus and occipital cortex, where they appear in clusters of up to six waves, closely related to gaze direction in dream imagery. Although the term PGO indicates their presence in the visual pathway, these potentials appear in many thalamic nuclei and cortical areas, outside the visual system. This widespread occurrence is due to the generalized thalamic projections of PGO generators, which are cholinergic neurons located in the pedunculopontine tegmental (PPT) nucleus and an adjacent nucleus at the mesopontine junction. Five distinct neuronal types have been recorded within cholinergic nuclei, whose discharges are temporally related to PGO field potentials in the visual thalamic nucleus.

The major input sources of PPT neurons are in adjacent areas of the upper brain stem reticular formation,

hypothalamic areas, and cerebral cortex. Impulses in the PGO generators can be triggered by any of these inputs through direct excitation or postinhibitory rebound excitation. Then, PPT neurons can be regarded as the final link in the brain stem–thalamic path that generates PGO waves. Because PGO potentials are thought to represent 'the stuff dreams are made of' and in view of the emotional nature of dreaming mentation in REM sleep, hypothalamic and forebrain structures that store some of the emotionally charged information may be most effective in driving the PGO brain stem neuronal generators.

In naturally sleeping cats – the species of choice for experimental studies because they display sleep patterns similar to those of humans – PGO waves precede other signs of REM sleep by approximately 30–90 s, and they appear during the last period of slow-wave sleep. Thus, there is a transitional period between EEG-synchronized and EEG-activated (REM) sleep that is called the pre-REM period, during which PGO waves appear over the background of a fully synchronized EEG. Thalamic neurons are hyperpolarized during the pre-REM period, when the sleep EEG is still fully synchronized, whereas they are depolarized during REM sleep. These two states (pre-REM and fully developed REM sleep) generate different PGO-related responses of LG neurons to brain stem cholinergic inputs, which influence the signal-to-noise ratio in the visual channel – that is, the ratio between the neuronal activity related to the PGO signal and the background firing of the same neuron. During pre-REM, the activity of thalamic LG cells starts with a short spike burst coinciding with a high-amplitude PGO wave and continues with a train of single spikes. By contrast, during full-blown REM sleep, the rate of LG cells' spontaneous firing is much higher than in pre-REM; LG cells do not fire spike bursts but, rather, single action potentials; and the amplitudes of PGO waves are much lower than during the pre-REM state. The peri-PGO histograms of LG neuronal activities show that the signal-to-noise ratio reaches values of approximately 7 during the pre-REM epoch, whereas the ratio values during REM sleep are approximately 2.

Why all these details? Because PGO waves are commonly regarded as the physiological correlate of dreaming, the greater signal-to-noise ratio in the LG-cortical channel during the pre-REM epoch than during REM sleep suggests that the vivid imagery associated with dreaming sleep may appear before fully developed REM sleep, during a period of apparent slow-wave sleep. Discussing dreaming from experiments in cats is justified in view of the behavioral repertoire typical for dreaming mentation that can be elicited in cats after having prevented muscular atonia by lesions of certain brain stem reticular structures. The idea that PGO waves with greater amplitudes during the pre-REM stage may reflect more vivid imagery during that epoch than even during REM sleep corroborates earlier data showing that after interrupting sleep immediately after the occurrence of the first PGO wave (in the pre-REM stage) and eliminating approximately 30 s of the slow-wave sleep stage that precedes REM sleep, the increased time of the REM sleep rebound was due to phasic events (PGO waves) rather than the loss of REM sleep *per se*. This observation fits in well with data on dream reports from the last epoch of EEG-synchronized sleep, which are indistinguishable from those obtained from REM sleep awakenings.

PGO waves are thought to play a role in development and structural maturation of the brain. PGO waves also exert specific excitatory effects on complex neurons in the visual cortex, and their deprivation may modify the expression of long-term potentiation in immature animals.

See also: Sleep Architecture; Sleep and Sleep States: PET Activation Patterns; Sleep and Sleep States: Hippocampus–Neocortex Dialog; Sleep and Sleep States: Network Reactivation; Sleep and Sleep States: Gene Expression; Sleep Oscillations; Sleep-Dependent Memory Processing.

Further Reading

Achermann P and Borbély A (1997) Low-frequency (<1 Hz) oscillations in the human sleep EEG. *Neuroscience* 81: 213–222.

Contreras D, Destexhe A, Sejnowski TJ, and Steriade M (1996) Control of spatiotemporal coherence of a thalamic oscillation by corticothalamic feedback. *Science* 274: 771–774.

Contreras D and Steriade M (1995) Cellular basis of EEG slow rhythms: A study of dynamic corticothalamic relationships. *Journal of Neuroscience* 15: 604–622.

Gais S, Plihal W, Wagner U, and Born J (2000) Early sleep triggers memory for early visual discrimination skills. *Nature Neuroscience* 3: 1335–1339.

Hobson JA and Pace-Schott EF (2002) The cognitive neuroscience of sleep: Neuronal systems, consciousness and learning. *Nature Reviews Neuroscience* 3: 679–693.

Huber R, Ghilardi MF, Massimini M, and Tononi G (2004) Local sleep and learning. *Nature* 430: 78–81.

Maquet P, Degueldre C, Delfiore G, et al. (1997) Functional neuroanatomy of human slow wave sleep. *Journal of Neuroscience* 17: 2807–2812.

Massimini M, Ferrarelli F, Huber R, Esser SK, Singh H, and Tononi G (2005) Breakdown of cortical effective connectivity during sleep. *Science* 309: 2228–2232.

Massimini M, Huber R, Ferrarelli F, and Tononi G (2004) The sleep slow oscillation as a traveling wave. *Journal of Neuroscience* 24: 6862–6870.

Mölle M, Marshall L, Gais S, and Born J (2002) Grouping of spindle activity during slow oscillations in human non-REM sleep. *Journal of Neuroscience* 22: 10941–10947.

Plihal W and Born J (1997) Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience* 9: 534–547.

Steriade M (2003) *Neuronal Substrates of Sleep and Epilepsy*. Cambridge, UK: Cambridge University Press.

Steriade M and Llinás RR (1988) The functional states of the thalamus and the associated neuronal interplay. *Physiological Reviews* 68: 649–742.

Steriade M and McCarley RW (2005) *Brain Control of Wakefulness and Sleep*. New York: Springer.

Steriade M, Nuñez A, and Amzica F (1993) A novel slow (<1 Hz) oscillation of neocortical neurons *in vivo*: Depolarizing and hyperpolarizing components. *Journal of Neuroscience* 13: 3252–3265.

Steriade M, Paré D, Bouhassira D, Deschênes M, and Oakson G (1989) Phasic activation of lateral geniculate and perigeniculate neurons during sleep with ponto-geniculo-occipital spikes. *Journal of Neuroscience* 9: 2215–2229.

Steriade M and Timofeev I (2003) Neuronal plasticity in thalamocortical networks during sleep and waking oscillations. *Neuron* 37: 563–576.

Steriade M, Timofeev I, and Grenier F (2001) Natural waking and sleep states: A view from inside neocortical neurons. *Journal of Neurophysiology* 85: 1969–1985.

Stickgold R, James L, and Hobson JA (2000) Visual discrimination learning requires sleep after training. *Nature Neuroscience* 3: 1237–1238.

Sleep Research and Sleep Medicine in Historical Perspective

R Pelayo and C Guilleminault, Stanford University Sleep Medicine Program, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

When the history of sleep research and sleep medicine began is a matter of debate, but modern sleep medicine may be related to development of sleep research tools. In 1929, J-H Berger recorded the electrical activity of the human brain and demonstrated differences in these rhythms during wakefulness and sleep. He called the signals that he recorded 'electroencephalogram' (EEG). At Harvard University, Loomis et al. studied sleep onset and overnight sleep, and they published the results of their investigation in several reports between 1937 and 1939, dividing sleep into five states (A–E). Blake et al. provided information on sleep subdivision. The description of specific EEG rhythms during sleep was the beginning of a new era in sleep research.

However, the overall development of neurosciences at the beginning of the twentieth century had other impacts on the clinical understanding of sleep. Constatin Von Economo probably provided the most insight on central nervous system structures involved in the control of sleep. While investigating the viral pandemic that began at the end of World War I and killed many people, he made important observations on patients affected by 'encephalitis lethargica.' He correlated the clinical symptoms of insomnia and continuous lethargy with the presence of two different sites of inflammatory lesions. Lesions in the preoptic area and anterior hypothalamus led to insomnia and hyperkinetic movements, whereas lesions of the posterior hypothalamus led to sleepiness, continuous lethargy, and eye movement abnormalities. The role of the ventrolateral preoptic nucleus in sleep induction has been demonstrated, as has the role of the hypocretin peptides along with the identification of the hypocretin-containing neurons in the posterolateral regions of the hypothalamus. However, in the early 1920s Von Economo had found that the anterior hypothalamic lesions were often extended to the diencephalon (basal ganglia), whereas the posterior hypothalamic lesions were associated with lesions in the periaqueductal gray matter with frequent extension to the oculomotor nuclei, which thus explained the oculomotor symptoms. This led Von Economo to speculate that the anterior hypothalamus contained a sleep-promoting area and that an area spanning from the posterior wall of the third ventricle to the third cranial nerve was involved in activity promoting wakefulness. He also speculated that the syndrome of narcolepsy involved this posterior hypothalamus area (a fact also mentioned by CP Symonds in 1926, based on Von Economo's findings), as reported again very recently.

In 1949, Moruzzi and Magoun made an important discovery that had a major impact on further electrophysiological research on sleep that began to attract increasingly more researchers. They showed that the brain stem netlike core of neurons, the reticular formation, had ascending pathways, leading to the discovery of the 'ascending reticular activating system.' Stimulation of this neuronal network induced long-lasting and widespread cortical activation, marked by replacement of cortical slow waves with fast activity. They also showed that lesions of the reticular formation, but not those of sensory pathways, produced a loss of cortical activation. The lesions with most marked effect were located in the oral pontine and midbrain reticular formation and the posterior hypothalamic–subthalamic regions where ascending pathways reach into the forebrain. Additional, smaller lesions in the midbrain tegmentum and caudal diencephalons showed that dissociation between cortical activation and behavioral arousal of wakefulness could be obtained.

Deficiency of cortical activation without prevention of behavioral responsiveness to sensory stimulation is obtained with lesions of central midbrain tegmentum, whereas lesions of the ventral tegmentum and hypothalamus lead to unresponsiveness without loss of cortical activation.

Following these initial studies, increasingly more lesions at different levels were performed, leading to a delimitation of structures responsible for activation of the cortex, more numerous than initially considered, including not only the reticular formation but also the posterior hypothalamus–subthalamus and basal forebrain systems (nucleus basalis of Meynert). Transection of the brain stem behind the oral pontine tegmentum by Batini and other collaborators of Moruzzi resulted in a complete insomnia.

The understanding of sleep was greatly changed by the discovery of what has been called rapid eye movement (REM) sleep, paradoxical sleep, or desynchronized sleep. It is related to the investigations by Nathaniel Kleitman on rest and activity cycles. Eugene Aserinsky, a graduate student of Kleitman, was collecting data on napping infants and designed a data sheet with 5 min periods with indication of 'periods of motility' or 'absence of motility.' Aserinsky and Kleitman obtained data suggesting the presence

of an 'eye motility rhythm' and decided to verify it in adults. In order to monitor the eye movements during the night, they decided to measure eye movements using electro-oculography (EOG). While monitoring these electro-oculograms, they noted for the first time the presence of two very different types of movements: movements with bursts different from slow eye movements seen at sleep onset. Initially thought to be artifacts, direct observations of the phenomenon indicated by the recordings confirmed the reality of the movement. The investigators also noted that simultaneously during these specific eye movements, irregular breathing and heart rate accelerations were seen. Based on their observations, they assumed that there was a 'lightening' of sleep with these movements and that they may represent dreaming. Aserinsky and Kleitman had no continuous EEG recording but woke up their subjects during periods of occurrence and absence of eye movements, obtaining very different findings as reported in 1953. It was with another student, William C Dement, that Kleitman monitored EEG continuously to further investigate the association of EOG and EEG. In 1957, Dement and Kleitman reported the cyclical variation of EEG during sleep and its relation to eye movements, body motility, and dreaming. They found that the cyclical variation from a period of eye movement to the successive one was between 90 and 100 min, and that there was a regular re-occurrence of rapid eye movements and dreaming. However, at that time, sleep was still considered as only one state. They called these periods with rapid eye movements 'emergent stage 1' sleep. After reading Dement's 1958 report on the occurrence of low-voltage, fast EEG patterns during behavioral sleep in cats, French researcher Michel Jouvet replicated the findings and indicated that the cat also presented a muscle atonia in association with the rapid eye movement and the characteristic EEG pattern. In 1962, Jouvet used the term 'paradoxical sleep,' with the 'paradox' being that although the studied cat had an EEG that resembled wakefulness, behaviorally the animal remained asleep and unresponsive. With Francois Michel, Jouvet also reported that this new 'state' was associated with the presence of large electrical potentials that were first seen in the pons but appeared a few milliseconds later and were able to be recorded most easily in the lateral geniculate nucleus and, with further but limited delay, in the occipital cortex, leading to the term ponto-geniculo-occipital spikes. Sleep was divided by the Lyon, France, team into two states: slow wave sleep and paradoxical (also called dysynchronized or REM) sleep.

Following these discoveries, under the initiative of WC Dement, a small group of researchers formed the Association of Psychophysiological Study of Sleep. This small group invited other North American, European, and Asian researchers to participate in its research and to built a 'common sleep scoring system.' By consensus, these researchers agreed on a final product that had a tremendous impact on human sleep research. An atlas was created that could be used as a reference to score sleep in a similar way throughout the world. This atlas also included recommendations on how to monitor sleep as well as how and what to tabulate. This atlas, edited by Alan Rechtschaffen and Anthony Kales, was published in 1968.

These different efforts opened the field for systematic investigation of pathology of sleep in humans. However, there had previously been publications on specific disorders during sleep. In 1685, Willis gave the best description of restless leg syndrome, long before the 1945 publication of Ekbom that gave the name to the syndrome. An associated syndrome was noted with small movements of the big toe that occurred at regular intervals during sleep. This motor manifestation was questioned, and called 'nocturnal myoclonus' and erroneously thought to be an epileptic equivalent by Charles P Symonds in 1953; it was renamed periodic limb movement syndrome after demonstration of its linkage to restless leg syndrome (RLS) by Lugaresi and colleagues. Jean Baptiste Gelineau described the narcolepsy syndrome in 1880. He named the syndrome that he considered as an independent entity. If the terminology 'narcolepsy' began in 1880, Caffe in 1862, as cited by Gelineau, had already described a 'sleeping sickness' that plagued a subject with short attacks of sleepiness. The patient also presented 'hallucinations' that had him hospitalized in a mental hospital. Also, Westphal in 1877 published a report on a patient with sleeping attacks associated with episodes of loss of motor and language abilities. He also indicated that the mother was also affected by the same problem. What would be called the 'Pickwickian syndrome' by William Oster was first reported in the nineteenth century by Robert Caton in England. He reported on 8 February 1861 to the Clinical Society of London about a 37-year-old poulterer with excessive sleepiness: "The moment he sat down on his chair sleep came on and even when standing and walking he would sink into sleep constantly while serving customers in shop, sleep would come as he stood by the counter, he would wake and find himself holding in his hand the duck or chicken which he had been selling to a customer a quarter of an hour before, the customer having in the meantime departed." Caton goes on to describe that "when in sound sleep a very peculiar state of the glottis is

observed, a spasmodic closure entirely suspending respiration (1889). The thorax and abdomen are seen to heave from fruitless contractions of the inspiratory and expiratory muscles; their efforts increase in violence for about a minute or a minute and a half, the skin mean time becoming more and more cyanosed, until at last, when the condition to an on-looker is most alarming, the glottic obstruction yields, a series of long inspirations and expirations follows and cyanosis disappears. This acute dyspneic attack does not awaken the patient. The night nurses stated that these attacks go on throughout the night." The usage of polygraphic recording during sleep led to a new understanding of syndromes during sleep.

The first 'polysomnogram' monitored during nocturnal sleep was probably by Bulow and Ingward, who studied breathing during sleep in normal subjects and in 1961 published a report stating that there was a clear change of breathing pattern during sleep onset, with the presence of a short respiratory pause and a change of carbon dioxide tension that occurred much earlier than the notion of the 'apnea threshold' and also a change in arterial CO2 set point when a transition from wakefulness to sleep was noted. Influenced by these researchers, Werner Gerardy in Heidelberg modified an EEG machine to monitor respiration and pulse rate. He performed the recording of an obese patient hospitalized for headaches and who presented episodes of 'periodic breathing' with intermittent absence of airflow and association with change of heart rate. Thereafter, Gerardy et al. monitored another 'Pickwickian' patient, noted a similar phenomenon, and reported their findings in 1960. They reported the first sleep recording of a Pickwikian patient and first demonstration of apnea during sleep. These different reports, followed by those of Gastaut et al. and others on Pickwickian subjects, were still the realm of neurologists and clinical neurophysiologists.

However, in 1970, Christian Guilleminault, who had benefited from the advances in understanding in sleep physiology, pharmacology, and biochemistry, believed that the existence of three different states of being – wakefulness, nonrapid eye movement (NREM) sleep, and REM sleep – with many important neuronal changes implied that the control of vital functions would be affected by passage from one state to another and that pathologies may be seen. For Guilleminault there was a different dimension: the states of sleep that led to different controls as those seen during wakefulness. The existence of these different controls during the two sleep states led to special clinical presentations, required treatment aimed at the dysfunction occurring during the sleep states, and justified the development of what he called 'sleep medicine.' He was given limited funding in 1970 by the Societe Medicale des Hopitaux de Paris for investigation of sleep medicine patients in La Salepétriere Hospital, Paris. The skepticism in France led him to accept the invitation from William C Dement and Stanford University to work there in 1972.

Also in 1972, Elio Lugaresi and Pierre Sadoul organized the first international meeting on 'hypersomnia with periodic breathing' focusing on the abnormalities of breathing during sleep in obese subjects. In 1973, Guilleminault et al. reported sleep apnea and insomnia, and in 1975 the definition of an 'obstructive sleep apnea syndrome' was given. It was followed by the publication of obstructive sleep apnea syndrome in children and a presentation of the state of the art on sleep apnea syndrome in a monograph titled *Sleep Apnea Syndromes* that included most researchers involved in the study of the syndrome. Colin Sullivan, in the pulmonary physiology laboratory in Sydney, designed equipment to apply continuous air pressure to avoid the collapse of the upper airway. Using an air compressor and an airtight mask, Sullivan and colleagues succeeded in eliminating the sleep-related obstruction in a severely affected patient. After observing the results obtained with their prototype, they not only began treating patients in the hospital but also made major advances with regard to the possibility of a permanent nonaggressive treatment of obstructive sleep apnea (OSA) with application of continuous positive airway pressure (CPAP) through nasal mask at home, as reported in 1981.

The search for new treatments for OSA was not limited to the creation of nasal CPAP. Dr. Ikematsu, an otolaryngologist in Japan, developed a surgical procedure performed on the soft palate. The procedure was aimed at eliminating loud chronic snoring that was disturbing to spouses. Fujita, an American-Japanese otolaryngologist from Detroit, extended the procedure under the name uvulopalatopharyngoplasty (UPPP) in 1981. Because the procedure was not completely successful in several cases, this led to a search for other anatomy-oriented treatment modalities. The development of what became known as the 'Stanford surgical protocol' based on work by Nelson Powell, Robert Riley, and Christian Guilleminault began in 1981 and initially involved only mandibular geniotubercle advancement, but later involved a 'phase I' and a 'phase II' surgical approach, with phase II consisting of bimaxillary advancement that could provide results similar to those noted with nasal CPAP treatment. To avoid performing orthognathic surgery, efforts were made to use dental-related

devices to protrude the tongue forward during sleep. Samelson and Cartright reported objective results obtained with the 'tongue retaining device.' Following these initial trials, different types of dental appliances were designed to help patients with the syndrome of mild to moderate severity or residual problems post-UPPP.

The understanding that OSA was, commonly, a familial syndrome, often with involvement of anatomical and particularly mild changes in craniofacial features, led to systematic investigation of nonsyndromic children not only in families of adult patients with OSA but also in children with chronic snoring. Orthodontic techniques were explored. Rapid maxillary distraction (RMD), a well-defined orthodontic procedure, was thus first used in adults by Cistulli et al. for mild OSA. It was applied thereafter to prepubertal children with residual OSA post-adenotonsillectomy and narrow hard palate by the Rome and Stanford teams. Also, mandibular distraction osteogenesis with or without RMD on the hard palate was introduced for treatment of teenagers. Finally, investigation of children and, later, adults with more sophisticated monitoring sensors allowed the description of an abnormal breathing pattern that may not be associated with the defined sleep apnea and hypopnea syndrome: the upper airway resistance syndrome.

The first international meeting on narcolepsy was held in 1975 at Montpellier, France, at which the modern definition of the syndrome was developed. In 1983, Fuji and Honda reported a dominant association between a specific group, HLA (DR2), and the presence of narcolepsy in Japanese people. A strong HLA association between a certain HLA typing and narcolepsy was quickly confirmed. Neely and colleagues investigated different ethnic groups, particularly African Americans, and did not find an association with DR2, and the Stanford group showed that the best and most common HLA association was with HLA DQB1–0602.

The discovery of a canine model for narcolepsy cataplexy led to many pharmacological studies and allowed the recognition that the narcoleptic Doberman had a genetic defect on a newly discovered brain system – the hypocretin/orexin system. This discovery was nearly simultaneously associated with the finding that mice with knockout hypocretin receptor presented cataplectic attacks. In 2000, Nishino et al. showed that human narcoleptics presented an absence of measurable hypocretin/orexine levels in cerebrospinal fluid, and it was found that genetic defect was very rare in human narcolepsy. Also in 2000, Peyron et al. and Thannickal et al. showed destruction of hypocretin cells in the posterolateral region of the hypothalamus (the same region as identified by Von Economo in the brains of human narcoleptics) was the most common cause of narcolepsy and identified the presence of primary and secondary narcolepsy. The cause of the primary narcolepsy, which is most common, is unknown.

After the association between RLS and periodic leg movement had been made by Lugaresi et al. in 1968, the presentation of periodic limb movement as an isolated syndrome was made by Guilleminault et al. in 1978 and extended in 1982 by Coleman et al. In 1982, Akpinar reported the beneficial effect of dopamine agonist in the treatment of RLS – a finding largely confirmed over time. However, the underlying defect associated with RLS is still elusive, despite the clear demonstration of a large group of familial cases and the identification in Quebec families of specific chromosomic loci associated with RLS occurrence.

Abnormal Behavior during Sleep

NREM Sleep Parasomnia

Broughton and Gastaut coined the term 'disorders of arousal' to qualify these clinical presentations and dissociate them from an epileptic disorder occurring during stages 3 and 4 of NREM sleep, usually during the first one-third of the night. Fisher et al. emphasized the presence of 'hypersynchronous' delta waves, which were later called 'delta bursts' or 'delta sequence' by others. The existence of a familial pattern of sleepwalking was noted by Kales et al., but data call into question whether the familial pattern is related directly to sleepwalking or it is related to an underlying cause of behavior. Analysis of sleep EEG patterns associated with NREM sleep parasomnia has used new approaches, particularly the investigation of the cyclic alternating pattern (CAP) initially described by Terzano et al. in adults. This analysis shows the presence of an abnormal CAP rate in patients, indicative of instability of NREM sleep. It is most often related to other underlying unrecognized and untreated sleep disorders, usually sleep-disordered breathing, particularly upper airway resistance syndrome or, rarely, RLS.

REM Behavior Disorder

Experimental lesions performed in cats on the brain stem induced an 'active' REM sleep: cats presented motor activity similar to dreaming (i.e., acting as if hunting or playing). They 'act out' their dreams due

to the absence of REM sleep motor atonia. In humans, Japanese teams observed similar findings with delirium. It was also observed in patients with a brain stem neurodegenerative disease, olivo-ponto-cerebellar degeneration, by Quera-Salva and Guilleminault. Finally, Schenck and Mahowald reported a series of 'idiopathic cases' and called the disorder 'REM behavior disorder' (RBD). Initially thought to be 'idiopathic,' it progressively appeared to be associated with other neurological disorders – not only narcolepsy, a syndrome of dissociated REM sleep, but also very prominently with Parkinson's disease, multiple system atrophy, and Lewy's dementia. The notion has emerged that RBD is more commonly associated with synucleopathy, raising the issue of idiopathic RBD being the first clinical indication of the presence of Lewy bodies located mostly in the lower brain stem before caudal extension of lesions.

Sleep investigations uncovered another neurodegenerative disorder: fatal familial insomnia (FFI). The genetic nature of the disorder, its direct linkage to a prion disease, and its inexorable evolution toward the complete disappearance of sleep and then death were identified by the continuous research led by Elio Lugaresi and involving both the Gambetti team in Cleveland and many researchers in Bologna. FFI is associated with the development of very severe dysautonomia, complete insomnia before death, severe dementia, and bilateral symmetrical degeneration of the thalamus.

Sleep–wake is part of chronobiology, and the history of sleep disorders is very much intertwined with the understanding of chronobiology and its defect leading to dyschronosis or chronobiologic disorders. Identification of genes such as CLOCK and PERIOD (or PER) led to research on circadian rhythms and sleep pathologies. Two syndromes, 'sleep phase delay' and 'sleep phase advance' syndromes, were described. The term sleep phase delay syndrome was introduced by Weitzman and colleagues and by Czeisler et al., who reported the first treatment attempt called 'chronotherapy.' The discovery of a very large family in Utah presenting phase advance syndrome led to the identification of a specific genetic marker for this syndrome. Advances in the treatment of dyschronosis have been based on the investigation of the role of light and dark cycles and placement of sleep in this cycle.

Conclusion

In 1994, the Center for Sleep Research of the National Heart, Lung, and Blood Institute of the National Institutes of Health was created after a major Congress-supported effort by a special committee chaired by WC Dement. The American Sleep Disorders Association and its successor, the American Academy of Sleep Medicine, have had their efforts recognized in accreditation of training centers in sleep medicine, and the accreditation process was handed over to the American Council on General Medical Education in 2005. The Sleep Medicine Board examination will be administered under the auspices of the American Board of Medical Specialties beginning in 2007. Sleep medicine is a living history.

See also: Cataplexy; Narcolepsy; Parasomnias; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Architecture; Sleep and Sleep States: PET Activation Patterns; Sleep Apnea; Sleeping Sickness.

Further Reading

Bulow K and Ingvar DH (1961) Respiration and state of wakefulness in normals, studied by spirography, capnography, and EEG: A preliminary report. *Acta Physiologica Scandinavia* 51: 230–238.

Czeisler CA, Richardson GS, Coleman RM, et al. (1981) Chronotherapy: Resetting the circadian clocks of patients with delayed sleep phase insomnia. *Sleep* 4: 1–21.

Ekbom KA (1945) Restless legs. *Acta Medica Scandinavia* 158 (supplement): 1–123.

Gastaut H, Lugaresi E, Berti-Ceroni G, and Coccagna G (eds.) (1968) *The Abnormalities of Sleep in Man*. Bologna: Aulo Gaggi.

Guilleminault C and Dement WC (1978) *Sleep Apnea Syndromes*. New York: Liss.

Guilleminault C, Dement WC, and Passouant P (eds.) (1976) *Narcolepsy*. New York: Spectrum.

Guilleminault C, Eldridge FL, and Dement WC (1973) Insomnia with sleep apnea: A new syndrome. *Science* 181: 856–858.

Jouvet M (1962) Research on the neural structures and responsible mechanisms in different phases of physiological sleep. *Archives Italiennes de Biologie* 100: 125–206.

Jouvet M and Michel F (1959) Corrélation électromyographiques du sommeil chez le chat décortique et mésencephalique chronique. *Comptes Rendus de la Société de Biologie (Paris)* 153: 29–45.

Juji T, Satake M, Honda Y, and Doi Y (1984) HLA antigens in Japanese patients with narcolepsy. All the patients were DR2 positive. *Tissue Antigens* 24: 316–319.

Kleitman N (1939) *Sleep and Wakefulness*. Chicago, IL: University of Chicago Press (Revised in 1963).

Lavie P (2003) *Restless Night*. pp. 29–45. New Heaven, CT: Yale University Press.

Loomis AL, Harvey FN, and Hobart G (1935) Further observations on the potential alpha-rhythms of the cerebral cortex during sleep. *Science* 2: 198.

Lugaresi E, Medori R, Montagna P, et al. (1986) Fatal familial insomnia and dysautonomia with selective degeneration of thalamic nuclei. *New England Journal of Medicine* 315: 997–1003.

Moruzzi G and Magoun HW (1949) Brain stem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology* 1: 455–473.

Peyron C, Faraco J, Rogers W, et al. (2000) A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nature Medicine* 6: 991–997.

Sadoul P and Lugaresi E (eds.) (1972) *Hypersomnia with Periodic Breathing: A Symposium, Bulletin de Physiopathologie Respiratoire* 8: 967–1200.

Schenck CH, Bundlie SR, Ettinger MG, and Mahowald MW (1986) Chronic behavioral disorders of human REM sleep: A new category of parasomnia. *Sleep* 9: 293–308.

Sullivan CE, Issa FG, Berthon-Jones M, and Eves L (1981) Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. *Lancet* 1: 862–865.

Von Economo C (1930) Sleep as a problem of localization. *Journal of Nervous and Mental Disease* 71: 249–259.

Sleep: Development and Circadian Control

H C Heller, Stanford University, Stanford, CA, USA
M G Frank, University of Pennsylvania, Philadelphia, PA, USA

© 2009 Published by Elsevier Ltd.

The states of arousal, wake, rapid eye movement (REM) sleep, and nonrapid eye movement (NREM) sleep, are physiologically and behaviorally well defined in adult mammals, and their temporal expression is clearly organized by the circadian system. The situation is not clear, however, in developing mammals. Comparing studies of the development of sleep/wake and circadian rhythms in different species of mammals is complicated because of the major differences in maturity at the time of birth. However, understanding the development of these systems can contribute to understanding their basic biology and their possible contributions to certain postnatal pathologies, such as sudden infant death syndrome.

Three Stages of Sleep Development

Altricial mammals are less fully developed at birth than are precocial mammals, and this difference is clearly reflected in the development of sleep/wake and circadian rhythmicity. Whereas precocial species have brain activity at birth that shows characteristic differences associated with adult arousal states, the electrical activity recorded from brains of altricial species is undifferentiated. This difference is not surprising given the fact that in altricial brains at the time of birth most axons have not yet reached their targets and much synaptogenesis has yet to occur. However, in both altricial and precocial species, the development of sleep/wake can be divided into three stages, some of which take place *in utero* in the precocial species. These stages are dissociation, concordance, and maturation.

Dissociation is the stage of development when the defining electroencephalography (EEG) characteristics of arousal states are absent but the other correlates of arousal states, such as changes in heart rate and respiratory rate and phasic or tonic motor activity, are present. However, these correlates are not coordinated with each other. In studies of premature human neonates, Hamburger applied the term 'presleep' to refer to the period prior to the emergence of the EEG components of sleep states. Similar situations have been observed *in utero* in studies of precocial species such as lambs and baboons. In altricial species such as the kitten, the absence of sleep EEG patterns and dissociation of the other hallmarks of sleep states are seen postnatally.

Concordance occurs with the emergence of differentiated EEG patterns characteristic of adult arousal states and the alignment of the other correlates of wake and sleep states with those EEG parameters. In precocial species, concordance occurs in late gestation and EEG definable arousal states are observable in the newborn. Concordance in humans begins between gestational weeks 28 and 32 and is completed sometime in the last month of fetal life in full-term infants. Throughout the period of concordance in precocial species, there is no evidence of transitional or precursor sleep states that are homologous with adult sleep states. Rather, the EEG characteristics of REM and NREM sleep emerge at approximately the same time and the other sleep parameters come into alignment with those EEG patterns.

Concordance in altricial species is not as clear and has generated some controversy. EEG patterns are undifferentiated in the newborn altricial animal, but motor patterns that appear similar to those associated with arousal states are evident. Specifically, these motor patterns are coordinated movements, jerky movements (twitches), or quiescence. Periods of tonic, coordinated motor activity are considered wake, and the rest of the time is considered to be sleep. Sleep is then divided into behavioral quiet sleep (BQS), characterized by quiescence, or behavioral active sleep (BAS), characterized by phasic twitches. Because the undifferentiated EEG pattern is relatively fast and low amplitude, it resembles the activated EEG that in adults is associated with wake and REM sleep. This association of an apparent activated EEG with phasic muscle activity (BAS) has been assumed to be a precursor of mature REM sleep, whereas quiescent periods (BQS) have been assumed to be the precursor of mature NREM sleep. However, the problem of equating an undifferentiated EEG with an activated EEG casts doubt on what appears to be an obvious continuity between these apparent precursor sleep states and adult sleep states defined by differentiated EEG patterns. We examine more critically the assumption that BAS is the precursor of REM sleep and that BQS is the precursor of NREM sleep later in this article. Note, however, that the terms 'active sleep' and 'quiet sleep' used to describe sleep states in premature and term human infants refer to different suites of variables than are described by these same terms in altricial species. The difference is that the EEG patterns of active and quiet sleep are clearly differentiated in human preterm and full-term infants. In contrast, BAS and BQS in altricial species are not distinguished by any differential EEG activity; rather, these states are distinguished primarily by the presence or absence of myoclonic twitches.

Following concordance of parameters into distinct sleep states, there are rapid and sometimes dramatic maturational changes in the amount and timing of those states and in some of the EEG characteristics that define them. Early in the maturation of sleep, the amount of REM sleep is much higher than adult levels and declines, whereas the amount of NREM sleep decreases to a lesser degree and wake increases. The characteristic brain activities that define the sleep states in the adult become increasingly well defined and in some cases altered. In altricial species, for example, the activated EEG of REM sleep precedes the development of the slower, highly regular theta rhythm, which has a frequency in the 4 or 5 Hz range early on and gradually increases to the characteristic adult frequency of 7 or 8 Hz. Pontine–geniculate–occipital waves do not appear until a week or more after the emergence of REM sleep, and the characteristic muscle atonia of REM sleep is even slower to develop. The two major EEG correlates of NREM sleep, slow waves (delta band, 0.5–4.5 Hz) and sleep spindles (7–14 Hz), develop at different rates. First to appear is slow wave activity, which in the rat steadily increases in amplitude from approximately postnatal day (P) 9 to P21, reaching higher levels than are seen in the adult. Sleep spindles appear in the third postnatal week. Relatively late to mature are the regulatory mechanisms governing the expression of sleep – the ultradian, circadian, and homeostatic controls.

BAS and BQS May Not Be Homologous with REM and NREM Sleep

In altricial species such as the rat, the undifferentiated EEG is the same in the behaviorally defined states of BAS and BQS. Since the undifferentiated EEG lacks high-amplitude wave patterns, it somewhat resembles the activated EEG of REM sleep. This resemblance, however, is superficial, as might be expected based on the fact that most axons traveling to the cortex have not yet reached their targets and most synaptogenesis has not yet occurred. Since REM sleep is a cholinergically mediated state, it is especially important to note that the cholinergic system is very immature at the time BAS is maximally expressed. There is an inverse relationship between BAS expression and (1) expression of muscarinic receptors and their signal transduction mechanisms, (2) the development of cholinergic nuclei and their efferents, and (3) neuronal responsiveness to cholinergic drugs. In fact, cholinergic blockade which eliminates adult REM sleep has no effect on BAS in the rat until late in the second postnatal week after differentiated EEG patterns have developed. Lesions of brain stem structures essential for mature REM sleep have no effect on BAS in neonatal rats or kittens. Even transection of the neonatal rat's spinal cord below the brain stem does not eliminate the phasic twitches that define BAS. In adult mammals, such transections eliminate all episodes of myoclonia that normally accompany REM sleep. These results pose the obvious question of what is the nature of the myoclonia of BAS. It is likely that the twitches of BAS are 'nonspecific fetal activity' that is associated with the development of spinal reflex circuits prior to the imposition of descending inhibitory control.

There is also no compelling evidence for assuming that BQS is a precursor state to NREM sleep. The comparison of different studies of BQS is complicated by the lack of a clear set of criteria that define BQS. If the definition of BQS is quiescent periods with regular respiration and no twitches, this amounts to only 2% of total recording time in 1-week-old rats. However, in other studies in which small movements or 'startles' are allowed in BQS, it can be more than 30% of total recording time. Similar variances in the designation of BQS in kittens exist, all depending on the behavioral criteria used by the investigators.

The first EEG hallmark of sleep states to appear in altricial species is slow waves, and when they appear, they are not restricted to episodes of BQS. In fact, when slow waves first appear in the neonatal rat, they are as likely, if not more likely, to occur during episodes of BAS than during episodes of BQS. By the third week of postnatal life, slow wave activity is mostly restricted to periods of motor quiescence and regular respiration, and the amplitude of the slow waves steadily grows until they reach maximum values at approximately the beginning of the fourth week. EEG theta activity, which marks REM sleep, appears in the neonatal rat at approximately the same time as slow waves, but it matures to adult levels more slowly than do slow waves.

In summary, the view most consistent with existing data is that both REM and NREM sleep emerge from an undifferentiated brain state, and they do so at approximately the same time even though NREM sleep patterns seem to mature more rapidly than REM sleep patterns. Thus, in both precocial and altricial species of mammals, there are not precursor sleep states that precede EEG differentiation. Rather, the development of the EEG hallmarks of sleep reflects the development of the underlying neuronal mechanisms, and presumably therefore the yet-to-be-discovered functions of the sleep states.

Sleep Regulation

There are three components of sleep regulation that develop after the concordance of EEG defined states: ultradian, circadian, and homeostatic. Ultradian refers

to the cyclicity of sleep states, or the sleep cycle. In most mammals, the distribution of total sleep time is approximately 80% NREM sleep and 20% REM sleep, but these sleep times are distributed into bouts that alternate with each other. The sleep cycle in the human adult is approximately 90 min and the sleep cycle in a rat is approximately 10 min. This cyclicity of sleep states is poorly developed at the time the states first emerge, and it develops gradually, with the cycles becoming more regular and longer.

Circadian refers to the daily cyclicity of sleep and wake which is dependent on the circadian pacemaker that resides in the suprachiasmatic nucleus (SCN) of the hypothalamus. Although the pacemaker is probably oscillating in the fetus and is probably entrained by the periodicity of the mother, its coupling to the systems controlling sleep states occurs considerably after the emergence of sleep states. In rats, the circadian organization of sleep begins to appear on approximately P15. In humans, circadian organization of sleep emerges in the second or third month after birth, but it takes longer to become entrained to the environmental cycle of light and dark.

Homeostatic regulation refers to the relationship between sleep and the duration of prior wake. In adult mammals, sleep homeostasis is robust. An increase in wake time results in an increase in sleep time, but most important, it results in an increase in the EEG slow wave activity in NREM sleep. Thus, whereas the effect of sleep deprivation on REM sleep is to increase its duration, the main effect of sleep deprivation on NREM sleep is to enhance the expression of slow wave activity, also called delta power. One striking developmental change is an increased tolerance of sleep pressure. The amount of wake is very low in neonatal humans and altricial species, and they are unable to sustain long episodes of wake. Short periods of sleep deprivation that would be insignificant in the adult result in clear sleep rebound in the neonate. However, the nature of the sleep rebound changes with age. Early on when there are robust rebounds in NREM sleep, there are none in REM sleep. In the rat, responses in REM sleep time to prior sleep deprivation first appear toward the end of the third postnatal week. Thus, as with the development of the EEG characteristics of sleep, the homeostatic characteristics of REM sleep seem to emerge later and develop more slowly than those of NREM sleep. A most striking change in the homeostatic regulation of NREM sleep occurs in the fourth postnatal week. Prior to this time, the response to sleep deprivation is increased NREM sleep time with no changes in NREM sleep slow wave activity. In the fourth postnatal week, the response of the rat pup to sleep deprivation is no longer an increase in NREM sleep time but, rather, an increase in slow wave activity or delta power. The significance of this sudden shift is not known, but it occurs after the time when normal delta power levels have reached their maximum and are declining and at the time when the ability of the animal to sustain longer episodes of wake is rapidly increasing. These observations indicate that the generation of EEG slow waves is a critical aspect of the restorative function of sleep.

Development of the Circadian System

The circadian system of mammals is practically synonymous with the SCN of the hypothalamus, but it also involves the pineal gland, which produces the hormone melatonin. In contrast to the complexity of the neural structures involved in sleep, the SCN and the pineal are discreet structures that can be studied anatomically, neurochemically, and functionally. Therefore, we can ask questions about the development of the circadian system at these levels.

The SCN is a compact but heterogeneous cluster of cells with a definite structure, yet the relationship between the organization of the SCN and its function as the circadian pacemaker is not understood. SCN cells arise from the germinal epithelium of the anterior, ventral diencephalon. In the rat, this begins on approximately embryonic day 14 (E14) and is complete by several days before birth. The fate determination of putative SCN cells is quite robust in that deafferentation of the area does not alter their differentiation, and even if they are excised and transplanted into the chamber of the eye, they still develop into SCN cells. Also, when SCN tissue is extracted in early postnatal life and placed in culture, the cells continue to express the properties of SCN cells, including circadian rhythmicity.

There are three major afferent pathways to the SCN, and their development has been described in the rodent. Information from photoreceptors reaches the SCN by a direct and an indirect route. The direct route is via the retinohypothalamic tract, in which the neurotransmitters are glutamate and pituitary adenylate cyclase-activating polypeptide (PCAP), and these connections develop between P1 and P10. The indirect pathway for photic information through the intergeniculate leaflet uses the neurotransmitter neuropeptide Y and develops between P4 and P11. The third set of projections are serotonergic from the midbrain raphe nuclei, and they develop between the day of birth and P10. These pathways are the major mechanisms of entrainment in the adult mammal, but as discussed later, entrainment of the fetal SCN occurs considerably before birth. Thus, additional mechanisms must be responsible for the initial entrainment of

the fetal SCN. Strong evidence shows that melatonin and dopamine may be two such factors that mediate the maternal entrainment of the circadian systems of her pups, and fetal SCN cells do express the receptors for these two neurochemicals.

Functional development of the circadian system involves: (1) pacemaker function, or the generation of rhythms; (2) entrainment of rhythms; and (3) the coupling of SCN rhythmicity to drive other variables, such as body temperature and sleep/wake. The first of these properties to develop, of course, is pacemaker function. Circadian cycles are observable in the SCN within days after the cells become postmitotic. The first evidence of rhythmicity in the fetal SCN was documented in rats by the differences in uptake of 2-deoxyglucose (2DG) by SCN cells when the 2DG was injected at two different times of day. The earliest these oscillations were observed in rats was approximately 3 days before birth. Of course, these changes in fetal SCN metabolism could have been solely imposed by a maternal influence. However, subsequent studies revealed daily changes in 2DG uptake and in electrical activity of the SCN in brain slices taken from fetuses. Such rhythms have also been observed in fetuses or in brain slices from fetuses after the dam received SCN lesions early in gestation and was herself arrhythmic. The evidence supports the view that pacemaking function develops in the SCN prior to birth.

The ability of the SCN oscillations to be entrained also develops before birth in rodents, but the entraining stimuli come from the mother. In normal circumstances, the newborn pups are in synchrony with each other and with the maternal circadian rhythm. If the dam has received an SCN lesion early in gestation and is therefore arrhythmic, the newborn pups express circadian rhythms, but they are not in synchrony with each other. Most telling are cross-fostering experiments in which the newborn pups are presented to mothers that are 180 degrees out of phase from their natural mothers. In these litters, when rhythmicity is expressed in the pups, it is in phase with the rhythm of their birth mothers and not their foster mothers. The fetal SCN seems to be highly sensitive to entraining signals, including melatonin, dopamine, and probably other compounds such as the cholinergic drug nicotine. Whether or not disruption of fetal rhythms by exposure to such compounds *in utero* has effects on postnatal health is not known. However, when pregnant ewes were kept in constant light, their circadian rhythms were disrupted, as evidenced by vasopressin levels in the cerebrospinal fluid, and their fetuses did not survive. When kept on a light/dark cycle, their circadian rhythms were normal and their fetuses survived. In the postnatal period, the sensitivity of the SCN to the entraining influence of light develops as the afferent pathways bringing photic information to the SCN develop.

The coupling of SCN oscillations to various physiological and behavioral properties of the newborn emerges at different postnatal ages. In the rat, a rhythm of pineal *N*-acetyltranferase (the rate-limiting enzyme for melatonin synthesis) activity develops in the first postnatal week, a rhythm of body temperature develops in the second postnatal week, and a rhythm of sleep/wake appears in the third postnatal week. In the human newborn, a clear day/night organization of sleep/wake develops by the third month of life.

Sleep and Circadian Development Continues throughout Life

Focusing on the human, it is known that many characteristics of sleep and circadian rhythmicity change from birth through childhood, adolescence, adult, middle age, and old age. Total sleep time (TST) is maximal in the neonate and may reach 17 or 18 h a day, with almost half of that being REM sleep. Circadian organization of sleep is absent in the neonate but gradually develops between the second and third months of life. When circadian organization appears, it is in phase with the environment, so entrainment of the pacemaker occurs before the coupling of the circadian system to sleep control. By age 5 years, TST declines to approximately 10–12 h per day, with only approximately 20% of that being REM sleep. Thus, the major change in sleep amounts over early childhood is the decrease in REM sleep.

Adolescence is a time of significant changes in characteristics of sleep and circadian rhythms. The spectral properties of the NREM sleep EEG show considerable change at this time. The power density in the slow wave band (delta power) is high during childhood and at a maximum at approximately 9 or 10 years. Delta power then declines significantly (approximately 25%) between ages 12 and 14 years. After that, delta power continues to fall slowly and steadily with age. Since delta power is the definitive characteristic of stages 3 and 4 NREM sleep, these stages decrease linearly with age and are practically nonexistent in the elderly. REM sleep amounts, in contrast, show no major change with adolescence and follow a modest, linear decline from an average of approximately 22% in childhood to approximately 18% in old age.

The major circadian changes that occur after infancy are a phase advance of the timing of sleep

onset at adolescence and a gradual dampening of the amplitude of circadian rhythms with advanced age. To a certain extent, this leveling of circadian organization of sleep/wake may be due to declining levels of physical activity during the day. Sleep during the day leads to a lessening of sleep pressure at night, and hence reduced quality of nocturnal sleep.

See also: Circadian Organization in Non-Mammalian Vertebrates; Circadian Regulation in Invertebrates; Circadian Rhythms in Sleepiness, Alertness, and Performance; Entrainment of Circadian Rhythms by Light; Mammalian Sleep and Circadian Rhythms: Flies; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Peripheral Circadian Oscillators.

Further Reading

Blumberg MS, Karlsson KA, Seelke AMH, and Mohns EJ (2005) The ontogeny of mammalian sleep: A response to Frank and Heller (2003). *Journal of Sleep Research* 14: 91–98.

Carskadon MA and Dement WC (2005) Normal human sleep: An overview. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, pp. 13–23. New York: Elsevier.

Davis FC, Frank MG, and Heller HC (1999) Ontogeny of sleep and circadian rhythms. In: Turek FW and Zee PC (eds.) *Regulation of Sleep and Circadian Rhythms*, pp. 19–79. New York: Dekker.

Frank MG and Heller HC (2003) The ontogeny of mammalian sleep: A reappraisal of alternative hypotheses. *Journal of Sleep Research* 12: 25–34.

Frank MG and Heller HC (2005) Unresolved issues in sleep ontogeny: A response to Blumberg et al. *Journal of Sleep Research* 14: 98–101.

Sleep-Dependent Memory Processing

M P Walker, University of California, Berkeley, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

An exciting renaissance is currently under way within the biological sciences, centered on the question of why we sleep, and focusing specifically on the dependence of memory and plasticity on sleep. Although this resurgence is relatively recent in the annals of sleep research, the topic has a surprisingly long history. In the early nineteenth century, the British psychologist David Hartley proposed that dreaming might alter the strength of associative memory links within the brain. Yet it was not until 1924 that Jenkins and Dallenbach performed the first systematic studies of sleep and memory to test Ebbinghaus' theory of memory decay. Their findings showed that memory retention was better following a night of sleep than after an equivalent amount of time awake. However, they concluded that the memory benefit following sleep was passive and resulted from a lack of sensory interference during sleep. They did not consider the possibility that the physiological state of sleep could actively orchestrate these memory modifications.

It is only in the past half century, following the discovery of rapid eye movement (REM) and nonREM (NREM) sleep, that research began testing the hypothesis that sleep, or even specific stages of sleep, actively participated in the process of memory development. In exploring what has become known as sleep-dependent memory processing – and its associated brain basis, sleep-dependent plasticity – we focus herein primarily on evidence in humans (for more extensive reviews covering animal data, see the publications listed in the section 'Further reading').

Delineations and Definitions

We begin our discussion of interactions between sleep and memory by clarifying the complexities that these terms encompass – specifically, sleep states, memory categories, and memory stages.

Sleep States

To begin, it is important to note that the brain does not remain in one single, physiological state across the 24-h day, but instead cycles through periods of differing neural and metabolic activity, associated with distinct biological states, most obviously divided into those of wake and sleep. Sleep has been broadly divided into REM and NREM sleep stages, which alternate across the night in humans in a 90-min cycle (Figure 1(a)). In primates and felines, NREM sleep has been further divided into substages 1–4, corresponding to increasingly deeper states of sleep (Figure 1(a)). The deepest NREM stages, stages 3 and 4, are collectively referred to as 'slow-wave sleep' (SWS), based on a prevalence of low-frequency cortical oscillations in the electroencephalogram (EEG). Dramatic changes in brain electrophysiology, neurochemistry, and functional anatomy accompany these sleep stages, making them biologically distinct from the waking brain, and dissociable from one another. Thus, sleep cannot be treated as a homogeneous state, which either does or does not affect memory. Instead, each sleep stage possesses a set of physiological and neurochemical mechanisms that may contribute uniquely to memory processing and plasticity.

Memory Categories

In the same way that sleep cannot be considered homogeneous, the spectrum of memory categories believed to exist in the human brain, and the processes that create and sustain memory appear equally diverse. Although often used as a unitary term, 'memory' is not a single entity. Human memory has been subject to several different classification schemes, the most popular based on the distinction between declarative and nondeclarative memory (Figure 1(b)). Declarative memory can be considered as the consciously accessible memories of fact-based information (i.e., knowing what). Several subcategories of the declarative system exist, including episodic memory (autobiographical memory for events of one's past) and semantic memory (memory for general knowledge, not tied to specific events). Current neural models of declarative memory formation emphasize the critical importance of structures in the medial temporal lobe, especially the hippocampus, a structure that is thought to form a temporally ordered retrieval code for neocortically stored information, and to bind together disparate perceptual elements of a single event. In contrast, nondeclarative memory is expressed through action and behavior, and includes procedural memory (i.e., knowing how), such as the learning of actions, habits, and skills, as well as implicit learning, and appears to depend on diverse neural anatomies.

Although these categories offer convenient and distinct separations, they rarely operate in isolation in real life. For example, language learning requires a combination of memory sources, ranging from

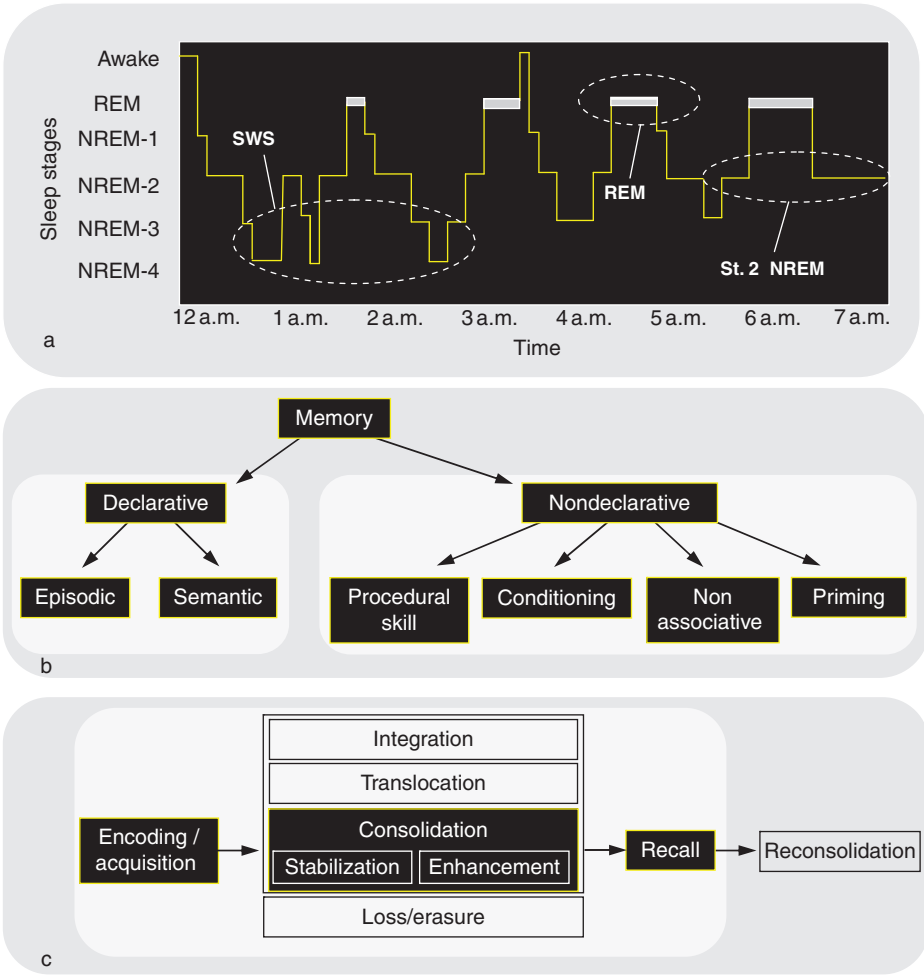

Figure 1 The sleep cycle, memory systems, and memory stages. (a) The human sleep cycle. Across the night, rapid eye movement (REM) and non-REM (NREM) sleep go through a cycle every 90 min in an ultradian manner, while the ratio of NREM to REM sleep shifts. During the first half of the night, NREM stages 3 and 4, slow wave sleep (SWS), dominate, whereas stage 2 NREM and REM sleep prevail in the latter half of the night. Electroencephalogram patterns also differ significantly between sleep stages, with electrical oscillations such as K-complexes and sleep spindles occurring during stage 2 NREM, slow (0.5–4 Hz) delta waves developing in SWS, and theta waves occurring during REM. (b) Memory systems. Human memory is most commonly divided into declarative forms, including episodic and semantic memory, and nondeclarative forms, including an array of different types (e.g., procedural skill memory). (c) Developing stages of memory. Following the initial encoding of a memory, several ensuing stages are proposed, beginning with consolidation and including integration of the memory representation and translocation of the representation, or erasure of the memory. Also, following later recall, the memory representation is believed to become unstable once again, requiring periods of reconsolidation.

nondeclarative memory for procedural motor programs to articulate speech, to memory of grammatical rules and structure, and to aspects of declarative memory for the source of word selection. This too must be kept in mind as we consider the role of sleep in learning and memory.

Memory Stages

Just as memory cannot be considered monolithic, there similarly does not appear to be one sole event that creates or sustains it. Instead, memory appears to develop in several unique stages over time (Figure 1(c)). For example, memories can be initially formed or encoded by engaging with an object or performing an action, leading to the formation of a representation of the object or action within the brain. Following encoding, the memory representation can undergo several subsequent stages of development, the most commonly recognized of which is consolidation. The term 'memory consolidation' classically refers to a process whereby a memory, through the simple passage of time, becomes increasingly resistant to interference from competing or disrupting factors in the absence of further practice . That is to say, the memory becomes more stable.

Recent findings have begun to extend the definition of consolidation. For example, consolidation can be

thought of as not only stabilizing memories, but also as enhancing them – two processes that may be mechanistically distinct. While the stabilization phase of consolidation appears to occur largely across time, the enhancement stage appears to occur primarily, if not exclusively, during sleep, either restoring previously lost memories or producing additional learning, both without the need for further practice. From this perspective, the enhancement phase of memory consolidation causes either the active restoration of a memory that had shown behavioral deterioration, or the enhancement of a memory over its simple maintenance. Thus, consolidation can be expanded to include more than one phase of postencoding memory processing, with each phase occurring in specific brain states such as wake or sleep, or even in specific stages of sleep. Finally, following its initial consolidation, a memory can be retained for days to years, during which time it can be recalled.

Although the focus here is primarily on the effects of sleep on encoding and consolidation, it is important to note that additional postencoding stages of memory processing should also be appreciated. These include the integration of recently acquired information with past experiences and knowledge (a process of memory association), the anatomical reorganization of memory representations (memory translocation), reactivation and reconsolidation of memory through recall, and even the active erasure of memory representations, all of which appear to occur outside of awareness and without additional training or exposure to the original stimuli. It is interesting to note that sleep has already been implicated in all of these steps.

Sleep and Memory Encoding

Some of the first studies to investigate the relationship between sleep and human memory examined the influence of sleep on posttraining consolidation (see later), rather than its influence on initial encoding. However, more recent data have described the detrimental consequence of inadequate pretraining sleep on successful memory encoding.

One of the earliest studies to report the effects of sleep deprivation on declarative memory encoding in humans found that 'temporal memory' (memory for when events occur) was significantly disrupted by a night of pretraining deprivation. These findings have been recently revisited in more rigorous studies, again using the temporal memory paradigm. The task comprised photographs of unknown faces, with the temporal memory component involving recency discrimination, together with a confidence judgment. Significant impairments of temporal memory were evident in a group deprived of sleep for 36 h, which scored significantly lower than did controls. Indeed, significant impairment was evident even in a subgroup that received caffeine to overcome nonspecific effects of lower arousal. Furthermore, the sleep-deprived participants displayed significantly worse insight into their memory-encoding performance.

Neuroimaging reports have begun to investigate the effects of similar periods of total sleep deprivation on verbal memory. As in previous studies, total sleep deprivation resulted in significantly worse acquisition of verbal learning. Surprisingly, however, individuals showed more prefrontal cortex activation during learning when sleep deprived than when not sleep deprived. In contrast, regions of the medial temporal lobe were significantly less activated during learning when sleep deprived. Perhaps most interesting, the parietal lobes, which were not activated during encoding following normal sleep, were significantly activated after sleep deprivation. These findings confirm that sleep deprivation induces a significant behavioral impairment in verbal memory learning, and suggest that these impairments are mediated by a dynamic set of bidirectional changes – overcompensation by prefrontal regions combined with a failure of the medial temporal lobe to engage normally, leading to compensatory activation in the parietal lobes.

We recently investigated the impact of sleep deprivation on declarative memory encoding of both emotional and nonemotional material. Individuals were either sleep deprived for 36 h or allowed to sleep normally prior to an incidental memory encoding session composed of sets of emotionally negative, positive, and neutral words. Following two subsequent nights of sleep, individuals returned for an unexpected recognition task. Overall, individuals in the sleep-deprived condition exhibited a 40% reduction in memory retention relative to individuals who had slept normally prior to encoding (**Figure 2(a)**), results that represent a striking impairment of declarative memory formation under conditions of sleep deprivation. When these data were separated into the three emotional categories (positive, negative, or neutral), the encoding deficit remained, although the magnitude of effect differed across the emotion categories (**Figure 2(b)**). Within the sleep control group, both positive and negative stimuli were associated with superior retention levels relative to the neutral condition, consonant with the notion that emotion facilitates memory encoding. However, there was severe disruption of encoding and hence later memory retention deficit for neutral and especially positive emotional memory in the sleep-deprived group, which exhibited a significant 59% retention deficit relative to the control condition for positive emotional words. Most interesting, however, was the resistance of negative emotional memory

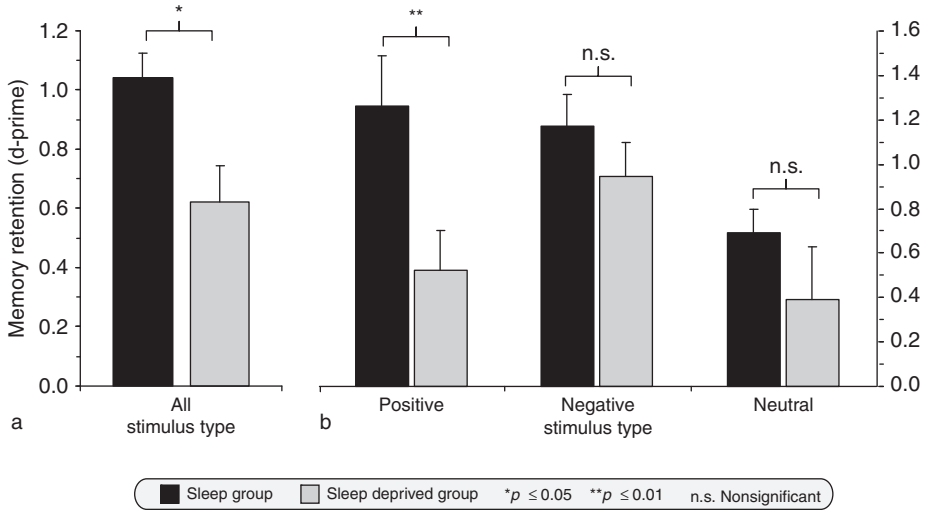

Figure 2 Sleep deprivation and encoding of emotional and nonemotional declarative memory. Effects of 36 h of total sleep deprivation on encoding of human declarative memory (a) when combined across all emotional and nonemotional categories. (b) Effects when separated into emotional (positive and negative valence) and nonemotional (neutral valance) categories.

to sleep deprivation, showing a markedly smaller (19%) and nonsignificant impairment. These data indicate that sleep deprivation severely impairs the encoding of declarative memories, resulting in significantly worse retention 2 days later. Although the effects of sleep deprivation are directionally consistent across subcategories, the most profound impact is on the encoding of positive emotional stimuli and, to a lesser degree, emotionally neutral stimuli, while the encoding of negative stimuli appears more resistant to the effects of prior sleep deprivation.

In summary, sleep deprivation in humans, of varying durations, significantly impairs the ability to effectively form human episodic memories, the effects of which appear be most pronounced for stimuli of positive emotional valence. Furthermore, early evidence suggests that these deficits may be associated with a failure of normal memory encoding networks to respond normally to information acquisition.

Sleep and Memory Consolidation

In addition to the impact of prior sleep deprivation on memory encoding, a plethora of work also demonstrates the critical requirement for sleep after learning for later memory consolidation. Using a variety of behavioral paradigms, evidence of sleep-dependent memory consolidation has now been found in numerous species, including humans, nonhuman primates, cats, rats, mice, and zebra finch.

Declarative Memory

Much of the early work investigating sleep and memory in humans focused on declarative learning tasks. These studies offered mixed conclusions, some in favor of sleep-dependent memory processing and others against it. For example, significant increases in posttraining REM sleep have been observed after intensive foreign language learning, with the degree of successful learning correlating with the extent of REM sleep increase. Such findings suggest that REM sleep plays an active role in memory consolidation, and that posttraining increases reflect a homeostatic response to the increased demands for such consolidation. However, other studies have reported no alteration of posttraining sleep architecture following learning of a verbal memory task. Similar inconsistencies have been reported both for the degree to which training on declarative tasks alters subsequent sleep-stage properties and for the degree of learning impairment that follows selective sleep deprivation. Recently, several studies by Born and his colleagues have shown improvement on a word-pair associates task after SWS-rich early-night sleep, as well as modification of this posttraining sleep. These findings are striking in the face of earlier studies that showed no effect. However, the discrepancy may reflect the nature of the word pairs used. Whereas older studies used unrelated word pairs, such as dog–leaf, Born used related word pairs, such as dog–bone. The nature of the learning task thus shifts from forming and retaining completely novel associations (dog–leaf) to the strengthening or tagging of well-formed associations (dog–bone) for subsequent recall. Thus, sleep's role in declarative memory consolidation, rather than being absolute, might depend on more specific aspects of the consolidation task, such as the degree of semantic association.

A further example of this subtlety pertains to the emotional nature of the material being learned. For example, it has been shown that late-night sleep, rich in REM sleep, selectively favors retention of previously learned declarative emotional texts relative to neutral texts. We have also investigated the time course of emotional and neutral episodic declarative memory consolidation across the day and overnight. Individuals performed an initial study session containing standardized emotional and neutral pictures, either in the evening or morning. Twelve hours later, after sleep or wake, individuals performed a recognition test, discriminating between the original and novel pictures. When the wake and sleep time periods were combined, significantly better emotional memory recognition was observed, relative to neutral memory – consistent with the notion that emotion enhances consolidation (Figure 3(a)). However, when separated according to the wake and sleep periods, a remarkable dissociation was evident (Figures 3(b) and 3(c)). While memory recognition was consistently superior following sleep, consolidation effects were particularly strong for emotional rather than neutral stimuli, with recognition accuracy for emotional pictures improving by 42% overnight, relative to the 12-h waking period. Indeed, emotional memory performance following the wake period was no different from neutral memory in either the wake or sleep conditions. These data indicate the selective facilitation of emotional declarative memory consolidation across sleep, rather than simply across time *per se*, resulting in the enhancement of memory retrieval.

It is important to note that such findings have only begun to test sleep-related effects at the most basic of memory levels – recognition and recall. Several recent studies have now explored more varied measures of memory, demonstrating that the strengths of associative memories are altered in a state-dependent manner. For example, it has been demonstrated that REM sleep provides a brain state in which access to weak associations is selectively facilitated, and flexible, creative processing of new information is enhanced. It has also been demonstrated that, following initial practice on a numeric-sequence problem-solving task, a night of sleep can trigger insight into a hidden rule that can enhance performance strategy the following morning.

Taken as a whole, these studies suggest a rich and multifaceted role for sleep in the processing of human declarative memories. Although contradictory evidence is found for a role of REM sleep in the processing of simple, emotion-free declarative memories, such as the learning of unrelated word pairs, a substantial body of evidence indicates that both SWS and REM sleep contribute to the consolidation of complex, emotionally salient declarative memories, embedded in networks of previously existing associative memories.

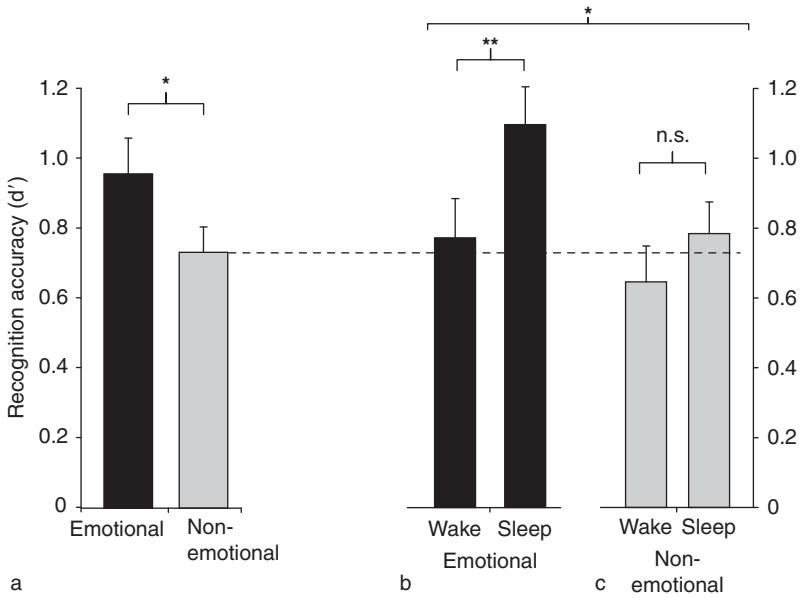

Figure 3 Consolidation of emotional and neutral declarative memory across wake and sleep. Individuals ($n = 14$) entered a repeated-measures design, composed of a sleep phase and a wake phase, presented in counterbalanced cross-over design. During the sleep phase, individuals performed the study session in the evening and 12 h later underwent the recognition test session following a night of sleep. In the wake phase, individuals underwent the study session in the morning and 12 h later performed the recognition test session that evening, without intervening sleep. These two phases were counterbalanced; so that half of the individuals experienced the wake phase first, followed by the sleep phase, while the remaining half experienced the sleep phase first, followed by the wake phase, with the two always being separated by a 1-week interval.

Nondeclarative Procedural Memory

The reliance of procedural, nondeclarative memory on sleep is now a robust finding. These data span a wide variety of functional domains, including both perceptual and motor skills.

Motor learning Motor skills have been broadly classified into two forms – motor adaptation (e.g., learning to use a computer mouse) and motor sequence learning (e.g., learning a piano scale). Beginning with motor sequence learning, we have shown that a night of sleep can trigger significant improvements in speed and accuracy on a sequential finger-tapping task, while equivalent periods of wake provide no significant benefit. These sleep-dependent benefits appear to be specific to both the motor sequence learned and the hand used to perform the task. Furthermore, overnight learning gains correlate with the amount of stage 2 NREM sleep, particularly late in the night (Figures 4(a)–4(c)). This sleep window corresponds to a time when sleep spindles – a defining electrophysiological characteristic of stage 2 NREM – reach peak density. Spindles have been proposed to trigger intracellular mechanisms required for synaptic plasticity, and they increase following training on a motor task. Thus, sleep spindles produced in late-night sleep may trigger key cellular events that in turn initiate mechanisms for neural plasticity.

At the behavioral level, the motor sequence task has been dissected to determine where in the motor program this sleep-dependent improvement occurs. More specifically, differences in transition speeds between each of the separate key-press movements, before and after sleep, were analyzed. For example, in the sequence 4–1–3–2–4, there are four unique key-press transitions: (a) from 4 to 1, (b) from 1 to 3, (c) from 3 to 2, and (d) from 2 to 4 (Figure 5(a)). When individual transition-speed profiles of study participants were analyzed before sleep, the speed of individual key-press transitions within the sequence was unequal (Figure 5(b); open circles), with some transitions seemingly easy (fast) and others problematic (slow), as if the entire sequence was being parsed into smaller subsequences during initial learning (a phenomenon termed 'chunking'). Surprisingly, after a night of sleep, the problematic slow transitions (problem points) were preferentially improved, whereas transitions that had already been effectively mastered prior to sleep did not change (Figure 5(b), filled circles). Most remarkable, however, for participants who were trained and retested after an 8-h waking interval across the day, no such improvement in the profile of key-press transitions, at any location within the sequence, was seen (Figure 5(c)).

These findings suggest that the sleep-dependent consolidation process involves the unification of smaller motor memory units into one single memory element by selectively improving problem regions of the sequence. This overnight process would therefore offer a greater degree of performance automation, effectively optimizing speed across the motor program, and would explain the sleep-dependent improvements in speed and accuracy previously reported. But more importantly, it again suggests that the role of

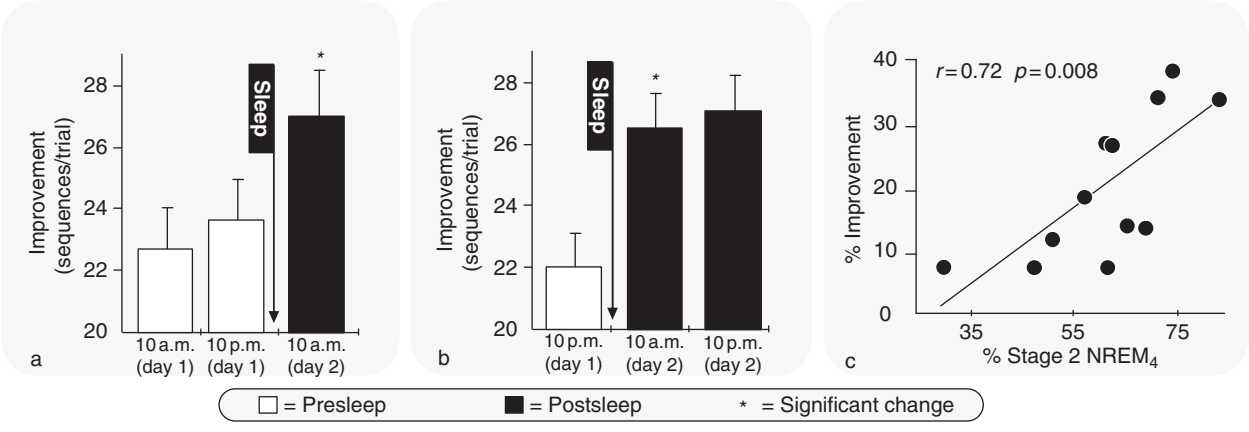

Figure 4 Sleep-dependent motor skill learning. (a) Wake first: after morning training (10 a.m., clear bar), individuals showed no significant change in performance when tested after 12 h of wake time (10 p.m., clear bar). However, when tested again following a night of sleep (10 a.m., filled bar), performance had improved significantly. (b) Sleep first: after evening training (10 p.m., clear bar), individuals displayed significant performance improvements just 12 h after training following a night of sleep (10 a.m., filled bar), yet expressed no further significant change in performance following an additional 12 h of wake time (10 p.m., filled bar). (c) sleep stage correlation; the amount of overnight improvement on the motor skill task correlated with the percentage of stage 2 non-rapid eye movement (NREM) sleep in the last (fourth) quarter of the night (stage 2 $NREM_4$). Asterisks indicate significant improvement relative to training, and error bars indicate standard error of the mean.

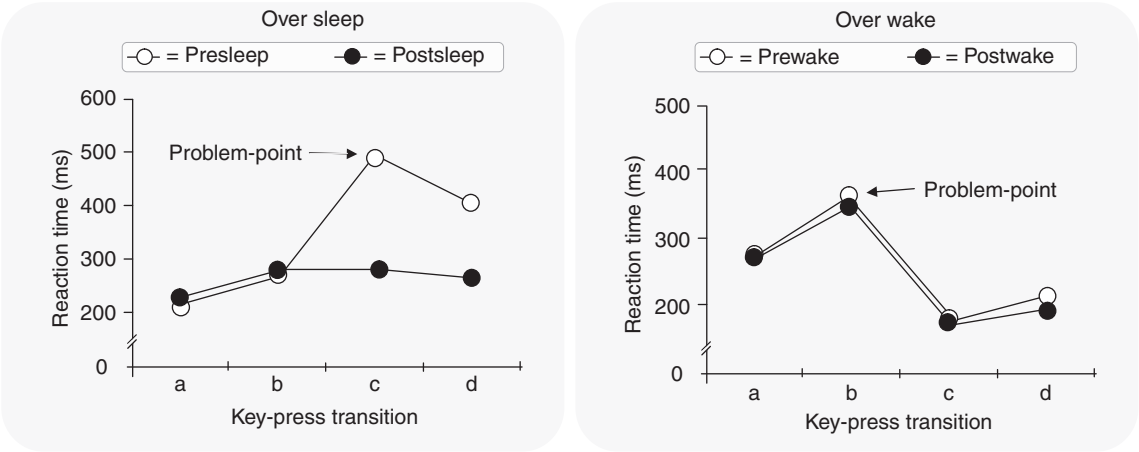

Figure 5 Single-subject examples of changes in transition speeds. Within a five-element motor sequence (e.g., 4–1–3–2–4), there are four unique key-press transitions: (a) from 4 to 1, (b) from 1 to 3, (c) from 3 to 2, and (d) from 2 to 4. (Left) The transition profile at the end of training before sleep (open circles) demonstrated considerable variability, with certain transitions being particularly slow (most difficult; problem points), whereas other transitions appear to be relatively rapid (easy). Following a night of sleep (filled circles), there was a specific reduction (improvement) in the time required for the slowest problem-point transition. (Right) Similarly, at the end of training before a waking interval, transition profiles were uneven (open circles), with some particularly slow transitions (problem points) and other relatively fast transitions (easy). However, in contrast to postsleep changes, no change in transition profile was observed following 8 h of wake (filled circles).

sleep is more subtle and complex than simply increasing the strength of existing memory representations.

Moving from motor sequence learning to motor adaptation learning, it has been demonstrated that selective sleep deprivation impairs retention of a visuomotor adaptation task. All participants were trained and tested on the task and were retested 1 week later. However, some individuals were either completely or selectively deprived of different sleep stages across the first night following memory acquisition. At later retest, individuals deprived of stage 2 NREM sleep showed the most pronounced deficits in motor performance, which again suggests that stage 2 NREM is a crucial determinant of successful motor memory enhancement. Using high-density EEG, more recent studies have been able to show that daytime motor skill practice on a similar task is accompanied by a discrete increase in the subsequent amount of NREM slow-wave EEG activity over the parietal cortex at the start of the night, and that this increase in slow-wave activity is proportional to the amount of delayed learning that developed overnight; individuals showing the greatest increase in slow-wave activity in the parietal cortex that night produced the largest motor skill enhancement the next day.

Visual perceptual learning Learning of a visual texture discrimination task, which does not benefit from 4–12 h of wake following training, has similarly been shown to improve significantly following a night of sleep and appears to require both SWS and REM sleep. For example, overnight improvement has been shown to be specifically sleep dependent, not time dependent (Figure 6(a)), and to correlate positively with the amount of both early-night SWS and late-night REM sleep. Indeed, the product of these two sleep parameters explains more than 80% of interindividual variance (Figure 6(c)). Selective disruption of REM sleep, or of either early sleep (normally dominated by SWS) or late-night sleep (normally dominated by REM and stage 2 NREM), results in a loss of overnight improvement, suggesting that consolidation is initiated by SWS-related processes, but that subsequent REM sleep then promotes additional enhancement. In addition, these delayed performance benefits have been shown to dependent absolutely on the first night of sleep following acquisition (Figure 6(b)).

Taken together, these reports build a convincing argument in support of sleep-dependent learning across several forms of procedural motor and perceptual skill memory. All of these studies indicate that a night of sleep triggers delayed learning, without the need for further training. In addition, overnight skill improvements consistently display a strong relationship to specific sleep stages, and, in some cases, to specific sleep-stage windows at specific times in the night.

Procedural memory and daytime naps Although the majority of sleep-dependent studies have investigated learning across a night of sleep, several reports have begun to examine the benefits of daytime naps on perceptual and motor skill tasks. Based on evidence that motor learning continues to develop overnight,

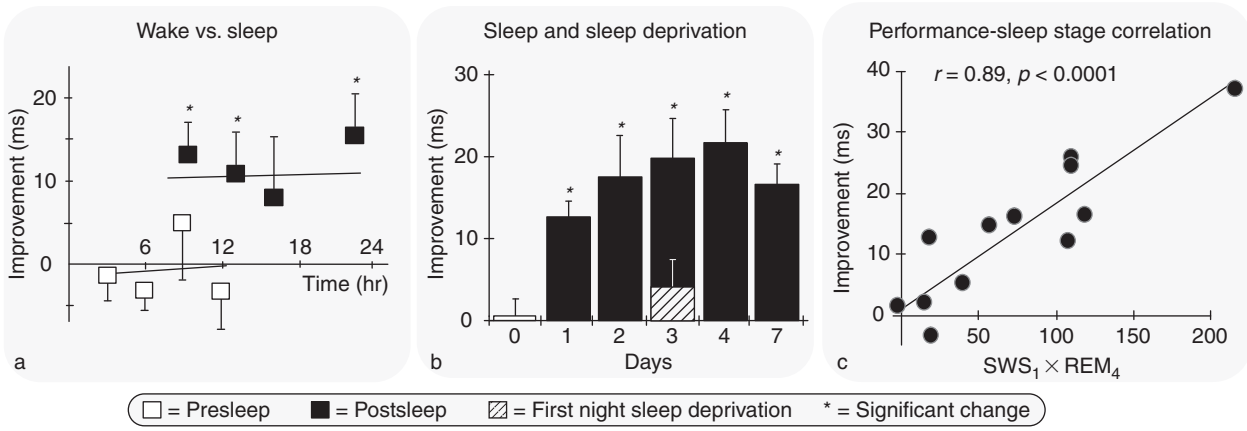

Figure 6 Sleep-dependent visual skill learning. Individuals were trained (during which baseline performance was obtained) and then retested at a later time; the respective improvement in performance (in milliseconds) is illustrated across time. Each individual was retested only once, and each point represents a separate group of individuals. (a) Wake versus sleep: individuals who were trained and then retested on the same day after either 3, 6, 9, or 12 h of subsequent wake (clear squares) showed no significant improvement as a consequence of the passage of waking time across at any of the four time points. In contrast, individuals who were trained and then retested 8, 12, 15, or 23 h after one night of sleep (filled squares) showed a significant improvement occurring as a consequence of sleep. In total, $n = 57$, with $n = 7$–9 for individual points. (b) Sleep and sleep deprivation. Individuals ($n = 89$) who were trained and retested 1–7 days later (filled bars) continued to improve after the first night, without additional practice. Individuals ($n = 11$) sleep deprived the first night after training showed no improvement (crosshatched bar), even after two nights of recovery sleep. (c) Overnight improvement was correlated with the percent of slow-wave sleep (SWS) in the first quarter of the night (SWS_1) and rapid eye movement (REM) sleep in the last quarter of the night (REM_4); $^*p < 0.05$; error bars indicate standard error of the mean.

we have explored the influence of daytime naps on motor sequence learning. Two groups trained on the task in the morning. One group subsequently took a 60- to 90-min midday nap, while the other group remained awake. When retested later that same day, individuals who experienced a 60- to 90-min nap displayed a significant learning enhancement of nearly 16%, while individuals who did not nap failed to show any significant improvement in performance speed across the day (Figure 7). Interestingly, however, when individuals were retested a second time following a subsequent full night of sleep, those individuals in the nap group showed only an additional 7% overnight increase in speed, while individuals in the control group, who had not napped the previous day, displayed speed enhancements of nearly 24% following the night of sleep. These results demonstrate that as little as 60–90 min of midday sleep is sufficient to produce large significant improvements in motor skill performance, while equivalent periods of wake produce no such enhancement. In addition, these data suggest that there may be a limit to how much sleep-dependent motor skill improvement can occur over the course of 24 h, such that napping changes the time course of when learning occurs, but not how much total delayed learning ultimately accrues. Thus, while both groups improved by approximately the same total amount 24 h later (Figure 4), the

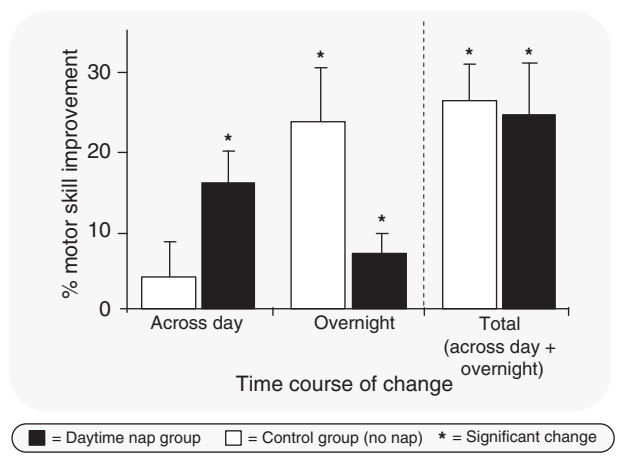

Figure 7 Daytime naps and motor skill learning. Individuals practiced the motor skill task in the morning and either took a 60- to 90-min midday nap or remained awake across the first day. When retested later that same day, individuals who experienced a 60- to 90-min nap (filled bar; across day) displayed significant performance speed improvements of 16%, whereas individuals who did not nap showed no significant enhancements (clear bar; across day). When retested a second time after a full night of sleep, individuals in the nap group showed only an additional 7% increase in speed overnight (filled bar; overnight), whereas individuals in the control group expressed a significant 24% overnight improvement following sleep (clear bar; overnight). Therefore, 24 h later, the groups averaged nearly the same total amount of delayed learning (filled and clear bars; total). Asterisks indicate significant improvement and error bars indicate standard error of the mean.

temporal evolution of this enhancement was modified by a daytime nap.

As with motor skill learning, daytime naps also appear to benefit visual skill learning, although the characteristics of these effects are subtly different. For example, if a visual skill task is repeatedly administered across the day, performance deteriorates rather than remaining stable or improving. This may reflect a selective fatigue of brain regions recruited during task performance, a characteristic not observed in the motor system. However, if a 30- to 60-min daytime nap is introduced among these repeated tests, the performance deterioration is ameliorated. If a longer nap period is introduced, ranging from 60 to 90 min and containing both SWS and REM sleep, performance not only returns to baseline, but is also enhanced. Furthermore, these benefits did not prevent additional significant improvements across the following night of sleep, in contrast to findings for motor skill task performance.

Together these studies build a cohesive argument that daytime naps confer a robust learning benefit to both visual and motor skills and, in the case of visual skill learning, are capable of restoring performance deterioration caused by repeated practice across the day.

Sleep and Brain Plasticity

Memory formation depends on brain plasticity – the brain's lasting structural and/or functional neural changes in response to stimuli (such as experiences). If sleep is to be considered a critical mediator of memory consolidation, then evidence of sleep-dependent plasticity would greatly strengthen this claim. In this section, we consider a mounting wealth of data describing sleep-dependent brain plasticity; our focus here is on neuroimaging studies in humans.

Modification of Posttraining Sleep and Brain Activation

Several studies have investigated whether daytime learning is capable of modifying functional brain activation during subsequent sleep. Based on animal studies, neuroimaging experiments have explored whether the signature pattern of brain activity elicited while practicing a memory task actually reemerges (i.e., is 'replayed') during subsequent sleep. Using positron emission tomography (PET), it has been shown that patterns of brain activity expressed during training on a motor sequence task reappear during subsequent REM sleep, whereas no such change in REM sleep brain activity occur in individuals who received no daytime training. Furthermore, the extent of learning during daytime practice exhibits a positive relationship to the amount of reactivation during REM sleep. As with previously described animal studies, these findings suggest that it is not simply experiencing the task that modifies subsequent sleep physiology, but rather it is the process of learning that does so. Similar findings have been reported using a virtual maze task. Daytime task learning was initially associated with hippocampal activity. Then, during posttraining sleep, there was a reemergence of hippocampal activation, this time specifically during SWS. The most compelling finding, however, was that the amount of SWS reactivation in the hippocampus was proportional to the amount of next-day task improvement, which suggests that this reactivation is functional, and leads to offline memory improvement. Such sleep-dependent replay may potentially modify synaptic connections established within specific brain networks during practice, strengthening some synaptic circuits while potentially weakening others in the endeavor of refining the memory.

Overnight Reorganization of Memory Representations

A second approach, which more directly examines sleep-dependent plasticity, compares patterns of brain activation before and after a night of sleep. In contrast to approaches that measure changes in functional activity during sleep, this technique aims to determine whether next-day improved performance results from an overnight, sleep-dependent restructuring of the neural representation of the memory. Using a sleep-dependent motor skill task, we have recently investigated differences between patterns of brain activation before and after sleep using functional magnetic resonance imaging (fMRI). Following a night of sleep, and relative to an equivalent intervening period of wake, increased activation was identified in motor control structures of the right primary motor cortex (**Figure 8(a)**) and left cerebellum (**Figure 8(b)**) – changes that allow more precise motor output and faster mapping of intention to key-press. There were also regions of increased activation in the medial prefrontal lobe and hippocampus (**Figures 8(c) and 8(d)**), structures recently identified as supporting improved sequencing of motor movements. In contrast, decreased activity postsleep was identified bilaterally in the parietal cortices (**Figure 8(e)**), possibly reflecting a reduced need for conscious spatial monitoring as a result of improved task automation, together with regions of signal decrease throughout the limbic system (**Figures 8(f)–8(h)**), indicating a decreased emotional task burden. In total, these findings suggest that sleep-dependent motor learning is associated with a large-scale plastic reorganization of memory throughout several brain regions, allowing skilled motor

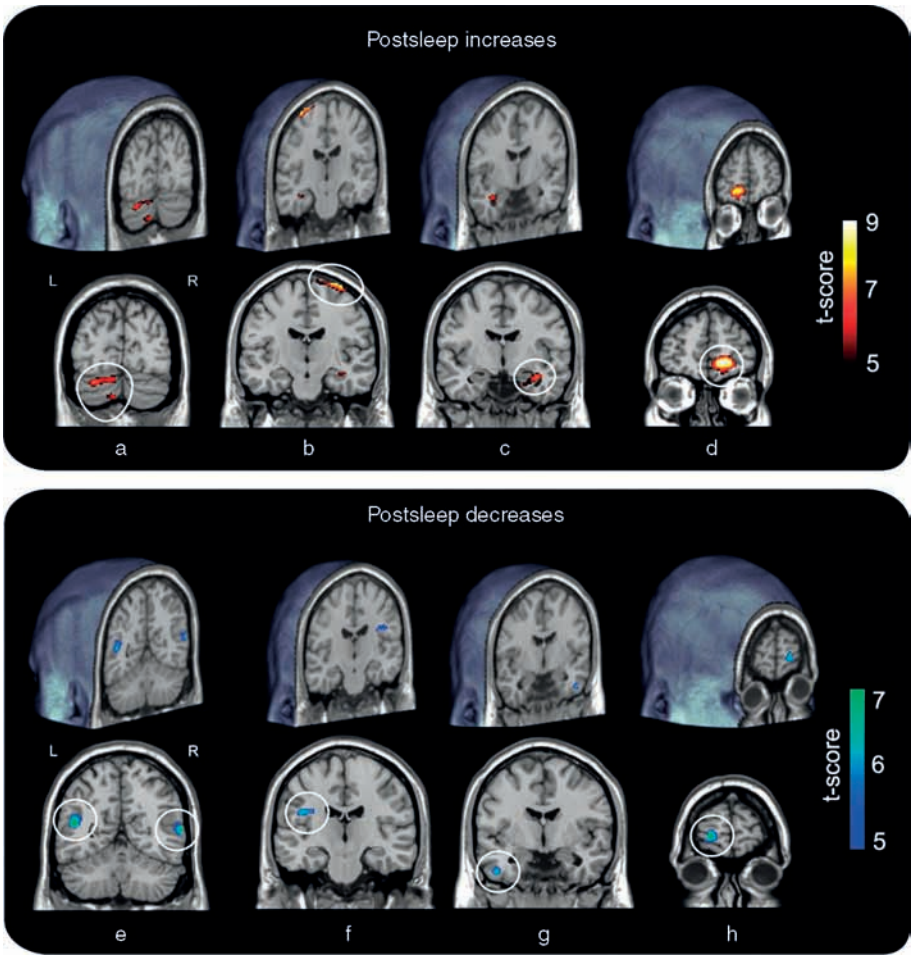

Figure 8 Sleep-dependent motor memory reorganization in the human brain. Individuals were trained on a sleep-dependent motor skill task and then tested 12 h later, either following a night of sleep or following intervening wake, during a functional magnetic resonance imaging brain-scanning session. Scans after sleep and wake were compared (subtracted), resulting in regions showing increased functional magnetic resonance imaging activity postsleep (in red/yellow; a–d) or decreased signal activity (in blue; e–h) postsleep, relative to postwake. Activation patterns are displayed on three-dimensional rendered brains (top row of each graphic), together with corresponding coronal sections (bottom row of each graphic). Following sleep, regions of increased activation were identified in the right primary motor cortex (a), the left cerebellum (b), the right hippocampus (c), and the right medial prefrontal cortex (d). Regions of decreased activity postsleep were expressed bilaterally in the parietal lobes (e), together with the left insula cortex (f), left temporal pole (g), and left frontopolar area (h), all regions of the extended limbic system. All data are displayed at a corrected threshold of $p < 0.05$.

movements to be executed more quickly, more accurately, and more automatically following sleep.

We have also investigated whether overnight reorganization similarly occurs in sensory–perceptual systems using a sleep-dependent visual texture discrimination task, described earlier. Individuals were trained and then tested with or without intervening sleep. Relative to the condition without sleep, testing following sleep was associated with significantly greater activation in an area of primary visual cortex corresponding to the visual target location. However, there were also several other regions of increased postsleep activity throughout both the ventral object recognition (inferior parietal and occipital–temporal junction) and dorsal object location (superior parietal lobe) pathways, together with corresponding decreases in the right temporal pole, a region involved in emotional visual processing. Thus, a night of sleep appears to reorganize the representation not only of procedural motor but also of visual skill memories. Specifically, greater activation throughout the visual processing streams likely offers improved identification of both the stimulus form and its location in space, with signal decreases in the temporal pole reflecting a reduced emotional task burden resulting from the overnight learning benefits.

While the aforementioned studies examine the benefit of a night of sleep, reports have also investigated the detrimental effects of a lack of sleep on underlying brain activity, using a visuomotor adaptation task. Individuals were trained on the task and

tested 3 days later, with half of the individuals deprived of sleep the first night. Controls, who slept all three nights, showed both enhanced behavioral performance at testing and a selective increase in activation in the superior temporal sulcus (a region involved in the evaluation of complex motion patterns), relative to individuals deprived of sleep the first night. In contrast, no such enhancement of either performance or brain activity was observed in these individuals, indicating that sleep deprivation had interfered with a latent process of plasticity and consolidation. These findings indicate that sleep deprivation disrupts not only consolidation, but also the underlying neural mechanisms that support it.

In summary, learning and memory are dependent on processes of brain plasticity, and sleep-dependent learning and memory processing must be mediated by these processes. Many examples of such plasticity across sleep have now been reported over a range of descriptive levels – from whole brain to molecules. As reviewed in this discussion, neuroimaging studies have now identified changes in (1) the functional patterns of brain activity during posttraining sleep periods and (2) the reorganization of newly formed memories following a night of sleep. These plastic brain changes likely contribute to the consolidation of previously formed memory representations, resulting in improved next-day behavioral performance.

Summary

Over the past 25 years, the field of sleep and memory has grown exponentially, with the number of publications per year doubling every 9–10 years, faster than the growth for study of either sleep or memory alone. These reports, ranging from studies of cellular and molecular processes in animals to behavioral studies in humans, have provided converging evidence that pretraining sleep plays a critical role in preparing the brain for initial memory encoding, while posttraining sleep is important in triggering offline memory consolidation and associated neural plasticity. In the end, the question appears not to be whether sleep mediates memory processing, but instead, how it does so. The future of the field is truly exciting, and the challenge to neuroscience will be both to further uncover the mechanisms of brain plasticity that underlie sleep-dependent memory consolidation and to expand our understanding of sleep's role in memory processes beyond simple encoding and consolidation, into the constellation of additional processes which are critical for efficient memory development. Work across the neurosciences will be necessary to answer these questions, but with the current rate of growth of research in the field, we can expect important advances in our understanding of this critical function of sleep. By way of such a multidisciplinary approach, and with a measured appreciation that sleep plays a fundamental role in forming, consolidating, integrating, and reforming memories, we can look forward to new advances in treating disorders of memory, and perhaps even in improving the capacity of memory in healthy individuals.

See also: Autonomic Dysregulation During REM Sleep; Memory Representation; Memory Disorders; Metabolic Syndrome and Sleep; Multiple Memory Systems; Sleep Oscillations and PGO Waves; Sleep and Sleep States: PET Activation Patterns; Sleep and Sleep States: Hippocampus–Neocortex Dialog; Sleep and Sleep States: Network Reactivation; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function.

Further Reading

Dudai Y (2004) The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology* 55: 51–86.

Gais S and Born J (2004) Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proceedings of the National Academy of Sciences of the United States of America* 101: 2140–2144.

Hobson JA and Pace-Schott EF (2002) The cognitive neuroscience of sleep: Neuronal systems, consciousness and learning. *Nature Reviews Neuroscience* 3: 679–693.

Karni A, Tanne D, Rubenstein BS, et al. (1994) Dependence on REM sleep of overnight improvement of a perceptual skill. *Science* 265: 679–682.

Maquet P, Laureys S, Peigneux P, et al. (2000) Experience-dependent changes in cerebral activation during human REM sleep. *Nature Neuroscience* 3: 831–836.

Peigneux P, Laureys S, Fuchs S, et al. (2004) Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron* 44: 535–545.

Stickgold R and Walker MP (2005) Memory consolidation and reconsolidation: What is the role of sleep? *Trends in Neurosciences* 28: 408–415.

Stickgold R, Whidbee D, Schirmer B, et al. (2000) Visual discrimination task improvement: A multi-step process occurring during sleep. *Journal of Cognitive Neuroscience* 12: 246–254.

Wagner U, Gais S, Haider H, et al. (2004) Sleep inspires insight. *Nature* 427: 352–355.

Walker MP, Brakefield T, Morgan A, et al. (2002) Practice with sleep makes perfect: Sleep dependent motor skill learning. *Neuron* 35: 205–211.

Walker MP and Stickgold R (2004) Sleep-dependent learning and memory consolidation. *Neuron* 44: 121–133.

Walker MP and Stickgold R (2006) Sleep, memory and plasticity. *Annual Review of Psychology* 10: 139–166.

Walker MP, Stickgold R, Alsop D, et al. (2005) Sleep-dependent motor memory plasticity in the human brain. *Neuroscience* 133: 911–917.

Walker MP, Stickgold R, Jolesz FA, et al. (2005) The functional anatomy of sleep-dependent visual skill learning. *Cerebral Cortex* 15: 1666–1675.

Sleeping Sickness

M Bentivoglio, University of Verona, Verona, Italy
K Kristensson, Karolinska Institutet, Stockholm, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Sleeping sickness or human African trypanosomiasis (HAT) is a re-emergent disease prevalent in sub-Saharan Africa. It is caused by the hemoflagellate protozoan *Trypanosoma brucei* (*T.b.*, also called African trypanosome), and it is transmitted by bites of flies of the genus *Glossina* (tsetse flies).

There are two clinically distinguishable forms of HAT caused by two different subspecies of African trypanosome and localized to separate geographical areas, although they overlap in many foci of central Africa. The Gambian or West African form of the disease occurs west to the Rift Valley and is caused by *T.b. gambiense*, with humans as primary host and main reservoir. The Rhodesian or East African form of sleeping sickness occurs east to the Rift Valley and is caused by *T.b. rhodesiense*. Rhodesian sleeping sickness is a zoonosis, for which cattle or wild game animals are primary hosts. Both forms are lethal if untreated, but their temporal profile differs. The infection caused by *T.b. gambiense* lasts for several months or even years, whereas that caused by *T.b. rhodesiense* can progress more rapidly in humans. This difference in temporal profile may reflect a relative inefficiency of the tsetse flies as vectors, which renders a long-term infection in the primary host necessary to increase the chances for transmission.

A third subspecies of the parasite, *T.b. brucei*, produces a relatively mild form of disease in native game animals but a severe condition in most domesticated animals. *T.b. brucei* is noninfective for humans, who have a serum protein that inhibits this parasite, but produces a lethal disease in rodents and is therefore used in experimental studies.

Although drugs for treatment of both forms of HAT early during the infection are relatively effective, treatment later during infection, when most of the patients may seek medical advice, still widely employs arsenical-based preparations, which have a high degree of toxicity. Better diagnostic tools for surveillance, early detection of HAT in the field, diagnostic tools for the distinction between early and late stages of the infection, as well as less toxic drugs for the cure of brain infection are therefore of high priority. Understanding the mechanisms by which the severe neurological disturbances develop during the disease is essential to prevent, diagnose, and cure neuroinvasion of the parasites. These disturbances are complex but hallmarked by alterations of the sleep pattern. Such alterations and experimental data on their pathogenesis are the focus of this article.

Historical Notes and Current Prevalence

Sizeable epidemics of sleeping sickness in humans were not identified until the turn of the twentieth century, although the disease had been recognized in West Africa for more than 600 years. A plausible explanation for these outbreaks of the disease is the migration of people caused by the European colonization of the continent. By this migration, individuals carrying the parasites were moved from areas with widely spread infections into novel areas. It has been estimated that half a million people died of sleeping sickness along the Congo River at the time of its exploration between 1896 and 1908. The most renowned outbreak of the disease occurred on the northern shore of Lake Victoria between 1898 and 1908. This epidemic may have killed two-thirds of the population in Uganda and the hunt for its causative agent was therefore intensified. In 1902, Robert M. Forde discovered in the blood of a patient with 'Gambian fever' the presence of parasites that were identified as trypanosomes by Joseph E. Dutton. Aldo Castellani found in 1903 the parasites in the cerebrospinal fluid (CSF) of cases of sleeping sickness in Uganda. This finally determined that the etiological agents of the disease were African trypanosomes. David Bruce and his wife Mary discovered in 1895 that tsetse flies were vectors for infections caused by these parasites in horses and cattle (*nagana*, from the Zulu language meaning 'feeble, useless').

In the following decades, the incidence of sleeping sickness gradually decreased and the disease had almost disappeared from several areas by the 1960s. This was due to surveillance and treatment of infected patients, as well as to tsetse fly control programs. Thus, the disease was almost forgotten during the first postcolonial period when other, and more urgent, problems came into focus. Because of the low level of awareness and due to political instability and civil unrest, a recrudescence of the disease occurred during the late twentieth century, most notably in Angola, R-D Congo, Uganda, and southern Sudan. In certain regions of these countries, the incidence of infection reached the level of the early decades of the century. This recrudescence reflects

shortcomings in surveillance and treatment, increased migration of human populations due to streams of refugees, as well as disruption of tsetse fly control programs.

Sleeping sickness is certainly the most dreadful of sleep diseases in terms of victims. HAT now occurs in approximately 250 foci in 36 sub-Saharan countries. Approximately 50–60 million people live in endemic areas and are therefore at risk, but only approximately 10% are under surveillance. The estimated number of 70 000 new cases detected and treated per year is therefore underscored. According to the World Health Organization (WHO), the actual annual number of cases may be in the range of 300 000–500 000. The disease was estimated to cause a loss of 2.1 million disability-adjusted life-years in 1999.

The Gambian form of the disease presents the major human health problem, whereas the Rhodesian form only occurs in a minority of patients. During the past few years, the situation has stabilized in areas with less political unrest and improved socioeconomic conditions, although the reemergence of sleeping sickness in many foci in Africa in the past few years is alarming. Cases of infections by *T.b. rhodesiense* and *T.b. gambiense* in travelers have been repeatedly reported in recent years, focusing attention on the increased risk of infection and on the need to consider this diagnostic possibility for imported cases.

With the recent awareness of the situation, eradication of the disease during the coming decades should become a realistic objective. To accomplish this, accurate and cost-effective diagnostic methods have to be developed and need to be implemented in remote and impoverished regions, in addition to re-institution of efficient programs for the vector control.

The Parasite and Human Infections

The Parasite

T.b. gambiense is transmitted mostly by the tsetse fly species *Glossina palpalis* and, as mentioned previously, humans are the major host, although domestic animals such as pigs may also be a reservoir. *T.b. rhodesiense* is transmitted mostly by *Glossina morsitans*, with wild game and domestic livestock as reservoir.

The single-celled *T.b.* parasites belong to the Salivarian group of the genus *Trypanosoma* and they complete their development in the proboscis/salivary glands of the tsetse flies. These flies attack and bite moving animals during the day and have a phototactic behavior, being attracted by blue and black surfaces. The blood-sucking female tsetse flies become infected when encountering a host animal with parasites circulating in the blood. In the tsetse fly, the parasites migrate from the hindgut to the salivary glands, in which they are transformed into metacyclic trypanosomes that are ready to be inoculated into a new mammalian host animal during a blood meal. Approximately 5000 metacyclic trypanosomes are inoculated into the dermal connective tissue of the host with the tsetse fly saliva during a feed. At the site of inoculation, an inflammatory reaction will ensue, sometimes forming a 'chancre' (a skin lesion with edema and also cell infiltration and proliferation). From the inoculation site, trypanosomes enter the lymphatics and bloodstream of the host, where the metacyclic forms are transformed into the slender bloodstream form. The slender parasites divide, resulting in high levels of parasitemia which fluctuate in the host in waves lasting in rodents for approximately 3 or 4 days. Troughs in the parasitemia are caused by the humoral immune response of the host.

The three strains of *T.b.* parasite are morphologically and mostly serologically indistinguishable. They are covered on their surface by a single species of variant surface glycoproteins (VSGs) and express a new species of VSG on their surface after each immune attack. By this mechanism, the parasites evade the humoral immune response and can avoid being eliminated in mammals. The VSG genes occupy 10% of the trypanosome genome and the mechanisms for the process of VSG switching well known. The sequence of the 26-megabase genome of *T.b.* was completed in 2005.

Stages of Sleeping Sickness, Treatment, and Neurological Signs

Sleeping sickness progresses through two clinical stages. The parasites first multiply in the lymph and blood to invade peripheral organs. This first (hemolymphatic) stage is followed by spread to the central nervous system. The second (meningoencephalitic) stage leads to the most severe, and ultimately fatal, consequences of the infection, and it requires treatment with drugs which pass the blood–brain barrier (BBB).

According to current WHO criteria, diagnosis of the second stage relies on the occurrence of trypanosomes, more than five white blood cells, and/or increased levels of protein in the CSF. However, these criteria are debated and need to be better validated considering the important therapeutic implications.

Early during infection, pentamidine is the drug of choice for clearance of *T.b. gambiense*, whereas suramin is used against *T.b. rhodesiense*. Both drugs are given parenterally, but formulas for oral drug delivery are currently being sought. Treatment of the second

stage with an arsenical derivative (melarsoprol, the trypanocide introduced in 1949), which crosses the BBB, kills the parasites but in 5–10% of cases causes a reactive arsenical encephalopathy which can be fatal in approximately half of the patients. Another problem is that melarsoprol resistance is increasingly reported, in certain areas in as many as 30% of patients. The relatively new drug difluoromethylornithine (DFMO) is used as an alternative against the West African form of HAT but requires long-term intravenous injections at high doses. Because of this, and also due to side effects and insufficient availability, the utility of DFMO in HAT treatment is limited.

The neurological symptoms and signs are most prominent in the Gambian form of the disease due to its more protracted course. There is considerable variation between afflicted individuals in the neurological picture of the disease, which involves a constellation of sensory, motor, and mental disturbances. Increase in muscle tone and abnormal movements, which include trembling of the hands and fingers, as well as choreiform and athetoid movements, are common. There are no objective sensory defects, but patients usually develop a peculiar pain syndrome in which a slight touch is sufficient to provoke an intense pain episode (Kerandel's sign).

States of mental confusion are common and patients may develop psychiatric disturbances, including emotional instability, indifference, aggression, asocial and stereotypic behavior, impulsive actions, fugue states, manic episodes, melancholia, and delirium. Dementia is common at the end stage, and patients will finally become comatose and die in a state of wasting. The most characteristic sign of sleeping sickness, however, is represented by sleep alterations.

Sleep and Circadian Rhythmicity Disturbances

Interestingly, sleeping sickness is not a hypersomnia, as its name may suggest. The name of the disease derives from the observation that patients develop frequent and sudden sleep episodes during the day. On the other hand, patients have wakefulness episodes during the night. Polysomnographic recordings in HAT patients have demonstrated that the total sleep time and total amount of each sleep stage during 24 h are not increased. Instead, the main alternations during sleeping sickness involve the sleep–wake cycle and the sleep pattern. The sleep–wake alteration tends to occur in short cycles, with sleeping and waking occurring equally during the day and night (Figure 1). The internal structure of sleep is altered, with frequent onset of rapid eye movement (REM) sleep periods and REM sleep episodes occurring soon after long wakefulness transitions with latency shorter than 15 min. Such sleep onset of REM sleep (SOREM) periods without any intermediary non-REM sleep, and with the occurrence of SOREM episodes during wakefulness, are considered characteristic of the disease. However, such alterations have also been documented in narcolepsy. In addition, abnormal electroencephalographic (EEG) events have been recorded in sleeping sickness patients, with slowing of EEG or periodic slow waves during wakefulness and sleep stages.

It has been reported that the sleep–wake disturbances in sleeping sickness are related to the severity of the disease, and that the occurrence of sleep alterations can be used as diagnostic criterion for the second stage. However, it has not been determined experimentally whether the onset of sleep changes corresponds to the brain parenchyma invasion by the parasites or stage I drug treatment failure.

The alterations in sleep and sleep–wake cycle generally ameliorate with arsenical drug treatment, but the recovery of these parameters in surviving patients is extremely slow and can occur months after the interruption of treatment. This indicates that sleep-regulatory structures are not destroyed by neurodegenerative phenomena during the disease, as also assessed experimentally, and that sleep changes are not directly related to the presence of the parasites in the brain parenchyma. Further potential long-term consequences have not been thoroughly investigated.

Loss in the circadian rhythmicity of cortisol and prolactin secretion, as well as loss of growth hormone release, has been recorded. The relationship between plasma hormone pulses and sleep–wake stages is preserved, however. Melatonin secretion maintains its circadian rhythmicity but exhibits a phase advance. Together, these findings, in the presence of a severe alteration of the sleep–wake cycle, point to the complexity of the interrelationships between different circadian rhythms.

Neuropathological Findings

The pathological changes in both forms of sleeping sickness are similar, and the Rhodesian form may be regarded as a compressed Gambian form. The neuropathological picture of the disease is relatively mild compared to its clinical severity. The alterations are mainly inflammatory, without marked neurodegenerative changes, and they are diffuse. The inflammatory changes are predominantly seen in the white matter of the cerebral hemispheres (leucoencephalitis), but areas of demyelination, when present, are confined to perivascular sites. Perivascular infiltration of monocytic inflammatory cells also occurs at several sites, including basal ganglia, brain stem, and cerebellar white matter. There is also an infiltration in the brain of peculiar cells, the so-called Mott cells,

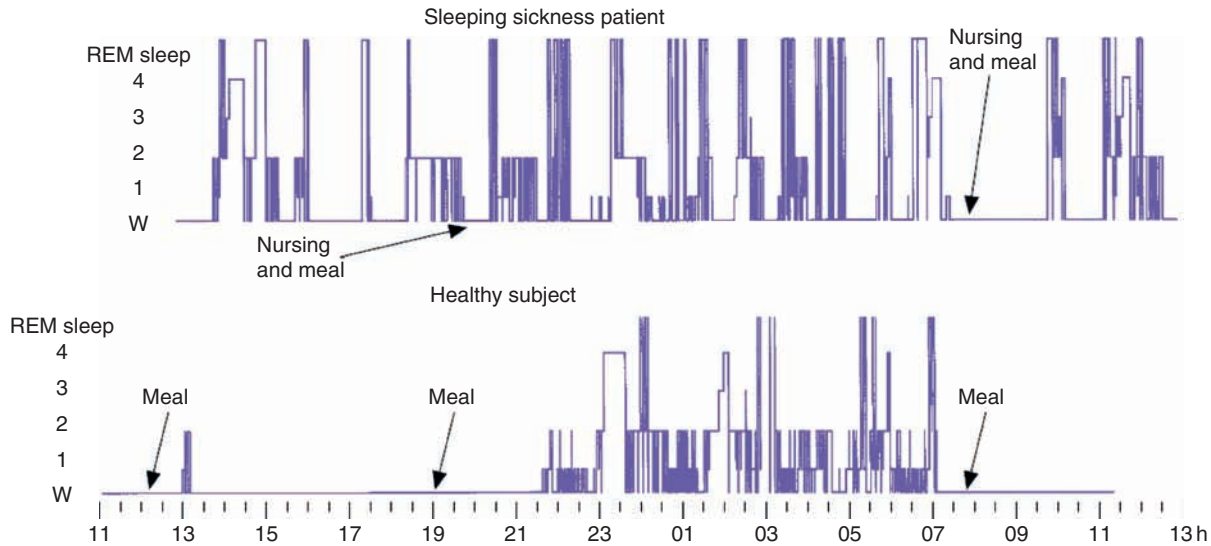

Figure 1 Hypnograms showing the 24 h distribution of wakefulness (W), rapid eye movement (REM) sleep, and stages 1–4 of non-REM sleep in an African patient affected by sleeping sickness (top) and in a healthy African subject (bottom). Hypnograms were recorded in both subjects at bed rest and with an intravenous catheter (due to which the sleep of the healthy subject had many interruptions). Note the disruption of the circadian distribution of sleep and wakefulness in the patient, with frequent occurrence of sleep-onset REM sleep episodes. Reproduced from Buguet A, Bourdon L, Bouteille B, et al. (2001) The duality of sleeping sickness: Focusing on sleep. *Sleep Medicine Reviews* 5: 139–153, with permission.

which are morular shaped and contain IgM; they are most numerous in long-standing cases. Marked activation of microglial cells and widespread astrocytosis in the brain of human victims have been reported in early descriptions of sleeping sickness.

In most of these studies, trypanosomes were not detected in the brain parenchyma. This may reflect a high degree of sensitivity of the parasites to post-mortem autolysis, or it could be due to arsenical drug treatment that may kill the parasites in the brain before patients die.

Experimental Studies and Hypotheses on Pathogenesis of Sleep Disruption during the Disease

The pathogenesis of sleeping sickness is complex and poorly understood. Because the causative agent is an extracellular parasite, neuronal dysregulation causing neurological symptoms should be due to factors released by the parasite, its interactions with the immune system of the host, or both.

In view of laboratory safety conditions, as mentioned previously, most experimental studies have employed *T.b. brucei*.

Neuroinflammation in *Trypanosoma brucei*-Infected Rodents

Experimental studies have revealed a temporal profile of *T.b.* neuroinvasion, which seems to occur stepwise. Early after infection in experimental rat and mouse models, *T.b. brucei* localizes to the choroid plexus and circumventricular organs, which lack a BBB. Later during infection, the parasites can penetrate into the brain parenchyma across the BBB. However, the parasites can cross the BBB with preserved endothelial cell tight junction proteins. This finding supports the clinical observation that drugs that cannot penetrate the BBB are inefficient in the treatment of HAT late during infection. As described later, a major unsolved pathogenetic problem is the identification of factors which, in the parasite and/or the host or their interaction, determine *T.b.* passage through the BBB.

The pathological picture in experimental infections in rodents is neuroinflammatory, as in humans, with only minor signs of neurodegeneration, and the reaction is predominant in the white matter, into which the parasites penetrate. There is a marked activation of astrocytes and microglia, which shows a predilection for certain areas including the hypothalamus, as well as production of cytokines in the brain.

The protagonists of the immune response to the infection (in experimental animals, as well as in humans) are proinflammatory cytokines, including interleukin (IL)-1, tumor necrosis factor (TNF)-α, and interfron (IFN)-γ. TNF-α was isolated in 1985 from *T.b.*-infected rabbits in the search for the cause of cachexia in trypanosomiasis, and it was initially named 'cachectin.' In the brain, cytokines can have an endogenous origin, being released by activated glial cells. The action of proinflammatory molecules,

however, can be counteracted by the release of anti-inflammatory cytokines such as IL-6 and IL-10 and also other extra- and intracellular factors that control inflammation. Free radicals, including nitric oxide, and prostaglandins are also induced in the brain during the infection. In particular, high levels of prostaglandin D2 have been reported in late-stage patients. A prostaglandin released by *T.b.* has also been identified, but its eventual role in sleep changes during the disease has not been examined.

Alterations of Sleep and Endogenous Rhythms in Rodent Models

In rats infected with *T.b. brucei*, the sleep pattern has been found to be markedly fragmented, with major dysregulation of the REM/non-REM sequence of sleep episodes on EEG recordings. Frequent awakenings during slow-wave sleep (SWS) episodes, a reduction of the average length of SWS episodes, and a reduced REM sleep latency have been observed. The sleep structure is marked by abnormal transitions and especially the appearance of SOREM-like episodes. However, the duration of REM sleep episodes is unaffected.

The experimental infections in rodents do not reproduce the full spectrum of the human disease, which mostly develops over a long period of time, whereas the rodent infections allow only a limited survival of a few weeks. However, 24 h EEG recordings have revealed that, in addition to sleep fragmentation, disruption of the sleep–wake cycle can also occur in late stages of the experimental infection in rats.

It should be emphasized that, as in humans, despite the pronounced secretion of somnogenic and pyrogenic cytokines (i.e., IL-1 and TNF-α), neither hypersomnia nor fever are characteristic of the infection in experimental animals. The reason for this paradox is unknown.

Experimental data also point to a disturbance in the melatonin-generating system in the pathogenesis of the disease, with decreased amplitude of the nocturnal peak of melatonin secretion in the urine of *T.b.*-infected rats. The EEG alterations are also ameliorated by treating infected animals with melatonin. However, the locomotor activity and body core temperature rhythms are preserved in *T.b.*-infected rats, but the amplitude of locomotor activity is markedly reduced even early during infection, and a phase advance in the temperature cycles has been noted.

Involvement of Neural Centers That Regulate Circadian Rhythms and Sleep

The sleep disturbances during *T.b.* infection in rodents are in part reminiscent of those observed after lesions of the hypothalamic suprachiasmatic nucleus (SCN), which is the master circadian pacemaker in the mammalian brain. Therefore, functional parameters of this nucleus have been investigated during experimental infections.

The SCN is entrained by environmental light through the retinohypothalamic tract, which utilizes glutamate as a neurotransmitter. The circadian timing system can interact with diencephalic sleep-regulatory centers, whereby circadian information can be transmitted to arousal- and sleep-promoting centers. In the posterior hypothalamus, the histamine-containing tuberomammillary nucleus and the dorsolateral group of neurons containing the peptide orexin/hypocretin (**Figure 2**) have been identified as key relay centers in the regulation of wakefulness. The ventrolateral preoptic nucleus of the anterior hypothalamus (**Figure 2**) has been identified as a sleep-promoting center. These regions have been proposed to be involved in a bistable 'flip-flop' model by which sleep and wakefulness can be switched by reciprocal inhibitory circuits. A decreased resistance in this circuit would weaken the stability of the switch and could therefore result in disruption of the sleep–wakefulness pattern. In particular, the orexin/hypocretin-containing neurons have been suggested to stabilize the switch, causing arousal through an excitatory effect on tuberomammillary neurons directly and by disinhibition.

Experimental data in rodents suggest an involvement of sleep–wakefulness-regulating hypothalamic cell groups during trypanosome infections (**Figure 2**). Marked impairment of the SCN response to a photic stimulus, in the absence of structural changes of retinohypothalamic fibers, was documented in infected rats, and this was accompanied by altered expression of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) glutamate receptor subunits. The amplitude of spontaneous firing in the SCN was also found to be severely altered in *T.b.*-infected rats. Alterations in the synaptic machinery of SCN neurons, particularly disturbances in excitatory neurotransmission, were found in slice preparations of the SCN from infected rats, and these were also caused by treatment with IFN-γ in combination with the endotoxin lipopolysaccharide and TNF-α. Data also indicate a dysregulation of orexin/hypocretin-containing neurons in infected rodents. Furthermore, effects of trypanosome infections on the brain stem serotoninergic raphe nuclei (**Figure 2**), which is also a major sleep-modulatory center, have been suggested.

Together, the findings indicate that due to targeting of trypanosomes in the brain to circumventricular organs close to diencephalic centers and prolonged release of proinflammatory molecules during the infection, trypanosomes could interfere with the function

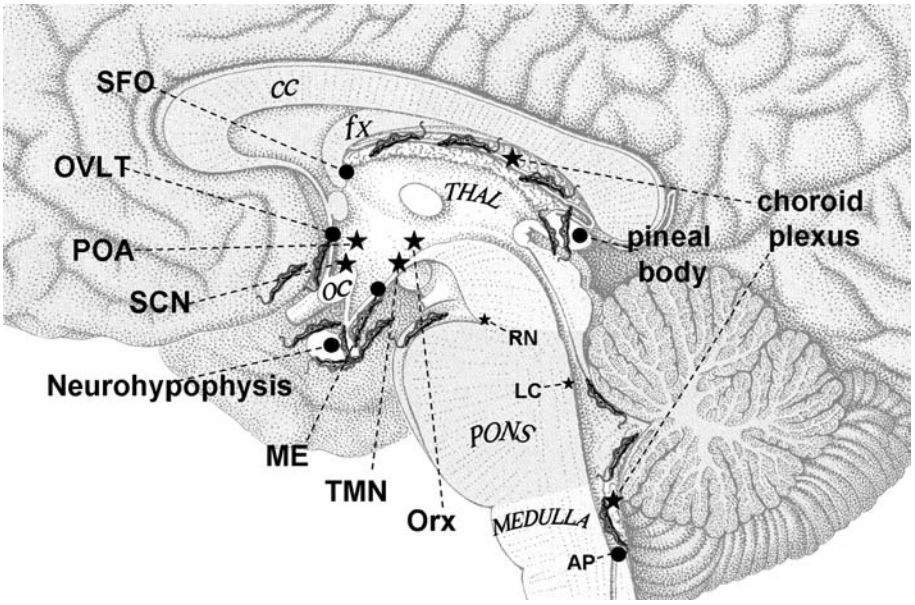

Figure 2 This drawing depicts on the medial surface of the human brain the localization of African trypanosomes at early stages of central nervous system infection, as derived from experimental studies on rodent models. Sitting in circumventricular organs and in the choroid plexus before penetration into the brain parenchyma, the parasites release factors and elicit response of the immune system of the host, with release of inflammatory molecules, including proinflammatory cytokines. Long-term exposure to these molecules could affect neural centers which regulate the sleep–wake alternation and organization. Over time, this could cause dysfunction of circadian timing and sleep-regulatory neuronal assemblies and/or their coupling. The solid circles indicate circumventricular organs; the stars indicate the choroid plexus in the lateral and fourth ventricles and also circadian timing and sleep-regulatory neural centers. AP, area postrema; cc, corpus callosum; fx, fornix; LC, locus coeruleus; ME, median eminence; oc, optic chiasm; Orx, orexin/hypocretin-containing cell groups; OVLT, vascular organ of lamina terminalis; POA, preoptic area; RN, raphe nuclei; SCN, suprachiasmatic nucleus; SFO, subfornical organ; THAL, thalamus; TMN, tuberomammillary nucleus.

of the master circadian pacemaker and sleep-regulatory centers and/or with the coupling between these structures. This neural functional derangement could be built up over time in the molecular interaction between the parasite and the host and/or reach a threshold revealed by the onset of sleep changes in the host.

The Blood–Brain Barrier and Implications for Treatment

As indicated previously, a major problem with regard to therapy of the encephalitis that evolves during sleeping sickness is that the only available drugs which penetrate the BBB are very toxic. Important questions to address are how the parasites cross the BBB, whether there are clinical correlates or markers for parasite penetration into the brain parenchyma, and whether the parasites can hide behind the BBB from the immune response or whether a continuous influx of parasites into the brain is needed to maintain the cerebral infection.

In experimental rodent models, trypanosomes do not move passively into the parenchyma through a disrupted BBB but, rather, follow processes similar to the infiltration of white blood cells (WBCs) into the brain during inflammation. Infiltration of WBCs into the brain can be modulated by cytokines, which also influence *T.b. brucei* neuroinvasion. Thus, the migration of trypanosomes from the lumen of cerebral blood vessels into the brain parenchyma is impeded in mutant mice deficient in IFN-γ, IFN-γ receptor, IL-12p40, and recombinant activating gene-1, despite higher levels of parasitemia in the mutant strains compared to wild-type mice. In the mutant mice, there is a striking accumulation of parasites in the perivascular compartment, confined between the endothelial and the parenchymal basement membranes. These data also indicate that a host immune response molecule, IFN-γ, paradoxically may facilitate penetration of trypanosomes across cerebral blood vessels.

That mechanisms for penetration of parasites into the brain parenchyma are related to those of WBCs is further supported by studies on the effect of the drug minocycline on parasite neuroinvasion. This tetracycline antibiotic impedes the penetration of WBCs into the brain in experimental allergic encephalomyelitis models and, similarly, it impedes neuroinvasion of trypanosomes in *T.b. brucei*-infected mice. Weight loss occurring during infection with *T.b. brucei* is not observed after minocycline treatment of the mice, which also survive longer than nontreated mice.

Due to the treatment hazards of the disease at a late stage, there is an urgent need for better

staging criteria of the infection. For instance, since experimental animals develop sleep pattern changes, a correlation between changes in sleep parameters and passage of trypanosomes across the BBB should be sought. Until there are novel drug discoveries for late-stage disease, early diagnosis and treatment are mandatory to control the infection.

See also: Cataplexy; Narcolepsy; Sleep Apnea; Sleep: Development and Circadian Control.

Further Reading

Bentivoglio M and Kristensson K (2007) Neural-immune interactions in disorders of sleep-wakefulness organization. *Trends in Neuroscienses* 30: 645–652.

Buguet A, Bert J, Tapie P, et al. (1993) Sleep–wake cycle in human African trypanosomiasis. *Journal of Clinical Neurophysiology* 10: 190–196.

Buguet A, Bourdon L, Bouteille B, et al. (2001) The duality of sleeping sickness: Focusing on sleep. *Sleep Medicine Reviews* 5: 139–153.

Dumas M and Bisser S (1999) Clinical aspects of human African trypanosomiasis. In: Dumas M, Bouteille B, and Buguet A (eds.) *Progress in Human African Trypanosomiasis, Sleeping Sickness*, pp. 215–233. Paris: Springer.

Kennedy PGE (2006) Human African trypanosomiasis – neurological aspects. *Journal of Neurology* 253: 411–416.

Kristensson K and Bentivoglio M (1999) Pathology of African trypanosomiasis. In: Dumas M, Bouteille B, and Buguet A (eds.) *Progress in Human African trypanosomiasis, Sleeping Sickness*, pp. 157–181. Paris: Springer.

Lundkvist GB, Kristensson K, and Bentivoglio M (2004) Why trypanosomes cause sleeping sickness. *Physiology* 19: 198–206.

Masocha W, Robertson B, Rottenberg ME, Mhlanga JDM, Sorokin L, and Kristensson K (2004) Cerebral vessel laminins and IFN-γ define *Trypanosoma brucei brucei* penetration of the blood–brain barrier. *Journal of Clinical Investigation* 114: 689–694.

Mhlanga JDM (1996) Sleeping sickness: Perspectives in African trypanosomiasis. *Science Progress* 79: 183–214.

Wang CC (1995) Molecular mechanisms and therapeutic approaches to the treatment of African trypanosomiasis. *Annual Review of Pharmacology & Toxicology* 35: 93–127.

Welburn SC, Coleman PG, Maudlin I, Fèvre EM, Odiit M, and Eisler MC (2006) Crisis, what crisis? Control of Rhodesian sleeping sickness. *Trends in Parasitology* 22: 123–128.

Sleep–Wake State Regulation by Acetylcholine

B E Jones, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Background

Since early histochemical and pharmacological studies in the 1960s, acetylcholine (ACh) has been known to comprise a key element of the activating systems in the brain. The catabolic enzyme acetylcholinesterase (AChE) was found to be contained in fibers which ascended from the brain stem reticular formation into the forebrain and from the basal forebrain into the cerebral cortex in what Shute and Lewis called the 'cholinergic reticular activating system.' This designation advanced the concept that the neurons which utilized ACh as a neurotransmitter formed the principal component of the 'ascending reticular activating system' previously delineated by Moruzzi and Magoun and colleagues in the 1940s and 1950s and shown to be critical for the cortical activation of waking. Enhancement of synaptic ACh through drugs such as physostigmine, which inhibits AChE, stimulated fast cortical activity on the electroencephalogram (EEG), which is indicative of cortical activation. The ACh receptor agonists, nicotine and muscarine, were long known as stimulants and also found to elicit cortical activation. Conversely, blocking nicotinic or muscarinic ACh receptors diminished fast cortical activity. Yet interestingly, atropine, which blocks muscarinic ACh receptors, produced slow wave activity on the cortex but did not prevent behavioral activity and arousal, and thus led to a disassociation of the normal EEG activity and behavior of waking. This dissociation revealed the selective role of ACh in stimulating cortical activation. Indeed, ACh was subsequently found to stimulate cortical activation during the behavioral quiescence and immobility that occurs during the state of rapid eye movement (REM) sleep (also called paradoxical sleep (PS); Figure 1). The cholinergic systems are thus well known to play an important role in cortical activation that occurs during both waking and REM sleep. Conditional upon the activity of other systems, they can also promote either waking or REM sleep states.

Cholinergic Systems in the Brain

The two major groups of cholinergic neurons with ascending projections are located in the pontomesencephalic tegmentum and the basal forebrain (Figure 1). They have been identified and localized definitively by employing immunohistochemical staining for choline acetyl transferase (ChAT), the synthetic enzyme for ACh. The enzyme is found in select neurons that synthesize ACh for use as a neurotransmitter and were thus presumed to be 'cholinergic.' As originally visualized by Shute and Lewis, AChE, the catabolic enzyme, is contained in many more neurons, including monoaminergic and glutamatergic reticular and thalamic neurons, which are likely target neurons of the cholinergic cells. Nonetheless, ChAT immunohistochemistry revealed the presence of cholinergic neurons within the brain stem reticular formation and basal forebrain as well as the presence of cholinergic fibers ascending through the brain and from the basal forebrain up to the cerebral cortex in parallel with other ascending systems to confirm the essential characterization of a cholinergic reticular activating system. The presence of the vesicular transporter protein for ACh (VAChT), which permits storage and synaptic release of ACh, in the brain stem and forebrain ChAT-immunostained cell bodies and terminals confirmed the use of ACh as a neurotransmitter by these accordingly established 'cholinergic' neurons.

The pontomesencephalic cholinergic neurons are located within the laterodorsal tegmental (LDTg) and pedunculopontine tegmental (PPTg) nuclei (Figure 1). Together with other neurons of the reticular formation (which predominantly utilize glutamate (Glu) as a neurotransmitter), these cholinergic cells send ascending projections dorsally up to the thalamus and ventrally to the hypothalamus and, to a lesser extent, the basal forebrain. In the thalamus, they innervate some of the specific thalamocortical relay nuclei (particularly the medial and lateral geniculate nuclei) but innervate most prominently and ubiquitously the neurons of the nonspecific thalamocortical projection system (the midline and intralaminar nuclei, which, like all thalamic projection neurons, utilize Glu), which in turn project in a widespread manner to the cerebral cortex to have the capacity to stimulate widespread cortical activation. Either through collaterals of the forebrain projecting cells or by unique projections, the pontomesencephalic cholinergic neurons also project into the brain stem reticular formation, where they can influence reticular neurons with ascending, local, or descending projections, which are important for REM sleep generation.

The basal forebrain cholinergic neurons are located within the medial septum and diagonal band of Broca (MS-DBB) rostrally and the magnocellular preoptic nucleus, substantia innominata, and globus pallidus (MCPO-SI-GP) more caudally (Figure 1). These two major collections of cells provide prominent innervation to the hippocampus and paleocortex (MS-DBB)

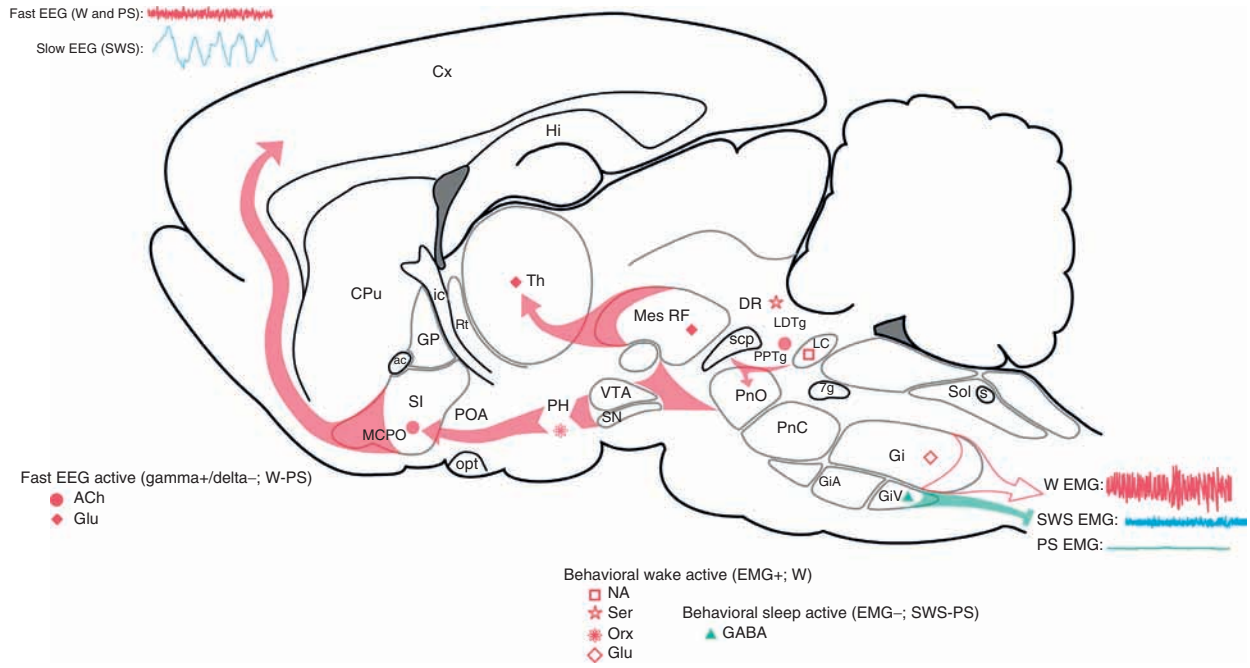

Figure 1 Cholinergic and other neurons involved in sleep–wake state control. Sagittal schematic view of the rat brain depicting neurons with their chemical neurotransmitters and pathways by which they influence cortical activity or behavior across the sleep–wake cycle. Wake (W) is characterized by fast gamma activity on the cortical electroencephalogram (EEG; top left) and high postural muscle tone on the neck electromyogram (EMG; bottom right), slow wave sleep (SWS) is characterized by slow delta EEG and low tone on the EMG, and paradoxical sleep (PS) is characterized by fast gamma EEG and atonia on the EMG. Neurons that are active during waking (red symbols) include cells with ascending projections toward the cortex which stimulate fast cortical activity and cells with descending projections toward the spinal cord which stimulate postural muscle tone and behavioral arousal. The cholinergic (ACh) neurons and other glutamatergic (Glu) neurons which have predominantly ascending projections discharge in association with fast EEG activity (gamma+) and cease firing with delta activity (delta-) to be active during W and PS (W-PS; solid red symbols). The serotonergic (Ser), noradrenergic (NA), orexinergic (Orx), and some putative glutamatergic (Glu) neurons which have more diffuse or descending projections discharge in association with behavioral arousal and EMG activity (EMG+) and cease firing with muscle atonia to be active during W and silent during PS (W; open red symbols). Neurons that are active during sleep (aqua symbols) include cells with descending projections into the brain stem and spinal cord which diminish behavioral arousal and muscle tone. Important among such cells are GABAergic (and glycinergic) neurons in the ventral medulla which project to the brain stem and spinal cord, where they can inhibit motor neurons during sleep. The cholinergic neurons have the capacity to stimulate cortical activation during waking and PS and also to promote PS with muscle atonia by acting in part upon such GABAergic neurons, when the monoaminergic and orexinergic neurons are silent. 7g, genu seventh nerve; ac, anterior commissure; ACh, acetylcholine; CPu, caudate putamen; Cx, cortex; DR, dorsal raphe nucleus; Gi, gigantocellular RF; GiA, gigantocellular, alpha part RF; GiV, gigantocellular, ventral part RF; Glu, glutamate; GP, globus pallidus; Hi, hippocampus; ic, internal capsule; LC, locus coeruleus nucleus; LDTg, laterodorsal tegmental nucleus; Mes RF, mesencephalic RF; MCPO, magnocellular preoptic nucleus; NA, noradrenaline; opt, optic tract; Orx, orexin; PH, posterior hypothalamus; PnC, pontine, caudal part RF; PnO, pontine, oral part RF; POA, preoptic area; PPTg, pedunculopontine tegmental nucleus; PS, paradoxical sleep (also called REM sleep); RF, reticular formation; Rt, reticularis nucleus of the thalamus; s, solitary tract; scp, superior cerebellar peduncle; Ser, serotonin; SI, substantia innominata; SN, substantia nigra; Sol, solitary tract nucleus; Th, thalamus; VTA, ventral tegmental area. Modified from Jones BE (2005) From waking to sleeping: Neuronal and chemical substrates. *Trends in Pharmacological Sciences* 26: 578–586, with permission from Elsevier.

and to the neocortex (MCPO-SI-GP). Projecting in a topographically organized, though widespread manner, the basal forebrain cholinergic neurons thus have the capacity to stimulate widespread cortical activation through both paleo- and neocortex.

Cholinergic Systems and Cortical Activation of Waking

The pontomesencephalic cholinergic neurons send their major projections into the thalamus by which they can excite thalamocortical neurons to in turn stimulate cortical activation. Release of ACh in the thalamus is maximal in association with fast cortical activity during waking and during REM sleep, indicating that the thalamic projecting cholinergic cells discharge during both waking and REM sleep. Acting on both nicotinic and muscarinic ACh receptors, ACh depolarizes and excites thalamocortical projection neurons. It thereby prevents their slow bursting discharge, which occurs during slow wave sleep (SWS), and stimulates their fast tonic discharge, which occurs during, while also stimulating, fast cortical activity of waking or REM sleep.

The basal forebrain cholinergic cells project directly to the hippocampus and neocortex, where they act on

cortical neurons to alter their discharge. Release of ACh from hippocampus and neocortex is high in association with fast cortical activity during both waking and REM sleep. ACh prevents the slow bursting discharge of cortical neurons, which occurs during SWS, and stimulates the fast tonic discharge that occurs during waking and REM sleep. ACh's action in the cortex occurs largely through muscarinic receptors of different types on pyramidal cells and interneurons. Some interneurons are hyperpolarized through muscarinic ACh receptors, whereas others, such as pyramidal cells, are depolarized and excited through muscarinic and nicotinic ACh receptors. Overall, however, ACh and the basal forebrain cholinergic neurons stimulate fast cortical activity particularly in the gamma frequency range (30–60 Hz), which is maximal during attentive or active waking and REM sleep. ACh also stimulates theta activity (4–10 Hz), which is similarly maximal during attentive or active waking and REM sleep. Indeed, it may stimulate these in parallel since the high-frequency gamma activity rides on the slow rhythmic theta waves. Local injections of glutamate or other agonists acting on the cholinergic cells in the basal forebrain evoke theta with enhanced gamma activity. The basal forebrain cholinergic neurons can modulate this rhythmic cortical activity by their discharge, which, as established for identified ChAT-immunostained cells, occurs in rhythmic bursts with theta along with high gamma activity during active waking and REM sleep (Figure 2(a)).

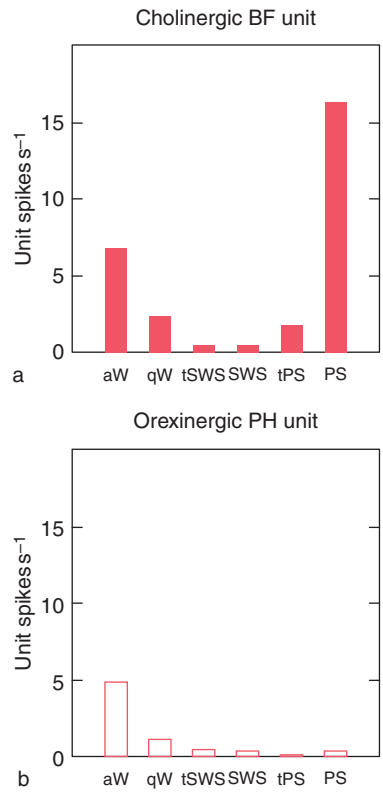

Figure 2 Discharge cholinergic and orexinergic units across sleep–wake states. Average discharge rate (spikes s^{-1}) across the sleep–wake states and transitions (t) of a cholinergic cell (a) in basal forebrain (BF; within MCPO-SI, identified by juxtacellular labeling with neurobiotin and immunohistochemical staining for ChAT) and an orexinergic neuron (b; in the posterior hypothalamus (PH) within the perifornical area, identified by juxtacellular labeling with neurobiotin and immunohistochemical staining for orexin). ACh neurons discharge in association with cortical activation (fast EEG active, gamma+/Delta−) during wake and paradoxical sleep (W-PS; solid red symbols in Figure 1), whereas orexinergic neurons discharge, like NA and Ser neurons, in association with behavioral activity and postural muscle tone (behavioral wake active, EMG+) during W (open red symbols in Figure 1). The discharge of cholinergic neurons when orexinergic and monoaminergic neurons are silent leads to PS. aW, active wake; qW, quiet wake; tSWS, transition to slow wave sleep; SWS, slow wave sleep; tPS, transition to paradoxical sleep; PS, paradoxical sleep (also called REM sleep). Data from Lee MG, Hassani OK, Alonso A, and Jones BE (2005) Cholinergic basal forebrain neurons burst with theta during waking and paradoxical sleep. *Journal of Neuroscience* 25: 4365–4369; and Lee MG, Hassani OK, and Jones BE (2005) Discharge of identified orexin/hypocretin neurons across the sleep–waking cycle. *Journal of Neuroscience* 25: 6716–6720.

Cholinergic Systems and REM Sleep

Following identification of the cholinergic neurons by ChAT immunohistochemistry, it became apparent that the major cell groups in the LDTg and PPTg were located in a region previously shown by Jouvet to be important in the generation of REM sleep, named by him 'paradoxical sleep.' Indeed, complete transections of the brain stem through the caudal midbrain which allowed the persistence of PS in the periphery (marked by postural muscle atonia on electromyogram) would have left intact the pontomesencephalic cholinergic cells innervating the brain stem (Figure 1). In later studies, neurotoxic lesions destroying the pontomesencephalic cholinergic neurons resulted in the loss of PS, including the phasic pontogeniculo-occipital (PGO) spikes accompanying REM sleep and tonic muscle atonia of that state. It had also been shown in both early and later studies that local injections of the cholinergic agonist carbachol into the pontine tegmentum could elicit PS, including PGO spikes with REM, theta activity, and muscle atonia.

On the other hand, the pharmacological elicitation of REM sleep through cholinergic agonists was not always automatic. AChE inhibitors such as physostigmine commonly stimulated cortical activation with wakefulness in otherwise untreated normal animals or humans. However, in early studies it was found that if physostigmine was administered following depletion of the monoamines (with reserpine), it elicited REM sleep. By these pharmacological studies, it was evident that the occurrence of REM sleep depends on

the activity of cholinergic neurons, but it also depends on the inactivity of monoaminergic neurons, a principle which was subsequently formalized in a model of reciprocal interaction by McCarley and Hobson.

Single unit recording studies indicated that whereas many neurons in the pontomesencephalic tegmentum discharged robustly during REM sleep, including those in the region of the cholinergic cells, the presumed serotonergic neurons in the raphe and presumed noradrenergic neurons in the locus coeruleus ceased firing during REM sleep (Figure 1). Although definite identification of recorded neurons awaits absolute confirmation of these principles, c-Fos expression, which reflects neural discharge, was found to be maximal in identified cholinergic pontomesencephalic neurons and minimal in serotonergic and noradrenergic neurons during REM sleep. Finally, ACh release is high during REM sleep in both the forebrain and the brain stem, whereas the release of the monoamines is minimal during REM sleep. Converging lines of evidence thus indicate that cholinergic neurons trigger REM sleep if monoaminergic neurons are silent. ACh can act in different ways on various brain stem and spinal neurons, including GABAergic (and glycinergic) neurons which can inhibit motor neurons, such as to provoke decreased activity and muscle tone. In contrast, serotonin and noradrenaline can stimulate behavioral arousal along with postural muscle tone through excitatory actions on brain stem and spinal neurons, including reticular glutamatergic and motor neurons.

It has also been revealed that hypothalamic neurons containing the peptide orexin (or hypocretin) normally stimulate behavioral arousal with postural muscle tone in a critical manner since in their absence narcolepsy with cataplexy results in humans and animals. Like the presumed monoaminergic neurons and in contrast to the cholinergic neurons, the orexin neurons fire selectively during waking and are silent during REM sleep (Figures 1 and 2(b)). Like monoaminergic neurons, orexin neurons would thus also oppose cholinergic neurons in the generation of REM sleep, although they can work together with cholinergic neurons to stimulate behavioral arousal together with cortical activation during waking.

Cholinergic Systems, Sleep, and Memory

The importance of ACh in memory processes has been known since early pharmacological studies showed serious decrements in memory consolidation following administration of the muscarinic ACh receptor antagonists atropine or scopolamine. Both short- and long-term memory processes have been shown to be enhanced by ACh. ACh enhances long-term potentiation (LTP) and stimulates theta, which in turn enhances LTP and plasticity. Stimulation of the basal forebrain has been shown to promote learning and plasticity. It thus follows that states with heightened ACh release and maximal cholinergic unit discharge would be associated with enhanced learning and memory. These states are active or attentive waking and REM sleep. Although it has long been known that learning is maximal during aroused or attentive waking, it has also been shown in multiple studies that memory consolidation for implicit and explicit or particularly procedural, emotional, and spatiotemporal elements is enhanced during REM sleep.

See also: Cholinergic Pathways in CNS; Dopamine Control of Arousal; Sleep and Sleep States: Thalamic Regulation; Sleep and Sleep States: Hypothalamic Regulation; Sleep and Sleep States: Histamine Role; Sleep and Sleep States: Cytokines and Neuromodulation; Sleep-Dependent Memory Processing; Sleep–Wake State Regulation by Noradrenaline and Serotonin.

Further Reading

Jones BE (2004) Paradoxical REM sleep promoting and permitting neuronal networks. *Archives Italiennes de Biologie* 142: 379–396.

Jones BE (2005) From waking to sleeping: Neuronal and chemical substrates. *Trends in Pharmacological Sciences* 26: 578–586.

Lee MG, Hassani OK, Alonso A, and Jones BE (2005) Cholinergic basal forebrain neurons burst with theta during waking and paradoxical sleep. *Journal of Neuroscience* 25: 4365–4369.

Lee MG, Hassani OK, and Jones BE (2005) Discharge of identified orexin/hypocretin neurons across the sleep–waking cycle. *Journal of Neuroscience* 25: 6716–6720.

McCormick DA (1992) Neurotransmitter actions in the thalamus and cerebral cortex and their role in neuromodulation of thalamocortical activity. *Progress in Neurobiology* 39: 337–388.

Nishino S and Mignot E (1997) Pharmacological aspects of human and canine narcolepsy. *Progress in Neurobiology* 52: 27–78.

Rauchs G, Desgranges B, Foret J, and Eustache F (2005) The relationships between memory systems and sleep stages. *Journal of Sleep Research* 14: 123–140.

Sarter M and Bruno JP (2000) Cortical cholinergic inputs mediating arousal, attentional processing and dreaming: Differential afferent regulation of the basal forebrain by telencephalic and brainstem afferents. *Neuroscience* 95: 933–952.

Siegel JM (2004) The neurotransmitters of sleep. *Journal of Clinical Psychiatry* 65(supplement 16): 4–7.

Steriade M (2004) Acetylcholine systems and rhythmic activities during the waking–sleep cycle. *Progress in Brain Research* 145: 179–196.

Stickgold R, Hobson JA, Fosse R, and Fosse M (2001) Sleep, learning, and dreams: Off-line memory reprocessing. *Science* 294: 1052–1057.

Weinberger NM (2003) The nucleus basalis and memory codes: Auditory cortical plasticity and the induction of specific, associative behavioral memory. *Neurobiology of Learning and Memory* 80: 268–284.

Zeitzer JM, Nishino S, and Mignot E (2006) The neurobiology of hypocretins (orexins), narcolepsy and related therapeutic interventions. *Trends in Pharmacological Sciences* 27: 368–374.

Sleep–Wake State Regulation by Noradrenaline and Serotonin

B E Jones, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Background

The monoamines, noradrenaline (NA) and serotonin (5-HT), were found to have the capacity to influence sleep–wake states in the 1960s through early pharmacological and histochemical, combined with lesion, studies. Particularly through the work of Jouvet with his colleagues, evidence was presented for an important role of NA and the major NA-containing locus coeruleus (LC) nucleus in wakefulness and for a role of 5-HT and the 5-HT-containing raphe nuclei in sleep, particularly slow-wave sleep (SWS). Pharmacological depletion of NA (with inhibition of its synthesis by α-methyl-paratyrosine (AMPT)) and lesions of the ascending pathway from the LC led to decreased wakefulness. Conversely, depletion of 5-HT (with inhibition of its synthesis by para-chloro-phenylalanine (PCPA)) and lesions of the raphe, including the dorsal raphe (DR) nucleus, led to decreased SWS. These early studies and conclusions indicating opposite roles for the noradrenergic and serotonergic systems in sleep–wake state control have since been further scrutinized by finer techniques, including single-unit recording, and revised to reveal different – but not opposite – roles of these systems in the sleep–wake cycle.

Projections and Role of Noradrenergic Locus Coeruleus Neurons in Arousal

The LC is a collection of NA-containing neurons located in the periventricular region of the rostral pons, which can be likened to a sympathetic ganglion because from it emerge multiple major projection pathways of varicose and highly collateralized axons that reach out to innervate the entire central nervous system (Figure 1). Although the neurons are somewhat segregated according to their predominant projections (such that the most dorsal cluster projects primarily to the thalamus and lateral cortex; the more intermediate projects to the hypothalamus, basal forebrain, and medial cortex; and the most ventral cluster projects to the brain stem and spinal cord), a single neuron can project through multiple collaterals to all regions. This highly diffuse noradrenergic system is thus uniquely organized to simultaneously influence the entire central nervous system in a way that could alter its fundamental mode of operation, as occurs in a change of state from sleeping to waking. It was indeed thus proposed in early studies that the noradrenergic LC neurons formed the most important contingent of an ascending activating system of the brain stem. In support of this concept, one of the major stimulants, amphetamine, was found to act by release of NA and also dopamine (DA). The early lesion studies involving multiple systems did diminish wakefulness and cortical activation which were thus proposed to depend on the noradrenergic LC innervation of the cortex. However, subsequent more selective lesions or destruction by neurotoxins (6-hydroxydopamine (6-OHDA)) of the noradrenergic system did not result in the same loss of cortical activation and waking as did the earlier nonselective lesions. These results thus indicated that, although the noradrenergic neurons have the capacity to stimulate cortical activation and wakefulness, they are not essential for these functions. They represent, instead, one of multiple partially redundant, although differentiated, arousal systems in the brain, which include most notably the acetylcholine (ACh)-releasing cholinergic neurons of the brain stem and basal forebrain and the orexin (Orx, or hypocretin) neurons of the posterior hypothalamus (Figure 1).

The NA LC neurons discharge during active waking, gradually decrease their firing during SWS, and cease firing during rapid eye movement (REM) sleep, also called paradoxical sleep (PS) (Figure 2(a)). They accordingly discharge during active or attentive waking associated with behavioral arousal and the presence of postural muscle tone recorded on the electromyogram (EMG; Figure 1). They discharge during orientation and response to sensory stimuli. They also respond to physiological stimuli which activate the sympathetic nervous system and situations which are associated with stress. Given their diffuse projections and actions, the noradrenergic neurons can simultaneously stimulate fast cortical activity on the electroencephalogram (EEG) and motor activity with postural muscle tone on the EMG during waking in a tonic and phasic manner. Acting through α_1-adrenergic receptors (α_1-AR), NA can depolarize and excite many neurons through the central nervous system, including thalamo-cortical and basalo-cortical projection neurons, which respectively release glutamate (Glu) and ACh in the cortex, and excitatory (Glu) reticulo-spinal projection or motor neurons (Figure 1). These actions appear to be very important or at least potent for stimulating and maintaining an aroused and attentive state in that the condition of hypersomnia and narcolepsy is best treated by drugs which enhance NA release, including amphetamine-like compounds or

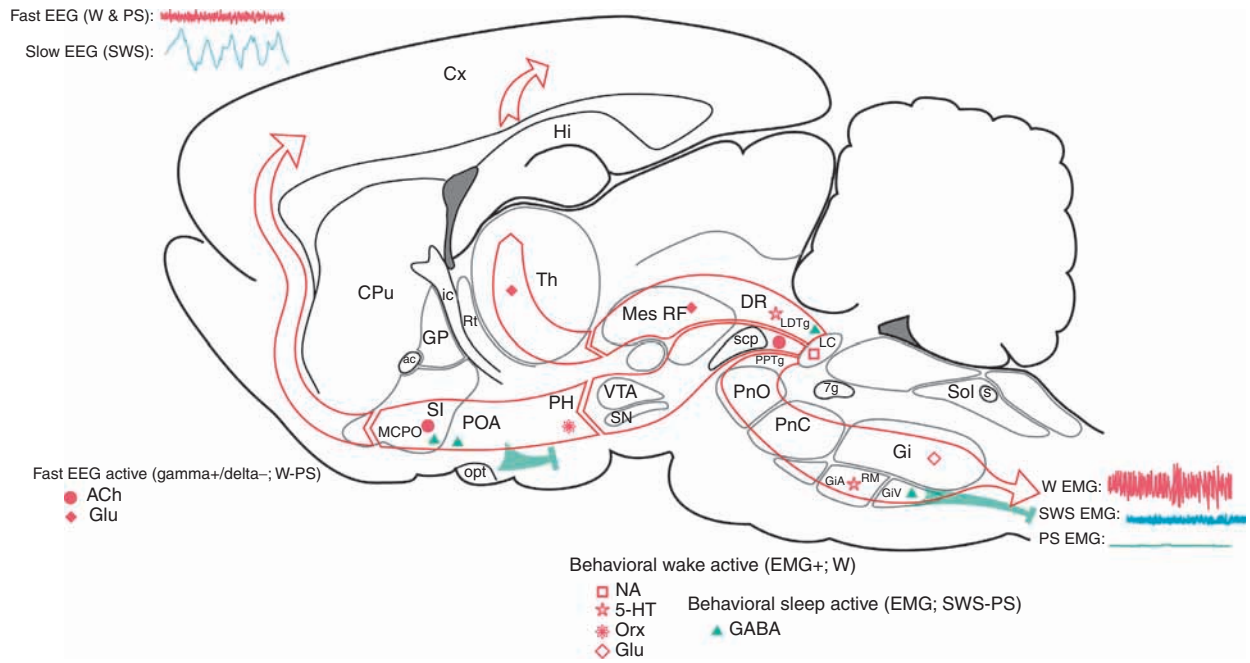

Figure 1 Noradrenergic (NA), locus coeruleus (LC), and serotonergic (5-HT) raphe nuclei along with other neurons involved in sleep–wake state control. Sagittal schematic view of the rat brain depicting neurons with their chemical neurotransmitters and pathways by which they influence cortical activity or behavior across the sleep–wake cycle. Wake (W) is characterized by fast gamma activity on the cortical EEG (upper left) and high postural muscle tone on the neck EMG (lower right); slow-wave sleep (SWS) is characterized by slow delta EEG and low tone on the EMG; and paradoxical sleep (PS) is characterized by fast gamma EEG and atonia on the EMG. Neurons that are active during waking (red symbols) include cells with ascending projections toward the cortex, which stimulate fast cortical activity, and cells with descending projections toward the spinal cord, which stimulate postural muscle tone and behavioral arousal. The NA LC neurons give rise to diffuse projections with highly collateralized and varicose axons that innervate multiple regions as way stations along their major pathways. One major projection pathway ascends dorsally through the reticular formation into the thalamus and beyond to the cortex. Another ascends ventrally through the ventral tegmentum into and through the hypothalamus via the medial forebrain bundle to reach the preoptic area and basal forebrain (MCPO and SI) and continue forward to the cortex, which is diffusely innervated. A third projection pathway passes ventrally into the pontine tegmentum to descend through the brain stem reticular formation and continues into the spinal cord, where multiple, including motor, neurons are innervated. 5-HT raphe nuclei also give rise to multiple and relatively diffuse projections, which, however, are segregated such that the dorsal raphe (DR) gives rise to prominent ascending projections and the medullary raphe (including the raphe magnus, RM) gives rise to prominent descending projections to the spinal cord. The NA and 5-HT neurons along with orexinergic (Orx) and some putative glutamatergic (Glu) neurons which commonly have relatively diffuse or descending projections discharge in association with behavioral arousal and EMG activity (EMG+) and cease firing with muscle atonia to be active during W and silent during PS (W, empty red symbols). The cholinergic (ACh) neurons and other Glu neurons which have predominantly ascending projections discharge in association with fast EEG activity (gamma+) and cease firing with delta activity (delta−) to be active during W and PS (W-PS, filled red symbols). Neurons that are active during sleep (aqua symbols) include cells with descending projections into the brain stem and spinal cord which diminish behavioral arousal and muscle tone. Important among such cells are GABAergic (and glycinergic) neurons in the ventral medulla which project to the brain stem and spinal cord, where they can inhibit motor neurons during sleep. The cholinergic neurons have the capacity to stimulate cortical activation during waking and PS and also to promote PS with muscle atonia by acting in part on such GABAergic neurons, when the monoaminergic and orexinergic neurons are silent and also inhibited by GABAergic neurons. 5-HT, serotonin; 7g, genu seventh nerve; ac, anterior commissure; ACh, acetylcholine; CPu, caudate putamen; Cx, cortex; DR, dorsal raphe nucleus; EEG, electroencephalogram; EMG, electromyogram; GABA, γ-aminobutyric acid; Glu, glutamate; Gi, gigantocellular RF; GiA, gigantocellular, alpha part RF; GiV, gigantocellular, ventral part RF; Glu, glutamate; GP, globus pallidus; Hi, hippocampus; ic, internal capsule; LC, locus coeruleus nucleus; LDTg, laterodorsal tegmental nucleus; Mes RF, mesencephalic RF; MCPO, magnocellular preoptic nucleus; NA, noradrenaline; opt, optic tract; Orx, orexin; PH, posterior hypothalamus; PnC, pontine, caudal part RF; PnO, pontine, oral part RF; POA, preoptic area; PPTg, pedunculopontine tegmental nucleus; PS, paradoxical sleep (also called REM sleep); RF, reticular formation; RM, raphe magnus; Rt, reticularis nucleus of the thalamus; s, solitary tract; scp, superior cerebellar peduncle; SI, substantia innominata; SN, substantia nigra; Sol, solitary tract nucleus; SWS, slow-wave sleep; Th, thalamus; VTA, ventral tegmental area; W, wake. Modified from Jones BE (2005) From waking to sleeping: Neuronal and chemical substrates. *Trends in Pharmacological Sciences* 26: 578–586.

catecholamine-uptake blockers. NA also excites and can thereby recruit Orx neurons, which in turn give rise, like the LC neurons, to highly diffuse projections and act by exciting multiple arousal systems together with cortical and spinal motor neurons. Orx, which also excites the LC neurons in return, is essential for the maintenance of waking because in its absence narcolepsy with cataplexy results.

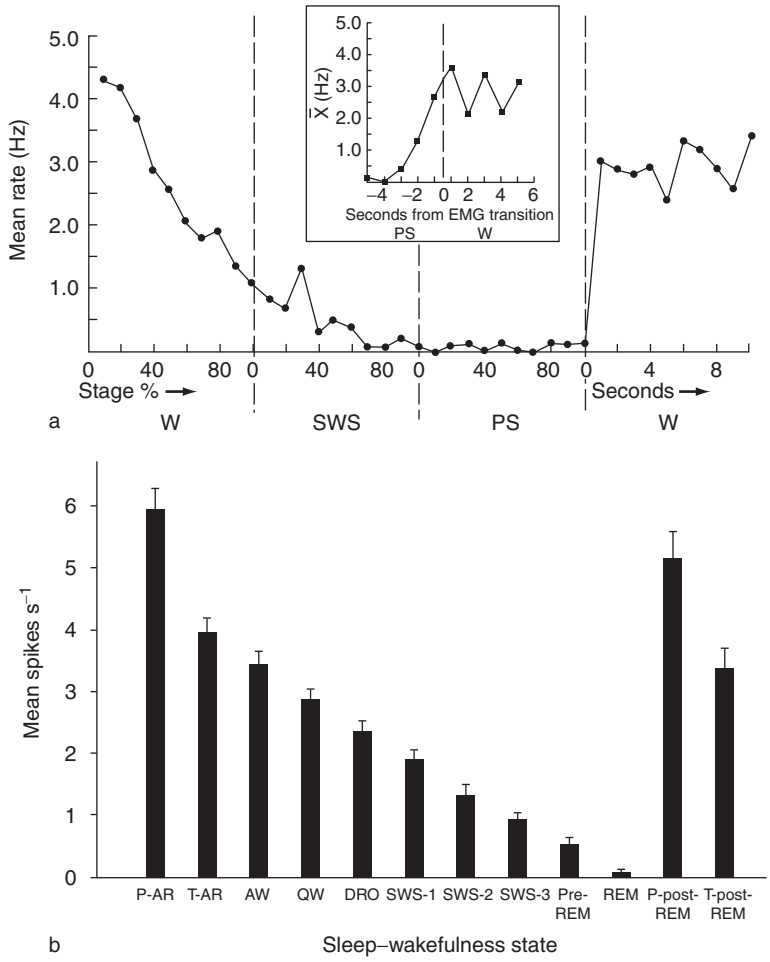

Figure 2 Discharge of (a) locus coeruleus (LC) and (b) dorsal raphe (DR) units across sleep–wake states: (a) mean firing rate (in hertz) per percentage of stage or seconds across a full sleep–wake cycle for multiple LC neurons (normalized according to completion of the entire cycle); (b) mean firing rate (in spikes s^{-1}) per state or stage across a full sleep–wake cycle for multiple DR neurons. As shown in (a), there is a progressive decrease in discharge with transition from waking into SWS and a cessation of firing during PS prior to an increase with the return of muscle tone (EMG) that actually anticipates the EMG change (inset). The graph in (b) also shows a progressive decrease with transition from aroused waking to late SWS and a virtual cessation of firing during PS prior to an increase with awakening. AW, active waking; DRO, drowsy; EMG, electromyogram; P-AR, phasic arousal; P-post-REM, phasic post-REM; Pre-REM, 60 s before REM onset; PS, paradoxical sleep; QW, quiet waking; REM, rapid eye movement sleep; SWS, slow wave sleep; SWS-1 early slow-wave sleep; SWS-2, middle slow-wave sleep; SWS-3, late slow-wave sleep; T-AR, tonic arousal; T-post-REM, tonic post-REM; W, waking. Vertical bars = SEM. (a) From Aston-Jones G and Bloom FE (1981) Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep-waking cycle. *Journal of Neuroscience* 1: 876–886. (b) From Trulson ME and Jacobs BL (1979) Raphe unit activity in freely moving cats: Correlation with level of behavioral arousal. *Brain Research* 163: 135–150.

NA also promotes waking by inhibiting sleep-promoting γ-aminobutyric acid (GABA)ergic neurons in the preoptic area (POA) and basal forebrain which bear α_2-AR (Figure 1). These GABAergic neurons discharge during SWS and PS in a reciprocal manner to the NA LC neurons and, accordingly, in negative correlation with muscle tone on the EMG. By descending projections, these sleep-promoting GABAergic neurons can inhibit Orx neurons in the posterior hypothalamus (PH) and also NA cells in the LC more caudally. The NA LC neurons thus promote arousal by synergistic activity with other arousal systems, including the Orx neurons, and prevent sleep by antagonistic activity with sleep-promoting systems, including GABAergic preoptic and basal forebrain neurons.

Specifically with regard to PS, the NA LC neurons were found long ago to have a reciprocal relationship in their discharge with respect to presumed cholinergic neurons of the brain stem, as formalized in a model by McCarley and Hobson. Whereas LC neurons cease firing during PS, presumed cholinergic pontomesencephalic neurons discharged at maximal rates during PS. From pharmacological and electrophysiological studies, it appeared that the monoaminergic, including, significantly, noradrenergic neurons, had to be off for the cholinergic neurons to stimulate the

constellation of events, including cortical activation with REMs and muscle atonia, of PS (Figure 1). Indeed, at least some cholinergic neurons are inhibited by NA through α_2-AR. On the other hand, ACh excites the NA LC neurons and thus must act to inhibit the NA cells through GABAergic interneurons in the region, which from c-Fos expression, as an indication of neural activity, are active together with the cholinergic neurons during PS (Figure 1). The NA LC neurons have thus been conceived as PS-permitting (by permitting PS to occur through their cessation of activity and resulting disinhibition of cholinergic cells along with disfacilitation of motor systems), whereas the cholinergic and local GABAergic neurons are PS-promoting (and act in part by inhibiting the NA LC neurons).

Projections and Role of Serotonergic Raphe Neurons in Waking and Sleeping

Following pharmacological and lesion studies which indicated that 5-HT and raphe neurons play a role in promoting sleep, single-unit recording studies were undertaken to determine when they discharged. These early studies by McGinty and Harper surprisingly revealed that presumed 5-HT neurons in the DR discharged during waking, decreased their firing during SWS, and ceased firing during PS, in a similar manner to the NA LC neurons (Figures 1 and 2). Subsequent studies examining the discharge of presumed 5-HT neurons in the other raphe nuclei through the brain stem similarly found the same profile of activity across the sleep–wake cycle, and studies examining release in the forebrain also indicated that 5-HT release was greatest during waking and minimal during sleep. So, it became clear through these results that 5-HT could not generate SWS, as was originally proposed.

However, many lines of pharmacological and physiological evidence indicated that 5-HT and raphe neurons did not act in the same way that the NA neurons did. First, as already stated, depletion of NA resulted in decreased waking, whereas depletion of 5-HT resulted in decreased sleep. Second, whereas NA LC neurons discharged during orientation to a novel sensory stimulus, presumed 5-HT DR neurons ceased firing. Moreover, whereas NA LC neurons discharged in response to physiological challenges which activated the sympathetic nervous system, 5-HT raphe neurons did not. So, perhaps 5-HT neurons promote a less attentive or responsive waking state from which sleep onset is facilitated.

The 5-HT DR neurons project along similar pathways as the NA LC neurons to reach the thalamus, hypothalamus, basal forebrain, and cortex while leaving collaterals along the way (Figure 1). Their ascending projections are complemented by those of the central superior nucleus. The major descending 5-HT projections originate from neurons in the medulla (raphe magnus, pallidus, and obscurus) and extend through the entire spinal cord. 5-HT acts on multiple, different receptors and thus has the capacity to excite or inhibit different cell populations through the central nervous system. Whereas, its actions on certain thalamo-cortical projection (i.e., Glu) neurons and basalo-cortical (i.e., ACh) neurons are often inhibitory through 5-HT1 receptors, they are excitatory on some cortical, including pyramidal, cells, through 5-HT2 or other receptors. In the brain stem and spinal cord, 5-HT depolarizes and excites, like NA, motor neurons and can thus enhance muscle tone and activity. It can inhibit sensory transmission.

With respect to PS, the presumed 5-HT neurons of the DR, like the NA LC neurons, cease firing in a reciprocal manner to the presumed cholinergic cells of the pontomesencephalic tegmentum. As for the NA LC neurons, the 5-HT DR neurons are inhibited by local GABAergic neurons during PS. 5-HT hyperpolarizes and inhibits cholinergic cells in both the brain stem and basal forebrain. It can thus prevent the occurrence of PS with muscle atonia promoted by brain stem cholinergic neurons and diminish cortical activation evoked by basal forebrain cholinergic neurons.

Given their discharge during waking behaviors involving rhythmic motor activities, such as locomotion or grooming, and their cessation of discharge during orientation to sensory stimuli, it has been proposed by Jacobs that 5-HT neurons are primarily involved in facilitating rhythmic motor patterns and attenuating responses to extraneous sensory stimuli during those activities. Their discharge thus facilitates muscle tone with motor activities, yet diminishes cortical activation in response to external sensory stimuli during those activities. In this way, they function in opposition to the cholinergic cells, which diminish muscle tone and motor activities and enhance cortical activation during attentive waking and PS. The discharge of the NA LC cells facilitates muscle tone and arousal, but also enhances cortical activation in response to external sensory stimuli during waking. NA, which excites and depolarizes the basal forebrain cholinergic cells and can act in synergy with certain cholinergic neurons during waking but operates in opposition to other pontomesencephalic cholinergic cells whose discharge occurs during PS. Collectively, both the NA and 5-HT systems prevent loss of muscle tone and the apparition of PS, and drugs which are used to treat narcolepsy include reuptake blockers, which act to enhance both 5-HT and NA transmission.

See also: Dopamine Control of Arousal; Sleep and Sleep States: Thalamic Regulation; Sleep and Sleep States: Hypothalamic Regulation; Sleep and Sleep States: Histamine Role; Sleep and Sleep States: Cytokines and Neuromodulation; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Aston-Jones G and Bloom FE (1981) Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep–waking cycle. *Journal of Neuroscience* 1: 876–886.

Jacobs BL and Fornal CA (1999) Activity of serotonergic neurons in behaving animals. *Neuropsychopharmacology* 21: 9S–15S.

Jones BE (2004) Paradoxical REM sleep promoting and permitting neuronal networks. *Archives of Italian Biology* 142: 379–396.

Jones BE (2005) Basic mechanisms of sleep–wake states. In: Kryger MH, Roth T, and Dement WC (eds.) *Principles and Practice of Sleep Medicine*, pp. 136–153. Philadelphia: Elsevier Saunders.

Jones BE (2005) From waking to sleeping: Neuronal and chemical substrates. *Trends in Pharmacological Sciences* 26: 578–586.

Jouvet M (1972) The role of monoamines and acetylcholine-containing neurons in the regulation of the sleep-waking cycle. *Ergebnisse der Physiologie* 64: 165–307.

McCormick DA (1992) Neurotransmitter actions in the thalamus and cerebral cortex and their role in neuromodulation of thalamocortical activity. *Progress in Neurobiology* 39: 337–388.

Nishino S and Mignot E (1997) Pharmacological aspects of human and canine narcolepsy. *Progress in Neurobiology* 52: 27–78.

Saper CB, Chou TC, and Scammell TE (2001) The sleep switch: Hypothalamic control of sleep and wakefulness. *Trends in Neuroscience* 24: 726–731.

Siegel JM (2004) The neurotransmitters of sleep. *Journal of Clinical Psychiatry* 65(supplement 16): 4–7.

Trulson ME and Jacobs BL (1979) Raphe unit activity in freely moving cats: Correlation with level of behavioral arousal. *Brain Research* 163: 135–150.

Zeitzer JM, Nishino S, and Mignot E (2006) The neurobiology of hypocretins (orexins), narcolepsy and related therapeutic interventions. *Trends in Pharmacological Sciences* 27: 368–374.

Slow Axonal Transport

A Brown, Ohio State University, Columbus, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

Axons can be 1 m or more in length in large animals, yet they are dependent on the cell body for the synthesis of most of their components. Materials destined for the axon are transported anterogradely, toward the axon tip, and materials destined to return are transported retrogradely, toward the cell body. This bidirectional transport process, known as axonal transport, is not fundamentally different from the pathways of macromolecular and membrane traffic found in other parts of the neuron, or indeed in any eukaryotic cell, but it is unique for the volume and scale of the traffic.

Kinetic studies on axonal transport *in vivo* using radioisotopic pulse labeling have defined two distinct patterns of movement, termed fast and slow. Fast axonal transport represents the movement of membranous organelles, including Golgi-derived vesicles, endocytic organelles, mitochondria, peroxisomes, lysosomes, and autophagosomes. For the most part, these organelles move rapidly and frequently at rates of up to several hundred millimeters per day, which corresponds to several micrometers per second. Some of these organelles move anterogradely, some retrogradely, and some (such as mitochondria) are capable of frequent reversals and move in both directions. Ribosomes and ribonucleoprotein particles, which are transported along developing and regenerating axons, probably also move by fast axonal transport.

In contrast to fast axonal transport, slow axonal transport represents the movement of proteins that make up the cytoskeleton and cytosol of axons. These proteins, which are not associated with membranous organelles, move in an intermittent manner at average rates of approximately millimeters or tenths of millimeters per day, which corresponds to several nanometers or tens of nanometers per second, much slower than fast axonal transport. The cargo structures of slow axonal transport are largely unknown, but they are believed to include cytoskeletal polymers and cytosolic macromolecular complexes. For many years, it was assumed that slow axonal transport is an exclusively anterograde movement, but it is now clear that at least some of the cargoes of slow axonal transport can move in both anterograde and retrograde directions, although the net direction of movement may still be anterograde.

The different rates of fast and slow axonal transport have long been assumed to reflect fundamentally distinct mechanisms of movement, but direct observations of membranous organelles and cytoskeletal polymers in living cells now suggest that all axonal cargoes are transported by fast motors. Thus, the principal difference between fast and slow transport is not thought to be the rate of movement *per se* but, rather, the frequency of movement and the nature of the cargo structures. The terms fast and slow refer to the average rate of movement on a timescale of days or weeks and do not reflect the actual velocity of the transport motors.

Methods for Studying Slow Axonal Transport

There are essentially three different kinds of experimental techniques for studying axonal transport: accumulation techniques, pulse-labeling techniques, and imaging techniques (Table 1). Although these approaches can all be applied to both fast and slow axonal transport, their use for studies on slow axonal transport is emphasized here. Each technique has strengths and limitations, and each reveals different aspects of the motility, so they should be regarded as complementary approaches, not as alternatives.

Accumulation Techniques

The oldest and simplest technique for studying axonal transport is to locally impede movement along an axon or nerve and then observe what cargo structures accumulate and the rate at which they do so. Retrogradely moving materials accumulate on the distal side of the blockade and become depleted on the proximal side, whereas anterogradely moving materials accumulate on the proximal side and become depleted on the distal side. The most commonly used approaches are to surgically ligate or crush a nerve *in vivo*. A variant on these approaches is the cold block, in which local cooling is applied to a nerve *in vivo*. This method is technically more challenging, but eliminates the tissue damage that is associated with the ligation and crush approaches. Another approach, which has been applied to isolated axons teased apart from peripheral nerves and to single axons of cultured neurons, is to locally compress axons with fine nylon or glass fibers (Figure 1).

Pulse-Labeling Techniques

The most widely used pulse-labeling technique is radioisotopic pulse labeling, which was first introduced

Table 1 Three experimental methods for studying slow axonal transport

	Accumulation techniques	*Pulse-labeling techniques*	*Imaging techniques*
Example	Ligate or constrict single axons or entire nerves and observe the accumulation of transported materials	Inject radioactive amino acids into the vicinity of nerve cell bodies *in vivo* and then sacrifice the animal and analyze the distribution of the radiolabeled proteins along the axons	Inject or transfect neurons with fluorescent proteins and observe the moving structures by fluorescence microscopy
Strengths	Simple; can be performed *in vivo* or in cell culture	Permits the kinetics and composition of axonal transport to be studied *in vivo*	High temporal resolution (seconds or milliseconds); high spatial resolution (micrometers); permits individual cargoes to be tracked
Limitations	Poor spatial resolution; poor temporal resolution	Poor spatial resolution (1 mm at best); poor temporal resolution (hours or days at best); each animal yields only one time point, so multiple animals are required for each analysis	Not well suited to long-timescale studies; cannot presently be performed *in vivo*

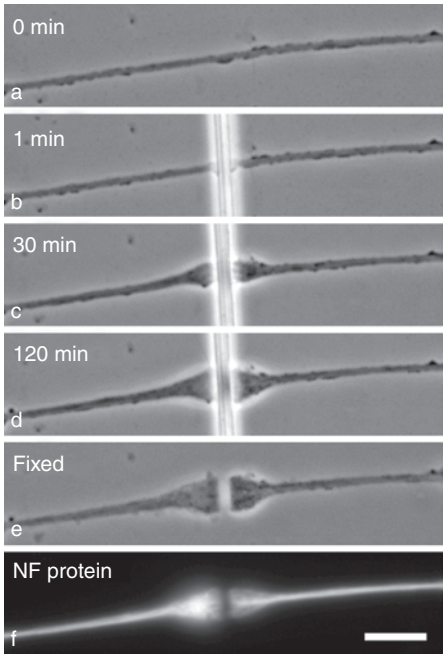

Figure 1 Constriction of an axon of a cultured neuron growing on a glass coverslip: (a) axon prior to constriction; (b) axon after 1 min constriction with a flexible glass fiber (~1 μm in diameter); (c) axon after 30 min; (d) axon after 120 min; (e) constricted axon after fixation and removal of the glass fiber; (f) distribution of neurofilament protein in the fixed axon, visualized by immunofluorescence microscopy. Note that the glass fiber compresses the axon without severing it. Anterogradely and retrogradely moving membranous organelles, which are cargoes of fast axonal transport, accumulate rapidly, causing the axon to swell on both sides of the constriction (proximal is left, distal is right). As shown in (f), the net anterograde movement of neurofilaments in these axons results in an accumulation of neurofilament protein proximal to the constriction, which is evident from the bright fluorescence in the proximal swelling. Scale bar = 5 μm. Reproduced from Koehnle TJ and Brown A (1999) Slow axonal transport of neurofilament proteins in cultured neurons. *Journal of Cell Biology* 144: 447–458.

in the late 1950s and 1960s. Most of what we know about the composition and kinetics of slow axonal transport has come from studies using this approach, and it remains to this day the preferred method for studying the composition and kinetics of slow axonal transport *in vivo*. Essentially, this technique involves the injection of radiolabeled precursors of macromolecules (amino acids, sugars, or nucleotides) into the vicinity of neuronal cell bodies in an animal. Most published studies have used radiolabeled amino acids, which permit the movement of proteins to be investigated. The radiolabeled amino acids are taken up by the cell bodies, creating a pulse of radiolabeled proteins. Those proteins destined for the axon move into and along the axon in association with distinct cargo structures. By injecting numerous animals and sacrificing them at various time intervals, the kinetics of transport can be analyzed (each animal yields a single time point). The most powerful approach is to dissect out the nerves containing the radiolabeled proteins and then cut them into contiguous segments, permitting biochemical analysis of the radiolabeled proteins by subcellular fractionation, immunoprecipitation, and/or electrophoresis (Figure 2).

Radioisotopic pulse-labeling studies of slow axonal transport can be performed on a variety of different nerve cell types. Those that are best suited are ones whose cell bodies are located in an anatomically discrete region, facilitating reproducible injection of the radioisotope, and ones whose axons extend in a nerve that can be dissected readily and that is long with minimal branching. Most radioisotopic pulse-labeling studies have been performed on rats or mice, and the most commonly used axons are the retinal ganglion cell axons of the optic nerve (radiolabeled precursor injected into the eye) and the motor

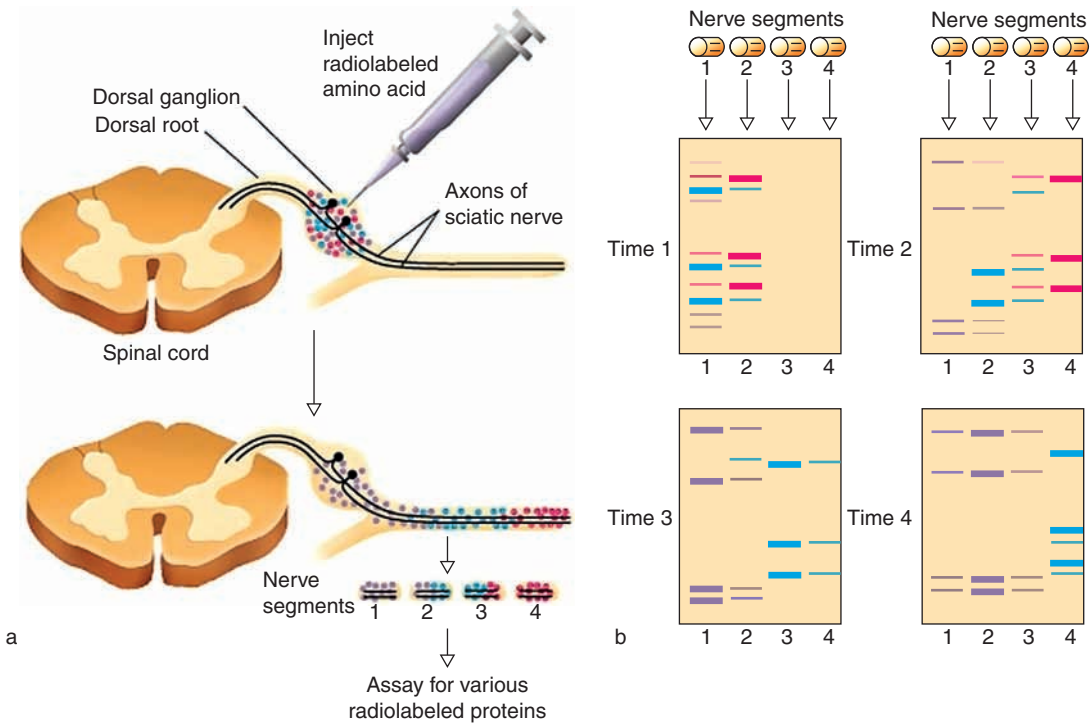

Figure 2 The radioisotopic pulse-labeling technique. (a) Radiolabeled amino acids are injected into the vicinity of nerve cell bodies (the dorsal root ganglion in this example), and the amino acids are incorporated into proteins in the nerve cell bodies. The proteins move out along the axons in association with distinct cargo structures (shown here as purple, blue, and red dots), which move at different rates. At early times the faster- and slower-moving cargo structures overlap considerably, whereas at later times they become spatially separated. At various time intervals, the animals are sacrificed and the nerves are cut into contiguous segments. (b) For each time interval (time 1–4), the proteins in each nerve segment are separated by gel electrophoresis and the amount of radioactivity associated with each protein is quantified. Labeled proteins associated with the purple, blue, and red cargo structures in (a) are shown as purple, blue, and red bands, respectively, in (b). Reproduced from Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, and Darnell J (2000) *Molecular Cell Biology*, 4th edn. New York: W. H. Freeman.

and sensory axons of the sciatic nerve (radiolabeled precursor injected into the ventral horn of the lumbar spinal cord or into the lumbar dorsal root ganglia, respectively).

Imaging Techniques

A third approach for studying axonal transport is direct observation of the movement of the cargo structures in living cells using light microscopy. Membranous organelles can be detected by video-enhanced differential interference contrast microscopy, but fluorescence microscopy is required to detect the cargoes of slow axonal transport because of their small size. Because fluorescence microscopy *in vivo* is problematic, experiments using this technique are presently limited to embryos or cultured nerve cells. The most common approach is to express fluorescent proteins in neurons or microinject them into cell bodies and then to observe the fluorescently labeled cargo structures by time-lapse imaging. The expression approach necessitates construction of a recombinant protein fused to green fluorescent protein or one of its variants, whereas the microinjection approach necessitates purification of the protein of interest and covalent modification with a fluorescent dye. The first attempts to detect slow axonal transport using these approaches date back to the late 1980s, but it was not until 2000 that the movement was demonstrated convincingly. To date, the movement of neurofilament proteins, tubulin, α-synuclein, synapsin I and glyceraldehyde-3-phosphate dehydrogenase have been observed in cultured nerve cells, but these are just a few of the many hundreds of proteins that are conveyed by slow axonal transport. The movements are fast, so imaging must be performed at intervals of several seconds or less. Because the cargo structures of slow axonal transport may be distributed continuously along the axon and cannot always be resolved from one another, it can be difficult to detect those that are moving. One way to

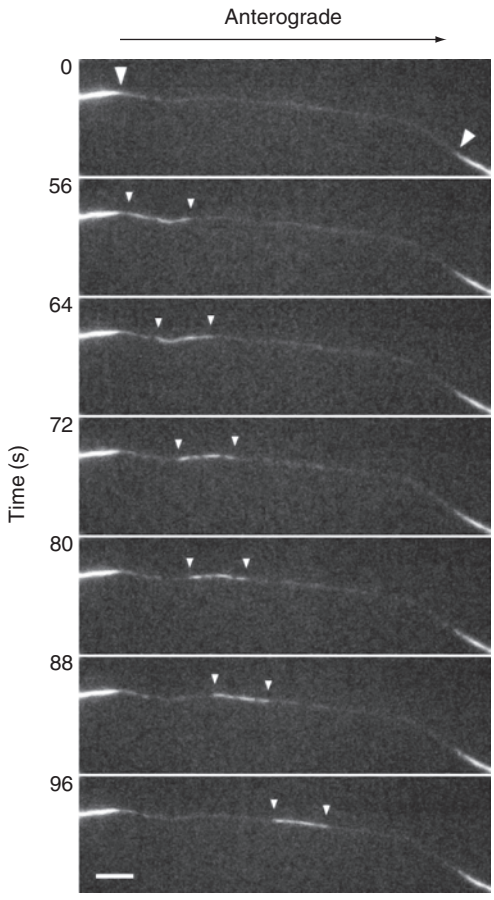

Figure 3 Movement of a single fluorescent neurofilament through a photobleached region in an axon of a cultured neuron. Neurofilaments were visualized by transfection with a neurofilament protein fused to green fluorescent protein, and the bleached region was observed by time-lapse fluorescence microscopy. The large arrowheads mark the edges of the photobleached region, and the small arrowheads mark the leading and trailing ends of the neurofilament, which is fluorescent because it originated from outside of the bleached region. Neurofilaments move at peak velocities of up to 3 μm s^{-1}, but the movements are infrequent and bidirectional. Scale bar = 5 μm. Reproduced from Wang L and Brown A (2001) Rapid intermittent of axonal neurofilaments observed by fluorescence photobleaching. *Molecular Biology of the Cell* 12: 3257–3267.

overcome this problem is to bleach the fluorescence in a segment of axon and then observe fluorescent cargo structures that originate from the flanking unbleached regions and move into the bleached region (Figure 3).

Two Distinct Subcomponents of Slow Axonal Transport

Radioisotopic pulse-labeling studies in a variety of neuronal cell types have shown that the proteins conveyed by slow axonal transport tend to resolve into two kinetically distinct subcomponents, termed slow

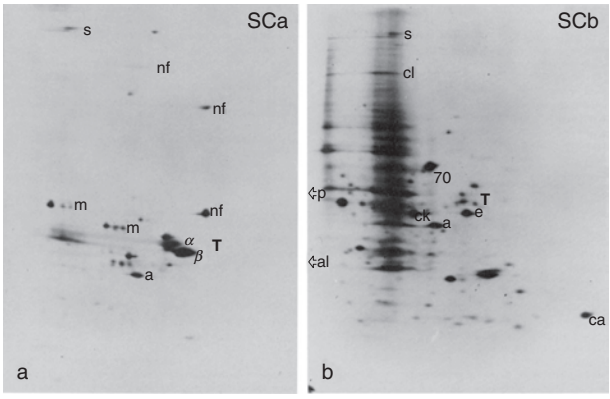

Figure 4 Autoradiographs of radiolabeled proteins in mouse optic nerve separated by two-dimensional gel electrophoresis: (a) slow component a (SCa); (b) slow component b (SCb). SCa (a) includes neurofilament triplet proteins (nf), tubulin (T; α and β subunits), tau (m), spectrin (s), and a small amount of actin (ac) trailing behind the SCb wave. SCb (b) includes actin (ac), calmodulin (ca), clathrin (cl), creatine kinase (ck), hsp70 (70), and nerve specific enolase (e). Some SCb proteins, such as aldolase (al) and pyruvate kinase (p), are not visible because they focused outside of the pH range of the isoelectric focusing gel. Reproduced from Lasek RJ, Garner JA, and Brady ST (1984), Axonal transport of the cytoplasmic matrix, *Journal of Cell Biology* 99: 212s–221s.

components a and b. Although some proteins are found in both subcomponents, most are unique to one or the other (Figure 4). Slow component a is slower and simpler in composition, being composed primarily of neuronal intermediate filament proteins, tubulin, and certain microtubule-associated proteins. In contrast, slow component b comprises more than 200 distinct proteins, indicating a remarkable complexity. Some of these proteins have been identified, and they include actin as well as the full spectrum of housekeeping proteins that make up neuronal cytosol, including proteins of glycolysis and intermediary metabolism, structural proteins, regulatory proteins, and motor proteins (Table 2).

The proteins that move in slow components a and b have generally been described as moving in a highly cohesive manner, but a careful reading of the literature suggests a more complex picture. For example, tubulin and neurofilament protein, which both move in slow component a, do not always move exactly in unison, and this is also the case for some of the proteins that move in slow component b. Moreover, although tubulin is transported predominantly in slow component a in some neurons, it is found in both slow components a and b in others. The explanation for this complexity is most likely that each subcomponent actually represents the movement of multiple distinct species of cargo structure that move independently. For example, tubulin and neurofilament proteins appear to be transported in the form of

Table 2 The composition and cargo structures of slow axonal transport

Rate class	*Average rate*	*Moving structures*	*Identified proteins*
Slow component a	0.2–1 mm day^{-1} ($\approx$ 0.002–0.02 μm s^{-1})	Neurofilaments and microtubules	Neurofilament proteins L, M, and H; internexin, peripherin, tubulin, neuronal spectrin (fodrin), tau, calcium/calmodulin-dependent protein kinase IIβ
Slow component b	2–8 mm day^{-1} ($\approx$ 0.02–0.09 μm s^{-1})	Unknown (may include microfilaments and supramolecular cytosolic protein complexes)	Actin, tubulin (in some neurons), dynein, dynactin, myosin Va, calmodulin, clathrin, clathrin uncoating protein hsc70, heat shock protein hsp70, cytosolic chaperonin containing T-complex polypeptide 1 (CCT), molecular chaperone hsc73, calcium/calmodulin-dependent protein kinase IIα, aldolase, creatine phosphokinase, enolase, glyceraldehyde-3-phosphate dehydrogenase, phosphofructokinase, superoxide dismutase 1, cyclophilin A, cofilin, actin depolymerizing factor, profilin, synapsin I, annexin VI, ubiquitin, ubiquitin C-terminal hydrolase PGP 9.5

microtubule and neurofilament polymers, which are distinct structures. It is possible that neurofilaments and microtubules could move at similar average rates in some axons and at different average rates in others. Moreover, the transport kinetics of any given protein is determined not just by the average rate of movement of the cargo structure with which it associates but also by the proportion of the time that the protein spends bound to that structure. Thus, proteins that associate with the same cargo structure may move with different kinetics if they differ in their affinity for that structure. Additional complexity can also arise if a given protein is capable of diffusion when it is not bound to its cargo structure.

The Kinetics of Slow Axonal Transport

When considering the rate of slow axonal transport, it is important to make a distinction between measurements made on the fast and slow timescales. In the radioisotopic pulse-labeling experiments, which are performed on a timescale of days or weeks, each protein conveyed by slow axonal transport moves out along the axon in the form of a unimodal bell-shaped wave that spreads as it moves distally. The rate of slow axonal transport is generally quoted as the rate of movement of the peak of the wave, which represents the modal transport rate, although the width and spreading of the waves clearly indicate that the proteins actually move at a range of rates. The modal rate differs in different neuronal cell types and may also vary spatially (along axons) as well as temporally (during development, maturation, and aging), but it is generally around 0.2–1 mm day^{-1} (0.002–0.02 μm s^{-1}) for slow component a and 2–8 mm day^{-1} (0.02–0.09 μm s^{-1}) for slow component b.

In contrast to the radioisotopic pulse-labeling experiments, direct observations on the movement of neurofilaments and microtubules in cultured neurons using fluorescence microscopy on a timescale of seconds or minutes indicate that these structures actually move very rapidly, at rates of up to several micrometers per second, and that these movements are also intermittent and infrequent (Figure 5). Thus, the bell-shaped waves that characterize slow axonal transport in the radioisotopic pulse-labeling experiments can be considered to represent the distribution of hundreds or thousands of pulse-labeled cargoes whose individual movements and pauses are summed over the days, weeks, or months that they spend traveling down the axon. Whether the movement is considered to be fast or slow depends on the timescale of the observations.

A Brief History of Experimental Studies on Slow Axonal Transport

The existence of material transport along axons was inferred by biologists more than 100 years ago, but it was not demonstrated experimentally until the classic study of Weiss and Hiscoe, published in 1948. These authors severed peripheral nerves in laboratory animals and then allowed the proximal axonal stumps to regenerate through segments of transplanted arteries. As the regenerating axons matured, they increased in caliber and experienced a gradually increasing constriction by the arterial sheath through which they passed. Over a period of days or weeks, the axons swelled proximal to the constrictions due to the accumulation of anterogradely transported materials. When the constrictions were released, the axonal swellings propagated distally at a rate of approximately 1 mm day^{-1}. The enlargement of the axons

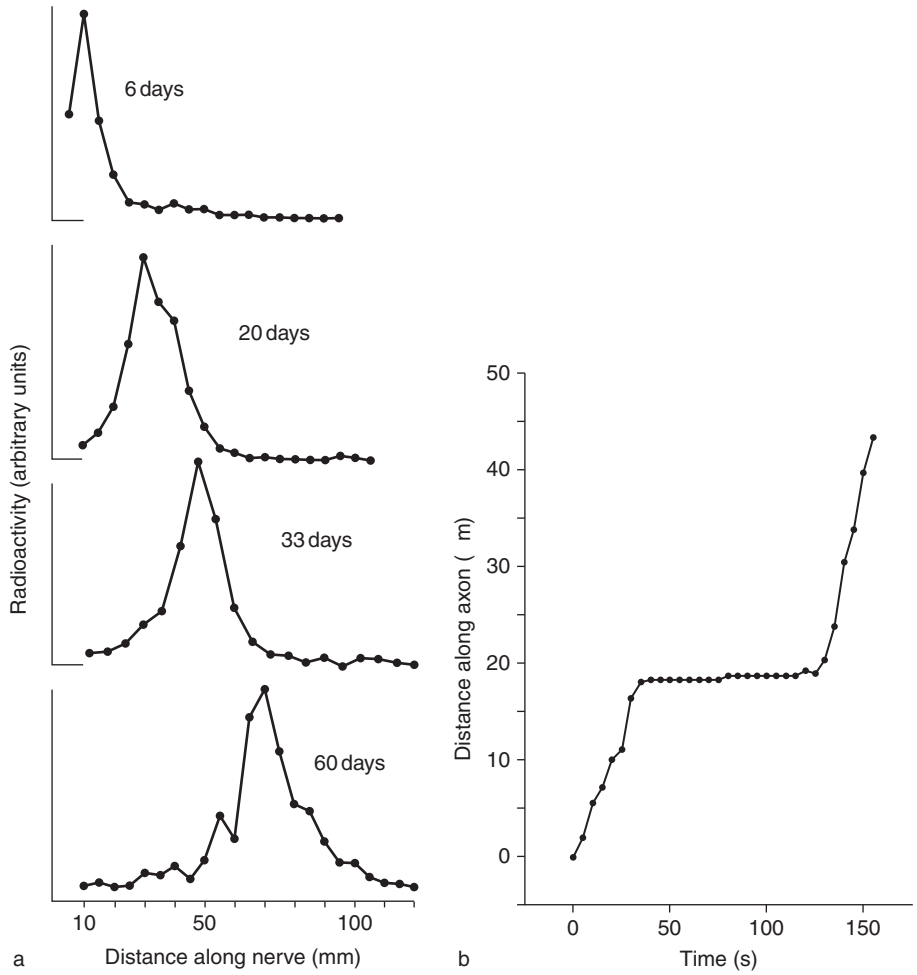

Figure 5 Kinetics of neurofilament transport on fast and slow timescales: (a) timescale of weeks to months as revealed by radioisotopic pulse labeling in rat spinal motor neurons *in vivo*; (b) timescale of seconds to minutes, as revealed by live cell fluorescence imaging in cultured nerve cells. In (a), the pulse of radiolabeled neurofilament proteins moves out along the sciatic nerve in the form of a bell-shaped wave that spreads as it moves distally. The graph in (b) represents a single neurofilament moving in an anterograde direction. The filament was tracked by time-lapse imaging (as described in **Figure 3**). Each point represents the position of the leading end of the neurofilament along the axon in one 5 s interval. Both anterogradely and retrogradely moving neurofilaments exhibit a similar behavior, characterized by bouts of rapid movement interrupted by pauses of varying durations. (a) Data from Hoffman PN, Griffin JW, Gold BG, and Price DL (1985) Slowing of neurofilament transport and the radial growth of developing nerve fibers. *Journal of Neuroscience* 5: 2920–2929, courtesy of Paul Hoffman. (b) Data from Wang L and Brown A (2001) Rapid intermittent of axonal neurofilaments observed by fluorescence photobleaching. *Molecular Biology of the Cell* 12: 3257–3267.

proximal to the constrictions was referred to as damming, and it was considered to be due to an 'axomotile' mechanism. The authors termed this movement axonal flow, but we know it today as axonal transport. The axoplasm was considered to move in a slow anterograde direction as a semirigid column propelled by peristaltic contractions of the axolemma. The fact that this model turned out to be incorrect should not detract from the importance of this paper, which was published long before the era of modern cell biology. In fact, the authors' appreciation of the significance of axonal transport for the growth and maintenance of axons was remarkably prescient, and this study remains a landmark in the field of axonal transport.

With the introduction of radioisotopic tracer techniques in the late 1950s, it became apparent that axonal transport was bidirectional, with distinct anterograde and retrograde components, and that many of the components moved at very fast rates. Based on these findings, the concept of axonal transport as a bulk anterograde cytoplasmic flow evolved into one of bidirectional movement of discrete cytological structures at a broad range of rates. In the 1970s and early 1980s,

two competing theories emerged. In the structural hypothesis of Lasek and colleagues, axonal transport was considered to be comprised of multiple discrete rate components, each corresponding to the cohesive movement of a distinct subset of cytological structures. Lasek and colleagues postulated that all axonal proteins move by association with, or as integral parts of, cytologically identifiable subcellular carrier structures. The strength of this perspective was the recognition that the fast and slow components of transport have a distinct polypeptide composition and that this distinct composition reflects the movement of distinct organelles and macromolecular structures. In fact, one of the original predictions of the structural hypothesis was that cytoskeletal polymers are among the cargoes of slow axonal transport, which is now known to be true. However, the emphasis of this hypothesis on the cohesive nature of the movement within each subcomponent of axonal transport gave rise to the widely held perception that cytoskeletal and cytosolic proteins move along the axon *en masse* (i.e., in a slow and synchronous manner), which is no longer thought to be the case. In a competing theory termed the unitary hypothesis, Sydney Ochs and colleagues questioned the existence of slow axonal transport and proposed that there is a single fast transport mechanism that explains all movement in axons. Ochs and colleagues proposed that proteins conveyed by fast axonal transport could become deposited along the axon by dropping off the transport mechanism, with turnover of these deposited proteins giving rise to the appearance of slowly moving waves. The strength of this hypothesis was the recognition that both fast and slow transport could be generated by fast movements. However, this hypothesis failed to recognize that slow axonal transport represents a genuine movement of cargoes that are structurally distinct from those of fast axonal transport.

During the 1980s, the notion that slow axonal transport is a slow and synchronous movement became widely accepted and had a profound influence on the design of experiments to observe this movement in living cells. For example, in the late 1980s and 1990s, there were numerous attempts to observe slow axonal transport directly in cultured neurons using fluorescence microscopy. All these experiments focused on the cytoskeletal subunit proteins, and most used fluorescence photobleaching or photoactivation to mark these proteins in a short segment of axon and time-lapse imaging to observe their movement. Because the movement was thought to be slow and synchronous, these studies were designed to detect a slow translocation of the marked zone toward the axon tip. When they failed to detect such a movement, this failure was widely interpreted as evidence that cytoskeletal polymers in axons do not move and that cytoskeletal subunit proteins must be transported in some other unassembled form. This perspective became known as the subunit transport hypothesis, and it attracted considerable controversy. However, the failure of the photobleaching and photoactivation studies to detect any movement, either of subunits or polymers, was perplexing.

The controversy surrounding the subunit transport hypothesis was resolved in 2000 with the direct observation of the movement of neurofilaments, and shortly thereafter microtubules, in cultured nerve cells. The solution to the controversy was disarmingly simple – cytoskeletal polymers do move in axons, but their movements are not slow after all. The key to detecting these movements was the development of direct imaging strategies capable of detecting the rapid movement of single cytoskeletal polymers (e.g., Figure 3). Experiments using this approach have shown that neurofilaments and microtubules move at fast rates, approaching the rate of movement of membranous organelles, but that the average rate of movement is slow because the movements are infrequent, bidirectional, and highly asynchronous. Thus, the slow rate of slow axonal transport is actually the result of rapid movements interrupted by prolonged pauses. For example, it has been estimated using computational modeling that neurofilaments in motor axons of the mouse ventral root and sciatic nerve spend approximately 97% of their time pausing during their journey along the axon, and similar extents of pausing have been observed experimentally in cultured neurons. It seems likely that the many other proteins conveyed by slow axonal transport exhibit a similar motile behavior, but this has not been shown.

The Mechanism of Slow Axonal Transport

Current research on slow axonal transport is focused on addressing four central questions: (1) What are the cargo structures of slow components a and b, (2) what are the tracks along which they move, (3) what are the motors that move them, and (4) how is their movement regulated? The observation of the rapid intermittent movement of neurofilament and microtubule polymers in axons indicates that these cytoskeletal polymers are among the cargo structures of slow axonal transport, although this does not preclude the possibility that cytoskeletal subunit proteins may also move in an unassembled form. The form in which the many other proteins in slow axonal transport move remains unclear. One hypothesis is that they are transported by binding to (and effectively riding piggyback on) moving cytoskeletal polymers. The relatively simple protein composition of slow

component a suggests that neurofilaments and microtubules could be the sole carrier structures for this rate component; all the proteins that move in slow component a are either integral parts of these cytoskeletal polymers or are known to associate with these polymers *in vivo*. The presence of actin in slow component b suggests that microfilaments could function as a carrier structure for this rate component, although it seems unlikely that the several hundred different proteins that move in slow component b all bind to microfilaments. More probably, many of these proteins form supramolecular complexes that, in turn, associate with the moving filaments or that associate with motor proteins independently of the moving filaments. However, this hypothesis is speculative because the movement of microfilaments and most other cytosolic proteins in axons has not yet been demonstrated.

The rapid rate of movement of neurofilaments and microtubules in axons indicates that they are transported by fast motors, perhaps similar to or even the same as the motors that move membranous organelles, but the identity of these motors and the tracks along which they move are not known. By analogy with the movement of membranous organelles, it seems likely that neurofilaments and microtubules and other cargoes of slow axonal transport move along either microtubule or microfilament tracks. For example, there is evidence that dynein may transport axonal microtubules anterogradely along microfilaments and that dynein and kinesin may transport axonal neurofilaments bidirectionally along microtubules. The frequency of neurofilament movement may be regulated by phosphorylation of the neurofilament proteins. Axonal neurofilaments can also interact with myosin Va, but the role of this interaction in neurofilament movement is presently unclear. Although much remains to be learned about the mechanism of movement, it seems likely that the cargo structures of slow axonal transport each interacts with a number of different motor proteins and that these motors act cooperatively to translocate and organize these polymers in both the longitudinal and radial dimensions of the axon.

A Unified Perspective for Fast and Slow Axonal Transport

The rapid intermittent movement of cytoskeletal and cytosolic proteins in axons when observed on a time scale of seconds or minutes suggests a unified perspective for fast and slow axonal transport which combines elements of the structural hypothesis of Lasek and colleagues and the unitary hypothesis of Ochs and colleagues. According to this perspective, multiple distinct classes of membranous and nonmembranous cargoes are all transported along axons by the same underlying mechanism, but they move at different rates due to differences in their duty ratio (the proportion of the time that they spend moving). Each molecule moves by virtue of its association, directly or indirectly, with a particular class of cargo structure. The transport rate of individual proteins is determined by the proportion of the time that they spend in association with their cargo structure as well as the velocity and duty ratio of that structure. The difference in the rate of fast and slow axonal transport is therefore due to differential regulation of the direction, duration, and frequency of movement of the cargo structures. Membranous organelles on the secretory and endocytic pathways move rapidly and continuously in a unidirectional manner, either anterograde or retrograde, pausing for only brief periods. Many of these cargoes function to deliver membrane and protein components to a particular site along the axon or at its tip, and the high duty ratio of these organelles ensures that these components are delivered rapidly. In contrast, the cargo structures of slow axonal transport move very infrequently, with a low duty ratio, pausing most of the time and sometimes reversing during their journey along the axon. Although we refer to these structures as cargoes, they are not simply the luggage of intracellular transport; these cytoskeletal polymers and cytosolic protein complexes are probably preassembled functional units that fulfill their architectural, physiological, and/or metabolic roles in the axon during their transit. For these cargoes, the journey is perhaps more important than the ultimate destination, and this may explain why they move so slowly.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal mRNA Transport and Functions; Axonal Transport and ALS; Axonal Transport and Alzheimer's Disease; Axonal Transport and Huntington's Disease; Axonal Transport and Neurodegenerative Diseases; Axonal Transport Disorders; Cytoskeletal Interactions in the Neuron; Microtubule Associated Proteins in Neurons; Microtubules: Organization and Function in Neurons; Mitochondrial Organization and Transport in Neurons; Myosin Transport and Neuronal Function; Neurofilaments: Organization and Function in Neurons; Peroxisomes: Organization and Transport in Neurons.

Furthur Reading

Ackerley S, Thornhill P, Grierson AJ, et al. (2003) Neurofilament heavy chain side arm phosphorylation regulates axonal transport of neurofilaments. *Journal of Cell Biology* 161(3): 489–495.

Baas PW and Buster DW (2004) Slow axonal transport and the genesis of neuronal morphology. *Journal of Neurobiology* 58(1): 3–17.

Brown A (2000) Slow axonal transport: Stop and go traffic in the axon. *Nature Reviews Molecular Cell Biology* 1: 153–156.

Brown A (2003) Axonal transport of membranous and nonmembranous cargoes: A unified perspective. *Journal of Cell Biology* 160(6): 817–821.

Grafstein B and Forman DS (1980) Intracellular transport in neurons. *Physiological Reviews* 60: 1167–1283.

Hirokawa N and Takemura R (2005) Molecular motors and mechanisms of directional transport in neurons. *Nature Reviews Neuroscience* 6(3): 201–214.

Hoffman PN, Griffin JW, Gold BG, and Price DL (1985) Slowing of neurofilament transport and the radial growth of developing nerve fibers. *Journal of Neuroscience* 5: 2920–2929.

Hoffman PN and Lasek RJ (1975) The slow component of axonal transport: Identification of major structural polypeptides of the axon and their generality among mammalian neurons. *Journal of Cell Biology* 66: 351–366.

Koehnle TJ and Brown A (1999) Slow axonal transport of neurofilament proteins in cultured neurons. *Journal of Cell Biology* 144: 447–458.

Lasek RJ, Garner JA, and Brady ST (1984) Axonal transport of the cytoplasmic matrix. *Journal of Cell Biology* 99(1, pt. 2): 212s–221s.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, and Darnell J (2000) *Molecular Cell Biology*, 4th edn. New York: W. H. Freeman.

Ochs S (1979) The early history of material transport in nerve. *Physiologist* 22(5): 16–19.

Rao MV and Nixon RA (2003) Defective neurofilament transport in mouse models of amyotrophic lateral sclerosis: A review. *Neurochemical Research* 28(7): 1041–1047.

Roy S, Winton M, Black MM, Trojanowski JQ, and Lee VMY (2007) Rapid intermittent cotransport of slow component b proteins. *Journal of Neuroscience* 27: 3131–3138.

Tytell M, Black MM, Garner JA, and Lasek RJ (1981) Axonal transport: Each major component reflects the movement of distinct macromolecular complexes. *Science* 214: 179–181.

Wang L and Brown A (2001) Rapid intermittent of axonal neurofilaments observed by fluorescence photobleaching. *Molecular Biology of the Cell* 12: 3257–3267.

Wang L, Ho C-L, Sun D, Liem RKH, and Brown A (2000) Rapid movement of axonal neurofilaments interrupted by prolonged pauses. *Nature Cell Biology* 2(3): 137–141.

Weiss PA and Hiscoe HB (1948) Experiments on the mechanism of nerve growth. *Journal of Experimental Zoology* 107: 315–396.

SNAREs

J Rizo, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Chemical synaptic transmission is mediated by neurotransmitters that are released by Ca^{2+}-triggered synaptic vesicle exocytosis. The exquisite spatial and temporal regulation of this process depends on a highly complex protein machinery that is formed in part by components with homologs in most types of intracellular membrane traffic, including *N*-ethylmaleimide sensitive factor (NSF), soluble NSF attachment proteins (SNAPs), SNAP receptors (SNAREs), Sec1/Munc18 (SM) proteins, and Rab guanosine triphosphatases (GTPases). Among these components, the SNAREs are believed to be at the heart of a conserved mechanism of membrane fusion by virtue of their ability to form tight SNARE complexes that bring two opposing membranes together. The neuronal SNAREs involved in neurotransmitter release are the synaptic vesicle protein synaptobrevin 2 (also known as vesicle associated protein 2 (VAMP2)) and the plasma membrane proteins syntaxin 1 (which in mammals includes two closely related isoforms, syntaxin 1A and 1B) and SNAP-25 (for synaptosomal associated protein of 25 kDa; no relation to SNAPs).

Neurotransmitter release depends also on proteins such as synaptotagmin 1 and complexins that play specialized roles in the tight regulation of this process and are not generally involved in other types of intracellular membrane traffic. Several of these proteins bind to the neuronal SNAREs and probably function in the regulation of SNARE complex assembly and/or in conjunction with the SNARE complex. Hence, synaptobrevin 2, syntaxin 1, and SNAP-25 have unique properties that have been adapted to the regulatory requirements of synaptic exocytosis, in addition to conserved features that are shared with SNAREs from other membrane compartments. This article focuses on the structural, biochemical, and functional properties of the neuronal SNAREs, but it also describes key findings on other SNAREs that have helped us to understand which of these properties are general and which are specialized.

SNARE Structure

SNAREs constitute a family of proteins characterized by sequences called SNARE motifs that comprise ~60–70 residues and have a high propensity to form coiled coils. Most SNAREs contain only one SNARE motif that is adjacent to a single C-terminal transmembrane (TM) region (e.g., synaptobrevin 2 and syntaxin 1). Some SNAREs contain two SNARE motifs connected by a long linker and do not have a TM sequence (e.g., SNAP-25) (Figure 1), but are attached to membranes through a posttranslational modification such as palmitoylation.

Circular dichroism (CD) and nuclear magnetic resonance (NMR) studies showed that isolated SNARE motifs are generally unstructured and adopt α-helical conformations when they bind to other SNARE motifs to form SNARE complexes. A SNARE complex contains four SNARE motifs that assemble into a parallel four-helix bundle, as first demonstrated for the neuronal SNAREs by electron paramagnetic resonance (EPR) spectroscopy and X-ray crystallography (Figure 2(a)). The neuronal SNARE complex is very stable (it is resistant to sodium dodecyl sulfate (SDS) and requires high temperatures for denaturation). This stability is conferred in large part by contacts between many hydrophobic residues that are arranged in layers, with each SNARE motif contributing one hydrophobic side chain to each layer. However, a polar, or zero layer, in the middle of the bundle is formed by one arginine (from synaptobrevin 2) and three glutamine side chains (one from syntaxin 1 and two from SNAP-25) (Figure 2(b)). These features are highly conserved in four different subfamilies of SNAREs and have led to their classification into Qa, Qb, Qc, and R SNAREs, depending on their homology with the SNARE motifs of syntaxin 1 (Qa), SNAP-25 (Qb and Qc for its N- and C-terminal SNARE motifs, respectively), or synaptobrevin (R). This classification is less ambiguous for SNAREs involved in homotypic membrane fusion than the original classification into v- and t-SNAREs (for vesicular and target membrane SNAREs, respectively). The assembly of the SNARE four-helix bundle is most likely at the center of the mechanism of membrane fusion. The functional importance of the polar layer is currently unclear, but it has been suggested to play a role in aiding SNARE complex disassembly or in dictating the proper register for SNARE complex formation.

In addition to a SNARE motif and a TM sequence, many SNAREs have an N-terminal region that spans more than half of its sequence. NMR studies have shown that the N-terminal regions of all SNAREs from the syntaxin subfamily, including syntaxin 1, contain an autonomously folded domain that adopts an antiparallel three-helix bundle structure, known as the H_{abc} domain (Figure 2(a)). This domain is connected to the SNARE motif through a linker region and is preceded by a short N-terminal

sequence (NTS) (Figure 1). The N-terminal regions of SNAREs from other subfamilies may also contain H_{abc} domains or autonomously folded domains with completely unrelated structures, such as longin domains or PX domains.

In the assembled SNARE complex, the H_{abc} domain of syntaxin 1 is flexibly linked to the four-helix bundle formed by the SNARE motifs. However, the H_{abc} domain binds intramolecularly to the SNARE motif in isolated syntaxin 1, forming a closed conformation that is incompatible with the SNARE complex. These findings suggested that the H_{abc} domain regulates SNARE complex formation and that syntaxin 1 must undergo a large conformational transition from a closed to an open conformation during exocytosis (Figure 3). However, only a few SNAREs from the syntaxin subfamily adopt a closed conformation, which thus appears to represent a specialization arising from unique regulatory requirements of neurotransmitter release and a few other membrane traffic processes. Conversely, some SNAREs outside the syntaxin family adopt closed conformations involving different types of autonomously folded N-terminal domains, indicating that intramolecular regulation of SNARE complex assembly is not restricted to syntaxins.

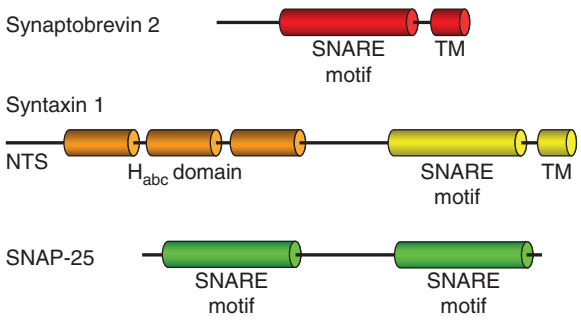

Figure 1 Domain structures of the neuronal SNAREs. The SNARE motifs, TM regions, and the helices of the H_{abc} domain are represented by cylinders; other sequences are represented by black lines. NTS, N-terminal sequence; SNAP-25, synaptosomal associated protein of 25 kDa; SNARE, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor; TM, transmembrane.

The closed conformation of syntaxin 1 binds tightly to the neuronal SM protein Munc18–1 and its structure at atomic resolution (Figure 2(c)) was revealed by X-ray analysis of a syntaxin 1–Munc18–1 complex. This mode of interaction is not generally conserved

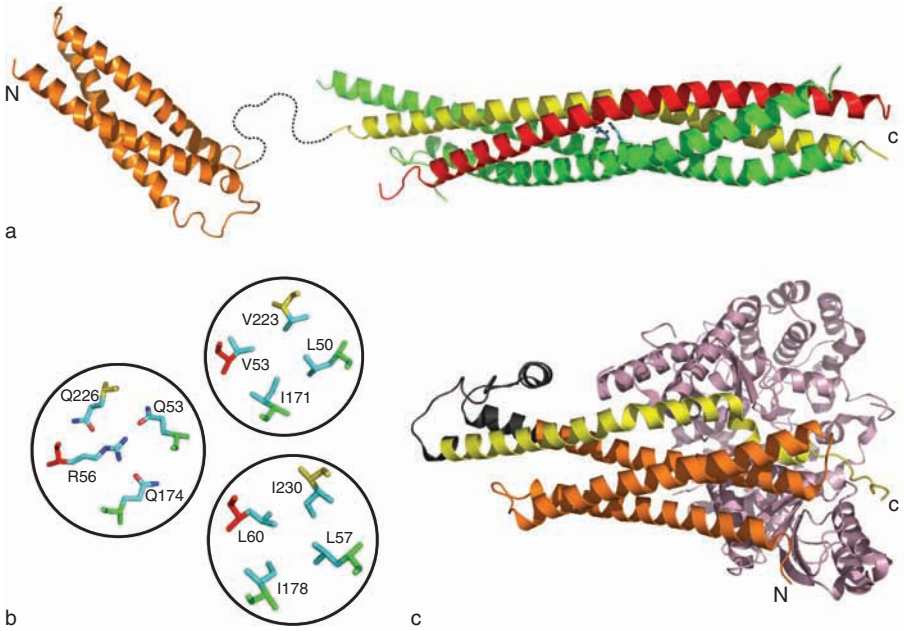

Figure 2 Three-dimensional structures of the neuronal SNAREs: (a) illustration of the three-dimensional structures of the syntaxin 1 H_{abc} domain determined by NMR spectroscopy (Fernandez et al.) (left) and the neuronal SNARE complex formed by the SNARE motifs of synaptobrevin 2, syntaxin 1, and SNAP-25 determined by X-ray crystallography (Sutton et al.) (right); (b) stick models of selected layers forming the neuronal SNARE complex; (c) illustration of the structure of the closed conformation of syntaxin 1 bound to Munc18–1 determined by X-ray crystallography. In (a), the dashed curved represents the flexible linker between the H_{abc} domain and SNARE motif of syntaxin 1. The color coding is the same as in Figure 1. In (b). the stick models include the polar layer formed by R56 of synaptobrevin 2, Q226 of syntaxin 1, and Q53 and Q174 of SNAP-25, as well as the preceding and ensuing layers, which are formed exclusively by hydrophobic residues. The backbone atoms are in the same color code as in Figure 1. Side chain atoms are in cyan for carbon, red for oxygen, and blue for nitrogen. In (c), the H_{abc} domain of syntaxin 1, its SNARE motif, and the linker region connecting them are in orange, yellow, and black, respectively. Munc18–1 is in violet (Misura et al.). C, C-terminus; N, N-terminus; NMR, nuclear magnetic resonance; SNAP-25, synaptosomal associated protein of 25 kDa; SNARE, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor.

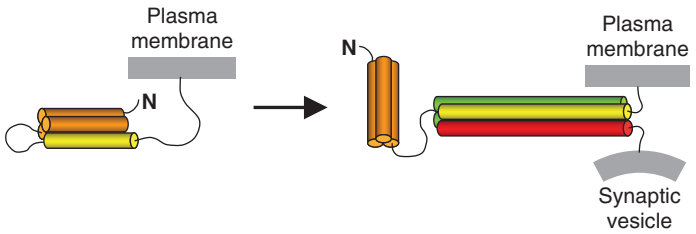

Figure 3 Diagram illustrating that syntaxin 1 must undergo a large conformational change during synaptic vesicle exocytosis. In the SNARE complex (right), the H_{abc} domain of syntaxin 1 is flexibly linked to the four-helix bundle formed by the SNARE motifs of syntaxin 1, SNAP-25, and synaptobrevin 2. In isolated syntaxin 1 (left), the H_{abc} domain binds intramolecularly to part of the SNARE motif, forming a closed conformation that is incompatible with SNARE complex formation. The color coding is the same as in **Figure 1**. N, N-terminus; SNAP-25, synaptosomal associated protein of 25 kDa; SNARE, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor.

in other syntaxins and SM proteins, but many SM protein/syntaxin interactions depend on the syntaxin N-terminal region, particularly the NTS. Hence, it is likely that the general function of the syntaxin N-terminal region is related to coupling with SM protein function.

Evolution of Ideas on SNARE Function

SNARE proteins have been investigated by a wide variety of techniques, providing a vivid example of the power of interdisciplinary research and a fascinating history with twists and turns that may continue for years to come. The first SNAREs were found in genetic screens for membrane traffic defects in yeast and in molecular cloning analyses of components from synaptic vesicle and synaptosomal membranes that identified syntaxin 1, synaptobrevin 2, and SNAP-25. The first evidence of the functional importance of these proteins for synaptic exocytosis was provided by the discovery in the early 1990s that they are the targets of clostridial neurotoxins, agents that potently inhibit neurotransmitter release. Around the same time, sequence analyses revealed a homology between these neuronal proteins and some of the proteins identified in the studies of membrane traffic defects in yeast. Moreover, NSF/Sec18p and SNAPs/Sec17p were identified as crucial factors for vesicle-mediated transport in yeast and mammalian cells. Homologies were also observed between yeast and neuronal Rab proteins, as well as between SM proteins involved in traffic at different membrane compartments. All these findings led to the now widely held notion that a conserved machinery mediates most types of intracellular membrane traffic.

Crucial discoveries in the understanding of SNARE function were also the observations in 1993 that synaptobrevin 2, syntaxin 1, and SNAP-25 form a tight 7S complex (the SNARE complex) that binds to SNAPs and NSF to form a larger, 20S complex, and that adenosine triphosphate (ATP) hydrolysis by NSF leads to SNARE complex disassembly. These findings, and the belief that NSF was directly involved in membrane fusion, led to the proposal that synaptobrevin 2, syntaxin 1, and SNAP-25 act as receptors for the membrane fusion apparatus and hence to their designation as SNAREs (for SNAP receptors). These results led also to a model of intracellular membrane traffic whereby binding of SNAREs in transport vesicles (v-SNAREs) to SNAREs on target membranes (t-SNAREs) mediates vesicle docking and target specificity (known as the SNARE hypothesis). In this model, the SNAREs were envisaged as binding in an antiparallel fashion and the disassembly of the SNARE complex by NSF/SNAPs was proposed to initiate fusion.

The finding that clostridial neurotoxins inhibit release by cleaving SNAREs from the synaptic vesicle and plasma membrane was consistent with the notion that SNAREs are required on opposing membranes for fusion, which was further supported by subsequent studies of yeast vacuolar fusion. However, it was noted that the SNAREs could not play a role in vesicle docking because docking is not affected on SNARE cleavage by the neurotoxins or by genetic ablation of syntaxin 1 in *Drosophila* (note that very recent data have suggested that syntaxin 1 may actually function in docking; these contradictory conclusions that still need to be resolved probably arise because redundant docking mechanisms may exist and/or because of differences in the methodology used to prepare samples for electron microscopy (EM)). The genetic studies in *Drosophila* also revealed that syntaxin 1 is critical for release and, together with the neurotoxin data, led to the proposal that SNAREs function downstream of docking. Moreover, studies of exocytosis in PC12 cells and of homotypic vacuolar fusion in yeast showed in 1996 that NSF functions at a prefusion step rather than fusion itself, and the SNARE motifs of synaptobrevin 2 and syntaxin 1 were shown in 1997 to interact in an antiparallel fashion by EM and fluorescence resonance energy transfer (FRET). Because the SNARE motifs are

adjacent to the TM regions of synaptobrevin 2 and syntaxin 1, which are anchored on the synaptic vesicle and plasma membrane, respectively, this key discovery led to the proposal that the energy of formation of the SNARE complex could be used to bring the membranes together and initiate membrane fusion (known as the zippering model; Figure 4(a)). In this model, disassembly of the SNARE complex by NSF/SNAPs after membrane fusion recycles the SNAREs for another round of fusion.

The zippering model is attractive because of its similarity to the mechanism of viral fusion and has gained wide acceptance. A similar model that in addition postulates that the SNAREs constitute a minimal membrane-fusion machinery was proposed based on reconstitution experiments, but this minimal model has been strongly challenged by subsequent reconstitutions. Further debate about the central role of the SNAREs in membrane fusion predicted by both models emerged from analyses of mice lacking synaptobrevin 2 or SNAP-25. The synaptobrevin knockout revealed a strong impairment of evoked release and a less marked decrease in spontaneous or hypertonic sucrose-induced release; evoked release was also impaired strongly in SNAP-25-knockout mice, but spontaneous release was increased. Although these results reinforced the notion that the SNAREs function downstream of vesicle docking, the persistence of release in these mice contrasts with the total abrogation of any form of release in Munc18–1-knockout mice. It seems likely that the release remaining in the absence of synaptobrevin 2 and SNAP-25 may arise from functional redundancy with other isoforms of these proteins, but this hypothesis remains to be validated experimentally. Despite these ongoing debates, there is currently a general consensus that SNARE function is related to their ability to bring two membranes together and that synaptobrevin 2, syntaxin 1, and SNAP-25 play a critical role in neurotransmitter release. In addition to resolving these debates, recent and ongoing research on SNAREs is focusing on investigating the mechanism of SNARE complex assembly and studying how the roles of other key proteins are coupled to SNARE function.

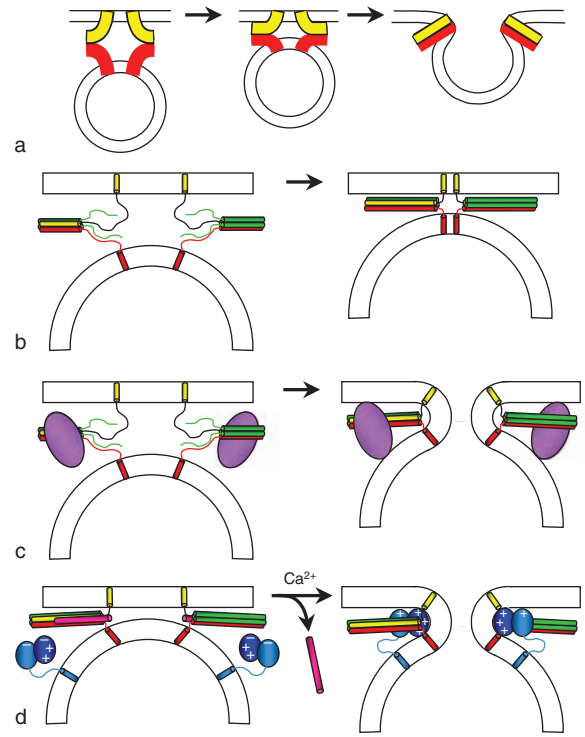

Figure 4 Models of SNARE function in membrane fusion: (a) original zippering model of SNARE-mediated membrane fusion based on EM data, showing that the SNARE motifs of syntaxin 1 and synaptobrevin 2 bind in a parallel fashion; (b) model showing how isolated assembling SNARE complexes might diffuse to the center of the space between the membranes, which requires much less energy than inducing membrane fusion; (c) same model as (b) with the diffusion prevented because the assembling SNARE complexes are bound to a bulky protein (in violet); (d) model illustrating the notion that complexin (in pink) may favor full or almost full SNARE complex formation and at the same time inhibit fusion to yield a metastable state. In (a), the model assumes that the SNARE motifs and TM regions of both proteins form continuous helices and that membrane fusion is forced as these continuous helices zipper from the N- to the C-terminus. In (b) and (c), the model assumes that the SNARE motifs and TM regions of synaptobrevin 2 and syntaxin 1 are connected by short flexible linkers. In (d), the metastable state (left) is predicted to be the substrate for synaptotagmin 1 to trigger neurotransmitter release by binding simultaneously to both membranes and to the SNARE complex, displacing complexin, on Ca^{2+} influx (right). Synaptotagmin 1 is in light blue, except for the C_2B domain, which is in dark blue. The – and + signs illustrate overall features of the surface electrostatic potential of the synaptotagmin 1 C_2 domains that may prevent phospholipid binding in the absence of Ca^{2+} or activate phospholipid binding and membrane fusion upon Ca^{2+} influx. Positive charges at the C-terminus of the SNARE complex that may also help to bend the membranes to induce membrane fusion are not shown. The N-terminal region of syntaxin 1 is not shown in any of the models. EM, electron microscopy; SNARE, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor; TM, transmembrane.

SNAREs and Membrane Fusion

The minimal model of SNARE-mediated membrane fusion was based on reconstitution experiments that demonstrated lipid mixing between proteoliposome populations containing synaptobrevin 2 or syntaxin 1/SNAP-25. This model has been accepted by many researchers, but has also been strongly challenged by others for diverse reasons. On one hand, there were multiple technical problems with the initial reconstitution experiments and mixing of the proteoliposome contents without leakiness, a key feature expected for physiological membrane fusion, has never been

demonstrated; on the other hand, a fundamental problem from a biological perspective is that the minimal model does not account for the strict requirement of SM proteins for all types of SNARE-dependent membrane traffic, including the essential nature of Munc18–1 for neurotransmitter release. Moreover, from a biophysical point of view, the structural-energetic basis of the minimal model is unclear. A key aspect of the original zippering model was the prediction that the SNARE motifs and TM regions of synaptobrevin 2 and syntaxin 1 could form continuous α-helices, which would therefore exert mechanical force on both membranes to induce fusion (Figure 4(a)). However, the available structural data suggest that the SNARE motifs and TM regions are connected by short flexible linkers, which is expected to hamper transduction of the energy of SNARE complex formation into the energy required to bend the membranes and initiate fusion. Hence, as they assemble, the SNARE complexes could diffuse into the space between the membranes and hinder (rather than facilitate) fusion (Figure 4(b)). In addition, introduction of helix-breaking residues or the insertion of long flexible linkers in the sequences between the SNARE motifs and TM regions had little or only moderate effects on the lipid mixing observed in the initial reconstitution experiments. Because the long linkers should have strongly uncoupled SNARE complex formation from its action on the membranes, the observed lipid mixing probably arose in part from factors beyond SNARE complex formation (e.g., instability of the vesicles).

Some of these factors became clear in subsequent reconstitution experiments. Thus, the proteoliposomes used in the initial reconstitutions may indeed have been unstable because they contained extremely high protein-to-lipid ratios (1:20 for synaptobrevin 2, which translates into a protein surface concentration comparable to a 12 mM concentration in bulk solution). Moreover, the proteoliposomes were prepared with a reconstitution method that relies on co-solubilization of proteins and lipids with detergent and yields heterogeneous vesicles (both in terms of size and protein density). Reconstitutions with this method using lower protein densities (e.g., 1:200) yielded less efficient lipid mixing, even though these densities are still comparable to a 1.2 mM protein concentration in bulk solution, and the observed lipid mixing could arise from populations of small vesicles with higher than average protein densities (note that the resulting high curvature destabilizes the vesicles because of membrane stress). Indeed, no significant lipid mixing was generally observed at these proteins densities when vesicles were prepared by a reconstitution method that involves the incorporation of detergent-solubilized SNAREs into preformed liposomes and yields more homogeneous proteoliposomes. Significantly, efficient SNARE complex formation without lipid mixing was observed using proteoliposomes prepared by this method, as predicted from the structural arguments described in Figure 4(b).

From the available reconstitution data, it seems clear that SNARE complex formation is not sufficient for membrane fusion, and lipid mixing can only be induced under conditions in which membranes are destabilized and a large number of SNARE complex force the two membranes to collapse. However, an important notion that emerged from some of the reconstitution experiments is that the TM regions of the SNAREs can destabilize the bilayer, which may contribute to the function of SNAREs in membrane fusion. Moreover, recent reconstitution data obtained with SNAREs, synaptotagmin, and complexin appear to reproduce events that may occur during neurotransmitter release. Hence, the reconstitution approach is providing insights into important features of the release machinery and, with the appropriate methodology, this approach promises to provide a critical tool to understand the mechanism of neurotransmitter release. Note also that, even though the minimal model now appears to be incorrect, the original idea arising from the zippering model that SNARE complex formation causes membrane fusion could still be valid without the requirement that the SNARE motifs and TM regions form continuous helices. Thus, a modified model predicts that the energy of SNARE complex assembly could be used to induce fusion if the SNAREs are kept apart by virtue of interactions with a bulky protein or proteins (Figure 4(c)), which would prevent the undesired diffusion of the SNARE four-helix bundle into the center of the intermembrane space (Figure 4(b)). This model remains to be tested but emphasizes the necessity of understanding the functions of additional factors that are crucial for release in order to fully understand the function of the SNAREs themselves.

SNARE Interactions with Other Conserved Components of the Fusion Machinery

A wide variety of proteins have been shown to play a role in synaptic vesicle exocytosis. Elucidating how the functions of these proteins are coupled to SNARE function is thus critical to understanding the mechanism of release. A difficulty in this endeavor has arisen because of the stickiness of the SNAREs, which has led to the identification of many interactions that are probably irrelevant (for instance, more than 40 proteins have been described that bind to syntaxin 1). Nevertheless, the complementary information provided by genetic and electrophysiological experiments, together with structural studies, have helped to unravel

the nature of some of these interactions and to support their physiological significance.

Among the interactions of the SNAREs with other conserved components of the intracellular membrane fusion machinery, those with SNAPs/NSF have a clear functional role, namely the disassembly of the SNARE complex after fusion to recycle the SNAREs for another round of fusion. Rab proteins, which have been implicated in vesicle docking or tethering, do not appear to interact directly with the SNAREs but have been shown to be functionally coupled to the SNAREs in diverse systems and may indirectly regulate the SNARE complex assembly. The most fascinating and at the same time enigmatic SNARE interactions with the conserved components of the fusion machinery are those with SM proteins. The strict requirement for SM proteins in all types of SNARE-dependent membrane fusion, their localization at sites of fusion, and their diverse interactions with SNAREs suggest that SM proteins may have a direct function in fusion. However, this function has remained elusive, in part because of the diversity of these interactions. Neuronal Munc18–1 was shown to bind tightly to the closed conformation of syntaxin-1, but this interaction hinders SNARE complex formation and hence does not explain the essential nature of Munc18–1 for neurotransmitter release. Conversely, the yeast plasma membrane SM protein Sec1p was found to bind to SNARE complexes instead of to the cognate syntaxin. Moreover, syntaxins from diverse intracellular compartments in yeast and mammals bind to SM proteins through their NTS, in an interaction that is compatible with SNARE complex assembly. Finally, the yeast vacuolar SM protein Vps33p appears to bind to SNARE complexes at least in part through the N-terminal PX domain of the SNAP-25 homolog Vam7p.

This diversity of SM protein–SNARE interaction modes has probably emerged from distinct regulatory requirements of traffic in different membrane compartments, but it seems likely that a common binding mechanism between these two conserved protein families must exist. In this context, increasing evidence has suggested that all SM proteins bind to SNARE complexes, and a direct interaction between Munc18–1 and neuronal SNARE complexes was recently identified. Hence, Munc18–1 exhibits at least two different modes of interactions with the SNAREs: one with the syntaxin 1 closed conformation, which probably represents a specialization that evolved for the tight regulatory requirements of neurotransmitter release, and another with the SNARE complex, which may be conserved in all types of intracellular membrane traffic. The former interaction may play a role in stabilizing syntaxin 1 or preventing this promiscuous protein from binding to other proteins. In addition, syntaxin 1 has been shown to form two different types of four-helix bundles with SNAP-25, containing two copies of syntaxin 1; these complexes represent kinetic traps that hinder SNARE complex formation, and binding of Munc18–1 to the syntaxin 1 closed conformation may prevent the formation of these traps. Because the closed conformation contains a four-helix bundle and Munc18–1 also binds to the SNARE complex, it has been proposed that Munc18–1 may provide a template for assembling the four-helix bundle formed by the SNARE motifs of syntaxin 1, SNAP-25, and synaptobrevin 2 in the same site. In this context, Munc18–1 could play the role of the bulky protein depicted in the model in **Figure 4(c)**, but this model remains to be tested and the precise nature of the Munc18–1–SNARE complex interaction still needs to be defined. Moreover, alternative models can be envisaged for the active role of Munc18–1 in release. For instance, the SNARE motifs of synaptobrevin 2 and syntaxin can bind in an antiparallel fashion, which hinders fusion; Munc18–1 may ensure assembly of the SNARE complex in the proper, parallel orientation.

The transition from the Munc18–1–syntaxin 1 complex to the Munc18–1–SNARE complex may be a central event in the priming reaction that makes docked synaptic vesicles readily releasable, but the mechanism of this transition is also unclear. Functional experiments in *Caenorhabditis elegans* have suggested that this transition may be mediated by Unc13/Munc13s and Unc10/Rab3 interacting molecules (RIMs), which are large proteins from presynaptic active zones with critical roles in release. Unc13/Munc13s were initially thought to bind to syntaxin 1, but recent evidence indicates that they do not form binary complexes with syntaxin 1. Hence, gaining insight into the biochemistry of these proteins, the interactions underlying the conformational transition of syntaxin 1, and the role of the Munc18–1–SNARE complex interaction will be crucial to the understanding of the mechanism of neurotransmitter release. It will also be critical to better understanding which aspects of the coupling mechanism between Munc18–1 and the neuronal SNAREs are conserved in other systems and may reflect the general function of SM proteins in membrane fusion, although it is becoming increasingly clear that this function is somehow coupled to SNARE complex assembly.

SNARE Interactions with Specific Components of the Release Machinery

Interactions of the SNAREs with components of the exocytotic machinery that have specialized roles in

release also appear to be associated with the regulation of SNARE complex assembly, but may also attract factors that facilitate and/or inhibit membrane fusion to confer the exquisite Ca^{2+} sensitivity of neurotransmitter release (the fast, synchronous components of release is triggered in less than 100 μs after Ca^{2+} influx in some synapses). An example of the former category is the interaction of syntaxin 1 and SNAP-25 with tomosyn, a large protein with a SNARE motif homologous to that of synaptobrevin 2. Indeed, the crystal structure of tomosyn bound to syntaxin 1 and SNAP-25 revealed a four-helix bundle analogous to the SNARE complex but with synaptobrevin replaced by the tomosyn SNARE motif (Figure 5(a)). Therefore, formation of this complex is expected to inhibit release by preventing the assembly of functional SNARE complexes containing synaptobrevin 2. This prediction has been confirmed by genetic experiments in *C. elegans*, although tomosyn may play additional functions through the long sequences preceding the SNARE motif. These experiments revealed at the same time a functional interplay between tomosyn and Unc13/Munc13s that may control the balance between unproductive tomosyn–syntaxin 2–SNAP-25 complexes and productive SNARE complexes containing synaptobrevin 2.

Diverse functional data have suggested that the SNAREs are directly or indirectly coupled to Ca^{2+} sensing during neurotransmitter release, including the alteration of the Ca^{2+} sensitivity of secretion in chromaffin cells caused by some point mutations in SNAP-25 and the finding that elevated Ca^{2+} can compensate the inhibition of release caused by botulinum neurotoxin A, which cleaves SNAP-25 close to its C-terminus. Whereas NMR studies showed that the SNARE complex does not contain specific Ca^{2+}-binding sites that could be directly involved in Ca^{2+} sensing during release, it has become increasingly clear that interactions of the SNAREs with synaptotagmin 1 and complexins are key for the Ca^{2+}-triggered step of neurotransmitter release.

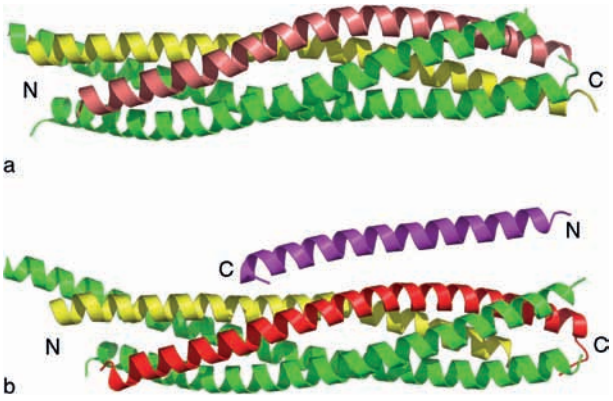

Figure 5 Structural basis for SNARE interactions: (a) with tomosyn; (b) with complexin. Crystal structures of the complex formed by the tomosyn, syntaxin 1, and SNAP-25 SNARE motifs (Pobbati et al.) (a) and the complexin–SNARE complex (Chen et al.) (b). Tomosyn is in salmon, complexin is in pink, and the remaining color coding is as in Figure 1. Note the similarity between the structures of the tomosyn–syntaxin 1–SNAP-25 complex (a) and the SNARE complex (Figure 2(a)), with the tomosyn and synaptobrevin 2 SNARE motifs occupying the same positions in the complexes. Note also that complexin forms an α-helix that binds to a groove between the syntaxin 1 and synaptobrevin 2 SNARE motifs in the SNARE complex (b). C, C-terminus; N, N-terminus; SNAP-25, synaptosomal associated protein of 25 kDa; SNARE, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor.

Synaptotagmin 1 is a synaptic vesicle protein that acts selectively as a Ca^{2+} sensor in synchronous release. This function depends on Ca^{2+}-dependent phospholipid binding to the two C_2 domains that form most of the synaptotagmin 1 cytoplasmic region (the C_2A and C_2B domain), with the C_2B domain playing a preponderant role. Many studies described interactions of the SNAREs with synaptotagmin 1, but it was unclear which of these interactions might be functionally relevant. Recent data may have yielded key insights into this issue. Thus, Ca^{2+}-bound synaptotagmin 1 was shown to bind simultaneously to the C-terminus of the SNARE complex and to phospholipids through the C_2B domain, and the C_2B domain was also found to interact simultaneously with two membranes on Ca^{2+} binding. Moreover, both the Ca^{2+}-bound C_2B domain and the C-terminus of the SNARE complex are highly positively charged. These observations suggest that the SNARE complex and the C_2B domain may cooperate in bringing the membranes together and in bending them to accelerate fusion (Figure 4(d), right). Although these ideas provide an attractive mechanism for coupling synaptotagmin 1 and SNARE function that explains the preponderant role of the C_2B domain, their validity remains to be demonstrated and alternative models can be envisaged. For instance, it is also possible that the highly positive electrostatic potential generated by the SNARE C-terminus and synaptotagmin 1 may help to open the fusion pore if the synaptic vesicle and plasma membranes are already partially merged (forming a hemifusion state) before Ca^{2+} influx.

Complexins are small soluble proteins that bind tightly to the neuronal SNARE complex in a Ca^{2+} independent manner. Knockout of the two major complexin isoforms in mice revealed a selective impairment of the Ca^{2+}-triggered step of synchronous release, but an excess of complexin can also inhibit release. X-ray crystallography and NMR spectroscopy showed that complexin binds in an antiparallel α-helical conformation to a groove between the SNARE motifs of synaptobrevin 2 and syntaxin 1 (Figure 5(b)), stabilizing the SNARE complex.

Because SNARE complexes are generally believed to assemble from the membrane-distal N-terminus and zippering toward the C-terminus is expected to be hindered by membrane repulsion, these findings led to a model whereby complexin helps to complete SNARE complex assembly, yielding a metastable state that is critical for the fast speed of synchronous release (Figure 4(d), left). Interestingly, recent reconstitution experiments showed that complexin inhibits SNARE-induced liposome fusion at a hemifusion state and that Ca^{2+} plus synaptotagmin 1 release the inhibition, resulting in a fast burst of fusion that cannot be observed with the SNAREs alone. In addition, simultaneous binding of the synaptotagmin 1 C_2B domain to phospholpids and SNARE complexes was shown to displace a complexin fragment bound to these complexes. Altogether, these data suggest that complexin may indeed play an active role in release by promoting SNARE zippering toward the C-terminus and thus bringing the system closer to fusion (perhaps to hemifusion), but at the same time plays an inhibitory role to prevent full fusion before Ca^{2+} influx; such inhibition is then released by simultaneous binding of synaptotagmin 1 to Ca^{2+}, to both membranes, and to the SNARE complex to active release (Figure 4(d)). Clearly, there are alternative models to explain the available data, including the active function of complexin for release, but these data show that there is a fascinating interplay among the SNAREs, complexin, synaptotagmin 1, Ca^{2+}, and phospholipids that is probably crucial for the exquisite regulation of neurotransmitter release. The relation between Munc18–1, complexin, and synaptotagmin 1 binding to the SNARE complex remains to be studied.

SNARE Specificity

The contribution of SNAREs to the specificity of membrane traffic has been another area of strong debate about SNARE function. This issue is important for interpreting the phenotypes of synaptobrevin 2 and SNAP-25 mice because, given the observation that neurotransmitter release can still be observed in these mice, an absolute specificity in SNARE pairing implies that the SNAREs are not essential for membrane fusion. Biochemical assays showed that some SNAREs have a tendency to bind to both cognate and noncognate SNAREs, and hence can be quite promiscuous. Conversely, reconstitution studies suggested a high specificity in SNARE-induced lipid mixing, but the noncognate SNARE pairings identified biochemically were not analyzed. A recent study showed that early endosome fusion is specifically mediated by a set of SNAREs in living cells, whereas the same set of SNAREs promiscuously induce liposome fusion with SNAREs from the plasma membrane or late endosomes.

These results can be interpreted in the framework of sequence analyses and the known structures of SNARE complexes. Because SNARE complex assembly involves protein–protein interactions, assembly must involve some degree of specificity. Moreover, the SNAREs are one of the most diverse among the protein families generally involved in intracellular membrane traffic (25 members in yeast and 36 in humans; only Rab proteins are more diverse). However, because coiled-coil interactions can be considerably promiscuous and the hydrophobic residues involved in SNARE complex formation are highly conserved in each subfamily, it is not surprising that some degree of nonspecificity exists in SNARE pairing. Hence, the picture that emerges is that SNAREs contribute to the specificity of traffic in distinct membrane compartments but that this specificity also involves other proteins. In particular, there is little doubt that an important contribution to specificity arises also from Rab protein interactions. Note also that the contribution of the neuronal SNAREs to specificity in neurotransmitter release most likely also involves interactions with factors specialized for release, such as synaptotagmin 1 and complexins.

Outlook

Research on SNARE proteins has provided an emphatic example of the power of interdisciplinary approaches to studying protein function and of the difficulties of exactly pinpointing the specific roles of proteins that act as part of complex macromolecular assemblies. Hence, the divide-and-conquer approach has provided crucial clues to the understanding of the functions of the SNAREs and other components of the release machinery, but increasing evidence illustrates the complexity and cooperativity of this system. Clearly, a true understanding of the mechanism of release requires further studies in which more and more of these components are examined together in the presence of membranes. A critical foundation for these studies has been provided by the astonishing advances in this field made during the past 20 years.

See also: Calcium Channels and SNARE Proteins; Munc13 and Associated Molecules; Munc18; Neurotransmitter Release: Synchronous and Asynchronous; Neurotransmitter Release from Astrocytes; Synaptic Vesicles.

Further Reading

Brunger AT (2005) Structure and function of SNARE and SNARE-interacting proteins. *Quarterly Review of Biophysics* 38: 1–47.

Dulubova I, Sugita S, Hill S, et al. (1999) A conformational switch in syntaxin during exocytosis: Role of Munc18. *EMBO Journal* 18: 4372–4382.

Hanson PI, Roth R, Morisaki H, Jahn R, and Heuser JE (1997) Structure and conformational changes in NSF and its membrane receptor complexes visualized by quick-freeze/deep-etch electron microscopy. *Cell* 90: 523–535.

Jahn R and Scheller RH (2006) SNAREs – engines for membrane fusion. *Nature Reviews Molecular and Cell Biology* 7: 631–643.

Lin RC and Scheller RH (1997) Structural organization of the synaptic exocytosis core complex. *Neuron* 19: 1087–1094.

Link E, Edelmann L, Chou JH, et al. (1992) Tetanus toxin action: Inhibition of neurotransmitter release linked to synaptobrevin proteolysis. *Biochemistry and Biophysics Research Communications* 189: 1017–1023.

Nichols BJ, Ungermann C, Pelham HR, Wickner WT, and Haas A (1997) Homotypic vacuolar fusion mediated by t- and v-SNAREs. *Nature* 387: 199–202.

Poirier MA, Xiao W, Macosko JC, et al. (1998) The synaptic SNARE complex is a parallel four-stranded helical bundle. *Nature Structural Biology* 5: 765–769.

Rizo J, Chen X, and Arac D (2006) Unraveling the mechanisms of synaptotagmin and SNARE function in neurotransmitter release. *Trends in Cell Biology* 16: 339–350.

Schiavo G, Benfenati F, Poulain B, et al. (1992) Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin. *Nature* 359: 832–835.

Schoch S, Deak F, Konigstorfer A, et al. (2001) SNARE function analyzed in synaptobrevin/VAMP knockout mice. *Science* 294: 1117–1122.

Schulze KL, Broadie K, Perin MS, and Bellen HJ (1995) Genetic and electrophysiological studies of *Drosophila* syntaxin-1A demonstrate its role in nonneuronal secretion and neurotransmission. *Cell* 80: 311–320.

Sollner T, Bennett MK, Whiteheart SW, Scheller RH, and Rothman JE (1993) A protein assembly-disassembly pathway in vitro that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion. *Cell* 75: 409–418.

Sollner T, Whiteheart SW, Brunner M, et al. (1993) SNAP receptors implicated in vesicle targeting and fusion. *Nature* 362: 318–324.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Social Brain: Evolution

R I M Dunbar, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Explaining Brain Evolution

Although neuroscientists do not often consider the evolutionary history of the brain, understanding how and why brains have evolved may help explain why they have come to have the structure they do. Vertebrate species, for example, vary enormously in both absolute brain size and brain size relative to body size, but some species (notably primates) have disproportionately large brains. Why should that be? Why do all mammals not have primate-like brains? A second issue is that brain researchers have conventionally assumed that the brain's main function is to acquire and process information about the environment, usually information directly relevant to the business of survival (i.e., finding food and avoiding predators). While obviously broadly correct, this view contains an implicit assumption that may be less defendable: the assumption that 'environment' here means the physical environment. Of course, a great deal of brain activity is about the physical environment, and, in the final analysis, it is the physical properties of the environment that constitute all perceptual inputs.

However, for more advanced vertebrates (birds and mammals), and in particular the more intensely social of these species, the environmental behavior that needs to be processed by the brain is that of the 'social' world. In the limiting case, the social world is a virtual world of behavior, the nature of which is understood and manipulated only in the animal's mind. Because the dynamic complexities of this virtual social world vastly exceed the complexities of the physical world, it is inevitably much harder to understand. Hence, intensely social species (such as primates) have brains that are specialized for dealing with this aspect of their world.

The emphasis on primates in the early literature on comparative brain evolution led to a natural focus on neocortex volume. The scaling relationships among mammalian brain units are such that neocortex volume increases much faster than that of all other parts of the brain, and especially among the primates. Although the neocortex accounts for as little as 10% of total brain volume in some of the more primitive insectivores and at most ~40% in nonprimate mammals, it accounts for a minimum of 50% in primates and rises to as high as 80% in humans. Thus, while the neocortex may be viewed as a mammalian invention, large neocortices are clearly a primate speciality – so much so, in fact, that when one asks why primates have big brains, one is really in effect asking why they have big neocortices.

In the following sections, the selective advantages of large brains are first discussed, in order to identify why large brains have evolved in various animal lineages. Then the role that cognition plays in these processes is considered. Finally, the question of whether these forms of social cognition relate to specific brain structures is briefly explored.

Theories for the Evolution of Large Brains

We might expect large-bodied species to have large brains simply because they have larger muscle masses to coordinate if they are to survive effectively. However, it has long been recognized that some groups of species (e.g., primates) have unusually large brains for their body size compared to all other groups of vertebrates. Why this should remain so is unclear, although it has been generally assumed that it probably has something to do with the ecology of these species.

Initial Explanations

Initial attention was drawn to the fact that, within many mammalian taxa, species with larger brains have larger territories, or have diets that are more difficult to locate or process (e.g., diets of ephemeral fruits as opposed to more widely available leaves, or of invertebrates that have to be extracted from hiding places). Others have pointed to correlations with altriciality (giving birth to babies that are not mature enough at birth to move and forage for themselves) and the scaling of maternal energy demands, and have argued that it was spare maternal energy capacity that made large brains possible. By the late 1980s, however, this assumption began to be questioned: it eventually became clear that the kinds of ecological tasks faced by primates do not differ much from those faced by other species. This realization begged the question as to why primates should need a substantially larger brain (in effect, more computing power) to solve the same foraging problems as, say, cows. The social brain hypothesis was proposed as an explanation.

The initial proposal, known as the 'Machiavellian intelligence hypothesis,' was intended to explain why primates had larger brains for body size than all other species. It was suggested that this might have something to do with the fact that primates live in unusually complex social systems. Primate sociality is characterized by kinds of behaviors (for example,

tactical deception, the formation of coalitions and alliances long before they are actually needed) that are rare or entirely absent in nonprimate species. This proposal was later renamed the 'social brain hypothesis' (or 'social intelligence hypothesis') in order to avoid unnecessary and unintended implications of sociopolitical scheming and to focus attention more closely on the positive socially integrating aspects that really underpin primate sociality.

Asking the Right Questions

An important consideration at the outset is to appreciate that there is an important distinction between functional hypotheses (those that provide a reason why large brains might be selected for) and explanations that are essentially about constraints. Most developmental hypotheses fall into the latter category: they identify trajectories along which brain evolution is forced to run, or system limitations that have to be overcome. Examples of the first include the fact that the way brain tissue is laid down necessarily results in unavoidable scaling of all brain parts: if such scaling is very tight, then it may result in all parts of the brain enlarging proportionately when evolution imposes selection on just one component. This might mean that in order to evolve more capacity in some key brain component, it is necessary to evolve everything on a larger scale. Examples of the second kind of limitation would be energetic constraints that limit brain growth: some diets may not provide sufficient surplus energy that can be channeled into brain growth. Consequently, species that need to evolve a larger brain may only be able to do so if they first change diet in order to bypass the energetic constraint.

While these kinds of developmental hypotheses are perfectly legitimate and tell us something about how large brains are evolved, they do not specify why they should evolve. Given that brain tissue is among the most 'expensive' in the body (the so-called expensive tissue hypothesis), to merely show that evolutionary change in a particular direction is possible is not enough. The costs of evolving (and indeed growing) a large brain are such that they provide an extremely steep gradient, up which selection has to push the organism. In the absence of any selective benefit (or advantage), the costs impose strong negative selection that will constantly favor a reduction in size (or, at best, a stable state). Some real advantage of having a large brain is needed to counteract that natural stabilizing effect.

Hypotheses for Brain Evolution

In considering likely selective factors that have been responsible for driving increases in brain size, we are left with two main contenders that can broadly be defined as ecological theories and social theories. Ecological hypotheses argue for a focus on foraging decisions (how to extract nutrients from the environment, how to find one's way about, etc.). In contrast, the social hypotheses argue that the computational load is created by the demands of tracking social relationships and calculating the pros and cons of alternative social actions.

In both cases, of course, the fundamental issue is the brain's problem-solving capabilities, but the focus of the problem and the mode of solution are different in each case. However, it is worth emphasizing that, either way, the root issue is essentially ecological: all animals have to overcome problems of day-to-day survival if they are to contribute offspring to the next generation (the root of all evolutionary processes). The difference lies in how this is done, and which particular computational demand imposes the highest cognitive load. Ecological hypotheses assume that the basis of an animal's decision-making in this respect is explicitly trial-and-error learning on an individual basis; the social hypotheses assume that the ecological problems are solved cooperatively (and that the limiting cognitive demand is created by the need to coordinate one's behavior with that of others in some way).

One way to test between these competing hypotheses is to correlate some measure of relative brain size with one or more behavioral indices that represent the hypothesis of interest. Relevant indices that have been used include the amount of fruit in the diet (representing the claim that the cognitive constraint lies in finding and monitoring high-quality food sources such as fruiting trees), the style of extractive foraging (whether or not food items are extracted from some physical matrix, such as a pod, or dug out of the ground, i.e., an index of complex food-processing decisions), day journey length and territory size (indices of mental mapping capacities), and social group size (a simple index of social complexity). When these indices have been correlated with indices of relative brain size, only group size is correlated with brain size (and specifically with neocortex size) (see **Figure 1**).

More recently, a number of new analyses have been undertaken that use new and more powerful statistical methods that allow the multivariate influences of a wider range of variables to be taken into account simultaneously. These have been able to show that, while other strictly ecological variables (notably, diet, diurnality, and foraging strata) play a role in brain evolution, sociality and/or social group size remain the strongest predictors of brain/neocortex volume in primates. In addition, the scope of the social brain

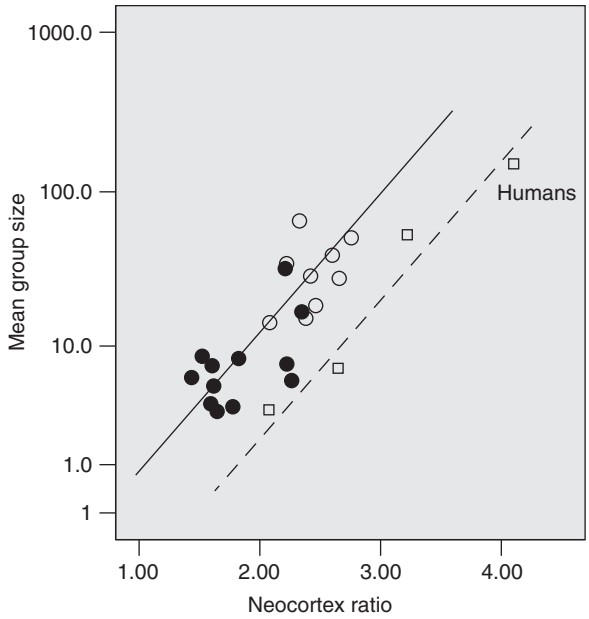

Figure 1 $\log_{10}$ Mean social group size for different primate species plotted against relative neocortex volume (indexed as neocortex volume divided by the rest of the brain, so as to remove scaling effects due to body and brain size). Solid symbols are New World monkeys; open circles are Old World monkeys; squares are apes. Note that the apes lie on a separate grade to the monkeys: they seem to require more computing power to handle groups of a particular size than monkeys do, suggesting that their social life is cognitively more complex.

hypothesis has now been convincingly extended to other mammalian orders, including carnivores, ungulates, whales, and bats, and even to birds. In all these cases, brain size is larger relative to body size in more social species than in relatively asocial species. However, there appears to be an important distinction between the primates and all other species in the way this process works. While the main effect of this relationship in primates is a quantitative effect with group size, in all other species the effect shows up most clearly as either a contrast between social and asocial species (those that live alone, except during the mating season) or a strong effect of monogamy (monogamous species have larger brains than do those living in other mating systems). The same tendency for monogamous (i.e., pairbonded) species to have disproportionately large brains is still present in primates, but it is much overshadowed by the strong quantitative group size effect. The implications of this are explored in the following sections.

In summary, the social brain hypothesis has proved to be an extremely robust explanation for the evolution of brain (but especially neocortex) size in primates, and more generally for the evolution of brain size in mammals as a whole. It has proved to be more successful in explaining the pattern of brain size evolution than have any of the alternative ecological hypotheses.

Social Complexity

Although the social brain hypothesis has been successfully tested using simple demographic indices like social group size or differences in mating system, its real focus lies in the subtleties of behavior. The argument is that group size is ultimately limited by the animals' abilities to maintain coherent social relationships with other individuals. Group size is thus an emergent property of animals' abilities to manage their social relationships.

This has been borne out by evidence that a number of indices of behavioral complexity also correlate with relative brain size in primates. These have included the use of coalitions, the size of alliances (indexed as grooming cliques), the frequency of tactical deception (behavior that misleads another individual), and the use of subtle social tactics that allow low-ranking males to undermine the dominant males' monopolization of matings. The fact that monogamous pairbonding (and especially lifelong pairbonding) is particularly strongly associated with large brain size across three major mammalian orders (primates, carnivores, and ungulates) likewise emphasizes the importance of social complexity, since pairbonded social systems are only possible if mates have the sociocognitive skills to negotiate a relationship in which each individual's needs are properly integrated into an effective partnership. This is most obvious, perhaps, in birds, where the pair has to coordinate and schedule two major activities: sitting on the nest (to keep eggs or fledglings warm and protected) and foraging.

The significance of social complexity as opposed merely to sociality, at least in primates, is given further emphasis by the fact that development can be shown to play an important role. While it has long been known that total brain size correlates with the level of parental investment (as indexed, for example, by the duration of pregnancy and lactation), neocortex volume correlates best with the period between weaning and puberty (i.e., the main period of socialization before the animal finally joins the adult world). This suggests that growing a big brain depends on how long the mother can afford to invest in direct parental care, but having a socially complex brain (and this really means neocortex) depends on the time you can afford to invest in socialization (in effect, the equivalent of the software programming). Further evidence is provided by the fact that, in primates, neocortex size also correlates with the proportion of all play that

is social (as opposed to instrumental or solitary): most social play takes place during this main period of socialization between weaning and puberty.

Thus, although brain size is the central issue in the social brain hypothesis, it is important to appreciate that the situation is rather different from that in the case of more conventional brain systems. A distinction needs to be drawn between what we might think of as first-order processing systems (those in which sensory inputs are analyzed for their intrinsic patterns) and those which involve secondary or even tertiary analyses of these inputs (e.g., those in which meaning is attached to the sensory patterns). While sensory processing mechanisms function more or less efficiently from the moment the animal engages with the world, the upstream neural systems responsible for adding the layers of social meaning (and which thus constitute the heart of the social brain) seem to require extensive tuning before they can operate at full capacity. That tuning presumably occurs in the social environment provided during development.

The Nature of Social Cognition

We do not have a clear idea of how cognition interfaces between the neural systems of the brain and the observable behavior of the animals in the social domain. However, there is general agreement that, at least in primates, it involves specialized processes that are collectively but loosely termed 'social cognition.' In humans, these are identified with the cognitive mechanism known as 'theory of mind' or 'mentalizing' (the ability to understand the minds of other individuals). Although the exact nature of this phenomenon is still a subject of debate, it provides at least a useful way of thinking about the general cognitive demands of sociality.

Theory of mind refers to the capacity to mentally represent another individual's mind state: "I *believe* that you *suppose*. . . ." This is generally understood to be equivalent to what philosophers of mind refer to as second-order intentionality, where intentionality refers to the capacity to use words that involve the 'intentional stance' (in other words, the ability to reflect on one's intentions, or more generally one's beliefs about the world). In principle, intentionality forms a reflexive hierarchy that is infinite: "I *believe* that you *suppose* that I *intend* that you *want* me to *understand*. . . ." Theory of mind is generally held to be acquired by children at around age 4 years. Prior to that, children, in common with all nonhuman animals, have only first-order intentionality: in effect, they understand their own minds, but not those of other individuals. The only exception to this is the chimpanzee (and possibly other great apes), for whom there is some (albeit weak) evidence for second-order intentionality. In contrast, normal adult humans are able to cope with fifth-order intentionality. These additional levels of intentionality appear to be acquired sequentially by the early teenage years.

Although we know quite a lot about theory of mind (but rather less about the higher levels of intentionality), in fact we do not really know what it actually is. There are two views on this: one views theory of mind as a specialized cognitive module; the other views it as simply a product of more conventional executive functions (memory, causal reasoning, etc.). An alternative view is that theory of mind is an emergent property of a number of other higher order cognitive abilities, including causal reasoning, analogical reasoning, the time frame over which consequences of actions can be forecast, the ability to compare two behavioral trajectories through time, etc. These might be viewed as 'middle-range' cognitive modules that intervene between the more conventional Fodorian modules (those associated with basic percept processing) and the higher order inferential processes associated with social cognition.

In effect, theory of mind is what 'pops' out when an individual can do all of these on a large-enough scale and focus them sufficiently onto a single problem. In other words, it is an emergent property of cognitive mechanisms that are both general and essential for everyday life. This being so, on a simple information-processing model, we might expect the ability to engage these middle-range processing capacities to be a function of brain size. Presumably, the capacity to engage in higher levels of intentionality beyond theory of mind (level 2) is simply a reflection of a correspondingly greater computational capacity.

Cognition and the Brain

This brings us to the last issue we need to consider, namely exactly how these higher order aspects of social cognition relate to the brain. There appear to be two possible positions on this. One is that several decades of research in neuroscience is enough to make it clear that the interconnectedness of the brain is of such an order that it does not make sense to consider the brain as modularized in any naïve sense: it functions as an integrated whole, in which processing capacities of different kinds are called up from different cortical and subcortical regions. The alternative view is that the specialized cognitive functions identified with theory of mind and social cognition can be equated with particular brain regions (specifically in the neocortex).

While it may be true that brain function does commonly involve activity in many parts of the brain that

are physically distant from each other, nonetheless there is convincing evidence that these higher-level skills are associated with specific neocortical regions, and specifically with the frontal lobe (an area of the brain that has proved to be singularly resistant to attempts to partition it into functional subunits). There is considerable clinical evidence, from the famous case of Phineas Gage in the 1890s onwards, that stroke or physical injury to the frontal lobes is particularly likely to disrupt social skills (and only social skills). However, several other sources of data also suggest that the frontal lobe of the neocortex plays a critical role in social cognition.

First, a detailed analysis of data for primates shows that when all other brain components are held constant, only neocortex size correlates with social group size; no other brain unit has any association with group size. Indeed, the correlation between neocortex volume and social group size improves measurably if the primary visual cortex is removed, and improves still further if only the frontal lobe volume is considered. In other words, the more one focuses on the frontal lobe, the better the relationship seems to be. Second, such differences as there are in mentalizing capacity between primate species (essentially first-order intentionality in monkeys, second order in apes, and fifth order in adult humans) are linearly related to frontal lobe volume, but not to any other part of the brain. Finally, neuroimaging studies have emphasized a particularly important role for frontal lobe units in social cognitive tasks (even though units elsewhere in the parietal and temporal lobes may also be involved).

One reason why the more occipital areas of the brain may not play a central role is that, as the principal visual areas, they are mainly concerned with primary and secondary visual processing, and the computational demands for this are largely dependent on the sensory input from the retina. Indeed, there is a near linear relationship in primates between orbit volume (and hence retinal area) and volumes of the successive segments of the visual pathway (optic nerve, lateral geniculate nucleus, and primary visual cortex in the occipital lobe). This is, in part at least, a consequence of the fact that the size of the retina is the rate-limiting factor in visual acuity, and there is little point in having a cortical visual processing capacity that is more powerful computationally than its input. Indeed, this seems to be a general principle in brain system organization. Hence, with orbit size relatively constant across primate species, any increases in brain volume are mainly associated with more frontal brain units.

Indeed, the brain (and neocortex in particular) is known to both evolve and develop from back (occipital lobe) to front, so that increases in brain volume are largely associated with increases in frontal lobe volume. Plotting 'spare' neural volume over and above that predicted by primary visual area (Figure 2) suggests that the volume of spare capacity (most of which is in the frontal lobe and thus supports executive function) starts to increase disproportionately at the brain size of great apes (i.e., just at the point where we have the first evidence for second-order intentionality) and is massive in modern humans (with their fifth-order capacities).

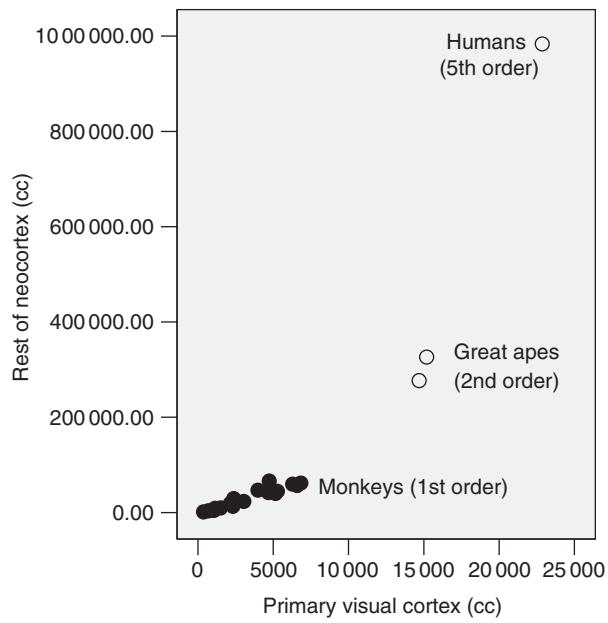

Figure 2 Neocortex volume excluding the primary visual area plotted against the primary visual area for different monkey and ape species. Solid symbols, monkeys; open symbols, apes. The maximum achievable level of mentalizing (intentionality order) is indicated beside each major group, and suggests that brains of the size seen in the great ape mark a critical juncture at which spare neural capacity in the more frontal units of the brain suddenly becomes available, thus allowing higher orders of mentalizing to develop.

See also: Brain Connectivity and Brain Size; Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brain Scaling Laws; Evolution of Vertebrate Brains; Neuroendocrinology of Social/Affiliative Behavior; Social Interaction; Social Interaction Effects on Reward and Cognitive Abilities in Monkeys.

Further Reading

Aiello LC and Wheeler P (1995) The expensive tissue hypothesis: The brain and the digestive system in human evolution. *Current Anthropology* 36: 199–221.

Barrett L, Henzi SP, and Dunbar RIM (2003) Primate cognition: From 'what now?' to 'what if?'. *Trends in Cognitive Sciences* 7: 494–497.

Barton RA (1996) Neocortex size and behavioural ecology in primates. *Proceedings of the Royal Society of London B* 263: 173–177.

Byrne RB and Whiten A (eds.) (1988) *Machiavellian Intelligence*. Oxford: Oxford University Press.

de Winter W and Oxnard CE (2001) Evolutionary radiations and convergences in the structural organisation of mammalian brains. *Nature* 409: 710–714.

Dunbar RIM (1998) The social brain hypothesis. *Evolutionary Anthropology* 6: 178–190.

Finlay BL, Darlington RB, and Nicastro N (2001) Developmental structure in brain evolution. *Behavioral and Brain Sciences* 24: 263–308.

Jerison HJ (1973) *Evolution of the Brain and Intelligence*. New York: Academic Press.

Keverne EB, Martel FL, and Nevison CM (1996) Primate brain evolution: Genetic and functional considerations. *Proceedings of Royal Society of London B* 262: 689–696.

Social Cognition

J L Marchant and C D Frith, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In everyday life we constantly interact with the people around us, whether it is to cooperate, compete, or simply to go about our day-to-day business. For such interactions to be successful we must be able to understand and predict the actions of these other people. We typically understand actions in terms of minds: beliefs, desires, and intentions. This is termed having an 'intentional stance' or having a 'theory of mind,' while our ability to 'read' mental states has been called 'mentalizing.' A range of functional imaging studies have attempted to identify the neural correlates of mentalizing when participants make 'off-line' inferences about the mental states of interacting characters in stories, cartoons, and animations. A number of brain regions have been consistently activated by these tasks, including the medial prefrontal cortex (MPFC), temporal poles, and temporoparietal junction.

Yet, like most psychological and neuroimaging studies, these studies were investigations of people in isolation. Even imaging studies of social cognition do not typically involve true interactions. Brain activity is measured in various social contexts or when an individual thinks about her own mental states or those of another person. In all of these paradigms the flow of information is one way. The participant simply responds to a facial expression or to a social scenario. In a true interaction the information flows both ways. We do not simply read the signals of others. We send back signals for them to read. We do not simply respond passively to other persons, we respond actively in order to change them: to make them trust us or fear us. For such interactions, it is not sufficient to represent our own mental state or the mental state of the other. We need also to represent the other's representation of our mental state. In a successful interaction our mental states are effectively shared. Perhaps the simplest example of such an interaction is joint attention, in which each participant knows that there is mutual attention to the same object. However, it is only recently that studies have started to investigate how brain systems are engaged when we directly interact with another person in order to perform 'on-line' mentalizing.

The most basic form of social interaction is joint action in which two people cooperate in order to achieve some common goal. In order to successfully achieve their shared goal, participants must understand how they and their partner view each other's roles within the partnership. This process involves at least four levels of mentalizing: (1) our belief about our partner's role, (2) our belief about how our partner views his role, (3) our belief about how our partner believes we view our role, and finally (4) our belief about how our partner believes we view him. If we are able to represent all levels of this mentalizing structure, we should be able to understand the intentions of our partner and predict his actions.

This section summarizes findings from recent studies in which two participants socially interact in a variety of tasks. We start by noting some important behavioral studies and then discuss neuroimaging findings that suggest an important role for the medial frontal cortex, including the anterior cingulate cortex (ACC), in on-line social interaction. Next, we highlight the new breed of 'hyperscanning' studies that go beyond analysis of a single brain to look at systems of interacting brains.

What Can Behavioral Studies Tell Us?

A variety of experiments have been used to explore cooperative behavior between humans and nonhuman animals. Tasks have included attainment of a reward that is not accessible to the individual, successful completion of a joint action task, and exchange of information in a communicative partnership. Such cooperation is highly prevalent in human society. Children as young as 18–24 months old are able to interact with an adult human in a cooperative activity, unlike their chimpanzee counterparts.

Even though language plays a vital role in most cooperative endeavors, it is not necessary for successful cooperation. As long as we can monitor the actions of others we can identify their goals and intentions and modify our own behavior accordingly. Indeed, the mere presence of another person performing a task with us can modify our own response. Sebanz and colleagues used a spatial compatibility task wherein participants made one of two spatially distinct actions in response to a color cue and in the presence of an irrelevant spatial cue. When participants worked alone and assigned one hand to each of two response alternatives, reaction times were slowed when there was an incompatibility between the desired response and the irrelevant spatial cue. This effect disappeared when the participant responded to only one of the two alternatives. However, the

incompatibility effect returned when a second participant was assigned to respond to the other alternative. This result suggests that we automatically represent the actions of the person sharing the task with us even though this impairs our own performance.

In the task used by Sebanz and colleagues the performance of one participant had no direct effect on the performance of the other. Reed and colleagues used a simple task in which a crank handle was used to align a cursor with a target. In the joint action condition both participants effectively held the same handle, but had no communication except via the force the other was applying. Participants perceived the contribution of the other as a hindrance, but, in practice, when they worked together, their performance became more efficient than when they worked alone. The force profiles of the participants' actions dramatically changed in the joint response as they developed strategies to work together. Most often one participant contributed more to acceleration and the other to deceleration of the movement of the crank, thereby achieving more efficient and faster movement.

Although cooperative behavior between nonrelated participants are uncommon in the animal kingdom, Mendres and DeWaal have shown that pairs of capuchin monkeys can work together to pull a tray forward in order to receive a reward. The monkeys made more communicational glances to their partner in the joint action task than when they acted alone, in an attempt to motivate their partner. Joint action performance became poorer when the partners were unable to see each other, and the authors suggest this was because the capuchins required visual contact to coordinate the timings of their responses.

The more complex tasks typical of human cooperative interactions are greatly helped by direct communication. Communication is not only a way to express our intentions to others, it also enables the exchange of information relevant to the task in hand. This bilateral process requires both participants to input their own information, monitor their partner, and update their action. Communicators must constantly evaluate and amend their actions to take account of the reactions of listeners, and listeners must inform communicators of any changes in their current understanding to make sure that both have common knowledge. Common knowledge acts to set a mutual base, where all participants are aware of the information and know that all the other participants are also informed. Much of this common knowledge emerges as a by-product of the interaction.

Communication does not rely solely on the monitoring of speech, but also on many other communicative signals, such as facial expressions, posture, and body movements, and what is happening in each participant's workspace and in the surrounding environment. When any part of this information is blocked, communication becomes less effective and less efficient. Understanding is further worsened if communication becomes a noninteractive process, as with written instructions. For example, Clark and Krych studied a task in which one participant gave directions while the other participant followed these directions in order to build a structure. Task efficiency was greatly reduced when a barrier obscured the view of each other's workspace. Furthermore, noninteractive communication using prerecorded directions produced the greatest number of errors.

Cooperative behavior is usually mutually beneficial to both participants, but often it requires at least one participant to invest their trust in the other with some degree of risk. To prevent betrayal of trust and loss of investment we must wisely choose our partners and invest the minimal amount to reap maximal reward. We must continually monitor our partner for trustworthiness and at the same time give out signals indicating that we are worthy of trust (reputation management). Trust and reciprocity have been studied in economic games such as the Prisoners' Dilemma. An optimal strategy in such games is 'tit-for-tat,' in which trust is repaid by trust and betrayal is repaid by betrayal. Players in economic games have a strong sense of the 'fairness' of interactions. Given the opportunity, they will punish those who make unfair offers, even at some cost to themselves (altruistic punishment). The occurrence of altruistic punishment increases group cooperation.

Neural Activity Associated with Social Interactions

To date the majority of imaging studies of on-line social interactions have involved various kinds of trust and reciprocity games in which the brain activity of each participant is recorded while each interacts with partners outside the scanner.

Some of these experiments have manipulated participants' 'intentional stance' by making them believe that they are interacting with a human or computer partner. In these studies, which have used both competitive and cooperative games, greater activity is seen in the MPFC and adjacent ACC when participants believe they are interacting with a human rather than a computer. This activity is located in the anterior rostral region of the MPFC (using the terminology of Amodio and Frith), where most of the MPFC activity associated with off-line mentalizing tasks has also been observed. A recent study in monkeys has shown that lesions in the ACC cause a loss of interest in social stimuli.

Brain regions concerned with emotion are strongly implicated in these trust and reciprocity games. For example, the ventral striatum and other regions concerned with reward are activated after episodes of mutual cooperation. Unfair offers, on the other hand, decrease activity in reward areas while activating the anterior insula. Insula activity also occurs in response to negative emotions such as disgust and pain. Given these effects, is not surprising that the faces of game partners who persistently cooperate or defect rapidly come to elicit positive and negative emotions, respectively. These effects are largely eliminated if the other player is believed to be a computer or a person who is not a free agent, but merely follows instructions.

When persistent defectors are seen to be punished, activity also increases in brain regions associated with reward. This involvement of reward systems suggests a mechanism whereby cooperation and altruistic punishment can be maintained during social interactions.

Two Brains, Two Scanners

Investigating the brain activations of one player only gives half the story, because another person (and their brain) is involved in a dyadic social interaction. To address this problem a new type of neuroimaging has been developed by Read Montague and his colleagues. This system, known as 'hyperscanning,' allows the simultaneous brain scanning of two socially interacting people by using an Internet link to temporally align two functional magnetic resonance imaging (fMRI) scanners and their presentation systems. Timing and presentation information, such as scanner onset or type and onset of stimuli and rewards, can be transmitted between different scanning laboratories. The Internet connection allows participants to interact with other participants in fMRI scanners located around the world and opens the opportunity to investigate the correlation between the activities of socially interacting participants.

The two hyperscanning studies published so far have both involved trust and reciprocity games. Participants played against the same human partner and were free to make their own choices to allow models of trust and reputation to be built. One participant, the investor, was given money and then decided how much money to share with the other participant, the trustee. Any shared money was tripled and then the trustee decided how much money to repay the investor.

As in previous studies, brain areas associated with reward (dorsal striatum) were activated when trust was reciprocated. However, in the course of the game this signal increasingly anticipated the time at which the investor knew that her trust had been reciprocated. The authors suggest this anticipatory signal is analogous to reward prediction in simple conditioning experiments. As trust builds during the course of the game, the investor becomes increasingly confident that her benevolent behavior will be rewarded by benevolent reciprocity.

However, the importance of the hyperscanning paradigm lies in the possibility of direct investigation of interactions between brains. For example, a cross-brain correlation analysis indicates whether activity in a region in one brain can predict activity in a region of the other brain. In the trust and reciprocity game a strong correlation was found between the investor's middle cingulate cortex and the trustee's ACC. These different locations in cingulate cortex could be related to representations of the 'self' (middle cingulate cortex active during submission of the participants' own responses) or the 'other' (ACC active during the revelation of their partners' responses). These representations were independent of the role of each participant or the type of reciprocity produced. The authors concluded that the cingulate cortex has a spatial map of response agency along the anteroposterior axis. Given this account of the representations in cingulate cortex, the cross-brain analysis indicates that the linked brain regions represent the same thing. Thus, the investor's representation of himself/herself (participant A) is linked to the trustee's representation of the other (participant A). This could be evidence, at the neural level, of the alignment between representations that is necessary for successful social interactions.

Conclusions

Investigations of the neural basis of interacting minds have only just begun. The possibilities are very exciting. As we have seen, there are many behavioral paradigms, in addition to economic games, that have yet to be adapted for brain imaging experiments. Hyperscanning is a particularly important development since interacting minds depend upon interacting brains. When we analyze activity in interacting brains, we need to treat the two (or more) brains as a single system, which has emergent properties that cannot be seen in the single brain. fMRI is perhaps not the ideal modality for such analysis, since we can only look at rather slow hemodynamic changes. Electroencephalography and magnetoencephalography, with their high temporal resolution and direct reflection of neural activity, will provide powerful

ways for looking at links between brains during social interactions. Analytic techniques already developed for looking at functional connectivity within brains, such as structural equation modeling and casual modeling, can be applied without modification to analysis of connectivity between brains.

See also: Cognition: An Overview of Neuroimaging Techniques; Decision-Making and Neuroeconomics; Emotion: Neuroimaging; Neuroeconomics: History; Neuroimaging; Reward Decision-Making; Reward Processing: Human Imaging; Social Brain: Evolution; Social Emotion: Neuroimaging; Social Interaction; Social Interaction Effects on Reward and Cognitive Abilities in Monkeys.

Further Reading

Amodio DM and Frith CD (2006) Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience* 7: 268–277.

Axelrod R (1984) *The Evolution of Cooperation*. New York: Basic Books.

Barr DJ (2004) Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science* 28: 937–962.

Clark HH and Krych MA (2004) Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50: 62–81.

Fehr E and Gachter S (2000) Altruistic punishment in humans. *Nature* 415: 137–140.

Frith CD and Frith U (2006) How we predict what other people are going to do. *Brain Research* 1079: 36–46.

King-Casas B, Tomlin D, Anen C, et al. (2005) Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308: 78–83.

Mendres KA and DeWaal FBM (2006) Capuchins do cooperate: The advantage of an intuitive task. *Animal Behaviour* 60: 523–529.

Reed K, Peshkin M, Hartmann MJ, et al. (2006) Haptically linked dyads – are two motor-control systems better than one? *Psychological Science* 17: 365–366.

Rudebeck PH, Buckley MJ, Walton ME, et al. (2006) A role for the macaque anterior cingulate gyrus in social valuation. *Science* 313: 1310–1312.

Rilling JK, Sanfey AG, Aronson JA, et al. (2004) Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *NeuroReport* 15: 2539–2543.

Sanfey AG, Rilling JK, Aronson JA, et al. (2003) The neural basis of economic decision-making in the Ultimatum Game. *Science* 300: 1755–1758.

Saxe R (2006) Uniquely human and social cognition. *Current Opinion in Neurobiology* 16: 235–239.

Sebanz N, Bekkering H, and Knoblich G (2006) Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences* 10: 70–76.

Tomlin D, Kayali MA, King-Casas B, et al. (2006) Agent-specific responses in the cingulate cortex during economic exchanges. *Science* 312: 1047–1050.

Warneken F, Chen F, and Tomasello M (2006) Cooperative activities in young children and chimpanzees. *Child Development* 77: 640–663.

Wu JZ and Axelrod R (1995) How to cope with noise in the Iterated Prisoners' Dilemma. *Journal of Conflict Resolution* 39: 183–189.

Social Communication in Whales and Dolphins

P L Tyack, Woods Hole Oceanographic Institution, Woods Hole, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Whales and dolphins are part of the mammalian order Cetacea. The primary division in the cetaceans is between the baleen whales (mysticetes) and the toothed whales (odontocetes), which includes dolphins and larger toothed whales, such as killer and sperm whales. As the ancestors of whales and dolphins entered the sea about 50 Ma, they encountered a sensory environment very different from life in air. Olfaction is a sense primarily for airborne chemicals, and the olfactory system is reduced in baleen whales and absent in toothed whales. By contrast, the sense of touch is well developed in marine mammals, which rely on tactile communication for close-range interactions in ways that are similar to terrestrial mammals. Vision is the optimal distance sense in air, but objects underwater can seldom be seen at ranges of more than tens of meters. Some low-frequency sounds, by contrast, can be heard at ranges of thousands of kilometers in the ocean. These differences between air and water as media for communication created new selection pressures for sensory systems in marine mammals. Toothed whales evolved high-frequency hearing and sound production mechanisms for communication at low frequencies specialized for echolocation. Baleen whales evolved hearing and sound production mechanisms for communication at low frequencies that propagate over hundreds of kilometers. Because of the critical importance of sound as a communication medium for marine mammals, this article will focus on acoustic communication.

All marine mammals produce sound by moving air past phonic lips. For terrestrial animals, this kind of pneumatic mechanism produces stable voice cues that can be used for individual identification. As a marine animal dives, however, air-filled cavities in the vocal tract change shape, leading to changes in voice cues. This suggests that the motor commands required to generate a vocalization with a specific acoustic structure may vary with the animal's depth. One solution to this problem involves learning mechanisms that can match vocal output to an auditory template. This problem with maintaining stable vocalizations may help explain how common vocal learning is among marine mammals.

Vocal Learning

There is only weak evidence for vocal learning among nonhuman primates, and it is rare among other terrestrial mammals, but three major marine mammal taxa, seals, toothed whales, and baleen whales, all show evidence of sophisticated vocal learning skills. Hoover, a harbor seal (*Phoca vitulina*) at the New England Aquarium, imitated human speech with a clear Maine accent. His mother had died soon after giving birth, and he was raised by a family in Maine for the first few months of life. Captive bottlenose dolphins (*Tursiops truncatus*) have repeatedly been shown to be capable of precise imitation of frequency-modulated tonal signals. The songs of humpback whales (*Megaptera novaeangliae*) are similar within a population at any one time, but they change continuously so that each sound is replaced within a few years. This 'hit parade' can be explained only by vocal learning in which whales copy the most popular songs at any one time.

The songs of humpback whales are made up of a series of 'themes' that are produced in such a well-defined order that each theme is given a number such that theme 3 always follows theme 2 and precedes theme 4. Each theme is made up of a variable number of repetitions of a 'phrase,' which is a sequence of about six to eight individual sound units. Humpback whales do not emit bubbles as they vocalize, and it is thought that at the end of each phrase, the whale recycles air from above the sound production organ back to the lungs. The sequences of sounds produced by a humpback singer cannot be replicated by Markov processes but rather are generated by a hierarchical mechanism with strong syntactic constraints. None of the changes in the song appear to correlate with changes in the meaning, so this sophisticated syntactic structure appears to be divorced from the semantic elements that are so well correlated with syntax in human language.

There is limited evidence for convergence of vocalizations in groups of terrestrial nonhuman mammals, but nothing like the evidence of imitative skills demonstrated for marine mammals. Most terrestrial mammals have a fixed vocal repertoire of communication signals, and the structure of these signals is inherited, developing normally even in the absence of exposure to auditory models. By contrast, many species of birds, humans, and a few mammals maintain an ability to learn new vocalizations throughout the lifetime. Several costs are associated with vocal learning. It may take longer for animals that require

auditory input to develop appropriate signals, and development can be derailed if an animal is not exposed to the appropriate auditory environment.

Animals that can learn their vocal signals have an open communication system that requires modes of understanding that are different from those of closed communication systems. Animals that produce vocalizations with no auditory feedback can be said to have a fixed vocal repertoire, each sound of which may evolve to meet specific communicative functions. Dolphin and avian vocal repertoires can be open and can differ significantly between individuals. In some species, different elements of the learned repertoire appear to be equivalent, suggesting that it is inappropriate to assume that each signal has a different function and raising questions about the significance of precise imitation versus copying errors.

Two communicative functions have been associated with vocal learning in marine mammals. These involve the development of songs for reproductive advertisement and of signature signals for individual or group recognition. Some species of seal and baleen whale use vocal learning to produce their reproductive advertisement displays. The best evidence among marine mammals for the uses of vocal learning in reproductive advertisement stems from humpback song because of the continuous change in so many features of this complex signal. Geographical dialects have been reported for songs of other whale species and for seals. Dialects are often cited as evidence for vocal learning, but vocal dialects in elephant seals from different islands are thought to result from founder effects. Regional dialects in animals such as blue whales are more likely a result of vocal learning, as dialect areas are not as isolated as breeding islands for seals. The songs of blue and fin whales are less complex and more repetitive than those of humpback whales, apparently as an adaptation for long-range communication. Humpback whales congregate to breed on traditional breeding grounds, while blue and fin whales are thought to disperse into tropical seas during the winter breeding season. The songs of blue and fin whales are simple low-frequency sounds that are regularly repeated. They can be detected at ranges from hundreds to more than 1000 km. Since these whales routinely swim more than 100 km day^{-1}, it is possible that the functional range of advertisement extends to these ranges. One of the adaptations required for such long-range communication involves specializations for production and reception of low-frequency signals, including an unusually wide, thin, and floppy apical end of the basilar membrane in the inner ear.

Many toothed whales use acoustic signature signals to maintain social cohesion between individuals that share strong social bonds. Bottlenose dolphins have societies characterized by strong bonds between individuals within fluid social groups. Even though calves are precocious in terms of ability to swim away from the mother, there is a prolonged period of dependency in which calves suckle and are sighted near the mother for years. As male dolphins mature, they tend to form alliances with one or two other males, and allied males tend to be sighted together 90% of the time. While groups of dolphins often change membership on timescales of minutes, social bonds between individuals often last a decade or longer. Even dolphins with strong social bonds separate out of sight of one another many times each day. After bonded dolphins separate, they use frequency-modulated whistle vocalizations to mediate reunions.

Infant dolphins produce simple unstereotyped whistles as soon as they are born. They develop an individually distinctive signature whistle during the first 1–2 years of life. As in some songbirds, the young tend to develop a whistle that differs from that of close associates but that resembles the whistle of a member of the population sharing low levels of association with the calf. Dolphins born in captivity tend to have whistles with less frequency modulation than those of wild dolphins, apparently because they incorporate acoustic features of the nonmodulated whistles used by trainers. Females retain a stable signature whistle for the rest of their life, but as males form an alliance, their whistles converge with those of their partners, becoming more similar while retaining distinctive elements.

A striking element of the communication system of dolphins involves one animal producing precise imitations of the signature whistle of another. Research with captive dolphins demonstrates that they are capable of learning to imitate artificial sounds and to produce these sounds when they are exposed to a manmade object. It is not known whether imitation of signature whistles represents naming other dolphins in the same way they can vocally label objects. The issue of defining how referential dolphin communication is remains a hot topic.

In contrast to the fluidity of dolphin social groups, sperm (*Physeter macrocephalus*) and killer whales (*Orcinus orca*) live in stable groups and produce group-specific repertoires of calls. Social structure in both sperm and killer whales is based on stable matrilineal groups. The best-studied killer whale population lives in inshore waters of the northeastern Pacific Ocean and feeds primarily on fish. One of the remarkable features of this resident population is that neither sons nor daughters have been seen to emigrate from their natal group. The basic social unit is a matriline consisting of one to four generations of

matrilineally related animals. The most commonly sighted group is a 'pod' composed of one or more related matrilines.

Killer whales produce a variety of sounds: echolocation clicks, frequency-modulated tonal whistles, and stereotyped pulsed calls called discrete calls. Each pod produces an average of ten types of discrete call, and the more closely related pods are, the more overlap there is in their discrete call repertoires. Pods do not use different discrete calls in different behavioral contexts, but rather the group-specific repertoire appears to function as an acoustic 'badge' to maintain pod cohesion. Animals within a pod often disperse beyond visual range, to forage separated by hundreds of meters, and different pods may form aggregations with animals from other pods. Discrete calls are most commonly produced when whales are separated while foraging or when different pods meet, settings when these animals need to reassociate to form their pods again.

There may be slight but predictable differences in calls shared by different matrilines within a pod. The acoustic features of some calls change over time, with different calls having different rates of change. This pattern of vocal change is thought to represent a form of cultural drift, leading over time to the observed pattern of repertoire differences between pods. However, while there are limited observations of vocal matching in captivity and the wild, there is no definitive evidence for vocal production learning in killer whales.

Sperm whales form matrilineal groups like those of killer whales, but in contrast to killer whales, maturing male sperm whales leave their natal groups. Female and immature sperm whales are usually found in groups of ten to 40 whales. These groups contain clusters of individuals that are closely related genetically, along with unrelated individuals. Individuals that associate over long periods of time (years) are called a unit, and most groups sighted at sea represent temporary combinations of several units that may stay together for a few days. One of the primary benefits of group living for sperm whales is thought to be social care of the young. The lability of size and composition of groups may allow units to balance optimal size of groups for local foraging conditions versus social defense against predators.

When sperm whales forage at depth, they produce regular clicks for echolocation. At other times, they produce rhythmic patterns of clicks, called 'codas,' that are thought to be used for communication. Each sperm whale unit produces a repertoire of codas. Analysis of coda repertoires from sperm whales throughout the South Pacific Ocean suggests that there are five sympatric 'clans' of sperm whales units that tend to share coda repertoires. These are similar to the vocal clans seen in resident killer whales but span on the order of 10 000 km, compared with 1000 km for killer whales, and involve many more individuals, perhaps 10 000 sperm whales, compared with about 100 killer whales in the entire resident population. Even though clans are sympatric, units appear preferentially to associate with other members of the same vocal clan. It is thought that these coda dialects allow units to identify units from the same clan, perhaps mediating communal defense and allomaternal care. While these coda dialects have been described as learned aspects of animal culture, there is no direct evidence that they are learned.

The directionality of signals of dolphins and toothed whales has led to a new appreciation of how directional signals can cue a receiver as to the orientation of a calling animal. The higher the frequency of a signal compared with the aperture of the sound production organ, the more directional the signal can be. This leads to a pattern in communication signals that parallels directional echolocation clicks. When the signaling animal is pointing directly toward a receiver, the ratio of high- to low-frequency energy is higher than when the signaler is pointing away. This means that when a dolphin whistles or a killer whale produces a discrete call, other animals nearby can detect whether the signaler is pointing toward or away from them. This information is particularly useful for cohesion signals used to coordinate movements between animals sharing a bond.

While most research on animal models of vocal learning in the past have focused on reproductive advertisement, there is growing evidence for the importance of vocal learning in the development of signature signals, including vocal convergence in the formation of group-distinctive calls. Sexual selection leads to the evolution of extreme displays, and the songs of birds and whales stand out as some of the most dramatic examples of learned displays. There are some similarities between language learning in humans and song learning in birds – normal vocal development in both songbirds and humans requires hearing species-specific vocalizations during a sensitive period early in life. But there are also significant differences. Vocal learning in many songbirds is largely limited to males and mainly to one type of vocalization – song.

Neurobiologists interested in comparative models of vocal learning with a prime focus on humans should consider vocal convergence among animals within a group in addition to sexual selection for vocal learning of advertisements. While there are no nonhuman primate models of vocal learning similar to songbirds, vocal convergence has been

shown in several nonhuman primates. Similar vocal convergence has been demonstrated for a wide variety of other mammals, including bats, toothed whales, baleen whales, seals, and elephants. Neurobiologists with broader comparative interests in the evolution of vocal learning should strive to embrace a broad taxonomic scope including groups whose phylogenetic relationships are distant enough to guarantee independent evolution of the traits of interest. Comparison of the ecological and social settings of a diverse array of animals will help provide a fuller understanding of the selection pressures for this important skill. For example, dolphins with stable bonds within fission–fusion societies use vocal learning to develop individually distinctive signatures, whereas killer whales with stable groups learn to develop group-distinctive repertoires of discrete calls.

There has been a growing appreciation of mechanisms that animals can employ to maintain effective communication in a noisy channel. One simple strategy is to increase the level of a signal in increased noise to improve the ratio of signal to noise. This can be achieved with relatively simple effects of auditory feedback. However, in situations where different conspecific signalers are competing for the attention of a receiver, such as when multiple males advertise to a female, more-complex strategies may come into play. Similarly, the diversity and increasing pace of introduction of novel sources of anthropogenic sounds may select for vocal plasticity that can help a signaler adapt to a changing acoustic environment. Birds and whales exposed to anthropogenic noise have been reported to shift the level, frequency, or repetitiveness of their calls to reduce interference. As humans alter the acoustic environment of animals that use sound for critical functions, vocal plasticity may become more important than during the preindustrial era, when the auditory environment was likely more stable. To the extent that successful use of vocal communication is required for growth, survival, and reproduction, it will be important for humans to understand the impact of disrupting the use of sound by animals.

See also: Auditory/Somatosensory Interactions; Cetacean Brains; Echolocation I: Behavior; Echolocation II: Neurophysiology; Hearing and Echolocation in Dolphins.

Further Reading

Janik VM and Slater PJB (1997) Vocal learning in mammals. *Advances in the Study of Behavior* 26: 59–99

Mann J, Connor R, Tyack PL, and Whitehead H (2000) *Cetacean Societies: Field Studies of Whales and Dolphins*. Chicago, IL: University of Chicago Press.

Tyack P (1998) Acoustic communication under the sea. In: Hopp AL, Owren MJ, and Evans CS (eds.) *Animal Acoustic Communication: Recent Technical Advances*, pp. 163–220. Heidelberg, Germany: Springer.

Tyack PL (1999) Communication and cognition. In: Reynolds JE III and Twiss JR Jr. (eds.) *Biology of Marine Mammals,* vol. 1, pp. 287–323. Washington, DC: Smithsonian Press.

Tyack PL and Clark CW (2000) Communication and acoustic behavior of dolphins and whales. In: Au W, Popper AS, and Fay R (eds.) *Hearing by Whales and Dolphins*, pp. 156–224. New York: Springer.

Social Emotion: Neuroimaging

T F Heatherton and A C Krendl, Dartmouth College, Hanover, NH, USA

© 2009 Elsevier Ltd. All rights reserved.

Humans are social beings who live in groups. Dependency on group living is not unique to humans, but the nature of relations among and between in-group and out-group members is especially complex in human societies. Decades of research in psychology has established the central role of emotional processes in facilitating social relationships and guiding group behavior. 'Social emotions,' which are complex subjective experiences (e.g., pride, admiration, jealousy, envy, irritation, and flirtatiousness), serve many important social functions that promote long-term relationships and group stability. From a functional perspective, social emotions enable people to be good group members, thereby avoiding rejection and enhancing their survival and reproduction. Thus, social emotions are essential to human social life.

Social emotions facilitate successful social relationships through two primary pathways: they provide incentives to engage in social interactions (e.g., affection, love, feelings of pride, or admiration for those with whom we interact), and they increase the likelihood that people will adhere to societal norms that are necessary for group living. When such norms are violated, people experience negative social emotions (e.g., feelings of guilt, embarrassment, or shame) that subsequently encourage them to act within the bounds of socially acceptable conduct, thereby reducing the risk of social exclusion and promoting positive social interactions. Moreover, long-lasting social emotions (such as remembering an embarrassing moment from adolescence) reduce the likelihood of repeat violations.

Although it is possible that nonhuman animals experience emotional states that modulate behavior, it is apparent to most researchers that complex social emotions are particularly characteristic of humans and some are possibly uniquely human. An open question, however, is whether social emotions are innate or learned behaviors. From an evolutionary perspective, social emotions serve the adaptive function of fostering social relationships within groups. Thus, it is possible that people have evolved dedicated neural circuits for social emotions, much as has been argued for the more basic emotions, such as fear. At the same time, some evidence suggests that many social emotions are influenced more by socialization than by genes, indicating that important aspects of social emotions develop within a social context. Indeed, research suggests that the influence of the shared environment on feelings of guilt increases over time, whereas the influence of genes disappears. Nonetheless, for most people, social emotions play a powerful and prominent role in their interactions with other people. Extensive evidence from patient research, as well as recent efforts using the methods of neuroscience to study human social emotions (i.e., neuroimaging), have started to provide important insights into the factors that produce and maintain social emotions.

Neuropsychological Evidence for Discrete Neural Networks in Social Emotions

Extensive evidence from patient research suggests that there are two primary neural regions involved in perceiving social emotions. The first (the amygdala) contributes to the identification of the social emotion, whereas the second (discrete regions of the prefrontal cortex) translates those emotions into information that can be used to facilitate social interactions. The process of garnering socially relevant information from social emotions involves three critical components. First, social emotions require people to be aware of their behavior, so as to gauge it against societal or group norms. Second, people need to understand how others are reacting to their behavior so as to predict how others will respond to them. To do so requires them to be able to make attributions about the mental states of others, a capacity often called 'theory of mind' (ToM). Finally, there need to be mechanisms for detecting discrepancies between self-knowledge and social expectations of norms, thereby motivating behavior to resolve any conflict that exists.

The first evidence that regions of the prefrontal cortex plays a critical role in social emotions came from the case of Phineas Gage, who suffered profound frontal lobe damage when a railroad tamping rod misfired into his head. Formerly described by friends as dependable, polite, and hardworking, Gage became capricious, volatile, and lewd after the accident, showing no evidence of embarrassment or remorse at his socially inappropriate behavior. Gage's radical personality change after his injury provided convincing evidence that the prefrontal cortex plays a crucial role in the social emotions that regulate behavior (e.g., guilt, embarrassment, and remorse). Since Gage's accident, abundant evidence from patient research, and more recently neuroimaging studies, has further identified discrete neural

networks involved in both identifying and using social emotions.

Neuroimaging and Patient Research on Self-Awareness

Social emotions require that people be aware of the effects of their behavior on others, which necessitates self-awareness. That is, when people commit a *faux pas*, they must be able to recognize that the behavior they performed is inappropriate and that the feelings they are experiencing are embarrassment at having committed the *faux pas*. Self-awareness occurs when people themselves are the objects of their attention, such as when they think about themselves or process other information in a personal manner. Many years of research in psychology has revealed that information processed with reference to self seems to be treated 'specially.' For instance, people show a robust enhancement in memory for material that has personal meaning. This self-reference memory enhancement effect has been observed in many contexts, such as when people remember information processed with reference to self better than information processed with reference to other people (i.e., friends and politicians). The use of functional neuroimaging has provided key insights into the brain regions that support self-awareness.

A series of functional magnetic resonance imaging (fMRI) studies conducted over the past 10 years have indicated that the medial region of the prefrontal cortex (MPFC) plays a vital role in self-awareness. This region is more active, for example, when people report on their personality traits, make self-relevant judgments about pictures, or retrieve autobiographical memories of past events. Recent research indicates that regions of the ventral anterior cingulate cortex (ACC) may work in conjunction with MPFC to reflect the valence of self-relevant information. Interestingly, the MPFC has also been identified as part of a 'default network,' which also includes the posterior cingulate gyrus and precuneus. This network is active when the brain is at rest (i.e., not engaged in an overt cognitive task). An abundance of positron emission tomography (PET) and fMRI data suggest that the default network plays a dominant role in self-awareness. As such, the default network supports important components of social emotions.

It is interesting to note that converging evidence from patient research indicates that frontal lobe lesions, particularly to the MPFC and adjacent structures, have a deleterious effect on personality, mood, motivation, and self-awareness. Patients with frontal lobe lesions show dramatic deficits in recognizing their own limbs, engaging in self-reflection and introspection, identifying a *faux pas* as being socially inappropriate, and even reflecting on personal knowledge. Indeed, frontal lobe patients are particularly impaired in social emotions, such as feeling embarrassed in appropriate contexts. The convergence of patient and imaging data support the conclusion that MPFC plays a prominent role in self-awareness, a necessary and critical contributor to the experience of social emotions.

ToM and the MPFC

In addition to recognizing our own mental states, social emotions require that we are able to accurately interpret the emotional states of others. For instance, to feel guilty about hurting a loved one, people need to understand that other people have feelings. Indeed, all social emotions related to empathy require the capacity to attribute specific mental states to others. Similarly, at the core of many social emotions is the belief that one is being evaluated by others (thereby giving rise to emotions such as embarrassment and social anxiety), which means recognizing that other people make evaluative judgments. The ability to infer the mental states of others is commonly referred to as mentalizing, or having the capacity for ToM. ToM enables the ability to empathize and cooperate with others, accurately interpret other people's behavior, and even deceive others when necessary. The rapidly emerging neuroimaging literature on ToM has consistently implicated MPFC as a central component of the neural systems that support mentalizing. Patient research in this area primarily focuses on people who have syndromes that involve impaired ToM (i.e., Asperger's or autism). This line of research indicates that such people have decreased neural activity and cerebral blood flow to the MPFC during ToM tasks. Such findings suggest that the MPFC plays a vital role in mentalizing tasks.

It is interesting to note that the ability to mentalize relies heavily on similar neural networks engaged in processing self-relevant information, notably MPFC. However, this region of MPFC tends to be more dorsal in ToM studies than in self-reference studies, although overlap in activity has been observed. In a typical neuroimaging study of ToM, participants are required to make judgments about other people's mental states after reading scripted scenarios or observing pictures of people with differing facial expressions. For instance, early investigations compared neural activity observed during trials in which participants were asked to judge the motivations of an actor's behavior (e.g., "he turned himself into the police because they knew he had committed a crime") with that observed during nonmentalizing tasks, such

as judging a person's physical actions, and revealed MPFC activation in the mentalizing task only. MPFC activation can also be observed when participants are asked to determine the underlying motivation of an actor as depicted by pictures that convey specific social behavior (i.e., stealing, sharing). This region is also engaged when perceivers are asked to infer the mental states of targets that are most similar to them. This latter finding suggests that self-reflection is possibly engaged in mentalizing tasks, illustrating perhaps a common role that MPFC is playing for both self-awareness and ToM. Indeed, deficits in this brain region are associated with a reduction in ToM, reinforcing the importance of MPFC in ToM tasks. Although activity in other brain regions has been observed during ToM tasks, notably the superior temporal sulcus (STS), the temporoparietal junction, and less often the amygdala, MPFC appears to play a central role in the ability to make mental state attributions about other people. This ability is critical for social emotions because it enables us to modify our behavior in order to promote smooth social interactions.

Social Emotions and the Amygdala

In addition to MPFC, a distributed network of brain regions has been implicated in selecting and processing social emotions. One such region is the amygdala. For decades, the amygdala was considered most often in the context of fear learning and attention tasks. More recently, however, amygdala activation has been observed in making social judgments, such as evaluating stigmatized individuals. Not surprisingly, then, evidence has emerged from patient and neuroimaging research to implicate the amygdala in perceiving social emotions. For instance, patients with unilateral or bilateral amygdala damage have been shown to be impaired at recognizing specific social emotions from facial expressions (e.g., arrogance, guilt, admiration, and flirtatiousness). Interestingly, this deficit appears to be limited to social emotions, and does not extend to other emotions, suggesting that the amygdala plays a central role in recognizing social emotions.

Neuroimaging studies have also demonstrated an important role for the amygdala in processing a variety of social emotions. For instance, when participants are asked to choose the appropriate reaction (from varying perspectives) to sentences that represented different social emotions (e.g., embarrassment, pride, shame, guilt, admiration, irritated, and impressed), as compared to sentences conveying nonsocial emotions, heightened amygdala activation is observed, regardless of perspective taken during the task. Interestingly, when participants are asked to examine intentional violations of social norms through stories (e.g., "You are invited for dinner at your friend's house") that have one of three endings: one that was descriptive of normal social behavior (e.g., "You have a bite of the first course and like it, congratulate your friend for her good cooking"); one that described an embarrassing conclusion (e.g., "You have a bite of the first course, you choke and spit out the food while you are coughing"); or one in which the protagonist violated a social norm (e.g., "You have a bite of the first course, but do not like it and spit the food back onto your plate"), a similar pattern of activation is observed. When taking their own perspective, participants showed greater amygdala activation in response to intentional violations of social norms. These results suggest that the amygdala is indeed involved in processing social emotions.

Based on the extant literature on the amygdala, it is not surprising that this area may be involved in processing social emotions. It has long been argued that the primary role of the amygdala is to automatically respond to arousing emotional stimuli, particularly fear-inducing or aversive items. From this perspective, the role of the amygdala in processing social emotions may simply be to respond to the highly arousing nature of the unique conditions that produce social emotions. More recently, however, this region also has been implicated in drawing attention to novel stimuli that have biological relevance. Under this argument, it is plausible that the amygdala plays a role in processing social emotions because they have direct relevance in maintaining long-term social relations, which is a fundamental biological need. Regardless, the role of the amygdala in processing social emotions suggests that an automatic response network exists in the brain for distilling social emotions.

Social Emotions and the ACC

In addition to the MPFC and amygdala, patient research and neuroimaging have also implicated the ACC in processing social emotions. Although the ACC is anatomically part of the prefrontal cortex, a wealth of evidence from patient and neuroimaging research illustrates that ACC activity is functionally distinct from MPFC regions involved in self-awareness and ToM. Evidence from patient studies suggests that the ACC is involved in integrating cognitive and affective information to guide behavior appropriately. For instance, imaging research has implicated the ACC in resolving cognitive conflicts. With respect to social emotions, the anterior cingulate appears to be engaged in social interactions that provide conflicting social feedback, thereby eliciting conflicting emotions.

Although emerging evidence from neuroimaging research has yielded disparate results as to the roles of discrete regions of the ACC in processing social feedback, it has provided a working framework for understanding the ACC. Rejection, for instance, is a social emotion that results from a violation of the expectancy that people believe they are valued. Some neuroimaging research has found that a region of the dorsal ACC (dACC) is responsive during tasks designed to elicit social rejection, such as when virtual interaction partners suddenly and surprisingly stop cooperating with research participants. It is possible, however, that the dACC activity observed in these types of tasks indicates cognitive conflict more than social pain resulting from rejection, *per se*. Indeed, being rejected may violate the fundamental expectancy of social inclusion. Other research has indicated that a more ventral region of the ACC (vACC) is important for the social emotion of feeling rejected. Interestingly, interpersonal distress, such as feeling unloved, is often involved in depression, and abnormalities in vACC activity have been observed in unipolar depression. For instance, imaging research has found differential patterns of vACC activity to emotional facial expressions between depressed and non-depressed subjects. Thus, one possible role for vACC reflects its involvement in general affective state, thereby affecting social emotions indirectly.

Another possible role for ACC in social emotions is its involvement in detecting and attempting to resolve cognitive conflict. Thus, information that conflicts with pre-established knowledge signals errors that must be resolved, a situation that produces ACC activity. From this perspective, then, the ACC may be involved in processing social emotions by detecting conflict between actions and outcomes. For instance, when a social norm is violated, the ACC may play a role in identifying this error because it detects differences between the expectation (social norm) and action (violation). The detection of errors in social behavior may produce social emotions that guide future appropriate behaviors.

Self-Conscious Emotions

An important distinction that emerges in the psychological literature is the subcategory of 'self-conscious emotions' from the more general social emotions. Self-conscious emotions, including embarrassment, guilt, and shame, occur when the emotional experience reflects an evaluation of the self. People feel embarrassed when they believe that others are evaluating them negatively, and ashamed when they do it to themselves. Either way, the self is implicated in a way that is more specific than emotions such as jealousy, admiration, or flirtatiousness. The primary adaptive value of self-conscious emotions is that they provide a clear disincentive for people to act in a socially inappropriate manner. Although, neuroimaging research on self-conscious emotions is still in its infancy, patient studies and a handful of imaging studies shed light on the neural networks involved in this specific category of social emotions.

Patient research has provided compelling evidence that damage to the orbitofrontal cortex, a region adjacent to the medial prefrontal cortex, severely impairs self-conscious emotions. Patients with lesions to this area exhibit extremely socially inappropriate behavior (e.g., making inappropriate jokes, kissing or hugging strangers, disclosing intimate information about themselves to strangers). Imaging studies of self-conscious emotions such as guilt and embarrassment implicate additional brain regions, including anterior cingulate gyrus, MPFC, STS, and insula. For example, asking people to recall the most guilt they ever experienced produces greater activity in these regions than does recalling events when they did not feel guilty. In research comparing guilt and embarrassment, both self-conscious emotions elicited MPFC activation, but guilt apparently does so more than embarrassment. Robust MPFC activity in self-conscious emotions may reflect the unique circumstances under which people mentalize about how others are evaluating them.

Imaging Moral Emotions

Another category of social emotions, namely, moral emotions, dictate what is ethically 'right' or 'wrong' in a given situation. As might be expected by the foregoing discussion, when participants are asked to resolve various moral dilemmas during imaging, MPFC activation is observed. Specifically, comparisons between making moral-personal (e.g., the footbridge dilemma), moral-impersonal (e.g., the trolley dilemma), and neutral (or nonmoral) decisions revealed that MPFC and posterior cingulate gyrus are significantly more active in the moral-personal condition than in the moral-impersonal and the nonmoral conditions. MPFC activation also emerges when morality is measured by eliciting disgust or indignation from the observer. For instance, participants who read a series of disgust-inducing (e.g., "A friend told you about a newspaper headline. A man had died after he ate a living rat") or indignation-causing (e.g., "As you arrived home, you saw that the nurse has put a spider on the baby's face") statements showed heightened MPFC and

orbitofrontal activation to both the disgust- and indignation-inducing statements. Disgust also elicited greater amygdala and anterior cingulate response as compared to indignation, whereas the latter elicited greater insula and right inferior frontal gyrus.

Substantial evidence has also been found in the patient literature that lends support to these findings. Patients who have experienced damage to regions of the prefrontal cortex, including MPFC and orbitofrontal cortex, early in life (i.e., before 2 years) show extensive deficits in their ability to engage in moral reasoning tasks. Together, the findings from emerging neuroscience and patient research suggest that the prefrontal cortex, particularly the MPFC, has a critical role in moral reasoning.

Conclusions

Although neuroimaging research on social emotions is in its early days, it has already revealed a strong overlap with the neural networks that give rise to social cognition and processing emotions. Imaging studies have indicated a prominent role of the MPFC for social emotions (especially self-conscious emotions) that require inferences to be made about the mental states for self or others. The extant literature also implicates the amygdala in social emotions, although it remains an open question as to whether the amygdala is selectively responsive to specific social emotions or whether it plays a more general role in emotional responding (i.e., arousal or vigilance). Moreover, initial imaging studies also indicate an important role for the anterior cingulate in detecting conflicting social information, such as detecting social rejection, and perhaps in getting behavior back on track. Indeed, burgeoning research on the neural correlates of social emotions include studies on the neural networks underlying attachment, love, and perceiving stereotypes. Considered as a whole, patient and neuroimaging research are producing important insights into social emotions. In order to successfully guide social interactions, social emotions require self-awareness, the ability to understand the mental states of others, and mechanisms to detect conflicts between personal behavior and appropriate social rules. It is likely that neuroimaging research will continue to provide insights into this important aspect of emotion.

See also: Cognition: An Overview of Neuroimaging Techniques; Emotion Systems and the Brain; Emotion: Neuroimaging; Emotion: Computational Modeling; Neuroimaging.

Further Reading

Adolphs R (2003) Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience* 4(3): 165–178.

Adolphs R, Baron-Cohen S, and Tranel D (2002) Impaired recognition of social emotions following amygdala damage. *Journal of Cognitive Neuroscience* 14(8): 1264–1274.

Amodio DM and Frith CD (2006) Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience* 7(4): 268–277.

Botvinick MM, Cohen JD, and Carter CS (2004) Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences* 8: 539–546.

Feinberg TE and Keenan JP (2005) *The Lost Self: Pathologies of the Brain and Identity.* New York: Oxford University Press.

Frith CD and Frith U (2006) How we predict what other people are going to do. *Brain Research* 1079: 36–46.

Gallagher HL and Frith CD (2003) Functional imaging of 'theory of mind' *Trends in Cognitive Sciences* 7(2): 77–83.

Gusnard DA (2005) Being a self: Considerations from functional imaging. *Consciousness and Cognition* 14: 679–697.

Heatherton TF, Macrae CN, and Kelley WM (2004) A social brain sciences approach to studying the self. *Current Directions in Psychological Science* 13: 190–193.

Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, and Carter CS (2004) Anterior cingulate conflict monitoring and adjustments in control. *Science* 303(5660): 1023–1026.

Moran JM, Macrae CN, Heatherton TF, Wyland CL, and Kelley WM (2006) Neuroanatomical evidence fordistinct cognitive and affective components of self. *Journal of Cognitive Neuroscience* 18: 1586–1594.

Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, and Shulman GL (2001) A default mode of brain function. *Proceedings of the National Academy of Sciences of the United State of America* 98(2): 676–682.

Singer T, Seymour B, O'Doherty JP, Stephan KE, Dolan RJ, and Frith CD (2006) Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439: 466–469.

Stuss DT and Levine B (2002) Adult clinical neuropsychology: Lessons from studies of the frontal lobes. *Annual Review of Psychology* 53(1): 401–433.

Whalen PJ (1998) Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science* 7(6): 177–188.

Social Interaction

J K Rilling, Emory University, Atlanta, GA, USA
A G Sanfey, University of Arizona, Tucson, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Decision Making and Social Interactions

The study of decision making has as its goal the understanding of our fundamental ability to process multiple alternatives and choose an optimal course of action. A good decision is one that chooses the best available course of action in the face of characteristic uncertainty about the consequences. Although some decisions are straightforward, many are more difficult, either because the stakes are high (e.g., the choice of a mate or deciding between potentially life-saving medical treatments) or because there are difficult trade-offs to make between competing options (e.g., the respective benefits and side effects of a drug). To complicate matters, often our decisions are, in addition, dependent on the possible responses of others, for example when we are deciding whether to ask someone on a date or when entering a business negotiation.

Traditional economic models of decision making, such as the Utility family of theories, propose several aspects of a decision that must be integrated in order to choose the best option. Broadly speaking, these are the subjective values of the potential outcomes that may occur and the likelihood of these outcomes actually occurring. These factors are then weighed to produce a utility for each option, with the alternative with the highest utility chosen. In social decisions, this computation is complicated by the fact that we must also attempt to infer the values and probabilities of our partner or opponent in attempting to reach the optimal decision.

Given that humans live in highly complex social environments, many of our most important decisions are made in the context of these social interactions. This article focuses on research investigating social decision making in two-person interactions, although social decision making often involves considerably more than this. Nevertheless, even in the course of dyadic social interactions, humans are regularly confronted with a wide range of decisions that have far-reaching consequences for the welfare of both the people involved and others. Should I affiliate with or trust this person? Should I compete with or be deferential toward this person? How should I respond to a breach of trust by this person? What is the socially appropriate thing to do in this situation?

An additional reason to focus on the neural basis of decisions in the context of social interactions stems from evolutionary considerations. Primate social behavior is more complex than that of most other mammals because primates form alliances and coalitions that compete with one another. On average, primates also have larger brains for their body size than other mammals. Moreover, relative neocortex size is positively correlated with social group size, a presumed proxy for social complexity, across anthropoid primates. These facts suggest that large primate brains may have evolved in response to social complexity in order to support greater social intelligence that would, in turn, allow an individual primate to successfully negotiate these complexities. If humans represent an extension of this primate trend, then a reasonable claim is that the human brain was designed in some measure to deal with complex social interactions and that our brain's decision-making circuitry probably evolved in part to support these types of decisions.

Finally, a number of psychiatric disorders involve deficits in social decision making. For example, depressed patients often withdraw from social interactions, choosing instead to be alone. Psychopathic individuals often manipulate and objectify others, making it difficult for them to establish stable, long-term bonds. Mania is often accompanied by impulsive and frequently damaging decisions about sexual behavior, shopping, travel, and driving. In addition, several axis II personality disorders involve poor choices within interpersonal relationships. Endeavoring to understand the neural bases of these disorders therefore rests in part on an understanding of the neural bases of social decision making.

Tasks

In a similar fashion to the framework provided by Utility Theory for studying individual decisions, Game Theory offers well-specified models for the investigation of social exchange. However, much of the decision behavior actually observed in these tasks deviates, often quite substantially, from the predictions of the standard game theoretic model, suggesting that psychological and neuroscientific approaches may help our understanding of these decisions and judgments. Recent research has combined these tasks with neuroimaging in an effort to gain a more detailed picture of social decision making. This research probes the neural basis of decision making in the context of social interactions and combines behavioral paradigms from

behavioral economics with a variety of methods from neuroscience.

A class of decisions has been investigated by experimental economics tasks, which, although beguilingly simple, require complex reasoning about the motivations and strategies employed by other players and thus offer a useful window into more complex forms of decision making – decisions that may better approximate many of the choices we make in real life.

One specific focus of these games is reciprocal exchange, exemplified by trust games and closely related Prisoner's Dilemma (PD) games. In the Trust Game (TG), a player (the Investor) must decide how much of an endowment to invest with a partner (the Trustee) in the game. Once transferred, this money is multiplied by some factor (usually tripled or quadrupled), and then the Trustee has the opportunity to return some or all of the amount to the Investor but, significantly, need not return any money if he or she decides against it. If the Trustee honors trust and returns money to the Investor, both players end up with a higher monetary payoff than was originally obtained. However, if the Trustee abuses trust and keeps the entire amount, the Investor ends up with a loss. Because the Investor and Trustee interact only once during the game, Game Theory predicts that a rational and selfish Trustee will never honor the trust given by the Investor. The Investor, realizing this, should never place trust in the first place and so will invest zero in the transaction. Despite these grim theoretical predictions, in most studies of the TG a majority of investors do send some amount of their money to the Trustee.

The well-studied PD game is similar to the trust game, except that both players simultaneously choose whether to trust one another without knowledge of their partner's choice (Figure 1). In the PD game, two players choose to either cooperate with one another or not (i.e., defection) and receive a payoff that depends on the interaction of their two choices. The largest payoff occurs when the player defects and the partner cooperates (DC = \$3). The second largest payoff is for mutual cooperation (CC = \$2), followed by mutual defection (DD = \$1), and finally player cooperation combined with partner defection (CD = \$0). Each outcome corresponds to a different outcome of a social interaction and typically elicits a different set of social emotions. Again, game theoretic predictions are that players will immediately default to mutual defection, but in most iterations of the game, players exhibit much more trust than expected, with mutual cooperation a common finding, at least in early rounds of the game. Mutual cooperation is often associated with friendship, love, trust, or obligation; mutual defection is often associated with feelings of rejection and hatred. Cooperation by one person and defection by the other typically results in the cooperator's feeling anger or indignation and in the defector's feeling anxiety, guilt, or elation from successfully exploiting his or her partner to advantage.

		You	
		Cooperate	Defect
Partner	Cooperate	\$2 (\$2)	\$3 (\$0)
	Defect	\$0 (\$3)	\$1 (\$1)

Figure 1 Payoff matrix used in PD game. The scanned player's choices are the column labels and nonscanned partner's choices are the row labels. Payoff to the scanned player is in black; payoff to the nonscanned partner is in red.

These games are good models for the reciprocal exchange of favors. Given that favors are rarely exchanged simultaneously in real life, the TG at first seems a better model for real-life behavior than the PD; however, the PD choices in round n can also be viewed as a response to the partner's choice in round $n - 1$. Both PD and TG can be played as either single-shot or iterated interactions. In the single-shot versions, cooperative moves by the trustee in the TG, or by either partner in the PD, are unambiguously altruistic insofar as individuals are sacrificing personal gain so that their partner can receive more. In the iterated versions of the games, in which several consecutive rounds are played with the same partner, cooperative choices are not necessarily selfless because consistent cooperation with a reciprocating partner generally yields the player higher cumulative earnings over the long term than does a noncooperative strategy that often settles into the equilibrium outcome of mutual selfishness. Whereas economists have emphasized the importance of these interactions being anonymous in order to eliminate reputation effects that could bias choices, others argue that humans are evolutionarily unprepared for such nonnaturalistic social interactions and that subjects should meet and know the identity of their playing partners. Both anonymous and nonanonymous versions of these games have been studied with neuroimaging.

Bargaining games are another common focus of experimental economics, with the family of Dictator and Ultimatum games often used to examine responses to equality and inequality. In the Dictator Game (DG), one player (the Proposer) decides how much of an endowment to award to the second player (the Responder). Allocations in this game measure pure altruism, in that the Proposer sacrifices personal

gain to share the endowment with his or her partner. The Ultimatum Game (UG) is a variant which examines more strategic thinking in the context of two-player bargaining. In the UG, the Proposer and Repsonder are also asked to divide a sum of money, with the Proposer specifying how this sum should be divided between the two. However, in this case the Responder has the option of accepting or rejecting the offer. If the offer is accepted, the sum is divided as proposed. However, if it is rejected, neither player receives anything. In either event, the game is over; that is, there are no subsequent rounds in which to reach agreement. The decision to reject an unfair offer is considered a form of altruistic punishment because the Responder chooses to receive no money rather than the amount offered by the Proposer, presumably to punish the Proposer for making a miserly offer. If people are motivated purely by self-interest, the Responder should accept any offer and, knowing this, the Proposer should offer the smallest nonzero amount. However, again Game Theory predictions are at odds with observed behavior, and in most industrialized cultures, low offers of less than 20% of the total amount are rejected about half the time. There are some interesting differences in more traditional cultures, but in general the probability of rejection increases substantially as offers decrease from 50% to 0%. Thus, people's choices in the UG do not conform to a model in which decisions are driven by immediate self-interest.

Methods

The neural correlates of decision making in these games have been investigated using a variety of methods. One approach uses functional neuroimaging, namely functional magnetic resonance imaging (fMRI) or positron emission tomography (PET), to image changes in blood flow while subjects are playing interactive games in the MRI or PET scanner. Subjects can view computer-projected visual stimuli from inside the scanner, either via goggles that display the visual stimuli or via a mirror that allows the subject to view a projection screen outside the scanner. Because verbal responses can create motion artifacts, subjects generally indicate choices by pressing specific buttons on a response box. With these imaging methods, it is possible to examine regional blood flow during the decision-making epochs of the task and to link these to specific choices.

Imaging studies of social interactions have emerged relatively recently within cognitive neuroscience. Many early fMRI studies presented subjects with stimuli of other human faces, given the obvious importance of faces in human social interactions. Typically, these stimuli were static, two-dimensional pictures of faces that subjects were instructed to either passively view or judge on some attribute (gender, age, etc.). Using similar types of stimuli, other studies examined face-processing deficits in patients with damage to specific brain regions such as the amygdala or the fusiform gyrus. Still others have attempted to probe social cognition by asking subjects to read stories or view cartoons and then make judgments about these hypothetical scenarios. For example, the neural correlates of both mentalizing and moral reasoning have been probed with this methodology. These studies have yielded valuable insights with respect to the neural underpinnings of human social cognition. However, for each, we can raise questions about the ecological validity of the stimuli. Does the pattern of brain activation in response to static two-dimensional face stimuli accurately reflect the brain's response to the dynamic embodied faces that we encounter in everyday life? Is the pattern of brain activation in response to reasoning about hypothetical fictitious scenarios the same as when grappling with significant real-life social problems? Is mentalizing about the actions of another person the same as making a consequential decision based on these actions?

One approach to improving the ecological validity of experiments in social cognitive neuroscience is to image brain function as subjects actually interact with other people in real social interactions. Recent innovative studies have imaged human subjects while playing both TG and bargaining games with partners communicating from outside the scanner. Potentially even more exciting, hyperscanning technology has been developed that makes it possible to image brain function in two or more interacting partners simultaneously, by using network connections between two separate scanners. Hyperscanning has obvious advantages in terms of data-collection efficiency (i.e., collecting twice as much data in the same amount of time), but will also open new vistas in social cognitive neuroscience. For example, it will allow the imaging of coordinated patterns of brain activity in people who are effectively working together toward a common goal. It could also be used to evaluate simulation theories of empathy, according to which we understand others by reproducing their neural states. Further applications for this method will undoubtedly emerge in the future.

Another approach to investigating the neural correlates of social decision making involves manipulating specific neurotransmitter systems and examining the effect on game-playing behavior. For example, dietary tryptophan depletion can be used to decrease brain serotonin levels, and central oxytocin (OT) levels can be elevated by intranasal self-administration of OT. Still another approach involves the use of transcranial magnetic stimulation (TMS) to temporarily activate

or deactivate a brain region and then examine the effect on a decision-making task. Finally, patients with circumscribed brain damage to particular regions can be tested in these games to see if the damaged brain area has an impact on social decision making.

Findings

Although the methods and tasks outlined here seem to offer exciting prospects for research on social decision making in years to come, the very recent development of this field has meant that the definitive findings revealed thus far are understandably limited. However, studies have outlined a number of tantalizing findings with respect to both brain regions and neurotransmitter systems that appear to be involved in decision making in the context of social interactions.

The Caudate Nucleus

There is good reason to suspect that mesencephalic dopamine projections to the caudate nucleus are involved in decision making in the context of reciprocal exchange (Figure 2(a)). Single-cell recording studies in monkeys demonstrated that midbrain dopamine neurons track reward prediction errors. Unexpected rewards increase the firing rate of midbrain dopamine cells, whereas the omission of expected rewards decreases their firing rate. These reward prediction error signals are hypothesized to play a critical role in learning stimulus–response associations, thereby shaping decision making. Midbrain dopamine cells project to both the ventral and dorsal striatum, including the caudate nucleus, and the activation in the human caudate nucleus is modulated as a function of trial-and-error learning with feedback, suggesting that the caudate nucleus may allow an organism to learn contingencies between its own responses and either rewarding or punishing outcomes. Several neuroimaging studies have demonstrated that the human caudate tracks a social partner's decision to reciprocate or not reciprocate cooperation in the TG or PD. This caudate response is also related to future game decisions. Specifically, reciprocated cooperation activates the caudate nucleus and unreciprocated cooperation deactivates this area, in line with the reward prediction error already described (Figure 3). In addition, caudate activation is associated with increased cooperation in subsequent rounds. These findings suggest that the caudate registers social prediction errors that guide decisions about reciprocity. Interestingly, these prediction error signals from partner feedback

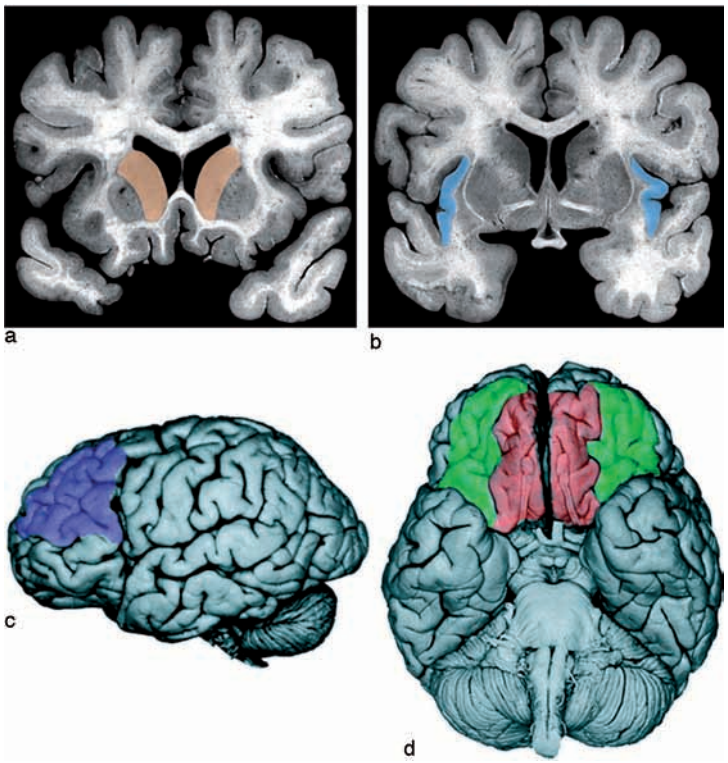

Figure 2 Brain regions involved in decision making in the context of social interactions: (a) caudate nucleus; (b) anterior insula; (c) dorsolateral prefrontal cortex; (d) ventromedial prefrontal cortex. (c, d) From Davidson RJ, Putnam KM, and Larson CL (2000) Dysfunction in the neural circuitry of emotion regulation–possible prelude to violence. *Science* 289(5479): 591–594.

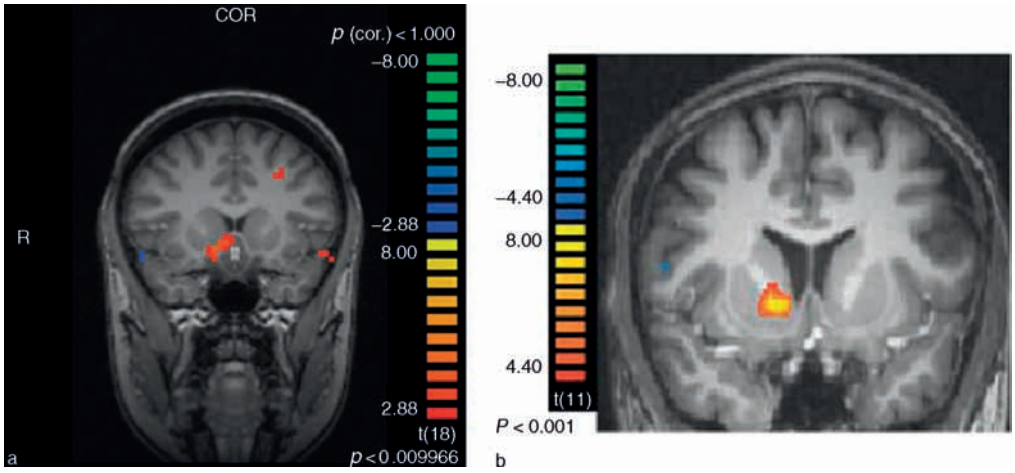

Figure 3 Ventral caudate activation for the contrast between partner reciprocation and nonreciprocation: (a) in an iterated PD game; (b) in an iterated Trust game. (a) From Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, and Cohen JD (2004) Opposing BOLD responses to reciprocated and unreciprocted altruism in putative reward pathways. *NeuroReport* 15: 2539–2543. (b) From Delgado MR, Frank RH, and Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience* 8: 1611–1618.

can be blunted or even absent when people instead base their decisions on the prior moral reputations of social partners. That is, moral reputations lead to a discounting of feedback information from social partners, demonstrating both top-down and bottom-up influences on the neural basis of social cooperation. Finally, activation in caudate nucleus is also related to satisfaction with punishing nonreciprocators for their norm violations, even when this punishment involves a financial loss to the subject.

Anterior Insula

Activation within the anterior insula has been observed in response to a variety of negative social interactions, from social exclusion in a virtual ball-tossing task to unreciprocated cooperation in an iterated PD game to receiving an unfair offer in the UG to watching a loved one receive a painful stimulus (Figure 2(b)). The anterior insula is also responsive to physically painful stimuli, and its activity is correlated with skin conductance responses. These results, and others, suggest that the anterior insula is involved in mapping physiological states of the body, including pain, touch, and visceral sensations of autonomic arousal. The right anterior insula, in particular, is thought to be a site of interoceptive awareness that gives rise to subjective feeling states that may play a role in decision making. Finally, recent fMRI data implicate the right anterior insula in aversive conditioning. In the UG, unfair offers that are subsequently rejected have a stronger average anterior insula response than unfair offers that are subsequently accepted (Figure 4), and activation in the anterior insula is related to the degree of unfairness of the offer. In the iterated PD game, individuals with a stronger anterior insula response to unreciprocated cooperation have a higher frequency of defection, suggesting that the anterior insula marks the social interaction as aversive and discourages trusting the nonreciprocating partner in the future. Thus, activation within the anterior insula may be involved in marking negative social interactions as aversive and biasing subsequent social decision making.

Although both the anterior insula and the caudate nucleus are involved in social decision making, both are also involved in nonsocial decision making such avoiding physically noxious stimuli and adaptive foraging, respectively. These neural systems probably initially evolved to support these more fundamental aspects of behavior. When social skills became more crucial with the evolution of primates, these systems may have been exapted for new functions, such as detecting harmful social stimuli and learning when, and to whom, altruism should be dispensed. However, in the process of adapting these old systems to novel demands, they may well have been modified and social pressures may have left their imprint. Indeed, this is the manner in which the evolutionary process typically unfolds. For example, some land mammals such as otters have evolved to fill aquatic niches. But rather than evolving completely novel structures to propel them through water, they modified their feet, the primary function of which is locomotion on land, to make them better suited to also travel through water. The end result is limbs designed for terrestrial locomotion combined with webbed feet designed for aquatic locomotion. That is, the otter limb is adapted to both forms of locomotion. Analogously, human brain systems may be designed for a combination of social and nonsocial functions.

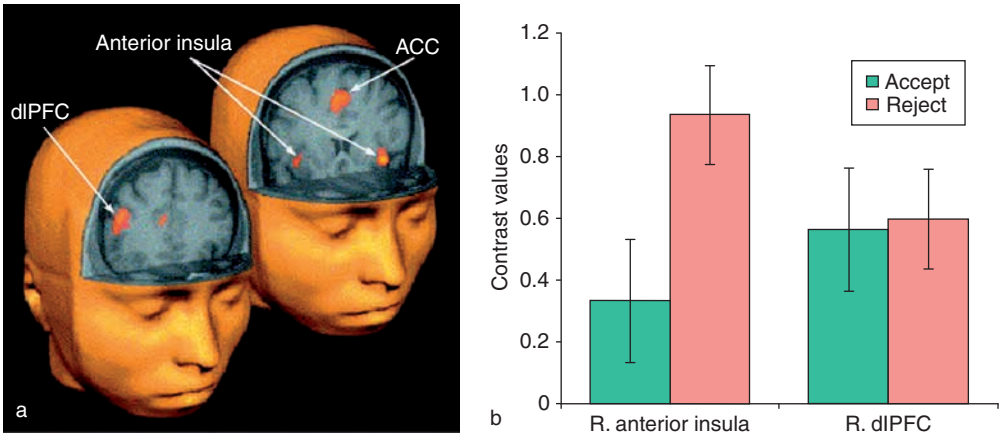

Figure 4 Brain responses to unfair offers: (a) activated brain regions in response to receiving an unfair (vs. fair) offer in the UG; (b) ratio of anterior insula to dorsolateral prefrontal cortex activation in response to unfair offers as a function of whether the offer was accepted or rejected. ACC, anterior cingulate cortex; dlPFC, dorsolateral prefrontal cortex; R., right. Reproduced from Sanfey AG, Loewenstein G, McClure SM, and Cohen JD (2006) Neuroeconomics: Cross-currents in research on decision-making. *Trends in Cognitive Science* 10: 108–116.

Dorsolateral Prefrontal Cortex

The dorsolateral prefrontal cortex (dlPFC) has long been recognized as critically involved in cognitive control, including cognitive control over emotions (Figure 2(c)). Recent studies suggest its specific involvement in overriding prepotent emotional biases, such as when delaying gratification or making utilitarian decisions in the context of moral dilemmas. The dlPFC is also activated by unfair offers in the UG, which subjects are more likely to accept when dlPFC activation exceeds anterior insula activation (Figure 4(b)). This has led to the hypothesis that UG decisions, and difficult social decisions more generally, often involve competition between emotional processing and higher-level controlled or deliberative processing that bias decision making in opposite ways. Emotional processes are driven by the subcortical, limbic, and paralimbic structures, whereas the deliberative processes rely on the anterior and dorsolateral regions of prefrontal cortex, as well as areas of posterior parietal cortex. Further evidence in support of the a role for the dlPFC in social decision making is provided by studies showing that application of repetitive transcranial magnetic stimulation (rTMS) to the dlPFC influences decision making in the UG. Finally, dlPFC activation is also associated with the choice to defect in the PD game, suggesting that most people may in fact have a prepotent emotional bias to cooperate in the game.

Ventromedial Prefrontal Cortex

Studies of patients who have suffered damage to the ventromedial prefrontal cortex indicate that this brain region is critical for appropriate social decision making. Previously well-adapted individuals who sustain damage to this region become unable to observe social conventions and unable to decide advantageously on matters pertaining to their own lives. Because these patients have generally well-preserved intellectual abilities and knowledge of appropriate social behavior, it has been hypothesized that their deficit lies in an inability to use emotions and feelings to guide social decision making. The result is that they often opt for behaviors that, although yielding immediate rewards, often have larger future negative consequences associated with them. This leads to the obvious but so far untested hypothesis that these patients should be less likely to endure the short-term cost associated with reciprocating altruism and should consequently have trouble sustaining long-term, mutually beneficial social relationships. This hypothesis could be formally evaluated by testing patients in an iterated PD game.

The Oxytocin System

In rodents, the neuropeptides oxytocin (OT) and vasopressin (AVP) facilitate social affiliation, making them attractive candidates for the neural modulation of human social relationships. Accordingly, in a TG, intranasal OT infusion was shown to increase initial monetary transfers by Investors. Recent research suggested that neuropeptides cross the blood–brain barrier after intranasal administration, bypassing the bloodstream, implying that these effects are due to OT's action in the brain. Intranasal OT administration has recently been combined with fMRI to assess the impact of OT on the neural response to fearful/threatening faces and scenes. Compared with

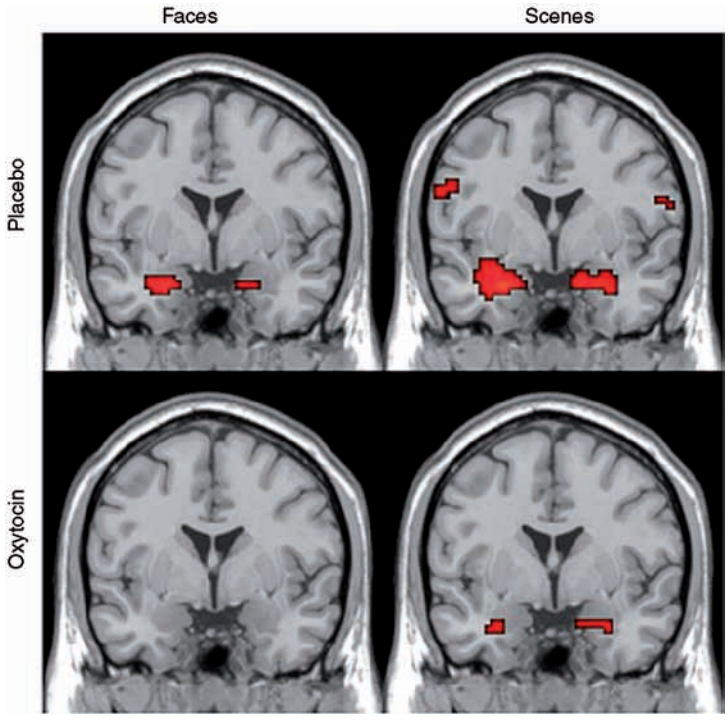

Figure 5 Intranasal oxytocin administration decreases amygdala response to threatening faces and scenes. Reproduced from Kirsch P, Esslinger C, Chen Q, et al. (2005) Oxytocin modulates neural circuitry for social cognition and fear in humans. *Journal of Neuroscience* 25: 11489–11493.

a placebo, OT decreased the activation in the amygdala for both types of fearful stimuli (Figure 5). These results are consistent with evidence that OT reduces stress and anxiety and are also consistent with the abundance of OT receptors in the rodent amygdala, although OT receptors have not yet been found in the human amygdala. This study suggests one potential mechanism by which OT could increase trust – by reducing anxiety about potential nonreciprocation. Another possibility suggested by animal experiments is that OT release in the nucleus accumbens renders social interactions more rewarding.

The Serotonin System

Serotonin plays a significant role in modulating prosocial behavior. Peripheral and central indices of serotonin function show negative associations with violent and aggressive behavior, but positive associations with socially affiliative behavior (such as grooming and approach) in primate models. Pharmacological interventions that increase serotonergic activity also facilitate affiliative behavior. These results suggest the possibility that serotonin levels could influence decision making in the context of social interactions. Indeed, depleting subjects of tryptophan, the amino acid precursor of serotonin, is associated with reduced rates of cooperation in the PD game. Given that serotonin is known to modulate dopamine reward pathways, these effects could be mediated by serotonin withdrawal's attenuating the rewarding and reinforcing effects of cooperation.

Conclusion

In summary, the neuroeconomics approach to social decision making has already yielded some interesting findings and hypotheses, and the combination of neuroscientific, economic, and psychological methods and tasks seems sure to uncover much more about the social brain in the coming years.

See also: Decision-Making and Neuroeconomics; Delayed Reinforcement: Economics; Game Theory and the Economics of Animal Communication; Games in Monkeys: Neurophysiology and Motor Decision-Making; Neuroeconomics: History; Neuroendocrinology of Social/Affiliative Behavior; Reward Decision-Making; Social Brain: Evolution; Social Cognition; Social Stress in Adult Primates.

Further Reading

Bechara A, Damasio H, and Damasio AR (2000) Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex* 10: 295–307.

Craig AD (2003) Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology* 13: 500–505.

Critchley H, Elliot R, Mathias C, and Dolan R (2000) Neural activity relating to the generation and representation of galvanic skin conductance responses: A functional magnetic resonance imaging study. *Journal of Neuroscience* 20: 3033–3040.

Davidson RJ, Putnam KM, and Larson CL (2000) Dysfunction in the neural circuitry of emotion regulation – possible prelude to violence. *Science* 289(5479): 591–594.

de Quervain DJ, Fischbacher U, Treyer V, et al. (2004) The neural basis of altruistic punishment. *Science* 305: 1254–1258.

Delgado MR, Frank RH, and Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience* 8: 1611–1618.

Henrich J, McElreath R, Barr A, et al. (2006) Costly punishment across human societies. *Science* 312: 1767–1770.

King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, and Montague PR (2005) Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308: 78–83.

Kirsch P, Esslinger C, Chen Q, et al. (2005) Oxytocin modulates neural circuitry for social cognition and fear in humans. *Journal of Neuroscience* 25: 11489–11493.

Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, and Fehr E (2005) Oxytocin increases trust in humans. *Nature* 435: 673–676.

Knoch D, Pascual-Leone A, Meyer K, Treyer V, and Fehr E (2006) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314: 829–832.

Lim MM and Young LJ (2006) Neuropeptidergic regulation of affiliative behavior and social bonding in animals. *Hormones and Behavior* 50: 506–517.

Miller E and Cohen J (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24: 167–202.

Montague PR, Berns GS, Cohen JD, et al. (2002) Hyperscanning: Simultaneous fMRI during linked social interactions. *Neuroimage* 16: 1159–1164.

Montague PR, Dayan P, and Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* 16: 1936–1947.

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, and Cohen JD (2004) Opposing BOLD responses to reciprocated and unreciprocted altruism in putative reward pathways. *NeuroReport* 15: 2539–2543.

Sanfey AG, Loewenstein G, McClure SM, and Cohen JD (2006) Neuroeconomics: Cross-currents in research on decision-making. *Trends in Cognitive Science* 10: 108–116.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, and Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. *Science* 300: 1755–1758.

Schultz W (1997) Dopamine neurons and their role in reward mechanisms. *Current Opinion in Neurobiology* 7: 191–197.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, and Frith CD (2004) Empathy for pain involves the affective but not sensory components of pain [see comment]. *Science* 303: 1157–1162.

van't Wout M, Kahn RS, Sanfey AG, and Aleman A (2005) Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *NeuroReport* 16: 1849–1852.

Wood RM, Billing JK, Sanfey AG, Bhagwagar Z and Rogers RD (2006) Effects of tryptophan depletion on the performance of iterated Prisoner's Dilemma game in healthy adults. *Neuropsychophamacology* 31(5): 1075–1084.

Social Interaction Effects on Reward and Cognitive Abilities in Monkeys

J R Stevens and M D Hauser, Harvard University, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When given a choice between receiving $100 in 30 days and $110 in 31 days, most people wait for the larger, more delayed reward. However, when the choice is between $100 today and $110 tomorrow, many people shift their preferences to the smaller, immediate reward despite the same difference of $10 and 1 day. Reward properties – such as time to receipt – play a crucial role in our processing of rewards. However, reward processing does not only occur for individual decisions but also for social interactions; we must decide whether to cooperate or compete with others for rewards. Humans stand out in the animal kingdom as exceptional cooperators, both in terms of the form that cooperation assumes as well as the nature of rewards attained. Regarding form, we are unique in the stability of our reciprocal interactions and in the scale of our cooperative coalitions, entailing multiple nation states in times of war. In terms of rewards, we are of course motivated, like all other animals, by the central survival payoffs such as food and water but also by abstract entities such as money, the promise of future support, and positions of power such as a king, president, or head of an academic department. To maintain these complicated interactions and evaluate the nature of reward, we must possess a number of prerequisite cognitive abilities. Here, we view this problem through the lens of evolutionary biology, asking which aspects of our cognitive machinery, and the social interactions it supports, are uniquely human and which are shared with other primates. Though we focus on lemurs, monkeys, and apes, we acknowledge that many of the processes we document are unlikely to be restricted to the primates, and, in many cases, there is already comparable evidence from other mammals and birds. We begin by reviewing a suite of cognitive mechanisms that are involved in both human and nonhuman primate reward processing. This article is particularly focused on the subset of situations with quantifiable rewards. We then describe the kinds of social interactions that are part and parcel of primate life, especially the highly social monkeys and apes. Finally, we merge these two sections and consider how constraints on primate cognition may limit the complexity of primate social interactions.

Primate Cognition

To set up the problem of reward processing in animals, we can distinguish between asocial, individually based processes and social ones. The primary difference is whether access to rewards entails competitive or cooperative social interactions or instead can be achieved by an individual on its own. Some processes operate across asocial and social contexts. For example, an individual seeking food as a reward must discriminate between food and nonfood objects, quantify the amount of food available and its relative value, recall where food is located in the environment as well as the timing of its seasonal availability, and inhibit the temptation to take a small, immediately available reward over a larger, more valuable but delayed reward. Social interactions add a layer of complexity to these processes, raising questions about the extent to which individuals can generate expectations about what others want and intend, their competitive abilities, and the value of a coalitionary partner.

Primates are an ideal group in which to study the comparative cognition of reward processing because they are phylogenetically closely related to humans and because primate species vary in their life histories, mating systems, and foraging ecologies. Thus, we can not only explore how these different socioecological parameters uniquely shape each species' capacity for reward processing but also explore the extent to which there are more general, derived mechanisms, having evolved before humans diverged from their phylogenetic ancestors.

Properties of Rewards

For folivorous, frugivorous, and omnivorous primates, individuals must distinguish between edible and inedible items, especially given the possibility of ingesting potentially toxic items. For many animals, recognition of edible foods entails using chemical cues, color patterns, textures, and shapes. And in some cases, these cues or features are detected in the context of social interactions. For example, in a study of rhesus macaques (*Macaca mulatta*), an experimenter ate one of two novel food items while a subject watched. The experimenter then placed new copies of each these food items on the ground and

allowed the subject to approach and select one. Consistently, subjects picked the food item eaten by the experimenter. To assess which features mediated response selection, a set of follow-up conditions put different features into competition, using artificially created foods. For example, in one condition, the experimenter ate an orange sphere while holding a blue cube and then presented a blue sphere and an orange cube. Subjects consistently picked the orange cube, showing that in the context of a social foraging problem, rhesus macaques use color over shape to identify valuable food.

Number

To reliably process rewards, animals must be able to quantify reward amounts, either using number or some continuous dimension of amount such as volume. There are at least two naturally occurring situations in which number would appear to matter for primates: aggressive competition between groups and foraging for food. In black howler monkeys (*Alouatta pigra*) and chimpanzees (*Pan troglodytes*), individuals attend to the number of competitors. Playback experiments demonstrate that a group of males is more likely to approach foreign males if the numerical odds are in their favor. Additionally, in many primate species, two or three individuals will form coalitions to defeat either a single dominant individual or a smaller coalition. Although these coalitions involve small numbers, they nonetheless require some capacity to quantify the number of competitors to acquire rewards such as mates or foraging areas.

Studies of foraging in animals show that individuals often maximize the rate of energetic returns, choosing patches with more over less food. Since estimates of rates of return depend on quantifying amount of food consumed over time, researchers have asked whether animals count the pieces, estimate the volume, or time the foraging periods in a patch. For instance, when given the choice between two different amounts of apple slices, rhesus macaques consistently picked two apples over one, three apples over two, and four apples over three. But when presented with five versus four apple pieces, some animals picked four and some five. Without training, rhesus monkeys can quantify the number of pieces of food up to about four but thereafter appear to rely on a general estimate of amount.

Based on an overwhelming number of carefully controlled experiments, it is now fair to say that animals have a number sense that likely consists of two naturally available systems. One system allows animals to quantify individuals up to about four with precision; the second allows them to approximate number but without any limits on magnitude. In the precise system, as evidenced by the rhesus macaque example, primates are limited to representing four items simultaneously in short-term memory. Human infants and adults demonstrate a similar limit in a variety of tasks, suggesting continuity across development and across species. In the approximate system, primates can discriminate between approximate quantities using analog magnitudes, in which performance is limited by the ratio between the quantities and is independent of absolute value. Other work on operantly trained rhesus macaques demonstrates that ratio determines numerical discrimination between quantities ranging from 1 to 30 items. Ratio is important because it suggests that primate numerical competence conforms to Weber's law, a psychophysical process in which the accuracy of discriminating different quantities scales with the magnitude of the quantities. Again, human infants and adults also represent large approximate numbers and show similar signature ratio limits.

Timing

Predicting the timing of reward availability is a critical capacity for primates. In many species, crucial elements of the diet are only seasonally available, requiring both timing and planning. For example, when chimpanzees discover a source of ripe palm nuts, they not only invest in the processing of these nuts with stone hammers and anvils, but often leave behind particularly good hammers, retrieving them on future visits. Studies conducted with captive apes reveal that they can recall the kinds of tools needed for certain tasks, and retrieve them in the service of planning for future tasks. Based on their patterns of movement, it is also clear that many primates store information about fruiting cycles, enabling them to time the optimum period of visitation and harvesting.

Like number, the capacity to time events is based on a mechanism that, both behaviorally and neurobiologically, shows the signature of the analog magnitude system that underlies number. Importantly, timing exhibits the scalar property of Weber's law: the variance of timing accuracy scales with interval magnitude and is therefore a function of the ratio between magnitudes. For instance, rhesus macaques showed similar accuracy levels between timing events with a 1:1.5 ratio, regardless of whether the discrimination was 4 versus 6 s, 6 versus 9 s, 8 versus 12 s, or 10 versus 15 s. Unfortunately, little work has explored primate timing abilities beyond a few seconds, with the exception of periodic timing of circadian cycles. Moreover, the work is largely restricted to studies involving operant training, with no work on the relationship between timing and the role of social or ecological contexts.

Impulsivity

Animals must often combine quantity and temporal information to choose between rewards available over different time intervals (intertemporal choice). For example, imagine a monkey encounters a tree full of unripe fruit but with only a few ripe fruit available: should it spend the time and energy to consume the few ripe fruit now or wait for the rest to ripen? Waiting would yield a higher reward amount, but the future is uncertain – another monkey may eat the fruit in the meantime, winds may knock them into a stream below, or a fungus may infest and spoil the perfectly good fruit. This uncertainty may have provided a strong adaptive benefit for impulsivity.

Decision makers can make intertemporal choices by means of at least two mechanisms. The first mechanism simply assumes that the forager maximizes its food intake rate over repeated foraging bouts. This rate maximization account predicts that foragers should maximize the total energy gained from rewards over the total time spent acquiring and processing the rewards. Evidence from operant experiments suggests that cotton-top tamarins (*Saguinus oedipus*) and common marmosets (*Callithrix jacchus*) may use rate maximization strategies when trading off future benefits. Temporal discounting represents the second mechanism and involves the subjective devaluation of future rewards. If animals discount the future (as the example in the introduction implied), delayed rewards are less valuable than immediate rewards, potentially resulting in impulsive choice. In experiments that estimate discounting levels, subjects are presented with two stimuli: one associated with a small, immediate reward and the other with a large, delayed reward. The discounting level is 'titrated' by incrementally increasing the delay-to-large until subjects are indifferent between choosing the large, delayed reward and the small, immediate reward. Therefore, researchers can find indifference points between immediate and delayed rewards over a range of small and large reward amounts. Psychologists have extensively studied discounting in rats and pigeons, but fewer studies have investigated primate discounting. Surprisingly, the handful of primate discounting studies that have been conducted demonstrate that cotton-top tamarins, common marmosets, brown capuchin monkeys (*Cebus apella*), and rhesus macaques show similar impulsivity levels to rats and pigeons.

The high levels of discounting observed for most nonhuman animals suggests that when rewards are distributed over time, individuals place a premium on immediacy, forgoing delayed rewards. This not only puts severe constraints on the foraging strategies of monkeys and apes, but constrains other behaviors such as cooperation, territory defense, and mate selection.

Inhibitory Control

Inhibiting prepotent responses to reward is a difficult task for nonhuman animals. Clearly, natural selection has favored a strong drive to obtain the largest reward available. This is evidenced by the performance of a number of primate species on the reversed-contingency task in which subjects must reach toward the smaller of two visible rewards to receive the larger reward. Brown and black lemurs (*Eulemur fulvus* and *Eulemur macaco*), tamarins, squirrel monkeys (*Saimiri sciureus*), rhesus macaques, orangutans (*Pongo pygmaeus*), gorillas (*Gorilla gorilla*), bonobos, and chimpanzees all invariably choose the large reward, only to receive the small reward – they seem incapable of controlling their desire to reach for the larger reward amount.

The reversed-contingency task has proven difficult for primates to solve but not impossible. A number of techniques can circumvent the inhibitory control problem. First, sheer repetition can allow subjects to choose the smaller and receive the larger food reward. Most subjects do not solve the task after 200 trials, though some are at chance rather than repeatedly choosing the smaller amount. After 400–2500 trials, however, six rhesus macaques received the larger reward in over 90% of trials. Thus, massive exposure to the reversed contingencies can eventually trump the inhibitory problems. Second, imposing costs for 'incorrect' choice can facilitate receiving the larger reward. By withholding all food when subjects pick the larger quantity, macaques and squirrel monkeys eventually learn to pick the smaller quantity; tamarins stick with the losing strategy, picking the larger and getting nothing at all. This suggests that part of the failure to point to the smaller quantity is due to the lack of costs associated with pointing to the larger quantity. Finally, including a mediator between seeing the food and receiving the food can curb the impulse to choose the larger reward amount. This has been most noticeably documented by an experiment testing chimpanzees trained on the Arabic number symbols from 1 to 6. Instead of choosing between one food treat versus four treats, the chimpanzees chose between a card with the numeral 1 written on its face and a card with the numeral 4; each card covered up the corresponding number of food treats. The chimpanzees quickly learned to pick the number 1 card and received four treats, indicating chimpanzees can learn a rule like "Point to the one you don't want to get the one you want." Surprisingly, once they learned

this rule with symbols, they were incapable of generalizing to the same problem with food presented on its own. Therefore, it seems as though the difficulty of this task results from the chimpanzees' strong motivation to reach for food rather than an inability to learn the reverse-contingency rule.

Memory

Memory is a multifaceted cognitive ability with clear evolutionary importance. Researchers have delineated several different types of memory that are applied to different types of information. Here we will focus on three types of memory – short-term, spatial, and episodic memory. One of the simplest tasks used to test short-term memory is the delayed-response task. In this task, a reward is placed under one of two (or more) identical but spatially separated stimuli. After a retention interval, subjects are allowed to chose one of the stimuli and receive the rewards if correct. Comparative data show that rhesus macaques are fairly good at remembering the location of the food item. They begin with about 80% accuracy after only a few seconds and drop to about 68% accuracy after a 15 s retention interval. Marmosets, however, though also starting at 80% accuracy, drop to 58% accuracy at 15 s, only slightly better than chance. Though some evidence suggests that the Old World monkeys may retain information better than the New World monkeys, this is by no means a universal, with capuchin monkeys being a notable exception to the rule.

Spatial memory plays an important role in primate reward acquisition. Recalling when and where food rewards are distributed is critical in foraging situations. In fact, some have suggested that foraging problems have posed strong evolutionary pressures on primate cognition, with species that feed on spatially and temporally variable foods showing higher levels of brain encephalization than species that feed on more stable food sources. Numerous studies have demonstrated strong spatial skills in primates, particularly in analogs to the radial arm maze. In this task, feeder stations placed throughout an enclosure are baited before subjects enter. After a subject finds all of the stations, researchers bait only some of the stations and, following a retention interval, record whether the subjects first attempt to feed at the previously baited stations. Most of the experiments on primates show high accuracy over retention intervals as long as 24 h. As predicted by the adaptive specialization hypothesis, species that search over longer distances for food in the wild tend to be more accurate at longer retention intervals (e.g., frugivorous golden lion tamarins (*Leontopithecus rosalia*) are more accurate than gummivorous Wied's marmosets (*Callithrix kulhi*)). Additionally, the contingency between space and reward plays an important role. Common marmosets, for instance, are more accurate in a win-stay paradigm in which previous food locations predict current locations than in a win-shift paradigm in which subjects are not rewarded for visiting previously visited food stations.

Probably one of the most interesting and controversial areas of primate memory is episodic memory. There is a feisty debate in the literature about whether nonhuman animals have the ability to mentally travel in time to recall previous experiences. The benchmark requirements for demonstrating episodic memory include showing that animals know 'what,' 'where,' and 'when' an experience occurred. Though this is difficult to study in nonlinguistic creatures, recent research provides suggestive evidence of episodic-like memory in corvids and primates. For example, up to 24 h after receiving a particular food item from an experimenter, a gorilla correctly identified the food type received and the experimenter from whom it was received.

Theory of Mind

Another highly contentious area of primate cognition is theory of mind. As mentioned earlier, this aspect of primate cognition is not strictly about reward. Yet, the capacity to make inferences about what others know, desire, believe, and intend often leads directly to cooperative and competitive rewards, and most of the experimental studies use food as a motivator to explore these mental states.

Early studies failed to find evidence of theory of mind in primates. One explanation for these failures was that the studies relied on animals making inferences about human mental states in contexts involving cooperation, as opposed to asking animals to make inferences about each other in the context of competition. Though several primate species do cooperate (see below), competition is far more common and may represent the most basic form of social interaction among animals. To test this hypothesis, recent studies of monkeys and apes have used both competitive and cooperative tasks, contrasting conspecific pairings as well as conspecific–heterospecific pairings. There is increasing evidence that monkeys and apes make inferences from seeing to knowing. Chimpanzees, for instance, can take the visual perspective of other individuals to understand whether they can see occluded food rewards: subordinate chimpanzees preferentially approach rewards that are obscured from the view of dominant individuals. Rhesus monkeys show similar abilities, stealing food primarily

from experimenters without visual access to the food rather than from those that do have access. Though visual perspective taking is only one component of having a theory of mind, it plays a significant role in the competitive and cooperative interactions of primates. What remains to be explored is whether primates go beyond this elemental capacity, attributing intentions, desires, and beliefs to others.

Primate Social Interactions

In socially living animals, such as some of the lemurs and most monkeys and apes, obtaining rewards is often coupled with socially competitive or cooperative interactions. In some instances, helping another individual yields payoffs for the helper, while in other cases, helping is costly for the helper. The existence of the latter poses a puzzle given the generally accepted Darwinian assumption that behavior evolves via natural selection to provide fitness benefits to the individual. Specifically, what selective pressure has favored individuals who provide benefits to other individuals?

Here, we address the evolution of cooperation by describing four different models (mutualism, kin selection, reciprocity, and punishment), focusing on evidence from monkeys and apes. We operationally define cooperation as social interactions resulting in a net benefit for all participants; here, benefit is measured in terms of increasing reproductive success.

Mutualism

The simplest type of cooperative behavior provides direct benefits to the cooperator, in addition to other individuals. Therefore, benefits to others are by-products of benefits to self. This model of cooperative behavior is termed by-product mutualism. Any individual that defects (does not cooperate) in mutualistic situations will, by definition, do worse than a cooperator; therefore, in the absence of a temptation to defect, cooperation provides the best option. Importantly, mutualism does not depend on the identity of one's partner and thus can occur between any members of the same species and even members of different species.

Mutualism is a common form of cooperation, with cooperative hunting providing a prime example. When multiple individuals cooperate to hunt the same prey, they can both increase the probability of successful prey capture and reduce the individual costs associated with hunting. Cooperative hunting provides mutualistic benefits only when the per capita intake rate increases with group size. Therefore, a pair of hunters would have to capture more than twice as many prey items as a solitary hunter. This typically requires the success rate of solitary hunters to be fairly low, making cooperation particularly successful. Increased success combined with lower hunting costs can lead to direct, immediate, and simultaneous fitness benefits for cooperative hunters, particularly when hunting small or difficult prey. For example, wild chimpanzees often cooperatively hunt arboreal monkeys. Although different populations vary, there is evidence that in some populations the per capita intake rate increases with group size, suggesting that the hunting is cooperative. Cheaters that do not hunt are discouraged by receiving only small amounts of meat in food-sharing bouts following a kill. Thus, it is in an individual's best interest to participate in a hunt to acquire the much-sought-after meat.

In the laboratory, mutualism has been documented in a variety of primates, including marmosets, capuchin monkeys, and chimpanzees. Typically, when joint action is required to receive separate benefits for each player, subjects succeed in cooperating. However, when food is monopolizable instead of divided between players, cooperation breaks down. That is, when immediate rewards are available, primates cooperate. But once competition reduces the probability of reward, there is no guarantee of immediate benefit and no apparent mechanism capable of maintaining cooperation. In a recent study of captive chimpanzees, individuals confronted an opportunity to manipulate an apparatus for a food reward; on some trials, the focal subject had to recruit the help of another individual in order to manipulate the apparatus and on other trials, the focal could operate the apparatus alone. Subjects only recruited help when two individuals were required to obtain the food reward. Moreover, in a second experiment, subjects selectively recruited effective helpers over ineffective helpers, demonstrating a keen knowledge of the role of others in the mutualistic acquisition of rewards.

Kin Selection

Whereas mutualism poses no problem for classic Darwinian theory, altruistic cooperation does as the actor incurs a cost while benefiting another. Kin selection provided the first clear theoretical solution to the paradox of altruism. Individuals may bias cooperation toward their genetic relatives because it helps propagate their own genes shared by common descent. What looks altruistic from an individual's perspective actually serves self-interest from the gene's view. Individuals have a certain probability of sharing a gene (r, the coefficient of relatedness) with

relatives due to common descent. If the benefits to kin discounted by this coefficient of relatedness outweigh the costs of helping, altruism toward kin can evolve.

Kin selection may be particularly powerful in primates because of the close-knit family groups found in many species. With a few notable exceptions, including chimpanzees, nonhuman primates follow the general mammalian pattern of philopatry in which males leave their natal groups upon reaching reproductive maturity whereas females stay in their natal groups for life; consequently, most primate groups consist of closely related female kin and genetically unrelated males. Kin selection theory predicts that individuals should preferentially help relatives and should help in proportion to their coefficient of relatedness. These types of quantitative predictions have proven difficult to test in primates, but there are data showing kin-directed helping, particularly in coalition formation and food sharing.

Primates frequently form coalitions in which multiple individuals form an alliance in agonistic interactions with other group members. This is particularly prevalent in primate species with matrilineal dominance hierarchies, such as macaques and baboons. In these species, females often assist relatives in aggressive encounters with other group members. Notably, the probability of coalitionary support tapers off with increasingly distantly related kin – mothers almost always aid offspring but are less likely to aid nieces and nephews. Therefore, as relatedness increases, these primates are more likely to pay costs associated with helping kin. Additionally, individuals are more likely to form coalitions with kin than nonkin against higher-ranking opponents, suggesting that they will pay a higher cost of helping related individuals. Male primates can also form coalitions, often to guard or acquire sexually receptive females. In situations in which few mating opportunities exist for males, these kin-based coalitions are especially relevant because aiding kin helps ensure the propagation of one's genes, even in the absence of direct mating opportunities. This is the case in red howler monkeys (*Alouatta seniculus*), where coalitions of related individuals have longer tenure in a troop (and therefore more mating opportunities) than unrelated individuals. Though many instances of coalitionary support are directed to kin, some are directed to nonkin and may be attributable to other benefits such as mutualism and reciprocity.

One of the most often-reported examples of apparent altruism is food sharing – voluntarily allowing another to consume food that one individual possesses. Food sharing is, of course, very common between mother and offspring for numerous primate species. However, it also occurs between grandparents and grandoffspring, siblings, and even adult relatives, often scaling with relatedness. In captive macaques, for instance, mother–daughter pairs ($r = 0.5$) frequently fed from the same food bin. Both grandmother–granddaughter pairs and sister–sister pairs ($r = 0.25$) co-fed at intermediate rates, and aunt–niece pairs ($r = 0.125$) co-fed at low levels indistinguishable from unrelated pairs. Though it is unclear what proportion of food transfers occurs between relatives, kin-based sharing likely explains the bulk of sharing in numerous species.

Reciprocity

Reciprocity occurs when an individual pays a short-term cost of cooperation for the future benefit of a social partner's reciprocated cooperation. As a form of altruistic cooperation, reciprocity aims to explain cooperative behavior among genetically unrelated individuals (thereby eliminating kin selection as an explanation) often using a unique type of social interaction termed the prisoner's dilemma. The keys aspects of the prisoner's dilemma are (1) cooperation maximizes the total payoff to everyone involved in the interaction (mutual cooperation provides more benefits than mutual defection); however, (2) any individual will receive a higher personal payoff by defecting, so a sizable temptation to cheat exists. Pursuing unilateral cooperation in this game is not a stable strategy.

Theoreticians suggested that reversing roles as donor and recipient of altruism may reduce the temptation to defect because it largely commits individuals to invest in future cooperation. When the fitness payoffs sum over a series of interactions with the same partner, reciprocal strategists can reap the benefits of mutual cooperation. The reciprocal strategy tit-for-tat, in which a player starts out cooperating and copies its opponent's behavior in previous interactions, can successfully dominate simulated populations of social partners engaging in prisoner's dilemma games, winning out over many alternative behavioral strategies. If the probability of interacting again exceeds a critical level, a reciprocal strategy can maintain cooperation.

Many authors have reported examples of reciprocity in animals, including primates, spanning such contexts as food sharing, grooming, and coalitions associated with mating opportunities. Unfortunately, in most cases these purported examples of reciprocity have either never been replicated or have been challenged by alternative explanations for the patterns of interaction, including especially kin selection and mutualism. Another alternative explanation, symmetry-based reciprocity, occurs when the reciprocal pattern can be explained by the symmetrical relationship between individuals. For instance, if individuals A and B interact frequently and cooperate, their behavior will appear to be reciprocal even though it is based on symmetrical interactions rather than contingent

cooperation. This contrasts with calculated reciprocity (such as tit-for-tat) in which individuals track debts owed and favors given when deciding whether to cooperate.

A field study of olive baboons (*Papio anubis*) provided one of the first reports of reciprocity. Males formed coalitions in order to drive off rival males and gain access to reproductively active females. Data suggested that males took turns reaping the benefits, implying evidence of reciprocity. The robustness of this finding, however, is unclear, given that a subsequent study of a different population of baboons did not find the same reciprocal patterns, and a study of a closely related species proposed an alternative, mutualistic explanation.

Another example comes from experimental tests of food sharing in captive capuchin monkeys. When pairs of capuchins were placed in adjacent chambers separated by a mesh partition, if individual A allowed individual B to consume dropped scraps of food in one particular trial, then B tended to allow A to gather scraps in the next trial. Because the food transfer is primarily passive and costs may be minimal, these results can be explained by a mutual tolerance between individuals. Therefore, the capuchins do not appear to be assessing the costs and benefits of their actions, making calculated reciprocity an unlikely explanation of capuchin food transfers.

Finally, in a series of experiments with captive cotton-top tamarins, subjects could pull a tool to give food to an unrelated recipient without getting any food for self. Tamarins pulled the tool most often for partners that always pulled and infrequently for partners that never pulled. The frequency of cooperation was, however, less than 50%, and as each game progressed, the amount of food given dropped. Tamarins, therefore, maintained moderate levels of cooperation with other cooperators, but the degradation of cooperation over time suggests an unstable system; in fact, game theoretic modeling of the data suggest that reciprocity crashes if an opponent fails to cooperate on two consecutive opportunities.

In summary, there are few examples of true, calculated reciprocity among monkeys and apes. Most putative examples can be explained by symmetry-based reciprocity, mutualism, or kin selection. Given the simplicity of the preconditions for reciprocity and the apparent ubiquity of reciprocity in humans, its rarity in primates (and other animals) is surprising. We take up this issue in the section on cognitive constraints on reward processing.

Punishment

Punishing individuals that defect can potentially impose enough costs to offset the temptation to cheat, and, like reciprocity, can elicit future cooperation. Punishment involves energetic costs, and, when accomplished by means of aggression, also involves the costs of risked injury. Consequently, as an adaptive behavior, punishment is a selfish strategy, favoring actions that ultimately benefit the punisher by eliciting cooperative behavior from the recipient of punishment. As defined, punishment resembles reciprocity, which should occur when it elicits cooperative behavior directed strictly at the reciprocator. Despite the theoretical interest, punishment is not well documented in animals; in fact, though animals act aggressively toward others in the context of resource defense (e.g., territoriality, dominance interactions), there are few examples in the context of cooperation.

Whereas punishment penalizes past behavior with the hope of future reward, a similar behavior called harassment penalizes present behavior to elicit immediate reward. Harassment occurs when an individual imposes costs such as aggression on another to induce immediate cooperation, thereby providing instant benefits to the harasser. Harassment may also influence future as well as current cooperation, suggesting that it can lead to punishment strategies. For instance, if a defector is in the process of consuming an entire food resource, a harasser could impose costs on the defector aimed at obtaining some portion of the food resource immediately. Whereas punishment appears relatively rare among animals, harassment may be more common. For instance, after capturing prey, chimpanzees frequently allow other individuals to consume part of the meat. Observational work has demonstrated that harassment best accounts for the pattern of food sharing, taking the form of individuals incessantly begging for food with hand gestures, grabbing at the food, and placing their hands over the mouths of the food possessors. In this context, harassment was costly for the food owner (food intake rate decreased as the number of beggars increased), owners shared more often when beggars harassed frequently and intensely, and when sharing occurred harassment levels decreased. Controlled experiments corroborated these findings with captive chimpanzees, as well as squirrel monkeys (*Saimiri boliviensis*), a species that rarely cooperates.

Harassment may, of course, influence future as well as current cooperation, suggesting that it may lead to punishment strategies. One example comes from rhesus macaques. When individuals find food, they often give food calls, an altruistic act that recruits others to the food source, thereby potentially benefiting kin. However, individuals sometimes find food and suppress their calls. If no one detects these silent discoverers, they reap more food rewards than those who announce their discoveries. If, in contrast, they are detected, then they are aggressively attacked, leading to an even larger loss of food to the attacker. Thus,

aggression yields an immediate benefit of accessing food for the attacker, clearly qualifying it as a case of harassment. In addition, the aggression may have a punishing effect on silence by eliciting future food calls. That is, aggressive behavior influences both immediate and future benefits.

The Cognitive Psychology of a Cooperator

Many researchers have argued that the frequency of complex social interactions in primates underlies the evolution of their cognitive abilities. Negotiating a complicated network of allies and enemies, dominants and subordinates, and mates and competitors has created selective pressure for cognitive abilities such as transitive inference, theory of mind, and deception. Though this connection between sociality and cognition has been well recognized, the cognitive constraints on social interaction have often been ignored. Put simply, much of the work on primate cognition has focused on how large brains enable rich social interactions instead of how limited cognitive abilities may constrain the range of social interactions. Here, we take up this 'cognitive constraints' perspective.

Mutualism and kin selection are both theoretically well-understood and empirically well-documented models of cooperation. By contrast, reciprocity and punishment, while theoretically feasible, are rare among primates (although harassment may be more widespread). Therefore, despite models purporting the evolutionary stability of all of these types of cooperation, some types occur much more frequently than others. Unfortunately, a strictly adaptive perspective has limited power to explain the frequency of mutualism and kin-biased cooperation and the rarity of reciprocity and punishment. A proximate perspective that links with the ultimate problem can, however, reveal how cognitive constraints limit or facilitate particular forms of cooperation.

The proximate approach emphasizes critical aspects of reciprocity and punishment that differ markedly from mutualism and kin-biased cooperation. In both reciprocity and punishment, the fitness benefits associated with cooperation depend on the partner's behavior: cooperation should only occur when the partner responds by reciprocating or punishing. When this contingent response occurs in the future, the temporal delay introduces cognitive challenges that may constrain the emergence and stability of cooperation. Animals can easily implement strategies that yield immediate benefits, such as mutualism and harassment, because individuals do not have to track benefits over time. With a time delay between cooperating and receiving return benefits, however, individuals must invest in an uncertain future. Delayed benefits impede learning the consequences of cooperation, place greater demands on the capacity to recall previous interactions, and trade off short-term fitness gains for long-term gains. Here we highlight how several key cognitive constraints play a role in mutualism, kin-biased cooperation, reciprocity, and punishment, thereby uniting the two sections of this essay.

Cognitive Constraints on Mutualism and Kin-Biased Cooperation

Because no temptation to cheat exists in mutualistic interactions, individuals should always cooperate. As a result, mutualism requires no special cognitive abilities above and beyond the challenges inherent in the cooperative behavior itself. Kin-biased cooperation does, however, require additional cognitive capacities. At a minimum, it requires the capacity to direct cooperative actions to related individuals.

Mechanisms of kin recognition include recognition alleles, phenotype matching, and spatial and familiarity cues. The recognition allele hypothesis predicts that individuals can compare a particular phenotypic cue (auditory, olfactory, visual, etc.) to an innately specified template. Such a model requires few cognitive skills other than discriminating the cue associated with relatedness. The closely related phenomenon of phenotypic matching occurs when an individual compares a conspecific's phenotypic cues to a learned template. This requires specialized perceptual and computational systems that detect cues at an early stage to form a template, then test cues against the template to discriminate kin. Though recognition alleles and phenotypic matching provide the most direct forms of kin recognition and have been observed in some animal species, they do not appear to be common mechanisms of primate kin recognition. Instead, primates seem to use a simple set of rules such as spatial and familiarity cues to discriminate kin. Often primates may use rules such as 'be nice to individuals near your home' or 'help those that you grew up with' to direct the benefits of cooperation toward kin. These familiarity mechanisms require little in the way of cognitive abilities, building on general laws of learning by association, and seem to be widespread throughout the primates.

Cognitive Constraints on Reciprocity and Punishment

The original formulation of reciprocity had three requirements for evolutionary stability: (1) the reciprocated benefit must outweigh the immediate cost, (2) individuals must interact repeatedly, and (3) individuals must recognize each other. These requirements, however, most likely underestimate what is cognitively necessary for both developing

and maintaining a system of stable reciprocity. In particular, the delay between the cost of a cooperative act and the benefit of reciprocated cooperation introduces a number of cognitive challenges. Like reciprocity, punishment can involve a delay between a costly act and a beneficial payoff, and in these cases it faces similar constraints. For this reason, we consider the constraints on reciprocity and punishment together, pointing to differences where appropriate.

Individual recognition Like kin selection, reciprocity and punishment require directing cooperation to others. Unlike kin selection, however, where altruistic acts are disseminated simply as a function of coefficients of genetic relatedness, reciprocity and punishment require targeting specific individuals. Therefore, the delayed, contingent response required for both reciprocity and punishment necessitates that individuals can distinguish different partners. Mechanisms for individual recognition appear to be the norm among animals, including lemurs, monkeys, and apes. Consequently, the paucity of evidence for reciprocity and punishment across primates cannot appeal to individual recognition as a constraint.

Number, amount, and time Numerical abilities can play a key role in reciprocity when individuals must precisely quantify the reward amounts being reciprocated. When primates engage in a bout of reciprocity, they will either be limited to small numbers of objects in cases where the exchange must be precise (a banana for a banana), or they will be freed from this constraint where approximate exchanges are tolerated. The same prediction holds for cases where the currency is time, such as the duration of a grooming bout. If one monkey grooms another for 10 min, the groomer will most likely accept as fair exchange a reciprocated grooming bout of 8–12 min. As reward quantity and time magnitudes increase, quantification accuracy decreases, making equitable exchange more difficult and leaving opportunities for cheaters to exploit the judgment errors. This constraint only applies when variance exists in the quantity and time magnitudes. If the exchange is always a banana for a banana, these quantification abilities are superfluous.

Inhibitory control and impulsivity Reciprocity is itself an inhibitory and impulsivity problem: can an individual inhibit the choice of the immediate large benefit to gain the delayed larger benefit? Like the reversed-contingency task, the chooser must first reach for the undesirable over the desirable reward. The first move can be likened to giving away food. By avoiding the temptation of the larger immediate reward, the first move is costly to self. As data on the reversed-contingency task in primates show, the cost often appears too great and the inhibitory system too weak.

The second move, waiting for the reciprocated benefit, is analogous to the intertemporal choices described earlier. Individuals must choose between the immediate reward of defecting and the long-term reward of cooperating. Indeed, a number of researchers have predicted that impulsivity can reduce the value of reciprocated benefits. Experimental data on variation in human impulsivity and cooperation validate the view that a preference for immediate rewards may inhibit reciprocity. Impulsivity correlates with cooperation such that individuals who prefer immediate gains also cooperate less frequently. In parallel, blue jays (*Cyanocitta cristata*) show stable cooperation in a prisoner's dilemma, only following an experimental reduction in their preference for immediacy. Although primates have not been tested in both impulsivity and cooperation tasks, it seems highly probable that a significant constraint on their capacity for reciprocity emerges from their steep discounting functions. Similar abilities are required for punishment: punishers must pay an immediate cost of punishing to achieve a future, discounted reward.

Memory Episodic memory may very well be required to implement reciprocity. Because individuals must track the actions ('what') of previous partners ('who') the last time that they interacted ('when'), fairly sophisticated memory systems may be required for reciprocity. Even if episodic memory itself is not required, limitations in memory decay, interference, and capacity can also constrain the frequency of reciprocity and punishment. Memory decay proceeds rapidly over time; therefore, longer time intervals between cooperative acts may make reciprocity and punishment more difficult. Even with short time delays between cooperative interactions and few distractions, every potential new partner increases the computational load of tracking debts owed, favors given, and costs imposed. Keeping score of reciprocal obligations and punishment with the extensive social network found in many primate species may place a computationally intensive burden on memory systems. Although few studies examine learning and memory constraints in primate cooperation, studies of human cooperation suggest that these constraints can pose challenges for maintaining stable cooperative relationships.

Theory of mind Understanding the knowledge of others is critical to strategic reasoning. 'I know that you know that I know that you know…' is a central assumption of economic decision making required

for rational agents in economic games. If, however, natural selection acts as the 'rational agent' by selecting for individuals that make good decisions, this capacity may not be required for individuals if they use simple rules that approximate the outcome of fully rational choice. Therefore, it is possible for primates to implement reciprocity without full theory of mind capacities. At present, only a few studies have attempted to integrate issues concerning mental state attribution with those involved in cooperation. For example, recent demonstrations of visual perspective-taking abilities in primates suggests that they might be able to attend to rewards that others seek, possibly aiding their ability to recognize cooperative situations. However, despite this evidence, recent work shows that chimpanzees do not demonstrate other-regarding preferences; that is, they do not preferentially acquire food rewards in such a way as to give food to others. Instead, they appear to be completely indifferent to how their own reward-based actions can benefit other individuals. In contrast, in the study of tamarins mentioned above, individuals were more likely to cooperate with individuals who intentionally gave food as an altruistic act as opposed to individuals who gave the same amount of food as an accidental by-product of an otherwise selfish act. Further study of what monkeys and apes understand of the goals, beliefs, and intentions of others will elucidate whether limitations in these areas constrain other-regarding preferences.

Conclusions

With increasingly sophisticated and ecologically relevant tasks, we are closing in on those aspects of reward processing and social cognition that have uniquely evolved in humans and those that evolved by common descent with other primates and, possibly, other nonprimate animals. We suggest that reward processing and social cognition are intimately linked, setting up cognitive constraints on the evolution of cooperative behavior. We do not propose that these reward-processing constraints are immutable. If the evolutionary benefit of cooperation is great, selection pressure may circumvent these constraints. This, however, is likely to occur in a domain-specific manner rather than a domain-general manner. For instance, food-caching species such as squirrels may suspend inhibitory and impulsivity problems when hiding nuts away for winter. Yet, this ability has evolved to solve a very specific problem and likely does not apply generally to the species' decision-making strategies. Therefore, domain-general cognitive systems must evolve to allow many of the more complex forms of cooperation. The possibly unique presence of reciprocal cooperation and punishment in humans may be directly attributable to a combination of our keen number sense, patience, inhibitory control, memory capacity, and theory of mind.

See also: Connectivity of Primate Reward Centers; Episodic Memory; Games in Monkeys: Neurophysiology and Motor Decision-Making; Goal-Directed Behavior Theories; Neuropsychology of Primate Reward Processes; Prefrontal Contributions to Reward Encoding; Primate Communication: Evolution; Representation of Reward; Reward and Learning; Reward Neurophysiology and Primate Cerebral Cortex; Social Brain: Evolution.

Further Reading

Axelrod R and Hamilton WD (1981) The evolution of cooperation. *Science* 211: 1390–1396.

Boesch C and Boesch H (1989) Hunting behavior of wild chimpanzees in the Tai National Park. *American Journal of Physical Anthropology* 78: 547–573.

Boysen ST and Berntson GG (1995) Responses to quantity: Perceptual versus cognitive mechanisms in chimpanzees (*Pan troglodytes*). *Journal of Experimental Psychology: Animal Behavior Processes* 21: 82–86.

Chapais B and Berman CM (2004) *Kinship and Behavior in Primates*. Oxford: Oxford University Press.

Hampton RR and Schwartz BL (2004) Episodic memory in nonhumans: What, and where, is when? *Current Opinion in Neurobiology* 14: 192–197.

Hare B, Call J, Agnetta B, and Tomasello M (2000) Chimpanzees know what conspecifics do and do not see. *Animal Behaviour* 59: 771–785.

Hauser MD, Chen MK, Chen F, and Chuang E (2003) Give unto others: Genetically unrelated cotton-top tamarin monkeys preferentially give food to those who altruistically give food back. *Proceedings of the Royal Society of London, Series B* 270: 2363–2370.

Hauser MD and Spelke ES (2004) Evolutionary and developmental foundations of human knowledge: A case study of mathematics. In: Gazzaniga M (ed.) *The Cognitive Neurosciences III*, pp. 853–864. Cambridge, MA: MIT Press.

Kacelnik A (2003) The evolution of patience. In: Loewenstein G, Read D, and Baumeister RF (eds.) *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, pp. 115–138. New York: Russell Sage Foundation.

Meck WH (2003) *Functional and Neural Mechanisms of Interval Timing*. Boca Raton, FL: CRC Press.

Stevens JR (2004) The selfish nature of generosity: Harassment and food sharing in primates. *Proceedings of the Royal Society of London, Series B* 271: 451–456.

Stevens JR, Hallinan EV, and Hauser MD (2005) The ecology and evolution of patience in two New World monkeys. *Biology Letters* 1: 223–226.

Tomasello M and Call J (1997) *Primate Cognition*. Oxford: Oxford University Press.

Trivers RL (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology* 46: 35–57.

Social Stress in Adult Primates

R Sapolsky, Stanford University, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

We westernized humans and our health are peculiar by the standards of most other species. With the exception of AIDS and tuberculosis, we are plagued by relatively few infectious diseases – we are no longer haunted by the specter of smallpox, malaria, or a minor cut gone septic. Moreover, we are mostly spared the diseases of poor hygiene and inadequate food. Instead, we have the remarkable luxury of living well enough and long enough to succumb to diseases of slow accumulation of damage. These include cardiovascular disease, cerebrovascular disease, diabetes, and metabolic syndrome.

A critical point is the recognition that many of the most prevalent diseases of westernized life can be significantly worsened by stress. A stressor can be considered to be any perturbation of physiological or psychological homeostasis, and such challenges trigger an array of endocrine adaptations. This 'stress response' is vital for surviving a short-term physical stressor, in that it mobilizes energy from storage sites to exercising muscle, increases cardiovascular tone, enhances immunity, suppresses inessentials such as growth and reproduction, and enhances aspects of cognition. However, if the stress response is mobilized chronically for reasons of psychosocial stress, these same adaptations can become deleterious, increasing the likelihood and severity of an array of maladies typical of westernized life.

Virtually any organism would activate the stress response following, say, breaking a limb. Yet, in the face of similar psychosocial circumstances, only a subset of individuals activates the stress response. As such, differences in personality and temperament are important for understanding individual differences in vulnerability to stress-related disease. Consider, for example, three disorders which can be framed as individuals perceiving stressors that others typically do not, namely general anxiety disorders, major depression, and Type A personality (now thought of as a style of responding to social interactions with a globalized hostility). Critically, all three carry an increased risk of cardiovascular disease.

As such, stress-related disease cannot be understood outside the context of the individual experiencing the disease. But an understanding of the role of stress in health requires expanding this view even further – disease also cannot be understood outside of the context of the society in which an individual dwells. This article focuses on the fact that the frequency and severity of stressors are not equally distributed among all the members of a society, and that this is pertinent to understanding patterns of health and disease.

The Socioeconomic Health Gradient in Humans

A clear demonstration of this pertinence in humans is the 'socioeconomic gradient' of health, where, throughout westernized societies, lower socioeconomic status (SES) is associated with a higher incidence of cardiovascular, respiratory, rheumatoid, and psychiatric diseases, lower birth weight, and higher infant mortality and mortality from all causes. Extensive research has revealed some of the key features of this SES/health relationship:

1. The vast majority of the SES/health relationship reflects the influence of SES on health, rather than the other way around.
2. This relationship is not a step function where health is roughly equal until reaching the poverty level; instead, it is a gradient where, with every step down the SES ladder from its highest stratum, health declines.
3. Differential health care access cannot explain the gradient, as it occurs in countries with universal health care, in instances where the poor, in fact, make more, rather than less, use of the health care system, and in diseases whose prevalencies are independent of health care access (such as juvenile diabetes).
4. SES differences in health risk factors account for only a minority of the variance in this relationship. Instead, a complex and subtle literature demonstrates that the bulk of the SES/health gradient arises from the psychosocial correlates of SES. As one example of these findings, while objective SES is a good predictor of many realms of health, subjective SES is at least as good; this makes the point that the SES/health relationship is not so much about being poor as it is about feeling poor.

Gleaning these insights has represented a triumph of public health, health psychology, and behavioral medicine research, as these are extraordinarily difficult issues to study in humans. As a result, many researchers have utilized nonhuman subjects to study

health/social status relationships. While SES is a human invention, hierarchical systems of status differences are common among many social species.

Work stretching back half a century has demonstrated a clear relationship between health and social status among rodents, namely that it is socially subordinate individuals who have the most elevated basal levels of stress hormones (such as glucocorticoids), the highest resting blood pressure, and the highest incidences of stress-related diseases. This seems quite logical, in that the cornerstones of psychological stress are lack of control and predictability, lack of outlets for the frustration caused by external challenges, and lack of social affiliation. These factors are all attributes of social subordination in a rodent (e.g., lack of control and predictability, in the form of being subject to random acts of displacement aggression at the teeth of more dominant individuals; lack of outlets, in the form of being unable to subsequently displace aggression onto another animal; lack of social affiliation, in the form of less access to social contact and grooming).

Prompted by these clear findings in rodents, investigators began to consider health/social status relations in nonhuman primates, with the obvious recognition that other social primates are far better models for understanding humans than are rodents. And here, the picture has been anything but clear, with vastly contradictory answers to the most basic question, namely whether it is dominant or subordinate individuals who show the greatest incidences of stress-related disease.

The resolution to this confusion rests on two key points:

1. There is enormous variability in the social systems of primates. Across the more than 150 different primate species, dominance systems vary dramatically as to how high rank is attained, maintained, and manifested, and as to how social subordination is experienced.
2. As a result of this, there is tremendous variation among different primate species as to whether it is high- or low-ranking individuals who experience the most psychological stress.

Sources of Variability in Health/Social Status Relations in Primates

What factors determine whether it is high or low rank that is most associated with psychological stress? A number have been identified:

- *Resource inequities*. Some primate dominance systems are marked by resources (e.g., food) being divided very unequally as a function of rank. In such cases, it is subordinate individuals who show the greatest pathophysiologic indices of stress. In contrast, this is not seen in more egalitarian dominance systems.
- *Maintenance of dominance*. While social rank is inherited and unchanging in some primate social systems (e.g., female rhesus monkeys), it changes over time in other primate systems. In some species (e.g., females of ring-tailed lemurs), dominance must be reasserted frequently, with high-ranking individuals having the highest rates of fighting and injury. In such cases, it is dominant individuals with the most pronounced pathophysiologic indices of stress. In contrast, the opposite is seen in species (such as the males among baboons, rhesus, and squirrel monkeys) where dominance in mostly maintained through intimidation, and where dominant individuals have among the lowest rates of fighting and injury.
- *Mating system*. Among some Old World primates, high-ranking females monopolize breeding access by way of harassing the mating efforts of subordinates. In such cases, social subordinate is associated with stress-related pathophysiology (including higher rates of reproductive impairments, a phenomenon called 'social contraception'). In contrast, in many New World primates (such as marmosets), while the alpha female is the only one to breed, subordinates are typically younger sisters, who help with child rearing and are not subject to harassment. In such cases, social subordination is not preferentially linked to stress pathophysiology.
- *Stability of social rank*. In a social system involving resource inequity, when the hierarchy is stable (i.e., where the bulk of dominance interactions reaffirm the status quo, rather than challenge it), it is social subordination which is most linked to stress-related disease. When hierarchies are unstable, the instability is most manifest in the upper echelon of ranks (where there are the highest rates of dominance interactions, including aggressive ones, the greatest fluctuations in outcome, and the most disruption of socially affiliative behaviors such as grooming), as are the stress-related pathologies. In feral populations, such instability is a rare event (secondary, perhaps, to the death of a key individual). In captive populations, it is characteristic of the first months after group formation, when dominance relationships are still being established.
- *Subordinate coping strategies*. Independent of the extent to which resources are distributed unequally, or of the frequency with which low-ranking individuals are subject to subordinating interactions with high-ranking animals, social systems differ as to what sorts of coping outlets are available to

subordinates (e.g., grooming with another individual, mating, displacing aggression onto a third party). The more such outlets exist for low-ranking individuals, the less subordination is associated with stress-related pathology.

- *Subordinate avoidance of dominant individuals.* The frequency with which a low-ranking individual is subject to subordinating interactions with a high-ranking one will increase, in part, with increased proximity between the two. This helps explain rank/health differences within the same species, depending on whether subjects are studied in captivity or in the wild. In the former, subordinate animals are less capable of avoiding dominant ones, and show greater indices of stress-related disease.
- *Alternative strategies to overt dominance competition.* It is coming to be recognized that among some primates, low-ranking individuals can pursue alternatives to overt (and unsuccessful) competition within the hierarchy. This can take the form, for example, of low-ranking animals foraging in a more solitary manner (e.g., female chimps). Fewer indices of stress-related disease are observed among subordinate male baboons who forgo overt male–male competition and, instead, increase reproductive success through covert social affiliation with females. The same is seen among subordinate male orangutans who forgo overt competition and, instead, increase reproductive success through forced mating with peripheral females.

Thus, a variety of factors help predict whether it will be dominant or subordinate individuals who show the most pathophysiological indices of stress in a particular primate species. These relationships can potentially be blunted or exacerbated by personality. For example, among male baboons, high-ranking individuals who react to nonthreatening interactions with rivals as if they are overt challenges have higher risks of stress-related disease than the high-ranking cohort in general. In contrast, low-ranking male baboons who are particularly adept at taking advantage of social support have a lower risk than their low-ranking cohort in general.

As emphasized, there is tremendous variability in social systems among different primate species. In addition, there is considerable variability within species. Different populations and social groups of the same species will differ in rates and intensities of dominance interactions, of aggression, of affiliative behaviors, of sexual behavior, and so on. Such variability can reflect factors external to the group, such as features of the ecosystem (e.g., extent of rainfall), density of predators, or frequency of flow of individuals to and from neighboring groups, as well as factors internal to the group, such as adult sex ratios or the temperament of the alpha individual.

The Adverse Health Consequences of a Stressful Social Rank in a Primate

The previous section reviews the factors that help predict whether it is high- or low-ranking individuals who show the most stress-related pathologies in a particular primate species and social group. What are those pathologies? A number have been identified.

Hypersecretion of Glucocorticoids

Glucocorticoid hormones (such as cortisol and corticosterone) are secreted in response to stress, and are essential for short-term adaptation to physical stressors. In contrast, excessive exposure to glucocorticoids in response to prolonged psychosocial stressors can have a number of adverse consequences for health, including insulin resistance, hypertension, and impaired growth, tissue repair, reproduction, and immunity. Numerous studies have shown there to be elevated glucocorticoid levels (basally, or during the recovery period from a stressor) in animals of the more stressed rank in various primate species. This includes dominant female ring-tailed lemurs (where dominance must be constantly reasserted) and dominant male baboons (during periods of rank instability), subordinate male baboons, squirrel and rhesus monkeys (in stable hierarchies), and male rhesus and female talapoins (in small captive enclosures).

Cardiovascular Pathologies

Studies have shown the social rank associated with more social stress to also be associated with a greater risk of cardiovascular pathologies, including basal hypertension, a sluggish activation of heart rate and blood pressure with challenge and a sluggish recovery thereafter, elevated low-density lipoprotein (LDL) and decreased high-density lipoprotein (HDL) cholesterol, and increased vulnerability to atherogenesis. These pathologies are more prevalent among subordinate macaques and baboons (in stable hierarchies) and dominant individuals of the same species (in unstable hierarchies).

Reproductive Impairments

Prolonged stress suppresses various facets of reproductive physiology in females, and a more stressful social rank in primates is associated with lower gonadal hormone levels and a greater likelihood of anovulation. This is observed among subordinate cynomolgus macaque monkeys (in stable hierarchies). Among males, there is (to some) a surprising lack of

relationship between the stressfulness associated with particular social ranks and basal testosterone levels.

Immune Function

While transient stress is immunostimulatory, prolonged stress is immunosuppressive, and the more stressful rank position is associated with suppressed circulating lymphocyte counts and a sluggish immune response to challenge. This can be dominant individuals (male captive chimpanzees in an unstable hierarchy) or subordinate ones (male baboons in a stable hierarchy).

Neurobiological Endpoints

Sustained stress can cause a variety of adverse consequences in the realm of neurobiological endpoints. Those associated with the more stressful social rank in primates include atrophy of dendritic processes in hippocampal neurons and inhibition of hippocampal neurogenesis. This has been observed in subordinate tree shrews (in stable, captive populations).

Conclusions

Studies such as these generate a number of broad conclusions. The first reflects the astonishing variety and flexibility of primate sociality. There is no such thing as 'a' typical primate or primate social group, a pleasing finding for primatologists and a cautionary one for those who might too readily overemphasize human universals. Intrinsic in that is a cautionary note for behavioral neurobiologists who would study a nonhuman species as 'the' model for some aspect of human behavior. This variety brings home a truism, namely that ethology, or, in this case, neuroethology, is the process of interviewing an animal, but in its own language.

Next, you cannot understand disease outside the context of the individual suffering from that disease, and relevant variables include the individual's social status, personality, and patterns of sociality. Finally, those same variables also must be understood in the larger context of the society in which they occur – that is, what such social status, personality, and patterns of sociality mean in that society.

See also: Mother–Infant Interaction in the Variable Foraging Demand Model; Stress and Neural Involvement in Metabolism; Stress Response: Sex Differences; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Abbott D, Keverne E, Bercovith F, et al. (2003) Are subordinates always stressed? A comparative analysis of rank differences in cortisol levels among primates. *Hormones and Behavior* 43: 67–82.

Capitano J, Mendoza S, Lerche N, et al. (1998) Social stress results in altered glucocorticoid regulation and shorter survival in SIV syndrome. *Proceedings of the National Academy of Sciences of the United States of America* 95: 4714–4719.

Cavigelli S (1999) Behavioural patterns associated with faecal cortisol levels in free-ranging female ring-tailed lemurs, *Lemur catta*. *Animal Behavior* 57: 935–944.

Clarke A and Boinski S (1995) Temperament in nonhuman primates. *American Journal of Primatology* 37: 103–114.

Cohen S, Line S, Manuck S, et al. (1997) Chronic social stress, social status and susceptibility to upper respiratory infections in nonhuman primates. *Psychosomatic Medicine* 59: 213–221.

Creel S (2001) Social dominance and stress hormones. *Trends in Ecology and Evolution* 16: 491–496.

Fuchs E, Flugge G, Ohl F, et al. (2001) Psychosocial stress, glucocorticoids, and structural alterations in the tree shrew hippocampus. *Physiology and Behavior* 73: 285–291.

Gust D, Gordon T, Hambright K, et al. (1993) Relationship between social factors and pituitary-adrenocortical activity in female rhesus monkeys (*Macaca mulatta*). *Hormones and Behavior* 27: 318–331.

Kaplan J, Adams M, Anthony M, et al. (1995) Dominant social status and contraceptive hormone treatment inhibit atherogenesis in premenopausal monkeys. *Arteriosclerosis, Thrombosis, and Vascular Biology* 15: 2094–2100.

Kawachi I and Kennedy B (2002) *The Health of Nations: Why Inequality is Harmful to Your Health*. New York: The New Press.

Levine S, Wiener S, and Coe C (1989) The psychoneuroendocrinology of stress: A psychobiological perspective. In: Levine S and Brush F (eds.) *Psychoendocrinology*, pp. 341–377. New York: Academic Press.

Magarinos A, McEwen B, Flugge G, et al. (1996) Chronic psychosocial stress causes apical dendritic atrophy of hippocampal CA3 pyramidal neurons in subordinate tree shrews. *Journal of Neuroscience* 16: 3534–3540.

Maggioncalda A, Czekala N, and Sapolsky R (2002) Male orangutan subadulthood: A new twist in the relationship between chronic stress and developmental arrest. *American Journal of Physical Anthropology* 118: 25–32.

Manuck S, Marsland A, Kaplan J, et al. (1995) The pathogenicity of behavior and its neuroendocrine mediation: An example from coronary artery disease. *Psychosomatic Medicine* 57: 275–283.

Masataka N (1990) Dominance and immunity in chimpanzees *(Pan troglodytes)*. *Ethology* 85: 147–155.

McEwen B (2002) *The End of Stress*. New York: Joseph Henry Press.

Ray J and Sapolsky R (1992) Styles of male social behavior and their endocrine correlates among high-ranking baboons. *American Journal of Primatology* 28: 231–250.

Sapolsky R (2005) The influence of social hierarchy on primate health. *Science* 308: 648.

Sapolsky R and Share L (2004) A pacific culture among wild baboons, its emergence and transmission. *Public Library of Science* 2: E106–E113.

Sapolsky RM, Romero LM, and Munck AU (2000) How do glucocorticosteroids influence stress responses? Integrating permissive, suppressive, stimulatory and preparative actions. *Endocrine Review* 21: 55–89.

Shively C and Clarkson T (1994) Social status and coronary artery atherosclerosis in female monkeys. *Arteriosclerosis and Thrombosis* 14: 721–726.

Singh-Manoux A, Adler N, and Marmot MG (2003) Subjective social status: Its determinants and its association with measures

of ill-health in the Whitehall II study. *Social Science and Medicine* 56: 1321–1333.

Steklis H (l986) Biochemical and hormonal correlates of dominance and social behavior in all-male groups of squirrel monkeys (*Saimiri sciureus*). *American Journal of Primatology* 11: 133–140.

Van Schaik C (1991) A pilot study of the social correlates of levels of urinary cortisol, prolactin, and testosterone in wild long-tailed macaques (*Macaca fascicularis*). *Primates* 32: 345–351.

Wilkinson R (2000) *Mind the Gap: Hierarchies, Health and Human Evolution*. London: Weidenfeld and Nicolson.

Sodium Channels

W A Catterall, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Voltage-gated sodium channels initiate action potentials in neurons and other excitable cells, and they are responsible for propagation of action potentials along nerve and muscle fibers. They are complexes of a large pore-forming α subunit and smaller β subunits. Multiple genes encode sodium channel subunits, and the distinct sodium channel subtypes have subtle differences in functional properties, differential expression in excitable cells, and differential distribution in subcellular compartments. These differences in function and localization contribute to the specialized functional roles of sodium channels in neuronal physiology and pharmacology.

Sodium Channel Subunit Structure

Na^+ channel proteins in mammalian brain are complexes of a 260 kDa α subunit in association with one or two auxiliary β subunits (β_1–β_4) of 33–36 kDa. The primary sequence predicts that the sodium channel α subunit folds into four internally repeated domains (I–IV), each of which contains six α-helical transmembrane segments (S1–S6) (Figure 1). In each domain, the S1 through S4 segments serve as the voltage-sensing module, and the S5 and S6 segments and the reentrant P loop that folds into the transmembrane region of the channel between them serve as the pore-forming module. A large extracellular loop connects the S5 or S6 transmembrane segments to the P loop in each domain, whereas the other extracellular loops are small. Large intracellular loops link the four homologous domains, and the large N-terminal and C-terminal domains also contribute substantially to the mass of the internal face of sodium channels. This view of sodium channel architecture originally derived from hydrophobicity analysis has been largely confirmed by biochemical, electrophysiological, and structural experiments.

Sodium Channel Genes

Sodium channels are the founding members of the ion channel superfamily that includes voltage-gated Ca^{2+} channels; TRP channels; voltage-gated, inward rectifying, and two-pore domain K^+ channels; and cyclic nucleotide-regulated CNG and HCN channels. In evolution, the four-domain sodium channel was last among the voltage-gated ion channels to appear, and it is only found in multicellular organisms. It is thought that sodium channels evolved by two rounds of gene duplication from ancestral single-domain bacterial sodium channels such as NaChBac found in *Baccillus halodurans,* which may also be the ancestral single-domain channel from which Ca^{2+} channels arose. Voltage-gated sodium channel genes are present in a variety of metazoan species, including fly, leech, squid, and jellyfish. The biophysical properties, pharmacology, gene organization, and even intron splice sites of these invertebrate sodium channels are largely similar to those of the mammalian sodium channels, adding further support for the idea that the primordial sodium channel was established before the evolution of vertebrates.

The human, mouse, and rat sodium channel genes and proteins are the best characterized. Ten related sodium channel genes are found in these species (Figure 2), which map to four paralogous chromosome segments within a HOX gene cluster. More than 20 exons comprise each of the nine sodium channel α subunit genes in vertebrates. Genes encoding sodium channels $Na_v1.1$, $Na_v1.2$, $Na_v1.3$, and $Na_v1.7$ are localized on chromosome 2 in human and mouse, and these channels share similarities in sequence, biophysical characteristics, block by nanomolar concentrations of tetrodotoxin, and broad expression in neurons. A second cluster of genes encoding $Na_v1.5$, $Na_v1.8$, and $Na_v1.9$ channels is localized to human chromosome 3p21–24 (and to the orthologous region of chromosome 3 in mouse). Although they are more than 75% identical in sequence to the group of channels on chromosome 2, these sodium channels all contain amino acid substitutions that confer varying degrees of resistance to the pore blocker tetrodotoxin. In $Na_v1.5$, the principal cardiac isoform, a single amino acid change from phenylalanine to cysteine in the pore region of domain I is responsible for a 200-fold reduction in TTX sensitivity compared to those channels on chromosome 2. At the identical position in $Na_v1.8$ and $Na_v1.9$, the amino acid residue is serine, and this change results in even greater resistance to TTX. These two channels are preferentially expressed in peripheral sensory neurons. In comparison to the sodium channels on chromosomes 2 and 3, $Na_v1.4$, which is expressed in skeletal muscle, and $Na_v1.6$, which is highly abundant in the central nervous system (CNS), have greater than 85% sequence identity and similar functional properties, including TTX sensitivity in the nanomolar concentration range. Despite

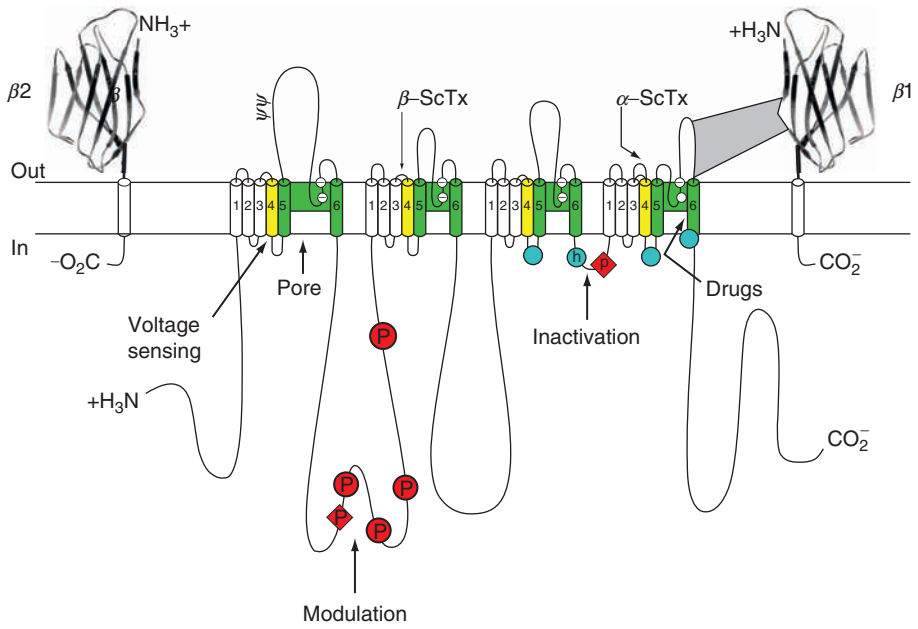

Figure 1 Transmembrane organization of sodium channel subunits. The primary structures of the subunits of the voltage-gated ion channels are illustrated as transmembrane folding diagrams. Cylinders represent probable α-helical segments. Bold lines represent the polypeptide chains of each subunit with length approximately proportional to the number of amino acid residues in the brain sodium channel subtypes. The extracellular domains of the β_1 and β_2 subunits are shown as immunoglobulin-like folds. ψ, sites of probable N-linked glycosylation; P, sites of demonstrated protein phosphorylation by PKA (circles) and PKC (diamonds); shaded, pore-lining S5–P–S6 segments; white circles, the outer (EEDD) and inner (DEKA) rings of amino residues that form the ion selectivity filter and the tetrodotoxin binding site; ++, S4 voltage sensors; h, inactivation particle in the inactivation gate loop; open shaded circles, sites implicated in forming the inactivation gate receptor. Sites of binding of α- and β-scorpion toxins and a site of interaction between α and β_1 subunits are also shown.

these similarities, phylogenetic analysis by parsimony comparisons suggests a more distant evolutionary relationship, consistent with their distinct chromosomal localizations (Figure 2). $Na_v1.4$ and $Na_v1.6$ are located on human chromosome 11 (mouse chromosome 17) and human chromosome 15 (mouse chromosome 12), respectively.

A tenth sodium channel, Na_x, whose gene is located near the sodium channels on chromosome 2, is evolutionarily more distant (Figure 2). Key differences in functionally important regions of the voltage sensor and inactivation gate and lack of functional expression of voltage-gated sodium currents in heterologous cells suggest that Na_x may not function as a voltage-dependent sodium channel. Consistent with this conclusion, targeted deletion of the Na_x gene in mice causes functional deficits in sensing plasma salt levels.

The $Na_v\beta$ subunits are encoded by four distinct genes in mammals. The gene encoding β_1 maps to chromosome 19q13, whereas β_2 and β_4 are located on chromosome 11q22–23 and β_3 is located nearby on chromosome 11q24. The β_1 and β_3 subunits associate noncovalently with the α subunits, whereas β_2 and β_4 are covalently linked by a disulfide bond. All four β subunits have a single transmembrane segment, a large N-terminal extracellular domain, and a small intracellular C-terminal domain (Figure 1). The $Na_v\beta$ subunits are dual-function proteins. Coexpression of $Na_v\beta$ subunits alters the kinetics and voltage dependence of activation and inactivation of sodium channel complexes. Moreover, their extracellular domain forms an immunoglobulin-like fold, similar to many cell adhesion molecules, and the β subunits are involved in interactions of sodium channels with proteins in the extracellular matrix, with other cell adhesion molecules, and with intracellular regulatory proteins such as protein kinases and phosphoprotein phosphatases. The importance of the functions of the $Na_v\beta$ subunits is underscored by the strong effects of deletion of the genes in mice encoding $Na_v\beta_1$ and $Na_v\beta_2$, which include impairments of myelination and axonal conduction, epileptic seizures, and premature death.

Expression and Localization of Sodium Channel Subtypes

The sodium channel subtypes are differentially expressed in tissues and differentially localized within

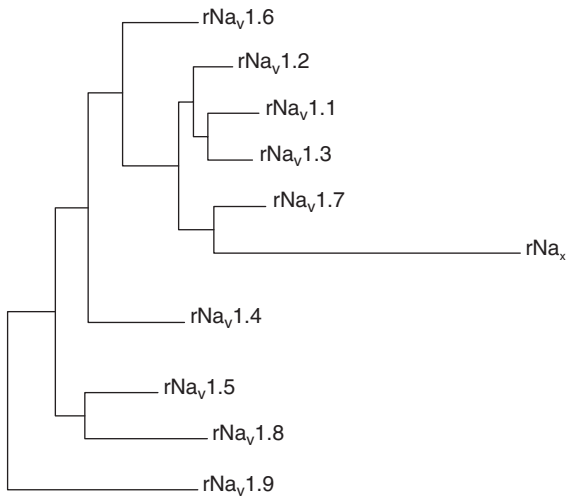

Figure 2 Amino acid sequence similarity and phylogenetic relationships of voltage-gated sodium channel α subunits. Phylogenetic relationships by maximum parsimony analysis of rat sodium channel sequences Na$_v$1.1–Na$_v$1.9 and Na$_x$. To perform the analysis, the amino acid sequences for all of the isoforms were aligned using ClustalW. The amino acid sequences in the alignments were then replaced with the published nucleotide sequences, and the nucleotide sequence alignments were subjected to analysis using the program PAUP*. Divergent portions of the terminal regions and the cytoplasmic loops between domains I–II and II–III were excluded from the PAUP* analysis. The tree was rooted by including the invertebrate sodium channel sequences during the generation of the tree, although these sequences are not shown in the figure.

individual cells. Na$_v$1.3 is highly expressed in fetal nervous tissues in rodents, whereas Na$_v$1.1, Na$_v$1.2, and Na$_v$1.6 are abundant in juvenile and adult CNS. Generally, Na$_v$1.1 and Na$_v$1.3 are primarily localized in neuronal cell bodies, where they may control neuronal excitability through integration of synaptic impulses to set the threshold for action potential generation and propagation to the dendritic and axonal compartments. Na$_v$1.2 is primarily localized in unmyelinated axons. During development, Na$_v$1.6 replaces Na$_v$1.2 in maturing nodes of Ranvier in myelinated axons. Na$_v$1.1, Na$_v$1.2, and Na$_v$1.6 are also expressed in the peripheral nervous system. However, the three isoforms that have been cloned from sympathetic and dorsal root ganglion neurons – Na$_v$1.7, Na$_v$1.8, and Na$_v$1.9 – are more abundant and are the principal sodium channels in the peripheral nervous system. The Na$_v$1.7 channel is localized in axons, where it functions in initiating and conducting the action potential. More restricted expression patterns are observed for Na$_v$1.8 and Na$_v$1.9; these channels are highly expressed in small sensory neurons of dorsal root ganglia and trigeminal ganglia, where they have a key role in nociception. Na$_v$1.4 and Na$_v$1.5 are the primary muscle sodium channels that control the excitability of the skeletal and cardiac myocytes, respectively. Na$_v$1.5 is transiently expressed in developing skeletal muscle but is replaced by Na$_v$1.4 in the adult.

Molecular Basis of Sodium Channel Function

Classical work by Hodgkin and Huxley defined the three key features of sodium channels: voltage-dependent activation, rapid inactivation, and selective ion conductance. Building on this foundation, structure–function studies employing molecular, biochemical, structural, and electrophysiological techniques have provided clear understanding of the molecular basis of sodium channel function. The narrow outer pore is formed by the reentrant P loops between transmembrane segments S5 and S6 of each domain (Figure 1). Mutation of a set of four residues in analogous positions in each domain (aspartate in domain I, glutamate in domain II, lysine in domain III, and alanine in domain IV; DEKA) to glutamates confers calcium selectivity, indicating that the side chains of these amino acid residues are likely to interact with sodium ions as they are conducted through the ion selectivity filter of the pore.

Similar to other voltage-gated ion channels, the voltage dependence of activation of sodium channels derives from outward movement of gating charges as a consequence of depolarization of the membrane and consequent reduction of membrane electric field. The S4 segments of each homologous domain serve as the primary voltage sensors for activation. They contain repeated motifs of a positively charged amino acid residue followed by two hydrophobic residues creating a transmembrane spiral of positive charges. Upon depolarization, outward movement and rotation of S4 is thought to initiate a conformational change that opens the sodium channel pore. This 'sliding helix' or 'helical screw' model is supported by strong evidence. For example, neutralization of the positively charged residues in S4 reduces the voltage dependence of gating. The outward and rotational gating movement of the S4 segment has been detected directly by reaction of substituted cysteine residues in S4 segments with extracellular sulfhydryl reagents following channel activation and by analysis of the movement of fluorescent probes incorporated into these substituted cysteine residues. A structural model of the voltage-sensing and pore-forming modules of sodium channels in the resting and activated states is illustrated in Figure 3.

Fast inactivation of the sodium channel is a critical process that occurs within milliseconds of channel opening. The generally accepted model of this process involves a conserved inactivation gate formed by the

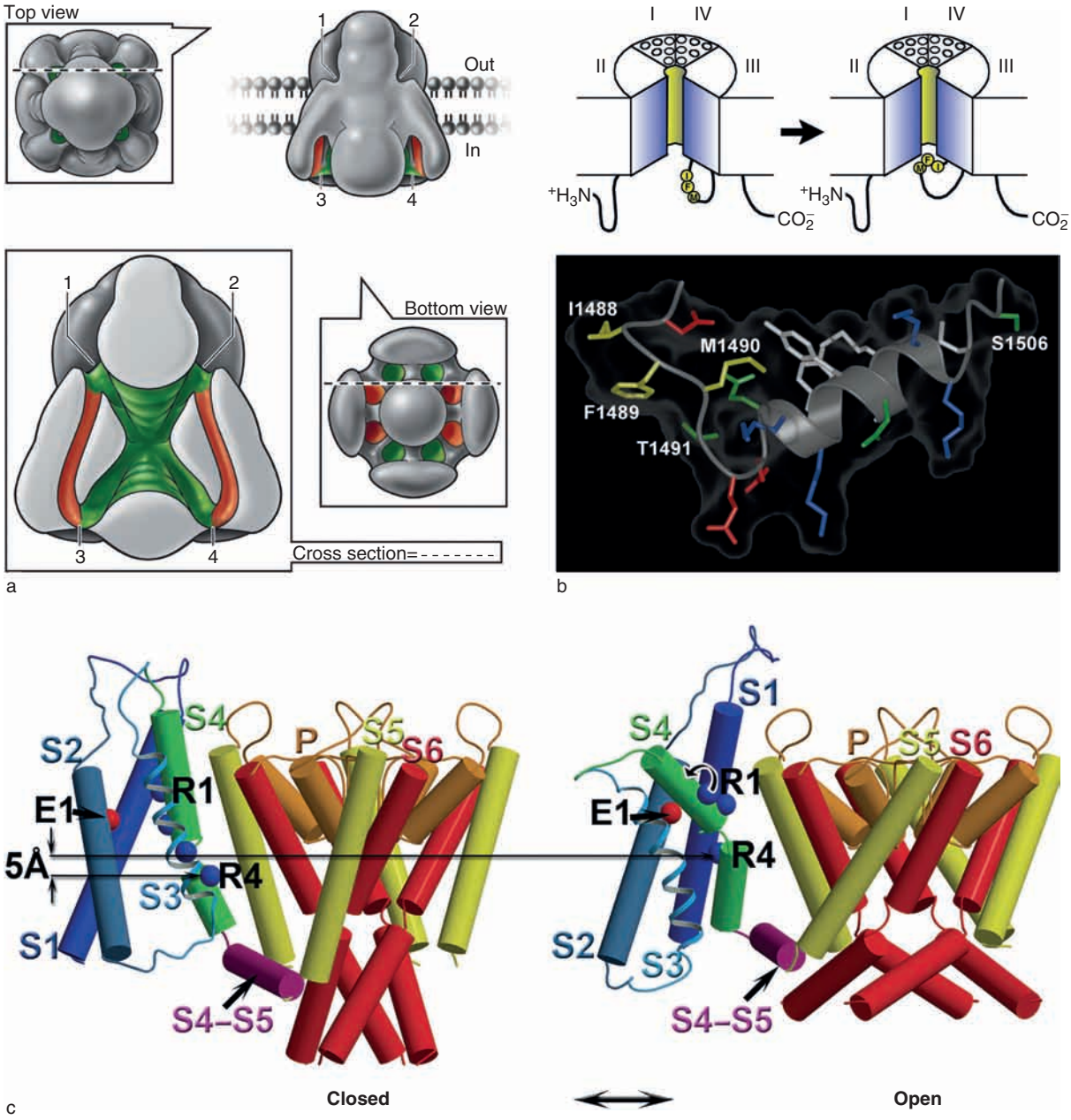

Figure 3 Three-dimensional structure of sodium channels and a related potassium channel. (a) The three-dimensional structure of the Na_v channel α subunit at 20 Å resolution, compiled from electron micrograph reconstructions. (b) (Top) The hinged-lid mechanism. The intracellular loop connecting domains III and IV of the sodium channel is depicted as forming a hinged lid with the critical phenylalanine (Phe1489) within the (IFM) motif shown occluding the mouth of the pore during the inactivation process. The circles represent the transmembrane helices. (Bottom) Three-dimensional structure of the central segment of the inactivation gate as determined by multidimensional NMR. Isoleucine 1488, Phe1489, and methionine 1490 IFM are illustrated in yellow. Threonine 1491, which is important for inactivation, and serine 1506, which is a site of phosphorylation and modulation by protein kinase C, are also indicated. (c) $K_v1.2$ models of the closed and open states shown in cylinder representation. Only a single voltage-sensing module is shown attached to the tetramer of the pore-forming module for clarity. Transmembrane segments S1–S6 and P loop are colored by rainbow scheme from blue to red. The S4–S5 linker is shown in purple. Approximate positions of the α carbon atoms of the first and fourth gating charge carrying arginines in the S4 segment (labeled R1 and R4 and colored blue) and E226 in the S2 segment (labeled E1 and colored red) are shown in sphere representation. All intracellular and extracellular loops, except for the S4–S5 linker, are represented by curved lines for simplicity. Vertical translation of R4 between the closed and open states along the membrane normal vector and relative to the plane of the membrane is indicated by arrows. S3 is represented by ribbon to clearly show the positions of the gating charge-carrying arginines in the S4 segment. (a) Adapted from Sato C, Ueno Y, Asai K, et al. (2001) The voltage-sensitive sodium channel is a bell-shaped molecule with several cavities. *Nature* 409: 1047–1051. (b) Adapted from Catterall WA (2000) From ionic currents to molecular mechanisms: The structure and function of voltage-gated sodium channels. *Neuron* 26: 13–25.

intracellular loop connecting domains III and IV, which serves as a hinged lid that binds to the intracellular end of the pore and blocks it (Figure 1). Intracellular perfusion of proteases or intracellular application of antibodies directed to this loop prevents fast inactivation. The latch of this fast inactivation gate is formed by three key hydrophobic residues, methionine 1490 (IFM), and an adjacent threonine (T). Mutations of these amino acid residues destabilize the inactivated state, and peptides harboring the IFM motif can restore inactivation to sodium channels having mutated inactivation gates. The closed inactivation gate is thought to make multiple interactions with hydrophobic amino acid residues near the intracellular mouth of the pore that may constitute the inactivation gate receptor. Scanning mutagenesis experiments implicate hydrophobic residues in intracellular S4–S5 loops of domains III and IV, as well as the intracellular end of the S6 transmembrane segment in domain IV, in forming the inactivation gate receptor.

Three-Dimensional Structure of Sodium Channels

The three-dimensional structure of the sodium channel is not completely known, but some limited structural information has emerged. Structural determination at 19 Å resolution by cryo-electron microscopy and image reconstruction techniques shows sodium channels as having a bell shape from a side view, a fourfold symmetry of transmembrane masses as viewed from above the membrane, and large intracellular and extracellular masses through which several inlets and outlets allow aqueous access (Figure 3(a)). Unexpectedly, this image of sodium channels reveals a central pore that does not directly connect the intracellular and extracellular spaces, instead splitting into four branches near the membrane surface. Moreover, it suggests the presence of four peripherally located narrow transmembrane pores, which may function as gating pores for voltage sensor movement as described later.

The only high-resolution structure of sodium channels is the solution structure of the inactivation gate in the intracellular loop connecting domains III and IV, as determined by nuclear magnetic resonance (NMR) analysis of this intracellular loop expressed in bacteria (Figure 3(b)). This structural analysis reveals a rigid α-helical structure, preceded by two turns that position the key hydrophobic sequence motif IFMT such that it can interact with and block the inner mouth of the pore.

The global fold of the sodium channel is predicted to be similar to that of voltage-gated potassium channels (Figure 3(c)). The structure of the open state of this channel has been determined by X-ray crystallography, and the structure of the closed state has been determined by structural modeling using the Rosetta method. This structural view shows the S5 and S6 region of the sodium channel arrayed in fourfold symmetry around the aqueous ion conduction pathway (Figure 3(c), red, tan, and yellow). The selectivity filter of the pore at its narrow extracellular end is formed by the P loops (Figure 3(c), tan), whereas the wider vestibule in the center of the pore is formed mostly by the S6 segments (Figure 3(c), red). A single voltage sensor module is illustrated for simplicity (Figure 3(c), blue and green). In the resting (closed) state of the channel, the S4 segment is in an inward position. In this sliding helix model, activation of the voltage sensor by depolarization causes outward and rotational movement of the S4 segment and complementary movement of the S1–S3 segments around it. These coupled molecular movements translocate the gating charges in the S4 segment from exposure to the intracellular milieu to exposure to the extracellular milieu, as required for voltage-dependent gating.

Comparison of the primary structures of the auxiliary β subunits to those of other proteins reveals a close structural relationship to the family of proteins that contain immunoglobulin-like folds, which include many cell adhesion molecules. The extracellular domains of these type I single membrane-spanning proteins are predicted to fold in a similar manner as myelin protein P0, whose immunoglobin-like fold is known to be formed by a sandwich of two β sheets held together by hydrophobic interactions (Figure 1). Myelin P0 protein is a cell adhesion molecule involved in tight wrapping of myelin sheets, and many related cell adhesion molecules with extracellular Ig folds and a single membrane-spanning segment are involved in cell–cell interactions among neurons and glia. As expected from their structure, $Na_v\beta$ subunits interact with extracellular matrix molecules, other cell adhesion molecules, and intracellular cytoskeletal proteins and signaling proteins. These interactions are thought to localize and stabilize sodium channels in specific subcellular compartments such as axon hillocks and nodes of Ranvier where they are present in high density and also to bring crucial signaling molecules to the sodium channel to regulate it.

Sodium Channel Pharmacology

Sodium channels are the molecular targets for drugs used in prevention of acute pain and in treatment of cardiac arrhythmias, epilepsy, and bipolar disorder. Sodium channel-blocking drugs are also in development for treatment of chronic pain. Local anesthetics bind to a specific receptor within the pore of sodium channels, formed by the S6 segments in domains I, III, and IV (Figure 4). Their binding blocks ion movement

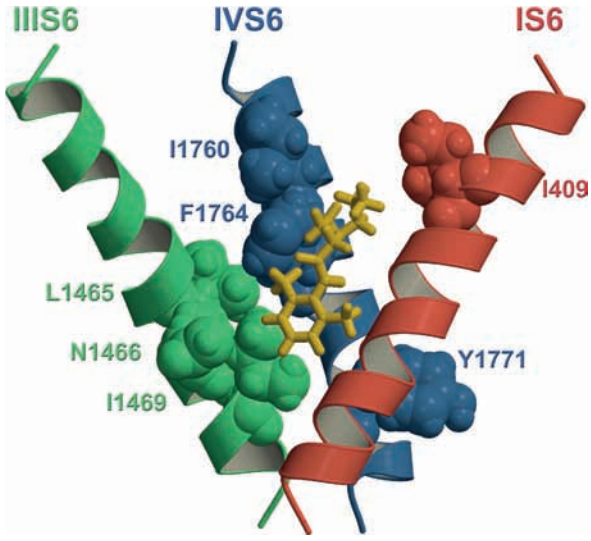

Figure 4 Model of etidocaine binding to the local anesthetic receptor site formed by transmembrane segments IS6, IIIS6, and IVS6 of the $Na_v1.2$ channel. Three-dimensional model of proposed orientation of amino acid residues within the Na^+ channel pore with respect to the local anesthetic etidocaine. Only transmembrane segments IS6 (red), IIIS6 (green), and IVS6 (blue) are shown. Residues important for etidocaine binding are shown in space-filling representation.

Table 1 Receptor sites on sodium channels

Receptor site	*Toxin or drug*	*Location (domain/segment)*
Neurotoxin receptor site 1	Tetrodotoxin	IP, IIP, IIIP, IVP
	Saxitoxin	IP, IIP, IIIP, IVP
	μ-Conotoxin	
Neurotoxin receptor site 2	Veratridine	IS6, IVS6
	Batrachotoxin	
	Grayanotoxin	
Neurotoxin receptor site 3	α-Scorpion toxins	IS5–IS6, IVS3–S4
	Sea anemone toxins	IVS3–S4
Neurotoxin receptor site 4	β-Scorpion toxins	IIS1–S2, IIS3–S4
Neurotoxin receptor site 5	Brevetoxins	IS6, IVS5
	Ciguatoxins	
Neurotoxin receptor site 6	δ-Conotoxins	IVS3–S4
Local anesthetic receptor site	Local anesthetic drugs	IS6, IIIS6, IVS6
Antiarrhythmic drugs		
Antiepileptic drugs		

through the pore and stabilizes the inactivated state of sodium channels. Antiarrhythmic drugs and antiepileptic drugs share similar, overlapping receptor sites. Complete block of sodium channels would be lethal. However, these drugs selectively block sodium channels in depolarized and/or rapidly firing cells, such as axons carrying high-intensity pain information and rapidly firing nerve and cardiac muscle cells that drive epileptic seizures or cardiac arrhythmias. This selective block arises because the drugs can reach their binding site in the pore of the sodium channel more rapidly when the pore is repetitively opened and they bind with high affinity to inactivated sodium channels that are generated in rapidly firing or depolarized cells. This use-dependent action of the sodium channel-blocking drugs is essential for their therapeutic efficacy.

In addition to these important pharmaceutical agents, sodium channels are the molecular targets for a large number of neurotoxins that paralyze prey by preventing neuromuscular function (Table 1). These toxins act at five or more distinct receptor sites and either block the pore of the channel or alter the kinetics or voltage dependence of gating. The pore blockers tetrodotoxin and saxitoxin bind to neurotoxin receptor site 1, which is formed by the P loops in the four domains. Alkaloids and other organic toxins from plants, frogs, and other species bind at neurotoxin receptor site 2 and cause persistent activation of sodium channels, which depolarizes nerves and muscles and causes repetitive firing and paralysis. Polypeptide toxins from scorpions and spiders bind at neurotoxin receptor sites 3 and 4 and either slow inactivation or enhance activation of sodium channels, which also causes repetitive firing and paralysis. Brevetoxins and ciguatoxins, which are large polyether compounds made by marine dinoflagellates, bind to neurotoxin receptor site 5 and cause persistent activation of sodium channels. The diversity of organisms that have developed toxins that target sodium channels is remarkable, highlighting the physiological importance of these channels in electrical signaling and neuromuscular function.

Sodium Channelopathies

Mutations that alter sodium channel function lead to human diseases of hyperexcitability (Table 2). Multiple mutations of $Na_v1.4$ that yield hyperactive skeletal sodium channels have been shown to cause hyperkalemic periodic paralysis and paramyotonia congenita. In contrast, mutations in S4 gating charges that cause reduced function of these channels and induce a sodium leak through their voltage sensor domains (termed 'gating pore current') cause hypokalemic periodic paralysis. Mutations of $Na_v1.5$ that impair inactivation or alter activation lead to

Table 2 Major inherited diseases of sodium channels

Disease	*Gene*	*Channel*
Generalized epilepsy with febrile seizures plus	SCN1A	$Na_v1.1$
Severe myoclonic epilepsy in infancy	SCN1A	$Na_v1.1$
Benign neonatal convulsions	SCN1A	$Na_v1.1$
Familial hemiplegic migraine type III	SCN1A	$Na_v1.1$
Hypokalemic periodic paralysis type II	SCN4A	$Na_v1.4$
K^+-sensitive normokalemic periodic paralysis	SCN4A	$Na_v1.4$
Hyperkalemic periodic paralysis	SCN4A	$Na_v1.4$
Paramyotonia congenita	SCN4A	$Na_v1.4$
Long QT syndrome type III	SCN5A	$Na_v1.5$
Brugada syndrome	SCN5A	$Na_v1.5$
Erythermalgia	SCN9A	$Na_v1.7$
Paroxysmal extreme pain disorder	SCN9A	$Na_v1.7$
Congenital indifference to pain	SCN9A	$Na_v1.7$

inherited long QT syndrome type 3 and Brugada syndrome, in which the risk of fatal cardiac arrhythmias is greatly increased. Generalized epilepsy with febrile seizure plus ($GEFS^+$) type II results from gain-of-function mutations of $Na_v1.1$ channels that enhance persistent inward sodium current, as well as from a mutation of the β_1 subunit. Gain-of-function mutations in $Na_V1.1$ channels also cause familial hemiplegic migraine type II. Severe myoclonic epilepsy in infancy is caused by inactivation of one allele of the gene encoding $Na_V1.1$ channels, and benign neonatal convulsions are caused by mutations in $Na_V1.2$ channels.

The requirement for intact sodium channel function in physiology is further highlighted by mouse knockout models, in which specific ablation of $Na_v1.1$, $Na_v1.2$, $Na_v1.5$, and $Na_v1.6$ led to premature lethality. In contrast, mutant mice lacking $Na_v1.8$ are viable and reveal defects in nociceptive response. These and other mouse models that target specific sodium channel genes should prove to be effective in delineating the significance of sodium channel function and its regulation in physiological contexts.

See also: Calcium Channels; Cyclic Nucleotide-Gated and Hyperpolarization-Activated Channels; Dorsal Root Ganglion Neurons; Epilepsy: Channelopathies; Inwardly Rectifying Potassium Channels; Ion Channel Localization in Cell Bodies and Dendrites; Neuromodulation of Sodium Channels; Transient Receptor Potential (TRP) Channels; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Armstrong CM (1981) Sodium channel and gating currents. *Physiological Reviews* 61: 644–682.

Cannon SC (2006) Pathomechanisms in channelopathies of skeletal muscle and brain. *Annual Review of Neuroscience* 29: 387–415.

Catterall WA (2000) From ionic currents to molecular mechanisms: The structure and function of voltage-gated sodium channels. *Neuron* 26: 13–25.

Catterall WA, Goldin AL, and Waxman SG (2005) International Union of Pharmacology: XLVII. Nomenclature and structure–function relationships of voltage-gated sodium channels. *Pharmacological Reviews* 57: 397–409.

Goldin AL (2002) Evolution of voltage-gated Na^+ channels. *Journal of Experimental Biology* 205: 575–584.

Hille B (2001) *Ionic Channels of Excitable Membranes*, 3rd edn. Sunderland, MA: Sinauer.

Hodgkin AL and Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* 117: 500–544.

Sato C, Ueno Y, Asai K, et al. (2001) The voltage-sensitive sodium channel is a bell-shaped molecule with several cavities. *Nature* 409: 1047–1051.

Tombola F, Pathak MM, and Isacoff EY (2006) How does voltage open an ion channel? *Annual Review of Cell and Developmental Biology* 22: 23–52.

Yarov-Yarovoy V, Baker D, and Catterall WA (2006) Voltage sensor conformations in the open and closed states in ROSETTA structural models of K^+ channels. *Proceedings of the National Academy of Sciences of the United States of America* 103: 7292–7297.

Relevant Website

http://www.pnas.org – *Proceedings of the National Academy of Sciences of the United States of America*; see Movie 1 and Movie 2.

Somatosensory Cortex

J H Kaas, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition,
© 2004, Elsevier B.V.

Somatosensory Cortex in Mammals with Little Neocortex

The organization of somatosensory cortex found in rats (Figure 1), is typical of that found in a wide range of mammalian species with relatively little phylogenetic expansion of neocortex. Much of somatosensory cortex in these mammals consists of two distinct systematic representations of the contralateral body surface, named in order of their discovery in cats, the first (primary) representation, or S-I, and the second representation, or S-II. Almost all neurons throughout both representations respond to gentle stimulation of the skin and body hairs. Thus, S-I and S-II appear to be related largely to cutaneous receptors. Neurons across the thickness of cortex at any location are all activated by stimulating the same region on the body surface. However, adjoining groups of neurons across the surface of the cortex are activated by different and typically nearby body regions so that all parts of the body eventually are represented in a pattern relative to the cortical surface. The larger S-I represents the body from tail to mouth in a mediolateral cortical sequence, while the smaller S-II has a head-to-tail mediolateral (or dorsoventral) cortical sequence. The two fields join along matched representations of the dorsal midline of the head. Thalamic input to both S-I and S-II is from a single representation of skin receptors in the ventroposterior nucleus. Axon pathways interconnect matched body parts of S-I and S-II of the same cerebral hemisphere and parts of S-I and S-II of the opposite cerebral hemisphere. S-I and S-II also have connections with other regions of cortex, including motor cortex.

Major features of organization of S-I in rats are common in S-I of other mammals:

1. A large proportion of S-I is devoted to the head. The representation of the head occupies half or more of S-I in rats, tree shrews, squirrels, opossums, rabbits, and several carnivores. Within this head representation, tree shrews and opossums emphasize the glabrous nose, rats emphasize the mystacial vibrissae, and squirrels and rabbits emphasize the lips. Raccoons and monkeys have enlarged representations of the hand. The disproportionate representation of body parts follows the general rule that those skin surfaces with the most receptors and the important tactile functions have the largest representations (cortical magnification).

2. The representation is largely continuous (somatotopic) so that skin surfaces that are adjacent are usually represented next to each other in cortex.

3. There are often partial disruptions, however, and the locations of these disruptions in the representation vary somewhat across species. For instance, a split representation of the dorsal and ventral trunk is an unusual feature found so far only in tree shrews. In primates, the representations of the anterior and posterior leg are commonly split in S-I to locations lateral and medial to the representation of the foot. Despite these disruptions, the representations of important sensory skin surfaces are typically continuous. The partial discontinuities seem forced by the requirements of disproportionately representing some skin surfaces, and by the need for representing the three-dimensional body surface within a two-dimensional cortical sheet.

4. Many somatotopic details of S-I in rats are widely found in S-I of other mammals. As in most other mammals, the head is represented with the dorsal surface posterior and the ventral surface anterior in cortex. The hand and foot are represented with the proximal surfaces posterior and the distal digit tips anterior. Digits 1–5 relate to lateromedial cortical sequences. The dorsal trunk is near the posterior border of S-I.

5. S-I is associated with cortex characterized by dense packing of small neurons in cortical layer IV (somatic koniocortex). In some mammals, discontinuities in the somatotopic map are associated with discontinuities in the dense packing layer IV cells. As a particularly striking anatomical specialization in mice and rats, the layer IV cells form ringlike aggregations (barrels) in the part of S-I representing vibrissae, with a barrel for each whisker.

The second somatosensory area, S-II, of rats also has an organization that is typical of that found in other mammals. S-II is oriented so that the top of the head joins the representation of the top of the head in S-I (a congruent border), and the distal limbs point ventrally away from S-I. The somatotopic organization of S-II is cruder than that of S-I in that receptive fields are considerably larger. However, the proportion of S-II devoted to different body parts roughly matches that of S-I. S-II is typically one-tenth to one-half the size of S-I. Rats also have a third representation of the body surface, the parietal ventral area,

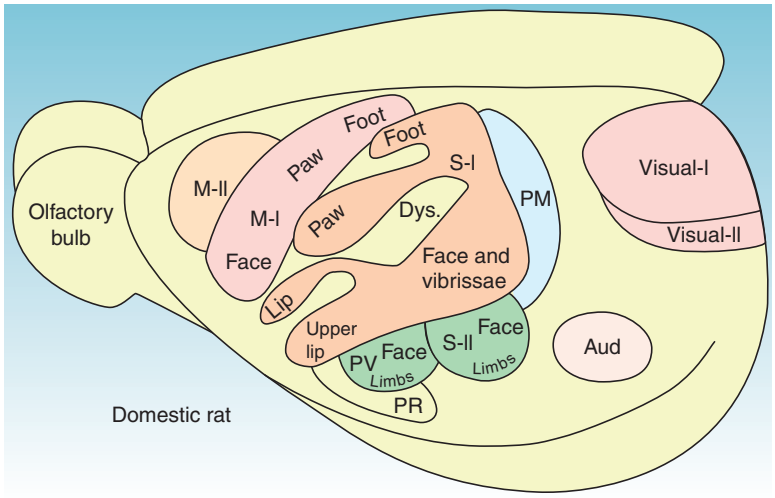

Figure 1 The locations of sensory and motor areas on a dorsolateral view of the brain of a rat. The traditional somatosensory areas are S-I, the primary area, and S-II, the secondary area. Each contains a detailed representation of cutaneous receptors, but only the representations of major body parts are shown here. The parietal ventral area, PV, contains a third systematic representation. The dysgranular cortex, Dys., invaginating rostral S-I and parietal-medial, PM, and parietal-rostral, PR, regions appear to be additional somatosensory fields. In addition, primary, M-I, and secondary, M-II, motor fields have connections with somatosensory fields. Auditory and primary and secondary visual fields are indicated for reference.

PV, first discovered in squirrels but now identified in a number of mammalian species. PV receives some inputs from S-I and some from S-II, thus having components of both parallel and serial processing.

Some regions of additional cortex in rats can be considered somatosensory. A narrow fringe of cortex between S-I and motor cortex responds to more intense stimulation of the body, and it is possible that this cortex relates to noncutaneous (deep) body receptors, such as those in muscles, but this is uncertain. A narrow zone of cortex posterior to S-I responds to stronger stimulation, while cortex posterior to this less responsive somatosensory zone is visual or auditory. Cortex just ventral to PV, a parietal rostral area or PR, has inputs from PV. PR may be a final processing station before outputs to limbic centers associated with the formation of long-term memories. Finally, rats have a primary motor area, M-I, and a premotor or supplementary motor area, M-II, which have mainly motor but also somatosensory functions. Thus rats have six or more cortical fields that are involved in processing somatosensory information. Cortical organization is similar in a number of other mammals, including cats, squirrels, tree shrews, and rabbits.

Prosimian Primates

The prosimian primates have less-developed brains than monkeys and other advanced primates, apes, and humans. Of prosimian primates, the organization of somatosensory cortex has been studied most extensively in galagos. Much of somatosensory cortex in galagos consists of S-I and S-II. S-I is organized much like S-I (area 3b) in monkeys (Figure 2), although the representation of the hand is smaller. As for other mammals, S-I in galagos responds to cutaneous stimuli throughout its extent, receives input from the ventroposterior nucleus, and is coextensive with somatic koniocortex. S-I is adjoined lateroventrally by a smaller S-II, which extends into the lateral fissure. As for nonprimate mammals, S-II forms a congruent border with S-I along matched representations of the dorsal midline of the head. The distal limbs in S-II point away from S-I, and the hand is represented in cortex anterior to that devoted to the foot. PV probably exists rostral to S-II, but this area has not yet been identified in prosimians. A narrow strip of cortex on the anterior border of S-I appears to be the homolog of area 3a of monkeys because of its structural similarity and its activation by deep body receptors, possibly muscle spindle receptors. A strip of cortex posterior to S-I responds to more intense stimuli, and this strip may be related to deep body receptors in joints and muscles. Thalamic input to this strip is from neurons dorsal to the ventroposterior nucleus. As in most nonprimates, somatosensory cortex in galagos does not appear to be extensive or to consist of many subdivisions.

Monkeys, Apes, and Humans

Somatosensory cortex in higher primates contains more subdivisions than somatosensory cortex in generalized nonprimates. Most of what is known about somatosensory cortex in higher primates stems from

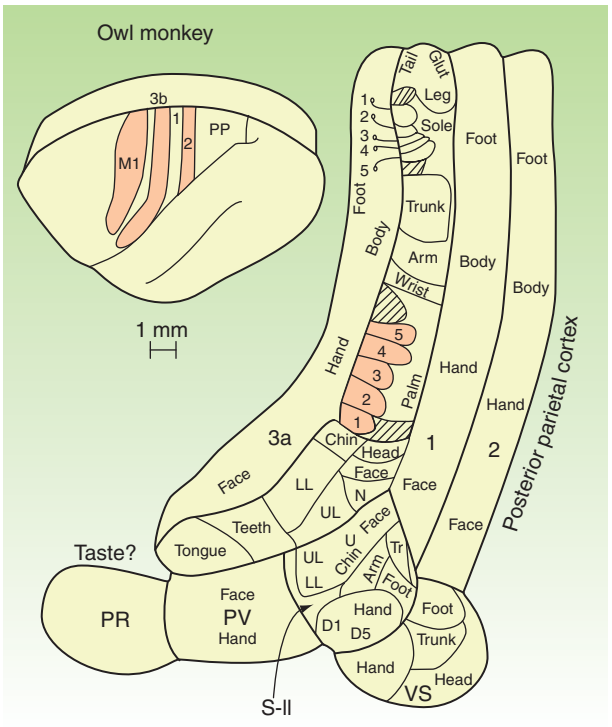

Figure 2 Somatosensory areas in owl monkeys, New World primates with small, relatively smooth brains. These monkeys lack a central fissure, so several sensory and motor areas can be indicated on a dorsolateral view of the brain (upper left). These exposed fields include areas 3a, 3b, 1, and 2 of anterior parietal cortex, posterior parietal cortex, PP, and primary motor cortex, M-I. Other somatosensory fields are hidden in cortex of the upper bank of the lateral fissure. Four of these hidden areas are shown in the expanded diagram of lateral and anterior parietal areas on the right. Little is known about the parietal rostral area, PR, but the other areas all contain systematic representations of body receptors. Detailed somatotopic features are indicated only for area 3b, where digit representations are numbered. Locations of major body part representations are noted for other fields. These areas appear to exist in monkeys, apes, and humans. Glut, gluteal region; N, nose; LL, lower lip; UL, upper lip; Tr, trunk; hatching, dorsal hand or foot.

experiments in monkeys, but it seems reasonable to conclude that somatosensory cortex contains at least as many subdivisions in apes and humans as in monkeys. Early experiments on the organization of anterior parietal cortex in macaque monkeys defined S-I as a broad region including architectonic areas 3 (3a and 3b), 1, and 2 of Brodmann. The locations of these four architectonic strips are shown for an owl monkey in Figure 2. It is now recognized from detailed microelectrode experiments on owl and other monkeys that each of the four architectonic fields 3a, 3b, 1, and 2 contains a separate representation of the body. Since these cortical areas roughly match each other in mediolateral representational sequence, early experiments using larger surface electrodes failed to distinguish them.

Because of the historical definition of S-I in monkeys, it is common to include areas 3a, 3b, 1, and 2 in S-I. However, only area 3b appears to be the homolog of S-I as generally defined in nonprimates and prosimians. Like S-I, area 3b is coextensive with cortex distinguished by densely packed small cells in layer IV, has dense input from the ventroposterior nucleus of the thalamus, is responsive throughout to light tactile stimuli, and has digit tips of the hand and foot represented along the rostral border. Area 3b forms a complete representation of the body surface. Major cortical projections of area 3b are to S-II, PV, area 1, and area 2.

Area 1 also contains a complete representation of the body surface. The somatotopic organization of area 1 is roughly a mirror reversal of that in area 3b (Figure 2). Thus, the digit tips of the hand and foot are represented posteriorly in area 1, and the somatotopic pattern is the reverse of that found in S-I of galagos and nonprimates. In comparison to area 3b, layer IV is less densely packed with small cells and the input from the ventroposterior nucleus is less dense. Area 1 also differs from area 3b by having fewer neurons with slowly adapting responses to maintained stimuli, neurons with larger and more complex receptive fields, and some neurons that appear to be activated by the rapidly adapting Pacinian receptor class. Major cortical projections are to areas 2, S-II, and PV. While area 1 is architectonically apparent in the cortex of monkeys, apes, and humans, there is no compelling evidence for this field in prosimians and nonprimates. Thus, it appears that the area 1 representation exists, or at least it is well developed, only in higher primates.

Area 3a is characterized by less dense packing of small layer IV cells and more layer V pyramidal cells than area 3b. Area 3a is activated predominantly by muscle spindle receptors and receives input from an anterodorsal division of the ventroposterior thalamic complex. The details of the somatotopic organization of area 3a have not been fully determined, but the field represents deep receptors in parallel with the adjoining representation of cutaneous receptors in area 3b. Major connections are with motor cortex. Similarities in responsiveness, histological structure, and even connections of area 3a in monkeys to the field termed 3a in cats are striking. Since a comparable 3a field has not been identified in a range of mammalian species, many of these similarities may be the result of parallel evolution. However, the reported presence of a somatosensory fringe along the anterior border of S-I of a number of mammals suggests that a less developed 3a field exists in many or most mammals.

Area 2 occupies a strip of cortex immediately caudal to area 1. Layers IV and VI are more densely

packed with cells than area 1. The posterior boundary of area 2 is less apparent, and there is disagreement on the location of this border. The somatotopic organization of area 2 is incompletely understood, but much or all of area 2 contains a body representation that is parallel to that in area 1, and at least partly mirrors that in area 1. Area 2 is activated by both cutaneous receptors and receptors in deep body tissues (muscles, joints, and tendons). The information from deep receptors appears to be relayed from dorsal divisions of the ventroposterior thalamic complex, and perhaps the anterior pulvinar, while information from cutaneous receptors appears to be relayed from areas 3b and 1 and, to a lesser extent, from the ventroposterior nucleus. Many neurons in area 2 are activated by both skin and deep receptors, and many respond best to complex stimuli involving shapes or direction of movement. Thus, area 2 seems to be a somewhat higher-order processing station. Major projections are to area 3a, S-II, PV, and posterior parietal cortex. There is no clear evidence for a homolog of area 2 in prosimians or nonprimates. However, in a number of mammals, cortex posterior to S-I receives input from part of the thalamus that is dorsal to the ventroposterior nucleus, and input from this location is suggestive of the input to area 2 from the dorsal parts of ventroposterior complex in monkeys.

Somatosensory cortex in monkeys also includes a large region of cortex on the upper bank and insular surface of the lateral fissure and the large posterior parietal region of cortex. The lateral fissure contains several somatosensory areas including an S-II proper that appears to be homologous to S-II in nonprimates. As in other mammals, S-II in monkeys responds to cutaneous stimuli and forms a systematic representation of the contralateral body surface. Cortical inputs are from areas 3b, 1, and 2 and thalamic inputs appear to be largely from the ventroposterior inferior nucleus and perhaps somewhat less from the ventroposterior nucleus. Pacinian receptors may relay to part of S-II or a subdivision of cortex outside of S-II.

One recent concept of S-II function in monkeys, and thereby in other mammals, is that S-II provides a hierarchical link in a relay of somatosensory information to limbic cortex that is critical in the permanent learning of tactile discriminations. This hierarchy also includes a PV and a ventral somatosensory area, VS. Both of these fields receive inputs from S-II and can be considered higher-order fields. Cortex rostral to PV, the PR region, receives inputs from PV and S-II. The organization of other regions of cortex in the lateral fissure of monkeys is not completely understood. Some of this cortex in the lateral fissure responds to somatosensory stimuli and contains subdivisions with different architectonic characteristics, types of neurons, and connection patterns with other fields and thalamic nuclei.

Posterior parietal cortex has been subdivided in various ways, but most investigators refer to subdivisions of area 5 and area 7 of Brodmann. These subdivisions are not completely satisfactory because there is no general agreement as to the exact locations and boundaries of cytoarchitectonic fields or on the criteria that defines them. In addition, and very importantly, these fields do not seem to be functionally homogeneous. Thus, the concepts of areas 5 and 7 have limited usefulness, and ultimately they may be redefined or replaced by more precisely defined subdivisions. Nevertheless, some general conclusions are possible. The more anterior part of posterior parietal cortex (area 5) receives cortical inputs from areas 1 and 2, and area 5 sends to more posterior parietal cortex (area 7). Area 7 also receives inputs from the lateral posterior nucleus of the thalamus and subdivisions of extrastriate visual cortex. The area 7 region appears to have higher-order functions since the response properties of neurons are complex and include types related to vision, eye movements, and visually guided movements. Lesions of area 7 produce visuospatial disorientation, defects in eye movements, misreaching, and a reluctance to use contralateral limbs.

Finally, as for other mammals, motor cortex has somatosensory inputs and thus somatosensory functions in monkeys. Cortical fields involved in sensorimotor functions include primary motor (M-I), supplementary motor (SMA), and dorsal (DPM) and ventral (VPM) premotor areas have been defined in monkeys.

Modular Organization

In early landmark studies of somatosensory cortex, Mountcastle proposed that cortical areas are not uniform in function, but are divided into small, functionally distinct processing modules, each containing a limited group of cells in a column extending from the surface of the cortex to fibers below. This concept has been widely accepted, but there have been only limited efforts to reveal the nature of processing units in any region of somatosensory cortex. The general validity of the concept is strongly supported by anatomical studies of connections, especially corticocortical connections, that cells in one location in the somatosensory system often project to a number of closely spaced separate foci in another area, thus suggesting a spacing of neurons involved in a particular function.

Recent electrophysiological evidence for one type of modular organization came from experiments where the local distributions of slowly adapting and

rapidly adapting cortical neurons were determined within area 3b of monkeys by recording from successive neurons across the depth of cortex in many closely spaced microelectrode penetrations. In the representation of the glabrous hand in area 3b of monkeys, most or all neurons respond to slight indentation of the skin with a probe. Some neurons respond only to the onset and offset of the indentation (rapidly adapting) while others continue to respond during a maintained indentation (slowly adapting). The slowly adapting neurons appear to be largely confined to the middle layers of cortex where they are clustered in extended foci that appear to form closely spaced bands coursing parallel to the surface. Cells in superficial and deep layers over and under these slowly adapting bands are rapidly adapting, and cells between the slowly adapting bands in layer IV form rapidly adapting bands. The reasons for this distribution are not clear. One possibility is that inputs relayed from slowly adapting and rapidly adapting receptor classes in the skin distribute in alternating bands in layer IV of cortex, and that slowly adapting responses are inhibited in cells outside the receiving layers. However, the observations demonstrate that processing modules in somatosensory cortex may be more band-like than column-like in form, and may be related to specific cortical layers, as are orientation-selective bands and ocular dominance bands in visual cortex.

The Reorganization of Cortical Maps after Injury

The somatotopic organizations of areas of somatosensory cortex do not reflect the full potential of the anatomical connections. Local features of organization in these maps can be altered by small lesions in the representations or by removal of the source of an input to part of a representation. Thus, cutting a sensory nerve to part of the hand or removing a finger does not permanently silence the cortex corresponding to the former representation of part of the hand or a finger. Rather, over a short period of hours to weeks, neurons throughout the deprived cortex become responsive to stimuli on the normally innervated parts of the hand. Likewise, following a small lesion of cortex that removes the representation of a given finger, neurons in cortex representing adjacent fingers come to be activated by inputs from the finger with the removed representation. These experiments indicate that somatotopic representations in adult mammals are subject to small degrees of change and reorganization, and that many local features of representations are dynamically maintained. Because some of the changes can be quite rapid, the anatomical connections mediating the rapid reorganization appear to be at least roughly in place, and the rapid reorganization appears to be one of changing the effectiveness of previously existing anatomical pathways. Other slowly emerging changes may involve the growth of new connections. The ability of these representations to change in internal organization may relate to recoveries after brain damage or sensory loss.

See also: Cerebral Cortex; Evolution of Vertebrate Brains; Neocortex: Origins; Neural Circuitry in the Somatosensory System; Somatosensory Cortex: Functional Architecture; Somatosensory Receptive Fields.

Further Reading

Johnson JI (1990) Comparative development of somatosensory cortex. *Cerebral Cortex* 8B: 335–449.

Kaas JH (1983) What, if anything, is SI? Organization of first somatosensory area of cortex. *Physiology Review* 63: 206–231.

Kaas JH (1991) Plasticity of sensory and motor maps in adult mammals. *Annual Review of Neuroscience* 14: 137–167.

Kaas JH (1993) The functional organization of somatosensory cortex in primates. *Annals of Anatomy* 175: 509–518.

Somatosensory Cortex: Functional Architecture

R J Nelson and and Y Liu, University of Tennessee Health Science Center, Memphis, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The phrase 'functional architecture' provides an interesting juxtaposition which, like 'mammalian somatosensory cortex,' must be explained in more detail. On one hand, dynamicism is implied by the use of the word 'functional,' suggesting something that is constantly changing in or being changed by the environment. On the other hand, architecture leads us to think about a fixed structure. It is safe to say that the somatosensory cortex has intrinsic scaffolding on which dynamic sensory processes act and that these processes are sculpted by inputs from the external sensory environment processed through the body surface (the exteroceptive system). Moreover, the somatosensory cortex plays double duty because it also receives sensory inputs from receptors deep within the body itself that signal the state of joints, muscles, and tendons (the proprioceptive system). The sculpting of processes may be accompanied by some changes in the architecture that can consist of changes in the way cortical neurons are connected or the strength of those connections. However, just as connections may contribute to functional plasticity, the extent of the plasticity may be limited by the architecture itself.

Vertical Architecture

The mammalian neocortex consists of six cellular layers stacked up on one another from the cortical surface (layer I) to the place where the deepest layer (layer VI) abuts the connectional fibers, called the cortical white matter (Figure 1). Each of these six layers contains neurons (some more than others) that are arranged vertically, that is, from the white matter border to the cortical surface. Neurons are arranged in units called minicolumn. Mountcastle describes a minicolumn as "... a narrow chain of neurons that extends vertically across the cellular layers II–VI, normal to the pial [cortical] surface" (1998, p. 165). By most accounts, in somatosensory cortex, these units consist of 80–100 neurons and form the building blocks from which larger, functionally specific cortical columns are made. As with all cortical sensory areas, the somatosensory cortex contains topographic representations of sensory inputs that project there and are thus made available to other cortical areas for processing.

The larger cortical columns that make up the somatosensory cortex consist of groups of minicolumns sharing a common feature, such as body surface location, but minicolumns within a cortical column may receive input predominantly from one type of peripheral sensory receptor type. As noted by Mountcastle, this is a way that several kinds of sensory signals can be represented in one cortical location. His pioneering work in the somatosensory cortex demonstrated that neurons in columns that are around 500 μm in size share similar functional properties and that properties change abruptly when examination moves from one column to the next. These property changes result because each column is driven by a small set of afferent fibers and afferents of several different types, or modalities, may be located at the same spot on the body. In somatosensory cortex, place is the primary representation and small zones on the body called receptive fields are best at activating neurons at a given cortical location.

Several features of the vertical architecture are unique and deserve additional mention. Although place is primarily represented in somatosensory cortex, afferents with different sensitivities and different ways of representing the same stimulus at the same peripheral location are often near each other in the cortex. Thus, contained within segments of a cortical place representation are several smaller islands in which inputs from one modality at that place are represented. The vertical organization of some parts of rodent somatosensory cortex stands out both visually and as a model for studying the relationships between peripheral receptors and their cortical representations. Cortical barrels can be seen in stained cortical cross sections and, in most instances, receive input from one peripheral hair, especially for those associated with whiskers. (This, as well as other model systems, is discussed later in this article.) Finally, some neurons in vertical units in the cortex are not so much limited to a columnar domain as they are relegated to spanning several columns. Neurons of this type may be either excitatory or inhibitory, but their function appears to be to shape the inputs and outputs of columns that may be near or at some distance from their location. Often times, inhibitory neurons are difficult to detect because, unlike neurons that are excited by stimuli, their influences are often seen in the activity of other cells whose activity patterns they influence.

Horizontal Architecture

Somatosensory cortex, like other regions of the brain, can be divided into areas based on a number of factors. Classically, Brodmann (*c.* 1909) divided the human neocortex into 52 areas based on the appearance or cytoarchitectonics of their cortical layers. Some functional regions consist of several areas. For example, the primary somatosensory cortex is thought to consist of areas 3, 1, and 2 of Brodmann. Area 3 was later subdivided into areas 3a and 3b based on different connections and functional properties of the cells located in these areas. Although many other systems for cortical classification have been offered, this one is the most commonly used.

Attempts to link cytoarchitectonics with specific functions have been, at times, instructive. Considering the primary somatosensory cortex, functions have been attributed to each of the four areas based on what functions seem impaired when one area is damaged while the others remain intact. Functions have also been inferred by recording the activity of single cells in response to control stimuli and based on connections with other brain structures having known functions. In general, areas 3b and 1 seem to process predominantly exteroceptive information, whereas areas 3a and 2 may be devoted more toward processing proprioceptive information (Figure 2). However, as with most attempts to link structure and function, this is a gross oversimplification. Certainly, each area has a different proportional distribution of submodality inputs to regions that represent a particular body part. What is now appreciated is that the activity induced in cortical neurons by stimuli cannot unequivocally be linked to select functions. This is because sensory processing of a given stimulus is itself dynamically altered by changes in attention, intention, and expectation in ways that are just now being appreciated. That appreciation has come because model systems have been used to study functional architecture.

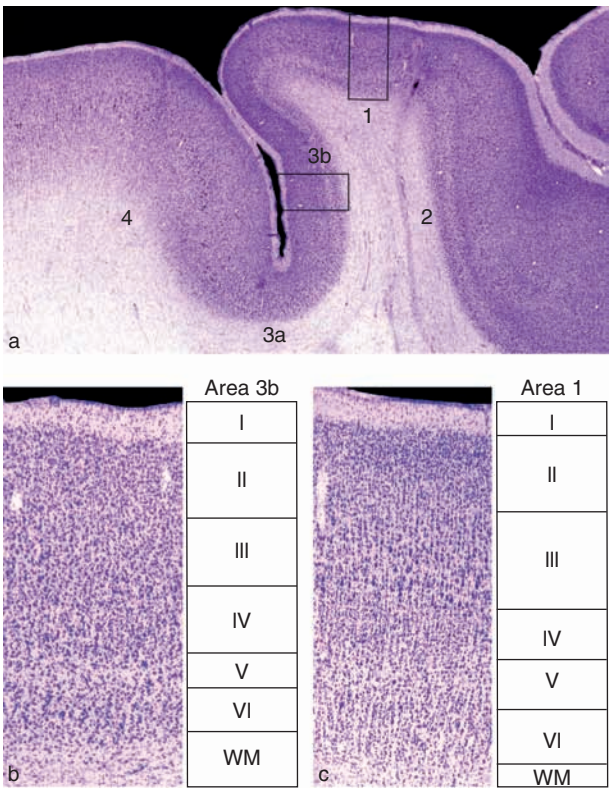

Figure 1 Cross sections through the cortex of a rhesus monkey: (a) cross-section through the left sensorimotor cortex with boxes indicating the regions enlarged in (b) and (c); (b) an enlarged section through area 3b and the approximate boundaries of the six cortical layers (I–VI) that make up the cortex; (c) an enlarged section through area 1 and the approximate boundaries of the six cortical layers. Panels (b) and (c) depict the differences and similarities in the vertical organization of somatosensory cortex. WM, white matter.

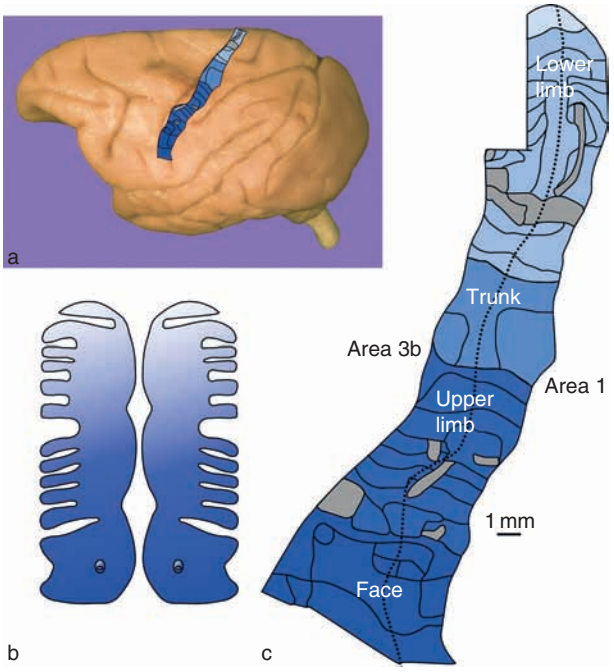

Figure 2 The somatosensory cortex and representations of the body in the cortex: (a) lateral view of the left cortical hemisphere of a rhesus monkey, with the approximate location of somatosensory cortex visible on the surface of the postcentral gyrus shown as a blue region superimposed on the picture of the cortical surface; (b) simianuculi, distorted drawings of the body in portions approximately equal to how they are represented in cortex; (c) two parts of the representations of the body within somatosensory cortex. In (b), lighter shades indicate body regions represented more medially and darker shades indicate regions represented more laterally in the cortex. In (c) (as schematized in (b)), representations in adjacent cortical areas are essentially mirror images of one another.

Model Systems of Functional Architecture

The somatosensory cortex of some mammals is so specialized that it serves as a unique model system to study structure and function together. Functional segregation may even be seen at the macroscopic level on the surface of the brain itself. One example is the hand representation of the raccoon in which each of the lobules located in the lateral part of the cortex contains the representation of a single digit of the hand. Raccoons have great dexterity and are surprisingly agile when using their hand to manipulate objects. Perhaps this should not be a surprise because their brains are specialized to not only accommodate the myriad sensory inputs that arise from their relentless exploration of their sensory environment but also to control the movements that guide their hands. Thus, their motor cortex is highly specialized too. In some instances, animals find out about the world around them more by tactile cues than by visual ones. An example of this is the star-nose mole, which makes up for its poor eyesight with a specialized nose containing 22 fleshy appendages (Figure 3). The number of receptors per unit area in these appendages and the way in which these receptors are arranged make this mole's nose several times more sensitive than the human hand. These creatures have elaborate specialization of their brain maps which allows them to identify objects as edible in less than 10 ms.

In other animals, the segregation of functional units is no less specialized, but it is best seen with special staining. An example of this is the rodent barrel system that represents mystacial vibrissae. The rodent barrel cortex, as a model system, provides unique opportunities and challenges for studying the linkage between structure and function in the somatosensory cortex. Cortex barrels are seen best when selective tissue stains are applied to horizontal section of cortical tissue cut from flattened brains. From above, the distorted profile of much of the animal's body surface representation can be seen in sections cut through layer IV, particularly in areas representing the whiskers, where barrels are easier to spot than in other areas. This is especially true in the posteromedial barrel subfield in which there seems to be one barrel for each peripheral whisker. These barrels have cell-rich walls and cell-poor hollows or centers that are filled with fibers that synapse with cells in the wall. In layer IV, each barrel

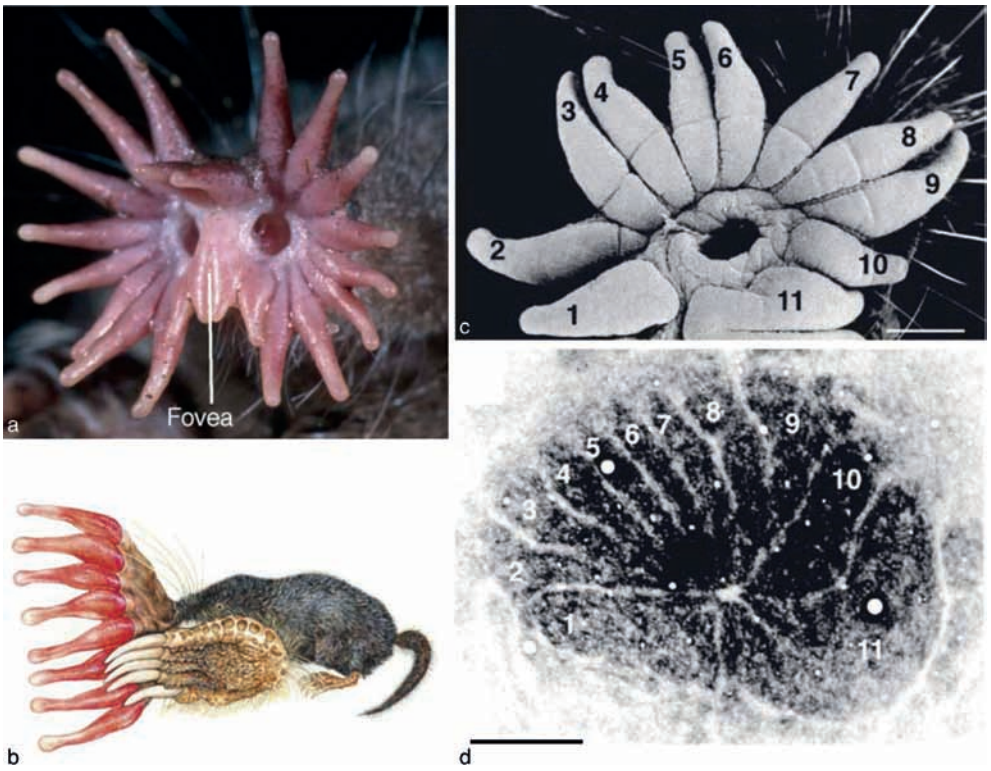

Figure 3 The nose of a mole: (a) the appendages of the mole's nostrils; (b) how the mole would look if the parts of the mole's body were the size of their relative representations in the mole's somatosensory cortex; (c) the right half of the mole's nose (rotated to match the S1 representation in (d)), (d) S1 representation of the nose star in the somatosensory cortex, In (d), each nasal appendage is represented by a cytochrome oxidase-dense stripe of cortex. Scale bar = 1 mm (c), 0.5 mm (d). Reproduced from Catania KC and Henry EC (2006) Touching on somatosensory specializations in mammals. *Current Opinion in Neurobiology* 16: 467–473, with permission from author.

represents predominantly one whisker, although above and below this layer inputs from adjacent whiskers can be found that shape the responses of the neurons of that barrel. Rodents explore their environment by whisking between 8 and 10 times per second. Thus, contact with objects sets up waves of afferent input that have unique spatial and temporal patterns. These can be demonstrated by recording the activity of many cortical cells simultaneously and displaying the activity patterns that arise in plots that describe what is happening spatially in the cortical representation over time. It has been shown that the waves of activation are not symmetrical but, rather, have directional biases that match well with the direction that rodents sweep their whiskers when exploring. Moreover, in time, one region of the somatosensory cortex may be highly active only to be suppressed in the next instant by sensory inputs that change the location of the focal points of activation in the cortex. In addition, when we carefully examine the activity of many neurons in different layers of the cortex, it is possible to see how information initially coming in from the thalamus is propagated through supra- and infra-granular layers of the cortex. Finally, alterations to the whiskers have been shown to result in the alteration of the sensory representations of those whiskers throughout the ascending somatosensory system. Because of the ease with which the structure of sensory representation can be demonstrated in rodent somatosensory cortex, it has been possible, using this model system, to show how excessive use or lack thereof changes the structure of the representations themselves. The ability to do this has led to theories about functional cortical plasticity that can be tested relatively easily in this model system, in which the demonstration of results is also somewhat easy to do.

Although many theories involving behavior can be tested with rodent models, those studying complex behaviors to unravel the functions that play out on the structure of the somatosensory cortex often turn to nonhuman primate models. Nonhuman primates have characteristics that make them uniquely suited for exploring the relationship between structure and function in the somatosensory cortex. Old World and New World primates have several somatosensory cortical areas that receive inputs from peripheral receptors but in different proportions. Because of this, it is possible to sample input from predominately cutaneous receptors in one region and proprioceptors in a regions just adjacent to the first. Thus, in a single experiment submodiality-specific responses to identical stimuli can be compared and contrasted. Nonhuman primates are the model choice for experiments involving complex sensory stimuli and behaviors in response to them.

Imaging the Functioning Somatosensory Cortex

The somatosensory cortex has not been imaged as much as other cortices such as the visual cortex or the frontal cortices. Those imaging studies that have involved the somatosensory cortex have provided us with a new appreciation of its complexity of function and the plasticity inherent in the system. As discussed by Mountcastle, imaging techniques in humans and animals are relatively good at answering questions about where within the somatosensory cortex functional processing is occurring and, to some extent, whether the region(s) involved stays relatively constant in size or expands and contracts as conditions change. For example, the central representations activated by peripheral stimulation have been imaged in experimental animals by 2-deoxyglucose. Often times, glucose use appears to reveal areas of activity that are many times larger than would be predicted or appreciated by conventional mapping experiments. With the use of positron emissions tomography (PET) and functional magnetic resonance imaging (fMRI) techniques has also come increasing resolution in describing regional specializations in cortical somatosensory representations. Increases in resolution have been accompanied by increases in understanding about the distributed nature of somatosensory processing and the fluidity of cortical activity. To a certain extent, these methods have shown that attention, intention to move, and expectation of reward can and do alter somatosensory cortical activity. Recent combinations of imaging techniques with electrophysiological recording within the sites of activation in experimental animals have yielded an appreciation for the potential of somatosensory cortical areas to simultaneously represent and process different types of information at the same cortical site.

Although cortical areas of activation seem to change based on stimulus conditions, there may be spatial limits to the expansion of the representation of sensory inputs imposed by the intrinsic structure of the cortex itself. Extreme examples of this are seen when sensory representations in monkeys were mapped after deafferentation. The somatosensory cortex can reorganize to represent what remains intact in the sensory periphery. But if the spatial extent of the deafferentation is too large, rather than filling in the gaps in the cortical representation, there are gaps. These gaps may belie the underlying distribution of cortical inputs to columns in the cortex. New inputs that seem to arise may actually have been there all along but have become unmasked. Spatial limitations and limits on the spread of inputs to the cortex may go hand in hand.

Plasticity in the cortex and limits on type of reorganization may themselves be different for different cortical layers. Recent evidence suggests that not all layers at a given cortical locus change in the same way in the face of stimulus conditions known to alter sensory responsiveness. Repetitive fictive whisking in rats causes facilitation of sensory responses in the cortical input layers (layers 4 and 5a) and suppression of responses in the superficial layers (layer 2/3). Moreover, repetitive touch stimulation results in stronger activation of layers 2/3 and 4 but not layer 5a. Facilitations were specific to protracting the whisker, whereas retraction was associated with depression of activity rates. This example illustrates that alterations in sensory input during active cortical processing may themselves alter the way in which the cortex responds. Moreover, this reorganization seems to be restricted by the intrinsic network dynamics which are superimposed on cortical structure.

Cortical Function during Movement

Although responses to sensory stimuli are sculpted during active use, certain similarities remain to what is seen in the quiescent condition. It is well known that the way neuronal responses under anesthesia often differ from those that can be observed during active behavior. As already noted, active areas of the somatosensory cortex depend on how and when the cortex receives input. To a certain extent, so does the extent of the region activated by a given stimulus. There are limits, however. When a portion of somatosensory cortex is deprived of functional innervation from the periphery for an extended period of time, that deprived region often can be shown to respond to inputs that were not originally there. Whether this is due to functional unmasking of inputs that were previously suppressed or the establishment of new connections to take the place of older ones is a subject of debate. What is true is that the somatosensory cortex rarely if ever reorganizes whereby new representations spring up in deprived areas that used to represent other major body parts. It seems that the line between body parts is not often crossed and may be a reflection of the intrinsic cortical connections underlying functional sensory representations.

When the activity of somatosensory cortical neurons is observed during purposeful movements, several things begin to become clear. Active movement is accompanied by the enhancement of some neurons and the suppression of others. These enhancements and suppressions vary over short time scales, of the order of tens of milliseconds, and are consistent from movement to movement if conditions remain the same. Many factors seem to influence whether somatosensory cortical neurons respond to sensory inputs. These include whether the inputs are useful in determining which type of movements should be made, whether the inputs may distract and thus compete for valuable cognitive resources needed to initiate desired behaviors, and whether inputs may get in the way of monitoring the execution of behaviors that have been initiated. As a result, for example, neurons activated by an input during the decision phase of a behavior may no longer respond as well, if at all, during the initial part of that behavior even though the activating stimulus is still present. Many have suggested that the parts of the brain that generate behaviors also send a copy of the generation signal (efference copy) to other brain regions. The purpose of this signal may be to inform the other regions that what is to come is self-generated, to cancel inputs that result from the behaviors themselves, or to establish a template against which feedback from behaviors can be compared to see if things are going according to plan. These centrally generated signals appear to alter somatosensory cortical neuronal responses to both central and peripheral inputs. The existence of these signals has been postulated, but proving unequivocally that they exist has been somewhat difficult. Theories behind them provide intriguing insights into the way the nervous system works when it is working.

Attention, Distractions, and Goals

Abrupt changes in attention, instruction and reward expectation, movement intention, and difficulty of making movements affect the activity of somatosensory cortical neurons in both profound and subtle ways. As already noted, responses in the somatosensory cortex are conditional. Attention directed toward or away from peripheral somatosensory inputs alters these responses in ways that suggest that, when the central nervous system is engaged fully in one task, things that are not important for that task have less chance of capturing the attentional spotlight. However, distracting or competing stimuli, if successful, may modify not only the response amplitudes of previously engaged cortical neurons, but may also alter the fidelity with which they represent the existence and characteristics of the stimuli to which they previously responded.

Goals instill value to inputs and allow those evaluating them to talk about the behavioral importance of both inputs to and outputs from the cortex. Most agree that representations of reward history exist in higher brain centers that are involved in decision making and strategy planning. These specialized regions of the brain ultimately connect with the somatosensory cortex, giving it access to plans and influencing the

responses to stimuli that are behaviorally important. When an individual receives positive reinforcement for a goal-oriented behavior, future behaviors often take on the same structure as those that were previously rewarded. Thus, enhanced responses to important stimuli and suppression of activity associated with inputs that are not important are maintained. However, as was recently pointed out, sometimes there is a failure to properly distinguish between previous and future goals. This situation may be associated with compulsive behavior whereby the nervous system constantly checks for behavioral completion or thinks that completed actions have not been. If higher-order brain regions involved in this sort of maladaptive behavior control the responsiveness of somatosensory cortical representations of peripheral inputs, an obvious but tragic mismatch between the actual and perceived may arise. Much remains to be known about how other brain regions influence the somatosensory cortex and how the intrinsic architecture of this cortex mediates these influences.

Changes in function of the somatosensory cortex may be momentary, as in the examples already presented, or prolonged, as in those thought to accompany aberrant motor outputs such as those associated with dystonia. Prolonged use results in prolonged sensory input which is virtually identical over time. In this worst-case scenario, local enhancements and suppressions of sensory information processing may spiral out of control until exuberant responses swamp the nervous system's ability to adjust to changing sensory and motor conditions. Forced alterations in behavior are at times effective in breaking the cycle which perpetuates inappropriate movements that may be the result of inappropriate sensory feedback about them. Although the dynamic regulations of somatosensory processing may be optimized normally and superimposed in the intrinsic structure of the cortex, things at times go drastically awry. Thus, the very processes that allow conditional adaptation may lock somatosensory processing into a cycle that must be actively broken.

Summary

Although much of the architectural framework of the somatosensory cortex is relatively fixed, dynamic changes in function occur. Even the primary somatosensory cortex does not always represent the sensory periphery in the same way. If changes in function are relatively short term, the dynamic modulations are probably accomplished without too much alteration in the structural framework. If, on the other hand, changes are profound and/or long term, structural changes may result as well. Proving that this is the case, however, has been extremely difficult because it is hard to mark what structures initially looked like to see how they have changed. In a very real sense, in the somatosensory cortex, structure and function are inextricably linked.

See also: Barrel Cortex Circuits; Brodmann's Areas; Neural Circuitry in the Somatosensory System; Sensorimotor Integration: Barrels, Vibrissae and Topographic Representations; Somatosensory Cortex; Somatosensory Pathways (Ascending): Functional Architecture; Somatosensory Plasticity; Somatosensory Receptive Fields.

Further Reading

Burton H (2002) Cerebral cortical regions devoted to the somatosensory system: Results from brain imaging studies in humans. In: Nelson RJ (ed.) *The Somatosensory System: Deciphering the Brain's Own Body Image*, pp. 27–72. Boca Raton, FL: CRC Press.

Catania KC and Henry EC (2006) Touching on somatosensory specializations in mammals. *Current Opinion in Neurobiology* 16: 467–473.

Derdikman D, Yu C, Haidarliu S, et al. (2006) Layer-specific touch dependent facilitation and depression in the somatosensory cortex during active whisking. *Journal of Neuroscience* 26(37): 9538–9547.

Flanders M (2005) Functional somatotopy in sensorimotor cortex. *NeuroReport* 16(4): 313–316.

Florence SL (2002) The changeful mind: Plasticity in the somatosensory system. In: Nelson RJ (ed.) *The Somatosensory System: Deciphering the Brain's Own Body Image*, pp. 335–366. Boca Raton, FL: CRC Press.

Genovesio A, Brasted PJ, and Wise SP (2006) Representation of future and previous spatial goals by separate neural populations in prefrontal cortex. *Journal of Neuroscience* 26(27): 7305–7316.

Hsiao SS and Vega-Bermudez F (2002) Attention in the somatosensory system. In: Nelson RJ (ed.) *The Somatosensory System: Deciphering the Brain's Own Body Image*, pp. 197–218. Boca Raton, FL: CRC Press.

Mountcastle VB (1998) *Perceptual Neuroscience: The Cerebral Cortex*. Cambridge, MA: Harvard University Press.

Mountcastle VB (2005) *The Sensory Hand: Neural Mechanisms of Somatic Sensation*. Cambridge, MA: Harvard University Press.

Nelson RJ (1996) Interactions between motor commands and somatic perceptions in sensorimotor cortex. *Current Opinion in Neurobiology* 6: 801–810.

Nelson RJ (ed.) (2002) *The Somatosensory System: Deciphering the Brain's Own Body Image*, Boca Raton, FL: CRC Press.

Tommerdahl M, Simons SB, Chiu JS, et al. (2006) Ipsilateral input modifies the primary somatosensory cortex response to contralateral skin flutter. *Journal of Neuroscience* 26(22): 5970–5977.

Somatosensory Pathways (Ascending): Functional Architecture

L Villanueva, INSERM U 677, Paris, France
L Monconduit and R Noseda, Université d'Auvergne, Clermont-Ferrand, France

© 2009 Elsevier Ltd. All rights reserved.

A number of animal and clinical studies indicate important differences in the functional architecture of ascending pathways that convey tactile and noxious inputs. Moreover, recent data demonstrate that the interactions between bottom-up and central, top-down mechanisms are probably crucial for the discrimination of innocuous and noxious information.

The Anterolateral Quadrant of the Cord

Experimental Findings in Animals

There are important differences in the anatomical and functional organization of the superficial and deep dorsal horn regions that contain the populations of spinal nociceptive neurons. These differences suggest that these areas probably play different roles in the processing of nociceptive information.

The vast majority of nociceptive primary afferent fibers terminate in the superficial dorsal horn (laminae I and IIo), some Aδ fibers terminate in lamina V, and C fibers of visceral origin terminate in laminae V–VII and X, sometimes bilaterally. Although there is no strict segregation of these two populations, it has been shown in several species that neurons activated specifically by noxious inputs are found mainly in the superficial dorsal horn (laminae I and II), whereas neurons activated by noxious and innocuous inputs are located mainly in deep laminae (V and VI). Moreover, there is also another population of neurons in the ventral horn (laminae VII and VIII) that respond to noxious inputs from widespread areas of the body.

Transections of the anterolateral quadrant (ALQ) that abolish pain arising from pathological conditions are associated with loss of thermal sensations at approximately the same levels where analgesia occurs. In this regard, ascending spinothalamic axons labeled following injections in dorsal horn lamina I neurons in monkeys cross at or just rostral to the segment of origin and ascend contralaterally in the middle of the lateral funiculus. There is in fact some variability between species and spinal cord levels with respect to the location of ascending axons from dorsal horn neurons, but most of their axons ascend within the ALQ. In the rat, these axons have a large dorsoventral dispersion within the lateral funiculus, and some are located quite dorsally. Some studies have suggested the possibility that thermoreceptive ALQ axons are located more dorsally than nociceptive axons originating from various dermatomes. Thus, it seems that thermal and pain sensations are close from both an anatomical and a functional standpoint, which could be related to the fact that both sensations are useful for homeostasis.

Although some controversy exists concerning the functional organization of nociceptive pathways in animals, the existence of an area within the ALQ that contains the main pathways giving rise to pain has been demonstrated in different species. The probability of misinterpreting the evidence obtained from experiments on species other than humans is always present since, obviously, animals cannot explain what they are feeling. Whether normal pain is always distinguished from other aversive sensations in experiments on animals is impossible to determine because the physiologist judges the presence of pain by obvious aversive reactions by the animal.

Clinical Findings

The locations of the spinothalamic and relevant spinoreticular fibers in the human spinal cord are shown in Figure 1. During the past century, this pathway was commonly transected to alleviate pain. Patients were unable to feel pain of a pathological condition on the opposite side of the body and could not feel pain or warmth or cold when suitable stimuli were applied. However, although the sensations one would expect to be evoked by such stimuli do not occur, there is evidence that other spinal pathways are conveying impulses to higher centers. For instance, repetitive stimulation with noxious hot or cold stimuli to the parts of the body rendered analgesic by a cordotomy can cause sweating and nausea, which suggests that impulses are activating autonomic centers via pathways other than those located in the ALQ.

When the ALQ is transected, spontaneous pain can occur. This type of pain is usually called neuropathic pain. It occurs in the part of the body that is no longer connected to ascending fibers of the ALQ. It rarely appears soon after section of the ALQ but may occur spontaneously weeks, months, or years after this lesion has been made. Thus, it is apparent that pathways other than the tracts of the ALQ can be involved in the transmission of impulses that can give rise to some sort of pain. Nevertheless, these sensations are often described as 'not pain but worse than pain.' This is a sensation not previously experienced, and the patients find no words with which to describe it.

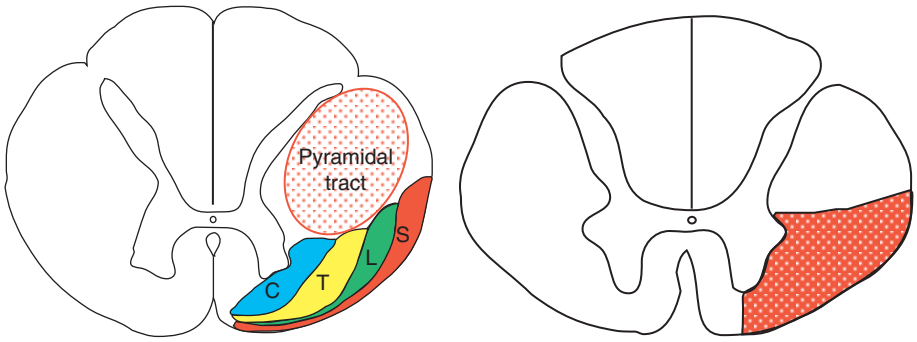

Figure 1 The anterolateral quadrant (ALQ) and pain. Distribution of ascending axons in the spinal cords of patients based on correlations of sensory losses with the location of transections. (Left) Diagram showing the somatotopic distribution of axons from sacral (S), lumbar (L), thoracic (T), and cervical (C) dermatomes in the ALQ. A considerable overlap of fibers is observed in its ventromedial portion. (Right) Diagram showing the location of ascending axons at the C3 level following transections that produced long-lasting contralateral analgesia in humans.

It is often referred to as dysaesthesiae or painful dysaesthesiae. It is poorly localized and may extend over half of the body. If it follows the operation of anterolateral cordotomy carried out for painful malignant disease, the sensations are felt in the area of the malignant lesion.

Another kind of pain that may occur following sectioning of the ALQ involves the sensation being referred. One of the features of this phenomenon is that the patient reports that a noxious stimulus applied at the analgesic part of the body is felt elsewhere. A pinprick applied to the analgesic skin is felt as touch and the referred pain is felt in normally innervated parts of the body as if it comes from within and goes outwards to the skin – a totally new type of sensation.

Following the operation of anterolateral cordotomy, there is usually a return of pain and thermal sensibility. It occurs within months or years, even though the entire tract has been divided. It is probable that this return of sensibility is due to conduction in the ALQ and not to other pathways becoming available since these sensations have their normal characteristics. It has been reported that performing a second cordotomy on the same side as the first 14 years later due to the return of the original pain removed all pain.

Some conclusions may be drawn from these facts. First, to have normal pain, one has to have an intact pathway running in the ALQ. Second, the sensation of 'not pain but worse than pain' does not occur when the pathways previously mentioned are intact; conduction in this system normally inhibits conduction in the other pathways used by impulses, giving rise to these abnormal sensations. If a cordotomy has rendered only part of the body analgesic, then these dysaesthesiae occur only in the analgesic part of the body and never in the part that is still connected to the tracts of the ALQ.

The Dorsolateral Funiculi

Some authors have reported that in various species, a large number of ascending axons from lamina I neurons are located in the dorsolateral white matter. In fact, it has been claimed that in the rat, some of the axons from lamina I neurons ascend in the most dorsal part of the lateral funiculus, an area that does not have to be transected to produce relief of pain in humans. Indeed, electrophysiological and behavioral data in animals show that dorsolateral lesions produce an increase in nociceptive reactions, probably by disrupting tonically active descending antinociceptive pathways. Moreover, dorsal spinothalamic axons lie at the level of the denticulate ligament and terminate preferentially in caudal ventral thalamic areas once referred to as the suprageniculate/posterior complex and now termed the ventromedial posterior nucleus. On the other hand, spinothalamic axons ventral to the denticulate ligament terminate predominantly in the main portion of the ventral posterolateral thalamus. The question of whether a 'dorsolateral' ascending nociceptive pathway exists may be merely a semantic problem since anatomical studies in monkeys have shown that lamina I spinothalamic axons are located more laterally than those arising from deep dorsal horn neurons. These axons are located in an area ventral to the region occupied by the pyramidal tract and are still within the ALQ.

The Dorsal Columns

Experimental Findings in Animals

The vast majority of fibers in the dorsal columns (DCs) are branches of myelinated primary afferent fibers. These fibers run medially and bifurcate in a Y-shaped manner to send one branch rostrally and another caudally that provide collaterals to the

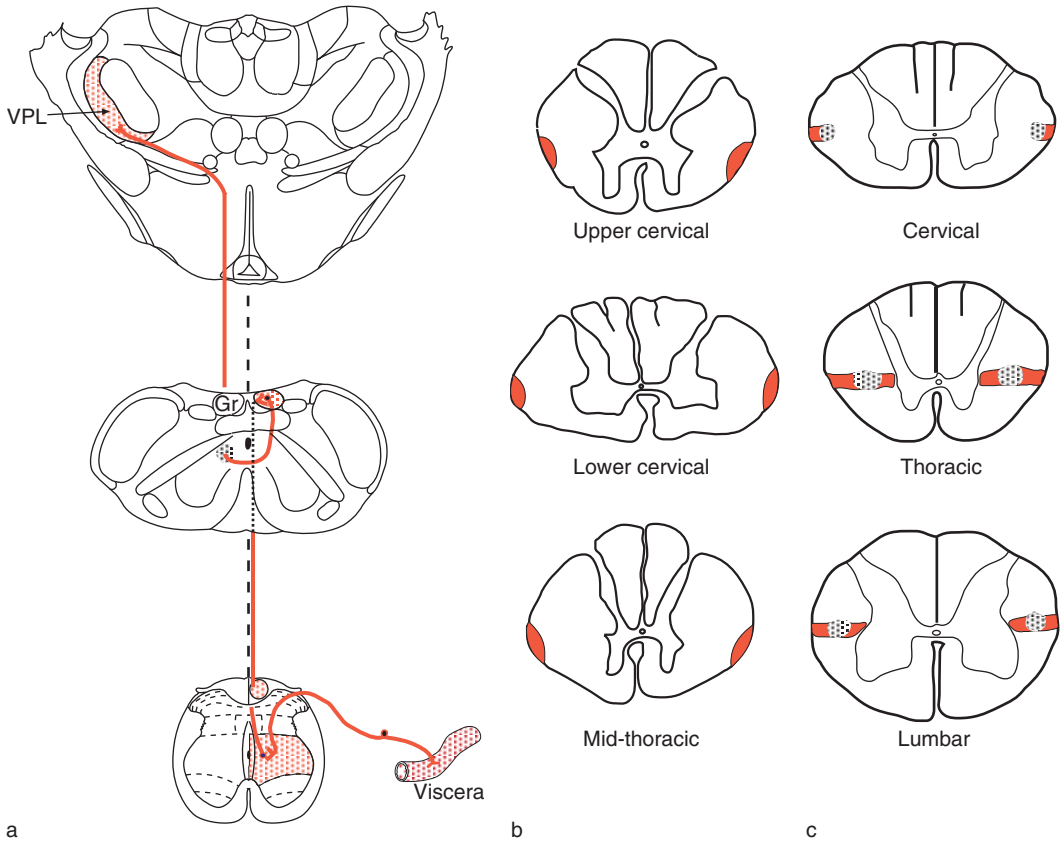

Figure 2 The dorsal columns. (a) The putative visceral nociceptive pathway in the dorsal columns. It has been proposed that nociceptive inputs from viscera are relayed in laminae VII and X neurons, whose ascending axons are confined near the midline of the dorsal columns. These axons terminate in the medullary nucleus gracilis (Gr), which conveys these inputs to the thalamic ventrobasal (VPL) nucleus. (b) The location of centripetal fibers subserving micturition and sensations of temperature and pain from the lower end of the ureter, the bladder, and the urethra in man. (c) The location of centripetal fibers subserving defecation (dots) and the sensation of pain (red area on the lateral surface of the cord) from the rectum and anus. It is probable that the efferent fibers for voluntary control lie in the dotted region and the efferent fibers for autonomic control lie in the more central red area. Sensations that micturition is impending, that it is progressing, and that it has ceased and also sensations of touch and stretch in the anal region are retained following a bilateral anterolateral cordotomy. Such impulses probably travel via the dorsal columns.

dorsal horn. Only some of these fibers reach the dorsal column nuclei (DCN), and a minority of the fibers in the DCs originate from dorsal horn laminae III–V. Most of these neurons are activated by low-threshold mechanoreceptors and some of them from nociceptors activated by mechanical stimuli of the skin or deep structures.

Several studies in rats have shown a visceral nociceptive pathway that activates gracile neurons. As illustrated in Figure 2(a), the ascending axons that activate these neurons following either stimulation of reproductive pelvic viscera or colorectal distension are confined to the fasciculus gracilis. In addition, the suppression of ventrobasal (VPL) thalamic responses to colorectal distension following dorsal column lesions led to the hypothesis that this pathway is more important than the ALQ for transmitting visceral nociceptive signals to the VPL, whereas the ALQ may be more important for transmitting cutaneous nociceptive information. It was concluded that the visceral inputs traveling in the DCs could either cooperate with direct spinal pathways to produce the perceptions of touch and pain or serve as an alternative nociceptive pathway to the spinothalamic tract.

Clinical Findings

Patients with lesions including the DCs have impaired vibratory and postural sensibility, diminished ability to distinguish two points, poor tactile localization, and ataxia. Thus, in the mid-eighteenth century the concept of a dorsal column syndrome was proposed on the basis that the DCs conveyed only tactile and proprioceptive sensations. However, both animal and clinical studies indicate that the role of the dorsal columns–lemniscal system is more complex.

If there are two nociceptive pathways conducting impulses causing visceral pain, then the removal of

one should either have no effect or have some effect in reducing the pain. This effect should be present immediately after one of the pathways has been cut. The first cordotomies were usually designed to remove visceral pain, and the success of this operation has been amply confirmed throughout the past century. This operation has been done far more frequently for pain arising in the viscera than for pain of somatic origin. Although in some cases the pain arose from the spread of carcinoma to the posterior thoracic or abdominal wall and the pleura or peritoneum, in the vast majority of cases the operation was performed to eliminate pain from the kidneys, ureter, bladder, urethra, testes, ovaries, uterus, vagina, stomach, duodenum, pancreas, small and large intestines, as well as from the malignancy invading bone, the sacral and brachial plexuses, and various peripheral afferent nerves.

Indeed, the DCs interact in a complex manner with the pain system. The pain of distorted posture that can occur as a phantom sensation after amputation is relieved by cutting the DCs. In fact, the DCs not only convey ascending inputs to the brain but also contain direct descending spinal projections originating from the DCN and terminating in dorsal horn laminae I and V. In addition, the ventral DCN region, where the majority of spinally projecting cells are located, is an important target for corticobulbar projections. Thus, the spinal output could be modulated not only directly by the DCN–spinally projecting neurons but also indirectly through cortico-DCN influences. The possibility that the information traveling in the DCN could be under the influence of inputs from several levels, including the cerebral cortex, suggests that DCN neurons integrate information from several ascending and descending systems.

It may well be that transections confined to the DCs not only disrupt an ascending pathway but also interfere with descending modulatory mechanisms. Perhaps improved knowledge of these influences might help us understand the complex modifications of sensory perception that have been reported to follow lesions confined to the DCs in humans, including changes in tactile sensations, tactile and postural hallucinations, and increases in sensations of pain, tickling, warmth, and cold. Thus, the clinical effects of lesions of the DCs are much more complicated than those of lesions of the ALQ.

The Primary Somatosensory Cortex (S1) as a Key Link between Ascending and Descending Somatosensory Pathways

Although there are species differences in the anatomical organization of somatosensory pathways in terms of density of inputs and topography, the ventrobasal thalamic targets that receive tactile and noxious inputs exhibit in all species a widespread convergence of inputs and encode precisely the intensities of different kinds of stimuli.

However, the fine encoding properties or the fact that thalamic neurons receive somatotopically organized spinal inputs cannot account for a role in a specific aspect of somatosensory processing. The encoding of the intensity of peripheral inputs shown by single unit recordings demonstrates the fine-tuning functions of the neurons, but it does not resolve neural coding functions in the context of normal information processing. Indeed, the brain processes information in real time; that is, it normally accomplishes its task without the need for multiple repetitions to provide statistical resolution of the data.

To address this issue, it is important to take into account the fact that the brain encodes information through the concurrent activity of neuronal populations rather than from activity in single neurons repeated over time. It also has many ways and strategies for modifying the incoming information it receives. The idea of specialized chains of neurons from the periphery to the cortex, or of separated unidirectional pathways which convey the information that inevitably produces the somatosensory experience, is as obsolete as the concept of the homunculus. Perhaps the most important reason lies in the fact that the brain is an active, selective recognition system that has different means and strategies for modifying afferent information, and thus the specificity of a sensation results from *a posteriori* selection. Moreover, the existence of an homunculus supposes the existence of 'something able to read it' which results in an infinite series of processing mechanisms encased for reading the resulting messages. The fact that when considered separately some ascending pathways act as labeled lines does not imply that a sensation is merely a facsimile in consciousness of every stimulus to the central nervous system. Selection is the main mechanism that generates a sensation, and thus by means of modulation the brain modifies the efficacy of certain inputs to the nervous system.

Cortical activity must be highly dependent on reciprocal interactions with thalamic relays since there are nearly 10 times as many fibers projecting back from the cortex to the thalamus as there are in the forward direction from thalamus to cortex. The function of this massive feedback network in the processing of somatosensory information has not been fully elucidated, but attempts to address this question showed that after several months of chronic delivery of an NMDA antagonist into the monkey primary somatosensory cortex, at the site where the hand is represented, the tactile receptive fields of ventrobasal thalamic cells dramatically increased but only in the

hand region. As in other sensory systems, corticothalamic controls seem to excite relay cells directly by reducing the strength of the intrinsic inhibitory mechanisms affecting these cells (i.e., by disinhibiting them). Overall, these studies suggest that the primary sensory cortices, via their corticofugal projections, provide positive feedback for the 'correct' input while at the same time suppressing irrelevant information (the 'egocentric selection').

The rat ventrobasal thalamus contains only excitatory neurons that project mainly to layer 4 of the primary somatosensory cortex. Descending projections in the thalamocortical loop originate primarily in layer 6 of the S1 cortex. The distal dendrites of ventrobasal neurons are densely innervated by these projections, which activate both ionotropic and metabotropic glutamate receptors. These corticothalamic neurons exert both excitatory and inhibitory influences on relay neurons, and the resultant balance of excitation and inhibition is thought to influence the activity patterns and sensory response properties of relay neurons. Moreover, top-down influences from the S1 cortex are required to specifically discriminate between innocuous and noxious cutaneous inputs in the spinothalamic system. Corticothalamic feedback from S1 to VPL is stimulus driven and modality specific, consisting of either an enhancement of innocuous- or a reduction of noxious-evoked cutaneous responses. Thus, S1 produces a selective modulation of VPL responses to somatosensory inputs by engaging specific, GABAergic, and glutamaergic-mediated excitatory and inhibitory mechanisms in the thalamus.

On their way to the thalamus, corticothalamic projections also send branches to reticular thalamic neurons (RTs) that are exclusively inhibitory and that project locally in the RT or to ventrobasal thalamus. It was suggested on the basis of the similar organization for the visual system that the primary sensory cortices and their thalamic targets do not form a strong, reciprocal loop. The reason for this (the 'no-strong-loop hypothesis') is that a reciprocal excitatory loop would be strong enough to throw the cortex into uncontrolled oscillations, as in epilepsy. Thus, it would be difficult for such a network to constrain its effects, probably because thalamic neurons are near the edge of excitation by the highly specific positive feedback loop circuitry. However, since epilepsy or phenomena related to cortical spreading depression do occur, it is tempting to propose that disturbances in normal sensory processing within these loops could lead to modifications in perception and thus to sensations produced in the absence of peripheral, somatosensory stimulation.

The precise somatotopic organization of S1 and the well-circumscribed anatomical and functional organization of thalamocortical reciprocal modulation lead to small focal activations within S1 that can only be detected with high-resolution approaches. Studies have shown that discrepancies in the interpretation of brain imaging studies were due mostly to the fact that S1 activation by either tactile or painful stimuli is strongly modulated by cognitive factors, including attention, that modify somatosensory perception. Moreover, hypnotic suggestion specifically directed toward changing the perceived intensity but not the unpleasantness of pain evoked by a noxious stimulus significantly modulated activity in S1. Interestingly, modulation of S1 activity seems to occur only with the concurrent presence of somatosensory activation, and thus part of the modulation may be subserved by corticofugal mechanisms driven by peripheral stimuli that are modality specific, such as those described previously. Consistent with this possibility is the observation that nociceptive input to S1 can modulate tactile perception, as shown by the observation that noxious heat is able to reduce the intrinsic optical imaging signal in S1 evoked by innocuous mechanical stimulation of the skin. Furthermore, psychophysical studies have shown that the presence of pain reduces tactile perception of stimuli in the vicinity of the noxious stimulus. This phenomenon was described as a 'touch gate' and could not be attributed to nonspecific changes elicited by generalized arousal or attentional reactions. These mechanisms may also explain earlier clinical observations showing that chronic pain relief resulted in sharpening of tactile perception.

Taken together, human and animal studies indicate that stimulus-evoked responses of S1 corticofugal neurons can contribute to somatosensory discrimination by selecting thalamic responses to preferred versus nonpreferred inputs in a modality-specific manner. In conjunction with mechanisms underlying topographic response sharpening, heightened activity in corticofugal S1 neurons may create a circumscribed zone of enhanced activity within the thalamocortical loop that mediates discrimination between tactile and painful sensations.

The central nervous system networks that integrate somatosensory information do not consist of only a bottom-up process that involves progressively different relays. As summarized in Figure 3, the most important and most widespread source of top-down modulation arises from the cortex. Several cortical regions mediate subtle forms of plasticity by adjusting both thalamic and prethalamic maps and consequently altering all the modulatory mechanisms downstream as a result of sensory experience. Accordingly, psychophysical, behavioral, and clinical evidence show that somatosensory perception is an active process.

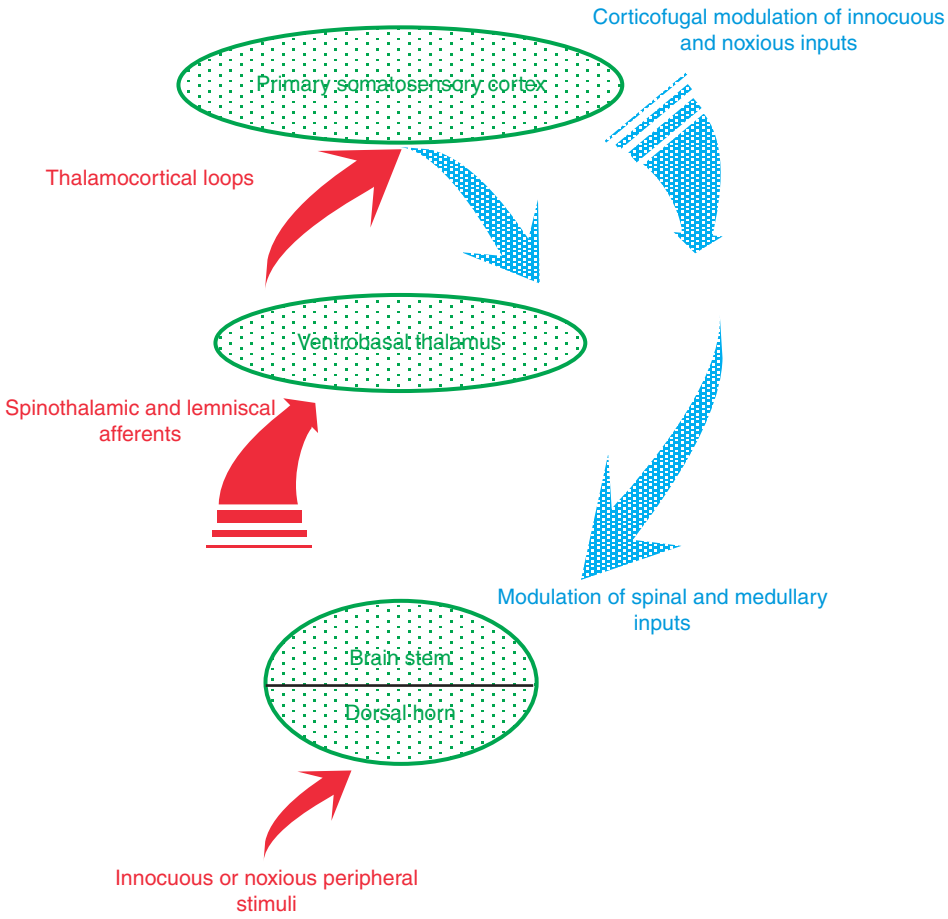

Figure 3 Diagram of the main corticofugal mechanisms of somatosensory modulation from the primary somatosensory cortex. Innocuous and noxious stimuli activate spinal and medullary relays. The most important, widespread source of top-down modulation arises from the cortex since both thalamic and prethalamic relays are under corticofugal modulation.

See also: Cross-Modal Interactions Between Vision and Touch; Neural Circuitry in the Somatosensory System; Nociceptor Responses; Pain: Neuroimaging; Somatosensory Cortex: Functional Architecture; Somatosensory Cortex; Somatosensory Perception; Somatosensory Plasticity; Somatosensory Receptive Fields; Spinal Cord Pain Systems; Temperature Sensation.

Further Reading

Apkarian AV (1995) Thalamic anatomy and physiology of pain perception. In: Besson JM, Guilbaud G, and Ollat H (eds.) *Forebrain Areas Involved in Pain Processing*, pp. 93–118. Paris: Libbey Eurotext.

Berkley KJ, Budell RJ, Blomqvist A, and Bull M (1986) Output systems of the dorsal column nuclei in the cat. *Brain Research Reviews* 11: 199–225.

Canedo A (1997) Primary motor cortex influences on the descending and ascending systems. *Progress in Neurobiology* 51: 287–335.

Craig AD (2003) Pain mechanisms: Labeled lines versus convergence in central processing. *Annual Reviews Neuroscience* 26: 1–30.

Davidoff RA (1989) The dorsal columns. *Neurology* 39: 1377–1385.

Deschênes M, Veinante P, and Zhang ZW (1998) The organization of corticothalamic projections: Reciprocity versus parity. *Brain Research Reviews* 28: 286–308.

Jones EG (2002) Thalamic circuitry and thalamocortical synchrony. *Philosophical Transactions of the Royal Society of London, Series B* 357: 1659–1673.

Monconduit L, Lopez-Avila A, Molat JL, Chalus M, and Villanueva L (2006) Corticothalamic feedback selectively modulates innocuous and noxious inputs in the rat spinothalamic system. *Journal of Neuroscience* 26: 8441–8450.

Nathan PW and Smith MC (1979) Clinico-anatomical correlation in anterolateral cordotomy. *Advances in Pain Research and Therapy* 3: 921–926.

Nathan PW, Smith MC, and Cook AW (1986) Sensory effects in man of lesions of the posterior columns and of some other afferent pathways. *Brain* 109: 1003–1041.

Rainville P (2002) Brain mechanisms of pain affect and pain modulation. *Current Opinion in Neurobiology* 12: 195–204.

Villanueva L and Fields HL (2004) Endogenous central mechanisms of pain modulation. *Progress in Pain Research and Management* 31: 223–246.

Villanueva L and Nathan PW (2000) Multiple pain pathways. *Progress in Pain Research and Management* 16: 371–386.

Somatosensory Perception

J C Craig, Indiana University, Bloomington, IN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

An important factor in developing an understanding of mechanoreception has been the close relationship between psychophysical and neurophysiological measurements. A number of studies have carried out psychophysical measurements with human participants and, using the same stimulus conditions, neurophysiological recordings with monkeys. Even when such measurements have not been carried out in the same study, the psychophysical and neurophysiological measures often use similar stimuli and techniques that facilitate comparisons between measures. Many of the measures of tactile sensitivity discussed in the following sections have been part of corresponding neural measures.

In this article various measures of tactile sensitivity to mechanical stimulation and the results of those measures are described first. In two subsequent sections the effects of aging on tactile sensitivity and the use of tactile stimulation in sensory aids for the blind and deaf are examined. Aging is, of course, a topic of increasing importance as our population ages. Work on sensory aids is an important topic on its own and has been the impetus for a number of studies of tactile sensitivity.

Measures of Spatial Sensitivity

The classic measure of spatial sensitivity is the two-point limen. This is typically defined as the minimum distance between two contact points – presented simultaneously to the skin of the arm, for example – that study participant reports feeling as two points of contact, rather than one point. As a reliable measure of spatial sensitivity, the two-point threshold has been shown to have a number of problems. A major problem with its use is that individuals have difficulty in adopting a stable criterion for responding 'two,' a problem that contributes to the unreliability of the measure. Studies in which individuals were given extensive training with the two-point threshold task showed dramatic changes in sensitivity with repeated testing. At the end of repeated testing, an individual's threshold might drop to one-tenth the size it was at the start of testing. Because the threshold on the arm not tested also declined, the changes in performance are very likely to be the result of instability in the individuals' criteria rather than neural changes associated with the site of stimulation. The use of the two-point threshold as a measure of spatial sensitivity has been likened to a clinical screening procedure in which a patient is asked whether they see one or two fingers.

Another measure of spatial sensitivity, grating orientation, avoids the criterion problem. With this measure individuals are touched with a square-wave grating having equal-width grooves and ridges (Figure 1). The gratings are presented in one of two orthogonal orientations. When testing on the finger pad, for example, the grooves are presented in either the proximal–distal orientation or the lateral–medial orientation. The individual's task is to indicate the orientation. Spatial sensitivity is measured by varying the width of grooves, larger grooves leading to higher levels of performance. In effect, the judgment becomes a two-alternative, forced-choice task, and the problem of the individual's criterion is avoided.

One would expect that a valid measure of spatial sensitivity should correlate well with the density of innervation at particular locations, and indeed grating orientation does. At the distal end of the finger where the spacing between the receptors responsible for fine spatial acuity is approximately 1 mm, the grating orientation threshold is also approximately 1 mm. The threshold at the base of the finger is more than 4 mm and on the palm is nearly 6 mm, reflecting the decline in density of innervation at more proximal locations. A second measure of spatial sensitivity, gap detection, has been found to be a useful measure of spatial sensitivity. In this measure, individuals are presented with one of two stimuli, a solid edge or an edge with a gap in it (Figure 2). The individual's task is to indicate which of the two stimuli was presented. The results with this gap task correlate well with the results from the grating orientation task. The gap task has the advantage that it can be used on less sensitive areas of the body, such as the arms and torso, where the grating orientation task is not feasible: the large groove widths needed to achieve threshold performance in the grating orientation task make the size of the contactors unwieldy.

Sensitivity and Site of Stimulation

The surface area of the skin is extensive, approximately 1.8 m^2 for adult males, and sensitivity varies considerably from one area to another. People's intuitions about which areas are most sensitive are generally correct: the most sensitive areas are the glabrous skin of the fingers, lips, and tongue. Some regional differences in sensitivity are shown in Figure 3. Two

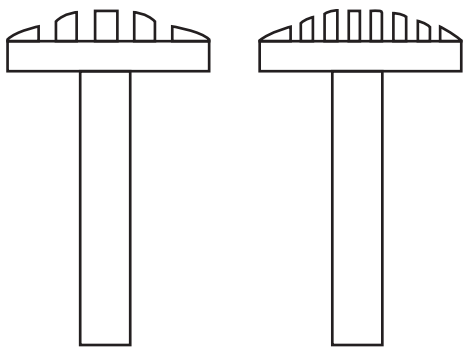

Figure 1 Drawings of two of the gratings used in the grating orientation task. The ability to discriminate the orientation of the gratings depends upon the widths of the ridges and grooves.

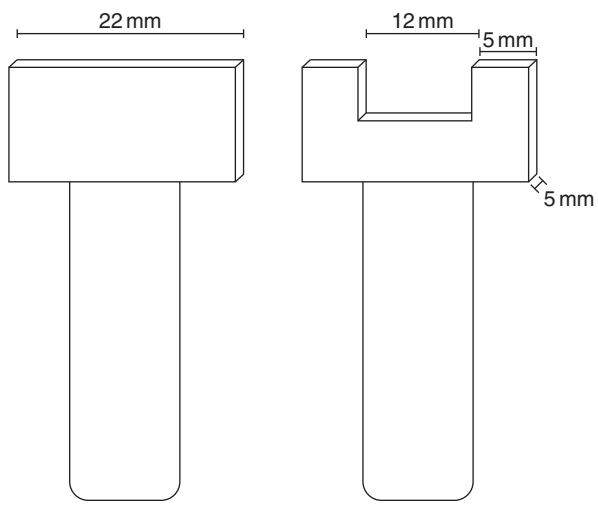

Figure 2 Drawings of the types of stimuli used in the gap detection task. The ability to detect whether the edge (left) or gap (right) was presented depends upon the width of the gap.

measures of sensitivity are shown, absolute sensitivity, as measured with a 50 Hz vibratory stimulus, and spatial sensitivity, as measured with the gap detection task. From these measures it can be seen that some areas of the body are much more sensitive than other areas. With higher frequencies of vibration, 200 Hz, the differences across regions of the body are even larger: the fingertip is 200 times more sensitive than the thigh. The gap detection task shows a reasonable correlation with the vibratory thresholds.

A traditional measure of force sensitivity involves the use of a von Frey hair as a stimulator. To produce a set of von Frey hairs, different lengths and thicknesses of hairs, often nylon bristles, are mounted at the end of a small holder. A hair is brought in contact with a scale until the hair bends, and the force exerted on the scale for that particular hair is recorded. With a set of such calibrated stimulators, the sensitivity of the body to force stimulation can be determined. What is somewhat surprising is that when measuring sensitivity with a von Frey hair, regional differences are much less pronounced and show less correlation with measures of spatial sensitivity. Sites on the hand have thresholds as high as those on the chest and thigh. The fact that vibratory stimuli produce more pronounced regional differences as compared to measures with von Frey hairs, and higher correlations with spatial measures, is likely due to the surface area of the vibratory contactor resulting in spatial summation and to the durative nature of the stimulus resulting in temporal summation. Also, vibratory stimuli may be more effective in eliciting neural activity, compared to the kind of 'step' produced by von Frey hairs.

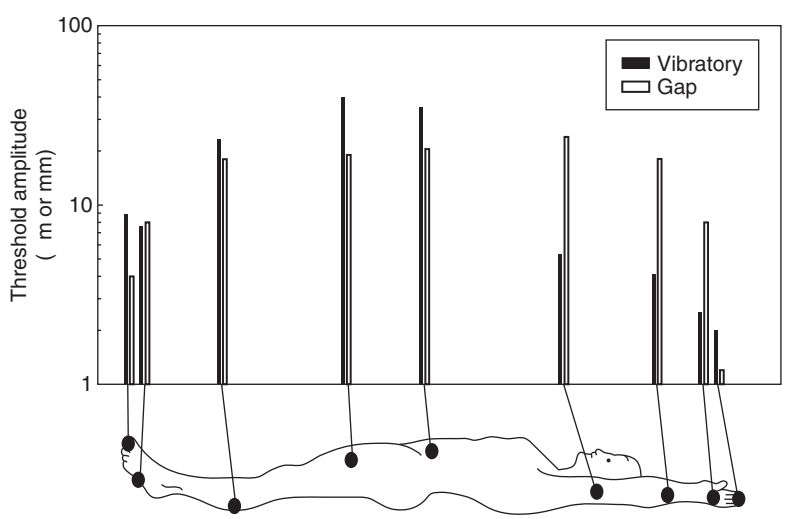

Figure 3 Thresholds for vibratory stimuli (solid bars) and for gap detection (open bars) at selected body sites. The vibratory thresholds are for a 50 Hz stimulus and are measured in micrometers. The gap detection thresholds are measured in millimeters. Based on vibratory data adapted from Wilska A (1954) On the vibrational sensitivity in different regions of the body surface. *Acta Physiologica Scandinavia* 31: 285–289, with permission; and on gap detection data adapted from Stevens JC and Choo KK (1996) Spatial acuity of the body surface over the life span. *Somatosensory Motor Research* 13: 153–166.

Measures of Vibratory Sensitivity

Similar to the measures that have been made with auditory stimuli, vibratory threshold sensitivity has been determined as a function of frequency. There is agreement that sensitivity, measured as a function of the amplitude of the displacement of the vibratory stimulus, is best around 200–300 Hz. Two typical threshold curves are shown in Figure 4. The shape of the threshold curve depends upon a number of variables, including the size of the contactor, the duration of the signal, and physical factors relating to the way in which the contactor touches the skin. Some studies have found considerably more spatial summation with higher vibratory frequencies as compared with lower vibratory frequencies, a result that has been attributed to engaging Pacinian corpuscles (encapsulated receptors in subcutaneous tissue).

Measuring sensitivity to vibratory stimuli extends to discriminating vibratory frequency changes. In making such measurements it is important to understand the basis on which the discrimination is being made. Changing the frequency of vibration produces changes not only in perceived 'rate,' but also in perceived intensity. Several strategies have been used to try and ensure that the discrimination is based on changes in perceived rate, akin to changes in auditory pitch, rather than changes in perceived intensity, or 'loudness.' To avoid this problem investigators have matched different vibratory frequencies for equal perceived intensity prior to the start of the frequency discrimination measurements. Estimates of the Weber fraction from different studies vary somewhat. Weber fractions in the range of 5–10% have been reported for low-frequency vibration, whereas larger Weber fractions are often found at higher frequencies: 20–25% in the range of 200–300 Hz. Difference thresholds of this size are significant not only for what they indicate about how frequency and changes in frequency might be encoded, but also what they indicate about the skin's capabilities in processing speech information (see the section titled 'Tactile aids for the blind and deaf').

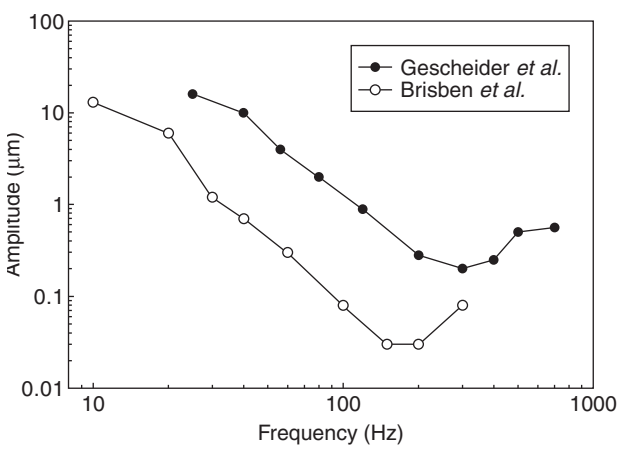

Figure 4 Thresholds for vibratory stimuli as a function of frequency. Two threshold functions are shown. A major difference in the two studies is the area of contact of the vibratory stimulus with the skin. In the study by Gescheider et al. (●) a contactor of 0.2 cm^2 was used. In the study by Brisben et al. (○) the area of contact was much larger. The study participants actively gripped the contactor, a tube, and the area of contact was a large area on the palm and fingers. Data adapted from Gescheider GA, Capraro AJ, Frisina RD, et al. (1978) The effects of a surround on vibrotactile thresholds. *Sensory Processes* 2: 99–115; and from Brisben AJ, Hsiao SS, and Johnson KO (1999) Detection of vibration transmitted through an object grasped in the hand. *Journal of Neurophysiology* 81: 1548–1558.

Measures of Temporal Sensitivity

Temporal Summation

Temporal summation for vibratory stimuli parallels the findings with spatial stimuli; that is, temporal summation is evident with stimulus conditions that favor stimulating Pacinian corpuscles. Higher temporal frequencies and larger contact surfaces yield the greatest amount of temporal summation. With these vibratory stimuli, threshold declines as the duration of stimulation increases out to 500 ms. Based on these and other results, it has been suggested that the Pacinian system (Pacinian receptors and their peripheral and central connections) is, in contrast to the other mechanoreceptor systems, capable of both spatial and temporal summation.

Although not as widely studied as temporal summation for vibratory stimuli, impulse stimuli also demonstrate temporal summation. Increasing the duration of a pulse from 0.35 ms out to 10 ms results in a more than threefold drop in threshold (amplitude of the pulse). Beyond 5 ms, the threshold remains nearly constant.

Temporal Masking

In studies of temporal masking, individuals are typically presented a target stimulus either preceded or followed by a nontarget stimulus, the 'masker.' In detection masking studies, the threshold for detecting the target is measured as a function of the time between the target and masker. Typically the masker is presented at a greater intensity than is the target stimulus. The general result is as one would expect: as the time between the target and masker increases, the amount of masking decreases. Two such masking functions for a tactile pulse are shown in Figure 5. One function shows the amount of masking when the target and masker are presented to the same location,

the right index finger. The overall amount of masking declines if the target and masker are presented to separate locations on the same hand, but as Figure 5 shows, masking is still seen even when the target is presented to one hand and the masker is presented to the opposite hand. This latter finding strongly suggests that the interaction between target and masker is not exclusively a peripheral event but has a central component as well. Detection masking studies with vibratory stimuli show similar functions, masking extending out several hundred milliseconds, and generally more forward than backward masking.

Masking studies with vibratory stimuli are useful for what they may reveal about processing by different types of receptors and their associated neural connections. When target and masking stimuli are presented at different frequencies, the amount of masking may be reduced or eliminated. For example, the threshold for a target stimulus at 15 Hz is unaffected by a masker of 250 Hz until the amplitude of the masker is quite high. Results such as these have been interpreted to suggest that there is no cross-channel masking, that is, when a target stimulus is confined to one mechanoreceptive channel, a stimulus that primarily activates another channel will not interfere with the detection of the target.

In addition to detection, target recognition is also subject to interference by masking stimuli. Spatial patterns such as those generated on the tactile display of the Optacon (see later section titled 'Tactile aids for the blind and deaf') are subject to both forward and backward masking. Using letterlike patterns, it has been shown that the interference in recognizing a target takes several forms. When the target is a spatial pattern presented in close temporal proximity to another spatial pattern (the potential masker), performance in identifying the target may be interfered with because of temporal integration: the study participants perceive the target as having additional features. In some cases individuals may perceive a composite of the target and masking patterns, and on other trials may respond with the masker rather than the target, an apparent difficulty in determining the temporal order of the two patterns.

There are several differences between the studies of recognition masking and detection masking. One is that in the former the masking stimulus does not have to be presented at an intensity greater than that of the target pattern: even when the target pattern is 'masked' it may remain detectable, but not identifiable. A second difference between recognition and detection masking is that there tends to be more backward than forward recognition masking, particularly at briefer temporal separations.

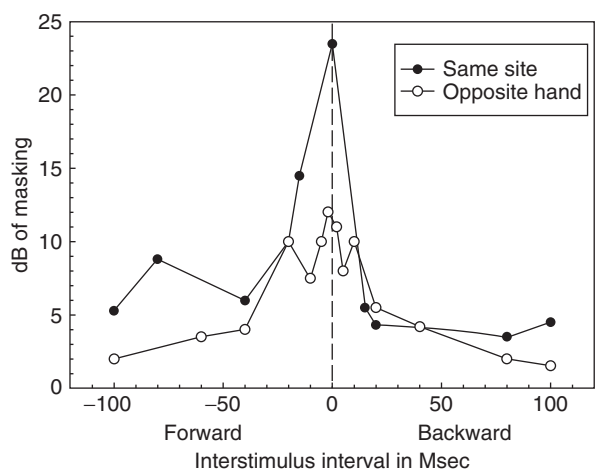

Figure 5 Amount of masking as a function of the interstimulus interval between the target and the masker. Negative time intervals refer to the condition in which the masker precedes the target (forward masking). Positive time intervals refer to the condition in which the masker trails the target pattern (backward masking). Masking is measured by the difference in amplitude (measured in decibels) required to detect the target stimulus in the presence of the masker and in the absence of the masker. Reproduced from Sherrick CE (1964) Effects of double simultaneous stimulation of the skin. *American Journal of Psychology* 77: 42–53, with permission of the University of Illinois Press. Copyright 1964 by the Board of Trustees of the University of Illinois.

Effects of Aging

Given the large number of studies that have shown changes in visual and auditory sensitivity with age, it should come as no surprise that various measures of sensitivity differ between old and young persons. Spatial sensitivity declines with age: spatial gap detection thresholds are about twice as large on the finger pad for individuals in their 70s as compared to those in their 20s. Absolute sensitivity also declines with age. The threshold, for example, for vibration at 250 Hz may be elevated by a factor of 10 for elders. Some studies report that sensitivity for lower frequency vibration remains virtually unchanged with age, whereas other studies show declines in sensitivity at both low- and high-frequency vibrations. The question (unresolved) of whether the loss in sensitivity is restricted to higher frequencies is important for determining the cause(s) for the age-related change: if the loss in sensitivity is restricted to higher frequencies, it suggests that changes in the Pacinian system may be responsible. A more uniform loss in sensitivity across frequencies would indicate changes in other receptor systems as well, or more central changes.

Temporal processing may also show changes with age. The effects of aging are revealed particularly in measures in which effects are seen when the representation of a stimulus persists beyond its offset. For

example, elders show more temporal masking than younger persons do in a forward masking paradigm – that is, when a vibratory masking stimulus precedes the target stimulus. Elders also have greater difficulty in detecting a temporal gap in vibratory noise. The effects of age appear to be greater with higher frequency vibratory stimuli, again suggesting that it is the Pacinian system that is showing more pronounced effects with age.

Spatial sensitivity has been measured at a number of locations on the body for younger and for older individuals. The decline in sensitivity with age was not uniform across locations: more peripheral locations tended to show greater declines in sensitivity. The largest decline in spatial sensitivity was on the great toe, with the elder persons' thresholds nearly five times the younger persons' thresholds. It has been speculated that the decline in sensitivity may be the result of a reduction in the number of receptors, which in turn might be due to decreases in peripheral blood flow. A discussion of possible mechanisms raises the issue of exactly what is meant when changes in sensitivity are attributed to age: older persons may have been exposed to more diseases, are more likely to be taking prescription drugs, have suffered more mechanical and thermal assaults to the skin, and so forth. It has also been observed that thresholds for older persons tend to be more variable than those for younger people – that is, same-age elders may show a greater range of sensitivities compared to same-age younger persons. Thus, age *per se* may not be the sole explanation for the well-documented declines in sensitivity seen in older persons.

Tactile Aids for the Blind and Deaf

The sense of touch has long been considered as a possible channel to aid blind and deaf individuals. Studies designed to examine the skin's capabilities as a communication channel have increased our knowledge of tactile sensitivity. Also, such systems can serve as proofs of the skin's ability to process information.

Tactile Aids for the Blind

Reading by means of braille provides an example that stimuli presented through a single finger pad are capable of being processed at relatively high rates. With braille, the reader is feeling individual cells containing from one to six raised dots. The dots are configured in a two-column by three-row matrix. Measured from the center of the dots, a cell is about 2.3 mm wide and 4.6 mm high. The spacing between dots is greater than the limits of spatial resolution for the finger pad. Proficient braille readers can read at rates up to 100 words per minute. In grade 1 braille, each cell stands for a single character, whereas grade 2 braille makes use of contractions and has cells that stand for short words, thus reducing the number of cells scanned per minute. A useful figure to keep in mind is that in reading grade 1 braille at 60 words per minute, the reader is processing a cell each 180 ms.

A second example demonstrating the rapidity with which stimuli presented through a single fingerpad can be processed is provided by the Optacon (so named for its optical to tactile conversion), a device developed as a reading aid for the blind. This device consists of a tactile display, a 6×24 array of tactors, measuring approximately 1.1×2.2 cm. Connected to the display is a camera which, when moved over printed material, provides an analog signal to the tactile display. If the camera is moved across the letter 'O,' it generates a corresponding circular pattern of vibration that moves across the tactile display. The blind user places his or her finger on the tactile display and moves the camera across the text. Some users of this device can read at rates exceeding 50 to 60 words per minute. Similar to reading grade 1 braille, a reading rate of 60 words per minute corresponds to the user processing the letters at rates of approximately 180 ms per letter. One of the advantages of the Optacon is that it provides the user direct access to printed material and does not require transcription, as does braille.

One potential limitation in reading with the Optacon, and indeed in reading braille, is the field of view, one letter at a time. One can get some sense of this by cutting a small slit, the width and height of a single letter, and then reading a paragraph. The reading rate will be slowed considerably, even though each individual letter is clearly perceived. This observation raises a question, both applied and theoretical, 'Can the tactile field of view be extended to two or more finger pads?' Attempts to alter the tactile array of the Optacon to accommodate two fingers and to allow two letters to be displayed simultaneously did not result in any improvement in reading speed.

Neurophysiological recordings from monkeys reveal that the tactile stimulation produced by the Optacon display does not stimulate Merkel (SA1) fibers; rather, both Meissner (RA1) fibers and Pacinian corpuscle (PC) fibers are stimulated. Because of the inability of PC fibers to transduce spatial patterns, it is likely that it is RA1 fibers that convey Optacon patterns. In light of the SA1 fibers' ability to resolve finer spatial patterns, as compared to RA1 fibers, one would expect that to be identified, letters displayed on the Optacon have to be larger relative to raised letters which activate SA1 fibers; and such is the case.

One of the recurrent questions about the blind and touch concerns whether blind people are more sensitive to tactile stimuli than sighted people are. Recent

interest in this question has been spurred by studies of brain functions that show activity in the visual cortex of blind persons as a result of tactile stimulation and that show that transcranial magnetic stimulation (TMS) directed at primary visual cortex can disrupt processing of spatial stimuli in sighted and in blind individuals. The question of improved tactile sensitivity for blind people appears straightforward, but, like many such questions, the answer is not straightforward. Some studies have found no difference in tactile sensitivity between sighted and blind persons; others have found differences favoring the blind, and still others have found that sighted individuals show better sensitivity. Three recent studies illustrate the complexity. All three of these studies tested spatial acuity with blind and sighted persons, all tested sensitivity on the fingers, and all used the grating orientation task. The fact that the grating orientation task was used is significant: First, as noted, it is a reliable measure. Second, of the various measures of sensitivity, spatial sensitivity might be expected to be affected by loss of sight. Two of the studies found blind persons to show significantly greater sensitivity compared to sighted controls. The third study found no difference in threshold between blind and sighted persons. Indeed, in this study the trend was slightly in the opposite direction. There is no obvious reason for the differences in results among these studies.

Tactile Aids for the Deaf

A variety of tactile aids have been developed for the profoundly deaf. The devices provide information about environmental sounds and aid in speech perception. The tactile information is typically vibratory. In some early attempts to use vibratory information, speech signals were amplified and imposed directly on the skin. The idea was that, with training, individuals would learn to associate particular patterns of vibration with the speech sounds that generated them. These attempts had limited success for several reasons. First, as noted, the skin is maximally sensitive to vibratory stimuli in the range of 200–300 Hz, and becomes relatively insensitive to vibrations above 500–700 Hz, well below the ear's sensitivity and below the range of much of speech information. Second, the auditory system is very sensitive to changes in frequency. At 1000 Hz changing the frequency by as little at 1 Hz is detectable. As noted, for touch the Weber fraction for frequency is about 25% at 300 Hz; an increment of 75 Hz is needed to detect a difference. The skin simply lacks the frequency-resolving power to understand speech directly imposed on it.

Various devices have attempted to circumvent these limitations. Often multichannel tactile aids are used and these employ a variety of ways to analyze the acoustic signals, particularly speech signals. Some analyze speech signals into various frequency bands, and, in a rough approximation of the analysis of frequency by the basilar membrane, the energies in the different bands of frequency are displayed at different sites of stimulation. Other tactile vocoders extract certain features of speech, such as the fundamental frequency and the first formant, and display analogs of the features through multiple sites on the skin. The Tickle-Talker is a device of this sort that generates the tactile signal by electrocutaneous stimulation.

When used as aids to speech perception, tactile aids are primarily used in conjunction with lipreading. The various devices have shown some success in this regard, generally producing significant but modest improvements in speech perception and being judged useful in responding to environmental sounds.

See also: Evolution of Sensory Receptor Specializations in the Glabrous Skin; Human Haptics; Mechanoreceptors; Neural Circuitry in the Somatosensory System; Proprioception; Sensation from the Face; Somatosensory Cortex: Functional Architecture; Somatosensory Pathways (Ascending): Functional Architecture; Somatosensory Plasticity; Somatosensory Receptive Fields; Tactile Coding in Peripheral Neural Populations; Tactile Texture.

Further Reading

Brisben AJ, Hsiao SS, and Johnson KO (1999) Detection of vibration transmitted through an object grasped in hand. *Journal of Neurophysiology* 81: 1548–1558.

Craig JC and Rollman GB (1999) Somesthesis. *Annual Review of Psychology* 50: 305–331.

Gescheider GA, Capraro AJ, Frisina RD, et al. (1978) The effects of a surround on vibrotactile thresholds. *Sensory Processes* 2: 99–115.

Goldreich D and Kanics IM (2003) Tactile acuity is enhanced in blindness. *Journal of Neuroscience* 23(8): 3439–3445.

Johnson KO and Hsiao SS (1992) Neural mechanisms of tactual form and texture perception. *Annual Review of Neuroscience* 15: 227–250.

Johnson KO, VanBoven RW, and Hsiao SS (1994) The perception of two points is not the spatial resolution threshold. In: Boivie J, Hansson P, and Thomas JP (eds.) *Touch, Temperature, and Pain in Health and Disease,* vol. 3, pp. 389–404. Seattle: IASP Press.

Loomis JM and Lederman SJ (1986) Tactual perception. In: Boff KR, Kaufman L, and Thomas JP (eds.) *Handbook of Perception and Human Performance. Cognitive Processes and Performance*, pp. 1–41. New York: Wiley.

Mason JL and Frost BJ (1992) Signal processing strategies for multichannel systems. In: Summers IR (ed.) *Tactile Aids for the Hearing Impaired*, pp. 128–145. London: Whurr Publishers.

Mountcastle VB (2005) *The Sensory Hand: Neural Mechanisms of Somatic Sensation.* Cambridge, MA: Harvard University Press.

Nolan CY and Kederis CJ (1969) *Perceptual Factors in Braille Word Recognition.* New York: American Foundation for the Blind.

Sherrick CE (1964) Effects of double simultaneous stimulation of the skin. *American Journal of Psychology* 77: 42–53.

Sherrick CE and Cholewiak RW (1986) Cutaneous sensitivity. In: Boff KR, Kaufman L, and Thomas JP (eds.) *Handbook of Perception and Human Performance. Cognitive Processes and Performance*, pp. 1–58. New York: Wiley.

Stevens JC and Choo KK (1996) Spatial acuity of the body surface over the life span. *Somatosensory Motor Research* 13: 153–166.

Summers IR (1992) Signal processing strategies for single-channel systems. In: Summers IR (ed.) *Tactile Aids for the Hearing Impaired,* pp. 110–127. London: Whurr Publishers.

Verrillo RT (1968) A duplex mechanism of mechanoreception. In: Kenshalo DR (ed.) *The Skin Senses*, pp. 139–159. Springfield, IL: Thomas.

Verrillo RT (1993) The effects of aging on the sense of touch. In: Verillo RT (ed.) *Sensory Research: Multimodal Perspectives*, pp. 285–298. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weinstein S (1968) Intensive and extensive aspects of tactile sensitivity as a function of body part, sex, and laterality. In: Kenshalo DR (ed.) *The Skin Senses*, pp. 195–222. Springfield, IL: Thomas.

Wilska A (1954) On the vibrational sensitivity in different regions of the body surface. *Acta Physiologica Scandinavia* 31: 285–289.

Somatosensory Plasticity

C Darian-Smith, Stanford University School of Medicine, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

This chapter focuses on somatosensory pathway reorganization following peripheral injury and injuries of the cord that affect primary afferent pathways. Particular emphasis is given to primate studies, as these have greatest clinical relevance. However, a great deal has been accomplished in nonprimate mammals, particularly in the field of cellular mechanisms, and the reader is encouraged to read related chapters (see end of article) for a broader perspective.

Somatosensory Pathway Organization and the Hand

The primary somatosensory pathway comprises the afferent projections of the sensorimotor system, from peripheral receptors (mechanoceptors, thermoreceptors, nociceptors, etc.) to the cortex. Information received by the receptors of the skin, muscles, tendons, and joints of, for example, the left hand, is transmitted via the dorsal roots to the dorsal horn of the spinal cord. The dorsal column projection relays in the left cuneate nucleus of the brain stem, and crosses the midline in the medial lemniscus to terminate within the contralateral ventral posterolateral nucleus (VPLc) of the thalamus. The spinothalamic neurons have their cell bodies in the cervical dorsal horn and cross the midline within 1–2 spinal segments before terminating in the same thalamic VPLc nucleus. Spinothalamic fibers also terminate more caudally in the anterior nucleus of the pulvinar (Pulo) of the thalamus, and more rostrally in VLp, the posterior nucleus of the ventral lateral complex. Thalamocortical projections from the thalamic VPLc nucleus then terminate within the primary somatosensory cortex (S1) (Brodmann's areas 3a, 3b, 1, and 2). Projections from the Pulo also terminate within these regions, the posterior parietal cortex, and even the caudal primary motor cortex. Figure 1 shows these pathways at their simplest level of organization. In reality, additional afferent projections transmit sensory information to higher levels (e.g., spinoreticular, spinocerebellar, spinocervical), and feedback loops from higher levels of the pathway can be found at every step (Figure 1). For example, the somatosensory cortex (hand region) projects directly to the cuneate nucleus, and to the dorsal horn of the cervical spinal cord, as part of the corticospinal tract. In the execution of fine manipulative tasks that involve continuous tactile and proprioceptive feedback from the hand, many of these pathways and feedback loops are continuously involved. Somatosensory and motor systems are also intrinsically linked with the somatosensory pathways, since the effective execution of directed exploratory movements of the digits, for example, will depend on continuous sensory feedback from the hand.

As with the retina of the visual pathway, and cochlea of the auditory system, the skin and fascia of the body surface are represented at each level of the somatosensory pathway. These representations are ordered or somatotopic, such that inputs from adjacent parts of the body surface are always located adjacent to each other at subsequent levels of the pathway (e.g., within the cuneate nucleus, the VPLc nucleus, and the S1). The primate hand is like the retinal fovea, in providing the sharpest somatosensory image of the world within reach. As such, the digits of the primate hand have a greater mechanoreceptor density than elsewhere, and the glabrous skin of the digit pads is represented disproportionately within S1, and subcortical nuclei. Representations can be mapped at any level of the pathway by recording from neurons, and by determining their receptive fields by stimulating the skin surface. It is these maps that have been used extensively over the last two decades to explore the mutability of the adult brain following injury, disease, and other manipulations (e.g., learning paradigms).

In adult mammals under normal conditions, these maps are stable over time. However, when the activity of the constituent neuron populations alters significantly, following, for example, a repeated learned behavior, or an injury that blocks normal input from peripheral receptors, then these central maps reorganize to adjust to the new circumstances. Since the early demonstration of cortical remodeling following a median nerve cut or digit amputation, there have been countless demonstrations of this phenomenon in all sensory and motor systems and in a broad range of mammalian species. Though first observed in the 1960s in the spinal dorsal horn, the finding that reorganization also occurs at subcortical levels has only fully been appreciated in the last decade. There may be clear reasons for this. Earlier studies generally focused on the cortical map because it represents the last port of call for the primary pathway, and is easily accessed for electophysiological recordings. Another reason may reflect differences in the organization of visual and somatosensory pathways, and a certain

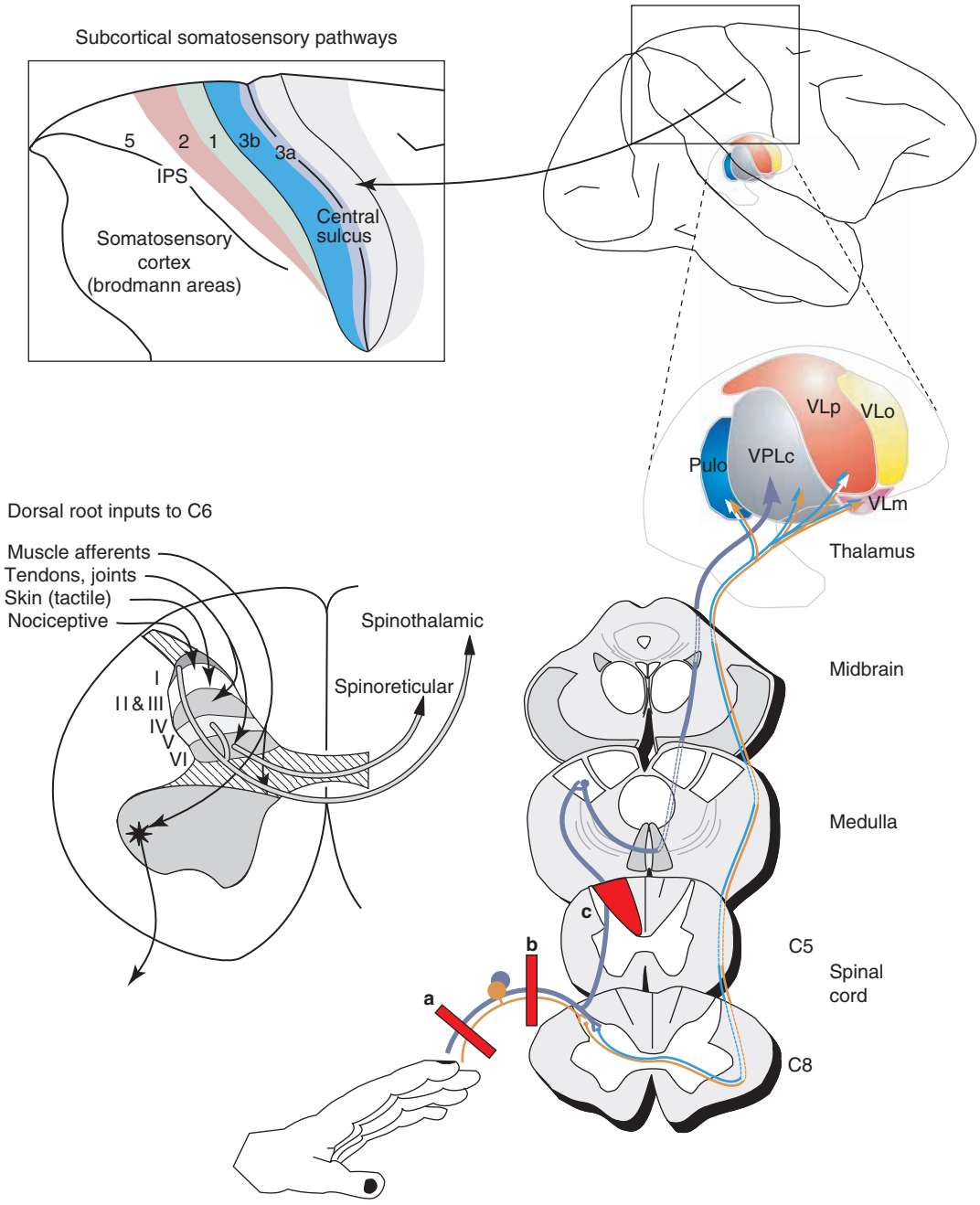

Figure 1 The primary somatosensory pathway showing the major afferent projections, and the different levels at which reorganization can occur following injury. a, b, and c indicate the three levels at which primary afferent neuronal populations can be severed following a peripheral nerve (a), dorsal root (b), or central dorsal column (c) injury. Inset shows terminal distributions for the different primary afferent subpopulations. In the macaque monkey and human, the primary somatosensory cortex (S1) comprises four regions, 3a, 3b, 1, and 2, each named according to Brodmann (1908). Adapted from Darian-Smith and Darian-Smith (2004).

inertia to fully appreciate the differences in these systems. In the visual pathway, there is a small receptor sheet (i.e., the retina) and tight afferent projection to the lateral geniculate nucleus (of the thalamus) and cortex. This means that following a retinal lesion, the most dramatic plastic changes have been shown to occur at the level of the cortex. In contrast, the receptor sheet for the somatosensory pathway is comparatively vast and multilayered (skin, muscles, and joints), and there is a spinal cord with its complex distributed parallel circuitry. This provides a greater opportunity for significant circuit reorganization at multiple subcortical levels. The central nervous system (CNS) appears to be opportunistic, in the sense

that neuronal reorganization can and will occur along a sensory pathway wherever the circumstances permit, so appreciating the underlying anatomy is critical for understanding where changes can occur.

How the somatosensory system responds depends on where the injury occurs, the extent to which fibers are spared by the lesion (the size of the injury), and the postlesion time period at which changes are examined.

With respect to the postlesion time period, there are several stages to the process of postinjury reorganization. These are generally divided into an acute phase (minutes to hours), a chronic phase (days, weeks, and early months), and an extended chronic phase (years). For example, immediately following a digit amputation or median nerve cut, there will be a 'silent region' within the corresponding contralateral cortical hand map, and a rapid expansion of the representation of parts of the hand lying adjacent to the silenced or deafferented region. This expansion can occur within minutes to hours, and may persist for weeks to months. During the chronic second stage (weeks to months), there is opportunity for the establishment of new connections from spared inputs and the consolidation of the cortical and subcortical body maps. A final stage (years) involves the eventual atrophy of chronically deafferented neuronal populations. This was evident more than a decade after monkeys received a dorsal rhizotomy that deafferented the forelimb. The cuneate nucleus of the brain stem and the VPLc nucleus of the thalamus reduced substantially in volume over this period, and an expanded face representation was observed within the somatosensory cortex. This suggests that when the deafferentation is complete and extensive, consolidation involving cell shrinkage and death can occur slowly over many years, and even decades.

Lesions of the Primary Afferent Neuron

The patterns of deafferentation that result from injuries to primary afferent neuron populations relaying sensory information from the hand to the cortex have been studied extensively. They include investigations into peripheral nerve injury, amputations, dorsal rhizotomies, and more central injuries to the spinal cord and dorsal column.

Whether the primary afferent neuron is injured peripherally, or centrally to the dorsal root ganglion, or centrally within the CNS (e.g., in the dorsal column), does make a difference to the postinjury changes that take place (Figure 1). The following sections briefly summarize what happens when injuries involve different components of the primary afferent, as shown in a–c (Figure 1). In the somatosensory pathway, subcortical structures have been shown to play a critical role in the postinjury responses, and this multilevel involvement is now generally accepted. For this reason, multilevel changes are discussed throughout the text and are not addressed separately.

Peripheral Nerve Injuries and Amputations

Peripheral nerves retain the capacity to regrow following injury, and will do so unless blocked by a physical or chemical barrier, so the central map changes that result following nerve cut are often relatively minor. Even so, studies in the macaque monkey have shown that if a nerve is cut and the cut ends sutured to enable nerve regeneration, the cortical somatotopic map may never completely recover. Many neurons in the reorganized cortex develop abnormal multiple receptive fields on nonadjacent skin regions, and the map becomes disorganized. Such mislocalization does not occur following nerve crush, where fibers regenerate to innervate their original peripheral target.

When nerves are unable to regenerate to their original target (e.g., following nerve ligation), reorganization is far more substantial. Following nerve transection, a 'silent zone' is initially observed within the corresponding cortical map where the digits and/or skin surface have been deafferented. Within minutes to hours, the 'silent zone' is surrounded by an expanded representation of the adjacent body parts, and over a period of weeks and months, there is a gradual 'filling in' or expansion of the representation of the adjacent skin surface until the new organization stabilizes. A specific example of this is seen in the squirrel monkey, where a median and ulnar nerve section results in a corresponding deprived region of cortex, that comes to respond over many months to inputs originating from the back of the hand.

Following the amputation of the hand or digits, where afferent and efferent nerves are both cut, the pattern of cortical reorganization mimics that seen following the permanent section of both sensory and motor nerves supplying the hand. Thus, following digit amputation, the representation of the stump region (or adjacent hand) expands into the deprived region of the cortical map. This can take weeks to months to occur, and parts of the cortex may remain silent during the postinjury period.

Similarities between amputation and nerve ligation diverge, most noticeably, when the amputation is large. For example, when the whole of the forelimb is amputated, there is some expansion of the face representation (which lies adjacent to hand representation) into the hand cortex. Imaging studies in humans with hand or forearm amputations have similarly reported face responses within the deprived

hand cortex many months to years following the amputation. Though the neural substrate for these observations remains unclear, the longer time frame suggests that local neuronal growth, synapse formation and modification, and perhaps even some transneuronal atrophy (where years have elapsed since the injury), may contribute.

In studies of nerve section or amputation, changes observed within the cortex reflect subcortical reorganization. Cutting nerves supplying the hand in the squirrel monkey, for example, results in the reorganization of the 'hand' map in both the brain stem cuneate nucleus and the thalamic ventroposterior lateral nucleus. Reorganization within the brain stem and thalamus was found to be similar in extent to that observed in the cortex of these monkeys.

Dorsal Rhizotomy – Root Cut

More than a century ago, Mott and Sherrington showed that cutting dorsal roots C4-T2 causes an extensive deafferentation of the macaque's hand and forearm, and a profound and lasting impairment of voluntary manipulative movements. In humans, the loss of sensory feedback from the hand (e.g., following brachial plexus injury), also typically results in a severe impairment in dexterity.

Recent investigations in our lab provide insight into a number of acute, chronic, and long-term changes (>8 months) that accompany experimentally induced dorsal root lesions. When dorsal roots that innervate the digits of the hand are sectioned, the central axons of the primary afferents that are cut, degenerate (while their cell bodies within the dorsal root ganglion and peripheral axons remain alive). In our experiments, only dorsal rootlets innervating the first two to three digits of one hand were cut (**Figures 2 and 3**). Immediately following one of these lesions, there is no detectable input from the experimentally 'deafferented' digits, and a 'silent region' is present in the corresponding spinal dermatomal map (**Figure 2(a)**), and in somatotopic maps at the level of the brain stem (**Figure 4**), and somatosensory cortex. These silent zones persist for many weeks at all levels of the pathway and correspond to an initially severe behavioral deficit (see below; **Figures 2 and 3**). Gradually, over the course of many weeks, and with constant hand use, there is some recovery of digit use (see 'Behavioral consequences of somatosensory plasticity'), and as this occurs there is a corresponding reorganization within the cuneate nucleus of the brain stem and cortex. Within dorsal root fascicles immediately adjacent to the lesion, cutaneous responses can also be elicited where this was not previously possible.

To investigate the neural basis of these changes, we conducted a series of anatomical experiments in monkeys with a similar lesion. It was shown that a small population of afferent fibers was spared by the lesion that continued to innervate the 'deafferented' digits. These entered the cord adjacent to the lesion site, and for many weeks following the lesion, they were apparently too few in number or too weak in their response to drive the somatotopic map at any level of the pathway. We showed that these primary afferent fibers sprout locally within the spinal dorsal horn and cuneate nucleus over the postlesion period. These 'sprouts' are likely to form additional connections with input-deprived postsynaptic target neurons (**Figure 5**) and play a role in driving reorganization at the spinal and higher levels of the pathway.

Our studies also demonstrate a limit to the amount of reorganization and recovery that can occur. Quite extensive postlesion reorganization can occur in the spinal dorsal horn (and at higher levels) as long as a small number (<5%) of uninjured primary afferents remain intact, that continue to transmit information from the periphery. For example, in monkeys with lesions that removed most input from the thumb and index finger, there was a complete re-emergence of the cortical map over the first 3–4 postlesion months (**Figure 2(b)**). However, in macaques with a slightly larger lesion (two to three segments; **Figure 3(a)**), the hand map never completely returned. This was true for the cuneate nucleus (**Figure 4**) and cortex (**Figures 2(a) and 3(a)**).

Earlier studies that provided insight into dorsal root injury in nonhuman primates focused on a group of macaque monkeys that received a massive dorsal root section (C2-T4) some 12–20 years earlier. These lesions completely deafferented the forelimb and, in some cases, part of the trunk, and the arm was rendered completely nonfunctional. These studies demonstrated several things. First, that chronic deafferentations of such a large magnitude result in significant transneuronal degeneration over many years. As might be expected, there was an extensive reorganization of the body map within the thalamic nucleus VPLc, and within the somatosensory cortex. In the cortex, no unresponsive zones were detected, but the face was represented over a larger region and abutted the trunk representation in the absence of any input from the deafferented limb. There was also significant atrophy within the cuneate fasciculus of the spinal cord, the cuneate nucleus of the brain stem (37–48% shrinkage by volume), the thalamic VPLc nucleus, and possibly even the somatosensory cortex, though the latter was not determined. Some atrophy in the cortex within the original forelimb

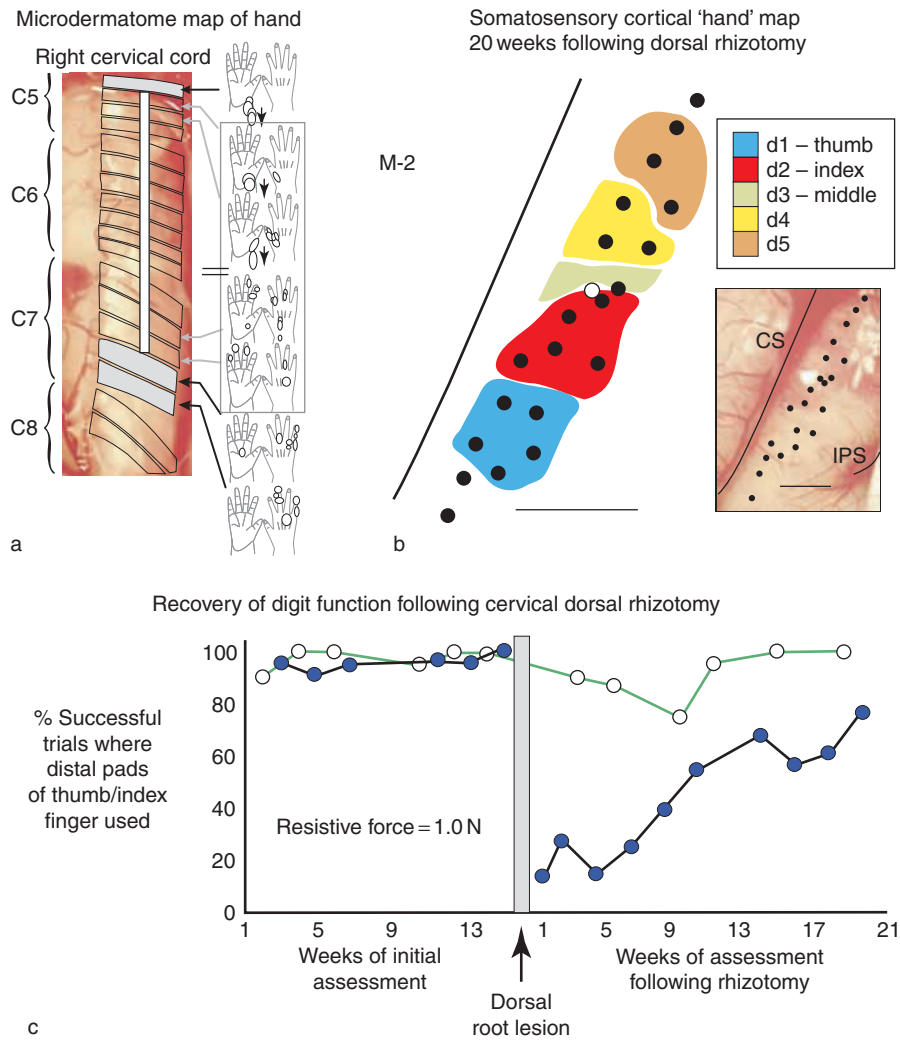

Figure 2 (a) A laminectomy showing a dorsal rootlet lesion in which tactile input was selectively removed from the thumb and index fingers. Cartoons of the hand show the distribution of cutaneous receptive fields on the palmar and dorsal surface of the hand in rootlets along the borders of the lesion zone, before the rootlets were cut. White line indicates which rootlets were sectioned and gray shaded rootlets identify fascicles that were left intact following the lesion. Afferent fibers within these rootlets altered in their response over several postinjury months such that cutaneous RFs could be mapped on the deafferented digits where this was not previously possible. See text. (b) The corresponding cortical 'hand map' for this monkey, 20 weeks following the lesion. By this time, a complete hand map had 're-emerged' and a silent region was not observed. Black dots indicate recording sites, and white dot identifies a site where no response could be elicited. (c) The accompanying recovery of performance of a reach–retrieval task in the same monkey over the course of 20 postinjury weeks. The opposition of the pads of the index finger and thumb was the preferred stratagem used by the monkey before the lesion, and in the control hand (open circles). The use of this stratagem was lost immediately after the lesion (filled circles) but was gradually recovered over 20 postlesion weeks. The time of the dorsal rootlet section is indicated with the gray line. The 1.0 N resistive force refers to the clamp force holding the target object. Adapted from Darian-Smith C and Ciferri M (2005) Loss and recovery of voluntary hand movements in the macaque following a cervical dorsal rhizotomy. *Journal of Comparative Neurology* 491: 27–45.

representation would help explain the cortical map changes. Perhaps surprisingly, only 5–15% of neurons were lost within the cuneate nucleus even up to 20 years post injury. This indicated that shrinkage by volume was associated with a loss of neuropil and axons, rather than a loss of neurons.

Spinal Cord – Dorsal Column Lesions

Injuries that involve the primary afferents at a central level provide additional insight into the role that spared primary afferents can play in reorganization. In one series of experiments in owl monkeys, a high dorsal column lesion (at C3–4) removed inputs from

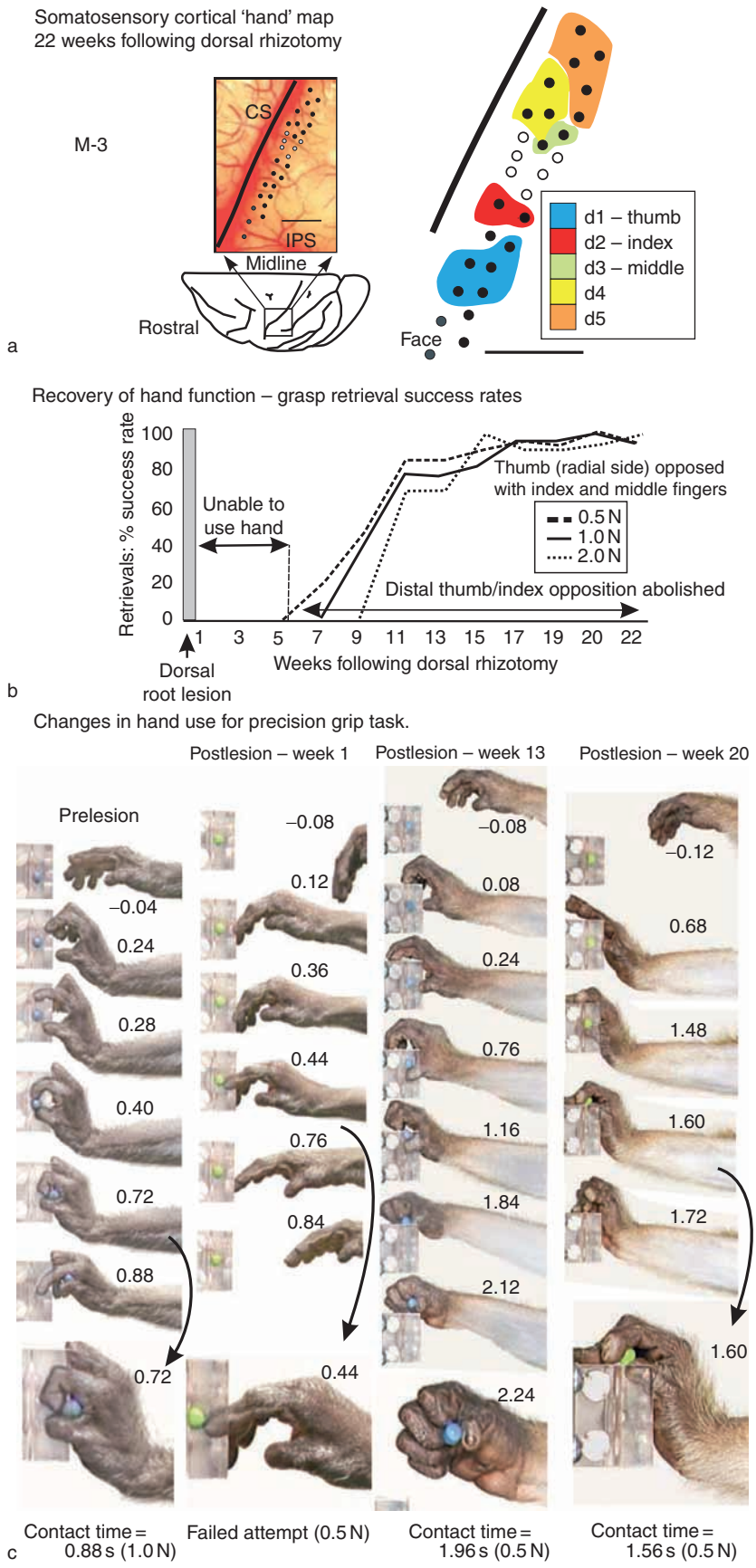
Somatosensory cortical 'hand' map
22 weeks following dorsal rhizotomy
M-3
CS
IPS
Midline
Rostral
Face
d1 – thumb
d2 – index
d3 – middle
d4
d5
a
Recovery of hand function – grasp retrieval success rates
Retrievals: % success rate
100
80
60
40
20
0
Unable to use hand
Thumb (radial side) opposed with index and middle fingers
0.5 N
1.0 N
2.0 N
Distal thumb/index opposition abolished
1 3 5 7 9 11 13 15 17 19 20 22
Weeks following dorsal rhizotomy
Dorsal root lesion
b
Changes in hand use for precision grip task.
Postlesion – week 1
Postlesion – week 13
Postlesion – week 20
Prelesion
–0.04
0.24
0.28
0.40
0.72
0.88
0.72
–0.08
0.12
0.36
0.44
0.76
0.84
0.44
–0.08
0.08
0.24
0.76
1.16
1.84
2.12
2.24
–0.12
0.68
1.48
1.60
1.72
1.60
Contact time = 0.88 s (1.0 N)
Failed attempt (0.5 N)
Contact time = 1.96 s (0.5 N)
Contact time = 1.56 s (0.5 N)
c

the forearm and lower body from the entire extent of the dorsal column nuclei of the brain stem. Though the spinothalamic pathway remained intact, these neurons were not able to independently transmit sufficient afferent information from the forelimb to reactivate the cortical map. When even a small afferent population was spared by the lesion, this input was found to reactivate the corresponding cortex and to expand its representation into the cortical map. When no afferents were spared within the dorsal columns, preserved inputs from the face appeared to grow into denervated regions of the cuneate nucleus from the adjacent trigeminal nucleus. This took 6 months to occur. Within the cortex, face inputs expanded to utilize the adjacent silent 'forelimb' cortex. This study suggested that the spinothalamic pathway does not play a defining role in the cortical reorganization observed, though some controversy regarding the compensatory role of mechanoreceptor afferents within the spinothalamic tract remains.

In humans, spinal injuries are rarely restricted to the dorsal column, so it is difficult to assess the neural bases for acute or chronic changes following injury.

Behavioral Consequences of Somatosensory Plasticity

Surprisingly, few nonhuman primate studies have systematically examined behavioral changes following somatosensory pathway injuries, and linked these back to neuronal reorganization.

Several investigations have looked into the effects of dorsal column lesions on hand and digit use over the last several decades, but these have been inconsistent in their findings. Deficits described have ranged from subtle and insignificant to chronic and debilitating. A likely reason for the range of results stems from the lack of injury reconstruction at the completion of many of these studies, and thus a lack of correlation between the anatomical and behavioral deficits. In studies by Vierck and colleagues, who did make this link, transection of the dorsal spinal column was found to initially impair the monkey's ability to discriminate textures, or specifically the frequency or duration of tactile stimulation. Some monkeys recovered this ability over several months of testing, and others did not. In the monkeys that did recover, it was proposed that monkeys learned to use alternate cues and pathways to assess and discriminate textures.

The behavioral effects of a spinal hemisection, which also involves the dorsal column, in addition to the corticospinal tract and other ascending and descending pathways on the lesion side of the cord, have also been studied in the macaque. Monkeys with a near-complete hemisection at spinal segments C3–4 initially presented with a severe hemiparesis, but regained the ability to pick up objects with the impaired hand within 1 month of the lesion. By the end of the third postlesion month, monkeys were able to use their hand and digits quite efficiently in a reach–grasp–retrieval task, though recovery was never fully complete (up to 3 years post injury). Persistent impairments included a less direct trajectory and preshaping of the hand during the reach component of the task, and some weakness of the opposing digits. The somatosensory component of the task was also clearly impaired, but not as completely as might have been expected. Somatotopic maps of the body representation showed a considerable reorganization of the deprived cortex and a greatly reduced hand map with abnormally large cutaneous receptive fields. In these studies, previously existing 'crossover' corticospinal projections were thought, at least in part, to mediate behavioral recovery. Localized sprouting of these or other pre-existing pathways may also have played some role. In all cases, recovery of hand function likely involved significant circuitry changes and synapse modification (see Mechanisms below).

Mott and Sherrington showed in 1895 that sectioning the dorsal roots of spinal segments C4-T4 in the macaque completely blocked somatic sensation in the hand and voluntary digit movements, while sparing some controlled movements of the elbow and

Figure 3 Data from a monkey that received a larger lesion than that shown in **Figure 2**. Here the dorsal root section removed all detectable sensory input from the thumb, index, middle finger, and thenar palm of one hand. This monkey was initially unable to use its hand at all and could not successfully perform a reach–retrieval task requiring cutaneous feedback for more than 2 months. (a) The partial re-emergence of the somatosensory cortical hand map in this monkey 22 weeks following the dorsal root lesion. Note that while there was a significant re-emergence of the digit map, a silent region remained. (b) The rate of success of retrieving a target reward as a function of time following the lesion. Data are given for three resistive forces (clamp forces holding the target), that the monkey had to overcome to retrieve the object. These correlate to increasing levels of difficulty with 2.0 N force being the most difficult and last to be successfully achieved. (c) Changes in the strategy used to perform the grasp–retrieval task. The prelesion stratagem, shown in column 1, shows the target being grasped and removed by the opposing index finger and thumb. At 1 week post lesion, the monkey failed to perform the task and was unable to feel the clamp or target object, or to preshape the digits. By the 13th postlesion week, the object could be grasped using an abnormal use of the first three digits, and by the 20th postlesion week, cutaneous sensation had recovered sufficiently that the monkey was able to retrieve the object 100% of the time, using opposition of the thumb and index fingers. Adapted from Darian-Smith C and Ciferri M (2005) Loss and recovery of voluntary hand movements in the macaque following a cervical dorsal rhizotomy. *Journal of Comparative Neurology* 491: 27–45.

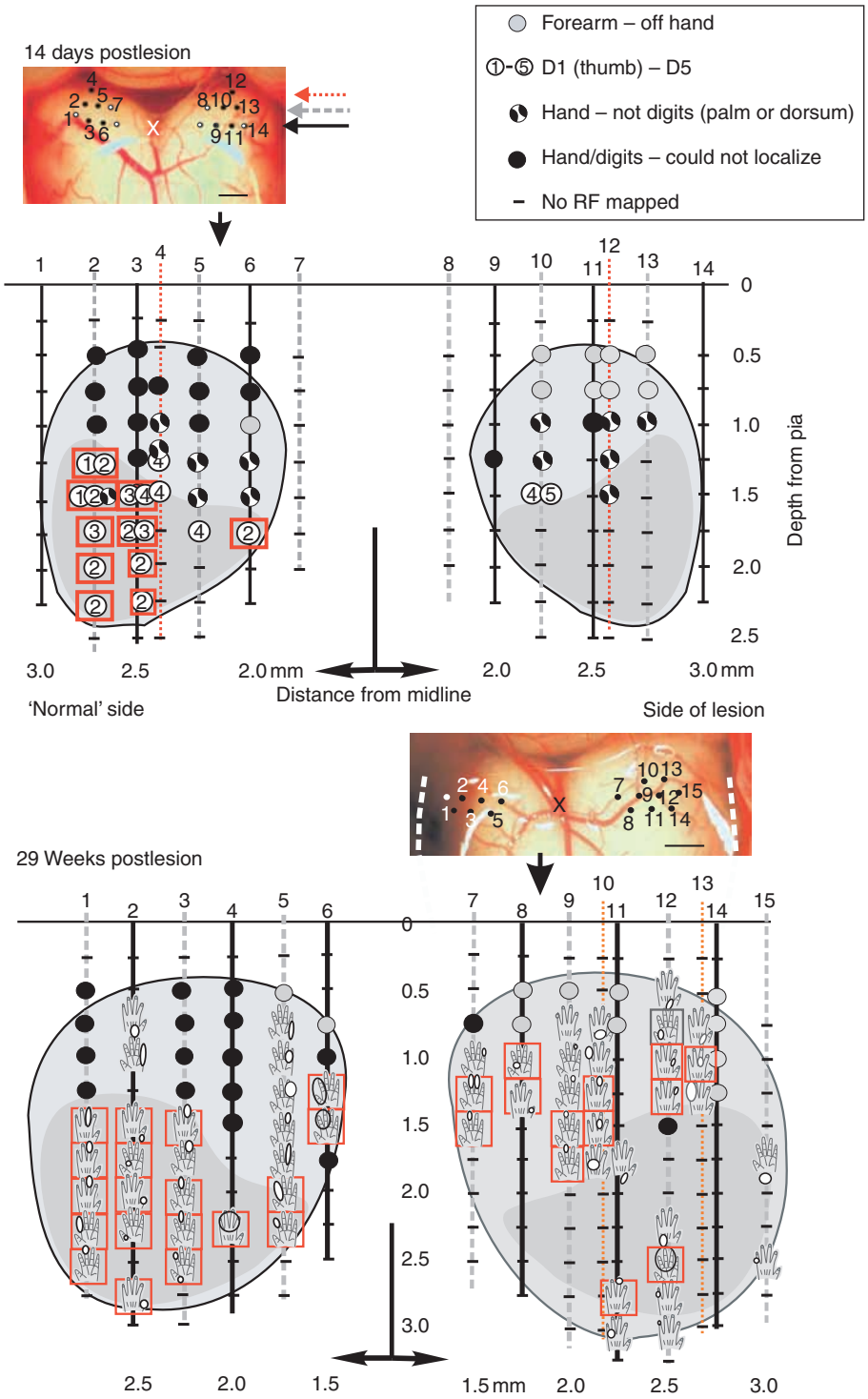

Figure 4 Changes in the pars rotunda of the cuneate nucleus of the brain stem of monkeys that received a DRL that removed input from the thumb, index, and middle fingers. Two weeks following the lesion, no inputs could be mapped within the cuneate nucleus on the first three digits of the hand. Twenty-nine weeks following the lesion, a silent region remained, but surrounding parts of the cuneate nucleus now responded to stimulation of the initially 'deafferented' digits. Monkeys with slightly smaller lesions also showed a substantial reorganization of the hand map in the cuneate >16 weeks following the lesion, but did not have a 'silent zone'. Electrode penetration sites are marked on photographs of the dorsal surface of the brain stem (insets). Different rostrocaudal rows of penetrations are identified with different line types. In upper panel, cutaneous RFs are shown according to their location, on the digits (D1–D5), the hand, and the forearm, as indicated in the legend. In the lower panel, RFs on the digits are shown on cartoons of the hand surface. Shaded regions in the cuneate map on the normal side outline the distribution of D1–D3 RFs. This input zone is transposed to the side of the lesion to help define the original region of input of the experimentally deafferented digits. Scale bar in photos = 1 mm; numbers refer to electrode penetrations. Adapted from Darian-Smith C and Ciferri M (2006) Cuneate nucleus reorganization following cervical dorsal rhizotomy in the macaque monkey: Its role in the recovery of manual dexterity. *Journal of Comparative Neurology* 498: 552–565.

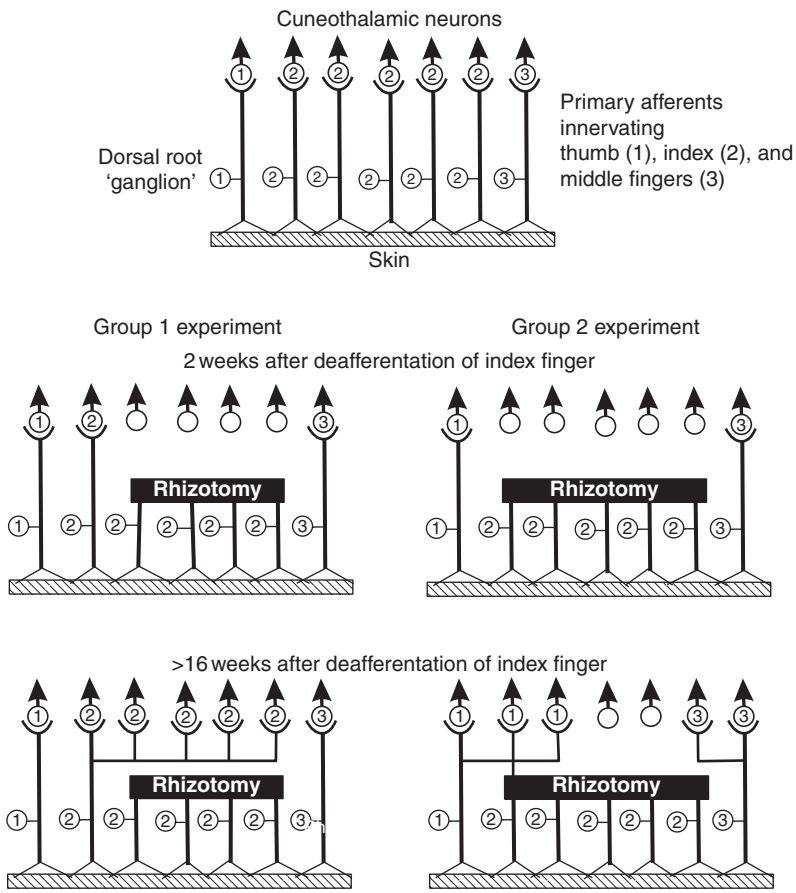

Figure 5 Schematic of projections from macaque's index finger (and also from the adjacent thumb and middle fingers) to neurons in the cuneate nucleus, before and following the section of selected dorsal rootlets: a summary of the findings of Darian-Smith C and Ciferri M (2006) Cuneate nucleus reorganization following cervical dorsal rhizotomy in the macaque monkey: Its role in the recovery of manual dexterity. *Journal of Comparative Neurology* 498: 552–565. In group 1 experiments (column on left), the rhizotomy was extensive, but spared ~5–10% of neurons innervating the index finger. However, in group 2 experiments (right column), following a more extended rhizotomy, the index finger was completely deafferented, which is reflected in the representational map of the hand (open, unlabeled circles). Sixteen weeks later (bottom row), the cuneate circuitry had changed profoundly: (a) the central terminals of the axotomized neurons had degenerated, and (b) the central terminals of primary neurons innervating the thumb or middle finger, and those spared neurons that innervated the index finger, had proliferated ('sprouted'), and some had established sufficient synapses on cuneothalamic neurons to re-establish the cuneate representational map of the thumb and fingers (bottom left panel). In group 2 monkeys, where the index finger was completely deafferented (bottom right panel), there was extensive proliferation of afferent terminals in the cuneate nucleus, though this could not compensate for the permanent loss of input from the index finger: the cuneate representation of these monkeys remained 'silent'. As a consequence, group 1 monkeys substantially recovered a precision grip over a period of weeks to months, whereas group 2 monkeys did not. Reorganization of cuneate circuitry was matched by cuneate axon terminal labeling (Darian-Smith C (2004) Primary afferent terminal sprouting after a cervical dorsal rootlet section in the macaque monkey. *Journal of Comparative Neurology* 470: 134–150) which suggested local sprouting of spared axon terminals, and the monkey's recovery of hand function (Darian-Smith C and Ciferri M (2005) Loss and recovery of voluntary hand movements in the macaque following a cervical dorsal rhizotomy. *Journal of Comparative Neurology* 491: 27–45 – see **Figures 2** and **3**). From Darian-Smith C and Ciferri M (2006) Cuneate nucleus reorganization following cervical dorsal rhizotomy in the macaque monkey: Its role in the recovery of manual dexterity. *Journal of Comparative Neurology* 498: 552–565.

shoulder. They also observed that if a single dorsal root was cut in the same range of cervical segments, there was no identifiable impairment of voluntary hand function. In more recent studies of reaching movements following cervical dorsal root transection, some return of gross limb movement was reported in the monkey forelimb. Vierck (1982) studied the effects of a dorsal rhizotomy (C3-T3) on forelimb reaching to points in space and found an initially severe deficit, followed by a gradual recovery of some directed limb movements, after many months of training. However, voluntary hand function was not systematically examined.

Recently, in our lab, we examined the behavioral effects of a restricted cervical dorsal rootlet lesion on the macaque's ability to perform a voluntary task requiring cutaneous feedback. Monkeys were divided into two groups, and trained to retrieve a target

object from a clamp (Figure 3(c)), using the thumb and index finger in opposition. In group 1, dorsal rootlets containing fibers supplying the thumb and index finger were cut, and in group 2, additional rootlets innervating the middle finger and thenar eminence were also cut. Performance parameters were assessed before and following the lesion, and included (1) percentage of successful retrievals, (2) digital stratagem (the pattern of digit opposition), and (3) the contact time (or duration of digit contact with the object prior to retrieval). During the first postoperative week, hand function was severely impaired in all monkeys, but over the following weeks, group 1 monkeys recovered the ability to retrieve the object by opposing the index finger and thumb in >80% of trials. Figure 2 shows the gradual recovery of the prelesion digit use over 21 postlesion weeks in one such monkey. Group 2 monkeys with the larger lesion also regained function in the impaired hand. Though this recovery was sometimes dramatic, as in the case of M-3 (Figure 3), it was never complete. Each of these monkeys quickly adopted a stratagem for grasping the target, using digits that were incompletely deafferented. At the end of these experiments, hand representations were mapped within the contralateral somatosensory cortex, and in the cuneate nucleus, to define hand deafferentation and cortical reactivation. In all cases, a tight correspondence was observed between these somatotopic maps and digit use. For example, if the monkey regained function in the thumb and middle finger but never again used the index finger, digit maps consistently showed a re-emergence of input from the thumb and middle finger, and a 'silent zone' for the index finger.

The behavioral studies described above, which focus on hand sensorimotor function, consistently demonstrate that wherever small numbers of somatosensory fibers are spared, quite dramatic behavioral recovery can occur during the early postinjury months.

Neural Substrates for Reorganization in the Somatosensory Pathways

Pathway Divergence and Disinhibition

The somatosensory pathways are divergent and hierarchical in their organization, from their point of entry to the spinal cord, to the brain stem nuclei, the thalamus, and cortex (Figure 1). This divergence is undoubtedly responsible for some plasticity within the system, particularly during the acute phase of the injury. This is well demonstrated in the cortical map, immediately following a peripheral nerve cut, where a number of groups have shown that receptive fields from adjacent body parts enlarge and 'fill in' the edges of the deafferented (silent) cortex (up to 1 mm). This can occur within minutes to hours and is enabled by the normal divergence of thalamocortical afferents (up to 2 mm) within the cortex, and the 'unmasking' or 'disinhibition' of inhibitory connections that normally hold these inputs (and receptive fields) in check. Unmasking during the acute phase occurs when excitatory inputs to inhibitory neurons within the cortex are removed by the lesion. This removes local GABAergic inhibition to excitatory synapses within adjacent cortex, and 'unmasks' the underlying pre-existing circuitry (GABA, gamma-aminobutyric acid). Available data also indicate that disinhibition can operate at subcortical levels. The demonstration that GABAergic pathways are modulated following peripheral lesions (or by pharmacological blockade) demonstrates the importance of this neurotransmitter in cortical and subcortical plasticity.

Terminal Sprouting and Synaptic Changes

When a peripheral injury is large, then additional mechanisms are recruited. These processes can involve structural modifications, such as local axonal branching or sprouting, and/or structural alterations of synapses, all of which take weeks to months to occur.

Central axon terminal sprouting has been reported in monkeys, rats, and cats following amputation or nerve transection. It has also been observed within the visual and somatosensory cortex, following retinal and peripheral nerve lesions, respectively. These findings suggest that central axonal growth over short distances, and perhaps even dendritic branching, may be something of a generalized injury induced response, which can occur whether or not remaining neurons are damaged or left fully intact following the injury.

Following the dorsal root lesion described in previous paragraphs, primary afferents that are spared by the injury sprout locally within the spinal dorsal horn and cuneate nucleus. This local terminal exuberance takes several months to occur and may underlie circuitry reorganization that leads to the eventual 'recovery' of dexterity (see above). It is thought to result in the formation of additional synapses with deprived postsynaptic neurons (Figure 5). One possibility is that this amplifies an initially weak signal from the periphery to suprathreshold levels, and may involve an LTP-like process. This provides a neural substrate for the complete or partial re-emergence of subcortical and cortical maps. Importantly, if the initial deafferentation of one or more digits is complete, then no subsequent proliferation

of primary afferent terminals can compensate for this loss. Figure 5 summarizes our recent findings, and underscores the limits of this reorganization.

Terminal growth is rarely observed to extend much beyond the original axonal territory, though there are exceptions to this rule. For example, one study showed that in monkeys with dorsal column lesions or arm amputation, 'face' afferents sprouted and grew into the cuneate nucleus from the trigeminal nucleus of the brain stem.

Many studies, primarily in the cortex, have shown that synaptic strength is greatly influenced by experience or learning. This activity dependence, which appears to follow Hebbian rules, is important for altering synaptic efficacy, at all levels of the somatosensory pathway following peripheral nerve and spinal cord injury. LTP, which is normally associated with learning in the hippocampus, has been demonstrated in rodent sensory and motor cortex, and either LTP or an LTP-like mechanism is likely to contribute to somatosensory plasticity, including plasticity that occurs following skill learning.

In addition to structural and even synaptic changes, damage to the CNS initiates a cascade of cellular and molecular events, many of which play a role in the regenerative and degenerative processes that follow.

Atrophy, Transneuronal Atrophy, and Degeneration

Atrophy and degeneration (e.g., Wallerian) occur in any chronic peripheral nerve or central injury that involves the death of axons entering, or coursing through the spinal cord.

In monkeys that received large-scale dorsal root injuries (C4-T2), 12–20 years earlier, transneuronal atrophy was considerable, and occurred at the level of the cuneate nucleus and the thalamus. In these monkeys, the brain stem cuneate and thalamic VPLc nuclei were reduced in overall volume by as much as 50%, though volume loss involved neuropil more than the total neuron numbers. As expected, somatotopic maps were greatly altered in these monkeys, but after this length of time, no inactive or 'silent' regions remained. These studies indicate that following a large peripheral injury, long-term transneuronal atrophy continues over many years and is a slow but dynamic process.

In the shorter term and to some extent, chronically, a number of non-neuronal cell types (e.g., macrophages, oligodendrocytes, microglia, and astrocytes) are recruited and activated. These cells perform many roles in addition to cleaning up and managing the structural integrity of the injury site. In chronic injuries, and most obviously in traumatic injuries of the spinal cord and CNS, astrocytes are known to play a major role in gliosis and scar formation. Activated astrocytes within the scar tissue secrete molecules, that can either permit or inhibit axon terminal growth and synapse formation. Their secretion of inhibitory factors can present a major barrier to axon growth and recovery following traumatic spinal injury, and the effects of these cells on neuronal reorganization forms an important focus of investigation. Their role in somatosensory plasticity, however, is complex, and requires further investigation.

Neurogenesis?

Though neurogenesis is not generally regarded as a contributor to plasticity in the somatosensory system, recent evidence suggests that it may play a potentially important part in the process of pathway reorganization following injury. Neurogenesis is well known to play a major role in hippocampal function, and several groups have reported adult neurogenesis following a number of neuropathologic conditions, including ischemia, and nonpathologic situations including various learning paradigms. Recently, we observed that adult neurogenesis also occurs in the spinal dorsal horn following the restricted dorsal root injury model (described above), and that it accompanies other neurophysiological and anatomical changes within the 'lesion zone' of the spinal dorsal horn. This indicates that neurogenesis occurs within the adult primate spinal cord in response to injury. Emerging evidence from work in the cortex suggests that newly formed neurons in the CNS integrate into local circuitry as GABAergic interneurons. However, the functional significance of neurogenesis in the reorganization of somatosensory pathways following injury is not currently known and awaits further research.

See also: Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Neural Circuitry in the Somatosensory System; Peripheral Nerve Regeneration: An Overview; Reaching and Grasping; Somatosensory Cortex: Functional Architecture; Somatosensory Cortex; Somatosensory Pathways (Ascending): Functional Architecture; Somatosensory Perception; Somatosensory Receptive Fields; Spinal Cord Injuries; Spinal Cord Regeneration and Functional Recovery: Strategies; Synaptic Plasticity: Neuronal Sprouting.

Further Reading

Bradbury EJ, McMahon SB, and Ramer SR (2000) Keeping in touch: Sensory neurone regeneration in the CNS. *Trends in Pharmacological Science* 21: 389–394.

Buonomano DV and Merzenich MM (1998) Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* 21: 149–186.

Darian-Smith C (2004) Primary afferent terminal sprouting after a cervical dorsal rootlet section in the macaque monkey. *Journal of Comparative Neurology* 470: 134–150.

Darian-Smith C and Brown S (2000) Functional changes at periphery and cortex following dorsal root lesions in adult monkeys. *Nature Neuroscience* 3: 476–481.

Darian-Smith C and Ciferri M (2005) Loss and recovery of voluntary hand movements in the macaque following a cervical dorsal rhizotomy. *Journal of Comparative Neurology* 491: 27–45.

Darian-Smith C and Ciferri M (2006) Cuneate nucleus reorganization following cervical dorsal rhizotomy in the macaque monkey: Its role in the recovery of manual dexterity. *Journal of Comparative Neurology* 498: 552–565.

Darian-Smith I, Galea M, Darian-Smith C, Sugitani M, Tan A, and Burman K (1996) The anatomy of manual dexterity: The new connectivity of primary sensorimotor thalamus and cerebral cortex. In: *Advances in Anatomy, Embryology and Cell Biology,* vol. 133, pp. 1–142. New York: Springer.

Ding Y, Kastin AJ, and Pan W (2005) Neural plasticity after spinal cord injury. *Current Pharmaceutical Design* 11: 1411–1450.

Jain N, Florence SL, Qi HX, and Kaas JH (2000) Growth of new brain stem connections in adult monkeys with massive sensory loss. *Proceedings of the National Academy of Sciences of the United States of America* 97: 5546–5550.

Jones EG (2000) Cortical and subcortical contributions to activity-dependent plasticity in primate somatosensory cortex. *Annual Review of Neuroscience* 23: 1–37.

Kaas JH (2002) Sensory loss and cortical reorganization in mature primates. *Progress in Brain Research* 138: 167–176.

Merzenich MM, Kaas JH, Wall JT, Sur M, Nelson RJ, and Fellerman DJ (1983) Progression of change following median nerve section in the cortical representation of the hand in areas 3b and 1 in adult owl and squirrel monkeys. *Neuroscience* 10: 639–665.

Mott FW and Sherrington CS (1895) Experiments upon the influence of sensory nerves upon movement and nutrition of the limbs. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 57: 481–488.

Wall JT, Xu J, and Wang X (2002) Human brain plasticity: An emerging view of the multiple substrates and mechanisms that cause cortical changes and related sensory dysfunctions after injuries of sensory inputs from the body. *Brain Research Reviews* 39: 181–215.

Vierck C (1982) Comparison of the effects of dorsal rhizotomy or dorsal column transection on motor performance of monkeys. *Experimental Neurology* 75: 566–575.

Vierck CJ and Cooper BY (1998) Cutaneous texture discrimination following transection of the dorsal spinal column in monkeys. *Somatosensory Motor Research* 15: 309–315.

Somatosensory Receptive Fields

S S Hsiao, P J Fitzgerald, P H Thakur, P Denchev, and T Yoshioka, The Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A marvel of human perception is the speed and accuracy with which humans recognize and manipulate objects with their hands. This is supported by psychophysical studies showing that blindfolded subjects perform at an accuracy rate of greater than 96% and mean response times of less than 5 s when asked to identify 100 common objects that they explore using their hands. It is also supported by studies of proficient braille readers who can easily read braille text at rates greater than 100 words per minute. These findings demonstrate that it takes less than 100 ms to recognize the spatial form and texture of patterns presented to the skin. One hundred milliseconds is very rapid and suggests that there is a high-capacity spatial processing system in touch that initially encodes a peripheral neural representation of patterns on the skin, transmits those representations to the cortex, and within cortex transforms and matches those representations against all previously stored memories of objects. The process must be highly dynamic since object identification is facilitated by dynamic exploration of the pattern, and must be performed by massively parallel processes. In touch, the encoding and processing of information begins with the activation of populations of neurons with receptors in the skin that have small receptive fields. Neurons further along the pathways have fields that are larger and more elaborate as the information progresses centrally toward regions in cortex involved in perception. The concept that neurons have a receptive field was originally proposed by Sherrington (in 1906), who coined the terms exteroceptive and interoceptive fields for distinguishing between neurons that respond to internal states of the body versus neurons that respond to external environmental stimuli. Sherrington proposed that the selectivity of the exteroceptive neurons was due to the particular accessory structures that are associated with each type of neuron. Sherrington also proposed that understanding the specific relationship between external stimuli and the neural responses was a key in understanding the basis of perception. The importance of receptive fields in neural processing of visual stimuli was further elaborated in the early 1930s with studies by Hartline and others who showed that neurons exhibit selectivity for specific kinds of visual stimuli and inferred that the specificity was based on neurons having specific receptive field structures. Later, Barlow proposed that these receptive fields were the basis for neural mechanisms of visual form processing whereby features of the external world are extracted by central neurons, and that this feature representation is the basis for the representation of form in the visual system. In the 1950s and 1960s, these ideas of visual receptive fields performing feature extraction were further elaborated on by Hubel and Wiesel who showed that neurons in primary visual cortex are orientation selective and hypothesized that the orientation tuning resulted from neurons having receptive fields with separate subregions of excitation and inhibition, and they called such neurons 'simple cells.' They further hypothesized that the receptive field structures become more elaborate and complex as the processing of visual information proceeds centrally to higher visual areas. The aim of numerous studies since then has been to understand the processing functions of those cortical neurons.

The receptive fields and processing functions of neurons in touch have received significantly less attention. Anatomical studies have shown that the processing of spatial form begins with the activation of four mechanoreceptive afferents that innervate the skin. From there the information progresses centrally to neurons in the dorsal column nuclei, then to the ventrobasal complex of the thalamus, and then to primary (SI) and secondary (SII) somatosensory cortex. The processing pathways beyond SII are not well understood. Here we review the receptive field properties of neurons in the somatosensory system. We begin with a discussion of the receptive fields of the peripheral afferents that innervate the skin. Then we discuss the receptive fields of neurons in SI and then SII cortex. Both of these areas have been shown in albation studies to lie directly along the pathway leading to tactile form perception.

Peripheral Receptive Fields

Tactile perception is based on four cutaneous mechanoreceptive afferent types: slowly adapting type 1 (SA1), rapidly adapting (RA), Pacinian (PC), and slowly adapting type 2 (SA2) afferents. Two of these, the RA and PC afferents, are classified as rapidly adapting because they respond transiently to sudden indentation and removal of stimuli and do not respond to sustained indentations. The other two, the SA1 and SA2 afferents, are classified as slowly

adapting because they respond to sustained skin deformation with a sustained discharge that declines slowly, but they, like the RA and PC afferents, are much more sensitive to skin movement than to static deformation. The neural response properties of the cutaneous SA1, RA, and PC afferents have been studied extensively in both human and nonhuman primates. SA2 afferents are not found in nonhuman primates.

Sample receptive fields for each of the four mechanoreceptive afferents from humans are shown in Figure 1. These responses were recorded from humans using percutaneous electrodes that recorded the activity of single afferent fibers while the skin was stimulated using a blunt probe that systematically indented the finger pad. As shown, the SA1 and RA afferents have small receptive fields with diameters of about 4–6 mm, whereas SA2 and PC afferents have extremely large receptive fields with diameters that essentially could not be measured using punctuate stimuli. There are several other aspects of the receptive fields that should be noted in the figure. First, for all afferent types the receptive field size increases with indentation amplitude. This is to be expected since all four afferent types receive their inputs from receptors that are imbedded in the epidermis (SA1 and RA) and deeper tissues (PC and SA2) and respond to changes in mechanical stresses and strains on the skin. Thus, the changes in receptive field size with increases in indentation amplitude are directly attributed to the changes in the stress and strain profiles in the skin.

SA1 afferent fibers branch repeatedly before they lose their myelin and terminate in the basal layer of the epidermis. There, they are enveloped by specialized epidermal cells, called Merkel cells, that enfold the unmyelinated ends of the SA1 axons. Each SA1 afferent fiber receives input from multiple Merkel cells which appear as local islands of excitation in Figure 1. SA1 afferents have peaked-like receptive fields with one or more central zones of strong excitation and response sensitivities that decline as one moves toward the edge of the field. Combined neurophysiological studies of nonhuman primates and psychophysical studies of humans have shown that these afferents are suited for detecting fine details of spatial stimuli and underlie form and texture perception. RA afferent fibers also branch repeatedly as they near the epidermis, and like the SA1 afferents, each of the 30–80 or so branches loses its myelin sheath before terminating in Meissner's corpuscles, which are found

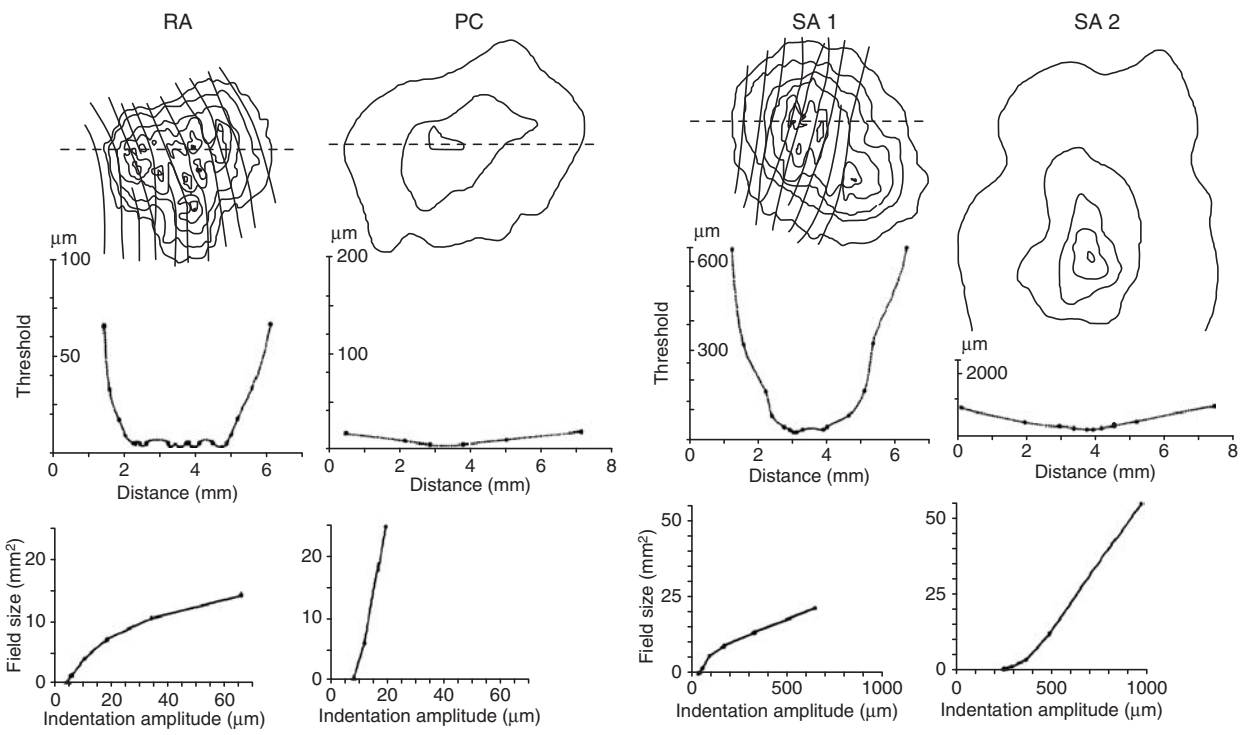

Figure 1 Receptive field characteristics of the four human mechanoreceptor afferents that innervate the glabrous skin. Top: Contour maps of receptive field sensitivities. The closed lines are isosensitivity contours that illustrate the threshold of the receptor for different amplitudes of mechanical indentation of the skin on the finger. The thin, parallel lines in the SA1 and RA plots mark the grooves between the papillary ridges. Middle: The threshold of the receptor at different distances from the center of a cross-section through the receptive field. Bottom: Relationship between indentation amplitude and receptive field size. Adapted from Johansson RS (1976) Receptive field sensitivity profile of mechanosensitive units innervating the glabrous skin of the human hand. *Brain Research* 104: 330–334.

in the spaces between the sweat-ducts and adhesive ridges. Evidence of these multiple branches can be seen in Figure 1. The sensitivity profile of the RA afferents is relatively flat compared to the SA1 afferents (Figure 1), suggesting that RA receptors are poorly suited for detecting the fine spatial details of stimuli but instead are well suited for detecting stimulus contact. Both RA and SA1 afferents innervate the skin of the finger pad densely. The estimates in humans and monkeys are not significantly different from each other with there being, at the fingertip, about 100 SA1 and 150 RA afferents per cm^2. These high innervation densities show that when the finger is in contact with an object both afferent types will encode in their discharge a dense two-dimensional representation of stimulus form indented on the skin.

A striking feature of RA responses is their sensitivity to minute skin motion. The effective operating range of indentations for RAs is about 4–400 μm; the comparable SA1 range is about 15–1500 μm or more. The neural response properties of RA afferents make them ideally suited for coding motion perception.

Each PC afferent terminates in a single Pacinian corpuscle, which occurs in the dermis or the deeper tissues. The most striking feature of the PC response is its extreme sensitivity, which derives from mechanosensitive ion channels located in the unmyelinated ending that lies at the center of the corpuscle. Pacinian corpuscles respond with action potentials to vibratory amplitudes as small as 3 nm applied directly to the corpuscle. Because of their extreme sensitivity to vibrations and movement, it is difficult to get accurate estimates of the receptive field sizes of these afferents. Some PCs have receptive fields that are restricted to one or a few digits, whereas others have receptive fields that span the entire arm. Like the RA afferents, the spatial sensitivity of PC afferents is fairly uniform across the receptive field (Figure 1), demonstrating that these afferents provide no useful information about spatial form. SA2 afferents also have very large receptive fields (Figure 1) and like PC afferents, they have unclear borders. These afferents are extremely sensitive to skin stretch and because of their large uniform receptive fields and low innervation densities, they do not provide useful information about spatial form. Interestingly, these afferents have not been found in nonhuman primates and recent studies in humans have brought into some doubt the underlying receptor and receptor mechanisms for these afferents.

The receptive field sizes of monkey SA1 afferents are significantly smaller and more uniformly shaped than the receptive fields of RA afferents. Although SA1 peripheral afferents tend to be smaller than RA afferents, there is overlap between the two populations in their sizes. The mean areas for these afferents are about 11 and 23 mm^2, respectively, with corresponding diameters of about 4.0 and 6.0 mm, which are similar to the receptive field sizes of these afferents measured in humans.

Recent studies have shown that experiments in which the receptive fields were derived using punctate probes (Figure 1) give an incomplete impression of the neuron's receptive field properties. Figure 2 shows the spatial temporal receptive fields (STRFs) for SA1 and RA neurons that were computed by performing a spike-triggered averaging of the neural response to a large array of independently controlled probes that randomly indented the skin. The receptive fields determined in this manner show in time whether a neuron is more or less likely to respond with an action potential when the probe is placed at any location within the receptive field. The response shown in Figure 2 illustrates that 5 ms after stimulus onset SA1 afferents have a central excitatory zone surrounded by a zone of mild suppression. The effect of the excitatory zone fades quickly (within 15–20 ms) and is replaced by a period of suppression where the neuron is less likely to fire action potentials. We refer to these effects as suppression rather than inhibition to distinguish them from active inhibitory mechanisms. For the peripheral afferents the initial surround suppression is due to skin mechanics, whereas the replacing suppression is due to the refractory period of the neuron. In contrast, the RA response, which also fades quickly, does not show edge-suppressed responses and like the SA1 response shows a period of delayed suppression following the excitatory response. Figure 3 shows the average total STRF area and average area for the excitatory and suppressed responses for a population of these afferents. The total receptive field area for RA afferents is similar to the area of the excitatory part alone. Further, Figure 3 shows that the inhibitory response due to the neuron's refractory period can affect the neural response for a significant period of time following excitation.

Receptive Fields of Neurons in SI Cortex

Sample receptive fields of 247 area 3b neurons mapped with scanned, random-dot stimuli are illustrated in Figure 4. In contrast to receptive fields of peripheral afferents, the receptive fields of most cortical neurons of SI cortex are characterized by having a single central region of excitation surrounded by inhibition on one, two, or three sides (Figures 4(a)–4(e)). Neurons with surround inhibition on all four sides (Figure 4(f)) were rare, as were neurons that had receptive fields like peripheral SA1 and RA afferents, which have a simple excitatory receptive field with

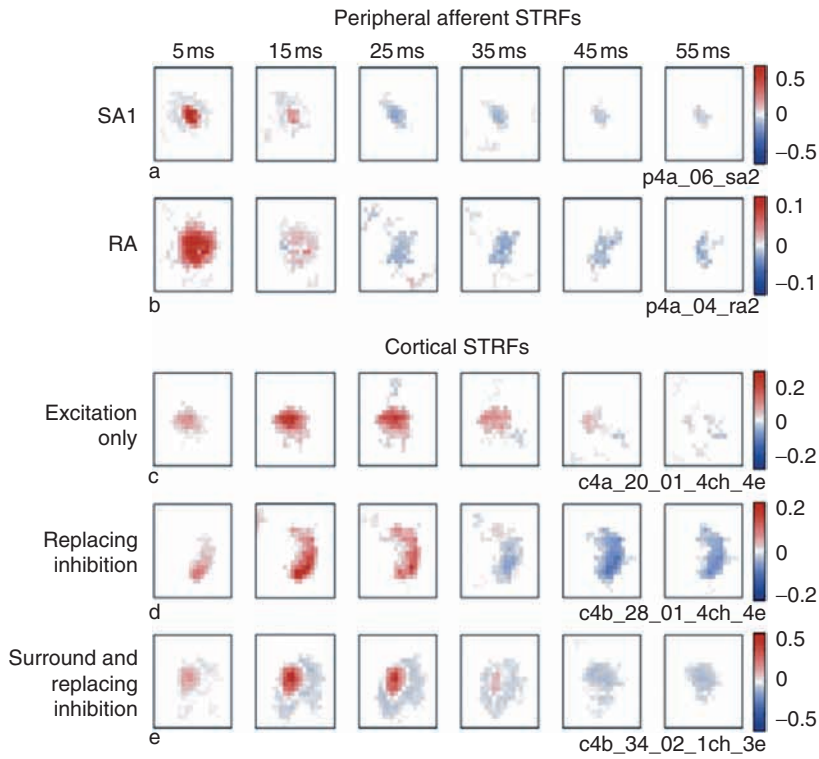

Figure 2 Peripheral and cortical spatiotemporal receptive fields (STRFs). Red areas denote excitation, blue areas denote inhibition/suppression, and the intensity of redness or blueness indicates the strength of the response. STRFs of a peripheral SA1 afferent (a) and a peripheral RA afferent (b). STRFs in SI cortex (areas 3b and 1) could be broadly categorized into the following three types: (c), an area 3b neuron with a purely excitatory STRF; (d), an area 3b neuron with replacing inhibition/suppression; (e), an area 1 neuron with surround and replacing inhibition/suppression. Colored bars indicate pixel intensities in spikes per second per micrometer. Adapted from Sripati AP, Yoshioka T, Denchev P, Hsiao SS, and Johnson KO (2006) Spatiotemporal receptive fields of peripheral afferents and cortical area 3b and 1 neurons in the primate somatosensory system. *The Journal of Neuroscience* 26: 2101–2114.

minimal surround inhibition (Figure 4(g)). The wide variety of receptive field structures in area 3b suggests that the representation of spatial form is significantly transformed from the peripheral response, with neurons in cortex extracting information about spatial features of the stimulus.

The relationship between the regions of excitation and inhibition is significantly more complex if one considers the STRF. This is illustrated in Figures 2(c)–2(e), which shows sample STRFs from three kinds of responses that are observed in cortex. First, for all cortical neurons the excitatory response becomes most prominent approximately 15–25 ms after stimulus onset and as expected is delayed relative to the peripheral response because of the conduction delays necessary to convey the response from the input to the cortex. The neurons shown in Figures 2(c) and 2(d) showed little surround inhibition whereas the neuron shown in Figure 2(e), which was more typical of the neural responses, showed significant surround inhibition, like the responses shown in Figure 4. The excitatory response shown in Figure 2(c) faded rapidly with little replacing inhibition, whereas the response shown in Figure 2(d) showed significant replacing inhibition that was still intense 50 ms following stimulus onset. For cortical neurons the amount and latency of the replacing inhibition cannot be easily explained by the refractory period of neurons, which suggests that it is due to active inhibitory mechanisms.

Figure 2(e) shows an example of the more common type of response that was observed in area 3b of SI cortex. This neuron showed significant surround inhibition and significant replacing inhibition. Notice in this figure that the surround inhibition evolves more slowly and persists longer than the excitatory region. The temporal evolution of regions of excitation and inhibition for cortical neurons can be more clearly seen in Figure 3(c). The figure shows that initially the response for the central neurons is dominated by excitation and that the size of the inhibitory response slowly increases. After about 25 ms the response switches to one where inhibition dominates excitation. The inhibition is due to both surround inhibitory mechanisms and replacing inhibition, with replacing inhibition becoming more dominant as time goes on. The time course of the response is

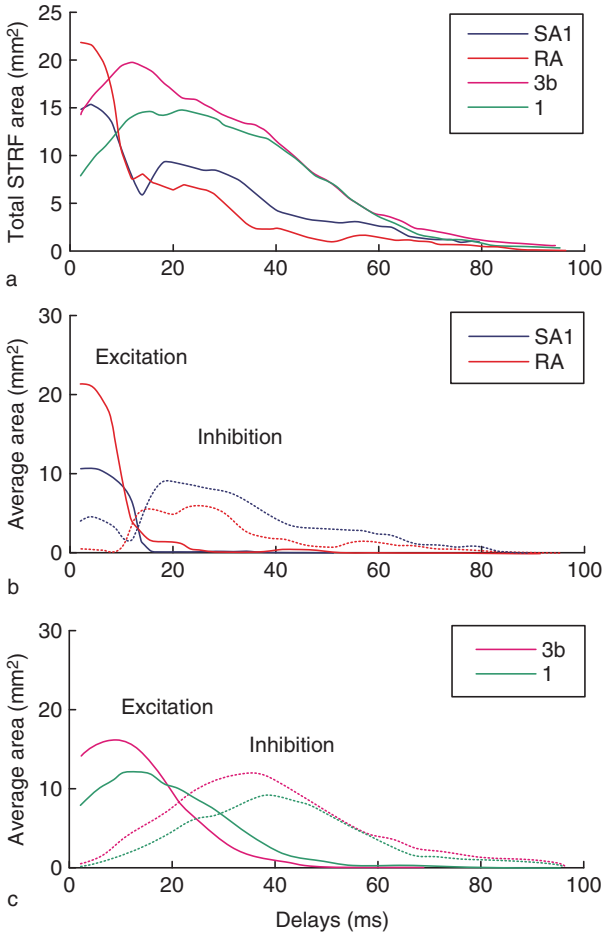

Figure 3 Temporal evolution of STRF areas. (a) Total (excitatory plus inhibitory/suppressed) STRF area averaged over all neurons of a given afferent type or cortical area. (b) Average areas of excitatory (solid lines) and suppressed (dotted lines) regions for peripheral SA1 and RA afferents. (c) Average areas of excitatory (solid lines) and inhibitory/suppressed (dotted lines) regions over time for cortical areas 3b (magenta) and 1 (green). Adapted from Sripati AP, Yoshioka T, Denchev P, Hsiao SS, and Johnson KO (2006) Spatiotemporal receptive fields of peripheral afferents and cortical area 3b and 1 neurons in the primate somatosensory system. *The Journal of Neuroscience* 26: 2101–2114.

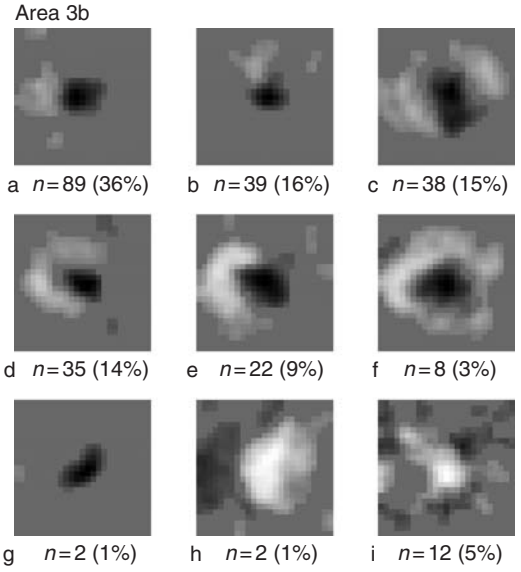

Figure 4 Receptive field (RF) structures of area 3b neurons studied with scanned dots. Each panel gives an example neuron of a structural type, the total number of RFs of that type, and their percent of the total RF sample ($n = 247$ neurons). The types are shown in decreasing order of frequency. Blackness denotes excitation, whiteness denotes inhibition, and the magnitude of excitation and inhibition is shown by the degree of blackness and whiteness, respectively. (a) A single inhibitory region located on the trailing (distal) side of the excitatory region. (b) A region of inhibition located on one of the three nontrailing sides of the excitatory region. (c) Two regions of inhibition on opposite sides of the excitatory region. (d) Inhibition on three sides of the excitatory region. (e) Inhibition on two contiguous sides of the excitatory region. (f) A complete inhibitory surround. (g) An excitatory region only. (h) RF dominated by inhibition. (i) RFs not easily assigned to one of the preceding categories. Adapted from DiCarlo JJ, Johnson KO, and Hsiao SS (1998) Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey. *The Journal of Neuroscience* 18: 2626–2645.

similar but not identical for neurons in area 3b versus area 1. In area 1 the excitation appears to evolve later, as does the inhibition, suggesting that area 3b neurons project to area 1 neurons, a hypothesis that is supported by anatomical evidence.

The total receptive field area for both peripheral and central neurons in SI cortex is shown in Figure 3(a). It is clear that neurons in SI cortex have larger receptive fields and that the differences between the peripheral and central responses are due to both increased excitation and the increased role that inhibition plays in the neural response.

The fact that neurons in SI show subregions with both excitation and inhibition demonstrates that these receptive fields are more elaborate than the receptive fields of peripheral neurons, which suggests that SI neurons are further along the pathway that underlies tactile form processing.

Figure 5 illustrates the receptive fields for a subset of the neurons that showed orientation selectivity. For these neurons the selectivity, like that of simple cells in the visual system, is based on the receptive field shape. As such the orientation tuning can be explained by these neurons having an oriented excitatory subregion flanked by an oriented inhibitory subregion. These kinds of responses may underlie the high percentage of neurons in SI (about 70%) that show orientation-tuned responses when bars are indented into the center of the finger pad.

A more stringent test of the validity of these linear receptive field types is their ability to account for the neural responses to complex spatial patterns scanned across the receptive field. This is shown in Figure 6 for two SI neurons, which shows the estimated receptive field, the predicted response to scanned embossed

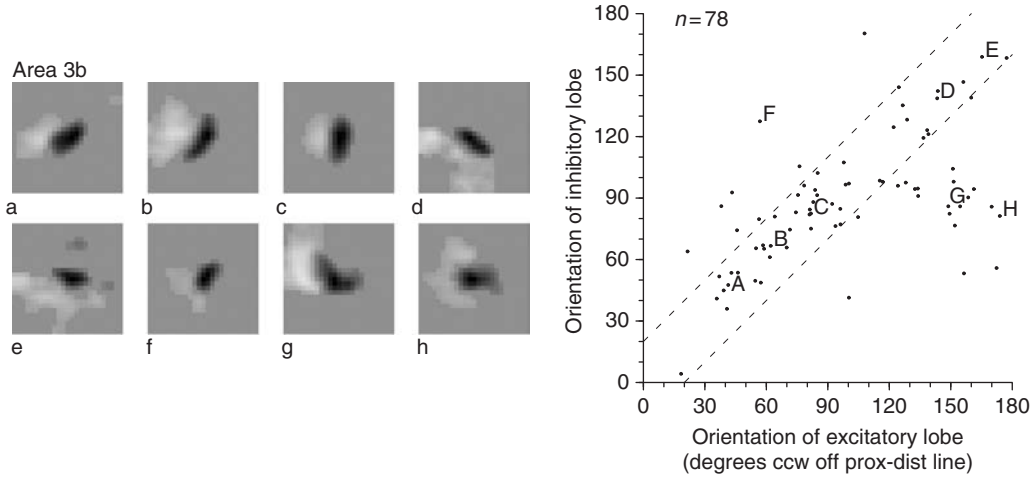

Figure 5 Relationship between excitatory and inhibitory receptive field lobe orientations. Only neurons with dominant excitatory and inhibitory lobes, both with aspect ratios >1.5, were included ($n = 78$ area 3b neurons). Right: Excitatory and inhibitory RF lobe orientations tended to be similar (correlation coefficient = 0.586). Sixty-five percent of the points lie within ±20° of the diagonal (within the dashed lines on the right). Left: Example neurons (a–h). Adapted from DiCarlo JJ, Johnson KO, and Hsiao SS (1998) Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey. *The Journal of Neuroscience* 18: 2626–2645.

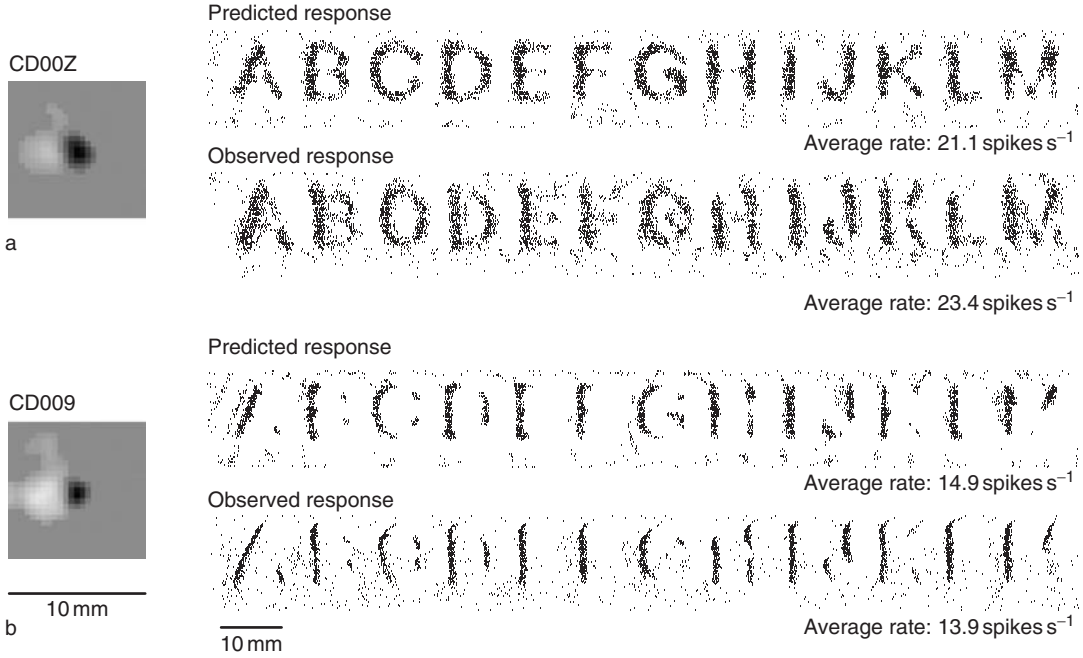

Figure 6 Generality of the receptive fields. The responses of two area 3b neurons studied with both random dots and embossed letters. The predicted responses to the embossed letters were determined by convolving the RF (left) determined from the random dot pattern with a template of the letter stimulus. Predicted firing rates were converted into action potentials using a Poisson spike generator, with the overall predicted spike rate adjusted to coincide with the observed spike rate. Adapted from DiCarlo JJ, Johnson KO, and Hsiao SS (1998) Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey. *The Journal of Neuroscience* 18: 2626–2645.

letters of the alphabet, and responses that were observed when neurons were presented with the same letter stimuli. The similarity between the observed and predicted responses is striking, which suggests that these receptive fields with regions of excitation and inhibition are good linear estimates of the processing kernels in this area of cortex.

Whereas the fixed excitatory and inhibitory components of the response provide the neurons with selectivity to spatial gradients and account for the complex spatial responses shown in Figure 6, the delayed replacing inhibitory component serves a different purpose. One hypothesis is that it serves as a compensatory mechanism that increases the firing

rate with increasing scanning velocity. As scanning velocity increases, the delayed inhibition lags progressively further behind in time to expose more excitation. The consequence is that the excitatory and inhibitory components of the receptive field grow in intensity with increases in velocity, having no effect on receptive field geometry. The result is a representation of spatial form that is invariant with scanning velocity.

Receptive Fields of Neurons in SII Cortex

Most neurophysiological studies support the idea that SII is further along the pathways responsible for processing tactile form. Recently, it has been shown that this cortical region consists of multiple areas that have been given a variety of names (OP1–OP4 in humans; anterior field (SIIa), central field (SIIc), and posterior field (SIIp); or SII, PV, and PR in nonhuman primates). While the precise organization of these areas and how they are interconnected with neurons in SI is not well understood, it is clear that neurons in these areas have larger and more elaborate receptive fields than do neurons in SI cortex. What is particularly striking is that the receptive fields are much larger in the three areas (fields) that make up SII cortex than for the neurons in SI cortex which typically only span one (area 3b) or a few (area 1) finger pad(s). Recently, SII receptive fields were studied using an oriented bar that systematically indented the distal, middle, and proximal finger pads of digits 2–5 of the macaque hand. For a given neuron, each pad was classified as either orientation tuned (T), untuned excitatory (UE), untuned inhibitory (UI), or unresponsive. Figure 7 shows a random sample of receptive fields for 15 SIIc neurons (top) that only had untuned excitatory pads, as well as a random sample of 15 SIIc neurons (bottom) that had one or more orientation-tuned pads. It is clear that neurons in SII have a variety of receptive field sizes and shapes, with some neurons having their receptive field confined to a single finger pad (Figure 7, row a of column 1) and others having receptive fields that spanned multiple fingers and finger pads. Although the receptive fields tended to be contiguous, with untuned excitatory pads adjacent to each other, this was not a rule, since some neurons had noncontiguous receptive fields (Figure 7, row d of column 3). In general, neurons in SII showed a similar variety of receptive field sizes and shapes, whether their responsive finger pads were untuned excitatory or untuned inhibitory. About 36% of the neurons only had untuned excitatory pads like the ones shown in Figure 7 (top). A total of 17% had only untuned inhibitory pads and the least commonly observed type were the 5% that had both untuned excitatory and untuned inhibitory pads.

The second most common type (about 23%) were neurons that had one or more tuned finger pads (Figure 7, bottom), that is, neurons that showed different responses when the bar was indented into the center of the pad at different orientations. As shown in the figure, most neurons had combinations of tuned and untuned pads and some had only tuned pads. While some neurons had only one tuned pad, many neurons, especially those in the central field of SII, had multiple tuned pads.

Although all stimulated parts of the hand were represented in the receptive fields, there was a bias in the central and anterior SII fields for the index and middle fingers to be better represented than the other two large digits, and a bias for the distal and middle pads to be better represented than the proximal pad. This is reasonable since these are the digits and pads that are used most often when exploring objects.

A close inspection of the tuning preferences for neurons with multiple tuned pads showed that the tuning was usually similar between pads. An example of one such neuron is shown in Figure 8. This neuron had seven tuned pads, all with similar tuning preferences (Figure 8, bottom). The similarity demonstrates that neurons in SII tend to show position-invariant responses to oriented bars and indicates that neurons in SII are further along the pathway for coding the shape of objects placed in the hand. One hypothesis is that these receptive field structures represent the underlying kernels that are used for representing object shape. Thus, orientation-tuned pads could signal the edges of objects and untuned pads could signal the flat surfaces of objects.

Figure 9 shows the responses from a single finger pad for two neurons in SII cortex that were stimulated with an oriented bar at multiple positions and orientations within that pad. Both of these neurons showed orientation-tuned responses to bars indented in the center of the finger pad; however, an analysis of the vector fields that estimate the local tuning across the pad demonstrates that the mechanisms underlying the tuned responses vary greatly between neurons. For example, the tuning response for the neuron shown on the right was explained well by a single excitatory region ('hot spot' which 'attracts' the vectors), which is predicted by the linear receptive field shown in the bottom right of this figure. In contrast many neurons showed invariant orientation-tuned responses like the one shown on the left in Figure 9. The response for this neuron is not easily explained by its linear receptive field. These results show that many neurons in SII cortex show nonlinear responses to spatial stimuli.

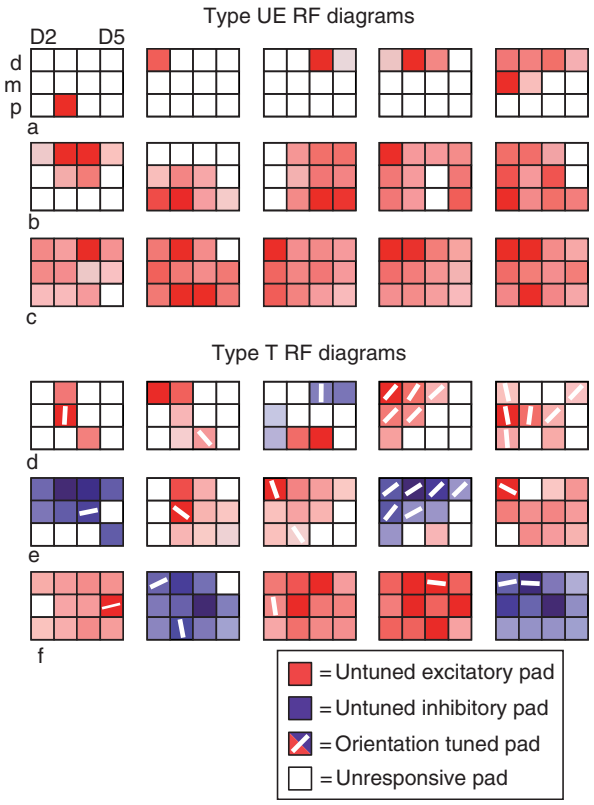

Figure 7 SII receptive field diagrams. Shown are a random sampling of 15 central field type UE (untuned excitatory) neurons (top), and 15 central field type T (orientation tuned) neurons (bottom). Each 3 × 4 grid represents the 12 pads of D2–5 for a single neuron (as shown for the upper left receptive field diagram), with the top row of each grid representing the distal pads, and the left column representing D2. For untuned excitatory pads, redness represents the peak excitatory rate; for untuned inhibitory pads, blueness represents the peak inhibitory rate; and for tuned pads, redness or blueness represents the maximum deviation from the spontaneous rate. The orientation of each white bar represents the preferred orientation (mean angular vector) of that tuned pad. Within each type, the neurons are arranged in an ascending order based on RF size (number of pads). D, digit; d, distal; m, middle; p, proximal. Adapted from Fitzgerald PJ, Lane JW, Thakur PH, and Hsiao SS (2006) Receptive field (RF) properties of the macaque second somatosensory cortex: RF size, shape, and somatotopic organization. *The Journal of Neuroscience* 26: 6485–6495.

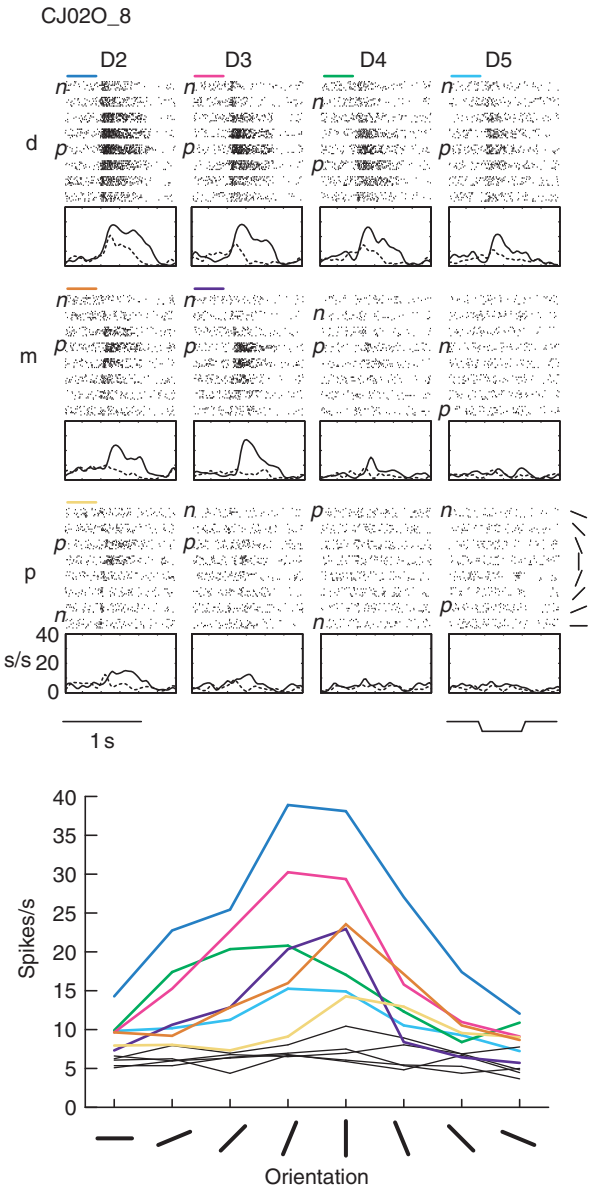

Figure 8 Example of an orientation-tuned SII neuron. Shown are raster plots and peristimulus time histograms (PSTHs) (top) for all 12 stimulated pads of this central field neuron, as well as orientation-tuning curves for all 12 pads (bottom). In the raster, stimulus trials are first sorted into the 12 stimulated finger pads (D2–5, d–p), and then within each pad are sorted by the orientation of the bar (see lower, extreme right). The bar stimulus indentation profile is also shown (lower right). Each PSTH graph is shown below its corresponding finger pad, where only the preferred (highest firing rate; solid line) and nonpreferred (lowest firing rate; dashed line) orientation rates are plotted after convolving these rates with a Gaussian. The short, colored bars above the raster plots indicate tuned pads, and correspond to the colored tuning curves shown in the orientation tuning plot at the bottom. Each point in each tuning curve corresponds to the mean evoked firing rate during the 500 ms indentation of the oriented bar at the given orientation on the given finger pad. The thin black lines in the orientation tuning plot represent untuned pads. D, digit; d, distal; m, middle; p, proximal; *p*, preferred orientation; *n*, nonpreferred orientation; s/s, spikes/second. Adapted from Fitzgerald PJ, Lane JW, Thakur PH, and Hsiao SS (2006) Receptive field (RF) properties of the macaque second somatosensory cortex: RF size, shape, and somatotopic organization. *The Journal of Neuroscience* 26: 6485–6495.

Conclusions and Summary

So what is a receptive field? In the simplest definition it is the area of stimulus space over which changes in the stimulus cause corresponding increases or decreases in the firing rate of the neuron. In this definition, neurons such as the SA1 afferents in the periphery simply provide information about whether a stimulus is present or absent at a particular location on the finger pad. For these neurons the responses are relatively linear and can be explained reasonably well by a central region of excitation with surround suppression. In cortex the receptive fields are significantly more elaborate and in SI can be roughly explained by neurons having spatiotemporal receptive

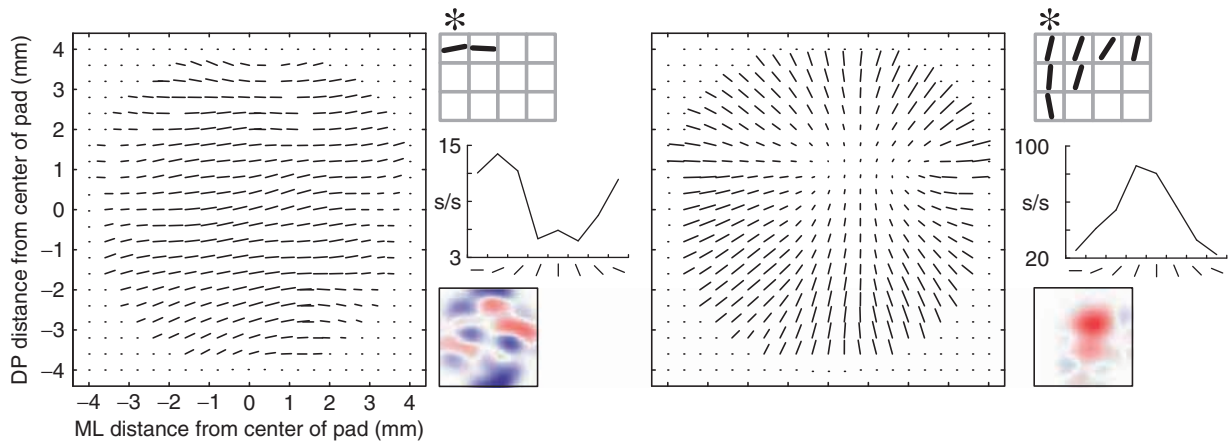

Figure 9 Example vector fields of SII neurons. These vector plots represent the orientation tuning at different points within a finger pad when an oriented bar was presented at different locations within that pad. Shown are an invariant neuron (left) and a hot-spot neuron (right). Within each vector field, each short line segment is a vector that represents the preferred orientation (direction) and strength (magnitude) of tuning at that point on the distal finger pad. Also shown for each neuron is the preferred orientation diagram (top right) from the 12 pad RF mapping protocol (**Figures 7** and **8**); the asterisk indicates the distal pad shown in this figure. The orientation tuning curve (middle right) is the tuning when the bar was centered on the pad and represents the mean firing rate during the 500 ms presentation of the bar at the particular orientation. The linear receptive field (lower right) for this figure was obtained using reverse correlation. Redness represents excitation and blueness represents inhibition, and the degree of redness or blueness is proportional to the magnitude of excitation or inhibition, respectively. Adapted from Thakur PH, Fitzgerald PJ, Lane JW, and Hsiao SS (2006) Receptive field properties of the macaque second somatosensory cortex: Nonlinear mechanisms underlying the representation of orientation within a finger pad. *The Journal of Neuroscience* 26: 13567–13575.

fields with a central excitatory region surrounded by one or more regions of inhibition. Furthermore, the excitatory regions are masked by replacing inhibition that gives these neurons a degree of velocity invariance. As one moves further along the pathway, the neurons tend to have receptive fields that are even larger and more elaborate, spanning multiple fingers and finger pads with complex groupings of tuned and untuned pads. A working hypothesis is that objects presented to the hand are represented across a population of such processing kernels. Future studies are needed to understand how these receptive fields are integrated to form central representations of object shape and texture.

See also: Mechanoreceptors; Neural Circuitry in the Somatosensory System; Proprioception; Sensorimotor Integration: Attention and the Premotor Theory; Sensorimotor Integration: Barrels, Vibrissae and Topographic Representations; Somatosensory Cortex: Functional Architecture; Somatosensory Cortex; Somatosensory Pathways (Ascending): Functional Architecture; Somatosensory Perception; Somatosensory Plasticity.

Further Reading

DiCarlo JJ and Johnson KO (1999) Velocity invariance of receptive field structure in somatosensory cortical area 3b of the alert monkey. *The Journal of Neuroscience* 19: 401–419.

DiCarlo JJ, Johnson KO, and Hsiao SS (1998) Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey. *The Journal of Neuroscience* 18: 2626–2645.

Fitzgerald PJ, Lane JW, Thakur PH, and Hsiao SS (2006) Receptive field (RF) properties of the macaque second somatosensory cortex: RF size, shape, and somatotopic organization. *The Journal of Neuroscience* 26: 6485–6495.

Fitzgerald PJ, Lane JW, Thakur PH, and Hsiao SS (2006) Receptive field properties of the macaque second somatosensory cortex: Representation of orientation on different finger pads. *The Journal of Neuroscience* 26: 6473–6484.

Hsiao SS (1998) Similarities between touch and vision. In: Morley JW (ed.) *Neural Aspects of Tactile Sensation*, pp. 131–165. Amsterdam: Elsevier.

Hsiao SS, Johnson KO, and Yoshioka T (2003) Processing of tactile information in the primate brain. In: Gallagher M and Nelson RJ (eds.) *Comprehensive Handbook of Psychology, Volume 3: Biological Psychology*, pp. 211–236. New York: Wiley.

Johansson RS (1976) Receptive field sensitivity profile of mechanosensitive units innervating the glabrous skin of the human hand. *Brain Research* 104: 330–334.

Johnson KO, Yoshioka T, and Vega-Bermudez F (2000) Tactile functions of mechanoreceptive afferents innervating the hand. *Journal of Clinical Neurophysiology* 17: 539–558.

Mountcastle VB (2005) *The Sensory Hand. Neural Mechanisms in Somatic Sensation.* Cambridge, MA: Harvard University Press.

Nelson RJ (2002) *The Somatosensory System: Deciphering the Brain's Own Body Image.* Boca Raton: CRC Press.

Sripati AP, Yoshioka T, Denchev P, Hsiao SS, and Johnson KO (2006) Spatiotemporal receptive fields of peripheral afferents and cortical area 3b and 1 neurons in the primate somatosensory system. *The Journal of Neuroscience* 26: 2101–2114.

Thakur PH, Fitzgerald PJ, Lane JW, and Hsiao SS (2006) Receptive field properties of the macaque second somatosensory cortex: Nonlinear mechanisms underlying the representation of orientation within a finger pad. *The Journal of Neuroscience* 26: 13567–13575.

Vega-Bermudez F and Johnson KO (1999) SA1 and RA receptive fields, response variability, and population responses mapped with a probe array. *The Journal of Neuroscience* 81: 2701–2710.

Somatostatin and Receptors

P C Emson, The Babraham Institute, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Somatostatin

Somatostatin, or somatotropin-releasing inhibitory factor (SRIF), is a cyclic tetradecapeptide with a disulfide bridge between the cysteine residues 3 and 14 (Figures 1(a) and 1(b)). The first evidence for its existence came in 1968 when Krulich and colleagues identified a hypothalamic factor (somatotropin) that reduced the release of growth hormone (GH) from rat pituitaries *in vivo*. Subsequent work by Guillemin's laboratory in 1972 attempting to purify GH led them to isolate somatostatin from 500 000 sheep hypothalamic extracts and then to sequence and synthesize it. The synthetic peptide caused a specific and concentration-dependent inhibition of GH release from the pituitaries consistent with its name. Following the synthesis of somatostatin, a number of groups developed antibodies to the cyclic peptide and further explored its physiological roles. Somatostatin is found in a wide range of central nervous system (CNS) and peripheral tissues outside the hypothalamus. In the CNS, somatostatin is found in number of neurons in the forebrain, including neurons in the cerebral cortex, hippocampus, basal ganglia, spinal cord, and hypothalamus. In the spinal cord, the neuropeptide is localized both in the dorsal root ganglia and dorsal horn (substantia gelatinosa) of the spinal cord, where the afferents from the dorsal root ganglia synapse. In the periphery, somatostatin is found in the gut neurons of the enteric nervous system and in the endocrine cells of the intestine and pancreas. A number of stable analogs of somatostatin have been produced, for example sandostatin (octreotide), which are used in the treatment of acromegaly (gigantism), and, due to somatostatin's marked antiproliferative action, as a treatment for a variety of endocrine and other tumors expressing somatostatin receptors.

Following the isolation and sequencing of somatostatin, the prohormone (precursor) sequence was elucidated and somatostatin-14 is produced from preprosomatostatin via a tryptic cleavage of the larger precursor somatostatin-28 (see Figure 1(a)). The 13-amino-acid N-terminus of somatostatin-28 corresponds to thrittene, a novel peptide recently isolated from monkey intestinal extracts. This peptide may represent a product of alternative processing of the somatostatin precursor.

The cloning of cortistatin, a peptide very closely related to somatostatin that also occurs in two short forms (cortistatin-17 and cortistatin-14) and a longer sequence (cortistatin-29), increased the complexity of the known somatostatin system (Figures 1(c) and 1(d)). Cortistatin-14 contains the core 11-amino-acid C-terminus of somatostatin (Figure 1(d)) and acts as a full agonist at somatostatin receptors. *In situ* hybridization studies with cortistatin probes reveals that cortistatin expression is mainly localized in the cerebral cortex and hippocampus, where it is found in parvalbumin-containing interneurons as well as coexisting with somatostatin itself in some populations of interneurons. Physiologically, as might be expected, cortistatin shares a number of functional properties with somatostatin, such as inhibiting neuronal activity, but inducing slow-wave sleep and reducing locomotor activity are unique properties of cortistatin.

Also related to somatostatin and cortistatin is the fish peptide urotensin II (UII) (Figure 1(e)). Urotensin II analogs are found in the mammalian CNS, but attempts to isolate the peptides expected to be formed from UII have so far been unsuccessful; urotensin-related peptide (URP), an octapeptide containing the C-terminal motif (CFWKYCV) found in UII, was the only UII-like peptide isolated, and this originates from a distinct URP precursor. The physiological effects of UII and URP are mediated through the serpentine G-protein-coupled receptor GRP14. Their effects may include specific actions on cardiomyocytes. Comparative genomic studies suggest that UII and URP genes and the somatostatin/cortistatin genes arise from segmental duplication of two ancestral genes that were physically linked together and that themselves arose from tandem duplication of a single ancestral somatostatin gene.

Somatostatin exerts its effects through some five somatostatin receptors, termed sst_{1-5} (Table 1), with two forms of the sst_2 receptor, termed sst_{2A} and sst_{2B}, which differ only in the length and composition of their C-terminus. Genomic studies suggest that there are only five somatostatin receptor genes. Cortistatin has thus far not been shown to have a separate receptor, although it has been suggested that an orphan receptor may be a candidate cortistatin receptor. Binding studies with labeled cortistatin in mice deficient in one of the main cortical somatostatin receptors (sst2A) showed that cortistatin binding remains in the hippocampus and habenula regions of the CNS of these knockout mice, suggesting these may be regions containing specific cortistatin receptors.

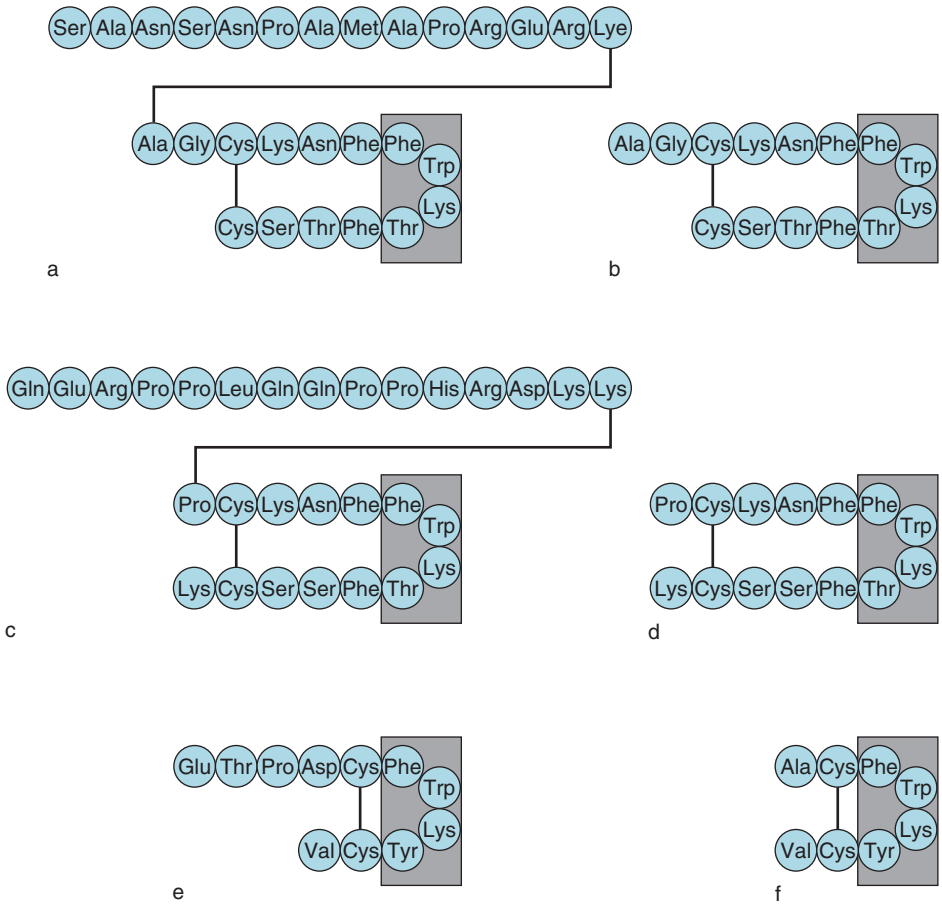

Figure 1 Diagrams of somatostatin and related peptides belonging to the cortistatin–somatostatin family: (a) $SRIF_{28}$, (b) $SRIF_{14}$, (c) $rCST_{29}$, (d) $rCST_{14}$, (e) urotensin II (UII), (f) urotensin-related peptide (URP). $SRIF_{28}$ and $SRIF_{14}$ are the two major peptides derived from tryptic processing of the somatostatin precursor. $rCST_{29}$ and $rCST_{14}$ are predicted sequences of cortistatin peptides derived from the cortistatin precursor by tryptic processing. Sequences of the urotensin peptides UII and URP are based on peptide purification data and sequencing, as are the sequences of the somatostatin peptides. The peptide thrittene corresponds to the N-terminal 13 amino acids of $SRIF_{28}$. CST, cortistatin; SRIF, somatotropin-releasing inhibitory factor.

However, the possibility that the labeled peptide is binding other members of the somatostatin receptor family which have not been knocked out or may be upregulated in the sst2A knockout mouse needs to be carefully considered.

Somatostatin Receptors

All the somatostatin receptors are typical G-protein-coupled receptors (Figure 2) with the classical seven transmembrane domains and sites for glycosylation and/or palmitoylation on the N-termini. They can be grouped into two subfamilies on the basis of pharmacology: receptors sst_1 and sst_4 in one group and sst_{2B} and sst_{2a}, sst_3, and sst_5 in the other group. The peptide analogs seglitide and octreotide can be used to distinguish the subfamilies; radioligand binding studies show that sst_1 and sst_4 possess only a low affinity for both analogs, whereas sst_2, sst_3, and sst_5 display a high affinity. The homology between the groups/subfamilies is less than 50%, but within the classes sequence similarities exceed 80% in transmembrane regions. The properties of the cloned human somatostatin receptors are summarized in Table 1. Somatostatin receptors couple to signal cascades through both pertussis toxin (PTX)-sensitive and pertussis-insensitive G-proteins (e.g., G_i/G_o and G_q). The precise coupling depending on the neuronal type and the different isoforms of G-proteins available in the target cell. It is also likely that somatostatin receptors may form oligomers (e.g., heterodimers) with other G-protein-coupled receptors such as the dopamine and opioid receptors; although this has been shown to occur only in cell lines overexpressing transfected receptor cDNAs, it may account for some of the unusual pharmacological properties of somatostatin observed in some experimental systems. All the receptors have been shown to inhibit adenylyl cyclase

Table 1 Summary of the characteristics of the cloned human somatostatin receptors

	sst_1	sst_2	sst_3	sst_4	sst_5
Amino acid number	391	369	418	388	363
Molecular mass (kDa)	42.6	41.3	45.7	41.9	39.1
Transcript size (kb)	4.8	8.5	5.0	4.0	4.0
Glycosylation sites	3	4	2	1	3
Phosphorylation sites	6	8	5	3	5
Chromosomal localization	14q13	17q24	22q13.1	20p11.2	16p13.3

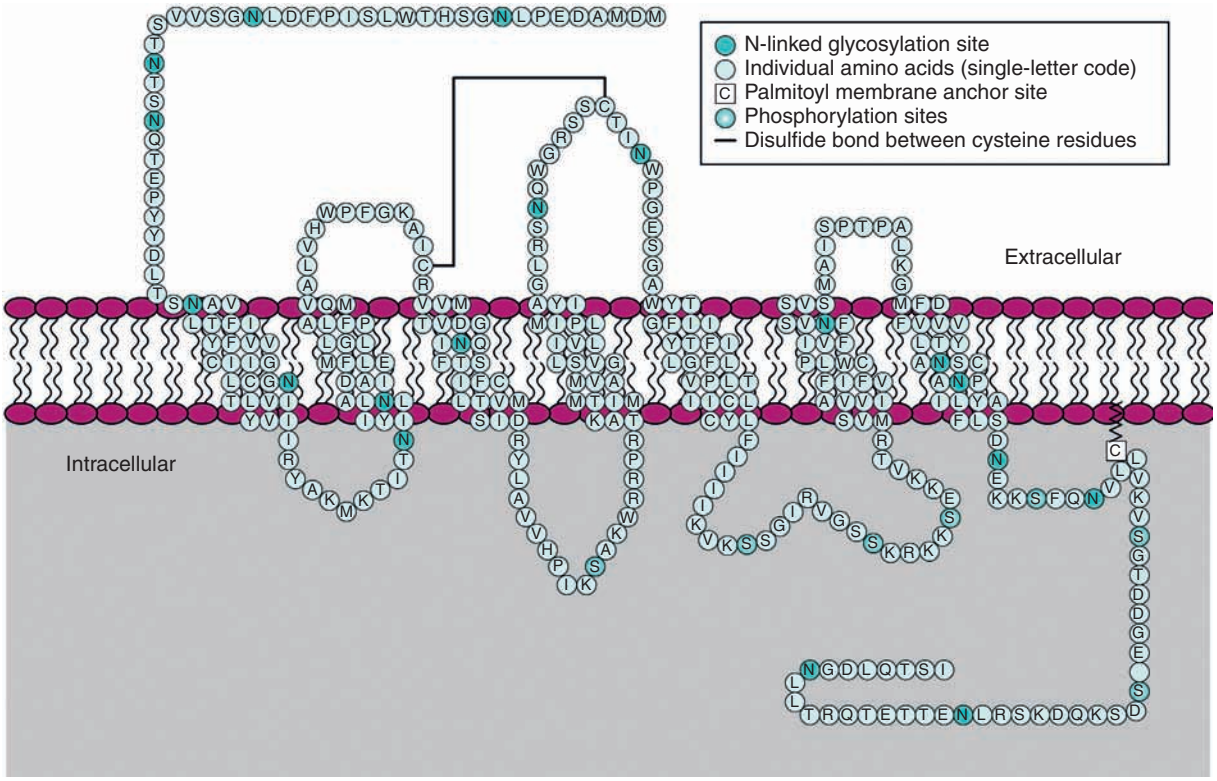

Figure 2 Diagram illustrating the structure of the somatostatin receptor sst_{2a}. Note the classical seven transmembrane structure of this typical G-protein-coupled receptor. Possible sites of gylcosylation, palmitoylation, and phosphorylation are shown.

activity in a PTX-sensitive fashion and to couple to phospholipase C. In addition, they also activate mitogen-activated protein kinase (MAPK) pathways influencing cell growth.

Distribution of Somatostatin

Following the isolation and sequencing of somatostatin, as noted earlier, the availability of antibodies for the peptide enabled the distribution and release of the peptide to be studied by using radioimmunoassay (RIA) and immunohistochemistry. As expected, somatostatin was found in neurons in the periventricular and paraventricular regions of the hypothalamus, from which it can be released to indirectly influence GH release by acting on growth hormone-releasing hormone (GHRH)-containing neurons in the hypothalamus and to directly act on the somatotrophs of the anterior pituitary. Outside the hypothalamus, somaostatin immunoreactivity was also found to be enriched in neurons in the forebrain, including the cerebral cortex, amygdala, hippocampus, and striatum. Somatostatin was released in a Ca^{2+}-dependent manner from brain slices, suggesting a possible role as a neurotransmitter/neuromodulator in the brain. Somatostatin neurons in the forebrain (except the hypothalamus) are now known to be primarily γ-aminobutyric acid (GABA)ergic interneurons which often contain a further vasoactive peptide neuropeptide Y (NPY) and the marker for the gaseous

transmitter nitric oxide (NO), neuronal nitric oxide synthase (nNOS), and in some neuronal populations the related peptide cortistatin. The distribution of the peptide in the forebrain suggests its involvement in neurological diseases affecting these regions, including Alzheimer's and Huntington's disease. Of particular interest is the recent observation that the β-amyloid-metabolizing enzyme neprilysin may be regulated by somatostatin such that a decline in somatostatin expression with age reduces the metabolism of β-amyloid, perhaps contributing to the accumulation of undegraded β-amyloid in senile plaques and eventually to the symptoms of dementia.

In Huntington's disease, an autosomal dominant condition characterized particularly by the degeneration of the majority of output cells (medium spiny neurons) from the caudate putamen (a region particularly involved in movement control), tissue studies of the brains of affected individuals established that the somatostatin/NPY/nNOS-containing interneurons of the caudate putamen are relatively preserved. It is not known why these cells are relatively resistant to damage, but a possible explanation may include a reduced sensitivity to the excitotoxic effects of the transmitter glutamate or the expression of protective proteins such as the Huntington-associated protein HAP1. Electrophysiological studies indicate that these striatal interneurons have distinct physiological properties with low threshold spikes (LTS) and occupy a position in the organization of the basal ganglia to provide feed-forward inhibition to output cells. It is becoming increasingly clear that each subpopulation of interneurons (see **Figure 3**) has distinct physiological properties selectively influencing output neurons. In Alzheimer's disease, the concentration of somatostatin peptide falls with the development of the characteristic hallmarks of this disease (the senile plaques and neuronal tangles), although the somatostatin-containing interneurons are not themselves the main targets for the primary pathology.

Paralleling the distribution of somatostatin immunoreactivity, somatostatin mRNA is also found in the forebrain in neurons, in excellent agreement with the earlier immunohistochemistry studies. Coexpression

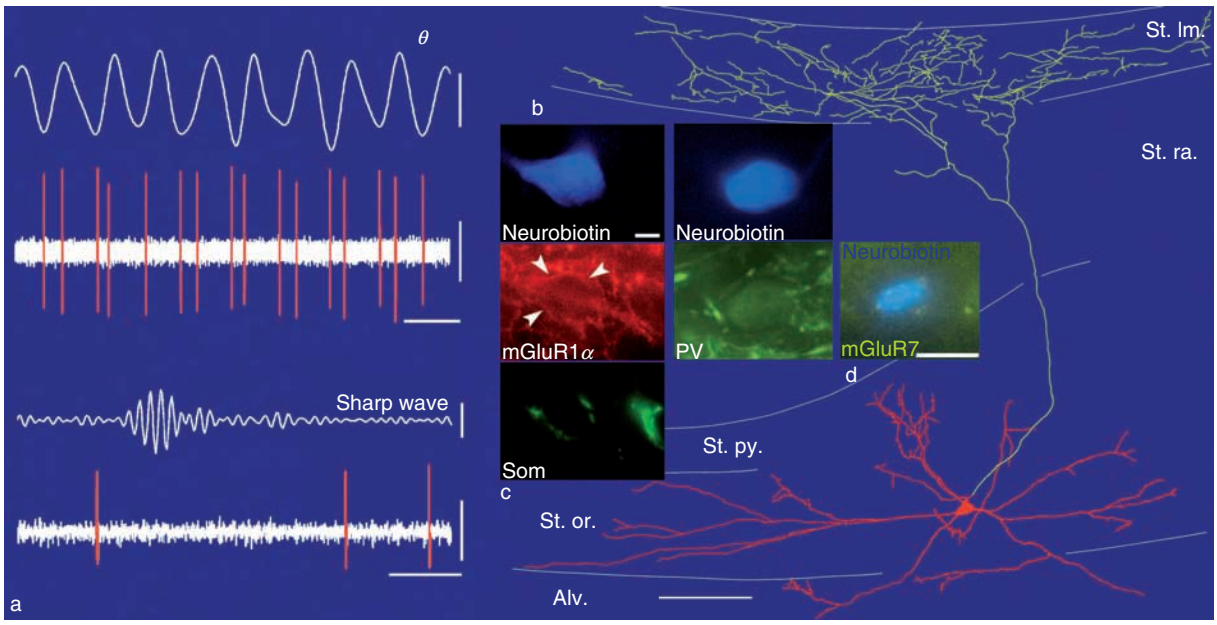

Figure 3 A somatostatin-positive O-LM cell in the rat hippocampus: (a) cell firing pattern; (b) reconstruction of the neurobiotin-filled cell showing the soma and dendrites in red; (c) soma cut into two sections; (d) dendrite of another recorded O-LM cell. The cell is identified on the basis of the distribution of its processes, the expression of its molecules as detected by immunocytochemistry, and its firing pattern *in vivo*. Following recording, the cell was labeled by the juxtacellular method. As shown in (a), the cell fired rhythmically on the trough of θ oscillations, but was silent during sharp wave-associated ripples. In (b), the axon restricted to str. lm. (yellow) is shown only from two 65-μm thick sections. In (c), one section of the soma shows somatostatin and mGluR1α immunoreactivity; the other demonstrates weak parvalbumin immunoreactivity. In (d), the dendrite of the O-LM cell is decorated by highly mGluR7a-immunopositive terminals (yellow). Other distinct types of somatostatin-positive cells, such as the bistratified cell, express different molecules in addition to somatostatin, establish synapses on other parts of pyramidal cells, and release transmitters at times different from the O-LM cell O-LM, Oriens-lacunosum moleculare cell; a, alveus; PV, parvalbumin. mGluR, metabolic glutamate receptor; Som, somatostatin; st. lm., stratum lacunosum-moleculare; st. or., stratum oriens; st. py., stratum pyramidale; st. ra., stratum radiatum. Scalebars = (unit) 0.3 mV; (θ) 0.2 mV, 300 ms; (shape wave) 0.05 mV, 50 ms (a), 100 μm (b), 5 μm (c, d). Reproduced from Klausberger T, Magill PJ, Marton LF, et al. (2003) Brain state- and cell type-specific firing of hippocampal interneurons *in vivo*. *Nature* 421(6925): 844–848, copyright Nature Publishing Group 2003.

studies (detection of two mRNAs in one cell) indicate that many of the somatostatin mRNA-containing neurons are, as shown by earlier immunohistochemistry studies, GABAergic neurons expressing the mRNA for the GABA synthetic enzyme glutamate decarboxylase (Gad67). GABA can influence the release of somatostatin, and this may allow both feedback and feed-forward effects on neuronal firing. In the hippocampus, somatostatin and cortistatin neurons correspond to distinct groups of partially overlapping interneuronal populations which, like the interneurons in the striatum, can coordinate the activity of their output cells (pyramidal cells) in a temporally distinct and brain state-dependent manner (Figure 3).

Distribution of Somatostatin Receptors

Like somatostatin itself, the mRNAs for several of the somatostatin receptors are concentrated in the forebrain, especially in the cerebral cortex, hippocampus, and amygdala (Figure 4), where the respective receptor proteins are also concentrated. The variant sst_{2b} in the mouse is concentrated in the pons/medulla, with only small amounts of receptor mRNA being found in the cerebral cortex. In contrast to sst_1, sst_{2a} and sst_4 mRNAs are found in the Purkinje and cerebellar granule cell layers and in the motor neurons of the spinal cord. Although small amounts of sst_5 mRNA are found in the brain, it seems to be mainly a peripheral receptor found in the lymphoid cells, pituitary, pancreas, and stomach.

Physiological Roles of Somatostatin

Somatostatin's major role in regulating the release of GH is mediated by the neurons of the hypothalamic periventricular and paraventricular nuclei. Somatostatin exerts most of its effects indirectly by regulating the release of GHRH. Based on the distribution of receptors in this region, the major receptor subtypes likely to be involved in the response of the GHRH-containing neurons are sst_1 and sst_2, with perhaps some involvement of sst_4 in the arcuate nucleus. Somatostatin can, of course, also act directly on receptors on the somatotrophs.

In the CNS, somatostatin and its receptors are important modulators of neuronal excitability, in which they influence GABA and acetylcholine release. Electrophysiological studies indicated somatostatin can inhibit neuronal Ca^{2+} currents, modulate inwardly rectifying K^+ currents, and depress excitatory postsynaptic glutamate currents. The ability of somatostatin to inhibit Ca^{2+} currents was one of the first electrophysiological effects observed for the peptide and is mediated through a PTX-sensitive receptor, indicating coupling to G_i or G_o. The most prominent electrophysiological effect of somatostatin, and also cortistatin, is probably due to the sst_2 receptor inducing a G-protein-coupled inwardly rectifying K^+ current, a response which leads eventually to the inhibition of neuronal firing. Likewise, the ability of somatostatin or its analogs to depress glutamatergic excitatory postsynaptic potentials (EPSPs) also reduces neuronal

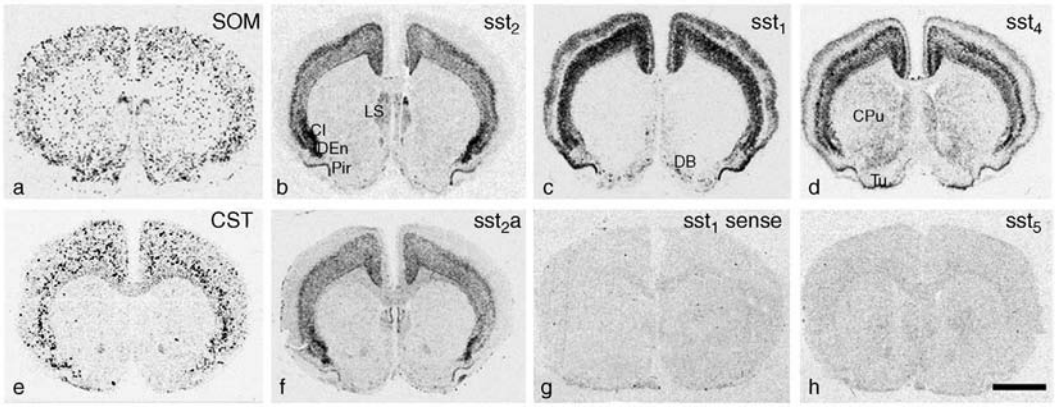

Figure 4 X-ray autoradiograms of hybridized coronal sections of the rat forebrain at the septal level: (a) somatostatin (SOM) mRNA expressed by scattered neurons throughout the forebrain; (b) sst_2 mRNA as detected by hybridization with a probe directed against the entire coding region; (c) sst_1 mRNA highly expressed in the supra- and infragranular layers of the cortex and in the area of the diagonal band (DB); (d) sst_4 mRNA highly expressed in the cortex and striatum (CPu); (e) cortistatin (CST) mRNA expressed selectively in cortex; (f) sst_2 mRNA as detected by hybridization with a sst_{2a}-selective probe directed against the entire coding region; (g) representative image showing control hybridization with a sst_1 sense probe; (h) specific labeling not detected after hybridization with a probe for sst_5 mRNA. In (b) and (f), sst_2 mRNA is expressed at high level in infragranular layers in the cortex and in the area of the lateral septum (LS). Scale bar = 3 mm (h). Cl, claustrum; DEn, dorsal endopiriform nucleus; Pir, piriform cortex; Tu, olfactory tubercle. From Stumm RK, Zhou C, Ara T, et al. (2004) Somatostatin receptor 2 is activated in cortical neurons and contributes to neurodegeneration after focal ischemia. *Journal of Neuroscience* 24(50): 11404–11415, with permission of the Society for Neuroscience.

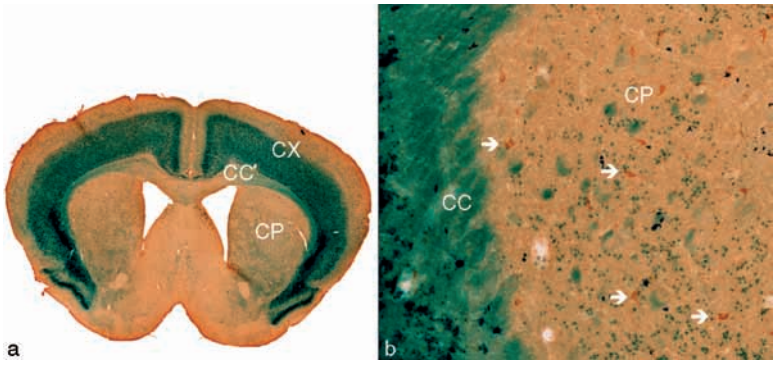

Figure 5 Localization of β-galactosidase staining in a sst_2 knockout–knock-in mouse: (a) cerebral cortex (CX) and corpus callosum (CC); (b) caudate putamen (CP). In the sst_2 knockout–knock-in mouse, the expression of the sst_2 (SSTR2) receptor has been replaced by the bacterial reporter gene LacZ or β-galactosidase. In (a), note the strong blue staining of neurons in the CX and blue β-galactosidase fibers running in the CC. In (b), immunohistochemical staining for the cholinergic marker enzyme choline acetyltransferase (brown peroxidase reaction) shows that some cholinergic neurons contain β-galactosidase (blue) positive inclusions, indicating that some of these CP neurons express the sst_2 receptor.

firing and has a general anticonvulsant and antiepileptic effect. The localization of somatostatin-containing neurons and receptors points to possible roles in epilepsy, motor-behavior cognition, and analgesia. Individual knockout mice in which the receptor or the peptide has been knocked out by homologous recombination in general produce evidence for these three broad physiological roles for somatostatin. For example, somatostatin-deficient mice were more vulnerable to kainic acid (an excitatory toxin) and had impaired motor learning, whereas mice with the sst_2 receptor knocked out were less susceptible to seizures but so far have not been investigated for their pain thresholds. Although anatomical studies indicate that the sst_2 receptor is likely to be the receptor involved in nociceptive transmission in the spinal cord, our study of our sst_2 knockout mice indicated an impairment in fine motor control mediated, probably, through cortical inputs to the striatum (Figure 5). The sst2 knockout mice that we used had a reporter bacterial enzyme β-galactosidase (LacZ) inserted, or knocked in, to replace the missing sst_2 receptor protein so that the enzyme activity of β-galactosidase revealed on the tissue section (LacZ) reports the sites of expression of the missing receptor (Figure 5). It is clear, however, that compensatory changes often occur in knockout mice and that the lack of an obvious phenotype for the various knockout mice indicates, as might be expected, that the diverse function of somatostatin makes the clear assignment of roles for receptor types difficult. It is to be hoped that regional and inducible knockouts, as well as other genetic techniques such as the use of short hairpin (Sh) or small interfering (Si) RNA to selectively disturb regional receptor expression, may enable the physiological roles of somatostatin, cortistatin, and their receptors to be further defined.

Summary

Somatostatin, or SRIF, is a cyclic peptide which inhibits the release of GH from the pituitary. Apart from its effects in the pituitary, somatostatin is found in a number of neurons in the CNS, where it may act as a neurotransmitter/neuromodulator. Peptides related to somatostatin include cortistatin (found principally in the forebrain) and the fish peptide UII. Somatostatin and cortistatin exert their effects in the CNS through G-protein-coupled receptors, of which five ($sstr_{1-5}$) have been cloned and localized in the CNS (although $sstr_5$ is found only in a small amount in the CNS). Outside the CNS, somatostatin is found in endocrine cells of the pancreas and in the intestine, where it is also found in neurons of the enteric nervous system. Somatostatin's effects are broadly inhibitory and a number of stable analogs have been developed to treat acromegaly and endocrine tumors; these may also be of value as analgesics or in treating neurological disease such as epilepsy and stroke.

See also: Alzheimer's Disease: Neurodegeneration; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: Transgenic Mouse Models; Animal Models of Alzheimer's Disease; Cell Replacement Therapy for Huntington's Disease; Huntington's Disease; Huntington's Disease: Neurodegeneration; Mammalian Neuropeptide Families; Neural Coding of Spatial Representations; Neuropeptide Release; Neuropeptides: Discovery; Transgenic Models of Neurodegenerative Disease.

Further Reading

Brazeau P, Vale W, Burgus R, et al. (1973) Hypothalamic polypeptide that inhibits the secretion of immunoreactive pituitary growth hormone. *Science* 179(68): 77–79.

de Lecea L (2005) Cortistatin: A natural somatostatin analog. *Journal of Endocrinological Investigations* 28(supplement 11): 10–14.

de Lecea L, Criado JR, Prospero-Garcia O, et al. (1996) A cortical neuropeptide with neuronal depressant and sleep-modulating properties. *Nature* 381(6579): 242–245.

Epelbaum J, Dournaud P, Fodor M, and Viollet C (1994) The neurobiology of somatostatin. *Critical Review of Neurobiology* 8(1–2): 25–44.

Fodor M, Kordon C, and Epelbaum J (2006) Anatomy of the hypophysiotropic somatostatinergic and growth hormone-releasing hormone system minireview. *Neurochemical Research* 31(2): 137–143.

Guillemin R (2005) Hypothalamic hormones a.k.a. hypothalamic releasing factors. *Journal of Endocrinology* 184(1): 11–28.

Hokfelt T, Efendic S, Johansson O, Luft R, and Arimura A (1974) Immunohistochemical localization of somatostatin (growth hormone release-inhibiting factor) in the guinea pig brain. *Brain Research* 80(1): 165–169.

Klausberger T, Magill PJ, Marton LF, et al. (2003) Brain-state- and cell-type-specific firing of hippocampal interneurons in vivo. *Nature* 421(6925): 844–848.

Maguire JJ (2003) Discovering orphan receptor function using human *in vitro* pharmacology. *Current Opinion in Pharmacology* 3(2): 135–139.

Mori M, Sugo T, Abe M, et al. (1999) Urotensin II is the endogenous ligand of a G-protein-coupled orphan receptor, SENR (GPR14). *Biochemical and Biophysical Research Communications* 265(1): 123–129.

Olias G, Viollet C, Kusserow H, Epelbaum J, and Meyerhof W (2004) Regulation and function of somatostatin receptors. *Journal of Neurochemistry* 89(5): 1057–1091.

Patel YC (1999) Somatostatin and its receptor family. *Frontiers in Neuroendocrinology* 20(3): 157–198.

Saito T, Iwata N, Tsubuki S, et al. (2005) Somatostatinregulates brain amyloid beta peptide Abeta42 through modulation of proteolytic degradation. *Nature Medicine* 11(4): 434–439.

Selmer I, Schindler M, Allen JP, Humphrey PPA, and Emson PC (2000) Advances in understanding neuronal somatostatin receptors. *Regulatory Peptides* 90(1–3): 1–18.

Somogyi P and Klausberger T (2005) Defined types of cortical interneuron structure space and spike timing in the hippocampus. *Journal of Physiology* 562(1): 9–26.

Stumm RK, Zhou C, Ara T, et al. (2004) Somatostatin receptor 2 is activated in cortical neurons and contributes to neurodegeneration after focal ischemia. *Journal of Neuroscience* 24(50): 11404–11415.

Sugo T, Murakami Y, Shimomura Y, et al. (2003) Identification of urotensin II-related peptide as the urotensin II-immunoreactive molecule in the rat brain. *Biochemical and Biophysical Research Communications* 310(3): 860–868.

Tostivint H, Joly L, Lihrmann I, et al. (2006) Comparative genomics provides evidence for close evolutionary relationships between the urotensin II and somatostatin gene families. *Proceedings of the National Academy of Sciences of the United States of America* 103(7): 2237–2242.

Weckbecker G, Lewis I, Albert R, Schmid HA, Hoyer D, and Bruns C (2003) Opportunities in somatostatin research: Biological, chemical and therapeutic aspects. *Nature Reviews Drug Discovery* 2(12): 999–1017.

Sonic Hedgehog and Neural Patterning

H Roelink, University of Washington, Seattle, WA, USA

© 2009 Elsevier Ltd. All rights reserved.

Ventralization of Naive Neural Epithelium by Signals from the Notochord and Prechordal Plate

During early development, the neural tube starts out as an epithelial sheet, the neural plate. At this early stage of development, cells in the neural plate epithelium are largely identical. Over time, this epithelial sheet rounds up to form the neural tube, and differentiated neural cells appear at stereotypic times and positions. Classic transplantation experiments have shown that the nature of the differentiated cells arising from the neural tube is dependent on molecular cues in the environment that are initially released by tissues surrounding the neural tube. In particular, the ectoderm overlying the neural tube at its dorsal aspect and the notochord located below its ventral aspect are prime sources of dorsalizing and ventralizing signals, respectively, for cells in the neural tube (Figure 1).

The notochord is a direct derivative of the node and is, like the node, an important signaling center in the developing embryo, both for the adjacent somites and for the overlying neural tissue. Transplantation of a notochord to a position dorsal to the neural tube causes a mirror image duplication of the ventral half of the neural tube, indicating that signals derived from the notochord are responsible for the induction of ventral cell types in the neural tube. The principal mediator of the inductive properties of the notochord is the signaling molecule Sonic hedgehog (Shh). Similarly, ectodermal explants can induce dorsal cell types in neural plate tissue, and it appears that most of the dorsalizing signals released by the ectoderm are members of the bone morphogenetic protein (BMP) family of signaling molecules.

Superimposed on the signals mediating the dorsal–ventral pattern are signals that mediate positional information along the anterior–posterior axis, restricting the subsequent development of the neural plate to unique identities, such as forebrain, midbrain, hindbrain, and spinal cord. Shh plays an important role in patterning of the neural tube, and at least early in development, but its main function is the induction of ventral cell types in the developing neural tube. The absence of Shh has little effect on patterning along the anterior–posterior axis, indicating that other signals are involved in inducing differentiation along that axis.

Shh Patterns the Developing Spinal Cord

The role of Shh in neural tube patterning has been most extensively studied in the developing spinal cord since this structure represents the simplest and best understood part of the central nervous system. In the spinal cord, it has become obvious that the role of Shh extends well beyond the induction of motor neurons.

In the ventral–medial aspect of the spinal cord, distinct cell types can be recognized. At the ventral midline is the floor plate, which itself starts expressing Shh. At increasing distances from the floor plate, five distinct groups of neurons can be identified: V3 interneurons, motor neurons, and V2, V1, and V0 interneurons. The ventral neural tube is subdivided into five progenitor domains by the combinatorial expression patterns of at least seven distinct homeodomain proteins. These progenitor domains generate the five ventral neuronal subtypes – V0–V3 interneurons and motor neurons. The homeodomain proteins that mediate this code are divided into class I and class II homeodomain proteins. The class I factors are induced by BMP signals originating from the dorsal neural tube, and the class II homeodomain proteins are induced by Shh. Different thresholds of induction of the class I and class II homeodomain proteins, combined with the mutual antagonism of pairs of these proteins, result in precise domains in the ventral neural tube characterized by stereotypic combinations of class I and class II homeodomain proteins. The combinations of class I and class II homeodomain proteins determine the type of neurons into which these precursor cells will differentiate (Figure 2).

The loss of the Shh gene in mice causes a severe embryonic lethal phenotype characterized in the neural tube by the almost complete absence of ventral cell types, resulting in a ventral expansion of intermediate/dorsal cell types to all positions in the neural tube (Figure 2). This loss of ventral cell types is evident along the entire anterior–posterior axis of the embryo. In the developing spinal cord, the loss of Shh results in the failure of all ventral cell types to form in the developing neural tube. The loss of multiple cells types, as a consequence of the absence of a single signaling molecule, supported the idea that Shh is present as an activity gradient away from its source, which is instrumental in the induction of pattern along the dorsal–ventral axis of the developing neural tube.

Shh Is a Morphogen

Immediately after its induction from the ectoderm, the neural epithelium is homogeneous without overt

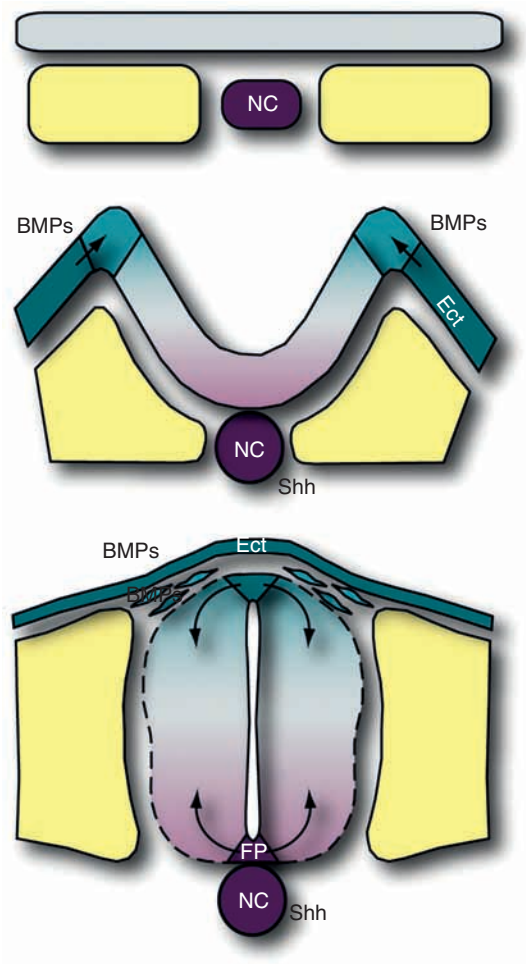

Figure 1 Neurulation diagram. The neural plate receives ventralizing signals from the notochord (NC) and dorsalizing signals from the adjacent/overlying ectoderm (Ect). Shh derived from the Notchord and floor plate (FP) is distributed in a ventral-to-dorsal gradient, and BMP activity derived from the ectoderm and dorsal neural tube is distributed in a dorsal-to-ventral gradient.

dorsal or ventral cell differentiation. Shh released from the notochord causes stereotypical changes in the overlying neural epithelium (Figure 3). At the position where the notochord is in direct contact with the neural plate, the floor plate is induced. The floor plate is a nonneuronal structure in the ventral midline of the neural tube which concomitant with its induction also begins to express Shh (Figure 4). At longer distances from the notochord and floor plate, motor neurons and interneurons are induced (Figures 2 and 4). Importantly, all these inductive events are mediated by Shh and not secondary signals induced by Shh. The explanation by which Shh can induce several distinct cell types is deceptively simple: Close to the sources (notochord and floor plate), cells in the neural epithelium are exposed to a high concentration of Shh, and this concentration gradually declines further away from the source due to limited transport and the presence of local 'sinks,' places where Shh is degraded (Figure 3). The resulting ventral-to-dorsal concentration gradient of Shh is interpreted by the cells in the neural epithelium such that cells exposed to a high concentration of Shh differentiate in floor plate cells and cells exposed to intermediate concentrations differentiate into motor neurons, whereas dorsal to the motor neurons, cells differentiate into interneurons in response to low levels of Shh.

Experimental support for this model came from experiments using explants from the neural plate in tissue culture. In these explants, all ventral cell types could be induced using increasing concentrations of a soluble form of Shh, ShhNp. Very high concentrations (>10 nM) resulted in the induction of floor plate cells. At decreasing ShhNp concentration thresholds, the predominant differentiation in the neural tube explants was V3 interneurons, motor neurons, and V2–V0 interneurons. The latter cell types require a Shh concentration less than 1 nM.

Although this morphogen model provides a reasonable explanation of how a single inducer can cause a stereotypic pattern in a bilaterally symmetrical way with columns of specific cells running parallel to the floor plate, a complication is caused by the physical properties of Shh since its mature form has two lipophilic modifications. First, an autocatalytic proteolysis necessary to yield an active N-terminal Shh peptide (ShhNp) results in the addition of a cholesterol moiety to the C-terminus of ShhNp. Second, an acyltransferase facilitates the addition of a palmitic acid. These two lipophilic moieties cause Shh to be associated with membranes. This membrane association is not easily reconciled with the free diffusion thought to be critical in creating a morphogenic gradient. Two models have been put forward to explain how ShhNp exerts its action at multiple cell diameters away from the source. One model proposes the formation of ShhNp of soluble multimers that diffuse through the extracellular space. The second model proposes an active mechanism whereby membrane-associated ShhNp is actively transported from cell to cell – a mechanism termed planar transcytosis (Figure 5).

The membrane association of Shh requires the existence of a regulated mechanism of export. Dispatched (Disp) is a transmembrane protein that contains a sterol-sensing domain and its expression is required in cells synthesizing the sterol-modified Shh but not necessary for the secretion of a form of Shh lacking the sterol conjugation. *Disp1* null mice resemble *Shh* null mice, indicating that Disp1 is required for proper Shh signaling. Since *Disp1* null cells are equally sensitive to a mutant form of Shh that is not membrane

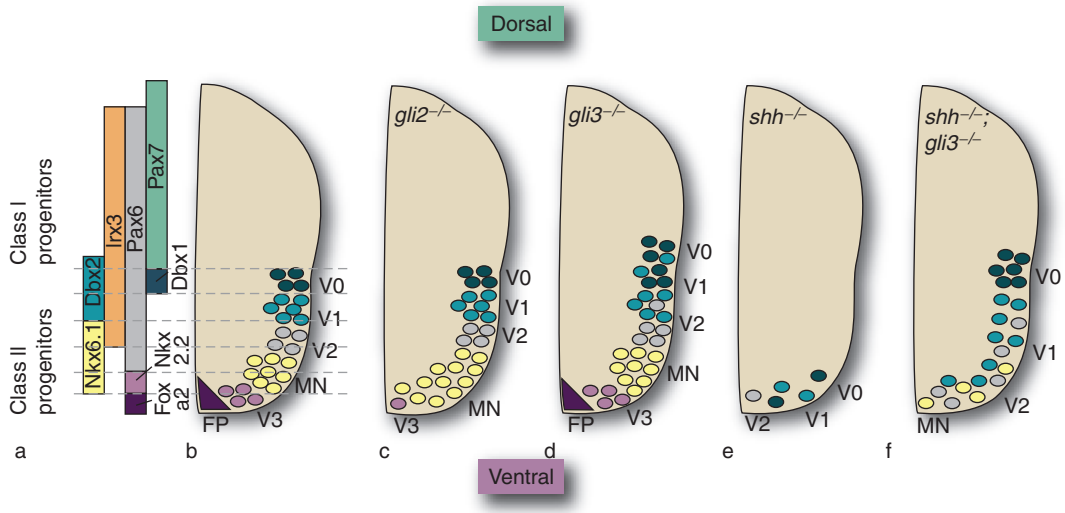

Figure 2 Diagram of class I/class II expression domains and the corresponding regions of ventral neuronal differentiation. A gradient of Shh signaling in the ventral spinal cord activates expression of class II progenitor proteins such as Nkx2.2 and Nkx6.1 while repressing expression of class I progenitor proteins such as Pax7, Pax6, Dbx1, Dbx2, and Irx3 (a). A gradient of BMP activity from the dorsal spinal cord is also able to alter the expression of progenitor proteins such as Pax6 and Pax7. Mutually repressive interactions between class I and class II proteins result in defined boundaries of protein expression, and the ensuing expression pattern of progenitor proteins within discrete regions establishes domains of neuronal differentiation. For example, the expression of Dbx2, Irx3, and Pax6, but not Dbx1, gives rise to the V1 progenitor domain (pV1) from which V1 interneurons (V1) differentiate. Differentiation of ventral neurons along the dorsoventral axis is marked by expression of Engrailed1 (En1) in V1 interneurons. *gli1*$^{-/-}$ mice have phenotypically wild-type spinal cords (b), whereas loss of *gli2* results in a loss of floor plate (FP) and most V3 interneurons (c). *gli3*$^{-/-}$ mice have a mild spinal cord phenotype with a dorsal expansion and intermingling of V0–V2 interneurons (d). Deletion of *shh* results in a loss of most ventral cell types (e), most of which are restored upon additional removal of *gli3* (*shh*$^{-/-}$; *gli3*$^{-/-}$) (f). This suggests that much of the *shh*$^{-/-}$ spinal cord phenotype is mediated by Gli3. MN, motor neurons.

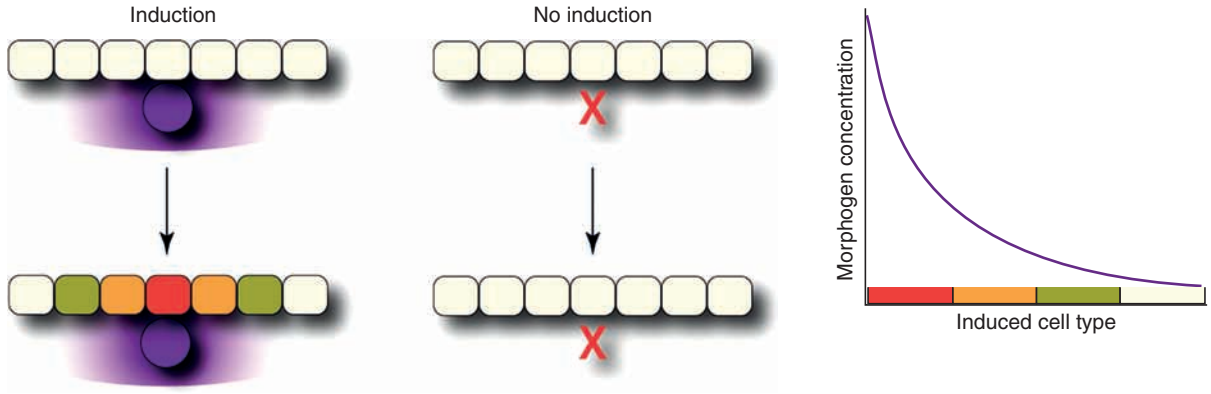

Figure 3 Diagram of morphogen action. Shh released from the notochord (purple) establishes a gradient along the neural plate. The local gradient is interpreted by cells in the neural plate which undergoes a bilaterally symmetric differentiation. In the absence of a notochord or Shh signaling, ventral cell types are not indicated (X).

bound, it has been postulated that Disp1 only plays a role in the cells that produce Shh, possibly mediating the release of Shh multimers from Shh source cells. It remains a possibility that Disp1 also plays this role in cells that have internalized Shh as part of their response. Disp1 could be an integral part of the planar transcytosis pathway by facilitating the export of Shh of cells that have taken it up and making Shh available to neighboring cells.

Shaping the Shh Gradient

Regardless of whether Shh travels from cell to cell or diffuses long distances, it interacts with components in the extracellular space and these molecules can shape the Shh gradient (Figure 4). Known molecules that bind Shh in the extracellular space are Patched1 (Ptc1, a part of the Shh receptor complex), Scube2, Hhip, the related molecules CDO and BOC, and HSPGs (Figure 6).

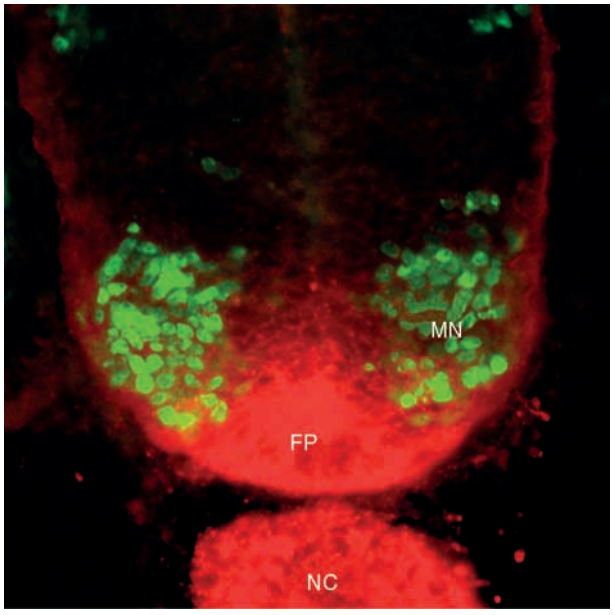

Figure 4 Shh and motor neuron distribution in the developing spinal cord. Cross-section of an embryonic chicken spinal cord, with the ventral/medial aspect shown. Significant amounts of Shh are present in its sites of synthesis, the floor plate (FP) and notochord (NC), but detectible levels of Shh are present in the ventral neural tube in the domain where motor neurons (MN) are differentiating.

Ptc1 plays a particularly interesting role in the distribution of Shh. Ptc1 binds Shh and regulates the activation of the Shh response. However, this binding also results in trafficking of the Ptc1/Shh complex to endosomes and lysosomes, providing a 'sink' for Shh. This role of Ptc1 in sharpening the Shh gradient is exacerbated by the observation that Ptc1 expression is upregulated by the Shh response, increasing the sink effect on the Shh gradient. CDO and BOC are transmembrane, cadherin-associated immunoglobulin superfamily members that bind Shh and are required for the presentation of Shh to Ptc1 and thus for initiating the Shh response. The range of Shh signaling is increased in mice that lack the feedback activities of Ptc1 and also when Hhip (hedgehog interacting protein) is absent. Thus, a Shh-dependent antagonism, mediated by Ptc1 and Hhip, is critical for normal dorsal–ventral development of the neural tube. In contrast, the loss of Scube2, a molecule present in the extracellular space of the neural plate, decreases the efficacy of Shh signaling. However, it is unclear whether scube2 acts by inhibiting the activity of BMPs, and thus reducing the activity of this Shh antagonist, or alternatively by enhancing the range of Shh signaling, possibly by binding to Shh resulting in a complex that is easier to transport in the extracellular space.

A rather complex picture arises with regard to the forms of Shh present in the extracellular space, where there appear to be multiple distinct molecules around molecules that bind to Shh and they are required either for a normal response or for normal distribution. Given that due to its lipophilic modifications, Shh is a nonsoluble molecule that nevertheless travels considerable distances, it is not surprising that many molecules are involved in this process. The nature of the complexes or multimers of Shh that travel from cell to cell remains obscure.

The Shh Response

The Shh receptor complex consists of at least two proteins, Ptc1 and Smo. Ptc1 is a large protein with 12 transmembrane segments sharing structural similarity with Disp1, whereas Smo may be an atypical member of the G-protein-coupled receptor family. Ptc1 is the ligand-binding component and in the absence of Shh it inhibits Smo activity, which is the signaling component that activates the Shh response pathway (Figure 6). Smo is constitutively active in *Ptc1*-deficient cells, and in *Ptc1*-deficient embryos the whole neural tube resembles the floor plate. Although the precise mechanism by which Ptc1 inhibits Smo is unknown, there is evidence that Ptc1 functions as a pump of small, vitamin D-like molecules that inhibit Smo. Consistent with this is the Shh-induced trafficking event culminating in the segregation of Smo into an endocytic compartment not containing Ptc1, which is a critical event in the initiation of the Shh response. Most likely, this process involves the function of intraflagellar transport proteins, indicating that transport along microtubules is an integral part of the Shh response. When segregated away from the inhibitory effect of Ptc1, Smo activates the Shh response likely via an inhibitory G-protein.

Inside the responding cells, three Gli proteins (Gli1–3) which are zinc finger-containing transcription factors, are the transcriptional effectors of the Shh response. In the absence of Shh signaling, Gli2 and Gli3 are proteolytically processed into repressors of Shh target genes. A protein complex that includes SuFu tags Gli3 for cleavage at a proteosome (Figure 6). An early event in the transduction of the Shh response involves the phosphorylation of this multiprotein complex within 15–30 min after Shh binding. These phosphorylation events result in the inhibition of Gli2 and Gli3 cleavage, and thus more of the activator forms of Gli2 and Gli3 are present at the expense of the cleaved inhibitor forms. The vast majority, if not all, of the effects of Shh signaling are mediated by Gli2 and Gli3. A simple model emerges of how cells interpret the Shh activity gradient. In the absence of Shh, all Gli3 is converted into the repressor

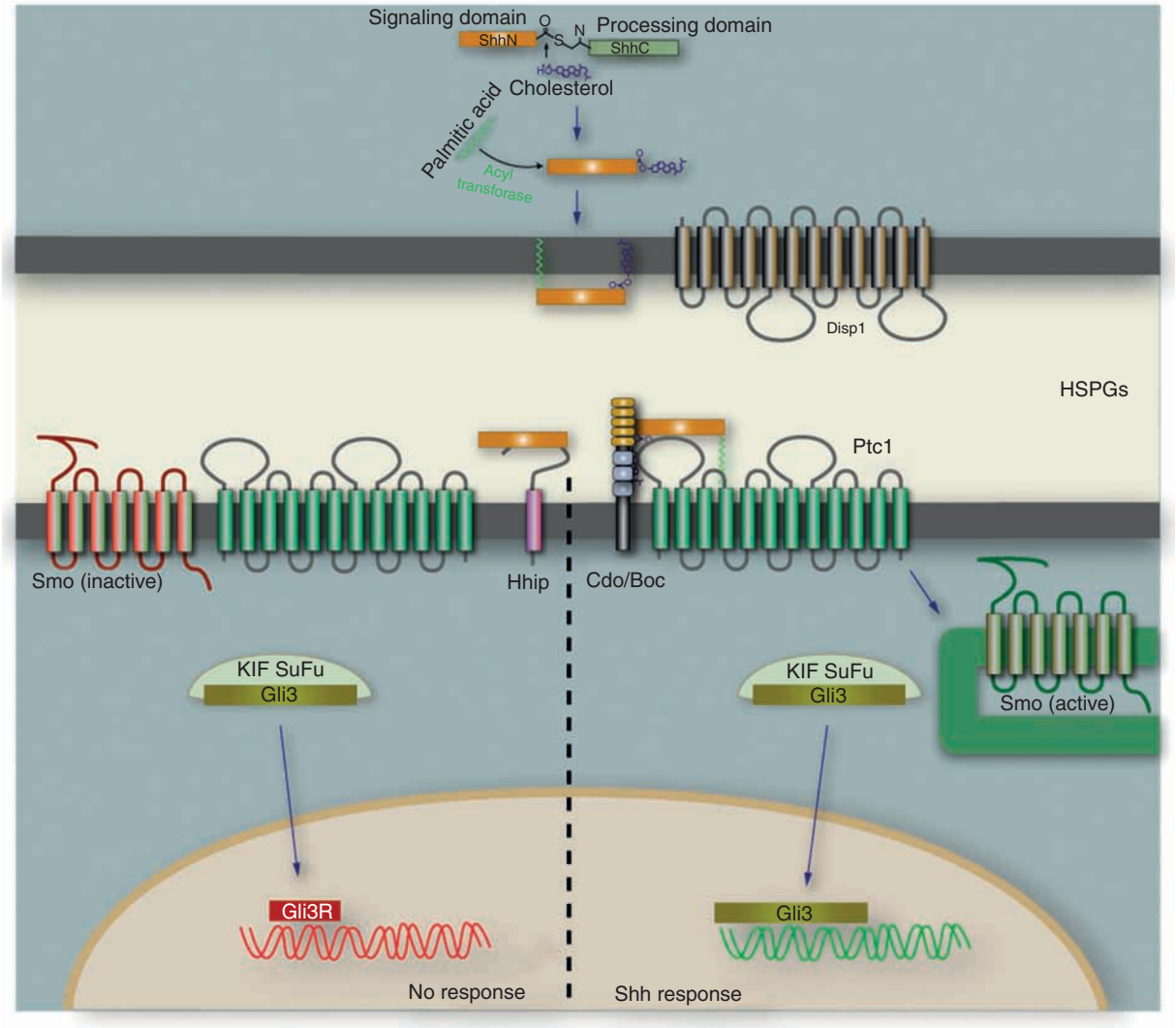

Figure 5 Models of Shh transport mechanisms. The two fundamental mechanisms are diffusion and planar transcytosis. The pathways have in common that Shh is released from the source by Disp1. The first pathway would involve the diffusion of a multimeric, soluble form of Shh across the developing neural epithelium. The second pathway evokes the activity of Ptc1 and Disp1 to create a 'bucket brigade' for Shh, effectively handing over internalized Shh to neighboring cells, which do the same.

form, and with increasing levels of Smo activation, the ratio of Gli2 and Gli3 activator forms over the Gli3 repressor form increases in a graded manner, reflecting the Shh morphogenetic gradient.

The phenotypes in the spinal cord caused by the loss of Gli2 and Gli3 are relatively mild (Figure 2). In the absence of Gli2, the floor plate and adjacent cells fail to form, but motor neurons are present, although their domain spans the ventral midline. The loss of Gli3 has no significant effect in the developing spinal cord, although an expansion of ventral cell types can be observed in the developing brain. These relatively minor effects might seem difficult to reconcile with the dramatic phenotype observed in the *Shh* null mice. However, Shh signaling would have two independent effects on Gli activation: The ventral-to-dorsal Shh gradient would establish a ventral-to-dorsal gradient of activated Gli2 as well as a dorsal-to-ventral gradient of Gli3 repressor activity. This would result in a somewhat redundant situation in which loss of either gene causes a much milder phenotype than the loss of Shh or Smo. However, the opposing gradients of activated Gli2 and the Gli3 repressor form are required for normal dorsal–ventral patterning in the neural tube.

Is Shh Required for Neural Tube Patterning?

The severity of the phenotype caused by Smo and Shh loss-of-function phenotype makes a compelling argument that Shh is critically important for the induction

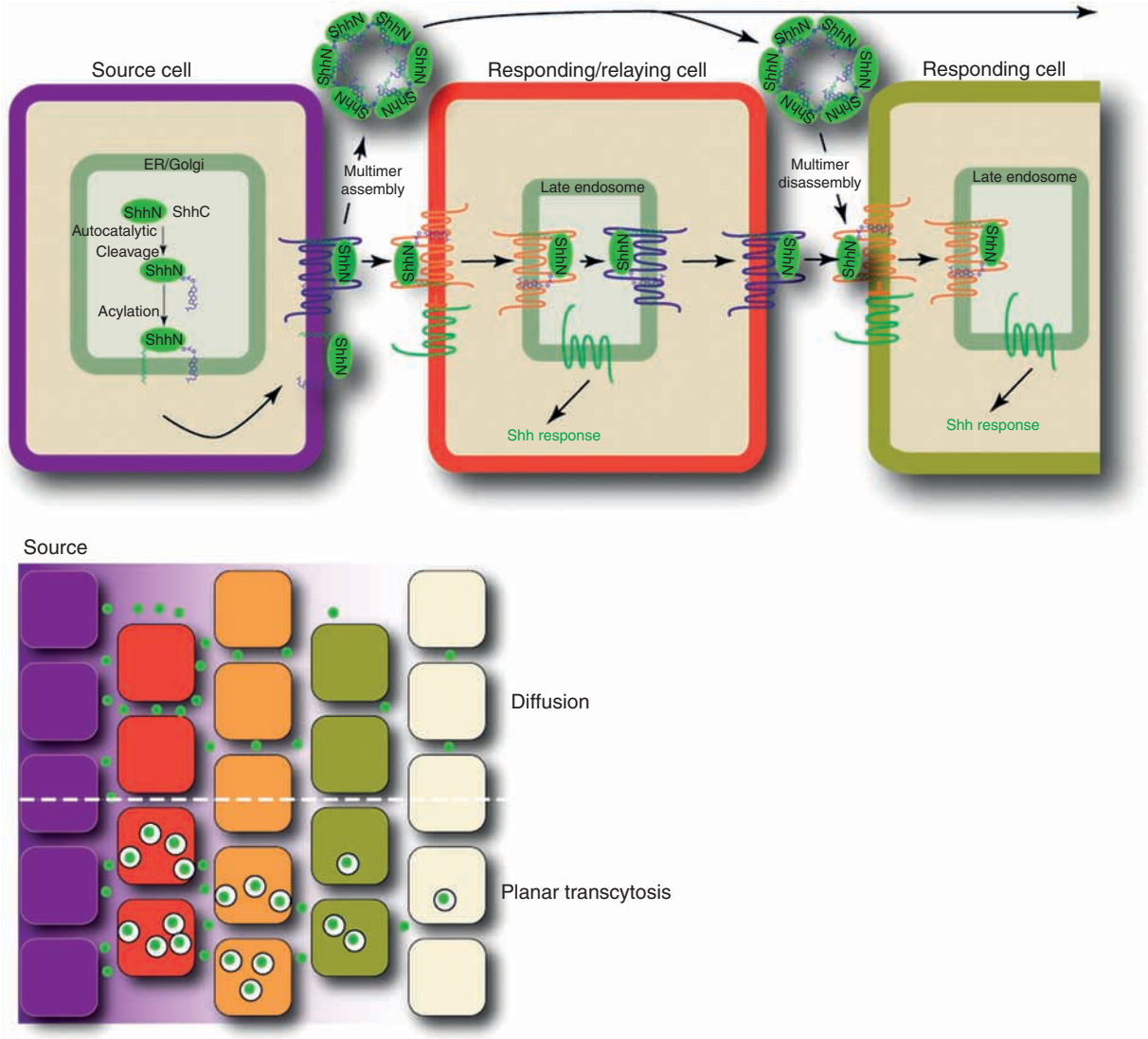

Figure 6 Diagram of Shh synthesis, posttranslational modification, export, and response. Normal processing of full-length Shh results in a molecule with N-terminal palmitic acid and a C-terminal cholesterol moiety. Disp1 is involved in the excretion of Shh. On the recipient cell, Hhip prevents binding to the receptor, whereas CDO/Boc are required for presentation of Shh to Ptc1. In the absence of hedgehog, Patched represses the activity of Smo. Gli3 is cleaved and converted to a repressor in a process that involves SuFu and scaffold proteins. The Patch-mediated inhibition of Smo is released upon binding of Shh to Patched. Gli3 is converted to a transcriptional activator and activates the transcription of hedgehog responsive genes.

of ventral cell types in the neural tube. However, in the absence of Shh signaling, some of the loss of ventral cell types can be rescued when Gli3 is also absent. This is consistent with the repressor activity mediated by Gli3 in the absence of an activated Shh response (Figure 2). In other words, the loss of Shh signaling can be partially overcome by removing the main repressor of the Shh response as well. In the absence of an activated Shh response and addition of Gli3 repressor activity, some ventral neurons such as motor neurons differentiate in the developing spinal cord. This demonstrates that the repressor activity of Gli3 is almost single-handedly responsible for the phenotype observed in the Shh or Smo null mice. Although the most ventral cell types, such as floor plate and V3 interneurons, are not present in such mice, other ventral cell types do appear, although not in the distinct dorsoventral domains characteristic for normal embryos. Nevertheless, this indicates that Shh-independent positional signals are present in the neural tube, and likely candidates for this activity are members of the BMP family, which are expressed in the dorsal neural tube and are predicted to be distributed in the neural tube in a dorsal-to-ventral gradient. It thus appears that although the Shh signaling system is not required for the formation of most

ventral cell types in the neural tube, the precise segregation of the differentiated cells into specific domains required the graded Shh signal.

During normal development of the neural tube, dorsally derived BMP signals antagonize the Shh response and vice versa. Activation of the BMP response is sufficient to induce dorsal cell types, whereas activation of the Shh response induces ventral cell types. Together, these overlapping and antagonistic gradients ensure a very robust mechanism that mediates dorsal–ventral patterning in the neural tube.

Shh and Patterning of the Brain

The forebrain (prosencephalon) is the most anterior aspect of the neural tube, and it is divided by the zona limitans intrathalamica into anterior and posterior parts. The prechordal plate is the rostral-most extension of the notochord, and it is located ventral to the anterior forebrain. The importance of the prechordal plate was already appreciated more than 70 years ago, when ablation experiments showed that the loss of the prechordal plate causes cyclopia, a disorder in which instead of two bilateral eyes, a single midline eye is present, as well as an array of associated malformations such as a single, undivided cerebral hemisphere (holoprosencephaly). The midline eye presents a barrier for the migration of the frontonasal process into its appropriate position to form the nose, philtrum, incisors, and part of the palate. Instead, this tissue accumulates in a proboscis–like structure above the single midline eye. The single midline eye is a consequence of the failure of the formation of ventral midline tissue in the anterior forebrain, which normally separates the eye fields into two bilateral domains. Homozygous loss of Shh causes cyclopia as well, indicating that much of the ventralizing activity of the prechordal plate is mediated via the Shh response pathway. As in the spinal cord, in the absence of Shh signaling the ventral forebrain acquires a more lateral phenotype.

Cyclopia is invariably associated with holoprosencephaly, a condition characterized by a single undivided cerebral hemisphere. However, holoprosencephaly is a much more developmental abnormality. Nevertheless, a principal cause of holoprosencephaly is an absent or attenuated Shh response. In line with the morphogenetic function of Shh, an attenuated Shh response can result in some loss of ventral structures, reflected in the face by malformations such as a single midline incisor, the absence of the nasal septum, and hypotelorism (closely spaced eyes). Holoprosencephaly is highly associated with these birth defects. It is unclear why the failure of the ventralizing activities of Shh to function normally is so closely correlated with holoprosencephaly, which involves the dorsal midline of the telencephalon, seemingly far away from the sources of Shh. In the ventral forebrain, prechordal plate and notochord are required for the induction of the hypothalamus, which is involved in many of the homeostatic functions of the organism.

Notochord-derived Shh also induces ventral structures in the midbrain, the tegmentum. Importantly, the ventral tegmentum is where dopaminergic neurons are localized in the substantia nigra. These neurons are preferentially lost in Parkinson's disease. Obviously, much research is being directed to recreate dopaminergic neurons *in vitro*, and invariably these approaches rely on the activation of the Shh response to generate such neurons from stem cells.

The third cranial (oculomotor) nerve also emerges from the ventral midbrain, and the motor neurons innervating the muscles that move the eyes are located in nuclei just ventral to the cerebral aqueduct. The oculomotor nuclei contain the most rostrally located motor neurons that connect to striated, voluntary muscles. Caudal to the oculomotor nuclei are a series of nuclei in the midbrain and hindbrain, which contain the motor neurons innervating the muscles in the face and neck, all of which require notochord and floor plate-derived Shh for their induction.

Small Molecule Inhibitors of Smo

More than a decade of research has demonstrated that Shh is critical for the formation of ventral cell types in the developing neural tube, and the nature of these ventral cell types is dependent on the position along the anterior–posterior axis. At any position along this axis, Shh functions as a morphogen and as such is responsible for the induction of a variety of cell types. However, because the loss of Shh signaling either through the loss of the ligand or through the loss of Smo causes the embryos to die before birth, not much is known about many of the later events mediated by Shh in neural development since this analysis requires the generation of multiple conditional knockouts.

An alternative approach has been made possible with the use of small molecule inhibitors of Smo, such as cyclopamine. Due to its ability to cause cyclopia in offspring of ruminants, this compound, which was isolated from *Veratrum californicum*, has served as a lead compound in the identification of several inhibitors of Smo. These molecules are used to address the requirement for Shh signaling at later stages of neural development as well as in normal function of neural stem cells.

See also: Anterior-Posterior Spinal Cord Patterning of the Motor Pool; Axonal Pathfinding: Guidance Activities of Sonic Hedgehog (Shh); Drosophila Apterous Neurons: from Stem Cell to Unique Neuron; Floor Plate Patterning of Ventral Cell Types: Ventral Patterning; Olfactory Neuron Patterning and Specification; Retinoic Acid Signaling and Neural Patterning; Segmentation: Spinal Cord Segmentation and A–P Somite Patterning; Wnt Pathway and Neural Patterning.

Further Reading

Cohen MM Jr. (2006) Holoprosencephaly: Clinical, anatomic, and molecular dimensions. *Birth Defects Research Part A: Clinical and Molecular Teratology* 76: 658–673.

Ishibashi M, Saitsu H, Komada M, and Shiota K (2005) Signaling cascade coordinating growth of dorsal and ventral tissues of the vertebrate brain, with special reference to the involvement of Sonic hedgehog signaling. *Anatomical Science International* 80: 30–36.

Jacob J and Briscoe J (2003) Gli proteins and the control of spinal-cord patterning. *EMBO Reports* 4: 761–765.

Litingtung Y and Chiang C (2000) Control of Shh activity and signaling in the neural tube. *Developmental Dynamics* 219: 143–154.

Marti E and Bovolenta P (2002) Sonic hedgehog in CNS development: One signal, multiple outputs. *Trends in Neurosciences* 25: 89–96.

Prakash N and Wurst W (2006) Genetic networks controlling the development of midbrain dopaminergic neurons. *Journal of Physiology* 575: 403–410.

Roussa E and Krieglstein K (2004) Induction and specification of midbrain dopaminergic cells: Focus on SHH, FGF8, and TGF-beta. *Cell and Tissue Research* 318: 23–33.

Ruiz i, Altaba A, Nguyen V, and Palma V (2003) The emergent design of the neural tube: Prepattern, SHH morphogen and GLI code. *Current Opinion in Genetics & Development* 13: 513–521.

Salie R, Niederkofler V, and Arber S (2005) Patterning molecules; Multitasking in the nervous system. *Neuron* 45: 189–192.

Wang Y, Ng EL, and Tang BL (2006) Rab23: What exactly does it traffic? *Traffic* 7: 746–750.

Wilson L and Maden M (2005) The mechanisms of dorsoventral patterning in the vertebrate neural tube. *Developmental Biology* 282: 1–13.

Sound Localization: Neural Mechanisms

D J Tollin, University of Colorado at Denver and Health Sciences Center, Aurora, CO, USA
T C T Yin, University of Wisconsin, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Auditory guided behavior depends on the ability of the nervous system to construct an accurate representation of the sounds in the environment. Toward this end, there are two primary functions of the auditory system. The first function is to determine what it was in the environment that produced a sound. The second function is to determine where in space the sound occurred. Knowledge of both the spatial position of a sound's source and what produced it facilitates the initiation of an appropriate behavioral response. For example, a predator like the cat might move toward the sound produced by a mouse rustling in the leaves while the mouse might freeze or move away from the sounds of the approaching cat.

The ability to locate sensory stimuli is not unique to the auditory system as it is also shared by the visual and somatosensory systems. The locations of visual and tactile objects are encoded directly in these systems due to the topographic, spatial organization of the receptors, the rods and cones of the retina, and the mechano-receptors of the skin, respectively. In contrast, in the auditory system, the hair cells of the cochlea are designed to encode sound frequency and intensity and cannot sense sound location directly. Therefore, sound location must be computed centrally in the auditory system based on the neural representations of the spectral and temporal characteristics of the acoustic stimuli arriving at the two ears.

In order to understand the neural mechanisms of sound location we first have to identify the acoustic information bearing elements of the sounds arriving at the two ears that specify location. There are three primary acoustical characteristics of sounds, or cues, which are used to determine the spatial position of a sound source. These acoustical cues are physically different because they are created by different ways that the propagating sound waveforms interact with the head and external ear. The fact that there are three distinct cues naturally suggests that three different kinds of neural mechanisms are needed to encode them. Indeed, the cues are encoded in anatomically separate, but parallel, pathways in the ascending auditory system. The evidence that each cue is encoded in a different pathway comes from experimental data from disparate scientific disciplines including neurophysiology, behavioral psychology, and comparative neuroanatomy.

Three Main Acoustical Cues to Sound Source Location

The acoustical cues for sound localization are generated by the interaction of the propagating sound waves with the anatomical structures supporting the receptors of the ear, namely the head and the external ear (or pinnae), and not by the topographic organization of the receptors themselves. Examples of the three cues are shown in Figure 1. Two of the cues, interaural time delays (ITDs) and interaural level differences (ILDs), result directly from the fact that the two ears themselves are sampling the sound waves at different locations due to their positions on opposite sides of the head. These two cues specify primarily sound location in azimuth. The third cue, spectral-shape, results from the way that the sound waves are distorted or filtered by the complicated convolutions and folds of the pinna. Spectral shape cues specify primarily sound location in elevation, and whether a sound is in the front of or behind the observer. Interpretation of the roles of the various auditory nuclei that process the sound localization cues requires knowledge of what the cues are and how those cues change with changes in source location.

Interaural Time Delays

ITDs result because the ears are physically separated in space by the head. Therefore, the direction-dependent differences in path lengths that sound must travel to reach each ear from the source will generate different times of arrival of the sound at the two ears, or ITDs (Figure 1(b)). For example, a sound on the right side will arrive at the right ear a short time before it arrives at the left ear, while a sound from directly ahead will arrive at the ears at the same time. The range of ITDs is directly proportional to head diameter so species with larger heads, hence larger distances between the two ears, exhibit a larger range of ITDs. For example, in cats, which have head diameters of 6–7 cm, ITDs increase monotonically with azimuth along the horizontal plane up to a maximum of 350–400 μs (Figure 1(d)). The diameter of the human head is nearly twice that of the cat, on the order of 15 cm. Consequently, the maximum ITD in humans is ~800 μs. Yet smaller mammals, like some species of bats and rodents, have head diameters <1 cm, resulting in maximum ITDs as small as 40 μs. Such small ITDs are not thought to be useful for accurate sound localization.

Interaural Level Differences

The fact that the two ears are separated by an obstacle, the head, results in the ILD cue to sound location.

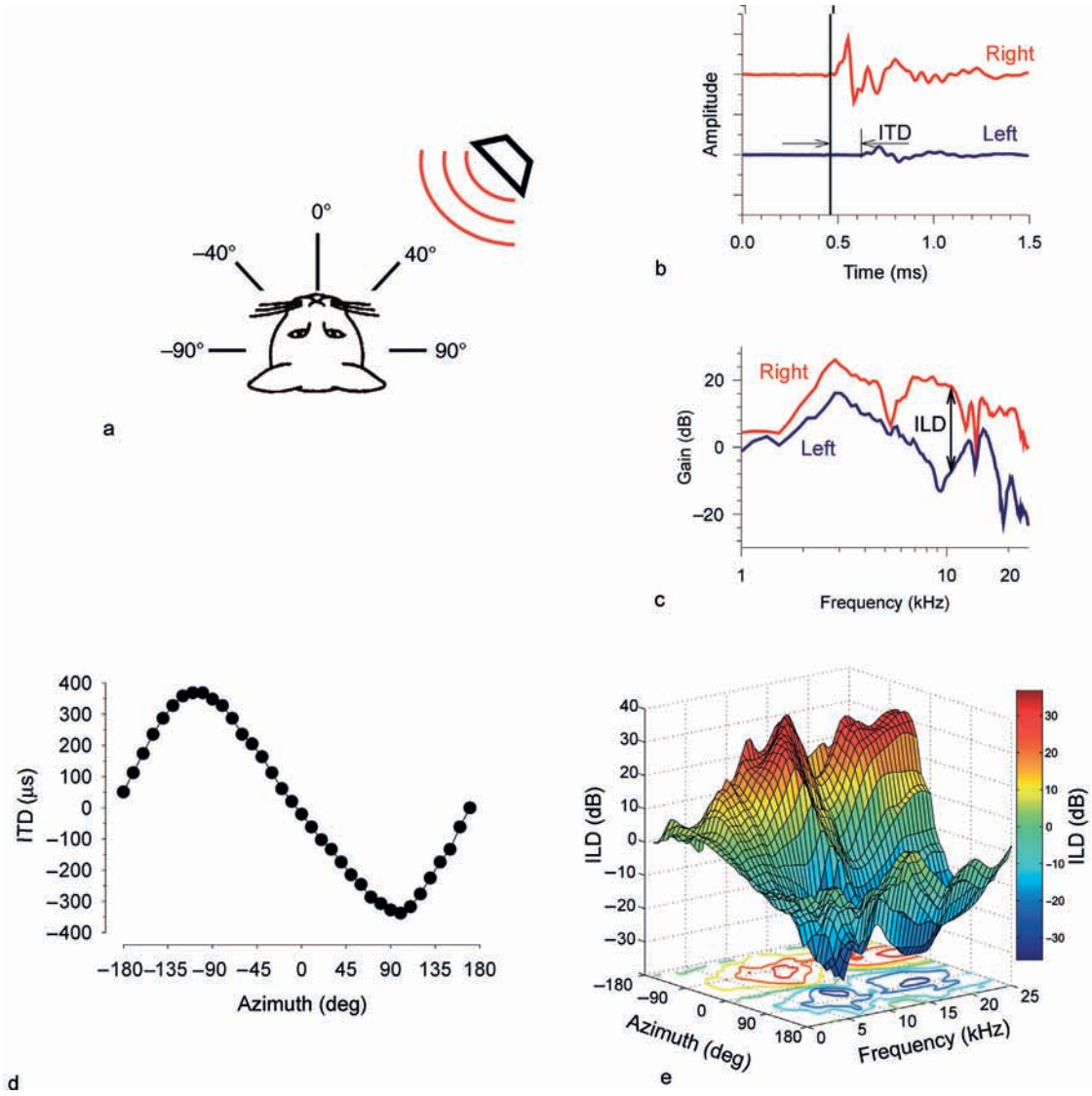

Figure 1 The acoustical cues to sound location. In the example, the cues are shown for the cat, but the same three cues are available in all species, although the magnitude of the cues will be different. (a) A broadband transient sound is presented from a loudspeaker at (40°, 0°). (b) The resulting acoustical responses to the transient near the eardrums in the left and right ears. The difference in the onset times of the sounds at each ear yields the interaural time difference (ITD) cue and difference in amplitudes reflect the interaural level difference (ILD) cue. (c) The frequency spectrum of the acoustical responses in (b). The spectral shape cues are captured by the changes in the shapes of the spectra as a function of source location. The ILD cue is the difference in sound level computed at each frequency. (d) The ITD plotted as a function of sound source azimuth. ITDs are minimal near the midline and maximal for sources to either side. (e) The joint spatial and frequency dependence of ILDs for sources varying along the horizontal plane. ILDs are a complicated function of azimuth and frequency for high stimulus frequencies. For illustrative purposes, sources were restricted to the horizontal plane and only in the frontal hemisphere for frequencies between 1 and 25 kHz.

For sounds of high frequency (or short wavelength), the head essentially creates an acoustic shadow for the far ear as sounds with wavelengths on the order of the diameter of the head and smaller are reflected off the near side of the head. Consequently, the resulting sound arriving at the ear farthest from the source is effectively attenuated thereby creating direction-dependent differences in the amplitudes, or levels, of the sounds that reach the two ears (Figure 1(c)). The sound level is greatest at the ear nearest the source, and lower in the ear farthest from the source. The ILD cue is a rather complicated function of both sound location and, at any one location, sound frequency (Figure 1(e)). As expected, ILDs are small in magnitude for low-frequency sounds and increase in magnitude for high-frequency sounds (Figure 1(e)). Therefore, ILDs are primarily useful for localization of high frequencies. ILDs that are large enough to be useful for sound

localization are available even in mammals with small heads provided that the sound source contains high frequencies with wavelengths shorter than the head diameter.

Monaural Spectral Shape

The ITD and ILD cues are not useful for sound localization in elevation because their values change little with variations in source elevation. The monaural spectral shape cues, however, do change systematically with source elevation. Spectral shape cues arise from direction- and frequency-dependent reflection and diffraction of the pressure waveforms of sounds by the head, torso, and pinna that result in broadband spectral patterns, or shapes, that change with location. The most prominent direction-dependent features in the spectra are the deep notches that occur at some locations (e.g., 5.5 and 9.5 kHz for the right and left ear, respectively, in Figure 1(c)). These spectral notch cues are used both for localization in elevation as well as facilitating the distinction between sound sources in front from sound sources behind of an observer. Spectral cues are created primarily for high-frequency sounds whose wavelengths are on the order of the dimensions of the pinna and its convolutions. The frequency at which the spectral notch occurs varies systematically with sound source elevation, with lower elevations yielding low-frequency notches and higher elevations yielding progressively high-frequency notches.

Neural Processing of Localization Cues in Three Parallel Pathways

For these three acoustical cues to location to be useful for spatial perception, there must be neural mechanisms in the ascending auditory pathways that effectively encodes each of them. Each of the three different cues is encoded in different areas of the auditory brain stem (Figures 2–4). That is, there are three parallel pathways through the brain stem that appear to encode separately the ITD, ILD, and spectral cues. Neurons in these pathways are the earliest places in the ascending auditory pathways to respond selectively to each of the three cues.

ITDs Are Processed in the Medial Superior Olive

The neurons comprising the medial superior olive (MSO) receive excitatory inputs from both ears via the bushy cells of the anteroventral cochlear nucleus (AVCN) on both sides (Figure 2(a)). The bushy cells of both sides receive excitatory inputs from the cochlea via the auditory nerve fibers (ANFs) via specialized synaptic terminals called Endbulbs of Held. The AVCN is organized tonotopically. Most MSO neurons are sensitive to ITDs in sounds presented to the two ears. ITD sensitivity in MSO results from a three-stage process as first hypothesized by a model proposed by Jeffress. First, the afferent inputs to MSO neurons from the ANFs and from the bushy cells of the AVCN carry timing information in the form of phase-locked neural responses. Phase locking refers to the tendency of a neuron to fire action potentials at particular phases of an ongoing periodic sound waveform, such as the sinusoidal waveforms that are typically used in physiological studies of the auditory system. Phase locking is the mechanism by which the peripheral auditory system keeps track of the times of occurrence of the ongoing amplitude fluctuations in sounds. Second, the MSO neurons behave like coincidence detectors, responding maximally only when action potentials from the SBC inputs from the left ear and the right ear arrive nearly simultaneously at the MSO neuron. Third, the axons of the afferent inputs to MSO from the bushy cells form delay lines along one axis of the nucleus due to differences in the neural path length from the bushy cells to the MSO neuron. Differences in neural path length result in differences in neural conduction times to the MSO which are then offset by a physical ITD cue. Since the projections to the contralateral MSO have a naturally longer path length than those to the ipsilateral MSO, most MSO neurons primarily encode ITDs produced by sounds on the contralateral side where the ITD can compensate for the longer conduction time from the contralateral ear (Figure 2(b)). In addition, different MSO neurons have different length delay-line inputs, so different MSO neurons will be maximally sensitive to different ITD values. Moreover, MSO neurons have a bias toward low-frequency sounds, as most MSO neurons are sensitive to predominantly low sound frequencies. Phase locking is best for neurons sensitive to low-frequency sounds. Behaviorally, ITDs are most useful for localization for low-frequency sounds. Most MSO neurons send excitatory afferents to the ipsilateral dorsal nucleus of the lateral lemniscus (DNLL) and the ipsilateral central nucleus of the inferior colliculus (ICC). As a consequence, low-frequency sensitive neurons in the DNLL and IC encode the ITDs of sounds primarily on the contralateral side.

In recent years the Jeffress model has been challenged on several grounds, chiefly by the finding that the MSO also receives inhibitory input from both sides. Current research efforts are aimed at understanding the role of this inhibition and what modifications to the simple model are needed.

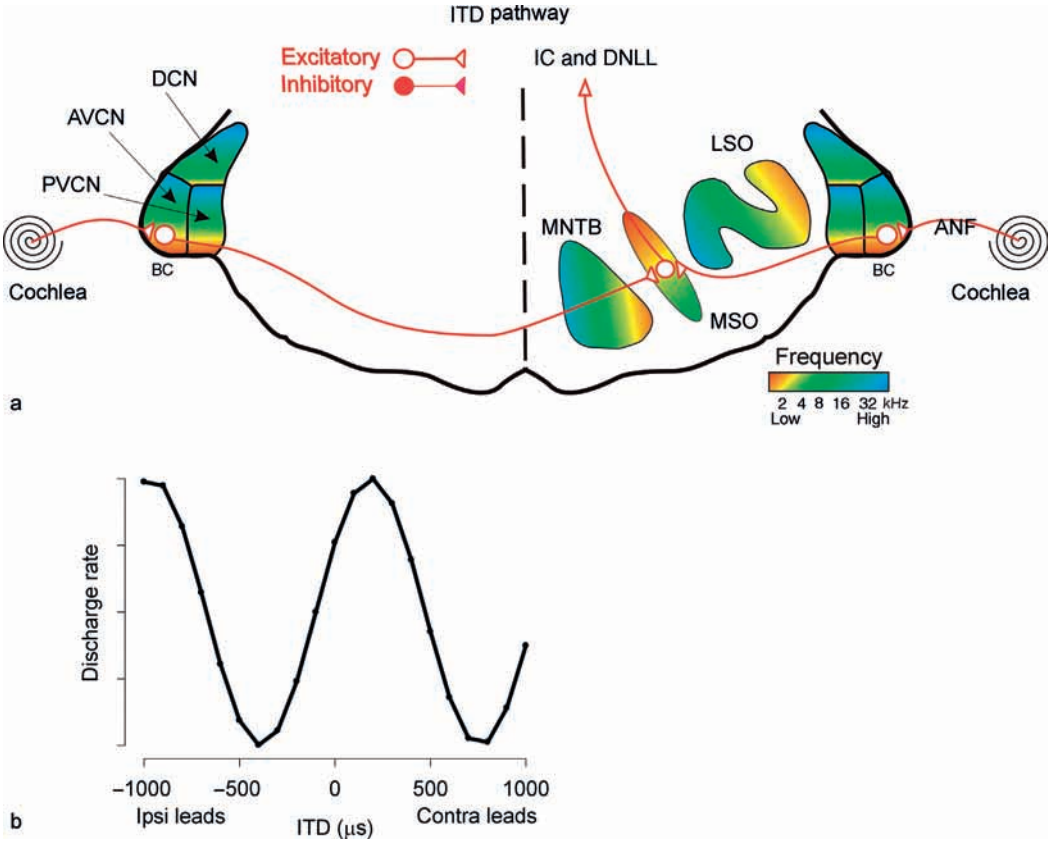

Figure 2 Illustration of a frontal section through the auditory brain stem showing the ascending pathways through the medial superior olive (MSO), which is believed to be responsible for encoding interaural time differences (ITDs). (a) MSO neurons receive bilateral excitatory inputs from both ears via the bushy cells (BC) of the anteroventral cochlear nucleus (AVCN). The bushy cells receive inputs from the auditory nerve fibers (ANFs). The AVCN and the MSO are tonotopically organized. The MSO is biased toward low-frequency sensitive neurons. The color bar and shading indicates, very roughly, the tonotopic organization and the approximate ranges of frequencies that the neurons in each nuclei are sensitive to in the cat. MSO neurons send an excitatory projection to the ipsilateral inferior colliculus (IC) and dorsal nucleus of the lateral lemniscus (DNLL). Open and closed symbols represent excitatory and inhibitory neurons and synapses, respectively. Also shown are the posteroventral cochlear nucleus (PVCN), the dorsal cochlear nucleus (DCN), the medial nucleus of the trapezoid body (MNTB), and the lateral superior olive (LSO). (b) MSO neurons encode ITDs in their response rates. Peak responses are usually found when the contralateral stimulus leads by 0–500 μs.

ILDs Are Processed in the Lateral Superior Olive

Lateral superior olive (LSO) neurons also receive afferent inputs from both ears (Figure 3(a)). The input from the ipsilateral ear is conveyed via auditory nerve synapses with the bushy cells of the AVCN, which then send an excitatory projection to the ipsilateral LSO. The afferent input from the contralateral ear also comes from bushy cells of the AVCN but in this case the bushy cells project across the midline to the neurons of the medial nucleus of the trapezoid body (MNTB). The bushy cell synapse onto MNTB neurons forms the massive calyx of Held; the large size of this unique synapse results in very fast and secure synaptic transmission. MNTB neurons are glycinergic, so their projection to the LSO has an inhibitory effect. In contrast to the MSO, both the MNTB and the LSO are biased predominantly to neurons that are sensitive to high sound frequencies where the ILD cues themselves are prominent. Because of the interplay of the ipsilateral excitation from the bushy cells and contralateral inhibition from the MNTB, LSO neurons essentially compute the difference between the neural representation of the levels of the sounds present at the two ears. Hence, LSO neurons encode ILDs. The sound level at each ear is believed to be encoded by a rate code, in that the number of action potentials elicited over a given time period is proportional to the level of the sound at the ear. Because LSO neurons are excited by sounds at the ipsilateral ear but inhibited by sounds at the contralateral ear, they will respond best to ipsilateral sound sources that produce ILDs favoring the excitatory ear (Figure 3(b)). LSO neurons project bilaterally to the DNLL and the ICC. In order to achieve the contralateral

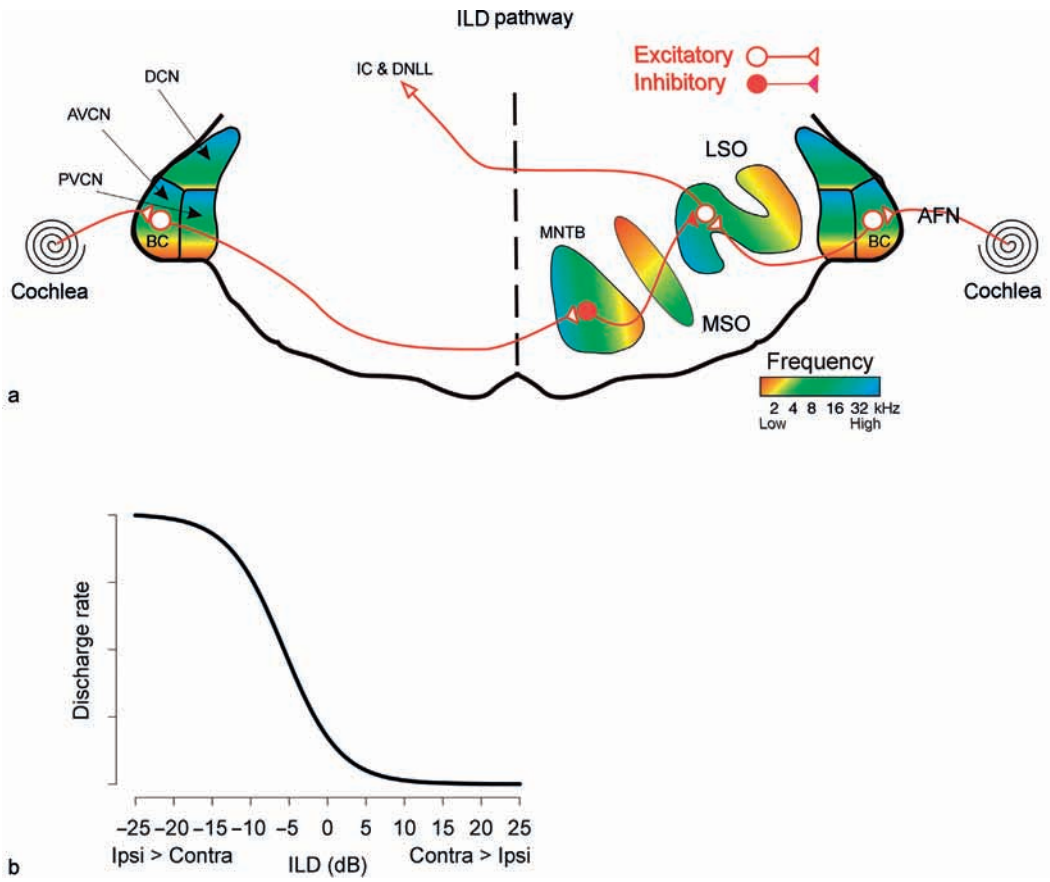

Figure 3 Illustration of a frontal section through the cat brain stem showing the ascending pathways through the LSO, which is believed to be responsible for encoding interaural level differences (ILDs). Symbols and abbreviations as in Figure 2. (a) LSO neurons receive bilateral inputs from both ears. The input from the ipsilateral ear via the bushy cells is excitatory but the input from the contralateral ear via the bushy cells of the contralateral AVCN is inhibitory due to the additional synapse in the ipsilateral MNTB. The ipsilateral excitation and contralateral inhibition allows LSO neurons to encode ILDs (b). LSO neurons send excitatory projections to the contralateral IC and DNLL and inhibitory projections (not shown) to the ipsilateral IC and DNLL. The neurons comprising the MNTB and LSO are predominantly sensitive to high-frequency sounds.

representation of space that is a basic feature of sensory and motor areas of the brain, the high-frequency sensitive neurons of the ventromedial LSO send excitatory projections to the contralateral DNLL and ICC, while the low-frequency sensitive dorsolateral LSO sends predominantly inhibitory projections to the ipsilateral DNLL and ICC. As a result high-frequency sensitive neurons in the DNLL and ICC respond best to contralateral sound sources that produce ILDs favoring the contralateral ear.

Spectral Shape Cues Are Processed in the Dorsal Cochlear Nucleus

Less secure is our understanding of the processing of spectral shape cues. It is hypothesized that the DCN analyzes the spectral shape cues created by the interaction of the sound with the external ear in a direction-dependent fashion. The principal output neurons of the DCN are the fusiform and giant cells (Figure 4). Neural responses of the fusiform and giant cells are similar, so they are often considered as a single neural group. As a neural circuit, the DCN is more complicated than the MSO and LSO/MNTB circuits. As such, it is easier to break down the DCN circuit into a simple qualitative model (Figure 4). In this general model, the type IV neurons, the fusiform and giant cells, represent the principal output neurons of the DCN that project to the contralateral ICC. The type IV neurons are selective for the broadband spectral notches due to the way they integrate excitatory and inhibitory inputs from a variety of sources. The main excitatory inputs arise from the ANFs. Due to the ANF inputs, the type IV neurons are tonotopically organized and are predominantly sensitive to high sound frequencies. Type IV neurons also receive inhibitory inputs from two different neurons. First, the type II neurons, corresponding to the tuberculoventral neurons of the DCN, receive an excitatory input

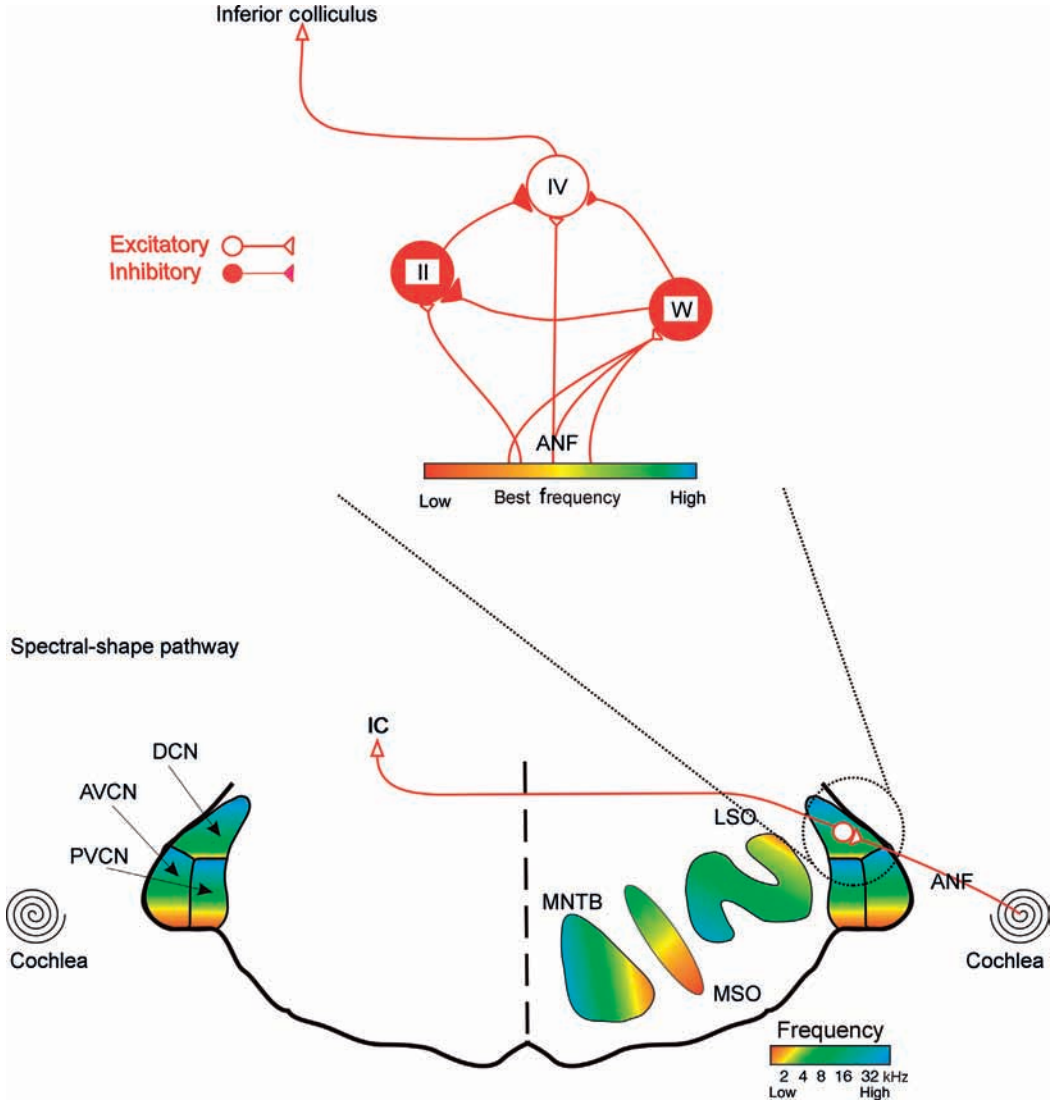

Figure 4 Illustration of a frontal section through the cat brain stem showing the ascending pathways through the DCN, which is believed to be responsible for encoding the spectral-shape cues. Symbols and abbreviations as in Figure 2. The inset shows a qualitative model of the DCN. Type IV neurons represent the principal neurons of the DCN, the fusiform, and giant cells, which receive excitatory inputs directly from the ANFs and inhibitory inputs from the type II and W neurons. Type II neurons, representing the tuberculoventral neurons of the DCN, receive excitatory input from ANFs sensitive to similar frequencies as the direct projection to the type IV neurons and project a strong inhibitory input to the type IV neurons. Type II neurons also receive a strong inhibitory input from the W neurons. The W neurons, representing the so-called wide-band inhibitor neurons, receive excitatory ANF inputs spanning a broad range of sound frequencies, both below and above those of the direct inputs to the type IV neurons. W neurons send a relatively weak inhibitory input to the type IV neurons. The result of this neural network of excitatory and inhibitory inputs allows the type IV neurons, which project to the contralateral inferior colliculus, to respond selectively to the spectral notches that constitute the spectral-shape cues.

from the auditory nerve and send a strong inhibitory input to the type IV neurons. Type II neurons are sensitive to a narrow range of frequencies that is similar to that of the type IV neurons. Type II neurons therefore effectively inhibit type IV responses to narrowband sounds. On the other hand, the wide-band inhibitor neuron, or W neuron corresponding to the D-Stellate neurons of the posteroventral cochlear nucleus (PVCN), receives excitatory inputs from many ANFs spanning a broad range of frequency sensitivity. W neurons in turn project a strong glycinergic inhibitory input to the type II neurons but a much weaker inhibitory input to the type IV neurons. Consequently, W neurons respond best to broadband sounds, which causes a strong inhibition of the type II neurons. Therefore, broadband sounds effectively release the strong inhibition on the type IV neurons from the type II. This complicated inhibitory network from the type II and W neurons, along with a single narrow frequency-tuned excitatory input from the auditory

nerve, allows type IV neurons to respond selectively to broadband spectral notches. When the frequency of the spectral notch cue to sound elevation is aligned with the best frequency of a type IV neuron, that neuron's response is maximally inhibited. The type IV neurons of the DCN send an excitatory projection to the contralateral ICC.

Further Evidence for Processing of Localization Cues in Three Parallel Pathways

The physiological studies reviewed above showed that there are three different nuclei in the auditory brain stem, the MSO, LSO, and DCN, that selectively process each of the three cues to sound location. In support of those studies, there is also evidence from other scientific disciplines that points more to the functional aspects of the encoding of the ITD, ILD, and spectral cues by those three circuits.

Comparative neuroanatomical studies in a large number of mammalian species have found that animals with large heads (like primates) tend to have a large and well-developed MSO, whereas the MNTB and LSO are relatively small and underdeveloped. In contrast, animals with small heads (like bats and most rodents) tend to have a small underdeveloped (in some cases nonexistent) MSO while the MNTB and LSO are large and well developed. Behaviorally, small-headed animals are not able to localize low-frequency sounds based on ITDs, likely due to the fact that the ITDs are quite small and the MSO, which processes ITDs, is also relatively small. On the other hand, mammals with large heads can localize accurately low-frequency sounds based on ITDs because the ITD cues are quite large and the MSO is also well developed. In contrast, small mammals can localize very high-frequency sounds presumably due to the fact that high-frequency sounds yield useful ILD cues to location and the MNTB and LSO which processes ILD cues are well developed. Small mammals tend to be able to hear higher frequencies than large mammals. However, large mammals can also localize high-frequency sounds. The data together implicate the MSO as necessary for encoding ITDs and the MNTB and LSO necessary for encoding ILDs.

Finally, the ability to localize a sound positioned along the horizontal plane, where ITD and ILD cues are predominantly used, is severely impaired in experimental animals where the input or output pathways to the MSO and LSO or the MSO and LSO themselves were selectively lesioned. These experiments show that the interaction of the afferent inputs to the MSO and LSO are necessary for the localization of sound based on ITDs and ILDs. Moreover, in animals where the DCN was selectively lesioned, sound localization ability was impaired chiefly for sources that varied in elevation while localization was nearly normal for sources varying in azimuth.

There are three primary acoustical cues to the spatial location of a sound source. Two cues, ITD and ILD, depend on the fact that the two ears are physically separated in space by the head, which acts as an acoustical barrier. ITD and ILD cues vary maximally for sources along the horizontal plane. Neurons in the MSO and LSO encode the ITD and ILD cues, respectively. Data from comparative as well as lesion studies support the separate processing of those two cues. Spectral shape cues are used for localizing sources varying in elevation. Spectral shape cues are encoded by neurons in the DCN. Lesion studies reveal that the DCN is critical for sound localization in elevation. The MSO, LSO, and DCN send projections to the ICC in such a way that neurons in the ICC encode predominantly sounds in the contralateral field. Neurons at higher auditory stations, including the medial geniculate body in the thalamus and the primary auditory cortex, have also been found to be sensitive to the three acoustical cues to location. However, the manner in which these cues are represented systematically in these higher-order areas as well as how sound source location itself is represented in the auditory system remains elusive.

See also: Auditory Cortex Structure and Circuitry; Auditory Scene Analysis; Auditory System: Central Pathways; Auditory Localization; Cochlear Development.

Further Reading

Boudreau JC and Tsuchitani C (1968) Binaural interaction in the cat superior olive S-segment. *Journal of Neurophysiology* 31: 442–454.

Glendenning KK and Masterton RB (1998) Comparative morphometry of mammalian central auditory systems: Variation in nuclei and form of the ascending system. *Brain Behavior and Evolution* 51: 59–89.

Goldberg JM and Brown PB (1969) Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: Some physiological mechanisms of sound localization. *Journal of Neurophysiology* 32: 613–636.

Heffner RS (1997) Comparative study of sound localization and its anatomical correlates in mammals. *Acta Otolaryngologica Supplement* 532: 46–53.

Jeffress LA (1948) A place theory of sound localization. *Journal of Comparative Physiological Psychology* 41: 35–39.

Masterton B, Thompson GC, and Diamond IT (1967) Role of brainstem auditory structures in sound localization. I: Trapezoid body, superior olive, and lateral lemniscus. *Journal of Neurophysiology* 30: 341–359.

May BJ (2000) Role of the dorsal cochlear nucleus in the sound localization behavior of cats. *Hearing Research* 148: 74–87.

Moore CN, Casseday JH, and Neff WD (1974) Sound localization: The role of the commissural pathways of the auditory system of the cat. *Brain Research* 82: 13–26.

Musicant AD, Chan JC, and Hind JE (1990) Direction-dependent spectral properties of cat external ear: New data and cross-species comparisons. *Journal of the Acoustical Society of America* 87: 757–781.

Oertel D and Young ED (2004) What's a cerebellar circuit doing in the auditory system? *Trends in Neuroscience* 27: 104–110.

Reiss LAJ and Young ED (2005) Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus. *Journal of Neuroscience* 25: 3680–3691.

Roth GL, Kochhar RK, and Hind JE (1980) Interaural time differences: Implications regarding the neurophysiology of sound localization. *Journal of the Acoustical Society of America* 68: 1643–1651.

Tollin DJ and Yin TCT (2002) The coding of spatial location by single units in the lateral superior olive of the cat. I: Spatial receptive fields in azimuth. *Journal of Neuroscience* 22: 1454–1467.

Yin TCT (2002) Neural mechanisms of encoding binaural localization cues in the auditory brainstem. In: Oertel D, Fay RR, and Popper AN (eds.) *Integrative Functions of the Mammalian Auditory Pathway*, pp. 99–159. New York: Springer.

Yin TCT and Chan JCK (1990) Interaural time sensitivity in medial superior olive of cat. *Journal of Neurophysiology* 64: 465–488.

Young ED and Davis KA (2002) Circuitry and function of the dorsal cochlear nucleus. In: Oertel D, Fay RR, and Popper AN (eds.) *Integrative Functions of the Mammalian Auditory Pathway*, pp. 160–206. New York: Springer.

Sox Gene Expression

S R Hutton and L H Pevny, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2009 Published by Elsevier Ltd.

Introduction

A suite of transcription factor families that are essential for the regulation of gene expression in both the developing central and peripheral nervous systems has been identified. One such group is the SOX family. SOX proteins serve many important functions throughout the development of the nervous system, including the induction of neural competence, the maintenance of progenitor characteristics, and cell fate determination. They were identified based upon their sequence identity to the founding member of the SOX family, the mammalian testis-determining factor SRY (sex-determining region Y). The *SRY* gene contains a single 79-amino-acid DNA-binding motif known as the high mobility group (HMG) domain. SOX family members are classified based upon their sequence similarity to the *SRY* HMG domain, with all members sharing at least 50% homology. SOX proteins are grouped into subfamilies based upon their degree of sequence similarity, with approximately 90% similarity in HMG domains within subfamilies. Currently in vertebrates, there are over 20 individual SOX proteins divided into eight subfamilies (A–H), with some subfamilies further subdivided based upon similarities outside of the HMG domain. A number of these SOX subfamilies are expressed in the developing nervous system (Table 1).

SOX factors regulate their downstream targets through multiple mechanisms, including chromatin modification, direct transcriptional activation or repression, posttranslational modification, and protein–protein interactions. All SOX proteins, as well as the related TCF/LEF family of transcription factors, contain a single 'HMG box'; these proteins have been shown to bind the DNA sequence $^{A}/_{T}$ $^{A}/_{T}$CAA$^{A}/_{T}$G with high sequence specificity compared to the classical HMG domain proteins HMGB1 and HMGB2, which contain multiple HMG domains but do not bind DNA with any sequence specificity. Generally, SOX proteins bind the minor groove of DNA, which in turn results in an acute bend in the DNA strand, suggesting that SOX proteins are capable of structurally modifying the DNA. However, SOX proteins have also been demonstrated to act as classical transcription factors, with their C-termini capable of acting as either an activator or repressor domain. In addition, posttranslational modification of SOX proteins also plays a role in their regulation of diverse genes. For example, the SUMO-1 (small ubiquitin-like modifier) protein can bind to SOX9 and SOX10 lysine residues in both the activation domain and immediately 5′ of the proposed E1 protein–protein interaction region, potentially modifying the proteins' stability, localization, and protein–protein interactions. These findings have propelled further studies to identify additional mechanisms that can modify SOX signaling.

The diverse functions of SOX proteins, as well as their broad and overlapping temporal and spatial expression patterns, suggest that SOX factors work in conjunction with tissue-specific cofactors to regulate the expression of their target genes. For instance, SOX2, a member of the SOXB1 subfamily, is expressed in multiple cell lineages and, in each tissue type, interacts with unique factors to regulate the expression of downstream targets. In embryonic stem (ES) cells, SOX2 has been shown to directly cooperate with OCT3/4 (octomer-binding factor) to regulate FGF4 (fibroblast growth factor) expression and thus maintain cell pluripotency. Similarly, in neuroepithelial (NEP) cells, SOX2 interacts with another POU factor, BRN2, in the regulation of the *Nestin* gene, a marker for neural stem/progenitor cells. On the other hand, in the developing lens, SOX2 directly interacts with PAX6 to bind and activate the expression of *δ1-Crystallin*. These data demonstrate the importance of identifying and understanding the interactions of SOX members and their binding partners and how they regulate transcription in the context of neural development.

Role of SOX Factors in Defining Neural Competence

Sox genes are expressed from the very onset of embryonic development. One family, the *SoxB*1 group (*Sox1*, *Sox2*, and *Sox3*), is essential for the acquisition of neural competence in primitive ectodermal cells and for their commitment to a neural fate. As mentioned previously, SOX2 is expressed in ES cells and is required to maintain the pluripotency of these cells in conjunction with other transcription factors, such as OCT3/4 and NANOG. During germ layer specification, both SOX2 and SOX3 are expressed in all ectodermal cells, including those that will constitute the prospective neuroectoderm. In the developing mouse embryo, the restriction of SOX2 and SOX3 expression to ectodermal cells committed to a neural fate coincides with the onset of SOX1 expression in these cells. As evidence for an evolutionarily conserved role of *Sox1*, the *Drosophila* orthologs of *Sox2* and *Sox1*, *Dichaete* (or *Fish-hook*) and *Sox-Neuro*, respectively, show similar confined expression

Table 1 SOX factor function in nervous system development

Family		*Member*	*Function*
B	1	1	Maintains neural progenitor characteristics in mouse embryonic stem and P19 cells
			Regulates migration of mouse telencephalic neurons
			Required for differentiation of neurons in mouse ventral striatum
		2	Maintenance of SOX2 expression biases *Xenopus* animal cap cells and mouse embryonic stem cells toward neural fate
			Maintains neural progenitor characteristics in embryonic stem cells, chick. spinal cord, and mouse retinal and rat oligodendrocyte progenitor cells
			Required for neuronal differentiation in anterior thalamus, dorsal striatum, and septum
			Required for maintenance of chick neural crest progenitor identity
		3	Maintains neural progenitor characteristics in chick spinal cord
			Required in a subset of mouse hypothalamic neurons that regulate the hormonal output of the anterior pituitary
	2	14	Promotes neuronal differentiation, cell cycle exit, and delamination of chick neural progenitors
		21	Promotes neuronal differentiation, cell cycle exit, and delamination of chick neural progenitors
C		4	Promotes activation of differentiated neuronal markers in chick embryonic spinal cord
		11	Promotes activation of differentiated neuronal markers in chick embryonic spinal cord
			Regulates neuronal survival and neurite outgrowth of chick neurons
D		5	Represses specification and terminal differentiation and influences mouse oligodendrocyte migration patterns
		6	Represses specification and terminal differentiation and influences mouse oligodendrocyte migration patterns
E		8	Regulates *Xenopus* and chick neural crest cell migration
		9	Required for neural crest induction in *Xenopus*, chick, and mouse embryos
			Promotes chick oligodendrocyte differentiation
		10	Maintenance and survival of neural crest progenitors in *Xenopus* and mouse embryos
			Induction and survival of glial lineages in *Xenopus* and mouse embryos
			Promotes oligodendrocyte differentiation

patterns in the developing fly. Ultimately, all three SOXB1 factors are then expressed in central nervous system (CNS) neural progenitors throughout ontogeny. Functional studies suggest that these proteins are essential for the establishment of the neural lineage, as inhibiting SOXB1 function in *Xenopus* embryos results in a lack of formation and differentiation of neural tissue. Conversely, overexpression of SOXB1 factors in naive ectodermal cells in *Xenopus* initiates neural differentiation in conjunction with FGF. The precise temporal and spatial restriction of the *SoxB1* genes to the neural ectoderm suggests that they are regulated by neural inducing signals. For instance, overexpression of the neural inducer chordin upregulates SOX2, while overexpression of the antagonistic target of chordin, bone morphogenetic protein (BMP), suppresses SOX2 expression in the *Xenopus* embryo. Similarly, Dpp and Sog, the fly counterparts of BMP4 and chordin, regulate the expression of the *Drosophila* ortholog of *SoxB1* genes, *SoxNeuro*. Furthermore, direct evidence for the regulation of SOX2 expression by factors that regulate neural induction stems from a series of elegant studies in the chick embryo showing that FGFs and WNTs can regulate a *Sox2* enhancer element and activate its expression to initiate neural plate development. Despite these findings, the precise roles of SOXB1 proteins in the extensive regulatory network controlling neural competence remain to be further elucidated.

Role of SOX Factors in the Maintenance of Identity and Differentiation of Neural Progenitor Cells of the Central Nervous System

Individual members of the SOX transcription factor family play essential roles not only in the acquisition of neural fate but also in the maintenance and differentiation of neural progenitor cells of the CNS (Figure 1). The SOXB1 proteins mark a common transcriptional state shared by diverse populations of neural progenitors throughout the CNS during development and in the adult. The expression of *SoxB1* genes directly correlates first with the commitment of cells to a neural fate. Next, after neural induction, all three genes are co-expressed in proliferating neural progenitor cells along the entire anteroposterior axis of the developing vertebrate CNS and are maintained in neurogenic regions of the postnatal and adult CNS. Furthermore, SOXB1 factors mark proliferating neural progenitors in derivatives of the CNS, including the neural retina, the olfactory epithelium, and the inner ear. The importance of SOXB1 factors in the nervous system has been highlighted both by results from misexpression and dominant interfering studies as well as by genetic analyses. Gain-of-function and dominant interference experiments in *Xenopus* and chick embryos, as well as in

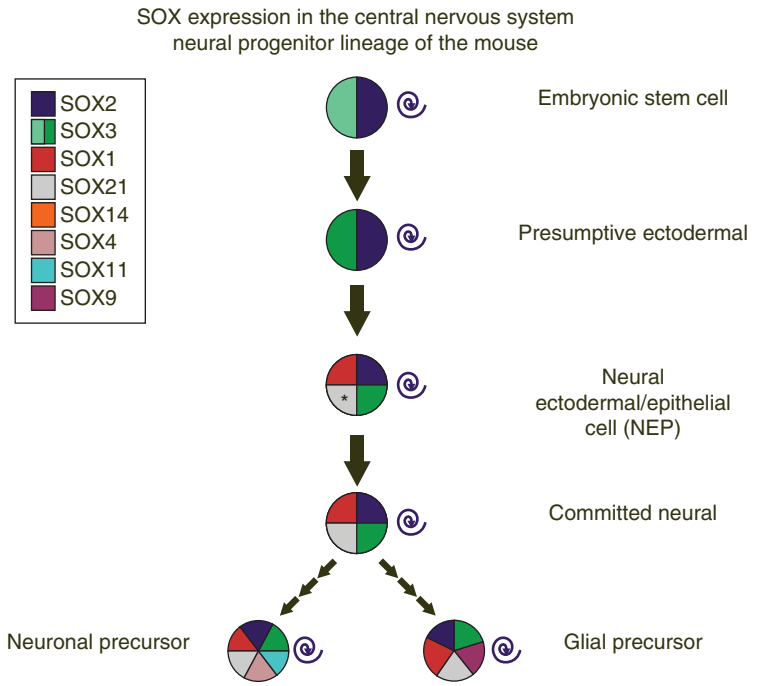

Figure 1 Temporal expression of different SOX proteins in neural progenitor cells of the mouse central nervous system. The asterisk denotes SOX21 expression in the ventricular zone. Light and dark green represent low and high levels of SOX3, respectively; spirals denote proliferative capacity.

mouse cell lines, have shown that SOXB1 signaling plays an essential role in maintenance of neural progenitor identity. These data provide evidence that inhibition of SOXB1 signaling in neural progenitor cells results in their premature delamination from the ventricular zone, their exit from the cell cycle, and the onset of their neuronal differentiation. Conversely, constitutive expression of SOX2 results in the maintenance of progenitor characteristics. These experiments also provide evidence that SOXB1 factors function by antagonizing the actions of proneural genes. Specifically, studies in the chick embryo have shown that the ability of proneural genes to promote neuronal differentiation inversely correlates with the level of SOXB1 expression.

The requirement of SOXB1 factors for the maintenance of neural progenitor identity has been confirmed by genetic studies in a number of species. Analyses of conditional and hypomorphic mutations of *SoxB1* genes in *Drosophila*, zebra fish, mice, and humans have not only verified the absolute requirement of SOXB1 factors but have revealed a dosage-dependent role for them in the maintenance of neural progenitor identity. The identification of hypomorphic mutations and the generation of compound mutations in the *Drosophila Dichaete* and *SoxNeuro* genes, as well as the generation of conditional, hypomorphic, and compound mutations in mouse *SoxB1* genes, permit the assessment of the function of SOXB1 factors in the CNS. In the fly, *SoxNeuro/Dichaete* double-mutant embryos show severe neural hypoplasia throughout the CNS, as well as dramatic loss of Achaete-expressing proneural clusters. These data suggest that members of the *Drosophila SoxB1* subfamily act upstream and in parallel to genes of the *Achaete–Scute* complex. Furthermore, analysis of a *Sox2* hypomorphic allele (Dr11) and directed transgene expression has demonstrated that Dichaete function is necessary for the correct development of midline glia of the CNS. In the mouse, analysis of the tissue-specific conditional ablation of SOX2 in neural progenitor cells, dependent on the Cre–loxP system, has provided the first genetic evidence for the requirement for SOX2 in the maintenance of neural progenitor cell identity. Specifically, retinal progenitor cells with conditionally ablated *Sox2* lose competence to both proliferate and terminally differentiate. In contrast, in *Sox2* hypomorphic/null mice, a reduction of SOX2 expression causes variable microphthalmia as a result of aberrant neural progenitor differentiation. In *Sox2* hypomorphic mutant retinas the decrease in levels of SOX2 expression directly correlates with a decrease in the levels of Notch1 and its direct downstream effector, HES-5, expression. Moreover, and consistent with the observations made in the fly embryo, in mouse neural retinal progenitor cells that express decreased levels of SOX2 due to germ line *Sox2* hypomorphic mutations, the expression of proneural genes such as *Atoh7* (*Math5*) and *Neurod1* is prematurely upregulated. It is important to note that although SOXB1

factors are generally restricted to and are clearly essential for the maintenance of neural progenitor cell identity, their expression has been detected in small subsets of postmitotic neurons in the developing and adult CNS. For example, it has been shown that SOX1 is not only expressed in, but also functions to regulate, the migration of telencephalic neurons of the ventral striatum, and deletion of SOX3 in mice leads to defects in pituitary function and abnormal development of Rathke's pouch.

Unlike the members of the SOXB1 subfamily, for which expression is generally downregulated as neural progenitor cells exit the cell cycle, expression of members of the SOXB2 subfamily, including SOX14 and SOX21, and members of the SOXC subfamily, SOX4 and SOX11, is maintained in postmitotic neurons. Consistent with this, these factors have been shown to play a crucial role to promote neuronal differentiation. Misexpression of either SOX14 or SOX21 in chick neural progenitor cells results in their premature exit from cell cycle, delamination to the differentiated mantle zone, and the onset of expression of markers of mature postmitotic neurons. In contrast, SOX4 and SOX11 function further downstream of proneural basic helix–loop–helix (bHLH) proteins and are essential for the establishment of panneuronal protein expression. Misexpression of SOXC factors results in the aberrant expression of markers of differentiated neurons but does not cause progenitor cells to exit the cell cycle or commit to neuronal differentiation.

SOX factor expression and function is maintained in restricted progenitor cells of both the neuronal and glial lineages. For example, it has been shown that members of the SOXD (SOX5 and SOX6) and SOXE (SOX8, SOX9, and SOX10) families are expressed in the oligodendrocyte lineage. Interestingly, and analogous to the antagonistic function of SOXB1 and SOXB2 factors during neuronal differentiation, both genetic and molecular data suggest that SOX5 and SOX6 directly interfere with the functions of group E SOX proteins. In contrast to SOX9 and SOX10, which promote oligodendrocyte specification and terminal differentiation, SOX5 and SOX6 repress the specification and terminal differentiation of oligodendrocytes and influence their migration.

Role of SOX Factors in the Maintenance and Differentiation of Neural Progenitor Cells of the Peripheral Nervous System

The peripheral nervous system (PNS) is derived from neural crest cells, a subset of dorsal neuroepithelial cells. Similar to the initial specification of neural fate during neural induction, neural crest cell fate is also defined by the functions of SOX family members. Human neuropathies, as well as spontaneous mutations found in the mouse, have provided evidence that one SOX family, SOXE, plays a pivotal role in the PNS. In addition, group C members SOX4 and SOX11, group D member SOX5, and the group B1 member SOX2 are also found expressed transiently during neural crest development (Figure 2).

All three SOXE members (SOX8, SOX9, and SOX10) are initially expressed in the developing neural crest in *Xenopus*, chick, and mouse, although the relative timing of expression onset and the expression level of each SOXE protein is species specific. Generally, as development proceeds, SOX9 is maintained in migrating cranial neural crest progenitors (which later form cranial neurons and glia), while SOX10 is maintained in trunk neural crest progenitors (which give rise to the PNS). SOX8 is maintained in both cell populations.

As mentioned previously, SOX10 is expressed in trunk neural crest progenitor cells that give rise to the PNS. Upon differentiation, SOX10 is maintained in all cells fated for the glial lineage but lost in all neuronal derivatives. Initially, SOX10 does not appear to be necessary for neural crest specification or migration, as SOX10 knockout mice are capable of generating some migrating neural crest cells. However, SOX10 does appear essential for the survival of the neural crest cells, as postmigratory neural crest cells undergo apoptosis in these mice. This, in addition to knockin experiments in which SOX10 was replaced with SOX8, demonstrates the inability of SOX8 to compensate for SOX10 loss. Interestingly, SOX10 morpholino knockdown in *Xenopus* blocks the expression of early neural crest makers SNAIL2 (SLUG), SOX9, and FOXD3. However, given that SOX10 expression succeeds that of these markers, it is suggested that SOX10 maintains the expression of these genes in progenitors.

In addition to regulating the survival of neural crest cells, SOX10 has also been demonstrated to maintain the pluripotency of these cells and to prevent premature neuronal differentiation. Interestingly, SOX10 is also necessary for the establishment of the neuronal lineage, as demonstrated by its functional requirement for the induction of the neurogenic transcription factors MASH1 and PHOX2B in the mouse, as well as Ngn1 in *Xenopus*. Therefore, how can the loss of SOX10 be required for neuronal differentiation, yet its expression is necessary for neurogenesis? This discrepancy appears to be dependent upon the level of SOX10 expression. In mice with a single functional *Sox10* allele, the maintenance of neurogenic potential is unaltered, as evidenced by the maintenance of MASH1 and PHOX2B expression. However, in these same

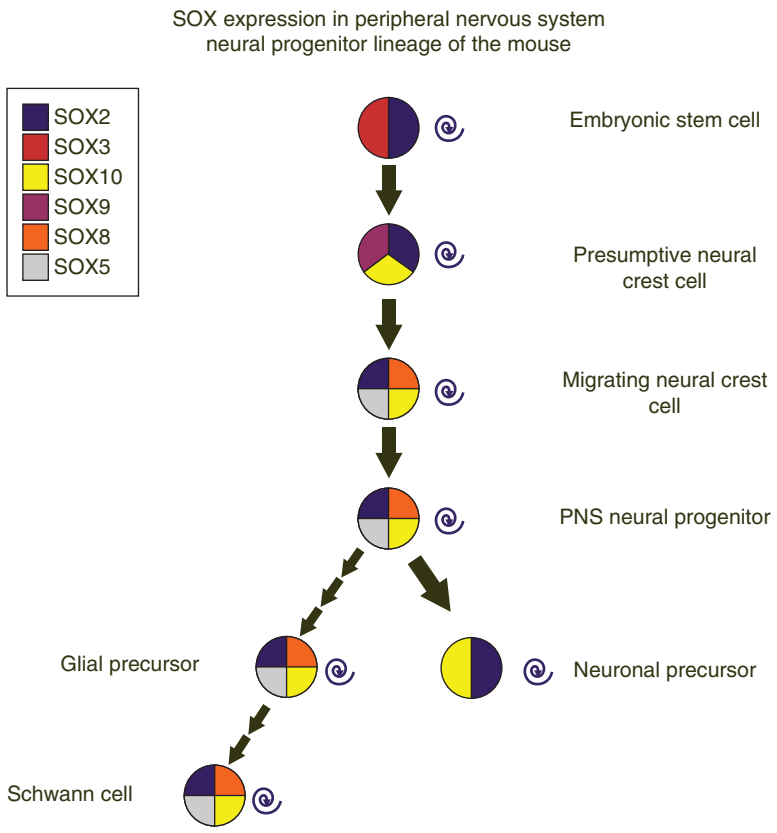

Figure 2 Temporal expression of different SOX proteins in neural progenitor cells of the mouse peripheral nervous system. Expression of SOX proteins in nonneural lineages is not shown; spirals denote proliferative capacity.

animals, the ability to inhibit neuronal differentiation is decreased, as observed by a reduction in the expression of PHOX2A, a marker for autonomic neuronal differentiation. This suggests that SOX10 is downregulated in cells that begin to differentiate but is maintained at sufficient levels to drive them toward a neuronal fate.

The maintenance of SOX10 in all glial cells of the PNS is important for glial fate, and its loss results in complete ablation of all peripheral glia. Molecularly, it has been shown that SOX10 can stimulate gliogenesis by regulating ERBB3 expression and neuregulin signaling. Furthermore, similar to its function in CNS oligodendrocytes, SOX10 also regulates myelination in PNS Schwann cells by controlling myelin protein-zero expression through interactions with the EGR/KROX20 transcription factor. Connexin-32 and ciliary neurotrophic factor (CNTF) are also regulated by SOX10, the latter being important for neuronal maintenance.

Consistent with other vertebrate species, SOX10 is required in human neural crest, and mutations in SOX10 are found in some human neuropathies. For instance, SOX10 mutations result in Waardenburg–Shah syndrome, which is characterized by skin hypopigmentation, deafness, and Hirschprung's disease (aganglionic megacolon), in addition to severe dysmyelination. Mice with little or no SOX10 expression present with similar phenotypes. Additionally, Charcot–Marie–Tooth syndrome may also involve SOX10 mutations.

SOX9, on the other hand, has been shown to be important for neural crest specification, though again species-specific differences are observed. For instance, early knockdown of SOX9 in *Xenopus* and zebra fish results in a lack of induction of neural crest markers such as FOXD3, whereas SOX9 knockout mice maintain normal FOXD3 expression. In addition, ectopic expression of SOX9 in the chick neural tube can activate expression of neural crest markers in the neuroepithelium, though these cells are unable to delaminate and migrate. These results suggest that SOX9 is involved in neural crest specification, but its contributions to this process may vary by species. This is supported by recent studies that have shown that SOX9 is able to bind and activate SNAIL2, a transcription factor necessary for the epithelial–mesenchymal transition and migration of neural crest cells, as well as by the identification of a neural crest specific enhancer for SOX9 that is dependent upon Wnt–β-catenin signaling. Currently, the

role of SOX9 in the survival of neural crest cells is still unclear. Conditional ablation of SOX9 specific to neural crest cells using Wnt1/Cre-expressing mice results in cranial neural crest defects but no observable abnormalities in trunk neural crest. This suggests that either SOX9 or SOX10 (which is expressed in trunk neural crest) is necessary to maintain neural crest survival.

In the mouse, SOX8 appears to serve a minimal role in neural crest development; SOX8 knockout mice do not exhibit any neural crest phenotype, potentially due to compensation by SOX9/SOX10 expression. However, in *Xenopus*, morpholino-mediated knockdown of SOX8 results in defects in proper neural crest migration. Therefore, the relative importance of SOX8 in the developing PNS remains to be determined.

SOXB1 proteins also appear to be involved in the establishment of the PNS, though to what extent remains unclear. The most studied of these proteins is SOX2. In the chick, SOX2 is downregulated, but not extinguished, in cells as they become biased toward the neural crest fate. This low level of expression is maintained in migrating neural crest cells, but upon reaching their targets SOX2 is upregulated in neural precursors in the developing dorsal root ganglia and is ultimately maintained in some glial lineages. Furthermore, neural populations of neural crest cells contain SOX2-specific enhancer activity, and SOX2 was demonstrated to inhibit neural crest formation and to regulate proliferation and differentiation in the PNS neural progenitors. Thus, SOX2 appears to play a similar neurogenic role in both the PNS and the CNS.

The functions of other SOX factors in PNS development are still unclear. Ectopic expression of SOX5 in the chick neural tube has been demonstrated to induce the expression of many neural crest transcription factors. Also, SOX11 has been shown to function in the regulation of neuronal survival and neurite outgrowth in both neural cell lines and primary dorsal root ganglia neuron cultures. Clearly, however, our understanding of the multiple functions of SOX proteins in PNS development is far from exhaustive. The upstream and downstream targets of individual SOX members are only beginning to be elucidated, as are the numerous cofactors with which SOX proteins interact. Only by identifying these mechanisms will the role of SOX proteins in the regulatory network controlling PNS development be understood.

Conclusion

The SOX family of transcription factors is essential for the establishment, development, and maintenance of the nervous system. Though highly conserved, each member appears to serve a distinct function, depending upon the tissue it is expressed in, the level of its expression, and the cofactors that are co-expressed in the cell. The mechanisms by which SOX proteins act are equally diverse, as are the downstream targets regulated by SOX function. These highlight the importance of SOX proteins in neural development and, in conjunction with *Sox* mutations associated with human neuropathies, demonstrate the necessity for further understanding of their functions.

See also: Neural Crest; Neural Crest Cell Diversification and Specification: Melanocytes; Neural Crest Cell Diversification and Specification: ErbB Role; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation; Neuropathy: Metabolically-Induced.

Further Reading

Bergsland M, Werme M, Malewicz M, et al. (2006) The establishment of neuronal properties is controlled by Sox4 and Sox11. *Genes & Development* 20(24): 3475–3486.

Dong C, Wilhelm D, and Koopman P (2004) Sox genes and cancer. *Cytogenetic Genome Research* 105(2–4): 442–447.

Donner AL, Episkopou V, and Maas RL (2006) Sox2 and Pou2f1 interact to control lens and olfactory placode development. *Developmental Biology* 303: 784–799.

Hong CS and Saint-Jeannet JP (2005) Sox proteins and neural crest development. *Seminars in Cell and Developmental Biology* 16(6): 694–703.

Ito Y, Wiese S, Funk N, et al. (2006) Sox10 regulates ciliary neurotrophic factor gene expression in Schwann cells. *Proceedings of the National Academy of Sciences of the United States of America* 103(20): 7871–7876.

Jankowski MP, Cornuet PK, McIlwrath S, et al. (2006) SRY-box containing gene 11 (Sox11) transcription factor is required for neuron survival and neurite growth. *Neuroscience* 143(2): 501–514.

Kamachi Y, Uchikawa M, and Kondoh H (2000) Pairing SOX off: With partners in the regulation of embryonic development. *Trends in Genetics* 16(4): 182–187.

Kelsh RN (2006) Sorting out Sox10 functions in neural crest development. *BioEssays* 28(8): 788–798.

Kiernan AE, Pelling AL, Leung KK, et al. (2005) Sox2 is required for sensory organ development in the mammalian inner ear. *Nature* 434(7036): 1031–1035.

Kordes U and Hagel C (2006) Expression of SOX9 and SOX10 in central neuroepithelial tumor. *Journal of Neurooncology* 80(2): 151–155.

Okubo T, Pevny LH, and Hogan BL (2006) Sox2 is required for development of taste bud sensory cells. *Genes & Development* 20(19): 2654–2659.

Orkin SH (2005) Chipping away at the embryonic stem cell network. *Cell* 122(6): 828–830.

Pevny L and Placzek M (2005) SOX genes and neural progenitor identity. *Current Opinion in Neurobiology* 15(1): 7–13.

Raible DW (2006) Development of the neural crest: Achieving specificity in regulatory pathways. *Current Opinion in Cell Biology* 18(6): 698–703.

Sasai Y (2001) Regulation of neural determination by evolutionarily conserved signals: Anti-BMP factors and what next? *Current Opinion in Neurobiology* 11(1): 22–26.

Sasai Y (2001) Roles of Sox factors in neural determination: Conserved signaling in evolution? *International Journal of Developmental Biology* 45(1): 321–326.

Stolt CC, Schlierf A, Lommes P, et al. (2006) SoxD proteins influence multiple stages of oligodendrocyte development and modulate SoxE protein function. *Developmental Cell* 11(5): 697–709.

Takemoto T, Uchikawa M, Kamachi Y, et al. (2006) Convergence of Wnt and FGF signals in the genesis of posterior neural plate through activation of the Sox2 enhancer N-1. *Development* 133(2): 297–306.

Taranova OV, Magness ST, Fagan BM, et al. (2006) SOX2 is a dose-dependent regulator of retinal neural progenitor competence. *Genes & Development* 20(9): 1187–1202.

Wegner M and Stolt CC (2006) Sox transcription factors in neural development. In: Thiel PDG (ed.) *Transcription Factors in the Nervous System*, pp. 181–203. Weinheim: Wiley-VCH.

Wegner M and Stolt CC (2005) From stem cells to neurons and glia: A Soxist's view of neural development. *Trends in Neurosciences* 28(11): 583–588.

Spasticity

A D Pandyan, H Hermens, and G R Johnson, Keele University, Keele, UK

Published by Elsevier Ltd.

Introduction

Any lesion in the descending tracts of the central nervous system (CNS), at any level (i.e., cortex, internal capsule, brain stem, or spinal cord), can lead to a combination of sensorimotor symptoms, including spasticity (Table 1). Spasticity, a neurological impairment, is a common feature in a variety of neurological disorders, such as stroke, cerebral palsy, brain injury, spinal cord injury, and multiple sclerosis. It is believed that spasticity contributes to contracture development, exacerbates pain, and hinders recovery of functional movement. Therefore, substantial effort has been invested in managing this clinical condition and in developing an understanding of it. However, progress has been relatively modest in both of these areas. It would appear that the main barriers to progress are imprecise definitions (of spasticity) and a paucity of valid and reliable, but clinically usable, outcome measures. Our current understanding of spasticity is based in part on existing, although imperfect, animal models. Further progress is dependent on the development of valid definitions and measures, and adequate neurophysiological models.

Pathophysiology

The positive features that result from CNS lesions affect tracts in both the pyramidal and parapyramidal systems. (Note: We specifically use the term 'parapyramidal,' as suggested by both Sheean and Burke, to differentiate between (1) tracts that run in close proximity to the corticospinal tract and arise at the cortex and (2) those that arise from the basal ganglia – that is, the extrapyramidal fibers, which are responsible for rigidity.) An isolated lesion of the corticospinal tract will normally result in loss in fine control of movement and, rarely, specificity. The positive symptoms arise as a result of plastic changes in response to the lesion (e.g., collateral sprouting) and/or release effects (disinhibition) as a result of the lesion. Irrespective of the cause of the lesion, the positive features primarily occur as a result of disinhibition at the level of the spinal cord, and the following pathways have been implicated (Figure 1):

- Increased excitation in the group Ia afferents from the muscle spindles.
- Increased excitation and reduced inhibition in the group II afferents from the muscle spindles.
- Reduced autogenic inhibition from the Ib afferent from the Golgi tendon organ.
- Reduced recurrent inhibition from the Renshaw cells and motor axon collaterals.
- Reduced reciprocal Ia inhibition from the muscle spindles from the antagonistic muscle groups.
- Reduced presynaptic inhibition of the Ia terminals.
- Increased facilitation of the flexor reflex afferents involving group II, group III, and group IV cutaneous afferents.

The other contributory factors are as follows:

- Increase in the fusimotor drive (i.e., γ-motor neurons innervating the muscle spindle); however, the limited evidence available suggests that this may not happen.
- Changes in the cell membrane properties of the motor neurons and possibly the interneurons. There is growing body of evidence demonstrating that the cell membranes are not the passive structures that they were once believed to be. In particular, prolonged depolarization (plateau potentials) resulting from localized persistent inward currents produced by voltage-dependent Ca^{2+} and Na^{+} channels may be responsible for some of the abnormal patterns of muscle activation. The effects of these inward currents are chemically modulated by serotonergic and/or noradrenergic motor neurons. (Plateau behavior has not been fully demonstrated in humans. In cat and rat models, it is primarily seen in chronic spinal preparations.)
- Reduction in the postactivation depression within the Ia terminals.
- Some patients will demonstrate a continuous drive to the α-motor neuron (this has been described by some as 'spastic dystonia' and by others as 'rigidity'; of the two, we would prefer the former, as the feeling of cogwheel-type resistance is not encountered in these patients during an externally imposed movement). This may result from a loss of corticoreticular inhibition of the pathways that descend from the brain stem and/or a reduction in the spinal reflex thresholds. From the available evidence it is not possible to delineate the exact pathophysiological mechanisms that result in the tonic (and continuous) activation of muscles.

In summary, abnormal muscle activation, associated with the positive features seen following a CNS lesion, can result from a combination of reduced thresholds and/or an increase in gain in both the descending pathways and reflex pathways that converge onto the α-motor neurons.

Table 1 Sensorimotor clinical features that present following a CNS lesion

Positive features	*Negative features*
Increased reflexes	Weakness
Spasticity	Fatigueability
Altered tone	Loss of dexterity (motor control)
Spasm (flexor and/or extensor)	
Clonus	
Abnormal movement patterns during attempted voluntary movement (e.g., associated reactions and dyssynergic co-contractions)	

Originally classified, by Hughlings Jackson, as (1) 'positive features' – excess/exaggerated activity (when compared to normal) which could occur spontaneously or in response to external stimuli, and (2) 'negative features' – abnormal release of function or a deficit in performance when compared to normal.

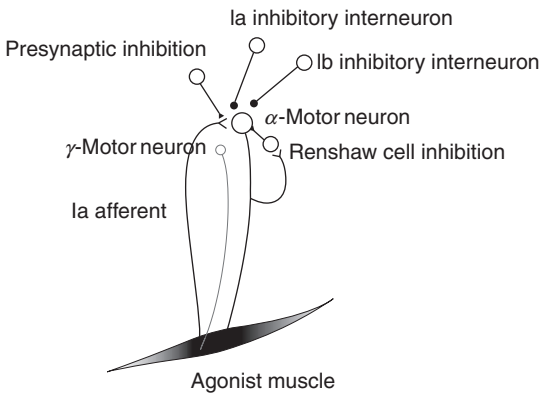

Figure 1 A schematic illustrating the potential spinal pathways implicated in spasticity. The group II afferents from the spindle and the group III/IV cutaneous afferents are not illustrated. From Nielsen JB, Crone C, and Hultborn H (2007) The spinal pathophysiology of spasticity – from a basic science point of view. *Acta Physiology* 189: 171–180.

Definitions

The term 'spasticity' has been about for a long time, and the first formal definition of sorts was that proposed by Denny-Brown in 1960. According to this definition, spasticity, following an upper motor neuron lesion, "could be identified by the presence of a soft yielding resistance that appears only towards the end of a passive stretch, and is associated with increased amplitude stretch reflex."

Subsequently, Lance, at a conference held in 1980, initially suggested that spasticity be defined as "a motor disorder characterized by a velocity dependent increase in tonic stretch reflexes and increased tendon jerks resulting from disinhibition of the stretch reflex, as one component of an upper motor neurone lesion." This definition specifically associates spasticity with the disinhibition of the stretch reflex pathway (see the sections titled 'Biomechanical methods' and 'Electrophysiological methods').

At a later stage in the conference, this original Lance definition was modified to "a motor disorder characterized by a velocity dependent increase in the tonic stretch reflex (muscle tone) with exaggerated tendon jerks, resulting from hyperexcitability of the stretch reflex, as one component of the upper motor neurone syndrome." In this revised definition (which will be described as the 'modified Lance definition'), one subtle and important assumption has been made – tonic stretch reflex activity has been equated to muscle tone. (There are two definitions of muscle tone. The first would appear to be based on Bernstien's 'innervational condition' and is defined as a state of readiness to act. The second equates muscle tone to the resistance to externally imposed movement.)

More recently, a fourth definition has emerged: the Task Force on Childhood Motor Disorders has recommended that spasticity be treated as a form of hypertonia (the other two forms being dystonia and rigidity), defined as "a velocity dependent increase in hypertonia with a catch when a threshold is exceeded." Of these four, fairly precise, definitions found in the literature, the modified Lance definition is perceived to be the most cited. However, this is by no means the most dominant definition, as only a third of the literature cites the modified Lance definition and many cite it inappropriately. Another third equates spasticity with hypertonia or increased muscle tone, defined as an increase in the resistance to passive movement. The remainder equate spasticity with one or many of the positive features associated with the CNS lesion (Table 1). In clinical practice, the term 'spasticity' is routinely used to describe one or many of the positive features of the upper motor neuron syndrome.

Resolving the Inconsistency

Data from current clinical trials, animal models, and clinical practice almost unanimously suggest that the clinical features associated with spasticity result from either intermittent or sustained involuntary (reflex) muscle activity. The original Lance definition associates one particular form of this abnormal activity (i.e., velocity-dependent increase in the tonic reflexes) with spasticity. Many others would appear to associate all forms of abnormal activity (e.g., spasm, clonus) with spasticity.

There are two other key assumptions made with respect to the Denny-Brown, modified Lance, and Task Force of Childhood Disorders definitions of spasticity: (1) the abnormal muscle activation will always contribute to a measurable increase in

resistance to passive movement and (2) the velocity-dependent increase in resistance to passive movement is an exclusive feature of spasticity. With respect to the former, from first principles, muscle activation will contribute to force production within the muscle and therefore will contribute to an increase in resistance to passive movement. The modeling of this relationship is, however, more complex, as the activation force relationship is being studied in an active muscle that is being lengthened. The more important question, which is often ignored, is related to validity of measurement. Most patients with a CNS lesion also have coexisting conditions that result in changes in the mechanical properties of soft tissue structures and joint structures. Therefore, using resistance to externally imposed passive movement to study spasticity is not valid, as this is a confounded measure. Resistance to passive movement and stiffness can be reliably measured; however, to then infer that one can reliably measure spasticity from these measurements is not correct (Figure 2).

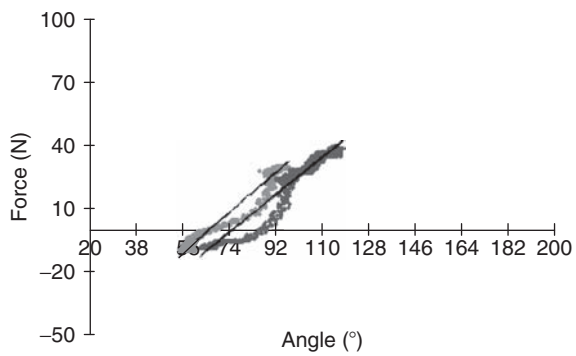

Figure 2 The stiffness and muscle electrical activity (not illustrated) were measured before (light gray) and 4 weeks (black/dark gray) after treatment with botulinum toxin (type A) to the elbow flexors (biceps and brachioradialis). The muscle activity was fully suppressed after treatment, but the stiffness showed almost no change (1.063 *N*/degree before treatment and 1.001 *N*/degree 4 weeks after treatment).

With respect to the latter, soft tissues (i.e., the muscle tendon complex) normally demonstrate viscoelastic behavior. Therefore, to exclusively attribute velocity-dependent increase in the resistance to externally imposed passive movement to a clinical feature of spasticity is wrong (Figure 3). Further, using a simple second-order system (as many of these definitions would suggest) may not be sufficient. Instead, it is likely that a second-order system with an added time delay (to account for the reflex delays) should be considered when measuring spasticity.

If there is to be substantial progress in our understanding of spasticity, then there is need to find a way of dealing with these inconsistencies in definition. There are three positions to take in identifying a way forward: (1) abandon the term spasticity, (2) standardize to the original Lance definition, and (3) redefine spasticity in a way that reflects current usage – that is, equates spasticity with many of the positive features identified with the upper motor neuron syndrome (increased reflexes, 'abnormal' muscle activity response to an externally imposed passive movement, altered tone of any sort, clonus, and/or spasms). This leads to spasticity being defined as "disordered sensorimotor control, resulting from an upper motor neuron lesion, presenting as intermittent or sustained involuntary activation of muscles." Based on current understanding, this definition may exclude abnormal movement patterns triggered during voluntary movement, and will exclude all the negative features associated with the upper motor

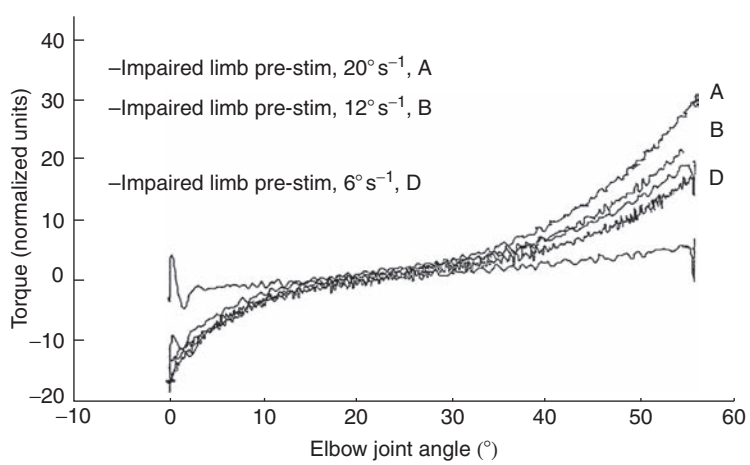

Figure 3 An illustration of stiffness about the elbow flexors at different velocities. The stiffness increases as velocity increases (curves: D, 6 °s^{-1}; B, 12 °s^{-1}; A, 20 °s^{-1}). Muscle activity was only observed at velocities of 20 °s^{-1}. From Dewald JPA and Given JD (1994) Electrical stimulation and spasticity reduction. Fact or fiction? *Physical Medicine & Rehabilitation*: *State of the Art Reviews* 8: 507–522.

neuron syndrome. (Note: The phenomenon of associated reactions – the overflow-type activity – can also be observed in neurologically intact individuals and in people with reduced muscular reserve capacity, when attempting tasks involving maximal voluntary contractions. Whether to equate these associated reactions and dyssynergic movement with spasticity needs to be debated.)

The first of the three suggested positions, although a preferred one, is unrealistic due to the pervasiveness of the term 'spasticity.' Of the two remaining options, redefinition may be the way forward, as the second has not led to any convergence over the past 25 years. However, if the third option is to be adopted then there is an obligation on both clinicians and researchers to be more precise in the description and measurement of the clinical features associated with what each individual identifies as 'spasticity'.

Measurement

> When you can measure ... you know something about it – but when you cannot measure ... your knowledge is of a meagre and unsatisfactory kind it may be the beginning of knowledge but you have scarcely, in your thought, advanced to the stage of science whatever the matter may be. (Lord Kelvin)

According to the current International Classification of Functioning (ICF) definition, spasticity (as newly defined) can be classified as an impairment. It is possible that this impairment may have an impact on activity (e.g., walking, feeding) and therefore on participation (e.g., quality of life, independence, career burden). From a patient perspective, the measurement of functional activity (e.g., gait), independence, and quality of life may be more important. However, these methods of surrogate measurement methods will not be reviewed further here, as the causal association between spasticity, functional limitations, and independence is not yet fully demonstrated.

The measures currently available for quantifying or classifying spasticity can be broadly classified into three categories: clinical scales, biomechanical methods, and electrophysiological methods.

Clinical Scales

Clinical scales are commonly used, for both research and clinical purposes, to either subjectively quantify or classify spasticity. Most of these scales measure one construct of this multifaceted impairment and can be further classified as follows:

1. Scales that subjectively quantify the resistance to externally imposed passive movement (e.g., modified Ashworth scale – Table 2, Tone Assessment Scale). A key, but as yet unvalidated, assumption made when using these scales is that the resistance encountered during the externally imposed passive movement results from a reflex response to the stretch. The reliability of these scales is variable – that is, it depends on the patient population, the limb segment under test, the inertia of the limb segment, the velocity of testing, and the testing position. A variety of other factors, such as temperature, previous history of activity, and time of day, may also confound. Commonly used scales can, at best, provide an ordinal level of measurement. It is not possible to recommend any of these scales as valid and clinically relevant measures of spasticity, in their respective current formats, as resistance to passive

Table 2 Properties and limitations of the Ashworth and modified Ashworth scales

Score	*Ashworth scale*	*Modified Ashworth scale*
0	No increase in tone	No increase in muscle tone
1	Slight increase in tone, giving a catch when the limb is moved in flexion or extension	Slight increase in muscle tone, manifested by a catch and release or by minimal resistance at the end of the range of motion when the affected part(s) is moved in flexion or extension
1+		Slight increase in muscle tone, manifested by a catch, followed by minimal resistance throughout the remainder (less than half) of the range of movement
2	More marked increase in tone but limb easily flexed	More marked increase in muscle tone through most of the range of movement, but affected part(s) easily moved
3	Considerable increase in tone – passive movement difficult	Considerable increase in muscle tone passive, movement difficult
4	Limb rigid in flexion or extension	Affected part(s) rigid in flexion or extension

An assessor is required to move the limb segment through full range of movement in 1 s and then subjectively grade the resistance encountered using the respective descriptors. Data from Ashworth B (1964) Preliminary trial of carisoprodal in multiple sclerosis. *Practitioner* 192: 540–542; Bohannon RW and Smith MB (1987) Interrater reliability of a modified Ashworth Scale of muscle spasticity. *Physical Therapy* 67: 206–207. Reproduced from Pandyan AD, Johnson GR, Price CIM, et al. (1999) A review of the properties and limitations of the Ashworth and modified Ashworth Scales. *Clinical Rehabilitation* 13: 373–383.

movement will be confounded by the mechanical properties of the soft tissue structures (muscles and tendons) and joints and the ability of the person to relax voluntarily.

2. Scales that classify a combination of the amplitude of the reflex response and the type of reflex response (e.g., National Institute of Neurological Disorders and Stroke myotatic reflex scale – Table 3), frequency of spasms (e.g., Penn spasm frequency scale, spasm frequency scale – Table 4), and clonus score. These scores have validity and provide an ordinal level of measurement of either spasms or clonus frequency. The diagnostic importance, sensitivity, reliability, responsiveness, and clinical usefulness of these scores have not been studied in great detail, especially in the context of their usefulness as an outcome measure that can inform evidence-based practice. Some of the scores may have potential as either nominal level classification measures or ordinal level measures. However, more work on validity and sensitivity is required. Attention will need to be paid to the relevance of the deep tendon reflexes (also termed 'stretch reflex' or 'tendon jerk reflex'), as evidence seems to suggest that the variability of the response in a population with spasticity overlaps with that of a normal population.

3. The final dichotomous measure that has been used in the diagnosis of people with an upper motor neuron syndrome and/or spasticity is the 'extensor toe sign.' The Babinski and the Gonda–Allen methods (Table 5) are the most reliable methods of eliciting this response. These methods have good predictive validity with respect to the presence of an upper motor neuron lesion and may provide a marker for one aspect of spasticity.

There are two other scales that attempt to measure multiple constructs associated with spasticity. The first of these produces a composite score of spasticity by multiplying the Ashworth score with the spasm frequency score. Until we have a better understanding of the component parts of this composite score, it is not advisable to use such methods of measurements. The second is the Tardieu method of measurement, which is often, inappropriately, reported as the 'Tardieu scale.' In this method of measurement the assessors are required to passively stretch a limb using two standardized velocities, one slow (which is done first) and one fast (which is done second) (Table 6). The assessor then classifies the reaction to the fast movement and quantifies the angle of the reaction in the fast movement with reference to the slow movement.

Table 3 Descriptor for the National Institute of Neurological Disorders and Stroke myotatic reflex scale

Score	*Descriptor*
0	Reflex absent
1	Reflex small, less than normal; includes a trace response, or a response brought out only by reinforcement
2	Reflex in lower half of normal range
3	Reflex in upper half of normal range
4	Reflex enhanced, more than normal; includes clonus if present, which optionally can be noted in an added verbal description of the reflex

Data from Hallet M (1993) NINDS myotatic reflex scale. *Neurology* 43: 2723.

Table 4 Descriptors associated with the ordinal level scales used to quantify the frequency of spasms

Score	*Penn spasm frequency scale*	*Spasm frequency score*
0	No spasms	No spasms
1	Mild spasms at stimulation	One or fewer spasms per day
2	Irregular strong spasms less than one time/hour	Between one and five spasms per day
3	Spasms more often than one time/hour	Five to less than 10 spasms per day
4	Spasms more than 10 times/hour	Ten or more spasms per day, or continuous contraction

Data from Penn RD, Savoy S, Corcos D, et al. (1989) Intrathecal baclofen for severe spinal spasticity. *New England Journal of Medicine* 320: 1517–1521; Snow BJ, Tsui JK, Bhatt MH, et al. (1990) Treatment of spasticity with botulinum toxin: A double blind study. *Annals of Neurology* 28: 512–515; Biering-Sørensen F, Nielsen JB, Klinge K, et al. (2005) Spasticity-assessment: A review. *Spinal Cord* 44: 708–722.

Table 5 Two possible methods for testing for the extensor toe sign

Stimulation methods considered to produce the reflex most reliably	*Reflex response*
Babinski: Stimulation of the plantar surface of the foot with a blunt point from the heel forward, crossing the metatarsal pad from the small to the great toe	The normal response is a flexion (curling down) of all toes ('downgoing toe')
Gonda–Allen: Downward snapping of the distal phalanx of the second or fourth toe	In people who have an upper motor neuron lesion there is an extension of the great toe ('upgoing toe') with fanning of the other toes (abnormal)

Reliable for identifying a corticospinal tract lesion.

Table 6 The Tardieu method of assessment

Muscle reaction (X)[a]	*Description*
0	No resistance throughout the course of the passive movement
1	Slight resistance throughout the course of the passive movement, with no clear catch at precise angle
2	Clear catch at precise angle, interrupting the passive movement, followed by release
3	Fatigable clonus (<10 s when maintaining pressure) occurring at precise angle
4	Infatigable clonus (>10 s when maintaining pressure) occurring at precise angle

[a] The velocities used when assessing with the Tardieu method are *V*1, as slow as possible so as not to trigger a stretch reflex; *V*2 the limb segment is allowed to fall under the influence of gravity; and *V*3, as fast as possible (faster than the rate of the natural drop of the limb segment under gravity). The assessor should use *V*1 (to assess range of movement) and either *V*2 or *V*3 (to assess spasticity). The reaction to stretch for each muscle group is rated at a specific stretch velocity with parameters *X* (muscle reaction) and *Y* (angle of muscle reaction). Scores *X*, from 0 to 4, are based on quality of muscle reaction when assessing spasticity using either *V*2 or *V*3. The angle of muscle reaction *Y* is measured relative to the position of minimal stretch of the muscle (corresponding to angle), where it is relative to the resting anatomic position. (The muscle reaction is expected to occur earlier in the range of movement with greater spasticity.)
Often, inappropriately, called the Tardieu Scale as currently described. This method of measurement was originally developed by Held and Pierrot-Deseilligny and was based on a method of assessment developed by Tardieu and colleagues. Data from Held JP and Pierrot-Deseilligny E (1969) *Reeducation Mortrice des Affections Neurologiques*. Paris: J.B. Baillière et fils; Gracies J-M (2001) Évaluation de la spasticité. Apport de l'échelle de tardieu. *Motricité Cérébrale* 22: 1–16; Morris S (2002) Ashworth and Tardieu scales: Their clinical relevance for measuring spasticity in adult and paediatric neurological populations. *Physical Therapy Reviews* 7: 53–62. Adapted from Haugh AB, Pandyan AD, and Johnson GR (2006) A systematic review of the Tardieu Scale for the measurement of spasticity. *Disability and Rehabilitation* 28: 899–907.

While this scale is growing in popularity within both clinical practice and research studies, there is a need to urge caution in use, as the validity, reliability, sensitivity, and responsiveness of this scale have not yet been studied in a systematic way.

In summary, clinical scales are very popular, as they are easy to use and cheap. However, most scales cannot be recommended for routine use, as validity remains unproved. The clinical scales that quantify resistance to passive movement are confounded measures and depend on characteristics which are poorly defined (e.g., catch, catch when a threshold is exceeded, and increased resistance end range of movement). The dichotomous scale for the extensor toe sign and the scales that quantify spasms and clonus have the potential to be of use for both clinical practice and research, but more fundamental work on the measures and descriptors is required.

Biomechanical Methods

The biomechanical approaches indirectly measure spasticity by quantifying the resistance to the externally imposed passive movement. As pointed out with clinical scales, until we are able to provide evidence of validity and sensitivity we will need to use extreme caution when using these methods for quantifying spasticity.

Biomechanical methods of measurement are underpinned by four basic principles. The first involves the use of a velocity-controlled external perturbation, during which the relationship between the imposed moment and angular displacement is quantified using standard curve fitting techniques. (The best fit lines used to estimate stiffness in Figure 2 uses the least-square estimate of the slope of the force angle curve.) The second involves the use of a constant moment external perturbation, during which the relationship between the velocity and angular displacement is studied (Figure 4). The third involves studying the damping of displacement when a relaxed limb is dropped against gravity (Wartenberg Pendulum Test and its variations – Figure 5). The fourth involves quantifying the kinematic and kinetic response of the stretch reflex to a controlled tap or vibratory stimulus.

With respect to the first two methods, the external perturbation can be applied using either a manual or motorized system. It is not entirely clear as to whether motorized methods are superior to manual methods of perturbation, and more comparative work is required to explore the relative merits of these approaches in terms of sensitivity and accuracy of measurement. (Note: The muscle spindle primarily responds to a linear stretch. However, it is not possible to ensure that the linear velocity of stretch presented to the muscle spindle ensemble is controlled when the angular rotation is controlled. The variations in linear velocity occur due to variations in the center of rotation, the radius of rotation, and the dimension of the tendon and ligamentous structures. Furthermore, the muscle–tendon unit dimension will vary as a function of time (postinjury) and the level of functional capabilities. This would imply that, even for the same individual, the stimulus to the muscle spindle ensemble will vary indeterministically in longitudinal studies.) The little evidence that is available to date would suggest that manually imposed external perturbations provide similar levels of measurement when compared to

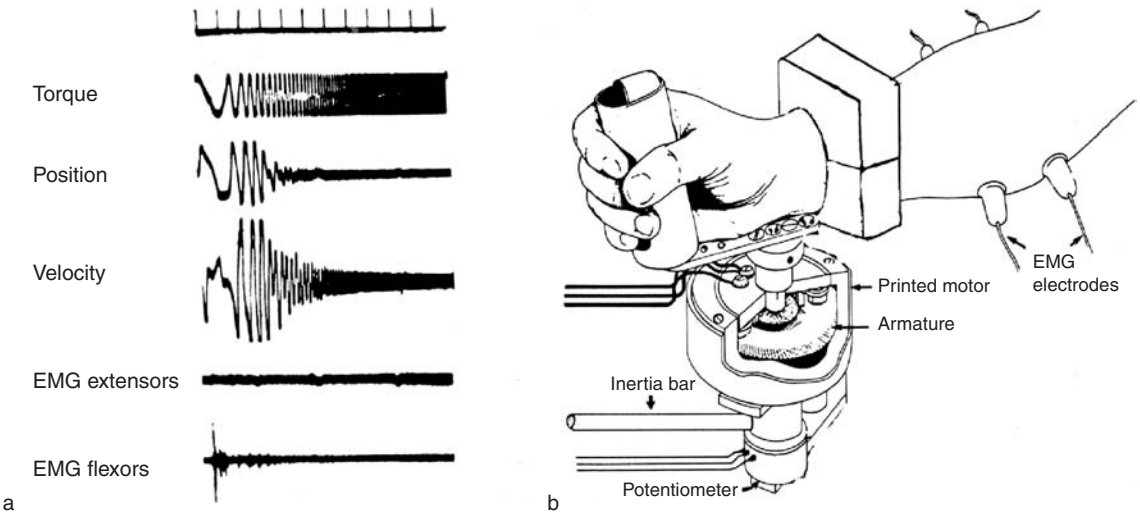

Figure 4 The output (a) from a measurement method (b) in which the externally imposed torque (moment) was standardized, and the resulting variations in both displacement and velocity were estimated (EMG, electromyelography). The resonant frequency method was developed by Walsh as a method of quantifying stiffness. From Walsh EG (1992) *Muscles, Masses and Motion. The Physiology of Normality, Hypotonicity, Spasticity and Rigidity*. London: Mac Keith Press. From Bajd T and Vodovnik L (1984) Pendulum testing of spasticity. *Journal of Biomedical Engineering* 6(1): 9–16.

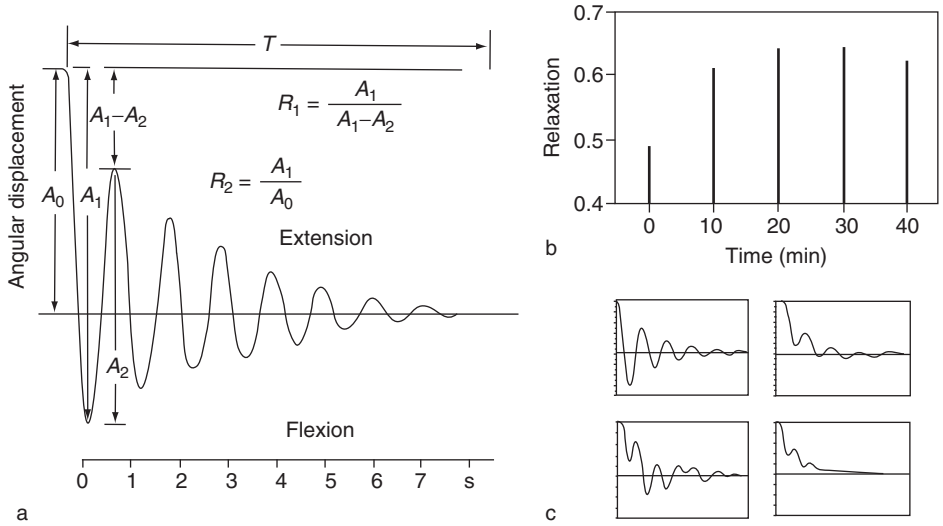

Figure 5 (a) Quantifying the response from the pendulum test when applied to the knee. There are two approaches to calculating the relaxation index (R_1 and R_2). (b) The (normalized) relaxation index is seen to vary with time. (c) Calculating the relaxation index is not straightforward, as the response to the pendulum test can be variable.

velocity-controlled displacement methods. The one advantage of using a motorized control system for perturbations is the ease at which random/pseudorandom perturbations can be applied, and these may help in the development of innovative methods of exploring the correlation between the neurological and biomechanical responses – a first step in studying validity of the indirect methods. More work is required to explore the reliability of manual techniques, as these have the potential to lead to the development of clinically relevant measures. While manual techniques are easy to apply for the distal joints of the upper and lower limb, they have limited value in taking measurements about the shoulder and hip. Furthermore, the mass of the lower limb segments and the inherent stiffness of the tendo achilles make manual perturbations in the lower limb more difficult when compared to the upper limb.

The gravitational method was first described by Wartenberg in the 1950s and was instrumented by Bajd and Vodovnik in the 1980s. In this test the examiner positions the individual such that the limb segment is hanging freely over an examination couch. The examiner then lifts the limb segment against gravity and lets it drop (the stimulus is the influence of gravity). The resulting damping is then quantified using relaxation indices (Figure 5). The damping will be influenced by two factors – the mechanical properties of the soft tissues and the neural activation of the muscle groups being stretched. The validity of this method has been tested only against unvalidated scales and therefore remains unexplored. The reports on reliability suggest a degree of variability, and the relaxation indices are known to vary with time, position of testing, time of testing, and a myriad of other conditions (e.g., ability to relax, temperature), emphasizing that a detailed protocol of the measurement is required. The primary assumption that the damping is primarily influenced by the quadriceps and hamstring group reflex activity needs to be further explored (Figure 6).

The instrumentation of the deep tendon reflex is also relatively easy, and displacement, velocity of displacement, and force of muscle contraction are all potential outcome measures. However, given the variability in response, more work needs to be done prior to adopting these outcome measures in routine use.

The criticism that biomechanical measures may be more cumbersome when compared to clinical scale-based methods may not be well founded. Some of these methods of measurement are relatively easy and inexpensive to carry out, especially the methods based on manual perturbations. The time investment within the clinical context is not significant (single joint measurements can be taken in less than 5 min). However, there is a need for more research to substantiate validity before these measures become clinically relevant.

Electrophysiological Methods

The only direct approach to measure spasticity is by studying the electromyographic (EMG) response (using surface or fine wire electrodes) of the neuromuscular system to an external stimulus or by monitoring EMG activity over a continuous 24 h period. While there are some advantages to using fine wire techniques to study motor unit properties and responses, there is a need to reflect on their clinical relevance. The discomfort associated with these recordings reduces their clinical potential significantly; more importantly, it is possible that the discomfort associated with these methods confounds the response measures obtained.

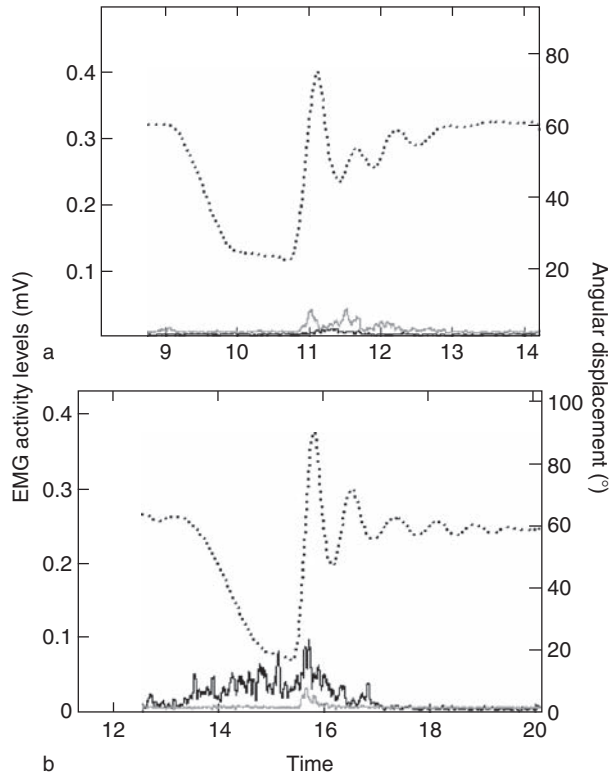

Figure 6 An illustration of the inconsistent relationship between the biomechanical measures of spasticity and the complex nature of the muscle response during the pendulum test. The muscle activation patterns observed for a child before treatment (a) and 4 weeks after treatment (b) for spasticity involving selective dorsal rhizotomy. The dotted lines represent angular displacement, the gray lines are extensor electromyelogram (EMG; rectus femoris) responses, and the black lines are the flexor EMG (hamstring) responses. The patterns of muscle activation pattern have changed: the hamstrings, which were silent during the preparation and hold phase prior to surgery, were active after treatment. The relaxation index (R_2) had increased and damping had reduced, showing a reduction in spasticity.

Whether one should take measurements from relaxed or active muscles is a moot point. From a clinical perspective, most people who are classified as having troublesome spasticity are often incapable of producing a controlled voluntary contraction, so that, by necessity, one will need to examine a relaxed muscle. Many patients with active, independent movement will probably not need to be treated for troublesome spasticity! So, on balance, there is a need to develop a combination of measurement protocols that can be used in a patient population with varying functional capabilities.

A variety of external perturbations can be used for studying the response of the neuromuscular system to an external perturbation. The perturbation signals can involve the stretching of a muscle (manual or motorized), a tendon tap (T-reflex), or an electrical stimulation (H-reflex and F-waves) (Figure 7). The

response can be quantified by using, amplitude (of EMG related to stimulus), time delay/response times, durations, and thresholds (related to either response or stimulus). If continuous activity is being monitored, the outcome measures could include a count/frequency of abnormal activity bursts (e.g., algorithms to identify clonus and spasms), duration of abnormal activation, characterization of activation patterns, and timing of these events.

There is a misconception that electrophysiological methods of measurement are complex. While this is true for some of the protocols involving fine wire recordings, most electrophysiological methods can be applied for routine clinical use (e.g., one effective method to select patients for intrathecal baclofen pump therapy is to identify the response of the H-reflex – which is normally abolished if a person responds to baclofen – to a bolus of intrathecal baclofen prior to surgically implanting the pump; further, this method can also be used to monitor effectiveness of the therapy over time). However there are some simple pitfalls that need to be avoided when using electrophysiological recordings:

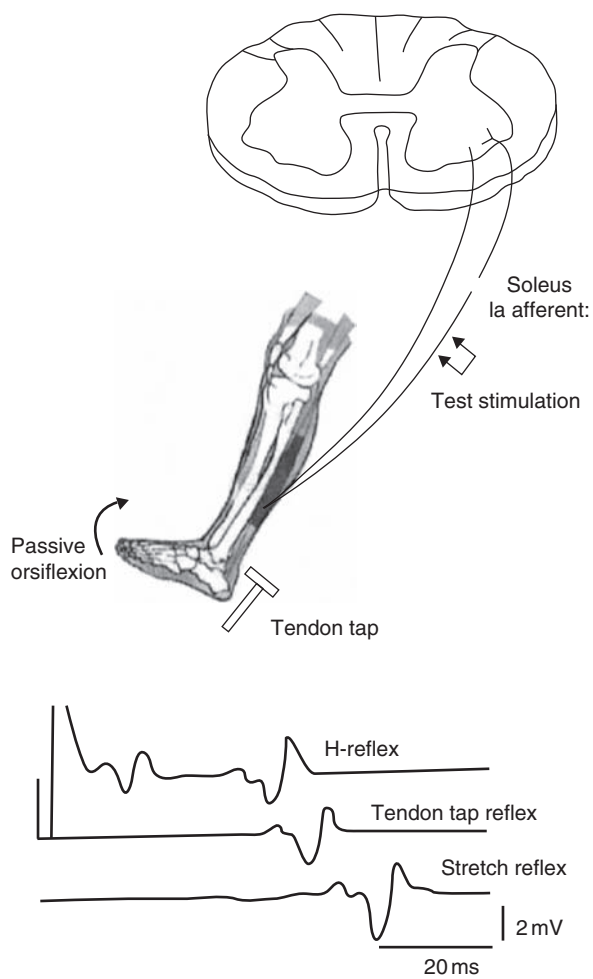

Figure 7 The muscle response to electrical stimulation, a tendon tap, and an externally imposed stretch. Note the relative delays in onset times. This results from the variations in the pathways stimulated by the signals. From Biering-Sørensen F, Nielsen JB, and Klinge K (2005) Spasticity-assessment: A review. *Spinal Cord* 44: 708–722.

- Using amplitude of the EMG response, in the absence of standardization, requires a very strict protocol to be used. Interindividual and intraindividual variabilities are due to the variabilities associated with the confounding factors, including skin impedance, muscle geometry, the placement of the electrodes with respect to motor endplate zones and tendons, interelectrode distance, possible muscle morphology changes, and joint position (i.e., resting length of the muscle). In addition, it has been shown that factors such as posture, joint position, and even head position and orientation do affect the EMG responses elicited by mechanical or electrical stimulation.
- It may be possible to reduce the variability of the amplitude response. Current methods involve referencing the amplitude of the EMG response to the amplitude of an EMG response to a maximal motor stimulus or EMG amplitude corresponding to a maximal voluntary contraction. However, using methods of standardization involving maximal voluntary contraction or a percentage of maximal voluntary contraction will be of limited use in patients who have weakness or paralysis.
- When using latencies (or onset delays), a bulk of the literature focuses on first deflection from baseline, and much of this identification is done manually. The development of reliable burst detection algorithms is essential to reduce impact of human error associated with visual identification.
- When measuring latencies, one should work within the constraints of the discrimination set by the sampling frequency (e.g., to measure latency to a sensitivity of 1 ms it is essential to measure at a sampling frequency of 2 kHz or greater – ideally, 4 kHz).
- Caution is urged when applying smoothing EMG data methods, as this can introduce variability and affect latency estimations (Figure 8).
- With some modern amplifiers, skin preparation is no longer critical (i.e., skin abrasion is no longer required). However, the site where the electrode is placed must be cleaned well.

Treatment

The management of spasticity is often multifaceted and can be broadly classified into pharmacological methods, surgical methods, and rehabilitation therapy-based methods. These are very briefly described in Table 7. The evidence base associated with each treatment method remains limited, as most sub-

stantial research has been underpinned by measures of limited validity and poor design. Most Cochrane reviews have also reported inconclusive findings. There is a need to reexamine, and in certain circumstances repeat, the studies exploring efficacy of individual treatments using valid measures of spasticity and patient-centered outcome measures.

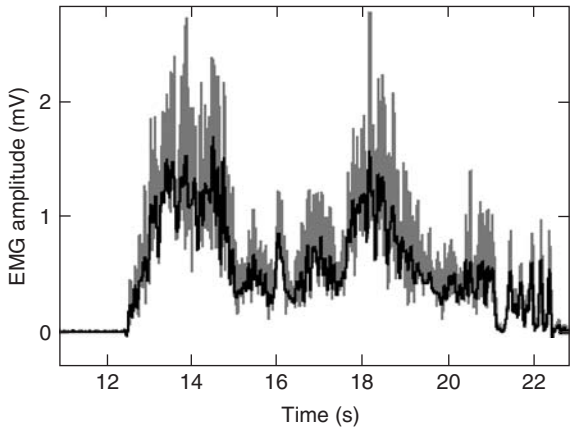

Figure 8 An illustration of differences when electromyelographic (EMG) signals are smoothed using two different techniques. The gray curve was obtained after notch filtering at 50 Hz and then smoothing using a root-mean-square procedure. The black curve was obtained after notch filtering, then rectifying and low-pass filtering at 10 Hz.

Future Work

A key finding from a century of research investment into the phenomenon of spasticity is the huge variability associated with the phenomenon. For example, the reflexes in people with spasticity are often documented to fall within the ranges of spasticity found in people with no known pathologies. It is possible that such variability results from the pathophysiological basis of this phenomenon being variable and the measures lacking specificity.

If we are to find a way forward, the first stage is to address this variability, which, in our view, has resulted from inadequate definitions. We have therefore suggested that spasticity be defined as "disordered sensorimotor control, resulting from an upper motor neuron lesion, presenting as intermittent or sustained involuntary activation of muscles." Clinicians and researchers will need to identify the specific features under treatment or investigation and use valid outcome measures.

Table 7 Currently available methods to treat spasticity

Method	*Description*
Pharmacological	
Baclofen	Acts pre- and postsynaptically at the level of the spinal cord; there are two methods of administration, oral and intrathecal
Benzodiazepines (diazepam)	Acts postsynaptically at the level of the spinal cord; oral administration
Dantrolene sodium	A muscle relaxant
Clonidine/tizanidine	Acts centrally on α2-adrenergic receptors; it is believed that this enhances central inhibition via the cerulospinal pathways
Cannabinoid	Mechanism associated with the reduction of spasticity not fully elucidated
Phenol	Neurolysis of the lower motor neuron
Botulinum toxin (types A and B)	Blocks the release of acetylcholine at the motor neuron junction with the muscle
Surgical	
Selective dorsal rhizotomy	Usually carried out at the level of the lumbo-sacral plexus
Peripheral neurectomy	This is rarely carried out as there are alternatives such as a treatment with botulinum toxin
Intrathecal baclofen therapy	Acts pre- and postsynaptically at the level of the spinal cord; there are two methods of administration, oral and intrathecal
Rehabilitation therapy	
Electrical stimulation	Electrical stimulation can be applied to a variety of sites in the central nervous system (deep brain stimulation, cortical stimulation, cerebellar stimulation) and the peripheral nervous system (muscle, antagonistic to those affected with spasticity, or muscles with spasticity); mechanism of action has not been elucidated
Orthotic devices	Traditional splints that hold muscles with spasticity in a stretched or neutral position, lycra splints, dynamic splints, or static progressive orthoses and serial casting; mechanism of action has not been elucidated
Rehabilitation therapies	There are many methods of treatment described under this category (e.g., Bobat method, strength training, peripheral neuromuscular facilitation, positioning and seating system), but the content and impact of these therapies are not clearly described; mechanism of action has not been elucidated

Some of the treatment methods are better described than others, but there is a paucity of evidence-based guidelines.

Evidence from some carefully controlled studies has shown that altered motor control, including the loss of reflex modulation and weakness, may be the key problem following CNS lesion. Despite this, the focus on spasticity has been driven by the assumption that persistent spasticity is responsible for contractures, pain, and loss of functional independence. While there are theoretical possibilities that this may be the case, experimental evidence (especially from robust longitudinal studies) is needed to substantiate these hypotheses. Such evidence will depend on valid measures which can be used in a clinically relevant population of patients – and these measures need to be developed. A substantial part of our understanding has developed from animal models, and it is likely that this work will still be required for some time to come. One would hope that our understanding will progress at a slightly faster pace than that in the past century. However, when collecting new evidence, it is important not to lose sight of the possibility that spasticity may be an 'epiphenomenon'!

See also: Axonal Injury in Demyelinating Disease and CNS Injury; Brain Trauma; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Hereditary Spastic Paraplegia; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation; Neuromuscular Junction (NMJ): Presynaptic Stretch Effects on Neuromuscular Transmission; Presynaptic Events in Neuromuscular Transmission; Spinal Cord Injuries; Stroke; Stroke: Injury Mechanisms.

Further Reading

Ashworth B (1964) Preliminary trial of carisoprodal in multiple sclerosis. *Practitioner* 192: 540–542.

Bajd T and Vodovnik L (1984) Pendulum testing of spasticity. *Journal of Biomedical Engineering* 6(1): 9–16.

Barnes MP and Johnson GR (eds.) (2001) *Upper Motor Neurone Syndrome and Spasticity: Clinical Management and Neurophysiology.* Cambridge, UK: Cambridge University Press.

Biering-Sørensen F, Nielsen JB, and Klinge K (2005) Spasticity-assessment: A review. *Spinal Cord* 44: 708–722.

Bohannon RW and Smith MB (1987) Interrater reliability of a modified Ashworth Scale of muscle spasticity. *Physical Therapy* 67: 206–207.

Brown JK, van Rensburg F, Walsh G, Lakie M, and Wright GW (1987) A neurological study of hand function of hemiplegic children. *Developmental Medicine and Clinical Neurology* 29: 287–304.

Denny-Brown D (1966) *The Cerebral Control of Movement*. Liverpool, UK: Liverpool University Press.

Dewald JPA and Given JD (1994) Electrical stimulation and spasticity reduction. Fact or fiction? *Physical Medicine & Rehabilitation*: *State of the Art Reviews* 8: 507–522.

Gracies J-M (2001) Évaluation de la spasticité. Apport de l'échelle de tardieu. *Motricité Cérébrale* 22: 1–16.

Hallet M (1993) NINDS myotatic reflex scale. *Neurology* 43: 2723.

Haugh AB, Pandyan AD, Johnson GR, et al. (2006) A systematic review of the Tardieu Scale for the me asurement of spasticity. *Disability and Rehabilitation* 28: 899–907.

Held JP and Pierrot-Deseilligny E (1969) *Reeducation Mortrice des Affections Neurologiques*. Paris: J.B. Baillière et fils.

Hornby TG, Rymer WZ, Benz EN, et al. (2003) Windup of flexion reflexes in chronic human spinal cord injury: A marker for neuronal plateau potentials? *Journal of Neurophysiology* 89: 416–426.

Kiehn O and Eken T (1997) Prolonged firing in motor units: Evidence of plateau potentials in human motoneurons? *Journal of Neurophysiology* 78: 3061–3068.

Lance JW, Feldman RG, Young RR, and Koella WP (eds.) (1980) *Spasticity: Disordered Motor Control*, pp 51–181 and 485–494. Chicago: Year Book.

Le Cavorzin PL, Poudens SA, Chagneau F, et al. (2001) A comprehensive model of spastic hypertonia derived from the pendulum test of the leg. *Muscle & Nerve* 24: 1612–1621.

Lieber RL, Steinman S, Barash IA, et al. (2004) Structural and functional changes in spastic skeletal muscle. *Muscle & Nerve* 29: 615–627.

Morris S (2002) Ashworth and Tardieu scales: Their clinical relevance for measuring spasticity in adult and paediatric neurological populations. *Physical Therapy Reviews* 7: 53–62.

Nielsen JB, Crone C, and Hultborn H (2007) The spinal pathophysiology of spasticity – from a basic science point of view. *Acta Physiology* 189: 171–180.

Pandyan AD, Johnson GR, Price CIM, et al. (1999) A review of the properties and limitations of the Ashworth and modified Ashworth Scales. *Clinical Rehabilitation* 13: 373–383.

Pandyan AD, Gregoric M, Barnes MP, et al. (2005) Spasticity: Clinical perceptions, neurological realities and meaningful measurement. *Disability and Rehabilitation* 27: 2–6.

Penn RD, Savoy S, Corcos D, et al. (1989) Intrathecal baclofen for severe spinal spasticity. *New England Journal of Medicine* 320: 1517–1521.

Pierrot-Deseilligny E and Burke D (2005) *The Circuitry of the Human Spinal Cord: Its Role in Motor Control and Movement Disorders*. Cambridge: Cambridge University Press.

Platz T, Eickhof C, Nuyens G, et al. (2005) Clinical scales for the assessment of spasticity, associated phenomena, and function: A systematic review of the literature. *Disability and Rehabilitation* 27: 7–18.

Rothwell J (1994) *Control of Human Voluntary Movement*, 2nd edn. London: Chapman and Hall.

Sheean GL (2001) Botulinum toxin treatment of spasticity: Why is it so difficult to show a functional benefit? *Current Opinion in Neurology* 14: 771–776.

Snow BJ, Tsui JK, Bhatt MH, et al. (1990) Treatment of spasticity with botulinum toxin: A double blind study. *Annals of Neurology* 28: 512–515.

Tardieu G, Shentoub S, and Delarue R (1954) A la recherche d'une technique de measure de la spasticite. *Revenue neurologique* 91: 143–144.

Voerman G, Gregoric M, and Hermens HJ (2005) Neurophysiological methods for the assessment of spasticity: The Hoffman reflex, the tendon reflex and the stretch reflex. *Disability and Rehabilitation* 27: 19–32.

Wood D, Burridge JH, van Wijck FM, et al. (2005) Biomechanical approaches applied to the lower and upper limb for the measurement of spasticity: A systematic review of the literature. *Disability and Rehabilitation* 27: 19–32.

Relevant Website

http://www.seniam.org – The SENIAM project (Surface Electromyography for the Noninvasive Assessment of Muscles).

Spatial Cognition

C L Colby, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Published by Elsevier Ltd.

Behavioral Studies of Spatial Cognition in Humans

Spatial Abilities

Spatial cognition involves not one but many specific abilities. These include locating points in space, determining the orientation of lines and objects, assessing location in depth, appreciating geometric relations between objects, and processing motion, including motion in depth. These spatial skills can be applied to imagined objects as well as to external stimuli, as in the classic experiments on mental rotation. These experiments showed that observers take longer to determine whether two objects are identical when the degree of mental rotation needed to align them increases. Similar kinds of basic processes contribute to spatial cognition in the auditory and somatosensory domains.

Multiple Frameworks for Spatial Representation

Our subjective experience strongly suggests that we have direct access to a single coherent and overarching representation of space. Whether we locate an object by sight, smell, hearing, or touch, we can respond to it with equal ease and with any motor system at our command. This introspection is misleading. There is no evidence for the existence of a single, explicit, topographic representation of space suitable for incorporating every kind of sensory input and generating every kind of motor output. On the contrary, the evidence points to multiple representations of space, in a variety of coordinate frames, each closely linked to separate output systems designed to guide specific motor effectors.

The particular representation of space in use at any time depends on the task the subject is trying to perform. For example, if you were to draw out a route for a hiking trip on a map, the route would be in the coordinates of the map, the piece of paper. If you were then going to walk along that route, you would have to begin by locating your current position within that coordinate frame and constructing a representation of the route with respect to your starting point. The first, map-based representation is an example of an allocentric representation, in which locations are represented in reference frames extrinsic to the observer. Allocentric representations include those centered on an object of interest (object-centered) and those in environmental (room-centered or world-centered) coordinates. The second representation, the one in the coordinates of the hiker's current position, is an example of an egocentric representation, in which locations are represented relative to the observer. Egocentric representations include those in eye-centered, head-centered, hand-centered, and body-centered coordinates. Experimental work in humans indicates that multiple reference frames can be activated simultaneously.

Neural Basis of Spatial Cognition

Impairments of Spatial Cognition in Humans

Much of our knowledge about the neural basis of spatial cognition comes from observations in patients with spatial deficits following brain damage. These include a wide range of perceptual and motor deficits, such as poor localization of visual, auditory, or tactile stimuli; inability to determine visual or tactile line orientation; impaired performance on mazes; impairment on tests of mental spatial transformations; right–left confusion; poor drawing; impaired eye movements to points in space; misreaching; defective locomotion in space; and amnesia for routes and locations. As can be seen from this partial list, spatial behavior involves many kinds of skills, and not surprisingly, a number of brain regions have been implicated in spatial cognition and performance.

Broadly speaking, the parietal lobe is responsible for spatial perception and representation of immediate extrapersonal space, whereas temporal and parahippocampal cortices are more involved in topographic memory and navigation. The frontal lobe receives input from both parietal and temporal cortex and is responsible for generating actions. An important point about deficits in spatial cognition in humans is that they are far more common after right hemisphere damage than left. Although patients with left hemisphere damage may also exhibit spatial deficits, it is clear that the right hemisphere in humans has a superordinate role in spatial processing and behavior.

Impact of Parietal Lesions in Humans on Spatial Cognition

Two kinds of spatial deficits following parietal lobe damage are particularly illuminating. First, a common sensorimotor deficit is difficulty in using visual information to guide arm movements, referred to as optic ataxia. Patients with optic ataxia have difficulty with everyday tasks that require accurate reaching under visual guidance, such as using a knife and fork. They both misdirect the hand and misorient it

with respect to the object, and they are most impaired when using the contralesional hand to reach for an object in the contralesional half of space. A second, classic disorder of spatial cognition in humans is the tendency to ignore one half of space, called hemispatial neglect. The most common form of neglect arises from damage to the right parietal lobe and is manifested as a failure to detect objects in the left half of space. Neglect is more than just a visual deficit, however. It can occur separately or jointly across many sensory modalities. Moreover, neglect occurs with respect to many different spatial reference frames. A patient with right parietal lobe damage typically neglects objects on the left, but left may be defined with respect to the body, or the line of sight, or with respect to an attended object. Furthermore, this neglect is dynamic, changing from moment to moment with changes in body posture and task demands.

Neglect is apparent even in the purely conceptual realm of internal images. Patients exhibit neglect when asked to imagine a familiar scene, such as a city square, and describe the buildings in it. The portion of space that is neglected changes when they are asked to imagine the scene from a different viewpoint. As Figure 1 illustrates, neglect can occur with respect to an internal image constructed by the individual. Patients with neglect show evidence of using multiple reference frames, just as intact individuals do.

Impact of Parietal Lesions in Monkeys on Spatial Cognition

The role of parietal cortex in spatial cognition has been explicitly tested in animal studies. Monkeys with posterior parietal lobe lesions exhibit many of the same deficits seen in humans, including deficits in the appreciation of spatial relations between objects and impairments in eye movements and reaching. They perform normally on tests of object discrimination but are selectively impaired on a spatial task that requires them to judge which of two identical objects is closer to a visual landmark. In contrast, monkeys with temporal lobe lesions are unimpaired on the spatial task but fail to discriminate between or recognize objects. This double dissociation between the effects of lesions, in combination with the discovery of distinctive cortical inputs to the parietal and temporal lobes, led to the concept of the dorsal and ventral visual-processing streams. The ventral stream processes information about object form and color and is necessary for object recognition. In contrast, the dorsal stream processes information

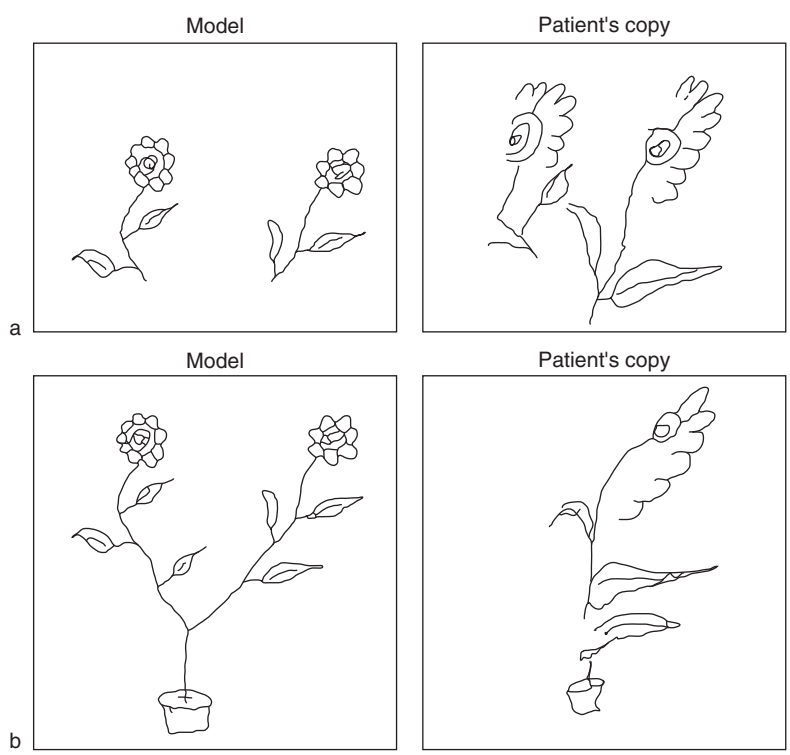

Figure 1 A test for object-centered neglect. When asked to copy the two drawings on the left, a patient made the two copies on the right. Detail is omitted from the left half of each object rather than from the left half of the drawing as a whole. Reproduced from Marshall JC and Halligan PW (1993) Visuo-spatial neglect: A new copying text to assess perceptual parsing. *Journal of Neurology* 240: 37–40, with permission.

about where objects are located and how to acquire or avoid them.

Neurons in the dorsal stream encode several types of visual information important for spatial perception. Neurons in specific dorsal stream areas are selective for orientation, depth, direction and speed of motion, rotation, and many other stimulus qualities appropriate for perceiving spatial information. The dorsal visual-processing stream leads to posterior parietal cortex, where many kinds of visual information converge, including information about stimulus shape. An equivalent set of somatosensory processing areas send tactile information to anterior parietal cortex.

Visual and somatosensory signals converge on single neurons within the intraparietal sulcus, which divides anterior and posterior parietal cortex. Auditory signals have also been demonstrated to contribute to spatial processing in monkey parietal cortex.

Spatial Information Is Encoded in Multiple Reference Frames

The standard approach for investigating the role of parietal neurons in spatial processing is to record electrical activity from individual neurons while the monkey performs a spatial task. Because brain tissue has no sensory receptors, fine microelectrodes can be introduced into the brain without disturbing the animal's performance. By recording neural responses during carefully designed tasks, neural activity can be related directly to the sensory and representational processes that underlie spatial behavior. The general conclusion from these studies is that the function of parietal cortex is to transform spatial information from its original sensory coordinates into the motor coordinates necessary for the guidance of action. The specific kind of coordinate transformation depends on the anatomical area within parietal cortex and the task to be accomplished. Several cortical areas have been defined within the parietal lobe in monkeys. Two examples of parietal areas with distinct forms of spatial representation are described next.

Head-centered spatial representation in the ventral intraparietal area The ventral intraparietal area (VIP) is located in the floor of the intraparietal sulcus, where inputs from high-order visual and somatosensory cortex converge. In the visual domain, VIP neurons are characterized by direction and speed selectivity, and they resemble neurons in other dorsal stream visual areas that process stimulus motion. Surprisingly, these same neurons were discovered to respond to light touch on the head and face. The somatosensory and visual receptive fields of individual neurons exhibit strong spatial correspondence: they match in location, size, and even their preferred direction of motion.

The existence of spatially matched receptive fields raises an interesting question: What happens when the eyes move away from primary position? If the visual receptive fields were simply retinotopic, they would have to move in space when the eyes do and so would no longer match the location of the somatosensory receptive field. Instead, for many VIP neurons, the visual receptive field moves to a new location on the retina when the eyes move away from the straight-ahead position. For example, a neuron that has a somatosensory receptive field near the mouth and responds best to a visual stimulus moving toward the mouth will continue to respond to that trajectory of motion regardless of where the monkey is looking. The conclusion is that these neurons have head-centered receptive fields: they respond to stimulation of a certain portion of the skin surface and to the visual stimulus aligned with it, no matter what part of the retina is activated. Neurons in area VIP send projections to the region of premotor cortex that generates head movements. Area VIP neurons contribute to the visual guidance of head movements and may play a special role in hand, eye, and mouth coordination. They operate in a head-centered reference frame in order to generate appropriate signals for a particular motor effector system, namely that which controls head movements.

Eye-centered spatial representation in the lateral intraparietal area In contrast to area VIP, neurons in the lateral intraparietal area (LIP) construct an eye-centered spatial representation. Individual neurons become active when a salient event occurs at the location of the receptive field. This can be a sensory event, such as the onset of a visual or auditory stimulus; a motor event, such as a saccade toward the receptive field; or even a cognitive event, such as the expectation that a stimulus is about to appear or the memory that one has recently appeared. The level of response reflects the degree to which attention has been allocated to the site of the receptive field.

As discussed for area VIP, the animal's ability to make eye movements raises an interesting question about spatial representation. Neural representations of space are maintained over time, and the brain must solve the problem of how to construct an accurate spatial representation when a receptor surface moves. With each eye movement, every object in our surroundings activates a new set of retinal neurons. Despite this constant change, we perceive the world as stable. Area LIP neurons contribute to this perceptual stability by using information about the metrics of the eye movement to update the spatial representation

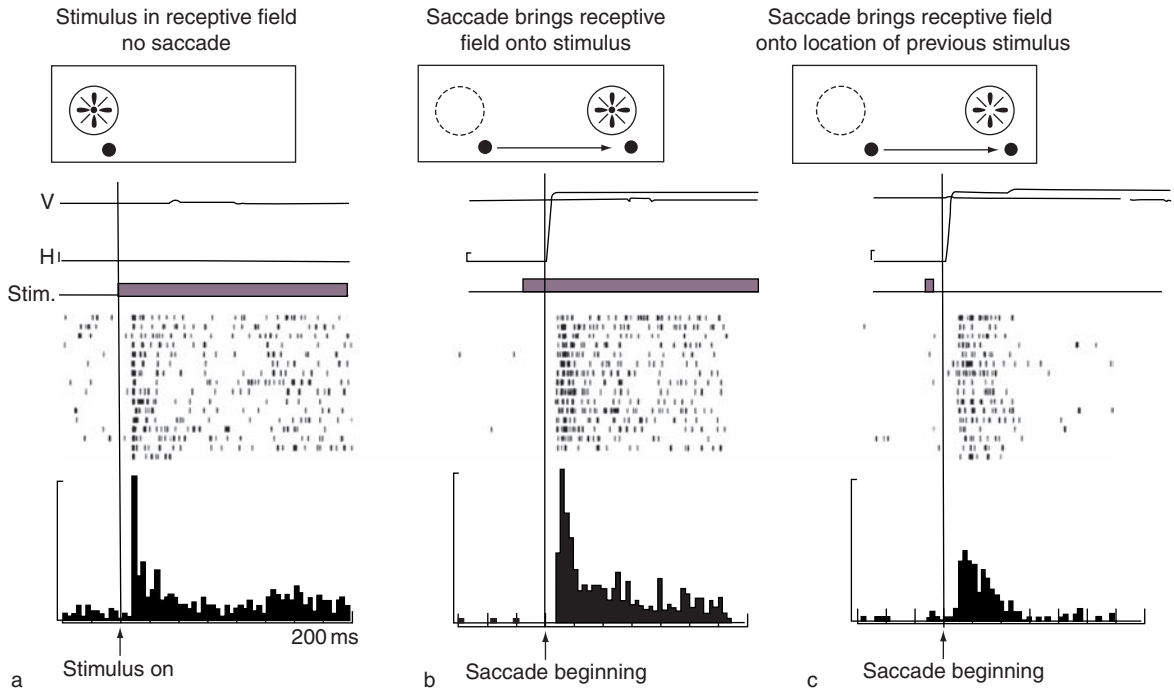

Figure 2 Remapping of visual memory trace activity in area LIP. The activity of a single neuron was recorded under three different conditions. (a) Simple visual response to a constant stimulus in the receptive field, presented while the monkey is fixating. The rasters and histogram are aligned on the time of stimulus onset. (b) Response following a saccade that brings the receptive field onto the location of a constant visual stimulus. (c) Response following a saccade that brings the receptive field onto the location where a stimulus was presented previously. The stimulus is extinguished before the saccade begins so it is never physically present in the receptive field. The neuron responds to the updated memory trace of the stimulus. H, horizontal eye position; V, vertical eye position. Reproduced from Duhamel JR, Colby CL, and Goldberg ME (1992) The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* 255: 90–92, with permission.

of a salient location. In a simple saccade task, LIP neurons are activated when the monkey makes an eye movement that brings a previously illuminated screen location into their receptive field. These neurons respond to the memory trace of the earlier stimulus: no stimulus is ever physically present in the receptive field, either before or after the saccade. The explanation for this striking finding is that the memory trace of the stimulus is updated at the time of the saccade. Before the saccade, while the monkey is looking straight ahead, the onset of the stimulus activates a set of LIP neurons whose receptive fields encompass the stimulated screen location. They continue to respond after the stimulus is extinguished, maintaining a memory trace of the stimulus. At the time of the saccade, a corollary discharge, or copy of the eye movement command, containing information about the metrics of the saccade arrives in parietal cortex. This corollary discharge signal causes the active LIP neurons to transmit their stored representation of the stimulus to the new set of LIP neurons whose receptive fields will encompass the stimulated screen location after the saccade. The representation of the stimulated location is dynamically updated from the coordinates of the initial eye position to the coordinates of the final eye position (Figure 2).

The significance of updating lies in what it reveals about spatial representation in area LIP. It demonstrates that the representation is dynamic and is always centered on the current position of the fovea. Instead of creating a spatial representation that is in purely retinotopic (sensory) coordinates, associated exclusively to the specific neurons initially activated by the stimulus, area LIP constructs a representation in eye-centered (motor) coordinates. The distinction is a subtle one but critical for the ability to generate accurate spatial behavior. By representing visual information in eye-centered coordinates, area LIP neurons tell the monkey not just where the stimulus was on the retina when it first appeared but also where it would be now if it were still visible. The result is that the monkey always has accurate spatial information with which to program an eye movement toward a current or remembered target. The transformation from sensory to motor coordinates puts the visual information in its most immediately useful form. Compared to a head-centered or world-centered representation, an eye-centered representation has the

significant advantage that it is already in the coordinates of the effector system (oculomotor) that will be used to acquire the target. Humans with unilateral parietal lobe damage fail on an eye movement task that depends on an eye-centered representation of a remembered target position. This failure presumably reflects an impairment of the updating mechanism in parietal cortex.

In summary, posterior parietal cortex plays a critical role in spatial cognition. Physiological studies in monkeys show that parietal neurons represent spatial locations relative to multiple reference frames, including those centered on the head and the eye. Individual neurons combine spatial information across different sensory modalities (in area VIP), and specific spatial reference frames are constructed by combining sensory and motor signals (in area LIP). In accord with the physiology, human neuropsychological studies show that neglect can be expressed with respect to several different reference frames.

Frontal Lobe Mechanisms of Spatial Cognition

The parietal lobe transforms sensory representations of attended objects into the motor coordinate frames most appropriate for action. It does not actually generate those actions. Dorsal stream outputs to frontal cortex provide the sensory basis, in the correct spatial framework, for producing specific motor outputs. The following sections describe different types of spatial representations found in distinct regions of frontal cortex.

Head-centered and hand-centered visual receptive fields in premotor cortex Two different forms of spatially organized visual responsiveness have been described in ventral premotor cortex. First, in a subdivision representing facial movements (area F4), neurons respond to visual stimuli at specific locations relative to the head, much like those described

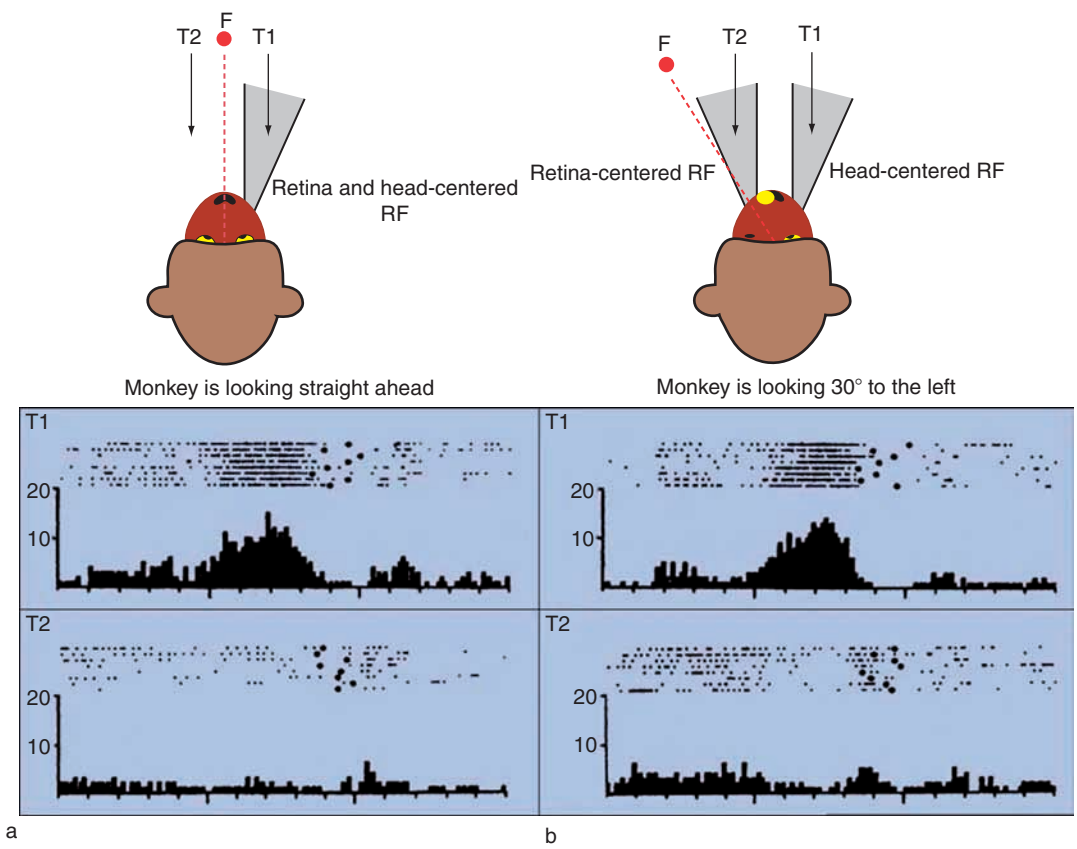

Figure 3 Data from a neuron in the premotor cortex with a head-centered visual receptive field (RF). (a) While the monkey was looking straight ahead at a fixation point (F), an object approached and receded, traveling either along trajectory 1 (T1, to the right of the head) or along trajectory 2 (T2, to the left of the head). The neuron fired strongly only when the object approached along trajectory 1. (b) When the monkey looked at a new fixation point on the far left, the neuron still responded only when the object approached along trajectory 1, indicating that the receptive field is tied to the location of the head and not to the location of the retina. The timelines above each raster and histogram show the horizontal and vertical eye position during one trial. The timeline below shows the position of the visual object as it comes toward and then moves away from the monkey. Adapted from Fogassi L, Gallese V, di Pellegrino G, et al. (1992) Space coding by premotor cortex. *Experimental Brain Research* 89: 686–690, with permission.

previously in area VIP. These neurons have been characterized by recording activity while objects approach the monkey's head along various trajectories. The preferred trajectory is constant with respect to the head, and it is not affected by changes in eye position. This specific zone of premotor cortex receives inputs from area VIP and uses the same form of spatial representation. In both cases, stimuli are represented in the motor coordinate frame that would be most useful for acquiring or avoiding stimuli near the face by means of a head movement (Figure 3).

A different kind of spatial representation has been observed in a subdivision of premotor cortex involved in generating arm and hand movements (area F5). Here, neurons respond to visual stimuli presented in the vicinity of the hand. When the hand moves to a new location, the visual receptive field moves with it. Moreover, the visual receptive field remains fixed to the hand regardless of where the monkey is looking, suggesting the existence of a hand-centered representation. A fascinating observation indicates that some neurons in this area are capable of even more abstract forms of representation. Some neurons in premotor cortex are activated both when the monkey grasps an object, such as a raisin, and when the monkey observes the experimenter performing the same action. These mirror neurons encode not just the monkey's own motor actions but also the meaning of actions performed by others.

Object-centered spatial representation in the supplementary eye field Actions are directed toward objects in the environment and toward specific locations on an object. Picking up your coffee cup requires that you locate both the cup in egocentric space and the handle in relation to the cup. The spatial reference frame that guides such movements is not limited to the egocentric representations described previously. Evidence from the supplementary eye field (SEF) demonstrates that single neurons can encode movement direction relative to the object. The SEF is a division of premotor cortex with attentional and oculomotor functions. Neurons here have visual responses and are also active before and during saccades. In monkeys trained to make eye movements to particular locations on an object, SEF neurons exhibit a unique form of spatial selectivity: they encode the direction of the impending eye movement

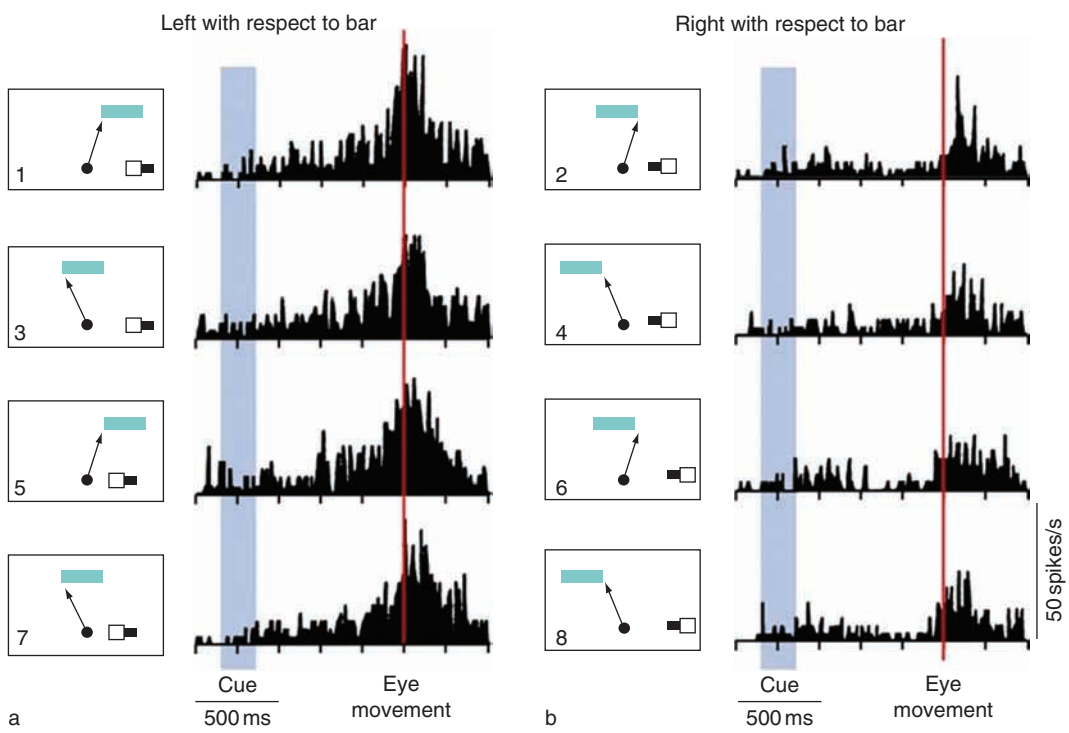

Figure 4 Data from a neuron in the SEF selective for the object-centered direction of eye movements. The monkey was trained to make eye movements to the right or left end of a horizontal bar. A cue appeared on the small sample bar shown in the lower right of each panel (1–8) to tell the monkey which end of the target bar (top in each panel) was relevant on a given trial. The arrow in the panel next to each histogram indicates the direction of the eye movement. The neuron fired strongly when the eye movement was directed to the left end of the target bar (left column) regardless of whether the physical movement of the eye was up and to the right (panels 1 and 5) or up and to the left (panels 3 and 7). Reproduced from Olson CR and Gettner SN (1995) Object-centered direction selectivity in the macaque supplementary eye field. *Science* 269: 985–988, with permission.

as defined relative to an object-centered reference frame. For example, a given neuron may fire when the monkey looks toward the right end of a bar placed at any of several different locations on the screen, regardless of whether the eye movement is a rightward or leftward saccade. Moreover, the same neuron will fail to respond when the monkey makes a physically identical eye movement toward the left end of the bar stimulus (Figure 4).

This remarkable result indicates that single neurons can make use of quite abstract spatial reference frames. Object-centered spatial information could potentially guide arm movements as well as eye movements. Moreover, neuropsychological evidence indicates that an object-centered reference frame can be used to direct attention: some patients exhibit object-centered neglect after parietal lobe damage. Parietal lobe lesions in monkeys likewise produce impairments on tasks that require an object-centered spatial representation.

Conclusion

Behavioral and neuropsychological studies indicate that we make use of multiple spatial representations in the perception of extrapersonal space. Neurophysiological studies are beginning to uncover the neural mechanisms underlying the construction of these egocentric and allocentric spatial representations. We can conclude that spatial cognition is the product of many brain areas with distinct representations of space.

See also: Attentional Networks in the Parietal Cortex; Multisensory Convergence and Integration; Neural Coding of Spatial Representations; Parietal Cortex and Spatial Attention; Prefrontal Cortex; Sensorimotor Integration: Attention and the Premotor Theory; Sensorimotor Integration: Barrels, Vibrissae and Topographic Representations; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Spatial Cognition and Executive Function; Spatial Transformations for Eye–Hand Coordination; Spatial Cognitive Maps; Vision for Action and Perception.

Further Reading

Barbieri C and De Renzi E (1989) Patterns of neglect dissociation. *Behavioural Neurology* 2: 13–24.

Behrmann M (2000) Spatial reference frames and hemispatial neglect. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Bisiach E and Luzzatti C (1978) Unilateral neglect of representational space. *Cortex* 14: 129–133.

Colby CL (1998) Action-oriented spatial reference frames in cortex. *Neuron* 20(1): 15–24.

Colby CL and Goldberg ME (1999) Space and attention in parietal cortex. *Annual Review of Neuroscience* 22: 319–349.

Duhamel JR, Colby CL, and Goldberg ME (1992) The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* 255: 90–92.

Fogassi L, Gallese V, di Pellegrino G, et al. (1992) Space coding by premotor cortex. *Experimental Brain Research* 89: 686–690.

Fuster JM (1997) *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe*, 3rd edn. Philadelphia: Lippincott-Raven.

Gallese V, Fadiga L, Fogassi L, and Rizzolatti G (1996) Action recognition in the premotor cortex. *Brain* 119: 593–609.

Graziano MS (2001) Is reaching eye-centered, body-centered, hand-centered, or a combination? *Review of Neuroscience* 12(2): 175–185.

Heilman KM, Watson RT, and Valenstein E (1985) Neglect and related disorders. In: Heilman KM and Valenstein E (eds.) *Clinical Neuropsychology*. New York: Oxford University Press.

Kakei S, Hoffman DS, and Strick PL (2003) Sensorimotor transformations in cortical motor areas. *Neuroscience Research* 46(1): 1–10.

Marshall JC and Halligan PW (1993) Visuo-spatial neglect: A new copying text to assess perceptual parsing. *Journal of Neurology* 240: 37–40.

Olson CR (2003) Brain representation of object-centered space in monkeys and humans. *Annual Review of Neuroscience* 26: 331–354.

Olson CR and Gettner SN (1995) Object-centered direction selectivity in the macaque supplementary eye field. *Science* 269: 985–988.

Rizzolatti G and Craighero L (2004) The mirror-neuron system. *Annual Review of Neuroscience* 27: 169–192.

Shepard R and Cooper L (1986) *Mental Images and Their Transformations*. Cambridge, MA: MIT Press.

Snyder LH (2000) Coordinate transformations for eye and arm movements in the brain. *Current Opinion in Neurobiology* 10(6): 747–754.

Ungerleider L and Mishkin M (1982) Two cortical visual systems. In: Ingle DJ and Mansfield RJW (eds.) *Analysis of Visual Behavior*, pp. 549–586. Cambridge, MA: MIT Press.

Spatial Cognition and Executive Function

M Behrmann, CarnegieMellon University and Center for the Neural Basis of Cognition, Pittsburgh, PA, USA
S Shomstein, George Washington University, Washington, DC, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Representing and exploring the space around us is fundamental to human cognition. The study of spatial cognition, both at the psychological and neural level, is extensive, encompassing parallel investigations of many different types of spatial knowledge and many different spatial tasks. For example, there is now a large literature concerned with the processes involved in deriving spatial relations between objects or between parts of objects, and with the way these processes differ from other mechanisms involved in object recognition, such as the representation of the parts alone.

A rather different form of spatial processing, which is also the topic of much current investigation, concerns the nature of the spatial representations engaged during the learning of the spatial layout of the environment. Two forms of spatial knowledge, both of which are deemed critical for navigation purposes, have been described in these investigations of spatial cognition: survey knowledge (based on an external perspective or an aerial or maplike view), which allows access to the global spatial layout, and route knowledge, which is the perspective represented by a ground-level observer. Recent functional neuroimaging studies have suggested that these two spatial processes recruit both a common network of cortical regions but also additional, differentiable regions of cortex. Consistent with this, in individuals with unilateral hemispheric lesions, several brain areas appear to be involved in retracing a route from the end to the beginning, including the right hippocampal formation, the right posterior parietal cortex, the right dorsolateral prefrontal cortex, and the right temporal lobe. The hippocampus also plays a particular role in path integration; specifically, individuals with lesions to the right (but not left) hippocampus are inconsistent in their navigation to previously viewed targets, and they also show a systematic underregistration of linear displacement (and/or velocity) during walking. Whereas the hippocampal coding appears to be specific to certain places in the field, parietal neurons code route maps in a more scalable and versatile form, independent of the size and spatial configuration of the navigational field.

A further strand of experimental studies concerned with spatial cognition has explored the psychological and neural mechanisms mediating spatial working memory. Keeping spatial information active online (even for extended periods of time) engages dorsal areas of lateral prefrontal cortex, posterior parietal cortex, and possibly even the hippocampal formation, and damage to these areas significantly impairs working memory. Finally, much research is concerned with delineating the nature of the underlying spatial representations, with specific emphasis on determining the reference frames or spatial coordinates employed for sensory information encoding. In some regions of cortex, the spatial position of the input is defined with respect to an eye-, head-, and/or body-centered frame of reference or sometimes some mixture of these different reference frames. While parietal cortex appears most involved in these egocentric forms of representation, in which the reference frame is centered on the observer, superior parietal cortices represent information allocentrically, independent of the observer, although, of note, these egocentric and allocentric representations appear to be somewhat interdependent. Of great interest, too, are the means whereby spatial information is updated – as the eyes move or as the observer navigates, the spatial positions of items in the environment change. To achieve a stable and robust representation, information must be constantly updated, and studies concerned with spatial updating have been (and continue to be) conducted with humans and nonhuman species, using a variety of methods, including lesion studies, functional magnetic resonance imaging, and single-unit recording.

As is evident from this brief overview, spatial cognition is multifaceted, and spatial representations are derived in the service of many different forms of cognition. Importantly, too, these different spatial representations are not mutually exclusive and much research remains to be done to specify the differences and commonalities among these representations and the spatial tasks in which they are engaged. This article focuses specifically on the spatial representations mediated by parietal cortex, often assumed to be the preeminent area involved in spatial processing. Two major lines of investigation are summarized, with one set of data primarily derived from studies of individuals with parietal lesions and the other set of data acquired from functional brain imaging studies (mostly functional magnetic resonance imaging).

Parietal Lobe and Spatial Representations

Anatomical Delineation of Parietal Cortex

The parietal lobe forms about 20% of the human cerebral cortex and is divided into two major regions, the somatosensory cortex and the posterior parietal cortex (Figure 1). Posterior parietal cortex, located at the junction of multiple sensory regions and projecting to several cortical and subcortical areas, is engaged in a number of cognitive operations, many of which involve spatial representations. Parietal cortex is separated from the frontal lobes by the central sulcus and from the occipital lobes by the parieto-occipital fissure. Since the 1940s, hundreds of studies have documented the behavioral consequences of parietal lesions in humans and monkeys, with primary emphasis on the perturbation of perceptual and action space following these lesions.

Parietal Lobe and Spatial Representations: Evidence from Hemispatial Neglect

In humans, the deficit that affects the representation of space following unilateral hemispheric lesion is referred to as hemispatial neglect ('neglect,' for short). Following brain damage, particularly to the right hemisphere (see Figure 2 for examples of lesions giving rise to neglect), individuals typically exhibit a variety of symptoms, including impaired representation of information appearing on the side of space contralateral to the lesion. These individuals may fail to report or to orient to sensory information on the contralateral side of space. Standard bedside tests, such as having the patients draw a clock from memory or copy a clock, bisect some lines, or cancel lines on a page, are typically used to elicit neglect (see Figure 2(b) for examples of patients' performance on such tests). Individuals may also show neglect of information appearing in contralesional personal space, in that they may eat, dress, and/or shave only the ipsilesional side. Furthermore, the eye and hand movements of such individuals may also be somewhat restricted to the ipsilesional side, and there is a perseverative tendency (perhaps induced by the failure to keep track of spatial locations) to revisit the same regions on the ipsilesional side. Interestingly, even in the absence of sensory input, neglect may be evident: the patients report only ipsilesional information in tasks of mental imagery and may show a paucity of contralesional saccades even when moving their eyes in the dark. Note that the neglect disorder does not arise from a primary motor or sensory deficit *per se* and is generally thought to arise from a deficit in attending to and representing the contralateral information.

Neglect also occurs more frequently and with greater severity following right than left hemisphere lesions (see also evidence from functional imaging studies on hemispheric asymmetry of parietal activation), and so we refer to neglect of left-sided information as a convenience throughout the following discussion. The scientific study of hemispatial neglect has been especially captivating over the past few decades and has continued to shed light on the mechanisms supporting spatial representation.

A critical question repeatedly addressed in the neglect literature is whether spatial information from different sensory modalities is processed in similar fashion. Although a majority of the studies have explored visual neglect, the question of interest is whether the underlying spatial map, perturbed by parietal damage, is specifically visual or utilized by other modalities as well. If the latter holds, then one might expect to observe neglect across multiple sensory modalities. Indeed, the consensus is that the representation of and awareness of space involve input from multiple sensory

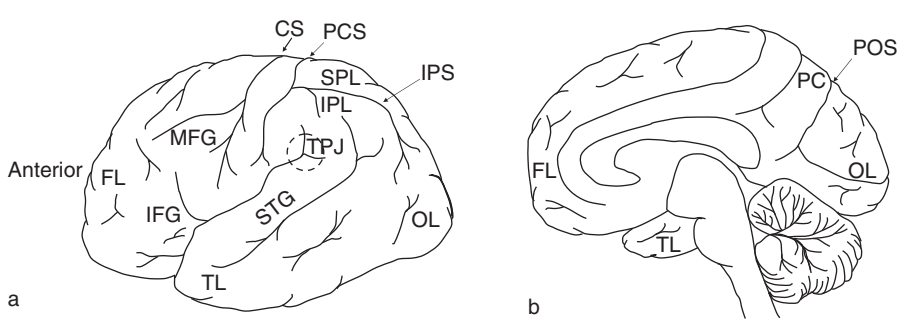

Figure 1 (a) Schematic depiction of relevant anatomical landmarks projected onto the (a) lateral and (b) medial surfaces of the human brain. Parietal cortex (PC) is located posterior to the postcentral sulcus (PCS), which lies posterior to the central sulcus (CS), and superior to the occipital lobe (OL). It is divided by the intraparietal sulcus (IPS) into the superior parietal lobule (SPL) and the inferior parietal lobule (IPL). The continuation of the SPL on the medial side, anterior to the parieto-occipital sulcus (POS), is called the cuneus. The frontal lobe (FL) is divided into the middle frontal gyrus (MFG) and the inferior frontal gyrus (IFG). The superior temporal gyrus (STG) runs along the superior extent of the temporal lobe (TL) and terminates at the temporoparietal junction (TPJ).

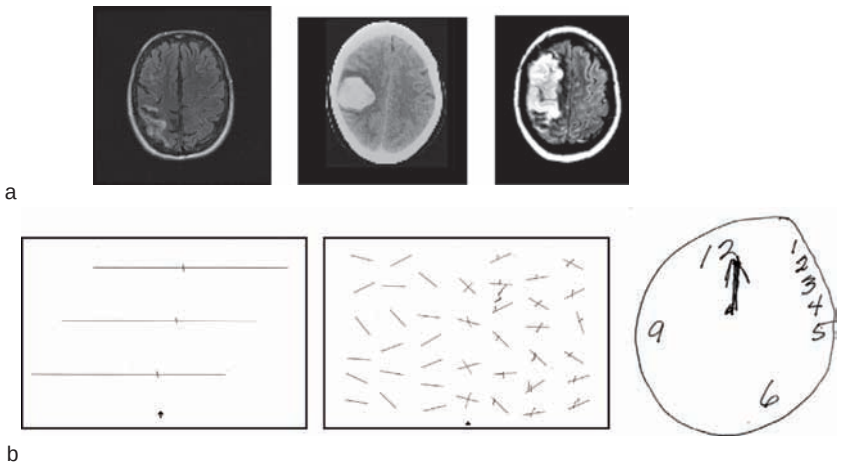

Figure 2 (a) Examples of structural scans from three different individuals, all of whom show hemispatial neglect following right parietal lobe lesions. Note that the right hemisphere is on the left of the image. (b) Examples of performance of three different patients with hemispatial neglect, using bedside tests of line bisection, line cancellation, and clock drawing from memory.

modalities: patients with visual neglect have difficulty localizing sounds and also experience problems in specifying the location of tactile input. Also, prism adaptation (see later for more details) appears to affect visual as well as proprioceptive input, further attesting to the cross-modal nature of spatial representation. Of course, one cautionary note in reaching this conclusion is that the lesions in humans are typically large and may implicate many regions of parietal cortex (in contrast to the lesions in nonhuman primates or the single-unit recording studies wherein the target areas are more circumscribed). With the advent of high-resolution structural imaging, we can start delineating better, in the neglect patient, the exact site of the lesion, the extent to which different parietal regions are affected, and the consequences of these more limited lesions on behavior.

Given that parietal cortex appears to derive a representation of space, one issue that has been under considerable investigation concerns the nature of competitive interactions between information appearing further on the ipsilesional or contralesional side of space. Consistent with ideas emerging from 'biased competition' accounts, there is a view that suggests that information appearing more contralesionally is assumed to compete with more rightward information, perhaps 'extinguishing' it entirely. If the contralesional information appears temporally in advance of the more rightward information or is much more salient than the rightward information, it might not be so readily extinguished and might ultimately be reported. It is also the case that if the contralesional information can be integrated or perceptually grouped with the rightward information, it, too, might survive the competitive interaction and be reportable. Although much of the emphasis has been on a spatial gradient with greater activation as one proceeds from left to right, there is also growing consensus that there is a temporal delay in representing spatial information (enhanced or prolonged 'attentional blink'), even when the information appears in the midline.

One recent domain in which considerable progress has been made concerns the nature of recovery of spatial attention deficits in individuals with hemispatial neglect. An intervention procedure used in some recent studies to ameliorate hemispatial neglect is to adapt the patients to prisms, which serves to induce a rightward displacement of the visual field. Several studies have successfully reported reduction in the severity of hemispatial neglect following prism adaptation and have shown that the eye movements postadaptation (in individuals with right-sided lesions and left-sided neglect) are significantly more rightward deviated than was the case prior to the intervention. Although many questions remain as to the mechanisms that give rise to this lateral-displacement postadaptation (Is it a redistribution of spatial attention? Is it visuomotor in nature?), and even though the consequences of adaptation are not very prolonged, the technique has proved to be remarkably valuable in flexibly altering the behavior of the patients with neglect, and much remains to be done in this domain.

Even without prism adaptation, following standard rehabilitation procedures, individuals with hemispatial neglect do show a decrement in the bias to orient rightward as well as showing a significant improvement in attention reorienting to the contralateral space. Of particular interest in recent investigations, as revealed in functional imaging studies, is that the

behavioral improvement even in these standard treatment procedures is correlated with reactivation of multiple cortical areas, which previously evinced weak or no task-related activity (even though some of these areas were anatomically intact). The opportunity to use functional imaging procedures to track recovery and perhaps to identify the optimal procedures for reinstituting cortical activation is promising and offers a new method for developing and assessing rehabilitation procedures.

Parietal Lobe and Spatial Representations: Evidence from Lesion Analyses

The advent of high-resolution imaging (and its relative ease of accessibility) has also enabled investigators to track the lesion site in greater detail than has been possible in the past and to reevaluate the anatomical underpinnings of hemispatial neglect. While neglect has been most commonly associated with the inferior parietal lobule and also the temporoparietal junction (TPJ; see Figure 1), the underlying anatomical substrate has come under increased scrutiny. For example, some findings suggest that the superior temporal gyrus (STG) is the critical site of cortical damage associated with neglect and subcortical sites, including the putamen, and, to a lesser extent, the caudate nucleus within the basal ganglia and the pulvinar within the thalamus have also been implicated. Yet others have argued that a disconnection between the frontal and the inferior parietal region, brought about by a lesion of the superior longitudinal fasciculus, gives rise to chronic neglect. A further set of investigations has suggested that parietal cortex has a more dorsal and a more ventral stream and that each may play a slightly different role, with more superior regions engaged in voluntary, goal-directed attention and more inferior regions more engaged in bottom-up detection of salience or novelty. Finally, large-scale studies report that neglect arises after lesions to any one of multiple nodes making up a distributed spatial-attentional network, including temporal, parietal, frontal, and occipital cortices as well as basal ganglia and thalamus (in order of incidence). The continued development of more sophisticated imaging techniques and methods of lesion analysis will be critical in outlining the neural substrate of spatial representation and possibly even in differentiating subtypes of neglect and their cortical signatures.

Parietal Lobe and Spatial Representations: Evidence from Functional Imaging Studies

Thus far, the role of parietal cortex in spatial representation has been considered based on findings from human lesion studies. Converging evidence from functional imaging studies has recently contributed enormously to elucidating the mechanisms underlying spatial representation. For example, one robust finding is that when cues direct an observer to a spatial location in which an upcoming target is likely to appear, activation is noted in superior frontal, inferior parietal, and superior temporal cortices. Spatial attentional selection serves to facilitate perception of stimuli (e.g., objects), as reflected in enhanced speed of processing of attended information. Given the increased access to functional imaging, researchers have spent considerable effort to shed light on the mechanism subserving spatial attentional orienting. A primary focus of such studies has been to determine the anatomical loci within the human cortex that give rise to the attentional biasing signal (i.e., the source or process that directs or shifts attention) that ultimately initiates the sensory enhancement of the selected stimulus (i.e., the effect). In this next section, we summarize findings from these studies, showing the differential involvement in the source versus the effects of attentional manipulations.

Effects of Spatial Selection: Enhancement of Sensory Signals

Studies of the effects of visual attentional selection have revealed behavioral facilitation and enhanced cortical responses to information appearing in attended spatial locations. Neurophysiological studies investigating the effects of top-down attentional bias have demonstrated that when a cell is presented with a stimulus in its receptive field (RF), the neural response to that stimulus is increased when attention is covertly directed to it as compared to when the stimulus is in the RF but attention is either unfocused or is diverted elsewhere. For example, in their seminal study, Moran and Desimone first identified the classical RF of a V4 neuron and its corresponding effective and ineffective stimuli (i.e., stimuli that drive the cell strongly and weakly, respectively). The monkeys were trained to attend to stimuli in a specific location within the visual field while ignoring stimuli at another location. Both effective and ineffective stimuli were presented within the RF and the responses of the V4 neurons were strongly modulated by the spatial attentional locus: the firing rate to the preferred stimulus was only one-third as great when the monkey did not attend to it as compared to when the preferred stimulus was attended. In fact, what was particularly interesting about this study is that when attention was directed to one of two stimuli in the receptive field of a V4 cell, the effect of the unattended stimulus was attenuated, as if the receptive field had contracted to include only the attended stimulus. This type of spatial

attentional enhancement was found to increase with task difficulty and was observed throughout early sensory regions within the visual cortex.

Several neuroimaging studies in humans have also documented early sensory enhancement following top-down spatial attentional allocation. In the majority of these experiments, individuals are presented with stimuli appearing to the left or right of fixation and are asked to direct their attention to either side of the fixation point (i.e., shifting their attention either to the left or the right side of space) (Figure 3). Directing spatial attention to the left hemifield results in increased stimulus-evoked neural activity of the early visual areas in the right hemisphere, whereas directed spatial attention to the right hemifield is accompanied by increase of activity in the early visual areas of the left hemisphere. Studies employing event-related potentials (ERPs) take advantage of the fact that visual stimuli typically elicit two early waveform components, which are termed P1 (first positive, occurring 100–200 ms after stimulus onset) and N1 (first negative). These effects are robust and are observed even during passive viewing of input. Of relevance, the P1 and N1 components are modulated (i.e., increase in negativity and/or positivity) when visual stimuli are being attended to, as opposed to when spatial attention is diverted elsewhere in the scene. The modulation is manifest by a change in the amplitude of these components, while the component latency remains the same when the stimuli are attended or ignored.

In summary, spatial attention serves to modulate sensory signals elicited by the stimuli in the early sensory cortex (i.e., areas V1–V4) and also in higher order areas such as the motion-sensitive region (MT+), fusiform face area (FFA) selective for face processing, and other higher order regions within the human cortex.

Sources of Spatial Orienting Signal

Several regions within monkey and human posterior parietal cortex (PPC) have been identified as coarsely representing spatial topography as it relates to the source or trigger of a spatial attentional orienting signal. For example, neurons in the lateral intraparietal area (LIP) of the monkey cortex fire in response to the location in space to which attention was cued, representing the voluntary allocation of attention to a specific region in space. Several regions within the human parietal and frontal cortices – inferior parietal sulcus (IPS), superior parietal lobule (SPL), and the frontal eye fields (FEFs) – are also identified as playing a critical role in spatial attentional selection. Functional magnetic resonance imaging (fMRI) studies have demonstrated that activity in these frontal and parietal regions is spatially selective, such that stronger activity is noted when attention is directed toward spatial locations in the contralateral field.

Several fMRI studies support the proposal that the top-down attentional signal (the implementation of the cue) arises at least partially from the parietal

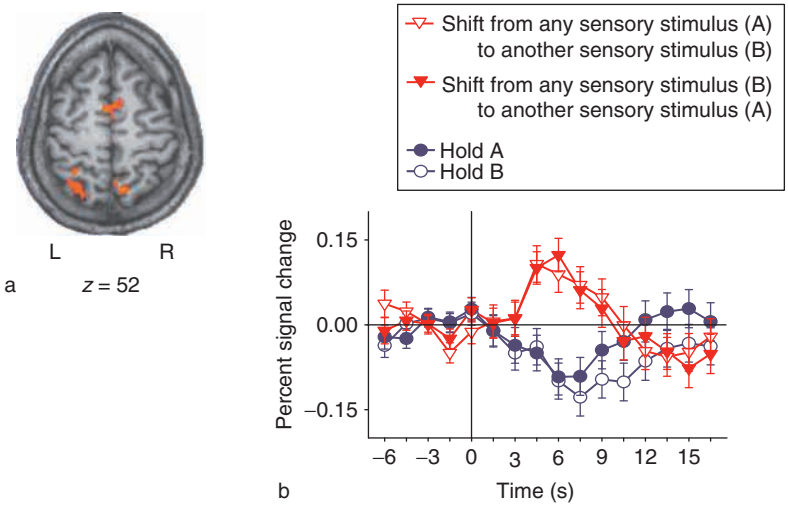

Figure 3 (a) A representative slice through a high-resolution reconstructed brain of an individual performing an attentional shifting task. The activation (shown in red) is obtained by contrasting activity during shifts of attention between the sensory stimuli (could be, for example, from left side of space to right side of space, or from auditory stimuli presented to the left ear and to the right ear) with activity when holding attention within a sensory stimulus. The cluster of activated voxels encompassed within the white circle is in the right superior parietal lobule. (b) Blood oxygen-level-dependent signal extracted from the right superior parietal lobule reflects transient activity elicited within this region following shifts of attention from one sensory stimulus to another. Note that activity within this region remains low for sustained attention (hold trials). From Shomstein S and Yantis S (2006) Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *Journal of Neuroscience* 26(2): 435–439.

cortex, with an additional contribution from the frontal cortex – what has been termed the frontoparietal attentional network. For example, when individuals are asked to covertly direct their attention following a cue to a spatial location in the visual display in preparation for a target, areas in the occipital lobe respond transiently to the cue, while areas in the posterior parietal cortex along the inferior parietal sulcus exhibit a more sustained response. The sustained activity in the PPC is assumed to reflect the source of the attentional orienting signal: it could not be linked to either visual stimuli or motor responses *per se* and its activity was time locked to the period during which individuals paid attention to the peripheral future target location.

It is important to understand the precise function of the parietal lobe's involvement in attentional control. At least two possibilities regarding the time course of parietal lobe activity following spatial shifts of attention must be considered. The first possibility is that the PPC provides a continuous signal that maintains the locus of attention at a specific spatial location – the 'sustained hypothesis.' In other words, once the individual has decided to change the locus of spatial attention, the PPC issues a signal that will be sustained until another change in the locus of spatial attention is made (i.e., sustained increase in neural activity). Another possibility is that the PPC issues a transient brief signal to shift the attentional state – the 'transient hypothesis,' such that once the locus of spatial attention needs to be redirected elsewhere, the PPC issues a transient signal that dissipates quickly once the attentional shift has been made. This latter hypothesis also predicts that the locus of attention is maintained by other areas, presumably frontal areas within the frontoparietal network.

Several recent studies examined the exact involvement of the PPC in attentional shifting tasks by adjudicating between these two hypotheses. In a typical task, individuals are shown two streams of input presented peripherally to the left and right of fixation and are initially instructed to monitor one stream for a cue (for example, a digit among the stream of letters). The identity of the cue indicates to the individuals whether they should maintain attention on the current stream or shift attention to the other stream. The critical finding is revealed when one compares cortical activity related to shifting the attentional focus from one spatial location to another with activity related to maintaining the attentional focus on a single stream. When attention was shifted between spatial locations, the time course of the blood oxygen-level-dependent (BOLD) signal in the right SPL and inferior parietal lobule (IPL) exhibited transient activity (see Figure 3). The transient nature of the signal elicited by the SPL suggests that this area of the parietal cortex is the source of a brief attentional control signal to shift attentive states to a new spatial position, and is not the source of a continuous signal to actively maintain the new attentive state.

Sources of Nonspatial Orienting Signal

Although the focus thus far has been on spatial representation and attending to spatial regions containing input, the top-down attention-shifting signal is not restricted to spatial shifts alone, and this region is activated when individuals shift their attention between any two dimensions of the input. For example, shifts between superimposed houses and faces, shifts between two different features of an object, or shifts between two different sensory modalities all selectively activate the PPC. Whereas spatial shifts are accompanied by increased activation in the SPL region of the parietal lobe, nonspatial shifts are accompanied by increased activity in the precuneus region, the continuation of the SPL on the medial side of the parietal lobe (see Figures 1 and 3). This anatomical distinction between spatial and nonspatial shifts of attention clearly deserves a more thorough investigation. We should note that additional attentional functions could be mediated by some of the same subareas of parietal cortex already described. For example, both an anterior part and an inferior part of the intraparietal sulcus appear to be activated in a visual conjunction task, even in the absence of multiple distractors, and this occurs to a greater extent in the left than in the right hemisphere. By contrast, activation of a more posterior region of the intraparietal sulcus appears to be contingent on the presence of distractors. Much work remains to be done to outline the various attentional functions mediated by the parietal cortex and also, in parallel, to map out the behavioral and neural mechanisms associated with spatial representation.

Conclusion

Spatial cognition is a complex, multifaceted set of processes that are engaged in a large variety of tasks, including, for example, mental rotation, spatial navigation, and spatial working memory. The exact nature of the spatial knowledge implicated in these different tasks remains to be determined. We have focused on behavioral and neural mechanisms involved in representation of spatial locations, with evidence drawn from neuropsychological studies of individuals with hemispatial neglect and neuroimaging studies of spatial attentional selection (and the ensuing sensory enhancement). As is evident, the

findings are probably more complicated than we might have thought, given the progress made to date on these topics. In particular, better use of structural and functional imaging has led to fine-grained reconsideration of previous findings. Some of the received wisdoms regarding hemispatial neglect have come under scrutiny, and the fMRI findings have opened new avenues of investigation. Among the outstanding issues are the mechanisms giving rise to the neglect in the first place (for example, a deficit in spatial attention, a problem in spatial updating, or competition between inputs). The neural locus of neglect has also come under reconsideration and many questions remain. Findings from functional imaging studies have also opened new domains of investigation. In particular, the imaging studies have suggested that spatial positions in the world can be selectively attended and that these attentional effects are robust and are present throughout the visual (and auditory) system, and attending to specific stimuli modulates the cortical areas involved in processing the input. The posterior parietal cortex appears to play a dominant role in signaling and implementing the attentional shift from one location (or feature or object) to another, as reflected in the transient activity in posterior parietal cortex.

See also: Attentional Networks in the Parietal Cortex; Cognition: An Overview of Neuroimaging Techniques; Executive Function and Higher-Order Cognition: Neuroimaging; Neglect Syndrome and the Spatial Attention Network; Neural Coding of Spatial Representations; Parietal Cortex and Spatial Attention; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Spatial Memory: Assessment in Animals; Spatial Transformations for Eye–Hand Coordination; Spatial Cognition; Spatial Cognitive Maps.

Further Reading

Bisley JW and Goldberg ME (2003) Neuronal activity in lateral intraparietal cortex and spatial attention. *Science* 299: 81–86.

Chambers CD and Mattingley JB (2005) Neurodisruption of selective attention: Insights and implications. *Trends in Cognitive Science* 9(11): 542–550.

Corbetta M, Kincade JM, and Shulman GL (2002) Neural systems for visual orienting and their relationships to spatial working memory. *Journal of Cognitive Neuroscience* 14(3): 508–523.

Corbetta M, Kincade MJ, Lewis C, et al. (2005) Neural basis and recovery of spatial attention deficits in spatial neglect. *Nature Neuroscience* 8(11): 1603–1610.

Luck SJ, Girelli M, McDermott MT, et al. (1997) Bridging the gap between monkey neurophysiology and human perception: An ambiguity resolution theory of visual selective attention. *Cognitive Psychology* 33: 64–87.

Malhotra P, Coulthard E, and Husain M (2006) Hemispatial neglect, balance and eye-movement control. *Current Opinion in Neurology* 19(1): 14–20.

Merriam EP, Genovese CR, and Colby CL (2003) Spatial updating in human parietal cortex. *Neuron* 39(2): 361–373.

Moran J and Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. *Science* 229: 782–784.

O'Craven KM, Rosen BR, Kwong KK, et al. (1997) Voluntary attention modulates fmri activity in human mt-mst. *Neuron* 18(4): 591–598.

Philbeck JW, Behrmann M, Levy L, et al. (2004) Path integration deficits during linear locomotion after human medial temporal lobectomy. *Journal of Cognitive Neuroscience* 16(4): 510–520.

Serences JT, Schwarzbach J, Courtney SM, et al. (2004) Control of object-based attention in human cortex. *Cerebral Cortex* 14(12): 1346–1357.

Shelton AL and Gabrieli JDE (2002) Neural correlates of encoding space from route and survey perspectives. *Journal of Neuroscience* 22(7): 2711–2717.

Shomstein S and Behrmann M (2006) Objects modulate competition in human parietal and extrastriate cortices. *Proceedings of the National Academy of Sciences of the United States of America* 103(30): 11387–11392.

Shomstein S and Yantis S (2006) Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *Journal of Neuroscience* 26(2): 435–439.

Treue S and Maunsell JHRM (1999) Effects of attention on the processing of motion in macaque visual cortical areas mt and mst. *Journal of Neuroscience* 19: 7603–7616.

Tversky B (1991) Spatial mental models. In: Bower GH (ed.) *The Psychology of Learning*, pp. 109–145. San Diego: Academic Press.

van Asselen M, Kessels RP, Kappelle LJ, et al. (2006) Neural correlates of human wayfinding in stroke patients. *Brain Research* 1067(1): 229–238.

Vandenberghe R, Gitelman DR, Parrish TB, et al. (2001) Functional specificity of superior parietal mediation of spatial shifting. *NeuroImage* 14(3): 661–673.

Wojciulik E and Kanwisher N (1999) The generality of parietal involvement in visual attention. *Neuron* 23: 747–764.

Yantis S, Schwarzbach J, Serences JT, et al. (2002) Transient neural activity in human parietal cortex during spatial attention shifts. *Nature Neuroscience* 5(10): 995–1002.

Spatial Cognitive Maps

B Poucet and E Save, Université de Provence, Marseille, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Several theories have been developed to account for spatial learning. For S–R theorists, spatial learning consists of chaining together a number of motor responses linking relevant external stimuli (stimulus–response associations). This view was challenged by Tolman in 1948, who argued that rats do not merely base their actions on S–R associations, but internally reorganize acquired spatial information so as to form cognitive representations of the environment. In other words, rats learn the location of a place where they have been rewarded independently of the specific movements necessary to reach it. Tolman explained this ability by introducing the notion of a cognitive map, a mental representation held by the rat in which it encodes the elements of a task and the spatial relationships between these elements. An important property of cognitive maps is that they allow animals to react to stimuli that are not immediately present since the relationship of such stimuli to those actually perceived is maintained in the representation. Thus, the rat can be aware of the properties of the environment beyond their field of perception. The major consequence of such a representation is that it bridges informational gaps about the environment, thus conferring greater flexibility and efficacy to behavior.

Tolman's notion of cognitive map was further developed by O'Keefe and Nadel in 1978, who proposed a theoretical framework to account for spatial behavior and its neural bases. They made a distinction between two main processes used by rats to perform spatial tasks. They proposed that rats can use either a taxon system or a locale system. The taxon system is based on S–R associations, whereas the locale system is based on spatial relationships in the form of a cognitive map (Figure 1). They further proposed that the brain regions subserving these forms of spatial navigation are different and at least partially independent.

Spatial Strategies

The number of studies which support the hypothesis that animals form a mental representation of their environment is quite impressive. This does not mean, however, that animals rely exclusively on spatial representations to orient in space. Performing specific responses, approaching particular cues, or even learning conditional associations about the sensory consequences of specific movement sequences may be efficient strategies in a number of situations. Nevertheless, in spite of their efficacy, such strategies do not confer behavior with the degree of flexibility afforded by the use of a cognitive map.

Selection of appropriate navigational strategies is primarily determined by the perception of space, that is, by the nature of the cues that can be used for navigation. There are two main categories of spatial cues: allothetic and idiothetic cues. Allothetic cues are provided by the environment and include visual, olfactory, and auditory information. Idiothetic cues are derived from the animal's own movements, including information provided by the vestibular, proprioceptive, and somatosensory systems, efference copies of motor commands, and external motion-related information such as optic flow. Although this article is about navigation based on cognitive maps, it is useful to briefly describe other strategies useful for moving efficiently in space.

Cue Navigation

The simplest navigational strategy is to move toward or away from a directly perceived cue. The use of such a system requires a radial stimulation gradient field centered on the source. The notion of gradient field refers to the variation of intensity of the source as a function of the distance. In the simplest case, intensity varies monotonically with the distance. The animal can therefore reach the goal, that is, the origin of the gradient, by navigating in the gradient direction or conversely, avoid a place by navigating in the opposite direction. A gradient field may originate from auditory and olfactory cues. Navigation toward a conspicuous visual cue has similar characteristics. The visual cue is considered to be at the center of a radial gradient where the gradient direction corresponds to the direction of the light flux. Then, reducing the distance from the visual cue (that leads to an increase of the apparent cue size) is equivalent to navigating along the axis of greatest intensity. The animal can also use a similar mechanism to reach a hidden goal that is closely associated with a salient landmark (called a beacon). A landmark is therefore considered as an intermediate goal whose spatial contiguity with the real goal has been learned through an associative process.

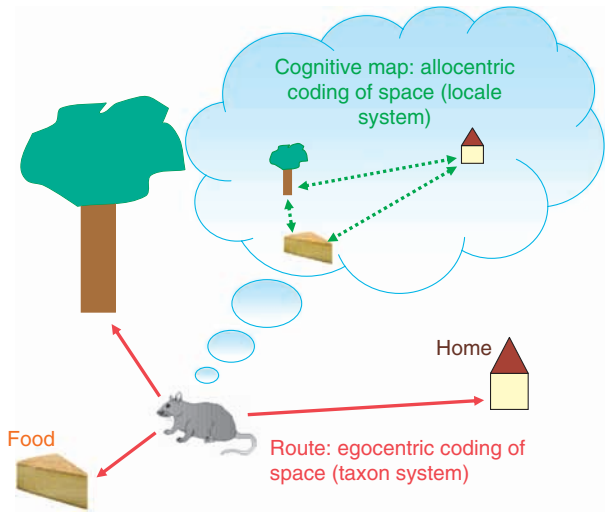

Figure 1 The two systems for encoding spatial information proposed by O'Keefe and Nadel in 1978. Cognitive mapping (locale system) involves building a representation of a part of space that encodes the relationships between environmental landmarks (e.g., the nest, a tree, and a food source), independently of the animal's location (allocentric coding). Using a route (taxon system) involves encoding the location of landmarks relative to the animal's position (egocentric coding).

Guidance

If the goal cannot be directly perceived and if there is no landmark closely associated to the goal, the animal must rely on environmental landmarks to navigate. One strategy is to use landmarks as directional cues. Guidance, therefore, requires that the animal directs its attention to particular landmarks and maintain some egocentric (i.e., relative to the animal) spatial relationships with respect to these landmarks to reach the goal. A place-recognition process is then needed to identify the goal location. Rats may be guided either by individual or configurations of landmarks. For example, they may be able to memorize a view (or snapshot) of the spatial arrangement of landmarks seen from the goal and so adjust movements until a current view matches a memorized view.

Routes

To reach a remote goal, cue navigation and guidance may be elaborated into route learning. In route learning, rats may memorize a sequence of associations coupling specific landmarks and motor responses, for example, "after reaching the rock, turn left." Each landmark is considered as an intermediary goal and the sequence of these goals constitutes a route. Routes allow rapid navigation at the expense of behavioral flexibility, however. Environmental changes altering the sequence of landmarks may be fatal to successful navigation. In addition, retracing the route cannot be achieved by simply inverting the order of the landmarks, but requires active learning of a new sequence.

Path Integration

Path integration, or dead reckoning, relies on an 'egocentric' (self-centered) coding process, which allows an animal to memorize a location (usually the starting point of its journey) in relation to its own position. This form of navigation relies on idiothetic (self-motion) cues, which are used to update a vector that specifies, in egocentric coordinates, the distance and head-referred direction of the goal from the animal's current location. Its major drawback is that the computation of these parameters is subject to cumulative error. If the cumulated error is great enough, the animal may miss its target. Although the egocentric coding process is unreliable, its precision can be improved if information is recalibrated from visual or other sensory information available from the environment. It is not clear, for rats or humans, whether any reliance is placed on path integration except when external sensory information is missing or inadequate.

Allocentric (Place) Navigation

Allocentric navigation refers to the process of determining and maintaining a course or trajectory from one place to another by using environmental cues. It requires formation and use of a map-like representation of the environment (a cognitive map) which encodes the allocentric relationships between landmarks and places, independently of the animal's position. The use of a mapping strategy allows an animal to reach a goal from a variety of points, including over trajectories that it has not previously used. A rat is thus able to infer its location relative to particular places whose position is encoded in the representation and compute optimal trajectories to reach these places. Place navigation is flexible because if a possible path is obstructed, or if a landmark is missing, the animal is still able to generate successful trajectories by performing detours and short-cuts and by adapting its navigational behavior to environmental changes.

One of the most versatile test to study place navigation is the Morris water maze task, in which the animal swims in a circular tank filled with opaque water in order to reach a submerged or visible platform, which is the only place to which it can escape from the water. In the place version of the task, there is no cue closely associated with the platform. Localization of the hidden platform can only be performed

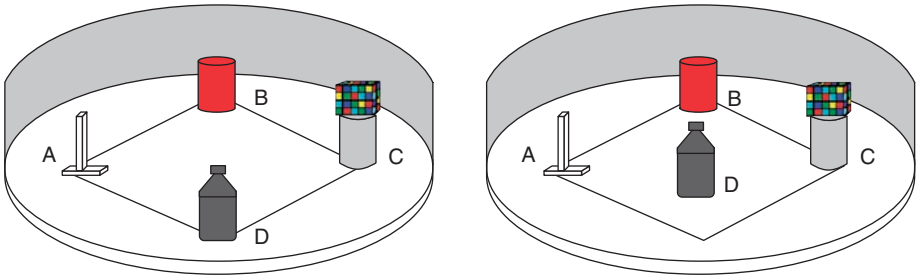

Figure 2 Role of exploration in updating cognitive maps. A typical protocol involves the rat exploring an enclosed apparatus containing a set of objects in stable locations. After the rat is familiar with this object configuration, a change is made in the spatial arrangement of objects (here a single object, D, is slightly moved from its original location). When reintroduced into the apparatus, animals usually display an exploratory reaction toward the moved object, indicating that they detect the change and update their cognitive map accordingly.

and learned by using distal room cues. To prevent the animal from using a local view for guidance, different starting places are used on successive trials. It is therefore assumed that learning results in the elaboration of a cognitive map. Acquisition of the Morris navigation task however also requires procedural memory and working memory. Procedural memory allows the rat to learn that it must swim, find an escape platform, climb on the platform, etc. Working memory results in learning that the platform is at a specific location, which rats can often do in a single trial after mastering the procedural details of the task.

Exploration and Spatial Cognitive Maps

Although an obvious function of cognitive maps is to indicate where potentially important objects are located in space, they can be used for other purposes. As a matter of fact, animals need to be familiar with the surroundings and will not behave according to their primary motivation, such as hunger or thirst, until they recognize and identify where they are in space. This ability to locate oneself in the current environment depends on prior exploration. In fact, navigation and exploration are so intimately related that accurate navigation cannot occur without preliminary exploration of the environment. The major function of exploration is information gathering and storage and so it is necessary for the formation of spatial representations. An advantage of such storage is that it may allow detection of unexpected new information in an otherwise familiar environment. For example, if place recognition is not achieved, exploration occurs again so that the cognitive map comes to match the real environment as closely as possible, therefore providing the animal with spatial invariants. Exploration may lead either to a new representation or to the updating of a former spatial representation.

An illustration of the updating process is provided by studies which show that changing the spatial configuration of objects in a previously explored arena induces a renewal of exploratory activity mostly aimed at the displaced objects (Figure 2). The protocol consists of two successive phases: habituation and spatial change. First, repeated exposure to the same configuration of objects results in habituation of exploratory activity, that is, a decrement in the response with increasing familiarity. It is hypothesized that during this phase, rats acquire some knowledge of the spatial characteristics of the environment. Habituation thus reflects the elaboration of a spatial representation. Once habituation is achieved, the animals are exposed to a new configuration of the same familiar objects. A renewal of exploration following this change is assumed to reflect that the animal (1) encodes certain spatial relationships and (2) is able to compare the previous and current situations. Such renewal is thus an adaptive behavior aimed at updating the spatial representation of the environment. It is therefore clear that there is a constant interplay between exploration and the cognitive map, with exploration allowing for initial buildup and subsequent updating of the map, which in turn drives further spatial behavior, including exploration, within the animal's current environment.

What Is the Information Stored in Spatial Cognitive Maps?

That animals are able to use representations of their environment does not mean that they encode all aspects of the spatial surroundings. In fact, evidence suggests that the information content of spatial cognitive maps can be quite selective.

Configurations of Cues and Geometry

In most spatial tasks, performance appears to rely on configural rather than on individual landmarks. Up to a point, removing specific cues in tasks such as the water maze place navigation task (or the radial arm

maze tasks) usually does not alter animal's performance to a great extent. In contrast, rearranging these cues can be quite detrimental to task solution. Based on the analysis of exploratory patterns, it can also be demonstrated that changing the spatial configuration of discrete objects in a familiar arena induces a renewal of exploratory activity. Further analyses of exploratory responses confirm that the animal keeps a record of the spatial arrangement of the objects, but that this record is specific to certain classes of spatial relationships. In general, changes that induce the strongest responses are those that affect either the overall geometrical arrangement of the object set, or the topological relationships among the objects. The configuration seems to be privileged over the absolute position of objects. Thus, spatial geometry seems to constitute an important piece of information contained in the map. In other studies, rats were placed in a rectangular chamber and were required to visit the four corners of the chamber. Each corner contained a different amount of food and was associated with a distinctive visual insert. Two opposite corners had the most bait while the remaining two opposite corners had the least bait. After the animals mastered the task (i.e., visited each corner according to a decreasing order from the most baited to the least baited arm), a number of probe tests were conducted based on transformations of the initial distribution of the inserts. The rats' patterns of visits to the food locations were remarkably insensitive to the modifications made in the spatial arrangement of the inserts. Rather, the animals used the rectangular shape of the experimental chamber as a means to locate the various food sources. Thus, rats ignore obvious landmarks and attend to landscape features instead in reorienting themselves in symmetric environments. Although they are aware of the landmarks, they simply do not appear to use them for certain purposes. The rules governing when landmarks or landscapes are used in behavior remain to be fully explicated. Current thinking suggests that local landmarks are not used to determine directions in space, nor to define a spatial framework within which such directions can be nested. Rather, landscape features are seemingly used for these purposes.

Distal versus Local Landmarks

Landscape features are not the sole markers of the spatial arrangement. Distal landmarks provide another major source of information on which spatial navigation can rely. In general, animals will preferentially use such distal information when it is available, to the detriment of local cues. As an example of the greater efficiency of distal cues in guiding spatial behavior, consider the navigation performance of rats trained in the water maze navigation task. In this task, a swimming rat can escape from water by locating a safe platform relative to the array of visual cues located 'outside' the swimming pool. After a few training trials (usually fewer than 30), the rat swims almost directly toward the platform. In contrast, when the task is changed so that rats have to rely on three objects placed directly in the swimming pool (local cues) to locate the platform, many more trials are required before the rat displays direct navigation paths.

Distal information also predominates when it conflicts with movement-related information such as that provided by path-integration mechanisms. While path-integration-based homing behavior in hamsters is barely altered by the manipulation of local cues, it can be easily altered by manipulating distal information. To understand why distal cues are so effective in controlling place navigation, it is necessary to realize that only distal cues maintain their reciprocal relationships with respect to the animal during its motion. Thus the perceived reciprocal relationships between distal cues are minimally affected by an animal's movements. On the contrary, the perceived reciprocal relationships between local cues are subject to strong changes during movements. Since a cognitive map contains absolute topographical information rather than information about egocentric locations relative to the animal, it becomes evident that distal cues provide a more reliable source of information for localization since they retain relatively stable relationships as the animal moves about the environment.

Metrics and Topology

The evidence reviewed above indicates that rats possess a representation of the configural aspects of the distal environment. This representation is not homogeneous, however: animals process certain locations in a more detailed way during exploration. This is not very surprising because space is not homogeneous. It is more surprising, however, that the topological relationships among such locations (e.g., whether they are in the same vicinity) often have stronger control over spatial behavior than do metric relationships (e.g., their absolute distance in Euclidean space). It is therefore possible that spatial representations can be both topological, therefore affording relatively unstructured information about the connectivity within space (e.g., place A is directly connected to place B but not to place C), and metric (e.g., place A is a certain distance and direction from place B), affording more detailed information about specific

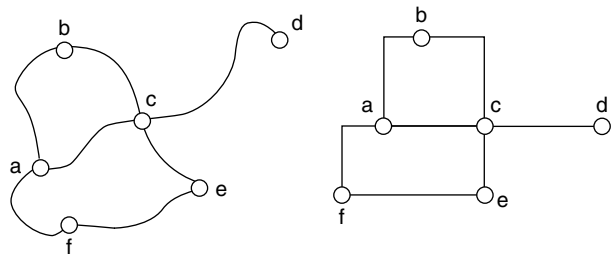

Figure 3 Metric and topological information can be used separately to describe space. The left part of the figure shows a environment in which a set of spatial locations (a–f) are connected to each other by specific paths. An arbitrary topological transform of the spatial network is shown in the right part of the figure. The topologically equivalent network preserves the topological properties of vicinity and connectivity between the locations, but not the metric relationships (i.e., angles and distances between locations).

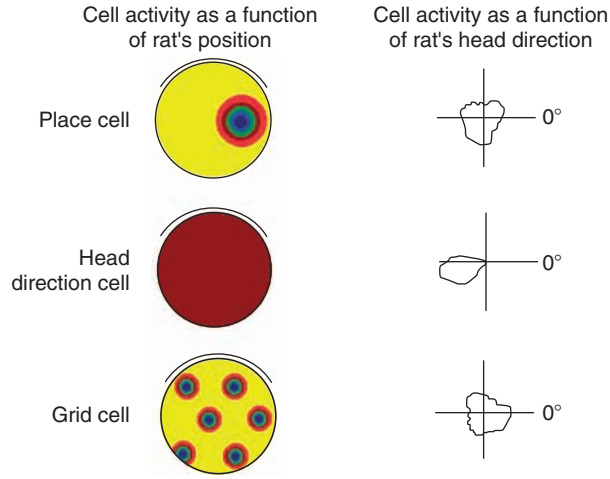

Figure 4 Main properties of place cells, head direction cells, and grid cells. Cells are usually recorded as a rat explores a cylindrical environment in which a single landmark (a cue card) is attached to the apparatus wall (shown as an arc at 12:00 in the figure). The left-hand and right-hand parts of the figure show the typical positional firing rate map and directional polar plot for each type of cell, respectively. Place cells discharge only at a specific location (shown as concentric colored circles at 3:00 in the positional firing rate map) but have no strong directional selectivity (the directional polar plot does not reveal strong direction variability). On the contrary, head direction cell firing depends on the heading of the animal (e.g., at 180°) independently of its location (no specific place is associated with increased firing). Finally, the discharge of grid cells resembles that of place cells but they have multiple, regularly spaced fields. Note that other cells exist that can combine these different properties (e.g., have both location and direction correlates).

relationships among places (Figure 3). The advantage of this dual format is that topological information is more rapidly acquired than metric information because metric encoding would largely rely on motion-related signals provided by repeated movements between places. However, direct empirical support for dissociation between topological and metric encoding of spatial information is still awaited.

Neural Bases of Spatial Cognitive Maps

Damage to almost any portion of the brain can result in impairments in spatial navigation. The question, however, is whether the brain is organized so that the two modes of spatial coding, egocentric and allocentric, are dissociated. This question is not trivial because it is not easy to disentangle the contribution of motion-related signals (which provide the basis for egocentric coding) to the allocentric coding of spatial information which likely involves movements in space. In spite of this difficulty, several studies have shown that the two systems may be distinct. Substantial evidence suggests that taxon navigation is subtended by basal forebrain structures such as the dorsal striatum, whereas locale navigation is subtended by temporal lobe structures including the hippocampus. Striatal lesions in rats impair the performance of egocentric localization tasks, while leaving unaffected the performance of equally difficult tasks that did not require egocentric localization. In contrast, hippocampal lesions produce place learning impairment in place navigation tasks, such as the water maze, but generally do not affect performance in egocentric tasks (e.g., cue navigation). Recent evidence for a functional dissociation between the effects of lesions of the striatum and lesions of the hippocampus confirm that these two brain regions are involved in egocentric coding processes (cue, guidance, and route strategies) and allocentric coding processes (place strategy), respectively.

The role of the hippocampus in spatial navigation based on a place strategy is further supported by the existence of cells that carry a spatial signal. Such cells are found in the hippocampus proper and closely related structures, and can roughly be classified as place cells, head direction cells, and the more recently discovered grid cells.

Place cells are hippocampal pyramidal cells whose firing is strongly correlated with the location of a freely moving rat in its environment. Head direction cells are primarily found in the postsubiculum, and the anterior and lateral dorsal thalamus. Their firing depends on the heading of the animal independently of its location. Finally, grid cells are found in the dorsocaudal medial entorhinal cortex. These cells are activated whenever the animal's position coincides with any vertex of a regular grid of equilateral triangles spanning the surface of the environment. Together, grid cells provide a directionally oriented, topographically organized neural map of the spatial environment (Figure 4). A common feature of all

these cell populations is that their firing is primarily controlled by salient visual cues and motion-related signals although other sensory cues can be used to maintain coherent activity. In addition to these populations of cells, the hippocampal formation contains neurons whose spatial signals are more complex. For example, cells in the subiculum have both location and direction correlates. It appears, therefore, that the hippocampus and related structures form a system where spatial information is processed and coherently organized and lies at the core of a widespread neural network concerned with the processing of spatial information. Although current evidence suggests that normal hippocampal function is required for place navigation, its exact role is still a matter for debate. In addition, it is not clear how the information processed in the hippocampus is passed onto other brain areas involved in spatial processing. A number of other cortical structures (e.g., entorhinal, retrosplenial, prefrontal, or parietal cortex) and subcortical structures such as several thalamic nuclei have been hypothesized to contribute to place navigation. The specific role of these structures in relation to the hippocampus remains to be elucidated. Nevertheless, recent findings suggest that the human hippocampus hosts cells that look very much like hippocampal place cells in the rat, thereby strengthening researchers' interest for a better understanding of the computational properties of these systems.

Conclusion

The hypothesis that the hippocampus is the locus of a cognitive map has resulted in a growing interest into the neural basis of animal spatial cognition. Current evidence suggests that rats are able to build spatial representations of their environment so as to navigate with efficiency. Data on the neural mechanisms potentially associated with the processing of spatial information are accumulating rapidly. We currently know many pieces of the puzzle, but are still far from understanding how they are put together. Nevertheless, neuropsychological studies of spatial cognition in the rat routinely provide new data likely to shed light on how spatial information is handled by the brain, and therefore to help us understand how complex human behaviors are implemented in neural activities.

See also: Magnetic Sense in Animal Navigation; Neural Coding of Spatial Representations; Parietal Cortex and Spatial Attention; Spatial Memory: Assessment in Animals; Spatial Cognition and Executive Function; Spatial Cognition; Synaptic Plasticity and Place Cell Formation.

Further Reading

Gallistel CR (1990) *The Organization of Learning*. Cambridge: MIT Press.

Hafting T, Fyhn M, Molden S, Moser M-B, and Moser EI (2005) Microstructure of a spatial map in the entorhinal cortex. *Nature* 436: 801–806.

Jeffery K (ed.) (2003) *The Neurobiology of Spatial Behaviour*. Oxford: Oxford University Press.

Muller RU (1996) A quarter of century of place cells. *Neuron* 17: 813–822.

O'Keefe J and Nadel L (1978) *Hippocampus as a Cognitive Map*. Oxford: Clarendon.

Poucet B (1993) Spatial cognitive maps in animals: New hypotheses on their structure and neural mechanisms. *Psychological Review* 100: 163–182.

Save E and Poucet B (2000) Involvement of the hippocampus and associative parietal cortex in the use of proximal and distal landmarks for navigation. *Behavioral Brain Research* 109: 195–206.

Taube JS (1998) Head direction cells and the neurophysiological basis for a sense of direction. *Progress in Neurobiology* 55: 225–256.

Thinus-Blanc C, Save E, and Poucet B (1998) Animal spatial cognition and exploration. In: Foreman N and Gillett R (eds.) *A Handbook of Spatial Research Paradigms and Methodologies*, pp. 59–86. Hove, UK: Psychology Press.

Tolman EC (1948) Cognitive maps in rats and men. *Psychological Review* 55: 189–208.

Wiener SI and Taube JS (eds.) (2005) *Head Direction Cells and the Neural Mechanisms of Spatial Orientation*. Cambridge: MIT Press.

Relevant Websites

http://www.bris.ac.uk – Neural Basis of Spatial Memory, Centre for Synaptic Plasticity, University of Bristol.

http://instruct1.cit.cornell.edu – Neuroethology of Rat Spatial Navigation: The Role of Place Cells, Academic Technologies

http://en.wikipedia.org – Place cell, Wikipedia.

Spatial Memory: Assessment in Animals

C M Bird and N Burgess, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Spatial memory supports retrieval of the locations of objects and places in the environment, an essential aspect of behavior in humans and other mobile animals. Foraging for food, returning to sites of storage or safety, and avoiding sites of danger within a complex and changing environment present a daily challenge to many species. The study of spatial memory in animals has also made important contributions to the understanding of human memory more generally, due to the straightforward transfer of spatial paradigms between species. Here we review several of the paradigms used and the insights they have provided. We start by focusing on paradigms developed to test rats, but also consider testing in other species, including humans. Many of these paradigms were developed to test hypotheses regarding the functions of brain regions associated with spatial memory, most notably the hippocampus, the results of which we comment upon in passing.

Is Spatial Memory Special?

In an influential review, Edward Tolman proposed in 1948 that rats are able to construct and store 'cognitive maps' of their environments. It had been observed that rats could learn to associate a location in space with a reward. Using a sunburst-shaped maze, Tolman and colleagues elegantly demonstrated that rats appeared to be able to make detours using maplike representations (see Figure 1). Latent learning of the layout of a maze, via exploration in the absence of externally provided rewards, could also be shown to improve navigation to subsequently rewarded locations.

Learning the locations of places within a cognitive map exemplified a type of cognitive memory, which contrasted with the learning of stimulus–response associations, the bedrock of classical models of animal learning. Perhaps unsurprisingly, many researchers were reluctant to accept that this reflected a qualitatively different process. One of the paradigms used to test whether rats remembered places, or simply responses, was the T-maze.

T-Mazes and 'Place' versus 'Response' Learning

A T-maze challenges the rat to choose either the left or right arm to collect a reward such as food (see Figure 2). The rat may learn to go to a specific location within the room, or to make a specific motor response at the choice point. However, if the rat approaches the choice point from the other direction in a 'probe trial,' the two behaviors will lead in opposite directions and so can be distinguished from each other. In such probe trials it is important neither to reward the animal nor violate its expectation of reward, if learning in the standard configuration is to continue unchanged. For this reason the animal might be removed from the maze after the choice point, or might be unrewarded if the experiment is run with a partial (i.e., intermittent) reinforcement schedule.

Using the T-maze, MG Packard and JL McGaugh have demonstrated a double dissociation between brain areas underpinning place and response learning as rats learn to approach a consistently baited arm from the start arm. In probe trials after 8 and 16 days of training, healthy rats shifted from approaching the 'place' associated with food after 8 days, to making the turn 'response' associated with food after 16 days. Injections of lidocaine to inactivate the hippocampus abolished place learning, while injections in the dorsal striatum (part of the basal ganglia) left place learning intact, but abolished the shift to turn responses. Thus, both place and response learning are present in parallel, with the former dominating behavior early in learning and requiring the hippocampus, and the latter dominating behavior after several trials and requiring the dorsal striatum.

Egocentric and Allocentric Representations of Location

Spatial locations can be represented 'egocentrically' in relation to the observer or 'allocentrically' in relation to environmental landmarks. Both types of representation appear necessary for normal behavior, with most aspects of perception, movement, and short-term memory associated with egocentric representations, and allocentric representations required for memory over longer durations or movements. In these terms, response learning would be egocentric while place learning would be allocentric.

Insight into the operation of these representations has come from recording the activity of individual

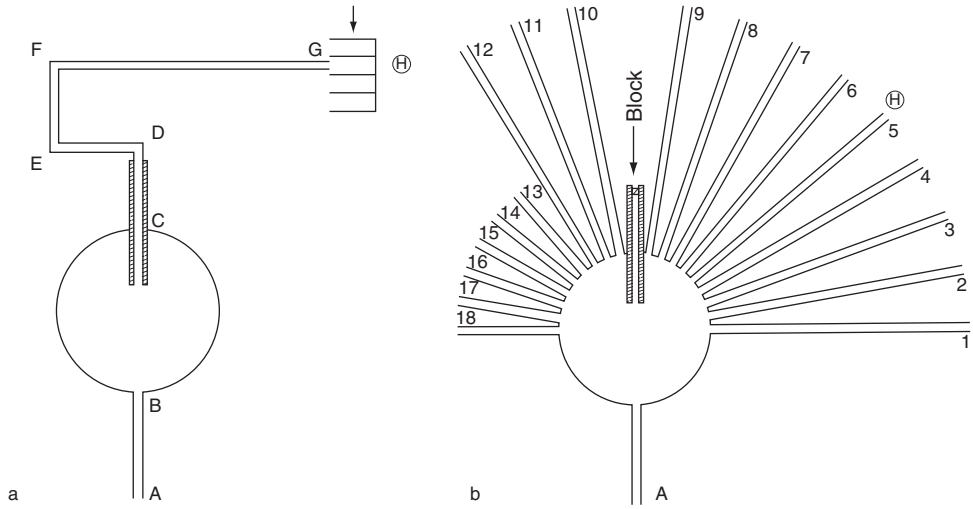

Figure 1 Sunburst maze, illustrating detour behavior. (a) Prior to testing, the rat learns to run via an elevated pathway to the goal location at position H. (b) At test, the elevated pathway is blocked and a number of straight pathways are opened. Most rats choose pathway 5, which leads directly to the goal location, displaying classic 'detour' behavior. However, in the original experiments the presence of a light at position H had left open the possibility that rats had learned only to run toward this. From Tolman EC, Ritchie BF, and Kalish D (1946) Studies in spatial learning. I: Orientation and the short-cut. *Journal of Experimental Pyschology* 36: 13–24.

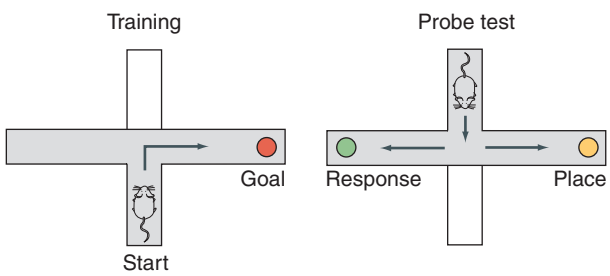

Figure 2 A T-maze, illustrating 'place' and 'response' behaviors. Rats are trained to retrieve food from one arm of a T-maze. The content of learning can be assessed by moving the starting arm to the other side of the maze on a probe test. The animal may enter the arm corresponding to the location of the reward during training (place strategy) or the arm corresponding to the turning response that was reinforced during training (response strategy). From Yin HH and Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* 7: 464–476.

neurons in behaving animals. Thus, for example, O'Keefe and colleagues have found the activity of neurons (place cells) in the hippocampi of freely moving rats to reflect an animal's location relative to the orientation cues and boundaries of its environment: this finding provided the impetus for O'Keefe and Nadel's proposal that the hippocampus supports a cognitive map of the environment. Equally, egocentric representations of the locations of objects relative to an animal's receptors and effectors in sensory and motor cortex, as well as representations in parietal and prefrontal cortices, reflect the short-term storage and manipulation of egocentric locations. In parallel with these developments, paradigms were developed to distinguish between the behavioral expression of these representations.

Radial-Arm Mazes: Dissociating Working, Reference, and Associative Memory

A classic paradigm for assessing spatial memory is the radial arm maze used by Olton and colleagues. In this test, a number of arms (e.g., eight) radiate outward from a central arena, extramaze cues are visible for orientation, and all arms are baited with food. Most healthy rats learn to collect the food without reentering a previously visited arm, and do not resort to obvious (nonspatial) strategies such as always taking the next arm to the left. Thus they are able to keep track of multiple locations in space so as not to revisit them, while rats with hippocampal lesions are severely impaired. This version of the task requires a 'win-shift' strategy: the rat must enter a different arm to get the next reward. In an alternative 'win-stay' version of the task, certain arms are always associated with reward within a trail, and rats acquire the habit of always visiting these arms. This type of learning is impaired by dorsal striatal lesions, again consistent with the dissociation between place and habit learning.

Despite the spatial nature of the standard task, it remains possible that rats use a nonspatial strategy to solve it. Thus, rats examine the entrances to the arms rather than simply moving directly between arms using distal cues for orientation, suggesting that more general deficits in working or associative

memory underpin the impairment shown by hippocampal rats. Accordingly, the task was adapted to distinguish between these possibilities. For example, long-term or reference memory and working memory can be dissociated by placing food rewards only in some of the arms. Long-term memory is required to learn which of the arms are never baited over several learning trials, whereas working memory is required to prevent reentry of previously visited arm within a single trail. Associative memory and spatial memory can be dissociated by making the arms distinct through the use of intramaze cues such as textured floor inserts. In 'spatial' trials, obvious fixed extramaze but not intramaze cues specify the rewarded arms, while the opposite is true in 'associative' trails. Rats appear to make use of both extramaze cues (or place learning, consistent with cognitive mapping) and intramaze cues (or cue learning, consistent with nonspatial associations). However, neurotoxic lesions of the hippocampus reveal greater impairment of place learning (encompassing both working memory and long-term memory), compared to cue learning, for which the effect was milder and restricted to working memory.

Recent advances in imaging neural activity have also been applied to these spatial questions. For example, expression of the immediate-early gene c-*fos* has been used to investigate which regions are active while rats perform the radial arm maze task. A critical issue here is the choice of control task – is there a task with sensory and motor demands similar to those of the radial arm maze, but without involving spatial learning? To address this issue, Aggleton, Brown, and colleagues compared c-*fos* expression in a group of test rats tested on a standard radial arm maze with a group of yoked control rats that only ran up and down one arm of a maze and received food rewards at the same frequency as experienced by the test rats. In another study, a test group learning the maze in a new room (with new extramaze cues) was compared to a group which had already learned it in the same room previously. Under these conditions, the test group would be expected to be learning more about the extramaze cues. In both studies, test rats showed significantly more c-*fos* expression in the hippocampus, as compared to control rats.

The Water Maze

Despite its successes, the radial arm maze also has some drawbacks. Behavior is constrained to the finite number of paths available (e.g., ruling out detours), while odor trails and other intramaze cues need to be carefully controlled and the desire of rats to explore can interfere with attempts to measure their memory for the target location. To address these issues, another paradigm, the water maze, has proved popular (see Figure 3).

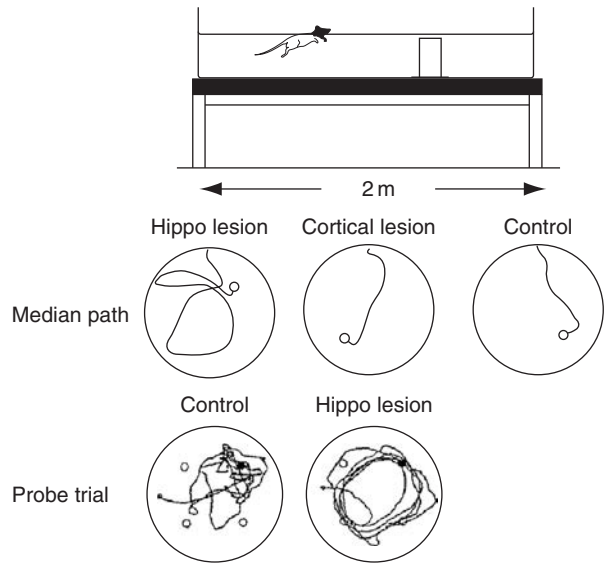

Figure 3 Water maze, for testing allocentric spatial memory. The rat or mouse must swim in a large circular tank of opaque water until it finds the escape platform, located just below the water surface. Healthy rodents learn to use extramaze cues to swiftly locate the platform, irrespective of their start position. Examples of typical paths swum by control rats, rats with cortical lesions, and rats with hippocampal lesions are shown. In probe trials, the platform is removed, and the search behavior is noted. Control rats concentrate their search in the area where the platform has been located on previous trails, whereas rats with hippocampal lesions tend to swim around the whole of the arena. From Morris RG, Garrud P, Rawlins JN, and O'Keefe J (1982) Place navigation impared in rats with hippocampal lessions. *Nature* 297 (5868): 681–683.

Hippocampal lesions dramatically impair performance on the water maze, although hippocampal rats can learn to locate the submerged platform if the starting position is held constant across sessions (suggesting intact egocentric memory) and can learn which of two differently colored visible platforms is safe (indicating that this form of cue learning is intact). They can also learn to find a submerged platform located at a fixed bearing from a single intramaze cue which is moved within the maze every four trials. MacDonald and White had rats learn to swim to a visible and a submerged platform. Damage to the hippocampal system impaired navigation to the submerged platform, but not to the visible platform, even when it moved within the maze. By contrast, damage to the dorsal striatum resulted in rats that could navigate to both platforms, but when the visible platform moved, they swam to the submerged platform. Thus, as with place and response learning in the T-maze, allocentric memory depended on the

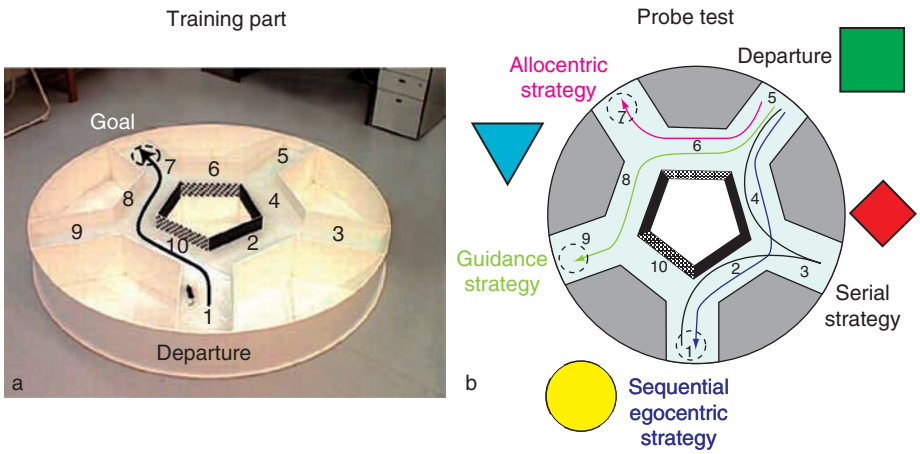

Figure 4 Star maze, showing multiple potential strategies. (a) All alleys are partially filled with water. The animals were placed in alley 1 and learned to find the hidden platform located in alley 7 by the shortest path (indicated with an arrow). To learn this trajectory, the animal could use different strategies that can be dissociated and identified during the probe test. (b) To distinguish which strategy the animal used to solve the task, the departure point was changed to alley 5. The positions of the extramaze and intramaze cues remained constant during the entire training session. When departing from alley 5, an animal had to choose between using the extramaze visual cues (5–6–7 trajectory, corresponding to an allocentric strategy), following the intramaze cues (5–6–8–9 trajectory, corresponding to guidance), executing a sequence of left–right–left body movements (5–4–2–1 trajectory, indicated as sequential–egocentric strategy), or visiting all radiating alleys successively (5–4–3–2–1 trajectory, indicated as serial strategy). Other behaviors were considered as representing no clear strategy. From Rondi-Reig L, Petit GH, Tobin C, et al. (2006) Impaired sequential egocentric and allocentric memories in forebrain-specific-NMDA receptor knock-out mice during a new task dissociating strategies of navigation. *Journal of Neuroscience* 26: 4071–4081.

hippocampus, while egocentric spatial memory depended on the dorsal striatum.

An ingenious combination of the radial arm and water mazes, the star maze, has been recently developed by Rondi-Reig and colleagues for testing spatial memory in mice (see Figure 4). The construction of the star maze allows use of the probe trials, to dissociate a number of possible strategies. After 20 days of training, healthy mice tended to use one of two strategies, navigating according to allocentric cues or executing a sequence of body turns. Mutant mice with a gene knockout targeting the *N*-methyl-D-aspartate (NMDA) receptors in the CA1 field of the hippocampus failed to learn the allocentric strategy, could learn an egocentric strategy when it involved just one turn, but were unable to learn sequences of two or more turns.

Spatial Manipulations of the Environment

Most of the aforementioned tasks involve static environments, although the animal's start position may vary. However, a powerful way to probe the representation that an animal is using is to change the positions of the potential cues and observe the effect upon behavior. For example, the positions of distal cues can be rotated relative to the local environment. Using a four-arm radial maze, it was shown that rats could learn to use six moving distal cues (that rotated *en masse*) to locate the rewarded arm. When hippocampal place cells were recorded in this task, they also rotated, showing a good correspondence to behavior, even on error trials. Aside from its overall orientation, the place cell representation of location seems to reflect location relative to extended barriers present in the environment: firing locations move and deform in order to maintain fixed distance from the nearer barriers following manipulations where the environment is deformed in shape. In contrast, firing locations do not move or reform when discrete intramaze cues are manipulated. By using a water maze with transparent sides in a room containing extramaze landmarks, Maurer and Derivaz showed that rats used the angular separation and apparent dimensions of landmarks to locate the escape platform, and that the distance of the platform from the wall was also a strong cue to its location.

Path Integration and Reorientation Paradigms

Path integration, sometimes known as dead reckoning, is the ability to keep track of a start location on the basis of subsequent self-motion information. Thus, a cumulative estimate of the distance and direction moved during foraging can allow an animal to return

to the nest. Desert ants provide some of the most impressive examples of this, aided by environmental orientation cues such as the polarization of ambient sunlight. In mammals, path integration may be used by animals foraging from the nest during darkness, but is severely constrained by the rapid accumulation of angular errors in the absence of orientation cues. Nevertheless, during navigation through an environment containing detectable landmarks, path integration provides an important moment-to-moment continuity to the sense of self-location.

Reorientation paradigms are used to investigate the mechanisms by which animals use environmental cues to reorient themselves in the absence of a reliable path integration signal. Figure 5 illustrates a typical setup. Following disorientation by rotation in the absence of sensory cues, the animal searches for a goal location.

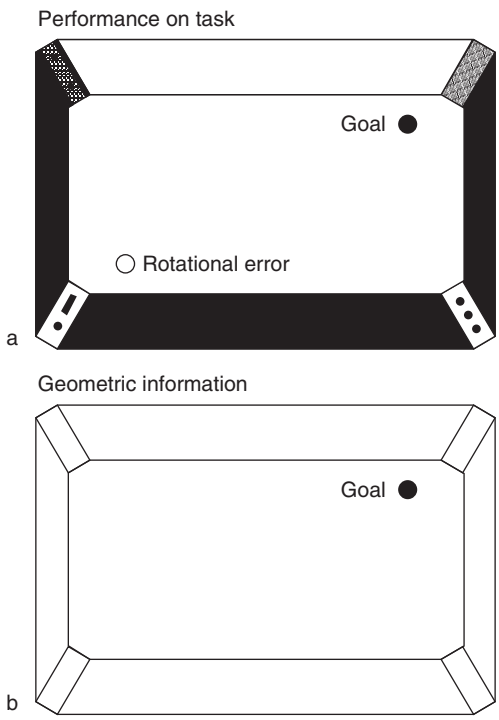

Figure 5 Reorientation testing. (a) The task in a rectangular arena as seen from above. In attempting to relocate the goal after disorientation, the rotational error is frequently committed. This is the location at 180° rotation through the center from the correct location. However, featural information (landmarks in the corners) specifies which is the correct location. (b) The geometric information is contained in the broad shape of the arena. The featural information is what is not shown: patterns on the panels, different brightnesses of walls, smells in the corner, and the like. In using a 'map' containing only geometric information, the goal and the rotational error cannot be distinguished. From Cheng K and Newcombe NS (2005) Is there a geometric module for spatial orientation? Squaring theory and evidence. *Psychonomic Bulletin and Review* 12: 1–23.

Reorientation paradigms have been used in a wide variety of animals, including rats, chicks, fish, monkeys, and humans. All of these animals are able to make use of the geometry of an environment to reorient. However, they differ in the extent to which they use featural information (e.g., landmarks, patterns and other information that can specify location). For example, Cheng and colleagues found that rats tend to make little use of featural cues, while chicks use them very effectively. When geometric cues and featural cues are pitted in opposition, Vallortigara and colleagues found that chicks can make use of both types of information, with featural information processed in the left hemisphere via the right eye and geometric information processed in the right hemisphere via the left eye. In humans, preverbal toddlers also appear to use only geometry, at least in small enclosures, while adults can also make use of features in most situations.

Exploration and Novelty

Most of the paradigms discussed so far have involved animals searching for a target – either a place rewarded by food, or a means of escape from an unpleasant environment, such as water. Many similar paradigms involve avoidance of a location associated with an unpleasant stimulus such as an electric shock to the foot. An alternative to these paradigms comes from some animals' spontaneous exploration of novelty in the environment. Eacott and Norman investigated the sensitivity of rats to changes in objects, their spatial locations, and the contexts in which they were encountered (altering the color and texture of the arena). The rats' spontaneous exploratory behavior demonstrated sensitivity to all three changes. Rats with lesions to the hippocampal system were insensitive to novel configurations of object locations in different contexts.

Primate Studies: Object-in-Place Tasks

The preceding paradigms were largely developed to test spatial memory in rats. They made use of the fact that rats forage for food, and the studies often employed designs that paralleled search behaviors in the wild. Convenient testing of spatial memory in primates requires more constrained paradigms. For example, a test tray containing food wells might be placed in front of a monkey, which would be required to remember the location of a rewarded well, or the object covering the baited well. Some more recent studies have employed computer touch screens, enabling the background to be changed trail by trail,

providing a rich selection of contexts within which locations or features had to be remembered (object-in-place tasks).

Single-unit recording from awake primates performing spatial memory tasks has provided valuable insights into how and where spatial information is represented in the brain. A popular paradigm is to train the monkey to fixate at a point while one or more locations are cued. After a delay, during which fixation is maintained, the monkey indicates the cued position with either an eye movement or an arm reach. Studies have revealed neural populations in regions of the lateral frontal lobe and posterior parietal lobe that respond during the delay, suggesting that they are encoding the target location, or at least the intention to make a motor response to that location. Careful manipulations of the target location and the position of the eyes, head, or limbs with respect to the rest of the body have revealed that the delay activity can reflect the intention to make a particular movement to the target location, rather than simply attending to the location. These manipulations can also reveal how the locations are encoded. Andersen and colleagues found neurons in the intraparietal sulcus that are tuned to the locations of targets with respect to several different coordinate frames. For example, in area 7a, the response to stimuli presented in a neuron's retinotopic receptive field can be modulated by the orientation of the eyes in the head, the head on the body, or even by the orientation of the monkey's body within the testing room. These multiply-modulated responses are consistent with translation of information between retinotopic, head-centered, body-centered, and world-centered reference frames. This type of translation is required to determine where a visual stimulus is located relative to an effector such as hand or mouth that might be required to act on it: this is consistent with parietal involvement in sensorimotor integration. In addition, egocentric to allocentric translation would be required for integration of information between the largely egocentric parietal representations and the largely allocentric representations in anterior parieto-occipital sulcus, parahippocampus, and hippocampus.

Recently, a study explicitly contrasted the performance of young humans with that of other species of great apes on a task that shares similarities with object-in-place tasks (Figure 6). The main question was to ascertain whether participants would use a spatial strategy or a landmark strategy to search for a reward. The results were striking. All of the great apes and prelinguistic children (aged 12–13 months) favored the spatial strategy. In contrast, 3-year-old children favored the landmark strategy. These results suggest that the common ancestor to all great apes

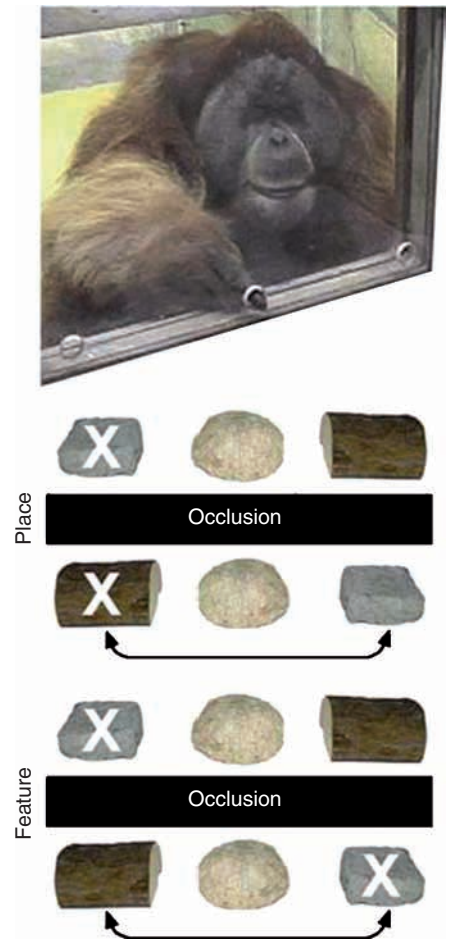

Figure 6 Testing spatial and featural representations in the Hominidae. (Top) Adult orangutan male performing the task. (Bottom) Three distinct objects are placed in front of the participant and a reward is hidden under one. During occlusion, the positions of two of the three objects are swapped. The X indicates the hidden reward before and after occlusion. Experimental conditions consist of place and feature. For place, two containers are swapped while the reward remains in its original location. For feature, two containers are swapped but the reward moves with its container. From Haun DB, Call J, Janzen G, et al. (2006) Evolutionary psychology of spatial representations in the Hominidae. *Current Biology* 16: 1736–1740.

would have also favored a spatial strategy when searching for food. However, at around the time of language development, humans switch strategies and associate reward more with specific objects rather than the location where reward was previously obtained.

A common factor to all of these object-in-place paradigms is that they can be solved using egocentric strategies. As reviewed earlier, the brain systems underpinning egocentric and allocentric spatial memory are not the same. This may account for the fact that some primate studies show little effect of hippocampal damage on spatial memory. In order to tease apart egocentric and allocentric representations, it is

necessary to move either the testee or the environmental landmarks between the initial encoding and subsequent test. Accordingly, a recent study involved monkeys foraging in an open-field arena for food located in hidden 'wells.' When the food wells were marked by local visual cues, monkeys with hippocampal lesions had no difficulties solving the task. However, when distal cues were the only means to locate the food wells, hippocampal lesions did impair performance. As with studies in freely moving rats, recordings from individual neurons in primates during motion through their environment have also revealed place cells. Other types of cells, termed 'spatial view' cells, have also been reported to respond whenever a monkey views a specific location in space, and may extend the function of place cells, to the representation of places remote from the animal's current location.

Translation to Assessment of Human Spatial Memory

Traditionally, human spatial memory tests have employed a two-dimensional format, involving, for example, remembering the locations of unrelated objects on a tabletop. Others have involved the learning by trial and error of a path across a regular grid of identical bolts. Similar to the majority of nonhuman primate tasks described previously, such paradigms can be solved by an egocentric strategy. Accordingly, some other tests have looked to the allocentric paradigms developed in rats such as the water maze. Recently, virtual reality (VR) has enabled spatial exploration and search behavior in specially designed, large-scale environments to be studied, and has allowed several paradigms developed for animals to be translated to humans.

Both the radial arm maze and water maze have been adapted for use with humans, including patients with brain damage. For example, in a version of the water maze, patients learn the position of a sensor hidden under the carpet. Patients with parahippocampal or hippocampal damage show impairments in various aspects of this task. Manipulations of the experimental configuration between presentation and test have been used in humans as well as in other animals. Wang and Simons had participants view an array of objects and subsequently report which object had moved. Between presentation and test, the participant might move to another viewpoint and the array might be rotated so that the test view is consistent either with egocentric perceptual representations or with egocentric representations of location that are automatically updated by self-motion. A further study by Burgess and colleagues also manipulated an environmental landmark to measure the contribution of allocentric representations of location. The performance benefit from consistency at test with all three types of representation indicates the parallel presence of (1) egocentric perceptual representations, (2) self-motion related representations, and (3) allocentric representations. These three representations all underlie our ability to keep track of locations as we move through the environment.

VR allows good experimental control and monitoring of responses in simulated navigation and can also allow participants to keep their head still, albeit at the expense of the vestibular component of navigation. One use for this is functional brain imaging. Using a VR version of an eight-arm radial maze, Bohbot, Iaria, and colleagues found that healthy adults who used an allocentric strategy (detected by increased errors when distal cues are removed) showed increased right hippocampal activation, while patients with lesions to this area either did not use this strategy or, if they did, performed very poorly. By contrast, healthy adults who used a nonspatial strategy activated parts of the dorsal striatum. Similar findings were obtained when Hartley, Spiers, Burgess, and colleagues investigated wayfinding compared to route-following in a virtual town in imaging and neuropsychological experiments, again paralleling the findings on place and response learning in rats. VR navigation has also allowed recording from individual neurons in epilepsy patients requiring intracranial electrodes for localization of the seizure focus. These studies showed responses resembling place cells, clustered in hippocampus, as well as spatial view cells and 'goal' cells responding to the current target of navigation.

VR also permits experimental manipulations that are impractical to carry out in reality. Hartley and colleagues had participants see an object in a rectangular arena surrounded by distant landmarks, and subsequently report where they thought it had been. On some trials, the dimensions of the arena were varied between presentation and test. The effects of these spatial distortions on participants' responses were consistent with a model of spatial memory based on storing hippocampal place cell representations. King and colleagues used the ability to instantaneously change viewpoint within VR to assess the use of allocentric representations of location in the absence of motion-related spatial updating. Participants viewed objects located within a 'town square' and subsequently indicated which of several copies of each object was in the correct location. While healthy volunteers could perform the task when tested from the same viewpoint as presentation, or from a suddenly imposed new viewpoint, a patient with hippocampal damage was severely impaired when tested from the new viewpoint.

Overall, these findings demonstrate both the power of spatial manipulations for testing models of memory and the cross-species and cross-methodology convergence of understanding in this field.

Conclusions

Spatial memory is essential for adaptive behavior in humans and many other animals. Studying spatial memory in animals has provided enormous insights into the processes and neural mechanisms of memory and provides a starting point for the understanding of human memory more generally. Our knowledge about how these systems interact has proceeded hand in hand with the development of different tests to tease them apart. Advances in neuroscientific techniques to assess the involvement of different brain areas in aspects of task performance have cross-validated and extended older findings. Similarly, new technologies such as VR have allowed the convenient extension of some paradigms from animals to humans.

See also: Neural Coding of Spatial Representations; Parietal Cortex and Spatial Attention; Procedural Learning: Striatum; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Spatial Cognition and Executive Function; Spatial Transformations for Eye–Hand Coordination; Spatial Cognition; Spatial Cognitive Maps.

Further Reading

Burgess N, Maguire E, and O'Keefe J (2002) The human hippocampus and spatial and episodic memory. *Neuron* 35: 625–641.

Burgess N (2006) Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences* 10(12): 551–557.

Cheng K and Newcombe NS (2005) Is there a geometric module for spatial orientation? Squaring theory and evidence. *Psychonomic Bulletin and Review* 12: 1–23.

Cohen YE and Andersen RA (2002) A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience* 3: 553–562.

Haun DB, Call J, Janzen G, et al. (2006) Evolutionary psychology of spatial representations in the Hominidae. *Current Biology* 16: 1736–1740.

Jarrard LE (1993) On the role of the hippocampus in learning and memory in the rat. *Behavioral and Neural Biology* 60: 9–26.

Lavenex PB, Amaral DG, and Lavenex P (2006) Hippocampal lesion prevents spatial relational learning in adult macaque monkeys. *Journal of Neuroscience* 26: 4546–4558.

Maurer R and Derivaz V (2000) Rats in a transparent Morris water maze use elemental and configural geometry of landmarks as well as distance to the pool wall. *Spatial Cognition and Computation* 2: 135–156.

Morris R (2007) Theories of hippocampal function. In: Andersen P, Morris R, Amaral D, et al. (eds.) *The Hippocampus Book*, pp. 581–713. New York: Oxford University Press.

Morris RG, Garrud P, Rawlins JN, and O'Keefe J (1982) Place navigation imparied in rats with hippocampal lession. *Nature* 297(5868): 681–683.

O'Keefe J and Nadel L (1978) *The Hippocampus as a Cognitive Map*. Oxford: Oxford University Press.

Rondi-Reig L, Petit GH, Tobin C, et al. (2006) Impaired sequential egocentric and allocentric memories in forebrain-specific-NMDA receptor knock-out mice during a new task dissociating strategies of navigation. *Journal of Neuroscience* 26: 4071–4081.

Tolman EC, Ritchie BF, and Kalish D (1946) Studies in spatial learning. I: Orientation and short-cut. *Journal of Experimental Psychology* 36: 13–24.

Tommasi L and Vallortigara G (2001) Encoding of geometric and landmark information in the left and right hemispheres of the avian brain. *Behavioral Neuroscience* 115: 602–613.

Yin HH and Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* 7: 464–476.

Spatial Orientation: Our Whole-Body Motion and Orientation Sense

D Merfeld, Massachusetts Eye & Ear Infirmary, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Our sense of body motion and orientation can be divided into two overlapping and interacting systems – one that senses head motion and head orientation with respect to the external world and a second that senses relative motion and orientation of the various body segments. This article focuses on our sense of motion and orientation with respect to the external world. This sense has nine degrees of freedom. It includes our sense of motion – three dimensions of rotational motion and three dimensions of linear motion. It also includes our sense of three-dimensional orientation, which we explicitly define as the combination of heading direction (i.e., which way am I facing? or yaw orientation) as well as tilt of the head with respect to gravity (e.g., pitch and roll with respect to the normal upright posture).

Although somatosensory cues provide a substantial influence on our sense of motion and orientation, the predominant contributors to our sense of motion and orientation are the vestibular and visual systems. This article focuses on the contributions of the vestibular system because this sensory system is often overlooked, compared to the more commonly known sensory systems. The vestibular system is made up of specialized sensory end organs dedicated to measuring head motion and orientation. In brief, the vestibular system in each inner ear consists of three semicircular canals, which sense three dimensions of head rotation, and two otolith organs, which sense three dimensions of head linear acceleration and gravity. The anatomy and physiology of the vestibular end organs are discussed very briefly next.

In addition to perceptual contributions, the vestibular system makes crucial contributions to several reflexive motor responses. These include (1) reflexive eye movements, called vestibuloocular reflexes, that improve visual acuity by helping stabilize images on the retina; (2) reflexive cardiac and respiratory responses, called vestibulo-autonomic reflexes, that help stabilize autonomic tone; and (3) reflexive postural responses, called vestibulo-spinal reflexes, that help stabilize our head and body with respect to gravity, assisting us as we walk and stand upright.

Although the vestibular system makes critical contributions to the forementioned reflexive responses, the vestibular system is also a sensory system that contributes to our perception of motion and orientation. This sense is often neglected relative to the commonly known five classical senses. This may be due to the inconspicuous nature of vestibular function, which almost always operates at an automated level. Patients suffering from rapid-onset vestibular disorders often become acutely aware of the fundamental contributions of the vestibular system to perception only when the vestibular system fails. The importance of the vestibular system is also emphasized by the fact that it appeared very early in evolutionary terms. The fossil record shows the presence of morphologically distinct vertical semicircular canals – not unlike mammalian vertical semicircular canals – dating back 400 million years.

Because the vestibular system is often overlooked as a sensory system, we compare the principal characteristics of the vestibular system to the principal characteristics shared by other sensory systems. First, as defined by Johannes Müller, physical stimuli activate specific nerve fibers that connect to the central nervous system to yield specific sensations – a concept referred to as specific sense energy. As a related concept, detailed in the *Principles of Neuroscience*, all sensory systems share "three common steps: (1) a *physical stimulus*, (2) a *chain of events* transforming the stimulus into nerve impulses, and (3) a response to this signal in the form of a *perception* or conscious experience of the sensation" (p. 412, 4th edn.). The relevant specific sense energy is discussed for the canals in the section titled 'Specific sense energy – semicircular canals' and for the otolith organs in the section titled 'Specific sense energy – otolith organs.' Second, all sensory systems mediate four stimulus characteristics – modality, location, intensity, and timing – that can be quantitatively correlated with perception. The relevant stimulus characteristics of the vestibular system are discussed in the section titled 'Vestibular senses information coding.' Because rotation yields changes in orientation, physical rotation and physical orientation are directly coupled to one another. If any observer knows the orientation of the head at some time point (time A) and also knows the head's complete rotation history between time A and a second time point (time B), we can correctly calculate its orientation at any time between times A and B). On the other hand, our perceptions of motion and orientation can, unlike true physical motion, be inconsistent.

Such inconsistencies are illustrated in the next section, which also provides compelling evidence for the contributions of the visual and vestibular systems to our sense of motion and orientation.

Example

Imagine that a human subject is upright with respect to gravity and looks at the inside of a sphere that is rotating in roll about an Earth-horizontal axis (Figure 1(a)). At first, the subject reports that he or she is stationary and that the visual display is rotating; both perceptions are veridical. However, if the subject continues to observe the rotating visual display for just a little longer (~10 s or more), he or she typically begins to perceive a sensation of self-rotation in the direction opposite the rotation of the visual display (Figure 1(b)). This illusory perception of self-motion evoked by visual motion is called vection, or more specifically rotational vection, and it demonstrates the visual contributions to our sense of self-rotation. Similarly, if subjects view a visual scene that is translating (e.g., the train next to theirs) while they are stationary, this often induces a perception of linear motion called linear vection, or, if subjects view a stationary display that is tilted (e.g., a tilted room), this can induce an illusion of self-tilt.

Paradoxically, subjects experiencing rotational vection, like that just described, seldom report that they are tumbling head over heals as they would if they truly were rotating to the extent that they perceive. In fact, subjects typically experience a simultaneous illusory sensation of tilt that gradually builds up to some relatively constant level (Figure 1(c)). These perceptions are often described as a sensation of motion without getting anywhere. These sensations are contradictory because if the perceived rotational motion were real it would yield continuously varying changes in perceived tilt, not a stable perception of some constant amount of tilt. This demonstrates that our sense of motion and orientation is not limited to perceptions that are consistent with physically possible motions and orientations.

This example can also be used to illustrate the role played by the vestibular system in our sense of motion and orientation. First, patients who have severe and complete damage in the vestibular periphery often report greater vection than normal subjects. Second, astronauts report greater vection when weightless than when on Earth. Third, astronauts experiencing rotational vection in weightlessness report a head-over-heals tumbling sensation, which is absent on Earth. This shows that the body's detection of gravity, which is measured by the otolith organs, inhibits the head-over-heels tumbling sensation here on Earth. Taken together, these findings show that the vestibular system makes a substantial contribution to our sense of motion and orientation. This is not surprising because the vestibular system is specialized to sense head motion and head tilt with respect to gravity.

Specific Sense Energy – Semicircular Canals

Like most sensory receptors, the semicircular canals respond primarily to a single physical stimulus, which for the semicircular canals is rotational motion. More specifically, the physical stimulus is angular acceleration, although the fluid mechanics of the canal yield a mechanical integration such that the neural response

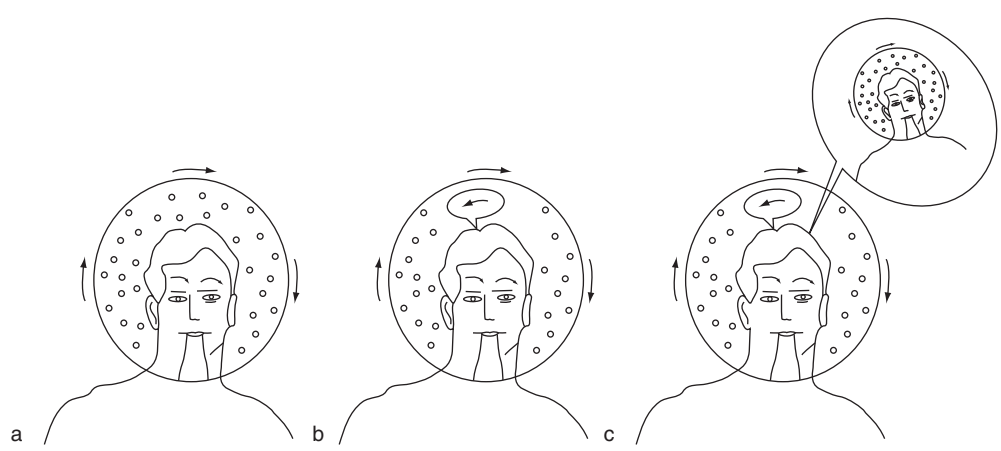

Figure 1 Roll vection: (a) sketch of a subject viewing a visual display rotating in roll; (b) subject beginning to experience vection (a sense of self-rotation in the direction opposite the rotation of the visual display); (c) roll vection is often accompanied by an illusion of roll tilt that is induced by the perceived roll rotation. For demonstration purposes, the visual display is shown as being transparent. Initially, subjects correctly sense that they are stationary and the visual display is rotating. In (c), the direction of the perceived tilt is consistent with the tilt direction that would have occurred if the subject were truly rotating in roll.

in first-order vestibular afferent neurons represents angular velocity over a broad range of physiological frequencies. A brief description of the physiological chain of events leading to peripheral transduction in the semicircular canal follows: (1) angular acceleration of the head yields fluid movement in the canals with respect to the head, which (2) deflects the cupula, which (3) deflects hair cell cilia, which (4) gates mechanically sensitive ion channels, which (5) alters the receptor potentials, which (6) alters the neurotransmitter release rate, which (7) alters the firing rate of afferent neurons innervating the canal.

Finally, there is no question that we perceive rotation. Skeptics can be rotated on a rotating chair (e.g., bar stool) with their eyes closed to show them that they perceive rotation even in the absence of cues providing a visual reference. Although the actual motion thresholds in the dark depend on the frequency content of the motion, at physiological frequencies, many normal subjects perceive angular velocities above 1–2 $^\circ s^{-1}$ and angular accelerations above approximately 0.2–1 $^\circ s^{-2}$. Consistent with Müller's notion of specific sense energy, a similar sense of rotation is evoked when the neurons innervating the canals are activated via electrical stimulation.

The perception of rotation about an Earth-vertical rotation axis has dynamic characteristics that mimic a high-pass filter (Figure 2); this is expected because the first-order afferent neurons from the canals are known to have high-pass response dynamics when the afferent response is compared to angular velocity stimuli. More specifically, at frequencies below approximately 0.025 Hz, little or no rotation is perceived. As the stimulation frequency increases, so does the perceived rotation. At moderately high frequencies (between 0.2 and 0.8 Hz), the sensation of rotation appears to plateau at a level that slightly underestimates the rotation.

That the semicircular canals are a predominant contributor to rotational sensation is demonstrated by the fact that patients in the dark with complete bilateral absence of vestibular function do not perceive rotational motion that would be detected easily by normal subjects. These patients do maintain the ability to perceive rotation in the light, when visual cues are available, especially at lower frequencies. In fact, it has been shown that signals from the visual periphery indicating visual motion (as opposed to foveal visual motion cues) converge with the afferent canal neurons in the vestibular nuclei, which is in the brain stem one synapse from the vestibular periphery.

The semicircular canals sense head rotation and evoke a perception of rotation. (Although we can argue that the perceived quantity is angular velocity, whereas the physical stimuli is actually angular acceleration, which is then mechanically integrated by the end organ such that the neural response most closely relates to angular velocity, this is not a critical distinction.) Hence, the dimensionality of the physical stimulus and the perceived quantity are the same. This direct relationship between the physical stimuli and the perceived quantity is interesting; no other sensory system has a more direct relationship between the perception and the primary physical stimulus. For example, the dimensionality of the perception evoked by physical stimulation (hearing, sight, smell, taste, and touch) and the actual physical stimulation (pressure waves, photons, chemical, and pressure) is not identical for any of the commonly known sensory systems.

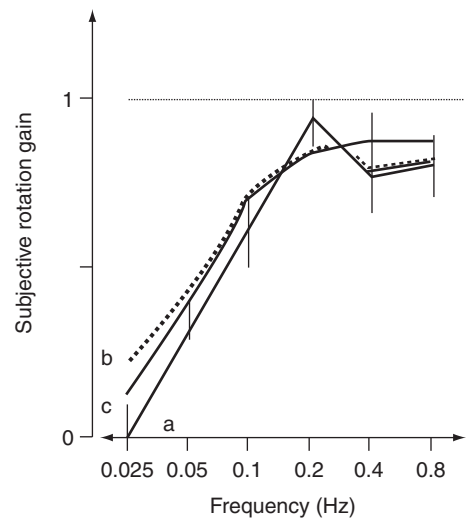

Figure 2 Magnitude of yaw rotation perception as a function of rotation frequency during sinusoidal rotations of the whole body. The three curves represent three different assays of perceived rotation, each of which yields a very similar measure of perceived rotation. Trace a shows perceived rotation in complete darkness; trace b shows perceived rotation when a single head-stationary object was viewed; and trace c shows perceived rotation when a single object rotated at twice the angular velocity of the subject. From Mergner T and Becker W (1990) Perception of horizontal self-rotation: multisensory and cognitive aspects. In: Warren R and Wertheim A (eds.) *Perception & Control of Self-Motion*, pp. 219–263. Hillsdale, NJ: Lawrence Erlbaum.

Specific Sense Energy – Otolith Organs

While most sensory receptors respond primarily to a single physical stimulus, the otolith organs, which sense both gravity and linear acceleration, provide an interesting exception. All linear accelerometers, including the otolith organs, respond equivalently both to gravity and linear acceleration, which is consistent with Einstein's equivalence principle which posits that linear acceleration and gravity have equivalent effects.

While probably important only to aficionados, it is interesting to note that one can define the physical stimulus to the otolith organs as the vector difference between gravity and linear acceleration. The corresponding force is called the gravito-inertial force. Under this definition, one could consider gravito-inertial force as the operating stimulus, which would mean that otolith organs, like the majority of sensory systems, respond primarily to a single stimulus. However, gravity and linear acceleration are different physical stimuli from a perceptual viewpoint. Nonetheless, they are indistinguishable in their effects from a physical viewpoint. There is no way, without separate information, to dissect the gravito-inertial force into portions due to each source. This is not an axiom of physics, but an inference from the observation that the gravitational and inertial masses (defined by Newton's law of gravitation and Newton's second law, respectively) have been measured to be equal to a very high degree of precision. Einstein made the assumption of perfect equality explicit in his equivalence principle, and built the theory of general relativity upon this assumption.

The physiological chain of events that yield neural responses in the otolith organs is similar to the cascade previously described for the semicircular canals. A brief description of the chain of events leading to peripheral transduction in the otolith organs follows: (1) linear acceleration of the head and/or the presence of a gravitational field deflects a membranous gel layer that is embedded with calcium carbonate crystals called otoconia, which (2) deflects hair cell cilia, which (3) gates mechanically sensitive ion channels, which (4) alters the receptor potentials, which (5) alters neurotransmitter release rate, which (6) alters the firing rate of afferent neurons innervating the otolith organ.

Because the otolith organs respond equally to two classically distinct forms of stimulation – linear acceleration and gravity – two distinct 'perceptual' responses are evoked by otolith stimulation. The issue of how this sensory ambiguity is resolved by the nervous system is discussed elsewhere.

The otolith organs contribute to tilt perception – the perception of the orientation of the head with respect to gravity. This sensation is both stable and relatively precise. Most young normal subjects can – even in a dark room – correctly estimate when tilted from upright by as little as 1–2° or so. Furthermore, for static tilts, subjective indications of perceived tilt more or less accurately reflect the actual physical tilt for tilts up to 90° (Figure 3), although consistent overestimates and underestimates of tilt angle are reported. If the subjects are tilted dynamically using sinusoidal stimuli, the subjective tilt responses remain somewhat constant across the frequency range of 0.005–0.8 Hz, although subjective tilt was accurately reported for high frequencies and somewhat overestimated at low frequencies (Figure 4). Because of the equivalence of gravity and linear acceleration, subjects report an illusion of tilt in the dark when they are dynamically accelerated using sinusoidal stimuli, even in the absence of actual physical tilt. This illusion is large at low frequencies (<0.05 Hz) before declining precipitously, such that almost no tilt illusion is experienced at frequencies of 0.2 Hz and higher (Figure 4).

The otolith organs also contribute to translation perception. In the absence of vision, normal subjects perceive linear motion having linear accelerations as small as approximately 0.05–0.10 m s^{-2} (510 mG) and linear velocities as small as approximately 0.02 ms^{-1} (<0.05 mph). In addition, when humans are exposed to steps of linear acceleration, the time that it takes to detect the direction of the motion (i.e., Is the motion to the left or right?) decreases as the magnitude of the linear acceleration increases (Figure 5).

Vestibular Senses Information Coding

As stated in the *Principles of Neuroscience*, "all sensory systems convey four basic types of information when stimulated – *modality*, *location*, *intensity*, and *timing*. Together these four elementary attributes yield sensation" (p. 412, 4th edn.). We look at vestibular function here as it relates to our sense of motion and orientation from the perspective provided by these four basic sensory system characteristics.

Generally, modality defines the general class of stimulus. As already discussed, the vestibular system has three modalities because it responds to rotational motion, linear acceleration, and gravity – with three dimensions for each of the three modalities. More specifically, the semicircular canals respond to rotation, with the nearly orthogonal canals sensing rotation in each of the three spatial dimensions, whereas the otolith organs respond to gravity in each of three dimensions and respond to linear acceleration in each of three dimensions.

Generally, location refers to the spatial distribution of the energy spectrum for each individual modality. For example, for touch and vision, the spatial distribution of the afferent neurons encodes the spatial location of the object (vision) or point of body contact (touch). For hearing, the spatial distribution of the afferent neurons represents the frequency of the acoustic stimulation. For the vestibular system, the spatial distribution of the afferent neurons represents the direction of motion or the direction of gravity. To use a specific example, neurons innervating the lateral canals respond nearly exclusively to yaw head

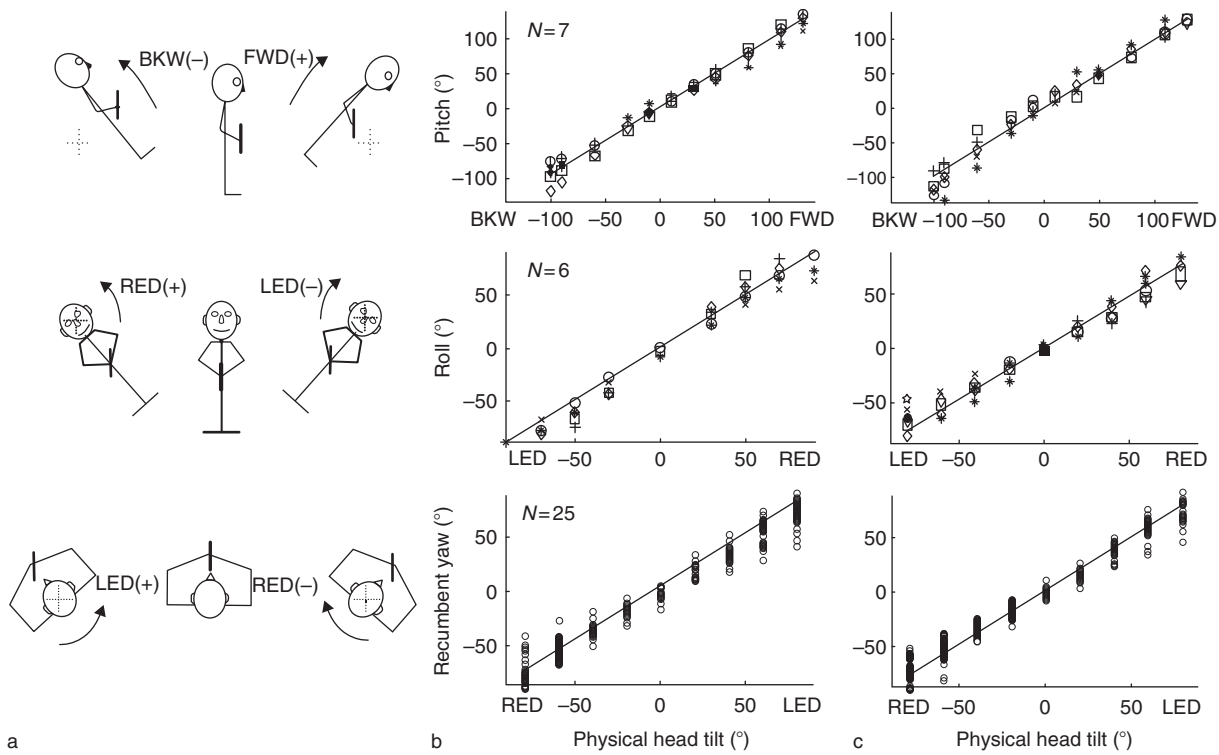

Figure 3 Tilt perception for static pitch (top row), roll (middle row), and yaw (bottom row) tilts: (a) sketches defining pitch, roll, and yaw and associated conventions; (b) tilt perception provided by subjects using a handheld haptic device; (c) tilt perception provided by subjects verbally. To first-order, perceived tilt accurately reflects the actual tilt for all three tilt directions, although roll tilts were slightly overestimated and yaw tilts were slightly underestimated. BKW, backward; FWD, forward; LED, left ear down; RED, right ear down. From Bortolami SB, Pierobon A, Dizio P, and Lackner JR (2006) Localization of the subjective vertical during roll, pitch, and recumbent yaw body tilt. *Experimental Brain Research* 173: 364–373.

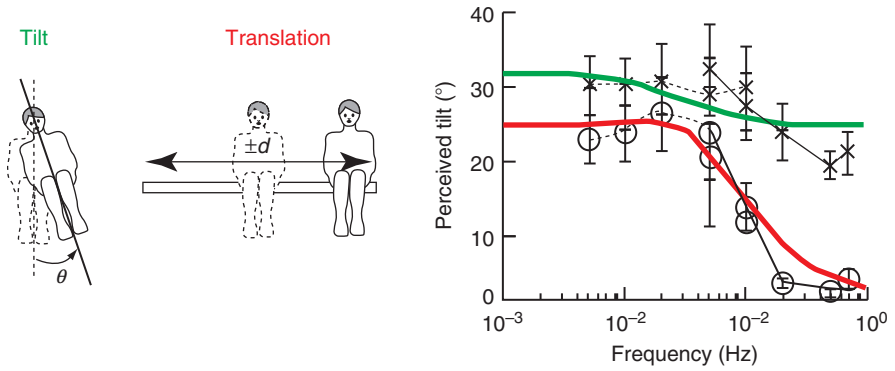

Figure 4 Magnitude of roll tilt perception, which is tilt to the left or right, provided by subjects using a two-handed task as a function of frequency between 0.005 and 0.7 Hz. Tilt stimuli: Roll tilt perception during applied sinusoidal roll tilts are shown (indicated by x) with a modeling prediction overlaid in green. Translation stimuli: Illusory roll tilt perception reported during sinusoidal linear accelerations (indicated by o), with no actual tilt stimuli, are shown with a modeling prediction overlaid in red. Adapted from Park S, Gianna-Poulin C, Black FO, Wood S, and Merfeld DM (2006) Roll rotation cues influence roll tilt perception assayed using a somatosensory technique. *Journal of Neurophysiology* 96: 486–491. Roll tilt modeling prediction from Merfeld DM, Park S, Gianna-Poulin C, Black FO, and Wood S (2005) Vestibular perception and action employ qualitatively different mechanisms I: Frequency response of VOR and perceptual responses during translation and tilt. *Journal of Neurophysiology* 94: 186–198.

rotations, responding minimally to rotations that do not include yaw head rotations. Similarly, the other semicircular canals, the anterior and posterior canals, respond only to rotational motion aligned with the anterior and posterior canal planes, respectively. A similar relationship holds true for the otolith organs. Individual neurons have spatial tuning characteristics that result in the neuron firing when a component of

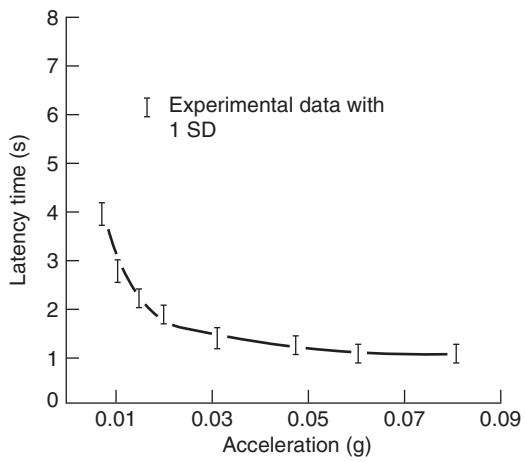

Figure 5 Latency between the start of a linear acceleration step and the subjective indication of the direction of linear motion, shown as a function of the level of linear acceleration; this latency is often called the time-to-detect. The time-to-detect decreases as the magnitude of the acceleration step increases. In this paradigm, the relationship between the time-to-detect and the linear acceleration magnitude covary in a manner consistent with the subjects correctly detecting the direction of motion when they attain a velocity of approximately 0.20 m s^{-1}. SD, standard deviation. From Young L (1984) Perception of the body in space: Mechanisms. In: Darian-Smith I (ed.) *Handbook of Physiology: The Nervous System*, pp. 1023–1066. Bethesda, MD: American Physiological Society.

gravity or linear acceleration aligns with that neuron's axis of maximal sensitivity, which is determined by the polarity of the innervated hair cell(s).

Generally, stimulus intensity is signaled by the firing rate of the afferent neurons. Neurons in the vestibular system of primates have a relatively high tonic baseline activity of approximately 100 action potentials per second. The firing rate of these neurons increases for motion in one direction and decreases for motion in the opposite direction. Within the normal physiological range, the magnitude of the change in the neural firing rate of each individual neuron represents the intensity of the motion. For rotations at physiological frequencies (circa 0.1–5 Hz), the magnitude of the neural firing rate for neurons innervating the semicircular canals scales nearly linearly with angular velocity until the stimulus becomes so large that saturation effects predominate. Similarly, for the otolith organs, the magnitude of the neural firing rate scales roughly linearly with (1) the component of linear acceleration aligned with the sensitive direction of each neuron innervating the otolith organ and/or (2) the gravitational component aligned with the sensitive axis of each otolith neuron. It is interesting to note that, saturation effects aside, the vestibular afferent responses change roughly linearly with stimulus intensity. This differs from most other sensory systems, in which the afferent response varies with the logarithm of the stimulus intensity.

The timing of the physical stimulus is nearly instantaneously indicated by changes in the neural firing rate of vestibular neurons. At physiological frequencies, the firing rate of each neuron innervating a semicircular canal increases and decreases in near synchrony with increases and decreases in the angular velocity. Similarly, the firing rate for neurons innervating the otolith organs increases and decreases in near synchrony with linear acceleration and/or gravitational stimulation.

Summary

The vestibular system is a sensory system that makes crucial contributions to our sense of motion and orientation. The subjective experiences evoked by the vestibular system include perceptions of angular velocity and linear velocity and perceptions of both tilt and heading. These subjective experiences are evoked by objective physical stimuli; the semicircular canals sense angular acceleration, whereas the otolith organs sense both linear acceleration and gravity. These physical stimuli lead to sensory transduction, which cause changes in the neural activation of the vestibular afferents. This neural activation is a first step in the physiological processes that lead to neural representations of the physical stimuli, which are sometimes called internal representations. Such internal representations are the brain's way of storing and processing the external physical stimuli and combining the vestibular information with other information – like information provided by visual cues, somatosensory cues, and/or cognitive cues – to best inform us about our motion and orientation.

See also: Auditory/Somatosensory Interactions; Canal–Otolith Interactions; Hair Cells: Sensory Transduction; Motion Sickness; Neural Coding of Spatial Representations; Spatial Cognition and Executive Function; Vestibular System; Vestibulo-Autonomic Responses; Vestibulo-Ocular Reflex; Vestibulospinal System and Eye–Head/Neck Movement; Vision: Mechanisms of Orientation, Direction and Depth; Visual–Vestibular Interactions.

Further Reading

Benson A (1982) The vestibular sensory system. In: Barlow H and Mollon J (eds.) *The Senses*, pp. 333–368. Cambridge, UK: Cambridge University Press.

Bortolami SB, Pierobon A, Dizio P, and Lackner JR (2006) Localization of the subjective vertical during roll, pitch, and recumbent yaw body tilt. *Experimental Brain Research* 173: 364–373.

Gardner E and Martin J (2000) Coding of sensory information. In: Kandel E, Schwartz J, and Jessel T (eds.) *Principles of Neural Science,* 4th edn., pp. 411–429. New York: McGraw-Hill.

Guedry F (1974) Psychophysics of vestibular sensation. In: Kornhuber HH (ed.) *Handbook of Sensory Physiology*, pp. 1–154. New York: Springer.

Howard I and Templeton W (1966) *Human Spatial Orientation.* New York: Wiley.

Kandel E, Schwartz J, and Jessell T (2000) *Principles of Neural Science,* 4th edn. Upper Saddle River, NJ: Prentice-Hall.

Merfeld DM, Park S, Gianna-Poulin C, Black FO, and Wood S (2005) Vestibular perception and action employ qualitatively different mechanisms I: Frequency response of VOR and perceptual responses during translation and tilt. *Journal of Neurophysiology* 94: 186–198.

Mergner T and Becker W (1990) Perception of horizontal self-rotation: Multisensory and cognitive aspects. In: Warren R and Wertheim A (eds.) *Perception & Control of Self-Motion*, pp. 219–263. Hillsdale, NJ: Lawrence Erlbaum.

Park S, Gianna-Poulin C, Black FO, Wood S, and Merfeld DM (2006) Roll rotation cues influence roll tilt perception assayed using a somatosensory technique. *Journal of Neurophysiology* 96: 486–491.

Young L (1984) Perception of the body in space: Mechanisms. In: Darian-Smith I (ed.) *Handbook of Physiology: The Nervous System*, pp. 1023–1066. Bethesda, MD: American Physiological Society.

Spatial Transformations for Eye–Hand Coordination

G Blohm, A Z Khan, and J D Crawford, York University, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

One of the most important ways in which humans interact with our environment is through the use of our hands, regardless of whether we manipulate objects directly or whether we use tools. In order to accurately reach and grasp objects of interest, we require precise egocentric information about where the object is relative to our body. The direction of an object is specified by the location on the retina that it stimulates. The distance of the object is provided through stereoscopic vision arising from the slightly different views of the two eyes. The eyes, however, only provide the spatial locations of objects relative to the retinas, which depends on the orientation of the eyes in space (gaze). In contrast, the arm is attached to the shoulder and, therefore, muscle contractions need to be specified with respect to this fixed point. Therefore, information about an object's location must be transformed from a gaze-centered representation into one that is centered on the insertion points of the arm muscles (shoulder and elbow). For simplicity, we will refer to the latter as a shoulder-centered representation.

We begin the following sections by discussing the gaze-centered encoding of reach space as well as some of the implications of this particular choice of representation. Next, we consider how motor output is specified in shoulder-centered coordinates by the brain. Finally we examine the transformation required to map the gaze-centered representation of space onto a shoulder-centered movement plan.

Visual Representations of Reach Space

Gaze-Centered Encoding of Reach Targets

Because the projection of a visual object on the retinas changes with the direction of the eyes, the spatial information that reaches visual cortex depends on gaze direction. This gaze-centered representation of reach targets is not only present in the early visual pathway (i.e., retina, lateral geniculate nucleus, striate cortex) but also later on in the visual pathway (e.g., extrastriate and parietal cortex). Thus, the spatial location of an object that we want to reach to or point at is initially encoded relative to the fovea (Figure 1(a)). The persistence of gaze-centered representations at intermediate levels such as parietal cortex might seem surprising considering that alternative representations, such as a head-centered coding, would be independent of intervening eye movements and thus more stable (in theory, a head-centered representation adds eye position to the gaze-centered target position to specify its position relative to the head). However, an advantage of gaze-centered representations is that, unlike head-centered ones, they do not require such an integration of eye position until one actually reaches, and so are computationally less expensive for the brain. It is worth mentioning that the sensorimotor system also encodes auditory and proprioceptive targets in gaze-centered coordinates, even though these senses are not fixed to the eye.

How is this gaze-centered representation encoded in the brain? In order to accurately reach for an object, the brain requires knowledge about the object's direction relative to the fovea (horizontal and vertical angular eccentricity) and how far it is from the eyes (egocentric distance). The visual field is projected onto the retina and is transmitted to the visual cortex in a topographically arranged fashion. Whereas eccentricity in this topographical map defines the direction of a potential reach target, the distance of this target is specified both by monocular (accommodation, relative object size, shading, perspective, etc.) and binocular (retinal disparity and convergence) information.

Early visual areas (yellow areas in Figure 1(b)) compute angular eccentricity based on the retinal images from both eyes, synthesized to provide target direction as seen by a virtual (cyclopean) eye located between the right and left eye. On the other hand, retinal disparity is calculated from the difference between the two eyes' images. Different areas in the early visual pathway carry out the computations related to target direction and disparity. In early movement planning, these signals appear to be recombined into a single 3-D representation of gaze-centered space in the posterior parietal cortex (orange area in Figure 1(b)); different neurons in the monkey parietal cortex have receptive fields that encode a specific angular eccentricity of a target and their activity is modulated by target distance. To summarize, early visual and visuomotor areas compute different spatial aspects of the 3-D location of reach targets in gaze-centered coordinates.

Updating of Spatial Memory

As mentioned in the previous section, the drawback of a gaze-centered representation of space is that it needs to be constantly adjusted as we move. This is

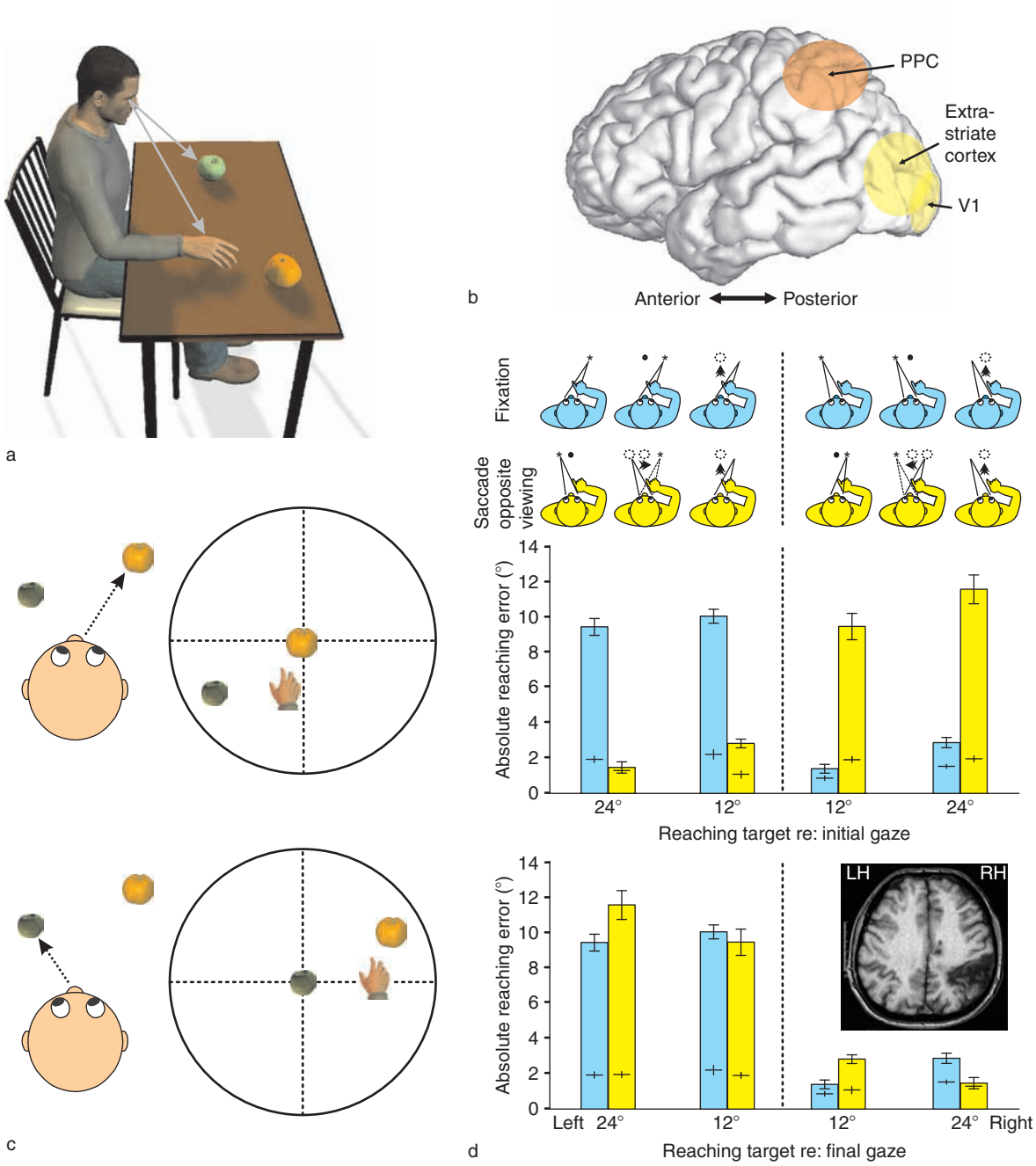

Figure 1 Gaze-centered encoding of reach space. (a) Drawing depicting the egocentric visual directions of the hand, an orange, and an apple, as shown by the gray arrows. (b) Side view of a human brain showing areas (highlighted in yellow and orange) that encode reach space in gaze-centered coordinates: V1, striate cortex; PPC, posterior parietal cortex. (c) The consequence of an eye movement on the gaze-centered representation of the visual field. The head/eye diagrams depict current gaze position; the circles represent the gaze-centered representation of this visual scene (dotted lines represent visual horizontal and vertical axes and intersect at the fovea). If the person in the upper diagram looks at the orange, the hand and apple are represented in the left visual field (upper circle). In contrast, if the person fixates the apple (lower diagram), the orange and hand are now represented in the right visual field (lower circle). If the orange and the hand were no longer visible when the eye movement occurred, the brain would need to remap their position by taking the intervening eye movement into account. (d) Remapping in optic ataxia. The fixation task (first row, blue) shows a typical pattern of reaching errors (identical blue bars in both graphs) for this disorder, where errors are greater when the reach target is in the contralesional (opposite to the lesion) visual field (left half of each graph). The remapping task (second row, yellow) required the person to view the target peripherally in one visual field and then make an eye movement to the opposite side before reaching to the target. If no remapping takes place, then errors (yellow bars) should be greater when plotted as a function of reach target location relative to initial gaze (first graph), whereas if the reach target is remapped with the eye movement, then errors should be greater when plotted against reach target relative to the final gaze position (second graph). The reaching errors in the second graph match very well to the control task errors, suggesting that the location of the reach target in the internal representation was indeed remapped during the eye movement. (d) Adapted from Khan AZ, Pisella L, Vighetto A, et al. (2005) Optic ataxia errors depend on remapped, not viewed, target location. *Nature Neuroscience* 8(4): 418–420, with permission.

necessary to maintain a stable representation of our environment despite shifts and rotations of the eyes. This adjustment of the gaze-centered representation through different types of movements is called updating or remapping. Remapping has been demonstrated to be a general phenomenon that occurs during different types of eye movements – that is, slow eye tracking movements (smooth pursuit) and fast orientation movements (saccades). Updating also occurs for both the rotation and the translational components during head and body movements. Consider the example in Figure 1(c); if the viewer initially looks at the orange (upper left panel), the apple is located to the left and down of it. In the internal representation of this visual scene (represented by the schematic in the upper right panel), the orange is represented at the center of the representation – that is, the fovea – and the apple is mapped onto the lower left quadrant of the visual field. After a saccade to the apple (lower left panel), both the apple and the orange are remapped by the same amount of rotation as the saccade, but in the opposite direction. As a result, the apple is now represented at the fovea and the orange is mapped onto the upper right quadrant. In everyday life, incoming visual information usually contributes to specify the new target location for these internal representations, but many experiments have shown that the brain can accomplish this remapping internally, even if vision is removed, by using eye movement signals from the oculomotor system. Also, it has recently been shown that remapping and vision give better performance compared to vision alone. Psychophysical evidence suggests that about four targets can be represented and remapped in this way.

Intuitively, it would make sense to simply shift the locations of the internal representation by the same magnitude as the eye movement, but in the opposite direction. However, recent studies have shown that because of the kinematics of eye movements (i.e., there is a nonzero torsional component about the line of gaze in noncentral eye positions), shifting the locations by the eye movement vector does not produce correct remapping. Moreover, during pointing to remembered targets, subjects perform in a manner that suggests that they do remap correctly after eye movements, which reveals that the brain does take the rotational component of the eye movement into account during remapping.

It is generally accepted that the areas in and around the medial bank of the intraparietal sulcus of the parietal cortex play a major role in representing and remapping reach targets in gaze-centered coordinates. The receptive fields (a receptive field corresponds to the visual locations in space that activate the neuron) of individual parietal neurons have been shown to shift during an eye movement to correspond to the remapped location of the target. At the population level, the remapping of individual receptive fields results in an updated memorized target position. Remapping of target location across saccades has also been shown in the posterior parietal cortex through human functional magnetic resonance imaging studies. Moreover, recent neural recordings in the monkey show evidence of updating in a number of extrastriate visual areas such as areas V3A, V3, and V2. This leads to the question of the specific roles these different areas could play in the remapping process for reach movements.

One way to demonstrate the role that parietal and occipital cortex play in remapping is to test unilateral optic ataxia patients with damage to the (right) superior parietal cortex. Optic ataxia generally manifests itself as a deficit in performing accurate visually guided movements to peripherally viewed objects in the contralesional (opposite side to the lesion) visual field. Figure 1(d) depicts reach errors in such a patient. After an intervening saccade, errors to previously memorized targets were better explained by the remapped gaze-centered representation than by the initial retinal encoding pattern. The fact that this patient was able to remap despite damage to the right superior parietal cortex is consistent with a distributed remapping process that involves both extrastriate visual and parietal areas. Occipital cortex probably plays a more general role for remapping in perception and action, whereas remapping signals in parietal cortex might be specific for actions like reaching and saccades. In summary, the complete 3-D locations of reach goals are represented in gaze-centered coordinates and these representations are updated during various movements of the body.

Motor Coding of Reach Movements

Shoulder-Centered Encoding of Reach Plans

In order to understand eye–hand coordination, we must also consider the physiology of limb control. A gaze-centered representation of a reach target is inappropriate for reaching because muscle activations which move the arm are determined with respect to the shoulder. Consequently, the brain needs to specify the desired movement vector in a coordinate system relative to the shoulder, which is defined as the difference between the current and desired hand position (Figure 2(a)). In this section, we illustrate how and where such shoulder-centered motor plans are encoded. The subsequent section briefly describes how these motor plans are transformed into the appropriate muscle activations.

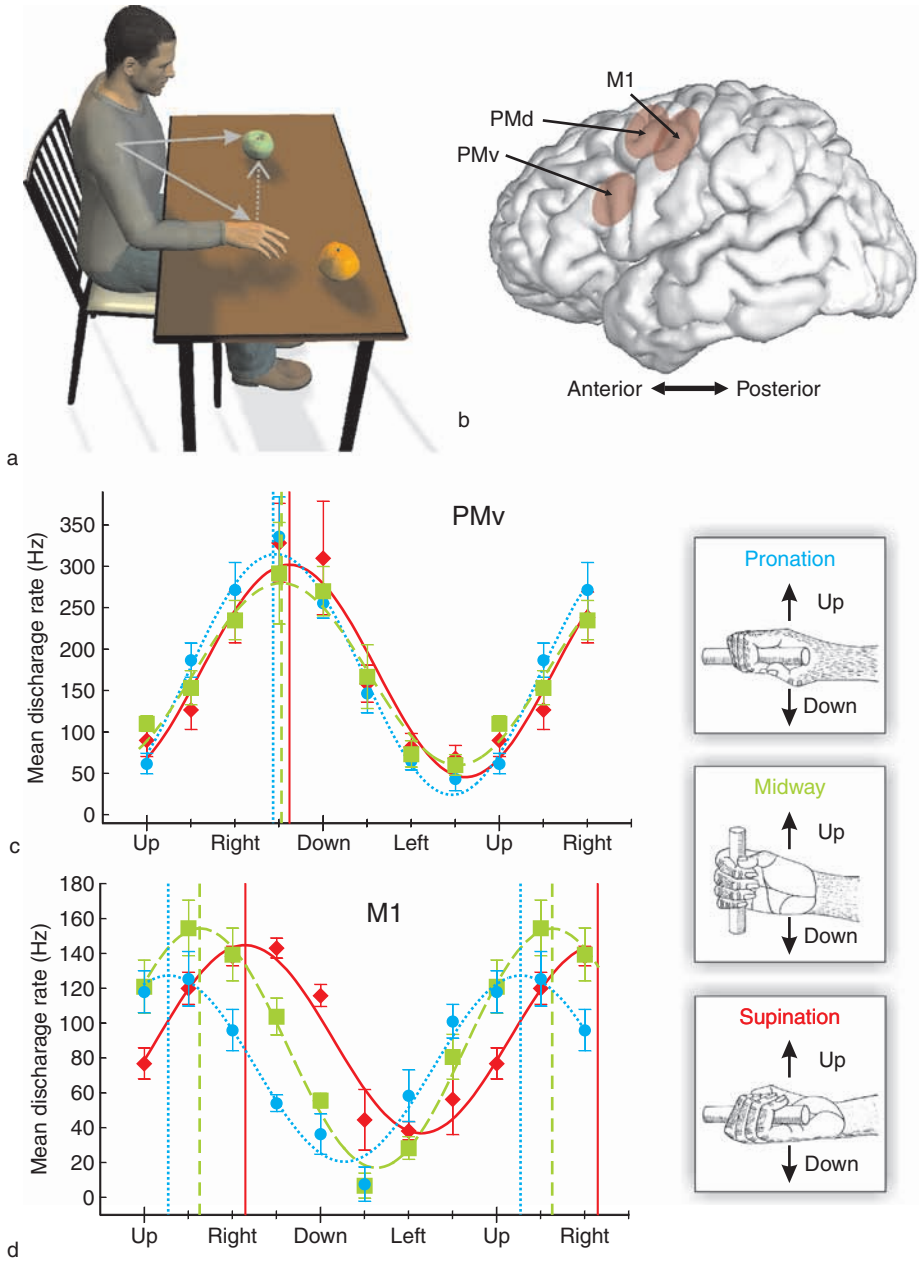

Figure 2 Motor representation of reaches. (a) The diagram depicts the position of the hand and apple encoded relative to the shoulder (gray solid arrows) as well as the desired movement vector (gray dotted arrow). (b) Schematic of the human brain (side view) showing areas (highlighted in red) that encode the desired movement vector in shoulder-centered or musclelike joint-centered coordinates: PMd, dorsal premotor cortex; PMv, ventral premotor cortex; M1, primary motor cortex. (c) Tuning curve of an example neuron in PMv. In this experiment, nonhuman primates where trained to perform wrist movements to visual targets in different spatial directions (x-axis) and the posture of the wrist was modified systematically. The three panels on the right show the three different postures used – pronation, midway (between pronation and supination), and supination. Each curve in the graph corresponds to the tuning curve of the same neuron recorded under those three different postural conditions. Clearly, this PMv neuron does not modulate its discharge rate with posture, which is consistent with a shoulder-centered representation of the desired reach. (d) Example neuron in M1 showing musclelike tuning curves. Here, the neuron's discharge was modified by the three different postural conditions. The directional shift of the tuning curve is in phase with the muscle activations required for a specific movement direction. Therefore, this M1 neuron shows a musclelike joint-centered encoding of the desired reach. (c, d) Adapted from Kakei S, Hoffman DS, and Strick PL (2003) Sensorimotor transformations in cortical motor areas. *Neuroscience Research* 46(1): 1–10, with permission from Elsevier.

Early neural recordings in the primary motor cortex (M1; Figure 2(b)) of the monkey during reach movements showed that individual neurons respond best to a certain movement direction, which is known as the neuron's preferred direction. The activity of such neurons decreases as the angular distance between the movement vector and the preferred direction increases. This pattern of activation forms

what is known as a cosine tuning of neural activity. The preferred directions of populations of neurons encoding movement seem to be uniformly distributed, thus all possible movements are represented. In summary, neurons with preferred directions and cosine-tuning properties provide the essential neurophysiology for movement encoding in the premotor and motor cortices.

To specify a complete 3-D reach plan one also needs to know the amplitude of the desired movement. Most psychophysical studies seem to indicate that direction and amplitude might be encoded independently, a conclusion drawn from the observation that errors in movement direction are generally smaller than errors in movement amplitude. On the other hand, neurophysiological recordings of neural activity in the dorsal premotor cortex (PMd; Figure 2(b)) during reach planning have shown that amplitude and direction are always coded together within the same neuron, where the activity related to a fixed movement direction simply increases or decreases depending on movement amplitude. One unifying solution for these seemingly contradictory findings might be that, at the neuronal level, direction and amplitude are coded together, but are read out by different mechanisms to generate muscle activation.

A large portion of neurons in the monkey ventral premotor cortex (PMv; Figure 2(b)) encodes extrinsic movement direction (i.e., the movement vector in an explicit shoulder-centered representation). Figure 2(c) shows this pattern for wrist movements where the tuning curve of a neuron in PMv reflects movement direction independent of hand posture (e.g., the palm turned up or down). Apart from the PMv, smaller portions of neurons in the PMd and the primary motor cortex (M1) also show such extrinsic spatial behavior. This shoulder-centered code coexists (in PMd and M1) with other, more muscle-related representations, which may reflect the gradual transformation of those codes into muscle activations. An abstract shoulder-centered programming of a motor plan is a fundamental stepping-stone for the brain to generate the appropriate muscle activations for the actual reach movement.

Planned Movement Vectors and Kinematics

As mentioned earlier, the shoulder-centered motor plan needs to be converted into the appropriate pattern of muscle activations to move the arm. The conversion has to correctly determine not only which muscles to activate (to specify direction) but also the correct force for each set of muscles (to specify amplitude). In addition, the brain has to overcome a series of dynamic constraints, such as the limb inertia and different arm loads due to the load of various hand-held objects. To generate accurate muscle activations, the brain must account for the current posture of the arm. Figure 2(d) shows a neuron in M1 with an activity that reflects the influence of posture on its cosine-tuning curve. This particular neuron does not code movement in shoulder-centered coordinates, but instead seems to represent a motor plan with respect to the muscles required for a particular movement. The shoulder-centered motor coding and the muscle-centered movement planning represent the two ends of a continuum, demonstrating the gradual conversion of extrinsic movement information into muscle activation patterns. This conversion involves large portions of neurons in PMd and M1 as well as the cerebellum.

To summarize, the brain represents motor plans in shoulder-centered coordinates in the premotor and motor cortices and then converts those representations of the desired reach into muscle activations. However, since the visual input to the brain specifies the location of a reach target in gaze-centered coordinates, the brain first must transform the early visual signals into appropriate motor commands for the arm in shoulder-centered coordinates. As we will see, simply using a gaze-centered movement vector to move the arm directly would produce large reach errors as a function of eye and head orientation. This visuomotor transformation is the topic of the following section.

The Visuomotor Transformation

Comparison between Hand and Target Positions

To reach out for a visual object, the brain needs to specify the required movement vector by computing the difference between the current hand location and the position of the object in space. This comparison between hand and target positions must satisfy two requirements. First, the areas in the brain that carry out this computation must contain signals providing the complete spatial information of both hand and target locations. Second, this comparison requires that hand and target locations are represented in the same reference frame in order for both to be combined in a spatially consistent manner. However, it is not yet clear which areas in the brain are involved in this comparison and in which reference frame the comparison is made.

There are three possibilities by which the brain could fulfill these requirements. One way could be to compare the hand and target positions as early as possible in the visuomotor pathway (i.e., compare the hand position to the target position in gaze-centered coordinates). As mentioned earlier, in the second section, target position has been shown to be encoded in gaze-centered coordinates at the level of the posterior

parietal cortex (PPC). Areas in the monkey PPC that code visual objects in gaze-centered coordinates have also been found to code hand position, even when the hand was not seen. This is particularly interesting because it means that the brain can derive the current hand position in gaze-centered coordinates from muscle-related proprioceptive signals. A neural substrate for a comparison between hand and target positions in gaze-centered coordinates has recently been identified in areas closer to the cortical surface of the intraparietal sulcus in the PPC. The activity of neurons in this area depends on both hand and target positions, which is consistent with the formation of a desired movement vector in gaze-centered coordinates.

Alternatively, hand and target positions could be compared at a much later stage (e.g., in shoulder-centered rather than in gaze-centered coordinates). This would require target position to be first transformed from gaze- to shoulder-centered coordinates. The hand position signal used for this comparison could be derived either from visual signals, which go through the same transformation process as target position, or from proprioceptive inputs. The latter joint- or muscle-related signals would require an intermediate transformation into shoulder-centered coordinates through an inverse model of the arm, and then a further transformation into gaze-centered coordinates would be unnecessary. Evidence for a late comparison of hand and target positions in shoulder-centered coordinates comes from several psychophysical experiments. If the hand is unseen, error of reaching toward memorized target locations are compatible with a shoulder-centered hand and target comparison. In contrast, simply providing visual hand position feedback shifted the errors toward a pattern compatible with the comparison of hand and target locations in gaze-centered coordinates.

Finally, the comparison between hand and target could be a distributed process involving multiple stages of representations taking place across several brain areas. This might seem redundant or contradictory, but it is well known that the brain can derive information from multiple sources (e.g., in sensory integration) and combine these inputs using Bayesian-like algorithms to compute the most reliable estimate. This scheme potentially enables the brain to access the same information encoded in different frames of reference, which would allow it to use the different representations, depending on task requirements. This is supported by behavioral findings which show that task context can alter the reference frame in which the comparison between hand and target locations is carried out. Regardless of where this comparison is done, a reference frame transformation between the gaze- and shoulder-centered representations is necessary. This visuomotor transformation could either be carried out for hand and target locations independently or for the combined signal (i.e., the desired movement vector).

Reference Frame Transformations

To transform gaze-centered into shoulder-centered signals, the brain has to perform a cascade of conversions which require knowledge about the 3-D orientation of body parts. Theoretically, the transformation involves a set of serial steps which gradually combine eye-in-head, then the head-on-neck, orientation, and subsequently the shoulder rotation with the visual representation, to produce a head-centered, body-centered, and ultimately shoulder-centered representation. But how do we know that these reference frame transformations are even necessary?

To answer this question, consider the example depicted in **Figure 3(a)**, where a study participant is asked to reach out to a target toward the left. When the head is upright and the eyes are fixating straight ahead, the gaze-centered coordinate system has the same orientation as the shoulder-centered coordinate system and is only translated with respect to the latter. Therefore, the desired movement vector is the same in both reference frames and no explicit reference frame transformation would be required. But what would happen if the head was tilted, say, toward the left shoulder, and the person maintained straight-ahead gaze? Since the eyes remain more or less in the same position relative to the head, the retinas are now tilted in space. Therefore, the projection of hand and target onto the retina would be rotated (in gaze-centered coordinates) in the opposite direction as the head – that is, toward the right. If this vector was directly used to guide the reach, the arm movement would be inaccurate and the move would be up and to the left instead of simply horizontally. Everyday experience and experimental evidence suggests that people generally do not make such errors, which demonstrates that the brain performs a reference frame transformation of the desired movement vector.

The preceding example illustrates a simplified portrayal of a reference frame transformation. However, this visuomotor transformation is actually much more complex. A large part of this complexity arises because body parts rotate around their anchor points; for example, the eye rotates in the ocular cavity and the head rotates around the base of the skull. Since rotations are nonlinear and noncommutative (sequence matters) operations, the relationship between gaze-centered and shoulder-centered coordinates can become quite complicated. In addition, the anchor points for different body parts are offset one from another. For example, if the head rotates, the eyes shift their position in space. Not taking translations into account could

result in a mislocalization of target position during the reach. In sum, an accurate visuomotor transformation has to take into account the rotational and translational aspects of body geometry.

Brain areas involved in this transformation would necessarily require extraretinal signals about the rotation of the eyes and head in addition to representing the 3-D hand and/or target locations. The

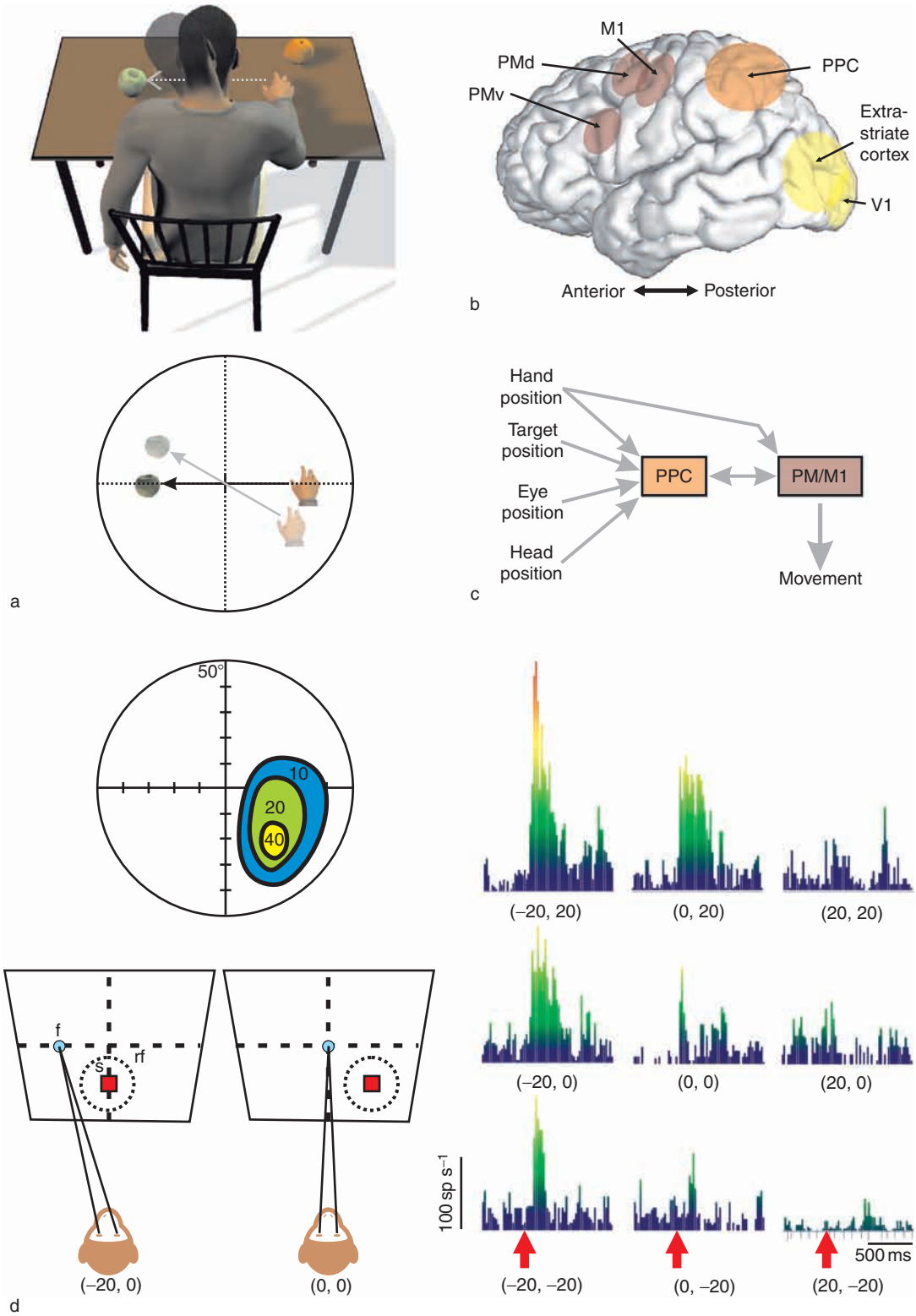

Figure 3 Continued

translational components are also implemented, probably acquired through experience. How would the brain purposefully combine these different signals in a population of neurons? A potential window into the distributed processing in neural populations is provided by the analysis of artificial neural networks trained to perform reference frame transformations. Artificial neural networks are highly simplified mathematical analogies of real networks of neurons and are believed to reflect fundamental mechanisms underlying brain function. Probably the most important finding is the discovery of neurons (both real and simulated) which alter their response magnitude based on extraretinal eye and/or head position signals, without changing the gaze-centered location of their visual receptive fields. This modulation of response magnitude is known as a gain field and is illustrated in Figure 3(d).

Mathematically speaking, gain fields operate through a nonlinear combination of different sources of information (e.g., visual input, eye position, head position) within the same neuron. If many different neurons incorporate multiple combinations of these signals, then the neural population has a complete distributed representation of space relative to different body parts. The individual multimodal neurons inside this population act like basis functions in that they provide a unique representational set and can be combined (by linear summation) in different ways to produce the desired reference frame output. Thus, using the same population of basis functions, it is possible to extract target position in multiple frames of reference.

The shortcoming of the concept of basis function neurons is that every possible combination of signals must be represented, which could lead to a large number of units in the population, a dilemma called the curse of dimensionality. Therefore, it is believed by some that the brain compromises between basis function neurons – implying the presence of gain fields – and explicit representations (e.g., shoulder-centered encoding of the desired movement vector) in various reference frames. Evidence for this comes from brain areas such as the PPC, which contains many neurons encoding reach space in gaze-centered coordinates. A significant proportion of PPC neurons with gaze-centered receptive fields also show eye-position-dependent gain fields. Neurons in the premotor and motor cortices show movement coding in shoulder-centered coordinates, which constitute another form of explicit representation of reach space. Gain fields are thought to be the implicit code that implements the visuomotor transformation between an explicit gaze-centered code and an explicit shoulder-centered reach plan.

Conclusions

Visually guided reaching involves the transformation of sensory information from the eyes and hand into movement commands for the arm. The early representation of sensory input is gaze-centered and appears to be remapped during body movements in order to ensure a stable internal representation of the visual environment. The arm motor plan is calculated by comparing the desired target position to the current hand location. This motor plan is spatially accurate because the brain accounts for eye and head posture through a distributed internal model of body geometry. As a result, the motor plan is specified in shoulder-centered coordinates and then further transformed into muscle activation patterns. This cascade of different representations and mechanisms (e.g., gaze centered, shoulder centered, updating) may seem complex and perhaps redundant at first glance. However,

Figure 3 Visuomotor transformation. (a) Diagram depicting two different possible head roll postures (head upright and rolled toward the left shoulder) while a person viewing an apple and an orange is planning to reach out for the apple (gray dotted arrow). The corresponding gaze-centered representation is shown schematically in the circle. Note that the same shoulder-centered movement plan (diagram of person, dotted arrow) leads to different gaze-centered desired movement vectors in the circle, depending on head posture (head upright, black arrow; head roll, faded gray arrow). (b) Schematic of the human brain (side view) showing the visuomotor transformation pathway from early visual to motor structures: PMd, dorsal premotor cortex; PMv, ventral premotor cortex; M1, primary motor cortex. (c) Model of the visuomotor transformation. Visual (target and hand position), nonvisual (proprioceptive hand position), and extraretinal (eye and head position) information is combined in the posterior parietal cortex (PPC) and also in the premotor and motor cortices (PM/M1) to produce an accurate movement. (d) Gain field modulations in area 7a of the posterior parietal cortex. The top circles depict the gaze-centered receptive field of one example neuron. The overlapping circular shapes show different levels of activation for this cell as a function of horizontal and vertical gaze-centered target location. Based on this information, a stimulus (red squares in the lower diagrams) was placed at the location of the highest activity in the neuron's receptive field. The animal fixated at one of nine locations and the activity in response to the stimulus was measured in the neuron's receptive field. The stimulus was always presented at the same location relative to gaze (i.e., in the receptive field, which moved with fixation), as can be seen in the two lower diagrams for two example fixation locations (center left and straight ahead). The graphs on the right show the cell's activity over time (red arrows signify stimulus onset) for the nine fixation locations. The figure shows that the cell's activity for the same stimulus location (relative to gaze) was modulated by eye position, as predicted by the gain field theory. (d) Data modified from Andersen RA, Essick GK, and Siegel RM (1985) Encoding of spatial location by posterior parietal neurons. *Science* 230(4724): 456–458, with permission. Copyright 1985 by AAAS.

this appears to be the most cost-efficient way to deal with the large amounts of information in the system – that is, by solving distinct computational problems at different stages of the visuomotor transformation, such as making cognitive comparisons between multiple targets at early visual stages and specifying detailed limb dynamics at later stages.

One of the remaining challenges for neuroscientists is to understand the theoretical requirements and mathematical algorithms the brain employs to carry out these transformations. Without such a theoretical framework, it is difficult to interpret the observed properties of individual neurons in the parietal to frontal network. This framework in turn is necessary to explain what goes wrong in patients with neurological disorders affecting visually guided reaching. Further, the field of visuomotor neuroscience is now at a point where we can begin to tackle more complex problems, such as the influence of other sensory modalities, mechanisms engaged when reaching to targets in movement, decision making, target selection, and the role of attention.

See also: Motor Psychophysics; Multisensory Convergence and Integration; Neural Coding of Spatial Representations; Oculomotor Control: Anatomical Pathways; Parietal Cortex and Spatial Attention; Posterior Parietal Cortex and Arm Movement; Posterior Parietal Cortex and Tool Usage and Hand Shape; Reaching and Grasping; Sensorimotor Integration: Models; Sensorimotor Control of Manipulation; Supplementary Eye Fields.

Further Reading

Andersen RA and Buneo CA (2002) Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience* 25: 189–220.

Andersen RA, Essick GK, and Siegel RM (1985) Encoding of spatial location by posterior parietal neurons. *Science* 230(4724): 456–458.

Batista AP, Buneo CA, Snyder LH, et al. (1999) Reach plans in eye-centered coordinates. *Science* 285: 257–260.

Buneo CA, Jarvis MR, Batista AP, et al. (2002) Direct visuomotor transformations for reaching. *Nature* 416(6881): 632–636.

Caminiti R, Ferraina S, and Mayer AB (1998) Visuomotor transformations: Early cortical mechanisms of reaching. *Current Opinion in Neurobiology* 8(6): 753–761.

Crawford JD, Medendorp WP, and Marotta JJ (2004) Spatial transformations for eye-hand coordination. *Journal of Neurophysiology* 92(1): 10–19.

Georgopoulos AP, Kettner RE, and Schwartz AB (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *Journal of Neuroscience* 8(8): 2928–2937.

Henriques DYP, Klier EM, Smith MA, et al. (1998) Gaze-centered re-mapping of remembered visual space in an open-loop pointing task. *Journal of Neuroscience* 18: 1583–1594.

Kakei S, Hoffman DS, and Strick PL (2003) Sensorimotor transformations in cortical motor areas. *Neuroscience Research* 46(1): 1–10.

Kalaska JF, Scott SH, Cisek P, et al. (1997) Cortical control of reaching movements. *Current Opinion in Neurobiology* 7(6): 849–859.

Khan AZ, Pisella L, Vighetto A, et al. (2005) Optic ataxia errors depend on remapped, not viewed, target location. *Nature Neuroscience* 8(4): 418–420.

Merriam EP and Colby CL (2005) Active vision in parietal and extrastriate cortex. *Neuroscientist* 11(5): 484–493.

Pouget A and Sejnowski TJ (1997) Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience* 9(2): 222–237.

Salinas E and Abbott LF (2001) Coordinate transformations in the visual system: How to generate gain fields and what to compute with them. *Progress in Brain Research* 10: 175–190.

Soechting JF and Flanders M (1992) Moving in three-dimensional space: Frames of reference, vectors, and coordinate systems. *Annual Review of Neuroscience* 15: 167–191.

Zipser D and Andersen RA (1988) A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature* 331(6158): 679–684.

Spectrin: Organization and Function in Neurons

A Das and R R Dubreuil, University of Illinois at Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Spectrin is an abundant protein in the nervous system, comprising 2.4% of total protein in mammalian brain homogenates. Spectrin is an unusually large protein with a native molecular weight greater than 1 million. The contour length of the purified protein when viewed by electron microscopy is ~200 nm. Spectrin is intimately associated with the cytoplasmic surface of the plasma membrane in many cell types, including neurons. There are also many cases in which spectrin is associated with intracellular structures, including the Golgi apparatus and nucleus.

Spectrin was originally discovered in human erythrocytes, where it is the major component of a submembrane protein scaffold known as the membrane skeleton. Biochemical studies of purified erythrocyte spectrin yielded a wealth of information on the structure and physical properties of the molecule as well as its interactions with other components of the erythrocyte membrane. The biochemical approach led to a model in which spectrin molecules are cross-linked by short actin filaments to form the submembrane scaffold. The model was eventually confirmed by electron microscopy of artificially spread preparations of the erythrocyte membrane (a simplified schematic is shown in Figure 1(b)). The abundance of spectrin in mammalian brain made it possible to purify the protein and subject it to the same kinds of biochemical approaches employed with its erythrocyte counterpart. That approach revealed that brain spectrin is capable of nearly all the protein interactions that occur in the erythrocyte. However, it has not been possible to establish the precise structural arrangement of spectrin molecules in any cells other than erythrocytes.

There are two kinds of linkages between neuronal spectrin and the plasma membrane. The best-characterized interactions occur through the peripheral membrane protein ankyrin, which indirectly couples spectrin to integral membrane proteins. There are also direct links between spectrin and membrane lipids as well as integral membrane proteins such as the neural cell adhesion molecule.

Structure

The native spectrin molecule is a tetramer composed of α and β subunits (Figure 1(a)). The subunits associate laterally to form antiparallel $\alpha\beta$ heterodimers. Dimers, in turn, associate head-to-head to form $\alpha_2\beta_2$ tetramers. Each subunit is composed of discrete structural domains, referred to here as segments, numbered in order from the N-terminus of each polypeptide (α1–22; β1–19).

Amino acid sequence analyses of the two spectrin subunits have produced considerable insight into the structure of the native protein. Much of the body of the protein is composed of a degenerate 106 amino acid spectrin repeat. Atomic structures for individual repeats established that they have a triple-barrel alpha helical structure. There are 20 repeats in α spectrin and 16 repeats in β spectrin. Another repeat is formed by the end-to-end interaction between α and β spectrin during tetramer formation. Here, a partial repeat near the C-terminus of β spectrin associates with a partial repeat near the N-terminus of α spectrin. Note that the numbering scheme for an α spectrin segment begins with its first full spectrin repeat. It has been suggested that the two subunits arose by fracturing of the coding sequence of a much larger ancestral gene. Apart from insertions of nonrepetitive sequence elements (described below), the individual repeats of α and β spectrin are distinguished from one another mainly by variations in their amino acid sequence. One exception is segment 16 of β spectrin (the fifteenth repeat; shaded in Figure 1(a)), which has been identified as the binding site for ankyrin.

There are several nonrepetitive sequence elements in the α and β spectrin subunits. The N-terminal region of β spectrin (~300 amino acids; segment 1) contains the actin binding site. The actin binding domain includes two copies of a calponin homology (CH) actin binding motif, which is present in many different cytoskeletal proteins that bind to filamentous actin. The C-terminal region of most β spectrin isoforms (~200–300 amino acids; segment 19) includes a pleckstrin homology domain (PH domain). Since the PH domain occurs downstream of the tetramer formation site, and since it is thought to interact directly with the plasma membrane, it is drawn here as a globular projection from the long axis of the spectrin molecule. The PH domain is likely to correspond to the small globular projections observed by electron microscopy of rotary shadowed specimens of native spectrin. However, the major splice variant of β spectrin expressed in human erythrocytes lacks that domain and consequently is small (2137 amino acids) compared with most other β spectrins (>2300 amino acids). The α spectrin subunit (2429 amino acids in human αI spectrin) includes an Src homology 3 (SH3) domain (~75 amino acids long; segment 10) inserted between the α-helical barrels of the ninth repeat.

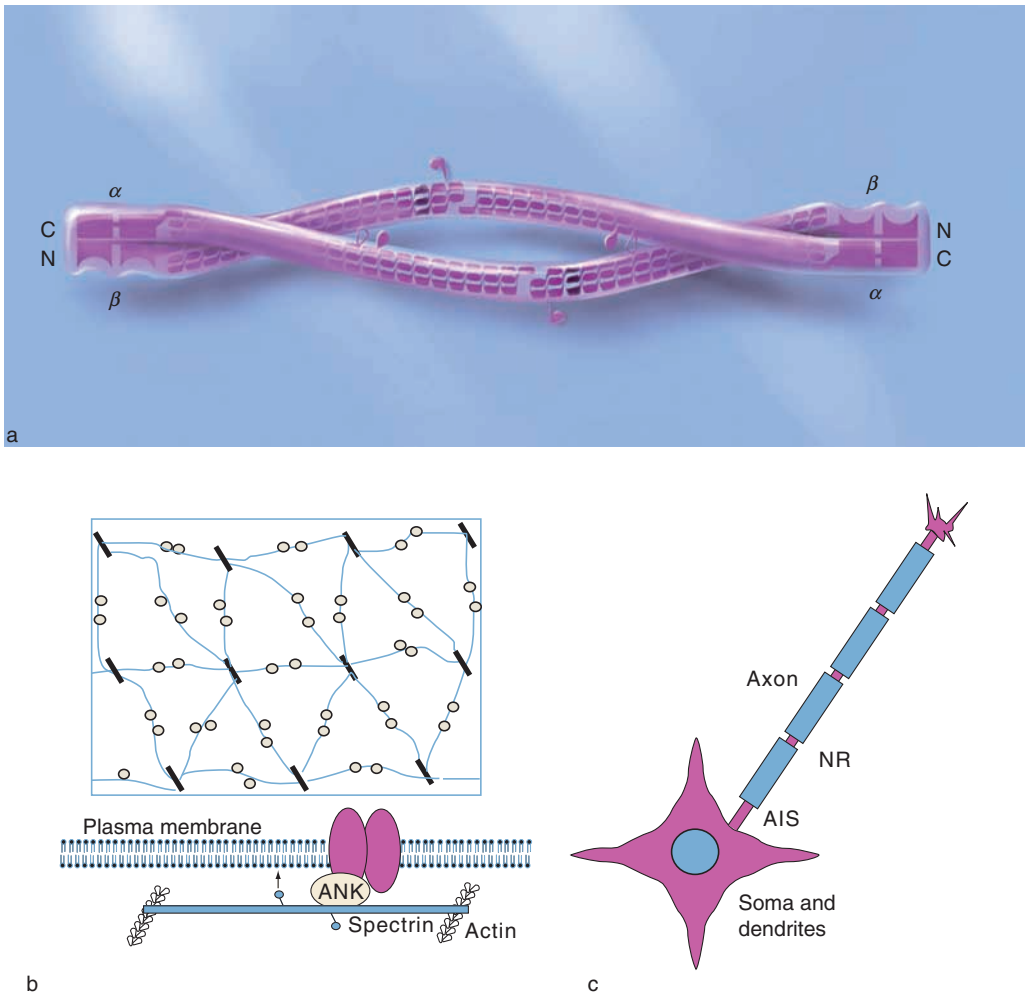

Figure 1 Spectrin structure, its interactions with the plasma membrane, and its distribution in neurons. (a)The spectrin tetramer. The typical spectrin molecule is a tetramer composed of two copies of the α and β subunits. Most of the sequence of each subunit consists of repetitive structural segments that are each ~106 amino acids long and folded into a three-barrel structure. A partial repeat near the C-terminus of β spectrin (2 barrels) associates with a partial repeat at the N-terminus of α spectrin (1 barrel) at the tetramer formation site near the center of the molecule. The sixteenth segment of β spectrin (shaded) binds to ankyrin. The C-terminal segment of β spectrin includes a pleckstrin homology (PH) domain, shown projecting from the spectrin tetramer near the center. The N-terminal domain (segment 1) of β spectrin includes 2 calponin homology domains that mediate binding to actin filaments. The adjacent C-terminal domain of α spectrin (segment 22) includes two EF hand domains that are aligned with the actin binding site. An SH3 domain (segment 10) is inserted between α-helical barrels of repetitive segment 9 of α spectrin. A loop containing a calmodulin binding site and a calpain cleavage site is found in segment 11 of α spectrin. (b) Spectrin in the human erythrocyte is arranged in a hexagonal submembrane network through interactions with short actin filaments. Ankyrin (ANK) serves as an adapter that links spectrin to the cytoplasmic domain of the major integral membrane protein in erythrocytes, band 3. Additional interactions link nonerythroid spectrins to the plasma membrane via the PH domain (arrow) and other ankyrin-independent membrane binding sites (not shown) (c) Spectrin is localized within discrete domains of the neuronal plasma membrane. Some spectrin isoforms are segregated to the soma and dendrites, and others are found in the axon and nerve terminal. βIV spectrin is concentrated within the axon initial segment (AIS) and the node of Ranvier (NR), and βII spectrin is concentrated within paranodal contacts.

The C-terminal domain of α spectrin (~155 amino acids; segment 22) includes two copies of an EF hand calcium binding motif. The EF hand domain is thought to be intimately associated with the N-terminal actin binding domain of β spectrin. The nonerythroid αII spectrin in humans is somewhat longer than αI (2472 amino acids), primarily because of inserted sequences found within segment 11 (described in the section titled 'Calpain cleavage site/calmodulin binding domain').

The Spectrin Gene Family

Spectrin is related to a large superfamily of actin cross-linking proteins that all share the CH domain actin binding motif. A closer-knit spectrin family of

proteins shares the same general domain structure and molecular shape as human erythrocyte spectrin. In humans there are five known β spectrin genes and two α genes. The αI and βI spectrin genes are primarily expressed in erythrocytes but are also expressed in muscle and some cells in the brain. The remaining α and β products are expressed in other cell types, such as neurons and epithelia. Invertebrates such as the fruit fly have a smaller repertoire of genes, including a single α spectrin that forms tetramers with two different β subunit gene products. The overall domain structures of spectrins from vertebrates and invertebrates are nearly identical to one another, with sequence identities in the range of 42–63%, suggesting that spectrins in different species are likely to have overlapping functions. Another, more distant relative of spectrin is the protein dystrophin, whose defects are responsible for Duchenne muscular dystrophy. While dystrophin is related to β spectrin, it appears to have evolved independently and is not known to associate with any other subunits or to form tetramers.

Spectrin Function

Mutations in the genes encoding erythrocyte α and β spectrin cause inherited anemias such as hereditary spherocytosis and hereditary elliptocytosis. Red blood cells from these patients have an abnormal shape and are unusually fragile, which ultimately causes anemia. These observations led to the conclusion that spectrin has important roles in cell shape and membrane stability.

Based on similarities to erythrocyte spectrin, the 'nonerythroid' forms of spectrin are also thought to contribute to the structure and stability of the plasma membrane. However, there are also some important differences between spectrins, and these differences led to an expanded model for spectrin function. First, whereas spectrin uniformly coats the plasma membrane in erythrocytes, spectrin is restricted to specialized membrane subdomains in most other cell types, such as the lateral domain of epithelial cells or the node of Ranvier in neurons. Second, there are only limited interactions between the spectrin cytoskeleton and integral membrane proteins in the erythrocyte. In contrast, nonerythroid spectrins are attached via ankyrin to many different integral membrane proteins. It is interesting that many of these proteins are involved in membrane transport (e.g., the Na,K ATPases, voltage-dependent sodium channels, potassium channels, and Na/Ca exchangers). Based on these observations, spectrins are thought to be involved in the establishment of plasma membrane polarity. Once assembled within a specific region, spectrin is thought to capture and stabilize other interacting membrane proteins to form a membrane domain with specialized composition and function.

Functional Sites in the Spectrin Molecule

Actin Binding Domain

The actin binding domain is the site of greatest amino acid sequence conservation among spectrins, presumably because of the importance of actin binding activity to spectrin function. This region of spectrin also interacts with other proteins, including members of the protein 4.1 family and synapsin. It is important to note that spectrin does not codistribute with bulk actin in cells. Instead, a subpopulation of actin filaments associates with spectrin to help form a plasma membrane-associated spectrin scaffold. A potential regulatory mechanism for the interaction of spectrin with actin is suggested by the behavior of the related actin cross-linker α actinin. There is a similar juxtaposition of actin binding and EF hand domains, and calcium binding to the EF hands of nonmuscle α actinin is known to downregulate the actin binding activity of the adjacent actin binding domain. It remains to be established whether the EF hand domain of spectrin modulates binding to actin.

Ankyrin Binding Domain

Most spectrins, with the exceptions of βV spectrin in mammals and β_H spectrins in invertebrates, possess binding sites for the adapter protein ankyrin. Ankyrin includes an N-terminal domain composed of 24 ankyrin repeats. The repeat domain is responsible for many of the known interactions between ankyrin and integral membrane proteins. Another domain of ankyrin, downstream of the membrane binding domain, has been shown to mediate binding to spectrin. *In vitro* binding assays were used to map the binding site for ankyrin to the sixteenth segment of β spectrin (sometimes referred to as the fifteenth repeat). Studies in mammalian heart suggest that spectrin targeting occurs through its interaction with ankyrin at this site, placing spectrin downstream of ankyrin. However, other evidence (described in the section titled 'PH domain') places ankyrin assembly downstream of spectrin.

A remarkable diversity of membrane proteins have been found to interact with the spectrin cytoskeleton through ankyrin. Many of these proteins are membrane transporters and channels that have important physiological activities in excitable cells such as neurons and cardiomyocytes. It is also remarkable that in several cases, defects in ankyrin and spectrin have phenotypic consequences as dire as defects in the

transporters themselves. Another class of ankyrin-interacting membrane proteins is the L1 family of neural cell adhesion molecules. L1 family members have diverse roles in axon fasciculation, axon guidance and neurite extension, synaptogenesis, and myelination. Mutations in human L1 are implicated in a host of symptoms, including mental retardation and hydrocephalus. Some of the initially described characteristics of the L1 family–ankyrin interaction made L1 proteins seem promising candidates in regulating spectrin cytoskeleton assembly. However, subsequent genetic analysis of ankyrin and spectrin mutants has suggested that L1 family members, like the other ankyrin binding membrane proteins, are downstream interacters that depend on spectrin and ankyrin for their accumulation and function.

Intersubunit Interactions

Two main classes of subunit interactions have been characterized in detail. One is found near the ends of the spectrin tetramer, mediating lateral interactions between terminal repeats of the α and β subunits. The second class is found near the center of the tetramer, where a partial repeat from β spectrin (helices A and B) interacts with a partial repeat from the N-terminus of α spectrin (helix C) to form an intact triple-barrel α-helical structure resembling the other spectrin repeats.

PH Domain

The PH domain found at the C-terminus of β spectrin resembles a broadly conserved sequence motif found in hundreds of other proteins. Many, but not all, PH domains have been shown to possess phosphoinositide binding activity and to participate in signal transduction pathways. The PH domain of β spectrin has been shown to bind to phosphatidylinositol bisphosphate and when expressed as a fusion protein with green fluorescent protein, was shown to have plasma membrane targeting activity *in vitro*.

The PH domain corresponds to one of the ankyrin-independent membrane binding sites identified in biochemical studies using purified spectrin and NaOH-stripped brain membrane vesicles. Additional binding sites were identified near the N-terminus of β spectrin. A recent genetic study in *Drosophila* demonstrated that the PH domain is required for targeting of spectrin to the plasma membrane in midgut epithelial cells. Whether targeting occurs through a lipid or protein target remains to be determined. Surprisingly, ankyrin binding activity was not required for targeting of spectrin to the plasma membrane. The same study revealed that another ankyrin-independent site in the β subunit was responsible for targeting spectrin in the salivary gland epithelium. That site may correspond to the N-terminal membrane association domain identified in biochemical studies. This line of evidence supports a model in which polarized assembly of spectrin is mediated by its direct interactions with membrane lipids and/or proteins, with effects on ankyrin and interacting membrane proteins occurring downstream.

SH3 Domain

The SH3 domain is another conserved sequence module found in many different signal transduction proteins. It was first identified in the Src oncogene. The SH3 domain provides another class of potential protein interaction sites in the spectrin molecule, and proteins that interact with the spectrin SH3 domain have been identified.

Calpain Cleavage Site/Calmodulin-Binding Domain

Calcium-dependent protease with papainlike activity (calpain) belongs to a family of cytoplasmic cysteine proteases that are activated by calcium ions and translocate to the plasma membrane. Calpain cleaves a range of substrates including spectrin and other membrane-associated proteins. The observations that spectrin is highly enriched in brain and undergoes calpain-mediated proteolysis in neurons have led researchers to use spectrin as a readout for the physiology and pathology of neuronal cells. A calpain cleavage site has been mapped between Tyr1176 and Gly1177 of segment 11 of α spectrin. Cleavage occurs near the middle of the αII spectrin polypeptide, and the cleavage products are easily identified in Western blots probed with sensitive antispectrin antibodies. Cleavage occurs rapidly in response to physiological stimuli, as in *N*-methyl-D-aspartate (NMDA)-induced excitotoxicity in which cleavage products can be detected as soon as 2 min after exposure to NMDA. While the physiological role of α spectrin cleavage remains to be established, it has proven to be a valuable and widely used marker of calpain activation.

The calpain-cleavage site in αII spectrin occurs within a 35 amino acid loop inserted between α-helical domains of segment 11 (**Figure 1(a)**). Neither the loop nor the cleavage site is found in human αI erythroid spectrin. This loop is also one of the main structural differences between mammalian αII spectrin and invertebrate α spectrin. The loop also includes a high-affinity calmodulin-binding site, and calmodulin binding enhances the susceptibility of spectrin to calpain cleavage. This calmodulin binding activity is absent in both αI spectrin and invertebrate α spectrin, suggesting that it too is a product of a relatively recent step in spectrin evolution.

EF Hand Domain

The C-terminal EF hand domain of α spectrin resembles the calcium-binding EF hand domain of the related protein α actinin. In both proteins, the EF hand domain is thought to interact with the adjacent actin-binding CH domain on the neighboring subunit. A mutation in *Drosophila* α spectrin that truncates the C-terminal 100 amino acids leads to essentially the same lethal phenotype as a null mutation, suggesting that the EF hand domain is critical to the structure and function of spectrin. The effect of the mutation is likely to be mediated through its effect on the adjoining actin-binding domain on the β subunit.

Distribution of Spectrin in Neurons

Soon after the discovery of spectrins in the nervous system, it became apparent that different isoforms of spectrin are differentially distributed within individual neurons. As a rule, antibodies recognizing erythroid forms of spectrin stained cell bodies and dendrites and antibodies recognizing 'nonerythroid' spectrins stained axons and presynaptic nerve terminals (Figure 1(c)). These observations provided the first clues that distinct targeting mechanisms acted on different isoforms of spectrin and raised the possibility that there were isoform-specific functions as well.

The nonerythroid form of spectrin in neurons was originally named fodrin, before its relationship to erythrocyte spectrin was recognized. In their elegant initial description, Levine and Willard established that fodrin was axonally transported. Now, three different 'nonerythroid' isoforms of spectrin appear to be present in axons. So far, isoforms containing the βIV subunit have been characterized in the greatest detail. The βIV spectrin is found in myelinated neurons, where it colocalizes with ankyrinG at axon-initial segments and the node of Ranvier in both central and peripheral neurons (Figure 1(c)). A membrane undercoat is visible at the node of Ranvier, and the undercoat was greatly diminished in neurons from βIV spectrin mutant mice, suggesting that spectrin forms a structure analogous to the erythrocyte membrane skeleton. It is interesting that loss of βIV spectrin was also associated with abnormal membrane shape and vesiculation in the vicinity of the node. A similar complex of αIIβII spectrin with ankyrin-B was recently identified in the paranodal region of central and peripheral nerves. The βIII isoform of spectrin is also present in axons and elsewhere, including intracellular sites such as the Golgi apparatus. The remaining β spectrin isoform, βV, was detected in the outer segments of photoreceptor rods and cones but is present at very low levels in mammalian brain.

Genetic Studies of Spectrin Function in the Nervous System

Invertebrates

Phenotypes have been analyzed for mutations in all three of the spectrin genes in *Drosophila* and *C. elegans*. Knockouts of α and β spectrin are lethal, establishing the essential function of spectrin. While the exact cause of death is not known, several defects have been described in the nervous system. First, physiological studies established that spectrin mutants have a presynaptic defect that blocks synaptic vesicle release. That study did not detect a significant effect of spectrin mutations on postsynaptic function at the embryonic neuromuscular junction. However, a subsequent study of neuromuscular junctions later in development revealed that spectrin is necessary for long-term stability of the synapse. Recently a defect in axonal pathfinding was observed in *Drosophila* β spectrin mutants. It is interesting that the spectrin mutation was phenotypically similar to defects in Robo and Slit, a ligand and receptor mediating repulsive guidance during development of the embryonic central nervous system. There was no apparent effect of the β spectrin mutation on Robo expression, and the deficit was not corrected by expression of β spectrin in glial cells which normally express the Slit ligand. Thus the β spectrin mutation may exert its effect on a downstream component of the Robo signaling pathway in neurons.

Human βIII Spectrin

Sequencing of candidate mutations revealed that defects in the βIII spectrin gene are responsible for human spinocerebellar ataxia type 5. Three mutations were identified in three independent lineages. Two affected repetitive segment 4, and the other affected the actin binding domain (Figure 1(a)). It is interesting that one of the mutations was traced through the pedigree of Abraham Lincoln, raising the possibility that the former US president may have suffered from this disease. βIII spectrin was originally described as a novel spectrin isoform associated with intracellular vesicles and the Golgi apparatus in tissue culture cells. However, less is known about the subcellular distribution of the protein *in vivo*. A yeast-2-hybrid study detected βIII spectrin in a screen for proteins that interact with the glutamate transporter EAAT4. The effects of the spinocerebellar ataxia mutations on EAAT4 behavior were assayed biochemically and in a tissue culture cell assay for protein mobility in the plasma membrane. The βIII spectrin mutations did not appear to affect the quantity of EAAT4 in autopsy samples of brain tissue, but the extraction properties of EAAT4 were

altered, suggesting a possible change in its subcellular distribution or perhaps altered interactions with other brain proteins. The mobility of glutamate transporters in the plasma membrane of HEK293 cells was greatly reduced with co-expression of wild-type βIII spectrin but not with the mutant version of spectrin. These results suggest that βIII spectrin has effects on interacting membrane activities at the plasma membrane, beyond any additional effects it may have within intracellular compartments.

Murine βIV Spectrin

Studies in the mouse established that the βIV subunit of spectrin codistributed with ankyrin-G at the node of Ranvier and at the axon initial segment. Truncations in mouse βIV spectrin were shown to be responsible for the mutant phenotype in quivering mice which exhibit ataxia, tremor, and deafness. Studies of βIV spectrin knockout mice revealed that βIV spectrin interacts with ankyrin-G and that neither ankyrin-G nor voltage-gated sodium channels were correctly targeted to the axon initial segment or node of Ranvier in the mutants. It is interesting that knockout studies also demonstrated that βIV spectrin was mislocalized in ankyrin-G mutant mice, indicating that the targeting of ankyrin and spectrin was mutually dependent. The phenotype of quivering mice appears to be caused by the effect of the βIV spectrin mutation on voltage-dependent sodium channel targeting or stability. These observations suggest that βIV spectrin and ankyrin exert their effects on interacting membrane proteins in a coordinated fashion.

There also appear to be multiple isoforms of βIV spectrin. In fact, the 'conventional'-size βIV product (288 kDa) was a relatively minor component relative to an unusual variant (~150 kDa) in which the N-terminal half (including the actin binding domain and ten repeats) was absent. Yet the phenotype observed at the node of Ranvier appeared similar whether knockouts affected all βIV spectrin isoforms or just the largest isoform. These observations raise the possibility that the novel small isoform may have unique functions that are distinct from conventional β spectrins.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Development of Drosophila Neuromuscular Junctions; Dystrophin, Associated Proteins, and Muscular Dystrophy; Fragile X Associated Tremor Ataxia Syndrome; Friedreich Ataxia; Neural Cell Adhesion Molecules and Synapse Regulation; Neuromuscular Junction (NMJ): Mammalian Development; Neuromuscular Junction (NMJ): Inherited and Acquired Disorders.

Further Reading

Bennett V and Baines AJ (2001) Spectrin and ankyrin-based pathways: Metazoan inventions for integrating cells into tissues. *Physiological Reviews* 81: 1353–1388.

Czogalla A and Sikorski AF (2005) Spectrin and calpain: A 'target' and a 'sniper' in the pathology of neuronal cells. *Cellular and Molecular Life Sciences* 62: 1913–1924.

Das A, Base C, Dhulipala S, and Dubreuil RR (2006) Spectrin functions upstream of ankyrin in a spectrin cytoskeleton assembly pathway. *Journal of Cell Biology* 175: 325–335.

Dubreuil RR (2006) Functional links between membrane transport and the spectrin cytoskeleton. *Journal of Membrane Biology* 211(3): 151–161.

Featherstone DE, Davis WS, Dubreuil RR, and Broadie K (2001) Drosophila alpha and beta spectrin mutations disrupt presynaptic neurotransmitter release. *Journal of Neuroscience* 21: 4215–4224.

Garbe DS, Das A, Dubreuil RR, and Bashaw GJ (2007) beta-Spectrin functions independently of Ankyrin to regulate the establishment and maintenance of axon connections in the Drosophila embryonic CNS. *Development* 134(2): 273–284.

Ikeda Y, Dick KA, Westherspoon MR, et al. (2006) Spectrin mutations cause spinocerebellar ataxia type 5. *Nature Genetics* 38: 184–190.

Komada M and Soriano P (2002) betaIV-Spectrin regulates sodium channel clustering through ankyrin-G at axon initial segments and nodes of Ranvier. *Journal of Cell Biology* 156: 337–348.

Lacas-Gervais S, Guo J, Strenzke N, et al. (2004) BetaIVSigma1 spectrin stabilizes the nodes of Ranvier and axon initial segments. *Journal of Cell Biology* 166: 983–990.

Levine J and Willard M (1981) Fodrin: Axonally transported polypeptides associated with the internal periphery of many cells. *Journal of Cell Biology* 90: 631–643.

Lux SE and Palek J (1995) Disorders of the red cell membrane. In: Handin RI, Lux SI, and Stossel TP (eds.) *Blood: Principles and Practice of Hematology*, pp. 1701–1818. Philadelphia: Lippincott.

Matteis MAD and Morrow JS (2000) Spectrin tethers and mesh in the biosynthetic pathway. *Journal of Cell Science* 113: 2331–2343.

Mohler PJ and Bennett V (2005) Ankyrin-based cardiac arrhythmias: A new class of channelopathies due to loss of cellular targeting. *Current Opinion in Cardiology* 20: 189–193.

Ogawa Y, Schafer DP, Horresh I, et al. (2006) Spectrins and ankryinB constitute a specialized paranodal cytoskeleton. *Journal of Neuroscience* 26: 5230–5239.

Pan Z, Kao T, Lemos ZHJ, et al. (2006) A common ankyrin-G-based mechanism retains KCNQ and Nav channels at electrically active domains of the axon. *Journal of Neuroscience* 26: 2599–2613.

Pielage J, Fetter RD, and Davis GW (2005) Presynaptic spectrin is essential for synapse stabilization. *Current Biology* 15: 918–928.

Pielage J, Fetter RD, and Davis GW (2006) A postsynaptic spectrin scaffold defines active zone size, spacing, and efficacy at the Drosophila neuromuscular junction. *Journal of Cell Biology* 175: 491–503.

Speech Perception: Adult

L L Holt, Carnegie Mellon University, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

As adult listeners, we comprehend the messages conveyed by spoken language with little effort. However, the ease of everyday conversation masks the perceptual and cognitive complexities involved in perceiving speech. Upon examination of the acoustic speech signal, it becomes immediately clear that this everyday skill is a demanding perceptual task.

A spoken syllable may persist in the world for a mere tenth of a second. Yet, as adult listeners, we are able to gather a great deal of information from these fleeting acoustic signals. We may apprehend the physical location of the speaker, the speaker's gender, regional dialect, age, emotional state, or identity. These spatial and indexical factors are conveyed by the acoustic speech signal in parallel with the linguistic message of the speaker. Although the spatial and indexical characteristics of the spoken message are of much interest in their own right, speech perception most commonly refers to an understanding of how the physical patterns of acoustic energy evoke perception of a linguistic message. Thus, the study of speech perception concerns the mapping between a time-varying acoustic signal received from a talker and linguistic meaning.

Units of Speech Perception

Some accounts of speech perception have suggested the possibility that adult listeners map the acoustic signal directly onto lexical entries, the words of the language. However, most models of language comprehension assume that the acoustic speech signal is translated into an intermediate unit that is used to access the words forming the lexicon. There has been much debate regarding the best characterization of these intermediate units and, for the most part, the possibilities have related closely to the natural classes described by linguistic theory, whether distinctive features, phonemes, or syllables. Despite these theoretical distinctions, the vast majority of research in speech perception has focused on the mapping from acoustics to phoneme or phonetic category as in the distinction between /b/ and /p/ in 'bin' versus 'pin.' It is likely that the concepts understood from this research apply similarly to other possible basic units. At all levels thus far investigated in-depth, research has demonstrated that the mapping from acoustics is very complex.

Context in Encoding

The complexities of speech perception begin with the nature of the acoustic signal. A great deal of early research documented the considerable variability inherent in the acoustics of speech. To summarize this broad literature very briefly, there do not appear to be acoustic signatures that uniquely specify phonetic categories, syllables, or words. Thus, listeners are faced with the perceptual challenge of mapping highly variable acoustic signals onto speech units in a many-to-one manner.

The variability inherent in the acoustics of speech arises from multiple sources. One notable source of acoustic variability is due to demands on speakers. Part of the efficiency of spoken language is the fluent nature with which it may be expressed. Speakers produce 10–15 consonants or vowels per second at typical rates of speech, and they may even double this rate in hurried speech. The rapid nature of speech production affects the acoustic signal that reaches the ear of a listener. In fluent speech, speakers articulate sounds such that movements of the vocal tract overlap in space and time across the production of different consonants and vowels. This overlap is known as coarticulation; its result on the acoustic speech signal is to cause the information for speech units to be spread across the acoustic signal. Phonetic elements of a spoken word (e.g., the three sounds in 'bin,' /bIn/) are not presented discretely as are the letters of a written word. Rather, the acoustic information for phonetic categories overlaps within the acoustic signal such that any particular time slice of an acoustic speech signal may contain information about multiple vowels and consonants. For example, the particular acoustic signature of a consonant /g/ as in 'gum' is greatly influenced by whether it is preceded by a word such as 'feel' or 'fear,' as in 'feel gum' and 'fear gum.' To describe this simply, consider the placement of the tongue body in these articulations. In articulating 'feel gum,' the tongue must move quickly from a more anterior position in the /l/ of 'feel' to the more posterior position of the /g/ in 'gum.' The resulting articulation is a compromise between where the tongue has been and where it is going. This is in part a consequence of physical constraints such as mass and inertia on the articulators, but speakers may also actively adjust articulation.

In the example, the result is that the articulators do not reach quite the same position in articulating the /g/ in 'gum' when it is preceded by 'feel' as when it is preceded by 'fear.' As a result, the acoustic signature of the /g/ in 'gum' is quite different across these two contexts. One issue for speech perception is how the perceptual system recognizes the functional equivalence of these two acoustically distinct /g/ sounds.

In addition to coarticulation, other sources also contribute to variability of the acoustic speech signal. Rate of speech, speaker idiosyncrasies, dialect, accent, emotion, gender, and room reverberance influence the acoustics of an utterance. One of the most significant challenges for understanding speech perception is that despite the radical effect these diverse sources of variability have on the acoustic signal, listeners somehow identify speakers' intended meaning. There has been considerable research investigating how listeners accomplish this feat. At the broadest level, it appears that listeners overcome the perceptual obstacles of acoustic variability by perceiving speech in a wholly context-dependent manner. The same speech signal may be perceived as belonging to different phonetic categories when it is heard in different contexts. Consider the example discussed previously. When listeners are presented with a set of speech stimuli that have been acoustically manipulated to vary perceptually along a /g/ to /d/ series as in 'gum' to 'dumb,' the endpoints of the series are perceived consistently as /g/ and /d/. The intermediate stimuli are perceptually ambiguous, being identified as both /g/ and /d/ some of the time. The boundary is steep, as is typical of identification of consonants. This pattern of perception is illustrated in **Figure 1** by the green line.

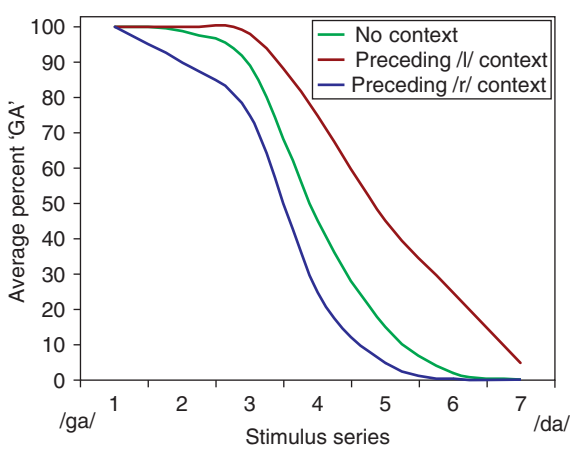

Figure 1 Listners identification of consonants varying perceptually from 'ga' to 'da' changes as a function of the speech that precedes it. A preceding 'l' shifs identification toward more 'ga' identifications whereas a preceding 'r' shifs identification of the same sounds toward 'da'. Data after Mann VA (1980) Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics* 28: 407–412.

Many experiments have demonstrated that listeners' perception of such sounds is influenced greatly by the preceding context. Preceded by /l/ as in 'feel,' listeners tend to label the sounds along the /g/ to /d/ series more often as /g/ as in 'gum' (**Figure 1**, red line). The same sounds are more often labeled as /d/ when /r/ is the context stimulus (**Figure 1,** blue line). Thus, physically identical acoustic speech stimuli are perceived very differently as a function of adjacent context.

The association between such phonetic context effects and coarticulation is intimate. Notice that the direction of these phonetic context effects in perception is opposite the coarticulatory effects on speech articulation described previously. In the utterance 'feel gum,' the /g/ is coarticulated such that it is articulated at a more anterior position than it is in isolation. This causes its acoustics to be more similar to the anterior articulation typical of /d/ in isolation; coarticulation of 'feel gum' results in a more /d/-like articulation of /g/. However, perception pulls in the opposing direction. A perceptually ambiguous /g/–/d/ speech token is pulled perceptually toward /g/ when preceded by an /l/ precursor, as in 'feel.' Thus, this contrastive perception works in opposition of coarticulation, in essence compensating for the assimilatory coarticulatory effects that draw neighboring utterances to be more similar to one another. Coarticulation causes /g/ to be acoustically more similar to /d/ in the context of a preceding /l/, whereas speech perception of consonants in the context of /l/ is shifted so that syllables are perceived more often as 'g' than 'd.' Context effects of this sort are ubiquitous in speech perception and have been documented for many speech segments. Moreover, they are observed in infants; therefore, neither experience producing coarticulated speech nor a well-developed lexicon appears to be prerequisite. Perhaps even more surprising, the influence of /l/ and /r/ on consonant perception is evident even among adult listeners for whom /l/ and /r/ are not phonologically distinct. Japanese listeners unable to accurately distinguish English /l/ from /r/ exhibit a phonetic context effect of the same magnitude as native English listeners. Listeners need not be able to classify the context syllables into distinct native phonetic categories (or even be able to hear a difference between precursors) for them to have an influence on perception of following speech.

Of interest in understanding the mechanisms that give rise to such effects, Japanese quail (*Coturnix coturnix japonica*) trained to peck a lighted key in response to presentation of /g/ endpoints of a /g/ to /d/ stimulus series pecked more vigorously to novel ambiguous midseries speech stimuli when they were preceded by /l/. A second set of birds trained to peck

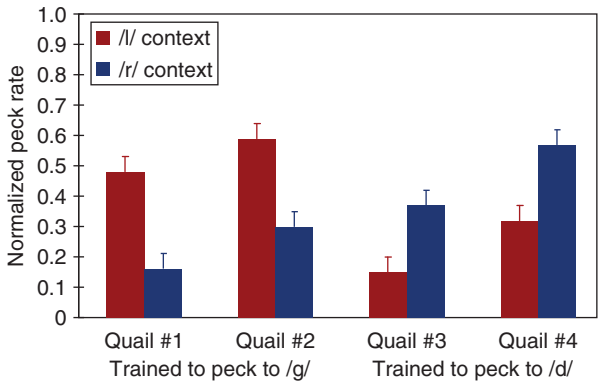

Figure 2 Ovail trained to peck to 'g' for 'd' sounds exhibit context-dependent pecking when 'l' and 'r' precede the target sounds. The influence of context is the same as that plotted for human listeners in Figure 1. Data from Lotto AJ, Kluender KR, and Holt LL (1997) Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America* 102: 1134–1140.

in response to /d/ responded more robustly to the same novel stimuli when they were preceded by /r/. Thus, as illustrated in Figure 2, birds exhibit context-dependent shifts in responses contingent on a speech precursor, and the directionality of this effect is the same as that observed for human adults and infants. The extension of phonetic context effects to a nonhuman species suggests that mechanisms specific to human speech may not be necessary to accommodate the complex variability present in speech. To pose the results in a general way, context sounds with higher frequency acoustic energy (e.g., /l/) shift perception of the following syllable toward the alternative with greater low-frequency energy, /g/. Thus, processes by which the auditory system exaggerates change in the acoustic signal may be sufficient to explain perceptual compensation for coarticulation. This conclusion is supported by research demonstrating that adult human listeners shift phonetic categorization responses not only as a function of neighboring speech contexts but also as a function of nonspeech tones, chirps, and noises that precede or follow speech. In the case of human and nonhuman perception of speech and nonspeech contexts, speech perception appears to be relative to and contrastive with the acoustics of context sounds, whether speech or nonspeech.

Thus, there is evidence that some of the perceptual challenges of the variability within the acoustic speech signal may be met with general perceptual processes that produce contrast. Experiments investigating listeners' sensitivity to distributional regularity in sentence-length acoustic contexts indicate that such contrastive processing may also provide a means of adjusting perception in response to variability in the acoustic speech signal introduced by different speakers. Overall, these results indicate that listeners retain information about context stimuli across multiple timescales and adjust speech perception in relation to this information. The exact nature of the mechanisms by which listeners retain context across multiple timescales remains to be determined, but empirical findings suggest that the mechanisms may rely on perceptual processing general to speech and nonspeech acoustic signals (and, indeed, available to even quail) rather than processing specialized to accommodate acoustic complexities of the speech signal.

Speech and Expertise

A great deal of research has been directed at identifying invariants, unique acoustic cues indicative of a particular phonetic category. However, despite these attempts, it appears that phonetic categories, by and large, are signaled by multiple cues that vary with the context of the spoken utterance, with no single cue being necessary or sufficient to signal phonetic identity. The case of English /b/ versus /p/ is illustrative: 16 distinct acoustic cues have been cataloged as differentiating these two categories in syllable-initial position. Recalling discussion in the previous section, it is also the case that acoustic cues signaling a particular phonetic category are quite dependent on the context in which the acoustic signal is presented. However, the existence of multiple acoustic cues to category identity does not imply that all of these cues are equally informative to listeners. For example, both spectral and temporal cues differentiate the English vowels /i/ and /I/, but listeners rely much more on the spectral cue than the temporal cue in categorizing these vowels. The precise nature of perceptual weighting of acoustic cues appears to depend, at least in part, on experience with the distribution of cues within the native language. There is evidence that the developmental trajectory of this learning is protracted; preschool-aged children apply very different perceptual weights to acoustic cues than do adults. Insight into how listeners learn to weight acoustic cues differentially may come from studies of other complex sounds, such as music and novel nonspeech categories, for which there is also evidence that listeners learn to weight acoustic cues based on cue distributions in the input.

The details of the acoustic distributions of the native language community play a very significant role in adult speech perception. In the first year of life, well before infants are speaking, they are already beginning to perceive speech in a native language-specific manner. As mentioned previously, this attunement to the native language continues into later childhood before reaching adult-like performance

and may, some argue, require learning to read to produce the kinds of phonetic categorization typically studied in investigations of adult speech perception. Among adults, perception of the acoustic speech signal is very significantly influenced by experience with the native language. Identical acoustic signals may be perceived very differently by listeners with different language or even dialect backgrounds. The classic example of this was alluded to previously. Japanese does not make a phonological distinction between the English sounds /r/ and /l/. Experience with the structure of Japanese shapes the way that adult Japanese native listeners hear speech such that many Japanese listeners are unable to distinguish English /r/ from /l/ even following significant training. Expertise with a native language has a strong influence on the manner by which speech is perceived. The relative interaction of native language experience with perception of nonnative sounds depends significantly on the degree to which the nonnative sounds overlap with native sounds in perceptual space. Understanding these phenomena and investigating the degree to which adult listeners may or may not exhibit plasticity in learning second-language sounds are active areas of research.

This is a significant challenge because adult listeners are experts in perceiving speech, having spent the greater part of their waking hours engaged in perception of speech and being exposed to the speech signal since their later prenatal months. Exposure to the acoustic systematicities of a language is thought to organize perception in terms of phonetic categories or equivalence classes. Due to the acoustic variability of speech, there is no one-to-one mapping between acoustics and phonetic category. The mapping of a set of variable acoustic signatures onto a common phonetic label (or syllable or word), as described previously, is an example of categorization; physically distinct acoustic signals may be treated by listeners as instances of the same phonetic category. These categories appear to have graded internal structure such that utterances of a particular vowel or consonant may be perceived to be more or less representative of the category. Despite general agreement on these observations, it has proven difficult to determine the precise learning mechanisms by which this organization occurs because adults, and even young infants, possess a great deal of phonetic experience. Without characterization or control of the acoustic distributions of this experience, it is difficult to tease apart how experience with different distributions of speech input may shape phonetic categorization. It is possible to observe that experience with different native languages produces very different phonetic organizations, but the means by which this occurs is not well understood. Several empirical inroads to these issues have been made by investigating learning in nonhuman animal models of speech perception and by investigating the boundaries and constraints on human adults' general ability to categorize complex nonspeech sounds. The findings from this growing literature suggest that listeners (both human and nonhuman) are able to extract regularities from distributions of complex acoustic signals such that categorical responses may be made. For example, European starlings trained to respond to the vowels /i/ and /I/ (as in 'beet' and 'bit') respond to the vowels in a graded manner, as do human adults. The strength of the starling responses to individual vowels correlates very strongly with native English human adults' ratings of the 'goodness' of these vowels. Thus, a nonhuman species is capable of learning to respond categorically to speech stimuli and does so in a way that mirrors the graded internal structure of human adult perception.

The virtue of nonhuman animal models is that there is a greater possibility of experimental control over the histories of experience such that firm conclusions about the course of learning may be drawn. A complementary approach that shares this virtue is training human adults to categorize novel nonspeech acoustic stimuli such that the constraints imposed by auditory categorization, in general, may be understood. Investigations of nonhuman animal models and adult human novel sound category learning may play an important role in understanding phonetic categorization because of the experimental control afforded. A challenge will be to balance this control with the ecological validity of studies of infant phonetic category acquisition and nonnative adult listeners' phonetic categorization to work toward an understanding of how adult perceivers become expert with the speech signals of their language community and how this expertise shapes speech perception.

Categorical Perception

Categorical perception is perhaps the best-known pattern of speech perception. The speech series illustrated in Figure 1 serves as an example. Although the speech sounds along the series vary gradually in their acoustic characteristics, listeners' pattern of identification changes abruptly, not gradually; this sharp identification function is one of three hallmarks of categorical perception. A second defining characteristic of categorical perception is the pattern of discrimination across the series. When listeners discriminate pairs of stimuli drawn from the series, the resulting discrimination function is discontinuous. Discrimination is nearly perfect for stimuli that lie on opposite sides of the sharp identification boundary, whereas it is very poor for pairs of stimuli that are equally

acoustically distinct but lie on the same side of the identification boundary. The final characteristic of categorical perception is that identification performance predicts discrimination performance; speech sounds that are identified with the same label are difficult to discriminate, whereas those identified with different labels are discriminated.

Categorical perception was formerly thought to be a peculiarity of speech perception. However, categorical perception has since been observed for perception of human faces and facial expressions, music intervals, and artificial stimuli that participants learn to categorize in laboratory tasks. It is observed in the behavior of nonhuman animals as well. Rather than a speech-specific phenomenon, categorical perception is a far more general characteristic of how perceptual systems respond to experience with regularities in the environment.

Influences from Other Sources of Information

As described previously, the mapping between acoustic speech signal and intermediate representation (acoustics to phonetics, in most characterizations) is complex. It is also the case that this mapping may be influenced by higher-order linguistic information, such as the knowledge of words within the language. An ambiguous sound that might be perceived as /g/ or /k/, for example, is more likely to be identified as /g/ if it is followed by 'ift.' The same sound is perceived as /k/ if it is followed by 'iss.' The direction of this influence is to shift phonetic categorization toward the alternative that forms a real English word. The influence of lexical information on phonetic perception has been established in a variety of paradigms, and the mechanisms by which these effects arise are hotly debated. Specifically, there is significant controversy between interactive accounts, which posit that such effects arise via feedback from the lexicon that directly influences the mapping from acoustics to phonetics, and strictly bottom-up autonomous accounts, whereby the influence of lexical information on phonetic responses emerges only at a later decision stage and does not directly influence the mapping from acoustics to phonetics (Figure 3). As in other domains of cognitive science that are modeled by these two approaches, distinguishing between interactive and autonomous accounts in speech perception has been difficult because the two models make many similar predictions. However, there is a domain in which the predictions of the models depart quite radically. Interactive accounts posit that lexical information may influence prelexical processing directly via feedback from the lexicon to prelexical processing. In autonomous models, this kind of direct modulation is impossible because information from phonetic and lexical processing is combined only at a later decision stage. If lexical information influences a prelexical phenomenon, it may be taken as evidence for interactive models.

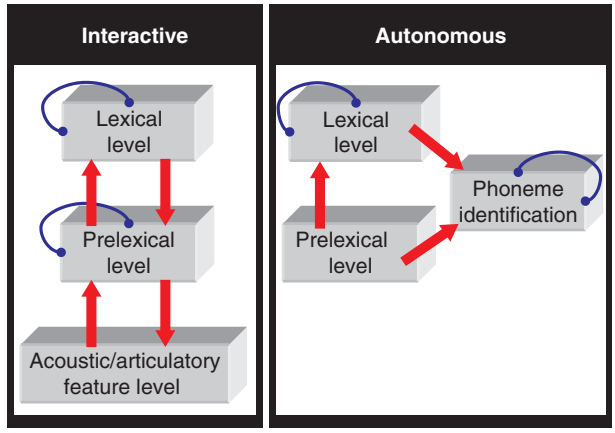

Figure 3 Schematic diagrams of interactive and autonomous models of speech perception.

There is one empirical paradigm in which this proposition has been tested, retested, and much debated. The phonetic context effects described previously are widely considered to arise from the mapping from acoustics to phonetic categories; that is, they are prelexical in the sense that they do not require information from the lexicon and operate at a level below that of words. Several lines of research have investigated whether lexical knowledge of the words of English influences phonetic context effects and thus whether lexical information modulates a prelexical effect, as predicted by interactive models. At the prelexical level, a preceding /s/ or /ʃ/ (as in 'sip' vs. 'ship') influences phonetic categorization of subsequent sounds as /t/ or /k/. Preceded by /s/, listeners more often categorize the speech as /k/, whereas preceded by /ʃ/, the same sounds are more often categorized as /t/. Of interest regarding the issue of the role of lexical processing in influencing prelexical effects, a single ambiguous speech sound between /s/ and /ʃ/ is heard as /ʃ/ if it is preceded by 'fooli__' but as /s/ if it is preceded by 'Christma__.' Knowledge of the English language, which includes the words 'foolish' and 'Christmas' but not 'fooliss' or 'Christmash,' serves to disambiguate the acoustically ambiguous speech sounds. Of critical importance, however, is the effect of this on listeners' perception of subsequent sounds as /t/ or /k/. When the ambiguous /s/–/ʃ/ sound is presented in the word contexts and precedes the /t/ to /k/ sounds, there is a shift in listeners' phonetic categorization such that preceded by 'Christma__' listeners respond /k/ more often,

whereas preceded by 'fooli__' the same sound is labeled as /t/. Thus, there is a phonetic context effect on the second sound that is driven by the lexical information, as predicted by interactive models. There has been considerable debate over findings such as these, resulting in replications and alternative conclusions more consistent with autonomous models. Recent investigations have tipped the scales back in favor of an interactive account because these effects have held up to empirical scrutiny.

Another source of influence on speech perception comes from the visual modality. Although speech perception is most often considered from the perspective of the acoustic signal, visual information from a speaker's face can also greatly influence perception of speech. The classic example of this is the McGurk effect, whereby an acoustic /ba/ and a visual /ga/ collaborate to produce a percept of /da/, unique from either unimodal source. A great deal of research illustrates the advantages of perceiving acoustic speech with synchronous, matched visual information of the face producing the speech. Speech presented in noise, for example, is more intelligible when a face producing the speech is visible. Some of the same issues and debates present in understanding why lexical influences on speech perception emerge in considering how the perceptual system integrates optic and acoustic information for speech perception; there are both autonomous and interactive models of the phenomenon. In addition, an important theoretical question that is under investigation is the degree to which auditory–visual integration for speech is a result of the two types of information sharing a common source or whether it is a result of learning about the natural covariance of these visual and auditory signals. This question is important in distinguishing among the theoretical accounts of speech perception. The question is very difficult to address because acoustic speech is intrinsically linked with the visual signals from articulating faces, but adult listeners also have vast experience with the covariation between articulating faces and acoustic speech. This has been investigated by examining whether listeners are able to learn to associate novel visual signals, not in any way associated with faces or with production of speech, with acoustic–phonetic categories. With sufficient training, listeners do, in fact, learn this relationship and, with learning, begin to integrate the novel visual information for the phonetic categories with acoustic speech. The phonetic category membership of an ambiguous acoustic stimulus is disambiguated by the newly learned visual signals. Moreover, the learning results in benefits to intelligibility of acoustic speech in noise when a newly learned visual signal associated with the phonetic category is paired with it. Further work is necessary to determine the relationship of speech expertise with auditory–visual integration in speech perception.

Theoretical Approaches

The challenges of speech perception have been addressed by multiple theoretical accounts of how listeners accommodate the complexities of the speech signal. Landmark studies detailing the relationship between speech acoustics and phonetic perception revealed the complex mapping. These data convinced many that this mapping was far too complex for acoustics to provide a plausible object for speech perception; the objects of speech perception were thought to need to be more or less invariant with phonetic categories. From this tenet arose the motor theory of speech perception. Given the radical variability in the mapping from acoustics to phonetic categories, the motor theory holds that the only tenable object of perception is the intended articulatory gesture of the speaker, as exemplified by the neuromotor command to the articulators. By this view, the objects of speech perception are the articulatory events of a speaker rather than the acoustic or auditory events presented to the listener. Another important aspect of the motor theory is that speech perception is uniquely human and cannot be ascribed to general auditory processing or perceptual learning. Instead, it is said to rely on a modular decoder that is part of the innately specified biological specialization for language, distinct from other forms of auditory processing.

Another theoretical approach, direct realism, shares with motor theory the claim that the objects of speech perception are articulatory rather than acoustic events. However, distinct from motor theory, direct realism posits that the articulatory objects of speech perception are the phonetically structured vocal tract movements, or gestures, and not the neuromotor commands. Direct realism is also very distinct from motor theory in that it denies that specialized processes are necessary to account for speech perception. Rather, following in the Gibsonian tradition of direct perception in the visual modality, this theory asserts that the articulatory gestures of a speaker that shape the acoustic signal serve as information for the listener to directly recover these gestures. This act of perception is not mediated by processes of inference or hypothesis testing but, rather, is direct in the sense that the gestures are apprehended without cognitive mediation from the rich information present in the acoustic signal that specifies a speaker's phonetic gestures. This theory is realist in the sense that perceivers are thought to recover the actual physical properties of the articulatory gestures from the acoustic signal.

The general approach to speech perception is distinguished from motor theory in that it does not invoke specialized mechanisms or modules to explain speech perception. Rather, its working hypothesis is that acoustic speech sounds are perceived with the same mechanisms of auditory perception and cognition that have evolved to handle other classes of complex environmental sounds. Insofar as speech is perceived as a multimodal event, the objects of perception in other modalities are hypothesized to also be general and not specific to speech. The general approach assumes that mapping from signal to meaning is not mediated by the perception of articulatory gestures but, rather, involves mapping the complex structure of the acoustic signal to regularities learned through experience with the distributions of the ambient language community. This general theoretical perspective embraces general perceptual and cognitive mechanisms, not specific to speech but neither limited to solely low-level sensory processing and psychophysics. The account is 'general' in the sense that it suggests that the broad perceptual/cognitive processing of the central nervous system and also the considerable feedback that higher centers have to lower levels of processing are brought to bear in perceiving spoken language.

A great deal of empirical research has called into question the necessity of specialized, modular mechanisms for speech perception. Nonhuman animals exhibit many of the hallmark perceptual phenomena previously thought to be indicative of specialized processing by humans for speech. Moreover, nonspeech acoustic contexts influence human adult speech categorization, a finding that would not be expected if speech processing is modular. There remains a very active debate among the proponents of direct realism and the general approach. However, both theories make predictions about the expected patterns of speech perception, so empirical research promises to increase understanding of speech perception and either further refine these accounts or suggest new thinking.

Neural Processing

One source that many believe will provide theory refinement is the study of the neural processing underlying speech perception. A great deal of research has been directed to understanding the neural bases of speech perception, and many of the efforts have been applied to the theoretical distinctions discussed previously. Research using neuroimaging techniques, for example, has compared auditory processing of speech versus processing of environmental sounds matched along multiple acoustic and perceptual dimensions, including rhythm, content, and duration. Common brain regions are activated by both stimulus types, but of possible theoretical interest, some regions are activated more by speech than by environmental sounds. Some have argued that these sorts of data demonstrate that human adults have specialized neural mechanisms for perceiving speech. However, it is worthwhile to keep in mind that these brain regions may take part in nonlinguistic processing but simply be more engaged by speech stimuli than other nonspeech sound sources. In fact, further investigation of the regions that are activated more by speech than nonlinguistic sound stimuli has found that these areas are also activated by pitch, melody, environmental sound, and/or nonauditory conceptual processing. In summary, caution must be used in interpreting brain regions as indicative of speech-specific processing based on limited controls. The differential activation of brain regions by speech versus nonspeech acoustic signals may arise from differential demands on auditory and conceptual processes brought to bear by speech versus nonspeech signals or by the vast differences in experience listeners have with speech compared to the nonspeech control stimuli.

See also: Connectionist Models of Language Processing; Sentence Comprehension; Speech Perception: Development; Speech Perception: Neural Encoding; Speech Perception: Cortical Processing; Statistical Learning of Language; Word Recognition.

Further Reading

Diehl RL, Lotto AJ, and Holt LL (2004) Speech perception. *Annual Review of Psychology* 55: 149–179.

Flege JE (2002) Interactions between the native and second-language phonetic systems. In: Burmeister TPP and Rohde A (eds.) *An Integrated View of Language Development: Papers in Honor of Henning Wode*, pp. 217–224. Trier, Germany: Wissenschaftlicher Verlag.

Fowler CA (1986) An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics* 14: 3–28.

Ganong WF (1980) Phonetic categorisation in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6: 110–125.

Hickok G and Poeppel D (2000) Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences* 4: 131–138.

Holt LL (2005) Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychological Science* 16: 305–312.

Johnson K (2003) *Acoustic & Auditory Phonetics*. Malden, MA: Blackwell.

Liberman AM (1996) *Speech: A Special Code*. Cambridge, MA: MIT Press.

Lotto AJ, Kluender KR, and Holt LL (1997) Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America* 102: 1134–1140.

Mann VA (1980) Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics* 28: 407–412.

McClelland J, Mirman D, and Holt LL (2006) Are there interactive processes in speech perception? *Trends in Cognitive Science* 10: 363–369.

McGurk H and McDonald J (1976) Hearing lips and seeing voices. *Nature* 263: 747–748.

Pickett JM (1999) *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology*. Needham Heights, MA: Allyn & Bacon.

Pisoni DB and Remez RE (eds.) (2005) *The Handbook of Speech Perception*. Malden, MA: Blackwell.

Price C, Guillaume T, and Griffiths T (2005) Speech-specific auditory processing: Where is it? *Trends in Cognitive Science* 9: 271–276.

Speech Perception: Cortical Processing

S K Scott, Institute of Cognitive Neuroscience, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The perceptual processing of speech has been long associated with the posterior temporal lobe. More recent work has suggested that Wernicke's area can be usefully fractionated into different functional streams of processing, and that this may be a helpful framework within which to think about speech perception. This approach has been motivated by developments in the neurophysiology of primate auditory processing and functional imaging studies of human speech perception. For practical reasons, work dissociating streams of processing in the primate auditory system has lagged behind the many developments in visual neurophysiology. The primate primary auditory cortex (PAC) is located on the supratemporal plane making it difficult to access and it does not show a strong response to pure sine tones. Recent technical developments have permitted these problems to be overcome, including the crucial realization that pure sine tones were not the optimal stimulus for driving auditory cortical responses, because most sounds encountered in the environment – especially communicative vocalizations – are highly complex. Unlike the visual system, the auditory system computes many aspects of sounds in the ascending auditory pathway, such that the representation at the cortical level is typically of a more complex auditory object. This has implications for the optimal stimuli to drive neurons in the PAC.

Since these developments, it has become clear that there are some general parallels between the cortical organization of sound and vision, in the sense that both hierarchical and parallel processing can be seen in auditory processing, and this has some degree of functional specialization. The core of the auditory cortex receives all its input from the auditory thalamus, and is organized as three distinct core fields in a rostral–caudal orientation. These core areas are connected to surrounding belt and parabelt regions. The rostral–caudal organization is maintained within these connections, such that the most rostral core field projects to the rostral and rostral lateral belt, but not to the caudal belt, and vice versa. The same is true for projections from the belt and core to parabelt regions.

In terms of connectivity, there is both hierarchical and parallel processing in the PAC. The rostral–caudal organization of the PAC is also maintained in the projections to frontal cortex. There are projections from rostral belt and parabelt to ventral prefrontal regions, and adjacent but nonoverlapping projections to the dorsal prefrontal and premotor cortex from caudal belt and parabelt. These two broad streams of processing are associated with different functions. Moving from medial to lateral auditory fields, responses are driven by progressively wider bandwidth sounds. It has been demonstrated that antero-lateral parabelt cells respond to conspecific vocalizations, while a greater number of more posterior lateral parabelt cells respond preferentially to the localization of calls. This is strong evidence for functional streams, perhaps similar to those in the visual system, with an anterior 'what' stream being functionally and anatomically distinct from a posterior 'where' stream. In addition to these, there is also a suggestion of further types of auditory processing. In the medial caudal belt regions, auditory neurons have been identified which respond to touch, and these may form a route for sensorimotor integration of sound and action.

Several recent approaches to the neural basis of speech perception have used evidence for anatomical and functional streams of processing as a framework for thinking about the early perceptual processing of speech. This approach has the strength of moving away from simplistic anatomical terms like Wernicke's area, and permits the results of functional imaging studies to be interpreted within an anatomical context. Arguably, an anatomical framework is essential when interpreting functional imaging data, because methods such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) are anatomical tools, and the locations of activations are of critical importance. Second, the work from primates indicates that the primate brain processes sounds in (at least two) different ways – for identity and for location. This suggests that, in the human brain, one might well not expect speech – which is first and foremost a sound – to be processed along one single route. Third, although language has no animal model, the finding that the 'what' pathway in primates is sensitive to conspecific vocalizations suggests that early acoustic processing of speech, distinct from language abilities, might have developed in a neural system already important in auditory communication. Indeed, there is even evidence that as speech and language are left lateralized in humans, so is the processing of conspecific vocalizations left lateralized in macaques. This suggests that understanding the neural systems important in processing vocalizations by

nonhuman primates may well be useful when interpreting data from functional imaging studies of human language.

The Processing of Intelligible Speech

Early functional imaging studies of speech showed a surprising degree of bilateral activation when speech was contrasted with silence. This was surprising because a left lateralized system was expected, given that left, not right, hemisphere lesions result in aphasia. Of course, the bilateral activation seen was because the results of functional imaging studies are highly dependent on the baselines used, and comparing speech to silence fails to control at all for acoustic stimulation, let alone the acoustic structure of speech.

Several more recent studies, utilizing more sophisticated baselines, have been able to delineate neural systems which seem to be important in early perceptual processing of intelligible speech. Regions within the superior temporal sulcus (STS), typically left lateralized, tend to show an enhanced response to intelligible speech in such contrasts. Some of these studies have argued that the response in left anterior STS shows the strongest response, and that this has a direct analog with the anterior 'what' pathway seen in the primate literature. Studies have also suggested that there is a hierarchical response to the speech signal. There is an area lateral to the PAC, in the superior temporal gyrus (STG), which shows a heterogeneous response to structure (e.g., harmonic structure) and modulations (e.g., amplitude modulations) in sound. Several studies have also shown a response here to speech sounds. From this response in early auditory areas, activation lateral and anterior to the PAC becomes progressively more selective to intelligible speech. Interestingly, in the STS, there is no differentiation between the response to intelligible speech from a human voice and the response to speech from a robotic-sounding harsh whisper. Thus the anterior STS seems to be strongly activated by intelligible speech and less sensitive to whether this sounded like a 'real' human talking. In parallel, several studies have associated perceptual processing of voice information (e.g., speaker identity) with right STS, an area which does not commonly show a sensitivity to intelligibility in speech.

However, because these studies used intelligible sentences, we cannot dissociate the precise representational characteristics of this response, which could include lexical, semantic, and syntactic factors, although similar activation is seen using unconnected single words. Intelligibility is a useful construct in speech processing, because it captures a global level of speech comprehension and can be easily quantified. More fine-grained approaches try to unpack how specific lexical factors are processed in these dorsolateral temporal lobe areas. For example, responses have been demonstrated to both semantic and syntactic violations along the left anterior dorsolateral temporal lobe, which might suggest that more linguistic elements are being processed in addition to purely phonological and acoustic processing. However, direct contrasts of these two types of violations did not reveal significant differences in the temporal lobes (although greater responses were seen in the basal ganglia for syntactic violations), suggesting that semantic and syntactic processes may be difficult to precisely distinguish in the temporal lobes (at least with the resolution of current fMRI techniques and subtraction methodologies). Consistent with this, a recent study addressing the effect of predictability on the intelligibility of speech has identified regions in the left inferior frontal gyrus (IFG), medial prefrontal cortex, and parietal lobe that are activated when context helps the comprehension of speech. This suggests that more high-level linguistic factors modify speech intelligibility from areas beyond the temporal lobe, and that these may represent candidate areas for 'top-down' modulation of speech perception.

These studies have identified an important role for the STS in the processing of speech, along with re-entrant processes in the left IFG, medial prefrontal cortex, parietal lobe, and basal ganglia in the processing of intelligible speech. The next sections address more specifically targeted processing of the speech signal along phonetic and psychological bases.

Phonological Information: Phonological Features

From a phonetic perspective, information in speech can be categorized in different ways, for example, by characterizing the manner or the place of articulation. This in turn is a candidate cue to which the speech perception system is sensitive. There is some evidence to support this as an aspect of the perceptual processing of speech. For example, both magnetencephalography (MEG) and fMRI studies have indicated that vowels and consonants are apparently processed perceptually using these abstract representations of the place of articulations, despite considerable surface variation due to speaker differences and articulatory variation. In these studies, multiple exemplars of consonants and vowels were used from different speakers, so the consistency in results was not due to basic acoustic properties of individual tokens. Thus, although different speakers produce quite different speech sounds with respect to the acoustic pattern,

there are responses in the lateral STG which are selective to the place of articulation across different vowels and consonants. This approach is attractive, as it suggests that we can successfully use theories of phonetics to drive investigations of the neural basis of speech perception. It also suggests that responses in the lateral STG, which is the auditory association cortex and contains the homolog of primate auditory parabelt regions, may well be tuned to specific properties of speech sounds. It also raises the possibility that we might be able to identify a topographic map of phonetic features in the auditory association cortex.

Phonological Information: Phonotactic Structure

A second theoretical issue in the speech sciences is the combinatorial structure of speech sounds – phonotactics. This describes how, in different languages, different sequences of speech sounds are possible, and can also govern overall syllable structure. For example, in British English, a syllable can be formed with up to three consonants before a vowel, and up to four afterward (e.g., 'strengths'). Such a structure would be impossible in some other languages; for example, Japanese does not permit clusters of consonants before or after a vowel. Phonotactics thus describes properties of syllable structure at both local and global levels.

These properties of syllable structure have been used to distinguish the neural bases of sensitivity to linguistic versus perceptual information in speech. The power of this approach is enhanced by using two language groups, such as French and Japanese. French is a language that does not distinguish vowels as different based on their length: thus, the nonwords ebuzu and ebuuzu sound like the 'same' nonword in French. In Japanese, however, vowel length is linguistically contrastive: ebuzu and ebuuzu are two different nonwords. However, as noted earlier, Japanese does not permit clusters of consonants, unlike French: when presented with the nonword ebzu, the bz sequence is illegal in Japanese, and an epenthetic vowel is heard – Japanese listeners presented with ebzu 'hear' ebuzu. Thus, to Japanese listeners, ebuzu and ebzu sound like the same nonword, while to French listeners these are two different nonwords. In an fMRI study, French and Japanese listeners were presented with pairs of nonwords, which had to be classified as the same or different. Elegantly, the nonwords that could be distinguished linguistically for one group were an acoustic difference for the other group, and vice versa. Using this design, neural responses can be identified in left STG, immediately lateral to the PAC, which were sensitive to the linguistically relevant stimuli comparisons, but not to the acoustically different but perceptually identical comparisons in both language groups.

What this study is thus able to show is a sensitivity to syllable-level structure that is from one's native language. Importantly, this activation is early in the perceptual processing of speech, in brain regions which would correspond to the parabelt in the primate brain. The hierarchical processing of speech not withstanding, this indicates there is some sensitivity to the structural properties of speech early in auditory cortical fields. As mentioned earlier, this region of the auditory cortex, lying lateral to the PAC, has also been shown to be sensitive to amplitude modulation (and other forms of auditory structure). Amplitude modulation is a very important aspect of the speech signal, reflecting global characteristics of the amplitude envelope of speech. That the neural processing of linguistically relevant changes in phonotactic structure overlaps with the neural processing of amplitude modulation suggests that early auditory areas may be tuned for linguistically relevant properties of sound. This opens the possibility that responses in the STG will reveal the processes and representations important in the fast, plastic, hierarchical, and parallel processing of the speech signal which must occur prior to comprehension.

Psychological Factors: Categorical Perception

More psychological approaches to speech perception have also been attempted in cognitive neuroscientific investigations. Categorical perception has been used as an approach to identify neural responses to consonant–vowel (CV) stimuli. Categorical perception describes the phenomenon of perceptual discrimination being best at the boundary between two classes of sounds, such that a dimension spanning both categories is not classified linearly. This is considered by psychologists to be an important property of speech perception and potentially explains how listeners cope with the variable speech output produced by speakers. Using fMRI, a recent study employed sounds which varied along a ba-da series, a series which is perceived categorically by English speakers. A form of spectral rotation was used to give an acoustically similar but nonintelligible series, which was established not to be perceived categorically. The results showed left lateralized response to the ba-da stimuli, in the mid STS, when contrasted with the rotated stimuli. This is an important study as it indicates that the STS response to speech is not driven by sentential or complex lexical structure. It will be interesting to establish whether the neural

regions identified demonstrate categorical neural responses to the speech series, which would be powerful evidence for the categorization of speech sounds in the auditory cortex. It would also permit a comparison with the data on featural/phonotactic structure, which tend to give activations somewhat earlier in the auditory cortex.

Psychological Factors: Auditory Word Forms

Another approach to the neural basis of speech perception is that of identifying candidate 'auditory word form' areas. This approach is motivated by cognitive neuroscience investigations and cognitive models and has gained support from PET and fMRI studies of reading, which have consistently identified a response in the left extrastriate cortex to written words and orthographically legal nonwords. While there is much less work concerning auditory word form areas, the phenomenon of repetition suppression, whereby the neural response is reduced to the second presentation of a stimulus relative to the first presentation of the same stimulus, has been used to try and identify candidate areas. This approach has revealed a selective response in the left anterior STS, in a region very similar to previous demonstrations of a neural response to intelligible speech signals. It seems plausible to consider that the two results are essentially addressing different aspects of the same perceptual phenomenon – not lexical identity *per se*, but a point along the anterior 'what' pathway in the auditory system, consequent to acoustic, sequential, and featural processing, which is sufficiently well specified that it can form the basis for access to the lexical and semantic system. From here, one might propose, there is access to and feedback from the lexical and semantic system in ventral temporal areas, and links with attentional and executive systems in the prefrontal and parietal cortex.

Speech Perception: Production Links

One of the most striking effects in speech perception is the link between speech perception and production. For example, if our language does not distinguish two speech sounds linguistically, we can have problems hearing or producing the sounds accurately. This is the basis for Japanese mispronouncements of r/l distinctions: these two sounds are two different consonants in English but two variations of the same consonant in Japanese, which leads to problems in accurate articulation. Some perspectives suggest a role for perception production links in human language organization. It has been demonstrated that silently mouthing a written word primes a later auditory lexical decision task, suggesting that articulation itself is sufficient to activate the representation of word forms. Indeed, one entire approach to modeling speech perception, the motor theory of speech perception, is focused on the primacy of motor representations in the perception of speech, such that a speaker's intentions for articulation are perceived.

A candidate route for such phenomena would be some point in the brain where there are shared resources for the perception and production of speech, be those representational, process based, or both. The previous sections of this article have suggested that functional imaging studies have been successful at delineating structures important in the neural basis of speech perception; the following section is targeted at the possibilities for the neural basis of speech perception and production links.

Perception Fields Important in Production

As outlined in the previous sections, early acoustic processing of speech involves regions in the dorsolateral temporal lobes. We have seen that a specific route for speech comprehension runs lateral and anterior to the PAC. In stark contrast to this, many previous studies have ascribed an important role for the planum temporale (PT) in speech perception. Lying lateral and posterior to the PAC, the PT has been shown by functional imaging studies not to be a site for speech-specific perceptual processing. Indeed, much of the interest in this region has derived from anatomical studies which have indicated that the PT is larger on the left than on the right, an approach which has suffered considerable criticism. However, more recent work has suggested that there may well be some speech-specific functions in the medial planum. A region of the auditory cortex, posterior and medial to the PAC, has been shown to respond to speech production – both speaking aloud and silent mouthing. This is a striking activation of the auditory cortex in a motor output task, and it is all the more impressive for it being seen when the output is mouthed – that is, the response is not solely to the subjects hearing their own voices. One interpretation of such activation is that it indicates a guidance of articulation by speech perception mechanisms – that the motor act of speech is guided by representations of the target sound of the utterance. The medial PT is in an optimal place to act as a sensorimotor interface in speech perception and production, being part of the posterior stream of auditory processing, linking into the parietal and motor cortex; it also lies just ventral to the secondary face area in the sensory cortex. In nonhuman primates, these medial and

caudal auditory areas are activated by multimodal sensory stimulation, again consistent with a role for sensorimotor integration. A recent review has suggested that this area may be representing 'possible' articulations, and it has been suggested that it may form a posterior 'how' stream of processing. This posterior–medial planum activation has also been identified as a component of a verbal working memory system, which may not be speech specific.

Production Fields Activated in Speech Perception

Primary speech motor recruitment has been suggested in another recent study in which participants were scanned listening to syllables and producing syllables aloud. A region-of-interest (ROI) analysis on the brain regions activated during speech production revealed peaks in the left motor and premotor cortex that were also activated during speech perception. This indicated that speech perception recruits the motor cortex in a very direct fashion, perhaps along the lines of the classic motor theory of speech. However, a study of nonverbal emotional vocalizations (such as laughs and screams) has also shown a strong response in the motor and premotor cortex to these nonlinguistic stimuli, suggesting that the recruitment of speech production areas during speech perception may not necessarily reflect linguistic processing of the speech signal.

Summary and Conclusions

Overall, therefore, we can see a crucial role for the human STS in speech perception, showing responses to CV syllables, words, and sentences, and potentially being the location of auditory word forms. Consistent with a hierarchical response in the auditory pathway, we can see phonological properties of speech processed in lateral STG areas; thus, the perceptual mapping from sound to acoustic-phonological forms arguably occurs in these areas. In posterior and medial auditory areas, as well as in frontal motor control areas, the speech signal seems to be processed as an action. These areas may form the basis for sensory motor integration of speech sounds with the motor representation of speech.

See also: Auditory Cortex Structure and Circuitry; Auditory System: Central Pathways; Auditory Cortex: Models; Connectionist Models of Language Processing; Sentence Comprehension; Speech Perception: Adult; Speech Perception: Development; Speech Perception: Neural Encoding; Statistical Learning of Language; Word Recognition.

Further Reading

Cohen L, Jobert A, Le Bihan D, and Dehaene S (2004) Distinct unimodal and multimodal regions for word processing in the left temporal cortex. *Neuroimage* 23: 1256–1270.

Davis MH and Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. *Journal of Neuroscience* 23: 3423–3431.

Friederici AD, Ruschemeyer SA, Hahne A, and Fiebach CJ (2003) The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cerebral Cortex* 13: 170–177.

Hickok G and Poeppel D (2004) Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition* 92: 67–99.

Jacquemot C, Pallier C, LeBihan D, Dehaene S, and Dupoux E (2003) Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *Journal of Neuroscience* 23: 9541–9546.

Liebenthal E, Binder JR, Spitzer SM, Possing ET, and Medler DA (2005) Neural substrates of phonemic perception. *Cerebral Cortex* 15: 1621–1631.

Obleser J, Lahiri A, and Eulitz C (2004) Magnetic brain response mirrors extraction of phonological features from spoken vowels. *Journal of Cognitive Neuroscience* 16: 31–39.

Obleser J, Wise RJ, Alex Dresner M, and Scott SK (2007) Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience* 27: 2283–2289.

Poremba A, Malloy M, Saunders RC, Carson RE, Herscovitch P, and Mishkin M (2004) Species-specific calls evoke asymmetric activity in the monkey's temporal poles. *Nature* 427: 448–451.

Scott SK, Blank SC, Rosen S, and Wise RJS (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123: 2400–2406.

Scott SK and Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences* 26: 100–107.

Tian B, Reser D, Durham A, Kustov A, and Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. *Science* 292: 290–293.

Warren JE, Wise RJ, and Warren JD (2005) Sounds do-able: Auditory-motor transformations and the posterior temporal plane. *Trends in Neurosciences* 28: 636–643.

Wilson SM, Saygin AP, Sereno MI, and Iacoboni M (2004) Listening to speech activates motor areas involved in speech production. *Nature Neuroscience* 7: 701–702.

Wise RJS, Scott SK, Blank SC, Mummery CJ, and Warburton E (2001) Identifying separate neural sub-systems within Wernicke's area. *Brain* 124: 83–95.

Speech Perception: Development

S Curtin and D Hufnagle, University of Calgary, Calgary, AB, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The speech signal is complex. There is no one-to-one mapping of acoustic signal to linguistic unit. Moreover, the acoustic signal is filled with environmental sounds that are not relevant to speech. Yet, speech perception abilities develop quite rapidly in young children. Research on infant speech and word perception supports the importance of audition and exposure to speech stimuli in resolving complex mappings and discovering relevant linguistic units. This article details the emergence of language-specific sound categories, sound combinations, rhythmic patterns, and words during the first year of life; describes several theoretical accounts of the development of speech perception; and comments on current issues in the field.

Perceptual Abilities at Birth

Infants demonstrate a number of speech perception capabilities at birth. Typically developing infants are able to discriminate, categorize, and store subtle phonetic speech contrasts in long-term memory. In order to facilitate speech perception, it is useful for infants to separate and identify different sound sources in the acoustic signal, beginning with distinguishing speech from nonspeech. From birth, infants listen longer to speech than to nonspeech sounds that are comparable in complexity and spectral frequency. This is evidence that infants possess some of the raw psychoacoustic and cognitive capacities necessary to perceive speech at or just before birth. To account for the special status of speech, it has been posited that infants possess a bias for listening to speech over other types of sounds. A predisposition to attend to speech would enable infants to tune in to important properties of the speech signal, thus guiding acquisition.

Newborns also demonstrate sensitivity to rhythmic properties of language. Any language falls into one of several rhythmical classes: stress-timed, syllable-timed, or mora-timed. In stress-timed languages, including English and German, speech is timed to create roughly equal intervals between stressed syllables. Syllable-timed languages, such as French and Catalan, have syllables that are relatively equal in duration, with lengthening rules for the ends of phrases or utterances. Mora-timed languages, such as Japanese, have subsyllabic timing units. As early as birth, infants are sensitive to rhythm information and show a preference for listening to languages that conform to the rhythmical class of their native language but do not distinguish their native language from others within the same rhythm class. Sensitivity to native-language rhythm patterns at birth is evidence for infants' use of knowledge acquired during prenatal language exposure.

Learning Language-Specific Sound Categories

Infants are able to perceive a number of speech contrasts from birth, but these contrasts may not coincide with the contrasts that are used by the individual languages that the infants are learning. Before 6 months of age, infants discriminate a range of consonant and vowel contrasts present in their native language as well as contrasts found in other languages. Young infants between 1 and 4 months of age, like English adults, are better able to discriminate two different /ba/ and /pa/ syllables that vary only in voice onset time (VOT; the time between the release of articulators and the onset of voicing) if the syllables are instances of two different adult phonemic categories than if the two syllables are from within a single adult category. This is true even though within- and across-category tokens differ by equal VOT intervals. In discriminating contrasts from unfamiliar languages, infants are better able to discriminate across- than within-category differences, even though adults often have difficulty discriminating nonnative contrasts. For example, Kikuyu-learning infants of 4 months are able to discriminate an English voiced-voiceless contrast (/ba/ vs. /pa/) even though this contrast is not used in the Kikuyu language. It is interesting that when tested on the voicing distinction in both English and Spanish, which uses different cues from English to distinguish voicing contrasts, 4.5- to 6-month-old Guatemalan infants were able to discriminate the English VOT contrast but not the Spanish one. This result might suggest that infants are not sensitive to fine-grained details in the signal. However, if this were the case, then Spanish-learning infants would not be able to acquire the language's voicing distinction. Rather, it is more likely that the English contrast, which is readily discriminated, is aligned with a language-general voicing boundary, and infants learning the Spanish voicing distinction require resetting or shifting of their perceptual categories. This would explain both the inability to discriminate the Spanish contrast by Guatemalan infants and the ability

to discriminate the voicing contrast by Kikuyu infants. Moreover, 3- to 6-month-old infants show graded, within-category perception of VOT under appropriate testing conditions, indicating that although categorical perception is most easily revealed, graded perception is also possible.

These classic findings in categorical perception laid the foundation for a broad picture of infant speech perception. Infants discriminate minimal phonetic differences not just in salient initial position in words but in medial position as well. They discriminate minimal differences in both consonants and vowels. Infants of 6–8 months can categorize multiple but distinct instances of stimuli from within a phonetic category as equivalent and distinct from a set of stimuli for a different phonetic category, but they are unable to make such divisions at arbitrary points along an acoustic continuum.

Functional reorganization from language-general to language-specific perception of consonant contrasts occurs by 10–12 months of age. English-learning 6- to 8-month-olds are able to distinguish contrasts that are not used in English, performing like adults who speak languages in which the contrasts are meaningful. On the other hand, English-learning 10- to 12-month-olds perform like English-speaking adults: they distinguish contrasts that are meaningful in English but do not demonstrate behavior consistent with discrimination for other contrasts. This decline in infants' ability to distinguish contrasts has been replicated using a number of contrasts. Similar to consonant perception, infants are able discriminate across a range of vowel categories within the first few months of life, even vowels that are acoustically similar or not phonemically relevant in the native language. Young infants are also able to detect vowel pattern differences across syllable sequences. However, by 6 months of age, looking-time data suggest that infants are already less sensitive to nonnative vowel contrasts. The reorganization process, the interaction between the perceptual/cognitive starting point and experience, appears to occur earlier for vowels than for consonants. One explanation for this is that vowels carry much of the basic prosodic information that infants are attracted to in early infancy. For example, English and Swedish 6-month-old infants exhibit a language-specific perceptual bias when tested on discrimination within vowel categories. Similarly, English-learning 4-month-old infants discriminate non-English vowel contrasts found in German, whereas 6- to 8-month-olds show more sensitivity to English than to German contrasts, and 10- to 12-month-old infants do not appear to discriminate the German contrasts. Though 4- to 12-month-olds can discriminate all vowel contrasts presented (native and nonnative), an asymmetry is observed by 6 months of age, with native prototypical vowels biasing the perception of nonprototypical vowels when the two are heard in a sequence.

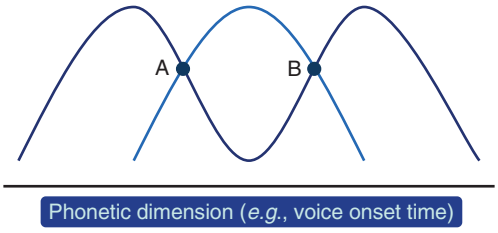

Figure 1 Learning distributions. Acoustic cues that have a bimodal distribution encourage learners to form two categories, while cues that have a unimodal distribution encourage the formation of a single category. When asked to discriminate two sounds along an acoustic continuum (A and B), infants exposed to the bimodal distribution distinguish the sounds, while infants exposed to the unimodal distribution do not categorize them differently, even though the sounds are equally frequent in the input.

During the first year of life, reorganization of speech perception takes place, but how might this be accomplished? One possibility is that infants are sensitive to statistical patterns and use this ability to track the distribution of speech sounds in a language. If so, they should be able to use this information to determine the linguistic relevance of various acoustic properties. Exposure to different distributional patterns should promote a category distinction (see Figure 1). Specifically, a group of sounds characterized by a unimodal distribution along some acoustic dimension should indicate that the acoustic property is uninformative for distinguishing speech sounds in a given language, and therefore that property need not be attended to. On the other hand, bimodal distributions should signal the linguistic importance of a contrast. In fact, infants are sensitive to the frequency distribution of speech sounds in the input, and this sensitivity is present during the age range in which developmental changes in speech perception are observed. Therefore, attention to the statistical distribution of speech sounds in the input is one factor driving the development of speech perception over the first year of life.

Learning Language-Specific Syllable Patterns

During the first year of life, infants not only begin to figure out the specific sound categories of their language, but also legal sound combinations (phonotactics) and prosodic properties (rhythm and stress). They also segment sequences that make up emergent syllable and word forms and begin to learn basic relationships between forms. Infants use their emerging speech perception abilities to learn the distributional properties of their native language. For example, phonetic distribution of the target language affects

the perception of vowels by 6 months of age, and by 9 months infants can discriminate between speech sound sequences based on the frequency with which the sounds occur in the language.

Infants learn the phonotactics of their language in the first year of life. Phonotactics refers to the co-occurrences allowed by specific languages in different syllable positions (i.e., pt does not occur word initially in English but can occur at the end of a word, such as in kept). Between 6 and 9 months, infants develop knowledge of phonotactic regularities in their language. For example, an English-learning infant will listen longer to a word beginning with the legal *str* sequence than the unacceptable *rst* sequence and will show the opposite pattern for word endings, just as an adult does. Infant's phonotactic preferences are most parsimoniously explained by frequency in the input. When presented with nonsense words with two equally legal phonotactic sequences, infants of 10 months consistently prefer stimuli with the more common sequences. Bilingual infants 10 months old prefer the phonotactic patterns of the language that they hear most often. Converging evidence from artificial language learning research also supports the use of distributional statistics to learn permissible phonotactic patterns.

Infants also learn the stress patterns of their native languages rapidly. Between 1 and 4 months of age, infants are sensitive to changes in stress patterns, and sometime between 6 and 9 months of age, infants begin to orient to the predominant stress pattern of their language. English is classified as a trochaic stress system (strong/weak syllable patterns), and English polysyllabic words tend to have strong initial syllables. Strong syllables in English have higher pitch, greater amplitude, and longer duration than weak syllables. Thus, by 9 months of age, English infants prefer trochaic patterns. Infants' sensitivity to the distributional and stress-related properties of language suggests that infants' exposure to syllables that differ in their acoustic properties (i.e., for lexical stress, the change in duration, amplitude, and pitch) results in differing perceptions of these syllable types.

Learning about Words

Parsing the continuous speech stream into individual words is a fundamental task for the language-learning child. Discovering word boundaries is a nontrivial problem due to the lack of direct marking of word boundaries in the acoustic signal, unlike the way that white spaces mark boundaries in written text. However, between 6 and 8 months of age, infants begin segmenting sequences from the speech stream. Strategies based on prosodic cues (including pauses, segmental lengthening, metrical patterns, and intonation contour) have been proposed as means of detecting word boundaries. Other recent proposals have focused on the statistical properties of the target language that might be utilized in early segmentation.

After being familiarized to words spoken in isolation, 7.5-month-old infants who hear passages containing familiar words (cup) and passages that contain novel forms that differ minimally from familiar words (tup) prefer to listen to the passages containing the familiar words, suggesting they extracted these words from the passage. Further evidence that infants use familiar words to help segment speech is seen when a highly frequent form such as mommy precedes an unfamiliar word. When this happens, infants' segmentation of the unfamiliar word is facilitated. This is not the case when two unknown adjacent words are presented. Similarly, when novel words are preceded by a highly frequent function word, such as the, infants are better able to segment and remember the novel words. Being able to use familiar word information facilitates the segmentation of the speech stream. There are a variety of other cues to segmentation, such as prosodic, rhythmic, and segmental information; transitional probabilities; phonotactics; and stress. These cues are part of the distributional properties of the linguistic input.

Statistical information about the co-occurrence of syllables aids infants in segmenting the speech stream. When exposed to an artificial language of trisyllabic nonwords with no cues to word boundaries other than probabilistic co-occurrence information, infants at 8 months were able to segment word forms from the speech stream. Even when co-occurrence is the only type of information present, infants can successfully segment the speech stream.

Syllables themselves carry other information, such as phonotactic and prominence information (word stress). Probabilistic information in the form of phonotactics is used for segmentation by 9 months of age. When infants are presented with CCVC sequences in a highly probable word boundary context (C#CVC) and a low probability one (CCVC), they prefer the CVC sequence that originated from the highly probable word boundary context. Prior to this, 7.5-month-old English-learning infants use strong syllables to determine the presence of certain words in fluent speech. When infants are exposed to only the strong syllables of a two-syllable word (e.g., king for kingdom) and then to passages containing the entire word, they do not listen any longer to these passages than to passages that contain novel targets. This is also the case when presented first with passages containing words like kingdom and then to just the strong syllable in isolation (king). These results suggest that

infants are doing more than just matching strong syllables; they are matching the entire word.

The items used in these experiments corresponded to a trochaic pattern, and it might be the case that the infants were segmenting the speech based on this pattern. This type of language-specific preference for a particular rhythmic pattern may guide segmentation. Infants exposed to a predominately iambic language, such as Canadian French, segment only weak-strong (wS) words. English infants, on the other hand, misparse iambic wS patterns and treat the S syllable as word initial. For example, in a word like guiTAR, infants segment the strong syllable, TAR, as a word-initial syllable. Moreover, if TAR is consistently followed by an unstressed word (e.g., is), infants will treat TAR#is as a single unit. However, if two strong-syllable monosyllabic words are adjacent to each other (COLD ICE or PACK ASH), infants will not missegment these items. By the time English-learning infants are 11 months old, they no longer missegment wS words. The preference for trochaic patterns has been further demonstrated using an artificial language-learning task. Both 9- and 7-month-old infants successfully parse the stream and prefer syllable sequences at test that correspond to an initially stressed syllable from the familiarization phase, even though the test sequences had equally prominent syllables. Taken together, these results support the hypothesis that word stress is a salient cue for segmentation that can be used by the infant to parse the continuous stream of speech.

Inevitably, any partially predictive cue to word segmentation is not completely reliable, and cues differentially influence segmentation throughout development. Stress, independent of other cues, can be used to segment sequences at 7 months of age, but co-occurrence cues dominate stress cues at this age. Coarticulation and stress cues take precedence over raw co-occurrence patterns in 8-month-old infants' parsing strategies. Regardless of the dominance of one cue or another at a given age, it is likely that infants use all available cues to segment speech throughout development.

Infants' capabilities for learning about language-specific sound categories and syllable patterns facilitate the segmentation of speech. As sequences are segmented from the speech stream, these units in turn help refine knowledge about the categories and structures being acquired. Though infants demonstrate some proficiency in speech perception, these abilities continue to evolve throughout development (see Figure 2).

Theoretical Approaches

In 1980, Aslin and Pisoni outlined four possible models for the development of speech perception. (1) Universal theory corresponds to the model in which infants are born with sensitivities to native and nonnative phonetic contrasts, and only those biases that receive support from experience persist. Without relevant listening experience, initial sensitivities are lost. (2) Attunement corresponds to a model in which general biases exist at birth, but experience functions to shape them more precisely to match adult categories. Early experience drives the ability to discriminate contrasts according to (3) the perceptual learning theory. Rate of development is dependent on frequency of contrasts as well as acoustic discriminability of contrasts and the infant's attention. Finally, (4) the maturational theory argues that development occurs according to a predetermined schedule; thus, discrimination of a contrast is independent of experience. None of these theories can account for all the findings related to speech sound contrasts. Rather, it is likely that a combination of these approaches is required to explain infants' abilities.

Other models followed, some specifically linguistic in nature, others focusing on the interface between acoustics (or articulatory gestures) and subsequent representations of speech sounds. Best's perceptual assimilation model (PAM) and Kuhl's native language magnet model (NLM) were proposed to explain general and language-specific perception of speech sounds. PAM provides a 'direct realist' account of native and nonnative speech perception, in which infants perceive speech categorically by recovering information about the speaker's vocal tract configuration. PAM predicts nonnative speech segments will be perceived according to how they might be assimilated to native categories. If the nonnative sounds are assimilated to two different native language phonological categories, they will remain discriminable. If, however, the nonnative sounds are assimilated to a single intermediate category, they will not be discriminable.

The NLM uses acoustic cues as the source of information available to the listener. According to NLM, categories are defined by groupings that emerge in multidimensional space. Initially, this type of category structure results in asymmetries in discrimination. Later, nonnative speech sounds will be collapsed into native ones. Repeated experience listening to speech helps create a category based on category goodness. The best instances of a phonetic category act as perceptual magnets and pull nearby token into their perceptual space, making them less discriminable for infants. Thus perceptual asymmetries are predicted. Poorer performance in a discrimination task is expected if a central member of the category is used as a standard than if a more peripheral token is used.

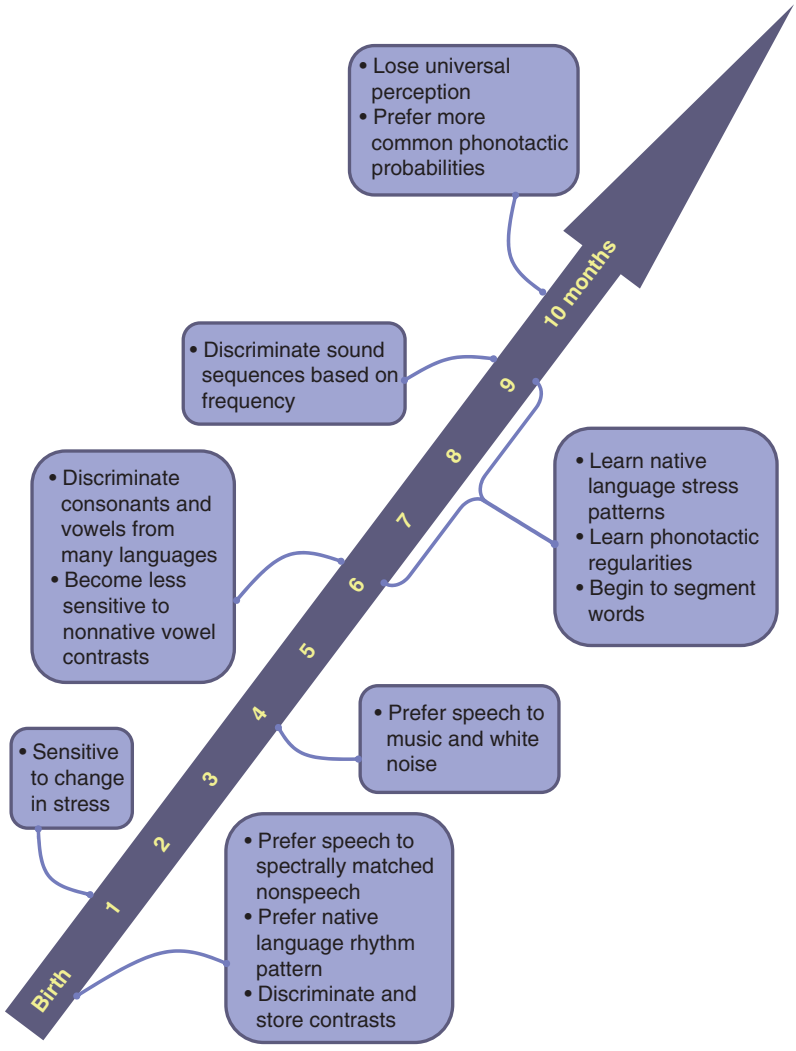

Figure 2 Milestones in speech perception in the first year.

Both PAM and NLM focus exclusively on speech perception of sound categories and do not consider the role of speech perception in learning about the structure of words. To provide a unified account of how language experience affects perception of native language sound categories and also word recognition in infancy, Jusczyk proposed word recognition and phonetic structure acquisition (WRAPSA). As the signal (speech and nonspeech input) enters the auditory system, a set of auditory analyzers provides a description of the signal. During acquisition, the output of the auditory analyzer is weighted to give prominence to features that are required for contrasting words. Once the signal has been weighted, pattern extraction takes place whereby the signal is segmented into units that temporally group together prominent features into syllabic units. WRAPSA assumes that infants first have access to prosodic information, then syllabic, and only later phonetic. Word recognition occurs by the use of representations that act as probes. Probes are matched against any existing representations of known words. If a close match is obtained, then a word is recognized, and its meaning, if represented, is accessed. If no match can be found, the probe will either be reprocessed or stored as a new entry with or without meaning.

Werker and Curtin recently proposed processing rich information from multidimensional interactive representations (PRIMIR), a theoretical framework designed to explain infant speech perception and word learning. PRIMIR assumes that the same general statistical learning mechanisms are operating over the different levels of analysis simultaneously. Thus, prosodic analysis, segmentation of the speech stream, extraction of syllables, formation of phonetic categories, and storage of word forms happen simultaneously, with each level further influencing category formation and information pickup at each other level. PRIMIR stores exemplar-based word forms in a similar fashion to WRAPSA. Word forms

are sequences that have been identified as possible lexical candidates that eventually or simultaneously become linked to meaning. In this way, WRAPSA and PRIMIR are comparable. The models diverge in the way phonemic categories are formed. WRAPSA assumes that phonemes emerge from the weighting scheme prior to the extraction of potential word forms. PRIMIR argues that phonemic categories are formed once distinguishing properties are detected after a number of word form and meaningful word linkages occur.

Challenges

During the first year of life, infants demonstrate a shift from general speech perception abilities to preferences for categories and structures that occur in the language they are learning. The past few decades have yielded numerous empirical findings that have provided the field with a solid foundation for understanding this development. However, some challenges remain. The speech stream is abundant with rich information, including acoustic properties and segments that pattern statistically. Researchers have identified some of the information that infants can use in experimental settings. Individual properties may not be necessary and sufficient for language and speech development. Research in this area has just begun to scratch the surface of what might and might not be available to and used by infants. Further challenges include understanding how developmental changes occur over time and whether there are maturational or other developmental events that affect the infants' ability to pick up and use the information available to them.

See also: Connectionist Models of Language Processing; Language Development; Sentence Comprehension; Speech Perception: Adult; Speech Perception: Neural Encoding; Speech Perception: Cortical Processing; Statistical Learning of Language; Word Recognition; Word Learning.

Further Reading

Aslin RN and Pisoni DB (1980) Some developmental processes in speech perception. In: Yeni-Komshian GH, Kavanagh JF, and Ferguson CA (eds.) *Child Phonology*, pp. 67–96. New York: Academic Press.

Best CT (1994) The emergence of native-language phonological influences in infants: A perceptual assimilation model. In: Goodman JC and Nusbaum HC (eds.) *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, pp. 167–224. Cambridge, MA: MIT Press.

Curtin S and Werker JF (2007) Perceptual foundations of phonological development. In: Gareth Gaskell M, Altmann GTM, Bloom P, Caramazza A, and Levelt P (eds.) *The Oxford Handbook of Psycholinguistics*, pp. 577–600. Oxford, UK: Oxford University Press.

Eimas PD (1975) Speech perception in early infancy. In: Cohen LB and Salapatek P (eds.) *Infant Perception: From Sensation to Cognition*, vol. II, pp. 193–231. New York: Academic Press.

Gómez RL (2007) Statistical learning in infant language development. In: Gareth GM, Altmann GTM, Bloom P, Caramazza A, and Levelt P (eds.) *The Oxford Handbook of Psycholinguistics*, pp. 601–616. Oxford, UK: Oxford University Press.

Jusczyk PW (1997) *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

Kuhl PK (1993) Innate predispositions and the effects of experience in speech perception: The native language magnet theory. In: de Boysson-Bardies B, de Schonen S, Jusczyk P, McNeilage P, and Morton J (eds.) *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*, pp. 259–274. Dordrecht, Netherlands: Kluwer Academic Press.

Kuhl PK (2004) Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience* 5: 831–843.

Maye J, Werker JF, and Gerken L (2002) Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82: B101–B111.

Saffran J, Werker JF, and Werner L (2006) The infants' auditory world: Hearing, speech, and the beginnings of language. In: Kuhn D and Siegler M (eds.) *Handbook of Child Psychology, Vol. 2: Cognition, Perception, and Language*, 6th edn., pp. 58–108. Hoboken, NJ: Wiley.

Saffran JR, Aslin RN, and Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274: 1926–1928.

Werker JF and Curtin S (2005) PRIMIR: A developmental framework of speech processing. *Language Learning and Development* 1: 197–234.

Werker JF and Tees RC (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7: 49–63.

Speech Perception: Neural Encoding

D G Sinex, Utah State University, Logan, UT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The responses of individual auditory neurons to speech sounds were first described in the 1970s. The earliest reported measurements were obtained from second-order neurons in the cochlear nucleus, followed a few years later by studies of the responses of large populations of auditory nerve fibers (ANFs). ANFs are primary auditory neurons that constitute the only pathway for conveying information from the cochlea to the brain stem. Many subsequent studies of the responses of ANFs have been carried out, but the influence of the early population studies extends to the present day. Much is known about the representation of speech by ANFs, but there have been relatively few definitive studies on the responses of central auditory neurons. This article provides some background information about the acoustics of speech sounds and presents a view of the neural encoding of speech that emphasizes the representation of the amplitude envelope by ANFs. It also addresses a topic in translational neuroscience by relating observations in experimental animals to the delivery of speech information by intracochlear electrical stimulation.

Speech Acoustics

Speech is often described as being produced by a source and a filter. The source is air from the lungs which is forced through the larynx. Forcing air through vocal folds that are held close together causes them to open and close repetitively, producing a glottal pulse on each cycle. Sounds such as vowels produced in this way are said to be voiced. A sequence of glottal pulses produces a quasi-periodic waveform. The rate at which glottal pulses occur establishes the fundamental frequency (f_0) of the vowel or other voiced sound, and f_0 determines the sound's pitch. In speech, f_0 varies within and across talkers, but it is generally on the order of 120 Hz for adult male talkers and 250 Hz for children and adult females.

During the production of stop consonants, there is an interval during which the vocal tract is constricted. Vocal folds may begin vibrating before this constriction is released, coincident with the release, or sometime after the release. The length of that interval, referred to as voice onset time (VOT), is under the speaker's control and is the basis for the perceptual distinction between pairs of consonants produced at the same place of articulation. Bilabial stop consonants, for example, are produced by constricting the vocal tract at the lips. If the consonant is produced with a short VOT, it is heard as /b/, a voiced stop. If the consonant is produced with a long VOT, it is heard as /p/, a voiceless stop. Other voiced–voiceless consonant pairs include /d/–/t/ and /g/–/k/. An important acoustic correlate of the onset of voicing is a significant increase in overall amplitude, especially in the lower frequencies.

The spectrum of the sound that is radiated from the lips may have a distinctive pattern of peaks and valleys as a result of filtering by the supraglottal vocal tract. The peaks are produced by resonances called formants. Formant frequencies are determined by the physical characteristics of the vocal tract and by the locations of constrictions produced primarily by the tongue, teeth, and lips. Formants are designated in frequency order as F1, F2, F3, and so on. Here, these abbreviations are used to refer to spectral peaks in the resulting acoustic signal. The locations of the first two to three spectral peaks provide cues that help distinguish one phoneme from another.

In natural speech, the configuration of the vocal tract rarely reaches a steady state. In addition, ongoing speech includes sequences of both voiceless and voiced segments. As a result, the spectrum and the temporal envelope of speech fluctuate at a rate typically between 5 and 50 Hz. Frequency and amplitude change are illustrated in Figures 1 and 2. Figure 1(a) is the waveform of a two-syllable vowel–consonant–vowel nonsense word, 'eetah,' spoken by a male talker. The waveform is scaled to represent sound pressure when the sound's overall sound pressure level (SPL) was 60 dB, which is within the range of conversational speech. Amplitude varied over time and was highest during the intervals from −210 to −120 ms (during the vowel /i/) and from 0 to +100 ms (during the vowel /a/). Periodicity associated with glottal pulses can also be seen during the vowel segments. In this example, the average f_0 was approximately 120 Hz. Between the two vowels, the tongue was placed behind the upper teeth to occlude the vocal tract briefly. During that interval, the waveform became aperiodic and its overall amplitude decreased substantially.

Figure 1(b) shows the spectrogram of the same word. As noted previously, ongoing changes in the configuration of the vocal tract lead to short-term changes in the formant frequencies. These changes, called formant transitions, are visible in the spectrogram,

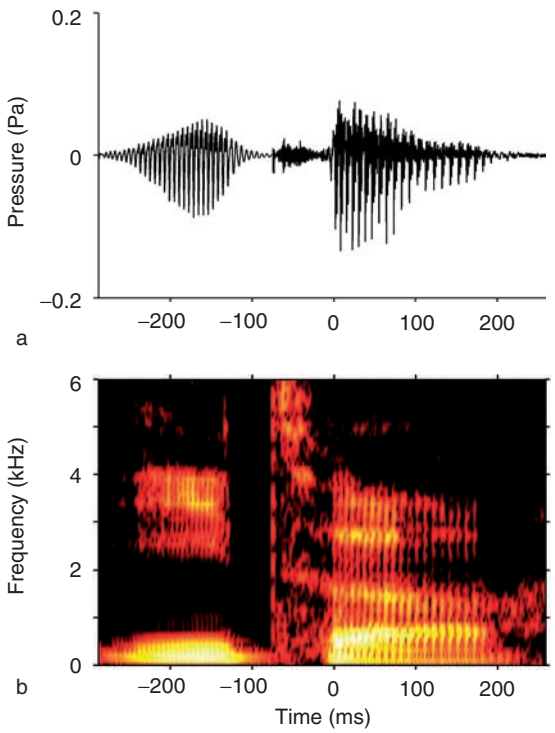

Figure 1 (a) Waveform of a two-syllable utterance consisting of the vowel /i/ as in 'heat,' the consonant /t/, and the vowel /a/ as in 'hot.' The timescale is normalized to the onset of voicing in the second vowel. (b) Spectrogram for the same utterance shown in (a). Relative level is color coded, from black to red to yellow.

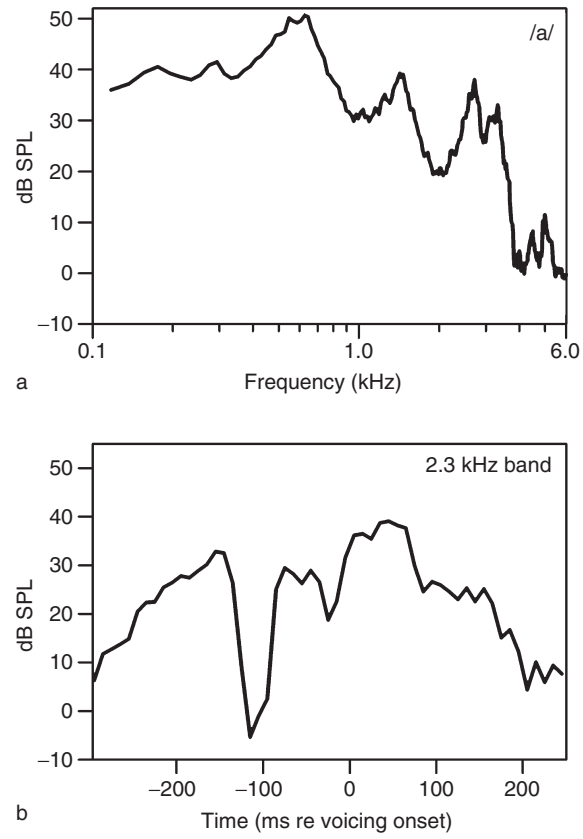

Figure 2 (a) Short-term spectrum of the vowel /a/, calculated over the interval from 20 to 40 ms of the waveform shown in Figure 1(a). Spectral peaks corresponding to the first three formants were located at 0.6, 1.4, and 2.6 kHz. (b) Temporal envelope in a limited frequency band. The envelope was computed with a gammatone filter centered at 2.3 kHz. A gammatone filter approximates the filtering that occurs at a single location along the cochlea.

especially in and around the interval between the vowels. An estimate of the short-term spectrum taken near the temporal center of the vowel /a/ is shown in Figure 2(a).

Because formants move, amplitude in any particular frequency band will also change across time. Within a band, amplitude changes over time may be larger than the amplitude changes observed across the frequency range of the first three formants (Figure 2(a)). This is illustrated in Figure 2(b) for a band centered at 2.3 kHz. In this example, the level within the band varied over a range of approximately 50 dB, and changes of 20–40 dB could occur within a few milliseconds. Other bands exhibit changes of the same or even greater magnitude, although with different temporal patterns. This kind of amplitude variation contributes to speech recognition, and the following sections discuss how it is represented in the responses of ANFs.

Basic Response Properties of Auditory Nerve Fibers

The responses of individual ANFs to speech sounds reflect intrinsic properties of the cochlea and of neurons, as well as the spectrotemporal pattern of the speech stimulus. Individual ANFs are tuned to a narrow range of frequencies at low SPLs, and the range expands asymmetrically toward lower frequencies as the level increases. The frequency at which the ANF is most sensitive is called its characteristic frequency (CF). CF is determined by mechanical properties of the cochlea, and it increases exponentially along the spatial array of ANFs from the apex to the base. Each individual ANF is driven (although in a nonlinear way) by the stimulus energy that falls within a frequency band around its CF.

ANFs can be categorized into two or three groups based on their intrinsic rates of firing in the absence of sound, the spontaneous discharge rates (SRs). Thresholds, defined as the stimulus levels that elicit a constant increase in discharge rate, are lowest for ANFs with SR greater than approximately 20 spikes s^{-1}; these are referred to as high-SR neurons. Thresholds are slightly higher for ANFs with SR between 1 and 20 spikes/s, referred to as medium-SR neurons. ANFs in the low-SR group generate less than 1 spike/s and

may have substantially higher thresholds. In cats, three SR groups can be distinguished, but in other mammals, medium- and low-SR ANFs are sometimes treated as a single group.

During the presentation of suprathreshold steady-state tones, discharge probability increases at stimulus onset and then adapts slowly over the duration of the stimulus. Discharge rate averaged over stimulus duration increases with SPL until it saturates. Rate saturation may be flat; that is, at some SPL, average discharge rate reaches a maximum and does not increase with further increases in SPL. Flat saturation is common in high-SR fibers. In sloping saturation, rate continues to increase with SPL, although it reaches an inflection point above which further increases in intensity produce more gradual increases in discharge rate. Sloping saturation is common in medium-SR fibers. The range of SPL between threshold and the level that produces rate saturation is termed the neuron's dynamic range. Within the dynamic range, changes in stimulus level can be encoded as changes in discharge rate; however, level changes that fall exclusively outside the dynamic range cannot. Dynamic range can be as narrow as 20 dB for high-SR fibers, but it is slightly larger for low- and medium-SR fibers. Rate saturation and the limited dynamic ranges of individual ANFs have significant consequences for the representation of speech information.

ANF discharges synchronize with the fine structure of the stimulus waveform, at least for frequencies less than 2–4 kHz. Because of the relation between the stimulus waveform and the neural response, information about the frequency content of the stimulus is present in the temporal discharge pattern. Discharge synchrony is important for the localization of low-frequency sounds based on interaural time differences, for the segregation of competing sounds, and for producing the sensation of pitch. Discharge synchrony has the potential to convey information about spectral shape, and this possibility is considered in the following sections.

Responses to stimuli more complex than pure tones are also affected by cochlear nonlinearities, including rate suppression and synchrony capture. Rate suppression is a reduction in the response to one tone or frequency band caused by energy in a tone or band at a different frequency; frequencies below CF are particularly effective suppressors. The component producing suppression may or may not elicit an excitatory response when presented alone. In responses to speech stimuli, spectral peaks associated with F1 may suppress the overall discharge rates of neurons that would otherwise respond to F2 or F3.

Synchrony capture occurs when the pattern of synchronized discharges is dominated by a high-amplitude component. Spectral peaks associated with formants may capture the synchronized discharges of neurons tuned to other frequency regions, especially those regions in which spectral peaks are lacking. In responses to speech, the synchronized response may be captured so completely by a component near a spectral peak that it is indistinguishable from the synchronized response to a pure tone at the same frequency.

Responses of Auditory Nerve Fibers to Speech

Steady-State Vowels

If discharge rate did not saturate, and responses were not affected by cochlear nonlinearities, each ANF would discharge at a rate proportional to the band-limited energy in the stimulus. That would lead to a relatively simple spatial representation of a stimulus such as the vowel in Figure 2(a), whose spectrum consists of a series of amplitude peaks and valleys. Neurons whose CFs were aligned with the peaks in the spectrum would respond at high rates. Neurons whose CFs were aligned with the valleys in the spectrum would respond at lower rates. Across the ordered array of ANFs, the spatial profile of discharge rate would match the shape of the spectrum. Increases in the overall level of the stimulus would be reflected in the overall discharge rates of the neurons, but the locations of peaks and valleys in the spectrum would be preserved in the spatial profile of discharge rate.

A representation similar to the ideal one described previously is observed in the responses of ANFs, but only when the effective level of the stimulus is low relative to the neurons' dynamic ranges. However, at levels that are higher but still within the range of normal conversational speech, the representation breaks down as a consequence of rate saturation and suppression. The effects of these factors on the representation of steady-state vowels can be seen clearly in the population response of the auditory nerve. In influential studies performed in the 1970s by Murray Sachs and Eric Young, responses of hundreds of ANFs were obtained from individual cats. The average discharge rate of each ANF was plotted as a function of its CF, a display referred to as a rate–place profile. Several rate–place profiles are illustrated in Figure 3. At low SPLs, the shape of the rate–place profile preserved the pattern of peaks and valleys in the vowel spectrum. However, as level increased, response peaks associated with the formants were lost in the rate–place profiles of the most sensitive, high-SR fibers (Figure 3(a)). Because the locations of spectral peaks became more difficult to discern, the rate–place profiles for two easily discriminable vowels looked similar.

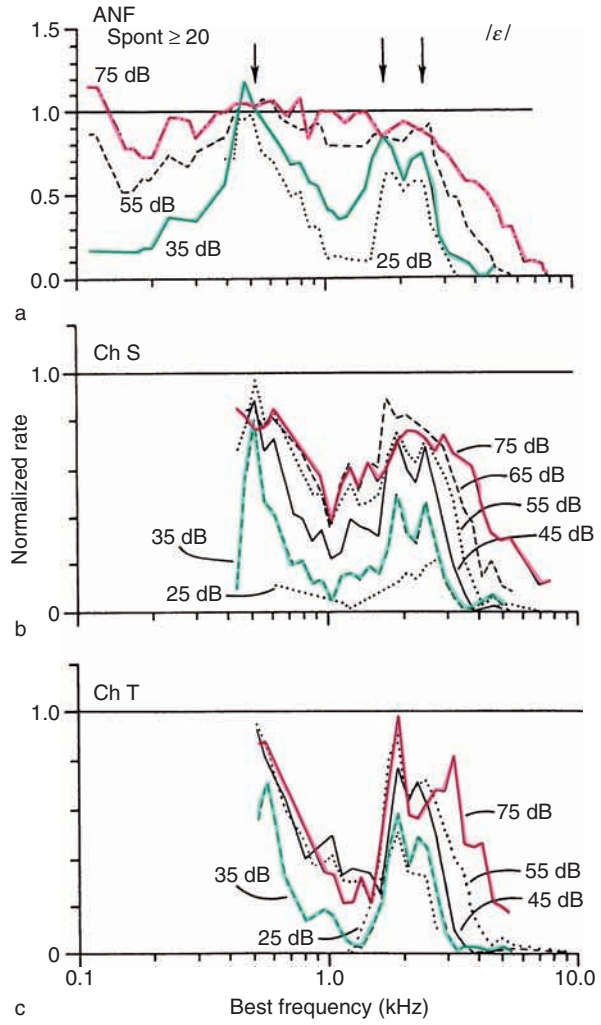

Figure 3 Rate–place profiles for high-SR ANFs (a) and cochlear nucleus chopper neurons (b and c) responding to the steady-state vowel /eh/, as in 'head.' The arrows in panel (a) mark the frequencies of the first three spectral peaks in the vowel. Responses obtained at 35 dB SPL are highlighted in green in each panel. At this level, the spectral peaks in the vowel were preserved as local peaks in discharge rate by all the neurons. Responses obtained at 75 dB SPL are highlighted in red. At this level, the locations of the spectral peaks were no longer identifiable in the responses of high-SR ANFs, but they were preserved by the chopper neurons. Adapted from Blackburn CC and Sachs MB (1990) The representations of the steady-state vowel sound /e/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *Journal of Neurophysiology* 63: 1191–1212, with permission.

At the time it was reported, this finding suggested that the simple discharge rate representation implied by the rate–place profile might be inadequate for representing speech information, and alternative coding mechanisms were considered. One possibility noted immediately was that the rate–place profiles of less sensitive, low- and medium-SR neurons preserved the shape of the vowel spectrum up to higher stimulus levels than did the profiles for high-SR ANFs. A rate–place representation of vowel spectra would be feasible if the responses of high-SR neurons could be processed separately from those of medium- and low-SR neurons in the central auditory system. Interestingly, one type of second-order neuron recorded in the ventral cochlear nucleus appears to combine information from all SR groups to produce a spatial representation of vowel spectra with a larger dynamic range than that of any single subgroup of ANFs. The rate–place profiles of these neurons, called choppers, resemble those of high-SR ANFs at low SPL and resemble those of low-SR ANFs at high SPL (**Figures 3(b) and 3(c)**). The rate–place profiles of chopper neurons preserve information about the locations of spectral peaks over a larger dynamic range than any subgroup of ANFs.

It was also noted that information about the locations of spectral peaks could be extracted by analyzing the synchronized responses of ANFs. The discharge patterns elicited by steady-state vowels and consonant–vowel syllables provided precise information about the spectral content of the stimulus, and the representation was robust over a large dynamic range. However, although the frequency content of synchronized responses is easily identified by techniques such as Fourier analysis, it is not known how or whether the central auditory system extracts comparable information to determine spectral shape. For this reason, potential coding strategies in which discharge synchrony conveys information about spectral shape lack the simplicity that is inherent in a code based on the spatial profile of average discharge rate.

Subsequent observations suggest that average discharge rate can convey more information about speech than was apparent in the original studies of the representation of steady-state vowels. For example, the olivocochlear efferent system acts to extend dynamic range, and this feedback system may exert a greater effect on the representation of speech in awake animals or humans than it does in anesthetized animals. Also, the dynamic range of the initial onset portion of the responses of an ANF is larger than the dynamic range of the same neuron's adapted response. This is potentially important because abrupt changes in the amplitude envelope (**Figure 2(b)**) can be viewed as a series of onsets. Both of these factors extend the effective dynamic ranges of ANFs, which would improve the representation of spectral shape at higher stimulus levels. Another observation is that small changes in the locations of spectral peaks are detectable as small changes in the discharge rates of individual ANFs. Relative discharge rate preserves information about spectral changes that are small enough to approach the psychophysical difference limen for the frequency of a spectral peak. Finally, as discussed previously, the frequency and amplitude

changes that are typical of ongoing speech are often more salient than those that distinguish one steady-state vowel from another. If discharge rate can preserve spectral information down to the range of psychophysical thresholds for steady-state vowels, then it can also be expected to preserve the highly salient frequency and amplitude cues that are present in normal, time-varying speech.

Time-Varying Stimuli

Additional evidence in support of the utility of the spatial pattern of average discharge rate has been obtained in studies of the neural representation of a specific amplitude cue that occurs at VOT in stop consonant–vowel syllables. As noted previously, VOT is under the speaker's control. However, the distribution of VOTs produced in normal speech is highly nonuniform. For example, English speakers produce the voiced consonant /d/ with VOT less than approximately 30 ms and produce the voiceless consonant /t/ with VOT of 50 ms or longer (a positive number indicates that the onset of voicing lags the release of the stop consonant). Speakers of English reliably do not produce VOTs in the intermediate region between 30 and 50 ms.

Series of syllables can be synthesized such that the correlate of VOT varies in equal steps across a range that includes the intermediate values that are rarely produced in normal speech. Humans, monkeys, chinchillas, and birds all exhibit nonmonotonic psychophysical acuity for VOT when tested with syllables from those series. Acuity is low for syllables with short VOTs in which the consonant is heard as /b/, /d/, or /g/ and for syllables with long VOTs in which the consonant is heard as /p/, /t/, or /k/. However, acuity is higher for syllables with VOT in the intermediate region where the label assigned to the consonant changes. This has led to the suggestion that a natural psychoacoustic or physiological boundary underlies the perceptual distinction between voiced and voiceless stop consonants.

Sinex and colleagues measured the responses of ANFs to tokens from synthesized VOT continua. They found that for low-CF neurons and for VOTs of 20 ms and longer, the onset of voicing produced robust discharge rate increases whose latencies increased with VOT. The rate increase was attributable to the rise in amplitude of 20 dB or more that occurs in the frequency region of F1 at voicing onset. There was no evidence in the responses of individual ANFs for a natural physiological boundary between syllables heard as /d/ and those heard as /t/. However, when responses were pooled across populations of low-CF neurons, a relation between the variance of the responses and psychophysical acuity was observed. Syllables for which psychophysical acuity is high elicited responses with low across-neuron variability. The low variability meant that it was easy to distinguish the neural responses elicited by pairs of syllables that are easily resolvable in psychophysical experiments (Figure 4(a)). However, syllables for which psychophysical acuity is low elicited responses with high variability across neurons; higher response variability meant that there was extensive overlap in the neural responses elicited by pairs of syllables that are difficult to resolve (Figure 4(b)). This pattern of results suggests that the amplitude cues in these syllables are represented by subpopulations of neurons in

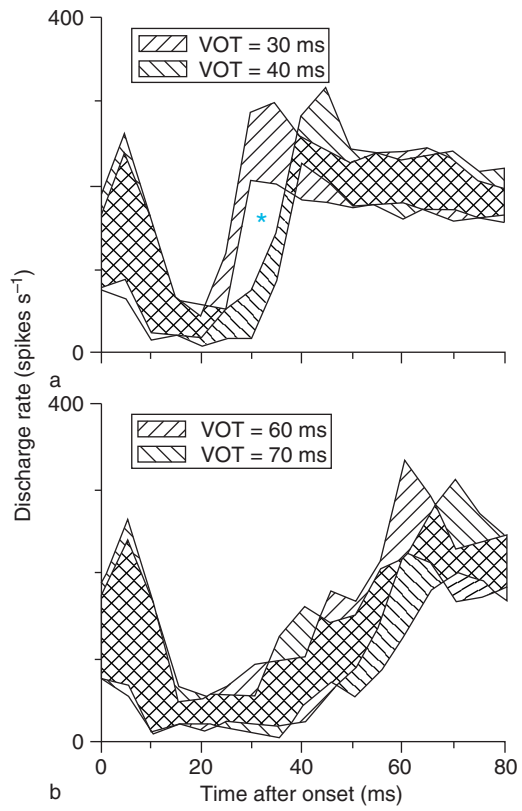

Figure 4 Differences in the precision with which the onset of voicing is represented by ANFs. Each cross-hatched area encloses the mean ±1 standard deviation of the average discharge rates of a subpopulation of ANFs from one animal. (a) Representations of two syllables with voice onset time (VOTs) of 30 and 40 ms, a 10 ms difference in VOT. For these syllables, which are perceptually distinct, there was little overlap between the responses around the onset of voicing (note the gap indicated by the asterisk). (b) Representations of two syllables with VOTs of 60 and 70 ms, again a 10 ms difference in VOT. In this case, the syllables are perceptually similar, and there was extensive overlap between the responses around the onset of voicing. Adapted from Sinex DG, McDonald LP, and Mott JB (1991) Neural correlates of nonmonotonic temporal acuity for voice onset time. *Journal of the Acoustical Society of America* 90: 2441–2449, with permission.

a way that promotes the perceptual separation of the VOT continuum into two easily distinguishable parts. There was no indication that this distribution of responses arose from any specialization on the part of the neurons to process these sounds; rather, it appeared that the patterns were a reflection of specific acoustic properties of the syllables that varied with VOT.

The representation of stop consonants differing in VOT by central auditory neurons has also been studied. Cortical neurons in primates have been reported to respond categorically to VOT syllables. Neurons responded to syllables with VOTs of 0 and 20 ms, which are heard as /da/, with a simple transient response at the syllable onset. The same neurons responded to syllables with VOTs of 40 and 60 ms, both heard as /ta/, with a pattern that included two transient components – one at syllable onset and another at voicing onset. However, cortical neurons in cats studied with syllables from a /ba/–/pa/ continuum did not exhibit comparable categorical response patterns. It is not likely that the use of a different consonant pair accounts for the difference in response patterns, nor does it seem likely that this is simply a species difference, given the apparent similarity of psychophysical processing of VOT by both mammals and birds.

Another possibility is that the responses of cortical neurons across studies seemed to be different because the stimuli were not comparable. Synthesized syllables with the same nominal VOT may differ in other important properties; for example, formant transitions may be simple or complicated. It follows that when one summarizes the response patterns elicited by a set of speech sounds, one is summarizing the responses to those particular tokens, which may or may not be representative of a larger set of instantiations of the same consonant or vowel. Generalizations from the responses to any limited set of speech sounds to larger issues in speech perception should be made with caution.

The increase in F1 amplitude associated with the onset of voicing in stop consonants is a special case in that the pattern of amplitude change is highly consistent across talkers within one language and, to a lesser extent, across languages. The representation of equally salient but less highly constrained amplitude information is also of interest, again because psychophysical observations suggest that the amplitude envelope plays an important role in speech perception. Studies by Robert Shannon and others indicate that speech may be highly intelligible even after processing that removes precise information about the locations of spectral peaks. These investigators found that in quiet listening conditions, as few as four to six frequency channels are sufficient to support high levels of speech recognition. The processing that removes frequency information, called vocoding, preserves only the amplitude envelope in each channel, and that information must account for the intelligibility of vocoded speech.

To a first approximation, the signal seen by an individual primary auditory neuron responding to speech resembles the signal transmitted in one channel of vocoded speech. When the input signal is normal time-varying speech, the range of amplitude variation in each channel is large relative to the dynamic ranges of individual ANFs, as shown in Figure 2(b). The magnitude of variation is large enough to minimize the effects of rate saturation and limited dynamic range that are significant for the representation of spectral details in steady-state vowels. When speech is presented at conversational levels, it is likely that most ANFs will respond at their saturation rates at some times, but it is also likely that no neuron will respond at its saturation rate all the time. In that case, discharge rate can convey quite good information about the amplitude envelope in a restricted frequency band. This is illustrated in Figure 5, which shows the time course of average discharge rate for three representative ANFs, each tuned to a different frequency band, responding to the stimulus shown in Figures 1 and 2. Each neuron exhibited a range of discharge rates in response to the stimulus. Changes in discharge rate generally followed the pattern of amplitude change and thus preserved the kind of amplitude information that supports the recognition of vocoded speech. Since amplitude changes over time are large relative to the dynamic ranges of neurons, they are reflected in discharge rate changes despite the existence of response properties such as rate saturation that limit the transmission of information about steady-state vowels.

Because discharge rate varies over time in a frequency-dependent manner, the spatial pattern of discharge rate in populations of neurons will also vary over time, as shown in Figure 6, which compares rate–place profiles generated from responses occurring at two different times during the word 'heed.' The location of the F1 peak in the vowel spectrum was preserved, but F2 and F3 were represented as one broad peak. Considered in isolation, the loss of resolution in this profile might be viewed as arguing against the utility of a discharge rate code. The second rate profile, generated from responses obtained during the time the vocal tract was constricted to form the consonant, is clearly different from the first, both in shape and in overall discharge rate. As implied previously, the change in the rate–place profile from vowel to consonant was larger than the changes across place during the steady-state vowel or the

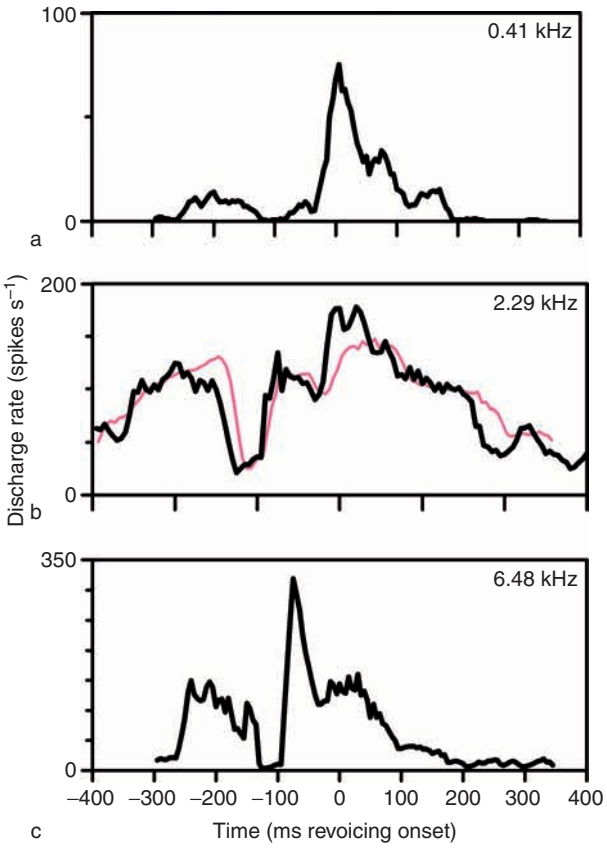

Figure 5 Responses of three ANFs to the two-syllable stimulus shown in **Figures 1** and **2**, presented at 60 dB SPL. Each ANF had a different characteristic frequency (CF), and each responded with a different temporal envelope. The CF of the neuron shown in panel (b) matched the center frequency of the analysis filter used to compute the stimulus envelope shown in **Figure 2(b)** and replotted here as the red line. Changes in the neuron's discharge rate generally followed the stimulus envelope. Adapted from Sinex DG and Narayan SS (1994) Auditory-nerve fiber representation of temporal cues to voicing in word-medial stop consonants. *Journal of the Acoustical Society of America* 95: 897–903, with permission.

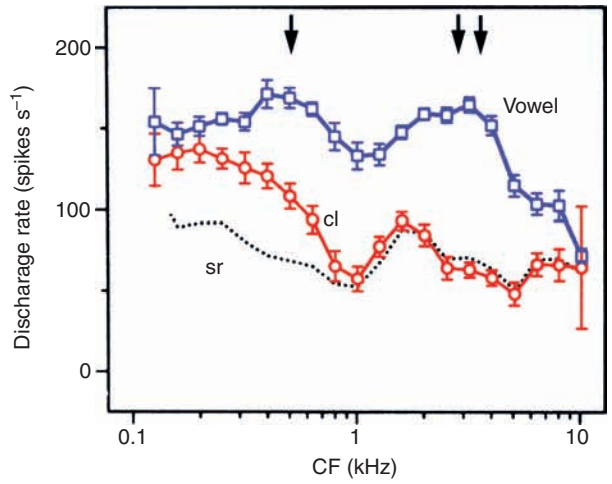

Figure 6 Rate–place profiles obtained at different times during a vowel–consonant syllable, presented at 60 dB SPL. The blue line labeled 'vowel' depicts the spatial pattern of responses during the vowel; the arrows indicate the frequencies of the first three spectral peaks at that time. The rate profile had one peak corresponding to F1 and another single peak in the region of F2–F3. The red line labeled 'cl' depicts the spatial pattern of responses approximately 150 ms later, during the consonant closure interval. Discharge rates were much lower during the consonant than during the vowel, especially for high characteristic frequency (CF) neurons whose rates did not exceed their spontaneous discharge rates (line labeled 'sr'). Adapted from Sinex DG (1993) Auditory nerve fiber representation of cues to voicing in syllable-final stop consonants. *Journal of the Acoustical Society of America* 94: 1351–1362, with permission.

changes in the rate–place profiles elicited by two steady-state vowels.

Representation of Speech by Cochlear Implants

It is reasonable to think that an understanding of how speech is encoded by auditory neurons could be applied to the practical problem of designing cochlear implants, devices intended to deliver sounds including speech to the profoundly deaf. Experiences of cochlear implant users may also provide insight into what is most important in the neural encoding of acoustic speech signals.

Speech undergoes extensive processing before being delivered to the electrode array of a cochlear implant. Modern processors begin by separating the signal into frequency bands, similar to the vocoding process described previously. Within each band, the temporal envelope is extracted and used to modulate the amplitudes of pulses presented at a constant rate at a single cochlear location. If the range of amplitudes is correctly matched to the dynamic ranges of neurons, then envelope peaks can be assumed to generate high discharge rates, and envelope valleys will generate lower rates or no response at all. Many implant users can achieve nearly normal speech recognition with this kind of stimulation, at least in the most favorable listening conditions. In these cases, speech presented via a cochlear implant apparently elicits a pattern of discharge rate increases and decreases that has much in common with the responses that the original acoustic sound would have elicited.

Cochlear implant users receive less benefit from their devices when listening to speech in unfavorable conditions: in the presence of noise, reverberation, or other competing sounds. Contemporary devices also are less successful at reproducing music than speech. In these listening situations, the normal auditory system likely makes use of the information about the fine structure of the stimulus that is present in the pattern

of discharge synchrony. That information is not presently available to implant users because of the loss of waveform fine structure during processing. These results and others suggest that synchronized discharges must play a critical role in the representation of speech in the presence of noise or other competing sounds. Discharge synchrony is known to contribute to neural processing that underlies the segregation of nonspeech sounds, and implant users have difficulty in noisy situations largely because they cannot segregate the competing signals as well as listeners with normal acoustic hearing. Pitch perception is also based in large part on patterns of discharge synchrony. Pitch obviously is critical for the perception of music, but pitch also contributes to the segregation of sounds with different f_0. Together, these observations suggest that the significance of discharge synchrony for the representation of speech is that it aids in the segregation of competing sound sources. It also contributes to generating the perception of pitch and is essential for the localization of sounds including speech. In this view, discharge synchrony is critically important, but it does not make a major direct contribution to the representation of spectral shape.

See also: Auditory System: Central Pathways; Cochlear Mechanics; Cochlear Prosthesis; Connectionist Models of Language Processing; Sentence Comprehension; Speech Perception: Adult; Speech Perception: Development; Speech Production: Adult; Speech Perception: Cortical Processing; Statistical Learning of Language; Temporal Processing in the Auditory Pathway; Word Recognition.

Further Reading

Blackburn CC and Sachs MB (1990) The representations of the steady-state vowel sound /e/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *Journal of Neurophysiology* 63: 1191–1212.

Carney LH and Geisler CD (1986) A temporal analysis of auditory-nerve fiber responses to spoken stop consonant–vowel syllables. *Journal of the Acoustical Society of America* 79: 1896–1914.

Delgutte B (1997) Auditory neural processing of speech. In: Hardcastle WJ and Laver J (eds.) *The Handbook of Phonetic Sciences*, pp. 507–538. Oxford: Blackwell.

Eggermont JJ (1995) Representation of a voice onset time continuum in primary auditory cortex of the cat. *Journal of the Acoustical Society of America* 98: 911–920.

Palmer AR and Shamma S (2004) Physiological representations of speech. In: Greenberg S, Ainsworth W, Popper A, and Fay R (eds.) *Speech Processing in the Auditory System*, pp. 163–230. New York: Springer.

Sachs MB, Voigt HF, and Young ED (1983) Auditory nerve representation of vowels in background noise. *Journal of Neurophysiology* 50: 27–45.

Sachs MB and Young ED (1979) Encoding of steady state sounds vowels in the auditory nerve: Representation in terms of discharge rate. *Journal of the Acoustical Society of America* 66: 470–479.

Sinex DG (1993) Auditory nerve fiber representation of cues to voicing in syllable-final stop consonants. *Journal of the Acoustical Society of America* 94: 1351–1362.

Sinex DG and Geisler CD (1983) Responses of auditory nerve fibers to consonant–vowel syllables. *Journal of the Acoustical Society of America* 73: 602–615.

Sinex DG, McDonald LP, and Mott JB (1991) Neural correlates of nonmonotonic temporal acuity for voice onset time. *Journal of the Acoustical Society of America* 90: 2441–2449.

Sinex DG and Narayan SS (1994) Auditory-nerve fiber representation of temporal cues to voicing in word-medial stop consonants. *Journal of the Acoustical Society of America* 95: 897–903.

Steinschneider M, Schroeder CE, Arezzo JC, and Vaughan HG Jr. (1995) Physiologic correlates of the voice onset time boundary in primary auditory cortex (A1) of the awake monkey: Temporal response patterns. *Brain and Language* 48: 326–340.

Stevens HE and Wickesberg RE (2005) Auditory nerve representation of naturally-produced vowels with variable acoustics. *Hearing Research* 205: 21–34.

Young ED and Sachs MB (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *Journal of the Acoustical Society of America* 66: 1381–1403.

Speech Production: Adult

S M Barlow and S Stumm, University of Kansas, Lawrence, KS, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The adult speech motor control system (including the chest wall (abdomen and rib cage), larynx, velopharynx, tongue, jaw, and lips) represents an anatomically diverse collection of connective tissue–muscle subsystems regulated by a phylogenetically elaborated and distributed neural system. More than 100 muscles are involved in the precise and rapid timing of these muscle subsystems to produce temporally complex speech sounds which are characterized by transitions as short as 10 ms, articulatory velocities approaching 20 cm s^{-1}, and an average speaking rate of approximately 8–12 sounds per second. As a sensorimotor skill, speech is performed with speed and accuracy, improved with practice, highly adaptive in achieving spatiotemporal goals, and relegated to automaticity in the adult speaker. Speech also exhibits the Hebbian principle of motor equivalence, which is defined as the capacity of the motor control system to achieve the same goal or end product with considerable variation among individual components that contribute to that output. In achieving a particular vocal tract goal, the specific contributions of individual speech articulators may vary from one production of a particular element to another, as long as the desired end product is achieved.

Compensatory vowel articulation – presumably involving the operation of motor equivalence between the lips, tongue, and jaw – is defined as the ability of speakers to generate acceptable vowel qualities despite a fixed position of the mandible or other structures. Producing vowels which require a relatively closed jaw position, such as /i/ or /u/, with a spacer held between the teeth fixing the jaw into a very open position confronts a speaker with a new, or at least unfamiliar, task compared with producing the same vowel under normal, nonconstrained conditions. Remarkably, speakers are able to produce vowels which are acoustically correct with regard to formant frequency locations at the first glottal pulse. These studies clearly demonstrate that the tongue is capable of reorganizing motor output to achieve the acoustic product. At least for adult speakers, goal-oriented motor reorganization appears to require no learning since speakers are able to produce the correct vowels under bite-block conditions on the very first attempt, only a few seconds after inserting the bite-block and being cued for what vowel to produce.

Damage to select areas of the nervous system involved in the selection, sequencing, and activation of articulatory muscles will degrade speech production and may reduce intelligibility. Measurement of articulatory muscle performance variables, including kinematics, is central to advancing our understanding of the development of speech movements over the life span and the response to neurological disease and traumatic injury.

Vocal Tract Dynamics

Generating the source excitation and shaping the anterior portion of the vocal tract to achieve a sequence of acoustic targets involves coordinated muscle actions and movements of the chest wall, larynx, velopharynx, tongue, jaw, and lips. The integrity of the underlying performance anatomy (including contractile elements, connective tissue, bone, and the neural substrate) is central to a discussion of motor proficiency during speech. In some instances, the accurate positioning of one structure (i.e., the lower lip) may be dependent on another structure (i.e., the mandible). Motor goal acquisition often involves reorganization of motor patterns for individual structures during the course of speech production. Kinematic studies of speech typically involve recording from multiple structures in an attempt to understand the trading relations between structures, patterns of organization, and reorganization during development or following brain injury or neurological disease. Feedback and predictive or 'forward looking' neural mechanisms are hypothesized to play an important role in the acquisition and maintenance of speech movements.

Measures of Muscle Output

Activation of the articulatory muscles yields a number of measurable outputs, including force, displacement, heat, vibration, and electrical activity. Speech kinematic variables typically studied include the amplitude of displacement, velocity, acceleration, phase and relative timing among multiple articulators, spatiotemporal variability, phase relations to electromyography muscle patterns, and spectral properties of movement (frequency domain).

Methods and Studies of Speech Kinematics

X-Ray Tracking of Orofacial Movements

The X-ray microbeam is a computer-controlled system available in the United States and Japan that uses

a narrow beam of X-rays to localize and track the two-dimensional movements of small gold pellets attached to select speech structures, including the lips, jaw, tongue, soft palate, and eustachian tube. Two additional pellets are attached to the bridge of the nose and a maxillary incisor to serve as reference points that are immobile relative to the skull. The focal X-ray beam (0.4 mm diameter) is scanned at a high rate across a pellet producing a recognizable 'shadow' that is registered on a sodium iodide crystal detector. The path of the X-ray beam toward a pellet is determined by predictions of the position of the pellet generated by current and previous locations.

The X-ray microbeam system has been used successfully to study the relation between speech rate and velocity profiles of movements of the lower lip and tongue tip during the production of stop consonants in young normal adults. Fast speaking rates yielded symmetrical, single-peaked velocity functions, whereas slow speech produced asymmetrical, multipeaked velocity profiles. Speech produced at fast rates appears to involve unitary movements that may be preprogrammed and executed with little or no dependence on sensorimotor integration, whereas articulatory gestures produced at slow speaking rates may be influenced by feedback mechanisms.

X-ray microbeam data have also been combined with cinefluorography to examine the displacement of the tongue body during opening articulatory gestures in deaf and hearing subjects. Speech samples consisted of consonant–vowel–consonant syllables embedded within a carrier phrase. Displacement patterns in deaf and hearing subjects were examined for variation in vocalic contexts between subjects. Deaf speakers had less flexible tongue bodies as a result of compensatory and incorrectly learned principles for constraining tongue movement during speech.

The question of functional regionality within the tongue during speech and swallowing was studied by quantifying the strength of coupling among four different tongue locations. Tongue-surface movement patterns characterized by calculating the covariance between the vertical displacement time histories of all possible pellet pairs show that speech and swallowing kinematics are clustered into distinct groups based on their coupling profiles.

Orofacial Magnetometry

Alternating magnetic field devices known as magnetometers have been used to capture elements of speech production where movements of the face, jaw, tongue, soft palate, and chest wall are tracked in real time. Part of the magnetometer's appeal in speech physiology research and clinical application is due to the fact that the kinematics of intraoral structures (tongue and velopharynx) can be observed and recorded in real time without the biohazards associated with radiological imaging methods (i.e., cineradiography). Historically, the study of speech kinematics has been problematic because movements within the mouth are difficult to measure and visualize.

Orofacial magnetometry is also a valuable therapeutic tool. For example, individuals with hearing loss can observe a visual feedback analog of their tongue movements on a computer in real time. This approach can be applied to the rehabilitation of certain forms of dysarthria, craniofacial disorders, or second language acquisition.

Optical Three-Dimensional Tracking of the Face

Developmental studies of lip and jaw coordination during speech production have benefited directly from the emergence of sophisticated multichannel video-based movement tracking systems. High-speed video cameras are used to register marker location using either reflective markers or infrared source tracking. An algorithm has been developed by researchers at Purdue University to index the spatial and temporal variability in the kinematic trajectory patterns for facial articulators during repeated productions of select speech utterances. When applied to the development of functional synergies among labial–mandibular systems, they found that the time course of development for speech motor coordination is protracted and does not reach 'adult-like' performance until after age 14 years for both males and females, with males lagging on the spatiotemporal variability index.

Tracking Tongue Movements

From the late 1940s through the mid-1970s, our understanding of tongue movements in adult speakers relied heavily on studies using radiography and tongue contact devices such as the dynamic palatometer. Since then, technological advances have been made in optics, magnetometry, ultrasound, X-ray microbeam, and functional magnetic resonance imaging (fMRI) for tracking the tongue during speech and swallowing.

Palatometry has been used to show that each vowel in English is associated with a unique stationary lingua-palatal contact map. For example, during the stable contact portion of [i], the tongue is in contact with sensors extending from the cuspid–bicuspid region of the palate to the posterior border of the alveolar ridge. During the [ae], the contact is against the most posteriolateral sensors. Diphthongs are

characterized by movements between two stable monophthong positions. Optical and contact pattern (tongue–palate) tracking devices have been used in training or retraining vowel space and consonant production in hearing-impaired adults.

Ultrasonic Imaging of the Tongue

Ultrasound provides real-time images of the tongue surface in a digital video format during speech and swallowing. Images are combined with estimates of the hard palate boundary using ultrasound during the production of command swallows to establish a reference within headspace for coregistration among subjects and calculation of select phonetic measures.

Ultrasound results indicate that cross-sectional tongue shape is directly related to the position of the tongue and the lateral and sagittal shape of the tongue. In general, midsagittal grooving is evident for all vowel types, with posterior grooves being deeper than anterior grooves. Posterior grooving is greater in the /p/ context than in the /s/ context. Grooving for vowels in the /p/ context demonstrated a continuum, whereas in the anterior /s/ context two groups of vowels were identified (high group/shallower grooves and back group with deeper grooves). In the anterior /s/ context, tongue shape for /i/ and /u/ was convex.

Considerable progress has been made with high-speed three-dimensional (3-D) reconstruction of tongue ultrasonic images. In adult speakers, this work has focused on defining predictable mathematical relations between midsagittal tongue contour shapes and up to five cross-sectional (coronal) contours. A concatenation of these coronal slices permits the reconstruction of a 3-D tongue surface. When compared among four tongue shape phonemic categories (i.e., front raising /i/, back raising /ŋ/, continuous grooving /ae/, and two point displacement /l/), it was found that transitional values exist that predict tongue arching versus midsagittal tongue grooving. Even better predictions are possible using three anterior coronal slices in which strong correlations exist between midline displacement and groove depth to arch height. Knowledge of both category and midsagittal displacement provides good prediction of coronal tongue shape.

Tracking Velar and Laryngeal Movements

Kinematic studies of the 'invisible valves' of speech, including the velopharynx and larynx, is technically challenging due to small size and inaccessibility. Acceptable sampling methods in humans usually reflect a concession between the invasiveness of the instrument and the quality of the acquired signals. Whereas it is acceptable to attach small sensors to the lips and jaw in the form of radio-opaque pellets, reflective balls or tape, infrared diodes, cantilevers, or magnetic coils to the rib cage or abdomen, no such electromechanical device or sensor is acceptable to the delicate tissue boundary of the human vocal fold because this would disrupt the behavior and health of the organism. Instead, techniques have evolved that rely on imaging, acoustics, and/or fluid mechanics. These procedures are described in the following sections.

Velopharynx

The velopharynx (VP) is a complex muscular valve that is strategically situated to divert acoustic and aerodynamic energy through the oral and nasal cavities. The size of the velar port determines the oral or nasal nature of speech sounds. Movements of the velum, lateral pharyngeal walls, and the posterior pharyngeal wall collectively determine the size of the velar port.

Methods aimed at measuring the size and/or movements of the velum and lateral and posterior pharyngeal walls roughly fall into one of two categories: direct and indirect. Direct methods include imaging techniques such as cineradiography, video nasendoscopy, and electromechanical, optomechanical transduction of velar displacement. These transducers for measuring velar activity and radiographic imaging techniques share a common limitation in only resolving movement in a single plane. In most radiographic studies, discrete points are tracked on a frame-by-frame basis, which is useful in resolving velocity and displacement profiles. However, because the radiographic methods are limited to one plane or slice through the velopharynx, one is never certain if closure has occurred. Velar apertures may exist at locations opposite the lateral pharyngeal walls on one or both sides. The patterns of VP closure are highly variable both within and across adult speakers.

Complete VP closure is not always obtained during vowel production. Pioneering work aimed at characterizing normal patterns of VP closure using cineflourographic techniques revealed that high vowels exhibit greater VP closure than low vowels, regardless of consonant context. Complete closure of the VP is not always present during production of the low vowels nor attained on vowels adjacent to a nasal consonant (assimilation). Speech rate is another factor which can significantly influence the degree of VP closure.

Flexible fiberoptic nasoendoscopy is a powerful imaging tool which helps to resolve some of the uncertainty regarding the dynamics of the velopharyngeal

port. Advanced camera, fiberoptics, and digital recording technology provides high-resolution color images of the velopharynx in real time during speech production. Image recognition software routines assist the investigator in accurate identification of tissue boundaries and kinematic properties of the velopharynx, including computation of portal area and boundary velocity, displacement, and calibration schemes for scaling distance. This information will be invaluable for studies of velopharyngeal motor control in patients with sensorimotor speech disorders.

Aerodynamic measures of velopharyngeal port function offer some unique perspectives on the fluid dynamics of this valve during speech in adults and children. Current aerodynamic protocols and instrumentation provide reasonably accurate estimates of the subglottal driving air pressures (P_s) and airflows acting on the velopharynx during speech in fluid mechanics terms without the biohazards associated with radiation or the invasiveness of a fiberoptic bundle inserted into the nasal cavity. Area functions, resistance (impedance) estimates, and temporal pattern studies have been used effectively to characterize the activity of the velopharynx during speech.

Numerous reports have described some of the temporal relations between pressure–flow variables during nasal-plosive blends to stress the velopharyngeal mechanism and reveal the coarticulatory dynamics between velopharynx and other upper airway structures in normal and cleft palate speakers. For example, studies of VP aerodynamics in repaired cleft palate adults and normal controls revealed that the magnitude of mean intraoral air pressures tends to be slightly less than that of control speakers. Articulatory timing errors were evident as the nasal airflow pulse overlapped into the rising phase and peak of the pressure pulse associated with /p/ in the word 'hamper.' A decrease in respiratory effort may represent a compensatory strategy used by patients with repaired cleft palates to achieve adequate velopharyngeal closure and minimize shunting (air leak) through the velopharyngeal port. This conclusion was based on careful study of the temporal relations between the airflow and pressure curves associated with production of the nasal-plosive blend.

Laryngeal Kinematics

The larynx represents a microcosm of the entire speech mechanism in that it provides a sound source in coordination with the respiratory system, acts as a dynamic 'articulator' capable of rapid adductor and abductor adjustments, modulates pitch (male f_0 ~125 Hz, female f_0 ~225 Hz), and conveys emotion and personal identity. Endoscopy is an important tool in the assessment of laryngeal movement disorders affecting speech/voice. Laryngeal aerodynamics is a useful adjunct which provides quantitative data on the articulatory dynamics of laryngeal engagement for consonant–vowel transitions, vocal efficiency, and chest wall function for P_s regulation. Use of fiberoptic nasopharyngoscopy, laryngoscopy, and stroboscopy in conjunction with low-light cinematography allows observation of the dynamic processes of speech production. This technique allows the investigator to visualize the overall articulatory dynamics of the larynx during speech and other behaviors such as swallowing, coughing, and respiration. Information obtained during the nasendoscopic examination has proved to be useful for biofeedback in the remediation of select laryngeal and velopharyngeal impairments affecting speech.

Videokymography

A relatively new technique that takes advantage of digital video recording and processing in strip kymography (the recording of wavelike motions or undulations) of the glottis is videokymography (VKG). This method permits visualization of the individual cycles of phonation by analyzing only a small segment of the vibrating vocal folds. Developed in The Netherlands by Drs. Harm Schutte and Jan Švec of the University of Groningen, VKG uses a modified charge-coupled device camera to record images of selected cross-sections of the vocal folds during motion. The position of the horizontal measuring line is maintained throughout the analysis. VKG systems use standard videolaryngoscopic setups with a continuous light source and record images to a digital storage device. In place of recording the entire image of the vocal fold, however, VKG systems allow for the collection of a single line of the image (orthogonal to the longitudinal axis of the fold) at rates of 8 kHz, thus allowing for sampling rates great enough to cover the entire frequency range of fold oscillation. Successive lines are temporally aligned and cascaded in real time on a conventional monitor, thus showing the vibratory pattern of the selected region of interest on the vocal fold for a preselected time period. The resulting 'kymogram' displays a spatiotemporal image of the vocal folds at the scanned line over time. Videokymography and digital kymography allow for true cycle-by-cycle analysis of vocal fold vibratory patterns. Analysis of the resulting kymograms can provide information on right–left vocal fold asymmetries, open and closed quotients, ventricular fold and/or mucous interference, mucosal wavefront propagation, frequency, amplitude, speed quotient, and within-cycle perturbations.

Electroglottography (EGG) has been used with success to assess laryngeal activity. This instrument consists of a flexible neck collar supporting an array of electrodes and a signal conditioning unit. In EGG, a small DC bias current is fed through the tissues of the neck and the conditioning unit senses changes in the electrical resistance in the region of the larynx. Essentially, the EGG signal is correlated with vocal fold contact area and is regarded as a useful assessment technique in drawing inferences about patterns of vocal fold vibration during speech.

Electroglottographic measures have been used to measure timing between voiced segments and for analysis of the cycle-by-cycle dynamics of voicing, and they are usually obtained in combination with inverse filtered flow signals (AC flow) for the purpose of detailing the organization and timing of the open and closed phases of the glottis during voice production. Because information derived from EGGs is based on the vocal segment and changes in vocal fold contact area, it offers limited information concerning the aeromechanical events underlying vocal fold engagement. Furthermore, the reliability and validity of the EGG signal associated with the early phases of laryngeal engagement are questionable. The beginning phases of arytenoid rotation are associated with displacement of the vocal folds in glottal space. During this interval, one would not expect appreciable increases in vocal fold contact area until arytenoid advancement results in actual tissue approximation. Therefore, the EGG output during this phase of engagement may yield little or no output signal even though significant displacements of the vocal folds have occurred.

Laryngeal Aerodynamics

The utility and noninvasive nature of aerodynamics is underscored in adults with neurogenic speech disorders. Voice disorders range from hyperfunctional (excessive medial compression, hard glottal attack, and groping voice quality) to hypofunctional (soft glottal attack, breathy, and soft or weak voice) and disturbances in the timing and coordination of arytenoid–VF medialization (basal ganglia (BG), cerebellar disease, upper motor neuron syndrome, and others). For example, speakers with Parkinson's disease often exhibit air loss when attempting to engage (adduct) the vocal folds for voice production. In contrast, patients with adductor spasmodic dysphonia manifest exaggerated medial compression of the vocal processes of the arytenoids. This pattern of hyperadduction may include the ventricular or false folds and results in a sudden interruption of the breath stream during voicing. Disruption of this important articulatory adjustment due to neural and/or biomechanical factors can dramatically influence the manner in which the vital capacity is used for speech. In cases of slow or mechanically limited engagement, significant portions of the lung vital capacity may be wasted by a defective laryngeal articulatory apparatus independently from the pressure–flow dynamics associated with voice production.

Real-time digital speech aerodynamic recording and analysis tools are available to measure articulatory proficiency during laryngeal engagement and phonation. This technique, based on the simultaneous sampling of intraoral air pressure, translaryngeal airflow, and voice acoustic signals, has been used effectively to reveal the reorganization of laryngeal function in patients with amyotrophic lateral sclerosis, idiopathic Parkinson's disease following posteroventral pallidotomy, and bilateral subthalamic nucleus deep brain stimulation. The magnitude and time course of translaryngeal flow, described mathematically as air volume and flow rate declination functions, are presumed to reflect the underlying kinematics of vocal fold adduction toward the phonatory phase. Measures of pressure–flow dynamics and laryngeal airway resistance can also be determined for the phonatory phase of syllable production to provide for a comprehensive evaluation of the articulatory dynamics of laryngeal behavior. The net result of a hypokinetic adductory mechanism is that lung volume is wasted during translation of the arytenoids to achieve vocal fold approximation. The lung volume available for speech is depleted and overall utterance length is decreased. Therefore, the evaluation of laryngeal function in patients with sensorimotor speech disorders benefits from an analysis of the kinematic properties of transitory (engagement–disengagement) and phonatory (voice efficiency) phases.

Chest Wall Kinematics

The chest wall includes, except for the lungs and airways, all parts of the respiratory apparatus – the rib cage, diaphragm, and the abdomen and its contents – and is regarded as one of the principal articulatory systems that provide a pressure–flow source to support speech and voice production. The rib cage and diaphragm constitute the thoracic cavity. The abdominal cavity is defined by the diaphragm and abdominal wall bounding an incompressible mass of liquid. According to Hixon's model, the chest wall is reduced to a two-structure model consisting of the rib cage and 'diaphragm-abdomen.' Isovolume maneuvers are

used to calibrate the chest wall size against lung volumes and involve equal and opposite volume displacement by the surfaces of the abdomen and rib cage under a closed system. Movement outputs from the rib cage and abdomen can be adjusted to account for the difference in volume displacements from these two parts. Once the chest wall has been calibrated, lung volume calibrations can be achieved by measuring integrated flow at the airway opening during simultaneous chest wall kinematic measurements to establish the relationship between chest wall movement and various static inspiratory or expiratory lung volumes.

Magnetometry

Changes in the anterior–posterior diameters of the rib cage and abdomen are transduced with two-channel electromagnetic devices consisting of small coils (2×0.5 cm) set up in pairs, one for signal generation and the other for sensing changes in inductance due to displacement. The generator coils are attached midline on the anterior surface of the chest wall at the level of the nipples and to the abdomen immediately above the umbilicus. An alternative transducer system that has been used to measure chest wall kinematics is the strain gage belt pneumograph. Setup is simple and involves a clip gage attached to a circumferential belt with an elastic strap.

Speech breathing kinematics using magnetometers has been studied in individuals with profound hearing impairments, voice disorders, motor neuron disease, flaccid paralysis, parkinsonian dysarthria, cerebellar ataxic dysarthria, traumatic brain and spinal cord injury, and in a variety of normal subgroups in which factors such as age, body type, and sex are the dependent variables.

Functional Neural Systems for Speech Motor Control

The encoding of force and movement by the brain is central to models of movement and highly relevant to our understanding of speech motor control. The relation between force control and neural firing patterns in primate motor cortex is strongly correlated with the rate and direction of force change and preferred torque directions. Primary motor cortex has also been shown to encode functional muscle synergies. The cortical mechanisms involved in the regulation of muscle dynamics during the early phases of force recruitment have been shown for orofacial muscles. The complex articulatory force dynamics evident among facial muscles during speech is similar in many ways to the precise movements and forces generated by hand and fingers during manipulation. During speech production, the muscles of the perioral system exhibit a streaming repertoire of phasic adjustments in force from resting tonic activation to approximately 10% of the maximum voluntary contraction level.

Mounting evidence also suggests that the sensorimotor cortex may participate in the modulation of the reflexive compensatory motor responses in lip muscles during speech. The orofacial apparatus appears to utilize mechanosensory information differentially for the early versus later phases of lip force output. During speech articulation, longer latency reflex actions presumably involving primary motor cortex correct for the effects of external movement disturbances. Mechanical inputs delivered to a dynamically contracting lip muscle produce a compound evoked perioral response characterized by phases of excitation and suppression that are absent during static force conditions. Under these conditions, compensatory lip movements demonstrate a well-calibrated readjustment.

Velocity is one kinematic variable which has been explored in some detail and appears to have a definable neural network. Positron emission tomography mapping of regional cerebral blood flow has demonstrated a velocity subcircuit within the motor control system, including the bilateral sensorimotor cortex, BG (putamen and globus pallidus (GP)), and the ipsilateral cerebellum. Increases in velocity are associated with parallel increases in BG activation. The BG motor circuit may be preferentially involved in controlling or monitoring the scale and/or dynamics of arm movements and may also control articulatory movements at lower frequencies. A dysfunctional BG circuit (i.e., Parkinson's disease (PD)) may result in hypophonia (lowered speech volume) and bradykinesia. Functional brain imaging of patients with PD during velocity control tasks reveals a significant disruption to the BG circuitry which may partially account for the pathophysiology of bradykinesia. Unilateral lesions of the GP internal and GP external result in slowness of movement and abnormal co-contraction of agonist and antagonist muscles and often impair movements of the orofacial, laryngeal, and chest wall systems for speech production.

A positive relation also exists between increases in speech rate and cerebellar activation, suggesting this brain structure plays a significant role in the velocity of articulatory movements within the frame of time-critical adjustments. Motor speech symptoms resulting from cerebellar pathology may reflect a distorted process of articulatory planning and coordination. Cerebellar posterior lobe syndrome is often

associated with speech that is perceived as having an 'intoxicated-like' quality.

The brain imaging and neurocomputational literature is proliferating rapidly with studies of speech production and sensorimotor control of the upper airway. The production of simple vowels and simple oral movements results in similar activation patterns among cortical and subcortical motor systems. More complex, polysyllabic utterances are associated with additional activation of the bilateral cerebellum, presumably reflecting increased demands on speech motor control, and additional activation of the bilateral temporal cortices thought to subserve phonological processing.

The hemodynamic effects associated with paced syllable production during fMRI have been compared with passive listening of pacing signals. Cerebral structures associated with speech motor control included the supplementary motor area (SMA), primary sensorimotor regions, dorsolateral prefrontal cortex (including Broca's area), anterior insula, thalamus, putamen/pallidum, and cerebellar hemispheres. Dorsolateral prefrontal and intrasylvian cortex displayed lateralization toward the left side, whereas the other components showed a bilateral activation pattern. The cerebral correlates of speech motor control based on connectivity analyses appear to be organized into two separate networks – motor preparation (SMA, dorsolateral frontal cortex (Broca's area), anterior insula, and superior cerebellum) and a motor execution processor (sensorimotor cortex, BG, thalamus, and inferior cerebellum). Broca's area is thought to contribute to orthographic-to-phonological transformation, whereas motor cortex, SMA, and the cerebellum play critical roles in phonetic encoding and articulation of speech. Functional imaging studies show that pre-SMA and SMA are engaged in the control of externally triggered movements. Moreover, SMA shows a frequency effect and may indicate that this region is involved in the generation of the motor program for an entire spoken word, with the level of activation in SMA influenced by how often this motor program is implemented. Motor cortex appears to represent articulatory gestures at the syllable or phoneme sequence level. Clinical data support this dichotomy of motor preparation/motor execution.

Speech production is dependent on auditory, somatosensory, and visual feedback for its development and maintenance. These mechanisms are also crucial

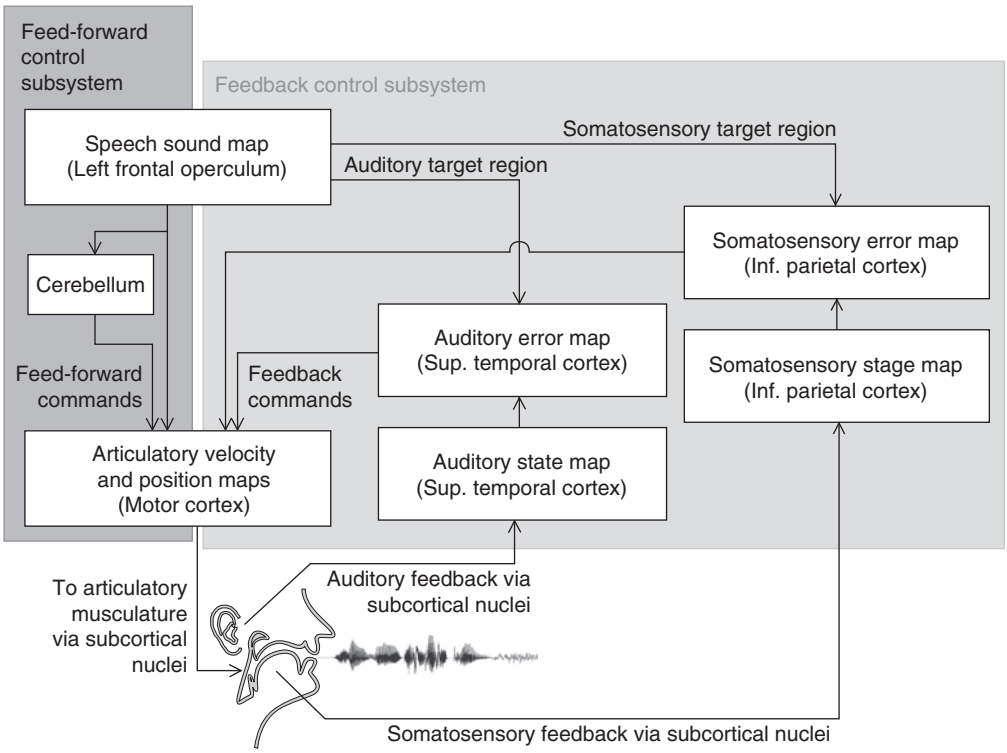

Figure 1 Schematic of the DIVA model of speech acquisition and production. Projections to and from the cerebellum are simplified for clarity. Reproduced from Guenther FH (2006) Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders* 39: 350–365, with permission.

to invoke mechanisms of plasticity following a sudden insult to the brain (ischemia reperfusion or hemorrhage) or during the course of a progressive neuromotor disease (i.e., PD). A simulation has been developed to test many of these conditions as given by the Directions into Velocities of Articulators (DIVA) model of speech motor control, shown schematically in Figure 1. The DIVA model includes a feed-forward network, composed of the premotor and primary motor cortices and cerebellum, and feedback control systems which encode auditory and somatosensory inputs with comparator functions to assist in the specification of speech articulatory velocity and position. A central feature of the DIVA model that differentiates it from other computational models of speech production is that all components of the model have been associated with specific neuroanatomical loci. The DIVA model generates simulated fMRI activations associated with small populations of synchronously firing neurons from the model's cell activities during computer simulation. It features adaptive control mechanisms to effect the learning of auditory and somatosensory consequences during babbling or syllable productions which are eventually incorporated into feed-forward mechanisms. According to the DIVA model, outputs from premotor cortex (left frontal operculum) and cerebellum to motor cortex constitute feed-forward motor commands for syllable production. In infants, feed-forward commands for a syllable are tuned on each production attempt. Initial attempts rely heavily on auditory feedback. With changes in body plan (i.e., growth of the speech articulators), the auditory feedback control system continues to provide corrective tuning of the feedback controller over the life span. Consistent with research in other laboratories, it is clear that somatosensory information is central to achieving the precision requirements of speech movements, including consonant and vowel position targets. The neural control of stiffness appears to be a key factor to consider for the somatosensory precision involved in speech production. Future work by the developers of the DIVA model will incorporate the anterior cingulate cortex, SMA, BG, and anterior insula into the scheme of feedback and feed-forward control of speech movements.

See also: Emotion in Speech; Evolution of Vertebrate Respiratory Control; Motor Sequences; Motor Timing; Orofacial Motor Control; Psycholinguistics; Sentence Production; Speech Production: Development; Trigeminal Motor System; Word Production.

Further Reading

Adams SG, Weismer G, and Kent RD (1993) Speaking rate and speech movement velocity profiles. *Journal of Speech and Hearing Research* 36: 41–54.

Baken RJ (1992) Electroglottography. *Journal of Voice* 6: 98–110.

Barlow SM, Farley GR, and Andreatta RD (1999) Neural systems in speech physiology. In: *Handbook of Clinical Speech Physiology*, pp. 101–165. San Diego: Singular Publishing.

Cunnington R, Windischberger C, Deecke L, and Moser E (2002) The preparation and execution of self-initiated and externally-triggered movement: A study of event-related fMRI. *NeuroImage* 15: 373–385.

Estep M and Barlow SM (2007) Modulation of the trigeminofacial pathway during syllabic speech. *Brain Research* 1171: 67–74.

Gracco VL, Tremblay P, and Pike B (2005) Imaging speech production using MRI. *NeuroImage* 26: 294–301.

Guenther FH, Ghosh SS, and Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96: 280–301.

Hixon TJ (1987) Respiratory function in speech. In: Hixon TJ (ed.) *Respiratory Function in Speech and Song*, pp. 1–54. Boston: Little, Brown.

Levelt WJM (1989) Speaking: *From Intention to Articulation*. Cambridge, MA: MIT Press.

Ludlow CL and Bassich CJ (1984) Relationships between perceptual ratings and acoustic measures of hypokinetic speech. In: McNeil MR, Rosenbet JC, and Aronson AE (eds.) *The Dysarthrias*, pp. 163–195. San Diego: College-Hill Press.

Munhall K (2006) Speech production: The force of your words. *Current Biology* 16(21): 922–923.

Nasir SM and Ostry DJ (2006) Somatosensory precision in speech production. *Current Biology* 16: 1918–1923.

Perkell JS, Cohen MH, Svirsky MA, Matthies ML, Garabieta I, and Jackson MTT (1992) Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America* 92: 3078–3096.

Riecker A, Mathiak K, Wildgruber D, et al. (2005) fMRI reveals two distinct cerebral networks subserving speech motor control. *Neurology* 64: 700–706.

Schutte HK, Svec JG, and Sram F (1998) First results of clinical application of videokymography. *Laryngoscope* 108 (8): 1206–1210.

Smith A, Johnson M, McGillem C, and Goffman L (2000) On the assessment of stability and patterning of speech movements. *Journal of Speech, Language, and Hearing Research* 43: 277–286.

Soros P, Sokoloff LG, Bose A, McIntosh AR, Graham SJ, and Stuss DT (2007) Clustered functional MRI of overt speech production. *NeuroImage* 32: 376–387.

Stone M, Epstein MA, and Iskarous K (2004) Functional segments in tongue movement. *Clinical Linguistics & Phonetics* 18: 507–522.

Turner RS, Desmurget M, Grethe J, Crutcher MD, and Grafton ST (2003) Motor subcircuits mediating the control of movement extent and speed. *Journal of Neurophysiology* 90(6): 3958–3966.

Turner RS, Grafton ST, McIntosh AR, DeLong MR, and Hoffman JM (2003) The functional anatomy of parkinsonian bradykinesia. *NeuroImage* 19: 163–179.

Speech Production: Development

R D Kent and K C Hustad, University of Wisconsin–Madison, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Speech may be defined as the production of vocal sounds for communication through the processes of respiration, phonation, resonation, and articulation. This definition links several fundamental concepts, including sound (the acoustic signal of speech), communication (the purpose of speech), and underlying physiology (the biological systems involved in speech production). The development of speech can be described in varying levels of detail for each of these components. From the point of view of the neurosciences, speech is particularly interesting for several reasons. First, speech is a remarkable motor skill. Ordinary conversational speech can be produced at rates of up to six to nine syllables (20–30 individual speech sounds) per second. No other discrete human motor performance rivals that rate. Speaking also commands more motor fibers than any other human mechanical activity, and there is growing evidence that the motor fibers are uniquely equipped for the requirements of speech. The muscles used in speech production differ from skeletal and limb muscles in their developmental patterns, functional properties, and molecular phenotypes. Perhaps most important, speech is the primary modality of language in the vast majority of human societies (one important exception being the manual communication systems used by the deaf). Speech and language are often confused by laypersons, partly because they seem inextricably linked in the learning and use of language. In fact, it is difficult to study speech separate from language.

At a relatively abstract level, speech development can be charted with respect to the child's mastery of its phonemes. A phoneme is a basic sound unit that has the linguistic function of distinguishing morphemes (the minimal units of meaning in a language). The words 'cat,' 'hat,' 'mat,' and 'sat' are four morphemes that differ in their initial phonemes. To become a competent speaker of American English, a child must master the 44 phonemes that are used to form the morphemes and words of the language. The primary goal of the perceptual characterization of speech is to describe what speech sounds like to listeners. To describe speech development using perceptual methods, the most important tool is the human ear. Perceptual features of speech can be considered according to a number of different dimensions related to the speech subsystems. However, for children who are neurologically normal or typically developing, a primary interest is how production of phonemes, or articulatory competence, emerges developmentally.

Studies of phonetic development have primarily taken two approaches. The first is a relational paradigm in which speech segments (phones) produced by children are compared with adult forms. This approach has yielded normative data that have been used to identify several acquisition points, primarily for consonants. These include the point at which children acquire 'mastery-level' production accuracy (i.e., the ability to produce an individual segment correctly in all occurrences) and the point at which children acquire 'customary' production accuracy (i.e., the ability to produce individual segments correctly more often than incorrectly). The second paradigm for studying speech development has been called the independent approach. In this paradigm, the phonetic repertoire of children's speech is described without comparison to adult forms; thus, mastery is not of interest. Instead, the goal is characterization of phones produced in the child's spontaneous speech. The primary vehicle for describing acquisition of speech sounds in both approaches is phonetic transcription of elicited and/or spontaneous productions. Similar methods have been used to describe the perceptual characteristics of speech from infancy through adulthood.

At a more detailed and concrete level, speech development can be specified in terms of a host of anatomic changes (such as those relating to the growth and maturation of the anatomy of speech production), a gradual refinement of perceptual systems, and an expanding capacity to produce sounds flexibly. Much of the recent progress in the understanding of speech development has come through the application of tools to study the anatomy, physiology, acoustics, and perception of speech. The discussion in this article synthesizes these sources of knowledge, using phonetic accomplishments at different ages as both developmental milestones and a conceptual framework.

Developmental Anatomy and Physiology of Speech

Speech is produced through the coordinated actions of muscles in three major systems – respiratory, laryngeal, and articulatory. Each of these is introduced briefly as background information for the subsequent sections on speech development. As will be seen, each system has its own developmental pattern so that speech development must be understood against the composite development of a complex biological system.

For basic life purposes, the respiratory system exchanges blood carbon dioxide (CO_2) with air oxygen (O_2), but it is also used to provide the aerodynamic energy for speech. The major adjustment for speech breathing, as opposed to rest breathing, is that the expiratory phase is prolonged for speech so that several syllables can be uttered on each expiration. That is, the basic pattern of speech breathing is for a quick inspiration and a long, controlled expiration. Major life span aspects of respiratory anatomy and physiology are summarized in Figures 1–4. Anatomic and functional changes in the respiratory system are substantial, especially in the infant and young child. By approximately 7 years of age, children have respiratory patterns resembling those of adults, and subsequent changes are mostly increases in capacity.

The larynx, or voice box, is not only the source of vibratory energy for speech sounds but also an important valve that regulates airflow into and out of the lungs. The birth cry is typically a child's first venture into vocalization. This sound is made with air that is expelled from the inflated lungs to produce vibrations of the vocal folds (often called vocal cords, but the term vocal folds is anatomically more accurate). Despite this early use of the larynx, its anatomy and physiology have a protracted development that includes some remarkable features such as the adolescent voice change. Major developmental events relating to the larynx are summarized in Figures 1–4. One point to be emphasized is that the laminated structure evident in the adult vocal folds (and a feature that is central to modern theories of vocal fold vibration, such as the body-cover model) is not completed in children until approximately 12 years of age.

If the respiratory organs and the larynx are considered to be the major source of energy for speech, then the articulatory apparatus, or vocal tract, can be regarded as the place where the distinctive properties of most speech sounds are formed. It is through the actions of the articulators (tongue, velum, lips, and jaw) that vowels and consonants are made. The growth and development of the articulators involves not only an increase in size but also a number of reconfigurations that affect their individual shapes and their relationships with one another. Humans begin life with a vocal tract resembling that of a chimpanzee, but this system is gradually remodeled during childhood until the distinctively human vocal tract is formed. The word 'distinctively' is apt because no other creature has a vocal tract with similar anatomic and physiologic properties. Moreover, the distinctiveness may even be found at the level of molecular biology, given that human speech muscles appear to have muscle fiber compositions unlike those of homologous muscles in nonhuman species. Some of the most important developmental features are identified in Figures 1–4. Growth and maturation of this system continues until at least 16 years in females and 18 years in males.

Methods for Studying Speech Development

Speech development has been studied with a variety of approaches, primarily the methods of perceptual phonetics, acoustics, and physiology (including aerodynamics, kinematics, and electrophysiology). Perceptual methods usually focus on sound mastery – the degree to which a child's productions match adult sound patterns. Mastery is described according to a statistical criterion, such as 75% correct, which means that a child's productions of a sound match the adult target 75% of the time. With such a criterion, it can be shown that certain sounds are mastered at an early age of 2 or 3 years, whereas other sounds may not be mastered until the age of 7 or 8 years. The perceptual method is easy and inexpensive but not necessarily valid or accurate, especially when applied to infant vocalizations such as babbling.

Acoustic analysis is typically done with the techniques of digital signal processing. Acoustic analysis has an objectivity and quantification that give it several advantages over perceptual methods such as phonetic transcription. However, it requires sophistication on the part of the user and can be rather tedious if a large set of measures is obtained. Acoustic analyses are an effective complement to perceptual analysis in that both are derived from the same acoustic signal.

Several physiological methods have been used to study speech in adults, but relatively few of these have been extensively used with children. Aerodynamic analysis is a type of physiological method that makes use of recordings of air pressures, airflows, or air volumes that are associated with the production of speech sounds. Like acoustic analysis, it offers advantages of objectivity and quantification. However, it requires special instrumentation that is sometimes difficult to use with young children. The most commonly used aerodynamic measures are intraoral air pressure for oral consonants, airflow for sustained vowels, and glottal (laryngeal resistance). Kinematic analysis is focused on movements. It attempts to record the positions, velocities, and accelerations of individual articulators, as well as the relative timing of different movements performed by two or more articulators. Kinematic data are obtained from specialized equipment, and most methods provide information on only

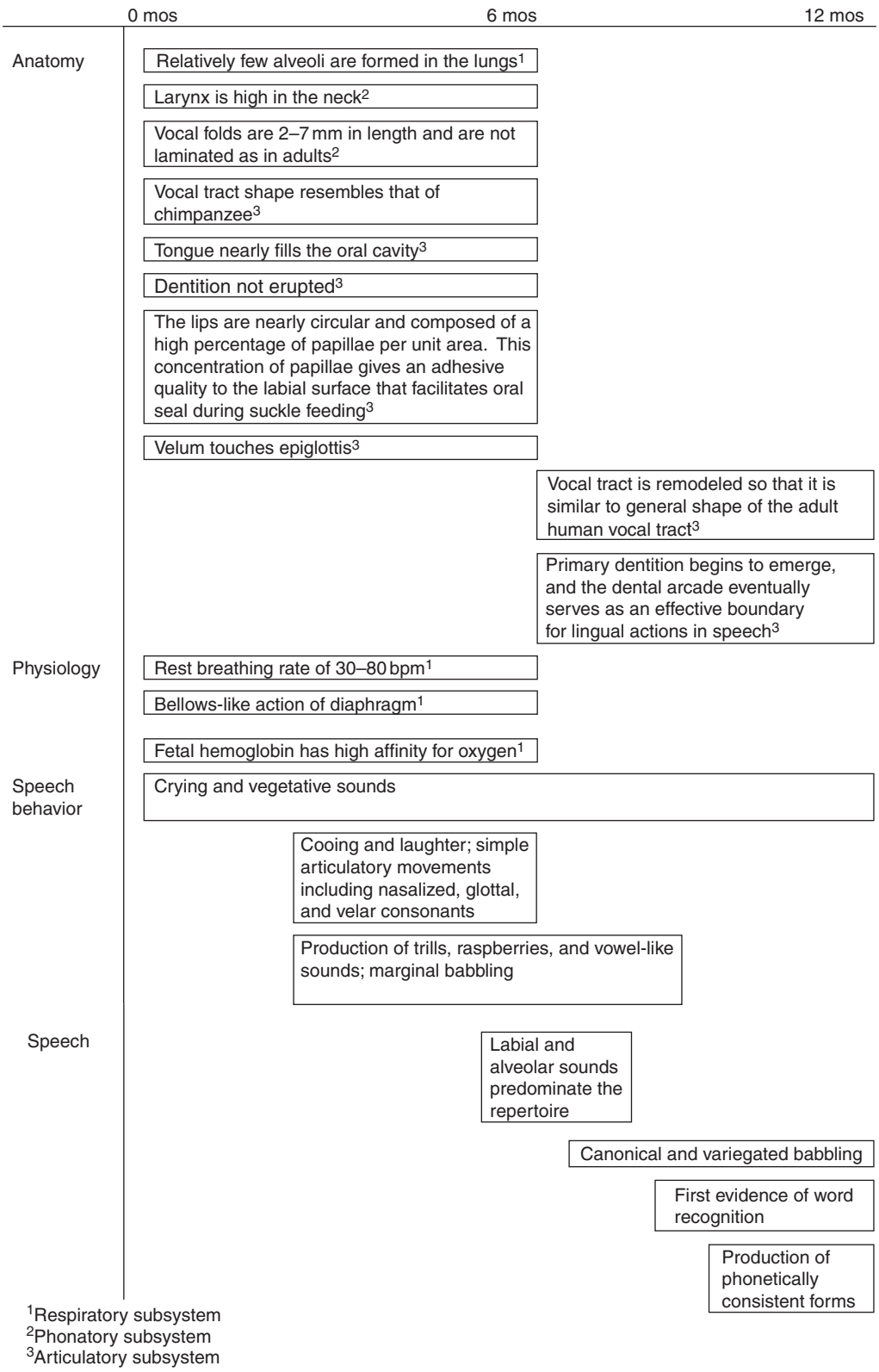

Figure 1 Schematic representation of major speech-related developmental milestones occurring in the first 12 months of life. Note that ages and age ranges are approximate and are meant to provide a simplified illustration of the co-occurrence of developmental events.

the most accessible articulators, especially the lips and jaw. Occasionally, kinematic studies are accompanied by electromyography so that movements can be interpreted relative to data on muscle contraction.

In the following discussion, acoustic and physiological methods are grouped together in the section titled 'Instrumental findings.'

Speech Development in Infancy (0–12 Months)

This period covers the progression from cries and vegetative sounds to the onset of the first words, usually at approximately 12 months. A major point to be made about this early period is that it involves

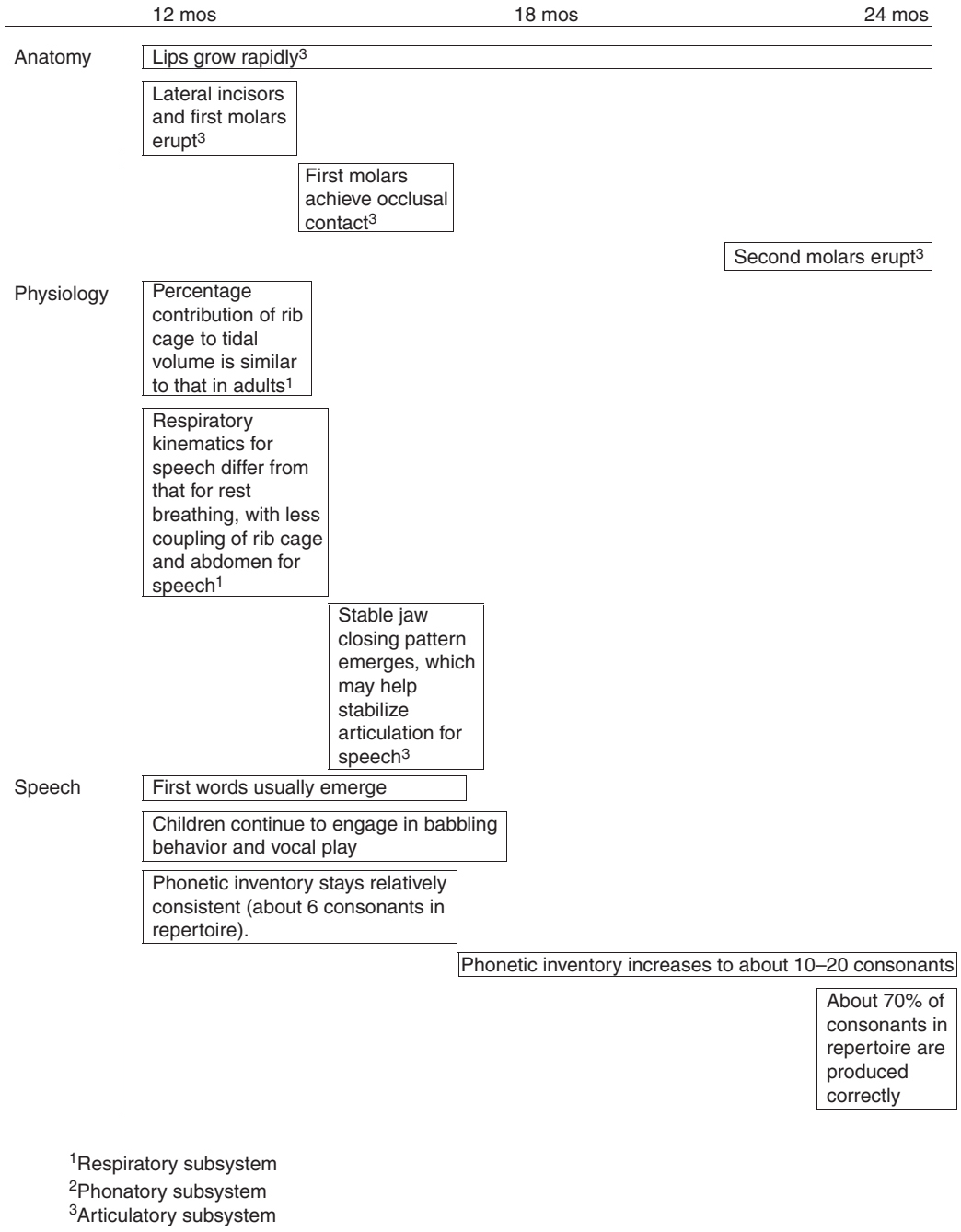

Figure 2 Schematic representation of major speech-related developmental milestones occurring between 12 and 24 months of age. Note that ages and age ranges are approximate and are meant to provide a simplified illustration of the co-occurrence of developmental events.

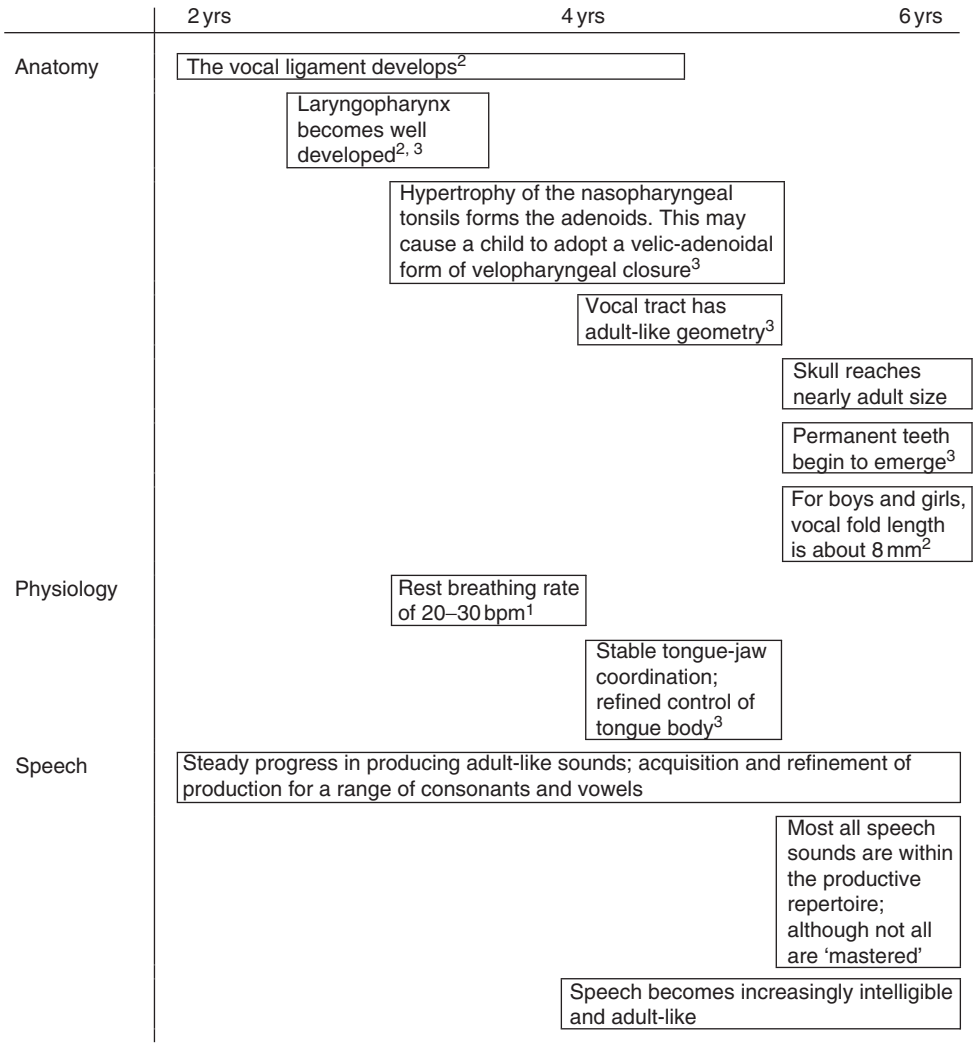

[1]Respiratory subsystem
[2]Phonatory subsystem
[3]Articulatory subsystem

Figure 3 Schematic representation of major speech-related developmental milestones occurring between 2 and 6 years of age. Note that ages and age ranges are approximate and are meant to provide a simplified illustration of the co-occurrence of developmental events.

a significant anatomic remodeling of the speech production system. As mentioned previously, humans start out with a vocal tract like that of nonhuman primates. The human vocal tract emerges developmentally through adjustments that include a descent of the larynx, formation of a right-angled tube in which the oral portion is perpendicular to the pharyngeal portion, and the separation of velum and epiglottis. Some of the changes would seem to contribute importantly to speech maturation, for example, the ability to make reliable oral versus nasal resonance contrasts (e.g., the initial contrast in the words 'buy' vs. 'my'). The anatomic differences are sufficiently great that some writers have asserted that speech production needs at least two theories – one for infants and another for adults. The first year of life typically begins with the birth cry and ends with a child's first words. The intervening vocalizations have been described mostly by perceptual methods, and the resulting data point to both universal and language-specific factors in sound patterns.

Perceptual Findings

Studies of early speech development using perceptual methods are problematic from several perspectives, but especially because of the inevitable tendency for

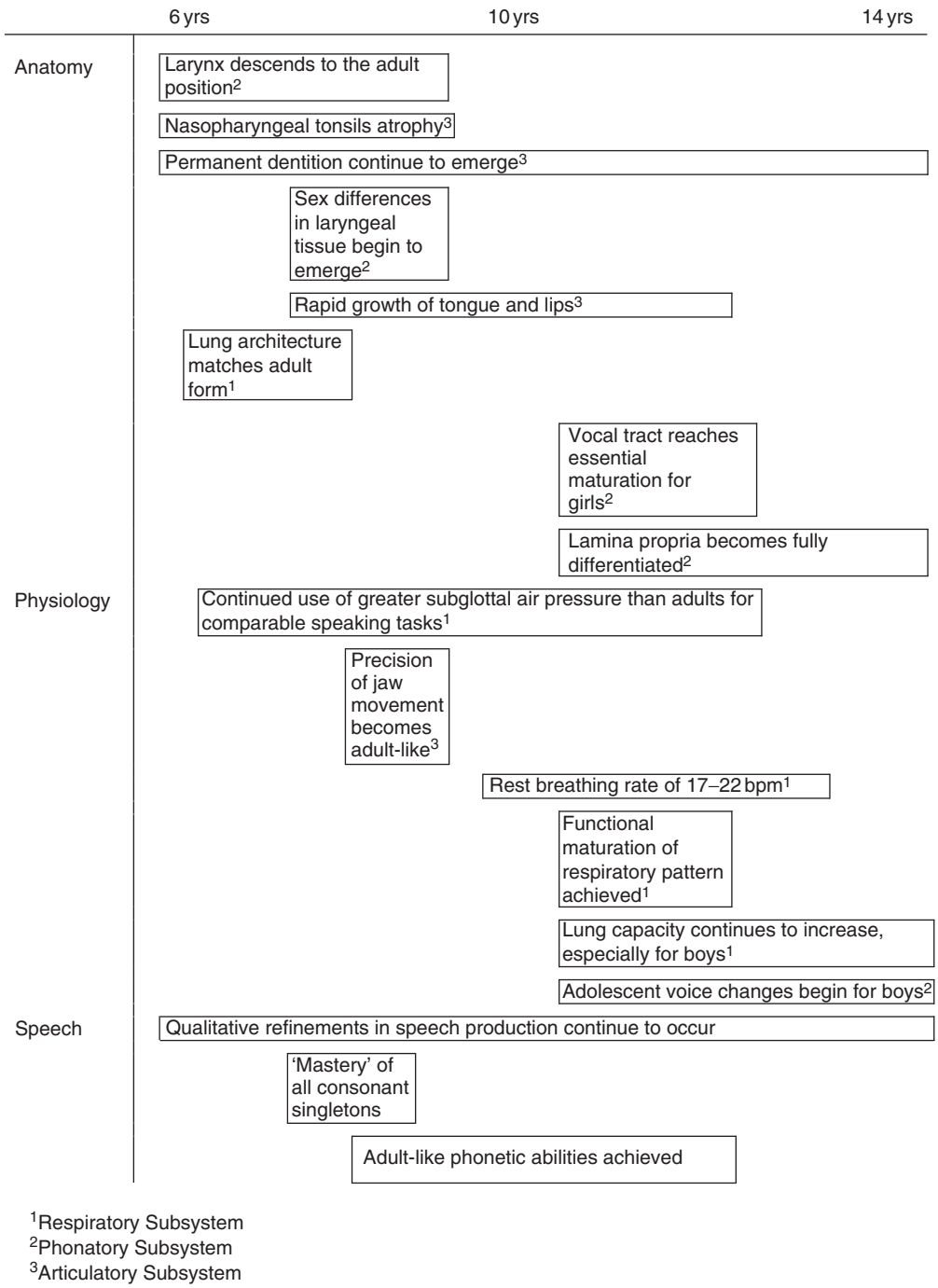

Figure 4 Schematic representation of major speech-related developmental milestones occurring between 6 and 14 years of age. Note that ages and age ranges are approximate and are meant to provide a simplified illustration of the co-occurrence of developmental events.

adult listeners to describe infant vocalizations in terms of their own linguistic experience. Nonetheless, considerable research has focused on describing pre-speech vocal behavior of infants using perceptual approaches, and these data have been fundamental to proposals of stages in vocal development. Although there are some differences among staging taxonomies identified by different researchers, there is substantial overlap in the behaviors described at different ages during the first year of life.

Generally, research suggests that from birth to 2 months, vocalizations consist primarily of crying and vegetative-type sounds with little or no articulation. Between the ages of 2 and 6 months,

infants begin to produce cooing sounds and laughter. In addition, they begin to make some simple articulatory movements during vocalization, including production of nasalized consonants, glottal and velar consonants, trills and raspberries, and vowel-like productions (primarily front and neutral vowels). Vowel and consonant sounds may be produced in vocal play-type contexts in which there is increasing control of phonation (evidenced in pitch variation during sustained vowel-type productions). Also appearing at this time is marginal babbling, a precursor to true babbling, that involves alternating opening and closing of the vocal tract during phonation.

Between 7 and 12 months of age, children begin canonical babbling, or true babbling. Canonical babbling consists of rapid adult-like alternation between consonant and vowel elements that are often reduplicated (e.g., ba-ba-ba). At approximately the same period, children begin to produce variegated babbling, which incorporates production of different consonant + vowel syllables in succession (e.g., ba-di-ga). Importantly, during this period children begin to produce a greater number of consonant-like sounds, with labial and alveolar sounds predominating the repertoire. By the end of this period, most children begin to produce their first words or word approximations, usually consisting of the same types of consonants and vowels that are present in babbling. One of the guidelines recommended for parents and pediatricians is to follow carefully any child who is not babbling, pointing, or using other gestures by 12 months.

The relationship between babbling and early word production is a much discussed topic of speech development. The discontinuity hypothesis (also known as the babbling independence hypothesis) holds that babbling has little or no connection with early speech development. In contrast, the continuity hypothesis (also known as the babbling drift hypothesis) maintains that there is substantial continuity between sound patterns in babbling and those in early word production. Studies favoring the continuity hypothesis have demonstrated the ability of adult listeners to classify babbling by ambient language, shown acoustic similarities between babbling and speech in adults, and described adaptation of infants' phonetic inventories to those in the ambient language. However, it should be noted that some studies have reported that listeners were unable to make a reliable auditory classification of babbling by identifying the ambient (parental) language of the child.

A common view is that biological factors exert a primary influence on phonetic patterns in the first year of life, and perhaps even into the second year, until language factors eventually begin to mold sound patterns. The primacy of biological factors is demonstrated in the fact that infants from diverse language backgrounds tend to produce the same kinds of sounds in babbling. Specifically, there is a predominance of bilabial and alveolar consonants (e.g., sounds resembling those that begin the words 'ball' and 'doll').

Continuity in vocal development does not necessarily mean progressive unidirectionality in developmental processes of speech. In making the transition from babbling to first words, some infants show a labial regression in which labial sounds are preferred in early words even though their babbling may have shown a growing predominance of alveolar sounds. Also, the pattern of intonation development has been described as U-shaped, meaning that the developmental path includes a phase of apparent regression.

At approximately 1 year of age, plus or minus 2 months, most children utter their first words. At approximately the same age, typically slightly earlier, infants also produce phonetically consistent forms (or vocables). These utterances have a sound–meaning relationship and they are produced with a fairly repeatable sound pattern that allows adults to recognize them. They are not based on words in the adult language, but they may signal an infant's readiness to produce words.

Instrumental Findings

Infants in the first year of life have small vocal folds and a short vocal tract, and both of these help to explain major acoustic properties of infant vocalizations. The infantile vocal folds are associated with a relatively high fundamental frequency of vocalization, with a mean value of approximately 400 Hz, compared to 225 Hz for an adult female and 125 Hz for an adult male. Because the vocal tract of an infant is only approximately one-third to one-half as long as that of an adult, it is expected from standard acoustic theory that infants will have much higher resonance frequencies. Indeed, acoustic studies show that vowel-like sounds have much higher formant frequencies than perceptually similar vowels produced by adults. Similarly, the noise spectra in infant fricatives have energy that extends to much higher frequencies than those seen for adults. Therefore, it is physically impossible for infants to produce vocal sounds that have exactly the same physical attributes of adult sounds. These differences prompt the need for normalization algorithms that minimize cross-talker differences. Most of these algorithms are based on logarithms or ratios.

Acoustic data have been reported on several aspects of vocalization during the first year of life, beginning with the birth cry and continuing to

babbling and first words at the end of this period. Acoustic patterns of the birth cry have been shown to have diagnostic value in that certain acoustic features may indicate developmental problems. Canonical babbling appears to be produced with syllables that prefigure adult syllables with respect to their major acoustic characteristics.

Acoustic analysis may be more sensitive than perceptuophonetic studies for some aspects of sound production. For example, acoustic studies have shown that occasionally young children will make differences among sounds that are not discriminated by adults, even those with phonetic training. As another example, language-specific vowel production in babbling has been demonstrated by showing that acoustic properties of vowel production in infants from different language environments mirrored the acoustic properties in the corresponding adults' models. This conclusion indicates that even before infants begin to produce recognizable words, they may have adapted their vocalizations to properties of the ambient language. As a related point, the babbling of deaf infants differs from that of normally hearing infants, especially in the age period of approximately 7–10 months. It appears that infants connect the perceptual and motor aspects of babbling during this period. The auditory–motor associations may be groundwork for later speech development.

Due to the difficulties of obtaining kinematic data from infants, only a few studies of speech movements have been published for this developmental period. It has been reported that breathing behavior varies considerably across four types of vocalization (cries, whimpers, grunts, and syllable utterances) and there is no consistent pattern of breathing by vocalization type. The major conclusion is that infants have numerous degrees of freedom in controlling the respiratory system to produce sounds.

Speech Development in Toddlers (12–24 Months)

Perceptual Findings

Although most typically developing children begin to produce words during this period, they also continue to engage in babbling behavior. Indeed, the phonetic inventory of toddlers stays relatively consistent until approximately 18 months of age. During this time, children produce approximately two or three times more consonants than vowels and have approximately six consonants in their repertoire (usually stops, nasals, and glides). In addition, the inventory of phones produced in the initial position of words tends to be larger than the inventory of phones produced in the final position of words. Phones produced in the initial position of words tend to involve the anterior portion of the mouth and tend to be voiced. Phones produced in the final position of words tend to be alveolar and voiceless.

Between 18 and 24 months, phonetic inventory size increases to approximately 10–20 consonants, and the inventory of consonants used in the syllable initial position increases more rapidly than the inventory for syllable final position. Although individual differences are present in the acquisition of speech sounds, the pattern of emergence of different sounds and sound classes tends to be relatively consistent.

One important question that has not been comprehensively studied relates to intelligibility of children's speech. In other words, how understandable are children at different periods during speech development? Clinical phoneticians and phonologists have tended to use a phone-and/or a phoneme-based approach to addressing this issue. One widely used metric is percentage of consonants correct (PCC). This measure provides information about how many consonants produced by a child match the adult form of the same consonants. However, it does not provide information about the information-bearing capacity of the speech signal produced by the child. Nonetheless, studies suggest that by the age of 2 years, children have an average PCC of approximately 70%. Although children at this age have a reduced repertoire of sounds relative to the adult, they seem to favor words that are within their phonetic repertoire and are on the path to achieving adult-like mastery.

Instrumental Findings

Studies converge on the general conclusion that children of this age have an increasing ability to produce contrasts in intonation and the phonetic properties of speech sounds. Rate of speech is quite slow compared to that in older children and adults, and air pressures and airflows used to produce speech sounds tend to be higher. As observed for infants in the first year of life, breathing behavior is highly variable and there is little consistency by vocalization type. In other words, speech breathing is not closely matched to the properties of utterances.

Speech Development in Preschoolers (2–6 Years)

Perceptual Findings

Children make important progress in their ability to produce a variety of speech sounds that are approaching adult-like between the ages of 2 and 6 years. There are considerable normative data on phonetic

and phonologic development during this period. By approximately 5 or 6 years of age, most children have all the speech sounds in the English language within their productive repertoire, although they likely have not 'mastered' production of all sounds. Studies suggest that among children between the ages of 3 and 11 years, approximately 92.5% develop articulation 'normally,' such that any developmental speech sound errors resolve spontaneously without the need for speech therapy. Although studies are limited, research also suggests that children between the ages of approximately 4 and 5 years produce continuous speech that is highly intelligible. It is during this period that many childhood disorders of communication can be identified. These include developmental stuttering and disorders of articulation and language.

Instrumental Findings

One of the fundamental questions about speech development pertains to the relationship between speech and feeding behaviors, which, after all, use much of the same biology. One line of thought is that movements for speech are derived from those for the ingestive behaviors of chewing and swallowing. This hypothesis is challenged by evidence of task specificity in infants' movements observed in speech, chewing, and lip protrusion. One of the main differences is that production of true words, as opposed to chewing or lip protrusion, is associated with a stronger coupling among all mandibular muscles. The implication of these kinematic studies is that speechlike movements and nonspeech movements are distinct at the earliest stages of word production. However, more research is needed to show to what degree speech and feeding draw from the same sensorimotor capabilities. This is a pivotal issue in understanding speech development, and it holds implications for theory and clinical application.

Vocal fundamental frequency declines between the ages of 12 months and 3 years so that by the age of 6 years, the mean in both males and females is approximately 250 Hz. Acoustic and physiologic studies generally agree in showing that speech production during this period becomes faster and more precise (less variable across repetitions of the same utterance).

Speech Development in Older Children (6+ Years)

Perceptual Findings

During this period, children are primarily making qualitative refinements in their phonetic development. By the age of approximately 6 years, typically developing children have all phonemes in the English language within their productive repertoire and can produce all consonant singletons accurately in more than 51% of occurrences. Children reach 90% 'mastery' of all consonant singletons by the age of approximately 8 years. Consonant clusters continue to undergo refinements toward adult mastery levels through 9 years of age. For all practical purposes, children have attained adult-like phonetic ability before the age of 10 years. If children begin to learn a second language before the age of 10 years, they have a much greater likelihood of attaining native-like proficiency (absence of accent) than if they begin at a later age. This is one kind of evidence that is used to support the critical period hypothesis for language.

Instrumental Findings

As noted previously, by perceptual standards, speech development can be regarded as complete (adult-like) in early childhood. However, instrumental studies lead to a somewhat different conclusion. Acoustic measurements of segment durations and spectral properties indicate that adult-like precision is not reached until the age of approximately 12 years. Physiological studies show that motor refinements for speech continue even longer – until approximately 16 years of age in both boys and girls. This protracted development of speech matches the development of language, which also has a long period of maturation.

During this period, boys and girls begin to differ slightly in the acoustic characteristics of speech. Sex-related differences emerge for fundamental frequency at 5–8 years and for vowel formant frequencies (i.e., resonance frequencies of the vocal tract) at approximately 8–10 years. However, the most conspicuous differences between boys and girls appear during puberty. Between the ages of 12 and 15 years, boys' vocal fundamental frequency drops markedly (approximately 80%). This voice change, or mutation, is related to a substantial growth of the larynx and its associated tissues.

Conclusions

When speech development is analyzed with respect to its component processes and systems, it is seen to be both complicated and protracted. The biology of speech encompasses variegated tissues and organs of the respiratory, laryngeal, and articulatory systems. These components have different schedules of growth and maturation, and putting them into a coherent framework is a basic challenge of research. Moreover, the control of these systems is linked to the requirements of language, and it appears that

maturation of speech as a motor behavior is in some ways tied to the maturation of language. However, some of the most dramatic advances are seen in infancy and early childhood as the basic capability for speech emerges from earlier coos and babbles. Cross-linguistic studies show that babbling has universal properties, and these are taken to reflect biological predispositions. Eventually, infant vocalizations show adaptations to the ambient language, with the first of these appearing at approximately 6–9 months. Phonetic mastery is achieved by the age of 8 years, but continuing refinement of speech motor control is evident until the age of approximately 16 years.

See also: Evolution of Vertebrate Respiratory Control; Language Development; Motor Sequences; Motor Timing; Orofacial Motor Control; Sentence Production; Speech Production: Adult; Trigeminal Motor System; Word Production.

Further Reading

Boysson-Bardies B de (1999) *How Language Comes to Children*. Cambridge, MA: MIT Press.

Fletcher P and MacWhinney B (eds.) (1995) *The Handbook of Child Language*. Oxford: Blackwell.

Jusczyk PW (1997) *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

Kent RD (2004) Models of speech motor control: Implications from recent developments in neurophysiological and neurobehavioral science. In: Maassen B, Kent RD, Peters HFM, van Lieshout PHHM, and Hulstijn W (eds.) *Speech Motor Control in Normal and Disordered Speech*, pp. 3–28. Oxford: Oxford University Press.

Kent RD and Vorperian HK (1995) Anatomic development of the craniofacial–oral–laryngeal systems: A review. *Journal of Medical Speech–Language Pathology* 3: 145–190.

Locke JL (1993) *The Child's Path to Spoken Language*. Cambridge, MA: Harvard University Press.

Oller DK (2000) *The Emergence of the Speech Capacity*. Mahwah, NJ: Erlbaum.

Smith A and Goffman L (2004) Interaction of motor and language factors in the development of speech production. In: Maassen B, Kent RD, Peters HFM, van Lieshout PHHM, and Hulstijn W (eds.) *Speech Motor Control in Normal and Disordered Speech*, pp. 225–252. Oxford: Oxford University Press.

Smith A, Goffman L, and Stark RE (1995) Speech motor development. *Seminars in Speech and Language* 16: 87–99.

Strand EA (1992) The integration of speech motor control and language formulation in process models of acquisition. In: Chapman RS (ed.) *Processes in Language Acquisition and Disorders*, pp. 86–107. St. Louis: Mosby.

Vihman MM (1996) *Phonological Development: The Origins of Language in the Child*. Cambridge, MA: Blackwell.

Relevant Website

http://www.asha.org – American Speech–Language–Hearing Association.

Spike-Timing-Dependent Plasticity (STDP)

D V Buonomano and T P Carvalho, University of California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Spike-timing-dependent plasticity (STDP) refers to a form of associative synaptic plasticity in which the temporal order of the presynaptic and postsynaptic action potentials determines the direction of plasticity, that is, whether synaptic depression or potentiation is induced. In the most common form of STDP, long-term potentiation (LTP) is induced if the presynaptic spike precedes the postsynaptic spike (pre→post). In contrast, if the postsynaptic spike happens before the presynaptic spike (post→pre), then long-term depression (LTD) is induced. In addition to the order of the pre- and postsynaptic spike, STDP is sensitive to the interspike interval (ISI), the time elapsed between the two spikes. In general, short intervals produce maximal plasticity, while longer intervals produce little or no change in synaptic strength.

STDP represents an important form of synaptic plasticity for both experimental and computational reasons. Experimentally, LTP and LTD have often been induced by distinct protocols, and the relationship between these protocols as well as their physiological relevance has not always been clear. STDP formalizes the conditions for the induction of both LTP and LTD into a single learning rule. STDP is attractive as a computational learning rule because the direction of synaptic plasticity reflects the causal relationship between pre- and postsynaptic activity. If the presynaptic spike consistently precedes a postsynaptic spike, it is likely that the former contributes in eliciting the latter, and this contribution will be 'reinforced' through potentiation. In contrast, if the presynaptic spike consistently fires after the postsynaptic spike, no causal relationship between pre- and postsynaptic activity is present, and depression will ensue.

STDP and Hebbian Plasticity

The psychologist Donald Hebb postulated that if a presynaptic neuron "repeatedly or persistently takes part in firing" a postsynaptic neuron the synapse between them would be potentiated. This form of plasticity has come to be known as Hebbian or associative synaptic plasticity. However, Hebb's original postulate left two important issues unaddressed. First, what is the learning rule that underlies 'decreases' in synaptic strength? Second, what is the effective time window between pre- and postsynaptic activity that will result in potentiation? STDP addresses both these issues by establishing the specific conditions and temporal requirements for the induction of both LTP and LTD.

In a typical protocol, a single presynaptic spike is repetitively (e.g., every 1 s for 1 min) elicited 0–25 ms before or after a postsynaptic action potential. Here, intervals consisting of pre→post pairings will be defined as 'positive' ISIs, while post→pre pairings will be referred to as 'negative' ISIs. By examining the direction and magnitude of plasticity over a range of ISIs, we can plot the so-called 'STDP function' (change in synaptic strength vs. ISI), schematized in Figure 1. The STDP function characterizes plasticity at a given synapse, and represents the 'time windows' for the induction of both LTP (green) and LTD (red). Because both LTP and LTD are maximal for short intervals, differences of only a few milliseconds result in a sharp discontinuity at ~0 ms. In addition to the amplitude and duration of both time windows, another important feature of the STDP function is their ratio. For example, under the simplified assumption that pre- and postsynaptic spikes occur in isolation and randomly at all possible ISIs, one can see that if the LTP and LTD windows are equal in area there will be no net change in the overall synaptic strength. However, if the LTD area is larger than its LTP counterpart, random pre- and postsynaptic events will lead to a decrease in synaptic strength toward zero. Conversely, synaptic strength will saturate if the LTP area is larger. While not all the experimental data are in agreement, it is clear that in some cases the STDP functions do exhibit larger LTD windows, and thus are consistent with a potential role of STDP in synaptic competition (see below).

STDP functions similar to the one schematized in Figure 1 have been characterized in a number of different preparations including excitatory neocortical synapses, the retinotectal projection in *Xenopus*, and dissociated hippocampal cell cultures. Nevertheless, STDP is not universally observed, and thus the following points should be noted.

1. Not all synapses seem to undergo STDP in response to the pairing of single pre- and postsynaptic action potentials. For example, most evidence suggests that pairing of single pre- and postsynaptic spikes in the Schaffer-collateral→CA1 synapse in hippocampal slices does not result in synaptic changes that follow a function similar to the one depicted in Figure 1. At these synapses, the induction of LTP seems to require a burst of postsynaptic spikes.

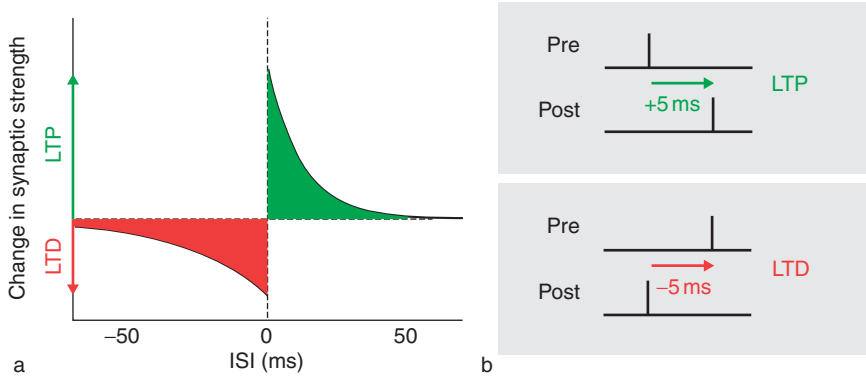

Figure 1 (a) A hypothetical STDP function is depicted. Positive or negative interstimulus intervals give rise to LTP or LTD, respectively. No appreciable plasticity is induced at the synapse for large ISIs. Note the sharp discontinuity at 0, and that a longer-lasting LTD window is depicted. (b) A positive ISI is defined as the time interval between an initial presynaptic spike and a consecutive postsynaptic spike, which will lead to LTP. For negative ISIs, the postsynaptic cell fires before the presynaptic, and LTD will ensue.

2. The shape of the STDP function is highly variable. The peak or total area of the LTP and LTD windows are dependent not only on the types of synapses being examined, but on the exact induction protocol, pre- and postsynaptic cell type, and even on the position of the synapse along the dendritic arbor.
3. The standard STDP function is based on isolated single pre- and postsynaptic spikes, and does not necessarily describe synaptic plasticity in response to more complex patterns of activity.

Linearity of STDP

What would happen if the presynaptic spike consistently preceded and followed the postsynaptic spike by ISIs that individually would contribute with the same, but opposite, amount of LTP and LTD? If the result of such an experiment yielded no change in the synaptic strength, it would be said that individual ISIs were independent and that the contributions of each of the intervals summate linearly. In contrast, if, for instance, LTP was observed, STDP would be nonlinear and LTP would be dominant. Under experimental conditions, it is possible to constrain the spike patterns to single isolated spikes; however, *in vivo* data show that bursts of spikes or complex ongoing temporal patterns of spikes are commonly elicited in response to a given stimuli. Thus, in order to understand the computational role of STDP *in vivo*, it is necessary to address the issue of linearity or nonlinearity of STDP, and define the rules governing plasticity in response to complex patterns of pre- and postsynaptic activity.

Experimental studies have examined the issue of linearity by using protocols with spike triplets (e.g., pre→post→pre or post→pre→post) or quadruplets. It is clear that STDP is not strictly linear; however, the precise nature of the interaction of complex spike patterns on the LTP and LTD components of STDP is still emerging. Empirically determining the higher-order features of STDP is a challenging task, because the potential arrangements of spike patterns undergo a combinatorial explosion. Thus, a detailed understanding of the higher-order properties of STDP will likely rely on the elucidation of the underlying mechanisms. Specifically, if the critical cellular, synaptic, and biochemical mechanisms governing STDP can be quantitatively characterized, it should be possible to accurately predict the direction and magnitude of synaptic plasticity in response to any arbitrary patterns of pre- and postsynaptic spikes.

Mechanisms

Over the past decades, considerable emphasis has been placed on elucidating the mechanisms underlying LTP and LTD; similar efforts are currently focused on understanding the mechanisms underlying STDP. Two critical features of STDP must be accounted for at the mechanistic level:

1. *Order*. How can a synapse detect differences of a few milliseconds regarding the temporal order of presynaptic and postsynaptic spikes, and account for the sharp discontinuity observed at approximately 0 ms ISIs?
2. *Interval*. Independently of the issue of temporal order, what are the mechanisms that allow synapses to be sensitive to the pre→post and post→pre intervals? Specifically, why does a long positive (or negative) ISI produce little LTP (or LTD), while a short interval produce stronger LTP (or LTD)?

The mechanisms responsible for the interval sensitivity of the LTP component of STDP are generally accepted as being the same as those underlying associative LTP. Specifically, the *N*-methyl-D-aspartate (NMDA) receptors function as coincidence detectors of pre- and postsynaptic activity. A presynaptic spike

results in the release of glutamate, which binds to the NMDA receptor. Subsequent depolarization caused by a postsynaptic spike produces a voltage-dependent extrusion of the Mg^{2+} ion that impedes Ca^{2+} influx at rest. Only when both of these events occur in close temporal proximity will the NMDA channels allow the critical influx of Ca^{2+}. Since glutamate can remain bound to the NMDA receptor for at least tens of milliseconds, if the postsynaptic spike occurs shortly after the presynaptic event, glutamate will still be bound and Ca^{2+} influx will ensue. Thus, it appears that the binding kinetics of the NMDA receptor can account for the timing of the LTP window of STDP. Specifically, at large positive ISIs, little or no glutamate will remain bound to the postsynaptic NMDA receptor resulting in little or no LTP; however, at short intervals, maximal glutamate will be bound, and high levels of Ca^{2+} influx will be achieved, thus inducing maximal LTP. Consistently, there is robust experimental data showing that the blockade of NMDA receptors blocks the LTP component of STDP. For these reasons, the properties of the NMDA receptor likely underlie both the order and interval sensitivity components of STDP potentiation.

It is not immediately clear how the NMDA receptors could account for both the interval and order sensitivity of the LTD window of STDP. If the postsynaptic neuron fires before the presynaptic one, the membrane of the postsynaptic cell should have returned to close to its resting potential before glutamate is released from the presynaptic terminal (particularly for long negative ISIs); thus, the released glutamate should be ineffective in leading to Ca^{2+} influx. It could be argued that a small degree of Ca^{2+} influx occurs in response to a presynaptic action potential even when it occurs after a postsynaptic spike, and that this is sufficient for the induction of LTD. This argument, however, leads to a scenario in which presynaptic stimulation by itself could induce LTD (which is generally not the case).

Conflicting results have been reported regarding the ability of NMDA antagonists to block the LTD component of STDP; however, at least in some preparations, it seems clear that postsynaptic NMDA receptors are not involved in the induction of LTD in response to post→pre pairings. If NMDA receptors are not necessary for detecting the order and interval of negative ISIs, a second coincidence detector would be required, that is, the presence of a postsynaptic followed by a presynaptic spike must be detected by another biochemical pathway that triggers the induction of LTD. This two-coincidence detector model of STDP assumes the presence of a second biochemical locus that is capable of detecting the order and interval of post→pre spikes. One candidate mechanism for the second coincidence detector is the metabotropic glutamate receptor pathway. In this model, post→pre spiking leads to a low-level Ca^{2+} rise entering through voltage-dependent Ca^{2+} channels (VDCCs) that would prime the activation of the G-protein-mediated cascade, which would in turn be triggered by the presynaptic action potential via metabotropic glutamate receptors.

Computational Relevance and Correlates of STDP *In Vivo*

Competition

In general, computational models of cortical plasticity require two critical components: (1) an associative form of synaptic plasticity that allows correlated inputs to be strengthened and (2) competition, which is required to prevent neurons from responding to multiple sets of correlated patterns (stimuli). In contrast to other models, it has been proposed that STDP may be able to satisfy both these conditions with a single learning rule.

In its simplest sense, in STDP, competition arises between synapses that fire before and after a postsynaptic spike. Synapses from presynaptic neurons that fire before the postsynaptic action potential will be strengthened and take control of the postsynaptic behavior, while the synapses from neurons that fire later will be weakened. A competitive role for STDP predicts that the area of the LTD window should be larger than that of LTP. As mentioned above, this insures that in response to random background activity the net synaptic change is negative. While STDP can clearly implement competition between synapses, it should be stressed that the competition is dependent on the temporal relationship between the stimuli, that is, for different sets of correlated inputs to compete they must occur in close temporal proximity. If the two stimuli are separated by an intervals well outside the STDP window, there can be no interaction (and thus no competition) between stimuli over these time frames (note that LTD produced by spontaneous activity would be common to both stimuli). Thus, while STDP does provide a potential mechanism for competition, it does not by itself provide a universal one.

Computational Relevance and Neural Correlates of STDP *In Vivo*

In vitro STDP is generally studied by directly eliciting action potentials in the pre- and postsynaptic neurons. However, a number of studies have examined the putative *in vivo* correlates of STDP in the context of external stimuli. The order sensitivity of STDP

predicts that if two stimuli, A and B, are reliably presented in the order A→B and in close temporal proximity, a receptive field asymmetry should develop. Specifically, responses to A should increase and those to B decrease. Indeed, several experimental studies have measured stimulus-induced shifts in receptive fields in response to exposure to ordered stimuli. These include experiments in which the stimuli A and B corresponded to orientation of bars of light, visual position, or spatial receptive (i.e., the receptive fields of hippocampal place cells). Taken together, these results support the notion that STDP operates *in vivo*, and plays an important role in shaping neural responses.

While STDP likely contributes to the order-sensitive shifts in receptive field properties, it does not generate order selectivity. In the above example, consistent pairing of stimuli A→B results in strengthening of synapses driven by stimulus A and weakening of those from B, but nevertheless the neuronal response to A→B should be approximately equal to the response to B→A. Thus, the order sensitivity of STDP does not by itself generate order-selective neuronal responses.

Experience-dependent changes in the receptive fields of cortical neurons have also been interpreted in the context of STDP. For example, plasticity can be induced by cutting all but one whisker of the mystacial pad of a rat. In the barrel of a cut whisker, two changes are observed: a decreased response to the normally dominant whisker and an increased response to the spared whisker. It has been proposed that STDP can contribute to this phenomenon. Specifically, the decreased response of layer II/III pyramidal neurons to the previously dominant whisker appears to be due to LTD of synapses from layer IV. It has been shown that this LTD is consistent with STDP and the order of activation of these neurons. Specifically, in the deprived barrel, L-II/III pyramidal neurons may fire before the L-IV neurons, resulting in LTD.

Timing

STDP is highly sensitive to the interval between the pre- and postsynaptic spikes; however, the role of STDP in generating timed responses remains unclear. If a presynaptic spike consistently occurs 25 ms before the postsynaptic spike, the synaptic strength between the two neurons will increase. Assuming that this potentiation can influence cell firing, the postsynaptic neuron will spike at progressively shorter latencies, until the synaptic strength saturates or the interval decreases toward the monosynaptic delay latency. Thus, the original temporal interval used in the training is not effectively reproduced. As a result, STDP maintains the order of activation and favors early responses, but the patterns produced after STDP do not capture the temporal structure of the training stimuli. Nevertheless, STDP could take part in setting up specific spatiotemporal patterns of network activity that could play a role in temporal processing.

See also: Adult Cortical Plasticity; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Hebbian Plasticity; Long-Term Depression: Cerebellum; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Metabotropic Glutamate Receptors (mGluRs): Functions; Spike-Timing-Dependent Plasticity Models; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Abbott LF and Nelson SB (2000) Synaptic plasticity: Taming the beast. *Nature Neuroscience* 3: 1178–1183.

Bender VA, Bender KJ, Brasier DJ, and Feldman DE (2006) Two coincidence detectors for spike timing-dependent plasticity in somatosensory cortex. *Journal of Neuroscience* 26: 4166–4177.

Bi G-Q and Poo M-M (1998) Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience* 18: 10464–10472.

Bi G-Q and Poo M-M (2000) Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience* 24: 139–166 (Review).

Dan Y and Poo MM (2004) Spike timing-dependent plasticity of neural circuits. *Neuron* 44: 23–30.

Debanne D, Gahwiler BH, and Thompson SM (1994) Asynchronous presynaptic and postsynaptic activity induces associative long-term depression in area CA1 of the rat hippocampus *in vitro*. *Proceedings of the National Academy of Sciences of the United States of America* 91: 1148–1152.

Feldman DE (2000) Timing-based LTP and LTD at vertical inputs to layer II/III pyramidal cells in rat barrel cortex. *Neuron* 27: 45–56.

Karmarkar UR and Buonomano DV (2002) A model of spike-timing dependent plasticity: One or two coincidence detectors? *Journal of Neurophysiology* 88: 507–513.

Levy WB and Steward O (1983) Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience* 8: 791–797.

Markram H, Lubke J, Frotscher M, and Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275: 213–215.

Mehta MR, Quirk MC, and Wilson MA (2000) Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron* 25: 707–715.

Song S, Miller KD, and Abbott LF (2000) Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience* 9: 919–926.

Yao H and Dan Y (2001) Stimulus timing-dependent plasticity in cortical processing of orientation. *Neuron* 2: 315–323.

Zhang LI, Huizhong WT, Holt CE, and Poo M-M (1998) A critical window for cooperation and competition among developing retinotectal synapses. *Nature* 395: 37–44.

Spike-Timing-Dependent Plasticity Models

J L van Hemmen, Technical University of Munich, Garching/Munich, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Activity in a neuronal network is spatiotemporal. That is, it changes in space all over the network and in time for each specific neuron separately. How, then, can a network learn such detailed activity? Furthermore, what precisely does learning mean? In this article we focus on spiking neurons so that spatiotemporal activity for a network of size N means a collection of spike times t_i^f, where f indicates the spikes of neuron i and $1 \leq i \leq N$ labels the neurons. Learning then means that, as a result of and starting with a specific set of initial states, the network repeats the same spatiotemporal activity pattern as time proceeds. Of course, we implicitly assume *ceteris paribus* (the other things and thus all inputs from outside being equal).

To understand learning, two fundamental principles are important. First, where does learning take place? The kind of learning we are studying here depends on the arrival of spikes and is strictly local, meaning that events that merely occur at or near a synapse where learning happens are important to the change of the synaptic strength (or efficacy). If this depends only on pre- and postsynaptic spikes, it is called Hebbian.

Second, learning is, in general, slow with respect to the neuronal timescale (milliseconds); for instance, learning associated with sensory modalities takes days, weeks, or even longer. That is (to express it sloppily), most learning relies on repetition: practice makes perfect. The organism repeats the very same or similar patterns or sequences of patterns several times and, in so doing, trains the synapses. Hence, it is bound to be slow. Furthermore, for sensory systems this suggests that interesting phenomena in map formation also happen slowly. We define a map to be a neuronal representation of the outside sensory world. If learning is slow, then long-time correlations are picked up but short-time ones, such as white or light-colored noise, are averaged out.

As for locality of learning, Donald Hebb's classic *The Organization of Behavior* (1949, p. 62), contains the now famous neurophysiological postulate: "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased". (Hebb, 1949:62)

We are then left wondering where the 'metabolic change' might take place, but Hebb already postulated explicitly that all this happens in synapses connecting the pre- and postsynaptic neurons. So, we focus here on pre- and postsynaptic spikes.

Learning a Spatiotemporal Pattern as Simply as Possible

To make our argument as clear as possible, we simplify it for the moment as much as we can and discretize time by units $\Delta t = 1$ ms, with the idea that a spike generically lasts 1 ms. Furthermore, we also discretize the states of our neurons $i, 1 \leq i \leq N$, through a state variable S_i, assuming two values: $S_i = 1$ if a neuron fires a spike at a certain time and $S_i = -1$ if it does not (the negative sign signaling that nothing happens). This type of neuron is a classic one, described in 1943 by McCulloch and Pitts, and equipped with universal computational capabilities.

To remember a pattern to come, given the network state at time t, we need the network's past – the more of it the better – so as to predict the future state at time $t + \Delta t$. How, then, can the network look backward in time to sample these data? That is both easy and natural – through delays. A neuronal network has lots of delays, most prominently the axonal ones. This is one of the key ideas behind the argument presented here. Interestingly, this idea is historical as well, and we therefore follow this line of thought for a while.

We focus on an excitatory synapse connecting the presynaptic neuron j to the postsynaptic neuron i. Its strength is denoted by J_{ij}. The axonal delay it takes a spike fired by j to arrive at a synapse on neuron i is $\Delta_{ij}^{\mathrm{ax}}$. Imprinting on the network a sequence of states $S(t) = \{S_i(t), 1 \leq i \leq N\}$ during a period $1 \leq t \leq T_l$, where T_l will soon be specified, our synapse should contribute to the learning process in an appropriate way. We can now state without further ado (and then verify that this makes sense) that the simplest possible way for a synaptic change ΔJ_{ij} is given by

$$\Delta J_{ij} = \eta_{ij} \frac{1}{T_l} \sum_{t=1}^{T_l} S_i(t + \Delta t)[S_j(t - \Delta_{ij}^{\mathrm{ax}}) + 1] \qquad [1]$$

This expression is to be added to the already existing synaptic efficacy J_{ij} of a synapse connecting neuron j to neuron i.

What does eqn [1] mean? At time t, a spike arrives at a synapse on neuron i, if at time $t - \Delta_{ij}^{\mathrm{ax}}$ neuron

j has fired. If it has not, then $[S_j(t\text{-}\Delta_{ij}^{ax})+1]=0$, in agreement with the fact that neuronal networks are characterized by low activity so that what is not active should not count. Only if a spike arrives does the synapse change by $\pm 2\eta_{ij}$ according to the postsynaptic activity $S_i(t+\Delta t)$ at the time $t+\Delta t$ following the arrival of the presynaptic spike at time t. Because learning is slow, the factor η_{ij} is to be small, meaning in suitable units $0<\eta_{ij}\ll 1$.

Now comes the key point. We focus on an excitatory synapse where a spike arrives at time t. The argument of the sum in eqn [1] tells us that the synapse is to be strengthened by $2\eta_{ij}$ if neuron i fires at time $t+\Delta t$, whereas it is to be weakened by $-2\eta_{ij}$ if neuron i does not fire. That is, if as an excitatory synapse it does its job, telling i to fire and thus causing the postsynaptic neuron to fire at time $t+\Delta t$, then the synapse is to be strengthened. If, on the other hand, at time $t+\Delta t$ the postsynaptic neuron does not fire, then the synapse does not do its job properly and it is to be weakened. "Those that come too late shall be punished." As for the rationale, it all fits.

Furthermore, we average over a time T_l, meaning we sum up, divide by T_l, and are done. How big or small, then, should T_l be? It is here that the adiabatic hypothesis $0<\eta_{ij}\ll 1$ works to advantage – to great advantage, even. Because in real life learning is slow, which is equivalent to saying $0<\eta_{ij}\ll 1$, we can choose, on the one hand, a T_l much bigger than all neuronal timescales (which are of the order of milliseconds) and, on the other hand, we can choose a T_l much smaller than the time needed to learn a specific task (which takes days or weeks). A gerbil needs 2 days and a barn owl needs 2 weeks to learn azimuthal sound localization – and for most humans learning to playing a piece by Chopin on the piano in 2 weeks is an optimistic estimate. Hence, we can imagine T_l to separate microscopic neuronal timescales (in milliseconds) from macroscopic learning times (in days and longer).

Finally, why does eqn [1] work as an algorithm to store a spatiotemporal pattern? **Figure 1** shows that it does work, but why does it? The key to understanding this is that the synapses use the very same pattern in conjunction with the very same wetware, that is, with the very same neuronal hardware, such as axons and synapses, and hence with the very same axonal delays Δ_{ij}^{ax}, both for learning and later on for replay. A closer look at **Figure 1** readily reveals that the total pattern consists of the disjoint union of parts occurring at subsequent times. Both eqn [1] and the upshot of the physiological learning process are the sum of subsequent parts. If these, or different subsequent patterns, are strongly correlated – roughly speaking, strongly resembling one another – there is a correlation problem needing an outside solution that is also outside the present context (e.g., through unlearning).

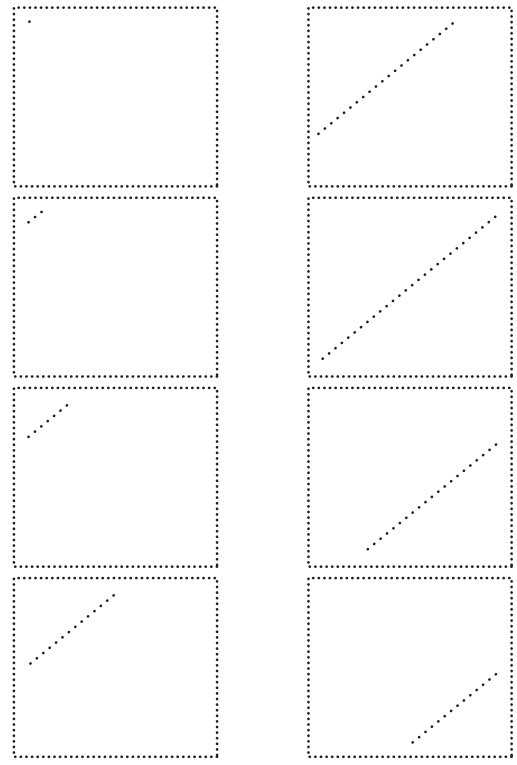

Figure 1 Motion of a 'phase boundary', a string of active neurons (black pixels), through a 20 × 20 storage layer. The system starts with a single point in the upper left-hand corner, and the string develops as time proceeds (top to bottom; first left, then right). During the motion, the number of black pixels varies between 1 and 20. From van Hemmen JL, Gerstner W, Herz AVM, Kühn R, and Vaas M (1990) Encoding and decoding of patterns which are correlated in space and time. In: Dorffner G (ed.) *Konnektionismus in Artificial Intelligence and Kognitionsforschung*, pp. 153–162. Berlin: Springer.

Spike-Timing-Dependent Plasticity

The inputs provided by many synapses decide what a neuron does, but, once it has fired, a local synaptic algorithm determines whether each of the synaptic efficacies increases or decreases. Also, without time discretization and when a synapse can remember more than the very last time step (i.e., over a period of time lasting longer than the time span Δt), there is a suitable generalization – the synaptic learning window, to which we now turn. This is a generalization of what we see in eqn [1], called (for good reason, as we will see) spike-timing-dependent plasticity (STDP). Each of the terms in eqn [2], below, has a

neurobiological origin. We designate the process they describe infinitesimal learning because the synaptic increments and decrements are small and, consequently, it takes quite a while before the organism builds up a noticeable effect.

For the sake of definiteness, we now study the waxing and waning of synaptic strengths associated with a single neuron, which therefore need not carry a label (**Figure 2**). The $1 \le i \le N$ synapses stemming from the neurons that we originally labeled j but now relabel i provide their input at times t_i^f. The firing times of the postsynaptic neuron are denoted by t^n, where n is a label like f. Given the firing times, the change $\Delta J_i(t) := J_i(t) - J_i(t - T_l)$ of the efficacy of synapse i (synaptic strength) during a learning session of duration T_l and ending at time t is governed by several factors:

$$\Delta J_i(t) = \eta \left[\sum_{t-T_l \le t_i^f < t} w^{\text{in}} + \sum_{t-T_l \le t^n < t} w^{\text{out}} + \sum_{t-T_l \le t_i^f, t^n < t} W(t_i^f - t^n) \right] \quad [2]$$

which exhibit STDP in the literal sense of the words. Here the firing times t^n of the postsynaptic neuron may, and in general will, depend on J_i; more precisely, they depend on all the J_i with $1 \le i \le N$.

We now focus on the individual terms of eqn [2] and treat them in turn. The prefactor $0 < \eta \ll 1$ reminds us explicitly that learning is slow on a neuronal time scale. Throughout what follows, we refer to this condition as the 'adiabatic hypothesis'. This holds in numerous biological situations and has been a mainstay of computational neuroscience ever since it was proposed. It may also play a beneficial role in an applied context. If the adiabatic hypothesis does not hold and η is not small, a numerical implementation of the learning rule (eqn [2]) is straightforward but an analytical treatment is not. Furthermore, each incoming spike and each action potential of the postsynaptic neuron change the synaptic efficacy in eqn [2] by ηw^{in} and ηw^{out}, respectively.

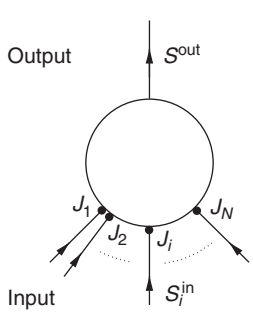

Figure 2 Single neuron. We study the development of synaptic weights J_i (small filled circles, $1 \le i \le N$) of a single neuron (large circle). The neuron receives input spike trains, denoted by S_i^{in}, and produces output spikes denoted by S^{out}.

The last term in eqn [2] represents the 'learning window' $W(s)$, which indicates the synaptic change in dependence on the time difference $s = t^f - t^n$ between an incoming spike t_i^f and an outgoing spike t^n. When the former precedes the latter, we have $s < 0 \Leftrightarrow t_i^f < t^n$, and the result is $W(s) > 0$, implying potentiation. This seems reasonable because N-methyl-D-aspartate (NMDA) receptors, which are important for long-term potentiation (LTP), need a strongly positive membrane voltage to become accessible by losing the Mg^{2+} ions that block their gate. A postsynaptic action potential induces a fast retrograde spike doing exactly this. Because the presynaptic spike arrived slightly earlier, the neurotransmitter is waiting to gain access, which is allowed after the Mg^{2+} ions are gone. The result is Ca^{2+} influx. On the other hand, if the incoming spike comes too late, then $s > 0$ and $W(s) < 0$, implying depression – which is in agreement with the rule "Those that come too late shall be punished". In neurobiological terms, there is no neurotransmitter waiting to be admitted. The learning rule (eqn [2]) is a direct extension of eqn [1], its time-discrete predecessor. There is, meanwhile, extensive neurobiological evidence in favor of this time-resolved Hebbian learning. An illustration of what a learning window does is given in **Figure 3** below.

Once new (infinitesimal) learning algorithms have been discovered, we can simply adapt W accordingly. For instance, for inhibitory synapses infinitesimal growth processes have been found that can be described qualitatively by putting $W := -W$ in **Figure 4**, which figure and eqn [3] show a typical learning window for an excitatory synapse:

$$W(s) = \eta \begin{cases} \exp(s/\tau^{\text{syn}})[A_+(1 - s/\tilde{\tau}_+) + A_-(1 - s/\tilde{\tau}_-)] & \text{for } s \le 0 \\ A_+ \exp(-s/\tau_+) + A_- \exp(-s/\tau_-) & \text{for } s > 0 \end{cases} \quad [3]$$

Here, as before, $s = t_i^f - t^n$ is the time difference between the arrival of the presynaptic spike and the postsynaptic firing, η is our small learning parameter, $\tilde{\tau}_+ := \tau^{\text{syn}}\tau_+/(\tau^{\text{syn}} + \tau_+)$, and $\tilde{\tau}_- := \tau^{\text{syn}}\tau_-/(\tau^{\text{syn}} + \tau_-)$. Typical parameter values used in numerical simulations are $\eta = 10^{-5}$, $A_+ = 1$, $A_- = -1$, $\tau^{\text{syn}} = 5$ ms, $\tau_+ = 1$ ms, and $\tau_- = 20$ ms.

The philosophy behind eqn [3] and **Figure 4** is straightforward. Neuronal spike generation happens at the axon hillock, where synaptic input is perceived through a postsynaptic potential of finite temporal width. Once a presynaptic spike has arrived at an excitatory synapse at time s, when $W(s)$ is

approximately maximal, then the synapse's influence is felt optimally at time $s = 0$, when the postsynaptic neuron fires. This makes the finite width of W readily understandable.

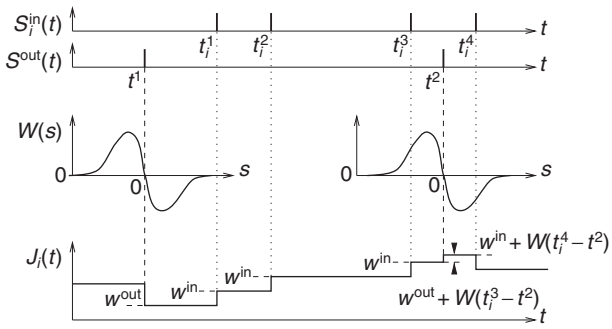

Figure 3 Schematic of Hebbian learning and spiking neurons. Shown in the bottom graph is the time course of the synaptic weight $J_i(t)$, evoked through input and output spikes (upper graphs, vertical bars). An output spike (e.g., at time t^1) induces the weight J_i to change by an amount w^{out}, which is negative here. To show the effect of correlations between input and output spikes, the learning window $W(s)$ (middle graphs) has been indicated around each output spike; $s = 0$ matches the output spike times (vertical dashed lines). The three input spikes at times $t_i^f = t_i^1, t_i^2$, and t_i^3 (vertical dotted lines) increase J_i by an amount w^{in} each. There are no correlations between these input spikes and the output spike at time t^1. This becomes clear once we look at them through the learning window W, centered at t^1; the input spikes are too far away in time. The next output spike at t^2, however, is close enough to the previous input spike at t_i^3. The weight J_i is changed by $w^{out} < 0$ plus the contribution $W(t_i^3 - t^2) > 0$, the sum of which is positive (arrowheads). Similarly, the input spike at time t_i^4 leads to a change $w^{in} + W(t_i^4 - t^2) < 0$. From Kempter R, Gerstner W, and van Hemmen JL (1999) Hebbian learning and spiking neurons. *Physical Review E* 59: 4498–4514.

On the basis of minimal assumptions in conjunction with the adiabatic hypothesis ($0 < \eta \ll 1$), we can now derive a general learning equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} J_i = \eta \left[w^{in} \nu_i^{in} + w^{out} \nu^{out} + \int_{-\infty}^{\infty} \mathrm{d}s\, W(s) C_i(s,t) \right] \quad [4]$$

In this form the learning equation is easy to remember – the input rate ν_i^{in} modifies the synaptic efficacy through w^{in}, the output rate ν^{out} does so through w^{out}, and the Hebbian correlation function C_i favors or disfavors it through the learning window W. In eqn [4] the correlation function C_i is defined through

$$C_i(s,t) := \frac{1}{T_l} \int_{t-T_l}^{t} \mathrm{d}t' \langle S_i^{in}(t'+s) S^{out}(t') \rangle \quad [5]$$

where S_i^{in} is the input spike train (a flow of delta functions), S^{out} is its analog for the output spike train (cf. Figure 2), and the angle brackets denote an average over stochasticity in the neuronal dynamics with short-range correlations of the order of milliseconds.

We now turn to two applications of STDP that show clearly why and how conjugate pre- and postsynaptic timing can together tune a collection of synapses on a neuron such as that in Figure 2.

Barn Owl's Laminar Nucleus

Our first example illustrates how STDP sampling singles out the right synapses. It is, in fact, what STDP had been made for. Azimuthal sound localization is based on phase locking, a correlated neuronal

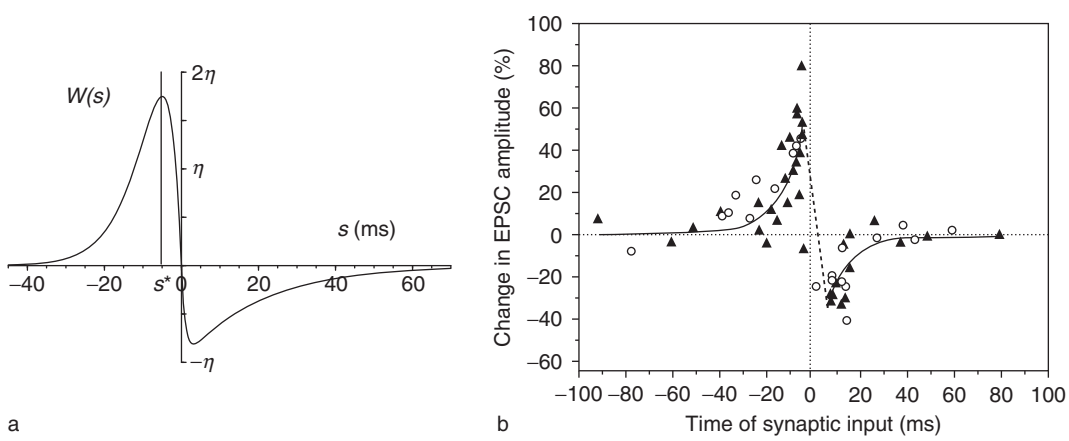

Figure 4 A. The learning window W: (a) in units of the learning parameter η as a function of the delay $s = t_i^f - t^n$ between the presynaptic spike arrival at synapse i at time t_i^f and the postsynaptic firing at time t^n; (b) experimentally obtained learning window of a cell in the rat hippocampus with EPSC as excitatory postsynaptic current. In (a), if $W(s)$ is positive (negative) for some s, the synaptic efficacy J_i is increased (decreased). The increase of J_i is most efficient if a presynaptic spike arrives a few milliseconds before the postsynaptic neuron starts firing (vertical dashed line at $s = s^*$). For $|s| \to \infty$ we have $W(s) \to 0$. The form of the learning window and parameter values are as described in eqn [3]. The similarity between (b) and (a) is evident. It is important to realize that the width of the learning window must be in agreement with the other neuronal time constants. In the auditory system, for instance, these are nearly two orders of magnitude smaller, so the learning window's width scales accordingly. (b) From Zhang LI, Tao HW, Holt CE, Harris WA, and Poo M (1998) A critical window for cooperation and competition among developing retinotectal synapses. *Nature* 395: 37–44.

response to cochlear input. In the barn owl, this happens up to 8 kHz. The animal's laminar nucleus is the first place where signals from both ears come together, but the response is completely smeared out because the axonal delays from left and right ears have a broad distribution, approximately 1 ms wide (Figure 5(a)). This is the situation 3 weeks after hatching, when the head of the young barn owl is adult but the owl cannot perform sound localization yet.

Two weeks later, the owl reaches a microsecond precision through an STDP tuning. The result of this is shown in Figures 5(b) and 5(c) for 2 and 5 kHz, respectively. Every axonal delay line preceding a synapse and equipped with the right delay is strengthened; the others are – so to speak – weeded out. Delay lines differing by one or several periods (here 500 or 200 μs) are effectively the same.

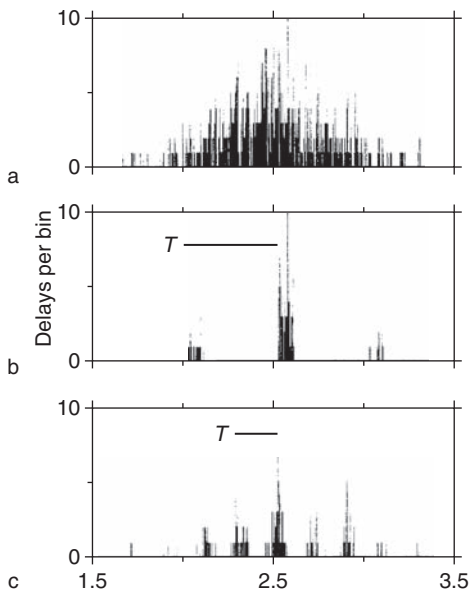

Figure 5 Modeling the selection of synapses connected to axonal delay lines during synaptic pattern formation of the synaptic connectivity on an integrate-and-fire neuron as in Figure 2: (a) 3 weeks after hatching a young barn owl's head is adult but the animal cannot perform azimuthal sound localization yet; in view of the broad (here Gaussian) distribution of delays, with a standard derivation of 0.5 ms, this is not surprising. The input through 300 axons from the left and 300 axons from the right ear is taken to be a periodic, inhomogeneous Poissonian input with frequency $\nu = 1/T$, where ν is the best frequency stemming from the cochlea and $T = \nu^{-1}$ is the corresponding period. After 2 more weeks the barn owl is able to localize a sound source and the resulting delay, i.e., synaptic distribution, is shown in (b) and (c) for best frequencies 2 and 5 kHz, respectively. We have obtained, so to speak, survival-of-the-fittest synapses with delays within in a narrow temporal domain or with corresponding delays shifted by one or more periods, which are identical as far as oscillations by period T are concerned. The rest are less fit and have been eliminated, a striking consequence of spike-*timing*-dependent plasticity. From Gerstner W, Kempter R, van Hemmen JL, and Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. *Nature* 383: 76–78.

The key mechanism involved in STDP is correlating incoming presynaptic and outgoing postsynaptic spikes. If the postsynaptic neuron fires, we often get a backpropagating spike guaranteeing a fast message telling the synapse, so to speak, that the neuron it is on has fired. In the auditory system, neurons are even more compact than usual, so synapses practically directly feel that their postsynaptic neuron has fired.

Let us now imagine a laminar neuron, as in Figure 2. We start with a collection of synaptic efficacies with a uniform value and with an axonal delay line connected to each synapse providing it with spikes coming from the left or right ear; the wide delay distribution is shown in Figure 5(a). Independent of the other delay lines, each delay line is fed with an inhomogeneous periodic Poisson process with frequency ν and angular frequency $\omega = 2\pi\nu$ as input, that is, a point process with rate function $\lambda(t) = c + \sin(\omega t)$ and $c \geq 1$, a constant. If $\lambda(t)$ is small, the probability of getting an input spike is small, if it is large, the probability is large (although always $\lambda \leq c + 1$), and cochlear input is always noisy. As we see from Figures 5(b) and 5(c), during a training period under the influence of external auditory input only few delay lines survive through a collective coupling of postsynaptic firing and well-timed arrival of presynaptic spikes. (In passing, we note that for the sake of convenience the asymptotic saturation of the synaptic efficacies was built in.)

Place Cells in the Rat's Hippocampus

A completely analogous argument to the one for the barn owl auditory system may well hold for hippocampal place cells too. There is, however, a neat difference. The typical time constant in the auditory system is 0.1 ms; on the other hand, it is 10 ms in the hippocampus, that is, roughly two orders of magnitude more. We may therefore expect that the temporal width of the learning window W depends on the physiological structures it lives in. That is, the width of W in the hippocampus is, indeed, 50–100 ms.

Conclusion

From a theoretical point of view, STDP is extremely appealing. Depending on the arrival time of a presynaptic spike at a synapse on the postsynaptic neuron, the synapse is strengthened or weakened depending on the firing time of the postsynaptic neuron. If the postsynaptic neuron fires at $t = 0$ and the synapse is excitatory, then it is strengthened if it receives a presynaptic spike at a time $t < 0$ and thus tells the

neuron to fire before the latter actually fires. That is, it is doing its job properly, and thus its synaptic strength increases.

If, on the other hand, the presynaptic spike arrives at $t > 0$, then it is too late because the postsynaptic neuron has already fired. "Those that come too late shall be punished," and, hence, the synaptic strength decreases. All this is taken care of by the function W, the learning window in Figure 4. Clearly, if the arrival time of the presynaptic spike and the firing time of the postsynaptic neuron are too far apart, then nothing happens.

What we have analyzed in this article is STDP as a stand-alone phenomenon, that is, in its full simplicity. Neurobiological reality is both more complicated and more fascinating than this, in that STDP automatically normalizes synaptic strengths and gives rise to a whole field of applications such as map formation. Let us now briefly consider these phenomena in turn.

The arguments presented so far hold for a single pair of spikes, a presynaptic one and a postsynaptic one. It has turned out experimentally that synaptic modification may also depend on the spiking pattern, that is, on the interspike intervals within each neuron. In addition, the precise position of a synapse on the dendritic tree may play a role as well.

Furthermore, STDP gives rise to a normalization of the synaptic efficacies so that they do not diverge as time proceeds. Finally, STDP allows for a comprehensive explanation of neuronal map formation, one of the most intricate, yet manageable problems in today's neuroscience. As we have seen, a map is a neuronal representation of the outside sensory world. Although STDP itself is local, map formation can be explained in detail by means of STDP in conjunction with more globally operating mechanisms.

In short, because the very same neuronal wetware with the very same delays is used both for learning and for replay, a reproduction (replay) of what is to be learned is possible. Because STDP is local, it cannot handle global correlations. In spite of that, STDP is a very potent mechanism for storing time-resolved spike data at a scale finer than a millisecond and in this way can explain a huge diversity of neuronal phenomena. As such, it exhibits the mathematical universality to explain neurobiological diversity.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Hebbian Plasticity; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Spike-Timing Dependent Plasticity (STDP); Synaptic Plasticity and Place Cell Formation; Synaptic Plasticity: Short-Term Mechanisms.

Further Reading

Abbott LF and Regehr WG (2004) Synaptic computation. *Nature* 431: 796–803.

Bi G and Poo M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience* 24: 139–166.

Caillard O, Ben-Ari Y, and Gaiarsa J-L (1999) Mechanisms of induction and expression of long-term depression at GABAergic synapses in the neonatal rat hippocampus. *Journal of Neuroscience* 19: 7568–7577.

Froemke RC, Poo M, and Dan Y (2005) Spike-timing-dependent synaptic plasticity depends on dendritic location. *Nature* 434: 221–225.

Gerstner W and Abbott LF (1997) Learning navigational maps through potentiation and modulation of hippocampal place cells. *Journal of Comparative Neuroscience* 4: 79–94.

Gerstner W, Kempter R, van Hemmen JL, and Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. *Nature* 383: 76–78.

Hebb DO (1949) The organization of behavior: A neuropsychological theory. Wiley: New York.

Herz AVM, Sulzer B, Kühn R, and van Hemmen JL (1989) Hebbian learning reconsidered: Representation of static and dynamic objects in associative neural nets. *Biological Cybernetics* 60: 457–467.

Kempter R, Gerstner W, and van Hemmen JL (1999) Hebbian learning and spiking neurons. *Physical Review E* 59: 4498–4514.

Kempter R, Gerstner W, and van Hemmen JL (2001) Intrinsic stabilization of output rates by spike-based Hebbian learning. *Neural Computation* 13: 2709–2741.

Kempter R, Leibold C, Wagner H, and van Hemmen JL (2001) Formation of temporal feature maps by axonal propagation of synaptic learning. *Proceedings of the National Academy of Sciences of the United States of America* 98: 4166–4171.

Koch C (1999) *Biophysics of Computation*. New York: Oxford University Press.

Linden DJ (1999) The return of the spike: Postsynaptic action potentials and the induction of LTP and LTD. *Neuron* 22: 661–666.

Lisman J and Spruston N (2005) Postsynaptic depolarization requirements for LTP and LTD: A critique of spike timing dependent plasticity. *Nature Neuroscience* 8: 839–841.

Morrison A, Diesmann M, and Gerstner W (2008) Phenomenological models of synaptic plasticity based on spike timing. *Biological Cybernetics* 98: 459–478.

Rubin JE, Gerkin RC, Bi G, and Chow CC (2005) Calcium time course as a signal for spike-timing-dependent plasticity. *Journal of Neurophysiology* 93: 2600–2613.

Shon AP, Rao RPN, and Sejnowski TJ (2004) Motion detection and prediction through spike-timing dependent plasticity. *Network: Computation in Neural Systems* 15: 179–198.

Song S, Miller KD, and Abbott LF (2000) Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience* 3: 919–926.

Stuart G, Spruston N, Sakmann B, and Häusser M (1997) Action potential initiation and backpropagation in neurons of the mammalian CNS. *Trends in Neuroscience* 20: 125–131.

van Hemmen JL (2001) Theory of synaptic plasticity. In: Moss F and Gielen S (eds.) *Handbook of Biophysics*, vol. 4, pp. 771–823. Amsterdam: Elsevier.

van Hemmen JL, Gerstner W, Herz AVM, Kühn R, and Vaas M (1990) Encoding and decoding of patterns which are correlated in space and time. In: Dorffner G (ed.) *Konnektionismus in*

Artificial Intelligence and Kognitionsforschung, pp. 153–162. Berlin: Springer.

Waters J, Schaefer A, and Sakmann B (2005) Backpropagating action potentials in neurons: Measurement, mechanisms and potential functions. *Progress in Biophysics & Molecular Biology* 87: 145–170.

Wenisch OG, Noll J, and van Hemmen JL (2005) Spontaneously emerging direction selectivity maps in visual cortex through STDP. *Biological Cybernetics* 93: 239–247.

Zhang LI, Tao HW, Holt CE, Harris WA, and Poo M (1998) A critical window for cooperation and competition among developing retinotectal synapses. *Nature* 395: 37–44.

Spiking Neuron Models

W Gerstner, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Introduction: Spikes and the Question of Neural Coding

Neurons communicate with each other by electrical pulses, called action potentials or spikes. Whether the exact timing of action potentials or only the mean firing rate plays a role in neuronal communication is a question of intense debate, often referred to as the problem of neural coding. The question is further complicated by the fact that neuronal firing rates can be defined in at least three different ways: (1) the mean firing rate of a single neuron in a single trial, defined as a temporal average over a sufficiently long time (e.g., number of spikes in a time window of 500 ms, divided by 500 ms); (2) the firing density of a single neuron averaged across several repetitions of the same stimulus, typically defined via the amplitude of a peristimulus-time-histogram as a function of time; and (3) the population rate in a group of neurons with similar properties, defined as an average across the group. From reaction time experiments, it is clear that the first definition of a rate (average across time) cannot be the code used by the neurons. Averaging across several repetitions, as in the second definition, is a useful experimental method but cannot be the strategy of an animal which has to respond to a new stimulus. Finally, while the concept of averaging over groups of equivalent neurons is theoretically appealing, it is hard to see how such groups could be defined in a living brain since the organization into groups might change from one moment to the next, depending on task and stimulus demands.

If rates are such a difficult concept, does that imply that the exact timing of individual spikes matters? Not necessarily. A typical cortical neuron receives input from thousands of other neurons. What matters is probably the number (and spatial distribution across the dendrite) of spike arrivals averaged over an interval of one or a few milliseconds. However, if one spike arrival at an excitatory synapse were removed and replaced by a spike arrival at a different excitatory synapse at a similar dendritic location, the response of the neuron would hardly change. Since the question of coding and of the relevance of exact spike timings for neuronal response is still an open issue, researchers in theoretical neuroscience often use models that take the pulsed (i.e., 'spiking') nature of neuronal signals into account. These are so-called spiking neuron models, as opposed to rate models, in which the activity of neurons is described only in terms of firing rates.

Rate Models in Theoretical Neuroscience

In the field of artificial neural networks, the traditional way of describing neuronal activity has been by means of rate models. In this framework, each unit in a neuronal network is characterized by its rate $r(t)$, which depends nonlinearly on the total stimulation I it received immediately before. In a standard model in discrete time, the rate $r_i(t)$ of a neuron i is given by

$$r_i(t) = g(I_i(t - \mathrm{d}t)) \qquad [1]$$

where $\mathrm{d}t$ is the time step of the simulation and g the nonlinear gain function. Typically, the function g takes a value of zero for a stimulation I below some threshold value and saturates at a maximum value $r^{\max}$ for very strong stimulation. The total stimulation I_i of neuron i has two contributions, that is, the sum over all inputs converging onto neuron i plus potentially some external input I^{ext}:

$$I_i(t) = \sum_k w_{ik} r_k(t) + I^{\mathrm{ext}}(t) \qquad [2]$$

where r_k are the firing rates of other neurons k. In continuous time, eqn [1] is often replaced by a differential equation:

$$\tau \frac{\mathrm{d}r_i}{\mathrm{d}t} = -r_i + g(I_i) \qquad [3]$$

with some time constant τ that describes the response time of the unit to a change in stimulation.

While artificial neural networks using rate models have been successfully used in tasks describing cortex development or memory retrieval, the biological interpretation of the network units is unclear. Each unit i and k in eqn [2] could be interpreted as a single neuron and its rate as a temporally averaged firing rate. In view of the limitations of the firing rate concept discussed above, it must then be concluded that such a single-neuron rate model cannot be used to describe the fast neuronal dynamics during, for example, the rapid reaction of an organism to a new stimulus. Alternatively, each unit could be interpreted as a population of cells and its activity as the population rate. However, in that case the connectivity matrix w_{ij} would refer to the connectivity between populations rather than neurons, and the relevance of model results for electrophysiological measurements is questionable.

Detailed Models of a Spiking Neuron

The classical model of a spiking neuron is the mathematical description of Hodgkin and Huxley of action potential generation in the giant axon of the squid. Generation of spikes in the model arises through the interplay of four nonlinear differential equations. The first equation summarizes the current conservation on a small piece of membrane. A current I^{ext} injected onto the membrane can either charge the capacity C of the membrane or pass through one of the ion channels. With three different ion channel types (one for sodium, one for potassium, and one for all remaining 'leak' components), the change of the voltage is given by

$$C\frac{\mathrm{d}u}{\mathrm{d}t} = I^{\text{ext}} - g_{\text{Na}}m^3h(u - E_{\text{Na}}) - g_{\text{K}}n^4(u - E_{\text{K}}) - g_L(u - E_{\text{L}}) \qquad [4]$$

where g_{Na} and g_{K} are the maximum conductance of the sodium and potassium channel, respectively; E_{Na} and E_{L} are their reversal potentials; and m, h, and n are additional 'gating' variables that describe the state of the channel as a function of time. Each of the three variables m, n, and h follows an equation of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{x - x_0(u)}{\tau(u)} \qquad [5]$$

with empirical function $x_0(u)$ and $\tau(u)$ derived from the experiments (and different for m, n, and h).

Computer simulations as well as mathematical analysis of the system of four nonlinear differential equations shows that action potentials are generated only if the total stimulation I^{ext} reaches a critical value. To a first degree of approximation, one may therefore conclude that action potential generation is an all-or-none process. However, a closer examination of the behavior shows that there is neither a strict voltage threshold nor a strict current threshold. In particular, for time-dependent inputs (e.g., current steps and ramps or randomly fluctuating current as input), not only the momentary voltage or current amplitude matters, but the stimulation history as well.

Formal Spiking Neuron Models

Although neither the Hodgkin–Huxley model nor real neurons have a strict firing threshold, in practice the process of neuronal action potential generation can often be well approximated as a threshold process. The simplest model in the class of formal spiking neuron models is the leaky integrate-and-fire model. In this model, spikes are triggered whenever the membrane potential u reaches a given threshold θ. Below threshold, the membrane potential is described by its capacity C and input resistance R. Current conservation gives the linear differential equation.

$$C\frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{u}{R} + I^{\text{ext}} \qquad [6]$$

If u reaches the threshold θ, the spike time t^f is noted and the membrane potential reset to a fixed value u_{reset}, before integration of eqn [6] is resumed. The essence of the leaky integrate-and-fire model is a clear separation between a completely linear (passive) subthreshold regime and a strict firing threshold. The leaky integrate-and-fire model is today the standard neuron model in many simulation studies concerning the dynamics of large neuronal networks or the question of neural coding.

Generalizations of the leaky integrate-and-fire model have been proposed in several directions. First, since some neurons exhibit in the subthreshold regime a damped oscillatory response to changes in the input (as opposed to a simple exponential decay), the voltage eqn [6] can be coupled with a second equation. As long as this second equation is linear, the subthreshold behavior of the neuron is still completely linear, that is, doubling the amplitude of an input generates a subthreshold response which is twice as large. Second, refractory effects (i.e., reduced responsiveness of a neuron immediately after firing) have been included. The combination of refractory effects with a linear subthreshold behavior leads to the spike response model with a membrane potential

$$u(t) = \eta(t - \hat{t}) + \int \kappa(t - \hat{t}, s)I^{\text{ext}}(t - s)\mathrm{d}s \qquad [7]$$

where $\hat{t}$ is the firing time of the last spike of the neuron, η describes the form of the action potential and its spike after-potential, and κ the linear response to an input pulse. The next spike occurs if the membrane potential u hits a threshold θ from below, in which case $\hat{t}$ is updated. Hence the main characteristics of the Spike Response Model are identical to those of the leaky integrate-and-fire model, that is, a linear subthreshold regime in combination with a strict firing threshold. However, the spike response model allows refractory effects to be included (since the functions η and κ depend on the time since the last firing) as well as subthreshold oscillations (expressed through an appropriate shape of η and κ).

A third generalization of the standard leaky integrate-and-fire model concerns adaptation. By introduction of a second equation that summarizes processes on a slower timescale, it is possible to describe the slow adaptation of the neuron after a step in the stimulating current.

Fourth, the spike triggered by a strict threshold process at the soma can be combined with one or several additional equations describing the (passive) properties of the dendrite. Finally, the strict threshold can be replaced by a smooth threshold process if eqn [6] is turned into a nonlinear equation:

$$C\frac{\mathrm{d}u}{\mathrm{d}t} = -f(u) + I^{\mathrm{ext}} \qquad [8]$$

where $f(u)$ describes the nonlinearities of the membrane in the vicinity of the firing threshold. Two standard choices of $f(u)$ are the quadratic integrate-and-fire model $f(u) = a\ (u-u_1)\ (u-u_2)$ and the exponential integrate-and-fire model $f(u) = -(u/R) + b\ \exp[(u-\theta)/\Delta]$. Both models have free parameters, that is, a, u_1, and u_2 for the quadratic and b, θ, Δ, and R for the exponential integrate-and-fire model, that can be used to adapt the model and put it into a desired firing regime.

Limitations of Formal Spiking Neuron Models

Formal spiking neuron models approximate the electrical properties of neurons by one or a few equations that summarize basic features of normal neuronal behavior with a small number of phenomenological parameters. Formal spiking neurons are therefore not suitable for predicting electrophysiological experiments under nonstandard stimulation conditions. In particular, since all electrical properties of a neuron are summarized in one or two phenomenological equations, these models cannot be used to predict changes of behavior caused by, for example, pharmacological blockage of specific ion channels. Moreover, since the spatial structure of real neurons is not represented in detail, all nonlinear dendritic effects cannot be incorporated into formal spiking neuron models (passive properties, however, can be). Similarly, since the model summarizes neuronal behavior under some reference condition, slow changes in neuronal behavior, caused by slow current ramps or accumulation of intracellular calcium or a simple fatigue of the neuron, for example, cannot be captured.

However, under some stimulation conditions, formal spiking neurons perform surprisingly well. A direct comparison of the spike response model with the Hodgkin–Huxley model during stimulation with random current shows that up to 90% of spike times of the Hodgkin–Huxley model are correctly predicted by the spike response model with a temporal precision of 2 ms. Moreover, a spike response model with a second equation describing adaptation has also been used to predict spike times of layer-V pyramidal neurons in rat cortex under random current injection. The same time-dependent input $I^{\mathrm{ext}}(t)$ was given to both a model neuron and a pyramidal neuron. If the fluctuating current had large amplitude, the neuron itself generated spikes reliably with the same timing across several repetitions of the experiments, but less so if the fluctuation amplitude was reduced. Similarly, the model neuron was able to correctly predict the spike times of the pyramidal neuron (up to 70%) if the fluctuation amplitude of the current was high, but much less well if it was low. Thus, under random current injection, formal spiking neurons describe action potential generation in pyramidal neurons to a high degree of accuracy.

Spiking Neurons in Large Networks

A major advantage of formal spiking neurons such as the leaky integrate-and-fire model is their simplicity, which has two important consequences. First, it is possible to simulate neural networks with a large number of neurons at a reasonable numerical cost. Second, network properties such as the mean firing rate of neurons in a network of randomly connected integrate-and-fire units can be studied analytically with tools from mathematical probability theory, statistical physics, and bifurcation theory.

One important insight that has arisen from studies with formal spiking neurons is the importance of the subthreshold regime for cortical activity. In order to have large networks of interacting spiking neurons working in a state in which individual neurons show irregular firing with broad interspike interval distributions, it turns out to be necessary that the total drive from thousands of excitatory inputs is approximately balanced by inhibition. If this is the case, the neuronal membrane potential hovers normally in the subthreshold regime but stays close to the threshold so that neurons are very responsive to small changes in the input.

Furthermore, studies with networks of spiking neurons have been used repeatedly to elucidate the potential of different neural codes. For example, the stability and reproducibility of spatiotemporal spike patterns in networks of integrate-and-fire neurons have been studied in the context of synfire chains, a potential coding mechanism relying on precise spike times. Similarly, it can be understood why and under what conditions a population of spiking neurons responds instantaneously to changes in the input, that is, much faster than the membrane time constant that characterizes the passive response of the membrane. Finally, formal spiking neurons have been used for studying the functional consequences of spike timing-dependent plasticity.

The example of oscillatory activity allows one to see how formal spiking neurons can be used to predict by a purely mathematical argument the firing activity in large networks. For the sake of simplicity, we suppose that the network consists of N neurons with identical properties and that every neuron is connected to all other neurons by synapses of the same strength. Since we are interested in oscillatory activity, we assume that all neurons in the network fire at the same time, except one that lags behind by a small amount. Will this neuron eventually join the group of the others? To answer this question, we calculate the total synaptic input to this specific neuron, generated by the action potentials in the group of synchronous neurons. Knowing the input, we can derive from eqn [6] or eqn [7] the time course of its membrane potential, and from the time course, we can predict its firing time, that is, the moment when the membrane potential hits the threshold. An analogous calculation is repeated to predict the firing time of the group of synchronous neurons and hence the timing difference between the activity of the single neuron and that of the group. The one neuron lagging behind will eventually join the group of synchronous neurons (i.e., the oscillation is stable) if the timing difference is reduced from one firing cycle to the next. This is just a simple example, but similar mathematical arguments can be used to predict the population activity in large networks and answer questions such as: What is the mean firing activity of neurons in the network? Do neurons fire asynchronously, or do the firing times tend to cluster? Does the network activity show spontaneous oscillations? Will the network switch to a different state on a new input? If yes, could this explain short-term memory in neural networks?

To summarize, formal spiking neuron models are a highly simplified and compressed way of describing action potential generation in real neurons. While such an approach has obvious limitations and cannot be used to study detailed properties of isolated neurons (such as active dendrites or effects of pharmacological blockage of specific channels), the models have been useful in the past to elucidate the essence of spike generation in single neurons as well as principles of neuronal coding in large neuronal networks, and they will probably remain important tools for modeling studies in the future.

See also: Action Potential Initiation and Conduction in Axons; Hodgkin–Huxley Models; Population Codes: Theoretic Aspects; Spike-Timing Dependent Plasticity (STDP); Spike-Timing-Dependent Plasticity Models.

Further Reading

Abeles M (1991) *Corticonics*. Cambridge, UK: Cambridge University Press.

Brette R and Gerstner W (2005) Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology* 94: 3637–3642.

Fourcaud-Trocme N, Hansel D, van Vreeswijk C, and Brunel N (2003) How spike generation mechanisms determine the neuronal response to fluctuating input. *Journal of Neuroscience* 23: 11628–11640.

Gerstner W and Kistler WK (2002) *Spiking Neuron Models*. Cambridge, UK: Cambridge University Press.

Izhikevich EM (2004) Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks* 15: 1063–1070.

Jolivet R, Rauch A, Lscher H-R, and Gerstner W (2006) Predicting spike timing of neocortical pyramidal neurons by simple threshold models. *Journal of Computational Neuroscience* 21: 35–49.

Koch C and Segev I (2000) The role of single neurons in information processing. *Nature Neuroscience* 3(supplement): 1171–1177.

Rieke F, Warland D, de Ruyter van Steveninck R, and Bialek W (1996) *Spikes Exploring the Neural Code*. Cambridge, MA: MIT Press.

Spinal Cord Injuries

A Gorio, G Marfia, D Bottai, and A M Di Giulio, University of Milan, Milan, Italy

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Throughout the ages spinal cord injury (SCI) has been known as a devastating affliction, with little hope of any treatment's resulting in a meaningful recovery. The unknown Egyptian author of the Edwin Smith Surgical Papyrus accurately described the symptoms of a complete cervical myelopathy. The hopelessness of treatment of SCI was already known 5000 years ago, with the author advising that it was "an ailment not to be treated." Hippocrates advocated the closed reduction of spinal deformities by traction on an extension bench, a method used by physicians throughout the centuries. The first surgical procedure for SCI was performed in the seventh century by Paul of Aegina, who was also considered the originator of the decompressive laminectomy. Sommering discussed the surgical treatment for trauma to the spinal cord in 1793. The advances in surgery in the nineteenth century, such as antisepsis, had little effect on the dismal fate of those suffering from SCI. As reported by Guttman, Cushing reported that during World War I, 80% of the spinal cord-injured died within the first week of injury.

The advent of antibiotic therapy and the availability of whole blood greatly reduced the acute mortality rate for SCI to approximately 20% during World War II. Furthermore, this war provided the impetus for the development of a center for comprehensive treatment, rehabilitation, and research of the spinal cord-injured. This pioneering facility, the Spinal Cord Injuries Centre of the Stoke Mandeville Hospital in Aylesbury, England, was headed by Sir Ludwig Guttmann. Many physicians who trained in this facility subsequently developed similar centers throughout the international medical community. In recent decades, the battle against SCI has been waged in all phases of health care. Prevention of injuries due to trauma has been a major public health concern. This is evident in public awareness programs such as those attempting to prevent drunk driving and diving accidents, as well as safety equipment improvements in motor vehicles. Unfortunately, these efforts have not satisfactorily reached the right target since the number of motor vehicle accidents has greatly increased in very recent years. Such accidents represent more than 70% of the causes of SCI. More than 60% of the affected people are between 15 and 30 years of age. Estimates of SCI prevalence range from 721 to 906 cases per million. These estimates correspond to between 36 000 and 40 000 spinal cord-injured persons alive in Italy. Currently, the mortality rate within 1 year after SCI is approximately 6%. Critical factors in this improved prognosis of SCI have been the advent of emergency medical services and improvements in the acute management of the spinal cord-injured, as well as of the multiple trauma patient.

Although uncommon, SCI is tragic since it completely alters the lives of many of the most productive members of society. Although SCI poses a significant health problem, social ills are linked directly and indirectly to its etiology. The economic cost of SCI is staggering. Despite continuously decreasing lengths of stay for initial hospitalization, mean total hospital charges are increasing. The following figures refer to the United States. In 1986 the mean initial hospital charge was estimated to be $91 400. Latest estimates of individual health care and living expenses attributable to SCI are $198 000 for the first year postinjury and $24 100 each subsequent year, with costs increasing with severity of injury. Based on the current estimates of incidence and prevalence, the annual direct cost of SCI to the United States is $8 billion. If indirect costs such as lost wages and productivity are added, the annual national cost of SCI is nearly $19 billion.

The need for effective acute intervention, both to limit the number of permanently impaired patients and to give real hope to the newly injured, has never been greater. Within the past 2 decades, the intense research into traumatic SCI has yielded only a controversial pharmacological treatment, high-dose methylprednisolone administered acutely after injury. Glucocorticoid therapy seems useful only when administered within 8–9 h of injury. The mechanisms by which glucocorticoids exert their moderately beneficial effects remain unclear, although the anti-inflammatory action of the steroids have been proposed. Yet despite the fundamental inadequacy of high-dose glucocorticoid treatment, it has remained the only available and mandatory therapy for SCI.

With significantly improved mortality rates after SCI, current research is focused on reducing the devastating morbidity after SCI, in both the acute and the chronic phases following injury. Cord injury caused by penetrating forces is usually circumscribed around an area of severe damage that corresponds to the level of the penetrating force. Damage to the spine is quite variable and generally less severe than that to the cord. The injuries due to blunt forces are generated by more complex mechanisms. The force affects the vertebral column and is modified by this impact

before it is transmitted to the cord. Thoracic cord injuries have the highest proportion of complete injuries when compared with cervical or thoracolumbar segments. Primary injury is that which is an immediate result of direct action of mechanical force on the spinal cord. Secondary injury consists of all pathological events initiated by the original mechanical insult. Research has flourished in the field of secondary injury involving both the brain and the spinal cord, and many mechanisms and their potential treatments are common to both organs. The secondary injury following SCI is fundamentally the result of a stroke-like ischemic insult; it is associated with dysregulated arterial flow, venous stasis, edema, and vasospasm, all of which serve to lower spinal arterial perfusion to critically low levels. The biochemical and metabolic events associated with vascular dysregulation are numerous and lead to the progressive degeneration of white and grey matter that may greatly exceed the early mechanical damage. Three major hypotheses have been proposed to account for such damage, namely, disturbances in cellular calcium, free radical production, and excitotoxic amino acid release. A disturbance in intracellular calcium homeostasis is widely acknowledged as being a critical element in the initiation of neuronal injury. Calcium is involved in the regulation of numerous cellular functions, including neurotransmitter release, neuronal excitability, and second messenger and cellular enzyme action. Ischemic insult rapidly results in energy failure, depolarization, and blockade of energy-dependent ionic pumps that normally control intracellular concentrations of sodium and calcium ions. The increased intracellular sodium ion concentration is thought to induce a reversal of calcium ion extrusion through a Na^+/Ca^{2+} antiporter, causing a further increase in intracellular Ca ions. The energy depletion and elevated intracellular Ca ions cause a massive release of excitatory amino acids (e.g., glutamate) that results in excitotoxic neuronal death. The formation of destructive reactive oxygen species (ROS) is strongly associated with damage derived from ischemic and traumatic injuries. These highly reactive radicals, which can cause the peroxidation of membrane phospholipids, cellular proteins, and nucleic acids, attack both neuronal and glial cell membranes as well as the spinal cord vasculature. The spinal cord appears to be particularly vulnerable to such oxidative injury. It contains relatively high concentrations of readily peroxidable fatty acids and is relatively poorly endowed with protective antioxidant enzymes and endogenous antioxidant compounds. Levels of arachidonic acid, generated as a result of membrane calcium-dependent lipolysis (including calcium activation of phospholipase 2), rises quickly during ischemia. Arachidonic acid can then be metabolized by either cyclo-oxygenases or lipo-oxygenases to produce a variety of vasoactive substances: prostaglandins, prostacyclins, thromboxanes, leukotrienes, and ROS. At site of injury, neutrophils are considered to be one of the primary sources of ROS. A complicated interplay of proinflammatory molecules, anti-inflammatory molecules, and migratory cells also mediates the pathophysiology of the spinal cord following injury. Within the first 12 h, neutrophils first appear, with lymphocytes and microglia then mobilized to the injury site within 24–48 h. Much interest has been focused on proinflammatory cytokines, specifically MIP-2, tumor necrosis factor-α (TNFα), interleukin (IL)-1β, and IL-6, which are key mediators of inflammation following CNS lesions. Overproduction of MIP-2 into injured spinal cord was associated with neutrophil infiltration, indicating a key role played by this mediator in inflammatory cell recruitment induced by SCI. In keeping with this suggestion, reparixin (a blocker of IL-8 receptor) treatment has been found to completely prevent neutrophil and ED-1-positive cell infiltration into injured spinal cord. The complex summation of primary and secondary degeneration events determines the extent of tissue damage, normally extending much farther than the limits of the mechanical impact. The attenuation of cellular and molecular abnormalities underlying the secondary injury may result, then, in a dramatic improvement of recovery from SCI-derived disability.

Neutrophils, and perhaps microglia, responding to tissue injury, may inappropriately destroy adjacent tissues, resulting in extension of the infarct. Neutrophils in high concentrations have been observed to accumulate in CNS regions with low blood flow during the early postischemic reperfusion phase, further contributing to delayed hypoperfusion. Soon after the traumatic injury, a marked increase of IL-8 at the lesion site begins and reaches its peak in 24 h, then subsides in 48 h. IL-8 is likely the most important chemokine leading to migration and accumulation of neutrophils in the affected area. The effect of microglia/macrophage activation in the injured spinal cord remains controversial. Suppression or depletion of microglia/macrophages improves tissue integrity and functional recovery after traumatic SCI. However, intraspinal injection of preactivated microglia or blood monocytes increases axon growth at the injury site and may promote improved locomotor function. Because the activation and migration of macrophages/microglia precedes most inflammatory events within the injured CNS, these cells may direct the overall magnitude, kinetics, and effects of downstream inflammatory processes, secondary degeneration, and endogenous mechanisms of spinal cord repair. Specifically, activated microglia has been linked

with regions of increased blood–spinal cord barrier permeability, axonal degeneration, and increased production of inflammatory neurotoxins, chemokines, and cytokines following traumatic CNS injury. Species comparisons have shown that the phagocytic capacity and release of cytokines and eicosanoids are greater in rats than in mice. Expression of CD4 and CD8 molecules on microglia and macrophages in the injured spinal cord of rats could contribute to species-specific differences in lesion pathology. Historically, CD4 and CD8 antigens define T-helper and cytotoxic T cell subsets, respectively. These molecules are essential for major histocompatibility complex (MHC)-restricted antigen recognition by T cells and T cell activation. In the CNS, these molecules are also expressed constitutively on a subset of rat microglia and are upregulated on some activated microglia/macrophages in models of peripheral nerve injury, cerebral ischemia, and multiple sclerosis (i.e., experimental autoimmune encephalomyelitis). Recently, CD8 was localized to hematogenous macrophages infiltrating the brain during the first week following an ischemic insult. Hematogenous macrophages are recruited to the rat spinal cord after contusion injury, and these cells contribute to the acute secondary pathology associated with SCI. Subsets of these macrophages express CD8. The demonstration that ligation of macrophage CD8 can trigger the release of proinflammatory cytokines such as IL-1, TNFα, and nitric oxide provides a mechanism whereby microglia/macrophages could contribute to secondary injury and cavitation in spinal injury. A complicated interplay of proinflammatory and anti-inflammatory molecules and migratory cells also mediates the pathophysiology of the spinal cord following injury. Within the first 12 h, neutrophils first appear, with lymphocytes and microglia then mobilized to the injury site within 24–48 h. Among proinflammatory cytokines, much interest has focused on MIP-2, TNFα, IL-1β, and IL-6, which are key mediators of inflammation following CNS lesions. Glial cells are a primary target of cytokines and are activated in response to many cytokines, including TNFα and IL-1. This activation can trigger further release of cytokines that might affect local inflammatory responses and neuronal survival. It was suggested that IL-1 might promote neurotoxin release from glia and that TNFα could induce apoptosis directly through the activation of Fas receptors. MIP-2 is a primary chemotactic and activating factor for neutrophil. Overproduction of MIP-2 into injured spinal cord was associated with neutrophil infiltration, indicating a key role played by this mediator in inflammatory cell recruitment induced by SCI. In keeping with this suggestion, reparixin treatment completely prevented neutrophil infiltration into injured spinal cord. Similarly, reparixin administration strongly reduced ED-1-positive cell invasion, suggesting that neutrophil infiltration induced by SCI could play an important role in subsequent mononuclear cell recruitment. In agreement with this hypothesis, the phagocytic activity of neutrophils plays an important role in damaged tissue, but activated neutrophils are also known to produce an array of inflammatory mediators including cytokines and chemokines and to regulate the cascade of events leading to TNFα production. In addition, the increase of proinflammatory cytokines occurring after traumatic SCI has been successfully counteracted by 7 days' reparixin treatment, a novel CXCR2 receptor blocker. The inhibitory effect of reparixin on IL-1β and IL-6 was already observed at day 1 after SCI, further sustaining a key role of infiltrated neutrophils in the production of cytokines observed after traumatic SCI. On the other hand, reparixin treatment reduced MIP-2 and TNFα production starting from day 2 after SCI, indicating that resident cells and infiltrated mononuclear cells could be the main source of these factors. Thus neuroinflammation is an important component of SCI, and drugs capable of modifying this picture may prove useful for the development of a much needed pharmacological control of the early degenerative events following traumatic injury of the cord.

The complex summation of primary and secondary degeneration events determines the extent of tissue damage, normally extending much farther than the limits of the mechanical impact. The degenerated tissues are enclosed in a cavity surrounded by reactive glia. Such a scar represents both a safe border for undamaged spinal cord tissue and a powerful limitation to axonal regeneration across the lesion site. The attenuation of cellular and molecular abnormalities underlying the secondary injury may result, then, in a dramatic improvement of recovery from SCI-derived disability. It is well known that more than 90% of the traumatic injuries to the cord are not complete and a portion of spinal axons are spared. Unfortunately, they are unable to conduct electric impulses. Limiting the progression of secondary injury should make it possible to preserve the excitability of spared axons and reduce the extent of cavity formation. The current commonly used standard therapy consists of the administration of methylprednisolone sodium succinate (MPSS) to attenuate SCI damage by reducing lipid peroxidation, free radical formation, and edema occurring during the posttraumatic ischemia and reperfusion. However, in spite of initial clinical data suggesting that treatment using MPSS within 8 h of injury was beneficial, the current consensus is that MPSS offers at best very limited benefit and may actually cause harm. The potential explanation of this recent outcome may be that MPSS treatment

of SCI may produce beneficial effects by reducing proinflammatory cytokines that are likely involved in the process of secondary injury, but detrimental effects by suppressing trophic factors such as nerve growth factor and glial-derived neurotrophic factor.

Erythropoietin and SCI

The cytokine erythropoietin (EPO), long known for its hormonal effects on bone marrow that sustain an adequate circulating red cell mass, has recently been recognized to function as a markedly potent and widespread endogenous molecule protective of multiple injuries in many tissues and mammalian species. Study has demonstrated robust protection of the spinal cord from ischemia, as well as mechanical injury in the rabbit, rat and mouse. EPO mediates protection via multiple, overlapping pathways, including reducing apoptotic cell death and secondary increases in proinflammatory cytokines, mobilizing endothelial progenitor cells to promote angiogenesis. An impressively broad therapeutic time window exists for a nonerythropoietic analogue administered in a multiple dosage regimen following compressive SCI and for a single dose of EPO administered intravenously at 5000 U per kg of bodyweight. With SCI produced by contusion, there is a temporal window for a single dose of EPO. Although an initial improvement is noted for animals treated with rhEPO with a delay of 24–48 h, this improvement is not maintained; the curve flattens, resulting in a final motor score no different from that following administration of saline. This observation suggests that a critical process occurs in the early phase following SCI which is not effected by subcutaneous EPO ultimately leads to a diminished motor function no different from that of saline-treated animals. Using a rat model of contusive injury, researchers determined the effect of EPO (500–5000 U per kg of bodyweight) administered by different routes and timing with respect to injury and of MPSS (30 mg per kg of bodyweight) for neurological outcome, proinflammatory cytokine production, and histological damage of the spinal cord. EPO (5000 U per kg of bodyweight) given intravenously, intraperitoneally, or subcutaneously reached equivalent serum levels only by 15 h, yet was equally effective when administered immediately following injury. This is consistent with a broad therapeutic time window. Although high-dose EPO and MPSS both suppressed the proinflammatory cytokines TNF, IL-1β, MIP-2, and IL-6 within the spinal cord, only EPO administration provided long-lasting neurological improvement; and suppression of inflammatory cytokines does not prevent the development of secondary injury. Surprisingly, MPSS antagonized the protective effects of EPO. This did not depend on reduced EPO receptor (EPOR) expression as EPOR was upregulated following injury whether or not MPSS was present. MPSS may suppress an EPO-mediated growth factor. These data suggest that MPSS should not be administered with EPO in clinical trials for SCI. EPO treatment also improved preservation of ventral white matter compared with saline-treated rats, and this is associated with a significantly increased number of oligodendrocyte precursors (OPCs), identified by NG2 expression. These findings are in line with previous observations that systemic delivery of rhEPO in an experimental model of autoimmune encephalomyelitis decreases demyelination and increases NG2-positive OPCs. Moreover, it has been shown that rhEPO promotes the formation of oligodendrocyte processes and increases the expression of myelin basic protein. Since OPCs are able to differentiate in myelinating oligodendrocytes both *in vitro* and *in vivo*, their enhanced proliferation likely contributes, together with reduction of apoptosis, to the restoration of oligodendrocyte and myelin density. The expression of Nogo-A, a myelin-associated protein containing domains with strong inhibitory activity of neurite growth that in adult CNS is mainly expressed in oligodendrocyte cell bodies and in the myelin membranes, is also affected by EPO in the injured cord. Nogo-A is concentrated at paranodes of myelinated fibers and contributes to the organization of the axolemmal-glial apparatus and to the segregation of potassium channels in juxtaparanodal regions. The remarkable reduction of Nogo-A labeling levels caused by injury reflects the severe structural alterations of myelinated fibers, while the better preservation of the ventral white matter in the rhEPO group is indicated by Nogo-A labeling levels comparable to those of sham-operated rats. In addition, rhEPO treatment not only attenuates astrogliosis but influences the expression of AQP4, an effect that may be beneficial for the resolution of the edema in SCI.

Stem Cell Administration to Spinal Cord Injuries

One of the main targets of research efforts on traumatic SCI is the transplantation of stem cells with the aim of supplying a favorable environment for the stabilization of the early pharmacological effects obtained with EPO, and possibly for axonal regeneration across the lesion, as some recent reports based on the use of chondroitinase ABC and anti-Nogo antibodies suggest. In most cases of CNS trauma and disease, both neurons and glia are lost. Cell replacement might be a vital step in CNS lesions, in which spared systems cannot supplant the function of lost cells. An excellent example is the loss of motor neurons in the case of injury to the

cervical spine. No matter how well the lesion is bridged or reinnervated, the target motor neurons that project to the muscles are irrevocably lost. Recognizing the need for cell replacement, several laboratories have taken advantage of fetal tissue grafts to replace cells after a variety of CNS insults. Somatic stem cells can be derived from fetal or adult stage in animal life and are tissue specific. The prevailing hypothesis on their physiological role is that they may constitute a reservoir of multipotent cells available for substituting aged or injured cells of the same tissue. Such a result has been shown in some instances, such as the cases of satellite cells in the regeneration of striated muscle, keratinocytes in skin growth, and bone marrow precursors in the generation of blood cells. Somatic stem cells are capable of active and powerful proliferative activity during an unlimited time. In this way, a constant number of progenitor stem cells is maintained throughout life and constantly employed, such as in the physiological replacement of skin and blood cells, or for a sudden need, as in the case of tissue repair. Thus some stem cells will proliferate and maintain the existing pool while others will undergo differentiation. It is still unclear what factors influence such a physiological choice of fate by stem cells lying in the same tissue, but several *in vitro* studies have suggested that the process might be regulated by both local and circulating factors. In the brain, transplanted bone marrow cells contributed to the population of cerebellar Purkinje neurons. These observations have disarmed long-held beliefs in lineage-restricted stem cells with a finite specialization repertoire. These lineage changes have been named 'cellular transdifferentiation' and represent a new life for the clinical interest in stem cell application. Work is under way to clarify whether transdifferentiation of stem cells occurs and is useful when these cells are applied to injured cord.

Neural Stem Cells

It is well known since Ramon y Cajal's pioneering work that injured neurons regrow the degenerated processes (axons and dendrites) rather than dividing into a new neural cell, as happens in other organs, such as liver. Thus the discovery that stem cells are present and proliferate under appropriate conditions in some brain areas represented a major scientific breakthrough. Once purified, neural stem cells (NSCs) show a high rate of cellular division and high proliferative activity. This type of cell has also been isolated and cloned using human brain as a source.

Skin-Derived Stem Cells

Skin-derived stem cells were transplanted into the cord of traumatic SCI-lesioned rats. The advantage of using skin-derived stem cells is that they represent a readily accessible source of neural progenitor cells that can be obtained without invasive procedures. These cells grow in a medium enriched with epidermal growth factor and acidic fibroblast growth factor (nonpermissive (NP) medium), and after 7 days form spheres that can be dissociated into single cells that gradually reform other spheres. The expression of nestin marker was maintained over 15 passages of spheres in NP medium. Single cells after adhesion on a laminin-coated dish have an asymmetric shape; 12% are positive for nestin, neurofilaments, and tubulin III; and more than 80% are positive for glial fibrillary acidic protein (GFAP). Cells dissociated from spheres grown in NP medium were injected into the gracilis and cuneatus fascicle 3 mm caudal or rostral or both to the impact site, performed at T9. Cells were transplanted within 30' of the impact or a week later. Within 3 days most of the transplanted cells migrated into the lesion site and were located at the most outer portion of the cavity, just in contact with the lesion scar. The early transplanted cells reversed their *in vitro* differentiation and became mac-1 positive cells with a macrophage-like cell shape (50% of total cells) within 1 week after transplantation, and at 90 days massive apoptosis was observed, as expected. The late transplanted cells differentiated into cells positive for GFAP (30%), neurofilament (18%), and platelet/endothelial cell adhesion molecule (PECAM; 2%). When immediately after transplantation animals were treated with EPO, the number of PECAM-positive cells increased to 14%, and several of them, as in the picture shown above, were, in continuity with nonexogenous endothelial cells, forming capillaries in the white matter and in the cavity at the site of injury. Whereas early transplants were totally ineffective on locomotor recovery as assessed with the Basso–Beattie–Bresnahan (BBB) scale, the late ones showed a small improvement that became significant 40 days after transplantation. When EPO was used with transplants, there was an immediate improvement of locomotion recovery that was significant as early as 4–7 days after transplantation and maintained thereafter.

Mouse NSCs were supplied within 24 h to the injured cord of mice by direct injection into the cord or via intravenous administration. The lesion was performed as described above for the rat, but the force was smaller, 40 g. The evaluation of the hind limb recovery of function was performed according to the BBB Ohio Scale for mice. Following the traumatic lesion, mice's ability of free locomotion by means of the hind limbs was temporarily lost; 4 days later a slow and gradual recovery began in saline-treated rats. In these animals the maximal extent of recovery (8 points on the BBB Scale) was achieved in 16–20 days. The intraspinal

injection of NSCs promoted a much faster recovery: a significant improvement (12 points BBB Scale) in 2 weeks. However when NSCs were supplied intravenously, the effect was earlier and greater. It must be born in mind that the intraspinal injection of NSCs is a second lesion. Cellular preservation was optimal, and none of the administered NSCs were GFAP-positive, even after 2–3 months of their application. The intravenous route of NSCs administration yielded a better recovery than the intraspinal one. About 2% of total intravenously administered NSCs homed in on the SCI site and survived almost undifferentiated; thus the positive effect of NSC treatment cannot be ascribed to damaged tissue substitution. The NSCs homing in on the injury site triggered within 48 h a local tremendous increase in the expression of neurotrophic factors and chemokines. These cells still maintained the ability to produce neurospheres 1 week after transplantation, when they were recovered from the lesion site and grown *in vitro*. Most NSCs were phagocytated by macrophages at a later time. The environmental modification of lesion site and the augmented expression of neurotrophic factors and chemokines by live NSCs may be the basis of the enhanced recovery of function. More recently, a new population of NSCs has been isolated that migrates to the site of SCI and transdifferentiates into cholinergic neurons bearing axonal and dendritic processes. These processes form throughout the lesion area a rich neuropile that constitutes a favorable environment for axonal regeneration, as shown by the abundant number of TH- and 5-HT-positive axons crossing the injury gap. This new achievement increases hopes of finding a proper cellular therapy for SCI.

Conclusions

Acute SCI-induced paraplegia may be prevented in experimental animals by administration of rhEPO within the 24 h, and the effect is secondary to the attenuation of secondary degeneration and improved myelination. The potential of somatic stem cells to transdifferentiate *in vivo* and to promote recovery when applied after traumatic injury of the spinal cord promises to revolutionize the treatment for such a terrible condition, one that is also extremely costly for public and family budgets. On the other hand, little is known about the mechanisms that regulate stem cell proliferation and differentiation, and a deeper knowledge of their physiological role is needed.

See also: Sensorimotor Plasticity and Control of Movement Following Spinal Cord Injury +D18; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Regeneration and Functional Recovery: Strategies; Spinal Motor Neurons: Properties; Spinal Cord Pain Systems; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Tumors of the Brain and the Spinal Cord.

Further Reading

Bracken MB, Shepard MJ, Collins WF Jr., et al. (1992) Methylprednisolone or naloxone treatment after acute spinal cord injury: 1-year follow-up data. Results of the second National Acute Spinal Cord Injury Study. *Journal of Neurosurgery* 76: 23–31.

Brines M and Cerami A (2005) Emerging biological roles for erythropoietin in the nervous system. *Nature Reviews Neuroscience* 6: 484–494.

Fu ES and Saporta S (2005) Methylprednisolone inhibits production of interleukin-1beta and interleukin-6 in the spinal cord following compression injury in rats. *Journal of Neurosurgical Anesthesiology* 17: 82–85.

Gorio A, Gokmen N, Erbayraktar S, et al. (2002) Recombinant human erythropoietin counteracts secondary injury and markedly enhances neurological recovery from experimental spinal cord trauma. *Proceedings of the National Academy of Sciences of the United States of America* 99: 9450–9455.

Gorio A, Maraschi M, Di Stefano F, et al. (2005) Methylprednisolone neutralizes the beneficial effects of erythropoietin in experimental spinal cord injury. *Proceedings of the National Academy of Sciences of the United States of America* 102: 16379–16384.

Gorio A, Maraschi L, Zadra G, et al. (2007) Reparixin, an inhibitor of CXCR2 function, attenuates inflammatory responses and promotes recovery of function after traumatic lesion to the spinal cord. *Journal of Pharmacology and Experimental Therapeutics* 322: 973–981.

Hall ED and Springer JE (2004) Neuroprotection and acute spinal cord injury: A reappraisal. *NeuroRx* 1: 80–100.

Hayashi M, Ueyama T, Nemoto K, Tamaki T, and Senba E (2000) Sequential mRNA expression for immediate early genes, cytokines, and neurotrophins in spinal cord injury. *Journal of Neurotrauma* 17: 203–218.

Kwon BK, Tetzlaff W, Grauer JN, Beiner J, and Vaccaro AR (2004) Pathophysiology and pharmacologic treatment of acute spinal cord injury. *Spine Journal* 4: 451–464.

Merola A, O'Brien MF, Castro BA, et al. (2002) Histologic characterization of acute spinal cord injury treated with intravenous methylprednisolone. *Journal of Orthopaedic Trauma* 16: 155–161.

Rabchevsky AG, Fugaccia I, Sullivan PG, Blades DA, and Scheff SW (2002) Efficacy of methylprednisolone therapy for the injured rat spinal cord. *Journal of Neuroscience Research* 68: 7–18.

Short DJ, El Masry WS, and Jones PW (2000) High dose methylprednisolone in the management of acute spinal cord injury: A systematic review from a clinical perspective. *Spinal Cord* 38: 273–286.

Szabo C (2003) Multiple pathways of peroxynitrite cytotoxicity. *Toxicology Letters* 140–141: 105–112.

Vitellaro-Zuccarello L, Mazzetti M, Maraschi L, Bosisio P, Gorio A, and De Biasi S (2007) Erythropoietin-mediated preservation of the white matter in rat spinal cord injury. *Neuroscience* 144: 865–877.

Yu Y, Matsuyama Y, Nakashima S, Yanase M, Kiuchi K, and Ishiguro N (2004) Effects of MPSS and a potent iNOS inhibitor on traumatic spinal cord injury. *Neuroreport* 15: 2103–2107.

Spinal Cord Pain Systems

W D Willis Jr., University of Texas Medical Branch, Galveston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Initial Processing of Pain Signals by Nociceptors

Nociceptors

Neural responses that lead to the perception of pain commonly begin with the activation of nociceptors, which are sensory receptors that respond to stimuli that cause or threaten to cause tissue damage (Figure 1). Essentially all peripheral tissues, including skin, muscles, joints, and visceral organs, are innervated by nociceptors. Nociceptive primary afferents are generally supplied by unmyelinated or finely myelinated axons, although a few have Aβ fibers. Noxious stimuli can be mechanical, thermal, or chemical in nature or combinations of these. Transduction of noxious stimuli depends on the activation of protein molecules located in the cell membranes of nociceptors. Examples of such transducer proteins include several transient receptor potential molecules, such as TRPV1s, which respond to noxious heat, lowered pH, and certain chemicals (notably capsaicin), but not to mechanical stimuli, and TRPV2s, which respond to higher temperatures than do TRPV1 receptors but which do not respond to capsaicin or reduced pH. Other transducer proteins encode noxious cold, noxious mechanical stimuli, reduced pH, or a variety of noxious chemicals.

It should be noted that not all pain responses are triggered by the direct activation of nociceptive afferent neurons. Pain can also result more indirectly following damage to nervous tissue. Peripheral neuropathic pain may develop following damage to peripheral nerve, and central neuropathic pain from damage to any of several central nervous system structures, including the spinal cord, the brain stem, and the somatosensory portions of the thalamus or cerebral cortex. The mechanisms underlying neuropathic pain are still unclear but are under intense scrutiny.

Nociceptive Neurons in Dorsal Root and Cranial Nerve Ganglia

The cell bodies of nociceptive primary afferent neurons reside in peripheral sensory ganglia, including the dorsal root ganglia and some cranial nerve ganglia, such as the trigeminal ganglion. The cell bodies of nociceptors tend to be small or medium in size, and their surface membranes contain a variety of active proteins, including neurotransmitter receptor molecules that are also present on the terminals of the neurons in the central nervous system or in the periphery. Activity in the peripheral branch of the axon of a nociceptor will propagate into the appropriate ganglion and may invade the cell body. More important, the action potential will reach synaptic terminals within the central nervous system by way of the central branch of the ganglion cell. This results in the release of excitatory transmitters (such as glutamate, substance P, and calcitonin gene-related peptide) in the dorsal horn of the spinal cord (or medulla) that activate central nociceptive neurons, some of which project to the thalamus or other rostral brain areas and initiate higher-level processing of pain signals.

Nociceptive Neurons in the Spinal Cord and Brain Stem

Nociceptive Interneurons

Recordings from interneurons in the dorsal horn of the spinal cord (or in the pars caudalis of the spinal nucleus of the trigeminal nerve, often designated the medullary dorsal horn) can be used to demonstrate their responses to various types of peripheral stimuli applied to their receptive fields. Often, the interneurons selected have cutaneous receptive fields, favoring the application of well-controlled stimuli. However, in some studies, they are activated by less well-controlled stimuli applied to other tissues, such as a muscle, joint, or visceral structure.

Many dorsal horn interneurons receive a convergent input from different types of receptors supplying the same tissue and/or from several different tissues. In general, workers in the field of somatosensory physiology have classified dorsal horn interneurons on the basis of their responses to graded intensities of mechanical stimuli applied to the skin. Thus, some dorsal horn neurons respond just to input from low-threshold cutaneous mechanoreceptors. Such neurons are classified as 'low threshold (LT)' neurons. Other interneurons can be activated to some extent by low threshold mechanoreceptors but even more effectively by nociceptors. These have been termed 'wide dynamic range (WDR)' neurons. Still other interneurons are excited exclusively (or nearly exclusively) by nociceptors. These are called 'nociceptive-specific (NS)' or 'high threshold (HT)' neurons (Figure 2(a)).

Despite its widespread use, this classification scheme is deficient in several ways. One problem is that the actions of thermoreceptors are ignored. For example, there are neurons in the superficial dorsal

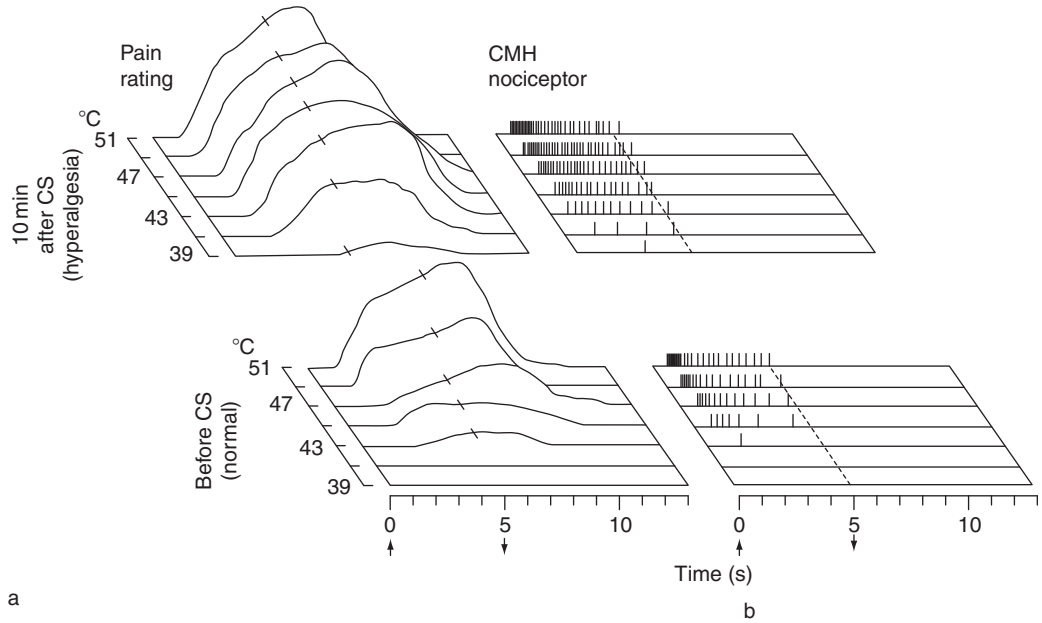

Figure 1 Sensory responses of a human and action potential responses of an unmyelinated C cutaneous nociceptor in a monkey to graded intensities of heat stimuli. In (a) are shown the time course of the pain ratings of the human to a series of heat stimuli. The skin temperature in degrees Celsius is plotted on the *y*-axis and time in seconds on the *x*-axis. The arrows indicate the times of onset and termination of the stimuli. The lower graph was recorded when the skin was normal and the upper graph after a mild burn (50 °C for 100 s). In (b) are peristimulus time histograms showing the responses of a C polymodal receptor that innervated the skin of a monkey subjected to the same stimulus regimen, including the mild burn. Note that the skin damage lowered the threshold and enhanced the responses of the nociceptor; that is, it caused peripheral sensitization of the afferent. The damage altered the pain ratings in a comparable way, resulting in primary hyperalgesia. CMH is C mechanoheat and CS is conditioning stimulus. Reproduced from LaMotte RH, Thalhammer JG, and Robinson CJ (1983) Peripheral neural correlates of magnitude of cutaneous pain and hyperalgesia: A comparison of neural events in monkey with sensory judgments in human. *Journal of Neurophysiology* 50: 1–26, used with permission from The American Physiological Society.

horn that respond to specific thermoreceptors (cold or warm receptors). These neurons are either left out of this classification scheme, or if the neurons have a convergent input from mechanoreceptors or mechanical nociceptors, they are classified on the basis of their responses to mechanical, but not to thermal, stimuli. However, their function may relate more to thermal than to mechanical sensation. Another defect in this classification system is that many nociceptive neurons are activated by stimulation not only of the skin but also of other structures, such as muscle, joints, or viscera.

Another approach to the classification of dorsal horn neurons is by application of cluster analysis. This involves the use of a cluster analysis technique, such as k-means cluster analysis, to separate the responses of a set of neurons according to their similarities to each other and to their differences from the responses of other sets of neurons (Figure 2(b)). The use of different intensities of mechanical stimuli applied to the skin leads to the demonstration of several clusters that resemble the sets of LT, WDR, and HT neurons already mentioned. It would theoretically be feasible to include also the responses of the neurons to other stimuli, such as thermal stimuli or stimuli applied to deep structures, to yield a more refined analysis. However, to do this, the mathematics becomes much more complex.

Nociceptive Spinal Reflex Circuits

Nociceptive spinal cord interneurons may participate either in motor circuits whose activity leads to spinal reflexes or in sensory circuits that signal pain by way of axonal projections that ascend to the brain.

The class of spinal reflex that is activated by nociceptive spinal interneurons is the flexor withdrawal reflex or flexion reflex. When a noxious stimulus is applied to the skin, as for example when a person steps on a tack, the limb directly affected by the noxious stimulus is flexed so that the limb is withdrawn from the stimulus. At the same time, the extensor muscles of the contralateral lower extremity contract in a crossed extension reflex, which functions to provide the increased postural support needed to prevent the individual from falling while the other lower extremity is flexed. Details of the pattern of reflex activity depend on the site of noxious stimulation, a phenomenon known as 'local sign.' The interneurons that participate

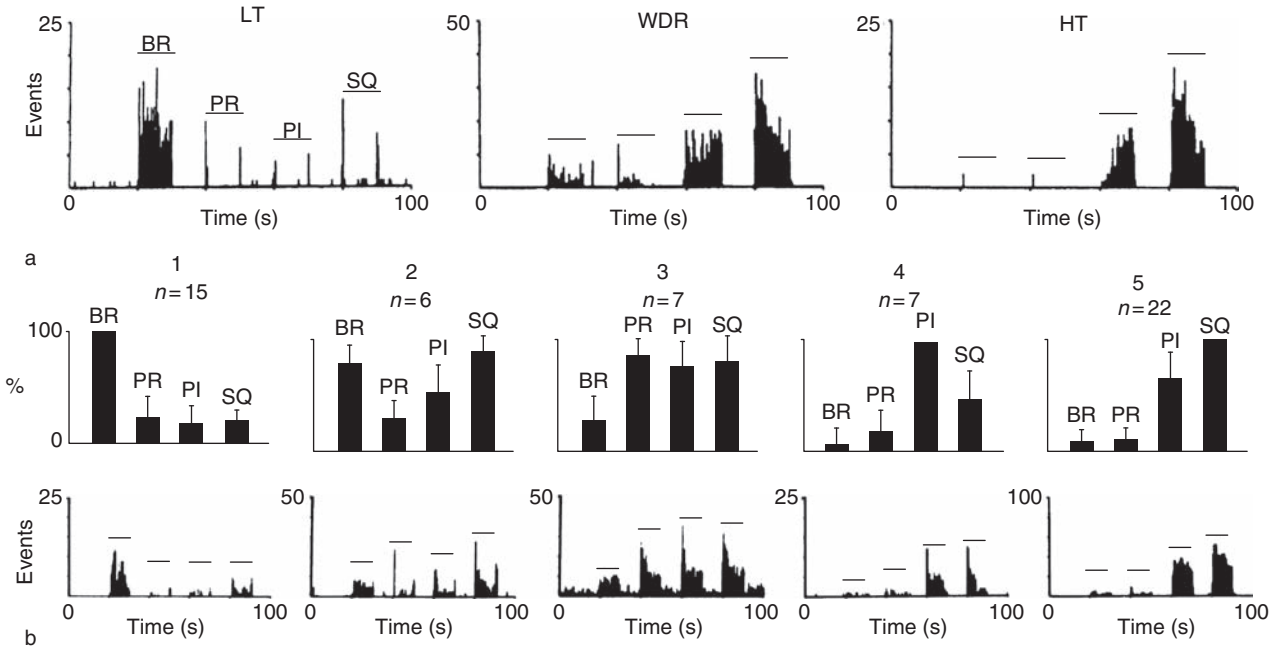

Figure 2 The classification scheme used for spinal cord neurons is based on the responses to graded intensities of mechanical stimulation. The stimuli were brushing (BR) with a camel hair artist brush, compression of the skin with arterial clips having two levels of grip force, the weaker causing a sensation of pressure (PR) when applied to human skin and the stronger (PI) causing overt pain but no visible damage. The final, most intense stimulus was squeezing (SQ) the skin with serrated forceps, a damaging stimulus. The three representative neurons whose responses are shown in the peristimulus time histograms in (a) were classified as low threshold (LT), wide dynamic range (WDR), and high threshold (HT) on the basis of their responses to the series of four mechanical stimuli (applied during the times indicated by the horizontal bars). The bar graphs in (b) show the response profiles of five classes of interneurons in the cervical spinal cord dorsal horn as determined by k-means cluster analysis, using the responses of these interneurons to the same four mechanical stimuli as in (a). The peristimulus time histograms at the bottom show the responses of representative interneurons that belonged to each cell class. From Sorkin LS, Ferrington DG, and Willis WD (1986) Somatotopic organization and response characteristics of dorsal horn neurons in the cervical spinal cord of the cat. *Somatosensory Research* 3: 323–338.

in the flexor withdrawal reflex are thought to include those activated by afferents known as the 'flexion reflex afferents.'

The flexion reflex is commonly used as a means of assessing responses to painful stimuli, either in the awake or in a lightly anesthetized state. However, since the flexion reflex occurs below a complete spinal cord transaction, it is clearly not a direct measure of pain sensation.

Projection Neurons

Spinal cord nociceptive neurons include not only interneurons but also neurons that project to a variety of targets in the brain (Figure 3). Several of these ascending spinal cord tracts have an important nociceptive component. These projection neurons are also generally classified as LT, WDR, or HT neurons, but in addition they can often be identified according to their projection targets. This can be accomplished in anatomical experiments by retrogradely labeling the neurons with a tracer substance that is injected into a site in the brain that receives synaptic terminals from the spinal projection neurons of interest, followed by retrograde axonal transport of the marker and its identification at the spinal cord level by an immunohistochemical staining procedure or by fluorescence microscopy.

Alternatively, the target of the axon of a projection neuron can be identified in physiological experiments by antidromic activation. This involves electrical stimulation of the axon terminals within the target structure and the recording of antidromically propagated action potentials from the vicinity of the cell body of the projection neuron. Antidromic activation is verified by several criteria: (1) the antidromic action potential occurs at a constant latency following repetitive stimulation; (2) antidromic action potentials can follow high frequencies of stimuli, unlike action potentials produced by orthodromic activation; and (3) orthodromic action potentials collide with antidromic action potentials over a predicted range of intervals.

Spinothalamic Tract

Clinical observations have suggested a prominent role for the spinothalamic tract (STT) in pain and in thermal sensations. There is also evidence of a role

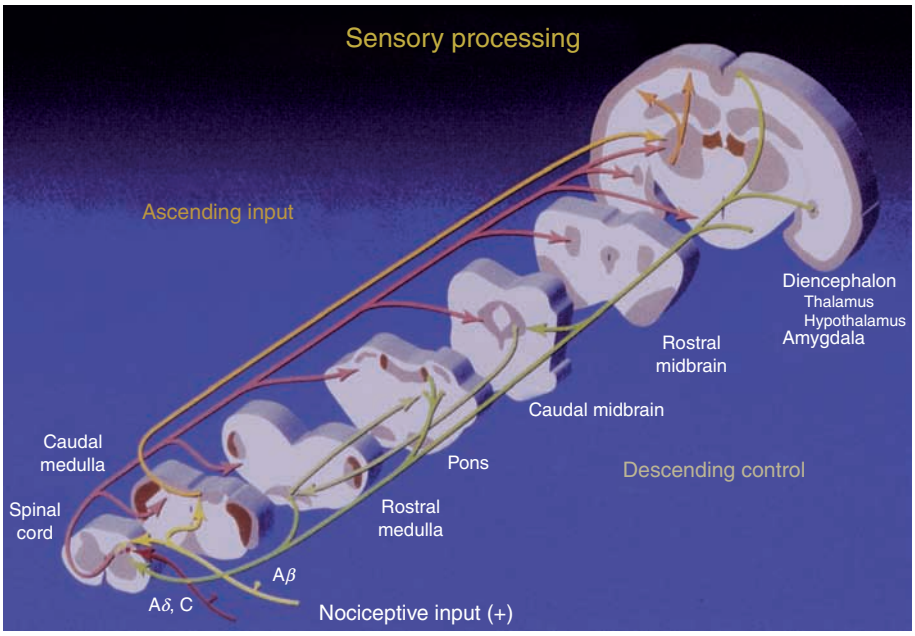

Figure 3 Sensory projection systems that ascend from the spinal cord to the brain (ascending input) are indicated on the left side of the illustration. Primary afferent input involving Aβ mechanoreceptors (yellow arrows at the bottom right) ends in the dorsal horn and also projects by way of collaterals through the dorsal column to the dorsal column nuclei in the caudal medulla. The dorsal column nuclei then project contralaterally to the ventral posterior lateral thalamic nucleus (light orange arrow), and this nucleus in turn projects to the somatosensory cortex (orange arrows). Aδ and C afferents (shown by the maroon arrow at the bottom right) synapse in the dorsal horn. Projections to the brain (red arrows at the left) are to the brain stem reticular formation, thalamus, hypothalamus, and amygdala. Descending control systems (green arrows on the right) include the corticospinal tracts, forebrain projections to the periaqueductal gray (PAG) and rostral ventral medulla (RVM), as well as projections of the PAG, locus coeruelus and RVM to the spinal cord dorsal horn. From Willis WD and Westlund KN (2004) Pain system. In: Paxinos G and Mai JK (eds.) *The Human Nervous System*, 2nd edn., Amsterdam: Elsevier/Academic Press.

of the STT in tactile sensation. On the basis of the effectiveness of cordotomies placed in slightly different parts of the anterolateral quadrant of the spinal cord in producing a loss of cutaneous pain and temperature sensations, it has been suggested that the component of the STT that mediates thermal sensation ascends in the lateral funiculus just dorsal to that which mediates pain.

Neurons that give rise to the STT are concentrated in Rexed's laminae I and IV–VI, although some are also found in other laminae, such as lamina VII (the intermediate region) and IX (which is in the ventral horn; Figure 4).

There is a special, numerous set of STT neurons in the upper cervical segments. However, the receptive fields of these are very large and often bilateral, and so they are unlikely to signal the location of a discrete somatosensory stimulus. Similarly, a population of STT neurons in the intermediate region and ventral horn of the spinal cord that projects to the central lateral intralaminar nucleus of the thalamus have large receptive fields and are more likely to be involved in arousal and in attentional mechanisms than in the sensory discrimination of pain. By contrast, many of the STT neurons in the dorsal horn have relatively discrete receptive fields (sometimes as small as those of the mechanoreceptive neurons found in the dorsal column nuclei) and so could contribute to the localization of a painful stimulus applied to the body surface, such as a pin-prick. Although a number of laboratories have reported that neurons in lamina I of the spinal cord are generally of the HT variety, studies in our laboratory on the monkey spinal cord indicate that primate lamina I STT cells are usually of the WDR type. Furthermore, many WDR STT cells with cell bodies located in the deeper layers of the dorsal horn have dorsally extending dendrites that may reach laminae I and II (Figure 5(c)). Other deeply positioned STT cells are of the HT type (Figure 5(a)). Rarely, an STT cell is classified as an LT cell (Figure 5(b)).

Craig and his colleagues have observed that some lamina I STT cells can be activated by weak thermal stimuli, suggesting that these play a role in thermal

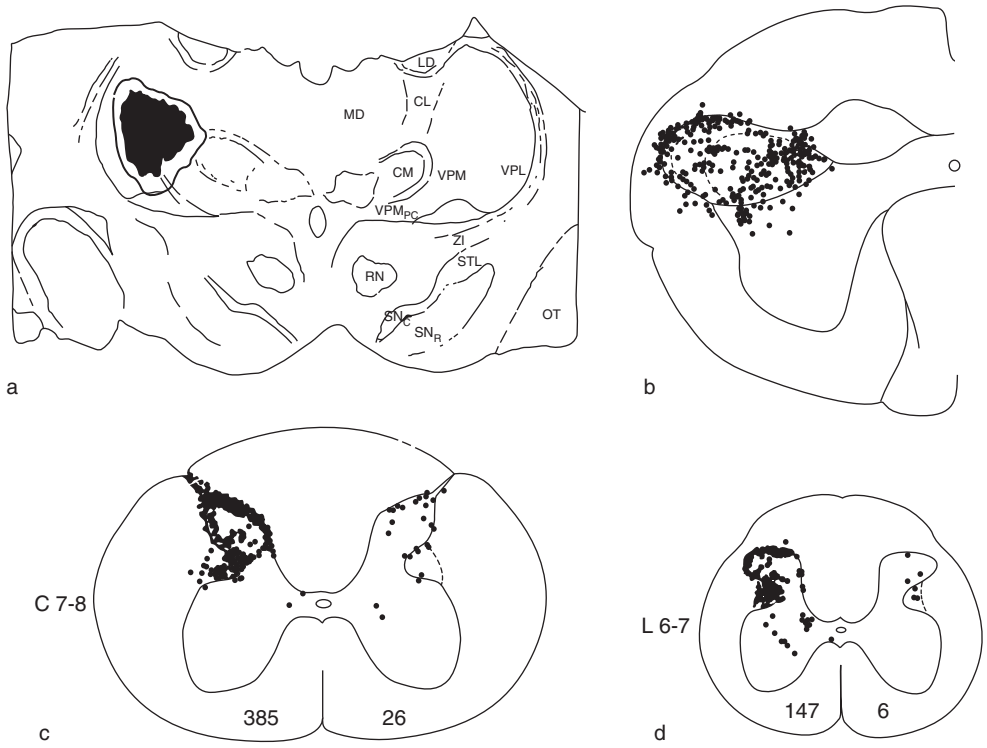

Figure 4 Projections of neurons of the spinothalamic tract and of the spinal nucleus of the trigeminal to the ventral posterior lateral and medial thalamic nuclei in a monkey. The site of injection of the retrograde tracer, cholera toxin subunit B, into the ventrobasal complex is shown in (a). Retrogradely labeled neurons concentrated in lamina I and in the deep dorsal horn are plotted in drawings of the trigeminal spinal nucleus (b) and of the cervical (C) and lumbosacral (L) enlargements ((c) and (d)). From Willis WD and Westlund KN (2004) Pain system. In: Paxinos G and Mai JK (eds.) *The Human Nervous System*, 2nd edn. Amsterdam: Elsevier/Academic Press.

sensation. The axons of these lamina I STT cells ascend in the lateral funiculus of the spinal cord white matter in a position that is consistent with the location of the thermosensitive spinal projection to the thalamus.

The STT provides a prominent input to the ventral posterior lateral nucleus, and there are also terminations of spinal cord projections in the ventral posterior inferior nucleus, the posterior complex, and the intralaminar complex, particularly in the central lateral nucleus. There is a recent controversy about the thalamic nuclei in which lamina I STT cells terminate. Craig and colleagues suggest that most or all lamina I STT cells terminate in a newly defined nucleus termed by them the VMpo nucleus (posterior part of the ventral medial nucleus). However, several other groups, including ours, have been able to stimulate the axons of lamina 1 STT cells antidromically and to label such neurons retrogradely from the ventral posterior lateral nucleus (Figure 4), and thus we do not believe that the claims of the Craig group reveal a balanced picture of the spinothalamic input to the thalamus. Other reservations about the proposed VMpo nucleus by the laboratory of Edward Jones are based on anatomical considerations.

Postsynaptic Dorsal Column Pathway

The postsynaptic dorsal column pathway originates from neurons of the dorsal horn located in laminae IV–VI and to a lesser extent from neurons in the area near the central canal. The axons of these neurons enter the dorsal column (and also the dorsal part of the lateral funiculus) and ascend to terminate in the dorsal column nuclei (n. gracilis or cuneatus, depending on whether the cell of origin is below or above the midthoracic level, respectively).

Postsynaptic dorsal column neurons in cats or monkeys are generally classified as WDR neurons, and rarely as LT or HT neurons. However, in addition to mechanical stimulation of the skin, these neurons can often be activated by stimulation of visceral organs. For example, many postsynaptic dorsal column neurons in the lumbosacral spinal cord can be excited by noxious colorectal distention. Similar responses can be recorded from neurons in the nucleus gracilis and in the ventral posterior lateral thalamic nucleus, and these responses are nearly eliminated following transection of the fasciculus gracilis (Figure 6). This arrangement appears to be the

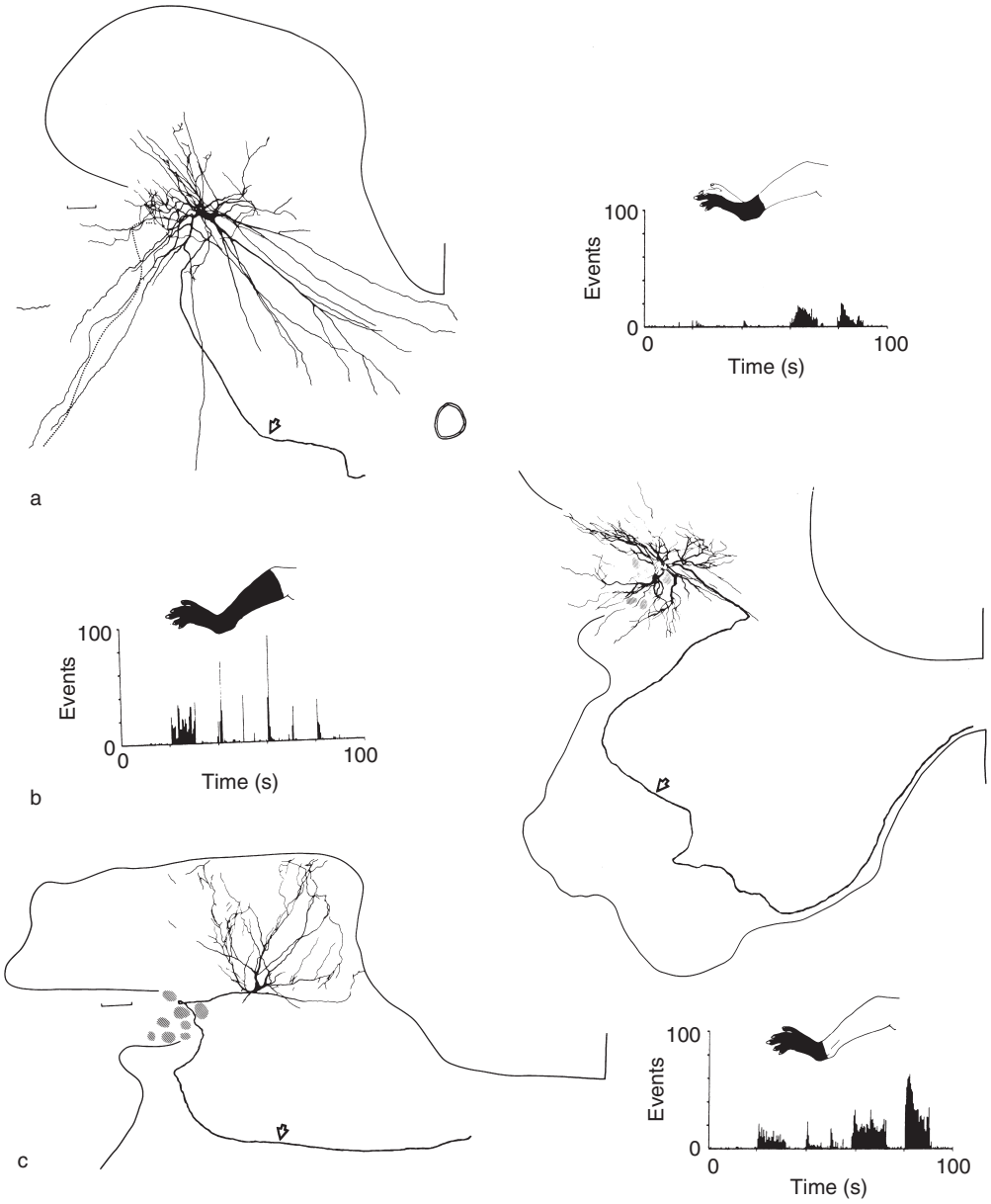

Figure 5 Primate spinothalamic tract (STT) neurons were injected intracellularly with horseradish peroxidase and later reconstructed histologically. Drawings of the soma, dendrites, and axons of three such STT cells are shown in (a)–(c), along with peristimulus time histograms of the responses of these neurons to graded intensities of mechanical stimuli. The receptive field of each is also shown on a drawing of the leg. The STT cells in (a), (b), and (c) were classified as high threshold, low threshold (LT), and wide dynamic range, respectively. The LT STT cell was one of very few recorded in our studies. Reproduced from Surmeier DJ, Honda CN, and Willis WD (1988) Natural groupings of primate spinothalamic neurons based on cutaneous stimulation: Physiological and anatomical features. *Journal of Neurophysiology* 59: 833–860, used with permission from The American Physiological Society.

explanation for the success of midline myelotomies at T10 in eliminating pelvic cancer pain in human patients, as described by Hirshberg, Nauta, and their colleagues.

Spinoreticular and Spinomesencephalic Tracts

There are abundant projections of spinal cord neurons to various parts of the brain stem reticular formation, including the medullary, pontine, and midbrain reticular formation (Figure 3). Particular nuclei in which these projections terminate in the medulla and pons include the dorsal reticular nucleus and the central nucleus of the medulla, the nucleus reticularis gigantocellularis and nucleus paragigantocellularis, and the nucleus pontis caudalis. The spinomesencephalic tract ends in several midbrain nuclei, including the nucleus cuneiformis, the periaqueductal gray, the parabrachial nuclei, and the superior colliculus.

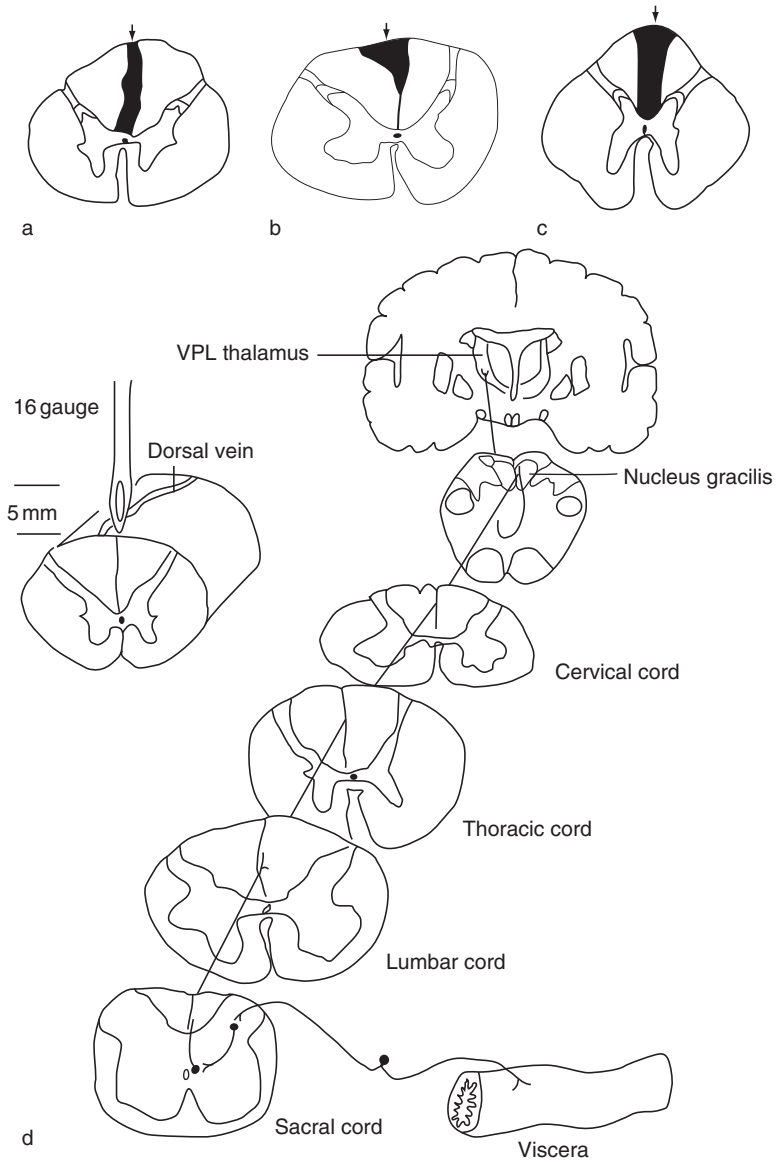

Figure 6 The postsynaptic dorsal column pathway transmits visceral pain signals, and interruption of the dorsal columns can relieve pelvic cancer pain. The drawings of transverse sections of the spinal cord in (a)–(c) show the locations of midline myelotomies that interrupted the medial part of the dorsal column in three patients. The patient whose spinal cord section at T10 is shown in (a) experienced complete relief from colon cancer pain during a 3-month survival period. The patient whose lesion at T7 is shown in (b) had lung cancer and had incomplete pain relief, perhaps because the lesion did not extend sufficiently far bilaterally or was not rostral enough. The patient whose lesion is shown in (c) had several years of relief of pelvic cancer pain. On the left in (d) is a drawing of one surgical technique for midline myelinotomy. At the right is a drawing of the postsynaptic dorsal column pathway, including a primary visceral afferent nociceptor, an interneuronal connection in the spinal cord, a projection from a postsynaptic dorsal column neuron to the nucleus gracilis, and projection from the nucleus gracilis to the contralateral ventral posterior lateral (VPL) thalamic nucleus. From Willis WD and Westlund KN (1997) Neuroanatomy of the pain system and of the pathways that modulate pain. *Journal of Clinical Neurophysiology* 14: 2–31.

Recordings have been made from spinoreticular neurons identified by antidromic activation from the reticular formation, and such neurons have also been retrogradely labeled following injections of tracers into particular parts of the reticular formation. In general, spinoreticular neurons that end in the pontomedullary reticular formation are located in deep layers of the dorsal horn or in the intermediate region and medial ventral horn. Spinomesencephalic neurons are often located in the dorsal horn in positions similar to those of STT cells that project to the lateral thalamus.

Spinoreticular neurons may have restricted cutaneous or deep receptive fields or large, complex receptive fields. They are generally nociceptive, either WDR or HT neurons, but there are some LT spinoreticular

neurons as well. Similarly, neurons of the spinomesencephalic tract are usually WDR or HT neurons, often with complex receptive fields.

It has been suggested that neurons of the reticular formation that are excited by pathways ascending from the spinal cord can in turn activate thalamic neurons that contribute to arousal and to mechanisms of attention. However, the significance of such spino-reticulo-thalamo-cortical pathways has been questioned.

Spinolimbic Pathways

Several pathways ascend from the spinal cord to terminate in brain structures that belong to the limbic system, including the hypothalamus and the amygdala (Figure 3). The spinohypothalamic tract has two components that ascend either laterally or medially through the brain stem. The origin of these projections is neurons located in the superficial and deep laminae of the dorsal horn, as well as in lamina X. Terminals are found bilaterally. It is interesting to note that the axons cross the midline in the hypothalamus and descend into the posterior thalamus and then caudally throughout the entire longitudinal extent of the brain stem. Most of the spinohypothalamic neurons from which recordings have been made were nociceptive, either WDR or HT neurons, but a small fraction were LT cells. Other spinolimbic pathways include a continuation of the spinoparabrachial pathway to the amygdala (the spino-parabrachio-amygdalar pathway), as well as a direct spinoamygdalar pathway and several other projections from the spinal cord to the septal nuclei, the nucleus accumbens, and other limbic structures. Presumably these spinolimbic pathways participate in the motivational-affective components of pain responses.

Plastic Changes in Pain Responses

Peripheral and Central Sensitization

Strong noxious stimuli can alter the responsiveness of peripheral and central nociceptive neurons. For example, damage or inflammation can cause nociceptive primary afferent neurons to develop an enhanced responsiveness (Figure 1). This change is known as peripheral sensitization. Nociceptive neurons in the spinal cord dorsal horn can also develop increased responsiveness following a strong noxious stimulus. This process is called central sensitization.

Peripheral sensitization can involve the activation of transducer proteins, such as TRPV1, by the action of sensitizing agents, including inflammatory mediators such as bradykinin, prostaglandins, monoamines, neurokinins, and growth factors, on signal transduction cascades. Central sensitization is triggered by the synaptic release of neurotransmitters such as glutamate, substance P, and calcitonin gene-related peptide onto dorsal horn neurons. An increase in intracellular sodium concentration results both from the influx of sodium ions through the surface membrane by way of *N*-methyl-D-aspartate (NMDA) receptor channels and voltage-gated sodium channels, as well as by release of sodium from intracellular stores following the activation of metabotropic receptors. The increased sodium concentration activates signaling molecules, including nitric oxide synthase (NOS) and sodium/calmodulin activated protein kinase II (CaMKII). A downstream consequence is the phosphorylation of various proteins, including subunits of NMDA (Figure 7) and alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors.

This leads to stronger excitation of the affected central neurons by several mechanisms, including (1) increased responsiveness of NMDA (Figure 8(a)) and AMPA glutamate receptors; (2) reduced effectiveness of γ-aminobutyric acid and glycine actions

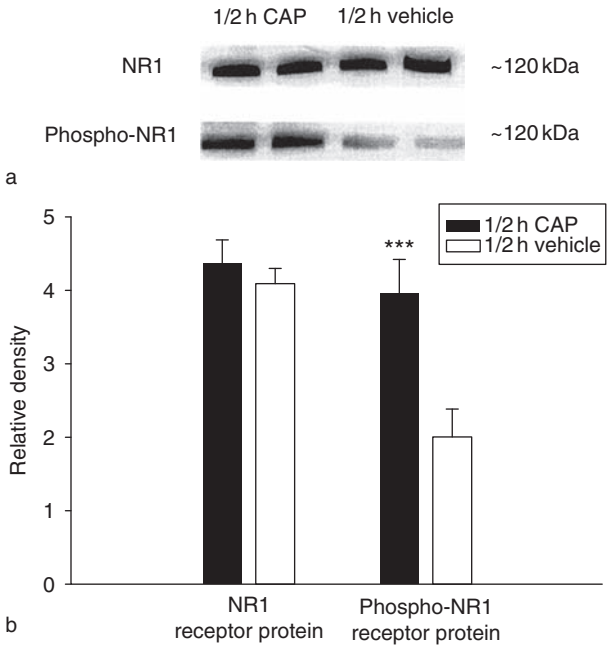

Figure 7 After an intradermal injection of capsaicin (CAP) into the rat foot, there is an increase in the amount of phosphorylated NR1 subunit protein of *N*-methyl-D-aspartate receptors in the lumbosacral spinal cord, as shown by Western blots. In (a) are immunoblots showing NR1 and phospho-NR1 protein 1/2 h after a vehicle injection (right lanes) in comparison with these proteins 1/2 h after CAP injection (left lanes). The bar graph in (b) shows the relative densities of the proteins. Note that phospho-NR1 increased significantly, but NR1 did not, following CAP injection. From Zou X, Lin Q, and Willis WD (2000) Enhanced phosphorylation of NMDA receptor NR1 subunits in spinal cord dorsal horn and spinothalamic tract neurons following intradermal injection of capsaicin in rats. *Journal of Neuroscience* 20: 6989–6997.

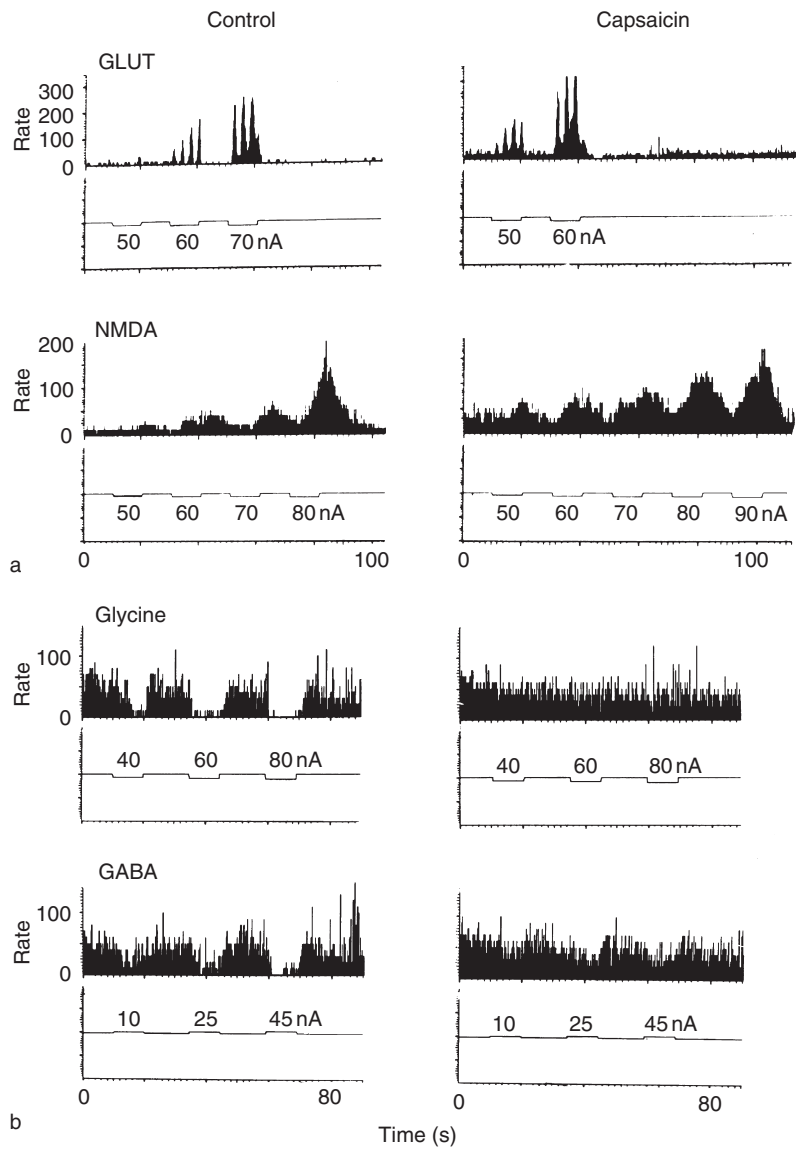

Figure 8 The responses of primate spinothalamic tract (STT) cells to iontophoretically released amino acid neurotransmitters are shown before (left column) and after (right column) the intradermal injection of capsaicin into the skin of the receptive field. The peristimulus time histograms in (a) show the responses of an STT cell to glutamate (GLUT) and to *N*-methyl-D-aspartate (NMDA). From Dougherty PM and Willis WD (1992) Enhanced responses of spinothalamic tract neurons to excitatory amino acids accompany capsaicin-induced sensitization in the monkey. *Journal of Neuroscience* 12: 883–894. The current monitors below each histogram indicate the current doses. The records in (b) are the effects of iontophoretically applied glycine and γ-aminobutyric acid (GABA) before and after capsaicin. Reproduced from Lin Q, Peng YB, and Willis WD (1996) Inhibition of primate spinothalamic tract neurons by spinal glycine and GABA is reduced during central sensitization. *Journal of Neurophysiology*. 76: 1005–1014, used with permission from The American Physiological Society.

on inhibitory amino acid receptors (Figure 8(b)); and (3) increased trafficking of AMPA receptors into the postsynaptic membrane. Central sensitization in the spinal cord can be compared to long-term potentiation in such brain structures as the hippocampal formation.

Behavioral Consequences of Sensitization

Peripheral sensitization is thought to account for the development of primary hyperalgesia following a strong noxious stimulus (see Figure 1). Central sensitization has been proposed to help account for secondary mechanical allodynia and hyperalgesia. These sensory changes have been well described in humans (Figure 9). Alterations in the excitability of primary afferent nociceptors have been demonstrated in both human and animal experiments, and changes in the responses of central nociceptive neurons to peripheral stimulation, including the responses of primate spinothalamic neurons, have been observed in animal

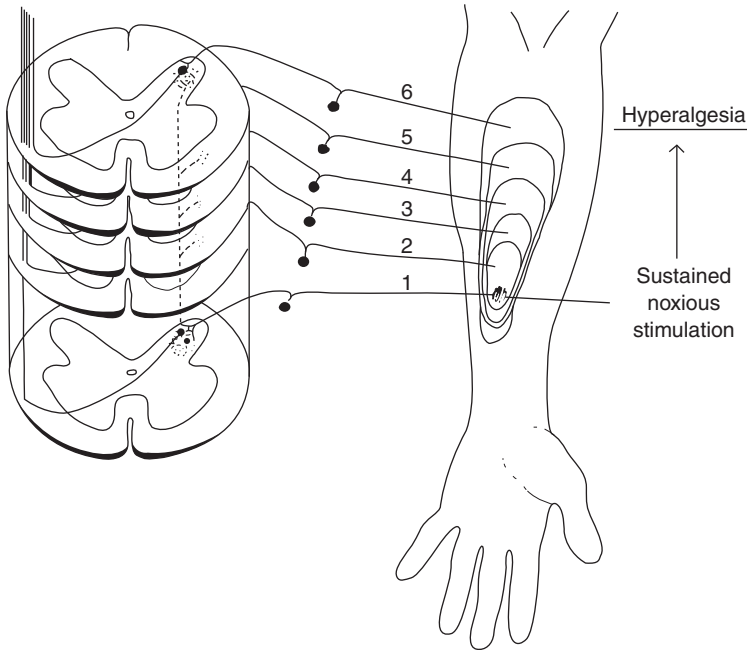

Figure 9 The spread of secondary hyperalgesia across the skin following application of a strong but localized noxious stimulus (such as a small burn or injection of capsaicin). The pain produced by this stimulus is transmitted to the contralateral side of the brain by nociceptive projection neurons that are activated by the stimulus. However, nociceptive circuits within the ipsilateral dorsal horn also transmit the signal to projection neurons at other levels of the spinal cord. These neurons undergo central sensitization, making them more responsive to inputs from normal skin. For example, application of a tactile stimulus in an area of skin several centimeters from the damaged area will now result in pain (tactile mechanical allodynia), and intense mechanical stimulation will now result in hyperalgesia. From Hardy JD, Wolff HG, and Goodell H (1967) *Pain Sensations and Reactions*. New York: Hafner.

experiments. Parallel changes in the behavioral responses of rats have also been reported. Secondary mechanical allodynia is of particular interest since this term describes a process by which touch becomes painful. Patients afflicted with this condition feel pain in response to contact with clothing or bedding. One mechanism by which a tactile stimulus can become painful would be an increase in the responses of WDR neurons to innocuous mechanical stimuli, with a resultant increase in action potential generation, leading to the transmission of pain signals to the brain.

See also: Nociceptor Responses; Pain Pathways: Descending Modulation; Pain and Plasticity; Pain and Genes; Pain Management; Pain: Central; Pain: Neuroimaging; Reticulospinal System; Visceral Pain.

Further Reading

Belmonte C and Cervero F (1996) *Neurobiology of Nociceptors*. New York: Oxford University Press.

Blomqvist A and Berkley KJ (1992) A re-examination of the spino-reticulo-diencephalic pathway in the cat. *Brain Research* 579: 17–31.

Burstein R and Giesler GJ (1989) Retrograde labeling of neurons in spinal cord that project directly to nucleus accumbens or the septal nuclei in the rat. *Brain Research* 497: 149–154.

Campbell JN, Basbaum AI, Dray A, et al. (eds.) (2006) *Emerging Strategies for the Treatment of Neuropathic Pain*. Seattle, WA: IASP Press.

Depaulis A and Bandler R (eds.) (1991) *The Midbrain Periaqueductal Gray Matter: Functional, Anatomical, and Neurochemical Organization*. New York: Plenum.

Dougherty PM and Willis WD (1992) Enhanced responses of spinothalamic tract neurons to excitatory amino acids accompany capsaicin-induced sensitization in the monkey. *Journal of Neuroscience* 12: 883–894.

Eccles RM and Lundberg A (1959) Synaptic actions in motoneurones by afferents which may evoke the flexion reflex. *Archives Italiennes de Biologie* 97: 199–221.

Fang L, Wu J, Lin Q, et al. (2002) Calcium-calmodulin-dependent protein kinase II contributes to spinal cord central sensitization. *Journal of Neuroscience* 22: 4196–4204.

Fang L, Wu J, Zhang X, et al. (2003) Increased phosphorylation of the GluR1 subunit of spinal cord AMPA receptor in rats following intradermal injection of capsaicin. *Neuroscience* 122: 237–245.

Geppetti P and Holzer P (eds.) (1996) *Neurogenic Inflammation*. Boca Raton, FL: CRC Press.

Graziano A and Jones EG (2004) Widespread thalamic terminations of fibers arising in the superficial medullary dorsal horn of monkeys and their relation to calbindin immunoreactivity. *Journal of Neuroscience* 24: 248–256.

Gybels JM and Sweet WH (1989) *Neurosurgical Treatment of Persistent Pain: Physiological and Pathological Mechanisms of Human Pain*. Basel: S. Karger.

Hardy JD, Wolff HG, and Goodell H (1967) *Pain Sensations and Reactions*. New York: Hafner.

Hirshberg RM, Al-Chaer ED, Lawand NB, et al. (1996) Is there a pathway in the posterior funiculus that signals visceral pain? *Pain* 67: 291–305.

LaMotte RH, Thalhammer JG, and Robinson CJ (1983) Peripheral neural correlates of magnitude of cutaneous pain and hyperalgesia: A comparison of neural events in monkey with sensory judgments in human. *Journal of Neurophysiology* 50: 1–26.

Lin Q, Peng YB, and Willis WD (1996) Inhibition of primate spinothalamic tract neurons by spinal glycine and GABA is reduced during central sensitization. *Journal of Neurophysiology* 76: 1005–1014.

Magoun HW (1963) *The Waking Brain,* 2nd edn., Springfield, MA: Thomas.

Mendell LM (1966) Physiological properties of unmyelinated fiber projection to the spinal cord. *Experimental Neurology* 16: 316–332.

Nauta HJW, Soukup VM, Fabian RH, et al. (2000) Punctate midline myelotomy for the relief of visceral cancer pain. *Journal of Neurosurgery* 92: 125–130.

Oh U (ed.) (2006) *The Nociceptive Membrane.* Amsterdam: Elsevier/Academic Press.

Owens CM, Zhang D, and Willis WD (1992) Changes in the response states of primate spinothalamic tract cell s caused by mechanical damage of the skin or activation of descending controls. *Journal of Neurophysiology* 67: 1509–1527.

Sherrington CS (1906) *The Integrative Action of the Nervous System.* New Haven, CT: Yale University Press.

Sorkin LS, Ferrington DG, and Willis WD (1986) Somatotopic organization and response characteristics of dorsal horn neurons in the cervical spinal cord of the cat. *Somatosensory Research* 3: 323–338.

Surmeier DJ, Honda CN, and Willis WD (1988) Natural groupings of primate spinothalamic neurons based on cutaneous stimulation: Physiological and anatomical features. *Journal of Neurophysiology* 59: 833–860.

Trevino DL, Coulter JD, and Willis WD (1973) Location of cells of origin of spinothalamic tract in lumbar enlargement of the monkey. *Journal of Neurophysiology* 36: 750–761.

Willis WD (1982) *Control of Nociceptive Transmission in the Spinal Cord.* Heidelberg, Germany: Springer.

Willis WD (1985) *The Pain System: The Neural Basis of Nociceptive Transmission in the Mammalian Nervous System.* Basel: Karger.

Willis WD (2002) Long-term potentiation in spinothalamic neurons. *Brain Research Reviews* 40: 202–214.

Willis WD, Al-Chaer ED, Quast MJ, et al. (1999) A visceral pain pathway in the dorsal column of the spinal cord. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7675–7679.

Willis WD and Coggeshall RE (2004) *Sensory Mechanisms of the Spinal Cord,* 3rd edn. New York: Kluwer.

Willis WD and Westlund KN (1997) Neuroanatomy of the pain system and of the pathways that modulate pain. *Journal of Clinical Neurophysiology* 14: 2–31.

Willis WD and Westlund KN (2004) Pain system. In: Paxinos G and Mai JK (eds.) *The Human Nervous System,* 2nd edn. Amsterdam: Elsevier/Academic Press.

Willis WD, Westlund KN, and Carlton SM (2004) Pain system. In: Paxinos G (ed.) *The Rat Nervous System,* 3rd edn., Amsterdam: Elsevier/Academic Press.

Willis WD, Zhang X, Honda CN, et al. (2001) Projections from the marginal zone and deep dorsal horn to the ventrobasal nuclei of the primate thalamus. *Pain* 92: 267–276.

Willis WD, Zhang X, Honda CN, et al. (2002) A critical review of the role of the proposed VMpo nucleus in pain. *Journal of Pain* 3: 79–94.

Yezierski RP and Burchiel KJ (eds.) (2002) *Spinal Cord Injury Pain: Assessment, Mechanisms, Management.* Seattle, WA: IASP Press.

Zhang X, Kostarczyk E, and Giesler GJ (1995) Spinohypothalamic tract neurons in the cervical enlargement of rats: Descending axons in the ipsuilateral brain. *Journal of Neuroscience* 15: 8393–8407.

Zou X, Lin Q, and Willis WD (2000) Enhanced phosphorylation of NMDA receptor NR1 subunits in spinal cord dorsal horn and spinothalamic tract neurons following intradermal injection of capsaicin in rats. *Journal of Neuroscience* 20: 6989–6997.

Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants

J W Fawcett, Cambridge University Center for Brain Repair, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Why Are Bridges Needed?

Where a spinal cord injury completely transects the cord, the two stumps form a new glia limitans at their ends with a glial scar underneath it, and they then retract apart. With time after injury the cord distal to the injury atrophies, and the distance between the stumps gets larger. In the gap between the two stumps is cerebrospinal fluid with meningeal and other fibroblastic tissue. It is clearly impossible for axons from either stump to regenerate through this environment and into the other stump, which in human spinal cord injuries may be more than 1 cm away. In order for any axon regeneration to occur, a bridge of a material that can promote axon growth will have to be placed across the gap. Most spinal cord injuries are incomplete, which means that there is some spinal cord tissue running across the injury site. The nature and quantity of this tissue varies greatly depending on the injury. Most commonly, fracture dislocations of the cord are associated with injuries in which the gray matter in the center of the cord is lost but there is some white matter sparing around the outside. In other forms of penetrating injury, some gray matter remains and the outer white matter is damaged. In all cases, there is a lesion cavity. In cases of incomplete injury, axon regeneration could in principle occur through the remaining gray or white matter. However, in human cord injuries in which the injury may be 1 cm or more, to expect axons to regenerate for this distance through the scar tissue that is found on the perimeter of the lesion and then on down the cord may be unrealistic. Providing a bridge of axon growth-permissive material across the lesion core or cavity may provide a better chance of regeneration. The more complete the lesion, the more likely it is that it will need to be bridged if there is to be any chance of significant axon regeneration.

Bridge Principles: The On-Ramp, the Off-Ramp, and the Bridge

A bridge that promotes axon regeneration across a spinal cord injury has three zones. The bridge itself is the material that carries regenerating axons across the lesion site. At either end, the bridge interfaces with the damaged spinal cord. Here, the growing axons have to enter the bridge from central nervous system (CNS) tissue or exit the bridge to re-enter CNS tissue; these are known as the on-ramp and the off-ramp, respectively. It is useful to think of bridges in this way, but in reality motor axons enter the bridge from rostral and exit it caudally, whereas sensory axons need to regenerate in the other direction. The two ends of the graft therefore function as both an on-ramp and an off-ramp for different sets of axons. What properties do these three regions of the bridge require?

The bridge must support the long-distance growth of axons and guide those axons toward the off-ramp. After the axons have grown, it must allow the axons to be myelinated where appropriate and provide sufficient trophic support to allow the axons to survive. In order to perform these functions, it needs to either contain myelinating glia or promote their entry after axons have grown.

The on-ramp must attract sprouting axons to leave the environment CNS tissue adjoining the bridge and guide them into the bridge. It is therefore necessary that an inhibitory scar barrier is not induced between bridge and spinal cord. It is also desirable that there be some form of locally active molecule that will attract axons toward the bridge, such as a trophic factor.

The off-ramp is the most problematic part of bridge design. At this point, axons must leave the bridge environment, which is generally attractive to axons, and pass into the more inhibitory CNS environment, passing through the reactive glial environment that may have formed at the interface between bridge and spinal cord. In order to achieve this, it may be necessary to arrange for molecules, such as trophic factors, that attract axons to be present in the spinal cord just beyond the bridge.

What Properties Should a Bridge Have?

The properties desirable for a spinal cord injury bridge must satisfy the requirements described previously. The bridge must integrate into the spinal cord tissue on either side. If axons are to pass between the graft and the spinal cord over the on- and off-ramps, it is not desirable to have a scarlike boundary between the graft and spinal cord tissue. Different glial cell types often do not intermix, leaving a boundary of reactive tissue, as occurs to some extent when Schwann cells and astrocytes interface with one

another. A graft material that integrates seamlessly into the spinal cord stumps allowing axons to cross is preferred. The central part of the bridge must promote long-distance axon growth, provide trophic support, and should either contain myelinating glia or promote their entry from the stumps at either end. At the on-ramp end of the graft, it is necessary that axons are attracted toward the bridge. This can be accomplished if the bridge secretes trophic factors or other growth-promoting molecules. At the off-ramp end of the graft, it is necessary for axons to leave the usually very attractive environment of the graft and pass into the rather inhibitory environment of the injured spinal cord. This can be facilitated in cellular bridges if the bridge cells migrate from the bridge into the spinal cord through scar tissue, bringing axons with them. Providing an attractive cue such as a source of trophic factors in the spinal cord just beyond the bridge is helpful. Treatment of the boundary between bridge and spinal cord to break down scar tissue (e.g., by digestion with chondroitinase) or with cells that ease axons across the interface (e.g., olfactory glia) has enabled axons to grow out of bridges and back into spinal cord.

Schwann Cell/Peripheral Nerve Bridges

The idea of using peripheral nerve grafts to repair injuries to the CNS dates back to the pioneering experiments of Tello and Cajal. The rationale is that axons regenerate well after peripheral nerve injury, and transplanting this obviously permissive environment into the CNS might promote regeneration there. This concept was revived in the 1980s when Albert Aguayo, Peter Richardson, and Sam David used modern axon tracing methods to examine axon regeneration through peripheral nerve bridges transplanted into spinal cord injuries. They observed that many axons grew through the bridges. However, very few axons were able to cross the off-ramp to re-enter CNS tissue. Moreover, almost all the neurons whose axons had grown into the grafts had their cell bodies very close to the graft interface. These findings demonstrate the main problems with the technique. First, CNS axons usually show a substantial regeneration response only when cut close to their cell body. Second, regenerating axons will preferentially grow in the most permissive and attractive environment, and they will therefore not leave the attractive environment of a peripheral nerve graft to enter the inhibitory CNS environment found on the other side of the off-ramp. There have been many subsequent experiments in which peripheral nerve grafts or grafts containing purified Schwann cells have been transplanted. These experiments demonstrate a third problem: Schwann cells do not migrate out of grafts to mix with the astrocytic environment of the spinal cord but, rather, remain in a circumscribed area, with a sharp interface between themselves and reactive hypertrophic astrocytes. The interface between Schwann cell bridges and the spinal cord is therefore not unlike the interface between Schwann cells and astrocytes at the peripheral root entry zones, and these are also a barrier for axon growth from peripheral to central nervous system territory. Schwann cell grafts have been prepared in a number of ways. In many experiments, a cell suspension is injected into the injury site. This is the method usually used to introduce the cells into impactor lesions. For repair of a complete transection, the Bunge laboratory, which has provided much of the recent literature on Schwann cell transplantation, incorporates Schwann cells into a matrigel support and then inserts this into polyacrylonitrile/polyvinylchloride copolymer tubes into which the ends of the cord are inserted. There is some uncertainty about the survival of transplanted Schwann cells. After Schwann cell transplants, there are many Schwann cells in the treated area even months after implantation. However, it appears that many of these cells may be host Schwann cells which have been recruited in some way due to the transplant of exogenous cells.

Olfactory Glia Bridges

In order to overcome some of the previously described problems with Schwann cell bridges for spinal injury repair, several groups have investigated the use of olfactory glia as an alternative. The rationale for the use of these cells is that in the normal adult CNS new axons are constantly growing from the olfactory epithelium into the CNS at the olfactory bulb. The cells over which the axons grow and which carry them into the CNS environment are olfactory ensheathing cells. These cells share many characteristics with Schwann cells, including their ability to promote axon regeneration. However, they appear to integrate better with astrocytes and may have some ability to migrate in the adult CNS environment. Several laboratories have implanted these cells into spinal cord injuries with variable results. The cells are derived either from the olfactory bulb or from the olfactory epithelium. The most positive results come from Raisman's laboratory. They transplant the cells from semipurified cultures, scraped off the dish with their own matrix and then implanted. Long-distance regeneration of the corticospinal and other tracts with dramatic functional recovery is

reported in rodent models. Ramon-Cueto also showed that olfactory glial transplants could promote recovery in a complete transection model. In other laboratories, the results have been less positive, with limited cell migration and little axon regeneration. In all cases, the cells have been implanted into partial lesions of the cord, usually as a cell suspension. The reasons for the different outcomes are not clear. As with Schwann cell transplants, host Schwann cells are recruited to olfactory glial transplants.

Two groups implant olfactory cells into human patients as experimental therapies. Huang's group in Beijing implants cells derived from the olfactory region of midstage human embryos, whereas Lima's group in Lisbon implants adult olfactory epithelium. In neither case are there control groups, and it is not possible to make judgments about the efficacy of the treatments. A small controlled phase I safety trial of olfactory glial transplantation into human spinal cord injuries has been conducted in Brisbane. Initial results suggest that the transplants are safe.

Stem Cell Bridges

As it became possible to grow neural stem cells *in vitro*, and to direct their differentiation to some extent toward neurons or glia, experiments were performed grafting these cells into spinal cord injuries. The expectation that the cells would differentiate to replace the lost spinal cord tissue has not been fulfilled. However, in several cases the transplants have shown protective ability, with the lesions being smaller than in controls. Some axon growth into the grafts has been seen, but this has been fairly modest unless the cells have been modified to secrete neurotrophins. Some progenitor transplants have led to remyelination of axons, many of which are demylinated by spinal cord injuries. This approach has been followed by Keirstead's laboratory and developed with industry support, and clinical trials of the treatment are planned.

Noncellular Bridges

There has been considerable interest in designing bridges for spinal injury repair produced from bioengineered polymers. Degradable biomaterials have included collagen, fibronectin mats, alginates, and poly (glycolic acid)/poly (lactic acid) implants. Nondegradable materials have included PHEMA, PHPMA, polyacrylonitrile/polyvinylchloride copolymer, and carbon nanotubes. These materials have been variously seeded with growth-promoting cells such as Schwann cells and with cells secreting trophic factors.

Combinatorial Treatments to Make Bridges More Effective

The previous discussion shows that there are several problems that need to be solved before transplantation strategies for spinal injury will work. It is necessary to attract axons into the grafts, particularly those whose cell bodies are far away from the site of injury, to promote the growth of axons through grafts and particularly to encourage axons to leave the grafts over the off-ramp to re-enter the spinal cord tissue.

Neurotrophins have been used to encourage axons to grow into grafts and also to help them leave the grafts. The axons entering Schwann cell grafts tend to be mainly those belonging to neurons close to the injury site. Tetzlaff et al. enabled regeneration of axons cut farther from the cell body into a nerve graft by infusing brain-derived neurotrophic factor (BDNF) around rubrospinal neurons, showing that they could then regenerate their axons into a nerve graft placed far from the neuronal cell bodies. Cells grafted into spinal cord injuries have also been genetically modified to secrete trophic factors. Fibroblasts or bone marrow stromal cells secreting BDNF or NT-3 grafted into the cord attract large numbers of host Schwann cells and regenerating axons. Neurotrophin infusion has been used to partially overcome the problem of persuading axons to exit bridges and re-enter the spinal cord across the off-ramp region, with the infusion being made on the spinal cord side of the off-ramp. Implants of NT-3 secreting fibroblasts below spinal cord injuries have been successful in promoting the growth of axons through injuries. However, axons tend to grow into the attractive region of high neurotrophin concentration and not beyond it. A trail of implants placed at intervals down the cord is a possible strategy for inducing longer distance regeneration.

The bacterial enzyme chondroitinase digests the glycosaminoglycan chains of chondroitin sulfate proteoglycans. These proteoglycans are a major inhibitory factor in glial scar_tissue, and the glycosaminoglycan chains are responsible for much of their activity. Because there is intense glial reactivity in spinal cord injuries, and at the ends of peripheral nerve and Schwann cell grafts, the enzyme has been infused on the rationale that it would aid the growth of axons into grafts and their growth back into spinal cord tissue. Fouad et al. showed increased axon growth into and through Schwann cell grafts, and Houle et al. showed much increased regeneration of axons through peripheral nerve grafts and back into spinal cord tissue following chondroitinase treatment, in both cases with increased behavioral recovery.

Olfactory glia have been implanted at the interface between Schwann cell grafts and spinal cord tissue. The cells allowed larger quantities of axons to grow out of the grafts into the spinal cord.

Various signaling pathways are involved in promoting axon growth and regeneration. A number of experiments have shown that elevation of the cAMP level in neurons can increase the ability of their axons to regenerate in the spinal cord. This strategy was also used as an adjunct to Schwann cell transplantation, resulting in improved axon growth, sparing of spinal cord tissue, and greater functional recovery.

Future Perspectives

Repair of human spinal cord injuries will require bridging strategies. Successful bridges will probably combine cellular therapies with bioengineered substrates designed to orient and position the cells. Attention will also have to be given to trophic support of the axons, promoting myelination, and enabling axons to enter and leave the bridges. The result will be complex combinatorial implants. However, for the first clinical trials, simpler designs will be used. It is hoped that a matrix containing Schwann cells or olfactory glia, possibly combined with a strategy to help axons enter and exit the grafts, will be sufficient to produce measurable benefit.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Regeneration: Role of the Extracellular Matrix and the Glial Scar; Spine Plasticity; Cell Replacement Therapy: Mechanisms of Functional Recovery; Nogo-A: Its Role in Axon Regeneration; Olfactory Ensheathing Cells: Regeneration Following Spinal Cord Injury; Peripheral Nerve Regeneration: An Overview; Spinal Cord Regeneration and Functional Recovery: Strategies; Spinal Cord Injuries; Spinal Cord Pain Systems.

Further Reading

Bunge MB and Pearse DD (2003) Transplantation strategies to promote repair of the injured spinal cord. *Journal of Rehabilitation Research and Development* 40: 55–62.

David S and Aguayo AJ (1981) Axonal elongation into peripheral nervous system bridges after central nervous system injury in adult rats. *Science* 241: 931–933.

Fawcett JW (2006) Overcoming inhibition in the damaged spinal cord. *Journal of Neurotrauma* 23: 371–383.

Fouad K, Schnell L, Bunge MB, Schwab ME, Liebscher T, and Pearse DD (2005) Combining Schwann cell bridges and olfactory-ensheathing glia grafts with chondroitinase promotes locomotor recovery after complete transection of the spinal cord. *Journal of Neuroscience* 25: 1169–1178.

Geller HM and Fawcett JW (2002) Building a bridge: Engineering spinal cord repair. *Experimental Neurology* 174: 125–136.

Guest JD, Hiester ED, and Bunge RP (2005) Demyelination and Schwann cell responses adjacent to injury epicenter cavities following chronic human spinal cord injury. *Experimental Neurology* 192: 384–393.

Houle JD, Tom VJ, Mayes D, Wagoner G, Phillips N, and Silver J (2006) Combining an autologous peripheral nervous system 'bridge' and matrix modification by chondroitinase allows robust, functional regeneration beyond a hemisection lesion of the adult rat spinal cord. *Journal of Neuroscience* 26: 7405–7415.

Jones LL, Oudega M, Bunge MB, and Tuszynski MH (2001) Neurotrophic factors, cellular bridges and gene therapy for spinal cord injury. *Journal of Physiology* 533: 83–89.

Kakulas BA (1999) A review of the neuropathology of human spinal cord injury with emphasis on special features. *Journal of Spinal Cord Medicine* 22: 119–124.

Keirstead HS, Nistor G, Bernal G, et al. (2005) Human embryonic stem cell-derived oligodendrocyte progenitor cell transplants remyelinate and restore locomotion after spinal cord injury. *Journal of Neuroscience* 25: 4694–4705.

McDonald JW, Liu XZ, Qu Y, et al. (1999) Transplanted embryonic stem cells survive, differentiate and promote recovery in injured rat spinal cord. *Nature Medicine* 5: 1410–1412.

Nomura H, Tator CH, and Shoichet MS (2006) Bioengineered strategies for spinal cord repair. *Journal of Neurotrauma* 23: 496–507.

Pearse DD, Pereira FC, Marcillo AE, et al. (2004) cAMP and Schwann cells promote axonal growth and functional recovery after spinal cord injury. *Nature Medicine* 10: 610–616.

Plunet W, Kwon BK, and Tetzlaff W (2002) Promoting axonal regeneration in the central nervous system by enhancing the cell body response to axotomy. *Journal of Neuroscience Research* 68: 1–6.

Raisman G (2001) Olfactory ensheathing cells – another miracle cure for spinal cord injury? *Nature Reviews Neuroscience* 2: 369–375.

Ramon-Cueto A, Cordero MI, Santos-Benito FF, and Avila J (2000) Functional recovery of paraplegic rats and motor axon regeneration in their spinal cords by olfactory ensheathing glia. *Neuron* 25: 425–435.

Ramon-Cueto A, Plant GW, Avila J, and Bunge MB (1998) Long-distance axonal regeneration in the transected adult rat spinal cord is promoted by olfactory ensheathing glia transplants. *Journal of Neuroscience* 18: 3803–3815.

Spinal Cord Regeneration and Functional Recovery: Strategies

J W Fawcett, University of Cambridge, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Overall Strategies for Spinal Cord Repair

Spinal cord injury (SCI) disconnects the cord distal to the injury from descending motor inputs, and prevents ascending sensory information from reaching the brain. Lesions above the thorax also block sympathetic outflow, and the sacral parasympathetic output is affected in most patients. In addition, the local damage at the region of injury affects motor neurons and other spinal circuits at that level. Most spinal injuries are incomplete, so in the majority of patients some function in some of these modalities remains. The aim of spinal cord repair is to replace the functions that have been lost, and by doing so to allow the patient to regain as much useful function as possible. The patients' own priorities for return of function depend on their level of the injury. For patients with high cervical injuries affecting the diaphragm, the priority is to regain sufficient motor ability to get off the ventilator, while for patients with slightly lower C4–C5 injuries the priority is to regain some voluntary arm movement. For patients with thoracic injuries, who have normal arm and hand control, the priorities are very different, with sexual function and bowel and bladder control heading the list. There is no treatment at present that comes close to promoting complete repair of the damaged spinal cord. However, there are various interventions currently under development that will probably bring back some useful function to patients. The rationale for these interventions is described in this section, then more detail is provided later.

The earliest stage of intervention after SCI is to attempt to protect the cord from further damage. Considerable additional damage can be caused by unskilled handling of patients by emergency teams, but the standard of training and equipment is now excellent in most countries. Following injury, there is an extended period of lesion extension, with death of neurons and glia caused by excitotoxic, inflammatory, and other mechanisms. There has been a long search for effective neuroprotective treatments, but none is yet proven to be effective in human patients. The National Acute Spinal Cord Injury Studies (NASCIS) trial suggested that high-dose methyl prednisolone, if given within 8 h of injury, had some protective action. However, recent studies suggest that the benefits and adverse effects of this treatment are evenly balanced, and it is not used in many countries.

Axon regeneration is a major target for spinal cord repair. The potential benefits of promoting axon regeneration through and beyond spinal cord injuries are obvious. In current animal experiments axon regeneration has been seen for a maximum of about 4 cm. This could produce a useful motor improvement over three or four spinal segments below the injury. However, there is also a possibility that motor axons regenerating through an injury might connect to intraspinal circuits, leading to a relay of their influence further down the cord. The extent to which these intraspinal relay circuits are functional in the human cord is at present uncertain. For sensory axon regeneration, it is unlikely using present technology that axons can be made to regenerate as far as the sensory nuclei of the medulla. However, as with motor circuits, many sensory axons do not connect directly to the medulla, but make their first synapse close to the level at which they enter the cord, making it possible that information from regenerating sensory axons could be relayed to the brain.

Plasticity is an important target for spinal cord repair. Most spinal injuries are incomplete, leaving some axons passing through the lesion. Promotion of plasticity may therefore allow the remaining connections to have greater effects. Some indication of the potential of this form of treatment can be inferred by studying spontaneous recovery after spinal cord injury in young children. The spinal cord, as with most parts of the nervous system, has a period of great plasticity after birth, which becomes much reduced with maturation. Young children show much greater recovery after equivalent SCI than adults do, demonstrating the potential effects of greater plasticity on spinal cord injury.

In addition to biological treatments to enhance regeneration and plasticity, there will be a considerable part to play for prosthetic devices. Biological treatments are likely to have their greatest effect on the spinal levels just below the injury, so enhancing functions further below these levels may require prosthetic devices.

Protecting the Cord from Further Damage

The mechanisms that increase the size of spinal injuries over time are probably similar to those that expand the size of strokes and other injuries. Excitotoxicity, inflammation, mitochondrial damage, and

other mechanisms are involved. In parallel with attempts to develop neuroprotective strategies for stroke, treatments have been applied to spinal injury models. Following encouraging animal data, high-dose methylprednisolone entered trials with the NASCIS consortium. Some benefit was shown in those patients that received the treatment 3–8 h after injury. In some countries this became the standard of care, but in recent years adverse effects from the treatment have emerged, and in most countries the treatment is now not routinely given. Good preclinical data promoted a clinical trial of the *N*-methyl-D-aspartate (NMDA) receptor blocker gacyclidine, but this did not show clinical benefit. Potential anti-inflammatory and other treatments are undergoing preclinical development, but currently there is no treatment that is generally accepted to be useful for neuroprotection after SCI.

Axon Regeneration Strategies

After lesions in the spinal cord or elsewhere in the central nervous system (CNS) there is no long-distance axon regeneration, unlike in the peripheral nervous system (PNS), where axons can regenerate over long distances. At the site of injury a glial scar forms, and the abortive axon growth terminates in this structure.

Why Does Axon Regeneration Fail in the Damaged Spinal Cord?

Myelin inhibitory molecules Where the spinal cord is damaged, myelin is disrupted, and degenerating axons beyond the lesion also generate myelin debris. This debris is cleared very slowly compared with myelin clearance in the damaged PNS, and remains for months after the injury. This and other differences between PNS Schwann cells (which promote axon regeneration) and oligodendrocytes led to studies examining the effects of CNS myelin on axon regeneration. Myelin from several mammalian species is very inhibitory to axon growth, and removal of oligodendrocytes from the region around a spinal cord injury promotes some axon regeneration. Four types of inhibitory molecules have now been identified. Nogo-A is a member of the reticulon family and is found in oligodendrocytes and some axons. It acts on the Nogo-66 receptor and also on a currently unidentified second receptor. Myelin-associated glycoprotein (MAG) is one of the most plentiful myelin proteins and also has inhibitory effects via the Nogo-66 receptor and via gangliosides. Oligodendrocyte myelin glycoprotein (OMgp), a third myelin molecule, also produces inhibitory effects via the Nogo-66 receptor. The Nogo receptor associates in *cis* with p75, LINGO-1, and TROY to produce inhibitory signaling, much of it via the small guanine triphosphate hydrolase (GTPase) Rho. In addition ephrin-B3 is expressed on oligodendrocytes, inhibiting growth through interactions with Eph receptors on axons. There is therefore a plethora of inhibitory molecules on oligodendrocytes. Many experiments have modified inhibition from just one of these molecules and shown some axon regeneration, but the effects of multiple treatments are not yet established.

The glial scar Wherever the CNS is damaged a glial scar forms, first evident within hours of injury and developing over a period of weeks. Initially microglia, oligodendrocyte precursors, and astrocytes hypertrophy and multiply. Later meningeal cells from the surface or from major blood vessels enter to form a fibroblastic lesion core, which also contains vascular endothelial cells and inflammatory cells. Within this structure there are several inhibitory influences. In the glial tissue surrounding the lesion core several inhibitory chondroitin sulfate proteoglycans (CSPGs) are upregulated, produced by astrocytes and oligodendrocyte precursors. This, together with upregulation of tenascin, leads to a highly inhibitory extracellular matrix through which axons would have to regenerate. Astrocytes also upregulate inhibitory ephrin-B2 and EphA4. In the lesion core the meningeal cells express high levels of members of the semaphorin 3 family, and also slit proteins.

Axon growth ability Axon growth cone activity is influenced by the integration of many signaling pathways. Many of the inhibitory molecules act via the small GTPase Rho, which in turn has effects on the axonal cytoskeleton through the control of actin polymerization and depolymerization. However, there are also positive effects on axon growth which can come from neurotrophins, integrins, adhesion molecules, and other sources. To some extent inhibitory influences from the previously discussed molecules can be counteracted by growth-promoting effects of adhesion molecules and tropic molecules and vice versa. In the damaged CNS the environment surrounding axons is overwhelmingly inhibitory, with the various inhibitory influences overwhelming any growth-promoting effects. In addition, axons vary greatly in their ability to regenerate, with adult CNS axons exhibiting lower growth ability than embryonic axons. Also, some CNS axons regenerate better than others. For instance, in the spinal cord the central branches of sensory axons generally regenerate more vigorously than do other damaged axons.

Treatments to Promote Spinal Cord Axon Regeneration

Neutralizing myelin inhibitory molecules Historically, Nogo-A was the first inhibitory molecule to be discovered in the CNS, and treatments based on its neutralization are therefore well advanced. Many experiments have shown that administration of antibodies able to block the inhibitory effects of Nogo-A can promote the regeneration of axons in the damaged spinal cord, and can promote functional recovery. Only a small proportion of the cut axons have been seen to regenerate, and only for distances of between 1 and 2 cm, but this has been associated with substantial functional recovery. The most recent experiments have used two monoclonal antibodies that bind to the middle section of the molecule, which does not interact with the Nogo-66 receptor. These have been applied to macaque partial cervical injuries and have produced considerable functional recovery. Other Nogo-related treatments have focused on the Nogo-66 receptor and the part of the Nogo-A molecule that interacts with it. Thus a soluble form of the receptor and a peptide that competes with Nogo-A for binding to it both promote axon regeneration after spinal cord injuries. Signaling from the Nogo-66 receptor involves p75, TROY, and LINGO-1. Therapies based on these molecules have not yet been tested in spinal cord injuries.

Overcoming inhibition in the glial scar Much of the inhibition of axon regeneration in the glial scar comes from the various CSPG molecules that are upregulated after injury. The inhibitory properties of these molecules depend largely on the sulfated sugar glycosaminoglycan chains attached to the core proteins. Bacterial enzymes called chondroitinases digest these glycosaminoglycans into disaccharides, and this removes much of the inhibitory activity from the CSPGs. This strategy has been applied to rodent spinal cord injuries, resulting in a degree of axon regeneration comparable to that seen after anti-Nogo-A antibody treatments, with good return of function. Mice lacking expression of EphA4 exhibit enhanced axon regeneration after injury, but there are at present no reagents that can be applied to the injured spinal cord to neutralize Eph/ephrin mechanisms or semaphorin 3 inhibition.

Treatments affecting the regenerative ability of axons Many of the inhibitory mechanisms discussed in the preceding sections converge on a small number of signaling pathways, which in turn affect the activity of axon growth cones via controls on cytoskeletal motility mechanisms. There have therefore been attempts to affect axon regeneration by manipulating these signaling processes. Many of the inhibitory molecules signal through Rho, and blocking Rho from its effector kinase can therefore block inhibitory influences on axon growth. Rho kinase inhibitors and the bacterial Rho ribosylation enzyme C3 have been used for this purpose. The C3 enzyme has been attached to a cell-permeant peptide, allowing it to enter cells, and this compound, named Cethrin, has recently started clinical trials. During development the regenerative vigor of axons declines, and this is associated with a decrease in cAMP levels. Treatments that increase neuronal cAMP have enhanced axon regeneration in tissue culture models and in the damaged rodent spinal cord. Other signaling pathways in which intervention has led to axon regeneration in animal spinal injury models are conventional forms of protein kinase C and the epidermal growth factor (EGF) receptor, and the target of inosine, serine–threonine kinase (N-kinase). In addition to treatments affecting signaling pathways, axon regeneration has been promoted by the application of neurotrophins, which stimulate axon regeneration. Neurotrophins applied at the injury site or beyond it have increased the number of regenerating axons and have acted as chemotropic attractants, bringing axons through the injury site. Neurotrophins have effects both on axons and at the level of the cell body, where they induce expression of regeneration-promoting genes. Brain-derived neurotrophic factor (BDNF) applied to the cell bodies of descending spinal cord tracts has prevented neuronal atrophy and promoted axon regeneration.

Transplantation of permissive cells Axons regenerate over long distances after peripheral nerve damage in association with Schwann cells. This observation led to experiments in which peripheral nerve tissue and Schwann cells were implanted into the CNS, with the result that many CNS axons are stimulated to regenerate in this Schwann cell environment over long distances. Regeneration into these Schwann cell grafts has been considerably enhanced by increasing cyclic adenosine monophosphate (cAMP) levels. However, a problem is that these axons are unwilling to cross the sharp boundary that forms between the implanted Schwann cells and the CNS astrocytes, and to leave the Schwann cell environment and reenter the CNS. To some extent this problem has been solved by using olfactory ensheathing cells, which are similar to Schwann cells and promote CNS axon growth. These cells have some ability to mix with astrocytes and migrate from the transplant site into the host CNS. Regenerating axons are able to cross these transplants and continue on into the host CNS, in some cases with good return of neurological functions. A third strategy has been to transplant embryonic CNS tissue

into spinal cord lesions, on the rationale that axons grow in this environment during development. These transplants have promoted regeneration of various classes of CNS axon, again with return of function. This form of transplant has an added dimension, which is that the neurons in the transplant can receive inputs from regenerating host axons and grow their own axons into the host tissue, thereby acting as relays.

Bridging the gap Human spinal cord injuries can be large, with damaged tissue extending over more than one spinal level. Even the most common injuries due to fracture dislocation produce damage over 1 cm or so. In most human injuries the cord lesion is incomplete, with some remaining spinal cord tissue continuing across the lesion. Experience from animal studies shows that regenerating axons generally grow through this remaining tissue rather than through the fibroblastic lesion core. However, for all lesions, and necessarily for complete lesions, it may be necessary to provide a bridge of permissive material to carry axons across the lesion area. Schwann cells or olfactory ensheathing cells are the obvious candidates for this permissive bridging material. However, given the size and disorganization of the lesion it will probably be necessary to implant them in some form of matrix. Discussion of potential designs is beyond the scope of this article, but progress has been made with oriented fibronectin fibers and with self-aggregating peptide fibrils.

Promoting CNS Plasticity

Most spinal cord injuries are incomplete, leaving some axons passing through the lesion site. In over half of patients these carry some useful function. A potential form of treatment is to enhance the natural plasticity of the spinal cord so that these remaining connections increase their efficacy. A possible problem with this concept is that remaining axons induced to sprout might make inappropriate connections and produce maladaptive changes. As described in the following sections, the general finding is that this has not been a problem and that the spinal cord is able to make good use of the new connections, and function is improved. However, sprouting of small sensory axons induced by nerve growth factor (NGF) has been shown to produce enhanced pain sensitivity.

Decline in Plasticity with Age

A general feature of the CNS is that there is a period of great plasticity immediately after birth, known as the critical period. These critical periods terminate in rats at around 35 days after birth, and in humans around 5 years, depending on the brain area. During these critical periods the compensation for injury and functional recovery from various forms of CNS damage, including SCI, is considerably greater than in later life. It is therefore a reasonable treatment objective to attempt to return to the adult CNS the degree of plasticity that was present in the young. The factors that terminate critical periods are not well understood, and much of what is known comes from investigations into ocular dominance plasticity in the visual cortex. It is generally accepted that increased inhibitory influences from GABAergic inhibitory interneurons plays a part, but removing this inhibition results in epileptic attacks and so is not a practicable treatment option. Recently two other mechanisms have been discovered. At the time that plasticity terminates many neurons become surrounded by condensed extracellular matrix coats known as perineuronal nets (PNNs), which contain several inhibitory CSPGs together with tenascin and hyaluronan. Removal of these structures by digestion with chondroitinase or prevention of their formation results in enhanced plasticity. There is also evidence for an involvement of Nogo-A in plasticity, because anti-Nogo antibodies have been shown to promote spinal cord plasticity, and animals lacking the Nogo-66 receptor display prolonged ocular dominance plasticity.

Treatments That Enhance Plasticity

The adult CNS displays a degree of plasticity which allows some functional recovery after damage. Recently three interventions have been developed that enhance this process: chondroitinase, anti-Nogo-A antibodies, and inosine. Chondroitinase digests the glycosaminoglycan chains on CSPGs and also has activity against hyaluronan. Following chondroitinase treatment of the brain or spinal cord, many of the PNN molecules are removed. This treatment has been shown to return plasticity both to the visual cortex and the spinal cord. Following spinal cord injury, chondroitinase treatment led to a rapid improvement in sensory–motor performance, at least some of which must have been due to promotion of plasticity, and the treatment has promoted recovery in a model in which plasticity must have been the mechanism after peripheral nerve repair with inaccurate connection of the motor and sensory axons. Chondroitinase also promotes sprouting of sensory axons, but without producing hyperalgesia. Anti-Nogo-A was applied to animals with lesions of one corticospinal tract, leading to sprouting of the remaining tract across the midline. The antibody

also produced rearrangement of topography in the motor cortex and recovery of function in stroke models, presumably by promoting plasticity. Manipulation of the Nogo-66 receptor has also promoted sprouting, and prolonged visual cortex plasticity is seen in mice lacking the Nogo-66 receptor. Inosine, whose action is thought to be via N-kinase, was given to rats with partial spinal injuries and promoted both axon regeneration and sprouting, and recovery of function in a stroke model.

Rehabilitation Treatments and Plasticity

As with other conditions in which there is physical damage to the CNS, rehabilitation therapy can improve the abilities of patients. After spinal cord injury, patients usually receive intensive physiotherapy to maximize useful functions. There is also extensive experience of the beneficial effects of physical therapy in other CNS disorders such as stroke. These treatments probably work in part by stimulating and guiding plasticity in the CNS, allowing intact parts of the CNS to take over lost functions. Constraint therapy, in which strapping of the unaffected limb is used to improve function in a disabled arm after stroke, has this aim. It is probable that the effectiveness and speed of action of physical therapy could be improved by treatments, as described previously, that promote plasticity. There is direct evidence for this from the use of cortical stimulation to promote plasticity in stroke models coupled with rehabilitative therapy.

Prostheses and Other Devices

Regeneration and plasticity therapies of the type described in the preceding sections are starting to enter clinical trials, and will eventually become an accepted part of the treatment for spinal injury. However, it seems probable that they will have their maximum effect in the spinal levels immediately below the lesion. Moreover, there is currently no treatment under development that promises to induce a complete repair of a spinal cord injury. Patients will therefore continue to suffer from disabilities for which they require treatment. Alongside advances in regeneration and plasticity treatments have been advances in prosthetic devices designed to aid bladder, bowel, arm, and hand function and devices to aid with standing and walking. Discussion of these approaches is

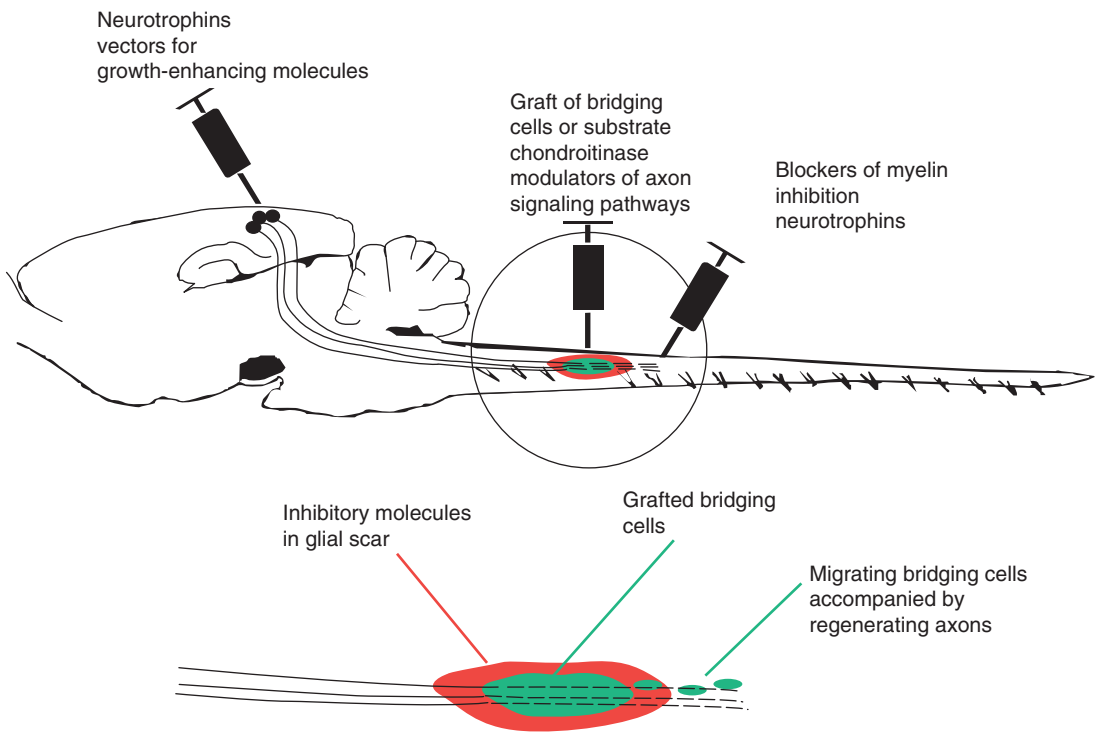

Figure 1 Summary of the classes of treatment that have been developed to promote axon regeneration in the damaged spinal cord. Some treatments, such as bridge grafts and treatments affecting the glial scar, must be applied to the site of injury. Myelin inhibitors are found throughout the central nervous system as well as at the injury site, and need to be applied both to the injury and more distally. Neurotrophins exert a major part of their effect through changes in gene expression in the neuronal cell body, and have been most effective applied here as well as at the injury site.

beyond the scope of this article. However, advances in the field are being made, as are improved methods for interfacing microelectronics with the nervous system. It is probable that in the foreseeable future spinal injury treatment will involve both regenerative treatments and prosthetic devices.

Clinical Trials

Of the treatments that have been successful at promoting recovery of function after spinal cord injury in animal models, two have now begun clinical trials, the Rho inhibitor Cephrin and a humanized anti-Nogo-A monoclonal antibody. Other treatments are in late preclinical development. The development of new regeneration and plasticity-inducing treatments, which have been effective in animal models of spinal cord injury, has necessitated the urgent development of methods for conducting clinical trials. Because of the unpredictability of outcome soon after injury, the limited number of patients available, and the difficulties in assessment, the development of trial protocols is challenging. This area is beyond the scope of this article, and readers are referred to the report by Steeves, Fawcett, and Tuszynski of the first international meeting on clinical trial design in SCI that was held in 2004 in Vancouver, Canada.

Future Prospects

For many years there have been progressive advances in the basic science of axon regeneration and plasticity. These have now led to several treatments that have promoted functional recovery in animal models of SCI, and the first of these treatments have now entered phase I clinical trials. Over the coming years more treatments will reach this point, and information on efficacy will come from phase II and phase III trials, the methodology for which is currently under development. It is probable that some of these treatments will show efficacy and will become part of the standard treatment for the condition. A condition of the complexity of SCI will require more than one intervention for optimum treatment, and it will be necessary to find out how best to combine therapies. These efforts will be preceded by animal experiments on combination treatments. Alongside this will go advances in physical therapy, probably aided by treatments to promote plasticity, and the development of new prostheses.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Axonal Injury: Neuronal Responses; Cell Replacement Therapy: Mechanisms of Functional Recovery; Nogo-A: Its Role in Axon Regeneration; Olfactory Ensheathing Cells: Regeneration Following Spinal Cord Injury; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Injuries; Spinal Cord Pain Systems; Spine Plasticity.

Further Reading

Anderson KD (2004) Targeting recovery: Priorities of the spinal cord-injured population. *Journal of Neurotrauma* 21: 1371–1383.

Berardi N, Pizzorusso T, and Maffei L (2004) Extracellular matrix and visual cortical plasticity: Freeing the synapse. *Neuron* 44: 905–908.

Bracken MB, Shepard MJ, Collins WF Jr., et al. (1990) A randomized, controlled trial of methylprednisolone or naloxone in the treatment of acute spinal-cord injury. Results of the Second National Acute Spinal Cord Injury Study. *New England Journal of Medicine* 322: 1405–1411.

Carulli D, Laabs T, Geller HM, et al. (2005) Chondroitin sulfate proteoglycans in neural development and regeneration. *Current Opinion in Neurobiology* 15: 116–120.

Fawcett JW (2006) Overcoming inhibition in the damaged spinal cord. *Journal of Neurotrauma* 23: 371–383.

He Z and Koprivica V (2004) The Nogo signaling pathway for regeneration block. *Annual Review of Neuroscience* 27: 341–368.

Hensch TK (2005) Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience* 6: 877–888.

Markus A, Patel TD, and Snider WD (2002) Neurotrophic factors and axonal growth. *Current Opinion in Neurobiology* 12: 523–531.

McKerracher L and David S (2004) Easing the brakes on spinal cord repair. *Nature Medicine* 10: 1052–1053.

Pearse DD and Bunge MB (2006) Designing cell- and gene-based regeneration strategies to repair the injured spinal cord. *Journal of Neurotrauma* 23: 437–452.

Raisman G (2001) Olfactory ensheathing cells – another miracle cure for spinal cord injury? *Nature Reviews Neuroscience* 2: 369–375.

Schwab ME (2002) Repairing the injured spinal cord. *Science* 295: 1029–1031.

Schwab ME (2002) Increasing plasticity and functional recovery of the lesioned spinal cord. *Progress in Brain Research* 137: 351–359.

Steeves JD, Fawcett JW, and Tuszynski M (2004) Report of International Clinical Trials Workshop on Spinal Injury, Feb. 20–21 2004, Vancouver Canada. *Spinal Cord* 42: 591–597.

Spinal Motor Neurons: Properties

H Hultborn, University of Copenhagen, Copenhagen, Denmark
B Fedirchuk, University of Manitoba, Winnipeg, MB, Canada

© 2009 Elsevier Ltd. All rights reserved.

The Motor Neuron

Spinal motor neurons are located in the ventral horn along the whole spinal cord, as well as in cranial motor nuclei. The large alpha (α-)motor neurons innervate skeletal muscle fibers (extrafusal fibers), whereas the smaller gamma (γ-)motor neurons (around 30% of the total number of motor neurons) innervate the intrafusal muscle fibers of the muscle spindles. Several of the large motor neurons innervating the extrafusal muscle fibers have now been shown to also innervate intrafusal muscle fibers. They are often referred to as β-motor neurons, although their properties are otherwise identical to the α-motor neurons. The dendritic tree of the α-motor neurons may extend for several millimeters, as can be visualized following the injection of markers through a recording microelectrode and subsequent reconstruction. The diameter of the soma region is only 50–100 μm. The spinal segments controlling the musculature of the limbs are thicker because of the many motor neurons and interneurons, as well as the extensive sensory information entering the segments from the limbs. The motor neurons innervating a given muscle are clustered in a motor nucleus, which can be seen as a specific area in a transverse section and as a cigar-shaped column along the rostro-caudal axis.

The Motor Unit

In the beginning of the twentieth century, Sherrington defined the motor unit as the "motorneurone axon and its adjunct muscle-fibres" and also described the idea of an orderly recruitment of such motor units in the control of a graded muscle contraction. The motor unit is, thus, composed of a motor neuron and the innervated muscle, as sketched in Figure 1(a). Each muscle cell is innervated by only a single motor axon in the adult animal, but each motor neuron innervates from as few as ten up to about 1000 muscle cells, depending on the muscle size and whether it is used in fine or coarse motor activities. Mammalian muscle cells are divided into slow- and fast-twitch muscle fibers. Indeed, they are usually divided into three classes: S (slow), FR (fast resistant), and FF (fast fatigue). Slow and fast refer to the contraction speed of the innervated muscle units; resistant and fatigable describe how the muscle unit contractility is sustained during prolonged activation. Based on the histochemical profiles and biochemical features of the muscle fibers, these three types are often referred to as slow oxidative (or type I), fast oxidative glycolytic (or type IIA), and fast glycolytic (or type IIB), respectively. All muscle fibers in a muscle unit are of the same type. It has been established that many parameters of the muscle unit depend on the innervation pattern by the motor neuron. Low-frequency discharges during prolonged periods of activity causes the muscle twitch to be slow and the muscle fibers to become resistant to fatigue when tested with long bursts of activity. Conversely, infrequent bursts of high firing frequency causes the motor units to become fast contracting and fatigue sensitive. This has been established by externally applying stimulus patterns to peripheral nerves and has been supported by experiments with cross-reinnervation of fast and slow muscles. It is thus established that motor neurons of different types (slow or fast) strongly influence the innervated muscle fibers with regard to their contractile properties (speed and endurance) and their associated biochemical profile. The degree to which a reinnervating motor neuron can reset the properties of all reinnervated muscle fibers can vary. Therefore, it may be more accurate to speak of an adaptive range of muscle properties which are susceptible to modification by the motor neuron. Some experiments also suggest that the good match between properties of the motor neuron and the reinnervated muscle fibers requires that the muscle fibers talk back to the innervating motor neuron to specify their intrinsic properties, influencing the motor neuron firing pattern. Thus, there is a mutual interdependency between the neuronal and muscular components in the motor unit.

Motor Neuronal Control of Muscle Force: Recruitment and Frequency Modulation

The output of the motor neuron is simple – all-or-none action potentials that are conducted to the end-plate region. Here the synaptic connection, with acetylcholine as the transmitter, is strong enough to ensure that each action potential in the motor neuron is causing one action potential in every connected muscle fiber (thus, the whole muscle unit) and thereby elicit a muscle twitch. The strength of muscle contraction therefore depends on how many motor neurons are activated (summation of muscle unit twitches) and the frequency of discharge (the higher the frequency, the more efficient the summation of the twitches within individual muscle units;

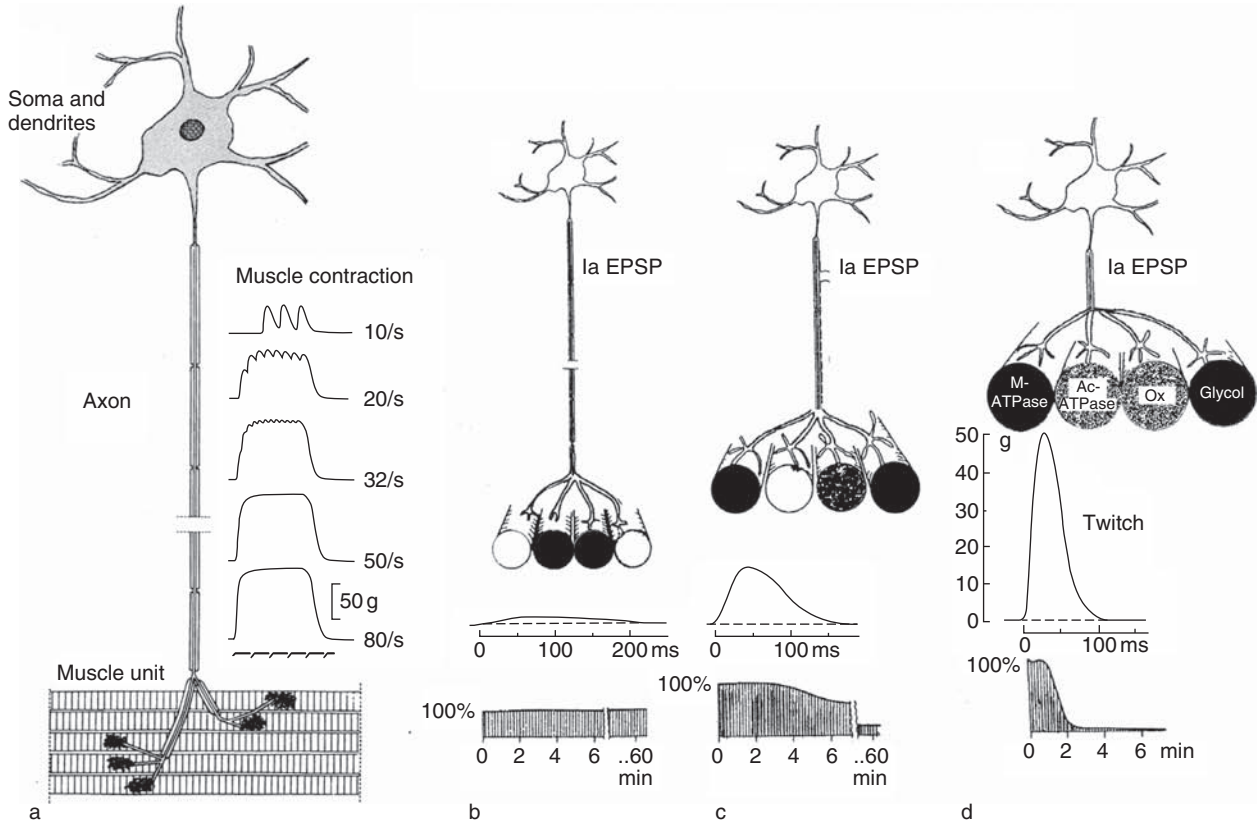

Figure 1 Drawings of the motor unit properties: (a) the motor neuron and its innervated muscle fibers (inset shows the summation of tension as the activation frequency increases); (b) motor unit properties for slow units; (c) motor unit properties for fast resistant units; (d) motor unit properties for fast fatigue units. In (b)–(d) the upper panels show that the size of the motor neurons is larger for fast than for slow units; the axons also have correspondingly a larger diameter for fast than for slow units. The profile of various enzymes is different for slow and fast fibers, as indicated by intensity of immunohistochemical staining of the muscle fibers (black > gray > white) for myofibrillar ATPase histochemistry (with alkaline preincubation, M-ATPase, related to fast contraction; with acid preincubation, Ac-ATPase, slow fibers are strongly stained; fast fatigue fibers are intermediately stained), oxidative enzymes (oxidative capacity, Ox), glycolytic enzymes (anaerobic capacity, Glycol). In (b–d) the middle panels show recordings of single twitches. Note the differences in peak amplitude as well as twitch time course. In (b–d) the lower panels show force maintenance with repetitive activation of the motor unit 0.33 s bursts of 40 Hz once per s for up to 1 h. ATP, adenosine triphosphate; EPSP, excitatory postsynaptic potential.

see Figure 1(a)). The recruitment order of motor neurons depends both on the intrinsic properties of the motor neurons as well as the distribution of the synaptic excitation projecting to them. The frequency of repetitive firing depends on the strength of synaptic excitation and on the intrinsic properties that convert the net excitation to a frequency code. The basic understanding that the muscle force depends on both the number of recruited motor units and their firing frequency goes back to the initial investigations by Adrian and Bronk in the 1920s.

Spike Threshold – Recruitment of the Motor Neuron

If an excitatory postsynaptic potential (EPSP) produced in the motor neuron is small, the movement of K^+ ions through leak channels is sufficient to restore the resting membrane potential; so the EPSP is seen to decay and no action potential is elicited. However, if the EPSP is larger, the K^+ ion leak cannot prevent the activation of voltage-dependent Na^+ channels. The Na^+ entry causes depolarization and activation of neighboring voltage-dependent Na^+ channels and forms a rapid positive feedback system underlying an explosive increase in Na^+ conductance. This transient Na^+ conductance underlies the all-or-none action potential and thereby causes the recruitment of the motor neuron into the motor output. The spike threshold (V_{th}) is therefore the membrane potential at which the inward Na^+ current (from Na^+ channels opened by the depolarization) just outweighs the outward currents (mainly the K^+ leak current) trying to stabilize the membrane potential by bringing it back to the resting membrane potential. If the concentration of Na^+ channels is high, only a smaller percentage need to open to reach the

point of balance between the inward and outward currents; thus, the V_{th} is lower when the number of Na^+ channels is high. Therefore, the part of the motor neuron with the highest density of Na^+ channels has the lowest threshold and is the location for spike initiation under physiological circumstances. Motor neurons have the highest Na^+-channel density within the initial segment (i.e., axon hillock). Therefore, even though the amplitude of the EPSPs are larger at the dendritic level, close to the synapse, the much lower threshold at the initial segment ensures that the spike is initiated at this place independent of the location of the excitatory synapses. When considering integration, this has the strategic advantage that it is a weighted sum of all synaptic input that determines whether the threshold will be passed and cause the recruitment of the motor unit.

It is obvious that the spike threshold is an important variable in deciding when a motor neuron is recruited into a motor task. The threshold depends on several parameters, including the speed at which the depolarization occurs. If the depolarization is slow and ramplike, it causes the Na^+ channels to open, but the opened channels also have time to inactivate. This accommodation causes an increased V_{th} (i.e., threshold is at a more depolarized membrane potential). It has been demonstrated that different motor neurons have different spike thresholds. Thus, the V_{th} is lower for motor neurons supplying slow muscle fibers than for those innervating fast muscle units (Table 1), which facilitates the recruitment of small slow units before large fast ones. For many years, aside from accommodation already described, the spike threshold was considered to be a fixed parameter for a given motor neuron. Therefore, it came as a surprise when it was discovered that V_{th} could be lowered in a state-dependent manner, as illustrated in Figure 2. In this example, V_{th} moved from −46 mV at rest to −55 mV during fictive locomotion elicited by the stimulation of the mesencephalic locomotor center (MLR) *in vivo*. This finding has subsequently been confirmed during motor activity in brain stem-attached *in vitro* preparations. It was determined that the lowering of V_{th} can be induced through a neuromodulatory action of monoamines, most likely accomplished by hyperpolarizing the voltage dependency of fast Na^+ channels.

Table 1 Properties of motor neurons

	S	*FR*	*FF*
Muscle unit properties			
Twitch contraction time	+++	+	+
Fatigue resistance	+++	++	+
Aerobic capacity	+++	++	+
Anaerobic capacity	+	++	++
Isometric force (twitch)	+	++	+++
Motor neuron properties			
Membrane area	+	++	+++
Conduction velocity	+	++	+++
Membrane resistivity, R_M	+++	+	+
Voltage threshold, V_{th}	+	++	+++
Rheobase current	+	++	+++
AHP duration	+++	+	+
Synaptic input			
Monosynaptic Ia EPSP	+++	++	+
Cutaneous and red nucleus EPSP	+	++	++
Cutaneous and red nucleus IPSP	+++	+	+

AHP, afterhyperpolarization; EPSP, excitatory postsynaptic potential; FF, fast fatigue; FR, fast resistant; IPSP, inhibitory postsynaptic potential; S, slow.

Firing Frequency Is Controlled by Summed Synaptic Input and the Postspike Afterhyperpolarization

Following recruitment, the motor neuron can respond to suprathreshold excitation with a train of action potentials. As discussed earlier, the frequency of the resulting firing can encode the strength of the synaptic excitation, and an increased frequency enhances the summation of the force produced by repetitive twitches (Figure 1). The postspike afterhyperpolarization (AHP) has a key function in transducing the summed result of processed synaptic input into a variable spike frequency. In the case of sensory axons, Adrian and Zotterman concluded as early as 1926 that the "... regular rhythm could be explained as the natural consequence of the refractory period and the subsequent return of excitability." It was known even then that stronger excitation of a sensory neuron overcomes its refractoriness and results in a higher spike frequency. In 1929, Adrian and Bronk extended the observations to motor axons and wrote, "... The likeness between the impulse discharge from a sense organ and that from a motor nerve cell suggests that both are due to some property common to axon or dendritic terminations in general." The direct demonstration of refractoriness in motor neurons did not occur until many years later, when it was uncovered by intracellular recording as the postspike AHP by Eccles and co-workers in the early 1950s (Figure 3(a)). They also compared the duration of the AHP in different types of motor units (slow and fast) and suggested that "... since it seems likely that the frequency of repetitive discharge from motor neurons is controlled by the AHP, durations of the AHPs are matched to the twitch durations of muscle." To further advance the understanding of the transduction process in the motor neurons, Granit, Kernell, and co-workers used current injections through the recording microelectrode to mimic synaptic currents (Figures 3(b) and 3(c)).

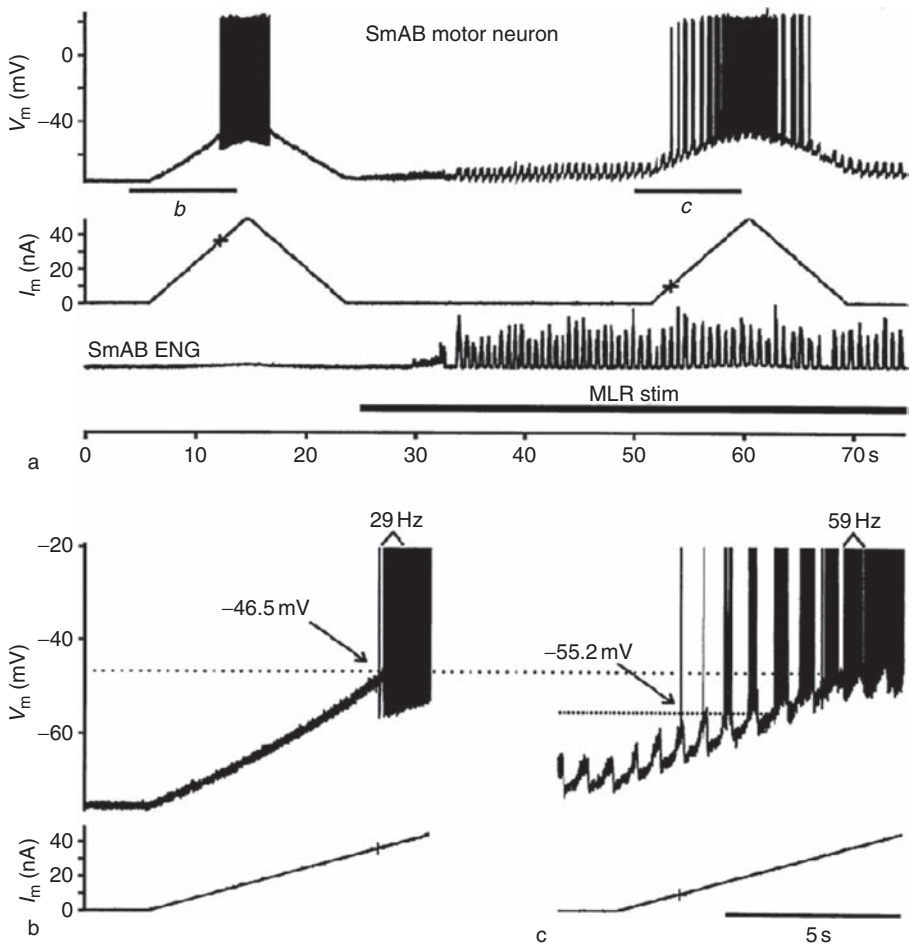

Figure 2 Lowering of voltage threshold for the production of action potentials during locomotion: (a) trial for a hip extensor motor neuron (bars labeled *b* and *c* denote the time periods expanded in panels (b) and (c)); (b) voltage threshold for the production of action potentials (V_{th}) before fictive locomotion; (c) voltage threshold for the production of action potentials (V_{th}) during fictive locomotion. Firing was elicited in lumbar motor neurons by intracellular current injection, prior to and during fictive locomotion evoked by stimulation of the mesencephalic locomotor region. In the trial in (a), a 50 nA ramp of current was injected, after which the brain stem stimulation was started (indicated by the bar under the ENG trace). Fictive locomotion was evident as rhythmic activity that alternated between extensor and flexor electroneurograms (ENGs, flexor not illustrated). Discontinuous current clamp recording allowed accurate measurement of the membrane potential during simultaneous current injection. As shown in (c), during fictive locomotion, less current was required to fire the neuron (compare current at the pluses (+) in (a)) and the V_{th} was hyperpolarized compared to (b). Note that the neuron fired at 29 Hz before locomotion and at 59 Hz at the same membrane potential during locomotion (see bracketed areas). ENG, electroneurogram; MLR stim, mesencephalic locomotor center stimulation. From Krawitz S, Fedirchuk B, Dai Y, Jordan LM, and McCrea DA (2001) State-dependent hyperpolarization of voltage threshold enhances motoneuron excitability during fictive locomotion in the cat. *Journal of Physiology* 532: 271–281.

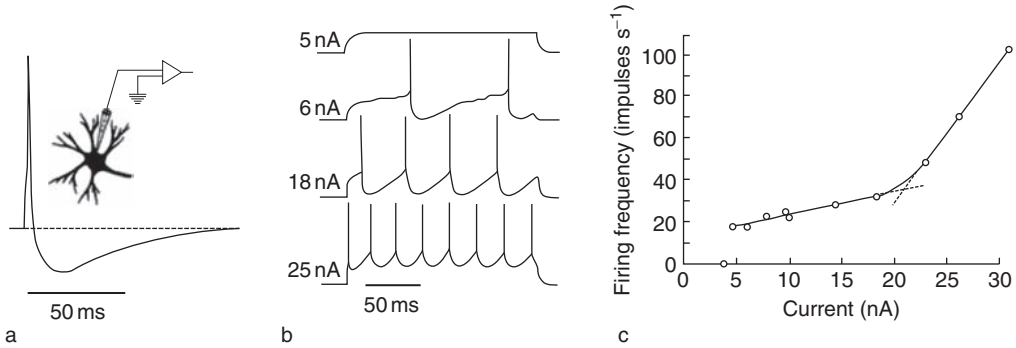

Figure 3 The postspike afterhyperpolarization and relation between injected current and firing frequency: (a) a single action potential with the subsequent afterhyperpolarization; (b) repetitive firing in a motor neuron with intracellular current pulses of increasing amplitude; (c) relation between injected current and firing frequency.

The relation between injected current (I) and the resulting firing frequency (f) typically had a slope of 1–2 imp/s/nA for a relatively long linear segment, which was subsequently named the primary range. In some cases, large current injections reached a level where a steeper linear slope was seen; this was called the secondary range. It was typically seen when the spike frequency reached ~50 imp s^{-1}.

The AHP is caused by a K^+ current ($I_{K,Ca(SK)}$) that is elicited by Ca^{2+} entry during the action potential. The time course and properties of the AHP seen after individual action potentials can explain much of the I–f relation, especially for the primary range. Accordingly, when the AHP amplitude is reduced by pharmacological means or by applying neuromodulators, the f/I slope becomes steeper. With the stronger depolarizing currents needed to reach the secondary range, persistent inward currents are activated. They effectively boost the current injected through the recording microelectrode and therefore contribute to the steeper f/I slope of the secondary range.

The minimum rate for a repetitive discharge is lower for slow motor units than for fast ones, and this is directly related to the longer duration of the AHP in slow motor neurons. There is no real difference in the f/I slopes between slow and fast motor neurons, although the rheobase is much lower for the slow units which are recruited before the fast units. With long-lasting constant current pulses, there is an evident adaptation of the firing frequency; that is, the frequency is highest in the beginning (especially for the first interspike interval) and gradually decreases during the train. The explanation for this early adaptation is mainly an enhancement of AHPs which delays following spikes. However, a progressive increase in the spike threshold (V_{th}) caused by partial Na^+ conductance inactivation may also contribute to the adaptation. Following longer-lasting repetitive discharges, a late adaptation is seen. There is evidence that an electrogenic Na^+–K^+ pump causing a hyperpolarization may contribute to this late adaptation.

The Influence of Dendrites on Synaptic Excitation and Inhibition

The dendritic tree of an α-motor neuron may extend for several millimeters, whereas the diameter of the soma region is only 50–100 μm (Figure 4(a)). The number of synaptic boutons is estimated to around 100 000, probably originating from around 10 000–20 000 presynaptic neurons. The dendritic membrane area accounts for around 97% of the total area, and the dendrites receive synapses over their entire length. Thus, the dendrites have large effects on the electrical properties of the motor neuron and influence the synaptic potentials as they spread from the dendritic synaptic sites to the initial segment where the action potentials are elicited (see the section titled 'Spike threshold – recruitment of the motor neuron'). The influence of the dendrites was first addressed by Rall in the 1950s. He found that the change in membrane potential resulting from a current pulse injected into the soma (through the recording microelectrode) did not follow a single exponential, as would be expected by a simple model of the membrane having a resistance and capacitance in parallel (an RC network). He demonstrated that the faster transients could be explained if the cable properties of dendrites were taken into account. He subsequently studied the effects of dendritic cable properties on synaptic potentials and synaptic integration. As illustrated in

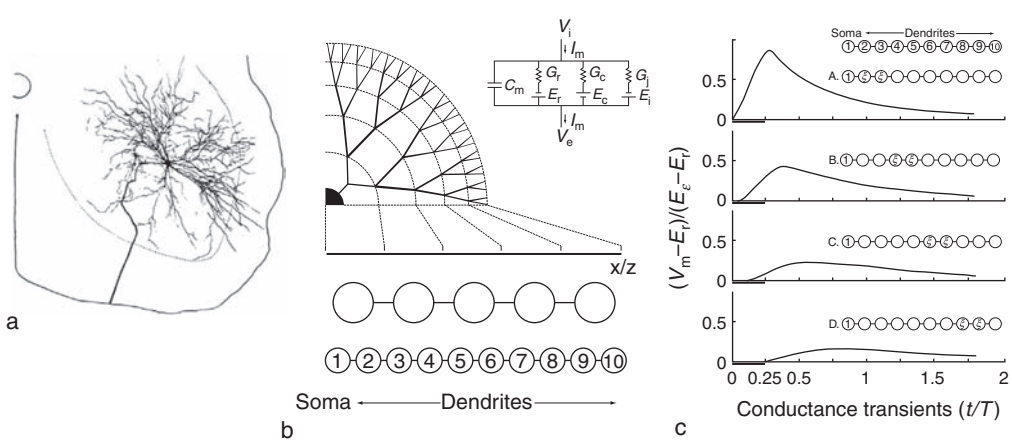

Figure 4 Dendtritic tree and models of nerves: (a) motor neuron visualizing the extensive dendritic tree; (b) diagram of the simple circuit model of a nerve membrane (insert: the RC circuit) and the Rall model of the dendritic tree mathematically transformed into an equivalent cylinder of five or ten equal compartments; (c) simulated EPSP shapes (as seen from the soma) as the same excitatory conductance transients were applied to different dendritic compartments. EPSP, excitatory postsynaptic potential; RC, resistance and capacitance. (b,c) From Rall W (1964) Theoretical significance of dendritic tree for input–output relation. In: Reiss RF (ed.) *Neural Theory and Modeling,* pp. 73–97. Stanford: Stanford University Press.

Figures 4(b) and 4(c), the depolarization seen at the soma was not only smaller but also slower when the same conductance was simulated at increasing distances from the soma. The attenuation may be smaller than expected because the membrane resistivity (R_m) is likely to be much higher at the distal dendrites than at the soma. Furthermore, the attenuation illustrated in Figure 4(c) was calculated for passive dendrites. However, as we describe in the next section, motor neurons contain voltage-dependent noninactivating inward currents that can amplify the summed net depolarization in the dendrites and therefore boost the passive spread from the dendrites to the soma/initial segment.

The Role of Persistent Inward Currents in Amplifying Synaptic Inputs

The classic view of the mammalian motor neurons, which emerged from the laboratories of Eccles and Granit in the 1950s and 1960s, was that the membrane was essentially passive in its areas of synaptic contact (i.e., largely in the dendrites). As a result, the relation between a motor neuron's synaptic excitation and its firing frequency was initially thought to be determined solely by the cell's postspike AHP. When Schwindt and Crill discovered in the late 1970s that motor neurons could display prolonged plateau potentials and self-sustained firing, it certainly opened new horizons in the exploration of motor neuron function. They demonstrated that plateau potentials were due to voltage-dependent, noninactivating Ca^{2+} currents. Compared to steady-state outward currents, the persistent inward currents were large enough to produce a maintained shift in the membrane potential (i.e., the plateau potential). In their experiments, penicillin and tetraethylammonium were used to block outward Cl^- and K^+ currents, respectively, thereby uncovering the inward Ca^{2+} current. A few years later, the Copenhagen group (Hultborn, Hounsgaard, Kiehn, and co-workers) continued the analysis of self-sustained discharge, including its mechanisms and function. The phenomenon, its dependence on physiological transmitters (monoamines), and its cellular mechanisms were unraveled in a series of experiments conducted initially in *in vivo* cat preparations and, subsequently, using an *in vitro* preparation of the turtle spinal cord. In addition, it is now established that a persistent Na^+ current also contributes to plateau potentials in mammalian motor neurons and that not only monoamines but a number of physiological transmitters can facilitate the appearance of plateau properties. Such neuromodulation may be accomplished by reducing outward currents and/or augmenting the persistent inward Ca^{2+} and Na^+ currents. Figure 5 illustrates how the persistent inward current manifests itself in current-clamp mode when activated by current injection through the recording microelectrode. Short depolarizing current pulses (Figure 5(a)) may trigger repetitive firing that outlasts the stimulating current pulse (i.e., self-sustained firing). Alternatively, persistent inward current can cause a sudden acceleration of the firing frequency seen during a linearly increasing ramp of current injection. When the threshold for the persistent inward current is reached (arrows in Figures 5(b) and 5(c)), there is a noticeable jump in the firing frequency. With triangular current pulses sufficient to activate the persistent inward current (Figure 5(b)), the firing frequency remains higher during the downslope of the current injection because the frequency is supported by the additional persistent inward current. This phenomenon is apparent as a counterclockwise hysteresis of the *f*/*I* relation (Figure 5(c)). When action potentials are blocked, the persistent inward current manifests itself as a sustained membrane depolarization (i.e., the plateau potential; Figure 5(d)) and a counterclockwise hysteresis of the *V*/*I* relation produced by a triangular current pulse (Figures 5(e) and 5(f)). Voltage-clamp recording *in vivo* and *in vitro* have greatly enhanced the analysis of the voltage dependency and the dynamics of the persistent inward currents. It has also yielded better evidence for the localization of these currents over the soma-dendritic membrane. Several lines of evidence suggest that the main origin of the persistent inward currents is dendritic. Therefore, dendritic synaptic inputs have a more direct access to the voltage-sensitive noninactivating Ca^{2+} and Na^+ channels underlying the persistent inward current than the current injected through a microelectrode located in the soma. Thus, synaptic current is more effective in turning the persistent inward current on and off than are currents injected experimentally into the soma through the recording microelectrode. It is likely that in normal conditions the persistent inward current is not an all-or-nothing phenomenon but rather might be activated regionally in the dendrites in a graded fashion determined by the local depolarization in the vicinity of the active synapses. That would optimize the conditions in which this persistent current can act as an amplifier for synaptic inputs.

Recruitment Order Is Controlled by Intrinsic Motor Neuron Properties and Synaptic Organization

As was first described by Denny-Brown in the late 1920s, small (slow-twitch) motor units are recruited before large (fast-twitch) ones. In the 1960s, Henneman and collaborators confirmed these observations in a large number of studies and also demonstrated

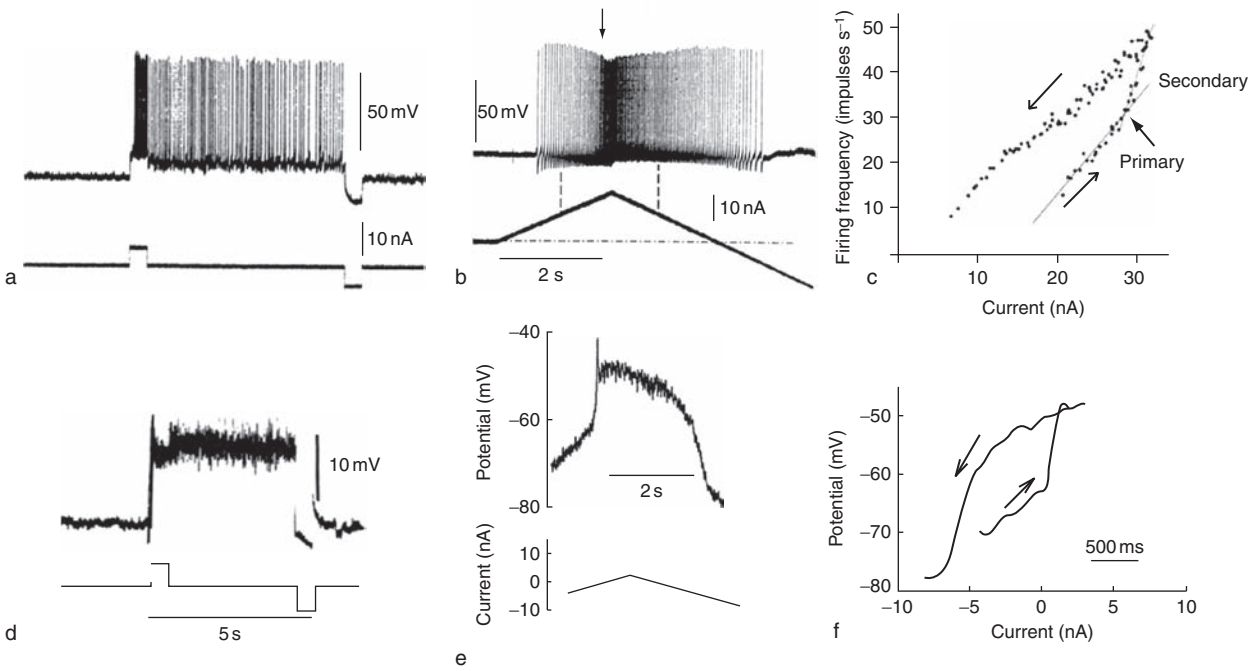

Figure 5 Signs of plateau potentials recorded in spinal motor neurons; sustained shifts in excitability triggered by depolarizing and hyperpolarizing currents injected intracellularly into cat spinal motor neurons in a decerebrate, unanesthetized preparation: (a) sustained repetitive firing in a motor neuron, which was initiated by a short depolarizing current pulse and terminated by a short hyperpolarizing pulse; (b) firing pattern and membrane potentials during injection of triangular current pulses (spike activity in a motor neuron (upper trace) and triangular profile of its injected current (second trace)); (c) graph of (b); (d) sustained repetitive firing in a motor neuron following an injection of QX 314 to block Na^+ channels; (e) firing pattern and membrane potentials during injection of triangular current pulses (showing the membrane potential without spikes); (f) graph of (e). In (a,b) and (d,e), the upper traces are intracellular recordings and the lower traces indicate the amount and timing of the injected current. In (b) the interrupted vertical lines point to the difference in firing frequency at similar current intensities during the ascending and descending phase; the interrupted horizontal line shows the zero level for injected current. In the graph in (c), the thin lines approximate the *f/I* slopes for Kernell's primary and secondary ranges of firing. Note the counterclockwise hysteresis. Modified from Hounsgaard J, Hultborn H, Jespersen B, and Kiehn O (1988) Bistability of alpha-motoneurons in the decerebrate cat and in the acute spinal cat after intravenous 5-hydroxytryptophan. *Journal of Physiology* 405: 345–367; Bennett DJ, Hultborn H, Fedirchuk B, and Gorassini M (1998) Synaptic activation of plateaus in hindlimb motoneurons of decerebrate cats. *Journal of Neurophysiology* 80: 2023–2037.

that derecruitment proceeds in the reverse order. As small (slow-twitch) motor units also have smaller-size motor neurons than large (fast-twitch) ones, they described the small-to-large recruitment order as a consequence of motor neuronal size, now known as the size principle. Inherent in this concept was the idea that somehow the anatomical size of motor neurons reflected their excitability. Indeed, less current is needed to bring the small S motor neurons to discharge (lower rheobase). However, if synaptic density and distribution over the membrane are the same for small and large motor neurons, size *per se* could not explain orderly recruitment. But size seems to correlate with other parameters of direct influence on excitability. For instance, the membrane resistivity, R_M, seems to be higher in type S than in types FR and FF motor neurons (Table 1). In addition, the absolute voltage threshold for the action potential seems to be lower in type S than in types FR and FF motor neurons (see Table 1). These built-in features in the motor neuron pool could account for motor units being recruited in a force-related manner (from small type S to large type FR and, finally, type FF) even under conditions of equal synaptic density and efficacy across the motor neuron types.

These findings do not exclude the possibility that variation in the distribution of synaptic inputs among the different motor neuron types of a given pool play an additional role for the orderly recruitment. Indeed, the evidence favors that the preferential projection from (for example) the monosynaptic Ia excitation from muscle spindles fortifies the recruitment order from type S to FR and FF. Exceptions to the normal recruitment order have been described, for example, a selective recruitment of fast motor units during the extremely fast movements seen with paw shake (activated in the cat when the paw is wetted). Synaptic inputs to motor neurons from skin afferents and the rubrospinal tract produce inhibition in motor neurons supplying slow-twitch motor units

and excitation to those innervating fast-twitch ones. The use of such pathways may explain the reversed recruitment order. There has been much discussion concerning to what extent this deviant projection pattern might change orderly motor-unit recruitment. Such a differential distribution of synaptic inputs not only changes the recruitment order but – perhaps more important – also the ease of graded recruitment (i.e., the recruitment gain in a motor neuronal pool). This argument is further illustrated in Figure 6. A motor neuron may be said to have an intrinsic excitability, here operationally defined as the amount of injected current required to discharge the motor neuron in the absence of background synaptic activity. At any given moment, the ability of a motor neuron to generate an action potential depends on both its intrinsic excitability and the sum of its synaptic inputs. If the distribution of some synaptic inputs deviate (pool bias) from the common drive (pool drive), the pool bias might expand or compress the range of functional thresholds among the motor neurons. This changes the ease with which a uniformly distributed excitatory system (pool drive in Figures 6(a)–6(c)) activates the pool. More precisely, the term recruitment gain has been used to indicate the relation between the intensity of the synaptic input (drive) to a motor neuron pool and the output from that pool (i.e., the number of active cells). It predicts that a biasing input (with the distributions to low- and high-threshold motor units different from that of the drive input) will indeed change this gain (Figure 6(d)). There is some experimental evidence, both in the cat and in humans, supporting this reasoning.

Intrinsic Motor Neuron Properties Are Controlled by Neuromodulators

Specific control of the timing of motor neuronal activity in relation to execution of movement is relying on fast ionotropic receptors at excitatory (mainly glutamate as the transmitter) and inhibitory (mainly glycine and γ-aminobutyric acid (GABA) as transmitters) synapses. However, more long-lasting and general control of motor neuronal excitability in relation to the sleep–wake cycle, exercise level, and affective state is mediated mainly by metabotropic receptors (mainly amines and peptides). These modulatory actions are mainly exerted by acting on intrinsic mechanisms, thereby indirectly changing the overall behavior of the motor neuron. In the preceding sections, we describe a number of functions that depend on specific voltage-dependent ionic channels, which build up the intrinsic response properties of the motor neurons. They include (1) the transient Na^+ channels ($I_{Na,t}$), responsible for the voltage threshold and the depolarizing phase of the action potential; (2) K^+ channels ($I_{K,Ca(SK)}$), causing the postspike AHP and thus controlling the relation between net synaptic excitation and firing frequency; and (3) persistent voltage-dependent noninactivating inward currents (Ca^{2+}, $Ca_v1.3$, and Na^+ channels), underlying the plateau potentials and serving to amplify synaptic inputs. These are only a few examples; there are many more. Here we briefly mention the following:

1. The leak currents I_{leak} (mainly K^+ currents) that stabilize the resting membrane potential at rest.

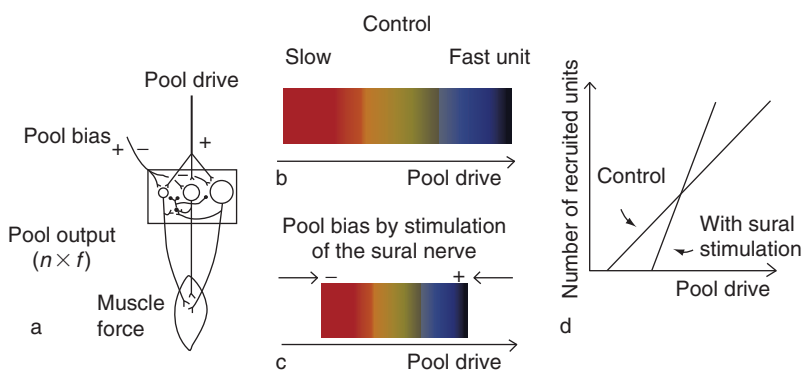

Figure 6 Testing the recruitment gain hypothesis, which concerns changes in recruitment gain produced by conditioning volleys with a non-uniform distribution to the pool (pool bias), which differs from the orderly distribution of the test volley (pool drive): (a) diagram showing inputs (pool drive and pool bias) to a motor neuron pool (with early- and late-recruited motor neurons shown as small and large, respectively); (b) recruiting of the motor neurons by the test volley in the normal orderly manner, from slow (low-threshold) to fast (high-threshold) units; (c) the addition of a conditioning volley (pool bias) that inhibits slow motor units and facilitates fast ones; (d) diagram illustrating the relation between the test volley and the number of recruited motor neurons in the control situation (as in (b)) vs. sural nerve stimulation (i.e., a pool bias-conditioning stimulus, as in (c)). As shown in (c), when a conditioning volley is added, the test volley still recruits motor neurons in the same order but within a smaller range of the test volley. This is seen as a compression of the excitability spectrum. The recruitment gain is increased because a smaller increment of the test volley is needed to recruit successive motor neurons. Modified from Hultborn H, Brownstone R, Toth TI, and Gossard JP (2004) Key mechanisms for setting the input-output gain across the motoneuron pool. *Progress in Brain Research* 143: 77–95; unpublished material by JB Nielsen, H Hultborn, and J-P Gossard.

2. A number of K^+ channels activated by depolarization that contribute to the rapid repolarization of the action potential.
3. A K^+ current (I_{Kir}) activated by hyperpolarization and referred to as an inward rectifier.
4. A mixed Na^+ and K^+ current (I_h) activated by hyperpolarization that produces a slow depolarization of importance for rhythmic behavior (i.e., a pacemaker potential).

All these voltage-dependent ion channels seem to be controlled by neuromodulators. The serotonergic system, in particular, is important for two reasons: much is known about the mechanisms underlying the actions of serotonin at motor neuronal level, and there is an emerging knowledge on the relation between the activity of the serotonergic system and motor behavior.

Serotonin increases motor neuronal excitability by decreasing the leak current, by lowering the spike threshold, by reducing the AHP, and by enhancing plateau properties (i.e., facilitating the persistent inward currents). All the serotonergic projections originate for the raphe nuclei in the brain stem, with the neurons located in the raphe pallidum and obscurus nuclei projecting to the spinal motor nuclei. Chronic recording from the raphe neurons in these nuclei in freely moving animals (cats) showed a strong correlation in their activity to the motor activity level. They have a low discharge rate at sleep, and during rapid eye movement (REM) sleep with muscular atonia they become almost silent. During waking, their frequency increases and correlate well to the level of motor activity (standing < walking < running on a treadmill). During forced running on a treadmill, the frequency started out very high but then slowly decreased as the animal became more tired. At exhaustion (i.e., the cat did not manage to follow the pace), the raphe neurons' firing frequency dropped dramatically until they became virtually silent as the animal stopped running, and their activity recovered during rest (up to 1 h). These observations have triggered discussions on the involvement of the raphe-spinal projection in central fatigue.

Motor Neuronal Plasticity Following Training or Disuse

Do α-motor neurons, like muscle fibers, adapt to long-lasting increases or decreases in activity? There is much evidence to support the idea that motor activity and motor learning indeed involve neuronal plasticity at both the spinal and supraspinal levels of the central nervous system. However, it is only during the last several years that changes in the properties of α-motor neuron themselves in response to altered levels of physical activity have been directly assessed. The present evidence originates from experiments in which rats were subjected to forced endurance training or spontaneous running in exercise wheels to increase the motor activity or were subjected to body-weight support (i.e., suspension of the hindlimbs) for a couple of weeks to reduce the motor activity. Only a few examples of the described changes are mentioned here. With increased motor activity, the slow motor neurons showed more negative resting membrane potential as well as firing threshold. It was proposed that these changes may prevent depolarization, accommodation, and firing adaptation during prolonged activity. During inactivity, the resting membrane potential and firing threshold became more depolarized. The motor neurons become less excitable and more prone to accommodation and frequency adaptation. The signals mediating the adaptation of these (and other) motor neuron properties are not known at this time. However, it has been proposed that increased levels of neurotrophins originating from the exercised muscle, or other neurotrophic factors released by afferent activity, may induce appropriate changes in motor neuronal gene expression.

Motor Neuronal Plasticity Following Chronic Spinal Lesions

Chronic spinal transection has also been used as a model of reduced activity, although it obviously involves much more than inactivity. For example, the partial denervation and secondary changes resulting from the degeneration of descending axons affect the environment of the motor neurons. Several studies reported that complete or partial spinal lesion may cause a small but significant reduction in the number of motor neurons caudal to the lesion, as well as morphological changes, such as changes in soma diameter and dendritic size and branching, in the surviving population. Electrophysiological examination of motor neuron properties revealed that the resting membrane potential and voltage threshold became depolarized and that the excitability for repetitive firing was reduced. These changes were reduced by passive exercise of the hindlimbs or transplantation of fetal spinal cord tissue.

As already described, the persistent inward currents underlying the plateau properties need descending monoaminergic facilitation for optimal function. Previous studies in cats showed that, following acute spinal transaction, plateau properties are lost but could be restored by the administration of precursors or agonists of monoamines. It was, therefore, surprising to find that plateau potentials, which were

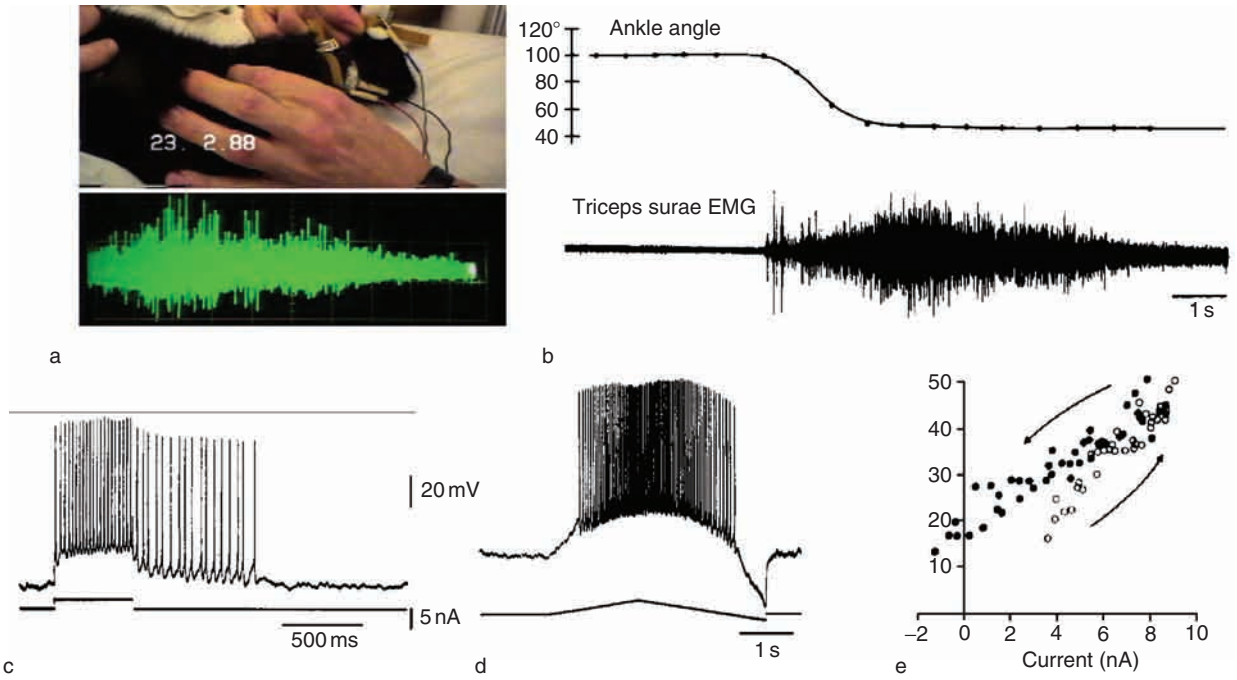

Figure 7 Tonic stretch reflexes and plateau potentials in motor neurons in the chronic spinal cat (recorded in a cat 3 weeks following a complete spinal transection at low thoracic level): (a) tonic stretch reflex in the triceps surae (upper) and raw EMG (recorded with surface electrodes) during a manual flexion of the ankle joint (lower); (b) ankle angle and EMG for (a); (c) responses of an extensor motor neuron to injection of rectangular current pulses; (d) responses of an extensor motor neuron to injection of triangular current pulses; (e) graph of (d). The upper traces in (c) and (d) illustrate intracellular recordings, and the lower traces indicate the injected current. Note in (c) the accelerating firing frequency during the rectangular current pulse and the afterdischarge following the termination of the current pulse. In (e) the instantaneous frequency measured in the record in (d) is plotted against current. The direction of arrows indicated the ascending (○) and descending (●) phase of the triangular waveform. The frequency–current relation shows a counterclockwise hysteresis. EMG, electromyogram. Modified from Eken T, Hultborn H, and Kiehn O (1989) Possible functions of transmitter-controlled plateau potentials in alpha motoneurones. *Progress in Brain Research* 80: 257–267.

lost in the acute spinal state, reappeared 2–3 weeks after spinalization without the addition of monoaminergic precursors or agonists. Based on this observation, it was suggested that reemerging plateau properties following spinal injury may play a role in re-adapting the overall excitability of the spinal network below the lesion and may be a contributor to the pathophysiology of spasticity. These chronic-spinal experiments were originally performed in cats with spinal transection at a low thoracic level (Figure 7). Bennett and colleagues later developed a rat preparation in which a very caudal chronic spinal lesion causes pronounced hyperreflexia of the tail without interfering with the normal hindlimb or bladder function. This caudal part of the spinal cord is also so thin that it can be investigated under *in vitro* conditions. This model has recently been used for studying the cellular mechanisms behind the development of spasticity and hyperreflexia following a chronic lesion.

Summarizing Remarks

In many respects, motor neurons resemble other neurons in the central nervous system. However, their output has to match the properties of the muscle fibers that they innervate and to deliver spike trains that are appropriate to cause the muscle forces necessary for the normal motor behavior. This requires relatively lower firing frequencies than most other central nervous system neurons, and therefore many intrinsic properties of motor neurons are geared to obtain that result. Accordingly, there are systematic differences between motor neurons innervating slow muscle units compared with those innervating fast muscle units. The underlying mechanisms for this match seem to depend on a mutual interdependence between the motor neurons and the muscle fibers they innervate. The recruitment of the motor neuron and the relation between the synaptic input and the firing frequency are set by the spike threshold, the postspike AHP, and the amplification of the net synaptic excitation at dendritic level. All these intrinsic properties of the motor neurons are under the control from a number of neuromodulatory control systems, typically using monoamines and neuropeptides as transmitters. In the case of the serotonergic system, we now know much about the behavioral situations in which it is activated, as well as its mechanisms of

action at motor neuronal level. These neuromodulatory systems govern a short-term, state-dependent plasticity of motor neuron function to match normal variations of environmental demands. Long-term changes in motor neuron properties in response to long-lasting increases of motor activity (exercise) or decreased activity (disuse and spinal lesions) have also been established, but the signals underlying this plasticity are still mainly unknown.

See also: Anterior-Posterior Spinal Cord Patterning of the Motor Pool; Motor Neuron Specification in Vertebrates; Pattern Generation; Reticulospinal System; Sensorimotor Plasticity and Control of Movement Following Spinal Cord Injury +D18.

Further Reading

Brownstone RM (2006) Beginning at the end: Repetitive firing properties in the final common pathway. *Progress in Neurobiology* 78: 156–172.

Burke RE (1981) Motor units: Anatomy, physiology, and functional organization. In: Brooks VB (ed.) *Handbook of Physiology, Sect. 1: The Nervous System, Vol. 3: Motor Control, Pt. 1*, pp. 345–422. Washington, DC: American Physiological Society.

Eccles JC (1957) *The Physiology of Nerve Cells*. Baltimore, MD: Johns Hopkins Press.

Gardiner P, Dai Y, and Heckman CJ (2000) Effects of exercise training on {alpha}-motoneurons. *Journal of Applied Physiology* 101: 1228–1236.

Granit R (1972) *Mechanisms Regulating the Discharge of Motoneurons*. Sherrington lectures XI. Springfield, IL: Liverpool University Press.

Heckmann CJ, Gorassini MA, and Bennett DJ (2005) Persistent inward currents in motoneuron dendrites: Implications for motor output. *Muscle & Nerve* 31: 135–156.

Hultborn H, Brownstone R, Toth TI, and Gossard JP (2004) Key mechanisms for setting the input-output gain across the motoneuron pool. *Progress in Brain Research* 143: 77–95.

Jacobs BL, Martin-Cora FJ, and Fornal CA (2002) Activity of medullary serotonergic neurons in freely moving animals. *Brain Research Brain Research Reviews* 40: 45–52.

Kernell D (2006) *The Motoneuron and Its Muscle Fibres*. Oxford: Oxford University Press.

Krawitz S, Fedirchuk B, Dai Y, Jordan LM, and McCrea DA (2001) State-dependent hyperpolarization of voltage threshold enhances motoneuron excitability during fictive locomotion in the cat. *Journal of Physiology* 532: 271–281.

Powers RK and Binder MD (2001) Input-output functions of mammalian motoneurons. *Reviews of Physiology, Biochemistry and Pharmacology* 143: 137–263.

Rall W (1992) Path to biophysical insights about dendrites and synaptic function. In: Samson F and Adelman G (eds.) *The Neurosciences: Paths of Discovery,* vol. 2, pp. 214–238. Boston, MA: Birkhäuser.

Rekling JC, Funk GD, Bayliss DA, Dong XW, and Feldman JL (2000) Synaptic control of motoneuronal excitability. *Physiological Review* 80: 767–852.

Segev I, Rinzel J, and Shepherd GM (eds.) (1995) *The Theoretical Foundation of Dendritic Function: Selected Papers of Wilfrid Rall*. Cambridge, MA: MIT Press.

Spine Plasticity

W-B Gan, N Kasthuri, and G Yang, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

In 1888, Ramon y Cajal, one of the greatest neuroanatomists, described a series of small protrusions extending from the dendrites of neurons. These protrusions, termed 'dendritic spines,' have since been found on both excitatory and inhibitory neurons throughout the central nervous system of diverse species (Figure 1). Electron microscopic studies indicate that the vast majority of dendritic spines (~98%) are associated with synaptic contacts between neurons. Because spines are excellent indicators of synaptic connections and can be visualized using light microscopy, their structure and function have been extensively studied. It is now well established that dendritic spines are specialized structural, biochemical, and electrical compartments in which neurons communicate with each other. Despite the fundamental role of dendritic spines in the nervous system, key aspects of their development and plasticity in living animals still remain unclear. For example, how do spines form? How long does a spine last in an animal? How does experience of the external world change the development and life time of spines? What are the roles of abnormal spine plasticity in various brain diseases? With the development of new molecular and cell biology tools, neuroscientists have recently begun to glimpse the answers to these questions.

The Development of Spines

Dendritic spines appear early during development soon after dendritic processes are extended from neurons. Time-lapse imaging of spine formation in neuronal cultures, brain slices, and living animals has shown that spines either form directly from dendritic shafts or are transformed from dendritic 'filopodia' – long and thin protrusions without bulbous heads (Figure 2). At early stages of development, synaptic contacts already exist between spines and presynaptic axons, suggesting that the process of spine formation is intimately associated with the process of contact formation between neurons. An important question about spine development is related to the temporal sequence of 'spinogenesis' and 'synaptogenesis': Do neurons make synaptic contact (synaptogenesis) before dendrites produce spines (spinogenesis), or is it the other way around? The answer to this question is important for understanding whether spinogenesis reflects intrinsic properties of postsynaptic dendrites or is governed by extrinsic factors associated with presynaptic axonal terminals.

There are several alternative ways in which spinogenesis and synaptogenesis might occur during development. In the first, dendrites produce long and thin dendritic filopodia that 'sample' some of the potential axonal contacts in nearby regions. Capturing and choosing the appropriate presynaptic axon eventually result in the maturation of the filopodia into the spines. Consistent with this view, dendritic filopodia are abundant in the young developing brain during the period of extensive spinogenesis. Electron microscopic (EM) studies indicate that some filopodia make several synaptic contacts with axons, whereas others do not have any synaptic contacts. Furthermore, time-lapse imaging of filopodia has shown that the vast majority of filopodia undergoes rapid extension and retraction within minutes to hours, consistent with their 'sampling' of the local 'synaptic neighborhood.' Importantly, studies both *in vitro* and *in vivo* have shown that a small percentage of filopodia are transformed into the more stable thin- or mushroom-like anatomy of dendritic spines, providing direct evidence for filopodia as spine precursors. The fact that most filopodia do not result in spines suggests perhaps the highly selective nature of finding appropriate axonal partners.

Although the theory in which dynamic filopodia find their partners and mature into spines seems to describe an important rule of spine development, other modes of spinogenesis may also occur. For example, presynaptic axons can find/recognize the shafts of dendrites and induce the postsynaptic cell to make a dendritic spine directly below that axonal terminal. In support of such a scenario, other EM studies have shown that the formation of shaft synapses often precedes the formation of spine synapses at early developmental stages. Furthermore, time-lapse imaging studies have also found that spines can occasionally form directly from dendritic shafts rather than from filopodia. This latter observation is compatible with an alternative model in which spines first emerge independently of synaptic contact and then axonal terminals find/recognize these pre-formed spines on dendrites to make specific contacts. Also consistent with this model, spinogenesis can occur in the absence of synaptogenesis under certain conditions, such as deafferentation in some parts of the brain. Thus, although there is a strong correlation between the production of postsynaptic dendritic spines and the establishment of connections with presynaptic axonal partners, the actual relationship between these two

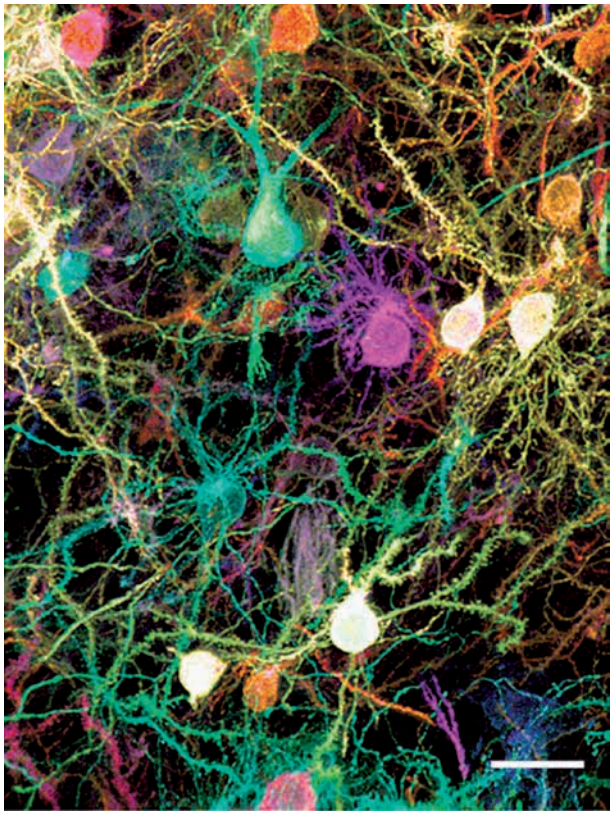

Figure 1 Dendritic spines as the primary contact sites of excitatory synapses are highly abundant in the brain. In this image, a few cortical neurons and their dendrites are labeled in different colors with lipophilic dyes from a mouse cortex at the age of postnatal day 20. A large number of dendritic spines can be seen along individual dendritic branches, demonstrating both the complexity of neuronal connections and the importance of dendritic spines as fundamental structural units in the brain. Scale bar = 25 μm. Adapted from Gan W-B, Grutzendler J, Wong W-T, Wong Rol, and Lichtman JW (2000) multicolor 'DiOlistic' labeling of neuronal circuits using lipophilic dye combinations. *Neuron* 27: 219–225, with permission.

phenomena may vary with different systems and under different conditions.

The definitive answer of when synaptogenesis occurs in relation to when spines develop requires simultaneous monitoring of pre- and postsynaptic cells as they are making contacts and subsequently identifying whether such contacts are synaptic in nature using EM. Although technically challenging, such experiments are now possible. Regardless of how spinogenesis correlates with synaptogenesis, a key aspect of spine development is that most of the newly formed spines are eliminated soon after they are formed, suggesting a highly selective process in the formation of stable spines. As described later, such a selective process goes well beyond the peak of spinogenesis and represents a prominent feature in the maturation of the developing brain.

Substantial Net Loss of Spines during Late Postnatal Development

Soon after Cajal described dendritic spines, he inferred that as the number of spines per unit length of a dendrite decreased with late postnatal development of the nervous system, neuronal circuits were probably rewiring by removing spines and hence the connections between neurons. This prophetic prediction, as many of Cajal's predictions, was subsequently confirmed by more quantitative analyses of the number of synapses and spines in fixed brains from different species at varying developmental ages. In the cerebral cortex of humans and other mammals, rapid synaptogenesis during early postnatal life is followed by a substantial (~50%) loss of synapses/spines that extends through adolescence. In adulthood, the number of spines appears relatively constant until the late stages of life. These studies allowed neuroscientists to infer, like Cajal, that indeed the number of spines seemed to decrease during late postnatal development and stabilize in adulthood.

However, evidence from the analysis of fixed postmortem tissue could not determine the dynamic behavior of spines and the nature of spine loss. For example, because spines are appearing and disappearing, the decrease in spine number could be due to more elimination of existing spines or to slower addition of new spines. The answer to this question would require the confluence of techniques in molecular biology and optical imaging, more than a century after the initial discovery of spine loss. The advent of transgenic mice expressing green fluorescent protein and its spectral variants in a subset of neurons and the use of two-photon microscopy to image spines in the living brains of these transgenic mice allowed neuroscientists for the first time to record the dynamic behavior of spines. It was found that in layer 5 pyramidal cells of different cortical regions, the number of dendritic filopodia is progressively reduced as development proceeds. Consistent with the role of filopodia in spinogenesis, the rate of spine formation is also reduced over the same interval. Importantly, in the first few postnatal months, the rate of spine removal was significantly higher than the rate of spine addition, leading to the net loss of spines observed in diverse regions of the cortex.

These *in vivo* imaging studies confirm Cajal's initial inference: The major rewiring of the cortex during late postnatal life is the elimination of connections between neurons. Because a substantial loss of synapses occurs across different regions of the developing brain in monkeys and humans, it is possible that spines of many cell types and cortical layers undergo a similar pruning process via similar mechanisms. The substantial loss of spines before adulthood correlates with

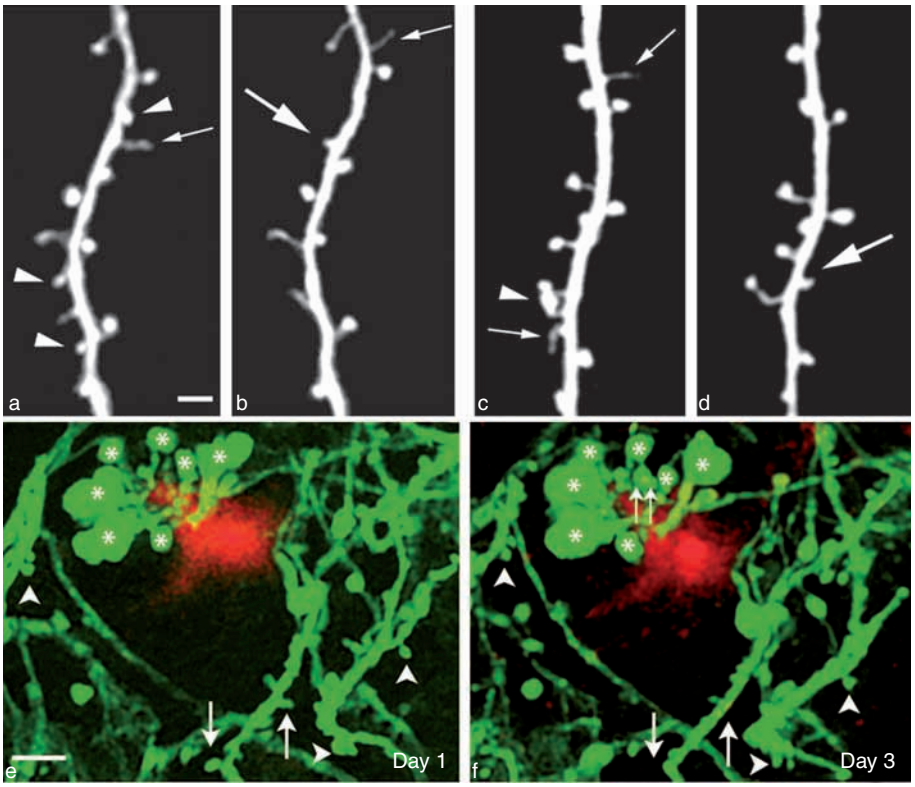

Figure 2 *In vivo* time-lapse imaging of dendritic spine formation and elimination in the mouse cortex. (a–d) Repeated imaging of two dendritic branches from 4 to 6 weeks of age reveals spine elimination (arrowheads) and formation (large arrows) as well as filopodium turnover (small arrows) in control (a and b) and sensory-deprived mouse barrel cortex (c and d). Long-term sensory deprivation via whisker trimming reduces spine elimination in young adolescent mouse barrel cortex from 4 to 6 weeks of age. (e and f) *In vivo* imaging of cortical dendrites and axons (green) near an amyloid deposit (red) in cortical layer 1 of a mouse model of Alzheimer's disease at 6 months of age. Although the majority of spines (arrowheads) and axonal varicosities (asterisks) were stable over 2 days, some structural changes (e.g., spine loss (arrows) and varicosity formation (double arrows)) occurred near the deposits. Scale bars = 2 μm (a–d) and 5 μm (e, f). Adapted from Zuo Y, Yang G, Kwon E, Gan WB (2005) Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* 436: 261–265, and Tsai J, Grutzendler J, Duff K, and Gan W-B (2004) Fibrillar amyloid deposition leads to local synaptic abnormalities and breaking of neuronal branches. *Nature Neuroscience* 7: 1181–1183, with permission.

the development of many cortical functions, suggesting that sculpting early established synaptic connections is a fundamental process in the developing brain.

Spine Stability in Mature Animals

Although many of the initial investigations into spine plasticity focused on development, a time when animals are undergoing remarkable changes in shape and behavior, little was understood about the degree of spine plasticity in adulthood until recently. The adult mammalian brain possesses remarkable plasticity essential for learning and equally remarkable stability for maintaining memory. Although it is generally believed that changes in synaptic strength and number underlie various forms of learning and plasticity in the brain, it is unknown whether or not synaptic connections are stable enough to serve as a structural basis for long-term memory. One way to address this question is to directly observe the degree of stability of synaptic connections in the mature brain. For example, if the majority of adult synapses remains throughout life, this would suggest that memory and basic cortical functions can be stably maintained by synaptic connections that are established during development. On the other hand, if adult synapses are highly dynamic and turn over multiple times within the lifetime of an animal, this would suggest that long-term information might instead be stored in a dynamic manner within highly plastic neuronal circuits.

As described previously, by actually observing the same dendritic spine over time, investigators can measure the dynamic nature of spines in living mouse cortex. By repeatedly imaging the adult mouse cortex over time, it was found that adult spines are remarkably stable on the apical dendrites of layer 5 pyramidal neurons in many regions of the cerebral cortex.

Only 4 or 5% of spines disappear or appear over 1 month in sensory, motor, and frontal cortices. In the somatosensory cortex, >70% of spines are present over a 19-month interval. Assuming that all the spines have a similar turnover rate, the average lifetime of an individual adult spine would be approximately 63 months. Because the average life span of a mouse is approximately 3 years, these results suggest that the majority of adult spines in the cerebral cortex can last throughout the life span of an animal. This cortical-wide spine stability could therefore provide a structural basis for lifelong information storage. Interestingly, whereas the vast majority of spines are stable in adulthood, spine morphology changes both during development and in adulthood. Because spine size correlates with synaptic strength, changes in spine morphology indicate that synaptic strength and, consequently, neuronal function might be readily modified without synapse turnover. Such alterations in the strength of previously established connections could be important for rapid adult plasticity seen in short-term learning. Over an extended period of time, changes in synaptic efficacy may ultimately result in a limited degree of spine turnover and circuit rewiring.

The results mentioned previously mirror studies on the peripheral nervous system in which mouse neuromuscular junctions and submandibular ganglia maintain axonal branching patterns over many months in adulthood. These observations suggest that synaptic stability dominates in the mature nervous system and raise the question of how synaptic connections can be kept largely stable in the presence of lifetime experience. It is important to note that studies so far have examined the dynamics of spines located in cortical layer I; it remains unknown whether the observations of spine stability can be generalized to the entire dendritic tree of layer 5 pyramidal cells or different neuronal types. In addition, whether and to what degree spines can remain stable over the life span in other species such as monkeys and humans require further investigation.

Genetic Control of Spine Formation and Elimination

One of the major goals of research on spine plasticity is to understand how and to what degree spine formation and elimination are regulated by intrinsic genetic programs. By removing or overexpressing individual genes, investigators have identified many genes that are important for regulating the number and size of dendritic spines. These genes encode a wide variety of proteins, including transcription factors, protein kinases, neurotransmitter receptors, adhesion molecules, postsynaptic density proteins, as well as actin cytoskeleton and its regulatory elements. Because spine formation is closely associated with synaptogenesis, genes participating in synapse formation generally have varying effects on the generation of dendritic filopodia and spines. Interestingly, the initial formation of spines can occur in the absence of sensory activity or even more extreme, in the cerebellum, spines of Purkinje neurons can form in the absence axonal terminals. Thus, it appears that some aspects of the ontogenesis of spines are genetically determined and do not rely on synaptic transmission. On the other hand, lack of various subunits of neurotransmitter receptors and ion channels has significant impact on spine density both *in vitro* and *in vivo*. Furthermore, modifying inhibitory neuronal activity either genetically or pharmacologically has profound effects on spine density in excitatory pyramidal neurons. These studies suggest a complicated relationship between gene transcription and neuronal activity in the generation and maintenance of new spines.

One emerging scenario on the genetic control of spine plasticity is that various gene products and patterns of neuronal activity converge to regulate actin polymerization and depolymerization in dendritic spines, leading to spine formation or elimination. Actin filaments form the main cytoskeleton of dendritic spines and are remarkably dynamic. The Rho family of small GTPases, including RhoA, RhoB, and Rac, regulates dynamics of the actin cytoskeleton and is an important contributor to spine formation and elimination. Overexpression or suppression of these molecules as well as their interacting proteins results in changes in the density of dendritic filopodia and spines *in vivo* and *in vitro*. In addition, neuronal activity can induce extensive remodeling of the actin cytoskeleton via calcium influx through glutamate receptors of the *N*-methyl-D-aspartate (NMDA) and the α-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA) subtypes that are highly concentrated at the postsynaptic spines. Many lines of evidence have shown that NMDA and AMPA receptor activation is essential for spine formation and/or maintenance, and that the effect of these receptors on actin dynamics is at least partly mediated by the calcium/calmodulin-dependent protein kinase II pathway.

Much progress has been made in the past few decades regarding the gene regulation of spine development and plasticity. However, most of the studies so far have focused on the role of individual genes in regulating spine density from a single time point

observation. Due to the dynamic nature of spinogenesis, it is important to combine molecular and time-lapse imaging approaches to determine the role of different genes and signaling pathways in various phases of spine formation and maintenance. Furthermore, gene expression is dynamically regulated *in vivo* and lack of a specific gene is likely to alter not only the function of a particular gene *per se* but also patterns of neuron activity in general, which in turn impact on spine plasticity. Thus, in living animals, better temporal and spatial control of gene expression, in combination with improved *in vivo* imaging and electrophysiology, is necessary to decipher the gene regulation of spine plasticity in the future.

Experience-Dependent Regulation of Spine Plasticity

A fundamental aspect of the nervous system is that the patterns of neuronal connectivity and the animal's behavior are profoundly influenced by experience throughout life. For example, long-term sensory deprivation starting at birth and continuing until young adulthood often reduces the number of synapses, whereas enriched environments increase the extent of dendritic branches and synapse number in sensory cortices. Furthermore, in cultured brain slices, induction of long-term potentiation, a synaptic model for learning and memory, can cause rapid outgrowth of dendritic protrusions as well as the formation of multiple synapses over hours. These results suggest that sensory experience and/or neuronal activity often leads to an increase in spine and synapse formation at early stages of development. Time-lapse imaging has been used to examine the long-term effect of sensory experience on the rates of dendritic spine elimination and formation in mouse barrel cortex, a cortical region in which sensory input can be easily manipulated (Figure 2). The results indicate that during adolescence when the peak of synaptogenesis is over and synapse elimination dominates, sensory deprivation preferentially reduces the rate of spine elimination rather than formation. Furthermore, as spines become stabilized in adulthood, the impact of sensory deprivation on spine elimination is reduced substantially. Therefore, sensory experience appears to have different roles in regulating the number of synapses depending on the animal's developmental stage.

It is well established that anatomical features such as dendrite and axon branching patterns in the visual system of the developing mammals are highly sensitive to changes in the external environment but that this sensitivity decreases into adulthood. The previously described studies indicate that experience-dependent spine plasticity also decreases as an animal's age increases. Specifically, there appears to be a temporal window after birth and before 'adulthood' when spine plasticity progressively diminishes and the malleability of the nervous system decreases. As spines become more stable, the ability of the cortex to rewire and, as a result, the ability of animals to change their behavior secondary to environmental changes become limited. Thus, one important insight from the studies on experience-dependent spine plasticity is that the period of heightened sensitivity to experience, termed the 'critical period' is closely associated with the period of spine plasticity.

Another insight is related to the role of spine pruning in the development of the mammalian brain. It is generally believed that a functional mature nervous system is formed from an initial pool of imprecise synaptic connections by selectively forming some synapses and eliminating others based on experience. Although many lines of evidence have demonstrated the role of experience in promoting spine formation, the dominant feature of spine plasticity during late postnatal development is experience-induced spine loss. These results again resonate with studies of the development of the connections between spinal motor neurons and their muscle targets in the peripheral nervous system. Here, it has been established that initially multiple motor neurons make synaptic contacts with individual muscle cells, and this initial exuberance is pruned by competition mediated by neuronal activity. Thus, experience-dependent sculpting of neuronal connectivity seems vital to the maturation of the brain. The same mechanism may also be important for learning and memory formation throughout life.

It is important to mention that many studies have shown substantial changes in the number of synapses following various learning and injury paradigms even in the mature brain. Soon after brain injury, animals, including humans, show a remarkable ability to relearn behavior, suggesting that the mature brain must retain some capacity to form new synapses and rewire its circuitry. Indeed, work has demonstrated that trauma to the skull in mature animals leads to a substantial increase in the turnover rates of adult spines. Thus, although *in vivo* imaging studies have shown that adult spines can form and retract to a limited degree under normal and sensory-deprived environments, it remains to be determined to what degree spine dynamics can be altered under enriched environment and learning conditions.

Dendritic Spine Pathology in Brain Disorders

A number of psychiatric and neurological diseases are associated with alterations in spine morphology or density. Because spine morphology and number are intimately linked to neuronal function, altered spine structures in disease states are likely to have diverse detrimental effects. For example, higher spine density and more immature, long, and thin dendritic spines have been found in many areas of the brain in fragile X syndrome (FXS), the most frequent form of inherited mental retardation. Patients with Down's syndrome have decreased spine density in the neocortex and hippocampus. Schizophrenic patients also exhibit decreased spine density in neocortical pyramidal neurons. Important progress has been made in the past few decades in identifying the genes responsible for psychiatric diseases and developing mouse models for mechanistic studies. In the case of FXS, it has been shown that the disease can be caused by a mutation of the *fragile mental retardation 1* (*Fmr1*) gene located on the X chromosome. The *Fmr1* gene encodes a protein, the fragile X mental retardation protein (FMRP), which binds to many mRNA ligands and acts as a translational suppressor in different subcellular locations, including dendrites and dendritic spines. Similar alterations in dendritic spine morphology and density, as well as behavioral deficits, have been found in *Fmr1* knockout mice and human patients, suggesting that this mouse model is excellent for studying mechanisms underlying the development of FXS. Emerging evidence suggests that many psychiatric diseases are fundamentally disorders of early development that become progressively worse and eventually manifest in early adulthood. Because of the heightened spine plasticity early in life, it is conceivable that the rates of spine formation and elimination in young developing brains are more sensitive to alterations due to mutations of genes related to psychiatric diseases. Even slight changes in spine plasticity may accumulate over time and lead to a substantial increase or decease in spine density as observed in many psychiatric disorders.

In addition to many hereditary mental illnesses, abnormal spine morphology and number have been found in other brain disorders, such as addiction, anxiety, and depression, that are often linked to environmental factors such as malnutrition, abnormal hormone level, and chronic drug abuse. For example, in animal models, repeated exposure to psychomotor stimulants such as amphetamine and cocaine causes an increase in both the dendritic branching and spine density on apical dendrites of pyramidal cells in the prefrontal cortex as well as on medium spiny neurons of the nucleus accumbens. Such an increase in spine density indicates that synaptic reorganization occurs in these brain regions upon repeated drug exposure. The significant and persistent drug-induced synaptic changes may play an essential role in the development of compulsive patterns of drug-seeking behavior.

Structural changes of dendritic spines also likely play an important role in the pathogenesis of many age-related neurodegenerative diseases. This is especially true in Alzheimer's disease (AD), in which the best correlate of cognitive dysfunction is thought to be the loss of synapses as seen by immunohistochemistry and EM in postmortem tissue. Amyloid peptide deposition and tangle formation, two of the hallmarks of AD, are commonly associated with dystrophic neurites and loss of synapses. By cross-breeding a transgenic mouse model of AD with mice expressing yellow fluorescent proteins in subsets of neurons, the degree and time course of synaptic disruption associated with fibrillar amyloid deposits have been studied *in vivo* with two-photon microcopy (**Figure 2**). Time-lapse imaging over days to weeks revealed extensive formation and elimination of dendritic spines near the periphery of the amyloid plaques, suggesting that amyloid deposition triggers continuous remodeling of nearby neuronal structures. Furthermore, local dendritic and axonal abnormalities associated with amyloid deposits can eventually result in synapse loss and the breakage of dendrites and axons. These studies indicate that the accumulation of fibrillar amyloid leads to large-scale and permanent disruption of synaptic connections, contributing to cognitive decline and memory loss in AD.

As described previously, neurological disorders and neurodegenerative diseases are frequently associated with aberrant spine structure and plasticity in critical brain regions. Because dendritic spines are the major sites of neuronal connections in the brain, it is perhaps not surprising that changes in spine plasticity would have a significant impact on disease pathogenesis and progression. The important issue is to determine whether change in spine plasticity is merely the consequence or the primary cause of the diseases and how abnormal spine plasticity contributes to the disease progression. Abnormal formation and/or pruning of spines during development may compromise the establishment of a functional brain. Because spines are remarkably stable in adults, such developmental defects may have long-lasting effects later in life. Furthermore, the failure of the nervous system to maintain stability of neuronal connections could also lead to cognitive and behavioral deficits in at least some brain diseases. The answers to such questions will undoubtedly provide important insight into the

onset, progression, and treatment of psychiatric and neurological diseases.

See also: Alzheimer's Disease: Neurodegeneration; Cell Adhesion Molecules at Synapses; Dendrite Development, Synapse Formation and Elimination; Dendritic Spine History; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Developmental Disability and Fragile X Syndrome: Clinical Overview; Fragile X Syndrome; Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin; Glutamate Regulation of Dendritic Spine Form and Function; Postsynaptic Development: Neuronal Molecular Scaffolds; Presynaptic Development and Active Zones; Receptor Trafficking; Retrograde Transsynaptic Influences; Synapse Formation: Competition and the Role of Activity; Synaptic Precursors: Filopodia; Ultrastructural Analysis of Spine Plasticity.

Further Reading

Calabrese B, Wilson MS, and Halpain S (2006) Development and regulation of dendritic spine synapses. *Physiology* 21: 38–47.

Carlisle HJ and Kennedy MB (2005) Spine architecture and synaptic plasticity. *Trends in Neurosciences* 28: 182–187.

Fiala JC, Spacek J, and Harris KM (2002) Dendritic spine pathology: Cause or consequence of neurological disorders? *Brain Research: Brain Research Reviews* 39: 29–54.

Gan W-B, Grutzendler J, Wong WT, Wong Rol, and Lichtman JW (2000) multicolor 'DiOlistic' labeling of neuronal circuits using lipophilie dye combinations. *Neuron* 27: 219–225

Grutzendler J, Kasthuri N, and Gan WB (2002) Long-term dendritic spine stability in the adult cortex. *Nature* 420: 812–816.

Huttenlocher PR (1990) Morphometric study of human cerebral cortex development. *Neuropsychologia* 28: 517–527.

Rakic P, Bourgeois JP, Eckenhoff MF, Zecevic N, and Goldman-Rakic PS (1986) Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science* 232: 232–235.

Shepherd GM (1996) The dendritic spine: A multifunctional integrative unit. *Journal of Neurophysiology* 75: 2197–2210.

Tsai J, Grutzendler J, Duff K, and Gan W-B (2004) Fibrillar amyloid deposition leads to local synaptic abnormalities and breaking of neuronal branches. *Nature* Neuroscience 7: 1181–1183

Xu HT, Pan F, Yang G, and Gan WB (2007) Choice of cranial window type for *in vivo* imaging affects dendritic spine turnover in the cortex. *Nature Neuroscience* 10: 549–551.

Yuste R and Bonhoeffer T (2001) Morphological changes in dendritic spines associated with long-term synaptic plasticity. *Annual Review of Neuroscience* 24: 1071–1089.

Yuste R and Bonhoeffer T (2004) Genesis of dendritic spines: Insights from ultrastructural and imaging studies. *Nature Reviews Neuroscience* 5: 24–34.

Ziv NE and Smith SJ (1996) Evidence for a role of dendritic filopodia in synaptogenesis and spine formation. *Neuron* 17: 91–102.

Zuo Y, Yang G, Kwon E, and Gan WB (2005) Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* 436: 261–265.

Spines and Mental Disorders

N Nadif Kasri and L Van Aelst, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Mental Retardation

Mental retardation (MR), defined by global deficiencies in cognitive abilities, affects approximately 2–3% of children and young adults. MR shows a wide range in the severity of cognitive disability and is commonly divided into intelligent quotient (IQ) ranges between 85 and 70 (borderline), 70 and 55 (mild MR), 55 and 40 (moderate MR), 40 and 25 (severe MR), and below 25 (profound MR). In addition, MR is often associated with deficits in adaptive behavior. These include at least two of the following skill areas: communication, self-care, ability to live independently, social and interpersonal skills, decision making, health, and safety. The underlying causes of MR are extremely heterogeneous and include nongenetic factors such as infectious disease, premature birth, and fetal alcohol syndrome, as well as genetic changes that include chromosomal abnormalities and single-gene mutations.

Among mentally handicapped families, an excess of males over females has long been noted. There is now evidence to support that this male bias is largely due to mutations in X-linked genes. X-linked MR (XLMR) has been divided into syndromic and nonsyndromic forms of MR. In syndromic XLMR (MRXS), the cognitive disability is associated with a fixed constellation of other manifestations, such as body and brain malformations. In contrast, nonsyndromic XLMR (MRX) is characterized by reduced cognitive function without any other clinical features. Notably, the boundaries between MRXS and MRX are not always clear. The subclassification of MRX versus MRXS largely relies on the clinical evaluation of the patients. Furthermore, mutations in the same gene have been found to produce both forms of XLMR. For instance, the Aristaless-related Homeobox (ARX) gene was initially identified as a MRXS gene; however, more recently some ARX mutations have also been identified in families with MRX. In this article, we kept the MRX/MRXS subclassification of genes as described in 'Euro–MRX European mental retardation consortium'(see 'Relevant Websites').

The search for genes associated with XLMR has been very active over the past few years. Approximately 50 genes linked to MRXS and 20 genes linked to MRX have been identified so far (for an update, see 'Relevant Websites' ('Greenwood Genetic Center')). For most of the XLMR genes identified, there is as yet little information available as to how mutations in these genes result in cognitive impairment. Moreover, their corresponding proteins seem to display very different functions. However, recent studies have revealed that several of the MRX gene products share in common that they all impinge on the actin cytoskeleton and affect dendritic spine morphogenesis. The dynamic actin-rich nature of dendritic spines pointed to Rho GTPase family members as central contributors because they are key regulators of actin dynamics and organization. Indeed, mutations in regulators and effectors of the Rho GTPases have been found to underlie various forms of MR. In this article, we first discuss briefly the role of Rho GTPases in spine and synapse formation and subsequently describe in more detail some of the Rho GTPase signaling pathways involved in MRX and MRXS, as well as in autosomal MR.

Mental Retardation, Spines, and Rho GTPases

Approximately three decades ago, Golgi staining of postmortem material of children with MR indicated for the first time that MR is associated with abnormalities in dendrites and dendritic spines, which are the main sites of excitatory synapses. Dendritic spines take on various shapes, including stubby, thin, or mushroom; and significantly a strong correlation between the size of the spine head and synaptic strength has been suggested. Studies on children with MRX revealed a reduction in the number of dendritic spines and a predominance of very long and thin spines at the expense of stubby and mushroomlike spines. Given that the long thin spine morphology resembled immature spines, the term spine dysgenesis was introduced. Subsequent studies have demonstrated that several types of MR and cognitive disorders are associated with abnormalities is spine density and/or morphology. For instance, patients with fragile-X syndrome (FRAXA) have an increased proportion of long tortuous thin spines, whereas patients with schizophrenia or Down syndrome more commonly reveal a reduction in spine size and spine density. These findings suggested that abnormal development of dendritic spines within the brain might contribute to or underlie changes in plasticity and cognitive impairment.

Morphology and rapid motility of dendritic spines are widely believed to be dependent on polymerization and/or depolymerization of actin. A number of studies have demonstrated that molecules that either

directly or indirectly modify the actin cytoskeleton, such as Rho GTPase family members, RhoA, Rac1, and Cdc42, influence the number and/or shape of dendritic spines and, ultimately, synaptic function. Rac1 and Cdc42 promote, whereas RhoA inhibits, the growth and/or stability of dendritic spines. Rac1 and RhoA also differentially regulate two different forms of rapid motility that are often observed in spines, namely head morphing and protrusive motility.

An expanding number of positive and negative regulators of the Rho GTPases, including guanine nucleotide exchange factors (GEFs) and GTPase-activating proteins (GAPs), and downstream effectors have been identified and shown to control Rho GTPases postsynaptically and to mediate their actions on spine morphogenesis. Significantly, mutations in several of these regulators and effectors have been associated with various forms of MR; these include MRX, MRXS, and autosomal MR (Figure 1).

Rho-Linked Proteins Involved in Mental Retardation

Nonsyndromic X-Linked Mental Retardation

Among the 20 identified genes that have been associated with MRX, three code for regulators and an effector of Rho GTPases: *oligophrenin-1* (*OPHN1*); *ARHGEF6*, also referred to as *αPIX* or *Cool-2*; and *PAK3*. How mutations in MRX genes result in defects in neuronal signaling and function *in vivo* remains largely unknown. However, recent studies have started to provide insight into the functions of these Rho-linked proteins.

Oligophrenin-1 *Oligophrenin-1 (OPHN1),* the first identified Rho-linked MRX gene, encodes a Rho-GAP protein that negatively regulates RhoA, Rac1, and Cdc42 *in vitro*, in nonneuronal and neuronal cells. The protein is expressed in multiple tissues, although with highest levels in the brain, where it is

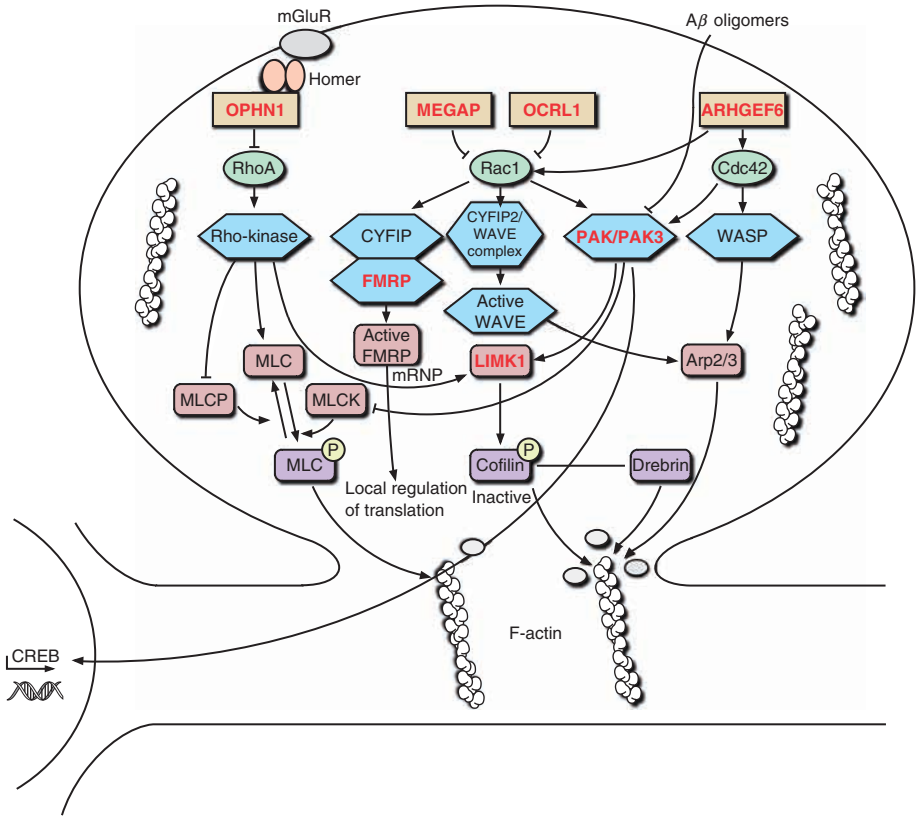

Figure 1 Rho-linked mental retardation proteins and effector pathways connecting Rho GTPases to actin dynamics. The proteins encoded by Rho-linked genes involved in different forms of MR are highlighted in red. Aβ, amyloid β; ARHGEF6, a Rho guanine nucleotide exchange factor 6; Arp2/3, actin-related proteins 2 and 3; CREB, cAMP-responsive element-binding protein; CYFIP, cytoplasmic FMR1 interacting protein; FMR1, fragile-X mental retardation-1; FMRP, fragile-X RNA-binding protein; GAP, GTPase-activating protein; LIMK, Lin-11, Isl-1, and Mec-3 kinase; MEGAP, mental disorder-associated GAP protein; mGluR, metabotropic glutamate receptor; MLC, myosin light chain; MLCK, myosin light chain kinase; MLCP, myosin light chain phosphatase; OCRL1, oculo-cerebrorenal syndrome of Lowe; OPHN1, oligophrenin-1; P, phosphorylation; PAK, p21-activated kinases; RhoA, Rac1, and Cdc42, Rho GTPase family members; WASP, Wiskott–Aldrich syndrome protein; WAVE, WASP family verprolin-homologous protein.

found in neurons of all major regions, including the hippocampus and cortex, and is present in axons, dendrites, and spines. Families with mutations in the *OPHN1* gene were initially described as having a MRX phenotype, but on reevaluation the affected males were found to have also cerebellar abnormalities, in particular, cerebellar hypoplasia. Independent reports describing additional families with mutations in *OPHN1* revealed that these individuals, in addition to cognitive impairment, also suffered from epilepsy and ataxia and showed strabismus, macrocephaly, and hypogenitalism.

All *OPHN1* mutations identified to date appear to result in a loss of protein function. A recent study has provided insight into to how mutations in *OPHN1* may impact neuronal function. By using RNA interference (RNAi) and antisense RNA approaches, it was found that the knockdown of *OPHN1* in CA1 pyramidal neurons in hippocampal slices results in a significant decrease in dendritic spine length. This phenotype was mimicked using a constitutive active (CA) RhoA mutant and was rescued by inhibiting a key effector of RhoA, termed Rho-kinase. These findings support a model in which the loss of OPHN1 causes changes in spine morphology during development as a result of changes in the actin cytoskeleton triggered by the elevation of RhoA and Rho-kinase activities.

Interestingly, OPHN1 has also been reported to associate with Homer, a postsynaptic adaptor protein that links glutamate receptors, such as type-1 metabotropic glutamate receptors (mGluRs), to multiple intracellular targets. Moreover, Homer proteins have been shown to influence dendritic spine morphogenesis and synaptic transmission. Although the function of this interaction remains to be established, it raises the intriguing possibility that OPHN1 acts downstream of glutamatergic receptors to regulate RhoA activity in spines and, in turn, spine morphogenesis. Taken together, this study supports the idea that cognitive impairment observed in *OPHN1* patients can be attributed to aberrations in dendritic spine morphology.

PAK3 A second Rho-linked MRX gene, *PAK3*, encodes a member of the larger family of p21-activating kinases (PAKs). PAK proteins function as Rac1 and Cdc42-specific effector molecules that mediate the actions of these GTPases on cytoskeletal remodeling and gene expression. PAK3 is highly expressed in the brain, particularly in postmitotic neurons of the cerebral cortex and the hippocampus, where it shows a diffuse distribution throughout the soma and proximal dendrites. Mutations in the *PAK3* gene have been identified in several MRX pedigrees. These mutations have been associated with either the loss of PAK3 protein or loss of its kinase activity.

Recent studies have provided evidence for a role of PAK3 in spine morphogenesis and synaptic plasticity. Antisense- and RNAi-mediated suppression of PAK3 levels, or the expression of a dominant negative PAK3 mutant carrying the human MRX30 mutation, in rat hippocampal organotypic slice cultures resulted in the formation of abnormally elongated dendritic spines and filopodia-like protrusions and in a decrease in mature spine synapses. These defects were associated with reduced expression of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPAR) at the synapse and defective long-term potentiation (LTP). In a parallel study, it was demonstrated that PAK regulates spine morphogenesis by triggering the phosphorylation of myosin light chain (MLC). Surprisingly, mice lacking the *PAK3* gene showed no obvious deficits in neuronal structures. However, these mice were selectively impaired in late-phase hippocampal LTP, a distinct form of long-term synaptic plasticity involving new gene expression. The mice also showed a dramatic decrease in the levels of the active form of cyclic adenosine-3′,5′-monophosphate (cAMP)-responsive element-binding protein (CREB) in the hippocampus. CREB is a transcription factor required for activity-dependent gene expression and L-LTP. These findings suggest a novel role for PAK3 in the regulation of CREB function and in synaptic plasticity and cognition. They also support a model in which abnormal CREB function may contribute to the cognitive deficits in MR patients. The differences seen with regard to spine morphology between knockout and RNAi studies may potentially reflect differences between a homogeneous and a heterogeneous cell population, respectively.

Interestingly, defects in PAK signaling not only result in MRX, but recently have also been associated with Alzheimer's disease (AD). This may not be that surprising considering the analogy between AD and MR. Both conditions share in common spine loss or spine alterations. AD is the most common form of degenerative dementia of the human central nervous system. It is defined clinically as a progressive loss of cognitive function with the onset of a slowly progressive impairment of memory during mid-to-late adult life. The neuropathologic hallmarks of this disease include amyloid deposits (Aβ), neurofibrillary tangles, and reductions in the number of neurons and synapses in many areas of the brain, but especially in the cerebral cortex and the hippocampus. Loss of postsynaptic proteins, such as PSD95 and drebrin, are also observed in neurons affected by AD. PAK1 and 3 protein levels and their activities were found to be markedly reduced in brains of AD patients.

The latter was accompanied by a prominent cofilin pathology and loss of the spine actin-regulatory protein drebrin, which cofilin removes from actin. Drebrin is localized at spines in adult brains and is required for active clustering and synaptic targeting of PSD95. Interestingly, the Aβ oligomers implicated in AD were shown to cause PAK signaling defects resulting in a loss of drebrin from the spines. The expression of active PAK could prevent the effects induced by Aβ oligomers, and significantly, pharmacological PAK inhibition in adult mice was sufficient to cause drebrin loss and memory impairment. These findings demonstrate that the loss of PAK3 and/or PAK1 is involved in both developmental-dependent and age-dependent cognitive deficits, such as observed in AD.

ARHGEF6 The *ARHGEF6* gene, also known as *αPIX* or *Cool-2*, codes for a Cdc42/Rac1 exchange factor, which was initially isolated as a PAK-interacting protein. More recent studies have unraveled the complex nature of its GEF activity toward Rac1 and Cdc42. In its dimeric form, αPIX/Cool-2 functions as a Rac1-specific GEF, whereas in its monomeric form, αPIX/Cool-2 can act as a GEF for Rac1 and Cdc42 but only on the binding of PAK or Cbl to its SH3 domain. An intriguing feature of αPIX/Cool-2 is its ability to interact with activated mutant forms of Rac1 and Cdc42, which is unusual for a GEF. Activated Cdc42 was found to enhance the GEF activity of dimeric αPIX/Cool-2 toward Rac1, whereas activated Cdc42 inhibited the GEF activity of monomeric αPIX/Cool-2.

The first mutation in *ARHGEF6* associated with MR was identified in a male carrying a reciprocal X;21 translocation breakpoint located between exons 10 and 11 of the *ARHGEF6* gene. Subsequently, mutations have been identified in the first intron of this gene. These mutations result in preferential skipping of exon 2 and a predicted protein product that lacks the first 28 amino acids in affected males of a large MRX family (MRX46). How mutations in *ARHGEF6* affect neuronal morphology and/or function and how they contribute to MRX remain unknown. However, recent studies examining the closely related family member of αPIX/Cool-2, namely βPix/Cool-1, have provided some clues to the function of ARHGEF6/αPIX/Cool-2 in neurons. It has been shown that the G-protein-coupled receptor kinase-interacting protein 1 (GIT1), an adaptor protein that is enriched at synapses of cultured hippocampal neurons, functions to localize βPIX to dendritic spines and to locally activate Rac1 and its downstream effector, PAK. Mislocalization of GIT1 leads to alterations in spine morphogenesis and synapse formation. Moreover, the GEF activity of βPIX toward Rac1 was found to be necessary for spine morphogenesis and synapse formation. Based on these observations, it was suggested that a signaling complex consisting of βPix, GIT1, Rac1, and PAK plays an essential role in spine and synapse formation.

Interestingly, the truncated ARHGEF6/αPIX/Cool-2 mutant found in MRX46 family members lacks the GIT1-binding domain. It is, therefore, tempting to speculate that this mutant protein is mislocalized and thereby causes aberrations in spine morphology. Consistent with a potential role for PIX family members in the formation and stabilization of synapse-bearing spines, the *Drosophila* homolog, dPIX, has been shown to play a major role in regulating postsynaptic structures and protein localization at the glutamatergic neuromuscular junction. In the future it will be of interest to investigate whether mutations in αPIX/Cool-2 affect spine morphology and postsynaptic signaling and whether this protein acts in a common pathway with PAK3 underlying cognitive function.

Syndromic X-Linked Mental Retardation

Rho-linked genes are also associated with syndromic forms of XLMR. These include FRAXA and Lowe syndrome.

FMR1/CYFIP/Rac FRAXA is the most common monogenic form of inherited MR (estimated prevalence of 1/4000 males and 1/8000 females). The manifestations include (in addition to cognitive deficits) poor eye contact, approach–avoidance behavior, tactile defensiveness, impulsivity, elongated facial features, and machroorchidism. Many patients also manifest attention-deficit hyperactivity disorder and autistic-like behavior. FRAXA is typically caused by an unstable expansion of the CGG trinucleotide repeat and hypermethylation of CpG dinucleotides in the 5′ untranslated region of the fragile-X mental retardation-1 gene (*FMR1*), which results in transcriptional silencing of *FMR1*. *FMR1* encodes an RNA-binding protein (FMRP) that is present at synapses, and evidence has been presented that it acts as a translational regulator. It binds to specific mRNAs through an RNA G-quartet structure.

As mentioned earlier, the shape and density of dendritic spines are altered in patients and in *FMR1*-deficient mice brains. These findings led to the postulation that a defect in maturation and/or function of synapses could be the basis of MR. A major challenge has been to define the mechanism(s) underlying the *FMR1*-associated abnormal spine morphology. A few reports suggested that FMRP could affect spine morphogenesis through the regulation of cargo mRNAs,

such as Map1B and profilin mRNAs. Recent studies, mainly performed in *Drosophila*, have linked FMRP's effect on spine morphology to the Rac1 GTPase signaling pathway. One group demonstrated that the mRNA encoding Rac1 is present in Fmr1-messenger ribonucleoprotein complexes *in vivo*. Moreover, evidence was provided that *Fmr1* and *Rac1* genetically interact and that Rac1 mediates, at least in part, the effects of Fmr1 (*Drosophila* fragile X-related protein) on dendritic branching. An independent study demonstrated a biochemical association between the Fmr1-interacting protein CYFIP and Rac1. The same study also provided data showing that Rac1 negatively regulates CYFIP, which in turn negatively regulates Fmr1, implying that Rac1 indirectly activates Fmr1. Together with the previous findings, these data suggest that there is a feedback loop between Rac1 and Fmr1 functions *in vivo*.

Interestingly, the mammalian homologs of *Drosophila* CYPIP, CYFIP1 and CYPIP2, have also been shown to interact with FMRP. In mammals, CYFIP1 (also known as p140/Sra-1) was initially identified as a target of Rac1, whereas CYFIP2 (also termed PIR121) was found to be part of the Wiskott–Aldrich syndrome protein family Verprolin-homologous protein (WAVE) complex, which mediates actin nucleation by Rac. In its inactive state, this complex contains WAVE and four other proteins: HSPC300, Nap125, Abi2, and PIR121. When active Rac1 is added, the complex dissociates, freeing WAVE and HSPC300, thereby allowing WAVE to activate the actin-related protein 2/3 (Arp2/3) complex to induce actin polymerization. In analogy to the mechanism of WAVE activation, a model was proposed in which CYFIP dissociates from FMRP/Fmr1 on interaction with activated Rac1, allowing released FMRP/Fmr1 to regulate local protein translation.

Taken together, these observations suggest a model in which Rac1, CYFIP, and FMRP act together in a dynamic signaling complex to regulate actin dynamics and control local protein translation, processes that are key to neuronal morphogenesis and connectivity.

OCRL1 Lowe oculocerebrorenal syndrome (OCRL) or Lowe syndrome is a rare X-linked developmental disorder characterized by a pleiotropic phenotype, including congenital cataracts, muscular hypotonia, MR, and renal Fanconi syndrome. The gene responsible for OCRL was identified by positional cloning of X chromosome breakpoints and encodes a protein termed OCRL1. The protein contains a central inositol polyphosphate-5-phosphatase domain and a Rho-GAP homology domain in its C-terminus. OCRL1 was found predominantly in the *trans*-Golgi network and functions as an active phosphatase with phosphatidylinositol 4,5-bisphosphate ($PI(4,5)P_2$) as its preferred substrate. The GAP domain of OCRL1 has been shown to interact with Rac1 and to possess weak GAP activity toward Rac1.

Mutations that cause Lowe syndrome have been mapped exclusively to the *OCRL1* gene. The observed mutations include premature stop codons, frameshift mutations, in-frame deletions, genomic or exon deletions, and missense mutations that yield single amino acid changes. It was observed that *OCRL1*-deficient fibroblasts derived from Lowe syndrome patients have increased $PI(4,5)P_2$ levels, alterations in the actin cytoskeleton, an increased sensitivity to actin depolymerizing agents, and mislocalization of the actin-binding proteins α-actinin and gelsolin. These findings suggest that increased $PI(4,5)P_2$ levels and abnormal cytoskeleton may contribute to the disease process. Whether $PI(4,5)P_2$ affects the cytoskeleton in a Rho GTPase-dependent or -independent manner remains to be established. The potential involvement of the Rho-GAP domain is of particular interest given that several mutations of *OCRL1* in Lowe syndrome patients have been found in the Rho-GAP domain. Surprisingly, *OCRL1* knockout mice do not develop Lowe syndrome. A potential explanation for this observation is that in mice the OCRL1 loss of function is compensated by the phosphatase, Inpp5b, which shares high homology with OCRL1. Notably, Inpp5b is expressed at significantly higher levels in mice than in humans, especially in the tissues affected by Lowe syndrome.

Autosomal Mental Retardation

Mutations in Rho-linked genes that give rise to MR are not only found on the X chromosome, but have also been identified on autosomes. These include the *MEGAP/WRP/srGAP3* gene, which encodes a Rho-GAP, and the *LIMK1* gene, which codes for a downstream effector of Rac1 and RhoA. Mutations in the *MEGAP* gene are associated with $3p^-$ syndrome, whereas the loss of *LIMK1*, along with several other genes, is associated with Williams syndrome.

MEGAP The *MEGAP* gene was identified as a gene that is disrupted in patients suffering from $3p^-$ syndrome. These patients exhibit (in addition to MR) microcephaly, growth failure, heart and renal defects, hypotonia, and facial abnormalities. Notably, the MEGAP (mental-disorder-associated GAP) gene product had previously been identified as a WAVE-associated protein (WRP) and as a ROBO-interacting protein (srGAP3). The MEGAP/WRP/srGAP3 mRNA

transcript is predominantly expressed in fetal and adult brains and is enriched in the neurons of the hippocampus, cortex, and amygdula.

Biochemical studies showed that this protein strongly enhances intrinsic Rac GTPase activity and, to a lesser extent, Cdc42. Together with the observation that MEGAP/WRP/srGAP3 directly binds to WAVE-1, a model was proposed in which MEGAP/WRP/srGAP3 functions in a negative feedback loop that inactivates Rac1 associated with WAVE-1, thereby controlling actin dynamics and spine morphogenesis. Two independent groups generated and characterized WAVE-1 knockout mice. One of the groups reported that WAVE-1 knockout mice exhibit defects in balance and coordination, reduced anxiety, and deficits in learning and memory. Interestingly, these phenotypes are strikingly similar to those observed in $3p^{-}$ syndrome patients. Together, these findings suggest that signaling through MEGAP/WRP/srGAP3 and WAVE-1 to the actin cytoskeleton is important for normal neuronal function.

LIMK1 Williams syndrome (WS; or Williams–Beuren syndrome) is a rare (1/25 000) genetic disorder characterized by mild MR, cardiovascular problems, distinctive craniofacial features, poor visual-motor integration, attention deficits, and, at times, hyperactivity. The genetic basis for WS has been attributed to a heterozygous deletion of approximately 1.5 Mb at chromosome 7q11.23. This deletion encompasses fewer than 20 genes, including *LIMK1*. The loss of *LIMK1* is believed to be responsible for some of the neurological aspects of WS, particularly the impaired visual-spatial cognition. *LIMK1*, which encodes a neuron-specific LIMK-family member, acts downstream of both Rac1 and RhoA and phosphorylates, and it inactivates the actin depolymerization factor ADF/cofilin, thus controlling actin dynamics.

Studies examining the phenotypes of *LIMK1* knockout mice provided the first insights into how the lack of *LIMK1* may contribute to some symptoms of WS. Mice with a homozygous deletion of *LIMK1* showed abnormal spine morphology: spine head size was dramatically decreased and spine necks were thicker. Moreover, these mice revealed enhanced hippocampal LTP following repeated high-frequency stimulation. Both aberrations in spine morphology and synaptic function were related to perturbations in ADF/cofilin phosphorylation and actin dynamics, thus potentially linking cytoskeletal defects to WS. Significantly, *LIMK1* knockout mice also displayed heightened locomotor activity, impaired spatial learning, and altered fear response. These phenotypes are similar to certain deficits found in human patients with WS.

Interestingly, a brain-specific microRNA, miR-134, has recently been localized to the synaptodendritic compartment of rat hippocampal neurons. This miRNA negatively regulates the size of dendritic spines, and this effect was mediated by miR-134 inhibition of the translation of an mRNA encoding LIMK1. This finding indicates that the level of LIMK1 can be highly regulated in dendritic spines.

Conclusion

The identification and characterization of Rho-linked genes associated with different forms of MR have contributed greatly to our current understanding of how defective cellular signaling can result in abnormal neuronal connectivity, which can give rise to impaired information processing underlying cognitive function. Aberrations in Rho-mediated signaling pathways (by, e.g., mutations in genes encoding regulators of Rho GTPases, including OPHN1, ARHGEF6, MEGAP, and OCRL1, or downstream effectors such as PAK3, LIMK1, and FMRP) lead to a deregulation of dendritic spine formation. Given the association between abnormalities in the size, shape, and number of dendritic spines and changes in LTP end LTD, this strongly supports the idea that structural changes of dendritic spines may underlie the cognitive deficits associated with MR. Taken together, Rho GTPase signaling is critical in the regulation of neuronal structure, which may explain why aberrations in Rho GTPase signaling can result in MR.

See also: Actin Cytoskeleton in Growth Cones, Nerve Terminals, and Dendritic Spines; Cytoskeleton in Plasticity; Dendrite Development, Synapse Formation and Elimination; Dendritic Spine History; Developmental Disability and Fragile X Syndrome: Clinical Overview; Fragile X Syndrome; LIM Kinase and Actin Regulation of Spines; Rho GTPases and Spines; Ultrastructural Analysis of Spine Plasticity.

Further Reading

Bear M, Huber KM, and Warren ST (2004) The mGluR theory of fragile X mental retardation. *Trends in Neuroscience* 27: 370–377.

Chelly J and Mandel JL (2001) Monogenic causes of X-linked mental retardation. *Nature Reviews Genetics* 2: 669–680.

Fiala JC, Spacek J, and Harris KM (2002) Dendritic spine pathology: Cause or consequence of neurological disorders? *Brain Research Reviews* 39: 29–54.

Frints SG, Froyen G, Marynen P, and Fryns JP (2002) X-linked mental retardation: Vanishing boundaries between non-specific (MRX) and syndromic (MRXS) forms. *Clinical Genetics* 62: 423–432.

Govek EE, Newey SE, and Van Aelst L (2005) The role of the Rho GTPases in neuronal development. *Genes & Development* 19: 1–49.

Lowe M (2005) Structure and function of the Lowe syndrome protein OCRL1. *Traffic* 6: 711–719.
Newey SE, Velamoor V, Govek EE, and Van Aelst L (2005) Rho GTPases, dendritic structure, and mental retardation. *Journal of Neurobiology* 64: 58–74.
Soderling SH and Scott JD (2006) WAVE signalling: From biochemistry to biology. *Biochemical Society Transactions* 34: 73–76.
van Galen EJ and Ramakers GJ (2005) Rho proteins, mental retardation and the neurobiological basis of intelligence. *Progress in Brain Research* 147: 295–317.
Zarnescu DC, Shan G, Warren ST, and Jin P (2005) Come FLY with us: Toward understanding fragile X syndrome. *Genes, Brain and Behavior* 4: 385–392.

Relevant Websites

http://www.euromrx.com – Euro-MRX European Mental Retardation Consortium.
http://www.ggc.org – Greenwood Genetic Center.
http://synapses.mcg.edu – Synapseweb.

Spinocerebellar Atrophy

M Manto and P Jissendi, Université Libre de Bruxelles, Bussels, Belgium

© 2009 Elsevier Ltd. All rights reserved.

Since the discovery of triplet repeat expansions in 1991, more than 13 human disorders sharing an expansion of simple DNA sequences have been identified. These diseases are clinically and genetically heterogeneous, characterized by an elemental feature of repetition of unstable repeats. Autosomal dominant spinocerebellar ataxias (SCAs), fragile X syndrome (*fragile site mental retardation 1* (*FMR1*) gene), and Friedreich ataxia (FA) are among this group of neurological diseases. Patients exhibit various combination of supraspinal and spinal symptoms. The current numbering of SCAs corresponds to the order of gene description (SCA1–SCA8, SCA10–SCA23, SCA25, mutations in the gene encoding fibroblast growth factor (FGF)14, dentatorubral-pallidoluysian atrophy (DRPLA)). The *FMR1* gene is involved in different syndromes, including fragile X syndrome and fragile X-associated tremor/ataxia syndrome (FXTAS). Fragile X syndrome is a childhood disease whose main clinical deficit is mental retardation, whereas FXTAS occurs in adults.

Epidemiology of Triplet Expansion Disorders

The prevalence of SCAs is estimated to be 1–4 per 100 000. However, values can be much higher in some regions of the world due to a founder effect. This is the case for SCA2 in Cuba and SCA3 in the Azores. The SCA10 families originated in Latin America, and DRPLA is observed mainly in Japan. The commonest forms of SCAs are SCA1, SCA2, SCA3, SCA6, and SCA7. SCA3 might be the most prevalent SCA in the world.

FA is the most common form of recessive ataxia. Atypical features are increasingly recognized in individuals with FA. Late-onset Friedreich ataxia (LOFA) and very-late-onset Friedreich ataxia (VLOFA) are characterized by a symptomatic onset after 25 and 40 years of age, respectively. FA is seen in individuals of European, North African, Middle Eastern, and Indian descent, and also in Southeast Asian populations.

Fragile X syndrome is the most common inherited form of mental impairment. Carrier frequencies are approximately 1/800 males and 1/260 females. FXTAS develops in adults over 50 years of age. In men with late-onset cerebellar ataxia, the prevalence of premutation alleles is 13 times greater than expected based on the prevalence in the general population; the estimated prevalence of FXTAS amongst men older than 50 in the general population is about 1/6000. The prevalence of the mutant allele in patients presenting an unexplained gait ataxia starting after 50 years is 1–4%, and the prevalence is approximately 2–3% in patients developing symptoms suggestive of the cerebellar type of multiple system atrophy. Penetrance increases with age, and penetrance among carriers of larger premutation alleles is greater. The age at death is inversely correlated with repeat size.

Anticipation

Genetic anticipation was initially described in myotonic dystrophy, associated with expansions of CTG repeats. The disorder is characterized by earlier onset and increased severity as the mutant gene is transmitted from one generation to the next. Genetic anticipation was subsequently found for Huntington's disease (CAG expansion), FA (GAA expansion), and various autosomal dominant SCAs. The penetrance of mutations increases in successive generations, an observation made initially in fragile X syndrome, in which a progressive intergenerational expansion of CGG repeats located within the 5′ untranslated region (UTR) region of the *FMR1* gene was discovered. This discovery was followed by the demonstration of CAG repeat expansions in spinal and bulbar muscular atrophy (SBMA, also called Kennedy disease), and others findings followed.

Clinical Description

Patients exhibit combinations of a cerebellar syndrome with extracerebellar signs (Table 1). The cerebellar syndrome includes gait ataxia, which is usually the main ataxic deficit; ataxia of stance; dysmetria and/or kinetic tremor in four limbs; oculomotor deficits; and dysarthria. Because the degenerative process is not limited to the cerebellum but extends to the retina, optic nerve, brain stem, basal ganglia, cerebral cortex, spinal cord, and peripheral nervous system, extracerebellar signs are common. It is noticeable that within a genetic subtype the clinical signs may be highly variable. Moreover, phenotypic overlap may be marked. Several clinical scales (such as the International Cooperative Ataxia Rating Scale (ICARS) or Friedreich Ataxia Rating Scale (FARS)) have been developed to quantify clinically the neurological deficits.

Table 1 Clinical spectrum of spinocerebellar ataxias associated with triplet repeats

Pure ataxia					
Pure cerebellar syndrome[a] SCA5*, SCA6*					
Eyes/oculomotor control	*Movement disorders*	Pyramidal signs	*Cortical signs Cognitive impairment behavioral symptoms*	*Seizures*	*Peripheral nervous system*[b]
Cerebellar ataxia plus					
Slow saccades:[c] SCA1, SCA2**, SCA3, SCA7*	Parkinsonism: SCA1, SCA2, SCA3, SCA12, SCA17, SCA21 FXTAS	SCA1, SCA2, SCA3* (spasticity), SCA4, SCA7, SCA8, SCA11,	SCA1, SCA2, SCA3, SCA13, SCA17*, SCA19, SCA21, SCA-FGF14, DRPLA*	SCA10*, SCA17, DRPLA*	SCA1, SCA2, SCA3, SCA4*, SCA6, SCA8, SCA-FGF14, SCA12, SCA17, SCA18*, SCA22, SCA25*
	Dystonia: SCA3, SCA17	SCA12, SCA13, SCA15	FXTAS		FXTAS
Downbeat nystagmus: SCA6*	Tremor: SCA8, SCA12, SCA16 (head + hand), SCA19, SCA20 (palatal)	FA	FA[d] (?)		FA[d]
Ophtalmoplegia: SCA1, SCA2, SCA3	Dyskinesias: FGF14*				
Pigmentary retinopathy: SCA7**	Myoclonus: SCA2, SCA14, SCA17, SCA19, DRPLA Chorea: SCA1, SCA17, DRPLA				

[a]The following SCAs may evolve from a pure cerebellar syndrome into a cerebellar ataxia plus syndrome: SCA8, SCA11, SCA14, SCA15, SCA16, SCA22.
[b]Sensory neuropathy and sensorimotor axonal neuropathy; dying-back axonopathy, and/or neuronopathy.
[c]Maximal saccade velocity influenced by polyglutamine size.
[d]Possible hypertrophic cardiomyopathy and diabetes mellitus.
*indicates clinical signs which are suggestive; ** indicates clinical signs which are highly suggestive and with some specificity in the context of an autosomal dominant progressive cerebellar syndrome. DRPLA, dentatorubral-pallidoluysian atrophy; FA, Friedreich ataxia; FGF, fibroblast growth factor; FXTAS, fragile X tremor ataxia syndrome; SCA, spinocerebellar ataxia.
Adapted from Manto M (2005) The wide spectrum of spinocerebellar ataxias. *Cerebellum* 4: 2–6.

Symptoms of Spinocerebellar Ataxias

Symptoms of SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, SCA13, SCA17, and SCA25 may start in the first decade, whereas ataxia can start after 65 years in SCA6 (Figure 1). As a general rule, there is an inverse correlation between the length of the repeat and the age of onset.

Some correlations between CAG repeat length and phenotypic presentation have been reported. For instance, SCA3 patients may present a syndrome of ataxia plus peripheral neuropathy or a combination of ataxia plus spasticity when the repeats are less than 73 CAG or more than 73 repeats, respectively.

The phenotype of SBMA includes gynecomastia, testicular atrophy, and muscle weakness due to the degeneration of the anterior horns in the spinal cord.

Symptoms of Friedreich Ataxia

Patients with FA exhibit combinations of oculomotor deficits (mainly fixation instability), dysarthria, gait and limb ataxia, loss of vibration sense and impaired joint position sense starting usually in lower limbs, loss of tendon reflexes with spasticity, and Babinski's sign. Spastic paraplegia may be a leading symptom. Preserved reflexes may occur in the Friedreich ataxia with retained reflexes (FARR). Atypical phenotypes are described, especially in compound heterozygosity. Symptoms related to concentric hypertrophic cardiomyopathy and diabetes mellitus may occur. Electrocardiographic repolarization abnormalities are common. Cardiomyopathy is a cause of premature death. Joint deformities are frequent. Approximately 40% of patients show a slight to moderate hearing deficit. Patients with relatively short expansions usually exhibit a clinical course which is more benign. The phenotype in VLOFA is classically milder.

Symptoms of Fragile X-Associated Tremor/Ataxia Syndrome

Fragile X syndrome is the most common cause of inherited mental retardation. The gene, called *FMR1*, codes for the fragile X mental retardation protein (FMRP) protein. Loss of function of *FMR1* is a model for neurodevelopmental and behavioral disorders such as autism (Table 2). Because Fragile X syndrome is an X-linked disorder, the production of FMRP by the normal allele generally results in less severe phenotypes in women. IQ correlates with blood concentrations of FMRP in both sexes. Overexpression of *FMR1* mRNA causes FXTAS in adults older than 50 years. The phenotype of FXTAS combines intention tremor/ataxia and various combinations of parkinsonism, dementia, neuropathy, and autonomic dysfunction. Cognitive symptoms range from mild frontal executive and memory deficits to global dementia. Anxiety, agitation, or depression are frequent psychiatric manifestations. Emotional problems were more common in adult premutation carriers compared to a control group, and their severity might be correlated with the concentrations of *FMR1* mRNA. The age of onset of tremor or ataxia inversely correlates with repeat size.

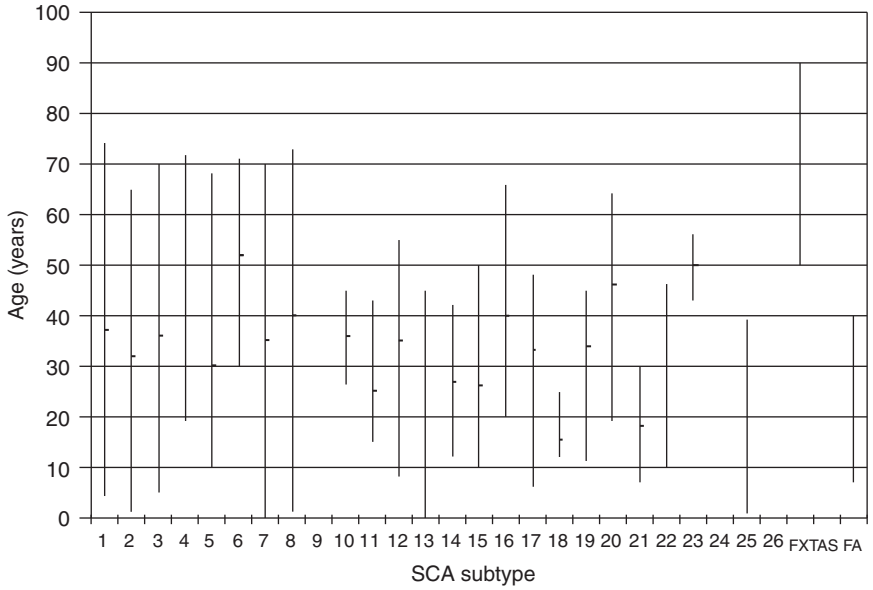

Figure 1 Age of onset of triplet expansion disorders. Dots on vertical bars indicate mean ages of onset. (Not illustrated: dentatorubral-pallidoluysian atrophy (DRPLA), range of 0–62 years (mean = 30); fibroblast growth factor (FGF)14, range of 27–40 years (mean = 34).) FXTAS, fragile X tremor ataxia syndrome; FA, Friedreich ataxia; SCA, spinocerebellar ataxia. Adapted from Manto M (2005) The wide spectrum of spinocerebellar ataxias. *Cerebellum* 4: 2–6.

Table 2 Phenotypes in full mutation or premutation of FMR1

	Symptoms	*Onset*	*Penetrance*
Full mutation			
FXS	Developmental delay	Neonate	M: 100%
	Autism		
	Anxiety		
	ADHD		
	Hypothalamic dysfunction		
	Facial abnormalities		
	Seizures		
	Otitis media		
	Macro-orchidism		
Premutation			
Reproductive symptoms	Primary ovarian failure	Adulthood	F: 20%
	Early menopause		F: 30%
FXTAS	Cognitive deficits	>50 years	M: 33%
	Depression		F: ?
	Cerebellar ataxia		
	Tremor		
	Parkinsonism		
	Peripheral neuropathy		
Developmental disorder	Autism	Childhood	8%
	ADHD		
	Developmental delay		

ADHD, attention deficit hyperactivity disorder; F, female; FXS, fragile X syndrome; FXTAS, fragile X tremor ataxia syndrome; M, male.
Adapted from Jacquemont S, Hagerman R, Hagerman P, and Leehey M (2007) Fragile-X syndrome and fragile X-associated tremor/ataxia syndrome: Two faces of FMR1. *Lancet Neurology* 6: 45–55.

Radiological Findings

Magnetic resonance imaging (MRI) is considered as the imaging of choice in SCAs. However, the scan may be normal when symptoms start. Four patterns of atrophy are found in SCAs: (1) a pattern of pure cerebellar atrophy (in SCA4, SCA5, SCA6, SCA8, SCA9, SCA10, SCA11, SCA14, SCA15, SCA16, SCA18, SCA21, and SCA22); (2) a pattern of olivopontocerebellar atrophy (in SCA1, SCA2, SCA3, SCA7, and SCA13); (3) a pattern of global cerebral atrophy (in SCA12, SCA17, SCA19, and DRPLA); and (4) an association of brain atrophy with spinal atrophy in upper segments (in SCA1, SCA2, and SCA3). Most of the SCAs start by a cerebellar atrophy which can progress subsequently to extracerebellar structures. **Figures 2–4** illustrate typical MRI findings in SCA1, SCA2, and SCA3, respectively. Volumetric studies have shown putaminal and caudate atrophy in SCA3 but not in SCA1 or SCA2. A notable finding is the presence of calcifications in the cerebellar nuclei in SCA20. White matter lesions may be observed, particularly in DRPLA.

Proton magnetic resonance spectroscopy (H^1-MRS) has been used to evaluate the metabolic counterpart of cerebellar or olivopontocerebellar volume loss in SCAs. In the spectrum at short echo time, *N*-acetylaspartate (NAA) decrease is well correlated with neuroaxonal loss or dysfunction and with clinical disability. Figure 3 illustrates the abnormal spectrum in a patient presenting with SCA2.

Positron emission tomography may show a reduced 18-fluoro-2-deoxyglucose (FDG) signal in the posterior fossa, as well as in the basal ganglia and cerebral cortex. In SCA3, increased glucose metabolism in the temporal and parietal cortices of asymptomatic carriers may correspond to compensation. Widespread hypometabolism may be observed in SCA6, contrasting with the clinical findings of an apparently pure cerebellar syndrome from the clinical point of view.

FA patients usually present with spinal cord atrophy (Figure 5). Slight to moderate cerebellar vermian atrophy is not exceptional. The neuroradiological features of FXTAS include prominent white matter disease. The involvement of middle cerebellar peduncles on T2-weighted images is suggestive but not specific (Figure 6). Global brain atrophy is more pronounced in the frontal and parietal regions, pons, and cerebellum. The degree of brain atrophy is correlated with the CGG repeat size.

Neuropsychological Deficits

Cognitive deficits occur at variable degrees. Executive deficits are a common finding in several SCAs. Poor performance on the Wisconsin Cart Sorting Test (WCST) is found in SCA1, SCA2, SCA3, and SCA6. Patients may develop a dementia, especially with SCA17. Depression and personality disorders may occur in up to 65% and 25% of patients, respectively.

Clinical Neurophysiology

Very mild to marked neuropathy may develop in SCAs. Sensory or sensorimotor neuropathy is found in approximately 50% of patients with SCA1, in 80% of patients with SCA2, and in 75% of patients with SCA3. Approximately half the patients with SCA6 exhibit a mild sensorimotor neuropathy.

An early peripheral sensory axonal neuronopathy, characterized by a reduction of the amplitude of sensory action potential (SNAP), is common in FA. Somatosensory-evoked potentials are markedly abnormal in most cases. There is an inverse correlation

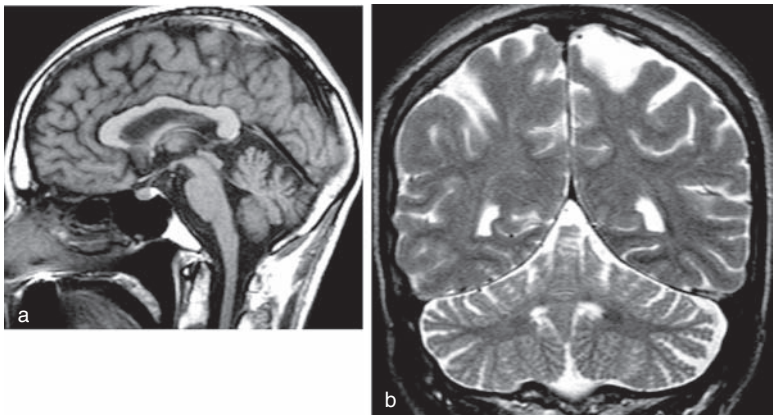

Figure 2 Brain MRI of a patient with SCA1: (a) sagittal T1-weighted image; (b) coronal-T2-weighted image. The images show vermian and cerebellar atrophy characterized by the enlargement of the fissures and the absence of pontine atrophy. Notice the atrophy of the supratentorial structures. MRI, magnetic resonance imaging; SCA, spinocerebellar ataxia.

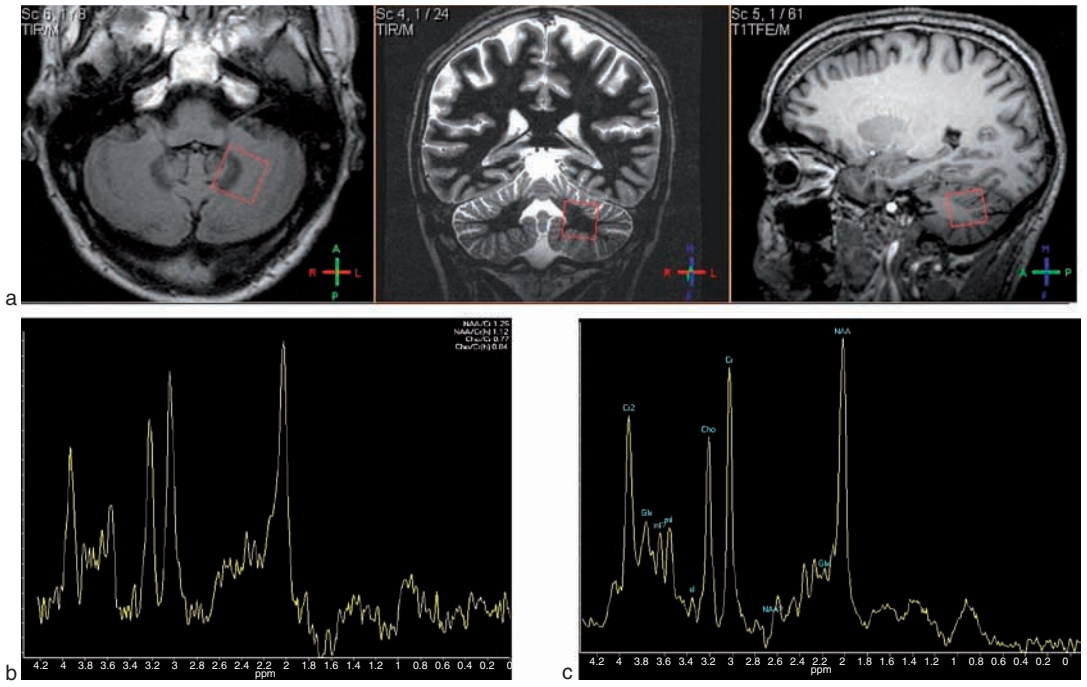

Figure 3 Brain MRI and H^1-MRS of a patient with SCA2: (a) MR images used to plan the single voxel acquisition; (b) spectrum acquired in the deep left cerebellum of patient; (c) spectrum acquired in the deep left cerebellum of age- and sex-matched control. The images in (a), axial FLAIR, coronal T2-weighted, and sagittal T1-weighted (from left to right), show global atrophy in the posterior fossa, as well as in the supratentorium The spectrum in (b) shows a decrease of the NAA/Cr ratio (1.26) compared to the control (1.53). Cho, choline; Cr, creatine; FLAIR, fluid-attenuated inversion recovery; H^1-MRS, proton magnetic resonance spectroscopy; MRI, magnetic resonance imaging; NAA, *N*-acetylaspartate; TIR/M, T inversion recovery-modular; TITFE/M, T1 turbo-fast echo-modular.

between GAA expansion size and the severity of the electrophysiological deficits. Peripheral motor nerve action potential amplitude, and motor conduction velocities are within the normal range in most patients. Central motor conduction times (CMCTs) are increased and worsen with time.

An electroencephalogram (EEG) may show a predominance of slow disorganized activities in hereditary ataxias. Interictal EEG may be abnormal is patients exhibiting seizures such as SCA10.

Diagnosis and Differential Diagnosis

As a result of the marked phenotypic variability of triplet expansion disorders even within single families, the diagnosis cannot rely solely on clinical evaluation;

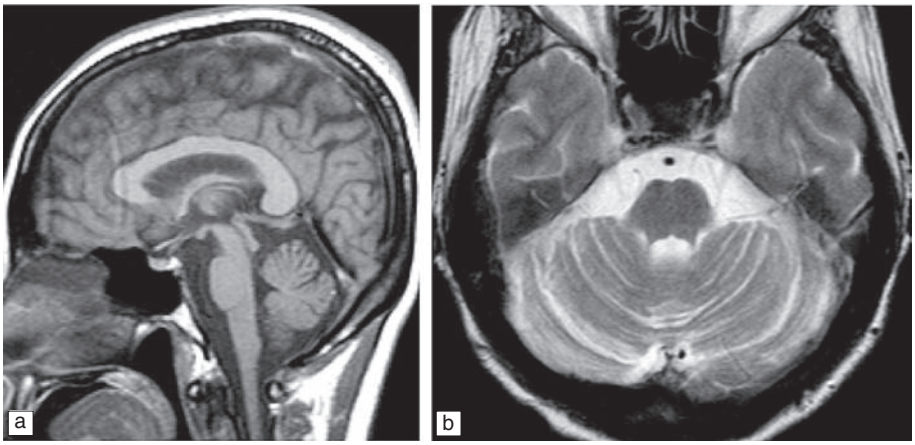

Figure 4 Brain MRI of a patient presenting SCA3: (a) sagittal T1-weighted image; (b) axial T2-weighted image. Images show enlargement of the cisterns in the posterior fossa with global cerebellar atrophy; an involvement of the brain stem is visible at the level of the medulla. MRI, magnetic resonance imaging; SCA, spinocerebellar ataxia.

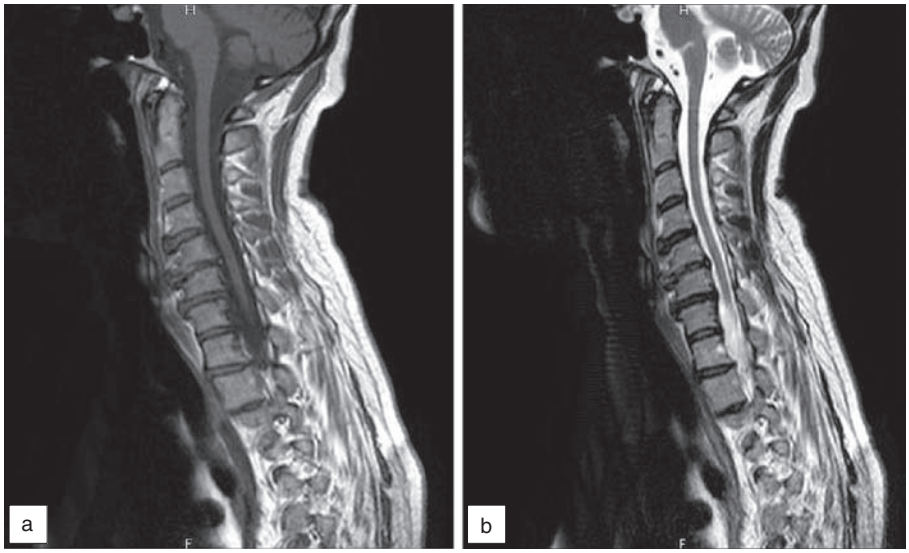

Figure 5 Spinal MRI (cervical segments) of a patient presenting Friedreich ataxia: (a) T2-weighted image; (b) T1-weighted image. Shrinkage of the spinal cord is visible. H, head (cranial direction); F, foot (caudal direction).

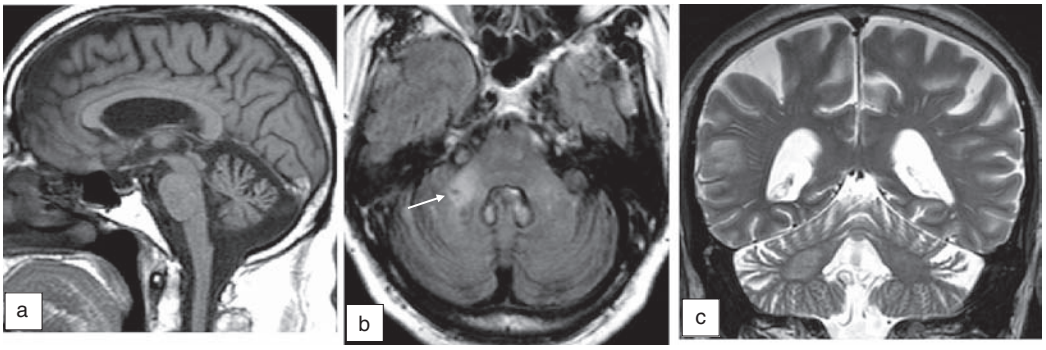

Figure 6 Brain MRI in a patient presenting FXTAS: (a) sagittal T1-weighted image; (b) axial FLAIR image; (c) coronal T2-weighted image. The image in (a) shows the vermian atrophy. The images in (b) and (c) show the cerebellar atrophy and suggestive hypersignal abnormalities in the white matter of middle cerebellar peduncles bilaterally; this is more obvious on the right (arrow). Note the atrophy of cerebral sulci in (c). FXTAS, fragile X tremor ataxia syndrome; FLAIR, fluid-attenuated inversion recovery.

a detailed genetic search is necessary. From the phenotypic point of view, several genetic diseases may overlap with triplet expansion disorders.

The following disorders need to be considered in the differential diagnosis, among others: episodic ataxias (EA), hereditary spastic paraplegias (HSP), Huntington's disease (CAG triplet expansion), essential tremor, hereditary sensory–motor neuropathy (HSMN), mitochondrial diseases (Kearns-Sayre syndrome; myoclonus epilepsy associated with ragged-red fibers (MERRF) syndrome; mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke (MELAS) syndrome; neuropathy, ataxia and retinitis pigmentosa (NARP); Leigh syndrome), gluten ataxia, vitamin E deficiency (AVED), ataxia with ocular apraxia (AOA), leukodystrophies, progressive myoclonic epilepsies, multiple system atrophy, parkinsonism, and cerebrovascular diseases. Family history can remain negative in SCAs with late-onset symptoms, such as SCA6; in case of severe anticipation, as in SCA7; in case of reduced penetrance, as in SCA17; in families with marked phenotypic variations among members; and in case of false paternity.

Pathological Findings

Dominant olivopontocerebellar atrophy is a characteristic feature of SCA1, SCA2, and SCA7. Various degrees of Purkinje cell degeneration, ranging from depleted dendritic trees to total loss, are regularly found (Figure 7). A characteristic feature of SCA with cortical cerebellar atrophy is the formation of empty baskets about the atrophic Purkinje cells. The marked involvement of the Purkinje neurons contrasts with the relative sparing of the stellate/basket cells and Golgi neurons. The granular layer of the cerebellar cortex is often relatively preserved. Grumose degeneration of neurons in dentate nuclei – which can occur also in FA – is frequently found in SCA3. Retrograde atrophy of the inferior olivary nuclei occurs in several SCAs and hypertrophic inferior olivary neurons may be found in DRPLA. Unlike in SCA6, which spares the spinal cord, gray and white matter abnormalities are found in the spinal tracts of SCA2 and SCA3 patients. Nigral lesions are common in SCA2 and SCA3.

Immunohistochemical studies show neuronal inclusions positive for antipolyglutamine antibodies (Figure 8). They are considered a pathological hallmark of SCAs. Inclusion bodies are most frequent in the neurons of the basis pontis in SCA1, SCA2, and SCA3.

In FA, pathological changes occur first in the dorsal root ganglia, with the loss of large sensory neurons, followed by axonal degeneration in the posterior

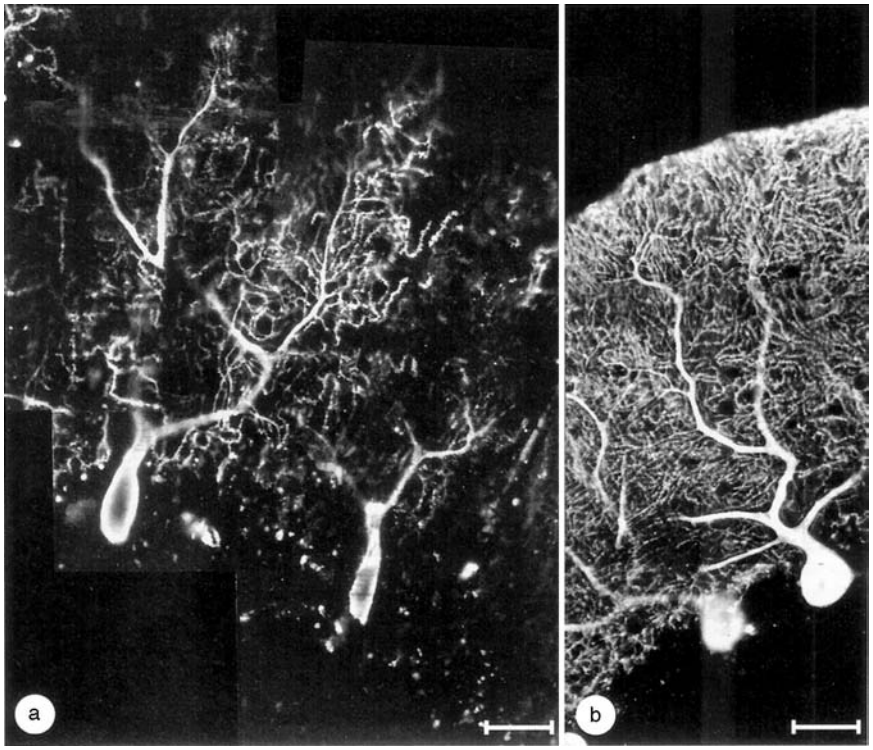

Figure 7 Immunofluorescence with anticalbindin: (a) patient with SCA6; (b) control. Scale bar = 50 μm. The image in (a) shows Purkinje cell atrophy. Reduction in the size of Purkinje cell bodies and simplified dendritic tree are observed in SCA-6. SCA, spinocerebellar ataxia. From Koeppen A (2005) The pathogenesis of spinocerebellar ataxia. *Cerebellum* 4: 62–73.

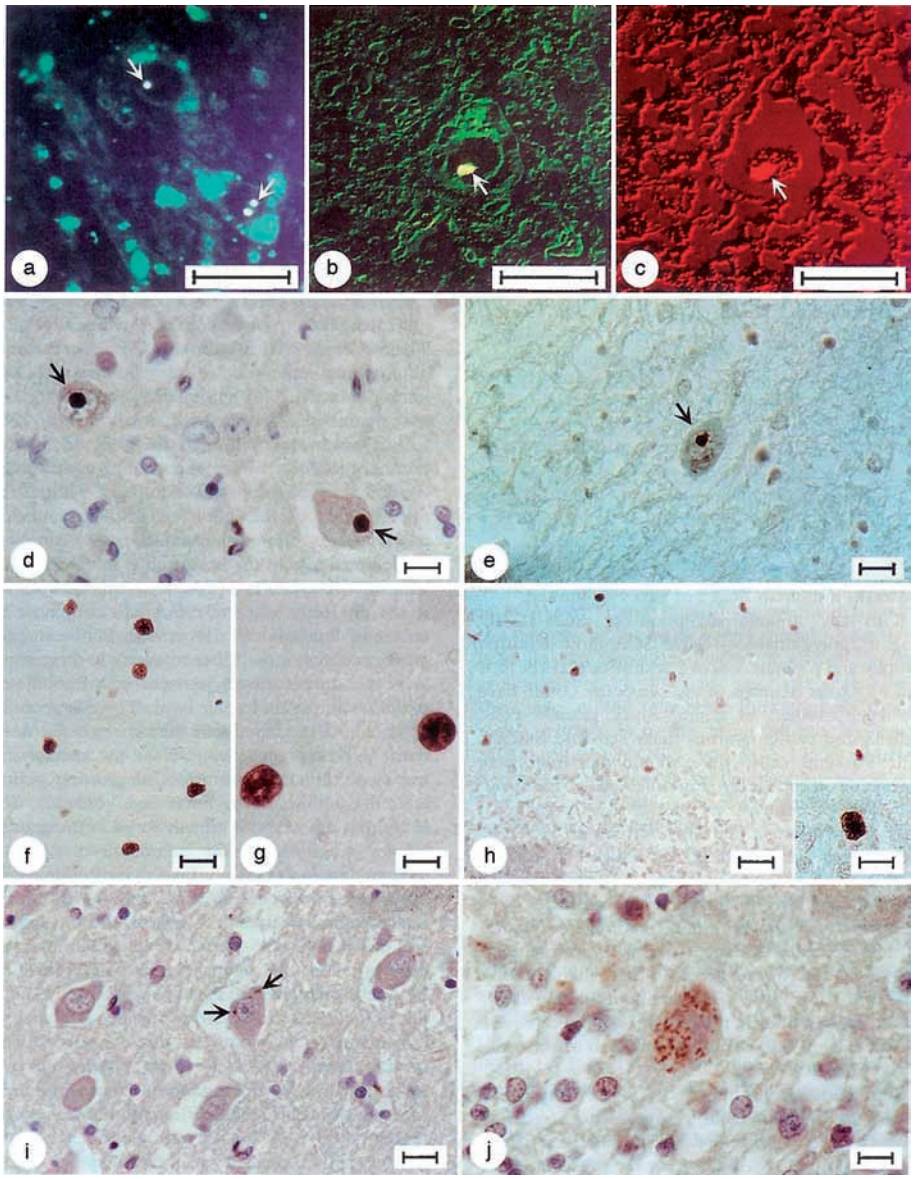

Figure 8 Inclusion bodies in spinocerebellar ataxias: (a–b) SCA3, anti-ataxin 3 (c) SCA3, anti-ubiquitin; (d) SCA1, antipolyglutamine; (e) SCA2, antipolyglutamine; (f–h) SCA17 (TATA-binding protein), antipolyglutamine; (i–j) SCA6, antipolyglutamine. Image (a) shows multiple small nucleoliform ataxin 3-reactive intranuclear inclusions in neurons of the basis pontis (arrows). Image (b) shows ataxin 3 reaction product (arrow). Image (c) shows co-localizing with ubiquitin (arrow). Image (d) shows positive–contrast reaction product in large intranuclear inclusions (arrows). Image (e) shows that the intranuclear inclusion body (arrow) is relatively small compared to those in SCA1 (d). In image (g), neurons of the basis pontis show pannuclear reaction product. Image (h) shows pannuclear inclusions in basket and stellate cells, and the inset shows an intranuclear inclusion in a surviving Purkinje cell. In image (i), pontine neurons show small intracytoplasmic inclusions (arrows). Image (j) shows Purkinje cell containing multiple cytoplasmic inclusion bodies. Scale bar = 25 μm (a–c, f, h); 10 μm (d, e, g, j, h inset); 15.9 μm (i). SCA, spinocerebellar ataxia. From Koeppen A (2005) The pathogenesis of spinocerebellar ataxia. *Cerebellum* 4: 62–73.

columns and pyramidal and spinocerebellar tracts, as well as atrophy/neuronal loss in Clarke's column. Mild degenerative changes may be found in the medulla, cerebellum, and pons. A biopsy of peripheral sensory nerves reveals an important loss of myelinated fibers, with the complete disappearance of large myelinated fibers, in absence of regenerative process.

A substantial degree of inflammation and a perinuclear accumulation of iron are observed in cases with hypertrophic cardiomyopathy. The analysis of iron and iron-responsive proteins suggests that cardiomyopathy results from iron-catalyzed mitochondrial damage followed by muscle-fiber necrosis and a chronic reactive myocarditis. Although iron accumulation in the heart is now an established feature for

FA, cardiomyopathy differs greatly from cardiac hemochromatosis, in which iron levels rise to much higher concentrations and iron does not accumulate in mitochondria. At the level of the heart, the thickness of interventricular septum and posterior wall and the ventricular mass indices are correlated with the GAA repeat length on the shorter allele.

FXTAS is characterized by the presence of eosinophilic ubiquitin-positive intranuclear inclusions in the neurons and in the astrocytes, both in the brain and spinal cord. Inclusions are τ-negative/α-synuclein-negative and contain *FMR1* mRNA, an argument in favor of the toxic RNA hypothesis. There is a correlation between the repeat size and the frequency of intranuclear inclusions. Spongiosis of the middle cerebellar peduncles and the loss of Purkinje cells are common. Weak, elongated, and immature synapses are found, hence the hypothesis that FMRP is involved in synaptogenesis.

Molecular Basis of Triplet Expansion Disorders

The repeats exert their detrimental effects through different molecular mechanisms, depending on whether they are located in coding or noncoding regions of the gene (Figure 9). On the basis of pathogenic mechanisms, disorders of unstable repeat expansions can be divided in four groups: (1) diseases caused by expansions of coding repeats which result in altered protein function, (2) diseases caused by expansions of noncoding repeats which result in loss of protein function, (3) diseases caused by expansions of noncoding repeats which result in impaired RNA function, and (4) diseases of unknown pathogenesis. The loci of trinucleotide expansions are given in Table 3.

Gain of Function of Proteins

Polyglutamine expansions are associated with transcriptional alterations. For several SCAs (SCA1, SCA2, SCA3, SCA6, SCA7, SCA8, and SCA17), an abnormally long polyglutamine tract has been reported as in Huntington's disease, hence the terminology of polyglutamine expansion disorders. Phenotypic differences among the diseases are determined not only by the repeat length but also by the instrinsic function of the disease-causing protein. Formation of unusual structures by repetitive DNA seems to be a core mechanism of repeat instability which leads to elongated polyglutamine tracts.

Proteins resulting from defective genes aggregate both in the cytoplasm and in the nucleus. They are cleaved before entering the nuclei. Transcriptional deregulation might be a key factor in the pathogenesis of most polyglutamine disorders, polyglutamine tracts interacting closely with well-defined transcriptional regulators such as cAMP responsive element-binding protein (CREB). For instance, the SCA7 gene product, ataxin 7, is part of a transcriptional coactivator complex, called STAGA, which has histone acetyltransferase activity. In SCA17, the CAG repeat expansion occurs in TATA-binding protein (TBP), a transcription factor. SBMA (X-linked) is caused by a polyglutamine expansion in the androgen receptor. Impaired protein conformation leads to protein accumulation and a cascade of aberrant protein–protein interactions, resulting finally in cell death. Chaperones and proteasomes are used to refold or dispose polyglutamine-containing fragments in order to prevent further aggregation. Accumulation of insoluble aggregates occurs in the nuclei for SCA1, SCA7, and SCA17; in the cytoplasm for SCA2 and SCA6; and in both the nuclei and cytoplasm for DRPLA, SBMA, and SCA3. Small aggregates sequester other proteins and protein complexes including proteasomes and transcription factors, leading to inclusion bodies. Formation of inclusions might be a protective mechanism against the toxicity of expanded polyglutamine proteins. A mechanism of impaired Ca^{2+} influx into neurons in SCA6, caused by mutations in the α-1A subunit of the voltage-gated neuronal Ca^{2+} channel, could contribute to the neurological deficits in this relatively pure cerebellar syndrome.

Loss of Protein Function

The decreased expression of molecules involved in translational control as in fragile X syndrome and in mitochondrial function as in FA has widespread pathogenic effects, as expected. Frataxin is a highly conserved nuclear-encoded protein found in prokaryotes and eukaryotes at the inner mitochondrial membrane. Frataxin appears necessary for the efficient regulation of cellular iron homeostasis. GAA trinucleotide repeat expansion in the frataxin gene results in frataxin deficiency, due to the inhibition of transcriptional elongation. Most patients have expansions in both alleles. The longer the repeat, the lower the level of frataxin. All patients carry at least one copy of an intronic GAA triplet-repeat expansion impairing frataxin transcription. Molecular diagnosis of FA is based on the detection of this expansion, which is present in homozygosity for the majority of cases. Some patients are heterozygous for the GAA expansion and show a frataxin point mutation at the other allele (compound heterozygotes). Several roles have been suggested for frataxin: as an iron-binding protein, in the protection

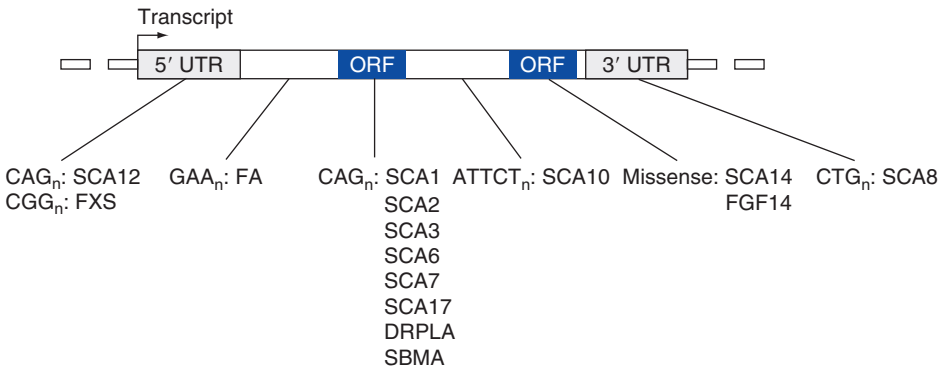

Figure 9 Trinucleotide repeat expansions in coding and noncoding DNA regions. (Not shown: CAG expansion in Huntington's disease (exon 1 of the Huntingtin gene).) DRPLA, dentatorubral-pallidoluysian atrophy; FA, Friedreich ataxia; FGF, fibroblast growth factor; FXS, fragile X syndrome; ORF, open reading frame; SBMA, spinal and bulbar muscular atrophy; SCA, spinocerebellar ataxia; UTR, untranslated region.

Table 3 Loci of triplet expansion disorders

Expansion subtype (*gene product*)	*Locus* (*normal/pathogenic repeat length*)	*Putative function*
Spinocerebellar ataxias		
SCA1 (ataxin 1)	6p23 (6–39/40–82)	Transcription
SCA2 (ataxin 2)	12q24.1 (15–24/32–200)	RNA metabolism
SCA3/Machado-Joseph disease (ataxin 3)[a]	14q24.3–q31 (13–36/61–84)	De-ubiquitylating activity
SCA4	16q22.1	
SCA5	11p12	
SCA6 (CACNA1A)	19p13 (4–20/20–29)	P/Q type α-1A Ca^{2+} channel subunit
SCA7 (ataxin 7)	3p12–13 (4–35/37–306)	Transcription
SCA8	13q21 (16–34/>74)	
SCA9		
SCA10 (ataxin 10)	22q13 (10–20/500–4500)	
SCA11	15q14–21.3	
SCA12 (PPP2R2B)	5q31–33 (7–45/55–78)	Phosphatase regulation
SCA13	19q13.3–q13.4	
SCA14 (PKC-γ)	19q13.4-qter	
SCA15	3p24.2-pter	
SCA16	8q22.1–q24.1	
SCA17 (TATA-box binding protein)	6q27 (25–42/47–63)	Transcription
SCA18	7q22–q32	
SCA19[b]	1p21–q21	
SCA20	11?	
SCA21	7p21.3–p15.1	
SCA22[b]	1p21–q23	
SCA23	20p13–12.3	
FGF14	13q34	
SCA25	2p15–p21	
DRPLA (atrophin 1)	12p13.31 (7–34/49–88)	Transcription
FX		
Fragile X syndrome (FMRP)	Xq27.3 (6–60/>200 full mutation)	Regulation of translation
FXTAS (FMRP)	Xq27.3 (6–60/60–200 premutation)	RNA-mediated
Fragile X syndrome E (FMR2)	Xq28 (4–39/200–900)	Transcription?
Friedreich ataxia (frataxin)		
	9q13–q21.1 (6–32/200–1700)	Iron metabolism in mitochondria

[a]Allelic to Machado-Joseph disease (MJD).
[b]CA19 might be allelic with SCA22.
DRPLA, dentatorubral-pallidoluysian atrophy; FXTAS, fragile X tremor ataxia syndrome; PKC, protein kinase C; SCA, spinocerebellar ataxia.

and synthesis of Fe–S clusters, and as an associate of ferrochelatase for heme metabolism. Tissue from patients is characterized by iron deposits and a generalized decrease in the activity of proteins containing Fe–S clusters. In the nervous system, frataxin is highly expressed in the spinal cord, particularly in the dorsal root ganglia and at lower levels in the cerebellum.

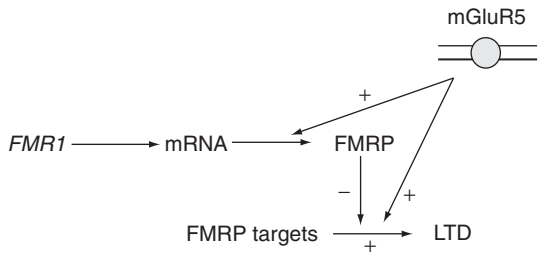

Figure 10 Interactions between *FMR1* gene and mGluR5 receptors. In subjects with a normal allele, mGluR5 stimulates the translation of a pool of *FMR1* mRNA. FMRP, the gene product, represses mRNA translation (FMRP targets the pool of mRNAs at the synapse). In case of full mutation, the translational repression is lost. The mGluR5-dependent long-term depression (LTD) is thus increased. mGluR, metabotropic glutamate receptor. Adapted from Jacquemont S, Hagerman R, Hagerman P, and Leehey M (2007) Fragile-X syndrome and fragile X-associated tremor/ataxia syndrome: Two faces of *FMR1*. *Lancet Neurology* 6: 45–55.

Fragile X syndrome is characterized by transcriptional silencing and loss of the protein product of *FMR1*, FMRP. Tracts with more than 200 CGG repeats cause fragile X syndrome, whereas intermediate alleles of 60–200 repeats (premutation) cause FXTAS. Fragile X syndrome may rarely result from deletions of *FMR1* or its promoter and not from the expansion of CGG repeats. FMRP is enriched in neurons. It associates with polyribosomes as part of a large messenger ribonucleoprotein particle (mRNP). FMRP inhibits translation. FMRP also localizes to the dendritic spines, where an upregulation occurs in response to stimulation by metabotropic glutamate receptors (mGluRs) (Figure 10). FMRP could be involved in translation of specific RNA targets at synapses and might modulate synaptic maturation and pruning (elimination of weak synapses) via a regulatory action of gene products involved in synaptic developments.

Patients with fragile X syndrome E (FRAXE) exhibit variable phenotypes, including learning deficits and mental retardation. Expansion of a CCG repeat of the *FMR2* gene results in transcriptional silencing and hypermethylation of the CpG Island. Deficits in the mitogen-activated protein kinase (MAPK) pathway are associated with X-linked mental retardation syndromes, suggesting that *FMR2* could play a role in these pathways. Both in fragile X syndromes and in FA, there is a correlation between repeat length and disease severity; transcriptional silencing being greater with longer repeats.

Impaired RNA Function

Carriers of premutation alleles (60–200 CGG repeats) in the *FMR1* gene may develop FXTAS. There are several lines of evidence indicating an RNA-mediated pathogenesis, with an RNA toxic gain of function. The elevation of *FMR1* mRNA to levels 2–8 times normal occurs, and the translation efficiency of *FMR1* mRNA is reduced. Inclusions co-localize with components of the protein degradation system. FXTAS shares with the next group the fact that expansion mutations occur in noncoding mRNA.

Unidentified Mechanism

SCA8 is caused by an expansion of CTG triplets in the 3′UTR of an untranslated RNA. The SCA8 transcript might function as an endogenous antisense RNA overlapping with the *KLHL1* gene encoding an actin-binding protein. Toxic effects of the expanded CUG RNA might participate in the pathogenesis.

SCA10 is characterized genetically by a pentanucleotide-repeat expansion in an intron of gene *E46L* (*ATXN10*). ATTCT repeats are unique in the high content of A-T, a feature of DNA sequences that form unpaired structures under torsional stress – called DNA unwinding elements (DUE). The length of expanded ATTCT repeats is variable in different tissues and highly unstable during paternal transmission. SCA10 repeat expansion is one of the largest microsatellite repeat expansions described in the human genome. Here also, the size of the expanded repeat shows an inverse correlation with the age of onset. SCA10 encodes a ATXN10 protein of 475 amino acids with a molecular weight of approximately 53 kDa. Transcriptional silencing of the gene is a plausible hypothesis.

A CAG repeat expansion has also been found in the 5′ region of the *PPP2R2B* gene in SCA12. *PPP2R2B* is a regulator of the protein phosphatase PP2A, modulating dephosphorylation activity for specific substrates such as vimentin or τ. The expansion does not encode a polyglutamine tract.

Experimental Models of Repeat Expansion Disorders

Mouse models have been generated for SCA1, SCA2, SCA3, SCA7, and SCA8. Transgenic mice expressing expanded polyglutamine tracts in a foreign protein develop a phenotype and neuropathological findings reminiscent of polyglutamine disorders. The fact that overexpression of molecular chaperones in fly and mice models reduces toxicity supports the hypothesis that expanded polyglutamines enhance mechanisms acting against protein misfolding. In *Drosophila*, histone deacetylase inhibitors have been shown to reduce toxicity.

Studies in yeast and mice have provided insights into FA pathogenesis. The deletion of the yeast

frataxin homolog YFH1 results in the accumulation of iron in the mitochondria; increased sensitivity to oxidative stress; and reduced activities of aconitase, which contains Fe–S clusters. The inactivation of the FA mouse gene leads to early embryonic lethality, in absence of iron deposits, showing an important role of frataxin in development. Oxidative stress has been pointed to as a main factor in the pathogenesis of FA, although a recent study in a mouse model challenged this hypothesis.

The *Drosophila* model of fragile X syndrome is characterized by a rescue of synaptic plasticity deficits with mGluR antagonists. Data in this model support the importance of the association of FMRP with the RNA interference machinery. FMRP might bind specific RNA ligands; this would recruit RNA-induced silencing complex (RISC) and micro-RNAs, facilitating translational control. *FMR2*−/− mice present with deficits in regulation of nervous system development, sensory perception, and synaptic plasticity. The fragile X premutation CGG repeats generate inclusion formation and progressive degeneration in the eye of *Drosophila melanogaster*. These inclusions are mainly found in the cytoplasm and are ubiquitin and HSP70 positive. Electrophysiological studies reveal an increased mGluR1-dependent long-term depression (LTD) in *FMR1* knockout mice, supporting the hypothesis of an upregulation of the downstream effectors of mGluRs.

Genetic Counseling

Diagnosis of a triplet expansion disorder has consequences for many family members. Therefore, genetic counseling is recommended when a triplet expansion disorder is suspected or demonstrated. Families of carriers of fragile-X alleles illustrate the complexity of counseling.

Therapies

Treatment of SCAs remain an open issue. We are still missing a therapy to prevent neuronal cell death. The benefits of pharmacological treatments remain limited, although some benefits regarding ataxic symptoms have been reported with 5-hydroxytryptophan, buspirone, acetazolamide in SCA6, and deep brain stimulation in patients presenting tremor in SCA2. *N*-methyl-D-aspartate (NMDA) modulators might be helpful in some cases. Amantadine/levodopa/dopamine agonists may improve extracerebellar symptoms in SCA2 and SCA3. Baclofen/tinazidine improves spasticity. Dystonia can be treated with botulinum toxin. However, patients with spinal cord involvement are at risk for developing muscle atrophy. Restless legs and periodic leg movements in sleep usually respond to dopaminergic treatment. Muscle cramps may be disabling, especially in SCA3. Cramps may respond to magnesium, quinine, or mexiletine. Urinary urgency is improved with spasmolytics or adrenergic α-receptor blockers. Depression may improve with antidepressant treatment. Seizures require the administration of antiepileptic drugs according to international guidelines.

Pilot studies with vitamin E, coenzyme Q10, or idebenone have shown the potential effects of antioxidant therapies for FA. Idebenone seems to play a protective effect against cardiomyopathy. If iron chelators are to be used, they must be able to enter in the mitochondria. Larger randomized clinical trials with antioxidants are ongoing. A combination of antioxidants, anti-inflammatory drug, and a chelating drug could be considered. β-Blockers at high doses could provide benefits in patients with heart disease.

mGluR5-coupled pathways are dysregulated in individuals with fragile X syndrome. Whether mGluR5 antagonists (such as derivatives of methylphenylethynylpyridine (MPEP)) could be useful in treating FXTAS needs to be investigated.

Speech rehabilitation and regular physiotherapy/occupational therapy are recommended. Most patients will have some improvements with the use of orthesis, sticks, or strollers. Unfortunately, many patients will become wheelchair-bound during the course of their illness. Swallowing difficulties are common in advanced cases of SCAs. Gastrostomy is usually recommended.

New drugs are currently tested in transgenic models (*Caenorhabditis elegans*, *Drosophila*, and mice). RNAi therapies with the goal of inhibiting polyglutamine-induced neurodegeneration is under evaluation. Transplantation of stem cells is another approach. Prevention of protein misfolding and aggregation by overexpressing chaperones, such as heat shock protein (HSP)70 or DNAJ1, and application of histone deacetylase (HDAC) inhibitors, increasing acetylation at particular residues on histones, are being assessed.

For specific queries about DNA studies, readers are referred to the Genetic Alliance website. Users will have access to supports groups (e.g., National Ataxia Foundation (NAF) and Friedreich Ataxia Research Alliance (FARA)). In Europe, the Eurosca consortium does the analysis; a list of universities performing the studies is on the website.

See also: Cerebellum: Clinical Pathology; Congenital Muscular Dystrophy; Fragile X Associated Tremor Ataxia Syndrome; Friedreich Ataxia; Sporadic Degenerative Ataxias and the Dominantly Inherited Spinocerebellar Ataxias; Triplicate Repeats: Huntington's disease.

Further Reading

Arrasate M, Mitra S, Schweitzer ES, Segal MR, and Finkbeiner S (2004) Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature* 431: 805–810.

Babcock M, de Silva D, Oaks R, et al. (1997) Regulation of mitochondrial iron accumulation by Yfh1p, a putative homolog of frataxin. *Science* 276: 1709–1712.

Bates GP and Gonitel R (2006) Mouse models of triplet repeat diseases. *Molecular Biotechnology* 32: 147–158.

Bhidayasiri R, Perlman SL, Pulst S-M, and Geschwind DH (2005) Late-onset Friedreich ataxia: Phenotypic analysis, magnetic resonance imaging findings, and review of the literature. *Archives of Neurology* 62: 1865–1869.

Brunberg JA, Jacquemont S, Hagerman RJ, et al. (2002) Fragile X premutation carriers: Characteristic MR imaging findings of adult male patients with progressive cerebellar and cognitive dysfunction. *American Journal of Neuroradiology* 23: 1757–1766.

Cohen S, Masyn K, Adams J, et al. (2006) Molecular and imaging correlates of the fragile X-associated tremor/ataxia syndrome. *Neurology* 67: 1426–1431.

Durr A, Cossée M, Agid Y, et al. (1996) Clinical and genetic abnormalities in patients with Friedreich's ataxia. *New England Journal of Medicine* 335: 1169–1175.

Gatchel JR and Zoghbi HY (2005) Diseases of unstable repeat expansion: Mechanisms and common principles. *Nature Reviews Genetics* 6: 743–755.

Helmlinger D, Tora L, and Devys D (2006) Transcriptional alterations and chromatin remodeling in polyglutamine diseases. *Trends in Genetics* 10: 562–570.

Jin P, Alisch RS, and Warren ST (2004) RNA and micro RNAsin fragile X mental retardation. *Nature Cell Biology* 6: 1048–1053.

Koeppen A (2005) The pathogenesis of spinocerebellar ataxia. *Cerebellum* 4: 62–73.

Lin X and Ashizawa T (2005) Recent progresses in spinocerebellar ataxia type 10. *Cerebellum* 4: 37–42.

Manto M (2005) The wide spectrum of spinocerebellar ataxias. *Cerebellum* 4: 2–6.

Mascalchi M, Cosottini M, Lolli F, et al. (2002) Proton MR spectroscopy of the cerebellum and pons in patients with degenerative ataxia. *Radiology* 223: 371–378.

Nakamura K, Jeong S-Y, Uchihara T, et al. (2001) SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Human Molecular Genetics* 10: 1441–1448.

Rotig A, deLonlay P, Chretien D, et al. (1997) Aconitase and mitochondrial iron-sulphur protein deficiency in Friedreich ataxia. *Nature Genetics* 17: 215–217.

Schols L, Bauer P, Schmidt T, Schulte T, and Riess O (2004) Autosomal dominant cerebellar ataxias: Clinical features, genetics and pathogenesis. *Lancet Neurology* 3: 291–304.

Steffan JS, Bodai L, Pallos J, et al. (2001) Histone deacetylase inhibitors arrest polyglutamine-dependent neurodegeneration in *Drosophila*. *Nature* 413: 739–743.

Verkerk AJMH, Pieretti M, Sutcliffe JS, et al. (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* 65: 905–914.

Willemsen R, Hoogeveen-Westerveld M, Reis S, et al. (2003) The *FMR1* CGG repeat mouse displays ubiquitin-positive intranuclear neuronal inclusions: Implications for the cerebellar tremor/ataxia syndrome. *Human Molecular Genetics* 12: 949–959.

Relevant Websites

http://www.eurosca.org – Eurosca consortium.
http://www.geneticalliance.org – Genetic Alliance.

Split-Brain Patients

M E Roser, University of Plymouth, Plymouth, UK
M S Gazzaniga, University of California at Santa Barbara, Santa Barbara, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Terminology

The term split brain is used to denote a patient in whom the cortical commissures have been severed, for the relief of epilepsy that cannot be controlled by medication, in an operation known as either a commissurotomy or a callosotomy. A commissurotomy involves severing the corpus callosum, the anterior commissure, and the hippocampal commissure. The massa intermedia, if present, is also severed. A callosotomy involves sectioning of the corpus callosum alone, either completely or partially. In a partial callosotomy, either the anterior or the posterior callosum may be cut. Complete callosal disconnection may result from a staged callosotomy in which the anterior and the posterior parts of the callosum are severed in two separate operations.

History

The split-brain operation in humans was first reported in the early 1940s by the neurosurgeons Van Waganen and Herren, who carried out more than 30 such operations as a treatment for intractable epilepsy. Severing the commissures between the hemispheres prevented the interhemispheric propagation of seizures. These patients were later tested by Akelaitis, who reported that the patients showed no marked effects on behavior beyond the reduction in their epileptic symptoms. This negative finding was congruent with earlier studies of callosal infarcts which suggested that the cortical commissures had little functional role, a state of affairs which prompted Karl Lashley to facetiously remark that the purpose of the corpus callosum was to "keep the hemispheres from sagging."

It was not until the systematic experiments carried out on commissurotomized cats and monkeys by Ronald Myers and Roger Sperry that the importance of the cortical commissures for the transmission of information between the hemispheres became strikingly apparent. Cutting the commissures disrupted the transfer of visual and tactile information, and learned associations, from one hemisphere to the other. The animal studies also suggested that division of the hemispheres would have no catastrophic consequences for human patients because the animals displayed normal behavior outside of the testing situation.

Following the publication of these studies, the commissurotomy operation was carried out on a number of human patients in the early 1960s by the neurosurgeons Philip Vogel and Joseph Bogen. These patients are commonly referred to as the California series and include patients W.J., N.G., and A.A. Other extensively tested patients, including P.S. and J.W., were operated on in the 1970s by Donald Wilson and David W Roberts at Dartmouth Medical School. Many of these patients have participated in a large number of experiments over decades.

The divided human brain offered a unique opportunity for each hemisphere to be studied in relative isolation. The initial investigations of hemispheric communication and asymmetry in humans were carried out by Michael Gazzaniga, working in the laboratory of Roger Sperry. These studies demonstrated the striking disconnection syndrome in which each hemisphere remains unaware of the sensory and gnostic properties of stimuli presented to its opposite number.

Testing Split-Brain Patients

In everyday situations, the split-brain patients appear normal. This is due to the freedom with which they can direct their gaze and movements around in the environment, allowing both hemispheres knowledge of their surroundings. Only when patients are tested in strictly controlled circumstances does the disconnection syndrome become apparent. To test each hemisphere individually, stimuli must be presented to only one hemisphere. Tactile shapes, for example, may be palpated by one hand but hidden from view.

Testing with visual stimuli usually involves the presentation of stimuli to each of the visual hemifields for a duration less than 200 ms, brief enough to preclude a saccade toward the stimulus while it is displayed. Each hemifield projects to the contralateral visual cortex, and thus a stimulus in the left hemifield will be seen by the right hemisphere. In order to achieve lateralized presentation of stimuli, patients are required to fixate on a central point on the display. Stimuli are displayed at locations eccentric to the foveated point by several degrees of visual angle. A chin rest is often used to minimize head movements and infrared eye tracking can ensure that the patient's gaze remains on the fixation point. Using the brief-presentation technique, simple pictures, words, shapes, and colors can be lateralized.

For presentations longer than a few hundred milliseconds, necessary for extended viewing of complex stimuli, steps must be taken to ensure the stimulus remains in one visual hemifield. One method is to prevent foveation by employing a stimulus deflector, which tracks eye movements and adjusts the view of the display accordingly so as to ensure that any saccade toward the stimulus is accompanied by a corresponding movement of the stimulus away from fixation. Another method allows for extended viewing in free, monocular, vision but restricts visual input to one hemisphere through the use of a contact lens that occludes one-half of the visual field. Finally, visual information may be presented in full view of the patient, and thus to both hemispheres, with a response required from only one hemisphere.

Responses are typically collected from the hemisphere to which the stimulus was presented, often by using the hand on the same side as the stimulus. Control of the distal effectors, such as fingers, is thought to be predominantly contralateral. Thus, button presses, as a measure of either simple reaction time or of categorical choice, are often used. In some experiments, a hemisphere may be required to respond by reaching for a picture representing the correct choice.

Care must be taken when testing patients that one hemisphere is not able to cross-cue, or tip off, its opposite number as to the nature of a stimulus or the correct response. Patients are seated with their hands and feet apart. Objects are placed on a soft surface so that they do not make identifying noises when moved.

Neuropsychological Profile

Most patients who have participated in multiple studies over several decades have been assessed as showing verbal, memory, and attentional capacities within normal limits, although scores are often below average. Memory, particularly free recall, shows the greatest deficit, and may result from section of both the corpus callosum and the hippocampal commissure.

Drawing conclusions about function in the brain on the basis of results from split-brain patients is complicated by their history of epilepsy and invasive surgery. Because split-brain patients have diverse medical histories, it is desirable to test as many patients as possible. Often, a multiple case study approach is taken because patients cannot be grouped for comparison with a control group. This heterogeneity does, however, increase the impressiveness of obtaining the same result in two or more patients because consistent patterns of asymmetry are unlikely to have resulted from common neurological history.

Although the brains of callosotomy patients differ greatly from intact brains, evidence from both behavioral testing and functional imaging studies suggests that the patterns of functional lateralization observed in the split brain are congruent with those found in neurologically intact participants in many different cognitive domains. It must be remembered, however, that the divided brain is an unusual situation and that cognitive processes may not operate in the same fashion in each isolated hemisphere as they do in the intact brain. Thus, comparison with a control sample is desirable.

The Disconnection Syndrome

One of the earliest findings from split-brain research was the importance of the corpus callosum for the transmission of information between the hemispheres. The relative isolation of each hemisphere is referred to as the disconnection syndrome and is manifested as poor performance at cross-comparisons of stimuli presented to different hemispheres and, for most patients, an inability to verbally name stimuli presented to the right hemisphere.

Split-brained patients often perform at chance level on tests in which they must decide whether patterns or objects in the two visual fields are the same or different. This includes letters, digits, words, colors, shapes, and luminances. The importance of the corpus callosum for the transfer of information between the visual cortices is also apparent using electrophysiological measures. Sensory evoked potentials are not seen over the hemisphere ipsilateral to a visual stimulus in patients in whom the posterior corpus callosum has been severed. Some patients, however, can judge whether lines on each side are aligned or not. These preserved abilities presumably rely on intact subcortical pathways between the hemispheres.

An inability to name stimuli presented to the right hemisphere can be demonstrated by having the split-brain patient fixate on a central point and flashing stimuli to the left visual field. If the patient is asked to verbally name the stimulus presented to his or her left visual field, he or she is unable to do so because the right hemisphere, in most patients, is not able to produce speech and is separated from the linguistically able left hemisphere. The right hemisphere can, however, control the left hand to find a covered object that corresponds to the stimulus presented. The patient is not able to verbally name the object until it is moved into the right visual field or the right hand.

Other evidence for perceptual disconnection comes from presenting, to each visual field, words that can be

combined to form the names of emergent objects. Patients are then asked to draw what they saw. One callosotomy patient, V.P., who has some residual fibers left in her corpus callosum, is able to integrate words presented in the two visual fields (e.g., the words 'head' and 'stone') and draw the emergent object, in this case a headstone. Another subject, J.W., whose corpus callosum was fully cut, did not draw emergent objects when tested but instead drew each object separately. Thus, the perception of complex visual information is divided between the hemispheres of the split brain.

Evidence for the division of attention between the hemispheres is more equivocal. Visual search of an array of objects can be carried out independently by each hemisphere. This was suggested by the observation that the addition of items to the stimulus array impacts to a lesser degree on performance when the objects to be searched are distributed between the two visual hemifields than when the items are all contained within one hemifield. The ability to carry out two independent search processes also leads to a rare advantage in performance for split-brain patients over callosally intact participants.

The ability to divide some attentional processes between the hemispheres does not, however, endow split-brain patients with twice the attentional resources available to neurologically normal individuals. Performance of a demanding task by one hemisphere has been shown to negatively impact upon performance by the other hemisphere, suggesting a finite, shared, pool of attentional resources.

Learned bimanual tasks are not adversely affected by callosotomy, although the acquisition of new skills involving complex bimanual coordination is impaired. Conversely, callosotomy patients can perform some coordination patterns which callosally intact subjects find difficult, such as drawing two different shapes simultaneously, one with each hand. This independence of action following callosotomy suggests that the corpus callosum contributes to the interhemispheric integration of bimanual movement. Some studies suggested that callosal integration is confined to information about movement direction and not timing because temporal coupling of the hands was preserved in callosotomy patients, although recent results have suggested that this only occurs for movements with discrete events.

The divided brain presents an opportunity to study each hemisphere in relative isolation. It can also allow the researcher to obtain double dissociations of cognitive processes if the two processes rely critically on different hemispheres. In one case, seemingly indivisible cognitive processes, such as spoken and written language output, can be separated into lateralized components.

Partial Callosotomy and Callosal Channels

Callosotomy is typically performed in a staged manner, with either the anterior callosum, the genu, or the posterior callosum, the splenium, sectioned first, followed by the remaining fibers if the first operation fails to sufficiently reduce the patient's epilepsy. Because the corpus callosum is organized in a roughly topographic fashion, tests of patients after partial callosotomy can reveal the functional specificity of regions of the corpus callosum. Section of the splenium and midbody disrupts the transmission of visual and somatosensory information between the hemispheres, resulting in poor performance on cross-comparison tasks. Intact anterior pathways are able to support the transfer of semantic information between the hemispheres, allowing patients to perform above chance on tests of naming left visual field stimuli, although performance is often slow and imprecise. Anterior section produces less evidence of disconnection, although the disruption to anterior midbody tracts can affect bimanual coordination for complex motions.

Subcortical Transfer

Each hemisphere of the split brain shows awareness of the presentation of a stimulus to the opposite hemisphere. Simple responses, such as button presses, can be made to stimuli presented in the visual field opposite the response hand. These are known as crossed responses and are subject to a delay, ranging from 30 to 70 ms, relative to uncrossed responses. This delay is thought to index the time taken for transmission of a signal from the hemisphere receiving the stimulus to that making the response, via subcortical channels, and contrasts with 5–15 ms delay for interhemispheric transfer in callosally intact subjects.

The affective content of a stimulus may also be transferred between the hemispheres, as shown by an experiment in which emotive pictorial stimuli were presented to the right hemisphere and were found to affect the judgment of emotionally neutral stimuli, and the patient's current environment, by the left hemisphere.

The complexity of subcortical visual transfer in the split brain is limited to basic spatial information that can support interhemispheric comparisons of stimulus location and alignment. Some experiments suggested that higher order information, such as category or numerical value, could be transferred subcortically and support comparative judgments such as whether a stimulus in one visual field was bigger than

a stimulus in the opposite field or whether two stimuli were both of the same categories. The deterioration in performance of stimulus cross-comparisons with increasing set size suggested, however, that this apparent ability was due not to the transfer of complex information but to the responding hemisphere adopting a response strategy based only on the information presented to it. Thus, care must be taken, when testing subcortical transmission in split-brain patients, to ensure that experimental situations truly require transmission for above-chance performance. The transfer of binary information, such as whether a number is odd or even, coupled with response strategies, seems sufficient to explain most instances of apparent higher order transfer.

Hemispheric Specialization

Testing the split brain can reveal whether each isolated hemisphere is competent at a particular task and the degree to which a process can be supported by a single hemisphere. The term specialization, when applied to one of the hemispheres in the split brain, is sometimes used to imply that one hemisphere bears primary responsibility for a particular perceptual, cognitive, or motor process, usually on the basis of a performance difference between the hemispheres. Asymmetries in behavior are often more stark in the divided brain than in lateralized responses from normal controls. Double dissociations between the performances of the two hemispheres on two tasks are sometimes observed, and most fully justify the interpretation that the different processes are predominantly handled by different hemispheres. This is especially true if each hemisphere performs at chance level on one task. Often, hemispheric differences in performance are relative because both hemispheres are capable of above-chance performance. This result is often interpreted as suggesting that the two hemispheres contribute to performance to different degrees. Findings suggesting different levels of competence in each hemisphere are congruent with studies in neurologically normal subjects that suggest that many hemispheric asymmetries are relative, rather than absolute, and can change as task demands change.

Language

Early testing of the split-brain patients revealed their inability to name stimuli in the left visual hemifield and confirmed the primacy of the left hemisphere for the understanding and production of language. In most patients, only the isolated left hemisphere is able to process complex grammar, although in some patients the right hemisphere is capable of making simple grammaticality judgments. The right hemisphere of most patients can understand simple verbal instructions, although these are often accompanied by visual examples presented in central vision. It can also understand simple phrases which, although they contain several words organized grammatically, may be best characterized as single items in a lexicon.

The right hemisphere is able to understand the meaning of words and can choose a picture that represents the item represented by an auditory or visual word stimulus, but in most patients it is unable to match rhyming words and pictures. This suggests that the right hemisphere does not generate the corresponding phonology for the item represented in the picture. However, several patients have demonstrated the development of the ability to verbally name stimuli presented in the left visual hemifield, in one case approximately 10 years after surgery.

Visual Perception

Both hemispheres in the split brain are capable of a range of visual processes, including object recognition, but the right hemisphere displays superiority for visual processes that include a spatial component. The spatial ability of the isolated right hemisphere was demonstrated early in the history of split-brain testing by requiring patients to replicate a pattern by arranging blocks with each hand. This task was performed better, by most patients tested, when using the left hand and patients were frequently unsuccessful when using their right hand. Similarly, tactile patterns are better recognized by the right hemisphere. In the visual domain, the difference in spatial ability between the hemispheres is apparent on a test of mental rotation, which requires the patient to decide whether two objects are mirror images or identical, but rotated, objects. Additional studies have found right hemisphere superiority for judging spatial relations, such as orientation and alignment, between objects. Comparisons of visual stimuli on nonspatial dimensions, such as luminance, can be performed equally well by the two hemispheres.

Based on observations of visual field differences in normal subjects, several dichotomies have been proposed for how visual processes differ between the two hemispheres. These include specializations for processing high- and low-spatial frequencies, the global and local arrangement of arrays, and categorical and coordinate representations of space. Testing in the split brain has not lent strong support to these hypotheses, with several studies failing to show the predicted hemispheric differences. The most common finding was that both hemispheres were capable of

performing both modes of processing included within each of the dichotomies, with occasional instances of superior performance by one hemisphere, in line with predictions. This finding is congruent with the view that hemispheric specialization is relative, and fluid, rather than absolute.

The right hemisphere has shown superiority over the left hemisphere for some types of perceptual grouping. Both hemispheres, in two callosotomy patients, were able to group visual stimuli and perceive illusory contours, but only the right was able to group stimuli and perceive an amodally completed object, in which the object's illusory boundaries were partly occluded. This finding suggests that simple perceptual grouping processes are supported by each hemisphere, but complex grouping processes are lateralized.

Differences between the hemispheres are also apparent in the processing of the timing of visual events. The left hemisphere of commissurotomy patient L.B. performed better than the right hemisphere at discriminating between simultaneous and successive lights when delays were very short. In contrast, L.B., and two callosotomy patients, showed a right hemispheric advantage for temporal discrimination of delays between the onsets of two visual stimuli when delays were in the range necessary to produce apparent movement. The two callosotomy patients also showed right hemisphere superiority for determining if the motion of a stimulus was caused by the impact of another stimulus. Thus, the right hemisphere seems superior to the left for the perception of interactions between objects in space and time.

Face Processing

Both hemispheres in the divided brain are able to recognize the self from pictures of the patient's face which are morphed with pictures of the faces of others, although different studies have found opposite hemispheric asymmetries in biases toward identifying faces as either self or other. The right hemisphere in several callosotomy patients has shown greater ability than the left hemisphere in matching pictures of similar faces to a sample, although tests of other patients have failed to find a right hemispheric advantage on tests of face perception. The left hemisphere is better able to voluntarily produce facial expressions of different emotions.

Memory

Some commissurotomy and callosotomy patients have shown lowered performance on tests of general memory ability, although others show little deficit. In affected patients, this deficit is greater than that seen in other domains investigated by standard neuropsychological assessment. Poor memory for new information, such as word associations, contrasts with the intactness of previously learned information and memories of events prior to the surgery. This deficit may result from interruption to both the corpus callosum and the hippocampal commissure and fornix. The involvement of the hippocampal commissure is suggested by the lack of memory deficits exhibited by some anterior callosotomy patients, in whom the hippocampal commissure is spared. Evidence from several patients also suggests that recognition memory may not be affected to the same degree as recall.

Relative hemispheric advantage on memory tests can depend on both the nature of the task and the stimuli. For instance, it was found that in several patients the right hemisphere outperformed the left hemisphere in a test of memory for tactile shapes, but the left hemisphere performed better when patients were required to remember sequences of tactile shapes. Material-specific depth-of-processing effects, in the two hemispheres of two callosotomy patients, and the lack of a catastrophic memory deficit suggest that both hemispheres are capable of memory encoding. The finding also suggests that different material, such as faces and words, are encoded preferentially by one of the two hemispheres.

A tendency for the left hemisphere to elaborate on information it receives may underlie its poor performance on tasks that involve distinguishing between previously seen and novel items with semantic content, such as words or pictorial representations of common scenes. The left hemisphere was found to falsely recognize previously unseen items that were semantically congruent with items shown during the encoding phase. The right hemisphere made no such systematic error.

Higher Cognition

Debate on whether the isolated hemispheres each possess consciousness, either complete or impoverished, has persisted since the publication of early observations from split-brain patients. Both hemispheres display a range of positive competencies, but the mounting evidence for, sometimes subtle, hemispheric differences suggests that consciousness is not identical in the two hemispheres.

The isolated right hemisphere performs poorly on many problem-solving tasks, whereas the left hemisphere remains capable of a level of abstract reasoning comparable to the patient's preoperative performance. The right hemisphere in two callosotomy patients was found to be unable to perform a task involving reasoning about causes, despite this

task being solvable by young infants, although the left hemispheres in the same patients performed well.

The left hemispheres of several callosotomy patients have shown a propensity for providing elaborative accounts of their experience and their own behavior, including those initiated by the right hemisphere. If the left hemisphere is challenged to provide an explanation for the behavior of the right hemisphere, patients often reply with interpretation that is consistent with the information presented to the patient's verbally able left hemisphere. The tendency for the left hemisphere to form generalizations about its environment has also been shown in a nonlinguistic test involving the prediction of the location of a forthcoming stimulus. Thus, the interpretive and linguistic processes that allow individuals to make sense of their conscious experience seem to be most fully actualized in the left hemisphere of the split brain.

See also: Attentional Functions in Learning and Memory; Brain Asymmetry: Evolution; Consciousness: Neurophysiology and Visual Awareness in; Consciousness: Neural Basis of Conscious Experience; Dichotic Listening Studies of Brain Asymmetry; Epilepsy; Hemispheric Specialization and Cognition; Memory Representation.

Further Reading

Bogen JE (1993) The callosal syndromes. In: Heilman KM and Valenstein E (eds.) *Clinical Neuropsychology,* 3rd edn., pp. 337–407. New York: Oxford University Press.

Corballis MC (1994) Split decisions: Problems in the interpretation of results from commissurotomized subjects. *Behavioural Brain Research* 64: 163–172.

Gazzaniga MS (2000) Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain* 123: 1293–1326.

Gazzaniga MS (2006) Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience* 6: 653–659.

Gazzaniga MS, Bogen JE, and Sperry RW (1962) Some functional effects of sectioning the cerebral commissures in man. *Proceedings of the National Academy of Sciences of the United States of America* 48: 1765–1769.

Geschwind N (1965) Disconnexion syndromes in animals and man: I. *Brain* 88: 237–294.

Geschwind N (1965) Disconnexion syndromes in animals and man: II. *Brain* 88: 585–644.

Hugdahl K and Davidson RJ (eds.) (2003) *The Asymmetrical Brain.* Cambridge: MIT Press.

Sperry R (1982) Some effects of disconnecting the cerebral hemispheres. *Science* 217: 1223–1226.

Trevarthen C (ed.) (1990) *Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry.* Cambridge, UK: Cambridge University Press.

Zaidel E and Iacoboni M (eds.) (2003) *The Parallel Brain: The Cognitive Neuroscience of the Corpus Callosum.* Cambridge: MIT Press.

Spontaneous Patterned Activity in Developing Neural Circuits

P Wenner, Emory University, Atlanta, GA, USA
M B Feller, University of California, at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

How do neuronal circuits become assembled during development? Mechanisms that govern the formation of neural circuits during development traditionally have been divided into activity-independent and activity-dependent categories. In this view, the early organization of the circuits, namely the phenotype, location, and early connections of neurons, is governed by activity-independent mechanisms (e.g., the genetic profile of the cell). Neural activity subsequently acts to refine initially crude connections into the detailed circuits that characterize the adult nervous system. However, in the past few years, challenges to these two basic assumptions have suggested the model may be oversimplified. First, there is growing evidence that environmental cues, mediated by chemical and electrical activity within the cell, influence the earliest stages of development. Second, immature circuits generate highly correlated activity patterns before the onset of sensory experience, raising the possibility that this activity can drive early developmental events. This article is devoted to describing spontaneous correlated activity in developing neural circuits and the cellular mechanisms that underlie this activity.

The Depolarizing Nature of GABA

One common feature of developing neural circuits is that in immature neurons, activation of γ-aminobutyric acid $(GABA)_A$ receptors is excitatory. The depolarizing action of GABA is due to the complement of chloride cotransporters that exist early in development. High levels of NKCC1, a Na–K–Cl cotransporter which move chloride ions into cells, and low levels of the cotransporter KCC2 which exports chloride, results in elevated intracellular chloride levels. Hence, the reversal potential for GABA is more depolarized than the resting membrane potential. Thus, when $GABA_A$ receptors are opened, chloride leaves the cell, producing a depolarization. Later in development, as KCC2 becomes the predominant cotransporter, intracellular chloride is lowered and GABA becomes hyperpolarizing, as in the adult. In the developing spinal cord and hippocampus, the depolarizing, excitatory nature of GABA is crucial to the existence of the spontaneously correlated activity.

Spinal Cord

Spontaneous network activity (SNA) has been studied in developing spinal networks in the embryonic chick, mouse, and rat. Extensive study in the chick embryo has provided the best understanding of the underlying mechanisms and functions of SNA and is therefore the main focus of this discussion; however, the same principles are likely to apply for SNA in most species. SNA begins at a stage corresponding to when the first synapses are formed (~embryonic day 4.5; E4.5), and continues into late embryonic development (E16), when chloride conductances are no longer excitatory. Through these stages limb motoneurons have innervated their targets, interneuronal and sensory circuits have formed, motor and sensory neuron cell death begins and ends, and significant descending input arrives in the lumbosacral cord.

SNA is experienced in many spinal neurons synchronously. Cells have 20–30 mV depolarizations upon which are superimposed bursts of spiking activity. SNA lasts for multiple seconds followed by periods of many minutes of inactivity (**Figure 1(a)**). A vast majority of spinal neurons are recruited into the bursts or episodes, including motoneurons, which drive embryonic limb movements. In the chick embryo, within a single episode of SNA there are cycles of depolarizations and associated bursting, the period of which becomes longer toward the end of the episode. The pattern of activity within a cycle is dependent on the kind of motoneuron and can alternate between antagonist motoneurons. Although the mechanism is unknown, it has been suggested that a fast depression and recovery may underlie the generation of the cycling. When the cord is no longer able to generate the next cycle the episode ends.

Episodes of spontaneous network activity occur as a result of a highly excitable, synaptically coupled circuit, where GABA, acting through $GABA_A$ receptors, is depolarizing and therefore excitatory. Thus, the network is hyperexcitable because both excitatory and what will become inhibitory neurons promote the activation of the cells to which they project. The recruitment sequence is as follows. Motoneurons begin to fire in the interval between episodes, and progressively fire more vigorously as the interval extends. Surprisingly, spinal interneurons do not appear to discharge in the interepisode interval. Rather, motoneuron discharge eventually intensifies

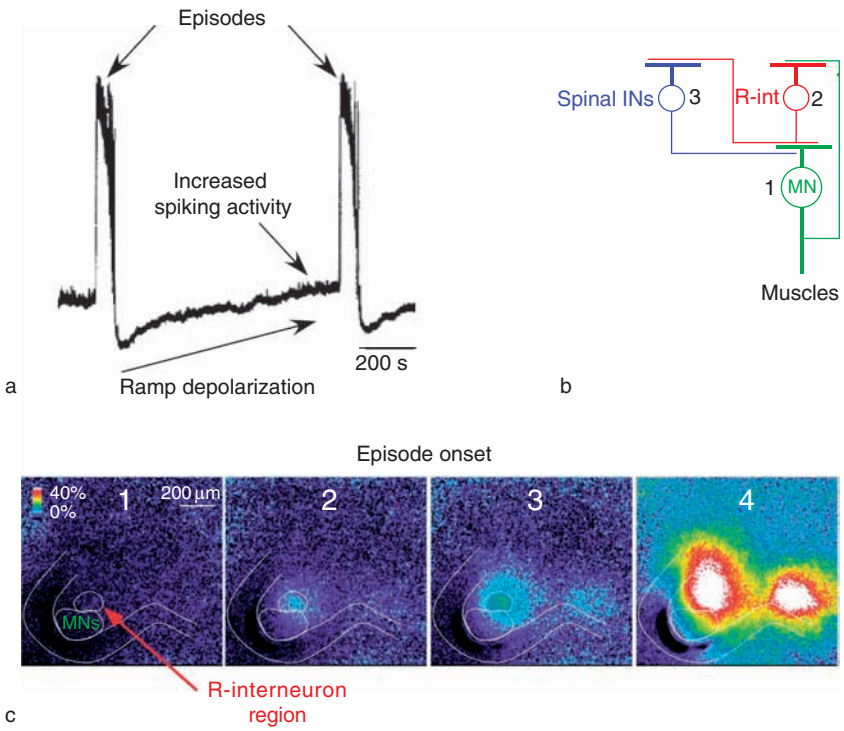

Figure 1 Spontaneous network activity in the spinal cord and its onset. (a) An example of two episodes of spontaneous network activity is shown via extracellular muscle nerve recordings. Episodes are associated with large depolarizations and spiking, and are followed by an interepisode interval consisting of a ramp depolarization and associated motoneuron discharge toward the end of the interval. (b) Schematic demonstrating the connectivity of motoneurons (MN), GABAergic R-interneurons (R-int), and other GABAergic and glutamatergic spinal interneurons. Numbers indicate the order in which each population is recruited at episode onset. (c) Recruitment order at episode onset is shown by imaging the cut face of the spinal cord with many spinal interneurons labeled with a calcium-sensitive dye. In panel 1, unlabeled MNs are active as determined by simultaneous muscle nerve recordings (not shown). In panel 2, cells in the location of R-interneurons become active, followed by other interneurons (panel 3), and finally the entire network of labeled interneurons (panel 4). Adapted from Wenner P and O'Donovan MJ (2001) Mechanisms that initiate spontaneous network activity in the developing chick spinal cord. *Journal of Neurophysiology* 86: 1481–1498, used with permission from the American Physiological Society.

to the point where it recruits GABAergic interneurons that receive direct input from the motoneuron recurrent collaterals. The excitatory GABAergic projections of these interneurons then excite other interneurons, eventually recruiting the entire network. As a result, spiking activity at episode onset is first observed in motoneurons, then recurrently activated interneurons, and finally all spinal neurons (Figures 1(b) and 1(c)).

During the episode, the spinal circuits become progressively depressed until they can no longer sustain the activity. Following an episode, several features of excitability of the cord are depressed. Spinal neurons are hyperpolarized relative to the period just before the episode, and then gradually experience a ramp depolarization in the interepisode interval (Figure 1(a)). Motoneurons no longer fire action potentials after an episode and only begin to fire minutes into the interval as the resting potential begins to recover to pre-episode levels. In addition, synaptic strength of many of the connections is weakened following the episode. Reflex synaptic responses, as well as miniature postsynaptic potentials thought to represent responses to a single vesicle of transmitter, are reduced in amplitude following the episode. As the cord recovers from the previous episode and the resting membrane potential depolarizes and synaptic strength increases, the spinal networks are once again competent to generate another episode.

Glutamatergic, nicotinic, glycinergic, and GABAergic transmitters all contribute to the drive underlying the episodes, and if any one of these transmitters is blocked the other transmitters will compensate for the loss. Very early in development (E4.5) nicotinic transmission and, to a lesser degree, GABAergic transmission, are most important for generating SNA. From E8 on, however, GABA appears to be the dominant transmitter in setting the frequency of episode occurrence. Why is GABAergic transmission so important to episode frequency? This is due to the modulatory nature of chloride-mediated $GABA_A$ currents, which significantly contribute to

the episode-induced depression and recovery of spinal excitability. During an episode of activity $GABA_A$ channels open and allow chloride ions to flow out of the cell to the extent that the reversal potential is altered and the driving force for chloride-mediated conductances is reduced. This is why evoked and miniature synaptic currents are smaller following an episode. Synaptic strength then recovers in the interepisode interval because chloride pumps (presumably NKCC1) progressively restore the higher intracellular chloride concentration, and therefore restore driving force for the GABAergic currents. Intracellular chloride modulation may also contribute to the hyperpolarization and recovery of the membrane potential following an episode, because this process is mediated by a tonic $GABA_A$ current. However, chloride regulation cannot completely account for the ramp depolarization because the tonic GABAergic current is completely abolished following the episode, even though $GABA_A$ agonists puffed on the cord after an episode produce a significant depolarization. This suggests that extracellular GABA builds up in the interval and then is cleared during the episode. Therefore GABA is critically important in the periodicity of SNA because it contributes significantly to the episode-induced depression and recovery of spinal excitability, through the ramp depolarization and chloride-mediated synaptic strengthening during the interepisode interval.

The embryonic movements driven by SNA are known to be important in bone, muscle, and joint development. Further, motoneuron pathfinding into the limb is dependent on normal levels of SNA and the development of locomotor circuits within the cord is thought to be dependent on nicotinic transmission associated with SNA. In addition, recent work has shown that SNA regulates the strength of both excitatory and soon-to-be inhibitory synapses. SNA is thought to provide the activity that allows the network to assess the synaptic strength of its excitatory and inhibitory components and thus adjust the gain of these synapses in order to maintain the activity at appropriate levels. In this way spontaneous activity can act to coordinately mature glutamatergic and GABAergic synaptic strength.

Retina

Immature retinal ganglion cells (RGCs) spontaneously fire bursts of action potentials. The spontaneous activity propagates across the ganglion cell layer, a pattern referred to as 'retinal waves.' Retinal waves have been detected in a broad range of vertebrate species, including chicks, turtles, mice, rats, rabbits, ferrets, cats, and primates. Though extensive studies have been conducted in most of these species, the focus here is on the mechanisms described for mammals, wherein retinal waves are detected during an extended period prior to the onset of light responses. That said, the spatial and temporal properties of retinal waves, and the synaptic circuits, are remarkably similar across all species.

Retinal waves initiate in random locations and propagate from one cell to another, correlating the firing of hundreds of cells (Figure 2). Though individual retinal waves have complex propagation properties, the firing properties of individual retinal ganglion cells is fairly stereotyped, with each cell firing a 1 to 2 s burst of action potentials that are separated by 1–2 min of silence. Retinal waves propagate with a speed of 100–200 $\mu m\, s^{-1}$. Following a retinal wave, a local region of the retina will go into a refractory period, during which retinal waves cannot be initiated. This refractoriness, a robust feature of the network, lasts 30–45 s and contributes to the complex geometry of the progression of the wave front.

Retinal waves persist for a long period of time perinatally in rodents and rabbits, first detected a few days before birth and lasting 2–3 weeks. During this time, the retina is rapidly developing, with dramatic changes in functional synaptic circuits. When retinal waves are first detected, an immature inner retina, containing synaptically connected RGCs and amacrine cells, has formed, but the ventricular zone still comprises most of the retina. By eye-opening (P14), all of the cell layers are fully developed and light responses are robust.

The maturation of retinal circuitry is reflected in the changing synaptic mechanisms that mediate waves. Retinal waves are divided into three distinct stages based upon the mechanisms that generate them (Figure 2). Stage I retinal waves are mediated before the formation of synaptic connections and these waves end at birth in mice, when amacrine cells and RGCs are synaptically connected, as determined by electron microscopy and electrophysiology. The mechanisms underlying stage II retinal waves are the most extensively studied. Stage II retinal waves are blocked by nicotinic acetylcholine receptors (nAChRs) antagonists. RGCs receive both cholinergic and GABAergic synaptic input during retinal waves. Only one class of cell in the retina, the starburst amacrine cell, releases acetylcholine (ACh). Very recent experiments in rabbit retina indicate that the properties of starburst amacrine cells and the starburst amacrine cell network are sufficient to explain most characteristics of stage II retinal waves. Starburst amacrine cells undergo spontaneous depolarizations, neighboring starburst amacrine cells monosynaptically release both ACh and GABA onto one another, and each starburst amacrine

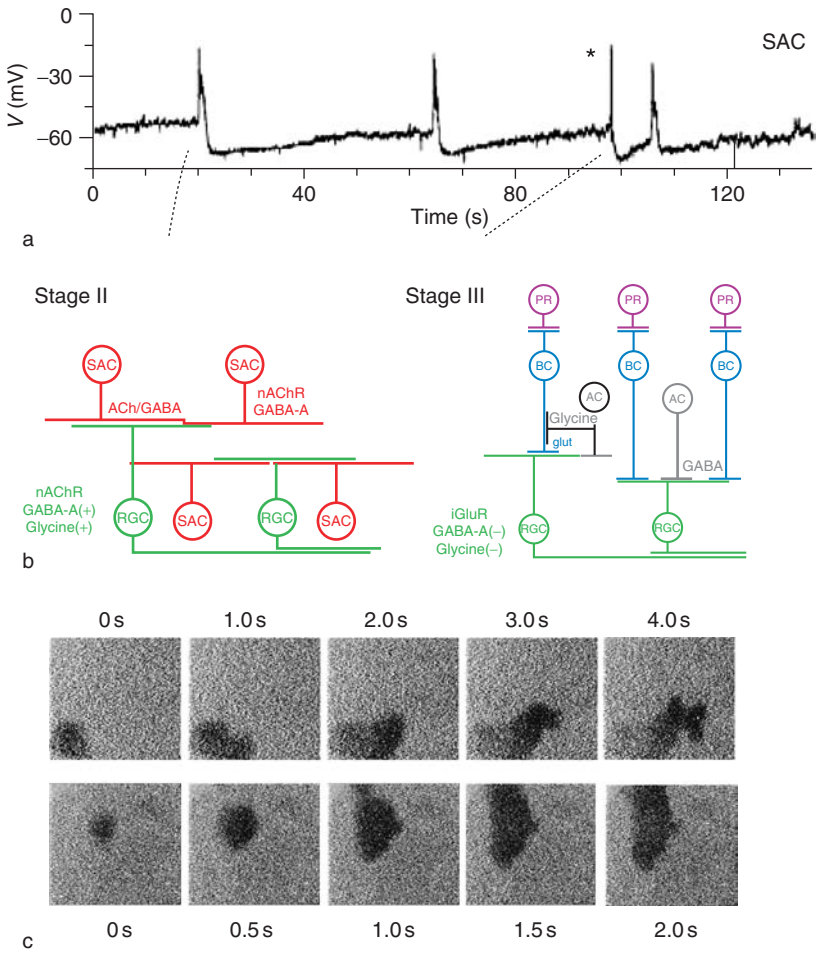

Figure 2 (a) Retinal waves. Current clamp recording from starburst amacrine cell (SAC), showing rhythmic compound bursts associated with retinal waves. (b) Schematic of wave-generating circuitry. Left: During stage II retinal waves, the cholinergic starburst amacrine cells (SAC) synapse onto one another and onto retinal ganglion cells (RGC). Not pictured are photoreceptors and bipolar cells, which are not yet connected to RGCs. RGCs and SACs receive excitatory synaptic inputs through nicotinic acetylcholine receptors (nAChRs) and γ-aminobutyric acid $(GABA)_A$ receptors. SACs release both ACh and GABA. Right: During stage III retinal waves, the retina's vertical circuit is complete. Calcium transients in RGCs are blocked by ionotropic glutamate receptor antagonists and are made more frequent by $GABA_A$ and glycine receptor antagonists. Bipolar cells (BC) synapse onto RGCs; inhibitory amacrine cells (AC) synapse onto RGCs and bipolar cells (PR, photoreceptor). (c) Time evolution of two retinal waves occurring in the same region 30 s apart. Dark areas indicate decreases in fura-2 fluorescence associated with the increased calcium evoked by waves. Images were acquired with a silicon intensifier target (SIT) camera. An initial background frame was acquired and was subtracted on a pixel-by-pixel basis from all subsequent frames to create a difference image. Each frame is 1.2×1.2 mm. (a) From Zheng J, Lee S, and Zhou ZJ (2006) A transient network of intrinsically bursting starburst cells underlies the generation of retinal waves. *Nature Neuroscience* 9: 363–371. (b) Reproduced from Firth SI, Wang C-T, and Feller MB (2005) Retinal waves: Mechanisms and function in visual system development. *Cell Calcium* 37(5): 425–432, with permission from Elsevier. (c) From Feller MB, Wellis DP, Stellwagen D, et al. (1996) Requirement for cholinergic synaptic transmission in the propagation of spontaneous retinal waves. *Science* 272: 1182–1187.

cell receives input from around 20 other starburst amacrine cells. (Starburst cells are one of a few neurons that secrete two neurotransmitters. GABA and ACh are not thought to be released from the same vesicles, or even the same location on starburst amacrine cell processes, but their release mechanisms have not yet been elucidated.) This connectivity among starburst amacrine cells may provide the substrate for wave propagation. Starburst amacrine cells also synapse onto RGCs, such that waves of depolarization propagating through the starburst amacrine cell network also activate nAChRs and $GABA_A$ receptors on RGCs, causing them to fire correlated bursts of action potentials. Though these findings have yet to be reproduced in other species, it is likely to be generalized and therefore is the best current model of retinal wave generation.

There are still many aspects of stage II retinal waves that are the focus of ongoing research efforts. During stage II retinal waves, activation of $GABA_A$ receptors

on retinal ganglion cells is depolarizing. However, $GABA_A$ receptor antagonists modulate but do not block retinal waves. Hence it has been hypothesized that GABA contributes to the depolarization of retinal ganglion cells, but not the propagation of activity through the starburst amacrine cell network. In addition, in many species, retinal waves are blocked by either gap junction receptor or adenosine receptor antagonists. Last, the spatiotemporal properties of retinal waves are strongly modulated by drugs that modulate the cyclic adenosine monophosphate/protein kinase A (cAMP/PKA) second messenger pathway by mechanisms that are not yet understood.

The circuits mediating stage III waves are also a topic of ongoing research. Stage III waves emerge when glutamatergic synapses between bipolar cells and retinal ganglion cells are first formed. The circuitry underlying stage III waves is known to be distinct from that mediating stage II waves. Ionotropic glutamate receptor antagonists but not nAChR antagonists block stage III retinal waves. In rabbits, during stage III waves, starburst amacrine cells no longer undergo spontaneous depolarizations and nAChRs are downregulated in starburst amacrine cells, indicating the starburst network is not the source of wave initiation or propagation. The observation that stage III waves are blocked by ionotropic glutamate receptor antagonists indicates that synaptic release from bipolar cells, the primary source of glutamatergic input to retinal ganglion cells, is a critical component of the network mediating stage III waves. However, bipolar cells are not directly coupled to each other and it is not known whether they function as pacemakers at this stage of development. Hence, many details regarding the mechanisms underlying stage III waves remain to be discovered.

Retinal waves are required for normal establishment of circuits with in the visual system. Many of the highly ordered connections in the visual system are refined at an early stage of development, prior to normal visual experience. For example, upon initial growth to the superior colliculus, retinal projections are broadly distributed. Subsequently, retinocollicular projections refine to form retinotopic maps. Similarly, upon initial growth to the lateral geniculate nucleus of visual thalamus, retinal projections from the two eyes significantly overlap. Subsequently, retinogeniculate axons refine into eye-specific regions. Blockade of retinal activity blocks the normal refinement process, and transgenic mice that do not have normal patterns of spontaneous activity do not undergo the normal refinement process. Though these findings indicate a role for spontaneous retinal activity, it remains a controversial question as to which of the unique spatiotemporal properties of retinal waves are required for driving refinement. Similar experimental paradigms indicate that early retinal activity may also be important for driving the establishment of adult circuitry in primary visual cortex as well. Hence, retinal waves have become a primary model for understanding how early neural activity drives the formation of adult circuitry before the onset of sensory experience.

Hippocampus

Spontaneous periodic network events are also a characteristic feature of the developing hippocampus. In the immature hippocampus, these events are referred to as giant depolarizing potentials (GDPs). GDPs have been recorded *in vivo* and extensively in *in vitro* hippocampal slices. GDPs invade all subfields of the hippocampus and cause correlated depolarization of both pyramidal cells as well as depolarizing inhibitory interneurons.

GDPs initiate in the CA3 subregion, with some recent evidence suggesting that they are initiated by intrinsic pacemaker activity of pyramidal cells. Activity in the CA3 region is propagated through the hippocampus via immature GABAergic and glutamatergic synapses. The pattern of spontaneous firing is more frequent than that previously described for spinal cord and retina. Both *in vivo*, and *in vitro*, GDPs last 0.5–3 s and occur every 2–10 s. There is no neuronal firing in between these population events, as observed in spinal cord and retina, but in sharp contrast to the adult hippocampus.

The generation of GDPs is due to synergistic interactions between immature $GABA_A$ and *N*-methyl-D-aspartate (NMDA) receptors. Blockade of either type of receptor prevents GDPs. Transmission through immature $GABA_A$ receptors is excitatory. Glutamate synapses are also immature during the developmental period when GDPs are expressed. In the mature hippocampus, glutamatergic synapses contain a mixture of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) and NMDA receptors. However, during development, there is a large percentage of NMDA-only synapses with this level, decreasing during the first postnatal week. NMDA receptors are considered 'coincidence detectors' because activation of these receptors requires both glutamate binding and depolarization of the membrane. It has been postulated that depolarization via $GABA_A$ receptors acts to relieve the Mg^{2+} block of NMDA receptors. Hence, excitatory GABAergic synapses from interneurons and glutamatergic inputs from pyramidal cells contribute to the generation of GDPs.

The immature pattern of $GABA_B$ receptors is also important in shaping GDPs. $GABA_B$ receptors are

G-protein-coupled receptors that activate potassium channels and therefore hyperpolarize membranes. $GABA_B$ receptors are absent on postsynaptic membranes during the first postnatal week, adding to the excitability of the networks. However, presynaptic $GABA_B$ receptors are mature, and therefore control the release of GABA and glutamate. Hence, blocking $GABA_B$ receptors dramatically changes the structure of GDPs such that they look more like seizure events.

Several functions for GDPs in the development of mature excitatory and inhibitory transmission have been postulated and are the focus of current research efforts. Some studies indicate that GDPs are critical for the normal maturation of glutamatergic and GABAergic synapses in at least three ways. First, GDPs are likely important in determining the timing of pre- and postsynaptic activity, which appears to be important in mediating the developmental switch for $GABA_A$ receptors as they transition from excitatory to inhibitory. Second, depolarization via $GABA_A$ receptors leads to the phosphorylation of certain transcription factors, which may be critical for development. Third, the activation of NMDAR-only synapses through the synergistic actions of $GABA_A$ and NMDA may provide that critical signaling necessary to drive the transport of AMPA receptor subunits. This 'AMPA-fication' of glutamatergic synapses is a critical step in the formation of adult hippocampal circuitry.

Conclusions

Following the formation of the first synaptic connections, developing networks express a spontaneous correlated activity that shares many characteristics in very different networks, while each also expresses its own unique features. We now consider the similarities and differences in the three networks discussed.

In spinal cord, retina, and hippocampal tissues, episodes of activity lasts seconds and are followed by long episodes of quiescence. During the episodes of activity, many cells experience depolarizations and increases in intracellular calcium, known to be important in several aspects of development. Depolarizing GABAergic transmission is critical for the generation of episodes of activity, especially in the spinal and hippocampal networks. Following the activity, the network is reduced in its excitability, though the source of this network depression may be different in the different circuits.

The existence of the spontaneous activity in multiple species and the similarities in the activity in different parts of the nervous system suggest a general importance of the activity toward some common functions. There is growing evidence for a role for spontaneous activity in the developing nervous system, by refining the connectivity of projection neurons, the development of local circuitry, or the activation of second messenger cascades that drive long-lasting transcriptional events. In the spinal cord, SNAs have been implicated in both how motoneurons select targets and in the balance of excitatory and inhibitory synaptic strength in local circuits. Retinal waves have a well-established role in influencing projections of retinal ganglion cells, but a role for the activity in the retina itself has not yet been elucidated. Hippocampal GDPs also have been implicated in the development of local circuitry and may be critical for establishing the synaptic relationships between the different subregions of the hippocampus.

One of the remaining controversies is whether spontaneous activity is 'instructive' or 'permissive' during these developmental events. An instructive role indicates that all of the relevant information driving a particular development process is contained in the firing pattern. For example, do SNAs influence targeting of motoneurons by ensuring that neighboring motoneurons are more correlated in their firing than are motoneurons in a different segment (an instructive signal)? Or do SNAs exert their actions by maintaining a certain average firing rate that allows different molecular interactions between motoneurons and their targets to take place (a permissive signal)? There are some clear examples where at least some features of the spontaneous firing patterns matter. For example, changing the frequency of SNAs influences the expression of different molecules. In another example, in transgenic mice that disrupt the endogenous pattern of retinal waves to disrupt nearest-neighbor correlations but maintain similar average firing rates, both retinotopic and eye-specific refinements are inhibited. Experiments with targeted manipulations that alter specific features of spontaneous activity patterns will determine what features of this early network activity are critical for the normal development of adult neural circuits. Hence, understanding the cellular basis of the different features of spontaneous activity patterns will be critical for designing these future studies.

See also: Enteric Nervous System: Neural Circuits and Chemical Coding; Gamma-Aminobutyric Acid (GABA); Hippocampus; Neural Circuitry in the Somatosensory System; Retinal Development: An Overview.

Further Reading

Ben-Ari Y, Khalilov I, Represa A, et al. (2004) Interneurons set the tune of developing networks. *Trends in Neurosciences* 27: 422–427.

Chub N and O'Donovan MJ (2001) Post-episode depression of GABAergic transmission in spinal neurons of the chick embryo. *Journal of Neurophysiology* 85: 2166–2176.

Feller MB, Wellis DP, Stellwagen D, et al. (1996) Requirement for cholinergic synaptic transmission in the propagation of spontaneous retinal waves. *Science* 272: 1182–1187.

Firth SI, Wang C-T, and Feller MB (2005) Retinal waves: Mechanisms and function in visual system development. *Cell Calcium* 37(5): 425–432.

Gonzalez-Islas C and Wenner P (2006) Spontaneous network activity in the embryonic spinal cord regulates AMPAergic and GABAergic synaptic strength. *Neuron* 49: 563–575.

Hanson MG and Landmesser LT (2004) Normal patterns of spontaneous activity are required for correct motor axon guidance and the expression of specific guidance molecules. *Neuron* 43: 687–701.

Le van Quyen M, Khalilov I, and Ben-Ari Y (2006) The dark side of high-frequency oscillations in the developing brain. *Trends in Nuerosciences* 29(7): 419–427.

Leinekugel X (2003) Developmental patterns and plasticities: The hippocampal model. *Journal of Physiology – Paris* 97: 27–37.

Milner LD and Landmesser LT (1999) Cholinergic and GABAergic inputs drive patterned spontaneous motoneuron activity before target contact. *Journal of Neuroscience* 19: 3007–3022.

Morgan JI and Wong RO (2006) Development of cell types and synaptic connections in the retina. In: Kolb H, Fernandez E, and Nelson R (eds.) *WEBVISION: The Organization of the Retina and the Visual System*. Available at http://webvision.med.utah.edu/.

Myers CP, Lewcock JW, Hanson MG, et al. (2005) Cholinergic input is required during embryonic development to mediate proper assembly of spinal locomotor circuits. *Neuron* 46: 37–49.

Sipila ST, Huttu K, Voipio J, et al. (2006) Intrinsic bursting of immature CA3 pyramidal neurons and consequent giant depolarizing potentials are driven by a persistent Na^+ current and terminated by a slow Ca^{2+}-activated K^+ current. *European Journal of Neuroscience* 23: 2330–2338.

Spitzer NC (2002) Activity-dependent neuronal differentiation prior to synapse formation: The functions of calcium transients. *Journal of Physiology – Paris* 96: 73–80.

Torborg CL and Feller MB (2005) Spontaneous patterned retinal activity and the refinement of retinal projections. *Progress in Neurobiology* 76: 213–235.

Wenner P and O'Donovan MJ (2001) Mechanisms that initiate spontaneous network activity in the developing chick spinal cord. *Journal of Neurophysiology* 86: 1481–1498.

Wong ROL (1999) Retinal waves and visual system development. *Annual Review of Neuroscience* 22: 29–47.

Zheng JJ, Lee S, and Zhou ZJ (2004) A developmental switch in the excitability and function of the starburst network in the mammalian retina. *Neuron* 44: 851–864.

Zheng J, Lee S, and Zhou ZJ (2006) A transient network of intrinsically bursting starburst cells underlies the generation of retinal waves. *Nature Neuroscience* 9: 363–371.

Sporadic Degenerative Ataxias and the Dominantly Inherited Spinocerebellar Ataxias

T Klockgether, University of Bonn, Bonn, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The sporadic degenerative ataxias, formerly named sporadic olivopontocerebellar atrophies, and the autosomal dominantly inherited spinocerebellar ataxias (SCAs) are adult-onset progressive diseases in which ataxia is the prominent clinical manifestation. In most of these diseases, ataxia is the consequence of degeneration of the cerebellar cortex and its afferent or efferent fiber connections. Additional pathology may be found in other parts of the nervous system, specifically in the brain stem, basal ganglia, spinal cord, and the peripheral nervous system. While the sporadic degenerative ataxias are diseases of unknown etiology, the SCAs are caused by dominant mutations of single genes, many of which have been identified. Both groups of diseases are distinct from the acquired ataxias, in which ataxia is due to a defined exogenous or endogenous nongenetic cause.

Patients with sporadic degenerative ataxia fall into two groups. In one group, the underlying pathology is that of multiple system atrophy (MSA), involving widespread degeneration in the central nervous system with presence of glial cytoplasmic inclusions (GCIs). In another group of patients, pathology and clinical manifestations remain largely confined to the cerebellum. At present, it is not known whether the latter manifestations represent a disease entity or rather are a collection of phenotypically similar syndromes. We use the name 'sporadic adult-onset ataxia of unknown etiology' (SAOA), while others prefer the older term 'idiopathic late-onset cerebellar ataxia' (IDLOCA).

Pathogenesis of MSA

MSA is a sporadic, adult-onset neurodegenerative disease that encompasses the former disease categories sporadic olivopontocerebellar atrophy, striatonigral degeneration, and Shy–Drager syndrome. Reasons to combine these diverse categories under the heading of MSA were the large clinical overlap between these disorders and the discovery that the brains of individuals affected by any of them contained GCIs, which are now considered as a highly specific marker of MSA (Figure 1). According to the clinical presentation, a cerebellar (MSA-C) and a parkinsonian (MSA-P) type of MSA are distinguished.

MSA is characterized by widespread degeneration in the central nervous system, with a particularly strong involvement of the cerebellar cortex, dentate nucleus, pontine nuclei, inferior olive, putamen, caudate nucleus, pallidum, substantia nigra, subthalamic nucleus, thalamus, pyramidal tract, preganglionic sympathetic neurons of the intermediolateral cell columns of the spinal cord, anterior horn, and Onuf's nucleus. GCIs are mainly found in oligodendroglial cells of the motor cortex, white matter underlying the cortex, corpus callosum, pyramidal tract, basal ganglia, pons, cerebellar white matter, cerebellar peduncles, and spinal cord. GCIs are accumulations of fibrils that are predominantly formed by polymerized α-synuclein (Figure 1). Although less prominent, there are also α-synuclein-positive inclusions in the cytoplasm and nuclei of neurons.

Although the etiology of MSA remains unknown, recent experimental work has provided new insights into the molecular pathogenesis. Overexpression of human α-synuclein under a myelin basic protein promoter in mice led to the formation of α-synuclein-positive inclusions in oligodendrocytes, astrogliosis of white matter tracts, and neurodegenerative changes in the basal ganglia and neocortex. In another mouse model, overexpression of α-synuclein in oligodendrocytes led to glial inclusions, neuronal cell loss, and progressive motor impairment. Both animal models support the view that accumulation of α-synuclein in oligodendrocytes contributes to the neurodegeneration in MSA. Together, the available autopsy findings and the observations in transgenic animal models suggest that the neurodegeneration of MSA is due to a profound disturbance of the microenvironment formed by oligodendroglia, axons, and neurons, with α-synuclein as a key player in the interaction between these cellular elements.

Clinical Features of MSA

MSA is a rare disorder, with an estimated prevalence of 4.4 in 100 000. Clinically, MSA patients present with various combinations of parkinsonism, cerebellar ataxia, pyramidal tract dysfunction, and autonomic failure. In a large series of European MSA cases, 87% had parkinsonism, 74% had autonomic failure, 54% had cerebellar ataxia, and 49% had pyramidal tract dysfunction. In 58%, the disease

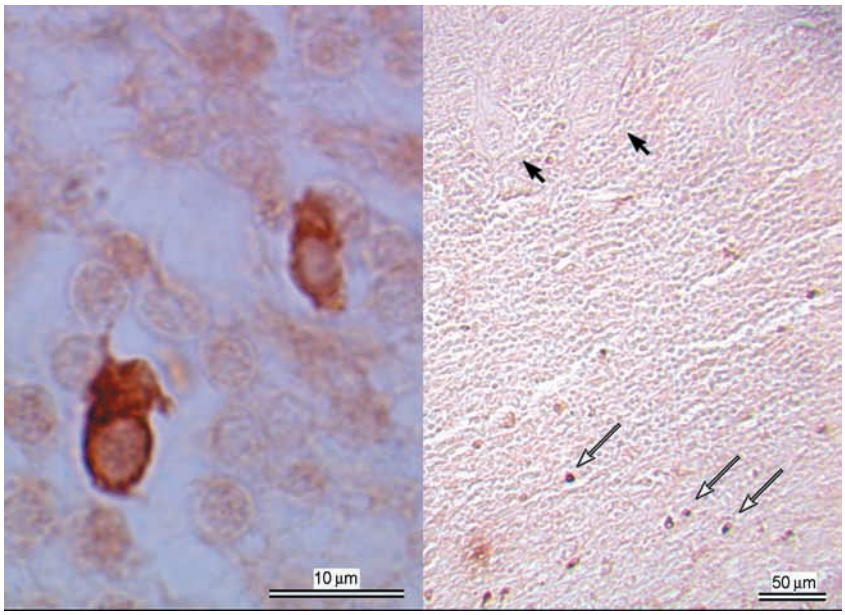

Figure 1 GCIs in cerebellar white matter of a MSA patient.

started with parkinsonism, while cerebellar ataxia was the first symptom in 29%. Interestingly, cerebellar ataxia is more prevalent among Japanese MSA patients. In a recent clinical study of 230 Japanese MSA patients, 67% were classified as MSA-C.

Most MSA patients with parkinsonism have an akinetic–rigid syndrome. When present, tremor is often irregular with myoclonic jerks. The ataxic syndrome in MSA points to a diffuse involvement of the cerebellum, with ataxia of gait and stance, ataxia of limb movements, action tremor, ataxic speech, and cerebellar oculomotor disturbances. This syndrome is undistinguishable from that in other degenerative disease of the cerebellum. Pyramidal involvement is usually manifested by increased muscle reflexes and extensor plantar responses. Proper spasticity is rarely encountered.

Orthostatic hypotension and urinary incontinency are the major manifestations of autonomic failure in MSA. Orthostatic hypotension is due to degeneration of preganglionic sympathetic neurons in the intermediolateral cell columns of the spinal cord, which results in an impairment of the baroreflex. A drop of systolic blood pressure of more than 20 mmHg is considered pathological. Symptoms of orthostatic hypotension may range from dizziness to syncope, with loss of consciousness, and are often precipitated by heavy meals. Orthostatic hypotension in MSA may be associated with increased blood pressure in supine position. Neurogenic bladder dysfunction is frequent in MSA and present in almost all patients in advanced disease stages. In male, it is always combined with erectile dysfunction. Typical clinical manifestation of neurogenic bladder dysfunction is urinary incontinence due to disturbed sphincter function caused by degeneration of preganglionic sympathetic neurons of the spinal cord and Onuf's nucleus. Many patients have additional urinary retention which is caused by parasympathetic dysfunction. MSA starts between 50 and 60 years of age and takes an unrelentlessly progressive course. After 5 to 6 years most MSA patients require a wheelchair. Median survival is 9 years. Death typically occurs due to bronchopneumonia.

Magnetic resonance imaging (MRI) has been extensively used to study the brain morphology of MSA patients *in vivo*. MRI abnormalities occurring in MSA include progressive atrophy of infra- and supratentorial brain structures as well as signal abnormalities. Morphometric studies reveal volume loss of the cerebellum, brain stem, putamen, and cortical areas, in particular the insular cortex. The signal changes typical for MSA are best seen on T_2-weighted images obtained by MRI scanners with high-field strength. They include increased signal in the middle cerebellar peduncles and the pons, decreased signal in the putamen, and a rimlike hyperintensity lateral to the putamen (Figure 2). These abnormalities have high specificity, but low sensitivity for the diagnosis of MSA. The apparent diffusion coefficient is increased in the putamina of MSA patients, suggesting neuronal loss with enlargement of the extracellular space. Positron emission tomography (PET) studies showed reduced metabolic activity in the basal ganglia, brain stem, and cerebellum.

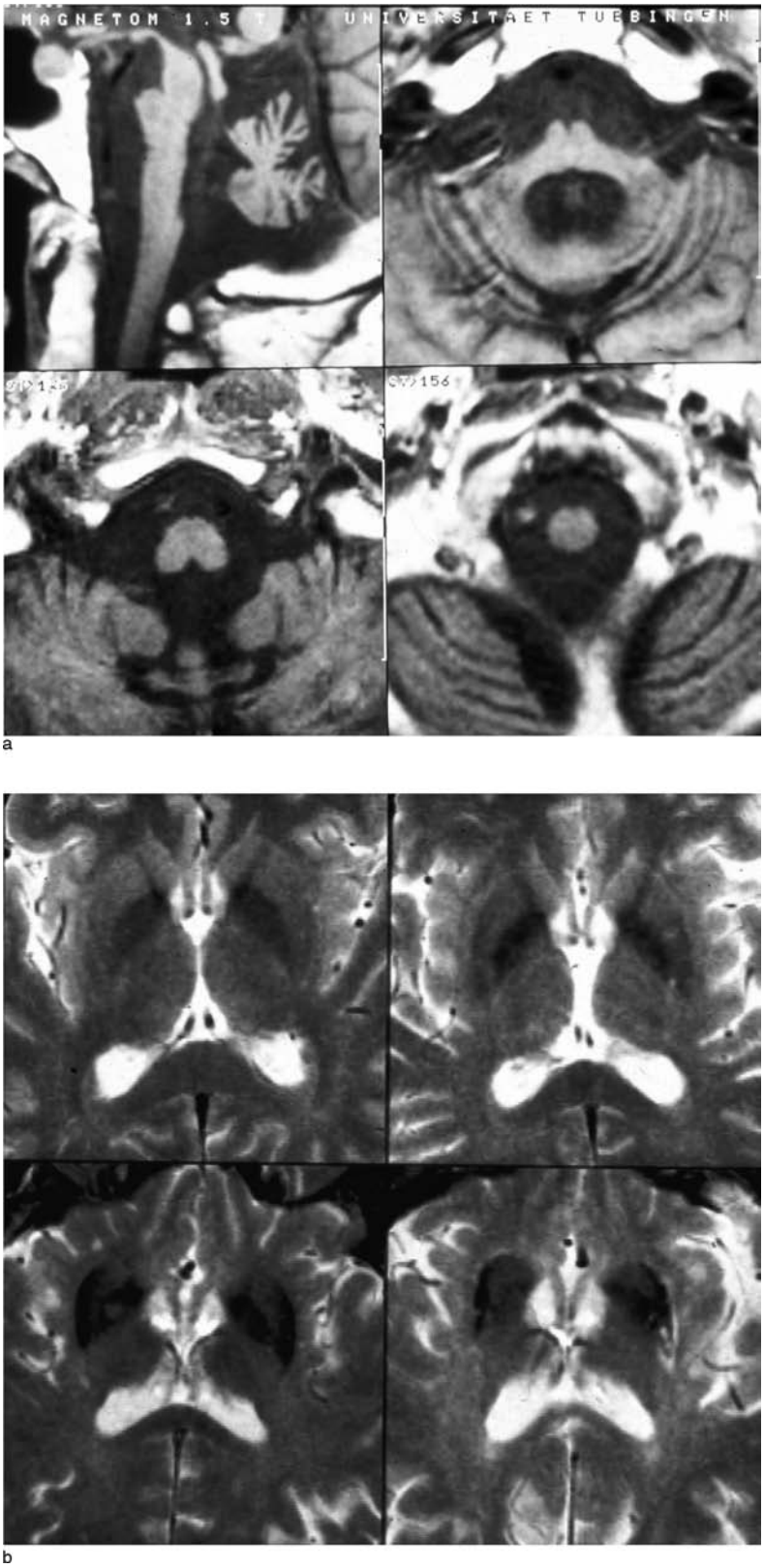

Figure 2 MRI in MSA. (a) T_1-weighted images with severe atrophy of the cerebellum and brain stem in an MSA-C patient. (b) Signal abnormalities in the basal ganglia on T_2-weighted images in MSA. Left upper panel: Normal. Right upper panel: Slight hypointensity of the putamina in a case of beginning MSA. Left lower panel: Mild atrophy of the basal ganglia and marked hypointensity of the putamina in an MSA-P patient. Right lower panel: Marked atrophy of the basal ganglia, marked hypointensity of the putamina, and unilateral rimlike hyperintensity lateral to the basal ganglia in a patient with advanced MSA.

There is no curative or preventive treatment for MSA. Parkinsonian symptoms may respond to levodopa, but the response is often transient and usually less robust than in idiopathic Parkinson's disease. Autonomic symptoms are treated in standard manner. There is no effective symptomatic treatment for ataxia.

Sporadic Adult-Onset Ataxia of Unknown Etiology

SAOA is a sporadic adult-onset ataxia disorder. It is clinically distinguished from MSA by the lasting absence of parkinsonism and severe autonomic failure. Epidemiological studies and clinical case series suggest that SAOA occurs with twice the frequency of MSA. An epidemiological survey in Wales found a prevalence of 8.4 in 100 000.

The etiology and pathogenesis of SAOA are unknown, and there is only limited knowledge of the neuropathology of this disorder. The few available autopsy studies suggest that the degeneration mainly affects the cerebellar cortex with or without involvement of the inferior olives. This degeneration pattern was previously designated as late cerebellar cortical atrophy. GCIs are not found in SAOA brains.

SAOA begins around the age of 50 years and affects men more frequently than women. Clinical presentation is characterized by progressive cerebellar ataxia accompanied by signs of peripheral neuropathy and pyramidal tract dysfunction in half of the patients. At first presentation, the differentiation between MSA-C and SAOA is often impossible. However, the more benign course and the absence of severe autonomic failure and parkinsonism allow differentiation of SAOA from MSA. The majority of SAOA patients are still ambulant 10 years after onset of ataxia. Median survival ranges from 20 to 25 years, resulting in an almost normal life expectancy. On MRI, SAOA patients usually have isolated cerebellar atrophy with little or no involvement of the brain stem. Currently, there are no treatment options for SAOA.

Genetics and Pathogenesis of SCAs

The SCAs are a genetically and clinically heterogeneous group of autosomal dominantly inherited progressive ataxia diseases. To date, almost 30 different gene loci have been found (Table 1). Of the mutations that have been identified, SCA1, SCA2, SCA3, SCA6, SCA7, and SCA17 are translated CAG repeat expansions coding for an elongated polyglutamine tract within the respective proteins. These diseases belong to a larger group of polyglutamine disorders that also include Huntington's disease, dentatorubro-pallidoluysian atrophy, and spinobulbar muscular atrophy. Three SCAs, SCA8, SCA10, and SCA12, are caused by untranslated repeat expansions in noncoding regions of the respective genes. In SCA5 (gene name: *spectrin, beta, non-erythrocytic 2*; *SPTBN2*), SCA13 (*potassium voltage-gated channel*; *KCNC3*), SCA14 (*protein kinase C, gamma*; *PRKCG*), SCA27 (*fibroblast growth factor 14*; *FGF14*), and 16q22-linked autosomal dominant cerebellar ataxia (*pleckstrin homology domain containing, family G*, or *puratrophin-1*; *PLEKHG4*), point mutations have been found in the respective genes. In all other SCAs, the affected genes and mutations have not yet been identified.

The genetic heterogeneity of the SCAs implies an even greater heterogeneity of the molecular pathogenesis. There has been some recent progress in the understanding of the pathogenesis of the polyglutamine disorders, while there is still very limited knowledge of the mechanisms that underlie the SCAs due to untranslated repeat expansions and to point mutations.

All polyglutamine diseases share common pathogenetic pathways which involve the deleterious actions of the expanded polyglutamine tracts. On the other hand, it is increasingly recognized that the actions of the respective disease proteins and their specific alterations due to mutations significantly contribute to the development of the disease phenotype. This dual pathogenetic model explains the common features shared by all polyglutamine diseases as well as the unique features of each particular disease.

That expanded polyglutamine tracts exert powerful neurotoxic actions has been convincingly shown in transgenic models in which overexpression of expanded polyglutamines in arbitrary proteins in neurons caused neurodegeneration. Several observations suggest that cleavage of the disease proteins and formation of smaller polyglutamine-containing protein fragments are prerequisites for polyglutamine-mediated neurodegeneration. However, the enzymes mediating protein cleavage have not yet been fully identified. In most polyglutamine diseases, a nuclear localization of the disease protein is required for its deleterious action.

Elongated polyglutamines have a strong tendency to misfold and aggregate. According to current view, oligomeric intermediates originating from the aggregation process are responsible for the pathogenesis, while the visible (via the light microscope) neuronal inclusions found in neurons of SCA brains are thought to be inert or even protective. Oligomeric intermediates exert their deleterious action by abnormally interacting with other cellular proteins. Potential interaction partners are transcription factors, which contain polyglutamine tracts. As a result,

Table 1 SCAs: molecular genetics and clinical phenotype

Disorder	*Mutation*	*Gene product*	*Clinical phenotype*
SCA1	Translated CAG repeat expansion	Ataxin-1	Ataxia, pyramidal signs, neuropathy, dysphagia, restless leg syndrome
SCA2	Translated CAG repeat expansion	Ataxin-2	Ataxia, slow saccades, neuropathy, restless leg syndrome
SCA3 (Machado–Joseph disease)	Translated CAG repeat expansion	Ataxin-3	Ataxia, pyramidal signs, ophthalmoplegia, neuropathy, dystonia, restless legs syndrome
SCA4	Unknown	Unknown	Ataxia, sensory neuropathy
SCA5	Point mutation	Beta-III spectrin (SPTBN2)	Almost purely cerebellar ataxia
SCA6	Translated CAG repeat expansion	Calcium channel subunit (CACNA1A)	Almost purely cerebellar ataxia
SCA7	Translated CAG repeat expansion	Ataxin-7	Ataxia, ophthalmoplegia, visual loss
SCA8	3′-Untranslated CTG repeat expansion	Ataxin-8	Ataxia, sensory neuropathy, spasticity
SCA10	Intronic ATTCT repeat expansion	Ataxin-10	Ataxia, epilepsy
SCA11	Unknown	Unknown	Almost purely cerebellar ataxia
SCA12	5′-Untranslated CAG repeat expansion	Phosphatase subunit (PP2A-PR55β)	Ataxia, tremor
SCA13	Point mutation	Potassium channel (KCNC3)	Ataxia, mental retardation
SCA14	Point mutation	Protein kinase Cγ (PKCγ)	Ataxia, myoclonus, dystonia, sensory loss
SCA15	Unknown	Unknown	Almost purely cerebellar ataxia
SCA16	Unknown	Unknown	Almost purely cerebellar ataxia
SCA17	Translated CAG repeat expansion	TATA-binding protein (TBP)	Ataxia, dystonia, chorea, dementia, psychiatric abnormalities
SCA18	Unknown	Unknown	Ataxia, sensory neuropathy, neurogenic muscle atrophy
SCA19	Unknown	Unknown	Ataxia, myoclonus, cognitive impairment
SCA20	Unknown	Unknown	Ataxia, dysphonia
SCA21	Unknown	Unknown	Ataxia, parkinsonism
SCA23	Unknown	Unknown	Ataxia, sensory neuropathy, pyramidal signs
SCA25	Unknown	Unknown	Ataxia, sensory neuropathy
SCA26	Unknown	Unknown	Almost purely cerebellar ataxia
SCA27	Point mutation	Fibroblast growth factor 14 (FGF14)	Ataxia, tremor, mental retardation
SCA28	Unknown	Unknown	Ataxia, ophthalmoparesis, pyramidal signs
16q22-linked ADCA	Point mutation	Puratrophin	Almost purely cerebellar ataxia

a general suppression of gene transcription has been found in models of polyglutamine diseases. The important role of transcriptional dysregulation in the pathogenesis of polyglutamine disorders is underlined by the discovery that SCA17 is caused by a CAG repeat expansion in the gene of an essential transcription factor, TATA-binding protein (TBP).

With the exception of SCA6 (*calcium channel, voltage-dependent, P/Q type, α 1A subunit*; *CACNA1A*) and SCA17 (*TATA box binding protein*; *TBP*), the functions of disease proteins of all SCAs of the polyglutamine group were completely unknown when the respective gene mutations were discovered. The physiological roles of these proteins, which have been named ataxins, are currently under intense investigation. In SCA1, the polyglutamine expansion stabilizes ataxin-1 and leads to its accumulation. Ataxin-1 interacts in a polyglutamine-independent way with the mammalian transcription factor senseless/Gfi-1, which is expressed in cerebellar Purkinje neurons. Loss of Gfi-1 leads to degeneration of cerebellar Purkinje neurons, suggesting that ataxin-1/Gfi-1 interaction contributes to the neurodegeneration in SCA1.

Ataxin-3, the disease protein of SCA3, binds polyubiquitin chains and acts as an ubiquitin protease. In addition, ataxin-3 has been shown to bind to DNA and form a transcriptional repressor complex. The transcriptional repressor function is lost in mutated ataxin-3. Overexpression of normal ataxin-3 prevents polyglutamine-induced neurodegeneration, raising the question whether a loss of the physiological function of ataxin-3 amplifies the polyglutamine-mediated toxicity in SCA3.

In SCA6, the polyglutamine expansion in the α_{1A} voltage-dependent calcium channel subunit impairs the physiological function of Ca^{2+} channels, resulting in reduced Ca^{2+} influx. As a consequence, cells overexpressing mutant Ca^{2+} channels are more susceptible

to cell death. The molecular mechanisms underlying the SCAs that are caused by untranslated repeat expansions, SCA8, SCA10, and SCA12, are poorly understood. Based on observations in a transgenic mouse model it has recently been suggested that SCA8 pathogenesis involves toxic gain-of-function mechanisms at both the protein and RNA levels.

It is generally thought that the pathogenesis of the SCAs caused by point mutations is more closely related to the physiological function of the affected proteins than is that of the SCAs which are due to repeat mutations. An example for this assumption is SCA5. SCA5 is caused by mutations of the *SPTBN2* gene, which is highly expressed in cerebellar Purkinje neurons. The *SPTBN2* protein, beta-III spectrin, acts as a stabilizer of the glutamate transporter EAAT4. Mutations might thus result in abnormal glutamate signaling and excitotoxicity. On the other hand, point mutations of the *PRKCG* gene in SCA14 have been shown to lead to the formation of intracytoplasmic aggregates, suggesting a disease mechanism resembling that in polyglutamine disorders.

Given the large genetic heterogeneity of SCAs it is not surprising that autopsy findings are extremely variable. Even within one genotype, neuropathology is not uniform. SCA1, SCA2, and SCA7 are prototypes of a degeneration pattern that was previously designated as olivopontocerebellar atrophy with degeneration of the cerebellar cortex, pontine nuclei, and inferior olives. In SCA2 brains there is almost always severe loss of dopaminergic neurons of the substantia nigra. The autopsy findings in these disorders may closely resemble those in MSA-C. However, GCIs are absent or present in only small numbers. Instead, in some polyglutamine disorders, particularly in SCA1, SCA3, SCA7, and SCA17, characteristic nuclear inclusions are found in neurons. These inclusions are formed by aggregated disease proteins encoded by the mutated genes. In addition, these inclusions contain additional proteins, such as chaperones, proteasomal subunits, and transcription factors.

In other SCAs, such as SCA6, the degeneration is largely confined to cerebellar Purkinje neurons without major involvement of other parts of the brains. For many of the less frequently occurring SCAs, systematic autopsy studies are lacking.

Clinical Features of SCAs

SCAs are rare disorders. Epidemiological studies conducted in different European regions have found prevalence rates ranging from 0.9–3.0 in 100 000. Due to founder effects, the prevalence is much higher in some geographically isolated areas. One example is the province of Holguin in Cuba, where SCA2 has an estimated prevalence of 43 in 100 000. Worldwide, SCA1, SCA2, SCA3, and SCA6 are the most common SCAs. Depending on the respective geographical region these mutations account for up to 75% of all SCA families. Although ataxia is the prominent symptom of all SCAs, clinical presentation is diverse. Ataxia is usually of cerebellar type, but there are also SCAs with predominantly sensory ataxia, such as SCA4, or with a strong sensory component, such as SCA3.

To account for the clinical heterogeneity of the dominantly inherited ataxias, a simple clinical classification was introduced by Harding. Although this classification has been replaced by the genetic SCA classification, the Harding classification is still useful for clinical purposes. Harding distinguished between dominant ataxias with a complex phenotype in which ataxia is accompanied by varying nonataxia symptoms, which she named autosomal dominant cerebellar ataxia type I (ADCA-I); a less frequent phenotype with a purely cerebellar syndrome, ADCA-III; and a rare phenotype in which ataxia is accompanied by progressive visual loss, ADCA-II. The majority of SCA disorders, including the most prevalent disorders, SCA1, SCA2, and SCA3, fall into the ADCA-I category. SCA6 is the prototype of the ADCA-III category, but other disorders, such as SCA5 and SCA11, are also purely cerebellar disorders. The ADCA-II phenotype is almost exclusively found in SCA7. Some SCA disorders have unusual phenotypes that cannot be readily classified with the Harding classification. These include SCA4, with prominent sensory neuropathy, and SCA10, with epilepsy.

SCA1 patients suffer from a progressive cerebellar syndrome with ataxia of gait and stance, ataxia of limb movements, dysarthria, and cerebellar oculomotor abnormalities. The oculomotor abnormalities include gaze-evoked nystagmus, saccade hypermetria, broken up smooth pursuit, reduced optokinetic nystagmus, and impaired suppression of vestibulo-ocular reflex by fixation. In the majority of patients, there are additional noncerebellar symptoms. About half of the patients have supranuclear gaze paresis and/or saccade slowing. Pyramidal tract signs with spasticity, extensor plantar responses, and hyperreflexia are found in more than 50% of SCA1 patients. Decreased vibration sense is found in up to 80% of the SCA1 patients. Dysphagia is a frequent complaint of SCA1 patients and is a particular clinical problem in late disease stages. Disturbances of sphincter control, mainly bladder dysfunction, occur less frequently, and are encountered in about 20% of the patients. Basal ganglia symptoms with parkinsonism

or dystonia are observed only occasionally. Mental disturbances are encountered in less than 10% of SCA1 patients.

SCA2 patients suffer from a progressive cerebellar syndrome with ataxia of gait and stance, ataxia of limb movements, and dysarthria. Saccade slowing is a highly characteristic feature that is observed in the majority of SCA2 patients. About half of the patients have vertical or horizontal gaze palsy. Cerebellar oculomotor abnormalities are rarely found in SCA2. Typically, tendon reflexes are absent or decreased. Pyramidal tract signs are present in less than 20% of the patients. Vibration sense is decreased in most patients, while sensation is otherwise normal. Atypical SCA2 phenotypes with prominent dementia, an amyotrophic lateral sclerosis (ALS)-like presentation, and levodopa-responsive parkinsonism are also encountered. Often, these patients have shorter expansions than the do more frequent ataxic SCA2 patients.

The clinical picture of SCA3 is characterized by a wide range of clinical manifestations, the precise nature of which partly depends on repeat length. All SCA3 patients suffer from a progressive syndrome with ataxia of gait and stance, ataxia of limb movements, and dysarthria. Vertical or horizontal gaze palsy is a frequent additional finding which occurs independently of age of onset. At least 40% of SCA3 patients have a levodopa-responsive restless leg syndrome. Saccade velocity is usually normal. Patients with a repeat length of more than 74 have an early disease onset and clinical features of pyramidal tract and basal ganglia involvement. Most of these patients have increased tendon reflexes, extensor plantar responses, spasticity, and dystonia. Patients with an intermediate repeat length of 71–74 units have a disease onset in middle age and show mainly ataxia and gaze palsy. Patients with a repeat length of less than 71 have a later disease onset and show signs of peripheral neuropathy with loss of tendon reflexes, amyotrophy, and decreased vibration sense. However, the boundaries between these clinical syndromes are vague, and the clinical phenotype of an individual may change with progression of the disease.

SCA6 patients suffer from a progressive cerebellar syndrome with ataxia of gait and stance, ataxia of limb movements, and dysarthria. Horizontal gaze-evoked nystagmus is almost universally present, and downbeat nystagmus is found in more than half of SCA6 patients. Other cerebellar oculomotor findings are also common, such as impaired smooth pursuit and dysmetric saccades. With disease progression, some SCA6 patients have clinical evidence of non-cerebellar involvement, including pyramidal signs and mild sensory disturbances.

SCA7 is clinically characterized by the combination of ataxia and visual loss. In patients with late disease onset, after the age of 40 years, cerebellar ataxia is the first symptom, followed by progressive loss of vision. In about half of the patients with late disease onset, there is no evidence of retinal degeneration or optic atrophy, suggesting that retinopathy only affects the macula. All patients with earlier disease onset, before the age of 40 years, have visual problems, starting either prior to or at the same time as cerebellar ataxia. A majority of these patients have retinal degeneration; some of them also have optic atrophy. Tendon reflexes are usually brisk. There are a number of additional symptoms that occur in less than half of the patients, and that tend occur more frequently in patients with long disease duration. These symptoms include gaze palsy, dysphagia, hearing loss, and muscle weakness.

Age of onset is highly variable in the various SCA disorders. Even within the same family the age at which ataxia is first noticed may range from childhood to late adulthood. In the polyglutamine disorders, SCA1, SCA2, SCA3, SCA6, SCA7, and SCA17, a part of this variability is explained the CAG repeat length. Patients with shorter alleles have a later disease onset, while the disease in those with longer alleles starts earlier. As the repeat length may increase from generation to generation, there is often a tendency to earlier disease onset in subsequent generations. This phenomenon is termed 'anticipation.' On average, SCA1 and SCA2 start between the ages of 30 and 40 years, SCA3 around 40 years, and SCA6 between 50 and 60 years. All SCA disorders take an unrelentlessly progressive course. On average, patients become wheelchair bound after 15 years and die 20 to 25 years after onset of ataxia.

Like the underlying neuropathology, MRI findings are heterogeneous. The most consistent finding is cerebellar atrophy, but in some SCA disorders with a strong sensory component of ataxia, such as SCA3 and SCA4, cerebellar atrophy may be absent or mild despite severe ataxia. In SCA1, SCA2, and SCA7, there is usually combined cerebellar and brain stem atrophy, resulting in a pattern of infratentorial atrophy similar to that in MSA (**Figure 3**). In contrast, SCA6 is the prototype of a SCA disorder with an isolated cerebellar atrophy on MRI (**Figure 4**). Morphometric MRI studies show additional shrinkage of the cervical spinal cord and volume loss of basal ganglia nuclei.

At present, there are no treatment options for the SCAs. Accompanying symptoms, such as restless legs syndrome, are treated in standard manner.

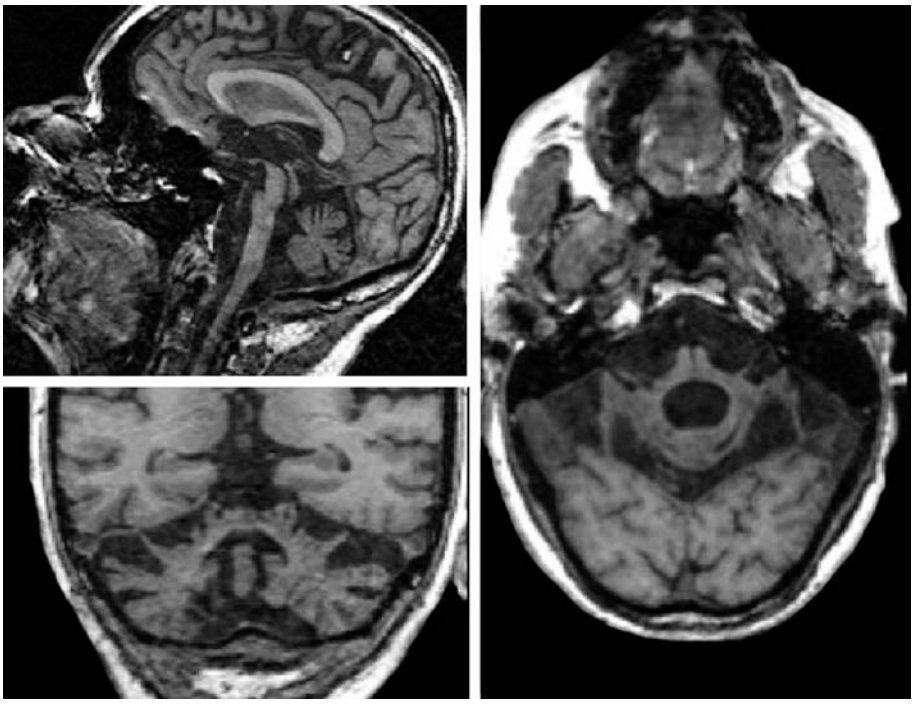

Figure 3 MRI in spinocerebellar ataxia type 2, showing severe atrophy of the cerebellum and brain stem on T_1-weighted images.

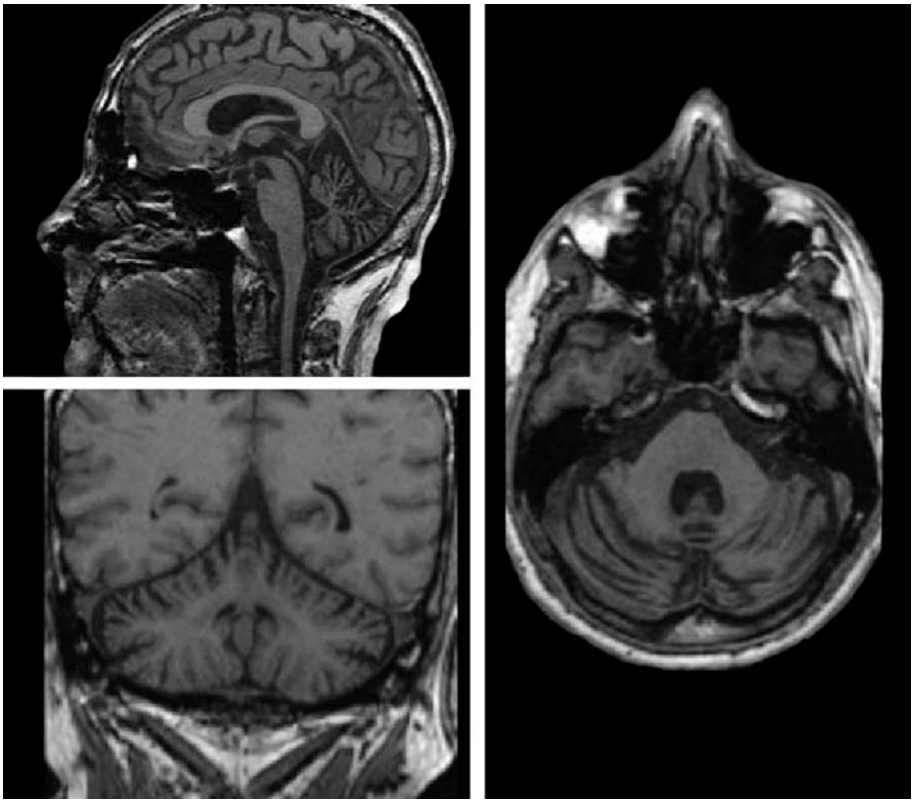

Figure 4 MRI in spinocerebellar ataxia type 6, showing isolated atrophy of the cerebellum on T_1-weighted images.

See also: Apoptosis in Neurodegenerative Disease; Axonal Transport and Neurodegenerative Diseases; Excitotoxicity in Neurodegenerative Disease; Lysosomal System; Protein Folding and the Role of Chaperone Proteins in Neurodegenerative Disease.

Further Reading

Abele M, Bürk K, Schöls L, et al. (2002) The aetiology of sporadic adult-onset ataxia. *Brain* 125: 961–968.

Chai Y, Berke SS, Cohen RE, et al. (2004) Poly-ubiquitin binding by the polyglutamine disease protein ataxin-3 links its normal function to protein surveillance pathways. *Journal of Biological Chemistry* 279: 3605–3611.

Gatchel JR and Zoghbi HY (2005) Diseases of unstable repeat expansion: Mechanisms and common principles. *Nature Reviews Genetics* 6: 743–755.

Gilman S, Low PA, Quinn N, et al. (1999) Consensus statement on the diagnosis of multiple system atrophy. *Journal of the Neurological Sciences* 163: 94–98.

Harding AE (1982) The clinical features and classification of the late onset autosomal dominant cerebellar ataxias. A study of 11 families including descendants of 'the Drew family of Walworth.' *Brain* 105: 1–28.

Ikeda Y, Dick KA, Weatherspoon MR, et al. (2006) Spectrin mutations cause spinocerebellar ataxia type 5. *Nature Genetics* 38: 184–190.

Klockgether T, Lüdtke R, Kramer B, et al. (1998) The natural history of degenerative ataxia: A retrospective study in 466 patients. *Brain* 121: 589–600.

Matsuyama Z, Yanagisawa NK, Aoki Y, et al. (2004) Polyglutamine repeats of spinocerebellar ataxia 6 impair the cell-death-preventing effect of CaV2.1 Ca^{2+} channel – Loss-of-function cellular model of SCA6. *Neurobiology of Disease* 17: 198–204.

Moseley ML, Zu T, Ikeda Y, et al. (2006) Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. *Nature Genetics* 38: 758–769.

Orr HT, Chung MY, Banfi S, et al. (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature Genetics* 4: 221–226.

Schaffar G, Breuer P, Boteva R, et al. (2004) Cellular toxicity of polyglutamine expansion proteins: Mechanism of transcription factor deactivation. *Molecular Cell* 15: 95–105.

Schöls L, Bauer P, Schmidt T, et al. (2004) Autosomal dominant cerebellar ataxias: Clinical features, genetics, and pathogenesis. *Lancet Neurology* 3: 291–304.

Shults CW, Rockenstein E, Crews L, et al. (2005) Neurological and neurodegenerative alterations in a transgenic mouse model expressing human alpha-synuclein under oligodendrocyte promoter: Implications for multiple system atrophy. *Journal of Neuroscience* 25: 10689–10699.

Tsuda H, Jafar-Nejad H, Patel AJ, et al. (2005) The AXH domain of Ataxin-1 mediates neurodegeneration through its interaction with Gfi-1/Senseless proteins. *Cell* 122: 633–644.

Wenning GK, Tison F, BenShlomo Y, et al. (1997) Multiple system atrophy: A review of 203 pathologically proven cases. *Movement Disorders* 12: 133–147.

Yazawa I, Giasson BI, Sasaki R, et al. (2005) Mouse model of multiple system atrophy alpha-synuclein expression in oligodendrocytes causes glial and neuronal degeneration. *Neuron* 45: 847–859.

Startle Response

J R Fetcho and D L McLean, Cornell University, Ithaca, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Startle Behavior

When a door unexpectedly slams, people are usually startled. In humans, the movements produced during this response include a rapid closing of the eyes and a hunching of the body as a result of a large contraction of the muscles in the neck and body (Figure 1). Such startle responses are nearly universal among animals and they occur in response to a strong, abrupt, and unexpected stimulus, such as a loud sound, an unexpected touch, or the sudden appearance of another animal or person. These startle responses are found widely among both invertebrates, including animals such as flies and crustaceans, and most vertebrates, including fish, amphibians, and mammals. The movements that occur during the response vary among different animals. In flies and in some mammals, the response can include a jumping movement. In crayfish and in fish, a flip of the tail moves the animal quickly through the water (Figure 2).

Behavioral Role of the Startle Response

Despite the varying forms of the movement during the startle response, its behavioral role is similar among the different species. The response serves a protective function that can move an animal rapidly out of the way of an attacking predator or protect vital parts of the body, such as the eyes, from injury that might be associated with an unexpected event. In many nonmammalian species, the startle response is referred to as an escape response because it often occurs in conjunction with a predatory attack.

Shared Features of Neuronal Circuits for Startle Responses

An appropriately timed, rapid, and powerful startle response can mean the difference between life and death for an animal. As a consequence, the neuronal networks underlying the startle response are well designed for both a rapid response and a powerful activation of muscle. The ways in which fast, powerful neuronal circuits are built are similar in different species, so the design of startle networks is similar in very diverse animals (Figure 3).

Startle responses are remarkably fast. In a larval zebra fish, which bends away from the location of a startling stimulus, the response can occur within approximately 5 ms of the onset of the stimulus. During the startle, the fish bends so that the head touches the tail in approximately 15 ms and the fish has moved several body lengths away from its starting location in 100 ms. Humans are slower but still remarkably fast, with the shortest muscle response latencies as little as 14 ms. The speed of response is accomplished by a very short and fast conducting pathway from the sensory input that triggers the response to the motor neurons that lead to the movement. The pathways typically include few neurons from input to output, which minimizes the delay introduced by synaptic connections. In fish, for example, the shortest pathway for an escape in response to an acoustic input involves only four neurons – the hair cells in the ear that connect to eighth nerve fibers which in turn connect to a neuron called the Mauthner cell (Figure 3), which directly connects to motor neurons in spinal cord to activate the startle response. The speed of the pathways is further optimized by the presence of electrical synapses, which allow direct current flow between one neuron and the next and reduce the delays associated with chemical neurotransmission. Finally, the startle circuits contain large, relatively fast conducting neurons in the pathways. These so-called giant neurons have relatively very large axons and, often, large cell bodies as well.

In invertebrates, fish, and amphibians, there are very few of these giant neurons, which can be present as a bilateral pair and are individually identifiable from animal to animal. The individual neurons receive input from many different sensory modalities, including the visual, auditory, and somatosensory systems, as well as more specialized senses such as the lateral line system, which senses water flow in fish. The presence of these large, multimodal interneurons has made the startle system especially good for electrophysiological studies because the large size allows relatively easy recording from the neurons.

A strong response during startle is ensured by the powerful connections of startle pathways to motor neurons that excite large numbers of very fast contracting muscle fibers. The motor neurons excited by the giant neurons in startle pathways are typically the largest motor neurons in the animal, with each motor neuron innervating many muscle fibers. This activates the fibers in the muscles nearly simultaneously to produce a powerful pulse of motor output that initiates the response.

The timing of startle responses can be critical. The response is typically initiated by a strong sensory input. The electrical properties of the giant neurons,

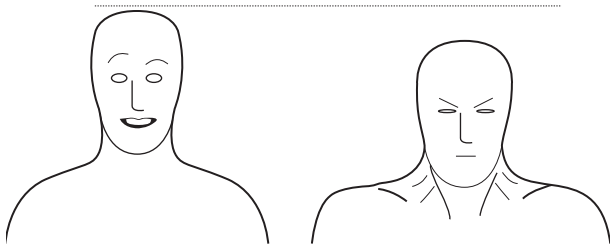

Figure 1 Human startle response. During the startle response the eyes and mouth close, the body is hunched, and the muscles of the neck and torso contract. These act to protect vital parts of the body from potential injury associated with the loud sounds or abrupt visual stimuli that might indicate a falling or flying object.

Figure 2 Escape response in a fish. Overlapping images of a fish doing a startle response after an abrupt touch of the body. The fish bends quickly to one side (in approximately 10 ms) and moves rapidly away from the stimulus.

which initiate the response, along with a powerful inhibition of these nerve cells at weak stimulus strengths give the neurons a high threshold for firing action potentials. This high threshold makes functional sense because the startle response is such a dramatic, energetically demanding movement that it would be detrimental to an animal to be constantly startling to every weak sensory stimulus in the environment. The relatively high threshold also helps to time the initiation of a startle response in a way that might maximize the ability of an animal to evade a predatory attack. For example, when a bird attacks a fish, if the fish performs the startle tail flip response too soon it might be possible for the bird to adjust its attack to the new location of the fish. The high threshold may help ensure that the startle does not occur until after the bird is committed to strike at a particular location, with the startle moving the fish out of that location at the last possible moment.

Important Observations Made Using Startle Circuits

Cellular and Synaptic

The remarkably large size of the neurons and their axons in startle circuits has made them a favorite preparation for basic studies of the physiology of nervous systems because of the ease of recording from the neurons. As a result, studies using startle neurons have provided a number of critical insights into the basic function of neurons and their synaptic connections. The most prominent of these is the

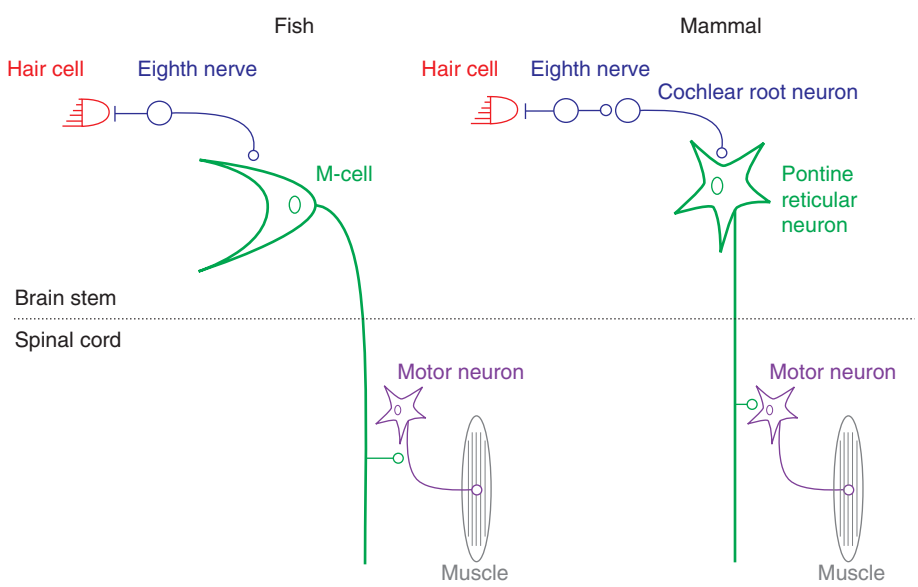

Figure 3 Similarities between startle neuronal circuits in fish and mammals. Inputs from the hair cells in the ear excite neurons in the brain stem that project to the spinal cord, where they directly excite motor neurons to cause the muscle contraction typical of the response. M-cell, Mauthner cell.

Nobel Prize-winning work of Hodgkin and Huxley, which described the ionic basis of the action potential by using a giant axon involved in the startle response of squid. The large size of the axon allowed them to thread a wire along inside it and control the membrane potential. This ability to properly clamp the transmembrane voltage and measure the currents required to do so during an action potential led to the discovery of the contribution of changes in membrane conductance to potassium and sodium to the generation of the action potential.

Other key contributions from studies of startle neurons include the first description of electrotonic transmission in a vertebrate nervous system. This was possible because the large size of the Mauthner neuron that generates the startle response in goldfish, along with some unique electrical properties of the cell, allowed for some of the earliest intracellular recordings from a neuron and its synaptic inputs in an intact brain. This ability to record reliably from a particular neuron in fish also led to the first description of field effect inhibition in the nervous system. This inhibition occurs not by a conventional synaptic interaction but, rather, as a result of current flows induced by the activated neurons, along with special morphological features that constraint the pathways along which the current can flow. Finally, studies of synaptic contacts onto the giant Mauthner neuron also provided striking evidence that each active zone in a presynaptic terminal releases at most one vesicle of neurotransmitter when the presynaptic neuron fires one action potential. These and other fundamental discoveries can all be attributed to the specializations of startle circuits that made them very accessible for even the earliest electrophysiological studies.

Network Level

One of the main challenges of neuroscience is to link neurons to the behavior they produce. Neurons in startle networks were some of the first to be shown to be active in a particular behavioral context. Once again, this was possible because the large size of the neurons allowed recording of the firing with extracellular electrodes even in freely moving animals. Such work showed a tight relationship between the firing of a giant neuron and the production of a rapid escape response, suggesting a causal role. Later studies in which the startle neurons were selectively deleted by using strong current injections or a laser confirmed a role for the giant neurons in the startle behavior.

Variability of the Startle Response

Although the startle response is often considered an almost reflexive response, there is evidence for variability in the behavior as a function of the sensory stimulus that elicits it. In crayfish, for example, a rostral stimulus activates one type of giant neuron (the medial giant interneuron) to produce a tail flip that moves the animal backward, away from the stimulus. A caudal stimulus, in contrast, activates another giant neuron (the lateral giant) that produces a different tail flip that directs the animal forward, away from the stimulus. Changes in the set of active neurons are also documented for the escape behavior in fish, where variation in the location of the sensory stimulus leads to activation of different large reticulospinal neurons and the production of different movements, all of which act to direct the fish away from a potential threat. Thus, there is considerable sophistication to the sensorimotor response, even though the network is relatively simple.

Plasticity of the Startle Response

A further level of sophistication in the response is evident in documented plasticity in synaptic connections in the circuits. There is evidence that both excitatory and inhibitory synapses in the circuits can undergo changes in strength depending on the pattern of activity in the network. This makes behavioral sense because the network must adjust to changes in the sensory environment. A fish in an environment that has considerable auditory noise (e.g., near a waterfall) might have an advantage if it reduced the efficacy of an auditory stimulus in producing an escape while maintaining the response to other sensory modalities. Otherwise, the fish would constantly be producing escapes to irrelevant stimuli. The ability of the escape behavior to show behavioral plasticity such as habituation is consistent with the importance of matching even this relatively simple behavior to the sensory environment.

Neuromodulatory systems such as those using dopamine and serotonin are intimately associated with neurons in startle networks, where they may plays roles in altering excitability and synaptic function to produce response plasticity. Neuromodulatory systems are often disrupted in neuropsychiatric diseases, which are also associated with alterations of the startle response.

Evolution of the Startle Response

The startle networks in invertebrates and aquatic vertebrates tend to involve few neurons, with at most only a handful of the giant neurons, which are critical for the response. In contrast, the network in mammals, although almost identical to that in fish

with regard to the number of synapses from auditory input to motor neuron, includes many more neurons at each step of the pathway (Figure 3). As in fish, the main giant neurons initiating the startle in mammals are located in the reticular formation (in the caudal pontine reticular nucleus) and their axons extend down the spinal cord to excite motor neurons. The evolution of the startle response thus appears to involve expansion of the number of neurons at each step in the pathway in more complex animals without much change in the number of steps, which might compromise the speed of response.

Pathology of the Startle Response

The startle response also plays an important role in studies of neuropsychiatric diseases of humans because disruptions in startle are often symptomatic of broader neurological problems. The startle response shows a well-known phenomenon called prepulse inhibition that is affected in a number of neurological disorders such as schizophrenia, Huntington's disease, and obsessive–compulsive disorder. In normal prepulse inhibition, when a weak sensory stimulus that does not elicit a startle is followed within less than approximately 0.5 s by a strong stimulus that would normally elicit a startle, the weak stimulus suppresses the response to the strong one. This phenomenon occurs in both invertebrates and vertebrates, including humans, and is thought to block later inputs to allow time for processing of the initial weak stimulus. This would seem counterproductive in a startle situation, but nonetheless prepulse inhibition does exist in startle circuits and is very important clinically because it is altered (often reduced) in diseases such as schizophrenia. The ease of testing for prepulse inhibition of startle in humans can provide an indication of a potential disease state. More important, the efficacy of various drugs in reducing deficits in prepulse inhibition might offer some indication of the effectiveness of the drug for treatment of the disease states. The effects on prepulse inhibition may well result from alteration of transmitter systems during disease, including inhibitory pathways known to be important for gating the flow of sensory information as well as changes in the neuromodulatory serotonergic and dopaminergic systems that are important both in startle responses and in many neurological diseases.

One indication of the critical role of inhibitory pathways in startle is an identified genetic disease of startle circuits called hyperekplexia, or familial startle disease. This is characterized by exaggerated startle responses and a reduction of habituation. The disease results from a mutation in the α subunit of the glycine receptor that alters the strength of inhibition. A balance between inhibition and excitation regulates the threshold of startle responses and a shift of this balance toward excitation by the mutation may explain the enhancement of startle in the disease.

Conclusion

The startle response is a fast, protective, fundamental sensorimotor response in animals that is often used to avoid predators. The circuits are simple and fast, with large neurons, which has allowed studies of neurons in startle networks to provide fundamental insights into the cellular and synaptic biology of the nervous system. The plasticity of the startle response and its disruption in disease states with broad neuropsychiatric consequences have made it a model system for studies of the neuronal generation of behavior and its disruption and treatment during disease.

See also: Huntington's Disease; Learning, Action, Inference and Neuromodulation; Obsessive–Compulsive Disorder; Prepulse Inhibition of Startle in Humans and Laboratory Models; Reticulospinal System; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome.

Further Reading

Allen MJ, Godenschwege TA, Tanouye MA, and Phelan P (2006) Making an escape: Development and function of the *Drosophila* giant fibre system. *Seminars in Cell and Developmental Biology* 17: 31–41.

Bakker MJ, van Dijk JG, van den Maagdenberg AM, and Tijssen MA (2006) Startle syndromes. *Lancet Neurology* 5: 513–524.

Dawson M, Schell A, and Bohmelt A (eds.) (1999) *Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science*. Cambridge, UK: Cambridge University Press.

Eaton RC (ed.) (1984) *Neural Mechanisms of Startle Behavior*. New York: Plenum Press.

Eaton RC, DiDomenico R, and Nissanov J (1991) Role of the Mauthner cell in sensorimotor integration by the brain stem escape network. *Brain, Behavior and Evolution* 37: 272–285.

Edwards DH, Heitler WJ, and Krasne FB (1999) Fifty years of a command neuron: The neurobiology of escape behavior in the crayfish. *Trends in Neurosciences* 22: 153–161.

Faber DS, Fetcho JR, and Korn H (1989) Neuronal networks underlying the escape response in goldfish. General implications for motor control. *Annals of the New York Academy of Sciences* 563: 11–33.

Faber DS and Korn H (eds.) (1978) *Neurobiology of the Mauthner Cell*. New York: Raven Press.

Korn H and Faber DS (2005) The Mauthner cell half a century later: A neurobiological model for decision-making? *Neuron* 47: 13–28.

Kumari V and Sharma T (2002) Effects of typical and atypical antipsychotics on prepulse inhibition in schizophrenia: A critical evaluation of current evidence and directions for future research. *Psychopharmacology (Berlin)* 162: 97–101.

Liu KS and Fetcho JR (1999) Laser ablations reveal functional relationships of segmental hindbrain neurons in zebrafish. *Neuron* 23: 325–335.

Rothwell JC (2006) The startle reflex, voluntary movement, and the reticulospinal tract. *Supplements to Clinical Neurophysiology* 58: 223–231.

Yeomans JS and Frankland PW (1996) The acoustic startle reflex: Neurons and connections. *Brain Research: Brain Research Reviews* 21: 301–314.

Statistical Analysis of Visual Perception

R B Lotto, University College London, London, UK

© 2009 Published by Elsevier Ltd.

Decoding a Code

If given the code *1-A-B-2-A-B-3-D-4-C-D-5-E-F-6-E-F*, 'cracking' it requires reducing the search space to a manageable size. The only way to achieve this reduction is to assume the nature of the information that code represents. Consider hieroglyphics. Although a code used by the Ancient Egyptians for 3500 years, it remained untranslatable for approximately 1500 years, even though much was known about the code and its history. Even when the 'Rosetta Stone' discovered by Napoleon's troops found its way to the University of Cambridge (by not altogether scrupulous means), Thomas Young – the father of modern color vision research – only managed to translate some of the text, particularly the proper nouns, even though the same decree on the stone was also inscribed in Greek and Demotic. Most of the hieroglyphic text was undecipherable because Young assumed – like most before him – that hieroglyphs were used to symbolize only objects and events in the natural and social world of the Egyptians. When Jean Francois Champollion put forward a new framework for thinking about hieroglyphs, namely that hieroglyphs had phonetic value as well as symbolic meaning and thus served as spoken language, he was able to translate the remaining text in 1821, thus initiating the modern field of Egyptology. The point, then, is that to fully understand hieroglyphics, one needed to describe not only its structures and organizations but also the information that it represented. This truism for hieroglyphs is true for any code, including the codes used in neural processing.

The spatial and temporal patterns of activity in the human visual brain are codes that ultimately transform the light that falls onto the eye into the sensations of color, space motion, form, etc. that guide our behavior. The aim of vision research, then, is to decipher this code. Fundamental to this aim is describing its underlying anatomy, physiology, and biochemistry. Of equal importance, however, is to deduce the nature of the information represented in that description. For many years, those in vision research assumed (usually implicitly) that visual encoding was a complex algorithm (e.g., sequence of filters) for accurately reconstructing the qualities of real-world objects (their reflectance, illumination, size, location, etc.) from elemental features of retinal images. In other words, the assumption was that the brain evolved to reconstruct and represent in perception the world as it is. Despite the intuitive appeal of this representationalist framework, exemplified in the important work of David Marr and colleagues in the 1970s and 1980s, it is beginning to lose ground – although not in all areas of investigation (e.g., some continue to argue that reflectance, illumination, and transmittance are represented explicitly as layers in visual processing). Replacing this perspective is a new framework, grounded in decades of computational, physiological, and behavioral studies on human and nonhuman visual processing, which suggests that vision encodes not the physical world as such but, rather, the statistics of its ecological influence on human observers.

Encoding Statistics

Why encode statistics? There are at least two reasons: energy resources for animals are limited, and the behavioral significance of visual information is ambiguous.

Energy Conservation

The survival of a visual animal requires generating useful behavior from stimulus information that falls onto its light sensors. One problem in doing this successfully is that neural processing is expensive. If more energy is spent on vision than is obtained through visually guided behavior, an animal's probability of survival decreases; spend less energy and the probability of survival increases. It is therefore reasonable to assume that a significant constraint in the evolution of vision has been the need to conserve the amount of energy an animal uses to process visual information.

One way to save energy is to see with as few cells as possible since each cell in the visual system uses energy to grow and be maintained. The smaller the visual network, the less energy the network requires. However, all things being equal, the fewer the cells, the less information the network can process. Thus, maintaining information capacity requires a commensurate increase in the amount of information encoded by each cell in a smaller network. This means smaller networks must work harder, fire more often and in more complex temporal patterns, and thus multiplex their functions. However, the cost of signaling with noisy synapses is nonlinearly related to signaling rate. The more often a cell 'fires,' the more expensive each response becomes. Thus, although there is pressure to evolve a system made up of fewer, frequently active cells, there is also selection

pressure to do just the opposite: spread visual processing across a larger network of infrequently firing cells – called 'sparse coding.'

From this we can hypothesize that visual processing evolved to maximize its information capacity while minimizing cell number and signaling frequency. How can it do all three? The answer is presumably computational efficiency achieved by eliminating redundancy in the code. There are two sources of redundancy in visual processing. The first is neuronal. If two cells fire simultaneously in response to the same sensory information, then their responses are correlated. Correlated signals are duplicate codes and therefore redundant. The costs associated with coding redundancy are the necessary increase in cell number and cell signaling. Thus, decorrelating responses early in visual processing by reducing the overlap in the information encoded by each cell can save energy by encoding the same information more efficiently.

The second source of redundancy is the natural image. To understand why, consider all the possible images the visual system could experience on a standard computer monitor set to a resolution of 1280×1024 pixels. Given 1.3 million pixels and a 32-bit color depth, the size of this theoretical image database is $10^{6\,619\,700\,000\,000}$. With the number of atoms in the universe being only 10^{21}, this is very large indeed (even larger if one were to equate the number of pixels of the monitor with the number of retinal receptors). Most of these images, however, would be fairly uninteresting, made up of random white noise (like 'snow' on television late at night) where each pixel is statistically independent from the quality of all other pixels in space and time. Images of natural scenes, on the other hand, are far more interesting, occupying a unique and relatively small part of the landscape of all possible images. What is interesting about natural images is that they have 'structure,' where pixel intensity at each point is correlated with the intensity of the neighboring pixels in space and time. The reason for these correlations is that surfaces and illuminants in the natural world are extended in space, and motion is continuous in time. Encoding information about individual pixels, therefore, would be highly inefficient since the relationship between pixels would lead to correlated (and thus redundant) neural responses. Instead, the visual system can increase the sparseness of its code by decorrelating its neural responses to images, which is achieved by taking advantage of the statistical relationships between pixels at multiple spatial and temporal scales.

Thus, assuming the visual system actually needs to re-encode the patterns of light that fall onto the retinae as efficiently and accurately as possible in its post-receptor processing, then clearly encoding the inherent spatial/temporal statistics of images is advantageous since it conserves energy.

Behavioral Robustness

It is important to ask why vision should re-encode the retinal image in the first place. After all, since all the information about an image is available in the pattern of retinal activation, then no matter how efficient and faithful its subsequent representation, the re-encoded information is redundant with the original retinal representation. There are several possible answers to this question. One is that the mammalian retina must send information about the activity of its 130 million receptors down an optic nerve that for whatever reason is constrained to only 1.2 million axons. Given this massive convergence of wiring, and assuming most visual processing of color, form, motion, etc. begins in the primary visual cortex (which is questionable, given the complexity of retinal and thalamic processing), coding efficiency may be a necessary step for maximizing image information en route to the cortex. Note, of course, that the seeming 'bottleneck' of the optical nerve may actually be a manifestation of efficient coding. On the other hand, the insect visual system does not suffer the same retina-to-brain wiring bottleneck, and yet much of the direct evidence for coding efficiency has been substantiated in insects. Thus, another possibility is that re-encoding the spatial and temporal patterns of activation across the retina enables information about the visual image to be mapped at minimal cost to multiple 'channels' (or modules) for independent parallel processing, thereby gaining the advantages that parallel processing confers on complexity of behavior.

Finally, and most critically, it is important to emphasize that an image of the world is just that – an image. Although real-world 'scenes' (defined in terms of the statistics of objects, illuminations, and three-dimensional distances) and their 'images' (defined in terms of lower and higher-order correlations between pixels) are frequently used interchangeably in the literature, the two are very different indeed. Thus, although research on statistics and/or entropy of images has been important for validating the ecological basis of vision, the analysis does not actually deal with the real challenge of visual ecology, which is not to 'see' the image that falls onto the eye (efficiently or otherwise) but, rather, to behave toward objects in the world that generated the image. It is therefore essential that vision research also explores directly the role that the behavioral significance of image information has played in visual

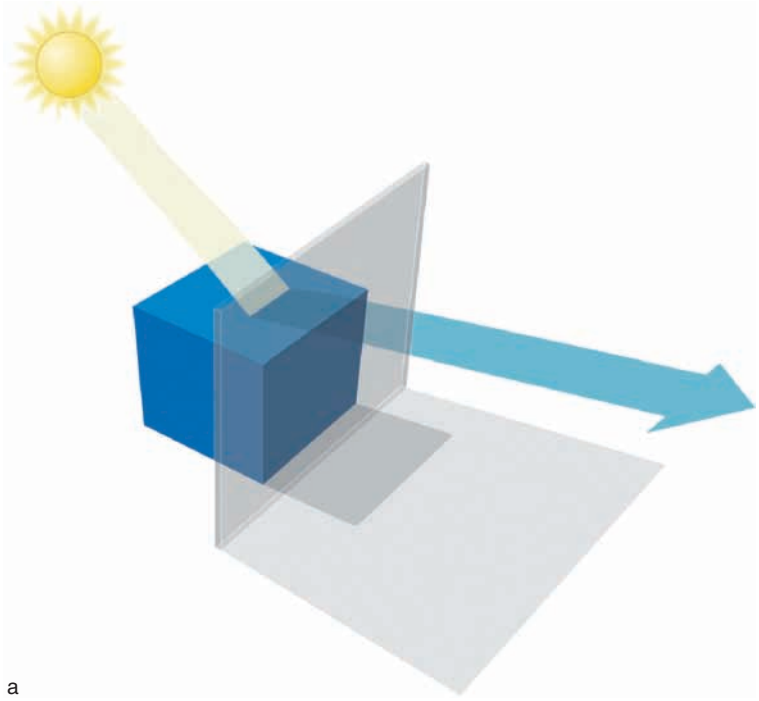

a

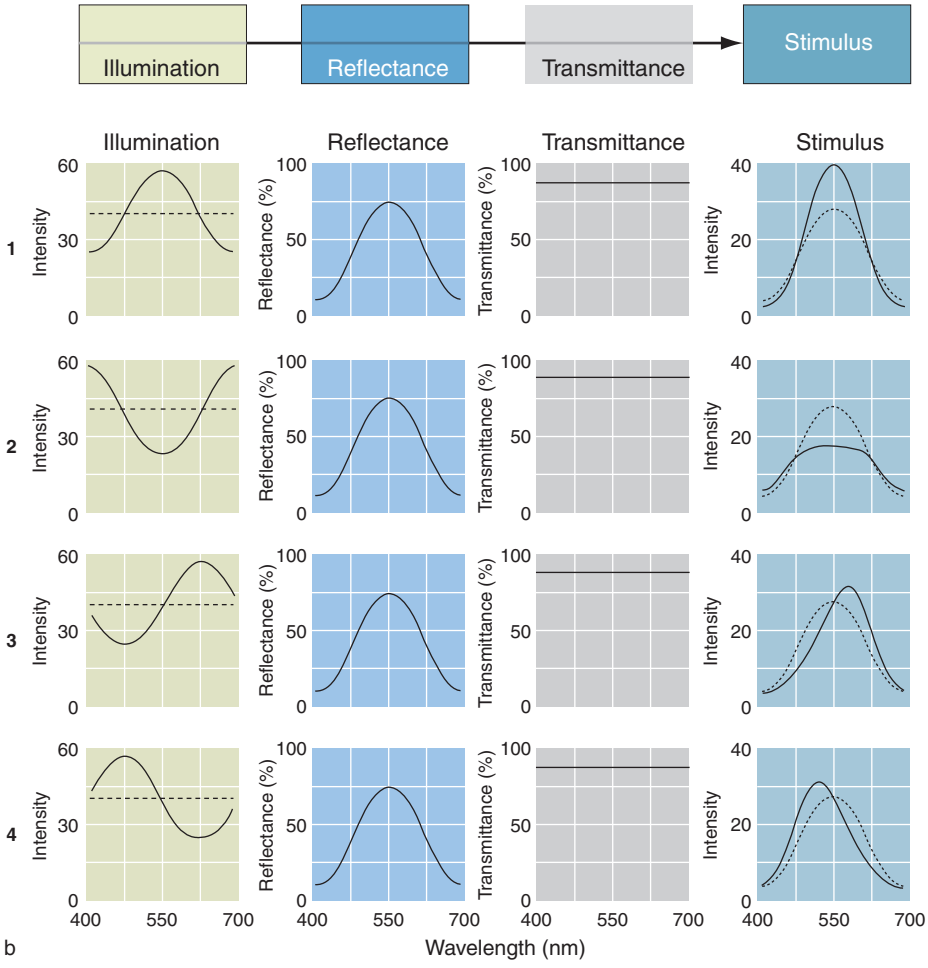

b

Figure 1 Continued

evolution and development. Evidence that it does play a role, even at the lowest level of visual processing, is indicated by the spectral sensitivity of retinal receptors. For instance, bees have three broadband receptors tuned to approximately 350, 450, and 550 nm. Computational analysis has shown that this combination of receptors is ideal for detecting stimulus information arising specifically from flowers (it is of course also possible – if not likely – that flowers evolved their reflectance characteristics according to the bee's spectral sensitivity). Similarly, the sensitivities of the 'red' and 'green' cones in the human eye are well matched to behaviorally relevant stimulus information arising from their historical environment. More generally, the relationship between the behavioral value of stimuli and receptor sensitivity has been thoroughly described in natural systems and quantitatively modeled in robotic and artificial life systems. Thus, the behavioral significance of images at the very least constrains the window through which the animal sees its world by determining what stimulus information to detect and what to ignore. However, it also does a good deal more. It actually shapes the way the visual system subsequently processes that information.

The reason is that there is not intrinsic significance for any image since each of its attributes is derived simultaneously from multiple attributes of the world (reflectance and illumination in the case of the spectral quality of a stimulus and size and distance in the case of its projected angle; **Figure 1**). Thus, if an animal is to survive, it must overcome the fact that the only direct information it has about the world is inherently ambiguous vis-à-vis its generative source. Little is known about how vision resolves this challenge. What is clear is that the only way to do so effectively is by encoding the statistical relationship between images, scenes, and behaviors in past experience. Originating with the work of Helmholtz and then Gibson, this empirical/ecological view of vision has been usefully explored quantitatively using Bayesian decision theory.

Work on visual illusions demonstrates the validity of this probabilistic framework particularly well. For instance, in **Figure 2** the two red crosses on the right and left faces of the cube appear to have different brightness, with the one in 'shadow' seemingly brighter than the one in 'light.' Despite their appearance, the crosses are in fact physically identical in that they reflect the same amount of light to the eye (**Figure 2(b)**); also note that all the small 'black' tiles in depicted light are physically the same as all the small 'white' tiles in depicted shadow). This perception of different colors in response to identical stimuli is an 'illusion' (like many others published previously): what is seen is not what is really there. Note, however, that had the identical stimuli originated from a real cube in light and shadow, then in the past experience of the animal the behavioral significance of the identical stimulus elements would have turned out to be differently reflective surfaces, with the one in shadow being a more reflective surface than the one in bright light. If vision evolved to see the world not as it is but, rather, with reference to statistical information about what an image turned out to be in the past (shown didactically in **Figure 1**), then the perception of the crosses in **Figure 2** should represent this experience and the crosses should appear to have different brightness, as indeed they do. Thus, these observations (and many similar observations) indicate that vision uses statistics not only to conserve energy but also to generate behaviors from image information that is inherently meaningless.

A corollary of this view is that a more complete understanding of vision will require quantitative descriptions of the probabilistic relationship between natural images, the patterns of activation in the relevant network generated by those images, and subsequent motor behavior, as well as the 'cost' to the animal/agent of generating that behavior, which can be modeled explicitly using Bayesian decision theory and/or empirical ranking theory. Although most of this information is unknown, if not unknowable, for human vision, in exploring the general biological and computational principles of vision it is not necessary to focus exclusively on primates (or even mammals for that matter) since simpler systems (e.g., insects) also evolved to overcome the same ecological challenges of vision. Also, since simple systems (both natural and artificial) are far more tractable,

Figure 1 Physical interactions in the generation of spectral stimuli. (a) Diagram illustrating the inevitable conflation of illumination, reflectance, and transmittance in the generation of spectral returns that elicit color sensations. (b) The typical consequences of the interaction of reflectance, illumination, and transmittance in this process. In this example, a 'surface' with a given reflectance efficiency function (illustrated here as a simple sine wave) is 'illuminated' by equiluminant spectra whose configuration, relative to that of the surface's reflectance, is the same (1), the opposite (2), or shifted toward long (3) or short (4) wavelengths. For comparison, the dashed line represents a 'white' light illuminant comprising a uniform distribution of wavelengths. For the sake of simplicity, transmittance is assumed to be uniform (although usually minor, the effects of atmospheric transmittance are neither negligible nor uniform). Note that the spectral returns given by the product of the reflectance efficiency functions and the illuminants vary systematically relative to the return generated by the same surface under white light.

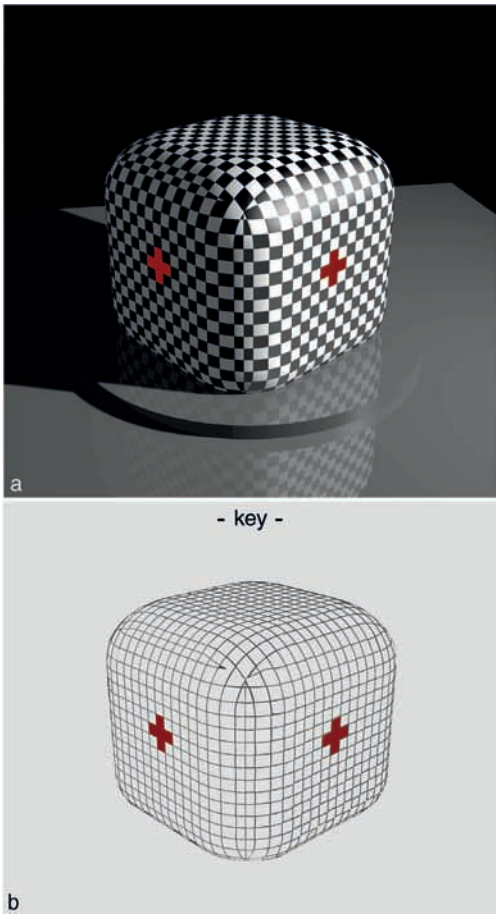

Figure 2 (a) A cube in light and shadow. Although the red crosses on the two sides of the cube appear to have different brightness, they are in fact physically identical, which is shown in the key (b). The illusion of brightness in a suggests that the human visual system encodes the statistical relationship between the patterns of light that fall on the retina and the surfaces and illuminants that generated those patterns in the past.

they provide an essential platform for rationalizing 'empirical' vision in quantitative terms – in parallel with the important work on humans and other primates.

Conclusion

There is nothing red about 'r-e-d'; 'l-o-n-g' is shorter than 's-h-o-r-t'; there is no '2-ness' in the letter string 't-w-o'; and there is nothing singular about 'o-n-e.' English is an emergent, social, and biological phenomenon, understandable only within its social and biological context. Information about syntactic structure, if divorced from the information encoded in the system, describes the code but does not explain it. Adaptive networks, like the visual brain, are also frequently described as codes that translate image information into behavior. As with all codes, understanding vision requires not only a clear description of its functional structure but also a clear understanding of (or at least coherent assumption about) the nature of the information encoded by vision. Answering the latter question, which is a key aim of theoretical and computational neuroscience, requires taking into account the challenges that are actually faced by vision in its evolution and development when interacting with the world. There are at least two: energy is limited and visual processing expensive, and image information is ambiguous and robust behavior essential for survival. Both issues have been the focus of much research on primates, insects, and synthetic visual systems, although usually pursued along parallel (i.e., noninteracting) lines of investigation. Nonetheless, both constraints may be resolved by the same process of encoding the statistics of images and the probabilistic relationship between images and their behavioral significance in an animal's ontogeny and phylogeny. This view, which is frequently prosecuted by a combination of information theory, Bayesian decision theory, and coding theory, provides a clear framework for rationalizing perception and processing based on explicit assumptions about the nature of the information encoded in the visual system.

See also: Contextual Interactions in Visual Perception; Information Coding; Natural Images: Coding Efficiency; Vision for Action and Perception; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging; Visual Cortex in Humans; Visual Cortical Models of Orientation Tuning.

Further Reading

Attick JJ (1992) Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems* 3: 213–251.

Baddeley R, Hancock P, and Foldiak P (eds.) (2000) *Information Theory and the Brain*. Cambridge, UK: Cambridge University Press.

Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Rosenblith WA (ed.) *Sensory Communication*, pp. 217–236. Cambridge, MA: MIT Press.

Gibson JJ (1979/1986) *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Erlbaum.

Knill DC and Pouget A (2004) The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences* 27: 712–729.

Laughlin SB (2001) Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology* 11: 475–480.

Laughlin SB and Sejnowski TJ (2003) Communication in neuronal networks. *Science* 301: 1870–1874.

Lythgoe JN (1979) *The Ecology of Vision*. Oxford, UK: Clarendon.

Marr D (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Olshausen BA and Field DJ (2004) Sparse coding of sensory inputs. *Current Opinion in Neurobiology* 14: 481–487.

Prete FR (2004) *Complex Worlds from Simpler Nervous Systems*. Cambridge, MA: MIT Press.

Purves D and Lotto B (2003) *Why We See What We Do: An Empirical Theory of Vision*. Sunderland, MA: Sinauer.

Rao RPN, Olshausen BA, and Lewicki MS (eds.) (2002) *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press.

Rieke F, Warland D, de Ruyter van Steveninck RD, and Bialek W (eds.) (1997) *Spikes – Exploring the Neural Code*. Cambridge, MA: MIT Press.

Simoncelli EP and Olshausen BA (2001) Natural image statistics and neural representation. *Annual Review of Neuroscience* 24: 1193–1216.

Statistical Learning of Language

J S F Hay, University of Wisconsin–Madison, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

The Problem of Language Acquisition

Learning language is a complex endeavor, both for infants learning their first language and adults learning their second, third, or fourth languages. What makes learning language so complex is that language itself is so complex. Elements in natural language, like sounds, words, and phrases, can be combined and recombined to form an infinite number of unique sentences. It is the task of language learners to figure out what the basic elements are and then to learn which combinations of these elements are permitted in their particular language.

How does this problem relate to learning the sounds (phonetics) and sound patterns (phonotactics) in language? In this case, finding the basic elements means figuring out which differences between speech sounds in a language environment matter within the language and which differences can or should be ignored. More than 800 speech sounds are found across the world's languages. Determining how a language breaks up the auditory space is not a simple task. Some languages use as few as 11 consonants and vowels, whereas others use as many as 141. It is the language learners' task to discover the relevant sounds in their native language. Learners then need to discover which sounds are allowed to go together, when they should go together, and in which order. For example, the sounds /t/ and /s/ are allowed to occur together at the beginning of words in Japanese, as in 'tsunami,' but in English they are allowed to occur together only in borrowed words. Because parents cannot directly communicate to their infants which sounds are important or which sounds go together, infants need to have a way to learn this through experience.

Infants are faced with a similar problem during word discovery. How can infants pick words out of continuous speech before they know any words? Similarly, they have to discover how words relate to things in their environment and how words themselves combine together to make phrases. Again, infants do not receive explicit instruction to help them solve these problems. To make matters even more challenging, the linguistic environment is noisy and imperfect. Although there is an essentially infinite number of possible combinations of sounds, words, and phrases, language learners can be exposed only to the subset of possible combinations permitted in their language. In addition, there is no explicit negative evidence provided in linguistic input to tell children that certain combinations are not allowed. How do children learn which combinations are permitted and which are not allowed from this seemingly impoverished input? Part of the answer lies in the fact that language is rich with regularities that language learners are well equipped to track.

What Is Statistical Learning?

Statistical learning is a process by which an individual can discover and make use of experienced regularities. Broadly defined, it involves detecting structure in the environment by tracking patterns in the input. Natural language contains myriad statistical information that can be captured by listeners. The majority of research on statistical learning has focused on how human adults, infants, and even nonhuman animals track regularities during phonetic acquisition, speech segmentation, and the learning of simple grammatical structures. Each of these areas is reviewed in the following sections.

Finding the Elements: Distributional Learning and Phonetic Acquisition

Statistical learning mechanisms have recently been invoked to explain how infants acquire their native-language distinctions between speech sounds. Newborn infants can detect sound contrasts found across the world's languages, often responding more to differences between sounds that are different consonants or vowels than to equivalent differences between instances of the same consonant or vowel. It is fairly straightforward to assess infants' responsiveness to different contrasts. When infants are presented with a sound that they find interesting, they tend to look in the direction that the sound is coming from. If the sound is played repeatedly, infants will become bored, and the amount of time they spend looking in the direction of the sound decreases. If the sound is changed and infants notice and find the change interesting, they will reorient toward the sound. If infants do not notice the change or do not find the change interesting, they will continue to pay little attention to the new sound. In this way infants' responsiveness to different sound contrasts can be measured. Over time, infants' perceptual abilities become tailored to their native language, and by 10 to 12 months of age, they show greater responsiveness

to native-language contrasts than to nonnative contrasts. For example, young infants can discriminate between English productions of /r/ and /l/, irrespective of their linguistic environment. Infants in an English-speaking environment maintain the ability to discriminate the /r/–/l/ contrast. However, by their first birthday, infants who are raised in Japanese-speaking environments, where there is a single consonant that is similar to both English /r/ and /l/, no longer show differential responding to this contrast. How do infants progress from possessing language-general perceptual systems to being native-language listeners?

This focusing of perceptual skills may be partially explained by all perceptual systems' propensity to develop according to the structure of the experienced environment. Specific sound inventories employed by different languages vary greatly. Within each language, speech sounds have substantial regularities, however, and it is these regularities that may help infants during acquisition of native-language consonants and vowels. To use the /r/–/l/ example, in English these two sounds are distributed bimodally in the multidimensional space that differentiates /r/ from /l/. This means that the majority of English speakers' productions of /r/ cluster around average acoustic values for /r/, and their productions of /l/ cluster around average acoustic values for /l/. There are productions of both /r/ and /l/ that are distributed further from average productions, but they occur with a lower frequency. If infants can track the distribution of sounds in their environment, English-learning infants should learn to attend to the differences between /r/ and /l/ because they conform to a bimodal distribution. In Japanese, there is a consonant that falls in the acoustic region between English /r/ and /l/. The majority of productions of this Japanese consonant cluster around the average acoustic value for it, and a smaller number of productions fall on either side of this cluster, overlapping some English productions of /r/ and /l/. Infants growing up in Japanese-speaking environments learn to ignore the difference between English /r/ and /l/ as both sounds fall within a single Japanese consonant category. Thus, distributional information could, in principle, provide the necessary information for phonetic acquisition.

Infants' ability to discover visual distributional and correlational information has been well demonstrated experimentally. Infants' ability to track distributional information for the purpose of learning native-language consonants and vowels has been suggested in only a small number of studies, however. The ability to track distributional information is not specific to humans; for example, it has also been shown in a rodent population. Responding to change depending on distributional information – a skill that may be a precursor to phonetic acquisition – is present across species.

Finding the Elements: Statistical Learning and Speech Segmentation

Imagine that you are in a foreign country and you hear the following sentence: *lalangueparleénecontientpasdesignauxphysiquesmarquantledébutetlafind'unmot* (spoken language does not contain physical signals that mark the beginning and end of a word). Unlike written language, in which words are separated by blank spaces, spoken language unfolds continuously, with few obvious cues to word boundaries. If you do not already know French, you probably would find that the sentence sounded as if it was spoken very quickly, and you would have no idea how many words were in the sentence. Preverbal infants are faced with a similar situation when first learning their native language. Parents do not generally speak to their children in single-word utterances. Although speech to infants tends to be acoustically exaggerated, compared to adult-directed speech, it also does not contain many discrete cues to word boundaries. How can infants learn the words in their language, given that there are few acoustic cues to word boundaries?

Part of the answer likely involves prosodic and rhythmic information, as well as intraphrase pauses. These cues, however, are not always present when infants are initially finding words in continuous speech. Fortunately, spoken language contains rich statistical information about words and their boundaries. Transitional probability (TP), also termed conditional probability, is the probability of one event given the occurrence of another event. This statistic refers to more than the frequency with which one element follows another because it adjusts for the base rate of the first event or element. The TP of Y given X is represented by the following equation:

$$\mathrm{P}(Y|X) = \text{frequency of pair } XY/\text{frequency of } X$$

TP between syllables provides a reliable cue to word boundaries in that within a word, the probability that a given syllable will be followed by another particular syllable is relatively high, whereas between words, the probability of one syllable following another is relatively low. Take, for example, the following series of likely two-word phrases that an infant might hear: pretty baby, happy baby, cute baby, pretty kitty, pretty doggy, pretty horsey. In this microcorpus, every time an infant hears *ba*, it is followed by *by*. Therefore, within the word 'baby,' the

syllable sequence has a high TP (TP = 1.0), and infants should treat 'baby' as a good word candidate. So although the syllable *ba* predicts the syllable *by*, hearing *ty* (from 'pretty') does not necessarily predict *ba* (from 'baby'). The TP between syllable sequences *ty* and *ba* is 0.25, and therefore '*tyba*' is a less good word candidate in English than 'baby.' Until recently it was unknown whether humans could track these regularities.

In the mid-1990s, Saffran, Aslin, and Newport developed a paradigm to test whether infants could track TPs. They created a language that contained four three-syllable nonsense words (i.e., tupiro, golabu, bidaku, and padoti), hereafter called 'words,' which were concatenated together in a quasi-random order (e.g., *bidakupadotigolabubidakutupiru* . . .). Words were defined by their high internal TP (TP = 1.0): each syllable within a word perfectly predicted the occurrence of the next syllable. The TP between syllables spanning word boundaries was much lower (TP = 0.33). The language was synthetically produced with no other markers to word boundaries. The researchers familiarized 8-month-old infants with 2 min of this artificial language, with the hypothesis that infants who can track TP should treat words from the language differently from nonwords (sequences of familiar syllables in a novel order, with a TP of zero). Using a head-turn preference procedure, they assessed the infants' preference for words from the language versus nonwords. Infants preferred listening to the nonwords (new sound sequences) from the artificial language. In a more stringent test of infants' abilities to track TP, the researchers presented a second group of infants with words and part-words (sequences made from combining the second two syllables from a word with the first syllable from another word in the language; part-words had a TP of 1.0 between the first and second syllables and a TP of 0.33 between the first and second syllables and a TP of 1.0 between the second and third syllables). They found that 8-month-old infants preferred listening to the part-words from the artificial language, suggesting that the infants had correctly segmented words from the language and now found the sequences with lower TPs more novel and hence more interesting.

These results have been replicated numerous times. With longer exposure to the artificial languages, older children and adults can also track TPs in more-complex artificial languages and use that information for speech segmentation.

The ability to track TPs is not specific to speech, sound, or even humans. Saffran and colleagues have found that human infants and adults are successful at segmenting tone 'words' – tone sequences with high internal TPs – from a continuous stream of tones, suggesting that this type of statistical learning is not tied to speech or language. Researchers have also found that cotton-top tamarins can segment trisyllabic words from an artificial language on the basis on TPs alone. Following 20 min of exposure to an artificial language, tamarins were more likely to orient in the direction of a concealed speaker if they were presented with a either a nonword or a part-word than if they were presented with a word from the artificial language. This finding suggests that the ability to track and segment speech on the basis of TPs is not special to humans and that the ability probably did not evolve specifically for language. Rats are also able to segment words from continuous speech on the basis of TP. Finally, both human infants and adults are successful at extracting sequences of visual shapes on the basis of their TPs. This type of statistical learning is not constrained to the auditory domain.

Given that the ability to track TP is so ubiquitous across perceptual domains and across species, how likely is it that the abilities demonstrated in these laboratory speech segmentation tasks are actually applicable to real-world speech segmentation and word learning? In a study designed to address this question, 17-month-old infants were familiarized with an artificial language that contained no acoustic cues to word boundaries and then tested for the ability to treat words from the language as labels for objects. Infants learned object–label associations only when the label was a statistical word from the previously familiarized artificial language. Infants failed to learn the object–label association if the label was either a part-word or a nonword. Thus, statistical words do, in fact, make good word candidates. Statistical learning is likely to be one of the learning strategies employed by infants during language acquisition.

The Transition from Language-General Statistical Cues to Language-Specific Linguistic Cues in Speech Segmentation

The ability to track purely statistical regularities (e.g., distributional shape and TPs) in the linguistic environment is essential for the initial stages of language acquisition. Once infants have discovered some basic linguistic elements, such as words, they can then use this information to bootstrap into discovering more linguistic regularities, such as phonotactic and rhythmic patterns. This bootstrapping also involves statistical learning. For example, infants raised in an English-speaking environment learn that some sound clusters, such as 'dl,' are not permitted at the beginning

of words in English but that they can be heard at word boundaries, as in the phrase 'good luck.' Once infants have learned the types of sounds that are allowed to occur together at the beginning of words, hearing an illegal sound cluster (such as dl) provides a good cue to word boundaries. Given a novel sequence such as 'murdlarn,' English-learning infants are more likely to segment it as murd + larn than as mur + dlarn because dl is an illegal initial sound cluster in English.

Stress can also provide a cue to word boundaries. In English, the majority of two-syllable words have stress on the first syllable (e.g., BAby, KItty, BOttle, TAble). In Spanish, the second to last syllable of a word is usually stressed (e.g., meSEta, gaTIto, caBAllo), and in French the final syllable of the phrase is stressed (e.g., petit beBE, grand chaTON, pain doRE). Once infants learn the association between stressed syllables and words, they can use stress as a cue to find words in speech. There is evidence that 9-month-old infants from English-speaking homes will use stress instead of TP to segment continuous speech when the two cues are in conflict, missegmenting statistical words in favor of extracting words that have a trochaic (strong–weak) stress pattern. In contrast, younger infants rely heavily on TP for speech segmentation, ignoring stress as a marker for word boundaries. When familiarized with a series of strong–weak nonsense words, however, even younger infants will switch to a stress-based segmentation strategy. Linguistic experience seems to be necessary for infants to violate purely statistical regularities in favor of more language-specific segmentation strategies.

Why would infants switch from using purely statistical cues, such as TP, to using more language-specific cues when the two cues are in conflict? One possibility is that using language-specific cues is ultimately easier. Tracking statistical regularities requires the learner to be able to perceive and retain information across instances and then perform computations on that information. As might be expected, statistical regularities and language-specific cues are rarely in conflict in natural languages. If a cue provides a good indicator to word boundaries, it does so precisely because its presence maps onto differences in TPs between and within words. It is the consistency of this mapping, and the redundancy it provides, that directs its usefulness for segmentation. If stress pattern (e.g., a strong–weak pattern in English or a weak–strong pattern in French) and syllable TP were uncorrelated, stress would likely not be used as a cue to word boundaries. These linguistic cues likely function in conjunction with purely statistical cues and provide a rich system of regularities for the language user to absorb.

Finding the Structure: Statistical Learning, Grammatical Structure, and Predictive Dependencies

Phonetic distribution shape and syllable TP provide only some of the statistical information necessary for infants to acquire language. Infants must also gather information about grammar and syntax. Since the 1960s, the orthodox view has been that infants are born with innate linguistic knowledge of a universal system of grammar (or universal grammar). According to the strong version of the theory, initially proposed by Chomsky, aspects of language that are learned, are learned through neural mechanisms that have evolved specifically for language. This neural mechanism is sometimes referred to as a language acquisition device. Research on statistical learning has suggested that, like phonetic acquisition and speech segmentation, learning at least some aspects of grammatical structure can be accomplished through general statistical learning mechanisms.

Tracking grammatical structure is substantially more complex than tracking distributional shape or TP. Learners must be able to abstract form classes (like determiners, adjectives, nouns, and verbs), have a rudimentary understanding of how form classes relate to how phrases are structured, and then be able to combine smaller phrasal units to form larger units – all within the structure allowed by a given language. Results from computational modeling suggest that word co-occurrence information available in large language corpora is sufficient for determining abstract form classes, such as parts of speech and the phrases in which they occur. Grammatical systems are also rife with predictive dependencies, such that the presence of one type of form class will predict the presence of another form class. For example, in English the presence of a determiner (e.g., 'a' or 'the') predicts the occurrence of a noun later in the sentence. The reverse relationship does not exist. Nouns often occur without determiners.

Saffran created an artificial language to investigate whether adults and children could learn rules about how phrases are structured based on the types of predictive dependencies found in natural languages. In Saffran's artificial language, sentences were made up of required and optional phrases. Each phrase was also made up of required and optional words. For example, in the language, A-words (i.e., biff, hep, mib, and rud) could occur with or without a following D-word (i.e., klor and pell), but a D-word was always preceded by an A-word. Saffran predicted that this type of statistical relationship should help learners discover phrase structure. Following exposure to

the artificial language, adult learners judged new sentences that followed the rules of the language to be more grammatical than new sentences that violated one of the rules. Predictive dependency was one of the primary factors that affected grammaticality judgments. That is, novel sentences that violated the predictive dependencies found in the artificial language were judged to be ungrammatical. Although there were a number of irrelevant surface features over which participants could have calculated statistical regularities, doing so would not have resulted in accurate grammaticality judgments. Success on the task demonstrates that human learners can use predictive dependencies in the service of acquiring simple phrase structure.

Infants and tamarins can also use predictive dependencies to discover structure in strings of words. After listening to a simplified artificial language, where predictive dependencies exist between individual words instead of between word classes, both infants and tamarins listened longer to an ungrammatical string of familiar words than to a different string made up of the same words that preserved the predictive dependency relations of the artificial language. When these two groups were presented with the more complex artificial language, in which the predictive dependencies were between word classes instead of individual words, only the human infants were capable of using dependencies among classes of words to acquire structured phrases.

Using a similar paradigm, it has been demonstrated that 12-month-old infants are able to learn novel grammatical structure following a short exposure (2 min) and that they can generalize the learned grammatical structure to new vocabulary. For example, if infants were exposed to a string such as FIM-SOG-FIM-FIM-TUP (with structure A X A A Y), they were able to abstract the learned grammar to the string PEL-TAM-PEL-PEL-JIG. In this experiment the TP of the test stimuli was zero because the vocabulary was new. This suggests that 12-month-old infants abstract beyond first-order word dependencies, even when the dependencies are only probabilistic in nature, to learn grammatical structure.

Constraints on Statistical Learning

It is evident that both human and nonhuman animal listeners are good at extracting regularities in the linguistic environment and that these abilities are relevant to the task of language learning. Natural languages contain statistical information that powerful statistical learning mechanisms could track but that would be irrelevant for the human language learner. For example, the distribution of syllables in English is such that it would be possible to track which words follow words whose second syllable begins with *th* (e.g., tether *ball*, mother *dearest*, lathered *up*) or which words are presented third in sentences (e.g., Once upon *a* . . . , The very *next* . . . , The white *snow* . . .), but such information is irrelevant to the task of language acquisition. The problem faced by the language learner remains essentially the same as discussed in the first section; there are infinitely many generalizations that are consistent with a given linguistic corpus over which the learner must compute the relevant statistics. Statistical learning must therefore be constrained or biased in order to reduce the acquisition problem.

Fortunately, language learners are biased to extract regularities that are relevant to natural languages. Artificial languages that have grammatical structures found in natural languages, like predictive dependencies, are easier to learn than artificial languages that do not have these naturally occurring structures or dependencies. These findings are true for adults, children, and infants. This suggests that natural languages might be structured in the way they are because predictability facilitates the discovery of underlying structure. There are additional types of constraints that may inform statistical learning and might also help structure language. For example, there are auditory constraints that make certain sounds easier to discriminate and easier to learn, and languages exploit such constraints when selecting their sound inventories. The study of constraints on statistical learning is even newer than the young field of statistical learning. It does, however, suggest that the constraints on learning may actually help structure language itself.

See also: Bilingualism; Connectionist Models of Language Processing; Language Development; Language: Cortical Processes; Language: Aphasia; Language: Learning Impairments; Lexical Impairments Following Brain Injury; Second Language Acquisition; Sentence Production; Sentence Comprehension; Word Production; Word Recognition; Word Learning.

Further Reading

Aslin RN, Saffran JR, and Newport EL (1998) Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9: 321–324.

Conway CM and Christiansen MH (2005) Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(1): 24–39.

Gomez RL and Gerken LA (1999) Artificial language learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 70: 109–135.

Graf-Estes K, Evans JL, Alibali MW, and Saffran JR (2006) Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science* 19(3): 254–260.

Hauser MD, Newport EL, and Aslin RN (2001) Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition* 78: B53–B64.

Maye J, Werker JF, and Gerken L (2002) Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 83(3): B101–B111.

Newport EL and Aslin RN (2000) Innately constrained learning: Blending old and new approaches to language acquisition. In: Howell SC, Fish SA, and Keith-Lucas T (eds.) *Proceedings of the 24th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Saffran J, Johnson E, Aslin R, and Newport E (1999) Statistical learning of tone sequences by human infants and adults. *Cognition* 70: 27–52.

Saffran JR (2001) The use of predictive dependencies in language learning. *Journal of Memory and Language* 44: 493–515.

Saffran JR (2002) Constraints on statistical language learning. *Journal of Memory and Language* 47: 172–196.

Saffran JR (2003) Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science* 12: 110–114.

Saffran JR, Aslin RN, and Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274: 1926–1928.

Thiessen ED and Saffran JR (2003) When cues collide: Statistical and stress cues in infant word segmentation. *Developmental Psychology* 39: 706–716.

Statistical Tests and Inferences

A L Shelton and A S Greenberg, Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Functional neuroimaging has become a popular method for investigating functional brain anatomy. As various neuroimaging techniques have developed, there has been a commensurate increase in the development of techniques for analysis. One of the critical issues surrounding neuroimaging analyses has been how to establish a valid assessment of statistical significance. Although details in the approach can vary, the general process involves taking the substantial amount of data collected and reducing it to meaningful patterns of functional brain activation. To discuss the typical procedures for accomplishing this task, this article focuses on the analysis of functional magnetic resonance imaging (fMRI) data; however, many of the concerns and techniques are applicable to other brain activation methods involving time series data. This article reviews the basic issues surrounding data reduction, preprocessing, individual analyses, group inference, threshold setting, and alternative analysis techniques.

fMRI Analyses as Massive Data Reduction

Like many neuroimaging techniques, fMRI involves collecting a massive amount of data. Data are in the form of snapshots of the brain in action taken over several time points (temporal frames). Each snapshot of the brain is made up of thousands of voxels or cubes of brain area; that is, each image of the brain contains a scanner intensity value for each voxel at the time the image was acquired. Taken together, each voxel can be characterized by a time course of values, with one value for each temporal frame. The goal of the analysis is to reduce this massive amount of data into meaningful information about brain function. For example, a whole-brain image taken at a resolution of 4 mm × 4 mm × 4 mm (30 slices, 24 cm field of view, 64 × 64 acquisition matrix) yields approximately 120 000 voxels. In a single 1 h scan session, ten functional runs of 3.6 min each, with a repetition time (TR) of 1000 ms, yield 2160 values for each voxel. These 2160 values at each of the 120 000 voxels must be organized according to (1) what was happening cognitively during particular periods of time (e.g., what task was the participant doing at any given time) and (2) where in the brain the activity was fluctuating during those different cognitive activities. This article presents the typical procedures involved in going from large volumes of data to meaningful statistical conclusions and the issues that arise in their application.

Preprocessing

During the acquisition of a typical fMRI scan, several technical details must be reconciled before the data can be analyzed for meaningful patterns. Although the particular order of these steps may vary according to the preferences of particular researchers, the purpose and processes remain similar. First, when a given volume of the brain is acquired in a single TR, that is, one complete scan through the whole brain, the individual slices differ in the time at which they were acquired. For example, if 20 slices are acquired sequentially from the top of the brain to the bottom of the brain at a TR of 1.5 s, the first slice and the last slice are acquired 1.425 s apart. Therefore, in order to better synchronize the brain activation changes that occur throughout the scan, the volumes are subjected to slice-timing correction. Second, during a scan session, participants are asked to remain still, but there will naturally be some small amount of head movement. To account for this movement, the volumes must be realigned to put all the acquired images into the same space. In addition to correcting for head motion, this step will provide data on how much movement occurred in each direction; such information can be used to assess effects due to motion. A similar realignment (also called coregistration) is performed to align anatomical images to the functional images for localization and display purposes. If group analyses are planned, it is further necessary to get all participants' brains into the same shape and alignment. This is generally accomplished through warping or normalization of each participant's brain images to a single standard or template brain. Neuroimaging data coming from the scanner are inherently noisy. One method for addressing this noise is to use spatial smoothing, which typically involves taking a moving average of voxel values to create larger voxels. The degree of smoothing depends on the size and type of filter applied. Smoothing is typically performed with a Gaussian-shaped smoothing kernel often specified by its full width at half maximum. This value is generally chosen to approximate the expected size of functional activation blobs in the final statistical parametric maps (SPMs). Although this procedure reduces the spatial resolution of the image, with the appropriate selection of a filter, it can substantially improve the sensitivity by

overcoming the considerable noise present in the data (i.e., increasing the signal-to-noise ratio). Smoothing helps whiten the noise (an important consideration both when employing Gaussian random field theory to correct for multiple comparisons and when subjecting the data to a general linear model, as discussed below). Finally, there are a number of ways that data can be filtered to remove systematic noise that is related to factors other than the cognitive tasks of interest. For example, low-frequency changes such as linear drift of the intensity values are customarily removed, either by applying a filter during preprocessing or as part of the multiple regression used during the analyses described in the next section.

Figure 1 shows the time course for a voxel in the parietal cortex in a single participant over the course of a single scan run. The values reflect the raw signal and after various stages of preprocessing including slice-time correction, motion correction/coregistration, normalization to percentage signal change (to eliminate differences in the raw signal intensity that occur across scans), and spatial smoothing.

Modeling Data for Individual Participant

In its simplest form, the question of interest when looking at an individual participant is whether the activation in each voxel is fluctuating in conjunction with the changes in the cognitive states that were engaged during the scan session. To model the data, each time point of the scan run can be 'labeled' according to the type of stimulus being presented and/or the type of mental manipulation in which the participant was supposed to be engaged. Figure 2 shows the scan scheme for an event-related design in which participants performed mental rotation over the course of five scan runs. In each run, participants were presented with four different trial types: no rotation, 60° rotation, 120° rotation, and mirror-image trials in random order, approximately every 6 s. (The mental rotation example used throughout this paper is for illustrative purposes only. The data are only partially represented.) To code these data for analyses, correct and incorrect trials were separated out. The different solid, colored lines in Figure 2 indicate the time at which there were correct instances of each of the different trial types throughout one run of the experiment; dashed lines show trials that were answered incorrectly. There are several methods for organizing and understanding the relationship between the time course of each voxel and that of the cognitive tasks; three of the common approaches are described here.

Multiple Regression

The most common method for modeling the data from a single participant is to conduct a multiple

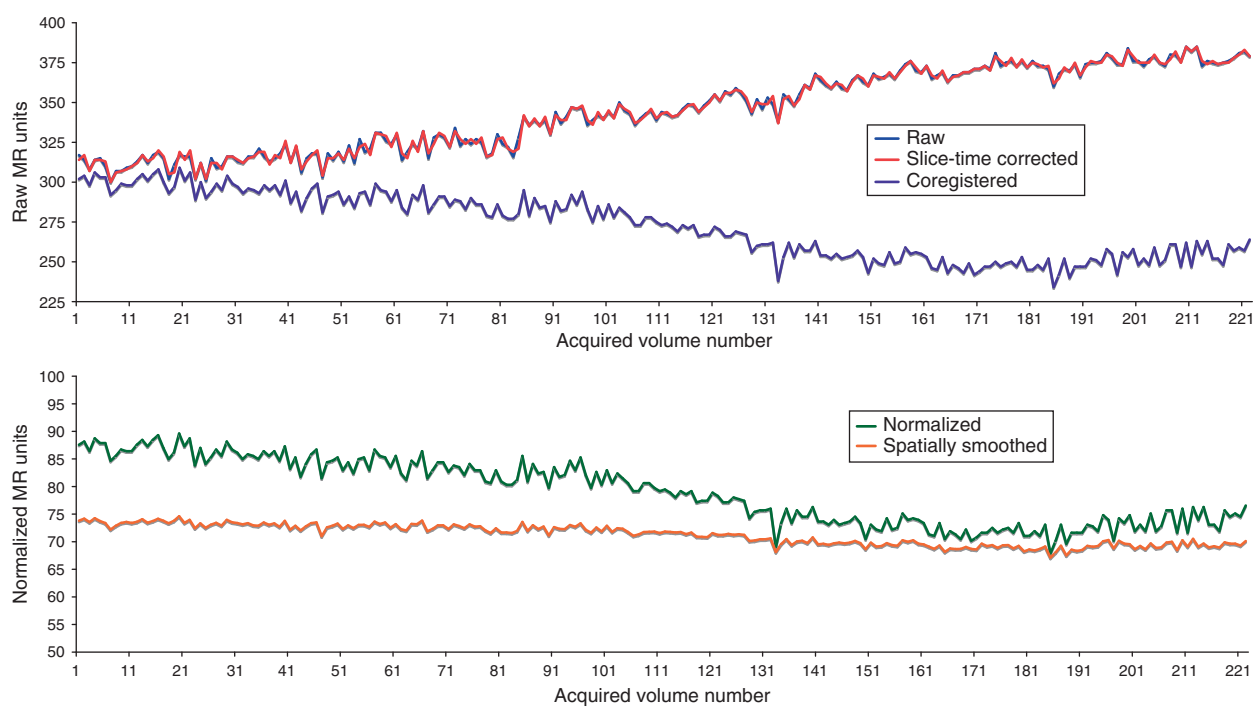

Figure 1 A single voxel time course from the left superior parietal lobule of a single subject over the course of one run. Blue line is the raw signal off the scanner; red line has been slice-time corrected, purple line has been motion corrected (coregistered), green line has been normalized by the mean of the run, and orange line has been spatially smoothed with nearby voxels in the brain.

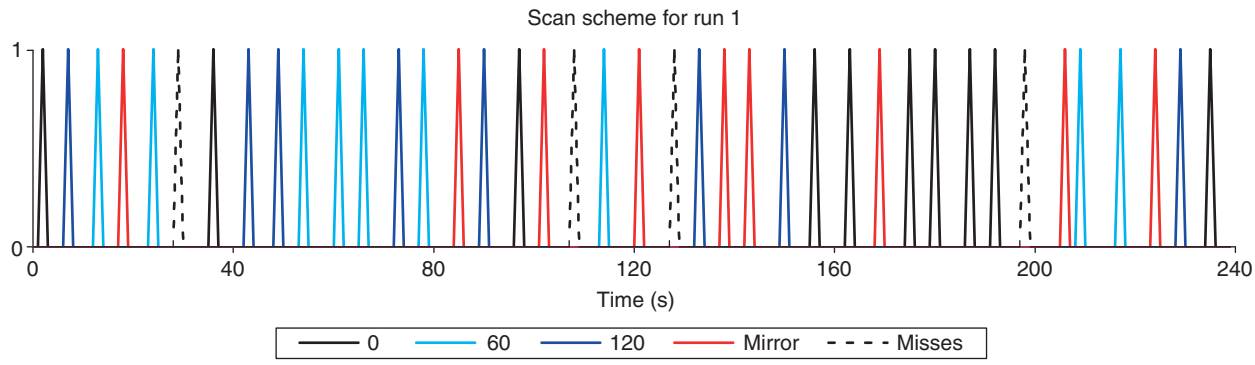

Figure 2 Scan scheme for a single 240 s run of an event-related mental rotation experiment. Peaks reflect the time point at which a given trial began, color coded to reflect the correct trials for no rotation (solid black), 60° rotation (cyan), 120° rotation (blue), and mirror-image trials (red), plus all incorrect trials (dashed black).

regression analysis at every voxel. The preprocessed voxel time course serves as the dependent measure, and a general linear model is built with regressors constructed to (1) control for extraneous variables of no interest (e.g., session-specific variations and linear drift) and (2) test the effects of interest. Each regressor is constructed according to the expected pattern of activation for the effect of interest. For example, Figure 3 shows the design matrix for a possible set of regressors in the mental rotation experiment in Figure 2. The first 48 regressors represent 12 different scan runs. The following six regressors represent the different possible trial types from the experimental conditions. This is equivalent to separating out each line in Figure 2 such that each regressor reflects the time points at which one trial type occurred during the scan runs.

In fMRI, the blood flow to a brain region is sluggish compared with the timing of the cognitive events and tends to show a smooth increase and decrease over time. As such, prior to modeling the voxel time course, regressors may be time shifted and smoothed to match the expected hemodynamic response function (HRF). For example, one might use a gamma function as the canonical HRF (Figure 4). This HRF convolution increases the likely similarity between the expected and observed time courses in voxels that are responding to particular cognitive events. These convolved regressors are then used to conduct the multiple regression on each individual voxel in the brain separately. The basic equation is as follows:

$$y(t) = \mathbf{BX} + \varepsilon_t \qquad [1]$$

where **B** represents the vector of to-be-solved regression weights (βs), one for each of the regressors, **X** represents the design matrix with the convolved regressors, and ε_t is the error.

The output of the multiple regression is a β at each voxel for each regressor. The magnitude of a given β

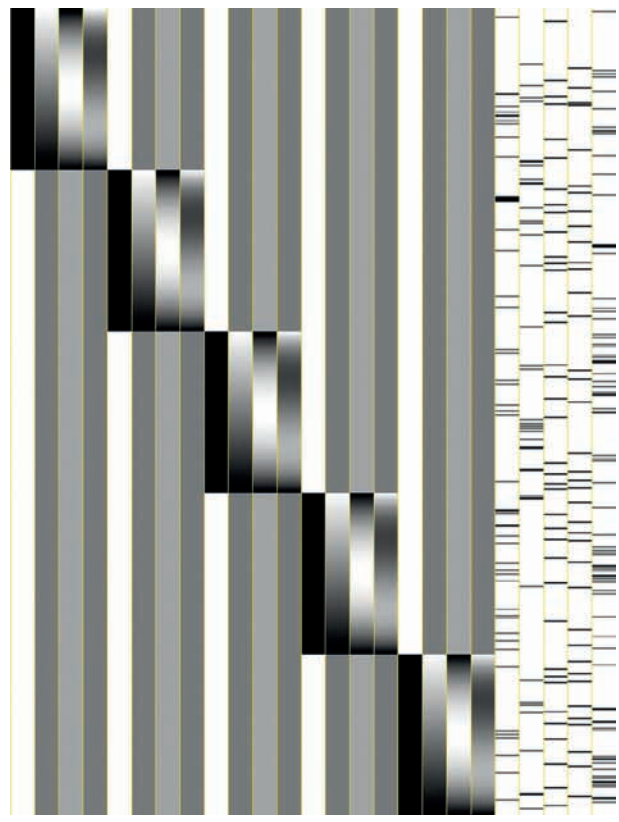

Figure 3 An image of the design matrix generated using AFNI from the example fMRI experiment in which there were five runs of data along with five regressors of interest. Each column is a graphical representation of a single predictor entered into the model. Each row corresponds to a time point from the beginning of run 1 through the end of run 5. White regressor time points correspond to a value of 0, black corresponds to 1, and grayscale colors are intermediate values. The first 20 regressors are used to account for trending components (0, first, second, and third order) from each of the five runs, separately. The final five regressors of interest correspond to correct responses to the four trial types (conditions) in the experiment plus one additional regressor to model all incorrect trials. Each regressor has been convolved with a hemodynamic response function beginning at each time point when a stimulus of that type occurred.

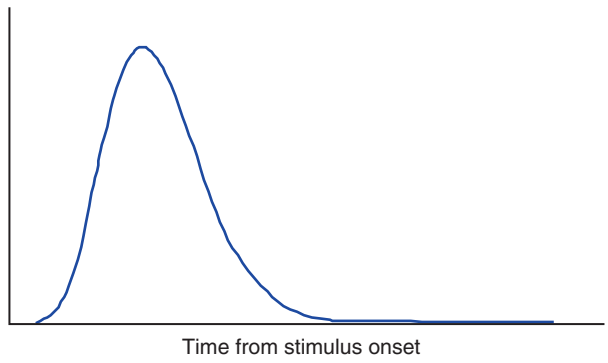

Figure 4 An example of a gamma function that might be used as the canonical hemodynamic response function to adjust regressors for the temporal lag of the brain's blood oxygen level-dependent response.

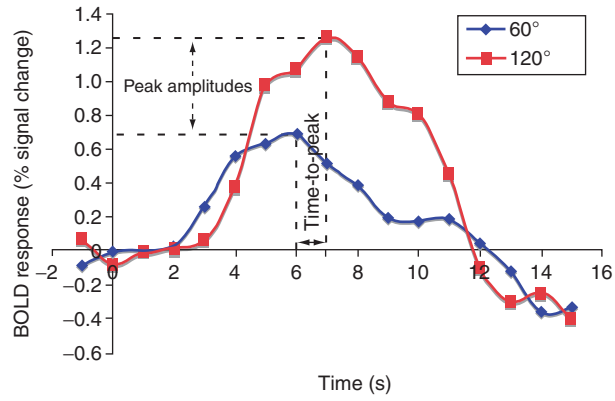

Figure 5 Examples of event-related time courses from the right superior parietal cortex in a single individual. In this case, the time course reflects as single voxel (30, −48, 51 in Montreal Neurological Institute (MNI) coordinates). Additional potential dependent variables are indicated with dashed lines.

reflects how well that regressor predicts the activation pattern in that voxel. The comparison of these β values for different conditions of interest can provide an index for increases and decreases in relative activation. For example, if a particular brain region is more involved in task A than in task B, then the β value for the task A regressor will be larger than the β value for the task B regressor. It is these relative relationships that provide the information needed to make statistical inferences about which brain regions are significantly active for one experimental condition versus another. As such, these β values (or various combinations of them) serve as the dependent variables for subsequent analyses. The statistical inferences will be based on placing thresholds on these β values or the differences among them. At this stage, one can create an SPM for the individual participant, showing the regions of the brain that meet a particular statistical threshold (see the section titled 'Thresholds and correction for multiple comparisons').

Event-Related Averaging and Deconvolution

When dealing with blocked designs in fMRI, multiple regression has remained a standard approach. However, concerns have been raised about whether the assumptions underlying the multiple regression approach – in particular, the need to specify the shape of the HRF – are more problematic when the design involves individual trials. Therefore, for analyzing event-related designs, other options are available that attempt to use the observed data to establish the hemodynamic response for each event type of interest directly from the data.

The first alternative, event-related averaging, involves segmenting the time course of each voxel for each event type by identifying a peri-event time window, defined by the time point of a given event, a specified number of preevent time points, and a specified number of postevent time points. For example, one might specify a peri-event window beginning 2 s before an event and including the 16 s including and following an event. This same 18 s window would be selected for each occurrence of a given event type throughout the scan session. For each of the frames in the 18 s window, there would be a mean value across all instances, resulting in an average 18 s time course for the event (Figure 5). The size of the window around the event must be carefully determined in order to correctly capture the event-related activation. In slow event-related designs, the window must be selected to avoid taking temporal frames that are associated with the trials before or after the trial of interest. In rapid event-related designs, the window must capture the trial, but one must also ensure that the history of each trial type is equivalent (i.e., that the trial types have been appropriately randomized or counterbalanced). The output for each individual is a set of average HRF time courses, one for each event type at each voxel. From these event-related averages, additional variables can be derived for each trial type: time for the HRF to peak, peak amplitude, area under the HRF curve, and so forth.

The second alternative, deconvolution, refers to taking the fMRI time course and breaking it into separable components. Rather than convolving the trial structure with a canonical HRF, a mutiple regression is constructed to determine the shape of the HRF for each trial type. Unlike simple event-related averaging, deconvolution uses modeling to provide an estimate

of the magnitude of the response at each time point following the onset of a stimulus (or condition). The output of the deconvolution operation is the time course based on these estimates for each event type. This time course can be passed back into the statistical model in order to further contrast conditions. Like event-related averaging, these estimated time courses can be used to calculate additional dependent measures for comparison.

Group Analyses

Regardless of the processing stream used to analyze individuals, the goal of a group analysis is to identify common sources across individuals in order to make inferences about how the brain functions in the relevant population. By combining data across individuals that have been warped to a common space, one can utilize many standard statistical tests to assess differences among conditions. In a simple blocked design, the individual participants' β weights for two conditions can be used to calculate a paired *t*-test at each voxel in the brain. By extension, one can calculate an analysis of variance at each voxel for more than two conditions or determine which voxels have significant correlations with independent behavioral measures. At each voxel, one can then determine whether the test is significant on the basis of the threshold one has set as a criterion (the chosen α level; see the section titled 'Thresholds and correction for multiple comparisons'). More recently, nonparametric analyses have become increasingly popular and can be used in a similar manner.

In the mental rotation example, the β values for no rotation, 60° rotations, and 120° rotations from each individual were subjected to an analysis of variance (ANOVA); the resulting group SPM is shown in Figure 6. Several areas showed significant activation at various thresholds, indicating that these regions have differential activation in the conditions of interest.

Thresholds and Correction for Multiple Comparisons

The goal of presenting the SPM from an individual or group analysis is to visualize how the brain is 'activated' for a particular comparison. The pretty pictures of colored blobs overlaid on top of anatomical brain images commonly seen in scientific journals these days are a result of statistical thresholding. In principle, this SPM tells the reader which locations in the brain enjoyed a statistically significant amount of signal modulation for the specific test or contrast between conditions. One critical issue in determining what one will or will not see in the SPM resulting from either an individual or group analysis is the selection of a statistical threshold.

Nominally, a statistic can be computed for any independent observation (say, a voxel β weight value) and an appropriate α level can be set. The α level is defined as the probability of finding a false positive (a type I error in statistical terms). When a single voxel is under consideration, the relevant threshold is the height or intensity threshold. Height thresholds are set based on the probability that a particular voxel will have a statistical value (e.g., the t- or z-value). For example, with an uncorrected α level of 0.001, any voxel that has a *t*-value with a probability of 0.001 or smaller will be shown in color in the resulting SPM. In the typical SPM, the particular color conveys information about the magnitude of the statistic or probability associated with that voxel. Returning to the example, Figure 6 shows activation for the rotation versus no rotation contrast at three different height thresholds.

Setting a height threshold for each voxel (an uncorrected α) can be misleading. It assumes that each voxel is an independent test; however, fMRI involves repeating the statistic test at thousands of different voxels in the brain. As such, the false positive rate is much higher because of this repetition. To address the inflated false positive rates, several methods for

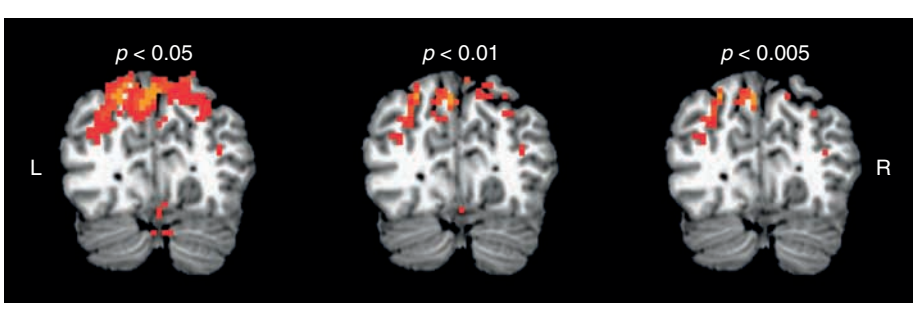

Figure 6 Example of a single slice ($x = -73$) from a group statistical parametric map showing areas of greater activation for mental rotation compared with no rotation at three different height thresholds (*p* values).

correcting the threshold have been developed. Some of the commonly used methods are discussed here.

The most stringent method is a Bonferroni correction, as is typically used in other realms of psychology and neuroscience. This correction is based on dividing the intended family-wise error rate (P_{FWE}) among the number of tests conducted. For any family of n values, the probability that all n tests will be less than α is $(1-\alpha)^n$. Therefore, the $P_{FWE} = 1-(1-\alpha)^n$. With small values of α, one can assume that $P_{FWE} \leq n^*\alpha$. To achieve the intended P_{FWE}, one simply tests each voxel at $\alpha_n = P_{FWE}\ n^{-1}$. The problem with such a correction harkens back to the wealth of data in standard fMRI. That is, if one is performing t-tests on 66 000 voxels, the appropriate correction for a desired $\alpha = 0.05$ would require voxels to be significant at height thresholds of 7.6×10^{-7}! An alternative use of the Bonferroni correction takes into account the spatial correlation in the brain; the number of independent values is actually smaller than the number of voxels because of correlations among neighboring voxels (due to physiology, preprocessing, etc.). Therefore, the correction can be adjusted according to the number of independent values, reducing the n of interest and increasing the height threshold per n.

One major issue with using the adjusted Bonferroni correction is the proper identification of the number of independent values. An alternative is to use random field theory (RFT) to find an appropriate correction. Put in simplified form, RFT correction involves first establishing the number of resolution elements (resels; Huettel et al.) on the basis of the total number of voxels and the smoothing kernal used in preprocessing ($n_{resels} << n_{voxels}$). Based on these resel counts, it is possible to estimate how many blobs of activation would be expected at different thresholds. These estimates reflect the expected Euler characteristics of the image and can be used to determine corrected thresholds. For example, if a particular Z threshold has a Euler characteristic of 0.05, then any observed activations can be said to have a probability of 0.05 or less due to chance. Thus, the corrected α becomes 0.05. In practice, RFT correction also accounts for the size and shape of the activated blobs to adjust the probabilities. RFT assumes that the noise present in the data is white (relatively equal contributions at all relevant frequencies), thus making it critical to perform spatial smoothing during preprocessing.

Bonferroni and RFT both control for error by adjusting the likelihood of any particular activation being a false positive. An alternative is to control the expected proportion of false positives using the false discovery rate (FDR). The FDR is determined from the distribution of probability values in the data. The basic approach is one of conditional probabilities. At a given threshold, some number of voxels will emerge as significant (N_S). FDR techniques then estimate how many of these are expected to be false positives. Multiple methods have been proposed for controlling these values, but they share the same basic approach. In general, this method is adaptive for neuroimaging because it depends on the amount of signal present. As such, it can allow for weaker designs. However, there are a number of concerns that have been raised about its practical application.

An additional approach to constraining the number of tests conducted is to constrain the search volume of interest. The use of small volume correction involves reducing the total number of voxels of interest by focusing on one or more specific regions. Unlike region of interest (ROI) analyses (see the section titled 'ROI analyses'), this approach imposes the restriction on the SPM. This approach can be effective when there are specific hypotheses about where activation is expected. Correction is still accomplished using one of the above corrections (Bonferroni, RFT, or FDR), but the total number of voxels in the volume of interest is substantially reduced. Conclusions are restricted to only those areas interrogated in the reduced brain volume.

The preceding methods of correction focus primarily on ways of adjusting the height threshold. Thresholds can be corrected by considering how meaningful it is to see activation in a single voxel. Even small brain regions (e.g., amygdala) are made up of multiple voxels. As such, we are typically looking for contiguous activation of voxels. To address this issue, one can set a threshold for contiguity: an extent threshold. Extent thresholds are set on the basis of the probability of having significant voxels (at a given height threshold) in close spatial proximity. Different combinations of height and extent can be used to establish the same corrected value, allowing some flexibility as a function of the regions one is most interested in observing.

ROI Analyses

In most neuroimaging studies, the researcher has clear hypotheses about the brain regions likely to be involved. As such, one approach to analyzing the data is to examine the fluctuations in activation that are occurring within those prespecified regions. ROI analyses can be used in a number of ways. The different methods are based on the way the ROIs are determined, and these methods can result in approaches that either are purely descriptive or allow for additional inferences.

For descriptive purposes, ROIs can be used to enhance and/or clarify other analyses. If a t-test or

ANOVA shows a significant cluster of activation, one can extract the mean β value for that region in each of the tasks and plot the values. Similarly, a region that shows a significant correlation with behavioral measures in a group SPM could be used as an ROI to extract the β values for each participant. This extraction would allow a plot of the correlation to help illustrate where the significant value came from. This ROI approach uses the observed differences to define the ROIs and is inherently for illustrative purposes only. Because the ROIs are not independently defined, they cannot be used for additional inferential tests.

When an analysis involves more than two conditions, the appropriate test at each voxel may be an ANOVA. Such an analysis can only determine voxels (and clusters) in which the conditions show a difference. To understand those differences, one must further analyze the differences between pairs of conditions. Rather than doing additional whole-brain contrasts for each pair, one can use the significant clusters defined by the ANOVA as ROIs. Using coordinates defined by the group analysis, one can extract the relevant β weights for each participant and use those values for *post hoc* graphing and contrasts. This approach still uses the original data set to determine the ROIs, but it can allow additional inferences about differences among conditions. For example, Figure 7 shows the activation in the superior parietal ROI that was significant in the ANOVA in Figure 6. By extracting the β values and subjecting them to ANOVA, one could easily make and visualize subsequent pairwise comparisons. Similar approaches may also involve using one contrast to define ROIs for other contrasts of interest.

Finally, ROIs can be used in place of a group analysis. For this type of analysis, ROIs must be determined independently of the data being analyzed. There are several ways that ROIs might be determined. First, one might use previous functional studies that have identified particular regions of activation that might be relevant. Provided these ROIs can be defined in the coordinates of the same standard space as the participants, then the data can be extracted for each individual in the defined region. For example, if previous studies have localized a subregion of the dorsolateral prefrontal cortex for object working memory, one could use the coordinates from those previous studies to define the ROI. For other regions, anatomical distinctions might be used to define ROIs in each person individually. For example, one might use known anatomical landmarks to identify the location and extent of each person's calcarine sulcus or hippocampus. Finally, one can also use localizer scans to define ROIs separately from the functional scans of interest. For example, if one is interested in attentional modulations in the visual cortex, one could first use flashing checkerboards to determine the retinotopy. In each of these cases, the ROIs are determined prior to analyses, and the main analyses are performed on the information extracted from the *a priori* ROIs.

In all the ROI analyses, the data of interest can take many forms. First, one can look at mean β weights for different conditions, as in Figure 7. This simple bar graph is appropriate for event-related designs and would be the primary outcome for any block design. Alternatively, for event-related designs, one could plot the group average of the event-related time courses for each condition in the ROI (Figure 8). From time courses, one could further evaluate different variables of interest, such as the time to peak, area under the curve, peak amplitude, or some combination of these measures; however, these measures should come from the individual event-related time courses rather than from the group (see Figure 4).

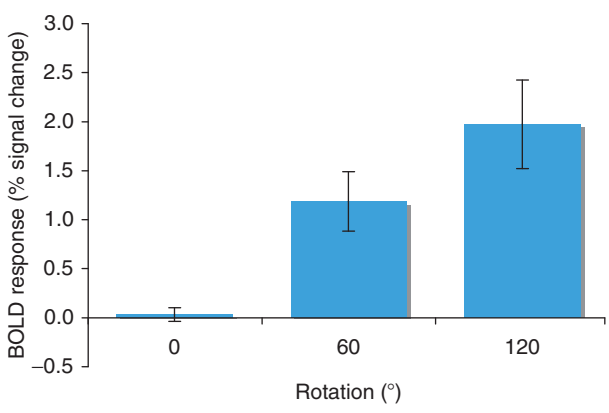

Figure 7 Mean b weight as a function of condition extracted from the right superior parietal cortex region of interest outlined in Figure 6. Error bars are SEM (std. error of mean).

Other Techniques

The preceding discussion is about general approaches to significance testing in fMRI. The majority of the topics apply to other types of time series data, particularly those with large data sets. However, there are numerous other techniques for making inferences about brain activation. Nonparametric tests have been suggested as a way of addressing concerns about the assumptions underlying the typical general linear model approaches. Techniques such as principal components analysis, independent components analysis, functional connectivity, and effective connectivity can be used in place of or in conjunction with the approach outlined above. These additional techniques attempt to provide information not

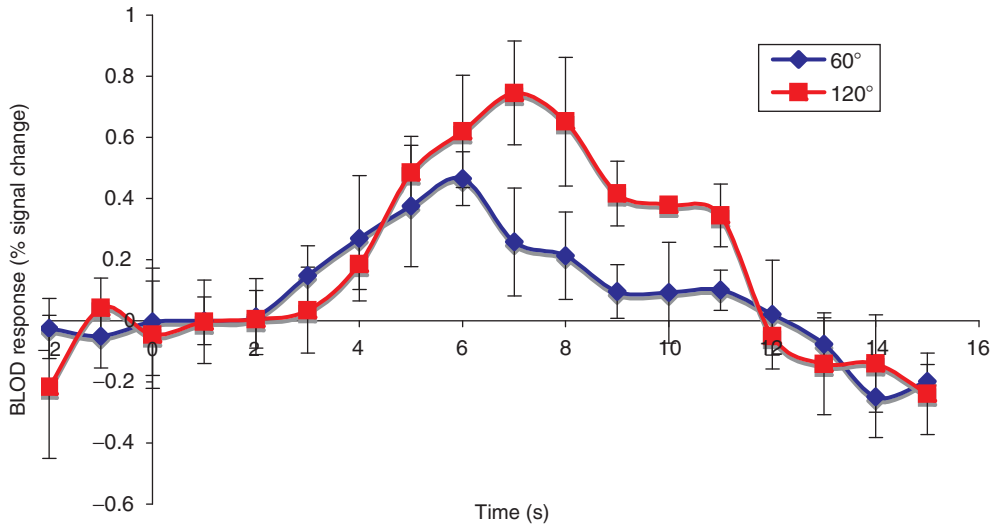

Figure 8 Group event-related average time courses taken from the peak voxel in the right superior parietal cortex region of interest outlined in Figure 6. Error bars are SEM (std. error of the mean).

readily available in a standard analysis. Other techniques will continue to emerge as the field evolves.

See also: Cognitive Neuroscience: An Overview; fMRI: BOLD Contrast; Neuroimaging; Perfusion MRI; Positron Emission Tomography (PET).

Further Reading

Buxton RB (2002) *Functional Magnetic Resonance Imaging: Principles and Techniques*. Cambridge, UK: Cambridge University Press.

Cox RW (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research* 29: 162–173.

Friston KJ, Homes AP, Worsley KJ, Poline JB, Frith CD, and Frackowiak RSJ (1995) Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping* 4: 189–210.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC (1994) Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping* 1: 214–220.

Genovese CR, Lazar NA, and Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* 15: 772–786.

Glover GH (1999) Deconvolution of impulse response in event-related BOLD fMRI. *Neuroimage* 9: 416–429.

Huettel SA, Song AW, and McCarthy G (2004) *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer Associates.

Nichols T and Hayasaka S (2003) Controlling the family-wise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research* 12: 419–446.

Serences JT (2004) A comparison of methods for characterizing the event-related BOLD timeseries in rapid fMRI. *NeuroImage* 21: 1690–1700.

Shelton AL and Pippitt HA (2006) Motion in the mind's eye: Comparing mental and visual rotation. *Cognitive, Affective, and Behavioral Neuroscience* 6(4): 323–332.

Toga AW and Mazziotta JC (1996) *Brain mapping: The methods*. San Diego, CA: Academic Press.

Worsley KJ, Liao CH, Aston J, Petre V, Duncan GH, Morales F, et al. (2002) A general statistical analysis for fMRI data. *NeuroImage* 15: 1–15.

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, and Evans AC (1996) A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* 4: 58–73.

Zarahn E, Aguirre GK, and D'Esposito M (1997) Empirical analyses of BOLD fMRI statistics. *NeuroImage* 5: 175–197.

Relevant Websites

http://www.brainvoyager.com – BrainVoyager fMRI analysis package.
http://imaging.mrc-cbu.cam.ac.uk – CBU Imaging.
http://afni.nimh.nih.gov – Deconvolution analysis of fMRI times series data, by B. Douglas Ward.
http://www.sph.umich.edu – False Discovery Rate.
http://psychology.uwo.ca – fMRI 4 Newbies.
http://afni.nimh.nih.gov – fMRI programs.
http://www.fmrib.ox.ac.uk – fMRI software.
http://www.fil.ion.ucl.ac.uk – statistical parametric mapping.
http://imaging.mrc-cbu.cam.ac.uk – Thresholding with Random Field Theory.

Stem Cells and CNS Repair

R Jandial, Burnham Institute for Medical Research, La Jolla, CA, USA; and University of California at San Diego, La Jolla, CA, USA
A Persson, University of California, San Francisco, San Francisco, CA, USA
A Ho, University of California at San Diego, La Jolla, CA, USA
M L Levy, Children's Hospital San Diego, San Diego, CA, USA
E Y Synder, Burnham Institute for Medical Research, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The myriad neurological pathology that can affect the human brain during prenatal development and throughout life and the brain's limited self-repair capacity call for new therapeutic strategies. Our increasing knowledge about the fundamental biology and therapeutic potential of various stem cell types opened a new chapter in regenerative medicine. The initial work on rodent stem cells during the past two decades is now being successfully continued with stem cells of human origin and from different developmental stages. We have learned about key genes and cellular mechanisms that maintain the status of stem cells or lead to differentiated progeny. We have also learned about the multiple roles of stem cells during development, disease, and aging. It is now well established that stem cells are not only a valuable tool for cell replacement but also equipped with important additional properties that may be harnessed for cell protection, detoxification, and gene therapy ('chaperone effects').

Stem cell biology is being recognized as a continuum of development, and developmental processes are tightly regulated, both temporally and spatially. Better understanding of these developmental events is considered to be a key strategy for the successful use of stem cells (endogenous and grafted) for central nervous system (CNS) repair and functional recovery. For instance, the generation of functional neurons and glial cells during brain development requires a concerted coordination of cell proliferation, migration, cell type specification, and synaptic integration, all of which are also crucial for successful stem cell therapy.

Stem Cell Prototypes

Stem cells give rise to organs and maintain tissue integrity and homeostasis in the adult organism. There are different types of stem cells, including embryonic and somatic (fetal or adult derived), from which new cells can be derived. To fulfill the criteria of a stem cell, as opposed to a 'progenitor' cell, a single clonal cell must have the following functional properties: (1) it should be able to generate the cell types from the organ from which it was derived, and (2) it should be capable of 'self-renewal' (i.e., the ability to produce daughter cells with identical properties). The ability to populate a developing or injured region with appropriate cell types upon transplantation is another important stem cell feature that is well established with hematopoetic stem cells and awaits standardization in other organ systems, including the brain. In the following sections, we introduce two prototypical stem cells, embryonic stem cells (ESCs) and neural stem cells (NSCs), and discuss their potentials for neural repair.

Embryonic Stem Cells

ESCs have been derived from the inner cell mass of blastocysts of different species including human. They are pluripotent and able to yield mature cell types from all different germ layers. Work performed with mouse ESCs has provided proof of principle that pluripotent cell lines can be harnessed for developmental biological studies as well as for new therapeutics. Since significant species differences exist between mouse and human ESCs regarding signaling pathways and molecular regulation of pluripotency, it is pivotal to fully characterize and define the molecular mechanisms in human ESCs. Our understanding of human ESCs is increasing and knowledge is accumulating on improved cell culture conditions, long-term propagation, controlled differentiation, and transplantation into animal models of human disease.

The list of various cell types differentiated from human ESCs (e.g., neurons, cardiomyocytes, and hepatocytes) is continuously increasing. Pluripotent ESCs can be stepwise differentiated in the culture dish by recapitulating aspects of *in vivo* development and the use of relevant epigenetic factors. Importantly, the acquisition of a particular developmental stage of a cell is best characterized by considering morphological, immunophenotypic, and functional criteria. Unlimited access to specific functional human cells is expected to play an important role not only in therapeutic cell replacement but also for disease modeling and drug screening.

Neural Stem Cells

In contrast to pluripotent human ESCs, somatic stem cells are believed to be multipotent and thus capable of generating the major cell types limited to the tissue

of origin. Typically, the NSC is capable of producing neurons, astrocytes, and oligodendrocytes. Somatic/tissue-specific stem cells are the building blocks of organs during development and survive in specialized microenvironments ('stem cell niche') contributing to new cells throughout life. NSCs are multipotent (i.e., they have the ability to yield mature cells in all three fundamental neural lineages throughout the nervous system: neurons, astrocytes, and oligodendrocytes), have the ability to populate a developing region and/or repopulate an ablated or degenerated region of the CNS with appropriate cell types, and undergo self-renewal (i.e., the ability to produce daughter cells with identical properties). NSCs have been identified *in vitro*. No study has demonstrated the existence of multipotent NSCs *in vivo*. NSCs are highly abundant during embryogenesis, with a sharp decline soon after birth. In the adult nervous system, NSCs are confined to the subgranular zone (SGZ) in the dentate gyrus of the hippocampus and the subventricular zone (SVZ) lining the lateral ventricles. The newly born neurons in hippocampus have been suggested to improve memory and play a role in mood behavior such as stress and depression. Neuroblasts born in the SVZ migrate along the rostral migratory stream to the olfactory bulbs, where they differentiate into periglomerular and granule neurons. In brain regions such as amygdala, substantia nigra, and cortex, cells with stem cell characteristics *in vitro* have been isolated. Morphologically, NSCs share properties with both astrocytes and radial glia. The main characteristic is a long process that extends radially. Although no definitive marker has been suggested for NSCs, a substantial amount of work shows that they are positive for nestin, an intermediate filament protein, and glial fibrillary acidic protein, traditionally used to identify astrocytes.

NSCs or progenitor cells with a more restricted developmental potential can be generated from human ESCs or directly isolated from the developing CNS as well as from neurogenic regions of the adult brain. Historically, the first established NSC lines exploited knowledge accumulated on tumor viruses and immortalization. These cell lines have been invaluable in expanding our experience on basic stem cell biology and neural repair. Some of these multipotent cell lines, such as the C17.2 NSC line, are still widely used. However, NSCs that have not been genetically modified can also be propagated *in vitro* for extended periods of time using high concentrations of mitogenic factors such as basis fibroblast growth factor (bFGF) and epidermal growth factor (EGF). Neural stem/progenitor cells have been cultured as monolayers on coated substrates or as free-floating spherical aggregated, termed neurospheres.

Stem Cell Repertoire

Chaperone Effects

Initially, stem cells were exclusively considered as tools for cell replacement. However, there is much evidence now for robust additional biological properties (chaperone effects) of stem cells that may be exploited therapeutically. Chaperone effects of stem cells include the natural delivery of neurotrophic, cytoprotective, and anti-inflammatory molecules (e.g., glial cell line-derived neurotrophic factor, brain-derived neurotrophic factor (BDNF), and neurotrophin-3) in order to rescue dysfunctional cells. This concept of stem cell-based chaperone effects was first demonstrated in the brain of aged and Parkinsonian mice and later confirmed and extended to other organ systems and various diseases (e.g., bone marrow-derived mesenchymal stem cells (MSCs) or ESCs for cardiac disease, and umbilical cord cells in stroke).

Environmental Cues

Increasingly, the microenvironments within the CNS are providing insight into the molecular milieu regulating stem cell biology. A specialized microenvironment in the neurogenic regions is responsible for the continued self-renewal and differentiation of the stem cell pool. For example, these regions have a higher density of blood vessels releasing factors, such as vascular endothelial growth factor (VEGF) and BDNF, contributing to increased proliferation and neurogenesis. Different levels of growth factors can have fundamentally different effects on proliferation and differentiation. For example, low levels of insulin-like growth factor-1 (IGF-1) promote neurogenesis, whereas higher levels increase proliferation and oligodendrogenesis. Established protocols for the isolation of NSCs include the growth factors EGF and bFGF that sustain self-renewal of NSCs. Injection of these factors intracranially into mice has been shown to regulate both proliferation and differentiation *in vivo*. Neurotrophins such as BDNF in collaboration with retinoic acid and Wnt signaling regulate neurogenesis, and ciliary neurotrophic factor/leukemia inhibitory factor or bone morphogenic proteins induce gliogenesis. It is also know that the cellular milieu contributes to fate determination and cortical development, specifically through the effects of resident astrocytes in both the SVZ and the hippocampus. Physical exercise and enriched environment have been shown to promote neurogenesis in the SGZ. The effects from physical activity are partly mediated by IGF-1, VEGF, BDNF, and endogenous opioids. The characterization of the stem cell microenvironment will provide the molecular and cellular

scaffold on which stem cell therapy, both endogenous and exogenous, can be built.

Stem Cells as Vectors for Gene Therapy

Brain lesions can be focal and restricted to a certain brain region or widely distributed in the parenchyma. Ideally, both lesion types would be targeted with a specific and efficient delivery of therapeutic molecules and drugs. In fact, efficient delivery is still a major hurdle in gene therapy. The finding that endogenous and grafted NSCs display an extensive migratory potential and tropism toward brain lesions founded the idea that these cells may be used as therapeutic vectors. Proof-of-principle experiments in animal models of lysosomal storage diseases (for a widely distributed brain lesion) and brain tumors (for a focal lesion) have shown that genetically modified NSCs are powerful therapeutics to cross-correct hereditary enzymatic deficiencies or to dramatically reduce a tumor mass. Thus, NSCs hold great promise for both cell and gene therapy.

Stem Cell-Based CNS Repair

It is manifest that stem cells can be used to replace neuronal, astrocytic, and oligodendroglial cells lost due to various brain diseases. However, it is important to note that successful use of stem cell is probably dependent on many factors, including the nature and degree of injury, disease history and age of the patient, primarily affected cell types, the type of stem cell chosen for transplantation, and the site of grafting. Deeper insight into these parameters will be important to tailor patient-specific treatment paradigms in a clinical context.

Stem cells have been explored in a number of rodent and primate models of human neurological diseases. These studies showed successful survival, differentiation, and synaptic integration of grafted stem cells, which led to functional restoration and amelioration of symptoms. In the following sections, we discuss examples of stem cell-based CNS repair and the challenges that remain to be addressed prior to clinical application.

Parkinson's Disease

Parkinson's disease (PD) is characterized by a progressive deterioration and loss of nigrostriatal dopaminergic neurons in the substantia nigra. The consequence of this cell death in the ventral midbrain is a deficient dopamine neurotransmission in the target region, the striatum. Clinically, PD presents with symptoms such as tremor, rigidity, and bradykinesia. Patients transplanted with fetal mesencephalic grafts in the early 1990s have demonstrated that an ectopic transplantation of dopamine-producing cells into the striatum can restore motor function and ameliorate clinical symptoms. Because of the limited availability of fetal tissue, stem cells are expected to provide unlimited numbers of transplantable dopamine neurons.

Several studies using rodent and primate models of PD have demonstrated successful integration and functional improvement after grafting of dopaminergic neurons derived from both ESCs and NSCs. In primate models, monkey ESCs have been transplanted and animals evaluated for behavioral improvements. As with rodent models, functional improvements occur. Furthermore, these behavior assessments can be corroborated with functional neuroimaging (Figure 1).

Although considerable progress has been made in stem cell-based treatment of PD inanimals, many challenges remain before clinical translation. Human ESCs differentiate to dopaminergic neurons under various protocols, yet the creation of a purified and homogeneous population of dopaminergic neurons is challenging and needs improvement. Animal models for future investigation should increasingly include primates in order to refine the mechanics and logistics of transplantation. Patients in whom stem cell therapy will be the most effective with the least side effects should be defined. The effects of posttransplantation training and rehabilitation need to be better understood; it appears that these contribute to improved functional outcome in experimental animal models. The clinical experience with fetal grafts suggests that the patient's disease history is an important parameter and that cell therapy will not be the method of choice for every Parkinsonian patient; thus, patient selection will be pivotal for clinical improvements after graft placement. Finally, the adverse side effects such as dyskinesias observed in some patients after transplantation of fetal grafts need careful consideration, and the safety of human ESCs needs to be established prior to clinical transplantation.

Stroke

Arterial occlusion within the brain can lead to ischemia and infarction of brain parenchyma. The current treatment of stroke remains limited and is focused on neuroprotection to limit the expansion of the infarct and to possibly recover the cells within the ischemic penumbra. The use of recombinant tissue plasminogen activator in select clinical situations within a critical time window after the stroke event has led to improved clinical outcome. Unfortunately, the time constraints in which this treatment can be offered limit potential application to a very small group of stroke patients.

Hypoxic/ischemic injury can lead to substantial tissue loss and the formation of infarction cavities, which would be a major therapeutic obstacle for the survival

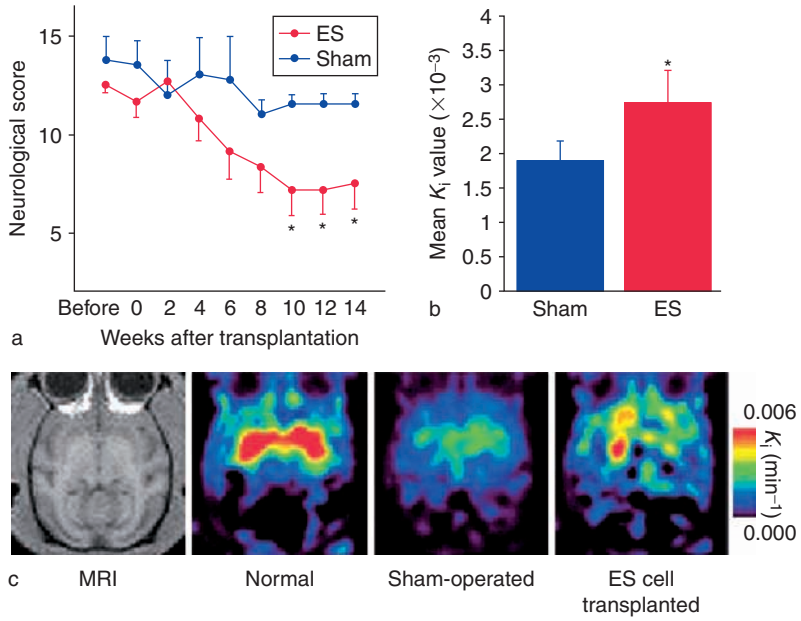

Figure 1 Function of embryonic stem (ES) cell-derived neurospheres in MPTP-treated monkeys. Behavioral scores (a) and PET study (b, c) of ES cell-transplanted ($n = 6$) and sham-operated animals ($n = 4$). (b) Mean K_i values from the entire putamen. (c) Increased ^{18}F-flourodopa uptake in the putamen of ES cell-transplanted animals. All values are mean ± SD.*$p < 0.05$. Reproduced from Takagi Y, Takahashi J, Saiki H, et al. (2005) Dopaminergic neurons generated from monkey embryonic stem cells function in a Parkinson primate model. *Journal of Clinical Investigation* 115: 102–109.

of newly seeded stem cells. Our group has demonstrated that extensively damaged brain areas can be repaired by the combined use of stem cells and polymer scaffolds that can be placed into the infarction cavity (Figure 2). It has been suggested that stem cells may be uniquely suited for stroke therapy given their inherent cytoprotective, anti-inflammatory, and restorative properties. Notably, early studies using other cellular therapies (e.g., human NT-2 teratocarcinoma line) demonstrated some functional improvement in stoke patients. However, the growth of stem cells is better controlled than any other immortalized cell line. Stem cell-derived and implanted neurons were shown to survive for more than 2 years in the human brain. Stem cell use in animal models of cerebral ischemia clearly demonstrates the ability of murine and human NSCs to engraft into the brain and survive, migrate, and differentiate leading to functional improvement. Other studies with murine NSCs have shown the potential of stem cell grafts to promote recovery in ischemic rats and recovery of sensorimotor deficits after transplantation into the striatum and cortex ipsilateral or contralateral to the stroke. In a clinical setting, cell transplants for stroke patients may be feasible even weeks after the ischemic event, allowing the patient to recover from the acute injury. Furthermore, several weeks may be needed to perform detailed neurophysiological and behavioral testing to allow selection of candidate patients. In accordance with a timetable that accounts for the most likely clinical scenario with patients, human somatic NSCs have led to functional recovery from stroke with improvement at both cortical and subcortical levels in various murine models of stroke. ESCs can be differentiated into NSCs following exposure to retinoic acid *in vitro* and have also demonstrated functional recovery in rodents.

Although regenerative cell therapy for stroke appears very promising, the use of stem cells is in its infancy. Clearly, the mechanisms that lead to beneficial effects after stem cell transplantation need to be better understood. For instance, it is important to determine whether the reported improvements are primarily the result of reconstitution of neural circuitry by cell replacement, the enhancement of intrinsic repair mechanisms (including the recruitment of endogenous stem cells), or both. Most likely, grafted and endogenous stem cells are effective through a multitude of mechanisms. It is possible that stem cells may be delivered to the injured brain not only by local intracerebral delivery but also by intravenous or intrathecal routes.

Spinal Cord Injury

Spinal cord injury (SCI) is a devastating ailment with little opportunity for treatment. The injury occurs from mechanical forces in the acute setting and is exacerbated by secondary inflammatory damage, both leading to neuronal death and demyelination. Accordingly, potential therapy would vary depending

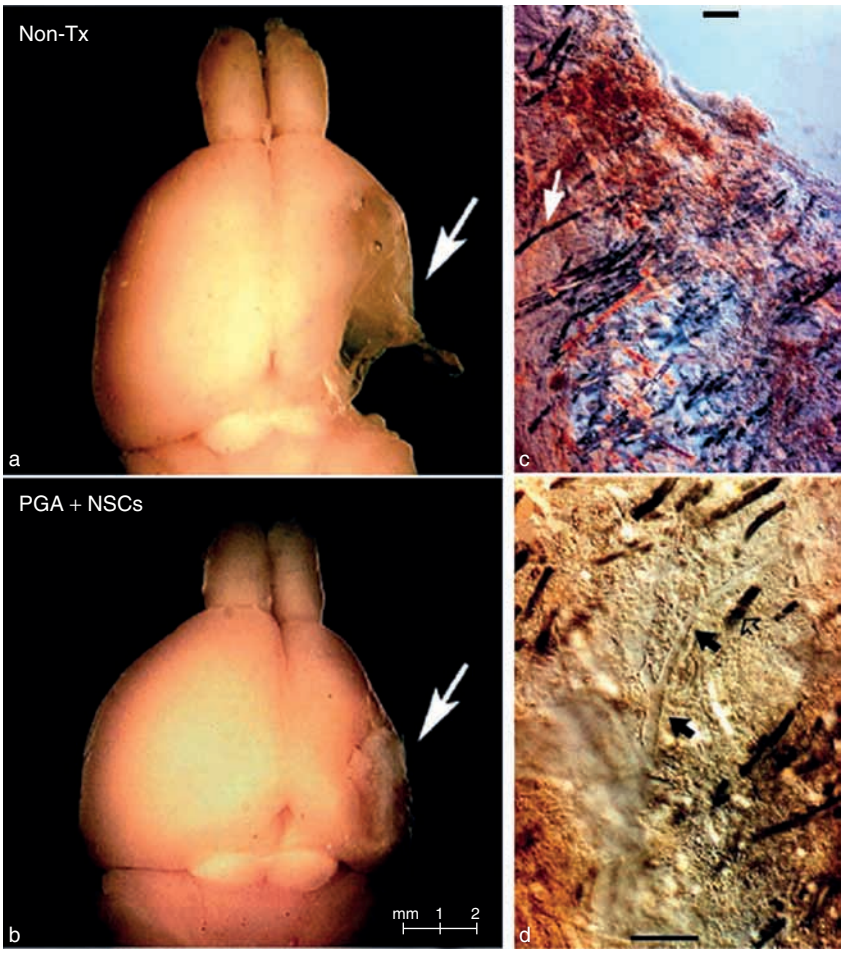

Figure 2 Implantation of neural stem cell–pluripotent embryonic stem cell (NSC–PGA) complexes into a region of cavity formation following extensive hypoxic/ischemic (HI) brain injury and necrosis. (a) Brain of an untransplanted (non-Tx) mouse subjected to right HI injury with extensive infarction and cavitation of the ipsilateral right cortex, striatum, thalamus, and hippocampus (arrow). (b) Contrasting with a, the brain of a similarly injured mouse implanted with an NSC–PGA complex (PGA + NSCs), generated *in vitro* as described in a, into the infarction cavity 7 days after the induction of HI (arrow; $n = 60$). At maturity (age matched to the animal shown in a), the NSC–scaffold complex appears in this whole mount to have filled the cavity (arrow) and become incorporated into the infarcted cerebrum. (c, d) Higher magnification of representative coronal sections through that region, in which parenchyma appears to have filled in spaces between the dissolving black polymer fibers (arrow in c) and even to support neovascularization by host tissues, as seen in d. A blood vessel is indicated by the black arrows in d; the open arrow in d points to degrading black polymer fiber. Scale bar = 100 μm (c, d). Reproduced from Park KI, Teng YD, and Snyder EY (2002) The injured brain interacts reciprocally with neural stem cells supported by scaffolds to reconstitute lost tissue. *Nature Biotechnology* 20: 1111–1117.

on the time frame after injury, with minimizing inflammation the primary concern early after injury and regeneration the major goal when injury is in its chronic phase. The pathobiology of SCI is highly dependent on the time course in which the injured spinal cord is examined. This directs the transplantation of NSCs or their derivatives (e.g., oligodendrocytes for myelination) to carefully account for and maximize the timing of transplantation with consideration of cell survival of grafted cells. Studies have addressed some of these issues demonstrating the importance of both timing of transplantation and the role of growth factors in murine models of SCI treated with NSCs. Injured rats received NSC transplants at different time points that would correspond clinically with subacute and chronic SCI, respectively. The administration of growth factors, including EGF, bFGF, and platelet-derived growth factor, resulted in increased numbers of cells grafted into the injured spinal cord by either enhanced survival or increased proliferation. To determine if NSCs together with growth factors can lead to neurological improvement, animals were evaluated using three independent behavioral tasks, all of which showed significant improvements compared to control mice, with even some long-term improvements (Figure 3). Human CNS fetal-derived stem cells have been shown to survive, engraft, differentiate, and improve

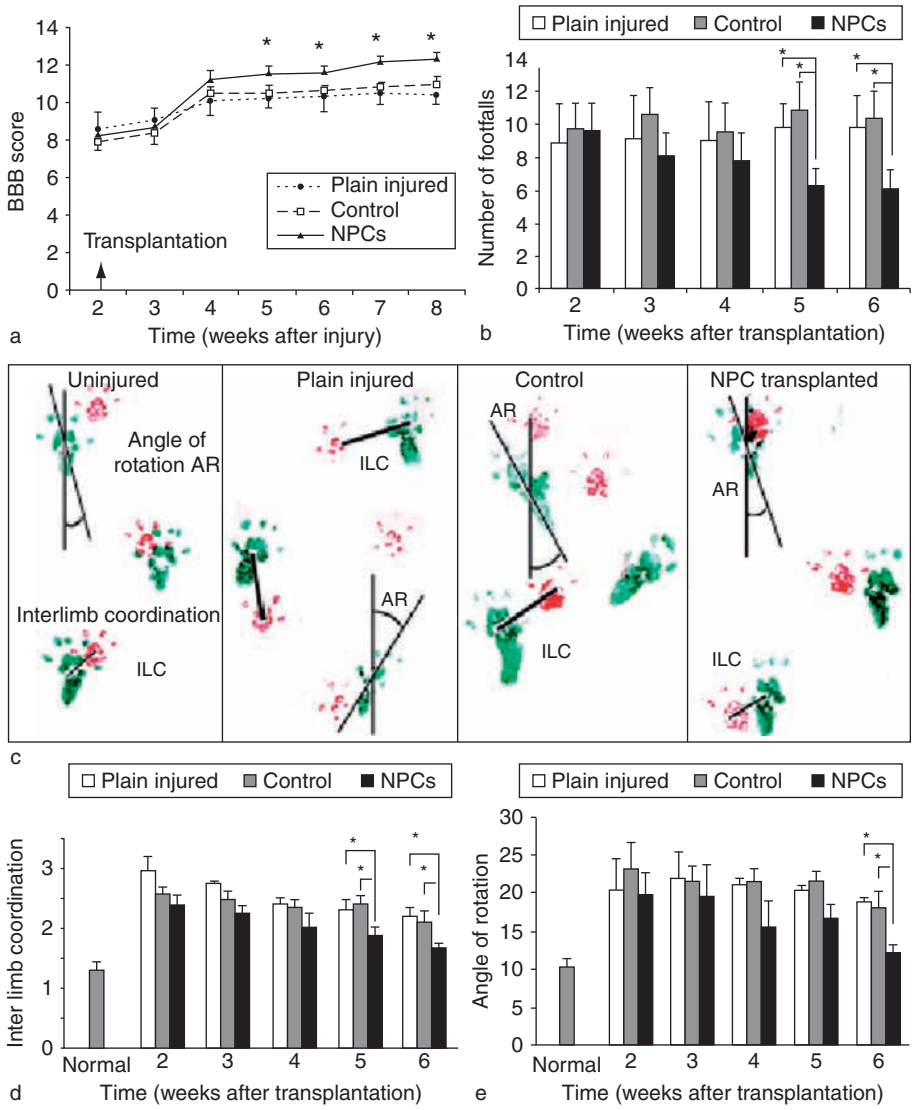

Figure 3 Subacute transplantation of YFP–NPCs resulted in a significant locomotor recovery compared with injured rats in the control group. (a) BBB rating scale showed a significant improvement in the locomotor BBB score in transplanted rats 3 weeks after transplantation compared with plain injured and control groups ($n = 5$ for plain injured group and $n = 8$ for other groups). (b) Using grid-walk analysis, transplanted rats also showed fewer errors in hindlimb placements 5 and 6 weeks after transplantation compared with plain injured and control groups ($n = 5$ for plain injured group and $n = 8$ for other groups). (c) Representative footprints of normal, plain injured, control, and grafted rats ($n = 5$ for plain injured group and $n = 8$ for other groups) show improvement in interlimb coordination as well as angle of rotation in the transplanted group compared with the plain injured and control groups. (d, e) Footprint analysis revealed that transplantation with adult NPCs significantly improved interlimb coordination and reduced the hindlimb angle of rotation 5 and 6 weeks after transplantation. The data show the mean ± SEM. $^*p < 0.05$. Reproduced from Karimi-Abdolrezaee S, Eftekharpour E, Wang J, et al. (2006) Delayed transplantation of adult neural precursor cells promotes remyelination and functional neurological recovery after spinal cord injury. *Journal of Neuroscience* 26: 3377–3389.

locomotor skills after traumatic SCI in mice. Contusive SCI of the thoracic cord was treated with injection of human NSCs. Functional recovery was assessed and shown to be improved. Also, selective ablation of grafted cells with diphtheria toxin (murine cells are 100 000 times less sensitive to diphtheria toxin than human cells) was used for the targeted killing of human NSCs. This selective ablation led to reversal of symptomatic and behavioral improvement, providing further support that human NSCs can mediate functional recovery in murine SCI models. Despite demonstrating functional recovery in murine models of SCI with both murine and human NSCs, further investigation using primate models is critical prior to any human trials. Indeed, spinal cord anatomy is different in the rodent compared to primates and humans.

Neural regeneration is not without pitfalls; some studies have shown the creation of aberrant axonal

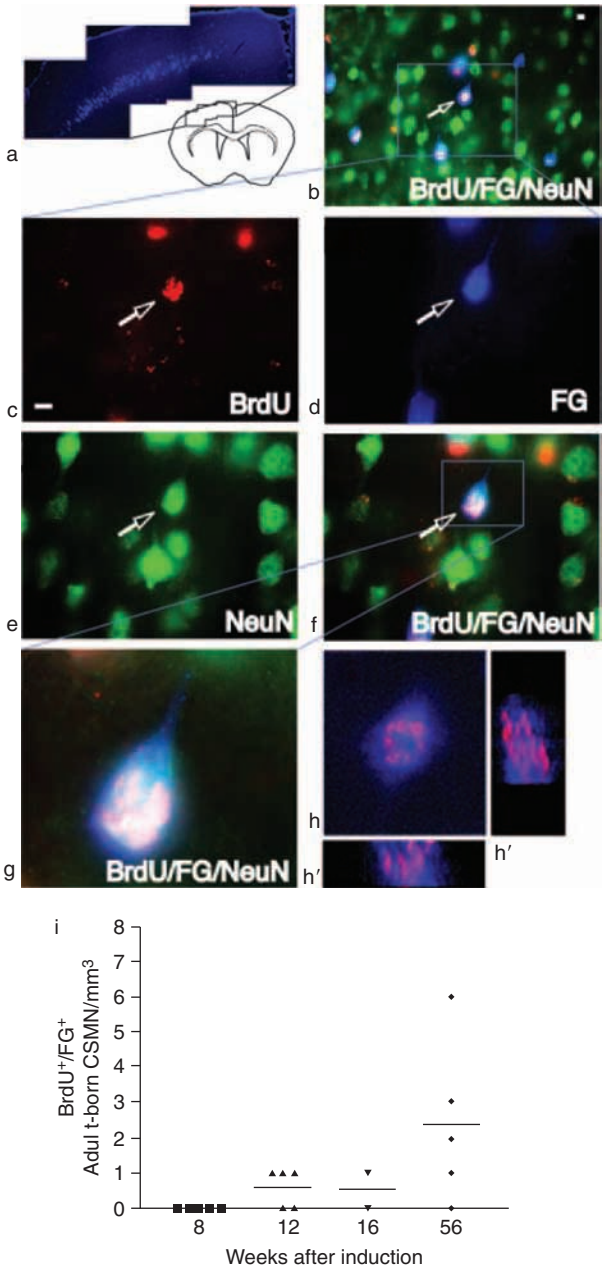

Figure 4 A subset of newborn layer V cortical neurons extends axons to the cervical spinal cord. (a) Both newborn and original layer V corticospinal motor neurons (CSMN) were retrogradely labeled by FG (blue). (b) Field expanded in (c)–(f) showing a BrdUrd^{+}/NeuN^{+}/FG^{+} triple-labeled adult-born neuron (arrow). (c–e) Individual images show that this BrdUrd^{+} nucleus (c, red) is located within this neuron, which is retrogradely labeled with FG from the cervical spinal cord (d, blue) and expresses NeuN (e, green). (f) Overlay showing BrdUrd^{+}/FG^{+}/NeuN^{+} neuron co-localization. (g) Higher magnification overlay of the same neuron from c–f. (h) A separate example of an adult-born neuron with a projection to the spinal cord. Laser-scanning confocal images were combined to produce three-dimensional reconstructions of newborn neurons. Viewing a BrdUrd^{+}/FG^{+} newborn neuron along its *x* (h′), *y* (h″), and *z* axes (h) unequivocally demonstrates the co-localization of BrdUrd and FG. (i) Quantification of BrdUrd^{+}/FG^{+} adult-born CSMN extending spinal projections from 12 to 56 weeks after induction of original CSMN apoptosis. Each point indicates the number of adult-born BrdUrd^{+}/FG^{+}neurons/mm^{3} in an individual animal; each bar indicates the mean Scale bar = 10 μm (b–e). Reproduced from Chen J, Magavi SS, Macklis JD, et al. (2004) Neurogenesis of corticospinal motor neurons extending spinal projection in adult mice. *Proceedings of the National Academy of Sciences of the United States of America* 46: 16357–16362.

sprouting which led to allodynia hypersensitivity. This may be related to excessive astroglial differentiation, highlighting the need for more controlled differentiation in order to both maximize functional recovery and minimize side effects. Along with increased understanding of cellular and molecular mechanisms, the timing and logistics of transplantation need to be improved, and noninvasive cellular imaging needs to be established.

Endogenous Stem Cell Recruitment for CNS Repair

The alternative or compliment to stem cell transplantation would be the manipulation of endogenous stem cells for therapeutic purposes. The advantages would include using the patient's own cells, not needing an invasive procedure, and obviating concern regarding the immunogenicity of transplanted cells. It appears that adult neurogenesis is restricted to the olfactory bulb and dentate gyrus of the hippocampus, but it is possible that some NSCs exist along the entire adult neuraxis.

Exploiting endogenous NSCs would require successful coordination of cell proliferation, differentiation, migration, and integration, and it does appear that some instructive signals remain in the adult CNS. However, most CNS regions are not permissive for neurogenesis under normal *in vivo* conditions, but some studies suggest that endogenous NSCs are primed to respond to environmental signals that exist primarily during pathological states. Accordingly, one approach to accentuate endogenous stem cell proliferation is the administration of growth factors. Intraventricular infusion of transforming growth factor-α into rodents with lesions of the substantia nigra dopaminergic neurons has led to functional improvements, putatively through recruitment of endogenous stem cells. Other likely bioactive molecules with potential to evoke a proliferative response in regions of the brain with multipotent cells are neurotrophins. Physiologically, neurotrophins are involved in cell cycle regulation, cell survival, and differentiation and are critical during normal development. Growth factor infusion can also promote proliferation of SVZ-derived progenitor cells, which gave rise to hippocampal CA1 pyramidal neurons in rodent stroke models, with

improvements in spatial orientation. Whether the neurogenic response creates neurons with long-term viability remains to be determined. Another candidate with efficacy in stroke models is erythropoietin, which has been shown to induce neurogenesis and functional improvement in rats. Furthermore, endogenous neural precursors can differentiate into new neurons that extend long-distance projections to the spinal cord in the adult rodent. Targeted apoptosis of corticospinal motor neurons was induced, and it was demonstrated that adult-born corticospinal motor neurons were generated extending from the motor cortex to the spinal cord (Figure 4).

Regarding neural repair, continued investigation into whether the diseased CNS can be treated with growth and differentiation factors to induce neural repair is required. As more is learned about the molecular signals and environmental cues, endogenous stem cells may prove to be a compliment or even replacement to transplantation of exogenous stem cells.

Other Stem Cells

The use of embryonic or somatic stem cells for brain repair is currently the focus of rigorous scientific investigation. Other stem cells have also been suggested as sources for cell therapy. For instance, some groups have found that mesenchymal stem cells (MSCs) can differentiate into astrocytes and neurons *in vitro* and *in vivo* and may have an advantage over ESCs or NSCs because they are a highly accessible source for the patient's own stem cells. However, there is controversy about the plasticity and developmental potential of MSCs. Some groups suggest that the findings with MSCs may be cell culture artifacts rather than true differentiation into unexpected cell types. Therefore, it is crucial to assay the differentiation of any stem cell into a particular cell type by combining morphological, immunophenotypic, and functional criteria. Currently, MSCs do not appear to be a realistic alternative to the use of ESCs or NSCs for neural repair.

Conclusion

Stem cell biology represents a strong foundation for neural repair. Experimental evidence suggests that this technology may be applicable to treat patients in the future. Since ESCs can be multiplied indefinitely and have the potential to give rise to a variety of functional human cells, it is conceivable that stem cells will play an important role in disease modeling and drug testing. Moreover, since stem cells mimic aspects of normal development, these cells may be used to study early steps of human development which would not be accessible for experimentation otherwise.

We have highlighted current problems in the rapidly progressing stem cell field, which involve safety issues, standardization of the protocols used, development of rigorous assays for characterization, and accumulation of experimental data in primate models of human disease. Realistic candidate diseases and patients who may benefit from stem cell therapy need to be defined before any clinical application. Since clinicians and stem cell biologists share a strong common interest to understand and treat human disease, stem cells have the true potential to transform modern medicine.

See also: Cell Replacement Therapy: Parkinson's Disease; Neural Stem Cells and CNS Diseases; Neural Stem Cells: Adult Neurogenesis; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Regeneration and Functional Recovery: Strategies; Spinal Cord Injuries; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Altman J and Das GD (1965) Autoradiographic and histological evidence of postnatal neurogenesis in rats. *Journal of Comparative Neurology* 124: 319–335.

Chen J, Magavi SS, Macklis JD, et al. (2004) Neurogenesis of corticospinal motor neurons extending spinal projection in adult mice. *Proceedings of the National Academy of Sciences of the United States of America* 46: 16357–16362.

Cummings BJ, Uchida N, Tamaki SJ, et al. (2005) Human neural stem cells differentiate and promote locomoter recovery in spinal cord-injured mice. *Proceedings of the National Academy of Sciences of the United States of America* 102: 14069–14074.

Falk A and Frisen J (2005) New neurons in old brains. *Annals of Medicine* 37: 480–486.

Gilbert S (2003) *Developmental Biology,* 7th edn. Sunderland, MA: Sinauer.

Karimi-Abdolrezaee S, Eftekharpour E, Wang J, et al. (2006) Delayed transplantation of adult neural precursor cells promotes remyelination and functional neurological recovery after spinal cord injury. *Journal of Neuroscience* 26: 3377–3389.

Lanza R, Blau H, Douglas M, et al. (2004) *Handbook of Stem Cells*. Oxford: Elsevier.

Lanza R, Gearhart J, and Hogan B (2006) *Essentials of Stem Cell Biology.* Oxford: Elsevier.

Lindvall OL, Kokaia Z, and Martinez-Serrano A (2004) Stem cell therapy for human disorders – How to make it work. *Nature Medicine* 10: S42–S50.

Nakatomi H, Kuriu T, Okabe T, et al. (2002) Regeneration of hippocampal pyramidal neurons after ischemic brain injury by recruitment of endogenous neural progenitors. *Cell* 110: 429–441.

Park KI, Teng YD, and Snyder EY (2002) The injured brain interacts reciprocally with neural stem cells supported by scaffolds to reconstitute lost tissue. *Nature Biotechnology* 20: 1111–1117.

Reyonals BA and Weiss S (1996) Clonal and population analyses demonstrate that an EGF-responsive mammalian embryonic CNS precursor is a stem cell. *Developmental Biology* 175: 1–13.

Singec I, Knoth R, Meyer R, et al. (2006) Defining the actual sensitivity and specificity of the neurosphere assay in stem cell biology. *Nature Methods* 3: 801–806.

Soderstrom K, O'Malley J, Steece-Collier K, and Kordower JH (2006) Neural repair strategies for Parkinson's disease: Insights from primate models. *Cell Transplantation* 15: 251–265.

Takagi Y, Takahashi J, Saiki H, et al. (2005) Dopaminergic neurons generated from monkey embryonic stem cells function in a Parkinson primate model. *Journal of Clinical Investigation* 115: 102–109.

Stereoscopic Vision

A J Parker, University of Oxford, Oxford, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Taken literally, stereoscopic vision describes the ability of the visual brain to register a sense of three-dimensional shape and form from visual inputs. In current usage, stereoscopic vision often refers uniquely to the sense of depth derived from the two eyes. This usage excludes a number of things that might be considered stereoscopic vision, such as the sense of depth arising from the motion parallax generated when subjects translate themselves through the visual environment. This article is primarily concerned with binocular stereoscopic vision.

Current usage also means that the term 'stereoscopic vision' tends to include a number of issues relevant to binocular vision that are unrelated to the perception of three dimensions. One example is singleness of vision: namely, the generation of perception of a single object by the two eyes by means of their co-coordinated use. Double vision is easily experienced for an individual object by raising a finger before the two eyes while continuing to look at a distant object, such as a church tower. Unusual circumstances, often a head injury or other neurological condition affecting the coordination of the two eyes, are required for all visual objects to look double.

Binocular Viewing Can Provide a Sense of Depth

When humans view a solid, three-dimensional object, each eye receives a slightly different view (see Figure 1) of the object. For the solid sphere (t) in Figure 1, the left eye (m) sees rather more of the left-hand side of the sphere before the surface of the sphere is occluded by its rim. Conversely, the right eye (n) sees a little more of the right-hand side of the sphere. Ptolemy, Leonardo da Vinci, and many in between understood such matters clearly. However, the ability of binocular geometry to provide a sense of depth was fundamentally misunderstood until the striking demonstrations due to Charles Wheatstone (1802–75; see Figure 2).

Wheatstone arranged for each eye to be given a different view in the most direct way, by showing each eye a physically different picture to look at. When the pictures viewed by the left and right eyes approximate the views that the eyes would receive when viewing a real object in natural binocular viewing, the visual brain responds to those views as if a real object in depth were being displayed.

Wheatstone's discovery generated intense scientific debate throughout the rest of the nineteenth century. For example, the need for real, physical movements of the eyes to gain a sense of depth was hotly contested until the critical experiment of Dove, who demonstrated that a sense of stereoscopic depth is produced by even the briefest flashed exposure that leaves the eyes no time to move. Integration of binocular stereo depth with other sources of information about three dimensions was intensively researched; for example, Helmholtz appreciated that there are strong similarities between depth from binocular stereopsis and depth from motion parallax but, critically, that one of these depth cues can supply information that is absent from the other. The link between binocular stereopsis and binocular eye movements was investigated, including the important links with changes in the focusing power of the eyes' lenses, known as accommodation. All this work was reviewed by Helmholtz and his acolytes toward the end of the nineteenth century.

Wheatstone's discovery also generated huge popular interest. Coupled with the development of photography, the stereoscope became a source of home entertainment. Many Victorian parlors housed one, and large collections of stereo pictures were generated. It is alleged that the stereoscope was put aside only when its persistent use to view salacious images became so widely known that its presence in a living room became an immediate source of embarrassment. (The same images are often regarded nowadays as signifying the innocence of a bygone age, and stereoscopic photography now has its place among nostalgia technology.)

Such a combination of high scientific and popular interest could not fail to challenge egos, and the discovery of the stereoscope was disputed hotly. There is no doubt that David Brewster's version that used lenses for viewing was much more convenient for popular viewing and thus became the canonical form, but Brewster's claims to priority were always regarded as misplaced.

Geometry of Binocular Viewing

The viewing geometry that Wheatstone had identified was subject to intensive analysis during the nineteenth and early twentieth centuries. This work

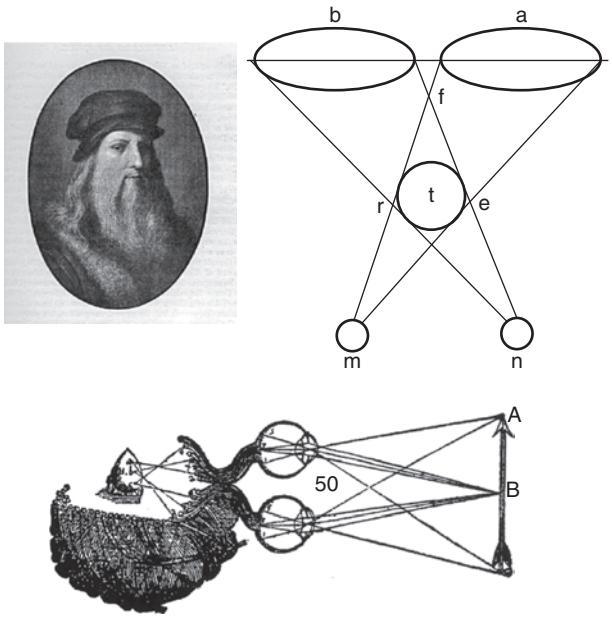

Figure 1 Leonardo da Vinci (top left) understood that each eye obtains a slightly different view of the world. For objects close to the head, some features on the surface of a solid object are occluded from one eye's view but are visible to the other eye. The diagram at the top right is derived from Leonardo's own drawing, adapted by Wade NJ, Ono H, and Lillakas L (2001) Leonardo da Vinci's struggles with representations of reality. *Leonardo* 34: 3, with permission from MIT Press Journals. The lower panel shows a figure from Descartes's *De Homine*, in which the schematic of connecting the left and right eyes to a single binocular target in the central nervous system is proposed.

established the concept of 'binocular disparity' as the fundamental measurement employed by the visual system to gain a sense of binocular stereoscopic depth. Binocular disparity is the angular difference between two landmarks viewed binocularly (see Figure 3).

Corresponding Points

This phase of study also introduced the concept of 'corresponding visual points' between the left and right eyes. For example, points that are 3 mm horizontal to the left of the fovea in the left and right eyes are corresponding points, whereas points that are 3 mm to the left in one eye and 2.9 mm to the left in the other eye are noncorresponding. If both eyes are looking at a single object in the world, then corresponding points on the retina project out to a set of geometrically defined locations in three-dimensional space. For the case of a plane, defined by the line joining the two eyes and the binocular fixation point, the locations of corresponding points defined a circle, the 'Vieth-Muller circle,' passing through the binocular fixation point and the optical centers of the two eyes.

Much of this development was directed toward understanding the perception of three-dimensional space by human observers. Imagine a person is looking at an array of objects lying in different directions and at different distances with respect to the binocular fixation point. If the person closes first one eye and then the other, given that each eye is at a different position, do these objects appear to lie in the same direction? The geometry of the Vieth-Muller circle provides a simple hypothesis, which is that objects will appear to lie in the same visual direction when they lie on anatomically corresponding points. This hypothesis is close to correct, but there are distinct patterns of deviation from the Vieth-Muller circle, and formulations of other curves and surfaces have been developed to describe the locations of points that are in visual, as opposed to anatomical, correspondence.

The concept of corresponding points is also important in considering the binocular coordination of the movements of the eyes. If the eyes are fixating an object at a particular distance from the observer's head, then the eyes will need to rotate inward with respect to each other if the object is near the head. Inspection of a new object at a different viewing distance therefore requires an adjustment, inward or outward, of the binocular alignment, if the new object is to fall on the foveae of both eyes. The size and direction of the adjustment is given by the binocular disparity between the currently viewed object and the new one.

Absolute and Relative Disparity

One way of defining an origin for measurement of disparity is therefore to use the current binocular fixation point. Anatomically corresponding points have a disparity of zero, and noncorresponding points may have a positive or negative disparity, depending on whether they project to points farther than or nearer than the distance to the fixation point. Since this definition can be employed even when the fixation point is just notional (a location in space with no visible object, toward which the foveae project), it is useful to give this concept of disparity the label 'absolute disparity.'

This definition of disparity may be contrasted with the concept of 'relative disparity,' which is defined as the disparity between two visible features. Here these two features are separated by one visual angle as seen from the left eye and a different visual angle as seen from the right eye: relative disparity is therefore the difference in these two angles. Note that, as the eyes move, the visual angle between the two features is not altered by rotations of the eye. Thus the relative disparity between two visible features is independent

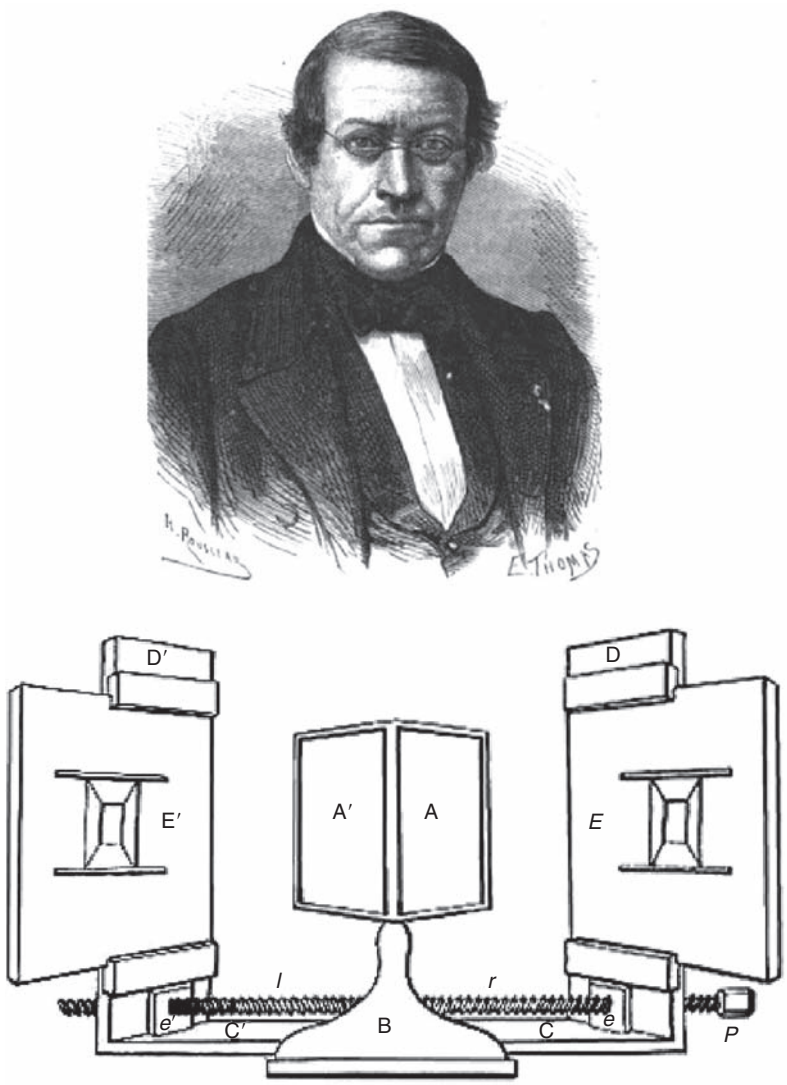

Figure 2 Charles Wheatstone and his mirror stereoscope. The drawing of the stereoscope shows drawings (E and E′) reflected by mirrors (A and A′) set at an angle so as to direct the reflected images toward a viewer. Wheatstone's observations with this device led him to conclude that binocular disparity was a sufficient stimulus by itself to give rise to a sensation of stereoscopic depth.

of the choice of binocular fixation point, which means that relative disparity is a useful description of the stimulus that is purely visual, not reliant on eye position.

In the psychological literature, the term 'disparity' is often taken to mean relative disparity, but in optometry, physiology, and ophthalmology, disparity is more likely to refer to absolute disparity.

First Steps to Neuronal Mechanisms

A major step forward in the understanding of binocular stereoscopic vision came in the 1960s. Two advances were especially significant. First, Bela Julesz, working at Bell Laboratories, invented the random dot stereogram (RDS; see Figure 4) and exploited this new paradigm to define and explore what he termed 'cyclopean vision.' Cyclopean refers to the mythical giant Cyclops, who had just one eye in the center of his forehead. Julesz's RDS consisted of fields of randomly bright and dark pixels, which were correlated between the left and right eyes' images. When placed in a stereoscope, a region of this figure was revealed as segregated from the background because it was at a different binocular depth (see Figure 4). Figures demonstrating stereoscopic segregation had been prepared before, but Julesz's development of a computer-based method was truly original and profoundly influential. Julesz argued that the perception of the revealed figure reflects central

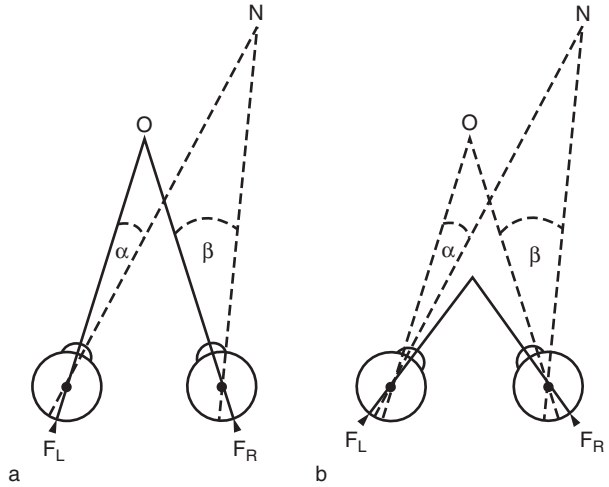

Figure 3 Diagram showing the definition of absolute and relative disparity. (a) The eyes are looking at point O. The angles α and β between the line of fixation to O and the projection of point N are different as viewed with the left (L) and right (R) eyes. N projects to a point in the right eye that is farther from the fovea (F) than is the point from the fovea in the left eye. N therefore has a nonzero absolute disparity whereas O has a zero absolute disparity. (b) The eyes have shifted to a new fixation point. Now, both O and N have nonzero absolute disparities. However, the relationship between O and N has not changed: Angles α and β still describe the angular separation of O and N. In both (a) and (b), the relative disparity between N and O is given by the quantity (β−α).

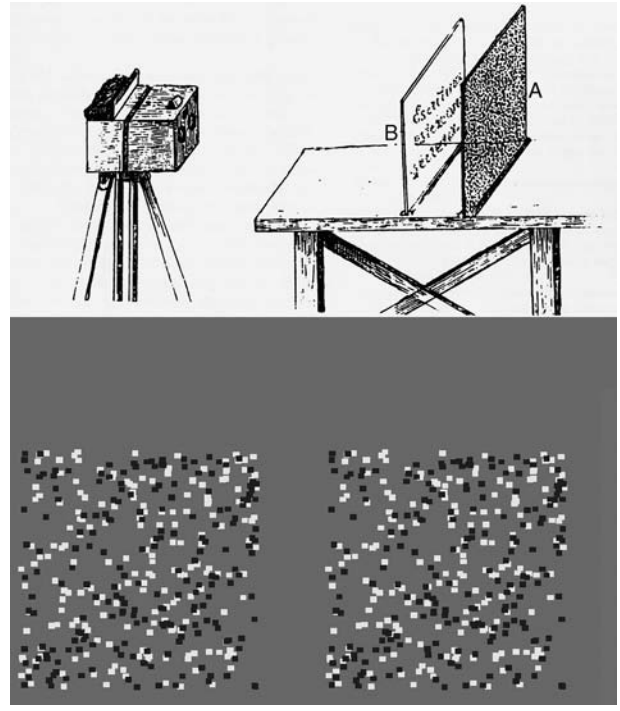

Figure 4 A random dot stereogram (RDS). When the left and right eye's images are viewed in a stereoscope, a region of different depth in the center emerges. The principle was apparently first published by Cajal, who devised it as a means of secret writing. The upper panel shows a stereoscopic camera taking the image of written text on a plate B at a different distance from the camouflage background A. If either the left or right photographic plate is viewed alone, the message cannot be decoded. Presumably, the intent was to send each image via a different route. The lower panel shows a modern RDS of the type invented by Julesz. When the left and right eye's images are fused binocularly, a circle segregates in depth from the background square.

processing by the brain, separate from the influence of either eye alone. He conducted a series of studies that investigated the classical visual illusions presented in their novel cyclopean form.

The other advance was the first identification of single neurons that are sensitive to the disparity of visual stimuli. In the late 1950s, Hubel and Wiesel had discovered that some neurons in the visual cortex often respond selectively to the orientation of visual contours and are sensitive to stimuli over a limited region of visual space (the receptive field) in both the left and right eyes. These neurons showed 'binocular summation': the response to the simultaneous stimulation of left and right eyes was stronger than the response to either eye alone. Two research groups, Nikara, Bishop, and Pettigrew, working in Australia, and Barlow, Blakemore, and Pettigrew, in the United States, presented new findings that some number of these visual neurons are specifically most sensitive when the independent stimulation of the left and right eyes falls on noncorresponding, as opposed to corresponding, points. Such neurons therefore give their greatest response to a nonzero disparity, and as the disparity is adjusted away from this optimal value, the response of the neuron declines. A neuron that responds optimally to stimulation of noncorresponding points is sensitive to the absolute disparity of the stimulus. This work established that there were neurons sensitive to different values within range of nonzero disparities, implying a population of neurons that is capable of encoding a range of binocular depths with respect to the binocular fixation point (see Figure 5). Another consequence was to give an unambiguous physiological interpretation to the concept of binocular corresponding points developed in earlier work.

Computational Stereo

These two developments set the scene for most modern experimental studies of stereoscopic vision. Julesz also identified a different conceptual aim; he sought to specify an automatic, computational method for processing a stereoscopic pair of images to extract the binocular depth. Marr developed an exceptionally broad view of the computational problem of early visual processing, which took in image processing, visual geometry, retinal and cortical physiology, artificial intelligence, and much more. Nonetheless, it

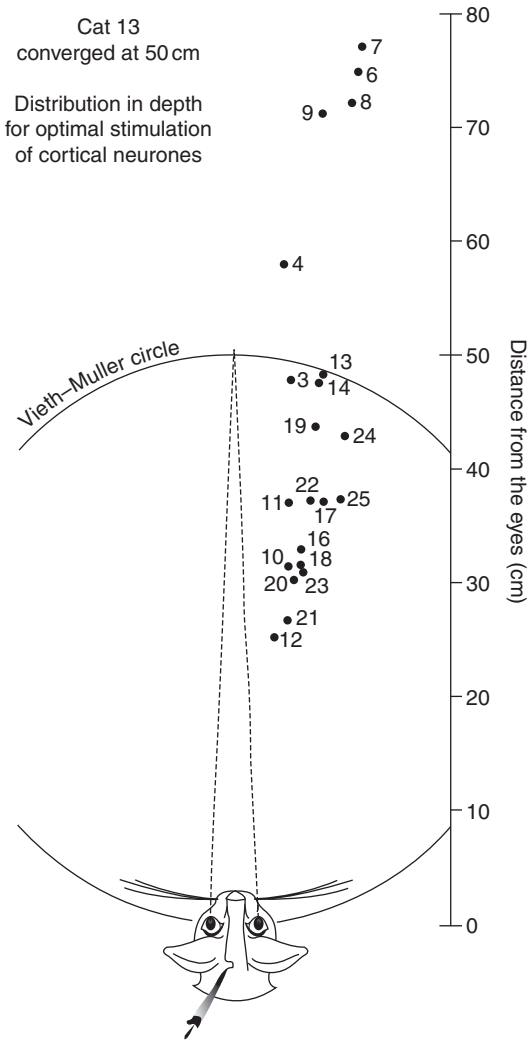

Figure 5 A range of neurons in cat visual cortex that have their greatest response at different binocular depths. The numbered dots represent visual stimuli. The dottled lines show the lines of right to the cat's fixation point. The Vieth–Muller circle is the circle passing through the fixation point and the optic centers of the cat's eyes. From Barlow HB, Blakemore C, and Pettigrew JD (1967) The neural mechanism of binocular depth discrimination. *Journal of Physiology* (*London*) 193(2): 327–342, with permission from Wiley-Blackwell Publishing Ltd.

is significant that Marr, working with T Poggio, took the problem of stereoscopic vision as a vehicle for carrying forward his more general program. A particularly exciting aim was that of mapping components of the computation onto the actions of individual nerve cells.

Development of Stereo Vision

An important element of Hubel and Wiesel's neurophysiological studies revealed how the visual system develops during early life. They argued that there is a period of high plasticity during early development. During this period, the neural apparatus must receive appropriate visual stimulation; otherwise the neurons fail to connect appropriately to one another. Much of the investigation of this so-called critical period was based on the study of binocularity in the visual cortex. Interventions such as depriving one eye of vision result in that eye's having a weaker, or even absent, influence on primary visual cortical (V1) neurons. The importance of concurrent input to both eyes was highlighted by the finding that alternating occlusion of each eye resulted in a visual cortex in which there were almost no binocular neurons but large numbers of neurons responsive exclusively to stimulation of either the left eye or the right eye. A similar pattern of cortical organization was found when a squint was induced by surgical intervention that resulted in the misalignment of the left and right eyes. Loss of binocularity is a central element of the human clinical syndrome of strabismus and associated amblyopia, in which the disruption of vision in childhood may lead in effect to blindness in one eye due to that eye's functional disconnection from the visual cortex. Hubel and Wiesel emphasized the importance of visual stimulation in selectively retaining cortical connections that had already been set up by genetic instruction, while others, notably Blakemore, highlighted the possibility that environmental influences could induce the formation of new connections.

Current Progress in Understanding Binocular Stereopsis

The Binocular Energy Model

At the physiological level, an important modern development was to understand the way in which the initial stages of disparity detection proceed. Ohzawa and Freeman, later joined by DeAngelis, found that many disparity-sensitive neurons in the V1 behave as if the initial stage of binocular combination is a simple addition of the excitation arising from each eye. The initial work on this problem used sinusoidal grating stimuli, in which the relative phases of the sinusoids in the left and right eye were adjusted parametrically to measure the binocular combination rule used by the neurons. A number of observations led them to propose that binocular disparity might be detected by means of either a mechanism that is specifically sensitive to interocular phase or (as originally implied) a mechanism that encodes a positional shift between the two eyes' receptive fields. Overall, their experimental findings pointed to a two-stage model of disparity detection. Disparity detection arises by thresholding and squaring the output of an initial detector and then combining

signals from multiple initial detectors, either phase- or position-sensitive; the combination of signals encompasses detectors with the same disparity sensitivity over a small range of locations in the visual field. This model was termed the 'energy model' since the operation of squaring that is embedded within it parallels the definition of – and method for calculating – the energy of an electrical signal. This model has been very successful in capturing a wide variety of findings about the functional properties of disparity-selective neurons in V1. Although a number of adjustments and elaborations to the model have been proposed, the substantial elements of the structure have remained.

Other Species

All these studies focused on the properties of neurons in the V1, which is the earliest stage at which disparity-sensitive neurons are found. During the 1970s and 1980s, the neural basis of stereopsis was extended considerably. First, other species were examined; for example, the owl's visual wulst also contains neurons selective for binocular disparity. Notwithstanding the fact that the anatomical projections from retina to brain are organized differently in owls and mammals, the forward-facing eyes of the owl are associated with the capacity to detect stereoscopic depth, and there are strong similarities in the neural apparatus involved in all species. Second, many different visual cortical areas were explored, mostly in the macaque monkey. The multiplicity of visual areas, initially identified by Zeki, proved to have a variety of specializations for processing different aspects of the visual image, such as color and motion. For stereoscopic vision, the work of G Poggio established unambiguously the presence of disparity selectivity in V1, V2, and V3, while the work of Hubel and Livingstone suggested that binocular neurons (by implication responsible for the perception of stereo depth) were especially concentrated in compartments of V1 and V2, identifiable on the basis of histological staining for the mitochondrial enzyme cytochrome oxidase; these were the regions between cytochrome oxidase blobs in V1 and the thick stripes in V2.

From Disparity Detection to Depth Perception

Although neurons both inside and outside the V1 were explored in detail, most experiments tended to carry over the paradigms established for V1 into other cortical areas. This ultimately proved to be a limitation. The reason was that disparity selectivity at the physiological level was dominated by the conceptual framework set up by the earliest investigations, which grounded their framework for understanding the concept of retinal correspondence described earlier. Work in human psychophysics, particularly due to Westheimer, pointed out that relative disparity was a much more relevant parameter for describing the outcome of perceptual experiments. This highlighted a gap in the conceptual framework for understanding binocular stereoscopic vision at the neuronal level. Neuronal studies were working with a framework linked to retinal correspondence and hence to absolute disparity (defined earlier), whereas perceptual studies pointed to relative disparity. A number of other perceptual properties of stereoscopic vision are at variance with the simple detection mechanisms implied by studies in V1, even those motivated by computational principles implemented in the energy model.

Cumming and Parker therefore initiated a program to compare the responses of single neurons at various levels in the visual system of nonhuman primates against the perceptual response to binocular stereo depth. The focus was to identify the extent to which neuronal responses in particular brain regions could explain the perceptual characteristics of binocular stereoscopic vision. This work demonstrated, for example, that neurons in V1 did not signal relative disparity whereas at least some neurons in cortical area V2 did indeed respond to relative disparity. Evidence from a number of laboratories has continued to develop the picture that stereoscopic depth processing is elaborated within the extrastriate visual areas beyond the early disparity-detection mechanisms in V1. This concept is also consistent with the conclusions from more recent developmental studies, in which visual processing has been disrupted experimentally or owing to human clinical conditions. These studies have indicated that some of the deficits in visual performance arise from disruptions of neuronal processing in extrastriate visual areas, which bring additional losses in performance over and above those attributable to the effects observed in V1.

Cortical area V5/MT, mostly studied in relation to the processing of visual motion, has proved to have an important role in binocular stereoscopic vision. In particular, DeAngelis, working initially in Newsome's laboratory, has revealed a number of neuronal signals identifiable in V5/MT that are specific to the performance of certain stereo tasks. These signals are separate from signals that are generated simply by the presence of certain binocular disparities within the stimulus. Thus, V5/MT carries not just simple sensory signals but also task-related signals for certain perceptual judgments of stereo depth. V5/MT lies in the dorsal stream of extrastriate visual pathways. Meanwhile, in the ventral stream, Janssen, working in Orban's group, has shown that brain area Tes, in the inferotemporal cortex,

carries signals about the three-dimensional shape of surfaces, specified by binocular disparity.

Current thinking suggests a broad division of functional involvement of dorsal and ventral extrastriate areas in different aspects of binocular stereoscopic vision. The ventral stream appears to be mainly concerned with exploiting stereoscopic depth for shape perception. The dorsal stream carries signals for controlling binocular eye movements, for orienting the observer in the visual environment, and for segregating different depth planes so that visuomotor tracking behavior can be sustained. The dorsal stream appears to make use of a simple correlation-based stereo algorithm, broadly similar to the energy model in V1, while the ventral stream may use a more sophisticated stereo algorithm that sustains detailed point-to-point matching of binocular features between the eyes.

See also: Representation of Reward; Shape Representation in Inferotemporal Cortex; Vision: Surface Segmentation; Vision: Light and Dark Adaptation.

Further Reading

Barlow HB, Blakemore C, and Pettigrew JD (1967) The neural mechanism of binocular depth discrimination. *Journal of Physiology (London)* 193(2): 327–342.

Bishop PO and Henry GH (1971) Spatial vision. *Annual Reviews of Psychology* 22: 119–160.

Bradley DC, Qian N, and Andersen RA (1995) Integration of motion and stereopsis in middle temporal cortical area of macaques. *Nature* 373(6515): 609–611.

Cumming BG and DeAngelis GC (2001) The physiology of stereopsis. *Annual Review of Neuroscience* 24: 203–238.

DeAngelis GC, Cumming BG, and Newsome WT (1998) Cortical area MT and the perception of stereoscopic depth. *Nature* 394 (6694): 677–680.

Helmholtz HV (1962) *Handbook of Physiological Optics*. New York: Dover (originally published 1909).

Janssen P, Vogels R, and Orban GA (2000) Three-dimensional shape coding in inferior temporal cortex. *Neuron* 27: 385–397.

Julesz B (1971) *Foundations of Cyclopean Perception*. Chicago: University of Chicago Press.

Marr D and Poggio T (1979) A computational theory of human stereo vision. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 204(1156): 301–328.

Ohzawa I, DeAngelis GC, and Freeman RD (1990) Stereoscopic depth discrimination in the visual-cortex: Neurons ideally suited as disparity detectors. *Science* 249(4972): 1037–1041.

Orban GA, Janssen P, and Vogels R (2006) Extracting 3D structure from disparity. *Trends in Neurosciences* 29(8): 466–473.

Parker AJ (2004) From binocular disparity to the perception of stereoscopic depth. In: Chalupa L and Werner JS (eds.) *The Visual Neurosciences*, vol. 1, pp. 779–792. Cambridge, MA: MIT Press.

Parker AJ (2007) Binocular vision and the cerebral cortex. *Nature Reviews Neuroscience* 8: 379–391.

Wade NJ, Ono H, and Liuakas (2001) Leonardo da Vinci's struggles with representations of reality. *Leonardo* 34(3): 231–235.

Wheatstone C (1838) Contributions to the physiology of vision: I. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London* 128: 371–394.

Stimulant and Wake-Promoting Substances

C Ballas and D F Dinges, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Modafinil

Modafinil is a nonamphetamine, wakefulness-promoting drug that has Food and Drug Administration (FDA) approval for the treatment of excessive sleepiness in the following disorders: narcolepsy, residual sleepiness in continuous positive airway pressure-treated obstructive sleep apnea, and shift work sleep disorder. Modafinil has been extensively studied for its effects in these disorders, as well as in sleep-deprived healthy adults. In the latter studies, modafinil was found to reduce physiological sleepiness and improve a wide range of cognitive deficits from sleep deprivation that varied from 36 to 88 h awake. Compared to placebo, modafinil reduced sleep-deprivation deficits on the following cognitive and behavioral tasks: reaction times, vigilant attention, alertness, short-term memory, mathematical calculation speed, logical reasoning, and simulated performance.

Modafinil is designated a controlled substance, Schedule IV, which means it is considered to have less potential for abuse than amphetamines and other schedule I–III stimulants. Although not FDA-approved for attention deficit hyperactivity disorder (ADHD), two clinical trials showed that modafinil resulted in significant improvement in ADHD symptoms, especially attention and concentration outcomes. However, it was associated with Stevens–Johnson syndrome in one child, and approval was rejected pending further investigation.

Modafinil Metabolism and Elimination

Modafinil is the metabolite of adrafinil, lacking adrafinil's terminal amide hydroxyl group. Modafinil exists in racemic form; the I-R-modafinil enantiomer is called armodafinil, and has separate FDA approval. The R enantiomer has a 10- to 14-h half-life, compared to 3 or 4 h for the S enantiomer. Modafinil has a protein binding of 60% and volume of distribution (V_d) of 0.8 l kg^{-1}. The half-life is approximately 12 h. Modafinil is converted primarily to the inactive modafinil acid and modafinil sulfone via CYP3A4/5. The *d*-form is cleared three times faster than the l-form.

Modafinil Pharmacology

The precise mechanism of action of modafinil is not well understood. It has weak affinity for the dopamine transporter and almost negligible affinity for the norepinephrine transporter. However, in the absence of functional dopamine transporters (e.g., DAT knockout mice), modafinil fails to exert any wakefulness-promoting actions. Medications that inhibit dopaminergic activity, including D1 and D2 blockers and tyrosine hydroxylase inhibitors, fail to inhibit modafinil.

Modafinil inhibits striatal and globus pallidus GABA activity. At doses equivalent to 50 times normal therapeutic doses, it can also enhance glutamate in the striatum. The medial preoptic area and posterior hypothalamus are primarily controlled by tonic GABA inhibition, and consequently it may be that modafinil's effect on glutamate could be blocked with $GABA_A$ antagonists. Modafinil was found at high doses to increase serotonin in the medial preoptic area and posterior hypothalamus, and it was found at lower doses in the cortex and amygdale. Dorsal raphe nuclei receive projections from orexin–hypocretin neurons in the lateral hypothalamus, and they show increases in serotonin with modafinil administration. At therapeutic doses modafinil has no direct effect on serotonin, but it has been found to synergistically augment serotonin increases to fluoxetine and imipramine. Modafinil appears to have minimal effects on the endocrine system, and it has not been associated with significant changes in circadian and sleep modulated hormones such as melatonin, growth hormone, and cortisol.

Loss of orexin–hypocretin cells is associated with narcolepsy, and orexin knockout animals exhibit narcoleptic phenotypes. Because of modafinil's demonstrated efficacy for treating excessive sleepiness in narcolepsy, it was thought that modafinil might exert its effect on orexin–hypocretin cells, which are located only in the lateral hypothalamus but have projections to a wide variety of brain areas. Modafinil was found to increase c-fos immunoreactivity in both orexin–hypocretin neurons and tuberomammilary neurons while having minimal effect in the suprachiasmatic nucleus and anterior hypothalamus. However, modafinil still retains its activity on wakefulness even in orexin–hypocretin receptor knockout animals. In fact, some researchers have found that modafinil has stronger wakefulness-promoting activity in such knockout animals. However, its putative effects via orexin–hypocretin neurons, which are also important in feeding behaviors, have not been associated with effects on food intake. Modafinil appears to also inhibit neurons in the ventrolateral preoptic (VLPO) area, which is the putative sleep switch. This enhances histaminergic activity (i.e., decreases VLPO

inhibition of histaminergic activity), which further promotes wakefulness.

Modafinil Adverse Reactions and Safety

Modafinil is generally well tolerated, with the most common reported side effects including headache and insomnia. In phase III trials, three patients with left ventricular hypertrophy and mitral valve prolapse experienced transient ischemic changes on electrocardiogram. As noted previously, one child in a controlled trial had a rash suggestive of Stevens–Johnson syndrome. Modafinil is otherwise relatively safe in overdose, and even chronic usage of high doses has not been found to be associated with significant adverse effects, even after abrupt discontinuation.

Amphetamines

The US Drug Enforcement Agency has listed amphetamines as Schedule II drugs under the Controlled Substances Act because of their high potential for abuse. They are approved by the FDA for the treatment of narcolepsy and ADHD, and several double-blind, placebo-controlled trials in adults and children have found efficacy for this purpose.

Amphetamines have been studied and used for conditions not formally approved by the FDA. These include cognitive dysfunctions associated with brain injury and stroke, and fatigue associated with multiple sclerosis and HIV. Amphetamines have also been studied as augmentation agents in the treatment of depression; however, their efficacy appears to be only moderate, and their effects are concentrated in the expected areas of fatigue and apathy.

Studies on healthy adults have found that amphetamines can improve performance on a variety of tests during sleep deprivation, including visual and auditory reaction times and short-term memory. Subjects in almost all trials have reported subjective feelings of improvements in attention, memory, concentration, and fatigue – even when these were not found on objective outcomes.

Amphetamines have long been known to improve wakefulness and decrease sleep. Whether taken 15 h before sleep or immediately before sleep, they decrease sleep duration. These effects may last into the second night, with further (although lesser) decreases in sleep duration. Amphetamines also affect sleep architecture by decreasing the duration of all sleep stages, and they increase the percentage of light stage 1 sleep. These effects may last into the second night after amphetamine dosing.

Amphetamine Structure–Activity Relationships

In general, amphetamines and their derivatives have three parts that influence their activity: the phenyl, isopropyl, and amine groups. Addition of a CH3 group to the amine carbon creates amphetamines and other chemicals with stimulant properties (e.g., bupropion, ephedrine, cathinone, and fenfluramine). Addition of methoxy groups to the phenyl groups creates hallucinogens (e.g., mescaline and 3,4-methylenedioxy-*N*-methylamphetamine (MDMA)). For example, 'Ecstasy' (MDMA) is built on a methamphetamine backbone with an aromatic substituted dimethoxy ring. The absence of the methyl group on terminal amine (i.e., 3,4-methylenedioxy-*N*-amphetamine (MDA), 'Love') further increases the hallucinogenic potency and duration. Collectively, these stimulant/hallucinogen hybrids are referred to as entactogens. Key differences in the activity of amphetamine versus methamphetamine are described later.

The most important structural alterations to the clinically relevant amphetamines consist of the creation of a secondary substituted amine and changes in stereospecificity. The addition of a methyl group to the amine creates levo (L-amphetamine) or dextro (D-amphetamine), which are considerably more potent than amphetamine. However, the L- and D-forms differ in their relative merits and strengths. The dextro form is twice as potent as the levo form in keeping one awake, but both have equivalent effects on rapid eye movement sleep and cataplexy. They do not differ in MAO-A or -B activity, or with regard to norepinephrine efflux. However, the dextro form is markedly more potent in dopamine efflux and reuptake. Importantly, rat experiments have shown that the levo form is considerably less potent (four times less potent) than the dextro form in reinforcement paradigms.

Amphetamine Metabolism and Elimination

Lipid solubility is high for amphetamines and more so for methamphetamine. This allows for rapid absorption and peak levels within 2 h. Its V_d is $5\,l\,kg^{-1}$, with an elimination half-life that varies greatly but is on the order of 20 h. The dextro form is more rapidly eliminated than the levo form. In general, the amount of unchanged amphetamine excreted in urine differs with the pH of the urine. Acidification of urine can double the unchanged fraction to approximately 60%. What is not excreted unchanged is metabolized to phenylacetone, which is then oxidized to benzoic acid and excreted as hippuric acid.

Amphetamine Pharmacology

Amphetamines exert two main effects on the dopaminergic system. Vesicles located inside the presynaptic

terminal contain dopamine. They release dopamine into the presynaptic terminal, which is then shuttled out into the synaptic cleft via dopamine transporter (DAT) molecules located on the surface of the presynaptic terminal. Simultaneously, the DAT transports an amount of dopamine back into the terminal. Thus, the transporter serves to release dopamine from the presynaptic terminal into the cleft and/or reuptake dopamine from the cleft into the presynaptic terminal. It is bidirectional, and its kinetics are dependent to a large degree on concentration gradients.

Amphetamine exerts actions in both places. It causes release of dopamine from the vesicles as well as blocking reuptake of dopamine at the DAT. Importantly, amphetamine promotes the outward release of dopamine through the transporter as well. Thus, amphetamine acts on both sides of the bidirectional DAT. Interestingly, amphetamine must first become a substrate of the DAT and be taken into the terminal in order for it to then act to promote efflux. It is this necessary property that has led some to examine the role of blocking the DAT in order to prevent neurotoxins (e.g., MPTP, a model for Parkinson's disease) from entering the terminal and subsequently destroying it.

Additional feedback control of dopamine regulation is provided by norepinephrine. Blockade of the norepinephrine transporter (NET) leads to increased dopamine release and increased amphetamine-related behaviors. Some confusion arises because of the often misleading names of the transporters. The NET and DAT are 70% homologous, and they are both sodium/chloride-dependent transporters. The norepinephrine transporter actually transports norepinephrine and dopamine in the prefrontal cortex, in contrast to the nucleus accumbens, in which the NET has little effect on dopamine. Consequently, NET inhibitors would have a major impact on dopamine levels in the prefrontal cortex but no effect in the nucleus accumbens. This partly explains why amphetamine, a potent NET inhibitor in the prefrontal cortex, raises dopamine there more than methamphetamine, which has minimal effects on the NET. In the nucleus accumbens, both amphetamine and methamphetamine raise dopamine levels equivalently because they are both acting solely through the strength of the DAT.

This same effect is responsible for the often confusing observation that atomoxetine, a 'selective' NET inhibitor, raises dopamine levels as much as clinically equivalent doses of amphetamine in the prefrontal cortex, but it has no effect on dopamine in the nucleus accumbens. From this, it can be seen how atomoxetine may improve cognition (through dopamine and/or norepinephrine in the prefrontal cortex) without causing reinforcement or addiction (the absence of dopamine effects in the nucleus accumbens).

Dopamine efflux is apparently mediated through a change in transporter confirmation. Amphetamine binding to the DAT allows for sodium influx, which in turn causes a calcium influx. This calcium then mediates protein kinase activation that results in the phosphorylation of the N-terminus of the DAT. This appears to relate only to dopamine efflux because deletion of N-terminal amino acids affects dopamine efflux but has no effect on reuptake.

Protein kinase C (PKC) is integral for mediating dopamine efflux and amphetamine's effects; blockade of PKC decreases or abolishes the locomotor activation of amphetamines, and PKC activators cause effects that are similar to those of amphetamine with a similar release of dopamine. PKC activity requires intracellular calcium, further identifying the specific isoform (classical over nonclassical isoforms) of PKC relevant to amphetamine activity.

Behavioral sensitization to the effects of amphetamines is known. However, even before behavioral sensitization, repeated exposure to amphetamine leads to a sensitization of the release of dopamine in the ventral tegmental area and striatal regions of the brain. Investigations have elucidated how dopamine efflux sensitization through the NET is calcium and calcium/calmodulin dependent. Amphetamines increase intracellular calcium through N- and L-type calcium channels. Calcium/calmodulin protein kinase II is activated, which promotes dopamine efflux. Inhibitors of calcium calmodulin protein kinase II block this efflux. This is in contrast to ordinary dopamine efflux after acute doses of amphetamines, which are neither calcium related nor affected by such inhibitors. Additionally, vesicular release of dopamine is not related to calcium or calcium/calmodulin.

Amphetamine Adverse Reactions and Safety

As noted previously, the high abuse potential and risk of dependence and side effects have resulted in the FDA listing amphetamines as Schedule II drugs. Their positive effects on alertness and cognitive functions must be carefully weighed against their unwanted effects on subjective state, disruptions of sleep physiology, cardiovascular and metabolic dysregulation, and abuse liability.

Caffeine

Caffeine is the most commonly used stimulant in the world. Coffee, its primary delivery vehicle, is the second largest commodity traded, after oil. Curiously, despite its popularity, important misconceptions about caffeine persist. The caffeine content in artificial beverages

such as soda is generally well controlled, ranging from 80 mg in an 8 oz. energy drink to 38 and 45 mg in two popular colas. The caffeine for these beverages is largely derived from the decaffeination of coffee. The caffeine content in coffee varies for numerous reasons. Each (fresh) coffee bean contains different amounts of caffeine. For example, Arabica – the 'higher quality' bean used in fine coffees and espresso – has half the caffeine of Robusta, the bean used in canned supermarket coffees. Simply roasting the coffee bean (to form dark roast or light roast) can decrease the caffeine content as the volatile caffeine is burned away. However, water is also lost in this process. Thus, by bean weight, dark roast contains more caffeine, but by bean volume (i.e., by measurement with tablespoons) a cup of dark roast will contain less caffeine than light roast. Amounts can vary significantly; light roast Colombia Supremo has a caffeine content of 1.3%, whereas dark roast Sumatras and mocha javas contain 0.02%. Similarly, a typical 1 oz. espresso drink will contain half the caffeine of a regular coffee due to the roasting, as well as the limited amount of water used in the preparation.

Espresso is not a type of bean; it is a fine grind of conceivably any coffee, although typically Arabica. Hot water is forced through the bean at high pressures (130 psi) to extract the oils in the coffee, including the caffeine. Because the 'espresso bean' is generally of the dark roast and Arabica variety, it will contain less caffeine than coffee. However, the primary determinant is the amount of water used and the length of time the water comes in contact with the grounds.

Finally, the caffeine content varies based on the subtleties of preparation. The same coffee from a coffee shop can have significant variability of caffeine content from day to day; differences in the amount of water or coffee used by each preparer can have a great impact. Additionally, home-brewed coffees are generally prepared using more water and consumed in greater volumes.

Caffeine has been shown to improve wakefulness and cognitive performance across a variety of measures (e.g., reaction time, attention, and memory), especially in sleep-deprived subjects. Such effects are found at a variety of caffeine doses. However, doses higher than 200 mg are generally superior to lower doses in improving performance on tasks and sustaining the improvement for longer time periods. Additionally, the longer the duration of sleep deprivation, the greater the caffeine dose necessary to reverse the effects.

Experiments in mice have established caffeine's reinforcing effects, and among prospective trials in healthy adults, a preference was demonstrated for caffeine over decaffeinated controls. Preference appears to be limited to dietary doses, such that increasingly higher doses lead to decreased preference, and avoidance occurs for doses higher than 400 mg. Caffeine reinforcement is also mood and context dependent, with those who prefer the effects of caffeine more likely to choose it. Similarly, negative mood states such as anxiety decrease caffeine preference. It is known that withdrawal effects such as headache or drowsiness reliably predict subsequent caffeine preference. Interestingly, a level of physical dependence (without withdrawal) predicts preference and enjoyment of caffeine, but in the absence of physical dependence this choice is significantly lessened.

There is evidence that caffeine may be protective against Parkinson's disease. It has been observed that caffeine consumption was inversely correlated to Parkinson's disease risk. Using various models of induced Parkinson's disease in mice (e.g., 6-OHDA and MPTP), caffeine was shown to have a protective effect on dopamine levels. This appears to be via its inhibition on A2a adenosine receptors. Decreases in dopamine result in increased glutaminergic activity, which causes destruction of dopaminergic neurons in the substantia nigra. Not inconsequentially, caffeine exerts positive pharmacokinetic effects on Parkinson's disease medications, shortening the time to maximal plasma concentration and the effects of levodopa. The effects of caffeine with levodopa appear to be additive or even synergistic.

An important source of confounding information about caffeine is the tendency to extrapolate acute dosing studies to the effects of chronic use and vice versa. For example, acute administration of 500 mg of caffeine lowers the seizure threshold and is a common maneuver prior to electroconvulsive therapy. However, chronic use (14 days) of caffeine in fact reduces the severity of seizures. Adenosine has anticonvulsant properties, perhaps via inhibition of glutamate or aspartate, and thus it is reasonable to expect that caffeine may alter seizure susceptibility via blockade of adenosine receptors.

Caffeine provides benefits in physical exertion on both performance and endurance. It has even been found that in dietary doses caffeine can reduce by almost half the subjective rating of pain induced by 30 min of exertion. Although its effects on performance are not profound, in competition sports these may be very significant, which may explain why caffeine was once banned by the International Anti-Doping Agency but subsequently taken off the list due to its widespread use. Importantly, the effects of caffeine on performance and endurance are not apparently related to glycogen use or muscle lactate

(i.e., anaerobic metabolism). Whether these gains are the direct result of increased sympathetic activity or through some other mechanism is not clear.

Caffeine Metabolism and Elimination

The solubility of caffeine varies substantially with temperature and is approximately 2.2 mg ml^{-1} (66 mg oz^{-1}.) at 25 °C and 180 mg ml^{-1} at 80 °C. Up to 90% of caffeine is hepatically metabolized, mostly to paraxanthine (80%) and theobromine (10%) and trivial theophylline (5%). Paraxanthine is generally further metabolized to inactive compounds before excretion. Importantly, this metabolism is species dependent and has some bearing on the interpretation of pharmacologic data. Rats metabolize less (30–40% excreted unchanged) and have markedly shorter half-lives (1 h for rat, and 3–5 h for human). Paraxanthine has similar effects as caffeine in terms of both its stimulant and its immune system effects. Theophylline is similarly active, but theobromine appears to be considerably less active.

Caffeine Pharmacology

Caffeine's clinical effect on alertness is mediated by nonselective antagonism of adenosine receptors (A1 more so than A2), which are widely distributed throughout the brain. Selectivity for the A2 receptor can be obtained by altering one methyl group with a larger alkyl group. Additionally, the potency can be increased by changing all of the methyls to larger groups.

Although generally not obtained in dietary doses, caffeine has a potent antagonistic effect on phosphodiesterase, which has the result of increasing cAMP and activating protein kinase A. This in turn blocks proinflammatory processes, such as lipopolysaccharide-induced production of tumor necrosis factor-α and interleukin (IL)-12, and promotes anti-inflammatory cytokines including IL-10 and mononuclear cells.

Caffeine's effects through adenosine antagonism are more predictable. Adenosine is released in response to tissue injury and induces collagen production and fibrosis in affected tissue (e.g., liver). Although caffeine is known to inhibit A1 adenosine receptors, data indicate that this may have little relationship to its wakefulness-promoting activity. Pure A1 antagonists will decrease sleep in mice, but A1 knockout mice do not respond to A1 antagonists. Although it may therefore seem logical to conclude that A1 is necessary for sleep, A1 knockout mice do not sleep any longer or shorter than wild-type mice; it is only their sensitivity to A1 antagonists that is different.

Caffeine promoted wakefulness in wild-type mice, but in contrast to the expectation, it also promoted wakefulness in A1 knockouts. It had no effect, however, on A2a knockouts. Taken together, these findings suggest that A1 and A2 receptors work in tandem to regulate wakefulness and sleep, and that a degree of compensation exists, especially for A1 by A2a. Additionally, it suggests that caffeine's wakefulness effects are not related to its majority effect on A1 but, rather, its 'minor' pharmacologic effect – which becomes the main clinical effect – on A2a receptors. This is further supported by location-specific effects of caffeine's stimulant properties in striatal neurons, in which A2a receptors are abundant.

Caffeine Adverse Reactions and Safety

The safety of caffeine as a food and beverage additive has been evaluated, with the conclusion that caffeine added to beverages at a level of 0.02% or less does not present a health risk. Caffeine has been shown to produce a transient increase in blood pressure, but a number of studies have failed to link caffeine with cardiovascular disease. Caffeine may be associated with a small increase in spontaneous abortion in the first trimester of pregnancy, and caffeine can significantly increase 24 h urine output. Because caffeine is not a prescription medication, standard dosing and side effect information is not available. The information in the *Physicians Desk Reference for Nonprescription Drugs and Dietary Supplements* for the product Nō-Dōz, which contains 200 mg of caffeine, indicates that the product may cause "nervousness, irritability, sleeplessness and, occasionally, rapid heart beat." Few research studies report carefully controlled side effect data from matched caffeine and placebo groups. Some investigators estimate that as many as 10% of adults develop the syndrome of caffeinism, which is defined as the daily consumption of large amounts of caffeine (usually in excess of 500 mg day^{-1}). Symptoms of caffeinism or caffeine intoxication include restlessness, nervousness, excitement, insomnia, flushed face, diuresis, gastrointestinal disturbance, muscle twitching, rambling speech, tachycardia, and agitation. The lethal acute adult dose of caffeine is approximately 5–10 g (approximately 75 cups of coffee).

See also: Amphetamines; Narcolepsy; Parasomnias; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Research and Sleep Medicine in Historical Perspective; Sleep Architecture; Sleep Apnea; Sleep Deprivation: Neurobehavioral Changes; Sleep Deprivation and Brain Function.

Further Reading

Bonnet MH, Balkin TJ, Dinges DF, et al. (2005) The use of stimulants to modify performance during sleep loss: A review by the Sleep Deprivation and Stimulant Task Force of the American Academy of Sleep Medicine. *Sleep* 28(9): 1163–1187.

Dinges DF, Arora S, Darwish M, and Niebler GE (2006) Pharmacodynamic effects on alertness of single doses of armodafinil in healthy subjects during a nocturnal period of acute sleep loss. *Current Medical Research and Opinion* 22: 159–167.

Kantor L, Zhang M, Guptaroy B, Park YH, and Gnegy ME (2004) Repeated amphetamine couples norepinephrine transporter and calcium channel activities in PC12 cells. *Journal of Pharmacology and Experimental Therapeutics* 311(3): 1044–1051.

Mobbs C (2005) Caffeine and the adenosine receptor: Genetics trumps pharmacology in understanding pharmacology. *Cell Science Reviews* 2.

Vanderveen JEC, Armstrong LE, Butterfield GE, et al. (2001) *Caffeine for the Sustainment of Mental Task Performance. Formulations for Military Operations*. Washington, DC: National Academy Press.

Stomatogastric Ganglion Models

A I Selverston and M I Rabinovich, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The stomatogastric system of crustaceans is one of a large number of invertebrate central pattern-generating circuits that have been investigated during the past 40 years. It can be considered a 'model' system with which to understand the mechanism underlying rhythmic behaviors in higher animals. The heart of the system is the 30-celled stomatogastric ganglion (STG) containing two central pattern generators (CPGs): the pyloric, which controls the pyloric region of the stomach, and the gastric, which innervates three ossicles (teeth) within the stomach. The pyloric CPG operates at a frequency of approximately 2 Hz and is composed of 13 motor neurons and one interneuron. The slower gastric CPG has an *in vitro* frequency of approximately 0.1 Hz. It is composed of 10 motor neurons and one interneuron. The synaptic connections between all of these neurons have been fully described, so the two CPGs are the most well-known neural circuits available in terms of cell-to-cell interactions. With so much information known about this system, one might assume that it is easy to understand exactly how it works. In fact, due to the complexity of the nonlinear interactions, it is not. Modeling may be the only way to understand the two networks and predict how they will respond to various inputs and perturbations in a way that allows them to retain the correct motor functions. The following are the principal modeling advantages of the preparation:

- The motor output patterns can be correlated with muscle contractions, thus directly linking the neuronal activity with behavior.
- 'Fictive' spatiotemporal patterns consisting of bursts of action potentials can be recorded from the motor nerves for many hours *in vitro* when the two 'higher' commissural ganglia are left attached.
- Neurons are large, easily identifiable, and display subthreshold excitatory and inhibitory potentials – important for identifying the integrative properties of the neurons.
- Up to eight neurons can be recorded simultaneously with intracellular electrodes.
- Command input from higher centers, sensory feedback, and neuromodulatory substances can all modify bursting and pattern formation, in some cases rather severely.

Detailed biophysical information for some of the identified stomatogastric neurons has been measured and the relative strengths and time courses of many of the chemical synapses are also known. Assembling all of this reductionistic data into a coherent system has proven challenging. In principle, one would like to explain the following features of this (or any) CPG:

1. What is the origin of the rhythmogenesis?
2. How are the patterns of bursts established?
3. Are rhythmogenic and pattern formation mechanisms separate?
4. How can the competing interests of robustness and flexibility be reconciled?
5. How is the system turned on and off?
6. How is the frequency established?
7. What is the mechanism for retaining optimal phase relationships at different frequencies?

Although researchers have attempted to answer these questions by trying to explain the experimental data verbally, such attempts are limited and cannot provide a quantitative description of the many phenomena observed. For example, a verbal model cannot predict what will occur when a selective perturbation to the system is imposed. Models that use the available experimental data to simulate the pyloric or gastric CPGs attempt to mimic the production of normal spatiotemporal patterns as well as have predictive value. Models of small CPGs can help to formulate the basic principles underlying central motor generators in general. For example, are there certain universal phenomena that can be found in all rhythmic CPGs? If there are general principles, it is probable that they will be found at the molecular level, whereas CPG networks in different species will have mostly idiosyncratic synaptic arrangements between neurons – that is, neural CPG circuit topology may vary for different operational requirements in different species. The stomatogastric CPGs, particularly the pyloric circuit, have been used to attempt to answer many questions about neural circuits. Both individual STG neurons and pairs of neurons and small networks have been modeled. In addition, the STG has been most widely used for computer/electronic neuron interfacing with biological neurons in so-called hybrid circuits.

Overview of the Stomatogastric System

To study the infrastructure underlying rhythmic behaviors, only invertebrate preparations are experimentally tractable at all levels of analysis. The

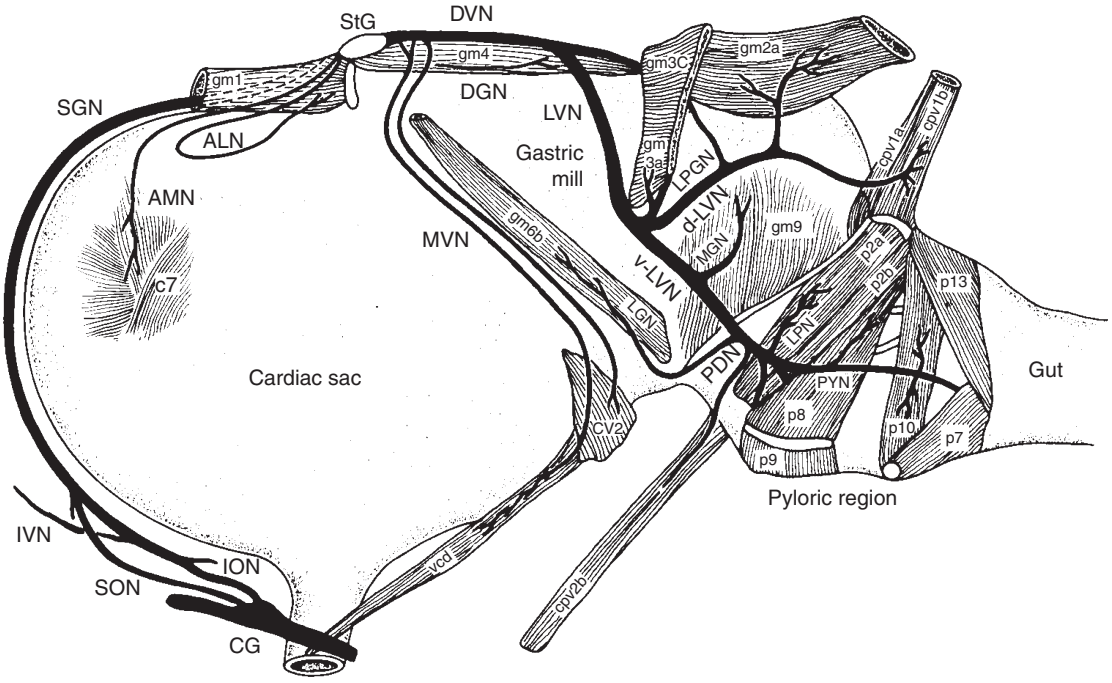

Figure 1 The crustacean stomatogastric system innervates the stomach shown here in side view. The nerves can be seen coursing along the surface of the stomach wall. The stomatogastric ganglion (STG) sits at the top of the stomach. CG, commissural ganglion; DVN, dorsal ventricular nerve; ION, inferior esophageal nerve; IVN, inferior esophageal nerve; LVN, lateral ventricular nerve; SGN, stomatogastric nerve; SON, superior esophageal nerve; v-LVN, ventral part of lateral ventricular nerve. Motor nerves: ALN, anterior lateral nerve to gm1; AMN, anterior median nerve to c7; DGN, dorsal gastric nerve to gm4; LGN, lateral gastric nerve to gm6b; LPGN, lateral posterior gastric nerve to gm3a; LPN, lateral pyloric nerve to p2a; MGN, median gastric nerve to gm9; MVN, median ventricular nerve to cv2 and vcd; PDN, pyloric dilator nerve to cpv2b; PYN, pyloric nerve to muscles p2b, p8–9, p10, p7, and p13.

stomatogastric circuits have thus been studied from the level of the molecular structure of ion channels to the modulation of gastric teeth behavior by neuromodulators. As mentioned previously, there are two rhythms, the gastric and the pyloric. The gastric rhythm is for 'chewing' food that has been swallowed and stored in the cardiac sac portion of the stomach (Figure 1). As with our own masticatory behavior, gastric mill (GM) behavior is episodic and highly flexible, depending on the consistency of the food present. The pyloric rhythm controls a pumping/filtering apparatus at the caudal end of the stomach, and although it's speed can be adjusted, it appears to work almost continuously. The neurons that comprise both CPGs, like many invertebrate CPGs, are few in number and the patterns are robust (Figure 2). The microcircuits formed by these neurons have been known for more than 30 years and are shown in Figure 3. The neurons in both CPGs have two functions. Their biophysical properties and synaptic interactions cooperate to generate multiphase rhythms within the ganglion (i.e., they function like premotor interneurons in vertebrates). All but two of them are also motor neurons that send bursts of impulses to the muscles. The two neurotransmitters in the ganglion – glutamate and acetylcholine – make mostly inhibitory connections onto other neurons within the ganglion but make excitatory neuromuscular synapses peripherally. In addition, like synapses in the retina, synapses in the STG release transmitter as a graded function of the membrane voltage in the presynaptic terminals and not just as pulsatile events following spike invasion of the terminal.

STG Output Patterns

When the STG and its associated nerves are removed from the wall of the stomach and the two commissural and the single esophageal ganglia are left connected, both the pyloric and gastric rhythms will burst for many hours (Figure 2). The patterns of spiking activity in the motor nerves correspond closely with the natural behaviors of the stomach. The pyloric dilator (PD) neuron burst is followed sequentially by lateral pyloric (LP) and pyloric (PY) bursts. This three-phase sequence acts on pyloric muscles to first dilate the pyloric region since pairs of pyloric muscles pull in opposite directions (Figure 1). This is followed by a rostral-to-caudal wave of muscle contractions with the whole system acting like a pump. The bursts in the GM units also correspond to the movements of the three teeth of the GM which have been studied extensively by video endoscopy. The lateral gastric (LG) and median gastric (MG) traces, whose burst

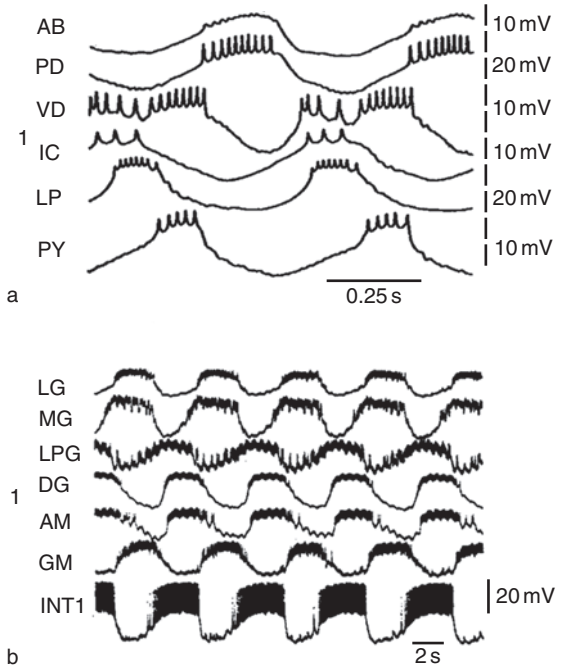

Figure 2 Simultaneous intracellular recordings from individual neurons of the pyloric and gastric CPGs. The spikes on the top of each burst form the patterns of activity seen in each motor nerve. The location of each neuron in the two circuits is shown in Figure 3. AB, anterior burster; AM, anterior median; DG, dorsal gastric; GM, gastric mill; IC, inferior cardiac; Int 1, interneuron 1; LG, lateral gastric; LP, lateral pyloric; LPG, lateral posterior gastric; MG, median gastric; PD, pyloric dilator; PY, pyloric; VD, ventricular dilator.

onsets are shifted slightly, close the two lateral teeth while strong lateral posterior gastric (LPG) bursts that occur in antiphase open the lateral teeth. The dorsal gastric (DG) and anterior median (AM) reset the medial tooth and help with contractions of the cardiac sac while the GMs act out of phase with DG to provide the power stroke movements of the medial tooth. Heinzel described a small variation of this behavior which occurs spontaneously. Instead of the lateral teeth coming together first and the medial tooth grinding over them, in some instances he observed all three teeth coming together at the same time – a behavior he referred to as 'squeeze.'

The CPGs and Some Associated Circuitry

The pyloric and GM neural circuits are shown in Figure 3. The pyloric circuit is composed of two PD cells, eight PY cells, and one each of the ventricular dilator (VD), LP, inferior cardiac (IC), and anterior burster (AB) cells. In a combined preparation with the two commissural ganglia and the single esophageal ganglion intact, all of the neurons in the pyloric network burst intrinsically. However, because they require neuromodulatory inputs from the commissurals, they are referred to as 'conditional bursters.' Without this modulatory input, the intrinsic conductances underlying bursting are not sufficiently activated and the cells are either silent or fire tonically but not in bursts. The AB neuron has the fastest rhythm, thus acting as a pacemaker for the other pyloric cells. By deleting neurons from the circuit sequentially, it has been shown how the synaptic connectivity and the intrinsic burstiness combine to produce a robust three-phase burst pattern. The ionic mechanism for bursting is basically a decrease in the potassium leak current that depolarizes the neuron until a voltage-gated calcium and sodium current is activated. This increases the rate of the depolarization and activates fast spiking and plateau currents. Inactivation of the sodium and calcium currents combined with calcium-activated potassium current brings the cell back to its resting potential. Other currents, particularly I_A and I_H, play an important role in burst timing and are considered later. It is important to note here that the rhythmicity is due to the intrinsic properties of the neurons. The network properties form the pattern by the appropriate use of inhibitory synapses and also may contribute to the rhythmicity.

The inhibitory synapses in the GM CPG play the determinant role in pattern formation and a major contributory role in rhythmogenesis. When neurons are isolated pharmacologically, only the DG and AM neurons appear to have intrinsic pacemaker properties. The key LG and DG neurons evoke two types of inhibitory postsynaptic potentials (IPSPs) in follower neurons. The fast type rises rapidly (100–300 ms), is mediated by an increased chloride and potassium conductance, and resembles the glutaminergic IPSP described for the pyloric CPG. The second, slow type of IPSPs has a long rise time (1 or 2 s) and is mediated by an increased conductance to potassium. The fast inhibition occurs alone at connections from DG and LG to power stroke motor neurons (MG and GM). Slow inhibition occurs in parallel with fast inhibition (producing dual-component responses) at LG to LPG and interneuron 1.

Although the two CPGs control separate oscillatory rhythms, there are many direct synaptic and electrotonic connections between neurons in each circuit. In addition, several neurons that are located in the commissural ganglia receive inhibitory synaptic inputs from the neurons in each CPG and therefore send pyloric and gastric timed inputs to neurons in both CPGs in a kind of closed feedback loop. This is most apparent on intracellular recordings from the DG and AM neurons which show bursts of excitatory postsynaptic potentials (EPSP) that originate from P cells that are modulated by the AB neuron. Since the pyloric and gastric CPGs operate at different

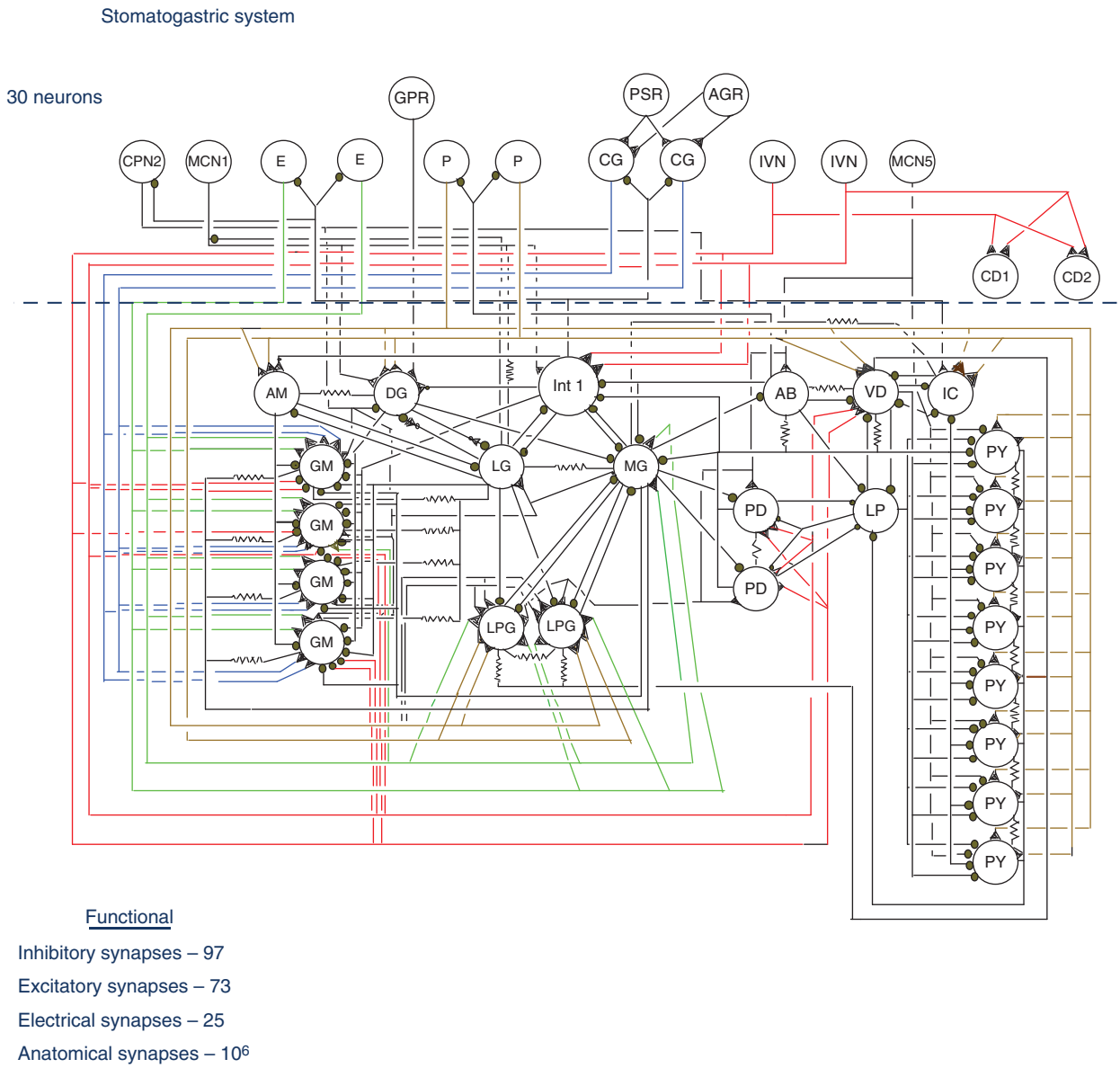

Figure 3 Pyloric and gastric mill neural circuits, including descending neurons from the commissural ganglia and sensory neurons on the stomach wall. Triangles represent chemical excitatory synapses, black dots represent chemical inhibitory synapses, and resistors are gap junctions. Neurons at the top of the figure (above the dashed line) are various excitatory neurons, the P cells and E cells, the IVN fibers from the brain, and various sensory receptor cells. Pyloric CPG: AB, anterior burster; IC, inferior cardiac; LP, lateral pyloric; PD, pyloric dilator; PY, pyloric; VD, ventricular dilator. Gastric CPG: AM, anterior median; DG, dorsal gastric; GM, gastric mill; Int 1, interneuron 1; LG, lateral gastric; LPG, lateral posterior gastric; MG, median gastric.

frequencies, the possible functions of these connections as well as the general computational problem of coupling oscillators running at different frequencies pose interesting questions.

Ion Channels in STG Neurons

As already mentioned, several voltage-gated conductances are responsible for the intrinsic burst-generating mechanisms found in STG neurons. Other channels are important in controlling various aspects of the neural patterns produced by the STG. The somata of STG neurons contain only outward current channels and are therefore inexcitable. Interestingly, when isolated neuronal somata are placed in culture, after a few days inward current-carrying channels are expressed and neurons begin to burst.

Two other currents which play an important role in determining the characteristics of the pattern are I_H and I_A. I_H is activated by hyperpolarization of the membrane and plays a role in determining the extent of postinhibitory rebound following inhibition. I_A is a transient depolarization activated K current that delays the onset of bursting. Voltage

clamp studies on individual pyloric neurons have shown that the distribution of I_A channels differs between cells. It is largest in PD and PY cells, smaller in AB, LP, and IC cells, and not detectable in the VD cell. When I_A currents are blocked with 4-AP, the cycle period is decreased by approximately 20%, there is a general enhancement of activity in all of the cells, and there is a change in the phasing of follower cells with respect to the PD/AB pacemaker group.

Neuromodulation

The gastric and pyloric CPGs are subject to the influence of neuromodulatory substances released into the ganglion nonspecifically or as a result of being transported by the blood. More than 20 neuromodulatory substances (mostly small peptides and amines) have been identified by immunohistochemistry in neurons that project to the STG. How these substances are used in the life of the animal is known only in a few cases (e.g., cholecystokinin). However, when they are applied to the ganglion in physiological dosages, they have profound effects on the motor patterns that last for many hours. The mechanism of action has been shown to be a result of G-protein activation of second-messenger cascades and an increased concentration of different kinases leading to the phosphorylation of specific channels in specific neurons. Because this changes the biophysical properties of the neurons and their synapses, there is a functional rewiring of the circuit. The consistent effects of particular neuromodulators depend on the fact that the biophysical properties and the types and densities of channels in the neurons comprising the STG circuitry are consistent from animal to animal.

Models of Observed Phenomena

Models of Single STG Neurons

If one considers the complexity of even a single neuron – its various ionic fluxes, intracellular mechanisms, responses to synaptic inputs, etc. – it is reasonable to assume that a single isolated cell is a high-dimensional system with some noise (i.e., requiring many equations to describe all of its functional variables). A nonlinear dynamical analysis using a global false nearest neighbor paradigm, however, showed that the synaptically isolated but still stationary bursting LP neuron has an attractor that demonstrates relatively low dimensional dynamics – on the order of 4 or 5 (Figure 4). In addition, the spikes in the bursts of this cell develop chaotic behavior toward the end of the bursts that could be shown by phase portraits or Lypunov exponent measurements. Not all neurons in the gastric or pyloric system were examined, but it is likely that other neurons have chaotic properties as well.

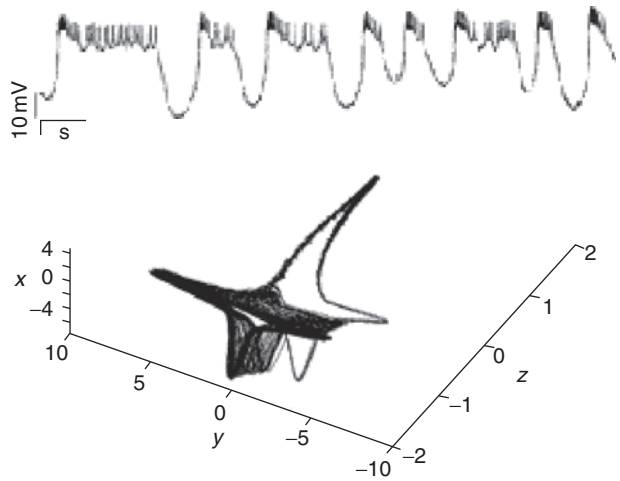

Figure 4 (Top) Chaotic membrane potential time series of a synaptically isolated LP neuron from the pyloric CPG. (Bottom) State space attractor reconstructed from the voltage measurements of the LP neuron shown in the top panel using delayed coordinates $[x(t), y(t), z(t)] = [V(t), V(t-T), V(t-2T)]$. This attractor is characterized by two positive Lyapunov exponents.

Many models of single neurons have been made, both simplified and conductance based. These have had varying degrees of success. It is important to emphasize that a simple reproduction of the stationary burst pattern is only the first step in creating a satisfactory model. Any model, to be adequate, has to be able to describe the transient dynamics as well. An isolated LP neuron model can be successfully made with a Hindmarsh–Rose (H–R) single compartment model consisting of only three equations. This model can show chaotic as well as regular behavior simply by adjusting the parameters. Adding a fourth polynomial equation to the original H–R formulation shown here produces chaotic behavior over a much more realistic regime range. The original equations shown here describe membrane voltage (x) and a fast (y) and a slow (z) current:

$$\frac{dx(t)}{dt} = ay(t) + bx^2(t) - cx^3(t) - dz(t) + I$$

$$\frac{dy(t)}{dt} = e - fx^2(t) - y(t) - gw(t)$$

$$\frac{dz(t)}{dt} = \mu(-z(t) + s(x(t) + h))$$

A much slower current (w) is given by

$$\frac{dw(t)}{dt} = v(-kw(t) + r(y(t) + l))$$

The slow current could represent any slow process, such as an inositol triphosphate-controlled movement

of internal calcium in and out of the endoplasmic reticulum calcium stores. It is the slower kinetics of this process, combined with the faster inward and outward membrane currents, that generates the appropriate chaotic behavior (Figure 5).

The AB neuron has also been numerically simulated using a Hodgkin–Huxley-type model that was able to produce various kinds of rhythmic behaviors in response to neuromodulators or channel blocking neurotoxins that were similar to the responses observed experimentally. By constructing two-dimensional maps of the parameter space, the changes in firing patterns of the models could be determined. Using a bifurcation analysis, regions in which the model's behavior transitioned from tonic firing to bursting could be shown. The model also mimicked the results of blocking I_A with 4-AP over a range of I_A reductions. Overexpression of I_H in PD neurons has also been modeled. The increase in this current causes a dramatic change in PD activity by depolarizing the minimum membrane potential, thereby reducing the amplitude of the oscillation. This leads to decreased time to the first spike in the burst, an increase in the duty cycle, and an increase in the number of spikes per burst.

Manipulation of both I_A and I_H using molecular biological techniques indicates that there is an activity-independent regulation of these channel proteins. An increase in either protein alone has little effect on cell firing properties in PD, LP, VD, and IC neurons. Two hypotheses might explain this remarkable observation. There might be a mislocalization of, for example, the protein Shal (for I_A) to the soma and initial neurites so as to be remote from the integrating region of the neuropile. Alternatively, the increase in I_A could be balanced by a homeostatic increase in I_H. A multicompartment model with the I_A current located in the soma and primary neurites produced simulations that were much greater than those observed experimentally with Shal-injected neurons. However, a model in which the ratios of I_A and I_H were kept constant showed complete compensation, suggesting that an activity-independent homeostatic mechanism maintained constant spiking–bursting neural activity after increasing the amount of I_A in the neuron.

The basis of most H–H-type models for identified invertebrate neurons is that the maximum conductances for each channel type remain constant. However, clearly there is channel protein turnover during the life of a neuron, so mechanisms must exist to keep the expression of each channel protein constant or, as has been suggested, at least keep the activity level of the neuron constant. Here, the idea is that protein

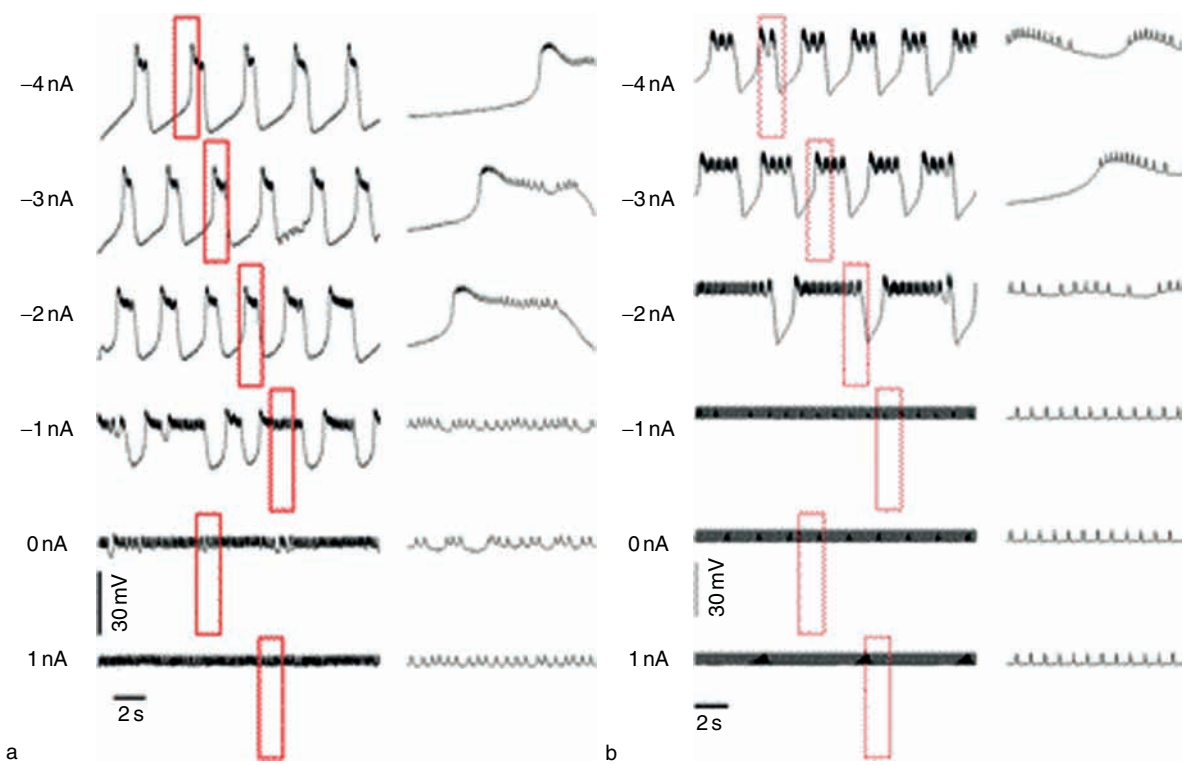

Figure 5 A conductance-based model of the LP neuron. Nine separate conductances were tuned to produce the output shown for different levels of applied DC current. (a) Responses in biological neuron to applied DC currents (b) are model responses to the same current levels. Red boxes show portions of the curve that are magnified in time.

expression is influenced by the electrical activity of the cell. There is experimental evidence from mammalian neurons that this is the case and evidence from STG work that there is also an activity-independent modification of channels that results from injection of specific channel mRNA into the neuron. Activity-dependent conductances in STG neurons have been modeled using internal calcium sensors as measures of cell activity and then up to three different equations have been used to transform the calcium concentration to changes in the maximum conductance values for the neuron. These models suggest that the maximal conductance values are not fixed parameters but, rather, dynamic and able to produce different cell behaviors by constantly adjusting the ratios of different ionic conductances. However, there is little evidence linking internal calcium concentration to changes in gene expression in STG neurons.

The idea, based on modeling, that there is a range of different conductance value ratios that can lead to the same activity profile has been extended to include very large numbers of combinations in a so-called database approach. Usually, conductance-based models use some form of tuning algorithm to obtain parameter values that produce the correct voltage responses to specific regimes. This can be a long and laborious process if the model is to be structurally stable under a variety of perturbations. There are two ways to avoid this problem. One is to determine the exact parameter values from stationary experimental data. However, this is usually not possible for various technical reasons associated with the complex morphology of neurons and the inability to voltage clamp at the precise location of voltage-sensitive currents. An alternative method, using modeling techniques, is to build a large database of many possible combinations (arbitrarily limited) of biophysical parameters and select those combinations which produce voltage patterns that match a template representing a time series of the biological voltage outputs for the cell. This suggests that neurons (and synapses) can have many different combinations of values for their conductances, some with as large as a 30-fold difference. The biological evidence to support this idea is weak, however, and there are many reasons why identified neurons could not have such a large degree of variability between animals, the most relevant of which would be that there is always a consistent response in particular identified neurons to a single neuromodulator or to a specific channel blocker.

The effect of the neuromodulator dopamine (DA) on PY and LP activity has been modeled in conjunction with experimental work showing that DA reduces the I_A current significantly. By reducing this current in PY neurons, a phenomenon known as postinhibitory rebound is altered in a way that causes a phase advance. The mechanism appears to be due to a reduction of the slowly inactivating component of the I_A current shifting the voltage activation curve in the depolarizing direction. A similar model of the LP neuron confirmed experimental data that indicated DA caused a reduction in both I_A and I_H and a significant delay in first spike latency.

Models of Synapses

Models of synapses in the STG have taken into account not only strength but also time course and nonlinear temporal changes. Models of chemical synapses have to automatically adjust the amplitude of the EPSP or IPSP to the voltage difference between the actual potential and the reversal potential. More developed models have incorporated nonlinear synaptic properties such as short-term facilitation and depression and long-term potentiation and long-term depression. Since there are also many electrical synapses, equations describing the bidirectional and unidirectional (or rectifying) flow of current have been constructed. Equations for chemical synapses typically take the following form:

$$I_{\text{syn}} = gS(V - V_{\text{rev}})$$

$$\tau \frac{\mathrm{d}S}{\mathrm{d}t} = R - S$$

$$\tau \frac{\mathrm{d}R}{\mathrm{d}t} = \Theta(V - V_{\text{th}}) - R$$

where the threshold potential for transmitter release is V_{th}, τ is the synaptic timescale, R and S are the presynaptic and bound transmitter, and Θ is the Heaviside step function.

These models have been used to examine the dynamics of mixed chemical and electrical synapses or the role of synapses in the functioning of the STG networks. For example, an important property of a motor network is the ability to keep optimal phase relationships over a variable frequency range. Depressing synapses may be able to achieve this function by changing synaptic strength as their frequency speeds up or slows down.

Neurons coupled with gap junctions have shown some surprising results. Two PD neurons, normally coupled very strongly, fire together in a bursting–spiking pattern (Figure 6(a)). Modeling two electrically coupled neurons shows that neurons with chaotic bursting activity can be synchronized and regularized by the electrical connection. Furthermore, in the model, changing the strength of the coupling can produce different regimes of in-phase and out-of-phase bursting. This modeling result was confirmed experimentally by using an artificial electrical synapse

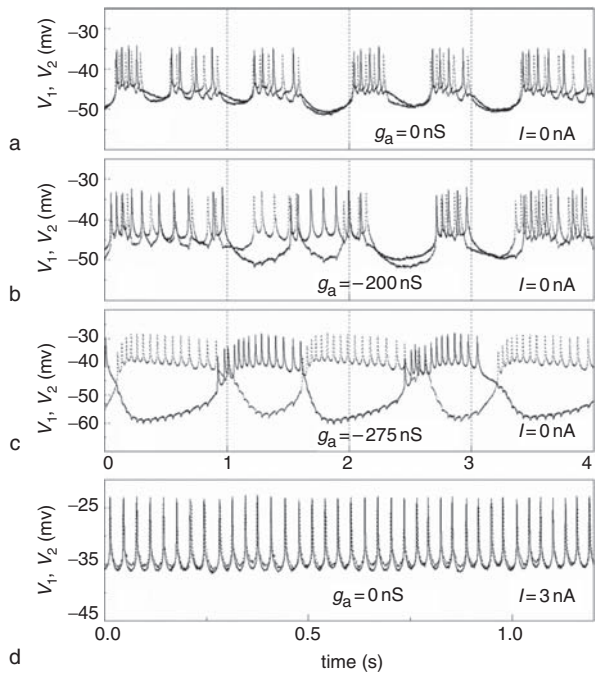

Figure 6 Artificial electrical coupling between two living chaotic PD cells of the stomatogastric ganglion of crustacean can demonstrate both synchronous and asynchronous regimes of activity. In this case, the artificial electrical synapse was built on top of the existing natural coupling between the two PD cells. The figure shows different synchronization levels (a–d) as a function of the artificial coupling g_a and a DC current injected I in one of the cells. (a) With their natural coupling $g_a = 0$, the two cells are synchronized and display irregular spiking–bursting activity. (b) With an artificial electrical coupling that changes the sign of the current $g_a = -200$ nS, and thus compensates the natural coupling, the two neurons behave independently. (c) Increasing the negative conductance leads to a regularized antiphase spiking activity (by mimicking mutual inhibitory synapses). (d) With no artificial coupling but adding a DC current, the two neurons are synchronized displaying tonic spiking activity.

between two PD cells. It is possible to study transitions between synchrony and asynchrony between the two neurons simply by changing the current flow between the cells (Figure 6).

Models of Pyloric Spatiotemporal Patterns

The two patterns shown in Figure 2 are the three-phase pyloric with a frequency of approximately 2 Hz and the five-phase gastric with a frequency of approximately 0.1 Hz. These spatiotemporal patterns are the result of the synaptic interactions and the biophysical properties of each neuron. The synaptic interactions can be seen in the simplified circuit diagrams shown in Figure 3. Because of the very large number of interactions, it is not possible to determine the mechanisms responsible for the performance of these CPGs by a simple inspection of the circuit. It is also not possible to predict the effects of perturbations even if the precise nature of the perturbation is known. Ideally, we want to be able to explain how the basic features of the pattern are produced (i.e., the genesis of the bursts and their phase relationships). In addition, we want to know how such parameters as burst length, frequency, and duty cycle are produced and maintained. Also, perhaps most difficult, we want to know how the combined cellular and synaptic properties remain robust while at the same time flexible enough to respond to sensory input and perturbations.

Most computer models of STG networks have thus far been limited to simplified subsets of neurons. For example, pairs or triplets of STG neurons coupled with inhibitory chemical synapses or pairs of neurons coupled with electrical synapses with the same topology as is found in the biological circuit can produce the three-phase pyloric pattern. For example, the pyloric network can be reduced to three oscillatory neurons representing the PD/AB group, the single LP neuron, and the eight PY neurons connected with inhibitory synapses.

Modeling Robustness and Adaptability

The CPG concept was built around the idea that behaviorally relevant spatiotemporal cyclic patterns are generated by groups of nerve cells without the need for rhythmic inputs from higher centers or sensory feedback from the structures that are moving. If activated, isolated invertebrate preparations can generate such rhythms for many hours and, as a result, have been extremely important in attempting to understand how the simultaneous cooperative interactions between many cellular and synaptic parameters can produce robust and stable spatiotemporal patterns whose mathematical image is a limit cycle.

Generally, neural oscillations can arise either through interactions among neurons (network-based mechanism) or through interactions among currents in individual neurons (pacemaker mechanism). The stomatogastric CPGs appear to use elements of both mechanisms. In the pyloric CPG, all of the neurons have intrinsic bursting activity but the fastest burster, the AB, acts as the pacemaker for the entire CPG circuit. The intrinsic currents are not constitutively active but require activation by neuromodulators; thus, they are called conditional bursters. In the gastric CPG, only one or two neurons have intrinsic bursting capability and the synaptic connections act to determine the pattern by exciting or inhibiting other neurons at the appropriate time. Networks with bursting neurons are extremely robust and have generally been thought to be present in systems in which the rhythmic activity is active more or less continuously. In the second case, the synaptic interactions between mostly nonbursty neurons generate the rhythmic activity, and many

schemes for the topology necessary to do this have been proposed. Usually, reciprocal inhibition serves as the basis for generating bursts in neurons that control antagonistic muscles, and there are many examples of cells in pattern-generating microcircuits connected this way. Network-based circuits are usually found for behaviors which are intermittent in nature and which require a greater degree of flexibility than those based on pacemaker cells. An example of a three-neuron CPG phase portrait is shown in Figure 7. The effect of a hyperpolarizing current in PD leads to changes in the pattern, as reflected by the phase portrait in Figures 7(b) and 7(c), but the fundamental activity of the network continues.

Despite the fact that the stomatogastric CPGs are the most thoroughly investigated neural networks, data on many important control parameters are lacking. This information is necessary to build a complete model, one that is able to predict the observable phenomena discussed previously. However, based on some important general ideas from nonlinear dynamics, it is possible to build a sufficiently realistic STG model capable of expressing dynamical principles, even without a detailed knowledge of all the morphological and neurophysiological details of the specific neurons and synapses. The success of such models is determined by the universality of the underlying dynamical principles: robustness, reproducibility, flexibility, and adaptivity. Based on these principles, CPGs have to be strongly dissipative (i.e., to return to the normal mode of activity quickly after a perturbation (robustness and reproducibility)) and to be structurally stable (i.e., to be able to keep the general form of the behavior even if the parameters of the network are changed slightly). At the same time, a CPG has to be sensitive (i.e., able to change the mode of activity) as a result of command inputs from above and/or sensory feedback. Mathematically, the robustness of a rhythmic CPG means that in the state space of the CPG there exists a limit cycle with a large basin of attraction. Structural stability means that the system is located far enough away from the bifurcation boundaries in the control parameter space that it will not be degraded by spontaneous variations in its parameters.

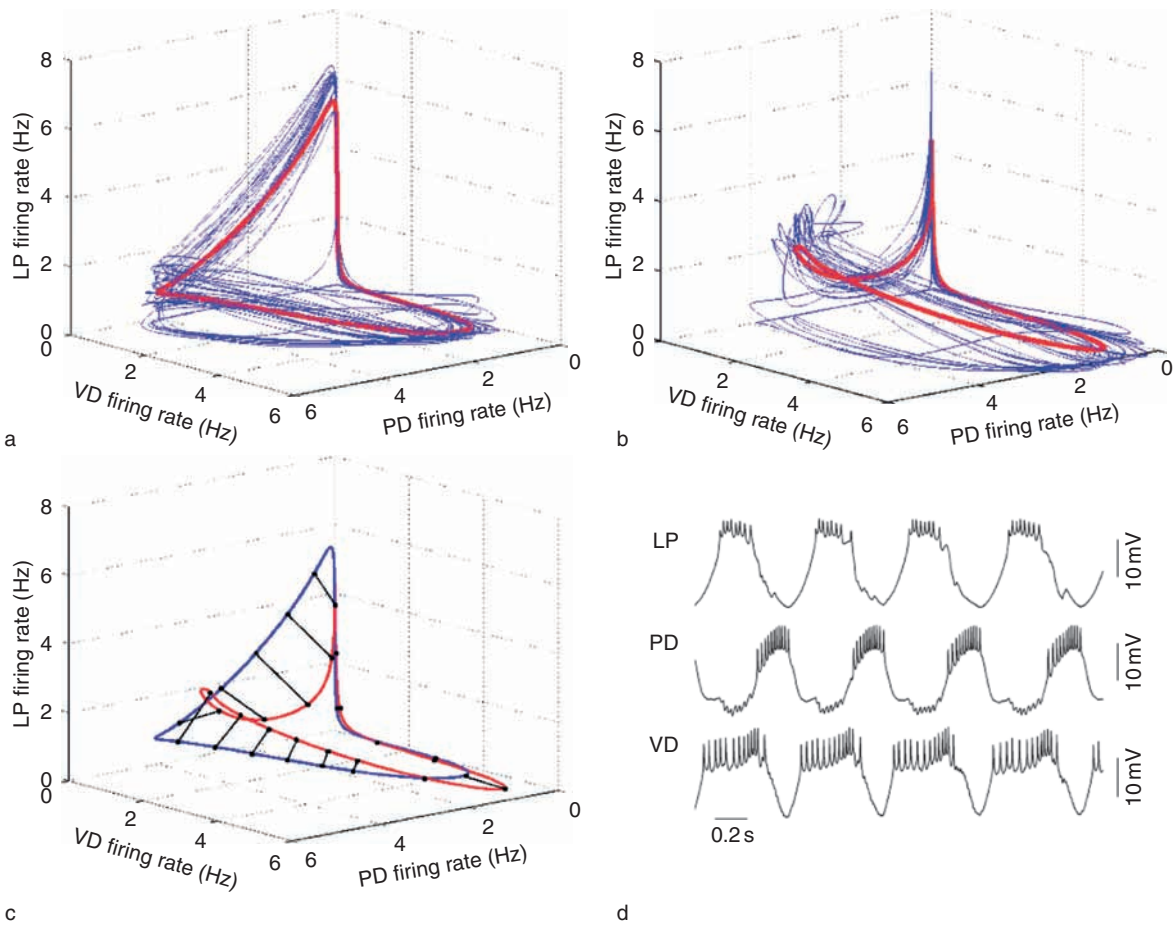

Figure 7 (a) Phase portrait of typical CPG output. The data were recorded in the pyloric CPG of the lobster stomatogastric ganglion. Each axis represents the firing rate of one of three pyloric neurons: LP, VD, and PD. The orbit of the oscillating pyloric system is shown in blue and the average orbit is shown in red. (b) The same as in a, but with a hyperpolarizing DC current injected into the PD. (c) The difference between the averaged orbits. (d) Time series of the membrane potentials of the three neurons.

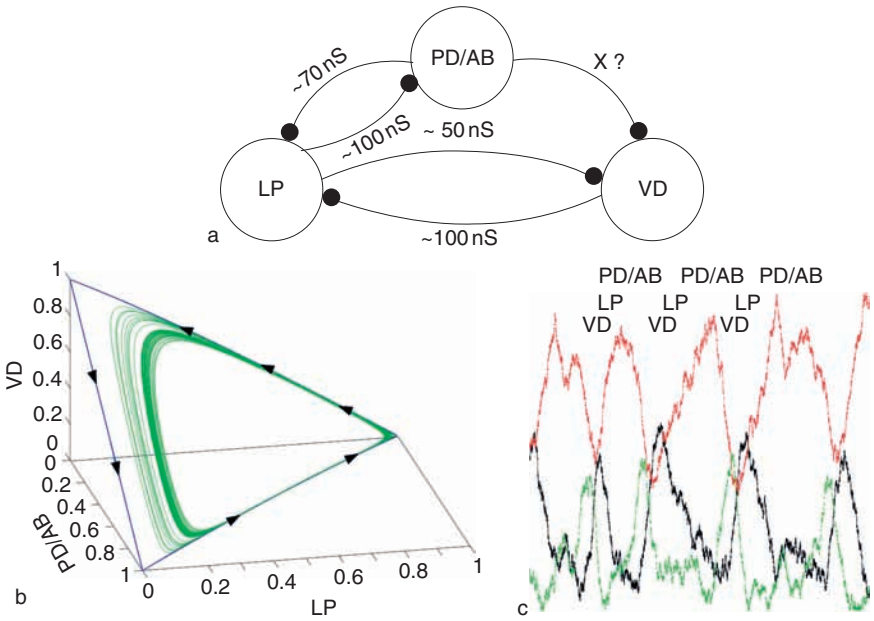

Figure 8 Minimal model of the pyloric CPG illustrates winnerless competition and the continuous sequencing of the bursts produced by the three-cell pyloric network. (a) Schematic diagram of the three-neuron network used for rate modeling. Black dots represent chemical inhibitory synapses with strengths given in nanoseconds. (b) Phase portrait of the model: the limit cycle corresponding to the rhythmic activity. (c) Robustness in the presence of noise: noise introduced into the model shows no effect on the order of activation for each type of neuron.

The dissipativity of the stomatogastric CPGs is guaranteed by the high percentage of inhibitory synaptic connections. Structural stability and sensitivity are maintained by the organization of the neural network. How nature solves the fundamental contradiction between robustness and sensitivity is a very difficult question. Nonlinear dynamics suggests this may be done by the coexistence of many attractors (multistability), which is good for sensitivity but bad for robustness: under the action of noise, the system will randomly move along the basins to different attractors. This means that each attractor has to be global, or at least have a big basin. In theory, however, it is possible to make a system with a single global heteroclinic attractor in such a way that if the informational input is changed, the topology of the attractor will be altered (i.e., new heteroclinic stable orbits will be formed depending on which neurons are affected by the inputs). The result is that new spatiotemporal patterns will be generated by the CPG for different signals (sensory, command, or neuromodulatory) that are external to the ganglion. A winnerless competition network is able to model these phenomena (Figure 8).

Summary

The stomatogastric nervous system and the two CPGs it contains have been modeled from the cellular to the network level. The primary goal has been to unify the enormous amount of reductionistic data that have been obtained for this preparation and use them to simulate the behavior of the preparation. In fact, it is not possible to handle such a large amount of data without some kind of modeling. It was originally hoped that by taking a modeling approach, some general principles of small network operation would be found that might extend to other small networks or possibility to complex microcircuit modules in the brain. To some extent, this has been achieved with phenomena first described in invertebrate preparations now used commonly to describe mechanisms in the nervous systems of higher animals. This is especially true for cellular mechanisms. Models for describing the operation of systems of neurons, however, are more likely to be idiosyncratic, and individualized models for these systems will probably be as close to an explanatory principle of CPG behavior as is possible to achieve. More general properties of CPGs, such as robustness and flexibility, appear to be more independent of specific topologies and may benefit from a dynamic analysis such as has been discussed here. Many of the questions posed at the beginning of this article have been answered for the STG. There are still many left unanswered, however, and these will be the direction of future experimental and modeling work on this amazing preparation.

See also: Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Hodgkin–Huxley Models;

Neural Oscillators and Dynamical Systems Models; Swim Oscillator Networks; Neural Integrator Models; Synaptic Transmission: Models.

Further Reading

Abarbanel HDI, Huerta R, Rabinovich M, Rulkov NF, Rowat PF, and Selverston AI (1996) Synchronized action of synaptically coupled chaotic model neurons. *Neural Computation* 8: 50–65.

Abbott L and Hooper SL (1991) Oscillating networks: Control of burst duration by electrically coupled neurons. *Neural Computation* 3: 487–497.

Abbott L and Marder E (1998) Modeling small networks. In: Koch C and Segev I (eds.) *Methods in Neuronal Modeling: From Ions to Networks*. Cambridge, MA: MIT Press.

Buchholtz F, Golowasch J, Epstein IR, and Marder E (1992) Mathematical model of an identified stomatogastric ganglion neuron. *Journal of Neurophysiology* 67: 332–340.

Denker M, Szucs A, Pinto RD, Abarbanel HDI, and Selverston AI (2005) A network of electronic neural oscillators reproduces the dynamics of the periodically forced pyloric pacemaker group. *IEEE Transactions on Biomedical Engineering* 52: 792–798.

Elson RC, Selverston AI, Huerta R, Rulkov NF, Rabinovich MI, and Abarbanel HDI (1998) Synchronous behavior of two coupled biological neurons. *Physical Review Letters* 81: 5692–5695.

Guckenheimer J, Gueron S, and Harris-Warrick RM (1993) Mapping the dynamics of a bursting neuron. *Philosophical Transactions: Biological Sciences* 341: 345–359.

Harris-Warrick RME, Marder AS, and Moulin M (eds.) (1992) *Dynamic Biological Networks: The Stomatogastric Nervous System*. Cambridge, MA: MIT Press.

Manor Y, Nadim F, Epstein S, Ritt J, Marder E, and Kopell N (1999) Network oscillations generated by balancing graded asymmetric reciprocal inhibition in passive neurons. *Journal of Neuroscience* 19: 2765–2779.

Pinto RD, Varona P, Volkovskii AR, Scucs A, Abarbanel HDI, and Rabinovich M (2000) Synchronous behavior of two coupled electronic neurons. *Physiological Reviews E* 62: 2644–2656.

Rabinovich M, Varona P, Selverston A, and Abarbanel HDI (2006) Dynamical principles in neuroscience. *Reviews of Modern Physics* 78(4): 1213.

Szucs A, Varona P, Volkovskii AR, Abarbanel HDI, Rabinovich MI, and Selverston AI (2000) Interacting biological and electronic neurons generate realistic oscillatory rhythms. *NeuroReport* 11: 563–569.

Strategic Control of Memory

B A Kuhl and A D Wagner, Stanford University, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Cognitive control mechanisms permit memory to be accessed strategically and so aid in bringing knowledge to mind that is relevant to current decisions and actions. A fundamental component of the strategic control of memory is the resolution of interference from competing, irrelevant representations. This article considers how the ventrolateral prefrontal cortex (VLPFC) regulates mnemonic competition in multiple memory systems. We initially discuss how damage to lateral prefrontal cortex impacts mnemonic function and then consider recent neuroimaging and focal lesion findings that highlight the distinct roles that subregions of the VLPFC play in the control of memory.

Lateral Prefrontal Cortex

The lateral prefrontal cortex (PFC) consists of ventral, dorsal, and frontopolar subregions. In this article, we primarily focus on the function of VLPFC. In the human (Figure 1(a)), VLPFC corresponds to the inferior frontal gyrus, which includes (moving caudally to rostrally) the inferior frontal pars opercularis (Brodmann area (BA) 44), inferior frontal pars triangularis (BA 45), and inferior frontal pars orbitalis (an area that Petrides and Pandya term area 47/12, which corresponds to the lateral portion of BA 47). Whereas Petrides and Pandya refer to area 47/12 and BA 45 collectively as the mid-VLPFC, distinguishing these regions from the caudally situated BA 44, in this article we functionally distinguish area 47/12 from BA 45. Thus, we use anterior VLPFC to refer to the inferior frontal pars orbitalis (area 47/12), mid-VLPFC to refer to the pars triangularis (BA 45), and posterior VLPFC to refer to the pars opercularis (BA 44). We note that the caudal portion of area 47/12 actually lies ventral (rather than rostral) to BA 45 (Figure 1(a)). The inferior frontal sulcus in humans (and the principal sulcus in monkeys) marks the approximate boundary between the VLPFC and the dorsolateral PFC (DLPFC). In terms of anatomical connectivity, the VLPFC is strongly connected with cortical areas in the lateral temporal lobe and medial temporal lobe (MTL), including (but not limited to) the inferotemporal cortex, superior temporal cortex, and, more medially, perirhinal and parahippocampal cortices.

Mnemonic Deficits Following Lateral Prefrontal Damage

Neuropsychological studies in humans and lesion studies in animals indicate that insult to the lateral PFC can produce memory impairments that are qualitatively distinct from those observed following MTL damage. Whereas MTL damage results in an amnesic condition that reflects the inability to encode and retrieve new declarative memories – long-term memories for events (episodic memory) and facts (semantic memory) – lateral PFC damage impairs the strategic regulation of multiple forms of memory, including declarative memory and working memory. Significantly, impairments in episodic memory, semantic memory, and working memory following lateral PFC damage are often most apparent when performance requires the resolution of interference, which led Moscovitch, Shimamura, and others to propose that lateral PFC subserves cognitive control mechanisms that regulate how we work with or dynamically filter memory.

Episodic Memory Deficits

Patients with lateral PFC damage show modest impairments in their ability to encode and retrieve episodic memories, with these deficits being particularly apparent when memory for target information is required in the face of distraction. Here we illustrate these deficits by highlighting a number of well-documented paradigms in which PFC patients are impaired.

First, PFC patients show disproportionate deficits on tests of free recall compared to item recognition. Theorists have argued that the unconstrained nature of free recall increases the likelihood of mnemonic competition or interference. Second, even within recognition tests, PFC patients show impairments when distracter items at test (foils) are similar to items from the study list. In other words, PFC patients suffer interference from items that are similar to those they studied, as evidenced by false claims of having studied these items. Third, in tests of source memory, in which memories for individual items must be attributed to a particular learning context – a discrimination that presents high interference because irrelevant contexts are often highly salient – PFC patients exhibit disproportionate deficits relative to tests of item recognition. Fourth, the use of AB–AC learning paradigms reveals that learning an AC association is disproportionately impaired in PFC patients (i.e., patients show a heightened sensitivity to proactive interference). In such paradigms, the patient's

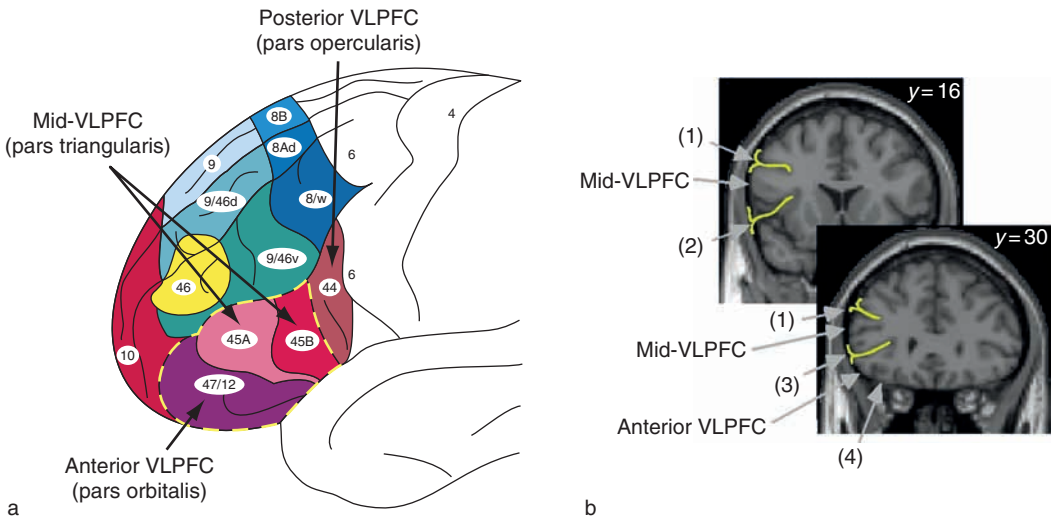

Figure 1 Anatomical divisions of the ventrolateral prefrontal cortex (VLPFC) in the human: (a) representation of the cytoarchitectonic subdivisions of the lateral prefrontal cortex (PFC); (b) coronal slices through the PFC depicting the anatomical boundaries that define the mid-VLPFC and anterior VLPFC. The subregions of the VLPFC include the pars orbitalis, pars triangularis, and pars opercularis, corresponding to the anterior VLPFC (area 47/12), mid-VLPFC (area 45), and posterior VLPFC (area 44), respectively. (Note that we use the term mid-VLPFC to refer to area 45 only, adopting anterior VLPFC to refer to area 47/12.) As shown in (b), both the anterior and mid-VLPFC lie ventral to the inferior frontal sulcus (1). In caudal slices, the mid-VLPFC is bounded ventrally by the insular sulcus (2), and in rostral slices it is bounded by the horizontal ramus of the lateral fissure (3). The anterior VLPFC is bounded ventrally and medially by the orbital gyrus (4). (a) Adapted from Petrides M and Pandya DN (2002) Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. *European Journal of Neuoscience* 16: 291–310. (b) From Badre D and Wagner AD (2007) Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* 45(13): 2883–2901.

memory for AC pairs is largely compromised by increased intrusions of the earlier learned AB associations. Moreover, even when sequentially studying two unrelated lists of items (i.e., the lists do not share common associates), PFC patients show disproportionate list 1 interference when trying to learn list 2. Finally, when PFC patients are provided with a subset of items from a previously studied list as retrieval cues (i.e., part-list cueing), the high salience of the provided items results in increased impairment in the patients' ability to recall the remaining (unpresented) items from the list. Collectively, these and other observations indicate that the ability to resolve interference in episodic memory is compromised following lateral PFC damage.

Semantic Memory Deficits

Just as PFC patients show impairments in unconstrained episodic memory tests, they also show deficits in unconstrained tests of semantic retrieval. For example, patients show reduced total output on verbal fluency tasks, which involve freely generating as many unique exemplars as possible in response to a semantic or orthographic cue. This impairment is thought to partially reflect patients' heightened sensitivity to output interference stemming from the initially retrieved exemplars. Significantly, this interference-dependent deficit during semantic or lexical retrieval stands in contrast to the typically unimpaired ability of PFC patients to recognize semantic structure or evaluate semantic relationships. Beyond verbal fluency tasks, patients with damage to the left lateral PFC fail to show normal semantic priming when the meaning of a semantic cue (the prime) is a context-dependent homograph, compared to words with less ambiguous meanings. This impairment may be due to PFC patients failing to retrieve the appropriate stimulus meaning when the prime is ambiguous, thereby preventing priming from occurring. Both these deficits are consistent with the perspective that the lateral PFC is recruited during controlled semantic retrieval, as well as when interference between competing semantic or lexical representations must be resolved.

Working Memory Deficits

In tests of working memory, PFC patients are often unimpaired, as measured by simple tests of verbal span. However, in tests in which maintenance is complicated by interference from irrelevant information, lateral PFC damage tends to result in impairment. For example, in delayed recognition tests, PFC patients are particularly sensitive to the presence of distracters during the delay. When compared to patients with damage to the MTL, PFC patients show impairments at all delay intervals as long as

distracters are present, whereas patients with MTL damage show impairments principally at long delays. A similar observation comes from studies of PFC-lesioned monkeys. Whereas it was initially believed that PFC-lesioned monkeys were unable to remember the location of a food item that was hidden for only a short delay, it was later shown that this apparent memory deficit could be eliminated if the delay period was held in total darkness. In other words, the monkey's memory deficit was a function of the interfering effects of visual input during the delay period. Strikingly, not only do PFC patients show impaired working memory performance due to distracters, but PFC patients also show exaggerated electrophysiological responses in cortical areas that process the modality of the distracting information, suggesting that when distraction is present, the PFC is responsible for gating activity in regions that would otherwise process the distracting information. Further evidence that lateral PFC damage results in an inability to filter or gate irrelevant distracting information comes from studies of negative priming. In such studies, healthy controls typically showed increased reaction times when processing a stimulus that was previously a distracter (i.e., these previous distracters showed negative priming). By contrast, frontal patients showed the reverse effect – facilitated processing of stimuli that previously served as distracters – which suggests that the patients had difficulty preventing the processing of the distracters.

Functional Specificity within Lateral Prefrontal Cortex

Although data from frontal patients provide compelling evidence that one contribution of the lateral PFC to the strategic control of memory is to regulate interference, the lack of anatomical specificity that typically accompanies naturally occurring PFC lesions in humans often precludes the determination of the specific functional contributions of particular PFC subregions. In an attempt to specify the mechanistic contributions of the subregions of the VLPFC to the control of memory, we consider in the next section evidence from functional neuroimaging studies and from recent reports of patients with lesions focused on subregions of VLPFC. As we discuss, extant data implicate the mid-VLPFC as a key component of the neural circuitry that resolves mnemonic interference, be it within working memory, semantic memory, or episodic memory.

Memory and the VLPFC

Focal lesion studies with monkeys have shown that damage to the inferior convexity (the monkey homolog of the human inferior frontal gyrus), but not to the DLPFC, causes perseverative tendencies in contexts in which the prior response learning must be reversed. Similar perseverative errors are a classic hallmark of human frontal lobe damage, as is often revealed through errors on the Wisconsin Card Sorting Task, when subjects must override a previously established, but now irrelevant, response set. As already noted, the strategic control of working memory, semantic memory, and episodic memory often requires conceptually similar mechanisms that permit the overriding of interference from irrelevant representations. Functional neuroimaging data in humans, complemented by focal lesion evidence, now indicate that the left mid-VLPFC regulates mnemonic interference by enabling the selection of relevant representations in the face of mnemonic conflict, whereas the left anterior VLPFC controls access to, and guides the retrieval of, long-term semantic knowledge. Response override or inhibition, by contrast, appears to differentially depend on regions in the human right VLPFC.

Left Mid-VLPFC and Interference in Working Memory

Extensive evidence for the role of the left mid-VLPFC (BA 45) in resolving mnemonic interference comes from a variant of the Sternberg working memory paradigm (Figure 2(a)). In this task, participants encode a target set of stimuli (e.g., four letters), which they then attempt to maintain in working memory across a brief delay. Following the delay, a probe (e.g., a letter) is presented and participants make a yes/no decision as to whether the probe is in the currently maintained memory set. Critically, some negative probes (i.e., probes to which the subject should respond no) come from the immediately preceding memory set (negative recent probes), whereas other negative probes have not appeared in either the present or the preceding memory sets (negative nonrecent probes) (Figure 2(a)). Given this structure, negative recent probes are associated with interference because subjects must attribute the familiarity of the probe to its having been in the preceding memory set to correctly reject the probe as not being a member of the current memory set. Thus, comparisons of negative recent to negative nonrecent trials provide leverage on the neural mechanisms that are engaged in the face of this mnemonic interference.

Jonides and colleagues were the first to demonstrate that functional activation in the left mid-VLPFC is greater during negative recent relative to negative nonrecent trials, a pattern that has been replicated and extended by others (Figure 2(b)). Subsequent data indicated that (1) the response in the left mid-VLPFC during negative recent trials is

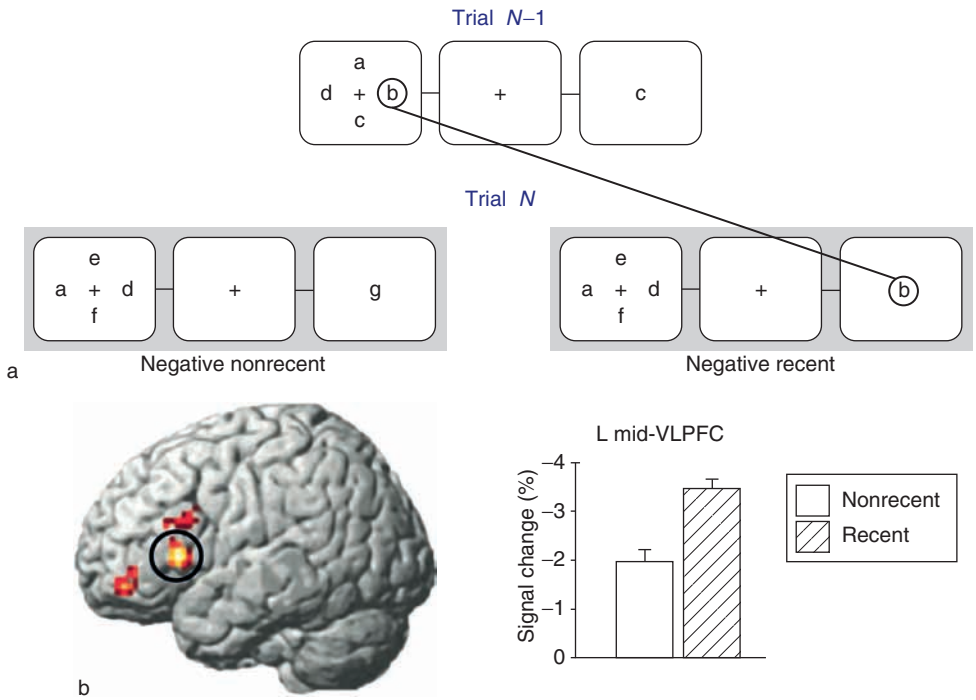

Figure 2 Interference resolution in the Sternberg working memory task: (a) a representative version of the task; (b) representative fMRI data. As shown in (a), during the task subjects maintain a set of four letters in working memory until a probe letter appears, at which point subjects indicate whether the probe is a member of the currently maintained set (positive probe) or is not a member of the current set (negative probe). Interference occurs when a probe is not a member of the current set, but is a member of the immediately preceding set (negative recent probe), relative to situations in which the probe is a member neither of the current set nor of the immediately preceding set (negative nonrecent probe). In (b), the fMRI data reveal greater activation during negative recent vs. negative nonrecent trials in the left mid-VLPFC (circled), reflecting that the presence of interference in working memory is accompanied by greater recruitment of the left mid-VLPFC. fMRI, functional magnetic resonance imaging; L, left; VLPFC, ventrolateral prefrontal cortex. Adapted from Badre D and Wagner AD (2005) Frontal lobe mechanisms that resolve proactive interference. *Cerebral Cortex* 15: 2003–2012.

restricted to the probe period of the trial, which is consistent with the fact that interference is a function of the probe's familiarity, and (2) when directly interrogating the DLPFC, a similar difference between negative recent and nonrecent trials is not typically observed, suggesting a specific role of the left mid-VLPFC in resolving interference in this task. Moreover, in a compelling study motivated by the aforementioned neuroimaging data, Thompson-Schill and colleagues showed that PFC patients with damage that spared the left VLPFC demonstrated interference effects that were comparable in magnitude to those in healthy controls, whereas one patient, R.C., with a focal lesion that damaged almost the entire extent of the left BA 45, exhibited an exacerbated interference effect despite relatively normal working memory performance when interference was minimal. Complementing these data, Postle and colleagues observed that when repetitive transcranial magnetic stimulation (TMS) – which transiently disrupts cortical function – was applied to the left VLPFC during the probe period in healthy humans there was a similar increase in susceptibility to interference to that seen in patient R.C. By contrast, TMS disruption of the primary motor, primary somatosensory, or supplemental motor area did not affect the ability to resolve interference in this task. Collectively, these data demonstrate that the left mid-VLPFC makes necessary contributions to resolving interference from currently irrelevant information in working memory. A leading hypothesis is that this region enables the selection of relevant active information over irrelevant active representations.

Left VLPFC and Declarative Memory Retrieval

Complementing the working memory literature implicating the left mid-VLPFC in resolving interference from recently activated but now irrelevant mnemonic representations are patient and imaging data that suggest that the left mid-VLPFC also supports selection during retrieval from declarative memory (i.e., semantic and episodic memory). For example, as already discussed, demands on interference resolution during semantic retrieval may be greater when available retrieval cues are unconstrained, and it is under these conditions that left PFC patients show impairments in generating semantic knowledge.

Such impairments may reflect an inability to select relevant representations from among competing alternatives.

Functional neuroimaging and focal lesion data converge on the left mid-VLPFC as being central for mediating the selection of relevant semantic representations. For example, in an influential functional magnetic resonance imaging (fMRI) study, Thompson-Schill and colleagues compared semantic processing under conditions that varied the demands placed on selection. Across several tasks, low-selection conditions were constructed such that they involved retrieving dominant, or prepotent, semantic information, whereas high-selection conditions involved retrieving nondominant semantic information from among competing representations. Significantly, the imaging data revealed that the left VLPFC (BAs 44 and 45) was consistently more active during the high-selection (vs. low-selection) conditions, demonstrating that this is a replicable characteristic of left mid-and/or posterior VLPFC function. Subsequent neuropsychological work by Thompson-Schill and colleagues revealed that it is the proportion of damage to this left VLPFC region, but not the gross lesion size, that strongly predicts the magnitude of behavioral impairment on high-selection semantic retrieval tasks.

Related data have increased the precision of our understanding of left mid-VLPFC function and its relation to the functions of the surrounding anterior and posterior VLPFC subregions. In particular, Badre and colleagues demonstrated that selection demands during semantic processing tasks are specifically associated with the same left mid-VLPFC subregion (BA 45; pars triangularis) that resolves interference in working memory tasks (**Figure 3(a)**). Moreover, these researchers observed a functional dissociation within the left VLPFC, indicating that the left mid-VLPFC supports selection from among active representations, whereas the more rostrally and ventrally situated left anterior VLPFC subregion (area 47/12; pars orbitalis) controls the retrieval of semantic representations stored in the lateral temporal cortical areas. This functional distinction between the left mid-VLPFC and anterior VLPFC appears replicable and generalizeable; Gold and colleagues reported a similar functional pattern using a lexical decision task – namely that the left anterior VLPFC was associated with controlled semantic retrieval and the left mid-VLPFC was associated with resolving interference from irrelevant active representations.

Evidence for functional segregation within the left VLPFC has also come from studies comparing semantic versus phonological control. For example, neuroimaging studies of phonological rehearsal and phonological analysis of stimuli have implicated the left posterior VLPFC (BA 44; pars opercularis), suggesting that this VLPFC subregion is critical for representing and maintaining phonological codes. On the other hand, neuroimaging studies of semantic retrieval and analysis of stimuli have implicated the left anterior VLPFC. Significantly, these dissociations between the left anterior and posterior VLPFC have been observed in studies that directly contrasted tasks that differentially depend on semantic and phonological control. Moreover, Devlin and colleagues have shown that TMS to the left anterior VLPFC differentially disrupts concurrent semantic processing, whereas TMS to the left posterior VLPFC differentially disrupts concurrent phonological processing, demonstrating that these subregions are necessary for distinct forms of control.

Functional distinctions between the left VLPFC subregions are also apparent during tasks that probe episodic memory. For example, Dobbins and colleagues observed that the left mid-VLPFC is associated with selecting between episodic details in the course of making a source judgment (**Figure 3(b)**), whereas the left anterior VLPFC is associated with semantically elaborating on the cues used to probe episodic memory. Further, during episodic encoding, Dolan, Fletcher, and colleagues demonstrated that left mid-VLPFC activity increases when prior learning interferes with to-be-encoded information (i.e., when the resolution of proactive interference is required). Collectively, these data suggest that the left mid-VLPFC contributes to the resolution of interference during all forms of declarative memory, whereas the left anterior VLPFC controls the retrieval of semantic knowledge that is not retrieved in an automatic (bottom-up) manner. Moreover, when considered along with the evidence that the left mid-VLPFC resolves interference within working memory, extant data favor the hypothesis that the left mid-VLPFC serves a domain-general role of resolving mnemonic interference.

Right VLPFC and the Control of Memory

We have thus far focused on how the left VLPFC contributes to interference resolution and other forms of cognitive control; however, it should be emphasized that the right VLPFC is sometimes coactive with the left VLPFC during tasks that require overcoming mnemonic conflict. Nevertheless, although the right VLPFC may support control functions that are conceptually analogous to those supported by the left VLPFC, it appears that (in the human) the right VLPFC at least partially differs from the left VLPFC in the domain of knowledge on which is operates. For example, Aron and colleagues, among others,

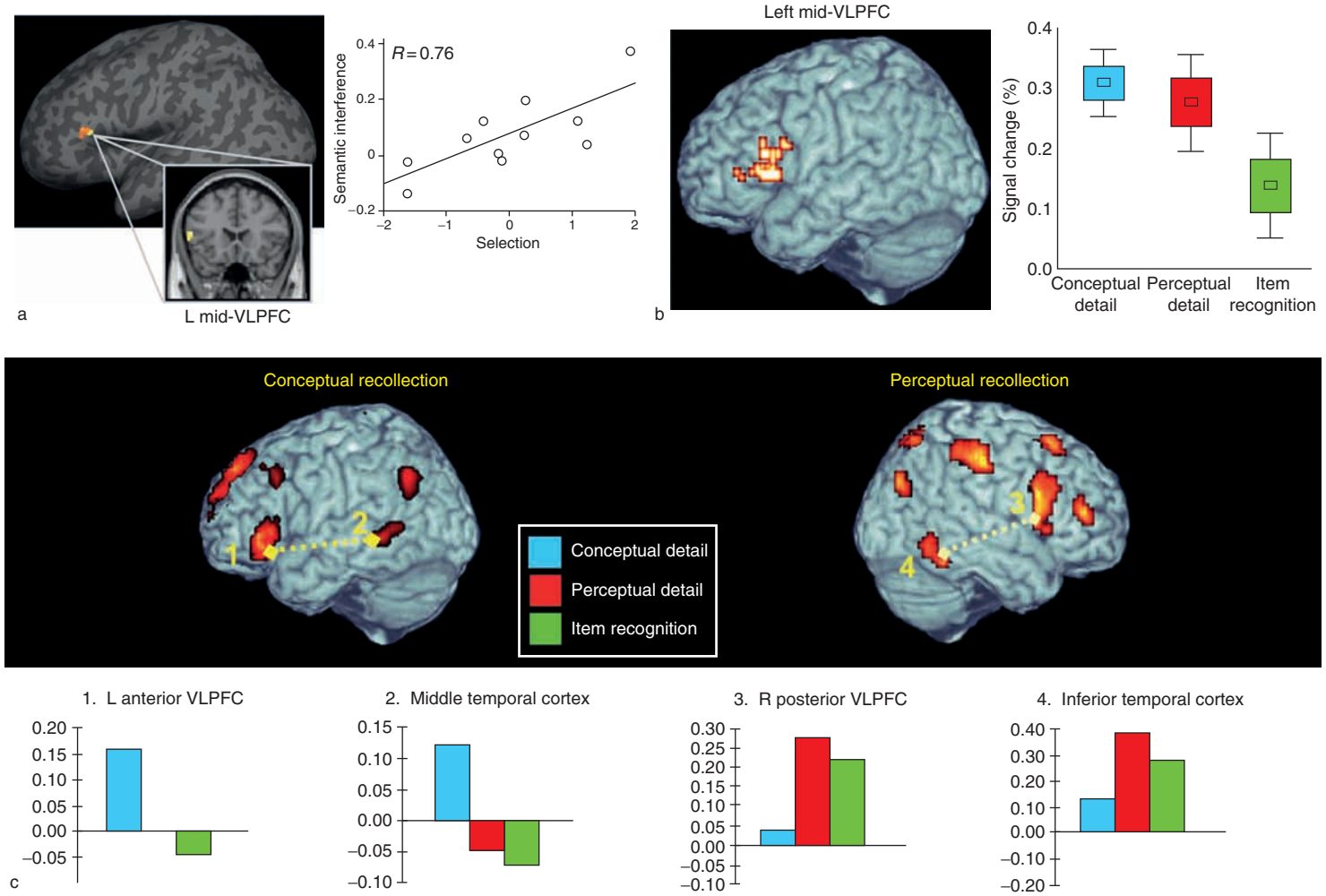

Figure 3 PFC contributions to the control of declarative memory: (a) a behavioral selection measure derived from performance during various semantic retrieval tasks; (b) the left mid-VLPFC during episodic retrieval; (c) the left anterior and right posterior VLPFC during episodic retrieval. In (a), the selection measure was exclusively associated with the magnitude of activation in the left mid-VLPFC (left). For example, this selection measure was tightly correlated with the magnitude of interference-related activation in the left mid-VLPFC during semantic retrieval (right). As shown in (b), during episodic retrieval, the left mid-VLPFC was differentially engaged when source details (conceptual or perceptual) were selectively retrieved, compared to simple item recognition. As shown in (c), during episodic retrieval, the left anterior VLPFC (1) was selectively engaged during recollection of conceptual source details, functionally coupling with the middle temporal cortex (2). By contrast, the right posterior VLPFC (3) was engaged during perceptual recollection and during recognition of familiar vs. novel objects, functionally coupling with the inferior temporal cortex (4). L, left; PFC, prefrontal cortex; VLPFC, ventrolateral prefrontal cortex. (a) Adapted from Badre D and Wagner AD (2005) Frontal lobe mechanisms that resolve proactive interference. *Cerebral Cortex* 15: 2003–2012. (b, c) Adapted from Dobbins IG and Wagner AD (2005) Domain-general and domain-sensitive prefrontal mechanisms for recollecting events and detecting novelty. *Cerebral Cortex* 15: 1768–1778.

reported that functional activation in the posterior aspect of the right VLPFC is consistently implicated in situations in which prepotent motor responses must be inhibited or when well-learned stimulus–response contingencies must be reconfigured. Moreover, these researchers demonstrated that when the right VLPFC is damaged, subjects display increased susceptibility to interference at the response level, highlighting the necessary contribution of this VLPFC subregion to response inhibition.

By contrast, other data indicate that the right VLPFC may contribute to the orienting of visual attention toward task-relevant object representations (and away from distracters). Significantly, with respect to the control of memory, this latter role of the right VLPFC in orienting to visuo-object representations has been observed in a number of mnemonic contexts, including discriminating novel from encountered visual objects, encoding faces and complex visual scenes into episodic memory, recollecting object–object associations, and recollecting perceptual details about previously encountered objects (Figure 3(c)). Thus, the extent to which the right VLPFC (as opposed to the left VLPFC) is engaged at least partially depends on the modality of the mnemonic representations attended, with the right VLPFC preferentially engaged during visuospatial processing and the left VLPFC engaged during phonological and semantic processing (Figure 3(c)). Significantly, although the right VLPFC has been implicated in visuo-object attention and in resolving competition at the response level, functional understanding of the subregions within the right VLPFC is not as well advanced as that of the homologous structures in the left VLPFC.

Concluding Comments

Interference resolution is a fundamental aspect of efficient mnemonic processing and may represent a principal way in which cognitive control interacts with memory. This interaction is critically enabled by control mechanisms that depend on the lateral PFC for their operation. This article focuses on the delineation of the mechanisms supported by the VLPFC, but it should be noted that the VLPFC must dynamically interact with the posterior neocortical association areas and with the MTL to implement the control of memory. Further, although VLPFC functions are clearly important for the resolution of mnemonic interference, the VLPFC also interacts with the DLPFC and frontopolar cortical areas that subserve other forms of cognitive control. Next we briefly illustrate each of these points.

To regulate memory-relevant processes, PFC control mechanisms must interact with systems that support or store long-term mnemonic representations. One line of evidence for such frontal-posterior interactions comes in the form of coactivations within functional neuroimaging studies. For example, fMRI studies of episodic remembering indicate that the left anterior VLPFC coactivates with the left middle temporal cortical areas that represent conceptual knowledge during controlled attempts to recollect such details about the past (Figure 3(c)). Analogously, the right VLPFC coactives with bilateral occipito-temporal areas that represent visuo-object form during attempts to recollect such details (Figure 3(c)). Moreover, as previously alluded to, lesions to the lateral PFC result in the failure to regulate processing in such posterior neocortical areas, demonstrating that the lateral PFC is necessary for gating the processing of distracting or interfering stimuli. Collectively, these data illustrate the top-down role of PFC control mechanisms in regulating perception, memory, and action.

With respect to cross-regional interactions within the PFC, it is important to emphasize that other structures in the DLPFC and frontopolar cortex are also often engaged in situations in which control must be implemented to accomplish a mnemonic goal. For example, DLPFC activation is frequently observed during complex working memory tasks, during attempts to remember past events, and during many tasks that require response selection or the resolution of response conflict. Although the specific nature of interactions between the DLPFC and VLPFC have yet to be well characterized, it is hypothesized that the DLPFC may operate at a higher stage in the processing hierarchy relative to the VLPFC, serving to perform operations on the products of VLPFC processing.

Thus, it is clear that the strategic control of memory is multifaceted. Future research promises to further illuminate how cognitive control emerges through lateral PFC function, allowing us to strategically wrest control of our memories, bringing them in line with current goals.

See also: Cognition: An Overview of Neuroimaging Techniques; Event-Related Potentials (ERPs); Inhibitory Control over Action and Memory; Neuroimaging; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Short Term and Working Memory; Spatial Cognition and Executive Function; Working Memory: Capacity Limitations.

Further Reading

Aron AR, Robbins TW, and Poldrack RA (2004) Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences* 8: 170–177.

Badre D, Poldrack RA, Paré-Blagoev EJ, Insler RZ, and Wagner AD (2005) Dissociable controlled retrieval and generalized selection

mechanisms in ventrolateral prefrontal cortex. *Neuron* 47: 907–918.

Badre D and Wagner AD (2005) Frontal lobe mechanisms that resolve proactive interference. *Cerebral Cortex* 15: 2003–2012.

Badre D and Wagner AD (2007) Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* 45(13): 2883–2901.

Dobbins IG and Wagner AD (2005) Domain-general and domain-sensitive prefrontal mechanisms for recollecting events and detecting novelty. *Cerebral Cortex* 15: 1768–1778.

Dolan RJ and Fletcher PC (1997) Dissociating prefrontal and hippocampal function in episodic memory encoding. *Nature* 388: 582–585.

Feredoes E, Tononi G, and Postle BR (2006) Direct evidence for a prefrontal contribution to the control of proactive interference in verbal working memory. *Proceedings of the National Academy of Sciences of the United States of America* 103: 19530–19534.

Gold BT, Balota DA, Jones SJ, Powell DK, Smith CD, and Anderson AH (2006) Dissociation of automatic and strategic lexical-semantics: Functional magnetic resonance imaging evidence for differing roles of multiple frontotemporal regions. *Journal of Neuroscience* 26: 6523–6532.

Gough PM, Nobre AC, and Devlin JT (2005) Dissociating linguistic processes in the left inferior frontal cortex with transcranial magnetic stimulation. *Journal of Neuroscience* 25: 8010–8016.

Henson RN, Shallice T, Josephs O, and Dolan RJ (2002) Functional magnetic resonance imaging of proactive interference during spoken cued recall. *Neuroimage* 17: 543–558.

Iversen SD and Mishkin M (1970) Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity. *Experimental Brain Research* 11: 376–386.

Jonides J, Smith EE, Marshuetz C, Koeppe RA, and Reuter-Lorenz PA (1998) Inhibition in verbal working memory revealed by brain activation. *Proceedings of the National Academy of Sciences of the United States of America* 95: 8410–8413.

Martin RC and Cheng Y (2006) Selection demands versus association strength in the verb generation task. *Psychonomic Bulletin & Review* 13: 396–401.

Moscovitch M and Melo B (1997) Strategic retrieval and the frontal lobes: Evidence from confabulation and amnesia. *Neuropsychologia* 35: 1017–1034.

Petrides M and Pandya DN (2002) Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. *European Journal of Neuoscience* 16: 291–310.

Shimamura AP (2000) The role of the prefrontal cortex in dynamic filtering. *Psychobiology* 28: 207–218.

Shimamura AP, Jurica PJ, Mangels JA, Gershberg FB, and Knight RT (1995) Susceptibility to memory interference effects following frontal lobe damage: findings from tests of paired-associate learning. *Journal of Cognitive Neuroscience* 7: 144–152.

Thompson-Schill SL, D'Esposito M, Aguirre GK, and Farah MJ (1997) Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences of the United States of America* 94: 14792–14797.

Thompson-Schill SL, Jonides J, Marshuetz C, et al. (2002) Effects of frontal lobe damage on interference effects in working memory. *Cognitive, Affective, & Behavioral Neuroscience* 2: 109–120.

Thompson-Schill SL, Swick D, Farah MJ, D'Esposito M, Kan IP, and Knight RT (1998) Verb generation in patients with focal frontal lesions: A neuropsychological test of neuroimaging findings. *Proceedings of the National Academy of Sciences of the United States of America* 95: 15855–15860.

Stress and Cognition

S J Lupien, Douglas Hospital Research Center/McGill University, Verdun, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Prior to becoming part of our day-to-day conversations, the term 'stress' was used by engineers to explain forces that can put strain on a structure. For example, one could place strain on a piece of metal in such a way that it would break like glass when it reached its stress level. In 1936, Hans Selye borrowed the term of stress from the field of engineering and talked about stress as being a nonspecific phenomenon representing the intersection of symptoms produced by a wide variety of noxious agents. For many years, Selye tested various conditions (e.g., fasting, extreme cold, operative injuries, and drug administration) that would produce morphological changes in the body that were representative of a stress response, such as enlargement of the adrenal gland, atrophy of the thymus, and gastric ulceration. Selye's view of the concept of stress was that the determinants of the stress response are nonspecific, that is, many unspecific conditions can put strain on the organism and lead to disease, the same way that many unspecific conditions can put strain on a piece of metal and break it like glass.

Not all researchers agreed with Selye's model, particularly with the notion that the determinants of the stress response are nonspecific. The reason for this was simple. While Selye spent his entire career working on physical stressors (e.g., heat, cold, and pain), we all know that some of the worst stressors we encounter in life are psychological in nature, and are induced by our interpretation of events. For this reason, a psychologist named John Mason spent many years evaluating stress hormone levels in humans subjected to various conditions that he thought would be stressful, in order to describe the psychological characteristics that would make any condition stressful, to anyone exposed to it.

By summarizing the results of studies measuring the circulating levels of these hormones before and after individuals were exposed to various jobs or situations that were deemed to be stressful (e.g., air-traffic controllers or parachute jumping), Mason was able to describe three main psychological determinants that would induce a stress response in any individual exposed to them. Using this methodology, he showed that in order for a situation to induce a stress response from the body, it has to be interpreted as being novel, and/or unpredictable, and/or the individual must have the feeling that he/she does not have control over the situation. Although this work led to a general debate between Selye and Mason, further studies confirmed that the determinants of the stress response are highly specific, and therefore, potentially predictable and measurable. More recently, a meta-analysis of the determinants of psychological stress in human added the component of social evaluative threat as a fourth characteristic of a situation that is a potent inducer of a physiological stress response in humans.

The Relativity of Stress

From this short historical background, we can now define stress as being a threat, real or implied, to homeostasis. In this sense, stress can be absolute (a real threat induced by an earthquake in a town, leading to a significant stress response in every person facing this threat), or it can be relative (an implied threat induced by the interpretation of a situation as being novel, and/or unpredictable and/or uncontrollable – for example, a public speaking task).

Absolute stressors serve adaptive purposes and are those events or situations that will necessarily lead to a stress response in the majority (if not the totality) of individuals when they are first confronted with it. Being in or witnessing an accident, confronting a dangerous animal, and being submitted to extreme cold or heat are all examples of absolute stressors. These extreme and particular situations constitute absolute stressors in that, due to their aversive nature, a stress response has to be elicited for one's survival and/or well-being. In our western societies, absolute stressors are rare, but are nonetheless those that elicit the greatest physiological response and most vivid memories. Conversely, relative stressors are those events or situations that will elicit a stress response only in a certain proportion of individuals. Moreover, this response may be mild or large. For example, having to unexpectedly deliver a speech may be very stressful for a given individual, and not at all for another. Consequently, large inter-individual variations in the stress response to psychological challenges have been frequently reported.

The stressor is the event itself, such as the earthquake in the case of the absolute stress, or the public speech in the case of the relative stress. The stress response is the body's reaction to the event, and it is this body's response to stress that is the foundation for the studies that determine the impact of stress on cognitive function. The reason for this is that the

stress hormones that are secreted in response to an absolute or relative stress, and particularly the glucocorticoids, are steroids that can easily cross the blood–brain barrier and access the brain, where they can influence learning and memory by binding to glucocorticoid receptors (GRs) localized in various brain regions known to be involved in learning and memory.

Stress Hormones

When a situation is interpreted as being stressful, it triggers the activation of the hypothalamic–pituitary–adrenal (HPA) axis whereby neurons in the hypothalamus, a brain structure often termed the 'master gland,' releases a hormone called corticotropin-releasing factor (CRF). The release of CRF triggers the subsequent secretion and release of another hormone called adrenocorticotropin (ACTH) from the pituitary gland, also located in the brain. When ACTH is secreted by the pituitary gland, it travels in the blood and reaches the adrenal glands, which are located above the kidneys, and triggers secretion of the so-called stress hormones. There are two main stress hormones, the glucocorticoids (called corticosterone in animals, and cortisol in humans) and the catecholamines (epinephrine and norepinephrine). The acute secretion of glucocorticoids and catecholamines in response to a stressor constitutes the primary mediators in the chain of hormonal events triggered in response to stress. When these two hormones are secreted in response to stress, they act on the body to give rise to the fight-or-flight response whereby one would, for instance, experience an increase in heart rate and blood pressure.

Glucocorticoids have a variety of different effects in target systems throughout the organism, which can be summarized as aiming to increase the availability of energy substrates in different parts of the body, and allow for optimal adaptation to changing demands of the environment. While the activation of the HPA axis can be regarded as a basic adaptive mechanism in response to change, prolonged activation of this system presents a health risk to the organism. The highly catabolic glucocorticoids antagonize insulin and increase blood pressure, thus increasing the risk for developing diabetes, hypertension, and arterial disease. Also, growth and tissue repair are impaired. Furthermore, activation of the HPA axis suppresses immune functions, which in a chronic state can be considered harmful for the organism, since it is associated with increased risk of infection.

Given their liposoluble characteristics, the glucocorticoids can easily cross the blood–brain barrier and access the brain where they bind to receptors. Three of the most important brain areas containing GRs are the hippocampus, amygdala, and frontal lobes, which are brain structures known to be involved in learning and memory. Although epinephrine does not readily access the brain due to absence of lipophilic properties, it can still act on the brain through its action on the sensory vagus outside of the blood–brain barrier, with information transmitted into the brain via the nucleus of the solitary tract. The most important brain area containing adrenergic receptors is the amygdala, which has been shown to play an important role in fear processing and memory for emotionally relevant information.

Important Characteristics of Glucocorticoids

Under basal conditions, glucocorticoid secretion exhibits a 24-h circadian profile in which glucocorticoid concentrations present a morning maximum in humans (the circadian peak), and slowly declining levels in the late afternoon, evening, and nocturnal period (the circadian trough), and an abrupt elevation after the first few hours of sleep. Circulating glucocorticoids bind with high affinity to two receptor subtypes, the mineralocorticoid receptor (MR or type I) and GR or type II.

Although both receptor types have been implicated in mediating glucocorticoid feedback effects, there are two major differences between MR and GR receptors. First, MRs bind glucocorticoids with an affinity that is about 6–10 times higher than that of GRs. This differential affinity results in a striking difference in occupation of the two receptor types under different conditions and time of day. Thus, during the circadian trough (the p.m. phase in humans and the a.m. phase in rats), the endogenous hormone occupies more than 90% of MRs, but only 10% of GRs. However, during stress and/or the circadian peak of glucocorticoid secretion (the a.m. phase in humans and the p.m. phase in rats), MRs are saturated, and there is occupation of approximately 67–74% of GRs.

The second major difference between these two receptor types is related to their distribution in the brain. The MR is exclusively present in the limbic system, with a preferential distribution in the hippocampus, parahippocampal gyrus, entorhinal, and insular cortices. On the contrary, the GR is present in both subcortical (paraventricular nucleus and other hypothalamic nuclei, the hippocampus, and parahippocampal gyrus) and cortical structures, with a preferential distribution in the prefrontal cortex. As we will see in the following sections, the impact of glucocorticoids on cognitive function can be best understood in terms of the differential effects of MR and GR activation.

Positive Effects of the Stress Response on Cognition

Because of their actions on brain structures known to be involved in fear detection and memory for emotionally relevant information, the stress hormones enhance the formation of the so-called 'flashbulb memories' of events related with strong emotions that are usually associated with an absolute stress. This process involves the amygdala, and the pathway for encoding these memories involves the interaction between neurotransmitters in the amygdala and in related brain areas such as the hippocampus, along with circulating levels of catecholamines and glucocorticoids. The importance of the stress hormones on memory for emotionally relevant information has been recently confirmed by studies in which blockade of either glucocorticoid or norepinephrine activity impaired the recall of emotionally relevant information. Consequently, secretion of these primary stress mediators is necessary for the adequate encoding of emotionally relevant information during an absolute stress. This enhancement of memory for stimuli inducing stressful and/or emotional responses may be essential for species' survival.

At the same time as the brain encodes information and controls the behavioral responses to stress, it is also changed structurally and chemically by those experiences. Although for a long time it was thought that the brain of adult mammals does not generate new nerve cells, more recent evidence showed that neurons are born in the adult mammalian brain. Interestingly, in the adult brain, the generation of new neurons (called neurogenesis) occurs in two regions. The first region is the subventricular zone in the wall of the lateral ventricle, where new interneurons are generated for the olfactory bulb, and the second region is the subgranular zone of the dentate gyrus of the hippocampus, which gives rise to the granule cells. Recent evidence show that adult neurogenesis in the dentate gyrus of the hippocampus is a feature of all mammalian species, occurring in rats, mice, tree shrews, marmosets, macaques, and humans. Although the functional significance of hippocampal neurogenesis has been questioned in the past, a study reported that neurogenesis in the adult is involved in the formation of trace memories, suggesting that the new neurons actually contribute to the function of the adult brain.

Comparative studies of different species show that hippocampal volume is increased in mammalian and avian species that depend critically on spatial memory for survival. Examples of such critical memory skills involve home-range navigation, migration, brood parasitism, and memory-based cache recovery in birds that hide food. This association between performance of spatial tasks and hippocampal size can be explained in terms of the impact of environmental demands (stress) on hippocampal volume, rather than the converse. For example, it has been reported that the hippocampus of titmice and chickadees increases in volume in association with the experience of storing and recovering food cache. This result shows that in animals, there can be an experience-dependent hippocampal growth that occurs at a relatively late stage in development. Similar results of experience-based changes in hippocampal volumes have been reported in humans. Maguire and collaborators tested the hypothesis that the ability of taxi cab drivers to navigate to a specific destination correlates with hippocampal volume. They showed that compared to age-matched controls, the taxi drivers had larger posterior hippocampi. Although these results surely do not confirm that the experience of driving a taxi in the complex streets of London has a significant effect on hippocampal volumes, they nonetheless raise the intriguing possibility of hippocampal plasticity in response to environmental demands.

Negative Effects of the Stress Response on Cognition

While short-term responses of the brain to novel and potentially threatening situations may be adaptive and result in new learning and acquired behavioral strategies for coping, as may be the case for absolute stress, repeated stress can cause both cognitive impairments and structural changes in the hippocampus. These processes may be occurring somewhat independently of each other and contribute in various degrees to different pathophysiological situations involving traumatic stress, depression, or aging.

The cognitive effects of long-term elevations of glucocorticoids in human populations have been studied in disorders affecting glucocorticoid levels and using exogenous administration of the synthetic compound to healthy subjects. Mental disturbances mimicking mild dementia (such as decrements in simple and complex attentional tasks, verbal and visual memory, encoding, storage and retrieval) have been described in depressed patients with hypercortisolism, and in steroid psychosis following glucocorticoid treatment. Similar cognitive deficits are also reported in patients suffering from Cushing's disease, a medical condition in which endogenous levels of glucocorticoids are chronically elevated. During human aging, a significant proportion of elderly individuals present an endogenous increase of glucocorticoid levels, and this increase has been related to impaired memory performance.

Chronic exposure to stress has also been shown to be associated with structural changes of the hippocampus. For example, treatment of adult rats with corticosterone decreases the proliferation of granule cell precursors, while removal of adrenal steroids stimulates the generation of new granule neurons. Given the suppressive actions of glucocorticoids on hippocampal neurogenesis, stressful experiences have been suggested to inhibit cell proliferation in adulthood, a hypothesis that has been demonstrated. Long-term stressful experience decreases the number of adult-generated neurons in the dentate gyrus in various species, including the rat, tree shrew, and marmoset. In a similar vein, repeated stress also produces prolonged suppression of cell proliferation in the dentate gyrus of the adult tree shrew. Stress-induced decreases in dentate gyrus cell proliferation could also contribute, in line with atrophy of the pyramidal cells of the hippocampus, to changes in hippocampal volumes observed after chronic exposure to stress.

Smaller hippocampal volumes associated with chronic elevations of glucocorticoids have also been reported in humans. Over the past two decades, many studies revealed the presence of smaller hippocampal volumes in various psychiatric disorders such as depression, posttraumatic stress disorder, and schizophrenia, all of which involve glucocorticoid level dysregulations. Hippocampal atrophy associated with chronic exposure to high levels of glucocorticoids is also reported in Cushing's patients and in elderly individuals. A recent study of Cushing's patients revealed that surgical treatment of hypercortisolism in these patients leads to a reversal of the glucocorticoid-induced hippocampal atrophy reported to occur in this population, with an average volume increase of 3.2%, and variations up to 10% in some patients. This is a significant finding since it implies the possibility of functional reorganization of the hippocampus, once the chronic stress has been taken away. Reversibility and/or preventability of such atrophy are major topics for future research, as is the implication of such treatment for cognitive function.

Positive Effects Meeting Negative Effects: The Inverted-U Shape Function

There is a great paradox in the field of stress research that relates to the fact that stress hormones are sometimes associated with positive effects and sometimes associated with negative effects on cognition. This apparent discrepancy in the data can be explained by the fact that in both animals and humans, many studies reveal the presence of an inverted-U shape function between circulating stress hormone levels and memory performance, so that up to an optimal level, glucocorticoids can enhance cognitive performance, while above this optimal level, glucocorticoids impair cognitive performance. A model has been developed and suggests the presence of an inverted-U shape function between circulating levels of glucocorticoids and cognitive performance (**Figure 1**). More importantly, this model is based on the two types of GRs found in the brain.

Remember that given their differential affinity for glucocorticoids, the MRs will always be almost saturated (100% activated) before the GRs can start to be activated. During periods of stress or high levels of glucocorticoids (the a.m. phase in humans), both MRs and GRs will thus be saturated. Many studies performed in rodents have reported that the ratio of MR/GR occupation is a major determinant of the direction of glucocorticoid-induced cognitive changes. For example, long-term potentiation (LTP), a proposed neurobiological substrate of memory formation, has been shown to be optimal when glucocorticoid levels are mildly elevated, that is, when the ratio of MR/GR occupation is high. In contrast, significant decreases in LTP are observed after adrenalectomy, when MR occupancy is very low, or after exogenous administration of synthetic glucocorticoids, which activate GRs and deplete cortisol, again resulting in low occupancy of MRs.

Some authors have reinterpreted the well-known inverted-U shape function between circulating levels of glucocorticoids and cognitive performance, in line with the MR/GR ratio hypothesis. In this view, cognitive function can be enhanced when most of the MRs and only part of the GRs are activated (top of the inverted-U shape function; increased MR/GR ratio). However, when circulating levels of glucocorticoids are significantly decreased or increased (extremes of the inverted-U shape function; low MR/GR ratio), cognitive impairments will result. The authors suggested that the negative view of glucocorticoid actions on human cognitive function could be partly explained by limitations in previous human experimental designs, which did not allow differential manipulation of MR and GR levels. In order to do this, such studies should measure cognitive function when glucocorticoid occupancy is decreased (rather than increased), thus allowing functional measures of MR/GR occupancy on learning and memory.

Recent studies have tested this hypothesis and confirmed it. A first study measured whether memory performance in young, normal adults can be modulated by a hormone removal–replacement protocol in which one pharmacologically manipulates circulating levels of cortisol and measures subsequent memory performance. In this protocol, the authors used a

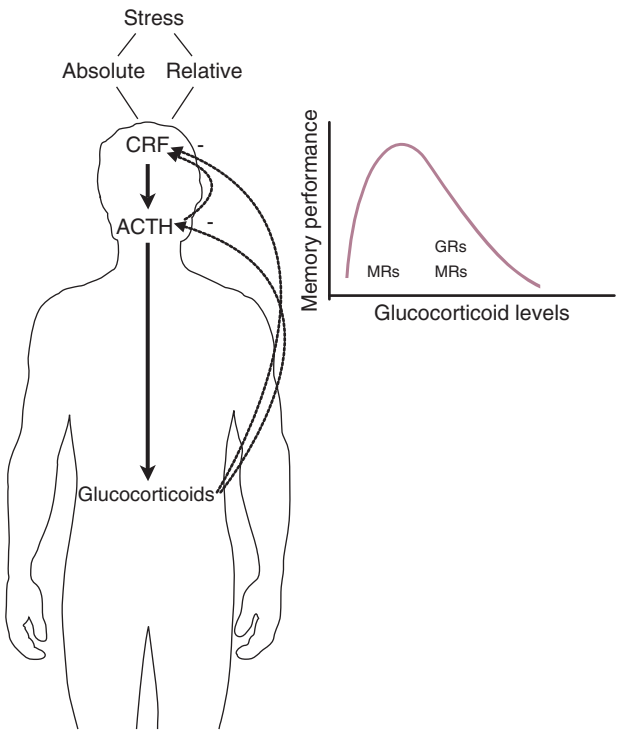

Figure 1 Schematic representation of the stress response and its impact on cognition. First, perception of an absolute or relative stressor triggers the secretion of corticotropin-releasing factor (CRF) by the hypothalamus, which then acts on the pituitary gland to induce secretion of adrenocorticotropin (ACTH). ACTH travels in the bloodstream and reaches the adrenal gland located on top of the kidneys and induces secretion of glucocorticoids. Because of their liposoluble properties, glucocorticoids easily cross the blood–brain barrier where they act on the hypothalamus and pituitary gland to shut down further secretion of CRF and ACTH, and they also access various brain regions where two types of glucocorticoid receptors (GRs) are located. The mineralocorticoid receptor (MR) has an affinity that is 6 to 10 times higher than that of the GR. Consequently, it has been proposed that the inverted-U shape function that relates circulating levels of glucocorticoids and memory performance is best explained by the differential activation of MR and GR receptors at different circulating levels of glucocorticoids. Low circulating levels of glucocorticoids mostly activate MRs and this is thought to be associated with the positive (left side of the inverted-U shape function) effects of glucocorticoids. In contrast, high circulating levels of glucocorticoids saturate MR and activate and/or saturate GRs (depending on the magnitude of the glucocorticoid increase) and this activation of MR and GR is thought to be associated with the impairing (right side of the inverted-U shape function) effects of glucocorticoids on cognition.

within-subject double-blind experimental protocol in which they first induced a chemical lowering of glucocorticoid levels by administration of metyrapone, a potent inhibitor of glucocorticoid synthesis, and then restored baseline circulating glucocorticoid levels with subsequent infusion of hydrocortisone. Memory performance of participants under each of these conditions was compared to that measured on a placebo day. The results showed that the metyrapone treatment significantly impaired memory, while hydrocortisone replacement restored performance at placebo level.

In a second study, the authors took advantage of the circadian variation in circulating levels of glucocorticoids and tested the impact of a bolus injection of 35 mg of hydrocortisone on memory performance in the afternoon. The idea behind this experiment was to test the impact of hydrocortisone in the late afternoon, at a time of very low cortisol concentrations, that is, at a time of low MR/GR ratio. It was postulated that if the ratio of MR/GR activation is involved in glucocorticoid-induced memory changes, administration of hydrocortisone in the late afternoon should increase the MR/GR ratio, and lead to increased memory performance when compared to placebo. The results confirmed the hypothesis, as it was shown that administration of hydrocortisone in the afternoon led to significantly faster detection times on the memory test when compared to administration of placebo.

In a third study, authors took advantage of the circadian variation in circulating levels of glucocorticoids and tested memory performance after exposure to a psychological stressor in participants who were stressed in the a.m. or p.m. phase. Here, the authors added a new variable in the protocol, and tested the potential differential impact of MR and GR activation on recall of neutral versus emotional information.

Young, normal adults were submitted to a psychological stress task before viewing a story composed of emotionally negative and neutral segments, and memory performance was assessed by a 1-week postlearning delayed recall. Results show that stress-induced increases in salivary glucocorticoid levels impaired delayed recall of emotionally arousing material in the a.m. group, but not in the p.m. group. There was no effect of stress on memory for neutral material. Altogether, these findings suggest that stressing participants in the a.m. phase, at a time of high circulating levels of corticosteroids, can overstimulate MR and GR receptors in the brain, impairing declarative memory for emotionally arousing material unrelated to the stressor. These findings suggest that the experimental context, that is, time of day at which the experiment occurs, the nature of the to-be-remembered material (remembering the stressful event itself or material unrelated to the stressor), and the valence of the to-be-remembered material (emotionally arousing vs. neutral), modulates the effects of stress on human memory.

Conclusion

After two centuries of research on the effects of glucocorticoids on human cognitive function, we have come to view the actions of this steroid as being modulatory rather than direct. History has taught us two important things. First, these steroids could access the brain and lead to changes in cognitive function. Second, glucocorticoids' action on cognitive process can be positive, and that the positive or negative effects of glucocorticoids depend on the balance of the two GR types known to exist today.

The story was supposed to become simpler as new discoveries were being made about the actions of glucocorticoids on the animal and human brains. Yet, it became more complex, dynamic, and difficult to grasp. Indeed, we now suspect that important variables related to the testing environment itself can have an important impact on the stress response and, consequently, can predict its impact on cognitive performance. Here, time of day, the valence, and the nature of the information to be learned were summarized as examples of the types of variables that have to be controlled when assessing the effects of glucocorticoids on cognition in humans, since these factors have been shown to impact on either the stress response itself and/or its effect on cognitive performance in humans populations. However, as complex as the story will get, it tells us that the glucocorticoid hormone is a very important player in the interface between the brain and the environment. This renders the study of its action more exciting than ever.

See also: Chronic (Repeated) Stress: Consequences, Adaptations; Hormones and Memory; Neurogenesis in the Intact Adult Brain; Stress and Vulnerability to Brain Damage; Stress and Neuronal Plasticity; Stress, Dopamine, and Puberty; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Cahill L (2000) Neurobiological mechanisms of emotionally influenced, long-term memory. *Progress in Brain Research* 126: 29–37.

de Kloet ER, Oitzl MS, and Joels M (1999) Stress and cognition: Are corticosteroids good or bad guys? *Trends in Neurosciences* 22: 422–426.

Dickerson SS and Kemeny ME (2002) Acute stressors and cortisol reactivity: A meta-analytic review. *Psychosomatic Medicine* 54: 105–123.

Fuchs E and Gould E (2000) Mini-review: *In vivo* neurogenesis in the adult brain: Regulation and functional implications. *European Journal of Neuroscience* 12: 2211–2214.

Het S, Ramlow G, and Wolf OT (2005) A meta-analytic review of the effects of acute cortisol administration on human memory. *Psychoneuroendocrinology* 30: 771–784.

Lupien SJ and Lepage M (2001) Stress, memory, and the hippocampus: Can't live with it, can't live without it. *Behavioral Brain Research* 127: 137–158.

Lupien SJ and McEwen BS (1997) The acute effects of corticosteroids on cognition: Integration of animal and human model studies. *Brain Research Reviews* 24: 1–27.

Mason JW (1968) A review of psychoendocrine research on the sympathetic-adrenal medullary system. *Psychosomatic Medicine* 30(supplement): 631–653.

McEwen BS (1998) Protective and damaging effects of stress mediators. *New England Journal of Medicine* 338: 171–179.

Roozendaal B (2000) Glucocorticoids and the regulation of memory consolidation. *Psychoneuroendocrinology* 25: 213–238.

Selye H (1975) Confusion and controversy in the stress field. *Journal of Human Stress* 1: 37–44.

Selye H (1998) A syndrome produced by diverse nocuous agents. 1936. *The Journal of Neuropsychiatry and Clinical Neurosciences* 10: 230–231.

Stress and Neural Involvement in Metabolism

M F Dallman, M T Foster, and J P Warne, University of California at San Francisco, San Francisco, CA, USA

© 2009 Published by Elsevier Ltd.

Stress

Stress was defined by Day in 2005 as "the multisystem response of the body to challenges that overwhelm or threaten to overwhelm the selective homeostatic mechanisms of the organism." This definition incorporates both Cannon's notion of homeostasis, an active defense of important body system variables, and Selye's original notion of stress, the body's nonspecific response to challenge. With such challenges there are immediate behavioral, autonomic, and neuroendocrine responses that appear to be have been designed to deal with the stressor, as effectively as possible. These occur through ensuring adequate muscular perfusion and tone, as well as energy supplies, particularly glucose and free fatty acids (FFAs), to support behavioral responses, whether escape, aggression, thermoregulation, or food intake. The multisystem response outputs are represented by the autonomic nervous system, including the adrenal medulla, which secretes catecholamines, and the adrenal cortex, which secretes glucocorticoids (GCs).

Thresholds

Day's definition of stress requires that the eliciting challenge must be of sufficiently high intensity so that discrete system-specific mechanisms are predicted to be, or are, inadequate to maintain homeostasis. Thresholds for responses of the stress systems are usually well beyond those that provoke specific responses, underlining this definition of stress. For example, with experience, a psychological threat that is introduced under conditions that allow behavioral escape does not cause a stress response, whereas the same stimulus administered under conditions that prevent escape does provoke elevated GCs. Similarly, experimentally induced decreases in blood volume, blood oxygen content, or blood glucose produce highly specific responses in the cardiovascular, respiratory, and glucose-regulating systems at initial steps, whereas at values beyond the threshold the same stimuli produce dose-related increases in catecholaminergic and glucocorticoid responses (Figure 1).The threshold for increasing autonomic outflow, however, is frequently different from that for glucocorticoid secretion. It is the brain stress response network, driven by corticotropin-releasing factor (CRF), that coordinates these multisystem responses to high intensity or persistent stressors.

CRF Stress Network

Clusters of cells that can express CRF are present in brain. Under conditions of stress and elevated GCs, CRF is induced in these neurons and at other sites in adults. Moreover, in adults that were stressed as neonates, these CRF cell groups appear to be permanently induced. In addition, CRF receptors (CRFRs) are found throughout the brain and play a key role in the multisystem responses to stress. The cluster of cells in the central nuclei of the amygdala (CeA) seems to be critical in the coordination of responses that alter activity of CRF-expressing cells in the bed nuclei of the stria terminalis as well as in the autonomic- and neuroendocrine-directed axons of cells in the hypothalamic paraventricular nuclei (PVN). Output from the CeA impinges directly on autonomic outflow from the lateral hypothalamus (LH), and indirectly on the hypothalamic PVN. Brain pretreatment with CRFR-1 antagonists diminishes or blocks many behavioral, autonomic, and neuroendocrine responses to stimuli that normally induce stress.

A single, punctate stimulus sufficient to engage the central stress network seems to have few long-term effects on the brain or metabolism. By contrast, very high-intensity single stimuli, or repeated or prolonged stimuli that sustain activity in the central stress network, may have considerable effects because of the persistent elevation in GCs and activity of CRF neurons that, in turn, interact with all monoaminergic cell groups to alter secretion of the monoamines throughout the brain, thus both sustaining and amplifying the initial effects of adequate stimuli.

In the following sections, we first discuss the metabolic consequences of increased activity in the central stress circuit – with respect to actions of GCs and the autonomic nervous system, including adrenal medullary secretion – on energy balance. Then we discuss some of the consequences of the autonomic and GC responses on the brain. Finally, we discuss the utility of these responses, both in the light of their evolutionary value and in the context of current, developed Western societies.

Stress Responses and Metabolism

The initial autonomic and endocrine responses to adequate stressors serve to maintain concentrations of critical fuels in the circulation for sustained use by the brain, heart, and exercising skeletal muscle.

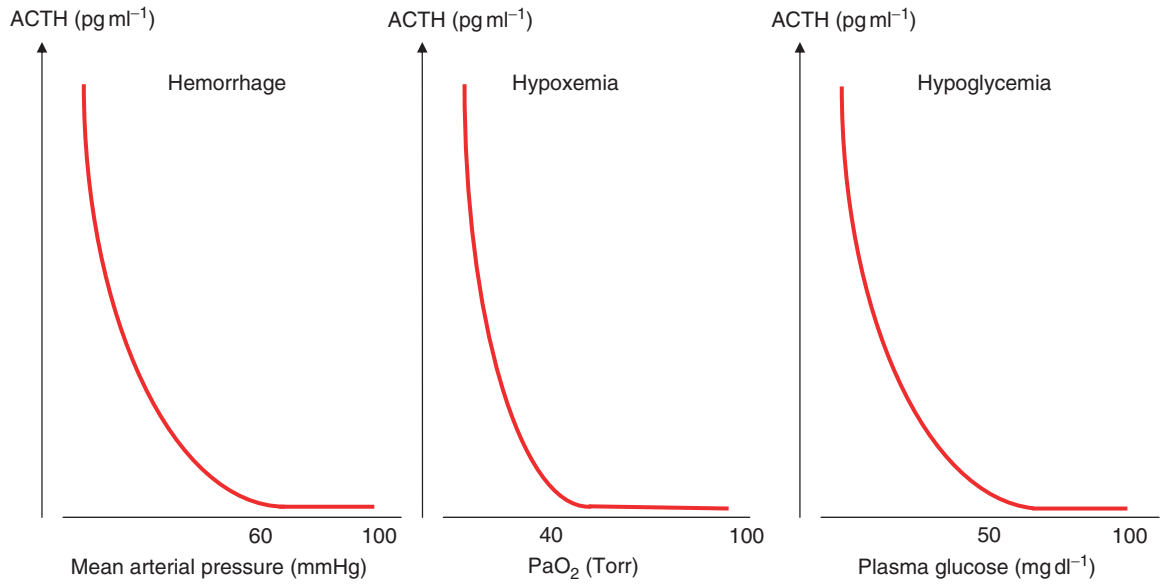

Figure 1 There are thresholds at which multisystem responses of the stress network occur (ACTH, adrenocorticotropic hormone). Below threshold, specific systems are capable of maintaining homeostasis. When stress responses occur, they are proportional to the magnitude of the insult.

Energy is stored in the body as glycogen, lipid, and protein, all of which are mobilized through actions of the autonomic nervous system and GCs in response to increased activity of the central stress response network. Thus, in the periphery, the multisystem stress responses are highly catabolic, resulting in loss of energy stores and body weight. Responses to sympathetic neural and adrenal medullary activation are rapid, and within minutes cells release preformed energy that is stored as glycogen and lipid. By contrast, most peripheral actions of the GCs are slower, increasing the rate of supply of glucose and FFAs into the vascular system by increasing the synthesis and activities of critical cellular enzymes.

Stress and Autonomic Outflow Pathways

Much of our knowledge of central autonomic pattern generators has come from the use of viral transneuronal tracer studies, in which the target innervated organ is injected with pseudorabies virus (PRV), and retrograde transport of the virus is determined over time. Using this technique, and others, a relatively small number of cell groups are seen to be directly involved in autonomic outflow from the hypothalamus to the innervated tissue. Because the live virus can replicate in cell bodies and then infect the axons afferent to the cell body, with the appropriate controls, synaptic chains of neurons can be distinguished over time after injection of virus into a specific organ. Virus is identified in the first-stage efferent sympathetic and parasympathetic cell groups in the spinal cord and medulla, as are neuronal retrograde tracers that cannot go beyond a single cell. Adrenergic neurons in C1 in the ventrolateral medulla, the medullary raphe nuclei, and adjacent medulla; the A5 group of noradrenergic neurons in the ventrolateral pons; and cells in the paraventricular nucleus and adjacent LH are filled with virus with increasing time after injections into a variety of organs; these structures are all part of the autonomic output limb of the central stress network. Although there are only a few cell groups that generally serve to supply autonomic outflow, both sympathetic outflow and parasympathetic outflow are highly selective, so that control can be manipulated organ by organ. In the PVN, for instance, there is a rough topographic relationship between the laterality of the virus-infected neurons and the body organ injected.

Stress and Secretion of GCs

The multisystem response to adequate stimuli, whether physical or psychological, involves activation of the hypothalamic–pituitary–adrenal (HPA) axis and, consequently, stimulation of the synthesis and release of GCs from the adrenal cortex. This response acts to prevent or reduce deleterious changes in essential homeostatic systems. Although metabolic changes are the focus here, it should be recognized that GCs also have major effects on the immune, cardiovascular, gastrointestinal, and skeletal systems as well. In the hypothalamus, major cell groups are heavily involved with integration inputs from stimuli and outputs through

the behavioral, autonomic, and neuroendocrine motor systems. These heavily interconnected cell groups have been designated the hypothalamic visceromotor pattern generators (HVPGs) by Thompson and Swanson. The dorsomedial hypothalamic (DMH) nuclei appear to be the dominant cell group of this network. Blockade of activity in the DMH eliminates or greatly reduces both HPA and autonomic responses to many stimuli.

The motor neurohormones involved in the HPA axis are CRF and arginine vasopressin (AVP). CRF is synthesized by neurons in the parvocellular region of the paraventricular nucleus. When activated, these neurons, which project into the median eminence, secrete CRF and this diffuses into the hypophyseal portal system to promote the release of adrenocorticotropic hormone (ACTH) from the corticotropes of the anterior pituitary. AVP is also expressed in the CRF-containing cells in the PVN and synergizes with the stimulatory actions of CRF at the level of the corticotrope. Secreted ACTH then acts at the adrenal gland to stimulate the rate of processing of cholesterol to pregnenolone, the rate-limiting step of GC synthesis, and this results in elevated circulating levels of cortisol (humans) or corticosterone (rodents).

GC Feedback

Activity of the HPA axis is tightly regulated through complex negative feedback and positive, feed-forward loops that restore circulating GC concentrations to those appropriate to the ongoing physiology of the organism. Negative feedback shows three temporally discrete domains: fast, early delayed, and late delayed feedback. Fast feedback occurs within the first few minutes of the stress response and appears to be nongenomic in nature, possibly involving the endocannabinoids. Early delayed feedback appears after 30 min and is usually maximal after 2 h postexposure, though the effect can persist for up to 24 h. This type of feedback critically requires *de novo* generation of second messengers to inhibit the release of CRF/AVP and ACTH. Thus, the duration of ACTH secreted in response to a punctate stimulus is limited by the GC negative feedback system. Late, delayed feedback is found when supraphysiological levels of GCs have persisted for days or even weeks, as in treatment with exogenous steroids or with Cushing's disease. This time domain is characterized by inhibition of CRF, AVP, and ACTH synthesis and release, and adrenal atrophy; recovery from chronic elevations in exogenously administered GC is very slow. But, with chronic stress, elevations in GC positively regulate recruited CRF neurons in the amygdala to stimulate further activity in the HPA axis. Thus, there are major distinctions between the chronic effects of GC that depend on whether the central stress network is engaged.

Stress and Maintenance of Blood Glucose

Stress has important effects on carbohydrate metabolism, affecting the synthesis, storage, and release of glucose, as well as the release and action of key regulators, notably insulin (Figure 2). Critical to this is the GC-induced development of insulin resistance, which is a consequence of the reduced action of insulin in target tissues (such as liver, adipose tissue, and muscle). The acute consequence of insulin resistance is much slower glucose removal from blood into insulin-sensitive tissues, and thus higher circulating glucose concentrations.

Liver

The liver is key to establishing blood glucose concentrations by maintaining a balance between the uptake and storage of glucose as glycogen (glycogenesis) and the synthesis and release of glucose (gluconeogenesis and glycogenolysis); the former is opposed by the process of glycolysis. There are several key regulatory enzymes involved in these processes and these form substrate cycles, in which cycling between the substrates and products of the enzymes occur. These cycles form a system in which the rate and direction of flux can be finely regulated.The enzymes specific for gluconeogenesis – phosphoenolpyruvate carboxykinase (PEPCK), fructose 1, 6-bisphosphatase, and glucose-6-phosphatase (G-6-Pase) – and for glycolysis – pyruvate kinase, 6-phosphofructo-1-kinase, and glucokinase – make sensitive control possible. Glucose homeostasis is usually controlled by insulin and the insulin-like growth factors. Increased activity of the brain CRF stress network stimulates sympathetic and HPA activity. GCs oppose the actions of insulin and promote gluconeogenesis. The net outcome of GC action on the liver is an increase in hepatic glucose output. Consequently, greater amounts of insulin are required to reduce glucose release from the liver. Treatment with exogenous GCs increases plasma insulin concentrations in relation to the dose of steroid given, and insulin secretion may increase markedly. The catecholamines (epinephrine and norepinephrine) rapidly stimulate glycogenolysis and the release of glucose into the circulation.

Prevailing Glucose Levels Are Also Influenced by Adipose Tissue

Glucose enters the adipocyte via glucose transporters, GLUT1 and GLUT4, in adipose tissue. Most of the

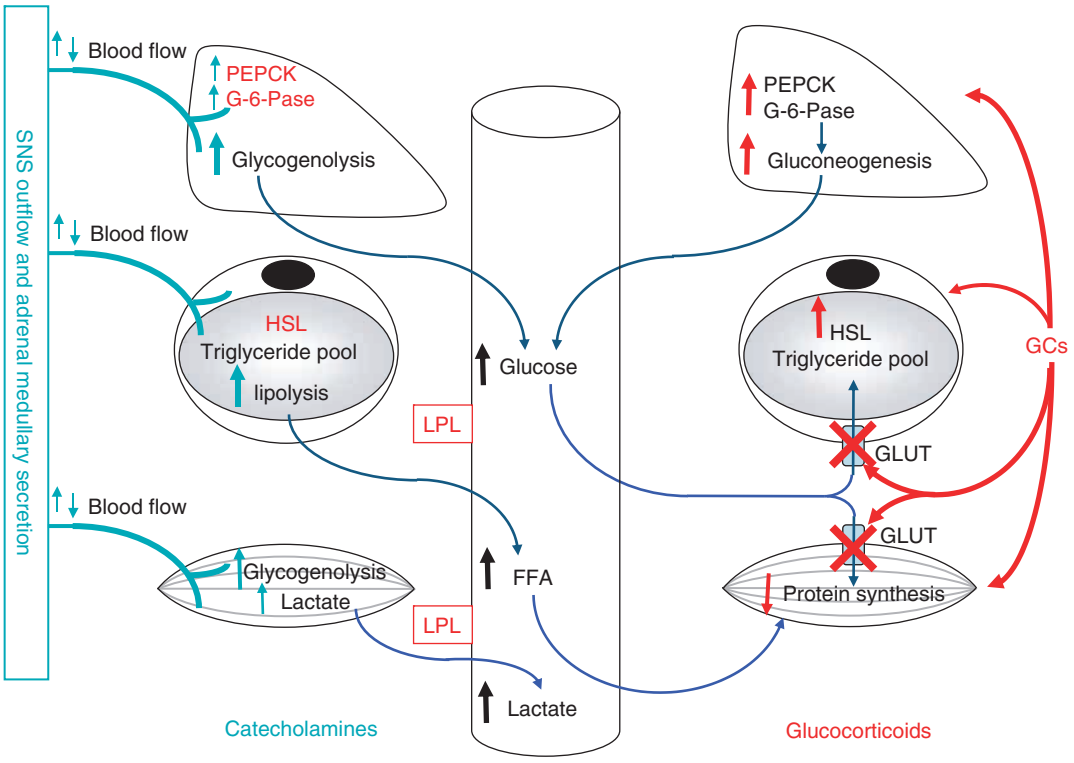

Figure 2 Multisystem stress response on metabolism mediated by the sympathetic nervous system (SNS) and HPA axis. PEPCK, phosphoenolpyruvate carboxykinase; G-6-Pase, glucose-6-phosphatase; HSL, hormone-sensitive lipase; LPL, lipoprotein lipase; GCs, glucocorticoids; GLUT, glucose transporter; FFA, free fatty acid.

glucose is metabolized to lactate or used as glycerol for subsequent conversion to triglycerides. Insulin promotes translocation of GLUT4 to the plasma membrane, thus facilitating glucose uptake. This action is antagonized by GCs. Similar GC-induced dysregulation of GLUT4 translocation also occurs in muscle. As in liver, catecholamines stimulate glycogenolysis in muscle; the glucose-6-phosphate generated is used by muscle for energy. Importantly, however, the amount of lactate released by muscle increases with increasing exercise. This lactate can be used by brain for energy in addition to the energy supplied by glucose.

GCs and catecholamines (acting via α-adrenergic receptors) also reduce insulin secretion from pancreatic beta cells in experimental animals and reduce the frequency of insulin release in fasted humans. However, this action is always modified by the circulating glucose concentrations, which must be controlled by insulin. Thus, at any given steady-state concentration of GC, the higher the GC concentration, the higher the insulin concentration will be, despite the fact that persistent stress reduces overall insulin concentrations. As with the HPA axis, long-term treatment with exogenous GCs inhibits sympathetic nervous system (SNS) outflow; however, when chronically elevated GCs are accompanied by a chronic stressor, SNS outflow is increased. Thus, the response to GCs in both systems depends entirely on whether an adequate stressor has recruited the central stress network and CRF.

Stress and Lipid Metabolism

White Adipose Tissue

White adipose tissue (WAT) cells are composed primarily of stored lipid; about 90% of the cell volume consists of a lipid droplet. Lipids are removed from the circulation through the action of lipoprotein lipase (LPL), which releases triacyl fatty acids from their complexes with phosphoprotein moieties of lipoproteins that allow them to circulate in watery solution (Figure 2). LPL is found in high density on capillary endothelia and its concentration and activity are increased by GCs and insulin, and its activity is (probably indirectly) inhibited through the actions of catecholamines on cytokines produced in WAT. Thus, when GC and insulin are elevated but sympathetic neural outflow is low, as in conditions of exogenous GC treatment, fatty acids from gut- and liver-generated lipoproteins can be readily stored in WAT. In adipocytes, the activity of the major fatty acid-mobilizing

enzyme, hormone-sensitive lipase (HSL), is also regulated by insulin, GCs, and the catecholamines. However, the degree of regulation is highly depot dependent, and intraabdominal fat stores that provide FFAs and glycerol via the portal vein directly to the liver appear to be the stores that are most sensitive to regulation by the central CRF stress network.

Brown Adipose Tissue

Specialized fat depots having cells with large numbers of mitochondria containing uncoupling protein-1 (UCP-1) is called brown adipose tissue (BAT); UCP-1 mediates an outward proton leak, resulting in nonshivering thermogenesis. BAT is present in small mammals and human newborns, but not adult humans, and is heavily innervated by sympathetic neural outflow. BAT has been studied in rodents as a model for the hypothalamic regulation of thermogenesis. Because much less is known about the central nervous system (CNS) regulation of the sympathetic outflow to WAT, and because many of the hypothalamic neurons that regulate BAT activity probably also regulate WAT activity, the neural regulation of BAT is addressed here.

Central Stress Network and BAT Function

CRF administered into the brain increases SNS outflow and affects blood pressure, heart rate, gastric acid secretion, and intestinal motility, in addition to causing alterations in lipid metabolism. Chronic intracerebroventricular (icv) infusions of CRF increase blood glucose, sympathetic neural outflow to interscapular BAT (iBAT), UCP-1 in iBAT, iBAT nonshivering thermogenesis, and catecholamine and FFA release from adipose tissue depots. Increased sympathetic neural outflow is directly responsible for the increase in thermogenesis that develops after icv injections of CRF; when neural drive to BAT persists, negative energy balance and weight loss result.

Central Hypothalamic Outputs to BAT

Viral transneuronal tract tracing confirms that neurons in the hypothalamic dorsomedial nucleus (DMN), PVN, and medial preoptic area (MPOA) contain multisynaptic connections to both WAT and BAT. The role of CRF on SNS drive is supported by several lines of evidence that focus on the hypothalamus as a major site for coordinating thermoregulation through its actions on downstream sympathetic premotor ganglion neurons (SPGNs) (Figure 3). CRF-synthesizing neurons, probably in the PVN, bed nucleus of the stria terminalis (BNST), and amygdala, seem to be responsible for stress-induced alterations in SNS outflow to iBAT. Recent studies show that direct injection of CRF into the DMH or the MPOA results in increased SNS outflow to iBAT. Inhibition of DMH activity blocks iBAT activation after injections of CRF into the MPOA, and inhibition of activity in neurons of the pontine anterior nucleus raphe pallidus blocks the stimulatory effects of CRF injected into the DMH. Therefore, in the sympathetic regulation of iBAT thermogenesis, CRF inputs into the hypothalamic MPOA and into the DMH stimulate output from raphe pallidus to the SPGNs in the spinal cord and thus regulate the activity of iBAT. It is likely that similar autonomic output pathways exist for the control of lipolysis in WAT.

Stress and Protein Metabolism

Because stress couples increased autonomic neural and HPA outflows, a typical result of chronic stress is loss of muscle protein mass. Although the amount of initial fat mass determines the amount of protein lost with stress, there is always some protein loss. Protein is degraded to amino acids, which enter the circulation and are taken up and used by liver for gluconeogenesis (Figure 2).

GC treatment of experimental animals causes loss of muscle protein associated with a reduction in the protein synthesis rate. In addition, increased protein breakdown rates have also been reported. Key to the nature of the loss of protein with GC treatment appears to be age, with a shift from GC-induced muscle protein breakdown to GC-induced reduction in muscle protein synthesis with advancing age. This has also been observed both in patients with Cushing's syndrome, who exhibit marked muscle wasting that results in profound weakness, and after intensive cortisol administration for a number of clinical conditions.

Stress and Body Weight

In male rodents, which grow throughout young adulthood, the administration of chronic or repeated stressors characteristically results in decreased chow intake and increased metabolic rate (both probably mediated by the central CRF stress response network); both responses combine to yield decreased rates of ponderal growth. Stressors result in loss of peripheral fat and protein and, with high levels of SNS outflow and GCs (and relatively high insulin levels), reallocation of energy storage into central fat depots. In part, this also occurs because the stress response inhibits the normal ongoing anabolic activity provided by both the growth hormone and gonadal axes – again, probably through central actions of CRF.

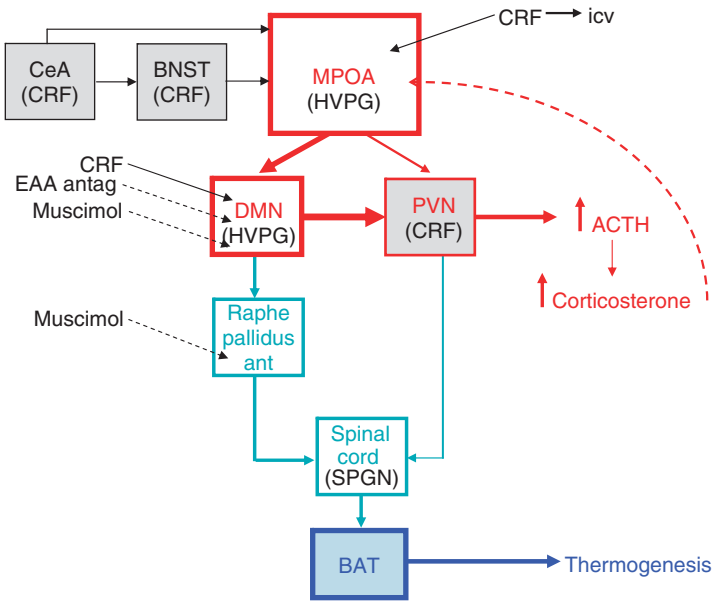

Figure 3 Corticotropin-releasing factor (CRF), via intracerebroventricular (icv) administration or into the medial preoptic area (MPOA – hypothalamic visceromotor pattern generator, HVPG) or dorsomedial nuclei (DMN), but not paraventricular nuclei (PVN), stimulates central SNS outflow to brown adipose tissue (BAT) and thermogenesis. This set of outputs mediating heat generation may be common for the central stress network that results in increased BAT activity seen after exposure to many stressors. CeA, central nuclei of the amygdala; BNST, bed nucleus of the stria terminalis; EEA, excitatory amino acid; ACTH, adrenocorticotropic hormone; SPGN, sympathetic premotor ganglion neuron.

Consequences of Stress-Induced Elevations of GCs and Catecholamines on the Brain

In addition to the actions of GCs and catecholamines on peripheral tissue targets, the brain is also a target for both of these stress response hormones. And both have actions in the processes of learning and memory: GCs markedly facilitate both search and aggressive behaviors as well as the preferential intake of calories from highly palatable sweet or fat foods. The latter effects appear to be mediated by association of GCs with relatively low-affinity glucocorticoid receptors, thus requiring high circulating concentrations of the steroid.

Learning and Memory

Catecholamines and GCs secreted during stress act through cells in the basolateral amygdala (BLA) to consolidate memories of events that are accompanied by emotional arousal. Although elevated circulating catecholamine concentrations do not enter the brain, they probably act through vagal afferent nerves to stimulate catecholamine secretion from the A2 cell group in the nucleus of the tractus solitarius, and/or from the A6 cell group in the locus coeruleus. Axons from both cell groups impinge on neurons in the BLA and, through β-adrenoceptors, activate cyclic adenosine monophosphate (cAMP)/protein kinase in these neurons. GCs interact with norepinephrine at the postsynaptic level in cells of the BLA, to increase the efficacy of the downstream cascade, and also facilitate noradrenergic brain stem cell groups that project to the BLA in consolidation of the memory of arousing stimuli.

However, high circulating concentrations of GCs also inhibit the retrieval of memories and working memory, which also depend on the integrity of the BLA and the secretion of norepinephrine into the BLA. GCs acting in the amygdala facilitate the rate of extinction of fear-induced startle. However, GCs appear to act throughout brain to modify memory consolidation and retrieval, whereas, in the tested paradigms that rely on emotional arousal, the BLA appears to be the critical site of action of the catecholamines. Thus, learning during stress is more efficient in part because of stress hormone secretion, but memory recall is damped with persistently elevated GCs, perhaps allowing more flexible choices to occur.

Running and Search Behaviors

In adrenalectomized rats, both running and search behaviors depend on GCs. In the absence of GCs, adrenalectomized animals run very little and exhibit no search behavior under conditions appropriate to increased searching; running and search behaviors are

restored to normal at low normal concentrations of GCs. In a study on search behavior, food was restricted so that the adrenalectomized rats were maintained at ~90% of their normal body weight in order to motivate them to both drink sucrose and search for it. In the absence of GCs, when sucrose in the solution was greatly reduced, search behavior was not increased, despite the increased need for calories, although in the presence of GCs, search behavior doubled. Extrapolating, it appears that GCs secreted in response to stress enable appropriate search behaviors that could reduce or eliminate the stressor.

Aggressive Behavior

As with learning and memory, GCs have bimodal effects on aggression; moreover, aggression stimulates GC secretion in vertebrate species. The rapid effects of GCs stimulate, or potentiate, aggressive behavior, but the chronic actions appear to reduce aggression, and these effects are primarily observed in social subordinates that have persistently elevated GC concentrations. It appears that, like the action of GCs on inhibition of the HPA axis, there are distinct temporal domains for the effects of GCs on aggression. The rapid effect of GCs is to increase aggressiveness, resulting in directed attacks, primarily to the more protected parts of the attacked animal, so that little permanent damage is done, and few attacks at the vulnerable sites (head, neck, and underbelly), which can result in severe, life-threatening damage. Interestingly, in the absence of high, stress-engendered GC concentrations, attacks focus to a greater extent on sites that can cause great damage to the recipient.

Chronic and persistent elevations in GCs are associated with depression-like and subordinate behaviors and serve to reduce overall activity and aggressive behavior, through mechanisms that are not well understood. The effects of GCs on aggression serve to provide more efficient competition, promoting the survival of the individual and transmittal of its genes under conditions of limited resources.

Calorie Ingestion

Stressors reduce calorie intake in vertebrates, probably through the increase in hypothalamic CRF secretion, which is anorexogenic. However, GCs modulate the choice and amount of substances ingested. In adrenalectomized rats without concomitant stress, GC replacement is dose related to ingestion of sweet (sucrose, saccharin) and fat (lard) substances, but without effect on ingestion of normal (relatively unpalatable) chow. These findings support those that show increased intake of palatable foods, generally dense in sugar and fat, in people who have undergone acute stress. Stressed rats with daily elevations in GCs, given their choice of chow, sucrose, and lard to ingest, may increase the proportion of the latter in their total intake. However, this is not always the case, and whether intake of lard and sugar is increased with stressors may represent the outcome of conflicting hedonic and memorial effects of the active central stress network and GCs. Nonetheless, persistent ingestion of calories from sucrose or lard does reduce acute HPA response to stressors and appears to reduce activity in the central stress response network. When GCs and insulin are elevated together as a result of increased lard and sugar intake, and then the abdominal fat depots preferentially store excess calories. Both hormones act to increase LPL (Figure 2), particularly in abdominal fat depots, and elevated insulin levels also counteract the lipid- and glycogen-mobilizing effects of the catecholamines.

However, because GCs affect learning and memory as well as calorie ingestion, the prior association between stress-induced ingestion of palatable foods may become habitual after the end of the stressor. Thus, the threshold for indulging in palatable foods with minor stress may decrease. Although this association does not appear to have much downside from an evolutionary standpoint, it has a distinct potential downside in current Western cultures.

Stress and Neural Involvement in Metabolism – Upside and Downside

The Upside

The central CRF-induced stress response network is designed to increase fitness of the animal to cope with adversity, as are the GC and catecholaminergic hormones. The actions of the central network and secreted hormones effect behaviors and metabolic changes that promote survival of the individual. These systems can be seen overall as a concerted effort to maintain calorie stores while serving calorie needs. It is highly likely throughout evolutionary time that the hunt for and utilization of calories have been a paramount necessity, and it is unlikely that overindulgence in calories occurred over the long run. The actions of the CRF network and its autonomic and endocrine outputs are clearly not detrimental in any major fashion. Longevity in rodents and subhuman primates is considerably extended by restricted feeding and limiting calories to 40–60% of normal *ad libitum* amounts. Under these conditions, GCs and some autonomic outflows are elevated during the entire extended life span of the rodent. Thus, *per se*, tonic activation of parts of the stress response system is not inimical to the duration of life, nor does it

promote the increased obesity, cardiovascular disease, metabolic syndrome, and type 2 diabetes that are seen with increasing frequency throughout our Western-style civilizations.

The Downside

The Western lifestyle includes the availability of plentiful, inexpensive foods – particularly those rich in sugar and fat. Also, there is far less physical activity demanded of those living the Western lifestyle. Thus, given the known central and peripheral actions of the catecholamines and GCs, and their interactions with high intake levels of sugary and fatty foods, it seems clear that this modern lifestyle can lead to the previously cited ills of our civilization. Current socioeconomic stressors are pervasive and not readily relievable, and pleasurable calories are easily available for consumption and have the effect of reducing the negative effects on mood (on activity in the stress response network). A sequence of chronic stressors, with elevated GCs, together with increased consumption of sugar and fat, will selectively increase abdominal adipose stores, because of the molding effects of elevated stress hormones on the deposition of excess calories. Large abdominal fat stores are strongly associated with cardiovascular disease, stroke, type 2 diabetes, and the metabolic syndrome. Stress reduction by means other than eating would seem to be an answer to the problem until evolutionary changes catch up with society and the mechanisms involved in calorie acquisition, stores, and mobilization become less important to the human organism.

See also: Activity-Dependent Regulation of Glucose Transporters; Adrenal Steroids: Biphasic Effects on Neurons; Norepinephrine: CNS Pathways and Neurophysiology; Stress and Cognition; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress and Parasympathetic Control; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress.

Further Reading

Cerri M and Morrison SF (2006) Corticotropin releasing factor increases in brown adipose tissue thermogenesis and heart rate through dorsomedial hypothalamus and medullary raphe pallidus. *Neuroscience* 140: 711–721.

Cook CJ (2001) Measuring of extracellular cortisol and corticotropin-releasing hormone in the amygdala using immunosensor coupled microdialysis. *Journal of Neuroscience Methods* 110: 95–101.

Dallman M, Akana S, Cascio CS, et al. (1987) Regulation of ACTH secretion: Variations on a theme of B. *Recent Progress in Hormone Research* 43: 113–173.

Dallman MF, Pecoraro N, Akana SF, et al. (2003) Chronic stress and obesity: A new view of 'comfort food.' *Proceedings of the National Academy of Sciences of the United States of America* 100(20): 11696–11701.

Day TA (2005) Defining stress as a prelude to mapping its neurocircuitries: No help from allostasis. *Progress in Neuropsychopharmacology & Biological Psychology* 29(8): 1195–1200.

DiMicco JA, Samuels BC, Zaretskaia MV, et al. (2002) The dorsomedial hypothalamus and the response to stress. Part renaissance, part revolution. *Pharmacology, Biochemistry and Behavior* 71: 469–480.

Haller J, Halasz J, Mikics E, et al. (2004) Chronic glucocorticoid deficiency-induced abnormal aggression, autonomic hypoarousal, and social defeat in rats. *Journal of Neuroendocrinology* 16: 550–557.

Heinrichs SC and Koob GF (2004) Corticotropin-releasing factor in brain: A role in activation, arousal and affect regulation. *Journal of Pharmacology and Experimental Therapeutics* 311: 427–440.

Leshner AI (1971) The adrenals and the regulatory nature of running wheel activity. *Physiology & Behavior* 6: 551–558.

Mikics E, Kruk MR, and Haller J (2004) Genomic and non-genomic effects of glucocorticoids on aggressive behavior in male rats. *Psychoneuroendocrinology* 29(5): 618–635.

Myers DA, Gibson M, Schulkin J, et al. (2005) Corticosterone implants to the amygdala and type 1 CRH receptor regulation: Effects on behavior and colonic sensitivity. *Behavioural Brain Research* 161(1): 39–44.

Owens MJ and Nemeroff CB (1991) Physiology and pharmacology of corticotropin-releasing factor. *Pharmacological Reviews* 43: 425–473.

Pacak K and Palkovits M (2001) Stressor specificity of central neuroendocrine responses: Implications for stress-related disorders. *Endocrine Reviews* 22: 502–548.

Pecoraro N, Dallman MF, Warne JP, et al. (2006) From malthus to motive: How the HPA axis engineers the phenotype, yoking needs to want. *Progress in Neurobiology* 79: 247–330.

Pellerin L (2005) How astrocytes feed hungry neurons. *Molecular Neurobiology* 32(1): 59–72.

Peters A, Pellerin L, Dallman MF, et al. (2007) Causes of obesity: looking beyond the hypothalamus. *Progress in Neurobiology* 81: 61–88.

Plotsky PM, Thrivikraman KV, Nemeroff CB, et al. (2005) Long-term consequences of neonatal rearing on central corticotropin-releasing factor systems in adult male rat offspring. *Neuropsychopharmacology* 30: 2192–2204.

Roozendaal B, Okuda S, de Quervain DJ, et al. (2006) Glucocorticoids interact with emotion-induced noradrenergic activation in influencing different memory functions. *Neuroscience* 138(3): 901–910.

Rothwell NJ (1990) Central effects of CRF on metabolism and energy balance. *Neuroscience and Biobehavioral Reviews* 14: 263–271.

Smagin GN, Howell LA, Redmann S Jr., et al. (1999) Prevention of stress-induced weight loss by third ventricle CRF receptor antagonist. *American Journal of Physiology* 276(5, Pt. 2): R146–R148.

Strack AM, Sawyer WB, Hughes JH, et al. (1989) A general pattern of CNS innervation of the sympathetic outflow demonstrated by transneuronal pseudorabies viral infection. *Brain Research* 491: 156–162.

Thompson RH and Swanson LW (2003) Structural characterization of a hypothalamic visceromotor pattern generator network. *Brain Research – Brain Research Reviews* 41(2–3): 153–202.

Stress and Neuronal Plasticity

B S McEwen, The Rockefeller University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The brain is the key organ of the stress response insofar as it processes information and determines whether an event or situation is threatening. Moreover, the outputs of the brain include regulation of autonomic, neuroendocrine and immune function, and behavioral responses that can be health promoting or health damaging. The brain is an adaptable organ, and both acute and chronic stress produce structural as well as neurochemical alterations that are, for the most part, reversible. Yet, severe insults such as seizures, strokes, and head trauma cause permanent damage, and there is also reason to believe that severe and prolonged stress, as well as disorders such as major depression, may also lead to permanent neural damage if they go on for a very long time.

Animal models have provided clues as to how the brain adapts to acute and repeated stress, and the hippocampus is the best-studied brain region. The hippocampus, which is an important brain structure for spatial, contextual, and declarative memory contains receptors for adrenal steroids, which regulate excitability and the morphological changes together with glutamate as well as other neurotransmitters and a host of endogenous neurochemicals.

This article first discusses the hippocampus and then turns to two other stress-sensitive brain structures, namely, the amygdala and prefrontal cortex. It then considers the behavioral consequences of remodeling of neural connections by stress and whether the remodeling increases or decreases the vulnerability of the brain to permanent damage. Finally, the relevance of the structural changes to the mood and anxiety disorders is discussed.

Adaptive Structural Plasticity

One of the ways that stress hormones modulate function within the brain is by changing the structure of neurons. Within the hippocampus, the input from the entorhinal cortex to the dentate gyrus is ramified by the connections between the dentate gyrus and the CA3 pyramidal neurons. One granule neuron innervates, on average, 12 CA3 neurons, and each CA3 neuron innervates, on average, 50 other CA3 neurons via axon collaterals, as well as 25 inhibitory cells via other axon collaterals. The net result is a 600-fold amplification of excitation, as well as a 300-fold amplification of inhibition, that provides some degree of control of the system.

As to why this type of circuitry exists, the dentate gyrus-CA3 system is believed to play a role in the memory of sequences of events, although long-term storage of memory occurs in other brain regions. But, because the DG-CA3 system is so delicately balanced in its function and vulnerability to damage, there is also adaptive structural plasticity, in that new neurons continue to be produced in the dentate gyrus throughout adult life, and CA3 pyramidal cells undergo a reversible remodeling of their dendrites in conditions such as hibernation and chronic stress. The role of this plasticity may be to protect against permanent damage.

Neurogenesis in the Dentate Gyrus

The subgranular layer of the dentate gyrus contains cells that have some properties of astrocytes (e.g., expression of glial fibrillary acidic protein) and which give rise to granule neurons. After bromodeoxyuridine (BrdU) administration to label DNA of dividing cells, these newly born cells appear as clusters in the inner part of the granule cell layer, where a substantial number of them will go on to differentiate into granule neurons within as little as 7 days. In the adult rat, 9000 new neurons are born per day and survive with a half-life of 28 days.

There are many hormonal, neurochemical, and behavioral modulators of neurogenesis and cell survival in the dentate gyrus including estradiol, insulin-like growth factor-1 (IGF-1), antidepressants, voluntary exercise, and hippocampal-dependent learning. With respect to stress, certain types of acute stress and many chronic stressors suppress neurogenesis or cell survival in the dentate gyrus, and the mediators of these inhibitory effects include excitatory amino acids acting via *N*-methyl-D-aspartate (NMDA) receptors and endogenous opioids.

Remodeling of Dendrites

Another form of structural plasticity is the remodeling of dendrites in the hippocampus. Chronic restraint stress causes retraction and simplification of dendrites in the CA3 region of the hippocampus. Such dendritic reorganization is found in both dominant and subordinate rats undergoing adaptation of psychosocial stress in the visible burrow system and it is independent of adrenal size.

What this result emphasizes is that it is not adrenal size or presumed amount of physiological stress *per se* that determines dendritic remodeling but a complex set of other factors that modulate neuronal structure. Indeed, in species of mammals that hibernate, dendritic remodeling is a reversible process and occurs within hours of the onset of hibernation in European hamsters and ground squirrels, and it is also reversible within hours of wakening of the animals from torpor. This implies that reorganization of the cytoskeleton is taking place rapidly and reversibly and that changes in dendrite length and branching are not 'damage' but a form of structural plasticity.

Mechanisms of Structural Remodeling

Adrenal steroids are important mediators of remodeling of hippocampal neurons during repeated stress, and exogenous adrenal steroids can also cause remodeling in the absence of an external stressor. The role of adrenal steroids may involve interactions with neurochemical systems in the hippocampus, including serotonin, γ-aminobutyric acid (GABA), and excitatory amino acids. Probably the most important interactions are those with excitatory amino acids such as glutamate. Excitatory amino acids released by the mossy fiber pathway play a key role in the remodeling of the CA3 region of the hippocampus, and regulation of glutamate release by adrenal steroids may play an important role.

Among the consequences of restraint stress is the elevation of extracellular glutamate levels, leading to induction of glial glutamate transporters as well as increased activation of the nuclear transcription factor, phosphoCREB. Moreover, 21d chronic restraint stress (21d CRS) leads to depletion of clear vesicles from mossy fiber terminals and increased expression of presynaptic proteins involved in vesicle release. Taken together with the fact that vesicles that remain in the mossy fiber terminal are near active synaptic zones, this implies that CRS increases the release of glutamate.

Extracellular molecules play a role in remodeling. Neural cell adhesion molecule (NCAM) and its polysialated-NCAM (PSA-NCAM) as well as L1 are expressed in the dentate gyrus and CA3 region and the expression of both NCAM, L1, and PSA-NCAM are regulated by 21d CRS. Tissue plasminogen activator (tPA) is an extracellular protease and signaling molecule that is released with neural activity and is required for chronic stress-induced loss of spines and NMDA receptor subunits on CA1 neurons.

Within the neuronal cytoskeleton, the remodeling of hippocampal neurons by chronic stress and hibernation alters the acetylation of microtubule subunits that is consistent with a more stable cytoskeleton and alters microtubule associated proteins, including the phosphorylation of a soluble form of tau, which is increased in hiberation and reversed when hibernation is terminated.

Neurotrophic factors also play a role in dendritic branching and length in that brain-derived neurotrophic factor (BDNF)+/− mice show a less branched dendritic tree and do not show a further reduction of CA3 dendrite length with chronic stress, whereas wild-type mice show reduced dendritic branching. However, there is contradictory information thus far concerning whether CRS reduces BDNF mRNA levels, some reporting a decrease and other studies reporting no change. This may reflect the balance of two opposing forces, namely, that stress triggers increased BDNF synthesis to replace depletion of BDNF caused by stress. BDNF and corticosteroids appear to oppose each other – with BDNF reversing reduced excitability in hippocampal neurons induced by stress levels of corticosterone.

Corticotrophin releasing factor (CRF) is a key mediator of many aspects related to stress. CRF in the paraventricular nucleus regulates adrenocorticotropin (ACTH) release from the anterior pituitary gland, whereas CRF in the central amygdala is involved in control of behavioral and autonomic responses to stress including the release of tPA that is an essential part of stress-induced anxiety and structural plasticity in the medial amygdala. CRF in the hippocampus is expressed in a subset of GABA neurons (Cajal-Retzius cells) in the developing hippocampus, and early life stress produces a delayed effect that reduces cognitive function and the number of CA3 neurons, as well as decreased branching of hippocampal pyramidal neurons. Indeed, CRH inhibits dendritic branching in hippocampal cultures *in vitro*.

What about Permanent Damage as a Result of Stress?

There is always concern that severe and prolonged stress may damage the brain. In the case of depressive illness, there is a high degree of life-time recurrence of major depression, which suggests that an underlying pathophysiological process is involved. What is being learned from animal models of stress and trauma effects in the hippocampus is that the remodeling of the hippocampus in response to stress is largely reversible if chronic stress is terminated at the end of 3 weeks. Yet, it is well established that glucocorticoids exacerbate damage to the hippocampus caused by ischemia and seizures. Nevertheless, the brain protects itself, as for example, in the phenomenon of

ischemia preconditioning: that is, prior stimulation of the hippocampus by a small ischemic event can induce a protective mechanism that may reduce the damage produced by a full-scale ischemic event. It is not clear if the same mechanisms might be operative when stress is applied, although there is evidence that 21d CRS increases the vulnerability to excitotoxic damage by ibotenic acid applied directly into the hippocampus. The unanswered question is whether the damage would have been worse if the dendritic retraction that occurs after 21d CRS had not occurred.

We do know that in the hippocampus there is a delicate balance between the generation of damage and destruction and the mechanisms that lead to dendrite retraction during more gradual repetitive stimulation, which may be a protective response that reduces the excitatory input without totally disconnecting the circuit functionally.

Stress-Induced Changes in the Prefrontal Cortex and Amygdala

Repeated stress also causes changes in other brain regions such as the prefrontal cortex and amygdala. Repeated stress causes dendritic shortening in medial prefrontal cortex but produces dendritic growth in neurons in amygdala, as well as in orbitofrontal cortex. Along with many other brain regions, the amygdala and prefrontal cortex also contain adrenal steroid receptors; however, the role of adrenal steroids, excitatory amino acids, and other mediators has not yet been studied in these brain regions. Nevertheless, in the amygdala, there is some evidence regarding mechanism, in that tPA is required for acute stress not only to activate indices of structural plasticity but also to enhance anxiety. These effects occur in the medial and central amygdala and not in basolateral amygdala, and the release of CRH acting via CRH1 receptors appears to be responsible.

Behavioral Effects of Acute and Chronic Stress

Acute stress induces spine synapses in CA1 region of hippocampus and chronic stress also increases spine synapse formation in amygdala but decreases it in hippocampus. Moreover, chronic stress for 21 days or longer impairs hippocampal-dependent cognitive function and enhances amygdala-dependent unlearned fear and fear conditioning, which are consistent with the opposite effects of stress on hippocampal and amygdala structure. Chronic stress also increases aggression between animals living in the same cage, and this is likely to reflect another aspect of hyperactivity of the amygdala. Behavioral correlates of remodeling in the prefrontal cortex include impairment in attention set shifting, possibly reflecting structural remodeling in the medial prefrontal cortex.

Relationship to Anxiety and Mood Disorders

Life events are known to precipitate depressive illness in individuals with certain genetic predispositions. Moreover, brain regions such as the hippocampus, amygdala, and prefrontal cortex show altered patterns of activity in positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) and also demonstrate changes in volume of these structures with recurrent depression: decreased volume of hippocampus and prefrontal cortex and amygdala. Interestingly, amygdala volume has been reported to increase in the first episode of depression, whereas hippocampal volume is not decreased. It has been known for some time that stress hormones such as cortisol are involved in psychopathology, reflecting emotional arousal and psychic disorganization rather than the specific disorder *per se*. We now know that adrenocortical hormones enter the brain and produce a wide range of effects upon it.

In Cushing's disease, there are depressive symptoms that can be relieved by surgical correction of the hypercortisolemia. Both major depression and Cushing's disease are associated with chronic elevation of cortisol that results in gradual loss of minerals from bone and abdominal obesity. In major depressive illness, as well as in Cushing's disease, the duration of the illness and not the age of the subjects predicts a progressive reduction in volume of the hippocampus, determined by structural magnetic resonance imaging. Moreover, there are a variety of other anxiety-related disorders, such as posttraumatic stress disorder (PTSD) and borderline personality disorder, in which atrophy of the hippocampus has been reported, suggesting that this is a common process reflecting chronic imbalance in the activity of adaptive systems, such as the hypothalamic–pitutary–adrenal (HPA) axis, but also including endogenous neurotransmitters, such as glutamate.

Conclusion

Animal stress models not only suggest how the human brain may change under repeated stress, but they also provide mechanistic clues about stress-induced anxiety and behavioral depression, which are relevant to human mood and anxiety disorders. For example, psychosocial stress in an animal model of depressive illness, resident-intruder stress in the tree shrew, suppresses neurogenesis and causes dendritic shrinkage in hippocampus.

Translational studies of brain changes in major mood and anxiety disorders (such as unipolar and bipolar depression and PTSD) show that changes in volume of structures (such as hippocampus, prefrontal cortex, and amygdala) must be considered as part of the neurobiological consequences of these illnesses. Structural remodeling in these brain regions is important for human psychiatric disorders because the altered circuitry is likely to contribute to impaired cognitive function and affect regulation. Moreover, stress is widely acknowledged as a predisposing and precipitating factor in psychiatric illness. Thus, animal models are relevant to human psychiatric disorders by showing the delicate balance between protection and damage and by providing mechanisms that raise the hopeful possibility that brain changes, in at least some major psychiatric disorders, may be treatable if we can find the right agents or therapies and intervene in time.

Acknowledgments

Research support has come from the National Institute of Mental Health Grants MH41256 and MH58911.

See also: Chronic (Repeated) Stress: Consequences, Adaptations; Gene Therapy and Protection from Stress-Induced Brain Damage; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Neuronal Plasticity after Cortical Damage; Stress and Vulnerability to Brain Damage; Synaptic Plasticity: Learning and Memory in Normal Aging.

Further Reading

Chen Y, Bender RA, Brunson KL, et al. (2004) Modulation of dendritic differentiation by corticotropin-releasing factor in the developing hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* 101: 15782–15787.

Conrad CD (2006) What is the functional significance of chronic stress-induced CA3 dendritic retraction within the hippocampus? *Behavioral and Cognitive Neuroscience Review* 5: 41–60.

Czeh B, Michaelis T, Watanabe T, et al. (2001) Stress-induced changes in cerebral metabolites, hippocampal volume, and cell proliferation are prevented by antidepressant treatment with tianeptine. *Proceedings of the National Academy of Sciences of the United States of America* 98: 12796–12801.

Isgor C, Kabbaj M, Akil H, and Watson SJ (2004) Delayed effects of chronic variable stress during peripubertal-juvenile period on hippocampal morphology and on cognitive and stress axis functions in rats. *Hippocampus* 14: 636–648.

Leuner B, Gould E, and Shors TJ (2006) Is there a link between adult neurogenesis and learning? *Hippocampus* 26: 216–224.

Liston C, Miller MM, Goldwater DS, et al. (2006) Stress-induced alterations in prefrontal cortical dendritic morphology predict selective impairments in perceptual attentional set-shifting. *Journal of Neuroscience* 26: 7870–7874.

Malberg JE, Eisch AJ, Nestler EJ, and Duman RS (2000) Chronic antidepressant treatment increases neurogenesis in adult rat hippocampus. *Journal of Neuroscience* 20: 9104–9110.

McEwen BS (1999) Stress and hippocampal plasticity. *Annual Review of Neuroscience* 22: 105–122.

Pawlak R, Magarinos AM, Melchor J, McEwen B, and Strickland S (2003) Tissue plasminogen activator in the amygdala is critical for stress-induced anxiety-like behavior. *Nature Neuroscience* 6: 168–174.

Sandi C (2004) Stress, cognitive impairment and cell adhesion molecules. *Nature Review/Neuroscience* 5: 917–930.

Sapolsky R (1992) *Stress, the Aging Brain and the Mechanisms of Neuron Death,* vol. 1, 423pp. Cambridge: MIT Press.

Sheline YI (2003) Neuroimaging studies of mood disorder effects on the brain. *Biological Psychiatry* 54: 338–352.

Sousa N, Lukoyanov NV, Madeira MD, Almeida OFX, and Paula-Barbosa MM (2000) Reorganization of the morphology of hippocampal neurites and synapses after stress-induced damage correlates with behavioral improvement. *Neuroscience* 97: 253–266.

Vyas A, Mitra R, Rao BSS, and Chattarji S (2002) Chronic stress induces contrasting patterns of dendritic remodeling in hippocampal and amygdaloid neurons. *Journal of Neuroscience* 22: 6810–6818.

Wellman CL (2001) Dendritic reorganization in pyramidal neurons in medial prefrontal cortex after chronic corticosterone administration. *Journal of Neurobiology* 49: 245–253.

Stress and Parasympathetic Control

S W Porges, University of Illinois at Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Traditionally, the autonomic nervous system is conceptualized as the paired antagonism between the sympathetic and the parasympathetic nervous system. These components of the autonomic nervous system, based on the region of the brain and spinal cord in which the autonomic nerves (i.e., preganglionic fibers) have their origin, compete functionally by either increasing or decreasing activity of specific target organs. The sympathetic system is defined by the autonomic fibers that exit thoracic and lumbar segments of the spinal cord. The parasympathetic system is defined by the autonomic fibers that exit either the brain stem via the cranial nerves (primarily via the vagus) or the sacral segments of the spinal cord. Since most organs in the viscera receive both sympathetic and vagal input, the regulation of the autonomic nervous system has been modeled as a balance system.

Central nervous system regulation of visceral organs is the focus of several historic publications. For example, in *The Expression of Emotions in Man and Animals*, Darwin described the dynamic neural relationship between the heart and the brain. He stated that

> when the heart is affected it reacts on the brain; and the state of the brain again reacts through the pneumogastric [vagus] nerve on the heart; so that under any excitement there will be much mutual action and reaction between these, the two most important organs of the body.

From Darwin's perspective, the vagus as a cranial nerve was a direct portal to and from the brain. Although Darwin acknowledged the bidirectional communication between the viscera and the brain, the subsequent formal description of the autonomic nervous system minimized the importance of central regulatory structures and afferents. A focus on the peripheral motor nerves with an emphasis on the paired antagonism between sympathetic and parasympathetic efferent pathways on the target visceral organs resulted in a lack of interest in afferent influences to the brain stem areas regulating specific efferent pathways. This early conceptualization of the vagus focused on an undifferentiated efferent pathway that was assumed to modulate 'tone' concurrently to several target organs. Thus, neural circuits regulating the supradiaphragmatic (e.g., myelinated vagal pathways originating in the nucleus ambiguus and terminating primarily above the diaphragm) were not functionally distinguished from the subdiaphragmatic (e.g., unmyelinated vagal pathways originating in the dorsal motor nucleus of the vagus and terminating primarily below the diaphragm). Without this distinction, research and theory focused on the paired antagonism between the parasympathetic and sympathetic innervation to target organs. The consequence of an emphasis on paired antagonism in physiology was an acceptance and use of global constructs focusing on peripheral physiology such as autonomic balance, sympathetic tone, and vagal tone. Ironically, the origin of modern research on stress physiology is often linked to the classical conditioning of autonomic activity, which, as Pavlov demonstrated, requires the involvement of higher brain structures in the modulation of visceral responses.

In the mid-1950s, W Hess proposed that the 'autonomic' nervous system was not solely 'vegetative' and automatic but, rather, was an integrated system with both peripheral and central neurons. In 1949, Hess was awarded the Nobel Prize in Physiology or Medicine. His Nobel lecture was titled 'The central control of activity of internal organs.' In the lecture, Hess acknowledged the importance of the prevailing model of the autonomic nervous system, which emphasized the paired antagonistic innervations of the internal organs and the definition of sympathetic and parasympathetic functions. However, he went well beyond this conceptualization to emphasize the importance of central structures in the regulation of visceral state by describing his studies that demonstrated the influence of the hypothalamus on the autonomic nervous system. By emphasizing the central mechanisms that mediate the dynamic regulation of peripheral organs, Hess anticipated the need for methodologies and technologies to continuously monitor the neural circuits involving both defined brain structures and peripheral nerves in the regulation of visceral function and state and to move from the prevailing conceptualization of autonomic nervous system as a peripheral system. The lecture provides a succinct statement: (1) to emphasize the importance of feedback circuits linking peripheral organs to brain structures and the bidirectionality of these feedback circuits and (2) to acknowledge that, although much can be learned about neural structures and functions via traditional experimental paradigms (e.g., neural blockade, surgery, and electrical stimulation), the dynamic feedback circuits cannot be adequately studied through these paradigms.

Arousal Theory and Stress Responses

For more than a century, researchers have measured autonomic variables (e.g., heart rate and palmar sweat gland activity) as indicators of emotional state related to perceived stress (e.g., fear, mental effort, workload, and anxiety). Historically, arousal theories provided scientists who studied brain–behavior relations with a model that assumed that activation of peripheral physiological measures regulated by the sympathetic branch of the autonomic nervous system provided 'sensitive' indicators of brain 'arousal' or 'activation.' This view was based on a rudimentary understanding of the autonomic nervous system in which changes in easily measured peripheral organs (e.g., sweat glands and heart) were assumed to be accurate indicators of how the brain was processing emotional stimuli. Usually, the emotional states were associated with fight-or-flight behaviors and the sympathetic–adrenal system (e.g., increases in heart rate, sweat gland activity, and circulating catecholamines) initially described by Cannon as well as the increased activity of the hypothalamic–pituitary–adrenal (HPA) axis with an emphasis on cortisol as a dependent variable in stress research as initially described by Selye.

Measures due, in part, to their availability and their neuroanatomical association with the target organs of the sympathetic nervous system (i.e., sweat and heart rate) and with secretions from the adrenal medulla (i.e., epinephrine and norepinephrine) and the adrenal cortex (i.e., cortisol) have become the primary physiological variables used to assess stress. An acceptance of a unitary arousal system created a research environment that neglected several important factors, including an understanding of the brain structures that regulate autonomic function; how these structures evolved from the most primitive vertebrates to mammals; how the autonomic nervous system interacts with the immune system, the HPA axis, and the neuropeptides, oxytocin, and vasopressin; and the co-evolution of stress and coping strategies with the increasing complexity of the autonomic nervous system. Missing from the dialog on stress is the role of the parasympathetic nervous system and especially the vagus with its bidirectional portal between the brain and specific visceral organs.

The Mammalian Nervous System Is a Product of Evolution

Embedded in the mammalian nervous system are neuroanatomical structures related to the expression and experience of stress. Via evolutionary processes, the mammalian nervous system has emerged with specific features that react to the challenge to maintain visceral homeostasis. Stress, simply defined, is the product of demands on the nervous system that result in a deviation from homeostasis. In general, the domains of homeostasis, which have been monitored, have focused on the visceral systems involved in cardiovascular, digestive, reproductive, and immune functions. Adaptive and successful coping to stress results in minimizing the magnitude and duration of this deviation. To survive, mammals must determine friend from foe, when an environment is safe, and communicate with their social unit. These survival-related behaviors limit the extent to which a mammal can be physically approached, whether vocalizations will be understood, and whether coalitions can be established. Moreover, these behavioral strategies, which are used to navigate through the 'stress of life,' form the bedrock on which social behaviors and higher cognitive processes can be developed and expressed. Thus, learning and other expansive mental processes must be structured, manipulated, and studied within the context of how the environment fosters or ameliorates stress-related physiological states.

Table 1 illustrates the phylogenetic differences in the structures that regulate the heart in vertebrates. The heart is selected because the regulation of the heart determines the availability of the metabolic resources required for mobilization as well as for growth and restoration. For example, cardiac output must be regulated to remain calm in safe environments, to mobilize for fight-or-flight behaviors, or to immobilize for death feigning or avoidance behaviors. To regulate cardiac output, several efferent structures have evolved. These structures represent two global and often opposing systems: (1) a sympathetic–catecholamine system including catecholamine-secreting chromaffin tissue and spinal sympathetic nerves and (2) a vagal system (a component of the parasympathetic nervous system)

Table 1 Phylogenetic shifts in vertebrate cardiac control

Group	*CHM*	*DVC*	*SNS*	*AD/m*	*VVC*
Jawless fish	X^+	(X^+)			
Cartilaginous fish	X^+	X^-			
Bony fish	X^+	X^-	X^+		
Amphibians	X^+	X^-	X^+		
Reptiles	X^+	X^-	X^+	X^+	
Mammals	X^+	X^-	X^+	X^+	X^-

CHM, chromaffin tissue; DVC, dorsal vagal complex with vagal efferent pathways originating in the dorsal motor nucleus of the vagus and vagal afferents terminating in the nucleus tractus solitarius; SNS, spinal sympathetic nervous system; AD/m, adrenal medulla; VVC, ventral vagal complex with efferent pathways originating in the nucleus ambiguus that regulate visceral structures (heart, bronchi, and thymus) and striated muscles via special visceral efferents and afferents via the tractus solitarius, trigeminal, and facial nerve.

with branches originating in the medullary source nuclei (i.e., dorsal motor nucleus of the vagus and nucleus ambiguus).

Unlike other vertebrates with a cardioinhibitory vagus, the mammalian vagus contains two branches. One branch originates in the dorsal motor nucleus of the vagus and provides the primary neural regulation of subdiaphragmatic organs such as the digestive tract. However, at the level of the heart and unlike more primitive vertebrates, the efferent vagal pathways that originate in the dorsal motor nucleus of the vagus do not play a major role in the normal dynamic regulation of cardiac output. Rather, during embryological development in mammals, cells from the dorsal motor nucleus of the vagus migrate ventrally and laterally to the nucleus ambiguus. There, they form the cell bodies for the visceromotor myelinated axons that provide potent inhibition of the sinoatrial node, the pacemaker for the heart.

Three principles can be extracted by investigating the phylogeny of the regulation of the vertebrate heart. First, there is a phylogenetic shift in the regulation of the heart from endocrine communication to unmyelinated nerves and, finally, myelinated nerves. Second, there is a development of opposing neural mechanisms of excitation and inhibition to provide rapid regulation of graded metabolic output. Third, with increased cortical development, the cortex exhibits greater control over the brain stem via direct (e.g., corticobulbar) and indirect (e.g., corticoreticular) neural pathways originating in motor cortex and terminating in the source nuclei of the myelinated motor nerves emerging from the brain stem (e.g., specific neural pathways embedded within cranial nerves V, VII, IX, X, and XI), controlling visceromotor structures (i.e., heart, bronchi, and thymus) and somatomotor structures (muscles of the face and head) that results in a neural circuit that functions to facilitate social behavior and to maintain calm behavioral states.

These phylogenetic principles provide a basis for investigating the parasympathetic contribution to stress reactions and recovery. In general, phylogenetic development results in increased neural control of the heart via the myelinated mammalian vagal system, which can promote transitory mobilization and the expression of sympathetic tone without requiring sympathetic or adrenal activation. With this new vagal system, transitory incursions into the environment or withdrawals from a potential predator can be initiated without the severe biological cost of the metabolic excitation associated with sympathetic–adrenal activation. Paralleling this change in neural control of the heart is an enhanced neural control of the face, larynx, and pharynx that enables complex facial gestures and vocalizations associated with social communication. This phylogenetic course results in greater central nervous system regulation of behavior, especially that needed to engage and disengage with environmental challenges.

Polyvagal Theory: Three Phylogenetic Systems Related to Stress Reactivity

The polyvagal theory emphasizes the phylogenetic origins of brain structures that regulate social and defensive behaviors. The polyvagal theory proposes that the evolution of the mammalian autonomic nervous system provides the neurophysiological substrates for emotional experiences, affective processes, and stress responses. The theory proposes that physiological state limits the range of adaptive behaviors and psychological experiences. In this context, the evolution of the nervous system determines the range of emotional expression, the quality of communication, and the ability to regulate bodily and behavioral state including the expression and recovery of stress-related responses. Relevant to adaptive stress responses, the theory makes the following assumptions:

1. Evolution has modified the structures of the autonomic nervous system.
2. The mammalian autonomic nervous system retains vestiges of phylogenetically older autonomic nervous systems.
3. Emotional regulation and social behavior are functional derivatives of structural changes in the autonomic nervous system due to evolutionary processes.
4. In mammals, the autonomic nervous system response strategy to challenge follows a phylogenetic hierarchy, starting with the newest structures and, when all else fails, reverting to the most primitive structural system.
5. The phylogenentic stage of the autonomic nervous system determines the behavioral, physiological, and affective features of stress reactivity.

The polyvagal theory emphasizes and documents the neurophysiological and neuroanatomical distinction between the two branches of the vagus (i.e., the tenth cranial nerve) and proposes that each vagal branch is associated with a different adaptive behavioral and physiological response strategy to stressful events. The theory describes three phylogenetic stages of the development of the mammalian autonomic nervous system (Table 2). These stages reflect the emergence of three distinct subsystems, which are phylogenetically ordered and behaviorally linked to communication (e.g., facial expression, vocalization, and listening), mobilization

Table 2 Phylogenetic stages of the neural control of the heart proposed by the polyvagal theory

Stage	*Autonomic nervous system component*	*Behavioral function*	*Lower motor neurons*
III	Myelinated vagus (ventral vagal complex)	Social communication, self-soothing and calming, inhibit 'arousal'	Nucleus ambiguus
II	Sympathetic–adrenal system	Mobilization (active avoidance)	Spinal cord
I	Unmyelinated vagus (dorsal vagal complex)	Immobilization (death feigning, passive avoidance)	Dorsal motor nucleus of the vagus

(e.g., fight-or-flight behaviors), and immobilization (e.g., death feigning, behavioral 'shutdown,' and syncope). The mobilization system is dependent on the functioning of the sympathetic nervous system. The most phylogenetically primitive component, the immobilization system, is dependent on the unmyelinated or 'vegetative' vagus, which is shared with most vertebrates. With increased neural complexity due to phylogenetic development, the organism's behavioral and affective repertoire is enriched. These subsystems provide a framework to study the vagal contribution to the regulation of stress responses.

The mammalian vagus functions as an active vagal brake in which rapid inhibition and disinhibition of vagal tone to the heart can support behavioral mobilization or self-sooth and calm an individual. When the vagal tone to the pacemaker is high, the vagus acts as a restraint or brake limiting heart rate. When vagal tone to the pacemaker is low, there is little or no inhibition of the pacemaker. Due to vagal influences to the sinoatrial node (i.e., the heart's pacemaker), resting heart rate is substantially lower than the intrinsic rate of the pacemaker. In mammals, the primary vagal inhibitory pathways occur through the myelinated vagus originating in the nucleus ambiguus. Consistent with assumptions of the polyvagal theory, the vagal brake contributes to the modulation of cardiac output by decreasing or increasing the inhibitory vagal control of the heart to influence rate and thereby adjust metabolic resources to support either mobilization or social engagement behaviors.

By transitory downregulation of the cardioinhibitory vagal tone to the heart (i.e., removing the vagal brake), the mammal is capable of rapid increases in cardiac output without activating the sympathetic–adrenal system. By engaging this system, rather than the sympathetic–adrenal system, mammals have an opportunity to rapidly increase metabolic output for immediate mobilization. Under prolonged challenge, the sympathetic system may also be activated. By rapidly re-engaging the vagal system, the sympathetic input to the heart is inhibited, resulting in a decrease in metabolic output and a calm behavioral state.

The Vagal Brake: Parasympathetic Regulation of the 'Stress' Response

Due to the tonic vagal influences to the sinoatrial node, resting heart rate is substantially lower than the intrinsic rate of the pacemaker. When the vagal tone to the pacemaker is high, the vagus acts as a brake on the rate at which the heart is beating. When vagal tone to the pacemaker is low, there is little or no inhibition of the pacemaker. Thus, the vagal brake may be used as a construct to describe the functional modulation of heart rate by the myelinated vagal efferent pathways. Neurophysiologically, the vagal brake provides a mechanism to support the metabolic requirements for mobilization and communication behaviors. Functionally, the vagal brake, by modulating visceral state, enables the individual to rapidly engage and disengage with objects and other individuals and to promote self-soothing behaviors and calm behavioral states. Thus, withdrawal of the vagal brake is associated with an adaptive stress response and a reinstatement of the vagal brake with recovery.

Withdrawal of the vagal brake will facilitate the recruitment of other neural mechanisms (e.g., excitation of sympathetic or the unmyelinated vagal pathways) and neural chemical mechanisms (e.g., stimulation of the HPA axis) to regulate physiological state. Thus, consistent with the polyvagal theory, if the vagal brake is not functioning or will not serve the survival needs of the organism, then the phylogenetically 'older' systems (e.g., the sympathetic–adrenal system or unmyelinated vagus originating in the dorsal motor nucleus of the vagus) will be recruited to regulate metabolic output to deal with environmental challenges. For example, if the vagal brake is not functioning, there is the potential for greater dependence on the sympathetic excitation of the cardiovascular system.

The Social Engagement System

Mammals have evolved to have a well-defined neurally mediated social engagement system. Embryologically, components of several cranial nerves develop

together (i.e., myelinated vagal pathways from the nucleus ambiguus and special visceral efferent pathways from cranial nerves V, VII, IX, X, and XI) to form the neural substrate of a social engagement system. This system provides the neural gatekeepers for social–emotional interactions. The social engagement system has a control component in the cortex (i.e., upper motor neurons) that regulates brain stem nuclei (i.e., lower motor neurons) to control eyelid opening (e.g., looking), facial muscles (e.g., emotional expression), middle ear muscles (e.g., extracting human voice from background noise), muscles of mastication (e.g., ingestion), laryngeal and pharyngeal muscles (e.g., vocalization and language), and head-turning muscles (e.g., social gesture and orientation). Collectively, these muscles function as filters and control social engagement with the environment. The neural control of these muscles determines social experiences. In addition, the source nuclei (i.e., lower motor neurons) of these nerves, which are located in the brain stem, communicate directly with an inhibitory neural circuit that slows heart rate, lowers blood pressure, and actively reduces arousal to promote calm states consistent with the metabolic demands of growth and restoration of neurophysiological systems. Direct corticobulbar pathways reflect the influence of frontal areas of the cortex (i.e., upper motor neurons) on the regulation of this system. Thus, the social nervous system is intimately related to stress reactivity. In addition, the anatomical structures involved in the social engagement system have neurophysiological interactions with the HPA axis, the neuropeptides of oxytocin and vasopressin, and the immune system. Thus, the social engagement system, as illustrated in Figure 1, provides a theoretical model to explain the interactive and stress-related functions of several physiological systems that have central regulatory components but are expressed in the periphery.

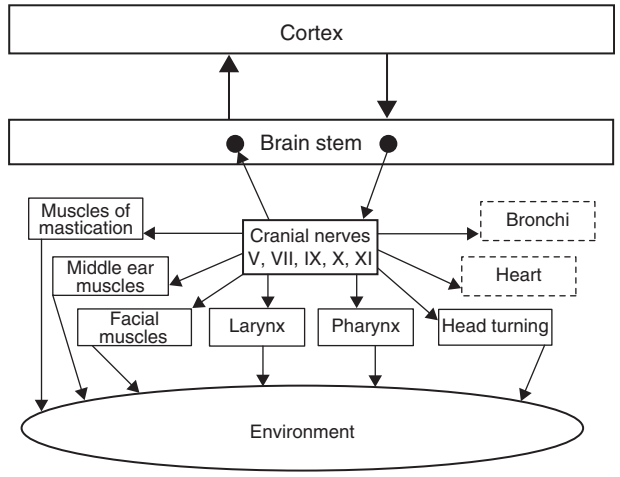

Figure 1 The social engagement system. Social communication is determined by the cortical regulation of medullary nuclei via corticobulbar pathways. The social engagement system consists of a somatomotor component (i.e., special visceral efferent pathways that regulate the striated muscles of the face and head) and a visceromotor component (i.e., the myelinated vagus that regulates the heart and bronchi). Solid rectangles indicate the somatomotor component, and dashed rectangles indicate the visceromotor component.

Vagal Regulation of the HPA Axis

The vagus is involved in the regulation of the HPA axis. Vagal afferents exhibit an inhibitory influence on the HPA axis and reduce cortisol secretion. Studies have demonstrated a covariation between increases in cortisol and decreases in cardiac vagal tone (i.e., measured by the amplitude of respiratory sinus arrhythmia). Thus, there appears to be a coordinated response that functions to promote metabolic activity and mobilization behaviors by withdrawal of vagal tone through the myelinated vagus and increasing both sympathetic activity and activation of the HPA axis. In general, the functioning of the adrenal cortex and the secretion of cortisol appear to be integrated into the mobilization function of the autonomic nervous system by increasing sympathetic activation and circulating catecholamines. These effects suggest that, consistent with the phylogenetic approach described in the polyvagal theory, cortisol secretion may be related to maintenance of mobilization (i.e., the conversion of norepinephrine into epinephrine) for fight-or-flight behaviors and in the recovery from the lactate buildup that may contribute to a functional oxygen debt (i.e., gluconeogenesis).

In addition, reports of dysregulation of the HPA axis, low cortisol, or low cortisol reactivity in disorders such as schizophrenia and posttraumatic stress disorder and also the consequences of neglect and abuse in children may be explained within the context of a phylogenetically ordered response strategy (i.e., polyvagal theory). For example, if mobilization strategies (i.e., fight-or-flight behaviors) are ineffective in removing the individual from the stressor and in modulating the effect of the stress, then the nervous system may degrade to a phylogenetically earlier level of organization. Thus, low cortisol or a hyporesponsive HPA axis may reflect a neural strategy associated with immobilization (i.e., passive avoidance) that would require a reduction in energy resources.

Immune Function

The vagus has been implicated in the immune system. Studies have demonstrated the important role of subdiaphragmatic vagal afferents in conveying information regarding visceral state to the brain.

The studies conceptualize the afferent vagus as providing a signal alerting the central structures regulating immune function. A few studies have described motor pathways via the vagus to the thymus. The link between the vagal function and the immune system is not clear. However, it might be plausible to speculate that the neural mediation of the myelinated vagus, via direct influence on thymus and direct inhibition of the sympathetic nervous system, may trigger a physiological state that would promote immune function. Likewise, withdrawal of vagal tone to the heart, increased sympathetic tone, and the release of cortisol have been associated with suppressed immune function.

Oxytocin/Vasopressin

Because the peripheral influences of oxytocin and vasopressin function through feedback, primarily via afferent vagal pathways, the peripheral effects of these peptides are less clear and may be level dependent or differ as a function of acute versus chronic exposure. For example, it is possible that peripheral vasopressin, by stimulating vagal afferents, may trigger massive vagal responses via the dorsal motor nucleus of the vagus. In support of this speculation, it is known that in humans, peripheral vasopressin, and not oxytocin, is related to the nausea experienced during motion sickness. In addition, systemic vasopressin may induce a baroreceptor-mediated bradycardia and a decrease in plasma concentration of norepinephrine.

Based on the polyvagal theory, the mammalian vagus, with myelinated motor fibers originating in the nucleus ambiguus, provides a system for voluntary engagement with the environment, with special features associated with the prosocial behaviors of communication. Paralleling this evolutionary shift in the vagus is a mammalian modification of the hypothalamic regulation of the dorsal vagal complex (i.e., source nucleus of the afferent and efferent pathways of the unmyelinated vagus originating in the dorsal motor nucleus of the vagus and the nucleus of the solitary tract) via both oxytocin and vasopressin. The advent of specific receptors for oxytocin and vasopressin has increased the range of adaptive functions involving the unmyelinated vagus. In mammals, the dorsal motor nucleus of the vagus, the motor component of the dorsal vagal complex, is sensitive to oxytocin and insensitive to vasopressin. In contrast, the sensory components of the vagus, the nucleus of the solitary tract and area postrema, are most sensitive to vasopressin. Although the nucleus of the solitary tract has receptors for oxytocin, area postrema may not be directly influenced by oxytocin. The differential sensitivity of specific components of the dorsal vagal complex to these two neuropeptides (the differential effects of central and systemic release on visceral function and a potential level dependency) results in a wider range of response options, including maximizing mobilization behaviors and the co-opting of the primitive vagal system associated with immobilization to support 'antistress' functions such as social engagement and growth and restoration.

Phylogenetic Approach to the Study of Stress

The phylogenetic orientation focuses our interest on the parasympathetic neural structures and neurobehavioral systems that we share with or have adapted from our phylogenetic ancestry. First, the three response systems proposed in the polyvagal theory (i.e., cranial nerves to regulate the face, sympathetic–adrenal system to increase metabolic output, and an inhibitory vagal system to decrease metabolic output and promote freezing and defecation) are the products of distinct neurophysiological systems. Second, these distinct neurophysiological systems represent a phylogenetically dependent hierarchy, with the use of cranial nerves to regulate facial expression emerging in mammals (well developed in primates); the sympathetic–adrenal system shared with other vertebrates including reptiles; and the inhibitory vagal system shared with more primitive vertebrates, including amphibians, bony fish, and cartilaginous fish. The three systems represent different phylogenetic stages of neural development. This phylogenetic development starts with a primitive behavioral inhibition system, progresses to a fight-or-flight system, and, in humans (and other primates), culminates in a complex facial gesture and vocalization system. Thus, from a phylogenetic perspective, the nervous system of vertebrates evolved to support a greater range of behaviors and physiological states, including states that are often associated with stress.

The physiological responses associated with stress in mammals do not concisely fit into a single neurophysiological system. Although cortisol has often been labeled as the stress hormone, other systems clearly respond to stress. Moreover, there are situations of severe stress in which cortisol is neither responsive nor at a high level. By expanding the phylogenetic model proposed in the polyvagal theory to emphasize the interactions between parasympathetic mechanisms and the HPA axis, the hyporesponsive HPA can be interpreted as reflecting a primitive passive avoidance system. Thus, several systems with target organs at the periphery contribute to the organism's adaptation to challenge and stress. These neurobiological systems are not solely stress systems. Rather, they are intimately

involved in the dynamic regulation of homeostasis, growth and restoration, metabolism, and mobilization. These systems form a complex set of mutually interacting neurophysiological pathways that communicate via nerves and neurally active chemicals (e.g., neurotransmitters, neuropeptides, and hormones) to cope with survival challenges.

See also: Autonomic Nervous System; Parasympathetic Nervous System; Stress Response and Self-Esteem; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress Response: Sex Differences; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Berlyne DE (1960) *Conflict, Arousal, and Curiosity.* New York: McGraw-Hill.

Bueno LM, Gue MJ, Fargeas M, et al. (1989) Vagally mediated inhibition of acoustic stress-induced cortisol release by orally administered kappa-opioid substances in dogs. *Endocrinology* 124: 1788–1793.

Bulloch K and Pomerantz W (1984) Autonomic nervous system innervation of thymic-related lymphoid tissue in wildtype and nude mice. *Journal of Comparative Neurology* 228: 58–68.

Cannon WB (1928) The mechanism of emotional disturbance of bodily functions. *New England Journal of Medicine* 198: 877–884.

Darrow CW (1943) Physiological and clinical tests of autonomic function and autonomic balance. *Physiological Reviews* 23: 1–36.

Darwin C (1872) *The Expression of Emotions in Man and Animals*. New York: Appleton.

Gray JA (1971) *The Psychology of Fear and Stress*. New York: McGraw-Hill.

Hess WR (1954) *Diencephalon, Autonomic and Extrapyramidal Functions.* New York: Grune & Stratton.

Koch KL, Summy-Long J, Bingaman S, Sperry N, and Stern RM (1990) Vasopressin and oxytocin responses to illusory self-motion and nausea in man. *Journal of Clinical Endocrinology and Metabolism* 71: 1269–1275.

Landgraf R, Mallkinson T, Horn T, et al. (1990) Release of vasopressin and oxytocin by paraventricular stimulation in rats. *American Journal of Physiology* 258: 155–159.

Langley JN (1921) *The Autonomic Nervous System,* vol. 1. Cambridge: Heffer and Sons.

Michelini LC (1994) Vasopressin in the nucleus tractus solitarius: A modulator of baroreceptor reflex control of heart rate. *Brazilian Journal of Medical and Biological Research* 27: 1017–1032.

Morris JL and Nilsson S (1994) The circulatory system. In: Nilsson S and Holmgren S (eds.) *Comparative Physiology and Evolution of the Autonomic Nervous System*, pp. 193–246. Chur, Switzerland: Harwood Academic.

Pavlov IP (1927) *Conditioned Reflexes.* London: Oxford University Press.

Porges SW (1995) Orienting in a defensive world: Mammalian modifications of our evolutionary heritage. A polyvagal theory. *Psychophysiology* 32: 301–318.

Porges SW (1998) Love: An emergent property of the mammalian autonomic nervous system. *Psychoneuroendocrinology* 23: 837–861.

Porges SW, Doussard-Roosevelt JA, Portales AL, and Greenspan SI (1996) Infant regulation of the vagal 'brake' predicts child behavior problems: A psychobiological model of social behavior. *Developmental Psychobiology* 29: 697–712.

Schwaber JS (1986) Neuroanatomical substrates of cardiovascular and emotional–autonomic regulation. In: Magro W, Osswald W, Reis D, and Vanhoutte P (eds.) *Central and Peripheral Mechanisms of Cardiovascular Regulation*, pp. 353–384. New York: Plenum.

Selye H (1956) *The Stress of Life*. New York: McGraw-Hill.

Vanhoutte PM and Levy MN (1979) Cholinergic inhibition of adrenergic neurotransmission in the cardiovascular system. In: Brooks CMcC, Koizumi K, and Sato A (eds.) *Integrative Functions of the Autonomic Nervous System*, pp. 159–176. Tokyo: University of Tokyo Press.

Watkins LR, Goehler LE, Relton JK, et al. (1995) Blockade of interleukin-1 induced hyperthermia by subdiaphragmatic vagotomy: Evidence for vagal mediation of immune-brain communication. *Neuroscience Letters* 183: 27–31.

Stress and Suicide

M Boldrini and V Arango, Columbia University College of Physicians and Surgeons, and New York State Psychiatric Institute, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Suicide is a worldwide public health problem accounting for nearly half of all violent deaths and resulting in 1 million fatalities each year. In 2001, the annual global toll from suicide exceeded the number of deaths by homicide (500 000) and war (230 000) (www.who.int; **Figure 1**). European nations have the highest suicide rates in the world, whereas Central and South American nations have the lowest, although this may be due to the notorious under-reporting that occurs in some countries.

Suicidal behavior is complex and multifaceted, involving genetics and psychosocial and biological factors. The presence of a psychiatric diagnosis, a family history of suicide, childhood abuse, and alcohol and drug abuse all play a central role in suicide. Physical illness and disabling pain can also increase suicide risk. Among psychosocial factors, poverty, unemployment, and social isolation can contribute to hopelessness, as can factors such as the loss of loved ones, breakdown in relationships, and legal or work-related problems.

Psychosocial–Environmental Stress Factors Contributing to Suicidal Behavior

When a change in suicide rates occurs within a short time, it is improbable that it is related to genetic factors, and it raises the possibility that stress is a major contributor to suicide risk. For example, Lithuania, a country with the highest suicide rate in the world (88.2 per 100 000), experienced a sharp decline in suicide rates in the 1980s, as did other Eastern European countries, as the result of the processes of democratization and an effective antialcohol campaign. The collapse of the Soviet Union and the independence of Lithuania and other Baltic states in the early 1990s resulted in a socially unstable situation and a dramatic increase in suicide rates.

Suicide rates increase in situations of extreme stress. During the Holocaust, annual suicide rates in concentration camps were approximately 25 000 per 100 000 and people were at greater risk for suicidal ideation. The high rate of suicide attempts among elderly patients who lived through the Holocaust and now reside in Israel highlights the need for intervention among those who experience traumatic events early in their lives.

Suicide rates are higher in rural than in urban areas in Lithuania, China, and the United States. Suicide predictors in rural China include lack of social support, stressful life events, lower education, poverty, and family disputes. Other predictors are depression, previous suicide attempt, high chronic stress, acute stress, low quality of life, severe interpersonal conflict, a family history of suicidal behavior, and a friend or associate with suicidal behavior. Social isolation, poverty, and unemployment have also been implicated in suicide in rural areas and can occur in the presence or absence of psychopathology.

More than 90% of people who commit suicide or suicide attempters have a diagnosable psychiatric illness. Approximately 60% of all suicides are related to mood disorders, and the rest are related to schizophrenia, alcoholism, substance abuse, and personality disorders. Most psychiatric patients do not commit suicide, so a psychiatric disorder alone is certainly a risk factor but not a sufficient condition to precipitate suicide.

Protective psychosocial factors against suicidal behavior include religious affiliation, parenthood, being married, having reasons for living, fear of social disapproval, moral objections to suicide, and greater coping skills. Psychopathological traits can account for different suicide rates: lower aggression/impulsivity has been identified as a factor for the lower suicide rate in Spain compared to New York.

Antistress factors: Resilience

In studying suicide, one must consider the variables related to resilience in the face of stress or a psychiatric disorder. Children raised in stressful settings who exhibit successful adaptation have good intellectual functioning, effective self-regulation of emotions, positive self-concept, optimism, altruism, the ability to convert traumatic helplessness into learned helpfulness, and an active coping style in the face of stressors. Data on adult resilience come from studies of men in combat and show the importance of bonding with a group sharing a mission and having the capacity to tolerate high levels of fear and still perform effectively. Glucocorticoids and dehydroepiandrosterone (DHEA) are secreted in response to stress. Cortisol has regulatory effects on the hippocampus, amygdala, and prefrontal cortex and facilitates the encoding of emotion-related

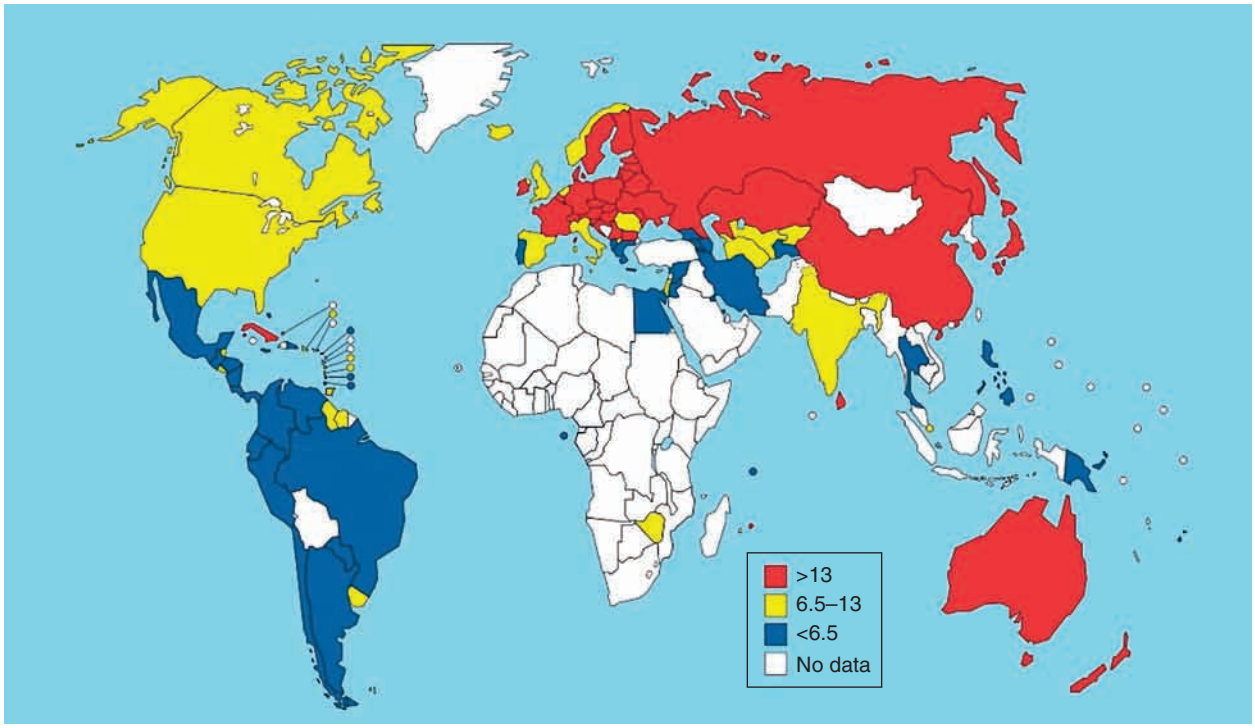

Figure 1 Map of suicide rates per 100 000 as of 2002, the most recent year available.

memory. DHEA promotes psychological resilience and there is a negative correlation between DHEA and the severity of posttraumatic stress disorder (PTSD). Also, several studies have found a negative correlation between plasma DHEA levels and depressive symptoms.

Animal Models

The first nonhuman primate study of early adversity used animals raised in social impoverishment with same-aged peers rather than with their mothers (Table 1). Peer-reared monkeys have lower serotonergic activity than maternally raised monkeys, with the defect persisting into adulthood and being associated with greater aggression and impulsivity. Appropriate parental input is critically important for normative development of the hypothalamic–pituitary–adrenal (HPA) axis and neurotransmitter system function. Primates removed from their mothers and deprived of early participation in complex social groups later exhibit aberrations in neurochemical function, neuroendocrine stress axis activity, and many aspects of social behavior. However, individual variation in susceptibility to environmental manipulation remains. A gene–environment interaction has been demonstrated in the rhesus macaque between a serotonin transporter gene promoter polymorphism and early rearing experience. Male monkeys raised without their mothers showed a monoamine oxidase A (MAOA) genotype–environment interaction effect on competitive behavior, with neither MAOA gene promoter polymorphism nor early rearing experience alone being sufficient to influence behavior, confirming that genetic susceptibility or vulnerability is critically dependent on environmental modulation.

Early Life Stressful Experiences

A history of child abuse is associated with a greater risk for depression and suicidal behavior in adult life independent of impulsivity, aggression, and borderline personality disorder. Risk for suicidal behavior in families of probands with mood disorders appears related to childhood abuse in probands, together with early onset of mood disorders and aggressive/impulsive traits. Individuals with a greater genetic risk for mood disorders and suicide are more likely to respond to stress by developing a depressive episode or exhibiting suicidal behavior.

There is a significant interaction between childhood maltreatment and low MAOA activity-associated alleles in modulating the risk for antisocial behavior, aggression, and violence. Low MAOA activity was a

Table 1 Early life environmental adversity and gene–environment interaction: effects on behavior

Study[a]	*Species*	*Test*	*Stress situation*	*Consequent alteration*
1	Monkeys	Early adversity	Social impoverishment, peer-reared	Lower serotonergic activity
2	Monkeys	Early adversity	Social impoverishment, peer-reared	Neurochemical function, neuroendocrine stress axis activity, social behavior
3	Monkeys	Early adversity	Peer-reared	Lower serotonergic activity, greater aggression and impulsivity
4	Human	Early adversity	History of child abuse	Risk for depression and suicidal behavior in adult
5, 6	Human	Early adversity	Child abuse	Suicidal behavior in adult
7	Human	Early adversity	Sexual abuse	Suicidal behavior in adult
8–10	Monkeys	Early adversity + genotype	Rearing adverse experience	Behavioral changes in serotonin transporter gene promoter polymorphism carriers
11, 12	Human	Early adversity + genotype	Childhood maltreatment	Antisocial behavior, aggressiveness, and violence in low-activity MAOA alleles carriers
13	Male human	Early adversity + genotype	History of abuse before 15 years	Higher impulsivity in low-activity MAOA alleles carriers
14	Male monkeys	Environmental modulation + genotype	Rearing adverse experience	Impaired social competition in low-activity MAOA gene promoter polymorphism carriers

[a]Studies: 1, Harlow HF and Harlow MK (1965) *The Affectional Systems.* New York: Academic Press; 2, Kraemer GW (1985) *Effects of Differences in Early Social Experience on Primate Neurobiological–Behavioral Development.* New York: Academic Press; 3, Higley JD, Thompson WW, Champoux M, et al. (1993) Paternal and maternal genetic and environmental contributions to cerebrospinal fluid monoamine metabolites in rhesus monkeys (*Macaca mulatta*). *Archives of General Psychiatry* 50: 615–623; 4, Young EA, Abelson JL, Curtis GC and Nesse RM (1997) Childhood adversity and vulnerability to mood and anxiety disorders. *Depression and Anxiety* 5: 66–72; 5, Brodsky BS, Oquendo MA, Ellis SP, Haas GL, Malone KM, and Mann JJ (2001) The relationship of childhood abuse to impulsivity and suicidal behavior in adults with major depression. *American Journal of Psychiatry* 158: 1871–1877; 6, Mann JJ, Bortinger J, Oquendo MA, Currier D, Li S, and Brent DA (2005) Family history of suicidal behavior and mood disorders in probands with mood disorders. *American Journal of Psychiatry* 162: 1672–1679; 7, Soloff PH, Lynch KG, and Kelly TM (2002) Childhood abuse as a risk factor for suicidal behavior in borderline personality disorder. *Journal of Personality Disorders* 16: 201–214; 8, Bennett AJ, Lesch KP, Heils A, et al. (2002) Early experience and serotonin transporter gene variation interact to influence primate CNS function. *Molecular Psychiatry* 7: 118–122; 9, Champoux M, Bennett A, Shannon C, Higley JD, Lesch KP, and Suomi SJ (2002) Serotonin transporter gene polymorphism, differential early rearing, and behavior in rhesus monkey neonates. *Molecular Psychiatry* 7: 1058–1063; 10, Barr CS, Newman TK, Becker ML, et al. (2003) Serotonin transporter gene variation is associated with alcohol sensitivity in rhesus macaques exposed to early-life stress. *Alcoholism, Clinical and Experimental Research* 27: 812–817; 11, Caspi A, McClay J, Moffitt TE, et al. (2002) Role of genotype in the cycle of violence in maltreated children. *Science* 297: 851–854; 12, Foley DL, Eaves LJ, Wormley B, et al. (2004) Childhood adversity, monoamine oxidase a genotype, and risk for conduct disorder. *Archives of General Psychiatry* 61: 738–744; 13, Huang YY, Battistuzzi C, Oquendo MA, et al. (2004) Human 5-HT1A receptor C(-1019)G polymorphism and psychopathology. *International Journal of Neuropsychopharmacology* 7: 441–451; 14, Newman TK, Syagailo YV, Barr CS, et al. (2005) Monoamine oxidase A gene promoter variation and rearing experience influences aggressive behavior in rhesus monkeys. *Biological Psychiatry* 57: 167–172.

risk factor only when combined with an adverse childhood environment, a finding replicated in a community-based sample. We reported that the lower expression allele of MAOA gene was associated with a history of early abuse (before 15 years of age) and with higher impulsivity in males, suggesting that the genetic variation may spare females of the effects of early abuse experiences on impulsive traits in adulthood.

Maltreatment, as defined by maternal rejection, loss of a primary caregiver, harsh discipline, physical or sexual abuse, interparental violence, parental neglect, or inconsistent discipline, is a risk factor for conduct disorder. Adverse parenting can involve neglect, but physical or sexual abuse has been more commonly studied for its impact on suicidal behavior. Severity of childhood sexual, but not physical, abuse predicted adult suicidal behavior independent of other known risk factors. The odds of a patient who was sexually abused as a child attempting suicide in adulthood are more than 10 times those of a patient who was never sexually abused.

Hypothalamic–Pituitary–Adrenal Axis

A great deal of work on depression and suicide has focused on the stress network, the HPA axis. In fact, hypercortisolism is a well-documented abnormality in mood disorders. People who commit suicide were found to have heavier adrenal glands than those who do not commit suicide, higher levels of corticotropin-releasing hormone (CRH) in the cerebrospinal fluid

Table 2 Cortisol and the hypothalamic–pituitary–adrenal (HPA) axis in stress responses and major depression

Study[a]	*Species*	*Test*	*Condition*	*Changes*
1–3	Human	Stress responses	Stress, major depression, and suicide	Cortisol and HPA axis
4–6	Animal	Neurobiological consequences of early neglect	Maternal separation	HPA overactivity, neuropeptide Y and CRH alterations
7, 8	Human	Anti-glucocorticoid effects of DHEA	PTSD and MDD	Negative correlation of DHEA with adrenal activation in PTSD and depressive symptoms in MDD
9	Human	Neurobiological consequences of abuse	Abuse and neglect	HPA overactivity, neuropeptide Y and CRH alterations
10	Rat	Early stress responses	Early life stress	Neuropeptide Y and CRH alterations

[a]Studies: 1, Bunney WE Jr., Fawcett JA, Davis JM, and Gifford S (1969) Further evaluation of urinary 17-hydroxycorticosteroids in suicidal patients. *Archives of General Psychiatry* 21: 138–150; 2, Carroll BJ, Greden JF, and Feinberg M (1981) Suicide, neuroendocrine dysfunction and CSF 5-H1AA concentrations in depression. In: Angris B (ed.) *Recent Advances in Neuropsychopharmacology: Selected Papers from the 12th Congress of the Collegium Internationale Neuro-Psychopharmacologicum, Ghoteborg, Sweden, 22–26 June 1980*, pp. 307–313. New York: Pergamon; 3, Nemeroff CB, Owens MJ, Bissette G, Andorn AC, and Stanley M (1988) Reduced corticotropin releasing factor binding sites in the frontal cortex of suicide victims. *Archives of General Psychiatry* 45: 577–579; 4, Hofer MA (1996) On the nature and consequences of early loss. *Psychosomatic Medicine* 58: 570–581; 5, Viau V, Sharma S, and Meaney MJ (1996) Changes in plasma adrenocorticotropin, corticosterone, corticosteroid-binding globulin, and hippocampal glucocorticoid receptor occupancy/translocation in rat pups in response to stress. *Journal of Neuroendocrinology* 8: 1–8; 6, Heim C and Nemeroff CB (2001) The role of childhood trauma in the neurobiology of mood and anxiety disorders: Preclinical and clinical studies. *Biological Psychiatry* 49: 1023–1039; 7, Goodyer IM, Herbert J, and Altham PM (1998) Adrenal steroid secretion and major depression in 8- to 16-year-olds. III: Influence of cortisol/DHEA ratio at presentation on subsequent rates of disappointing life events and persistent major depression. *Psychological Medicine* 28: 265–273; 8, Wolkowitz OM, Reus VI, Keebler A, et al. (1999) Double-blind treatment of major depression with dehydroepiandrosterone. *American Journal of Psychiatry* 156: 646–649; 9, Heim C, Newport DJ, Bonsall R, Miller AH, and Nemeroff CB (2001) Altered pituitary–adrenal axis responses to provocative challenge tests in adult survivors of childhood abuse. *American Journal of Psychiatry* 158: 575–581; 10, Husum H and Mathe AA (2002) Early life stress changes concentrations of neuropeptide Y and corticotropin-releasing hormone in adult rat brain. Lithium treatment modifies these changes. *Neuropsychopharmacology* 27: 756–764. CRH, corticotropin-releasing hormone; DHEA, dehydroepiandrosterone; MDD, major depressive disorder; PTSD, posttraumatic stress disorder.

(CSF), and fewer CRH receptors in the cortex. Patients with major depression also show adrenal gland hypertrophy by computerized tomography and more CRH-synthesizing neurons and mRNA in the paraventricular hypothalamic nucleus. The overactivity of the HPA axis and alterations in neuropeptide systems have also been reported in abused populations, as also demonstrated in animal studies of maternal separation. Early life stress-induced changes of neuropeptide Y and CRH in adult rat brain can be modified by lithium treatment (Table 2).

Extrahypothalamic CRH System

CRH neurons are localized in central nervous system (CNS) areas other than the hypothalamus in several species, including nonhuman primates and humans. We were the first to describe a group of CRH immunoreactive neurons in the pedunculopontine tegmental nucleus and laterodorsal complex in the brain stem of humans that also expressed CRH mRNA. These neurons are located close to the locus coeruleus (LC), the dorsal raphe nucleus (DRN) and median raphe nucleus, and the substantia nigra – all monoaminergic nuclei that are differentially innervated by CRH axons. The noradrenergic LC receives the strongest CRH innervation, followed by the serotonergic DRN, with the dopaminergic substantia nigra having a very light innervation. The specificity of this anatomical localization may dictate the differential influence that CRH may exert on a specific neurotransmitter system.

Noradrenergic System

People who commit suicide have less norepinephrine (NE) in the brain stem but more α_2-adrenergic receptors, perhaps secondary to lower NE. In depressed suicides fewer LC neurons were found compared to controls (Table 3). There are conflicting results regarding tyrosine hydroxylase (TH; the rate-limiting enzyme in the biosynthesis of NE), which has been reported to be higher by some and lower by others. TH increases were reported when increased NE release led to transmitter depletion. Therefore, TH immunoreactivity could be state dependent, which would explain the conflicting results in the literature. In the prefrontal cortex, suicides have higher β-adrenergic binding, higher levels of NE, and higher α_1-adrenergic binding in specific cortical layers.

Table 3 Stress and suicide-related brain changes

Study[a]	*Species*	*Condition*	*Changes*
1	Rat	Chronic restraint stress	Apical dendrites of CA3 pyramidal neurons atrophy
2	Rat	Exposure to excess glucocorticoids	Decreased dendritic length and number of apical dendritic branch points
3	Human	Suicide	Serotonin transporter and 5-HT_{1A} binding changes localized to the ventral prefrontal cortex
4	Human	Suicide	Fewer locus coeruleus neurons
5–12	Human	Depression and anxiety	Fewer neurons and/or glia, lower size and/or cell density and dendritic length in prefrontal cortex and/or hippocampus
13	Human	Impulsive/violent subjects	Lower frontal and midbrain serotonin transporter binding
14	Rodents	Chronic stress	Reduced 5-HT_{1A} receptor protein and mRNA in hippocampus
15	Rat	Behavioral stress-induced depression	Cell death and dendritic shrinkage in the hippocampus
16	Human	Suicide	Fewer SERT mRNA expressing neurons in dorsal raphe
17	Human	Depression	Lower orbital frontal cortex serotonin 5-HT_{1A} receptor binding in higher lifetime aggression subjects
18	Human	Psychological trauma	Vulnerability to develop psychopathology in subjects with smaller hippocampus
19	Rat	Immobilization stress	Increased tryptophan hydroxylase
20	Human	Suicide	Increased tryptophan hydroxylase in dorsal raphe
21	Human	Suicide	Higher mRNA levels of tryptophan hydroxylase
22	Human	Suicide	More (and denser) serotonergic neurons in dorsal raphe

[a]Studies: 1, Watanabe Y, Gould E, and McEwen BS (1992) Stress induces atrophy of apical dendrites of hippocampal CA3 pyramidal neurons. *Brain Research* 588: 341–345; 2, Woolley CS, Gould E, and McEwen BS (1990) Exposure to excess glucocorticoids alters dendritic morphology of adult hippocampal pyramidal neurons. *Brain Research* 531: 225–231; 3, Arango V, Underwood MD, Gubbi AV, and Mann JJ (1995) Localized alterations in pre- and postsynaptic serotonin binding sites in the ventrolateral prefrontal cortex of suicide victims. *Brain Research* 688: 121–133; 4, Arango V, Underwood MD, and Mann JJ (1996) Fewer pigmented locus coeruleus neurons in suicide victims: Preliminary results. *Biological Psychiatry* 39: 112–120; 5, Sheline YI, Wang PW, Gado MH, Csernansky JG, and Vannier MW (1996) Hippocampal atrophy in recurrent major depression. *Proceedings of the National Academy of Sciences of the United States of America* 93: 3908–3913; 6, Sheline YI, Sanghavi M, Mintun MA, and Gado MH (1999) Depression duration but not age predicts hippocampal volume loss in medically healthy women with recurrent major depression. *Journal of Neuroscience* 19: 5034–5043; 7, Benes FM, Kwok EW, Vincent SL, and Todtenkopf MS (1998) A reduction of nonpyramidal cells in sector CA2 of schizophrenics and manic depressives. *Biological Psychiatry* 44: 88–97; 8, Öngür D, Drevets WC, and Price JL (1998) Glial reduction in the subgenual prefrontal cortex in mood disorders. *Proceedings of the National Academy of Sciences of the United States of America* 95: 13290–13295; 9, Rajkowska G, Miguel-Hidalgo JJ, Wei J, et al. (1999) Morphometric evidence for neuronal and glial prefrontal cell pathology in major depression. *Biological Psychiatry* 45: 1085–1098; 10, Rajkowska G (2000) Postmortem studies in mood disorders indicate altered numbers of neurons and glial cells. *Biological Psychiatry* 48: 766–777; 11, Sapolsky RM (2000) The possibility of neurotoxicity in the hippocampus in major depression: A primer on neuron death. *Biological Psychiatry* 48: 755–765; 12, Arango V, Underwood MD, and Mann JJ (2002) Serotonin brain circuits involved in major depression and suicide. *Progress in Brain Research* 136: 443–453; 13, Tiihonen J, Kuikka JT, Bergström KA, et al. (1997) Single-photon emission tomography imaging of monoamine transporters in impulsive violent behaviour. *European Journal of Nuclear Medicine* 24: 1253–1260; 14, Lopez JF, Chalmers DT, Little KY, and Watson SJ (1998) A. E. Bennett Research Award. Regulation of serotonin1A, glucocorticoid, and mineralocorticoid receptor in rat and human hippocampus: Implications for the neurobiology of depression. *Biological Psychiatry* 43: 547–573; 15, McEwen BS (1999) Stress and hippocampal plasticity. *Annual Review of Neuroscience* 22: 105–122; 16, Arango V, Underwood MD, Boldrini M, et al. (2001) Serotonin 1A receptors, serotonin transporter binding and serotonin transporter mRNA expression in the brainstem of depressed suicide victims. *Neuropsychopharmacology* 25: 892–903; 17, Parsey RV, Oquendo MA, Simpson NK, et al. (2002) Effects of Sex, age, and aggressive trails in man on brain serotonin 5-HT_{1A} receptor binding potential measured by PET using [C-11] WAY-100635, *Brain Research* 954(2): 173–182. 18, Gilbertson MW, Shenton ME, Ciszewski A, et al. (2002) Smaller hippocampal volume predicts pathologic vulnerability to psychological trauma. *Nature Neuroscience* 5: 1242–1247; 19, Chamas FM, Underwood MD, Arango V, et al. (2004) Immobilization stress elevates tryptophan hydroxylase mRNA and protein in the rat raphe nuclei. *Biological Psychiatry* 55: 278–283; 20, Boldrini M, Underwood MD, Mann JJ, and Arango V (2005) More tryptophan hydroxylase in the brainstem dorsal raphe nucleus in depressed suicides. *Brain Research* 1041: 19–28; 21, Bach-Mizrachi H, Underwood MD, Kassir SA, et al. (2006) Neuronal tryptophan hydroxylase mRNA expression in the human dorsal and median raphe nuclei: Major depression and suicide. *Neuropsychopharmacology* 31: 814–824; 22, Underwood MD, Khaibulina AA, Ellis SP, et al. (1999) Morphometry of the dorsal raphe nucleus serotonergic neurons in suicide victims. *Biological Psychiatry* 46: 473–483. SERT, serotonin transport protein.

These noradrenergic changes in suicide may be stress related. In fact, increased noradrenergic and cortisol-dependent stress responses are reported in depression. Moreover, exposure to adverse childhood experiences can result in exaggerated sympathetic responses to stress in adulthood and such an effect would further deplete NE. Therefore, TH upregulation in the LC of suicide victims could potentially be a temporary response of the noradrenergic system to excessive NE release associated with the stress of the suicidal ideation. Animal studies show that both prolonged and brief maternal separation that occurs early produces persistent changes in tissue levels of brain monoamines and their metabolites in middle-aged female rats.

There is no correlation between CSF levels of NE, epinephrine, or their metabolites and suicidal behavior. Severe anxiety or agitation is associated with NE overactivity and also higher suicide risk and overactivity of the HPA axis. Urinary levels of 3-methoxy-4-hydroxyphenylglycol are higher in inpatients 24 h after a suicide attempt, indicating increased NE turnover following an attempt.

Children with PTSD have more suicidal ideation and attempts, excrete more NE and dopamine, and have more urinary free cortisol than overanxious children and healthy controls. Levels are correlated with duration and severity of illness.

In summary, studies on people who commit suicide and on patients with PTSD indicate that these stressed patients have higher NE activity, a finding not present in unstressed patients or animals. Increased noradrenergic activity together with fewer LC noradrenergic neurons and tone lead to NE depletion, increasing hopelessness. Thus, NE hyperactivity is state/stress related and probably has a precipitating role in suicidal behavior.

Serotonergic System

The HPA axis has a bidirectional relationship with the serotonergic system. There is a dose-dependent increase in tryptophan hydroxylase (TPH) protein and mRNA in rats subjected to immobilization stress. Higher TPH immunoreactivity and higher mRNA levels of the brain-expressed TPH gene, TPH2, were found in the DRN of depressed suicides compared to controls. In the prefrontal cortex of suicides, TPH immunoreactivity was also higher, as were TPH2 mRNA levels (Table 3). This may be a stress effect because dexamethasone and stress raise TPH immunoreactivity in the raphe nuclei.

Genetic studies on the TPH2 gene report an association with suicide but not with schizophrenia or schizophrenic subjects who died by suicide. Other investigators have searched for associations between variants of the TPH2 gene and psychiatric disorders and suicide. However, only one functional ARG441HIS missense variant was identified in the coding region of the TPH2 gene and found to be present in approximately 10% of a population of elder major depressives, and it was not detected in a cohort affected with bipolar disorder. When expressed in PC12 cells, this genetic variant was found to decrease serotonin levels by 80% compared to wild type, demonstrating that this single nucleotide polymorphism (SNP) is a loss-of-function mutation. This SNP appears to represent a rare functional polymorphism since it has not been found in several other populations studied.

The TPH1 gene is not expressed in the brain as abundantly as in the periphery, and a polymorphism (A779C) has been associated with suicide-related behavior. This association was found in a meta-analysis that also suggested that this specific allele has a dose-dependent effect on the risk of suicidal behavior, at least in Caucasians. Courtet suggested that TPH may be a quantitative risk factor because a greater effect of the gene leads to more pronounced serotonergic dysfunction with higher levels of anger and more severe suicidal acts. TPH1 alleles were associated with a blunted prolactin response to fenfluramine and with low CSF levels of 5-hydroxyindolacetic acid in healthy volunteers, reflecting possible serotonergic alterations outside the nervous system. Therefore, in stress situations TPH expression might be increased, but in psychiatric populations there might be a genetic impairment of the gene in the brain, which may lead to altered serotonergic responses in the CNS that may facilitate suicidal acts. Outside the CNS, TPH1 alterations may impair autonomic responses to fear and stress, resulting in a failure to cope with these situations and consequent proportional anger and more severe suicidal acts.

Chronic stress reduces 5-HT_{1A} receptor protein and mRNA in the hippocampus of rodents, an effect prevented by adrenalectomy (Table 3). 5-HT_{1A} receptor knockout mice are anxious, but when the receptor is rescued during the perinatal period, rather than during early adulthood, the mice do not exhibit anxious behavior. This reversal may be related to some structural changes in the brain that can impair resilience to stress. Serotonin and the 5-HT_{1A} receptor have an important role in development, including neural growth, cellular homeostasis, and antiapoptosis. Gene polymorphisms of the human 5-HT_{1A} receptor have been studied in psychiatric populations. The 5-HT_{1A} receptor promoter region C(-1019)G polymorphism was found to have a significant association with schizophrenia and substance use disorder but not with major depression, bipolar disorder, alcoholism, panic disorder, or suicide attempt. Therefore, it seems that the mechanism potentially responsible for 5-HT_{1A} receptor changes in suicide is not a genetic alteration but, rather, an expression alteration perhaps related to plasticity responses to stress.

Serotonin and Cell Proliferation and Survival

Stimulation of 5-HT_{1A} receptor is necessary for the proliferative response of cells treated with 5-HT. 5-HT_{1A} receptor knockout mice do not show increased neurogenesis and behavioral changes after

treatment with antidepressant drugs. In addition, behavioral stress-induced depression in animals is associated with cell death and dendritic shrinkage that can be reversed in the hippocampus by treatment with antidepressants and neurotrophins. It has been proposed that cell survival can be mediated by the 5-HT_{1A} receptor via inhibition of caspase-3 and other proapoptotic proteins. Remarkably, one study suggests that the expression of Akt1, Bcl-2, and other antiapoptotic proteins is decreased in the cortex of schizophrenics, and that the antimania drug Li^+ increases the number of Bcl-2 immunoreactive cells in several layers of the frontal cortex, dentate gyrus, and striatum of rats.

Dopaminergic System

Significant strides have been made in understanding how the brain regulates reward and motivation (hedonia), response to fear (effective behaviors despite fear), and social behaviors (altruism, and bonding and teamwork). The neural mechanisms that mediate these functions are relevant to how an individual responds to extreme stress and may account for the resilience to suicidal thoughts. The mesolimbic dopamine pathways, originating in the ventral tegmental area (VTA), are involved in reward and hedonic tone. Other regions involved in dopamine signaling are the dorsal and ventral striatum (i.e., nucleus accumbens) and the amygdala. Functional interactions between dopamine, glutamate, *N*-methyl-D-aspartate (NMDA), and dopamine receptors are critical to the proper functioning of reward circuits. Neurons of the orbital frontal cortex, a target of the VTA, have the ability to discriminate among rewards based on their motivational value.

The dopaminergic system appears abnormal in depression, and it may be associated with suicidal behavior. A low level of homovanillic acid in CSF is found in depressed suicide attempters, and this abnormality is specifically associated with the attempter status. Preclinical studies indicate that there may be a genetically determined mesocortical and mesoaccumbens dopamine response to stress that relates to learned helplessness. The severity of depression was shown to correlate with the rewarding effects of amphetamine, via a mechanism possibly related to synaptic dopamine depletion with upregulation of dopamine receptors.

No alteration of mRNA for dopamine D1 and D2 receptors was found in the caudate nucleus of suicide victims, nor dopamine D4 receptor binding. Our studies have found no differences in D1, D2, or dopamine transporter in suicides compared to controls. The role of dopamine in suicide is not compelling, and low dopaminergic tone seems to be associated with depression.

Sex Hormones

Another mechanism of defective versus resilient response to stress might be related to sex hormones. Aggression is the aspect of human behavior most often linked to testosterone concentrations, and suicide is more frequent in men. Psychological stress is associated with decreases in testosterone levels, and soldiers participating in physically and psychologically stressful training exercises show a lowering of testosterone levels. Also, lower plasma testosterone is associated with depression in men. On the other hand, estrogen (as well as progesterone) produces a relative resistance to glucocorticoid feedback. These data suggest that physiological doses of estradiol are inhibitory to stress responsiveness. These hormonal differences might play a role in the different rates of suicide in men and women.

Neuroanatomical Studies

Brain imaging and postmortem studies in depressed and anxious individuals have demonstrated atrophy of the prefrontal cortex and hippocampus, as indicated by fewer neurons and glia, lower size and/or cell density, and reduced dendritic length. A study of monozygotic twins discordant for trauma exposure found evidence that smaller hippocampal volume may constitute a risk factor for the development of stress-related psychopathology (Table 3).

The cognitive framework for decision making is relevant to suicidal behavior and volitional acts. Both implicate the ventromedial prefrontal cortex, as well as the amygdala and somatosensory and insular cortex, and the underlying white matter. There are reports of changes in amygdala volume in depression. The amygdala also modulates conditioned responses to rewarding stimuli, being part of a circuit involving subiculum, bed nucleus of the stria terminalis, nucleus accumbens, and medial prefrontal cortex. The cyclic adenosine monophosphate (cAMP) pathway and cAMP response element binding protein in the amygdala promote both aversive and rewarding associations. These neural networks establish the emotional value of a reward memory, as well as its strength and persistence. The molecular basis for such plasticity is just beginning to be understood.

During the expression of fear-related behaviors, the lateral amygdala engages the central nucleus, which projects to areas of the hypothalamus and brain

stem that mediate the autonomic, endocrine, and behavioral responses associated with fear. Failure to achieve an adequate level of activation of the medial prefrontal cortex after memory extinction might lead to persistent fear responses. Individuals with the capacity to function well after experiencing states of high fear may have potent medial prefrontal cortex inhibition of amygdala responsiveness.

The ventral prefrontal cortex is believed to play a role in behavioral inhibition and in the expression of emotions. Damage to the ventromedial prefrontal cortex results in defects of social behavior, including a failure to observe social conventions and a tendency to make poor decisions. The orbitofrontal cortex has been implicated in the neurobiology of suicide. This region is activated during response inhibition and decision making. Therefore, the prefrontal cortex provides insight into consequences of future actions, and its dysfunction might contribute to a reduced judgment of the consequence of a final act, facilitating the possibility of acting on suicidal thoughts in these patients. Impaired decision making and impulsive behavior are associated with suicidality. The orbital prefrontal and anterior cingulate cortices are also part of an anatomical circuitry of emotion and aggression. The activation of these regions is part of an effort to suppress emotions and enrolls an inhibitory projection to the amygdala. Imaging studies using positron emission tomography (PET) also indicate an inverse relationship between the activity of the prefrontal cortex and that of the amygdala.

A single photon emission computed tomography study of borderline or antisocial personality disorder patients found lower regional cerebral blood flow in brain areas associated with self-harm or harm to others. These regions included the right lateral temporal cortex (BA21), the right frontopolar cortex (BA10), and the right ventrolateral prefrontal cortex (BA47). Females with impulsive borderline personality disorder who attempted suicide showed PET hypometabolism in the medial orbital frontal cortex bilaterally, including BA9–BA11, compared to healthy controls. Lower frontal and midbrain serotonin transporter binding was observed in impulsive/violent subjects compared to controls, and an inverse correlation was found between 5-HT_{1A} binding and 5-HT_{2A} binding in the orbitofrontal cortex and lifetime aggression. Inferior frontal white matter microstructural defects, as shown by lower fractional anisotropy, have been found in association with aggressive impulsivity in males with schizophrenia. In suicide victims, the changes in serotonin transport protein (SERT) and 5-HT_{1A} binding are localized to the ventral prefrontal cortex, whereas SERT changes in depression are widespread (Table 3). Therefore, the implication of ventral prefrontal cortex in suicide appears to be related to behavioral control and not to depression.

Conclusions

Only a small fraction of patients with suicidal thoughts attempt suicide, few attempts prove to be fatal, and risk factors for suicide attempts (e.g., younger and female) and completions (e.g., older and male) differ markedly. The assessment of suicide risk should include the evaluation of the socio/environmental status and whether protective/precipitating factors exist, including characteristics of resilience.

Genes and environment interact in stress responses leading to phenotype expression (suicidal behavior). The interaction between genes and environment has long been acknowledged to account for significant individual variation in susceptibility or resilience to risk factors. This interaction seems to change the function of the neurotransmitters systems. One of the environmental factors thought to play a critical role in leading to the risk for complex behavioral disorders is early experience, particularly during infancy and early childhood. Stress seems to have an important role in the expression of psychobiological factors that contribute to suicide. Suicide may be the result of not only gene–gene interactions but also gene–environmental interactions, but little attention has been paid to this possible interplay.

Psychosocial factors are often associated with and not independent from psychiatric illness. Psychiatric disorders can provoke job or relationship failure or prevent the beginning of stable employment or stable relationships. Moreover, psychiatric illness and psychosocial adversity can combine to increase a person's hopelessness and thereby potentially increase the risk for suicidal behavior. It is difficult to separate the impact of psychosocial adversity from that of psychiatric illness.

The concept of allostatic load as a cumulative measure of physiological dysregulation over multiple systems has proven to be useful as a predictor of functional decline in elderly men and women. This concept could be used in the study of suicide in order to identify cumulative neurobiological/stress-related abnormalities that may be prognostic of poor outcome.

See also: Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Dopamine; Dopamine in Perspective; Norepinephrine: CNS Pathways and Neurophysiology; Posttraumatic Stress Disorder: Overview; Posttraumatic Stress Disorder: Neurobiology; Posttraumatic Stress Disorder as an Emotional Disorder; Prefrontal Cortex: Structure and Anatomy;

Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Stress Response: Genetic Consequences; Stress Response: Sex Differences; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load; Suicide.

Further Reading

Arango V, Ernsberger P, Marzuk PM, et al. (1990) Autoradiographic demonstration of increased serotonin 5-HT2 and beta-adrenergic receptor binding sites in the brain of suicide victims. *Archives of General Psychiatry* 47: 1038–1047.

Arato M, Banki CM, Bissette G, and Nemeroff CB (1989) Elevated CSF CRF in suicide victims. *Biological Psychiatry* 25: 355–359.

Baca-Garcia E, Oquendo MA, Saiz-Ruiz J, Mann JJ, and de Leon JA (2006) A pilot study on differences in aggression in New York City and Madrid, Spain, and their possible impact on suicidal behavior. *Journal of Clinical Psychiatry* 67: 375–380.

Barak Y, Aizenberg D, Szor H, Swartz M, Maor R, and Knobler HY (2005) Increased risk of attempted suicide among aging holocaust survivors. *American Journal of Geriatric Psychiatry* 13: 701–704.

Caspi A, Sugden K, Moffitt TE, et al. (2003) Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* 301: 386–389.

Caspi A, Taylor A, Moffitt TE, and Plomin R (2000) Neighborhood deprivation affects children's mental health: Environmental risks identified in a genetic design. *Psychological Science* 11: 338–342.

Chamas F, Serova L, and Sabban EL (1999) Tryptophan hydroxylase mRNA levels are elevated by repeated immobilization stress in rat raphe nuclei but not in pineal gland. *Neuroscience Letters* 267: 157–160.

Coryell W, Young E, and Carroll B (2006) Hyperactivity of the hypothalamic–pituitary–adrenal axis and mortality in major depressive disorder. *Psychiatry Research* 142: 99–104.

Hastings RS, Parsey RV, Oquendo MA, Arango V, and Mann JJ (1994) Volumetric analysis of the prefrontal cortex, amygdala, and hippocampus in major depression. *Neuropsychopharmacology* 29: 952–959

Higley JD, Mehlman PT, Taub D, et al. (1994) Are stress hormones and serotonin related to aggression in primates? Reply. *Archives of General Psychiatry* 51: 73

Mann JJ (2003) Neurobiology of suicidal behaviour. *Nature Reviews Neuroscience* 4: 819–828.

Mann JJ and Currier D (2006) Effects of genes and stress on the neurobiology of depression.*International Review of Neurobiology* 73: 153–189.

Mann JJ, Hung YY, Underwood MD, et al. (2000) A serotonin transport gene promotor polymorphism (5-HTTLPR) and prefrontal cortical binding in major depression and suicide. *Archives of General Psychiatry* 57: 729–738.

Nemeroff CB, Krishnan KR, Reed D, Leder R, Beam C, and Dunnick NR (1992) Adrenal gland enlargement in major depression. A computed tomographic study. *Archives of General Psychiatry* 49: 384–387

Oquendo MA, Russo SA, Underwood MD, et al. (2006) Higher postmortem prefrontal 5-HT2A receptor binding correlates with lifetime aggression in suicide. *Biological Psychiatry* 59: 235–243.

Ordway GA, Widdowson PS, Smith KS, and Halaris A (1994) Agonist binding to α_2-adrenoceptors is elevated in the locus coeruleus from victims of suicide. *Journal of Neurochemistry* 63: 617–624.

Parsey RV, Oquendo MA, Simpson NK, et al. (2002) Effects of Sex, age, and aggressive trails in man on brain serotonin 5-HT_{1A} receptor binding potential measured by PET using [C-11] WAY-100635. *Brain Research* 954(2): 173–182.

Raadsheer FC, Hoogendijk WJG, Stam FC, Tilders FJH, and Swaab DF (1994) Increased numbers of corticotropin-releasing hormone expressing neurons in the hypothalamic paraventricular nucleus of depressed patients. *Neuroendocrinology* 60: 436–444.

Richardson GE (2002) The metatheory of resilience and resiliency. *Journal of Clinical Psychology* 58: 307–321.

Ventura R, Cabib S, and Puglisi-Allegra S (2002) Genetic susceptibility of mesocortical dopamine to stress determines liability to inhibition of mesoaccumbens dopamine and to behavioral 'despair' in a mouse model of depression. *Neuroscience*115: 999–1007.

Weiss GK, Ratner A, Voltura A, Savage D, Lucero K, and Castillo N (1994) The effect of two different types of stress on locus coeruleus alpha-2 receptor binding. *Brain Research Bulletin* 33: 219–221.

Relevant Website

http://www.who.int – World Health Organization.

Stress and Vulnerability to Brain Damage

C D Conrad, R L Wright, and K J McLaughlin, Arizona State University, Tempe, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The stress response is critical for the success of an individual and helps one adapt to changes in the environment. The hippocampus, a brain region involved with learning and memory, is particularly sensitive to stress hormones and responds to them through changes in structure, neurochemistry, and excitability. In isolation, acute stress does not impact overall hippocampal health, but acute stress does exacerbate hippocampal damage when combined with a metabolic challenge. On the other hand, chronic stress is directly responsible for hippocampal dendritic retraction, which can compromise hippocampal function, such as spatial memory. When chronic stress ends, hippocampal dendritic retraction can reverse to its prestress condition, which also restores hippocampal function. However, hippocampal volume reductions or neuronal dendritic retraction may increase the risk for hippocampal cell death from other insults or conditions. Understanding these risks and taking steps to avoid or attenuate the impact of stress can help mitigate the potential negative effects of stress on the body and brain.

The Stress Response

The resources of the body are partitioned among many competing needs for an organism's long-term survival; these include digestion, reproduction, and growth. A stress response allows energy to be commandeered for immediate use because an organism will not survive unless energy is redirected to serve these urgent needs, such as escape or defense against real or perceived danger. The immediate response to a stressor, an event that triggers a stress response, requires energy sources to be diverted away from digestion, reproduction, and growth and to be redirected toward the heart, skeletal muscles, and brain. All of these changes enhance an individual's immediate needs for surviving a threat.

The hormonal changes in response to a stressor involve two main components: the sympathetic nervous system (SNS) and the hypothalamic–pituitary–adrenal (HPA) axis. The SNS is activated relatively quickly in response to a stressor because neural activation triggers the adrenal medulla to release epinephrine within seconds after stressor onset. Epinephrine increases heart rate, dilates the pupils, and enhances blood flow to the muscles, while decreasing blood flow to the digestive and reproductive tracts. In contrast to the fast-acting SNS, the HPA axis eventually triggers the adrenal cortex to release glucocorticoids, which are stress steroids that include cortisol in humans and corticosterone in rats and mice. Since HPA axis activation is hormonally mediated, glucocorticoid release takes much longer than epinephrine release, with peak glucocorticoid levels in the bloodstream occurring 15–30 min after stressor onset. Glucocorticoids make energy available, mainly through gluconeogenesis, which is the breakdown of protein into glucose, and inhibit the storage of sugar by blocking insulin actions. The HPA axis complements the SNS by providing energy resources after the initial reserves are depleted.

Hippocampal Vulnerability to Damage Following Acute Stress

While acute stress is essential for individuals to respond to their environment, acute stress can have negative ramifications on the brain under certain circumstances. One brain region that is particularly vulnerable to stress is the hippocampus, which is thought to be susceptible to stress for at least two reasons. First, the hippocampus contains a high number of receptors for glucocorticoids, which may make the hippocampus sensitive to changes in glucocorticoid levels. Indeed, many of the changes found in the hippocampus following acute stress can be produced by exogenous glucocorticoids and prevented by drugs that interfere with glucocorticoid release and/or actions. Second, the role of the hippocampus in learning and memory necessitates that the hippocampus remain labile and this plasticity may make it vulnerable to the influence of acute stress.

Acute stress or glucocorticoid elevations can influence the hippocampus in different ways, which may include neurogenesis inhibition, neurochemistry alterations, neuron excitability changes, and even neuron death. The hippocampus is one of the few brain regions whereby neurogenesis continues beyond development into adulthood. However, proliferating cells in the hippocampus are markedly reduced following a single exposure to a stressful experience or glucocorticoids. Several neurotransmitter systems and neurotrophic factors are altered by acute stress and glucocorticoids. Acute stress and glucocorticoids facilitate the release or responsiveness of serotonin (5-HT), glutamate, norepinephrine (noradrenaline), and nerve growth factors (NGFs), alters the inhibitory tone of the γ-aminobutyric acid (GABA) system, and

decreases brain-derived nerve growth factor (BDNF). The changes in each of these systems can have profound effects on the hippocampus. For example, enhanced 5-HT tends to negatively modulate hippocampal functioning and reduced BDNF can hinder structural plasticity within the hippocampus. Acute stress and glucocorticoids can also influence electrophysiological properties such as calcium conductance, a process that influences neurotransmitter release and cell communication. All of these processes are highly regulated under a tightly coordinated system so that alterations following acute stress are quickly returned to the prestress condition.

The mechanisms underlying acute stress effects on hippocampal neurogenesis, neurochemistry, and neuron excitability differ from those underlying hippocampal cell death. First, changes in hippocampal neurogenesis, neurochemistry, and neuron excitability are reversible, whereas hippocampal cell loss is permanent. Second, glucocorticoid elevations directly mediate the actions of acute stress on hippocampal neurogenesis, neurochemistry, and electrophysiology, but glucocorticoids alone fail to kill hippocampal neurons. Instead, acute stress compromises hippocampal neurons by reducing their ability to withstand a concurrent metabolic challenge. A metabolic challenge refers to abnormal metabolic events such as ischemia (localized blood flow obstruction), stroke (generalized blood flow obstruction), hypoxia (lack of oxygen), hypoglycemia/hyperglycemia (abnormal sugar levels), and malnutrition (insufficient nutrition, minerals, and vitamins). When severe, any of these metabolic events by themselves can be damaging to the hippocampus and other brain regions. Although, when one of these events occurs in a small brief episode, individuals may recover without noticeable brain damage. Moreover, the individuals may not even recognize that they experienced a metabolic event. However, when the small, typically uneventful metabolic event coincides with acute stress and elevated glucocorticoids, the hippocampus can become sufficiently compromised, such that brain damage occurs. Individuals experiencing stroke or ischemic challenges showed more neuron loss within the hippocampus following acute stress or elevated glucocorticoids than when ischemia occurred alone. Studies with animals have shown that neurotoxins applied to the hippocampus at subthreshold levels of neurotoxicity caused severe hippocampal neuronal death when the subthreshold level of neurotoxin coincided with acute stress or glucocorticoid elevations. Thus, changes in hippocampal neurogenesis, neurochemistry, and electrophysiology can be altered by acute stress alone, whereas hippocampal cell death requires acute stress to coincide with a metabolic challenge.

As people age, the probability of experiencing a metabolic challenge increases, and the likelihood that acute stress exacerbates damage to the hippocampus is enhanced. Moreover, the hippocampus participates in the regulation of the HPA axis and damage to the hippocampus disrupts HPA axis activity, producing a cascading cycle that leads to more damage. For example, hippocampal damage elevates basal levels of glucocorticoids and delays glucocorticoid recovery to baseline following a stressor. Consequently, a metabolic challenge that produces a small degree of hippocampal damage will increase the probability of further damage because this individual will have elevated glucocorticoids that can exacerbate hippocampal damage following a recurrent metabolic challenge. The relevance of small, subthreshold metabolic challenges is hypothesized to influence other neurological conditions. For example, small inconspicuous strokes or cerebrovascular accidents, which are referred to as silent cerebral infarctions (SCIs), are hypothesized to underlie late-onset depression. The risk factors for SCI are cigarette smoking and hypertension. These findings can be extrapolated to indicate that cigarette smoking and hypertension can further increase one's vulnerability to metabolic challenges and the likelihood of experiencing hippocampal damage exacerbated by stress. Thus, acute stress is important and essential for proper functioning and survival of an individual, but can have negative ramifications including irreversible brain damage under some circumstances.

Hippocampal Vulnerability to Damage Following Chronic Stress

Early, ground-breaking studies in the 1970s and 1980s by Aus Der Mühlen, Landfield, Sapolsky, and McEwen first revealed that the brain is negatively impacted by chronic stress, which is defined as the persistence of acute stress or repeated exposure to separate episodes of acute stress. These seminal studies demonstrated that chronic stress can potentially damage the brain in guinea pigs and rodents. Subsequent studies showed that prolonged and elevated glucocorticoids mediated the neuronal loss within the hippocampus caused by chronic stress in laboratory animals. However, recent work in the 1990s and early 2000s indicated that chronic stress or glucocorticoids, alone, were insufficient to produce hippocampal damage in nonhuman primates and humans. Consequently, chronic stress and glucocorticoids were hypothesized to contribute to eventual hippocampal cell death, but they were not the sole determinant for killing hippocampal neurons.

A less extreme change in the hippocampus following chronic stress, but far more reliable than cell

death, is dendritic retraction. Chronic stress consistently causes hippocampal dendritic retraction, which is characterized by decreased dendritic branch points and total dendritic length. Dendritic arbors are specialized regions of the neuron that receive information from neuronal afferents and alterations in dendritic trees can influence neuronal communication. The primary regions of the hippocampus are the cornu ammon regions (CA1, CA2, CA3, CA4) and the dentate gyrus, with the CA3 region being the most vulnerable because it expresses dendritic retraction first in response to chronic stress. In addition to the hippocampus, chronic stress produces dendritic retraction in neurons of the prefrontal cortex and increases dendritic complexity in neurons of the amygdala. Thus, chronic stress influences neuronal architecture in several brain regions.

Many of the neurotransmitters that are altered by acute stress are involved in producing hippocampal dendritic retraction following chronic stress (Figure 1). Chronic stress causes hippocampal CA3 dendritic retraction through glucocorticoids that most likely act on glucocorticoid receptors (GR or type II). Chronic stress reduces dendritic arborization through mechanisms that include enhancing 5-HT and glutamate transmission, and altering GABA and neurotrophic expression. Any one or a combination of changes in hippocampal neurochemistry following chronic stress may contribute to hippocampal dendritic retraction.

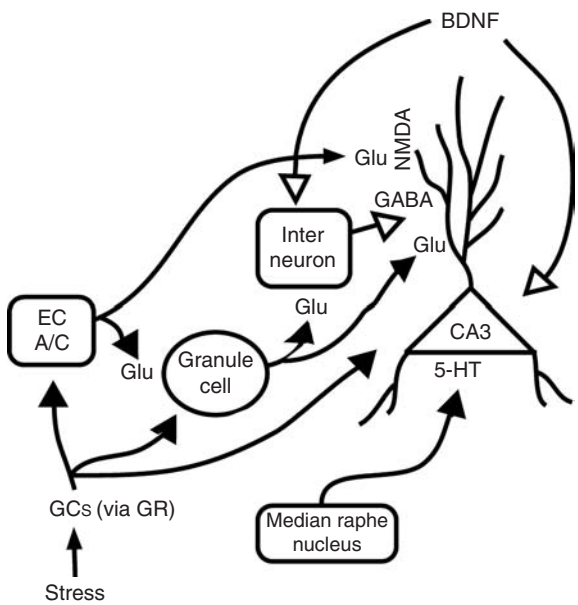

Figure 1 Repeated glucocorticoid (GC) elevations from chronic stress directly influence CA3 hippocampal neurons and their afferents (dentate gyrus granule cells, associational/commissural fibers (A/C), entorhinal cortex (EC)), because all of these cells express glucocorticoid receptors (GR). Excess glutamate (Glu via *N*-methyl-D-aspartate receptor (NMDA)) and serotonin (5-HT) as well as altered inhibitory tone from interneurons and gamma-aminobutyric acid (GABA) modulate CA3 dendritic retraction. Reduced levels of brain-derived neurotropic factor (BDNF), which is retrogradely transported to CA3 neurons may permit CA3 dendritic retraction. Solid arrows: enhanced tone permits CA3 dendritic retraction; open arrows: reduced tone permits CA3 dendritic retraction. Reproduced from Conrad CD (2006) What is the functional significance of chronic stress-induced CA3 dendritic retraction within the hippocampus? *Behavioral and Cognitive Neuroscience Reviews* 5: 41–60, with permission from Sage Press.

Hippocampal dendritic retraction following chronic stress may be a compensatory response to protect against extended excitatory amino acid stimulation, which can kill neurons. Consequently, dendritic restructuring may be a mechanism that allows the hippocampus to maintain stability through change (termed homeostasis, and recently termed allostasis). The price of maintaining homeostasis/allostasis under adverse conditions is termed, allostatic load. The consequence of this dendritic restructuring and allostatic load can be detrimental. For example, the presence of hippocampal dendritic retraction has typically corresponded to impaired hippocampal function, such as impaired spatial memory. Interventions that block hippocampal dendritic retraction also prevent spatial memory deficits. Moreover, CA3 hippocampal neurons expressing dendritic retraction following chronic stress are more likely to be killed following a neurotoxin challenge compared to neurons that fail to express dendritic retraction. Another hypothesis is that neurons expressing dendritic retraction are progressing toward neuronal death; however, chronic stress does not always lead to brain damage. Moreover, other brain regions, such as the amygdala, show enhanced dendritic arborization. So, it seems unlikely that both decreases and increases in dendritic complexity will progress to cell death. Thus, hippocampal neurons expressing dendritic retraction may express this type of plasticity as a compensatory response to continual stress and glucocorticoids, but at the risk of compromising efficient function and even the survival of neurons within the hippocampus following a metabolic challenge.

Factors That Influence Hippocampal Vulnerability to Chronic Stress

When an individual is under chronic stress and expresses hippocampal dendritic retraction or volume reductions, then this individual may be at an increased risk for irreversible brain damage. Although not every individual exposed to chronic stress shows debilitating symptoms, as some individuals fully recover following a bout of chronic stress, others exhibit long-term symptoms. Thus, a group of individuals can show a wide range of responses to chronic stress

conditions that appear similar. What produces these individual differences, and why do some individuals recover while others do not? Several factors can influence an individual's vulnerability to chronic stress; these include genetics, early life exposure to adversity, gender, and coping strategies.

Influence of Genetics on the Hippocampus Vulnerability to Chronic Stress

Factors that can facilitate or mitigate the expression of dendritic retraction following chronic stress will influence the hippocampus' vulnerability to cognitive dysfunction and susceptibility to damage. Neurochemical alterations can have a profound influence on hippocampal sensitivity to brain damage. As examples, individuals expressing few receptors for neurotropic factors or those with a high sensitivity to glutamate, which can lower the threshold for seizures, may be more inclined to experience structural changes to the hippocampus that will increase the risk for hippocampal damage than individuals with a high sensitivity for neurotropins and low sensitivity to glutamate. Anxiety predisposition can also guide one's susceptibility for increased risk of hippocampal damage. Rats that are prescreened for anxiety profile and then subjected to chronic stress show a very interesting pattern. Those rats characterized as high anxiety experience poor hippocampal cognitive function and altered hippocampal neurochemical profiles following chronic stress, whereas rats characterized as low anxiety performed well on hippocampal-dependent tasks and had nearly stable hippocampal neurochemical profiles. Particular strains of mice that are more fearful than normal and have elevated HPA axis activity show an increased risk of hippocampus to damage. Even maternal care can significantly influence the function of the HPA axis of the developing pups and this may be a mechanism by which genetics can influence maternal behavior (e.g., amount of maternal licking and grooming in rats), which can lead to dramatic differences in the pup's stress reactivity.

Influence of Early Life Events on Hippocampal Vulnerability to Chronic Stress

Early life experiences can influence the development of the HPA axis and subsequent HPA axis function later in life that includes alterations in the hippocampus. Removing rat pups from the mother (dam) for various periods early in life is a well-established model of early life variations in parental attention. When rat pups are briefly removed from the dam (3–15 min), the dam spends more time licking and grooming the pups when they are returned. Increased licking and grooming induced by brief maternal separation can alter the reactivity of the HPA axis. These pups exposed to brief separation show in adulthood increased GR in the hippocampus, decreased corticotrophin-releasing factor (CRF) synthesis and reduced secretion of adrenocorticotrophin hormone (ACTH), and glucocorticoids in response to stress. Consequently, the hippocampus has an enhanced capacity to detect glucocorticoids and can efficiently contribute to HPA axis negative feedback. In contrast, when the dam–pup separation is extended (3–6 h), the pups show an increased hormonal response to stressors later in life characterized by increased CRF mRNA and decreased glucocorticoid negative feedback on the HPA axis following a stressor. This reduced capacity to detect glucocorticoids contributes to inefficient negative feedback on the HPA axis, which can further contribute to increased CRF, ACTH, and glucocorticoid release from the hypothalamus, pituitary, and adrenal, respectively. A critical difference between the brief dam–pup separation (3–15 min) and the extended dam–pup separation (3–6 h) is the increased attention of the dams toward their pups. Dams that have been separated from their pups for extended durations are unable to provide sufficient attention to overcome the negative impact of the separation on the pups. Humans show similar patterns as observed with rodents in that individuals whose parents emotionally neglected them, or experienced harsh and inconsistent parenting when they were children, show a greater risk for depression, alcoholism, and anxiety disorders. Therefore, it appears that early life experiences can alter the function of the adult brain across animals and humans, with the hippocampus being particularly susceptible to the influence of early life exposure to stress hormones.

Maternal separation for extended periods also alters the serotonergic system. In monkeys and rats, extended maternal separation increases serotonin levels in adulthood. Since serotonin contributes to hippocampal dendritic retraction, enhanced serotonin levels following this developmental manipulation may contribute to hippocampal susceptibility to damage following insults in adulthood. Changes in the serotonin system early in life may work together with increased activity of the HPA axis to make the adult more susceptible to dendritic retraction as well as more permanent brain damage from other insults.

The prenatal period is another developmental stage that is susceptible to the effects of stress and glucocorticoids on the HPA axis and the hippocampus. Generally, the placenta protects the fetus from stress and high levels of glucocorticoids that the mother may experience. The placenta inactivates glucocorticoids to prevent the majority of circulating glucocorticoids from reaching the fetus. However, the fetus

can exhibit long-lasting effects following exposure to high levels of glucocorticoids through the administration of synthetic glucocorticoids, drugs with some glucocorticoid activity, or from extremely high levels of maternal glucocorticoids. In nonhuman primates, exposure to synthetic glucocorticoids, such as dexamethasone, during gestation produces hippocampal volume reductions that last into adulthood. In rats, prenatal exposure to glucocorticoids reduces the number of hippocampal receptors for glucocorticoids (GR and mineralocorticoid (MR)), and elevates the basal secretion of glucocorticoids even in adulthood. These findings show that exposure of the developing fetus to elevated glucocorticoids can permanently alter the HPA axis. These HPA axis changes can increase the exposure of the hippocampus to glucocorticoids, which may further compromise the hippocampus (Figure 2).

Influence of Gender on Hippocampal Vulnerability to Chronic Stress

The hippocampus is sexually dimorphic in both its anatomy and neurochemistry and these dimorphisms contribute to sex differences in stress sensitivity. In general, females have a larger hippocampus than males, while males have a more complex CA1 dendritic tree than females until late in life. The receptors that detect glucocorticoids (GR and MR) are abundant in the hippocampus of both sexes. Yet, females appear to have a greater overall binding capacity for glucocorticoids and show nearly 20% higher basal and stress levels of glucocorticoids than males. In contrast, the male hippocampus has a high affinity for glucocorticoids and hippocampal GR downregulates following adrenalectomy. Moreover, androgen administration downregulates GR without altering MR. Yet, MR may contribute to sex differences in response to glucocorticoids because ovarian hormones exert their actions primarily on the MR. All of these variables contribute to sexual dimorphisms within the hippocampus following chronic stress.

Chronic stress produces clear sex differences in the expression of GR, which may contribute to sex distinctions in hippocampal CA3 dendritic retraction. In males, chronic stress downregulates hippocampal GR and MR mRNA, while in females, chronic stress fails to alter hippocampal GR mRNA and even upregulates MR mRNA. These changes in receptor expression for males and females may influence the sensitivity of the hippocampus to chronic stress. Indeed, male rats express robust CA3 hippocampal dendritic retraction, while female rats show mild CA3 dendritic retraction. Males are thought to show a higher degree of hippocampal dendritic retraction than females because the dendritic tree affected by chronic stress in males is the apical region, which is three times larger than the basal region that is influenced by chronic stress in females. Moreover, two measures are typically used to detect changes in dendritic arbors (dendritic length and branch points)

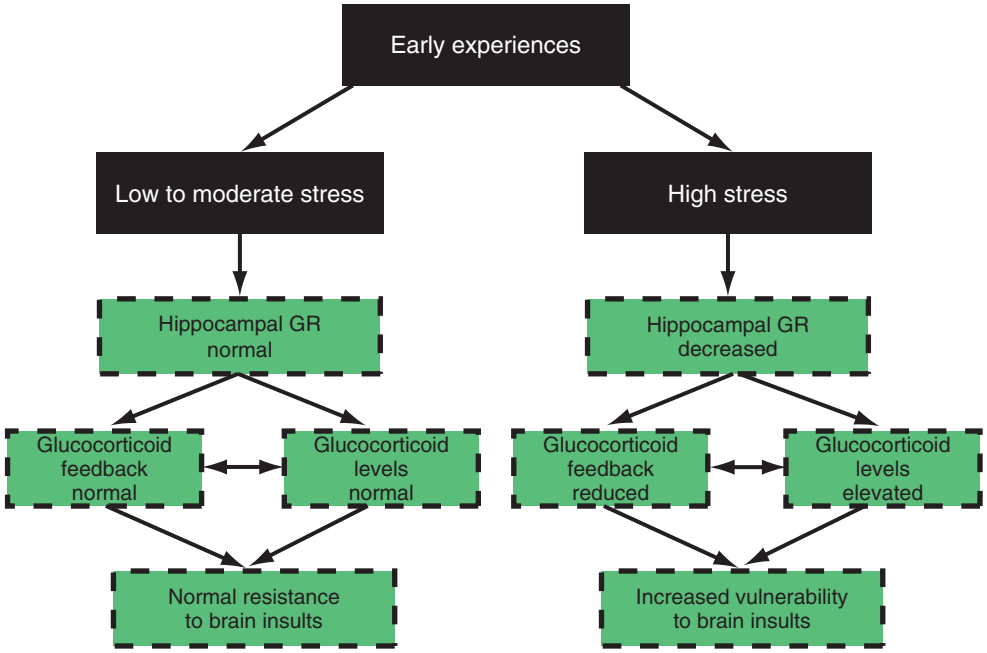

Figure 2 Diagram of the effects of early life experiences on the development of the HPA axis and the possible outcomes for hippocampal brain damage. The black boxes outlined in solid lines with white lettering indicate events that occur in early life. The shaded boxes outlined in dashed lines indicate outcomes that persist into adulthood. GR, glucocorticoid receptors.

and chronic stress alters both measures in males, while only reducing dendritic branch points in females. In another example, when the hippocampus of chronically stressed rats is challenged with a neurotoxin, the male hippocampus is more susceptible to cell death than the female hippocampus. These findings are taken to suggest that chronic stress makes the hippocampus of males more vulnerable to brain damage than the hippocampus of females.

Sex differences in chronic stress-induced changes in neurotransmitter systems can also contribute to sexual dimorphisms in hippocampal vulnerability to chronic stress. Males display decreased levels of dopamine in both the medial prefrontal cortex and the amygdala and show increased levels of GABA in the CA3 region of the hippocampus in response to chronic stress. In contrast, chronic stress increases norepinephrine and serotonin levels in female CA3 neurons, and no significant changes are evident in the prefrontal cortex (including any changes in GABA). Therefore, chronic stress exerts sex-specific alterations on several neurotransmitter systems, and these changes may contribute to the differences seen in CA3 dendritic retraction.

Gonadal hormones are an important component underlying CA3 dendritic retraction following chronic stress. In gonadally intact, female rats, chronic stress produces mild CA3 dendritic retraction, which is exacerbated following ovariectomy, the removal of endogenous ovarian hormones (Figure 3). These findings suggest that estrogen may protect against CA3 dendritic retraction following chronic stress and corroborates findings showing that estrogen attenuates brain damage following ischemia, oxidative stress, and neurological disorders, including Alzheimer's and Parkinson's disease. While ovariectomy nearly eliminates estrogen and progesterone, changes in cholinergic and dopaminergic systems are also observed. Estrogen replacement reverses these alterations and may protect the brain through the activation of these neurotransmitter systems. Moreover, in males, testosterone can attenuate hippocampal damage caused by severe stress. Thus, gonadal hormones may dampen the potentially negative effects of chronic stress on the hippocampus.

Coping and Behaviors that Mitigate Hippocampal Vulnerability to Chronic Stress

The negative effects of stress on the brain, endocrine system, and behavior can be mitigated by a healthy lifestyle and positive social interactions. A high-fat diet increases glucocorticoid levels, reduces the ability of rats to habituate to stressors, and makes rats more susceptible to dendritic retraction. So, a low-fat diet may help to shield one from the adverse consequences of stress. Getting enough sleep also helps to maintain a healthy stress response, as people who sleep less than 4 h a night for several consecutive nights have high glucocorticoid levels in the evening when these glucocorticoid levels should be at their lowest. After a full night's sleep, glucocorticoid levels return to normal. Social support is another powerful regulator of the HPA axis in social animals including humans. In humans, social support consists of positive interactions with other individuals. Close relationships such as with a spouse or close friend seem be more valuable than the support of acquaintances. In animals exposed to a laboratory stressor, social support reduces the glucocorticoid response with important gender differences. Males are more likely to be stressed by social crowding, whereas females are stressed by social isolation. Thus, females may benefit more from group interactions during a stressful period than males. Moreover, stable social hierarchies and being in close proximity to others of the social group lead to low stress hormone levels compared to situations of social unrest or separation. Positive interactions, as perceived by the individual, can help mediate the stress response by allowing emotions associated with that stressor to be expressed. Thus, lifestyle can influence the HPA axis, which leads to a healthy stress response and reduced susceptibility to the adverse effects of stress.

An important consideration of these findings is that the type of activity that one is engaged in at the time of glucocorticoid elevations is more important in determining brain susceptibility to damage than whether glucocorticoids are solely elevated. For instance, exercise protects the body and brain against the ravages of stress by (1) increasing dendritic complexity and regional brain volume, (2) stimulating neurogenesis within the hippocampus, and (3) alleviating symptoms of depressed patients. However, exercise elevates stress hormones and the glucocorticoid response. Thus, exercise produces advantageous effects on hippocampal health while elevating glucocorticoids. In another example, exposure of a rat to a predator, a highly stressful experience, impairs subsequent memory ability on a task that requires hippocampal function. The predator exposure elevates glucocorticoids levels and attenuating glucocorticoid elevations can prevent the memory deficits. While glucocorticoid elevations appear to mediate the memory impairment, elevating glucocorticoids using a different activity, such as allowing access to sexually receptive females, fails to alter cognitive performance. These studies indicate that the susceptibility of the brain to damage following glucocorticoid elevations is highly intricate and most likely involves a complex interplay between glucocorticoid

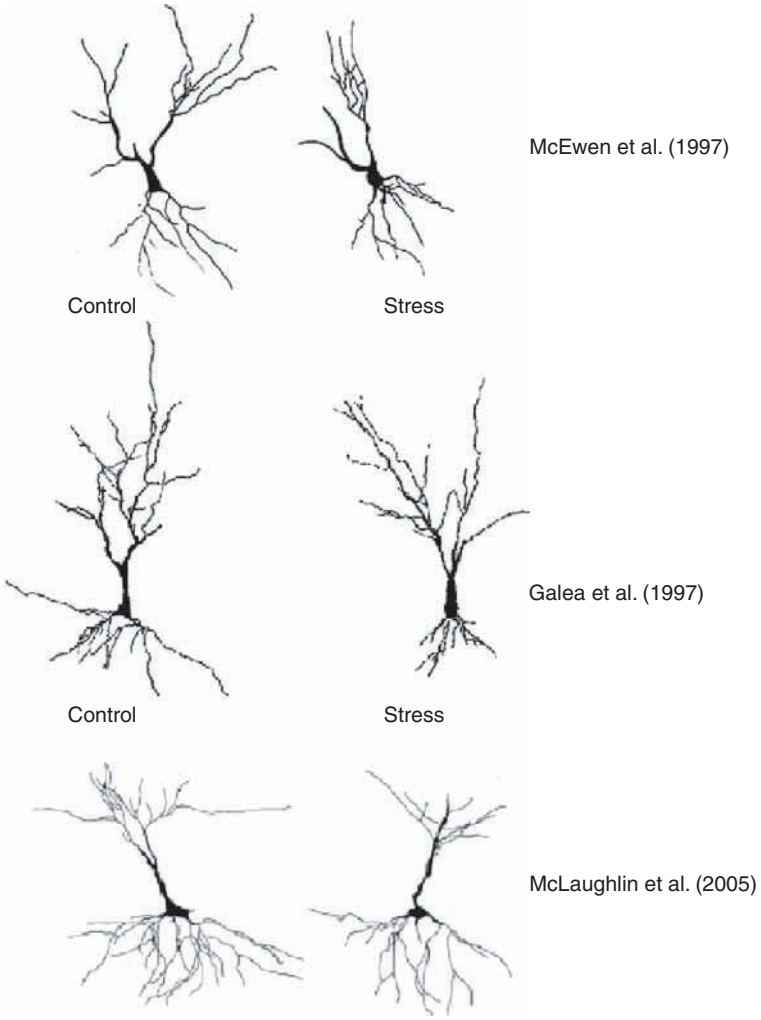

Figure 3 Camera lucida tracings of CA3 hippocampal neurons. The top two neurons show that chronic stress (6 h restraint/day for 21 days) produces CA3 dendritic retraction in the apical (top) region of males. The middle two neurons show that a similar chronic stress manipulation produces CA3 dendritic retraction in the basal (bottom) region in gonadally intact females. However, CA3 dendritic retraction is mild because chronic stress reduced dendritic branch points without affecting dendritic length. The bottom two neurons show that chronic stress causes robust dendritic retraction in chronically stressed females following ovariectomy. Dendritic retraction is considered robust because both apical and basal dendritic trees are altered and dendritic retraction is reduced as determined by both measures of dendritic branch points and length. Note that the bottom two neurons represent the long-shafted type of neuron, which is more likely to express dendritic retraction following chronic stress than the short-shafted neuron type, which is represented in the top and middle panels. Top two neurons are reproduced from McEwen BS, Conrad CD, Kuroda Y, Frankfurt M, Maqarinos AM, and McKittrick C (1997) Prevention of stress-induced morphological and cognitive consequences. *European Neuropsychopharmacology* 7: S323–S328, with permission from Elsevier. Middle two neurons are reproduced from Galea LAM, McEwen BS, Tanapat P, Deak T, Spencer RL, and Dhabar FS (1997) Sex differences in dendritic atrophy of CA3 pyramidal neurons in response to chronic restraint stress. *Neuroscience* 81: 689–697, with permission from Elsevier. Bottom two neurons are reproduced from McLaughlin K, Baran SE, Wright RL, and Conrad CD (2005) Chronic stress enhances spatial memory in ovariectomized female rats despite CA3 dendritic retraction: Possible involvement of CA1 neurons. *Neuroscience* 135: 1045–1054, with permission from Elsevier.

levels and the brain regions activated at the time of glucocorticoid elevations, as opposed to being solely dependent upon glucocorticoid increases.

Conclusion and Parallels between the Animal and Human Literature

Hippocampal dendritic retraction in animal studies has important parallels with the literature on humans. Ten days following the termination of chronic stress in rodents, CA3 dendritic retraction reverses to the prestress condition and is essentially indistinguishable from nonstressed controls. Individuals diagnosed with Cushing's disease, characterized by the hypersecretion of glucocorticoids, show reduced hippocampal volumes based upon functional magnetic resonance imaging (fMRI) techniques. After treatment, the hippocampal volumes of Cushing's

patients show significant increases. A common characteristic of individuals with major depression is glucocorticoid hypersecretion and hippocampal volume reduction, which is corrected with antidepressant treatment. The restoration or partial reversal of hippocampal volume shrinkage following treatment suggests that irreversible cell death is an unlikely mechanism and instead, reversible processes such as dendritic retraction may be involved. Another parallel between the animal and human literature is that chronic stress exacerbates hippocampal damage in vulnerable conditions. In rodents, hippocampal damage by neurotoxin challenge is worsened following chronic stress when hippocampal dendritic retraction is present. Humans show similar vulnerabilities in that the hippocampus is susceptible to damage when chronic stress precedes or coincides with conditions such as AIDS, obesity, depression, and Alzheimer's disease. Thus, chronic stress can increase the risk for hippocampal vulnerability to permanent damage.

See also: Adrenal Steroids: Biphasic Effects on Neurons; Chronic (Repeated) Stress: Consequences, Adaptations; Gene Therapy and Protection from Stress-Induced Brain Damage; Social Stress in Adult Primates; Stress and Cognition; Stress Response: Sex Differences; Stress and Neuronal Plasticity; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Cahill L (2006) Why sex matters for neuroscience. *Nature Reviews Neuroscience* Published online 10 May 2006, doi:10.1038/nrn1909.

Caldi C, Diorio J, and Meaney MJ (2000) Variations in maternal care in infancy regulate the development of stress reactivity. *Biological Psychiatry* 48: 1164–1174.

Conrad CD (2006) What is the functional significance of chronic stress-induced CA3 dendritic retraction within the hippocampus? *Behavioral and Cognitive Neuroscience Reviews* 5: 41–60.

de Kloet ER, Joëls M, and Holsboer F (2005) Stress and the brain: From adaptation to diseases. *Nature Reviews Neuroscience* 6: 463–475.

Diamond DM, Campbell A, Park CR, and Vouimba RM (2004) Preclinical research on stress, memory, and the brain in the development of pharmacotherapy for depression. *European Neuropsychopharmacology* 14(supplement 5): S491–S495.

Galea LAM, McEwen BS, Tanapet P, Deak T, Spenor RL, and Dhabar FS (1997) Sex differences in dendritic atrophy of CA3 pyramidal neurons in response to chronic restraint stress. *Neuroscience* 81: 689–697.

Lee SJ and McEwen BS (2001) Neurotrophic and neuroprotective actions of estrogens and their therapeutic implications. *Annual Reviews in Pharmacological Toxicology* 41: 569–591.

Luine V (2002) Sex differences in chronic stress effects on memory in rats. *Stress* 5: 205–216.

Lupien SJ, Nair NPV, Briére S, et al. (1999) Increased cortisol levels and impaired cognition in human aging: Implication for depression and dementia in later life. *Reviews in Neurosciences* 10: 117–139.

McEwen BS (2000) Allostasis, allostatic load, and the aging nervous system: Role of excitatory amino acids and excitotoxicity. *Neurochemical Research* 25: 1219–1231.

McEwen BS, Conard CD, Kuroda Y, Frankburt M, Maqarinos AM, and Mc Kittrick C (1997) Prevention of stress-induced morphological and cognitive consequences. *European Neuropsychopharmacology* 7: S353–S328.

McEwen BS and Norton-Lasley E (2002) *The End of Stress As We Know It*. Washington, DC: Dana Press.

Mclaughlin K, Baran SE, Wright RL, and Conrad CD (2005) Chronic stress enhances spatial memory in ovariectomized female rats despite CA3 dendritic retraction: Possible involvement of CA1 neurons. *Neuroscience* 135: 1045–1054.

Sapolsky RM (1992) *Stress, The Aging Brain, and the Mechanisms of Neuron Death*. Cambridge, MA: MIT Press.

Shors TJ (2006) Stressful experience and learning across the lifespan. *Annual Reviews in Psychology* 57: 55–85.

Smith MA (1996) Hippocampal vulnerability to stress and aging: Possible role of neurotrophic factors. *Behavioural Brain Research* 78: 25–36.

Will B, Galani R, Kelche C, and Rozenzweig MR (2004) Recovery from brain injury in animals: Relative efficacy of environmental enrichment, physical exercise or formal training (1990–2002). *Progress in Neurobiology* 72: 167–182.

Zautra AJ (2003) *Emotions, Stress, and Health*. New York: Oxford University Press.

Stress Response and Self-Esteem

J C Pruessner, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

The Situational Characteristics of a Stressful Situation

One of the most influential articles on the characteristics of stress in the field of psychoneuroendocrinology was contributed in 1968 by John Mason, who had reviewed and summarized the available stress literature at the time. He concluded that novelty, unpredictability, high ego involvement, anticipation of negative consequences, and uncontrollability are the five main factors that contribute to the perception of a situation as stressful. Although not all of these factors are solely situational characteristics, Mason's hallmark article was often referred to as defining the aspects of the situation that contribute to the experience of stress. In the late 1980s and early 1990s, the psychoneuroendocrine and stress research field received an increase in attention by the advent of noninvasive and inexpensive hormonal measurement techniques that made the inclusion of hormonal measures feasible for a wide range of researchers and studies. Thus, at the end of the last millennium, Bondi and Picardi reviewed over 100 research articles investigating psychological aspects of stress and concluded that the uncontrollability of the situation significantly contributes to the experience of stress, and that a stressful situation can contribute to perceptions causing irritability and tension. Finally, an influential and important review was contributed by Dickerson and Kemeny in 2004, who again reviewed the available stress literature at the time and, through the means of conducting a meta-analysis, concluded that it is the situational aspect of 'social evaluative threat' that determines the stressfulness of a situation.

The best known psychosocial stressor, the public speaking task, is likely to contain all the elements that these three reviews have identified, but it is the social evaluative threat that seems to be at the heart of the stress experience. The authors of these reviews formulated a conceptual framework within which the social evaluative threat had a special connotation: they argued that throughout evolution, a social self-preservation system was important for the survival of the individual as well as the species. It can be further argued that the hormonal stress system might be the physiological correlate of the self-preservation system, to provide increased metabolic resources in times of social threat and thus increased demand. Independent of the validation of this model, however, it is likely that a situation that contains a social evaluative threat component will also have elements of uncontrollability, unpredictability, and high ego involvement, together with the other factors that were identified early on, and as such the recent refinements to the elements defining a stressful situation integrate well with the earlier findings made in that regard.

Personality Characteristics That Contribute to the Perception of Stress

One of the most influential figures in stress research, Lazarus, noted early on that stressful conditions per se fail to produce reliable effects on task performance. Keeping all situational variables constant, the same stressor can have minimal effect, lead to performance improvement, or result in significant performance impairment across different subjects. This led Lazarus and others to suggest that individual differences in psychological variables are likely to interact with situational components, and what is considered threatening for some, and thus stressful and performance impairing, might be perceived as stimulating by others, and thus produce beneficial effects.

Subsequent research by Lazarus and others identified the process of evaluation as crucial in explaining the impact of psychological variables on the stress response. Any internal or external stimulus is perceived as stressful only if it is evaluated as harmful, or threatening. This evaluation process is simplistically explained by a comparison of the demands of the situation with the available resources within the individual. If the resources exceed the demands, the situation is judged positively (challenging, stimulating), but if the demands exceed the resources, the situation is likely to be considered negatively (threatening, dangerous). Personality characteristics, especially self-esteem and locus of control, are likely to influence the perception of one's own resources as well as the perception of the demands.

Self-esteem is defined as the value we place on ourselves. Epidemiological studies have shown that low self-esteem is associated with negative life outcomes, including substance abuse, delinquency, unhappiness, depression, and worsened recovery after illnesses. On the other hand, high self-esteem has been linked to increased health and longevity. In studies of aging, a positive self-concept has been identified to play a key role in successful aging, predicting independence, cognitive stability, and general health in old age.

Internal locus of control can be defined as the perception that one's outcomes are determined by one's actions. Not surprisingly, internal locus of control and self-esteem are usually highly correlated. The key link of these variables to the experience of stress lies in their impact on the evaluation of any given situation. We postulate that in the evaluation of whether a given situation is perceived as threatening or challenging, self-esteem and internal locus of control systematically interact with situational factors. If a person attributes little importance to him/herself, and thinks he/she has little impact on the outcome of his or her own actions, this person will find more situations uncontrollable and unpredictable, and consequently will evaluate more situations as threatening and harmful.

Psychoneuroendocrinological Studies Showing the Role of Self-Esteem and Locus of Control in the Perception of Stress

We have been able to find support for this model in several psychoneuroendocrinological studies, using different laboratory stressors. In a computerized mental arithmetic task, we manipulated the performance (thus inducing failure) to invoke stress. In the setup used in this task, 52 students performed the task on computer terminals in front of them. Half of the students were exposed to a difficult version leading to low performance, compared to an easy version of the task with high performance for the other half. The students played the task in three 3-min segments, and had to announce their performance score after each segment to the investigator, who wrote the scores down on a board for everybody to see. Saliva sampling before, throughout, and after the task allowed us to assess the cortisol dynamics in relation to this paradigm. Interestingly, this task only triggered a significant cortisol release in the subjects who were in the low-performance group, and had low self-esteem and low internal locus of control. Neither low performance alone, nor low self-esteem and internal locus of control alone, were significant predictors of cortisol release, supporting the notion that these personality variables produce effects only in interaction with a potentially stressful situation. We suggest that the evaluation of the situation is at the core of this interaction.

In a second study, we used the 'public speaking' paradigm, where subjects have to give a speech and perform serial subtraction tasks in front of an audience, usually for a total of about 10 min. The audience consists of two to three persons who maintain a neutral facial expression throughout the subject's presentation. During the speech, the audience interacts with the subject only to indicate the amount of time that is left to talk, or to ask specific questions. During the serial subtraction task, the subject is interrupted only when making a mistake. The subject is then corrected and instructed to start the task over. The task has been shown to be a powerful stressor, stimulating the hypothalamic–pituitary–adrenal (HPA) axis and leading to significant free cortisol increases within 15–30 min following the onset of the task. In this study, we aimed to validate the hypothesis that in humans, repeated exposure to the same stressor will lead to quick habituation of the stress response. In order to test the habituation of the stress response, a total of 20 young healthy male college students were exposed to the public speaking stressor on five subsequent days. For this purpose, the public speaking was modified using different speech topics and serial subtraction tasks on each day. Interestingly, we found that only 13 of the 20 students showed the typical habituation pattern, with a normal stress response on day 1 being significantly reduced on day 2, and no longer present on the subsequent days. In the seven remaining students, however, the cortisol stress response continued to be present on all days. When analyzing the available psychological variables, it became apparent that low internal locus of control and low self-esteem were the best predictors of a lack of habituation of the cortisol stress response to repeated stress exposure. This can be interpreted as a sign that these personality variables systematically interact with the evaluation of a situation during repeated exposure.

The Hippocampus as a Possible Mediator of the Relationship between Self-Esteem, Locus of Control, and Stress

Although only very few studies on brain correlates of personality variables and endocrine function in humans are available, a picture emerges that puts the spotlight on the hippocampus as a likely mediator between the reported personality variations and stress responses. On the one hand, the hippocampus is one of the major limbic system structures involved in the regulation of the stress response. Psychological stress involves ambivalent signals that require considerable evaluation before a stress response is initiated. The limbic system and the hippocampus are implicated in assessing the salience of stimuli. Within the HPA system, one relevant limbic connection involves the hippocampus to the hypothalamus via the bed-nucleus of the stria terminalis, and the paraventricular nucleus. Interestingly, these connections are mostly inhibitory, and numerous studies show the role of the

hippocampus in inhibiting the HPA. This observation led researchers to suggest that a smaller hippocampus might be linked to higher HPA axis activity, since the hippocampus serves to inhibit HPA axis activity. Indeed, numerous studies over the past two decades have contributed evidence to show how excessive HPA axis activity is linked to smaller hippocampal volumes. Early models postulated that associations between hippocampal volumes and HPA axis activity might represent the consequence of excessive exposure of the hippocampus to glucocorticoids, due to their powerful neurotoxic properties.

More recently however, evidence has surfaced that allows for a link between naturally occurring variations in hippocampal volumes and HPA function. Functionally, the hippocampus is the primary structure for memory contextualization, and it is here that a link to self-esteem and locus of control could occur. When faced with a potentially threatening and harmful situation, memories of past events are employed to contribute to the evaluation of the current event. However, if specific situational and environmental characteristics associated with negative past events cannot be recalled, this lack of awareness of situational circumstance can lead to an overgeneralization of negative past events, and thus an increased likelihood to consider the current situations as stressful as well. Thus, poor contextualization mediated by the hippocampus could be linked to higher stress responses on the one hand, and lower self-esteem on the other. A hippocampus smaller in size might be at higher risk for poor memory contextualization and thus overgeneralization of negative events. We recently tested this hypothesis of a link between hippocampal volumes, personality variables, and stress responses in 20 healthy young male college students. All students underwent extensive psychological assessment for personality variables, including self-esteem and locus of control. In addition, they underwent magnetic resonance imaging employing a structural T1-weighted acquisition protocol, which produces high-resolution images of the cerebrum. A recently developed manual segmentation protocol was then applied to all images; this included non-uniformity correction, signal-intensity normalization, and Tailarach-like transformation for standardization of brain size. This protocol then allows for manual assessment of hippocampal volumes. Finally, all students were exposed to a computerized stress task, similar to the previously described mental arithmetic task. Saliva samples before, during, and after the task accompanied the testing to capture the cortisol stress response. Results first replicated earlier findings of mental arithmetic stress: only the students with low self-esteem and low internal locus of control showed a significant release of cortisol. Extending earlier findings, however, a correlation emerged between the cortisol stress response and self-esteem ($r = -0.45$, $p < 0.05$). Furthermore, a significant correlation emerged between the cortisol stress response and the total hippocampal volume ($r = -0.53$, $p = 0.03$), suggesting that the size of the hippocampus is related to the regulation of the HPA. Finally, supporting our initial hypothesis, we also observed a link between internal locus of control and hippocampal volume ($r = 0.66$, $p = 0.006$), and self-esteem and hippocampal volume ($r = 0.58$, $p = 0.02$). Using a recently developed voxel-based regression analysis, we were able to show subregions of the hippocampus specifically correlated with the interindividual variations in self-esteem and locus of control (**Figure 1**). These analyses can be interpreted as study participants scoring higher in self-esteem, and locus of control having higher signal intensities in the hippocampal regions that are significant in the analysis, showing which areas of the hippocampus are likely larger in the presence of high self-esteem and locus of control.

There is also some evidence that these effects are specific for the hippocampus. When employing total brain gray matter as control structure in correlations with both personality variables and cortisol response, none of the tested correlations was significant, suggesting that the observed relationships were indeed specific to this medial temporal lobe area that contains the hippocampus.

Conclusion: Personality Variables, Brain Structures, and Stress Responses

It will be a challenging task for future studies to ascertain the origin of the relationship observed here between personality variables, hippocampal volume, and stress response. There are a number of possibilities how these associations could develop. One such model puts naturally occurring variations in hippocampal volume at the origin of these functional relationships. It can be speculated that hippocampal volume is a consequence of multiple interacting factors during critical developmental stages, leading to the variation in hippocampal volumes that follows a normal distribution pattern. If, however, the size of this structure determines memory contextualization capabilities, this variation would then have personality and stress-related consequences. The impaired source monitoring would lead to an overgeneralization of the experience of failure and rejection, and consequently to a self-perception of being a failure or being socially rejected in general, which might produce low self-esteem and low locus of control during childhood and adolescence. The observed

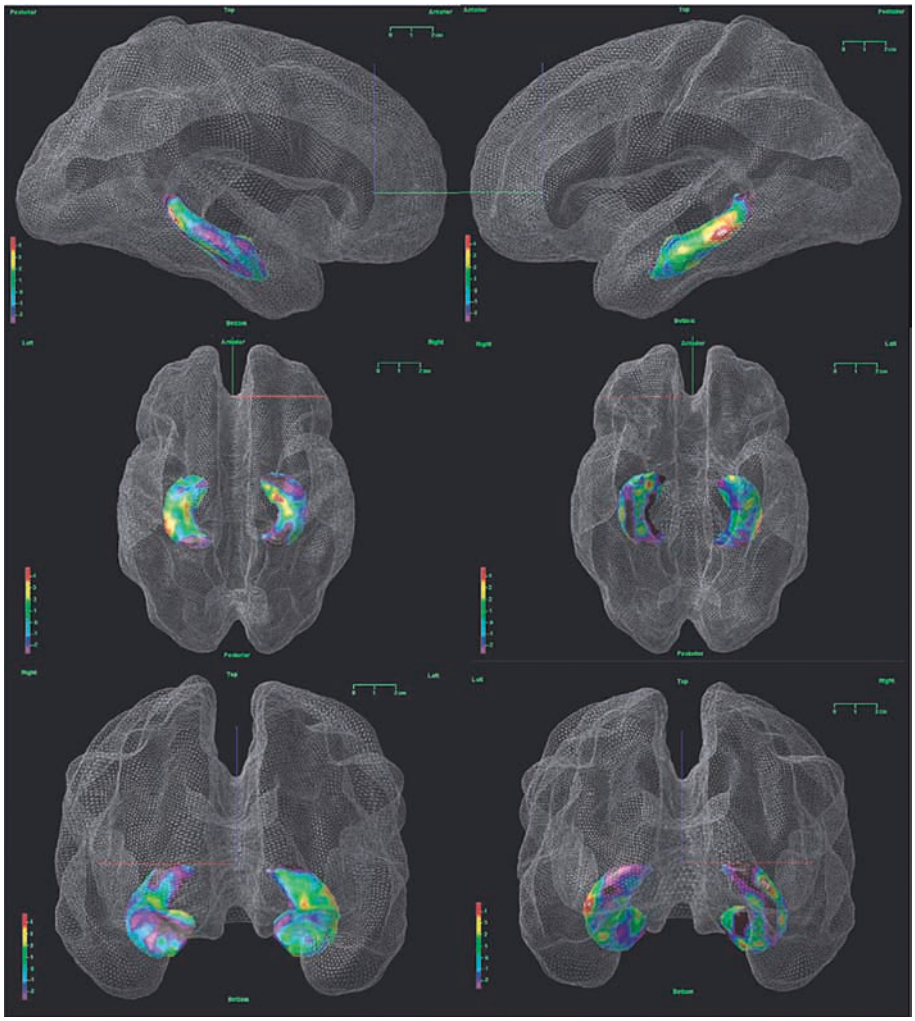

Figure 1 Voxel-based regressional analysis between scores of self-esteem and locus of control and signal intensity in the hippocampus in a population of 20 healthy young men. Areas exceeding a threshold of $3 < t < -3$ are significant and suggest that high self-esteem and high internal locus of control are associated with larger areas of hippocampal gray matter, especially in the regions next to the quadrigeminal cistern, and the inferior horn of the lateral ventricle.

higher cortisol responses would then not be a direct consequence of small hippocampal volumes, but of the increased stress perception due to consistent unfavorable evaluations of ambiguous situations. Of course, the question arises why overgeneralization should then not also be an issue for positive life events – in other words, a general lack of source monitoring for both positive and negative life events, with no consequences on self-perception. One possible explanation here could be seen in the stress-specific function of the hippocampus within the limbic system. There is evidence that it is specifically the hippocampus (together with the anterior cingulate) that is involved in the stress response. It is also known that the anterior cingulate is particularly involved in error monitoring, which means its involvement is restricted to situations where a mismatch between expectations and reality occurs. These circumstances would make a specific role of the hippocampus in reaction to negative life events more probable.

Based on our findings to date, we suggest that reduced hippocampal volumes might play a role in the development of low self-esteem. This may in turn produce a higher susceptibility to perceive ambiguous situations as threatening, and thus stressful. The idea that hippocampal volume variation might be the cause for, rather than a consequence of, adverse functional and behavioral consequences has recently been proposed in conjunction with risk factors for psychopathology. Studies are emerging showing that smaller hippocampal volumes might be present prior to the onset of an illness. In our own studies, consistent differences in hippocampal volumes were present even among relatively young adults. This supports a model in which variations in brain morphology could become a pathway to certain personality

characteristics and behavioral phenotypes early in life, but longitudinal studies will be needed to examine these changes over time, and provide validation for the assumed associations.

See also: Hippocampus; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Social Stress in Adult Primates; Stress and Parasympathetic Control; Stress Response: Sex Differences; Stress, Sex and Adolescent Nicotine Response; Stress, Dopamine, and Puberty; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Baltes MM and Baltes PB (1990) *Successful Aging*. Cambridge: Cambridge University Press.

Baumeister RF, Campbell JK, Krueger JI, et al. (2003) Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest* 4: 1–44.

Biondi M and Picardi A (1999) Psychological stress and neuroendocrine function in humans: The last two decades of research. *Psychotherapy & Psychosomatics* 68: 114–150.

Dickerson SS and Kemeny ME (2004) Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin* 130(3): 355–391.

Gilbertson MW, Shenton ME, Ciszewski A, et al. (2002) Smaller hippocampal volume predicts pathologic vulnerability to psychological trauma. *Nature Neuroscience* 5(11): 1242–1247.

Kirschbaum C, Prussner JC, Stone AA, et al. (1995) Persistent high cortisol responses to repeated psychological stress in a subpopulation of healthy men. *Psychosomatic Medicine* 57: 468–474.

Lazarus RS (1999) *Stress and Emotion: A New Synthesis*. New York: Springer.

Leary MR and McDonald G (2003) Individual differences in self-esteem: A review and theoretical integration. In: Leary MR and Tangney JP (eds.) *Handbook of Self and Identity*, pp. 401–420. New York: Guilford Press.

Mason JW (1968) A review of psychoendocrine research on the pituitary-adrenal cortical system. *Psychosomatic Medicine* 30 (supplement): 576–607.

Petrie K and Rotheram MJ (1982) Insulators against stress: Self-esteem and assertiveness. *Psychological Reports* 50(3, Pt 1): 963–966.

Pruessner JC, Baldwin MW, Dedovic K, et al. (2005) Self-esteem, locus of control, hippocampal volume, and cortisol regulation in young and old adulthood. *Neuroimage* 28: 815–826.

Pruessner JC, Collins DL, Pruessner M, et al. (2001) Age and gender predict volume decline in the anterior and posterior hippocampus in early adulthood. *The Journal of Neuroscience* 21(1): 194–200.

Pruessner JC, Hellhammer DH, and Kirschbaum C (1999) Low self-esteem, induced failure and the adrenocortical stress response. *Personality and Individual Differences* 27(3): 477–489.

Sapolsky RM (1999) Glucocorticoids, stress, and their adverse neurological effects: Relevance to aging. *Experimental Gerontology* 34(6): 721–732.

Sinha R, Lacadie C, Skudlarski P, et al. (2004) Neural circuits underlying emotional distress in humans. *Annals of the New York Academy of Sciences* 1032: 254–257.

Stress Response: Genetic Consequences

E Edwards, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA
J A King, University of Massachusetts Medical Center, Worcester, MA, USA

Published by Elsevier Ltd.

Introduction

The development of depression is a multifactorial process involving interactions of environmental and genetic factors. It imposes an enormous burden of illness on the medical system, as it affects about 15% of the population during the life span. The prominent clinical features of depression range from significant decrease of pleasure, interest, and cognitive abilities, to psychomotor dysfunction, to suicidal ideation, feeling of guilt, and psychosis. In addition, a host of vegetative symptoms, including anxiety, sleep disturbances, loss of appetite, weight loss, cardiovascular abnormalities, and gastrointestinal disturbances, have been demonstrated in depressed individuals.

Understanding the molecular and genetic bases of such a complex human condition has been hindered by the intricacy and inaccessibility of the human central nervous system. However, since the 1950s, the serendipitous discovery of agents to treat depression has inspired many hypotheses of the pathogenesis of depression. Most antidepressant drugs exert their therapeutic effects by modulating the presynaptic uptake transporters and degrading enzymes of serotonin and norepinephrine pathways. Consequently, investigators have focused research concerning the causality and treatment of depression on the metabolism of functional biogenic amines and the capacity of their respective receptors to alter intracellular signaling pathways and ultimately induce changes in gene activity.

Stress hormones have also been strongly linked to the pathogenesis of affective disorders, and exposure to a variety of stressors early in development modifies the regulation of hypothalamic and extrahypothalamic neurons, resulting in exaggerated and harmful endocrine-based physiological stress response later in life. Consequently, several research groups have formulated a hypothesis relating a dysfunction in stress hormone regulation to the causality of depression. Symptoms that are characteristics of depression include changes in functioning of the hypothalamic–pituitary–adrenocortical (HPA) system, resulting, in most depressed patients, in altered regulation of adrenocorticotropic hormone (ACTH) and cortisol secretory activity, impairment in corticosteroid receptor signaling, and increased secretion of corticotropin-releasing hormone.

Traditionally, researchers have encountered difficulties in developing and validating animal models of psychopathology because of the difficulties in providing direct evidence that the brain changes demonstrated in these models are identical to those in the human brain. Nevertheless, a variety of models of depression have been developed and they have advanced the study of the mechanisms of specific behaviors and their psychopathophysiology and have served as important screens for pharmacologic agents to treat depression. A wide range of animal models have been developed to date. They include models in which 'depressive behavior' is the result of pharmacologic treatment, the result of exposure to environmental stressors in development and in adulthood, and, more recently, the result of genetic selection or genetic manipulation. As discussed earlier, depression is heterogeneous in nature and there is great variability in the clinical responses to current antidepressants. A promising approach has been to identify key symptoms that can be studied in animals rather than trying to approximate the full spectrum of the clinical diagnosis of depression. A genetic model of affective disorder would allow the genetic analysis of polygenic traits that seem to underlie depression.

Key Considerations in Developing Models of Depression

Without a doubt, the basic goal for developing preclinical models is the overall relevance of the model to the human condition. Therefore the applicability of such models resides in their reflection of some intrinsic feature(s) of the human disorder, including heritability (see Tables 1 and 2). Reliability and validity are two key considerations to be considered when evaluating the possible effectiveness of an animal model. While the reliability of an animal model refers to the stability and reproducibility of the model, validity is established when criteria for at least one of several modes – namely face, construct, predictive, and etiological characteristic(s) – can be demonstrated. To meet the criteria for face validity the animal model should adhere to behaviors that are similar to those established as symptoms of the disorder. This particular validity marker as a sole mode of demonstration of a face validity animal model is sometimes difficult, relying on anthropomorphism; however, in other instances similarities can be rather meaningful. Another validity criterion is construct validity, which focuses on the relatedness of underlying mechanisms to the pathology and symptomatology of the human condition or disorder. On the other hand, predictive validity describes

Table 1 Behavioral features of some current models of depression[a]

Model[b]	*Major symptoms*[c]					
	LA	*MP*	*IMP*	*ANH*	*ANX*	*CI*
Genetic models						
cLH	↓	↓	↑	+	↑	+
FSL	↓	?	?	+	−	±
WKY	↓	?	?	+	↑	?
Genomic models						
$FBGR_{KO}$	↑	?	↑	?	?	?
HPA mutants	?	?	?	?	+	?
SERT	↓	↓	↓	−	↑	?

[a]While the proposed strength of an animal model is often directly related to the number of deficits observed, each model has inherent advantages and disadvantages based on various validity measures of interest. In assessing the strength of any animal model, researchers need to address the dimensions of predictive, face, and construct validities. Face validity refers to phenomenological similarity between behaviors exhibited in the animal model and behaviors that are established as symptoms of the disorder being modeled. Predictive validity is considered by many researchers as particularly salient when determining the effectiveness of a model, since it is determined largely by response of a model to antidepressant drugs. Construct validity for an animal model relates to similarity in neurobiological mechanisms, etiology, or psychological mechanisms. As is the case in humans, behavioral characteristics associated with depressive disorders encompass a range of cognitive deficits and emotional behaviors. The behavioral features listed in this table are among the most important measures used in monitoring treatment effectiveness. A model with several behavioral deficits akin to those seen in human depression would be preferable for global assessment of drug efficacy. On the other hand, models with a single behavioral deficit may be more reliable for assessment of neurochemical mechanisms.

[b]cLH, congenital learned helpless; FSL, Flinders Sensitive Line; WKY, Wistar Kyoto; $FBGR_{KO}$, forebrain, glucocorticoid receptor, knockout; HPA mutants, hypothalamic–pituitary–adrenal system mutation model; SERT, serotonin transporter knockout.

[c]Changes in the listed behavioral features are indicated by the '+' symbol; no changes are indicated by the '−' symbol; a '±' indicate that some studies reported increases while some other studies reported a decrease; a question mark indicates that there are no published results of changes in this particular feature for the depression model listed. Note that the + or − symbol are not indicative of increase or decrease in the parameter featured. Increases and decreases are indicated by ↑ and ↓, respectively; LA, locomotor activity; MP, motivation/persistence; IMP, impulsivity; ANH, anhedonia; ANX, anxiety-like behavior; CI, cognitive impairment.

the extent to which an animal model will facilitate predictions in terms of responses to pharmacological agents used as treatment options. Etiological validity focuses on the underlying mechanisms that are responsible for the biological cause (including hereditability) or development of the behavior in the animal model and human.

Since it is established that genetic and environmental factors are known contributors to depressive disorders, the phenomenon in question in aspiring to the produce the most reliable preclinical model is the weight attributed to each factor, and how these factors act and interact remains to be clarified. Genetic studies with first-degree biological relatives of individuals with major depression demonstrate an enhanced likelihood (1.5–3 times) of suffering from this disorder as compared to the general population, while estimates of the heritability of unipolar depression range from 40% to 70%, which strongly underscores a genetic component. Although other studies demonstrate possible transmission of parental depression, such analysis maybe confounded by environmental contributions, since vulnerable genotypes may facilitate susceptibility to aversive life events and mediate individual differences in depression and its severity.

Given the weight of genotype in this disorder, one of the stronger and more relevant considerations in establishing a preclinical model of depression is heritability. As such, the various genetic animal models of depression have individually augmented our understanding of the neurobiological basis of the disorder and, in some cases, putative treatment paradigms.

Genetics Models of Depression: What Have We Learned So Far?

Animal models of depression have held the promise of providing a dynamic and integrated system that can be manipulated in order to increase our understanding of human mood disorders, test hypotheses about the etiology of these disorders, and potentially develop more effective therapies. However, appropriate models of human depression should fulfill as closely as possible the criteria described in the preceding section: strong phenomenology and similar path physiology (face validity), and comparable etiology (construct validity) and common treatment (predictive validity).

Table 2 Neurochemical and neuroendocrine features of some current models of depression[a]

Process[b]	*Models of depression*[c]					
	cLH	*FSL*	*WKY*	*$FBGR_{KO}$*	*HPA mutants*	*SERT*
Serotonergic mechanisms						
5-HT uptake	+	+	−	?	?	+
5-HT receptors	+	+	−	?	?	+
Noradrenergic mechanisms						
β-Adrenergic receptors	+	−	−	?	?	−
HPA axis regulation						
Basal ACTH levels	+	−	+	+	+	?
Basal CORT levels	+	−	+	+	+	?
Stress ACTH	+	−	+	+	+	?
Stress CORT	+	−	+	+	+	?
Corticosteroid receptor	+	−	+	+	+	?
Receptor signaling	+	−	?	?	+	?
CRH secretion	?	−	?	?	+	?

[a]This table shows neurochemical and neuroendocrine features of the potential animal models discussed in the text. Monoaminergic mechanisms are highlighted, since most antidepressant drugs exert their therapeutic effects by modulating the presynaptic uptake transporters and degrading enzymes of serotonin and norepinephrine, and research on the causality and treatment of depression have focused on the metabolism of functional biogenic amines and the capacity of their respective receptors to alter intracellular signaling pathways and ultimately induce changes in gene activity. Moreover, stress hormones have also been strongly linked to the pathogenesis of affective disorders. Several research groups have formulated a hypothesis relating a dysfunction in stress hormone regulation to the causality of depression. Symptoms that are characteristics of depression include changes in functioning of the hypothalamic-pituitary-adrenocortical system, resulting, in most depressed patients, in altered regulation of adrenocorticotropin and cortisol secretory activity, impairment in corticosteroid receptor signaling, and increased secretion of corticotropin-releasing hormone.
[b]5-HT, 5-hydroxytryptamine (serotonin); HPA, hypothalamic–pituitary–adrenal; ACTH, adrenocorticotropic hormone; CORT, corticosterone; CRH, corticotropin-releasing hormone.
[c]Changes in the listed neurochemical and neuroendocrine features are indicated by the '+' symbol; no change is indicated by the '−' symbol; a question mark indicates that there are no published results of changes in this particular feature for the depression model listed. Note that the + or – symbol is not indicative of increase or decrease in the parameter featured. cLH, congenital learned helpless; FSL, Flinders Sensitive Line; WKY, Wistar Kyoto; $FBGR_{KO}$, forebrain, glucocorticoid receptor, knockout; HPA mutants, hypothalamic–pituitary–adrenal system mutation model; SERT, serotonin transporter knockout.

Twin, adoption, and family studies have suggested that genetic factors are important in the etiology of mood disorders. Hence, one approach to developing genetic models of depression has been to exploit long-term manipulations that simulate predisposition to depression, rather than a depressive response to a precipitating event. Other approaches have been designed by introduction and/or disruption of genes that are involved in the main molecular and biochemical mechanisms associated with depression: the monoamine system, the HPA system and the neurotrophin system.

Genetic Models of Depression

Congenital Learned Helpless Model

Exposure to uncontrollable stress in animals has been linked to significant changes in endocrine activity and central nervous system neurochemistry, increased autonomic system activation, immunosuppression, and a host of behavioral outcomes, including a deficit in learning to escape the uncontrollable stressor. This 'learned helplessness' (LH) phenomenon has been used as a model of depression and has shown good predictive validity for the treatment of human depression, as it is reversed by the administration of electroconvulsive therapy (ECT) as well as by monoamine oxidase (MAO) inhibitors, tricyclic antidepressants, and serotonin reuptake blockers. In many laboratories, learned helplessness is induced by exposure to an inescapable or uncontrollable electric footshock, and when naive Sprague-Dawley rats are subjected to this paradigm, approximately 20% will fail to exhibit escape avoidance behavior. Interestingly, the development of LH behavior is determined by a strong genetic component, as the vulnerability to LH is markedly increased by selective breeding. Indeed, after several generations of selective breeding, a strain of rats, termed congenital learned helpless (cLH), emerged; cLH rats exhibit an increased vulnerability to helplessness even in the absence of prior exposure to uncontrollable stress. In addition to the cLH strain, a control strain of rats was concurrently bred to exhibit a relative resistance to helpless behavior (cNLH). cLH and cNLH rats present a similar behavioral phenotype until challenged

by inescapable shock, and this model simulates an increased risk for precipitation of depression. The congenital learned helpless rat has been validated as a good animal model of depression and has provided a unique model system to identify specific biological markers of clinical psychopathology and to gain further understanding of the etiology of affective disorders.

Systematic investigations of the biochemical, neurochemical, behavioral, and genetic features of congenital learned helplessness over the past 15 years have resulted in the identification of a number of significant differences between the cLH and cNLH strains. For example, the upregulation of hippocampal β-adrenergic receptors and a corresponding downregulation in hypothalamic β-adrenergic receptors, both of which return to normal after antidepressant treatment, have been demonstrated. An imbalance in the serotonin system and a differential regulation of serotonin (5-hydroxytryptamine; 5-HT) receptors are also hallmarks of the cLH strain: increase in 5-HT_{1B} and $5HT_2$ receptors and changes in serotonin uptake and serotonin release are noted. Disturbances in serotonergic mechanisms have been demonstrated in depressed patients. Although it remains unclear by which mechanism serotonin reuptake inhibitors (SSRIs) exert their antidepressant action, there is a clear relationship between chronic SSRI and increased 5-HT levels in the frontal cortex. There is growing evidence to support a key role of the serotonin type 1A (5-HT_{1A}) receptor as a regulator of treatment response. Ressler and Nemeroff proposed that monoaminergic alterations, whether serotonergic or adrenergic, maybe more relevant when these changes are confined to specific brain regions subserving selective symptomatology. For example, these investigators contended that cortical disruption may mediate cognitive deficits, that thalamic sites may account for deficits in arousal status, and that hypothalamic abnormalities are associated with alterations in appetite and libido, and that the cortical–hippocampal–amygdala circuitry accounts for the stress and fear responses. Since the monoaminergic neurotransmitter systems are part of this entire neural circuitry, such a thesis is not unreasonable.

Abnormalities in the regulation of the HPA axis and endogenous glucocorticoid (GC) receptor-mediated responses in the brain have been demonstrated in the cLH strain rats: cLH rats have a selective desensitization of GC-mediated gene expression in the limbic system; in addition, cLH rats exhibit significantly lower endogenous plasma corticosterone levels as compared to cNLH rats and also display a diminished stress response to a variety of stressors. Also, cLH rats undergo a differential response in the development of the HPA axis and the phenotypic/genetic profiles of this rat strain influence the long-term response to challenge after early exposure to stress. Using quantitative cytochrome oxidase histochemistry, we have reported that cortical metabolism in cLH rats does closely resemble that of depressed humans. In the cLH rats, metabolism was found to be decreased in dorsal frontal, medial orbital, and anterior cingulate cortex, but increased in the subgenual region of the cingulate cortex and the same changes have been reported in depressed patients, as determined by positron emission tomography and magnetic resonance imaging techniques. A recent comparison of the levels of gene expression in different signaling molecules and various brain regions of the cLH and cNLH strains has identified interesting changes in the intracellular signaling that may confer vulnerability to mood disorders: reduced cAMP-dependent protein kinase A (PKACα) mRNA expression in hippocampus, prefrontal cortex and striatum; reduced levels of cAMP response element-binding (CREB) protein in the hippocampus; and reduced expression of glycogen synthase-3β (GSK-3β) mRNA. These findings agree well with most previous human and animal studies.

Flinder Sensitive Line Rat Model

Although the Flinders Sensitive Line (FSL) rat developed in 1993 by Overstreet and colleagues emerged as an eminent selectively bred animal model of depression, the researchers were initially in pursuit of another goal. Indeed the FSL project was initiated to develop an animal resistant to the pesticide diisopropylfluorophosphate (DFP), a long-acting cholinesterase inhibitor. However, the cholinergic–adrenergic theory of depression and mania, which proposed that cholinergic abnormalities such as supersensitivity to anticholinesterase agents were one of the underlying causes of depression, led to the assessment of the FSL model for the reaction of FSL rats to antidepressant drugs.

Over time, the FSL strain has exhibited many relevant features of depression, including psychomotor retardation, reduced appetite and weight repression, elevated rapid eye movement (REM) sleep, and reduced REM sleep latency. A key feature of this model is its excellent predictive validity. The FSL animal is less mobile on the forced swim test (FST), in which increased immobility is associated with 'depressive-like' behavior. More importantly this strain shows decrements in immobility following chronic, but not acute, treatment with desipramine and sertraline, compared to controls. This pharmacological response has been shown to be selective to antidepressants, since psychomotor stimulants such as amphetamines are

ineffective. Furthermore, rolipram and a melatonin agonist were shown to have antiimmobility effects in the FSL rats and later to have antidepressant effects in humans. Thus, the FSL rat model of depression exhibits some behavioral, neurochemical, and pharmacological features that have been reported in depressed individuals, and has been very effective in detecting antidepressants.

Finally, neurochemical and/or pharmacological evidence suggests that the FSL rat exhibits changes consistent with the cholinergic, serotonergic, dopaminergic, neuropeptide Y (NPY), and circadian rhythm models, but not with the noradrenergic, HPA axis, or γ-aminobutyric acid (GABA)ergic models of depression. Since many of these abnormalities have been demonstrated in humans, and the Flinders rats respond well to many classes of antidepressant drugs, the FSL rat model has been put forth as a particularly effective predictive model, and as such may be used for detecting novel antidepressants.

Wistar Kyoto Model

The Wistar Kyoto (WKY) rats were originally bred as normotensive controls for the spontaneous hypertensive line (SHR) rats. These rats have been used extensively in physiological studies relating to hypertension, but were also shown to exhibit a battery of distinct behavioral features. These behavioral alterations appeared concordant with accepted face validity markers of depression. For example, the WKY rat exhibits increased immobility duration in the forced swim test (a test used as a marker for 'behavioral despair'). More importantly, the increased immobility duration in the swim test was prevented by chronic (but not acute) antidepressant treatment. WKY rats also exhibit decreased activity in the open-field test (a test for novelty exploration and anxiety) and enhanced 'anhedonia' (as measured by decreased consumption of a sugar-based solution) in response to chronic mild stress and acute stress.

In addition to behavioral measures, the WKY strain also demonstrates endogenous central neurotransmitter alterations, including monoaminergic abnormalities, with dopaminergic and noradrenergic neurotransmitter systems and thyroid-stimulating hormone systems, compared with control animals. However, a critical feature of this animal model is HPA dysregulation, manifested as alterations in basal plasma levels of ACTH and corticosterone (CORT), as well as enhanced responses to stressors as measured by exaggerated ACTH and CORT levels postexposure to both acute and chronic stress. Such a composite of neuroendocrine and behavioral hyperresponsiveness to stress , along with the co-occurrence of anxiety features in this model, facilitates the uniqueness of the WKY rat as both a genetic and stress-induced animal model of depression.

Genomic Models of Depression

Corticosteroid Receptor Transgenic and Knockout Mice

A dysfunctional HPA axis is a common feature of patients with major depression, and this HPA dysfunction has been associated with a reduction in glucocorticoid receptor (GR) number and function. Clinical studies that support this hypothesis include reports of a reduction in GRs in the peripheral tissues of depressed patients and the successful treatment of psychotic depression with the glucocorticoid receptor antagonist, mifepristone.

Some recent preclinical studies hypothesize that a primary defect in forebrain GRs will alter both behavioral and physiological responses to stress. Consequently, investigators have used the homologous recombination approach to generate mice that are heterozygous for GR deletion. These mice display a marked reduction in the expression of GRs but also exhibit a behavioral phenotype that is sensitive to stress and parallels depression. Recent studies have also described the production of knockout mice with conditional disruption of forebrain (FB) GRs (thus named $FBGR_{KO}$ mice) that show increased depression-related behavior and impairments in negative feedback regulation of the HPA axis. In these mice, GRs are disrupted in the hippocampus and cortex but not in the central nucleus of the amygdala, paraventricular nucleus (PVN), or pituitary. $FBGR_{KO}$ mice exhibit increased locomotor activity, increased impulsivity when exposed to stress, and increased despair-related behavior in the tail suspension and forced swim test.

Transgenic mice expressing GR antisense mainly in neuronal tissue have also been tested as potential animal models of depression. These mice exhibit significant impairment in GR function in the brain. Moreover, several behavioral and neuroendocrine symptoms common to human depression have been exhibited in these mice, including reduced negative feedback sensitivity to dexamethasone and enhanced stress response to pharmacological and emotional stimuli.

HPA-System Mutation Models

HPA dysregulation is a central recurring theme in most genetically engineered models currently utilized to study depression. Given the prominence of HPA dysregulation in depression and the absence of a specific gene target linked to depression, this approach

has been valuable, yet challenging. Several mutants with distinct alterations in the corticotropin-releasing hormone (CRH) neuronal system have been generated and tested in the past decade, and the feasibility of generating a homozygous CRH knockout mouse was demonstrated as early as 1995 by Muglia and colleagues. Characterization of this model has demonstrated no gross behavioral deficits or viability issues. As expected, these animals exhibited a compromised but functional HPA axis, with compensatory mechanisms facilitating vasopressin release in response to decreases in ACTH. Similarly, in models with CRH overexpression, there are no correlates of specific depressive-like behaviors, but links to anxiety- and fear-related behaviors have emerged. These findings have lead researchers to develop and test the impact of a CRH antagonist on depression. Outside of CRH, there are other neuropeptides in the CRH family, namely the urocortins, that researchers propose may be orchestrating some of these responses. Given that urocortin 1 has a high affinity for both CRH receptors CRHR1 and CRHR2, its action is similar to that of CRH. Mutations in urocortin 1, however, did not alter the HPA axis, and its impact on anxiety remains debatable. Due to the complexity of depressive disorders, the use of mice models with a single gene variant is proving challenging as a viable construct to explore depression. Newer strategies for conditional inactivation or overexpression of factors involved in the HPA axis are emerging to facilitate better understanding of the molecular targets involved. The identification and characterization of mice deficient for abcb1-type (multidrug resistance gene1-type) P glycoprotein – which controls access of endogenous glucocorticoids into the central nervous system, thereby regulating all components of the HPA axis – may prove useful in the exploration of genetics, HPA regulation, and depression.

Serotonin Transporter Knockout

Among the three monoamines implicated in depression, serotonin (5-HT) has been the most extensively studied. Hence, alterations of the molecular mechanisms that modulate the levels of 5-HT at the synaptic cleft have provided a new avenue to create different mouse lines with potential as models of depression. One such line, the serotonin transporter (SERT) knockout, is based on disruption of the serotonin reuptake transporter (5-HTT) gene, leading to critical changes in the inactivation and control of serotonergic neurotransmission. Mice lacking the SERT gene exhibit excess 5-HT during embryonic development but display a marked reduction in 5-HT in adulthood. The SERT knockout mice also show marked reduction from wild-type mice in 5-HT_{1A} receptor protein and mRNA expression in the dorsal raphe, hypothalamus, amygdala, and septum, but marked increase in the hippocampus. Similarly, 5-HT_{2A} receptor density is also reduced in 5-HTT knockout mice.

Clinical and epidemiology studies have supported the involvement of serotonin neurotransmission in the etiology of depression. Hence, in certain clinical populations, a 44-base-pair sequence in the serotonin transporter promoter (SERT-linked promoter; SERTLPR) has been associated with increased vulnerability to affective disorder. Individuals with one or two copies of the short allele of the SERT promoter polymorphism show more depressive symptoms. Moreover, SERT is downregulated in the brains of depressed patients. These studies have prompted the systematic assessment of the behavioral phenotype of the SERT knockout as a potential model of human depression and anxiety. The SERT knockout mouse display a pattern of hypolocomotion, increased anxiety-like behaviors, reduced aggression, and reduced social interaction. Interestingly, the SERT mouse showed an increased propensity to learned helpless behavior. However, there were no differences between SERT knockout and wild-type mice in an anhedonia-based paradigm, the sucrose preference test.

In summary, while each animal model discussed herein has some behavioral features that parallel human depression, the obvious limitation of mutant models is that they have a single, possibly region-specific, mutation, whereas human depression has a multifactorial and multigenetic pathology. An additional caveat is the lack of a comprehensive assessment of the impact of genetic predisposition, early life stress, and acute or chronic stress exposure in these models. So far, none of the mouse mutants could be viewed as animal models of depression that fulfill all three criteria that reproduce aspects of human pathology: face validity (phenomenological and pathophysiological similarity), construct validity (comparable etiology), and predictive validity response to common treatment. However, each of the three mutant lines discussed herein offer opportunities to study selected symptoms such as anxiety, stress response, and a number of complex behaviors associated with HPA disturbances.

Future Directions

Translational Research Approaches: Optimal Utilization of Genetic Models of Depression

While we have made significant strides in understanding the molecular, cellular, and genetic bases of mood disorders, including depression and manic depression mechanisms, the development of new drugs that would provide effective therapies for a heterogeneous

population of depressed subjects is still lagging. Large subsets of depressed patients do not respond to any of the currently available therapies, and the root of failure in new drug development has been attributed to a mismatch of appropriate drug and patient, inappropriate dose and dosing schedule, undetectable effects, and species differences. Translational research in the genomic era seeks to minimize the risk associated with failure in drug development and to ensure that development and registration of new compounds are allocated optimally.

It will be difficult to fully realize the benefits of the genomic era, without the establishment of strong partnerships between the academic communities and preclinical and clinical researchers from industry, all having the common goal to identify the relevant endophenotypes of chronic depression and to develop appropriate test models for this condition. It is also necessary that clinical trial designs go beyond the standard measurement scales (e.g., Hamilton Rating Scale of Depression; HAM-D) and include endpoints more relevant to the behavioral and physiological measures done in animal models, so that better estimates of causality can be tested.

The current emphasis on translational research promises to be an important development in achieving preclinical and clinical congruence of results. We currently have evidence that some procedures used in animal research can be used in humans. For example, the use of progressive ratios for monetary reward in depressed patients can be related to the use of this schedule of reinforcement for more basic rewards in animals. Similarly, anhedonia is considered as a candidate behavioral endophenotype of depression, and rodent models of learned helplessness and chronic mild stress (CMS) are characterized by specific behavioral abnormalities such as anhedonia, reduced activity, and passivity.

Three basic technologies are currently being used in translational research programs focused on therapy development: gene-based target discovery, combinatorial chemistry, and high-throughput screening. In translational research for neuropsychiatric disorders, a major goal of gene-based target selection has been to break out of the box of monoamine-based therapies. The bioinformatics screening of the human genome provides a rich alternative source of potential targets and validation strategies. For drug discovery efforts for depression, the use of species such as the flatworm (*Caenorhabditis elegans*) has led to the identification of new genes that can suppress behaviors such as egg-laying and nose contraction that are not affected in serotonin-defective mutants. In addition, the use of combinatorial chemistry allows the systematic generation of multiple variants of small organic molecules that can be assayed for interaction against a chosen drug target. High-throughput screening also allows the screening of several hundred thousand small molecules in a short time and provides a useful start for analysis of structure–function relationships that will guide the optimization of lead compounds.

The use of these technologies needs to be coupled with testing of new targets in disease models in rodents, including those of directed mutants. Promising directions include collaborative efforts between industry, academia, and funding agencies, to provide a systematic phenotypic evaluation of large numbers of novel mutant mouse strain; increasing use of genetic models to identify genes associated with variable responses to drugs; combination of forward and reverse genetics and pharmacology to identify and validate drug targets, as well as to discover and develop new drugs; and strategies that identify individuals vulnerable to specific phenotypes and relate these to animal models and then exploit this vulnerability as a factor to discover candidate genes.

Gene Expression Profiling

A promising approach in increasing our understanding of the pathophysiology of depression is the generation of gene expression profiles in cellular and animal models as well as in clinical samples. Before the development of microarray technology, investigators examined the mRNA or protein levels of only a few genes at a time. Typically, these genes were selected on the basis of a hypothesis or a prediction regarding a possible mechanism and potential therapeutic strategies. With current advances in mRNA profiling, such as DNA microarray, it is possible to perform genome-wide gene expression analysis in a hypothesis-free manner. One caveat of such studies in complex human disorders, such mood disorders, will be the large number of genes that would have to be validated for their potential as causative factors or as novel drug targets. The use of gene expression profiling has already allowed us to identify unknown molecular cascades involved in mental disorders. For example, Aston and others have reported that oligodendrocyte-related genes are downregulated in major depression. Choudary and others have hypothesized that astroglial dysfunction plays a role in major depression, based on their findings of downregulation of glial high-affinity glutamate transporters SLC1A2 and SLC1A3, as well as glutamate ammonia ligase, which are predominantly in astroglia. By contrast, no changes in gene expression have been identified in the prefrontal cortex of any patient with major depression who committed suicide.

While the analysis of expression profiling holds the promise of generating comprehensive and comparable data sets in complex disorders, microarray studies of the brain are quite challenging. There are still significant limitations in detection sensitivity and dilution of signal differences due to cellular heterogeneity masking disease effects. With new technology such as laser-capture microdissection, and signal amplification protocols that require small amounts of starting materials, the technique should overcome the limitations of postmortem brain samples. One area that needs special attention is the utilization of gene profiling data to reveal drug treatment effect on gene expression. A number of key components, such as selection of animals, choice of drug, dosage, method of administration, duration of treatment, and environment, will have to be fully investigated for their potential effect on gene expression profiles.

Despite all of the shortcomings of the currently available gene profiling studies, this field holds great promise for the treatment of depression by facilitating genetic prediction to response and identification of drug targets. Moreover, the technology of expression profiling at the protein level is rapidly developing and we should soon be able to superimpose gene expression profiling on data sets obtained from proteomics, epigenetic, and polymorphism data sets.

Imaging Approaches

There are many different imaging approaches being utilized to study depression in humans, including magnetic resonance imaging (MRI), functional MRI (fMRI), and positron emission tomography (PET). Imaging approaches are particularly useful for exploring the circuitry, chemistry, and neuroanatomy underlying this complex disorder. Current MRI studies have led to a debate about the role of hippocampal volume in depression from lack of association to a reduction in volume. Several PET studies show an association between enhanced severity of depressive symptoms with decreased metabolism of the anterior cingulate cortex and dysfunction in the monoaminergic systems. However, many of the functional imaging studies in depression assess emotional responses. Such studies suggest greater activation of the left amygdala by fearful faces in acutely depressed patients while others demonstrate greater bilateral amygdala activation with aversive visual stimuli in patients with depression. In general, fMRI studies have proposed that experimental or pathological changes in mood state are associated with reciprocal changes in activity of limbic–subcortical and frontoparietal systems. Researchers hypothesize that the limbic–subcortical regions are more activated, and frontoparietal circuits are less activated, by sadness, and that effective antidepressant agents correct this abnormal responsivity in patients with depression.

A few imaging studies have provided additional support for the validity of some animal models of depression. For example, the congenital learned helpless rat shows activation patterns in frontal and cingulated regions of the brain that are comparable to those seen in humans. PET studies in the cynomolgus monkey model of depression demonstrate a correlation between behavioral depression and change in serotonin receptor binding potentials. Imaging studies offer the ability to use noninvasive methods in genetically vulnerable individuals and animal models to explore further the underlying biological mechanisms and neurocircuitry subserving complex disorders. It is, however, important to be aware of the importance of selecting the appropriate genetic background for the development of these animal models, due to the potential variability of morphometric and metabolic endophenotypes.

See also: Depression and the Brain; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness.

Further Reading

Crowley JJ and Lucki I (2005) Opportunities to discover genes regulating depression and antidepressant response from rodent behavioral genetics. *Current Pharmaceutical Design* 11(2): 157–169.

De La Garza R and Mahoney JJ (2004) A distinct neurochemical profile in WKY rats at baseline and in response to acute stress: Implications for animal models of anxiety and depression. *Brain Research* 1021: 209–218.

Edwards E (1992) *In vitro* neurotransmitter release in an animal model of depression. *Neurochemistry International* 21: 29–35.

Edwards E, Harkins K, and Henn FA (1991) Learned helplessness modulation of ^{3}H-paroxetine binding in the rat brain. *Journal of Neurochemistry* 56: 1581–1586.

Edwards E, King JA, and Fray JC (1999) Increased basal activity of the HPA axis and renin–angiotensin system in congenital learned helpless rats exposed to stress early in development. *International Journal of Developmental Neuroscience* 17(8): 805–812.

Hasler G, Drevets WC, Manji HK, et al. (2004) Discovering endophenotypes for major depression. *Neuropsychopharmacology* 29(10): 1765–1781.

Henn FA, Edwards E, and Johnson JO (1987) Research directions in behavioral medicine. In: Hellhammer D, Florin I, and Weiner H (eds.) *Neurobiological Approaches to Human Disease*, pp. 215–224. Toronto: Hans-Huber.

Henn FA and Edwards E (1994) Animal models in the study of genetic factors in human psychopathology. In: Papolos DF and Lachman HM (eds.) *Genetics of Affective Disorders*, pp. 177–192. New York: Brunner-Mazel.

Holsboer F (2000) The corticosteroid receptor hypothesis of depression. *Neuropsychopharmacology* 23(5): 477–501.

King JA, Campbell D, and Edwards E (1993) Differential development of the stress response in congenital learned helplessness. *International Journal of Developmental Neuroscience* 11(4): 435–442.

McArthur R and Borsini F (2006) Animal models of depression in drug discovery: A historical perspective. *Pharmacology Biochemistry and Behavior* 84(3): 436–452.

Nestler EJ and Carlezon WA Jr. (2006) The mesolimbic dopamine reward circuit in depression. *Biological Psychiatry* 59(12): 1151–1159.

Overstreet DH, Pucilowski O, Rezvani AH, et al. (1995) Administration of antidepressants, diazepam and psychomotor stimulants further confirms the utility of Flinders Sensitive Line rats as an animal model of depression. *Psychopharmacology (Berlin)* 121(1): 27–37.

Shumake J, Poremba A, Edwards E, et al. (2000) Congenital helpless rats as a genetic model for cortex metabolism in depression. *NeuroReport* 11: 3793–3798.

Willner P and Mitchell PJ (2002) The validity of animal models of predisposition to depression. *Behavioral Pharmacology* 13(3): 169–188.

Stress Response: Neural and Feedback Regulation of the HPA Axis

J P Herman, University of Cincinnati, Cincinnati, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

The Hypothalamic–Pituitary–Adrenal Axis

The hypothalamic–pituitary–adrenal (HPA) axis is essential for adaptation to physical or psychological stress. Glucocorticoid hormones, which are the ultimate product of HPA axis activation, act on multiple bodily systems to maintain homeostasis. Many of the effects of glucocorticoids are beneficial for short-term survival but can be counterproductive or even deleterious if prolonged. Therefore, the activation and inhibition of glucocorticoid is a temporally regulated process involving rapid neuronal activation and efficient feedback inhibition.

Stimulation of the HPA axis occurs in reaction to or in anticipation of physiologic challenge. Reactive or 'reflexive' release of glucocorticoids occurs in response to physiological challenges (e.g., blood loss, hemorrhage). This type of response is initiated by stimuli that signal a direct threat to homeostasis or survival and can be triggered without conscious perception. However, it is important to note that organisms also generate 'anticipatory' HPA axis responses to environmental cues associated with threats to survival or homeostasis. Thus, stimuli that predict adversity can generate a glucocorticoid response in the absence of an existing physiologic insult. These responses are generated by comparison of environmental stimuli with either innate programs (e.g., instinctual fear of predators) or memories (e.g., prior experience with a painful or fearful stimulus). The relevance of the anticipatory response hinges on the predicted need for an adaptive hormonal response, with the glucocorticoid secretory process begun in preparation of a survival challenge (prolonged flight, fight, and injury).

Glucocorticoid release is commonly associated with an inferred state of 'stress.' From the perspective of this article, 'stress' can be defined as 'a real or perceived threat to homeostasis or well-being.' There are four somewhat distinct 'stress' responses: behavioral change, activation of the sympathoadrenomedullary system, activation of the HPA axis, and activation of the immune system, roughly in order of threshold and timing. The glucocorticoid stress response addressed in this article is but one component of an integrated set of responses to stress.

The hypothalamic paraventricular nucleus (PVN) is responsible for regulation of HPA axis stress responses (Figure 1). Activation of the HPA axis is mediated by neurosecretory neurons localized in the medial parvocellular zone of this nucleus. These neurons project to blood vessels in the external lamina of the median eminence, where they release numerous adrenocorticotropin (ACTH) 'secretagogues,' the most important of which are corticotropin-releasing hormone (CRH) and arginine vasopressin.

The medial parvocellular PVN is in receipt of synaptic innervation from a relatively circumscribed set of central nervous system structures, summarized in Figure 2. In general, PVN projection neurons are localized in regions known to receive first- or second-order inputs from somatic nociceptors, visceral afferents, or humoral sensory pathways. Thus, the majority of PVN-projecting neurons are positioned to evoke rapid, reflexive activation of the HPA axis. These PVN projection neurons can in turn be influenced by inputs from cortical and limbic structures, which are likely to transduce anticipatory activation.

Glucocorticoids released following stress exert classical negative feedback effects on the HPA axis. Feedback effects of glucocorticoids occur in distinct time domains. Fast feedback occurs within minutes and is nongenomic in nature. Fast feedback is sensitive to the rate of change in circulating glucocorticoid levels and is essential for rapid termination of PVN activation and ACTH release normally seen during an acute stress response. The nongenomic mechanism involves membrane glucocorticoid receptors. Delayed feedback occurs over hours to days and involves genomic mechanisms. Delayed feedback is sensitive to the levels of circulating glucocorticoids rather than the rate of rise.

Glucocorticoid negative feedback may be mediated at multiple levels. The nuclear glucocorticoid receptor is though to regulate genomic negative feedback, perhaps in conjunction with permissive actions of the mineralocorticoid receptor, which binds glucocorticoids with high affinity and is extensively bound even under unstressed conditions. The glucocorticoid receptor is abundantly expressed in the pituitary, PVN, and numerous brain regions implicated in HPA axis regulation, all of which may contribute to negative feedback.

It is important that regulation of the HPA axis can be mediated by peripheral feedback factors other than glucocorticoids. For example, ingestion of a caloric load (e.g., sucrose consumption) can inhibit activation of the HPA axis independent of glucocorticoid secretion. The neural mechanism subserving this alternative feedback pathway remains to be determined.

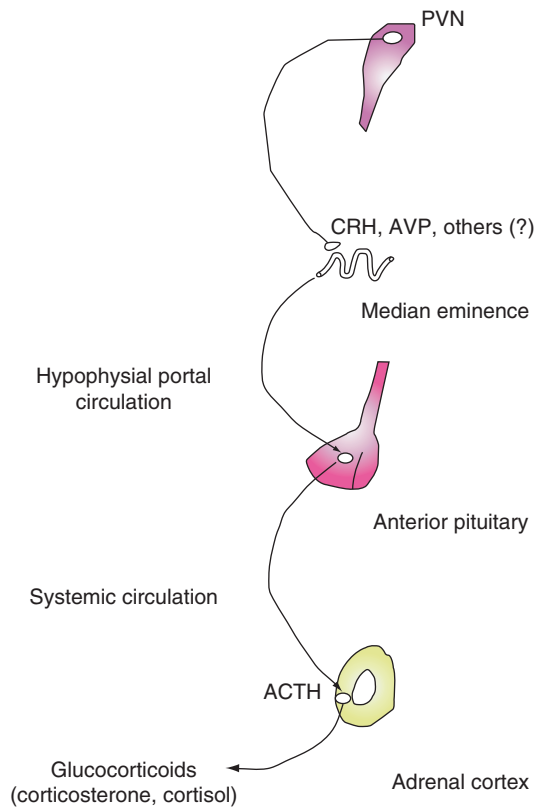

Figure 1 Schematic of the hypothalamic–pituitary–adrenal (HPA) axis. Stress triggers activation of paraventricular hypothalamic nucleus (PVN) neurons, causing release of corticotropin releasing hormone (CRH), and co-secretagogues (such as arginine vasopressin (AVP) in the portal vessels of the median eminence. Secretagogues travel via the portal blood to the anterior pituitary, where they elicit secretion of adrenocorticotrophic hormone (ACTH). ACTH travels via the systemic circulation to the adrenal cortex, causing synthesis and secretion of glucocorticoids.

Direct Pathways Controlling HPA Axis Activation

Brain Stem

The PVN receives a large input from noradrenergic and adrenergic neurons of the nucleus tractus solitarius (NTS). Lesions of these neurons or their terminals attenuate HPA axis stress responses to reflexive stressors, whereas direct adrenergic receptor stimulation activates CRH and ACTH secretion, confirming that norepinephrine (and perhaps epinephrine) neurons are necessary and sufficient for stress activation of the HPA axis. However, the influence of NTS norepinephrine–epinephrine on anticipatory stress responses is not observed consistently, suggesting that anticipatory excitation occurs by other mechanisms.

The NTS also contains a noncatecholaminergic PVN-projecting cell population. Many of these cells contain important peptidergic neuromodulators (e.g., glucagon-like peptide-1, neuropeptide Y) that are HPA axis excitatory, and there is evidence that these peptides may participate in generation of both reflexive and anticipatory stress responses.

Serotonergic systems also participate in HPA axis activation by way of projections from the midbrain and pontine raphe nuclei. Serotonergic innervation of the PVN is more circumscribed than that of the NTS, and modulation of extra-PVN serotonin can also have potent actions on HPA axis activity. The relative strengths of direct versus indirect serotonin serotonergic modulation of the stress response remain to be resolved.

Lamina Terminalis System

The PVN is in direct receipt of information on fluid and electrolyte status by way of the subfornical organ–lamina terminalis system. Signals are relayed through direct projections of the subfornical organ, median preoptic nucleus, and organum vasculosum of the lamina terminalis to medial parvocellular PVN neurons. These regions are critically involved in control of osmoregulation and drinking behavior, and this pathway may be critical for initiation of HPA responses to thirst or dehydration.

Hypothalamus

The PVN receives massive input from numerous regions of the hypothalamus, a large component of which is γ-aminobutyricacidergic (GABAergic; inhibitory). GABA inputs to the PVN originate in numerous hypothalamic nuclei concerned with homeostatic regulation, including the dorsomedial hypothalamus (autonomic function), preoptic area (thermoregulation), arcuate nucleus (energy balance), and lateral hypothalamus (ingestive behavior).

The PVN also receives input from hypothalamic neuropeptidergic (e.g., corticotrophin releasing factor, melanocortin) and glutamatergic neurons. These inputs are presumably excitatory and may serve to activate the HPA axis in response to homeostatic imbalance. Thus, hypothalamic inputs are sufficient to both excite and inhibit the PVN, consistent with a role in maintenance of homeostasis.

Bed Nucleus of the Stria Terminalis

Direct nonhypothalamic forebrain inputs to the parvocellular PVN are largely confined to discrete subnuclei of the bed nucleus of the stria terminalis (BST). The influence of the BST on HPA function is complex. Posterior regions of the BST inhibit HPA axis responses to stress, whereas ventrolateral PVN projecting neurons provide excitatory input. This complexity may be related to differential limbic inputs to these regions, as noted below.

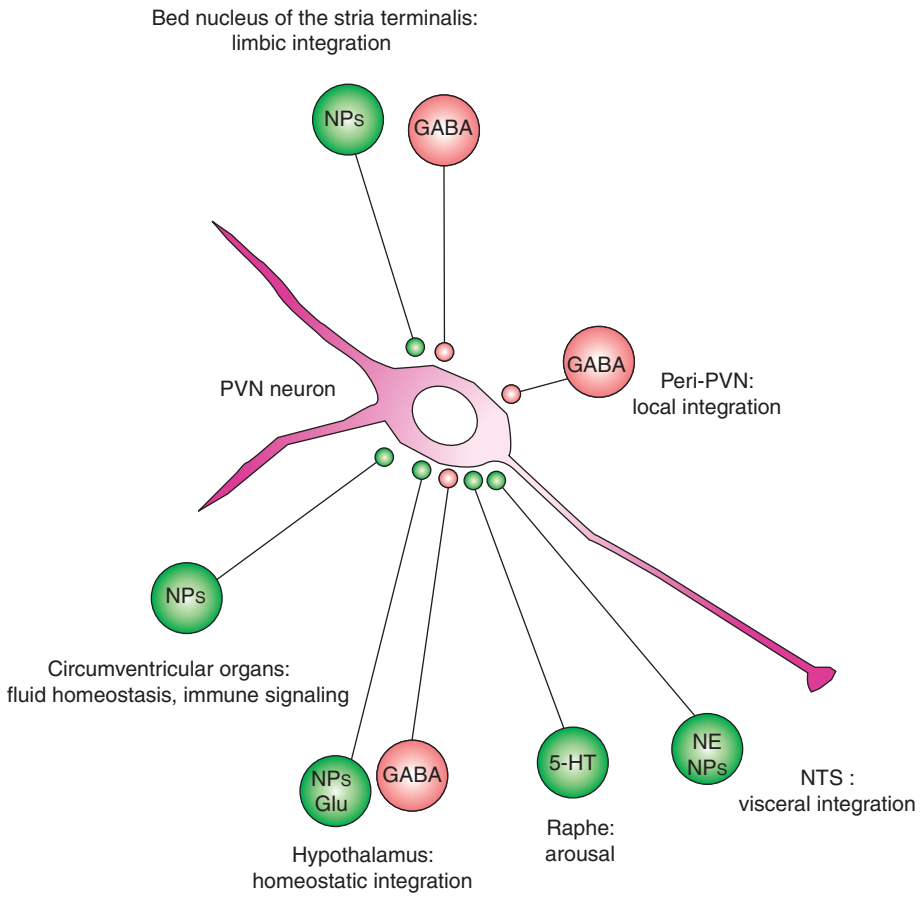

Figure 2 Direct projections to the paraventricular hypothalamic nucleus (PVN). Excitatory inputs are in green, inhibitory in pink. The PVN receives predominantly excitatory inputs from brain stem noradrenergic (NE), neuropeptidergic (NP), and serotonergic (5-hydroxytryptamine; 5-HT) cell groups, as well as from circumventircular structures (such as the subfornical organ). Inputs from the hypothalamus and bed nucleus of the stria terminalis can be excitatory (glutamate (Glu) or NPs) or inhibitory (γ-aminobutyric acid; GABA), depending on location. Neurons in the immediate vicinity of the PVN may serve as local inhibitors of HPA axis activation. NTS, nucleus tractus solitarius.

Thalamus

Neurons in auditory-responsive posterior thalamic nuclei, including the subparafascicular and posterior intralaminar regions, send direct projections to the medial parvocellular PVN and are likely involved in pathways mediating activation of the HPA axis by auditory stimuli.

Local PVN Information Processing

The PVN is well positioned to receive direct input from blood- and CSF-borne factors. The PVN has a dense nonfenestrated capillary plexus that can allow ready access to blood–brain barrier permeable factors, including glucocorticoids. Thus, the PVN is one site of potential glucocorticoid negative feedback regulation of the HPA axis. Brain microvessels may also play a major role in communicating immune (cytokine) signals to the CNS, through activation of prostaglandin biosynthesis. Notably, the medial parvocellular PVN richly expresses prostaglandin receptors, suggesting a capacity for local integration of immune signals within the PVN proper.

Neurons of the PVN have limited dendritic extensions that largely remain within the nucleus proper. This morphological arrangement supports intranuclear communication among different PVN subdivisions (e.g., magnocellular divisions projecting to the neurohypophysis), which may subserve intrinsic modulation of excitability. The restricted dendritic branching scheme also prevents parvocellular PVN neurons from sampling the peri-PVN region, which is richly invested with terminals and axons from limbic structures such as the hippocampus, prefrontal cortex, and lateral septum. This arrangement likely prevents direct modulation of HPA axis function by descending limbic structures.

The peri-PVN region contains large numbers of GABAergic neurons, some of which project into the PVN proper. The peri-PVN region receives rich input

from limbic structures, hypothalamic cell groups, and serotonin neurons, suggesting that it may play a role in 'gating' of information coming into the PVN.

Indirect Paths to the PVN

Limbic brain structures modulate anticipatory HPA responses to stress. These structures, including regions such as the hippocampus, prefrontal cortex, amygdala, septum, and midline thalamus, are critical for emotional responses and memory and are logical candidates for modulating pituitary–adrenal secretions with respect to prior experience. These regions have no substantial direct interactions with medial parvocellular PVN neurons, requiring intermediary neurons to relay their influence on ACTH secretagogue release (see **Figure 3**).

Hippocampus

The hippocampus is involved in terminating HPA axis responses to stress. Accumulating evidence indicates that the hippocampus is involved in inhibiting anticipatory but not reflexive stress responses, perhaps consistent with its role in memory and information processing. The observation that hippocampal damage can delay termination of corticosterone release is consistent with a role in glucocorticoid negative feedback inhibition of the HPA axis. This hypothesis is supported by rich localization of glucocorticoid receptors in the hippocampus and by functional studies demonstrating diminished feedback efficacy following hippocampal lesions, local glucocorticoid receptor inactivation, or hippocampal–cortical glucocorticoid receptor gene deletion.

Anatomical tracing studies reveal that the hippocampus can communicate with the PVN by way of relay neurons in the hypothalamus and BST. Given that the output of the hippocampus is largely excitatory, it is likely that inhibition is mediated by activation of inhibitory, PVN-projecting GABAergic neurons in these regions.

Medial Prefrontal Cortex

The medial prefrontal cortex also modulates HPA axis activation, albeit in a more complex fashion. Numerous studies indicate that dorsal components of the medial prefrontal cortex inhibit anticipatory but not reflexive HPA axis responses and may be another site for glucocorticoid negative feedback signaling. Conversely, the more ventral components of

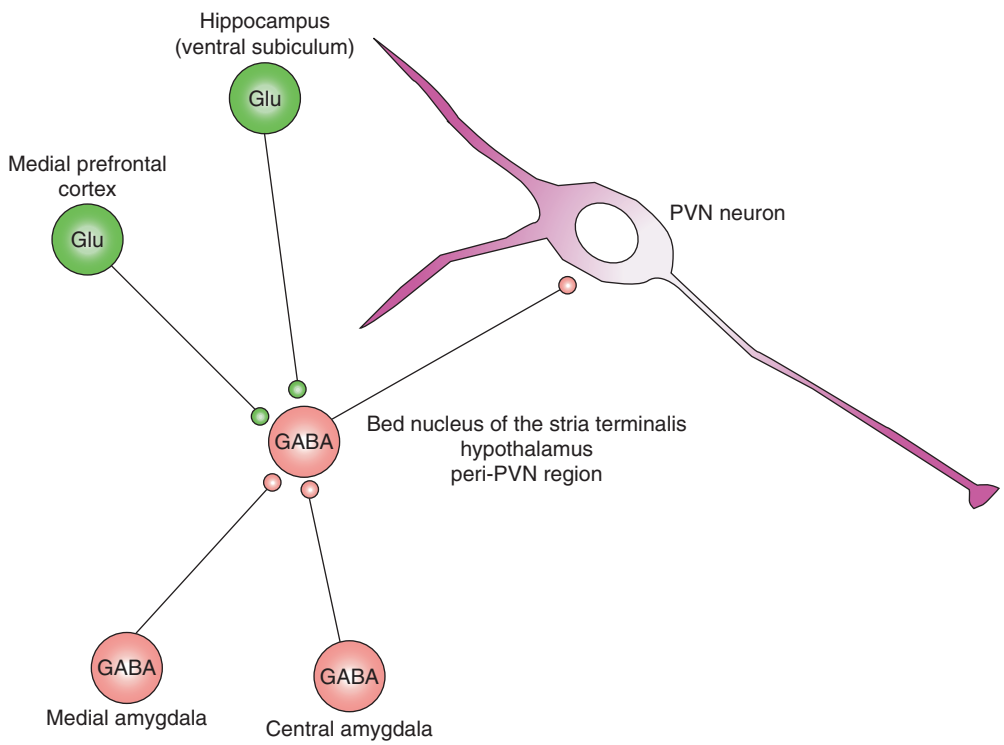

Figure 3 Simplified schematic of possible bisynaptic limbic connections with the paraventricular hypothalamic nucleus (PVN). Regions such as the hippocampus and medial prefrontal cortex inhibit the HPA axis by supplying excitatory glutamatergic innervation of PVN-projecting GABAergic neurons in the bed nucleus of the stria terminalis, hypothalamus, and peri-PVN region (see **Figure 2**). In contrast, amygdalar activation of the HPA axis is likely to occur by disinhibition, involving GABAergic inhibition of GABA projecting PVN neurons. Glu, glutamate.

the medial prefrontal cortex (e.g., infralimbic cortex) heavily innervate subcortical autonomic sites and may be involved in stress excitation. There is also some evidence that ventral medial prefrontal cortex excitation of the stress response is lateralized in the rat (driven by the right hemisphere).

Anatomical studies indicate that the prefrontal cortex does not innervate the PVN to any substantial degree. Dorsal regions of the medial prefrontal cortex project to regions such as the BST as well as a number of non-PVN projecting regulatory regions, whereas ventral regions (encompassing the infralimbic cortex) innervate autonomic regions such as the ventral BST, central amygdala, and NTS. Thus, like the hippocampus, HPA regulation by the medial prefrontal cortex is indirect.

Amygdala

In contrast with the hippocampus and medial prefrontal cortex, amygdaloid nuclei appear to stimulate the HPA axis. However, there is considerable functional differentiation among individual amygdalar regions. The central amygdaloid nucleus is involved in activation of the HPA axis to reflexive but not anticipatory stressors, whereas the medial amygdala appears to selectively involved in generation of anticipatory responses.

Projection neurons of the central and medial amygdala are primarily GABAergic and do not directly innervate the PVN. However, both regions project heavily to the BST and preoptic area, which contain populations of GABAergic PVN-projecting neurons. Therefore, excitatory effects on HPA axis stress responses are likely mediated by inhibition of GABA neurons in the hypothalamus or BST, effectively resulting in activation by disinhibition.

Lateral Septum

Lateral septal lesions are known to produce a constellation of behavioral and autonomic symptoms consistent with extreme anxiety ('septal rage'). Not surprisingly, this region inhibits HPA axis stress responses. Neurons in this area are activated by a variety of 'anticipatory' stressors but show little induction following activation of the 'reactive' stress pathway, suggesting selective involvement in the former process.

The anatomy of the septal region predicts prominent interactions with hypothalamic and midbrain PVN-projecting regions. As the majority of lateral septal neurons express GABA as their primary transmitter, it is likely that the HPA inhibitory effects of this region are mediated by inhibition of excitatory projections to the PVN.

Thalamus

Several recent reports implicate midline limbic thalamic nuclei in HPA axis regulation. Of the known stress-responsive thalamic nuclei, the paraventricular thalamus appears to play a major role in integrating HPA axis responses to repeated stressors, mediating both stress habituation and sensitization of responses to new stressors. The influence of the paraventicular thalamus is likely to be communicated by way of hypothalamic nuclei (e.g., dorsomedial nucleus).

Hypothalamus

There are several hypothalamic regions that appear able to interact with the PVN through intrahypothalamic relays. Prominent among these is the suprachiasmatic nucleus, which is the major integrator of circadian HPA rhythmicity, and the ventromedial nucleus, which is concerned with energy balance. Both regions interconnect with the PVN through intrahypothalamic relays, including the dorsomedial nucleus and lateral hypothalamus.

Brain Stem

Noradrenergic neurons of the locus coeruleus do not project directly to the medial parvocellular PVN. Nonetheless, lesions of this cell group reduce HPA axis stress responses, implying a transsynaptic excitatory action. This may also be the case for serotonin, as reduction of serotonergic neurotransmission in the amygdala reduces HPA axis responses to stress. In addition, the excitatory action of brain stem acetylcholine on CRH and ACTH release appears to be mediated by norepinephrine neurons of the NTS.

Principles Governing Neurocircuit Regulation of the HPA Axis

The organization of PVN inputs indicates that stress responses are dependent on the overall 'set' of central sensorial input, involving synthesis of homeostatic and cognitive information. Circuits that process reflexive responses have direct access to PVN neurons and tend to be excitatory in nature. These pathways are positioned to communicate sensory information regarding homeostatic imbalance, consistent with a mechanism for rapid adaptation to adversity. \In contrast, anticipatory responses appear to be activated by disinhibition, whereby amygdalar neurons 'turn off' GABAergic projections to the PVN. These GABAergic neurons are strategically positioned in hypothalamic nuclei regulating homeostatic function, providing a possible means for polysensorial cognitive inputs to override ongoing homeostatic regulatory information.

Stress-induced HPA axis activity is temporally limited by inhibitory inputs from the hypothalamus and forebrain. Hypothalamic inputs can directly relay inhibitory information from homeostatic effector pathways. In contrast, forebrain outflow from hippocampus and medial prefrontal cortex mediate inhibition by intermediary synapses, many of which involve hypothalamic homeostatic effector circuits. Thus, as was the case for the amygdala, the impact of hippocampal and cortical circuitry is positioned to act through modulating the ongoing activity of homeostatic pathways to the PVN.

Glucocorticoid negative feedback can be exerted at multiple levels of the neuraxis. Glucocorticoid receptors are present on PVN neurons, and these are involved in rapid as well as sustained feedback inhibition of the HPA axis. However, forebrain glucocorticoid receptors also play a role in inhibition of anticipatory HPA axis stress responses, mediated by effects of glucocorticoids on hippocampal and cortical output circuits.

Overall, regulation of the HPA axis appears to involve hierarchical processing of stressful information and glucocorticoid negative feedback. Reflexive responses represent a first line of defense and are directly wired to rapidly induce ACTH release. Anticipatory responses relay through nuclei that modulate reflexive HPA activation, providing a means by which psychological information can modulate or interact with homeostatic integration. Thus, the reflexive stress response is the default and is subject to modification by prior experience or innate programs. The hierarchical organization of brain stress integration is consistent with theories of brain evolution whereby increasingly complex functions are layered atop simpler reflex pathways.

Control of the HPA axis stress response is of substantial health significance. The impact of gross dysregulation of glucocorticoid secretion on metabolism and cardiovascular function is well known. In addition, more-subtle dysregulation of the HPA axis occurs in stress-related affective disease states, such as depression (hypersecretion) and posttraumatic stress disorder (hyposecretion). The HPA axis pathologies are intimately associated with altered activation or volume of the hippocampus, amygdala, and prefrontal cortices, suggesting that the disease process involves the anticipatory stress response. Thus, appropriate initiation and appropriate cessation of the HPA axis stress responses are a clear priority for both homeostatic balance and adaptation to adversity.

See also: Hypothalamic–Pituitary–Adrenal (HPA) Axis; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress.

Further Reading

Boyle MP, Brewer JA, Funatsu M, et al. (2005) Acquired deficit of forebrain glucocorticoid receptor produces depression-like changes in adrenal axis regulation and behavior. *Proceedings of the National Academy of Sciences of the United States of America* 102: 473–478.

Dayas CV, Buller KM, Crane JW, Xu Y, and Day TA (2001) Stressor categorization: Acute physical and psychological stressors elicit distinctive recruitment patterns in the amygdala and in medullary noradrenergic cell groups. *European Journal of Neuroscience* 14: 1143–1152.

De Kloet ER, Vreugdenhil E, Oitzl MS, and Joels M (1998) Brain corticosteroid receptor balance in health and disease. *Endocrine Reviews* 19: 269–301.

Herman JP and Cullinan WE (1997) Neurocircuitry of stress: Central control of the hypothalamo–pituitary–adrenocortical axis. *Trends in Neuroscience* 20: 78–83.

Herman JP, Figueiredo H, Mueller NK, et al. (2003) Central mechanisms of stress integration: Hierarchical circuitry controlling hypothalamo–pituitary–adrenocortical responsiveness. *Frontiers in Neuroendocrinology* 24: 151–180.

Herman JP, Tasker JG, Ziegler DR, and Cullinan WE (2002) Local circuit regulation of paraventricular nucleus stress integration: Glutamate-GABA connections. *Pharmacology, Biochemistry, and Behavior* 71: 457–468.

Keller-Wood M and Dallman MF (1984) Corticosteroid inhibition of ACTH secretion. *Endocrine Reviews* 5: 1–24.

Laugero KD, Bell ME, Bhatnagar S, Soriano L, and Dallman MF (2001) Sucrose ingestion normalizes central expression of corticotropin-releasing-factor messenger ribonucleic acid and energy balance in adrenalectomized rats: A glucocorticoid-metabolic-brain axis? *Endocrinology* 142: 2796–2804.

McEwen BS and Stellar E (1993) Stress and the individual: Mechanisms leading to disease. *Archives of Internal Medicine* 153: 2093–2101.

Munck A, Guyre PM, and Holbrook NJ (1984) Physiological functions of glucocorticoids in stress and their relations to pharmacological actions. *Endocrine Reviews* 5: 25–44.

Rivest S (2001) How circulating cytokines trigger the neural circuits that control the hypothalamic–pituitary–adrenal axis. *Psychoneuroendocrinology* 26: 761–788.

Sapolsky RM (1994) *Why Zebras Don't Get Ulcers: A Guide to Stress, Stress Related Diseases, and Coping*. New York: Freeman.

Sawchenko PE, Li HY, and Ericsson A (2000) Circuits and mechanisms governing hypothalamic responses to stress: A tale of two paradigms. *Progress in Brain Research* 122: 61–78.

Selye H (1936) A syndrome produced by diverse nocuous agents. *Nature* 138: 32.

Selye H (1976) *The Stress of Life*. New York: McGraw-Hill.

Whitnall MH (1993) Regulation of the hypothalamic corticotropin-releasing hormone neurosecretory system. *Progress in Neurobiology* 40: 573–629.

Stress Response: Sex Differences

R J Handa, Colorado State University, Fort Collins, CO, USA
R F McGivern, San Diego State University, San Diego, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The pioneering work of Hans Selye in the 1940s provided the first anatomical evidence suggesting a sex difference in the stress response by showing that the adrenal gland is larger in female animals than males. Subsequent studies show that the difference corresponds with a higher rate of basal- and stress-induced adrenal hormone secretion. Females exhibit a greater rise in adrenal glucocorticoid secretion in response to physical and psychologically stressful situations, as well as a larger diurnal fluctuation in plasma concentrations of glucocorticoids. In rodents, the magnitude of this sex difference is quite remarkable, with females often achieving peak levels of cortisol or corticosterone that are 1.5 to 2 times greater than that of males. In humans, gender-related stress responses are more complex and may depend upon the type of stressor. Current evidence suggests that women are more responsive to stressors that involve social interactions, whereas men are more responsive to those that involve achievement. Summaries of essential sex differences in the stress response of rodents and humans are provided in Tables 1 and 2.

The Hypothalamo-Pituitary-Adrenal Axis

In mammals, the secretion of glucocorticoid hormones from the adrenal cortex is controlled by the activity of the hypothalamo-pituitary-adrenal (HPA) axis. This neuroendocrine axis represents a cascade of neural and humoral signals driven by both the circadian pacemaker as well as the environment. Changing environmental conditions or perceived threats to homeostasis activate the HPA axis through neurons located in the paraventricular nucleus of the hypothalamus (PVN). These neurons integrate positive and negative inputs arising from other brain areas to coordinate the final response of the neuroendocrine brain. Central in this response are neurons located in the parvocellular part of the PVN that contain the neuropeptide, corticotropin-releasing hormone (CRH). The release of CRH to the hypophyseal portal system enhances synthesis and release of adrenocorticotrophic hormone (ACTH) from the anterior pituitary. ACTH acts in turn on the adrenal cortex to cause a rise in plasma corticosteroids. Circulating corticosteroids subsequently act at the level of the pituitary, hypothalamus, and higher brain areas in a negative feedback loop to limit further hormone secretion.

Sex Differences in the Function of the HPA Axis

Although sex differences in the pituitary ACTH and adrenal corticosteroid responses to a stressor have been consistently reported in the literature, the mechanisms underlying these differences are largely unexplored. In general, the neuroendocrine response of females to stress is different from that of males. The female response is characterized by a greater and prolonged secretion of ACTH and adrenal corticosteroids suggesting enhanced stimulus salience as well as a reduced negative feedback. Gonadectomy of both males and females reduces the sex difference suggesting an essential role for circulating sex steroid hormones. Hormone replacement of gonadectomized animals can reinstate the sex difference. Estrogen enhances whereas testosterone treatment inhibits HPA reactivity. The mechanisms by which estrogen and testosterone act to influence HPA function are not completely resolved and will be addressed in some of the following sections of this article. Evidence for estrogen and testosterone acting at the adrenal gland, anterior pituitary, and hypothalamus has been reported.

Pituitary Hormone Secretion

The sex difference in adrenal function is in part, due to sex differences in the neuroendocrine axis that regulates adrenal hormone secretion. Similar to the sex differences in adrenal glucocorticoid release, sex differences exist in the release of ACTH, a hormone synthesized by the anterior pituitary gland that stimulates adrenal corticosteroid synthesis and secretion. Animal studies have demonstrated increases in ACTH following physical or psychological stressors and that this rise is greater and more prolonged in females compared to males. An enhanced ACTH response to CRH administration has been noted in female rats, suggesting that part of the sex difference in ACTH may arise as a result of differences in pituitary responsivity to CRH.

Hypothalamic Function

The secretion of ACTH from the anterior pituitary gland is in turn regulated by hormones which are synthesized by neurons in the paraventricular n. of

Table 1 Summary of selected sex differences in the rodent stress response

Parameter	*Females*	*Males*
ACTH secretion	+	
Corticosterone levels	+	
Response to CRH	+	
# of CRH neurons	+	
Amplitude of diurnal corticosterone rise	+	
Neural vasopressin levels		+
Corticosteroid binding globulin	+	
Vasopressin secretion	+	
Neural GR/MR levels		+
Behavioral responses		
Freezing/defecation		+
Activity	+	
Struggling in FST	+	

ACTH, adrenocorticotropic hormone; CRH, corticotrophin-releasing hormone; GR, glucocorticoid receptor; MR, mineralocorticoid receptor; FST, forced swim test.

Table 2 Summary of selected sex differences in the human stress response

Parameter	*Females*	*Males*
Cardiovascular reactivity to emotional stress		+
ACTH response to exogenous CRH	+	
HPA reactivity to social stress	+	
HPA reactivity to performance stress		+
Clinical consequences related to stress		
Cardiovascular disease		+
Autoimmune disease	+	
Clinical depression	+	
Posttraumatic stress disorder	+	

the hypothalamus and released into the hypothalamo-hypophyseal portal vasculature. These hormones access the anterior pituitary gland to stimulate ACTH synthesis and secretion. The most important releasing factors for ACTH appear to be CRH and vasopressin. Both CRH and vasopressin are potent ACTH secretogogs and vasopressin potentiates the ACTH releasing activity of CRH several fold. Animal studies have shown that females have greater numbers of detectable CRH neurons within the paraventricular n. of the hypothalamus than males. Females also exhibit a more pronounced diurnal variation in hypothalamic CRH immunoreactivity than males. These data are consistent with the physiological data demonstrating a greater activation of the HPA axis in females.

In contrast, sex differences in vasopressin immunoreactivity have not been reported for the paraventricular n. of the hypothalamus; however, sex differences in vasopressin levels have been reported for other brain areas such as the bed nucleus of the stria terminalis. When sex differences in vasopressin are present, the amount of vasopressin is much greater in males than females. This perhaps indicates that sex differences in vasopressin are more relevant to behavior differences rather than hormone secretion. However, studies have demonstrated sex differences in the release of vasopressin to the circulation following a stressor. Whereas male rats do not release vasopressin following immobilization stress, females do. This sex difference in the functional response of vasopressin cells in the paraventricular n. of the hypothalamus to stressors, even in the absence of morphological gender differences in this nucleus may help explain sex differences in neuroendocrine responses.

Negative Feedback Regulation

The stimulatory limb of the HPA axis (CRH, vasopressin, and ACTH) is kept under continuous check by feedback inhibition. Thus, it is thought that one factor underlying sex differences in adrenal hormone secretion are gender differences in the intensity of feedback regulation. The negative feedback action of corticosteroids is mediated by two receptor types designated glucocorticoid receptor (GR) and mineralocorticoid receptor (MR) that reside in neurons in the hypothalamus, hippocampus, and anterior pituitary gland, as well as other brain areas. Both receptors appear to be involved in differing aspects of feedback regulation.

Mineralocorticoid receptors, because of their high affinity for corticosteroids, are occupied during, and consequently regulate, basal hormone secretion. GRs, in contrast, exhibit a tenfold lower affinity for corticosteroids and are occupied following elevations in hormone. These receptors thereby function to return stress responsive elevations in corticosteroids back to baseline. Assuming that receptor number corresponds to the sensitivity of a tissue for hormone, a conclusion could be drawn that sex differences in the HPA axis may arise as a result of sex differences in the number of receptors in a given cell. More receptors would indicate greater negative feedback sensitivity and hence, lower basal- and stress-activated hormone secretion. Lower levels of receptor would indicate weaker negative feedback and thus, higher basal- and stress-responsive hormone secretion. Evidence for sex differences in the concentration of GR and MR and the function of GR and MR has been demonstrated with females tending to have lower levels of GR and MR in many brain regions including the hypothalamus and hippocampus. This pattern is consistent with that found in other tissues such as the thymus and liver. Given similar levels of hormone, GR- or MR-mediated functions (such as autoregulation

of the receptor) have been reported to be less sensitive to corticosteroid modulation in females compared to males.

Estrogen Modulation of HPA Axis Function

Previous studies have demonstrated that the estrogen status of females affect their endocrine response to stress. In rats, there is a relationship between the day of the estrous cycle and their plasma corticosteroid levels. Basal- and stress-responsive corticosteroid levels are higher on proestrus when estrogen levels are correspondingly high. Hormone replacement studies show that estrogen treatment of gonadectomized female and male rats enhances the basal- and stress-responsive secretion of ACTH and corticosteroids. These findings have been confirmed for humans in preclinical trials as well. Estrogen treatment of women and men enhances heart-rate and blood pressure increases as well as the ACTH and cortisol response to psychosocial or mental stressors.

The actual mechanism(s) by which estrogen alters stress-responsive ACTH and corticosterone (CORT) secretion is not known. At the hypothalamic level, increases in stress hormones could be a consequence of increased stimulation of PVN neurons, or reductions in inhibitory tone. Such possibilities are suggested by data showing increased expression of stress-inducible c-fos mRNA following E-treatment. The prolonged stress-responsive secretion of ACTH and corticosteroids, coupled with the elevated baseline secretion of ACTH and corticosteroids in female rats following estrogen treatment, also suggests the possibility of a disruption in HPA negative feedback mechanisms. Consistent with this, the suppression of CORT secretion by dexamethasone is impaired by E-treatment of male and female rats.

Progesterone Modulation of HPA Axis Function

Females have higher progesterone (P) levels than males and these levels increase following ovulation. P inhibits HPA function, perhaps to counteract the enhancing effects of estrogen at certain times of the female cycle. Progesterone prevents E-induced increases in CRH mRNA in the PVN of female monkeys. Progesterone also potentiates the inhibitory effects of CORT on ACTH release in ewes, and prevents the adrenalectomy (ADX)-induced upregulation of AVP mRNA in rats. Whether the actions of P are mediated by binding inhibitory neurotransmitter receptors, such as γ-aminobutyric acid (GABA), remains to be determined. However, since gonadectomy of male and female rats can reduce sex differences in HPA function, and P is higher in intact females and inhibits HPA function, this apparently rules out a role for P in establishing sex differences in HPA function.

Androgen Modulation of HPA Axis Function

Testosterone may contribute to the sex differences in HPA reactivity since testosterone has an action opposite that of estrogen in regulating the HPA axis. Gonadectomy of adult male rats increases CORT and ACTH responses to stress, and correspondingly, the expression of the immediate early gene, c-fos, in PVN neurons. Hormone replacement of gonadectomized male rats with testosterone or the non-aromatizable androgen, dihydrotestosterone, returns stress-responsive CORT and ACTH back to that of the intact male. Treatment of GDX's animals with dihydroxytryptamine (DHT) also inhibits the stress induction of c-fos mRNA in the PVN demonstrating that T's effects are not due to aromatization of testosterone to estrogen.

Although androgen can inhibit HPA axis function and reduce CRH immunoreactivity (ir) in the PVN, androgen receptors are not localized in CRH or AVP neurons within the PVN. Low levels of androgen receptors are found in the PVN, but these neurons are located in the parts of the PVN that project to spinal cord and brain stem autonomic nuclei. Consequently, it is assumed that androgen regulates PVN neuropeptide expression and secretion transsynaptically. Recent studies using stereotaxic surgery to locally apply DHT to the PVN of males demonstrate that DHT can affect PVN neurons, an effect that does not appear to be mediated by the androgen receptor, but rather through estrogen receptor-β. Local application of DHT to the PVN is as effective as peripherally administered DHT in inhibiting HPA function. Local application of 5α-androstane-3β, 17β-diol (3β-diol), a metabolite of DHT, was also as effective as DHT in inhibiting HPA function, as is local application of ERβ selective agonists. Counter-intuitively, the effects of DHT and 3β-diol can be blocked by simultaneous administration of an ER antagonist, but not an AR antagonist. Since 3β-diol can bind ERβ with moderate affinity, this seems to indicate that DHT is metabolized to ligands, such as 3β-diol, that preferentially bind ERβ, and that ERβ occupancy works to inhibit HPA function in the male.

Sex Differences in Behavioral Responses to Stress

Animal studies have provided much of what we know regarding sex differences in response to stress.

The acute response to stress is often studied in rats using paradigms, such as the elevated plus maze and the novel open-feild, that place them in a situation where their natural tendency to explore is opposed by their aversion to open, well-lighted spaces. In such circumstances, rats exhibit a series of well-described behaviors associated with the degree of fear or anxiety that is being experienced. Freezing, defecation, activity, and vocalization are a few good examples of these behaviors. Compared to males, females exhibit a smaller behavioral response to novel or aversive environments, showing less freezing and defecation, and more activity. Females are also more apt to return to explore a compartment in which they have been previously shocked.

Under some conditions, females show more adaptive responses to long-term stress exposure as well. Animal models of anxiety often utilize 'learned helplessness' paradigms to measure behavioral responses to acute and chronic stress. Such paradigms are useful to study the efficacy of antidepressant treatments in reversing the effects of stress. One of these paradigms is the Porsolt test, a widely employed forced-swim paradigm that has proved to be an effective assessment tool to measure the efficacy of antidepressant actions. The test involves placing an animal in a circular, inescapable chamber of warm water and measuring the latency to stop swimming, and length of time spent in floating immobility. Studies consistently show that females swim (struggle to escape) longer than males in this test. This difference occurs with acute stress exposure, as well as chronic exposure for several weeks.

Opposing effects have been reported when using 'learned helplessness' paradigms that employ repeated, intermittent, and/or inescapable shock. In such paradigms, the shock session is subsequently followed by a session in which animals are exposed to the same environment and given an opportunity to escape. When the paradigm involves exposing animals to one or two sessions of inescapable shock before allowing them access to escape, females have faster latencies to escape than males. However, when exposed to inescapable shock over many sessions before being allowed access to an escape chamber, some studies have found that males escape faster than females. It appears that this sex difference in adaptability may be dependent on levels of corticosterone, since reducing the poststress level of corticosterone in the female to the level attained in the male allows the adaptation to repeated restraint.

The role of the HPA axis in accounting for behavioral sex differences in response to stress is complex, and the use of animal models has yielded an incomplete picture to date. However, sex differences in stress-related behaviors in rodents may involve steroid regulation of serotonin. Serotonin synthesis and receptor expression are regulated by both glucocorticoids and gonadal steroids. Compared to males, females are reported to have higher levels of 5-HT in brain as well as increased rates of synthesis. These data are consistent with those showing a more rapid response in females to antidepressant actions in the Porsolt test.

Gender Differences in the Human Response to Stress

Like experimental animals, humans also exhibit sex differences in their stress response, including physiological, endocrine, and behavioral reactions. From a perspective of general health, the gender difference in cardiovascular reactivity is one of the most important. The resting heart rate in women is higher than in men, which is similar to male and female laboratory animals. However, in spite of their higher baseline, women exhibit less cardiovascular reactivity to stressful emotional situations than men, as reflected by smaller changes from baseline for both heart rate and blood pressure. The greater cardiovascular lability to emotional stressors in men is considered a major contributing factor to their increased risk of cardiovascular morbidity.

Gender differences in the cardiovascular response to physical stress are not as pronounced. Studies involving moderate exercise show that healthy young men and women exhibit no significant differences in exercise-induced increases in heart rate. However, increases in systolic blood pressure under these conditions are often reported to be greater in men than women.

Although the animal literature shows large and consistent sex differences in HPA activation, gender-related data in humans are much more equivocal. In response to physiological stress such as exercise, young men and women show no reliable differences in HPA activation. Other studies have reported an increased release of ACTH and prolonged cortisol response in women compared with men following CRH administration, which might be expected from the animal literature. However, this pattern may reflect the direct, physiological stimulation of pituitary corticotrophs, in contrast to an endogenous response following a psychological or physical stressor. As discussed below, gender differences in the endocrine stress response to psychological and physiological events are situational, and thus not likely to be a simple reflection of differences in CRH sensitivity at the level of the pituitary.

Early studies of the endocrine responses of men and women produced conflicting results, with some reporting endocrine patterns to stress that paralleled results in laboratory animals, and other reporting no gender differences. The work was complicated by the methodological problem of the need to use invasive procedures to obtain blood samples. In the past decade, stress reactivity studies of the HPA have been greatly improved by the advent of the salivary cortisol assay. This allows a stress-free, noninvasive method of collecting multiple cortisol samples from an individual over time in response to an experimental situation. Moreover, the assay measures the free cortisol, which is the biologically active fraction. This avoids the interpretive problems associated with measure total cortisol in plasma, of which approximately 90–95% is attached to cortisol binding globulin (CBG) and biologically unavailable.

The results from more recent studies indicate that gender differences in human HPA responsiveness to psychological stress depend upon the type of stress and the environmental circumstances in which it occurs. Studies of HPA activation during psychological stress that involves achievement situations show that ACTH and cortisol levels are higher in men than women, although women report greater subjective feelings of stress. This achievement-related pattern of increased HPA activation in men has been observed in a range of experimental and real life situations, including public speaking, mirror drawing, problem solving, and marital conflict. However, when the stressor involves social rejection, females have been found to exhibit higher cortisol levels than males. This is consistent with other studies showing that psychological factors associated with social stress, such as taking care of an elderly parent, have more impact on HPA reactivity in women than in men. Some investigators have suggested that gender differences in stress reactivity along social/achievement dimensions may reflect the evolutionary pressures of sex roles. Studies examining stress in the workplace tend to find few gender differences in HPA reactivity when the situation does not involve conflict with co-workers.

Compared to laboratory animals, much less is known about the influence of gonadal steroids on stress-related activation of the HPA in humans. A stimulatory role for estrogen on HPA activation is suggested by data from studies conducted during the menstrual cycle. During the follicular phase, when estrogen levels rise and peak at ovulation, stress-induced rises in cortisol are generally greater than those observed in the luteal phase. In the luteal phase, rising progesterone levels may counteract the influence of estrogen on HPA activation. Removal of the ovaries prior to menopause reduces the HPA sensitivity to CRH stimulation. However, estrogen's effect on HPA reactivity is not restricted to women, since normal men administered a short-term regimen of estrogen also exhibit increased HPA activity. One possible explanation for estrogen's effect on HPA reactivity comes from observations that estrogen therapy will increase circulating levels of CBG. Increases in CBG would result in a decrease in free, bioavailable cortisol, potentially reducing negative feedback regulation.

The modulatory role of gonadal hormones on human HPA reactivity has been examined in volunteers undergoing a reversible, chemical castration procedure. Men and women volunteers treated for 2 months with leuprolide, a synthetic analog of gonadotropin releasing hormone, which suppressed the secretion of gonadal steroids. This treatment permitted the examination of the HPA response to stress under conditions where estrogen and testosterone levels were low in both sexes. The results showed that men sex exhibited a significantly greater plasma ACTH response to CRH stimulation than women and a greater cortisol response to exercise. However, no gender differences were observed in HPA responsiveness using the ratio of ACTH to cortisol. With subsequent testosterone replacement in the men, CRH stimulation of ACTH was significantly increased but cortisol significantly decreased. These findings suggest that there may be an underlying organizational effect in brain as the source of gender differences in HPA reactivity.

Gender-Related Clinical Consequences of Stress

The mechanisms linking stress to disease states in humans are not well understood, but hormonal changes, induced by the stress response, are recognized as a major influence. Stress is not viewed as the cause of disease, but rather as a factor increasing an individual's susceptibility to it. Stress-induced elevations of plasma glucocorticoids have direct effects on behavior and physiology which can be modulated by gonadal steroids. It is the interaction between glucocorticoid and gonadal harmones on cellular responses that is believed to be the major factor in gender differences in stress-related disease. Receptors for glucocorticoids, androgens, and estrogens are widely distributed tissues that are stress-responsive, including brain, heart muscle, blood vessels, and immune tissues such as spleen, thymus, and bone marrow. In brain these receptors are located in regions that mediate emotional and autonomic responses to stress such as the hippocampus, hypothalamus, amygdala, and

prefontal cortex. It is the interaction between glucocorticoid and gonadal hormones on cellular responses that is believed to be the major factor in gender differences in stress-related disease.

Males have a significantly higher risk for having cardiovascular disease at an earlier age than women, and this risk is exacerbated by stress. Women are afforded a significant degree of protection through the actions of estrogen against arteriosclerosis. The gender difference subsides after menopause, with the result that risk of heart disease increases in older women.

In contrast, women are significantly more likely to suffer from autoimmune disorders such as lupus, rheumatoid arthritis, and irritable bowel syndrome. An interaction between estrogen and HPA activation has been proposed as a mediator of the sex-linked ratio of autoimmune disorders, which are approximately threefold more common in women. In contrast to the protective role estrogen plays in cardiovascular disease, estrogen may exacerbate immune responses through receptors present on thymocytes, macrophages, and endothelial cells, all of which also express GRs. The interaction between estrogen and glucocorticoids on immunomodulation appears to underlie the gender difference in autoimmune disease susceptibility.

Women are 2–3 times more likely to suffer from depression than men. The gender ratio is consistent across cultures in western society, indicating that biological factors, as opposed to local variations in psychosocial attitudes, are a primary contributor. Clinical studies consistently show that the HPA system is tonically hyperactive in endogenously depressed patients. However, no clear differences between men and women have emerged in HPA activity that might account for the gender difference in depression. The onset of clinical depression is well known to be exacerbated by life stress and data from several studies suggest that the gender bias in mood disorders involves an increased susceptibility in some women to the combined effects of stress and hormonal status. Thus, fluctuations in estrogen and progesterone across the menstrual cycle have been proposed to confer susceptibility to depression in women and this can be exacerbated by life stress. Finally, serotonin synthesis in brain is significantly greater in men than women, a factor that has been implicated in the increased risk for depression in women. This pattern of synthesis is opposite that reported in animals.

The odds ratio for women developing posttraumatic stress disorder (PTSD) is similar to depression. Following exposure to prolonged psychological trauma, women exhibit PTSD symptoms twice as frequently as men, even when the type of trauma is equivalent. Some studies have found that experiencing psychologically traumatizing events prior to puberty greatly increases the susceptibility to PTSD in adulthood in females compared to males. This suggests that organizational effects of hormones may play a significant role in the increased susceptibility of women. Hippocampal damage, which can be induced by chronically high levels of glucocorticoids, may play a role in PTSD based on studies showing a reduction in volume in PTSD patients. Consistent with this, the hippocampus has been shown to play a critical role in neural circuitry assessing threat. In animals, uncontrollable stress has been found to reduce spine density, a measure of synapse numbers, in the hippocampus of females, but increase it in males.

In summary, studies over the years have identified consistent sex differences in the hormonal and behavioral responses to stress in laboratory animals and humans. In general, the reactivity of the HPA axis to a stressor is greater in females, whereas behavioral stress responses are often more severe in males. Such data indicate that coping strategies are different between the sexes. Underlying these sex differences are both organizational and activational actions of gonadal steroid hormones on brain areas controlling neuroendocrine function and fear- or anxiety-related behaviors. While there is broad agreement between animal and human studies, human sex differences in stress reactivity are move dependent upon the type of stress and the response measured.

See also: Hypothalamic–Pituitary–Adrenal (HPA) Axis; Social Stress in Adult Primates; Stress and Neural Involvement in Metabolism; Stress and Vulnerability to Brain Damage; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress and Neuronal Plasticity; Stress, Sex and Adolescent Nicotine Response; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress.

Further Reading

Harris HA (2006) Estrogen receptor-β: Recent lessons from *in vivo* studies. *Molecular Endocrinology* 21: 1–13.

Kajantie E and Phillips DI (2006) The effects of sex and hormonal status on the physiological response to acute psychosocial stress. *Psychoneuroendocrinology* 31(2): 151–178.

Klein LC and Corwin EJ (2002) Seeing the unexpected: How sex differences in stress responses may provide a new perspective on the manifestation of psychiatric disorders. *Current Psychiatry Reports* 4: 441–448.

Kudielka BM and Kirschbaum C (2000) Sex differences in HPA axis responses to stress: A review. *Biological Psychology* 69: 113–132.

Kudielka BM and Kirschbaum C (2005) Sex differences in HPA axis responses to stress: A review. *Biological Psychology* 69: 113–132.

Lundberg U (2005) Stress hormones in health and illness: The roles of work and gender. *Psychoneuroendocrinology* 30: 1017–1021.

Nemeroff CB, Bremner JD, Foa EB, Mayberg HS, North CS, and Stein MB (2006) Posttraumatic stress disorder: A state-of-the-science review. *Journal of Psychiatric Research* 40: 1–21.

Roca CA, Schmidt PJ, Deuster PA, et al. (2005) Sex-related differences in stimulated hypothalamic–pituitary–adrenal axis during induced gonadal suppression. *Journal of Clinical Endocrinology Metabolism* 90: 4224–4231.

Rubinow DR, Schmidt PJ, and Roca CA (1998) Estrogen–serotonin interactions: Implications for affective regulation. *Biology Psychiatry* 44: 839–850.

Taylor SE, Klein LC, Lewis BP, Gruenewald TL, Gurung RA, and Updegraff JA (2000) Biobehavioral responses to stress in females: Tend-and-befriend, not fight-or-flight. *Psychological Review* 107: 411–429.

Williamson M, Bingham B, and Viau V (2005) Central organization of androgen-sensitive pathways to the hypothalamic–pituitary–adrenal axis: Implications for individual differences in responses to homeostatic threat and predisposition to disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 29(8): 1239–1248.

Stress, Cytokines and Depressive Illness

M K Cowles and A H Miller, Emory University School of Medicine, Atlanta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Stress has long been accepted as a trigger for the onset of a number of diseases, including major depression. Nevertheless, the mechanisms by which stress translates into disease have yet to be fully elucidated. One possibility is the effect of stress on the immune system. Although much attention has been paid to the immunosuppressive effects of stress, recent data indicate that stress can activate certain aspects of immunity, specifically the innate inflammatory immune response. While inflammatory processes are well known to contribute to the development and progression of allergic disorders, autoimmune diseases, and infections, there has been increasing interest in the role of inflammatory responses in a number of other common medical conditions, including cardiovascular disease, diabetes, and cancer. New data suggest that inflammation may play a role in the pathophysiology of depression as well. Therefore, stress effects on the immune response, specifically the innate immune inflammatory response, may be an important and relevant pathway in the contribution of stress to disease, including psychiatric diseases such as depression.

Innate Immunity and the Inflammatory Response

The immune response can be divided into two broad categories – innate immunity (described later) and acquired immunity (comprising T and B lymphocytes that, respectively, mediate the cellular and humoral (antibody) responses to pathogen exposure). Both aspects of immunity are carried out by a series of cells and soluble factors, which work together to coordinate the response to foreign pathogens, neoplastically transformed cells, and cells and tissues that have been damaged or destroyed. All immune cells originate as hematopoietic stem cells that ultimately differentiate along either myeloid or lymphoid cell lines. Immune cell differentiation and function are guided by a number of factors, including cytokines. Cytokines are intercellular signaling molecules produced by a variety of cells, including activated neutrophils, monocytes/macrophages, lymphocytes, natural killer (NK) cells, glia (astrocytes and microglia), endothelial cells, fibroblasts, and adipocytes. Cytokine effects are both local and systemic, and for a particular cytokine there may be multiple cellular targets, resulting in a myriad of relevant effector functions. There are three general categories of cytokines: (1) mediators and regulators of innate immunity, (2) mediators and regulators of acquired immunity, and (3) stimulators of hematopoiesis. Table 1 summarizes the roles of these various cytokines.

Initial work with respect to the impact of stress and depression on immune function focused on inhibition of the acquired immune response. More recent evidence, however, supports the notion that stress and depression are also associated with activation of innate immune responses resulting in inflammation, which in turn, may contribute to the development of a number of medical disorders (see later) as well as alterations in behavior. The cellular components of the innate immune system include monocytes (which further differentiate into macrophages), neutrophils, and NK cells. Monocytes and neutrophils are derived from myeloid stem line whereas NK cells are derived from the lymphoid stem line. Monocytes, macrophages, and neutrophils are collectively referred to as phagocytes. Circulating mediators of innate immunity include complement, acute-phase proteins such as C-reactive protein (CRP), and cytokines including type I interferons (IFNs)-α and-β, tumor necrosis factor-α (TNF-α), IL-1, and IL-6. TNF-α, IL-1, and IL-6 are also referred to as proinflammatory cytokines.

Inflammation is characterized by a cellular response whereby the cytokines of the innate immune response promote the expression and release of chemokines and adhesion molecules in the local environment (Figure 1). Chemokines and adhesion molecules in turn support the collection of various innate immune cell subtypes at the site of pathogen invasion and/or tissue disruption/destruction; these subtypes include neutrophils and macrophages as well as elements of the acquired immune response such as activated T-helper and T-cytotoxic cells. In addition to the local response, cytokines can enter the peripheral circulation and stimulate the production of acute-phase proteins by the liver and interact with nervous system elements, leading to behavioral changes referred to as 'sickness behavior,' including depression, fatigue, and cognitive dysfunction (see later). The acute-phase response is geared to promote elimination of invading pathogens while at the same time inactivating potentially destructive enzymes (proteases) and the generation of oxygen radicals. Cytokine effects on the central nervous system may ultimately help to conserve energy in the service of

fever and other metabolic processes that support endogenous antibiosis and wound healing.

At the signal transduction level, activation of the transcription factor, nuclear factor-kappa B (NF-κB), plays a fundamental role in initiating the innate immune cascade, with the resultant production of multiple inflammatory mediators, including cytokines, chemokines, cellular adhesion molecules, acute-phase reactants, prostaglandins, nitric oxide, and the cyclooxygenase COX-2, as well as activating other stress responsive genes. NF-κB is activated by the interaction of pathogens with toll-like receptors on the surface of immune cells, notably macrophages. Toll-like receptors are pattern recognition molecules on innate immune cells that serve to recognize a few highly conserved structures present in many different microorganisms. Cytokine receptors (e.g., IL-1) containing the toll-like receptor binding domain also activate the NF-κB pathway. In addition to NF-κB, the mitogen-activated protein kinase (MAPK) cascade – including p38, extracellular signal-regulated kinase-1/2 (ERK-1/2), and c-Jun N-terminal kinase (JNK) – as well as the Janus kinase (JAK)-signal transducers and activators of transcription (STAT) signaling pathway also play an important role in initiating and maintaining the inflammatory response.

Although the inflammatory response is pivotal in providing a front-line defense against pathogen

Table 1 Roles of cytokines in the immune system

Modulation of differentiation of hematopoietic stem cells
Activation of immune cell growth and proliferation
Stimulation of antibody production and cell-mediated immunity
Inhibition of pathogen replication
Activation of the acute-phase response and inflammation
Stimulation of chemotaxis
Induction of cell death

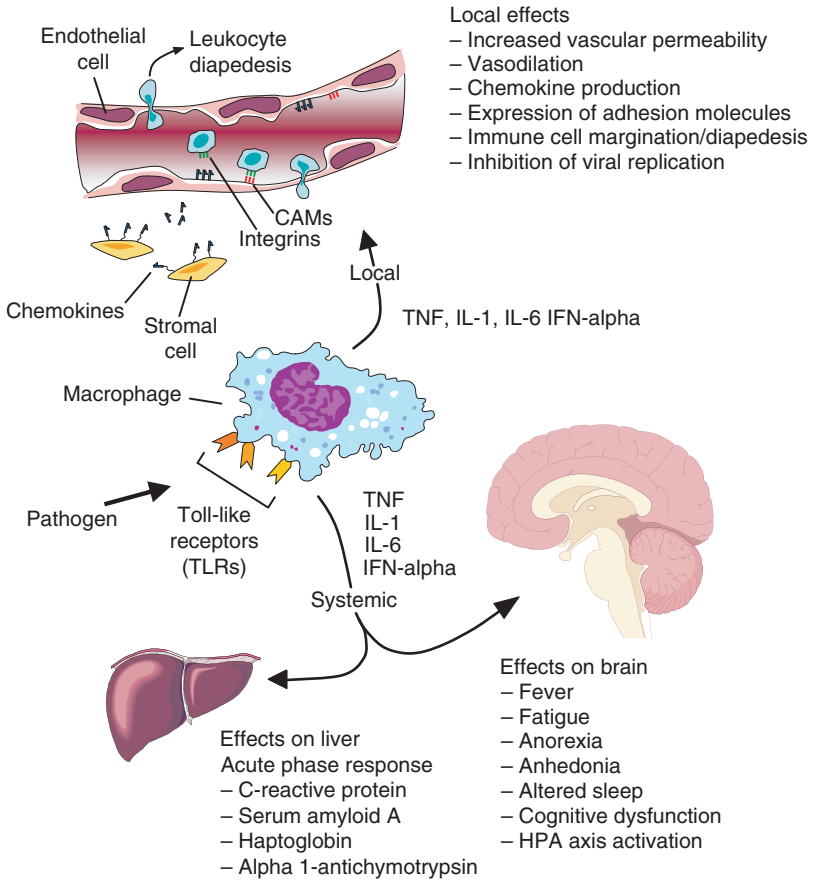

Figure 1 Innate immunity: local and systemic responses to cytokine release secondary to tissue injury/infection. Locally, cytokines act on endothelial and tissue stromal cells. Tissue stromal cells produce chemotactic factors, recruiting other immune cells to the site of injury. The endothelial cells produce adhesion molecules, enhancing immune cell margination and diapedesis. In the brain, proinflammatory cytokines – including interleukins-1 and -6 (IL-1, IL-6) and tumor necrosis factor-α (TNF) – activate the hypothalamic–pituitary–adrenal (HPA) axis and induce behavioral changes that subserve the metabolic demands of fever and inflammation. The proinflammatory cytokines also induce the liver to produce acute-phase proteins. CAM, cellular adhesion molecule; IFN, interferon. Reprinted with permission from *The American Psychiatric Publishing Textbook of Neuropsychiatry and Clinical Neurosciences*, copyright (2002). American Psychiatric Publishing, Inc.

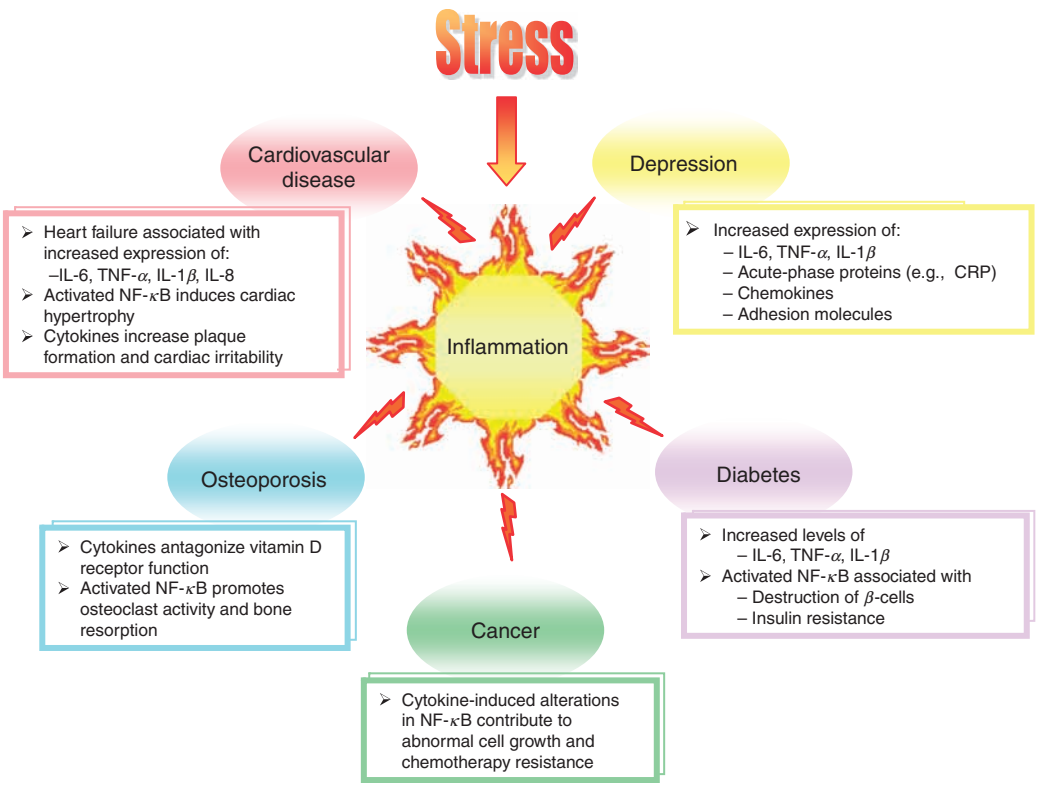

Figure 2 Potential effects of stress on inflammation and disease. IL, interleukin; TNF, tumor necrosis factor; NF-κB, nuclear factor-kappa, B; CRP, C-reactive protein.

invasion and tissue destruction, there is increasing appreciation that when the response is excessive or chronic, inflammation may have significant deleterious effects (Figure 2). Indeed, cardiovascular disease, including atherosclerotic heart disease and congestive heart failure, has been associated with elevated levels of CRP, TNF-α, IL-6, and IL-1β. These cytokines are believed to contribute to atherosclerotic plaque formation and cardiac irritability. Patients with diabetes have also been found to exhibit elevations of CRP, IL-6, and TNF-α, and inflammatory cytokines have been found to disrupt insulin receptor signaling and are associated with insulin resistance. Interestingly in this regard, genetically manipulated mice vulnerable to obesity and insulin resistance via overexpression of the gene for leptin exhibit normalization of glucose metabolic responses following disruption of inflammatory signaling pathways. Finally, inflammatory signaling pathways have been implicated in the neoplastic process, contributing to tumor cell growth and proliferation as well as to resistance to chemotherapeutic agents. Given the capacity of stress to activate inflammatory responses, there has been increasing interest as to whether stress-induced changes in the immune system may play a role in a number of diseases, including not only medical disorders (as noted previously) but also psychiatric illnesses such as depression (see later).

Stress and the Immune System

A wide variety of stressors have been shown to alter immune responses. In general, initial studies have indicated that exposure to acute and chronic stressors is associated with inhibition of acquired immune responses, especially the ability of T and B cells to proliferate *in vitro*. In the case of NK cells, chronic stress was shown to decrease NK cell numbers and activity, while acute stress was associated with increases in NK cell numbers. Although early studies on the effects of stress on immunity emphasized the immunosuppressive aspects of this relationship, more recently it has become apparent that acute and chronic stressors can also activate certain aspects of immunity, specifically the innate immune response. These findings have been apparent in both laboratory animals and humans and include the induction of proinflammatory cytokines, acute-phase reactants, chemokines, and adhesion molecules. In a recent study in humans, it was also shown that exposure to an acute psychosocial stressor (public speaking and mental arithmetic) can lead to the activation of

inflammatory signaling pathways. Indeed, within 10 min following stressor exposure, peripheral blood mononuclear cells from stressed individuals exhibited significantly increased NF-κB DNA binding as measured by electromobility shift assay. In attempts to reconcile the inhibitory effects of stress on acquired immune responses versus stress-induced activation of innate immune responses, the relationship between acquired and innate immune responses has been examined in rats immunized with keyhole limpet hemocyanin (KLH), which induces an antibody (B cell) response (acquired immunity). Interestingly, stress-induced decreases in antibody responses to KLH were reversed in animals pretreated with the IL-1 receptor antagonist. These findings suggest that activation of innate immune responses by stress may in turn contribute to stress-induced inhibition of acquired immune responses.

Regarding the mechanism of the effects of stress on the innate immune response, much attention has been paid to the sympathetic nervous system (**Figure 3**). Both alpha and beta adrenergic antagonists have been used to demonstrate that stress-induced catecholamines acting through alpha 1- and beta 2-adrenergic receptors are capable of inducing an inflammatory response, including the induction of proinflammatory cytokines in both the periphery and the brain. Of note, while alpha-adrenergic antagonists were effective in abrogating peripheral stress-induced proinflammatory cytokines, only beta-adrenergic receptor antagonists were effective in the brain.

Given the relative balance between sympathetic and parasympathetic nervous system responses at rest and during stress, it is interesting to note that activation of efferent parasympathetic nervous system pathways has been found to reduce inflammatory responses. This cholinergic anti-inflammatory pathway is mediated in the periphery by nicotinic acetylcholine receptors that contain the alpha 7 subunit and by muscarinic acetylcholine receptors in the brain. Indeed, both electrical stimulation of the vagus nerve (and thus parasympathetic outflow pathways) and peripheral administration of alpha 7 nicotinic acetylcholine receptor agonists have been shown to significantly reduce mortality and/or blood concentrations of TNF-α in rats exposed to high doses of endotoxin (which produces a massive inflammatory response). Similar effects have been found following central administration of muscarinic receptor agonists. Of note, vagal nerve stimulation, as currently employed to treat depression, stimulates afferent and not efferent parasympathetic pathways and therefore is not specifically designed to influence peripheral immune pathways. Finally, reduced heart rate variability, an indicator of relatively increased sympathetic versus parasympathetic tone, has been associated with evidence of an increased inflammatory response. Taken together, the data support the notion of a 'yin-yang' influence of autonomic nervous system pathways on the regulation of innate immune responses. The findings suggest that the effects of stress on immunity may in part be mediated by an impact on the balance of sympathetic and parasympathetic nervous system activity.

Depression and the Immune System

As with stress, initial work with respect to the immune system in depression revealed diminished functioning of the acquired immune response, primarily manifested by reduced T and B cell proliferative responses and reduced NK cell activity. More recent evidence, however, has supported the notion that in addition to alterations in acquired immune responses and NK cell activity, depression is associated with a hyperactivity of certain aspects of the innate immune response. Specifically, depressed patients have been found to exhibit increased proinflammatory cytokines in both peripheral blood and cerebrospinal fluid, increased acute-phase reactants, and increased expression of chemokines and adhesion molecules. Probably the most reproducible innate immune finding in patients with depression is increased levels of IL-6 and its downstream product from the liver, CRP. Not only have mean differences in inflammatory mediators been found in depressed patients versus control subjects, a number of studies have reported a dose-response relationship between severity of depression and commensurate increases in the concentrations of various inflammatory mediators. Interestingly, compared to other populations of depressed subjects, patients with treatment-resistant depression have been shown to be especially likely to exhibit elevations in inflammatory markers, including IL-6. In addition, IL-6 and TNF-α have been shown to be predictors of nonresponse to standard antidepressant medications. Finally, preliminary studies have indicated that polymorphisms in relevant cytokine genes may predict treatment response in populations of depressed patients.

In addition to evidence of increased inflammatory markers in depressed individuals, a large literature has demonstrated that cytokines of the innate immune response have the capacity to induce a depressive-like state. Indeed, peripherally or centrally administered cytokines induce a syndrome of 'sickness behavior,' which includes symptoms of anhedonia, anorexia, impaired sleep, fatigue, psychomotor slowing, and increased sensitivity to pain, all of which overlap with symptoms experienced by patients with major

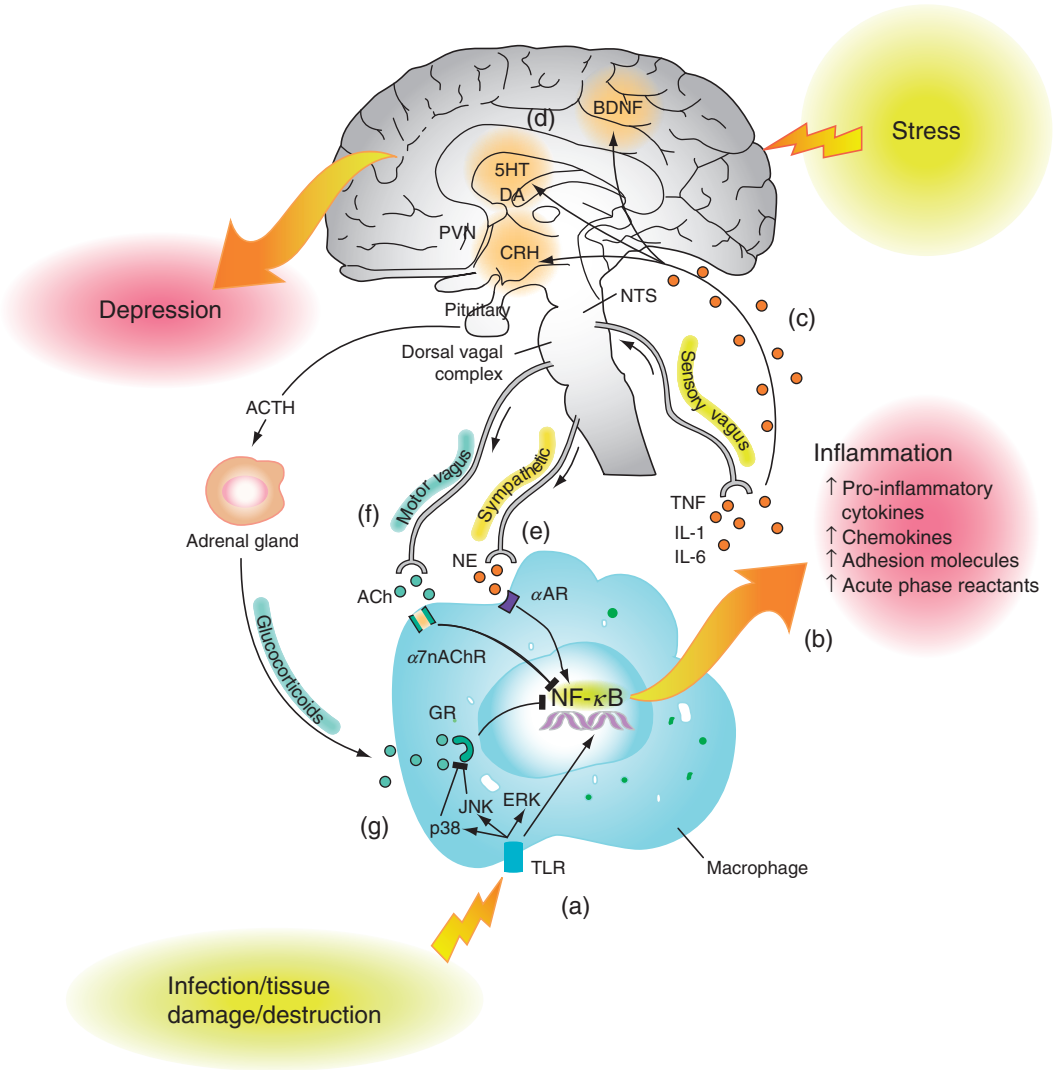

Figure 3 Stress–immune interactions and depression. (a) Activation of nuclear factor-kappaB (NF-κB) through Toll-like receptors (TLR) during immune challenge leads to an inflammatory response (b), including the release of the proinflammatory cytokines tumor necrosis factor and interleukins (TNF-α, IL-1, and IL-6). (c) These cytokines, in turn, access the brain via leaky regions in the blood–brain barrier, active transport molecules, and afferent nerve fibers (e.g., sensory vagus), which relay information through the nucleus tractus solitarius (NTS). (d) Once in the brain, cytokine signals participate in pathways known to be involved in the development of depression, including altered metabolism of relevant neurotransmitters such as serotonin (5HT) and dopamine (DA), activation of corticotropin-releasing hormone (CRH) in the paraventricular nucleus (PVN) and the subsequent production and/or release of adrenocorticotropin (ACTH) and glucocorticoids (cortisol), and disruption of synaptic plasticity through alterations in relevant growth factors, such as brain-derived neurotropic factor (BDNF). (e) Exposure to environmental stressors promotes activation of inflammatory signaling (NF-κB) through increased outflow of proinflammatory sympathetic nervous system responses (release of norepinephrine (NE, orange), which binds to α and β adrenoceptors (αAR, βAR)). (f) Stressors also induce withdrawal of inhibitory motor vagal input responses (release of acetylcholine (ACh, blue), which binds to the α7 subunit of the nicotinic acetylcholine receptor (α7nAChR)). (g) Activation of the MAPK pathways, including p38 and Jun N-terminal kinase (JNK), inhibits the function of glucocorticoid receptors (GR), thereby releasing NF-κB from negative regulation by glucocorticoids (blue) released as a result of the HPA axis in response to stress; ERK, extracellular signal-regulated kinase. Reproduced from Raison CL, Capuron L, and Miller AH (2006) Cytokines sing the blues: Inflammation and the pathogenesis of depression. *Trends in Immunology* 227: 24–31, with permission from Elsevier.

depression. Although there are a vast number of studies in laboratory animals demonstrating that the administration of proinflammatory cytokines or their inducers (e.g., lipopolysaccharide) can reliably contribute to the development of sickness behavior, much of the literature in humans derives from patients being treated with the cytokine, IFN-α. IFN-α is an early innate cytokine released during viral infection and has both antiviral and antiproliferative properties. As such, IFN-α is used to treat viral infections and cancer. Despite being a useful therapeutic agent, IFN-α is notorious for causing

Table 2 Depression in the medically ill

Illness	*Prevalence of depression (%)*
Cardiovascular disease	16–27
Cancer	25–42
Diabetes mellitus	9–27
Cerebrovascular accident	20–27
Human immunodeficiency virus	6–12

significant depressive symptomatology in 30–50% of patients, depending on the dose. Furthermore, it has been shown that the depression that develops during IFN-α therapy is responsive to treatment with standard antidepressants, suggesting that this cytokine-induced depressive syndrome is analogous to depressive disorders found in medically healthy individuals.

Further supporting the relationship between cytokines and depression is the high rate of depressive disorders in patients with a wide range of medical disorders, the majority of which are associated with some element of tissue damage or destruction and resultant inflammation. As shown in Table 2, rates of depression in the medically ill are up to five times higher than that observed in the general population (5%). These rates vary depending on the type, severity, and number of comorbid medical illnesses. With regard to proinflammatory cytokines and depression in medically ill subjects, it is particularly interesting that recent placebo-controlled data demonstrated that administration of an antagonist to TNF-α was associated with a significant reduction in depressive symptoms in patients with psoriasis. Of note, the reduction of depressive symptoms in patients receiving the cytokine antagonist appeared to be independent of the effect of the antagonist on disease activity.

Regarding the mechanisms by which cytokines influence behavior, attention has been focused on how cytokine signals access the brain, what happens once they reach the brain, and what pathways they interact with to contribute to the development of depression. Because cytokines are too large to freely pass through the blood–brain barrier, investigators have examined how cytokine signals are communicated to the brain. Mechanisms include (1) passage through leaky regions in the blood–brain barrier, such as the circumventricular organs, (2) binding to cytokine-specific transport molecules expressed on brain endothelium, and (3) activation of vagal afferent fibers that transmit cytokine signals to specific brain nuclei, including the nucleus of the solitary tract and the paraventricular nucleus in the hypothalamus.

Once cytokine signals have reached the brain, there is a cytokine network within the brain characterized by protein and mRNA for cytokines and their receptors in multiple cell types. This network includes a rich expression in brain regions relevant to depression and cognition, including the limbic and reward (striatum) circuitry, hypothalamus, and prefrontal cortex. Most central nervous system cytokines emanate from microglia, astrocytes, and endothelial cells of the cerebrovasculature; however, cytokine and cytokine receptor expression in neurons has also been described. Microglia, astrocytes, and endothelial cells are also capable of producing inflammatory mediators, including chemokines, adhesion molecules, prostaglandins, nitric oxide, and COX-2.

Cytokines and the Pathophysiology of Depression

Activation of the cytokine network within the brain has been associated with effects on many of the pathways that have been shown to be relevant to the pathophysiology of depression, including neurotransmitter metabolism, neuroendocrine function, synaptic plasticity, and information processing (Figure 3).

Cytokine effects on neurotransmitter metabolism

Cytokine effects on neurotransmitter metabolism include alterations in the metabolism of dopamine, norepinephrine, and serotonin. The overlap between the effects of stress and the effects of cytokines on neurotransmitter metabolism are particularly intriguing and suggest that shared mechanisms may be involved.

The influence of inflammatory processes on serotonin metabolism is of particular note, especially given the relationship between inflammation and the activation of the metabolic enzyme, indolamine 2,3-dioxygenase(IDO), and the p38 MAPK cascade. IDO is responsible for breaking down tryptophan into kynurenine and quinolinic acid, thus decreasing the availability of tryptophan (the primary precursor of serotonin) and increasing the concentration of quinolinic acid, which has neurotoxic properties. In addition, recent data have shown that activation of the p38 MAPK cascade by proinflammatory cytokines, including IL-1, is associated with upregulation of the activity and expression of the serotonin transporter, thereby further decreasing serotonin availability in the synapse. Thus, by decreasing serotonin availability both at the level of its metabolism as well as its presence in the synapse, inflammatory processes can contribute to marked decreases in serotonin neurotransmission, a hallmark of the pathophysiology of depression. Indeed, the efficacy of serotonin reuptake inhibitors in cytokine (IFN-α)-induced depression is consistent with the relevance of serotonin pathways in the mechanism of cytokine-induced behavioral changes.

Cytokine effects on neuroendocrine function Probably the most studied of the effects of inflammation and proinflammatory cytokines on the nervous system are the effects on the neuroendocrine system. Proinflammatory cytokines have been shown in multiple studies in laboratory animals to activate the expression, production, and release of the neuropeptide, corticotropin-releasing hormone (CRH), a primary regulator of the stress response. Given the capacity of CRH to induce behavioral changes that overlap with sickness behavior, including altered locomotor activity, anorexia, and impaired sleep, it has been suggested that cytokine-induced increases in CRH may in part contribute to the behavioral effects of cytokines. Indeed, increased CRH activity has been a consistent finding in major depression as manifested by increased CRH in the cerebrospinal fluid, increased CRH mRNA and protein in the hypothalamus, and decreased CRH receptors in the prefrontal cortex (likely secondary to increased CRH production).

Another neuroendocrine alteration that can be induced by proinflammatory cytokines and may contribute to the pathophysiology of depression is glucocorticoid resistance. Glucocorticoid resistance, as manifested by nonsuppression of the hypothalamic-pituitary-adrenal (HPA) axis following administration of the synthetic glucocorticoid, dexamethasone (DEX) (determined in the DEX suppression test or the DEX/CRH test), is a reliable finding in patients with depression. In addition, glucocorticoid resistance is believed to contribute to the hypersecretion of CRH, which is normally under negative feedback regulation by glucocorticoids. Glucocorticoid resistance in depression is believed to stem from alterations in the expression and/or activity of the glucocorticoid receptor (GR). Indeed, transgenic mouse models of decreased GR availability have been associated with vulnerability to depressive-like behaviors, including learned helplessness. Regarding cytokines, a number of studies have demonstrated that activation of p38 MAPK pathways as well the STAT5 pathway is associated with decreased GR function. Activation of p38 signaling leading to glucocorticoid resistance has been posited as a major mechanism involved in steroid-resistant asthma. p38 appears to phosphorylate the GR, leading to disruption of GR nuclear translocation, whereas STAT5 has been shown to dysregulate the GR through direct protein–protein interactions. Thus, proinflammatory cytokines can not only stimulate CRH expression, production, and release, but also, through inhibitory effects on the GR, disrupt GR-mediated negative feedback pathways on CRH, all of which contribute to a feed-forward cascade of inflammation and activation of CRH.

Cytokine effects on synaptic plasticity Recent research has focused on the role of synaptic plasticity and related growth factors, including brain-derived neurotropic factor (BDNF) in the pathophysiology and treatment of depressive disorders. Data suggest that activation of cytokine pathways in the brain during stress may influence growth factors, thereby representing another mechanism by which stress via cytokines may contribute to depression. For example, IL-1 has been shown to mediate the inhibition of BDNF expression in the hippocampus of rats following social isolation stress. TNF-α and activation of NF-κB signaling pathways have also been associated with stress-induced production of nitric oxide within the CNS, which in turn may contribute to alterations in neuronal growth and survival through the induction of oxidative stress. Oxidative stress has been shown to alter promoter function for several genes central to synaptic plasticity. Consistent with these findings is that stress-induced neuronal cell loss in rodents has been linked to increased concentrations of TNF-α and NF-κB.

Cytokine effects on information processing Recent data indicate that another major pathway by which cytokines may influence behavior is through their effects on the activity and function of neurocircuitry relevant to cognition. Indeed, using functional magnetic resonance imaging (fMRI) and visuospatial attention task, investigators have demonstrated that patients chronically administered the innate immune cytokine, IFN-α, exhibit significantly greater activation of the dorsal anterior cingulate cortex (dACC) than do control subjects not undergoing IFN-α treatment. The dACC is a brain region that plays a fundamental role in conflict monitoring and error processing and participates in the allocation of mental resources for accomplishing tasks and reaching goals. Important interconnections also exist between the dACC and the striatum, which previous studies have shown is a target for cytokine activity in patients with HIV and in patients undergoing IFN-α therapy. Increased activation of the dACC may indicate cytokine-induced increases in mental effort or may reflect an increased sensitivity (decreased threshold) for processing negative events. Increased activity in the dACC has also been observed in subjects with high-trait anxiety, neuroticism, and obsessive–compulsive disorder, all conditions of increased sensitivity to environmental and interoceptive stimuli and all conditions associated with an increased vulnerability to the development of mood and anxiety disorders. Thus, altered dACC activity as a function of exposure to innate immune cytokines may represent a cognitive pathway to psychopathology.

Integrating Stress, Cytokines, and Depression

From the literature reviewed in the preceding sections, it is clear that stress activates the innate immune response and that an activated innate immune response can in turn contribute to many of the pathologies that characterize depression, including both biological and behavioral features. In addition, there is a rich literature demonstrating that stress is a reliable precipitating factor for the development of depression. Despite the logical connections which can be drawn from these individual relationships, it may be difficult to establish that all of these factors together conspire in some demonstrable linear fashion to lead to disease. More likely, stress effects on the immune system and the immune system's contribution to the pathophysiology of depression represent two of many convergent pathways which together contribute to a multi-determined disease endpoint. Nevertheless, from the standpoint of treatment and prevention, several translational implications are apparent. The first has to do with the role of the autonomic nervous system in regulating the immunologic response to stress. Research and interventional strategies that address the relative balance of sympathetic and parasympathetic outflow to the immune system are particularly relevant in this regard and require focus on the transmitters and signal transduction pathways involved. Special attention needs to be paid to the catecholamines and acetylcholine and to their signaling pathways, as well as to their interaction with inflammatory signaling pathways, including NF-κB and MAPK. As for the immune system, it has become increasingly apparent that further knowledge about inflammatory signaling molecules that are involved in reproducing pathologies that characterize depression (including p38 and NF-κB) may be particularly relevant regarding the identification of therapeutic targets, as well as in determining the full clinical implications of the immune system as it relates to behavior. Finally, further detailing of the neurocircuitry of cytokine effects on the brain may help identify brain regions that subserve cytokine effects on behavior, and thus represent circuits for therapeutic intervention through either somatic or psychotherapeutic means.

Summary

In conclusion, the data indicate that depression is not an isolated disease state of the brain; rather, it is a systemic illness invoking a complex interaction of the nervous, endocrine, and immune systems. As such, the innate immune response appears to be a potential pivotal mediator in the relationship between stress and depression. Further elucidation of the interactions between stress, cytokines, and depression may ultimately provide a novel conceptual framework with the potential to guide future developments in research and therapy as they relate to the multidimensional pathophysiologies that lead to psychiatric diseases such as depression.

See also: Chronic (Repeated) Stress: Consequences, Adaptations; Depression and the Brain; Posttraumatic Stress Disorder: Overview; Sleep and Sleep States: Cytokines and Neuromodulation; Stress and Suicide; Stress and Vulnerability to Brain Damage; Stress Response: Genetic Consequences; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Abbas AK and Lichtman AH (eds.) (2005) *Cellular and Molecular Immunology,* 5th edn. Philadelphia: W.B. Saunders Co.

Bierhaus A, Wolf J, Andrassy M, et al. (2003) A mechanism converting psychosocial stress into mononuclear cell activation. *Proceedings of the National Academy of Sciences of the United States of America* 100(4): 1920–1925.

Capuron L, Pagnoni G, Demetrashvili M, et al. (2005) Anterior cingulate activation and error processing during interferon-alpha treatment. *Biological Psychiatry* 58: 190–196.

Dantzer R (2004) Cytokine-induced sickness behaviour: A neuroimmune response to activation of innate immunity. *European Journal of Pharmacology* 500(1–3): 399–411.

Glaser R and Kiecolt-Glaser JK (2005) Stress-induced immune dysfunction: Implications for health. *Nature Reviews Immunology* 5(3): 243–251.

Hayley S, Poulter MO, Merali Z, et al. (2005) The pathogenesis of clinical depression: Stressor- and cytokine-induced alterations of neuroplasticity. *Neuroscience* 135(3): 659–678.

Johnson JD, Campisi J, Sharkey CM, et al. (2005) Catecholamines mediate stress-induced increases in peripheral and central inflammatory cytokines. *Neuroscience* 135(4): 1295–1307.

Moraska A, Campisi J, Nguyen KT, et al. (2002) Elevated IL-1beta contributes to antibody suppression produced by stress. *Journal of Applied Physiology* 93(1): 207–215.

Nadjar A, Bluthe RM, May MJ, et al. (2005) Inactivation of the cerebral NF kappaB pathway inhibits interleukin-1beta-induced sickness behavior and cFos expression in various brain nuclei. *Neuropsychopharmacology* 30(8): 1492–1499.

Pavlov VA, Ochani M, Gallowitsch-Puerta M, et al. (2006) Central muscarinic cholinergic regulation of the systemic inflammatory response during endotoxemia. *Proceedings of the National Academy of Sciences of the United States of America* 103(13): 5219–5223.

Raison CL, Capuron L, and Miller AH (2006) Cytokines sing the blues: Inflammation and the pathogenesis of depression. *Trends in Immunology* 227: 24–31.

Rivest S (2003) Molecular insights on the cerebral innate immune system. *Brain, Behavior and Immunity* 17(1): 13–19.

Tyring S, Gottlieb A, Papp K, et al. (2006) Etanercept and clinical outcomes, fatigue, and depression in psoriasis: Double-blind placebo-controlled randomised phase III trial. *Lancet* 367(9504): 29–35.

Wang X, Wu H, and Miller AH (2004) Interleukin-1 alpha-induced activation of p38 mitogen-activated kinase inhibits glucocorticoid receptor function. *Molecular Psychiatry* 9: 65–75.

Zhu CB, Blakely RD, and Hewlett WA (2006) The proinflammatory cytokines interleukin 1beta and tumor necrosis factor-alpha activate serotonin transporters. *Neuropsychopharmacology* 31(10): 2121–2131.

Stress, Dopamine, and Puberty

S L Andersen and M P Leussis, McLean Hospital, Belmont, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Stress, Puberty, and Adolescence

A stressor is anything that is perceived by the individual to cause emotional, mental, or physical strain. This definition suggests that awareness of the stress is vital for its impact on both emotional and cognitive processing. Thus, what is considered stressful needs to be placed in developmental context when it is applied to immature systems. For example, conflict and missed deadlines are stressors for adolescents and adults, whereas neglectful parents or no friends are stressors for children. Relevant examples for adult animals in a species-specific context include choice conflicts or delay of reinforcement paradigms, whereas maternal deprivation or social isolation model childhood and young adolescent stressors. Age-appropriate stressors are important for correct interpretation of behavioral and neurochemical responses at different stages of development.

Development occurs in a series of phases, defined physiologically or socially. The term puberty refers to the achievement of reproductive fitness hailed by secondary sex characteristics and maturity of the reproductive systems. As part of this process, adolescence describes the period between the onset of puberty and adulthood when the individual defines and refines the sense of self in terms of social values, personal values, and future goals. The use of the terms puberty versus adolescence in this article will reflect whether the topic refers to physiological or social changes and the timing of the developmental period discussed. For example, stress during adolescence may depend more on social context than physiological deprivation given the importance of social development during this period. Adolescence will also be used to refer to the period of development that spans between puberty onset and adulthood.

As outlined in this article, understanding how the dopamine system develops during the transitions between childhood, adolescence, and adulthood sets the stage for framing the uniqueness of the impact of stress on behavior, neurochemistry, and neuroanatomy.

The Dopamine System

Neuroanatomy

The catecholaminergic dopamine system is more localized in its innervation patterns relative to many other neurotransmitter systems. Dopaminergic cell bodies, also known as A9 and A10, emanate from the ventral tegmental nucleus (VTA) and the substantia nigra, respectively. Smaller, less well-characterized dopamine terminals have been identified in the amygdala, hippocampus, and cerebellum. From the VTA, separate dopaminergic neurons project to either the nucleus accumbens and comprise the mesolimbic dopamine system, or the prefrontal cortex and comprise the mesocortical system. The substantia nigra sends projections to the caudate and putamen (collectively referred to as the striatum), and forms the nigrostriatal dopamine system.

Signaling Mechanisms in the Dopamine System

The signaling mechanisms of the dopamine system are associated with five main receptor subtypes, referred to as D1–D5 (Figure 1). The majority of these receptors are G-protein-coupled receptors (GCPRs). D1 and D5 are more excitatory in nature, are linked to stimulatory G-proteins, and classically increase the activity of the second messenger cyclic adenosine monophosphate (cAMP). In contrast, D2–D4 dopamine receptors are inhibitory in nature, couple with inhibitory G-proteins, and reduce cAMP activity. This is merely the beginning of the signaling cascade, as other second messengers (protein kinases) and the inositol triphosphate (IP_3) system are also involved.

Dopamine receptor signaling mechanisms, activated by environmental changes, can induce transcriptional regulating proteins, which ultimately leads to changes in gene transcription. Enhanced expression of cyclic AMP response element-binding protein (CREB) and delta-Fos subunit B (ΔFosB), have been extensively studied. It is believed that changes in these proteins signal the beginning of long-term alterations, or plasticity, that occur in response to drugs of addiction or stress. Changes in immediate early gene expression, notably the Fos and Jun genes, are among the most commonly reported. Their expression, in turn, increases the production of numerous proteins that produce the enduring actions of

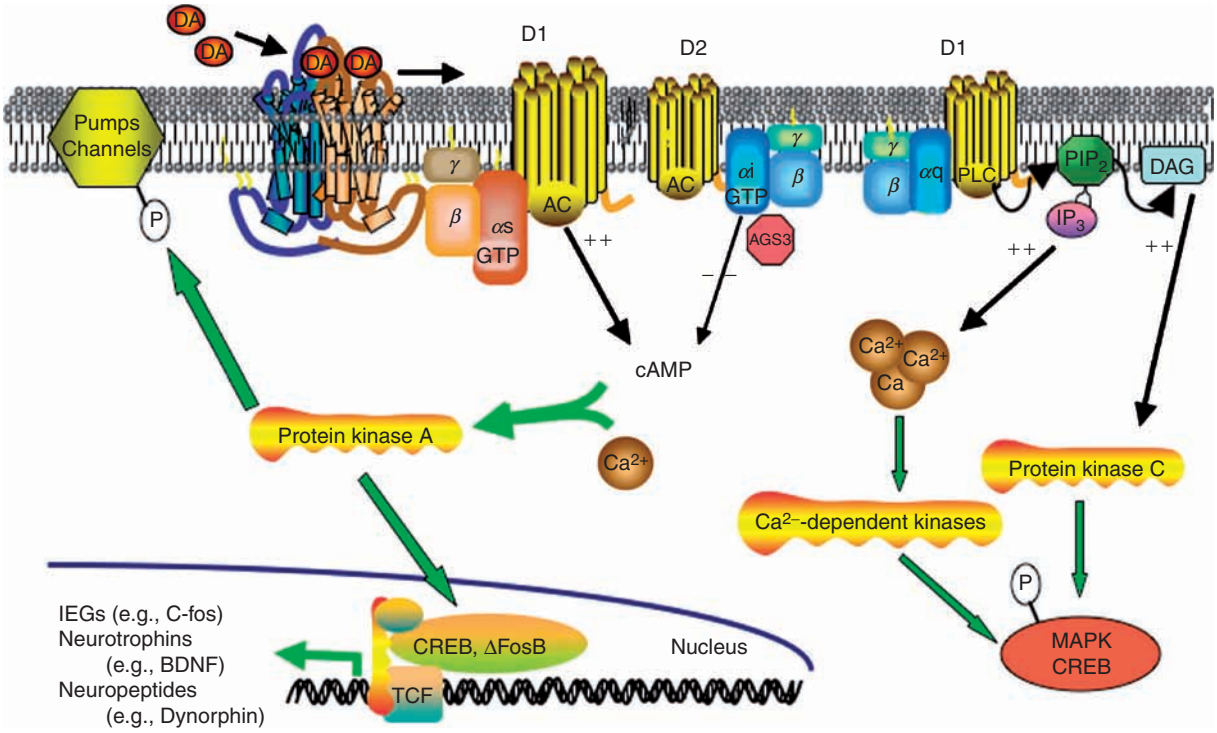

Figure 1 Dopamine receptor signaling cascades. Dopamine acts at G-protein-coupled receptors to produce activational (via D1/D5 receptors) or inhibitory effects (via the D2 family (D2–D4)) overall. Dopamine acting at D1/D5 receptors predominantly activates the stimulatory G-protein-coupled receptors (Gs), which upregulate adenylyl cyclase activity, leading to an increase in cAMP. Higher levels of cAMP activate protein kinase A, which phosphorylates many proteins, including transcriptional activators, and membrane-bound ion pumps and channels. Phosphorylation of ion pumps/channels produce short-term alterations in the membrane potential, causing the cell to become more or less likely to fire. In contrast, longer-term effects are seen through the phosphorylation of transcription factors such as CREB and ΔFosB. These transcription factors ultimately increase the production of proteins, including *c-fos*, neurotrophins, and neuropeptides. Dopamine at D2–D4 receptors activates the inhibitory G-protein-coupled receptors (Gi), which decreases adenylyl cyclase activity. Dopamine can also act through a secondary D1/D5 pathway, via Gq receptors and the phosphatidylinositol pathway. In this pathway, dopamine binding increases phospholipase C activity, which produces inositol triphosphate and diacylglycerol from phosphatidylinositol. Increased levels of diacylglycerol activate protein kinase C, which phosphorylates other kinases such as mitogen-activated protein kinase, and transcription factors such as CREB. Higher levels of inositol triphosphate increase cytoplasmic Ca^{2+} levels, thereby activating calcium-dependent kinases, which further continue the process of protein phosphorylation and activation. AC, adenylyl cyclase; AGS3, activator of G-protein signaling 3; BDNF, brain-derived neurotrophic factor; cAMP, cyclic adenosine monophosphate; CREB, cAMP response element-binding protein; DA, dopamine; DAG, diacylglycerol; ΔFosB, delta-Fos subunit B; GTP, guanosine triphosphate; IEGs, immediate early genes; IP_3, inositol triphosphate; MAPK, mitogen-activated protein kinase; PIP_2, phosphatidylinositol; PLC, phospholipase C; TCF, transcription factor.

dopamine receptor activation. Most of this work has been performed in animals exposed to drugs of addiction. However, because stress and addictive agents often work by the same mechanistic cascades, the information from stimulant studies is likely to transfer to stress exposure. This represents an area that is tremendously understudied to date.

Brain Development and the Dopamine System

Overproduction and Elimination of Synapses during Puberty

Profound brain changes occur during the course of postnatal development and adolescence and represent the final phase of this transition. Cortical structures, including the prefrontal cortex, and subcortical structures, including the striatum and nucleus accumbens, undergo an exuberant overproduction of synapses and associated signaling molecules between childhood and adolescence. This is subsequently followed by the selective elimination, or pruning, of synapses before adulthood. This pruning is similar to what occurs with trees: branches (dendrites) are selectively cut back to optimize growth potential (or in the case of the brain, efficiency) to match the needs of the environment. In humans, cortical gray matter volume reaches peak density at 11.6 years in females and 12.2 years in males. In rats, similar changes in density are observed at approximately 40 days of age. For all mammalian species, the timing of these synaptic

density changes coincides with the onset of puberty. However, it is noteworthy that different brain regions have their own unique course of development. For example, the striatum reaches peak dopamine receptor density at pubertal onset, and prunes shortly thereafter. The later-developing prefrontal cortex attains maximum receptor density at the same age, but delays pruning until young adulthood. This differential time course has been used to explain why motor disorders (attention-deficit/hyperactivity disorder and Tourette's syndrome) affect children, whereas schizophrenia typically emerges during young adulthood. Different periods of vulnerability are associated with the individual trajectory of a given brain region. Thus, in order to understand the impact of stress exposure during development, it is imperative to appreciate the state of maturity of the underlying trajectory at the time of insult.

Functional Development of the Dopamine System

Striatum and nucleus accumbens The subcortical structures of the dopamine system, the striatum and the nucleus accumbens, demonstrate functional changes during the onset of puberty. Dopamine D1 and D2 receptors are overproduced in the striatum, but not in the nucleus accumbens. Extracellular levels of dopamine decrease modestly during puberty in these regions. Dopamine autoreceptors regulate the rate of dopamine synthesis and release much like a thermostat functions to control fluctuations in temperature. Despite this regulatory role, the sensitivity of the dopamine autoreceptors does not change during puberty. These data suggest that changes within other systems during puberty mediate the observed functional changes in subcortical dopamine structures.

The impact of acute stress on either the striatum or the accumbens is not well understood. The striatum is associated with motor patterns, including stereotypies, habit learning, and initiation of movement. The nucleus accumbens plays vital roles in activity and reward. During pubertal onset, spontaneous locomotor activity reaches its highest point (females are more active than males), and reward systems are vulnerable to the addictive effects of drug exposure. Animal handling prior to testing enhances cocaine-induced locomotion relative to adults, suggesting enhanced reactivity. These behaviors reflect changes within the dopamine system, but are not well characterized as they relate to acute stress *per se*. The effects of chronic, early life stress on this system are of interest and are discussed below.

Prefrontal cortex The functional activity of the prefrontal cortex does not come on-line fully until the onset of puberty. This increase in dopaminergic activity is important for understanding a number of adolescent transitions. Immunohistochemistry shows that dopaminergic innervation of the prefrontal cortex approaches adult levels during puberty. Along with the increase in innervation, functional rearrangements also occur that increase dopamine activity in response to external stimuli. Dopamine D1 and D2 receptor density peaks during this period before being pruned back in adulthood. Dopamine receptors increase by 38%, plateau during puberty, and are ultimately pruned 33% in adulthood in the rat. Cyclic AMP levels are 35% higher during adolescence than adulthood. This enhancement of dopamine function during puberty may be partly related to the simultaneous loss of regulatory processes. Although poorly understood, modulation of dopamine synthesis in the prefrontal cortex wanes during pubertal onset and is absent by adulthood. Taken together, the cortical dopamine system rapidly reaches an adultlike state during this time.

Circuitry within these regions becomes increasingly complex, as innervation from the dopaminergic cell bodies and other neurotransmitter systems continues during puberty. Connections between stress-related brain regions also become more adultlike during puberty. Once the perception of stress has occurred, activational systems, such as glutamatergic efferents to various emotional and cognitive brain regions, are responsible for the course of action.

Sex Differences in the Dopamine System during Puberty

Sex differences exist in the immature dopamine system. It is often the case that females overproduce (and prune) less than males. Dopamine receptors are overproduced in the striatum of males during puberty, whereas little change in either D1 or D2 receptor density is observed in females. The pruning of D1 and D2 receptors shows that density changes by 54% and 63%, respectively, in the striatum, and approximately 25% in the accumbens of males between 40 and 80 days of age. This is not to imply that females do not significantly reorganize during adolescence. For instance, signaling mechanisms, including cAMP, in females are elevated during this period before pruning in adulthood.

In preparation for reaching sexual maturity, levels of the gonadal hormones estrogen and testosterone rise in the prepubescent subject. However, their dramatic rise seems to have little impact on the neuroanatomy of the dopamine system. Given the profound sex differences that exist in the dopamine system, it is plausible to hypothesize that testosterone increases the overproduction of dopamine receptors or that

estrogen suppresses it. Neither seems to be the case. Estrogen in adults or cell cultures increases dendritic branching, yet neuroanatomical changes in female adolescents are not observed. Based on immunohistochemistry with tyrosine hydroxylase antibodies (a marker that stains the rate-limiting enzyme in dopamine production), dopaminergic innervation of the three major terminal fields does not change with age in females. In addition, removal of the gonadal hormones well before they surge does not alter dopamine expression. These data suggest that gonadal hormones are not involved in the activational phases of changes within the dopamine system. Rather, sex differences may be organized during the prenatal period.

Stress Reactivity during Puberty

The Hypothalamic–Pituitary–Adrenal Axis and Stress Response

In order to understand stress reactivity, it is necessary to step back and briefly discuss the maturation of the hormonal system of stress responses. The hypothalamic–pituitary–adrenal (HPA) axis is involved in the secretion of stress hormones. Activation of the paraventricular nucleus of the hypothalamus (PVN) is involved in the release of corticotropin-releasing hormone (CRH), which activates the cascade of stress-related changes at the hormonal level. The PVN receives dopaminergic innervation, and thus is modulated by higher-level processing of stressors. CRH travels to the pituitary, causing the release of adrenocorticotropin hormone, which ultimately signals the release of the stress effector hormone cortisol (human and nonhuman primates) and corticosterone (rodents) from the adrenal glands. Cort (abbreviated as such to capture both hormone versions) then travels back to the brain to signal stress-related changes.

The onset of puberty is a transition point in HPA reactivity. Cort activation following acute restraint stress is more pronounced and prolonged in adolescent than in adult rats. Adolescent rats secrete more Cort in response to acute restraint stress than adult rats. Moreover, the response is larger in female than male adolescent rats. The Cort response following chronic stress, however, differs between adolescents and adults. Adolescent rats with a history of chronic stress exposure return to baseline levels faster than adults.

Age differences in stress-related reactivity are observable with *c-fos* immunoreactivity in the PVN. *c-fos* is often used as a marker of neuronal activity and can be used to map where changes occur. Prepubertal male rats have increased *c-fos* immunoreactivity in CRH-containing neurons of the PVN relative to adults in response to an acute stressor. These findings suggest that the prepubertal rat is more sensitive to the effects of stressful events than adults. They also raise the issue of what is initiating this cascade of stress-related effects.

Dopamine Systems and Stress Responsiveness

Of the three primary ascending dopamine systems, the mesocortical dopamine system is the one commonly associated with stress responses. The prefrontal cortex has an abundant amount of glucocorticoid receptors, second only to the hippocampus, which makes the prefrontal cortex particularly vulnerable to stress and its cascade of effects. In adults, stressors including footshock, restraint, and hotplate preferentially increase dopamine release in the prefrontal cortex. The nucleus accumbens is also activated by these stressors, but to a lesser extent than the prefrontal cortex. The striatum is minimally activated in response to acute stressors, but may be more involved in long-term compensations involving habits.

Stress Reactivity in the Dopamine System Emerges with Age

It is during pubertal onset that the cortex becomes increasingly vulnerable to the acute effects of stress relative to younger and older ages. Evidence from a pharmacological stressor, the GABAergic antagonist FG-7142, demonstrates that neuronal activity within the three major dopamine systems shifts in dominance. In this study, changes in the immediate early gene *c-fos* were quantified following FG-7142 and converted to a percentage of total *c-fos* immunolabeled neurons to derive an index of maturation. The subcortical structures are significantly activated before puberty and may be more vulnerable to stress at this age. These younger animals do not show the adult-typical pattern of dominant prefrontal cortex activation. Not surprisingly, the adult-typical pattern matures between 18 and 45 days of age in the rat – coincident with the onset of puberty. Further, *c-fos* immunolabeling is enhanced in the prefrontal cortex at 45 days of age relative to adult rats, demonstrating that an acute stressor during adolescence has a greater impact on cortical structures than comparable exposure in adulthood. These findings offer a neurochemical explanation for why children are more likely to increase their activity levels in response to stress. In contrast, as cortically based stress reactivity emerges, adolescents and adults experience stress on a more cognitive basis.

Circuits Involving Dopaminergic Modulation of Stress

Prefrontal cortex The functional nature of the prefrontal cortex places it in a vulnerable niche. The cortex is involved in motivation and cognition, in part by integrating environmental and emotional information and deciding the next course of action. As an example, abnormality or immaturity of the system is associated with impulsive or inappropriate decisions for a given context.

The cortex, however, does not act alone. Connections to other brain regions play a critical role in stress responses during puberty. One such circuit loop is the prefrontal cortex to dorsal raphe. Here, the prefrontal cortex has the important role of differentiating controllable from uncontrollable stress (Figure 2). The learned helplessness paradigm models depressive-like behaviors in rats. This test was designed to parse controllable from uncontrollable stress effects, and relies on cortical modulation. Normal, pubertal female rats demonstrate a higher level of helplessness following one day of conditioning to a series of escalating tail shocks. This sex difference was reversed, however, following a 5-day social stressor that occurs immediately before the onset of puberty. After this subchronic stress exposure, males demonstrate increased learned helplessness, whereas females show some degree of improvement in this behavior.

Amygdalar Contributions to Stress

The amygdala also plays an important role in stress responses, but more at the level of mediating fear responses. Fear, by definition, is a reaction to a future,

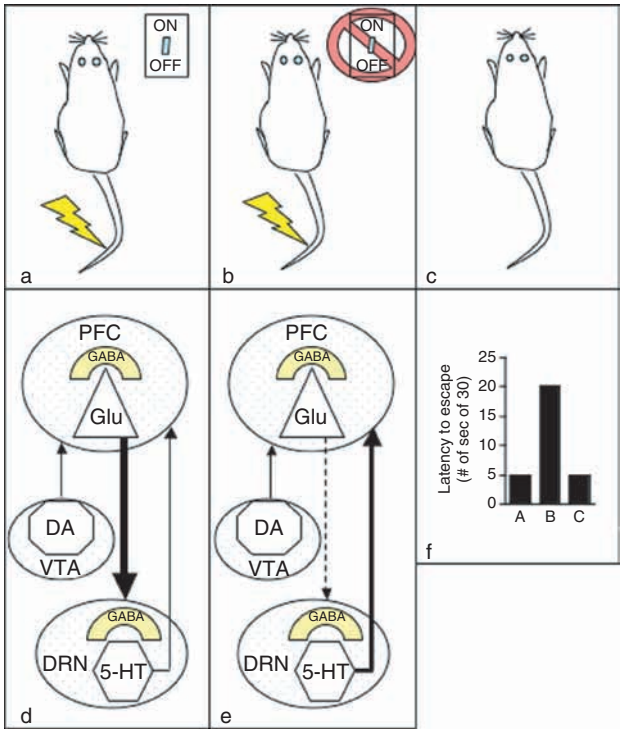

Figure 2 Learned helplessness paradigm. The learned helplessness model of depression is based on the differential response of rats to (a) controllable shock, (b) uncontrollable shock, or (c) no shock. Controllable (escapable) shock rats can terminate the tail shock by completing a behavioral response (turning a wheel). Uncontrollable (inescapable) shock rats are yoked to the controllable shock rat, so they receive the same number and duration of shocks, but are unable to control the termination of each shock. No shock rats are restrained similarly to the other two rats, but receive no shock during the training phase. The latency to escape a footshock (out of 30 s) and the number of escape failures (out of 30 trials) are recorded in a shuttlebox the next day. Leaned helplessness is defined as longer latencies to escape (f) and a greater number of escape failures than the controllable shock animals (A) or no shock controls (C). The anatomical circuit involved in dissociating controllable from uncontrollable stress involves the prefrontal cortex (PFC) and dorsal raphe nucleus (DRN). (d) Following controllable stress, glutamatergic projections from the PFC act on GABAergic interneurons in the DRN to suppress activation of serotonergic neurons. (e) Following uncontrollable stress, there is an increase in activation of serotonergic (5-HT) DRN neurons that project to the PFC, which serves in part to inhibit the glutamatergic neurons projecting back to the DRN. In the case of either controllable or uncontrollable stress, dopamine (DA) from the ventral tegmental nucleus (VTA) serves to signal the presence of a stressor. This figure and description is based on the seminal work of Maier and colleagues: Amat J, Baratta MV, Paul E, Bland ST, Watkins LR, and Maier SF (2005) Medial prefrontal cortex determines how stressor controllability affects behavior and dorsal raphe nucleus. *Nature Neuroscience* 8: 365–371.

probabilistic event based on past, conditioned responses. As such, fear responses are highly dependent on the context. Dopamine, via innervation from the VTA, plays an important role in the plasticity that underlies the associations between the context and the unconditioned, fear-eliciting stimulus. Increases in cortical dopamine are correlated with increased anxiety that increases during puberty. How levels or innervation of the amygdalar dopamine system change in response to stress during puberty is not known and remains another area of future investigation.

More is known, however, on the innervation of fibers from the amygdala to the prefrontal cortex. The prefrontal cortex receives input from the amygdala to integrate environmental contextual information with emotional processing. This innervation occurs throughout postnatal brain development, but begins to plateau during the onset of puberty. Studies with systemic and microinjections of dopamine agonists, in addition to dopamine depletions, show that these effects are modulated by dopamine. Again, the developmental profile of this effect is not known.

Synthesis

Taken together, data from Cort changes, *c-fos* localization, and behavioral paradigms suggest that the responses to stressful events rapidly shift during puberty onset. Increased dopaminergic activity in the prefrontal cortex is likely to play a significant role in the processing of environmental events, and thus facilitate enhanced responsiveness to stressful events. Acute stress during puberty onset has a greater impact on females compared with males. However, chronic stress exposure immediately before testing changes this relationship. Males have increased responsiveness within the HPA axis and in learned helplessness following chronic stress exposure when challenged during the pubertal period. The effect on adolescents is short lived, and animals show a return to baseline faster than adult cohorts. Studies such as these further suggest that the stress system is programmed by previous experiences differently during development when compared with adults.

Programming of Immature Stress-Responsive Systems

Sensitive Periods of Development

As discussed above, the brain develops in a series of stages. Overproduction and pruning of synaptic connections and signaling mechanisms program the brain to match the demands of the environment. What this means is that stress responses are appropriate for the level of stress that one has experienced. The immature brain, in turn, encodes these demands by pruning appropriately. However, the timing of stress experiences significantly influences which brain areas are vulnerable. Brain areas undergoing the most change are likely to be affected by stress exposure to a greater degree than those that have yet to overproduce connections or have already pruned. These periods of selective vulnerability are known as sensitive periods. The impact of stressful events on the immature brain will depend on the sensitive period during which the stressor occurred.

Early Postnatal Stressors and Brain Development

In order for something to be perceived as stressful, it has to involve the manipulation of something that is highly relevant to the organism. During early postnatal life, one of the single most important variables is the level of parental (predominantly maternal) care. Therefore, it should not be surprising that manipulation of the level of maternal care significantly affects the maturation of stress-related systems. All mammalian species studied to date that have significant variations in maternal care experience changes in Cort responses later in life, although exceptions occur for individual differences.

Events set in motion early in life program the formation of brain anatomy and functional impact follows suit. While detailed timecourses of functional changes within the dopamine system have not been assessed during puberty, it is likely that the impact of early life events manifests during this period.

Rats have been the most-studied species in maternal deprivation paradigms, and the discussion here will be limited to rodent studies. Early maternal deprivation alters the overproduction of synapses that occurs during adolescence. The most vulnerable region affected is the hippocampus, which fails to overproduce synapses between puberty and adulthood in animals that were stressed early in life. Early stress, however, has little impact on the dopamine system in the prefrontal cortex.

The nucleus accumbens shell, on the other hand, fails to overproduce dopaminergic terminals between puberty and adulthood in maternally deprived animals. Dopamine D1 receptors are elevated in the accumbens, with no observable change in the striatum during puberty. The rest of the results that are discussed are based on adult outcomes of the deprivation. Maternal deprivation decreases the density of dopamine transporters in the accumbens. This may explain why cocaine-stimulated dopamine release is higher in these animals. To compensate, animals with a history of maternal deprivation will self-administer

more cocaine relative to controls in adulthood. As the propensity that stimulant abuse in humans emerges during the adolescent period, these changes most likely emerge during this period.

Late Postnatal Stressors and Brain Development

The alternative paradigm to early maternal deprivation involves social isolation beginning from the time of weaning (prepubertal) into adulthood. This paradigm produces a number of behavioral changes linked with reduced cortical dopamine that include deficits in sensorimotor gating of the acoustic startle response, increased locomotion in response to novelty, and enhanced dopamine levels in the prefrontal cortex. Since this paradigm spans across puberty into adulthood, it is difficult to speculate as to when this social stressor exerts its effects on the dopamine system.

Clinical Impact of Stress during Puberty

The transition between childhood and adolescence is a unique period of vulnerability. It is during this period that adolescents (males more so than females) engage in risky behaviors, novelty-seeking, and peer-directed behavior including the abuse of substances. Postpubertal changes also serve as a permissive factor for the expression of mental illnesses that are associated with stress, including schizophrenia and depression. For example, stressful events during adolescence often precipitate the first psychotic episode, which on average occurs in males an average of 6 years earlier than in females.

Traumatic stress also effects adolescent transitions in brain development, especially within the late-developing prefrontal cortex. Exposure to stress is associated with an increase or a decrease in the size of this region, if the timing of the exposure is before or after puberty, respectively. In clear-cut cases of significant, traumatizing, stress, males are affected more than females as cognition is often impaired. For females, the onset of puberty brings with it its own unique challenges. As social roles are defined, the emergence of mood disorders affects females 2.3-fold more than males and rises from 3% of the population to an estimated 8% during adolescence. It is possible that stress during puberty affects the amygdala/cortical development in females, which would increase depression and anxiety.

Part of the emerging role of the prefrontal cortex during this time is to modulate stress responses. Mentally healthy individuals are able to separate stressors into these two categories of controllable and uncontrollable and deal with them accordingly. In contrast, individuals with depression, anxiety, and schizophrenia perceive most stressors as uncontrollable and become overwhelmed (Figure 2). These disorders are associated with elevated levels of cortical dopamine activity in response to stress. The parallel timing of the significant rise in dopamine activity and the prevalence of these disorders during puberty is more than just coincidence. Although this has yet to be demonstrated conclusively, it is likely that increased cortical dopamine plays a role in the perceptual 'switch' between controllable versus uncontrollable stress.

On the other hand, the pubertal transition, dopamine, and its relationship to stress provide a window of opportunity for some. The waning of motor tics in Tourette's syndrome during puberty reflects the benefit of positive changes associated with enhanced dopamine activity in the cortex. Generally, tics are exacerbated in response to stress in Tourette's syndrome patients. The prevalence of tics (or their willful modulation) improves during adolescence.

Conclusions

Despite the universal nature of the stressful transitions that are associated with the onset of puberty, relatively little research has been conducted on this important topic. The emerging role of cortical dopamine increases the perception of stress during puberty. Whether this translates into more or less controllability over life's strife and stress most likely depends on genetic and stress history. What is becoming increasingly clearer, however, is that events that happen before this transition can drastically shape the ability to deal with stressful events later on. The occasion to intervene during this potential window of opportunity beckons more researchers to explore what this area may offer.

See also: Stress and Cognition; Stress and Suicide; Stress and Vulnerability to Brain Damage; Stress Response: Sex Differences; Stress Response and Self-Esteem; Stress, Sex and Adolescent Nicotine Response; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Amat J, Baratta MV, Paul E, Bland ST, Watkins LR, and Maier SF (2005) Medial prefrontal cortex determines how stressor controllability affects behavior and dorsal raphe nucleus. *Nature Neuroscience* 8: 365–371.

Andersen SL (2002) Changes in the second messenger cyclic AMP during development may underlie motoric symptoms in attention deficit/hyperactivity disorder (ADHD). *Behavioral Brain Research* 130: 197–201.

Andersen SL (2003) Trajectories of brain development: Point of vulnerability or window of opportunity? *Neuroscience and Biobehavioral Reviews* 27: 3–18.

Andersen SL and Teicher MH (2004) Delayed effects of early stress on hippocampal development. *Neuropsychopharmacology* 29: 1988–1993.

Kosten TA, Zhang XY, and Kehoe P (2006) Heightened cocaine and food self-administration in female rats with neonatal isolation experience. *Neuropsychopharmacology* 31: 70–76.

Lyss PJ, Andersen SL, LeBlanc CJ, and Teicher MH (1999) Degree of neuronal activation following FG-7142 changes across regions during development. *Developmental Brain Research* 116: 201–203.

Meaney MJ and Aitken DH (1985) The effects of early postnatal handling on hippocampal glucocorticoid receptor concentrations: Temporal parameters. *Brain Research* 354: 301–304.

Romeo RD, Lee SJ, and McEwen BS (2004) Differential stress reactivity in intact and ovariectomized prepubertal and adult female rats. *Neuroendocrinology* 80: 387–393.

Spear L (2000) The adolescent brain and age-related behavioral manifestations. *Neuroscience and Biobehavioral Reviews* 24: 417–463.

Teicher MH (2002) Scars that won't heal: The neurobiology of child abuse. *Scientific American* 286: 68–75.

Wommack JC and Delville Y (2002) Chronic social stress during puberty enhances tyrosine hydroxylase immunoreactivity within the limbic system in golden hamsters. *Brain Research* 933: 139–143.

Young EA and Altemus M (2004) Puberty, ovarian steroids, and stress. *Annals of the New York Academy of Science* 1021: 124–133.

Stress, Sex and Adolescent Nicotine Response

F M Leslie and M K Park, University of California at Irvine, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Epidemiological studies have shown that tobacco use begins primarily during adolescence. Adolescence is a period of major biological changes, including maturation of limbic brain regions that regulate emotional responding, and alterations in hormonal levels. It is therefore not surprising that cognitive and inhibitory control processes and stress responses continue to mature throughout this period. Active brain growth and plasticity during adolescence are important in that these processes permit teenagers to adapt their behavioral responses to the surrounding environment, at a time when teens typically leave the family and become independent, and to learn from new experiences. However, these processes also represent a period of high vulnerability to behavioral control disorders, including substance abuse.

Uniqueness of Nicotine-Induced Responses during Adolescence

Although it has long been believed that teen smoking is the result of social influences, recent studies have shown that adolescence is a time of heightened biological sensitivity to tobacco constituents. Many teenage smokers rapidly progress to dependence after they start to inhale, and have difficulty quitting smoking when they do. This is particularly evident in those who begin to smoke during their early teens.

Rodents, which have been used extensively as an experimental model to examine age differences in drug response, exhibit during their fifth and sixth weeks after birth (postnatal days (P) 28–42) many of the behavioral and hormonal changes evident in human teenagers. During this developmental period, defined as rodent adolescence, mice and rats exhibit altered behavioral and neuronal responses to nicotine, the major psychoactive component of tobacco. Moreover, there is greater sensitivity to nicotine during early adolescence as compared to late, indicating that dramatic developmental changes occur within this time frame. During early adolescence, but not later adolescence or adulthood, male rodents exhibit increased sensitivity to the rewarding effect of nicotine, as evidenced in conditioned place preference, two-bottle choice, and self-administration tests. They also show decreased sensitivity to the aversive effects of nicotine. Furthermore, there are differences in other behavioral effects of nicotine, including locomotor activation and anxiety reduction, and in patterns of drug-induced neuronal activation in limbic and sensory brain regions, as measured by expression of immediate-early genes.

Epidemiological studies have shown a developmental progression of drug use during adolescence, from tobacco and alcohol during early teens through marijuana to later use of cocaine, amphetamine, and heroin. Such findings have led to the suggestion that tobacco may act as a 'gateway' drug by altering the adolescent brain, making it more responsive to the effects of other addictive substances. Animal studies provide considerable support for the view that nicotine produces irreversible changes in adolescent brain structure and function. Brief treatments with low doses of nicotine during early adolescence produce long-term changes in a number of neurochemical systems, including monoamine, acetylcholine, glutamate, and peptide neurons. Neurotoxic effects, including cell loss in the hippocampus, result from treatment with higher doses of nicotine. Adolescent nicotine exposure also produces corresponding long-term changes in a number of behaviors, including locomotion, learning, emotional responding, and nicotine self-administration. In all of these reports, the consequences of adolescent nicotine treatment are very different from those of a similar treatment in adults. Thus adolescent nicotine exposure produces profound changes in brain plasticity during a sensitive period for limbic system development.

Recent brain imaging studies in humans have shown that some of the structural changes that occur during adolescence are sex specific. In particular, there are significant sex differences in the development of the amygdala and hippocampus. Whereas the growth of the amygdala is enhanced in males as compared to females, the opposite is true for the hippocampus. These sex-dependent changes in regional growth are consistent with the enhanced expression of androgen receptors in the amygdala and estrogen receptors in the hippocampus. Neurochemical and structural studies in animals have yielded similar findings, and show that the unique actions of nicotine on adolescent brain may differ in males and females. Although not all studies have systematically examined whether there are sex differences in adolescent responses to nicotine, in those that have, substantial sex differences have been found. Whereas self-administration of a mixture of tobacco constituents, nicotine and acetaldehyde, decreases with age in males, this age-dependent decline is not observed in

females. Sex differences have also been observed in the anxiolytic and locomotor sensitizing effects of nicotine. Many of the long-term teratogenic effects that result from nicotine exposure during adolescence also differ in males and females; in particular, the loss of hippocampal cells resulting from adolescent nicotine exposure has been shown to occur only in females. Such findings suggest that there are sex differences in factors that may be critical for initiating and maintaining smoking.

Nicotine Interaction with Stress Circuitry

The central and peripheral stress systems are critical for adaptation to environmental change, with their activation resulting in improved chances of survival for the individual. Although these adaptations are normally positive and time limited, the central stress system has been heavily implicated in addictive processes, and there is a high correlation between stress-related disorders and drug abuse. Both a stressful environment and subsequent activation of the hypothalamic–pituitary–adrenal (HPA) axis facilitates the acquisition and reinstatement of use of addictive drugs, including nicotine and tobacco. Indeed, recruitment of brain stress systems has been proposed to provide the negative motivational state that drives addiction, and produces the vulnerability toward relapse (Figure 1).

The physiological stress response system consists of two components – glucocorticoid release from the adrenal cortex via activation of the HPA axis and catecholamine release from the adrenal medulla via activation of the splanchnic nerve (Figure 1). Nicotinic acetylcholine receptors (nAChRs) play a critical physiological role in both systems. Activation of the HPA axis, and subsequent glucocorticoid release, are mediated by nAChRs on noradrenergic brain stem cell bodies within the nucleus tractus solitarius (NTS); these provide excitatory innervation to the hypothalamic paraventricular nucleus (PVN), the site of corticotropin-releasing factor (CRF) neurons that are crucial for the release of adrenocorticotropic hormone (ACTH). Excitatory nAChRs on chromaffin cells of the adrenal medulla also regulate circulating levels of catecholamines. Nicotine can increase both plasma glucocorticoid and catecholamine levels via activation of nAChRs in the NTS and adrenal medulla, respectively. However, animal studies have shown that nicotine-induced activation of the two components of the stress system is dependent on the contingency of drug delivery; whereas experimenter-administered nicotine activates both the adrenal medulla and the HPA axis, self-administered nicotine activates only the latter.

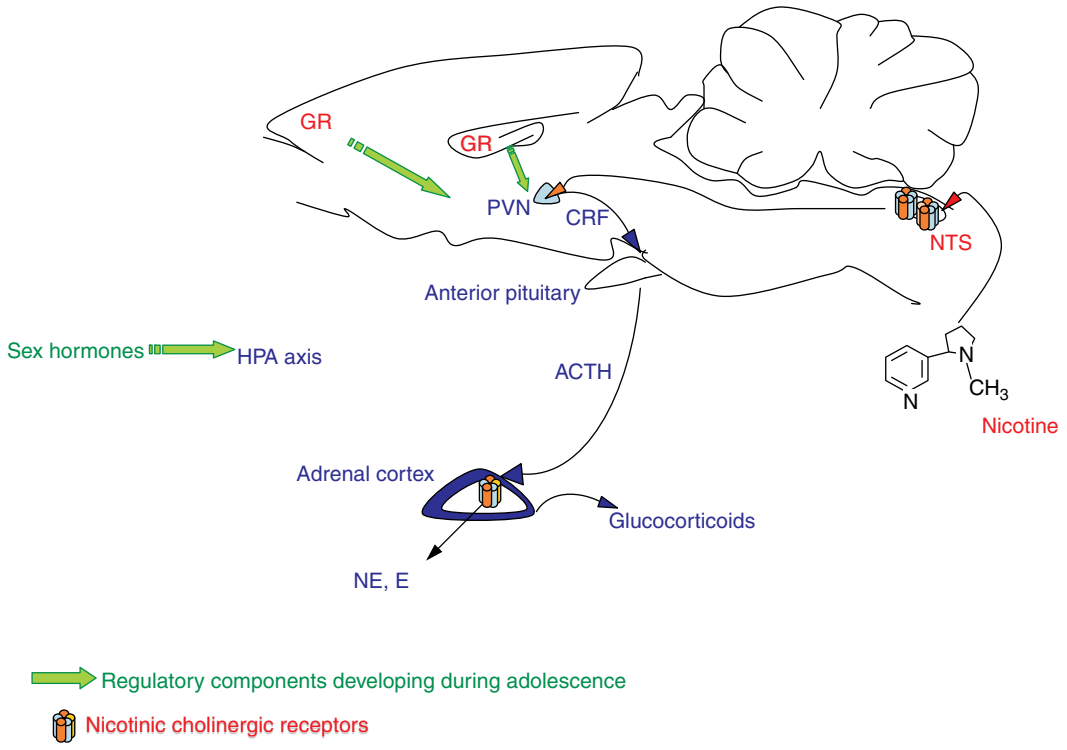

Figure 1 Schematic representation of the central and peripheral stress systems and their regulation by nicotine acting on nicotinic cholinergic receptors. ACTH, adrenocorticotropic hormone; CRF, corticotropin-releasing factor; E, epinephrine; GR, glucocorticoid receptors; NE, norepinephrine; NTS, nucleus tractus solitarius; PVN, paraventricular nucleus of the hypothalamus.

HPA axis regulation in adult rodents is sex specific, with males responding to acute nicotine injection with higher vasopressin release, whereas females respond with greater ACTH and glucocorticoid release. This finding is consistent with earlier reports from both clinical and preclinical studies that females respond to stressors with higher levels of ACTH and glucocorticoids than males do. This reflects an enhancement of HPA activation by estradiol and inhibition by testosterone. Although nicotine produces activation of the HPA axis when administered acutely, this effect diminishes over time with chronic exposure. Again, however, this decrease in the HPA-activating effects of chronic nicotine is sex specific, with some evidence from both clinical and preclinical studies that there may be less tolerance to chronic nicotine in females.

The interaction of nicotine and the stress system is dynamic and bidirectional. Whereas nicotine activates the HPA axis, glucocorticoids also modulate the physiological and behavioral effects of nicotine in a complex manner. Adrenalectomy potentiates a number of responses to nicotine, including heart rate, body temperature, and antinociception; such changes in nicotine sensitivity are prevented by glucocorticoid replacement. However, other effects of nicotine, such as locomotor activation, are attenuated by adrenalectomy. Glucocorticoids have also been implicated in the neural plasticity that results from repeated nicotine exposure, and which may contribute to tolerance and dependence. Whereas many studies have suggested that HPA activation promotes a form of learned tolerance to nicotine, glucocorticoids have also been shown to be necessary for the development of nicotine-induced locomotor sensitization, a form of behavioral plasticity that is implicated in addiction. There is some evidence to suggest that environmental context may be a critical determinant of whether tolerance or sensitization occurs. Further studies are needed to elucidate the underlying mechanisms.

The clinical literature indicates that there is a wide variance in individual response to the effects of tobacco, and that this may correlate, to some extent, with individual differences in stress reactivity. Preclinical studies have also shown that individual differences in sensitivity to both the rewarding and aversive effects of nicotine may reflect variations in HPA responsiveness to nicotine (Figure 2). In adult mice, there are strain differences in nicotine-induced glucocorticoid secretion, with the C57BL6 strain showing the lowest responsiveness. Conversely, this mouse strain exhibits the greatest propensity to self-administer nicotine, an effect which has been correlated with decreased sensitivity to the aversive effects of the drug. Similarly, Lewis and Fisher F344 inbred strains of rat exhibit substantial differences in HPA response to nicotine, with the F344 strain showing an enhanced drug-induced glucocorticoid secretion as compared to the Lewis rat. The Lewis rat is much more sensitive to the rewarding effect of nicotine, as determined in self-administration and conditioned place preference tests, and is less sensitive to nicotine-induced conditioned taste aversion. Lewis rats are also more sensitive to the discriminative stimulus properties of nicotine, and, unlike F344 rats, exhibit a substantial antagonist-precipitated withdrawal syndrome after chronic nicotine treatment. Although intriguing, these findings of an inverse association between nicotine reward and HPA activation are purely correlational. More direct experimental evidence will be required to determine whether this association reflects 'cause and effect.'

In this context it is of interest to note that, in clinical studies, attenuated HPA reactivity is associated with increased withdrawal symptoms. Cortisol levels are higher in smokers than in nonsmokers, and the HPA response to psychosocial stress is significantly blunted in chronic smokers of either sex. Upon quitting, there is a rapid and sustained decline in plasma cortisol levels, the magnitude of which is

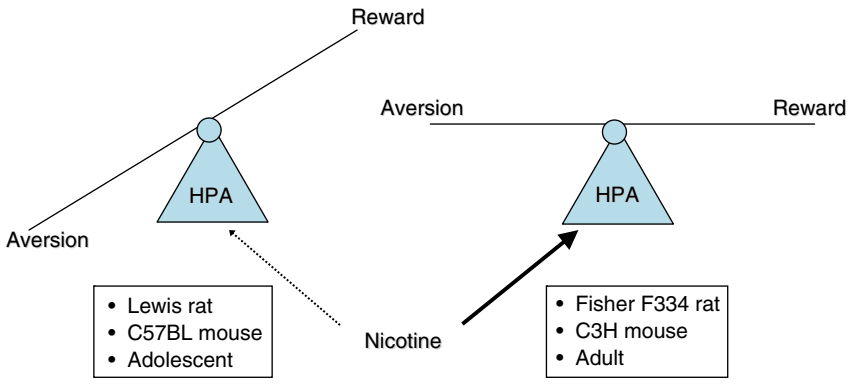

Figure 2 Schematic model of the possible interrelationship between nicotine-induced hypothalamic–pituitary–adrenal (HPA) activation and the balance of rewarding and aversive effects of the drug.

associated with subsequent relapse. Significant individual variation is evident in withdrawal-induced changes in HPA response, with those who relapse quickly showing the most attenuated HPA responses to acute stress and the most exaggerated withdrawal symptoms. Thus, decreased stress response may predict increased vulnerability for smoking relapse. Further studies need to be undertaken to determine whether this absence of nicotine-induced HPA activation underlies the reduced aversive properties of the drug during adolescence.

A number of different mechanisms may underlie the influence of glucocorticoids on nicotine or tobacco-induced physiological and behavioral changes. Glucocorticoids have been widely shown to regulate the neural systems which mediate addiction, including dopamine pathways, amygdala, and prefrontal cortex. In this respect, the effects of stress are generalizable to many drugs of abuse and are not specific to nicotine or tobacco. Glucocorticoids also have some selective effects on nicotine signaling, however, via modulation of nicotinic receptors. Corticosterone not only acts as an acute allosteric modulator of nACHRs, but also induces downregulation of some nicotinic receptor types, particularly the $\alpha 7$ receptor, which has been implicated in brain plasticity and cognitive function.

Adolescent Development of Stress Circuitry

Whereas the HPA axis is the final common pathway for glucocorticoid secretion, there is an extensive network of higher order neural circuits that regulate stress hormone secretion. Adolescence is a period of active remodeling of many of the brain regions which modulate the HPA response, including prefrontal cortex, hippocampus, and amygdala. Remodeling of the adolescent limbic system involves the same proliferative and regressive events that are the hallmarks of earlier periods of brain development, including neurogenesis, apoptosis, axonal sprouting, synapse formation and elimination, and myelination. In addition to these morphological changes, there are dynamic alterations in monoamine inputs that regulate higher order stress circuits. Such neuronal changes occur in a regionally and time-dependent manner and, although they reflect an innate developmental program, they are actively modified by hormonal changes and environmental experience.

Adolescence is also characterized by substantial changes in circulating steroid hormone levels. Puberty results in elevated levels of sex steroids, which produce the peripheral manifestations of pubertal growth and development and which also have marked effects on brain function. Gonadal steroid actions during adolescence play critical roles in the maturation of brain circuitry underlying adult sexual behavior and modulate emotional responding and social behaviors, as well as stress responses. In contrast to the sex steroids, adolescence is not marked by dramatic alterations in corticosteroid secretion. As with adults, there is a normal diurnal variation in glucocorticoid secretion, with highest levels upon waking. However, upon closer examination, there are significant age differences in the regulation of adolescent glucocorticoid levels.

In clinical studies in which diurnal cortisol secretion patterns in adolescents have been observed in naturalistic settings, an age-dependent increase in basal cortisol levels has been observed. Although basal cortisol levels are elevated with increasing age, there is also a flattening of diurnal rhythms, with a smaller increase in hormonal secretion upon waking. In contrast to adults, cortisol levels are also elevated when teenagers are alone, suggesting that social interaction is a critical regulator of the adolescent stress response. Preclinical studies have also shown age differences in the regulation of the HPA axis. One common finding is that there is decreased negative feedback control of glucocorticoid secretion in adolescents, which results in a prolonged HPA activation in response to an acute stress, as compared to adults. There are also substantial age differences in response to chronic stress. Whereas repeated exposure to a stressor results in a habituation of the peak corticosterone response in adults, this is not the case in adolescents; however, the duration of the stress response following chronic stress in adolescents is much shorter than in adults. Since experience-dependent glucocorticoid secretion can induce substantial changes in brain structure and function, these age differences in HPA response may have important functional consequences. A prolonged release of glucocorticoids in an adolescent after a single stressful experience may facilitate learning and memory of the event. In contrast, the more rapid decline of hormonal responses following chronic stress in the young may protect them from the deleterious effects of chronically high levels of glucocorticoid exposure.

There is an important interaction between pubertal age and stress in the regulation of both the HPA and the hypothalamic–pituitary–gonadal (HPG) axes. Adolescent and adult rats exhibit strikingly different patterns of regulation of both gonadal and adrenocortical systems in response to chronic stress. It has been proposed that, when they are stressed, adolescents protect body growth at the expense of sexual function, whereas the opposite is true in adults. Evidence for this comes from studies in which stress has been shown to decrease testosterone secretion in

adolescent rats, but to increase it in adults; conversely, adolescent rats maintain body weight following repeated stress, whereas adults do not. Treatment of male rats with the testosterone receptor antagonist flutamide also alters ACTH responses to stress in an age-dependent manner; whereas no drug effect is observed peripubertal rats, flutamide increases ACTH response to repeated stress in adults. In addition, there are striking age differences in the interaction of flutamide and stress on the numbers of vasopressin and CRF neurons within various components of the central stress circuitry. Sex differences in both hormonal and neuronal responses to stress also emerge at puberty, with females becoming hyperresponsive in comparison to both adult males and adolescent females. Such findings add to a growing body of literature that suggests that there is a major reorganization of central stress circuitry during puberty, a process in which sex hormones have influential roles.

Given the substantial remodeling of stress circuitry that occurs during adolescence, it is not surprising that various psychopathologies associated with limbic dysfunction emerge during this developmental period. Many cognitive and emotional disorders, such as substance abuse, eating disorders, and schizophrenia, are rarely seen before adolescence. Others, such as major depression, anxiety, and bipolar illness, greatly increase in prevalence during the teenage years. There is substantial comorbidity between tobacco use and many psychiatric disorders, including anxiety, depression, and schizophrenia, which may reflect either common neural substrates or attempts at symptomatic relief through self-medication. However, in at least one report, there is evidence that tobacco use during early adolescence contributes to the etiology of panic attacks. These clinical data are consistent with the finding of preclinical studies that chronic nicotine treatment in adolescent rodents increases subsequent anxiety in adults. Thus, nicotine exposure during the sensitive period of adolescence may interfere with the remodeling of central stress circuitry.

Nicotine Interaction with Adolescent Stress Circuitry

A number of factors, including stress, have been identified as contributing to the uptake of smoking in teenagers. In longitudinal clinical studies, both higher basal HPA activity and decreased anticipatory response to stress have been associated with the subsequent initiation of smoking in adolescents. In preclinical studies, differences in adolescent stress responses can be induced by differential handling and housing conditions during the early postnatal period. An interaction between acute nicotine treatment and early handling has been found, such that those adolescents with a higher basal HPA activity tend to exhibit a decrease in corticosterone levels in response to nicotine. Interestingly, acute intravenous nicotine administration, which models the route of delivery in drug self-administration studies, and which most closely models human smoking, does not induce the same profound glucocorticoid release in adolescent rats as it does in adult rats. Thus, HPA axis response to acute nicotine exposure is quite different during the adolescent sensitive period than it is in adults (Figure 2). Although the exact mechanism of this age difference is not yet understood, anatomical studies have shown age differences in nicotine activation of immediate-early gene expression in the hypothalamic PVN.

In contrast to the absence of an acute drug effect on adolescent glucocorticoid release, chronic nicotine treatment during adolescence induces a long-term adaptation in HPA responding. Chronic adolescent nicotine exposure has been reported to increase basal ACTH and corticosterone levels in a dose-dependent manner, and to increase levels of CRF peptide immunoreactivity in both the hypothalamus and prefrontal cortex. Behavioral correlates of this elevated HPA activity include increased locomotor responding in a novel environment, increased anxiety, and increased depression as measured on a forced swim test. Social factors play an important role, however, both in the adolescent reorganization of central stress circuitry and in the effects of nicotine on adults. Social stress during adolescence influences locomotor sensitization to nicotine and nicotine-induced activation of the HPA axis in a sex-dependent manner. This finding of a critical influence of social factors during adolescence on subsequent drug sensitivity has important implications for the design and interpretation of both clinical and preclinical studies. Social factors must be carefully controlled for in order to reduce interpretational confounds.

Conclusions

Both clinical and preclinical studies have shown that nicotine activates the HPA axis in adults. As with other stressors, females are hypersensitive to the HPA-activating effects of the drug. In contrast to adults, both male and female adolescents exhibit minimal HPA activation in response to nicotine, which may influence the balance of rewarding and aversive effects of the drug (Figure 2). Thus, decreased activation of stress circuitry may be one mechanism that underlies the increased rewarding and decreased aversive effect of nicotine that has been widely seen in preclinical studies of adolescents.

Adolescence is a period of dynamic plasticity and reorganization of central stress circuitry. During this sensitive period, hormonal and environmental influences can have profound and long-term influences on stress responding. Nicotine exposure during this sensitive period results in long-term alterations of central stress pathways, with resulting adaptations in both endocrine and behavioral response. Such drug-induced adaptations in HPA regulation may underlie the long-term addictive effect of smoking in those who begin during early adolescence.

See also: Chronic (Repeated) Stress: Consequences, Adaptations; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Stress and Cognition; Stress Response: Sex Differences; Stress, Dopamine, and Puberty; Stress: Definition and History.

Further Reading

Adam EK (2006) Transactions among adolescent trait and state emotion and diurnal and momentary cortisol activity in naturalistic settings. *Psychoneuroendocrinology* 31: 664–679.

Barron S, White A, Swartzwelder HS, et al. (2005) Adolescent vulnerabilities to chronic alcohol or nicotine exposure: Findings from rodent models. *Alcoholism: Clinical and Experimental Research* 29: 1720–1725.

Caggiula AR, Donny EC, Epstein LH, et al. (1998) The role of corticosteroids in nicotine's physiological and behavioral effects. *Psychoneuroendocrinology* 23: 143–159.

Cruz FC, Delucia R, and Planeta CS (2005) Differential behavioral and neuroendocrine effects of repeated nicotine in adolescent and adult rats. *Pharmacology, Biochemistry and Behavior* 80: 411–417.

Jamner LD, Whalen CK, Loughlin SE, et al. (2003) Tobacco use across the formative years: A roadmap to developmental vulnerabilities. *Nicotine and Tobacco Research* 5(supplement 1): S71–S78.

Koob GF (1999) Stress, corticotropin-releasing factor, and drug addiction. *Annals of the New York Academy of Science* 897: 27–45.

Leslie FM, Loughlin SE, Wang R, et al. (2004) Adolescent development of forebrain stimulant responsiveness: Insights from animal studies. *Annals of the New York Academy of Science* 1021: 148–159.

Lovallo WR (ed.) (2006) Cortisol and the addictions. *International Journal of Psychophysiology* 59: 193–264.

Matta SG, Fu Y, Valentine JD, et al. (1998) Response of the hypothalamo-pituitary-adrenal axis to nicotine. *Psychoneuroendocrinology* 23: 3–13.

Pechmann C, Levine L, Loughlin S, et al. (2005) Self-conscious and impulsive: Adolescents' vulnerability to advertising and promotion. *Journal of Public Policy & Marketing* 24: 202–221.

Romeo RD, Bellani R, Karatsoreos IN, et al. (2006) Stress history and pubertal development interact to shape hypothalamic–pituitary–adrenal axis plasticity. *Endocrinology* 147: 1664–1674.

Sisk CL and Zehr JL (2005) Pubertal hormones organize the adolescent brain and behavior. *Frontiers in Neuroendocrinology* 26: 163–174.

Slotkin TA (2002) Nicotine and the adolescent brain: Insights from an animal model. *Neurotoxicology and Teratology* 24: 369–384.

Spear LP (2000) The adolescent brain and age-related behavioral manifestations. *Neuroscience & Biobehavioral Reviews* 24: 417–463.

Viau V, Bingham B, Davis J, et al. (2005) Gender and puberty interact on the stress-induced activation of parvocellular neurosecretory neurons and corticotropin-releasing hormone messenger ribonucleic acid expression in the rat. *Endocrinology* 146: 137–146.

Relevant Website

http://www.nida.nih.gov – National Institute on Drug Abuse.

Stress, the HPA Axis and Depressive Illness

E Young, Molecular and Behavioral Science Institute, Ann Arbor, MI, USA
A Korszun, Barts and The London School of Medicine, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Depression can be conceptualized as a maladaptive, exaggerated response to stress. Stressful life events, particularly those related to loss, such as divorce and bereavement, have a strong causal relationship with depressive episodes. However, not all people who experience such events develop depression, and an individual's sensitivity to stress depends to a large extent on his or her genetic makeup. There is also compelling evidence that early life stress constitutes a major risk factor for depression in adult life, and thus the combination of genetic predisposition, early life stress, and ongoing stress determines an individual's susceptibility to developing depression. The main component of the physiological stress response is the stress hormone axis, the hypothalamic–pituitary–adrenal (HPA) axis, and it is now well established that abnormalities of the HPA axis play a major causal role in the development of depression.

HPA Axis Regulation

The HPA axis transforms stressful input into hormonal messages enabling the organism to adapt to environmental change and to maintain the body's homeostasis. Corticotropin-releasing hormone (CRH) is synthesized in the paraventricular nucleus (PVN) of the hypothalamus and secreted into the hypophyseal portal system via the median eminence. The stimulus for CRH secretion can be either 'physical' (e.g., exercise or starvation) or 'psychological' (e.g., perceived danger or stressful life events). The HPA axis is closely linked to the autonomic nervous system, and brain stem catecholamine systems are critical circuits involved in the 'activation' pathways for CRH release. The release of CRH stimulates secretion of pituitary adrenocorticotropic hormone (ACTH) which, in turn, results in the secretion of glucocorticoids by the adrenal cortex in a feed-forward cascade. Cortisol is the main glucocorticoid, and its secretion is tightly controlled by negative feedback effects of glucocorticoids at both pituitary and brain sites. Negative feedback on secretion can occur very rapidly, within 5–10 min, and this provides real-time inhibition to limit the stress response and prevents oversecretion of glucocorticoids. In addition to these rapid effects on secretion, glucocorticoids exert negative feedback effects on mRNA and subsequent protein stores for both CRH and the ACTH precursor, proopiomelanocortin (POMC). Regulation of gene transcription can occur directly via glucocorticoid responsive elements, as has been shown for POMC, or by transsynaptic influences and 'inhibition' circuits. There is strong evidence that these transsynaptic inhibition circuits are critical in regulating CRH mRNA levels in the PVN.

Stressful stimuli activate all levels of the HPA axis, causing increases in CRH, ACTH, and cortisol secretion. However, these increases are superimposed on an intrinsic circadian pattern of HPA activity driven by the suprachiasmatic nucleus (SCN). ACTH secretion is pulsatile in nature, with the trough of integrated secretion occurring in the evening and early night and the peak of secretion occurring just before awakening; active secretion continues through the morning and early afternoon. Even in the absence of corticosteroid feedback (e.g., adrenalectomy), ACTH continues to demonstrate a circadian rhythm which is believed to be secondary to the circadian rhythm of CRH production and results in a similar circadian pattern of cortisol secretion. There is evidence that intrinsic neural elements are responsible for both the initiation and the inhibition of the CRH/ACTH/cortisol circadian rhythm and that glucocorticoids act to dampen the overall amount of secretion. It is possible that glucocorticoids interact with neural elements to alter rhythm characteristics indirectly, although they are not themselves directly responsible for the inhibition of PVN CRH secretion during the quiescent phase.

HPA Axis Abnormalities in Depression

Overactivity of the HPA axis as manifested by an increase in cortisol secretion is a well-established phenomenon in depression. The original studies of Sachar and colleagues showed that up to 50% of depressed patients have higher mean plasma cortisol concentrations and an increased number and duration of cortisol secretory episodes, suggesting increased cortisol secretory activity. Numerous later studies have validated hypercortisolemia in depression; however, although it was expected that the increased cortisol would be accompanied by an increased level of plasma ACTH, this has proved more difficult to demonstrate. There are also conflicting reports of alterations in sensitivity to ACTH 1–24 in depressed patients. Also, it is unclear whether the hypercortisolemia of depression is due to excessive drive with increased ACTH secretion or to enhanced adrenal sensitivity to ACTH, although the balance of

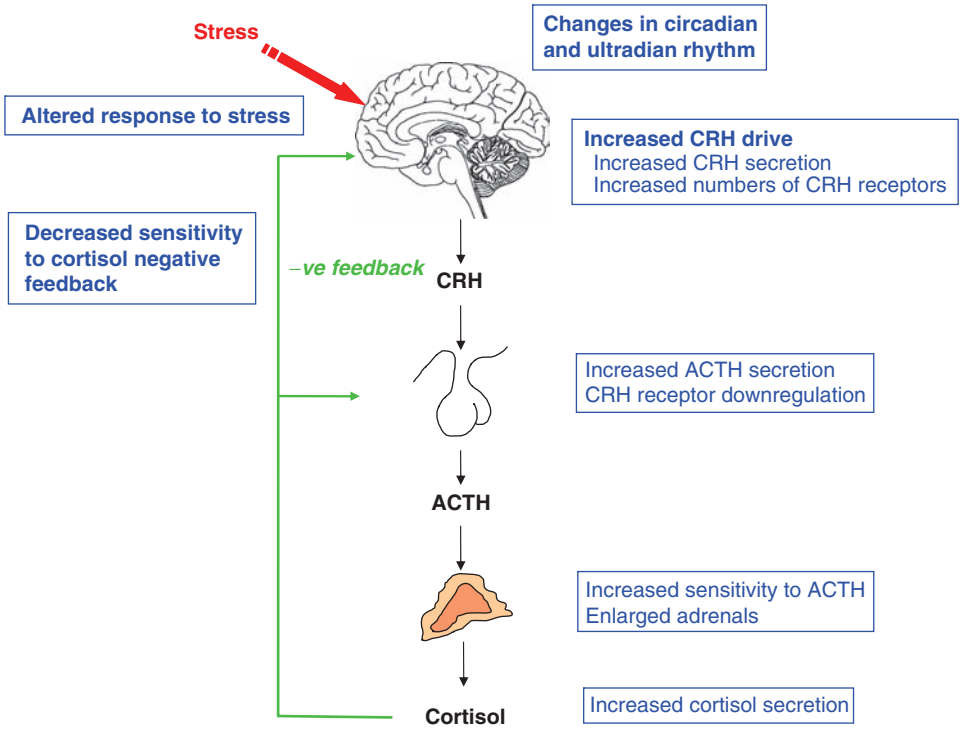

Figure 1 HPA axis abnormalities in depression.

evidence points strongly to the brain as the source of HPA axis hyperactivity.

Enlarged adrenals and increased sensitivity to ACTH are both common consequences of chronic adrenal stimulation by ACTH and, thus, stress or depression-associated enhanced adrenal sensitivity may be secondary to increased central drive. Following the development of hypercortisolemia, feedback mechanisms may be engaged that attempt to limit the degree of excess cortisol secretion by decreasing ACTH secretion. It is possible that more than one pattern of abnormality exists, that is, increased ACTH secretion accompanied by increased cortisol secretion, and normal ACTH secretion accompanied by increased cortisol secretion. In both cases, glucocorticoid feedback acts to restrain CRH and ACTH secretion, thus blunting the true degree of CRH overdrive in depression, analogous to the blunted response to exogenous CRH administration seen in depression.

Thus, HPA overactivity in depression could be due to one or a combination of factors: increased CRH drive, decreased sensitivity to glucocorticoid feedback, altered secretory rhythms, and altered response to stress (Figure 1).

Increased CRH Drive

Early investigators reporting hypercortisolemia in depression hypothesized that it resulted from an increased central CRH secretion. The subsequent isolation and sequencing of CRH allowed further studies, including evaluation of both behavioral and endocrine responses to exogenous CRH, as well as radioimmunoassay measurement of CRH in several brain areas in chronically stressed animals and in postmortem samples from depressed patients. It had been noted earlier that central CRH administration produces behavioral effects similar to the symptoms of depression, including anorexia, weight loss, motor hyperactivity in familiar environments, freezing behavior in novel environments, increased awakenings during sleep, and decreased libido. Endocrine challenge studies with exogenous CRH administration in depressed patients demonstrated a blunted ACTH response to CRH, a finding compatible with CRH receptor downregulation. However, the interpretation of these studies is complicated by the higher baseline cortisol, which inhibits responsiveness to exogenous CRH. Strategies to equalize the degree of negative feedback between patients and controls have used metyrapone or dexamethasone, to inhibit cortisol production, plus CRH challenge. Because metyrapone stimulates the HPA axis differentially in patients and controls, this adds an additional confounding factor. However, studies using the administration of dexamethasone followed by CRH challenge show greater ACTH and cortisol responses to CRH, consistent with a CRH overdrive hypothesis.

Some studies examining CRH in the cerebrospinal fluid (CSF) of depressed patients and normal controls have also found increased CRH in the CSF in

depressed patients. Furthermore, increased POMC, the ACTH precursor, has been demonstrated in suicide victims with an indication of oversecretion of CRH from the PVN, although it should be noted that suicide victims are not necessarily all depressed. Raadsheer and colleagues reported increased CRH and increased numbers of CRH neurons in the PVN in postmortem samples from depressed patients; both findings are compatible with increased CRH drive. In addition, decreased CRH receptor binding in a number of brain regions has also been observed in suicide victims, again compatible with the hypothesis of increased CRH release. Interestingly, several basic science studies have also demonstrated that early trauma may lead to lifelong changes in stress responsiveness and CRH content in hypothalamic and extrahypothalamic brain nuclei.

Studies of metyrapone in major depression support the presence of increased central nervous system (CNS) drive in depression, at least in the evening, and it appears that there is increased CRF/ACTH secretion, which is likely amplified at the adrenal to lead to increased cortisol. These changes in cortisol secretion are commonly considered to be 'state' changes and resolve when the depression resolves. However, almost all studies of the HPA axis in major depression in euthymic subjects have examined patients on tricyclic antidepressants, which exert direct effects on the HPA axis. Three studies of epidemiological samples from the United States and one from the United Kingdom found that salivary cortisol is increased in subjects with lifetime major depression, most of whom had no current mood symptoms.

Despite this evidence, only approximately 30–50% of depressed patients manifest hypercortisolemia. Thus, although CRH overdrive in depression remains an attractive hypothesis, it has not been conclusively demonstrated. This may in part be due to compensatory increases in cortisol acting to turn off both CRH drive and ACTH secretion so that the true extent of CRH hypersecretion is obscured by the effects of negative feedback.

Decreased Sensitivity to Glucocorticoid Feedback

The failure of some depressed patients to suppress cortisol after dexamethasone is a well-accepted phenomenon, but its pathophysiology is not clear. Although an insensitivity to glucocorticoid negative feedback is the hypothesized defect, studies suggest that relatively low plasma dexamethasone concentrations may contribute to the failure to suppress cortisol. Although plasma dexamethasone concentrations may explain part of the variance, some individuals who do not suppress cortisol show adequate plasma dexamethasone concentrations. Also, because dexamethasone is presumably acting through intermediate and delayed feedback mechanisms, the concentration of dexamethasone bound to the receptors in the CNS and pituitary feedback sites may actually be the relevant physiological parameter, rather than the dexamethasone plasma concentration.

It was commonly assumed that cortisol nonsuppression reflected pituitary nonsuppression of ACTH. The majority of studies tested this hypothesis by comparing plasma ACTH values between cortisol suppressors and cortisol nonsuppressors. In most studies, the plasma ACTH measures were taken in the morning (8 a.m.), whereas the cortisol samples were drawn at 4 p.m.; the reason for this discrepancy in sample timing is not clear. The majority of these studies concluded that ACTH concentrations are higher in nonsuppressors than suppressors. Subsequent studies have demonstrated that depression is accompanied by an increased cortisol secretory response following ACTH 1-24 infusion and, consequently, cortisol nonsuppression to dexamethasone is not necessarily linked to ACTH hypersecretion. Because cortisol nonsuppression is defined by a 'cut-off' criterion, rather than a decrease from baseline secretion, it is possible that ACTH hypersecretion may not be accompanied by cortisol nonsuppression. Thus, cortisol hypersecretion or ACTH hypersecretion or both could be involved in the HPA axis disturbances demonstrated by the dexamethasone suppression test.

Evidence from several studies shows that more depressed patients demonstrate HPA axis dysregulation at the pituitary corticotroph level than at the adrenal level, and that cortisol plays an important role in regulating changes in negative feedback at the pituitary. These findings are in agreement with data from rats demonstrating that age and elevated glucocorticoid concentrations interact to lead to alterations in glucocorticoid negative feedback that is associated with decreases in hippocampal glucocorticoid receptors. This nonsuppression is associated with higher baseline cortisol, particularly in the afternoon, and advancing age or menopausal status in women. Even small elevations in cortisol concentrations can lead to alterations in negative feedback, and decreased sensitivity to negative feedback occurs in more than half of depressed patients and can be demonstrated above the level of the adrenal.

Altered Secretory Rhythms

Circadian dysregulation hypothesis Although the HPA axis is conceptualized as the 'stress' hormone axis, in the absence of stressful stimuli, its secretion is driven by the circadian oscillator – the SCN. In practice, it is difficult to distinguish between secretion

occurring as a response to stress and secretion driven by the intrinsic circadian pacemaker: disruption of either of these major inputs may affect the functioning of the other. Previous observations in depression of diurnal variation in mood and a seasonal pattern suggested abnormalities in the circadian oscillator and disruption in the circadian rhythm of cortisol in depressed patients. Sachar's early report of a phase advance in the circadian rhythm of cortisol was further supported by observations of early morning awakening and the altered distribution of rapid eye movement (REM) sleep (shortened REM latency and increased REM density in the first half of the sleep period), suggesting a phase advance of multiple circadian systems in depression. However, a newer understanding of circadian physiology revealed that studies carried out under normal day/night conditions do not take into account masking of the endogenous circadian rhythm by entrainment to multiple environmental Zeitgebers. Also, introduction of the constant routine conditions that are necessary for accurate endogenous pacemaker assessment leads to sleep deprivation that frequently improves depression, thereby making it difficult to evaluate completely the circadian dysregulation hypothesis of depression.

Later studies testing the phase advance hypothesis of cortisol secretion produced ambiguous results. Two studies failed to find evidence for phase advance of the cortisol rhythm in depressed subjects and normal controls undergoing 24 h cortisol sampling, but they did observe an earlier onset of the first cortisol secretory episode in depressed subjects compared to normal controls. Another study, using age, sex, race, and menstrual status-matched controls for depressed subjects, observed no change in the circadian rhythm or in the onset of the first secretory episode. Also, in these same subjects melatonin onset and offset times did not differ between depressed patients and controls. However, these studies were conducted under ambient room light rather than dim light, so light-induced suppression of melatonin onset may have occurred. In addition, in several of these studies, hypercortisolemic depressed patients were older than eucortisolemic depressed patients, suggesting a possible interaction between age and HPA axis dysfunction.

A meta-analysis by van Cauter and co-workers of the pooled cortisol data from the controls in these and additional studies demonstrated clear effects of age on the circadian rhythm of cortisol secretion, with an earlier onset of secretion in older subjects. This suggests that the absence of strict age matching of controls to depressed patients, which was the case in all but one of the studies, skews the analysis of circadian rhythm. In another meta-analysis of the depressed patient cortisol data from all these studies, van Cauter found clear evidence for an earlier onset of secretion but no evidence of a change in the timing of the acrophase or nadir. Taken together, these studies suggest that a shift or phase advance in the whole circadian rhythm of cortisol does not occur in depressed patients, but that the onset of circadian drive may begin earlier in these patients. To complicate matters further, studies demonstrate that partial sleep deprivation can lead to an increase in cortisol secretion during the nadir, raising the question of whether the reported earlier onset of secretion observed in depressed patients may be secondary to sleep disturbances. The earlier onset of cortisol secretion could also be explained by a change in set point for cortisol in the CNS of depressed patients that would result in relative cortisol insufficiency in the late evening and subsequent activation of the SCN-CRH drive at an earlier time in the circadian period.

Studies involving evening administration of metyrapone, a cortisol synthesis inhibitor, have found that blockade of cortisol in the late afternoon and evening (4–10 p.m.) leads to greater rebound in ACTH secretion in patients with major depression compared to matched control subjects. Whereas these data may suggest increased central CRH drive, studies conducted with metyrapone in the morning and early afternoon (8 a.m. to 2 p.m.) found a normal ACTH response to metyrapone in patients with major depression. This suggests that CRH drive in the evening is different from that in the morning and may represent an earlier onset of the active phase of secretion in the presence of diminished negative feedback. Longer-term studies with metyrapone blockade are necessary to determine if this finding supports a change in set point or an earlier onset of secretion.

Ultradian rhythms and pulsatile secretion In addition to a 24 h circadian rhythm in the HPA axis hormones, ultradian rhythms of ACTH and cortisol have been identified, most frequently 90–110 min. A similar rhythm has been demonstrated for ACTH in rats that disappears with CRH immunoneutralization, linking this rhythm to the CNS. Other studies have found a coupling between cortisol secretory activity and EEG ultradian rhythms during sleep, with a significant relationship between daytime EEG β activity and cortisol secretory activity. The frequency of 110 min for β activity and 117 min for cortisol, with a lag time of 10 min for cortisol by cross-correlation, suggests that this ultradian rhythm of cortisol is linked to central alertness/arousal mechanisms, and our own data demonstrate a strong ultradian rhythm of 120 min. In addition, a number of studies have found that the onset of sleep, during

both night and daytime, results in a decrease in cortisol secretion, and failure to inhibit cortisol secretion at sleep onset has been reported in depressed patients. Following identification of the pulsatile nature of ACTH and cortisol secretion, emphasis has shifted to identifying pulses and to performing time domain-based analyses rather than frequency domain-based analyses. Less is known about the physiology of pulse frequency in the HPA axis than in other hormone systems such as the reproductive axis. Of the four studies that have examined cortisol pulsatility in depression, there is only one report of an increased number of pulses in depressed patients. The overall evidence suggests that increased cortisol secretion is not due to increased pulse number but, rather, to either increased amplitude or increased baseline cortisol not secreted in a pulsatile manner.

Altered Response to Stress

Increased cortisol is found in only a proportion of patients, and there is evidence that there may be two different types of depression. One is characterized by increased basal circadian drive, decreased negative feedback, and normal stress reactivity (classic depression). The other is associated with early trauma and characterized by enhanced stress reactivity and enhanced negative feedback resulting in normal to low cortisol secretion.

The National Comorbidity Study found that childhood adversities were consistently associated with the onset (but not necessarily persistence) of mood disorders. Data from this survey indicated a significant association between childhood sexual abuse and mood disorders, and childhood adversity increased the risk for recent episodes of major depression. Another large population-based study found that multiple childhood adverse experiences increased the risk for lifetime depression disorders (odds ratio, 2.7), and there was a dose relationship between the cumulative events and probability of lifetime and recent depressive disorders, even after controlling for a mentally ill household member. A Finnish study also found a similar cumulative effect of adult traumas, with odds ratios for persistent depression of 6.0 for men and 6.99 for women with three or more traumatic experiences.

Childhood sexual abuse has also been found to sensitize women to the depressogenic effects of subsequent stressful life events and to be associated with an increased likelihood of suicide or self-harm attempts in depressed women. Heim and colleagues examined multiple HPA axis function in depressed subjects with or without a history of childhood abuse. They found an effect of early abuse and depression on stress reactivity, with increased ACTH and cortisol responses to the stressor compared to either controls or depressed patients without childhood abuse. In these same subjects, they found a blunted response to CRH challenge in depressed patients with or without childhood abuse but an increased response to CRH in abused individuals without depression. The abused subjects also showed a blunted cortisol response to ACTH 1–24. Thus, childhood abuse resulted in an "enhanced pituitary response with counter-regulatory adrenal adaptations" – changes compatible with low or normal basal cortisol. Furthermore, they found lower cortisol and enhanced feedback to low-dose dexamethasone in the same subjects. The abused groups also demonstrated lower baseline and stimulated cortisol in response to CRH challenge as well as following ACTH infusion. These same groups of women showed a significantly greater HPA axis response to the Trier Stress Test despite smaller responses to CRH.

See also: Adrenal Steroids: Biphasic Effects on Neurons; Depression and the Brain; Diabetes Type 2 and Stress: Impact on Memory and the Hippocampus; Gene Therapy and Protection from Stress-Induced Brain Damage; Hypothalamic–Pituitary–Adrenal (HPA) Axis; Social Stress in Adult Primates; Stress and Neural Involvement in Metabolism; Stress and Cognition; Stress and Suicide; Stress and Vulnerability to Brain Damage; Stress Response: Neural and Feedback Regulation of the HPA Axis; Stress, Sex and Adolescent Nicotine Response; Stress, Dopamine, and Puberty; Stress, Cytokines and Depressive Illness; Stress: Definition and History.

Further Reading

Barden N (2004) Implication of the hypothalamic–pituitary–adrenal axis in the physiopathology of depression. *Journal of Psychiatry & Neuroscience* 29: 185–193.

Gillespie CF and Nemeroff CB (2005) Hypercortisolemia and depression. *Psychosomatic Medicine* 67(supplement 1): S26–S28.

Gold PW and Chrousos GP (2002) Organization of the stress system and its dysregulation in melancholic and atypical depression: High vs low CRH/NE states. *Molecular Psychiatry* 7: 254–275.

Keck ME, Ohl F, Holsboer F, and Muller MB (2005) Listening to mutant mice: A spotlight on the role of CRF/CRF receptor systems in affective disorders. *Neuroscience & Biobehavioral Reviews* 29: 867–889.

Nemeroff CB (2004) Neurobiological consequences of childhood trauma. *Journal of Clinical Psychiatry* 65(supplement 1): 18–28.

Pariante CM (2006) The glucocorticoid receptor: Part of the solution or part of the problem? *Journal of Psychopharmacology* 20 (supplement 4): 79–84.

Parker KJ, Schatzberg AF, and Lyons DM (2003) Neuroendocrine aspects of hypercortisolism in major depression. *Hormones and Behavior* 43: 60–66.

Swaab DF, Bao AM, and Lucassen PJ (2005) The stress system in the human brain in depression and neurodegeneration. *Ageing Research Reviews* 4: 141–194.

VanCauter E, Leproult R, and Kupfer DJ (1996) Effects of gender and age on the levels and circadian rhythmicity of plasma cortisol. *Journal of Clinical Endocrinology and Metabolism* 81: 2468–2473.

VanCauter E, Plat L, Leproult R, and Copinschi G (1998) Alterations of circadian rhythmicity and sleep in aging: Endocrine consequences. *Hormone Research* 49: 147–145.

Stress: Definition and History

G Fink, The Mental Health Research Institute of Victoria, Melbourne, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Historical Outline of Several Stress Concepts

Aristotle, Hippocrates, and the other Ancients were aware of stress and its adverse effects. However, Claude Bernard was the first to formally explain how cells and tissues in multicelled organisms might be protected from stress. One of the world's greatest physiologists, Bernard, working in Paris during the second half of the nineteenth century, first pointed out (1859) that the internal medium of the living organism is not merely a vehicle for carrying nourishment to cells. Rather, "it is the fixity of the *milieu intérieur* which is the condition of free and independent life." That is, cells are surrounded by an internal medium that buffers changes in acid-base, gaseous (O_2 and CO_2), and ion concentrations and other biochemical modalities to minimize changes around biologically determined set points, thereby providing a steady state. Fifty years later, Walter Bradford Cannon, working at Harvard, suggested the designation homeostasis (from the Greek *homoios*, or similar, and *stasis*, or position) for the coordinated physiological processes that maintain most of the steady states in the organism. Cannon popularized the concept of 'homeostasis' in his 1932 book, *Wisdom of the Body*.

Cannon also coined the term fight or flight to describe an animal's response to threat. The concept of fight or flight proposes that animals react to threats with a general discharge of the sympathetic nervous system, priming the animal for fighting or fleeing. This response was later recognized as the first stage (acute stress response) of a general adaptation syndrome (GAS) postulated by Hans Selye to be a universal stress response among vertebrates and other organisms.

Born in Vienna in 1907, Hans Hugo Bruno Selye, also known as the 'father of stress,' began his stress research while still a medical student in 1926. He observed that patients with a variety of illnesses had many of the same 'nonspecific' symptoms that were a common response to stressful stimuli experienced by the body. These clinical observations together with experiments on laboratory rats underpinned Selye's concept of GAS, which led Selye to assert that prolonged exposure to stress resulted in 'diseases of adaptation.' That is, chronic stress, by causing the overproduction of chemicals and hormones, produced gastroduodenal ulcers and high blood pressure. Although the GAS hypothesis was subsequently shown to be incorrect, it did put stress on the map and also highlighted the fact that stress had major effects on the immune system as well as on the adrenal glands.

In addition to providing the first clear definition of stress, Hans Selye was also the first to recognize that homeostasis could not by itself ensure stability of body systems under stress. He coined the term heterostasis (from the Greek *heteros*, or other) as the process by which a new steady state was achieved by treatment with agents that stimulate the physiological adaptive mechanisms. Heterostasis could be regarded as the precursor for the concept of allostasis, first advanced by Peter Sterling and Joseph Eyer in the 1980s. That is, homeostasis, which has dominated physiological and medical thinking since the nineteenth century, is thought to provide 'stability through constancy.' Allostasis, on the other hand, provides 'stability through change' brought about by central nervous regulation of the set points that adjust physiological parameters to meet the stress/challenge.

A different tack, focused on cognition, was taken by Richard Lazarus, the eminent and influential Berkeley University psychologist. At a time when psychology tried to understand human behavior by first understanding simple organisms engaging in simple behaviors learned by associations, rewards, or punishments, Lazarus instead emphasized the importance of studying cognition, which he extended into stress and coping.

Epidemiology has been and remains central to stress studies. Numerous epidemiological studies have tried to define the effects of social, workplace, and lifestyle on stress, health, and well-being. However, of the recent epidemiological studies that have generated new stress concepts, the most important perhaps is that of David Barker, which led to the hypothesis that fetal undernutrition in middle to late gestation programs later coronary heart disease. This concept was soon extended by Hales and Barker in their 'thrifty phenotype hypothesis.' The latter proposes an association between poor fetal and infant growth and the subsequent development of type 2 diabetes and the metabolic syndrome, which afflict communities in epidemic proportions. Poor nutrition in early life, it is postulated, produces permanent changes in glucose-insulin metabolism. These changes include insulin resistance (and possibly defective insulin secretion), which combined with effects of obesity, aging, and physical inactivity are the most important factors in

determining type 2 diabetes. Many studies worldwide have confirmed Barker's initial epidemiological evidence, although the strength of the relationships has varied between studies.

With advances in genomics, the concept of susceptibility genes that increase the vulnerability of individuals to stressful life events has attracted considerable research interest. Thus, for example, the work of Avshalom Caspi and associates suggests that a polymorphism in the monoamine oxidase A (MAOA) gene promoter, which reduces MAOA expression, influences vulnerability to environmental stress, and that this biological process can be initiated by childhood abuse. Furthermore, Caspi and associates, working at the Institute of Psychiatry, London, have demonstrated that a polymorphism in the promoter of the serotonin transporter gene can render individuals more susceptible to stressful life events. In addition to their disorder-specific value, these findings have heuristic value for further thinking and research on genetic–environmental interactions that determine the response to stress and the development of mental disorders.

In parallel with these stress concepts, neuroendocrine advances revealed the physiological substrate for homeostasis, allostasis, and the stress response mechanisms. The autonomic nervous and the hypothalamic–pituitary–adrenocortical (HPA) systems subserve the afferent and efferent limbs of the stress response in vertebrates and are also central to maintaining homeostasis and effecting allostasis. The term autonomic nervous system was coined in 1898 by the Cambridge physiologist John Newport Langley, who was also renowned for his development (in parallel with Paul Ehrlich) of receptor theory. Controlled by the brain, and utilizing as neurotransmitters epinephrine and norepinephrine (sympathetic nervous component) or acetylcholine (parasympathetic component), the role of the autonomic nervous system in fight-or-flight and homeostasis (especially cardiovascular) was clearly explained by Walter Cannon.

The story of our understanding of the HPA, and the concept of the neurohumoral hypothesis of anterior pituitary control, is tortuous. The pituitary gland had long been regarded, by luminaries such as the great Harvard neurosurgeon Harvey Cushing, as the autonomous controller of the adrenal cortex, the thyroid, and the gonads. That is, the anterior pituitary gland was considered to be the 'conductor of the endocrine orchestra.' This view was reinforced by the dramatic effects of experimental pituitary removal (hypophysectomy) in rodents made feasible by the parapharyngeal surgical approach to the pituitary developed by PE Smith in approximately 1930. However, at approximately the same time, experiments were in progress that would eventually prove that the anterior pituitary gland is not autonomous; rather, it is controlled by the brain. The first of these experiments, carried out by William Rowan working alone at −50 °C in Edmonton, Alberta, in the late 1920s, showed that migration in birds was controlled by the gonads, and that gonadal size in birds was increased many-fold by increases in day length. Day length and the effects of other exteroreceptive factors, such as stress, on endocrine function together with the effects of brain tumors and trauma in the human led to an acceptance of the then (1930s) revolutionary concept that the pituitary gland is under central nervous system (CNS) control.

The neural lobe of the pituitary gland is composed of nerve projections from the paraventricular and supraoptic nuclei of the hypothalamus: these projections terminate on systemic blood vessels into which they release the nanopeptide neurohormones, vasopressin and oxytocin. In contrast, the anterior pituitary gland receives no direct innervation from the brain. Rather, the CNS control of the anterior pituitary gland is mediated by neurohormones synthesized and released from hypothalamic neurons and transported to the anterior pituitary gland by the hypophysial portal vessels. Proof of the neurohumoral hypothesis of anterior pituitary control came, first, from the elegant pituitary stalk section and pituitary grafting experiments of Geoffrey Harris and Dora Jacobsohn; second, from the characterization of some of the neurohormones by Andrew Schally and Roger Guillemin, for which they were awarded the 1977 Nobel Prize for physiology and medicine; and third, from the demonstration, first by my group, that these neurohormones were indeed released into hypophysial portal blood. Corticotropin-releasing factor (CRF), a 41-amino-acid peptide that mediates neural control of adrenocorticotropic hormone (ACTH) release from pituitary corticotropes, was isolated and sequenced by Wylie Vale and associates in 1981. A series of physiological studies, including measurements of neurohormone release into hypophysial portal blood *in vivo*, have confirmed earlier views that arginine vasopressin acts synergistically with CRF to control ACTH release.

Finally, no outline of the history of stress concepts would be complete without mention of the characterization of the adrenocortical glucocorticoids and their function. The glucocorticoids are steroid hormones whose secretion by the adrenal cortex is controlled by ACTH. The hormones of the adrenal cortex were isolated, identified, and synthesized independently by Edward Kendall (at the Mayo Foundation) and Tadeus Reichstein (at Zurich) and their associates. The availability of large amounts of synthetic steroids

enabled their physiological effects to be studied. Ultimately, Philip Hench was able to test the glucocorticoid, cortisone, in the human and demonstrate that it is a powerful anti-inflammatory agent. Hench, at the Mayo Foundation, had previously observed that rheumatoid arthritis was sometimes relieved during pregnancy and in some patients with jaundice, leading him to conclude that the pain-alleviating substance was a steroid. Kendall, Hench, and Reichstein were jointly awarded the Nobel Prize for physiology and medicine for 1950, and synthetic glucocorticoids continue to be used to treat arthritis, asthma, autoimmune conditions, and other inflammatory disorders in humans.

Definitions of Stress

Stress has a different meaning for different people under different conditions. The first and most generic definition of stress is that proposed by Hans Selye: "Stress is the nonspecific response of the body to any demand." Selye repeatedly emphasized the fact that the continued use of the word stress as a nonspecific-response to any demand was most appropriate. Selye argued that stress is not identical to emotional arousal or nervous tension since stress can occur under or in response to anesthesia in man and animals, and it can also occur in plants and bacteria that have no nervous system. This point is elaborated later in the context of stress-induced heat shock proteins (Hsps) that play a key role in cytoprotection across all three phylogenetic domains of organisms on Earth. The word stress, as used by Selye, is accepted in all foreign languages, including those in which no such word existed previously.

Stress, Selye underscored,

> is not something to be avoided. Indeed, it cannot be avoided, since just staying alive creates some demand for life-maintaining energy. Even when man is asleep, his heart, respiratory apparatus, digestive tract, nervous system, and other organs must continue to function. Complete freedom from stress can be expected only after death.

There has been much controversy and debate about Selye's concepts and particularly Selye's view that stress is best regarded as a nonspecific response. Because of their heuristic value, these points will be further considered later.

Other definitions, reviewed in detail by Selye in his treatise *Stress in Health and Disease*, include the following:

1. In behavioral sciences, stress is regarded as the "perception of threat, with resulting anxiety discomfort, emotional tension, and difficulty in adjustment."

2. In the group situation, lack of structure or loss of anchor "makes it difficult or impossible for the group to cope with the requirements of the situation, and the problem of leadership and interpersonal behavior becomes one of evolving or supplying a structure or anchor and of supplying the expertness for coping with the demands of the situation."

3. Stress can also be defined in terms of pure neuroendocrinology. Eugene Yates, for example, defined stress as any stimulus that will provoke the release of ACTH and adrenal glucocorticoids. Presumably, the same might apply to the equally powerful sympathetic markers of stress, underscored earlier by Walter Cannon.

4. Finally, Selye also mentions Richard Lazarus, famous for his work in cognitive psychology and focus on the emotions. Lazarus underscores the difficulties of reaching a precise overarching definition of stress by setting out the following different meanings of the term: "In spite of consistent confusion about the precise meaning of the term, stress is widely recognized as a central problem in human life. Scientists of many disciplines have conceptualized stress but each field appears to have something different in mind concerning its meaning. For the sociologist, it is social disequilibrium, that is, disturbances in the social structure within which people live. Engineers conceive of stress as some external force which produces strain in the materials exposed to it. Physiologists deal with the physical stressors that include a wide range of stimulus conditions that are noxious to the body. In the history of psychological stress research, there has been no clear separation between physical stressors which attack biological tissue systems and psychological stressors which produce their effects purely because of their psychological significance."

Selye's Definition of Stress – A Further Consideration

Selye's definition and concept of stress has remained controversial. For some, his definition is too biological and ignores cognitive and psychological factors, a criticism that seems to stem from the mistaken idea that cognition is not a function of the brain (a reversion to Rene Descartes' outmoded doctrine that mind and body are separate). For others, Selye's definition is too general. This section reviews the basis for Selye's definition of stress and considers whether the criticisms leveled at Selye's stress concept are valid. Overall, our observations suggest that Selye fully understood so-called psychological or cognitive stress, and that the generality of Selye's stress definition has facilitated the molecular, genotypic and phenotypic analysis of stress and stress responses across all species from bacteria to man.

General Adaptation Syndrome

Selye first put stress on the map with GAS. In search of a new hormone, Selye injected extracts of cattle ovaries into rats. The injection caused the following characteristic triad:

1. The adrenal cortex became enlarged and discharged lipid secretory granules.
2. The thymus, spleen, lymph nodes, and all other lymphatic structures showed severe involution.
3. Deep bleeding ulcers appeared in the stomach and duodenum.

Selye at first thought that these effects were due to a new hormone in the extracts but soon found that all toxic substances – extracts of kidneys, spleen, and even toxicant not derived from living tissue produced the same syndrome.

Selye surmised that the response to the injection of toxic substance reflected his "classroom concept" of "the syndrome of just being sick." That is, adrenal enlargement, thymicolymphatic involution, and gastrointestinal ulcers were the omnipresent signs of damage to the body when under attack. The three changes thus became (for Selye) the objective indices of stress and the basis for the development of the entire stress concept.

First described in a note to *Nature* in 1936, GAS has three stages: alarm, resistance, and exhaustion. In the alarm stage, the body shows changes characteristic of the first exposure to the stressor; these changes generally coincide with the sympathetic discharge that enables the fight-or-flight phenomenon of Cannon. If the stressor continues and is compatible with adaptation, features of the alarm reaction disappear and resistance develops. Prolonged exposure to the stressor may result in exhaustion and finally death.

One of the most important findings of GAS is the stress-induced thymicolymphatic involution, which highlighted for the first time that stress has a major impact on the immune system. This concept preceded by more than 20 years the discoveries of lymphocyte recirculation by James Gowans and acquired immunological tolerance by MacFarlane Burnet and Peter Medawar. Selye's discovery began the field of neuroimmunomodulation.

Selye soon became aware of the fact that the adrenal enlargement of GAS was associated with increased secretion of glucocorticoids (cortisol or corticosterone) that

> induce glycogenolysis, thereby supplying a readily available source of energy for the adaptive reactions necessary to meet the demands made by the stressors. In addition, they facilitate various other enzymatically regulated adaptive metabolic responses and suppress immune reactions as well as inflammation, assisting the body to coexist with potential pathogens.

Selye asserted that glucocorticoids are needed for adaptation to stress primarily during the alarm reaction. Selye's view that glucocorticoids enhance and mediate the stress response has been upheld with the additional concepts that glucocorticoids play a permissive role that primes the body's stress response systems and also prevent overshoot of the defense systems. Overshoots in the body's defense system are perhaps most dramatically seen in major inflammatory cataclysms called cytokine storms and the consequent systemic inflammatory response syndromes that play a key role in the lethality of avian influenza and have also occurred in response to the injection of certain antibodies. Exogenous synthetic glucocorticoids such as methylprednisolone remain a mainstay of the treatment of cytokine storms.

Although GAS is sometimes manifest in extreme stress, the three components of GAS have not withstood the test of time as indices of stress as Selye had originally proposed. Rather, the main biological markers of stress have long been behavioral observations and tests and measures of sympathetic and HPA activation. In the case of the latter measurements of glucocorticoid concentrations in blood, either alone or in parallel with plasma concentrations of ACTH, have been used as the main biological indices of stress. So, despite its heuristic value, especially for stress-induced neuroimmunomodulation, the concept of GAS has lost scientific currency.

Stressors: Features of and Lack of Stressor Sign

In *Stress in Health and Disease*, Selye underscored the fact that

> stress is part of our daily human experience, but it is associated with a great variety of essentially dissimilar problems, such as surgical trauma, burns, emotional arousal, mental or physical effort, fatigue, pain, fear, the need for concentration, the humiliation of frustration, the loss of blood, intoxication with drugs or environmental pollutants, or even the kind of unexpected success that requires an individual to reformulate his lifestyle. Stress is present in the businessman under constant pressure; in the athlete straining to win a race; in the air-traffic controller who bears continuous responsibility for hundreds of lives; in the husband helplessly watching his wife's slow, painful death from cancer; in a race horse, its jockey, and the spectator who bets on them.

Selye went on to argue that "while all these subjects face quite different problems they respond with a stereotyped pattern of biochemical, functional, and structural changes essentially involved in coping with

any type of increased demand upon vital activity, particularly adaptation to new situations."

Selye also asserted that stressors have no sign as far as evoking the stress response. That is, the response will be the same "whether the agent or situation being faced is pleasant or unpleasant; all that counts is the intensity of the demand for readjustment or adaptation that it creates." Selye underscored this point with the following poignant example:

> The mother who is suddenly told that her only son died in battle suffers a terrible mental shock; if years later, it turns out that the news was false and the son unexpectedly walks into her room alive and well, she experiences extreme joy. The specific results of the two events, sorrow and joy, are completely different, in fact they are opposite to each other, yet their stressor effect – the nonspecific demand for readjustment to a new situation – is the same.

Nonspecificity of Stress Response

Selye seemed to have been driven to find specificity in the response to different types of stressors. Thus, again in *Stress in Health and Disease* he wrote,

> It is difficult to see at first how such essentially different things as cold, heat, drugs, hormones, sorrow, and joy could provoke an identical biologic reaction. Nevertheless this is the case; it can now be demonstrated by highly objective, quantitative biochemical and morphologic parameters that certain reactions are totally nonspecific and common to all types of agents, whatever their superimposed specific effects may be.

Criticism of Selye's definition of stress has been put to experimental test. Pacak, Palkovits, and associates, for example, demonstrated that different stressors activate different stress biomarkers and different regions of the brain. Thus, low blood glucose concentrations (glucopenia) or hemorrhage activate both sympathetic and HPA systems; hyperthermia, cold, and formalin injection selectively activate the sympathetic system. On the basis of these data, Pacak and Palkovits conclude that each stressor has its own specific neurochemical signature. However, since these stress indices are limited to just two neurohumoral systems, and since for most stressors there is at least some overlap in response, it is not clear that this approach invalidates Selye's definition, "stress is the nonspecific response of the body to any demand," which would probably be unassailable had Selye omitted the term 'nonspecific.'

Heat Shock (Stress) Proteins

Whatever the shortcomings of Selye's definition of stress for the human, it is probably appropriate for the vast majority of living organisms. Living cells are classified into three main evolutionary lines or phylogenetic domains: Bacteria (eubacteria), Archaea (formerly archaebacteria), and Eucarya (eukaryotes, which encompass all plants and animals through to man). The cellular response to stress in all three phylogenetic domains is represented at the molecular level by the stress-induced synthesis of stress or Hsps, of which molecular chaperones and proteases represent two well-characterized families. The heat shock response was discovered in 1962 by Ritossa, who observed a pattern of *Drosophila* salivary gland chromosome puffs that were induced in response to transient exposure to elevated temperatures. Since then, many studies have shown that the heat shock response is ubiquitous and highly conserved in all organisms from bacteria to plants and animals. It is an essential defense mechanism for protection of cells from a wide range of stressors, including heat shock, alcohols, ischemia, inhibitors of energy metabolism, heavy metals, oxidative stress, fever, or inflammation, which depending on amplitude and duration can all cause cell death by apoptosis or necrosis. The heat shock response can protect against stress-induced cell death by way of a cell-protective process known as thermotolerance or cytoprotection, in which exposure of cells to mild stress conditions, sufficient to induce the expression and accumulation of Hsps, protects against a subsequent challenge from another stress that is, by itself, lethal. Although their precise function remains to be determined, the high degree of conservation of these Hsps across species, coupled with their importance in cell survival in various conditions, suggests that Hsps are critical for both normal cellular function and survival after a stress. Several cytoprotective functions have been attributed to Hsps and, in particular, the HSP70 family. These include (1) the folding of proteins in various intracellular compartments, (2) the maintenance of structural proteins, (3) the refolding of misfolded proteins, (4) translocation of proteins across membranes and into various cellular compartments, (5) the prevention of protein aggregation, and (6) the degradation of unstable proteins. Hsps also serve as modulating signals for immune and inflammatory responses, and they may have a role in cytokine production.

So, for the heat shock response to stressful stimuli, Selye's "stress is the nonspecific response of the body [or cell] to any demand" would appear to be appropriate.

Concepts of Stress and Disease

There is a vast literature on the role or possible role of stress in the causation and/or exacerbation of disease in most organ systems of the body. Here, attention

is focused on mental disorders. The *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV) of the American Psychiatric Association recognizes two stress disorders: acute stress disorder and posttraumatic stress disorder (PTSD). For the diagnosis of acute stress disorder, the individual, while experiencing the trauma or after the event, must have at least three of several dissociative symptoms, such as a subjective sense of numbing, detachment, or absence of emotional responsiveness; reduction in awareness of surroundings; depersonalization; or dissociative-amnesia. Following the trauma, the traumatic event is persistently reexperienced, the individual avoids stimuli that may arouse recollections of the traumatic event, and he or she has anxiety or increased arousal. The trauma causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

PTSD is defined as a condition in which a traumatic event is persistently reexperienced in the form of intrusive recollections, dreams, or dissociative flashback episodes. Cues to the event lead to distress and are avoided, and there are symptoms of increased arousal. To meet the diagnostic criteria of the DSM-IV, the full symptom picture must be present for more than 1 month, and the disturbance must cause clinically significant distress or impairment in social, occupational, or other areas of functioning.

PTSD has only been accepted officially as a mental disorder since 1980, when it was included, amid considerable controversy, in the DSM-III. References to the aftereffects of psychological trauma date back as far as the third century BC; achieved prominence during the early period of the railroad in Britain when rail travel, then precarious and physically traumatic, gave rise to a syndrome called railway spine or postconcussion syndrome; and were regarded as the basis for hysteria at the turn of the nineteenth century by neurologists and psychiatrists such as Jean-Martin Charcot, Pierre Janet, and Sigmund Freud. Long before PTSD was included in any diagnostic system, Charles Dickens wrote *A Tale of Two Cities* (1859), which can be considered as an early case report of PTSD. Interest in PTSD increased dramatically during World War I: Charles Samuel Myers was the first to coin the term and report case histories of 'shell shock,' which described a condition that afflicted many troops who screamed and wept uncontrollably, froze and could not move, became mute and unresponsive, and lost their memory, sensations, and capacity to feel. Pat Barker's monumental trilogy, *Regeneration*, deals poignantly with the psychological traumas of war and the nature of shell shock. The condition occurred again in vast numbers of people as a consequence of World War II. However, it was the psychological trauma experienced by Vietnam veterans and their demand for compensation that led to the inclusion of PTSD in the DSM-III as a condition that occurred both in civilian (e.g., rape trauma syndrome, battered woman syndrome, and abused child syndrome) and in military trauma response syndromes.

Future Developments of Stress Concepts

This brief account of some of the definitions and concepts of stress shows that there is still much to be done in the area of stress research. Thus, with respect to gene–environmental interactions, work on susceptibility genes has just begun, and there is still much room for expanding our knowledge about the role in stress of epigenetic factors and other mechanisms of gene control such as RNA interference. On the basis of previous experience, it seems likely that Barker's 'fetal origins' hypothesis might be honed and revised and may lead to a robust understanding of the metabolic syndrome and diabetes type 2. New powerful computer analysis of brain imaging and electrical recording may help to resolve the many questions that surround consciousness and cognition. The vexed cause-and-effect questions regarding the influence of stress in mental disorders, cancer, and other diseases need to be answered. In the past 25 years, we have witnessed how Selye's cherished and widely accepted axiom that stress is the cause of gastric ulceration (one of the three components of GAS) was undone by the careful observations and courageous perseverance of Barry Marshall and Robin Warren, who demonstrated that, in fact, most gastroduodenal ulcers are caused by a microbe, *Helicobacter pyloris*, that is readily amenable to treatment with antibiotics. And yet, there is evidence that stress does play a role in gastric ulceration, so the stress – *H. pyloris* interaction needs to be worked out. These and other biological questions about stress are likely to be tractable – rational and rigorous biology will almost certainly win the day.

However, the causes of human stress, acute and posttraumatic, are commonly social and sociological, political and irrational (beliefs in religious and racial superiority). Our continuing strife and conflict reflect in part man's innate, often irrational (limbic-brain generated) drive for conquest, territory, acquisition, and reproduction of the species. Perhaps the fact that man is such a smart political animal works against us and could ultimately lead to our stressful self-destruction. It is not clear that anything can protect man from this fate.

See also: Adrenal Steroids: Biphasic Effects on Neurons; Chronic (Repeated) Stress: Consequences, Adaptations; Corticotropin-Releasing Hormone: Integration of Adaptive Responses to Stress; Posttraumatic Stress Disorder: Neurobiology; Stress and Neural Involvement in Metabolism; Stress and Cognition; Stress and Suicide; Stress and Vulnerability to Brain Damage; Stress and Parasympathetic Control; Stress Response: Sex Differences; Stress and Neuronal Plasticity; Stress, Sex and Adolescent Nicotine Response; Stress, Dopamine, and Puberty; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load.

Further Reading

Armitage JA, Khan IY, Taylor PD, Nathanielsz PW, and Poston L (2004) Developmental programming of the metabolic syndrome by maternal nutritional imbalance: How strong is the evidence from experimental models in mammals? *Journal of Physiology* 561: 355–377.

Arnetz BB and Ekman R (eds.) (2006) *Stress in Health and Disease*. Weinheim: Wiley–VCH Verlag.

Barker DJP (1995) Fetal origins of coronary heart disease. *British Medical Journal* 311: 171–174.

Cannon WB (1915) *Bodily Changes in Pain, Hunger, Fear and Rage: An Account of Recent Researches into the Function of Emotional Excitement*. New York: Appleton.

Cannon WB (1932) *The Wisdom of the Body*. New York: Norton.

Caspi A, McClay J, Moffitt TE, et al. (2002) Role of genotype in the cycle of violence in maltreated children. *Science* 297(5582): 851–854.

Caspi A and Moffitt TE (2006) Gene–environment interactions in psychiatry: Joining forces with neuroscience. *Nature Reviews Neuroscience* 7: 583–590.

Caspi A, Sudgen K, Moffit TE, et al. (2003) Influence of life stress on depression: Moderation by a polymorphism in the *5-HTT* gene. *Science* 301: 386–389.

Eriksson JG (2005) The fetal origins hypothesis—10 years on. *British Medical Journal* 330: 1096–1097.

Fink G (ed.) (2007) *Encyclopedia of Stress*, 2nd edn. Oxford: Elsevier.

Fink G, Smith JR, and Tibballs J (1971) Corticotrophin releasing factor in hypophysial portal blood of rats. *Nature* 203: 467–468.

Hales NC and Barker DJP (2001) The thrifty phenotype hypothesis. *British Medical Bulletin* 60: 5–20.

Huber TJ and te Wildt AT (2005) Charles Dickens' *A Tale of Two Cities*: A case report of posttraumatic stress disorder. *Psychopathology* 38(6): 334–337.

Jolly C and Morimoto RI (2000) Role of the heat shock response and molecular chaperones in oncogenesis and cell death. *Journal of the National Cancer Institute* 92: 1564–1572.

Kendler KS, Kuhn JW, Vittum J, Prescott CA, and Riley B (2005) The interaction of stressful life events and a serotonin transporter polymorphism in the prediction of episodes of major depression. *Archives of General Psychiatry* 62: 529–535.

Kregel KC (2002) Heat shock proteins: Modifying factors in physiological stress responses and acquired thermotolerance. *Journal of Applied Physiology* 92: 2177–2186.

Lasiuk GC and Hegadoren KM (2006) Posttraumatic stress disorder. Part I: Historical development of the concept. *Perspectives in Psychiatric Care* 42: 13–20.

Lazarus RS (2000) Toward better research on stress and coping. *American Journal of Psychology* 55(6): 665–673.

Lazarus RS (2006) Emotions and interpersonal relationships: Toward a person-centered conceptualization of emotions and coping. *Journal of Personality* 74: 9–46.

Macario AJ and Conway De Macario E (2007) Molecular Chaperones: Multiple functions, pathologies, and potential applications. *Frontiers in Bioscience* 12: 2588–2600.

Marmot MG (2005) Social determinants of health inequalities. *The Lancet* 365: 1099–1104.

McMillen C and Robinson JS (2005) Developmental origins of the metabolic syndrome: Prediction, plasticity, and programming. *Physiological Reviews* 85: 571–633.

Myers CS (1915) A contribution to the study of shell shock. *The Lancet* 1(February 13): 316–320.

Pacak K and Palkovits M (2001) Stressor specificity of central neuroendocrine responses: Implications for stress-related disorders. *Endocrine Reviewical* 22: 502–548.

Schulkin J (ed.) (2004) *Allostasis, Homeostasis, and the Costs of Physiological Adaptation*. Cambridge, UK: Cambridge University Press.

Selye H (1936) A syndrome produced by diverse nocuous agents. *Nature* 138: 32.

Selye H (1975) Confusion and controversy in the stress field. *Journal of Human Stress* 1(2): 37–44.

Selye H (1976) *Stress in Health and Disease*. Stoneham, MA: Butterworth.

Suntharalingam G, Perry MR, Ward S, et al. (2006) Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. *New England Journal of Medicine* 355(10): 1018–1028.

Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load

B S McEwen, The Rockefeller University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The brain mediates adaptation to changes in the physical and social environment through the autonomic, neuroendocrine, and immune systems as well as through behavioral responses that include fighting or fleeing, as well as health-promoting or health-damaging responses. Adaptation to stressful events or environmental changes is an active process that involves the output of mediators such as neurotransmitters and modulators, as well as many hormones and cytokines and chemokines of the immune system. The goal of this adaptation is to maintain homeostasis and promote survival of the organism. However, the process of adaptation also produces an almost inevitable wear and tear on the body and brain, and this wear and tear is exacerbated if there are many stressful events and/or if the mediators that normally promote adaptation are dysregulated, that is, not turned on when needed or not turned off efficiently when no longer needed. The description and analysis of this scenario has entailed the addition of some new terminology, such as allostasis, and the refinement of some classical terminology, such as homeostasis and stress.

Characteristics of Homeostatic Systems

Homeostasis refers to the ability of an organism to maintain the internal environment of the body within limits that allow it to survive. Homeostasis also refers to self-regulating processes that return critical systems of the body to a set point within a narrow range of operation, consistent with survival of the organism.

Homeostasis is highly developed in warm-blooded animals living on land, which must maintain body temperature, fluid balance, blood pH, and oxygen tension within rather narrow limits, while at the same time obtaining nutrition to provide the energy to maintain homeostasis. This is because maintaining homeostasis requires the expenditure of energy. Energy is used for locomotion, as the animal seeks and consumes food and water, for maintaining body temperature via the controlled release of calories from metabolism of food or fat stores, and for sustaining cell membrane function as it resorbs electrolytes in the kidney and intestine and maintains neutral blood pH.

Homeostasis also refers to the body's defensive mechanisms. These include protective reflexes against such things as inhaling matter into the lungs, the vomiting reflex as a protection to expel toxic materials from the esophagus or stomach, the eye blink reflex, and the withdrawal response to hot or otherwise painful skin sensations. There is also the defense against pathogens through innate and acquired immunity.

According to Walter B Cannon,

> Bodily homeostasis ... results in liberating those functions of the nervous systems that adapt the organism to new situations, from the necessity of paying routing attention to the management of the details of bare existence. Without homeostatic devices, we should be in constant danger of disaster, unless we were always on the alert to correct voluntarily what normally is corrected automatically. With homeostatic devices, however, that keep essential body processes steady, we as individuals are free from such slavery – free to enter into agreeable relations with our fellows, free to enjoy beautiful things, to explore and understand the wonders of the world about use, to develop new ideas and interests, and to work and play, untrammeled by the anxieties concerning our bodily affairs. (Cannon, 1929)

The body has a considerable redundancy, or safety margin, when it comes to many of these mechanisms. There are two kidneys and adrenal glands, in case one fails, and the digestive tract is able to function adequately in food absorption, even if intestinal length is shortened. Moreover, individuals are able to survive with one lung. Also, the cardiovascular system has a number of devices to maintain blood pressure after considerable blood loss. Finally, as Cannon notes, the body displays considerable resilience in the face of injury and after surgery, in order to compensate for lost function.

Need for Refocusing of the Concept of Homeostasis and Additional Terminology

Yet, there are paradoxes and problems with the notion of a set point that is always maintained. In discussing obesity, Cannon states "What leads to the storage of fat in large amounts – sometimes in grotesquely large amounts – in some persons, and in slight amounts in others, is not well-understood." This and other aspects of pathophysiology, particularly in response to stress, has forced a re-evaluation of the processes that reestablish homeostasis in the

face of a challenge and which, under some circumstances, exacerbate disease. This is one of the reasons why the terms allostasis and allostatic load (see below) have been introduced.

Another problem with 'homeostasis' is that, taken literally, it is a rather static concept, akin to the notion of equilibrium in classical thermodynamics, whereas, in real life, an organism experiences changing set points around which it maintains stability over a finite period of time. Moreover, unlike the closed systems of classical thermodynamics, a living organism is an open system in which energy and matter flow in and out, with equilibrium being replaced by steady state. Cannon recognized this and used the term steady state and even suggested that 'homeodynamics' might be a better term than 'homeostasis.'

The same problem was also recognized by a number of other authors. Nicolaides introduced the term homeorheusis, in which 'stasis' is replaced by 'rheusis,' meaning 'something flowing.' Mrozovsky introduced the term rheostasis to mean "a condition or state in which, at any one instant, homeostatic defenses are still present but over a span of time there is a change in the regulated level, or set point of a system." Mrozovsky considers this concept advantageous in explaining how organisms adjust to changing environments such as seasons of the year by storing body fat or changing reproductive physiology, and he believes that rheostasis allows for the kinds of physiological plasticity that is involved in evolution.

What Is Stress?

Stress may be defined as a real or interpreted threat to the physiological or psychological integrity of an individual that results in physiological and/or behavioral responses. In biomedicine, stress often refers to situations in which the adrenal glucocorticoids and catecholamines are elevated because of an experience. Stress is also a subjective experience that may or may not correspond to physiological responses, and the word stress is widely used in many languages as part of daily discourse. There is 'good stress' and 'bad stress,' and people talk about bad stress as 'being stressed out.'

Stress involves a stressor and a stress response. A stressor may be a physical insult, such as trauma or injury, or physical exertion, particularly when the body is being forced to operate beyond its capacity. Other physical stressors include noise, overcrowding, excessive heat or cold. Stressors also include primarily psychological experiences such as time-pressured tasks, interpersonal conflict, unexpected events, frustration, isolation, and traumatic life events, and all of these types of stressors may produce behavioral responses and evoke physiological consequences such as increased blood pressure, elevated heart rate, increased cortisol levels, impaired cognitive function, and altered metabolism.

Behavioral responses to stressors may decrease risk and get the individual out of trouble or involve health-promoting activities such as a good diet and regular exercise, but they may also include responses that exacerbate the physiological consequences of stress, for example, self-damaging behaviors like smoking, drinking, overeating, or consuming a rich diet, or risk-taking behaviors like driving an automobile recklessly. The physiological stress responses include primarily the activation of the autonomic nervous system and the hypothalamo–pituitary–adrenal (HPA) axis, leading to increased blood and tissue levels of catecholamines and glucocorticoids. It is these physiological responses that have both protective and damaging effects (see the sections on allostasis and allostatic load).

There are two important features of the physiological stress response: the first involves turning it on in amounts that are adequate to the challenge (see the section 'Allostasis'). The second is turning off the response when it is no longer needed. The physiological mediators of the stress response, namely, the catecholamines of the sympathetic nervous system and the glucocorticoids from the adrenal cortex, initiate cellular events that promote adaptive changes in cells and tissues throughout the body, which in turn protect the organism and promote survival. However, too much stress, or inefficient operation of the acute responses to stress, can cause wear and tear and exacerbate disease processes.

There are enormous individual differences in interpreting and responding to what is stressful, as well as individual differences in the susceptibility to diseases, in which stress may play a role. Genetic predispostions exist which increase the risk of certain disorders. In addition, developmental process, such as prenatal stress or nurturing postnatal experiences, contribute to the lifelong responsiveness of the behavioral and physiological responses to stressors. Furthermore, experiences throughout the life course resulting in memories of particularly unpleasant or pleasant situations combine with the genetic and developmental influences to produce large differences among individuals in how they react to stress and what the long-term consequences may be.

Allostasis

Allostasis means 'achieving stability through change'; it was introduced by P Sterling and J Eyer in 1988.

Allostasis refers to the process that maintains homeostasis, as defined above, and it recognizes that 'set points' and other boundaries of control may change with environmental conditions. There are primary mediators of allostasis such as, but not confined to, hormones of the HPA axis, catecholamines, and cytokines. Allostasis also clarifies the inherent ambiguity in the term homeostasis and distinguishes between the systems that are essential for life (homeostasis) and those that maintain these systems in balance (allostasis). Allostatic systems enable an organism to respond to its physical state (e.g., awake, asleep, supine, standing, exercising) and to cope with noise, crowding, isolation, hunger, extremes of temperature, physical danger, psychosocial stress, and microbial or parasitic infections. For example, blood pressure rises when we get up in the morning and fluctuates with exercise and other demands of the waking period, and these fluctuations make it possible for the individual to function in response to a changing social and physical environment.

Whether exposed to danger, an infection, a crowded and noisy neighborhood, or having to give a speech in public, the body responds to challenge by turning on an allostatic response, thus initiating a complex pathway for adaptation and coping, and then shutting off this response when the challenge has passed. Two primary allostatic responses involve the sympathetic nervous system and HPA axis. 'Turning on' these systems leads to release of catecholamines from sympathetic nerves and the adrenal medulla and the secretion of adrenocorticotropin (ACTH) from the anterior pituitary gland, which stimulates release of glucocorticoids from the adrenal cortex.

'Turning off' these responses leads to a return to baseline concentrations of cortisol and catecholamines, and this normally happens when the danger is past, the infection contained, the living environment is improved or the talk has been given. As long as the allostatic response is limited to the period of challenge, protection via adaptation predominates over adverse consequences. However, over weeks, months, or even years, exposure to elevated levels of stress hormones (an allostatic state) can result in allostatic load and overload, with resultant pathophysiological consequences.

Allostatic States

An allostatic state refers to the altered and sustained activity levels of the primary mediators, for example, glucocorticoids, that integrate energetic and associated behaviors in response to changing environments, challenges such as social interactions, weather, disease, predators, pollution, etc. An allostatic state results in an imbalance of the primary mediators reflecting excessive production of some and inadequate production of others. Examples are hypertension, a perturbed cortisol rhythm in major depression or after chronic sleep deprivation, chronic elevation of inflammatory cytokines and low cortisol in chronic fatigue syndrome, imbalance of cortisol, corticosteroid-releasing factor (CRF), and cytokines in the Lewis rat that increases risk for autoimmune and inflammatory disorders. Allostatic states can be sustained for limited periods if food intake and/or stored energy, such as fat, can fuel homeostatic mechanisms (e.g., bears and other hibernating animals preparing for the winter). If imbalance continues for longer periods and becomes independent of maintaining adequate energy reserves, then symptoms of allostatic overload appear. Abdominal obesity is an example of this condition.

Allostatic Load and Overload

Allostatic load and allostatic overload refers to the cumulative result of an allostatic state. For example, fat deposition in a bear preparing for the winter, a bird preparing to migrate or a fish preparing to spawn are examples of animals experiencing an allostatic load. This is a largely beneficial and adaptive condition and can be considered the result of the daily and seasonal routines which organisms use to obtain food and survive, and obtain the extra energy needed to migrate, molt, and breed. Within limits, they are adaptive responses to seasonal and other demands. However, if one superimposes on this additional load of unpredictable events in the environment, disease, human disturbance (for animals in the wild), and social interactions, then allostatic load can increase dramatically and become allostatic overload, which can become a pathophysiological condition.

There are two distinctly different outcomes of an allostatic state in terms of allostatic load or overload. First, if energy demands exceed energy income, and also exceeds what can be mobilized from stores, then type 1 allostatic overload occurs. For example, breeding birds use increasing food abundance in spring to reproduce and raise their young. If inclement weather then increases costs of maintaining homeostasis and the allostatic load of breeding, and at the same time reduces food available to fuel that allostatic load, then negative energy balance results in loss of body mass and suppression of reproduction. Another example is the mass movement of seabirds to islands in the face of a severe storm that limit access to food. Second, if energy demands are not exceeded and the organism continues to take in or store as much or even more energy than it needs, perhaps as a result of

stress-related food consumption, choice of a fat-rich diet, or metabolic imbalances (prediabetic state) that favor fat deposition, then type 2 allostatic overload occurs. Besides fat deposition, there are other cumulative changes in other systems that can result from repeated stress, for example, neuronal remodeling or loss in hippocampus, atherosclerotic plaques, left ventricular hypertrophy of the heart, glycosylated hemoglobin and other proteins by advanced glycosylation end products as a measure of sustained hyperglycemia, high cholesterol with low high-density lipoprotein (HDL), increased oxidative stress, elevated proinflammatory mediators, and chronic pain and fatigue, for example, in arthritis or psoriasis, associated with imbalance of immune mediators.

Types of Allostatic Overload

Four situations can lead to allostatic overload, resulting in overexposure to stress hormones over time. The first is frequent stress, for example, blood pressure surges can trigger myocardial infarction in susceptible persons, and experiments in primates have shown that repeated blood pressure elevations over weeks and months accelerates atherosclerosis, which increases the risk of myocardial infarction. Thus, the frequency of stressful events determines the degree of this type of allostatic load, leading to increased exposure to hormones and other allostatic mediators.

A second type of allostatic overload is the failure to habituate to repeated challenges, such as in some individuals' cortisol response during a repeated public speaking challenge.

The third type of allostatic overload is caused by an inability to shut off allostatic responses. This pertains to the termination of stress and also to the decline of these same hormones during the natural diurnal cycle. With respect to the stress response, blood pressure in some persons fails to recover after acute mental arithmetic stress, and hypertension accelerates atherosclerosis. Intense athletic training also induces allostatic load in the form of elevated sympathetic and HPA activity, which results in reduced body weight and amenorrhea and the often-related condition of anorexia nervosa. Regarding the diurnal cycle, women with a history of depressive illness have decreased bone mineral density because the allostatic load of chronic, moderately elevated glucocorticoid levels, especially at night, that are associated with depression cause chronic, reduced levels of osteocalcin.

The fourth type of allostatic overload is caused by inadequate allostatic responses that trigger compensatory increases in other allostatic systems. When one system does not respond adequately to a stressful stimulus, the activity of other systems is increased because the underactive system is not providing the usual counter-regulation. For example, if adrenal steroid secretion does not increase appropriately in response to stress, secretion of inflammatory cytokines (which are counter-regulated by adrenal steroids) increases. In Lewis rats, the failure to mount an adequate HPA response results in increased vulnerability to autoimmune and inflammatory disturbances; thus, an infection or wound that causes an inflammation will lead to an increased inflammatory response. Human counterparts of the state of HPA hyporesponsiveness include individuals with fibromyalgia and chronic fatigue syndrome.

Examples of allostatic states leading to allostatic overload include hypertension, reduced parasympathetic activity, increased levels of inflammatory cytokines and oxidative stress, elevated insulin and glucose, and triglycerides. In the brain, the remodeling of neurons in the hippocampus, amygdala, and prefrontal cortex, as a result of chronic stress, are examples of allostatic overload, which impair memory and enhance anxiety and aggression. For the immune system, acute stress enhances immune responses, whereas chronic stress has the opposite effect. Chronic stress is reported to shorten teleomeres and reduce telomerase activity on white blood cells (and possibly other cells, as well).

Early stress and neonatal handling influence the course of aging and age-related cognitive impairment in animals. Early experiences are believed to set the level of responsiveness of the HPA axis and autonomic nervous system. These systems overreact in animals subject to early unpredictable stress and underreact in animals exposed to the neonatal handling procedure. In the former condition, brain aging is accelerated, whereas in the latter, brain aging is reduced. Lifespan is also influenced by early-life differences in HPA reactivity: rats that show increased HPA reactivity as neonates to novelty have shorter life spans compared to rats with lesser HPA reactivity to novelty.

Role of Behavior in Allostatic Overload

Anticipation and worry can also contribute to allostatic overload. Anticipation is involved in the reflex that prevents us from blacking out when we get out of bed in the morning and is also a component of worry, anxiety, and cognitive preparation for a threat. Anticipatory anxiety can drive the output of mediators like ACTH, cortisol, and adrenalin; thus, prolonged anxiety and anticipation is likely to result in allostatic overload. For example, salivary cortisol levels increase within 30 min after waking in individuals who are under considerable psychological stress due to work or family matters. Moreover, intrusive

memories from a traumatic event (e.g., in posttraumatic stress disorder) can produce a form of chronic, internal stress and can drive physiological responses.

Allostasis, allostatic states, and allostatic overload are also affected by health-damaging behaviors such as smoking, drinking, excess calorie intake, or, on the opposite side, health-promoting behaviors such as a healthy diet and regular moderate exercise. These behaviors are integral to the overall notion of allostasis – how individuals cope with a challenge – and also contribute to increasing or decreasing allostatic overload by known pathways. For example, a rich diet accelerates atherosclerosis and progression to noninsulin-dependent diabetes by increasing cortisol, leading to fat deposition and insulin resistance; smoking elevates blood pressure and atherogenesis. Yet, regular, moderate exercise protects one against cardiovascular disease, prevents diabetes, and increases attention and other aspects of cognitive function.

See also: Adrenal Steroids: Biphasic Effects on Neurons; Chronic (Repeated) Stress: Consequences, Adaptations; Stress and Neural Involvement in Metabolism; Stress and Cognition; Stress and Suicide; Stress and Vulnerability to Brain Damage; Stress and Parasympathetic Control; Stress Response: Sex Differences; Stress Response and Self-Esteem; Stress, Sex and Adolescent Nicotine Response; Stress, Dopamine, and Puberty; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History.

Further Reading

Cannon WB (1939) *The Wisdom of the Body.* New York: W. W. Norton and Co.

Cannon W (1929) The Wisdom of the Body. *Physiological Reviews* 9: 399–431.

Cavigelli SA and McClintock MK (2003) Fear of novelty in infant rats predicts adult corticosterone dynamics and an early death. *Proceedings of the National Academy of Sciences of the United States of America* 100: 16131–16136.

Epel ES, Blackburn EH, Lin J, et al. (2004) Accelerated telomere shortening in response to life stress. *Proceedings of the National Academy of Sciences of the United States of America* 101: 17312–17315.

Kirschbaum C, Prussner JC, Stone AA, et al. (1995) Persistent high cortisol responses to repeated psychological stress in a subpopulation of healthy men. *Psychosomatic Medicine* 57: 468–474.

Mason J (1959) Psychological influences on the pituitary–adrenal cortical system. In: Pincus G (ed.) *Recent Progress in Hormone Research*, pp. 345–389. New York: Academic Press.

McEwen BS (1998) Protective and damaging effects of stress mediators. *New England Journal of Medicine* 33: 171–179.

McEwen BS and Stellar E (1993) Stress and the individual: Mechanisms leading to disease. *Archives of Internal Medicine* 153: 2093–2101.

McEwen BS and Wingfield JC (2003) The concept of allostasis in biology and biomedicine. *Hormones and Behavior* 4: 2–15.

Meaney M, Aitken D, Berkel H, Bhatnager S, and Sapolsky R (1988) Effect of neonatal handling of age-related impairments associated with the hippocampus. *Science* 239: 766–768.

Mrosovsky N (1990) *Rheostasis: The Physiology of Change.* New York: Oxford University Press.

Nicolaides S (1977) Physiologie du comportement alimentaire. In: Meyer P (ed.) *Physiologie Humaine*, ch. 2, pp. 908–922. Paris: Flammarion Medecine-Sciences.

Sapolsky R (1992) *Stress, the Aging Brain and the Mechanisms of Neuron Death.* Cambridge: MIT Press.

Schulkin J, McEwen BS, and Gold PW (1994) Allostasis, amygdala, and anticipatory angst. *Neuroscience and Biobehavioral Reviews* 18: 385–396.

Selye H (1936) A syndrome produced by diverse nocuous agents. *Nature* 138: 32.

Selye H (1955) Stress and disease. *Science* 122: 625–631.

Sterling P and Eyer J (1988) Allostasis: A new paradigm to explain arousal pathology. In: Fisher S and Reason J (eds.) *Handbook of Life Stress, Cognition and Health*, pp. 629–649. New York: Wiley.

Striatum: Internal Physiology

C J Wilson, University of Texas at San Antonio, San Antonio, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

The Physiological Function of the Striatum

The striatum is a part of the thalamocortical system, residing in a loop that connects the cortex to the thalamus. Inputs from nearly every part of the cortical mantle make connections in the striatum. The output of the striatum, after one connection in the basal ganglia output nuclei, influences neurons in some parts of the thalamus which project back to the cortex. There have been many proposed functions for this transcortical loop. Most of them have in common the idea that the striatum detects combinations of cortical inputs. Cortical areas can communicate directly, but they do so in a very limited way, through specific corticocortical projections. The striatum receives cortical inputs in a more general way, with some striatal cells receiving convergent input from a wide range of combinations of cortical regions. Cortical areas project to the striatum in a roughly topographical order, with some cortical areas projecting to widely separated regions of the striatum and others showing greater or lesser overlap. Within these broad domains, cortical inputs form complex, discontinuous innervation zones so that cortical areas that project to the same general region of the striatum may converge to a greater or lesser degree. In general, the patterns of convergence follow function-related rules. For example, the somatosensory cortex projections from multiple body maps converge in the striatum so that different body parts have nonoverlapping representations, but different cortical representations of the same body part overlap extensively. The axonal arborizations of individual corticostriatal axons are often very large, and the volume of striatum occupied by the dendritic field of a single striatal neuron is innervated by an enormous number of cortical cells located in widespread regions of the cortex. Thus, even cortical areas that project to mostly nonoverlapping regions of the striatum have some small region of convergence with some proportion of striatal cells receiving both inputs. Therefore, the striatum may be able to identify specific combinations of activity in functionally related cortical circuits that could not be detected in the cortex, or in the thalamus, where cortical afferents are also organized in a way that allows only limited overlap of cortical projections from diverse functional areas.

Of course, the simplest version of this combinatorial view of the striatum cannot be completely accurate. The striatum is smaller than the cortex and has fewer neurons, and there are many more combinations of activity in cortical cells than there are cortical cells. The striatum cannot be designed to detect and distinguish every possible arbitrary combination of activity in the cortex. To make a combinatorial idea credible, one must specify which combinations of cortical patterns the striatum is designed to detect and what combinations are disregarded or confounded with others. Possibly, the striatum extracts a low-resolution view of the cortex, detecting only the large-scale patterns that involve many neurons in each cortical region. This implies that single corticostriatal neurons are good, but low-resolution, representatives of the average activity signal for their cortical areas. Alternatively, the striatum may specifically detect some limited number of classes of coincident activity in cortical regions. If this second view is correct, and the striatum is selecting out some specific set of transcortical events, how are the correct convergent connections of the relevant cortical inputs onto specific striatal neurons specified so that each striatal neuron receives the correct set of synaptic connections? Perhaps the connections are established by exercise of a learning rule, and the large number of potential combinations of cortical activity patterns is the raw material from which the striatum would sculpt a representation of cortical activity based on some particular set of criteria. If these criteria could be discovered, they would constitute a compact description of the function of the striatum.

Afferents to the Striatum

Like most brain regions, the striatum has a large number of different sources of afferents, only a few of which have been studied in detail. The well-studied striatal afferents are the corticostriatal, thalamostriatal, and nigrostriatal pathways. Among the less studied are projections from the amygdala, the dorsal raphe nucleus, and a projection from the external segment of the globus pallidus.

The Corticostriatal Pathway

The corticostriatal pathway is enormous. It consists of at least two different projections, arising from different kinds of cortical neurons, located in different

cortical layers and with different axonal branching patterns. One corticostriatal pathway arises from collaterals of large pyramidal cells located mostly in layer V that project to the brain stem and perhaps even to the spinal cord. These cells have exclusively ipsilateral projections to the striatum. A second corticostriatal pathway arises from the collaterals of corticocortical cells, including those that form the callosal connections between the hemispheres. Bilaterally projecting corticocortical cells also project bilaterally to the striatum. Cortical efferents are diverse, so there may be a variety of subcategories of corticostriatal neurons that have very specific patterns of axonal collateralization. However, in any case, the corticostriatal neurons are not a unique kind of cortical cell with properties and connections specifically adapted for communication with the striatum. They are the same cells that carry cortical output signals to the brain stem and that subserve communication among cortical areas ipsilaterally and bilaterally.

The cortical projection neurons generally employ glutamate as their neurotransmitter, and glutamate serves this role in the corticostriatal projection. Stimulation of the cortex *in situ* or in slices (after eliminating complications due to inadvertent stimulation of other axons and polysynaptic pathways) produces brief, simple AMPA receptor-mediated excitatory postsynaptic potentials (EPSPs) in most striatal neurons. With removal of magnesium block (i.e., by depolarization), an NMDA receptor-mediated component is also seen. Cortically evoked glutamate release can also activate metabotropic glutamate receptors, for which a variety of effects have been reported.

The main target of the corticostriatal projection is the dendritic spine of the striatal spiny neuron. Approximately 90% of all corticostriatal synapses are made onto dendritic spines. Cortical inputs make a higher proportion of their synapses on spines than do other inputs, including those from the thalamus, which are also glutamatergic and often end on spines. Overall, approximately 80% of all asymmetrical synapses with round synaptic vesicles (the type of synapse formed by the corticostriatal projection) are formed on spines. Some neurons in the striatum are not strongly excited by stimulation of the cerebral cortex. For example, the aspiny cholinergic interneurons have been reported to respond weakly or not at all to cortical stimuli that were very effective at exciting spiny projection neurons located nearby. Another class of aspiny interneuron, the parvalbumin-positive fast-spiking interneuron, is strongly excited by cortical stimulation. This cell receives fewer synapses of cortical origin than the spiny cells; its strong response is probably due to the high input resistance and greater excitability of this neuron.

The Thalamostriatal Pathway

Like the corticostriatal pathway, the thalamostriatal projection is actually at least two separate pathways. One thalamostriatal projection, arising from the intralaminar nuclei proper (e.g., the central lateral nucleus), makes synapses on dendritic spines of spiny neurons and strongly resembles the corticostriatal projection. The other, arising from the parafascicular nucleus and center median, makes asymmetrical synapses, but mostly on dendritic shafts rather than on dendritic spines. The physiological responses to these two pathways have not been studied in isolation, but stimulation of the thalamus *in vivo* generates EPSPs that strongly resemble those of the corticostriatal pathway. The magnitude of the thalamostriatal projection is often overlooked. The maximal EPSPs evoked by stimulation of the cortex or thalamus are approximately the same size, implying approximate parity in effect on striatal neurons. However, the thalamostriatal projection must arise from a much smaller number of cells, so the rules governing convergence of thalamic inputs must be much different. The functions of the intralaminar thalamic nuclei are obscure. It has been difficult to incorporate the thalamostriatal projection into theoretical ideas about the function of the basal ganglia, despite its physiological prominence.

The Nigrostriatal Pathway

Because of its apparent importance in a wide range of functional disorders, from Parkinson's disease to schizophrenia, more effort has been expended on understanding the physiological function of dopamine in the striatum than on any other aspect. After decades of work, it is clear that dopamine does not directly control ligand-gated ion channels, at least on the spiny neurons of the striatum, and does not directly generate a change in their membrane potential that can control firing rate. On the other hand, dopamine has been shown to affect many other aspects of striatal cell function, including efficacy of corticostriatal synapses, spike threshold, calcium entry and disposition in the cell, and corticostriatal (perhaps also thalamostriatal) synaptic plasticity. Thus, it is difficult to discern a pattern in dopamine's effects in the striatum. Most work has focused on three general categories of dopamine effects: modulation of voltage-sensitive ion channels, presynaptic effects on transmitter release, and effects on synaptic plasticity.

It has been well demonstrated that dopamine can alter the voltage sensitivity or maximal conductance

of a host of voltage-sensitive ion channels on projection cells and interneurons in the striatum. What is missing from this information is an organizing principle that can help us to understand the net effect of all the dopamine-mediated modulations. For example, dopamine alters the voltage sensitivity of inactivation of sodium channels in spiny cells, shifting the spike threshold in the positive direction. Through the D1 receptor, dopamine can enhance activation of the voltage-sensitive L-type calcium channels in spiny cells possessing this receptor, whereas D2 receptors on other spiny cells have an inhibitory effect on L channels. Likewise, dopamine has a host of effects on release of neurotransmitters and other modulators, but the net effect of these modulations does not immediately suggest a single function for dopamine in the circuit. Dopamine is a prerequisite for synaptic plasticity of glutamate synaptic potentials in the striatum, which complement theoretical studies showing that the firing of dopamine neurons is very similar to that predicted for a critical part of the reinforcement–learning process.

The Striatal Spiny Projection Cell

Spiny projection cells outnumber all other types of striatal neurons by approximately 20 to 1. They receive the vast majority of afferent synapses, and their axon collaterals are the largest source of intrinsic synapses in the striatum. They are also the source of the efferent fibers of the striatum. Thus, the striatal network can be said to consist of the spiny cells, their afferents, and their interconnections. The spiny neurons are recognized in extracellular recordings by their phasic firing patterns. They have a low rate of background activity, often remaining silent for minutes at a time in an animal that is sitting quietly, and fire in brief episodes of activity that may be related to events of a motor, sensory, or cognitive nature. In many studies, investigators have attempted to characterize the nature of the circumstances that give rise to responses of spiny cells. These studies have identified striatal spiny cells that respond in nearly any kind of motor, sensory, or cognitive task, but the responses of single neurons are very specifically related to some distinct aspect of a task. This is consistent with the view, described previously, that individual striatal spiny neurons respond to some specific set of combinations of cortical activity, but that the striatum as a whole can represent a wide spectrum of cortical activity states.

The responses of spiny neurons apparently reflect the convergent activity of large numbers of cortical and thalamic afferents. In the absence of strong synaptic input, the cells are deeply hyperpolarized because of the action of a hyperpolarization-activated potassium conductance (KIR) that holds the cell near the potassium equilibrium potential. It is this ion channel that is primarily responsible for the low level of background firing among spiny neurons and also for the low input resistance of the spiny cell. To activate the spiny cell, synaptic input must be strong enough to overcome the KIR current, which requires depolarizing the cell by 10–15 mV, at which point the potassium channel deactivates and the cell enters a high-input resistance state in which it is very sensitive to excitation. If depolarized further, an additional set of depolarization-activated potassium channels are activated, which oppose synaptic excitation. The combination of excitatory synaptic inputs and depolarization-activated K^+ current creates another high-conductance state at approximately −50 to −55 mV, near but still slightly negative to the action potential threshold. Thus, the striatal spiny neurons spend most of their time in one of the two high-conductance states – the hyperpolarized 'down' state when the level of excitatory input is too low to overcome the KIR current, and the more depolarized 'up' state during sustained excitation. In the up state, action potentials may be evoked by strong excitatory inputs that drive the cell the remaining few millivolts to the spike threshold. The existence of the up and down states is determined by ionic conductances known to be present in the spiny neuron membrane, but the pattern of transitions between the states reflects the pattern of convergent synaptic input to the cells. In locally anesthetized, immobilized rats, striatal neurons remain in the down state for many seconds or minutes, and they have up state transitions that may last for 0.1 to several seconds. In rats anesthetized with barbituates, up and down state transitions are rhythmic at 5–7 Hz, corresponding to the spindles seen in cortical and thalamic neurons, including those that form the corticostriatal input. In rats anesthetized with urethane, up and down state transitions are rhythmic at a much lower frequency of approximately 1 Hz or less, corresponding to the slow oscillations seen in the cortex and thalamus under these conditions.

The dependence of spiny cell membrane potential fluctuations and firing patterns on the pattern of inputs suggests that the striatal cell is not directly responsible for any of these oscillations. That is, the cell's membrane potential is not bistable but, rather, has a single stable state, the down state; the up state is a driven state that requires synaptic input for its maintenance. This is demonstrably so, at least in urethane-anesthetized animals and in slices of the striatum (Figure 1). When forced into the membrane

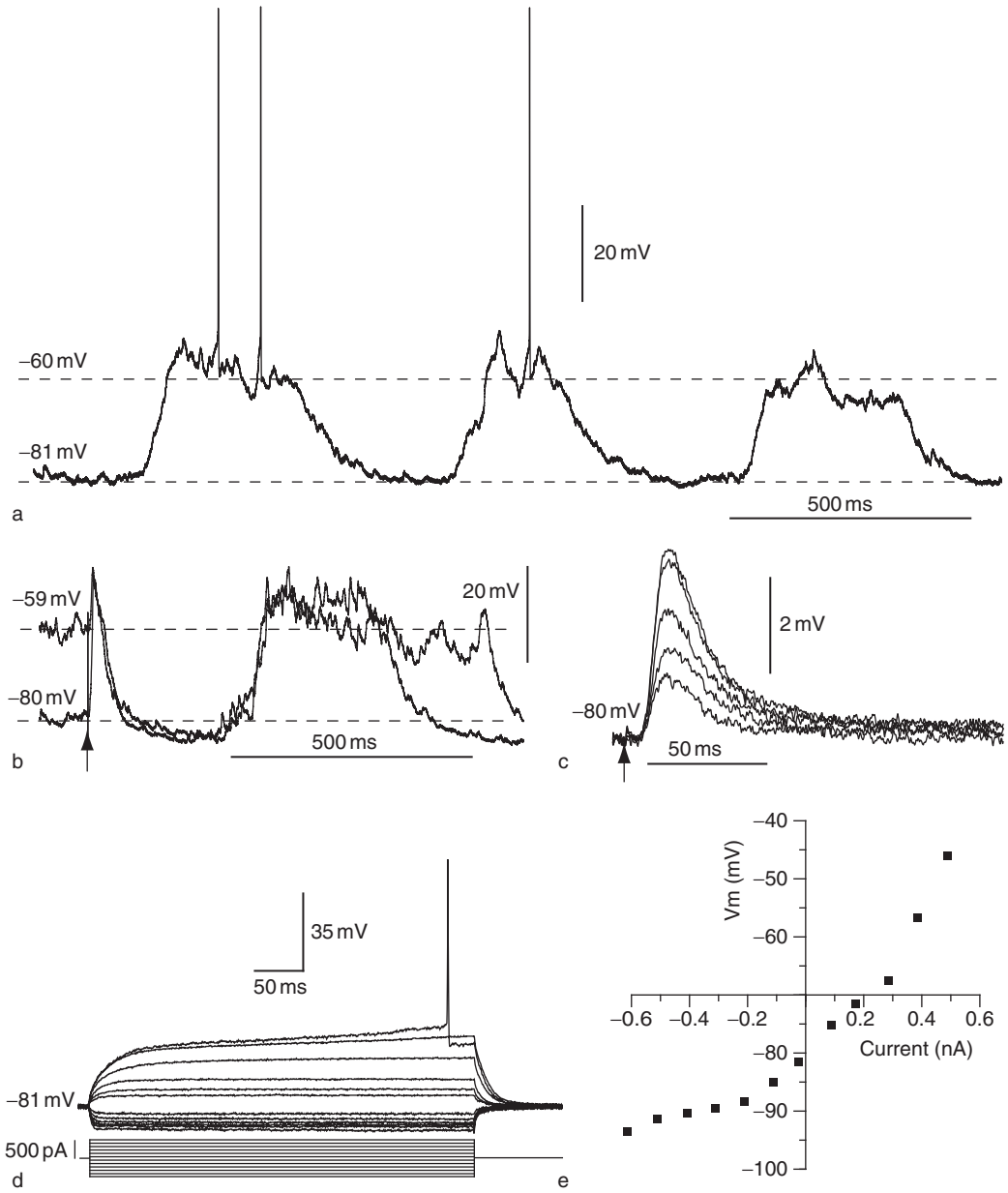

Figure 1 Physiological properties and synaptic responses of spiny projection neurons. (a) Spontaneous membrane potential fluctuations and firing *in vivo* (urethane anesthesia). Note the large membrane potential fluctuations between a hyperpolarized down state and a depolarized up state. Firing only occurs in the up state, but not every up state evokes action potentials. (b) Response to thalamic stimulation (arrow) in the up and down states. The EPSP is followed by a transition to the down state, followed 250–500 ms later by an up-state transition. Responses to cortical stimulation are similar (not shown). (c) Responses to a range of intensities of local stimulation in striatal slices. Note that cells in the slice are in the down state. Local stimulation produces an EPSP that is finely graded with stimulus intensity. (d) Responses to a graded series of constant current pulses. Note the low input resistance in the hyperpolarizing direction from rest (the down state) and increasing membrane resistance in the depolarizing direction. Firing is accompanied by a slow ramp depolarization and can occur arbitrarily late. (e) Current–voltage curve generated from the traces in (d). Note the strong inward rectification.

potential range of the up state by an experimentally applied input, spiny neurons in the down state return to the down state immediately upon removal of the depolarizing influence. However, the powerful voltage-dependent ion channels in the membrane of the spiny neuron shape the membrane potential response to excitatory and inhibitory inputs, and they impose a requirement for cooperation among the excitatory inputs from the cortex and thalamus. An excitatory input that arrives by itself in the down state is completely overcome by the hyperpolarizing influence of the KIR current. The same input arriving

in conjunction with other excitatory inputs that drive the cell into the low-conductance region between the states may be empowered to trigger an up-state transition or even and action potential.

How many extrinsic inputs are required to trigger an up-state transition or action potentials in a spiny neuron? Each spiny neuron receives approximately 10 000 excitatory axospinous inputs from the cerebral cortex and thalamus. Anatomical studies suggest that these arise from a nearly equal number of different cortical and thalamic neurons. Computer simulations and studies of explant cultures of the striatum suggest that the number of coincident inputs required to drive the spiny cell to the up state may be as low as 100–300. If that proves to be the case *in vivo*, then there may be a very large number of different cortical and thalamic input combinations that are equivalent in driving the spiny striatal cell. Each cell may not be searching for just one transcortical combination of inputs but, rather, a huge number of such combinations, that are considered equivalent by the spiny neuron. Alternatively, the large number of input combinations might be viewed as a list of possible cortical patterns from which the spiny neuron should select a correct set by the exercise of a synaptic learning rule that disables all the other combinations.

Striatal spiny neurons are GABAergic and inhibitory in their effects on their target structures in the globus pallidus and substantia nigra. Before exiting the striatum, the axons of spiny neurons give rise to axon collaterals that arborize in the general vicinity of the cells of origin, forming a dense plexus of locally derived inhibitory synapses. Many of the synapses formed by these spiny cell axons are made onto the dendrites and dendritic spine shafts of neighboring spiny cells. These mutually inhibitory synaptic connections among spiny neurons were originally proposed to be the cause of the low background firing pattern of the spiny cells. However, extracellular recording studies of the effects of local GABA blockade on firing patterns of spiny neurons showed that inhibition was not responsible for either the low background rate or the phasic firing pattern of the cells. Firing rates during the episodes of firing (probably corresponding to up states seen in intracellular recordings) were increased by blockade of all striatal $GABA_A$ receptors in the vicinity of the recorded cells. These observations and others suggest that collateral inhibition (and other GABAergic synaptic inhibition in the striatum) is more effective in controlling the firing rate and patterns of cells in the up state than in sculpting the temporal patterns of up and down states. Studies employing intracellularly recorded pairs of cells have shown that within a neighborhood corresponding roughly to the dendritic field of the cells, spiny cells have a very high level of connectivity; each cell makes a synapse on approximately one of every six of its neighbors and receives a synapse from approximately the same proportion. Reciprocally connected pairs are relatively rare. The connections among spiny cells are very weak when viewed from the soma, largely because of their distal location. Like the excitatory synapses on the spiny neuron (which are similarly distributed on the cells), the strength of the inhibitory synapses depends critically on the conductance state of the spiny cell.

Direct and Indirect Pathway

There are two types of spiny striatal neurons that can be distinguished by their axonal projections and by a number of cytochemical markers. The cell subtypes are named according to their axonal projections. The cells that form the direct pathway are those whose axons project directly to the basal ganglia output neurons of the globus pallidus internal segment and substantia nigra pars reticulata, as well as the external segment of globus pallidus. The indirect pathway spiny neurons project to the external segment of globus pallidus but do not connect directly to the output nuclei. These cells differ in their cotransmitters, with the direct pathway releasing substance P as well as GABA, whereas the indirect pathway neurons employ enkephalin and dynorphin as cotransmitters. They also differ in dopamine receptors, with the direct pathway expressing the D1 receptor and the indirect pathway expressing the D2 receptor. The two types of spiny cells are similar in their overall physiological properties, and both receive cortical and thalamic excitation. The axon collaterals of both types of neurons make synapses on both direct and indirect pathway cells. There are some differences in cortical innervation of direct and indirect pathway neurons, with collaterals of brain-stem-projecting cortical cells preferentially innervating indirect pathway neurons, whereas the collaterals of bilaterally projecting cells preferentially innervate the direct pathway cells. A number of questions have been posed based on the differential distribution of dopamine receptors on direct and indirect pathway spiny neurons. Generally, both direct and indirect pathway neurons respond to D1- and D2-dependent processes. For example, corticostriatal long-term depression is exhibited by all spiny neurons, but it depends critically on D1 dopamine receptors. Spiny cell responses to dopamine in such cases are primarily mediated by interneurons that possess both receptors.

Striosomes and Matrix

Another categorization of spiny neurons arises from the spatial compartmentalization of the striatum into striosomes (patches) and matrix. These compartments can be distinguished on the basis of cytochemical markers, such as calbindin (the expression of which is much higher in the matrix spiny cells) and μ opiate receptors (the expression of which is much higher in the neuropil of the striosomes). The corticostriatal innervations of striosomes and matrix are also different, with each following a separate topographical scheme. Both compartments contain direct and indirect pathway neurons, so there are four different kinds of spiny neurons. No physiological differences have been discovered between striosome and matrix cells, but the dendrites and axons of the spiny cells observe the boundaries of these compartments, so the local inhibition among spiny cells is restricted to their respective compartments. Striosome and matrix neurons have been reported to differ in their axonal projections, with direct pathway striosomal neurons projecting preferentially to the substantia nigra pars compacta.

The Cholinergic Interneuron

Extracellular recording studies of the striatum have always included a sample of cells that fire continuously, in contrast to the phasic firing pattern of the spiny cells. These cells, called tonically active neurons, also have unusual responses in behaving animals. Their responses are much less diverse than those of spiny neurons. They respond to sensory stimuli of high salience, including strong or meaningful unconditioned stimuli, or stimuli which have acquired a high salience by association. Their responses consist of a long-lasting interruption of their background firing, often flanked by preceding and subsequent moderate rate increases. Intracellular recording of cells firing in this pattern has revealed them to be large cholinergic interneurons, one of the best-known interneurons of the striatum (Figure 2). Although few in number, these cells have extremely dense axonal projections that provide the striatum with one of the densest cholinergic innervations in the brain.

The background firing of cholinergic interneurons is autonomous. When isolated from inputs, they fire at approximately the same rate as (although more rhythmically than) they do *in situ*. In slices, their firing is driven by a combination of a persistent sodium current that turns on at approximately -60 mV, action potential-triggered calcium influx through N-type calcium channels, and a calcium-dependent potassium current that repolarizes the cell after each action potential and delays the depolarization that leads to the next spike. A hyperpolarization-activated inward current (HCN) ensures that the cell will depolarize into the range of the persistent Na^+ current when the calcium-dependent potassium current ebbs sufficiently because of removal of calcium from the cytoplasm.

The cholinergic interneuron is also seen to fire in irregular and rhythmic bursting firing patterns, both in slices and *in situ*. Cholinergic interneurons have two different bursting mechanisms, which can both act at the same time. The first is an unusual mechanism consisting of a regenerative hyperpolarization that interrupts ongoing firing. This occurs because of a hyperpolarization-activated potassium current (KIR) that at some membrane potentials is strong enough to drive the cell membrane to a deeply hyperpolarized level. At the bottom of this hyperpolarization, a strong HCN current is activated and restores the membrane potential. This mechanism, which is basically a very slow hyperpolarizing spike, can lead to rhythmic bursting or can amplify hyperpolarizations that are triggered in other ways. One such source of hyperpolarizations is a slow afterhyperpolarization that occurs after a period of sustained depolarization (or a burst of action potentials). The slow afterhyperpolarization is caused by an unidentified calcium-dependent potassium current that is linked to L-type calcium current and calcium release from intracellular stores. Although strong depolarizations are not required to activate this hyperpolarizing current, prolonged depolarizations (longer than 20 ms) are required so the current is not triggered by the individual action potentials during single spiking. Cholinergic cells fire in either the single spiking or bursting patterns, depending on the relative strength of the mechanisms described previously. These mechanisms are sufficiently fluid that cholinergic cells can switch between firing patterns in a matter of seconds. The synaptic mechanism of the pause response in cholinergic interneurons is not understood, but some hints are available from the *in vivo* literature. *In vivo* studies have shown that an intact dopamine innervation and an intact thalamic input are required. However, the most interesting aspect of this response is its ability to be conditioned to nearly any stimulus that is used as a signal, and it is unknown how this occurs.

Like dopaminergic synapses, cholinergic synapses on spiny cells do not produce any strong ionic currents in spiny cells but act mainly through muscarinic receptors that activate G-proteins and have a wide variety of effects on voltage-sensitive channels, synaptic transmission, and synaptic plasticity. Effects on synaptic plasticity have received much attention because of the similarity between the responses of

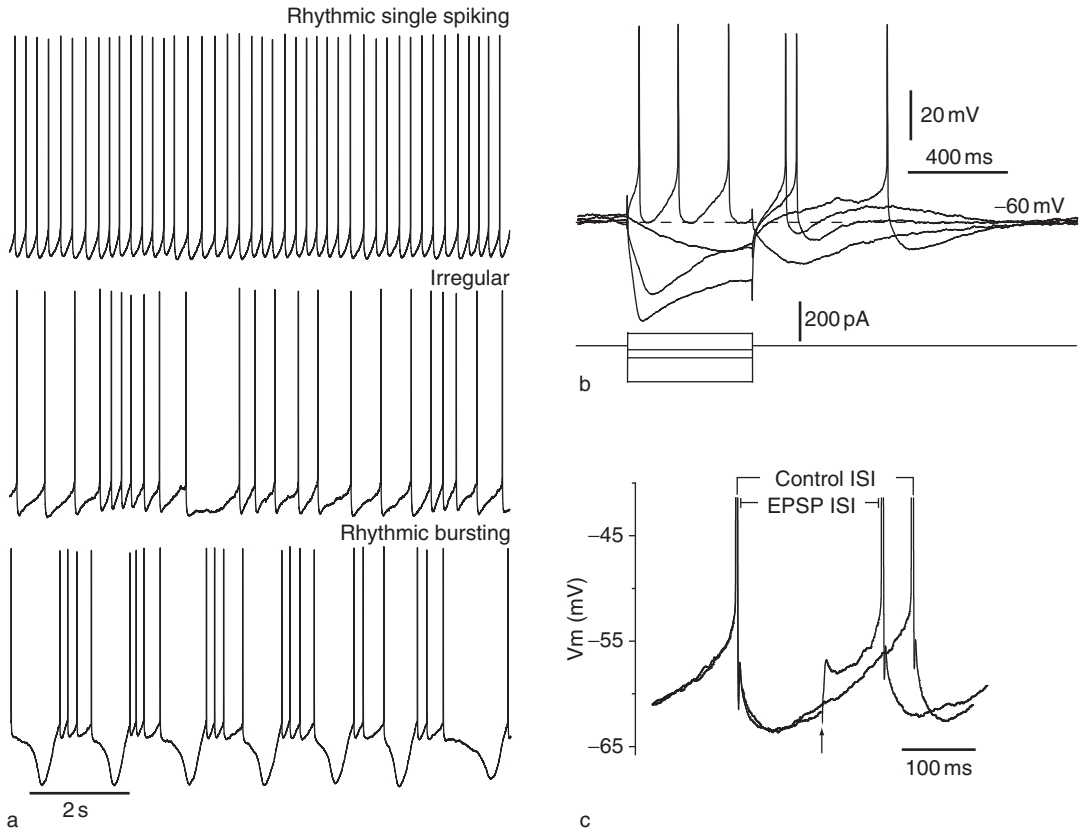

Figure 2 Physiological properties and synaptic responses of the cholinergic striatal interneuron. (a) The range of spontaneous firing patterns seen in slices. (b) Responses to constant current pulses. Note the strong time-dependent inward rectification and rebound in response to hyperpolarizing pulses and the slow afterhyperpolarization following depolarizing pulses. (c) Response to local stimulation in slices (arrow). Two traces are superimposed, one showing the spontaneous firing of a rhythmic single spiking neuron and one in which excitation evoked by local stimulation occurs approximately half way between two spontaneous action potentials. Note that the effect of synaptic stimulation is not primarily to evoke or prevent action potentials but, rather, to change their timing. The action potential occurs long after the synaptic potential is over, but it occurs earlier than it would have if the EPSP had been absent. In spontaneously active neurons such as the cholinergic cell, synaptic inputs can interact with the oscillation to change spike timing rather than to directly drive firing.

dopaminergic and cholinergic cells to reward and reward-related cues. Acetylcholine acts on a variety of striatal targets in addition to the spiny cell. Its effects on dopaminergic axon terminals, mediated via nicotinic receptors, may be important for the well-known interaction between cholinergic and dopaminergic drugs. Its effects on the fast-spiking interneuron may also be indirectly influential on spiny cells and the output of the striatum.

Fast-Spiking Parvalbumin-Positive Interneuron

Another striatal interneuron is easily recognized anatomically because of its strong expression of the calcium-binding protein parvalbumin. These cells are GABAergic and make their synapses mainly on the somata and proximal dendrites of the spiny projection cells. Physiologically, they can be recognized by their brief action potentials and ability to fire at high rates with minimal spike frequency adaptation. Because of these properties, which they share with parvalbumin-positive neurons in other forebrain regions, they are usually called fast-spiking (FS) interneurons (Figure 3). FS neurons are not spontaneously active, and they are quiet when isolated from excitatory inputs, as occurs in slices. Like spiny neurons, they receive a strong cortical (and probably thalamic) innervation, which causes them to fire phasically *in vivo*. FS cells are the most likely candidates for the inhibition that controls spiny cell firing in the up state. Although the total number of FS GABAergic synapses received by each spiny cell is less than that of the Sp→Sp synapse, individual synapses from FS cells are much more powerful than those formed by other spiny cells, and each FS cell makes more synapses on any one spiny cell. There are many fewer FS cells; thus, when a single FS cell fires (as these cells readily

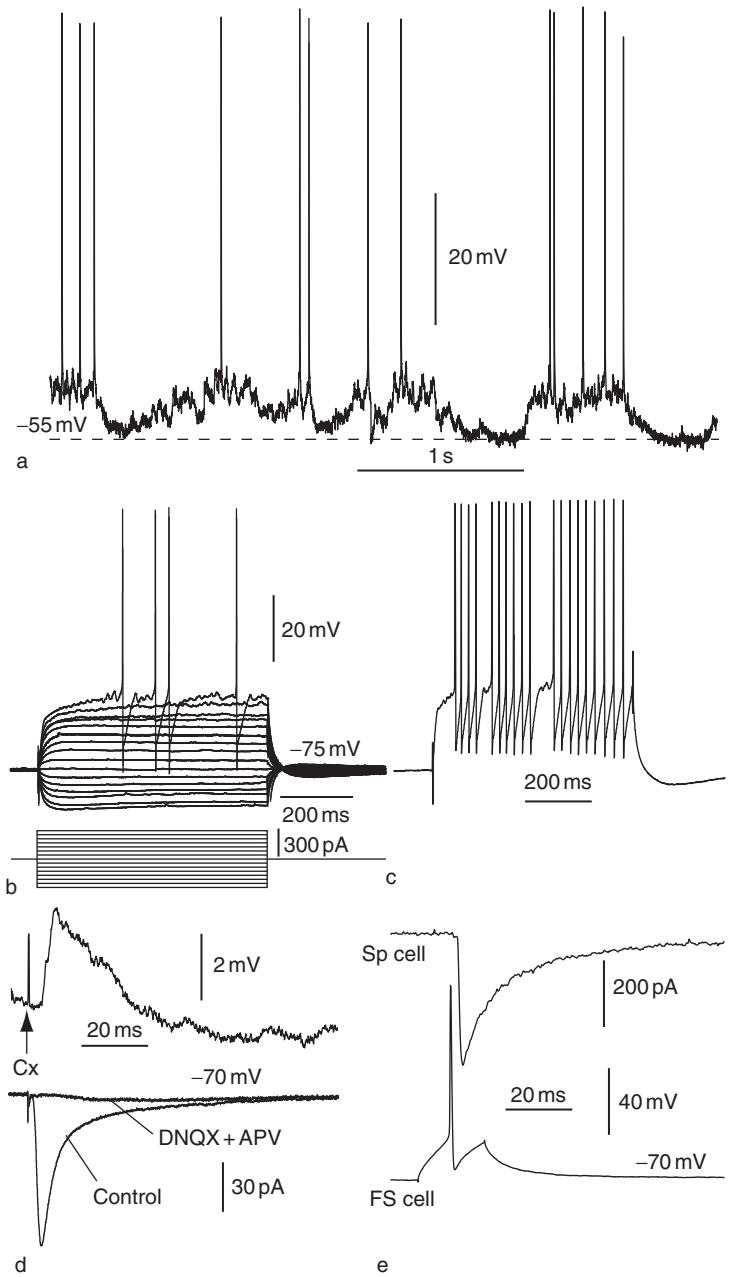

Figure 3 Physiological properties, synaptic responses, and synaptic effects of the fast-spiking (FS) striatal interneuron. (a) Spontaneous membrane potential fluctuations and firing of an FS neuron *in vivo* (urethane anesthesia). Phasic synaptic excitation of the FS cell gives rise to episodes of firing but not to distinct up and down states of the membrane potential. (b) Responses to a graded series of constant current pulses. Note the absence of strong rectification. (c) Response to strong depolarizing current includes episodic firing. (d) Synaptic responses to excitatory inputs. (Top trace) EPSP response to cortical stimulation *in vivo* (urethane anesthesia). (Bottom trace) Excitatory postsynaptic current evoked by local stimulation holding at −70 mV, superimposed on the response to the same stimulus after treatment with glutamate antagonists DNQX and APV. (e) Inhibitory postsynaptic current evoked in a striatal spiny neuron (top trace) by a single action potential in a nearby FS cell (bottom trace). The spiny neuron was loaded with chloride to move the reversal potential to 0 mV, and the spiny (Sp) cell was held at −70 mV.

do when corticostriatal inputs are active) it exerts a powerful effect on all spiny cells in its vicinity. FS cell axons do not extend over very large volumes of the striatum, so their influence can be spatially discrete. However, FS cells are interconnected by gap junctions that are likely to promote synchrony among those cells and which may sometimes allow for coordinated inhibition over large portions of the striatum.

Somatostatin/Nitric Oxide Synthetase-Containing Low-Threshold Spike Interneuron

The physiological properties of one other interneuron have been described. Like the other striatal interneurons, our knowledge of the low-threshold spike (LTS) neuron derives mostly from the studies of Kawaguchi and associates, who used immunochemical identification to establish the correspondence between physiological properties and the morphological and cytochemical cell types. Although they do not contain a high level of somatic GAD or GABA, the somatostatin-containing interneurons are GABAergic and make many of their synapses on the dendrites of spiny cells. The physiological properties of these cells are known only in outline. Their most striking feature is the presence of a large regenerative posthyperpolarization rebound (low-threshold spike) resembling that seen in thalamic relay neurons. At least some of these cells also demonstrate membrane bistability – that is, self-sustaining plateau potentials that can be triggered by depolarizations but long outlast the triggering stimulus. Like the FS cell, these neurons exert powerful inhibitory effects on spiny neurons within their axonal fields.

Conclusions

The striatum is fundamentally a feed-forward single-layer network consisting of the spiny neurons and their afferents. Because there are four categories of spiny neurons, the direct and indirect cells in the matrix and striosomes, the striatum should be considered as four such networks. The topographical arrangement of both cortical and thalamic afferents imposes a diversity of information on these striatal networks according to location within the striatum and also among cells at a given location. Cross-talk among spiny neurons is weak, and there are no long-range collaterals of spiny neurons, so the internal structure of the spiny cell network does not impose strong spatial intrinsic patterns on the output of the nucleus. Powerful feed-forward inhibitory interactions arise from the activity of GABAergic interneurons. Dopaminergic and cholinergic inputs act indirectly on the spiny cells, modulating synaptic plasticity and altering the excitability of inhibitory interneurons.

See also: Basal Ganglia: Internal Organization; Basal Ganglia: Physiological Circuits; Procedural Learning: Striatum; Somatostatin and Receptors.

Further Reading

Aosaki T, Graybiel AM, and Kimura M (1994) Effect of the nigrostriatal dopamine system on acquired neural responses in the striatum of behaving monkeys. *Science* 265: 412–415.

Bennett BD, Callaway JC, and Wilson CJ (2000) Intrinsic membrane properties underlying spontaneous tonic firing in neostriatal cholinergic interneurons. *Journal of Neuroscience* 20: 8493–8503.

Blackwell KT, Czubayko U, and Plenz D (2003) Quantitative estimate of synaptic inputs to striatal neurons during up and down states *in vitro*. *Journal of Neuroscience* 23: 9123–9132.

Centonze D, Picconi B, Gubellini P, Bernardi G, and Calabresi P (2001) Dopaminergic control of synaptic plasticity in the dorsal striatum. *European Journal of Neuroscience* 13: 1071–1077.

Flaherty AW and Graybiel AM (1991) Corticostriatal transformations in the primate somatosensory system. Projections from physiologically mapped body-part representations. *Journal of Neurophysiology* 66: 1249–1263.

Graybiel AM, Aosaki T, Flaherty A, and Kimura M (1994) The basal ganglia and adaptive motor control. *Science* 265: 1826–1831.

Gustafson N, Gireesh-Dharmaraj E, Czubayko U, Blackwell KT, and Plenz D (2006) A comparative voltage and current-clamp analysis of feedback and feedforward synaptic transmission in the striatal microcircuit *in vitro*. *Journal of Neurophysiology* 95: 737–752.

Kawaguchi Y (1993) Physiological, morphological and histochemical characterization of three classes of interneurons in rat neostriatum. *Journal of Neuroscience* 13: 4908–4923.

Kita H (1996) Glutamatergic and GABAergic postsynaptic responses of striatal spiny neurons to intrastriatal and cortical stimulation recorded in slice preparations. *Neuroscience* 70: 925–940.

Koos T and Tepper JM (1999) Inhibitory control of neostriatal projection neurons by GABAergic interneurons. *Nature Neuroscience* 2: 467–472.

Koos T, Tepper JM, and Wilson CJ (2004) Comparison of IPSCs evoked by spiny and fast-spiking neurons in the neostriatum. *Journal of Neuroscience* 24: 7916–7922.

Lei W, Jiao Y, Del Mar N, and Reiner A (2004) Evidence for differential cortical input to direct pathway versus indirect pathway striatal projection neurons in rats. *Journal of Neuroscience* 24: 8289–8299.

Mahon S, Deniau J-M, and Charpier S (2003) Various synaptic activities and firing patterns in cortico-striatal and striatal neurons *in vivo*. *Journal of Physiology (Paris)* 97: 557–566.

Mallet N, Le Moine C, Charpier S, and Gonon F (2005) Feedforward inhibition of projection neurons by fast-spiking GABA interneurons in the rat striatum *in vivo*. *Journal of Neuroscience* 25: 3857–3869.

Matsumoto N, Minamimoto T, Graybiel AM, and Kimura M (2001) Neurons in the thalamic CM-Pf complex supply striatal neurons with information about behaviorally significant sensory events. *Journal of Neurophysiology* 85: 960–976.

Nambu A, Kaneda K, Tokuno H, and Takada M (2002) Organization of corticostriatal motor inputs in monkey putamen. *Journal of Neurophysiology* 88: 1830–1842.

Nicola SM, Surmeier J, and Malenka RC (2000) Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annual Review of Neuroscience* 23: 185–215.

Nisenbaum ES and Wilson CJ (1995) Potassium currents responsible for inward and outward rectification in rat neostriatal spiny projection neurons. *Journal of Neuroscience* 15: 4449–4463.

Plotkin JL, Wu N, Chesselet M-F, and Levine MS (2005) Functional and molecular development of striatal fast-spiking GABAergic interneurons and their cortical inputs. *European Journal of Neuroscience* 22: 1097–1108.

Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology* 57: 87–115.

Tepper JM, Koos T, and Wilson CJ (2004) GABAergic microcircuits in the neostriatum. *Trends in Neurosciences* 27: 662–669.

Tunstall MJ, Oorschot DE, Kean A, and Wickens JR (2002) Inhibitory interactions between spiny projection neurons in the rat striatum. *Journal of Neurophysiology* 88: 1263–1269.

Wickens JR, Reynolds JN, and Hyland BI (2003) Neural mechanisms of reward-related motor learning. *Current Opinion in Neurobiology* 13: 685–690.

Wilson CJ (1986) Postsynaptic potentials evoked in spiny neostriatal projection neurons by stimulation of ipsilateral and contralateral neocortex. *Brain Research* 367: 201–213.

Wilson CJ (2005) The mechanism of intrinsic amplification of hyperpolarizations and spontaneous bursting in striatal cholinergic interneurons. *Neuron* 45: 575–585.

Wilson CJ, Chang HT, and Kitai ST (1983) Origins of postsynaptic potentials evoked in spiny neostriatal projection neurons by thalamic stimulation in the rat. *Experimental Brain Research* 51: 217–226.

Wilson CJ and Kawaguchi Y (1996) The origins of two-state spontaneous membrane potential fluctuations of neostriatal spiny neurons. *Journal of Neuroscience* 16: 2397–2410.

Zheng T and Wilson CJ (2002) Corticostriatal combinatorics: The implications of corticostriatal axonal arborizations. *Journal of Neurophysiology* 87: 1007–1017.

Stroke

D Leifer, Weill Cornell Medical College; and New York Presbyterian Hospital–Weill Cornell Medical Center, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Definition

Strokes are acute focal injuries to the brain that have a vascular cause and occur because of interruptions to the blood supply of the brain (ischemic strokes) or because of bleeding in or around the brain (hemorrhagic strokes). Approximately 85% of strokes are ischemic, and the others are hemorrhagic. Sometimes, the word stroke is used with a more limited definition that includes only ischemic events.

Epidemiology

Stroke is a major public health problem throughout the world. In the United States, there are more than 700 000 strokes per year. Stroke is the third leading cause of death after heart disease and cancer, and it is a leading cause of disability as well. Of those who survive a stroke, 90% have a deficit. Stroke is usually considered a disease of the elderly, and in the United States approximately two-thirds of strokes occur in people older than the age of 65 years, but strokes can occur at any age and their consequences in terms of lost productivity are even greater in younger patients than in older ones.

Stroke versus Transient Ischemic Attack

Strokes are distinguished from transient ischemic attacks (TIAs), which are traditionally defined as events that occur when there is a transient interruption of the blood supply to a part of the brain with focal deficits that resolve within 24 h. TIAs defined in this way were thought not to cause any permanent damage to the brain, but magnetic resonance imaging (MRI) with diffusion-weighted imaging has demonstrated that as many as half of all so-called TIAs are actually small strokes. Permanent damage is more likely the longer that symptoms persist. True TIAs that are not associated with any permanent damage mostly resolve within a few hours or less and may last for only a few minutes. A new definition of TIA has been proposed that classifies transient events with MRI evidence of damage as strokes.

In any case, the same processes that cause strokes also cause TIAs, and clinically, TIAs should be evaluated with the same rigor as strokes to identify their causes because there is a significant risk of a stroke after a TIA, especially within the first few days after the TIA.

Hemorrhagic Strokes

There are two major categories of hemorrhagic stroke: intracerebral hemorrhages within the brain and hemorrhages in the space around the brain. Intracerebral hemorrhages occur because of bleeding from small blood vessels within the brain, or less frequently from vascular malformations, in which there are tangles of abnormal blood vessels with direct connection from small arteries to veins. Small blood vessels may bleed because of chronic damage from hypertension or because they have been weakened by amyloid deposition, which is particularly a problem in elderly people older than approximately 80 years.

Intracerebral hemorrhage can also result from bleeding into an ischemic stroke. Such bleeding may be petechial in nature with only tiny patches of blood that are sometimes only visible at autopsy but can sometimes be seen on computerized tomography (CT) or MRI. Larger areas of hemorrhage are referred to as areas of hemorrhagic conversion, and even more extensive bleeding into an ischemic infarct may result in the formation of a cavity filled with blood (i.e., a hematoma) that can damage surrounding brain by compressing it.

Hemorrhages can also occur in the subarachnoid, subdural, or epidural spaces. Subarachnoid hemorrhages most commonly result from rupture of an aneurysm – a balloon-like protrusion from one of the main arteries around the base of the brain. Aneurysmal hemorrhages may recur with catastrophic or fatal bleeding within the first few days after an initial bleed, so rapid imaging to identify an aneurysm and then emergent treatment to obliterate the aneurysm are critically important. Aneurysms can be treated by surgically clipping the so-called neck that connects them to the artery from which they arise or by angiographic procedures in which they are filled with tiny metal coils delivered through a special catheter threaded into the artery. In addition to repeat bleeding, aneurysmal hemorrhages can also lead to vasospasm that causes constriction of arteries primarily near the skull base and may lead to ischemic strokes. Vasospasm is treated by aggressive hydration and the use of vasopressor drugs that elevate blood pressure to improve perfusion through the constricted arteries. Sometimes, it is also treated by intra-arterial vasodilating drugs or by angioplasty.

Subarachnoid hemorrhage can also occur spontaneously in the absence of an aneurysm; such hemorrhages may be venous in origin, are typically perimesencephalic in location, and have a benign prognosis. Subarachnoid hemorrhage can also occur if a parenchymal hemorrhage ruptures through the surface of the brain or if there is head trauma.

Bleeding may also occur into the subdural and epidural spaces. Bleeding into these spaces usually results from trauma, in which case it is usually not considered a stroke, but is occasionally spontaneous. These types of hemorrhage often require surgical drainage.

Treatment of acute hemorrhages involves the use of medications to reverse any abnormal bleeding tendency and to reduce reactive swelling around the hemorrhage, when appropriate, and surgical drainage or resection if the hemorrhage excessively compresses adjacent parts of the brain.

Ischemic Strokes

Approximately 85% of strokes are ischemic. Ischemic strokes are the focus of the remainder of this article. Several different mechanisms lead to the interruptions in blood flow that cause ischemic strokes.

Arteries that supply the brain may become blocked because of atherosclerotic plaques in which cholesterol is deposited. These plaques gradually narrow the artery. Often, blood clots form on top of the plaque and may completely close off the artery. In addition, these clots may break off (or, in other words, embolize) and travel into smaller arteries until they ultimately lodge in an artery that is too small for them to go any further.

Cholesterol deposition can also occur in small arterioles and leads to progressive narrowing with lipohyalinosis, in which there is not only formation of lipid clefts in the vessel walls but also hyalinosis, which may reflect chronic stresses to the vessel from hypertension.

Arteries may also become blocked if a clot forms in the heart and travels through the great vessels and then into progressively smaller arteries until it lodges in one that is too small for it to pass through. The most common cause for clot to form in the heart is atrial fibrillation, an arrhythmia in which the atria do not contract normally. Left ventricular thrombus, which forms when parts of the heart do not contract properly, for example, after a myocardial infarction, can also embolize and cause a stroke. In addition, in patients who have endocarditis, an infection of the heart valves, infected material or vegetations can develop on the valves and can also embolize to cause a stroke.

Uncommon Causes of Ischemic Stroke

Atherothrombotic and cardioembolic mechanisms are by far the most important causes of stroke, but several other causes of stroke, although rare, are worth noting.

Strokes are also caused by vasculitis with inflammation that is limited to the cerebral circulation in primary central nervous system (CNS) vasculitis or that is part of a systemic vasculitis. CNS vasculitis was originally defined in 1959. Pathologically, there are mononuclear infiltrates and fibrinoid necrosis of small leptomeningeal and parenchymal vessels. Granulomas are present in approximately 85% of patients. It is rare, with an incidence of less than 1:2 million. Headache is the most common presenting feature and is present in two-thirds of patients. Patients may also present with a diffuse encephalopathy and, less frequently, with a focal stroke. The diagnosis should particularly be considered when a patient presents stroke on a background of encephalopathy.

There is typically some lymphocytic pleocytosis and often an elevated cerebrospinal fluid (CSF) protein, but the CSF may be normal. CNS vasculitis is sometimes associated with infections, including mycoplasma, varicella-zoster, arboviruses, HIV, fungal, syphilis, and Lyme disease. It is also sometimes associated with lymphoma, malignant histiocytosis, and hairy cell leukemia.

CNS vasculitis may also be part of a systemic vasculitis, such as polyarteritis nodosa, Churg–Strauss syndrome, Wegener's granulomatosis, giant cell arteritis, hypersensitivity vasculitis, Behest's disease, systemic lupus erythematosus, Sjogren's syndrome, sarcoidosis, rheumatoid arthritis, scleroderma, mixed connective tissue disease, dermatomyositis, ulcerative colitis, and celiac disease.

Other strokes may be caused by paradoxical emboli that form in leg veins or in pelvic veins and travel through a patent foramen ovale or through a pulmonary arteriovenous malformation.

Strokes can also be caused by arterial dissection, in which the inner part of the arterial wall lifts up like a flap from the outer wall. The flap may occlude the artery or clot may form between the flap and the exposed inner surface of the outer wall, and pieces of this clot may embolize to block more distal arteries. Dissections of the carotid or vertebral arteries are the most likely to cause stroke, but aortic dissections can also cause stroke if they extend into origins of the carotids or vertebrals or if embolization occurs. In addition, dissections can also involve the large intracranial vessels. Dissections are often caused by trauma, although for some patients no definite precipitant can be identified. Certain connective tissue abnormalities, particularly Ehlers–Danlos syndrome

type IV and Marfan's syndrome, are associated with an increased risk of dissection. Dissections often heal with antithrombotic therapy, and the dissected artery may return to having normal or nearly normal flow as the flap reattaches to the outer wall.

Although most ischemic strokes result from a blocked artery, thrombosis of cerebral veins or sinuses can occasionally lead to a stroke. Such thromboses are more common in people who are hypercoagulable – that is, those who have an underlying tendency of the blood to clot excessively. Hypercoagulability may contribute to arterial strokes, but it seems to be more important in venous strokes.

Among the factors that are usually tested when hypercoagulability is searched for are anti-cardiolipin antibodies, PTT, protein C, protein S, factor V Leiden, and antithrombin III. There are also tendencies toward hypercoagulability late in pregnancy, in the postpartum period, and in association with systemic malignancy.

Despite intensive evaluation to classify ischemic strokes into one of these groups, it is not possible to assign every stroke to one of the categories, and most studies find that the cause cannot be identified definitively for approximately 30% of patients. As more potential risk factors are identified, it is anticipated that the number of cryptogenic strokes will decrease.

Stroke Syndromes

The symptoms that a stroke causes depends on its location in the brain. The most common finding in stroke is weakness on one side of the body. Weakness may involve the face, arm, or leg separately or in combination. For example, strokes affecting the portion of the posterior frontal primary motor cortex in the middle cerebral artery territory preferentially cause weakness of the face and arm and, to a lesser extent, weakness of the leg, which is sometimes spared entirely. In contrast, strokes involving the anterior cerebral artery preferentially affect the leg, and strokes that damage the internal capsule or pons where all motor fibers are close together tend to involve the face, arm, and leg equally.

Strokes affecting the language areas, which are in the frontal, temporal, and parietal regions around the Sylvian fissure and are usually in the left hemisphere, produce different forms of aphasia. Strokes affecting the corresponding areas in the right hemisphere tend to produce problems with attention and left-sided neglect. Strokes affecting the occipital cortex and the optic radiations produce partial or complete visual field defects. Strokes affecting the cerebellum and the fibers connecting with it tend to produce ataxia. Similar ataxia sometimes also results from damage to sensory areas in the parietal lobe. Double vision occurs in strokes that affect the brain stem. Slurred speech and swallowing difficulties are common after strokes in many areas of the brain.

Imaging Evaluation of Stroke

The initial tool of evaluation of stroke patients is CT scan. CT does not generally provide as much detail as MRI, but it is faster than MRI and generally more accessible. It is easier to monitor patients during a CT scan than during an MRI. The main disadvantage of CT is that acute strokes are generally not visualized and may not become visible until the infarcted tissue becomes sufficiently edematous or necrotic, and some small strokes are never seen on CT. In contrast, MRI is extremely sensitive for small acute strokes when diffusion-weighted imaging is used, which can detect acute strokes within 1 h from onset.

CT and MRI can also be used to visualize the arterial supply of the brain from the aortic arch to the second- or third-order branches inside the skull in procedures known as CT angiography (CTA) and MR angiography (MRA). These tests help to identify any fixed stenoses that may have caused the patient's stroke and may put the patient at risk for another stroke.

Carotid ultrasound and transcranial Doppler tests are also used to assess blood flow in the arteries that supply the brain. They provide information that complements the results of CTA or MRA.

It is also important to search for cardiac causes of stroke. In this regard, most stroke and TIA patients are studied by transthoracic echocardiography to search for intracardiac thrombus, to assess left ventricular function, and to search for valvular abnormalities. Telemetry or Holter monitoring are used mainly to search for atrial fibrillation, which is the leading cause of cardioembolic stroke, but also to monitor for other arrhythmias because many stroke patients also have cardiac disease.

Laboratory Evaluation of Stroke

Blood tests that are important specifically for risk factor evaluation include a fasting lipid panel and fasting glucose. Hemoglobin A1c levels, which measure the glycosylation of hemoglobin and provide a measure of the average blood glucose over the approximately 60-day half-life of red blood cells, may also be advisable in some patients with diabetes mellitus or suspected diabetes. Blood cell counts are also important because extremely high red cell counts may make blood viscous and predispose to thrombus formation. Elevated platelet counts

may also lead to increased risk of thrombosis. Coagulation parameters are also assessed routinely in stroke patients.

Acute Ischemic Stroke Treatment

The only approved treatment for acute stroke in the United States is intravenous tissue plasminogen activator (TPA), also known as alteplase. It is approved for use within the first 3 h after the patient was last known to be at baseline. A study sponsored by the National Institutes of Health demonstrated that 34% of patients treated with TPA had excellent outcomes compared to only 21% of control patients ($p < 0.002$). The benefits come with an increased risk of hemorrhage. A total of 6.4% of TPA-treated patients had a symptomatic intracerebral hemorrhage within the first 36 h compared to only 0.6% of the control patients ($p < 0.001$); however many of these hemorrhages occur in patients with large strokes who would do poorly.

Other options for acute treatment of stroke are available at specialized centers. Intra-arterial delivery of thrombolytic drugs such as TPA can sometimes dissolve clots that are resistant to intravenous TPA, and this is used for patients who have not responded to intravenous TPA or are outside of the 3 h window but still within 6 h from the time when they were last normal.

Sometimes, special corkscrew-like catheters can be inserted into a blocked artery and used to pull the clot out. These devices are used up to approximately 8 h after onset. Another intra-arterial option for opening occluded or stenotic arteries is angioplasty, which is usually followed by stenting.

At many centers, decisions to proceed with intra-arterial procedures are made only after MRA and CTA, and frequently these studies are performed in combination with perfusion imaging to measure delivery of blood to the brain parenchyma. Perfusion studies can identify tissue that is at risk but not yet infracted. Development of algorithms to use perfusion and other imaging data to identify those most likely to benefit from intra-arterial procedures is the subject of ongoing research.

Another option for patients who are not eligible for intravenous thrombolysis or for intra-arterial procedures is the use of hypertensive therapy. Its rationale derives from the existence of tissue at risk in the periphery of an infarct. The goal is to improve perfusion of this tissue by raising blood pressure with hydration and/or pressers. Anecdotal evidence suggests that this approach can stabilize patients with fluctuating deficits that worsen when blood pressure is lower.

In addition, some very large strokes may produce dangerous amounts of swelling that compress adjacent parts of the brain and necessitate surgical decompression.

Neuroprotection for Acute Ischemic Stroke

Numerous drugs have been found to reduce the size of acute ischemic strokes in animals by so-called neuroprotectant mechanisms that interfere with the generation of toxins as a result of the initial ischemic injury. The basic premise behind neuroprotection is that ischemia leads to a cascade of events that results in increases in intracellular calcium, which triggers the release of excitatory neurotransmitters that in turn lead to generation of more intracellular calcium. Increases in intracellular calcium lead to generation of free radicals and activation of proteases and lipases that cause additional damage. The processes of necrosis and apoptosis (programmed cell death) are triggered and lead to damage to additional cells.

Unfortunately, a growing list of clinical trials have failed to demonstrate any benefit for stroke patients from neuroprotectants. There are several possible reasons for this failure. Some trials have had overly long windows for treatment so that patients had already suffered too much irreversible damage for the drugs to work. In addition, human strokes are heterogeneous and may not all respond to the same neuroprotectants, but trials have generally included all types of stroke; for example, patients with strokes primarily involving white matter have been included in trials of drugs acting at synapses in gray matter structures. In addition, most drugs have been tested only in rodents with small brains, which may differ in important ways from the human brain. Also, studies have included patients with severe deficits who simply may have had too much damage to be helped by any treatment and other patients with mild deficits who recover too well without treatment for an effect to be detectable. Additional trials with other neuroprotectants are still in progress, and their designs have been modified because of the experiences with previous unsuccessful trials.

Secondary Prevention of Ischemic Stroke

After the acute event, one major goal of treatment is to reduce the risk of recurrent stroke and other ischemic events. The main therapeutic approaches are antithrombotic drugs and treatment of risk factors that increase stroke risk. With regard to antithrombotic

therapy, the two options are anticoagulation – usually with warfarin – and antiplatelet therapy. Numerous studies have shown that warfarin is the treatment of choice in patients who have atrial fibrillation. Other cardiac sources of emboli are sometimes also treated with warfarin. For other stroke patients, antiplatelet therapy is used. The Warfarin–Aspirin Recurrent Stroke Study showed several years ago that warfarin is not better than aspirin for stroke patients who do not have a definite cardioembolic explanation for their symptoms, and there is a significant increase in the rate of bleeding with warfarin.

Options for antiplatelet therapy include aspirin at doses of 81–325 mg day^{-1}, clopidogrel 75 mg day^{-1}, and the combination of aspirin 25 mg and dipyridamole combination 200 mg twice daily. Clinical trial data suggest that the aspirin/dipyridamole significantly reduces stroke risk compared to aspirin alone. There is also evidence that clopidogrel is slightly more effective than aspirin. An ongoing trial is comparing the combination of aspirin and dipyridamole to clopidogrel. Currently, there is no solid data to rely on in choosing among these alternatives.

The other important issue in stroke prevention is rigorous control of risk factors. Hypertension, hyperlipidemia, and diabetes mellitus need to be treated aggressively, and patients who smoke must be counseled about techniques for smoking cessation. Throughout the years, the goals of therapy have been progressively tightened. Systolic blood pressures below 160 were once considered acceptable; the usual standard now accepts blood pressures below 140/90, but there is evidence that especially for diabetic patients, there is added benefit in getting the pressure below 130/80, and systolic pressures below 120 may confer additional protection. Similarly, the target for low-density lipoprotein (LDL) cholesterol has been lowered; in recent years, guidelines have suggested keeping the LDL below 130, but data suggest a target of 100 for patients with significant atherosclerosis, including many stroke patients, and there is evidence from the coronary literature that plaque may actually regress if the LDL is below 70.

Rehabilitation after Stroke

Another goal of treatment after stroke is optimization of recovery. In this regard, the first step is the prevention of medical complications, such as infection, deep venous thrombosis, and bedsores. Intervention by trained therapists can help to reduce the risk of these problems and also help to maximize recovery of function. Depending on the deficits of stroke patients, they may need physical and occupational therapy, speech therapy, and swallowing therapy. Rehabilitation has long been thought to benefit stroke patients by helping them to regain strength and learn to compensate for their deficits, but until recently there has been little scientific foundation for rehabilitation.

Some justification for rehabilitation therapy comes from functional imaging studies that have visualized changes in brain activity as patients recover from a stroke. Additional areas that are not normally used for simple motor and language tasks are activated in stroke patients. In some cases, activity patterns tend to return more toward normal as patients recover. These changes in neuronal activity over time suggest that even the adult brain has some degree of plasticity and they provide motivation for developing therapeutic modalities that enhance recovery.

One promising approach is constraint-induced therapy, in which patients who are partially paralyzed on one side receive intensive physical therapy for the affected arm while the unaffected arm is immobilized to force them to use the affected limb more. Another new approach is the use of robotic devices that can guide a weakened limb in practice sessions. This may aid in recovery by helping to teach the movement to a patient in a different way from the help that a therapist gives and also by simply providing more repetitions of a task than therapists have time to do in standard therapy.

Current research is also investigating the role of stimulant drugs in promoting recovery when given prior to therapy sessions. Such drugs are being studied clinically and in experimental models of stroke. Growth factors that may enhance formation of new neuronal connections are also being studied in experimental models.

Finally, ongoing research is also studying electrical stimulation of the brain with implanted electrodes or with scalp electrodes as an adjuvant to physical therapy.

See also: Animal Models of Stroke; Cerebrovascular Disease; Exercise in Neurodegenerative Disease and Stroke; Map Plasticity and Recovery from Stroke; Stroke: Injury Mechanisms; Stroke: Neonate vs. Adult; Vascular Issues in Neurodegeneration and Injury.

Further Reading

Adams HP Jr, del Zoppo G, Alberts MJ, et al. (2007) Guidelines for the early management of adults with ischemic stroke. *Stroke* 38: 1655–1711.

Bougousslavsky J and Caplan LR (eds.) (2001) *Stroke Syndromes*, 2nd edn. Cambridge, UK: Cambridge University Press.

Bougousslavsky J and Caplan LR (eds.) (2001) *Uncommon Causes of Stroke*. Cambridge, UK: Cambridge University Press.

Caplan LR (2000) *Caplan's Stroke*, 3rd edn. Boston: Butterworth-Heinemann.

Mohr JP, Choi DW, Grotta JC, Weir B, and Wolf PA (eds.) (2004) *Stroke: Pathophysiology, Diagnosis, and Management*, 4th edn. Philadelphia: Churchill Livingstone.

National Institute of Neurological Disorders and Stroke, rt-PA Stroke Study Group (1995) Tissue plasminogen activator for acute ischemic stroke. *New England Journal of Medicine* 333: 1581–1587.

Sacco RL, Adams R, Albers GR, et al. (2006) Guidelines for prevention of stroke in patients with ischemic stroke or transient ischemic attack. *Stroke* 37: 577–617.

Relevant Websites

http://www.strokeassociation.org – American Stroke Association.
http://www.strokecenter.org – Internet Stroke Center.
http://www.stroke.org – National Stroke Association.

Stroke: Injury Mechanisms

A B Singhal and E H Lo, Harvard Medical School, Boston, MA, USA
J M Ren and S P Finklestein, Biotrofix, Inc., Needham, MA, USA

© 2009 Published by Elsevier Ltd.

Introduction

Stroke is a clinical syndrome of sudden-onset neurological deficits resulting from vascular etiology. Ischemic stroke accounts for 80% of all strokes; hemorrhagic stroke accounts for the remaining 20%. Brain injury after stroke results from sudden cessation of blood flow, leading to impaired delivery of oxygen, glucose, and other nutrients to brain cells, triggering a cascade of ultimately lethal cellular events.

Several mechanisms of stroke-induced cell death have been delineated, the most important of which include glutamate excitotoxicity, acidotoxicity, oxidative/nitrative stress, apoptosis, inflammation, and peri-infarct depolarization. These fundamental pathways are interdependent, overlapping, and evolve over minutes, hours, or days to mediate injury in neuronal cells, glial cells, and vascular structures. The individual contribution of individual mechanisms toward the net stroke-related injury (Figure 1) is dependent on the degree and duration of ischemia, the timing of reperfusion, the type of brain tissue affected, physiological variables such as brain temperature, and genetic influences. For example, excitotoxicity has a dominant role in the 'core' of the ischemic territory where the reduction in blood flow is severe (less than 20% of baseline, typically 10–12 ml $100\,g^{-1}\,min^{-2}$). In peripheral zones termed the 'ischemic penumbra,' where the blood flow reduction is abnormal but not as severe (20–60 ml $100\,g^{-1}min^{-2}$), cell death occurs relatively slowly via mechanisms such as inflammation, apoptosis, and peri-infarct depolarization. In these penumbral regions, brain electrical activity ceases, but the cells remain viable for an extended time period, constituting a target for stroke therapy.

In addition to the severity of ischemia, stroke-related injury is dependent on the type of brain tissue involved. Gray matter and white matter have different susceptibilities to ischemic injury. Among neurons, CA1 hippocampal pyramidal neurons, cortical projection neurons in layer 3, neurons in dorsolateral striatum, and cerebellar Purkinje cells are particularly susceptible. Among cell types comprising the white matter, oligodendrocytes are more vulnerable than are astroglial or endothelial cells. White-matter ischemia is typically severe, with rapid cell swelling and tissue edema because there is little collateral blood supply in deep white matter. Cell death mechanisms such as excitotoxicity have a different pathophysiology in white-matter ischemia, as compared to gray-matter ischemia. It is important to note that 'global' cerebral ischemia (which occurs after cardiac arrest) and 'focal' cerebral ischemia (which occurs due to thrombosis or embolism to a brain artery) result in different histological and behavioral outcomes and have different molecular mechanisms and treatments.

While restoration of blood flow can salvage penumbral brain regions, pharmacological interruption of one or more pathways of cell death can also mitigate stroke-related brain injury. Indeed, a major goal of uncovering pathways of cellular injury is to uncover targets for treatment. Numerous drugs have proved effective in reducing infarct size and functional deficits in animal stroke models. Although no neuroprotective drug to date has proved efficacious in humans, much knowledge has been gained. There is now a better understanding of factors such as the therapeutic time window and optimal patient selection, which are critical for success. Rapid advances in neuroimaging have made it possible to estimate the degree of ischemia and quantify penumbral tissue within minutes after symptom onset. These advances raise hope that successful stroke neuroprotection is imminent. In the following sections we review the major mechanisms of neuronal injury after stroke, emphasizing targets for intervention.

Excitotoxicity and Ionic Imbalance

Glutamate, the main excitatory neurotransmitter, promotes the activity of virtually every neuron within the brain and spinal cord. It is synthesized within neurons, stored in presynaptic vesicles, released in a calcium-dependent manner to bind with specific receptors, and inactivated after release by high-affinity reuptake mechanisms. Glutamate flux is tightly regulated by its receptors and transporters. There are two categories of glutamate receptors: ionotropic *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoxalleprionate (AMPA), and kainate receptors, and metabotropic glutamate receptors (mGluRs). In the context of stroke-related injury, NMDAR and AMPAR activation mechanisms have been extensively studied.

Stroke-induced reductions in oxygen and glucose supply lead to the failure of energy-dependent processes such as that involving Na^+, K^+-ATPase,

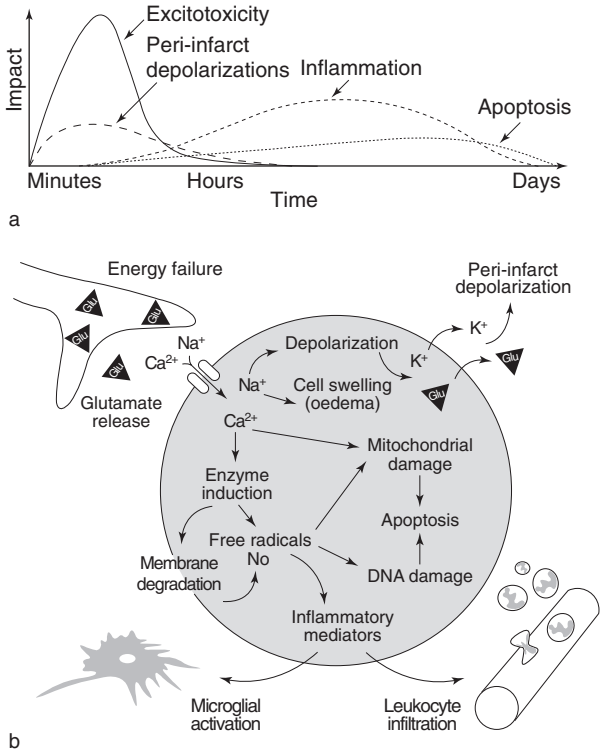

Figure 1 Contribution of individual mechanisms toward the net stroke-related injury.

processes which normally maintain cellular homeostasis. This causes an abrupt cessation in the exchange of Na^+, K^+, and other ions across cell membranes; release of excitatory neurotransmitters such as glutamate; and inhibition of glutamate-reuptake mechanisms. Excessive glutamate binding to ionotropic NMDA and AMPA receptors promotes the deleterious influx of calcium into the cytoplasm and mitochondria, with activation of downstream phospholipases and proteases that degrade membranes and proteins.

Ca^{2+} influx leads to activation of intracellular proteases such as calpains, which are involved in pro-death pathways such as those involving free radical-mediated mitochondrial damage, activation of Bad and subsequent release of cytochrome *c*, and cleavage of the Na^+/Ca^{2+} exchanger, leading to further Ca^{2+} influx. Calpains activate several caspases that in turn promote apoptosis, demonstrating the overlap between excitotoxic and apoptotic cell death mechanisms. Glutamate excitotoxicity and calcium influx also directly or indirectly influence intracellular signaling pathways. Three mitogen-activated protein kinase (MAPK) signaling pathways have been characterized: the extracellular signal-related kinase (ERK) MAPK pathway, the p38 MAPK pathway, and the c-jun N-terminal kinase (JNK) MAPK pathway. Growth factors, cytokines, and free radicals released after stroke all stimulate the MAPK pathways, providing further evidence for the overlap between the different fundamental pathways. ERK pathway activation is known to be neuroprotective; however, emerging data show that the ERK pathway is involved in mitochondrial cytochrome *c* release, activates caspase-3, and interacts with mGluRs to make them more susceptible to excitotoxic injury.

Because the endoplasmic reticulum (ER) participates in redistributing Ca^{2+} ions within the cells, Ca^{2+} influx also contributes to ER stress, which further exacerbates cell death. With regard to the mitochondria, Ca^{2+} overload leads to the opening of the nonspecific mitochondrial permeability transition pore, which allows protons and other small molecules to cross the mitochondrial membrane, resulting in ATP depletion and cell death via necrosis or apoptosis. Whether necrosis or apoptosis dominates is determined by the duration of pore opening. If the pore reseals quickly, ATP production may be rescued, resulting in apoptotic death via cytochrome *c* release rather than necrotic cell death. In addition to the aforementioned processes, glutamate-dependent processes include the activation of transcription factors cAMP response element-binding protein (CREB) and nuclear factor-kappa B (NF-κB), trophic factor signaling, and regulation of Bcl-2. Ionotropic glutamate receptors also promote ionic homeostasis perturbations that play a critical role in cerebral ischemia. Glutamate receptor activation is not uniformly deleterious, since different subtypes of receptors have prosurvival and proapoptotic effects. As explained next, this leads to a new level of complexity in finding suitable pharmaceutical agents for stroke.

Pharmaceutical agents that block the glutamate, NMDA, AMPA, and metabotropic subfamily receptors have been shown to attenuate infarct volumes in animal stroke models. However, these approaches were effective within a narrow 1 h time window, and a short therapeutic time window might explain the failure of several glutamate antagonists to show efficacy in clinical trials. Drugs that have failed clinical testing include lubeluzole, the competitive NMDA-antagonist selfotel, and several noncompetitive NMDA antagonists (dextrorphan, gavestinel, aptiganel, and eliprodil). The failure of these drugs is also attributable to drug toxicity, possibly due to inhibition of excitatory neurotransmission, and the fact that the NMDA receptor has variable functions depending on its location (synaptic or extrasynaptic), so that nonselective blockade of NMDA receptors may not be fruitful. For example, the activation of synaptic NMDA receptors promotes cell survival via mechanisms such as inactivation of the proapoptotic protein Bad, and the activation of CREB and expression of its target genes such as brain-derived

neurotrophic factor (BDNF). Activation of extrasynaptic NMDA receptors causes inactivation of the CREB pathway and counters the increase in BDNF expression, leading to a proapoptotic effect. Therefore, the challenge for the future is to uncouple the essential excitatory neurotransmitter functions from deleterious effects – for example, by blocking downstream neurotoxic signaling pathways without blocking the receptors.

Acidotoxicity

As compared to cell types in other organs, neurons are more vulnerable to ischemic injury. Experimental research has shown that the metabolic acidosis accompanying stroke can exacerbate neuronal death via excitotoxic mechanisms that are distinct from glutamate- and Ca^{2+}-mediated excitotoxicity. Neuronal acid-sensing ion channels (ASICs) have recently been characterized in the brain as ligand-gated channels belonging to the epithelial Na^{+} channel superfamily. ASICs have at least six subunits, which form voltage-insensitive, amiloride-sensitive channels on the surface membrane. In contrast to other neurotransmitter-receptor channels that are inactivated with acidosis, these multimeric channels become activated when pH falls below 7.0, and their activation is half-maximum at a pH of 6.2, which is consistently achieved within ischemic regions. ASIC activation is also increased by arachidonic acid and lactic acid production, as well as membrane stretching and decreased extracellular Ca^{2+} (all processes that routinely occur after stroke). The major role of ASICs is to gate amiloride-sensitive Na^{+} currents. However, some subunits facilitate a glutamate-independent influx of Ca^{2+} in the presence of acidosis. In experiments utilizing cell culture, patch clamp, and *in vivo* rodent stroke models, the tarantula toxin PcTX1 (an ASIC1a-specific antagonist) and amiloride have been shown to inhibit excitotoxic Ca^{2+} currents via the ASIC1a ligand-gated ion channel. Moreover, ASIC1a channel inhibition has been documented to decrease infarct volumes by over 60%, acid-mediated toxicity is shown to be absent in mice lacking the ASIC1 gene, and ASIC transfected cells show increases in intracellular Ca^{2+} following acidosis. These and other data indicate that ASIC1 causes a significant acid-dependent increase in excitotoxic cell death and that the ASIC1a subunit represents a promising stroke therapeutic target.

Apoptosis

Apoptosis is an evolutionarily conserved process of programmed cell death. It is characterized histologically by nuclear and cytoplasmic condensation, nuclear fragmentation, and cell shrinkage. Apoptotic cells are terminal deoxynucleotidyl transferase-mediated dUTP nick-end labeling (TUNEL)-positive cells that exhibit DNA laddering. Cell type, age, and brain location render them more or less resistant to apoptosis or necrosis. Since apoptotic pathways require energy in the form of ATP, apoptosis predominantly occurs in the ischemic penumbra that sustains milder injury, rather than in the core, where ATP levels are rapidly depleted. Nevertheless, ionic imbalances can also trigger apoptosis under certain conditions, suggesting a close mechanistic interrelationship between glutamate-mediated excitotoxicity and apoptosis. The normal human brain expresses caspases 1, 3, 8, and 9, apoptosis protease-activating factor-1 (APAF-1), death receptors, P53, and a number of Bcl-2 family members, all of which are implicated in apoptosis.

Caspases are protein-cleaving enzymes (zymogens) that belong to a family of cysteine aspartases constitutively expressed in adult and newborn brain cells, particularly neurons. Apoptosis occurs via caspase-dependent as well as caspase-independent mechanisms. Caspases are sequentially cleaved and activated via two pathways: an extrinsic pathway dependent on death receptors, and an intrinsic mitochondrial pathway. With regard to the extrinsic pathway, several apoptotic triggers of have been identified. These include oxygen free radicals, Bid cleavage, death receptor ligation, DNA damage, and lysosomal protease activation. The tumor necrosis factor (TNF) superfamily of death receptors, most importantly Fas, powerfully regulates upstream caspase processes. Emerging data suggest that the cell nucleus is involved in releasing signals for apoptosis, and that the mitochondrion plays a central role in mediating apoptosis. At least four mitochondrial molecules mediate downstream cell death pathways: cytochrome *c*, secondary mitochondria-derived activator of caspase (Smac/Diablo), apoptosis-inducing factor (AIF), and endonuclease G. While cytochrome *c* and Smac/Diablo mediate caspase-dependent apoptosis, AIF and endonuclease G mediate caspase-independent apoptosis. Cytochrome *c* binds to Apaf-1, which combines with procaspase 9 to form the 'apoptosome' that activates caspase 9 in the presence of dATP. In turn, caspase 9 activates caspase 3. It is important to note that formation of the apoptosome complex is inhibited by Bcl-2, which becomes upregulated very early after stroke. Smac/Diablo binds to inhibitors of activated caspases and causes further caspase activation. After activation, caspases 3 and 7 (the executioner caspases) attack and degrade various substrate proteins, ultimately leading to DNA fragmentation and irreversible cell death. Over 30

proteins can be cleaved, including the DNA-repairing enzyme poly(ADP-ribose) polymerase (PARP), the cytoskeletal protein gelsolin, actin, presenilins, and other caspases. Caspase 3, the most prominent cysteine protease, is present in the infarct core as well as in the penumbra and is cleaved acutely after stroke. Subsequent caspase cleavage occurring after hours or days leads to delayed cell death.

Caspase-independent apoptosis has been shown to develop in cultured neurons by NMDA-receptor induced activation of PARP-1, and in fibroblasts exposed to oxidative stress. AIF is implicated as a key molecule in this cascade. AIF is released from mitochondria after PARP-1 activation, then relocates to the nucleus, where it binds DNA, promotes chromatin condensation, and leads to cell death via incompletely understood mechanisms. The exact significance of caspase-independent apoptosis remains to be determined.

Overexpression of Bcl-2, deletions of genes for Bid or caspase-3, and the use of caspase-3 inhibitors, peptide inhibitors, and antisense oligonucleotides that suppress the expression and activity of apoptosis-promoting genes have been shown to reduce infarct volume or decrease neurological deficits after stroke. A single intracerebroventricular injection of zDEVD-FMK, a relatively selective caspase-3 inhibitor, is shown to be neuroprotective when administered up to 9 h after transient focal stroke. However, caspase inhibitors do not reduce infarct size in all brain ischemia models. This variable effect may be attributable to different severities of ischemia, to limited potency or inability of caspase inhibitors to cross the blood–brain barrier, to the relatively minor impact of apoptosis as compared to excitotoxicity or other mechanisms on stroke outcome, and to upregulation of caspase-independent or redundant cell death pathways.

Oxidative and Nitrosamine Stress

Free radicals subserve an important role in virtually all fundamental cell death pathways. Under resting conditions, mitochondria produce reactive oxygen species, such as superoxide and hydroxyl radicals, during electron transport. Oxidative and nitrosamine stresses are routinely modulated by endogenous antioxidant enzyme systems such as superoxide dismutase (SOD) and the nitric oxide synthase (NOS) family. Moreover, toxic free radicals are normally scavenged by SOD, catalane, and glutathione (GSH), and antioxidant vitamins (e.g., α-tocopherol and ascorbic acid). Ischemia-induced accumulation of intracellular Ca^{2+}, Na^{+}, and ADP stimulates excessive mitochondrial oxygen free radical production, which overwhelms endogenous scavenging mechanisms. Reoxygenation during reperfusion provides large amounts of oxygen as a substrate for numerous enzymatic oxidation reactions in the cytosolic compartments, subcellular organelles, and mitochondria. There is a burst of oxygen free radicals from enzymatic conversions, including the cyclooxygenase-dependent conversion of arachidonic acid to prostanoids and degradation of hypoxanthine, the superoxide-mediated reduction of iron, and the generation of nitric oxide. Substantial amounts of free radicals are also generated during the inflammatory response after ischemia.

Oxygen radicals and oxidative stress facilitate mitochondrial transition pore (MTP) formation, which dissipates the proton-motive force required for oxidative phosphorylation and ATP generation. As a result, mitochondria release apoptosis-related proteins and other constituents within the inner and outer mitochondrial membranes. Free radical-mediated mechanisms signal the release of mitochondrial cytochrome *c* by mechanisms probably related to Bcl-2 and translocation of Bax. After its release, cytochrome *c* binds to Apaf-1 followed by caspase-9 to form the apoptosome complex that subsequently activates caspase-3 and other caspases. Activated caspase-3 goes on to cleave several nuclear DNA repair enzymes, culminating in apoptosis. Another major signaling pathway of oxygen radicals involves NF-κB, an important transcription factor. NF-κB activation results in its translocation into the nucleus and binding to the NF-κB site of inducible genes such as cyclooxygenase-2 (COX-2), inducible nitric oxide synthase (iNOS), matrix metalloproteinases (MMPs), intercellular adhesion molecules (ICAM-1), and cytokines. Furthermore, after ischemia, the activation of NMDA receptors and formation of oxygen free radicals and nitric oxide by neuronal nitric oxide directly signal the mitochondrial release of cytochrome *c* or formation of peroxynitrite. Subsequent hydroxyl radical production results in lipid peroxidation and directly damages proteins, carbohydrates, and DNA, ultimately leading to necrotic cell death.

The significant contribution of free radical-mediated injury is apparent from studies showing reduced stroke-related tissue injury in mice overexpressing SOD, and in mice with reduced expression of the neuronal and inducible NOS isoforms. Treatment with free radical scavengers such as *N-tert*-butyl-α-phenylnitrone (α-PBN) is also shown to attenuate stroke-related injury. However, in clinical trials, free radical scavengers such as NXY-059 have failed to show any benefit.

Inflammation

Inflammation is intricately related to atherosclerosis and stroke onset, as well as subsequent stroke-related tissue damage. Elevated stroke risk has been linked to high blood levels of inflammatory markers, such as C-reactive protein, interleukin-6 (IL-6), TNF-α, and soluble intercellular adhesion molecule (sICAM). Arterial thrombosis associated with plaque ulceration is triggered by multiple processes involving endothelial activation and proinflammatory and prothrombotic interactions between the vessel wall and circulating blood elements. Ischemic stroke-related brain injury itself triggers inflammatory cascades within the parenchyma that further amplify tissue damage. Within a few hours after the onset of ischemia, the expression of several cytokines (including TNF-α and IL-1β) is upregulated in brain. By 24 h, brain microglia become activated. Adhesion molecules (including ICAM, and P- and E-selectins) are expressed in vascular endothelial cells, facilitating the attraction, rolling, and entering of blood-borne inflammatory cells into the ischemic brain – first neutrophils, at 24–48 h after ischemia, followed by lymphocytes and macrophages, peaking at 5–7 days after ischemia. Lymphocytes express inflammatory cytokines and chemokines (including TNF-α and IL-8), while macrophages participate in the scavenging of necrotic cell debris. The enzymes iNOS and COX-2 are also upregulated after cerebral ischemia, contributing to the formation of nitric oxide and oxygen free radicals.

There is evidence that the inflammatory cascade occurring after cerebral ischemia contributes to acute and delayed cell death. Blocking inflammatory cell infiltration or blocking or deleting the genes for various cytokines, adhesion molecules, or other inflammatory mediators reduces cell death in animal models of cerebral ischemia. On the other hand, infiltration of blood-borne cells following ischemia likely includes hematopoietic and other stemlike and progenitor cells, which may play a role in the repair of brain injury following stroke.

Peri-infarct Depolarizations

Research in mouse, rat, and cat stroke models has shown that neurons and astrocytes in ischemic penumbral regions exhibit waves of spreading depression, detected as spreading negative slow-voltage variations. These waves, termed peri-infarct depolarizations (PIDs), exacerbate tissue damage due to the increased energy requirements for reestablishing ionic equilibrium in metabolically compromised ischemic tissues. In the initial 2–6 h after experimental stroke, PIDs result in a stepwise increase of core-infarcted tissue into adjacent penumbral regions. The frequency and total duration of spreading depression correlate with infarct size. Recent evidence suggests that PIDs contribute to the expansion of the infarct core throughout the period of infarct maturation. The significance of PIDs to human stroke pathophysiology is unclear; however, a recent study of patients with subarachnoid hemorrhage showed strong correlation between periods of spreading electrocorticographic depression and delayed ischemic neurological deficits and brain infarcts. These results suggest that spreading depolarizations with prolonged depressions may be a promising target for stroke treatment with pharmaceutical agents such as NMDA antagonists, or with physiological approaches such as hypothermia and hyperoxia.

Contemporary Views on Stroke Pathophysiology: A Clinical Perspective

While each step in every cell death pathway provides an opportunity for pharmaceutical intervention, no drug has proved effective in clinical trials. Agents that have failed clinical testing include the lipid-peroxidation inhibitor tirilazad mesylate, the ICAM-1 antibody enlimomab, the calcium channel blocker nimodipine, the γ-aminobutyric acid (GABA) agonist clomethiazole, the glutamate antagonist and sodium channel blocker lubeluzole, the competitive NMDA antagonist selfotel, and noncompetitive NMDA antagonists such as dextrorphan, gavestinel, aptiganel, and eliprodil. The lack of efficacy can be related to diverse factors, including inadequate preclinical studies, inclusion of patients without evidence for salvageable tissue on brain imaging, and deficiencies in clinical trial methodology, particularly the use of crude 'global' outcome measures rather than outcome measures designed to assess specific modalities of neurological function, which may be more sensitive. From a basic science standpoint, the failure of clinical trials has exposed the shortcomings of traditional approaches that relied on targeting a single cell death pathway.

The overlapping features and interactions between pathways are now better appreciated, and newer approaches have focused on combining or adding drugs in series so as to target distinct pathways during the evolution of ischemic injury. Neuroprotective combinations that have shown success in animal models include the coadministration of an NMDA receptor antagonist with GABA receptor agonists, citicoline, free radical scavengers, cyclohexamide,

and caspase inhibitors. While growth factors such as basic fibroblast growth factor (bFGF) have shown promise in promoting recovery after stroke, recent studies indicate that growth factors also exert powerful neuroprotective effects via downstream mechanisms, including upregulation of free radical-scavenging enzymes and antiapoptotic proteins. Therefore, it is not surprising that synergy is observed when combining citicoline with bFGF, and that caspase inhibitors given with bFGF or an NMDA receptor antagonist extend the therapeutic window with lower effective doses. Since early reperfusion with tissue plasminogen activator (tPA) has proved efficacy, attention is also being paid to neuroprotective strategies that decrease reperfusion injury, reduce postischemic hemorrhage, and inhibit downstream targets in cell death cascades. In animal models, thrombolysis has shown synergistic or additive effects with oxygen radical scavengers, AMPA receptor antagonists, NMDA blockers, MMP inhibitors, citicoline, topiramate, antileukocytic adhesion antibodies, and antithrombotics. Clinical trials have reported the feasibility and safety of tPA used in combination with clomethiazole or lubeluzole. However, in recent major clinical trials, the combination of the free radical scavenger NXY-059 with tPA showed no benefit.

Given the importance of targeting multiple pathways, more attention is being paid to physiological approaches such as modulation of brain temperature, optimizing blood glucose levels, improving brain oxygenation, and enhancing cerebral blood flow. While less specific, these approaches offer potential to interrupt nearly all deleterious biochemical and molecular pathways. For example, a reduction in brain temperature reduces cellular physiological processes and metabolic demands, as well as slows the rate of pathological processes such as lipid peroxidation and reduces the activity of certain cysteine or serine proteases. Hyperoxia favorably alters the levels of glutamate, lactate, Bcl-2, manganese SOD, and COX-2, and inhibits cell death mechanisms such as apoptosis. These physiological interventions are now being carefully tested in ongoing clinical trials.

Finally, it is now recognized that stroke-related injury is not simply restricted to cellular elements, but also involves the endothelial and vascular smooth muscle cells, the astro- and microglia, and the associated tissue matrix proteins – the integrated concept of the 'neurovascular unit.' Endothelial/astrocyte/matrix interactions are considered vital in protecting the integrity of brain tissue, and it is recognized that disruption of the neurovascular matrix, which includes basement membrane components such as type IV collagen, heparan sulfate proteoglycan, laminin, and fibronectin, upsets cell matrix and cell–cell signaling mechanisms that maintain neurovascular homeostasis. Emerging data show important linkages between tPA, MMPs, and brain edema and hemorrhage. The potential efficacy of MMP inhibitors in mitigating ischemic damage after stroke is being actively investigated.

Summary

The current pathophysiological understanding of stroke is substantially based on experimental studies. Brain injury after cerebral ischemia results from a complex signaling cascade that evolves in a spatiotemporal pattern and differentially affects gray and white matter. Excitotoxicity leads to acute cellular necrosis predominantly in the core of the infarction. The surrounding ischemic penumbra suffers milder insults, with excitotoxic and inflammatory mechanisms leading to delayed cell death showing characteristics of apoptosis. Research to date has led to an improved understanding of the importance, timing, and functional relationships of individual mechanisms, and to a better understanding of the differences between humans and the animal models in which these cell death mechanisms were discovered. Despite advances in the knowledge of individual injury mechanisms and despite intense efforts to translate this knowledge to effective stroke treatments, no therapeutic strategy has proved successful. The traditional 'neurocentric' approach is quickly yielding to a modern strategy of targeting neurons and glial cells, in particular astrocytes, as well as the neurovascular unit. It is anticipated that advances in basic science research and molecular neuroimaging, combined with advances in the fields of genomics and proteomics – as well as refinements in patient selection, improved methods of drug delivery, use of more clinically relevant animal stroke models, and use of combination therapies that target the entire neurovascular unit – will soon make stroke neuroprotection an achievable goal.

See also: Animal Models of Stroke; Apoptosis in Nervous System Injury; Axonal Injury in Demyelinating Disease and CNS Injury; Blood-Brain Barrier and Neurovascular Mechanisms of Neurodegeneration and Injury; Excitotoxicity in Neurodegenerative Disease; Glutamate; Inflammation in Neurodegenerative Disease and Injury; Map Plasticity and Recovery from Stroke; Mitochondrial Dysfunction in Nervous System Injury; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Oxidative Damage in Neurodegeneration and Injury; Stroke; Stroke: Neonate vs. Adult; Vascular Issues in Neurodegeneration and Injury.

Further Reading

Barone FC and Feuerstein GZ (1999) Inflammatory mediators and stroke: New opportunities for novel therapeutics. *Journal of Cerebral Blood Flow & Metabolism* 19: 819–834.

Chan PH (2001) Reactive oxygen radicals in signaling and damage in the ischemic brain. *Journal of Cerebral Blood Flow & Metabolism* 21(1): 2–14.

Cramer SC, Koroshetz WJ, and Finklestein SP (2007) The case for modality-specific outcome measures in clinical trials of stroke recovery-promoting agents. *Stroke* 38: 1393–1395.

Dirnagl U, Iadecola C, and Moskowitz MA (1999) Pathobiology of ischaemic stroke: An integrated view. *Trends in Neuroscience* 22(9): 391–397.

Endres M and Dirnagl U (2002) Ischemia and stroke. *Advances in Experimental Medicine & Biology* 513: 455–473.

Fisher M (2004) The ischemic penumbra: Identification, evolution and treatment concepts. *Cerebrovascular Disease* 17(supplement 1): 1–6.

Gidday JM (2006) Cerebral preconditioning and ischaemic tolerance. *Nature Reviews Neuroscience* 7(6): 437–448.

Graham SH and Chen J (2001) Programmed cell death in cerebral ischemia. *Journal of Cerebral Blood Flow & Metabolism* 21: 99–109.

Hossmann KA (1996) Periinfarct depolarizations. *Cerebrovascular and Brain Metabolism Reviews* 8: 195–208.

Iadecola C, Cho S, Feuerstein GZ, et al. (2004) Cerebral ischemia and inflammation. In: Mohr JP, Choi DW, Grotta JC, et al. (eds.) *Stroke: Pathophysiology, Diagnosis, and Management*, pp. 883–893. Edinburgh: Churchill Livingstone.

Kroemer G and Reed JC (2000) Mitochondrial control of cell death. *Nature Medicine* 6: 513–519.

Lo EH, Dalkara T, and Moskowitz MA (2003) Mechanisms, challenges and opportunities in stroke. *Nature Reviews Neuroscience* 4(5): 399–415.

Lo EH, Moskowitz MA, and Jacobs T (2005) Exciting, radical, suicidal: How brain cells die after stroke. Review. *Stroke* 36(2): 189–192.

Mergenthaler P, Dirnagl U, and Meisel A (2004) Pathophysiology of stroke: Lessons from animal models. *Metabolic Brain Disease* 19(3–4): 151–167.

Mitsios N, Gaffney J, Kumar P, et al. (2006) Pathophysiology of acute ischaemic stroke: An analysis of common signaling mechanisms and identification of new molecular targets. *Pathobiology* 73(4): 159–175.

Nedergaard M and Dirnagl U (2005) Role of glial cells in cerebral ischemia. *Glia* 50(4): 281–286.

Petty MA and Wettstein JG (1999) White matter ischaemia. *Brain Research – Brain Research Reviews* 31: 58–64.

Ren JM and Finklestein SP (2005) Growth factor treatment of stroke. Review. *Current Drug Targets – CNS & Neurological Disorders* 4(2): 121–125.

Singhal AB, Lo EH, Dalkara T, and Moskowitz MA (2005) Advances in stroke neuroprotection: Hyperoxia and beyond. *Neuroimaging Clinics of North America* 15(3): 697–720, xii–xiii.

Syntichaki P and Tavernarakis N (2003) The biochemistry of neuronal necrosis: Rogue biology? *Nature Reviews Neuroscience* 4: 672–684.

Zukin RS, Jover T, Yokota H, et al. (2004) Molecular and cellular mechanisms of ischemia-induced cell death. In: Mohr JP, Choi DW, Grotta JC, et al. (eds.) *Stroke: Pathophysiology, Diagnosis, and Management*, pp. 829–854. Edinburgh: Churchill Livingstone.

Stroke: Neonate vs. Adult

A M Comi and M V Johnston, Kennedy Krieger Institute, Johns Hopkins University, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Definition

Stroke is defined as brain injury occurring as a result of insufficient glucose and oxygen delivery to meet the needs of the involved brain tissue. Ischemic stroke is defined as brain injury resulting from blockage or other impairment of arterial blood delivery. The causes, molecular pathways, available treatments, and clinical outcomes are age dependant. This article focuses on contrasting neonatal arterial ischemic stroke with that in adults (see Table 1). Venous and hemorrhagic causes are not included. Neonatal arterial ischemic stroke is defined in this article as an arterial stroke presenting either (1) in the term infant during the first month of life, or (2) in the first year of life with early handedness and magnetic resonance imaging (MRI) evidence of a stroke presumed to have occurred in the neonatal period.

Epidemiology

Overall, stoke is the third leading cause of death and the leading cause of disability in the United States. Although the incidence is highest in the elderly, about one-fifth of all strokes occur in individuals younger than 65 years. Recently, data have demonstrated that neonatal stroke occurs in approximately 1 in 4000 term live births. About 33% of children with cerebral palsy have the hemiplegic form affecting the motor functions of one side of the body. In many children, the hemiplegic cerebral palsy results from a stroke in the pre-, peri-, or postnatal period. Overall, the costs of stroke are projected to be in the tens of billions of dollars per year. Stroke in the developing brain is also an important cause of neurologic impairment that often persists into adulthood.

Presentation

When recognized in the neonatal period, the most common presentation of stroke is seizures. Other less common presenting signs are lethargy or asymmetric limb movements. Other evidence of encephalopathy may also be seen, including poor suck, apnea, and hypotonia. In addition, pediatric patients with stroke often present with seizures. Children with ischemic middle cerebral artery stroke who present with seizures may be at increased risk for poor neurologic and functional outcome. When neonatal stroke presents later in the first year of life, the most common presentation is early handedness or other evidence of hemiparesis, with or without other developmental delays.

Stroke in adults is much less likely (compared with neonates) to present with acute seizures. Adult stroke is much more likely to present with the acute onset of a neurologic deficit such as a motor paresis (weakness of a limb), aphasia (inability to speak), alexia (inability to read), or apraxia (the inability to perform a previously learned motor task despite normal strength and coordination). The precise deficit in adult stroke is dependent on the location of the stroke and severity of the ischemic insult. Elegant localization of the stroke in an adult is often possible based on the constellation of neurologic deficits. This is usually not the case in neonates. Small vessel ischemia may also present with slowly evolving dementia and/or motor impairments in older adults.

Location, Etiologies, Workup, and Management

The most common site of neonatal arterial stroke is in the distribution of the middle cerebral artery and on the left side. Less commonly, neonatal ischemic stroke may occur in a watershed distribution. Stroke in the adult is more variable and depends on the age and risk factors of the individual, and whether it is secondary to large or small vessel disease or embolism.

In elderly adults, the most common causes of ischemic stroke are cardiovascular, resulting in thrombotic, lacunar, or embolic strokes that may involve the cortical, subcortical, or brain stem structures. Diagnostic workup often includes an MRI (including diffusion and perfusion) of the brain, carotid ultrasound, and cardiac echo, depending on the location and type of stroke. In adults younger than 65 years, underlying causes of ischemic stroke are generally more diverse and include thrombophilias, infection, dissection, autoimmune and inflammatory causes, and malignancy. Workup in these younger adults therefore more often includes testing such as magnetic resonance angiogram (MRA) or computerized tomography angiogram (CTA) of the neck, a thrombophilia workup, a lumbar puncture, and testing for sources of autoimmunity or cancer.

Ischemic stroke in the neonate deserves a workup as well, although a specific etiology is less often found. Inherited thrombophilias may be sought, although a recent study of candidate gene polymorphisms in genes regulating thrombosis and thrombolysis did

Table 1 Differences between neonatal and adult stroke

	Neonate	*Adult*
Frequently presents with seizures	Yes	No
Frequently presents with focal deficits	No	Yes
Sex-related differences	Yes, but different from adults	Yes, but different from neonates
Prolonged prominent apoptosis after stroke	Yes	Not so much
Oxidative reserves	Less	More
Specific treatment available	No	Yes: tissue plasminogen activator (TPA) in some cases
Recurrence risk	Very low	Generally much higher

not identify any variant allele more common in newborns with stroke than in the normal controls. An MRI/MRA with diffusion imaging is needed to determine the location and nature of the stroke and extent of brain damage. Infection should be excluded and an echocardiogram considered as well. Maternal medical issues and placental pathology also have important roles in the etiology of neonatal stroke.

Recurrence risk for ischemic stroke is related to the underlying etiology and ongoing risk factors for stroke. The recurrence risk for neonatal stroke overall is very low. It is likely that in most neonates the stroke occurs as a result of a convergence of factors related to the process of transitioning from the intranatal state to the postnatal state. As a result, these children are not likely to experience recurrence of stroke. It is not clear to what extent this recurrence risk is modified by the existence of one or more thrombophilias. Major congenital cardiovascular anomalies do carry a substantial risk for stroke recurrence. In young adults, the recurrence risk depends on whether the underlying cause is ongoing but in general is higher than in neonates. Ischemic stroke in the elderly is most often caused by multivessel disease, and overall recurrence risk is high.

The acute management of ischemic stroke in adults includes supportive measures such as preventing fevers, managing intracranial pressure and edema, and ventilation if required. Specific treatment is available in the form of tissue plasminogen activator (TPA) given either peripherally or intravascularly to break up the clot producing the ischemia. TPA is limited to the subset of patients presenting in the first few hours after stroke with no contraindication, but in this setting can be very helpful in reducing the morbidity from stroke. Aspirin is another antithrombotic treatment shown to be beneficial for the acute treatment of stroke in adults. No randomized clinical trials have addressed the acute management of perinatal stroke. Supportive care is also provided to neonates after a stroke, but anticoagulation is generally not used except for treatment of a known cardioembolic stroke.

Gender Differences

Clinical ischemic stroke has long been recognized to be sexually dimorphic in adults; most studies have shown that while women have lower stroke incidence relative to men, early outcome may be more favorable in men. This difference has been largely attributed to a protective estrogen effect in women; the risk in postmenopausal women equals that in age-matched men. Until recently, little data were available from clinical trials in adults regarding gender and response to treatment. Female sex is a mildly unfavorable prognostic factor in rehabilitation outcome after stroke.

In the pediatric population, neonatal arterial ischemic strokes are more commonly diagnosed in boys. Boys are at higher risk for all stroke types than girls. In addition, male gender may be a prognostic factor for poor outcome after stroke. Little attention, however, has been given to whether or how clinical interventions for neonatal or pediatric stroke should be tailored for patient gender.

Very recent studies suggest that sex may influence the response to treatment of ischemic brain injury. The impact of animal sex on susceptibility to brain injury has recently been examined *in vitro*. Du and colleagues studied sex-segregated neuronal cultures and found that XX and XY cultures responded differently to cytotoxic and pharmacological manipulation. They reported that XY neurons were more sensitive to nitrosative stress and excitotoxicity than XX neurons. After $ONOO^-$ treatment, XY neurons demonstrated apoptosis-inducing factor (AIF)-mediated programmed cell death, but XX neurons underwent predominantly caspase-dependent apoptosis. Sex-related differences in response to hypoxia-ischemia have also been shown experimentally *in vivo*. Results of experiments with hypoxia-ischemia in 7-day-old mice showed that knocking out Parp-1 protected males but not females. These sex-related differences could be related to differences in expression of genes known to be on the X chromosome, such as AIF or X-linked inhibitor of apoptosis (XIAP). Poly(ADP-ribose) polymerase-1 (PARP-1) is a nuclear enzyme activated by DNA strand breaks that participates in DNA repair and

replication. On the other hand, a similar experiment in adult Parp-1 knockout mice demonstrated that female adults sustained worse injury rather than less injury. Recently, Zhu et al. showed that in P9 C57/BL6J mice, neuronal death after hypoxic-ischemic injury proceeded predominantly via an AIF-mediated pathway in males versus a caspase-dependent pathway in females. These findings support the conclusion that the pathways mediating ischemic brain injury and neuroprotection may differ in males and females. This sex-related difference is relevant to both adults and neonates, but manifests differently in these two age groups.

Mechanisms of Neuronal Injury

Both adult and neonatal animal models for stroke have been developed, and the basic pathways involved are similar. Broadly speaking, a core of severely injured tissue is formed at the epicenter of the ischemia that develops into pan-necrosis. Surrounding this core is an area of underperfused tissue, the penumbra. The penumbral tissue is at risk of becoming included in the infarcted core. Cerebral ischemia triggers the release of the excitatory neurotransmitter glutamate, and the energy uncoupling produced by ischemia prevents the active removal of glutamate from the synaptic cleft. Excess glutamate overstimulates postsynaptic receptors, thereby opening the associated ion channels. This allows sodium and calcium ions to enter the cell, while potassium flows outward. Excessive calcium influx initiates a cascade of energy-consuming and ultimately lethal intracellular events.

However, *in vivo* models have revealed important differences in the immature brain as compared to the adult brain exposed to ischemic injury. In the neonatal animal, apoptosis evolves over the week following the initial ischemic insult or longer. This contrasts with a much shorter period of apoptosis following ischemic injury in the adult brain. A range of morphologies from necrosis to apoptosis is more common in the neonatal brain. The oxidative reserve of the neonatal brain, in terms of its high oxygen demand and low concentration of antioxidants, is also less compared to that in the mature animal. This in conjunction with the high concentrations of unsaturated fatty acids and bioavailable iron suggests that the neonatal brain is more susceptible to oxidative damage compared with the mature brain. In the neonate, activation of the Cl^- permeable γ-aminobutyric acid type A ($GABA_A$) receptors is excitatory and has been linked to a higher propensity for seizures. The immature brain also differs greatly from the mature brain in terms of the ion channel subunits expressed, leading to important differences in excitation and patterns of regional vulnerability to ischemic injury. These important differences make it essential that developmentally appropriate models are used for mechanistic and preclinical studies and emphasize the need for clinical trials in neonates as well as adults with stroke.

Seizures induced by global hypoxia, global ischemia, or hypoxic-ischemic injury have been noted in adult and immature rats. Studies examining interactions between seizures and hypoxic-ischemic brain injury in developing rats have produced various results. Status epilepticus induced by the GABA antagonist bicuculline after hypoxia-ischemia did not exacerbate brain damage in postnatal day 7 (P7) rats. On the other hand, in P10 rats subjected to unilateral carotid ligation and brief hypoxia, seizures induced by kainate significantly exacerbated brain injury. It is controversial and difficult to determine, however, whether seizures occur only secondary to the injury or in fact contribute to the ischemic brain injury.

Outcome

Outcome in neonatal stroke is variable. Although the majority of children survive their stroke, about 75% have sequelae including cerebral palsy. In one study, 46% of children followed after neonatal stroke developed cerebral palsy. The assessment of motor movements in the weeks following birth can be useful in the early identification of hemiparesis. A delayed presentation is associated with the developmental of cerebral palsy, as are the radiologic findings of large stroke size and injury to Broca's area, the internal capsule, Wernicke's area, or the basal ganglia. Other long-term complications include epilepsy and learning disabilities. Outcome in adults depends on the age, associated medical comorbidities, type of stroke, and recurrence risk of the individual involved but is generally less good than in neonates. Nevertheless, the potential for lifelong neurologic impairments makes stroke in the neonate an important national health issue.

Prevention

Prevention of stroke in adults is aimed at decreasing or treating the risk factors for stroke such as hypertension or hypercholesterolemia, eliminating sources of embolic clots such as the heart or carotid artery, and anticoagulation. In contrast, prevention of neonatal stroke risk factors (such as preeclampsia, extracorporeal membrane oxygenation (ECMO), maternal diabetes, chorioamnionitis, placental pathology, trauma, infection, or polycythemia) is often currently not possible and recurrence risk is low in most cases. No randomized clinical trials have been done to address the issue of primary or secondary prevention

in neonatal stroke. Maternal thromboprophylaxis for maternal thrombophilia remains speculative, and aspirin use is generally not recommended after neonatal stroke.

Future Directions

More rigorous preclinical study and clinical trial design may produce future success in neuroprotection. Other promising future approaches include enhancing the endogenous regenerative response with growth factors. A recent pilot trial with granulocyte colony-stimulating factor (GCSF), given to adults after stroke, reported that CD34+ cells were mobilized and the treatment was safe and well tolerated. Recent genetic studies have identified loci that confer susceptibility to stroke and have future application in the management of ischemic stroke patients. Advances in the treatment of stroke in adults will require careful study in developmental models and neonatal trials before they can be applied in neonates.

See also: Animal Models of Stroke; Cerebrovascular Disease; Map Plasticity and Recovery from Stroke; Stroke; Stroke: Injury Mechanisms.

Further Reading

Ashwal S, Russman BS, Blasco PA, et al. (2004) Practice parameter: Diagnostic assessment of the child with cerebral palsy: Report of the quality standards subcommittee of the American Academy of Neurology and the Practice Committee of the Child Neurology Society. *Neurology* 62: 851–863.

Chalmers EA (2005) Perinatal stroke-risk factors and management. *British Journal of Haematology* 130: 333–343.

Du L, Bayir H, Lai Y, et al. (2004) Innate gender-based proclivity in response to cytotoxicity and programmed cell death pathway. *Journal of Biological Chemistry* 279: 38563–38570.

Dzhala VI, Talos DM, Sdrulla DA, et al. (2005) NKCC1 transporter facilitates seizures in the developing brain. *Nature Medicine* 11(11): 1205–1213.

Ferriero DM (2004) Neonatal brain injury. *New England Journal of Medicine* 351: 1985–1995.

Hagberg B, Hagberg G, Beckung E, and Uvebrant P (2001) Changing panorama of cerebral palsy in Sweden. VIII: Prevalence and origin in the birth year period 1991–94. *Acta Paediatrica* 90: 271–277.

Hagberg H, Wilson MA, Matsushita H, et al. (2004) PARP-1 gene disruption in mice preferentially protects males from perinatal brain injury. *Journal of Neurochemistry* 90: 1068–1075.

Hunt RW and Inder TE (2006) Perinatal and ischaemic stroke: A review. *Thrombosis Research* 118: 39–48.

Jensen FE (2002) The role of glutamate receptor maturation in perinatal seizures and brain injury. *International Journal of Developmental Neuroscience* 20: 339–347.

Lee JL, Croen LA, Lindan C, et al. (2005) Predictors of outcome in perinatal arterial stroke: A population-based study. *Annals of Neurology* 58: 303–308.

Lynch JK, Hirtz DG, de Veber G, and Nelson KB (2002) Report of the National Institute of Neurological Disorders and Stroke workshop on perinatal and childhood stroke. *Pediatrics* 109: 116–123.

Miller SP, Wu YW, Lee J, et al. (2006) Candidate gene polymorphisms do not differ between newborns with stroke and normal controls. *Stroke* 37: 2678–2683.

Nakajima W, Ishida A, Lange MS, et al. (2000) Apoptosis has a prolonged role in the neurodegeneration after hypoxic ischemia in the newborn rat. *Journal of Neuroscience* 20: 7994–8004.

Paolucci S, Bragoni M, Coiro P, et al. (2006) Is sex a prognostic factor in stroke rehabilitation? A matched comparison. *Stroke* 37: 2989–2994.

Rice JE III, Vannucci RC, and Brierley JB (1981) The influence of immaturity on hypoxic-ischemic brain damage in the rat. *Annals of Neurology* 9: 131–141.

Sprigg N, Bath PM, Zhao L, et al. (2006) Granulocyte-colony-stimulating factor mobilizes bone marrow cells in patients with subacute ischemic stroke. *Stroke* 37: 2979–2983.

Towfighi J, Mauger D, Vannucci RC, and Vannucci SJ (1997) Influence of age on the cerebral lesions in an immature rat model of cerebral hypoxia-ischemia: A light microscopic study. *Brain Research Developmental Brain Research* 100: 149–160.

Wang MM (2006) Genetics of ischemic stroke: Future clinical applications. *Seminars in Neurology* 26(5): 523–530.

Wu YW, Lynch JK, and Nelson KB (2005) Perinatal arterial stroke: Understanding mechanisms and outcomes. *Seminars in Neurology* 25(4): 424–434.

Zhu C, Xu F, Wang X, et al. (2006) Different apoptotic mechanisms are activated in male and female brains after neonatal hypoxia-ischemia. *Journal of Neurochemistry* 96(4): 1016–1027.

Substance Abuse and Dependence

J A Brewer and M N Potenza, Yale University School of Medicine, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Addiction involves compulsive substance use despite negative consequences and incorporates components of both substance abuse and dependence, as defined by the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV) and the *International Classification of Diseases*.

The DSM-IV diagnostic criteria for substance abuse require the presence of a "maladaptive pattern of substance use leading to clinically significant impairment or distress" as manifested by recurrent substance use resulting in a failure to fulfill major role obligations at work, school, or home; situations in which it is physically hazardous; related legal problems; or continued substance use despite having persistent or recurrent social or interpersonal problems caused or exacerbated by the effects of the substance.

The DSM-IV diagnostic criteria for substance dependence incorporate a combination of three or more of substance-related manifestations including tolerance, withdrawal, larger amounts of drug taken for longer than intended, a persistent desire or unsuccessful attempts to quit, a large amount of time/activity spent obtaining, using, or recovering from the substance, important activities given up or reduced due to substance use, and continued use despite knowledge of adverse consequence.

Generally speaking, ingestion of a psychoactive substance leads to a pleasurable feeling and/or body sensation (the 'high' of cocaine or the 'buzz' of alcohol, for example). As with other goal-directed or rewarding behaviors, memory of drug-related pleasurable experiences leads to the desire for their repetition, giving these substances reinforcing properties.

On a molecular level, psychoactive substances have diverse mechanisms of actions (see specific classes below) but typically share a common final pathway: increased synaptic dopamine within the basal ganglia. Specifically, projections from the ventral tegmental area release dopamine in the ventral pallidum, nucleus accumbens, prefrontal cortex, and amygdala, all part of a 'motivational' or reward circuit. Dopamine activity within this circuitry facilitates the learning of associations between behavior and perceived sensations.

Repeated drug use induces reorganization in the motivational circuit leading to the establishment of patterns of behaviors characteristic of addiction, namely 'cravings,' persistent use despite negative consequences, and a perceived 'loss of control.' These patterns are startling in their power to supplant most other goals.

Additionally, repetitive use of many drugs leads to the development of tolerance, in which larger amounts need to be ingested to produce the desired effects. This may be due to induction of liver enzymes that break down the drug, thus increasing drug metabolism, the downregulation of receptors on which the drug works, leading to a decrease in sensitivity, and/or other mechanisms, such as changes in the expression patterns of intracellular proteins.

With this as an introduction, the following sections briefly review the major classes of commonly abused drugs. Estimates of prevalence of use, abuse, and dependence related to these drugs are found in Tables 1 and 2.

The Opioids

The opioid agonists include both naturally occurring alkaloids (e.g., codeine, morphine, and opium), the semisynthetic derivatives (e.g., hydromorphone, hydrocodone, and heroin), and the synthetic compounds (fentanyl, tramadol, methadone, meperidine). There exist partial agonist (agonist–antagonist) opioids as well, for example, butorphanol, nalbuphine, and buprenorphine. Opiates may be ingested orally, inhaled, injected intravenously or intramuscularly. Additionally, they are commonly taken in combination with other psychoactive substances (e.g., heroin mixed with cocaine and injected intravenously, aka a 'speedball'). Rapid elevations in blood levels correlate with the associated 'high' or degree of intoxication. Once in the blood, opiates travel to the brain where they bind to opioid receptors (of primary importance is the mu-opioid receptor (reclassified as OP3) which modulates euphoria and analgesia) in the brain and other organ systems such as the gastrointestinal (GI) tract. In the brain, opioids activate the mesolimbic reward system generating signals in the ventral tegmental area, causing release of dopamine in the nucleus accumbens and feelings of pleasure. Binding to the mu-opioid receptor in the locus coeruleus suppresses norepinephrine (noradrenaline) release to other parts of the brain where its normal function is to stimulate wakefulness, breathing, and general alertness, among other functions.

Clinically, acute opioid intoxication manifests as euphoria (the acuity of which depends on the drug and route of administration), relaxation, drowsiness, pupillary constriction (miosis), motor retardation,

Table 1 Lifetime prevalence estimates of substance use, abuse, and dependence

	Lifetime use[a] (%)	*Lifetime abuse without dependence[b] (%)*	*Lifetime dependence[b] (%)*	*Lifetime abuse and dependence[b] (%)*
Opiates	1.3	1.1	0.07	0.3
Alcohol	82.4	17.8	1.7	10.8
Cocaine	14.2	1.8	0.7	0.9
Stimulants	8.3	1.4	0.1	0.5
Hallucinogens	14.3	1.5	0.03	0.2
Marijuana	40.2	7.1	0.1	1.2
Tobacco	71.4	N/A	24.0	N/A

[a]Age 12 or older from the 2004 National Survey on Drug Use and Health.
[b]From the 2001–02 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC).
N/A, not applicable as criteria for nicotine abuse have not been operationalized.

Table 2 Lifetime use of substances among students

Lifetime use (including abuse/ dependence)	*8th grade (%)*	*10th grade (%)*	*12th grade (%)*
Any illicit drug	21.4	38.2	50.4
Marijuana/hashish	16.5	34.1	44.8
Inhalants	17.1	13.1	11.4
Hallucinogens	3.8	5.8	8.8
MDMA	2.8	4.0	5.4
Cocaine	3.7	5.2	8.0
Crack	2.4	2.5	3.5
Heroin	1.5	1.5	1.5
Amphetamines	7.4	11.1	13.1
Methamphetamine	3.1	4.1	4.5
Alcohol	41.0	63.2	75.1
Cigarettes	25.9	38.9	50.0

From the 2005 Monitoring the Future survey.

slowing of the GI tract, and decreased pulse, blood pressure, and respiratory rate. Overdose can be fatal if untreated, with the three hallmark signs being central nervous system (CNS) depression (ranging from drowsiness to coma), miosis, and respiratory depression.

Continued or repeated exposure to opioids leads to tolerance and dependence in part via decreased receptor responsiveness. Importantly, abrupt cessation of these drugs or the administration of an opioid antagonist such as naloxone in dependent individuals leads to a withdrawal state, characterized by anxiety, muscle cramping, nausea, vomiting, diarrhea, pupillary dilation (mydriasis), diaphoresis, and yawning. Physiologically, this withdrawal state is thought to be mediated in part by increased noradrenergic output in the locus coeruleus. The onset of a withdrawal state is dependent upon the pharmacological half-life of the opioid involved: the longer the half-life, the later the onset of withdrawal. Typical withdrawal from heroin (half-life ~2 h), for example, begins 6–12 h after its last use, whereas that from methadone (half-life 15 h) begins 36–72 h after last use. Whereas withdrawal symptoms are generally uncomfortable, they are typically not life threatening (exceptions include at-risk vulnerable groups like newborns, older adults, and medically compromised individuals), and can be alleviated by administration of an opioid (the most common being the long-acting methadone). In the hospital or other treatment settings, if further opiate administration is tried to be avoided, withdrawal symptoms are typically treated with the alpha-2 agonist clonidine. Clonidine activates the locus coeruleus and is thought to dampen the overactive noradrenergic output that occurs during withdrawal. Clonidine may be augmented with lorazepam (or other anxiolytics), immodium, and tramadol for symptomatic relief of anxiety, diarrhea, and pain, respectively.

Current pharmacologic treatments for opiate addiction include methadone, buprenorphine, buprenorphine/naloxone combinations, and naltrexone. Methadone, a long-acting synthetic opiate (see above), is a schedule II controlled substance used to reduce opiate use and retain patients in treatment. In the United States, methadone is administered via federally licensed 'methadone clinics,' where it is provided in liquid form and ingestion is generally observed by a clinician. It is currently the mainstay of treatment and has been shown to reduce both mortality and morbidity. Buprenorphine, an opioid receptor partial agonist, was approved by the United States Food and Drug Administration (FDA) in 2002 for the management of opiate dependence. It is administered sublingually due to significantly reduced availability when taken orally due to first-pass effects. Additionally, it is available as a combination pill with naloxone (an opioid antagonist) to reduce its abuse potential: naloxone is poorly absorbed through the GI tract, thus has little to no effect when taken orally as intended, but if crushed and injected, it blocks binding of the buprenorphine to the opiate receptors. Naltrexone, another opioid antagonist which can be taken orally, can block opiate effects at the opiate

receptor for over 24 h, but is typically not abused because it lacks the positive effects afforded by opioid agonists or partial agonists. It is generally used in settings of teenage or newly narcotic addicted or within highly motivated groups (e.g., professionals such as doctors or nurses with licensing privileges at stake), with the hope of avoiding the need for methadone (which is often a more chronic therapy). A long-acting depot form of naltrexone appears safe and efficacious and was recently approved by the FDA for alcohol dependence (see below).

Nonpharmacological approaches include individual and group drug counseling, self-help groups, contingency management techniques ('earning' money for drug-free urines or the opportunity to receive take-home medications in return for prolonged abstinence), and therapeutic communities. Therapeutic communities are often long-term (6–18 months) residential settings where residents share responsibilities such as cooking, cleaning, and leading group therapy.

CNS Depressants

The CNS depressants include alcohol, benzodiazepines, barbiturates, and sedative hypnotics. As the class depressant implies, this group of drugs induces sedation or sleep. These drugs can be administered orally or intravenously, and once in the bloodstream, act on various receptors in the brain.

Benzodiazepines are relatively lipid soluble, with their degree of solubility determining their relative onset (more lipophilic = faster onset). They are thought to act at the benzodiazepine receptor complex which includes the γ-aminobutyric acid (GABA) receptor and a chloride channel. When bound, they increase the sensitivity of the GABA receptor complex to GABA, an inhibitory neurotransmitter, either by increased affinity for the receptor or increased coupling with its associated chloride channel. Neurotransmission at this receptor complex and associated GABA pathways leads to characteristic sedation, anxiolysis, and striated muscle relaxation. Benzodiazepine overdose is generally not fatal in otherwise healthy adults, unless combined with other drugs or taken parenterally.

Ethanol acts at a variety of sites in the brain. Specifically, it facilitates $GABA_A$ receptor function, inhibits *N*-methyl-D-aspartate (NMDA) receptors, and may influence voltage-sensitive calcium channels, serotonin receptors, and dopamine release. Intoxication is characterized by a wide range of behavioral and cognitive effects including sedation, clouded memory and judgment, slowed information processing and reaction time, ataxia, and slurred speech. Alcohol overdose or 'poisoning' often manifests as coma and respiratory depression, and can lead to death (often by coincident aspiration or trauma).

Barbiturates are largely absorbed in the stomach and small intestine, and generally act by facilitating GABA transmission at the GABA receptor. Additionally, they may act by inhibiting calcium-mediated excitatory neurotransmitter release. Intoxication is similar to alcohol and benzodiazepines, with overdose manifesting in stupor to coma. Additional signs of overdose include hypothermia, respiratory depression, and hypotension.

CNS depressants all have the propensity for tolerance and dependence. Due to a shared mechanism of action on the GABA receptor, there is a degree of cross-tolerance that can also develop between groups (alcohol and benzodiazepines most commonly).

Mild withdrawal from alcohol in dependent individuals manifests as anxiety, irritability, nausea, vomiting, tremor, anorexia, and diaphoresis. More serious effects include hallucinosis (often characterized by tactile and/or visual hallucinations), delirium tremens (disorientation, autonomic instability, fever, agitation), and seizures. Withdrawal generally begins 6–36 h after cessation, followed by hallucinosis and delirium tremens (at 3–5 days, but lasting up to 10 days). Seizures usually peak at 1–3 days after the last drink, but have been reported out to 14 days. Benzodiazepine and barbiturate withdrawal shows similar symptoms, but the timecourse may be different depending on the half-life of the agent involved.

Treatment of acute withdrawal from CNS depressants is usually undertaken in a hospital or acute detoxification facility. Initially, patients are provided supportive treatment (vitamin, electrolyte, and fluid repletion), while a taper or cross-taper (e.g., given chlordiazepoxide for alcohol dependence) of the class of drug is initiated. Anticonvulsants have also been used for alcohol withdrawal.

Pharmacological relapse prevention strategies for alcohol dependence include disulfiram, opioid antagonists, and acamprosate. Disulfiram has been available since the 1950s and works by inhibiting the liver enzyme aldehyde dehydrogenase, causing the buildup of acetaldehyde, a byproduct of alcohol, and leading to a variety of unpleasant physiological effects, including nausea and flushing. It is often suboptimal in maintaining abstinence due to treatment nonadherence. Naltrexone, an opioid receptor antagonist (see 'The Opioids'), is also used, and this medication is thought to reduce cravings and substance use as an anticraving medication. Naltrexone has been shown in a majority of trials to be superior to placebo in preventing relapse, but adherence is variable. A depot formulation has been developed, tested, and approved by the FDA in 2006. Acamprosate has

multiple actions in the brain, including decreased presynaptic glutamate release, decreased NMDA receptor excitability, and decreased calcium flux through the voltage-dependent calcium channels. Approved by the FDA in 2004, it (in combination with naltrexone, due to their theoretically different and complementary mechanisms) is currently undergoing a large clinical trial in the United States.

Behavioral treatments and self-help approaches that are used for relapse prevention include inpatient rehabilitation programs, self-help groups (most notably the '12-step' program Alcoholics Anonymous), cognitive behavioral therapies, social skills training, stress management, and community reinforcement, among others. Some evidence suggests that in an outpatient setting, 12-step programs are as effective as cognitive behavioral treatments. However, 12-step programs and behavioral therapies also show high rates of early dropout, suggesting that compliance is a concern across treatment modalities for individuals with alcohol dependence.

CNS Stimulants

CNS stimulants include cocaine, methylphenidate, amphetamine, methamphetamine, and methylenedioxymethamphetamine (MDMA). Most are injected, swallowed, smoked, or snorted depending on the compound. Methylphenidate and amphetamine, when prescribed by physicians, are often used to great therapeutic benefit for attention-deficit hyperactivity disorder (ADHD). They are also used to treat narcolepsy. Cocaine is derived from the leaves of the *Erythroxylon cocais* plant and used medically as a local anesthetic. On the street, cocaine is available in two forms: the water-soluble powder compound formed by dissolving the alkaloid in hydrochloric acid, and the 'free base' compound produced by alkalinization with sodium bicarbonate or ammonia (crack).

Cocaine acts peripherally by activating α- and β-adrenergic receptors, causing vasodilation. Additionally, it activates sigma receptors and blocks the formation of nitric oxide resulting in vasoconstriction. In the brain, it acts by blocking dopamine, norepinephrine, and serotonin reuptake at their respective transporters. Chronic abuse is thought to induce relative dopamine hypofunction, changing the functional coupling of GABA receptors and $GABA_A$ receptor subunit composition. Amphetamine increases extracellular catecholamine and serotonin concentrations via vesicular depletion, reuptake inhibition, enhancement of dopamine synthesis, and monoamine oxidase (MAO) inhibition. MDMA acts similarly to amphetamine, though it causes greater serotonin release than dopamine release. Additionally, it induces vasopressin (ADH) release from the hypothalamus.

Stimulant intoxication may manifest as hyperarousal, hypervigilance, psychomotor agitation, and hypersexuality. Increased amounts and/or length of ingestion can lead to confusion and even stimulant-induced psychosis, characterized by paranoia, delusional thought processes, and compulsive behavior (seen more commonly with amphetamines than cocaine). Due to its vasoactive properties, as well as influences on platelets, cocaine ingestion can lead to ischemic cardiomyopathy and associated chest pain, myocardial infarctions (heart attacks), strokes and other vascular problems. MDMA ingestion can initially cause agitation, time distortion, and diminished thirst, followed by euphoria and a sense of well-being. Additionally, intoxication can lead to diaphoresis, tremor, hypertension, and even serotonin syndrome (characterized by mental status change, autonomic hyperactivity, and neuromuscular abnormalities).

Abrupt cessation after continued use of a stimulant can produce a wide range of dysphoric symptomatology, commonly known as a 'crash.' During this time, anxiety, agitation, intense craving, and depressive symptoms can manifest, although controlled studies have often shown minimal withdrawal symptoms.

A number of medications from stimulants to antidepressants have been evaluated in relapse prevention studies, but none have shown substantial efficacy and currently there are no FDA-approved medications for the treatment of cocaine dependence. In targeting dopaminergic systems, among the most promising drugs for treatment of cocaine dependence has been disulfiram, a drug that inhibits dopamine β-hydroxylase (DBH), the enzyme that converts dopamine to norepinephrine. Novel therapeutics are currently being evaluated and include drugs that act on the GABA system (tiagabine, gabapentin, baclofen) and immunotherapies (cocaine vaccine).

Psychotherapies that have been employed for cocaine and other stimulant abuse include cognitive behavioral therapy, drug counseling, and contingency management procedures (presenting rewards or incentives for abstinence and withdrawing privileges for continued drug use).

Hallucinogens

The hallucinogens include lysergic acid diethylamide (LSD), mescaline, and psilocybin. LSD is swallowed or absorbed through oral mucosa, mescaline is smoked or swallowed, and psilocybin is generally swallowed. Once ingested, hallucinogens act as

agonists at the postsynaptic serotonin (5-HT_2)and NMDA receptors, among others.

Hallucinogens can produce an altered state of perception including synesthesia (mixing of sensory modalities; e.g., "I feel orange"), distortion of time, increased body temperature, pulse and blood pressure, loss of appetite, numbness, weakness, nervousness, panic, and paranoia. Visual disturbances, such as hallucinations, are also common. After-imagery, geometric pseudo-hallucinations, and halos can sometimes persist for years after drug use, and this effect may be due to an unusual sensitivity of the visual system compared to other cortical pathways. Studies of LSD administration also show an incidence of subsequent psychosis in up to 5% of samples studied.

There is little evidence for documented hallucinogen withdrawal. Benzodiazepines are used for treatment of panic induced by acute intoxication, and neuroleptics are used for psychotic components of ingestion.

Marijuana

Marijuana is the most commonly used illegal drug in the US. It is commonly swallowed or smoked. Its active ingredient is delta-9-tetrahydrocannabinol (THC), with an average joint containing between 5 and 20 mg of THC, although the concentrations may be increasing with new varieties of marijuana being grown. Inhalation provides a more rapid rise in plasma THC levels than does ingestion.

Acutely, marijuana has peripheral, as well as central effects. THC is thought to act at the cannabinoid receptors (thus far, there have been three subtypes identified: CB1, CB2, and CB3), which are prominent in the hippocampus, thalamus, brain stem, and basal ganglia. Although further characterization is required, activation of these receptors produces blockade of voltage-activated calcium channels, and activation of adenylyl cyclase via G-protein coupling. CB1 and CB3 receptors regulate GABA, glutamate, dopamine, norepinephrine, serotonin, and acetylcholine release in the brain. THC can produce mild euphoria, relaxation, perceptual distortions, and enhancement of ordinary sensory experiences. It can also impair short-term memory, motor skills, and reaction time. It is commonly known to stimulate appetite, and medicinal THC preparations are used to varying degrees of efficacy for anorexia associated with chronic diseases such as AIDS. Peripherally, THC can cause tachycardia and, at high doses, orthostatic hypotension. Although controversial, evidence also suggests an association between cannabis use and psychosis. This association has biological support from studies, suggesting that multiple actions are mediated through CB receptor activation.

With chronic use of marijuana, evidence suggests the development of tolerance, including decreases in autonomic functions, increased intraocular pressure, and mood changes. Abrupt cessation does not result in severe withdrawal, although subjective symptoms can include increased irritability and restlessness, insomnia, anorexia, and mild nausea. Signs can also include tremor, increased body temperature, and weight loss. Some studies estimate the risk of dependence is up to 30% with use over five times.

Relapse prevention strategies for marijuana dependence have only recently been studied. Cognitive behavioral therapy, motivational enhancement therapy, and case management have been demonstrated to be efficacious in reducing use in a limited number of controlled trials, with fewer data available about the efficacy of pharmacotherapies, although some drugs (antidepressants, anxiolytics) have shown potential in open-label trials.

Club Drugs and Dissociative Anesthetics

This group of drugs includes ketamine, phencyclidine (PCP), MDMA (see 'CNS Stimulants'), and gamma-hydroxybutyrate (GHB). Ketamine is taken intravenously, snorted, or smoked, PCP is injected, swallowed, or smoked (often dissolved in formaldehyde and soaked into marijuana, known as 'illy' or 'wet'), and GHB is swallowed.

PCP was first synthesized in the 1950s and marketed as a surgical anesthetic, but withdrawn from the market 15 years later due to associated behavioral disturbances including dysphoria, paranoia, hallucinations, rage, and violence. The psychotic symptoms from PCP anesthesia were noted to persist for more than a week in many patients. It acts as a noncompetitive antagonist at the NMDA receptor, a cholinergic antagonist centrally, and dopamine 2 receptor antagonist peripherally. Additionally, it increases dopamine release from synaptic vesicles. These actions manifest, as noted above, in a mixture of anesthetic, stimulant, depressant, and hallucinogenic effects. At low doses, anticholinergic symptoms are prominent (red dry skin, amnesia, disorganization), and, at high doses, dopaminergic effects predominate (paranoia and hallucinations). In susceptible individuals, PCP may 'induce' a permanent psychosis impossible to distinguish from schizophrenia, although it is often difficult to differentiate direct causality versus precipitation of an underlying psychotic illness by the drug.

Ketamine, a less potent structural analog of PCP introduced into the pharmacological armamentarium

about 30 years ago, is a dissociative anesthetic and continues to be used clinically for a number of applications (mainly in pediatric populations due to often intolerable side effects in adults, including nightmares). Arguably, its best-known action is as an NMDA receptor antagonist (binding to the same site as PCP), and it continues to be used in clinical studies as a model of induced psychosis. Additionally, ketamine binds to mu- and kappa-opioid receptors, blocks monoamine reuptake, and influences muscarinic receptors and voltage-sensitive calcium channels. When used clinically or as a club drug, ketamine produces dissociative anesthesia (floating outside of one's body), a 'dream-like' state, analgesia, perceptual changes (including hallucinations), delusions, impaired memory, increased pulse, and increased blood pressure.

GHB was sold until 1991 and used for fat reduction and muscle building. Structurally similar to GABA, GHB acts through the $GABA_B$ receptor complex, on second messenger pathways and through dopamine signaling blockade. Effects of GHB ingestion are dose dependent, ranging from amnesia, somnulence, and hypotonia to dizziness, confusion, respiratory depression, hypotension, bradycardia, seizures, loss of consciousness, and even coma at higher levels. As the effects of GHB generally subside in a few hours, treatment of GHB intoxication is typically supportive, although physostigmine has also been used.

There is some evidence of development of withdrawal symptoms with extended use of ketamine, and, in some severe cases, detoxification may be warranted. Chronic use of GHB may produce dependence and a withdrawal syndrome that includes anxiety, insomnia, tremor, and, in severe cases, treatment-resistant psychoses.

Tobacco

Tobacco is available as snuff, chewing tobacco, cigars, and cigarettes, although the majority is consumed as cigarettes. The main active ingredient is nicotine, a widely distributed plant alkaloid found in highest concentrations in the species *Nicotinia tobacum* and *Nicotinia risticum*. Once it enters the bloodstream, nicotine, a tertiary amine similar to acetylcholine, binds to nicotinic cholinergic receptors in the brain and elsewhere. In the brain, its sympathomimetic activity enhances release of dopamine, norepinephrine, acetylcholine, serotonin, vasopressin, glutamate, and GABA, among other neurotransmitters. In this respect, it acts as a CNS stimulant. With tobacco use, people report improved concentration, elevated mood, and reduced anger, stress, and tension, although it is sometimes difficult to differentiate intrinsic effects on the brain from those related to relief from abstinence. However, nicotine administration improves attention, reaction time, and problem solving, suggesting the former is at least contributory.

With repeated exposure to nicotine, tolerance and dependence develop which are likely mediated physiologically by regulation of the nicotinic acetylcholine receptor and psychologically by the mesolimbic dopamine system (described above). Abstinence leads to symptoms of lethargy, irritability, restlessness, inability to concentrate, and dysphoria.

When asked, about 70% of smokers would like to quit and most adults require four or more attempts before being successful. Nicotine replacement therapies (NRTs), in the form of patches, gum, nasal spray, and inhalers, have been shown to be effective in enhancing smoking cessation by twofold compared to placebo, although up to 90% of attempts to quit fail despite NRT. Bupropion has also shown benefit and is FDA approved for smoking cessation, and, when combined with NRT, has been shown to be more effective than NRT alone. Most recently, varenicline, a partial nicotinic receptor agonist, has received FDA approval for smoking cessation.

Aversion therapies, acupuncture, hypnotherapy, and exercise have been used, but further studies need to be done to determine their efficacies.

Conclusions

In sum, substance abuse and dependence exact a tremendous toll economically, socially, and personally on individuals around the world. Although drugs of abuse vary widely in their effects, research into the neurobiological basis of addiction has revealed some striking commonality in their action which leads to addiction. Additionally, new knowledge of specific mechanisms of action of drugs continues to lead to novel, targeted therapies that may prove to be more efficacious than those currently available. Just as importantly, as the neurobiological underpinnings of addiction continue to be elucidated, it is coming to be understood as a disease rather than just a social problem.

See also: Addiction: Neurobiological Mechanism; Alcoholism; Amphetamines; Drug Addiction: Behavioral Neurophysiology; Drug Addiction: Behavioral Pharmacology of Drug Addiction in Rats; Drug Addiction: Cellular Mechanisms; Drug Addiction: Neuroimaging; Drugs Addiction: Actions; Nicotine.

Further Reading

Abraham HD, McCann UD, and Ricaurte GA (2002) Psychedelic drugs. In: Davis KL, Charney D, Coyle JT, and Nemeroff C (eds.) *Neuropsychopharmacology: The Fifth Generation of Progress: An Official Publication of the American College of Neuropsychopharmacology*, pp. 1545–1556. Philadelphia: Lippincott Williams & Wilkins.

Gallagher HL and Frith CD (2003) Functional imaging of 'theory of mind' *Trends in Cognitive Science* 7(2): 77–83.

Hyman SE (2005) Addiction: A disease of learning and memory. *American Journal of Psychiatry* 162(8): 1414–1422.

Kalivas PW and Volkow ND (2005) The neural basis of addiction: A pathology of motivation and choice. *American Journal of Psychiatry* 162(8): 1403–1413.

Kosten TR and George TP (2002) The neurobiology of opioid dependence: Implications for treatment. *Science and Practice Perspectives* 1: 13–20.

Krystal JH, Karper LP, Seibyl JP, et al. (1994) Subanesthetic effects of the noncompetitive NMDA antagonist, ketamine, in humans. Psychotomimetic, perceptual, cognitive, and neuroendocrine responses. *Archives of General Psychiatry* 51(3): 199–214.

Krystal JH, Perry EB, Jr., Belger A, et al. (2005) Comparative and interactive human psychopharmacologic effects of ketamine and amphetamine: Implications for glutamatergic and dopaminergic model psychoses and cognitive function. *Archives of General Psychiatry* 62(9): 985–994.

Krystal JH and Tabakoff B (2002) Ethanol abuse, dependence, and withdrawal: Neurobiology and clinical implications. In: Davis KL, Charney D, Coyle JT, and Nemeroff C (eds.) *Neuropsychopharmacology: The Fifth Generation of Progress: An Official Publication of the American College of Neuropsychopharmacology*, pp. 1425–1433. Philadelphia: Lippincott Williams & Wilkins.

Nestler E and Aghajanian GK (1997) Molecular and cellular basis of addiction. *Science* 278: 58–62.

O'Brien CP (2005) Anticraving medications for relapse prevention: A possible new class of psychoactive medications. *American Journal of Psychiatry* 162(8): 1423–1431.

Sofuoglu M and Kosten TR (2005) Novel approaches to the treatment of cocaine addiction. *CNS Drugs* 19(1): 13–25.

Sulzer D, Sonders MS, Poulsen NW, et al. (2005) Mechanisms of neurotransmitter release by amphetamines: A review. *Progress in Neurobiology* 75(6): 406–433.

Teter CJ and Guthrie SK (2001) A comprehensive review of MDMA and GHB: Two common club drugs. *Pharmacotherapy* 21(12): 1486–1513.

Volpicelli JR, Krishnan-Sarin S, and O'Malley SS (2002) Alcoholism pharmacotherapy. In: Davis KL, Charney D, Coyle JT, and Nemeroff C (eds.) *Neuropsychopharmacology: The Fifth Generation of Progress: An Official Publication of the American College of Neuropsychopharmacology*, pp. 1445–1460. Philadelphia: Lippincott Williams & Wilkins.

Substance P/Tachykinins and its/their Receptors

A Lecci and C A Maggi, Menarini Ricerche, Florence, Italy

© 2009 Elsevier Ltd. All rights reserved.

Tachykinins: Genes, Peptides, and Their Receptors

Tachykinin Peptides

Substance P (SP), neurokinin A (NKA), and neurokinin B (NKB), the first peptides of the family of mammalian tachykinins (TKs) to be characterized, were initially found to be specifically expressed in the neurons of the peripheral districts and central nervous system (CNS) and therefore were included in the broad definition of neuropeptides. Nowadays, it is unanimously agreed that TKs are not exclusively localized in neuronal cells; nevertheless, the peripheral nervous system (PNS) and CNS are the districts in which the effects of TKs have been more extensively investigated. Another mammalian TK has been recently discovered. This was termed hemokinin-1 (HK-1) because it was initially localized in the hematopoietic tissue; however, it soon became clear that HK-1 was also expressed into the CNS. Unlike the other TKs, the sequences of mouse and rat HK-1, on the one hand, and human HK-1, on the other, are not identical; however, no striking pharmacological differences have been detected between these two peptides.

SP, HK-1, NKA, and NKB are short peptides containing 11, 11, 10, and 10 amino acids (aa), respectively; elongated forms of NKA (neuropeptide K with 36 aa and neuropeptide γ with 21 aa) and of HK-1 (endokinin A with 47 aa and endokinin B with 41 aa) have also been identified (Table 1).

Tachykinin Genes

All these TKs are encoded by three genes, termed TAC1 (or preprotachykinin-A, PPT-A, or PPT-I), TAC3 (or PPT-B or PPT-II), and TAC4 (or PPT-C). TAC genes are transcribed into primary mRNAs which are subjected to alternative posttranscriptional splicing, and the derived peptides are then subjected to posttranslational processing. This process produces different phenotypes from just one genotype (Table 2).

As with other peptides that are released extracellularly, TKs are synthesized in the rough endoplasmic reticulum as prepro-TKs (prepropeptide); large peptide sequences directly derived from the translation of exons contained in the mRNAs are listed in Table 2. The cleavage of a signal sequence in the prepro-TKs leads to the pro-TKs (propeptide), which is now stored in the Golgi apparatus; this process occurs in the soma of neurons. The maturation of the Golgi vesicles into secretory vesicles occurs during their transport along axons and parallels a further proteolytic cleavage of the peptide precursor (the propeptide) into the mature peptide that is now ready to be released from the axon terminals. In a subclass of TK-expressing primary afferent neurons (termed capsaicin-sensitive), this process is bidirectional; that is, the release of TKs (together with their co-transmitters) occurs at both the axonal (in the spinal cord) and dendritic (at the peripheral level) pools of the neuron.

All TKs, except endokinin C and D, share a common C-terminal amidated amino acid region, Phe-X-Gly-Leu-Met-NH_2, which is the minimal sequence for exerting biological activity at the TK receptors and therefore for exerting the function of neurotransmitters. Indeed, a truncated form of NKA, NKA(3–10), has been detected in some tissues, and it retains activity at the TK receptors (at NK2, in particular). The N-terminal part of SP also exerts several biological effects, especially at the CNS level; however, these effects are not mediated by known TK receptors. The receptor involved in the effect of the N-terminal fragment of SP (e.g., SP(1–5)) has not been identified yet; however, the recent finding that the SP domain contained in the protachykinin peptide (derived from β-preprotachykinin, or β-TAC1) interacts with δ-opioid receptor to regulate its trafficking at nerve terminals renders this receptor the likely target accounting for the biological effects exerted by SP(1–5).

Tachykinin Receptors

TKs exert their biological effect by acting at three different G-protein-coupled receptors, termed TK NK1 (coded by the TACR1 gene), NK2 (coded by the TACR2 gene), and NK3 (coded by the TACR3 gene). The NK1 receptor can be expressed in two different forms due to alternative TACR1 splicing; these forms have been defined as long and short because of the different lengths of the intracellular C-terminal domain. The short isoform of the NK1 receptor is less prone to desensitization and internalization than the long isoform, and it has been recently shown that the activation of the short isoform is a much more efficient signal for an autocrine induction of β-TAC1 than the stimulation of the long isoform. NK2 receptor is also expressed in two different isoforms: the α-isoform is the classic NK2 receptor, whereas the β-isoform is not a functional TK receptor because it lacks an amino acid sequence forming an extracellular (EC3) loop, an intracellular (IC2) loop, and a transmembrane domain (TM4) and so cannot be activated by TKs.

Table 1 Peptide sequence of mammalian tachykinins[a]

Peptides	*Aminoacid sequence*
Substance P	Arg-Pro-Lys-Pro-Gln-Gln-**Phe**-Phe-**Gly-Leu-Met**-NH_2
Neurokinin A	His-Lys-Thr-Asp-Ser-**Phe**-Val-**Gly-Leu-Met**-NH_2
Neuropeptide-γ	Asp-Ala-Gly-His-Gly-Gln-Ile-Ser-His-Lys-Arg-His-Lys-Thr-Asp-Ser-**Phe**-Val-**Gly-Leu-Met**-NH_2
Neuropeptide-κ	Asp-Ala-Asp-Ser-Ser-Ile-Glu-Lys-Gln-Val-Ala-Leu-Leu-Lys-Ala-Leu-Tyr-Gly-His-Gly-Gln-Ile-Ser-His-Lys-Arg-His-Lys-Thr-Asp-Ser-**Phe**-Val-**Gly-Leu-Met**-NH_2
Neurokinin B	Asp-Met-His-Asp-Phe-**Phe**-Val-**Gly-Leu-Met**-NH_2
Hemokinin-1	Thr-Gly-Lys-Ala-Ser-Gln-**Phe**-Phe-**Gly-Leu-Met**-NH_2
Endokinin A	Asp-Gly-Gly-Glu-Glu-Gln-Thr-Leu-Ser-Thr-Glu-Ala-Glu-Thr-Trp-Val-Ile-Val-Ala-Leu-Glu-Glu-Gly-Ala-Gly-Pro-Ser-Ile-Gln-Leu-Gln-Leu-Gln-Glu-Val-Lys-Thr-Gly-Lys-Ala-Ser-Gln-**Phe**-Phe-**Gly-Leu-Met**-NH_2
Endokinin B	Asp-Gly-Gly-Glu-Glu-Gln-Thr-Leu-Ser-Thr-Glu-Ala-Glu-Thr-Trp-Glu-Gly-Ala-Gly-Pro-Ser-Ile-Gln-Leu-Gln-Leu-Gln-Glu-Val-Lys-Thr-Gly-Lys-Ala-Ser-Gln-**Phe**-Phe-**Gly-Leu-Met**-NH_2
Endokinin C	Lys-Lys-Ala-Tyr-Gln-Leu-Glu-His-Thr-**Phe**-Gln-**Gly-Leu**-Leu-NH_2
Endokinin D	Val-Gly-Ala-Tyr-Gln-Leu-Glu-His-Thr-**Phe**-Gln-**Gly-Leu**-Leu-NH_2
Hemokinin-1[b]	Arg-Ser-Arg-Thr-Arg-Gln-**Phe**-Tyr-**Gly-Leu-Met**-NH_2

[a]The common C-terminal amino acid sequence required for exerting activity at tachykinin receptors is shown in bold: endokinin C and D lack the C-terminal Met and are almost devoid of affinity at these receptors; because of this they have been termed as tachykinin-related peptides. Red indicates the sequence of neurokinin A, of which neuropeptide-γ and neuropeptide-κ are elongated forms; blue indicates the sequence of human hemokinin-1, of which endokinin A and B are elongated forms.
[b]Rat and mouse sequence.

Table 2 Tachykinin genes, tachykinin-encoding messenger RNAs, and tachykinin peptides[a]

TK gene	*TK mRNA*	*TK peptides*
TAC1	α-TAC1	Substance P
	β-TAC1	Substance P, neurokinin A, neuropeptide K
	γ-TAC1	Substance P, neurokinin A, neuropeptide-γ
	δ-TAC1	Substance P
TAC3	α-TAC3	Neurokinin B
	β-TAC3	Neurokinin B
	γ-TAC3	None
TAC4	α-TAC4	Hemokinin-1, endokinin A, endokinin C[b]
	β-TAC4	Hemokinin-1, endokinin B, endokinin D[b]
	γ-TAC4	Hemokinin-1, endokinin B
	δ-TAC4	Hemokinin-1, endokinin B

[a]TK, tachykinin.
[b]TK gene-related peptides (see **Table 1**).

All TKs act as full agonists at the NK1, NK2, and NK3 receptors, although with different estimated affinities. The potency orders of TKs at the receptors are:

- NK1: SP = hHK-1 > NKA > NKB
- NK2: NKA > NKB > SP > hHK-1
- NK3: NKB > NKA > hHK-1 > SP

However, significant deviations to these rankings can occur, mainly depending on (1) the animal species investigated (large species-related differences in the affinity of nonpeptide TK receptor antagonists are well known, but smaller differences are also present among peptide agonists) and (2) the kind of bioassay (if more than one TK receptor is expressed or is in any way involved in the response measured, the kind of signal transduction activated, or the presence of peptidases). Selective TK receptor agonists (peptides only) and antagonists (both peptides and nonpeptides) are also available. These selective ligands have been fundamental tools in the characterization of the receptors involved in a given function. However, there are several studies in which a response is mediated by more than one TK receptors, so that the application of only one selective receptor antagonist might not be sufficient to block, or even to reduce, the studied function unless other TK receptors are blocked.

Knockout mice lacking the TAC1 or TACR1 gene have been produced. These knockout mice have also been instrumental in clarifying several aspects of the neurophysiological processes mediated by TKs; however, the possibility of developmental remodeling to compensate for the lack of function in the system can never be excluded when adopting this strategy. Indeed, as with other neurotransmitter systems, the tachykinergic system must not be thought of as static; there are several examples showing that developmental and environmental changes affect the pattern of expression of TKs and their receptors in both neurons and glia. Finally, the importance of TKs in neurophysiological processes regulating sensory, emotional, and neurovegetative functions is highlighted by a large number of preclinical studies.

Tachykinins and Their Receptors in the Central Nervous System

All the TAC genes described so far are expressed in the CNS; however, it is uncertain whether TAC4 is expressed exclusively in the glia or in both the glia and neurons. β- and γ-TAC1 mRNAs are the most abundant TAC1 transcripts in the CNS and have been quantified as appearing in a ratio of 1:3; α-TAC1 is expressed at a low level in the olfactory bulb and striatum (approximately 1% of the total transcripts). The most conspicuous source of TAC1 products in the CNS is represented by neurons located in the extra-pyramidal (mainly basal ganglia), limbic (mainly accumbens), thalamic (hypothalamus), and brain stem autonomic areas (mainly the raphe nuclei, periacqueductal gray matter, and lateral dorsal tegmentum). The substantia nigra and dorsal horns of the spinal cord contain a high density of SP- and NKA-positive fibers. Numerous NKB-positive neurons have been detected in the bed nucleus of stria terminalis, olfactory areas, and hypothalamic nuclei; the two latter areas also contain a dense network of NKB-positive fibers.

All the TACR genes described so far have been found to be expressed in the CNS. Despite the species-related variations in the distribution of TK receptors, a general pattern of TK NK1, NK2, and NK3 receptor expression can be discerned. NK1 receptors, of which the long isoform preponderates in the brain, are well represented in the olfactory system and related cortical areas, hippocampus, septum, amygdala, pons, and medulla. NK2 receptors are expressed at low levels in the limbic area of the cortex, hippocampus, thalamic nuclei, and substantia nigra (although only TACR2 mRNA was found, functional evidence for the expression of NK2 receptors is available). NK3 receptors are particularly dense in the deep cortical layers; hippocampus; amygdala; and selected hypothalamic, mesencephalic, and medullary nuclei. The dorsal horns of the spinal cord mainly express NK1 and NK3 receptors; NK2 receptors have also been described, although at a low level of expression. CNS NK2 receptors have described as developmentally regulated because autoradiographic studies have described a moderate expression in newborn animals but low, if any, expression in adult animals. Despite this, functional evidence clearly indicates that the adult brain expresses NK2 receptors in the same areas as newborn animals.

NK1, NK2, and NK3 can be expressed in glial cells which also express all the four TAC1 mRNAs during development. This expression is again turned on in the case of infection or injury, and SP acting through NK1 receptors can support the cytokine-mediated proliferation of glia. This mechanism is also important for the growth of gliomas; in these tumors, TAC4 mRNA expression was also detected. Recently, it has been shown that spinal glial cells can play a role in nocifensive behavior induced by intraplantar carragenan or intrathecal SP, whereas the SP-induced loss of dopaminergic neurons is mediated by the activation of microglial nicotinamide adenine dinucleotide phosphate (NADPH) oxidase.

Several drugs targeting CNS NK1, NK2, and NK3 receptors have been tested in clinics. Promising results have been obtained with an NK3 antagonist in schizophrenia, whereas the same antagonist was ineffective in a pilot trial on levodopa-induced dyskinesia. Likewise, results from large trials on NK1 receptor antagonists in pain and depression have been disappointing; however, aprepritant has been approved for the treatment of nausea and vomiting associated with chemotherapy or surgery. Most of the anatomical and functional information available derives from animal studies, the rat being the most extensively investigated. The main nerve TK pathways in the rat brain are shown in Figure 1 in more details on the distribution of TK peptides and their receptors in the CNS are presented in Table 3.

Rhinencephalon

In macrosomatic animals, the olfactory areas are involved in social behavior, reproduction, and emotional and aggressive behavior, but in microsomatic species, other brain areas are much more important for the regulation of these functions. Because of this, numerous species-related differences have been found in the cytoarchitecture and neuronal chemical coding of this area.

In invertebrates (*Drosophila*), TKs are involved in the modulation of olfactory perception, and a correlation between the degree of anosmia and the loss of SP-positive trigeminal neurons has been observed in a rodent model. However, no major deficits in odor perception have been detected in mice lacking the TAC1 gene. In the rat olfactory tubercle, TAC1 levels are modulated by sodium depletion, alcohol preference, and aging. Olfactory bulbs have also been implicated in the modulation of aggressive behavior; bulbectomy increases aggressions and locomotor activity, effects which are antagonized by antidepressants. Interestingly, bulbectomy-induced hyperactivity was not observed in TAC1 knockout mice, whereas cage-residents NK1 knockout mice displayed reduced aggressions against intruders. However, these effects could also be dependent on other brain areas (the accumbens, ventral pallidum, substantia nigra, or hypothalamus).

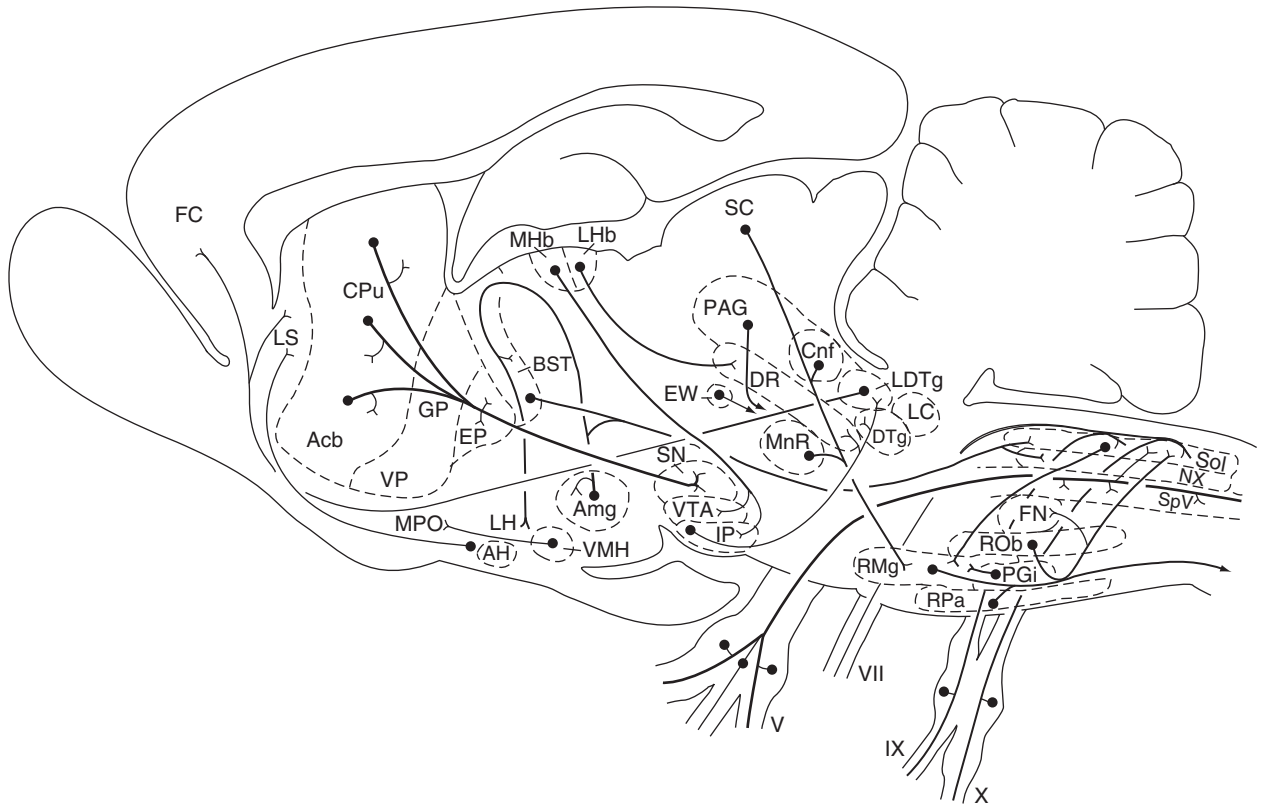

Figure 1 Drawing of main tachykinin (TK) pathways in a sagittal section of the rat brain. (1) Medium-size TK-containing neurons in the striatum, also termed as caudate putamen (CPu), send collaterals within the caudate putamen and project to the substantia nigra (SN). Striatal TK-containing neurons also send axons to the entopeduncular nucleus (EP), which is the rodent equivalent of the internal segment of the primate globus pallidus (GP). (2) TK-containing neurons in the ventromedial hypothalamic nucleus (VMH) send axons to the medial preoptic area (MPO), and those in the area between anterior hypothalamic nucleus (AH) and lateral hypothalamus (LH) project to the lateral septal nucleus (LS). The amygdala (Amg) contains a number of TK-containing neurons, which give rise to an intrinsic plexus in the amygdala and also project via the stria terminalis to the bed nucleus of this bundle (BST) and LH. TK-containing neurons in the medial part of the central nucleus of amygdala, together with those in the ventral bed nucleus of stria terminalis, project to dorsal motor nucleus of the vagus nerve (NX), and solitary tract nucleus (Sol). (4) TK-containing neurons in medial habenula (MHb) project via habenulointerpeduncular tract to the lateral subnuclei of interpeduncular nucleus (IP) and ventral tegmental area (VTA). Those in lateral habenula (LHb) project to dorsal raphe nucleus (DR). (5) TK-containing neurons found in rostral subnucleus of interpeduncular nucleus project to dorsal tegmental nucleus (DTg) and laterodorsal tegmental nucleus (LDTg) and to the rostral extension of dorsal raphe nucleus. (6) A group of cholinergic TK-containing neurons in laterodorsal tegmental nucleus projects via the medial forebrain bundle to the lateral septal nucleus and to medial parts of the frontal cortex (FC). (7) TK-containing neurons in the Edinger–Westphal nucleus (EW) project to the facial nucleus (FN) and spinal cord. (8) TK- and cholecystokinin-positive neurons in the periaqueductal central gray matter (PAG) send descending axons to the facial nucleus and spinal cord. (9) TK-containing neurons also co-expressing serotonin and thryrotropin-releasing hormone in raphe pallidus (RPa), raphe obscurus (ROb), and possibly raphe magnus (RMg) send axons to the facial nucleus, solitary tract nucleus, and spinal cord. (10) Raphe magnus receives TK projections from a number of brain stem nuclei including the cuneiform nucleus (Cnf), superior colliculus (SC), median raphe nucleus (MnR), paragigantocellular reticular nucleus (PGi), and solitary tract nucleus. (11) TK-containing primary afferent neurons in the ganglion of the trigeminal nerve (V) receive inputs from various tissues of the orofacial region and terminate in the caudal part of the spinal trigeminal nucleus (SpV). (12) The solitary tract nucleus is innervated by TK-containing primary afferent fibers from the facial (VII), glossopharyngeal (IX), vagal (X), and trigeminal nerves. Acb, nucleus accumbens; LC, locus coeruleus; VP, ventral pallidum. Adapted from Otsuka M and Yoshioka K (1993) Neurotransmitter functions of mammalian tachykinins. *Physiological Review* 73: 229–308.

Telencephalon

The expression of TKs in cortical areas is sparse in rodents but more prominent in other species, including humans. In the rat neocortex, TAC1 and TAC3 are expressed in separate populations of γ-aminobutyric acid (GABA)egic neurons, NKB often co-localizes with acetylcholine, vasointestinal peptide, colecystokinin (CCK), and corticotropin-releasing hormone (CRH). In the enthorinal cortex, bicuculline-induced epileptiform activity was reduced by the activation of NK1, whereas NK3 receptor-mediated enhancement of this activity was due to glutamtergic mechanism. In primates, both NK1 and NK3 receptors are distributed throughout all cortical areas; however, unlike in

Table 3 Tachykinin-positive neurons, fibers, and receptors in CNS areas

Area	*TK-positive neurons*	*TK-positive fibers*	*TK receptors*
Rhinencephalon			
Olfactory bulbs	SP-, NKA-positive mitral and granule cells NKB-positive perigranular cells	SP-, NKA-, NKB-positive fibers. Trigeminal SP and NKA-positive sensory fibers	NK1, NK3
Accessory olfactory bulbs	SP-, NKA-positive granule cells. NKB-positive neurons projecting to the amygdala	SP and NKA-positive sensory fibers	NK1, NK3
Olfactory tubercles	SP-, NKA-positive neurons projecting to the ventral pallidum. NKB-positive neurons projecting to the striatum	SP-, NKA-, NKB-positive fibers	NK1, NK3
Piriform cortex	SP-, NKA-positive neurons	SP-, NKA-, NKB-positive fibers	NK1, NK2, NK3
Telencephalon			
Neocortex	SP-, NKA-positive nonpyramidal cells of layers 4-6. SP-, NKA-positive pyramidal cells projecting to subcortical targets. NKB-positive neurons in layers 2, 3, and 4	SP-, NKA-, NKB-positive fibers. SP, NKA-positive fibers from the medial forebrain bundle (cell bodies in the dorsolateral tegmentum and from septal nuclei)	NK1, NK2, NK3
Nucleus accumbens	SP-, NKA-positive neurons co-express CART peptide and project to the ventral pallidum. Other SP-, NKA-, NKB-positive neurons project ventral tegmental area (some contain GABA). NKB- and mu opioid receptor-positive neurons project to the substantia innominata	SP- and NKA-positive fibers originate from solitary tract nucleus and from the periaqueductal gray area. NKB-positive fibers	NK1, NK2, NK3
Globus pallidus		SP- and NKA-positive fibers (contain dynorphin) from the striatum	NK1, NK3
Entopeduncular nucleus	SP-, NKA-positive neurons (some contain somatostatin) project to the lateral habenular nucleus	SP- and NKA-positive fibers from the striatum	NK1, NK3
Ventral pallidum		SP- and NKA-positive fibers from the accumbens	NK1, NK3
Striatum: matrix	SP-, NKA-positive neurons (contain GABA) project to the substantia nigra pars reticulata, the ventral tegmental area, the entopeduncular nucleus, and to the striatum itself	SP-, NKA-, NKB-positive fibers. NKB-positive fibers from olfactory tubercles	NK1, NK2, NK3
Striatum: striosomes	SP-, NKA-, NKB-positive neurons (contain GABA, dynorphin, and D1 receptors, some also met-enkephalin) project to the substantia nigra pars compacta and to globus pallidus NKB-positive neurons (contain mu opioid receptors) project to the substantia innominata	SP-, NKA-, NKB-positive fibers	NK1, NK2, NK3
Amygdala	SP-, NKA-, NKB-positive neurons project to the bed nucleus of stria terminalis and hypothalamus. SP- and NKA-positive neurons project to the periaqueductal gray	SP-, NKA-, NKB-positive fibers. SP-, NKA-positive fibers from lateral parabrachial and lateral dorsal tegmental nucleus and NKB- from the accessory olfactory bulbs	NK1, NK3
Bed nucleus of the stria terminalis	SP-, NKA-, NKB-positive neurons project to the periaqueductal gray, motor nucleus of vagus, and solitary tract nucleus	SP-, NKA-, NKB-positive fibers from hypothalamus and amygdala	NK3
Hippocampus	SP- and NKA-positive neurons in pyramidal layer and stratum oriens (CA1). NKB-positive neurons in dentate gyrus neurons (CA3)	SP-, NKA-, NKB-positive fibers from the hyothalamus, septo-hippocampal nucleus and local circuitries	NK1, NK2, NK3
Septum and cholinergic forebrain nuclei	SP, NKA-, and NKB-positive neurons, some projecting to the cortex some to the hippocampus	SP- NKA-, and NKB-positive fibers from the hypothalamus and lateral dorsal tegmental nucleus and from local circuitries	NK1, NK2, NK3
Diencephalon			

Continued

Table 3 Continued

Area	*TK-positive neurons*	*TK-positive fibers*	*TK receptors*
Thalamus	SP-, NKA-, and NKB-positive neurons in habenular nuclei projecting to the midbrain interpeduncular nucleus, dorsal raphe, ventral tegmental area, and the pineal gland	SP-, NKA-, and NKB-positive fibers in habenular nuclei from the entopeduncular nucleus. SP-, NKA-positive fibers in paraventricular thalamic nucleus from Edinger–Westphal and medullary raphe nuclei and mesopontine tegmentum. SP-, NKA-positive fibers in somatosensory thalamic nuclei from spinal cord neurons. SP-, NKA-positive fibers in the intergeniculate leaflet from dorsal raphe. SP-, NKA-, and NKB-positive fibers in the geniculate nucleus from superior colliculus. SP- and NKA-positive fibers in the medial and posterior part of the thalamus from the parabrachial nucleus	NK1, NK2, NK3
Hypothalamus	SP- and NKA-positive neurons in dorsomedial, ventromedial, dorsolateral, lateral, posterior hypothalamus, mammillary nuclei projecting to the hippocampus, preoptic, suprachiasmatic and arcuate, nuclei. NKB-positive neurons in anterior hypothalamus mammillary, arcuate, parabrachial nuclei, dorsal motor nucleus of the vagus, median eminence, and dorsal part of the hypothalamus	SP- and NKA-positive fibers in suprachiasmatic nucleus from retinic afferents. SP-, NKA-, and NKB-positive fibers from the bed nucleus of stria terminalis, the pedunculopontine and the laterodorsal tegmental nucleus (both cholinergic) and medullary cathecolaminergic nuclei (C1, C2, C3, A1, A2)	NK1, NK3
Mesencephalon			
Substantia nigra (A9) pars reticulata		SP- and NKA-positive fibers from ipsilateral striatum matrix area	NK1, NK2, NK3
Substantia nigra (A9) pars compacta		SP-, NKA-, and NKB-positive fibers from ipsilateral striatum striosomal area	NK1, NK2, NK3
Ventrotegmental area (A10)	SP- and NKA-positive neurons	SP-, NKA-, and NKB-positive fibers from habenular and accumbens nuclei	NK1, NK2, NK3
Interpeduncular nucleus	SP- and NKA-positive neurons projecting to the dorsal and laterodorsal tegmentum	SP-, NKA-, and NKB-positive fibers from the habenular nuclei	NK1, NK2, NK3
Superior and inferior colliculi	SP-, NKA-, and NKB-positive neurons projecting to raphe magnus (B3) and to geniculate nuclei and pulvinar	SP- and NKA-positive fibers from retinic ganglionic cells	NK1, NK3
Periaqueductal gray matter	SP- and NKA-positive neurons project to the thalamus, hypothalamus, accumbens, nucleus of tractus solitarius and the spinal cord. NKB-postive neurons project to the facial nucleus	SP-, NKA-, and NKB-positive from bed nucleus of stria terminalis	NK1, NK2, NK3
Edinger–Westphal nucleus	SP- and NKA-positive neurons project to the paraventricular thalamic nucleus, cerebellum, trigeminal motor nucleus and spinal cord	SP- and NKA-positive fibers from the suprachiasmatic nucleus	
Dorsal raphe (B7)	SP- and NKA-positive neurons project to thalamic nuclei intergeniculate leaflet	SP-, NKA-, and NKB-positive fibers from thalamic and habenular nuclei	NK1, NK3
Metencephalon			
Parabrachial nuclei	SP- and NKA-positive neurons project to the amygdala, bed nucleus of stria terminalis, and thalamic nuclei	SP-, NKA-, and NKB-positive fibers from the spinal cord, nucleus of tractus solitarius the reticular formation, the laterodorsal tegmental nucleus, hypothalamic nuclei, bed nucleus of stria terminalis, and amygdala	NK1, NK3
Dorsal and laterodorsal tegmental nuclei	SP- and NKA-positive neurons (some acetylcholine and CRH or ANP) project to the thalamus, hypothalamus, septum, and prefrontal cortex via the medial forebrain bundle	SP- and NKA-positive fibers from the interpeduncular nucleus	NK1, NK3
Pontine raphe nuclei (B5, B6)	SP- and NKA-positive neurons project to NTS	SP- and NKA-positive fibers from medulla (reticular formation, trigeminal, cuneate nuclei, NTS), peduncolopontine tegmental nucleus, other raphe nuclei, and PAG	NK1, NK3

Continued

Table 3 Continued

Area	*TK-positive neurons*	*TK-positive fibers*	*TK receptors*
Locus coeruleus (A6)	SP- and NKA-positive neurons	SP- and NKA-positive fibers from raphe nuclei and reticular formation	NK1, NK2, NK3
Cerebellum	SP-, NKA-, NKB-positive neurons	SP-, NKA-, NKB-positive fibers	NK1, NK2, NK3
Midbrain reticular formation	SP- and NKA-positive cholinergic neurons project to the cerebral cortex	SP-, NKA, and NKB-positive fibers	NK1, NK3
Myelencephalon			
Medullary reticular formation (A7, A5)	SP- and NKA-positive neurons project to raphe, parabrachial nuclei, spinal cord	SP-, NKA- NKB-positive fibers from mesopontine tegmentum	NK1, NK3
Raphe Magnus (B3)	SP- and NKA-positive neurons project to the reticular formation, NTS	SP-, NKA-, NKB-positive fibers from reticular formation, NTS, cuneate	
Medullary raphe nuclei (B1, B2)	SP- and NKA-positive neurons (co-expressing 5-HT and TRH) project to the NTS, spinal cord	SP- and NKA-positive fibers from reticular formation, NTS, cuneate, midbrain and pontine raphe, pedunculopontine tegmental nuclei and PAG.	NK1, NK3
Nucleus of tractus solitarius	SP- and NKA-positive neurons	SP-, NKA-, NKB-positive fibers from caudal raphe. SP- and NKA-positive fibers from trigeminal nodose and petrose ganglia	NK1, NK2, NK3
Dorsal motor vagal nuclei		SP- and NKA-positive fibers from caudal raphe, bed nucleus of stria terminalis, amygdala	NK1, NK2, NK3
Cuneate nucleus	SP-, NKA-, NKB-positive neurons	SP, NKA, and NKB-positive	
Spinal trigeminal nucleus	SP-, NKA-, NKB-positive neurons	SP-, NKA-, and NKB-positive fibers from raphe magnus and PAG	NK1
Spinal cord	NKB-positive interneurons. SP- and NKA-positive neurons projecting to the thalamus and hypothalamus	SP- and NKA-positive fibers from periaqueductal gray, raphe pallidus B1, obscurus B2, and nuclei reticularis gigantocellularis and intrafasciculari hypoglossi. SP- and NKA-positive fibers from primary afferent neurons. NKB-positive fibers	NK1, NK2, NK3

5-HT, serotonin; ANP, atrial natriuretic peptide; CART, cocaine- and amphetamine-regulated transcript; CNS, central nervous sytem; CRH, corticotropin-releasing hormine; GABA, γ-aminobutyric acid; NKA, neurokinin A; NKB, neurokinin B; NTS, nucleus of tractus solitarius; PAG, periaqueductal gray; SP, substance P; TK, tachykinin; TRH, thyrotropin-releasing hormone.

rodents, NK1 exceeds NK3 density. The highest expression of NK3 receptors occurs in the prefrontal cortex. In local circuitries, NK1 receptors are located on neurons expressing nitric oxide, GABA, and glutamate, whereas NK3 sites are specifically linked to glutamatergic transmission. NK3 receptors are also expressed on fibers of GABAergic neurons located in the substantia innominata. The relatively dense expression of NK3 receptors in the prefrontal cortex could be related to the promising results obtained in clinical trial of osanetant, a selective NK3 receptor antagonist, in schizophrenia. Modulatory effects on the TK system have been reported following various conditions such as ischemia, alcohol preference, schizophrenia, and depression. Interestingly, the systemic administration of NK1 or NK2 receptors attenuates stress-induced cortical noradrenaline release; however, these effects could be exerted at the locus coeruleus. NK1 receptor blockade also reduces stress-induced cortical dopamine release by acting putatively at the ventral tegmental area.

TKs and their receptors are expressed in the basal ganglia, such as the nucleus accumbens, globus pallidus, entopeduncular nucleus, and striatum (caudate putamen). The nucleus accumbens is the projection area of the mesolimbic dopaminergic pathway originating in the ventral tegmental area and also receives glutamatergic inputs from cortical areas. The accumbens is believed to serve as interface among emotions, motivation, and movements. Intra-accumbens injection of selective NK1 or NK3 receptor agonists induce a naltrexone-resistant analgesia in the tonic phase of the rat formalin test, suggesting that TKs activate a midbrain dopaminergic component of stress- or pain-induced analgesia. Local stimulation of the NK1 receptor

produced an haloperidol-sensitive increase in locomotor activity, further highlighting the interaction of these receptors with dopaminergic mesolimbic pathway.

The globus pallidus and the entopeduncular nucleus contain a high density of TK fibers, mainly originating from the striatum and creating part of a corticofugal pathway regulating motor activity. Reduced levels of TKs have been described in the postmortem brains of patients affected by Parkinson's and Huntington's diseases.

An increased locomotor activity was observed following NK1 receptor stimulation in the ventral pallidum, a region receiving tachykininergic input from the accumbens. The ventral pallidum is one of the brain areas that has been associated with reinforcing effects mediated by the stimulation of NK1 receptors. Although the injection of C-terminal fragments of SP into the rat ventral pallidum induces an NK1-mediated anxiolytic-like effects in the rodent plus-maze test, the possibility of an unspecific effect due to an increased locomotion cannot be ruled out. Indeed, a large body of evidence indicates that the stimulation of NK1 receptors induces an anxiogenic effects, whereas their blockade induces anxiolytic effects in several animal models.

The striatum is part of the extra-pyramidal motor system that is also involved in the regulation of motivational and cognitive functions. The striatum is the projection area of the mesostriatal dopaminergic pathway; it can be divided into morphologically (but not topographically) distinct areas: the matrix and striosomes. Striatal TK-containing neurons appear to be intimately connected with nigral dopaminergic neurons; dopamine mimetics acutely increase and antagonists decrease striatal TK release, whereas the chronic administration of dopaminergic drugs regulates the expression of TAC1 through a D1 receptor-mediated effect. On the other hand, TKs enhance striatal dopaminergic transmission. In the matrix, NK1 receptors are localized on cholinergic neurons and on somatostatin–nitric oxide synthase-positive neurons, whereas NK3 receptors are probably expressed on both these and dopaminergic neurons. In fact, NK3 receptor stimulation produces a tetrodotoxin (TTX)-resistant dopamine release, whereas that induced by NK1 and NK2 receptor stimulation is totally and partially TTX resistant, respectively. In contrast, in the striosomal area, NK1 receptor stimulation produces a TTX-resistant dopamine release, whereas NK2 receptors control the release of acetylcholine in which the modulation can be positive or negative depending on the viability of the dopamine receptors. The intrastriatal injection of a selective NK1 or NK2 receptor agonist induces atropine-sensitive controlateral rotations in mice, whereas NK1 or NK2 antagonists reduced rotations evoked by the striatal stimulation of D1 receptors. Likewise, the blockade of striatal NK1 receptors reduces amphetamine- or cocaine-induced dopamine release and protects striatal dopaminergic fibers from the toxic effect produced by amphetamine. Alterations in the striatal TK system have been associated with movement disorders in Parkinson's and Huntington's diseases. Reduced SP levels were observed in the striatum of Parkinson's patients, and the degree of reduction of TAC1 expression has been associated to the severity of motor impairment in a model of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridire (MPTP)-induced Parkinsonism. Reduction of SP in the striatal projection to the substantia nigra pars reticulate has been detected in Huntington patients.

As with other brain areas, amygdala is involved in a number of vital functions such as memory, emotions, sexual (one amygdaloid nucleus is more conspicuous in males than females) and aggressive behaviors, and neuroendocrine and autonomic regulations. In this area, NK3 receptors are abundant in rodents but relatively scarce in primates, whereas NK1 receptors predominate. The release of TK acting at the NK1 receptors in the amygdala has been associated with phobias and anxiety in both humans and rodents, and it is an important site where NK1 antagonists exert their anxiolytic effect. NK1 receptors have also been implicated in stress-induced hypertension. Accordingly, SP expression was increased in the medial amygdaloid nucleus (and lateral hypothalamus) in a rodent model of chronic mild stress. NK1 receptor-expressing (i.e., GABAergic) neurons have also been involved in opiate reward because the selective ablation of these neurons prevents morphine-induced conditioned place preference. Local activation of NK3 receptors suppresses salt intake in sodium-depleted rats.

The bed nucleus of the stria terminalis, which is also sexually dimorphic, is the area with the densest network of NKB-positive fibers. Local injection of NKA or its elongated forms suppresses male copulatory behavior. The local stimulation of NK3 receptors suppressed alcohol, water, and food intake in deprived rats. Altogether, these effects suggest that TKs exert a potent rewarding effects at this level.

The hippocampus is associated with memory generation and processing. The role of local NK1, NK2, and NK3 receptors in the regulation hippocampal circuitries is uncertain. It is known that the blockade of NK2 or NK3 receptor reduces the hippocampal release of acetylcholine induced by stress or CRH; however, these effects are likely to be exerted at the cholinergic basal forebrain nuclei such as the septal nuclei and/or nucleus basalis of Meynert rather than locally. Systemic administration of a selective NK1 receptor antagonist produces a dose-dependent increase of

hippocampal acetylcholine release and facilitates inhibitory avoidance learning and open field habituation. However, NK1 receptor antagonists also block SP-induced long-term potentiation, and decreased hippocampal SP levels have been observed following chronic β-amyloid infusion, suggesting that the stimulation of these receptors could elicit both pro- and anti-amnestic effects. Long-term treatment with NK1 antagonists enhances hippocampal serotonin (5-HT) transmission; this effect is believed to underlie the effect of these antagonists in animal models of depression. Hippocampal TKs have also been related to the development of epilepsy; status epilepticus increased hippocampal levels of both SP and NKB, and NK1 antagonists offered some protection against seizures.

The basal forebrain and septal cholinergic nuclei are part of the limbic system regulating a variety of behavioral (cognitive and emotional) and autonomic (cardiovascular, osmoregulation, and thermoregulation) functions. Stimulation of NK1, NK2, and NK3 receptors located on these nuclei elicits hippocampal acetylcholine release; however, NK2 antagonists reduced it only when elicited through CRH or behavioral stimulation. Local stimulation of NK1 receptors produces anxiogenic effects and increases blood pressure; however, given the results discussed so far, it is likely that this area is involved in the anxiolytic effects of NK2 but not of NK1 antagonists.

Diencephalon

The thalamus integrates somatosensory and motor information from both the spinal cord and brain stem nuclei and projects to the basal ganglia and cerebral cortex. In addition, an area close to the geniculate nuclei is part of the pathway regulating circadian rhythms. Systemic administration of a NK2 antagonist selectively reduces the firing of rat posterior thalamic neurons induced by thermal but not mechanical stimuli; however, this effect could be exerted at the spinal cord level. Systemic administration of NK1 antagonist inhibited light-induced phase advancement of locomotor activity in hamsters, indicating that NK1 receptors are the targets of retinic input; however, this effect could involve hypothalamic or mesencephalic, in addition to thalamic sites.

The hypothalamus, the first structure from which SP was isolated, controls endocrine functions through the pituitary, but it is also an important relay for autonomic (cardiovascular and thermoregulatory) and emotional (aggressive-defensive and ingestive) responses. Application of NK1 antagonists in the medial hypothalamus reduces aggressive behavior induced by amygdaloid stimulation.

Several neurohormonal effects are mediated by TKs acting through NK1 and/or NK3 receptors, although it should be pointed out that TKs can act at all levels of the hypothalamic–pituitary–gonadal or hypothalamic–pituitary–adrenal axes to produce neurohormonal effects. A dense network of TK-containing axon terminals are close to the blood vessel in the median eminence. SP acts on the suparoptic nucleus to release both vasopressin and oxytocin through the stimulation of both the NK3 and NK1 receptors. Interestingly, prolonged hypotension alters the dynamics of NK3 receptors at this level, supporting the involvement of NKB in the CNS control of cardiovascular homeostasis. NK3 receptors also control the secretion of gonadotropin-releasing hormone. NK1 receptors are involved in the control of luteinizing hormone and follicle-stimulating hormone release with direction of effects which depend on the follicular phase. Prolactin secretion can also be modulated by TKs. They exert a direct stimulatory effect on the pituitary gland, but at hypothalamic level TKs interact with both inhibitory and stimulatory mechanisms. SP, acting through NK1 receptors, tonically inhibits both the release and the synthesis of CRH by acting on the rat paraventricular nucleus. An NK1 receptor-mediated control of growth hormone secretion has also been postulated. In turn, as a feedback mechanism, a variety of hormones or releasing factors (CRH, ghrelin, TRH, testosterone) exert a regulatory effect on the expression of TK peptides and receptors.

Mesencephalon

The substantia nigra has been traditionally associated with disorders of the extra-pyramidal motor system, but this area also plays an important role in regulating autonomic, neuroendocrine, and emotional processes. Dopaminergic neurons of the substantia nigra (A9) pars reticulata and pars compacta receive direct excitatory input from ipsilateral striatal TK-containing neurons from the matrix and striosomes, respectively. Nigral TK fibers also target nondopaminergic (possibly GABAergic) neurons. In rodents, NK3 receptors, which are located on dopaminergic neurons, are the most abundant; in humans, dopaminergic neurons express NK1, but not NK3, receptors, and the formers are the most represented. The intranigral administration of nonselective TK agonists reproduces several electrophysiological, neurochemical, and behavioral effects of dopaminergic drugs; however, the dissection of the role of each TK receptor has revealed an unexpected complexity of the regulation of mesostriatal dopamine system by TKs. In rat studies, the unilateral intranigral infusion of NK1 receptor antagonists induced ipsilateral rotations, whereas NK1 agonists had no effect on locomotion but elicited vacuous oral movements, an effect enhanced by the chronic blockade of dopamine D2 receptors that suggests

that this mechanism is linked to tardive dyskinesia. Indeed, the infusion of an NK1 agonist in the pars reticulata induced a short-term (1–20 min) decrease of striatal dopamine release followed by a long-term (60–180 min) increase, suggesting that NK1 receptors can either inhibit or excite A9 dopaminergic neurons. However, when locomotor activity was increased by striatal D1 receptor stimulation, this effect was reduced by the nigral blockade of both NK1 and NK3 receptors. The stimulation of NK3 receptors in the substantia nigra compacta elicited locomotion, rearing, and sniffing; these receptors are located on A9 cell bodies, as determined by tyrosine hydroxylase co-localization and selective neurochemical lesions. A similar dopaminergic activation was elicited by uni- or bilateral stimulation of NK2 receptors in the substantia nigra reticulata that was paralleled to an increased metabolism of dopamine in the striatum. Interestingly, the selective stimulation of nigral NK1, NK2, or NK3 receptors elicits tachycardia in normotensive rats, whereas the blockade of these receptors decreases blood pressure in spontaneously hypertensive animals.

The ventral tegmental area (VTA) is the origin of dopaminergic neurons (A10) of the mesolimbic and mesocortical systems, which project to the nucleus accumbens, amygdala, olfactory tubercle, and prefrontal cortex. VTA has been associated with the regulation of locomotion, stress, reward, and the cardiovascular system. There is evidence that NK3 and possibly NK1 receptors are expressed on dopaminergic neurons because selective NK3 and NK1 agonists increase dopamine metabolism in the prefrontal cortex and nucleus accumbens and activate olfactory tubercle neurons. These neurochemical effects are associated with hyperlocomotion in the open field or, when tested in a familiar environment, with a characteristic stereotypic behavior consisting of sniffing, digging, face washings (NK1-mediated), head scratching (NK2-mediated), rearings, and wet-dog shakes (NK3-mediated). A parallel NK1 or NK3 receptor-mediated increase in blood pressure and heart rate has been also reported, whereas the stimulation of NK2 receptors increased heart rate only. These autonomic effects have been postulated to be the cardiovascular component of the defensive reaction. VTA NK1 receptors are also critical for the stress-induced analgesia because the reduction in the tonic phase of the formalin test induced by foot shock stress was prevented by VTA injection of NK1 antagonist. Furthermore, VTA stimulation of TK receptors induces analgesia that was reduced by the blockade of dopamine receptors in the nucleus accumbens, implying that TKs are the key mediators for the activation of the mesolimbic dopaminergic pathway during pain- or stress-induced analgesia. Like in the VTA, stress affects the TK system in the interpeduncular nucleus. The stimulation of NK1 receptors in the VTA has been associated with reinforcement properties of drugs of abuse in rodents, whereas in primates this effect has been attributed to the stimulation of NK3 receptors.

The periaqueductal gray matter (PAG) is another brain area involved in behavioral and autonomic manifestations related to stress, pain, and defense. PAG SP release has been observed following noxious stimuli, and the stimulation of PAG NK1 receptors, which are located on glutamatergic neurons, produces analgesia, produces anxiogenic-like effects, and increases fear-related behavior. NK1 receptor blockade in the nucleus of tractus solitarius (NTS) also prevents baroceptor reflex inhibition elicited by the electrical stimulation of the PAG, indicating that TK-positive neurons in this latter area integrate cardiovascular control exerted through the NTS.

The midbrain dorsal and medial raphe nuclei (B7 and B8) contain both TK-containing fibers and neurons which largely co-express SP and 5-HT in primates but not in rodents. In the dorsal raphe, NK1 receptors are located on glutamatergic neurons, which in turn excite 5-HT neurons (B7). The released 5-HT acts at the somatodendritic 5-HT1A receptors to provide inhibitory feedback. The fact that in some animal species (mouse and guinea pig) the acute administration of NK1 antagonists increase the neuronal activity of 5-HT neurons suggests that the described mechanism is tonically active. In contrast, the administration of NK2 (but not NK1) antagonist decreases the firing of rat 5-HT neurons. Dorsal raphe 5-HT neurons, which project to the hippocampus, are believed to be involved in the antidepressant/anxiolytic properties of NK1 receptor antagonists because hippocampal 5-HT transmission is enhanced following the chronic administration of NK1 antagonists as well as other antidepressant treatments. Activation of NK2 receptors in the median raphe increases rat locomotor activity, an effect dependent on the integrity of 5-HT neurons, whereas the stimulation of NK3 receptors increases locomotor activity and induced straub-tail, hindlimb splaying, and oral movements, suggesting that these receptors control the activation of the mesolimbic and mesostriatal dopaminergic pathways.

Metencephalon

In the pons, the highest density of TK-positive fibers (including NKB) is observed in the parabrachial nuclei, which are important centers for the integrating afferent inputs, including gustatory and visceral nociceptive stimuli, to the thalamic nuclei for the regulation of autonomic functions. Interestingly, the local blockade of NK1 receptors reduces nociception-induced

tachycardia, indicating that the supraspinal sympathetic drive is regulated by TKs.

Noradrenergic locus coeruleus neurons (A6) are activated by SP via NK1 receptors; NK1 antagonists do not acutely affect the spontaneous firing of these neurons, although they can prevent the inhibitory effect of adrenergic α-2 agonists. An increase in the spontaneous firing of A6 can be elicited following the chronic blockade of NK1 receptors. Stimulation of NK3 receptors also activate A6 neurons, and, at least in guinea pigs, these receptors have a permissive role on the NK1-mediated stimulation because noradrenaline release in the prefrontal cortex induced by NK1 agonists can be blocked by selective NK3 antagonists. The blockade of NK2 receptors also reduces stress-induced firing of A6 neurons and noradrenaline release in the prefrontal cortex, suggesting the expression of these receptors on these neurons. Locus coeruleus is involved in a number of physiological responses related to stress, including arousal and associated autonomic manifestations. The main center regulating the micturition reflex is also located close to the locus coeruleus; NK1, NK2, and NK3 receptors exert a supraspinal modulation of this reflex, but it is uncertain whether TKs act in this area.

The cerebellum contains only a few TK-expressing neurons and fibers. NKA and NKB are expressed in granule cells, and the stimulation of either NK2 or NK3 receptors can enhance glutamate neurotoxicity, whereas SP is more abundantly expressed in the glial cells. Dynamic changes in the cerebellar TK system have been observed in models of epilepsy.

Myelencephalon

TKs have been involved in the regulation of a number of autonomic functions at the medullary level, such as the regulation of cardiovascular reflexes (the baroceptor reflex, in particular), respiratory drive, and gastric motility (emesis) and secretions. SP- and NKA-positive afferent fibers from baroceptors and carotid body chemoceptors project to the NTS and ventrolateral medulla and are released during hypoxia; TKs applied at these sites evoke both respiratory and cardiovascular responses. The selective stimulation of either NK1, NK2, or NK3 receptors in the NTS produces bradypnoea, whereas the stimulation of either NK1 or NK3 receptors increases only the tidal volume. Exposure of neonatal rats to hypoxia produced an increase in respiratory frequency, an effect that was prevented by blocking NK1 receptors in the ventrolateral medulla. In adult rats, the hypoxia-induced hyperventilation of NK1 receptor knockout mice was attenuated compared to wild-type counterparts, indicating that TKs acting at these receptors modulate but are not essential for respiratory reflexes, at least in adulthood.

Baroceptor activation also induces the release of TKs in the NTS, where NK1 receptors are also involved in the modulation of baroceptor sensitivity following emotional noxious stimuli or physical exercise because the blockade of NK1 receptors differentially affects the symapathetic and parasympathetic arms of the vascular and cardiac components of the baroceptor reflex.

Medullary TKs are involved in the regulation of motility and secretions of the upper gastrointestinal tract (GI). CNS-acting NK1 antagonists block the emetic reflex induced by a wide array of chemical or physical stimuli in selected animal species. The stimulation of NK1 receptors in the dorsal motor nucleus of the vagus inhibits gastric motility; however, other medullary centers, such as the NTS and area postrema, could also participate in the anti-emetic effects of NK1 antagonists. C-fos expression in the NTS and area postrema induced by acid challenge into the stomach is modulated in a complex manner by excitatory amino acid and NK1 and NK2 receptors; the CNS blockade of these latter receptors prevents stomach acid secretion evoked by the CNS stimulation of transient receptor potential cation channel subfamily V (TRPV)1 receptors, the capsaicin receptors. Interestingly, the stimulation of CNS NK2 receptors can also evoke antisecretory effects (possibly through a dopaminergic mechanism) when gastric secretion is induced by pylorus ligation.

Finally, beyond the amygdala, hypothalamus, and PAG, the medullary areas receiving sensory TK inputs, such as the rostroventral medulla, provide a descending inhibitory control on noxious afferent processing in the spinal cord. This inhibitory control is disrupted in NK1 knockout mice or by blocking NK1 receptors in wild-type counterparts; however, it is still unclear whether spinal or supraspinal NK1 receptors play a pivotal role in this effect.

Spinal Cord

Most of the TK content in the spinal cord originates from fibers of the bulbospinal descending pathways and primary afferent neurons, which mainly project to the dorsal horn and autonomic areas. TK-positive spino-thalamic neurons, some of which co-express enkephalin, have also been detected. Capsaicin-sensitive (TRPV1-expressing) primary afferent fibers contain SP and NKA but not NKB, and they co-localize glutamate and calcitonin gene-related peptide (CGRP). The quantity of TK-positive sensory fibers is much higher in visceral than in somatic districts, so that TKs were postulated to play a more important role in the transmission of noxious information from the former than the latter domains. The vast majority of these fibers are unmyelinated (C fibers), whereas

fewer SP- and NKA-positive fibers are thinly myelinated (A-δ), with small or medium-size cellular bodies (respectively) located in the dorsal root ganglia (DRG). Most of these fibers project to the dorsal horn neurons located in lamina I and II; fewer in laminae III, IV, V; along the central canal (lamina X); and in autonomic areas. These capsaicin-sensitive neurons have been classified as polymodal nociceptors because they are preferentially stimulated by chemical and thermal stimuli and are also involved in encoding mechanosensitivity. Indeed, a certain deficit in encoding the intensity of paw pressure and colonic or bladder distension has been observed in NK1 or TAC1 knockout mice, respectively. Despite this, chemically induced but not hot- or pressure-induced nociceptive behavior was consistently reduced in either TAC1 or NK1 knockout mice, indicating that mechanical and thermal sensitivities were not altered despite the mentioned deficit in the electrophysiological coding of these stimuli. Interestingly, the enhancement of activity of the dorsal horn neurons induced by the repetitive stimulation of C fibers, a phenomenon defined as wind-up, was reduced in NK1 knockout mice, supporting an important role of TKs released from afferent fibers in central sensitization. NK1 receptors are expressed throughout all levels of the spinal cord, and their density exceeds that of NK2 and NK3 in both the dorsal (laminae I, III, IV, V) and ventral (lamina VII, VIII, IX, and X) parts and in the autonomic areas, whereas fewer receptors were found in the outer part of lamina II. NK3 receptors are denser in superficial laminae (I and II), where they co-localize with NOS and are scattered in other regions. A few NK2 receptors were detected in the spinal cord; these are mainly located in lamina I, V, and X and could be expressed by astrocytes rather than neurons.

There is massive electrophysiological and neurochemical evidence that TKs are released in the spinal cord upon peripheral noxious stimulation when NK1 receptors play a prominent role in the excitation of spinal cord nociceptive projection neurons. Despite this, the antinociceptive (or antihyperalgesic) activity of NK1 antagonists is restricted to a few animal models and is absent in humans. Several hypotheses, which are not mutually exclusive, have been put forward to explain this discrepancy.

1. The blockade of NK1 receptors alone is not sufficient to block the transmission of noxious signals, whereas the ablation of NK1 receptor-expressing spinal neurons is an effective antihyperalgesic treatment.
2. NK1 receptors are also located on inhibitory GABAergic neurons located in lamina V, whose excitation provide a strong and a long-lasting (feed-forward) inhibition to nociceptive projection neurons (i.e., SP-induced antinociception).
3. The excitation of NK1-expressing spinal cord nociceptive projection neurons activates supraspinal dopaminergic systems, which, in turn, activate a descending inhibitory control on nociceptive projection neurons (i.e., stress-induced analgesia).

Beyond the processing of nociceptive inputs, spinal cord TKs regulate neurovegetative functions. At the thoracolumbar level, NK1 and NK3 receptors affect preganglionic sympathetic neurons differently and evoke cardiovascular effects in spontaneously hypertensive rats which are qualitatively different from those observed in normal counterparts. Stimulation of NK1 receptors inhibits natriuresis and diuresis by enhancing sympathetic nerve activity. At the lumbosacral level, NK1 antagonists modulate the micturition reflex induced by bladder distension or irritation, whereas the stimulation of NK3 receptors produces an inhibitory effect on this reflex and affects urethral motility, probably through a direct action on Onuf's nuclei. The expression of TKs and TK receptors in CNS areas is detailed in Table 3.

Tachykinins and Their Receptors in the Peripheral Nervous System

In the PNS, TKs are expressed in peripheral branches of a subpopulation of capsaicin-sensitive afferent neurons, enteric neurons, and, under certain circumstances (i.e., axotomy), also in postganglionic sympathetic neurons. Peripheral branches of TK-positive capsaicin-sensitive afferent neurons, which originate from the DRG and trigeminal ganglia (which express TAC1 but not TAC3), are ubiquitous, although represented more in the visceral than somatic regions. Although the three TK receptors can be potentially expressed on the soma of DRG neurons, evidence for their expression on sensory fibers and terminals is scanty. In contrast, TK receptors are located on sympathetic and parasympathetic postganglionic neurons both in their soma within ganglia and in terminals regulating the activity of viscera. Likewise, NK1 and NK2 receptors are expressed by several cell types in peripheral tissues innervated by capsaicin-sensitive sensory neurons; the activation of these receptors mediates part of the an inflammatory response that has been defined as neurogenic inflammation. Several cell types express TAC1 in the enteric nervous system, and TAC3 expression has also been detected, although the amount of NKB was much less than the amount of SP and NKA at this level. NK1, NK2, and NK3 receptors have been found to be expressed on enteric neurons or nerve fibers, in addition to

effector cells. At the peripheral level, TAC4 expression has been shown to occur in nonneuronal cells (mainly leucocytes and reproductive cells).

Peripheral Branches of Afferent Fibers

TK-positive capsaicin-sensitive afferent neurons, which co-express glutamate and CGRP, terminate as free nerve endings in the skin and in the cardiovascular, bronchopulmonary, GI, genitourinary, and hematopoietic systems. The action of capsaicin on these nerve terminals is twofold in that an early excitatory effect is followed by a late neurotoxicity with a consequent depletion of TKs in innervated organs. These peripheral nerve terminals, which can be considered dendrites of the DRG, have the unique property of producing an antidromic release of both SP and NKA (and other mediators) on their stimulation (Figure 2).

The activation of NK1 receptors by peripherally released TKs can induce vasodilatation, promote plasma leakage and glandular secretions, and target a number of inflammatory/immune cells such as mast cells, neutrophils, macrophages, and lymphocytes. The stimulation of NK2 receptors has been associated with smooth-muscle contraction and increases the intraluminal permeability of hollow organs. Fewer effects have been associated with the stimulation of NK3 receptors by TKs released from sensory nerves. At the bronchopulmonary level, they have been shown to modulate parasympathetic outflow, whereas the source of TKs capable of evoking NK3-mediated contraction of the portal vein is still uncertain. Although peripheral terminals of sensory neurons represent an important source of tachykinins in the periphery, note that the role of nonneuronal TKs from inflammatory, epithelial, and neuroendocrine cells has been recently exploited.

In several animal models, the release of TKs at the bronchopulmonary level occurs following stimuli of chemical (irritants such as capsaicin, acrolein, and pollutants), physical (acid and cold), or biological

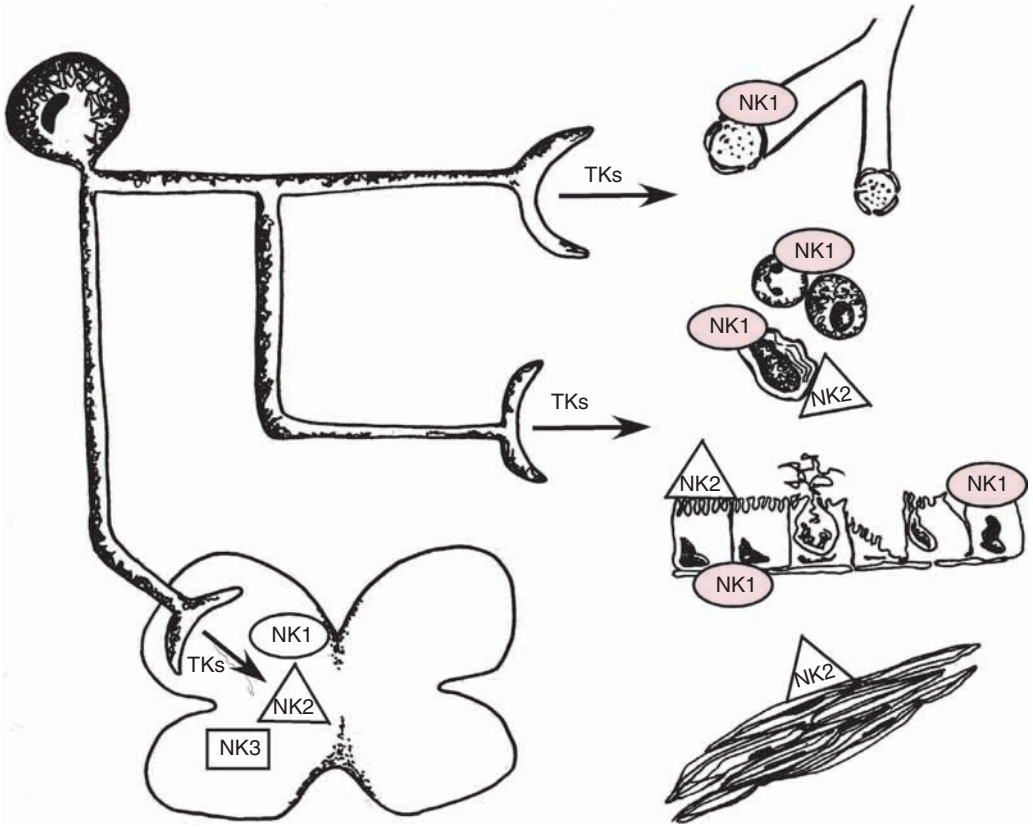

Figure 2 Drawing of a tachykinin-expressing capsaicin-sensitive sensory neuron and its targets at the peripheral and spinal cord level. Tachykinin NK1 receptors are located on (1) blood vessels, where they mediate vasodilatation and plasma leakage; (2) inflammatory cells such as lymphocytes (where they induce the expression of Th1-type cytokines), dendritic cells, macrophages; and neutrophils (where they promote respiratory bursts); (3) epithelial cells, where they favor proliferation; (4) epithelial secretory cells (i.e., globet cells), where they elicit mucous secretion; and (5) the spinal cord, where they participate in the transmission of nociceptive signals. Tachykinin NK2 receptors are located on (1) smooth muscle cells, where they elicit contraction; (2) epithelial cells, where they favor luminal permeability; (3) macrophages, where they favor respiratory burst; and (4) the spinal cord, where they participate in the transmission of nociceptive signals. NK3 receptors are rarely found at the peripheral level in the field of innervation of capsaicin-sensitive sensory neuron, but they can be stimulated at the spinal cord level. NK, neurokinin; TK, tachykinin.

(allergens, viruses, and products of bacterial metabolism) origin. In some animal species (guinea pigs and rabbits) but not in other ones (rats and mice), TKs induce bronchoconstriction through a direct stimulation of NK1 and NK2 receptors located on bronchial smooth muscle, promote airway hyperreactivity, and enhance motor responses driven by parasympathetic nerves (this latter effect is also promoted by NK3 receptors). An important role of NK1 receptors in eliciting mucus secretion and globet cell hyperplasia has also been demonstrated. Although bronchial inflammation (both plasma protein extravasation and recruitment of inflammatory cells) has been largely attributed to the stimulation of NK1 receptors, a minor NK2-mediated component has also been highlighted (intraluminal permeability of airways and recruitment of inflammatory cells). NK1 receptors are also expressed on airway epithelium, fibroblasts, and myoepithelial cells. In humans, the density of TK-positive fibers is much lower than in animals, and exploratory short-term clinical trials with NK1 or NK2 antagonists failed to show a significant TK-mediated component of asthmatic bronchoconstriction.

Only a minor component of the TK content in the GI tract originates from extrinsic capsaicin-sensitive afferent neurons. This applies also for salivary glands in which SP originates from both trigeminal (capsaicin-sensitive) and otic ganglia, in which it is co-expressed with acetylcholine in parasympathetic neurons (capsaicin resistant).

Primary afferent neurons provide the main source of SP and NKA in the urinary tract organs such as the kidney, ureter, urinary bladder, and urethra, and in associated peripheral ganglia (the pelvic and inferior mesenteric ganglia). The stimulation of these nerve terminals produces local motor effects in which NK2 receptors located on smooth muscle play a prominent role in most species, whereas NK1 receptors mediate spasmogenic effects mainly in rats (bladder and urethra) and in guinea pigs (ureter, bladder, and urethra). The stimulation of NK3 receptors can also elicit contractile responses in the rat ureter and urinary bladder; however, in the last organ the response is mediated by the release of TKs from nerve terminals which act through NK1 and NK2 receptors located on smooth muscle. Although this result provides evidence for the expression of NK3 receptors on peripheral terminals of capsaicin-sensitive neurons, the source of the endogenous-preferred agonist for these receptors (e.g., NKB) remains elusive. A role for both the NK1 and NK2 receptors in the enhancement of contractile effects mediated by the parasympathetic nerves has also been demonstrated. Because it occurs in the airways, the stimulation of NK1 receptors produces prominent proinflammatory responses such as vasodilatation, plasma leakage, leukocyte infiltration, mast cell degranulation, production of reactive oxygen species, and the expression of adhesion molecules. A role of NK1 receptors in promoting inflammatory responses in animal models of chemical cystitis has been demonstrated, whereas NK2 receptors play a role in bladder overactivity in models of bacterial cystitis and detrusor dyssinergia that follows spinal cord transection. Again, the contribution of these mechanisms in human diseases is uncertain because TK-positive nerve fibers are more prominent in animals than humans.

Primary afferent neurons provide a source of SP and NKA in both the male and female reproductive systems, although placenta and cumulus granulosa cells, blastocysts, and oocytes represent an alternative source of TKs (including products of the TAC3 and TAC4 genes) at this level. The effects of TKs include the enhancement of uterine and fallopian tube contractility, cervical dilatation, vasodilatation and plasma leakage of uterine vasculature, electrolyte transport across the endometrium, and secretion of steroids from the corpora lutea. The receptors involved in these effects are subjected to both species- and ovarian cycle-dependent variations. TKs have been implicated in stress-induced abortion, but a role in physiological pregnancy can also be postulated because capsaicin pretreatment or chronic treatment with an NK3 antagonist reduces rat litter size.

Enteric Nervous System

In the GI tract, most TKs (SP and NKA) are expressed by several classes of intrinsic neurons which are resistant to the neurotoxic effect of capsaicin. Evidence for the presence of NKB in the GI tract is scanty; its expression overlaps that of SP and NKA. A schematic representation of GI neurons expressing TKs is shown in Figure 3.

From a functional point of view, TKs provide inputs to smooth-muscle cells, interstitial cells of Cajal, myenteric and submucous neurons, glands, and secretory cells. Both NK1 and NK2 receptors mediate a noncholinergic excitatory component to muscular transmission and provide a prejunctional enhancement of cholinergic drive. The relative role of NK1 and NK2 receptors can vary depending on the intestinal segment and the species examined, but in humans the latter receptors play a pivotal role in the postjunctional effects mediated by endogenous TKs. In contrast, in all species examined so far, NK1 receptors are expressed on the interstitial cells of Cajal; these cells are believed to play dual motor and sensory roles because both are involved in the generation of slow waves and could behave also as stretch-sensitive

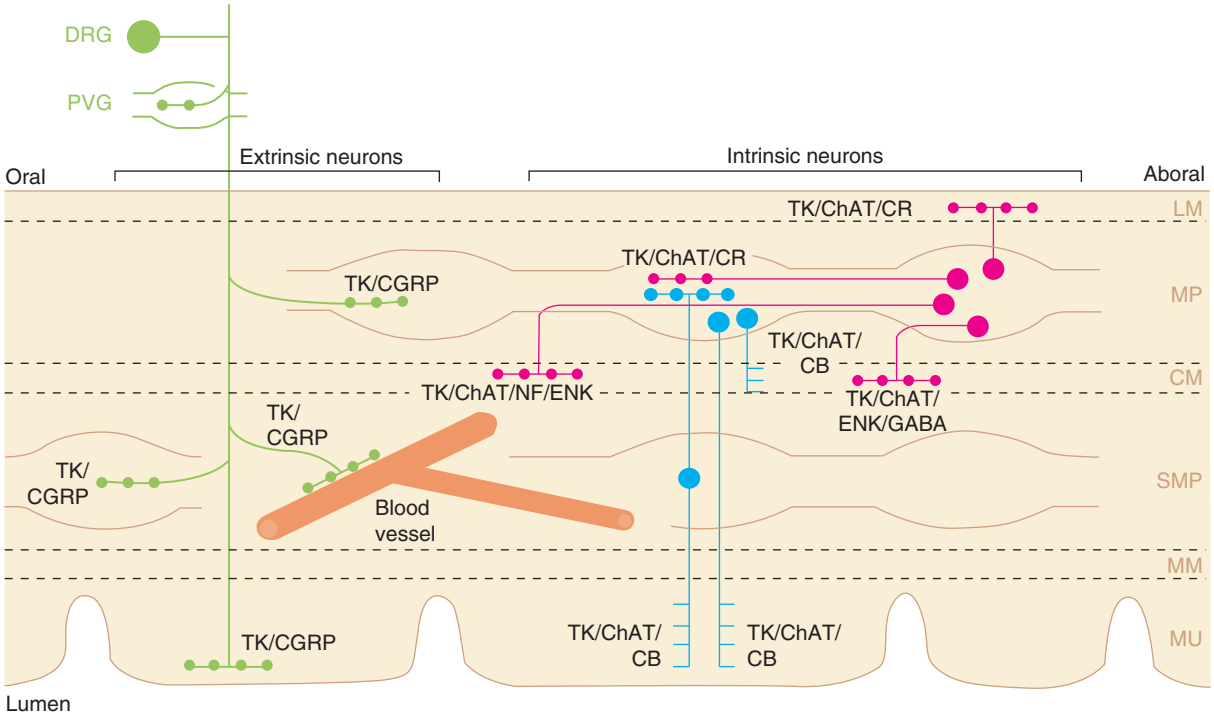

Figure 3 Schematic summary of the most important classes of TK-containing neurons and their polarized (oral-aboral) projections within the guinea pig small intestine. Extrinsic afferent neurons (in green) expressing calcitonin gene-related peptide (CGRP) arise from the dorsal root ganglion (DRG) and on their way to the gut pass through the prevertebral ganglion (PVG). Intrinsic neurons arise from the myenteric plexus (MP) and the submucosal plexus (SMP). Intrinsic primary afferent neurons (in light blue) express cholineacetyltransferase (ChAT) and calbindin (CB); long-projecting ascending excitatory motor neurons (in red) to the circular muscle (CM) express ChAT, neurofilament protein (NF), and enkephalin (ENK); and short-projecting ascending excitatory motor neurons (in red) to the circular muscle express ChAT, ENK, and γ-aminobutyric acid (GABA). Excitatory motor neurons (in red) to the longitudinal muscle (LM) express ChAT and calretinin (CR); ascending interneurons (in red) express ChAT and CR. MM, muscularis mucosae; MU, mucosa; TK, tachykinin. Reproduced from Holzer P and Holzer-Petsche U (2001) Tachykinin receptors in the gut: Physiological and pathological implications. *Current Opinion in Pharmacology*, 1: 583–590, with permission from Elsevier.

structures. NK1 and NK3 receptors are expressed on the soma of descending motor neurons, where their activation produces relaxant effects on smooth muscle. TK-mediated motor effects can also be regulated through an effect on myenteric neurons. NK3 receptors are involved in (slow) excitatory inputs between intrinsic primary afferent neurons, whereas NK1 receptors mediate slow excitatory inputs from intrinsic primary afferent neurons to interneurons. The cell body of ascending interneurons expresses both NK1 and NK3 receptors, whereas both varicosities and the cell body of descending inhibitory interneurons express NK2 receptors. A schematic representation of the GI expression of TK receptors is shown in Figure 4.

This complexity is well reflected in the variety of functional effects exerted by TK antagonists which depend on the species, the intestinal segment examined, and the experimental setting. Despite this, the general role of TKs in the regulation of GI tract physiology can be outlined. Peristalsis of isolated small or large intestine induced by distension is not consistently affected by NK1, NK2, or NK3 antagonists, but inhibitory effects can be observed when more than one TK receptor is blocked. A synergism between muscarinic and NK2 antagonists blocking peristalsis was observed in both the small intestine and colon of several species, whereas a similar inhibitory synergism between NK3 and nicotinic antagonists has been observed in the colon. These results indicate that TKs acting through NK1 and NK2 receptors (the latter is more important in humans) act as co-transmitters of acetylcholine at the neuromuscular junction, and NK3 receptors facilitate cholinergic transmission at synapses. Together the excitatory effects and the selective stimulation of each TK receptor can elicit inhibitory effects on intestinal motility. This effect could have pathophysiological relevance because the selective blockade of NK1, NK2, or NK3 receptors rescued an inhibited motility in rat models of postoperative ileus. Dilatation of mucosal arterioles induced by distension or mucosal stroking is partly mediated by TKs acting at NK3 receptors.

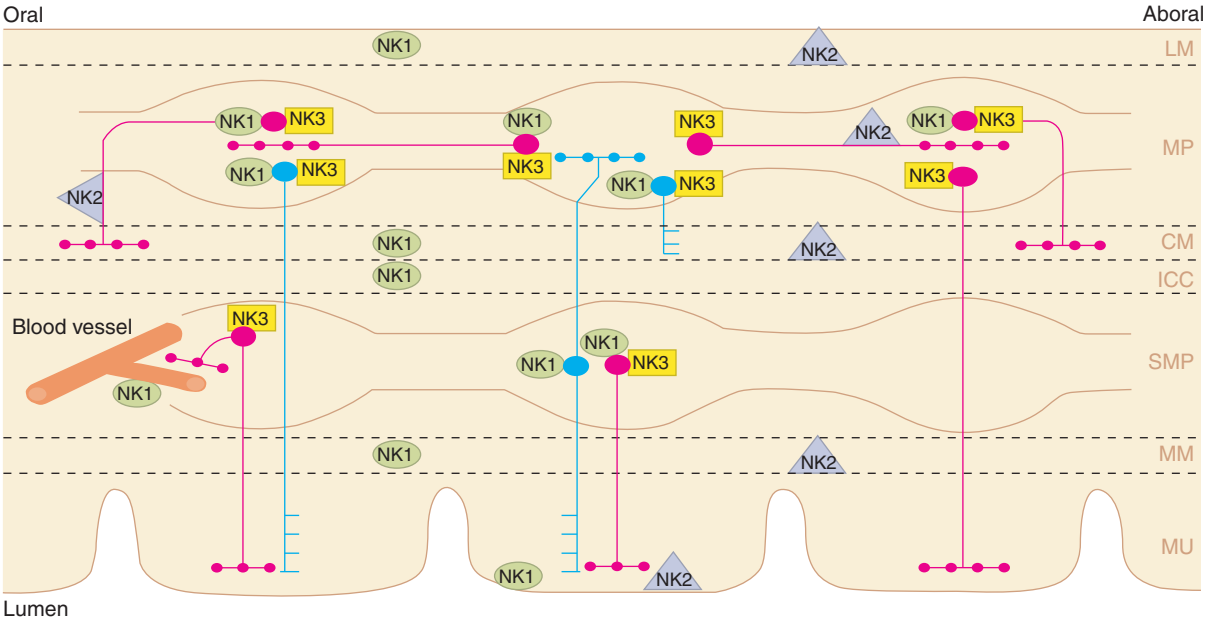

Figure 4 Schematic summary of the cellular expression of tachykinin NK1, NK2, and NK3 receptors in the guinea pig, rat, mouse, and human gut, based on immunocytochemical and pharmacological evidence. Filled circles depict neuronal cell bodies. Neurons in light blue are intrinsic primary afferent neurons. Ascending/descending motor neurons, interneurons, secretomotor, and vasomotor neurons are in red. Green ovals represent tachykinin NK1 receptors, violet triangles represent NK2 receptors, and yellow rectangles represent NK3 receptors. CM, circular muscle; ICC, interstitial cells of Cajal; LM, longitudinal muscle; MM, muscularis mucosae; MP, myenteric plexus; MU, mucosa; NK, neurokinin; SMP, submucosal plexus; TK, tachykinin. Reproduced from Holzer P and Holzer-Petsche U (2001) Tachykinin receptors in the gut: Physiological and pathological implications. *Current Opinion in Pharmacology*, 1: 583–590, with permission from Elsevier.

TKs also activate secretory processes through NK1, NK2, and NK3 receptors. NK1 or NK3 receptors exert this control through secretomotor neurons, whereas NK2 receptors are expressed on effector cells. All TK receptors have been involved in the peripheral modulation of visceral sensitivity; however, the precise mechanism through which TKs sensitize visceral afferents remains to be elucidated. As in other visceral organs, NK1 receptors mediate most of proinflammatory effects induced by TKs; these include vasodilation, promotion of plasma leakage, leukocyte recruitment, mast cell degranulation, stimulation of phagocytosis, and immunoglobulin and cytokine (especially T-helper (Th1) type, those linked to innate immunity) production. NK1 receptors have also been associated with enterocyte proliferation.

A few clinical studies have described the role of TK receptors in humans. NKA switches the small intestine motility from a fasted to a fed pattern through the stimulation of NK2 receptors. These receptors could be also involved in colonic urgency and in stress-induced inhibition of bowel movements, whereas the stimulation of NK1 receptors increases small intestine blood flow. Several drugs targeting the NK1, NK2, and NK3 receptors are being developed for GI diseases such as inflammatory bowel disease and irritable bowel syndrome.

See also: Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Gastrointestinal Control; Cardiovascular Function: Central Nervous System Control; Depression and the Brain; Mammalian Neuropeptide Families; Neuropeptides and Receptors in Glia; Neuropeptides: Discovery; Neuropeptides: Enteric Nervous System; Neuropeptides: Mental Disease; Neuropeptides: Sensory Systems; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Stress Response: Genetic Consequences; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness.

Further Reading

Beaujouan J-C, Torrens Y, Saffroy M, Kemel M-L, and Glowinski J (2004) A 25 year adventure in the field of tachykinins. *Peptides* 25: 339–357.

Couture R, Picard P, Poulat P, and Prat A (1995) Characterisation of the tachykinin receptors involved in spinal and supraspinal

cardiovascular regulation. *Canadian Journal of Physiology and Pharmacology* 73: 892–902.

Culman J and Unger T (1995) Central tachykinins: Mediators of defence reaction and stress reactions. *Canadian Journal of Physiology and Pharmacology* 73: 885–891.

Debeljuk L and Lasaga M (1999) Modulation of the hypothalamo-pituitary-gonadal axis and pineal gland by neurokinin A, neuropeptide K and neuropeptide-gamma. *Peptides* 20: 285–299.

Furuta T and Kaneko T (2006) Third pathway in the cortico-basal ganglia loop: Neurokinin B-producing striatal neurons modulate cortical activity via striato-innominato-cortical projection. *Neuroscience Research* 54: 1–10.

Halliday G, Harding A, and Paxinos G (1995) Serotonin and tachykinin systems. In: Paxinos G (ed.) *The Rat Nervous System*, 2nd edn., pp. 929–974. Sydney: Academic Press.

Helke CJ and Seagard JL (2004) Substance P in the baroceptor reflex: 25 years. *Peptides* 25: 413–423.

Holzer P and Holzer-Petsche U (2001) Tachykinin receptors in the gut: Physiological and pathological implications. *Current Opinion in Pharmacology* 1: 583–590.

Maggi CA (1995) Tachykinins and calcitonin gene-related peptide (CGRP) as co-transmitters released from peripheral endings of sensory neurons. *Progress in Neurobiology* 45: 1–98.

Millan MJ (2003) The neurobiology and control of anxious states. *Progress in Neurobiology* 70: 83–224.

Nussdorfer G and Malendowicz LK (1998) Role of tachykinins in the regulation of the hypothalamo-pituitary-adrenal axis. *Peptides* 19: 949–968.

Otsuka M and Yoshioka K (1993) Neurotransmitter functions of mammalian tachykinins. *Physiological Reviews* 73: 229–308.

Ribeiro da Silva A and Hokfelt T (2000) Neuroanatomical localisation of substance P in the CNS and sensory neurons. *Neuropeptides* 34: 256–271.

Rigby M, O'Donnell R, and Rupniak NMJ (2005) Species differences in tachykinin receptor distribution: Further evidence that the substance P (NK1) receptor predominates in human brain. *Journal of Comparative Neurology* 490: 335–353.

Saffroy M, Torrens Y, Glowinski J, and Beaujouan J-C (2003) Autoradiographic distribution of tachykinin NK2 binding sites in the rat brain: Comparison with NK1 and NK3 binding sites. *Neuroscience* 116: 761–773.

Relevant Websites

http://www.gpcr.org – Molecular Class-Specific Information System (MCSIS) project (sequences of cloned tachykinin receptors from various species).

http://www.ncbi.nlm.nih.gov – US National Institutes of Health.

Suicide

H S Sudak, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Incidence and Epidemiology

There are approximately 30 000 completed suicides per year in the United States and approximately 1 million worldwide. No one knows the actual number of attempts annually, but a conservative estimate is seven attempts per every completion – hence, at least 200 000 in the United States and 7 million in the world. Suicide ranges between the eighth and twelfth leading causes of death in the United States annually and it occurs at the rate of approximately ten per 100 000 in the United States (combined for both genders, all ages and races). The aftereffects of completed suicide extend far beyond the immediate family, leaving devastated relatives, colleagues, and friends. When studying suicide, comparing various groups, it is helpful to compare rates of completions or attempts per 100 000 of a given population at risk. For example, the fact that suicide is the second or third leading cause of death for college-age students is tragic, but this does not mean that their rate of suicide is higher than the average rate for all ages combined; it means that most young adults in such a healthier young population are not dying of heart disease, cancer, etc., so the number of deaths in this 'healthier' population will rank higher as a cause of death for this subpopulation than for a less healthy population. It is important to keep in mind that age-adjusted data are more reliable than crude death rates when comparing populations (unless one is studying rates for specific ages). Finally, when the International Classification of Diseases (ICD) was changed from ICD-9 to ICD-10 in 1997, slight differences in the classification of suicide ensued, making comparisons of pre-1997 data to post-1997 data not entirely comparable.

Gender Differences

Males complete suicide more than 3 times as frequently as females despite the fact that females attempt suicide 3 times more frequently than males. This appears to hold true for all countries throughout the world with the exception of China, where female completions equal or exceed male rates. Female completion rates peak in the middle years, resulting in a convex curve.

Racial and Ethnic Differences

White completion rates are greater than black rates; Hispanic rates are intermediate (except that young Hispanic women have particularly high attempt rates). Native Americans and Alaskan natives have the highest completion rates. South Asian completion rates vary along religious lines: Muslim's have remarkably low completion rates, whereas Hindu rates are relatively high.

Age Effects

Completed suicide is rare in children before puberty. Regarding US rates, from age 15 years through old age, in white males, rates are almost directly proportional to age with the exception that the slope is mildly biphasic – with an initial peak at approximately age 30–45 years and then slightly dipping before continuing to increase. In old age, rates are highest of all. Black US male rates are roughly comparable to white male rates until approximately age 35–40 years, after which black rates level off or decrease, whereas white male rates continue to increase. Female attempt rates are much greater than male attempt rates, especially in younger women (15–40 years old), and the white/black disparity is also seen in female attempt rates (Figures 1–3).

Religious and Socioeconomic Effects

Aside from the low rates in Muslims, religious and socioeconomic effects are not profound. In the United States, Protestants and Jews have slightly higher completion rates per 100 000 of a given at-risk population than do Catholics. Lower socioeconomic status is slightly more associated with completed suicide than is higher status.

Theories Possibly Accounting for These Differences

Very young children lack the means to kill themselves and may also lack the will and despair which accompany suicide. Furthermore, their dependence on their parents for their happiness and survival may also mitigate against suicide, as may their cognitive immaturity vis-à-vis existential matters. Women are more likely to choose suicide methods which are less lethal, less painful, and less disfiguring than those chosen by males (however, both sexes use guns as the most frequent method in the United States, but men choose guns more, proportionately). Thus, the use of pills is greater in females and this allows a larger margin of safety. Another explanation which has been put forth is that

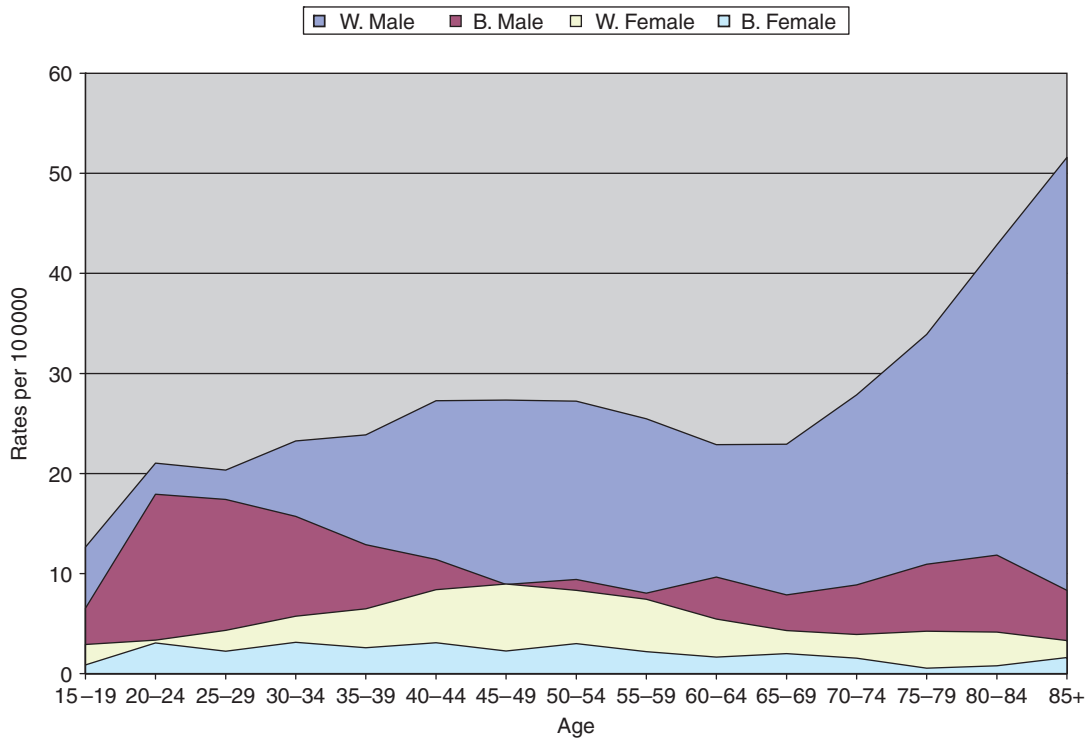

Figure 1 2003 suicide rates by age, race, and gender. Data from www.cdc.gov, Morbidity & Mortality Data, Wisqars program, September 2006.

the greater amount of alcohol abuse and dependence in males accounts for the differences (so that controlling for alcohol use makes male and female completion rates much more comparable). Comorbid alcoholism has also been used to explain the inordinately high completion rates in Native Americans and Alaskan natives. Why women in China have rates comparable to or exceeding male rates may relate to the fact that completed suicide primarily occurs in rural China, where guns and tall buildings are not readily available but agrarian poisons are, and this represents a highly lethal method favored by women. (Furthermore, the male/female disparity in alcohol abuse is probably less in China than in the West.) Explaining lower female rates in general by a focus on methodology is unsatisfactory (i.e., women die by suicide less often because they opt for methods of lower lethality) since questions remain regarding why they opt for such methods. Since rates of depression in general appear to be higher in women than men, and since women make many more suicide attempts than men, the higher male completion rates are all the more remarkable.

There are multiple theories regarding white/black disparities, none of which are very convincing; nor are there satisfactory arguments explaining why black male rates are essentially level or decrease from midlife on, whereas counterpart white male rates continue to increase throughout the life cycle.

Trends during the Past Century

Although the overall rate of completed suicide in the United States has remained remarkably stable since 1900, there have been some dramatic shifts within subgroups. Rates dropped significantly during both World Wars and increased significantly during the Great Depression. For 15- to 24-year-olds, US rates increased 300% for white and black males and for black females and 200% for white females during the period 1950–80 (the rates increased from approximately 8 to 24×10^{-5}, 4 to 12×10^{-5}, 2 to 6×10^{-5}, and 4 to 8×10^{5}, respectively). Overall rates (i.e., rates for whites and blacks and males and females and all ages combined) changed very little during that same period, however, due to a comparable decrease (although less dramatic since it was spread out over a larger age range than the 10-year adolescent/young adult period) in elderly (i.e., 55–85+ years) rates. The cause of this marked adolescent increase in the United States is unknown (guns, drugs, alcohol, and a breakdown in family and religious values were all postulated), but a disproportionate number of adolescents vis-à-vis the remainder of the population engendered by the postwar US baby boom, which disadvantaged these children once they became adolescents, appears to some to be a more plausible explanation. Conversely, the growing population of elderly was advantaged by the advent of Social Security and better

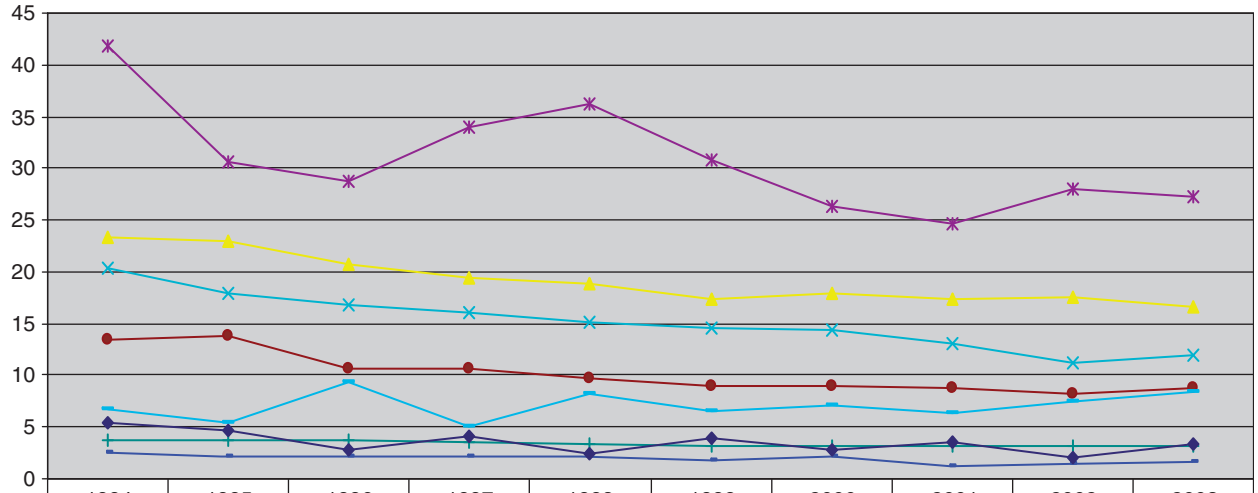

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
White males	23.4	22.89	20.65	19.33	18.95	17.45	17.85	17.42	17.46	16.67
Black males	20.35	17.87	16.8	16.12	15.12	14.48	14.31	13.06	11.26	12.01
Am Indian/AK native males	41.86	30.68	28.78	33.9	36.24	30.75	26.30	24.67	27.99	27.30
Asian/Pac islander males	13.4	13.78	10.71	10.58	9.67	8.87	9.04	8.76	8.21	8.73
White females	3.71	3.82	3.74	3.59	3.43	3.16	3.09	3.09	3.12	3.12
Black females	2.64	2.16	2.21	2.32	2.19	1.91	2.18	1.25	1.66	1.93
Am Indian/AK native females	6.64	5.42	9.38	5.03	8.14	6.61	7.07	6.44	7.46	8.32
Asian/Pac islander females	5.33	4.7	2.75	4.11	2.44	3.88	2.77	3.48	2.03	3.28

Figure 2 Suicide rates for 15- to 24-year-olds from 1994 to 2003; rates per 100 000, by race, gender, and ethnicity. Data from www.cdc.gov, Morbidity & Mortality Data, Wisqars program, October 2006.

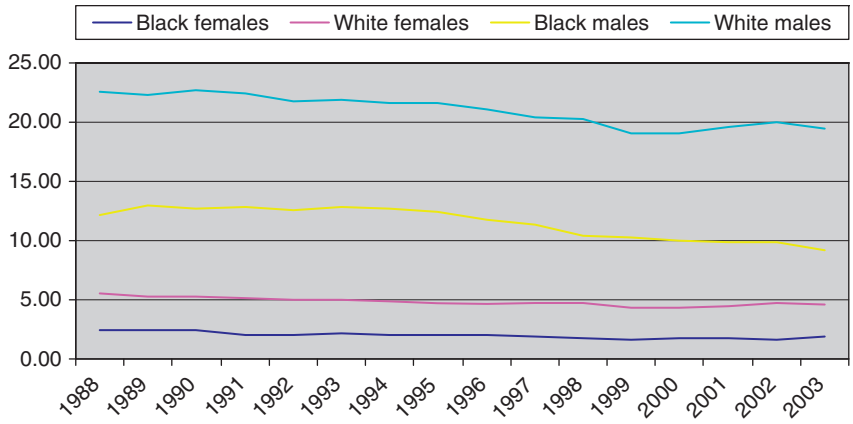

Figure 3 Age-adjusted suicide rates, 1988–2003, by race and gender. Data from www.cdc.gov, Morbidity & Mortality Data, Wisqars program, October 2006.

physical and mental health, which may have improved their quality of life sufficiently to favorably impact their suicide rates.

The 1990 US rates were generally higher than ever, and by 2000 the rates had returned to 1980 levels. Since then, completion rates have continued to fall by approximately 15–35% (up to 2003, the most recent year for which US data are available). Explanations abound for this also; for example, the baby boomers have grown up and no similar disproportionately large bolus of adolescents is moving through the life cycle, or the increased use of antidepressant medications has impacted favorably on suicide rates (supported also by studies of US county-by-county and international country-by-country utilization of such medications revealing inverse correlations between those counties and countries using the most antidepressants having lower completion rates than those that used less such medication). Rates have fallen more for the young and elderly, leading some to question the selective serotonin reuptake inhibitor (SSRI)/serotonin–norepinephrine reuptake inhibitor (SNRI) hypothesis accounting for decreasing rates. Since overall rates are essentially driven by white male rates and, for at least the past 30 years, the white male versus age rates have shown a biphasic pattern with a slight dip in early midlife, this may be a relatively lower risk period which might make this age group less sensitive to antidepressant/antisuicidal effects of SSRI/SNRIs. A similar argument may help explain the smaller decrease in female completion rates.

Etiology

It is helpful to view suicide as a symptom, like fever, rather than a disease. As such, one still may view it from a biopsychosocial perspective. Biologically, although no definitive gene or genes have yet been discovered which are uniquely linked to suicide, there are a number of studies which indicate that there may be significant genetic contributions to suicide. These genetic predispositions are independent from the genes resulting in depression, as suggested by Roy's finding of a concordance rate for completed suicide of 13.2% for monozygotic twins compared to 0.7% for dizygotic twins. Psychological explanations for suicide focus on internalized aggression and/or states of hopelessness, helplessness, desperation, etc., which often accompany depression. Socially, there are marked differences in the frequency of suicide across ethnic and religious boundaries (e.g., very low suicide rates in Muslims and very high rates in Chinese women and Native Americans). More than 90% of those who complete suicide meet criteria for *DSM-IV* axis I or axis II disorders, and of those who do, 80% or more meet criteria for one of the depressive disorders. Consequently, effective treatment of depression is the primary focus for suicide reduction efforts. Bipolar depressive disorders carry the highest lifetime prevalence of death by suicide (15–20%), followed by major depression (10–15%) and schizophrenia (10%). Comorbidities, particularly alcohol and substance abuse, and also various axis II disorders raise these risks markedly.

Two hypotheses utilize stress–diathesis models which combine nature and nurture effects regarding suicide. Both posit a predisposition to impulsivity as the genetic diathesis (alternatively, very early traumatic life experiences such as an infant's prolonged separation from his or her mother can become the underlying diathesis). When such a predisposed individual becomes stressed later in life, this stress, coupled with the tendency to

impulsivity, is more likely to result in suicide. John Mann and colleague at Columbia focus on biogenic–amine deficiencies as the primary pathway mediating this, whereas Charles Nemeroff and associates at Emory University view cortical-hypothalamic axis dysregulation as the primary pathway.

The Suicide Assessment

Every patient who presents with signs or symptoms of depression needs to be asked about suicide (this includes patients who deny depression yet present with vegetative signs of depression – sleep or appetite or other autonomic or hormonal changes – in the absence of clear medical etiology). The clinician should be alert for opportunities to tactfully inquire about suicidality. For example, if the patient complains of feeling sad, one might ask, "Does it get so bad you wish you were dead?" If the patient says he or she does, one might then ask directly, "Have you been thinking about hurting yourself?" A positive response requires further inquiry about a method and the means to implement it: is the patient afraid that he or she might actually harm him- or herself? Although a positive response to that question is clearly of immediate concern, a negative response is only reassuring if the clinician is convinced that the patient is 'not concerned' because he or she has decided to end it all and is relieved by finally making such a decision.

Even if the apparently depressed patient denies being suicidal, a thorough family history (of depression and suicide attempts/completions) is needed and a careful investigation of the patient's past history of depression and suicide is warranted. For patients who have made prior suicide attempts, in addition to questions regarding any current attempt, the following questions should be addressed: how lethal were any past attempts? how likely was the possibility of rescue each time? what were the precipitants of each or did they appear completely inexplicable and impulsive? what was the most serious of the attempts? was the intention really to die?

A patient may deny that he or she is suicidal but the patient's presentation or past history may concern the clinician sufficiently to require contact with family members or close friends in order to ascertain if the patient has expressed suicidal wishes to them (which is more likely than the patient expressing them to the clinician). Should a patient present in a clinically sufficient alarming manner but deny suicide, the clinician may (and should) waive confidentiality if the patient refuses permission to contact the family, in order to obtain history and/or to help protect the patient from harming himself or herself.

Suicide assessments need to be performed as often as clinically indicated on a given patient (e.g., for the hospitalized psychiatric patient, particularly during the first few months of antidepressant administration).

'Why now in and why now out?' provides a useful reminder when evaluating patients for inpatient status and when one is considering them for discharge. That is, what precipitated a patient's need for admission? Also, when evaluating the safety of discharging such a patient (assuming he or she had been admitted), what has now changed enough to make discharge reasonable?

Inpatient psychiatric staff, especially nurses, have come increasingly to rely on 'no-suicide' or 'safety contracts' whereby the patient verbally or orally contracts with staff to notify them if he or she begins to feel acutely suicidal. Staff feel more confident about patient safety with these contracts because they provide quick, often written, documentation that the patient has been asked about suicide and indicated that he or she is not suicidal and promises to notify staff if he or she becomes so. They do not take the place of a suicide risk assessment. Psychiatrists and other mental health professionals should be wary of these since they lull staff into thinking the patients are safe, misleading them into feeling that a more meaningful risk evaluation is unnecessary and consequently may do more harm than good. I believe they are only useful in the negative – that is, when a patient refuses to 'contract,' it should be viewed as a red flag warning of elevated risk.

Prevention

Universal prevention strategies include generally improving the quality of people's lives, thereby reducing stress and unhappiness (cf. decreased rates in the elderly), and decreasing the availability of lethal means, such as control of guns, placing barriers on bridges, and limiting the amounts of acetaminophen per packet. Selective strategies include gatekeeper training at schools and institutions so that depressed and suicidal individuals can be identified and treated before they harm themselves and focusing on high-risk groups such as bereaved individuals or publicly shamed individuals in positions of high trust, such as disbarred attorneys, defrocked clergy, and de-licensed physicians. Indicated prevention refers to treating those already diagnosed as depressed and suicidal or those who have already made an attempt in order to minimize morbidity and prevent death. A roughly comparable schema divides prevention efforts into primary (i.e., changing the conditions which give rise to the disease), secondary (treating the disease), and tertiary (minimizing the progression of the

disease). Suicidologists often match these to prevention, intervention, and postvention.

It is useful to remember that the profile of the typical completer is male, white, older, and with an axis I diagnosis. For attempters, being younger, female, and having an axis II diagnosis are more characteristic. However, attempters and completers clearly overlap: one cannot dismiss a young woman with borderline personality disorder as not seriously suicidal merely because she does not fit the completer profile. It is estimated that 8–10% of attempters do ultimately complete suicide.

It is impossible to prevent all suicides and impossible to expect psychiatrists to predict all suicides. The standard of care for psychiatrists (and for non-psychiatrist physicians treating such patients for their psychiatric conditions) is for them to appropriately evaluate the degree of suicidal risk for a patient and to act accordingly.

Risk Factors and Warning Signs

A working group of the American Association of Suicidology has published a consensus set of warning signs for suicide. These signs, in contrast to risk factors, are more subjective, clinically derived, and imply imminent risk (risk factors being more objective, empirically derived, and less proximate in terms of time). The main warning signs are someone threatening to hurt or kill him- or herself; seeking ways (means) to hurt him- or herself; and talking or writing about death, dying, or suicide. These warrant the witness calling 9-1-1 or seeking immediate help from a mental health provider. Other warning signs overlap with many of the following risk factors: hopelessness, angry, recklessness, trapped feelings, increased alcohol or drug use, withdrawal, anxiety or agitation, sleep changes, dramatic mood changes, and no reasons for living.

Although the greater the number of risk factors, the greater the risk, risk factors are not simply additive – not all risk factors are equally risky. For example, a history of a prior attempt has the greatest predictive value, whereas a diagnosis of a major depression or bipolar depression and comorbid substance abuse also substantially increase the risk, but slightly less than a prior attempt. Hopelessness has traditionally been viewed as the best affective predictor of suicide, but other intense affects, especially desperation but also marked anxiety and agitation, may be equally, if not more, predictive. Asking about guns in the home should be an essential part of any evaluation of a depressed patient.

Risk factors can also be classified as static (e.g., race, gender, and past and family history of depression and suicidal behavior) as opposed to dynamic (e.g., degree of depression and the presence of psychosis). Another classification is based on the likely proximity to a suicidal act occurring (e.g., desperation, hopelessness, extreme agitation, and anxiety are viewed as requiring immediate steps to protect the patient; having suicidal ideation without a specific plan and having coexistent physical health problems are viewed as indicating less proximate risk). These distinctions are approximate and a careful interview of the patient and family members is essential when appraising risk. The examiner's subjective sense of how much psychic pain the patient is experiencing is another important indicator (with two exceptions: the patient who has made up his or her mind and feels calmed by having decided to die, and a retarded, rather than an agitated, depression, where the actual degree of internal distress may be masked).

Treatment

As with the previous discussion of etiologies, treatments are best via a comprehensive, biopsychosocial approach, but for simplicity we consider the biological and the psychological separately.

Suicidal behavior viewed as a symptom mandates that we treat the underlying disease. However, since the suicidal patient is at risk to act on his or her feelings, we need to ensure his or her safety during the period required before a serious depression can be brought into remission by medication and/or psychotherapy. Aside from physically protecting the patient via inpatient or partial hospitalization or charging the family with protecting the patient, as appropriate, there are two critical, simple interventions which can provide prompt relief to the agitation/anxiety and insomnia from which such a large proportion of suicidally depressed patients suffer. Liberal prescribing of antianxiety agents, such as clonazepam or other benzodiazepines, can quickly diminish anxiety (and they, or other agents, may also be prescribed to help the patient sleep). Manic or hypomanic patients can benefit enormously from relief of insomnia. A number of second-generation antipsychotics (SGAs) have begun to be used for rapid relief of anxiety as well. Furthermore, some SGAs have been approved as first-line treatments of bipolar depressed patients. Psychoses also generally respond fairly quickly to SGAs.

The pharmacological treatment of depressions, *per se*, is via antidepressants, generally SSRIs or SNRIs with or without a mood stabilizer such as lithium, valproate, carbemazepine, or lamotrigine. For bipolar depressed patients, caution regarding the use of antidepressants is advised since they may result in a switch to a manic state. More detail regarding pharmacological or other treatments for treatment-resistant patients

can be found in other articles in this encyclopedia. It should be noted, however, that there are few treatments for suicidally depressed individuals which have a convincing evidence basis, but two that do are clozapine for schizophrenic patients and lithium for bipolar patients. Long-term lithium use greatly reduces (up to sevenfold) the rate of completed suicide in such patients (but suicide rates markedly increase if lithium is abruptly discontinued, necessitating a gradual tapering of this agent when it is to be discontinued).

The major controversy in the pharmacological treatment of depression centers on the dilemma of possibly increasing suicidal behavior by the use of SSRIs/SNRIs. In the United States, the Food and Drug Administration mandated black box warnings on SSRI/SNRI labels indicating that suicidal behaviors may be increased in children and adolescents taking such medications. This was broadened to include anyone younger than 25 years of age in December 2006. In almost all of the studies leading up to the warnings, the suicidal behaviors which increased referred to thoughts or attempts, not completions. One study followed children and adolescents who had been hospitalized for depression and compared suicidal behaviors subsequent to discharge in subjects who had taken or not taken SSRIs/SNRIs. The study found a significant difference for attempts and a less robust, although possibly still significant, difference for completed suicides. There were nine completed suicides and this small number resulted in a confidence interval from 1.65 to infinity. As discussed previously, increased use of these agents correlates dramatically with lower completion rates both nationally and internationally, and since the black box warnings were instituted, the completed suicide rates for children and adolescent have begun to rise again – apparently substantiating fears that the warnings would dissuade parents and others from using these medications and reverse the downward trend in suicide rates for which these medications were probably responsible. A Finnish study replicated the finding of markedly lower completion rates in patients treated with antidepressant coupled with markedly higher attempt rates in such patients. All authorities agree that parents, physicians, and adults taking these medications should be alert to the possibility that these medications may have this paradoxical effect so that patients need to be closely monitored for the emergence of such feelings, particularly in the early phases of therapy.

Psychological approaches to suicidally depressed patients encompass the usual modalities of individual, group, and family approaches and various schools of therapy (e.g., psychodynamic, interpersonal, supportive, and cognitive–behavioral). Of these approaches, only cognitive–behavioral therapy (CBT) has clearly demonstrated effectiveness in reducing suicide attempts (and, possibly, completions) in depressed patients. Dialectical behavior therapy (DBT) has also demonstrated effectiveness in decreasing suicide attempts in borderline patients.

Most mental health professionals subscribe to the view that combined psychotherapy/pharmacotherapy represents the optimal approach for suicidal patients. In depression, either approach results in a 50–60% improvement rate which can be increased to approximately 60–70% when the two are combined. There are also reports that CBT improvements in depression may last longer than those resulting from medication. This may also be true for other psychotherapies (e.g., (interpersonal therapy)).

Special Populations

Adolescents

Depressed adolescents may not present with the same symptoms as do adults. Changes in behavior, friends, or grades may be more likely; irritability rather than sadness is typical; and increased or decreased alcohol and drug use is not uncommon. A refusal to comply with parental requests to see a psychiatrist or psychologist is far more likely than agreement (so that parents often need to insist on compliance and, if indicated, continued treatment). It is not 'cool' to be in such treatment, but it is less uncool to say 'my parents are making me' and most such coerced adolescents later express gratitude for having been so pressed. Adolescents appear more vulnerable to imitative or copy-cat suicidal behavior, resulting in various national organizations publishing guidelines for how the media should report suicide (e.g., do not romanticize the event, do not memorialize the decedent in the school, and do not describe the methods used) and for how schools should intervene following a suicide in order to hopefully prevent additional suicides.

Elderly

The elderly also may not manifest depression in the classical depressed manner. For example, they may simply become more withdrawn or may develop vague somatic concerns for which no apparent organic cause can be found. Although the elderly comprise only 12% of the population, they constitute 19% of the total number of annual completed suicides, making them the group with the highest rate. In addition, they have the lowest number of attempts per completion of any group. This means that they provide the fewest warnings of their suicide (via nonfatal

attempts) of any group and also that when they do attempt, it is more likely to be an attempt of high lethality. An additional problem confronting the depressed elderly is that their depression too often gets dismissed by their physicians and caretakers as to be expected in the old or as untreatable because 'it's depressing to get old.'

The Medically Ill

Those with axis III *DSM* disorders, especially those with chronic, painful, crippling, or lethal illnesses, are at greatly increased risk of suicide, particularly if they are also suffering from depression. Central nervous system (CNS) diseases, such as multiple sclerosis, Huntington's disease, AIDS dementia, head injuries, and malignancies, appear to carry very high risk, compounded by any accompanying CNS disinhibition.

Survivors

'Survivors' of suicide refers to the families and close friends of those who have completed suicide (not to people who have attempted and survived, although this is a population which receives insufficient support as well). It is generally believed that due to the guilt and stigmatization borne by such survivors, grieving a suicidal death is the most prolonged and painful of all. Among such survivors, the grief of parents over a child lost by suicide may be the most intense of all. Parents feel responsible for their children's health and happiness and believe they have failed; they have also lost a part of themselves. Groups such as Survivors of Suicide exist in most US cities, where survivors provide support to others so bereaved. These are usually, although not universally, lay led.

Psychiatrists, psychologists, and other mental health professionals constitute another group who often become survivors of their patients' suicides and who are in need of support (which they rarely receive) following such tragic professional events.

Future Directions

With 30 000 deaths per year (2001) by suicide (compared to 20 000 per year via homicide), suicide is clearly a public health problem and should be addressed in public health terms. Former US Surgeon General David Satcher accordingly issued *Preventing Suicide: A Call for Action* in 2002, in which he set down specific goals for suicide reduction. Since depression is found in 70% or more of suicide decedents and since most depressions are treated by primary care physicians rather than psychiatrists, it makes eminent sense to better educate such physicians to more accurately recognize depression and suicide risk and to treat such depression more adequately than they typically do. Reviews have revealed that of patients receiving antidepressant treatment by their primary care physicians, less than 25% receive adequately aggressive pharmacotherapy. Additionally, better educating the public and gatekeepers about the signs of suicide and how to respond to them should be useful. Similarly, educating the media about the need to not romanticize articles about suicide or publishing details about the methods used may help to decrease the number of imitative or copy-cat suicides. Surely, more effective antidepressants than the current SSRIs and SNRIs will be developed (current ones appear no better than tricyclics, although they are safer and better tolerated). Preliminary data suggest that faster treatments (e.g., intravenous ketamine or scopolamine hydrobromide) or newer approaches (vagal nerve stimulation and transcranial magnetic stimulation) may yet prove effective. Also, the increased use of CBT and DBT by increasingly larger groups of providers should have very beneficial effects on suicide rates. Pooling of data from larger collection sources (e.g., National Institute of Mental Health/American Foundation for Suicide Prevention's Developing Centers on Interventions for the Prevention of Suicide at large universities and medical centers) may finally enable us to have a large enough population base to allow for randomly controlled, prospective, placebo versus treatment arm studies of completed suicides so that we can actually learn what works the best, the fastest, and the safest.

See also: Adolescent Brain Development and the Risk of Psychiatric Disorders; Anxiety: Drug Therapy; Depression and the Brain; Stress and Suicide; Stress Response: Sex Differences.

Further Reading

American Psychiatric, Association (2003) *Practice Guidelines for the Assessment and Treatment of Suicidal Behaviors*. Washington, DC: APA Press.

Brown GK, Ten Have T, Henriques GR, Xie SR, Hollander JE, and Beck AT (2005) Cognitive therapy for the prevention of suicide attempts. *Journal of the American Medical Association* 294: 563–570.

Conwell Y, Duberstein PR, and Caine ED (2002) Risk factors for suicide in later life. *Biological Psychiatry* 52: 193–204.

Dubicka B, Hadley S, and Roberts C (2006) Suicidal behavior in youths with depression treated with new-generation antidepressants. *British Journal of Psychiatry* 189: 393–398.

Fawcett J (2006) Depressive disorders. In: Simon R and Hales R (eds.) *Textbook of Suicide Assessment and Management*, pp. 255–276. Washington, DC: American Psychiatric Publishing.

Geddes JR, Burgess S, Hawton K, Jamison K, and Goodwin GM (2004) Long-term lithium therapy for bipolar disorder:

Systematic review and meta-analysis of randomized control trials. *American Journal of Psychiatry* 161: 217–222.

Gibbons RD, Hur K, Bhaumik DK, and Mann JJ (2006) The relationship between antidepressant prescription rates and rate of early adolescent suicide. *American Journal of Psychiatry* 163: 1898–1904.

Hawton K, Townsend E, Deeks J, et al. (2001) Effects of legislation restricting pack sizes of paracetamol and salicylate on self-poisoning in the United Kingdom: Before and after study. *British Medical Journal* 322(7296): 1203–1207.

Holinger PC and Offer D (1984) Toward the prediction of violent deaths in the young. In: Sudak HS, Ford AB, and Rushforth NB (eds.) *Suicide in the Young*, pp. 15–29. Littleton, MA: John Wright/PSG.

Mann JJ, Waternaux C, Haas GL, and Malone KM (1999) Toward a clinical model of suicidal behavior in psychiatric patients. *American Journal of Psychiatry* 156: 181–189.

McKeown RE, Cuffe SP, and Schulz RM (2006) U.S. suicide rates by age group, 1970–2002: An examination of recent trends. *American Journal of Public Health* 96: 1744–1751.

Meltzer HY and Okayli G (1995) Reduction of suicidality during treatment of neuroleptically resistant schizophrenic patients: Impact on risk–benefit assessment. *American Journal of Psychiatry* 152: 183–190.

Newport DJ, Stowe ZN, and Nemeroff CB (2002) Parental depression: Animal models of an adverse life event. *American Journal of Psychiatry* 159: 1265–1283.

Olfson M, Marcus SC, and Shaffer D (2006) Antidepressant drug therapy and suicide in severely depressed children and adults: A case–control study. *Archives of General Psychiatry* 63: 865–872.

Roy A, Rylander G, and Sarchiapone M (1997) Genetic studies of suicidal behavior. *Psychiatric Clinics of North America* 20: 595–611.

Simon RI and Hales RE (2006) *Textbook of Suicide Assessment and Management*. Washington, DC: American Psychiatric Publishing.

Sudak HS (2005) Suicide. In: Sadock B and Sadock V (eds.) *Comprehensive Textbook of Psychiatry,* 8th edn., pp. 2442–2453. Philadelphia: Lippincott Williams & Wilkins.

Relevant Websites

http://www.afsp.org – American Foundation for Suicide Prevention.

http://www.cdc.gov – Centers for Disease Control and Prevention, Morbidity and Mortality Data, Wisqars Program.

http://www.sprc.org – Suicide Prevention Resource Center.

Superior Colliculus

R H Wurtz, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

The superior colliculus (SC) is a structure in the midbrain that is part of the brain circuit for the transformation of sensory input into movement output. Its major function is orienting the animal, particularly with eye movements, to objects of interest in the outside world. This article first summarizes the basic functional architecture of the SC, then considers how the activities of neurons in the SC lead to the generation of eye movements, and finally addresses questions about the role of the SC in primates, including its role in providing information about impending movements to the cerebral cortex.

Functional Structure of the SC

The SC lies on the roof of the midbrain of the brain stem (Figure 1(a)). In mammals it has six layers and is the direct homolog of the optic tectum of birds, fish, reptiles, and amphibians. In these lower animals it is the major center in the brain for the receipt of the output of the retina. In mammals, and particularly in primates which we concentrate on here, the SC continues to receive retinal inputs but it also receives massive inputs from cerebral cortex and subcortical structures such as the basal ganglia.

The monkey SC has six alternating gray and white layers with an additional small fiber layer on its surface (the stratum zonale). These layers have different characteristics and distinguishing them is critical to describing and understanding the functional organization of the SC. The superficial layers consist of a gray and a deeper white layer (the superficial gray and the optic layers) followed by the intermediate gray and white layers and the deep gray and white layers (Figure 1(b)).

These layers have different inputs and outputs as well, and Figure 1(c) indicates the major ones. The input to the neurons in the dorsal part of these superficial layers is directly from the retina, but with increasing depth in the superficial layers the visual responses are more and more dependent upon the input from visual cortex: ablation of the primary visual cortex greatly reduces the visual responses of the neurons. While the terms gray and white layers imply pure neuronal and fiber layers as is the case in cerebral cortex, there are neurons scattered in the white layers as there are fibers of passage in the gray layers.

A major output of the visual layers is the descending one onto the intermediate layers below, which has recently been demonstrated in slice preparations of the rodent SC. The major ascending projections are up to the cerebral cortex after a synapse in the pulvinar nucleus (Figure 1(c)). One possibility is that the superficial layers of the SC lie on a second pathway from retina to cortex in addition to the prominent retina, lateral geniculate, visual cortex pathway. An alternative view is that the SC to pulvinar to visual cortex pathway provides a combination of sensory–motor information that provides a modulator input to the cerebral cortex rather than a purely visual input.

The major inputs to the intermediate layers of the SC are from the superficial layers, from regions of cortex beyond the primary visual areas such as the lateral intraparietal area of parietal cortex and the frontal eye field of frontal cortex. A major subcortical input is from the substantia nigra pars reticulata of the basal ganglia. A major output of the intermediate layers is a pathway, which crosses from one side of the brain to the other (the predorsal bundle) that conveys the collicular output to neurons in the deeper midbrain and pons. The critical point is that the pathway crosses the midline at this point so that activity in and above the SC is related primarily to saccades made toward the visual field contralateral to the side of the brain of the neuron, whereas neuronal activity below the SC is related to saccades made to the ipsilateral side. A major ascending output is to frontal cortex through the medial dorsal nucleus of the thalamus.

Neurons in the superficial gray and optic layers give a short latency (about 40 ms) response to the onset of a visual stimulus (Figure 2). The optimal stimulus is usually a small spot or object in one part of the visual field contralateral to the SC in which the neuron is located. Neurons in the intermediate gray and white layers also show a brief burst of neuronal activity after the onset of a visual stimulus, but in addition show increased activity beginning before the monkey makes rapid or saccadic eye movements. Many neurons in the deeper layers of the SC (deep gray and white layers) have properties similar to those in the intermediate layers, and they also frequently respond to sensory modalities other than vision such as auditory and somatosensory stimuli. The multimodal nature of these neurons indicates that a movement initiated by activity in the SC may result from input other than the visual or in conjunction with the visual. Less is known about the detailed organization

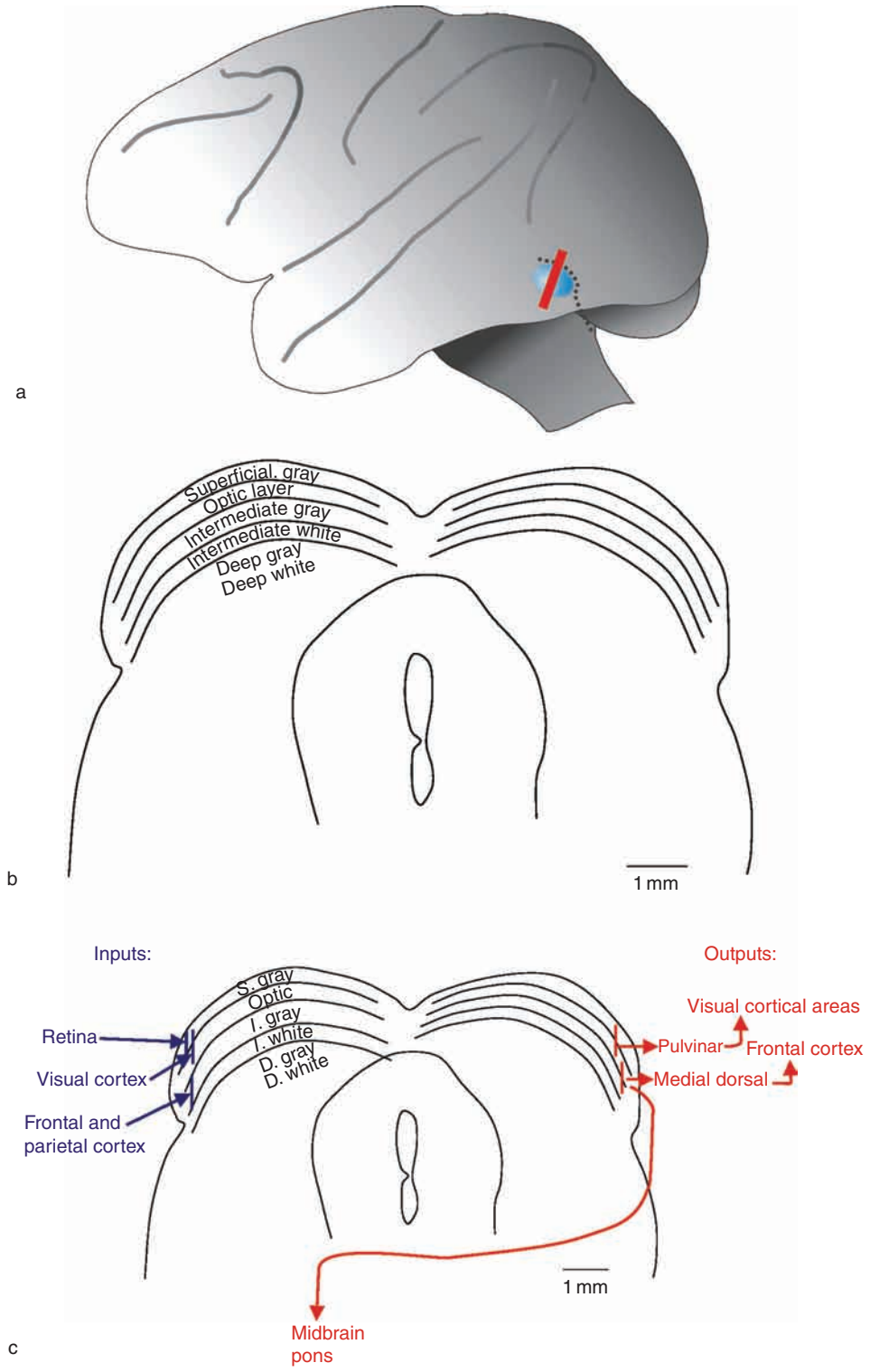

Figure 1 Anatomical structure of the superior colliculus (SC). (a) Location of the SC in the brain. The SC lies on the roof of the midbrain which is covered by the overlying cerebral cortex. The red vertical line indicates the plane of section through the SC (frontal or coronal) shown in (b). (b) Layers of the SC shown on the coronal section. The two sides of the SC are symmetrically placed across the midline. (c) Major inputs and outputs of the SC layers.

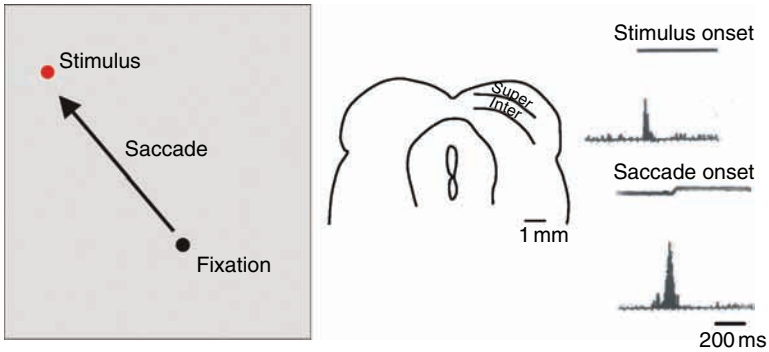

Figure 2 Visual activity in the SC superficial layers and movement-related activity in deeper layers. Histograms of activity on the right show the visual response (top) aligned on the onset of a visual stimulus and the saccade-related activity aligned on the saccade onset. Cartoon on the left shows the visual stimulus and saccade to it that led to this activity. The current fixation is indicated by a crosshair, the onset of the stimulus by the black spot, and the ensuing saccade to it by the arrow.

of these deep layers in relation to movement, and consequently the four intermediate and deep layers are frequently grouped together as the deep layers. We will concentrate primarily on the intermediate layers since this is the layer where the sensory to motor transition is most prominent. The samples of neuronal activity in Figure 2 indicate why the superficial layers are usually referred to as the visual layers and the deeper layers are referred to as the movement layers.

The final and probably most critical point regarding the SC structure is the presence of a retinotopic map of the visual field spread out on its surface and extending down in register through the lower layers. As shown in Figure 3, the contralateral visual hemifield is represented in the SC on one side of the brain, as is the case in primary visual cortex. This map is a nonlinear one with the central visual field represented in more detail than the peripheral field; hence, the map is compressed toward the peripheral edge of the visual field. Visual stimuli close to the center of the visual field are represented rostrally in the SC (black circle in Figure 3) and targets in the periphery are represented caudally (black triangle).

Each neuron in the superficial layers of the SC has a visual receptive field restricted to one part of the visual field, with the response of the neurons strongest in the field center, and with that center located on the map as predicted by the overall visual field map shown in Figure 3. The neurons in the intermediate layers of the SC also have receptive fields that map onto the same map as the neurons in the superficial layers. But in addition, these neurons become active before saccades made to just one region of the visual field, their movement field. There is a gradient of activity within each movement field just as in the visual receptive field (Figure 4), and a neuron's maximum discharge is associated with saccades to the

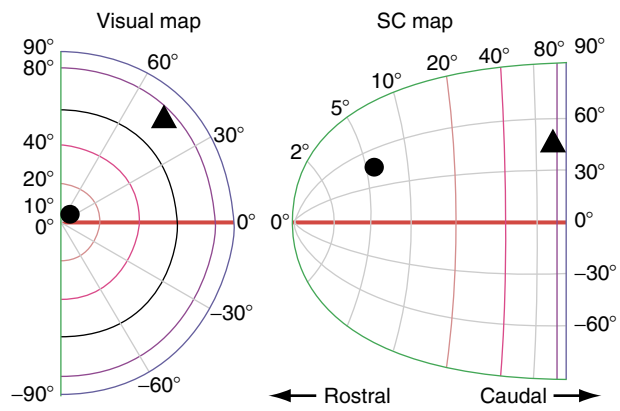

Figure 3 Projection of the visual field onto the SC. The projection of the right visual is to the left SC. The color code indicates the eccentricity. Courtesy of Nakahara, Morita, Wurtz, Optican (2006).

center of this field which conveys information about the specific amplitude and direction of an impending saccade. A neuron's optimal saccadic vector is usually closely aligned with the center of its visual receptive field. Different neurons have different vectors and the neurons are organized in a highly regular fashion to produce a neural map of saccadic vectors covering the contralateral visual field. On this map, large saccades are represented in the caudal SC, small are in the rostral SC, upward are more medial, and downward are more lateral, just as was the case for the visual map.

A key point of this map is that it is registering differences. For the visual responses, neurons with receptive fields 10° of arc away from the center of gaze (the center of the fovea), will indicate with their maximal response that the target is 10° away in a particular direction. Similarly for a saccade target at the same location, the activity of the saccade-related neurons with maximal activity for this point indicates that the change in the eye position to bring the eye to

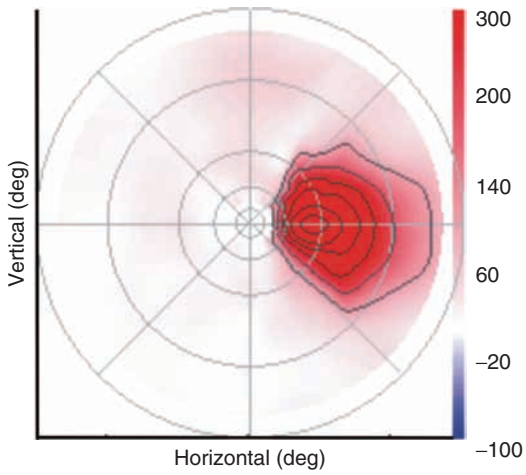

Figure 4 Two-dimensional plot of a SC movement field showing the gradient of activity of neurons at one SC recording site. The color scale indicates the activity in spikes per second, but because the color indicator saturates, the contour lines indicate the higher frequencies in steps of 100 spikes/second with a small center frequency of 600 spikes/second. The map extends out horizontally and vertically to 30° of visual angle.

the target is 10° in that direction. Both indicate the distance from the fovea of the retina and are thus both retinotopic maps.

A second key point is that activity on this map produces eye movements not by the action of a few localized neurons but by a large population of neurons centered on one region of the map. If we consider neurons along one meridian, one estimate is that during fixation approximately 15% of the neurons (those centered on the rostral pole) are active and that during a saccade about 25% of the burst neurons and 40% of the buildup neurons are active.

Neuronal Activity for Saccades and Fixation

Activity leading to the generation of a saccadic eye movement is centered in the intermediate layers because it is here that neurons are active before the generation of saccades, where electrical stimulation generates them, and where inactivation leads to their disruption. The same logic holds for the control of the fixation between these movements. Not all intermediate layer neurons, however, are the same, and it is useful to consider these neurons as falling into three groups even though they almost certainly form a continuous distribution of neuronal types.

Burst neurons are the saccade-related neurons that have been most extensively studied in the SC, and are the ones that first flagged the SC as related to the generation of saccades. Their defining feature is a brief burst of firing time locked to saccades made into their movement fields (Figure 5, right). The burst neurons are found throughout the rostral to caudal extent of the collicular map, and their activity indicates an impending saccade with a vector coded by the position of that neuron on the SC map. They also frequently have a visual response time locked to the onset of the visual target.

Buildup neurons are similar to burst neurons in having visual responses and usually increased activity just before the saccade, but their distinctive characteristic is an increase of activity during the period between the visual response and the saccadic burst (the delay period – Figure 5, middle). Because of this delay in activity, they have also been referred to as long lead neurons or prelude bursters, and because the activity tends to increase as the time of the saccade approaches, they are referred to as buildup neurons. Like the burst neurons, the buildup neurons are found throughout the collicular map and represent the amplitude and direction of impending saccades related to that region of the map. The working hypothesis is that their buildup or delay activity represents the preparation to make a saccade, and their activation may facilitate saccade generation.

The neurons usually referred to as fixation neurons lie at the rostral pole of the SC. They are tonically active when the monkey fixates its eyes on a visual target (Figure 5, left) They are not dependent on the presence of the target, however, because the firing persists when the target is briefly removed, as long as fixation persists. There is a decrease in the discharge rate when a saccade occurs. All these characteristics led to referring to these rostral neurons as fixation neurons although further investigation suggests that they are a rostral extension of the buildup neurons (see below).

Saccade Generation

The elements in the SC show activity in an orderly sequence before the onset of saccadic eye movements. While the monkey is looking at a point and not moving its eyes, the fixation neurons are tonically active. During this period of fixation, when a target appears in the visual receptive field of a burst or buildup neuron, there is brief visual activity followed by a gradual increase in activity in buildup neurons that is taken to indicate preparation to make a saccade to the target. At some point the activity becomes strong enough to produce the high-frequency burst in the buildup and the burst neurons, and their output provides a strong drive on the saccade-related neurons in the lower midbrain and pons. These brain stem neurons receive what is essentially a retinotopic signal, and over a series of steps they provide a signal

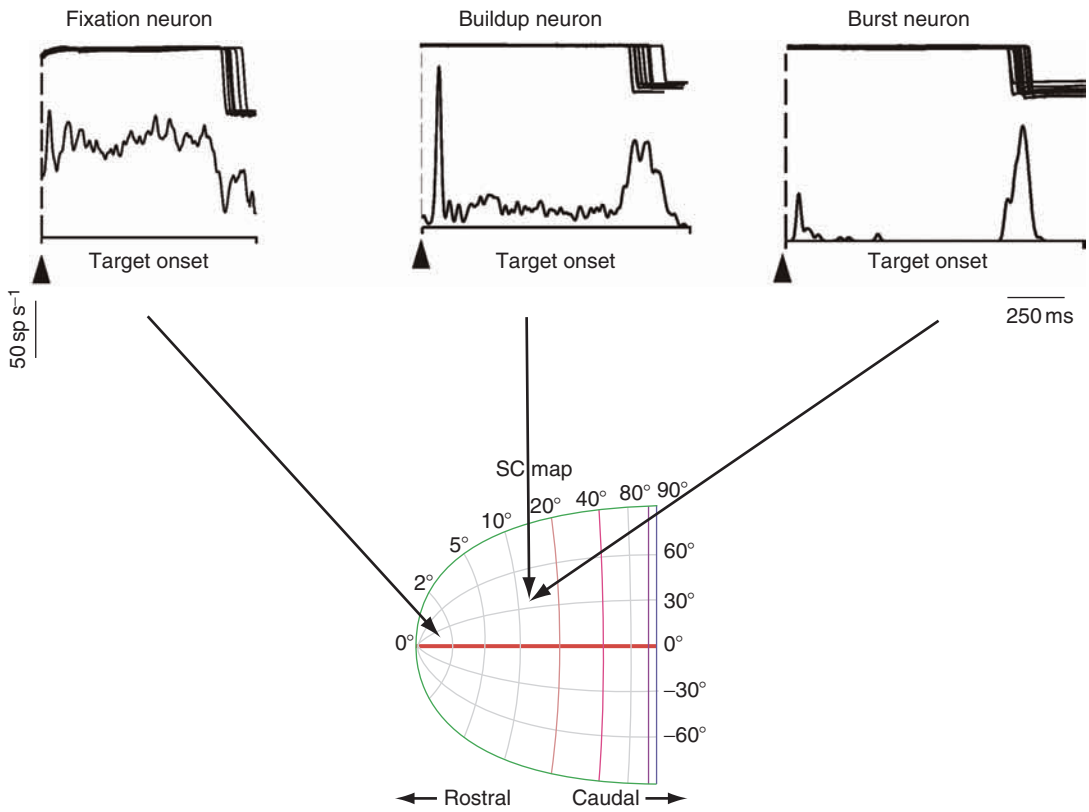

Figure 5 Examples of histograms for burst, buildup, and rostral neurons all aligned on the onset of the target (start of each record) for the saccade (indicated by the deflection of the eye position traces above the histogram). The location of each neuron is indicated on the SC map with the fixation neurons being at the rostral pole and the burst and buildup neurons being more caudal in the SC.

appropriate for the pulling direction of the six eye muscles so that the eye moves to direct the fovea to the target that has led to the activation of the SC neurons. The output of the SC also acts to move the head to the target as well; most shifts to a new target involve the coordinated movement of both eyes and head and is referred to as a shift in gaze.

The SC is the last step where the active neurons on a retinotopic map indicate the amplitude and direction of the impending movement. In contrast, neurons below the SC indicate which muscles should contract and how much and for how long they should do so. Thus the signal to move the eyes is transformed from a spatial code on a map where the vigor of the activity is secondary to their place on the map to one in which there is no longer any map and it is a frequency code where the frequency and duration of the activity determines the eye movement.

What leads to the activity within the SC? From the frontal eye field region of the frontal cortex (refer to Figure 1(b)) the direct projection centered on the intermediate layers provides an excitatory drive to the SC. The frontal cortex provides a similar drive through the basal ganglia to the SC, but the output via this route through the substantia nigra pars reticulata is inhibitory. The substantia nigra input to the SC has a tonic high discharge rate and a pause at the time of the saccade just the opposite of that from the frontal eye field which has a relatively low discharge rate and a burst of activity. Thus these two inputs act as a push–pull control on the saccade-related activity of the SC. These two inputs exert a major effect on the SC. The input from parietal cortex, particularly lateral intraparietal area (LIP), is also a major excitatory input to the SC. Thus the activity in the SC related to saccades is shaped by these excitatory and inhibitory inputs.

Fixation

While it is critical for saccades to move the eye to different parts of the visual field to examine areas of interest, most of our useful vision occurs in the periods of fixation between the saccades. There is a continual interaction between periods of fixation for viewing one scene and saccades to shift from one view to another. This alteration of fixation and saccades forms the very root of our visual perception. It should

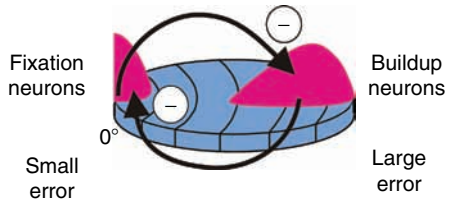

Figure 6 Schematic of the interaction between neurons at the rostral pole of the SC and buildup neurons in the caudal SC. Activity of each population on the other is inhibitory.

not be surprising that there are neuronal mechanisms that insure visual stability during these periods, the SC neurons that are critical are the neurons in the rostral SC, those referred to as fixation neurons.

The action of these rostral neurons has been demonstrated by showing that saccades are suppressed by their activation by either electrical stimulation or injection of bicuculine, a γ-aminobutyric acid (GABA) antagonist, that frees these rostral neurons from tonic inhibition. In contrast, saccades are facilitated by the inactivation of the rostral neurons by injection of the GABA agonist muscimol. Both activation and inactivation experiments show that the rostral SC alters when saccades are generated. Figure 6 shows schematically the interaction between fixation and saccadic mechanisms in the SC. The interpretation is that when the rostral neurons are active they inhibit all other SC neurons including the burst and buildup neurons in the caudal SC, and when these later neurons are active during a saccade they inhibit the rostral SC neurons.

Simplifying the SC Neuron Categorization

A possible continuity between the rostral neurons (the fixation neurons) and buildup neurons substantially simplifies the categorization of the SC neurons. Both neuron types have delay activity between the time of target presentation and the saccade; for the buildup neurons this is the buildup before the saccade and for the rostral neurons it is the continuing activity during fixation. Both types also have movement fields, that is, they have a burst of activity just before a saccade to the target as described for the buildup neurons but which applies equally well for the rostral neurons which show this burst before very small saccades (consistent with their position in the rostral pole of the SC). More importantly, the increase in activity of the rostral neurons occurs whether a saccade is made or not. This makes the rostral neurons just like the buildup neurons because they show increased activity long before saccade onset even if the saccade is not actually made. But the key point is that both these rostral neurons and the buildup neurons indicate an error between where the eye is now and where it should be. For the buildup neurons the error can be large depending on their location on the SC map, and the buildup or delay activity indicates this error (Figure 6). The same is true of the rostral neurons, but their activity indicates a small error, frequently only a fraction of a degree and too small to make a saccade. The working hypothesis is that the activity of these rostral and buildup neurons indicates an error between where the eye is and where the target is, and the size of the error represented by the neuron depends upon the location of the neuron on the SC movement map. This does not change the functional interaction between rostral and buildup neurons described above, nor the role of the rostral neurons in maintaining fixation, but instead provides a mechanism for the interactions. The hypothesis also raises the possibility that this error signal can be used for movements other than saccades, particularly for the generation of pursuit movements.

SC Determines Movement Goal, Not Metrics

The organization of the SC, its relationship to other structures, and the sequence of activity preceding saccadic eye movements form the core of our knowledge of the SC, for which there is little dispute. More recent experiments, however, have shown that the SC is as closely related to selecting the goal of a saccade as it is to determining the metrics of the impending saccade.

The first point is that even the earliest experiments on the primate SC showed that, although short-term SC inactivation largely eliminates saccade generation, ablation limited to the SC does not. This only serves to emphasize that while we are considering the SC, it is the brain that generates the saccade, and to develop a viable model of saccades other areas such as the cerebellum must be included. In addition, while the SC neurons provide the amplitude and direction information required to bring the fovea onto a new target of interest, knowing this does not allow prediction of the metrics of the saccade actually made. For example, by using an adaptation procedure, monkeys can be made to make saccades that are shorter than that required to get the eye to the visual target, but the peak activity in the SC remains centered on the region representing the target not the saccade actually made.

In contrast, SC activity is modulated by shifting the goal of a saccade. The importance of the goal first became apparent when it was recognized that the visual responses of SC superficial layer neurons were enhanced when the monkey attended to the stimulus

that was to be the goal of the impending saccade. More recent experiments have shown that when monkeys are given two visual targets in rapid succession, the monkey can shift its goal from one target to the other and make a strongly curved saccade from one target to the other. The SC activity is consistent with the successive selection of the two targets rather than the shape of the saccade executed. In addition, there are substantial modulations of the SC activity related to the selection of the target itself. If the probability of one of several visual stimuli becoming the target for the next saccade is varied, the activity in the SC changes long before the burst of activity related to a specific saccade. Therefore, the SC is providing an error that indicates the goal of the saccade, and via the fixation neurons, when the saccade will occur, but it is only part of the system determining the metrics of that movement.

Corollary of Impending Saccades

Everything considered so far summarizes the extensive information on the SC contribution to the generation of movement. But it has long been known that the SC not only has downward projections to the regions of the oculomotor nuclei for eye movement and the spinal cord for head movement but that it has projections that ascend upward through the thalamus to the cerebral cortex. Recent experiments have begun to indicate the signals carried by these pathways and the possible functions for them.

One prominent pathway extends from the SC (primarily from the intermediate layers) to the medial dorsal nucleus of the thalamus and from there to the frontal cortex including the frontal eye fields (Figure 7). The specific medial dorsal neurons that receive information from SC and relay it to the frontal eye field (FEF) have been identified by using the classic physiological techniques of antidromic and orthodromic stimulation, and they have been found to convey the visual and presaccadic activity described for the SC. One obvious possibility is that these neurons are providing a signal for the guidance of saccades, but this has been rejected because inactivation of the medial dorsal neurons does not significantly alter saccades. The other salient possibility is that the activity is conveying information for keeping track of the movement rather than generating it, that is, it is carrying a corollary discharge or efference copy of the intended movement. This turns out to be the case; inactivation of the medial dorsal neurons impairs saccadic accuracy in a task (the double step task) that requires the internal knowledge of the preceding saccade that would be provided by a corollary discharge. This sparing of guidance and disruption of registration is clearly what would be expected if this path from SC up to frontal cortex carried a corollary discharge and provides strong evidence that the SC provides information back up to the cerebral cortex, as well as receiving information from the cortex.

A corollary discharge is not the only type of information that the SC provides to cerebral cortex. Recent experiments have shown that visual spatial attention can be produced by stimulating oculomotor regions of the brain including the SC. Electrical stimulation of the SC improved performance on a visual motion task in ways consistent with a covert shift of attention. Because visual motion perception is largely a result of cortical processing, its modulation implies that the action of the SC is on cortex. Whether this effect reaches cortex via the pathway connecting SC and the frontal eye field or via the pathway from the SC through the thalamic pulvinar nucleus and then directly to motion areas in cortex remains unknown.

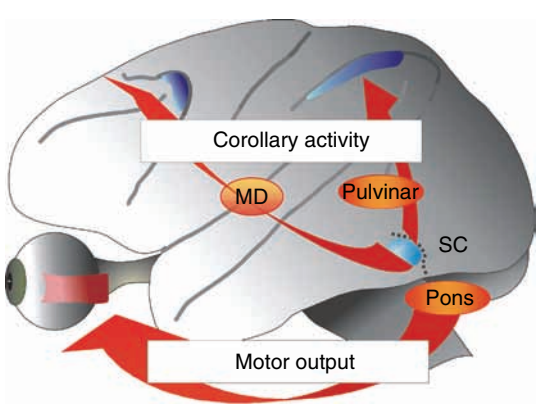

Figure 7 Comparison of the two major SC outputs, one for the generation of saccadic eye movements and the other to convey a corollary of these movements to cerebral cortex.

Conclusion

The SC is one of those structures in the brain whose function has shifted over evolution. In the primate, it still has activity related to the reflex movements made to the sudden appearance of a target in the visual field. But it is also a key way station in the oculomotor circuits that descend from the visual and premotor areas of the cerebral cortex to the movement centers in the brain stem. It is the last node where instructions for movement are represented as an error on a spatial retinotopic map before being broken up into instructions related to muscles. It provides information back to the cortex on the movement to be executed, the long sought corollary discharge. While it is embedded in the larger visual and oculomotor control system, it

has provided a superb model to study sensory–motor integration for both the control and the registration of movement.

See also: Attention and Eye Movements; Basal Ganglia and Oculomotor Control; Brainstem Control of Eye Movements; Eye and Head Movements; Frontal Eye Fields; Multisensory Convergence and Integration; Oculomotor Control: Anatomical Pathways; Oculomotor System: Models; Parietal Cortex and Spatial Attention; Saccade–Pursuit Interactions; Saccades and Visual Search; Saccadic Eye Movements; Supplementary Eye Fields; Target Selection for Pursuit and Saccades.

Further Reading

Isa TKY and Saito Y (2004) Dynamic modulation of signal transmission through local circuits. In: Hall WC and Moschovakis A (eds.) *The Superior Colliculus*, pp. 159–172. New York: CRC Press.

Keller ER (2004) Distributed neural processing in the saccadic system. In: Hall WC and Moschovakis A (eds.) *The Superior Colliculus,* vol. 21, pp. 277–301. New York: CRC Press.

Munoz DP and Fecteau JH (2002) Vying for dominance: Dynamic interactions control visual fixation and saccadic initiation in the superior colliculus. *Progress in Brain Research* 140: 3–19.

Ozen G, Helms M, and Hall WC (2004) The intracollicular neuronal network. In: Hall WC and Moschovakis A (eds.) *The Superior Colliculus,* vol. 21, pp. 147–158. New York: CRC Press.

Quaia C, Lefevre P, and Optican LM (1999) Model of the control of saccades by superior colliculus and cerebellum. *Journal of Neurophysiology* 82: 999–1018.

Robinson DL and McClurkin JW (1989) The neurobiology of saccadic eye movements The visual superior colliculus and pulvinar. *Reviews of Oculomotor Research* 3: 337–360.

Sommer MA and Wurtz RH (2003) The dialogue between cerebral cortex and superior colliculus: Implications for saccadic target selection and corollary discharge. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 1466–1484. Cambridge: MIT Press.

Sparks DL (2004) Commands for coordinated eye and head movements. In: Hall WC and Moschovakis A (eds.) *The Superior Colliculus*, pp. 303–318. New York: CRC Press.

Sparks DL and Hartwich-Young R (1989) The deep layers of the superior colliculus. In: Goldberg ME (ed.) *The Neurobiology of Saccadic Eye Movements, Reviews of Oculomotor Research,* vol. III, pp. 213–256. Amsterdam: Elsevier.

Wurtz RH, Basso MA, Paré M, and Sommer MA (2000) The superior colliculus and the cognitive control of movement. In: Gazzaniga MS (ed.) *The New Cognitive Neurosciences,* 2nd edn., pp. 573–587. Cambridge, MA: MIT Press.

Wurtz RH, Sommer MA, and Cavanaugh J (2005) Drivers from the deep: The contribution of collicular input to thalamocortical processing. *Progress in Brain Research* 149: 207–225.

Supplementary Eye Fields

M A Sommer, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

The supplementary eye fields (SEFs) were discovered in 1985 by John Schlag and Madeleine Schlag-Rey, who were informed by prior findings spanning the preceding century. In 1885, David Ferrier demonstrated that frontal cortex near the midline in monkeys is electrically excitable. By stimulating there, Ferrier evoked movements of the head and eyes toward contralateral space. Homologous sites in the human brain were reported by the neurosurgeons Foerster and Penfield. A breakthrough was provided by Woolsey, who carefully mapped the supplementary motor area in many animals and showed that it had a regular topography that included rostral eye representation. Brinkman and Porter qualitatively explored this zone in 1979. Unlike all of the prior studies, Schlag and Schlag-Rey explored the region in a highly quantitative manner using accurate and precise eye movement measurements. After their seminal work, studies on the function of SEF went in surprising directions.

Early work on the SEF first tested the hypothesis that it is important for generating eye movements, particularly the fast ones known as saccades. This hypothesis quickly fell by the wayside. In its place, other intriguing ideas about SEF function have been promoted and supported. It now appears that the SEF is unnecessary for generating eye movements but is important for higher-order control of eye movement behavior. The SEF might be likened to the foreman of a construction crew for assembling behavior; the SEF does not do the work but guides it.

The unveiling of the SEF as a distinct region in 1985 had special impact due to another landmark set of results published the same year by Bruce and Goldberg that formally delineated the nearby frontal eye field (FEF). Since then, a useful tactic in the field has been to analyze the SEF in relation to its historical twin, the FEF. The present article follows this comparative approach.

Anatomy

The SEF is located in dorsomedial frontal cortex (Figure 1(a)) and represents the ocular extension of the body map in the supplementary motor area (Figure 1(b)). The exact boundaries of the SEF have been a subject of debate. The functionally defined SEF corresponds roughly to a zone (F7) delineated by Matelli, Luppino, and Rizzolatti using cytoarchitectonics, but associating functional regions with cytoarchitectonic zones is not always straightforward. The FEF, for example, straddles two distinct cytoarchitectonic zones (Walker's 8A and 45) even though its functions seem continuous across them. It is generally accepted that eye-movement-related neurons can be found throughout a large swath of peri-midline cortex, leading some laboratories to select this entire region as their focus of study (using the term dorsomedial frontal cortex rather than SEF). Other laboratories are guided by the principle of studying the zone where eye movements can be elicited with least current threshold using trains of electrical current pulses. In the FEF, a criterion of <50 μA was established, and in an effort to make a fair comparison between SEF and FEF the same criterion was transferred to SEF. The location of SEF in humans is homologous to that in monkeys – in dorsomedial cortex at the rostral end of the supplementary motor area. In terms of connections, the SEF is very similar to the FEF. Both receive projections from, and send efferents to, a plethora of visual- and saccade-related areas with minor differences, such as slightly different patterns of thalamic connections. Functionally, however, the SEF seems more distantly connected than the FEF to the final saccade-generating circuits because the ability to evoke saccades from the SEF requires, in general, relatively higher currents, longer train durations, lower frequencies, and more care in minimizing behavioral obstacles such as intense fixation.

Function

Generating Saccades and Fixations

Schlag and Schlag-Rey proposed that the SEF was a distinct eye field because stimulating it elicited saccades and recording from it revealed vision- and saccade-related activity. Follow-up studies established that it also contains many fixation-related cells. Hence, the first hypothesis about SEF was a simple one: perhaps it helps to generate saccades and fixations. At the outset, however, this hypothesis seemed doubtful. A lesion study by Schiller had shown already that when other major oculomotor structures were ablated (the FEF and superior colliculi), animals were permanently devastated in generating saccades even though the SEF remained intact. It may have been that the SEF was indirectly damaged in that study, or that the SEF is important but needs the FEF or superior colliculus to transfer its signals downstream. Later experiments seemed to refute these possibilities, however. Stimulation in the

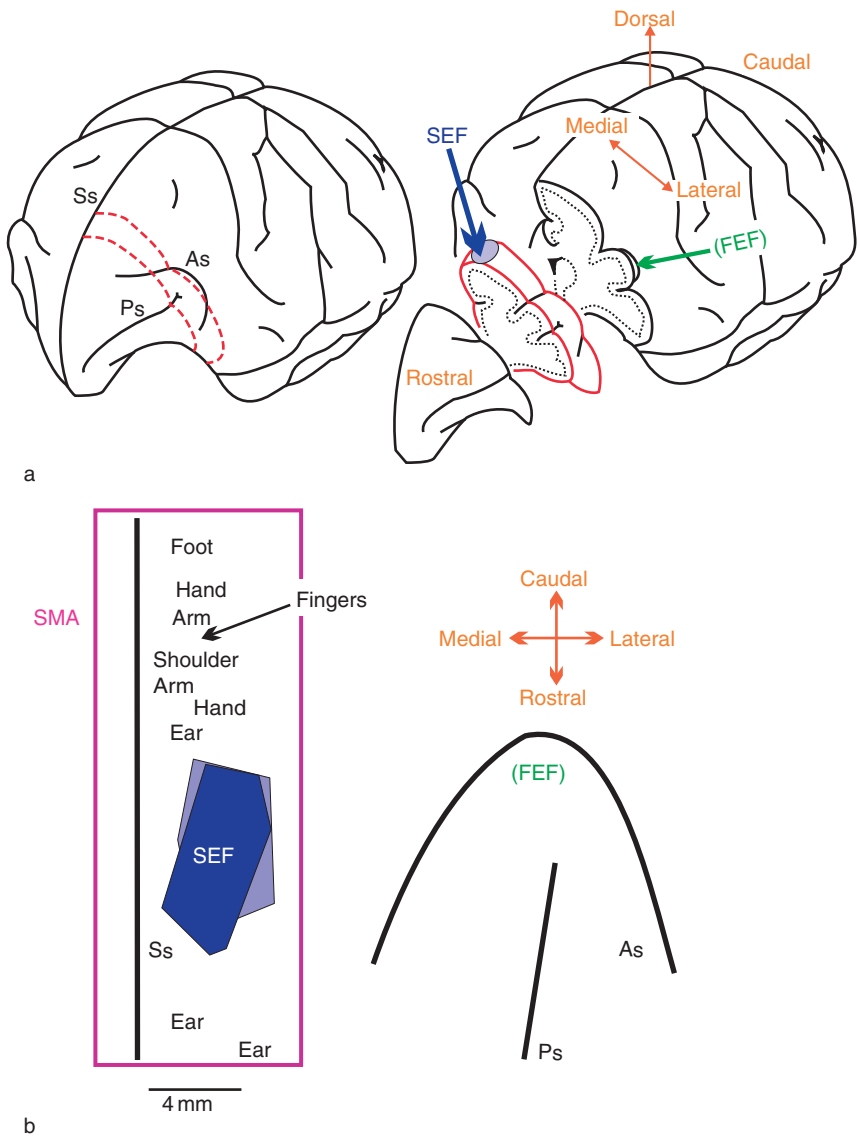

Figure 1 Location of the SEF. (a) Frontal, lateral perspective of monkey brain with general location of SEF and FEF shown in cutaway view at right. As, arcuate sulcus; Ps, principal sulcus; Ss, sagittal sinus. Anatomical directions are shown in orange. (b) Range of SEF recording and stimulation sites (blue) as reported by Schlag and Schlag-Rey. Data from two hemispheres (the two polygons) are superimposed. The SEF lies in the rostral supplementary motor area (SMA), in the head part of the body map. SEF and SMA are functional terms; the anatomical term for the region is the dorsomedial frontal cortex. Location of the FEF is shown for reference to the right. (a) Adapted with kind permission of Springer Science and Business Media from *Experimental Brain Research* 58, 1985, pp. 208–211, 'Unit activity related to spontaneous saccades in frontal dorsomedial cortex of monkey,' Schlag and Schlag-Rey, fig. 1, © Springer-Verlag 1985. (b) Modified and used with permission from *Journal of Neurophysiology* (Schlag J and Schlag-Rey M (1987) Evidence for a supplementary eye field. *Journal of Neurophysiology* 57: 179–200).

SEF evokes saccades after lesions of FEF or superior colliculus, implying that the SEF independently accesses saccadic-generating circuits. More to the point, reversible inactivation of the SEF has no discernible effect on saccades and fixations (even though inactivation of the FEF or superior colliculus causes enormous deficits). A study of SEF lesions by the Schiller group found deficits only in an extremely sensitive test of visuosaccadic behavior. The impairments were weak and short-lasting. It is safe to say that the SEF has little or no influence on the basic generation of saccades and fixations.

Stimulating the SEF also elicits head movements and smooth eye movements. No studies have silenced the SEF to test whether it is needed for these actions.

Providing Frames of Reference

If the SEF is inconsequential for fundamental oculomotor behaviors, then why is it so packed with visual- and eye-related neurons? Why does electrical

stimulation of it evoke saccades and fixations? These have been the salient questions driving SEF research in the past decade. The mystery is whether the SEF does something important even if it is unconcerned with the details of making individual movements.

From the first studies of Schlag and Schlag-Rey it was recognized that the SEF might play a special role in providing frames of reference. Electrical stimulation of the FEF, superior colliculus, and most other eye movement areas evoke vectors of saccades (i.e., constant displacements regardless of initial eye position). Schlag and Schlag-Rey showed that stimulation of the SEF, in contrast, may evoke goal-directed saccades. Such movements travel to a 'termination zone' in space that stays constant regardless of where the eyes initially fixate (**Figure 2(a)**). A debate ensued between Tehovnik and colleagues, who confirmed the goal-directed findings and extended them by describing a topographic map of termination zones across the SEF and surrounding dorsomedial frontal cortex, and Bruce and colleagues, who explained the appearance of goal-directedness as natural variation, whereas vector saccades were the norm (**Figure 2(b)**). The topographic stimulation map of Tehovnik was matched by a topographic map of tuning in SEF fixation-related neurons; similarly, the vector coding of Bruce was matched by the vector coding of SEF saccade-related neurons. It is possible that Tehovnik preferentially activated fixation neurons and Bruce saccade neurons, but this is speculative. On a related note, two laboratories (the Crawford and Schlag groups) showed that some SEF sites produce vector saccades and other SEF sites goal-directed saccades when stimulated, and that vector-to-goal-directed representations in SEF may be organized in a rostral–caudal gradient. The debate about coordinate frames in the SEF has spilled over into a nearly identical controversy regarding whether goal-directed or more limited movements are encoded in primary motor cortex.

A goal-directed or craniotopic (referenced to the head) representation is not the only type of nonstandard coordinate frame encoded by the SEF. Single neuron recording studies by Olson and colleagues found that many SEF neurons use an object-centered frame of reference (**Figure 2(c)**). Such neurons fire preferentially for saccades made to a particular part of an image (e.g., the left end of a small bar) regardless of where that image is presented. As argued by Olson, object-directed frames of reference are especially critical for us in everyday life as we survey, handle, and inspect the items that surround us. It is unknown to what extent object-directed frames of reference are unique to the SEF.

The bottom line is that there is broad support that the SEF plays a special role in representing frames of reference that are not directly extractable from the retinal image but which require further information: the location of eye in head for craniotopic frames of reference and the segregation of image clutter into distinct entities for object-centered frames of reference.

Supervising Behavior

The paradox of the SEF – that it seems so saccade related but is so blatantly unnecessary for generating saccades – and the early findings on multiple coordinate systems in the SEF inspired a new generation of experiments that have been very informative. This nascent research posits a higher level function for the SEF. The premise is that the SEF is well poised for controlling saccadic behavior due to its variety of neuronal signals as revealed by recording, its tight linkage to other saccadic centers as revealed by stimulation, and its fluency in multiple ways of spatially encoding visual stimuli and eye movements. The first fundamental question about an executive role for SEF logically would be the following: Can the SEF turn on and off saccades, even if it cannot generate them by itself?

Schall and colleagues studied this question using a task in which a monkey is instructed to make a saccade but then, randomly, may be instructed to cancel it. Nimble supervision of the saccadic system is necessary in this 'countermanding' task, and SEF neurons carry signals that may reflect this supervision. When errors of commission are made (saccades are made when they should have been canceled), many SEF neurons burst vigorously just after the movement (**Figure 3(a)**). Neurons in FEF, in contrast, seem to have a more proactive function: their activity is well suited to cause saccade generation or cancellation but is poor at representing task outcome.

Whether the error-related activity in SEF influences behavior cannot be answered by neuronal recordings and requires, instead, causal experiments. Schall and Stuphorn performed such a study, electrically stimulating the SEF in the countermanding task. In support of the role of SEF in supervising behavior, activating it with low currents lengthened saccadic reaction times in the countermanding task so that monkeys had more time to cancel their movements (**Figure 3(b)**). Outside of the countermanding task, however, in a visually guided saccade task, reaction times were made shorter instead of longer. This combination of neuronal recording and stimulation provides complementary evidence for a supervisory role of SEF in controlling whether a saccade is made.

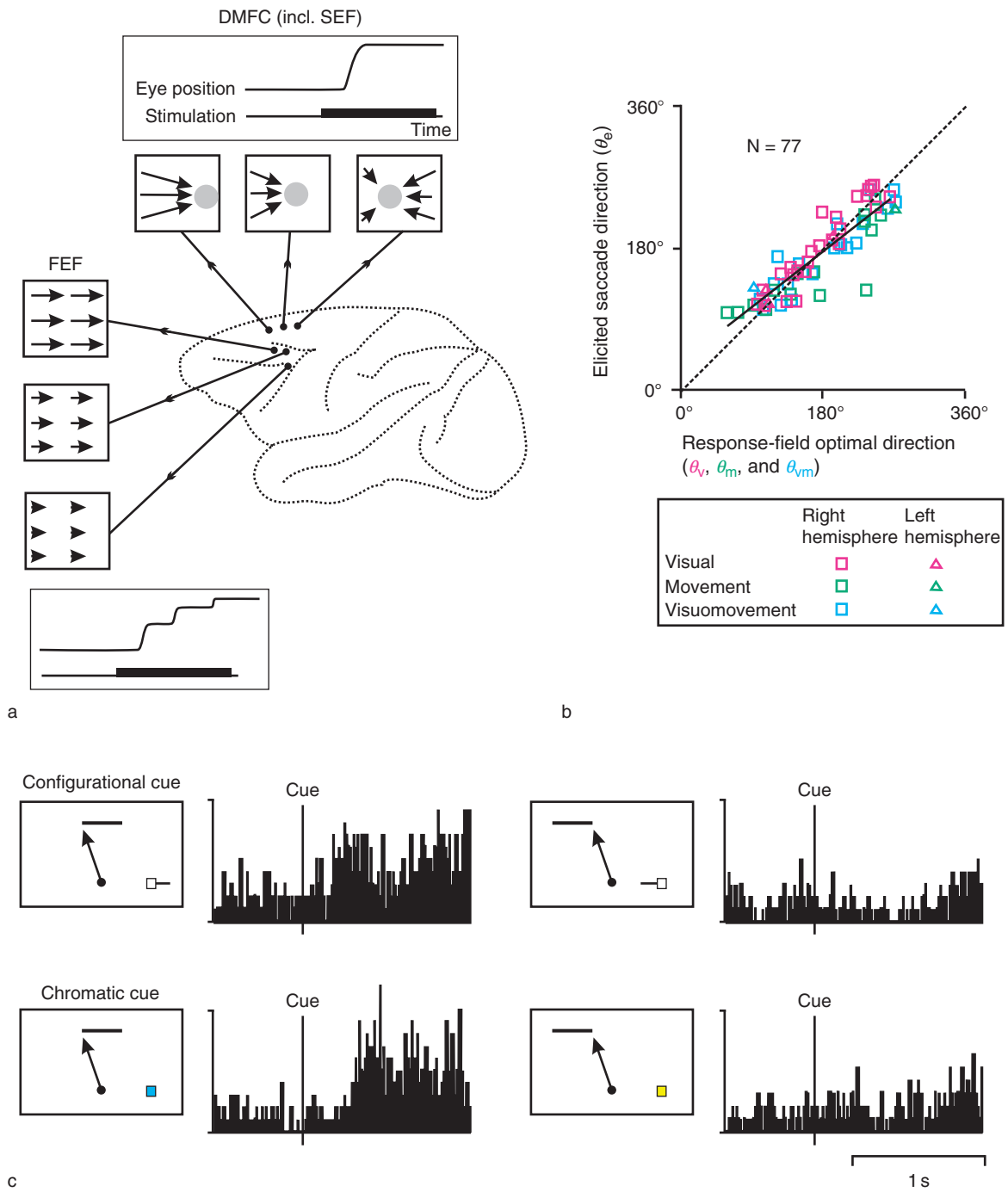

Figure 2 Reference frames in the SEF. (a) Head-centered coordinate frame. Saccades (arrows) evoked by stimulation of the dorsomedial frontal cortex (DMFC) including the SEF (top) compared with saccades evoked from the FEF (left). In the DMFC, stimulation-evoked saccades go to a termination zone (gray circle) as a function of stimulated site. Long durations of stimulation cause a single saccade (inset). In FEF, stimulation-evoked saccades exhibit a constant displacement regardless of initial eye position. Long durations of stimulation cause similar saccades to occur repeatedly. (b) Eye-centered coordinate frame. In the core of the SEF, brief train durations evoke directions of saccades (*y*-axis) that match the directional tuning (*x*-axis) of visual, movement, or visuomovement neurons found at the stimulated site. Little evidence for termination zones was found in this experiment. (c) Object-centered coordinate frame. Activity of an SEF neuron is strong for saccades made to the left side of a bar (left) but not the right side (right), regardless of the instruction that cues the monkey (square on bar, top; color, bottom). (a) Adapted and reprinted from *Behavioral Research Reviews*, vol. 32, Tehovnik et al., 'Eye fields in the frontal lobes of primates,' pp. 413–448, © 2000, with permission from Elsevier. (b) Modified and used with permission from *Journal of Neurophysiology* (Russo & Bruce, 84: 2605–2621, 2000). (c) Modified and used with permission from *Journal of Neurophysiology* (Olson & Gettner, 81: 2340–2346, 1999).

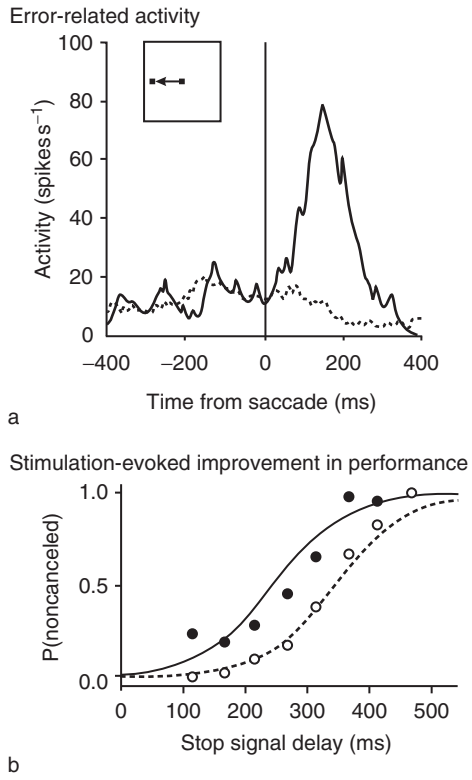

Figure 3 Supervisory function of SEF. (a) Average activity of an SEF neuron is shown for saccades made to the left (into ipsilateral space; inset) when the movement was appropriate (dotted line) and when it was errant because it should have been withheld (solid line). The SEF neuron fired vigorously after the error. The same effect occurred for rightward saccades. (b) Improvement in performance when stimulating the SEF (white circles) relative to not stimulating (black circles). Stimulation allowed the monkey to wait even longer – that is, nearer to the last possible moment (rightward shift to longer stop signal delays) before canceling the movement. (a) Adapted by permission from Macmillan Publishers Ltd: *Nature* 408: 857–860, © 2000. (b) Adapted by permission from Macmillan Publishers Ltd: *Nature Neuroscience* 9: 925–931, © 2006.

Signaling Conflict

An effective foreman not only gives instructions but also makes wise judgments in the face of conflicting information. So it is with the saccadic system. The visual world provides little information about where to look next, and the signals that are available can provide mixed messages. A traffic light 'walk' signal is good to notice for an instant, but its continual illumination (and sudden change to a flashing 'don't walk' signal) provides a conflict for behavior by demanding your attention when you should, instead, maintain a vigilant watch to the left and right. SEF neurons seem to play a special role in dealing with visuosaccadic conflicts so that one looks in the correct direction, not just the attention-grabbing direction.

One form of visuosaccadic conflict was described previously: in the countermanding task, there is a conflict between the sudden need to fixate and the ongoing plan to make a saccade. Indeed, Schall and colleagues found that a second major cell type in SEF, in addition to the error-encoding type described previously, signaled conflict defined as the successful recognition of a stop signal when saccades were inhibited.

A particularly influential study on visuosaccadic conflict was performed by the discoverers of the SEF, Schlag and colleagues. In their task, a visual stimulus appeared in the periphery but then monkeys had to look away from it. The conflict was between the stimulus location and the required saccadic endpoint. SEF neurons had significantly higher activity in such 'antisaccade' trials than in control trials in which they were allowed to look toward the stimulus (Figure 4(a)). In addition, this antisaccade-related activity was higher in the SEF than in the FEF, providing a rare situation in which signals in the SEF are stronger than those in its more lateral neighbor. One problem in interpreting the results was that the higher activity may have been related to monitoring the conflict, to the different rules implicit in making a saccade toward a target versus away from it, or to other more subtle factors.

To tease apart the reason for elevated SEF activity in the antisaccade task, Olson and colleagues created a task in which the color of a cue, regardless of where it appeared in the visual field, informed the monkey of what vector of saccade to make (left, right, up, or down). The cue was placed randomly to the left, right, up, or down relative to initial fixation. The result was that rules were held constant (color–vector association never varied) but conflict changed from trial to trial (saccades might be made toward, away from, or at right angles to the visual stimulus). In another version of the task, the rule changed so that the monkey simply looked at the stimulus wherever it appeared. The end result was that SEF neurons fired more actively when conflict was high (saccades away from, not toward, stimuli) and rules were difficult (saccades determined by stimulus color, not location; Figure 4(b)). This explains the intense activity in antisaccades versus prosaccades, because in the former conflict is higher and the rule is harder.

Predicting Events

A common type of neuronal activity found throughout the cerebral cortex is predictive or anticipatory signals. The activity of many neurons ramps up if task context implies that an event, such as visual stimulus onset or a cue to move, is imminent. The SEF seems to

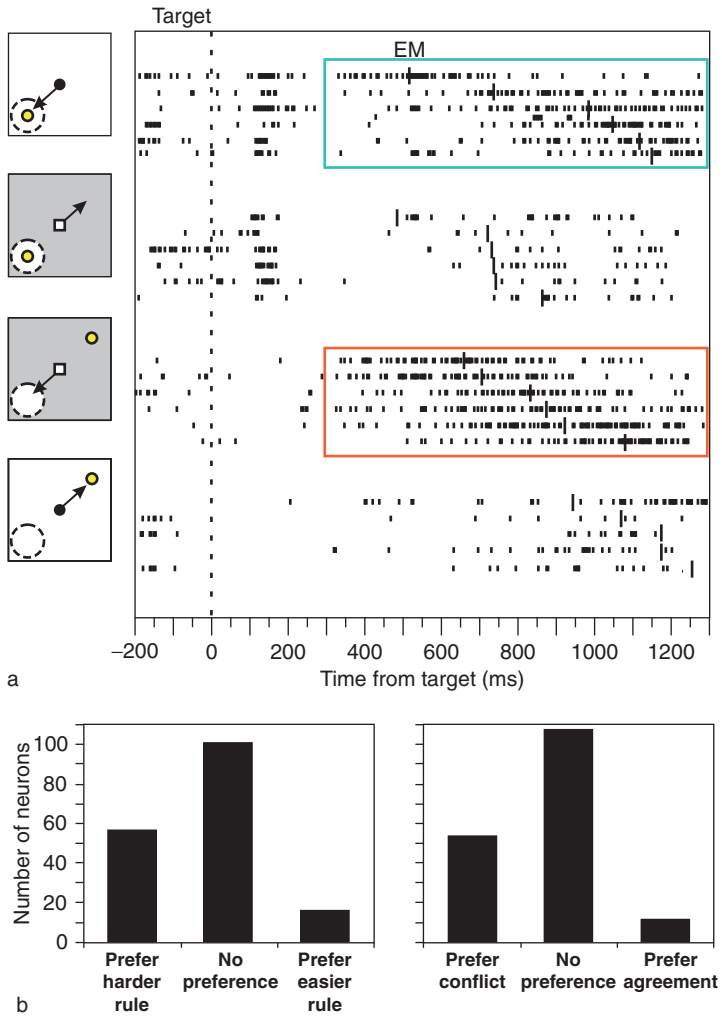

Figure 4 Conflict and rule functions of SEF. (a) SEF activity during antisaccade task. Considering only the trials in which the saccade (arrow) enters the receptive field (dotted circle), the SEF neuron fires more strongly when the saccade is an antisaccade (orange box) than when it is a prosaccade (blue box). The yellow circle shows visual stimulus. Other data show permutations of the task when saccade goes away from the receptive field. (b) The antisaccade task varies from the prosaccade task in that it involves conflict and a harder rule. To tease apart the influence of each factor, conflicts and rules can be varied independently. Both factors are important: SEF neurons fire more strongly for harder rules than easier ones (left) and more strongly for stimulus-response conflict than agreement (right). (a) Adapted by permission from Macmillan Publishers Ltd: *Nature* 390: 398–401, © 1997. (b) Adapted and reprinted from *Physiology & Behavior*, vol. 77, Olson and Gettner, 'Neuronal activity related to rule and conflict in macaque supplementary eye field,' pp. 663–670, © 2002, with permission from Elsevier.

be especially predictive in nature, exhibiting more prevalent anticipatory activity than the FEF or an oculomotor region of parietal cortex (the lateral intraparietal area), as shown in a comparative study by the Hikosaka laboratory. The predictive capacity of the SEF was studied in most detail by Heinen and colleagues in relation to smooth pursuit movements. The smooth pursuit system is exquisitely sensitive to regularities in motion and seems to promote prediction avidly. The result is to keep the fovea centered as much as possible on an object of interest rather than having it lag behind due to afferent visual delays. Heinen took advantage of this feature of the smooth pursuit system to quantify the predictive nature of the SEF. In agreement with its other higher-order functions, the SEF was found to be highly active when anticipatory pursuit occurred, and electrical stimulation in SEF promoted anticipatory pursuit. The predictive activity, like the other signals in SEF, seems to be sensitive to rules and other supervisory factors, as shown in a clever task by this group that required monkeys to judge whether a pursuit target would

cross a 'strike zone'; in this ocular baseball task, anticipatory activity was stronger for strikes than balls. The ability to modulate one's anticipation as a function of context – the point of their findings – clearly is important not only for baseball but also for many everyday tasks (e.g., driving a vehicle).

Learning

Little is known about learning at the neuronal level in the SEF. The only detailed study was performed by Chen and Wise, who demonstrated remarkably strong activity modulations as monkeys learned pairings between foveal-presented stimuli (arbitrary symbols) and direction of saccade required (right, left, up, or down) on a daily basis. For comparison, the monkeys also performed stimulus–saccade pairings that were constant from day to day. One type of neuron seemed to be a 'surprise' detector or learning selective, having activity at the start of learning that disappeared once learning was achieved (Figure 5(a)). Such neurons were not simple error detectors as described by Schall and colleagues because the modulation could occur at any time during the trial – for example, immediately after stimulus onset and not just after an incorrect response. Another type of neuron was learning-dependent, changing its activity in correlation with the learning curve as if mediated by long-term plasticity (Figure 5(b)). Such neurons could in principle store the new stimulus–response association. A striking consequence of their change in activity was that their directional tuning curves shifted during learning. This ties together with the flexibility in spatial coordinate frames represented by SEF neurons, and it raises the question of whether the SEF could represent any coordinate frame with enough training. Whereas learning-related neurons were prevalent in the SEF, they were significantly less common in the FEF.

Learning is a function of many variables, such as detection of stimuli, resolution of conflicts between looking at stimuli and places instructed by stimuli, trial-and-error attempts at finding responses that yield reward, and detection of reward or punishment. The SEF has activity related to sensory input, oculomotor response, conflict resolution, and supervisory control

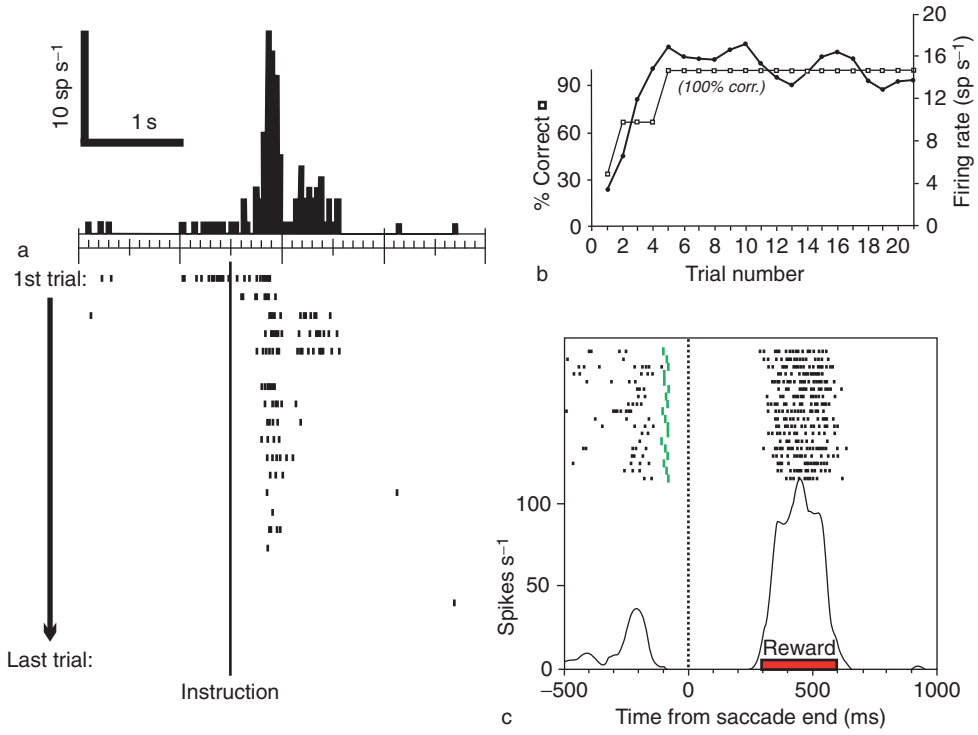

Figure 5 Components of learning in the SEF. (a) For some neurons, activity surges and then disappears as learning progresses (learning-selective neurons). (b) For others, such as the one depicted in this summary graph, activity is correlated with behavioral improvement and then remains at a new level once learning is achieved (learning-dependent neurons). (c) A necessary signal for learning is acknowledgment of reward, and many SEF neurons fire for reward as shown by this example. (a,b) Adapted and used with permission from *Journal of Neurophysiology* (Chen and Wise, 73: 1101–1121, 1995). (c) Adapted with permission from *Journal of Neurophysiology* (Amador et al., 84: 2166–2170, 2000).

of actions. What about reward? Results have been mixed. During a trial, some SEF neurons seem to anticipate reward delivery, but Olson and colleagues suggest that this is uncommon and may be related primarily to motivational state rather than to reward *per se*. After the behavior necessitated by a trial, and at the time of reward delivery, many SEF neurons do seem to be unambiguously sensitive to detection of reward, as demonstrated by the Schlag laboratory (**Figure 5(c)**). In summary, the SEF seems to have essentially all of the apparatus necessary to accomplish the learning of visuomotor associations. Still needed are causal experiments to test the necessity of SEF for learning.

Planning Sequences

The only oculomotor deficit found after reversible inactivation of the SEF and its surrounding cortex has been a mild but significant impairment in performing sequences of saccades, as shown by Sommer and Tehovnik in the Schiller laboratory. In lesion studies, Chou and Schiller demonstrated that damage to the area results in mild, short-term impairments in the ability to process and respond normally to the presentation of multiple visual stimuli. In both studies, saccades to single targets were affected barely or not at all. As with learning, the ability to react correctly when faced with multiple stimuli is the product of many underlying factors. An integrative hypothesis would be that the SEF is necessary for the correct coordination of sequences because of the underlying operations that it performs, such as supervisory control.

An alternative view is that the SEF plays a special role in overseeing the production of sequences. This is supported by further evidence. Hikosaka and colleagues showed that many SEF neurons become strikingly selective for various segments of complex saccadic sequences ('hypersets' of five sequential, learned binary choices). Similar results were found by Tanji and colleagues. Transcranial magnetic stimulation of human SEF reorders saccades of sequences as shown by Tobler and Müri, and Histed and Miller showed that electrical stimulation causes a similar effect in monkeys. In the latter study, reordering of sequences favored a final termination location in contralateral space, thus linking the effect with the coordinate frame results discussed previously.

Coordinating Limbs and Eyes

By definition, the SEF is an eye field. But how absolute is this eye specificity? The SEF represents the rostral end of a topographic map of the body that extends through the entire supplementary motor area. Edges are blurred everywhere in this map, and it is well documented that skeletomotor representation extends into the oculomotor zone and vice versa. This suggests a supramodal role for the SEF: simply stated, perhaps it is involved in hand–eye coordination. As noted by Tehovnik in a review on the topic, neurosurgeon Penfield and colleagues found that stimulating the area around human SEF caused "the impression that the contralateral hand is raised … and that meanwhile the gaze is directed by movement of head and eyes toward the hand." By comparing the gross anatomy of where eye-related neurons and arm-related neurons were found in monkey dorsomedial frontal cortex, Tehovnik concluded that overlap is extensive (**Figure 6(a)**).

Specific studies of eye- and arm-related interactions in neuronal activity were performed by Tanji and colleagues. They had monkeys look at visual targets, reach to them, or look and reach. Exploring the entire dorsomedial frontal cortex, they concluded that eye-specific activity was largely confined to SEF and arm-specific activity to more caudal supplementary motor regions, but that there was appreciable overlap (**Figure 6(b)**). Approximately 7% of neurons outside of SEF proper were eye-specific, and approximately 14% of neurons within SEF proper were limb-specific. Other neurons showed complex contingencies (e.g., firing for eye movements if limb movements were withheld but going silent if eye and limb movements co-occurred). In contrast, the FEF was overwhelmingly eye-specific. More work needs to be done to explore multimodality of the SEF and its environs, but the concept that the area plays a special role in hand–eye coordination has received preliminary support.

Conclusion

The SEF has proven to be one of the more challenging cortical areas to study. Investigators have had to resort to intricate, psychologically based tasks in order to describe SEF activity, but the results have been fruitful. In parallel with the monkey work described here, studies on humans (using metabolic imaging, electroencephalograms, and studies of lesions) have arrived at comparable conclusions regarding human SEF. The human work has been neglected for brevity here, but it is just as important. Much of the progress in understanding the SEF has arisen from systematic application of simple tasks (countermanding and antisaccade tasks) to both the FEF and SEF and comparing the results. Findings from these simple tasks then inspired more complex ones specific to interpreting the functions of the SEF. This slow and steady approach has changed the

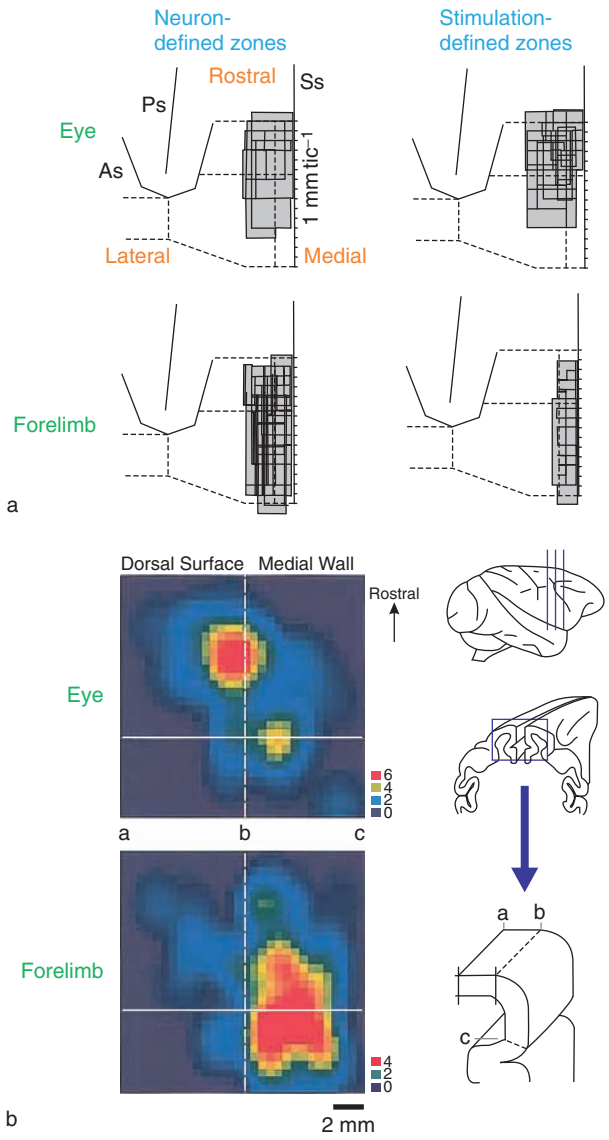

Figure 6 Overlapping limb and eye representation in and around the SEF. (a) Summary maps from a review on eye and forelimb fields in dorsomedial frontal cortex. The left column shows ranges of cortex within which eye-related or forelimb-related neurons were found in several studies. The right column shows ranges within which eye movements or forelimb movements were evoked by electrical stimulation. There is considerable overlap between the eye and forelimb fields, particularly as defined by neuronal activity. (b) Example map from one study that distinguished eye-related neurons from forelimb-related neurons. Color map (left) represents density of each type of neuron (number of neurons recorded per mm^2) as a function of cortical space. Orientation and layout of the map is explained by the anatomical diagrams at right. Although there is noticeable segregation, there is overlap as well. (a) Adapted and reprinted from *Behavioral Brain Research*, vol. 67, Tehovnik, 'The dorsomedial frontal cortex: eye and forelimb fields,' pp. 147–163, © 1995, with permission from Elsevier. (b) Adapted and used with permission from *Journal of Neurophysiology* (Fujii et al., 87: 2158–2166, 2002).

general view of the SEF from a frustrating area of study to a rather beautiful, intriguing area. The downside is that hypotheses about SEF function have sprouted wildly. A necessity in future work will be the rigorous application of causal tests, especially reversible inactivation, to establish the functions for which the SEF is truly needed.

See also: Cortical Control of Eye Movements; Frontal Eye Fields; Oculomotor Control: Anatomical Pathways; Spatial Transformations for Eye–Hand Coordination.

Further Reading

Berman RA, Colby CL, Genovese CR, et al. (1999) Cortical networks subserving pursuit and saccadic eye movements in humans: An fMRI study. *Human Brain Mapping* 8: 209–225.

Kennard C, Mannan SK, Nachev P, et al. (2005) Cognitive processes in saccade generation. *Annals of the New York Academy of Sciences* 1039: 176–183.

Lynch JC and Tian JR (2005) Cortico-cortical networks and cortico-subcortical loops for the higher control of eye movements. *Progress in Brain Research* 151: 461–501.

Martinez-Trujillo JC, Medendorp WP, Wang H, and Crawford JD (2004) Frames of reference for eye–head gaze commands in primate supplementary eye fields. *Neuron* 44: 1057–1066.

Olson CR (2003) Brain representation of object-centered space in monkeys and humans. *Annual Review of Neuroscience* 26: 331–354.

Pierrot-Deseilligny C, Rivaud S, Gaymard B, Müri R, and Vermersch AI (1995) Cortical control of saccades. *Annals of Neurology* 37: 557–567.

Schall JD (1991) Neuronal activity related to visually guided saccadic eye movements in the supplementary motor area of rhesus monkeys. *Journal of Neurophysiology* 66: 530–558.

Schall JD (1997) Visuomotor areas of the frontal lobe. In: Rockland K, Peters A, and Kaas J (eds.) *Extrastriate Cortex of Primates. Cerebral Cortex*, vol. 12, pp. 527–638. New York: Plenum.

Schall JD, Stuphorn V, and Brown JW (2002) Monitoring and control of action by the frontal lobes. *Neuron* 36: 309–322.

Schlag J and Schlag-Rey M (1985) Unit activity related to spontaneous saccades in frontal dorsomedial cortex of monkey. *Experimental Brain Research* 58: 208–211.

Schlag J and Schlag-Rey M (1987) Evidence for a supplementary eye field. *Journal of Neurophysiology* 57: 179–200.

Sommer MA and Tehovnik EJ (1999) Reversible inactivation of macaque dorsomedial frontal cortex: Effects on saccades and fixations. *Experimental Brain Research* 124: 429–446.

Tehovnik EJ (1995) The dorsomedial frontal cortex: Eye and forelimb fields. *Behavioural Brain Research* 67: 147–163.

Tehovnik EJ and Slocum WM (2004) Behavioural state affects saccades elicited electrically from neocortex. *Neuroscience & Biobehavioral Reviews* 28: 13–25.

Tehovnik EJ, Sommer MA, Chou IH, Slocum WM, and Schiller PH (2000) Eye fields in the frontal lobes of primates. *Brain Research Reviews* 32: 413–448.

Swim Oscillator Networks

A H Cohen and T Kiemel, University of Maryland, College Park, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

We begin our discussion of modeling swim oscillator networks with a brief review of the uses to which models of such networks are put. After this overview, we examine some types of models in the context of the central pattern generator (CPG) for locomotion of the lamprey. We also mention models of other swim networks.

Models of the lamprey swim network differ in the level or granularity of their description. Biophysical models focus on the specifics at the level of the membrane channel currents and details of the intra- and intersegmental connectivity. Such models have multiple variables per cell or cell type. (Often, one model neuron represents all neurons of a given type in each hemisegment.) Connectionist models have a single variable for each cell type and focus primarily on intra- and intersegmental connectivity. Phase models have a single variable for each segment and focus on intersegmental connectivity.

The choice of level of description for a model depends on both the research question being addressed and the trade-off between detail and simplicity. For example, a biophysical model is required to fully understand the impact of neuromodulators on the CPG. On the other hand, any model type can be used to investigate the production of intersegmental phase lags. Developing a biophysical model can be considered the ultimate goal since it accounts for the CPG's behavior on multiple levels of description. However, using a biophysical model requires estimating many parameters, only some of which can be directly estimated. Connectionist and phase models have fewer parameters, which simplifies both parameter selection and understanding of the resulting model. Given the lack of detailed empirical information, a model will not necessarily reproduce aspects of the CPG's behavior that were not considered when selecting its parameters. For example, all models are tuned to reproduce correct intersegmental phase lags. However, only some models reproduce correct functional intersegmental coupling strength or the details of the functional anatomy of the coordinating system.

The behavior to be captured by the models of swim networks is relatively simple. All swim oscillator networks produce some form of rhythmic output produced by a CPG. The rhythmic output of any swimming network is alternation among the muscles that move the body left/right or up/down, as well as a feature such as a traveling wave to aid in propulsion. In lamprey, the behavioral components of the motor output pattern are clearly defined as alternation between left and right muscles within a single spinal segment and a constant phase delay among the muscles along the side of the body (Figure 1). The traveling wave generated by this pattern typically forms a single wavelength along the body, which has been shown to be most efficient for this eel-like body form .

Modeling the Oscillator

We begin with the rhythm generation, or the mechanism for the oscillation between off and on states that accompany the oscillation of the membrane potential seen in the output elements of the CPG, the motor neurons. The oscillatory mechanism is often included with the mechanism for the alternation between the two sides of a single segment, but the two need not be linked. In modeling the oscillator, the first issue is whether the system depends on a single pacemaker or a number of distributed oscillators. The second issue is whether or not the oscillation is a function of membrane properties that generate the periodic changes in the membrane potential or whether it is a function of network connections. Indeed, most often it is a function of both membrane potential mechanisms and network connections, as has been found in the stomatogastric ganglion of Crustacea and the heart ganglia of the leech. In the lamprey, the evidence appears to support this double method as well. There are clearly membrane oscillations that appear in response to NMDA, and there are also network connections that are capable of forming an alternating output in combination with sufficient channel properties that are highly common.

It is also known that the CPG responsible for the oscillation is a distributed system, with each segment, hemisegment, or small group of segments capable of producing a rhythmic discharge of its own motor neurons. The exact minimal unit is still under debate, but there is evidence that it could be a hemisegment. On the other hand, the frequency appears to be a product of the ensemble of oscillators and apparently is not governed by any single oscillator.

Thus, any model of the lamprey swim network must account for rhythm generation in a single segment or hemisegment. There are several models for the network

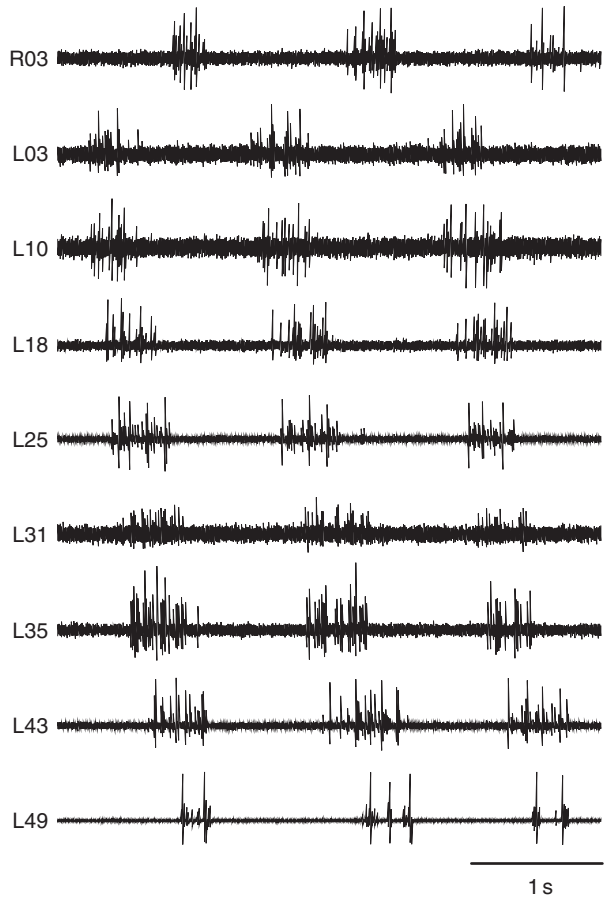

Figure 1 Neural activity recorded from the motor nerves of an isolated 50-segment piece of lamprey spinal cord. Activity from the left and right motor nerves of the same segment exhibits alternation (L03 and R03). Activity from different segments exhibits a rostrocaudal traveling wave. Numbers indicate the segment numbers.

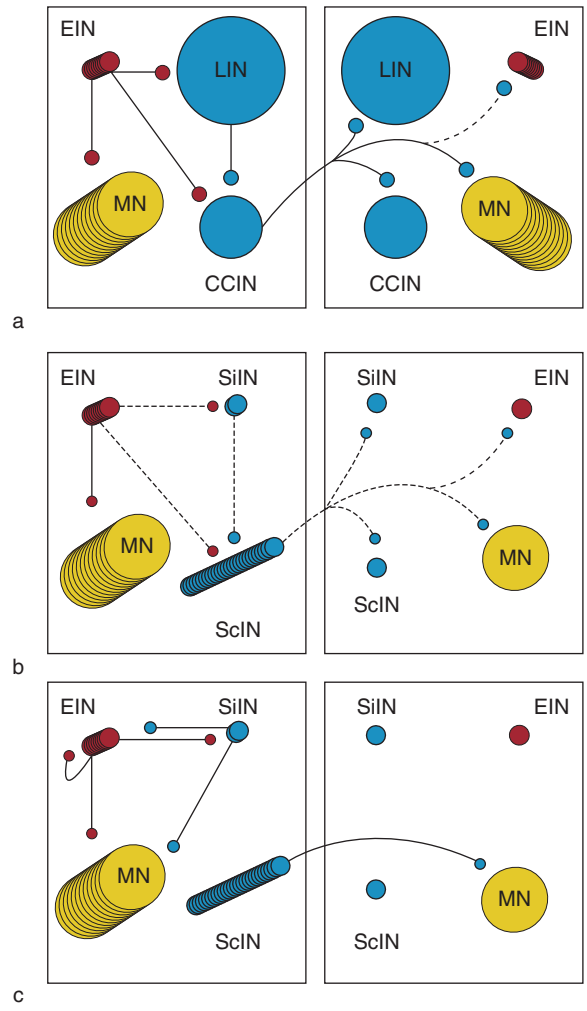

Figure 2 Known (solid lines) and presumed (dashed lines) connections in the lamprey CPG. Cell types are motor neurons (MN), excitatory interneurons (EIN), lateral inhibitory interneurons (LIN), crossed caudal interneurons (CCIN), small ipsilateral inhibitory interneurons (SiIN), and small crossing inhibitory interneurons (ScIN). (a) Connectivity scheme of Buchanan and Grillner based on pairwise recordings from different segments. (b) Segmental network if the LIN and CC interneurons are removed, based on the argument that these neurons provide primarily intersegmental coupling. (c) Evidence of intrasegmental connections. Figure provided by David Parker.

of the segmental unit CPG. The simplest was the original proposed by Buchanan and Grillner based on the experimental evidence of single cell recordings. The current network includes some additional neurons, again based on recordings of single neurons during rhythm generation by the isolated spinal cord, or what has been called 'fictive locomotion.' The basic oscillator model adds channel currents and some modulation. Such a model will almost certainly be missing some important elements of the oscillator, such as neuronal connections and/or channel currents. Nor can it include currents that have not yet been found that are either ligand gated or voltage gated. The current model does include some rather novel currents, such as potassium-dependent sodium channels.

There are many ways to model even a simple system such as the lamprey spinal cord. The issues are the neuronal composition and its connectivity. We present here, for comparison, the model put forth by David Parker, shown in Figure 2. The dashed lines represent those connections that are presumed to exist but for which there is no direct evidence. The solid lines represent those for which there is direct evidence. The inference of the connections from the large interneurons, such as the lateral interneurons and crossed caudal interneurons (CCINs), has been based on recordings between segments some distance apart. Although there have been attempts to prove that the large neurons do or do not have connections within their own segment, such evidence is very difficult to obtain directly. It is unclear how the

model will be affected by whether or not the inhibitory connections are from the larger CCINs or from the smaller crossed inhibitory interneurons. Unfortunately, answering these questions requires obtaining good recordings of each of the small interneurons in combination with their posited target neurons within a single segment; that task is very difficult. Obtaining the parameter values for the channel currents for each neuron throughout the dynamical changes that are likely to occur within a single cycle is even more difficult. Thus, although the model is a useful way to keep track of the latest view of the oscillator, one must be cautious of strong conclusions based on this or, indeed, any model.

The biophysically based model, in its current configuration, could be used to test the role of a variety of modulatory currents, such as those proposed by Parker and colleagues. This analysis would require that one perform a sensitivity analysis of those currents using perhaps two parameter values simultaneously. Unfortunately, there are far too many parameters in the current model for this kind of analysis to be practical. However, an abstracted form of the model could be developed in which such an analysis could be performed. The results of such a study would suggest which currents and modulators to vary in order to determine more accurately their roles in the development of the rhythm.

The oscillator networks have been modeled in several model systems, including the tadpole of *Xenopus* and the salamander. There are also examples among the invertebrates, including the leech. In some cases, the issues relate to the development of the network, whereas in others they relate to the rhythm generation alone.

In summary, a major difficulty with biophysically detailed models in which the parameters values are not fully known is that there are many 'free' parameters. That is, there are many parameters whose values during the motor behavior are dynamically changing but are known only from static experiments, at best. When one tweaks the parameters to achieve the desired motor output pattern, one can relatively easily change enough parameters to obtain the desired behavioral regime.

We demonstrate this with a discussion of a totally different kind of modeling by Jung and colleagues for the oscillator which used a bifurcation analysis to illustrate the range of potential output patterns that can be assumed by the simple six-neuron network of Williams. The model we consider makes no claims to realism. It has few parameters except the weights of the connections among the classes of the neurons and the descending drive strength that serves as the excitation to the network. In this simple network, we found that a parameter unnecessary for the oscillation was the connectivity of the excitatory neurons. Experimental evidence from the Grillner group has since been presented to counter this possibility, but the results of the model do suggest that one can, with a simple network, produce a relatively large range of behaviors. One must be cautious about any conclusions one draws from any model because the realm of its possible behaviors is manyfold and typically largely unexplored.

Modeling the Traveling Wave Motion

In modeling the traveling wave motion, there are other fundamental issues that must be addressed besides rhythm. The first is the necessity of sensory feedback. Does sensory feedback produce the sequential activation of the segments? In the lamprey CPG and several other CPGs (e.g., leech and *Xenopus*), it has been shown that the traveling wave among the segments appears in the absence of sensory feedback. Sensory feedback may contribute to the traveling wave motion, but it is not necessary for the motion to occur. The next question is whether the delays among the segments are a constant phase of the cycle duration or a constant delay. If they are a constant delay, the mechanism can be relatively easily explained, potentially, as due to conduction delays along the axons connecting the segmental oscillators. However, in most systems the delays are a constant phase angle of the cycle and not a constant delay. Thus, they are due to a computation wherein the delay is changed for each frequency at which the oscillators burst. In the lamprey, for example, this has been shown to be the case. The delays among the segments are a constant phase of the cycle regardless of the frequency even in the complete absence of sensory feedback. That is, in the case of 'fictive swimming,' the delays are roughly appropriate for the frequency seen that is elicited by the addition of excitatory amino acids to the bath holding the isolated spinal cord. The mechanism for the computation of the constant phase delay has been the focus when modeling the intersegmental coordination of the segmental oscillators.

Modeling the traveling wave motion began in 1982 with a paper by Cohen et al. Their model was a phase approximation of a chain of coupled stable limit cycle oscillators. A stable limit cycle oscillator has the property that, following a perturbation, the trajectories of the oscillator's variables return to their original waveforms; the only lasting effect of the perturbation is a phase shift. If stable limit cycle oscillators with similar frequencies are weakly coupled, then it is possible to derive a phase model that approximates the behavior of the system. The state of each oscillator is described by a single variable, its

absolute phase, and coupling between oscillators depends only on the relative phases of the oscillators. The behavior of a phase model is determined by the uncoupled frequency of each oscillator and the relative phase coupling functions. There can be a different coupling function for each pair of pre- and postsynaptic oscillators, although typically the coupling function is assumed to depend only on the relative position of the two oscillators in the chain.

Use of a phase approximation simplifies the understanding of intersegmental coordination by dividing the problem into two parts: (1) understanding how the details of the neural dynamics and neural circuitry affect uncoupled oscillator frequencies and the relative phase coupling functions and (2) understanding how the properties of the oscillator frequencies and coupling functions relate to intersegmental coordination. Cohen et al. addressed the second part of the problem. They assumed that coupling functions were sine functions. An important property of a coupling function is its preferred phase, the relative phase that would be produced by unidirectional coupling between two oscillators with equal uncoupled frequencies. The preferred phase of a sine coupling function is either in-phase or anti-phase, depending on the sign of the sine wave's coefficient. Cohen et al. showed that a traveling wave could be produced in two ways: (1) by a gradient of uncoupled oscillator frequencies and nearest-neighbor in-phase coupling and (2) by equal uncoupled oscillator frequencies, nearest-neighbor in-phase coupling, and anti-phase coupling between the first and last oscillator in the chain. These two examples correspond to the two basic mechanisms for producing a traveling wave in a chain of coupled oscillators: a systematic variation in oscillator frequencies (the oscillator frequency mechanism) or non-in-phase coupling functions (the coupling function mechanism). For the lamprey CPG, experimental evidence supports the coupling function mechanism as the source of the traveling wave. For example, when a piece of spinal cord is cut into smaller parts, the relative frequencies of these parts do not show a systematic pattern.

Work by Ermentrout and Kopell has provided a general understanding of the coupling function mechanism for phase models with short-distance coupling. Allowing preferred phases to take on any value, not just in-phase or anti-phase, they showed that the relative strength of ascending and descending coupling plays a critical role in determining intersegmental phase lags. For example, in the case of nearest-neighbor coupling with ascending coupling stronger than descending coupling, phase lags between adjacent oscillators are close to the preferred phase of the ascending coupling, except near the rostral end of the chain. This coupling dominance can be seen in certain swim models, such as the lamprey CPG models of Williams and Ekeberg.

Although the behavior of many swim models can be understood using a phase approximation, there are two important limitations. First, a phase approximation is only guaranteed to be valid in general if coupling is weak. For strong coupling, a phase approximation will be valid only in very restrictive special cases. Second, swim models only exhibit certain properties, such as coupling dominance, if long-distance coupling is absent or very weak. To test whether these two assumptions, weak and predominantly short coupling, hold for the lamprey CPG, we developed a method of estimating functional coupling strength based on recordings from two motor nerves. The method is based on estimating parameters in a model of two coupled phase oscillators and gleans information from cycle-to-cycle variation in cycle periods and intersegmental phase lags. Functional coupling strength is quantified by the measure β, which varies from 0 (no coupling) to 1 (infinitely strong coupling). For a 50-segment piece of spinal cord, functional coupling between the rostral and caudal ends is very strong (Figure 3). Also, by inhibiting 20 middle segments, we found that long-distance connections provided substantial coupling. Therefore, as summarized in Figure 3, coupling in the lamprey CPG is neither weak nor predominantly short.

The functional coupling strength measure β can be estimated for any model by adding noise to create cycle-to-cycle variations in cycle periods and intersegmental phase lags. We have estimated functional coupling strength for three models: the connectionist model of Williams; the model of Ekeberg, which was

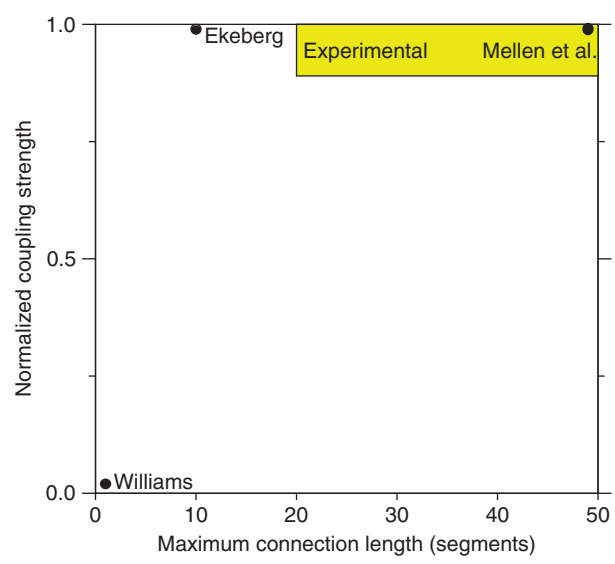

Figure 3 Comparison of experimental estimates of coupling length and function strength to those of various models.

intended as a simplified version of the more-detailed biophysical models of Grillner and colleagues; and the phase model of Mellen et al. Figure 3 shows β for each model along with its maximum connection length. Note that the Williams model, which has only nearest-neighbor connections, exhibits weak functional coupling, whereas the other two models, which contain longer-distance connections, exhibit strong functional coupling.

The traveling wave of muscle activation along the body is a necessary element of the behavioral motor pattern underlying swimming. A full model for swim behavior should include the known properties of the intersegmental coordinating system fiber anatomy and physiology. Most models have focused on reproducing intersegmental phase lags, not functional intersegmental coupling strength.

Conclusion

It is critically important that models are never mistaken for truth. Regardless of how realistic they appear, they are unlikely to ever include all critical parameters, cells, and connections, at least until new methods for determination of neurons, their channels, and currents as well as their connections can be found. Until then, one can only state that a model contains sufficient anatomical and physiological elements for the generation of the behavioral features of swimming motor output. This holds true for both the rhythmic oscillations and the traveling wave.

One important conclusion from modeling work on the lamprey CPG is that it is advantageous to simultaneously model the system at different levels of description. For systems consisting of coupled oscillators, such as the lamprey and salamander spinal cords and leech nerve cord, phase models can provide insight into how factors such as frequency differences, coupling length, and coupling asymmetry lead to the system's observed behavior. For example, phase models of the lamprey CPG suggest the importance of long connections, but not frequency differences, in setting phase lags. In contrast to phase models, more-detailed models allow us to understand the underlying basis for frequency differences and coupling properties. More-detailed models also potentially allow us to check the validity of a phase model approximation. Thus, the complementary use of detailed models and high-level models often offers the best chance of gaining insight into a complex system's behavior. This should remain true regardless of the details of the underlying physiology of the system. Such systems are notoriously difficult to understand intuitively. Modeling often offers the only sensible approach, as has been shown by a robotic implementation of a salamander by Auke Ijspeert and colleagues in which they demonstrated a potential mechanism for changing gaits in tetrapods.

See also: Central Pattern Generators; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Computational Approaches to Motor Control; Learning, Action, Inference and Neuromodulation; Pattern Generation; Swimming: Neural Mechanisms .

Further Reading

Buchanan JT (1992) Neural network simulations of coupled locomotor oscillators in the lamprey spinal cord. *Biological Cybernetics* 66(4): 367–374.

Cang J and Friesen WO (2002) Model for intersegmental coordination of leech swimming: Central and sensory mechanisms. *Journal of Neurophysiology* 87(6): 2760–2769.

Cohen AH, Ermentrout GB, Kiemel T, Kopell N, Sigvardt K, and Williams T (1992) Modelling of intersegmental coordination in the lamprey central pattern or for locomotion. *Trends in Neurosciences* 15: 434–438.

Cohen AH, Holmes PJ, and Rand RH (1982) The nature of the coupling between segmental oscillators of the lamprey spinal generator for locomotion: A mathematical model. *Journal of Mathematical Biology* 13: 345–369.

Cohen AH and Wallén P (1980) The neuronal correlate to locomotion in fish: 'Fictive swimming' induced in an *in vitro* preparation of the lamprey spinal cord. *Experimental Brain Research* 41: 11–18.

Ekeberg Ö (1993) A combined neuronal and mechanical model of fish swimming. *Biological Cybernetics* 69: 363–374.

Ijspeert AJ (2001) A connectionist central pattern generator for the aquatic and terrestrial gaits of a simulated salamander. *Biological Cybernetics* 84(5): 331–348.

Jung R, Kiemel T, and Cohen AH (1996) Dynamical behavior of a neural network model of locomotor control in the lamprey. *Journal of Neurophysiology* 75: 1074–1086.

Kiemel T, Gormley KM, Guan L, Williams TL, and Cohen AH (2003) Estimating the strength and direction of functional coupling in the lamprey spinal cord. *Journal of Computational Neuroscience* 13: 233–243.

Kopell N and Ermentrout GB (1990) Phase transitions and other phenomena in chains of coupled oscillators. *SIAM Journal on Applied Mathematics* 50: 1014.

Parker D (2000) Spinal-cord plasticity: Independent and interactive effects of neuromodulator and activity-dependent plasticity. *Molecular Neurobiology* 22(1–3): 55–80.

Roberts A, Tunstall MJ, and Wolf E (1995) Properties of networks controlling locomotion and significance of voltage dependency of NMDA channels: Stimulation study of rhythm generation sustained by positive feedback. *Journal of Neurophysiology* 73(2): 485–495.

Wallén P, Ekeberg Ö, Lansner A, Brodin L, Tråvén H, and Grillner S (1992) A computer-based model for realistic simulations of neural networks: II. The segmental network generating locomotor rhythmicity in the lamprey. *Journal of Neurophysiology* 68: 1939–1950.

Wallén P and Williams TL (1984) Fictive locomotion in the lamprey spinal cord *in vitro* compared with swimming in the intact and spinal animal. *Journal of Physiology* 347: 225–239.

Williams TL (1992) Phase coupling by synaptic spread in chains of coupled neuronal oscillators. *Science* 258(5082): 662–665.

Swimming: Neural Mechanisms

Y I Arshavsky, University of California at San Diego, La Jolla, CA, USA
T G Deliagina and G N Orlovsky, Karolinska Institute, Stockholm, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Swimming is a mode of locomotion in most aquatic animals. In the case of microscopic planktonic animals, swimming is performed through the rhythmic beating of body cilia. Larger animals swim using periodic muscle contractions, which are controlled by the nervous system. Swimming is usually adapted to the environment. Particularly when swimming, animals are oriented in space. Like other automatic rhythmic movements, swimming is controlled by ensembles of neurons – central pattern generators (CPG) – which are capable of producing the basic spatiotemporal motor output in the absence of sensory feedback. Therefore, the neural mechanisms underlying swimming can be studied in terms of the activity of single neurons and neuron-to-neuron interactions using *in vitro* preparations (*in vitro*-generated swimming-like rhythmic activity is called fictive swimming). These studies have shown that invertebrates and vertebrates share some general principles of swim control, regardless of the mode of organization and the complexity of their nervous systems.

Generation of Rhythmic Swimming Activity

To understand the basic neural mechanisms that underlie swim movements, it is important to answer two main questions:

- What is the source of rhythmic activity?
- What mechanisms shape the pattern of motor output (i.e., determine the number of phases in a cycle and the transitions from one phase to another)?

These two aspects of swimming CPG activity are closely interconnected and, in general, are determined by the endogenous properties of generator neurons and their interactions. However, the relative importance of cellular and network mechanisms in generating a swimming rhythm, and in shaping the motor output, is different. The patterning of the motor output is mainly determined by interactions between the neurons. In contrast, experiments with physical and functional isolation of neurons show that the decisive role in rhythm generation is played by specific groups of interneurons (INs) that are endogenous pacemakers. Unlike rhythmic movements (e.g., heartbeat and respiration) that continue throughout the life span and are controlled by constitutive pacemakers, swimming is an episodic movement. Therefore, swimming rhythms are generated by conditional pacemakers, which are transferred from a nonactive to a rhythm-generating state by command signals. *In vitro*, pacemaker activity can be initiated by applying the neurotransmitters that are used by command inputs. Several examples of swim control systems are described for illustration.

Jellyfish Swimming

Jellyfish swim by periodically contracting the musculature. Jellyfish's musculature is composed of myoepithelial cells whose contractile processes form a circular sheet. Due to the jellyfish's radial symmetry, their nervous system has annular organization. It includes pacemaker neurons that continue to generate a swimming rhythm after isolation of a small part of the nerve ring. Putative command neurons are nitrergic neurons that initiate pacemaker activity through the NO/cGMP signaling pathway. The pacemakers form excitatory synapses on motor neurons (MNs), which innervate circular musculature.

Swimming cycles in jellyfish have only one active phase of muscle contraction, followed by passive relaxation. Therefore, jellyfish exhibit the simplest mode of interneural interactions. Pacemaker neurons interact via electrical synapses, as well as via unique bidirectional chemical excitatory synapses, where both synaptic terminals can serve as pre- and postsynaptic elements; inhibitory interaction between neurons is absent. This mode of neuron interaction performs two functions. First, it synchronizes the activity of generator neurons and therefore (along with electrical connections between the myoepithelial cells) allows the simultaneous contraction of circular musculature. Second, the frequency with which individual pacemakers fire is variable, and in the system of neurons with bidirectional connections there is a constant shifting of the leading function between pacemakers; each following cycle is triggered by the fastest element. Thus, interneural interactions raise both the frequency and the regularity of the generator output.

In addition to regular (slow) swimming, the stimulation of sensory cells evokes a fast escape reaction. This reaction involves the giant ring axon, which in turn excites MNs, producing fast, strong muscle contractions. The different strength of muscle contractions during regular and escape swimming is due to the ability of MNs to conduct two types of

spikes. Pacemaker neurons produce low-threshold calcium spikes of a small amplitude, whereas the giant axon produces high-threshold sodium spikes. The two types of impulses propagate independently because the peak of the calcium spike is below the threshold of the sodium spike.

Pteropod Mollusk

The pelagic marine mollusk *Clione limacina* (**Figure 1(a)**) swims by rhythmically moving its two wings. Each swim cycle consists of two phases, dorsal and ventral wing flexions, which are produced by alternating contractions of antagonistic wing muscles (**Figures 1(b) and 1(d)**). The swimming CPG is located in the pedal ganglia that give rise to the wing nerves. Two antagonistic groups of INs (INs 7 and 8 in **Figures 1(c) and 1(d)**) play the decisive role in rhythm generation. The INs generate prolonged (~100 ms), tetrodotoxin-resistant action potentials; there is one potential per cycle (**Figure 1(c)**). These action potentials determine the duration of the cycle phases. In young mollusks that have a higher frequency of wing oscillations (due to biomechanical reasons), the duration of IN action potentials is much shorter than in adults.

The INs of both groups are endogenous pacemakers. After extraction from the ganglion, they continue (under the influence of serotonin) to generate periodic action potentials with the frequency typical of wing beating (**Figures 1(e) and 1(f)**). Each neuron discharge is preceded by a pacemaker potential (**Figure 1(e2)**, arrowhead). Extraexcitation of the isolated IN shifts the phase of rhythmic activity (**Figure 1(f)**). One of the command systems, which transfers these pacemakers to a rhythm-generating state, is formed by serotonergic neurons located in the cerebral ganglia. The initiation or acceleration of swimming activity induced by the serotonergic neurons are not accompanied by the depolarization of pacemaker INs (**Figure 1(g)**). This suggests that the effect of serotonin is realized through metabotropic receptors. Simultaneously, the serotonergic neurons depolarize MNs, which results in an increase of their spike discharges in the proper phases of the cycle (**Figure 1(g)**). The serotonergic mechanisms, in addition to swim activation, enhance wing muscle contractility and accelerate the heartbeat. Therefore, these mechanisms perform both command and integrative functions.

The other command system is formed by the hunting neurons located in the cerebral ganglia. During *Clione* hunting, these neurons initiate the activity of different motor systems, including the swim control system. The putative transmitter used by the hunting neurons is acetylcholine because *in vitro* fictive hunting behavior can be evoked by the cholinesterase inhibitor. In addition to excitatory command neurons, there are inhibitory command neurons producing an abrupt halt of the swimming rhythm (e.g., in response to *Clione* head stimulation).

The pacemaker INs of each group are electrically interconnected and inhibit antagonistic INs (**Figure 1(d)**). A system of neurons with this type of interconnections generates biphasic output with synchronous excitation of synergetic neurons and a half cycle shift between phases (**Figures 1(b) and 1(c)**). Mutual inhibition of antagonistic INs not only determines their alternating excitation but also, due to postinhibitory rebound, assists the transition from one phase to another and, therefore, enhances the reliability of the rhythm generation. The rebound phenomenon facilitates the beginning of each following phase of the cycle through the hyperpolarization-activated inward current (I_h) occurring in the INs after the termination of the inhibitory postsynaptic potentials.

An additional mechanism contributes to the reliable transition from one phase of the cycle to another. The swimming CPG includes one more group of INs (INs 12 in **Figure 1(c)**). They are nonspiking cells that generate long-lasting plateau potentials. INs 12 are not endogenous oscillators, and their rhythmical activity is determined by inputs from the pacemaker INs. They receive an excitatory input from the ventral phase INs and an inhibitory input from the dorsal phase INs (**Figure 1(d)**). When excited, INs 12 produce a slight recurrent inhibitory effect on the ventral phase INs and an excitatory effect on the dorsal phase INs. Thus, INs 12 promote the termination of the ventral phase and initiation of the dorsal phase of the cycle.

The pacemaker INs control the activity of wing muscle MNs. They excite the synergetic MNs and inhibit the antagonistic ones (**Figure 1(d)**). Simultaneous excitation of synergetic MNs is determined by both their common inputs from INs and electrical interconnections.

Unlike *Clione*, rhythm generation in the other marine mollusk, *Tritonia*, is solely based on interneural interactions. The main reason for this difference is the fact that this benthos mollusk does not demonstrate sustained swimming. When *Tritonia* are touched by the ambulacral feet of a predatory starfish, this contact initiates a short series of alternating dorsal/ventral flexions of the body, which cause it to rise up from the substrate and be carried away by water current.

Medicinal Leech

To swim, a leech flattens itself by contracting its dorsoventral muscles; it then undulates by alternating contractions of dorsal and ventral longitudinal

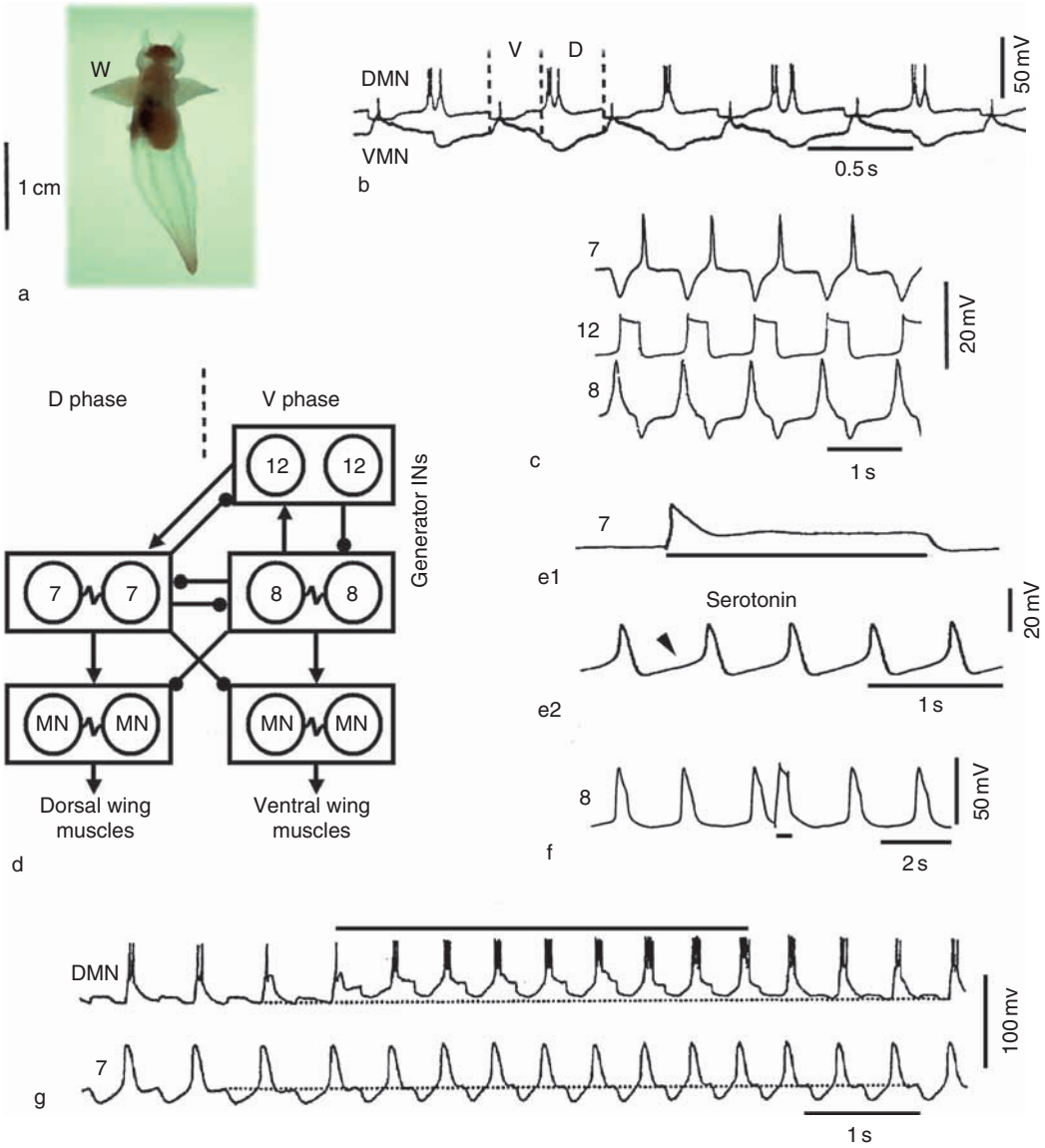

Figure 1 Neural mechanisms controlling swimming in the pteropod mollusk. (a) Photo of *Clione* during swimming. W, wings. (b and c) Fictive swimming recorded in preparations of the isolated nervous system; simultaneous intracellular recordings from two antagonistic MNs active in the ventral (V) and dorsal (D) phases of the cycle (b) and from the generator INs 7, 8, and 12 (c). (d) Organization of the swimming CPG; arrows, filled circles, and resistor symbols show chemical excitatory and inhibitory connections and nonrectifying electrical connections, respectively. (e, f) The INs 7 and 8 are conditional pacemakers. An IN extracted from the pedal ganglion does not generate rhythmic activity, spontaneous or in response to prolonged depolarization (horizontal line in e1), but is transferred in a rhythm-generating state by serotonin application (e2). Extraexcitation of the isolated IN by current injection shifts the phase of rhythmic activity (f). (g) Intracellular stimulation of a cerebral serotonergic command neuron (horizontal line) results in an acceleration of the ongoing rhythmic activity and an increase of MN discharges; dotted lines indicate initial levels of the membrane potential. (b) Adapted with permission from Arshavsky YI, Beloozerova IN, Orlovsky GN, Panchin YV, and Pavlova GA (1985) Control of locomotion in marine mollusc *Clione limacina.* 1: Efferent activity during actual and fictitious swimming. *Experimental Brain Research* 58: 255–262. (c) Adapted with permission from Panchin YV, Sadreev RI, and Arshavsky YI (1995) Control of locomotion in the marine mollusk *Clione limacina.* 10: Effects of acetylcholine antagonists. *Experimental Brain Research* 106: 135–144. (e, g) Adapted with permission from Panchin YV, Arshavsky YI, Deliagina TG, et al. (1996) Control of locomotion in the marine mollusk *Clione limacina.* 11: Effects of serotonin. *Experimental Brain Research* 109: 361–365. (f) Adapted with permission from Arshavsky YI, Deliagina TG, Orlovsky GN, et al. (1986) Control of locomotion in marine mollusc *Clione limacina.* 6: Activity of isolated neurons of pedal ganglia. *Experimental Brain Research* 63: 106–112.

muscles, which produce quasi-sinusoidal, rearward-traveling waves, pushing the animal forward through the water. When the frequency of the undulations increases, the velocity of wave spreading also increases. As a result, there is no change in intersegmental phase lag and, therefore, in wavelength. The swimming CPG is distributed throughout the segmental nerve cord (Figures 2(a)–2(c)).

Rhythmic activity can be generated by single ganglia (Figure 2(d)). Each ganglion includes one unpaired and six paired putative oscillator INs (Figure 2(e)). There is no direct evidence that oscillator INs are endogenous pacemakers. Indirectly, the suggestion that at least some of them are endogenous oscillators is supported by the fact that intracellular injections of current shift the phase of the cycle (Figure 2(f)), which is typical of pacemakers (Figure 1(f)).

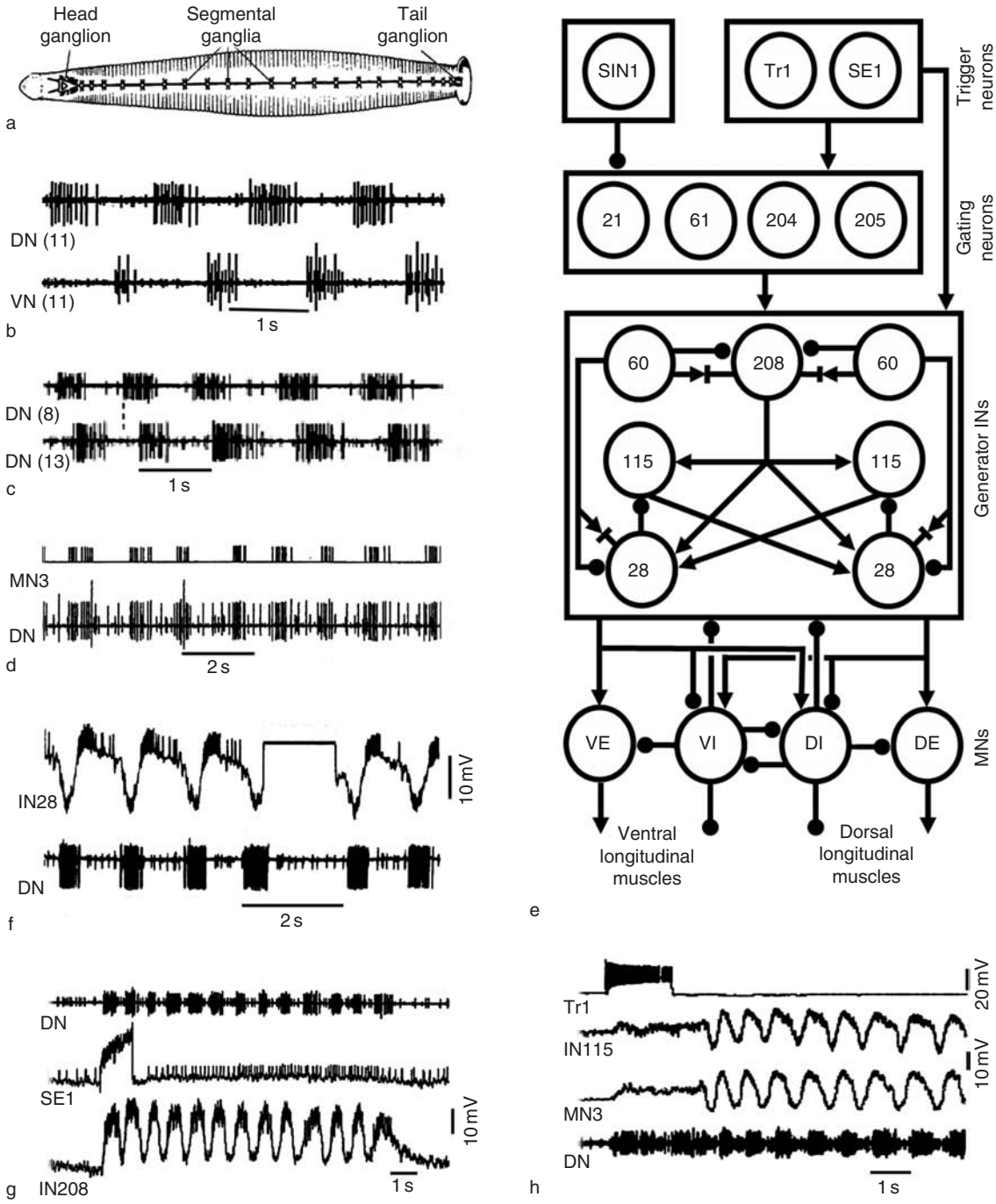

Figure 2 Neural mechanisms controlling swimming in the medical leech. The segmentally organized nervous cord (a) can generate fictive swimming after isolation (b, c). (b) Extracellular recording from the nerves supplying dorsal (DN) and ventral (VN) longitudinal muscles (the numbers of the segments are indicated in parentheses). (c) Activity in DNs from two segments (intersegmental lag is indicated by the dotted line). (d) Rhythmic activity generated by a single segment that had been isolated for 25 days; the upper trace indicates the time of occurrence of impulses in MN3 supplying a dorsal muscle. (e) The main groups of neurons involved in the swimming control; only one unpaired and three paired generator INs are shown; all homologous of INs are electrically coupled (not shown).

Interactions between oscillator INs within a ganglion are very complex (Figure 2(e)). All paired neurons are electrically connected. They communicate through inhibitory and rectifying electric synapses. The unpaired IN is an excitatory one. Some INs participate in the control of alternating activity in segmental antagonistic MNs (Figure 2(e)). In addition to excitatory MNs producing muscle contractions (dorsal and ventral excitors), there are output neurons designed as dorsal and ventral inhibitors. These MNs both inhibit the antagonistic excitors and reduce muscle contractions induced by the excitors. Thus, inhibitory MNs promote alternating contractions of antagonistic muscles. In addition, the inhibitors produce recurrent inhibition of some oscillator INs and, therefore, may be involved in the rhythm generation.

Unlike MNs that demonstrate biphasic activity (Figure 2(b)), phases of activity in INs are more dispersed. This is related to the fact that the oscillator INs not only control MN activity but also perform an additional function. They send axons either rostrally or caudally, and these intersegmental interactions determine the travel of the waves of excitation in the rostrocaudal direction (Figure 2(c)). Along with these central mechanisms, stretch receptors (which are embedded in the longitudinal muscles) also play an important role in organizing undulatory movements with proper intersegmental phase lags.

Command neurons initiating leech swimming have a two-level organization (Figure 2(e)). Neurons of a higher level (triggers) are located in the head ganglion. They receive inputs from body receptors, whose stimulation initiates swimming. The trigger neurons activate segmental command neurons (gaiting neurons) and, to some extent, oscillator neurons. The transmitters used by command neurons are glutamate and serotonin. Brief stimulations of command neurons initiate prolonged swimming bouts (Figures 2(g) and 2(h)). This suggests that they act through metabotropic synapses, although there is no direct evidence. Some command neurons perform integrative functions. In addition to activation of the swimming CPG, they excite flattener MNs. The command system in the leech, as in *Clione*, includes not only excitatory but also inhibitory neurons.

Swimming in Vertebrates

Among vertebrates, the lamprey's neural mechanisms of swimming have been studied in greatest depth. Lampreys swim by alternating contractions of antagonistic lateral muscles, which produce rearward-traveling undulatory waves. As in the leech, the undulatory waves spread with a constant intersegmental phase lag. The swimming CPG is distributed throughout the spinal cord. It is activated by the glutamatergic reticulospinal pathway – the main command system activating locomotion in all vertebrates. Rhythmic activity is generated due to intrinsic membrane properties of excitatory INs. Synchronous activity of the unilateral INs is determined by their mutual interconnections. These INs excite ipsilateral MNs and, through commissural INs, inhibit neurons of the contralateral side. This contralateral inhibition leads to alternating activity of right/left MNs. Although an inhibitory interaction is not necessary for rhythm generation, it may promote a transition from one phase to another due to postinhibitory rebound. In addition to central mechanisms, reliable rhythm generation is supported by signals from the stretch receptors located at the lateral margin of the spinal cord. The rostrocaudal travel of the waves of excitation is mainly determined by a higher inherent frequency of the rostral segmental CPGs. There is a similar organization in the swim control system in teleost fish swimming due to undulatory movements.

In addition to regular swimming, both Cyclostomata and teleosts exhibit startle/escape reactions that occur during prey capture or in response to threatening stimuli. These reactions are realized through the specific reticulospinal command system, including the giant Mauthner cells with high-diameter axons. Descending signals activate trunk MNs, which produce a unilateral bend of the body, followed by its fast straightening and, thus, by a propulsive tail stroke. Spinal circuits

(f) Current injection into the generator IN28 shifts the phase of rhythmic activity of the whole system. (g, h) Short stimulations of trigger (command) neurons SE1 (g) and Tr1 (h) initiate prolonged bouts of fictive swimming as monitored by rhythmic activity of MNs and generator INs. (a, c) Adapted with permission from Brodfuehrer PD and Thorogood MS (2001) Identified neurons and leech swimming behavior. *Progress in Neurobiology* 63: 371–381. (b) Adapted with permission from Shaw BK and Kristan WB (1977) The neuronal basis of behavioral choice between swimming and shortening in the leech. *Journal of Neuroscience* 17: 786–795. (d) Adapted with permission from Kristan WB and Guthrie PB (1977) Acquisition of swimming behavior in chronically isolated single segments of the leech. *Brain Research* 131: 191–195. (e) Adapted with permission from Kristan WB, Calabrese RL, and Friesen WO (2005) Neuronal control of leech behavior. *Progress in Neurobiology* 76: 279–327. (f) Adapted with permission from Friesen WO, Poon M, and Stent GS (1976) An oscillatory neuronal circuit generating a locomotory rhythm. *Proceedings of the National Academy of Sciences of the United States of America* 73: 3734–3738. (g) Adapted with permission from Brodfuehrer PD, Parker HJ, Burns A, and Berg M (1995) Regulation of the segmental swim-generating system by a pair of identified interneurons in the leech head ganglion. *Journal of Neurophysiology* 73: 983–992; and Brodfuehrer PD and Friesen WO (1986) Initiation of swimming activity by trigger neurons in the leech subesophageal ganglion. *Journal of Comparative Physiology A* 159: 489–502.

involved in regular swimming and escape behaviors partly overlap but are not identical. In particular, the activation of red muscles is dominant during regular swimming, whereas the escape reaction involves massive activation of fast, white muscles.

The neural mechanisms of swimming in amphibians were mainly studied on tadpoles, including the *Xenopus laevis* embryo released from the egg before normal hatching. As in the lamprey, the propulsive force arises from lateral body undulations, produced by alternating contractions of right/left muscles (**Figures 3(a) and 3(b)**). Neural organization of the swimming CPG in tadpoles (**Figure 3(c)**) is similar to that in *Clione* (**Figure 1(d)**) and in the lamprey. The core role in rhythm generation is played by excitatory INs that, together with ipsilateral commissural INs and MNs, form two antagonistic half-centers. The generator INs excite each other and, via commissural INs, inhibit neurons of the contralateral half-center, which provides a half-cycle shift between right/left efferent discharges (**Figure 3(b)**). As in *Clione*, synchronous activity of homonymous MNs is determined not only by the common input from the generator INs but also by their electrical coupling.

A swimming-like rhythm can be generated by a surgically isolated longitudinal half of the rostral (but not caudal) spinal cord taken from the *Xenopus* embryo (**Figure 3(d1)**). This suggests that generator neurons located in the rostral part of the spinal cord, and probably in the caudal hindbrain, are endogenous pacemakers capable of rhythm generation in the absence of inhibitory interactions between antagonistic neurons. Due to the rostral localization of the pacemaker neurons, an intersegmental delay of the wave of excitation underlying undulatory movements does not depend on the cycle period (as it does in the leech and lamprey). An additional mechanism which contributes to rhythm generation is reciprocal right/left inhibition followed by a postinhibitory rebound. An important role of the postinhibitory rebound follows from the fact that the glutamate agonist NMDA initiates rhythmic activity in the isolated hemicord only in the presence of Mg^{2+}, whereas the rhythmogenic effect of NMDA on the intact spinal cord does not depend on Mg^{2+}.

The decisive role in transferring the pacemaker INs from a passive to a rhythm-generating state is played by the glutamatergic command input(s) from higher centers. The application of the glutamate antagonist to the isolated hemicord blocks generation of swimming activity (**Figure 2(d2)**). Glutamate is also used as a transmitter by the excitatory generator INs. Therefore, mutual interactions between the excitatory INs not only lead to their synchronous activity but also strengthen the command signal. In addition to glutamatergic mechanisms, serotonergic mechanisms participate in transferring the conditional pacemakers into a rhythm-generating state. The application of NMDA combined with serotonin to the tadpole's spinal cord, which was treated with tetrodotoxin (which blocks spike activity and, therefore, disconnects synaptic interactions), produced rhythmic oscillations of the membrane potential in swim-control neurons (**Figure 3(e)**). This supports the idea that the generator neurons are endogenous oscillators.

As in other animals, the swimming CPG in tadpoles receives not only excitatory but also inhibitory command signals. These signals are transmitted through the GABAergic reticulospinal pathway that produces prompt termination of swimming in response to bumping into an obstruction.

Small tadpole embryos demonstrate a high frequency of undulations (up to 25 Hz). Correspondingly, an embryo's spinal neurons generate only one spike per cycle (**Figure 3(b)**). The mechanism preventing the neurons from repetitive discharges is a slowly inactivating increase of K^+ conductance, which maintains prolonged membrane refractoriness (**Figure 3(f1)**). After hatching, membrane properties start to change, and the neurons gain the ability to generate bursts of spikes (**Figure 3(f2)**). This leads to a decrease in the frequency of undulations in a growing tadpole (**Figure 3(g)**).

Thus, even the relatively simple neural networks that control swimming in both invertebrates and lower vertebrates have a redundant organization. All the important aspects of the network activity (rhythm generation, synchronous activity of synergetic neurons and alternating activity of antagonistic ones, transition from one phase of the cycle to another, etc.) are performed by multiple mechanisms that complement each other. This ensures a high reliability of swim control.

Spatial Orientation during Swimming

Many species actively stabilize a particular body orientation in space during swimming. Analysis of the systems for postural orientation and equilibrium control has been performed at behavioral and network levels in two animals, mollusk *Clione* and lamprey. It was shown that both actively maintain a specific body orientation due to gravitational reflexes driven by signals from the vestibular organs (lamprey) or from their functional analogs in invertebrates, the statocysts (*Clione*).

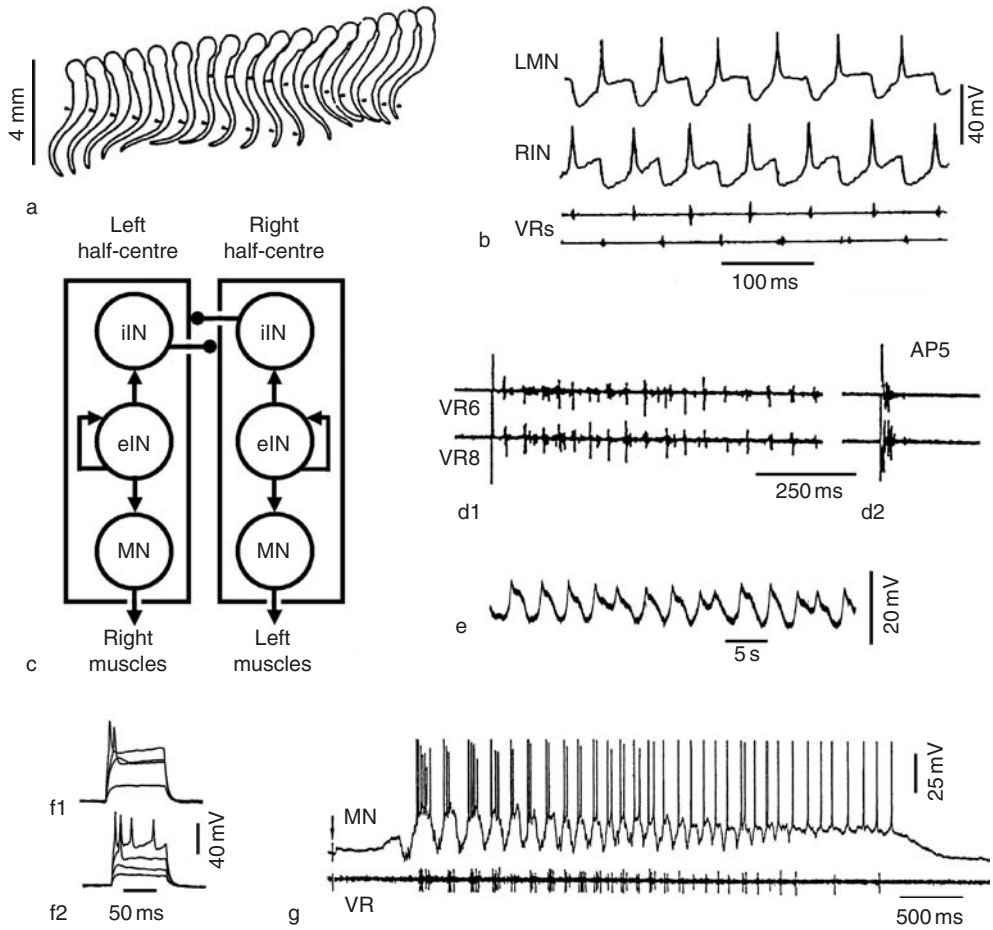

Figure 3 Neural mechanisms controlling swimming in tadpoles. (a) Tracing of swimming movements of a *Xenopus* embryo (intervals between traces are 3.3 ms); arrowheads indicate waves of bending, which progress toward the tail. (b) Fictive swimming in an embryo immobilized with *d*-tubocurarine; simultaneous recordings from a left spinal MN (LMN), right IN (RIN), and two symmetrical ventral roots (VR). (c) Organization of the swimming CPG; eIN, excitatory INs; iIN, inhibitory commissural INs. (d) A longitudinally split rostral part of the spinal cord isolated with a caudal part of the hindbrain generates swimming-like rhythmic activity as monitored by recordings from the sixth and eighth ventral roots in response to hindbrain stimulation (d1); the NMDA antagonist, 2-amino-5-phosphonovaleric acid (AP5), blocks the rhythm generation (d2). (e) Application of NMDA and serotonin to the tetrodotoxin-treated spinal cord of a *Rana temporaria* tadpole produces rhythmic oscillations of the membrane potential in spinal neurons. (f) Injection of depolarizing current of different strength into a spinal neuron of the *Xenopus* embryo evokes only single spikes (f1), whereas current injection into a neuron of the tadpole (24 h after hatching) can produce a burst of impulses (f2). (g) An episode of rhythmic activity evoked in an immobilized *R. temporaria* tadpole by electrical stimulation of the skin; the transition from the generation of bursts of impulses to the generation of single impulses is accompanied by an increase in the frequency of the swimming rhythm. (a–c) Adapted with permission from Arshavsky YI, Orlovsky GN, Panchin YV, Roberts A, and Soffe SR (1993) Neuronal control of swimming locomotion: Analysis of the pteropod mollusc *Clione* and embryos of the amphibian *Xenopus. Trends in Neurosciences* 16: 227–233. (d) Adapted from Soffe SR (1989) Roles of glycinergic inhibition and *N*-methyl-D-aspartate receptor mediated excitation in the locomotor rhythmicity of one half of the *Xenopus* embryo central nervous system. *European Journal of Neuroscience* 1: 561–571. (e) Adapted from Sillar KT and Simmers AJ (1994) 5HT induces NMDA receptor-mediated intrinsic oscillators in embryonic amphibian spinal neurons. *Proceedings of the Royal Society of London, Series B* 255: 139–145. (f) Adapted from Sillar KT, Simmers AJ, and Wedderburn JF (1992) The post-embryonic development of cell properties and synaptic drive underlying locomotor rhythm generation in *Xenopus* larvae. *Proceedings of the Royal Society of London, Series B* 249: 65–70. (g) Adapted from Soffe SR and Sillar KT (1991) Patterns of synaptic drive to ventrally located spinal neurons in *Rana temporaria* embryos during rhythmic and non-rhythmic motor responses. *Journal of Experimental Biology* 156: 101–118.

Postural Neural Mechanisms in *Clione*

Normally, *Clione* maintains a vertical, head-up orientation (**Figures 1(a) and 4(a)**). Deviation from this orientation evokes postural correcting response. The tail is bent upward in the plane of deviation and the wing oscillations become asymmetrical, with the increased amplitude in the down-directed wing and retraction of the up-directed wing (**Figures 4(b) and 4(c)**, arrows). These postural reflexes lead to turning of the animal toward the vertical. The reflexes are driven by the gravity-sensing organs, the statocysts, that send signals to two types of cerebropedal

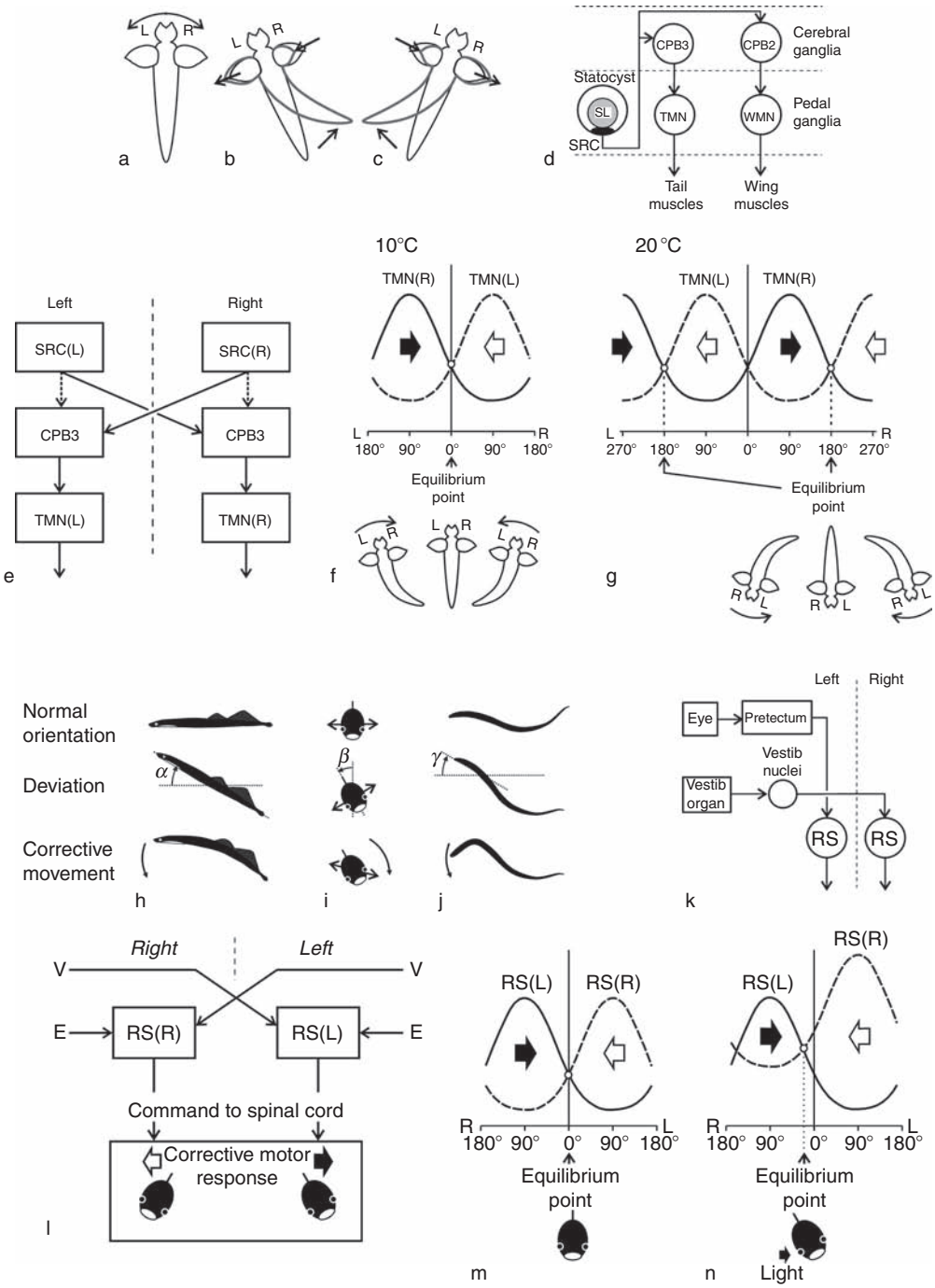

Figure 4 Control of spatial orientation in *Clione* (a–g) and lamprey (h–n). (a) The most common orientation of the swimming *Clione* is vertical, with the head up. L and R, the left and right aspects of the body, respectively. (b, c) Corrective motor responses to deviations from vertical in the frontal plane (arrows show directions of tail bending as well as increased activity of the down-directed wing and retraction of the up-directed wing). (d) The main groups of neurons involved in gravitational reflexes. CPB2 and CPB3, cerebropedal neurons; SL, statolith; SRC, statocyst receptor cell; TMN, tail MNs; WMN, wing MNs. (e) Neuronal network controlling *Clione* orientation in the frontal plane. Solid and interrupted lines are connections at lower and higher temperatures, respectively. (f and g) Operation of the network at lower temperature (f) and higher temperature (g). Due to a switch of connections from SRCs to CPB3 shown in d, the equilibrium point of the system changes from 0° (head-up orientation) at 10 °C (f) to 180° (head-down orientation) at 20 °C (g). The black and white arrows show directions of *Clione* rotation caused by the right and left TMNs, respectively. (h–j) Lamprey spatial orientation and its stabilization in the sagittal (pitch) plane (h; lateral view), transversal (roll) plane (i; frontal view), and horizontal (yaw) plane (j; dorsal view). Deviations from the stabilized orientation in these planes (angles α, β, and γ, respectively) evoke corrective motor responses (large arrows) aimed at restoration of the initial orientation. Symmetrical small arrows in (i) show directions of locomotor body undulations. (k) Main components of the brain stem orientation-controlling system. RS, reticulospinal neurons. (l) Organization of the roll-controlling network. Two groups of RS neurons, RS(R) and RS(L), receive inputs from the labyrinths (V) and eyes (E); they affect the spinal networks to evoke rolling of the

neurons, CPB3 and CPB2. The cerebropedal neurons, in turn, control the tail and wing muscle MNs (Figure 4(d)). After removal of the statocysts, *Clione* is not able to maintain any definite orientation in space.

The tail has the main role in stabilization of the body orientation. For example, deviations of the body in the frontal plane activate right or left statocyst receptor cells that excite the contralateral cerebropedal neurons (Figure 4(e), solid arrows). As a result, the contralateral tail MNs are excited and the tail bends up. The antagonistic gravitational reflexes driven by right and left statocyst receptors usually compensate for each other at the vertical head-up orientation (the equilibrium point of the control system in Figure 4(f)). Interneural interactions within the orientation-controlling network can change in relation to different behaviors. For example, the strength of the antagonistic gravitational reflexes can change asymmetrically. This will lead to a shift in the equilibrium point of the control system and, therefore, to a change in the swimming direction. This mechanism can be used during hunting behavior. A radical change in the interneural interaction occurs with an increase of water temperature. At higher temperature, statocyst receptors excite not the contralateral but, rather, the ipsilateral cerebropedal neurons (Figure 4(e), dashed arrows). This leads to a change in the equilibrium point in the system at 180° (from the head-up to the head-down orientation), which allows *Clione* to leave the superficial warm-water layers (Figure 4(g)).

Postural Neural Mechanisms in Lamprey

When swimming, the lamprey stabilizes the body orientation in the transversal (roll), sagittal (pitch), and horizontal (yaw) planes. Any deviation from the stabilized orientation causes corrective responses. In the pitch and yaw planes, the corrections occur due to the body bending in the corresponding plane (Figures 4(h) and 4(j)). In the roll plane, the corrections occur due to a change in the direction of locomotor body undulations from lateral to oblique (Figure 4(i)). These different types of motor responses are caused by four motor neuron pools in each spinal segment that innervate the dorsal and ventral muscles on the two sides.

The system for stabilization of orientation is mainly driven by vestibular input (Figure 4(k)). Deviation of the body from the stabilized orientation activates a specific group of vestibular afferents that, through neurons of vestibular (octavomotor) nuclei, affect contralateral reticulospinal (RS) neurons. These neurons, which constitute the main descending pathway in the lamprey, transmit commands for changing body orientation. In response to RS commands, the spinal cord forms output motor signals producing corrective movements (Figures 4(h)–4(j), bottom panels).

In addition to vestibular signals, RS neurons receive an ipsilateral visual input through pretectal neurons (Figure 4(k)). Visual signals exert a modulatory influence on RS neurons and thus can modify effects of the vestibular input. Natural sunlight assists in maintaining the normal dorsal-up orientation, whereas unilateral illumination causes a roll tilt toward the source of light (Figure 4(n)). Similar postural reactions to visual signals have been described in teleosts (the dorsal light response). Unlike the lamprey, some vestibuloectomized bony fish can maintain the dorsal-up orientation solely due to visual signals.

As in *Clione*, operations of the control systems, which stabilize orientation of the lamprey in each of the three planes, are based on the interactions between antagonistic vestibular reflexes mediated by two groups of RS neurons (Figures 4(l)–4(n)). The reflexes causing animal rotations in the roll plane are determined by the right and left RS neurons. Due to the vestibular input, the activity of the right and left RS neurons and, therefore, the activity of corresponding MNs depend on the animal orientation (Figure 4(m)). The antagonistic vestibular reflexes usually compensate for each other at the dorsal side-up orientation (the equilibrium point in Figure 4(m)). Unilateral eye illumination affects the two antagonistic groups of RS neurons differently and causes a shift in the equilibrium point of the system, which leads to a change in the stabilized orientation (Figure 4(n)). In the pitch control system, the stabilized orientation also corresponds to the equal activities in two bilateral antagonistic groups of RS neurons sensitive to the nose-up and nose-down tilt, respectively. The

lamprey. (m) Operation of the roll-controlling system when driven only by vestibular inputs. Ordinate, activity in RS(L) and RS(R) neurons; abscissa, angle of right and left roll tilting. Directions of rolling caused by activity of RS(L) and RS(R) neurons are indicated by the black and white arrows, respectively. The system has an equilibrium point at 0° (dorsal-side-up orientation). (n) Operation of the system when the right eye is illuminated. The visual input causes a shift of the equilibrium point to the left and the corresponding tilt of the animal. (a–d) Adapted with permission from Deliagina TG, Orlovsky GN, Selverston AI, and Arshavsky YI (1999) Neuronal mechanisms for the control of body orientation in Clione I: Spatial zones of activity of different neuron groups. *Journal of Neurophysiology* 82: 687–699. (e–g, l–n) Adapted with permission from Deliagina TG, Orlovsky GN, Zelenin PV, and Beloozerova IN (2006) Neural bases of postural control. *Physiology* 21: 216–225. (h–j) Adapted with permission from Zelenin PV, Orlovsky GN, and Deliagina TG (2007) Sensory-motor transformation by individual command neurons. *Journal of Neuroscience* 27: 1024–1032.

stabilized angle can be changed by increasing the water temperature, which affects the two groups differently and thus shifts the equilibrium point toward the nose-down orientation. Similarly, in the yaw control system, the stabilized orientation corresponds to the equal activity in the two antagonistic groups of RS neurons sensitive to the left and right lateral turn, respectively.

A control system for lamprey spatial orientation described previously at a functional level (Figures 4(h)–4(n)) has been analyzed in detail, with characterization of both sensory inputs and motor effects of individual RS neurons. A strong correlation has been found between vestibular inputs to individual RS neurons and effects on an efferent spinal output exerted by the same RS neurons. If a neuron is activated by a turn in a given plane, its motor effect results in a torque opposing a turn in this plane.

In conclusion, despite different structure of postural networks in the two evolutionary remote species (*Clione* and lamprey), the functioning of their orientation-controlling systems is based on the same principle – interaction between two antagonistic gravitational postural reflexes. Both animals stabilize the body orientation at which the antagonistic reflexes compensate each other. Both animals are able to gradually change the stabilized orientation through the change in the gain in one of the reflex chains. In addition, *Clione* can maintain two distinct orientations due to the reconfiguration of connections between elements of the postural network.

See also: Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Command Systems; Neuromodulation; Pattern Generation; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Startle Response; Swim Oscillator Networks; Vestibular System; Visual–Vestibular Interactions.

Further Reading

Arshavsky YI, Beloozerova IN, Orlovsky GN, Panchin YV, and Pavlova GA (1985) Control of locomotion in marine mollusc *Clione limacina*. 1: Efferent activity during actual and fictitious swimming. *Experimental Brain Research* 58: 255–262.

Arshavsky YI, Deliagina TG, Orlovsky GN, et al. (1986) Control of locomotion in marine mollusc *Clione limacina*. 6: Activity of isolated neurons of pedal ganglia. *Experimental Brain Research* 63: 106–112.

Arshavsky YI, Deliagina TG, Orlovsky GN, et al. (1998) Analysis of the central pattern generator for swimming in the mollusk *Clione*. *Annals of the New York Academy of Sciences* 860: 51–69.

Arshavsky YI, Orlovsky GN, Panchin YV, Roberts A, and Soffe SR (1993) Neuronal control of swimming locomotion: Analysis of the pteropod mollusc *Clione* and embryos of the amphibian *Xenopus*. *Trends in Neurosciences* 16: 227–233.

Brodfuehrer PD and Friesen WO (1986) Initiation of swimming activity by trigger neurons in the leech subesophageal ganglion. *Journal of Comparative Physiology A* 159: 489–502.

Brodfuehrer PD, Parker HJ, Burns A, and Berg M (1995) Regulation of the segmental swim-generating system by a pair of identified interneurons in the leech head ganglion. *Journal of Neurophysiology* 73: 983–992.

Brodfuehrer PD and Thorogood MS (2001) Identified neurons and leech swimming behavior. *Progress in Neurobiology* 63: 371–381.

Buchanan JT (2001) Contributions of identifiable neurons and neuron classes to lamprey vertebrate neurobiology. *Progress in Neurobiology* 63: 441–466.

Deliagina TG, Arshavsky YI, and Orlovsky GN (1998) Control of spatial orientation in a mollusc. *Nature* 393: 172–175.

Deliagina TG, Orlovsky GN, Selverston AI, and Arshavsky YI (1999) Neuronal mechanisms for the control of body orientation in Clione I: Spatial zones of activity of different neuron groups. *Journal of Neurophysiology* 82: 687–699.

Deliagina TG, Orlovsky GN, Zelenin PV, and Beloozerova IN (2006) Neural bases of postural control. *Physiology* 21: 216–225.

Eaton RC, Lee RK, and Foreman MB (2001) The Mauthner cell and other identified neurons of the brainstem escape network of fish. *Progress in Neurobiology* 63: 467–485.

Friesen WO, Poon M, and Stent GS (1976) An oscillatory neuronal circuit generating a locomotory rhythm. *Proceedings of the National Academy of Sciences of the United States of America* 73: 3734–3738.

Grillner S (2006) Biological pattern generation: The cellular and computational logic of networks in motion. *Neuron* 52: 751–766.

Kristan WB, Calabrese RL, and Friesen WO (2005) Neuronal control of leech behavior. *Progress in Neurobiology* 76: 279–327.

Kristan WB and Guthrie PB (1977) Acquisition of swimming behavior in chronically isolated single segments of the leech. *Brain Research* 131: 191–195.

Mackie GO (2004) Central neural circuitry in the jellyfish Aglantha: A model 'simple nervous system.' *Neurosignals* 13: 5–19.

Orlovsky GN, Deliagina TG, and Grillner S (1999) *Neuronal Control of Locomotion. From Mollusc to Man*. Oxford: Oxford University Press.

Panchin YV, Arshavsky YI, Deliagina TG, et al. (1996) Control of locomotion in the marine mollusk *Clione limacina*. 11: Effects of serotonin. *Experimental Brain Research* 109: 361–365.

Panchin YV, Sadreev RI, and Arshavsky YI (1995) Control of locomotion in the marine mollusk *Clione limacina*. 10: Effects of acetylcholine antagonists. *Experimental Brain Research* 106: 135–144.

Roberts A, Soffe SR, Wolf ES, Yoshida M, and Zhao FY (1998) Central circuits controlling locomotion in young frog tadpoles. *Annals of the New York Academy of Sciences* 860: 19–34.

Satterlie RA (2002) Neuronal control of swimming in jellyfish: A comparative story. *Canadian Journal of Zoology* 80: 1654–1669.

Shaw BK and Kristan WB (1977) The neuronal basis of behavioral choice between swimming and shortening in the leech. *Journal of Neuroscience* 17: 786–795.

Sillar KT, Simmers AJ, and Wedderburn JF (1992) The post-embryonic development of cell properties and synaptic drive

underlying locomotor rhythm generation in *Xenopus* larvae. *Proceedings of the Royal Society of London, Series B* 249: 65–70.

Sillar KT and Simmers AJ (1994) 5HT induces NMDA receptor-mediated intrinsic oscillators in embryonic amphibian spinal neurons. *Proceedings of the Royal Society of London, Series B* 255: 139–145.

Soffe SR (1989) Roles of glycinergic inhibition and *N*-methyl-D-aspartate receptor mediated excitation in the locomotor rhythmicity of one half of the *Xenopus* embryo central nervous system. *European Journal of Neuroscience* 1: 561–571.

Soffe SR and Sillar KT (1991) Patterns of synaptic drive to ventrally located spinal neurons in *Rana temporaria* embryos during rhythmic and non-rhythmic motor responses. *Journal of Experimental Biology* 156: 101–118.

Zelenin PV, Orlovsky GN, and Deliagina TG (2007) Sensory-motor transformation by individual command neurons. *Journal of Neuroscience* 31: 1024–1032.

Sympathetic Nervous System

A Dahlström, Göteborg University, Goteborg, Sweden

© 2009 Elsevier Ltd. All rights reserved.

Introduction: Historical Overview

The peripheral autonomic nervous system was classically divided into two systems, sympathetic and parasympathetic, on a physiological basis. In 1904 TR Elliott, experimenting on frogs, suggested that adrenaline (epinephrine; E) was the chemical substance that mediated the effect of nerve impulses to the effector cells from sympathetic nerves. After the concept of chemical neurotransmission was accepted, these nerves were named 'adrenergic' by Sir Henry Dale in 1933, the suffix-ergic meaning 'using as transmitter.' In mammals the peripheral sympathetic nerves use noradrenaline (norepinephrine; NE) for transmitter instead of E, as demonstrated by Ulf von Euler in 1946. It is, however, important to stress that 'sympathetic' is not synonymous with 'adrenergic,' since sympathetic nerves use adenosine 5′-triphosphate (ATP) and neuropeptide Y (NPY) as cotransmitters and some use acetylcholine (ACh) as transmitter. Thus, peripheral adrenergic nerves contain and release NE in mammals (E in amphibians), but in the central nervous system (CNS) NE-containing as well as E-containing neurons exist; the terms 'adrenergic transmission' and 'noradrenergic transmission' therefore have to be used when dealing with the CNS.

The Peripheral Sympathetic System Is Composed of Two Neuron Chains

The sympathetic system in mammals is structured with two neurons connected in series. The first neuron (preganglionic), located in the intermediolateral gray column of the spinal cord, exits from the CNS at spinal cord levels Th1–L3–4 via the anterior roots. The preganglionic axons then deviate from the root via the ramus communicans alba to reach the sympathetic ganglia in the paravertebral chain of ganglia. The axons of this first neuron make synaptic contacts with the cell body of the second neuron in the peripheral sympathetic ganglia, named the postganglionic neuron. The postganglionic axons, which are thin and unmyelinated, return to the main root and accompany the peripheral nerves to the effector sites. The most cranial sympathetic ganglia of this paired chain of ganglia, the superior and middle cervical ganglia, and the stellate ganglia instead send their postganglionic axons directly to the head, neck, and cardiac tissues via thin adrenergic axon bundles. While the preganglionic neuron uses ACh as a classical transmitter in the synapses of sympathetic ganglia, the postganglionic neuron in most cases uses NE, ATP, and NPY as cotransmitters at its very widely ramified nerve terminal varicose plexus.

Structure of the Sympathetic Postganglionic Neuron

Von Euler's studies, published in 1946, demonstrated that tissue homogenates from sympathetic ganglia and sympathetically innervated tissues contained a substance that pharmacologically was identified as NE. However, the exact cellular localization of this catecholamine was not known. Until that time peripheral innervation of tissues was thought to be mediated by a 'syncytium,' since no microscopically identified structures that could be connected to a transmitter function had been observed until then. However, in 1946 Nils-Åke Hillarp was able to present microscopic pictures of a widely ramified network of thin, unmyelinated fibers stained with methylene blue (MB), using a special modification of the then generally practiced MB staining protocol (Figure 1). Also, sympathetic ganglia were found to contain such nets of varicose fibers (Figure 2). Hillarp, in his thesis of 1948, presented strong arguments that this net of varicose fibers was the morphological component that mediated transmitter effects.

The full structure of the sympathetic neuron was clarified in the 1960s, when Hillarp and co-workers succeeded in developing a histochemical fluorescence method that made it possible, for the first time in scientific history, to exactly pinpoint the cellular and intracellular localization of a transmitter. This histochemical fluorescence method used gaseous paraformaldehyde to initiate a ring closure of the side arm of the NE molecule, resulting in a strongly fluorescent isoquinoline. After using this method on an air-dried preparation of rat iris, a brilliantly green fluorescent varicose network of NE-containing sympathetic nerve terminals could be seen in the fluorescence microscope (Figure 3). This method could also be used to demonstrate the other catecholamines (dopamine (DA) and E) with a similar fluorescence spectra, as well as serotonin (5-HT), whose reaction product emitted a strong yellow fluorescence. This method was subsequently used for all kinds of tissues, and was the basis for the first mapping of the brain neuronal pathways that use NE, DA, or 5-HT as a transmitter. These studies were carried out during 1962–69 at the

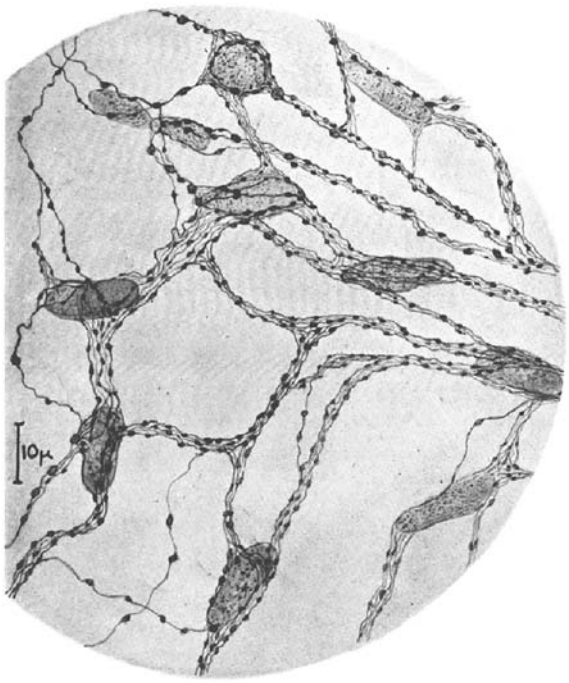

Figure 1 Net of varicose nerve terminals in spread preparated rat tissue, stained with methylene blue. Small varicosities are seen along terminal branches which intermingle and anastomose in bundles. Large oval structures are Schwann cell nuclei, but the tubular sheaths of Schwann cell cytoplasm, covering the terminal branches, cannot be seen. From Hillarp (1946).

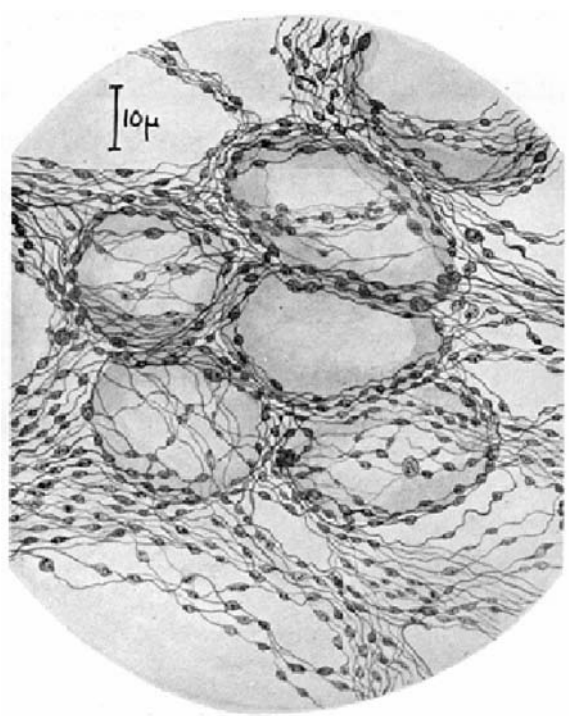

Figure 2 Net of varicose nerve terminal branches penetrating in between ganglion cells, forming baskets around cell bodies. Methylene blue stain. From Hillarp (1946).

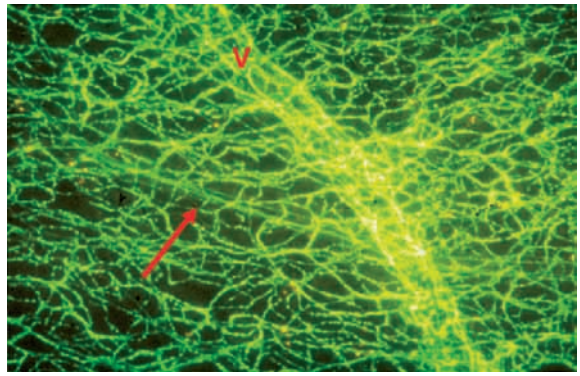

Figure 3 Spread preparation of rat iris, exposed to formaldehyde vapor after air drying. Fluorescence micrograph demonstrating the dense net of green fluorescent nerve terminal branches with strongly fluorescent varicosities. V denotes a blood vessel with especially dense innervation of the fluorescent NE-containing nerve terminals. Arrow points to a bundle of preterminal axons with low fluorescence intensity, which have not reached their area of innervation. From Dahlström A (1971) Axoplasmic transport (with particular respect to adrenergic neurons). *Philosophical Transactions of the Royal Society of London B* 261: 325–358.

Figure 4 Section of a sympathetic ganglion demonstrating a postganglionic cell body with many branching dendrites. From Hillarp (1948).

Department of Histology, Karolinska Institute in Stockholm, Sweden.

Using this fluorescence method the full extension of the postganglionic sympathetic neuron was clarified. In the sympathetic ganglia the multipolar nerve cell bodies with many long dendritic neurites (Figure 4)

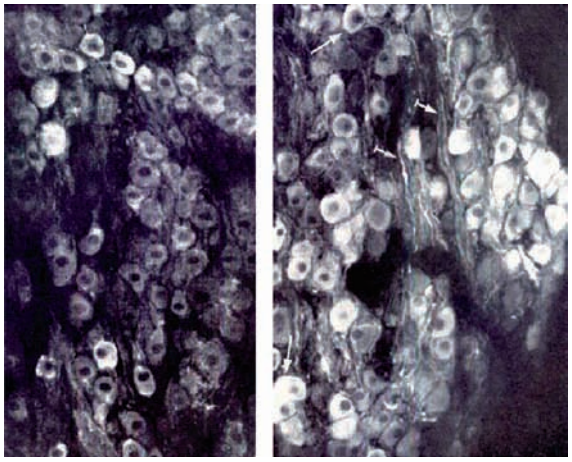

Figure 5 Fluorescence micrographs from rat sympathetic ganglia. To the left, from a normal rat. Cell bodies contain varying intensities of NE-induced fluorescence, leaving the nucleus nonfluorescent. To the right, a section from a ganglion treated with colchicine, which disrupts microtubules and blocks the export from the cell body of NE storing granules. Markedly increased fluorescence in enlarged cell bodies indicate block in export but continued production. Small arrows indicate swollen postganglionic axons with accumulations of fluorescent NE, due to interrupted axonal transport. From Dahlström (1968).

were found to contain varying levels of the fluorescent NE product, located in the cytoplasm and extending into the neuritic processes (Figure 5). Using the substance reserpine, known earlier to deplete tissues of catecholamines, it was found that the fluorescence was completely abolished, suggesting that the NE in ganglion cells and in nerve terminals was stored in some subcellular compartment which was sensitive to reserpine. Similar depletion of catecholamines was found around 1960 to occur in adrenal glands after reserpine treatment, and this was shown to be due to the blockade of a norepinephrine storage mechanism in subcellular granules, which could be isolated by centrifugation procedures. The storage mechanism for catecholamines in the adrenergic nerves was also reserpine sensitive, and it was thought in the 1960s to be located in the small synaptic vesicles observed by Eccles and others in the electron microscope (EM).

Hökfelt and co-workers, using a modified $KMnO_4$ fixation technique, demonstrated in 1968 that small synaptic vesicles in the ganglion cells and in the nerve terminals were the NE storage organelles for the sympathetic nerves. In normal tissues the vesicles in adrenergic structures contained a small electron-dense core, which was no longer observable after reserpine treatment. With this fixation protocol other nerve terminal structures were found to contain small clear vesicles without a dense core, presumably representing ACh-containing vesicles (Figure 6).

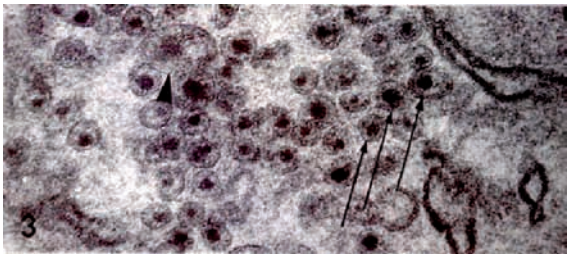

Figure 6 Electron micrograph of adrenergic nerve terminal, fixed in potassium permanganate ($KMnO_4$). A group of granular/vesicular structures can be seen, most of which carry an electron-dense precipitate in the matrix (arrows). From Hökfelt (1968).

The Sympathetic Neuron Has a Widespread Innervation Pattern

With combinations of quantitative biochemical and microscopic methods, the full extension of a sympathetic neuron in rat and cat tissues was investigated (Figure 7). The total NE content of rat irides, as well as a fixed length of rat vas deferens, was measured with the trihydroxyindole (THI) method. Then the total number of fluorescent varicosities in the irides was counted manually in stretch preparations. The lengths of vas deferens were sectioned and the varicosities were counted manually. Since the nonterminal axons contained very weak fluorescence (see arrows in Figure 3) and thus minute amounts of NE, the vast majority of NE would be present in the small, intensely fluorescent varicosities with a diameter of 0.5–1 μm. It was calculated that each varicosity of the sympathetic innervation in the rat iris and vas deferens, and in the cat head structures, contained in the order of 0.005 pg of NE.

The superior cervical ganglia (SCG) give rise to the sympathetic innervation in head and neck structures. In order to measure the NE content in the terminals emanating from this ganglion, all soft tissues of the cat head structures were dissected (in the cold). Five heads were used, and the total NE content in the tissues of each cat's head was found to be about 32 μg of NE. The histochemical formaldehyde fluorescence method demonstrated a linear relationship between NE concentration and fluorescence intensity in model experiments in concentrations well within physiological limits. Since, with the fluorescence microscope, the fluorescence intensity of nerve terminals in the cat was the same as in the rat, we could assume that the content/concentration of the rat and cat tissue varicosities was about the same. The number of nerve cells in a cat's SCG was carefully counted by Billingsley and Ranson in 1918. The majority of the ganglion cells (247 000, as indicated using the microscope) contained 104 ng of NE per pair. The SCG gave rise to a nerve terminal area containing

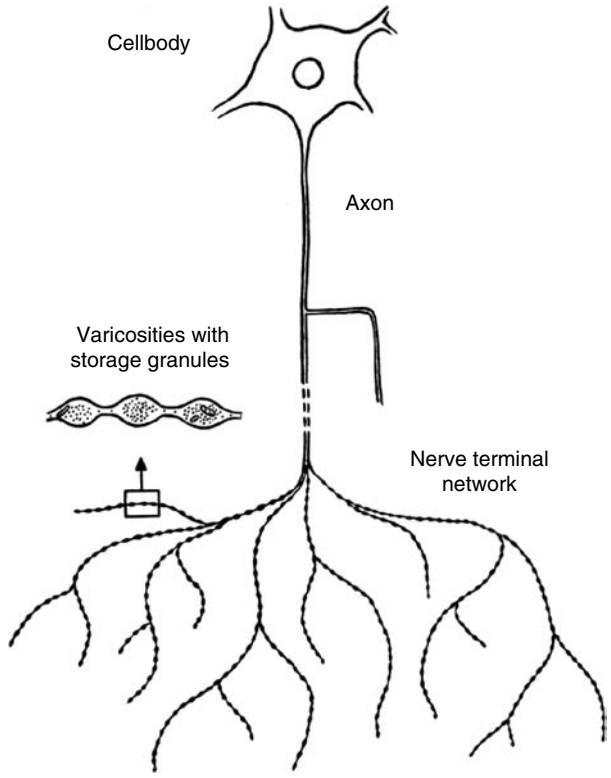

Figure 7 Schematic drawing illustrating the morphology of a postganglionic sympathetic neuron. From a multipolar cell body, a thin, unmyelinated axon, which can give off collaterals, passes to the area of innervation, where an extensive branching takes place. The peripheral nerve terminal branches carry presynaptic structures, varicosities, which are filled with transmitter organelles. To the left, three varicosities are drawn enlarged to illustrate the packed transmitter organelles in the varicose swelling, while the segments between varicosirites are very thin and devoid of transmitter granules. From Dahlström A (1971) Axoplasmic transport (with particular respect to adrenergic neurons). *Philosophical Transactions of the Royal Society of London B* 261: 325–358.

32 ng of NE. If each varicosity contained 0.005 pg of NE, each ganglion cell, on average, would gave rise to about 30 000 varicosities, corresponding to a total length of the added branches of the nerve terminal tree of about 15 cm. (Later in the 1970s it was reported that a mature SCG of rat contains approximately 30 000 principal ganglion cells.)

This historical work demonstrates the divergence of the sympathetic innervation pattern. At the same time, however, the innervation pattern is converging, since the effector cells (smooth muscle cells, gland cells, cardiac muscle cells, etc.) are usually reached by transmitter release from several neurons. The nerve terminals arrange themselves in a basketlike arrangement near, for instance, salivary gland acini, with a net of intermingling terminals from different neurons, both sympathetic and cholinergic. The ramifications of this intermingled net are coated by thin, fenestrated layers of Schwann cell cytoplasm. This arrangement, observed by Hillarp in his MB preparations, was given the name 'the autonomic ground plexus.'

Also in the ganglia the innervation from the preganglionic neurons is both diverging and converging. As pointed out by Burnstock (see **Figure 8**), each sympathetic ganglion cell receives innervation from many preganglionic axons (convergence), which in turn branch (divergence) to innervate many postganglionic neurons. This organization, together with the fact that the preganglionic axons can ascend or descend to other levels in the sympathetic chain before synapsing in the ganglion, is the reason for the slow, general activation of this emergency system. The physiological effect is one of slowly (some seconds) engaging the whole organism in the reaction. The sympathetic system is normally silent and is called into action during alarming situations. The release of E and NE from the suprarenal medulla is the quickest route by which the organism is alerted, the other parts of the sympathetic system becoming enrolled some seconds later.

NE Storage Organelles Contain Specific Matrix and Vesicle Membrane Components

The matrix of the NE storage organelles contains ATP and NE in a complex similar to that in the adrenal medullary granules. The population of organelles is not uniform in size or composition; both small vesicles (50 nm) and larger (<80 nm) NE-containing organelles ('granules') are observed with the EM. Usually the fraction of large organelles is only about 2–5% of the whole vesicle population in each varicosity, but this varies in different organs. These larger organelles also contain dopamine β-hydroxylase (DBH), which converts DA to NE, and have been proposed to represent the newly arrived, by fast axonal transport, organelles. With each round of exocytotic release, as suggested by Dahlström and Häggendal in 1982, the DBH-containing organelles lose part of the matrix, and the retrieved membrane is used to form the small NE-containing vesicles, which is the fraction that stores most of the NE in the nerve terminals. Studies using reserpine, which causes an irreversible block of NE storage in sympathetic nerves, indicate that the newly formed (after reserpine clearance) NE-storing organelles, recently arrived at the terminals by fast axonal transport, are the main 'NE-releasers,' and that only a few percent of the normal NE content in the nerve terminals is needed for full stimulation-induced effect. It was proposed that this small percentage of NE, sufficient for stimulation response after reserpine, is present in the newly transported granules. Since DBH in the large granules is needed for the NE

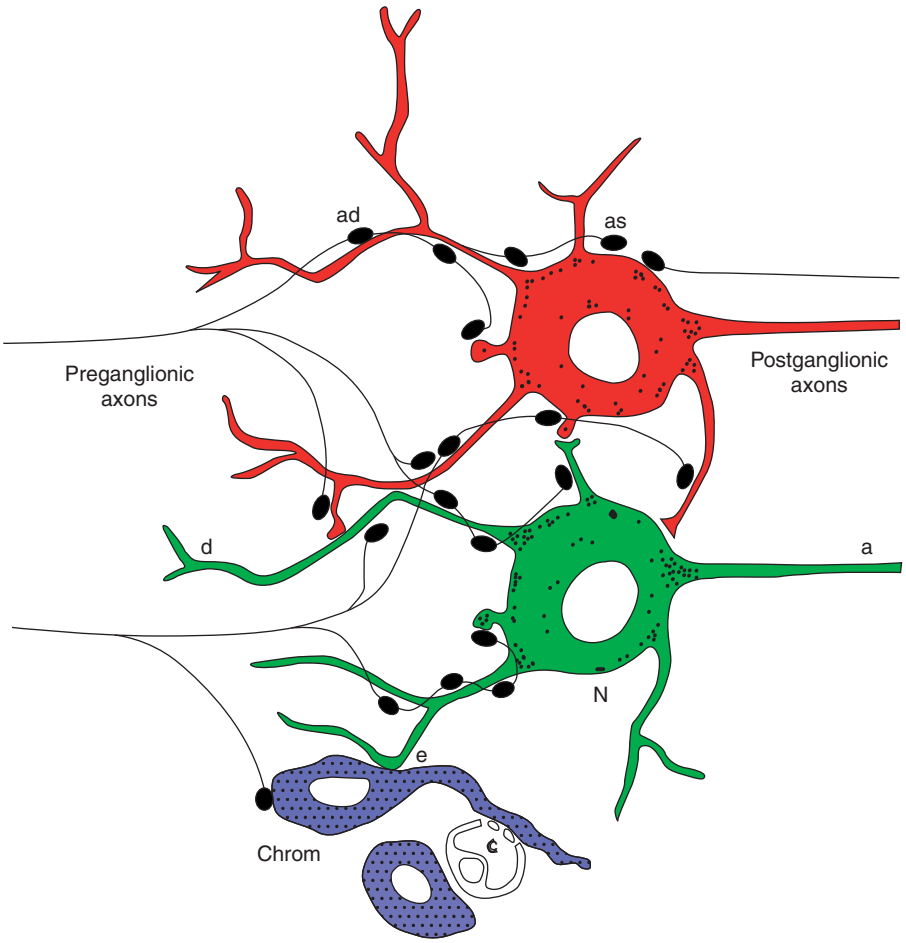

Figure 8 Schematic drawing of two postganglionic, sympathetic, ganglion cells which are innervated by two preganglionic axons seen to the left. Both axons branch and form varicose boutons (contacts) on somas and dendrites of both the green and the red ganglion cell, exemplifying 'diverging' innervation. The upper axon contacts both the green and the red neuron. Since the green cell is contacted by both axons and the red cell in innervated by both axons, this represents 'convergence' in innervation. The two blue structures are SIF-cells which are innervated and also lie close to blood vessels, creating a basis for release of chemicals (e.g., DA) into the circulation. From Burnstock G and Costa M (1975) *Adrenergic Neurons*. London: Chapman & Hall.

synthesis in the nerve terminal varicosities, it is obvious that the organelle turnover in sympathetic neurons differs mechanistically from that in cholinergic nerve terminals, where the ACh is synthesized by choline acetyltransferase (ChAT) in the cytosol before being taken up into the synaptic ACh-releasing vesicles by the vesicular ACh transporter (VACT).

The membranes of synaptic vesicles in the sympathetic neuron contain many identified integral proteins – for instance, synaptophysin, synaptic vesicle protein 2 (SV2), synaptotagmin I, synaptobrevin II, and the NE transporter, vesicular monoamine transporter 2 (VMAT2). In addition to the integral membrane proteins, many proteins have been found attached to the surface of the NE-containing vesicles, including synapsin, Rab3a, and clathrin light chain (LC). It is also likely that the enzyme tyrosine hydroxylase (TH), which is present in the soluble fraction when sympathetically innervated tissues are homogenized, attaches to transport organelles when exported from the cell body. This enzyme is transported intra-axonally at fast rates, which would not be the case if it were in the cytoplasmic phase during transport distally.

Peptides and Transmitter Presence Vary in Different Ganglia and in Different Species

In addition to the 'classical transmitter' NE, other small transmitter molecules are present in sympathetic ganglia. Generally, the enzyme TH, which is the rate-limiting step in catecholamine synthesis (a marker for sympathetic structures), is present in the majority of neurons in sympathetic ganglia. Some neurons, however, appear to be cholinergic in nature, since they instead express the enzyme ChAT in adult

animals. In many cases, neurons, which are nonadrenergic (TH-negative) and noncholinergic (VACT- and ChAT-negative), are present. These may contain histamine, but the presence of γ-aminobutyric acid (GABA) ergic neurons has also been reported. These GABA-containing neurons form basketlike nerve terminal connections around other ganglion cells, possibly representing an inhibitory intrinsic system in the ganglia.

Peptides like vasoactive intestinal peptide (VIP), NPY, galanin, enkephalin, somatostatin, and chromogranins may be present together with TH in sympathetic neurons. However, the amine and the peptides are stored in separate subcellular compartments. NPY and other peptides are seen to be stored in larger granules with an electron-dense matrix in regular EM preparations (in contrast to the NE in small synaptic vesicles, which requires fixation in $KMnO_4$ for visibility). NPY has been shown to be expressed in sympathetic neurons that innervate blood vessels, and somatostatin is present in some cells within the prevertebral ganglia, innervating the gut. Interestingly, combined retrograde tracing with triple immunohistochemistry has shown that the target tissue appears to determine which peptide is coexpressed with TH in the ganglion cells. Members of the family of chromogranins are frequently expressed in the sympathetic neurons. Recently the newly identified chromogranin NESP55 was demonstrated to be stored in a subpopulation of ganglion cells, concentrated to the Golgi region. As seen in Figure 9 the NESP55-positive cells were mainly scattered, except for in the thoracic chain ganglia, where groups of NESP55-containing neurons were observed. In only a small percentage, the peptide was colocalized with TH immunoreactivity. Recently it was reported that mouse sympathetic ganglia contain a population of neurons expressing one of the seven identified zinc transporters, ZnT3, co-localized with TH but never with NPY (Figure 10).

Thus, the presence and distribution of peptides vary in different species and in the various ganglia, and the expression is regulated most probably mainly by the target tissue receiving innervation from an individual ganglion cell. Some peptides are not visualized in control ganglia but are upregulated in response to axotomy or denervation. A dense pericellular net of varicose nerve terminals is present in sympathetic ganglia (Figure 2). Most of these fibers represent cholinergic preganglionic terminals storing different peptides – for example, enkephalin. Nerve terminals containing the peptide substance P also originate from sensory dorsal root ganglia. In prevertebral

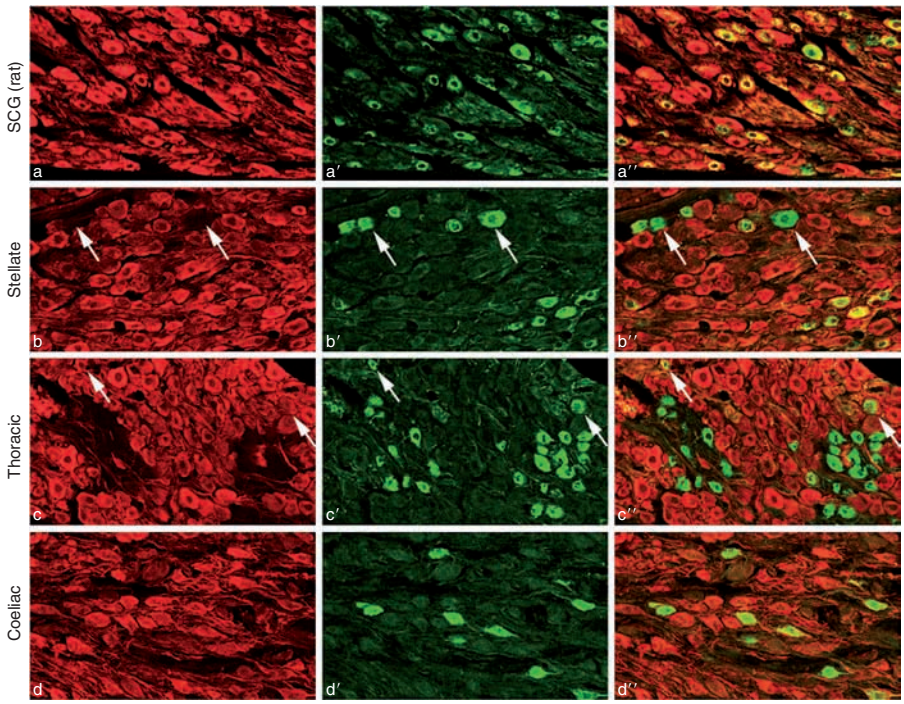

Figure 9 Rat sympathetic ganglia (label to the left) incubated with anti-TH (red) abd Anti-NESP55 (green). Most ganglion cells contain TH-immunoreactive material while only a small number of cells contain green NESP55-like material. In thoracic ganglia the NESP55-containing cells are often seen in groups, while in other ganglia the green cells are scattered. Some TH-positive cells also contain the NESP55-peptide (yellow cells in the right lane indicating co-localization) while some NESP55 containing cells appear devoid of TH (arrows in lane 2, stellate ganglion). These sympathetic ganglion cells are a heterogenous population as regards transmitter-related substances. From Li and Dahlström (2007).

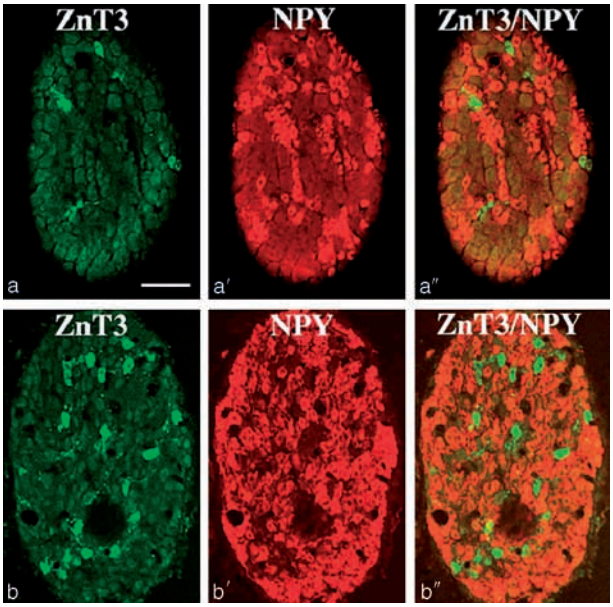

Figure 10 Sections of mouse sympathetic ganglia incubated with anti-ZnT3 (green) and anti-NPY (red). Many NPY-positive cells and a few ZnT3-containing cells are seen, and there is no indication of co-localization. From Wang et al. (2003).

ganglia, nerve terminals with bombesin-like material are also observed, probably representing a reciprocal innervation from the intramural ganglia in the gut.

The sympathetic neuron in normal adult animals expresses high levels of the growth-associated protein 43 (GAP43), observable with immunohistochemistry in all parts of the neuron. This protein is expressed in other types of neurons mainly as a response to trauma. The high level of GAP43 in sympathetic neurons is likely to reflect the high degree of plasticity of sympathetic neurons.

Intraneuronal Events in the Sympathetic Neuron

In the sympathetic neuron, as in other neurons, all materials necessary for the function of the neuron are produced in the cell body where the nucleus is located. The first observations that demonstrated a fast, distally directed, intraneuronal transport of NE were published in 1965. The rate of accumulation of material proximal to a peripheral nerve crush clearly indicated a rate of 5–10 mm h^{-1} in the sciatic nerve adrenergic axons of rat and cat, respectively. This figure was based on the rate of accumulation proximal to a crush/ligation, but has since been verified in isotope experiments using labeled amino acids incorporated into transported protein.

Later studies by Jia-Yi Li and co-workers, between 1994 and 1997, using immunofluorescence demonstrated a fast anterograde transport of the integral membrane proteins, of matrix peptides, and of surface-associated peptides. It is interesting to note that the amount of retrogradely transported peptides varies between these three classes of organelle components; the integral proteins showed the highest proportion of retrograde transport (about 50–80%, suggesting a recycling of membrane), about 20–30% of matrix peptides were found to recycle, while the surface-adsorbed peptides showed the lowest degree of retrograde transport. These figures probably mirror dynamic events in the nerve terminals; surface-attached proteins detach from vesicles in the nerve endings, matrix peptides are partially released via exocytosis during transmitter release, while integral membrane proteins remain in the membrane and are fully recycled.

Embryology/Development

The autonomic ganglia emanate from the neural crest, like the sensory ganglia and the innervation of the gut. The migration of the neural crest cells is dependent upon chemotaxis, and nerve growth factor (NGF) has been shown to be a prerequisite for the development of sympathetic and sensory innervation. Interestingly, during embryonic development, future sympathetic ganglion cells are initially cholinergic in nature, but alter the phenotype upon maturation. Thus, initially immunoreactive for ChAT, most ganglion cells downregulate the expression of ChAT and instead express TH, the catecholamine-synthesizing enzyme. However, a small proportion of the ganglion cells usually remain cholinergic in the adult animal, as demonstrated by high levels of acetylcholinesterase (AChE) and ChAT, as well as immunopositivity of VACT, the integral membrane protein that transports ACh into the synaptic cholinergic vesicles.

For the development of sympathetic as well as of sensory ganglia the neurotrophin NGF is essential. This peptide couples to two receptors, TrkA and p75; p75 is a common receptor for many neurotrophins, but for the trophic effect of NGF, both the specific TrkA receptor and p75 must be activated. NGF is thought to be supplied by the target area cells, which, thus, attract the outgrowing sympathetic fibers by secreting NGF. The NGF is internalized at the tip of growth cones and transported by retrograde axonal transport to the cell body, where it influences gene transcription and supports outgrowth. By administrating an NGF antiserum to mammals, sympathetic denervation may be carried out, as first demonstrated by Levi-Montalcini and Booker in 1960. Recently it was reported that sympathetic ganglion cells can

synthesize and secrete pro-NGF, indicating the possibility for autocrine stimulation of NGF.

In some lumbar sympathetic ganglia, neurons that innervate the sweat glands of the lower extremities have remained cholinergic, and in addition coexpress the peptide VIP, the vasodilatory peptide, and in some species also the calcitonin gene-related peptide. When stimulated, these sympathetic cholinergic neurons release ACh, which triggers the gland cell receptors to induce secretion, while the coreleased VIP dilates the surrounding blood vessels, thereby ensuring access of the sweat 'precursor,' serum.

Small, Intensely Fluorescent Cells

In addition to the principal ganglion cells, sympathetic ganglia contain so-called small, intensely fluorescent (SIF) cells. These small cells are usually distributed along blood vessels, have short processes, and contain in most cases high concentrations of DA, giving them a strong green fluorescence using the classical paraformaldehyde histochemical method (see previous section). They also may contain serotonin or histamine, usually together with a peptide (e.g., substance P). The catecholamine storage granules and the principal ganglion cells contain the same integral membrane proteins, including VMAT2. The SIF cells originate in the neural crest, but seem to have a somewhat different developmental differentiation scheme compared with principal ganglion cells.

Glia, Satellite Cells

In addition to principal ganglion neurons and SIF cells, a third cell type, called satellite cells, is present, usually located close to the cell body of the multipolar ganglion cells (Figure 11). These cells represent a type of glia cell, but are separate from Schwann cells in the periphery, and from astrocytes and oligodendroglia in the CNS. They do not myelinate, since the sympathetic ganglion cells have unmyelinated thin axons, less than 1 μm in diameter. Satellite cells express S-100 protein but not glial fibrillary acidic protein, and neuropeptides have not been observed to be present in these cells. Satellite cells are probably of importance for the metabolism of the ganglion cells, and *in vitro* studies indicate that they support the outgrowth of dendrites from the perikarya of ganglion cells, possibly by expressing neurotrophic factors. Bone-derived neurotrophic factor immunoreactivity has been reported to occur in SCG of rat, but so far no other trophic factor has been observed. Comparatively little is known at present about the function of satellite cells. Connective tissue and blood vessels are naturally present in sympathetic ganglia; vascular endothelial cells may contribute to the functional complexity of sympathetic ganglia by secreting nitric oxide (NO), for instance.

NO Synthesis in Ganglia

Many studies implicate NO in a multitude of processes in the nervous system. The sympathetic ganglia contain considerable levels of NO synthase (NOS) and NADPH activity. Many studies indicate that most of the NOS appears to be linked to the preganglionic neuron, and, thus, is present mainly in the preganglionic nerve terminals in the ganglia. However, nitrergic nerve cell bodies have been reported in dog SCG, but were not found in rat ganglia. Most likely, the vascular endothelium in sympathetic ganglia also expresses inducible NOS (iNOS), which may be involved in the high degree of plasticity typical for the sympathetic system.

Sex Differences in Sympathetic Ganglia

In the rat, sympathetic ganglia are generally smaller in the female than in the male. This may be explained by the effect of perinatal testosterone in the male, which inhibits the postnatally occurring apoptotic processes, thus reducing the number of ganglionic cells in the female. This sexual dimorphism is not only seen in pelvic ganglia, where a difference may seem obvious, but also in the SCG, where the difference in cell number is found mainly in the population of cells innervating the salivary glands, not in those innervating, for instance, the iris.

The influence of these differences on performance has not been studied in detail, but possibly it may explain a lower physical endurance in females than in males, in line with the higher level of offensive aggression in males, relative to females, in most rat

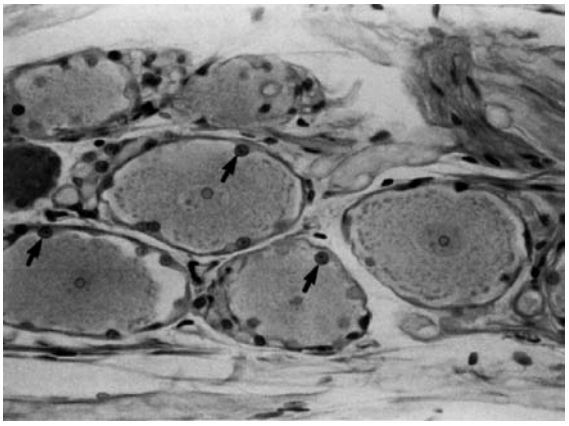

Figure 11 Routine stained section of sympathetic ganglion of rat, showing large ganglion cells with a few satellite cells clinging to the cell body (arrows). From Hillarp (1949).

strains. Recently it was reported that estrogen alters the expression of TrkA and p75 neurotrophin receptors in sympathetic neurons, probably contributing to sexual differences in functions of the sympathetic system and influencing the survival and function of ganglion cells.

See also: Adenosine Triphosphate (ATP); Adenosine Triphosphate (ATP) as a Neurotransmitter; Autonomic Nervous System; Cotransmission; Norepinephrine: Adrenergic Receptors; Pain and the Sympathetic Nervous System; Sensory Ganglia.

Further Reading

Billingsley PR and Ranson AW (1918) On the number of nerve cells in the ganglion cervicale superius and of the nerve fibres in the cephalic end of the truncus sympaticus in the cat and on the numerical relations of preganglionic and postganglionic neurons. *Journal of Comparative Neurology* 29: 359–366.

Burnstock G and Costa M (1975) *Adrenergic Neurons*. London: Chapman & Hall.

Dahlström A (1965) Observations on the accumulation of noradrenaline in the proximal and distal parts of peripheral adrenergic nerves after compression. *Journal of Anatomy (London)* 99: 677–689.

Dahlström A (1971) Axoplasmic transport (with particular respect to adrenergic neurons). *Philosophical Transactions of the Royal Society of London B* 261: 325–358.

Dahlström A (1982) Axoplasmic transport and synaptic transmission. In: Weiss DG and Gorio A (eds.) *Axoplasmic Transport in Physiology and Pathology*, pp. 16–20. Berlin-Heidelberg: Springer .

Dahlström A (1996) Nils-Åke Hillarp – the life of a great scientist. In: Fuxe K, Hökfelt T, Olson L, et al. (eds.) *Molecular Mechanisms of Neuronal Communication. A Tribute to Nils-Åke Hillarp*, pp. 15–39. Oxford: Pergamon.

Dahlström A and Carlsson A (1986) Making visible the invisible. In: Parnham MJ and Bruinvels J (eds.) *Discoveries in Pharmacology*, vol. 3, pp. 97–125. Amsterdam: Elsevier.

Eränkö O (1976) *SIF-Cells: Structure and Function of Small, Intensely Fluorescent Sympathetic Cells*. Washington DC: US Government Printing Office (DHEW).

Gibbins I (1995) Chemical neuroanatomy of sympathetic ganglia. In: McLachlan EM (ed.) *Autonomic Ganglia*, pp. 73–121. Switzerland: Harwood Academic Publishers.

Häggendal J and Dahlström A (1971) The importance of axoplasmic transport of amine granules for the functions of adrenergic neurons. *Acta Neuropathologica* V: 238–248.

Hökfelt T (1971) Ultrastructural localization of intraneuronal monoamines. Some aspects on methodology. *Progress in Brain Research* 34: 213–222.

Levi-Montalcini R and Booker B (1960) Destruction of the sympathetic ganglia in mammals by an antiserum to a nerve growth protein. *Proceedings of the National Academy of Sciences of the United States of America* 46: 384–391.

McLachlan EM (1995) *Autonomic Ganglia*. Switzerland: Harwood Academic Publishers.

Schultzberg M (1983) The peripheral nervous system. In: Emson P (ed.) *Chemical Neuroanatomy*, pp. 1–51. New York: Raven Press.

Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress

D S Goldstein, National Institutes of Health, Bethesda, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

The body has three peripheral catecholamine systems, corresponding to the three endogenous catecholamines – norepinephrine, epinephrine (synonymous with adrenaline), and dopamine. Each system works differently. Norepinephrine is the main neurotransmitter in sympathetic nervous system regulation of the circulation. In contrast, in the sympathetic cholinergic system, acetylcholine is the neurotransmitter stimulating secretion by sweat glands. In humans, epinephrine is the main hormone of the adrenomedullary hormonal system. Outside the brain, dopamine seems to function as an autocrine–paracrine substance, such as in the kidneys. This article is about the roles of the sympathetic noradrenergic and adrenomedullary hormonal systems in stress and distress. Understanding these roles requires discussion of several concepts – the sympathoadrenal system, stress as a scientific idea, characteristics of distress, allostasis and allostatic load, and scientific integrative medicine. Most of this article is a brief overview of these concepts.

The Sympathoadrenal System

Walter B Cannon proposed that the sympathetic nervous system and the adrenal gland work together as a unit to maintain homeostasis in emergencies – the sympathoadrenal (or sympathoadrenomedullary or sympathico-adrenal) system. This probably was the first suggestion of the existence of an integrated nervous-hormone, or neuroendocrine, system. The notion of a unitary sympathoadrenal system continues in medical thinking, despite evidence for differential changes in sympathetic noradrenergic and adrenomedullary hormonal activities in different forms of stress, as a function of variables such as aging, obesity, and clinical pathophysiologic states.

According to Cannon, in emergencies, rapid activation of the sympathoadrenal system produce compensatory and anticipatory adjustments enhancing the likelihood of survival. Sheltered in a laboratory, with controlled temperature and *ad libitum* water, nutrients, and calories, mammals do not seem to require an intact sympathetic nervous system.

By now it is appreciated that pulse-synchronous bursts of skeletal muscle sympathetic nerve activity and plasma levels of norepinephrine normally are readily detectable even under resting conditions. Norepinephrine continuously enters the venous drainage of most organs. It is also by now clear that continual alterations in sympathetic nervous system outflows attend activities of daily life, such as meal ingestion, changing posture, and locomotion. Each of these activities is associated with a somewhat different set of 'normal' apparent steady states, directed by the brain and determined by coordinated actions of a variety of effector systems. This principle leads directly to the concept of allostasis, discussed below.

Stress and Science

Hans Selye popularized stress as a medical scientific idea. He elaborated a theory based on stress defined as the nonspecific response of the body to any demand, a stereotyped syndrome shared by all organisms in their interaction with the environment. The stress response was nonspecific, meaning that it could be brought on regardless of the stimulus. After specific responses were removed from consideration, the nonspecific syndrome would remain.

Although nonspecific with respect to the inciting agents, the stress response itself was thought to consist of a specific pathological pattern, with three components: enlargement of the adrenal glands, shrinkage of the thymus gland (associated with atrophy of the lymph nodes and inhibition of inflammatory or immune responses), and ulcers or bleeding in the stomach or gastrointestinal tract. The stress response was also thought to have characteristic stages, in a triphasic general adaptation syndrome. The first phase would be a rapid 'alarm' reaction. After an undefined length of time, a longer-lasting stage of resistance would ensue, with increased resistance to the particular stressor but decreased resistance to other stimuli. The stage of resistance would end with depletion of adaptation energy, ushering in the final stage, exhaustion, characterized by resurgence of activity of the adrenal gland, gastrointestinal ulcers, immunological failure, and death. Chrousos and Gold modified the doctrine of nonspecificity by proposing that above a threshold intensity, any stressor elicits the stress syndrome.

More than a half century elapsed before Selye's doctrine of nonspecificity underwent experimental

testing, which failed to confirm it. By now, researchers have largely abandoned both Cannon's and Selye's notions of stereotyped, nonspecific neuroendocrine responses regardless of the stressor. More modern theories view stress as a sensed threat to homeostasis, in which the response has a degree of specificity, depending among other things on the particular challenge to homeostasis and the organism's perception of the stressor and ability to cope with it. Nevertheless, Selye's theory figures prominently in current lay literature on stress.

According to the homeostat theory, stress is not an environmental stimulus or a stereotyped response but a condition, in which expectations, whether genetically programmed, established by prior learning, or deduced from circumstances, do not match current or anticipated perceptions of the internal or external environment, and this discrepancy between what is observed or sensed and what is expected or programmed elicits patterned, compensatory responses.

This sensation requires a comparative process, in which the brain compares available information with set points for responding. Central to the homeostat theory is that the body possesses numerous homeostatic comparators, homeostats. Each homeostat compares information with a set point for responding, determined by a regulator. Homeostatic systems typically use multiple effectors to influence values for the monitored variable. The loop is closed by afferent information about the status of the monitored variables. It is the state of sensed discrepancy that defines stress. The organism responds not to the external or internal perturbation itself but to changes in values of the variable that the afferent information actually delivers to the homeostat, usually located in the central nervous system. According to this definition, even a simple homeostatic reflex, such as the baroreflex, can reflect stress, when a perceived discrepancy between a set point for the monitored variable and information about the level of that variable elicits compensatory responses that decrease the discrepancy.

Homeostatic systems of higher organisms always include regulation by negative feedback. This is the key feature that maintains stable levels of the monitored variable. Increases in values of the monitored variable result in changes in effector activity that oppose and thereby buffer changes in the level of that variable. In homeostatic diagrams, a plus sign denotes a stimulatory and a minus sign an inhibitory relationship. Each minus changes the sign of the next relationship in the loop, whereas each plus does not affect the sign of the next relationship. In a system in which the monitored variable is maintained at a plateau level, the loops involving that variable have an odd number of minus signs.

If a loop has only plus relationships and no minus relationships (a positive feedback loop), the system, if otherwise unchecked, would 'explode'; levels of the monitored variable would increase or decrease exponentially. Positive feedback loops are inherently unstable. If a complex system included both a positive feedback loop and a 'nested' negative feedback loop, then the level of the monitored variable would change but to a stable plateau value.

Homeostatic systems use multiple effectors. This redundancy has several consequences. Disabling an effector compensatorily activates the others, enabling at least partial maintenance of the level of the monitored variable. Multiple effectors improve cost efficiency, measured by the energy used to maintain appropriate levels of the monitored variable. By decreasing wear and tear on effectors, having multiple effectors also prolongs the average useful life of the entire system. Finally, having multiple effectors enables patterning of effector responses. Patterning of hormonal, physiological, and behavioral effectors maximizes the likelihood of responses being appropriate for the particular threat to homeostasis.

The body's responses to different forms of stress occur in relatively specific patterns that help maintain appropriate levels of monitored variables such as temperature, blood pressure, blood glucose, and blood oxygen tension. A general theme of these patterns is that they are automatic, involuntary, and unconscious, but once the organism senses that the compensatory adjustments are not maintaining or will not maintain levels of the monitored variables, then the situation may reach consciousness, with the brain directing the adrenal gland to release epinephrine into the bloodstream and the individual experiencing distress, discussed below.

More than one homeostat can regulate activity of the same effector a phenomenon called effector sharing. Examples of effector sharing abound in physiology. For instance, both increased serum osmolality and decreased filling of the heart increase vasopressin secretion, and both decreased delivery of sodium to the kidneys and decreased blood pressure increase activity of the renin–angiotensin–aldosterone system.

Effectors interact with each other, both indirectly, such as via compensatory activation, and directly. Much of physiology research involves studying effector interactions. Examples are effects of adrenocortical steroids on adrenomedullary function and of angiotensin II on sympathetic nervous system outflows.

Disruption of a negative feedback loop, whether by blockade of afferent information, destruction of

homeostats, or failure of effectors, increases variability of the monitored variable.

Characteristics of Distress

According to the homeostat theory, distress is a form of stress with the additional characteristics of consciousness, aversiveness, observable signs, and adrenal gland activation. Distress is aversive to the organism, as evidenced by motivation for learning to escape or avoid the stressor.

The homeostat theory does not assume that distress causes disease. In contrast, Selye characterized distress as unpleasant or harmful, without separating these two very different characteristics. He never incorporated the relationship between distress and disease explicitly in his theory. Selye's theory emphasizes the nonspecificity of the stress response, whereas according to the homeostat theory, the experience of distress depends on the character, intensity, and meaning of the stressor as perceived by the organism and on the organism's perceived ability to cope with it. Distress responses, as all stress responses, have a 'purpose,' mitigating effects of a stressor in some way. This applies not only to neuroendocrine aspects of those responses but also to psychological aspects. Distress responses evolved and probably continue to be expressed even in higher organisms, including humans who actually are only rarely exposed to true fight-or-flight agonistic encounters, because of the importance of those responses in instinctive communication. Selye's theory does not consider the communication aspect of distress.

Homeostasis versus Allostasis

Claude Bernard propounded the founding concept of scientific integrative medicine, when he theorized that the overall goal of body systems is to maintain a constant internal cellular environment – the *milieu intérieur*. His *Lectures on the Phenomena of Life Common to Animals and Vegetables*, translated in 1974 from the French publication of 1876, contains one of the most famous passages in the history of physiology: "The constancy of the internal environment is the condition for free and independent life.... All the vital mechanisms, however varied they might be, always have one purpose, that of maintaining the integrity of the conditions of life within the internal environment."

Walter B Cannon coined the word homeostasis. By this, he referred to the stability of the internal environment. The concept of homeostasis was a direct extension from Bernard's *milieu intérieur*. According to Cannon, the brain coordinates body systems, with the aim of maintaining a set of goal values for key variables. Internal or external disturbances threatening homeostasis, by causing large enough deviations from the goal values, arouse nervous and hormone systems, induce emotional and motivational states, and generate externally observable behaviors, all of which have the goal of reestablishing homeostasis. The brain would respond to all emergencies in the same way, by evoking increased secretion of epinephrine.

According to the more recent concept of allostasis, no single set of ideal values exists for levels of internal variables. Allostasis comes from two Greek words that together mean 'other sameness.' The term was used by Sterling and Eyer in 1988, referring to levels of activity required for the individual to "maintain stability through change," that is, to adapt. In terms of the homeostat theory, allostasis is the set of apparent steady states maintained by multiple effectors. In the analogy of the home temperature control system, one can keep temperature at constant but different levels by appropriate use of effectors.

Homeostat resetting redefines the conditions required to maintain homeostasis. Regulation around an altered apparent steady state is the essence of allostasis. An example would be the hyperglycemia of exercise. Even in anticipation of the need for metabolic fuel, by activation of 'central command,' the blood glucose level increases. Resetting alters activities of multiple effector systems required to maintain allostasis, at least for short durations. During stress, short-term changes in homeostatic settings generally enhance the long-term well-being and survival of the organism. Responses during exercise provide an obvious example. When superimposed on a substrate of pathology, however, homeostatic resetting can cause harm. For instance, in the setting of ischemic heart disease, global or patterned increases in sympathetic outflows from homeostat resetting increase cardiac work, the resulting imbalance between oxygen supply and demand precipitating angina pectoris, myocardial infarction, or sudden death.

Allostatic Load and Chronic Disease

It is by way of prolonged activation of effectors to maintain allostasis that chronic stress can contribute to the development of chronic degenerative diseases. Allostatic load refers to effects of prolonged continuous or intermittent activation of effectors involved in allostasis. In the analogy of the home temperature control system, allostatic load increases if a window or door is left open. In this situation, one or more effectors are activated frequently or even continuously.

Continued use of the furnace and air conditioner in opposition to one another, an example of an inefficient allostatic state, consumes fuel and contributes to wear and tear on both pieces of equipment. Long-term allostatic load – the wear-and-tear cost of adaptation – provides a conceptual basis for studying long-term health consequences of stress.

Distinct Roles of the Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems

Differences in the extents of activation of the adrenomedullary hormonal and sympathetic noradrenergic systems in different stress responses provide convincing evidence for distinctive neuroendocrine stress response patterns (Table 1). This is ironic, because it was Cannon who was so instrumental in describing and studying these components and yet who advocated the combined and undifferentiated activation of the sympathoadrenal system in response to stress.

After adequately sensitive assay methods for plasma levels of norepinephrine and epinephrine became available, evidence rapidly accumulated for different noradrenergic versus adrenergic responses in different situations. A new concept began to emerge, in which the sympathetic noradrenergic system would help distribute blood flows appropriately, such as during orthostasis, cold exposure, mild blood loss, locomotion, exercise, altered salt intake, and water immersion. The adrenomedullary hormonal system would respond to global or metabolic threats, such as hypoglycemia, hemorrhagic hypotension, exercise beyond an anaerobic threshold, asphyxiation, emotional distress, and shock. Evidence also accumulated for an association between norepinephrine and active escape, avoidance, or attack, and an association in epinephrine and passive, immobile fear. Table 1 provides some examples of different patterns of sympathetic noradrenergic, adrenomedullary hormonal, and hypothalamic–pituitary–adrenocortical responses to different stressors.

Table 1 Hypothalamic–pituitary–adrenocortical (HPA), adrenomedullary hormonal system (AHS), and sympathetic noradrenergic system (SNS) responses to different stressors

Condition	*HPA*	*AHS*	*SNS*
Active escape/avoidance in rats	+	+	++
Cardiac arrest (or bypass)	+++	++++	++
Cold exposure, hypothermia	+	++	++++
Cold exposure, no hypothermia	0	+	+++
Exercise	+	++	+++
Exercise to exhaustion	++	+++	++++
Fainting	++	++++	0
Glucoprivation	+++	++++	+
Handling in rats	++	++	+
Hemorrhagic hypotension	+++	+++	+
Hemorrhage, no hypotension	+	+	++
Immobilization in rats	++++	++++	++++
Laboratory mental challenge	++	++	+
Pain	++	+++	++
Passive/immobile fear	++	+++	+
Public performance	++	+++	+
Social stress in rhesus monkeys	++	++	++
Surgery	+	+	++

Different intensities are indicated from 0 through ++++. There is a generally closer association between AHS and HPA than between AHS and SNS responses.

Thus, in contrast with the doctrine of nonspecificity, according to the homeostat theory of stress, activities of effector systems are coordinated in relatively specific patterns, including neuroendocrine patterns. These patterns serve different needs, and the sympathetic noradrenergic and adrenomedullary hormonal systems play important roles in many of them. For instance, sympathetic noradrenergic system activation predominates in response to orthostasis, moderate exercise, and exposure to cold, whereas adrenomedullary hormonal system activation predominates in response to glucoprivation and emotional distress (Table 1).

In terms of the body's thermostat, studies of humans exposed to cold or with mild core hypothermia have provided support for the notion of primitive specificity of neuroendocrine stress responses. Cold exposure increases plasma norepinephrine levels, with little if any increases in plasma epinephrine levels, consistent with sympathetic neuronal activation and relatively less adrenomedullary hormonal activation. Mild core hypothermia also increases antecubital venous levels of norepinephrine but not epinephrine. Both norepinephrine and epinephrine levels in arterial plasma increase in this setting, but with larger norepinephrine responses. These findings make sense, in that one can maintain body temperature effectively by sympathetically mediated cutaneous vasoconstriction, piloerection, and shivering. When these mechanisms give way, and core temperature falls, then high circulating epinephrine levels increase generation of calories, associated with the experience of distress, which motivates escape and avoidance and augments norepinephrine release from sympathetic nerve terminals for a given amount of nerve traffic.

For each stress, neuroendocrine and physiological changes are coupled with behavioral changes. For instance, the regulation of total body water in humans depends on an interplay between behavior (the search for water and drinking), an internal experience or feeling (thirst), and the elicitation of a neurohumoral response pattern (in this case

dominated by vasopressin, the antidiuretic hormone, and to a lesser extent angiotensin, a potent stimulator of drinking). Evoked changes in homeostat function often produce not only neuroendocrine and physiological effects but also behavioral responses; however, because of traditional boundaries among physiology, endocrinology, and psychology, interactions producing integrated patterns of response remain incompletely understood.

Scientific Integrative Medicine

Scientific integrative medicine is not a discipline, a group of disorders, or a method of treatment, but an approach, a way of thinking. Scientific integrative medicine uses systems concepts to explain disease processes and develop strategies to treat, prevent, or palliate them. It emphasizes disorders of the multiple interacting systems that regulate the body's "inner world." Scientific integrative medicine asks researchers and clinicians to consider more than one system at a time, as assessed by measuring levels of more than one chemical effector at a time. The scientific integrative medicine approach provides a framework for understanding complex and dynamic challenges to our integrity as organisms.

Scientific integrative medicine builds on systems biology, but it is also distinct in several ways. First, scientific integrative medicine recognizes that in higher organisms, including humans, the brain dominates regulation of the inner world. The brain regulates the many internal monitored variables of the body in parallel – analogous to a computer's multitasking. Second, the brain has plasticity, which enables modifications in the step-by-step instructions for cellular, tissue, organ, and systemic processes – the algorithms of life. According to the concept of allostasis, about which you will learn later in this article, set points and other elements of these algorithms vary, depending on recollections, sensations, and anticipations by the brain. Third, scientific integrative medicine is medical; its overall mission is to understand and rationally treat disorders and diseases. The systems that maintain the stability of the inner world eventually degenerate. Their efficiencies decline, and as they decline, the likelihood of deleterious, self-reinforcing positive feedback loops increases, threatening organismic stability and survival. Clinicians rarely cure; they manage. They exploit negative feedback loops and attempt to forestall or counter positive feedback loops. Finally, the medications and treatments clinicians prescribe interact with their patients' internal systems. Multiple, simultaneous degenerations, combined with multiple effects of multiple drugs and remedies and myriad interactions among those processes and the drug used to treat them, have come to the fore.

What renders scientific integrative medicine new, different, and powerful is the focus on disorders of the multiple, interacting, feedback-controlled systems that regulate the monitored variables of the body. Scientific integrative medicine recognizes that the operating characteristics of internal systems change over time, based on numerous life exposures and experiences that range from carrying out genetic instructions (especially in fetal development through adolescence) to aging-related degenerative disorders and interactions between medications and those disorders.

Integrative medicine has gained popularity recently. The word integrative has been used synonymously with holistic, mind–body, complementary, or alternative. The scientific integrative medicine approach, however, actually fits quite well with conventional, mainstream clinical medicine. Effective clinicians apply such an approach all the time. One such approach is to "exploit the lesion," where a treatment's success depends on an abnormality that occurs as part of the disease process itself. An example would be treatment of angina pectoris by nitroglycerine in patients with coronary artery disease. Nitroglycerine relaxes veins, decreasing venous return to the heart and cardiac stroke volume. Normally, decreased cardiac filling 'unloads' baroreceptors, and heart rate increases reflexively. The work of the heart changes little, because while the heart rate increases, the systolic blood pressure falls from the direct effects of the drug. Patients with angina pectoris from coronary artery disease, however, typically have low baroreflex-cardiovagal gain. When they take nitroglycerine, the drug reduces systolic blood pressure, but because of the low baroreflex-cardiovagal gain, the heart rate increases relatively little. The work of the heart therefore decreases, relieving the angina pectoris.

Another adage of scientific integrative medicine is "treat patients, not numbers." For hyperglycemia in gastrointestinal bleeding, or hyponatremia in heart failure, astute clinicians focus on treating the underlying problem. Appropriate treatment eliminates secondary effects of effector sharing, in these cases of the adrenomedullary hormonal system and vasopressin system, whereas attempting to reverse the abnormal laboratory findings directly can make matters worse.

One can also try to prevent compensatory damages, deleterious effects of reflexive effector activation. For instance, patients awakening from general anesthesia typically have postoperative hypothermia. Sympathetic noradrenergic and adrenomedullary hormonal system activation to counter hypothermia increases

myocardial oxygen consumption. This can contribute to the relatively high frequency of morbid events in patients with coronary artery disease during the immediate postoperative period.

The major goal of scientific integrative medicine is to treat effectively or prevent diseases before they become symptomatic. Beyond practices that promote health in general, such as regular exercise, dietary moderation, and refreshing sleep, the rapidly expanding availability of genetic, physiological, and biochemical information enables prevention of allostatic load and timely institution of effective treatments that are specific for the individual.

Induction of a positive feedback loop in a homeostatic system evokes instability. For instance, activation of the renin–angiotensin–aldosterone system in congestive heart failure increases sodium retention and vascular tone, leading to increased cardiac preload and afterload that worsen the congestive heart failure. Therefore, treatment with an angiotensin-converting enzyme inhibitor or angiotensin II receptor blocker can successfully treat congestive heart failure.

The homeostatic theory of stress and the concept of allostasis can help understand chronic as well as acute medical consequences of stress. Chronic activation of allostatic effectors in allostatic states promotes allostatic load. For instance, chronic elevations in adrenomedullary and hypothalamic–pituitary–adrenocortical outflows worsens insulin resistance, and chronic cardiac sympathetic activation accelerates cardiovascular hypertrophy and development of heart failure.

Another application of the homeostatic idea to medical consequences of stress is in terms of the perceived ability to cope. An organism experiences distress upon sensing that the effector responses will not be sufficient to restore or maintain allostasis. In contrast with distress, stress does not imply a conscious experience. Thus, even heavily sedated humans have substantial adrenomedullary stimulation in response to acute glucoprivation. Distress instinctively elicits observable signs and pituitary–adrenocortical and adrenomedullary activation. Via these neuroendocrine changes, distress could worsen pathophysiologic processes. For instance, because of adrenomedullary activation, in a patient with coronary artery stenosis, distress could elicit cardiovascular stimulation and produce an excess of myocardial oxygen consumption over supply, precipitating myocardial infarction or lethal ventricular arrhythmias. Moreover, long-term distress could augment both the risk of a mood disorder and the risk of worsening coronary disease.

Stress is an interdisciplinary topic, and understanding health consequences of stress requires an integrative approach. Research and ideas about stress must move beyond considering only one effector system, such as the pituitary adrenocortical system, and only one monitored variable, such as serum glucose levels, to incorporate multiple effectors that are regulated in parallel by multiple homeostats. They must also move beyond the notion of a single set of ideal values for monitored variables – homeostasis – to incorporate dynamic changes in homeostatic settings – allostasis. Merging of the homeostatic definitions of stress and distress with the concept of allostasis should provide a better understanding of the roles of catecholamine systems in stress, distress, and disease and provide a conceptual basis for the further development of scientific integrative medicine.

See also: Adrenal Steroids: Biphasic Effects on Neurons; Stress and Neural Involvement in Metabolism; Stress and Parasympathetic Control; Stress, Cytokines and Depressive Illness; Stress, the HPA Axis and Depressive Illness; Stress: Definition and History; Stress: Homeostasis, Rheostasis, Allostasis and Allostatic Load; Sympathoadrenal System: Neural Arm of the Stress Response.

Further Reading

Cannon WB (1939) *The Wisdom of the Body*. New York: W.W. Norton.

Chrousos GP and Gold PW (1992) The concepts of stress and stress system disorders. Overview of physical and behavioral homeostasis. *Journal of American Medical Association* 267: 1244–1252.

Goldstein DS (1995) *Stress, Catecholamines, and Cardiovascular Disease*. New York: Oxford University Press.

Goldstein DS (2001) *The Autonomic Nervous System in Health and Disease*. New York: Marcel Dekker, Inc.

Goldstein DS (2003) Catecholamines and stress. *Endocrine Regulations* 37(2): 69–80.

Goldstein DS (2006) *Adrenaline and the Inner World: An Introduction to Scientific Integrative Medicine*. Baltimore, MD: The Johns Hopkins University Press.

Goldstein DS and McEwen B (2002) Allostasis, homeostats, and the nature of stress. *Stress* 5(1): 55–58.

Kvetnansky R, Pacak K, Fukuhara K, et al. (1995) Sympathoadrenal system in stress. Interaction with the hypothalamic–pituitary–adrenocortical system. *Annals of the New York Academy of Sciences* 771: 131–158.

McEwen B and Stellar E (1993) Stress and the individual. Mechanisms leading to disease. *Archives of Internal Medicine* 153: 2093–2101.

Pacak K, Palkovits M, Yadid G, Kvetnansky R, Kopin IJ, and Goldstein DS (1998) Heterogeneous neurochemical responses to different stressors: A test of Selye's doctrine of nonspecificity. *American Journal of Physiology* 275: R1247–R1255.

Selye H (1956) *The Stress of Life*. New York: McGraw-Hill.

Sympathoadrenal System: Neural Arm of the Stress Response

M Palkovits, Semmelweis University, Budapest, Hungary

© 2009 Elsevier Ltd. All rights reserved.

The sympathetic nervous system, like the other major regulatory mechanisms of the central and peripheral nervous systems, can be divided into afferent and efferent arms. The somatosensory and viscerosensory projections carrying external and internal signals to the spinal and supraspinal compartments of the sympathetic nervous system represent its afferent arm. The efferent arm of the sympathetic nervous system (sympathetic outflow) appears to be organized into two major divisions: sympathoadrenal (SAS) and sympathoneural (SNS) systems. The term sympathoadrenal system is frequently used as a synonym for the sympathetic outflow or for the efferent arm of the sympathetic nervous system in general. It should be emphasized that according to their anatomical organization, neurotransmitter characteristics, and their functional role in the sympathetic outflow, the two systems are different: They may be independently activated, and they independently respond to stressful stimuli.

The activation of these two divisions of the efferent arm of the sympathetic nervous system is organized in a hierarchical manner (Figure 1). First, supraspinal (premotor sympathetic) neurons project to sympathetic preganglionic neurons in the thoracic spinal cord. Second, the sympathetic preganglionic neurons that represent the first compartment in both systems provide anatomical substrate through which selective regulation of sympathetic nervous system targets is accomplished. Fibers of the sympathetic preganglionic neurons innervate the adrenal medulla (SAS) and neurons in the sympathetic ganglia (SNS) (Figure 2). Third, the adrenal medullary chromaffin cells synthesize and release epinephrine (EPI) into the circulation. EPI in the plasma is exclusively of adrenomedullary origin. Neurons in the sympathetic ganglia project to a variety of peripheral targets in which norepinephrine (NE) is stored and released from the postganglionic sympathetic nerve endings, called varicosities. Almost all NE in the plasma originates from these nerve endings.

Activation of the sympathetic nervous system is accompanied by an increase in the outflow of both NE and EPI. Variation in the sensitivity and the number of pre- and postsynaptic receptors in the sympathetic nervous system may account for regional and functional differences in sympathetic activation, especially in response to different stressors.

Sympathetic Preganglionic Neurons

The sympathetic preganglionic neurons provide the anatomical substrate through which selective regulation of sympathetic nervous system targets is accomplished. The adrenal medulla and the sympathetic ganglion cells are directly innervated by sympathetic preganglionic neurons (Figure 3).

The sympathetic preganglionic neurons are cholinergic and form a longitudinal cell line, the intermediolateral cell column (IML), in the lateral portion of the spinal central gray matter, between the first thoracic (T1) and the second lumbar (L2) segments of the spinal cord. Within the IML, the sympathetic preganglionic neurons are topographically organized with respect to their tissue/organ targets. The adrenal medullary sympathetic preganglionic neurons (neurons that selectively project to the adrenal medulla, also called sympathoadrenergic cells) occupy the IML in segments T5–T8, whereas the sympathoneural sympathetic preganglionic neurons occupy, with some overlap, the IML in the other segments. In addition to the IML, scattered sympathetic preganglionic neurons are found in the nucleus intermediomedialis (called the central autonomic area) and in the lateral funiculus, immediately lateral to the IML.

Axons of the sympathetic preganglionic neurons leave the spinal cord through the ventral roots, at the level where their perikarya are located, and initially they run together with somatomotor axons within the spinal nerve. After a short distance, they separate from the spinal nerve and travel in small bundles, called white rami communicantes, to the paravertebral sympathetic chain (Figure 2). The sympathetic chain extends from the base of the skull down to the sacrum, on the two sides of the vertebral column. Some of the sympathetic preganglionic fibers form synaptic contact with ganglion cells within the chain (i.e., in paravertebral sympathetic ganglia). A group of the sympathetic preganglionic fibers establish synaptic contacts with ganglion cells at the same level where they enter the chain, whereas others travel upward or downward in the chain and terminate on paravertebral ganglion cells at more rostral or caudal levels (Figure 2). The vast majority of the ascending presynaptic sympathetic fibers in the sympathetic chain terminate in the cervical paravertebral sympathetic (superior and inferior (stellate)) ganglia.

The other portion of the sympathetic preganglionic fibers pass the chain without establishing synaptic contacts with any of the paravertebral ganglion cells. They enter some of the branches of the sympathetic chain and terminate on ganglion cells in one of the prevertebral

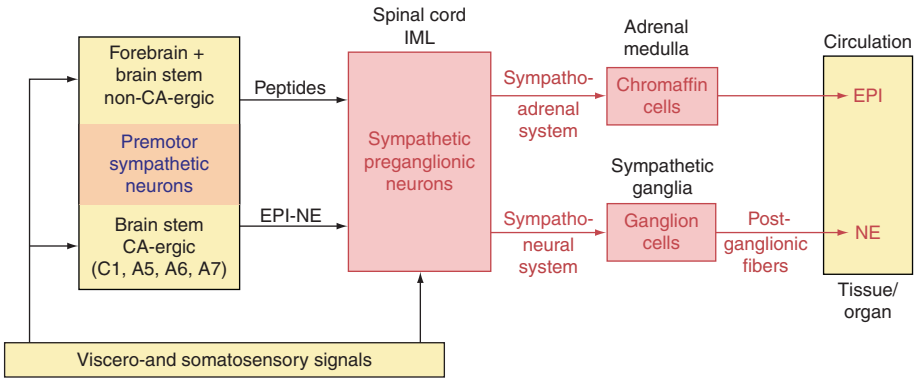

Figure 1 Neuroanatomical organization of the sympathetic nervous system. EPI, epinephrine; IML, intermediolateral cell column; NE, norepinephrine.

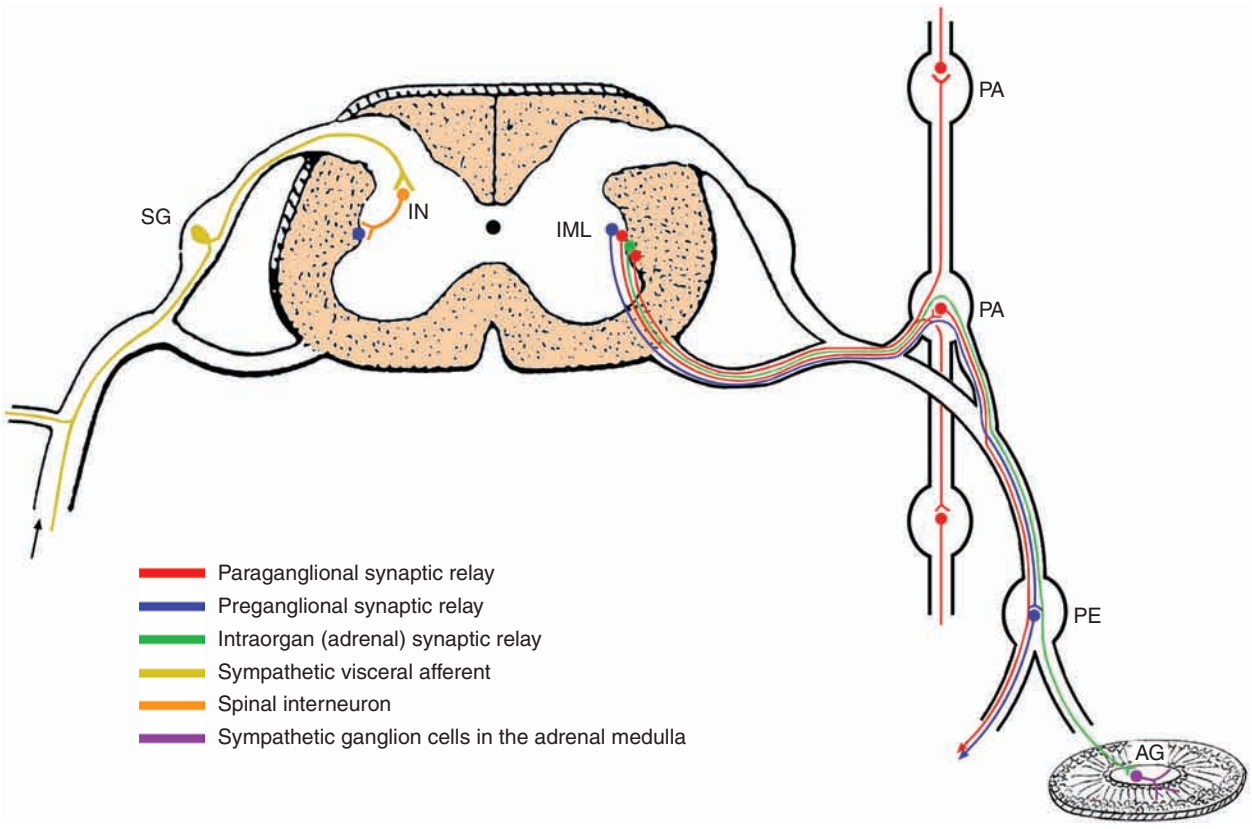

Figure 2 Viscerosensory inputs to the sympathetic preganglionic neurons (left) and the organization of pre- and postganglionic fibers in the sympathoadrenal and sympathoneural systems (right). AG, adrenomedullary ganglion cells; IML, intermediolateral cell column; IN, interneurons in the dorsal horn; PA, paravertebral ganglion; PE, prevertebral ganglion; SG, sympathetic ganglion.

ganglion in the abdomen, such as the celiac/superior mesenteric or inferior mesenteric ganglion (Figure 2). The sympathetic preganglionic neurons with synaptic contacts with either para- or prevertebral ganglion cells constitute the SNS.

A group of the sympathetic preganglionic fibers may pass the sympathetic chain, as well as the para- and prevertebral ganglia, without establishing synaptic contacts with ganglion cells (Figure 2). These presynaptic fibers leave the sympathetic chain for the splanchnic nerves; they enter the adrenal gland and terminate directly on the chromaffin cells in the adrenal medulla. They constitute the SAS.

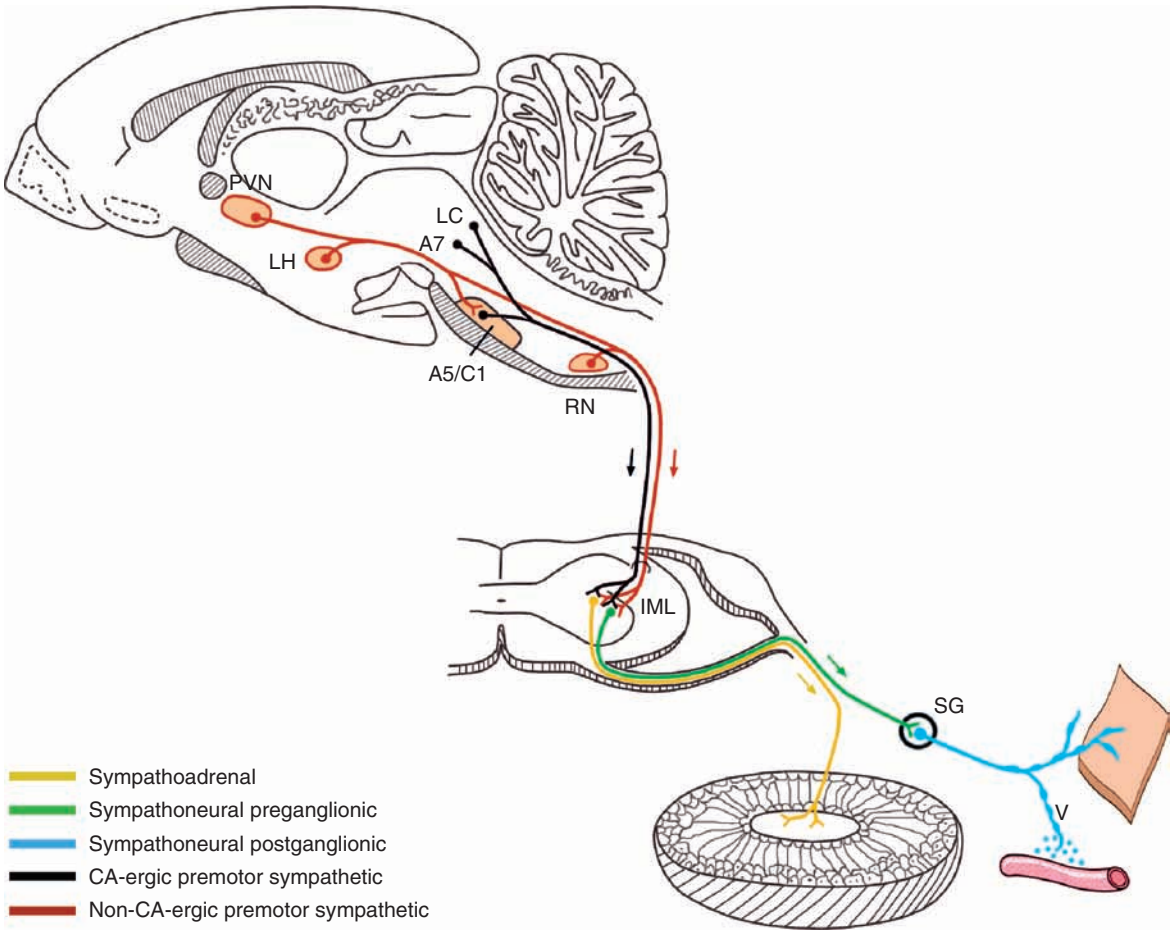

Figure 3 Afferent and efferent projections of spinal sympathetic preganglionic neurons. A5, A5 noradrenergic cell group; A7, A7 noradrenergic cell group; C1, C1 adrenergic cell group; IML, intermediolateral cell column; LC, locus coeruleus; LH, lateral hypothalamic area; RN, medullary raphe nuclei; PVN, paraventricular nucleus; SG, sympathetic ganglion; V, varicosities.

The Sympathoadrenal System

The SAS is composed of sympathetic preganglionic neurons and the chromaffin cells of the adrenal medulla. Chromaffin cells, like principal sympathetic ganglion cells, are members of the sympathoadrenal lineage derived from the neural crest. They share a common precursor, the sympathoadrenal progenitor cells. The sympathetic preganglionic neurons that innervate chromaffin cells are located in the IML of the thoracic spinal cord, mainly between segments T5 and T9. These neurons are cholinergic; acetylcholine is released from their terminals. Stimulation of the adrenal medulla leads to release of EPI and NE into the blood.

Two populations of adrenal chromaffin cells have been identified: EPI and NE secreting cells. The ratio of EPI- to NE-synthesizing chromaffin cells is approximately 4:1 in humans and slightly different in mammals. Although this ratio varies among species, EPI-synthesizing chromaffin cells predominate in the adrenal medulla. Note that a small number of chromaffin cells are also localized extra-adrenally, inside or around the sympathetic ganglia (paraganglia).

Each chromaffin cell contains approximately 10 000 chromaffin granules, in which EPI or NE are stored. They are secreted via exocytosis, carried by capillaries that drain from the adrenal medulla into the adrenal vein. Then, EPI and NE are transported throughout the body and exert their effects on most of the major regulatory systems. Besides catecholamines, the chromaffin granules contain numerous other transmitters, granins, and neuropeptides, which are co-localized and co-released with catecholamines.

The NE-synthesizing chromaffin cells possess the full repertoire of functions of EPI-synthesizing cells except the capacity to synthesize EPI. The two types of chromaffin cells are innervated by different sympathetic preganglionic fibers. This arrangement provides a basis for distinct sympathetic circuits underlying the differential regulation of EPI and NE secretion from

the adrenal medulla in response to physiological challenges and stressful stimuli.

In addition to chromaffin cells, the adrenal medulla contains a small number of ganglion cells, which synthesize NE. Like chromaffin cells, these ganglion cells are also innervated by sympathetic preganglionic fibers (Figure 2), but they contribute little to NE content in plasma and do not synthesize EPI. Their postganglionic axons distribute in the adrenal and terminate both in the medulla and in the adrenal cortex. Thus, ganglion cells in the adrenal medulla have similar characteristics as ganglion cells in the pre- or paravertebral ganglia, neverthelen they are embedded in the innervated organ, that is, in the adrenal gland. Despite their location, the adrenal medullary ganglion cells functionally belong to the SNS rather than the SAS.

The Sympathoneural System

Neuroanatomically, the SNS is organized in a two-neuron chain. The first neuron (as in the SAS) is referred to as the sympathetic preganglionic neuron, whereas the second neuron is called the ganglionic (or, frequently, postganglionic) neuron, which projects a so-called postganglionic axon. The preganglionic neurons in the IML are cholinergic and innervate the sympathetic ganglion cells by release of acetylcholine as the primary neurotransmitter that acts on nicotinic and, to a lesser extent, muscarinic receptors. Various biologically active neuropeptides, such as substance P, vasoactive intestinal polypeptide, and enkephalins, are co-localized with acetylcholine within the sympathetic preganglionic terminals. They may act as neuromodulators of cholinergic transmission.

The ganglion cells reside in sympathetic paravertebral (superior cervical, stellate ganglia, and ganglion cells in the sympathetic chain) and prevertebral (celiac and superior and inferior mesenteric) ganglia (Figure 2). With terminal branches, a single sympathetic preganglionic axon terminates on several para- or prevertebral ganglion cells. The divergence ratio of the pre- and postsynaptic neurons, depending on the species, is between 20:1 and 5:1. This divergence permits coordinated activity in sympathetic neurons at several different spinal levels. In contrast, ganglion cells establish synaptic contacts with more than one sympathetic preganglionic neuron. Therefore, a substantial convergence of synaptic inputs exists on each ganglion cell.

The sympathetic postganglionic nerve originating from either a pre- or paravertebral ganglion cell establishes a plexus of terminal fibers within a short distance to the target tissue/organ (Figure 2). The length of a terminal fiber may be several millimeters. The postganglionic fibers show a beaded appearance due to local swellings called varicosities. The varicosities are specialized structures involved in the storage and release of NE. They contain a high concentration of large and small granular vesicles that store mainly NE. A single postganglionic fiber may contain more than 20 000 varicosities with more than 1000 granular vesicles in each. Each granule stores approximately 15 000 molecules of NE. The varicosities do not display synaptic junction; NE is released to the extracellular space by exocytosis and diffuses over a large territory of the target area or is drained directly into the blood. Almost all of the NE in the plasma originates from the varicosities of the postganglionic fibers. In certain conditions, the nerve endings may take up and store a significant amount of NE from the extracellular space. The process of synthesis, storage, and release of NE from the varicosities is under the control of the sympathetic preganglionic neurons. In addition to the para- and prevertebral sympathetic ganglia, sympathetic neurons are also found in a small number in ganglia in or near the pelvic organs (pelvic ganglia).

NE is not the sole neurotransmitter in the sympathetic ganglion cells; they express other cotransmitters and neuropeptides, such as neuropeptide Y and galanin. The co-expression and the selective outflow of one or more neurotransmitters or neuropeptides from the sympathetic postganglionic nerve terminals allow the organism to differentiate between sympathetic activities and responses at the level of the target organ.

In addition to ganglion cells, sympathetic ganglia contain small granular cells called intensely fluorescent (SIF) cells. These cells express and store predominantly dopamine, and some of them also express and store EPI. SIF cells, like the sympathetic ganglion cells and the chromaffin cells, are members of the sympathoadrenal lineage derived from the neural crest.

Premotor Sympathetic Neurons

Neurons at multiple levels of the neuraxis, ranging from the lower brain stem to the forebrain, regulate the sympathetic outflow. These neurons, called premotor sympathetic neurons, innervate the sympathetic preganglionic neurons in the IML (Figure 3). Neurotransmitters that regulate the preganglionic neurons include NE, EPI, dopamine, serotonin, and neuropeptides. Thus, sympathetic activities elicited by the activation of sympathetic preganglionic neurons are controlled by distinct neuronal circuits and probably by distinct neurotransmitters. Accordingly, the premotor sympathetic neurons used to be classified as catecholaminergic and noncatecholaminergic (Figure 1).

The hypothalamus is regarded as the principal brain region of central autonomic regulation. Most of the forebrain regions, such as the amygdala, hippocampus, and limbic cortical areas, are involved in the control of the autonomic nervous system outflow. They do not directly innervate the IML; rather, they exert their influence on preganglionic sympathetic neurons by way of the hypothalamus. Hypothalamic neurons projecting to the IML of the thoracic spinal cord reside mainly in the paraventricular nucleus, but some reside in the lateral hypothalamic area and the arcuate nucleus (Figure 3). The hypothalamic premotor sympathetic neurons are peptidergic. Serotoninergic neurons in the rostral ventromedial medulla and the caudal raphe nuclei also project to sympathetic preganglionic neurons. These cells, together with the hypothalamic neurons, represent the noncatecholaminergic premotor sympathetic neurons (Figure 1).

In the lower brain stem, catecholaminergic neurons in the rostral ventrolateral medulla (in the A5 noradrenergic and C1 adrenergic cell groups) constitute the major portion of the catecholaminergic premotor sympathetic neurons that preferentially terminate on the sympathetic preganglionic neurons (Figure 3). In addition, neurons of the A7 noradrenergic cell group in the pons and neurons in the caudal portion of the locus coeruleus belong to this system. The distribution of the descending premotor sympathetic axons in the IML is topographically and functionally organized: The target-specific IML neurons are selectively innervated by different catecholamine premotor neurons. The NE or EPI nerve terminals establish conventional synaptic contacts with sympathetic preganglionic neurons, especially in the segments T5–T9, where the sympathoadrenergic preganglionic cells are located. The IML cell column also receives dopamine innervations, mainly from the A11 dopamine cell group in the posterior diencephalon.

In certain conditions, the two premotor sympathetic systems may act synergistically: Hypothalamic neurons that project to the IML send axon collaterals to catecholaminergic neurons in the lower brain stem (Figure 3).

Sensory Arms of the Sympathetic Nervous System

To maintain homeostasis, the sympathetic nervous system responds to several different types of sensory inputs. Some of these inputs are somatosensory, including nociceptive signals, whereas others – viscerosensory inputs – arise from interoceptors in the internal organs. They may reach spinal and supraspinal receptors by neural or humoral ways. Inputs which influence the activity of the sympathetic preganglionic or premotor sympathetic neurons (Figure 1) are called sympathetic sensory afferents. They represent the sensory (also called ascending or afferent) arm of the sympathetic regulatory circuit.

The sympathetic visceral afferent fibers enter the dorsal horn of the spinal cord through the dorsal roots and innervate the sympathetic preganglionic neurons in the IML via interneurons located in lamina V and lamina VII of the spinal cord (Figure 2). Some sensory fibers that project to the spinal cord also send a branch to sympathetic ganglia forming reflex circuits that may control some visceral autonomic functions.

Viscerosensory signals to supraspinal premotor sympathetic neurons are relayed in the nucleus of the solitary tract (NTS), which represents the primary viscerosensory center in the central nervous system. Viscerosensory fibers travel in cranial nerves (mainly in the trigeminal, glossopharyngeal, and vagal nerves) or in ascending spinal fibers (spino-solitary tract) and terminate in the NTS. The sensory afferents terminate in a topographically and functionally organized manner in the different subdivisions of the nucleus. The ascending fibers from the NTS may reach both the catecholaminergic and the noncatecholaminergic premotor sympathetic neurons in the brain stem and the hypothalamus. Some of them influence the activity of limbic and cortical structures that may respond to these signals through the activation of hypothalamic or brain stem premotor sympathetic neurons.

Sympathetic Nervous System in the Organization of Stress Response

The sympathetic nervous system helps to maintain homeostasis during exposure to stressors. Activation of the sympathetic nervous system leads to the outflow of catecholamines from the adrenal medulla and from the varicosities of the sympathetic postganglionic axons. These two systems constitute the final compartments of the effector loop of the sympathetic nervous system that participate in the organization of stress response.

Under resting conditions, low levels of EPI and NE are released into the blood flow from the adrenal medulla and sympathetic postganglionic nerve terminals, respectively. During stressful situations, large amounts of EPI and NE are released. Depending on the nature, intensity, and duration of stressful stimuli, levels of NE and EPI in the plasma may increase 10- to 100-fold (or even higher) within a very short period of time after the stressful stimuli occur. The tissue levels of NE measured in various organs are reduced in acute stress since approximately 70% or more of the circulating NE is derived from the

varicosities of the postganglionic sympathetic nerve fibers, which provide the sympathetic innervations of the organs. Similarly, levels of NE and EPI are reduced in the adrenal medulla, depending on the intensity and the duration of acute stressors. After long-term or repeated stress exposure, adrenal medullary catecholamine stores are increased in the chromaffin cells. In this situation, the increased catecholamine synthesis, as a consequence of the stress-induced marked elevation of gene expression of catecholamine-synthesizing enzymes (tyrosine hydroxylase, dopamine β-hydroxylase, and phenylethanolamine *N*-methyltransferase (PNMT)) in the adrenal medulla, exceeds the release of NE and EPI.

Based on measurements of plasma levels of EPI, NE, catecholamine precursors, and metabolites in stress, it is generally accepted that SAS and SNS are specifically, and somewhat independently, activated by different stressors. Exposures to cold evoke a selective activation of SNS, as indicated by elevated plasma NE levels, whereas insulin-induced glucoprivation or emotional stressors activate SAS, as indicated by plasma EPI responses. Other stressors, such as restraint or formalin-induced pain, activate both systems. Immobilization represents a severe stress that elicits an extremely large increase in plasma NE and EPI. Results from a variety of studies indicate that not only the peripheral but also the central catecholamine system and the hypothalamic–pituitary–adrenal (HPA) axis respond specifically to different stressful stimuli. An intact HPA axis is essential for the activity of PNMT, the enzyme responsible for EPI synthesis in the adrenal medulla. Thus, depending on the nature and the target sites of various stressors, SAS and SNS may be regulated by different central pathways. For example, a single subcutaneous formalin injection, which strongly increases EPI levels in the plasma in intact animals, may reduce the immobilization-induced high EPI levels whether it was injected before or during immobilization.

Whereas our knowledge of the stress-related activity of SAS is continuously increasing, SNS has been much less studied than the adrenal medulla and relatively little is known about the mechanisms of gene expression of catecholamine biosynthetic enzymes in sympathetic ganglion cells following stressful stimulation.

See also: Chromaffin Cells: Model Cells for Neuronal Cell Biology; Chronic (Repeated) Stress: Consequences, Adaptations; Norepinephrine: CNS Pathways and Neurophysiology; Stress: Definition and History; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress; Sympathetic Nervous System.

Further Reading

Björklund A, Hökfelt T, and Owman C (eds.) (1988) *Handbook of Chemical Neuroanatomy, Vol. 6: The Peripheral Nervous System.* Amsterdam: Elsevier.

Burnstock G and Hoyle CHV (eds.) (1995) *The Autonomic Nervous System, Vol. 1: Autonomic Neuroeffector Mechanisms.* London: Harwood Academic.

Elfvin LG, Lindh B, and Hökfelt T (1993) The chemical neuroanatomy of sympathetic ganglia. *Annual Review of Neuroscience* 16: 471–507.

Gabella G (1995) Autonomic nervous system. In: Paxinos G (ed.) *Rat Nervous System,* 2nd edn., pp. 81–103. San Dieg, CA: Academic Press.

Goldstein DS (2001) *The Autonomic Nervous System in Health and Disease.* Oxford: Oxford University Press.

Kandel ER, Schwartz JH, and Jessell TM (2000) *Principles of Neural Science*, 3rd edn., pp. 961–981. New York: McGraw-Hill.

Langley K and Grant NJ (1999) Molecular markers of sympathoadrenal cells. *Cell and Tissue Research* 298: 185–206.

Loewy AD and Spyer KM (eds.) (1990) *Central Regulation of Autonomic Function.* New York: Oxford University Press.

Pacak K and Palkovits M (2001) Stressor specificity of central neuroendocrine responses: Implications for stress-related disorders. *Endocrine Reviews* 22: 502–548.

Palkovits M (1999) Interconnections between the neuroendocrine hypothalamus and the central autonomic system. *Frontiers in Neuroendocrinology* 20: 270–295.

Strack AM, Sawyer WB, Platt KB, and Loewy AD (1989) CNS cell groups regulating the sympathetic outflow to adrenal gland as revealed by transneuronal cell body labelling with pseudorabies virus. *Brain Research* 491: 274–296.

Sun MK (1995) Central neural organization and control of sympathetic nervous system in mammals. *Progress in Neurobiology* 47: 157–233.

Sympathomimetic Drugs and Adrenergic Receptor Antagonists

T C Westfall, Saint Louis University School of Medicine, St. Louis, MO, USA

© 2009 Published by Elsevier Ltd.

Sympathomimetics

Sympathomimetic drugs are agents which in general mimic responses due to stimulation of sympathetic nerves. These agents are able to directly activate adrenergic receptors or to indirectly activate them by increasing norepinephrine and epinephrine (mediators of the sympathoadrenal system) levels. Chemically these drugs are catecholamines (Figure 1) or are structurally similar to catecholamines. The parent structure is β-phenylethylamine. When hydroxyls are placed on β-phenylethylamine at the 3 and 4 positions of the aromatic ring (Figure 1), the drugs are called a catecholamine. There are three naturally occurring catecholamines (dopamine, norepinephrine, and epinephrine) as well as various synthetic amines (e.g., isoproterenol and dobutamine).

Adrenergic Receptors

Adrenergic receptors are the specific molecular structures in or on effector cells with which catecholamines or sympathomimetic drugs react in order to elicit the characteristic response(s) of the cells. Adrenergic antagonists also bind to these receptors and prevent the response(s) mediated by endogenous or exogenous agonists. Adrenergic receptors were originally defined by the rank order of potency of a series of agonists, or specificity of antagonists. More recently they are also identified by their molecular structure and cloning of the gene. Currently, there are two general types of adrenergic receptors, α and β, with various subtypes (α_{1A}, α_{1B}, α_{1D}; α_{2A}, α_{2B}, α_{2C}; β_1, β_2, and β_3). All of the adrenergic receptors belong to the family of G-protein-coupled receptors (GPCRs) that link to heterotrimeric G-proteins. Although structurally related, different receptors regulate distinct physiological processes by controlling the synthesis or release of a variety of second messengers or the activity of ion channels. Each major type of receptor shows preference for a particular class of G-protein (e.g., α_1 to G_q; α_2 to G_i). Stimulation of α_1 adrenergic receptors results in the regulation of multiple effector systems. A primary mode of signal transduction involves activation of the G_q/phospholipase Cβ/inositol 1,4,5-trisphosphate/Ca^{2+} and calmodulin-sensitive pathways, such as calmodulin (CaM) kinase. In addition to these pathways, α_1 adrenergic receptor-induced transduction of G_q can lead to elevations of phospholipase D_1, phospholipase A_2, and activation of ligand-gated Ca^{2+} channels. Activation of α_2 adrenergic receptors also couples to a variety of effectors and is transduced through $G_{i1,2,3}$, resulting in a decrease in adenylyl cyclase and cyclic AMP, and through the $\beta\gamma$ subunits of G_i to activate K^+ channels, and through G_o to inhibit L- and N-type Ca^{2+} channels. All three β adrenergic receptors couple to G_s and activate adenylyl cyclase and increase cyclic AMP.

Physiological/Pharmacological Response

The actions of sympathomimetics can be classified into at least seven broad types of responses: (1) excitatory action on smooth muscle (e.g., contraction of vascular smooth muscle), (2) inhibitory action on smooth muscle (e.g., relaxation of bronchiolar smooth muscle), (3) stimulation of the heart, leading to increased heart rate, contraction, and conduction, (4) metabolic actions (e.g., glycogenolysis in liver, resulting in an increase in glucose), (5) endocrine actions (e.g., increasing or decreasing insulin release), (6) actions in the central nervous system, and (7) prejunctional action on nerves, increasing or decreasing transmitter synthesis or release. The response to sympathomimetics often can be predicted on the basis of knowledge of the selectivity of the agent for reacting with a specific α or β receptor and the physiological response governed by these adrenergic receptors in specific cells, tissues, or organs. Table 1 depicts the responses produced following activation of adrenergic receptors or stimulation of sympathetic nerves.

The response of any cell or organ to sympathomimetic drugs is dependent on the density and proportion of α and β adrenergic receptors. An example is bronchiolar smooth muscle, in which the principal adrenergic receptor is β_2. Norepinephrine has little effect on bronchiolar smooth muscle since it has little activity on the β_2 receptor. On the other hand, epinephrine or isoproterenol will relax bronchiolar smooth muscle due to the ability of these compounds to activate β_2 adrenergic receptors. The ultimate response of an effector cell or organ to a sympathomimetic drug is dictated not only by the direct effect of the agents, but also by reflex homeostatic mechanisms. This is illustrated well by α_1 adrenergic agonists. Such drugs will constrict blood vessels, therefore increasing blood pressure. This, in turn, elicits a compensatory decrease in heart rate resulting from the carotid sinus–aortic arch baroreflex system. An increase in blood pressure results

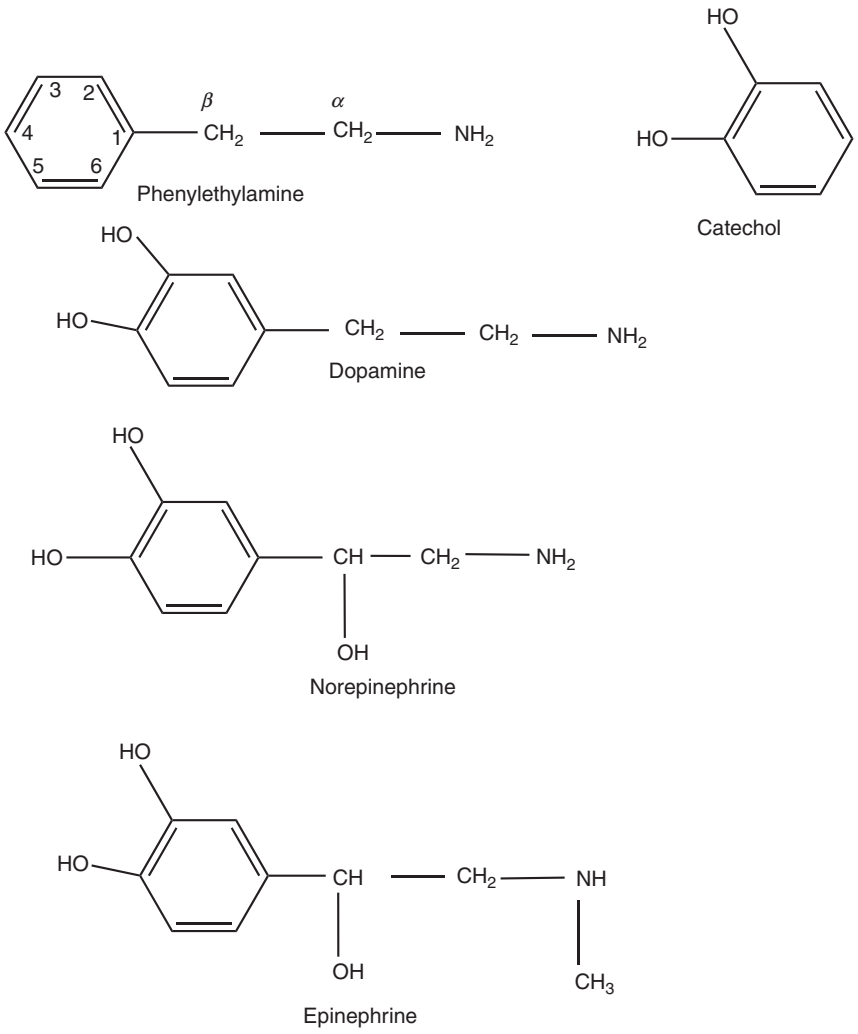

Figure 1 Chemical structure of catechol β-phenylethylamine and naturally occurring catecholeamines.

in a decrease in sympathetic outflow from the central nervous system and an increase in parasympathetic outflow. Decreases in blood pressure result in the opposite: an increase in sympathetic outflow and a decrease in parasympathetic outflow. Although a drug such as phenylephrine has little direct effect on the heart, it can actually decrease heart rate indirectly following its increase in blood pressure (via α_1 activation) and activation of the baroreflex resulting in an increase in parasympathetic outflow.

Classification of Sympathomimetic Amines

Catecholamines and sympathomimetic drugs are classified as directly acting, indirectly acting, or mixed acting (Figure 2). Directly acting drugs may have selectivity for a specific adrenergic receptor subtype (e.g., phenylephrine for α_1, clonidine for α_2) or may act on numerous receptor subtypes (e.g., epinephrine for α_1, α_2 , β_1, β_2, β_3). Indirectly acting sympathomimetic amines increase the amount of norepinephrine available to stimulate adrenergic receptors (e.g., amphetamine, methamphetamine).

Some drugs have the ability to increase the amount of norepinephrine and therefore produce sympathomimetic actions (e.g., cocaine, entacapone) but chemically are not sympathomimetic drugs. In general, drugs can increase the amount of norepinephrine by (1) releasing norepinephrine from sympathetic terminals (amphetamine), (2) blocking the reuptake of norepinephrine into sympathetic neurons (cocaine), or (3) inhibiting the biotransforming enzymes monoamine oxidase (pargyline) or catechol-*O*-methyltransferase (entacapone). Mixed-acting drugs (ephedrine, dopamine) can directly activate adrenergic receptors as well as increase the amount of norepinephrine.

Figure 2 lists some prototypical drugs which act by one or more of these mechanisms. Although the classification of sympathomimetic drugs into directly,

Table 1 Peripheral adrenergic receptors and responses

Organ (site)	*Response*	*Receptor*
Eye		
Dilator or radial smooth muscle of the iris	Contraction (mydriasis)	α_1
Ciliary muscle	Relaxation (accommodation for vision)	β
Heart		
SA node	Increase in rate (positive chronotropic)	β_1 (β_2)
Atria	Increase in contractility (positive inotropic)	β_1 (β_2)
AV node	Increase in conduction velocity (positive dromotropic effects)	β_1 (β_2)
Ventricle	Increase in contraction (positive inotropic)	β_1 (β_2)
Blood vessels		
Arteries (cutaneous, visceral, skeletal, pulmonary)	Contraction	α_1 (α_2)
Some arteries (skeletal and visceral)	Relaxation	β_2
Veins	Contraction	α_1 (α_2)
Veins	Relaxation	β_2
Lung		
Bronchial muscle	Relaxation (bronchodilation)	β_2
Mast cells	Inhibition of mediators	β_2
Mast cells	Stimulation of mediators	α
Gastrointestinal tract		
Motility	Decrease	α_1, α_2, β_1, β_2
Sphincters	Contraction	α_1 (α_2)
Sphincters	Relaxation	β_2
Secretions	Decrease	α_2 (β_2)
Urinary bladder		
Detrusor muscle	Relaxation	β_2
Trigone and sphincter	Contraction	α_1
Uterus	Relaxation	β_2
Uterus	Contraction	α_1
Salivary gland	$\uparrow K^+$ release	α_1
	$\uparrow$ Amylase secretion	β_1
Metabolic or endocrine		
Liver	Increase glycogenolysis	α_1, β_2
Skeletal muscle	Increase glycogenolysis	β_2
Fat cells	Lipolysis (thermogenesis)	α, β_1, β_2, β_3
Pancreatic islets	Inhibition	α_2
Pancreatic islets	Stimulation	β_2
JG cells (kidney)	Increase renin release	β_1
Spleen	Contraction	α_1
Pilomotor muscle	Stimulation	α_1
Nerve terminals	Inhibition of transmitter release	α_2
Autonomic ganglia	Hyperpolarization	α_2

indirectly, or mixed acting is convenient, it is not absolute and there is a continuum from predominantly directly acting, on the one hand, to predominantly indirectly acting, on the other.

Endogenous Catecholamines

Epinephrine; Norepinephrine

Epinephrine is a hormone released from the adrenal medulla as well as a neurotransmitter in the central nervous system (CNS). Norepinephrine is the principal neurotransmitter released from postganglionic sympathetic nerve terminals, a hormone released from the adrenal medulla, and a neurotransmitter in the CNS. The endogenous catecholamines (dopamine, norepinephrine, and epinephrine) are rapidly metabolized by catechol-O-methyltransferase (COMT) and monoamine oxidase (MAO). They are charged molecules at physiological pH and are inactive when given by the oral route. When released from nerve endings they are rapidly and effectively transported by the norepinephrine transporter (NET) or dopamine transporter (DAT) into presynaptic terminals. The uptake may also occur when the three amines are given as drugs. All have short duration of action and do not penetrate into the CNS (across the blood–brain barrier).

As mentioned, epinephrine is nonselective and it can directly activate all adrenergic receptors (α_1, α_2, β_1, β_2, and β_3), whereas norepinephrine activates α_1, α_2, and β_1 receptors but has very little activity on β_2 receptors. This highlights the biggest difference between epinephrine and norepinephrine. The pharmacological actions of the catecholamines can be

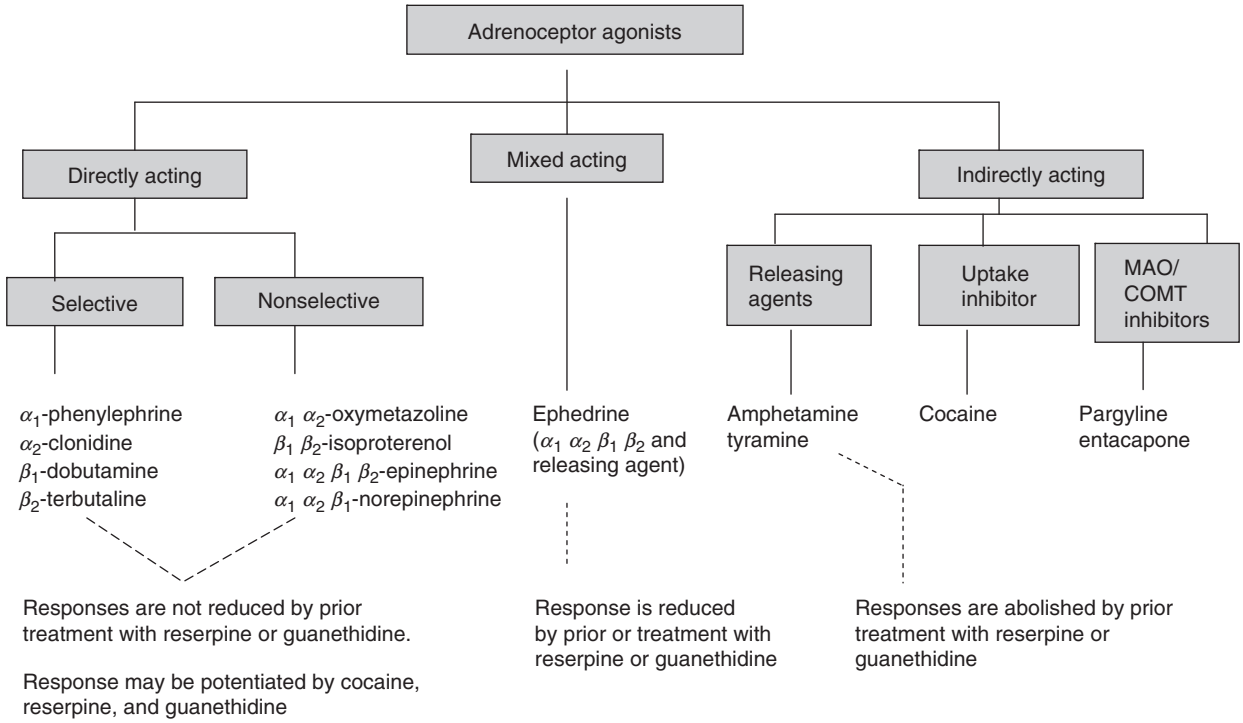

Figure 2 Classification of adrenergic receptor agonists (sympathomimetic amines) or drugs that produce sympathomimetic-like effects. For each category, a prototypical drug is shown. (Cocaine, pargyline, and entacapone are not actually sympathetic drugs but do produce sympathomimetic-like effects.) Adapted from Westfall TC and Westfall DP (2006) In: Brunton L, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th edn., ch. 10. New York: McGraw-Hill.

deduced by examining Table 1. Both norepinephrine and epinephrine have prominent cardiovascular effects, especially after parenteral administration. Since epinephrine is a potent stimulant of both α and β receptors, its effects on target organs are complex. Epinephrine is a potent pressor drug. The mechanism of the rise in blood pressure by epinephrine is due to (1) direct myocardial stimulation which increases ventricular contraction (positive inotropic effect), (2) increased heart rate (positive chronotropic effect), and (3) vasoconstriction in many vascular beds – especially in the precapillary resistance vessels of skin, mucosa, and kidney, along with marked constriction of veins. Depending on the route and dose, epinephrine may also decrease peripheral resistance due to its activation of β_2 receptors in skeletal muscle blood vessels, resulting in vasodilators and an enhancement of blood flow. In this case, mean blood pressure is not greatly elevated and therefore the compensatory baroreflex does not appreciably antagonize the direct cardiac actions. Heart rate, cardiac output, stroke volume, and ventricular work are all increased due to direct cardiac stimulation and increased venous return. The direct effects of epinephrine on the heart are one of a powerful stimulant due to activation of β receptors. All aspects of myocardial function are increased (heart rate, contraction, conduction velocity). Arrhythmias may be seen with high epinephrine concentrations.

The effects on nonvascular smooth muscle depend upon the type of receptor (e.g., relaxation of bronchiolar smooth muscle via activation of β_2 receptors; see Table 1). Epinephrine has been used clinically to treat acute asthma, open-angle glaucoma, and anaphylactic shock, and is used with local anesthetics to prolong their duration of actions. It has also been used occasionally in complete heart block or cardiac arrest.

Norepinephrine is also a potent pressor agent and increases both systolic and diastolic blood pressure and peripheral resistance. Unlike epinephrine, norepinephrine does not activate β_2 receptors in skeletal blood vessels and there is no decrease in diastolic blood pressure. The marked increase in peripheral resistance and mean arterial blood pressure following intravenous administration can result in a compensatory slowing of the heart rate, which overrides the direct effects of norepinephrine on heart rate. Like epinephrine, norepinephrine directly increases all aspects of myocardial function due to activation of β receptors. Norepinephrine is used to treat various forms of hypotension, including shock. Side effects of

epinephrine and norepinephrine include palpitations of the heart, cardiac arrhythmias, cerebral hemorrhage, headache, tremor, and restlessness.

Dopamine

Dopamine is the third naturally occurring catecholamine. It is the immediate precursor of norepinephrine, an important CNS neurotransmitter, an interneuron mediator in some autonomic ganglia, and perhaps a local hormone in the kidney. Dopamine can directly activate α_1 and β receptors and can also evoke the release norepinephrine from sympathetic neurons and is therefore classified as a mixed-acting sympathomimetic. Dopamine is a substrate for both MAO and COMT and is inactive orally. In addition to its effects on α_1 and β receptors it can activate dopamine D_1 receptors located in renal, mesenteric, and coronary blood vessels, leading to dilation. Activation of D_1 receptors on renal tubular cells decreases sodium transport and produces natriuresis. Dopamine is used clinically to treat cardiogenic shock, congestive heart failure, and acute renal failure. A side effect of high doses is vasoconstriction.

Nonselective β Agonists

Isoproterenol

This is a synthetic catecholamine with activity on all β receptors ($\beta_1, \beta_2, \beta_3$). It is readily absorbed when given parenterally or as an aerosol. It is primarily metabolized by COMT, is a poor substrate for MAO, and is poorly taken up by sympathetic neurons. Isoproterenol produces positive inotropic and chronotropic effects, leading to an increase in cardiac output, unchanged or increase in systolic blood pressure, and a decrease in diastolic blood pressure. It will relax nonvascular smooth muscle containing β_2 receptors.

Dobutamine

Although often thought of as the prototypical β_1 selective agonist, its pharmacological effects are complex. It has two enantioimeric forms present in the racemic mixture used clinically. The (−) isomer is a potent α_1 receptor agonist that can cause pressor effects. The (+) isomer is a potent α_1 receptor antagonist which can block the effects of (−) dobutamine. The (+) isomer is a more potent β receptor agonist than is the (−) isomer. Dobutamine is a myocardial stimulant with more prominent inotropic than chronotropic effects. The drug is indicated for the short-term treatment of cardiac decompensation that may occur after cardiac surgery or in patients with congestive heart failure or acute myocardial infarction.

Adrenergic Receptor Agonists

β_2-Selective Adrenergic Receptor Agonists

Metaproterenol, terbutaline, albuterol, isoetharine, pirbuterol, bitolterol, fenoterol, fomoterol, procaterol, salmeterol, and ritodrine comprise a series of drugs which have preferential affinity for β_2 receptors, compared with β_1 receptors. However, this selectivity is not absolute and is lost at high concentrations of these drugs. Chemical modification has resulted in lower rates of biotransformation and enhanced oral bioavailability relative to catecholamines. The selectively for β_2 receptors theoretically allows these drugs to relax airway smooth muscle at doses producing minimal activation of the heart.

Many of these drugs have been found useful in the treatment of asthma, bronchospasm, chronic obstructive pulmonary disease (COPD), and premature labor. Side effects result from excessive activation of β receptors and include tachycardia, arrhythmias, myocardial ischemia, tremor, and increased bronchial hyperreactivity.

α_1-Selective Adrenergic Receptor Agonists

The major effects of phenylephrine, mephenteramine, metaraminol, midodrine, and other α_1-selective adrenergic receptor agonist drugs are to activate α adrenergic receptors in vascular smooth muscle. This results in an increase in peripheral resistance and elevation of blood pressure. These drugs possess some selectivity of α_1 receptors over α_2 or β receptors. Most can be given orally and can be used to treat hypotension, including orthostatic hypotension or shock. They can also be used as nasal decongestants and as mydriatics (dilation of the pupil) (e.g., phenylephrine). Midodrine is an orally effective α_1 agonist and is a prodrug; it is converted to an active metabolite, desglymidodrine, and is used to treat orthostatic hypotension.

α_2-Selective Adrenergic Receptor Agonists

Clonidine, apraclonidine, brimonidine, guanffacine, guanabenz, methyldopa, and tizanidine are α_2-selective adrenergic receptor agonists. Although the principal receptor in vascular smooth muscle is the α_1 type, there are α_2 adrenergic receptors as well, which when activated promote vasoconstriction. Therefore, it is somewhat ironic that many of the α_2-selective adrenergic receptor agonist drugs are used to treat hypertension. The ability to lower blood pressure, however, is the result of activation of α_2 adrenergic receptors in the cardiovascular control centers in the CNS. Activation of α_2 receptors at these sites decreases the outflow of sympathetic nerve activity from the CNS. There is some evidence that clonidine's sympathoinhibitory action is

partly mediated by an action on imidazoline receptors in the brain stem.

A second major use of some of these drugs is to decrease the production of aqueous humor, thus reducing intraocular pressure in the eye in patients with wide-angle glaucoma (praclonidine and brimonidine). These drugs also have the ability to decrease the evoked release of sympathetic transmitters from neurons by acting on the inhibitory α_2 autoreceptors or heteroreceptors on sympathetic or other neurons.

Indirect-Acting or Mixed-Acting Sympathomimetics

Indirect-acting or mixed-acting sympathomimetics (e.g., amphetamine, methamphetamine, methylphenidate, dexmethylphenidate, and ephedrine) have powerful CNS stimulant actions (especially amphetamine and methamphetamine) in addition to peripheral α and β actions. Most can be given orally and have long duration of actions relative to catecholamines. Amphetamines increase both systolic and diastolic blood pressure, directly stimulate the heart, and on occasion may cause a reflex slowing of the heart. CNS effects include wakefulness, alertness, decreased feeling of fatigue, elevation of mood, and increased initiative, self-confidence, and ability to concentrate, in addition to elation, euphoria, and increase in motor and speech activities. Prolonged use may be followed by depression and fatigue.

These drugs are indirectly acting or mixed acting, meaning they release norepinephrine (from sympathetic neurons) and a variety of biogenic amines (from neurons in the CNS), including norepinephrine, dopamine, and serotonin. Side effects are a result of the extension of their pharmacological effects and include increases in blood pressure, anginal pain, and arrhythmias, as well as other cardiovascular effects. CNS effects may include restlessness, dizziness, tremor, hyperactive reflexes, talkativeness, tenseness, irritability, weakness, insomnia, fever, and euphoria. More serious effects include confusion, aggressiveness, anxiety delirium, paranoid hallucinations, panic states, and suicidal or homicidal tendencies. Chronic use can also lead to tolerance and dependence.

Therapeutic uses include attention-deficit/hyperactivity disorder (ADHD) and narcolepsy. Although amphetamine and related drugs can reduce appetite, the use in treating obesity is controversial and considered inappropriate. Many herbal preparations contain ephedrine (ma huang, ephedra) and are used around the world. The safety of many of these products is of concern and therefore the United States Food and Drug Administration has banned the sale of dietary supplements containing ephedra in the United States effectively from April 2004.

Adrenergic Receptor Antagonists

Adrenergic receptor antagonists are drugs that bind to α or β adrenergic receptors and prevent or block the responses produced by catecholamines or sympathomimetic drugs. Almost all of these drugs are competitive antagonists with the exception of phenoxybenzamine. By definition these drugs bind to the receptor site (affinity) but cannot produce a biological response. However, if an endogenous transmitter such as norepinephrine is present, the adrenergic receptor antagonist will result in a physiological response. The pharmacological response to α or β adrenergic receptor antagonists depends upon a composite of (1) degrees of existing sympathetic tone in the cell or tissue at the time of blockade (amount of norepinephrine interacting with the receptor, (2) the type of receptor inhibited (e.g., α_1, β_1), (3) degree and duration of blockade, (4) other drugs present, and (5) actions of the blocker not associated with actions on adrenergic receptors. As with sympathomimetics, antagonists may be selective for one type of receptor (e.g., α_1, β_1) or may be nonselective (β_1 and β_2). Figure 3 classifies numerous adrenergic receptor antagonists as to the receptor or receptor subtype with which they interact.

α Adrenergic Receptor Antagonists

As depicted in Figure 3, α adrenergic receptor antagonists are nonselective (phenoxybenzamine, phentolamine), meaning that they block both α_1 and α_2 adrenergic receptors or that they are α_1 selective (prazosin) or α_2 selective (yohimbine). Some drugs have selectivity for the α_1 or α_2 subtypes (α_{1A}, α_{1B}, α_{2A}, etc.) but such a discussion is beyond the focus of the present article.

A summary of the effects of α adrenergic antagonists is depicted in Table 2. Some of the most important effects of α adrenergic receptor antagonists are on the cardiovascular system. Blockade of α_1 adrenergic receptors inhibits vasoconstriction, and therefore vasodilation in both arteriolar resistance vessels and veins may occur. The fall in blood pressure depends upon the degree of sympathetic tone (higher in erect vs. supine position). For most α blockers the decrease in blood pressure is opposed by baroreceptor reflexes that cause an increase in heart rate and cardiac output.

Blockade of α_2 receptors interferes with prejunctional modulation of norepinephrine release by inhibitory α_2 receptors. The result is enhanced norepinephrine release. As mentioned previously, activation of α_2 receptors in central cardiovascular control sites leads to a decrease in sympathetic outflow and a decrease in blood pressure. Blockade of such receptors will result in an increase in sympathetic

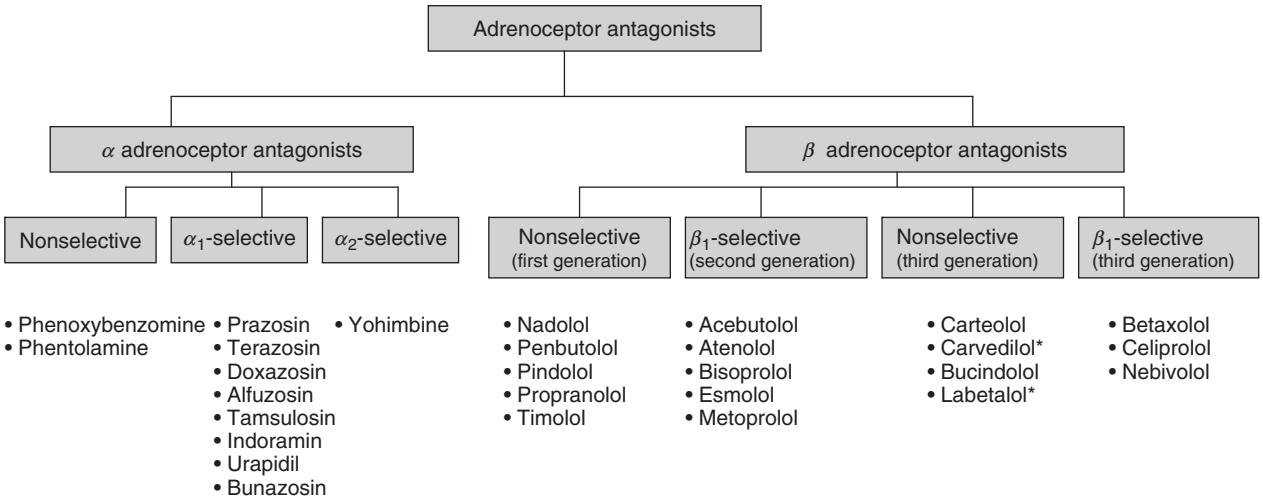

Figure 3 Classification of adrenergic receptor antagonists. The nonselective third-generation β adrenoceptor antagonist drugs carvedilol and labetalol also block α_1 receptors. Adapted from Westfall TC and Westfall DP (2006) In: Brunton L, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th edn., ch. 10. New York: McGraw-Hill.

outflow and an increase in sympathetically mediated responses (e.g., increased blood pressure).

Phenoxybenzamine is unlike other α adrenergic receptor antagonists in that the tertiary amine cyclizes to an ethylenimonium intermediate. The three-member ring breaks and forms highly reactive carbonium ions that can alkylate various sites. This forms a covalent bond with the α receptors and this produces a nonequilibrium rather than a competitive blockade. Other names given to this type of blockade are 'unsurmountable' or 'competitive irreversible.' Once the blockade has fully developed, it cannot be surmounted by the presence of an α agonist, as takes place with competitive antagonists. The blockade has a long duration of action. Phenoxybenzamine is a very reactive drug and can also block other types of receptors, such as cholinergic, serotonergic, and histaminergic receptors. In addition, it can block norepinephrine uptake into neurons (uptake 1; NET) and extraneuronal sites (uptake 2). The most useful α adrenergic receptor antagonists clinically are the α_1 antagonists that are used to treat essential hypertension. In addition, contractile α_1 receptors are present in the prostate gland and in benign prostatic hyperplasia (BHP) and impede the outflow of urine from the bladder through the urethra. The α_1 adrenergic antagonists are used to relax the prostate in such situations, giving symptomatic relief. In many cases, α_1 adrenergic receptor antagonists are used in combination with 5α-reductase inhibitors.

β Adrenergic Receptor Antagonists

The β adrenergic receptor antagonists are among the most important drugs developed over the past 40–50 years because of their efficacy in treating numerous pathological conditions, including hypertension, ischemic heart disease, heart failure, and certain arrhythmias, to name but a few. Since the development of the first clinically useful beta blocker, propranolol, in the 1960s, there have been numerous additional drugs developed (listed in Figure 3). As with α adrenergic receptor antagonists, drugs are nonselective (blocking β_1 and β_2 receptors) or β_1 selective. In addition to being nonselective or β_1 selective, some beta blockers also have other actions, such as a membrane-stabilizing (local anesthetic) action, or intrinsic sympathomimetic activity. In the latter case such agents are partial agonists, meaning that in the absence of other β agonists they can activate β receptors, but in the presence of other agonists they behave like antagonists. Moreover, the so-called third-generation drugs are also vasodilators and produce this effect by a variety of mechanisms, including β_2 agonism, α_1 receptor antagonism, production of nitric oxide, Ca^{2+} entry blockade, or opening K^+ channels (Table 3).

Pharmacological Actions and Side Effects

As is the case for α adrenergic receptor antagonists, the pharmacological properties of β receptor antagonists can be understood largely from knowledge of the responses elicited by the receptors in various tissues and the activity of the sympathetic nerves that innervate these tissues (Table 1). A summary of β adrenergic antagonists is included in Table 4.

Beta blockers have prominent effects on the heart, including decreasing heart rate, decreasing myocardial contractility, and decreasing cardiac output. The

Table 2 Summary of α adrenoceptor antagonists

Class	*Drug*	*Prominent pharmacological action*	*Principal therapeutic application*	*Untoward effects*	*Comments*
α Blockers Nonselective (classical α blockers)	• Phenoxybenzamine • Phentolamine • Tolazoline	• Decrease in PVR[a] and blood pressure • Venodilation	• Treatment of catecholamine excess (e.g., pheochromocytoma)	• Postural hypotension • Failure of ejaculation	• Cardiac stimulation due to initiation of reflexes and due to enhanced release of NE via α_2 receptor blockade
α_1 Selective	• Prazosin • Terazosin • Doxazosin • Trimazosin • Alfuzosin • Tamsulosin	• Decrease in PVR and blood pressure • Relax smooth muscles in neck of urinary bladder and in prostate	• Primary hypertension • Increased urine flow in BPH[b]	• Postural hypotension when therapy instituted	• Phenoxybenzamine produces long-lasting α receptor blockade at high doses; can block neuronal and extraneuronal uptake of amines • Prazosin and related quinazoline compounds are selective for α_1 receptors but not among subtypes • Tamsulosin exhibits some selectively for α_{1A} receptors

[a]Peripheral vascular resistance.
[b]Benign prostatic hyperplasia.
Adapted from Westfall TC and Westfall DP (2006) In: Brunton L, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th edn., ch. 10. New York: McGraw-Hill.

Table 3 Third-generation β receptor antagonists with additional cardiovascular actions: Proposed mechanisms contributing to vasodilation

Nitric oxide production	*β_2 receptor agonism*	*α_1 receptor antagonism*	*Ca^{2+} entry blockade*	*K^+ channel opening*	*Antioxidant activity*
Celiprolol[a]	Celiprolol[a]	Carvedilol	Carvedilol	Tilisolol[a]	Carvedilol
Nebivolol	Carteolol	Bucindolol[a]	Betaxolol		
Carteolol	Bopindolol[a]	Bevantolol[a]	Bevantolol[a]		
Bopindolol[a]		Nipradilol[a]			
Nipradilol[a]		Labetalol			

[a]Not currently available in the United States; most are under investigation for use in the United States.
Adapted from Westfall TC and Westfall DP (2006) In: Brunton L, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th edn., ch. 10. New York: McGraw-Hill.

short-term effect of beta blockers is to increase peripheral resistance; however, long-term use leads to a decrease in or produces no change in total peripheral resistance. These drugs have significant effects on cardiac rhythm and automaticity. Due to activation of β receptors (see Table 1), beta blockers can also antagonize other effects, such as blocking bronchiolar smooth muscle, leading to an increase in airway resistance, reduction of intraocular pressure, inhibition of sympathetic nervous system, stimulation of lipolysis in fat cells and glycogenolysis in the liver, and antagonism of the sympathetic nerve-induced release of renin in the kidney and thus angiotensin formation in the circulation. Many of these effects are considered side effects when beta blockers are being used clinically.

Therapeutic Uses

The β adrenergic receptor antagonists are very useful clinically and have been used to treat a wide variety of conditions, including angina pectoris (ischemic heart disease), essential hypertension, cardiac arrhythmias, idiopathic hypertrophic subaortic stenosis, congestive heart failure, glaucoma, hyperthyroidism, surgical removal of pheochromocytoma, nonparkinsonian tremor, and migraine (prophylaxis), and are used in a wide variety of anxiety situations (e.g., stage fright) and in alcohol and sedative–hypnotic drug withdrawal.

Angina pectoris (ischemic heart disease) Angina pectoris is a syndrome which produces a sensation of strangulation, squeezing, and pressure in the chest region. It is caused by inadequate coronary blood flow that fails to meet oxygen demands of the heart tissue and is associated with acute myocardial ischemia. The most common form is chronic stable angina caused by fixed narrowing or stenosis of the coronary blood vessels. It is often precipitated by exertion, exercise, or stress. Beta blockers are useful in the long-term and prophylactic treatment of angina. These drugs reduce the frequency of angina episodes and improve exercise tolerance. These effects are due to β adrenergic receptor blockade, resulting in a decrease in cardiac work and a reduction in oxygen demand.

Essential hypertension The pathophysiological mechanisms causing essential hypertension are obscure. It is well established that left untreated it can lead to morbid events such as coronary heart disease, renal disease, or stroke. Beta blockers have been proved to be effective in treating essential hypertension and are well tolerated. They reduce both supine and standing blood pressure during long-term treatment, although they are not usually first-choice drugs. They are valuable when used together with diuretics, angiotensin-converting enzyme inhibitor, angiotensin receptor antagonists, and other vasodilators. The precise mechanisms of action of beta blockers in treating essential hypertension are unclear. Suggested mechanisms include reduction in cardiac output, inhibition of renin release and thus angiotensin II production, sympathoinhibition via a CNS mechanism, or blockade of prejunctional excitatory β receptors on sympathetic neurons.

Cardiac arrhythmias The β adrenergic receptor antagonists are effective in treating both supraventricular and ventricular arrhythmias. Indications for use include catecholamine-induced arrhythmias as well as atrial flutter, atrial fibrillation, paroxysmal supraventricular tachycardia, and selected ventricular arrhythmia. These drugs are used to control sinus tachycardia and frequent premature ventricular contractions in patients with catecholamine-secreting tumors and pheochromocytoma, as well as in the surgical management of the tumor.

Glaucoma Beta blockers were found serendipitously to reduce intraocular pressure in glaucoma, and work following either systemic or topical application. They appear to work by reducing

Table 4 Summary of β adrenoceptor antagonists

Class	*Drug*	*Prominent pharmacological action*	*Principal therapeutic applications*	*Untoward effects*	*Comments*
β blockers Nonselective (first generation) β_1 Selective (second generation) Nonselective (third generation) vasodilators β_1 Selective (third generation) vasodilators	• Nadolol • Pindolol • Propranolol • Timolol • Bisoprolol • Metoprolol • Carvedilol • Labetalol • Celiprolol	• Decrease in heart rate • Decrease in contractility • Decrease in cardiac output • Slowed conduction in atrial and AV node • Increased refractory period-AV node • Bronchoconstriction • Prolonged hypoglycemia • Decrease in plasma free fatty acids • Reduction in HDL cholesterol • Increase in LDL cholesterol and triglycerides • Hypokalemia (membrane-stabilizing effect); intrinsic sympathomentic activity (vasodilation)	• Angina pectoris • Hypertension • Cardiac arrhythmias • Congestive heart failure • Pheochromocytoma • Glaucoma • Hypertropic-obstructive cardiomyopathy • Hyperthyroidism • Migraine prophylaxis • Acute panic symptoms • Substance abuse withdrawal • Variceal bleeding in portal hypertension	• Bradycardia • Negative inotropic effect • Decrease in cardiac output • Bradyarrhythmias • Reduction in AV conduction • Bronchoconstriction • Fatigue • Sleep disturbances (insomnia; nightmares) • Prolongation of hypoglycemia • Sexual dysfunction in men • Drug interactions	• Pharmacological effects depend largely on degree of sympathoadrenal tone • Bronchoconstriction (concern only in asthmatics and chronic obstructive pulmonary disease) • Hypoglycemia (concern in hypoglycemics and diabetics) • Membrane-stabilizing effect (propranolol; acebutolol; carvedilol and betaxolol only) • Intrinsic sympathomimetic activity (strong for pindolol; weak for penbutolol; carteolol; labetalol and betaxolol) • Vasodilation-seen in third-generation drugs; multiple mechanisms (α_1 antagonism, β_2 agonism, release of NO, Ca^{2+} channel blockade, opening of K^+ channels, others)

Adapted from Westfall TC and Westfall DP (2006) In: Brunton L, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th edn., ch. 10. New York: McGraw-Hill.

the production of aqueous humor. They are often first-choice drugs and are devoid of the side effects of blurred vision and have little effect on pupil size or accommodation.

Congestive heart failure Although it is a common observation that acute administration of beta blockers can worsen or even precipitate congestive heart failure in compensated patients, numerous well-designed randomized clinical trials have demonstrated that certain beta blockers are a highly effective treatment for patients with all grades of heart failure secondary to left ventricular systolic dysfunction. The drugs have been shown to improve myocardial function and to improve quality of life and prolong life. Long-term beta blockade has been shown to reduce cardiac volume, myocardial hypertrophy, and filling pressure, and to increase ejection fraction. Large trials have been carried out with carvedilol, bisoprolol, metoprolol, xamoterol, bincindolol, betaxolol, nebivolol, and talinolol.

Other uses Beta blockers have also been shown to be effective in reducing the clinical symptoms of tachycardia, tremors, and anxiety often seen in hyperthyroidism, alcohol and sedative–hypnotic drug withdrawal, and acute situational anxiety (stage fright). They are also useful in the prophylaxis of migraine headache, nonparkinsonian tremor, and urinary incontinence.

See also: Brain Adrenergic Neurons; Norepinephrine: CNS Pathways and Neurophysiology; Norepinephrine: Adrenergic Receptors; Stress and Parasympathetic Control; Sympathetic Noradrenergic and Adrenomedullary Hormonal Systems in Stress and Distress; Sympathetic Nervous System.

Further Reading

Alward WL (1998) Medical management of glaucoma. *New England Journal of Medicine* 339: 1298–1307.

Cazzola M, Matera MG, and Donner CF (2006) Inhaled β_2-adrenoceptor agonists: Cardiovascular safety in patients with obstructive lung disease. *Drugs* 65: 1595–1610.

Cleland JG (2003) β-Blockers for heart failure: Why, which, when and where. *Medical Clinics of North America* 87: 339–371.

Coats AJS (2006) β-Adrenoceptor antagonists in elderly patients with congestive heart failure. *Drugs & Aging* 23: 93–99.

Goldsmith DR and Plaskser GL (2005) Duxazosin gastrointestinal therapeutic system. A review of its use in benign prostatic hyperplasia. *Drugs* 65: 2037–2047.

Ioannides LL, Proietto J, and McNeil JJ (2005) Pharmacotherapy of obesity. *Drugs* 65: 1391–1418.

Matera LG and Cazzola M (2007) Ultra-long-acting β-adrenoreceptor agonists. An emerging therapeutic option for asthma and COPD? *Drugs* 67: 503–515.

Missale C, Nash SR, Robinson SW, et al. (1998) Dopamine receptors: From structure to function. *Physiological. Reviews* 78: 189–225.

Robidoux J, Martin TL, and Collins S (2004) β-Adrenergic receptors and regulation of energy expenditure, a family affair. *Annual Review of Pharmacology and Toxicology* 44: 297–323.

Rosei EA and Rizzoni D (2007) Metabolic profile of nebivolol, a β-adrenoceptor antagonist with unique characteristics. *Drugs* 67: 1097–1107.

Ruffolo RR Jr. (1987) The pharmacology of dobutamine. *American Journal of the Medical Sciences* 294: 244–248.

Santillo VM and Lowe FC (2006) Treatment of benign prostatic hyperplasia in patients with cardiovascular disease. *Drugs & Aging* 23: 795–805.

Toda N (2003) Vasodilating β-adrenoceptor blockers as cardiovascular therapeutics. *Pharmacology & Therapeutics* 100: 215–234.

VanDijk MM, delaRosette JMCH, and Michel MC (2006) Effects of α_1-adrenoceptor antagonists on male sexual function. *Drugs* 66: 287–301.

Westfall TC and Westfall DP (2006) Adrenergic agonists and antagonists. In: Brunton L, Lazo JS, and Parker KL (eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics,* 11th edn., pp. 237–295. New York: McGraw-Hill.

Synapse Formation: Competition and the Role of Activity

L Cancedda and M-M Poo, University of California at Berkeley, Berkeley, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Synaptic competition is a cellular process by which the presence of one synapse affects the stability or survival of other synapses on the same postsynaptic cell. Synapse elimination refers to synapse loss due to either a low 'intrinsic merit' of the synapse for survival or to its failure in winning synaptic competition with other synapses on the same cell. Such competition and elimination can be driven by either neuronal/synaptic activity or other activity-independent processes. The idea that the strength of synaptic connections between neurons may be modified can be traced back to Ramon y Cajal, who proposed that such modification serves as a cellular mechanism for learning and memory. That memory formation involves making and breaking existing synaptic connections has been a popular idea over the past century, but solid experimental evidence in support of this idea has been elusive. Recent morphological studies of synapse stability in the adult rodent brain have revealed rather limited synaptic remodeling under normal conditions. In contrast, initial synaptic connections established early in the developing nervous system undergo substantial remodeling, with some connections stabilized and others eliminated, as a result of experience. This developmental remodeling of connectivity often involves cooperative and competitive interactions between converging synapses on the postsynaptic cell and in many cases depends on the pattern of electrical activity. We here summarize the evidence of activity-dependent synaptic competition and elimination in various brain regions, together with potential underlying cellular mechanisms.

Visual System

Activity-dependent synaptic competition and elimination in the central nervous system was first indicated by work of Hubel and Wiesel on the development of ocular dominance columns (ODCs) in the visual cortex. Cortical neurons preferentially responding to one eye or the other are normally found to be segregated into alternating columns in the primary visual cortex, representing sorting of geniculocortical projections serving two eyes during postnatal development. The sorting process occurs prior to eye opening and is highly sensitive to visual experience. Depriving visual input to one eye of a newborn cat or monkey by suturing the eyelid during a critical period after birth leads to retraction of geniculocortical projections serving that eye. Interestingly, the consequences of visual deprivation during early postnatal period are much more severe in cats subjected to monocular deprivation than binocular deprivation, suggesting that competition driven by activity between two eyes, rather than the reduced activity in one eye, is responsible for sorting geniculocoritical projections. The ODCs develop by a progressive segregation of initially overlapping geniculocortical projections serving two eyes. While the initial alternating pattern of projections is established by mechanisms independent of the retina activity, pruning of extensive overlapping projections is likely to be driven by activity, including spontaneous or visually evoked retinal activity, because blocking the firing of retinal ganglion cells (RGCs) in both eyes prevents complete ODC formation.

Early in development there is also substantial overlap of retinogeniculate projections from the two eyes in the lateral geniculate nucleus (LGN). Segregation of these projections into eye-specific layers occurs before the onset of vision, but at a time when spontaneous activity waves are prominent in the retina. Blocking all activity in both eyes prevents segregation of retinogeniculate projections and can desegregate already existing eye-specific projections later in development, although it is unclear whether the specific pattern of the waves is required for eye-specific segregation. When the activity is altered in one eye, the more active eye acquires more synaptic territory than the less active eye, suggesting activity-mediated competition. Furthermore, even after eye-specific segregation is completed, activity in one eye can further prune retinogeniculate projections by reducing the convergence of RGC inputs onto a single LGN neuron from 12–20 to 1. This within-eye pruning based on activity-dependent competition among RGCs faciliates sharpening of the receptive field of LGN neurons. Such activity-dependent competition may be attributed to the competition in the growth or stability of RGC axonal arbors. In the developing optic tectum of zebra fish, reducing neuronal and synaptic activity in a subset of RGCs leads to suppression of their axon growth and branching and this suppression is relieved by blocking the activity of nearby RGC axons as well.

The successive levels of the mammalian visual system are organized into retinotopic maps that preserve an orderly representation of visual inputs at the retina, through topographically precise retinogeniculate,

retinocollicular, and geniculocoritical projections. Formation of these maps requires specific patterns of spontaneous activity in the retina because disrupting these waves affects map formation at all levels. While the initial projections may establish a crude retinotopic map via axon guidance based on molecular cues, interfering with both axon guidance cues and spontaneous wave activity in the same animal results in a dramatic cumulative effect in disrupting the map in the superior colliculus. Finally, the development of retina circuitry itself also depends on activity. Depriving the retinal activity by blocking spontaneous activity or dark rearing blocks both the normal maturational loss of ON–OFF responsive RGCs and the pruning of dendrites at the stratified ON and OFF layers of the inner plexiform layer, although it is unclear whether dendritic pruning in this layer results from competitive interaction among bipolar inputs.

Somatosensory System

Somatosensory projections in the brain exhibit a somatotopic map, with axons of peripheral receptor sheets projecting in an orderly manner onto central brain structures. The most intensively studied system is the rodent trigeminal pathway where the patterned array of the whisker is replicated in the patchy distribution of afferents and the modular organization of their postsynaptic counterparts along the pathway from the periphery to the primary somatosensory cortex. The whisker-related patterns are first established in the brain stem nuclei, then in the ventroposteromedial nucleus of the thalamus, and finally in the somatosensory cortex ('barrels'), where neurons respond best or exclusively to deflection of the corresponding facial whisker. These somatotopic maps emerge during early development in ascending order along the neuroaxis, with a sequence that includes afferent fibers segregation before rearrangement of their target neurons in discrete zones along the pathway. It is generally believed that the initial crude topographic projection of the afferent fibers is independent of sensory experience, but the presence of segregated afferent fibers and postsynaptic glutamate receptor activities is required for the subsequent parcellation of their postsynaptic targets.

That cortical map formation depends on a competitive process is suggested by the finding that damage of the branch of the trigeminal nerve supplying the whisker pad prior to birth results in a reduction of the cortical representation of the whiskers and a concomitant increase in the representation of other peripheral receptor surfaces. Whether and how such competition occurs through synaptic competition and elimination are unknown. Synapse elimination has been observed in the ventral posteromedial thalamic nucleus of young animals, where multiple afferents on each neuron are reduced to one or two afferents as the animal matures. However, this synapse elimination and remodeling occur even in animals deprived of sensory experience from birth, indicating that the process is independent of sensory experience, although spontaneous neuronal activity may still be required.

Morphological organization of somatosensory cortical neurons into barrels during development depends on signals conveyed by invading thalamic axons and sensory activity. Indeed, primordial visual cortex transplanted into the neonatal somatosensory cortex forms barrels when invaded by axons from somatosensory thalamic nuclei, and genetic manipulations that interfere with the appropriate segregation of these axons disrupt barrel formation. Moreover, that sensory peripheral sensory signals may be instructive in sculpting the somatosensory cortical map is also evidenced by lesion or genetic studies showing that altering the number of functioning whiskers leads to shrinkage, expansion, or addition of barrels. However, such aberrant barrel formation may result from sprouting and retraction of afferents and dendrites due to neuronal degeneration and alteration in the number of afferents, involving competitive synapse formation, rather than activity-dependent synaptic competition. Notably, changes in whisker use drive functional changes in the barrel neurons without affecting anatomical appearance of barrels, suggesting no gross reorganization of synaptic connections. Nevertheless, sensory deprivation results in a specific loss of GABAergic synapses and impairment of secondary dendritic branches in the barrel cortex. Thus, activity may lead to synapse elimination of at least a subset of synapses in the somatosensory cortex, presumably through a competitive process.

Auditory System

Sound entering the ear stimulates cochlear hair cells, which make synaptic connection with spiral ganglion cells that give rise to the primary auditory nerve for carrying auditory information into the brain stem. From there the signals converge in the inferior colliculus that sends projections to the thalamus, which further relays auditory signals to the auditory cortex. In these relay areas, neurons are arranged in a topographic manner according to the sound frequencies to which they are most sensitive. Although the topography of these connections is apparent early in development, the precision of the map is refined later in development through an activity-dependent process.

Developmental pruning of both axonal arbors and dendritic branches has been widely observed in the auditory system. Cochlear nerve axons and their target neurons in nucleus magnocellularis (NM) undergo extensive parallel structural transformations involving pruning of cochlear axonal arbor, massive reduction of dendrites in NM neurons, and elimination of polyneuronal innervation. Since otocyst removal has no effect on the extent, timing, and pattern of dendritic loss, this extensive synapse elimination is independent of the sensory activity. Interestingly, after this early massive remodeling, NM neurons undergo further dendritic growth before maturation in a sensory activity-dependent manner.

In contrast to that in the NM, developmental remodeling of axons and dendrites in the nucleus laminaris and superior olivary nucleus (SON) after the onset of hearing appears to be activity dependent. Cochlea removal or blockade of glycinergic transmission impairs remodeling of axonal and dendritic morphology in SON. Since most of the remodeling occurs after the onset of hearing, acoustically evoked activity is likely to be involved, although spontaneous activity may also contribute. Indeed, correlated spontaneous activity is present in the embryonic brain stem and auditory nerve. The spatiotemporal pattern of spontaneous firing could provide developmental cues for the spatial ordering of auditory projections, as suggested by the presence of a systematic relationship between the rate of rhythmic bursting and the tonotopic location in the chick. Such activity-dependent remodeling of connectivity may contribute to the tonotopic map refinement at many different levels in the auditory system.

After the onset of hearing, the auditory cortex undergoes a transition from a tonotopic map dominated by broadly tuned, high-frequency-selective neurons to the adult tonotopic map consisting of neurons that represent the full spectrum of acoustic inputs. Sensory-evoked activity is responsible for this transition. Early acoustic environment is critical for the maturation of tonotopic maps, because exposing rat pups to pulsed white noise or rearing them in continuous moderate-level noise impairs the emergence of adultlike tonotopic map, whereas exposure to pulsed tones of specific frequencies results in accelerated emergence and expansion of auditory cortex representations of those frequencies. This activity-dependent remodeling of cortical maps is likely to involve synaptic competition and elimination, although whether functional refinement of the map directly reflects structural reorganization of synaptic connectivity remains to be determined.

Olfactory System

The olfactory sensory neurons in mammals express only one of about 1000 odorant receptor genes, and neurons expressing a given receptor are randomly dispersed within one of four broad zones in the olfactory epithelium. The axons of these sensory neurons converge upon spatially conserved glomerulus within the olfactory bulb. The topographic mapping between sensory neurons and specific glomeruli may depend on the expression of specific molecules along their projection pathways or in themselves. There is evidence, however, that activity in these olfactory neurons may also play a role. Although the patterns of axon convergence in the bulb are largely intact in mice lacking functional olfactory cyclic nucleotide-gated channels, hence no odorant-evoked activity occurs in these neurons, noncorrelated spontaneous activity may still be required for the correct mapping process. Indeed, sensory map is not affected when conditional expression of tetanus toxin light chain inhibits synaptic transmission in the majority of olfactory sensory neurons. However, inhibition of synaptic release in a small subpopulation of neurons expressing the P2 receptor results in correct targeting of the sensory axons initially, but the P2 glomerulus is not maintained and P2 neurons ultimately diminish. Preventing excitation of the neuron has a similar effect on the formation of the olfactory map. Thus, spontaneous neuronal activity may play a role in pruning or stabilization of axon terminals of olfactory neurons on their glomerular target cells, but there is little evidence that synaptic competition is involved in olfactory map formation.

Cerebellum

The two main afferent systems in the cerebellar cortex are the climbing fibers (CF) originating from the inferior olivary nucleus and the mossy fibers (MF) from various nuclei in the spinal cord, brain stem, and deep cerebellar nuclei. Each CF directly contacts the proximal dendritic compartment of a single Purkinje cell (PC), whereas the MFs influence PCs indirectly through granule cells, whose axons form the parallel fibers (PFs) that synapse onto the dendrites of many PCs. There is evidence for a complex topographic map of these afferents fibers. For example, cutaneous inputs carried by CFs are topographically organized to form a map of peripheral body, with CF axonal arbors in register with cortical parasagittal bands of chemically heterogeneous PCs. The formation of these precise topographic maps

involves both activity-independent and dependent steps. First, positional information shared between CFs and PCs during embryonic development provides the molecular code for the formation of coarse-grained maps independent of neuronal activity. Activity-dependent mechanisms are later required for the transition to a fine-grained map, by pruning CF terminal arbors on each PC from multiple to single CF innervation. This pruning involves strengthening one CF while weakening all other CFs during the first postnatal week, leading to the elimination of the latter. Moreover, alteration of the temporal pattern of CF activities specifically during development impairs such CF elimination *in vivo,* suggesting an activity pattern-dependent synaptic competition.

In addition to potential competition among homologous CFs, there is also heterologous competition between CFs and PFs that results in their segregation into different dendritic domains of each PC. Weakening of the CF input leads to the reduction of its dendritic territory and concomitant strengthening and expansion of the PF input, and vice versa. Furthermore, elimination of CFs depends on the activity of developing PF–PC synapses, because CF elimination is affected by reducing PF inputs, through granule cell degeneration, impairment of granule cell function, or genetic manipulation of PF–PC synapse formation or efficacy. The similarity in the consequence of reducing CF and PF activities in CF elimination suggests that similar mechanisms may underlie synaptic competition/elimination among homologous versus heterologous inputs.

Autonomic Ganglia

Ganglionic cells in the autonomic nervous system are innervated by preganglionic neurons of the spinal cord or brain stem and send projections to target tissues via spinal nerves to control involuntary functions of the body. In mammals, characteristic patterns of sympathetic and organ responses elicited by the activity of individual spinal axons are due to mapping between specific spinal segments and peripheral targets. This mapping requires a stereotyped and selective innervation of ganglion cells by preganglionic axons, with each ganglion cell innervated by one or few axons from specific contiguous spinal cord segments. During development there is a transition from initially exuberant innervation of each ganglion cell by many preganglionic axons to innervation by only one or a few axons. Since the spinal segment responsible for activating the mature and neonatal ganglion cell is the same, developmental synapse elimination involves reduction of preganglionic axons from the same spinal segment, suggesting competition occurs among axons of nearby spinal neurons. Furthermore, transection of a portion of preganglionic nerve innervating a developing ganglion leads to sprouting of residual preganglionic axons and partial restoration of original multiplicity of innervation. These results are all consistent with the idea that synaptic elimination is not simply due to the intrinsic merit of the input, but involves competition among inputs, presumably for a postsynaptic factor of limited supply, for example, locally secreted trophic factor.

The competitive interaction that determines the final number of preganglionic axons converging upon a single ganglion cell depends on the proximity of competing synapses, with each surviving synaptic terminal claiming a certain territory on the dendritic or somatic surface and neurons with more extensive dendritic arbors receiving more axons. This distance-dependent synapse competition can be explained by a limited amount of available postsynaptic factors. This synapse competition appears to be activity dependent, because among converging inputs to a single ganglion neuron, strong synapses become further strengthened and weak synapses further weakened during synapse elimination. Moreover, the synaptic strength for each synaptic input of a multiply innervated cell is, on average, weaker than that of a singly innervated cell, suggesting that the total synaptic strength of a postsynaptic ganglion cell is conserved, due to a limited amount of a synapse-related factor.

Neuromuscular Junction

Each muscle fiber in the neonatal animal is multiply innervated by axon collaterals of several motor neurons, but becomes singly innervated during early postnatal life. This process depends on synapse elimination involving withdrawal of a subset of nerve terminals of each motor neuron innervating a given muscle (i.e., reduction of the size of motor units) rather than reduction in the number of motor neurons innervating a muscle. Partial denervation of the muscle at birth results in the retention of the large motor unit size without substantial collateral motor axon sprouting. Synapse elimination is competitive rather than a random process of withdrawal, since muscle fiber without a single axon is never observed. The elimination is also activity dependent. Blocking motor neuronal activity prevents elimination, while elevating activity accelerates it. Importantly, when the relative synaptic efficacy of two competing axons at a single neuromuscular junction is impaired by genetic deletion of acetylcholine in one axon, the latter loses the competition, suggesting that the strength of the synapse is predictive of the outcome of the competition.

Mechanisms of Activity-Driven Synapse Competition

The Hebb's Rule for Synapse Competition

Hebb postulated that strengthening of a synapse might be achieved by repetitive presynaptic activation that leads to postsynaptic firing. This postulate was later transformed into a simple rule – coincident pre- and postsynaptic activity leads to synapse strengthening and stabilization. To account for the finding of Hubel and Wiesel on activity-dependent remodeling of connectivity in the developing visual system, Stent further extended the Hebb's rule by assuming that noncoincident pre- and postsynaptic activity leads to synapse weakening and elimination. Modeling studies showed that such correlation-based Hebb's rule could explain activity-dependent refinement of developing visual circuits. In the past decade, a temporally specific form of Hebb's rule has been proposed, based on findings of spike-timing-dependent synaptic plasticity in a variety of systems. The temporal order in the spiking of pre- and postsynaptic neurons was shown to be critical for synaptic modification, in addition to the extent of coincidence in spiking: 'Pre-before-post' results in strengthening, and 'post-before-pre' leads to weakening of the synapse. This spike-timing-dependent plasticity offers an element of causality in the activity-induced synaptic competition: inputs that contribute to (and cause) the postsynaptic spiking are advantageous in synaptic competition over those inputs arriving after postsynaptic spiking has just occurred. Importantly, experimental evidence for the validity of various forms of the Hebb's rule mainly came from studies of activity-induced functional changes of synaptic efficacy, for example, long-term potentiation (LTP) or long-term depression (LTD), rather than changes in the morphological connectivity.

Importance of Temporal Pattern of Activity

The importance of the pattern of activity in synapse competition has been demonstrated mainly in the development of ocular dominance and orientation selectivity of the primary visual cortex. Rearing kittens with induced squint (strabismus), which alters the pattern but not the absolute level of activity, results in striking changes in the binocular property of cortical cells, reflecting altered synapse competition of geniculocortical inputs. Artificially imposing synchronous activity on optic nerves from the two eyes prevents segregation of thalamocortical projections into ODCs, whereas asynchronous activity allows segregation. Similarly, synchronous activation of optic nerves blocks the development of topographic maps in the optic tectum and reduces orientation selectivity in the cortex. Spike-timing-dependent synaptic modification provides a natural basis for such pattern-dependent competition. In the developing *Xenopus* visual system, spike-timing-dependent induction of LTP and LTD has been demonstrated, but whether such persistent changes in functional efficacy of synapse is causally related to structural changes in connectivity remains unknown.

Contribution by GABAergic Inhibition

Activity in the brain depends on a proper balance of excitation and inhibition. It is thus not surprising that GABAergic activity plays a regulatory role in the refinement of developing circuits. In mice lacking one form of the enzyme (GAD65) responsible for GABA synthesis, the ocular dominance plasticity resulting from monocular deprivation is absent, and increasing GABA inhibition in diazepam-treated mutant mice allows the appearance of ocular dominance plasticity. In the developing retinotectal system, a proper level of GABAergic inhibition is required for normal refinement of the retinotopic map in the tectum, both the increase and decrease of inhibition impede map refinement. Interestingly, the role of GABAergic inhibition appears not only to reduce the overall neuronal excitation, but also to sharpen the temporal pattern of the neuronal activity by shortening stimulus-evoked discharges, potentially facilitating spike-timing-dependent refinement of neural circuits. Shunting of specific excitatory inputs may also be achieved by selective distribution of inhibitory synapses on the dendrite. Finally, inhibitory synapses may also undergo competition and refinement as well, because they are integral part of the neural circuit. At present, it is unknown how nascent inhibitory connections, while playing important regulatory roles in refining excitatory connections, can themselves undergo activity-driven refinement and be properly consolidated into the mature circuit.

Causal Link between LTP/LTD and Synapse Competition

At some synapses repetitive coactivation of the pre- and postsynaptic cell leads to not only homosynaptic LTP, but also heterosynaptic LTD of non-coactive converging synapses onto the postsynaptic cell, thus providing a potential competitive mechanism for synaptic elimination. The hypothesis that synapse stabilization and elimination are mechanistically linked to or even result from activity-induced LTP and LTD, respectively, remains to be fully tested. Many lines of correlative evidence support this hypothesis. Blocking NMDA receptor activation, which abolishes many forms of LTP/LTD, impedes refinement of developmental circuits.

Repetitive visual stimuli that modify developing visual circuits can induce NMDA receptor-dependent LTP/LTD of retinotectal synapses. During the postnatal critical period, the composition of NMDA receptors undergoes experience-dependent developmental regulation in the visual system, and there is a correlation between the susceptibility for LTP/LTD induction and for circuit refinement. However, whether LTP/LTD is relevant or a prelude to structural refinement in the visual system remains unknown. Synapse elimination (structurally) will certainly eliminate the synaptic function, but whether LTD will lead to synapse elimination is unclear. Recent studies in hippocampal slices has provided evidence that LTP/LTD induction is followed by a swelling/shrinkage of dendritic spines, supporting the linkage between synaptic efficacy and synaptic structure.

The Trophic Factor Hypothesis

Purves and Lichtman have proposed that synaptic competition involves the competition between co-innervating presynaptic terminals for a limited amount of 'trophic factors' derived from the postsynaptic cell. This hypothesis can be extended to factors in the postsynaptic cytoplasm or plasmalemma, together with a localized retrograde signaling to the presynaptic nerve terminal. Competition for the trophic factor can be regulated by activity. The activity can serve a permissive role for synapse competition by regulating the synthesis and release of trophic factors/retrograde signals, whereas other activity-independent mechanisms determine the competitive advantage of a synapse. For activity to serve an instructive role, the pattern of activity in the co-innervating nerve terminals may determine the outcome of competition by controlling the uptake or the efficacy of the trophic factor at the nerve terminal. The local release of trophic factors may also depend on local synaptic activation, which is in turn driven by the pattern of activity, including the timing of pre- and postsynaptic spiking. The neurotrophin family of proteins, which are known to regulate synaptic function and axon/dendrite morphology, are attractive candidates for the trophic factor in synapse competition. The expression and secretion of neurotrophins, and their potentiating actions on the synapse, are all activity dependent. Neurotrophins are required for the development of ODC and for the induction of activity-induced LTP in several systems. Furthermore, neurotrophin secretion is activity pattern dependent, and the secreted neurotrophins are likely to act locally by its binding to cell surfaces at the synapse, allowing them to serve as local synaptic modulators. Activity-dependent depletion of available neurotrophins in the local environment of the synapse may lead to the functional and structural modification underlying synapse elimination.

Concluding Remarks

Studies of synapse competition and elimination in the near future are likely to be facilitated greatly by the availability of many new technologies for selective neuronal labeling and optical imaging in the living animal. Transgenic animals with selective populations of fluorescence-tagged neurons, together with *in vivo* multiphoton imaging, will allow us to directly monitor changes in the morphology of axons and dendrites during the process of synapse competition and refinement over prolonged period in the intact brain. Fluorescent Ca^{2+} or membrane-voltage sensors will allow us to monitor neuronal activities, and photo-activated probes expressed in selective neuronal populations will allow us to directly manipulate neuronal activity and to examine how activity drives synapse competition and elimination.

See also: Axonal Regeneration: Role of Growth and Guidance Cues; Cochlear Development; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Hebbian Plasticity; Interstitial Axon Branching/Collateral Elimination; Long-Term Depression: Cerebellum; Neurotrophins: Physiology and Pharmacology; Olfactory Bulb Anatomy; Peripheral Nerve Regeneration: An Overview; Somatosensory Cortex: Functional Architecture; Somatosensory Cortex; Topographic Maps: Molecular Mechanisms; Visual Development.

Further Reading

Bi G and Poo M-M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience* 24: 139–166.

Buffelli M, Burgess RW, Feng G, Lobe CG, Lichtman JW, and Sanes JR (2003) Genetic evidence that relative synaptic efficacy biases the outcome of synaptic competition. *Nature* 424: 430–434.

Feldman DE and Brecht M (2005) Map plasticity in somatosensory cortex. *Science* 310: 810–815.

Hua JY, Smear MC, Baier H, Stephen J, and Smith SJ (2005) Regulation of axon growth *in vivo* by activity-based competition. *Nature* 434: 1022–1026.

Hubel DH, Wiesel TN, and LeVay S (1977) Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical Transactions of the Royal Society of London B: Biological Science* 278: 377–409.

Jansen JK, Van Essen DC, and Brown MC (1976) Formation and elimination of synapses in skeletal muscles of rat. *Cold Spring Harbor Symposia on Quantitative Biology* 40: 425–434.

Katz LC and Shatz CJ (1996) Synaptic activity and the construction of cortical circuits. *Science* 274: 1133–1138.

Lichtman JW and Purves D (1980) The elimination of redundant preganglionic innervation to hamster sympathetic ganglion cells in early post-natal life. *Journal of Physiology* 301: 213–228.

Lohof AM, Delhaye-Bouchaud N, and Mariani J (1996) Synapse elimination in the central nervous system: Functional significance and cellular mechanisms. *Reviews of Neuroscience* 7: 85–101.

Poo M-M (2001) Neurotrophins as synaptic modulators. *Nature Reviews Neuroscience* 2: 24–32.

Purves D and Lichtman JW (1980) Elimination of synapses in the developing nervous system. *Science* 210: 153–157.

Rubel EW and Fritzsch B (2002) Auditory system development: Primary auditory neurons and their targets. *Annual Review of Neuroscience* 25: 51–101.

Sanes JR and Lichtman JW (1999) Development of the vertebrate neuromuscular junction. *Annual Review of Neuroscience* 22: 389–442.

Sotelo C (2004) Cellular and genetic regulation of the development of the cerebellar system. *Progress in Neurobiology* 72: 295–339.

Yu CR, Power J, Barnea G, et al. (2004) Spontaneous neural activity is required for the establishment and maintenance of the olfactory sensory map. *Neuron* 42: 553–566.

Synapsins

M V Khvotchev and J Sun, University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

Introduction

More than 90% of neurotransmission in the brain occurs at synapses via fast Ca^{2+}-regulated exocytosis of small synaptic vesicles. Synaptic vesicles are relatively stable and do not randomly fuse with the plasma membrane, other organelles, or each other. Instead, they undergo tightly controlled fusion with a specialized area of the plasma membrane called the active zone. After exocytosis, synaptic vesicles are locally reformed via multiple endocytic pathways and reused for subsequent rounds of fusion. Although uniform in size and shape, synaptic vesicles are not functionally equal. At many synapses, they appear to be organized in discrete pools that release neurotransmitters with vastly different probabilities. Thus, rapid recruitment of synaptic vesicles from the reserve pool to the readily releasable pool enables synapses to adapt to a broad range of activities. Molecular mechanisms that govern the size and maintenance of synaptic vesicle pools are not well understood. Extensive studies during the past 30 years suggest that a family of abundant synaptic phosphoproteins called synapsins participates in maintaining stability and pool organization of synaptic vesicles.

Structure of Synapsins

Synapsins are an evolutionary conserved protein family found in most species throughout the animal kingdom, ranging from *Caenorhabditis elegans* to humans. In invertebrates, a single gene for synapsin has been identified. Most vertebrate species contain three synapsin genes, termed I, II, and III. Each synapsin gene appears to include a nested gene for tissue inhibitor of metalloproteinases.

The mRNAs transcribed from synapsin genes are alternatively spliced at the 3′ ends. In mammals, alternative splicing gives rise to two isoforms for each synapsin I and II: the long forms, called synapsins Ia and IIa, and the short forms, called synapsins Ib and IIb. For synapsin III, only the long IIIa isoform has been characterized. However, analysis of gene organization and mRNA transcripts in humans indicates that at least six different synapsin III isoforms may be present.

Sequence comparison reveals that all synapsins consist of an assortment of conserved and variable domains (Figure 1). The first 400 amino acids of all synapsins form three signature domains: a short conserved domain A; a variable domain B that is enriched in prolines, serines, and glycines; and a large highly conserved domain C. The C-terminal parts of synapsins are highly divergent and contain a variety of short domains from D to J (Figure 1). The only domain shared by all three synapsins in this area is domain E, found at the C-termini of synapsins Ia, IIa, and IIIa.

The three-dimensional structure of the largest and most conserved domain in synapsins, domain C, has been determined (Figure 2). Domain C shows remarkable structural similarity to a family of bacterial ligases and synthetases, suggesting that synapsins may be adenosine triphosphate (ATP)-utilizing enzymes.

All synapsins contain several conserved and isoform-specific phosphorylation consensus sites for multiple protein kinases. Domain A in all synapsins carries a phosphorylation site for protein kinase A (PKA) and CaM kinase I/IV. Domain C contains a single conserved site for tyrosine kinase Src. In addition, domains B and D in synapsin I contain multiple sites for MAPK, cdk1, cdk5, and CaMK II kinases.

Expression and Localization of Synapsins

Synapsins are abundant proteins expressed primarily in the brain. Synapsins I and II are expressed at the highest levels and are specifically localized to presynaptic terminals, whereas synapsin III is expressed at very low levels and is also found in cell bodies and growth cones. Synapsins are present at most synapses, with the notable exception of ribbon synapses, and show isoform-specific differential distribution. Synapsin II appears to be relatively enriched in excitatory glutamatergic synapses, whereas synapsin I is more abundant at inhibitory GABAergic synapses. At presynaptic terminals, synapsins dynamically associate with the surface of synaptic vesicles and constitute approximately 9% of total synaptic vesicle protein (~15–20 synapsin molecules per vesicle).

Biochemistry of Synapsins

Synapsins bind to synaptic vesicles with nanomolar affinities. This interaction may occur via binding to lipids, proteins, or a combination of both. *In vitro*, synapsins show saturable binding to synthetic vesicles containing negatively charged phospholipids and are able to stabilize phospholipid bilayers. The N-terminal

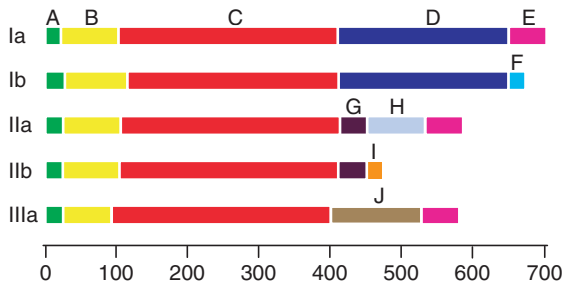

Figure 1 Domain organization of the five synapsins. Domains are color coded and labeled on top. Amino acid residue scale is shown below.

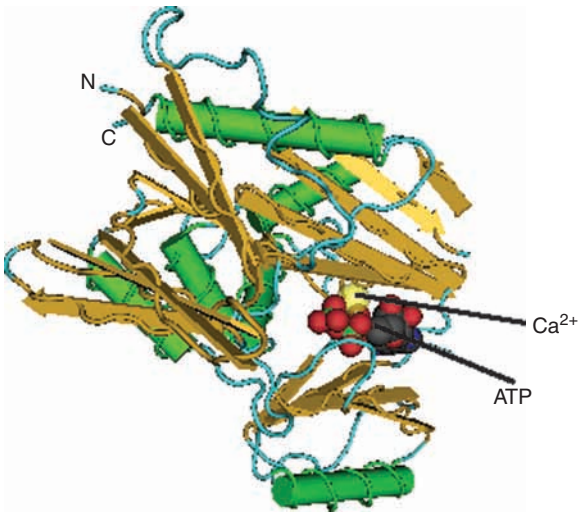

Figure 2 Ribbon diagram of the complex of the rat synapsin I domain C complexed with ATP and calcium (PDB entry 1PX2).

domain A appears to be responsible for the electrostatic interaction between synapsins and phospholipids. In addition, the central domain C may mediate hydrophobic protein–lipid interactions.

A plethora of protein interactions for synapsins have been identified *in vitro*. These include components of cytoskeleton (actin and actin filaments, spectrin, microtubules, and neurofilaments), scaffolding proteins (Grb2), lipid-binding proteins (annexins), and protein kinases (c-src and CaM kinase II). Synapsins also bind each other and form stable homo- and heterodimers and tetramers via the domain C. However, the *in vivo* significance of these protein interactions is not clear.

The three-dimensional structure of the domain C resembles that of a family of ATP-utilizing enzymes (Figure 2). Domain C in all synapsins binds ATP with low micromolar affinity (0.15–0.45 μM) and forms a stable complex with ATP and calcium. Surprisingly, ATP binding to synapsins is regulated by calcium in an isoform-specific manner. At low micromolar concentration, calcium facilitates ATP binding to the domain C of synapsin I, has no effect on ATP binding to synapsin II, and inhibits binding of ATP to synapsin III. Synapsins have no detectable ATPase activity *in vitro*. The function of ATP binding to domain C in synapsins is unknown. It has been proposed that synapsins can be ATP-dependent enzymes or ATP sensors at synapses.

Synapsins are phosphorylated *in vivo* and *in vitro* by multiple protein kinases and are major substrates for PKA in brain. Phosphosynapsins show altered biochemical properties and synaptic localization, suggesting that phosphorylation may play an important regulatory role. Phosphorylation by PKA at the conserved site within the lipid-binding domain A appears to cause dissociation of synapsins from synaptic vesicles and accumulation of phosphosynapsins in the cytosol. Interestingly, electrical or chemical stimulation of synapses increases both phosphorylation of synapsins and their dissociation from synaptic vesicles. Thus, the association of synapsins with synaptic vesicles may be dynamically regulated by phosphorylation at the single conserved site. Phosphorylation of synapsins often change their conformation and the apparent affinity of interactions with other proteins. For example, synapsin I that is phosphorylated by CaM kinase II within domain D displays significantly reduced binding to actin.

Functions of Synapsins

Several approaches have been used throughout the years to test functions of synapsins in the regulation of neurotransmitter release. Injection of the whole synapsin I (but not the CaM kinase II phosphorylated synapsin I) into presynaptic terminals inhibits neurotransmitter release. Furthermore, injection of anti-synapsin antibodies or peptides derived from domain E into lamprey reticulospinal axons or giant squid synapses results in depletion of synaptic vesicles away from the plasma membrane but has no effect on the number of docked vesicles. Injected nerve terminals show normal activity under low-frequency stimulation but are rapidly depressed upon high-frequency stimulation. These results combined with several biochemical observations have led to a model proposing that synapsins reversibly cross-link synaptic vesicles to each other and to cytoskeletal elements at synapses, and they regulate their availability for exocytosis. In this model, phosphorylation of synapsins performs a crucial regulatory role by affecting interactions of synapsins with phospholipids and cytoskeletal proteins.

Further evidence for an important role of synapsins in regulating neurotransmitter release is derived from mouse genetic experiments. Mouse lines lacking synapsins I, II, and III individually and double or triple

knockout lines have been described. All knockout mice are viable, fertile, and show no gross morphologic abnormalities. In addition, mice have no apparent change in life span and exhibit a range of normal complex behaviors, suggesting that synapsins are not required for general brain functions such as synaptogenesis or axonal transport. Of all generated lines, synapsin I and II double knockout mice display the strongest abnormalities in synaptic functions. These animals and also single synapsin I and II but not synapsin III knockouts develop frequent epileptoid seizures precipitated by sensory stimuli. The size and density of synapses and active zones in double knockout mice are normal; however, the total number of synaptic vesicles per synapse is reduced by more than half. The remaining synaptic vesicles still form clusters at the release sites. The levels of many intrinsic synaptic vesicle proteins, such as synaptotagmin I, synaptophysin I, and synaptobrevin II, are also reduced in the double knockout line, suggesting that synapsins may be important for the stability of synaptic vesicles. The reduction in synaptic vesicle density appears to be uniform both away from the plasma membrane and in the areas close to the active zone. In electrophysiologic tests, double knockout mice exhibit normal neurotransmitter release under low-frequency stimulation and normal paired pulse facilitation. Thus, synapsins are not required for synaptic exocytosis as such, and they do not significantly modify the release probability of synaptic vesicles at the active zone. However, repetitive high-frequency stimulation of synapses lacking synapsins I and II reveals a severe defect in their ability to sustain high levels of activity. For example, a form of short-term plasticity, posttetanic potentiation, is significantly decreased in these knockouts, and synaptic depression is drastically increased. High-resolution electrophysiological studies of neurotransmitter release at the calyx of Held synapse by direct capacitance measurements did not uncover any defects in spontaneous or single action potential evoked release in the mutants. However, use-dependent synaptic depression induced by high-frequency stimulation was increased in a CaM kinase II-dependent manner. The contribution of synapsin III to these phenotypes is insignificant and the function of synapsin III remains unknown.

Conclusions

Synapsins are a family of abundant evolutionary conserved phosphoproteins found at nearly all synapses in the nervous system. At presynaptic terminals, they reversibly associate with the surface of synaptic vesicles and presumably with cytoskeletal proteins such as actin. These interactions of synapsins are regulated by phosphorylation at multiple sites. A combination of functional and genetic approaches established that synapsins are not essential constituents of the synaptic vesicle cycle. Instead, they appear to regulate the maintenance of the reserve vesicle pool and recruitment of synaptic vesicles from the reserve pool to the readily releasable pool. Synapsins appear to boost the release probability at the synapses under high-frequency stimulation. Thus, synapsins may be especially important for synapses that undergo rapid changes in their levels of activity. The molecular mechanisms by which synapsins exert their action are not fully understood. In particular, it is not clear whether synapsin isoforms generated by alternative splicing have similar or different functions at the synapse. The discovery of the structural homology between synapsins and a family of ATP-utilizing enzymes raises an intriguing possibility that synapsins may have enzymatic functions at the synapse or their function may be regulated by ATP. Future functional studies on synapsins will address these questions.

See also: Active Zone; Adenosine Triphosphate (ATP); Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neurotransmitter Release: Synchronous and Asynchronous; Presynaptic Development and Active Zones; Synapsins and Regulation of the Reserve Pool; Synaptic Plasticity: Short-Term Mechanisms; Synaptic Vesicles.

Further Reading

Bahler M, Benfenati F, Valtorta F, and Greengard P (1990) The synapsins and the regulation of synaptic function. *Bioessays* 12: 259–263.

Benfenati F, Bahler M, Jahn R, and Greengard P (1989) Interactions of synapsin I with small synaptic vesicles: Distinct sites in synapsin I bind to vesicle phospholipids and vesicle proteins. *Journal of Cell Biology* 108: 1863–1872.

Cheetham JJ, Hilfiker S, Benfenati F, et al. (2001) Identification of synapsin I peptides that insert into lipid membranes. *Biochemical Journal* 354: 57–66.

De Camilli P, Benfenati F, Valtorta F, and Greengard P (1990) The synapsins. *Annual Review of Cell Biology* 6: 433–460.

Esser L, Wang CR, Hosaka M, et al. (1998) Synapsin I is structurally similar to ATP-utilizing enzymes. *EMBO Journal* 17: 977–984.

Feng J, Chi P, Blanpied TA, et al. (2002) Regulation of neurotransmitter release by synapsin III. *Journal of Neuroscience* 22: 4372–4380.

Ferreira A and Rapoport M (2002) The synapsins: Beyond the regulation of neurotransmitter release. *Cellular and Molecular Life Sciences* 59: 589–595.

Greengard P, Valtorta F, Czernik AJ, and Benfenati F (1993) Synaptic vesicle phosphoproteins and regulation of synaptic function. *Science* 259: 780–785.

Hilfiker S, Pieribone VA, Czernik AJ, et al. (1999) Synapsins as regulators of neurotransmitter release. *Philosophical*

Transactions of the Royal Society of London, Series B: Biological Science 354: 269–279.

Hosaka M, Hammer RE, and Südhof TC (1999) A phospho-switch controls the dynamic association of synapsins with synaptic vesicles. *Neuron* 24: 377–387.

Hosaka M and Südhof TC (1998) Synapsins I and II are ATP-binding proteins with differential Ca^{2+} regulation. *Journal of Biological Chemistry* 273: 1425–1429.

Kao HT, Porton B, Hilfiker S, et al. (1999) Molecular evolution of the synapsin gene family. *Journal of Experimental Zoology* 285: 360–377.

Rosahl TW, Geppert M, Spillane D, et al. (1993) Short-term synaptic plasticity is altered in mice lacking synapsin I. *Cell* 75: 661–670.

Rosahl TW, Spillane D, Missler M, et al. (1995) Essential functions of synapsins I and II in synaptic vesicle regulation. *Nature* 375: 488–493.

Sihra TS, Wang JK, Gorelick FS, and Greengard P (1989) Translocation of synapsin I in response to depolarization of isolated nerve terminals. *Proceedings of the National Academy of Sciences of the United States of America* 86: 8108–8112.

Südhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Südhof TC, Czernik AJ, Kao HT, et al. (1989) Synapsins: Mosaics of shared and individual domains in a family of synaptic vesicle phosphoproteins. *Science* 245: 1474–1480.

Sun J, Bronk P, Liu X, Han W, and Südhof TC (2007) Synapsins regulate use-dependent synaptic plasticity in the calyx of Held by a Ca^{2+}/calmodulin-dependent pathway. *Proceedings of the National Academy of Sciences of the United States of America* 103: 2880–2885.

Torri Tarelli F, Bossi M, Fesce R, et al. (1992) Synapsin I partially dissociates from synaptic vesicles during exocytosis induced by electrical stimulation. *Neuron* 9: 1143–1153.

Synapsins and Regulation of the Reserve Pool

D Gitler, Ben Gurion University of the Negev, Beer-Sheva, Israel
G J Augustine, Duke University Medical Center, Durham, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

Neurotransmitter Release Depends on Several Pools of Synaptic Vesicles

Presynaptic terminals rely on a complex cycle of membrane trafficking to release neurotransmitters. After synaptic vesicles (SVs) are synthesized, they undergo repeated cycles of fusion (exocytosis) and subsequent retrieval (endocytosis) within the presynaptic terminal. SVs apparently are not equivalent and can be divided into pools on the basis of the ease with which they undergo exocytosis in response to neuronal activity. One fraction of the SVs is easily mobilized for release, and this fraction is defined as the recycling pool. At most synapses, SVs released from the recycling pool return to this pool after endocytosis, in effect retaining their status. Within the recycling pool, a subgroup of SVs is docked at the plasma membrane at the active zone, poised and ready for fusion with the membrane. Such SVs require no (or minimal) additional molecular adjustments to fuse once an action potential arrives; they constitute the readily releasable pool (RRP) of SVs. Essentially, the RRP is the source of SVs for exocytosis. As its members fuse with the plasma membrane, they are replaced from the recycling pool.

During periods of intense synaptic activity, when demand for transmitter-laden SVs exceeds the supply available in the recycling pool, SVs are mobilized from a reserve pool (RP), where the majority of SVs in the terminal reside. Vesicle mobilization occurs through a poorly defined mechanism that senses synaptic activity. This requirement for activity effectively segregates the RP SVs into a separate pool. Recent results indicate that in many synapse types, the RP and the recycling pool (but not the RRP) are not spatially segregated, suggesting that the division of the SVs into pools is determined by molecular identity rather than by spatial location.

The organization of SVs into pools and their mobilization from the RP to the RRP allows neurons to release neurotransmitters even when the rate of release exceeds the rate at which SVs are recycled. Because the integrative properties of synaptic transmission depend on the dynamics of neurotransmitter release from the presynaptic terminal, it is important to understand the molecular origins of the partitioning of SVs into pools, the spatial manifestation of this division, and the interplay between these vesicle pools. This article concentrates on the role that the synapsins play in controlling SV dynamics.

Synapsins Are a Family of Phosphoproteins Found on SVs

In the 1970s, synapsins were identified by Paul Greengard and colleagues as the most abundant phosphoproteins in the brain. Originally termed 'protein I' and 'protein III,' they were later renamed synapsin I and synapsin II. We now know that there are three mammalian synapsin genes, named synapsin I, II, and III. Each of these genes is subject to alternative splicing, producing at least ten isoforms: synapsins Ia, Ib, IIa, IIb, IIIa, IIIb, IIIc, IIId, and the atypical IIIe and IIIf. The synapsin isoforms arise from the combinatorial arrangement of numerous domains (Figure 1). The N-terminal portion of the synapsins is highly conserved and consists of domains A, B and C, whereas the C-terminal portion is variable because of heterogeneous combinations of domains D–J.

Synapsins are expressed throughout the central and peripheral nervous systems, although their expression varies across synapse types and brain regions. In contrast to synapsins I and II, which are stably expressed in adult presynaptic terminals, synapsin III expression is developmentally controlled and is not confined to synaptic terminals.

Synapsins are associated with SVs and exhibit many intriguing biochemical properties. They are peripheral proteins of SVs that bind to both the membrane phospholipids and the protein constituents of the SV. Synapsins also bind to cytoskeletal elements, such as actin, tubulin, and spectrin, and it is thought that these connections between SVs and the cytoskeleton may be important for the functioning of synapsins. As already mentioned, synapsins are the most abundant phosphoproteins in the brain and serve as substrates for several protein kinases, including protein kinase A (PKA), calcium/calmodulin dependent protein kinase type I (CaMKI), CaMKII, and mitogen-activated protein (MAP) kinase. Phosphorylation by CaMKII and PKA causes synapsins to dissociate from SVs. Synapsins also interact with other cellular signaling elements, such as phosphatidylinositol 3-kinase and regulators of G-protein signaling. Synapsins bind to adenosine triphosphate (ATP) and

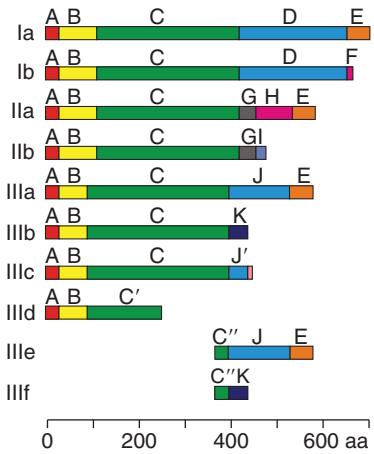

Figure 1 Domain structure of the ten known isoforms produced by the three synapsin genes.

have structural similarities to ATP-utilizing enzymes; this suggests that synapsins may possess an as yet unidentified enzymatic function.

Synapsins Regulate SV Trafficking

Several lines of research have investigated the functional role of synapsins in intact synapses. Their role in synaptic transmission has been examined in mice in which some or all synapsins have been deleted, in synapses where synapsin function was disrupted by injection of inhibitory peptides, and in synapses into which purified synapsins were injected. Such treatments produce varied effects on neurotransmitter release, on short-term synaptic plasticity, and/or on the distribution of SVs in terminals. Although the results of these studies present a complex – and at times contradictory – picture of the role played by the synapsins, several consistent themes have emerged. First, synapsins have been shown to control the number and distribution of SVs within presynaptic terminals, in most cases predominantly affecting SVs far from the active zone. This suggests that synapsins maintain SVs in the RP. Second, synapsins have been shown to help sustain synaptic transmission during intense stimulation, also consistent with a role in maintaining the RP. In some synapses, synapsins also appear to regulate fusion of SVs not originating in the RP. Other roles for synapsins have been proposed, including regulation of fusion kinetics, stabilization of SVs, and/or regulation of late steps of endocytosis. Synapsins also have been implicated in neuronal development, synaptogenesis, and maintenance of mature synapses.

Deletion of individual synapsin genes in mice causes varied effects on synaptic transmission. Synaptic depression, a form of plasticity that requires mobilization of SVs from the RP, is not affected by deletion of the synapsin I gene but is accelerated by loss of synapsin II and is slowed by deletion of synapsin III. Since synapsins may have redundant functions, deletion of a single gene does not define its physiological roles because the remaining synapsins may compensate for its loss. In addition, deleting individual genes does not define the importance of the various isoforms produced from each gene by alternative splicing. Thus, deleting all three synapsin genes was a necessary starting point in defining the roles that synapsins play in synaptic function.

Synapsin-Deficient Mice Are Viable

Synapsin triple knockout (TKO) mice were generated by crossing preexisting lines of mice harboring null mutations in single synapsin genes. To serve as a genetically similar control, a line of matching wild-type mice, termed synapsin 'triple wild-type' (TWT) mice, was derived in parallel. The synapsin TKO mice have a normal life span and are fertile, though to a lower degree than the TWT mice. The TKO mice do not express obvious signs of morbidity, except for transient, mild seizures. Thus, synapsins are not 'essential' genes. Gross brain morphology is not affected noticeably by deleting the synapsins, indicating that these proteins do not play a crucial role in the development of the nervous system. However, synapsin TKO mice have various deficiencies in basic neurophysiological performance, such as muscle strength, posture, and reflexes. In addition, the mice show impairments in learning and memory tests, as exemplified by defects in performance in the radial arm maze and the Morris water maze.

Synapsins Regulate the Dynamics of Glutamatergic Synaptic Transmission

To determine the role of synapsins in synaptic function, properties of synaptic transmission have been compared in TKO and TWT neurons, mainly in hippocampal neurons grown in cell culture. The ability of synapsin TKO neurons to form synapses in culture did not differ from that of TWT neurons, which allows any differences in synaptic transmission to be attributed to changes in the properties of release from individual synapses rather than changes in the number of synapses.

The role of synapsins in release of glutamate from pyramidal neurons, the main type of neuron in the hippocampus, has been studied by using stimulation paradigms that report on the properties of the RRP and RP. Spontaneous transmitter release and release

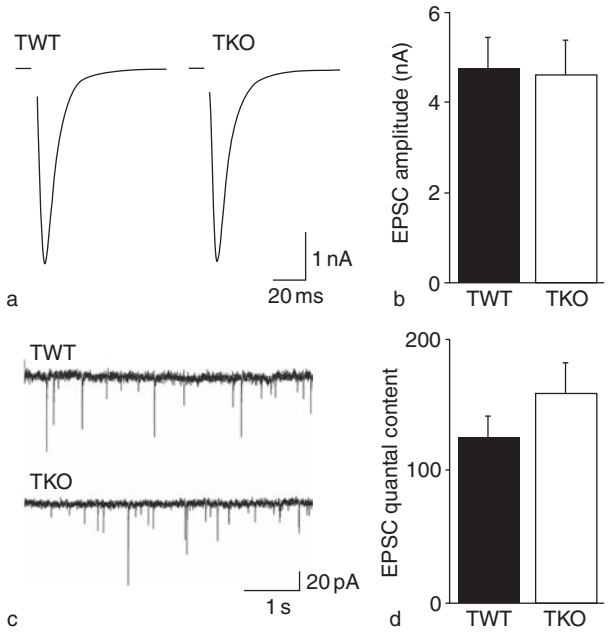

Figure 2 Characteristics of glutamatergic neurotransmission in cultured synapsin triple wild-type (TWT) and synapsin triple knockout (TKO) neurons. (a) Excitatory postsynaptic currents (EPSCs) recorded from cultured glutamatergic autaptic neurons. (b) Mean amplitudes of evoked EPSC of TWT and TKO neurons are not different. (c) Representative traces of spontaneous EPSCs in TWT and TKO neurons. (d) Mean quantal content of EPSCs is not different between TWT and TKO neurons. From Gitler D, Takagishi Y, Feng J, et al. (2004) Different presynaptic roles of synapsins at excitatory and inhibitory synapses. *Journal of Neuroscience* 24: 11368–11380.

evoked by single stimuli draw SVs from the RRP. In contrast, long trains of stimuli elicit synaptic depression, which results from the depletion of SVs from the RRP and the recycling pool, resulting in the mobilization of SVs from the RP.

Properties of synaptic transmission that are related to the size of the RRP – such as the amplitude of evoked excitatory postsynaptic currents (EPSCs; Figures 2(a) and 2(b)), the frequency of spontaneous EPSCs (Figure 2(c)), and the quantal content of evoked EPSCs (Figure 2(d)) – are not different in TKO and TWT neurons. This indicates that loss of synapsins does not affect the RRP of glutamatergic neurons. In addition, stimulation with pairs of stimuli elicits synaptic facilitation that is similar in TKO and TWT neurons, suggesting that synapsins do not affect the probability of release from the RRP.

On the other hand, the rate of synaptic depression is accelerated by the deletion of synapsins (Figures 3(a) and 3(c)). It is important to note that the difference between TWT and TKO neurons is evident only after the neurons have been stimulated 30 or more times (Figure 3(b)), a point at which the synaptic terminal has released most of its RRP. Thus, it appears that the RP of glutamatergic SVs is reduced in the TKO neurons. To quantify the number of exocytotic events evoked during such stimuli, the total amount of synaptic charge produced during the stimulus train can be calculated by integrating

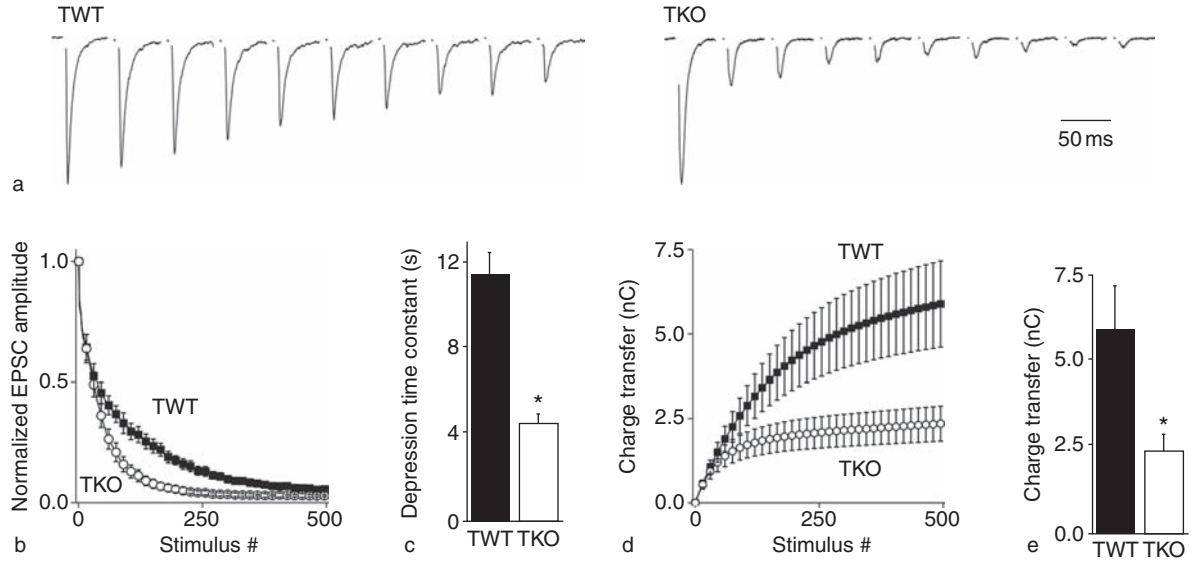

Figure 3 (a) Representative traces showing every 15th excitatory postsynaptic current (EPSC) evoked by a 10-Hz train of stimuli in synapsin triple wild-type (TWT) and synapsin triple knockout (TKO) neurons. (b) Mean EPSC amplitudes during the 500-event trains. Symbols represent EPSC amplitudes normalized to the first response in the train, and every 15th response is plotted. (c) Mean time constants for synaptic depression in TWT and TKO neurons. (d) Cumulative EPSC charge transfer in TWT and TKO neurons. (e) Mean values for total EPSC charge transfer after 500 stimuli were applied to TWT and TKO neurons. Asterisks denote a statistically significant difference. All error bars represent ± SEM. Adapted from Gitler D, Xu Y, Kao H-T, et al. (2004) Molecular determinants of synapsin targeting to presynaptic terminals. *Journal of Neuroscience* 24: 3711–3720.

each individual synaptic response (Figure 3(d)). TKO neurons generate much less synaptic charge than TWT neurons do, indicating that loss of synapsins approximately halves the number of SVs mobilized from the RP (Figure 3(e)).

Deleting Synapsins Reduces the Number of SVs

These electrophysiological indications of defects in the RP of TKO neurons can be correlated with other approaches that examine the number and distribution of SVs within presynaptic terminals. For example, the relative quantities of SV proteins (synaptobrevin 2, synaptotagmin I, and synaptophysin I) are reduced in brains from TKO mice, while representative postsynaptic proteins (PSD95 and *N*-methyl-D-aspartate receptor subunits NR1 and NR2B), structural proteins (β-tubulin), and presynaptic plasma membrane proteins (syntaxin I) are not affected. This indicates a specific reduction in the amount of SV proteins in the terminals of TKO mice, consistent with a loss of SVs as a consequence of synapsin deletion.

The density of SVs can also be assessed in live neurons by examining the localization of fluorescently tagged synaptobrevin 2, an integral protein of the SV membrane. Fusing this protein to the green fluorescent protein (GFP) permits visualization of SV clusters in individual presynaptic terminals of cultured neurons. The relative amount of GFP–synaptobrevin 2 in presynaptic terminals is significantly reduced in TKO neurons, again consistent with a reduction in the number of SVs.

Finally, electron microscopy can be used to determine the absolute number and spatial distribution of SVs (Figure 4(a)). Quantitative analysis reveals that TKO terminals have approximately half as many SVs as TWT neurons do (Figure 4(e)). Analysis of the spatial distribution of SVs shows that there is a specific reduction in the number of the SVs located far from the active zone in TKO neurons (Figures 4(b) and 4(c)), although the number of SVs docked at the presynaptic active zone – which presumably includes (and may comprise) the RRP – is unchanged (Figure 4(d)). Thus, several lines of evidence indicate that the absence of synapsins causes the RP to be diminished. This leads to the conclusion that synapsins maintain SVs in the RP, accounting for the observation that TKO neurons cannot maintain glutamatergic transmission during periods of intense synaptic activity.

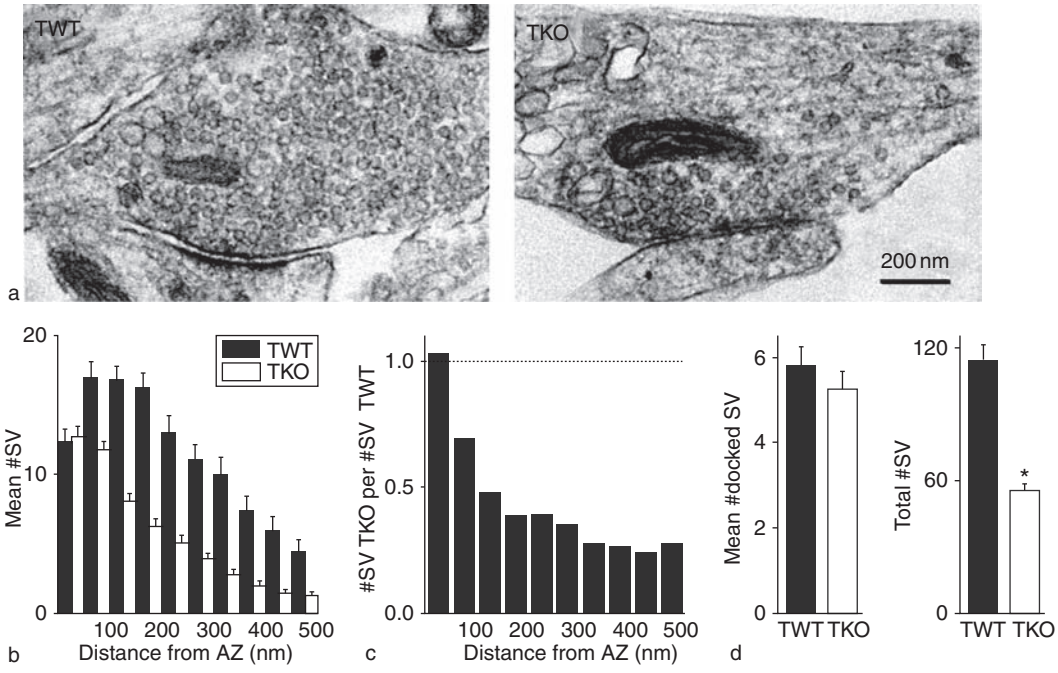

Figure 4 Spatial distribution of synaptic vesicles (SVs) in glutamatergic presynaptic terminals. (a) Representative electron micrographs of excitatory terminals of synapsin triple wild-type (TWT) and synapsin triple knockout (TKO) neurons. (b) Mean number of SVs located within 50-nm-wide concentric compartments centered at the active zone (AZ) of TWT and TKO neurons. (c) Ratio of the number of SVs in the various compartments, calculated from the data shown in (b) by dividing TKO values by the TWT values. Dashed line marks equality between TWT and TKO data. (d) Mean number of docked SVs in TWT and TKO terminals. (e) Mean total number of SVs within 500 nm from the active zone in TWT and TKO terminals. Asterisk denotes a statistically significant difference. All error bars represent ± SEM. From Gitler D, Xu Y, Kao H-T, et al. (2004) Molecular determinants of synapsin targeting to presynaptic terminals. *Journal of Neuroscience* 24: 3711–3720.

Synapsin IIa Maintains the RP of Glutamatergic Vesicles

Although these studies of synapsin TKO mice indicate that synapsins regulate the dynamics of glutamatergic synaptic transmission, it is not clear why so many different isoforms of synapsin exist. Previous experimental approaches using mice in which single synapsin genes are deleted suggest that the three synapsin genes are not equivalent in their ability to support the RP of glutamatergic SVs. However, knocking out entire genes cannot distinguish the roles of the various isoforms generated from each gene. Further, such studies cannot take into account the compensatory influences of remaining synapsin gene products.

TKO mice provide an opportunity to determine the functions of individual synapsin isoforms. On this null background, individual synapsin isoforms can be transfected into cultured TKO neurons, and electrophysiological methods then can be used to examine the effect of the introduced synapsins on synaptic depression. This approach indicates that different synapsin genes are not equivalent in function; indeed even isoforms derived from the same gene differ in their function.

Synapsin IIa slows synaptic depression when expressed in TKO neurons, illustrating that synapsin IIa functions to support the RP of glutamatergic SVs (Figure 5(a)). In contrast, synapsin IIb has no effect on synaptic depression (Figure 5(a)), indicating that isoforms from the same gene are not equivalent in function even though they differ only in their C-terminal tails. In fact, among the several synapsin isoforms that have been examined (synapsins Ia, Ib, IIa, IIb, and IIIa), synapsin IIa is the only one that slows synaptic depression when expressed in TKO neurons. Overexpression of synapsin IIa in TWT neurons also further slows synaptic depression (Figure 5(a)). This result indicates that synapsin IIa can increase the size of the RP. In summary, synapsin IIa apparently is the isoform responsible for the actions of synapsins on the RP of glutamatergic SVs.

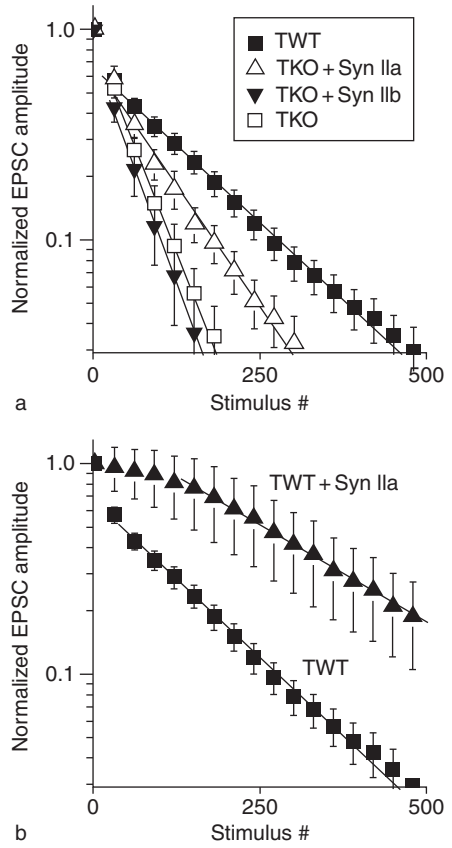

Figure 5 Effect of synapsin II isoforms on kinetics of synaptic depression. (a) Normalized excitatory postsynaptic current (EPSC) amplitudes during trains of 500 stimuli delivered at 10 Hz. Shown are mean amplitudes for untransfected synapsin triple wild-type (TWT) and synapsin triple knockout (TKO) neurons, and for TKO neurons transfected with green fluorescent protein (GFP)–synapsins IIa or IIb. Solid lines represent exponential decay curves. (b) As in (a) except data are from control untransfected TWT neurons and from TWT neurons transfected with GFP–synapsin Ib. Syn, synapsin.

Synapsin Ib Serves as a Dominant Negative Isoform

What are the functions of the remaining synapsin isoforms? Hints about the function of synapsin Ib first came from studies of the targeting of synapsins to presynaptic terminals. GFP-tagged versions of any synapsin isoform, including synapsin Ib, properly target presynaptic terminals in TWT neurons. However, in TKO neurons synapsin Ib is not able to target even though other synapsins still target normally (Figures 6(a)–6(c)). This suggests that targeting of synapsin Ib requires the presence of other synapsin isoforms. Previous *in vitro* work shows that synapsin I is capable of dimerizing, either with itself or with synapsin II. Co-transfection of synapsin Ia or synapsin IIa rescues the targeting of synapsin Ib (Figure 6(d)), indicating that targeting of this isoform is mediated by heterodimerization with other synapsin isoforms.

Given that synapsin Ib alone cannot target to presynaptic terminals, it is possible that this isoform could form heterodimers that interfere with the targeting and subsequent presynaptic function of other synapsin isoforms that dimerize with synapsin Ib. This hypothesis has been considered by examining the influence of synapsin Ib on targeting of other synapsins expressed in TKO neurons. Synapsin Ib reduces the targeting of synapsins, particularly targeting of synapsin IIa, the isoform most important for determining synaptic

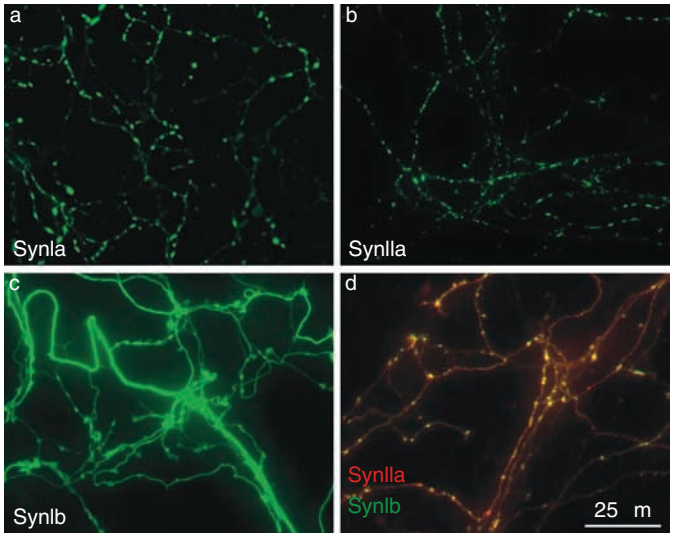

Figure 6 Synaptic targeting capabilities of synapsins. (a) Green fluorescent protein (GFP)–synapsin Ia and (b) GFP–synapsin IIa localize to presynaptic terminals in cultured synapsin triple knockout (TKO) neurons. (c) GFP–synapsin Ib does not localize to the presynaptic terminals but instead concentrates in the axons of transfected neurons. (d) On co-transfection with yellow fluorescent protein (YFP)–synapsin IIa (red), cyan fluorescent protein (CFP)–synapsin Ib (green) co-localizes (yellow) to presynaptic terminals. Syn, synapsin.

depression kinetics (Figure 7(a)). Likewise, synapsin Ib reduces presynaptic targeting of endogenous synapsins. This was established by immunolabeling TWT neurons with an antibody that recognizes synapsins Ia, IIa, and IIIa yet does not recognize synapsin Ib. Overexpressing synapsin Ib reduces the amount of these endogenous synapsins in the synaptic terminals while causing an uncharacteristic increase in their abundance within axons. Thus, synapsin Ib does interfere with targeting of other synapsin isoforms.

Given that synapsin Ib interferes with the targeting of other synapsins and that reducing the amount of synapsins reduces the size of the RP (in synapsin TKO neurons), it is possible that synapsin Ib reduces the RP. This possibility was examined by evaluating the effect of synapsin Ib on the targeting of GFP–synaptobrevin 2. Synapsin Ib caused a substantial reduction in the targeting of synaptobrevin 2 (Figure 7(b)), indicating that synapsin Ib reduces the size of the RP. To further evaluate this possibility, synapsin Ib has been overexpressed in TWT neurons. In these conditions, synapsin Ib significantly enhances the rate of synaptic depression so that depression kinetics in synapsin Ib-transfected TWT neurons was indistinguishable from that in TKO neurons (Figure 7(c)). Because synapsin Ib has no effect when expressed in TKO neurons, an interaction between synapsin Ib and endogenous synapsins must be responsible for the inhibitory effects of synapsin Ib.

This reduction in the RP seems to be responsible for the acceleration of depression by synapsin Ib because there is a very good correlation between RP size and the rate of depression in neurons expressing synapsin Ib (Figure 7(d)). In summary, synapsin Ib appears to serve as a dominant negative isoform by heterodimerizing with other synapsin isoforms, reducing the targeting of these isoforms and thereby reducing the size of the RP of SVs.

Synapsin Dimerization May Control Synaptic Dynamics

The observations described above suggest that dimerization of synapsins controls the extent of cross-linking of SVs in the presynaptic terminal (Figure 8). This proposal is based on proposed differences in the ability of different synapsin isoforms to cross-link SVs (Figure 8(a)). Synapsin IIa acts as the prototypical synapsin by clustering SVs in the terminals through formation of divalent dimers that can interact with more than one vesicle, either directly with the vesicle or through interactions with cytoskeletal elements. On the other hand, synapsin Ib negates this effect by forming heterodimers with other synapsins that are monovalent for interaction with SVs and thus does not allow cross-linking and vesicle clustering.

This hypothesis predicts that the relative abundance of the synapsin isoforms determines the net behavior of the RP. An increase in the abundance of synapsin IIa will shift the balance toward divalent dimers and thus enlarge the RP and slow synaptic depression (Figure 8(b)). Conversely, a relative increase in synapsin Ib will raise the proportion of

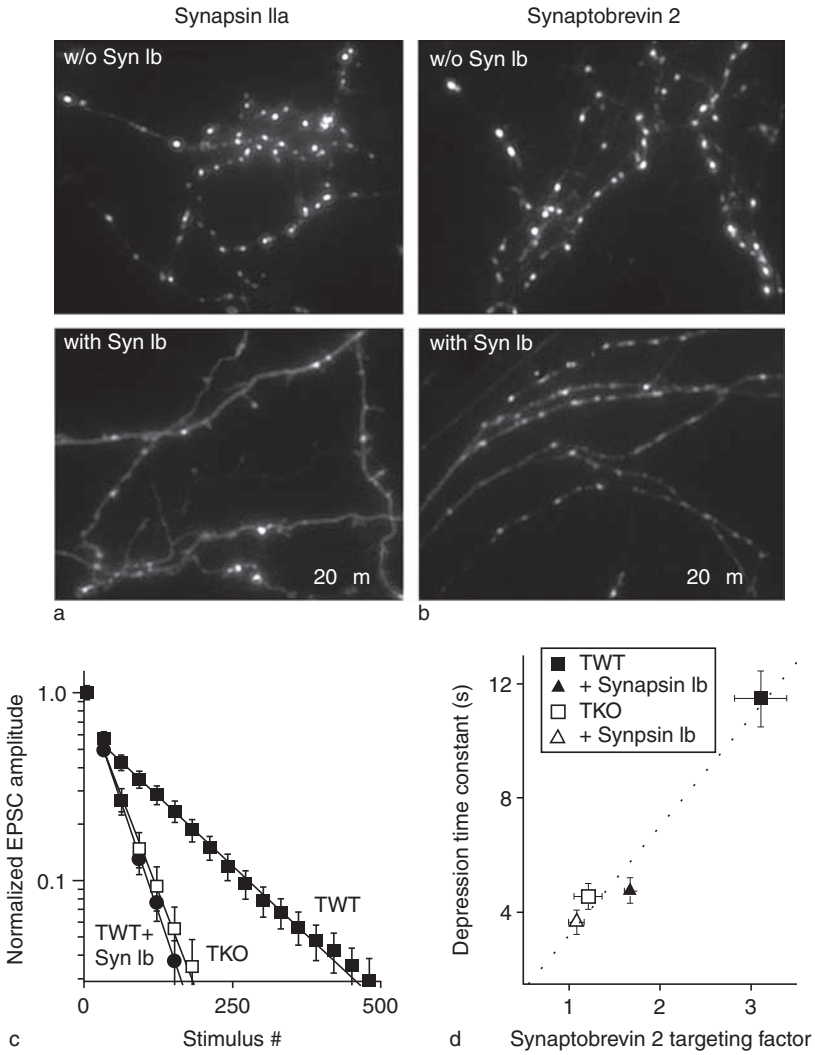

Figure 7 Synapsin Ib interferes with synaptic targeting of other synapsins and decreases the size of the reserve pool. (a) Synaptic localization of yellow fluorescent protein (YFP)–synapsin IIa expressed in synapsin triple knockout (TKO) neurons, in the presence (bottom) or absence (top) of coexpressed cyan fluorescent protein (CFP)–synapsin Ib (not shown in these images). (b) Synaptic localization of green fluorescent protein (GFP)–synaptobrevin 2 when expressed in synapsin triple wild-type (TWT) neurons, in the presence (bottom) or absence (top) of coexpressed CFP–synapsin Ib (not shown in images). (c) Normalized excitatory postsynaptic current (EPSC) amplitudes during trains of 500 stimuli delivered at 10 Hz. Shown are mean amplitudes for untransfected TWT and TKO neurons, and for TWT neurons transfected with GFP–synapsin IIa. Solid lines represent exponential decay curves. (d) Correlation of synaptic depression time constant with synaptic targeting of GFP–synaptobrevin 2 in TKO and TWT neurons. Slow synaptic depression is correlated with a larger vesicle pool in TWT neurons. Expression of synapsin Ib accelerates synaptic depression and reduces the size of the vesicle pool in TWT neurons to the level of the TKO neurons. Syn, synapsin.

ineffective monovalent dimers, thus dispersing the RP and accelerating synaptic depression (Figure 8(c)). Hence, synapsins can bidirectionally fine-tune the characteristics of synaptic depression.

Synapsins Have Different Functions at Other Synapse Types

Many lines of evidence indicate that synapsins regulate the RP of glutamatergic SVs. However, synapsins have other functions at other synapses. At hippocampal γ-aminobutyricacidergic (GABAergic synapses), loss of synapsins also reduces the number of SVs, but this effect is not restricted to the RP. This homogeneous decrease in the number of GABAergic SVs throughout the terminal also reduces the number of docked SVs. As a result, basal evoked release of GABA is diminished, but the kinetics of synaptic depression is unaffected. Thus, loss of synapsins has different structural and physiological effects at glutamatergic and GABAergic neurons. Synapsins appear to play a role also in regulating the RP of striatal and caudate–putamen

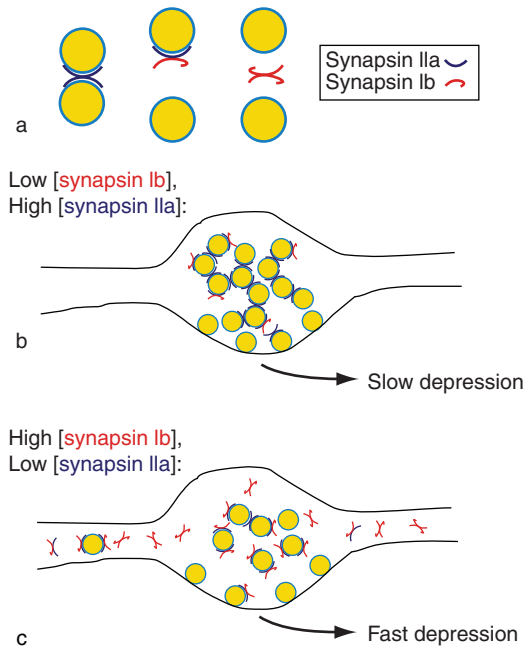

Figure 8 Model for synapsin control of vesicle pools in presynaptic terminals. (a) Synapsin IIa forms a divalent dimer that can bind and cross-link synaptic vesicles (SVs), whereas heterodimerization of synapsin IIa with synapsin Ib disrupts this clustering ability. (b) A predominance of synaspsin IIa in the presynaptic terminal favors an increase in the size of the reserve pool, whereas a predominance of synapsin Ib favors a dispersal of this pool (c).

dopaminergic neurons. It will be interesting to determine the molecular mechanisms underlying these differences, as well as to examine the functions of synapsins at synapses that use other neurotransmitters.

Summary and Conclusions

Studies of TKO mice devoid of all three synapsin genes reveal that synapsins have a variety of presynaptic functions which depend on the type of neurotransmitter employed. At hippocampal glutamatergic synapses, synapsins are responsible for maintaining an RP of SVs that are mobilized during synaptic depression, enhancing the duration and frequency range of glutamatergic synaptic transmission. Synapsin IIa seems to be mainly responsible for this action, with synapsin Ib serving as a dominant negative isoform that reduces the RP by heterodimerizing with synapsin IIa and reducing vesicle cross-linking. At GABAergic synapses, synapsins influence both the RRP and the RP of SVs. The synapsin isoforms involved in these functions are not yet clear. At dopaminergic synapses, synapsins also appear to be responsible for maintaining a reserve vesicle pool that is mobilized by cocaine and other psychostimulant drugs. The fact that synapsins appear to play different roles in various synapse types indicates that the molecular mechanisms governing the makeup of functionally defined vesicular populations must differ between these synapses in multiple ways. Elucidation of these mechanisms is one of the targets of intense contemporary research concerning synaptic physiology. The identity of the synapsin isoforms participating in defining the physiology of the synapses is of prime importance in this context, as are the precise molecular pathways by which the synapsins exert their effect.

Gaps and Future Directions

Although the work summarized here provides a coherent picture of the functions of the synapsin protein family, much remains to be understood. First, transmission has been examined at only a very small number of synapses, and it will be important to determine whether the conclusions derived from these initial studies can be generalized to the many thousands of other synapse types found within the brain. Second, synapsins have transmitter-specific functions, and it will be important to find out whether these proteins serve additional functions at synapses employing transmitters other than glutamate, GABA, or dopamine. Third, among the ten known synapsin isoforms, clear-cut roles have been identified for only two, namely synapsins Ib and IIa, which serve opposing functions in regulating the RP of glutamatergic SVs. Future studies certainly will need to identify the function of the eight other isoforms, as well as to consider whether synapsins Ib and IIa have additional roles. Finally, while the synaptic defects in the synapsin TKO mice provide plausible explanations for some of the behavioral phenotypes of these mice, additional work will be required to actually establish cause–effect relationships in these cases.

See also: Presynaptic Development: Functional and Morphological Organization; Presynaptic Regulation by Liprins; Synapsins; Synaptic Vesicles; Synaptic Transmission: Models; Vesicle Pools.

Further Reading

Gitler D, Takagishi Y, Feng J, et al. (2004) Different presynaptic roles of synapsins at excitatory and inhibitory synapses. *Journal of Neuroscience* 24: 11368–11380.

Gitler D, Xu Y, Kao H-T, et al. (2004) Molecular determinants of synapsin targeting to presynaptic terminals. *Journal of Neuroscience* 24: 3711–3720.

Hilfiker S, Pieribone VA, Czernik AJ, et al. (1999) Synapsins as regulators of neurotransmitter release. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 354: 269–279.

Hosaka M and Sudhof TC (1999) Homo- and heterodimerization of synapsins. *Journal of Biological Chemistry* 274: 16747–16753.

Llinas R, Gruner JA, Sugimori M, McGuinness TL, and Greengard P (1991) Regulation by synapsin I and Ca(2+)-calmodulin-dependent protein kinase II of the transmitter release in squid giant synapse. *Journal of Physiology* 436: 257–282.

Rizzoli SO and Betz WJ (2005) Synaptic vesicle pools. *Nature Reviews Neuroscience* 6: 57–69.

Rosahl TW, Spillane D, Missler M, et al. (1995) Essential functions of synapsins I and II in synaptic vesicle regulation. *Nature* 375: 488–493.

Ryan TA, Li L, Chin LS, Greengard P, and Smith SJ (1996) Synaptic vesicle recycling in synapsin I knock-out mice. *Journal of Cell Biology* 134: 1219–1227.

Südhof TC, Czernik AJ, Kao H-T, et al. (1989) Synapsins: Mosaics of shared and individual domains in a family of synaptic vesicle phosphoproteins. *Science* 245: 1474–1480.

Venton BJ, Seipel AT, Phillips PEM, et al. (2006) Cocaine increases dopamine release by mobilization of a synapsin-dependent reserve pool. *Journal of Neuroscience* 26: 3206–3209.

Synaptic Capture and Tagging

K C Martin, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Synaptic plasticity, the process whereby neurons alter the strength of their synaptic connections with experience, provides a cellular model for learning and memory in the brain. Since the vertebrate central nervous system consists of approximately 10^{11} neurons and 10^{14} synapses, using the synapse as the unit of plasticity greatly expands the information-processing capacity of the brain. The fact that long-lasting forms of plasticity, like long-lasting forms of memory, require changes in gene expression at the level of transcription, however, raises a cell biological paradox: how can a transcription-dependent process function to alter efficacy at some but not all synapses made by a given neuron? Although it not difficult to envision how short-term changes in synaptic strength, which are dependent on covalent modifications, might occur in a synapse-specific manner, long-lasting synaptic plasticity, which is dependent on transcription within a single, shared nucleus, presents a greater challenge. The synaptic tagging hypothesis provides a resolution to this challenge by proposing that the products of gene expression are delivered throughout the cell but only function to increase synaptic strength at synapses that have been 'tagged' or 'marked' by previous synaptic activity. Studies in a number of systems, including *Aplysia* sensory–motor synapse, crayfish neuromuscular junction, and rodent hippocampal synapses, have described the phenomenon of synaptic tagging and have defined roles for and broad characteristics of the products of transcription ('plasticity factors' or 'plasticity-related proteins') that are 'captured' at synapses tagged by previous synaptic activity. While there is a wealth of evidence showing that synaptic tagging occurs and contributes to synapse-specific, transcription-dependent plasticity, the nature and identity of the synaptic tag remain under investigation. Perhaps not surprisingly, initial studies indicate that a variety of 'tags' exist and can be recruited in distinct cell types by distinct stimuli.

Evidence for Synaptic Tagging

Synaptic Tagging in Rodent Hippocampal Synapses

The first evidence for synaptic tagging during long-term potentiation (LTP) in rodent hippocampus came from a classic series of studies done by Frey and Morris in 1997. In these experiments, the authors showed that when stimuli that produce protein synthesis-dependent, late-phase LTP (L-LTP) were applied to one input (S1) to CA1 neurons in the Schaffer collateral pathway of the hippocampus, and then stimuli that normally would produce only early-phase LTP (E-LTP) were applied to a second input (S2), L-LTP was observed at both S1 and S2, as long as the stimuli to S1 and S2 were delivered within a discrete time window. Specifically, the weaker stimulus to S2 had to be delivered 1–2 h before, or less than 2.5–3 h after, the stronger stimulus was delivered to S1. These results indicated that the weaker stimulus received by S2 created a synaptic tag that could 'capture' the products of gene expression, resulting in persistent synaptic strengthening at S2. They further indicated that the synaptic tag persisted for at least 1–2 h and that the products of transcription, referred to as plasticity-related proteins, persisted for at least 2.5–3 h following stimulation. It is worth noting that similar synaptic tagging experiments performed using *in vivo* electrical recording (i.e., at physiological temperature) have indicated that the tag persists for approximately 30 min.

Kauderer and Kandel, and more recently, Sajikumar and Frey, subsequently demonstrated synaptic tagging during long-term depression (LTD) of hippocampal synapses. In these studies, they found that a strong low-frequency stimulation induced long-lasting, protein synthesis-dependent, input-specific LTD (L-LTD) of hippocampal synapses and that this L-LTD could be captured at a second synaptic pathway receiving a weak low-frequency stimulus that on its own induced only transient, protein synthesis-independent LTD. As was the case with LTP, experiments in which the strong and weak low-frequency stimuli were given with various time delays revealed that the synaptic tag persisted for 1–2 h. Since previous studies had shown a requirement for dopamine during L-LTD, Sajikumar and Frey tested a role for dopamine during synaptic tagging during LTD and found that dopamine application was sufficient to induce synthesis of plasticity-related proteins, which could then be captured by weak low-frequency stimulation to produce L-LTD. They referred to this process as heterosynaptic tagging, reflecting the role of modulatory, heterosynaptic dopaminergic inputs in synaptic tagging of glutamatergic hippocampal synapses. The emerging finding is that modulatory inputs, such as dopamine, can contribute to the synthesis of plasticity-related proteins, but that the synaptic tag is set by the glutamatergic input.

Synaptic Tagging at *Aplysia* Sensory–Motor Synapses

To test the synaptic tag hypothesis at the level of a single cell, Martin and colleagues developed a culture system in which a single *Aplysia* sensory neuron formed synaptic contacts with two spatially separated motor neurons. They found that local application of five pulses of serotonin to the connections made onto one of the motor neurons produced an increase in synaptic strength (measured as an increase in excitatory postsynaptic potential (EPSP) amplitude) that depended on transcription in the sensory neuron. With this culture system of branch-specific long-term facilitation (LTF) of *Aplysia* sensory–motor synapses in hand, they then tested the synaptic tagging hypothesis by asking whether, following induction of long-term, transcription-dependent facilitation at one branch of a bifurcated *Aplysia* sensory neuron, a synaptic tag could be generated at the opposite branch using a subthreshold stimulus. As a subthreshold stimulus, they utilized a single application of serotonin, which on its own produced only a transient increase in synaptic strength. When five pulses of serotonin were applied to the connection made onto one motor neuron, and then a single pulse of serotonin was applied to the opposite connection, that single application of serotonin became capable of producing an increase in EPSP amplitude that persisted for 72 h. This facilitation was referred to as synaptic capture, since the single pulse of serotonin was capable of 'capturing' a transcription-dependent, long-term increase in synaptic strength. LTF could be captured if the single pulse of serotonin was applied 1 h after the five pulses of serotonin were applied to the opposite branch, but could not be captured if the single pulse was applied 4 h after the five applications of serotonin. A converse set of experiments revealed that synaptic capture occurred if the single pulse was given immediately or 1 or 2 h before, but not 4 h before the five pulses of serotonin. These results indicated that the synaptic tag persisted for at least 2 h and that the products of gene expression (plasticity-related factors) persisted for at least 1 h following synaptic stimulation.

Temporal Synaptic Tagging at the Crayfish Neuromuscular Junction

Zucker and colleagues described a form of 'temporal' tagging at the crayfish neuromuscular junction, in which the neuron integrates synaptic stimulation at a single synaptic connection over time. Specifically, they found that presynaptic hyperpolarization-activated cation channels, or I_h channels, become activated during a tetanus and that this activation generates a synaptic tag, which allows the motor neuron to increase transmitter output by subsequently responding to additional stimuli that would otherwise be ineffective. If the activation of the I_h channels occurs concomitant with elevations in presynaptic calcium, tagged synapses can undergo protein-synthesis-dependent, LTF in response to subthreshold stimuli applied at least 1 h after the time the tag was set. While defined as a synaptic tag, there are two important differences between this type of tag and the tag defined for LTF in *Aplysia* and LTP in rodent hippocampus. First, LTF at the crayfish neuromuscular junction does not depend on transcription, but rather on local protein synthesis, and thus this tag does not interact with the cell-wide products of transcription but instead with locally synthesized proteins to produce persistent synaptic strengthening. Second, this tag does not provide spatial specificity (although in another context, it could do so), but rather serves as a 'temporal tag' that provide a history of activity at the synapse, and can combine with subthreshold stimulus over a discrete time window to produce long-lasting changes in synaptic efficacy.

Cross-Tagging

The discovery that both L-LTP and L-LTD can be captured by synaptic tags raised the question of how a single neuron would integrate distinct tags generated by LTP- or LTD-inducing stimuli. Sajikumar and Frey performed a set of experiments to address the functional interactions between LTP and LTD synaptic tagging, which they referred to as cross-tagging. These experiments showed that the products of transcription, plasticity-related proteins, generated following induction of either L-LTP or L-LTD at one pathway could be captured by either LTP- or LTD-induced synaptic tags at a second pathway. These studies indicated that the plasticity-related proteins induced during L-LTP or L-LTD are nonspecific and that the specificity of plasticity is generated by the synaptic tag. This, in turn, focuses attention on the particular nature of the synaptic tag that is produced in response to specific stimuli. A number of studies have begun to identify molecules that function as specific tags during LTD or LTP.

In studying the effects of low-frequency stimulation (LFS) on L-LTP, Young and Nguyen found that LFS applied before a capturing stimulus (a single 100 Hz tetanus) inhibited the capture of L-LTP at that pathway. Further studies indicated that this inhibition occurred at the level of the synaptic tag and indicated that the tag undergoes a period during which it is vulnerable to disruption. Their results, however, were consistent with the rapid generation of a synaptic

tag and the capture of plasticity-related proteins within minutes of synaptic stimulation.

While cross-tagging experiments similar to those performed by Frey and colleagues in hippocampus have not been done at the level of single cells in *Aplysia*, Guan and colleagues undertook a series of experiments to determine how stimuli producing LTF at one branch of an *Aplysia* sensory–motor synapses (five pulses of serotonin) were integrated with stimuli producing LTD (five pulses of the neuropeptide FMRFamide) at another branch. They found that simultaneous application of the facilitatory neurotransmitter serotonin at one branch with the inhibitory neuropeptide FMRFamide at an opposite branch led to inhibition of LTF at the branch receiving serotonin but expression of LTD at the branch receiving FMRFamide. Further investigation indicated that this integration of serotonin- and FMRFamide-induced signals occurred in the nucleus, with FMRFamide inducing a pattern of histone deacetylation that promotes depression and inhibits facilitation. These studies thus indicate that plasticity-related proteins are specific for LTF or LTD in this system.

Synaptic Tagging and Competitive Maintenance

In a further test of the synaptic tagging hypothesis, Fonseca and colleagues reasoned that if long-lasting synaptic plasticity involved both the setting of synaptic tags and the induction of plasticity-related proteins, which they referred to as plasticity factors, then reducing the concentration of plasticity-related proteins would lead to a competition between tagged synapses. They tested this idea using distinct doses of protein synthesis inhibitors to decrease the concentration of plasticity factors and found that under these conditions, capture of LTP at one pathway decreased the magnitude of LTP at the pathway where L-LTP had been induced in a dose-dependent manner, suggesting that the captured synapses were competing with the initially stimulated synapses for a limited pool of plasticity factors. In addition, these investigators found that the greater the distance between inputs, the less the competition for plasticity factors, a finding also consistent with the idea of competitive maintenance. Thus, when L-LTP was induced in apical dendrites and captured in basal dendrites, competition was not observed. These findings not only support the synaptic tagging hypothesis, but also indicate that synapses must compete for plasticity-related factors. They further indicate that such competition may be effectively restricted to larger compartments of neurons, for example, to apical and/or basal dendrites.

Restriction of Synaptic Tagging within Dendritic Compartments

After demonstrating that synaptic tagging and capture could occur in basilar dendritic compartments of CA1 hippocampal neurons in a manner similar to that described in apical dendrites, Alarcon et al. asked whether synaptic capture could occur across apical and basilar dendritic compartments. They found that capture could occur across dendritic compartments only when the intensity of the capturing stimulus was increased, such that capture occurred only when a stimulus that induces intermediate-term potentiation (two 100 Hz tetani with a 30 s intertrain interval) was used and not when a stimulus that would induce early LTP (one 100 Hz tetanus) was used. Together, these results point to a spatial specificity of synaptic capture, in which tags can capture plasticity-related proteins within discrete functional dendritic compartments. This process has been referred to by Tonegawa and colleagues as localized capture associativity.

What Is the Nature of the Synaptic Tag?

While the phenomenon of synaptic tagging and capture has been described by many research groups, identifying the molecules that form the tag, or the plasticity-related proteins that are captured by the tag, has proven more challenging. In part, this difficulty stems from the fact that synaptic tagging and capture occur within spatially restricted compartments and are thus not amenable to biochemical or molecular biological investigation of large populations of neurons. Nonetheless, some progress has been made in identifying molecular components of tagging and capture.

Is Tagging Dependent on Local Translation or Local Protein Degradation at the Synapse?

The activation of local translation meets many of the requirements for a tag that can integrate activity in both temporal and spatial domains. Protein translation at the synapse can alter the composition of the synapse over extended time periods and thereby leave a relatively persistent trace of synaptic activity. In addition, activating translation locally provides a spatial tag for gene expression, because messenger RNAs that are delivered from the cell soma will be preferentially translated at sites at which translation is active. Notably, Ostroff and colleagues have found that the percentage of spines that contain polyribosomes increases markedly in CA1 neurons after LTP induction. Furthermore, after the induction of LTP, the area of the postsynaptic density is significantly larger in polyribosome-containing spines than in spines that lack them.

Despite evidence in hippocampus and in *Aplysia* sensory–motor neurons that local translation plays a role in synapse-specific long-term plasticity, it is also clear that local translation is not required to generate a synaptic tag. Thus, the initial studies demonstrating synaptic tagging in hippocampus by Frey and Morris directly demonstrated that translation could be inhibited during generation of a synaptic tag. Similarly, in the *Aplysia* bifurcated sensory–motor culture system, local translation is not required at the captured branch for 24 h facilitation. It is, however, required for facilitation at 72 h, which is accompanied by the growth of new synaptic connections. These findings indicate that local translation is required not for the initial setting of the synaptic tag, but rather for stabilizing new synaptic growth to create persistent synaptic plasticity. Local translation may thus be found to be required for synaptic tagging at hippocampal synapses when the captured LTP is examined over longer time domains.

Regulated protein degradation provides another means of altering the macromolecular composition of synapses. A role for ubiquitin-mediated protein degradation via the proteasome during plasticity of rodent hippocampal synapses as well as during LTF of *Aplysia* sensory–motor synapses has been described. Zhao and colleagues have shown that serotonin induces increases in polyubiquitination of synaptic proteins in sensory–motor cultures. It is therefore possible that local, ubiquitin-mediated protein degradation of specific molecules in response to synaptic stimulation creates a synaptic tag by altering the composition of the synapse and that this tag combines with the products of transcription to produce a persistent type of synaptic plasticity.

A Role for Specific Kinases in the Generation of Synaptic Tags

Roles for a number of kinases during synaptic tagging have been mathematically modeled by Byrne and colleagues and experimentally demonstrated by a number of groups. Casadio and colleagues found that although 24 h synaptic capture was not dependent on local translation in *Aplysia* sensory–motor synapses, it did depend on activation of protein kinase A (PKA). Studies in hippocampal synapses have also revealed a requirement for PKA during synaptic tagging and capture. Nguyen and colleagues have shown that genetic or pharmacological reduction of PKA activity blocks synaptic capture and that activation of PKA is sufficient to produce a synaptic tag that can capture L-LTP of hippocampal synapses. Abel and colleagues went on to show that the anchoring of PKA to synaptic sites, mediated by interactions with the A-kinase anchoring protein (AKAP), was required for efficient synaptic tagging.

Taking advantage of the finding that synaptic tagging and capture are restricted to discrete dendritic compartments (apical and basilar), and that cross-tagging between LTP and LTD could occur, Frey and colleagues found that calcium/calmodulin-dependent protein kinase II (CaMKII) activity was required for LTP but not LTD tagging, and that p42/44 mitogen-activated protein kinase (MAPK) activity was required for LTD but not LTP tagging in apical dendrites of CA1 hippocampal neurons. In contrast, they found that either PKA or protein kinase Mz, a persistently active PKC isoform, served as LTP-specific tags in basilar dendrites. Together, these findings indicate that distinct molecular tags may be recruited by distinct stimuli and within distinct dendritic compartments.

Conclusions

Synaptic tagging allows the synapse, rather than the nucleus, to be the unit of long-lasting neuronal plasticity. Studies of the phenomenon of synaptic tagging have not only confirmed its existence but also revealed a number of important characteristics, including cross-tagging, competitive maintenance, and localized capture associativity, in which tagged synapses within a dendritic subcompartment compete for a common group of plasticity-related proteins. Together, these findings provide insight into the mechanisms whereby information is stored in the brain. They have given rise to refined models of information storage, including the clustered plasticity model of long-term memory engrams proposed by Govindarajan et al. and the late-associative reinforcement model proposed by Frey and colleagues.

The evidence that synaptic tagging does occur has also kindled deep interest in identifying the molecular nature of the tag, and in understanding the function of synaptic tagging during learning and memory. Progress is being made on both fronts. Specific roles for local translation and for specific kinases have been identified during synaptic tagging and capture, and future years promise to increase the number of identified tags and plasticity-related proteins. Behavioral and *in vivo* electrophysiological recording paradigms are being developed by a number of groups to investigate the existence and function of synaptic tagging in living animals. Together, the results of these investigations promise to shed light on the mechanisms whereby synaptic tagging functions to allow information to be stored at the level of the synapse.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Depression; Transcription Factors in Synaptic Plasticity and Learning and Memory; Translational Regulation at the Synapse.

Further Reading

Barco A, Alarcon JM, and Kandel ER (2002) Expression of constitutively active CREB protein facilitates the late phase of long-term potentiation by enhancing synaptic capture. *Cell* 108: 689–703.

Beaumont V, Zhong N, Froemke RC, Ball RW, and Zucker RS (2002) Temporal synaptic tagging by I_h activation and actin: Involvement in long-term facilitation and cAMP-induced synaptic enhancement. *Neuron* 33: 601–613.

Casadio A, Martin KC, Giustetto M, et al. (1999) A transient, neuron-wide form of CREB-mediated long-term facilitation can be stabilized at specific synapses by local protein synthesis. *Cell* 99: 221–227.

Fonseca R, Nägerl UV, Morris RGM, and Bonhoeffer T (2004) Competing for memory: Hippocampal LTP under regimes of reduced protein synthesis. *Neuron* 44: 1011–1020.

Frey U and Morris RGM (1997) Synaptic tagging and long-term potentiation. *Nature* 385: 53–56.

Frey U and Morris RGM (1998) Weak before strong: Dissociating synaptic tagging and plasticity-factor accounts of late-LTP. *Neuropharmacology* 37: 545–552.

Frey U and Morris RGM (1998) Synaptic tagging: Implications for late maintenance of hippocampal long-term potentiation. *Trends in Neuroscience* 5: 181–188.

Govindarajan A, Kelleher RJ, and Tonegawa S (2006) A clustered plasticity model of long-term memory engrams. *Nature Reviews Neuroscience* 7: 575–583.

Martin KC, Casadio A, Zhu H, et al. (1997) Synapse-specific, long-term facilitation of Aplysia sensory to motor synapses: A function for local protein synthesis in memory storage. *Cell* 91: 927–938.

Martin KC and Kosik KS (2002) Synaptic tagging—who's it? *Nature Reviews Neuroscience* 3: 813–820.

Morris RGM (2006) Elements of a neurobiological theory of hippocampal function: The role of synaptic plasticity, synaptic tagging and schemas. *European Journal of Neuroscience* 23: 2829–2846.

Reymann KG and Frey JU (2007) The late maintenance of hippocampal LTP: Requirements, phases, "synaptic tagging," "late-associativity" and implications. *Neuropharmacology* 52: 24–40.

Sajikumar S and Frey JU (2004) Late-associativity, synaptic tagging, and the role of dopamine during LTP and LTD. *Neurobiology Learning and Memory* 82: 12–25.

Sajikumar S, Navakkode S, and Frey JU (2007) Identification of compartment- and process-specific molecules required for "synaptic-tagging" during long-term potentiation and long-term depression in hippocampal CA1. *Journal of Neuroscience* 27: 5068–5080.

Sossin WS (1996) Mechanisms for the generation of synapse specificity in long-term memory: The implications of a requirement for transcription. *Trends in Neuroscience* 19: 215–218.

Synaptic Depression

M A Xu-Friedman, University at Buffalo, State University of New York, Buffalo, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Definition

Synapses, the connections between neurons, are highly plastic. Synaptic plasticity arises from a number of mechanisms, with many possible triggers and targets, that can lead to increases or decreases in synaptic efficacy over different timescales. This article focuses on forms of plasticity that cause decreases in synaptic efficacy over short timescales that are triggered by presynaptic activity (Figure 1). These forms of plasticity are collectively known as short-term synaptic depression, and represent a group of mechanisms that can result from a variety of changes both presynaptically and postsynaptically. Other forms of synaptic plasticity (e.g., short-term facilitation, posttetanic potentiation, long-term potentiation, long-term depression, neuromodulation) have different triggers and different targets, and are covered elsewhere in this Encyclopedia.

Short-term depression is widespread throughout the nervous system. Because depression is triggered merely by activity, and takes several seconds to recover, it affects synapses constantly. As a result, depression plays a critical role in the moment-by-moment processing of information that the brain is responsible for during normal behavior. Therefore, in order to understand how the brain processes information, depression and other forms of short-term plasticity must be considered in addition to the longer-term plasticities that may match the more conventional definitions of 'memory.'

Which Synapses Show Depression?

Depression was found in the earliest recordings of synaptic potentials by Feng and by Eccles, Katz, and Kuffler at the neuromuscular junction in the early 1940s. Depression has since been observed at many synapses throughout the nervous system. It is most easily characterized in reduced preparations such as brain slices by stimulating a presynaptic axon with one or more conditioning pulses, waiting some interval, and then stimulating with a test pulse (Figure 1). The postsynaptic response can be quantified using current clamp (to measure the postsynaptic potential) or voltage clamp (to measure the postsynaptic conductance changes).

The amount of depression differs for different synapses, and depends on the recording conditions. Typically, depression is greater at synapses that release neurotransmitter vesicles with high probability. This condition can be mimicked even at synapses that normally show little or no depression, by increasing the external calcium concentration (e.g., from 2 to 4 mM). In a complementary way, depression can be reduced and even eliminated at synapses that normally show depression, by decreasing external calcium (e.g., from 2 to 0.5 mM). These manipulations are thought to change the amount of presynaptic calcium influx, and thus the probability of vesicle release, which affects all the mechanisms that are proposed to underlie depression (described later). These experimental manipulations are relevant to the conditions in the intact brain, as for example neuromodulators may cause presynaptic calcium influx to decrease, which can flip a synapse from depressing to facilitating.

Mechanisms of Depression

Depletion

One of the first mechanisms to be proposed for depression was depletion of neurotransmitter, put forth by Liley and North in 1953 (Figure 2(a)) As more has been learned about synaptic structure, this idea has been further refined. Neurotransmitter is packaged in vesicles, and there is only a limited pool of vesicles (the 'readily releasable pool') that are docked at the presynaptic membrane and ready to release their contents into the synaptic cleft. Action potentials in the presynaptic axon cause vesicles within this pool to release their contents, thereby reducing the size of the pool (Figure 2(a)). The presynaptic terminal recovers from depletion as the pool is restocked through multiple pathways of reuptake or priming of reserve pools of vesicles. Because replenishment takes time, the readily releasable pool remains at least partially depleted for a period usually estimated as being on the scale of a few seconds. If a second action potential arrives during that period of recovery, then fewer vesicles are ready to be released. The reduced pool size causes the amount of neurotransmitter release to be smaller. Thus the characteristics of synapses that are likely to enhance depletion would be a high initial probability of vesicle fusion (such that many of the releasable vesicles are released), and a slow recovery rate (such that there is an opportunity for the smaller pool to have a measurable impact).

Despite the simplicity of the idea, the small scale of synaptic processes makes it difficult to observe directly. The difficulties are compounded by the fact that, as yet, it is not known for even the best-studied synapses how many vesicles are actually ready to be

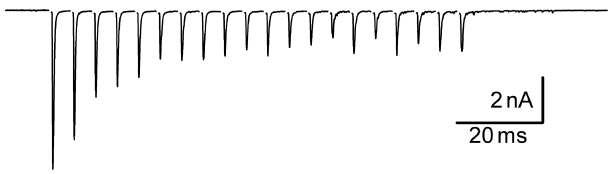

Figure 1 An example of a depressing synapse. A bushy cell in the anteroventral cochlear nucleus was recorded from in voltage clamp, and a single presynaptic auditory nerve fiber was stimulated at 100 Hz. The sharp negative peaks are inward currents carried by AMPA-type glutamate receptors, which start out very large, but get progressively smaller. Several of the mechanisms discussed in the main text have been proposed at various times to account for this depression. This is an average of four traces recorded in voltage clamp at 34 °C in a brain slice taken from a P21 mouse. Data from Xu-Friedman MA and Regehr WG (2005) Dynamic-clamp analysis of the effects of convergence on spike timing. II: Few synaptic inputs. *Journal of Neurophysiology* 94: 2526–2534, used with permission.

released. Electron micrographs show some vesicles very close to the membrane at active zones, but it is not likely that physical location is the only criterion. In addition, it is not known how many vesicles are released after an action potential invades the presynaptic terminal. There has been some debate over whether only a single vesicle can be released per synapse, or whether multiple vesicles can be released. Part of the reason this is difficult to resolve is that most researchers measure vesicle fusion using the simplest technique, which is to measure the postsynaptic current in voltage clamp. However, this is only an indirect measurement, which is really of the number of activated, open postsynaptic receptors. The relationship between receptor activation and the amount of neurotransmitter released is not well known.

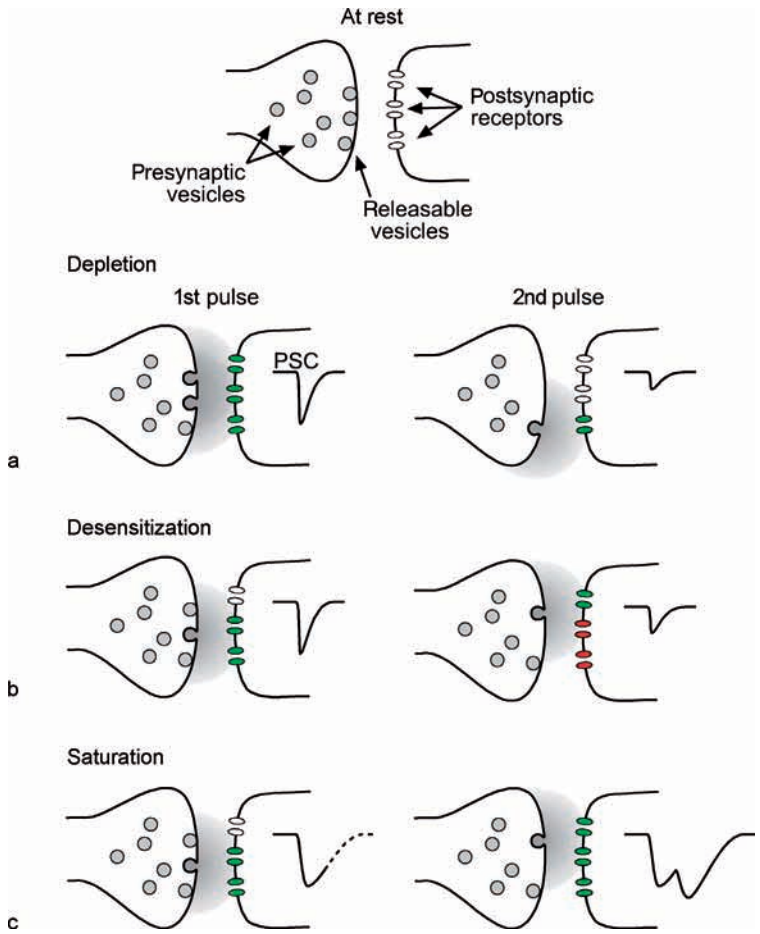

Figure 2 Principle mechanisms of depression. Top, at rest, the presynaptic terminal has a pool of releasable vesicles near the membrane, as well as a reserve pool, and the postsynaptic side has a set of closed ligand-gated receptors. (a) Depletion. The first action potential triggers the fusion of a large fraction of the releasable pool with the presynaptic membrane, and the activation of postsynaptic receptors (in green). If the second action potential arrives before the releasable pool has been replenished, fewer vesicles will be released. (b) Desensitization. Neurotransmitter released during the first action potential first activates (green), then desensitizes (red) the postsynaptic receptors. If a second action potential arrives before the receptors return to the resting state, then fewer will be able to conduct current, and the PSC will be depressed. (c) Saturation. Neurotransmitter released during the first action potential binds to the receptors for a prolonged period (green). If a second action potential arrives before ligand unbinds, only initially unbound receptors will be available to provide additional current.

To get around these difficulties, different techniques have been applied. Some of the best evidence for depletion comes from recordings made in a very reduced preparation from the retina by Gary Matthews and colleagues. The retinal cells they studied (retinal bipolar cells) have large presynaptic terminals, which can be directly studied with electrophysiology techniques. These terminals also have a characteristic ultrastructure, called the synaptic ribbon, which makes it easier to identify vesicles that are likely to be in the releasable pool. Matthews and colleagues quantified the amount of vesicle fusion using the capacitance of the cell, which increases as a result of the small neurotransmitter vesicles fusing with the membrane, thereby adding their capacitance to it. By depolarizing the retinal cell for different periods of time and measuring transient increases in capacitance, they estimated the number of vesicles that fuse with the membrane. They could cause fusion of larger number of vesicles using longer depolarizations, but only up to a certain limit, suggesting that the releasable pool size was limited. In addition, Matthews and colleagues counted the number of vesicles positioned along the ribbon. The correlation between the size of pool measured anatomically and physiologically was remarkably strong. These results suggested that different pools of vesicles do exist, have a physical basis, and can be depleted through activity.

One question is whether the findings in the rather specialized retinal bipolar cells apply to other cells in the nervous system. Retinal bipolar cells are unusual in that they release neurotransmitter in a graded fashion, rather than in response to action potentials. In addition, most cells lack the bipolar cell's synaptic ribbon. Some attempt has been made to examine cells in the rest of the brain. Capacitance techniques have been successfully applied to the calyx of Held synapse in the auditory pathway. However, the morphological correlate of the releasable pool has been more difficult to distinguish than in the retinal synapse studied by Matthews and colleagues. Another direct method to look for depletion is to drive a synapse to release its complement of vesicles, and then examine the synapse using electron microscopy to look for changes in the spatial locations or numbers of vesicles. Unfortunately, it is not possible to get 'before' and 'after' pictures of a synapse, because the fixation required for electron microscopy destroys the tissue. In addition, even comparisons between synapses are complicated by the presence of recovery mechanisms that operate on the timescale of seconds, and it may be difficult to deplete a synapse to a measurable extent. One method has been to eliminate recovery mechanisms entirely by using temperature-sensitive mutants of the recovery machinery. A temperature-sensitive mutation of the dynamin protein has been described in *Drosophila* (called *shibire*). When these organisms are warmed, vesicles cannot be taken back up, and electron micrographs reveal a lack of vesicles, which matches a drop in the size of synaptic potentials. While this shows that vesicles are consumed by activity, it has not been shown yet how this relates to a readily releasable pool.

For some synapses, an indirect approach can also be used to examine the loss of vesicles, called 'variance-mean analysis.' This method makes certain assumptions about the statistics of vesicle release, and uses the mean and variance of synaptic currents or potentials to estimate the total number of vesicles. This approach has been used by Wade Regehr and colleagues at the climbing fiber synapse in the cerebellum to show that the number of vesicles available for release is lower after bouts of activity. However, the preconditions that allow this statistical analysis to be used are difficult to meet experimentally, and it has provided mixed results at other synapses.

Desensitization

A second form of depression relies on a postsynaptic mechanism – receptor desensitization. This occurs when a ligand-gated receptor is exposed to its ligand for an extended period, whereby it transitions through the open state to a fully bound, but nonconducting state, referred to as 'desensitized.' Acetylcholine receptors at the neuromuscular junction were among the first to be shown to exhibit desensitization, and, subsequently, all ionotropic neurotransmitter receptors were also found to desensitize. Receptor desensitization can be examined in great detail using the outside-out patch technique. These experiments have demonstrated that α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA)-type glutamate receptors desensitize particularly rapidly compared to other ligand-gated ion channels, and among the AMPA receptors, some subunit types desensitize more rapidly than others.

It is possible that postsynaptic receptor desensitization could play a role in depression. Neurotransmitter released during the first action potential initially activates receptors, which causes synaptic currents and potentials (**Figure 2(b)**). Then, the receptors may enter a desensitized state, provided ligand does not unbind before the channel changes conformation and enters the nonconducting state. If a second action potential were to arrive during this period, then postsynaptic receptors may not be ready to respond to the fresh elevations in ligand concentration. Thus,

synapses would tend to be most subject to desensitization provided receptor channels enter desensitized states rapidly, and persist in them for a long time. In addition, desensitization is more likely to be found at synapses where neurotransmitter lingers in the synaptic cleft. Recovery would depend on the rate of neurotransmitter removal from the cleft, as well as the rate of ligand unbinding.

The first evidence that desensitization could actually play a role in depression came from the work of Laurence Trussell and colleagues. They examined a synapse in the chick auditory brain stem that had both characteristics mentioned above. First, the synapse expresses AMPA receptors that desensitize particularly rapidly ($GluR4_{flop}$). Second, the synapse is calyceal in structure, with many active zones clustered together. These active zones do not appear to express glutamate transporters, so the clearance of glutamate probably takes place through simple diffusion. A number of experimental approaches lent support to the idea that desensitization contributes to depression there, including (1) that depression increases with the amount of neurotransmitter released, (2) that depression recovers relatively rapidly (50–100 ms), (3) that miniature postsynaptic currents (mPSCs) that occur right after a large synchronous PSC are reduced in size, and (4) that pharmacological compounds that prevent desensitization also reduce depression. These drugs include cyclothiazide and aniracetam, which appear to interfere with various kinetic properties of the AMPA receptor, including the conformational changes that lead to desensitization.

Since then, desensitization has been found to contribute to depression at a number of additional synapses, in the thalamus, cortex, and cerebellum, as well as in invertebrates. These synapses have been investigated in less detail than the chick calyx, but do appear to share some characteristics. Further investigation of these as well as comparison with nondesensitizing synapses will help reveal the critical conditions. In addition it remains to be seen whether desensitization contributes to depression at synapses expressing different receptor types, such as γ-aminobutyric acid (GABA), glycine, and *N*-methyl-D-aspartate (NMDA)-type glutamate receptors. Pharmacological tools that interfere with desensitization of these receptor types would be particularly helpful.

Saturation

A third mechanism that affects depression is postsynaptic receptor saturation. This occurs when an action potential leads to neurotransmitter release, and the binding and activation of a pool of receptors. If a second action potential arrives while these receptors are still bound, then they provide no additional synaptic current because they are already occupied (**Figure 2(c)**). This state is referred to as saturation. In the extreme case, all the receptors are occupied, but saturation would also have an impact even if only a subset of receptors in the active zone are claimed by the first action potential.

Saturation is most likely to be important at synapses with certain properties. First, synapses are more likely to be saturated if they express receptor types that remain bound to ligand or are exposed to ligand for an extended period. In addition, saturation should depend on the spatial distribution of postsynaptic receptors and their physical distance from the site of neurotransmitter release. For example, a small synapse would likely undergo significant saturation even if only one vesicle were to be released. By contrast, a large synapse may not experience saturation unless vesicles are consistently released in a limited area, or if many vesicles are released at once. As yet, there is little experimental evidence to support the relationship between saturation and synapse size, as the ultrastructure has been well measured for only a few synapses, and the amount of saturation has proven difficult, or at least controversial, to estimate.

The most likely example of saturation causing depression is with NMDA-type glutamate receptors. NMDA receptors typically have very slow kinetics, and the synaptic currents passed by NMDA receptors last for several hundred milliseconds. If a second action potential arrives during that period, it typically has a smaller amplitude. It must be pointed out that the second synaptic current rides on top of the first, so the net effect may be a summed synaptic current of similar amplitude to the first. Thus, saturation, while technically resulting in depression, does not really suppress synaptic strength in the same spirit as do depletion and desensitization. Rather, saturation effectively results in a prolongation of synaptic currents.

There is still considerable disagreement whether NMDA receptors actually are saturated. Low-affinity antagonists have been used to reduce saturation, and the evidence appears fairly clear that NMDA receptors are saturated at some synapses, such as the retinal ganglion cell synapse in the lateral geniculate nucleus. GABA and glycine receptors are usually somewhat faster than NMDA, but can show a similar phenotype (i.e., a relatively smaller second peak that rises to the same absolute height). Further study with low-affinity antagonists may help shed light on this issue.

Under special conditions, saturation can actually play a role in reducing depression. This happens

because saturation reduces the amount of receptor activation for a given amount of neurotransmitter released. It is most easily understood with a very artificial example. Suppose a synapse is completely saturated by the total neurotransmitter contained in two vesicles. Then there is no difference in postsynaptic response whether the presynaptic terminal releases eight vesicles (say from a fully stocked releasable pool) or just two. That means that a depleted synapse will show no obvious depression, provided it can still squeeze out two vesicles. The principles of this example would hold under less extreme conditions, as long as there is saturation and multivesicular release at individual release sites. Evidence for this phenomenon was found by Wade Regehr and colleagues at the climbing fiber synapse in the cerebellum. The overall effect of saturation is to reduce the effects of depletion, and accelerate the synapse's recovery.

Other Mechanisms

The mechanisms described above have received the strongest experimental support at a number of synapses. There are additional mechanisms that have been proposed to underlie depression in more special conditions, or have not been described as widely – depletion of extracellular calcium, calcium channel inactivation, and incomplete vesicle refilling.

Depletion of extracellular calcium Synapses are usually present in a dense neuropil, where the amount of extracellular space is quite small. Thus, the activity of neurons could, in principle, affect the ionic concentrations in extracellular space. Some researchers have proposed that calcium flux through voltage-gated calcium channels or calcium-permeable receptors can deplete the calcium in the extracellular compartment, which could reduce calcium influx and neurotransmitter release. Borst and Sakmann suggested that strong bouts of activity could deplete calcium around the calyx of Held synapse, though it is not yet clear that this would take place under physiological conditions, or whether other forms of depression would dominate with such intensive activity. It has proven difficult to measure extracellular calcium depletion experimentally, in part because the domains where it might take place form such a small percentage of the total extracellular space.

Calcium channel inactivation Voltage-gated calcium channels can be inactivated by high levels of internal calcium. Since neurotransmitter release is highly sensitive to the calcium concentration, even a slight reduction in available calcium channels could contribute significantly to depression. Ling-Gang Wu and colleague provided convincing evidence that this can cause depression at the calyx of Held synapse, by measuring presynaptic calcium currents and postsynaptic currents. The form of depression they studied had a very slow onset, taking ~100 ms to reach a peak, which is consistent with the slow onset of inactivation. However, the calyx is normally active at very high rates, in which case depletion and desensitization appear to have somewhat greater influence over depression. Additional study at other synapses will be necessary to determine if it can play a significant role at more slowly firing synapses.

Incomplete vesicle refilling After a vesicle fuses with the membrane and releases its contents, many steps are required before the vesicle can again become readily releasable. In addition to the physical reuptake of the vesicle membrane proteins and translocation to the active zone, it is necessary that the vesicle be reloaded with neurotransmitter. Current estimates suggest that reloading takes several seconds. If vesicles enter the releasable pool prematurely, then they could be incompletely filled. If these vesicles were to release their contents, then fewer postsynaptic receptors would be activated resulting in smaller postsynaptic currents. While this seems a reasonable possibility, there is little evidence for it contributing to depression. In fact, it is unknown whether variability in vesicle content might not be common even with resting synapses, as the variability in size of spontaneous mPSCs is notoriously unaccounted for.

Functional Role of Depression

A natural question that arises is, does depression play a role in the everyday function of the nervous system? It is hard to imagine that it does not, especially as depression can be profound at some highly active synapses. But does depression actually contribute useful computational properties to synapses? Or is it just an undesirable side effect of synaptic transmission? These questions are difficult to answer. Ideally depression could be prevented, and the ensuing changes in nervous system function studied in its absence. But because depression currently appears to be an inevitable consequence of vesicle fusion and neurotransmitter release, blocking it is impossible.

Instead, several workers have used modeling approaches to try to understand the effects of depression. This has led to a number of insights into potential roles of synaptic plasticity. One theme that runs through the proposed functions of depression is that it keeps a highly active presynaptic input from

swamping the postsynaptic cell and dominating its output. This allows smaller signals to emerge, and helps to keep neurons responsive over a wider range of stimuli. These effects have been examined in cortical neurons by Laurence Abbott and colleagues, in auditory neurons responsible for sound localization by William Spain and colleagues, and in *Aplysia* motor neurons by John Byrne and colleagues. In addition, Harunori Ohmori and colleagues have suggested that depression could enhance the specificity of coincidence detection by keeping the postsynaptic potentials small, so only precise coincidences can trigger a postsynaptic response.

These explorations of function represent only the beginning, and many issues remain to be resolved. For example, desensitizing synapses recover much more rapidly than depleted synapses. Does this have functional consequences? Do selection pressures underlie the differences between synapses? Is desensitization selected for because it has different computational properties from depletion, or are these forms of plasticity side effects of other functions? Also, some synapses show profound depression with a single conditioning pulse, whereas others do not, and may even facilitate. What are the mechanisms that determine the level of depression at different synapses, and what are the functional consequences? These are only a few of the many functional issues surrounding short-term plasticity.

See also: AMPA Receptor Cell Biology/Trafficking; BDNF in Synaptic Plasticity and Memory; Cytoskeleton in Plasticity; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; NMDA Receptors, Cell Biology and Trafficking; Plasticity of Intrinsic Excitability; Spike-Timing Dependent Plasticity (STDP); Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Abbott LF and Regehr WG (2004) Synaptic computation. *Nature* 431: 796–803.

Blitz DM, Foster KA, and Regehr WG (2004) Short-term synaptic plasticity: A comparison of two synapses. *Nature Reviews Neuroscience* 5: 630–640.

Byrne JH, Fioravante D, and Antzoulatos EG (2006) Cellular and molecular mechanisms of associative and non-associative learning. In: Selzer M, Clarke S, Cohen LG, Duncan PW, and Gage FH (eds.) *Textbook of Neural Repair and Rehabilitation*, pp. 79–94. Cambridge, UK: Cambridge University Press.

Eccles JC, Katz B, and Kuffler SW (1941) Nature of the 'endplate potential' in curarized muscle. *Journal of Neurophysiology* 4: 362–387.

Feng TP (1941) Studies on the neuromuscular junction. XXVI: The changes in the endplate potential during and after prolonged stimulation. *Chinese Journal of Physiology* 16: 341–372.

Jones MC and Westbrook GL (1996) The impact of receptor desensitization on fast synaptic transmission. *Trends in Neurosciences* 19: 96–101.

Koenig JH, Saito K, and Ikeda K (1983) Reversible control of synaptic transmission in a single gene mutant of *Drosophila melanogaster. Journal of Cell Biology 96: 1517–1522.*

Liley A and North K (1953) An electrical investigation of effects of repetitive stimulation on mammalian neuromuscular junction. *Journal of Neurophysiology* 16: 509–527.

Regehr WG and Stevens CF (2001) Physiology of synaptic transmission and short-term plasticity. In: Cowan WM, Sudhof TC, and Stevens CF (eds.) *Synapses*, pp. 135–175. Baltimore, MD: Johns Hopkins University Press.

Silver RA (2003) Estimation of nonuniform quantal parameters with multiple-probability fluctuation analysis: Theory, application and limitations. *Journal of Neuroscience Methods* 130: 127–141.

Trussell LO, Brenowitz S, and Otis T (1998) Postsynaptic mechanisms underlying synaptic depression. In: Faber DS, Korn H, Redman SJ, Thompson SM, and Altman JS (eds.) *Central Synapses: Quantal Mechanisms and Plasticity*, pp. 149–158. Strasbourg: Human Frontier Science Program.

von Gersdorff H and Borst JG (2002) Short-term plasticity at the calyx of Held. *Nature Reviews Neuroscience* 3: 53–64.

von Gersdorff H and Matthews G (1999) Electrophysiology of synaptic vesicle cycling. *Annual Review of Physiology* 61: 725–752.

Xu-Friedman MA and Regehr WG (2004) Structural contributions to short-term synaptic plasticity. *Physiological Reviews* 84: 69–85.

Xu-Friedman MA and Regehr WG (2005).

Zucker RS and Regehr WG (2002) Short-term synaptic plasticity. *Annual Review of Physiology* 64: 355–405.

Synaptic Mechanisms of Learning

J A Kleim, University of Florida and Malcom Randal VA Hospital, Gainesville, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Most adult behavior involves the performance of skilled movement, and we spend much of our lives expanding our repertoire of motor skills. Motor skills are characterized by the fact that they develop slowly through repetition (practice) and are highly resistant to decay over time. This suggests that skills are progressively encoded as enduring neurophysiological changes within motor areas of the central nervous system. Although numerous areas of brain and spinal cord are certainly involved in skill learning, primary motor cortex (M1) is arguably the key brain structure performing skilled limb movement. The axons of the corticospinal neurons along the bank of the precentral gyrus can be considered the final common path for motor output. Other motor structures, such as basal ganglia and cerebellum, modulate movement indirectly by adjusting the activity of neurons within motor cortex. Further, there are progressive changes in motor cortex activity during motor learning that parallel the development of skilled action patterns. Motor learning also causes enduring changes in cortical organization that can be observed outside the training experience. There are profound changes in the topography of movement representations ('motor maps') within motor cortex such that areas of cortex controlling trained movements become larger and more overlapping. All this suggests that skill learning drives changes in cortical circuitry. To this end, there is now a large body of evidence showing that (1) synaptic plasticity can be induced within motor cortex circuits in response to electrical stimulation, (2) similar synaptic changes occur in association with motor learning, and (3) synaptic plasticity is the mechanism underlying learning-dependent reorganization of cortical movement representations.

Evidence for Synaptic Plasticity in Motor Cortex

Long-Term Potentiation and Long-Term Depression of Cortical Synapses

Much of the evidence for synaptic plasticity within motor cortex has come from studies of long-term potentiation (LTP) and long-term depression (LTD) of various extra- and intracortical pathways. Experiments in animal preparations have shown that synapses within layer II/III can be potentiated via tetanic (high-frequency) stimulation of thalamic and somatosensory afferents. It is interesting that thalamocortical inputs can be potentiated only when activity is paired with stimulation of cortico–cortico afferents. Synapses onto neurons in layer V can be potentiated in response to white matter stimulation, presumably because the stimulated fibers contain both thalamocortical and cortico–cortico afferents. This associative LTP is thought to represent a neural mechanism by which sensory inputs act to facilitate skill learning.

In vitro studies of cortical slice preparations have shown that intracortical synapses can also be modified through stimulation. Tetanic stimulation of horizontal fibers within layer II/III and V induces potentiation of synapses along these fibers up to several millimeters away. Further, stimulation of vertical afferents from layer II/III can enhance synaptic responses in layer V. Thus, local circuits can be modified by coordinating activity of both horizontal and vertical pathways. These forms of synaptic plasticity are also believed to represent the mechanism by which movement representations reorganize during skill learning (see the section titled 'How does synaptic plasticity in motor cortex support motor skill learning?').

The neurochemical and intracellular signals required for LTP in motor cortex have also been investigated using motor cortex slice preparations in which specific pharmacological manipulations can be made. This approach has demonstrated that stimulation-induced LTP in layer II/III requires the presence of γ-aminobutyric acid ($GABA_A$) receptor antagonists. This suggests that a minimal amount of neuronal depolarization is required for LTP to occur. Application of potassium channel blockers or increasing extracellular calcium levels can also facilitate the induction of LTP. Both glutamate and acetylcholine are critical neurotransmitters as both *N*-methyl-D-aspartate and acetylcholinergic receptor antagonists block LTP of horizontal afferents. These same systems may also be engaged during skill learning as similar manipulations also impair skill learning. LTP is also dependent on the activation of specific intracellular signaling pathways, including extracellular signal-regulated kinase and protein kinase A. Blockade of either pathway prevents synaptic potentiation in response to stimulation. These same signaling pathways have been shown to be involved in orchestrating synaptic plasticity in several

brain areas and likely act to coordinate the expression of various genes required for long-term changes in synapse number or morphology.

Although much of what is known about LTP in motor cortex has come from *in vitro* preparations, LTP can also be induced in awake, freely behaving rats. However, unlike the slice preparation, LTP in the whole animal cannot be induced in a single session. Daily stimulation must be given for several days before significant changes in synaptic responses are observed. Further, synaptic potentiation continues to increase each day and does not reach asymptote for 10–14 days. This observation is consistent with the fact that motor skill learning also requires several days of training before significant motor improvements are observed. In addition, the learning-dependent changes in motor maps, believed to be mediated by synaptic plasticity in motor cortex, also require several days to appear (see the section titled 'Learning-dependent synaptic plasticity in motor cortex').

If skill learning occurs only through the strengthening of synapses within motor cortex, then with continued skill learning, all synapses would eventually become saturated. Therefore, synapses within motor cortex must be capable of not only potentiation but also depression. In support of this idea, LTD of synaptic efficacy can also be induced within motor cortex. Low-frequency stimulation of horizontal afferents reduces synaptic responses within layer II/III. The depression can also be reversed through high-frequency stimulation of the same pathways. Thus, cortical circuitry can be altered through bidirectional changes in synaptic strength.

Synaptic Plasticity in Human Motor Cortex

Demonstrating synaptic plasticity in human motor cortex is somewhat more difficult than in animal preparations. However, such plasticity can be inferred by measuring motor-evoked potentials (MEPs) in response to transcranial magnetic stimulation (TMS). TMS involves sending an electrical charge carried on a magnetic pulse through the skull and into the cortex to activate corticospinal neurons. It is a noninvasive method for stimulating cortex that can be done repeatedly in unanesthetized participants. Corticospinal output is then assessed by measuring small changes in MEPs in target muscles. Because TMS drives corticospinal neurons transsynaptically, the changes in MEP amplitude are thought to be due to changes in the strength of synapses onto corticospinal neurons. MEP facilitation and inhibition can be elicited in response to different TMS protocols. Paired pulse stimulation has been used to show that MEP amplitude can be increased or decreased by varying interstimulus interval. Intracortical facilitation is observed with longer (25 ms) interstimulus intervals, whereas intracortical inhibition occurs with shorter (2 ms) interstimulus intervals. Repetitive TMS (rTMS) is a stimulation protocol in which repeated pulses are presented over a short time. This is reminiscent of the tetanic stimulation given in animal models of synaptic plasticity. Low-frequency stimulation reduces MEPs, whereas high-frequency stimulation enhances them. An associative form of synaptic facilitation can also be demonstrated by pairing TMS pulses with peripheral nerve stimulation that activates somatosensory inputs into motor cortex. The pairing of direct TMS stimulation with sensory nerve stimulation increases the amplitude of TMS-evoked MEPs. This phenomenon is reminiscent of the associative LTP observed in animal preparations in which potentiation of layer V and layer II/III synapses can be induced via conjoint stimulation with somatosensory cortex afferents. Finally, the level of cortical excitability prior to stimulation can also influence the capacity and direction of stimulation-induced synaptic plasticity. Preconditioning with transcranial direct current stimulation prior to rTMS can increase or decreases cortical excitability depending on the stimulation parameters used. If the cortex is previously facilitated, rTMS decreases excitability, whereas if the cortex is previously inhibited, rTMS increases excitability. This is further evidence that synaptic strength within motor cortex can be increased or decreased.

Learning-Dependent Synaptic Plasticity in Motor Cortex

If synaptic plasticity such as that observed in the above studies of human and animal motor cortex stimulation does underlie learning, then similar plasticity should be observed in response to motor skill training. There is now convincing evidence from animal and human experiments to support this idea. Training on a skilled reaching task induces an increase in synaptic strength of horizontal afferents in slices of rat motor cortex contralateral to the trained paw but not in slices from ipsilateral cortex. It is interesting that the amount of LTP that can be induced after training is reduced in the trained hemisphere whereas the amount of LTD is enhanced. This is consistent with the notion that skill learning alters cortical connectivity by changing the range of efficacy within which individual synapses can operate and allows for a greater degree of specificity for synaptic changes within cortical circuits.

Chronic studies in awake behaving animals have revealed similar results. Training on a skilled reaching task induces a transient increase in synaptic potential

within layer V of motor cortex in response to white matter stimulation that coincides with an increase in task proficiency during later stages of training. Further, compared with the ipsilateral side, less potentiation but more depression can be induced on the side contralateral to the reaching forelimb. Enhanced synaptic responses of intracortical afferents also appear at approximately the time when animals reach behavioral asymptote, and the potentiation is maintained with continued training. Thus, different pathways within cortical circuitry may undergo changes in synaptic strength at different stages of training.

Motor skill training can also induce synaptic potentiation in human motor cortex. Skill training increases MEP amplitudes in trained muscles of naïve participants, and MEP amplitudes are consistently larger in professional athletes and musicians. For example, volleyball players have larger MEPs in shoulder muscles and racquet ball players have larger MEPs in wrist muscles. Motor learning also occludes stimulation-induced enhancement and augments depression of MEPs. These changes are specific to motor learning and are not observed in participants who experience a similar type of muscular activity without learning. These findings are strikingly similar to the changes in capacity for synaptic potentiation observed in animal experiments of skill learning and provide further evidence for the role of adaptations in the range of synaptic strength as a mechanism of learning in motor cortex.

In addition to physiological measures of synaptic plasticity, changes in synapse number also occur with skill training. Layer V and layer II/III neurons of rats trained to perform a skilled reaching task have more-extensive dendritic arborizations in the motor cortex contralateral to the trained forelimb. The increased dendritic arbor likely reflects an increase in synapse number as there are more synapses per neuron in motor cortex of reach-trained animals than in untrained and activity controls. Further, the synaptogenesis paralleling the acquisition of skill is detectable only after significant improvements in performance. This is consistent with studies of synaptic physiology showing potentiation of synaptic responses only after several days of training. Thus, it is not simply the acquisition of skilled movement that drives synaptogenesis but the continued repetition of the novel movement sequences once acquired.

How Does Synaptic Plasticity in Motor Cortex Support Motor Skill Learning?

Skill Learning Alters Motor Map Organization

Motor learning induces changes in synaptic strength within motor cortical circuits. But how does such plasticity contribute to the encoding of skilled movement? Some insight can be gained from examining how movement is organized within motor cortex and how motor learning alters this organization. First, movement dexterity is reflected in the topography of movement representations within motor cortex (the 'motor map'). Body parts such as the hands, that have a high degree of dexterity, are represented across larger regions of cortex than body parts with low dexterity, such as the leg. Furthermore, motor map organization is not fixed and is capable of reorganization in response to a variety of manipulations, including motor skill training. Studies in rats, monkeys, and humans have all shown that skill training leads to an increase in the proportion of motor cortex containing representations of trained movements. Skilled reach training in rats, for example, causes an increase in the area of cortex from which distal forepaw movements can be elicited. Training on a skilled digit movement task in monkeys causes an increase in the area of digit representations. In addition to causing changes in the representation area, training causes the maps to become more fractured (intermixing of representations) and the number of sites where stimulation evokes multiple movements to increase. Professional athletes and musicians have larger representations of trained muscles that are more overlapping than those of amateurs. Therefore, movement representations become more overlapped and interdigitated with skill training. This is thought to represent the enhanced capacity for producing skilled movement sequences.

Synaptic Plasticity Supports Motor Map Reorganization

It is hypothesized that changes in synaptic strength within intracortical afferents are the mechanism by which movement representations in motor cortex reorganize. Neurons within motor cortex are aggregated such that small groups of cells appear to encode elementary movement representations. Further, these pools of neurons are interconnected by dense horizontal connections that can extend several millimeters. Both intracortical microstimulation and TMS evoke muscle contraction via transsynaptic activation of corticospinal neurons. Thus, any change in motor map topography must therefore involve changes in the pattern of intracortical connectivity.

The development of skilled movement sequences involves strengthening the spatiotemporal relationship between specific muscle synergies. This process may occur through changing the connectivity between specific sets of corticospinal neurons through changes in synaptic efficacy. Movements that are repeatedly produced as part of an acquired sequence would then

become more connected within motor cortex. This would result in a biasing of cortical circuitry such that trained movements are more likely than untrained movements to be evoked in response to cortical stimulation. In support of this hypothesis, learning-dependent synaptogenesis is colocalized with regions of motor cortex that exhibit map reorganization. During skill learning, synaptic potentiation appears at approximately the same time as changes in motor map topography appear. Inducing LTP through tetanic stimulation also induces synaptogenesis and map expansion, whereas LTD causes synapse loss and map shrinkage. Finally, pharmacological manipulations that impair induction of synaptic potentiation also impair map reorganization and motor skill learning.

The key factor in driving learning-dependent synaptic plasticity within cortical circuits is therefore timing. Hebb postulated that synaptic efficacy between two neurons is enhanced when the first repeatedly contributes to the firing of the second. Further, if the second neuron is coincidently active when the synaptic inputs from the first arrive, the connectivity is further strengthened. This form of associative LTP has been demonstrated clearly in motor cortex. Stable reorganization of motor output can be induced by creating an artificial connection between two sites in motor cortex of freely behaving monkeys. That is, the output produced by stimulation of a cortical site can shift toward that of a neighboring site when the two sites are coincidently activated. During skill training, this kind of coincident activity is observed as a progressive increase in the synchrony of neural activity within motor cortex. The coactivation of circuits during practice would lead to plasticity of intracortical synapses resulting in a strengthening of connections between neurons. Cortical representations of trained movements would therefore become larger and more overlapping. Thus, synaptic plasticity provides a neurophysiological mechanism for enhancing dexterity and creating the muscle synergies required for performing skilled movement.

See also: Computational Approaches to Motor Control; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Functional Imaging of the Motor System; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP); Motor Skill Learning; Motor Sequences.

Further Reading

Adkins DL, Boychuk J, Remple MS, and Kleim JA (2006) Motor training induces experience-specific patterns of plasticity across motor cortex and spinal cord. *Journal of Applied Physiology* 101: 1776–1782.

Hess G (2003) Synaptic plasticity of local connections in rat motor cortex. *Acta Neurobiologicia Experimentalis* 63: 271–276.

Hess G and Donoghue JP (1994) Long-term potentiation of horizontal connections provides a mechanism to reorganize cortical motor maps. *Journal of Neurophysiology* 71: 2543–2547.

Hess G and Donoghue JP (1996) Long-term depression of horizontal connections in rat motor cortex. *European Journal of Neuroscience* 8: 658–665.

Hodgson RA, Ji Z, Standish S, et al. (2005) Training-induced and electrically induced potentiation in the neocortex. *Neurobiology of Learning and Memory* 83: 22–32.

Iriki A, Pavlides C, Keller A, and Asanuma H (1991) Long-term potentiation of thalamic input to the motor cortex induced by coactivation of thalamocortical and corticocortical afferents. *Journal of Neurophysiology* 65: 1435–1441.

Jackson A, Mavoori J, and Fetz EE (2006) Long-term motor cortex plasticity induced by an electronic neural implant. *Nature* 444: 56–60.

Kleim JA, Barbay S, Cooper NR, et al. (2002) Motor learning-dependent synaptogenesis is localized to functionally reorganized motor cortex. *Neurobiology of Learning and Memory* 77: 63–77.

Kleim JA, Hogg TM, VandenBerg PM, Cooper NR, Bruneau R, and Remple M (2004) Cortical synaptogenesis and motor map reorganization occur during late, but not early, phase of motor skill learning. *Journal of Neuroscience* 24: 628–633.

Ljubisavljevic M (2006) Transcranial magnetic stimulation and the motor learning-associated cortical plasticity. *Experimental Brain Research* 173: 215–222.

Monfils MH and Teskey GC (2004) Skilled-learning-induced potentiation in rat sensorimotor cortex: A transient form of behavioural long-term potentiation. *Neuroscience* 125: 329–336.

Rioult-Pedotti MS, Friedman D, and Donoghue JP (2000) Learning-induced LTP in neocortex. *Science* 290: 533–536.

Rioult-Pedotti MS, Friedman D, Hess G, and Donoghue JP (1998) Strengthening of horizontal cortical connections following skill learning. *Nature Neuroscience* 1: 230–234.

Schieber MH (2002) Training and synchrony in the motor system. *Journal of Neuroscience* 22: 5277–5281.

Ziemann U, Ilic TV, and Jung P (2006) Long-term potentiation (LTP)-like plasticity and learning in human motor cortex: Investigations with transcranial magnetic stimulation (TMS). *Supplements to Clinical Neurophysiology* 59: 19–25.

Synaptic Plasticity and Place Cell Formation

J J Knierim, University of Texas Medical School, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In the mid-1950s, an epileptic patient with the initials HM underwent a neurosurgical operation to remove his hippocampus and surrounding parts of his medial temporal lobe in order to excise the source of his life-threatening seizures. As an unexpected consequence of this procedure, HM lost the ability to form new, declarative (conscious) memories. His general intellect and personality were unaffected, and his memories of his past were generally intact. However, from that day onward, HM has not been able to create any new, long-lasting memories about the events that happen to him. He can retain information for brief periods, but as soon as he is distracted or a few minutes go by, he loses all memory of what just occurred. To this day, he has little, if any, memory of his life history after the operation. This type of amnesia has since been studied in numerous patients that have suffered hippocampal and medial temporal lobe damage as the result of stroke, infection, brain trauma, or neurodegenerative disease (e.g., Alzheimer's disease).

From the initial case report of HM's condition by William Scoville and Brenda Milner, the hippocampus has become the premier site to investigate the mechanisms of learning and memory in the mammalian brain. The decade after Scoville and Milner's report witnessed a number of laboratories engaged in attempts to discover physiological correlates of learning and memory in the hippocampus. The early 1970s produced two seminal discoveries about hippocampal physiology that still dominate this research three decades later. In 1971, John O'Keefe and Jonathan Dostrovsky reported that, in freely moving rats, hippocampal pyramidal neurons fire selectively in specific locations in an environment (Figure 1). Subsequent research on these cells and a thorough review of the experimental literature on hippocampal lesions led O'Keefe and Lynn Nadel in 1978 to propose a theory that these hippocampal 'place cells' were the building blocks of an internal, cognitive map of the environment. This map was necessary not only for flexible, spatial navigation, but also as a framework to organize the items and events of an organism's experience into a long-term memory store.

In 1973, another landmark paper by Timothy Bliss and Terje Lømo presented compelling evidence that strong, electrical stimulation of hippocampal input fibers can cause a long-lasting potentiation of the synaptic efficacy of these presynaptic fibers onto their hippocampal targets. This phenomenon, called long-term potentiation (LTP), has become the predominant model for how neural activity can cause changes in the strength of synapses between neurons, thus serving as a cellular mechanism for information storage (i.e., learning and memory). The initial discovery of this phenomenon in the hippocampus was thought to perhaps reflect the special role of the hippocampus in learning and memory. It is now known that there are multiple forms of learning in different brain systems, and that LTP (and its counterpart, long-term depression, or LTD) can be elicited in many different brain structures (including the neocortex and cerebellum). LTP and LTD are fundamental physiological mechanisms that fine-tune the synaptic connectivity of the brain as it wires itself during development and as it adapts to the constantly changing circumstances that an organism experiences throughout its lifetime.

In the first couple of decades after the appearance of these two papers, surprisingly little published research directly investigated the relationship between LTP and place cells. Place cell research focused on the questions of what external and internally generated cues generated the spatial specificity of these neurons and whether hippocampal neurons were sensitive to nonspatial variables as well as spatial location. LTP research focused on establishing a direct link between LTP and learning in behavioral studies and on investigating the cellular and molecular mechanisms of LTP induction and maintenance. In recent years, a number of investigations have begun to focus explicitly on the role of synaptic plasticity in place field formation and function.

Experience-Dependent Properties of Place Cells

Given the strong link between the hippocampus and LTP, it is perhaps surprising that much of the spatially selective firing of place cells does not appear to be learned. Rather, place cells can demonstrate their spatially selective place fields from the very first time an animal enters the place field. These fields are not static, however, which suggests that some form of plasticity occurs in these networks. One demonstration of this plasticity comes from the application of multielectrode recordings of dozens of CA1 place cells, which allows the offline reconstruction of the animal's trajectories through an environment by

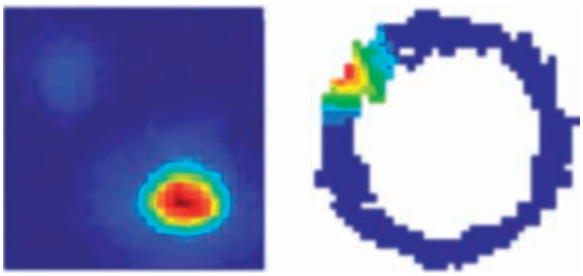

Figure 1 Examples of place fields from the rat hippocampus. Each figure is a spatial firing rate map for a CA1 place cell, in which red indicates a location in which the cell fired at a high rate and blue indicates a region where the cell was silent. The example from the left is from a rat foraging for food on a square platform. The cell fires selectively in the middle of the southeast quadrant of the platform. The example on the right is from a rat running clockwise on a circular track. The cell fires selectively at the northwest section of the track.

analyzing the population activity of the recorded cells (i.e., by knowing the subset of active place cells at a given moment, one can accurately determine the location of the animal at that moment). In a novel environment, the accuracy of this reconstruction increases over the first 10 min of exploration. Thus, some fine-tuning of the place fields must occur, such as an increase in the specificity or reliability of individual firing fields. Although the place field 'map' is evident immediately in a novel environment, the first few minutes of exploration cause changes to the cells that improve the spatial signal at the population level.

The place fields appear to become bound to external sensory landmarks through a learning process. In a familiar, stable environment, the firing locations of place cells can be controlled by external landmarks, such as visual cues at the periphery of an environment, in that rotation of these salient landmarks can cause an equal rotation of the place cell map. However, if an animal is disoriented each time it enters an environment, such that it perceives that the landmarks in that environment are unstable relative to its own, internal direction sense, the ability of the landmarks to control the location of the place fields becomes weakened. Similarly, if the landmarks are moved around in the presence of the animal, they can lose the ability to control the place fields. Individual place cells maintain their firing locations relative to each other, however. Under these conditions, and under conditions of darkness or other external sensory deprivation, the cells appear to be under the dominant control of path integration mechanisms, presumably provided by the input from the medial entorhinal cortex. This input is probably always at work as a major drive onto place cells, and it endows the cells with their immediate spatial specificity in a novel environment. As the animal explores the new environment, the place cells are thought to form associations with other inputs that represent the landmarks in that environment through an LTP-dependent mechanism. This learned association eventually allows the landmarks to control the place fields, and can be disrupted when the animal perceives that the landmarks are unstable directional cues. Furthermore, these associations allow the hippocampus to activate the proper subset of place cells when the animal reenters that environment in the future.

Changes to the environment can trigger complex alterations in the ensemble activity of place cells. Cells that were previously silent can form new place fields when new objects are placed in the environment or when expected rewards are omitted. In many cases, changes in the environment can produce wholesale changes in the populations of active place cells. This phenomenon, called 'remapping' by Robert Muller and John Kubie, can occur immediately when the change is first made; can occur abruptly some time after the change is made; or can develop slowly and continuously over many experiences in the altered environment. The phenomenon of place cell remapping is not entirely understood, and it certainly encompasses a variety of different neural mechanisms. As such, it is a prime candidate to address the relationship between place cell function and synaptic plasticity.

Place field plasticity has also been correlated with the temporal dynamics of LTP induction. As first demonstrated by William Levy and Oswald Steward, LTP is temporally asymmetric. That is, presynaptic cells must fire before postsynaptic cells in order to produce LTP. If the postsynaptic cell fires first, then LTD is produced. This spike-timing relationship preserves the causal relationship of firing. If the presynaptic cell fires shortly before the firing of the postsynaptic cell, then presumably the presynaptic cell had an impact on causing the postsynaptic cell to fire and, according to Hebb's rule, should increase its strength of connection. If the timing is reversed, however, then the firing of the presynaptic cell could not have caused the firing of the postsynaptic cell, and therefore the connection strength should be decreased. Evidence that such timing rules might be useful for encoding spatiotemporal sequences of activity was found in place cells. When rats run stereotyped routes, the place fields expand backward (i.e, in the direction opposite to the direction of travel of the rat) with experience, as predicted by computational models that applied spike-timing-dependent plasticity rules toward place cells. As illustrated in Figure 2, as

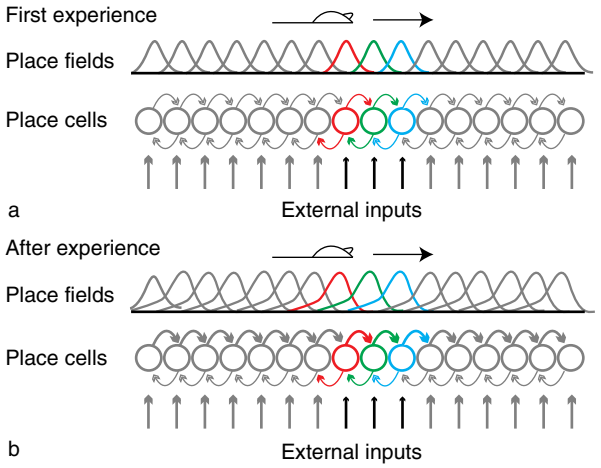

Figure 2 Backward expansion of place fields. Originally discovered by Mehta and colleagues, this NMDA receptor dependent phenomenon suggests a relationship between temporally asymmetric induction of LTP and the possible encoding of spatial sequences in the hippocampus. (a) The first time a rat runs on a linear track, the linear place fields are symmetric and are presumably driven by external spatial inputs (landmarks, self-motion cues, etc.), indicated by thick arrows. The place cells are also interconnected weakly (thin, curved arrows). The rat is in the middle of the track, where the red place cell is active, and the green and blue place cells are about to be activated as the rat moves to the right. (b) After a number of laps where the rat runs from left to right, LTP has strengthened the connections between pairs of cells that fired in temporal order (e.g., the connection of the green place cell onto the blue cell was strengthened, because the green cell always fired before the blue cell, but its connection onto the red cell was not strengthened, because it always fired after the red cell). As a result, each cell now gets strong external input as well as strong input from its preceding neighbor. This added input causes the cell to fire earlier in the sequence than it did in (a).

the rat runs from left to right, the red place cells always fire before the green place cells, which fire before the blue place cells. As a result, the red place cell over time increases its connection strength onto the green place cell, causing the green place cell to fire earlier than before (i.e., the green place field expands backward). The same expansion happens for all of the place fields on the track. After experiencing these stereotyped sequences of locations, the earliest spikes of each place field may serve as a prediction of where the animal will be if it continues on its learned path, rather than a representation of the animal's current location. That is, in Figure 2(b), the late spikes from the red cell indicate the current location of the animal, whereas the early spikes from the green cell indicate that the animal will arrive at the green location next, which will then predict the blue location, and so on. This backward expansion is apparent in both CA1 and CA3 place fields. CA3 appears to be the location where the sequences are initially learned (in as little as one trial) and stored for the long term. In CA1, the sequences are stored only for a few hours, suggesting a role for CA1 in the online encoding of sequences and comparison with the long-term stored sequences in CA3.

Place Cells in Aged Rats

Some of the earliest evidence linking LTP to place cells came from studies of aged rats. Old rats have well-documented deficits in LTP and in spatial learning. However, the specificity of individual place fields is only modestly affected in aged rats, which is consistent with the immediate presence of place fields in a new environment and implies that place fields themselves do not always require synaptic plasticity to form. A number of secondary properties of place fields that are thought to depend on synaptic plasticity are disrupted in aged rats. Aged rats appear to have deficits in activating the proper subset of place cells when entering a familiar environment. In parallel, they have deficits in changing stable representations when faced with environmental alterations. In other words, place cells of aged rats or more likely than young rats to inappropriately remap a stable environment, even though no salient changes have been made to the environment. Whereas young rats will display the same place fields in a familiar environment day after day, aged rats are more likely to display different place fields on different visits to the same environment. The binding of place fields to individual landmarks allows the young rat to recognize a familiar environment and activate the proper set of place cells. This binding is a learning process that appears to be impaired in aged rats.

Conversely, place cells of aged rats are less adaptable to environmental changes or to behavioral task changes than are place cells of young rats. As mentioned before, place cells of young rats can remap an environment in response to salient changes, as the hippocampus creates new, context-specific representations of different environments or even the same environment under different circumstances. The place cells of aged rats are more rigid and inflexible, however, as they are less likely to remap under these conditions. This rigidity may be a reflection of deficits in synaptic plasticity mechanisms.

Place cells of aged rats also fail to show the backward expansion/shift effect when the rats run a stereotyped route. Taken together, these results are consistent with the notion that deficits in LTP in aged rats impair the ability of place cells to encode relevant information about changes to the environment and also impair the rats' ability to retrieve the proper hippocampal representations for familiar environments. If this retrieval failure subsequently prevents retrieval of the nonspatial information that has become associated with the

hippocampal map, this result may explain many of the learning impairments that are seen in senescent animals (including humans).

Place Cells and LTP

The work on place cells and aged rats provides striking evidence of a relationship between synaptic plasticity and high-order cognitive representations. A major advantage of this work is that it capitalizes on a natural experimental manipulation. That is, the well-characterized deficits in LTP that occur naturally in the aging process can be correlated with changes in place cell properties without the need for experimental manipulations. A disadvantage, however, is that the results are by nature correlational and indirect, and one would also like to study the question under conditions that are more explicitly under experimental control. These experiments can be performed with the use of pharmacological agents that block LTP and genetic techniques that can alter specific molecular aspects of the LTP induction, expression, and maintenance pathways.

LTP blockers have been used to study the remapping phenomenon. Rats injected with pharmacological blockers of LTP can remap an altered environment similar to controls. That is, when the environment is altered, new place fields emerge and old place fields become silent or scramble. This result demonstrates once again that the expression of place cell spatial selectivity does not depend on LTP mechanisms. However, in contrast to control rats, rats with LTP blockers are not able to maintain a stable representation of the altered environment when they are reintroduced into it a day later. It appears that LTP is necessary for the sensory landmarks that allow the rat to recognize the new environment to become associated with the new place field map and reactivate it when the rat is placed back into the environment. Without this binding in the initial visit to the environment, the hippocampus treats the second visit to the environment as if it were a completely novel environment and activates a subset of place cells that is different from the subset active in the first exposure (Figure 3). This result demonstrates explicitly that LTP is necessary for the binding of a particular subset of place cells to represent a particular environment, and supports the argument that LTP deficits in aged rats cause the improper 'remapping' of familiar environments due to a deficit in this binding.

LTP blockers have also been shown to block the backward expansion of place fields, providing direct evidence that this phenomenon is dependent on the proper functioning of *N*-methyl-D-aspartate (NMDA) receptors (the type of glutamate receptor that is essential for LTP induction). This result lends credence to Hebb's ideas that synaptic plasticity can generate phase sequences: cell assemblies that are serially linked together in a learned temporal order. Such phase sequences endow the brain with two important properties: predictability and temporal recall. That is, if synaptic strengthening causes one cell assembly (e.g., a cell assembly representing location A) to activate the cell assembly that tends to follow it (e.g., a cell assembly representing the next location on a familiar route), which activates the next assemblies in order, this sequential activation becomes a prediction of what events are likely to follow given the current state of the animal. This predictability can then be used to alter behavior adaptively (i.e., if the end of the phase sequence predicts a negative consequence, alter one's behavior; if the end predicts a positive consequence, maintain the behavior). Second, storing these sequences allows the cell assemblies to be reactivated in order, thus allowing the recall of the temporal sequence of events, which is a hallmark of episodic memory that is thought to rely critically on hippocampal processing. Evidence for these sequences arises from studies of rats during sleep, in which the temporal order of place cell firing that occurs during the behavior of the animal is preserved when these cells

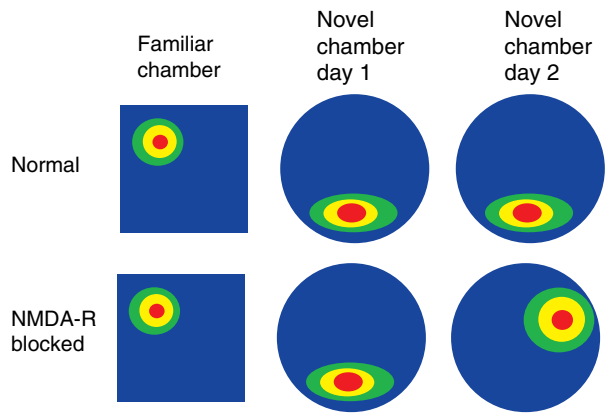

Figure 3 NMDA receptor antagonists prevent the binding of place fields to the cues that define a particular environment. This figure illustrates schematically the demonstration by Kentros and colleagues that, under normal conditions, a place cell will show different place fields in a familiar environment (the square) and a novel environment (the circle). When reintroduced into the novel chamber the next day, the place cell 'remembers' where it should fire in the new environment, presumably due to an associative binding between the neural representations of landmarks in that environment and the place cell. When NMDA receptors are blocked, however, this binding does not take place. The initial formation of a new place field is unaffected in the novel chamber on the first day. However, the next day, the place field fires in yet another location. Apparently, the blockage of LTP the day before prevented the landmarks in the novel chamber from activating the cell in the proper location a day later. The cell has a place field, but the population representation is substantially different from the first exposure.

fire during subsequent sleep periods. This temporal reactivation has been proposed to reflect an offline process by which the hippocampus 'replays' to the neocortex the activation patterns it experienced in the recent hours, thus allowing the slower plasticity mechanisms of the neocortex to gracefully integrate the hippocampus-mediated new learning into the long-term memory stores of the neocortex.

The emergence of site-specific, inducible genetic knockouts and transgenic animals has opened a new avenue of investigation of the relationship between synaptic plasticity and place cells. Different knockout lines have targeted NMDA receptor subunits in the CA1 and other forebrain regions and in the CA3 region. Other genetically altered mouse lines have eliminated or mutated other molecules that are necessary for normal LTP, such as calmodulin (CaM) kinase II and cAMP response element binding (CREB). In general, the mutant mice tend to have place fields that are unstable or less specific than wild type, although they show clear spatial specificity. These results support the findings that place cells do not require LTP to show spatial selectivity, but that the selectivity is fine-tuned by synaptic plasticity. Particularly interesting is the CA3 NMDA receptor knockout, which inhibits one-trial learning in a match-to-place version of the Morris water maze and reduces the selectivity of downstream place fields in CA1 during first exposure to a novel environment. In subsequent exposures, the animals are able to perform the task and the CA1 place fields are normal. These results suggest that CA3 plays a critical role for rapid, one-trial learning, and that CA1 is able to support learning only at a slower time course, consistent with other pharmacological investigations of site-specific NMDA receptor blockade on the relative contributions of CA3 and CA1 plasticity in learning and with differences between CA3 and CA1 place fields in spatial sequence encoding when the animal experiences an altered environment.

Associative Memory: Theory and Experiment

An autoassociative network is one in which the output patterns of activity can be associated with themselves through feedback pathways. The recurrent collateral pathway of CA3 has for a long time suggested to theoreticians that CA3 may form an autoassociative network for memory storage. An important property of such a network is called 'pattern completion,' whereby a degraded or incomplete input pattern can reactivate the entire, originally stored input pattern via the recurrent collaterals. Such an organization can also be used to generalize input patterns, as similar input patterns can cause the same output pattern to be activated in the network. In some cases, however, if similar input patterns represent distinct states that must be kept segregated, a memory storage system should be capable of 'pattern separation,' whereby similar input patterns are made more dissimilar before being stored in the synaptic weights of the network. Pattern separation allows the network to store more memories before they begin to interfere with each other.

Place cell recordings have shown evidence for both pattern separation and pattern completion in the CA3 region. In agreement with computational models that suggest a sigmoidal relationship between input pattern similarity and output pattern similarity, when changes are made to local cues in a familiar environment, the CA3 region outputs a representation that is similar to the original representation. In contrast, the CA1 region outputs a representation that is less similar to the original representation than CA3. When the animal is brought into an entirely different environment, the CA3 region outputs a completely new representation. In contrast, the CA1 region outputs a representation that maintains a degree of similarity with the representation of the first environment. Thus, CA3 appears to reflect a pattern completion process when inputs are modified to a limited degree, whereas it reflects a pattern separation process when the inputs are altered more extensively.

Further evidence for the role of CA3 in pattern completion comes from genetic studies. CA3 mice with NMDA receptor knockout show deficits in pattern completion. Place cells recorded from CA1 in these mice had larger place fields compared to wild-type mice when a number of the salient spatial cues in an environment were removed, suggesting that the mutant mice had a reduction in the ability to pattern complete the partial input. These mice also showed deficits in LTP in CA3 and in spatial learning under the partial cue conditions.

Summary

Place cells of the hippocampus are a prime example of a robust, high-order, cognitive representation. Because of the well-documented role of the hippocampus in learning and memory and the extensive knowledge of LTP in the hippocampus, place cells are becoming an important model system for neural plasticity as well. These studies rely heavily on the use of multi-electrode recording technologies, which can provide critical data on the changes that occur in many neurons in a small period of time. Because the hippocampus is thought by many to be particularly important for one-trial, episodic memory, these ensemble recordings will become increasingly essential toward understanding the role of synaptic

plasticity in hippocampal function, and the role of hippocampal function in the neuropsychology of memory.

See also: Hippocampus and Neural Representations; Hippocampus; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation (LTP): NMDA Receptor Role; Spatial Cognition and Executive Function; Spatial Cognitive Maps; Spike-Timing Dependent Plasticity (STDP); Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging; Synaptic Depression.

Further Reading

Barnes CA (2003) Long-term potentiation and the ageing brain. *Philosophical Transactions of the Royal Society of London, Series B. Biological Sciences* 358: 765–772.

Best PJ, White AM, and Minai A (2001) Spatial processing in the brain: The activity of hippocampal place cells. *Annual Review of Neuroscience* 24: 459–486.

Bliss TV and Lømo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* 232: 331–356.

Ekstrom AD, Meltzer J, McNaughton BL, and Barnes CA (2001) NMDA receptor antagonism blocks experience-dependent expansion of hippocampal 'place fields.' *Neuron* 31: 631–638.

Guzowski JF, Knierim JJ, and Moser EI (2004) Ensemble dynamics of hippocampal regions CA3 and CA1. *Neuron* 44: 581–584.

Hebb DO (1949) *The Organization of Behavior.* New York: Wiley.

Kentros C, Hargreaves E, Hawkins RD, Kandel ER, Shapiro M, and Muller RU (1998) Abolition of long-term stability of new hippocampal place cell maps by NMDA receptor blockade. *Science* 280: 2121–2126.

McClelland JL, McNaughton BL, and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102: 419–457.

Mehta MR, Barnes CA, and McNaughton BL (1997) Experience-dependent, asymmetric expansion of hippocampal place fields. *Proceedings of the National Academy of Sciences of the United States of America* 94: 8918–8921.

Nakazawa K, McHugh TJ, Wilson MA, and Tonegawa S (2004) NMDA receptors, place cells and hippocampal spatial memory. *Nature Reviews Neuroscience* 5: 361–372.

O'Keefe J and Nadel L (1978) *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

Rapp PR (1998) Representational organization in the aged hippocampus. *Hippocampus* 8: 432–435.

Redish AD, McNaughton BL, and Barnes CA (1998) Reconciling Barnes et al. (1997) and Tanila et al. (1997a, b). *Hippocampus* 8: 438–443.

Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry* 20: 11–21.

Wilson MA and McNaughton BL (1993) Dynamics of the hippocampal ensemble code for space. *Science* 261: 1055–1058.

Synaptic Plasticity: Cerebellum

E Hosy, A Belmeguenai, and C Hansel, Erasmus University Medical Center, Rotterdam, The Netherlands

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In neural networks activity-dependent synaptic plasticity can be long lasting and bidirectional. Synapses can be potentiated (long-term potentiation (LTP)) or depressed (long-term depression (LTD)), enabling the storage (and erasure) of information at selected sets of synapses. LTP and LTD provide the most convincing cellular candidate mechanisms for learning and memory ('synaptic memory'), which explains the high level of attention that these synaptic alteration processes have received during the past three decades. In the cerebellum, LTD at parallel fiber (PF) to Purkinje cell synapses is believed to provide a cellular basis for forms of cerebellar motor learning, such as associative eyeblink conditioning or gain adaptation of the vestibulo-ocular reflex. More recently described forms of LTP might, in turn, provide reversal mechanisms. This article describes cerebellar LTD and LTP from a cellular/molecular perspective and points out similarities and differences to their counterparts at hippocampal synapses.

LTD at Parallel Fiber–Purkinje Cell Synapses

Purkinje cells form inhibitory synapses with their target cells in the deep cerebellar nuclei (or the vestibular nuclei). Therefore, theories of cerebellar function have described PF-LTD, rather than PF-LTP, as a central component of motor learning because LTD at excitatory synaptic inputs to Purkinje cells was expected to cause a disinhibition of the target neurons in the cerebellar or vestibular nuclei.

PF-LTD can be monitored as a downregulation of PF-evoked excitatory postsynaptic currents (EPSCs). In the example shown in Figure 1(a), LTD was measured using somatic whole-cell patch-clamp recordings from Purkinje cells in rat (postnatal days 18–28) cerebellar slices. Typically, in these experiments spontaneous spike firing is prevented by the injection of small hyperpolarizing currents and fast inhibitory transmission is blocked by bath application of $GABA_A$ receptor antagonists. Both measures allow us to characterize excitatory synaptic events in isolation.

Next to the synapses provided by PFs, Purkinje cells receive a second type of excitatory synaptic input, provided by a single climbing fiber (CF), which originates from the inferior olive. PF-LTD induction depends on paired PF and CF activation (Figure 1(a)). LTD is optimally induced when PF activity slightly precedes the activation of the CF input (by 50–200 ms), which corresponds well to the timing requirements for the application of conditioned and unconditioned stimuli in associative eyeblink conditioning.

CF activity results in an all-or-none complex spike in Purkinje cells, which is associated with a widespread calcium transient in Purkinje cell dendrites. This heterosynaptically evoked calcium transient is required for PF-LTD induction, reflecting the involvement of a threshold level of postsynaptic $[Ca^{2+}]_i$. At the PF synapses, both α-amino-3-hydroxy-5-methyl-4-isoxazoleproprionic acid (AMPA) receptors and type 1 metabotropic glutamate (mGluR1) receptors need to be activated for LTD induction. Glutamate binding to mGluR1 receptors results in the activation of phospholipase C and subsequently in the production of inositol-1,4,5-trisphosphate (IP_3) and 1,2-diacylglycerol (DAG). IP_3 signaling might contribute to the total calcium transient by releasing calcium from internal stores. The participation of this signaling pathway, however, has been controversial. In contrast, the second product of phospholipase C activity, DAG, could be unequivocally linked to LTD induction. DAG and calcium (provided by voltage-gated calcium channels opened by AMPA receptor-mediated depolarization) synergistically activate protein kinase C (PKC). The critical involvement of PKC in LTD induction has been demonstrated using pharmacological approaches as well as transgenic mice (L7-PKCI), in which a PKC inhibitory peptide is specifically expressed in Purkinje cells. It has also been shown that LTD is impaired in cerebellar slices obtained from *αCaMKII−/−* mice, suggesting that not only PKC but also αCaMKII is involved in LTD induction. Finally, it has been suggested that the gaseous second messenger nitric oxide (NO) is involved in LTD induction by triggering an NO/cGMP/PKG cascade.

How does activation of various kinases mediate LTD at PF synapses? The chain of events leading from PKC activation to LTD has been characterized in detail. In Purkinje cells, the majority of AMPA receptors are constituted by GluR2–GluR3 heteromeres. PF-LTD results from a clathrin-mediated endocytosis of GluR2, which requires PKCα-dependent phosphorylation of GluR2 at serine 880. The internalization of GluR2 involves unbinding from glutamate receptor-interacting protein (GRIP) and binding to protein interacting with C-kinase 1 (PICK1). In contrast, the role of αCaMKII in GluR2 endocytosis

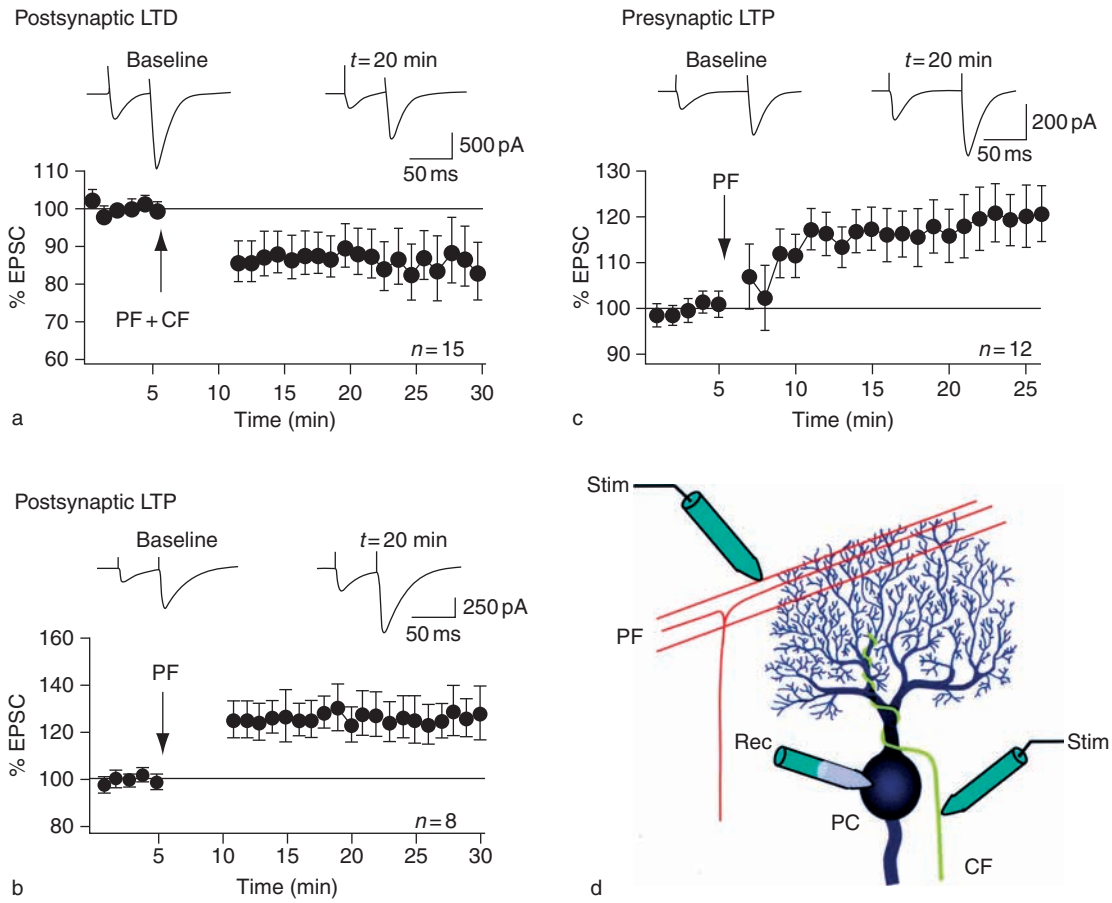

Figure 1 Three forms of long-term plasticity at cerebellar parallel fiber synapses. (a) Postsynaptic LTD is induced by paired parallel fiber (PF) + climbing fiber (CF) activation at 1 Hz for 5 min ($n = 15$). The traces show EPSC pairs (50 ms delay) before and after LTD induction. PF-LTD is not associated with changes in the PPF ratio. (b) Postsynaptic LTP is induced by isolated PF stimulation at 1 Hz for 5 min ($n = 8$). Traces show EPSC pairs (50 ms delay) before and after LTP induction. This form of LTP is not associated with changes in the PPF ratio. (c) Presynaptic LTP is induced by isolated PF activation at 8 Hz for 15 s ($n = 12$). Traces show EPSC pairs (100 ms delay) before and after LTP induction. The PPF ratio is reduced after tetanization. (d) Diagram showing the electrode arrangement used for LTP and LTD experiments *in vitro*. Whole-cell patch-clamp recording is used to monitor electrical responses to PF and/or CF stimulation. For extracellular stimulation, glass pipettes are used (Stim) that are filled with ACSF.

and LTD remains to be examined. PKG likely has an indirect role. It seems possible that phosphorylation of G substrate by PKG can cause an inhibition of protein phosphatases that might dephosphorylate the same substrates that are phosphorylated by PKC or αCaMKII.

PF-LTD is mediated by an endocytosis of AMPA receptor subunits and, consequently, in a reduction of the postsynaptic AMPA receptor density. The postsynaptic expression side is confirmed by an absence of changes in the paired-pulse facilitation (PPF) ratio.

LTP at Parallel Fiber–Purkinje Cell Synapses

If PF-LTD is a correlate of forms of cerebellar motor learning, it can be predicted that at least one type of LTP exists to enable reversibility. In fact, both pre- and postsynaptically expressed forms of PF-LTP have been described. Postsynaptic LTP can be induced by the same low-frequency PF activation protocol used for LTD induction (1 Hz, 5 min), when the CF input is not coactivated (Figure 1(b)). LTP requires a dendritic calcium transient for induction, but the calcium threshold for LTP induction is lower than that for LTD induction (hence, CF activity is needed for LTD induction). This type of LTP is postsynaptically expressed, as indicated by the absence of changes in the PPF ratio. Moreover, the low-affinity competitive AMPA receptor antagonist γ-D-glutamylglycine, which can be used as a reporter for changes in glutamate release, reduces EPSC amplitudes to the same level after LTP induction as it does before, suggesting that transmitter release does not change after tetanization.

Unlike LTP at most other types of synapses, postsynaptic LTP at PF synapses does not require kinase activation. PF-LTP is insensitive to bath application

of the PKC inhibitor chelerythrine or the general kinase inhibitor H-7. Moreover, LTP is intact in slices prepared from *αCaMKII*−/− mice. In contrast, LTP depends on the activation of protein phosphatases 1, 2A, and 2B, as suggested by the blockade of LTP in the presence of the phosphatase inhibitors okadaic acid (PP1/2A), microcystin LR (PP1/2A), inhibitor-2 (PP1), fostriecin (PP2A), and cyclosporin A (PP2B).

The phosphatase involvement in synaptic potentiation suggests that LTP and the kinase-dependent LTD can indeed reverse each other. Electrophysiological saturation experiments confirm this hypothesis. When LTP is saturated first by application of two subsequent LTP protocols, application of the LTD protocol leads to a depotentiation. This synaptic depression is truly a 'depotentiation' (rather than just a masking effect) because the subsequent application of a third LTP protocol allows for a potentiation of the previously depressed EPSC amplitudes. Such a re-potentiation would not be possible when the LTP mechanism would still be saturated. As discussed previously, LTD is mediated by GluR2 endocytosis. LTD reversal might thus involve a mechanism enabling GluR2 membrane insertion. Indeed, it has been shown that PF-LTP is associated with GluR2 insertion following NO-mediated *N*-ethylmaleimide-sensitive factor binding to GluR2.

PF stimulation alone can also elicit a presynaptic form of LTP. For induction, PF stimuli are applied at higher frequencies and for shorter durations (e.g., 8 Hz, 15 s) compared to the protocols used for induction of postsynaptically expressed LTP. The presynaptic expression side of this second form of LTP is evident from a reduction in the PPF ratio (Figure 1(c)). Moreover, injection of the calcium chelator BAPTA at a concentration high enough to block all postsynaptic forms of plasticity (30 mM) does not affect LTP resulting from 8 Hz, 15 s PF stimulation. The latter finding suggests the involvement of a presynaptic induction cascade. Indeed, presynaptic LTP depends on the activation of calcium/calmodulin-sensitive adenylyl cyclase I and the subsequent activation of protein kinase A (PKA). A PKA substrate involved in LTP induction is the active zone protein RIM1α, whose phosphorylation by PKA promotes a long-lasting increase in transmitter release.

LTD and LTP at Parallel Fiber Synapses under Heterosynaptic Climbing Fiber Control

Adult Purkinje cells receive synaptic input from only one CF, which is the result of an activity-dependent elimination process that is completed at approximately postnatal day 18. CF activation evokes in Purkinje cells an all-or-none response, the so-called 'complex spike,' which is associated with a widespread calcium transient in Purkinje cell dendrites. This heterosynaptically evoked calcium surge is crucial for PF synaptic plasticity. Studies have revealed that the CF input, through complex spike-associated calcium transients, in fact controls all forms of long-term plasticity at PF synapses that have been characterized so far. Three ways of interference have been observed:

1. Prolonged PF tetanization at low frequencies can induce either LTD or LTP (postsynaptically), depending on whether or not the CF was coactivated. Paired PF and CF activation results in LTD induction, whereas LTP can be observed after PF stimulation alone. The CF controls the direction of synaptic gain change through the calcium signal that it adds on top of the calcium transient locally elicited at the PF input sites. Indeed, a higher calcium threshold level needs to be reached for LTD induction than for LTP induction (Figure 2).
2. Synaptic plasticity is not only a feature of PF synapses but also a feature of the CF input. Brief isolated CF tetanization at slightly elevated frequencies (5 Hz, 30 s) results in CF-LTD, which actually is a multicomponent phenomenon: a depression can be seen at the level of CF-EPSCs (voltage-clamp mode), the complex spikes (current-clamp mode), complex spike-associated calcium transients, as well as the afterhyperpolarization following complex spikes. The long-term reduction in calcium transients that accompanies CF-LTD has consequences for PF plasticity. When the PF-LTD protocol is applied after previous CF-LTD induction, PF-LTP is induced instead. The most likely explanation is that the overall reduction in the calcium transient shifted the relative probabilities for PF-LTD and PF-LTP induction. Thus, this phenomenon describes a new form of heterosynaptic metaplasticity.
3. CF activity, while postsynaptically promoting PF-LTD, blocks presynaptic LTP induction, thus providing a 'safety-lock' mechanism that prevents a postsynaptic downregulation of AMPA receptor function (postsynaptic LTD) is counterbalanced by an upregulation of transmitter release (presynaptic LTP). The CF blocks presynaptic LTP by promoting the release of endocannabinoids from Purkinje cell dendrites. The endocannabinoids act as retrograde messengers and bind to presynaptic CB1 receptors, whose activation blocks adenylyl cyclase activity. Because LTP is triggered by a signaling cascade involving adenylyl cyclase 1 activation, the production of cAMP, and subsequent PKA activation, CB1 receptors can efficiently control this presynaptic form of plasticity.

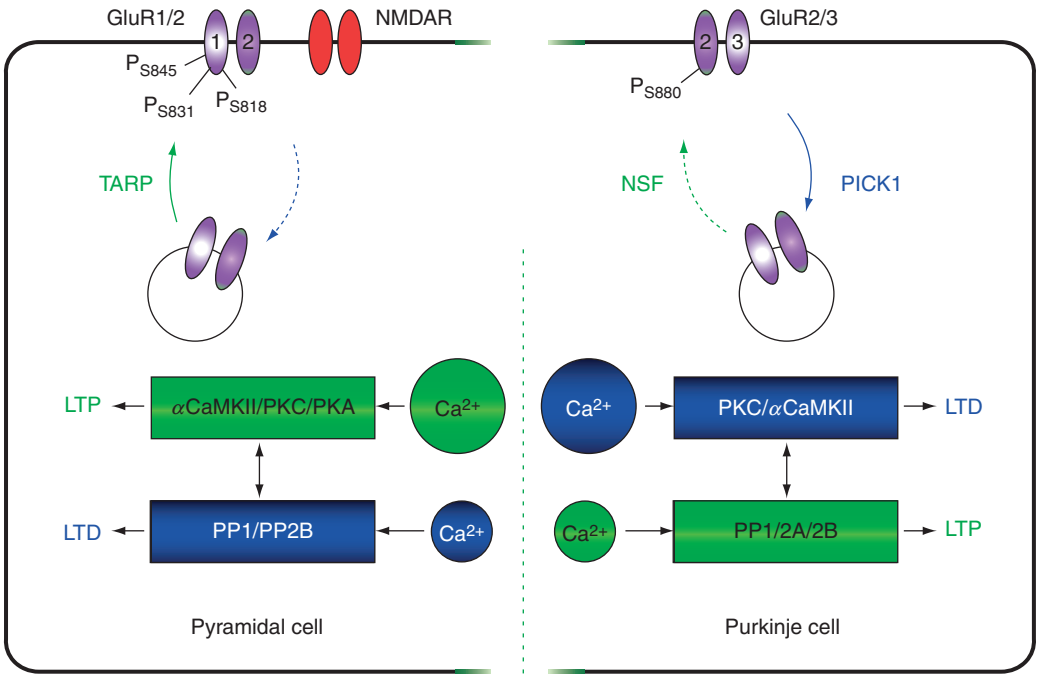

Figure 2 LTP and LTD induction cascades at hippocampal (left) and cerebellar synapses (right). LTD induction cascades are shown in blue, and LTP induction cascades are shown in green. The dashed line separating the 'pyramidal cell' from the 'Purkinje cell' does not continue all the way to the top because PICK1-controlled GluR2 endocytosis (right) occurs in both types of neurons. *N*-methyl-D-aspartate receptor (NMDAR) TARP, transmembrane AMPA receptor regulatory protein. Reproduced from Jörntell H and Hansel C (2006) Synaptic memories upside down: Bidirectional plasticity at cerebellar parallel fiber–Purkinje cell synapses. *Neuron* 52: 227–238, with permission from Elsevier.

Cerebellar LTD and LTP: Mirror Image of Hippocampal Plasticity

The signaling cascades involved in LTP and LTD induction at excitatory synapses onto hippocampal and neocortical pyramidal cells have been particularly well studied. In fact, the synapses between Schaffer collaterals and CA1 pyramidal cells are considered as reference synapses for LTP and LTD studies in general.

At these types of synapses, both LTP and LTD can be triggered by dendritic calcium transients. In 1982, Bienenstock, Cooper, and Munro suggested that synaptic modification is controlled by an activity-related two-threshold mechanism. In an updated version, in which the general term 'activity' was more specifically replaced by amplitudes of postsynaptic calcium signaling, the 'BCM rule' suggests that a higher calcium threshold amplitude has to be reached for LTP than for LTD induction. In contrast, LTP and LTD at cerebellar PF synapses are governed by an 'inverse' calcium threshold mechanism: a higher calcium threshold amplitude has to be reached for LTD than for LTP induction (Figure 2).

Can this mirror-image-like arrangement of induction cascades also be found downstream of the calcium transients? AMPA receptors are heteromeric complexes of the four homologous subunits GluR1–GluR4. In hippocampal pyramidal cells, the majority of AMPA receptors consists of GluR1–GluR2 and GluR2–GluR3 heteromeric complexes. Hippocampal plasticity is assumed to primarily rest on modifications of GluR1 subunits, resulting in single channel conductance changes and/or altered GluR1 subunit trafficking. LTP induction is promoted by large calcium transients that activate αCaMKII, PKC, and PKA, whereas LTD induction is promoted by direct or indirect activation of protein phosphatases 1 and 2B by prolonged, low-amplitude calcium transients. Although these findings suggest that LTP and LTD, respectively, are determined by the phosphorylation level of only one GluR1 phosphorylation site, it currently seems that at least three different phosphorylation sites on GluR1 are involved as well as several scaffolding/anchoring proteins.

In cerebellar Purkinje cells, GluR1 subunits are only weakly expressed and the majority of AMPA receptors consist of GluR2–GluR3 heteromeric complexes. At the level of kinase/phosphatase requirements, bidirectional cerebellar plasticity provides an almost perfect mirror image of its hippocampal counterpart. PF-LTD induction is dependent on the activation of PKC and αCaMKII, whereas PP1, PP2A,

and PP2B are involved in LTP induction (Figure 2). In the case of PKC, the subsequent steps of the signaling cascade have been described in detail: PKCα phosphorylates GluR2 at serine 880, which initiates GluR2 endocytosis. Evidence suggests that a similar PKC-triggered cascade leading to GluR2 endocytosis is also involved in hippocampal LTD induction.

Conclusion

It seems that LTP and LTD at PF synapses in the cerebellum are governed by induction cascades that, at the level of calcium and kinase/phosphatase dependencies, operate 'inverse' in comparison to their hippocampal counterparts. At the level of AMPA receptor trafficking, the 'mirror image' does not describe perfectly well the difference between pyramidal cells and Purkinje cells but, rather, that between GluR1–GluR2 heteromeres and GluR2–GluR3 heteromeres.

See also: Calcium-Calmodulin Kinase II (CaMKII) in Learning and Memory; Cerebellum: Models; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Depression: Cerebellum; Long-Term Depression (LTD): Metabotropic Glutamate Receptor (mGluR) and NMDAR-Dependent Forms; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Depression (LTD): Endocannabinoids and Cerebellar LTD; Long-Term Potentiation (LTP); Synaptic Plasticity: Diacylglycerol Signalling role; Synaptic Plasticity: Short-Term Mechanisms.

Further Reading

Bienenstock EL, Cooper LN, and Munro PW (1982) Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience* 2: 32–48.

Boyden ES, Katoh A, and Raymond JL (2004) Cerebellum-dependent learning: The role of multiple plasticity mechanisms. *Annual Review of Neuroscience* 27: 581–609.

Chung HJ, Steinberg JP, Huganir RL, and Linden DJ (2003) Requirement of AMPA receptor GluR2 phosphorylation for cerebellar long-term depression. *Science* 300: 1751–1755.

Coesmans M, Weber JT, DeZeeuw CI, and Hansel C (2004) Bidirectional parallel fiber plasticity in the cerebellum under climbing fiber control. *Neuron* 44: 691–700.

Eccles JC, Ito M, and Szentagothai J (1967) *The Cerebellum as a Neuronal Machine*. Berlin: Springer.

Hansel C (2005) When the B-team runs plasticity: GluR2 receptor trafficking in cerebellar long-term potentiation. *Proceedings of the National Academy of Sciences of the United States of America* 102: 18245–18246.

Ito M (2002) The molecular organization of cerebellar long-term depression. *Nature Reviews Neuroscience* 3: 896–902.

Ito M, Sakurai M, and Tongroach P (1982) Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells. *Journal of Physiology* 324: 113–134.

Jörntell H and Hansel C (2006) Synaptic memories upside down: Bidirectional plasticity at cerebellar parallel fiber–Purkinje cell synapses. *Neuron* 52: 227–238.

Lev-Ram V, Wong ST, Storm DR, and Tsien RY (2002) A new form of cerebellar long-term potentiation is postsynaptic and depends on nitric oxide, but not cAMP. *Proceedings of the National Academy of Sciences of the United States of America* 99: 8389–8393.

Sacchetti B, Scelfo B, Tempia F, and Strata P (2004) Long-term synaptic changes induced in the cerebellar cortex by fear conditioning. *Neuron* 42: 973–982.

Salin PA, Malenka RC, and Nicoll RA (1996) Cyclic AMP mediates a presynaptic form of LTP at cerebellar parallel fiber synapses. *Neuron* 16: 797–803.

Schmolesky MT, Weber JT, DeZeeuw CI, and Hansel C (2002) The making of a complex spike: Ionic composition and plasticity. *Annals of the New York Academy of Sciences* 978: 359–390.

Steuber V, Mittmann W, Hoebeek FE, et al. (2007) Cerebellar LTD and pattern recognition by Purkinje cells. *Neuron* 54: 121–136.

Wang SSH, Denk W, and Häusser M (2000) Coincidence detection in single dendritic spines mediated by calcium release. *Nature Neuroscience* 3: 1266–1273.

Synaptic Plasticity: Diacylglycerol Signalling role

W S Sossin and C A Farah, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Diacylglycerol and Synaptic Plasticity

Diacylglycerol (DAG) is one of the most important second messengers produced by neurotransmitters. In this article, we highlight the mechanisms for production and removal of DAG by neurotransmitters, the targets of DAG, and the role that these targets play in neuronal plasticity.

Phospholipase Cs are the Major Source of DAG Production

DAG is a glycerol derivative that has two hydroxyl groups esterified by fatty acids (groups R1 and R2 in Figure 1). While DAG is a normal metabolite produced during the synthesis of many membrane lipids, steady-state levels of DAG are low and its production is stimulated upon activation of different cellular cascades. The critical enzyme that generates DAG after extracellular signals is the phosphatidylinositol-4,5-biphosphate (PIP_2)–specific phospholipase C (PLC) which produces DAG and inositol-1,4,5-triphosphate (IP_3) from PIP_2 (Figure 1). Another way for signals to produce DAG is from phosphatidylcholine (PC) by two reactions involving PC-specific phospholipase D (PLD) isozymes and phosphatidic acid (PA) phosphohydrolase (PAP) (Figure 1). This second pathway has been proposed to be important for more persistent increases in DAG but has not yet been established in any plasticity paradigms. It should be noted that the DAG produced from PC differs from that produced by PIP_2 in the saturation of positions R1 and R2. It is not yet known if different DAGs can differentially activate distinct targets.

Phospholipase C

There are four classes of PLCs: β, γ, δ, and ε (Figure 2). PLCs β, γ, and δ have in common a PH domain, connected through an EF hand motif to the catalytic region (Figure 2). C-terminal extensions and inserts in the catalytic domain are isoform and class specific. Distinct processes regulate different classes of PLC: PLCβ isoforms are activated downstream of G protein-coupled receptors (GPCRs); PLCγ is activated downstream of receptor tyrosine kinases (RTKs); PLCδ is activated by calcium ions; and PLCε is activated by small GTPases including Ras and Rho. Below is a brief description of the regulation of each class of PLC and its relevance for synaptic plasticity.

PLCβ PLCβ isoforms are activated both by $G\alpha_q$ and $\beta\gamma$ subunits; however, different isoforms show relative preferences for activation by either $G\alpha_q$ or $\beta\gamma$ with PLCβ1, β3 and β4 best activated by $G\alpha_q$ and PLCβ2 and β3 being best activated by $\beta\gamma$. The two effectors bind to distinct regions of PLCs, with $G\alpha_q$ binding to the C-terminal extension and $\beta\gamma$ binding to a loop in the catalytic domain. Thus the two pathways can be additive. PLCβ is also an effective GTPase for $G\alpha_q$, stimulating GTP hydrolysis and inactivation of the effector. This is one reason why PLC signaling and DAG production downstream of GPCRs is usually transient.

In the brain, most PIP_2 turnover is mediated through $G\alpha_q$ due to the abundance of PLC isoforms β1 and β4, and neurotransmitters that signal through $G\alpha_q$-coupled receptors, particularly 5-HT 2 receptors, type I metabotropic glutamate receptors (mGLURs) 1 and 5, and muscarinic receptors (m1 and m3), all activate PLCβ isoforms in the brain. Interestingly, mGLURs are more coupled to PLCβ4 while muscarinic receptors are more coupled to PLCβ1. Thus, PLCβ4 knockout animals have major deficits in mGLUR-mediated PIP_2 turnover, and several forms of neuronal plasticity are mediated by mGLURs, such as developmental pruning of climbing fibers in the cerebellum, cerebellar long-term depression (LTD), forms of neuropathic pain and circadian rhythmicity. In contrast, PLCβ1 knockout animals have major developmental disorders and seizures through loss of developmental plasticity in the cortex and a major deficit in muscarinic-mediated PIP_2 turnover, with no change in mGLUR PIP_2 turnover.

In *Drosophila* photoreception, a specific isoform of PLC is part of a signaling complex held together through protein–protein interactions with both upstream and downstream effectors, and these types of signaling complexes may also be important in the mammalian brain to provide for additional specificity such as described above for differential regulation of PLC isoforms.

PLCγ PLCγ isoforms contain SH2 domains that directly bind to phosphorylated tyrosines of activated RTKs. This binding both localizes PLCγ to membranes where its substrate PIP_2 exists and leads to phosphorylation of PLCγ on tyrosines, which

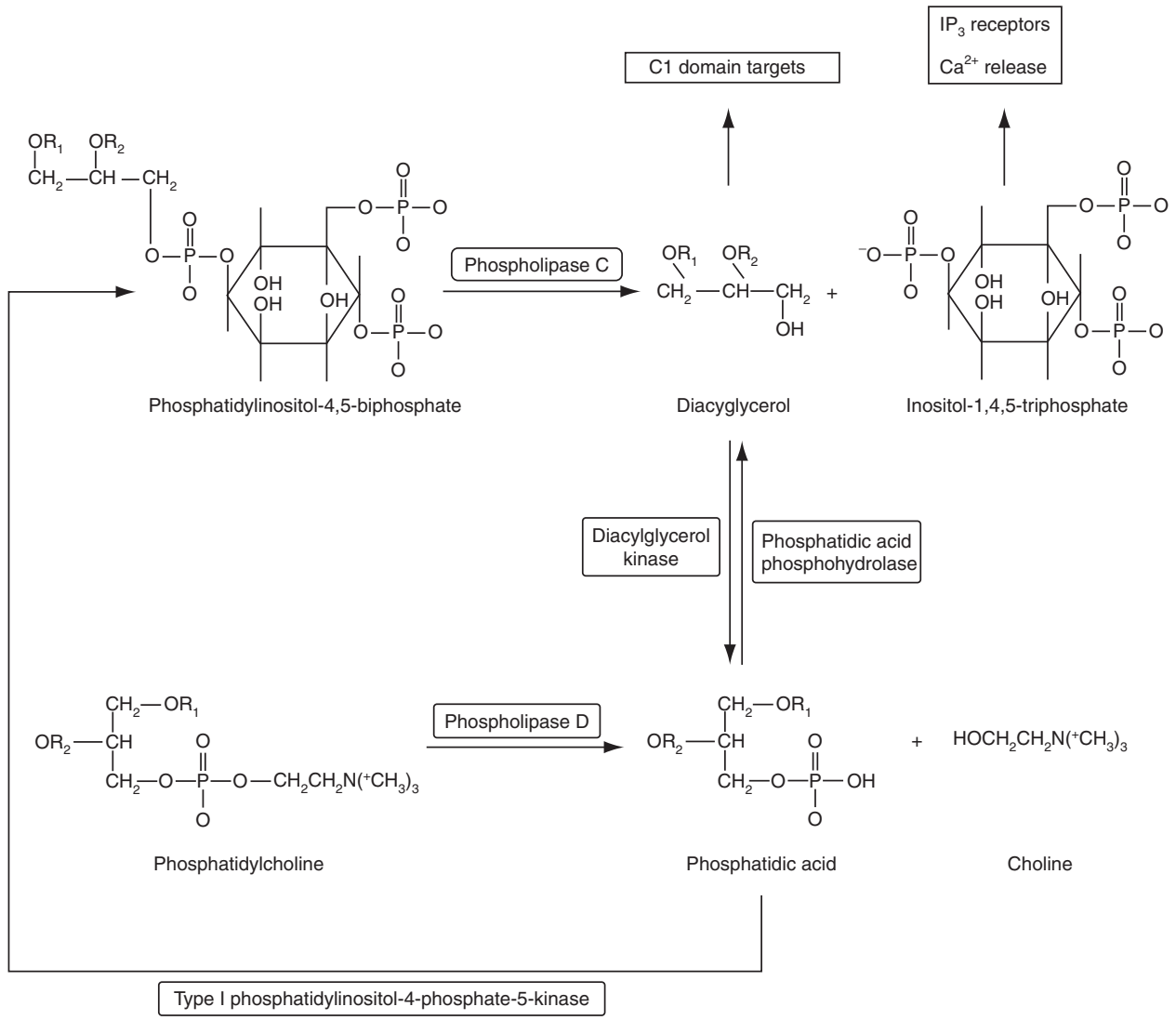

Figure 1 *Sn*-1,2-Diacylglycerol (DAG) can be generated from two different pathways. The first one involves hydrolysis of PIP_2 into DAG and IP_3 by phospholipase C. IP_3 acts on IP_3 receptors to trigger calcium release from endoplasmic reticulum stores. The second pathway, which generates DAG from PC, comprises a two-step reaction involving PLD and PAP, respectively. PA generated by the action of PLD or DAG kinase can be converted back to PIP_2 by the action of type I phosphatidylinositol-4-phosphate-5-kinase. DAG kinases terminate DAG signal by phosphorylating DAG to generate PA. Groups R1 and R2 are fatty acids.

increases its activity. Interestingly, activation of PLCγ downstream of tropomyosin receptor kinase (Trk) activation is particularly important for Trk signaling in synaptic plasticity such as long-term potentiation (LTP).

PLCδ PLCδ is particularly sensitive to regulation by calcium ions. While all PLCs can be regulated by calcium binding to the EF hand motif, PLCδ is the only isoform regulated by physiological levels of calcium. PLCδ is responsible for DAG production that is observed downstream of *N*-methyl-D-aspartate (NMDA) receptor activation and activation of voltage-gated calcium channels independently of GPCR activation.

PLCε PLCε has two unique regions responsible for activation by small GTPases. Ras activates PLCε through a C-terminal Ras-association domain, while Rho activates PLCε through an insertion in the catalytic domain. Modulators that act through $G\alpha_{12,13}$ such as thrombin and lyosphosphatidic acid (LPA) activate DAG production through Rho and PLCε.

DAG Kinases Are the Key Enzymes Responsible for Termination of the DAG Signal

As a second messenger, DAG can regulate the function of many signaling proteins by recruiting them and

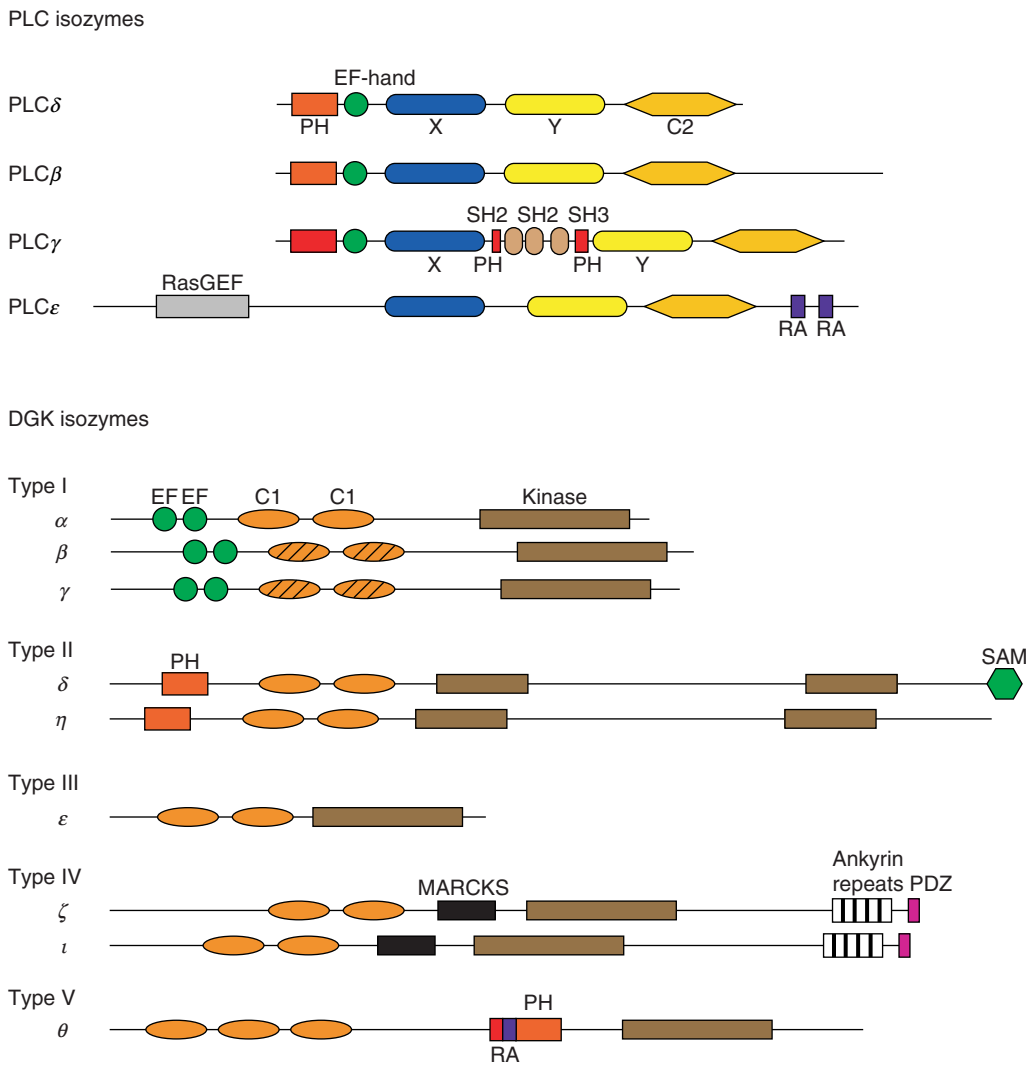

Figure 2 Schematic representation of the PLC and DGK isozymes. PH, pleckstrin homology domain; SH2 and SH3, Src homology domains 2 and 3; X and Y, split catalytic domain of PLC; RA, Ras association domain; RasGEF, Ras guanine nucleotide exchange factor-like domain; SAM, sterile alpha motif; MARCKS, myristoylated alanine-rich C-kinase substrate like-domain. The shaded C1 domains of DGKβ and DGKγ are typical C1 domains, which bind to phorbol-esters, while the other DGK isoforms contain atypical C1 domains, PDZ (postsynaptic density protein 95/discs large/zonula occludens protein 1 domain).

activating them. As for all second messengers, mechanisms exist for removing the second messenger and this pathway is highly regulated. This is best illustrated by the tumor-promoting agents phorbol esters, which are DAG analogs that are only slowly metabolized and lead to the promotion of tumor formation by persistently activating DAG targets. DAG is metabolized in three ways: addition of cytidine diphosphate–choline (CDP–choline) or –ethanolamine to form PC or phosphatidylethanolamine, hydrolysis of a fatty acyl chain by DAG lipase to produce a monacylglycerol and a free fatty acid, and phosphorylation of the free hydroxyl group by DAG kinase (DGK) to generate PA. DGK is considered to be the major pathway for removal of the second messenger pool of DAG.

DAG Kinase

There are nine known mammalian DGK isoforms (Figure 2). Most, if not all of them, are expressed abundantly in the brain. Based on the structure of their regulatory domain, the DGKs are grouped into five subtypes. DGKs need to be translocated to DAG-rich membranes to be effective, and isoforms differ in their mechanisms for membrane translocation. For example, type II DGKs are translocated through their PH and SAM domains, while type IV DGKs have PDZ (postsynaptic density protein 95/discs large/zonula occludens protein 1 domain) and ankyrin binding domains. All DGKs also have C1 domains that bind phorbol esters/DAG in other

proteins (see below). However, in most DGKs, C1 domains do not bind phorbol esters (with the exception of DGKβ and DGKγ), and probably do not bind DAG, but allow regulation by other lipids. Genetic evidence suggests that DGKε participates in the regulation of synaptic transmission. Mice lacking DGKε were shown to have reduced LTP in the perforant path and are resistant to electroconvulsive shock. However, these results are probably not due to acute affects of signal prolongation, but to long-term changes in levels of various lipids that are altered due to feedback mechanisms that occur when DAG levels are unregulated.

DAG Targets Have C1 Domains

For DAG to act as a second messenger, it needs to bind to downstream effector modules. Other than the possible exception of Trp channels (see below), all DAG targets contain a high-affinity module known as a C1 domain based on its initial identification as PKC-conserved domain 1. C1 domains are also the targets of the DAG analogs, phorbol esters. The C1 domain is a zinc-finger-like motif of 50–51 amino acids that is rich in cysteine residues and possesses the pattern $HX_{12}CX_2CX_{13/14}CX_2CX_4HX_2CX_7C$ (H is histidine, C is cysteine, and X is any other amino acid). The two histidines and five of the cysteines are responsible for coordinating two Zn^{2+} ions. The zinc-finger motif forms a single ligand-binding site for DAG or phorbol-ester. Some proteins possess two copies of the C1 domain named C1A and C1B, but each C1 domain can bind DAG independently. The structure of the C1B domain of PKCδ in complex with the high-affinity DAG analog, phorbol-ester, has been elucidated. The domain consists of two small β-sheets and a C-terminal α-helix. Phorbol-ester binding to the C1 domains of PKC does not induce conformational changes. It rather alters the nature of the protein surface and stabilizes the membrane-inserted state. Phorbol ester inserts lengthwise into a groove between the two pulled-apart β-strands at one tip of the domain forming a contiguous hydrophobic surface that mediates membrane targeting. Figure 3(a) shows the molecular surface properties of the C1 domain, and a model for the membrane-inserted form of the C1-phorbol ester complex is shown in Figure 3(b). Thus, the C1 domain allows protein to target themselves to DAG-rich membranes when DAG is produced as a second messenger. A fairly similar structure is seen in the C1 domain of Munc13-1, another important target of DAG (Figures 3(c) and 3(d)). The major difference is in the orientation of a tryptophan residue that partially occludes the DAG and phorbol ester-binding site. This may lead to lower affinity for these targets, or to more substantial conformational changes after DAG binding.

Not all C1 domains bind DAG, as noted above for DGKs. These include Raf-1 and PKCζ, whose C1 domains lack essential properties for DAG/phorbol ester recognition and are thus called atypical C1 domains. Atypical C1 domains often bind related lipids such as PA.

DAG Targets

Below we list the known DAG targets and summarize how DAG binding affects their activity. In some cases, such as for PKC, the major role for DAG is to translocate the protein to the membrane, while for others, such as myotonic dystrophy kinase-related cdc42-binding kinase (MRCK), DAG directly stimulates enzyme activity. In many cases, the target is a substrate for PKC, and thus localized pools of DAG may act to co-localize the kinase and substrate leading to efficient substrate phosphorylation.

Protein Kinase C's

The PKC isozymes are the most well known targets of DAG. PKCs are grouped into three subclasses: the classical PKCs or cPKCs, the novel PKCs or nPKCs, and the atypical PKCs or aPKCs (Figure 4). aPKCs have a single C1 domain that does not bind DAG, and thus is not a direct target of DAG. All PKCs consist of a regulatory domain, containing an autoinhibitory pseudosubstrate sequence and a catalytic domain. Lipid binding to the regulatory domain (phosphatidylserine (PS) for most PKCs) causes a conformational change, removing inhibition from the pseudosubstrate and allowing the catalytic domain to phosphorylate downstream targets. However, the affinity of the resting enzyme for PS is low and activation of the enzyme requires targeting to PS-containing membranes, mainly through binding of the C1 domain to DAG. There are four cPKCs in vertebrates: PKCα, $\beta1$, $\beta2$, and γ, where $\beta1$ and $\beta2$ are spliced isoforms that differ in their C-terminal domain. cPKCs also contain a C2 domain following the C1 domains that allows for calcium-dependent phospholipid binding and synergistic activation by the two messengers. The nPKCs comprise PKCδ, ε, η, and θ. These isoforms also contain a C2 domain; however it differs from the C2 domain of cPKCs in its position (at the N-terminal of the enzyme) and its inability to bind to calcium ions. The C2 domain plays several roles in regulating PKCs: it inhibits binding of DAG to the C1 domain; acts to localize PKC through protein–protein interactions, and, particularly in the cPKCs, it contributes to

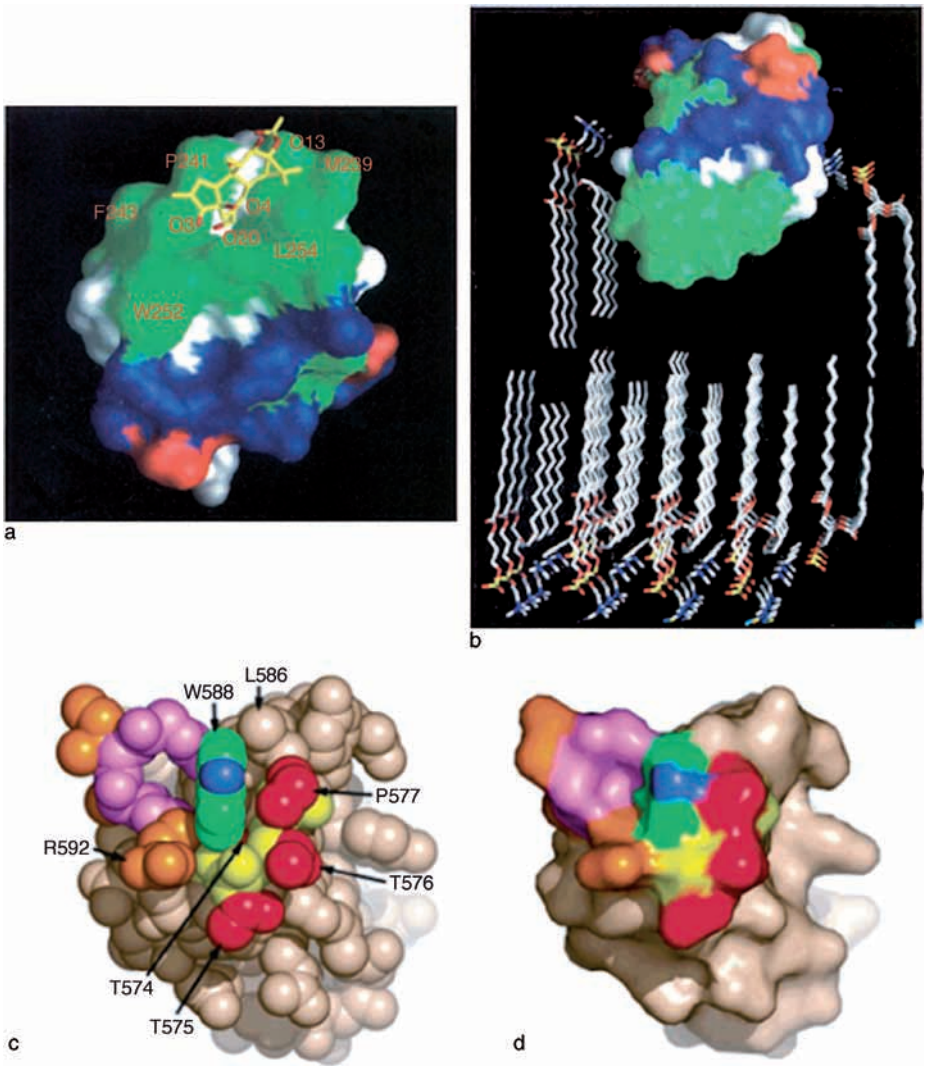

Figure 3 (a) The molecular surface properties of the C1B domain of PKCδ in complex with phorbol-ester (carbons colored yellow and oxygens colored red) binding in the deep groove at the top of the structure. On the protein surface of the C1B domain, hydrophobic residues are green, positively charged residues are blue, negatively charged residues are red, and neutral polar residues are white. (b) represents a model for the membrane-inserted form of the cys2–phorbol-ester complex based on the molecular surface of the protein–phorbol complex and the crystal structure of dimyristoyl phosphorylcholine. The molecular surface of the protein–phorbol complex is colored as in (a), except that phorbol-ester is also colored green. The membrane model is intended only to represent the estimated depth of penetration of the protein into the bilayer. Lipids in front and in back of the domain have been removed for clarity. (c) and (d) represent the occlusion of the ligand binding-site on the Munc13-1 C1 domain by the W588 side chain. (c) is a space filling model and (d) is the accessible surface of the Munc13-1 C1 domain, viewed from the top of the domain, illustrating how the W588 side chain packs against the rest of the domain. Backbone atoms of loops 1 and 2 are colored yellow and violet, respectively, and side chain atoms from these loops are colored red and orange, respectively, except for those of the W588 side chain, which are colored green (for carbon atoms) and blue (for nitrogen atoms). The positions of some of the side chains are labeled. (a, b) Reproduced from Zhang G, Kazanietz MG, Blumberg PM, and Hurley JH (1995) Crystal structure of the cys2 activator-binding domain of protein kinase Cδ in complex with phorbol ester. *Cell* 81: 917–924, with permission from Elsevier. (c, d) Reprinted with permission from Shen N, Guryev O, and Rizo J (2005) Intramolecular occlusion of the diacylglycerol-binding site in the C1 domain of Munc13-1. *Biochemistry* 44(4): 1089–1096, Copyright (2005) American Chemical Society.

membrane binding and specificity. PKCs are also regulated by phosphorylation. Three ordered phosphorylations are required for the enzyme to gain catalytic competence. The first phosphorylation occurs on a loop near the beginning of the catalytic site called the activation loop by phosphoinositide-dependent kinase (PDK1). This phosphorylation correctly positions residues for catalysis and unmasks the entrance to the substrate-binding site. Phosphorylation at the activation loop rapidly triggers the autophosphorylation at the turn motif in a proline-rich region. Activity is regulated by phosphorylation at the final site known

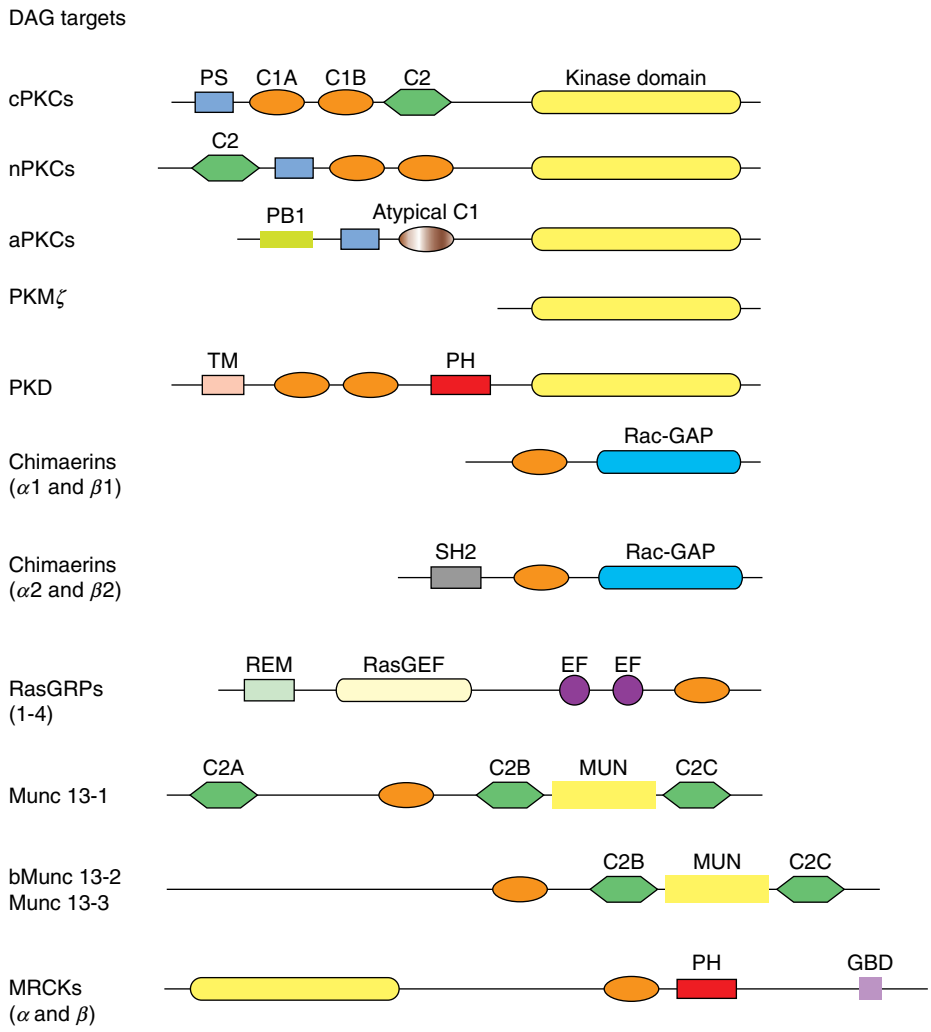

Figure 4 Schematic representation of the DAG targets. PS, pseudosubstrate sequence; TM, transmembrane domain; PH, pleckstrin homology domain; SH2, Src homology domain; RasGEF, Ras guanine nucleotide exchange factor-like domain; Rac-GAP, GTPase activating protein domain; EF, EF hand; REM, Ras exchange motif; MUN, Munc autonomously folded α-helical domain; GBD, GTPase binding domain; PB1, Phox and Bem1 domain.

as the hydrophobic motif. This site may be autophosphorylated or phosphorylated by the kinase, target of rapamycin (TOR). However, unlike many TOR substrates, phosphorylation of PKC is rapamycin insensitive and is a product of the TOR-Rictor complex.

Finally, PKC is regulated through protein–protein interactions. Some PKCs bind directly to actin, which is important for PKC regulation of the cytoskeleton. Scaffolding proteins like receptors for activated C kinase (RACKs), A kinase anchoring proteins (AKAPs), and PDZ domain-containing proteins organize PKC into signaling complexes, localizing them close to particular substrates.

Munc13s

Munc13 proteins consist of four mammalian homologs (Munc13-1, -2 and -3, and -4) of the *Caenorhabditis elegans* Unc13 protein that was isolated in a screen for uncoordinated worms. Munc13-2 is expressed as a brain specific (b) and ubiquitous (ub) splice variant, whereas Munc13-1 and -3 are only found in the brain and Munc13-4 is expressed mainly outside the brain. Munc13s contain a central C1–C2 tandem domain, a large Mun domain, and a C-terminal C2 domain (Figure 4). Munc13–1 also contains an additional N-terminal C2 domain (Figure 4). Munc13 isoforms are specifically localized to presynaptic active zones, the neurotransmitter secreting compartment, and are critical for synaptic vesicle priming. In the absence of Munc13, there is no transmitter release. The Mun domain is sufficient to reconstitute transmitter release in Munc13-negative cells, while the other domains act to regulate the level of Munc13 activity. At the molecular level, Munc13s are thought to act by unfolding and

activating the SNARE protein syntaxin through regulating its interactions with another priming factor Munc18. All Munc13s bind to DAG/phorbol-esters through their C1 domains, and this leads to translocation to the plasma membrane in response to DAG binding. Translocation to the membrane and possible conformational changes that occur upon binding to DAG enhance priming. As evidence for this, phorbol-ester-mediated increases in transmitter release in hippocampal CA1 neurons are lost when Munc13-1 is replaced by a version that does not bind to DAG/phorbol ester. Moreover, mice containing this version of Munc13-1 do not thrive, suggesting a critical role for regulation of Munc13 by DAG for ongoing synaptic transmission.

Chimaerins

Chimaerins were the first family of non-PKC phorbol-ester receptors to be discovered. These proteins resemble a 'chimaera' between the typical C1 domain of PKC isozymes and the GTPase-activating protein (GAP) domain of the breakpoint cluster region protein (BCR), involved in chronic myelogenous leukemia (Figure 4). The GAP domain has activity for Rac and was shown to promote the hydrolysis of GTP to GDP from this small GTP-binding protein. *n*-Chimaerin (later named α1-chimaerin) was first isolated from the brain. Three other isoforms were discovered later: α2-, β1-, and β2-chimaerins produced from alternative splicing of the α- and β-chimaerin genes. The C1 domains of α- and β-chimaerins are almost identical. The main difference between the spliced variants is an SH2 domain located at the N-terminal of α2- and β2-chimaerins (Figure 4). Phorbol-ester treatment was shown to induce translocation of β2-chimaerin from the cytosolic compartment to the Golgi and the plasma membrane. However, phorbol-esters do not affect the GTPase-activating function of chimaerins, suggesting that the function of DAG binding is primarily to translocate chimaerins to membranes, thus spatially restricting their Rac-GTPase-activating effects and to place them near active PKC, since phosphorylation of chimaerin by PKC is required for chimaerin activity.

Ras Guanyl Nucleotide-Releasing Proteins (RasGRPs)

The RasGRP family is composed of four isoforms, RasGRP1–4. All but RasGRP4 are highly expressed in the brain. RasGRPs contain a guanine nucleotide exchange factor domain for Ras-like GTPases (RasGEFN/RasGEF) at their N-terminal. They also contain two EF hand motifs and a C-terminal C1 domain (Figure 4). RasGRPs promote GDP/GTP exchange and activation of Ras and related small GTPases through their RasGEFN/RasGEF domains. This leads to the stimulation of the Raf-MEK-ERK cascade and to the regulation of other signaling pathways. All the RasGRPs contain C1 domains that bind DAG/phorbol-esters. RasGRP1, -3, and -4 were shown to translocate to membrane compartments in response to phorbol-esters and, similar to other Ras guanine nucleotide exchange factors, this may be sufficient to activate the enzyme. In addition, translocation to DAG rich membranes may be important to facilitate phosphorylation of RasGRPs by PKCs that activates some, but not all isoforms of RasGRP. RasGRPs are essential for DAG-mediated increases in ERK phosphorylation in immune cells, but the contribution of RasGRPs to ERK activation downstream of DAG production in neurons is not known.

Protein Kinase D

The protein kinase D (PKD) family consists of PKD1, PKD2, and PKD3. All PKDs contain two C1 domains, a negatively charged central domain, a PH domain, and a serine/threonine protein kinase catalytic domain (Figure 4). PKD1 contains an additional N-terminal apolar domain. Similar to RasGRPs and chimaerins, the C1 domains of PKD1 are not only important for translocation, but also to localize them near PKCs; since phosphorylation of PKD by PKC is required for enzyme activity. Interestingly, the majority of PKD translocation is to the Golgi apparatus and translocation to the Golgi apparatus is also seen in some isoforms of PKC, such as PKCε and η. PKDs are important for vesicle budding from the Golgi and Golgi structure. It is not clear whether DAG production in the Golgi compartment is regulated by extracellular messenger or may be more important in intracellular homeostatic pathways.

Myotonic Dystrophy Kinase-Related cdc42-Binding Kinase

MRCKα and β, are kinases related to Rho kinase (Figure 4) and are implicated in regulation of neurite outgrowth and other cytoskeletal rearrangements through phosphorylation of the myosin light chain phosphatase. While MRCKs bind to cdc42, this does not increase enzyme activity; instead, MRCKs are activated by phorbol-esters binding to their C1 domain.

Trp Channels and the Visual Response in *Drosophila*

Trp channels are cation channels involved in many types of neuronal processes including sensory

transduction such as mechano- and heat sensation. Some forms of Trp channels appear to be regulated by DAG even in the absence of a C1 domain. It is not yet clear if Trp channels have a unique DAG binding domain, or whether activation by DAG is indirect. The most powerful example of this regulation is seen in visual transduction in *Drosophila*, where Trp channel opening is the end result of the phototransduction cascade. In *Drosophila* photoreceptors, the Trp channel, PKC, and PLC are all organized into a signaling complex by a scaffolding PDZ protein. The Trp channel is either directly activated by DAG or by a further metabolite of DAG, while PKC activation is required for signal termination and adaptation. The importance of DGK is also illustrated by this example since in the absence of DGK, the light response is constitutively on leading to retinal degeneration.

Synaptic Plasticity and Behavior

Many transmitters act through GPCRs or RTKs to produce DAG. In the following sections, we review several of the important forms of plasticity regulated through this pathway and, when appropriate, discuss behaviors that are regulated by this plasticity.

Increases in the Probability of Transmitter Release

Many neurotransmitters increase transmitter release through DAG production. There are two major targets of DAG in the presynaptic terminal, PKC and Munc13, and their importance for DAG actions can vary widely in different neurons. Activation of Munc13 by DAG increases synaptic vesicle priming, as noted above, and increased priming will lead to an increased probability of transmitter release. PKCs regulate transmitter release in several ways. PKC phosphorylation of calcium channels relieves inhibition by $\beta\gamma$ subunits, although this regulation is only important for the reversal of previous inhibitory transmitter action. In *Aplysia* neurons, PKC also regulates transmitter release in a state-dependent manner; activation of PKC by DAG does not increase transmitter release at synapses that had not recently fired, but can reverse synaptic depression at synapses that have recently been active. This is thought to mediate behavioral dishabituation. PKC phosphorylates the SNARE protein SNAP-25, and this can increase the size of the readily releasable pool. PKC regulation of the cytoskeleton may also regulate transmitter release. Activation of PKCε increases glutamate release from synaptosomes, and this requires a specific actin-binding site located between the two C1 domains.

Ion Channels

Slow neurotransmission is the process by which neurotransmitters alter membrane properties through activation of second messengers, such as DAG. Most slow neurotransmission downstream of DAG is due to the direct phosphorylation of ion channels by PKC. There are numerous ion channels regulated by PKC phosphorylation including voltage-gated channels (potassium, sodium), cation channels, calcium-dependent potassium channels, and ligand-gated channels such as AMPA and NMDA receptors. PKC regulates many aspects of channel functions including efficacy, inactivation rates, and cycling between the plasma membrane and internal stores. While there are many forms of synaptic plasticity that utilize this mechanism, two examples are described below where there is a clear link between synaptic plasticity and behavior.

Pain sensitization One particularly important example is the regulation of primary nociceptor channels, such as vanilloid receptors by PKCε. Acute sensitization of these receptors is mediated by peptides that act through PLC such as epinephrine and bradykinin, and increased levels of PKCε may mediate the prolonged sensitization of these receptors observed in persistent pain models. Thus, in animals lacking PKCε, there is a decrease in the ability to sensitize nociceptors through a number of different stimuli.

Cerebellar LTD One attraction of the dual role of calcium and DAG in activation of cPKCs is that PKC can be an associative molecule. This is exemplified in Purkinje neurons where PKC is activated by the conjunction of granule cell and climbing fiber activity. The activation of PKC requires DAG produced by activation of mGLUR1 after granule cell activity and calcium produced by depolarization and opening of voltage-gated calcium channels after climbing fiber activity. PKC then phosphorylates the ionotropic glutamate receptor GluR2 at a residue near the C-terminal that causes endocytosis of the receptors through regulating the binding partners of this C-terminal sequence. This endocytosis leads to a LTD of the synaptic response from the granule cells. Interestingly, this plasticity is mediated by the cPKC, PKCα, and requires the PDZ binding C-terminal of PKCα. Thus generation of a signaling complex is critical for this plasticity. This type of cerebellar LTD is thought to underlie several forms of behavioral plasticity that require Purkinje cells such as the VOR reflex and eyeblink conditioning. Indeed, mice that constitutively express a PKC inhibitor only in Purkinje cells have major problems in learning the timing needed for proper eyeblink conditioning.

Protein Synthesis and Gene Expression

Many neurotransmitters activate gene expression through activation of a cascade including ERK, and this leads to long-term changes in neuronal properties and gene-expression-dependent long-term memories. DAG pathways activate ERK through both PKC phosphorylation of upstream kinases such as β-Raf and through RasGRPs. Regulation of local protein synthesis is also an important mechanism for certain types of synaptic plasticity. Activation of mGLUR1 or -5 in hipppocampal neurons causes a LTD that requires local synaptic translation, and regulation of the translational process is mediated through PKC.

Morphological Plasticity

DAG signaling is important for various types of morphological plasticity, and this is mediated by a number of DAG targets that regulate the cytoskeleton. For example, stimulation of PLCβ-coupled cell surface receptors recruits β1-chimaerin to the plasma membrane resulting in the pruning of dendritic spines and branches. This pruning activity required both the DAG-binding and Rac GAP activity of β1-chimaerin. PKC phosphorylates many cytoskeletal regulators such as growth-associated protein-43 (GAP43), MARCKS, and adducin. Neurite outgrowth is also regulated through other C1 domain proteins including MRCK and RasGRPs, suggesting many overlapping pathways between DAG and the cytoskeleton in order to orchestrate complex morphological changes during neuronal development and perhaps during learning and memory as well.

Persistent Activation of PKCs

Several persistent types of plasticity and behavior are caused by persistent activation of PKC. A form of intermediate sensitization in *Aplysia*, associative learning in *Hermissenda*, and hippocampal LTP can all be blocked in the maintenance phase by PKC inhibitors. However, there is no evidence in these paradigms for persistent increases in DAG levels. Rather, other mechanisms of PKC activation, such as formation of PKMs (PKCs that lack the regulatory domain), or alteration in phosphorylation of PKC are required for the persistent activation of kinase activity. In particular, a form of PKC known as PKMζ (Figure 4) is expressed in the nervous system from an alternative promoter. Specific inhibitors of PKMζ can block LTP even 5 h after induction, suggesting that this form of plasticity requires persistent phosphorylation of target proteins. Thus, at least during the initial phases of memory formation, persistent activation of PKCs can act as a memory trace.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Depression (LTD): Endocannabinoids and Cerebellar LTD; Synaptic Plasticity: Learning and Memory in Normal Aging; Synaptic Mechanisms of Learning.

Further Reading

Khasar SG, Lin YH, Martin A, et al. (1999) A novel nociceptor signaling pathway revealed in protein kinase C epsilon mutant mice. *Neuron* 24: 253–260.

Kim D, Jun KS, Lee SB, et al. (1997) Phospholipase C isozymes selectively couple to specific neurotransmitter receptors. *Nature* 389: 290–293.

Leitges M, Kovac J, Plomann M, and Linden DJ (2004) A unique PDZ ligand in PKC alpha confers induction of cerebellar long-term synaptic depression. *Neuron* 44: 585–594.

Newton AC (1995) Protein kinase C: Structure, function, and regulation. *Journal of Biological Chemistry* 270: 28495–28498.

Rhee JS, Betz A, Pyott S, et al. (2002) Beta phorbol ester- and diacylglycerol-induced augmentation of transmitter release is mediated by Munc13s and not by PKCs. *Cell* 108: 121–133.

Rhee SG (2001) Regulation of phosphoinositide-specific phospholipase C. *Annual Review of Biochemistry* 70: 281–312.

Serrano P, Yao Y, and Sacktor TC (2005) Persistent phosphorylation by protein kinase Mzeta maintains late-phase long-term potentiation. *Journal of Neuroscience* 25: 1979–1984.

Shen N, Guryev O, and Rizo J (2005) Intramolecular occlusion of the diacylglycerol-binding site in the C1 domain of Munc13-1. *Biochemistry* 44(4): 1089–1096.

Tanaka C and Nishizuka Y (1994) The protein kinase C family for neuronal signaling. *Annual Review of Neuroscience* 17: 551–567.

Topham MK (2006) Signaling roles of diacylglycerol kinases. *Journal of Cellular Biochemistry* 97: 474–484.

Yang C and Kazanietz MG (2003) Divergence and complexities in DAG signalling: Looking beyond PKC. *Trends in Pharmacological Sciences* 24(11): 602–608.

Zhang G, Kazanietz MG, Blumberg PM, and Hurley JH (1995) Crystal structure of the cys2 activator-binding domain of protein kinase Cδ in complex with phorbol ester. *Cell* 81: 917–924.

Synaptic Plasticity: Learning and Memory in Normal Aging

M Baudry, University of Southern California, Los Angeles, CA, USA

© 2009 Published by Elsevier Ltd.

LTP/LTD and Learning and Memory

Since the first publication reporting the existence of long-term potentiation (LTP) by Bliss and Lomo, there has been a plethora of papers, books, symposia, and meetings devoted to the discussion of the relationship between this molecular/cellular form of synaptic plasticity and some behavioral forms of learning and memory. Although the issue may occasionally still be debated, there is little doubt that LTP indeed represents one of the cellular mechanisms that is used in a number of neuronal networks for the storage of new information. This has been particularly evidenced in the hippocampal formation, to which a wealth of pharmacological and genetic manipulations have been applied to attempt to demonstrate that LTP induction and maintenance are causally related to memory formation and duration. According to this body of work, LTP in hippocampus would be ideally suited for the formation of episodic or declarative memory. LTP has been observed in a large variety of mammalian species, including humans, clearly indicating the universality of this phenomenon.

On the other hand, the relationship between long-term depression (LTD) and learning and memory has not been as clearly established until now. The best case for such an association has been made in cerebellar learning, in which LTD of synaptic transmission at the parallel fibers to Purkinje cell synapses is likely to participate in certain forms of motor learning. In other brain structures, the potential link between LTD and learning of specific types of information is much more tenuous, and further work is needed to establish convincingly that LTD is necessary and sufficient for any particular form of learning. Note that another form of synaptic plasticity referred to as depotentiation has also been discussed as a potential mechanism for memory erasure. Whereas LTP is generally elicited by trains of high-frequency stimulation, depotentiation is generally elicited by trains of low-frequency stimulation delivered to pathways previously potentiated. Whether this form of plasticity is related to 'forgetting' is still a matter of speculation.

This article focuses on age-related alterations in LTP/LTD, on the one hand, and on learning and memory, on the other hand, and we argue that age-related cellular modifications leading to changes in LTP/LTD properties are also responsible for age-related changes in properties of some forms of memory. We first summarize the main findings regarding age-related changes in learning and memory. Changes in LTP/LTD properties are then reviewed. Finally, we discuss the potential cellular origins of age-related modifications in LTP/LTD and illustrate how these changes reflect basic age-related changes in cellular function.

Learning and Memory and Aging

It has long been recognized that aging is associated with a loss of learning and memory; this has led to the definition of age-associated memory impairment and has been a strong stimulus for the development of new drugs that could remedy this impairment first in the extreme case of this condition – that is, Alzheimer's disease – but also in its milder version, known as mild cognitive impairment (MCI). However, it is only recently that this view has been modified with the appreciation that the decline in learning and memory is in fact taking place throughout the life span and is initiated very soon after the peak in learning and memory at the beginning of adulthood. It has been argued that mechanisms of learning and memory are critical during the developmental period when the organism needs to adapt to the environment, whereas the focus of adulthood is on the transmission of genes to the next generation. The extreme form of this argument is that evolution of memory was associated with that of the aging process. Such a view is similar to the idea that plasticity and age-related deterioration of neuronal function involve the opposite requirements for neurons to be stable to resist injuries and insults and to be plastic to continue the adaptation of the organism to a continuously changing environment. According to this view, the existence of mismatch between resources and requirements to satisfy these two sets of constraints results in the progressive impairment of synaptic plasticity during aging.

A wealth of experiments have documented age-related impairments in learning and memory in a wide variety of tasks in experimental animals as well as in humans. Barnes and colleagues in the late 1970s used learning to escape a circular platform to show age-related impairment in spatial learning in rats. Using the Morris water maze to study behavioral learning, numerous laboratories have since reported

that cohorts of aged rats can be subdivided into learning-impaired and learning-nonimpaired animals. Reasons for this dichotomy are not clear, although it is intimately related to a similar dichotomy in plasticity properties at the cellular level. Age-related learning deficits have also been reported in mice both in spatial tasks and in fear conditioning. Trace eyeblink conditioning also requires hippocampal function, and impairment of this form of learning has been found in aged mice, rats, rabbits, and humans. Much less has been done regarding progressive memory decline in experimental animals during the life span, although results indicate that environmental events in early life can produce learning impairment in middle-aged rats.

Age-related impairment in delay eyeblink conditioning, a cerebellar-dependent form of learning, has also been found and, interestingly, is already present in middle-aged mice. It is also present in humans and, in fact, it might even be more dramatic than impairment in a verbal task, which would be considered more declarative and therefore more hippocampus dependent.

Age-Related Changes in LTP/LTD

Although it was initially thought that cognitive decline with aging was due to massive neuronal loss and alterations in neuronal morphology, particularly dendritic branching, reexamination of this issue has led to the conclusion that aging is not associated with a general loss of neurons, dendritic alterations, or loss of synaptic contacts. Likewise, aging is not associated with gross alterations in biophysical properties of neurons in hippocampus. The most significant change in membrane biophysics associated with aging is an increased conductance for calcium, possibly due to an increased number of L-type voltage-dependent calcium channels. As discussed later, this change in calcium conductance may have a critical role in age-related changes in synaptic plasticity and learning and memory.

Although the intimate details of the mechanisms involved in LTP and LTD are not completely known, there is a growing consensus for the major events responsible for the induction and expression of the long-term increase (LTP) or decrease (LTD) in synaptic transmission elicited by the specific patterns of electrical stimulation required to produce them (Figure 1). In the case of LTP, brief bursts of electrical stimulation result in sufficient postsynaptic depolarization to relieve NMDA receptors from the voltage-dependent magnesium blockade of their associated channels. In turn, activation of NMDA receptors results in calcium influx in postsynaptic spines and the stimulation of a host of calcium-dependent processes, including calcium-dependent protein kinases (in particular, calcium/calmodulin protein kinase (CamKII)), calcium-dependent proteases, and calcium-dependent phospholipases. The combined action of these processes results in changes in spine cytoskeleton, the insertion of AMPA receptors in postsynaptic density, and the modification of cell adhesion proteins involved in pre/postsynaptic interactions. It has thus been proposed that dendritic spines modify their structures providing for increased insertion of AMPA receptors or, conversely, that a regulated process of receptor insertion imposes increased surface area for the postsynaptic density. Moreover, integrin activation has been proposed to represent a form of consolidation because interfering with this class of adhesion proteins within a short time after tetanus is capable of preventing LTP maintenance. It has also been proposed that protein synthesis as well as gene transcription participate in the late phase of activity-dependent changes in synaptic structure and function, although the exact nature of the genes activated and the proteins translated remains unknown. Numerous studies point to a critical role for brain-derived nerve growth factor (BDNF) in the translation of brief bursts of electrical activity into long-lasting changes in synaptic efficacy. BDNF would be released as a result of NMDA receptor activation and activate its membrane-associated receptor, trkB, in an autocrine-type manner. Stimulation of trkB would trigger a number of signaling pathways and, in particular, by regulating protein synthesis localized in dendrites, BDNF appears to be a regulator of synaptic consolidation. It is also noteworthy to mention the existence of non-NMDA receptor-dependent LTP that requires activation of the voltage-dependent calcium channels for its induction.

Although the detailed mechanisms underlying LTD are not as clearly established as those for LTP, several of the same processes participating in LTP have also been shown to be activated in LTD. Two major forms of LTD have been particularly well studied – an NMDA receptor-dependent form and a metabotropic glutamate receptor (mGluR)-dependent one. In both cases, internalization of AMPA receptors appears to be responsible for the maintenance of decreased synaptic efficacy.

Several studies have indicated alterations in both LTP and LTD properties in old animals, although there are differences in the details of these age-related modifications depending on the anatomical regions evaluated and the protocols used to elicit LTP or LTD. The first studies performed approximately 30 years ago indicated that following a strong stimulus to elicit LTP at the synapses of the perforant path to dentate gyrus or to CA3 pathways, LTP induction was similar

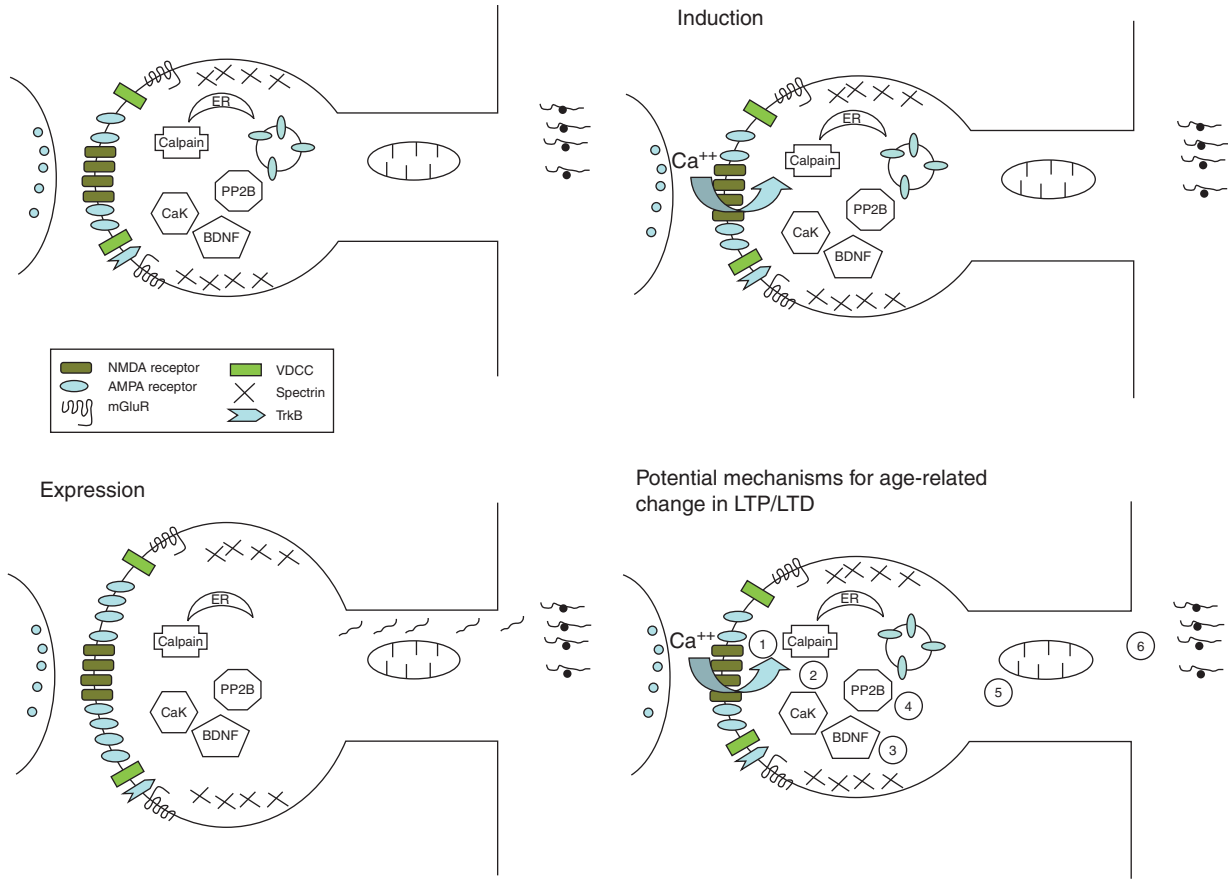

Figure 1 Schematic representation of a glutamatergic synapse and the events involved in induction and expression of synaptic plasticity and potential mechanisms accounting for age-related impairment in synaptic plasticity. (Top left) Dendritic spine with AMPA and NMDA receptors as well as mGluRs and voltage-dependent calcium channels (VDCC). Also represented are calmodulin kinase (CaK), calcineurin (PP2B), and calpain. In addition, an intracellular pool of AMPA receptors is shown. BDNF and the trkB receptors are also present postsynaptically, as are mitochondria and endoplasmic reticulum (ER). Finally, ribosomes with mRNAs are represented in the base of the spine. (Top right) LTP/LTD induction. An influx of calcium is the major trigger for induction of synaptic plasticity. Depending on the amplitude and duration of the calcium signal, LTP or LTD will be elicited due to changes in the balance between protein kinase and protein phosphatase activity and calpain activation. Stimulation of BDNF release and activation of trkB also participate in the regulation of the changes in synaptic structure and function. Note that VDCC, ER, and mitochondria will participate in the regulation of the calcium signal generated by the stimulation protocol. (Bottom left) LTP expression. Increase in synaptic efficacy is due to an increase in the available number of AMPA receptors produced by the insertion of AMPA receptors in the postsynaptic membrane. This effect might be associated with a change in the morphology of the dendritic spine (note the enlarged spine head). Local translation of mRNAs might provide for newly synthesized proteins that could stabilize the modifications produced by proteolytic activation. (Bottom right) Potential mechanisms responsible for age-related impairment in synaptic plasticity. 1, Changes in NMDA receptor function or signaling; 2, changes in calcium regulation (VDCC, ER, and mitochondria); 3, changes in BDNF signaling; 4, changes in calcineurin activity; 5, changes in free radical production or elimination; 6, changes in protein synthesis.

in aged and young rats but decayed faster in the former than in the latter. On the other hand, altered LTP induction was observed at the Schaffer collateral to CA1 pathway. More recent studies have used induction protocols more closely related to natural stimulation frequencies, such as the so-called theta burst stimulation (TBS), in which brief bursts of high-frequency stimulation (generally 100 Hz) are delivered at the frequency of the theta rhythm (5–7 Hz). Under these conditions, LTP induction was impaired at the Schaffer collateral to CA1 synapses. Interestingly, using a 5-Hz stimulation protocol to induce LTP revealed the same dichotomy between subgroups of rats previously mentioned, with rats exhibiting learning impairment in the Morris water maze also exhibiting decreased amplitude of potentiation at these synapses. In addition, another form of NMDA receptor-dependent LTP elicited by prolonged paired-pulse stimulation delivered at 1 Hz appears in aged animals. Finally, one study indicated that, similar to the gradually occurring learning deficits over the life span, deficits in LTP can also be found in middle-aged rats; thus, although LTP induction was not modified in the basal dendrites of

CA1 pyramidal neurons of 7- to 10-month-old rats, it failed to stabilize and excitatory postsynaptic potentials returned to baseline levels within 30 min after TBS.

Age-related changes in LTD are almost the mirror images of those found for LTP. Thus, whereas LTD is difficult to elicit in hippocampal slices prepared from adult and middle-aged rats using prolonged trains of low-frequency stimulation (generally 1 Hz for 15 min), it is readily produced in hippocampal slices from aged rats. Again, an interesting dichotomy was reported regarding LTD properties in hippocampus of learning-impaired versus learning-unimpaired aged rats: learning-unimpaired rats 'switched' from NMDA receptor-dependent LTD to non-NMDA receptor-dependent LTD. No data are available regarding potential age-related changes in LTD in cerebellum.

Potential Cellular Mechanisms Underlying Age-Related Changes in LTP/LTD

Changes in NMDA Receptors

In view of the critical role of NMDA receptors in the induction of LTP and LTD, numerous studies have evaluated potential changes in NMDA receptor properties with age. Although several developmental changes in NMDA receptor properties have been described, it does not appear that changes in NMDA receptors account for age-related changes in LTP/LTD. However, because the NMDA receptor is part of a large complex of more than 200 proteins, it is plausible that changes in the organization/function of this complex take place during aging and modify some intracellular signaling cascades associated with activation of the NMDA receptors.

Changes in Calcium Regulation

As mentioned previously, one of the most conspicuous age-related changes in neuron biophysical properties consists of a change in calcium regulation and, in particular, in an increased number of L-type voltage-dependent calcium channels. Although the increase in calcium channels does not affect basal intracellular calcium concentration, it produces a significantly higher calcium concentration as a result of synaptic activity in neurons from aged rats compared to that in young rats. In turn, this results in increased afterhyperpolarization (AHP) resulting from the activation of calcium-dependent potassium channels. Interestingly, the amplitude of the AHP is a good predictor of learning ability in rabbits, and learning results in decreased amplitude of the AHP. AHP decreases excitability and may thus be responsible for the age-related decrease in LTP because increased AHP would decrease the ability of trains of stimulation to depolarize the postsynaptic membrane sufficiently to activate the NMDA receptors. Results indicate that calcium-induced calcium release from intracellular stores is actually responsible for the increased AHP observed in aged animals, further incriminating calcium dysregulation as the primary cause for age-related impairment in synaptic plasticity. Interestingly, pharmacological or genetic manipulations directed at calcium-activated potassium channels eliminate age-related impairment in LTP and in cognitive function.

Changes in BDNF Signaling

As discussed previously, BDNF has been implicated as a critical regulator of synaptic efficacy in hippocampus and some age-related alterations in synaptic plasticity could be due to age-related changes in BDNF levels or signaling. Conflicting reports regarding changes in expression and levels of BDNF have been published, but the overall view is that there are no significant changes in the levels of BDNF in hippocampus between 2 and 24 months of age. On the other hand, there appears to be a consensus regarding age-related decrease in trkB, the major receptor for BDNF, in hippocampus. Accordingly, BDNF-mediated effects should be significantly reduced in old animals. Understanding the precise role that this decreased signaling plays in age-related decrease in LTP/LTD requires further investigation. Interestingly, whereas increased physical activity does not seem to reverse age-related decrease in BDNF signaling, enriched environment does seem to do so.

Changes in Calcineurin

Calcineurin is a calcium-dependent phosphatase that also plays a critical role in synaptic plasticity because the balance between protein kinase and protein phosphatase activity is a major determinant of the direction and amplitude of activity-dependent changes in synaptic efficacy. It is clear that calcineurin activity is increased with aging, and this has major consequences for the regulation of the state of phosphorylation of several targets of this phosphatase at synapses. Previous results have shown that calcineurin inhibition results in LTP that occludes tetanus-induced potentiation. Moreover, in aged rats, inhibition of calcium release from intracellular stores results in LTP, an effect that is blocked by calcineurin inhibition. It has thus been proposed that calcineurin indirectly regulates the phosphorylation of voltage-dependent calcium channels, thereby contributing to the regulation of cell excitability and LTP induction.

Role of Oxygen Radicals

One of the key contributors to age-related changes in cellular function is oxidative stress, and it has been repeatedly shown that aging is associated with increased protein oxidation and lipid peroxidation as a result of increased formation of reactive oxygen species in the absence of increased levels of detoxifying enzymes (superoxide dismutase, catalase, and peroxidases) and scavengers of free radicals. In recent years, it has been proposed that free radicals exhibit opposite effects on synaptic plasticity, LTP in particular, in adult and aged rats. In the adult, low levels of reactive oxygen species, specifically H_2O_2, would be required for the normal induction of LTP, whereas high concentrations would be detrimental. Moreover, the effects of low concentrations of H_2O_2 on LTP would be mediated through the stimulation of calcineurin and the regulation of voltage-dependent calcium channels. On the other hand, reactive oxygen species would be mostly detrimental to LTP induction in the aged animals. In this case, high concentrations of H_2O_2 are presumed to stimulate calcium release from internal stores, resulting in activation of calcineurin and thus interfering with LTP induction. Furthermore, reactive oxygen species have been shown to interfere with proteolytic mechanisms such as the proteosome and the lysosomes. Studies have shown a critical role for the proteosome/ubiquitin proteolytic pathway in synaptic plasticity.

Changes in Protein Synthesis

Whereas aging does not appear to be associated with major changes in LTP-induced gene expression, the transcription of c-fos appears to be decreased in CA1 pyramidal neurons from old rats. As discussed previously, BDNF enhances translation rate in dendrites, and the decrease in BDNF responsiveness with aging is likely to affect protein synthesis, thereby further contributing to age-related impairment in synaptic plasticity.

Conclusions

It is clear that the data strengthen the argument that LTP and, to some extent, LTD represent cellular correlates of learning and memory. In fact, in several instances, the parallel is striking, as in the studies demonstrating the dichotomy in subgroups of rats between learning-impaired and learning-unimpaired with the similar dichotomy with impaired LTP or normal LTP. The data also provide better evidence for a role of LTD in learning and memory; in particular, in young rats, there is a significant correlation between the amplitude of NMDA receptor-dependent LTD and an index of performance in the learning task. Interestingly, in aged rats, successful learners switch from NMDA receptor-dependent LTD to non-NMDA receptor-dependent LTD. As more information has become available regarding the cellular and molecular mechanisms underlying LTP/LTD, it has become easier to assess which components of the complex machinery linking selective pattern of electrical activity to long-lasting changes in synaptic structure and function are exhibiting age-related alterations. Thus, it has become clear that since LTP/LTD use a number of basic mechanisms of cell biology involved in cell motility, cell signaling, and cell communication, age-related changes in basic cell function have significant impact on the features of LTP/LTD in aged organisms.

See also: Aging of the Brain; Aging: Brain Potential Measures and Reaction Time Studies; Cognition in Aging and Age-Related Disease; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Long-Term Potentiation (LTP): NMDA Receptor Role; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP).

Further Reading

Barnes CA (1979) Memory deficits associated with senescence: A neurophysiological and behavioral study in the rat. *Journal of Comparative Physiology and Psychology* 93: 74–104.

Barnes CA (2001) Plasticity in the aging central nervous system. *International Review of Neurobiology* 45: 339–354.

Barnes CA (2003) Long-term potentiation and the ageing brain. *Philosophical Transactions of the Royal Society of London* 358: 765–772.

Baudry M and Thompson RF (1999) Synaptic plasticity: From molecules to behavior. In: Baudry M, Thompson RF, and Davis JL (eds.) *Advances in Synaptic Plasticity*, pp. 299–319. Cambridge, MA: MIT Press.

Bear MF and Malenka RC (1994) Synaptic plasticity: LTP and LTD. *Current Opinion in Neurobiology* 4: 389–399.

Bergado JA and Almaguer W (2002) Aging and synaptic plasticity: A review. *Neural Plasticity* 9: 217–232.

Blank T, Nijholt I, Kye MJ, Radulovic J, and Spiess J (2003) Small-conductance, Ca^{2+}-activated K^+ channel SK3 generates age-related memory and LTP deficits. *Nature Neuroscience* 6: 911–912.

Bramham CR and Messaoudi E (2005) BDNF function in adult synaptic plasticity: The synaptic consolidation hypothesis. *Progress in Neurobiology* 76: 99–125.

Burke SN and Barnes CA (2006) Neural plasticity in the ageing brain. *Nature Reviews Neuroscience* 7: 30–40.

Chapman PF (2005) Cognitive aging: Recapturing the excitation of youth? *Current Biology* 15: R31–R33.

Cooke SF and Bliss TVP (2006) Plasticity in the human central nervous system. *Brain* 129: 1659–1673.

Davis HP, Small SA, Stern Y, et al. (2003) Acquisition, recall, and forgetting of verbal information in long-term memory by young, middle-aged, and elderly individuals. *Cortex* 39: 1063–1091.

de Magalhaes JP and Sandberg A (2005) Cognitive aging as an extension of brain development: A model linking learning, brain plasticity, and neurodegeneration. *Mechanisms of Ageing and Development* 126: 1026–1033.

Driscoll I and Sutherland RJ (2005) The aging hippocampus: Navigating between rat and human experiments. *Review of Neuroscience* 16: 87–121.

Foster TC and Kumar A (2002) Calcium dysregulation in the aging brain. *Neuroscientist* 8: 297–301.

Kamsler A and Segal M (2003) Hydrogen peroxide modulation of synaptic plasticity. *Journal of Neuroscience* 23: 269–276.

Landfield PW (1994) Increased hippocampal Ca^{2+} channel activity in brain aging and dementia. Hormonal and pharmacologic modulation. *Annals of the New York Academy of Sciences* 747: 351–364.

Lee HK, Min SS, Gallagher M, and Kirkwood A (2005) NMDA receptor-independent long-term depression correlates with successful aging in rats. *Nature Neuroscience* 8: 1657–1659.

Lu T, Pan Y, Kao SY, et al. (2004) Gene regulation and DNA damage in the ageing human brain. *Nature* 429: 883–891.

Malinow R and Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* 25: 103–126.

Mattson M (ed.) (2002) *Diet–Brain Connection: Impact on Memory, Mood, Aging and Disease*. Dordrecht: Kluwer.

Murphy GG, Fedorov NB, Giese KP, et al. (2004) Increased neuronal excitability, synaptic plasticity, and learning in aged Kvbeta1.1 knockout mice. *Current Biology* 14: 1907–1915.

Pfeiffer BE and Huber KM (2006) Current advances in local protein synthesis and synaptic plasticity. *Journal of Neuroscience* 26: 7147–7150.

Serrano F and Klann E (2004) Reactive oxygen species and synaptic plasticity in the aging hippocampus. *Ageing Research Reviews* 3: 431–443.

Soule J, Messaoudi E, and Bramham CR (2006) Brain-derived neurotrophic factor and control of synaptic consolidation in the adult brain. *Biochemical Society Transactions* 34: 600–604.

Thibault O, Hadley R, and Landfield PW (2001) Elevated postsynaptic $[Ca^{2+}]_i$ and L-type calcium channel activity in aged hippocampal neurons: Relationship to impaired synaptic plasticity. *Journal of Neuroscience* 21: 9744–9756.

Tombaugh GC, Rowe WB, Chow AR, Michael TH, and Rose GM (2002) Theta-frequency synaptic potentiation in CA1 *in vitro* distinguishes cognitively impaired from unimpaired aged Fischer 344 rats. *Journal of Neuroscience* 22: 9932–9940.

Woodruff-Pak DS and Finkbiner RG (1995) Larger nondeclarative than declarative deficits in learning and memory in human aging. *Psychology of Aging* 10: 416–426.

Synaptic Plasticity: Neuronal Sprouting

J Nilsen, University of Southern California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In neuroscience the term 'sprouting' refers to any phenomena invoking neurite growth; sprouting occurs through the life span in the mammalian central nervous system (CNS) and in nonmammalian brains and the peripheral nervous system (PNS). These changes in neurite growth are usually represented by modifications to synapses, evidenced as either alterations in synapse number or junctional area. Such alterations can occur in response to normal and pathological stimuli as part of the overall plasticity of nervous system circuitry, believed to underlie changes in behavioral activity. The structural synaptic changes are referred to as 'synaptic plasticity,' as a reference to the term 'behavioral plasticity.' The latter term was first defined by William James in 1890 as any meaningful change in behavior of an organism.

Synaptic plasticity was first recognized during observations of reactive synaptogenesis and sprouting that occurred in response to extension lesions. It was demonstrated that in response to neuritic loss the nervous system was able to form new synapses, and undamaged neurons were able to form new branches (sprouts) with the capability of making new synaptic connections. This led directly to the examination of synaptic plasticity in neurodegenerative diseases and injury. Although the field started with the examination of plastic responses to neuronal injury, it is clear that plasticity occurs throughout life as a normal response and it is recognized that synaptic plasticity is more pervasive than just a compensatory mechanism. It is a subtle, ongoing process in the normal organism, responding to the environment and endogenous rhythms.

Normal Adult Plasticity

The nervous system is very dynamic, even down to the level of its synaptic connections. Proper mental function is based upon a dynamic organization of synaptic structure. The lability of these neuronal networks is highest in the phylogenetically young brain structures involved in realization of perception and self-awareness, which rely upon continual readjustment and refinement. This lifelong self-optimization process underlying neuronal complexity is remodeled to meet environmental demands. Experience-dependent learning combines both additive and subtractive processes. In response to an organism's attempt to control aspects of a new, complex environment there is the creation of new synaptic connections. Then, through the process of pruning, the brain selects the strongest neural circuits, fine-tuning the connections in a subtractive process. Thus experience-dependent learning couples sprouting with selective pruning, with a net gain of synapses. This constant, dynamic synaptic turnover makes it possible to adapt to complex, changing environments. Thus these synaptic connections are in a constant state of flux, responding to hormonal and environmental alterations with adaptive changes in neural circuitry and resulting behavioral alterations. However, the selective dynamic stabilization and destabilization of synaptic connections are accompanied by increasing inherent potential of failure and neuronal vulnerability during the aging process.

Neuronal plasticity mediated by sprouting can involve alteration in dendritic spine number and/or size or alterations in neuritic complexity (length and branching). Dendritic spines are micrometer-sized protrusions of the dendritic membrane that serve as the postsynaptic component for the vast majority of central nervous system excitatory synapses. Spine synapses, in contrast to shaft synapses, act to create spatially isolated compartments, confining biochemical signals and membrane trafficking to localized regions, allowing for fine-tuning of individual synaptic responses. Spine turnover and morphological changes in existing spines are important for the modulation of neuronal circuits during development and plasticity. Dendritic spines are highly dynamic in size, shape, and number, with highest density occurring during late development and decreasing to a relatively stable level throughout adulthood. Changes in spine number and shape are observed in response to high-frequency synaptic activity, behavioral stimuli, or endogenous hormonal cycles.

Further refinement of the neuronal circuitry can be imposed by the function of neighboring glial cells. Soluble glia-derived factors are also important in synaptogenesis, synapse maturation, and plasticity. Neuronal sprouting is dependent upon astrocyte-derived apolipoprotein E (apoE) and cholesterol, which complex into apoE-containing lipoproteins involved in membrane remodeling, repair, and lipid redistribution. Further, neuritic sprouting is impaired by activated glial cells and can be enhanced by reduction in the expression of glial fibrillary acidic protein (GFAP). Increased astrocyte contact with neurons and coverage

of neuronal surface area can modify synaptic activity by controlling local neurotransmitter concentrations and by spatially restricting neuron–neuron contact.

The structural plasticity described by synaptic plasticity is beginning to be linked to functional plasticity. Such functional plasticity can be seen as alterations of synaptic transmission that are strengthened or weakened in response to previous activity. These alterations in synaptic transmission can be of short duration, as in paired-pulse depression and posttetanic potentiation, or long-lasting, as in long-term depression and long-term potentiation (LTP). LTP, generated in the hippocampus, is thought to be a synaptic component of learning and memory, one form of behavioral plasticity. LTP is expressed as an increase in synaptic strength in response to a stimulus train, with increased amplitude being observed after 10–20 s and for many hours, or even up to days or weeks. Accompanying the change in synaptic transmission there is an expansion of the area of the synaptic junction. Although the number of synapses and the shape of dendritic spines do not change after the induction of LTP, stimuli that alter spine number or shape have been shown to also regulate LTP generation. Thus, the functional plasticity represented by LTP is tightly linked to structural changes associated with synaptic plasticity.

Synapse turnover is a constant process of loss and replacement of synapses, with total synapse number determined by the relative rates of synaptogenesis and synaptic regression. One set of stimuli may result in a net increase in synapse number by increasing synaptogenesis. Likewise, competing stimuli may induce synaptic regression, resulting in a net decline in synapse number. The subtle balance between stimuli may represent a mechanism underlying the synaptic plasticity observed as part of the normal life cycle, dependent upon developmental stage, experience, and age.

Neurons have the capacity to respond to perturbations to maintain their function within normal physiological range, such that the neuron can regulate the function of each impinging synaptic terminal, decreasing or increasing the strength of each synapse, resulting in a normal sum total excitation despite inappropriate innervation. Such homeostatic regulation of neuronal activity would maintain the robust function of the nervous system during the restructuring and refinement of neural circuits during the postembryonic developmental period. Such homeostatic regulation would establish limits beyond which activity-dependant changes in synaptic number would not reasonably alter cellular activity. This allows for maintenance of cellular physiological responses during the tumultuous periods of synaptic pruning required for establishment of synaptic structure and function. As the dendrites of the central nervous system, preexisting synaptic connections grow physically and chemically more removed from the cell soma, and more surface area is exposed for the inclusion of new synaptic contacts, leading to new functionality imposed by the changing response properties of the neuron.

The continual turnover of synapses, coupling sprouting with selective pruning underlying experience-dependent learning, allows for efficient memory storage in the brain, balancing the energy and metabolic demands of synapse maintenance with the extra energy cost of synaptic overgrowth required to obtain selective removal. Such a scheme allows for the judicious refinement of neural circuits required for proper memory storage within a highly adaptable context that is modifiable by endogenous and extrinsic environmental cues. However, the reliance of this system on exquisite balance between continual synaptogenesis and synaptic regression is vulnerable to maladaptive alterations in either side of the equation, leading to declines in cognitive function. The importance of correct pruning of the neural circuitry is observed in seizure patients. In the dentate gyrus, aberrant development of recurrent collaterals branching from granule cell axons (the mossy fibers) may contribute the scrambling of neural circuitry and to the epileptogenic activity in patients with temporal lobe epilepsy.

Puberty, Reproduction, and Sex Steroids in Synaptic Sprouting

Synaptic density peaks in the early postnatal period; this is followed by an extended time period of continuous synaptic turnover balanced toward synaptic removal or pruning, leading to a gradual decline to adult levels of synapses by puberty. Puberty is the point of transition, marking the metamorphosis of the child into the adult, with accompanying changes in reproductive maturity and alterations in body growth and composition, as well as profound changes in brain function. Not only does puberty mark the end of the extensive period of synaptic pruning that occurs during development, but it is also represented by the initiation of adult patterns of synaptic plasticity and sprouting. Neuronal connectivity is not fixed once development is complete but instead continues to change throughout life. However, there is a reduction in synaptic elimination, shifting the balance of the continuous synaptic turnover from a state of pruning to a state of maintenance.

Most dramatic are the alterations in synaptic connectivity in the hypothalamic centers governing sexual physiology and behavior. Here, steroid hormones both remodel and activate neuronal circuits during

adolescent brain development, establishing recurrent interactions between steroid hormones and the central nervous system. The release of gonadotropin-releasing hormone (GnRH) from GnRH neurons, a process that couples brain activity and gonadal function, is accomplished by these recurrent interactions, establishing a controlled cyclic gametogenesis and gonadal hormone production. In females, the midovulatory gonadotropin surge that stimulates ovulation is preceded by an estrogen-inducible retraction of axosomatic synapses in the arcuate nucleus. Following ovulation there is a recovery to baseline levels of synaptic connections, representing defined cyclical synaptic plasticity under control of sex hormones governing reproductive physiology. Due to the lack of alterations in spine or neurite morphology, these events are not sprouting *per se*; however, they remain reminiscent of the sex hormone-induced sprouting involved in the development of these same neural circuits.

Although the ovarian cycle-related synaptic changes in the hypothalamus do not involve neuritic outgrowth, in the hippocampus and prefrontal cortex there is an accompanying alteration in dendritic spine density. Specifically, there is significantly lower spine density during estrus than during proestrus. In addition, removal of circulating gonadal steroids via ovariectomy results in a significant decrease in spine density of the hippocampus in adult female rats, and spine density increases upon replacement of estrogen. This effect is reminiscent of the estrogen-induced sprouting during development of hypothalamic circuits underlying reproductive physiology and behavior.

Further supporting a role of sex hormones as environmental regulators of synaptic sprouting in the adult animal, gonadal steroids have been shown to be involved in regulating lesion-induced axonal sprouting. Although lesion-induced sprouting is more akin to pathological alterations in neuronal function, it is a useful paradigm to study the interaction of aging and environmental factors on normal aging-related synaptic plasticity and sprouting. These effects are often too subtle to isolate the mechanisms underlying normal synaptic turnover that are revealed by magnifying the effect via reactive regenerative processes. Sprouting is increased in the presence of sex hormones and in response to environmental enrichment or exercise *in vivo* following perforant path lesions of the hippocampus, and *in vitro* in the wounding-in-a dish model. The role of glial cells in the sprouting response is supported by the inverse relationship between astrocytic glial fibrillary acidic protein expression and neuritic sprouting. However, the influence of sex steroids on synaptic sprouting invites another point of vulnerability to the proper maintenance of the synaptic circuitry with the decline in hormonal control in aged animals, emphasized by the loss of sex hormone production in reproductive senescence, as discussed later.

In addition to the intrinsic regulation of synaptic density by gonadal hormones, extrinsic environmental cues can support sprouting as well. Environmental enrichment and exercise have been well established as one means of promoting neuritic sprouting and the incorporation of newly generated neurons into existing neuronal circuits. Exposing rodents to an enriched environment, consisting of large cages with toys ladders, mazes, and social interactions, results in increased performance on spatial learning tasks, compared with rodents housed in impoverished conditions. Similar results are obtained with increased physical exercise. These behavioral modifications are complemented by physical brain changes consisting of increased synaptic density and increased complexity of dendritic branching. While these adaptive responses are most pronounced in the young brain, environmental stimulation produces effects in the aging brain as well.

Neuronal Sprouting and Aging

Aging is associated with a decline in cognitive function that can be explained in part by alterations in cellular response directly affecting plasticity, in which the ability to learn new tasks decreases with age. On the cellular level, synaptic contacts, synaptic strength, and plasticity are reduced with age. These changes are much more subtle than are the dramatic alterations in neuronal morphology and survival that occur in age-associated neural disorders such as Alzheimer's disease and Parkinson's disease. In contrast to these pathological states, neuron loss does not appear to be an important contribution to age-related functional decline. Rather, subtle shifts in dendritic branching and spine density occur in region-specific patterns.

Although there is little evidence of significant neuronal loss during normal aging, alterations in dendritic extension of the neuronal soma occur in a large number of cells during aging. Many neurons show progressive restriction and atrophy of their more peripheral dendrite branches and, especially in cortical pyramidal cells, among the basilar shafts. The remaining dendritic branches often demonstrate beaded swellings, in correlation with the irregular dendritic spine loss. However, during these aging-related losses to dendritic systems, the potential for neuronal growth is not lost. Possibly as a compensatory response to increase available synaptic area, other neurons grow further dendritic extensions.

Even though there are these dendritic alterations in many neurons, they do not represent a gross regression of dendrites and remain region specific. There is

a reduction in dendritic branching with age in the prefrontal cortex in humans and rodents. In contrast, the hippocampus of humans and rodents does not experience significant changes in dendritic length with age, with net stabilization of dendritic extent in CA1, CA2, CA3, and the subiculum. In fact, in many brain regions, including the parahippocampal gyrus and the dentate gyrus, the aging brain is associated with increased dendritic branching, until very old age, when there is a consistent regression of the dendritic tree back to that of mature young adults. The region specificity of the dendritic losses indicates that the declines in neuronal plasticity are not merely consequences of wear and tear or global aging. This variability in aging that occurs within subpopulations of neurons reflects the different demands throughout the life span.

The observed region-specific increases in dendritic plasticity may be a compensatory response to partial cell loss that occurs during normal aging. New connections would be formed by healthy neurons within the population, assuming parallel functions. Such an extension of existing inputs would maintain circuit function, but would limit the amount of redundancy in the system, making it more vulnerable to further alterations in synaptic connectivity. Alternatively, the new connections could be formed by fibers from converging pathways that can act to boost weakened signals and maintain functional stability. This response may be a recapitulation of the developmental environment in which many neurons are programmed to form a certain number of terminals or to synapse, and if they cannot do so in one region, they will tend to increase terminal growth elsewhere. The increase in available synaptic area in response to the progressive regression of neighboring dendritic systems works to maintain the total postsynaptic surface per neuron. This leads to a concept of two neuronal responses to aging, one involving dendritic retraction and one involving reactive dendritic expansion. This two-stage response of dendritic spines (loss of total number but enhanced sprouting of remaining spines) represents a neuronal response to the decreased allostatic load, defined as a decrease in the capacity of neurons to oppose the damaging effects of strong, excessive stressors, compatible with full function at normal levels of load. The decreased allostatic load, or homeostatic reserve, is a characteristic of normal aging, resulting from continual or repeated stress.

Even though the structural and behavioral plasticities associated with sprouting are present in the normally aged brain, functional decline and its associated synapse loss are still observed. The aged rodent brain has a remarkable capacity for sprouting and synaptogenesis. However, aged synapses respond to plasticity-inducing stimuli differently than do young synapses, and the degree of sprouting is often blunted with respect to young animals. Further, even though in many brain regions the absolute number of spines is often not altered, there are significant changes in the shape and distribution of synaptic contacts. With increasing age there is a shift from L-type (lollipop shape) to N-type (nubby-like shape) synapses. The spine head becoming smaller, the spine shaft thickening, and a decrease in spine length characterize this shift. Cognitive-sparing diet alterations have been demonstrated to prevent the loss in L-type synapses. The aging-related shift in spine makeup and response is demonstrated by the interaction between vulnerability to aging and sex steroids in female rats. The estrogen-induced spine increase seen in CA1 of young rats does not occur in response to estrogen in aged rats. However, in the aged rat, estrogen increases the number of *N*-methyl-D-aspartate (NMDA) receptor per synapse, restoring the more youthful receptor profile that is lost in normal aging. Thus estrogen may help preserve hippocampal function in the context of a decreased synaptic density. This is another case in which the aged response to environmental stimuli maintains the circuit function, but with a decrease in the system's redundancy. This alteration in estrogen responsiveness is species specific. In contrast to the rodent, the CA1 region of the aged monkey hippocampus is still as responsive to estrogen-induced spine density and synapse number increases as is the young monkey.

Many of the other age-related changes in neuronal plasticity are generally subtle and may only be observationally manifested under conditions of perturbation. This was first observed in 1978 by Scheff and colleagues, who demonstrated that there is a clear reduction in collateral sprouting and regeneration in the aged nervous system. This limited response of reactive sprouting in aged animals has been well characterized by many groups since. Following entorhinal lesion, both young and aged rats can replace synapses lost in the dentate molecular layer of the hippocampus. However, this reinnervation, which begins very rapidly in young animals, is delayed in aged animals, with impairment in the initial phases of synaptic sprouting. Thus the rate of reactive sprouting is decreased as a function of age.

The age-related alterations in neuronal plasticity are emphasized again with the interplay between sex steroids and reactive sprouting. As discussed earlier, estrogen greatly enhances neurite outgrowth in various models of sprouting. However, in the aged female rat the induction of neuronal sprouting is impaired and, unlike in young adults, the sprouting is not sensitive to estradiol. This effect may be due to the increased glial activation associated with aging,

as evidenced by an increase in the activation marker GFAP. GFAP expression is usually regarded as secondary to neurodegenerative processes. However, glia become activated early in aging without concurrent clinical manifestations of pathology. Since neuritic sprouting is inversely related to GFAP expression, normal aging-related glial activation may have a profound impact on synaptic functions. Increased glial activation and GFAP expression may shift the balance of sprouting and retraction unfavorably during the ongoing synaptic turnover, leading to age-related declines in cognitive function.

Age-related alterations in synaptic plasticity may be due to changes in the properties of individual neurons or due to changes in the nervous system itself, as the individual neurons exist within a highly organized system. The decrements in neuronal plasticity could be due to inherent changes in the neuron's ability to extend neurites or form new spines. Alternatively, the decrement could be due to alterations in the complex system of extrinsic signals that trigger the requisite outgrowth. However, there has not been a consistent observation of age-related changes in neurotrophic factors to support the latter scenario. Another alternative to intrinsic changes to the aged neuron diminishing its capacity for growth involves alterations in the target neural circuitry in a way that lessens the accommodation or stimulation of sprouting. This is a model supported by the fact that the aged hippocampus supports less neuron ingrowth than does young hippocampus, whereas aged transplants show robust innervation in young host hippocampal tissue. This is an effect that does not appear to be due to reductions in baseline substrate properties that promote neurite outgrowth. Further support is provided by the observation that sprouting in the wound-in-a-dish model is diminished when neurons are co-cultured with astrocytes from aged animals, as compared to astrocytes from young animals. Thus, in the target region preventing efficient sprouting, there are likely changes that are marked by increased glial activation. This increased glial activation could result from the increase in allostatic load imposed by the results of cumulative reactive oxygen species (ROS) production or excessive stress. Both increased oxidative load and stress (or its hormonal effector, glucocorticoid) have been shown to reduce neuronal plasticity, LTP, and behavioral learning. Further implicating oxidative load in aging-associated cognitive decline, spine losses can be attenuated by superoxide dismutase overexpression (antioxidant defense).

In addition to reduced oxidative load and hormonal modification, synaptic sprouting can be altered favorably by other means. Calorically restricted diets can retard synapse loss, mostly through maintenance of L-type synapses, rather than preventing a decrease in total spine number. The processes of synaptic plasticity appear to be strengthened by increased neuronal activity, driven by enriched environment or behaviors such as exercise. Such age-associated dendritic morphologies are also reflected in the region-specific alterations in spine density. This continued structural plasticity of the aged brain is reflected behaviorally, in that age-related loss in speed of mental processing can be offset by continual environmental enrichment and challenge so as to promote synaptic plasticity. The fact that aging-vulnerable circuits are still responsive to environmental stimuli indicates that the aged synapse remains plastic. This is in contrast to cognitive deficits due to neuronal loss, in which case little can be done to restore circuit function. If the cognitive deficit of normal aging is due to the frailty of the synapse, the mechanisms of synaptic plasticity might be harnessed to for therapeutic intervention.

See also: Adult Cortical Plasticity; Aging of the Brain; Axonal and Dendritic Identity and Structure: Control of; Cognition in Aging and Age-Related Disease; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Gene Expression Regulation: Steroid Hormone Effects; Neuronal Plasticity after Cortical Damage; Spine Plasticity; Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging.

Further Reading

Bennett EL, Diamond MC, Krech D, et al. (1964) Chemical and anatomical plasticity in brain. *Science* 146: 610–619.

Bertoni-Freddari C, Fattoretti P, Paoloni R, et al. (1996) Synaptic structural dynamics and aging. *Gerontology* 42(3): 170–180.

Buell SJ and Coleman PD (1979) Dendritic growth in the aged human brain and failure of growth in senile dementia. *Science* 206(4420): 854–856.

Burke SN and Barnes CA (2006) Neural plasticity in the ageing brain. *Nature Reviews Neuroscience* 7(1): 30–40.

Calabrese B, Wilson MS, and Halpain S (2006) Development and regulation of dendritic spine synapses. *Physiology* 21: 38–47.

Crutcher KA (2002) Aging and neuronal plasticity: Lessons from a model. *Autonomic Neuroscience* 96(1): 25–32.

Finch CE (2003) Neurons, glia, and plasticity in normal brain aging. *Neurobiology of Aging* 24(supplement 1): S123–S127; discussion S131.

Foster TC (2002) Regulation of synaptic plasticity in memory and memory decline with aging. *Progress in Brain Research* 138: 283–303.

Hof PR and Morrison JH (2004) The aging brain: Morphomolecular senescence of cortical circuits. *Trends in Neurosciences* 27(10): 607–613.

Serrano F and Klann E (2004) Reactive oxygen species and synaptic plasticity in the aging hippocampus. *Ageing Research Reviews* 3(4): 431–443.

Toescu EC (2005) Normal brain ageing: Models and mechanisms. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 360(1464): 2347–2354.

Synaptic Plasticity: Neuronogenesis and Stem Cells in Normal Brain Aging

G M Bernal and D A Peterson, Chicago Medical School at Rosalind Franklin University of Medicine and Science, North Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Although the postnatal brain had long been thought to lack the capacity to generate new neurons, the past decade's research has radically changed this assumption with ample evidence that new neurons are generated in some regions of the adult brain. This process is referred to as neuronogenesis for the specific production of neurons and more broadly as neurogenesis when all neural cell types may be generated. In many cases the terms neuronogenesis and neurogenesis are used interchangeably, as they will in this article. Neural stem cells and neurogenesis are restricted to specific areas in the adult brain: the subventricular zone/rostral migratory stream/olfactory bulb system and the dentate gyrus of the hippocampus. This spatial restriction means that most of the brain does not support neurogenesis under normal conditions and is thus referred to as nonneurogenic. Understanding the capacity of different regions to support survival and integration of neural stem cells will be of importance for the potential use of stem cells for brain repair, as it is this wider, nonneurogenic portion of the brain that will most frequently need to be the target of repair strategies.

Within the young adult brain, neurogenesis routinely contributes large numbers of new neurons to the olfactory bulb and small numbers of new neurons to the dentate gyrus. Many studies have shown that the rate of neurogenesis in the young adult brain is not fixed, but subject to stimulation by experimental manipulation. Neuropathological studies of the aging central nervous system (CNS) frequently utilize young adult, rather than aged, animals in experimental models or transgenic mutations of senescence-associated diseases, such as Parkinson's, Alzheimer's, or Huntington's diseases. Less is known about the underlying changes leading to the decline in normal (nonpathological) brain aging, yet this characterization will be necessary to interpret the contribution of specific disease pathology to impaired brain function. Here we look at the effects of normal aging on neurogenesis in the adult mammalian brain, the process in which resident neural stem cells give rise to new functional neurons that are incorporated into the circuitry. We discuss how age-induced changes may modulate neural stem cell activity resulting in a significant decline in neurogenesis and the implications of this for the aging brain.

Characterization of Neural Stem Cells

Many postnatal mammalian tissue systems contain cells that possess the features consistent with the definition of a stem cell, that is, the capacity for unlimited self-renewal and the ability of their progeny to differentiate into different cellular types. Stem cells present during embryonic development have considerable flexibility in sculpting tissues; however, this capacity may be restricted in postnatal tissue. In fact, it is a matter of debate as to whether stem cells found in certain adult tissues, such as the CNS, are truly stem cells or would be better termed progenitor cells, reflecting their more restricted proliferate capacity and differentiation potential to the same tissue in which they reside.

At present, there is no single definitive marker by which neural stem cells in tissue can be identified. Identification of stem/progenitor cells in adult tissue is usually accomplished through labeling the proliferating cells and assessing their subsequent lineage outcome. Most *in vivo* studies of neurogenesis have used administration of the thymidine analog bromodeoxyuridine (BrdU) to identify new cells. The duration of BrdU delivery and the interval between administration and examination are both of critical importance in interpreting the cell type under study. For example, sustained delivery of BrdU over several days labels a population of proliferating cells, but obscures the ability to distinguish whether those proliferating cells are slowly dividing or rapidly dividing, or whether the labeled cohort contains a heterogeneous population of both cell types. Therefore, many studies with longer BrdU administration do not provide information discriminating between slowly cycling stem cells and rapidly cycling progenitor cells. The ability to ascertain the frequency of reentry into cell cycle is of importance for understanding the contribution of this parameter to age-related changes in neurogenesis.

Another important parameter to establish when assessing neurogenesis is to extend beyond the quantification of changes in BrdU-positive cell number by also assessing their fate to determine the frequency at which they turn into neurons. As no single definitive marker exists for neural stem cells, identification of progression through lineage commitment is accomplished by combinatorial expression of various

markers. There has been considerable progress in defining the sequential expression of lineage markers in the young adult CNS, but our knowledge of whether the same sequence of markers and duration of their expression are invariably expressed in the aging brain is less than complete. In addition, parameters of neurogenesis differ between mice and rats, suggesting that species differences should be kept in mind when extrapolating results from animal studies.

Neurogenesis Is Decreased with Aging

Neurogenesis does persist in the aged brain, including that of elderly humans. Quantitative studies evaluating age-related changes have consistently reported lower levels of neurogenesis in both the subventricular zone and the dentate gyrus that become statistically significant by mid-age. In aged dentate gyrus of various rat or mouse strains, neurogenesis has been reported to be at the level of 10–20% of young adult animals. Olfactory bulb neurogenesis is significantly reduced to approximately 30% of neuronal production in the young brain and is accompanied by decreased proliferation in both the subventricular zone and the rostral migratory stream to approximately 20–40% of levels in the young brain. The decline of hippocampal neurogenesis is accompanied by a decrease in cell proliferation in addition to a decline in the number of cells expressing doublecortin, a marker associated with early neuronal lineage commitment. However, despite the age-related reduction in neurogenesis, the capacity of newly generated cells in the aged brain to migrate, survive, and differentiate into mature neurons is maintained. Furthermore, early work demonstrated that the age-related decline in dentate neurogenesis could be partially reversed in some conditions, such as environmental enrichment or alteration of glucocorticoid levels. Those reports suggest that neurogenesis in the aged brain is not absolutely diminished, but there remains a capacity to elevate neurogenesis, at least in the hippocampus.

Despite the reduced proliferation of new cells in the aged hippocampus, the proportion of those newly generated cells that differentiate into neurons is comparable to the young hippocampus. One interpretation of these data was that there are fewer neural stem cells in the aged brain generating the basal level of cell production. Studies using BrdU labeling of cell proliferation could not address this possibility, as they fail to discriminate between slowly dividing stem cells and rapidly dividing progenitor cells. However, recent work assessing expression of a stem cell marker, Sox2, in combination with other lineage markers indicates that the subpopulation of Sox2-positive cells that correspond most closely to the likely neural stem cell population is not reduced in the aging brain. These results suggest that the age-related reduction in neurogenesis is not due to there being fewer neural stem cells, but rather that these cells may be more quiescent in the aged CNS. In addition to quiescent stem cells producing fewer new cells in the aged brain, the maturation of newly generated neurons appears delayed and the extent of their dendritic arborization is less in the aged hippocampus.

Modulation of Neurogenesis in the Aged Brain

The observed age-related decrease in neurogenesis may result from a combination of environmental changes and cell-intrinsic limitations with increasing age. The quiescence of neural stem cells in the aged hippocampus may be a cell-autonomous property, but proliferation in aged hippocampus can be stimulated under certain conditions. Thus, there is substantial evidence that environmental factors play an important role in the age-related reduction in neurogenesis.

Overexpression of trophic factors has been shown to enhance neurogenesis in the young brain. Protein levels for some of the key trophic factors (such as brain-derived neurotrophic factor (BDNF), fibroblast growth factor-2 (FGF-2), insulin-like growth factor 1 (IGF-1), and vascular endothelial growth factor (VEGF)) promoting neural stem cell proliferation, differentiation, and survival show a decline as early as mid-age, correlating with the observed decrease in the proliferation of stem cells and neurogenesis in the aged brain. These changes may be accounted for by alterations in gene expression and/or posttranslational modification. Enhanced availability of trophic factors has been explored as an attractive target for therapeutic manipulation to promote neurogenesis in the young brain. Increased neurogenesis in the young brain has been reported following delivery of BDNF, FGF-2, IGF-1, VEGF, and epidermal growth factor (EGF). The aged brain, likewise, can be stimulated to increase cell proliferation and possibly neurogenesis following delivery of FGF-2, BDNF, IGF-1, and heparin-binding EGF (HB-EGF). However, in neurogenic regions, any enhancement of neurogenesis appears to be modest and does not restore neurogenesis to the level seen in the young brain. These results suggest that the decline of trophic factor support alone may not be responsible for age-related changes.

Trophic support of cell proliferation and differentiation may be regulated by the availability of both ligand and the receptors to bind to those ligands. Changes in the expression of trophic factor receptors, either on the neural stem cells themselves or on other cells in the neurogenic niche, may contribute to reduced neurogenesis in the aged brain. Relatively little is known, at present, about possible changes in

receptor expression on neural stem cells themselves, but there are indications that alteration of receptor expression occurs in other types of cells residing in the neurogenic niche, suggesting possible secondary effects on neural stem cell proliferation. Age-related changes in receptor expression would support possible cell-autonomous mechanisms for modulating neurogenesis through downstream alteration of the local environment. For example, there is reduced expression of FGFR-2 on astrocytes of the aged neurogenic niche, which may reflect an impairment of these cells to respond to FGF-2 signaling and/or in their autocrine regulation to produce FGF-2. This may, in turn, reduce the paracrine effect of FGF-2 on neural stem cells in the aged neurogenic niche.

Structural plasticity of dendrites and spines in the aging hippocampus can be significantly modified by alteration of hormone levels. Neurogenesis may also be modulated by other factors such as hormones and inflammatory molecules that show age-related changes. Despite the strong evidence for age-related hormone-induced plasticity in the hippocampus, relatively few studies have directly examined modulation of neurogenesis in this context (Table 1). Changes in glucocorticoid regulation do appear to influence neurogenesis in the aging brain. Stress has a potent effect on neurogenesis in both the young and aged brain, although the precise linkage with glucocorticoid as the proximal mechanism of that change is a matter of debate. Unfortunately, few studies on the important estrogen-related changes in the aging brain have investigated neurogenesis. There is evidence that neurosteroids, such as allopregnanolone, can promote neurogenesis in an Alzheimer's disease (AD) mouse model, although these mice were still relatively young. The role of hormonal modulation is an active area of investigation with important clinical implications, and assessment of neurogenesis is likely to be incorporated into future reports.

Implications of Decreased Neurogenesis in the Aging Brain

While numerous studies have demonstrated that neurogenesis can be enhanced, the advancement of our understanding about the biology regulating neurogenesis has been hampered by an inability to experimentally block neurogenesis. In this regard, studying age-related neurogenesis provides a natural model of reduction that may reveal the key regulatory elements for sustaining the generation of new neurons.

Table 1 Reported effects of hormonal alteration on neural stem cell proliferation and neurogenesis in the hippocampal dentate gyrus

Hormone	*Natural age-related change*	*Experimental manipulation of hormonal levels*	*Outcomes*
Corticosterone	Increased	*Rodents*	
		Intact young (exogenous corticosterone)	Decreased neurogenesis
		Intact young (following various stressors)	Decreased cell proliferation and/or neurogenesis
		Intact aged	Decreased neurogenesis
		Intact aged (following psychosocial stress)	Additional decrease in neurogenesis
		Adrenalectomy (decreased corticosterone)	
		Early postnatal; tested young and aged	No effect on neurogenesis at either age
		Young	Increased neurogenesis
		Mid-aged- tested when aged	Increased neurogenesis
		Aged	Increased neurogenesis (reversed by corticosterone)
		Primates	
		Young adult (following psychosocial stress)	Decreased cell proliferation
Gonadal hormones (estrogen)	Decreased in females	*Young female rodents*	
		Ovariectomy (OVX; results in decreased estrogen):	
		OVX	Decreased cell proliferation
		OVX + single estradiol administration	At 1 week, increased cell proliferation At 4 weeks, no effect on cell proliferation
		OVX + single estradiol and progesterone administration	No effect on cell proliferation
		OVX + chronic or cyclic estradiol administration	No effect on cell proliferation
		Aged female rodents	
		OVX ± chronic or cyclic estradiol administration	No effect on neurogenesis reported
		Aged female primates	
		OVX ± chronic or cyclic estradiol administration	No effect on neurogenesis reported

While individual new neurons are generated in the adult brain that assume functional characteristics of mature neurons, the extent to which their integration in a larger neuronal network directly contributes to a functional outcome remains unclear. Numerous studies have sought to establish a link between the generation and addition of new neurons in the olfactory bulb and hippocampus and functional output from those systems. With its natural reduction in neurogenesis, the study of aging animals may provide one approach to addressing the question of function. Unfortunately, studies attempting to correlate the level of impairment in hippocampal-dependent learning tasks and the extent of age-related reduction in neurogenesis have yielded mixed results. Some results have suggested that the aged animals most impaired in behavioral performance had fewer new hippocampal neurons, and other studies have indicated that animals in this group had the highest levels of neurogenesis. However, stimulation of hippocampal neurogenesis in aged animals, such as enriched environment or running, has been reported to correlate with improved performance in hippocampal-dependent tasks. It remains to be demonstrated if this correlation of stimulated neurogenesis and improved performance is causal (i.e., due to increased neurogenesis) or if it is a secondary effect to other biological changes, such as improved blood flow, elevation of circulating cytokines, etc. Although models of hippocampal-dependent learning are relevant to assessing function of hippocampal neurogenesis, there is no evidence to date suggesting that neurogenesis plays a role in age-related mild cognitive impairment in humans.

The incidence of anosmia (decreased or absent sense of smell) is increased in elderly patients with attendant risks for malnutrition and food poisoning. Newly generated olfactory bulb neurons in young animals have been shown to integrate and possibly participate in processing of olfactory input, and it had been postulated that the age-related reduction of olfactory bulb neurogenesis may contribute to the development of anosmia. However, it has been reported recently that humans lack a defined rostral migratory stream, making the extent to which stem cells in the subventricular zone contribute to olfactory bulb neurogenesis in humans unclear. This report also creates uncertainty as to the extent to which age-reduced neurogenesis may contribute to anosmia in elderly humans.

The aging brain is vulnerable to a variety of insults and neurodegenerative disease. With our understanding that neurogenesis persists in the adult and aging brain, there has been enthusiasm for utilizing neural stem cells for structural brain repair, where dysfunctional or dead neurons are replaced by grafted neural stem cells. Much of the work in grafting of neural stem cells has been performed in young animal models. It will be important to assess the physiological changes in the aged brain that may constrain the ability of grafted neural stem cells to survive, differentiate, and integrate into the aged brain. In particular, more information is needed about the ability of nonneurogenic regions of the aged brain to support neural stem cell survival and differentiation, as it is primarily the nonneurogenic regions that will be the target of stem cell therapy. Establishment of these parameters will be critical for discriminating between nonpathological aging and the contribution of pathology in animal models of neurodegenerative disease. Further study of the regulation of neurogenesis in the aged brain should provide useful insights into the development of therapeutic strategies for repairing the aged brain.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Neural Stem Cells and CNS Diseases; Neural Stem Cells: Adult Neurogenesis; Neurogenesis in the Intact Adult Brain; Stem Cells and CNS Repair; Synaptic Plasticity: Neuronal Sprouting; Synaptic Plasticity: Short-Term Mechanisms.

Further Reading

Bernal GM and Peterson DA (2004) Neural stem cells as therapeutic agents for age-related brain repair. *Aging Cell* 3: 345–351.

Carson CT, Aigner S, and Gage FH (2006) Stem cells: The good, bad and barely in control. *Nature Medicine* 12: 1237–1238.

Eriksson PS, Perfilieva E, Bjork-Eriksson T, et al. (1998) Neurogenesis in the adult human hippocampus. *Nature Medicines* 4: 1313–1317.

Hattiangady B and Shetty AK (2006) Aging does not alter the number or phenotype of putative stem/progenitor cells in the neurogenic region of the hippocampus. *Neurobiology of Aging* Nov 6 [Epub ahead of print].

Jin K, Sun Y, Xie L, et al. (2003) Neurogenesis and aging: FGF-2 and HB-EGF restore neurogenesis in hippocampus and subventricular zone of aged mice. *Aging Cell* 2: 175–183.

Kempermann G, Kuhn HG, and Gage FH (1998) Experience-induced neurogenesis in the senescent dentate gyrus. *Journal of Neuroscience* 18: 3206–3212.

Rao MS, Hattiangady B, Abdel-Rahman A, Stanley DP, and Shetty AK (2005) Newly born cells in the ageing dentate gyrus display normal migration, survival and neuronal fate choice but endure retarded early maturation. *European Journal of Neuroscience* 21: 464–476.

Shetty AK, Hattiangady B, and Shetty GA (2005) Stem/progenitor cell proliferation factors FGF-2, IGF-1, and VEGF exhibit early decline during the course of aging in the hippocampus: Role of astrocytes. *Glia* 51: 173–186.

van Praag H, Shubert T, Zhao C, and Gage FH (2005) Exercise enhances learning and hippocampal neurogenesis in aged mice. *Journal of Neuroscience* 25: 8680–8685.

Synaptic Plasticity: Short-Term Mechanisms

J S Dittman, Weill Cornell Medical College, New York, NY, USA
A C Kreitzer, University of California at San Francisco, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

One of the fascinating features of chemical synapses is their capacity to alter their output based on past history. This temporal plasticity is thought to contribute to the brain's faculty for computation, learning, and memory. Short-term synaptic plasticity (STSP) occurs over a range of milliseconds to minutes and involves temporary enhancement or depression of synaptic strength. The molecular mechanisms that underlie STSP involve regulation of neurotransmitter release (presynaptic) and changes in sensitivity to neurotransmitter (postsynaptic). This article focuses on the mechanisms and molecules that underlie presynaptic STSP.

After an action potential (or series of action potentials) invades a presynaptic terminal and triggers calcium-dependent fusion of one or more synaptic vesicles (SVs), the likelihood of releasing neurotransmitter in response to subsequent action potentials may either increase or decrease. These increases and decreases in likelihood are referred to as short-term enhancement (STE) and short-term depression (STD), respectively. Historically, STE has been divided into several distinct processes, such as paired-pulse facilitation (PPF), augmentation, and posttetanic potentiation (PTP), based on kinetics and pharmacology. These forms of plasticity are in large part driven by the residual intracellular calcium that accumulates in presynaptic terminals following depolarization. Presynaptic STD may simply reflect a transient depletion of available SVs, although there is evidence for a refractory period during which SVs cannot fuse. Recovery from STD is regulated by residual calcium and therefore represents another type of STSP.

The molecular processes that account for presynaptic STSP are only just beginning to be revealed as experimental synaptic preparations are becoming more amenable to genetic perturbations and imaging techniques. Hundreds of proteins play a role in coordinating neurotransmitter release, and mutations in some of these proteins can result in changes in both basal synaptic transmission and STSP. It is difficult to ascribe specific contributions of these proteins solely to STSP because any change in synaptic transmission can indirectly affect STSP as well. For instance, many perturbations that raise the probability of transmitter release will concomitantly decrease the amount of PPF or PTP simply because there is less room for further increase in release probability. These changes may incorrectly be interpreted as playing a crucial role in STE when in fact they may be entirely independent of facilitation. However, most of the proteins that coordinate SV docking and priming with calcium-triggered fusion have multiple interactions with other proteins and lipids, and removal of any one of these will likely affect numerous processes involved in both basal synaptic transmission and presynaptic STSP. Thus, the selective removal of a particular form of plasticity without altering basal transmission may be the exception to the rule when analyzing synaptic mutations. Despite these limitations, several molecules have been implicated in STSP and regulation of the SV cycle.

Activity-Dependent Regulation of Neurotransmitter Release

The centerpiece of presynaptic function is the SV cycle, which is composed of vesicle biogenesis and trafficking to the plasma membrane, docking and priming steps which determine the SVs that are competent for fusion, calcium-dependent fusion of the SV, retrieval of SV membrane and proteins, and, finally, refilling with neurotransmitter and reentry into the SV cycle. A synaptic varicosity typically contains hundreds of SVs, but only a small number are competent to fuse with the plasma membrane at any particular time. This subset of SVs is called the readily releasable pool (RRP), and the size of the RRP may be a key determinant of synaptic strength. Processes that modulate the RRP are therefore wellsuited to account for various forms of presynaptic STSP.

Active Zone Proteins

SVs fuse at a specialized site within the synapse known as the active zone (AZ) which is composed of a variety of protein families, including RIMs, Munc13s, ELKS proteins, Liprin-αs, Piccolo, and Bassoon. Munc13-1 and -2 are large multidomain cytoplasmic proteins that play an important role in the priming of SVs in part by facilitating the formation of the ternary SNARE complex (synaptobrevin on SVs, and syntaxin and SNAP-25 on the plasma membrane) which brings SVs in close apposition with the plasma membrane. In addition, Munc13s appear to serve other functions in supporting SV fusion. They interact

with RIMs, syntaxin, spectrin, the ARF-GEF Msec7-1, diacyl glycerol (DAG), and calmodulin.

The relative abundance of Munc13-1 versus Munc13-2 may determine whether a synapse expresses STE or STD when challenged with a high-frequency stimulus. Residual calcium bound to calmodulin has been shown to interact with a conserved region near the N-terminus of Munc13s. This interaction appears to increase STE (Munc13-2) and recovery from STD (Munc13-1). In both of these cases, the underlying mechanism of Munc13-dependent enhancement of release is thought to be increased SV priming. The details of this process are not fully understood but may involve interactions with the plasma membrane SNARE protein, syntaxin 1. Syntaxin 1 can exist in a 'closed' conformation which cannot participate in forming complexes with the vesicle SNARE, synaptobrevin. Munc13 interactions may convert syntaxin 1 into an 'open' or permissive conformation thereby facilitating a complete SNARE complex formation required for maturation of the SV to a release-ready state. There is also evidence that Munc13s facilitate vesicle fusion at an additional step downstream from open syntaxin. Phorbol esters are known to enhance transmitter release in large part through their interaction with C1 domain-containing proteins such as protein kinase C (PKC) and Munc13s. This protein–lipid interaction is thought to be a pharmacological mimetic of DAG–C1 interactions occurring *in vivo* to recruit C1 domains during signaling. Thus, local DAG production near AZs may recruit and activate Munc13s, which in turn increase SV priming and release probability.

NCS-1/Frequenin

Another molecule implicated in facilitation of neurotransmitter release is NCS-1 (also known as frequenin). NCS-1 is a small cytoplasmic protein possessing an EF-hand domain that binds calcium. This protein localizes to presynaptic terminals and enhances transmitter release when overexpressed. In hippocampal cultures, overexpression of NCS-1 converts the form of STSP from depression to facilitation without altering basal release probability, perhaps by mobilizing additional SVs or activating silent release sites. This observation implies that at least in some cases, a form of STSP can be separated from the core process of synaptic transmission. Interactions with phophatidylinositol 4-kinase and extracellular signal-related kinase (ERK1/2) appear to be important for the function of NCS-1. These interactions may connect residual calcium with lipid signaling, IP3-gated intracellular calcium stores, small GTPases such as Ras, and the MAP kinase pathway, which have been implicated in STSP in *Aplysia* and mice. There is also evidence that NCS-1 can affect P/Q-type voltage-dependent calcium channels, which in turn provide the calcium that drives SV fusion. Thus, NCS-1 appears to have a central role in coordinating calcium increases with a variety of intracellular signaling pathways that modulate transmitter release.

SNARE Regulation

Interactions between vesicular and plasma membrane SNAREs play a critical role in determining both where an SV can fuse and how likely it is to fuse. Furthermore, after fusion, plasma membrane ternary SNARE complexes must be separated and sorted in order to provide free synaptobrevin upon endocytosis. Several molecules have been suggested to inhibit synaptic transmission via inhibition of SNARE complexes and thus provide a means of modulating synaptic transmission. The transmembrane vesicular protein synaptotagmin IV (SytIV) can act to inhibit SV fusion by direct interaction with SNAREs, perhaps by competing with synaptotagmin I (SytI). SytI and SytIV are members of a family of transmembrane proteins containing tandem C2 domains (C2 domains are Ca^{2+}-dependent membrane-targeting protein modules). SytI is required for normal calcium-triggered SV fusion and believed to be the primary 'sensor' of the calcium that triggers synchronous release by acting on the SNARE complex. Presynaptic expression of SytIV can be rapidly induced in particular brain regions of rodents, and increased SytIV decreases basal synaptic transmission and leads to enhanced PPF and PTP. This inhibition is hypothesized to result from SytIV competing with SytI for SNARE binding. Thus, SytIV can modulate synaptic transmission by inhibiting normal SNARE function.

Another molecule which is proposed to inhibit SNARE function is the syntaxin-interacting protein tomosyn. Tomosyn is a large cytoplasmic protein that contains a SNARE motif at its C-terminus. By binding to syntaxin and SNAP-25 in place of synaptobrevin, tomosyn prevents assembly of the ternary SNARE complex. Regulation of tomosyn function would therefore be predicted to modulate the priming of SVs. Overexpression of tomosyn in chromaffin cells and pancreatic β cells inhibits fusion of large dense-core vesicles. In *Caenorhabditis elegans*, removal of tomosyn increases the size of the RRP and prolongs the decay time of evoked EPSCs. Protein kinase A (PKA) has been shown to phosphorylate tomosyn,

decreasing its affinity for syntaxin. These observations suggest that tomosyn is a critical regulator of SNARE complex formation and may be an important target of PKA-mediated enhancement of synaptic transmission.

SNAP-25 has been studied as a regulator of facilitation. In mammals, there are two splice variants of SNAP-25 (SNAP-25a and -b), and their expression changes during development such that there is a shift from SNAP-25a to SNAP-25b. Artificial sustainment of high levels of SNAP-25a results in an enhancement of PPF without altering basal transmission. SNAP-25 has also been shown to be a target of PKC, and blockade of this phosphorylation appears to inhibit SNARE assembly. The isoform and phosphorylation state of one of the SNAREs therefore contributes to STSP in addition to its role as a core component of SV fusion.

Finally, recycling of SNAREs may also be an important site of SV regulation. SNAP-29 is a small cytoplasmic protein that binds to plasma membrane ternary SNARE complexes and prevents disassembly by competing with α-SNAP. Inhibition of disassembly ultimately may slow the rate of recovery from STD. Indeed, decreased SNAP-29 levels have been found to correlate with increased recovery from STD, whereas increased SNAP-29 slowed recovery from STD, and this slowing could be reversed by overexpression of α-SNAP.

Kinase Pathways

Signaling complexes associated with receptor tyrosine kinases (RTKs) have been proposed to regulate synaptic strength and STSP. In particular, the brain-derived neurotropic factor receptor TrkB is complexed with an actin-based motor protein myosin VI (Myo6) and a Myo6-interacting protein GIPC1. Loss of either Myo6 or GIPC1 results in decreased basal synaptic transmission and increased PPF, whereas PTP is decreased. The Abl family nonreceptor tyrosine kinases Abl and Arg have also been implicated in STSP. *abl*−/− and *arg*−/− mice show normal basal synaptic transmission and PTP, whereas PPF is selectively reduced. Since RTKs and non-RTKs such as Abl and Arg will be unlikely to phosphorylate their targets on the rapid timescale of facilitation, it is more likely that loss of these kinases alters the function of many synaptic proteins that require constitutive phosphorylation for optimal function. Other kinases, such as CaMKII, PKA, and PKC, may also contribute to STSP, and it will be important to identify the phosphorylation states of the molecules that regulate synaptic transmission so that the pleiotropic effects of disrupting kinases and phosphatases can be interpreted in terms of their net effects on critical downstream targets.

G-Protein-Mediated Regulation of Neurotransmitter Release

Short-term plasticity of neurotransmitter release is not only triggered by activity-dependent changes within the terminal itself but also can be induced by activation of presynaptic G-protein-coupled receptors (GPCRs). GPCRs represent a large family of seven transmembrane domain receptor proteins that, when activated by a ligand, function as guanine nucleotide exchange factors (GEFs) for heterotrimeric G-proteins. This GEF activity leads to the replacement of bound GDP by GTP on Gα subunits, resulting in the dissociation of the α and β/γ subunits and their subsequent interaction with effector molecules. G-protein signaling is terminated by the hydrolysis of α-bound GTP to GDP and the reassociation of α and β/γ subunits. Although more than 300 different types of GPCRs have been described, they can be generally classified by their interactions with different types of G-proteins.

$G_{i/o}$-Coupled Receptors

Presynaptic modulation of neurotransmitter release is best characterized for $G_{i/o}$-coupled receptors, which include group II and III metabotropic glutamate receptors (mGluRs), $GABA_B$ receptors, adenosine A1 receptors, cannabinoid CB1 receptors, opioid receptors, and muscarinic M2 and M4 receptors. These GPCRs are widely expressed and are often found presynaptically adjacent to active zones. They can be activated through the spillover of neurotransmitter from the active zone or through the actions of modulators released from other neurons, including the postsynaptic cell. Although presynaptic activation of $G_{i/o}$-coupled receptors almost invariably leads to inhibition of neurotransmitter release, this may be in large part due to their localization near the release machinery, which would permit certain target interactions, primarily via β/γ signaling mechanisms. Signaling through α subunits is also likely to play an important role through modulation of cAMP levels and kinase activity. Additional functions of Gα subunits may include determining the specificity of β/γ subunit interactions, as well as the termination of G-protein signaling through reassociation with β/γ subunits following GTP hydrolysis.

$G_{q/11}$-Coupled Receptors

$G_{q/11}$-coupled receptors, such as group I mGluRs and muscarinic M1 and M3 receptors, are also critical for short-term presynaptic modulation, although the best-described example comes from their postsynaptic role in the production of endocannabinoids. In numerous brain regions, including the cortex, striatum, hippocampus, and cerebellum, group I mGluR-mediated activation of $G_{q/11}$ signaling leads to the postsynaptic biosynthesis of endocannabinoids, at least in part due to the activation of phospholipase C beta (PLCβ) and DAG production. A similar effect of M1/M3 receptor activation has been observed in the striatum and hippocampus, indicating that activation of PLCβ represents a common target for enhancing endocannabinoid release through different $G_{q/11}$-coupled receptors. After synthesis and release from postsynaptic neurons, endocannabinoids can bind to presynaptic $G_{i/o}$-coupled receptors and inhibit neurotransmitter release for tens of seconds.

Direct presynaptic actions of $G_{q/11}$-coupled receptors have been more difficult to establish, particularly given the need to rule out endocannabinoid signaling in cases of apparent direct presynaptic inhibition. Because no firm evidence exists for presynaptic G_q-coupled receptor modulation, it is difficult to predict the effects of such signaling. Although it is possible that G_q-coupled receptors could have inhibitory actions via β/γ subunit signaling, $G_{q/11}$-mediated DAG production via PLCβ as well as subsequent PKC activation are predicted to have powerful excitatory presynaptic effects, given the actions of phorbol esters and PKC-mediated phosphorylation on proteins involved in vesicle priming and exocytosis. If presynaptic G_q signaling does occur, its precise nature may depend on the localization of the receptors relative to their possible signaling targets.

G_s-Coupled Receptors

Presynaptic G_s-coupled receptors are far less characterized in the central nervous system, but evidence suggests that these receptors can enhance neurotransmitter release, presumably through cAMP-dependent mechanisms. Paradoxically, some of the reported presynaptic actions of G_s-coupled receptors, such as dopamine D1 receptors, are actually inhibitory. For example, presynaptic G_s-coupled dopamine D1 receptors in the nucleus accumbens and cortex have been reported to reduce glutamate release. One explanation for these findings is that β/γ subunits released following activation of G_s-coupled receptors can inhibit release through mechanisms described later, although this has never been demonstrated in neurons.

Another peculiar finding is that some metabotropic receptors seem to be able to switch their G-protein coupling in response to different conditions. For example, $G_{i/o}$-coupled cannabinoid CB1 receptors may switch to G_s signaling when activated with low agonist concentrations or when associated with dopamine D2 receptors. In the cortex, $G_{q/11}$-coupled 5-HT2A receptors appear to also signal through $G_{i/o}$ when activated by class 5-HT2A agonists that have hallucinogenic properties in humans. Thus, metabotropic receptors may be capable of multiple conformational states that can modulate their coupling to G-protein signaling cascades.

Effectors and Targets of Presynaptic G-Protein Signaling

Many of the signaling targets of presynaptic modulation are the same proteins involved in activity-dependent short-term plasticity described in previous sections. Indeed, some signaling pathways may require both presynaptic activity and G-protein signaling. Thus, Gα signaling can lead to increases or decreases in important second messengers that control the phosphorylation state and function of presynaptic proteins involved in the docking, priming, and exocytosis of vesicles at the active zone. On more rapid timescales, Gβ/γ signaling can cause rapid and transient modulation of voltage-gated conductances as well as the protein complexes involved in SV priming and fusion.

Presynaptic Gα Signaling

Among the first signal transduction pathways described was the stimulation (by G_s) and the inhibition (by G_i) of adenylyl cyclase (AC) and cAMP. A striking feature of G_s activation and other similar signaling cascades is the capacity for robust signal amplification. In the case of G_s, only one or two GPCRs is required to stimulate GDP turnover on tens of G_s subunits, and each G_s-activated AC can in turn produce thousands of cAMP molecules. Another important aspect of signaling cascades is the divergence of signaling targets. In the case of the cAMP cascade, signaling can proceed in parallel through different mechanisms, including activation of PKA as well as activation of cAMP-GEFs (Epacs), both of which have been shown to mediate enhancement of neurotransmitter release at central nervous system synapses. PKA can phosphorylate numerous proteins, such as synapsin, snapin, α-SNAP, syntaphilin, tomosyn, CSP, SNAP-25, and RIM. All of these proteins are involved in the SV cycle, and their phosphorylation may alter

docking, SNARE complex formation, and release of SVs. In addition to these actions, Epac activation of monomeric G-proteins such as Rap and Ras can lead to additional signaling through MAP kinases and other signaling pathways, although the precise presynaptic actions of Epac have yet to be determined.

A second major signal transduction cascade in the presynaptic terminal is mediated by $G_{q/11}$ activation of PLCβ. The same principle of amplification applies for PLC-mediated synthesis of DAG and IP3, although membrane availability of the DAG precursor PIP2 can play a limiting role in this reaction, and PIP2 may be under dynamic control. DAG can activate PKC, which, like PKA, has numerous synaptic protein targets. For example, the syntaxin-binding protein Munc18, which binds to and stabilizes the closed form of syntaxin, is phosphorylated by PKC, leading to a decrease in its affinity for syntaxin and an increase in the probability of SNARE formation. PKC also phosphorylates other key synaptic proteins, including calcium channels and SNAP-25. DAG may also be important for presynaptic signaling independent of PKC via its interactions with Munc13, which were discussed previously.

Presynaptic Gβ/γ Signaling

Given the critical role of calcium channels in the triggering of neurotransmitter release, it is perhaps not surprising that modulation of calcium channels represents a major target for G-protein signaling. However, what initially surprised investigators was that this inhibition was not mediated by Gα subunits but instead by direct binding of G-protein β/γ subunits to calcium channels. In contrast to Gα signaling, no amplification occurs with β/γ signaling, suggesting that such signaling is likely to be more local and also have faster kinetics. Interestingly, Gα subunits may play an important role not only in the localization, specification, and timing of Gβ/γ signaling but also through downstream kinase signaling. For example, β/γ binding to calcium channels can be blocked by phosphorylation of the channel by PKC, suggesting the possibility of metaplastic control by $G_{q/11}$ signaling.

G-protein β/γ subunits can also directly bind and activate a class of potassium channels known as G-protein-activated inward rectifying K-channels (GIRKs). Although in theory, such a mechanism could reduce calcium influx by speeding action potential repolarization, little evidence for a presynaptic role of GIRKs has been found. It has also been recognized that β/γ subunits can exert an inhibitory influence on AC activity, particularly calcium- and calmodulin-dependent AC1 and AC8, suggesting that metabotropic receptor signaling can override activity-dependent changes in cAMP. Perhaps the most intriguing target of Gβ/γ is SNAP-25, a critical component of the SNARE complex that binds synaptotagmin and mediates vesicle fusion. Gβ/γ subunits can directly bind to SNAP-25 and inhibit release, possibly by disrupting the association of SNARE complexes with synaptotagmin 1, providing a direct mechanism for G-protein-mediated modulation of the release machinery independently of calcium influx.

Regulation of G-Protein Signaling

Initially, it was thought that G-protein signaling was activated solely by metabotropic receptors and terminated by the intrinsic GTPase activity of Gα subunits, but the discovery of activators of G-protein signaling (AGS) and regulators of G-protein signaling (RGS) proteins suggests that these concepts need revision. AGS proteins represent a family of receptor-independent G-protein activators that function through either GEF activity or by sequestering Gα or Gβ/γ subunits. A number of different AGS family members are expressed in neurons, where they could modulate receptor-mediated signaling in unexpected ways, such as uncoupling metabotropic receptors from $G_{i/o}$ or G_s pathways or enhancing basal inhibition by Gβ/γ through sequestration of Gα-GDP. However, their specific roles in presynaptic neuromodulation have yet to be characterized.

On the other hand, RGS and related family proteins have already been shown to play important roles in neuronal modulation and presynaptic plasticity, both through their GTPase-accelerating protein (GAP) activity and through other domain–domain interactions. For example, RGS6, -7, -9, and -11 have a G-protein γ-subunit-like domain that can bind to Gβ5 subunits. Gβ5 acts with RGS as an obligate heterodimer to drive GAP activity rather than as a classical β subunit in the β/γ complex. RGS12 and -14 contain Ras binding domains (RBDs) as well as $G\alpha_i$ binding domains. Interestingly, the RGS14 RBD binds Rap, suggesting a possible interaction between Epac and RGS signaling. Functionally, RGS proteins have been shown to modulate both basal synaptic properties and presynaptic modulation. In cultured hippocampal neurons, RGS2 overexpression gives rise to enhanced release probability and paired-pulse depression, whereas RGS2 knockout mice exhibit decreased release probability and enhanced paired-pulse facilitation, without any alterations in vesicle cycling, indicating that RGS2 is likely modulating basal G-protein inhibition of calcium channels. In the striatum, RGS9-2 has been reported to reduce modulation by dopamine

D2 receptors in both cholinergic and medium spiny neurons, and RGS4 reduces the efficacy of muscarinic M4 autoreceptor inhibition from cholinergic interneurons. Thus, the RGS family likely provides a widespread mechanism for the regulation of G-protein signaling in the central nervous system.

Conclusion

Short-term plasticity can arise from numerous sources, including presynaptic calcium-dependent signaling pathways, changes in the distributions and associations of presynaptic proteins following vesicle fusion, as well as transient cell surface signaling through G-protein-coupled receptors. However, all of these signaling mechanisms appear to converge onto common targets representing the proteins mediating SV trafficking, docking, priming, fusion, and recovery. In order to fully understand the mechanisms of neurotransmitter release and STSP, a comprehensive list of presynaptic proteins involved in neurotransmitter will have to be determined. Furthermore, three-dimensional models will have to be developed that incorporate the massive web of intermolecular associations among these proteins, as well as their regulation by calcium and phosphorylation, in both space and time. Although such a description is many years away, some of the most important proteins and their interactions are beginning to be characterized, giving us a glimpse into the vast complexity underlying regulated neurotransmitter release.

See also: Active Zone; Endocytosis and Presynaptic Scaffolds; Exocytosis: Ca^{2+}-Sensitivity; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neurotransmitter Release: Synchronous and Asynchronous; Presynaptic Facilitation; Presynaptic Inhibition; SNAREs; Synaptic Depression; Synaptic Vesicles; Synaptic Transmission: Models; Vesicle Pools.

Further Reading

Baba T, Sakisaka T, Mochida S, and Takai Y (2005) PKA-catalyzed phosphorylation of tomosyn and its implication in calcium-dependent exocytosis of neurotransmitter. *Journal of Cell Biology* 170: 1113–1125.

Blumer JB, Cismowski MJ, Sato M, and Lanier SM (2005) AGS proteins: Receptor-independent activators of G-protein signaling. *Trends in Pharmacological Sciences* 26: 470–476.

Ferguson GD, Wang H, Herschman HR, and Storm DR (2004) Altered hippocampal short-term plasticity and associative memory in synaptotagmin IV (−/−) mice. *Hippocampus* 14: 964–974.

Gerachshenko T, Blackmer T, Yoon E-J, et al. (2005) Gβ/γ acts at the C terminus of SNAP-25 to mediate presynaptic inhibition. *Nature Neuroscience* 8: 597–605.

Han J, Mark MD, Li X, et al. (2006) RGS2 determines short-term synaptic plasticity in hippocampal neurons by regulating Gi/o-mediated inhibition of presynaptic calcium channels. *Neuron* 51: 575–586.

Junge HJ, Rhee JS, Jahn O, et al. (2004) Calmodulin and Munc13 form a calcium sensor/effector complex that controls short-term plasticity. *Cell* 118: 389–401.

McCudden CR, Hains MD, Kimple RJ, Siderovski DP, and Willard FS (2005) G-protein signaling: Back to the future. *Cellular and Molecular Life Sciences* 62: 551–577.

McEwen JM, Madison JM, Dybbs M, Kaplan JM, et al. (2006) Antagonistic regulation of synaptic vesicle priming by Tomosyn and UNC-13. *Neuron* 51: 303–315.

Rhee JS, Betz A, Pyott S, et al. (2002) Beta phorbol ester- and diacylglycerol-induced augmentation of transmitter release is mediated by Munc13s and not by PKCs. *Cell* 108: 121–133.

Schoch S and Gundelfinger ED (2006) Molecular organization of the presynaptic active zone. *Cell Tissue Research* 326: 379–391.

Sippy T, Cruz-Martin A, Jeromin A, and Schweizer FE (2003) Acute changes in short-term plasticity at synapses with elevated levels of neuronal calcium sensor 1. *Nature Neuroscience* 6: 1031–1038.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Takamori S, Holt M, Stenius K, et al. (2006) Molecular anatomy of a trafficking organelle. *Cell* 127: 831–846.

Tedford HW and Zamponi GW (2006) Direct G protein modulation of Cav2 calcium channels. *Pharmacological Reviews* 58: 837–862.

Zucker RS and Regehr WG (2002) Short-term synaptic plasticity. *Annual Review of Physiology* 64: 355–405.

Synaptic Precursors: Filopodia

E Robles and S J Smith, Stanford University, Stanford, CA, USA
M P Meyer, Kings College London, London, UK

© 2009 Published by Elsevier Ltd.

Filopodia as Synaptic Precursors

The vertebrate central nervous system (CNS) contains billions of neurons connected by an even greater number of synapses. Despite this complexity, the CNS develops reliably and precisely. The goal of developmental neurobiologists is to understand the mechanisms that control the accurate formation of synaptic connections in the developing brain. This article will review evidence that axonal and dendritic filopodia are fundamental players in the assembly of functional neural circuits. As will become apparent during the course of this article, the highly motile nature of filopodia and the relationship of filopodia to synaptogenesis are central to their role in generating a precisely ordered nervous system. This article is structured, therefore, around the participation of filopodia in synapse formation, the regulation of filopodial motility, and the way such regulation is fundamental to the genesis of appropriate and ordered synaptic connections. In part one we concentrate on a phenomenological discussion of the role of filopodia in synaptogenesis and growth and branching of axonal and dendritic arbors. This is followed by a discussion of intrinsic and extrinsic factors that may regulate the formation of synaptic contacts by altering filopodial motility in developing neurons.

Dendritic Filopodia as Spine Precursors

Dendritic spines were discovered in 1888 by Ramon y Cajal, who noticed that Purkinje cells' dendrites were decorated with small thorns ('espinas'). He later proposed that these small, submicrometer protrusions from the parent dendrite were the site of axodendritic contact. Cajal's hypothesis was proved correct 60 years later, when the electron microscope was employed to show that spines were indeed sites of synaptic contact between axons and dendrites. In fact, more recent estimates suggest that more than 90% of excitatory axodendritic synapses in the mature CNS occur on dendritic spines. Spines must therefore be key elements in neuronal circuitry, and their structure suggests that one fundamental function of spines may be to bridge physical gaps between densely packed dendrites and axons. Hence, the mechanisms governing spine morphogenesis likely play a fundamental role in the selection of synaptic partners and establishment of functional connectivity.

A clue to the developmental origins of spines comes from the observations that prior to the overt presence of spines, many elongated filopodia extend from dendritic shafts. These filopodia are long (2–20 μm), thin (<0.3 μm) structures and lack the bulbous head that is characteristic of spines. The elongated morphology of dendritic filopodia hints at an exploratory role, and further evidence for this has come from time-lapse imaging studies showing that filopodia are highly motile structures. For example, in early postnatal hippocampal slice cultures, filopodia extend and retract rapidly from dendritic shafts and have a mean lifetime of only 10 min or so. At later developmental times, there are progressively fewer filopodia, a trend accompanied by a steady increase in the number of more-stable spinelike structures. The scarcity of filopodia on mature dendrites suggests a mainly developmental role, and the structural similarities and sequential appearance of filopodia and spines suggests that filopodia might be spine precursors. Consistent with this hypothesis, in dissociated hippocampal neurons maintained *in vitro*, the appearance of spines coincides with an increase in the number of functional synapses and a reduction in the number of motile filopodia. Furthermore, the total number of filopodia formed is more than sufficient to account for the entire spine population. Formation of a contact between a dendritic filopodium and a nearby axon has also been observed directly in hippocampal neurons *in vitro*. This contact event resulted in the stabilization of the filopodium and the formation of presynaptic bouton at the contact site. These *in vitro* studies have been complemented by several electron microscopy studies of developing tissues that demonstrate the presence of synapses on filopodia, clearly demonstrating that filopodia also form synapses *in vivo*. These findings have been incorporated into a filopodial model of synaptogenesis and spine formation: (1) dendritic filopodia on immature dendrites are highly dynamic, facilitating short-range exploration and contact initiation; (2) contact between a filopodium and a neighboring axon may result in the formation of a nascent synapse; and (3) maintenance of the nascent synapse stabilizes the filopodium on which it is located, transforming it into a dendritic spine. In contrast, filopodia that fail to contact an axon, or contact an axon but fail to form a stable synapse would then be rapidly eliminated.

Caveats and Alternative Models of Spine Formation

Presented with the evidence above, one could be forgiven for thinking that spinogenesis is essentially a

process of trial and error. Highly labile filopodia extend randomly from the dendritic shaft, and those that establish a stable synaptic contact with a presynaptic axon are transformed into a more stable spine. While this may be the case for some neurons, it does not appear to be true for all neuronal cell types. A number of lines of evidence suggest that spine formation can occur independently of a presynaptic cell. For example, in 'weaver' mutant mice, granule cells – the presynaptic partners of the vast majority of Purkinje cells – are absent. In these animals, the main Purkinje cell dendrites are abnormal but are nevertheless decorated with spines. Even in the normal cerebellum, the distal dendritic branches of Purkinje neurons develop spines before synaptic contact with parallel fibers' axons. Thus, the normal developmental sequence of events suggests that Purkinje cell spines can exist in the absence of presynaptic axons. Such uninnervated spines have also been observed in the barrel cortex of adult mice. Time-lapse imaging *in vivo* has demonstrated the emergence of new spines from layer 5 pyramidal neuron dendrites. Reconstruction of new spines using serial electron microscopy has shown that at least some of them lacked a presynaptic partner. These data suggest a model in which spine formation is an intrinsic property of some neurons and so can occur without interaction with a presynaptic cell. However, given the importance of reciprocal pre- and postsynaptic interactions during nervous system development, it is unlikely that uninnervated spines are completely normal.

Filopodia and Formation of Dendritic Branches

The data described previously suggest that filopodia are important for establishing synaptic contact with nearby axons and that successful filopodia are stabilized as spine precursors. However, synapses are also present on dendritic shafts, and shaft synapses are particularly prevalent at early developmental stages. How are shaft synapses formed? Also, filopodia occur in neurons that do not form spines. What is the function of filopodia in nonspiny neurons? One intriguing idea that could answer both of these questions is that filopodia are important not only for establishing spines and spine synapses but also for development of dendritic branches and the shaft synapses on them. Support for this idea has come from time-lapse imaging studies of nonspiny tectal cell dendrites as they arborized in the tectum of live zebra fish larvae. Engineering zebra fish tectal neurons to express a fluorescent postsynaptic marker has enabled time-lapse imaging to be used to study the relationship between dendritic filopodia, synaptogenesis, and growth of dendritic arbors *in vivo*. Growth of tectal dendrites is highly dynamic, involving prolific extension and retraction of filopodia, and the use of a fluorescent synaptic marker has revealed that virtually all tectal synapses form on newly extended filopodia. It is surprising that the majority of these nascent synapses and the filopodia that bore them were transient, lasting only tens of minutes or so, indicating a high turnover rate for both synapses and filopodia. The crucial finding, though, was that the fraction of synapses that were maintained in turn stabilized the filopodia on which they were located, and these stabilized filopodia matured into dendritic branches (see **Figure 1**). Successive iterations of filopodial extension and stabilization by synapses resulted in growth and branching of the entire dendritic arbor. Thus, in nonspiny tectal dendrites, filopodia are important for establishing synaptic contacts (as they do in spiny neurons), but in these cells, stabilized filopodia go on to form dendritic branches. Thus, the dendrite essentially grows out to meet axons via filopodial stabilization at the point of synaptic contact. It is not yet known whether shaft synapses and dendritic branches form this way in dendrites that go on to form spines.

Filopodia and Synaptotropic Growth of Dendrites

The data obtained from zebra fish tectal neurons are interesting for several reasons: first, the high turnover of nascent synapses implies that an important function of filopodial dynamics is to establish many trial synapses with nearby axons. Those synapses which are maintained, and presumably 'correct,' in turn stabilize growth of the dendrite in regions of appropriate synaptic partners. Conversely, filopodia which do not form stable synapses are rapidly retracted, and further growth in these regions is prevented. Second, because growth is stabilized only in correct target regions, filopodia that extend from a previously stabilized one are more likely to make contact with correct targets. Thus, once a correct synapse forms, a positive feedback is established that guides growth of the arbor into appropriate target regions (see **Figure 1**). Synaptogenesis in this context is effectively a growth guidance mechanism that provides a means to match the extent of a dendrite to the set of connections it forms, that is, there are no branches without synapses. This mechanism relies on randomly distributed filopodial extension and a local stochastic process of synapse stabilization. Modeling studies have shown that such a process can generate precise patterns of connectivity independently of innate dendrite branching programs or expression of dendritic guidance molecules. This is potentially very important as it is unlikely that the genome can encode

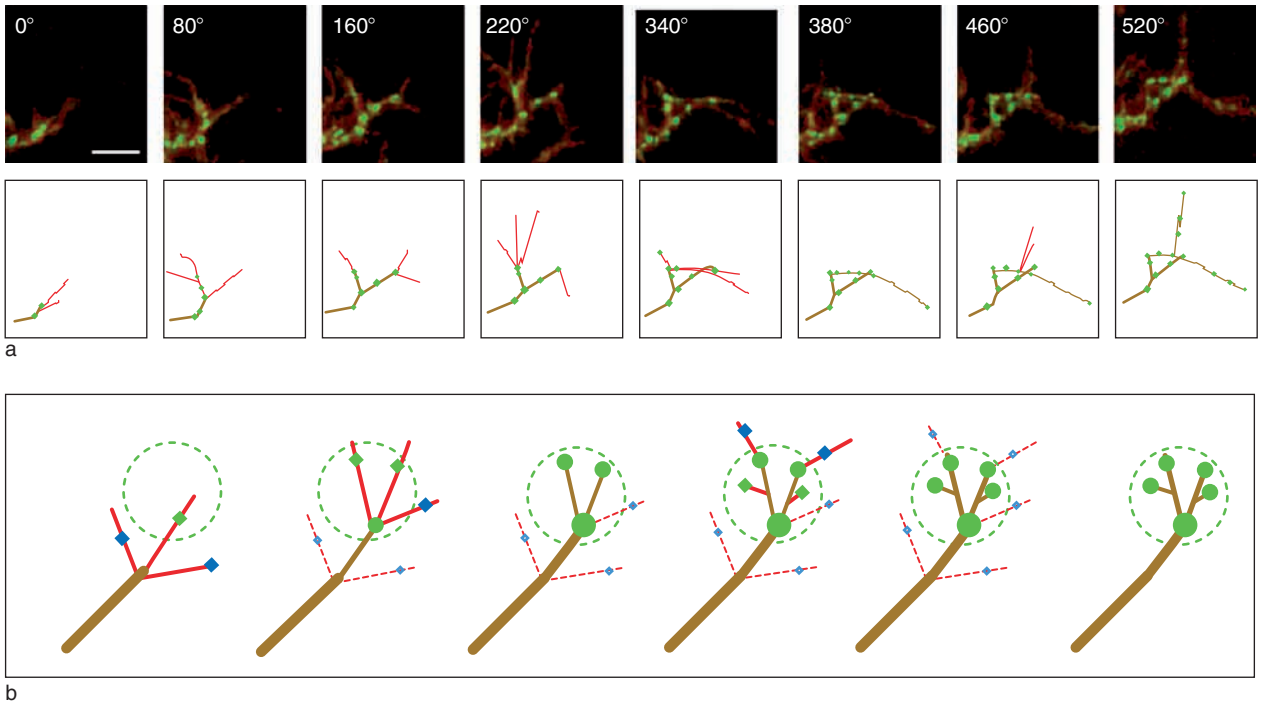

Figure 1 (a) Growth of zebra fish tectal cell dendrites *in vivo*. Growth occurs by an interactive sequence of selective filopodial stabilization and synapse formation. Still images from time-lapse series, accompanied by a schematic rendering for clarity. Green represents accumulations of the postsynaptic marker PSD-95: green fluorescent protein (GFP), and red areas are newly formed (often transient) branches. Brown in the schematic represents persistent branches. (b) Model of synaptotropic guidance of dendrite arbor growth. A number of filopodia (solid red) extend from a dendritic branch. Those that encounter correct partners and form synaptic connections (green dots) are stabilized as new branches (brown), whereas those that establish inappropriate contacts (blue dots) are retracted (dashed red). Successive rounds of selective stabilization result in arborization within a field of appropriate synaptic connections (dashed green region).

explicit instructions to form every synapse in the nervous system. The 'synaptotropic hypothesis' of dendrite growth described above was first proposed based on electron microscopy studies of developing rat spinal cord. The presence of synapses on filopodia and the fact that dendrites did not extend until presynaptic partners had arrived are both consistent with the idea that formation of a synapse could stabilize nascent dendritic branches and bias further growth in that direction. A very similar pattern of dendritic growth has been observed in real time in the zebra fish retina. Similar to tectal cell dendrites, retinal ganglion cell (RGC) dendrites were highly dynamic at early developmental stages and demonstrated little net growth indicative of unguided exploration. However, as the laminar organization of amacrine cell processes (the presynaptic partners of RGC dendrites) began to develop, RGC dendrites became progressively more stable and ramified sequentially in amacrine plexuses. Thus, growth and targeting of prepatterned afferents by RGC dendrites appear to occur by a process of selective stabilization of nascent branches, perhaps by a synaptotropic mechanism.

However, there is ample evidence indicating that many dendrites do not rely on a synaptotropic mechanism. For example, hippocampal neurons form complex dendrites *in vitro* in the absence of presynaptic partners, and in the drosophila olfactory system, second order projection neurons elaborate dendrites in a highly stereotyped manner prior to the arrival of the olfactory sensory neurons. Therefore, during development, multiple strategies are employed to generate appropriate dendritic arbor morphology. Further studies should allow the relative contribution of each strategy to be tested and the contexts in which a particular strategy predominates to be elucidated.

Axonal Filopodia, Synaptogenesis, and Growth of Axonal Arbors

So far this article has focused on the role of dendritic filopodia. However, filopodia also extend from shafts of axonal branches. These filopodia are therefore distinct from the filopodia that emanate from the leading edge of axonal growth cones. While growth cone filopodia mediate axonal navigation, the filopodia that emerge from axonal branches appear to play a

role similar to that of dendritic filopodia, that is, in synaptogenesis and the formation of axon branches. Since the role of axonal filopodia has been investigated less extensively, the data that support this idea will be summarized only briefly. Time-lapse imaging has shown that axonal filopodia are highly motile *in vitro* and *in vivo*, consistent with an exploratory role. Most axonal filopodia are transient, lasting only a few minutes or so. Second, motility of axonal filopodia is downregulated as development and synaptogenesis proceed. Third, a number of different techniques have been employed to show that axonal filopodia are able to establish synaptogenic contact *in vitro* and *in vivo*. These studies have also indicated that formation of a stable synapse stabilizes the axonal filopodium on which it is located. As is the case with zebra fish tectal cell dendrites, axonal filopodia of retinal axons (the presynaptic partners of tectal cells) that are stabilized by synapses go on to form new axon branches from which further growth can occur. Selective stabilization of axonal branches by synapses has also been observed in the *Xenopus* retinotectal projection. These data suggest that as is the case with tectal cell dendrites, there is also a synaptotropic component to growth of retinal axonal arbors. Thus, stabilization of filopodia by synapses is important for guiding the formation of retinotectal connectivity from both the pre- and postsynaptic sides.

Despite the obvious similarities between axonal and dendritic filopodia, there is one notable difference. Axonal filopodia contain clusters of vesicles that are capable of exo-endocytic recycling, whereas dendritic filopodia do not. Thus axonal filopodia may be sources of neurotransmitter release prior to the formation of obvious synapses. Local release of neurotransmitters (and neurotrophic factors) may have an important role in regulating synapse formation by locally modulating filopodial dynamics.

Regulation of Filopodial Motility

Filopodia serve an exploratory function during synaptogenesis, actively probing the extracellular environment for instructive cues located in the extracellular environment and on the surface of other cells. Filopodial motility increases the probability of synaptic contact between pre- and postsynaptic cells simply by increasing the area sampled in search of an appropriate synaptic partner. Filopodial motility also provides a mechanism that allows neurons to form many trial synapses with multiple potential synaptic partners. Because of the complexity of nervous system connectivity, this element of trial and error is indispensable as it allows for the formation of precise circuitry without the need to explicitly encode every synaptic connection within the genome. In contrast to the trial-and-error nature of synaptogenesis, filopodial motility is highly regulated by both intrinsic and extrinsic signals. While intrinsic signals may function to promote filopodial motility throughout the entire neuron, extrinsic signals may function to guide synapse formation and arbor growth through local regulation of filopodial motility. In the following section we describe well-characterized intrinsic and extrinsic factors that regulate filopodial motility and discuss the potential role of these factors during the process of synaptogenesis.

Intrinsic Regulation of Filopodial Motility

One potential explanation for the early prevalence of filopodia in developing neurons is that filopodium formation during this developmental period is increased as a result of an intrinsic genetic program. This idea is supported by evidence that cytoskeletal proteins associated with neurite growth are expressed at high levels during periods of neurite growth and synapse formation and are subsequently downregulated in the mature nervous system. Filopodial protrusions consist primarily of parallel-bundled actin filaments oriented with the growing 'plus' end toward the filopodial tip. Filopodial initiation and extension occur through regulated actin monomer addition at the plus ends of preexisting actin filaments. Neuronal actin consists of two distinct isoforms, γ- and β-actin, of which β-actin is strongly enriched at sites of dynamic cytoskeletal remodeling, such as filopodia and spines. High expression levels of neuronal β-actin during periods of neurite growth and synaptogenesis are consistent with actin-dependent filopodial dynamics' being controlled by an intrinsic genetic program. In addition to regulation by protein expression, actin dynamics in developing axons and dendrites could be elevated through upregulation of cytoskeletal regulatory proteins. Several types of signaling proteins have been shown to regulate actin dynamics and neuronal architecture, including actin-binding proteins, Rho family guanosine triphosphatases (GTPases), tyrosine and serine-threonine kinases, and motor proteins. In axonal growth cones, several cytoskeletal regulators have been localized to a filopodial tip complex located at the interface between the distal, growing ends of filopodial actin filaments and the plasma membrane. Recent studies also suggest that local intracellular signals generated within this specialized tip complex, function to direct local actin polymerization and drive filopodial extension. Proteins localized to the tip complex in developing

neurons include members of the Ena/VASP family of actin-binding proteins, Src-family tyrosine kinases, the Rho family GTPase Cdc42, and the Cdc42 effector protein p21-activated kinase (PAK). Although these mechanisms have been best characterized in axonal filopodia, Cdc42 activation has also been shown to correlate with increased motility of dendritic filopodia. This indicates that, in addition to obvious structural similarities, the motility of axonal and dendritic filopodia is regulated by common signaling pathways. Intrinsic regulation of growth-associated mechanisms may contribute to cellwide filopodial motility in neurons during development. However, a synaptotropic model of synapse formation suggests that filopodial dynamics be locally regulated, implying the involvement of extrinsic regulators of filopodial dynamics.

Regulation of Filopodial Dynamics by Brain-Derived Neurotrophic Factor

Activity-dependent secretion of neurotrophins may guide synaptotropic arbor growth by promoting filopodial motility at synapse-rich regions of developing neuronal arbors. Brain-derived neurotrophic factor (BDNF), signaling through its receptors TrkB and p75, may serve this function within the CNS. It is secreted in an activity-dependent manner and has been shown to alter neuronal morphology and synapse density. Furthermore, BDNF application has also been shown to promote filopodial motility in axons and dendrites of several types of neurons *in vitro*. BDNF stimulation has been shown to modulate filopodial dynamics by stimulating the GTPase activity of Cdc42, which likely acts through downstream targets such as PAKs, which regulate filopodial motility in growth cones and synaptogenesis and spinogenesis in hippocampal neurons. Recent findings have also demonstrated that clusters of TrkB receptors are transported into both axonal and dendritic filopodia, consistent with the idea that filopodia can sense extracellular BDNF. In *Xenopus laevis* tadpoles, local BDNF administration in the optic tectum has been shown to increase the density of synapses on retinal axons and tectal neuron dendrites. This increased synapse formation correlates with an increase in retinal arbor complexity. Together these studies suggest that activity-dependent secretion of BDNF may promote filopodial motility locally and in so doing promote synaptogenic contact between axons and dendrites.

Regulation of Filopodial Motility by Neurotransmitters and Neuronal Activity

It is well established that neuronal activity plays an important role in the accurate formation and refinement of neuronal connections during development. Given the central role of filopodia during synaptogenesis, an attractive model is one in which neuronal activity promotes filopodial exploration and stability. Globally, action potential firing may operate in conjunction with intrinsic genetic programs to promote filopodial motility in immature neurons during development. In addition, synaptic transmission and local depolarization of specific dendritic regions may promote synaptotropic growth by promoting filopodial extension in the vicinity of functional synapses. Evidence for this model was first provided by reports that application of the excitatory neurotransmitter glutamate stimulates rapid filopodial growth from hippocampal neuron dendrites (see Figure 2). These findings suggested that neurotransmitter secretion could promote synaptogenic contact by triggering local filopodial growth from potential postsynaptic partners. However, it should be noted that glutamate has been shown to either increase or decrease the motility of axonal and dendritic filopodia, suggesting that its effects are highly context dependent. Synaptically evoked action potential firing has also been shown to promote the formation of motile dendritic filopodia. This suggests a mechanism that promotes dendritic filopodial exploration and subsequent synaptogenesis and arbor growth in neurons receiving strong synaptic inputs. Such a mechanism could locally promote further synaptogenic contact between appropriately matched axonal and dendritic arbors. In support of this model, *in vivo* imaging has provided evidence that sensory experience promotes the protrusion of dendritic filopodia from pyramidal neurons in the developing rat barrel cortex and *Xenopus* tectal neurons. In the barrel cortex, reduced filopodial motility induced by whisker trimming correlates with the formation of less refined receptive fields, suggesting that activity-dependent filopodial exploration is necessary for accurate formation of sensory maps during development. Neuronal activity may promote filopodial motility by direct activation of second-messenger signaling cascades or by heightening sensitivity to locally secreted cues such as BDNF.

Role of Calcium Signaling in Filopodial Motility

The regulation of filopodial motility by action potential firing and glutamate application suggests that intracellular calcium elevations may mediate observed changes in neuronal structure and motility. This possibility is supported by several reports of neuronal growth and morphology being regulated by calcium-sensitive proteins such as calmodulin, calmodulin-dependent protein kinase II (CaMKII), the intracellular protease calpain, and the phosphatase calcineurin. The β isoform of CaMKII has been shown to specifically regulate the motility of dendritic filopodia and the formation of

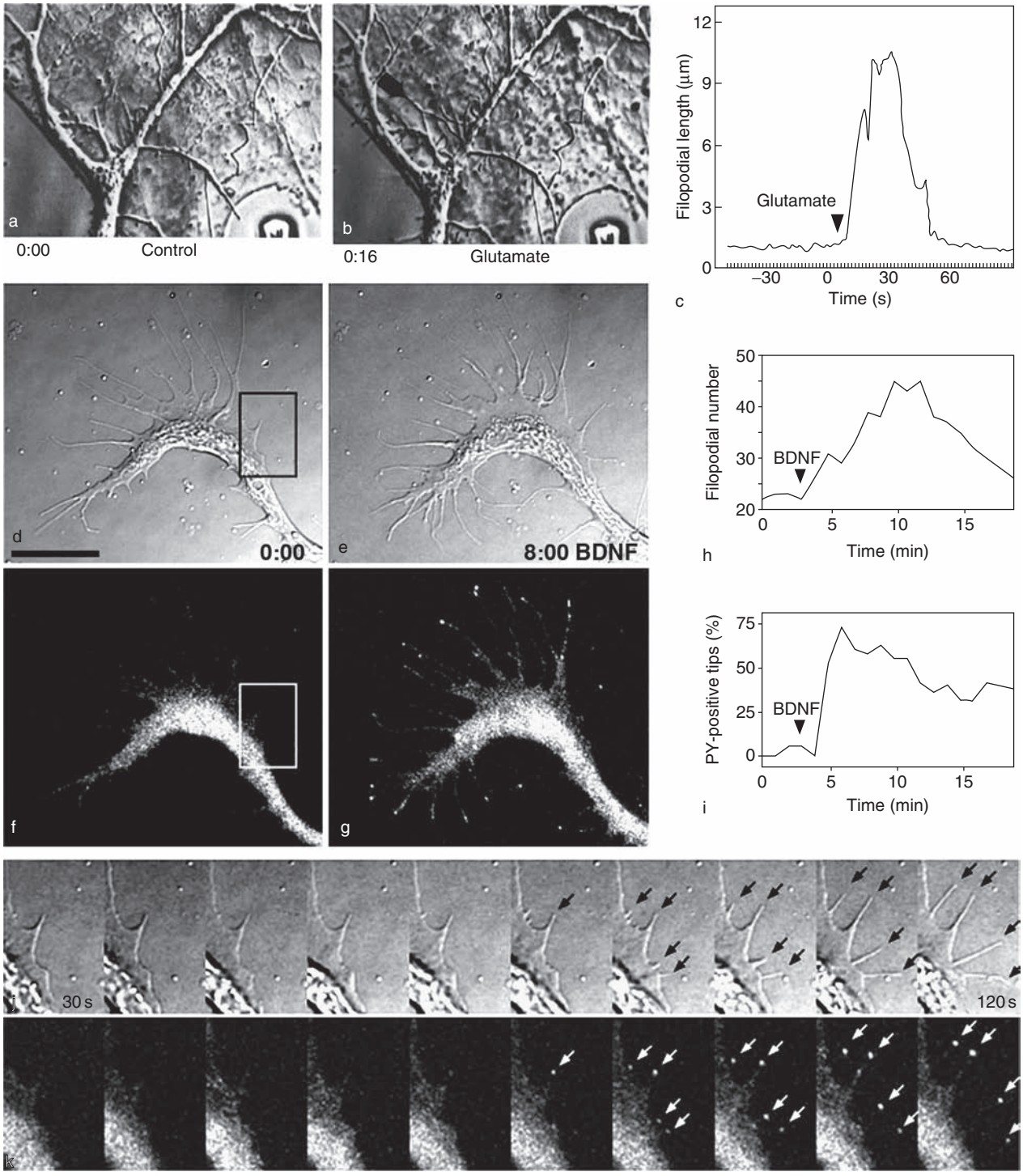

Figure 2 Rapid modulation of filopodial motility by glutamate and brain-derived neurotrophic factor (BDNF). (a, b) High-magnification view of a hippocampal neuron dendrite at indicated times after local pulatile application of glutamate. Note rapid increase in filopodial number and length. Arrow indicates a newly sprouted filopodium. (c) Measurement of filopodial length as a function of time in response to glutamate application for the filopodium indicated by arrow in (b); time in seconds. (d–g) Time-lapse differential interference contrast (DIC) and fluorescence images of a *Xenopus* spinal neuron expressing green fluorescent protein (GFP)-dSH2, a fluorescent indicator of tyrosine phosphorylation (PY), at indicated times after bath application of BDNF. Note that many filopodia lack enriched GFP-dSH2 at their tip before BDNF application but accumulate GFP-dSH2 after BDNF stimulation. (j), (k) Time-lapse DIC and GFP-dSH2 fluorescence images of boxed regions indicated in (d)and (f)at indicated times after BDNF application. Time between images is 10 s. Arrows point to one preexisting filopodium that acquires tip PY at 80 s after BDNF stimulation and three newly extended, PY-containing filopodia. (h, i) Kinetics of increased filopodium formation and increased GFP-dSH2 accumulation at filopodial tips in response to BDNF. Scale bar = 5 μm (d).

excitatory synapses by hippocampal neurons *in vitro*. The beta isoform of CaMKII, which is the predominant isoform expressed during early development, has also been specifically shown to bind to F-actin in hippocampal dendrites and to localize to dendritic regions undergoing dynamic structural rearrangements. Although the specific mechanism by which CaMKIIβ functions to promote filopodial motility is unknown, one possibility is that actin binding functions to localize CaMKIIβ to a multimolecular signaling complex. Within such a complex, CaMKIIβ activation by intracellular calcium elevation and calmodulin binding may lead to spatially precise activation of downstream signaling proteins that regulate the actin cytoskeleton. Another mechanism by which calcium signaling may regulate local motility of filopodia is through the generation of local calcium transients restricted to small regions of a dendritic or axonal arbor. The function of dendritic spines as discrete calcium signaling compartments has been demonstrated by monitoring intracellular calcium levels in response to synaptic transmission. However, similar observations have been made in axonal and dendritic filopodia, suggesting that calcium signaling may regulate filopodial motility and thereby influence synaptogenesis. Consistent with this possibility, calcium imaging in RGCs has revealed that newly formed dendritic branches also exhibit local neurotransmitter-evoked calcium elevations, and these signals may function to stabilize nascent synaptic contacts. This model for calcium-dependent stabilization of dendritic branches is consistent with calcium imaging studies examining the function of local calcium transients in filopodia of hippocampal neuron dendrites and *Xenopus* spinal neuron growth cones. In both systems, increased frequencies of local calcium transients correlated with the stabilization of active filopodia, suggesting that extracellular signals that promote the stabilization of synaptic contacts may do so through modulation of intracellular calcium levels. Although calcium-sensitive signaling cascades can control neuronal motility, the way in which calcium signals regulate distinct aspects of filopodial motility such as growth and stabilization remains unclear.

Filopodial Stabilization by Adhesion Molecules

During synaptogenesis, filopodial exploration and recognition of a putative synaptic partner lead to filopodial stabilization through an adhesive interaction with the target cell. Initial contact is mediated by the binding of transmembrane adhesion receptors such as members of the cadherin, protocadherin, and SynCAM protein families as well as neuroligin and β-neurexin. Of these adhesion molecules, N-cadherin has been localized to early contacts between dendritic filopodia and axons, where it functions to stabilize synapses and promote the formation of mature spine synapses. Following transmembrane engagement, structural stabilization of this contact site arises through direct interaction between n-cadherin and α-catenin, an actin-binding protein that forms a functional linkage between the transmembrane adhesion protein and the actin cytoskeleton. Disruption of either n-cadherin or α-catenin function results in altered spine morphologies characterized by the appearance of filopodia-like motile protrusions extending from the spine head. This suggests that in a system in which filopodia function as spine precursors, synapse stabilization by cadherins and catenins normally inhibits filopodial exploration. In addition to serving a structural function in maintaining physical contact between pre- and postsynaptic cells, transmembrane adhesion proteins can serve a signaling role in directing the assembly of synaptic specializations at the original site of adhesion. The homophilic adhesion molecule SynCAM has been shown to be necessary for the formation of glutamatergic synapses by directing the formation of presynaptic specializations. Similarly, presynaptic neuroligin binding to postsynaptic β-neurexin stimulates the local recruitment of both pre- and postsynaptic components. Although the dynamic localization of SynCAM and β-neurexin–neuroligin during synaptogenesis has not yet been determined, the fact that filopodia are known to make synaptic contacts suggests that these synaptogenic molecules may well be expressed on them. Expression of synaptic cell adhesion molecules on the surface of filopodia may function to halt filopodial exploration on formation of a synaptic contact, thereby allowing differentiation of these early contacts into mature, stable synapses.

See also: Cell Adhesion Molecules at Synapses; Dendritic Spine History; Postsynaptic Development: Neuronal Molecular Scaffolds; Spine Plasticity.

Further Reading

Abe K, Chisaka O, VanRoy F, and Takeichi M (2004) Stability of dendritic spines and synaptic contacts is controlled by alpha N-catenin. *Nature Neuroscience* 7: 357–363.

Alsina B, Vu T, and Cohen-Cory S (2001) Visualizing synapse formation in arborizing optic axons *in vivo*: Dynamics and modulation by BDNF. *Nature Neuroscience* 4: 1093–1101.

Biederer T, Sara Y, Mozhayeva M, et al. (2002) SynCAM, a synaptic adhesion molecule that drives synapse assembly. *Science* 297: 1525–1531.

Boda B, Alberi S, Nikonenko I, et al. (2004) The mental retardation protein PAK3 contributes to synapse formation and plasticity in hippocampus. *Journal of Neuroscience* 24: 10816–10825.

Bond JF and Farmer SR (1983) Regulation of tubulin and actin mRNA production in rat brain: Expression of a new beta-tubulin mRNA with development. *Molecular & Cellular Biology* 3: 1333–1342.

Dailey ME and Smith SJ (1996) The dynamics of dendritic structure in developing hippocampal slices. *Journal of Neuroscience* 16: 2983–2994.

Eom T, Antar LN, Singer RH, and Bassell GJ (2003) Localization of a beta-actin messenger ribonucleoprotein complex with zip-code-binding protein modulates the density of dendritic filopodia and filopodial synapses. *Journal of Neuroscience* 23: 10433–10444.

Fiala JC, Feinberg M, Popov V, and Harris KM (1998) Synaptogenesis via dendritic filopodia in developing hippocampal area CA1. *Journal of Neuroscience* 18: 8900–8911.

Fink CC, Bayer KU, Myers JW, Ferrell JE Jr., Schulman H, and Meyer T (2003) Selective regulation of neurite extension and synapse formation by the beta but not the alpha isoform of CaMKII. *Neuron* 39: 283–297.

Gehler S, Gallo G, Veien E, and Letourneau PC (2004) p75 neurotrophin receptor signaling regulates growth cone filopodial dynamics through modulating RhoA activity. *Journal of Neuroscience* 24: 4363–4372.

Gibney J and Zheng JQ (2003) Cytoskeletal dynamics underlying collateral membrane protrusions induced by neurotrophins in cultured *Xenopus* embryonic neurons. *Journal of Neurobiology* 54: 393–405.

Gomes RA, Hampton C, El-Sabeawy F, Sabo SL, and McAllister AK (2006) The dynamic distribution of TrkB receptors before, during, and after synapse formation between cortical neurons. *Journal of Neuroscience* 26: 11487–11500.

Gomez TM, Robles E, Poo M, and Spitzer NC (2001) Filopodial calcium transients promote substrate-dependent growth cone turning. *Science* 291: 1983–1987.

Graf ER, Zhang X, Jin SX, Linhoff MW, and Craig AM (2004) Neurexins induce differentiation of GABA and glutamate postsynaptic specializations via neuroligins. *Cell* 119: 1013–1026.

Hartmann M, Heumann R, and Lessmann V (2001) Synaptic secretion of BDNF after high-frequency stimulation of glutamatergic synapses. *EMBO Journal* 20: 5887–5897.

Jontes JD and Smith SJ (2000) Filopodia, spines, and the generation of synaptic diversity. *Neuron* 27: 11–14.

Katz LC and Shatz CJ (1996) Synaptic activity and the construction of cortical circuits. *Science* 274: 1133–1138.

Lanier LM, Gates MA, Witke W, et al. (1999) Mena is required for neurulation and commissure formation. *Neuron* 22: 313–325.

Lendvai B, Stern EA, Chen B, and Svoboda K (2000) Experience-dependent plasticity of dendritic spines in the developing rat barrel cortex *in vivo*. *Nature* 404: 876–881.

Lohmann C, Finski A, and Bonhoeffer T (2005) Local calcium transients regulate the spontaneous motility of dendritic filopodia. *Nature Neuroscience* 8: 305–312.

Lohmann C, Myhr KL, and Wong RO (2002) Transmitter-evoked local calcium release stabilizes developing dendrites. *Nature* 418: 177–181.

Maletic-Savatic M, Malinow R, and Svoboda K (1999) Rapid dendritic morphogenesis in CA1 hippocampal dendrites induced by synaptic activity. *Science* 283: 1923–1927.

Meyer MP and Smith SJ (2006) Evidence from *in vivo* imaging that synaptogenesis guides the growth and branching of axonal arbors by two distinct mechanisms. *Journal of Neuroscience* 26: 3604–3614.

Niell CM, Meyer MP, and Smith SJ (2004) *In vivo* imaging of synapse formation on a growing dendritic arbor. *Nature Neuroscience* 7: 254–260.

Robles E, Woo S, and Gomez TM (2005) Src-dependent tyrosine phosphorylation at the tips of growth cone filopodia promotes extension. *Journal of Neuroscience* 25: 7669–7681.

Sanchez AL, Matthews BJ, Meynard MM, Hu B, Javed S, and CohenCory S (2006) BDNF increases synapse density in dendrites of developing tectal neurons *in vivo*. *Development* 133: 2477–2486.

Sin WC, Haas K, Ruthazer ES, and Cline HT (2002) Dendrite growth increased by visual activity requires NMDA receptor and Rho GTPases. *Nature* 419: 475–480.

Smith SJ and Jahr CE (1992) Rapid induction of filopodial sprouting by application of glutamate to hippocampal neurons. In: Letourneau PC, Kater SB, and Macagno ER (eds.) *The Nerve Growth Cone*, pp. 19–26. New York: Raven Press.

Togashi H, Abe K, Mizoguchi A, Takaoka K, Chisaka O, and Takeichi M (2002) Cadherin regulates dendritic spine morphogenesis. *Neuron* 35: 77–89.

Waites CL, Craig AM, and Garner CC (2005) Mechanisms of vertebrate synaptogenesis. *Annual Review of Neuroscience* 28: 251–274.

Yuan XB, Jin M, Xu X, et al. (2003) Signalling and crosstalk of Rho GTPases in mediating axon guidance. *Nature Cell Biology* 5: 38–45.

Yuste R and Bonhoeffer T (2004) Genesis of dendritic spines: Insights from ultrastructural and imaging studies. *Nature Reviews Neuroscience* 5: 24–34.

Ziv NE and Smith SJ (1996) Evidence for a role of dendritic filopodia in synaptogenesis and spine formation. *Neuron* 17: 91–102.

Synaptic Transmission: Models

S Raghavachari, Duke University Medical Center, Durham, NC, USA
J Lisman, Brandeis University, Waltham, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Chemical signaling at a synapse occurs when a synaptic vesicle fuses with the presynaptic membrane in response to calcium influx through voltage-gated calcium channels. Vesicle fusion results in the formation of a fusion pore through which neurotransmitter packaged in the vesicle can escape into the synaptic cleft. The released neurotransmitter then diffuses through the synaptic cleft and binds to ligand-gated receptors on the postsynaptic membrane, which then lead to current flow across the postsynaptic membrane. To understand synaptic transmission, it is necessary to determine the spatiotemporal pattern of channel opening. However, physical limitations preclude direct observation of the dynamics of neurotransmitter release and receptor activation at single synapses. With the vast amount of biophysical and structural data that are now available, it is possible to make constrained models with sufficient biological realism that can be used to answer the following questions about the generation of synaptic responses:

1. How does the structure of the synapse shape the response?
2. How does the biophysics of receptor activation contribute to the response?
3. What is the influence of the kinetics of release of neurotransmitter and its reuptake or degradation?

We first discuss modeling methods that can be used for the study of synapses and then discuss the application of these methods to the study of excitatory transmission at the frog neuromuscular junction and at central glutamatergic synapses.

Modeling Approaches

The physical processes involved in chemical synaptic transmission are neurotransmitter diffusion and binding to receptors followed by conformational transitions of receptors, neurotransmitter binding to transporters for re-uptake (at central synapses), and neurotransmitter hydrolysis (in case of the neuromuscular junction).

The average time evolution of the neurotransmitter concentration in the cleft subject to these processes can be succinctly written as

$$\frac{\partial C(\mathbf{r},t)}{\partial t} = \left.\frac{\partial C(\mathbf{r},t)}{\partial t}\right|_{\text{diffusion}} + \left.\frac{\partial C(\mathbf{r},t)}{\partial t}\right|_{\text{reaction}} + \left.\frac{\partial C(\mathbf{r},t)}{\partial t}\right|_{\text{uptake}} \quad [1]$$

where $C(\mathbf{r}, t)$ is the concentration of the neurotransmitter at the spatial location $\mathbf{r}$ and time t. The diffusion of neurotransmitter is given by

$$\left.\frac{\partial C(\mathbf{r},t)}{\partial t}\right|_{\text{diffusion}} = D^2 C(\mathbf{r},t) \quad [2]$$

where ∇^2 is the diffusion operator, and D is the diffusion constant (measured in $\text{mm}^2\,\text{s}^{-1}$).

These processes can be simulated on computers using two fundamentally different mathematical approaches. In the first, the molecular nature of the neurotransmitter and receptors is ignored. The neurotransmitter concentration and the probability of receptor activation are treated as continuous quantities that vary smoothly in space and time. The time evolution of these quantities is then modeled using analytic approximations or partial differential equations. A second approach respects the discrete, probabilistic nature of neurotransmitter diffusion and receptor interaction. The ligand and receptor molecules are represented individually and undergo random motions (corresponding to Brownian diffusion) and stochastic interactions and transitions (corresponding to the reactions).

Continuum Methods

The main advantage of these continuum methods is their rapid implementation and the precise prediction of average quantities such as the time course of neurotransmitter clearance or receptor activation. Whereas early work using continuum models used simplified geometries, newer methods have allowed the incorporation of complex geometries, such as that of the neuromuscular junction.

In continuum models, the equations for neurotransmitter diffusion, uptake, and binding are supplemented by equations for the concentrations of receptors in different states. This system of differential equations is subject to the appropriate initial conditions, specifying the concentration of the neurotransmitter and the receptor states at time $t = 0$ at different spatial locations. The information on the synapse geometry is contained in the boundary conditions that govern the behavior of the neurotransmitter at reaction

boundaries as well as the inert surfaces. The latter represent physical barriers that the diffusing neurotransmitter cannot cross. These are mathematically known as zero flux boundaries, where the spatial gradient of the concentration $\partial C/\partial x = 0$. Uptake mechanisms or receptors located on the boundaries of the reaction space are instead modeled as absorbing boundaries where the concentration vanishes-that is, $C(r, t) = f(C)|_{\text{surface}}$.

In the absence of reaction and degradation/uptake terms in eqn [1], closed-form solutions of the diffusion equation are only possible for idealized geometries, which have a high degree of symmetry. Some example geometries include a cube, a cylinder, and a sphere. For example, a thin circular disk, with one face representing the presynaptic terminal and the other the postsynaptic membrane, is a good approximation of small central synapses. The solution of the diffusion equation in cylindrical coordinates is given in the form of Bessel functions, which can then, in principle, be used to study the dynamics of receptor opening. However, simple analytic models do not capture the complexity of the diffusion and reaction processes, and one must resort to numerical techniques for the solution of the full equation, including reaction terms and neurotransmitter uptake/degradation. A number of techniques have been developed for the solution of these equations under various approximations. Those relevant for synaptic transmission are discussed next.

Finite difference methods Finite difference (FD) methods are a powerful set of techniques to solve nonlinear reaction-diffusion equations (e.g., eqn [1]). These methods represent the concentration $C(\mathbf{x})$ by its values at discrete points in the reaction volume and approximate the derivatives (in space and time) as differences between adjacent points. This discretization results in a set of algebraic equations for the concentration of the diffusing neurotransmitter that can be solved using standard methods of linear algebra which have been implemented in software packages such as MATLAB. This approach is particularly useful for studying the synaptic current due to receptors with complex multistate kinetics as well as the effects of antagonists and modulators of receptors. Due to the long history of FD methods and their wide application, many optimization techniques are available for the simultaneous solution of the neurotransmitter diffusion, uptake, and receptor binding and even complex receptor kinetics. However, these methods are best suited for simplified geometries, but even here, the amount of discretization places severe restrictions on the accuracy of the method as well as its implementation on readily available personal computers.

Finite element methods Finite element (FE) methods are widely used in engineering applications to study how material structures behave under applied loads. These methods have been extended to a wide variety of partial-differential equations, including nonlinear reaction-diffusion equations (e.g., eqn [1]). In a strict mathematical sense, FE methods find a piecewise approximation to the solution of eqn [1]. This is done by subdividing the arbitrarily complex reaction space into small (finite) elements and decomposing the solution as a linear weighted sum of functions that are defined only over the finite elements. This decomposition then results in a nonlinear matrix equation that must be solved in order to determine the coefficients. In three-dimensional space, these elements are usually trapezoidal (other more complex shapes are possible). The accuracy of the method, as for the FD method, depends on the discretization (with finer subdivisions yielding better accuracy.) Similar to FD methods, the discretization can be adaptive, with high spatial resolution where the concentration of neurotransmitter varies rapidly.

The diffusion of acetylcholine (ACh) in complex geometries can be implemented using publicly available FE packages to gain insight into the structure-function relationship at neuromuscular junctions (NMJs). Similarly, the effect of vesicular release and acetylcholinesterase (AChE) distribution on miniature excitatory postsynaptic current (mEPSC) kinetics has also been studied using such models. The key advantage of the FE method is that the computational cost is proportional to the number of finite elements. Thus, a full endplate current (representing the release of approximately 200–300 quanta in a geometrically complex NMJ) can be simulated. However, the implementation of the reaction terms in FE models is nontrivial. The esterase and the receptors must be represented as spatially delimited sinks resulting in the boundary condition

$$\hat{n}(\mathrm{x})D\Delta C(\mathbf{x}, t) = -kC(\mathbf{x}, t) \qquad [3]$$

where $\hat{n}(\mathbf{x})$ is the surface normal, and k is a constant that is related to the steady-state constant of ACh to the esterase or receptor. The kinetics of the ACh receptor are relatively simple (the only isomerization step involves the opening of the channel), which makes its implementation possible within the FE formalism. More complex reaction schemes (e.g., glutamate receptors) are much more difficult to represent using this method. This method has been used in a few cases to study the effect of differences in fold geometry on the kinetics of the mEPSC. Ultrastructural observations show that fast-twitch and slow-twitch muscles have different structures. Muscular

dystrophy is also known to distort NMJ structure, leading to differences in mEPSC kinetics.

Monte Carlo Approaches

Continuum methods such as FD and FE outlined previously predict the average properties of chemical reactions. Such models are accurate (within the limits imposed by stability and accuracy considerations) as long as the number of reactions modeled is large. However, the chemical reactions that underlie neurotransmission at a single central synapse present a challenge to such models. Experiments show that the synaptic current at small central synapses is due to the opening of tens of receptors. This implies that each vesicle only contains a few thousand molecules of neurotransmitter, of which several tens participate in receptor activation. Finally, the amplitude of the individual synaptic response is often variable from trial to trial. These considerations imply that the stochastic or 'noisy' nature of chemical reactions must be accounted for in modeling synaptic transmission. The main source of this noise is the ubiquitous 'thermal noise' that characterizes the diffusive (Brownian) motion of molecules and conformational transitions of receptor proteins. The Monte Carlo method assumes that the physical processes in synaptic transmission are inherently probabilistic. The dynamics of the individual reactions are then followed by rolling a dice at each time step (hence the name Monte Carlo) to decide how much a given ligand molecule will move and whether a particular reaction will occur or not. Thus, variability is an inherent part of the simulation procedure. The average behavior of the system is obtained by averaging the results of the simulations across multiple trial runs. One strength of the Monte Carlo method is the high level of spatial and temporal detail that can be achieved.

The heart of the Monte Carlo method is its approach to modeling the diffusion of neurotransmitter in complex reaction spaces. In general, a diffusing molecule at a point P at time $t=0$ has some thermal velocity and undergoes collisions with water molecules on the sub-picosecond timescale at room temperature. After many collisions (over a nanosecond timescale), the molecule may have moved to a new location in a random direction. The distribution of displacements from the original location can be calculated from the diffusion equation (eqn [1]). For diffusion in free space, the solution of this equation is given by

$$C(r,\Delta t) = \frac{C_0}{(4\pi D\Delta t)^{3/2}}\,\mathrm{e}^{-r^2/4D\Delta t} \qquad [4]$$

Thus, the fraction of molecules in a spherical shell (of volume $4\pi r^2 dr$) is

$$f_r = \frac{1}{(4\pi D\Delta t)^{3/2}}\,\mathrm{e}^{-r^2/4D\Delta t}(4\pi r^2 dr) \qquad [5]$$

which is the same as the probability of a radial displacement of a single molecule. Thus, diffusion of a single neurotransmitter molecule is simulated by choosing a random number with the previous probability distribution. A second random number specifies the direction of this displacement in three dimensions. When the diffusing molecule encounters an impenetrable barrier, it is reflected back into the reaction volume by an elastic collision. In this way, the zero flux boundary conditions specified previously are naturally handled. Note that collisions need not occur with nonreactive boundaries but also with reactive boundaries (representing a surface receptor). In this case, the reflection occurs with a probability that is related to the macroscopic on-rate constant of the ligand with its receptor. This probability is given by

$$p_{\mathrm{b}} = k_{\mathrm{on}}\frac{1}{2N_{\mathrm{av}}A_{\mathrm{tile}}}\left(\frac{\pi\Delta t}{D}\right)^{1/2} \qquad [6]$$

where k_{on} is the macroscopic on-rate constant for the ligand to bind to the receptor, N_{av} is Avogadro's number, and A_{tile} is the area of the surface element occupied by the receptor. Unimolecular reactions, such as conformational changes of receptors, also occur in a probabilistic manner, with the probability of any given transition at a time step given as

$$p_{\mathrm{u}} = 1 - \exp\left[-\left(\sum_i k_i\right)\Delta t\right] \qquad [7]$$

where k_i is the macroscopic rate constants of the ith possible reaction for a receptor at any given state (e.g., a ligand can unbind from a receptor or the receptor might isomerize to a new state). The choice of the time step is dictated by the fact that the reaction probabilities must be $<<1$. The power of the method comes from the realization that fixing the time step automatically fixes the spatial resolution of the method. The size of the reactive and nonreactive surfaces, on the other hand, can be specified to a much higher precision (a small multiple of the area of a receptor, ~10–20 nm). The Monte Carlo technique is thus a powerful approach to simulating ligand–receptor interactions in arbitrarily complex geometries as well as receptors (and ligands) with complex kinetic properties. However, this power comes at a high computational cost. The advent of modern computers and the powerful program MCell, which includes many optimizations that speed up the calculations necessary to track diffusion and ligand–receptor interactions, makes the use of this approach entirely feasible.

Measurement of synaptic response: Quantal analysis The postsynaptic response to stimulating a single input axon may be quite complex; several synapses may be involved and multiple vesicles may contribute to the response, even at a single synapse. The factors that determine the dispersion of vesicle-release times are complex, and it is desirable to simplify the issue by studying the response generated by a single vesicle. Miniature synaptic currents (mEPSCs) occur spontaneously and are thought to result from the release of a single vesicle. Computational modeling of the synaptic response has thus focused on whether the properties of mEPSCs can be accounted for. These include the rise time, fall time, and variance. The rise time of the mEPSC at the NMJ is approximately 100 ms. It initially appeared that the rise time of the mEPSC at glutamatergic synapses of the hippocampus was considerably longer, but with the development of dendritic recording methods that eliminated distortions due to inadequate voltage clamp, rise times similar to those at the NMJ were found. The decay times are also similar (5–10 ms), but the amplitudes differ enormously: 4 nA in the frog compared to 15pA at hippocampal synapses. Interestingly, the geometries of the two synapses, although having obvious differences, are similar in size. Each presynaptic active zone at the NMJ is organized in a linear array approximately 1 μm long; the largest hippocampal synapses have similar dimensions (the smallest are ~0.2 μm in diameter).

Information needed for modeling A full consideration of the various numerical values that must be known to model synaptic function is beyond the scope of this article. Here, we list the various constants that must be estimated:

1. The transmitter content of the vesicles is estimated to be approximately 2000–4000 for glutamate and approximately 10 000 for ACh.
2. This transmitter exits the vesicle into the cleft through a fusion pore. There is considerable research on the size of the fusion pore; it may be much smaller during kiss-and-run release than during full fusion. Importantly, several methods are available that can provide at least a rough estimate of the size of the fusion pore.
3. The geometry of the cleft can be determined from electron micrographs.
4. The diffusion constant of transmitter in the cleft has been estimated and shown to be close to the diffusion limit; apparently, occlusion of the cleft by transsynaptic structures is not a major impediment.
5. Transmitter may bind to proteins other than the transmitter channels (e.g., transport proteins). The concentration of these sites is probably sufficiently low that their influence on the rising edge of the response is negligible.
6. The density of transmitter-activated channels has been measured directly. Interestingly, the density of ACh receptors (the channel at the NMJ) is approximately 10^4 mm^{-2}, which is approximately 10 times higher than that of AMPA channels (the major type of glutamate channels).
7. A kinetic model of the transmitter-activated channel is required. This will specify the on-rate constant for binding, the rate constant for opening and desensitization, and so on. Generally, the development of such models is based on data from systems in which transmitter can be applied rapidly and under controlled conditions.

Modeling Cholinergic Transmission at the Neuromuscular Junction

The vertebrate NMJ is one of the best-studied model synapses because of its large size, which makes it easy to obtain electrical recordings. The basic rules of synaptic transmission were worked out in this system, and a large amount of physiological and morphological data are available. The NMJ is composed of three distinct compartments: (1)the presynaptic nerve terminal, which contains a large number of vesicles containing the neurotransmitter ACh-a subgroup of the these vesicles is located directly adjacent to the active zones; (2) an extensively folded postsynaptic membrane; and (3) the intervening 50 nm-wide synaptic cleft. The junctional folds (JFs) have variable depth (0.5–1 μm) and are contiguous with the cleft. The acetylcholine receptor (AChR) is a pentamer (subunit composition) and is distributed on the crests of the JFs with a density of approximately 7000–12 000 mm^{-2}, declining by 30–50% at the bottom of the folds. In addition to the receptors, the enzyme AChE, which degrades ACh, is present on the postsynaptic membrane at densities of 2000–3000 mm^{-2} on the primary cleft and the JFs in clusters of 12 (three tetramers).

Insights from Modeling

The key challenge in modeling synaptic response due to the release of ACh at the NMJ is to incorporate the intricate geometry of the cleft. The physical properties of the NMJ place strong constraints on the time course and magnitude of the mEPSC. Simulations incorporating the synaptic geometry, receptor properties, and speed of vesicle release led to the concept of a saturated disk-a region on the postsynaptic membrane where nearly all the ACh contained in a single vesicle can be bound to receptors. An important

implication of this finding is that the synapse has evolved for a near perfect postsynaptic detection of each molecule of ACh release (this is not the case for glutamatergic synapse). The area of this saturated disk, $a_{\text{disk}} \propto N/\sigma$, where σ is the density of AChRs and N is the number of ACh molecules in the vesicle. For typical values of σ of approximately 10 000 mm^{-2}, the area of the saturated disk is approximately 0.5 mm^2. Simulations of the mEPSC in a comblike approximation of the NMJ or a detailed reconstruction demonstrated the existence of this saturated disk. The idea of a saturated disk can be used to estimate the contribution of diffusion, binding, and receptor kinetics to the rise time of the mEPSC. Assuming that the entire content of the vesicle is released instantaneously into the cleft, the time required for this ACh to diffuse over the saturated disk is

$$t_{\text{diff}} \propto \frac{N}{\sigma D} \qquad [8]$$

where D is the diffusion constant of ACh in the cleft, N is the number of ACh molecules, and the numerical factor C arises due to the cleft geometry. For typical values in the NMJ, this corresponds to approximately 50 μs. Similarly, the contribution of binding to the rise time is given by the time taken for a given concentration of ACh in the cleft to bind to the receptor:

$$t_{\text{bind}} \propto \frac{h}{\sigma k_+} \qquad [9]$$

where h is the height of the cleft, σ is the concentration of AChR in the cleft, and k_+ is the binding constant of ACh to the receptor ($\sim 4 \times 10^7\,\text{M}\,\text{s}^{-1}$). This value ($\sim 40$ μs) is comparable to the diffusion time t_{diff}. Since ACh does not leave the cleft over the rise time, binding can keep pace with diffusion and all the ACh can be optimally bound to receptors in a small area, ensuring saturation. The third contribution to the rise time is due to the finite time taken for the receptor to change conformation from the doubly bound state to the conducting state. The relevant time constant for this conformational change is approximately 50 μs. These considerations give a rather straightforward account of the 100 μs rise time of the mEPSC.

One assumption in the models is that the concentration of ACh builds up rapidly (nearly instantaneously) in the cleft. Ultrastructural studies on the NMJ have led to a model of neurotransmission in which rapid fusion of a vesicle with the presynaptic membrane releases neurotransmitter into the synaptic cleft. Following fusion, a membrane pore approximately 10–15 nm long forms, connecting the vesicle interior to the cleft. The neurotransmitter then flows down the concentration gradient (~200–500 mM in the vesicle and ~0 mM in the cleft). The diameter of the pore has not been directly measured. However, AC capacitance measurements have been used to measure the conductance of fusion pores formed by small synaptic vesicles. These typically estimate the pore conductance to be >500 ps. Given this geometry, one can use computational models of diffusion through the fusion pore and within the cleft to determine the time course of ACh within the cleft. One can then determine whether this time course, together with the properties of the ACh channel, can account for the measured 100 μs rise time of the mEPSC. Monte Carlo simulations of the NMJ with realistic geometries have shown that the fast rise time is accounted for by passive diffusion provided that the pore expands in diameter at a rate greater than 25 nm ms^{-1} (i.e., the pore would expand by 2.5 nm over the rise time of the mEPSC), which is well within the working stroke of many molecular motors.

The power of the Monte Carlo method is best illustrated when studying the factors governing the variability of the quantal response. The stochastic opening and closing of the channels as well as randomness in the release location lead to a CV of 0.02–0.03, which is much smaller than observed (~0.2). What are the additional factors? One factor is the variation in the amount of ACh in a single vesicle. However, experimental measurements suggest that the mEPSC amplitude is uncorrelated with its rise time, whereas the simple considerations discussed previously suggest that large mEPSC (corresponding to larger amounts of ACh in a vesicle) would have a longer rise time. However, simulations using detailed spatial reconstruction of the NMJ have been used to show that variations in the cleft height and branching of the JFs contribute a significant amount of the variability.

Hippocampal Synapses

Morphological Constraints

There is generally a single synapse on spines of CA1 pyramidal cells. This synapse has a single presynaptic active zone and a single postsynaptic density. There is a large variation in the size of these synapses. The average nonperforated synapse on mushroom or stubby spines on P15 CA1 pyramidal cells is 0.1 mm^2; perforated synapses are larger, whereas synapses on thin spines are considerably smaller (0.05–0.07 mm^2 area). Immunogold labeling suggests that glutamate receptors of α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA) type are fairly uniformly distributed within the synapse, with a density of approximately 1000 mm^{-2}. Thus, large mushroom or stubby

spines can contain approximately 100 receptors, whereas thin spines have fewer than 20. On the other hand, *N*-methyl-D-aspartate type glutamate receptors are found on almost all synapses and their number is weakly correlated with the area of the synapse. The presynaptic bouton typically contains a single active zone whose size is matched to the postsynaptic density. The intervening synaptic cleft is typically 15–20 nm in width.

Central synapses lack enzymatic processes analogous to AChE to degrade and thus dilute the neurotransmitter released in the cleft. At central synapses, diffusion is the major mechanism for the clearance of neurotransmitter from the synaptic cleft. Thus, a good estimate of the diffusion constant of glutamate is essential for modeling the efficacy of receptor activation at synapses. Most neurotransmitters are small molecules whose diffusion coefficient in free solution is approximately $1\,\mathrm{mm}^2/\mathrm{ms}$. However, in the brain, their diffusion is likely to be slower for two reasons: The extracellular space resembles a porous medium (much like a sponge) and diffusion is restricted to small (nanomolar-sized) channels, and the synaptic cleft and the extrasynaptic space are packed with both charged and neutral macromolecules which further impede diffusion by volume exclusion. The hindrance of diffusion due to the porous nature of extracellular space as well as macromolecular crowding can be accounted for by estimating an effective diffusion coefficient, $D_{\mathrm{eff}} = D/\lambda^2$, where $\lambda > 1$ is the tortuosity of the medium, a measure of the hindrance (geometric or otherwise) experienced by the diffusing molecules. Measurements suggest that glutamate diffusion in the cleft is similar to its value in solution (decreased by approximately a factor of 2).

In the hippocampus, electron microscopy studies show that most functional synapses seem to have glial processes at the perimeter of the axon-spine interface. At most CA1 synapses, the astrocytic processes only partially surround the synapse, allowing glutamate to freely escape into the extracellular space. The astrocytes make up nearly 10% of the total membrane in CA1 and are lined with glutamate transporters (proteins that move glutamate from the extracellular space into the glial cell) at a density of approximately $10\,000\,\mathrm{mm}^{-2}$. These transporters also bind glutamate with high affinity (10–20 μM) and serve to either buffer or remove synaptically released glutamate. From a modeling standpoint, it is not correct to incorporate the effect of transporters by slowing the diffusion of glutamate in the extracellular space. This is because the buffering action of the transporters means that glutamate actually binds to them, resulting in a loss of glutamate from the extracellular space. Instead, the effects of transporters should be included by explicit modeling of the binding and unbinding of glutamate rather than a reduction of its diffusion constant.

Kinetic Constraints

AMPA receptors act on a rapid timescale and display weak affinity to glutamate (the dissociation constant, $K_D \sim 400\,\mu\mathrm{M}$). In recent years, these channels have been cloned and expressed in heterologous systems in order to study their electrophysiological properties under controlled conditions, leading to a deep understanding of the structure and kinetic properties of AMPA channels. These channels are tetramers composed of several stoichiometric combinations of glutamate receptor subunits (GluR) 1–4. Thus, channels may be homomers composed of identical subunits ($\mathrm{GluR1}_4$) or heteromers composed of distinct subunits in a strict stoichiometric manner ($\mathrm{GluR1}_2$–$\mathrm{GluR2}_2$, $\mathrm{GluR2}_2$–$\mathrm{GluR3}_2$, etc.). Depending on the subunit composition, AMPA channels may have distinct kinetic properties that shape channel kinetics and even permeability to different cations ($\mathrm{GluR1}_4$ are calcium permeable). Each subunit has a binding domain which can bind glutamate independent of whether glutamate is bound to any of the other subunits (i.e., there appears to be no cooperativity in glutamate binding). Structural studies indicate that agonist binding causes a closure of the binding domain and initiates conformational changes that serve to activate the channel. It appears that at least two subunits need to have glutamate bound in order for the channel to conduct ions. Importantly, additional binding greatly speeds channel opening and enhances the single channel conductance (typical values of conductance are 5–8, 7–10, and 12–15 ps, depending on subunit occupancy). However, it is not entirely clear whether these conductance steps occur for channels natively expressed in synapses. Furthermore, it appears that glutamate binding and unbinding can occur from both closed and open states, but binding and unbinding from the open states is slow.

A hallmark of glutamate channels is their ability to rapidly desensitize–that is, switch to a nonconducting conformation upon binding glutamate. The exact conformational changes that lead to desensitization are a subject of active research, but it is widely accepted that glutamate binding to the subunit can set in motion conformational changes leading to either activation or desensitization. Additionally, desensitization as well as recovery from desensitization show a concentration dependence that can be best explained by considering that each subunit is capable of desensitization upon binding glutamate (i.e., desensitization is not a cooperative phenomenon). These

considerations lead to models of AMPA receptor gating that may typically have 40–50 distinct states. Once the state diagram for the receptors is chosen, the transition rates between these states can be determined by fitting the data on channel activation and desensitization produced by the controlled application of glutamate to patches excised from neurons. Many of these rate constants are fixed to be multiples of each other due to the assumption of subunit independence. Further reduction is possible due to the presence of loops in the state diagram (different paths through a loop should be equivalent according to the principle of microscopic reversibility). The model that best fits data from controlled glutamate application to AMPA channels from CA1 pyramidal cells suggests that at low concentrations, desensitization competes effectively with channel opening, whereas at higher concentrations the cooperative process of channel opening makes the opening rate much higher than the desensitization rate.

Insights from Modeling

This model of receptor kinetics was used in a geometrically constrained model of a synapse to determine the influence of various parameters on synaptic transmission.

Time course and amplitude of the mEPSC At large mushroom spines in CA1, the number of neurotransmitter molecules vastly exceeds the number of receptors, and neurotransmitter diffusion and clearance is rapid. Thus, the loss of neurotransmitter due to receptor binding can be largely neglected. Moreover, the uptake mechanisms at central synapses are located outside the cleft. With these simplifications, eqn [1] can be solved analytically for a cylindrical cleft where glutamate is released at the center of the top face of the disk with absorbing boundary conditions at the cleft edges. Modeling the escape of neurotransmitter from the vesicle through a narrow fusion pore leads to a brief 'concentration spike' during which a large amount of neurotransmitter is confined to a small volume of the cleft near the release site. This spike extends approximately 100 nm from the release site and is dissipated within 100 μs (i.e., the concentration declines by $1/e$ of its peak value). During this spike, the glutamate concentration builds up to nearly 3 mM near the fusion pore. The rapid diffusion of glutamate dilutes this spike. The efficacy of the spike in producing receptor occupancy is made clear by a calculation analogous to the neuromuscular junction. If we take the area of the spine to be the upper limit in the expression for t_{diff} ($\sim$0.1 μm^2 for larger spines), we obtain a diffusion time of approximately 10 μs, which is much shorter than that in the NMJ. Assuming typical values of forward binding rate of 2×10^7 $M\,s^{-1}$, t_{bind} is 80 μs. This implies that the rise time at central synapses would almost entirely be determined by the binding of glutamate to the receptors as well as the receptor kinetics.

Solving the equations for receptor activation using the concentration of glutamate calculated previously results in a simulated mEPSC with an amplitude of approximately 10–15 pA and a 20–80% rise time of $<$100 μs, which matches experimental observations. Most of the current (nearly 80%) at the peak of the mEPSC is carried by a small number of channels in a 250-nm hot spot around the release site ($\sim$25). The multiple glutamate binding sites on the channel and the concentration dependence of desensitization play an important role in determining the size of this hot spot. Near the site of release, the concentration is in the millimolar range, which is required to effectively load the multiple glutamate binding sites on AMPA channels; the channel then opens efficiently because channel opening is greatly favored over desensitization. Outside the hot spot, opening is inefficient because site occupancy is low and the opening rate is therefore slower than the desensitization rate. If desensitization is blocked with cyclothiazide (a drug that inhibits desensitization of AMPA receptors), channels outside the hot spot can open, albeit slowly, and enhance the tail of the mEPSC.

Unlike the NMJ, the majority of the glutamate content of a vesicle at hippocampal synapses leaves the cleft without activating receptor channels. Moreover, a single vesicle is only able to activate a fraction of channels at a medium to large synapse. This nonsaturation has been observed experimentally. The implication of this nonsaturation is that hippocampal synapses, despite their low receptor density, can respond in a graded manner, which has consequences for the dynamic range of hippocampal synapses.

Origins of quantal variability A critical issue regarding the characteristics of transmission at central synapses is the variability of the mEPSC amplitude. At the NMJ, the distribution of the peak mEPSC is roughly Gaussian, with a CV of approximately 0.2. At CNS neurons (specifically at hippocampal CA1 synapses), early work showed that the mEPSC amplitude distribution exhibited a broad tail. This broad distribution could be due to the variation in the amount of glutamate contained in vesicles. Modeling and experimental studies that have varied the vesicular glutamate content do not show the distinct rise time–amplitude correlation that is observed experimentally.

On the other hand, histograms of responses evoked by minimal stimulation (stimulation of a single axon synapsing onto a target cell) often show peaks that are narrow, indicative of low quantal variation,

whereas the overall distribution is broad and skewed. Simulations using Monte Carlo models show that amplitude histograms with multiple narrow peaks could arise from single synapses releasing multiple vesicles. These simulations show that effective quantal summation can occur at single synapses and that fluctuations in release time and position are not sufficient to strongly smear the peaks. Although summation is nearly linear, there are interactions of the glutamate released by multiple quanta that produce some nonlinearity. These may explain why the spacing between quantal peaks may not be exactly even. Furthermore, the CV of high peaks is not higher than that of small peaks as would be expected if quantal responses were independent, and it may decrease and can even be less than that of small peaks. There is extensive experimental evidence of multivesicular release at hippocampal synapses. These considerations suggest that the broad distribution of mEPSCs could also arise from the spontaneous release of multiple vesicles at single synapses.

One implication of these results is that the quantal size can be relatively insensitive to the number of channels at the synapse (except at the smallest synapses, which have the dimension of the hot spot). The main postsynaptic determinant of quantal size is the local density of AMPA channels in the region of the synapse where a vesicle has been released. It is necessary to use the word 'local' because of the possibility that density may not be uniform within a synapse, as has been found at GABAergic synapses. The unimportance of total AMPA channel number in determining quantal size is illustrated by mossy fiber inputs to CA3 pyramidal cells. This connection occurs on a single giant spine that contains many synapses. These have a nearly identical AMPA channel density but highly variable size and AMPA receptor content. Consistent with importance of density rather than size, quantal size at these synapses is restricted to a narrow range.

Glutamate diffusion in extracellular space Synapses in the CNS are packed very tightly. Ultrastructural studies suggest that a cubic micrometer of cortex contains approximately one or two synapses. Unlike the NMJ, a typical hippocampal synapse contains a small number of receptors (1–200). Given a maximum of four binding sites per receptor, the number of glutamate molecules in small vesicles is far in excess of the number of receptors. If a large number of vesicles are released at different synapses, this excess glutamate will build up, leading to a nonnegligible amount of ambient glutamate available that can affect synaptic transmission. Since neurotransmitter released at central synapses is not degraded enzymatically, it can escape from the cleft by diffusion and activate receptors at nearby synapses. This heterosynaptic activation would appear as a slowly decaying tail in the AMPA receptor responses. Direct manipulations of glutamate diffusivity by large-molecular-weight polymers dramatically alter the tail of the AMPA response while minimally affecting the peak of the response. The process of glutamate diffusion and receptor activation in a geometrically complex space has been modeled in simplified geometries to show that extrasynaptic spillover plays an important role in excitatory transmission in the neocortex, hippocampus, and the cerebellum.

Further Directions in Modeling of Synaptic Transmission

Synaptic transmission involves a large array of molecules that show complex regulation and intricate localization properties. It is becoming increasingly apparent that realistic models of synaptic transmission hold great promise in making quantitative predictions. However, these models are only as good as the input parameters, such as accurately reconstructed diffusion volumes, measurements of receptor densities and kinetics, as well as dynamics of the release process. Many of these technologies are under rapid development. For instance, electron tomography in conjunction with improved sample preparation techniques can generate highly detailed (~3–8 nm resolution) three-dimensional structures of synapses. Highly sensitive freeze–fracture replica labeling has been used to determine the density of AMPA receptors in synaptic sites. Finally, new recording techniques allow for the detection of sub-millisecond vesicle fusion events. Numerical algorithms that can integrate this level of structural detail in computer models are also a subject of active research. However, these tend to be scattered across many disciplines (e.g., computational geometry and parallel computing). The main challenge is to combine these disparate approaches into a unified model of synaptic transmission including both presynaptic and postsynaptic aspects to obtain a deep quantitative understanding of synaptic physiology.

See also: Active Zone; AMPA Receptors: Molecular Biology and Pharmacology; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Enteric Nervous System: Neural Circuits and Chemical Coding; Fusion Pore; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions; Postsynaptic Density/Architecture at Excitatory Synapses; Synaptic Vesicles.

Further Reading

Barbour B and Hausser M (1997) Intersynaptic diffusion of neurotransmitter. *Trends in Neurosciences* 20(9): 377–384.

Bartol TM, Land BR, Salpeter EE, and Salpeter MM (1991) Monte Carlo simulation of miniature endplate current generation in the vertebrate neuromuscular junction. *Biophysical Journal* 59(6): 1290–1307.

Crank J (1980) *Mathematics of Diffusion*. Oxford: Clarendon.

Faber DS, Young WS, Legendre P, and Korn H (1992) Intrinsic quantal variability due to stochastic properties of receptor–transmitter interactions. *Science* 258(5087): 1494–1498.

Franks KM, Bartol TM, and Sejnowski TJ (2002) A Monte Carlo model reveals independent signaling at central glutamatergic synapses. *Biophysical Journal* 83(5): 2333–2348.

Franks KM, Stevens CF, and Sejnowski TJ (2003) Independent sources of quantal variability at single glutamatergic synapses. *Journal of Neuroscience* 23(8): 3186–3195.

Nielsen TA, DiGregorio DA, and Silver RA (2004) Modulation of glutamate mobility reveals the mechanism underlying slow-rising AMPAR EPSCs and the diffusion coefficient in the synaptic cleft. *Neuron* 42(5): 757–771.

Raghavachari S and Lisman JE (2004) Properties of quantal transmission at CA1 synapses. *Journal of Neurophysiology* 92(4): 2456–2467.

Smart JL and McCammon JA (1998) Analysis of synaptic transmission in the neuromuscular junction using a continuum finite element model. *Biophysical Journal* 75(4): 1679–1688.

Stiles J and Bartol T (2001) *Computational Neuroscience: Realistic Modeling for Experimentalists*, pp. 87–127. Boca Raton, FL: CRC Press.

Stiles JR, Helden DV, Bartol TM, Salpeter EE, and Salpeter MM (1996) Miniature endplate current rise times less than 100 microseconds from improved dual recordings can be modeled with passive acetylcholine diffusion from a synaptic vesicle. *Proceedings of the National Academy of Sciences of the United States of America* 93(12): 5747–5752.

Wahl LM, Pouzat C, and Stratford KJ (1996) Monte Carlo simulation of fast excitatory synaptic transmission at a hippocampal synapse. *Journal of Neurophysiology* 75(2): 597–608.

Wathey JC, Nass MM, and Lester HA (1979) Numerical reconstruction of the quantal event at nicotinic synapses. *Biophysical Journal* 27(1): 145–164.

Relevant Website

http://mcell.cnl.salk.edu – Salk Institute for Biological Studies; MCell: A Monte Carlo Simulator of Cellular Microphysiology.

Synaptic Vesicle Protein-2 (SV2)

R Janz, University of Texas Medical School, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Synaptic vesicle protein-2 (SV2) was originally identified with a monoclonal antibody generated against synaptic vesicles purified from the electric organ of the ray *Discopyge ommata*. SV2 was characterized as a glycoprotein that was associated with synaptic vesicles from neurons as well as with secretory vesicles in neuroendocrine cells. The SV2 antibody reacts with all types of vertebrate synapses, from fish to mammals, and is used widely as a synaptic marker in immunohistology.

There are three homologous isoforms of SV2 in mammals: SV2A, SV2B, and SV2C. Each isoform is encoded by a different gene, and all three isoforms react with the SV2 monoclonal antibody. All three SV2 proteins are integral membrane proteins with predicted 12 transmembrane domains and N- and C-termini facing the cytoplasm. An alignment of the protein sequences of the three SV2 isoforms is depicted in Figure 1. The three proteins share a high degree of homology, with approximately 60% identical sequences. SV2 proteins share significant homology with proteins of the major facilitator superfamily (MFS) of transporters, which has lead to the hypothesis that SV2 proteins are also transporters.

Structure of the Proteins

The cytoplasmatic N-terminus is the most divergent domain among the three SV2 isoforms. This domain is relatively similar between SV2A and SV2C, with homologous stretches of amino acids interrupted by nonconserved blocks of amino acids. In contrast to SV2A and SV2C, the N-terminal domain of SV2B is much shorter and shares less homology with the other isoforms. The N-terminal domain of all SV2s contains a short stretch of amino acids that is conserved among all SV2 proteins. This sequence is probably the epitope for the monoclonal SV2 antibody. Most loops connecting the 12 transmembrane domains are relatively short. Exceptions are the cytoplasmic loop connecting transmembrane domains 6 and 7 and the intravesicular loop connecting transmembrane domains 7 and 8. SV2 proteins have five conserved cysteine residues in intravesicular loops that are potentially involved in disulfide bonds. However, the actual location of the disulfide bonds is not currently known. All SV2 isoforms are highly glycosylated. Three predicted N-glycosylation sites are located in the large intravesicular loop between transmembrane domains 7 and 8. An interesting feature in the SV2 protein sequence is the presence of two conserved negatively charged amino acids (D179 and E182 in SV2A) located in the first transmembrane domain.

SV2 proteins share significant homology with the large gene family MFS. This gene family codes for a large group of proteins that catalyze the transport of specific substrates across biological membranes. Most MFS proteins have 12 transmembrane domains with the same topological orientation as SV2 proteins. Recently, high-resolution structures of two members of the MFS family, the glycerol-3-phosphate antiporter and lactose permease, have been solved. The orientation of the transmembrane domains is remarkably similar in these two bacterial proteins, indicating a conserved structure for this gene family. These high-resolution structures reveal that the 12 transmembrane domains form two 6-helical bundles. These two helical bundles comprise transmembrane domains 1–6 and transmembrane domains 7–12. A hydrophilic cavity is formed at the interface between the two helical bundles that is open to the cytoplasmic side. This cavity most likely forms the binding site for the substrate to be transported.

Based on the orientation of the transmembrane domains of the MFS proteins, a corresponding model for the structure of SV2 was generated, depicted in Figure 2. In this model, the two negatively charged amino acid residues in transmembrane domain 1 are facing the hydrophilic cavity. These residues could potentially mediate the binding of a positively charged substrate. It should be pointed out that this structural model is purely based on the homology of the SV2 proteins to the MFS and has not been tested experimentally.

Distribution of SV2 Proteins

The expression of the SV2 isoforms has been well characterized in the rat brain. SV2A is widely expressed and found in all synaptic layers of the brain. SV2B is also widely distributed in the brain. However, some areas of the brain do not express SV2B. SV2B is not expressed at mossy fiber synapses of the hippocampus, which are projections from pyramidal cells of the dentate gyrus to the CA3 region. In addition, SV2B is not expressed in synapses from projections of Purkinje cells to the deep cerebellar nuclei.

In contrast to the widespread distribution of SV2A and SV2B, SV2C has more restricted expression. SV2C is detectable at very low levels in the neocortex and the hippocampus and at high levels in synapses of

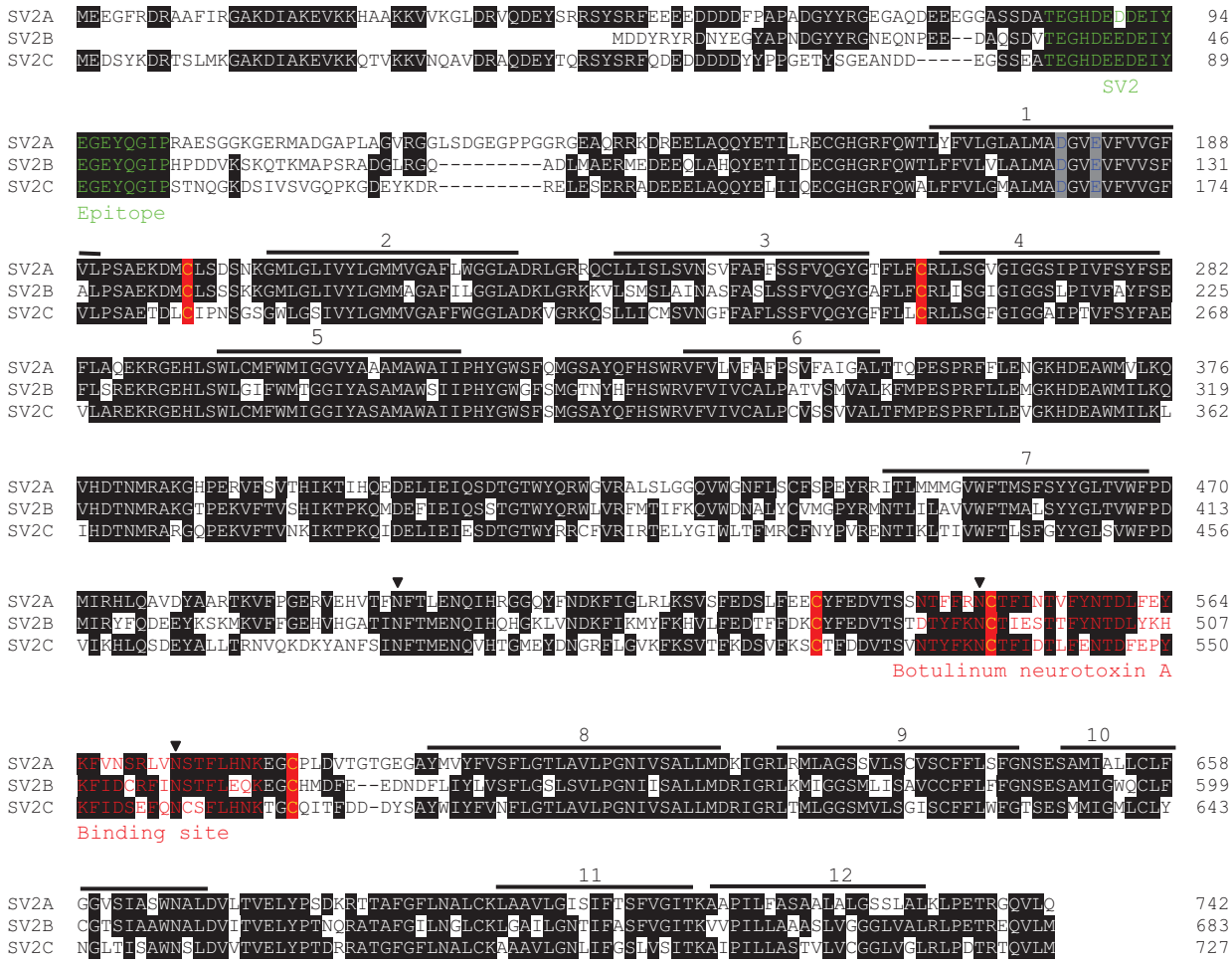

Figure 1 Structure of SV2 proteins. The protein sequences of rat SV2A, SV2B, and SV2C are aligned for optimal homology. The sequences are identified on the left and numbered on the right. Residues that are conserved between at least two of the isoforms are highlighted in white on a black background. At the N-terminus, the predicted epitope sequence for the SV2 monoclonal antibody is marked in green. The 12 predicted transmembrane regions are numbered and marked with lines above the sequences. The conserved negatively charged residues in the first transmembrane domain are marked in blue on a gray background. The three N-glycosylation consensus sequences located in the intravesicular loop between transmembrane domains 7 and 8 are marked by triangles above the sequences. Conserved cysteine residues in intravesicular loops that are potentially involved in disulfide bonds are labeled in yellow on red background. The binding domain for botulinum toxin A is labeled in red.

the basal ganglia and brain stem. In addition, SV2C protein is found in projections from Purkinje cells to deep cerebellar nuclei.

Outside the central nervous system, all SV2 isoforms are found at the neuromuscular junction. SV2A is also located in adrenal chromaffin cells and the pheochromocytoma cell line PC12. SV2B is not detectable in these two cell types. SV2C is expressed in adrenal chromaffin cells but not detected in PC12 cells.

Role of SV2 in Ribbon Synapses of the Retina

The analysis of the distribution of SV2 proteins in the mouse retina revealed that SV2A is found in all conventional synapses, but is absent in the specialized ribbon synapses of the rod photoreceptors and the bipolar cells. Ribbon synapses are characterized by a specialized presynaptic structure, the synaptic ribbon. These synapses release synaptic vesicles at an extremely high rate compared to conventional synapses.

Ribbon synapses from rod photoreceptors and bipolar cells express only SV2B. Interestingly, SV2B is absent from conventional synapses of the retina. This distribution indicates that SV2B might have a specialized role in ribbon synapses. Interestingly, rod photoreceptor synapses from SV2B-knockout mice show a reduction of the levels of the SV2-interacting protein synaptotagmin 1.

Cone photoreceptor synapses are a special exception to the typical retinal ribbon synapses because they also express SV2A. In contrast to the other SV2 isoforms, SV2C is detectable only in a small subset of retinal synapses formed by amacrine cell projections.

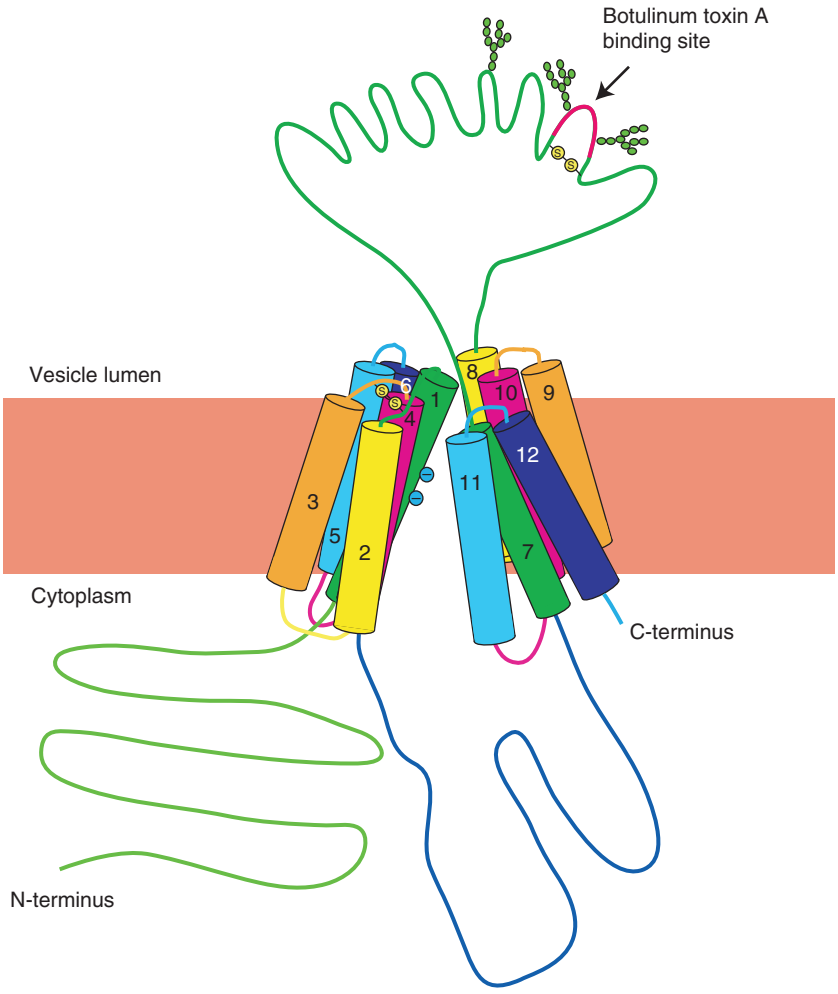

Figure 2 Model of the structure of SV2. proteins in the synaptic vesicle membrane, generated by adapting the arrangement of the transmembrane domains from the known structures of two bacterial MFS proteins. The predicted transmembrane helices are numbered and two potential disulfide bonds between conserved cysteine residues in the intravesicular loops are depicted. The binding domain for botulinum toxin A and the three predicted glycosylation sites in the intravesicular loop between transmembrane helices 7 and 8 are marked. The positions of the two negatively charged amino acid residues in transmembrane domain 1 are labeled. MFS, major facilitator superfamily.

Interacting Proteins

The exocytosis of synaptic vesicles is thought to be catalyzed and regulated by a controlled set of protein–protein interactions between synaptic vesicle proteins and plasma membrane molecules. One possible molecular mechanism for the role of SV2 in the regulation of synaptic vesicle exocytosis is through interaction with other synaptic proteins. Studies using native and recombinant proteins have established an interaction between SV2 proteins and the Ca^{2+} sensor for fast, Ca^{2+}-triggered exocytosis, synaptotagmin 1. However, the physiological role of this interaction is still unclear.

Analysis of synaptic vesicle exocytosis in hippocampal neurons from SV2A/B double knockout mice do not indicate a change in the Ca^{2+} dependence of synaptic vesicle exocytosis, indicating that the Ca^{2+}-sensing function of synaptotagmin 1 is unaffected by the removal of SV2. However, as previously mentioned, the removal of SV2B from photoreceptor synapses is correlated with a strong reduction in synaptotagmin 1 protein levels. These results indicate that SV2B may stabilize synaptotagomin 1 in this special set of synapses.

Functional Studies

The analysis of SV2A- and SV2B-knockout mice has demonstrated a role for SV2 proteins in the regulation of synaptic vesicle exocytosis. SV2A-knockout mice develop severe seizures after 1 postnatal week. The onset of these seizures is followed by weight loss and death of the mutant pups within 2–3 postnatal weeks.

Studies of synaptic transmission in cultured SV2A/B double knockout hippocampal neurons show reduced

synaptic depression after repetitive stimulation compared to wild-type mice. This phenotype was reversed by buffering residual synaptic Ca^{2+} using the Ca^{2+} chelator ethylene glycol tetraacetic acid (EGTA). These results suggest that SV2 proteins may regulate presynaptic Ca^{2+} levels. Because homology suggests SV2s may be transporters, they may regulate Ca^{2+} levels by transporting Ca^{2+} into synaptic vesicles.

Further studies from cultured SV2A/B double knockout hippocampal neurons show the removal of SV2 results in reduced synaptic vesicle exocytosis with low-frequency stimulation and reduced size of the readily releasable pool of synaptic vesicles. Recovery of the knockout phenotype to wild-type levels with high-frequency stimulation suggests that SV2 may regulate synaptic transmission in quiescent neurons.

Role of SV2 Proteins in Epilepsy

A first indication that SV2s might be involved in epilepsy came from the observation that SV2A-knockout mice develop severe seizures. In addition, SV2A was recently identified as the binding site for the antiepileptic drug levetiracetam (Keppra™, USB Pharmaceuticals). Levetiracetam is an FDA-approved drug that successfully treats drug-resistant, chronic epilepsy. The mechanism of action is currently not known. Because knockout of SV2A leads to seizures in mice, it is unlikely that levetiracetam works as an inhibitor of SV2A function. A more likely mechanism of action is that levetiracetam binding to SV2A activates or stabilizes the protein.

SV2 Proteins as Receptors for Botulinum Toxin A

Botulinum toxin A (botox A) is one of the highly effective neurotoxins produced by the bacterium *Clostridium botulinum*. Botox A consists of a heavy chain that mediates entry of the toxin into the presynaptic terminal and a light chain that cleaves synaptosomal-associated protein of 25 kDa (SNAP-25), thereby blocking synaptic vesicle exocytosis. The uptake of botox A into synapses is dependent on SV2 proteins. Botox A binds the loop between the transmembrane domains 7 and 8 of SV2 (see Figures 1 and 2) when that loop is exposed to the synaptic cleft after synaptic vesicle exocytosis. SV2-bound botox A is then endocytosed when SV2 is recycled into a synaptic vesicle. After endocytosis, the regulatory domain of botox A facilitates the translocation of its catalytic domain from the synaptic vesicle lumen into the cytoplasm of the presynaptic terminal. Once translocated, botox A wreaks havoc on a synaptic terminal through the cleavage of SNAP-25, completely shutting down neurotransmission. All SV2 isoforms are capable of mediating entry of botox A into synapses.

See also: Botulinum and Tetanus Toxins; Epilepsy; Epilepsy: Neuronal Death; Epilepsy: Channelopathies; Neuropeptides: Epilepsy; Ribbon Synapses; Synaptic Vesicles.

Further Reading

Bajjalieh SM, Frantz GD, Weimann JM, McConnell SK, and Scheller RH (1994) Differential expression of synaptic vesicle protein 2 (SV2) isoforms. *Journal of Neuroscience* 14: 5223–5235.

Bajjalieh SM, Peterson K, Linial M, and Scheller RH (1993) Brain contains two forms of synaptic vesicle protein 2. *Proceedings of the National Academy of Sciences of the United States of America* 90: 2150–2154.

Bajjalieh SM, Peterson K, Shinghal R, and Scheller RH (1992) SV2, a brain synaptic vesicle protein homologous to bacterial transporters. *Science* 257: 1271–1273.

Buckley K and Kelly RB (1985) Identification of a transmembrane glycoprotein specific for secretory vesicles of neural and endocrine cells. *Journal of Cell Biology* 100: 1284–1294.

Crowder KM, Gunther JM, Jones TA, et al. (1999) Abnormal neurotransmission in mice lacking synaptic vesicle protein 2A (SV2A). *Proceedings of the National Academy of Sciences of the United States of America* 96: 15268–15273.

Custer KL, Austin NS, Sullivan JM, and Bajjalieh SM (2006) Synaptic vesicle protein 2 enhances release probability at quiescent synapses. *Journal of Neuroscience* 26: 1303–1313.

Dong M, Yeh F, Tepp WH, et al. (2006) SV2 is the protein receptor for botulinum neurotoxin A. *Science* 312: 592–596.

Feany MB, Lee S, Edwards RH, and Buckley KM (1992) The synaptic vesicle protein SV2 is a novel type of transmembrane transporter. *Cell* 70: 861–867.

Janz R, Goda Y, Geppert M, Missler M, and Sudhof TC (1999) SV2A and SV2B function as redundant Ca^{2+} regulators in neurotransmitter release. *Neuron* 24: 1003–1016.

Janz R and Sudhof TC (1999) SV2C is a synaptic vesicle protein with an unusually restricted localization: Anatomy of a synaptic vesicle protein family. *Neuroscience* 94: 1279–1290.

Lazzell DR, Belizaire R, Thakur P, Sherry DM, and Janz R (2004) SV2B regulates synaptotagmin 1 by direct interaction. *Journal of Biological Chemistry* 279: 52124–52131.

Lynch BA, Lambeng N, Nocka K, et al. (2004) The synaptic vesicle protein SV2A is the binding site for the antiepileptic drug levetiracetam. *Proceedings of the National Academy of Sciences of the United States of America* 101: 9861–9866.

Schivell AE, Batchelor RH, and Bajjalieh SM (1996) Isoform-specific, calcium-regulated interaction of the synaptic vesicle proteins SV2 and synaptotagmin. *Journal of Biological Chemistry* 271: 27770–27775.

Wang MM, Janz R, Belizaire R, Frishman LJ, and Sherry DM (2003) Differential distribution and developmental expression of synaptic vesicle protein 2 isoforms in the mouse retina. *Journal of Comparative Neurology* 460: 106–122.

Synaptic Vesicles

S Takamori, Tokyo Medical and Dental University, Tokyo, Japan

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Communication between neurons or from neurons to their target tissues takes place at a specialized structure called the 'synapse' (Greek meaning 'to clasp'). Synapses consist of two functionally and morphologically distinct components: the presynapse, from which neurotransmitter molecules are released, and the postsynapse, where specific receptors for the respective neurotransmitters are localized on the surface and the complex signaling cascades proceed. As such, one neuron can activate or inactivate connected neurons or target tissues, depending on the transmitter molecule they utilize for their signal transmission. Based on electrophysiological experiments pioneered by Katz and colleagues in the 1960s, it has been postulated that neurotransmitters are released from presynaptic terminals in discrete packets termed 'quanta.' Electron microscopy revealed that at presynaptic terminals, hundreds of small and round membranous vesicles – synaptic vesicles (SVs) – are clustered, which can be reasonably linked to the quanta (Figure 1(a)). Furthermore, biochemical experiments using isolated vesicles from mammalian brains have proven that SVs exhibit a variety of transport activities for major chemical transmitters. Based on these findings, it is generally believed that SVs store neurotransmitters and the fusion of SV membrane to the plasma membrane elicits the quantal release of neurotransmitters. Because of their important roles in basic neural functions, much effort has been focused on understanding the molecular machinery of SVs. This article focuses on general features of SVs in terms of morphology, biogenesis, and recycling and their overall molecular composition.

General Features of Synaptic Vesicles as an Organelle

Clustered at the nerve endings, SVs are one of the most striking morphological hallmarks of presynaptic terminals in electron micrographs. They appear to be roughly homogeneous in size and shape, but it has been determined that some heterogeneity exists among them. Under certain experimental conditions, their shape appeared to be of two types – almost spherical and oval or flat. This difference may be an experimental artifact presumably from the fixation process. Interestingly, this difference in shape may be associated with their transmitter content. It was found that most excitatory, asymmetric glutamatergic synapses contain spherical vesicles, and inhibitory, symmetric, GABAergic synapses contain the latter. Regardless of the functional significance, this observation has provided a possible morphological characteristic to distinguish glutamatergic SVs and GABAergic SVs.

Due to their uniformity in size and abundance in the brain, it is feasible to isolate SVs with high purity and in large amounts – a prerequisite for biochemical and biophysical experiments (Figure 1(b)). Electron microscopic observations of isolated SVs from rat brains have revealed that the surface is uneven. They are decorated with one or more prominent globular structures and several spiky or amorphous substructures are observable under electron microscope. After proteolytic digestion these structures are removed and the rims of the lipid bilayer become clearly visible, demonstrating that the vesicles are coated with proteins on the surface (Figures 1(c) and 1(d)). As measured by electron microscopy, their size varies substantially: The diameter of the outer bilayer ranges from 35 to 50 nm, with an average peak at 42 nm. The total dry mass of an average SV was deduced from a combination of protein quantitation, lipid quantitation, and SV particle counting, resulting in approximately 30 attograms (ag) per vesicle, which consists of 17 ag of proteins and 12 ag of lipids. Assuming the thickness of a lipid bilayer is 4 nm, the average inner volume can be estimated as approximately 20×10^{-21} l, which provides enough space for approximately 1800 transmitter molecules at a concentration of 150 mM. Within these physical and molecular constraints, SVs are effectively equipped with a unique set of proteins and lipids necessary for executing the fundamental tasks in neurotransmitter release.

Biogenesis of Synaptic Vesicles

Like other proteins of the secretory pathway, most SV proteins are synthesized at the endoplasmic reticulum and processed through the Golgi apparatus for maturation in the cell body of neurons (Figure 2(a)). Conceptually, subdomains which selectively collect the SV proteins bud off from the Golgi apparatus and then travel along the axon to the presynaptic terminals. However, since no vesicles are as small as mature SVs, and some larger nonuniform tubulovesicular structures are seen in the axons, SV constituents are thought to travel along the processes on heterogeneous membranes termed SV precursors.

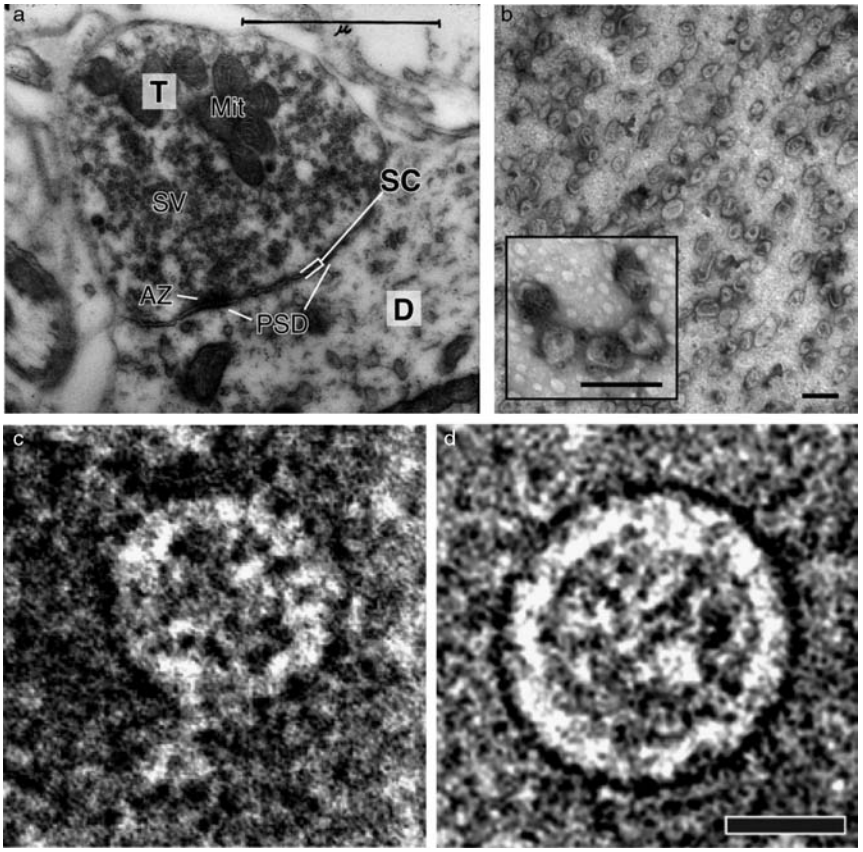

Figure 1 Morphology of synapse and synaptic vesicles. (a) A transmission electron micrograph of synapse. The presynaptic terminal (T) contains numerous synaptic vesicles (SVs) and mitochondria (Mit). A small portion of SVs is attached to electron dense structures at the presynaptic plasma membrane termed the active zone (AZ), where exocytosis of SVs takes place. Opposite the AZ, there are electron dense structures beneath the postsynaptic membrane termed postsynaptic density (PSD), where neurotransmitter receptors form signaling complexes. D, dendrite; SC, synaptic cleft. (b) An electron micrograph of synaptic vesicle fraction. SVs were purified from mouse whole brains, negatively stained, and imaged by transmission electron microscopy. Inset shows immunogold labeling of the vesicles with an antibody against an SV marker protein, synaptophysin. Approximately 95% of the membranous structures are labeled. (c, d) SVs before (c) and after (d) proteolysis imaged by cryo-electron microscopy. Scale bar = 1 μm (a), 100 nm (b, inset), 20 nm (c). (a) Adapted from the George E. Palade EM Slide Collection at Yale University School of Medicine. (c, d) Reproduced from Takamori S, Holt M, Stenius K, et al. (2006) Molecular anatomy of a trafficking organelle. *Cell* 127: 831–846, with permission from Elsevier.

Once formed, the SV precursor needs a pathway to reach its destination. Microtubules are known to serve as the roads for transporting cargo. Motor proteins of the kinesin superfamily (KIFs) bind to both the microtubules and the cargo, and they control the directional transport of intracellular transporting cargo. Among the KIFs, KIF1A and KIF1Bβ are known to participate in axonal transport of SV precursors. Gene disruption of either KIF1A or KIF1Bβ results in reduction of SV density at the presynaptic terminals and therefore impaired neurotransmission.

The KIF1A- and KIF1Bβ-carrying SV precursors contain the major SV marker proteins, such as synaptophysin, synaptobrevin, synaptotagmin, and rab3A, but not presynaptic plasma membrane proteins, such as syntaxin 1 and SNAP-25. In addition, proteins that form the cytomatrix at the active zone, such as piccolo and bassoon, are transported to the axon on distinct cargo, suggesting that segregation of the presynaptic proteins is initiated before they arrive at the terminals. The mechanism by which the SV constituents are selectively recruited into the SV precursor is unknown. It is possible that SV proteins have a specific signal sequence that is assembled and forms microdomains at the exit of the *trans*-Golgi network, but no such common signal sequence has been found. Furthermore, the molecular mechanism regarding how the constituents of the SV precursors can be recognized by KIF1A/1Bβ proteins is poorly understood.

When SV precursors arrive at the presynaptic terminal, they fuse with the plasma membrane where the main route for SV biogenesis is initiated. This route, the AP-2-dependent pathway, is mediated by several essential proteins for vesicle endocytosis, namely clathrin, dynamin, and AP-2. An alternative route by

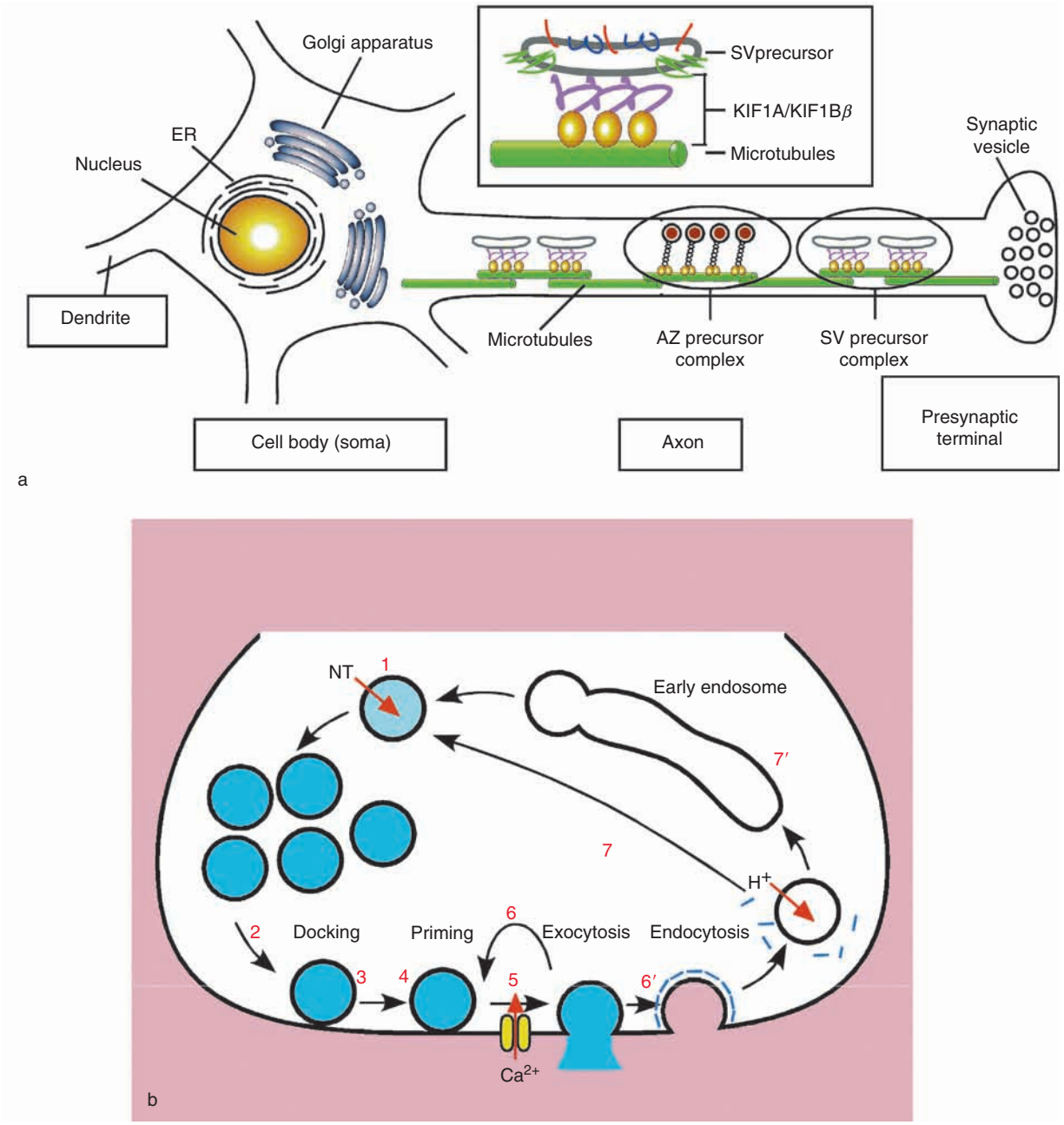

Figure 2 Biosynthesis and recycling of SVs. (a) Biosynthesis of SVs. SV proteins are synthesized in the cell body of neurons and the SV precursors bud off from the *trans*-Golgi network. The SV precursors are transported along the axon, guided by the microtubules and kinesin motor proteins (KIF1A and KIF1Bβ). (b) SV cycle at presynaptic terminal. (1) SVs are filled with neurotransmitter. (2) The SVs are then transported to the active zone and docked to the presynaptic plasma membrane (docking) (3). The docked SVs become fusion competent by molecular events called priming (4). When an action potential arrives at the terminal, calcium influx through voltage-dependent calcium channels triggers fusion of SV membrane with the plasma membrane (exocytosis), causing discharge of SV content (5). Exocytosed SVs are regenerated either by clathrin-independent fast endocytosis (6) or by clathrin-dependent slow endocytosis (6′). The newly regenerated SVs are immediately refilled with neurotransmitters (7) or, in some case, they undergo fusion steps with early endosomes (7′).

which SVs are generated from early endosomal membranes at the presynaptic terminal is dependent on AP-3. The latter pathway does not seem to account for the majority of SVs since there are no remarkable alterations in SV morphology and numbers in the *mocha* mouse, which lacks functional AP-3. How, then, can SV proteins, but not plasma membrane residents, be recruited into newly generated SVs? Several amino acid sequence motifs or molecular determinants for each SV protein have been proposed to

explain this mechanism. First, a dileucine motif located in the cytoplasmic tails of vesicular transporters (VMAT2, VAChT, and VGLUT1) is important for precise sorting and fast recruitment of those SV proteins. A VAChT mutant which lacks the dileucine motif was trapped at the plasma membrane of the cell body of the differentiated cholinergic SN56 cell lines, and the targeting of the mutant protein to neurites and varicosities (which resemble axon terminals) was dramatically reduced, indicating that the SV precursor cargo carrying VAChT is not directly transported through the axon but is delivered to the plasma membrane of the cell body before being transported along the axon. Second, the intravesicular N-glycosylation of synaptotagmin 1 is essential for its vesicular targeting. Interestingly, synaptotagmin 7, one of the synaptotagmin family members which resides preferentially on the plasma membrane, is targeted to vesicles when the intravesicular portion is replaced with a portion containing the N-glycosylation site of synaptotagmin 1. The mechanistic basis and involvement of other factors is not clear, but the interaction of glycoresidues of synaptotagmin 1 with other proteins at the cell surface may govern the targeting of synaptotagmin 1 to vesicles.

It is likely that multiple factors, including protein–protein interactions and glycoresidue–protein interactions, modulate the sorting of SV proteins. The general principle of the segregation of SV proteins from plasma membrane proteins during SV biosynthesis has not been clarified.

The Synaptic Vesicle Cycle

At the presynaptic terminals, SVs are not of 'single use.' They are regenerated at the terminals independent from protein synthesis in the cell body. The SV cycle can be outlined, with the uptake of neurotransmitter into SVs as a first step (**Figure 2(b)**, step 1). Neurotransmitters in the central nervous system (CNS) are synthesized locally in the cytoplasm of the presynaptic terminals and are actively transported into SVs. Away from the plasma membrane, the majority of neurotransmitter-filled SVs are either diffusively floating in the cytoplasm or tethered with cytoskeleton components such as actin and spectrin (step 2). For exocytosis, SVs must come into physical contact with the plasma membrane (docking; step 3). SVs do not evenly dock to the whole area of the presynaptic plasma membrane but, rather, to a restricted area called the active zone. There, large cytomatrix proteins, such as bassoon, piccolo, and munc13, form huge protein complexes that appear as electron-dense structures on electron micrographs. Docked SVs are then transformed into fusion-competent SVs via a process called priming (step 4). As soon as an electrical stimulus arrives at the terminal, voltage-dependent calcium channels at the active zone open, resulting in a rapid and local increase in Ca^{2+} concentration. Ca^{2+} ions trigger the fusion of the SV membrane with the plasma membrane in less than 100 μs. Since other exocytotic reactions (i.e., hormone secretion from endocrine cells) take much longer (seconds to minutes), there must be unique factors present only in neurons to perform such a rapid membrane fusion reaction.

After exocytosis, SV components that are incorporated into the plasma membrane are retrieved to form a new SV by endocytosis. There are at least two kinetically distinct modes of endocytosis. The time constants of the fast and slow phase are approximately 1 and 10 s, respectively. Whereas the molecular machinery for the fast phase is not well understood, the slow phase is mediated by the formation of clathrin-coated pits. In both modes, GTP hydrolysis by the GTPase protein dynamin is indispensable for the fission of the invaginated membranes of newly formed SVs. Although various endocytosis-related proteins, such as AP-2, endophilin, amphiphysin, and synaptojanin, have been identified and implicated in controlling endocytosis, their precise roles are a matter of intense research.

The reformed SVs then either recycle back and are refilled with neurotransmitters (step 1) or pass through the early endosomal intermediates before recycling back to step 1. An alternative pathway has been proposed that is similar to the exocytosis of secretory granules; that is, SVs do not fully collapse with the plasma membrane upon fusion but instead form a narrow and transient fusion pore which does not allow a complete discharge of neurotransmitter content. As soon as the pore closes, the half-empty SV can either be immediately engaged in another round of exocytosis or go back to step 1. The existence of such a rapid recycling mode in the CNS, termed the 'kiss-and-run' mechanism, is under debate.

The SVs clustering at the presynaptic terminals can be divided into two functional pools. The first pool contains a small fraction of SVs (5–10% of the total SVs at the presynaptic terminals) that can be released rapidly by a brief high-frequency train of action potentials or by stimulation with hypertonic solution. This pool is thought to be release-ready and is therefore referred as to the readily releasable pool (RRP). The second pool, the reserve pool (RP), represents a vesicle fraction that does not immediate participate in exocytosis. Instead of participating in exocytosis, the RP vesicles replenish the RRP pool after the RRP vesicles undergo exocytosis. Both the amount of the RRP and the rate of replenishment of RRP with RP are critical parameters to determine the availability of vesicles for exocytosis, thereby affecting the

characteristics of short- and long-term plasticity of a given neuron. Classically, the RRP was related to a fraction of morphologically docked vesicles at the plasma membrane and the RP was thought to be spatially distant from the plasma membrane. However, one study suggested that the RRP vesicles do not necessarily correlate with the morphologically docked vesicles but are randomly distributed in the SV cluster at the terminal, indicating that there are no correlations between the anatomical locations of SVs and their functional fusion competence. Mechanisms underlying how the mobility and the fate of an individual SV can be molecularly defined are uncertain.

Molecular Composition of Synaptic Vesicles

With respect to SV functions as described previously, SVs should be equipped with two classes of essential protein components: transport proteins involved in neurotransmitter uptake into SVs and membrane trafficking proteins involved in the regulation of vesicle cycle and exo-endocytosis. A large body of work has focused on identifying protein components on SVs, leading to an almost complete identification of SV proteins. The stoichiometry of SV proteins may be flexible because many cytoplasmic proteins are temporarily attached to and detached from SVs depending on the status of a specific SV during recycling. On the other hand, a basic set of essential proteins must be present on all SVs for vesicle functions, and the numbers of individual proteins and the proportions within an SV might proteins within an SV to be maintained (Figure 3(a)). Furthermore, immunohistochemical studies on SV proteins have demonstrated that most of the major SV proteins consist of multiple isoforms whose expressions in the CNS are partially overlapping or mutually segregated, the combinations of which would create heterogeneity in protein composition in each SV. In contrast to the identification of SV proteins which became relatively handy with advancements in mass spectroscopy, understanding the mechanistic function(s) of each protein has been not so trivial, mainly because of technical limitations with regard to manipulating and measuring the intracellular events with high temporal and spatial resolution. The following sections introduce some of the essential and abundant SV proteins, which are stoichiometric components with at least one copy per SV.

Proteins for Neurotransmitter Uptake

The accumulation of neurotransmitters in SVs is driven by a proton electrochemical gradient which is generated by a vacuolar-type proton ATPase. The proton pump consists of at least 13 subunits and is the largest functional protein complex in the SV membrane. It contains two functional units – a larger peripheral protein complex (V_1) which catalyzes ATP hydrolysis and an integral membrane protein complex (V_o) which builds up a ring structure in the membrane and mediates proton translocation. The molecular weight of the entire complex is approximately 800 kDa; therefore, a single complex accounts for approximately 10% of the total SV protein. An SV contains a single copy of the proton pump complex, which may be sufficient to energize neurotransmitter uptake into the SVs.

With the aid of a proton electrochemical gradient, vesicular transporters specific for neurotransmitter-type mediate neurotransmitter uptake into the vesicles. There are four distinct uptake systems for neurotransmitters in the CNS. Three isoforms of vesicular glutamate transporters (VGLUT1–3) transport glutamate into SVs. GABA and glycine share the same transporter, the vesicular inhibitory amino acid transporter (VIAAT; initially called vesicular GABA transporter (VGAT)). Two monoamine transporters (VMAT1 and VMAT2) transport all biogenic amines, with VMAT2 preferentially expressed in the brain. The uptake of acetylcholine is mediated by the vesicular acetylcholine transporter VAChT. All four transporter families belong to the solute carrier protein family, but there are no sequence homologies among the transporters for different neurotransmitter types. There are differences in energetics of the transport; some of them preferentially utilize the membrane potential (VGLUTs) and others use the pH gradient (VMATs and VAChT) as the main driving force. The expression of a particular transporter in the SV membrane is the ultimate determinant for the type of neurotransmitter which is released from a given neuron. Moreover, there are indications that the expression level of the transporter per SV modulates the amount of neurotransmitters accumulated in an SV, thereby regulating quantal size.

In addition to the neurotransmitter transporters discussed previously, SVs contain a chloride channel to facilitate the acidification of SVs. The voltage-dependent chloride channel ClC-3 was proposed to confer this activity on SVs. Although several channel activities, such as a cation selective channel, have been demonstrated by electrophysiological methods on reconstituted systems, the molecular identities of the activities have been elusive. The SV2 protein family and tetraspan vesicle membrane proteins (synaptophysin, synaptogyrin, and SCAMPs) have been identified as SV-specific proteins, and their predicted protein structures suggest their role as a transporter or a channel.

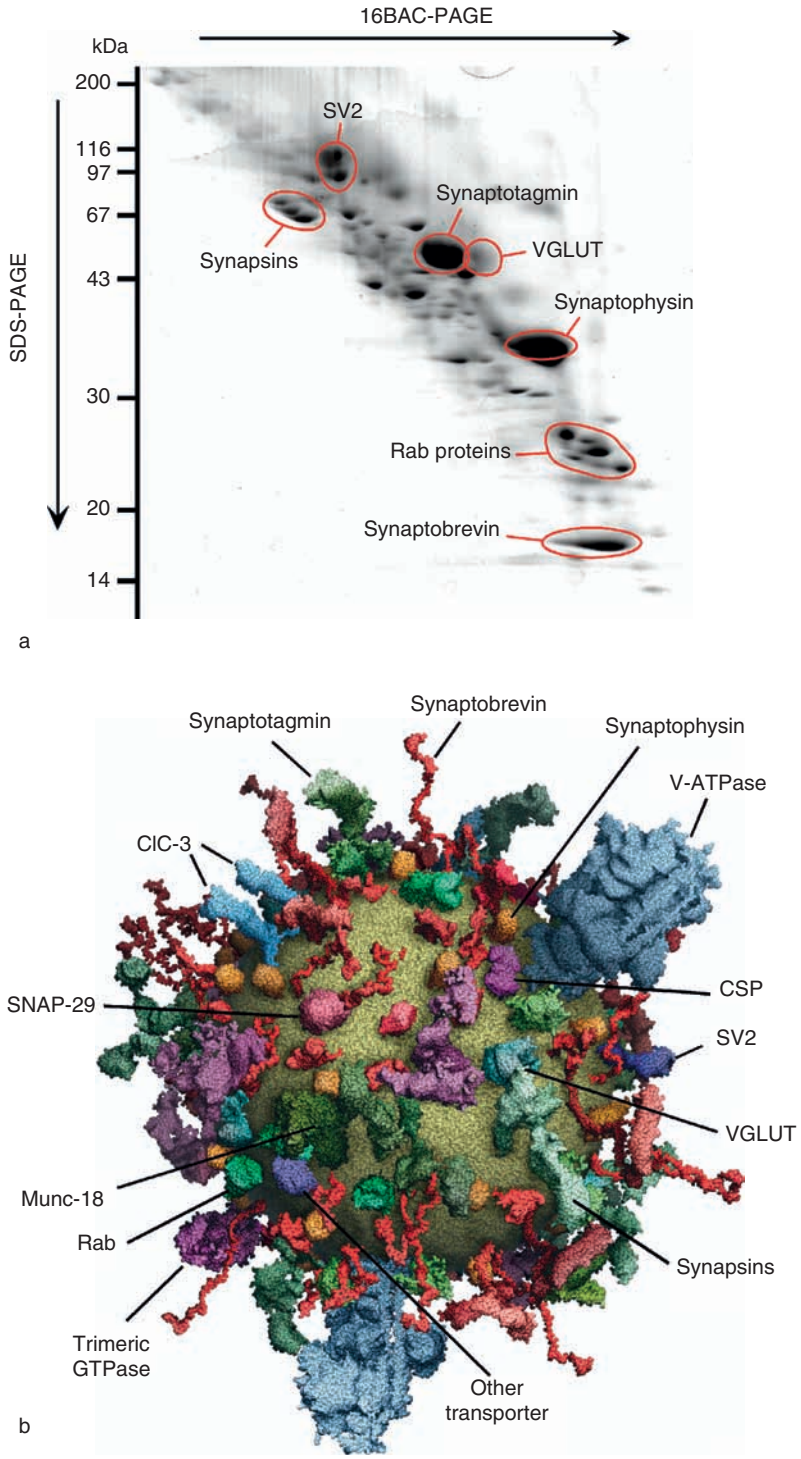

Figure 3 Molecular composition of SV. (a) Protein composition of SVs visualized by 16-BAC/SDS two-dimensional gel electrophoresis. A total of 500 μg of SV proteins was applied. The spots containing the major SV proteins are circled. (b) Structural model of an average SV. Based on quantitative measurements, a three-dimensional model of an average SV was constructed. An average SV contains 70 copies of synaptobrevin, 30 copies of synaptophysin, 10 copies of neurotransmitter transporter (VGLUT), 8 copies of synapsin, 15 copies of synaptotagmin, 25 copies of Rab GTPase, and 1 or 2 copies of SV2, synaptogyrin, SCAMP, and V-ATPase. The numbers of phospholipids and cholesterol are estimated to be approximately 7000 and 6000, respectively.

However, no clear transport activities associated with these proteins have been demonstrated. It should be noted that although not related to an intrinsic function of SV2s, they have been shown to function as a protein receptor for botulinum neurotoxin A.

Proteins for Membrane Trafficking

Neural exocytosis is a tightly regulated process elicited by a rapid increase in Ca^{2+} concentration. Since Ca^{2+} ions induce neurotransmitter release within 100 μs, a series of complex enzymatic reactions could not be expected for exocytosis within the short time frame. Thus, SVs that undergo exocytosis not only should be docked but also primed for membrane fusion, waiting for the final stimulus for exocytosis. Two abundant integral membrane proteins on SVs, synaptobrevin (also referred as to VAMP2) and synaptotagmin 1, have been established as the most important players in Ca^{2+}-triggered exocytosis. Furthermore, SVs contain two abundant peripheral protein families, rab-GTPases and synapsins, which are implicated in the modulation of vesicle pools.

Synaptobrevin is a small integral membrane protein that contains a SNARE motif and is thus called a v-SNARE (vesicular SNARE, also referred to as R-SNARE because of a conserved arginine in the middle of the SNARE motif). Synaptobrevin forms a tight four-helix bundle, the SNARE complex, with two other SNARE proteins at the plasma membrane – syntaxin 1 and SNAP-25 (t-SNAREs named after target-SNAREs; also referred as to Q-SNARE, which has a conserved glutamine in the middle of the SNARE motif). The SNARE hypothesis regarding membrane fusion in general proposes that the assembly and tight complex formation of appropriate pairs of SNARE proteins pulls the two membranes close together so that the two membranes become competent for fusion. Analogously, SVs become 'primed' when synaptobrevin forms the SNARE complex with syntaxin 1 and SNAP-25. The importance of SNARE proteins in neuronal exocytosis is emphasized by the fact that the clostridial neurotoxins, a family of metalloproteases which specifically cleave the neuronal SNARE proteins, abolish Ca^{2+}-dependent neurotransmitter release while the docked vesicles remain unchanged by the toxin treatment. The formation of the SNARE complex is affected by other binding factors of the SNARE proteins. For instance, it has been proposed that Munc18 protein binds to syntaxin, keeping the syntaxin molecule in a close conformation and thereby preventing syntaxin from forming the SNARE complex. Biochemical experiments using the SNARE-reconstituted proteoliposomes suggest that the SNARE complex indeed mediates membrane fusion. However, the liposome fusion with SNAREs alone is slow and spontaneously occurs without any regulation, indicating that other factors might play a role in synchronizing the fusion reaction.

Synaptotagmin 1 is known as a vesicular Ca^{2+} sensor for exocytosis that accomplishes rapid membrane fusion in response to Ca^{2+} ions. It has an N-terminal highly glycosylated intravesicular domain and two cytoplasmic C_2 domains that interact with Ca^{2+} ions. Ca^{2+}-triggered exocytosis in synaptotagmin 1-deficient neurons is severely impaired, indicating its role as the Ca^{2+} sensor for exocytosis. The affinity of C_2 domains of synaptotagmin 1 for Ca^{2+} ions is very low, but it increases dramatically (up to 0.5–5.0 μM) when the C_2 domains bind to phospholipids. Its apparent affinity to Ca^{2+} meets the requirement that it should have a Ca^{2+} concentration sufficient for triggering exocytosis *in vivo* (in the calyx of Held or in chromaffin cells). However, investigations of synaptotagmin 1-deficient neurons and chromaffin cells have suggested that synaptotagmin 1 is a Ca^{2+} sensor only for the fast, synchronous exocytosis and not for the slow, asynchronous component. It is not clear if other synaptotagmin isoforms (at least 15 members in mammals which differ in affinity to Ca^{2+} and subcellular localizations) function as a Ca^{2+} sensor for the latter.

Rab proteins are known to be involved in a variety of intracellular membrane trafficking events. Among the Rabs, Rab3A, Rab5, and Rab11 are abundant isoforms on SVs. They associate with SVs in the GTP-bound forms, whereas they dissociate from SVs in the GDP-bound state. The association and dissociation cycle of Rab3A occurs in parallel with SV exocytosis, although Rab proteins are not directly involved in exocytosis or membrane fusion. In Rab3A-deficient neurons, some forms of synaptic plasticity, such as long-term potentiation in hippocampal mossy fibers, are affected, indicating that Rab3A modulates the efficacy of neurotransmission, probably by changing the vesicle pools available for exocytosis.

In addition to Rab proteins, the synapsin family represents abundant peripheral membrane proteins which may interact with SVs in an activity-dependent manner. The synapsin family consists of five isoforms (I/IIA, IIB, IIIA, and IIIB) and they form homo- or heterodimer on the surface of SVs. Originally, it was discovered as a neuron-specific abundant protein substrate for the cyclic AMP-dependent protein kinase (protein kinase A) at presynaptic terminals. Their function in synaptic transmission remains unclear. It has been proposed that synapsins function as an anchoring protein to cytoskeleton components to make a portion of vesicles immobile because they exhibit a binding property to actin, tubulin, and spectrin. Such a view is partially supported by the fact that although

Ca^{2+}-triggered exocytosis is intact, short-term synaptic plasticity is impaired in the absence of synapsin 1 and 2. Therefore, synapsins might play a modulating role in controlling the vesicle pools at the terminals.

Conclusion

Although significant progress has been made in understanding molecular mechanisms of neurotransmitter release, a mechanistic understanding of SVs is yet to be achieved. Until recently, even elementary quantitative information about SV components was lacking. Quantitative analyses of SV proteins and lipids have allowed, for the first time, the proposal of a structural model of an average SV shown in Figure 3(b). The model demonstrates that SVs are highly decorated with proteins. Numerous copies of the essential proteins for exocytosis, such as synaptobrevin and synaptotagmin, are present on a vesicle indicating that the surface densities of these proteins are not rate limiting for fusion. Furthermore, SVs contain large numbers of the neurotransmitter transporter, supporting a high-speed refilling of the vesicles upon strong repetitive stimulations. One exception among the essential proteins is the V-ATPase, which is estimated to have one or two copies per vesicle, further indicating that the retrieval of SV constituents by endocytosis has to be guaranteed by precise processes. More studies are needed to explore the structure–function relationships of each SV constituent to understand the function of this intriguing organelle.

See also: Active Zone; Endocytosis: Kiss and Run; Exocytosis: Ca^{2+}-Sensitivity; Fusion Pore; Large Dense Core Vesicles (LDCVs); NSF and SNAPs; Rab3; SNAREs; Synapsins; Synaptotagmins; Vesicle Pools; Vesicular Neurotransmitter Transporters.

Further Reading

Becherer U and Rettig J (2006) Vesicle pools, docking, priming and release. *Cell and Tissue Research* 326: 393–407.

Bonanomi D, Benfenati F, and Valtorta F (2006) Protein sorting in the synaptic vesicle life cycle. *Progress in Neurobiology* 80: 177–217.

Fernandez-Chacon R and Sudhof TC (1999) Genetics of synaptic vesicle function: Toward the complete functional anatomy of an organelle. *Annual Review of Physiology* 61: 753–776.

Jahn R (2004) Principles of exocytosis and membrane fusion. *Annals of the New York Academy of Sciences* 1014: 170–178.

Montecucco C, Schiavo G, and Pantano S (2005) SNARE complexes and neuroexocytosis: How many, how close? *Trends in Biochemical Sciences* 30: 367–372.

Rissoli SO and Betz WJ (2005) Synaptic vesicle pools. *Nature Reviews Neuroscience* 6: 57–69.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Takamori S, Holt M, Stenius K, et al. (2006) Molecular anatomy of a trafficking organelle. *Cell* 127: 831–846.

Synaptojanin

B Chang-Ileto and G Di Paolo, Columbia University Medical Center, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Optimal neurotransmission relies on the proper trafficking of synaptic vesicles in nerve terminals. At the beginning of the cycle, synaptic vesicles dock at the active zone of nerve terminals where they undergo a priming step that makes them competent for Ca^{2+}-dependent fusion. When action potentials reach nerve terminals, the entry of Ca^{2+} triggers the fusion of synaptic vesicles with the plasma membrane, resulting in the release of neurotransmitters into the synaptic cleft, which in turn stimulate postsynaptic receptors. Following exocytosis, the synaptic vesicle membrane is rapidly retrieved from the cell surface and recycled to generate new synaptic vesicles. This process is crucial to prevent plasma membrane expansion and synaptic vesicle depletion, both of which would impair neurotransmission. A major recycling pathway is clathrin-mediated endocytosis, although faster recycling parallel pathways (e.g., 'kiss-and-run') have been described. After pinching off from the plasma membrane, clathrin-coated vesicles rapidly shed their coat and are directed back to the synaptic vesicle cluster, where they are refilled and can participate in new rounds of release. The recycling machinery includes coat components as well as a variety of accessory proteins, such as membrane deforming factors, scaffolding molecules, and signaling components. In addition, a variety of enzymes modifying the lipid content of membranes have been implicated in the recycling process. The best-characterized family of lipid enzymes involved in membrane trafficking at the synapse is that of synaptojanins, the focus of this article. These proteins largely function as phosphatases that dephosphorylate phosphoinositides (i.e., phosphorylated derivatives of phosphatidylinositol). These phospholipids are major membrane-bound signaling molecules that regulate intracellular organelle transport, cytoskeletal dynamics, the function of ion channels and transporters, as well as signal transduction mechanisms.

Structure of Synaptojanins

Synaptojanins constitute a protein family that is conserved from yeast to humans. Whereas yeast expresses three synaptojanin-like genes (*Sjlp1, -2,* and *-3* or *INP51, -52,* and *-53*), the worm and fly genomes contain only one synaptojanin gene. The mammalian genome contains two genes, *Synj1* and *Synj2* (Figure 1). All synaptojanins consist of a central inositol 5-phosphatase domain, which is flanked by an NH_2-terminal Sac1 domain and a COOH-terminal proline-rich region (Figure 1). Both inositol 5-phosphatase and Sac1 domains are highly conserved within the family and possess polyphosphoinositide phosphatase activity, whereas the variable proline-rich tail serves as a targeting domain. The term 'synaptojanin' was derived from the two-headed Roman god Janus to reflect the unique properties of synaptojanins (among the family of inositol 5-phosphatases), namely the presence of two domains involved in inositol phosphate metabolism.

Based on genetic studies in independent organisms, a fundamental function of synaptojanins is to catalyze the dephosphorylation of phosphatidylinositol-4,5-bisphosphate [PtdIns(4,5)P2]. Thus, the inositol 5-phosphatase domain is likely the business part of the molecule, though the Sac1 domain has also been shown to play a role in some of synaptojanin's functions. Currently, structural information on synaptojanins is only available for this catalytic domain. Although the primary sequence of the 5-phosphatase catalytic domain shares no resemblance to any other classes of enzymes, the catalytic residues are highly conserved with those of nucleases, particularly the alkaline phosphatase (AP) endonucleases, and thus the structures of the catalytic domains were predicted to be similar. The inositol 5-phosphatase catalytic domain from the *Schizosaccharomyces pombe* synaptojanin ortholog, SPsynaptojanin, was crystallized in both the apo- and product-bound states, revealing that not only is the positioning of the critical catalytic residues of the 5-phosphatases and AP endonucleases similar but also the overall structure is similar. The SPsynaptojanin catalytic domain structure reveals a center composed of two β sheets, positioned like the slices of bread in a sandwich, flanked on either side by several α helices (Figure 2). The active site is formed by the funnel-shaped depression located at one end of the β sheet sandwich. Mutagenesis targets suggested by the structure revealed that the surface composed of loops from β sheet II that surrounds the active site is involved in specificity toward lipid and/or soluble inositol phosphates and that the α_7–β_9 loop determines dephosphorylation site selectivity (Figure 2).

Expression, Localization, and Interacting Partners of Synaptojanins

Of the two genes encoding mammalian synaptojanins, *Synj1* is the best characterized. *Synj1* occurs as

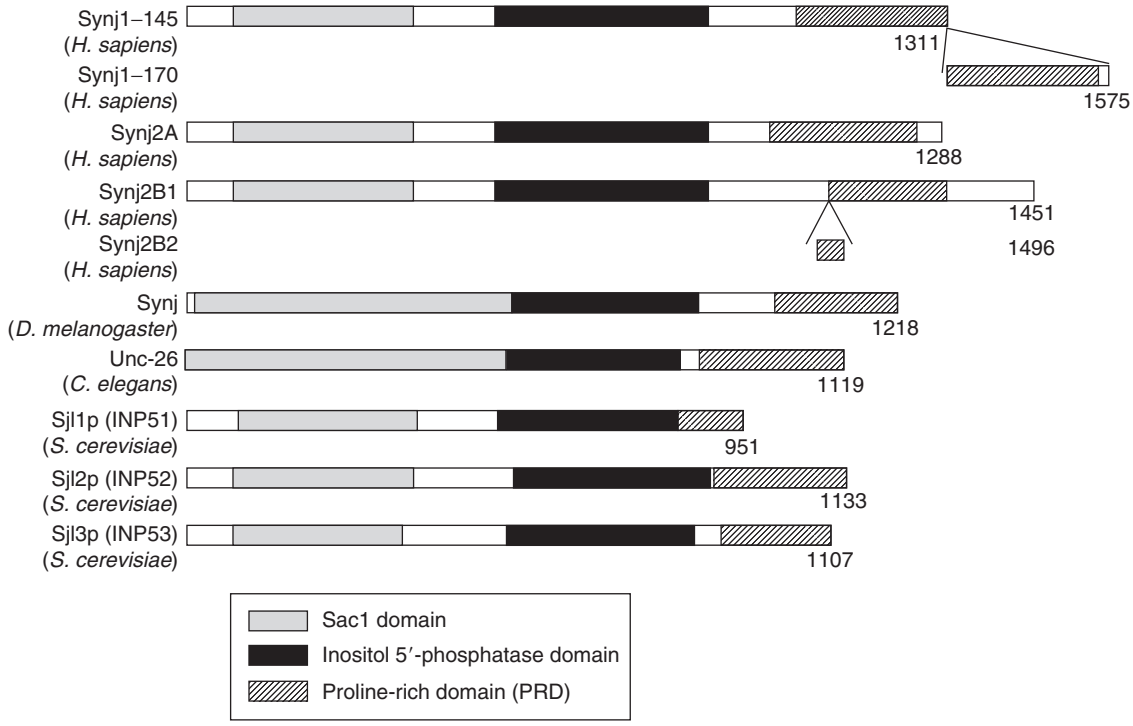

Figure 1 Domain structure of synaptojanins. The domain structure of the synaptojanin proteins is conserved from yeast to humans. The NH_2-terminus contains a Sac1 domain, followed by a central inositol 5-phosphatase domain and a COOH-terminal proline-rich region (PRD). Human Synj1 has two major isoforms, Synj1-145 and Synj1-170. The Synj1 gene is transcribed into a cDNA containing two open reading frames (ORFs) separated by an in-frame stop codon; the first ORF is translated into the 145-kDa product, Synj1-145, whereas Synj1-170 is an alternative splice product from both ORFs that splices out the in-frame stop codon. An additional alternatively spliced product of rat Synj1 yields Synj1-145 variant with an additional exon of 16 amino acids (not shown). The human Synj2 gene encodes two alternative splice forms – Synj2A, which is localized to mitochondria, and Synj2B, which more closely resembles Synj1 proteins. Synj2B is further alternatively spliced into Synj2B1 and Synj2B2. Both flies and worms possess only one synaptojanin gene. Protein sequences were obtained from the Swiss-Prot database.

two main isoforms – a 145-kDa protein (Synj1-145) and a 170-kDa protein (Synj1-170). These two isoforms are expressed in various tissues; Synj1-145 is heavily enriched in the adult brain, whereas Synj1-170 is predominantly found in the developing brain as well as in nonneuronal cells. Synj1 was originally isolated as a major Grb2-binding protein via the interaction of its PRD domain with the Grb2–SH3 domain. In addition, the Synj1 PRD domain has been shown to interact with other SH3-containing proteins involved in endocytic processes, such as endophilin and amphiphysin. The two main Synj1 isoforms arise from alternative splicing of its mRNA that contains two open reading frames (ORFs) separated by an in-frame stop codon. Synj1-145 is the protein product of the first ORF, whereas Synj1-170 is the product when the stop codon is spliced out and both ORFs are translated. Thus, in addition to the Synj1-145 sequence, Synj1-170 possesses a COOH-terminal extension with unique properties. Importantly, it contains NPF repeats, a motif involved in binding to the EH domains of endocytic protein Eps15. This same COOH-terminal extension of Synj1-170 has also been shown to bind to clathrin and to have three distinct AP-2 α-subunit binding motifs – DPF, FXDXF, and WXXF – that help coordinate the precise temporal recruitment of this and other endocytic accessory proteins to clathrin coats. Although the subcellular localization of endogenous Synj1-170 has not been clearly determined, transfection studies suggest that the protein is localized to endocytic pits, consistent with its ability to bind to the clathrin coat. In contrast, the localization of adult brain-enriched isoform Synj1-145 is well characterized and has been shown to coincide with mature nerve terminals in various regions, such as the hippocampus (Figure 3(a)), the granule cell layer of the cerebellum, and the brain stem. Based on immunogold electron microscopic analysis of broken nerve terminals, Synj1-145 appears to be concentrated at clathrin- and dynamin-coated endocytic intermediates, consistent with a role in endocytosis (Figure 3(b)).

Similar to Synj1-170, the various isoforms of Synj2 exhibit a wide tissue distribution. Subcellular fractionation shows that Synj2 is predominantly in the particulate fraction. The *Synj2* gene has three alternatively spliced isoforms – Synj2A, Synj2B1, and Synj2B2 – and their localization is dependent on the activity of

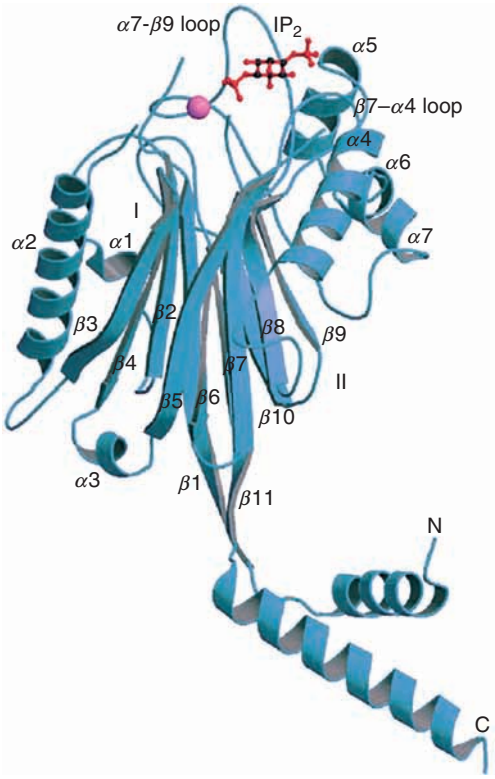

Figure 2 Crystal structure of SPSynj1. A ribbon diagram of the SPsynaptojanin inositol 5-phosphatase catalytic domain with bound Ca^{2+} and Ins(1,4)P_2 molecules. The structure reveals a center composed of two β sheets, positioned like two slices of bread in a sandwich, flanked on either side by several α helices. The active site is formed by the funnel-shaped depression located at one end of the β sheet sandwich. The active site was determined by soaking the apo crystals in a solution containing the substrate Ins(1,4,5)P_3 and 1 mM $CaCl_2$. The electron density showed that the reaction product, Ins(1,4)P_2, was bound to the molecule instead of Ins(1,4,5)P_3, and structures obtained with a shortened substrate incubation yielded a complex electron density map that was consistent with having a mixture of bound Ins(1,4,5)P_3 and Ins(1,4)P_2. Reprinted from *Cell*, Vol. 105, Tsujishita Y, Guo S, Stolz LE , York JD, and Hurley JH, Specificity determinants in phosphoinositide dephosphorylation: Crystal structure of an archetypal inositol polyphosphate 5-phosphatase, pp. 379–389, Copyright 2001, with permission from Elsevier.

the Rac1 GTPase. The COOH-terminus of Synj2A interacts with the PDZ domain of the mitochondrial outer membrane protein OMP25, thereby localizing this isoform to mitochondria. Synj2B, which undergoes further alternative splicing, more closely resembles the Synj1 proteins, with the longer isoform, Synj2B2, having binding sites for both endophilin and amphiphysin and the shorter isoform, Synj2B1, having a binding site only for amphiphysin. Synj2B has been localized to both nerve terminals and the spermatide manchettes in testes. The overlapping biochemical properties and localization of Synj2B and Synj1 may account for some of the residual synaptic vesicle recycling capability observed in Synj1-deficient neurons.

In worms and flies, the synaptojanin ortholog is enriched in nerve terminals, where its localization is controlled by endophilin. Based on this feature as well as analyses of synaptojanin function in these two organisms, it appears that these orthologs may be closely related to mammalian Synj1-145.

Substrate Specificity of Synaptojanins

Synaptojanins are part of the inositol 5-phosphatase family of enzymes, which catalyze the hydrolysis of phosphate groups attached to the 5′ position of the inositol ring in either soluble or lipid-bound inositol phosphates. Whereas the central inositol 5-phosphatase domain removes the phosphate from the 5′ position of PtdIns(4,5)P_2 and PtdIns(3,4,5)P_3, the NH2-terminal Sac1 domain can dephosphorylate a variety of phosphoinositides, such as PtdIns(3)P, PtdIns (4)P, PtdIns(5)P, and PtdIns(3,5)P_2. The 5-phosphatase domain can also dephosphorylate (soluble) inositol polyphosphates on the 5′ position, such as the PI-phospholipase C enzymatic reaction product inositol-1,4,5-trisphosphate [Ins(1,4,5)P_3]. Based on lipid analyses carried out on brain tissue and cultured neurons from Synj1 knockout mice, the main physiological substrate appears to be PtdIns(4,5)P_2, thereby denoting a dominant role for the inositol 5-phosphatase domain. The physiological substrates for the Sac1 domain are unknown, although a study on yeast synaptojanin-like protein 2 (Sjlp1/INP52) suggests that at least one of these substrates may be PtdIn(3)P on endosomal membranes. However, the Sac domain of synaptojanins may function redundantly with that of other Sac domain-containing proteins as well as with other 3-phosphatases, such as myotubularins. Whether synaptojanins can physiologically dephosphorylate inositol polyphosphates, such as Ins(1,4,5)P3, is unclear.

Regulation of Synaptojanin Function

Synaptojanins are soluble (cytosolic) lipid enzymes and thus their recruitment to membranes largely relies on their interaction with other proteins. Generally, the targeting of synaptojanins to cell membranes occurs through an interaction of their COOH-terminal proline-rich tail with a variety of binding partners, specifically with Src homology 3 (SH3) domain-containing proteins. Although synaptojanin 1 was originally identified through its affinity to the SH3 domain of Grb2, it appears that the main physiological interactors of synaptojanin 1 may be endophilin A family members. Other interactors include amphiphysin, syndapin, intersectin, and sorting nexin 9, although the precise roles of such interactions are unclear. Importantly, synaptojanins possess multiple independent binding sites for SH3 proteins via

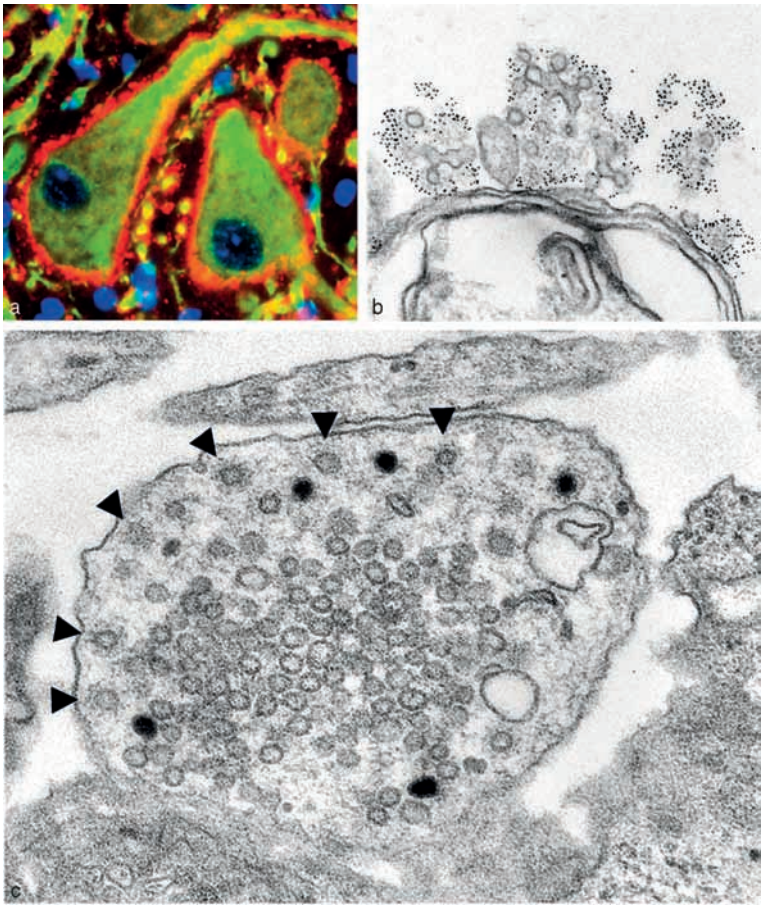

Figure 3 Localization of Synaptojanin 1 and consequence of its ablation. (a) Synj1-145 has been shown to be localized at nerve terminals in various regions of the adult brain, including the hippocampus, cerebellum and brain stem. Anti-Synj1 immunoreactivity is visualized with red fluorescence displaying the specific localization of this protein in nerve terminals in the deep cerebellar nuclei of rat brain. Preparations were counterstained with the dendritic marker MAP-2 (green) and nuclear marker (DAPI in blue). Reprinted by permission from Macmillian Publishers Ltd: *Nature* (McPherson PS, Garcia EP, Slepnev VI et al. (1996) A presynaptic inositol-5-phosphatase. *Nature* 353–357), copyright 1996. (b) Ultrastructural studies using immunogold electron microscopy have demonstrated that Synj1 can be found at clathrin- and dynamin-coated pits. Here, anti-synaptojanin polyclonal antibody conjugated with gold particles label gently homogenized rat retina. Reprinted by permission of Federation of the European Biochemical Societies from Synaptojanin 1: Localization on coated endocytic intermediates in nerve terminals and interaction of its 170 kDa isoform with Eps15, by Haffner C, Takei K, Chen H, et al. *FEBS Letters*, vol. 419, pp. 175–180, copyright 1997. (c) Accumulation of clathrin-coated vesicles (arrowheads) in Synj1 knockout mouse cortical neurons is observed by electron microscopy. Reproduced from *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, Kim WT , Chang S, Daniell L, Cremona O, Di Paolo, G, and De Camilli P, Delayed reentry of recycling vesicles into the fusion-competent synaptic vesicle pool in synaptojanin 1 knockout mice, pp. 17143–17148, copyright 2002, National Academy of Sciences, USA.

sequences that contain variations of the canonic proline-rich peptides that determine their affinity for the respective SH3 proteins. For instance, synaptojanin 1 contains independent binding sites for endophilin and amphiphysin, and the alternatively spliced isoforms of synaptojanin 2 possess binding sites for amphiphysin, either alone or in combination with those for endophilin (Figure 1).

The function of synaptojanin 1 is largely controlled by a cycle of phosphorylation/dephosphorylation. This mechanism both modulates its intrinsic catalytic activity and regulates its interaction with SH3 proteins, which in turn mediate the recruitment of synaptojanin to membranes. Synaptojanin 1 belongs to the 'dephosphin' group of proteins, which are constitutively phosphorylated in resting nerve terminals and undergo depolarization-dependent dephosphorylation. This process allows for efficient spatiotemporal control of protein–protein interactions to occur between dephosphins (e.g., dynamin and amphiphysin) as well as other endocytic proteins (e.g., endophilin) and mediates their recruitment to sites of endocytosis. After a cycle of exocytosis–endocytosis has occurred, dephosphins are rephosphorylated during repolarization to allow the next round of exocytosis–endocytosis to occur upon the next stimulation.

Although pharmacological studies indicate that the dephosphorylation of dephosphins, including synaptojanin 1, is largely mediated by the Ca^{2+}-dependent protein phosphatase calcineurin, various protein kinases have been shown to be responsible for their phosphorylation. A key enzyme appears to be cyclin-dependent kinase 5 (Cdk5), which is a proline-directed Ser/Thr kinase that is highly expressed in neurons and in nerve terminals. This kinase has been implicated in the phosphorylation of dynamin, amphiphysin, and Synj1. The 5-phosphatase activity of Synj1 was shown to be reduced as a result of Cdk5-mediated phosphorylation, and Cdk5 inhibitors, which block the phosphorylation of Synj1, were able to prevent this reduction in Synj1 catalytic activity. Ser1144 was mapped as the site of phosphorylation by Cdk5, although additional sites may be present. In addition, Cdk5 phosphorylation of Synj1 decreased its ability to interact with both endophilin 1 and amphiphysin 1, suggesting that its membrane recruitment to sites of endocytosis is regulated by phosphorylation.

Another kinase that is potentially involved in the Ser/Thr phosphorylation of synaptojanin is protein kinase C (PKC). Inhibitors of PKC function, Ro31-8220 and Go7874, were shown to obstruct both dynamin 1 and synaptojanin rephosphorylation upon nerve terminal repolarization, thereby preventing endocytosis after a second stimulus. However, the sites of PKC phosphorylation on dynamin 1 found *in vitro* were shown not be phosphorylated *in vivo* from material isolated from intact synaptosomes. The sites of PKC phosphorylation on synaptojanin and whether these sites are physiologically relevant in synaptic vesicle recycling remain to be determined.

Minibrain kinase/dual-specificity tyrosine phosphorylation-regulated kinase (Mnb/Dyrk1A) is a candidate Ser/Thr kinase implicated in the regulation of endocytic proteins, including Synj1. Mnb/Dyrk1A is found within the Down syndrome critical region of human chromosome 21 and its expression is elevated in Down syndrome patients. This proline-directed kinase has been shown to phosphorylate both amphiphysin and dynamin. It has also been suggested to phosphorylate Synj1, affecting its ability to bind amphiphysin and intersectin with minimal impact on its catalytic activity.

In addition to Ser/Thr phosphorylation, tyrosine phosphorylation of Synj1 may play a critical role, as suggested by a study of the EphB2 receptor signaling pathway. This receptor tyrosine kinase, whose ligand is ephrinB2, is believed to function in modulating the actin cytoskeleton resulting in changes in cell morphology. Synaptojanin 1 was identified as a substrate in a screen of proteins phosphorylated after treatment of cells with ephrinB2. Tyrosine phosphorylation of synaptojanin appears to play an important regulatory role, blocking its ability to bind endophilin and affecting its catalytic activity. A tyrosine phosphatase for synaptojanin has yet to be identified in this process.

Role of Synaptojanins in Neurons

The identification of synaptojanin 1 as a partner for the SH3 domain of Grb2 (and amphiphysin), along with the endocytic fission factor dynamin, indirectly suggested an involvement of this enzyme in endocytosis. However, conclusive evidence for a direct role of synaptojanin in this process derived from a genetic study in the mouse, in which a null mutant was shown to produce defects in synaptic vesicle recycling. At the $Synj1^{-/-}$ synapse, increased PtdIns(4,5)P2 levels correlated with an excess of clathrin-coated vesicles, suggesting a delay in the uncoating reaction (**Figure 3c**). In agreement with defects of recycling, a deeper depression of neurotransmission was observed in hippocampal slices as well as in primary cortical cultures from null mice during a prolonged stimulation. Fluorescent dye (FM1-43) uptake and release assays in cultured neurons confirmed that newly endocytosed vesicles recycle with slower kinetics in nerve terminals, likely reflecting the accumulation of clathrin-coated vesicles (**Figure 3(c)**). Strikingly, endocytic defects resulting from inactivation of synaptojanin function were also observed in other species, such as the budding yeast, the worm, and the fly. In addition, defects in synaptic vesicle trafficking were observed in the ribbon synapses of fish photoreceptors as a result of the Nrc mutation, which targets the synaptojanin gene. Thus, the function of synaptojanin appears to be largely conserved across evolution, thereby implicating its role in fundamental biological processes in eukaryotes. Along with studies showing that key endocytic molecules, such as plasma membrane-associated clathrin adaptors and dynamin, bind PtdIns(4,5)P2, work on synaptojanin has provided strong evidence for involvement of PtdIns(4,5)P2 in endocytosis. Consistent with the ability of this lipid to also promote actin polymerization, abnormally high levels of actin-like cytomatrix were found in nerve terminals from $Synj1^{-/-}$ mice and in the giant synapse from the lamprey following a microinjection of inhibitory antibodies against synaptojanin.

Based on studies of Synj1 function in various organisms as well as the role of PtdIns(4,5)P2 in multiple experimental systems, a main function of synaptojanin appears to be the elimination of PtdIns(4,5)P2 from membranes during the endocytic process. This lipid is essential for multiple processes occurring at the plasma membrane, including signal transduction,

regulation of cytoskeletal dynamics, and ion channel/transporter function as well as exo-endocytosis. To a large extent, the implication of PtdIns(4,5)P2 in these processes reflects its ability to function as a compartment-specific identity tag for the plasma membrane. Importantly, this tag must be eliminated as membrane material flows to other cellular compartments, such as endosomes, during endocytosis. Failure to carry out this PtdIns(4,5)P2 conversion via Synj1-mediated dephosphorylation produces defects in the endocytic phase of the synaptic vesicle cycle, including an accumulation of clathrin-coated vesicles suggestive of an uncoating delay. In yeast, failure to hydrolyze PtdIns(4,5)P2 during endocytosis results in the aberrant accumulation of this lipid in intracellular organelles, where it is normally de-enriched. Thus, the main function of synaptojanin might be to spatially restrict PtdIns(4,5)P2 to the plasma membrane as part of homeostatic mechanisms ensuring the maintenance of membranous compartment identity in cells.

Conclusions

Studies disrupting synaptojanin function in neurons, either genetically or via peptide competition, have shown that this molecule plays a critical role in synaptic vesicle recycling. Whereas the most well-characterized function of synaptojanin is the dephosphorylation of PtdIns(4,5)P2 by the 5-phosphatase domain, the exact site and time of action of this enzyme during the endocytic process are not known. The accumulation of clathrin-coated vesicles in null animals suggests involvement of this molecule at a late stage, although it may also modulate the endocytic process at an early stage (e.g., for the regulation of vesicle fission), as suggested by work on Synj2 in nonneuronal cells.

A fundamental question is whether all the defects present in Synj1 null animals are caused by aberrant PtdIns(4,5)P2 metabolism and, thus, by a defect of the 5-phosphatase activity. Indeed, the Sac1 domain has been shown to contribute to the endocytic process while the proline-rich domain is responsible for interactions with other proteins that help target the molecule to sites of action. Reconstitution experiments from Synj1 knockout cultured neurons have addressed the contribution of the various regions of the molecule to specific stages of the endocytic process. Furthermore, based on the large number of biological phenomena regulated by PtdIns(4,5)P_2 in cells, it will be important to determine the contribution of synaptojanin family members to these processes. For instance, the function of PtdIns(4,5)P2-sensitive ion channels or transporters may be regulated by rapid hydrolysis of PtdIns(4,5)P2, either by dephosphorylation or by phospholipase C. Similarly, synaptojanin may regulate the availability of PtdIns(4,5)P_2 for conversion into active metabolites, such as diacylglycerol/Ins(1,4,5)P_3 and PtdIns(3,4,5)P_3.

Because PtdIns(4,5)P2 metabolism plays an essential role in neurophysiology, it would not be surprising that its dyshomeostasis in humans may lead to behavioral abnormalities, such as cognitive deficits. Importantly, the SYNJ1 gene is mapped to chromosome 21 and is overexpressed in a gene dosage-dependent manner in brain tissue from Down syndrome-affected individuals. Studies in mice have suggested that the trisomy of *Synj1* may contribute to subtle neurobehavioral deficits in tasks assessing spatial memory. Additionally, *Synj1* haploinsufficiency may confer protection against some of the synaptotoxic effects of amyloid beta peptide, which has been recently shown to decrease the levels of PtdIns(4,5)P_2. In addition, the ch21q22 region on which Synj1 is found has also been linked to bipolar disorder, and studies have identified potential Synj1 variants that may be involved in this psychiatric disorder. Work using animal models should increase our understanding of the contribution of synaptojanin dysfunction to brain disorders.

See also: Down Syndrome; Dynamin; Lipids and Membranes in Brain Aging; Neuronal Motility and Structure: Cdk5 Pathways; Synaptic Vesicles.

Further Reading

Cremona O and De Camilli P (2001) Phosphoinositides in membrane traffic at the synapse. *Journal of Cell Science* 114: 1041–1052.

Di Paolo G and De Camilli P (2006) Phosphoinositides in cell regulation and membrane dynamics. *Nature* 443: 651–657.

Fernandez-Alfonso T and Ryan TA (2006) The efficiency of the synaptic vesicle cycle at central nervous system synapses. *Trends in Cell Biology* 16: 413–420.

Halstead JR, Jalink K, and Divecha N (2005) An emerging role for PtdIns(4,5)P2-mediated signalling in human disease. *Trends in Pharmacological Sciences* 26: 654–660.

Harris TW, Schuske K, and Jorgensen EM (2001) Studies of synaptic vesicle endocytosis in the nematode *C. elegans*. *Traffic* 2: 597–605.

Murthy VN and De Camilli P (2003) Cell biology of the presynaptic terminal. *Annual Review of Neuroscience* 26: 701–728.

Richmond JE and Broadie KS (2002) The synaptic vesicle cycle: Exocytosis and endocytosis in *Drosophila* and *C. elegans*. *Current Opinion in Neurobiology* 12: 499–507.

Rohrbough J and Broadie K (2005) Lipid regulation of the synaptic vesicle cycle. *Nature Reviews Neuroscience* 6: 139–150.

Ryan TA (2006) A pre-synaptic to-do list for coupling exocytosis to endocytosis. *Current Opinion in Cell Biology* 18: 416–421.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Wenk MR and De Camilli P (2004) Protein–lipid interactions and phosphoinositide metabolism in membrane traffic: Insights from vesicle recycling in nerve terminals. *Proceedings of the National Academy of Sciences of the United States of America* 101: 8262–8269.

Synaptosomes

I J Weiler, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Definition

Synaptosomes are a subcellular fraction of resealed synaptic junctions prepared from brain tissue. On gentle homogenization of brain cells, lipid bilayers reseal, forming subcellular fragments, some of which contain synapses. The synaptic junction is hardy, making it possible to fractionate the subcellular fragments to enrich for synapse-containing elements on the basis of size or density. Preparations are thus characterized by a large preponderance of resealed pre- and postsynaptic entities.

Modes of Separation

Centrifugation

Initial gentle homogenization is important in the disruption of cell constituents. Glass–glass manual homogenizers and low speeds seem to yield more synaptosomes, as judged by the yield of neurotransmitters such as acetylcholine. Most laboratories employ buffers containing 0.32 M sucrose as the homogenizing medium. Centrifugation is then used to separate particles on the basis of sedimentation rate, in which size plays an important role. Early descriptions by Gray and Whittaker, De Robertis, and others demonstrated by electron microscopy and neurotransmitter analysis that centrifugation procedures could yield a population of subcellular particles largely derived from nerve endings.

In a typical procedure, the homogenate is pelleted at 1000 g for 10 min to remove the nuclear pellet; the supernate is spun for 1 h at 17 000 g to yield a mitochondrial pellet, which is then redispersed on a sucrose density gradient (0.8 M, 1.0 M, 1.2 M, and 1.4 M sucrose), centrifuged at 50 000 g for 2 h to separate mitochondria and myelin fragments from synaptosomes. Centrifugation procedures have been subsequently refined by the use of Percoll and Ficoll gradients that reduce osmotic damage.

Moreover, using synaptosomes as a base material, it has been possible to further purify synaptic vesicles (as a source for neurotransmitters), synaptic plasma membranes, and synaptic junctional complexes. A thorough review of the literature in 1975 was provided by DG Jones in his book *Synapses and Synaptosomes*.

A major issue with all methods, including the synaptoneurosomes (discussed later), is assessing the contamination of the fractions by axonal and glial components of the initial tissue homogenate. Although the close functional association of neurons and glia is basic to intact brain tissue function, fragments of glia can skew the analysis of enzyme levels in preparations of isolated nerve terminals. The assessment of purity is based ideally on electron microscopy of pelleted fractions and depends on recognizing a synapse on the basis of the postsynaptic density and the presynaptic vesicles. Serial sectioning must be used to distinguish particles on this morphological basis, and quantitative determination is difficult because of the complexity of tissue fragments. Most laboratories have reported over 60% purity of the synaptosomes.

Specialized Preparations

Considerable effort has been put into achieving either purer or more active preparations. Percoll gradients have been helpful in reducing contamination by myelin and extrasynaptosomal mitochondria; however, the procedures are lengthy (1 h or more) and inflict stress on the cell membrane by repeated pelleting and pipetting procedures. Phase partition, using dextran T500 and polyethylene glycol 4000, has been reported to yield synaptosomes of 62.9% purity based on morphological criteria. A novel method for purification based on synaptic function is described below.

Synaptodendrosomes

Further subfractionation of synaptosomes to the level of the synaptic plasma membrane and synaptic junctional complex has been undertaken; this method relies on two centrifugations followed by ultracentrifugation through a discontinuous Ficoll gradient, preserving some postsynaptic polyribosomes. Despite the lengthy procedure, the synaptodendrosomes are robust enough to exhibit dendritic protein translation that can be shown not to be due to mitochondrial protein synthesis.

Synaptoneurosomes

Hollingsworth and colleagues made the important step of developing a shorter, milder procedure based on simple size fractionation, eliminating centrifugation entirely. A gently homogenized neuronal tissue suspension is passed through a series of filters, yielding a preparation of synaptosomes unusually rich in attached resealed postsynaptic entities; these were

termed synaptoneurosomes and were found to be more homogeneous than previous types of preparations and to yield a larger proportion of postsynaptic configurations (Figure 1). Because of the mild conditions used, it has been possible to obtain functioning synaptoneurosomes from single rat hippocampal slices and to observe metabolic changes during an hour or more of incubation.

Function-Based Purification

To achieve purification of synaptoneurosomes based on functional criteria, we have taken advantage of a lipophilic dye, FM 1–43. This and similar dyes have been used to tag active synapses and to examine exocytotic release from hippocampal synaptosomes. After release of neurotransmitter and endocytosis by synaptic vesicles, the FM 1–43 molecule is trapped in the lipid layers of synaptic vesicles, at which point its fluorescence becomes 50-fold stronger. It is of an intensity such that it can be used in high-performance fluorescence-activated cell sorters. Presynaptic fluorescence can be distinguished from passive absorption to lipid membranes by its dependence on metabolic processes; thus, poisoning the cells with tetrodotoxin abolishes only the active uptake. The particles are loaded with FM 1–43 in the dark, stimulated by addition of 40 mM K^+, and washed twice in cold phosphate-buffered saline (PBS)/10^{-4}M Cd^{2+} to remove excess dye from the medium. The level of fluorescence remains steady for at least 1 h at 4 °C, during which time the particles are sorted at a velocity of about 1000 s^{-1} by means of fluorescence emission at 625 nm. The population distribution shows a large peak of moderate brightness, with a shoulder representing a smaller population (~15%) that is tenfold brighter (Figure 2). The brightest 15% of particles can be selected, concentrated by centrifugation, lysed, and analyzed. The disadvantage of very small yield is compensated for by the increased purity of the preparation.

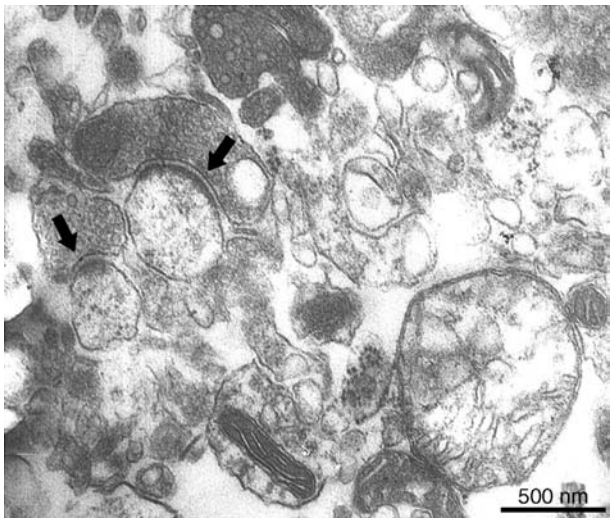

Figure 1 Electron microscope photo of a precipitated and fixed synaptoneurosome preparation. Arrows indicate synaptic contacts sectioned in the transverse plane. Thanks to Anna Klintsova and Georgina Aldridge for assistance.

Applications

Viability

Depolarization of synaptosomes can trigger enzyme reactions leading to increased protein synthesis, suggesting that many metabolic processes remain intact

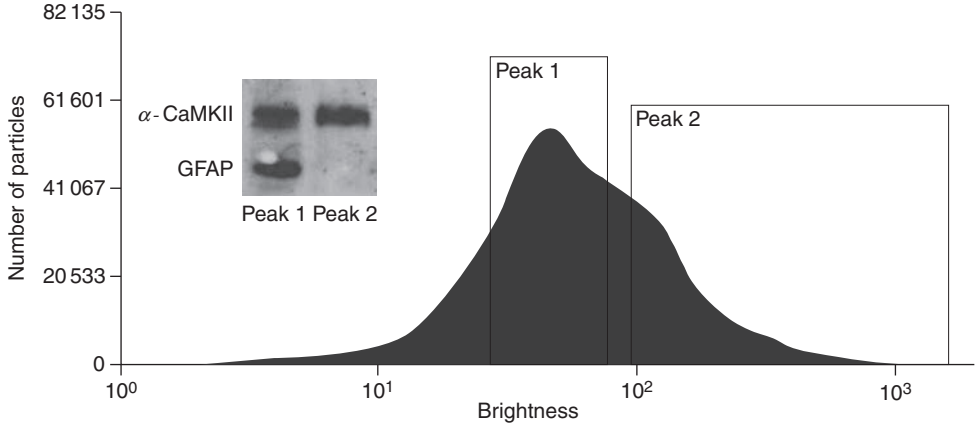

Figure 2 Biphasic distribution of synaptoneurosomes sorted according to brightness. Peak 1 contains particles labeled passively by FM 1–43; after K^+ stimulation and incorporation of dye into presynaptic vesicles, a second peak (peak 2), tenfold brighter, appears as a shoulder on the original distribution. Particles harvested from the peaks analyzed on western blots (inset) show either the presence (peak 1) or absence (peak 2) of glial fibrillary acidic protein (GFAP); the postsynaptic marker α-calmodulin-dependent protein kinase II (α-CaMKII) is present in both fractions. Thanks to Soong Ho Kim for assistance.

for several hours. A careful study of synaptosomal viability, based on measurements of lactate dehydrogenase as well as Ca^{2+} regulation, was made in 1988 by Verhage and colleagues, who concluded that viability remains high for up to 4 h when neocortical synaptosomes are stored on ice; it declines rapidly after 30 min at 37 °C.

Synaptosomes have been an important source of information about neurotransmitters, membrane polarization, and ion entry and exit. They maintain a membrane potential that is sensitive to the addition of 40 mM K^+, a treatment that is known to cause both depolarization of neuronal membranes and an increase in permeability to Ca^{2+} ions.

Neurotransmitter Release

Neocortical synaptosomes have been extensively used to monitor the release of neurotransmitters such as glutamate, dynorphin, norepinephrin, met-enkephalin and cholecystokinin-8, as well as insulin, from isolated nerve terminals. Typically, synaptosomes are applied to a filter and superfused with buffer to which agonists can be added in a controlled manner and the rinse volume analyzed for released transmitters. A plasma membrane Ca^{2+}-adenosine triphosphatase (ATPase) extrudes Ca^{2+} and maintains the polarization of the membrane; depolarization may be caused either by an elevated K^+ concentration (activating Ca^{2+} channels) or veratridine (activating voltage-dependent Na^+ channels). The advantage of synaptosomes for studies of the release of neurotransmitters is the relative simplicity of the preparation, combined with homogeneity such that ion transport, membrane potentials, and bioenergetics can be measured in a complex biological system. Nicholls and others, as reviewed in 1989, demonstrated ATP-dependent exocytotic release of glutamate and γ-aminobutyric acid (GABA) in mammals using this system.

Protein Synthesis

Postsynaptic protein translation was demonstrated in several types of synaptosome preparations and correlated with the presence of ribosomes. The surprising longevity of the synaptoneurosomal preparation and the particulate nature of the suspension make this preparation advantageous for the analysis of reaction dynamics based on repeated rapid sampling from the same aliquots.

It had been shown by Scheetz, using frog tectum synaptoneurosomes, that specific mRNAs, such as that coding α-calmodulin-dependent protein kinase II (CaMKII), are gradually translated during the 75 min following *N*-methyl-D-aspartate (NMDA) activation, against a background of decreased translation of total protein. To look at more rapid events, we took advantage of the uniform nature of a neocortical synaptoneurosome suspension to take serial samples from a single preparation, divided into untreated and agonist-treated aliquots. We found that depolarization by 40 mM K^+ rapidly triggered protein translation, as shown by a shift of RNA into polyribosomal fractions within 1 min of stimulation. By applying extracellular and intracellular Ca^{2+} chelators to the aliquots of a single preparation and using agonists and antagonists specific for ionotropic or metabotropic glutamate receptors, we were able to show that the group I metabotropic glutamate receptors are most effective in triggering rapid postsynaptic protein synthesis by activating a protein kinase C (PKC)-centered second-messenger cascade. We next compared the mRNAs found in the polyribosome fractions derived from stimulated and unstimulated synaptoneurosomes and discovered that the fragile X mental retardation protein (FMRP) is among the proteins being rapidly synthesized after synaptic stimulation. Subsequent studies have revealed that rapid synaptic FMRP synthesis is dependent on the activation of mitogen-activated protein kinase (MAPK) signaling pathways and that this pathway does not trigger rapid protein synthesis in neocortical synaptoneurosomes from mice lacking a functional *Fmr1* gene. Thus, at least some translation that takes place at the synapse requires functional FMRP, suggesting that the altered transport and/or translation of the mRNAs regulated by FMRP may underlie some symptoms of fragile X syndrome.

The surprising robustness of synaptosomes makes them an attractive candidate for the analysis of biochemical pathways. The synapse-bound enzyme systems are stable enough that the synaptodendrosome preparation can exhibit phosphorylation of eIF4E and enhanced peptide elongation of α-CaMKII after brain-derived neurotropic factor (BDNF) treatment. The same preparation has also been used to demonstrate synaptic localization of α-CaMKII mRNA and to measure its increase subsequent to high-frequency *in vivo* stimulation of rat dentate gyrus. Synaptosomes from mice lacking a tyrosine phosphatase (PTPα) have been used to dissect the role of NMDA receptor phosphorylation, which affects learning and memory. Chloride fluxes were studied in synaptoneurosomes from different parts of the rat brain; PKC was shown to facilitate synaptic vessel turnover. Moreover, the effectiveness of Ca^{2+} buffering within brain terminals following hypoxia was studied in synaptosomes from neonatal and adult brain. Finally, because of the enrichment of synaptic elements in these preparations, they can be used to

compare levels of key proteins between dendritic arbors and synaptic regions, and their changes after neural activation, either after *in vivo* treatment, such as whisker stimulation, or after neurotransmitter activation of *in vitro* systems.

See also: Acetylcholine Neurotransmission in CNS; Acetylcholinesterase Inhibitors and Alzheimer's Disease; Acetylcholinesterase; Basal Ganglia: Acetylcholine Interactions and Behavior; Cholinergic Neurotransmission in the Autonomic and Somatic Motor Nervous System; Muscarinic Receptors: Autonomic Neurons; Neuromodulation; Nicotinic Receptors: Autonomic Neurons; Nicotinic Acetylcholine Receptors; Sleep–Wake State Regulation by Acetylcholine.

Further Reading

Betz WJ and Bewick GS (1992) Optical analysis of synaptic vesicle recycling at the frog neuromuscular junction. *Science* 255: 200–203.

De Robertis E, Rodriguez de LoresArnaiz G, and Pellegrino de Iraldi A (1962) Isolation of synaptic vesicles from nerve endings of the rat brain. *Nature* 194: 794–795.

Gray EG and Whittaker VP (1962) The isolation of nerve endings from brain: An electron-microscopic study of cell fragments derived by homogenization and centrifugation. *Journal of Anatomy* 96: 79–88.

Hollingsworth EB, McNeal ET, Burton JS, Williams RJ, Daly JW, and Creveling CR (1985) Biochemical characterization of a filtered synaptoneurosome preparation from guinea pig cerebral cortex: Cyclic adenosine 3′:5′-monophosphate-generating systems, receptors, and enzymes. *Journal of Neuroscience* 8: 2240–2253.

Johnson MW, Chotiner JK, and Watson JB (1997) Isolation and characterization of synaptoneurosomes from single rat hippocampal slices. *Journal of Neuroscience Methods* 77: 151–156.

Jones DG (1975) *Synapses and Synaptosomes: Morphological Aspects*. London: Chapman and Hall.

Kanhema T, Dagestad G, Panja D, et al. (2006) Dual regulation of translation initiation and peptide chain elongation during BDNF-induced LTP *in vivo*: Evidence for compartment-specific translation control. *Journal of Neurochemistry* 99: 1328–1337.

Keelan J, Bates TE, and Clark JB (1996) Intrasynaptosomal free calcium concentration during rat brain development: Effects of hypoxia, aglycaemia, and ischaemia. *Journal of Neurochemistry* 66: 2460–2467.

Nicholls DG (1989) Release of glutamate, aspartate, and γ-aminobutyric acid from isolated nerve terminals. *Journal of Neurochemistry* 52: 331–341.

Rao A and Steward O (1991) Evidence that protein constituents of postsynaptic membrane specializations are locally synthesized: Analysis of proteins synthesized within synaptosomes. *Journal of Neuroscience* 11: 2881–2895.

Scheetz AJ, Nairn AC, and Constantine-Paton M (2000) NMDA receptor-mediated control of protein synthesis at developing synapses. *Nature Neuroscience* 3: 211–216.

Todd PK and Mack KJ (2000) Sensory stimulation increases cortical expression of the fragile X mental retardation protein *in vivo*. *Molecular Brain Research* 80: 17–25.

Verhage M, Besselsen E, Lopes Da Silva FH, and Ghijsen WEJM (1988) Evaluation of the Ca^{2+} concentration in purified nerve terminals: Relationship between Ca^{2+} homeostasis and synaptosomal preparation. *Journal of Neurochemistry* 51: 1667–1674.

Verity MA, Brown WJ, and Cheung M (1980) Isolation of ribosome containing synaptosome subpopulation with active *in vitro* protein synthesis. *Journal of Neuroscience Research* 5: 143–153.

Weiler IJ and Greenough WT (1991) Potassium ion stimulation triggers protein translation in synaptoneurosomal polyribosomes. *Molecular and Cellular Neuroscience* 2: 305–314.

Weiler IJ, Irwin SA, Klintsova AY, et al. (1997) Fragile X mental retardation protein is translated near synapses in response to neurotransmitter activation. *Proceedings of the National Academy of Sciences of the United States of America* 94: 5395–5400.

Whittaker VP (1965) The application of subcellular fractionation techniques to the study of brain function. *Progress in Biophysics & Molecular Biology* 15: 39–96.

Synaptotagmins

A Maximov, University of Texas Southwestern Medical Center, Dallas TX, USA

© 2009 Published by Elsevier Ltd.

Introduction

Synaptic transmission plays a central role in interneuronal communications. In nerve terminals, fast release of neurotransmitters is mediated by synaptic vesicles, intracellular trafficking organelles that rapidly fuse with the plasma membrane in response to action-potential-driven Ca^{2} influx. Over the past two decades, soluble *N*-ethylmaleimide-sensitive factor attached protein receptor (SNARE) proteins and synaptotagmins have emerged as critical components of the release machinery that mediates synaptic vesicle fusion and exocytosis.

Synaptotagmins: Calcium Sensors for Membrane Fusion

Synaptotagmins is a family of membrane-trafficking proteins composed of a short N-terminus, a single transmembrane domain, a variable linker domain, and two C_2 domains (the C_2A and C_2B domains; Figure 1). The mammalian synaptotagmin family includes 16 genes, whereas the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes encode eight and seven synaptotagmin genes, respectively. In *Drosophila* and mice, the expression of most synaptotagmin isoforms is restricted to the nervous system. Despite the similarity in their principal domain organization, different members of the synaptotagmin family display distinct expression patterns, subcellular localizations, and biochemical properties. Of the 16 synaptotagmins identified in the mouse genome, four isoforms, synaptotagmins-1, -2, -9, and -12, are localized on the synaptic vesicles, whereas the remaining isoforms are likely to be targeted to the various intracellular membrane compartments. The prototypical member of the synaptotagmin family, synaptotagmin-1, was originally identified in a monoclonal antibody screen for synaptic vesicle proteins. Subsequent genetic and functional studies showed that synaptotagmin-1 acts as a Ca^{2} sensor for fast synaptic vesicle exocytosis. In *Drosophila* and mice, deletion of the synaptotagmin-1 gene results in lethal phenotypes associated with the specific loss of fast (synchronous) Ca^{2} -triggered neurotransmitter release (Figure 2); whereas the mutations in the mouse synaptotagmin-1 gene that alter overall synaptotagmin-1 Ca^{2} -binding affinity produce a similar shift in a Ca^{2} sensitivity of transmitter release. Recent studies showed that, similar to synaptotagmin-1, synaptotagmins-2 and -9 act as Ca^{2} sensors for fast synaptic vesicle exocytosis. However, the functional roles of most nonsynaptic vesicle synaptotagmin isoforms remain controversial.

Biochemistry and the Mechanism of Action of Synaptotagmin-1

The average synaptic vesicle contains approximately 15 synaptotagmin-1 molecules that are oriented with their N-termini facing the vesicle lumen and their C_2 domains facing the cytosol. The intraluminal domain of synaptotagmin-1 contains a site for N-glycosilation that is essential for vesicular targeting. The structure of synaptotagmin C_2 domains consists of a compact β-sandwich composed of two four-stranded β-sheets and the top loops that form the sites for Ca^{2} binding (Figure 1(b)). The Ca^{2} binding sites of C_2 domains are coordinated by several negatively charged residues that are widely separated in the primary sequence. The C_2A domain binds three Ca^{2} ions, whereas the C_2B binds two Ca^{2} ions. The intrinsic Ca^{2} affinities of the synaptotagmin-1 C_2 domains are low because the coordination spheres for the Ca^{2} ions are incomplete. The apparent Ca^{2} affinity of the C_2 domains increases dramatically when the C_2 domains bind to phospholipids whose negatively charged head groups provide additional coordination sites for the bound Ca^{2} ions. Apart from Ca^{2} -dependent phospholipid binding, C_2 domains of synaptotagmin-1 mediate Ca^{2} -dependent and Ca^{2} -independent interactions with the SNARE proteins syntaxin-1 and synaptosomal-associated protein of 25 kDa (SNAP-25) and with assembled SNARE complexes. In addition, a number of studies using recombinant and native proteins reported interactions between synaptotagmin-1 and the various components of the presynaptic active zone and proteins implicated to regulation of synaptic vesicle cycle.

Substantial evidence has accumulated that Ca^{2} -dependent binding of synaptotagmin-1 to the lipid membranes and SNARE complexes plays a central role in the function of synaptotagmin-1 in triggering fast synaptic vesicle exocytosis. In reconstituted systems, the C_2 domains of synaptotagmin-1 accelerate the rate of SNARE-mediated lipid mixing in the presence of Ca^{2} , suggesting that the SNAREs and synaptotagmin-1 may cooperate in mediating vesicle fusion. This model is supported by the notions that the purified C_2B domain of synaptotagmin-1 is able to bring the lipid vesicles into close proximity by simultaneously binding to two membranes in a Ca^{2} -dependent manner. Analysis of Sr^{2} -induced

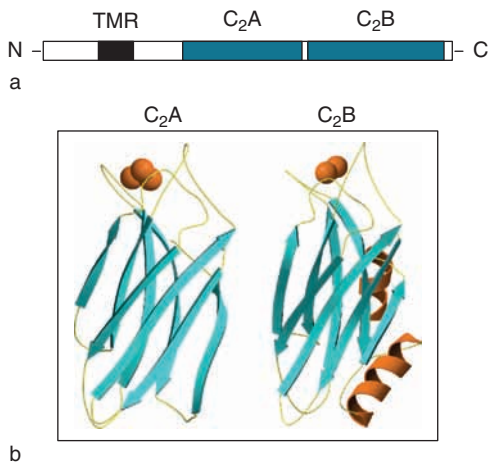

Figure 1 Synaptotagmins: (a) principal domain organization of synaptotagmins illustrating the position of the transmembrane region (TMR) and C_2 domains (C_2A and C_2B); (b) structure of synaptotagmin-1 C_2A and C_2B domains depicting the β-sandwich composed of two four-stranded β-sheets and the top loops that form the sites for Ca^{2+} binding. In (b), the spheres on the top illustrate bound Ca^{2+} ions. Published with permission of Dr. Josep Rizo.

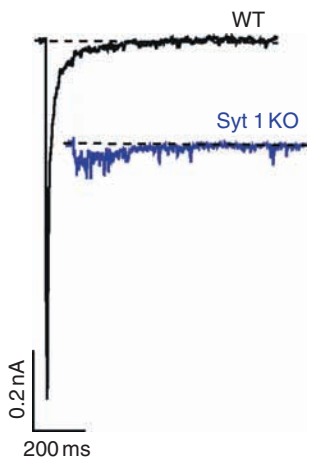

Figure 2 Fast synaptic transmission in wild-type (WT) cortical neurons and neurons isolated from synaptotagmin-1 knockout (Syt 1 KO) mice. Excitatory postsynaptic currents were triggered by single-action potentials and recorded in a whole-cell mode in the presence of 2 mM extracellular Ca^{2+}.

synaptic transmission in cultured hippocampal neurons indicates that synaptotagmin-1 binding to phospholipids may alone be sufficient for triggering synaptic vesicle exocytosis. However, an independent reconstitution study concluded that synaptotagmin-1 triggers fusion by direct interactions with the SNAREs and not solely through the aggregation of vesicles. This conclusion is supported by functional studies of synaptic transmission in neurons isolated from the mutant synaptotagmin-1 mouse line that carries a single-point mutation that alters the affinity of Ca^{2+}-dependent synaptotagmin-1 binding to SNAREs complexes without affecting its phospholipids binding activity. The later study has shown that the enhanced affinity of synaptotagmin-1-SNARE binding (D232N mutation in the C_2A domain) leads to an increase in the probability of evoked neurotransmitter release. Most likely, interaction between synaptotagmin-1 and SNAREs occurs during the last step of synaptic vesicle priming with the plasma membrane, when Ca^{2+}-bound synaptotagmin-1 competes for SNARE complex binding with complexins, small soluble proteins that have been shown to promote the full assembly of the SNARE complexes. In summary, the possible mechanism of synaptotagmin-1 action in triggering fast synaptic vesicle exocytosis involves Ca^{2+}-dependent phospholipid binding and displacement of complexins from the fully assembled SNARE complexes, followed by cooperation with SNAREs in completing the fusion reaction.

Synaptotagmins-1, -2, and -9: Fast Calcium Sensors with Distinct Kinetic Properties

Apart from synaptotagmin-1, three additional synaptotagmin isoforms are localized on the synaptic vesicles. Two of these isoforms, synaptotagmins-2 and -9 (also called synaptotagmin-5), contain functional Ca^{2+}-binding sites in their C_2 domains and display Ca^{2+}-dependent phospholipid and SNARE-binding activities. In the mouse brain, expression patterns of synaptotagmins-1, -2, and -9 are partially segregated, with synaptotagmin-1 being predominantly expressed in the forebrain; synaptotagmin-2 in the cerebellum, brain stem, and spinal cord; and synaptotagmin-9 in the limbic system and striatum. Recent genetic analyses have shown that deletions of the synaptotagmin-2 and -9 genes in mice impair fast Ca^{2+}-triggered neurotransmitter release in the synapses of the Calyx of Held, neuromuscular junctions, and cultured striatal neurons. In rescue experiments, lentiviral-mediated expression of synaptotagmins-2 and -9 restored the fast synaptic responses in cortical neurons lacking synaptotagmin-1, indicating that synaptotagmins-2 and -9 act as Ca^{2+} sensors for fast synaptic vesicle exocytosis. Interestingly, the evoked synaptic responses triggered by three different synaptotagmin isoforms displayed substantial differences in kinetics and Ca^{2+} sensitivity. When compared to synaptotagmin-1, fast synaptic responses triggered by synaptotagmin-2 displayed similar Ca^{2+} dependence but had significantly faster rise and decay times. In contrast, synaptic responses triggered by synaptotagmin-9 displayed much slower rise and decay times and lower Ca^{2+} dependence. Hence, in

subsets of neurons, distinct release properties may be defined by differential expression of fast Ca^{2+} sensors.

Conclusion and Perspectives

Combinations of biochemical, structural, genetic, and electrophysiological approaches advanced our understanding of the functional roles of several members of synaptotagmin family in triggering Ca^{2+}-dependent vesicle fusion. However, a number of questions remain to be answered. First, the interplay between the fast (synchronous) and slow (asynchronous) forms of synaptic vesicle exocysotis needs to be established. Indeed, the synapses of neurons lacking synaptotagmis-1 and -2 exhibit a dramatic increase in the magnitude of delayed asynchronous release and in the frequency of miniature neurotransmission, suggesting that fast Ca^{2+} sensors may regulate asynchronous vesicle exocytosis and minis. Second, the physiological roles of other members of synaptotagmin family need to be clarified. Several nonsynaptic vesicle synaptotagmin isoforms contain functional Ca^{2+}-binding sites and have been implicated in regulating various forms of membrane fusion. For example, synaptotagmin-7 was proposed as mediating Ca^{2+}-triggered exocytosis of dense-core vesicles in PC12 cells and acting as a Ca^{2+} sensor in lysosome exocytosis in nonneuronal cells, whereas synaptotagmin-4 has been recently shown to serve as a Ca^{2+} sensor for postsynaptic vesicle exocytosis in *Drosophila*. Despite these initial clues, the exact subcellular localizations and the functional roles of various nonsynaptic vesicle synaptotagmin isoforms need to be further elucidated.

See also: Calcium Channel Subtypes Involved in Neurotransmitter Release; Calcium Channels and SNARE Proteins; Learning and Memory in Invertebrates: *C. Elegans*; Learning and Memory in Invertebrates: *Drosophila*; Neurotransmitter Release: Synchronous and Asynchronous; SNAREs; Synaptic Transmission: Models.

Further Reading

Adolfsen B, Saraswati S, Yoshihara M, and Littleton JT (2004) Synaptotagmins are trafficked to distinct subcellular domains including the postsynaptic compartment. *Journal of Cell Biology* 166(2): 249–260.

Arac D, Chen X, Khant HA, et al. (2006) Close membrane-membrane proximity induced by Ca(2+)-dependent multivalent binding of synaptotagmin-1 to phospholipids. *Nature Structural & Molecular Biology* 13(3): 209–217.

Bhalla A, Chicka MC, Tucker WC, and Chapman ER (2006) Ca(2+)-synaptotagmin directly regulates t-SNARE function during reconstituted membrane fusion. *Nature Structural & Molecular Biology* 13(4): 323–330.

Chai Q, Arndt JW, Dong M, et al. (2006) Structural basis of cell surface receptor recognition by botulinum neurotoxin B. *Nature* 444(7122): 1096–1100.

Fernandez I, Arac D, Ubach J, et al. (2001) Three-dimensional structure of the synaptotagmin 1 C2B-domain: Synaptotagmin 1 as a phospholipid binding machine. *Neuron* 32(6): 1057–1069.

Fernandez-Chacon R, Konigstorfer A, Gerber SH, et al. (2001) Synaptotagmin I functions as a calcium regulator of release probability. *Nature* 410(6824): 41–49.

Geppert M, Goda Y, Hammer ER, et al. (1994) Synaptotagmin I: A major Ca2+ sensor for transmitter release at a central synapse. *Cell* 79(4): 717–727.

Giraudo CG, Eng WS, Melia TJ, and Rothman JE (2006) A clamping mechanism involved in SNARE-dependent exocytosis. *Science* 313(5787): 676–680.

Littleton JT, Stern M, Schulze K, Perin M, and Bellen HJ (1993) Mutational analysis of *Drosophila* synaptotagmin demonstrates its essential role in Ca(2+)-activated neurotransmitter release. *Cell* 74(6): 1125–1134.

Martens S, Kozlov MM, and McMahon HT (2007) How synaptotagmin promotes membrane fusion. *Science* 316(5828): 1205–1208.

Pang ZP, Melicoff E, Padgett D, et al. (2006) Synaptotagmin-2 is essential for survival and contributes to Ca^{2+} triggering of neurotransmitter release in central and neuromuscular synapses. *Journal of Neuroscience* 26(52): 13493–13504.

Pang ZP, Shin OH, Meyer AC, Rosenmund C, and Sudhof TC (2006) A gain-of-function mutation in synaptotagmin-1 reveals a critical role of Ca^{2+}-dependent soluble *N-ethylmaleimide*-sensitive factor attachment protein receptor complex binding in synaptic exocytosis. *Journal of Neuroscience* 26(48): 12556–12565.

Rizo J, Chen X, and Arac D (2006) Unraveling the mechanisms of synaptotagmin and SNARE function in neurotransmitter release. *Trends in Cell Biology* 16(7): 339–350.

Sudhof TC (2004) The synaptic vesicle cycle. *Annual Review of Neuroscience* 27: 509–547.

Takamori S, Holt M, Lemke EA, et al. (2006) Molecular anatomy of a trafficking organelle. *Cell* 127(4): 831–846.

Tang J, Maximov A, Shin OH, Dai H, Rizo J, and Sudhof TC (2006) A complexin/synaptotagmin 1 switch controls fast synaptic vesicle exocytosis. *Cell* 126(6): 1175–1187.

Xu J, Mashimo T, and Sudhof TC (2007) Synaptotagmin-1, -2, and -9: Ca(2+) sensors for fast release that specify distinct presynaptic properties in subsets of neurons. *Neuron* 54(4): 567–581.

Yoshihara M, Adolfsen B, Galle K, and Littleton JT (2005) Retrograde signaling by Syt 4 induces presynaptic release and synapse-specific growth. *Science* 310(5749): 858–863.

SynCAMs

A I Fogel and T Biederer, Yale University, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Key Molecules: SynCAM Proteins

Synapses in the central nervous system (CNS) are asymmetric sites of interneuronal adhesion at which the presynaptic release machinery is localized opposite the postsynaptic receptor apparatus. Both pre- and postsynaptic specializations are tightly aligned with each other. The proper formation of synaptic contacts is endpoint to a highly complex process. It is preceded by signaling and guidance steps at multiple levels, including axonal pathfinding, physical contact of pre and postsynaptic neurons, and interactions between neurons and glia cells surrounding synaptic sites. Once intercellular contact of an incoming axon with a target neuron is achieved, specific signals are necessary to instruct both pre- and postsynaptic neurons to assemble specializations capable of controlled neurotransmitter release and detection. These instructive signals need to be specific both spatially and temporally to achieve the concomitant formation of overlapping pre- and postsynaptic specializations. Cell adhesion molecules are likely candidates to mediate this signaling as they are both capable of demarcating a membrane area as a nascent synapse and of mediating signaling between cells. It can be hypothesized that these activities of synaptic cell adhesion molecules assemble the appropriate intracellular molecular machinery at demarcated sites of synaptic contact.

At mammalian central synapses, several transsynaptic interactions have been described. Roles for adhesion molecules at synapses include stabilization and maturation of synaptic contacts, as described for N-cadherin, and long-term potentiation, in which neural cell adhesion molecule is involved. Receptors such as Eph receptors and their ephrin ligands have been implicated in dendritic spine morphogenesis and synapse differentiation. However, only two transsynaptic adhesion systems are known to play instructive roles in the organization of nascent synapses; these are the neuroligin/neurexin cell adhesion pair and the SynCAM (synaptic cell adhesion molecule) proteins. The four SynCAM proteins are encoded by cell adhesion molecule genes 1–4 and have also been named nectin-like molecules and tumor suppressor in lung cancer (TSLC) proteins as a result of their identification in different contexts. It is likely that additional adhesion molecules are involved in this process, and it will ultimately require detailed studies of each of the proteins to distinguish their roles in synapse formation, stabilization, and maintenance. This article reviews current knowledge of the roles the recently identified SynCAM proteins play in these processes.

Identification and Basic Properties of SynCAM Proteins

SynCAM 1 was initially identified in a bioinformatic search for proteins containing both extracellular immunoglobulin (Ig)-like domains and intracellular PDZ-domain interaction sequences. The search for proteins with this domain organization was motivated by research on the invertebrate proteins fasciclin II in *Drosophila* and apCAM in *Aplysia*, which contribute critically to proper synapse formation and share the presence of these domains and sequences with SynCAM 1. No related proteins had been investigated at vertebrate synapses prior to the identification of SynCAM 1. This domain organization is well suited for a synapse-organizing adhesion molecule as extracellular Ig-like domains are known to mediate homo- and heterophilic adhesion, while the intracellular PDZ interaction sequence is capable of tethering the adhesion molecule to scaffolding molecules found abundantly in the postsynaptic density and the presynaptic active zone. SynCAM 1 is the founding member of a small family of four SynCAM proteins that all share this domain organization.

SynCAMs are single-pass transmembrane proteins of relatively small size (just over 400 amino acids, about 45 kDa in predicted molecular weight depending on family member identity and splice variant). SynCAM proteins contain three extracellular Ig-like domains of the V-set, C1-set, and I-set subclasses (Figure 1(a)). Intracellularly, SynCAM proteins have a short cytosolic tail similar to the cytosolic tails of neurexins, another family of synaptic adhesion molecules involved in synaptogenesis (Figure 1(b)). The cytosolic tails of SynCAMs and neurexins share two motifs. One motif allows for interactions with proteins containing type II PDZ domains, which include a number of synaptically localized adaptor molecules. This motif is located at the extreme C-terminus of SynCAM proteins. The second motif is a sequence of amino acids that can bind members of the protein 4.1 family, which promote assembly of actin/spectrin filaments. One isoform, protein 4.1N, is neuronally expressed and could constitute a molecular link for SynCAMs with the actin cytoskeleton. This motif is located in the N-terminal part of the cytosolic tail, near the transmembrane domain.

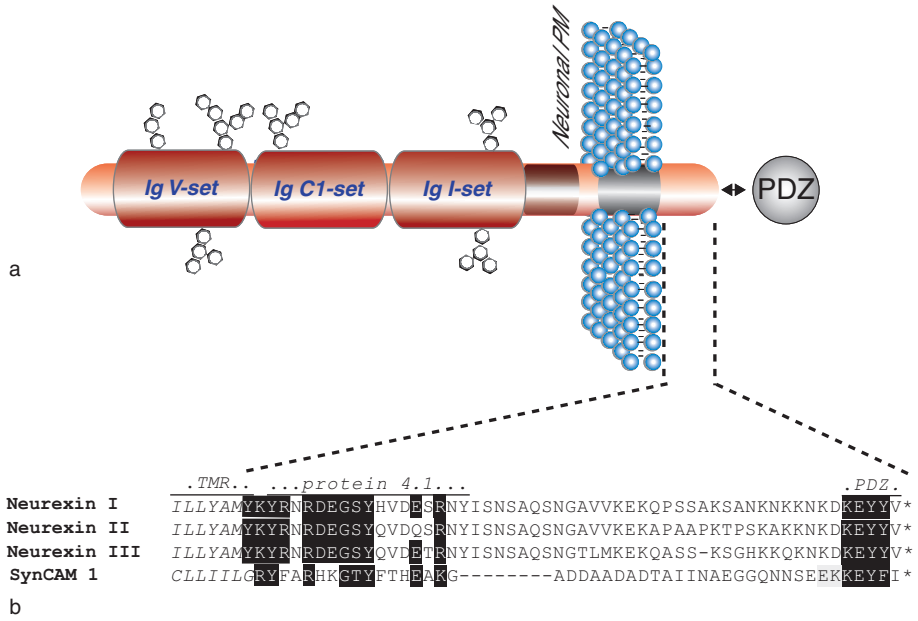

Figure 1 Domain map of synaptic cell adhesion molecule (SynCAM) 1 and sequence alignment of SynCAM 1 with neurexins. (a) Domain map of SynCAM 1. SynCAM proteins contain three extracellular Ig-like domains that are N-glycosylated and mediate its adhesive contacts. Positions of N-glycosylation sites are shown as they are located in SynCAM 1. Following the single transmembrane region, SynCAM 1 has a short cytoplasmic sequence which mediates interactions with protein 4.1 family members and with PDZ domain-containing proteins. (b) The C-terminal sequences of SynCAM proteins and neurexins are highly similar, suggesting that the proteins share converging signaling pathways. Interaction sequences with protein 4.1 and PDZ type II domain-containing adaptor molecules are indicated. Ig, immunoglobulin; PM, plasma membrane; TMR, transmembrane region.

The SynCAM protein family is encoded by four *CADM* (cell adhesion molecule) genes designated *CADM 1–4*. SynCAM orthologs are highly conserved throughout the vertebrate phylum. In addition, individual SynCAM family members are highly related to each other and share up to 70% amino acid similarity. No genes encoding SynCAM proteins are found in invertebrates, suggesting the function of the encoded proteins is specific to vertebrates. A protein family with comparable domain organization, the nectins, has been described in vertebrates. Nectins also contain three extracellular Ig-like domains and a short cytoplasmic tail, but they lack sequence similarity to the SynCAM proteins and are not homologous. Most significantly, SynCAMs and nectins have distinct cytoplasmic sequences and interact with different sets of adaptor proteins. SynCAMs have at times been included in a group of proteins called nectin-like proteins, but this nomenclature is not sufficiently suitable as the term nectin-like molecule subsumes multiple proteins not homologous to each other, is not functionally descriptive, and does not clearly distinguish SynCAMs from the nonhomologous nectins. Other nomenclatures for SynCAM 1 have also been utilized as a result of its identification in different contexts. These include RA175, as named after identification in a screen for retinoic acid inducible proteins, TSLC1, for its role in tumor suppression in lung carcinoma, and spermatogenic IGSF, being a member of the immunoglobulin superfamily expressed in testis. Of these, TSLC1 is the only alternative name that describes a function of a SynCAM protein. This function is performed outside the nervous system and is only briefly referred to in this article (see the next section).

Expression and Localization Studies of SynCAM Proteins

All four SynCAM proteins are expressed in the CNS. SynCAM proteins 1 and 4 are also expressed in nonneuronal tissue. In addition to its expression in the CNS, SynCAM 1 is expressed in lung and was identified as a tumor suppressor in non-small-cell lung cancer. SynCAM 1 is also expressed in testes, where it is necessary for proper spermatid adhesion and maturation, and in T cell zone dendritic cells of the immune system. SynCAM 4 is also expressed in the immune system, where its role is unclear. SynCAMs 2 or 3 have been identified exclusively within the CNS.

The expression of SynCAMs 2, 3, and 4 in mouse brain begins early after birth and increases over the next 2 postnatal weeks, an expression profile consistent with a role in synapse formation. SynCAM 1 is expressed already in late embryonic stages, and its expression continues throughout postnatal development into adulthood, indicating that it may have

functions preceding synapse organization in the brain. It is interesting that SynCAM 1 is differentially expressed as two distinct molecular weight species at different developmental stages, suggesting that different splice variants and/or glycosylation patterns occur for SynCAM 1 in a regulated manner. The molecular and functional consequences of altered SynCAM 1 splicing and posttranslational modification are not yet understood.

Biochemical, immunohistochemical, and immunoelectron microscopy studies demonstrate that SynCAMs are synaptically localized. Biochemical fractionation of total brain shows that all four SynCAM proteins strongly enrich in synaptic plasma membrane preparations. Co-localization studies with synaptic marker proteins confirm that SynCAM proteins are synaptic. By means of immunoelectron microscopy, SynCAMs have been shown to localize to both pre- and postsynaptic specializations. These immunohistochemical and immunoelectron microscopy studies used an antibody that recognizes SynCAMs 1, 2, and 3, and as such, no data are yet available for SynCAM 4. However, the synaptic enrichment of SynCAM 4 in biochemical preparations, in addition to its domain organization and interaction motifs shared with other SynCAM proteins, suggests that it will likely be sorted to the same subcellular compartments and confirmed as synaptic protein as well.

The regional expression patterns of SynCAM in the developing CNS are distinct, suggesting that they exert differing functions in the developing brain. Expression studies have been carried out using *in situ* hybridization, RT-PCR, and immunoblotting. Notably, those subsets of SynCAM proteins interacting strongly *in vitro* (see next section) also are co-expressed in the same brain regions during development. One example are SynCAMs 1 and 2, which are both strongly expressed in the cortex and hippocampus during the early postnatal period. This suggests that pairs of SynCAMs may interact not only biochemically but also functionally *in vivo*. Detailed analysis of protein localization at the cellular and synaptic level in conjunction with functional studies will be necessary to determine whether cognate pairs of SynCAMs are in physical proximity and capable of mediating intercellular adhesion to promote synapse formation of neighboring neurons.

Extracellular Interactions of SynCAM Proteins

Heterophilic Interactions of SynCAM Proteins

The adhesive interactions of the extracellular domains of SynCAM proteins have been studied with a number of different techniques. Clustering assays, in which cells or beads display the extracellular domain of different family members, suggest that all four SynCAM proteins are capable of mediating homophilic interactions, but to differing degrees. While homophilic binding of SynCAMs 1, 2, and 3 can be readily demonstrated, SynCAM 4 homophilic interactions appear much weaker. Heterophilic interactions within the family have been studied using both affinity chromatography experiments and in cell overlay assays, in which the soluble extracellular domain of one SynCAM is allowed to adhere to a cell expressing another membrane-bound SynCAM. While SynCAM proteins are capable of interacting homophilically, these studies suggest that they preferentially engage in heterophilic interactions between family members. In these assays, strong binding occurs between SynCAMs 1 and 2 and between SynCAMs 3 and 4. These interactions are summarized in Table 1. The interaction of SynCAM 1 and 2 is affected by glycosylation, indicating regulation of this adhesion complex by posttranslational modification. As the extracellular interactions of SynCAM 1 are necessary for its activity to induce synapses, these studies of SynCAM interactions with each other will be critical to address their roles in synapse formation.

Table 1 Extracellular interactions of the SynCAM family

Protein	*Homophilic interaction*	*Other known interaction partners*
SynCAM 1	Yes	SynCAM 2, SynCAM 3, CRTAM, nectins
SynCAM 2	Yes	SynCAM 1, SynCAM 4
SynCAM 3	Yes	SynCAM 4, SynCAM 1, nectins
SynCAM 4	Very weak	SynCAM 3, SynCAM 2

SynCAM, synaptic cell adhesion molecule; CRTAM, class-1 restricted T cell-associated molecule.

Interactions of SynCAM Proteins with Non-SynCAM Proteins

SynCAM proteins also interact with other immunoglobulin superfamily members outside the SynCAM family, both in the nervous system and in nonneuronal tissues. The extracellular domains of SynCAMs 1 and 3 have been demonstrated to interact with members of the nectin family of proteins in cell clustering assays. However, since it is not known whether these two protein families share subcellular localization, it is unclear what the functional significance of these interactions is. In the immune system, SynCAM 1 is expressed in dendritic cells, where it interacts with the cell adhesion molecule 'class-1 restricted T cell-associated molecule (CRTAM),' as shown by cell overlay assays.

Intracellular Interactions of SynCAM Proteins

Compared to the extracellular analysis of SynCAM binding, the interactions of the cytoplasmic domains of the SynCAM proteins are less well characterized. In part, the promiscuity of the type II PDZ interaction sequence, which SynCAM proteins feature at their extreme C-termini, can mediate *in vitro* interactions with a number of partners that may or may not occur or be significant *in vivo*. Affinity chromatography studies have demonstrated that the cytoplasmic tail of SynCAM 1 is capable of interacting with the brain expressed adaptor molecules CASK, Mint1, and syntenin, all of which are PDZ domain-containing proteins found to differing extents at central synapses. Signaling of SynCAM 1 in the nervous system is plausibly mediated through these adaptor proteins as these are also synaptically enriched. It is likely that these interactions partners of SynCAM 1 also bind cytoplasmic sequences of the other SynCAM family members, but this remains to be confirmed. As expected, SynCAM 1 does not interact with type I PDZ domain-containing proteins such as PSD-95. Further, SynCAM 1 does not bind the nectin-interacting PDZ domain-containing protein afadin.

Interactions of SynCAM 1 cytoplasmic sequences have also been studied outside the CNS. SynCAM 1 interacts with DAL1, a member of the protein 4.1 family expressed in the lung. This interaction was confirmed with both biochemical and cell biological techniques. In addition, the PDZ domain-containing protein called 'protein associated with Lin7–2' was identified from testes as a protein capable of interaction with SynCAM 1 in a yeast two-hybrid screen.

The sequence similarity of the cytoplasmic domains of SynCAMs and neurexins (Figure 1(b)) suggests they function by engaging similar adaptor molecules on the presynaptic side. SynCAMs and neurexins share a protein 4.1 interaction sequence and a PDZ type II interaction motif. Notably, protein 4.1 is also capable of interaction with the scaffolding molecule CASK, providing for the potentially preferred formation of a trimeric complex of SynCAMs, protein 4.1, and CASK (Figure 2). A similar complex has been described for the neurexin proteins. The SynCAM cytoplasmic domains can therefore be imagined as multivalent nucleation sites for scaffolding and cytoskeletal molecules, which, when clustered, assemble molecular complexes such as the vesicle release machinery. This clustering mechanism provides an attractive hypothesis for the way cell adhesion molecules mediate assembly of synaptic membrane complexes.

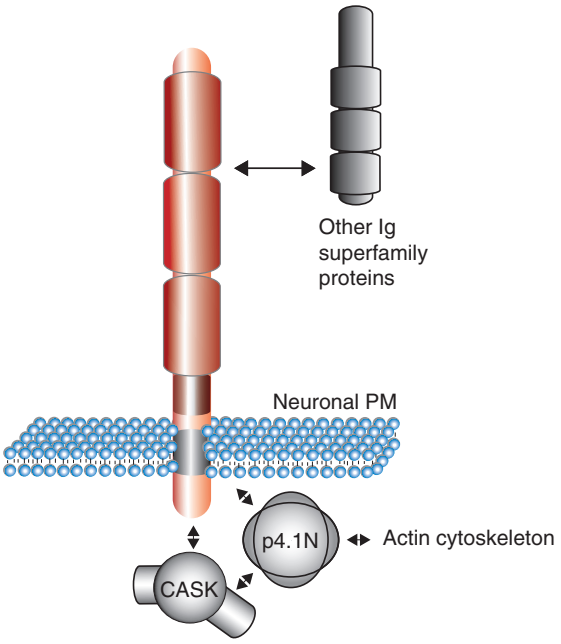

Figure 2 Model of protein–protein interactions of the synaptic cell adhesion molecule (SynCAM) family. SynCAM 1 interacts with different immunoglobulin (Ig) superfamily members through its extracellular domains. Intracellularly, SynCAM 1 interacts with PDZ domain-containing adaptor molecules such as CASK and forms a trimeric complex with p4.1N which could nucleate cytoskeletal structures at synapse. PM, plasma membrane.

Functional Characterization of SynCAM Proteins in Synapse Organization

Multiple lines of evidence show roles of SynCAM 1 in synapse organization. The first of these is based on a co-culture assay for studying synaptogenesis in a reduced system. Here, expression of SynCAM 1 in epithelial cells co-cultured with hippocampal neurons causes the neurons to form functional synapses atop the epithelial cells. This has been demonstrated by three different approaches. First, immunostaining for presynaptic marker proteins, including synaptic vesicle proteins, shows a specific induction of puncta of presynaptic marker proteins formed by axonal membranes atop the surface of SynCAM 1-expressing epithelial cells (Figure 3). These results suggest that SynCAM 1 is capable of instructing neurons to recruit the synaptic vesicle release machinery to sites of contact. Further, by means of live cell imaging in fluorescent marker dye-recycling assays, it has been demonstrated that these specializations formed by neurons atop SynCAM 1-expressing epithelial cells indeed contain recycling synaptic vesicles. Notably, the synaptic vesicle recycling properties of these SynCAM 1-induced presynaptic specializations are indistinguishable from those of interneuronal synapses. Thus, SynCAM 1-induced presynaptic specializations

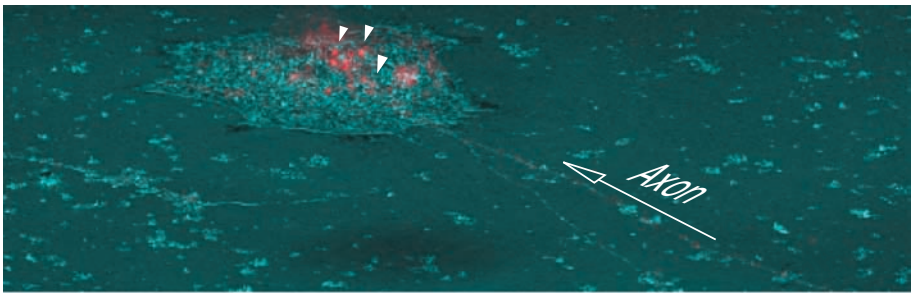

Figure 3 Synaptic cell adhesion molecule (SynCAM) 1 induces presynaptic specializations. Three-dimensional reconstruction of a confocal image of an epithelial-like HEK293 cell expressing SynCAM 1 (top left) that was co-cultured for 2 days with hippocampal neurons (not shown). The HEK293 cell is identified by co-expression of ECFP. The co-culture was stained for the presynaptic marker synaptophysin (red). Contact of an axon with the SynCAM 1-expressing HEK293 induces the axon to form multiple punctate presynaptic specializations (filled arrowheads). An axon extending toward the HEK293 cell is visible due to diffuse synaptophysin staining (arrow). Reproduced from Biederer T (2006) SynCAM in formation and function of synaptic specializations. In: Dityatev A and El-Husseini A (eds.) *Molecular Mechanisms of Synaptogenesis*, ch. 8, pp. 125–136, figure 3B. Copyright Springer Science and Business Media, with permission of Springer Science and Business Media.

are capable of evoked vesicle fusion and recycling. In this co-culture assay, SynCAM 1 has also been tested for the capability of recruiting postsynaptic protein clusters, a property demonstrated for the neurexins through their interactions with neuroligins. However, SynCAM 1 is not capable of assembling postsynaptic densities on contact with dendrites, suggesting that SynCAM 1-mediated transsynaptic signaling occurs in a single direction or does not occur through homophilic contacts.

The most direct demonstration that SynCAM 1-induced presynaptic terminals are functionally active utilized electrophysiological recordings from an artificial postsynaptic element. Electrophysiological recordings were obtained from epithelial cells co-expressing SynCAM 1 and a glutamate receptor formed by GluR2 subunits that were co-cultured with dissociated hippocampal neurons. These studies revealed that evoked excitatory neurotransmitter release occurs from the SynCAM 1-induced presynaptic specializations onto the surface of the SynCAM 1-expressing epithelial cell. Excitatory miniature events could be reconstituted as well. The similarities of SynCAM 1-induced presynaptic terminals to those of interneuronal synapses suggest that this reconstitution approach of neurotransmission might serve as a useful general approach on which the roles of other synaptic proteins might be studied.

Functions of SynCAM 1 in synapse formation and synaptic transmission have also been studied using overexpression analysis in cultured neurons. Overexpression of full-length SynCAM 1 in dissociated neurons increases frequency but not amplitude of excitatory miniature potentials that can be recorded from the neurons, suggesting an increase in release sites formed atop the SynCAM 1-overexpressing neurons. Alternatively, existing release sites may undergo plastic changes on postsynaptic overexpression of SynCAM 1. This effect was abrogated by the removal of the three extracellular Ig-like domains of SynCAM 1, demonstrating that its transsynaptic interactions are necessary for this effect. In a converse loss-of-function approach, overexpression of the cytoplasmic domain of SynCAM 1 in cultured hippocampal neurons reduces the number of presynaptic terminals compared with control neurons. This effect is abrogated when the C-terminal three amino acids are removed from the cytoplasmic tail, suggesting that PDZ interactions are critical to SynCAM 1 function. Since the cytoplasmic domains of neurexins and SynCAMs are similar, this experiment suggests that the interactions of scaffolding molecules with these intracellular adhesion molecules are critical for synapse formation. It is likely that the presynaptic signaling mechanisms of neurexins and SynCAMs converge by interaction with the same downstream partner proteins.

The activities of SynCAM 1 likely involve its interactions with SynCAM 2. Both proteins assemble into a heterophilic synaptic adhesion complex during the peak period of synaptogenesis, and can recruit each other across the synaptic cleft to form this SynCAM 1/2 complex. Concurringly, SynCAM 2 promotes functional synapses as it increases the number of active presynaptic terminals and enhances excitatory neurotransmission similar to SynCAM 1. Interestingly, both proteins are differentially expressed in hippocampus, indicating that they may contribute to specifying pre- and postsynaptic interactions in synapse formation.

The roles of SynCAMs at mature synapses have not yet been investigated. However, fractionation studies on SynCAM proteins are consistent with their synaptic localization in adult animals. Given the known

importance of adhesion molecules in the maturation and plasticity of synapses, it may be that SynCAM proteins play roles in these processes as well. Alternatively, SynCAMs might play an active role only in the initial induction of synapses and afterwards assume a static, structural function in synapse maintenance. These questions will need to be addressed in future experiments using mouse models.

Functions of SynCAM 1 *In Vivo*

Analysis of SynCAMs *in vivo* requires the use of mouse models in which the expression of one or more SynCAMs is altered. Thus far, only SynCAM 1 has been studied in this manner. SynCAM 1 knockout mice have been generated, and the main phenotype noted is infertility in the male homozygotes due to lack of retention of maturing spermatids at Sertoli cells. This condition is characterized by low sperm number and motility and aberrant sperm morphology (oligo-astheno-teratozoospermia) and is consistent with SynCAM 1's mediating cell adhesion during spermatid maturation. Similar phenotypes in testes have been noted for other Ig superfamily members.

No obvious phenotypes were noted for mice lacking SynCAM 1 in brain and lung, where it has characterized functions as synapse-inducing molecule and tumor suppressor, respectively. However, further studies will be needed to rigorously characterize synapse formation in mice deleted for SynCAM 1. It is plausible that SynCAM 1 knockout mice show no obvious defect in synapse formation because of redundancy *in vivo* between SynCAM 1 and other SynCAM proteins and because of compensatory mechanisms mediated by other synaptic adhesion molecules. In these cases, multiple knockouts will be necessary to probe the role of SynCAMs *in vivo*.

Summary and Outlook

SynCAM proteins form a family of four adhesion molecules belonging to the Ig superfamily. SynCAM 1 functions in synapse organization of developing neurons. Additional roles for SynCAM 1 exist outside the neurons. The expression of each SynCAM is unique during brain development, suggesting nonredundant roles for all four isoforms in the formation of central synapses. It is important to note that SynCAM proteins preferentially engage in heterophilic interactions with each other, suggesting these may be the key interactions involved in synapse formation. To date, only SynCAMs 1 and 2 have been studied extensively in synapse formation, and their role *in vitro* is well established. The interaction of SynCAM family members *in vitro* is also well understood, as is their basic expression. In order to understand the relevance of these described SynCAM interactions, more detailed expression analysis and functional studies will need to be performed. Such studies will likely provide leads for specific roles of each SynCAM protein in the developing CNS. Ultimately, it will be necessary to develop animal models to truly understand the role of these SynCAM proteins in synapse organization.

See also: Axonal Pathfinding: Netrins; Deafness; Dendrite Development, Synapse Formation and Elimination; Differentiation: the Cell Cycle Instead; Glutamate Receptor Clusters: Narp, EphB2 Receptor, Stargazin; Neurexins; Neuroligins; Vestibulo-Ocular Reflex.

Further Reading

Biederer T (2006) Bioinformatic characterization of the SynCAM family of immunoglobulin-like domain-containing adhesion molecules. *Genomics* 87: 139–150.

Biederer T (2006) SynCAM in formation and function of synaptic specializations. In: Dityatev A and EL-Husseini A (eds.) *Molecular Mechanisms of Synaptogenesis,* ch. 8, pp. 125–136.

Biederer T, Sara Y, Mozhayeva M, et al. (2002) SynCAM, a synaptic adhesion molecule that drives synapse assembly. *Science* 297: 1525–1531.

Butz S, Okamoto M, and Südhof TC (1998) A tripartite protein complex with the potential to couple synaptic vesicle exocytosis to cell adhesion in brain. *Cell* 94: 773–782.

Fogel AI, Akins MR, Krupp A, Stagi M, Stein V, and Biederer T (2007) SynCAMs organize synapses through heterophilic adhesion. *Journal of Neuroscience* 27: 12516–12530.

Fujita E, Kouroku Y, Ozeki S, et al. (2006) Oligo-asthenoteratozoospermia in mice lacking RA175/TSLC1/SynCAM/IGSF4A, a cell adhesion molecule in the immunoglobulin superfamily. *Molecular Cell Biology* 26: 718–726.

Kuramochi M, Fukuhara H, Nobukuni T, et al. (2001) TSLC1 is a tumor-suppressor gene in human non-small-cell lung cancer. *Nature Genetics* 27: 427–430.

Sara Y, Biederer T, Atasoy D, et al. (2005) Selective capability of SynCAM and neuroligin for functional synapse assembly. *Journal of Neuroscience* 25: 260–270.

Thomas LA, Akins MR, and Biederer T (in press) Expression and adhesion profiles of SynCAM molecules indicate distinct neuronal functions. *Journal of Comparative Neurology.*

Yamagata M, Sanes JR, and Weiner JA (2003) Synaptic adhesion molecules. *Current Opinion in Cell Biology* 15: 621–632.

Synfire Chains

M Abeles, Bar-Ilan University, Ramat-Gan, Israel

© 2009 Elsevier Ltd. All rights reserved.

What Is a Synfire Chain?

A synfire chain is essentially a feed-forward network of neurons with many layers (or pools). Each neuron in one pool feeds many excitatory connections to neurons in the next pool, and each neuron in the receiving pool is excited by many neurons in the previous pool. When activity in such a cascade of pools is arranged like a volley of spikes propagating synchronously from pool to pool, it is called a synfire chain. Figure 1 is an illustration of the connectivity in a synfire chain. The mean number of neurons in each pool is the width (w) of the chain. The mean number of synapses that each neuron in one pool receives from neurons in the previous pool (or by which it connects to neurons in the next pool) is the multiplicity of connections (m). In Figure 1, $w = 6$ and $m = 4$. In real cortical networks, w and m are assumed to be much larger, with $w \approx 100$ and $m \approx 50$.

In principle, one may envisage that activity along a synfire chain may propagate in either a synchronous or an asynchronous mode. In the asynchronous mode, an elevated firing rate in one pool will summate over space and time and cause an elevation in the firing rate in the next pool. In the synchronous mode, a synchronous volley in one pool will elicit a synchronous volley in the other after one synaptic delay. Abeles argued that the only stable mode of transmission is the synchronous mode. This was shown to be true theoretically, by numerical calculations, by simulations, and in the slice.

Thus, in most circumstances, in a multilayered feed-forward network of spiking neurons, the only stable mode of transmission is by synchronous volleys. Within each volley there is a small random dispersion of spike times, but this dispersion does not increase as the volleys propagate along the chain. This is so because at the transmission from pool to pool, the dispersion averages itself so that the distribution of firing times remains stable.

A synfire chain is a very simplified structure. There might be other structures, with less specific connections, which can generate precise firing sequences. Indeed, any network that can produce and recognize precise firing sequences would have all the advantages of a synfire chain. In a random diluted network, if one chooses w and m appropriately, one can find an almost infinite number of connectivity schemes similar to synfire chains by chance. However, the activity in such a network is not reproducible. If we start the activity by activating the same w neurons twice, even a very small amount of noise will cause the activity patterns to diverge very fast.

Historical Notes

The term synfire chain was coined by Abeles to account for the appearance of precise firing sequences with long interspike delays which resisted explanation in terms of the known properties of cortical physiology. This structure, with every neuron in one pool exciting all neurons in the second pool, was suggested 20 years earlier by Griffith as a structure that could guarantee a fixed level of activity in a network of excitatory neurons. He called this structure a complete transmission line. Griffith did not study its properties in any detail. Similar structures were presumably also used by Grossberg to learn and reproduce complicated space–time patterns. Bienenstock suggested that the transmission delays need not be equal. In such a system, a connection with a long delay may skip over one or more pools. He called such a network synfire braids. The governing rule for synchronous transmission in such a braid is that for any two neurons along the braid, the sum of all delays along each of the multitude (multineuronal) trajectories, by which one neuron may excite the other, adds up to the same value. A very simplified version of the synfire braid, in which each neuron is activated by a few strong excitatory inputs, was termed polychrony by Izhikevitch.

Memory Capacity for Synfire Chains

A single neuron may take part in more than one synfire pool. When activated with its members in a pool of one synfire chain, it will participate in activating the neurons of the next pool in this chain. When activated within a pool of another chain, it will activate the next pool in the other chain. Obviously, if all the neurons take part in all the pools of a network, we end up with an all-to-all connectivity. The problem of how many synfire pools can be embedded in a network of N neurons without interfering with the reproducibility of the activation was examined by Bienenstock and also by Herz and collaborators. Although they used two very different approaches, both came to similar conclusions: that the number of pools (P) is proportional to the network size (N), it decreases when the average firing rate in the network increases, and it depends on the width (w) of the pools. A cortical network with N neurons, an average

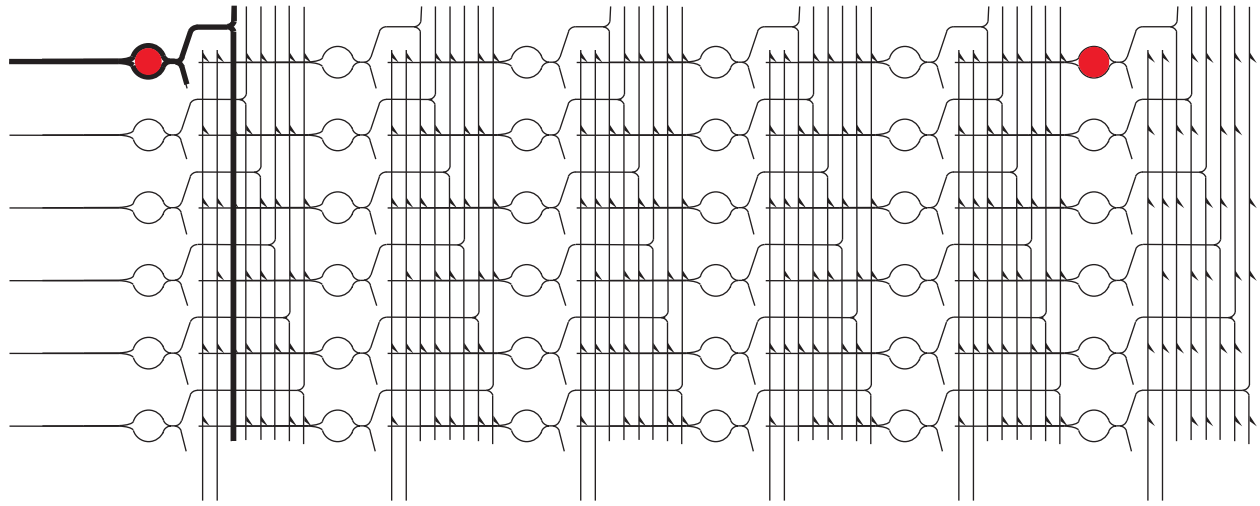

Figure 1 A synfire chain. Each neuron is depicted as a circle (the cell body), with a dendrite protruding to the left and a branching axon to the right. Synapses are symbolized by small triangles. The boldfaced drawing at the top left depicts one neuron. The neurons of each pool are drawn one under the other. The layout of the chain does not represent the anatomy but, rather, the order of activation in time. The same neuron can be part of more than one pool. For example, the two red neurons may be physically the same neuron. Repeated participation of the same neurons can occur up to the limit of the synfire chain's memory capacity.

firing rate of $5\,s^{-1}$, and with $w = 100$ may sustain N synfire pools ($P = N$). This means that each neuron can take part in 100 synfire pools.

These pools can be concatenated into one huge synfire chain or into many shorter chains. As noted by Maass, any structure that can be wired from individual neurons could just as easily be composed of synfire pools. The advantages would be more reliable operation, shorter input–output delays, and immunity to fallout.

Thus, an average cortical hypercolumn of 20 000 pyramidal neurons may contain 200 synfire chains, each 100 pools long and 100 neurons wide. At 5 spikes per second per neuron, this column would have 300 neurons firing every 3 ms. If the average transmission time between two pools is 3 ms, there could be three synfire waves active concomitantly. These may be distributed over three synfire chains or confined to repeated activation of one chain or any combinations thereof.

Although appealing, this idea failed to materialize when tested in a simulation of a large network of integrate and fire neuromimes, in which excitation and inhibition are balanced. Even when only one synfire chain was embedded in such a balanced network, the network went into a mode of synchronous oscillations. Attempts to remedy this failure took two forms. One solution was to also include inhibitory neurons in each pool. The output of these inhibitory neurons was not exclusively fed forward to the subsequent pool but, rather, fed back to the entire network. The other solution, put forward by Diesmann, was to use dispersed parameters for the neuromimes. In particular, using dispersed thresholds, synaptic potential sizes, and delays for the inhibitory neurons was very useful in avoiding global synchronization. It is interesting that in the past, anatomists classified cortical neurons into more than 60 types, only one or two of which were excitatory neurons (comprising 70–90% of cortical neurons), whereas all the others were nonpyramidal cells – almost all of which are the inhibitory neurons in the cortex.

When the same neuron participates in several pools within the same synfire chain, the chain is not a strictly feed-forward network anymore. At some level of such feedback connections, activity may start to reverberate within the synfire chain. However, here too, the reverberating activity is organized in synchronous volleys.

Computing with Synfire Chains

There have only been a few applications using synfire chains, including finding invariances in a picture, parsing an auditory scene, and parsing French sentences. However, the main application seems to lie in Bienenstock's suggestion to use synfire chains for implementing 'compositional' systems. He posits that only a few cross-links between two synfire chains that obey appropriate timing constraints are necessary to ensure that both synfire chains will act as a wider and more stable synfire chain. This wider synfire chain can in turn lock-in to activities of other synfire chains, etc. Thus, a large structure of synfire chains can be dynamically generated to represent binding of many simple components into a meaningful composite mental representation. This type of binding is dynamic so that a single synfire chain may, under different conditions,

lock-in to a number of other chains and therefore be part of numerous distinct composite representations. In Bienenstock's view, the most important feature of such dynamic binding is that it can be 'vertical,' in which a higher level concept binds to the lower level elements that comprise it.

Hayon has shown by way of simulation and by numerical analysis that hierarchical systems of synfire chains can be used to separate figure from ground, to implement a minimal description length principle, and to exhibit the property of compositionality. However, these demonstrations were for 'toy problems.' It remains to be shown whether synfire chains can efficiently solve such problems in real-world situations.

Abeles showed by way of simulations that two synfire chains with random connections may learn to lock-in to each other if activated synchronously several times and if the synaptic modifications follow a time-dependent synaptic plasticity (i.e., synapses are strengthened if the presynaptic spike precedes the postsynaptic one, and they are weakened if the time order is reversed). Such learning rules were experimentally reported and are dubbed TDSP. It should be noted that to a limited extent even a single neuron can learn to recognize specific spatiotemporal firing patterns, as demonstrated by Gutig and Sompolinsky's tempotron.

Self-Organization of Synfire Chains

Bienenstock and Doursat found that in a random network of excitatory neurons, with simple learning rules, if one seeds synchrony by occasionally exciting w neurons synchronously, a synfire chain will spontaneously be formed. In their model, the total synaptic strength that a neuron can produce or accept was limited. However, other researchers have had difficulties in growing synfire chains in a random network. Typically, one obtains short cyclical structures even with fine-tuning of the model parameters. Levy and Horn called the activity in such cyclic synfire chains distributed synchrony. Apparently, the optimal solution is to use the TDSP learning rule and limit the total cumulative synaptic strength that neurons may produce and receive to spontaneously grow long synfire chains. This was recently achieved by Jun and Jin. Itzhikevich easily grew a synfire braidlike structure in a random network when synapses were very strong so that synchronous excitation by very few neurons (two to four) was enough to reach threshold in the postsynaptic neuron. He called activity in such a network polychrony.

Experimental Evidence

Direct experimental support for the existence of synfire chains requires simultaneous recording of several neurons from one pool and several from the next pool under conditions in which the synfire chain is repeatedly activated. Using current techniques, one can simultaneously record up to approximately 100 neurons; however, these will be dispersed over a few cortical hypercolumns. This sampling is far too diluted. Thus, the most one can hope for is to record a few neurons from the same chain simultaneously. In this case, whenever this chain is activated, one expects to see a precise firing sequence of these neurons. Indeed, there have been multiple reports in the literature about observations of precisely repeating firing patterns.

With the synfire memory capacity described previously, each neuron may take part in up to 100 chains within a hypercolumn, which contains 200 chains. Thus, the chance that a given neuron will participate in a given chain is approximately 0.5. The chance that three neurons will participate in this chain is 0.125. Considering the fact that there are many different synfire chains and that one can record from more than three neurons simultaneously, there could be numerous different precise firing sequences in any recording site. This was reported to be the case in the slice and in behaving monkeys.

It should be noted that activity in synfire chains must result in precisely repeating firing patterns, but the existence of such patterns does not prove that activity is organized along synfire chains. There could be other networks that produce precise firing patterns. When recording intracellular activity from a neuron that is strongly affected by a repeatedly activated synfire chain, one expects to see repeating patterns of membrane voltage fluctuations. This was reported by Yuste to be the case in the slice.

Criticism

The claim that the reported precise firing patterns and repeated voltage fluctuations were real was challenged by several research groups that argued that the statistics used to detect such a pattern were flawed. An elegant solution was provided by Bienenstock and Geman. In their approach, the null hypothesis to be tested is that there is nothing precise about spike timing to within a time window of W ms. If true, then any statistic derived from the spike trains should not be significantly different from the same statistic derived after teetering each spike at random within W ms.

In order to test the null hypothesis, the spike trains need to be teetered many times, a preferred statistic must be computed after each teetering, and a histogram of the statistics must be built. If the statistic of the real data falls within the α percentile of the teetered values, then the null hypothesis can be rejected with a significance level of $1-\alpha$. This method also

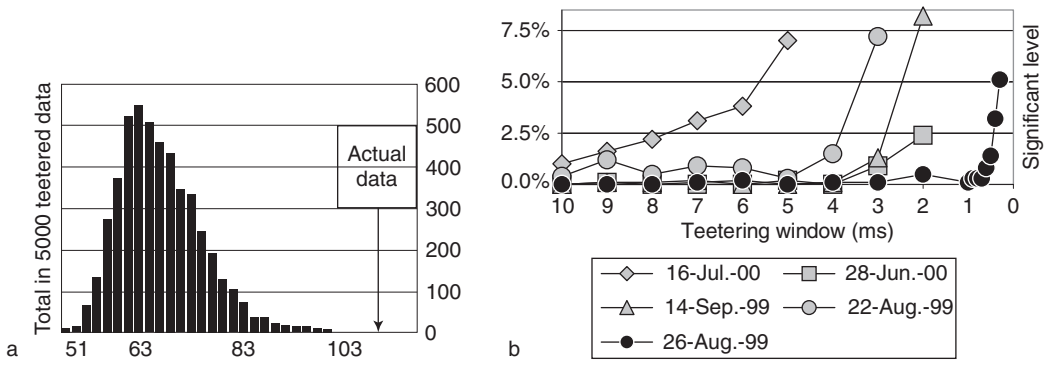

Figure 2 Estimating the precision of spikes. A statistic based on the most prominent peaks in cross-correlations was extracted to describe the time precision in the data sets. (a) The distribution of 5000 estimations of the statistic when times were teetered within 10 ms and the statistic of the actual data. Clearly, the probability of obtaining the statistic for real data by chance is less than 1 in 5000. (b) The probability that the statistic was obtained by chance as a function of the teetering window for 5 recording days.

allows for estimating the time precision of the spikes. By making the teetering window W increasingly smaller, one reaches a critical W at which the null hypothesis can no longer be rejected. This critical W is an upper bound for precision. Potentially, with another statistic, better precision could be found. Figure 2, based on the work of Shmiel et al., illustrates these ideas.

Gerstein suggested that a better method of teetering is to use the joint interspike interval for teetering each spike while considering the timing of the spike that preceded it and the spike that followed it. All types of teetering destroy precise timing relations; thus, hasty conclusions about precise firing patterns should be avoided. For example, a perfectly periodic spike train will contain a multitude of multispike patterns that repeat, and their number will be reduced considerably by random teetering.

Summary

Classical anatomy and physiology of the cortex support the idea that activity may be organized in synfire chains. Synfire chains generate and can learn to recognize sequences of coactivated neurons. Therefore, one can create compositional systems from synfire chains. It remains to be determined if, when, and where cortical activity is organized in synfire-like modes and whether real-world problems can be solved by synfire chains.

See also: Connectionist Models of Language Processing; Neurotransmitter Release: Synchronous and Asynchronous.

Further Reading

Abeles M (1982) *Local Cortical Circuits*. Berlin: Springer.

Abeles M (1991) *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge, UK: Cambridge University Press.

Bienenstock E (1995) A model of neocortex. *Network* 6: 179–224.

Bienenstock E (1996) Composition. In: Aertsen A and Braitenberg V (eds.) *Brain Theory: Biological Basis and Computational Properties*, pp. 269–300. Amsterdam: Elsevier.

Diesmann M, Gewaltig M-O, and Aertsen A (1999) Stable propagation of synchronous spiking in cortical neural networks. *Nature* 402: 529–533.

Hermann M, Herz J, and Prugel-Bennett A (1995) Analysis of synfire chains. *Network* 6: 403–414.

Synucleins

S Chandra, Yale University, New Haven, CT, USA

© 2009 Elsevier Ltd. All rights reserved.

Identification of Synucleins

The synuclein family is made up of three distinct genes, α-synuclein, β-synuclein, and γ-synuclein. The cardinal member, α-synuclein, was initially discovered in *Torpedo California*, by expression cloning, using antibodies against cholinergic synaptic vesicles. It was named as such as it was localized to presynaptic nerve terminals and artifactually to nuclei. Later, α-synuclein was also identified as the non-amyloid-β component of Alzheimer's disease plaques. Subsequently, a 14 kDa phosphoprotein identified in bovine and rat brains was shown to be another member of the synuclein family, namely β-synuclein. More recently, γ-synuclein was identified as breast cancer-specific gene 1, a gene upregulated in breast carcinomas.

Gene Organization

Synucleins are a vertebrate-specific gene family with no known members in other phyla. Human α-synuclein is located on chromosome 4q21.3–q22, β-synuclein on chromosome 5q35, and γ-synuclein on chromosome 10q23.2–q23.3. Phylogenic analyses of all synucleins show that α- and β-synucleins are more closely related to each other than to γ-synuclein. Still, all three genes have five coding exons. The human α-synuclein gene has two additional 5′ noncoding exons; the β-synuclein gene, one 5′ noncoding exon; and the γ-synuclein gene none. Human β- and γ-synuclein are not alternatively spliced. In contrast, human α-synuclein is alternatively spliced at the first two noncoding exons and at exons 4 and 6. The function of these splice forms is not yet known.

Tissue Expression Pattern

Synucleins are mainly found in neural tissue. Synucleins are abundant proteins and by some estimates account for 0.1% of total brain proteins. Both α- and β-synuclein are mainly found in brain, where they are enriched in presynaptic nerve terminals. They have a wide expression pattern and are expressed in the neocortex, hippocampus, striatum, thalamus, and cerebellum. In songbirds, α-synuclein expression is specifically increased, during periods of song acquisition, in key nuclei that control song learning. Meanwhile, γ-synuclein has a distinct expression pattern and is mainly found in peripheral nervous system, that is, in sensory, sympathetic, and motor neurons as well as the retina. Expression of γ-synuclein is also seen in breast carcinomas. This atypical expression is used as marker of progression of breast cancers as it is especially upregulated in metastatic cancers.

At a subcellular level, all synucleins are found in soluble synaptic fractions as well as associated with synaptic vesicle fractions. Immunoelectron microscopic analysis shows that antibodies to synucleins clearly decorate synaptic vesicles. The sequences that target synucleins to presynaptic terminals are currently unknown.

Structural Characteristics

Human α-synuclein has 140 amino acids, β-synuclein 134, and γ-synuclein 127 amino acids. Synucleins are natively unfolded proteins; that is, they have no secondary or tertiary structure in solution and hence are heat stable. The unusual biophysical properties of synucleins are rooted in their characteristic primary sequence (Figure 1). All synuclein proteins have a highly conserved N-terminal domain with an 11-residue repeat and a less conserved, highly acidic C-terminal domain. The seven imperfect 11-residue repeats have the consensus sequence XKTKEGVXXXX and are highly reminiscent of amphipathic helices of apolipoproteins (class A2). The only divergence from this N-terminal repeat structure occurs in β-synuclein, which has an 11-amino-acid deletion (amino acids 73–83), with profound consequences to its aggregation properties (see the section titled 'Oligomerization and fibrillization'). The central portion of α-synuclein is hydrophobic and corresponds to a fragment known as non-amyloid component (amino acids 61–95; Figure 1), previously isolated from Alzheimer's diseased brains. The C-terminus of synucleins is highly acidic and has many prolines.

Though synucleins are natively unfolded proteins, synucleins can bind and fold on acidic phospholipid surfaces via their N-terminal domain. This folding can also be replicated in anionic detergents such as sodium dodecylsulfate. The structure of the folded conformation of human α-synuclein has been determined using protein nuclear magnetic resonance and electron paramagnetic resonance methods. On sodium dodecylsulfate micelles, the folded conformation consists of two amphipathic α-helices with an unfolded C-terminus. The two α-helices are oriented in an antiparallel arrangement. Recently, human β- and γ-synuclein were also shown to fold into this two-helix conformation on sodium dodecylsulfate

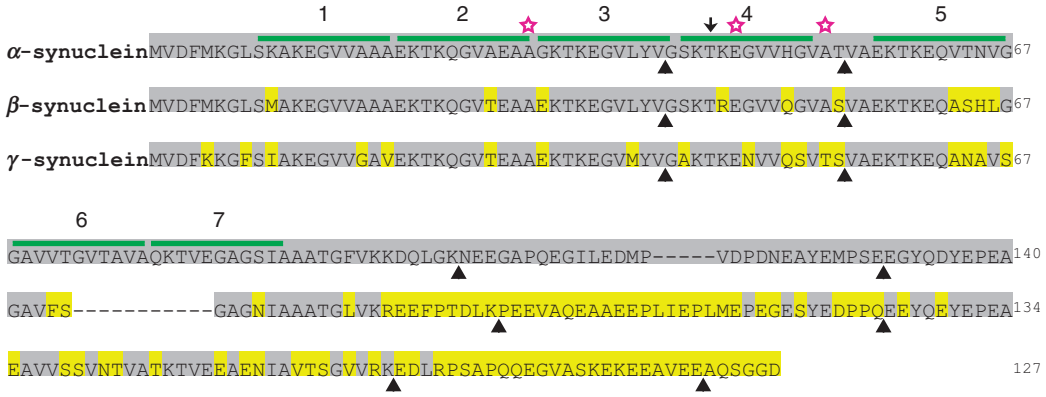

Figure1 Sequence alignment of human α-synuclein, β-synuclein, and γ-synuclein protein sequences. Residues are colored gray to indicate identity across all three synucleins. The green lines outline the seven 11-amino-acid repeats, and the arrowheads show exon boundaries. The stars indicate the position of the three known PD mutations, A30P, A53T, and E46K and the arrow indicates position of the break between the two helices in the α-helical conformation of human α-synuclein.

micelles. On liposomes which have a shallower curvature, human α-synuclein still adopts this two-helix conformation. In this conformation, the inner surface of the helices is apolar, with the basic residues pointing sideways from the helix. This allows the inner surface to interact with apolar detergent/lipid tails and concomitantly for the basic side chains to interact with the acidic head groups. These interactions explain the acidic lipid specificity of α-synuclein binding. In this structure, the C-terminus does not bind lipids/detergents and remains unfolded. As this conformation is conserved among all synucleins, it is likely to be relevant for their physiological functions.

α-Synuclein is phosphorylated, most prominently at Ser 129, which increases its ability to aggregate and form fibrils. α-Synuclein and β-Synuclein are also O-glycosylated. The function of this modification is not well understood.

Oligomerization and Fibrillization

Human α-synuclein can readily oligomerize to higher-order species: first to dimers, then trimers, and then oligomers, culminating in large fibrils or amyloids. Conditions that promote the helical conformation of α-synuclein, such as acidic lipids and polyunsaturated fatty acids, also promote oligomerization of α-synuclein. Posttranslational modifications such as phosphorylation and C-terminal truncations, incubation with metal ions such as iron, and oxidative/nitrative challenges also accelerate fibril formation. It is interesting that human γ-synuclein oligomerizes very slowly, and human β-synuclein does not oligomerize or aggregate. In fact, addition of β-synuclein decreases the rate at which α-synuclein oligomerizes. Sequence divergences, such as the 11-amino-acid deletion in the non-amyloid component region of β-synucleins, have been shown to underlie these differences. Another interesting observation is that mouse α-synuclein oligomerizes very poorly. There are seven amino acid changes between these two species that probably account for this difference.

Protofibrils are transient β-sheet-containing oligomers that are formed during fibrillization. Protofibrils have heterogeneous morphologies, including spherical, annular, porelike, tubelike, and chainlike structures. Protofibrils are the precursors to stable macroscopic β-sheeted fibrils or amyloids seen in α-synuclein pathology (see the sections titled 'Alpha-synuclein and Parkinson's disease' and 'Synucleinopathies'). Protofibrillar α-synuclein, in contrast to the monomeric and the fibrillar forms, can bind liposomes via a β-sheet-rich structure and may even act as pores in the membrane. This finding has led to the hypothesis that α-synuclein protofibrils are neurotoxic.

Thus, human α-synuclein is unique among synucleins in the distinct conformational states it can occupy. Several of these conformations are physiological while others are pathological (Figure 2).

Alpha-Synuclein and Parkinson's Disease

Strong genetic and pathological evidence links the human α-synuclein gene to Parkinson's disease (PD). Three independent mutations in the human α-synuclein gene, namely the A30P, A53T, and E46K, have been linked to early-onset familial PD. These mutations are dominant in nature and hence represent a 'gain of function.' Recently, families with duplication and triplication of the α-synuclein locus have also been identified. Even though these patients have a normal wild-type gene, they develop an autosomal dominant

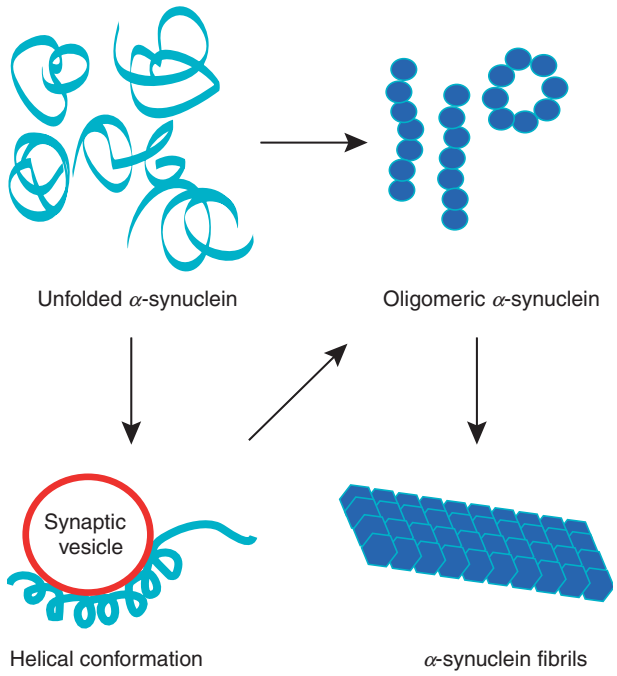

Figure 2 Model for fibrillization pathway of human α-synuclein. Unfolded monomeric α-synuclein folds on synaptic vesicles in a helical conformation. α-Synuclein can also oligomerize to higher-order species including β-sheeted protofibrils. Conditions that promote the helical conformation of α-synuclein also promote its oligomerization. Eventually, these oligomers form large β-sheeted fibrils or amyloids such as those seen in Lewy bodies.

form of PD. The severity and onset of the disease in these patients appears to depend on levels of α-synuclein protein. The discovery of these mutations and duplications has prompted the search for genetic polymorphisms within the α-synuclein locus and the subsequent testing to determine whether these are risk factors for the disease. So far the data are partly contradictory and far from conclusive.

α-Synuclein is a component of Lewy bodies, the defining pathological hallmark of PD for both familial and sporadic cases. Lewy bodies are eosinophilic cytoplasmic inclusions containing α-synuclein, usually in its ubiquitinated form, and synuclein-interacting proteins such as synphilin, as well as heat shock proteins. Lewy bodies may represent end points in a neuron's attempt to cope with an aggregation-prone protein. There is much debate in the literature over whether Lewy bodies themselves are toxic to dopaminergic neurons. Increasingly, recent data suggest that soluble oligomers of α-synuclein (protofibrils) are the culprits and not large aggregates, as represented by Lewy bodies. One of the strongest pieces of evidence is that the A30P mutation actually slows fibrillization, while the A53T mutation accelerates fibrillization *in vitro*. However, both the A53T and A30P mutants promote protofibril formation, compared with wild-type α-synuclein. Interestingly, dopamine reacts with α-synuclien to form covalent adducts that slow the conversion of protofibrils to fibrils, giving further credence to this hypothesis.

The identification of point mutations in the human α-synuclein gene that cause familial PD has led to many of animal models for PD overexpressing wild-type and parkinsonian mutants of α-synuclein. These models have been generated in a variety of organisms, such as mouse, rat, *Drosophila*, *C. elegans*, zebra fish, and even yeast. It should be noted that several of these lower organisms do not have the α-synuclein gene in their genomes. These models have had mixed success in replicating the characteristics of PD, namely, the dopaminergic neuron selectivity, the α-synuclein pathology, and age dependence. The murine models, depending on the promoter used, show late-onset neuronal cell death and some synuclein pathology/inclusions, but the selectivity of nigrostriatal loss is largely absent. It is interesting that in one such α-synuclein transgenic mouse line, co-expression of β-synuclein inhibited aggregation of α-synuclein, mirroring the *in vitro* oligomerization data. Surprisingly, the best model to date that replicates all three characteristics is the Drosophila model that overexpresses human α-synuclein. Co-expression of the chaperone Hsp70 or pharmacological upregulation of heat-shock factors in this model ameliorates dopaminergic cell loss but does not affect inclusion formation.

In addition to transgenics, viral expression systems have also been used to introduce the α-synuclein gene to specific regions of the midbrain. Lentiviral-based expression of human α-synuclein in rat substantia nigra resulted in selective dopaminergic toxicity with nonfibrillar inclusions. Thus transgenic and viral expression systems, in spite of their limitations, have been very useful to test various hypotheses on the causes of PD, such as protein aggregation, mitochondrial dysfunction, and oxidative stress. These models will continue to be very useful as drug discovery platforms.

The last decade has been very fruitful in finding genes linked to rare familial forms of PD. Starting with the identification of the A53T mutation in α-synuclein, several genes, such as parkin, PINK, DJ-1, and LRRK2, have been shown to cause both recessive and dominant forms of PD. These genes have diverse functions in neurons; for example, parkin is an E3 ubiquitin ligase; PINK, a mitochondrial kinase; LRRK2, a leucine-rich-repeat containing tyrosine kinase; and DJ-1 appears to be a mitochondrial redox sensor. These discoveries have renewed interest

in understanding the ubiquitin–proteasome and mitochondrial pathways and their relationship to the nigrostriatal dopaminergic loss seen in PD. The challenge is to determine whether the various PD genes interact with α-synuclein and lead to a pathway of dopaminergic neuron death.

Synucleinopathies

α-Synuclein aggregates, and pathology is associated with a variety of neurodegenerative disorders besides PD. These diseases include Alzheimer's disease, especially the Lewy body variant, dementia with Lewy bodies; multiple system atrophy; Down syndrome; and Halloverden–Spatz disease. Pathologically, these diseases are now known as synucleinopathies. The pathology seen in these patients' brains is diverse and includes Lewy bodies, amyloid-β plaques, and Lewy neurites. In many of these lesions, synuclein is ubiquitinated or nitrosylated or exists as adducts of dopamine. In spite of their diversity, these pathologies represent common pathogenic pathways and as such are useful to study mechanistically.

Physiological Functions

The presynaptic localization of synucleins and the fact that they are vertebrate-specific suggest that synucleins have a modulatory, rather than an essential, synaptic function. The initial discovery of α-synuclein as a gene upregulated in key nuclei that control songbird learning also suggests that synucleins are important for synaptic plasticity. This premise has been tested in single- and double-knockout mice for α-synuclein and β-synuclein. While, some of the data are contradictory, the overall consensus is that synucleins are not essential for neurotransmission but negatively regulate dopamine transmission in certain paradigms. In keeping with this, α- and β-synuclein double-knockout animals have decreased levels of striatal dopamine levels.

The dopaminergic neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) causes parkinsonism in both humans and mice. MPTP is transported via the dopamine transporter into nigrostriatal neurons, where it inhibits mitochondrial electron transport. Strikingly, all synuclein knockout mice tested to date (α- and γ-synuclein) are completely resistant to MPTP. Unearthing the mechanisms of this resistance will greatly help in understanding synuclein function and may also shed light on the dopamine selectivity of PD.

Surprisingly, the best clues to the physiological function of synucleins have come from analyzing transgenic mice that overexpress α-synuclein. Paraquat is a herbicide that when administered to mice causes the formation of cytosolic α-synuclein aggregates and the degeneration of nigrostriatal neurons. Mice that express human α-synuclein driven by the tyrosine hydroxylase promoter have been shown to be resistant to the herbicide paraquat. These data were the first to hint at a neuroprotective function of synucleins. More recently, in a genetic study, transgenic overexpression of α-synuclein was shown to completely rescue the neurodegeneration and lethality of knockouts for the presynaptic cochaperone CSPα. Phospholipid binding of synucleins was shown to be important for this rescue. Conversely, deletion of endogenous synucleins exacerbated the CSPα knockout phenotypes. Biochemical analysis of these crosses revealed that synucleins function downstream of CSPα to stabilize/fold substrates of CSPα. These data collectively support a neuroprotective role for synucleins and suggest that the involvement of α-synuclein in human neurodegenerative diseases arises not only from a gain of toxic function but also from a loss of synapse-protective properties.

Prospects

Future research on synucleins needs to address aspects of both their physiological and their pathological functions. In particular, it should (1) define the neuroprotective function of synucleins mechanistically, (2) elucidate the pathological conformations of human α-synuclein, and (3) determine the interplay between human α-synuclein and other familial PD genes.

See also: Alzheimer's Disease: Neurodegeneration; Amyloid: Vascular and Parenchymal; Animal Models of Parkinson's Disease; Cell Replacement Therapy: Parkinson's Disease; MPTP Parkinsonism Model; Oxidative Damage in Neurodegeneration and Injury; Parkinsonian Syndromes; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Role of NO in Neurodegeneration; Vascular Issues in Neurodegeneration and Injury.

Further Reading

Abeliovich A, Schmitz Y, Farinas I, et al. (2000) Mice lacking alpha-synuclein display functional deficits in the nigrostriatal dopamine system. *Neuron* 25: 239–252.

Chandra S, Gallardo G, Fernandez-Chacon R, Schluter OM, and Sudhof TC (2005) Alpha-synuclein cooperates with CSPalpha in preventing neurodegeneration. *Cell* 123: 383–396.

Clayton DF and George JM (1998) The synucleins: A family of proteins involved in synaptic function, plasticity, neurodegeneration and disease. *Trends in Neurosciences* 21: 249–254.

Conway KA, Lee SJ, Rochet J-C, Ding TT, Williamson RE, and Lansbury PT Jr. (2000) Acceleration of oligomerization, not fibrillization, is a shared property of both alpha-synuclein mutations linked to early-onset Parkinson's disease: Implications for pathogenesis and therapy. *Proceedings of the*

National Academy of Sciences of the United States of America 97: 571–576.

Dauer W, Kholodilov N, Vila M, et al. (2002) Resistance of alpha-synuclein null mice to the parkinsonian neurotoxin MPTP. *Proceedings of the National Academy of Sciences of the United States of America* 99: 14524–14529.

Dauer W and Przedborski S (2003) Parkinson's disease: Mechanisms and models. *Neuron* 39: 889–909.

Feany MB and Bender WW (2000) A Drosophila model of Parkinson's disease. *Nature* 404: 394–398.

Iwai A, Masliah E, Yoshimoto M, et al. (1995) The precursor protein of non-A beta component of Alzheimer's disease amyloid is a presynaptic protein of the central nervous system. *Neuron* 14: 467–475.

Jao CC, Der-Sarkissian A, Chen J, and Langen R (2004) Structure of membrane-bound alpha-synuclein studied by site-directed spin labeling. *Proceedings of the National Academy of Sciences of the United States of America* 101: 8331–8336.

Ji H, Liu YE, Jia T, et al. (1997) Identification of a breast cancer-specific gene, BCSG1, by direct differential cDNA sequencing. *Cancer Research* 57: 759–764.

Kruger R, Kuhn W, Muller T, et al. (1998) Ala30Pro mutation in the gene encoding alpha-synuclein in Parkinson's disease. *Nature Genetics* 18: 106–108.

Lee VM-Y and Trojanowski JQ (2006) Mechanisms of Parkinson's disease linked to pathological alpha-synuclein: New targets for drug discovery. *Neuron* 52: 33–38.

Manning-Bog AB, McCormack AL, Purisai MG, Bolin LM, and Di Monte DA (2003) Alpha-synuclein overexpression protects against paraquat-induced neurodegeneration. *Journal of Neuroscience* 23: 3095–3099.

Polymeropoulos MH, Lavedan C, Leroy E, et al. (1997) Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* 276: 2045–2047.

Singleton AB, Farrer M, Johnson J, et al. (2003) alpha-Synuclein locus triplication causes Parkinson's disease. *Science* 302: 841.

Spillantini MG, Schmidt ML, Lee VM, Trojanowski JQ, Jakes R, and Goedert M (1997) Alpha-synuclein in Lewy bodies. *Nature* 388: 839–840.

Ulmer TS, Bax A, Cole NB, and Nussbaum RL (2005) Structure and dynamics of micelle-bound human alpha-synuclein. *Journal of Biological Chemistry* 280: 9595–9603.

Syringomyelia

S Sgouros, Birmingham Children's Hospital, Birmingham, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Syringomyelia is the cystic cavitation of the spinal cord extending over a distance greater than two spinal segments (see Figure 1). Often the cavity is referred to with the term 'syrinx,' an ancient Greek word meaning 'tube,' also used for an ancient Greek wooden musical instrument similar to a flute. The syringomyelia cavity contains fluid with a consistency similar to cerebrospinal fluid (CSF). Although the condition was described in 1876 by Leyden and in 1891 by Abbe and Coley, the pathophysiology surrounding the formation of the syrinx and its propagation is still unclear. The advent of magnetic resonance imaging (MRI) has contributed greatly to the understanding of syringomyelia, and its wide use has resulted in an apparent increase in the incidence of diagnosis in adults and children.

Syringomyelia is most frequently associated with hindbrain herniation. Hindbrain herniation, or Chiari I malformation, is the prolapse of the cerebellar tonsils for 5 mm or more beyond the foramen magnum (see Figure 1). In large clinical series, 50–80% of patients with symptomatic Chiari I malformation have syringomyelia. In most patients, Chiari I malformation is regarded as congenital. Acquired Chiari I malformation can develop following lumbar or ventricular shunting. Syringomyelia may be associated also with various other conditions, such as spinal cord tumors, spinal arachnoid cysts, traumatic paraplegia, postinfectious or postinflammatory meningitis (tuberculous meningitis, chemical insults such as antibiotics or Myodil (Pantopaque), epidural abscess, Pott's disease, or idiopathic meningeal fibrosis), various forms of spinal dysraphism (diastematomyelia, spinal lipoma, and recurrent tethering after myelomeningocele repair), and rare malformations affecting the foramen magnum region (achondroplasia, posterior fossa arachnoid cysts, Chiari II malformation, and syndromic craniosynostosis). Syringomyelia commonly affects the cervical and upper thoracic spinal cord, although it can extend throughout the entire spinal cord, and in that case is called holocord syringomyelia. If untreated it follows a progressive course over a number of years. The neurological function of arms and legs is at threat, and patients may eventually become quadriplegic if the syringomyelic cavity continues to extend despite treatment. The cavity may extend both proximally and distally, and the tension of the fluid inside it continues to increase. Long-standing syringomyelia provokes gliosis formation in the cavity walls and the surrounding cord, seen in postmortem examinations. Gliosis inside the cord may lead to progressive neurological loss even after successful surgery and collapse of the cavity. The syringomyelic cavity can reach the upper cervical cord and medulla, in which case the condition is called syringobulbia (see Figure 2), and eventually affect the function of the lower brain stem, producing respiratory failure and death.

Epidemiology

The incidence of Chiari I malformation in the general population is estimated at 7% although not all patients are symptomatic. Syringomyelia has a prevalence of 8.4 cases per 100 000 population. The incidence of posttraumatic syringomyelia is around 1–3% of all paraplegics; most of these patients are male and relatively young. The average time between spinal injury and symptomatic presentation is 7 years (range: 3 months to 30 years).

Pathophysiology of Syringomyelia

In the majority of patients, syringomyelia is associated with an obstruction of CSF flow in the subarachnoid space. Surgical treatments are designed to remove such a subarachnoid obstruction. The block in CSF flow is in most patients in the region of the foramen magnum due to hindbrain hernia (Chiari I mostly, Chiari II infrequently) or in other sites in any of the other conditions mentioned already.

The mechanism of formation of syringomyelia has not been adequately explained yet. In experimental studies, dye injected in the spinal subarachnoid space has been found in the central canal and is believed to have traveled there along the Virchow-Robin spaces beside the vessels of the cord. There have been reports of a presyringomyelic state, seen in MRI scans of patients with Chiari I malformation and upper cervical spinal cord swelling, which subsequently developed syringomyelia. Several theories have been proposed on the formation and propagation of syringomyelia. The most notable ones, in chronological order, are Gardner's hydrodynamic theory, Williams's theory of 'pressure dissociation,' and Oldfield's theory of increased subarachnoid CSF pressure.

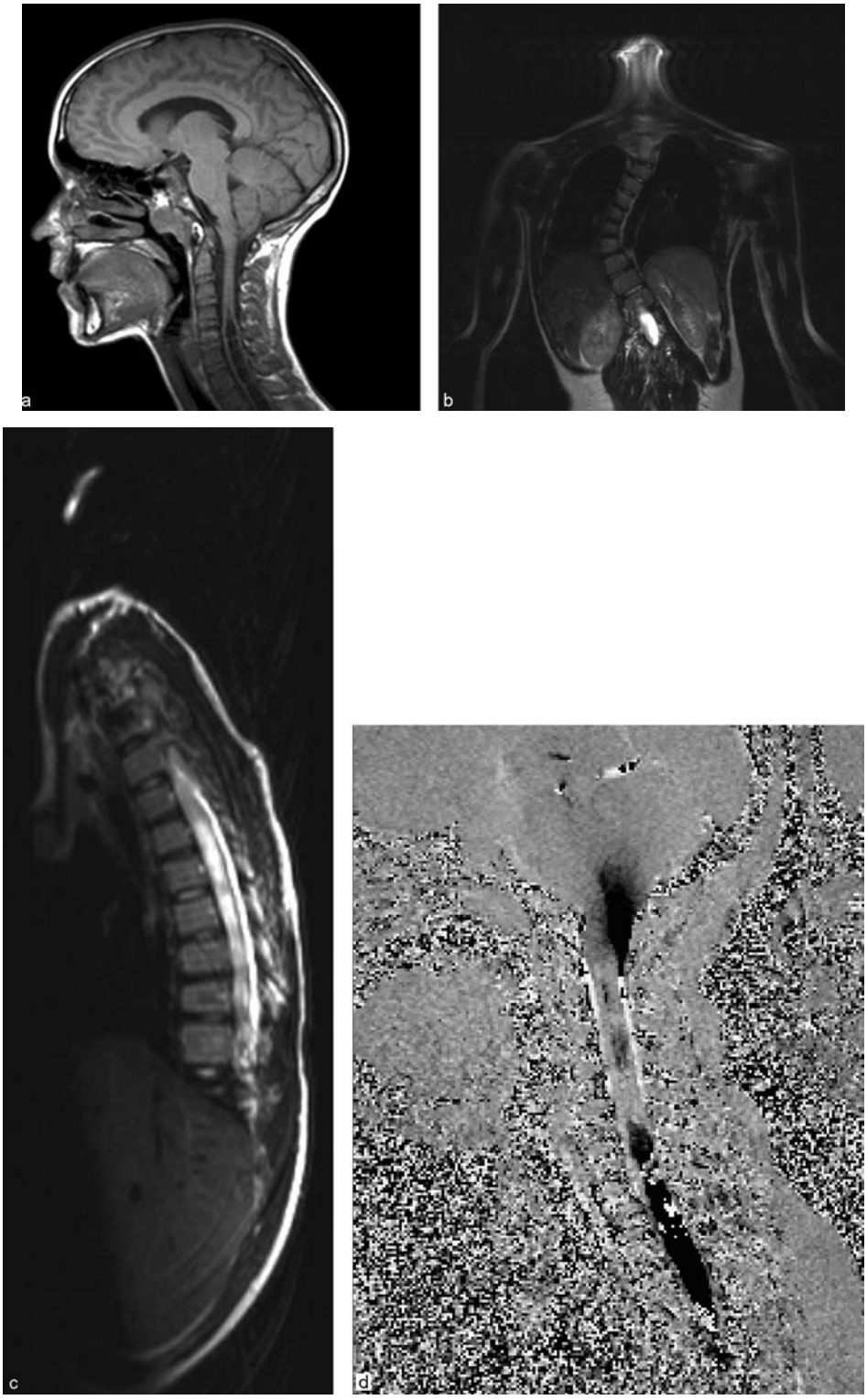

Figure 1 MRI scans of a 14-year-old girl presenting with clumsy walking, occipital headaches, and scoliosis. (a) Sagittal T1w image. There is a Chiari I malformation with tonsillar descent down to the arch of C1 vertebra and a large syringomyelia cavity extending from the C5 vertebra to the lower thoracic region. (b) Coronal body view showing marked scoliosis. (c) Sagittal T2w image through the spine. Because of the scoliosis, the spinal cord cannot be visualized in one sagittal image. The syringomyelia cavity is seen but the exact lower end of it cannot be estimated. (d) Still image from the cine phase contrast magnetic resonance imaging loop: abnormal CSF flow can be seen at the craniovertebral junction and inside the syringomyelic cavity.

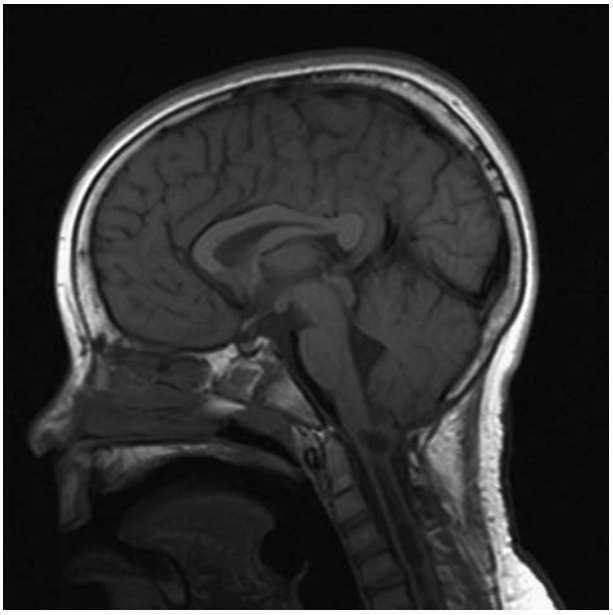

Figure 2 Sagittal T1w magnetic resonance imaging scan of an 11-year-old boy with previously treated Chiari I malformation and syringomyelia who represented with double vision and unilateral facial palsy. There is recurrent arachnoiditis at the foramen magnum, and the syringomyelia has extended upwards and invaded the medulla, becoming syringobulbia.

- In the 1960s Gardner observed that in patients with hindbrain hernia, the foramen of Magendie was partially obstructed by membrane. He believed that fluid entered the central canal in the obex at the floor of the fourth ventricle due to the 'water hammer' effect of the arterial pulsations on CSF flow. Fluid accumulated inside the central canal, creating the syringomyelic cavity – a condition called 'communicating syringomyelia' or 'hydromyelia.' This led to the operative concept of obex plugging, in which a piece of muscle or cotton wool was used to plug the obex. Syringomyelia not communicating with the central canal was called noncommunicating. MRI studies have shown that in most patients, there is no direct communication between the obex and the syringomyelic cavity. Cardiac-gated cine phase contrast MRI (PC-MRI) studies have shown that during systole, the syrinx cavity contracts, whereas if Gardner's theory were correct, it should expand under the force of inflowing CSF. Gardner's theory is regarded as unsatisfactory nowadays, and the term 'communicating syringomyelia' has been abandoned.
- In the early 1980s, Williams described the concept of 'pressure dissociation' between the head and the spine secondary to the block of hindbrain hernia, based on *in vivo* CSF pressure measurements and *in vitro* biomechanical studies. Every time a Valsalva maneuver takes place (e.g., coughing), the pressure from the chest and abdomen is transmitted to the spinal canal through the valveless venous plexus around the vertebral bodies. Normally pressures are equalized rapidly within the spine. In the presence of a partial subarachnoid block, fluid is forced upwards past the block more efficiently than it can run down again. This leads to a collapsed theca below the block, which exerts a suction effect on the spinal cord (the 'suck' mechanism). Fluid that has entered the cord cavity can extend the cavity (the 'slosh' mechanism). Williams's theories were partly verified by MRI scan observations. Cine PC-MRI has shown that in the presence of hindbrain hernia, CSF movement in the region of the foramen magnum differs between systole and diastole. Williams's observations of pressure differential were not confirmed by other studies of intraoperative CSF pressure measurements, however, and is believed that the sitting position his patients used when he obtained his readings contributed to exaggerated cranial-spinal pressure difference readings due to CSF leakage from the site of the spinal measuring needle.
- Using preoperative PC-MRI scanning and intraoperative CSF pressure measurements in patients with syringomyelia and Chiari I malformation, Oldfield's group observed that CSF velocity was increased at the foramen magnum but CSF flow was decreased, cervical subarachnoid CSF pressure was increased, spinal CSF compliance was decreased, and syrinx CSF flowed caudally during systole and cranially during diastole. The group surmised that the prolapsed cerebellar tonsils act like a piston, partially occluding the subarachnoid space and create a state of increased subarachnoid pressure waves, which compress the spinal cord from without and force CSF into the syrinx cavity and propagate it with each heartbeat.

All these theories do not explain exactly how and where CSF enters the cord and the mechanism of formation of septations inside the syringomyelic cavity.

In patients who suffer spinal cord trauma, cord contusion leads to edema, blood effusion, and subsequent liquefaction, leading in up to 50% of patients to cavity formation, the 'primary cyst.' Commonly the cyst at the site of the injury is separated from the syringomyelia by an intact septum. The syrinx may go flat above and below the injury site following treatment, and the primary cysts may remain unchanged. As the acute injury settles, inflammation is followed by adhesions in the subarachnoid space and gliosis in the cord. Obstruction of the CSF pathways may be compounded by narrowing of the bony spinal canal due to the spinal fracture. Fluid may enter the cord at the fracture site, where the cord is usually thin-walled.

For forms of syringomyelia associated with other pathologies (dysraphic states, spinal cord tumors, and spinal arachnoid cysts), there has been no satisfactory explanation of how CSF enters the spinal cord, and an analogy to hindbrain-related syringomyelia is currently accepted. The presence of arachnoid adhesions appears important in the formation of syringomyelia.

Posterior Fossa Volume

It is postulated that hindbrain hernia is caused by a small posterior fossa. Early studies on lateral skull radiographs and later studies on computed tomography or MRI scans showed that posterior fossa volume and dimensions are smaller than normal in patients with Chiari I malformation due to presumed maldevelopment of the occipital enchondrium. A recent study focusing on children demonstrated that patients with Chiari I and syringomyelia have posterior fossa volume significantly smaller than normal whereas those without syringomyelia have normal posterior fossa volume, especially in the first 10 years of life. These findings challenge the view that the development of syringomyelia represents an evolutionary stage of the Chiari I malformation.

Clinical Features

Presenting Symptoms

Patients with hindbrain-related syringomyelia often present with symptoms of Chiari I malformation, and the syringomyelia is discovered on MRI scan. The symptoms can be grouped into those of pain, sensory and motor deficits, and disordered autonomic functions. Severity of symptoms can vary, and there are patients who remain completely symptom free for long periods.

Symptoms associated with Chiari I malformation include headaches that are usually, but not always, occipital and that worsen on physical exertion and straining; occasional clumsiness and poor coordination with hands; and ocular symptoms such as blurred vision, diplopia, or oscillopsia. Other, rare symptoms are dizziness, unsteadiness (disequilibrium), tinnitus, vertigo, hyperacusis, nystagmus, dysphagia, sleep apnea, dysarthria, tremor, facial pain or numbness, shortness of breath, and syncopal (drop) attacks.

Symptoms related to syringomyelia include weakness or clumsiness in arms and legs, paresthesia, dysesthesia and pain, unsteady gait, and muscle atrophy and spasticity. Pain is commonly a prominent feature. In a small percentage of patients there may be urinary and fecal incontinence and impotence in male patients. Syringobulbia commonly presents with spreading facial numbness. Swallowing and voice may also be affected, then vision, hearing and finally respiratory function.

Clinical Examination Findings

In patients with Chiari I malformation or syringobulbia, clinical examination findings include lower cranial nerve and cerebellar signs. A small minority have impaired visual acuity, extraocular muscle palsy, nystagmus, and papilledema. Other signs from lower cranial nerves include facial sensory loss, sensorineural or conductive hearing loss, impaired vestibular function, vocal cord palsy, and impaired gag reflex. Cerebellar signs such as dysmetria or truncal ataxia are not uncommon.

In patients with syringomyelia, long tract signs are common. Patients can have sensory deficit over arms and/or legs of spinothalamic and/or dorsal column type, not always following dermatomal distribution. It is usually unilateral and tends to involve the upper limb early. The traditionally described pain–temperature dissociation due to loss of pain and temperature sensation with preserved light touch and joint position sense is rare. Patients can have weakness of arms and/or legs, often unilateral, not always following a myotomal distribution. In advanced cases, there may be claw deformity in the hand with wasting of all the forearm and arm musculature and fasciculations. Dystrophic (Charcot) joints can be seen. Hyper- or hypo-reflexia can be present, often asymmetrical. Spasticity can affect any limb. Scoliosis is common in children and young adults with syringomyelia. Short neck and low hair line are seen in patients with basilar invagination. Horner's syndrome is seen infrequently. Gross disturbance of sweating can occur, either dry skin or hyperhidrosis.

In patients with spinal cord injury, incomplete paraplegia that deteriorates or ascends and the development of new autonomic features may imply syringomyelia. In complete paraplegia, improvement of leg spasms, alteration of sweating patterns, and impairment of other autonomic functions such as bladder and bowel control and sexual function could imply downward extension of the syrinx. Pain is common in posttraumatic syringomyelia; a paraplegic complaining of severe pain has syringomyelia until proven otherwise. Syringomyelia can be confused with chest or abdominal pathology.

Radiology

MRI scan is the investigation of choice and confirms the diagnosis in most cases. Computed tomography myelography, acquiring late views to visualize contrast in the syringomyelia cavity, remains useful in patients

with severe claustrophobia or magnetic implants. Plain radiographs of the spine are useful in patients with scoliosis, bony malformations, or posttraumatic kyphosis. Flexion–extension views can provide information on possible spinal instability.

MRI scanning shows the Chiari malformation and the extent of the syringomyelia cavity and septa, especially T2-weighted sequences (see Figure 1). Around 15% of patients have hydrocephalus. Other pathologies are clearly demonstrated. Children with profound scoliosis can be difficult to visualize in MRI scanning. As the curvature of the spine can be extensive, it may not be possible to have the entire spinal cord in a single sagittal view, and it may be difficult to correctly count the vertebral levels and identify the full extent of the syrinx (see Figures 1(b) and 1(c)). Children with significant scoliosis can have syringomyelia in the absence of Chiari I malformation. Posttraumatic syringomyelia and its relation to primary cysts shows well in MRI scans (see Figure 3), although in some patients with extensive spinal fractures, it may be difficult to identify the complexity of the bony injury. Diastematomyelia and spinal lipomas are clearly seen, and commonly the syringomyelia cavity is not extensive (see Figure 4). Spinal cord tumors are clearly identified in contrast-enhanced MRI sequences. Whenever a patient is found to have syringomyelia in the absence of a clearly identifiable associated lesion, contrast-enhanced sequences should be obtained to exclude the presence of spinal cord tumors. In a small minority of patients, no associated lesion is identified. Information relating to CSF circulation can be obtained with cardiac-gated cine PC-MRI (see Figure 1(d)), and its correlation with clinical outcome investigated. Sequential MRI imaging is useful in assessing the results of treatment or in observing patients over time.

Figure 3 Sagittal T1w magnetic resonance imaging scan of a quadriparetic 24-year-old man who suffered a cervical spine fracture during a car accident. His arm function deteriorated 2 years later. There is a wedge fracture of the body of C5 vertebra. There is a large syringomyelia cavity extending from C4 to T1 vertebrae.

Neurophysiological Tests

Nerve conduction studies, brain stem evoked potentials (BSEP), and somatosensory evoked potentials (SSEP) have not proved very useful in clinical practice as usually they are abnormal after clinical symptoms or signs have been established. BSEP and SSEP can be useful intraoperatively in selected patients, such as patients with spinal cord tumors.

Natural History of Hindbrain Hernia and Syringomyelia

Controversy still surrounds the management of hindbrain hernia and syringomyelia, largely due to the lack of prospective studies outlining their natural history. Nobody has ever followed 'asymptomatic' patients with hindbrain hernia for several years to see when they will develop symptoms and how many of them will develop syringomyelia. Experience suggests that completely asymptomatic patients with Chiari I malformation only do not develop symptoms meriting operation later and do not develop syringomyelia.

The general consensus is that a completely asymptomatic patient with hindbrain hernia does not need surgery. The controversy focuses on what constitutes asymptomatic hindbrain hernia. Most patients, on close questioning, present clear evidence of cough-induced headache or occipital pain. Some even admit to arm pain on exertion. In small children, mild motor developmental delay is frequently seen. Other, more subtle symptoms and signs can go unnoticed. There is

Figure 4 Sagittal T1w magnetic resonance imaging scan of a 4-year-old girl with newly presented limp and a hairy patch at the bottom of her spine. There is a diastematomyelia peg at the level of L3 vertebral body tethering the spinal cord. There is a syringomyelia cavity at T11-L1, which did not alter significantly after a cord untethering operation. Scale is in centimeter.

a spectrum of symptom intensity, and the surgeon's expertise in the field will influence which patients would benefit from surgery.

Although sudden neurological deterioration has been described, the percentage of patients with syringomyelia who deteriorate and the time frame involved are not known. Subtle symptoms and signs are easy to miss. This problem is more serious in young children, in whom mild scoliosis and subtle motor developmental delay are difficult to evaluate. Considering that long-standing syringomyelia leads to gliosis inside the spinal cord and irreversible damage, it would make sense to have a low threshold for surgical treatment. Often, though, mild or moderate syringomyelia cavities can stay unchanged for months or even years.

Indications for Surgical Treatment

- In patients with Chiari I malformation with or without syringomyelia, the presence of relevant symptoms and clinical findings indicates craniovertebral decompression. A tense syrinx occupying the cervical cord, even if asymptomatic, merits surgical treatment as it potentially threatens vital neurological functions. In contrast, the patient with a small, two-level cavity and no relevant symptoms can be watched with repeat MRI scans annually.
- The presence of scoliosis in children, even in the absence of other neurological symptoms, requires surgical treatment to stabilize the progress of scoliosis. Most orthopedic surgeons would not contemplate surgical correction of scoliosis in the presence of untreated syringomyelia as the risk of postoperative paraplegia or quadriplegia is significant.
- Patients with posttraumatic syringomyelia merit early surgery as syringomyelia can silently and insidiously threaten their residual neurological function. Surgical decompression is performed on the site of the subarachnoid block, which is commonly the fracture site.
- In patients with dysraphic states, syringomyelia rarely constitutes a problem and an indication for surgery and remains unchanged after successful cord untethering.
- Patients with spinal cord tumors usually require surgical excision for their tumor, which leads to reduction of the syringomyelia cavity. Recurrence of syringomyelia commonly implies tumor recurrence.
- Patients with spinal arachnoid cysts commonly have neurological symptoms from cord compression and require surgical excision of the cyst.
- Patients with postinflammatory syringomyelia usually have significant neurological deficits even before syringomyelia develops, and it is not always easy to determine whether the development of syringomyelia is associated with the development of new symptoms. In general, these patients merit early surgery to avoid clinical deterioration. They need careful radiological evaluation to determine the site of the subarachnoid block and tailored surgical strategy.
- Patients with recurrent syringomyelia and unclear neurological signs require careful assessment. Repeat surgery for such patients often has poor results, and caution should be exercised. Often an initial period of observation and twice yearly MRI scans is a good strategy, with surgery for clear clinical or radiological deterioration.

Surgical Management of Hindbrain Hernia–Chiari I Malformation

Craniovertebral decompression for Chiari I malformation includes occipital craniectomy and removal of the arch of C1 at least, depending on the extent of the hindbrain hernia. This aims to remove the block of CSF flow in the craniovertebral junction due to the

prolapsed tonsils. This technique leads to radiological improvement in up to 80% of the patients. Syringopleural or syringosubarachnoid shunts, when used as primary treatment, have a very poor long-term success rate in the presence of an active subarachnoid block, which acts as a filling mechanism for the syringomyelia.

Controversy exists over whether the arachnoid matter should be opened. Most neurosurgeons prefer to open the arachnoid. The proponents of opening the arachnoid maintain that dissection of the arachnoid adhesions on the foramen of Magendie safeguards against recurrence. The proponents of leaving the arachnoid intact believe that breaching the arachnoid can promote creation of further adhesions due to the intraoperative spillage of blood. Recently, intraoperative ultrasound scanning has indicated that absence of CSF flow at the foramen of Magendie after craniectomy may be an indication for opening the arachnoid membrane.

Another controversial issue is the use of grafting to enlarge the cisterna magna. Most neurosurgeons in North America are in favor of grafting because of a lower incidence of CSF leak. Arguing against grafting, however, is the observation that grafts promote adhesion formation in the cisterna magna. This can be worse in young children, who tend to regrow part of the removed foramen magnum.

Surgical Management of Posttraumatic Syringomyelia

In the 1970s and 1980s, syringosubarachnoid and syringopleural shunts were used widely, but in the past decade, their long-term effectiveness has been questioned. Mechanical complications and infection have been observed. The drain functions for some time, but then the cord collapses around it, and the draining holes become occluded by gliotic tissue. Subsequently the syringomyelia recurs if the filling mechanism has not been corrected. The current treatment of posttraumatic syringomyelia is decompressive laminectomy and subarachnoid space reconstruction by adhesiolysis at the fracture site. This technique results in radiological improvement of the syrinx in 80% of patients.

Surgical Management of Residual or Recurrent Syringomyelia

When failure to reduce syrinx size or recurrence has been observed, reexploration of the craniovertebral junction and extensive dissection of the arachnoid adhesions in all the foramina of the fourth ventricle are necessary. Stents can be placed in the foramen of Magendie to avoid recurrent adhesions. The use of endoscopy to break down septations inside the syrinx cavity was employed in the 1990s but has been abandoned. Recurrent posttraumatic or postinflammatory syringomyelia requires repeat exploration and adhesiolysis. Syringopleural or syringosubarachnoid shunts can be reserved for patients who show poor syrinx response despite adequate surgical decompression.

Management of the Associated Hydrocephalus

The connection between hydrocephalus and syringomyelia has not been elucidated fully yet. Traditionally, hydrocephalus is treated first, before a craniovertebral decompression. Cine PC-MRI evidence suggests that hydrocephalus is due to fourth ventricle obstruction. In practice, though, the ventricles commonly do not decrease in size after craniovertebral decompression. On the other hand, when ventricular shunting is performed first, the syrinx cavity reduces in size, but the need for craniovertebral decompression is not avoided. Recently, endoscopic third ventriculostomy has been used to treat hydrocephalus, with good results for syrinx resolution.

In patients with a ventriculoperitoneal shunt and syringomyelia that has been controlled, recurrence of the syringomyelia implies shunt obstruction until proven otherwise. Shunt revision should be performed before any other surgery is contemplated.

Complications

The complication rate for craniovertebral or spinal decompression is low. Perioperative mortality is rare nowadays. Low rates of wound infection, CSF leak, meningitis, subdural hematoma, urinary tract infection, and thromboembolic complications have been observed. Self-limiting sterile meningitis affects many patients after craniovertebral decompression, most often when the arachnoid has been opened. A very small percentage of patients suffer catastrophic complications such as posterior fossa hematoma, transverse sinus thrombosis, or brain stem infarction due to posterior inferior cerebellar artery damage. Cerebellar sag or slump has been reported as a delayed side effect of craniovertebral decompression. It was more common in the 1980s and 1990s and presented with intractable headaches and lower cranial nerve signs. In recent years there has been a tendency for smaller occipital craniectomy, which has reduced the incidence of this complication.

Treatment Outcome

Up to 80% of patients with hindbrain-related syringomyelia experience clinical and radiological improvement after craniovertebral decompression. Most symptoms improve, but an unfortunate minority of patients experience long-standing incapacitating chronic headaches. Established neurological deficits such as sensory loss and weakness improve little. Muscle wasting and severe lower cranial nerve palsies do not improve. Early myelopathy improves variably. Established spasticity improves little, and often other surgical options have to be used (e.g., intrathecal Baclofen).

In paraplegics, several problems may persist despite resolution of syringomyelia. Loss of pain sensitivity leads to trophic changes in skin or joints. Damage to joints without pain sensation may be severe, leading to neuropathic or Charcot joints. In the shoulder, the head of the humerus may rapidly absorb. The elbow joint can develop osteophytes and limitation of movement. Paraplegic patients have increased risk of developing cervical spondylosis, possibly related to increased use of their arms as a result of the paraplegia.

Up to 15% of patients do not experience postoperative radiological resolution of their syringomyelia. Most of them end up having reexploration of the craniovertebral junction or the spinal decompression site and division of residual arachnoid adhesions, with variable result. Patients with recurrent syringomyelia and/or recurrent arachnoiditis in the foramen magnum or the spinal decompression site tend to deteriorate with time. Despite radiological control of the syrinx, a significant number of patients continue to deteriorate clinically, possibly due to progressive gliosis around the syrinx walls.

Intracranial Hypertension and Syringomyelia

Intracranial hypertension in the absence of hydrocephalus may be more common among these patients than previously realized. There have been reports of incidence of papilledema up to 13% in patients with hindbrain hernia, and some patients have intracranial hypertension without papilledema. It is possible that those who do not demonstrate resolution of syrinx following successful surgery may harbor intracranial hypertension.

Future Development

Advances in imaging and noninvasive methods of measuring increased intracranial pressure may help the neurosurgeon better identify which patients need surgical treatment. Patients with syringomyelia benefit from support systems and help groups.

See also: Evoked Potentials: Recording Methods; Perfusion MRI; Spinal Cord Injuries; Tumors of the Brain and the Spinal Cord.

Further Reading

Batzdorf U (2005) Primary spinal syringomyelia. Invited submission from the joint section meeting on disorders of the spine and peripheral nerves. *Journal of Neurosurgery Spine* 3(6): 429–435.

Batzdorf U, Klekamp J, and Johnson JP (1998) A critical appraisal of syrinx cavity shunting procedures. *Journal of Neurosurgery* 89: 382–388.

Gardner JW (1965) Hydrodynamic mechanism of syringomyelia: Its relationship to myelocele. *Journal of Neurology Neurosurgery and Psychiatry* 28: 247–259.

Heiss JD, Patronas N, DeVroom HL, et al. (1999) Elucidating the pathophysiology of syringomyelia. *Journal of Neurosurgery* 91: 553–562.

Klekamp J and Samii M (2002) *Syringomyelia: Diagnosis and treatment*. Berlin: Springer.

Menezes AH, Greenlee JD, and Donovan KA (2005) Honoured guest presentation: Lifetime experiences and where we are going: Chiari I with syringohydromyelia: Controversies and development of decision trees. *Clinical Neurosurgery* 52: 297–305.

Milhorat TH, Chou MW, Trinidad EM, et al. (1999) Chiari I malformation redefined: Clinical and radiographic findings for 364 symptomatic patients. *Neurosurgery* 44: 1005–1017.

Milhorat TH, Johnson WD, Miller JI, Bergland RM, and Hollenberg-Sher J (1992) Surgical treatment of syringomyelia based on magnetic resonance imaging criteria. *Neurosurgery* 31: 231–245.

Oldfield EH, Muraszko K, Shawker TH, and Patronas NJ (1994) Pathophysiology of syringomyelia associated with Chiari I malformation of the cerebellar tonsils: Implications for diagnosis and treatment. *Journal of Neurosurgery* 80: 3–15.

Sgouros S, Kountouri M, and Natarajan K (2006) Posterior fossa volume in children with Chiari malformation type I. *Journal of Neurosurgery: Pediatrics* 105(supplement 2): 101–106.

Sgouros S and Williams B (1995) A critical appraisal of drainage in syringomyelia. *Journal of Neurosurgery* 82: 1–10.

Sgouros S and Williams B (1996) Management and outcome of post-traumatic syringomyelia. *Journal of Neurosurgery* 85: 197–205.

Tubbs RS, McGirt MJ, and Oakes WJ (2003) Surgical experience in 130 pediatric patients with Chiari I malformations. *Journal of Neurosurgery* 99: 291–296.

Williams B (1981) Simultaneous cerebral and spinal fluid pressure recordings: Cerebrospinal dissociation with lesions at the foramen magnum. *Acta Neurochirurgica* 59: 123–142.

Williams B (1990) Syringomyelia. *Neurosurgery Clinics of North America* 1: 653–685.

Relevant Websites

http://www.apaiser.asso.fr – APAISER, Association Pour Aider et Informer les Syringomyéliques Européens Réunis, an association of patients with syringomyelia in France.

http://www.conquerchiari.org – Conquer Chiari, also known as C&S Patient Education Foundation, an association of patients with Chiari malformation and syringomyelia in the USA.

http://www.srhsb.org – Society for Research into Hydrocephalus and Spina Bifida, Clinicians international research group meeting once a year, encompassing syringomyelia within its interest areas.

http://www.theannconroytrust.org.uk – The Ann Conroy Trust, an association of patients with Chiari malformation and syringomyelia in the United Kingdom.

http://www.wacma.com – World Arnold Chiari Malformation Association, an association of patients with Chiari Malformation in New York.

Tactile Coding in Peripheral Neural Populations

A W Goodwin and H E Wheat, University of Melbourne, Parkville, VIC, Australia

© 2009 Elsevier Ltd. All rights reserved.

The sense of touch is used by humans to provide a vast array of information about objects in the environment. There are many features, such as the compliance or softness of an object, that can only be determined by using touch. Other features, such as the texture of a surface, can be determined visually, but resolution is greater with touch, which is thus the preferred mode.

Invariably tactile perception requires hand movements. In some cases, the motor control component is relatively simple, principally providing contact with the object or providing relative motion between the skin and the surface; an example is determining the texture of a surface by rubbing the fingers over the surface. Such touch is often termed active because the relative motion is generated by the observer, but tactile sensation is also possible via passive touch, in which the relative motion is generated by a machine or a second person. In contrast, many operations are true sensorimotor tasks in which tactile sensation and hand movements are interdependent; two simple examples are operating a computer mouse or inserting a catheter into a vein.

During complex manipulations, tactile input may give rise to conscious sensation and perception or it may be used by the motor control system without conscious sensation. In general, the sensory signals lead to both perception and motor control.

The Sensors

The palmar (front) surface of the hand is covered by specialized skin called glabrous skin (Figure 1). The more superficial epidermis and the deeper dermis are arranged in an interlocking pattern of grooves and ridges. This undulating pattern is visible in the thick layer of keratin that covers the epidermis, particularly on the fingertips, where it forms the characteristic fingerprint patterns which vary from individual to individual.

There are four classes of mechanoreceptors in glabrous skin; these are the sensors that mediate touch. Two of the receptor types, Meissner corpuscles and Merkel complexes, are located relatively close to the surface of the skin. Meissner corpuscles are located in the most superficial part of the dermis, in the crests of the waves called the dermal papillae. Merkel complexes are located at the epidermal–dermal junction at the base of the downwardly protruding limiting ridges of the epidermis. The remaining two receptor types are located further from the skin surface. Ruffini organs are located within the dermis and Pacinian corpuscles are located both deep within the dermis and also below the dermis in subcutaneous tissues. This difference in the depth of the receptors is reflected in a difference in their spatial properties.

Based on their structure, the mechanoreceptors can be divided into two groups. Each Meissner corpuscle and Pacinian corpuscle is surrounded by a capsule, and within the capsule the structure is laminated with layers of nerve terminals or connective tissue separated by layers of fluid. In contrast, the Ruffini organs and Merkel complexes do not have a laminated structure. This structural difference between the two groups results in a difference in their temporal characteristics.

The mechanoreceptors are the terminations of sensory neurons with somas or cell bodies located in the dorsal root ganglia. The axons which extend from the receptors to the cell bodies form the sensory component of the peripheral nerves. These fibers, sometimes referred to as primary afferent fibers, are large myelinated fibers ($\sim$5–12 μm in diameter) which conduct action potentials rapidly ($\sim$30–70 m s^{-1}). A single fiber may end in a single mechanoreceptor, but more commonly it will innervate many mechanoreceptors. Conversely, a single mechanoreceptor may be innervated by a single fiber or by more than one fiber.

Single-Fiber Responses to Simple Stimuli

Recording responses in single peripheral nerve fibers, in both monkeys and humans, is relatively straightforward. In anesthetized monkeys, the peripheral nerve is exposed surgically and microdissected, or successively split, until a single functional axon can be draped over an electrode and its action potentials recorded. In awake humans, a sharpened microelectrode is inserted through the skin into the peripheral nerve and positioned so that action potentials can be recorded from a single axon.

In early studies of cutaneous mechanoreceptors, the stimulus consisted of a handheld probe indented into the skin surface. Sometimes the probe was rigid, such as a glass rod, but more commonly it was a monofilament (von Frey hair), which provided some rudimentary control over contact force.

Limited though these stimuli were, they did enable classification of the primary afferent fibers into four unambiguous groups. Two classes respond during the dynamic phase, when the probe is indented or

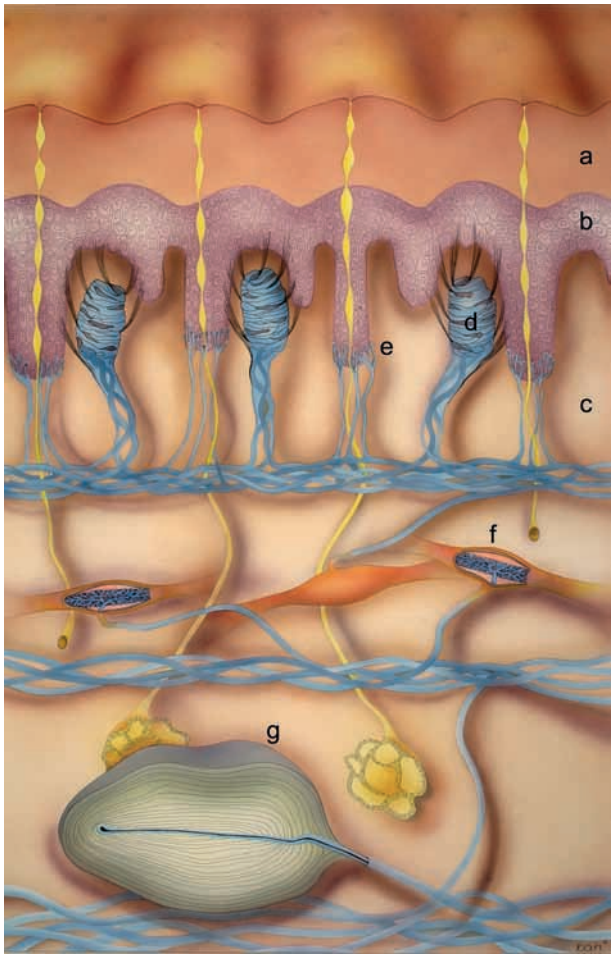

Figure 1 A cross section through glabrous skin. The upper layer, or epidermis (b), is covered in thick keratin (a). Both the epidermis and the underlying layer, the dermis (c), have a wave-like structure. Peripheral nerve fibers, shown in blue, terminate on four types of mechanoreceptors: the Meissner corpuscles (d), the Merkel endings (e), the Ruffini organs (f), and the Pacinian corpuscles (g). Reproduced from Darian-Smith I (1984) The sense of touch: Performance and peripheral neural processes. In: Brookhart JM, Mountcastle VB, Darian-Smith I, and Geiger SR (eds.) *Handbook of Physiology, Vol. 3: The Nervous System,* pp. 739–788. Bethesda, MD: American Physiological Society, used with permission from American Physiological Society.

withdrawn from the skin, and during the static phase, when the probe is in contact with the skin without change in indentation. These are termed slowly adapting (SA) afferents. One class, with small well-defined receptive fields, is termed SAI ,and the second class, with larger less well-defined receptive fields, is termed SAII. The remaining two classes of afferents respond only during the dynamic phase of stimulation and are silent during the static phase. These are termed fast-adapting (FA) afferents; FAI afferents have small receptive fields, and FAII afferents large receptive fields. There is clear evidence that SAI afferents terminate in Merkel complexes and strong evidence that FAII afferents terminate in Pacinian corpuscles. There is reasonable evidence that FAI afferents terminate in Meissner corpuscles and, by exclusion, it is generally accepted that SAII afferents terminate in Ruffini organs. The more superficial receptors (Meissner corpuscles and Merkel complexes) give rise to small well-defined receptive fields (type I afferents), and the deeper receptors (Pacinian corpuscles and Ruffini organs) are associated with larger less well-delineated receptive fields (type II afferents). The receptors with a laminated structure (Meissner and Pacinian corpuscles) result in a transient response in their innervating axons.

The introduction of electronically controlled stimulators enabled the temporal profile of the probe indentation to be varied accurately and also allowed for the indentation amplitude or the contact force to be controlled. Using such stimulators, it was shown that FA afferents responded to the velocity as well as the depth of indentation and therefore responded well to vibrating probes. For any stimulus parameter, individual afferents differ in their sensitivity. For example, individual SAI afferents show a range of indentation thresholds and individual FAI afferents show a range of velocity thresholds.

A key factor that limits tactile resolution is the density of the mechanoreceptors and, more critically, the innervation density of the associated primary afferent fibers that convey the tactile information to the central nervous system. Density varies with the afferent type and with the location on the hand. Density of the FAI and SAI afferents increases distally, reaching a maximum on the fingertips of approximately 1.4 and 0.7 afferents mm^{-2}, respectively. Density of the FAII and SAII afferents is significantly lower, approximately 0.14 afferents mm^{-2}, and is fairly uniform over the hand.

Coding of Simple Stimuli

Accurately controlled indenting probes allowed afferent responses to be characterized quantitatively and allowed human perception to be quantified in matching psychophysics experiments. By comparing the neural and psychophysical responses, conclusions were drawn about how the characteristics of the simple probe stimuli were coded in the neural responses. As an illustration, at frequencies around 10 Hz the human threshold for detecting vibration on glabrous skin matches FAI afferent thresholds and is well below FAII afferent thresholds. Conversely, at 300 Hz the human threshold matches the FAII afferent thresholds and is below the FAI afferent thresholds. Thus, it can be concluded that high-frequency tactile vibration is coded by FAII afferents and low-frequency vibration (sometimes called flutter) is coded by FAI afferents.

Although they are seminal and extremely useful, it is important to recognize that studies such as this only scratch the surface of the highly complex issue of coding. The next step requires measuring the statistical distribution of the afferent properties (in this case, threshold), some model of how the responses of the many responding afferents are combined, an allowance for noise or neural variability, and finally a precise comparison of the combined neural responses with the human performance. An obvious and perhaps more important limitation of these early studies is the nature of such stimuli. They cannot readily be related to the tactile stimuli we encounter daily and rely on for successful perception and manipulation of objects.

The Nature of Tactile Stimuli

When we handle objects in everyday life, there are a number of properties of those objects that we are aware of and that we can make judgments about should we wish to. Three of the more obvious are: surface texture (which encompasses roughness), size, and shape. This is true for immobile objects that we explore with active touch, such as a large statue, and also for mobile objects that we manipulate, such as a piece of fruit. When we manipulate a mobile object, several additional parameters affect the cutaneous mechanoreceptor responses; these include the grip forces (normal to the skin) that we employ, the load forces (tangential to the skin) imposed by gravity, and the distribution of the contact points on the hand. All commonly encountered tactile stimuli have both temporal and spatial components.

These properties of the manipulated object are far more complex than the simple indenting probes that were used initially to study the mechanoreceptors. In order to gain a more realistic appreciation of the nature of mechanoreceptor responses, it is therefore necessary to employ complex stimuli that can be directly related to real, everyday tactile events.

How is it possible for the cutaneous mechanoreceptors with apparently simple transduction properties to convey information about such complex stimulus properties, and how do they convey, simultaneously, multiple properties without ambiguity? As discussed next, such rich information is contained in the responses of whole populations of afferents.

What Is a Population Response?

If all the active neurons had identical responses to the stimulus and there was no variability in the neural responses, then a single neuron would convey as much information as a group of neurons. But this is never the case. There is always some degree of neural noise or variability, and the statistical reliability of the combined (summed or averaged) responses of a group of neurons is greater than the reliability of the individual neural responses. This improvement in signal-to-noise ratio is the simplest way in which information transmission by a single neuron is enhanced by a population of multiple neurons.

In most cases, the power of the population response is much greater than a simple enhancement of statistical reliability. Rather, the spatial and temporal patterns within the population contain a wealth of information about the stimulus and can simultaneously encode multiple stimulus parameters.

To measure a population response experimentally, it would be necessary to record the single-unit activity of all the responding neurons simultaneously. This is not possible with existing technology. Therefore, the only way to study population responses currently is to model them in some way by reconstructing the population response from individual single-unit recordings.

A Simple Population Model

In the tactile system, the earliest examples of population models stemmed from mapping the spatial characteristics of receptive fields of primary afferent fibers with an indenting probe. A typical receptive field, for an SAI afferent, is illustrated in Figure 2. The neuron

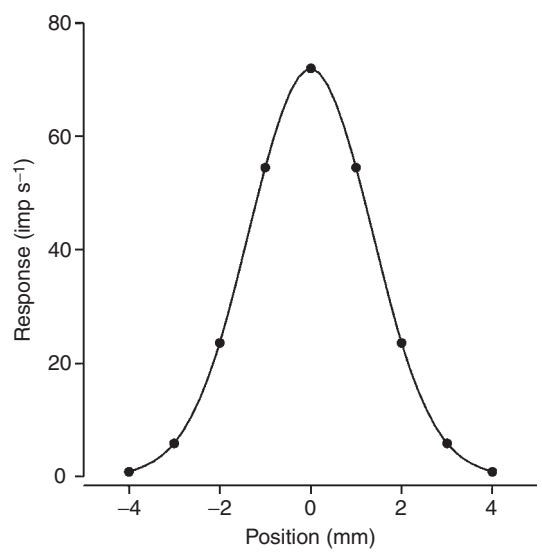

Figure 2 The receptive field profile of a single SAI afferent mapped using an indenting probe with a spherical tip of radius 1.44 mm (curvature 694 m^{-1}). The probe was positioned successively at 1 mm intervals on a line through the receptive field center. Alternatively, the profile may be viewed as the response of a population of identical SAI afferents with their receptive fields located along a line on the skin. imp, impulses; SAI, slowly adapting type I.

is most responsive to indentation at the center of the receptive field, and the response reduces successively as the probe is applied at points on the skin successively further from the center. This is the receptive field profile of a single SAI afferent, measured at 1 mm intervals, but it can also be interpreted as a population profile as follows. Consider an array of identical SAI fibers, with receptive field centers separated by 1 mm, on a region of skin in contact with an indenting probe. The afferent with a receptive field center directly under the probe responds most vigorously, and afferents with receptive field centers successively further from the probe respond successively less vigorously.

Texture

Sometimes we wish to determine the texture of an object's surface, and we do so by scanning our fingertips over the surface. We can make fine discriminations, for example in the roughness of a piece of wood we are sanding. Texture is also an important parameter for the motor control system in object manipulation because of the relationship between texture and the effective coefficient of friction between a surface and the skin.

For experimental purposes, surfaces with controlled and easily quantifiable textures are readily achieved by gratings of alternating grooves and ridges or by embossed patterns of raised dots, both of which are easy to manufacture. Commonly, the peripheral neural population response for such textures is approximated by reciprocally interpreting a single afferent's receptive field properties. The approach is analogous to that illustrated in **Figure 2,** except that, instead of stepping the stimulus across the receptive field, it is scanned over the receptive field at constant velocity. The embossed dots in **Figure 3(a)** are scanned over the afferent's receptive field along the surface's long axis (i.e., in the direction of the arrow), with small shifts between scans at right angles to this axis. In this way, the entire pattern is scanned over the receptive field center. Corresponding two-dimensional plots of evoked action potentials (**Figure 3(b)**) can be interpreted as a simplified population response. At a qualitative level, such plots show the extent to which the population of that afferent class represents the stimulus. In addition, some quantitative analysis can be used to compare these neural responses to psychophysical measures of human perception when the same surfaces are scanned over the subjects' fingertips.

Notwithstanding the power of this approach, it is important to recognize the substantial difference between these response plots and the true population response. First, these plots imply infinite innervation density but, in reality, they are sampled by the afferents with a finite density of some 1.4 or 0.7 mm^{-2} for the FAI and SAI populations respectively. Second, the inherent assumption in such representations is that all the afferents have identical properties. In particularly, no account is taken of the varying sensitivities of afferents, even though it has been shown that SAI sensitivities vary by a factor of around 5. Third,

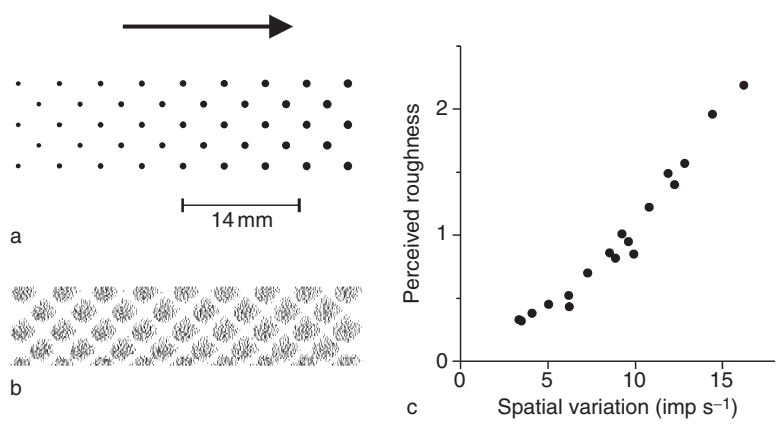

Figure 3 Responses to a textured surface: (a) surface of embossed dots of graded size; (b) two-dimensional plot of evoked action potentials; (c) plot of degrees of spatial variation in afferent impulse rate. In (a), the surface of embossed dots is scanned across an afferent's receptive field in the direction of the arrow; between scans the stimulus is shifted a small amount in a direction at right angles to the scanning direction. In (b), each dot in the two-dimensional plot marks an impulse evoked in an SAI afferent by the surface. In (c), parts of the surface with different dot sizes produce different degrees of spatial variation in the SAI afferents' impulse rates; this correlates with the human perception of the roughness of those parts of the surface. imp, impulses; SAI, slowly adapting type I. Adapted from Blake DT, Hsiao SS, and Johnson KO (1997) Neural coding mechanisms in tactile pattern recognition: The relative contributions of slowly and rapidly adapting mechanoreceptors to perceived roughness. *Journal of Neuroscience* 17: 7480–7489.

response variability or neural noise is not accounted for, but that will have an obvious impact on resolution in the population.

Shape

When we touch an object, the cutaneous mechanoreceptors relay precise information about the local shape of the object. When we touch the mouse and keyboard during typing, it is very obvious that the key surfaces are concave and the mouse surfaces convex. It is also obvious that the central wheel of the mouse is more convex than the buttons. In human psychophysics experiments, the resolution of the cutaneous mechanoreceptors for such shapes can be measured by using a machine to apply variously shaped stimuli to an immobilized fingertip so that only those receptors are activated. In such experiments humans can discriminate two spherical surfaces that differ in curvature by approximately 10%. We may use such shape information consciously, but it is also important for the motor control system because the characteristics and consequences of load and grip forces vary with the shape of a grasped object. Also, for manipulations in which the fingers move over the object, the trajectories of finger movements depend on the object's shape.

To reconstruct a simplified SAI afferent population response to a sphere, the sphere is indented into the skin at contact points over the two-dimensional receptive field (Figure 4(a)); this is the two-dimensional equivalent of Figure 2. Comparing Figures 4(a) and 4(b), we see how a change in the curvature (shape) of the sphere results in a change in the shape of the population response. In contrast, comparing Figures 4(a) and 4(d), we see how changing a different parameter of the stimulus, namely the contact force between the spherical object and the finger, does not affect the curvature of the population response but results in a corresponding scaling of the response. A third independent parameter of the stimulus is the position of contact of the sphere on the skin. When this parameter is changed (compare Figures 4(a) and 4(c)), neither the curvature nor the scale of the population response is affected but the entire response is shifted, within the population, by a corresponding amount.

In Figure 4 three different parameters have been varied: the curvature (shape) of the sphere, the force between the sphere and the skin, and the position of contact on the skin. From these idealized population

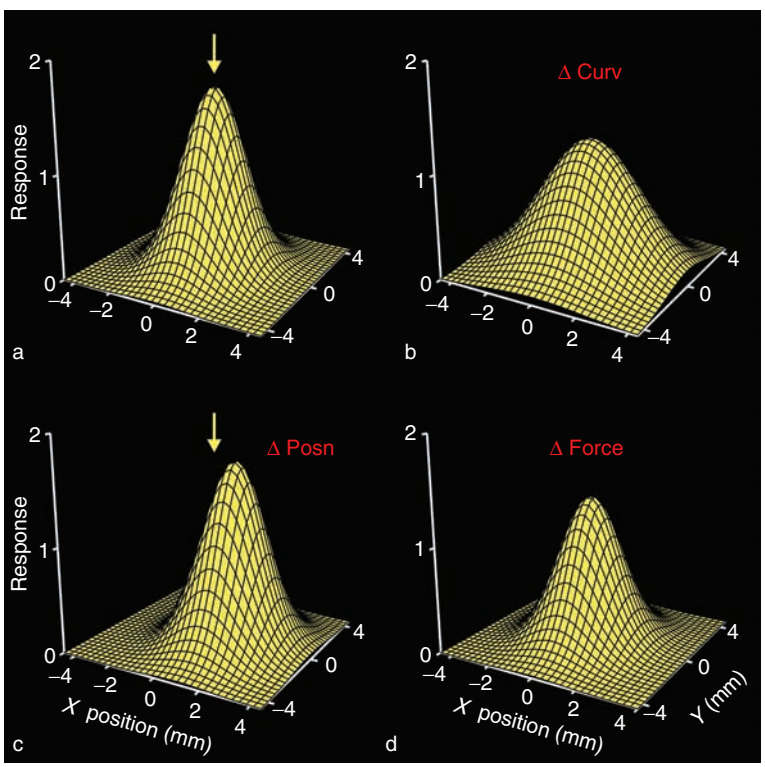

Figure 4 Representation of multiple stimulus parameters in the SAI afferent population response: (a) sphere has a curvature of 694 m^{-1} (radius 1.44 mm); (b) stimulus shape changes to a curvature of 256 m^{-1} (radius 3.9 mm); (c) shifting the stimulus by 1 mm results in the population response shifting by 1 mm (compare to arrow and arrow in (a)); (d) a reduction in contact force scales the entire response correspondingly. ΔCurv, change in curvature (shape); ΔForce, change in contact force; ΔPosn, change in position of contact; SAI, slowly-adapting type I. Adapted from Goodwin AW and Wheat HE (1999) Effects of nonuniform fiber sensitivity, innervation geometry, and noise on information relayed by a population of slowly adapting type I primary afferents from the fingerpad. *Journal of Neuroscience* 19: 8057–8070.

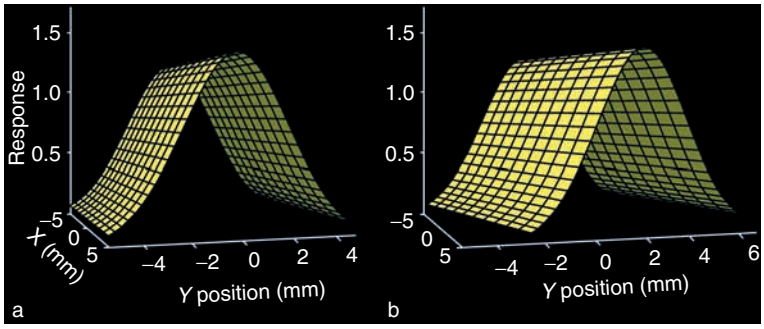

Figure 5 Idealized slowly adapting type I (SAI) afferent population response to a cylinder contacting the skin (radius 5.8 mm, curvature 172 m^{-1}) for two orientations of the cylinder: (a) 0°; (b) 15°. Adapted from Dodson MJ, Goodwin AW, Browning AS, and Gehring HM (1998) Peripheral neural mechanisms determining the orientation of cylindersgrasped by the digits. *Journal of Neuroscience* 18: 521–530.

responses, it is clearly evident how a single population can provide unambiguous information about the three stimulus parameters. The brain can extract the shape of the sphere from the shape of the population response, the contact force from the overall magnitude of the response, and the position on the skin from the location of the response. All three parameters can be extracted independently.

In a similar way, Figure 5 shows how a cylinder in contact with the skin will result in a cylindrically shaped population response. A change in the orientation of the cylinder is evident from a corresponding rotation of the population response. As was the case for the sphere in Figure 4, changing the curvature, position, or contact force of the cylinder will uniquely change the curvature, position, or scaling of the population response. All four parameters of the cylindrical stimulus are clearly and independently signaled to the brain.

Population Codes

So far we have drawn only qualitative conclusions from the population responses. This has been possible because of their isomorphic nature; for example, the impulse pattern in Figure 3(b) resembles the pattern of dots in the stimulus. In order to quantify the population responses and compare them with human performance, it is necessary to extract some response measure relating to each stimulus parameter. Such a measure may be thought of as a population code for that stimulus parameter. How does the pattern of impulses in Figure 3(b) encode the roughness of the textured surface evoking it? Experimental data show that the smoother the surface, the more uniformly the impulses are distributed across the two spatial dimensions of the response. Thus, one potential code for surface roughness is the spatial variation in impulse density. As we see in Figure 3(c), this code corresponds well with the measured human perception of the surface roughness. In a similar way, an inspection of Figures 4(a) and 4(c) suggests that a potential code for the position of the stimulus on the skin is the centroid of the population response. The centroid is independent of the curvature of the sphere and independent of the contact force and adequately explains the human capacity to discriminate differences in position of less than 0.5 mm.

More Realistic Population Responses

Although useful, the population responses shown so far may overestimate the precision of stimulus representation because of their idealistic nature. A neural code extracted from such a population may match human performance and therefore be deemed a viable code, whereas in a more realistic population that same code may not have sufficient resolution to explain human performance and must therefore be rejected as a potential code.

Building on the underlying profiles shown in Figure 4, mathematical and statistical modeling, often in the form of computer simulations, can be used to generate credible population responses and potential codes for the stimulus parameters. A powerful feature of such modeling is that it enables a systematic analysis of the effect of the inherent properties of the population.

One problem with idealized population profiles is the assumption of equal sensitivity or responsiveness for all afferents within a class, which is not the case. In computer models, it is simple to randomly assign a sensitivity to each afferent using estimates derived from experimental data of the sensitivity distribution for a specific afferent class. One set of random sensitivities for afferents with receptive fields along a line through the center of the stimulus is shown in Figure 6(a). The distorting effect this has on the ideal response profile is shown in Figure 6(b). The two-dimensional contour plot in Figure 6(c) shows the

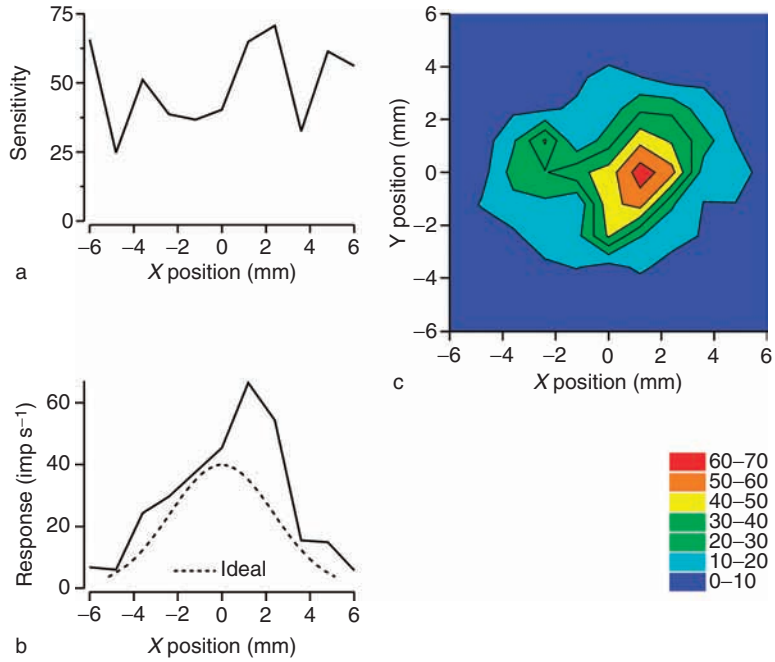

Figure 6 Variation in sensitivity distorts the SAI afferent population response to a sphere contacting the skin: (a) sensitivities along the line $y=0$ were randomly allocated according to the experimentally measured distribution; (b) plot of the resultant distortion (solid line) of the ideal response profile (dashed line); (c) contour plot showing the distorted population response (in impulses s^{-1}) in two dimensions. Adapted from Goodwin AW and Wheat HE (1999) Effects of nonuniform fiber sensitivity, innervation geometry, and noise on information relayed by a population of slowly adapting type I primary afferents from the fingerpad. *Journal of Neuroscience* 19: 8057–8070.

distortion of the ideal profiles of Figure 4. When we view the ideal population profile, it might seem reasonable to propose the position of the peak response as a potential code for the position of the stimulus on the skin. However, when variation in afferent sensitivity is included, the position of the peak shifts and is an unreliable predictor of human performance. The centroid is a more robust measure which averages out the individual afferent sensitivities. Modeling also allows more speculative schemes to be considered. For example, could the brain compensate for varying afferent sensitivity using some prior knowledge or estimate of each afferent's sensitivity? Such knowledge could be accurate or could have varying amounts of uncertainty.

Neural noise, or variability in responses to the same stimulus, is an intrinsic property of all neurons. In reality, the response of each afferent in Figure 4 has a random, or noise, component added, which will distort the profile. If the same stimulus is presented many times, each time the noise component on each neuron will be different and the resultant profile will be different to an extent dependent on the magnitude of the noise. Little is known about the magnitude of such noise in the multitude of neural populations in the nervous system. An additional complication, which is difficult to measure experimentally, is the degree of interdependence of noise or covariance among the members of a population. These are important considerations because different neural codes will be affected differently by different types of noise and different degrees of covariance. An interesting outcome of simulations which incorporate noise is that neural codes akin to population averages, such as the total volume under the curve in Figure 4, are degraded by covariance of noise among the constituent neurons, whereas codes of a spatial nature, such as the centroid, benefit from covariance.

An additional factor that limits resolution in the population response is the finite innervation density of the afferents. Profiles such as those in Figure 6 are not continuous but rather are sampled by the afferents. The best estimates of innervation density for the fingertip are 1.4 and 0.7 afferents mm^{-2} for the FAI and SAI afferents, which are equivalent to a separation between receptive field centers of 0.85 and 1.2 mm, respectively, assuming a uniform pattern of innervation. However, there is some uncertainty about these figures; the estimates are based on a single study in humans and a single study in monkeys, and they result from indirect calculations necessitating a number of assumptions. Moreover, the pattern of innervation is unknown. The Meissner corpuscles and Merkel complexes conform to the pattern of fingerprints (Figure 1), which vary over the fingertip and vary between individuals, but the pattern of convergence between

receptors and innervating afferents is unknown. Models can sample the simulated response profiles at any desired density and in any desired pattern to explore the effect on the stimulus representation and on the resolution of any putative codes. This is particularly pertinent in common neurological diseases, such as diabetic neuropathy, in which some of the afferents cease to function, perhaps in a patchy pattern.

An alternative approach to the computational analysis of neural populations is to simulate a neural network of some sort. One advantage of neural networks is that they can be trained, with known inputs and outputs, without knowledge of or assumptions about the details of the intervening neural processes. This, of course, is also a disadvantage. A more serious criticism of neural networks is that the constituent neurons do not behave like real neurons and the simulation may therefore be of limited value for understanding the human nervous system.

See also: Human Haptics; Mechanoreceptors; Neural Circuitry in the Somatosensory System; Reaching and Grasping; Sensorimotor Integration: Models; Sensorimotor Control of Manipulation; Tactile Texture; Temperature Sensation.

Further Reading

Blake DT, Hsiao SS, and Johnson KO (1997) Neural coding mechanisms in tactile pattern recognition: The relative contributions of slowly and rapidly adapting mechanoreceptors to perceived roughness. *Journal of Neuroscience* 17: 7480–7489.

Craig JC and Rollman GB (1999) Somesthesis. *Annual Review of Psychology* 50: 305–331.

Darian-Smith I (1984) The sense of touch: Performance and peripheral neural processes. In: Brookhart JM, Mountcastle VB, Darian-Smith I, and Geiger SR (eds.) *Handbook of Physiology, Vol. 3: The Nervous System*, pp. 739–788. Bethesda, MD: American Physiological Society.

Dodson MJ, Goodwin AW, Browning AS, and Gehring HM (1998) Peripheral neural mechanisms determining the orientation of cylinders grasped by the digits. *Journal of Neuroscience* 18: 521–530.

Georgopoulos AP, Schwartz AB, and Kettner RE (1986) Neuronal population coding of movement direction. *Science* 233: 1416–1419.

Goodwin AW and Wheat HE (1999) Effects of nonuniform fiber sensitivity, innervation geometry, and noise on information relayed by a population of slowly adapting type I primary afferents from the fingerpad. *Journal of Neuroscience* 19: 8057–8070.

Goodwin AW and Wheat HE (2004) Sensory signals in neural populations underlying tactile perception and manipulation. *Annual Review of Neuroscience* 27: 53–77.

Johansson RS and Vallbo AB (1983) Tactile sensory coding in the glabrous skin of the human hand. *Trends in Neuroscience* 6: 27–32.

Johnson KO and Hsiao SS (1992) Neural mechanisms of tactual form and texture perception. *Annual Review of Neuroscience* 15: 227–250.

Johnson KO, Hsiao SS, and Twombly AI (1995) Neural mechanisms of tactile form recognition. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 253–267. Cambridge, MA: MIT Press.

Khalsa PS, Friedman RM, Srinivasan MA, and LaMotte RH (1998) Encoding of shape and orientation of objects indented into the monkey fingerpad by populations of slowly and rapidly adapting mechanoreceptors. *Journal of Neurophysiology* 79: 3238–3251.

Klatzky RL and Lederman SJ (2003) Touch. In: Weiner IB (ed.) *Handbook of Psychology*, pp. 147–176. New York: Wiley.

Romo R and Salinas E (2003) Flutter discrimination: Neural codes, perception, memory and decision making. *Nature Reviews Neuroscience* 4: 203–218.

Vallbo AB (1995) Single-afferent neurons and somatic sensation in humans. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences*, pp. 237–252. Cambridge, MA: MIT Press.

Tactile Texture

C E Chapman and A M Smith, Université de Montréal, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Tactile surface texture varies along a continuum from smooth to rough. Surface asperities are absent for smooth surfaces such as polished glass or fabrics like satin or silk; rougher surfaces (abrasive papers or burlap), in contrast, are characterized by the presence of surface asperities. The term 'rough' is defined in the Oxford English Dictionary as a "surface diversified with small projections, points, bristles etc; not even or smooth." Roughness is thus a function of multiple tactile elements which may vary in their size, shape (pointed vs. round), and distribution (periodic vs. nonperiodic; distance between elements) on the surface. The material properties of the surface (hardness/softness or stiffness/elasticity) can also contribute to the overall impression of texture: some surfaces are nondeformable (abrasive papers); others are deformable (human skin, textiles). Yet, deformability is not a key element for roughness perception since subjects can scale the roughness of a continuum of abrasive papers or fabrics.

The importance of the physical characteristics of the surface to the perception of roughness has been investigated experimentally using surfaces with randomly disposed tactile elements (abrasive papers) and also simple periodic patterns (gratings formed by alternating ridges and grooves, raised dots similar to Braille dots). The latter approach has an advantage in that the size and spacing, and so the density, of the tactile elements can be varied independently. In contrast, particle diameter, spacing, and density all change with grit value across a range of abrasive papers ('grit' refers to the number of openings per inch (2.54 cm) in the sieves used to produce abrasive papers).

Tactile Roughness Psychophysics

In everyday life, impressions about surface texture are normally obtained by exploring the surface or object with your hand/fingers, using what is termed active touch. The first systematic evaluations of perceived roughness employed abrasive papers. These studies demonstrated that roughness increases with a decrease in the grit value of abrasive surfaces, corresponding to larger particle diameters and lower particle densities. The relationship between perceived roughness and particle diameter has been described as a power function with a slope of approximately 1.5. The Weber fraction (the minimal difference that can be discriminated relative to the magnitude of the standard stimulus) is approximately 0.30 for very fine abrasive papers. This relatively high Weber fraction, which corresponds to an ability to discriminate a 3 μm change in particle density and/or size for a standard surface of 2000 grit (average particle size, 9 μm), may reflect the well-known tendency of the Weber fraction to increase at very low intensities.

As particle density, size, and spacing covary across abrasive papers, however, such studies cannot determine the relative importance of each factor to the perception of roughness. For this reason, subsequent studies have used simple manufactured patterns (grooved metal plates, etched arrays of gratings, or raised dots) to systematically determine the contribution of various physical parameters to roughness appreciation.

Using grooved metal plates, in which case a single dimension can be varied independently, it has been shown that roughness increases as a function of groove width (ridge width constant). In contrast, perceived roughness declines modestly as ridge width is increased (groove width constant). Early experiments were restricted to a small range of surfaces (groove or ridge widths of 0.125–1 mm). In these early studies, spatial period (SP, distance center-to-center between adjacent ridges) *per se* was not the determining factor for roughness because the range of SPs in each series (varying groove or ridge width) was identical (0.375–1.25 mm). These results have since been extended to larger spacings, up to a SP of 3.1 mm, and also finer spacings (0.1 mm SP). The Weber fraction for discriminating gratings is ~0.05, corresponding to a 50 μm threshold for a standard SP of 1 mm. As with fine abrasive papers, the Weber fraction is higher for smaller SPs (0.30–0.50 for a 40 μm SP standard), corresponding to finely textured surfaces.

Further studies to determine the contribution of the physical structure of the surface to perceived roughness have been undertaken using patterns of raised dots, the spacing of which can be varied in one dimension, as with mechanical gratings, or in two dimensions. When spacing is increased in a single dimension, along the direction of scanning, then subjective tactile roughness shows a linear relationship with SP up to 8.5 mm (Figure 1(a)). This relationship is similar to that seen with mechanical gratings, extending the results to a wider range of SPs. When, in contrast, dot spacing is changed in two dimensions (tetragonal patterns), then an inverted U-shaped relation between perceived roughness and dot spacing is obtained (Figure 1(b)): roughness increases with

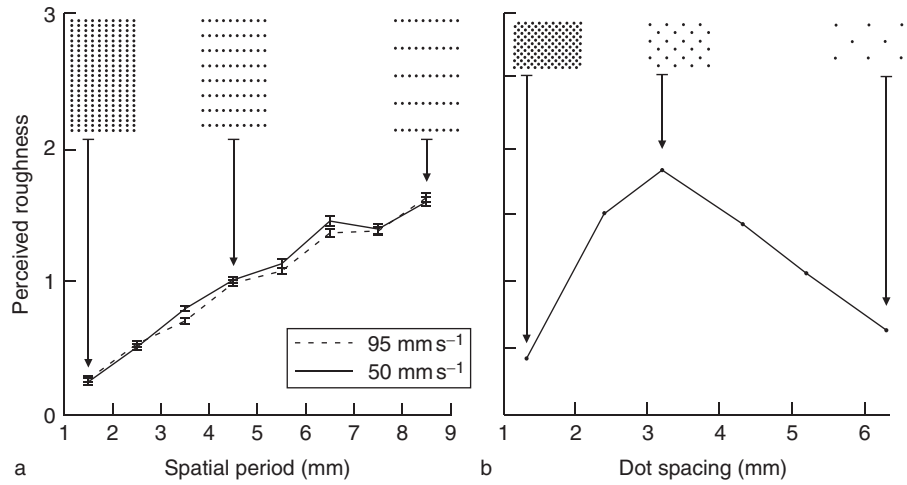

Figure 1 (a) Mean normalized roughness estimates (± SEM) as a function of longitudinal SP for raised dot surfaces displaced under the finger tip (passive touch). SP was increased in one dimension, corresponding to the direction of the scan (see representative surfaces). A monotonic relation between roughness and SP was found, and this was independent of the scanning speed (trials intermixed during acquisition). (b) Mean normalized roughness estimates ($n = 21$) plotted as a function of the dot spacing of tetragonal arrays of raised dots (active touch). Dot spacing was increased in two dimensions (see representative surfaces). Roughness showed a U-shaped relation with dot spacing. (a) Adapted from figure 4 in Meftah E-M, Belingard L, and Chapman CE (2000) Relative effects of spatial and temporal characteristics of scanned surfaces on human perception of tactile roughness using passive touch. *Experimental Brain Research* 132: 351–361, and reproduced with kind permission of Springer Science and Business Media. (b) Adapted and reprinted, with permission, from the *Annual Review of Neuroscience*, Volume 15 ©1992 by Annual Reviews.

increased dot spacing up to 3.2 mm, and thereafter declines. The apparent discrepancy between the two sets of results reflects, in large part, the fact that the height of the dots in the tetragonal arrays was very low (0.35 mm vs. 1.8 mm for the one-dimensional arrays in A), and it is known that subjective roughness increases with increased dot height (see below). Thus, there may be a family of curves relating roughness perception to the spacing of regular arrays of raised dots.

Several factors related to the physical characteristics of the raised dot surfaces modify overall roughness. Thus, when dot diameter is increased then perceived roughness declines. The latter observation is consistent with the decrease in roughness found when the ridge width of mechanical gratings is increased. In addition, as mentioned above, roughness increases with the height of the raised dots. This observation likely reflects both decreased contact between the skin and the smooth surface forming the floor between the dots and increased mechanical deformation of the skin by the raised dots as the skin penetrates more deeply into the spaces between adjacent raised dots. Finally, lubricating the surface decreases its subjective roughness.

Importance of the Mode, and Parameters, of Touch

One key consideration for tactile roughness is that discrimination of surface roughness is better when there is tangential movement between the skin and the surface, or what can be termed dynamic touch, as compared to no movement or static touch. For example, roughness discrimination thresholds are approximately doubled when using static touch as compared to dynamic touch. The superiority of dynamic touch over static touch can be explained by several factors. First, dynamic touch recruits rapidly adapting mechanoreceptors (RA and PC or Pacinian afferents) as well as slowly adapting mechanoreceptors (SA types I and II), while static touch only activates SA receptors. Second, the discharge rates for SA afferents are higher for dynamic stimuli than for static stimuli, so that the signal-to-noise ratio is increased. Finally, neurons in primary somatosensory cortex show a bias for dynamic as opposed to static stimuli, so that the population of neurons contributing to the processing of the tactile input is potentially larger for dynamic inputs.

There has been a long-standing debate in the literature as to whether or not tactile perception is equivalent with active touch (movement made by the subject; the stimulus remains stationary) and passive touch (surface moved over the skin by an external agent; the subject remains stationary). Intuitively, it seems obvious that active touch should show an advantage, in that the exploration is controlled and presumably optimized by the subject. Passive touch, in contrast, is restricted by the parameters of stimulation imposed by the external agent. Nevertheless, current evidence indicates that if exploration conditions

are suitably matched, then roughness discrimination and magnitude estimates of surface roughness are similar for active and passive touch.

Systematic changes in roughness perception can, nevertheless, be produced by changing the physical parameters of the movement. For example, the psychophysical curve for perceived roughness of mechanical gratings is shifted up when contact force is increased from very light (<0.3 N) to very heavy (~7 N) contact forces. Intermediate estimates are obtained with a contact force closer to that employed during active tactile exploration (~1.5 N). This observation, along with the fact that roughness is a function of multiple tactile elements and that roughness increases with increased dot height (above), leads to the possibility that the change in friction (ratio of the tangential to normal force) offered by the surface contributes substantially to roughness. Recent results indicate that while subjective roughness decreases when kinetic friction (i.e., the friction observed during the actual scanning movement) is reduced by lubricating the surface, the most important variable appears to be fluctuations in tangential force, as reflected in the root mean square of the first derivative of the tangential force, dF/dt. Tangential force variations likely also contribute to the human ability to detect a single raised dot of 1–3 μm on a smooth background.

Another important parameter of exploration, scanning speed, has little influence on roughness estimates, at least for coarser textures. As shown in Figure 1(a), there is perceptual constancy for roughness over a range of scanning speeds (50 and 95 mm s^{-1} here; 10 to 100 mm s^{-1} in other studies). To situate these speeds, fair to good Braille readers scan text at 60–125 mm s^{-1}, while excellent Braille readers use faster scans of up to 190 mm s^{-1}. When very rapid scanning speeds are employed (~200 mm s^{-1}), roughness estimates show a modest decline, likely reflecting the fact that the stimulus is less effective, as there is less time for mechanical deformation of the skin by the tactile elements.

Neuronal Codes Underlying Tactile Roughness

The neuronal codes underlying tactile roughness have been the subject of a number of investigations, mostly concentrating on characterizing the responses of the cutaneous mechanoreceptive afferents in the monkey, and relating these to the results of human psychophysical experiments. This is a difficult task because roughness is critically dependent on surface structure, and this in a number of ways, so that no single physical parameter can describe surface roughness. Moreover, surface texture is a continuum from the micron to the millimeter level. Indeed, it may be overly optimistic to expect that a single neuronal code can explain the subjective impression of roughness across the entire range.

A number of candidate neuronal codes exist. One of the simplest is mean discharge rate, often referred to as an intensive code. Experimental evidence in favor of such a code has been obtained using fabrics, mechanical gratings, and raised dot surfaces. These data are, necessarily, restricted to a consideration of three types of mechanoreceptive afferents – SAI, RA, and PC – because glabrous skin in the monkey lacks SAII innervation. For fabrics, there is limited information suggesting that RA afferents but not SAI afferents show a linear relation between firing rate and surface texture. For gratings displaced sinusoidally over the skin, the mean cyclic discharge rate of all three types of afferent increases with groove width, as does perceived roughness. SAI afferents appear to be particularly important for the tactile roughness of gratings, an observation recently extended down to very fine gratings. But, only PC afferents, and not SAI and RA afferents, show a decrease in discharge rate with increased ridge width as required to explain the decrease in roughness with increased ridge width. This leaves open the possibility that ridge width may be signaled by another code, specifically the temporal frequency (speed/SP) of discharge of the population of afferents activated by moving gratings. For raised dot arrays with two-dimensional changes in spacing, in contrast, the mean discharge frequencies of SAI, RA, and PC afferents take a U-shaped form with increased dot spacing, similar to the psychophysical curve shown in Figure 1(b). But these neural functions all peak at a smaller spacing value than the peak in perceived roughness, so that the correlations with subjective roughness are relatively poor. This observation has been interpreted as suggesting that an intensive code may not be applicable to these surfaces, and by extension to other textured surfaces. An alternative explanation is, however, that the shift in the peak of the U-shaped function may reflect species differences related to differences in the size of the stimulated digit relative to the dot patterns (humans vs. monkeys).

Another candidate is a modal code whereby roughness reflects some combination of inputs from the various mechanoreceptive afferents activated by textured stimuli, an interesting idea since all major types of afferents found in monkey glabrous skin (SAI, RA, and PC) are activated when a textured surface is displaced over the skin. One variant of this, based principally on psychophysical data, is the duplex theory whereby spatial cues are thought to contribute to the appreciation of coarser surfaces (see below),

but vibratory cues (presumably via PC afferents with their exquisite sensitivity to mechanical vibration) are thought to be critical for the appreciation of finer surfaces. Unfortunately, recordings from PC afferents have not confirmed that PC discharge covaries with the subjective roughness of fine textures.

Another possibility is some form of temporal code, an observation consistent with the demonstration that peripheral afferents fire in phase-locked bursts to the temporal frequency of mechanical gratings, and that this periodicity is preserved within the firing pattern of both thalamic and somatosensory cortical neurons. The observation that the subjective roughness of ridge-varying gratings is critically dependent on temporal and not spatial variables suggests that a temporal code may well contribute to roughness appreciation. The close correlation between tangential force variations and roughness also supports the importance of a temporal code. A temporal code is not likely to be exclusive since it is difficult to reconcile with the observation that roughness estimates are independent of scanning speed. Nevertheless, temporal coding could complement an intensive code.

Finally, a spatial code has also been proposed, and is supported by the results of a number of studies. This hypothesis was developed to explain the inverted U-shaped relation between subjective roughness and two-dimensional raised dot arrays (**Figure 1(b)**). The spatial code is defined as the mean absolute difference in firing rate between afferent fibers separated by 2 mm. Such a neuronal code has the interesting property of being insensitive to changes in scanning speed, and so consistent with observations that roughness perception is independent of scanning speed (**Figure 1(a)**). According to the spatial variation hypothesis, differences in the firing rates of SAI afferents innervating skin regions separated by 2 mm are converted centrally, at some unspecified location, into an intensive code that signals roughness independent of the scanning conditions. Critical to this hypothesis, the required fine-grained spatial detail is indeed preserved in the discharge of neurons in area 3b of primary somatosensory cortex. Other studies have, on the other hand, also shown the existence of neurons, including in area 3b, whose mean discharge rate covaries with the SP of either gratings or raised dot surfaces, as predicted by the intensive code. At present, it is not clear if such cells reflect the result of the central transformation suggested for the spatial code, or a simpler intensive code.

The SAI spatial code has been extended to other types of surfaces (gratings, random raised dot surfaces). While SAI spatial variation may be the neuronal code employed across the smooth-rough continuum, SAI mean impulse rate may also contribute since there is evidence that SAI mean impulse rate, as well as SAI spatial variation, both covary with the perceived roughness of fine textures.

Remote Sensing of Roughness

The coding mechanisms described above are applicable to textures explored with the bare finger, that is, direct contact. There is now a growing body of knowledge about remote sensing of texture, using a hand-held probe, particularly important for new technical applications such as haptic interfaces for the remote operation of instruments and training surgical skills within virtual-reality environments. The various approaches taken all share a common theme in using force-feedback through a hand-held tool or implement to guide the operator in interacting with the remote or virtual environment. When using a probe, roughness discrimination has variously been reported to be unaffected or impaired. In contrast, subjective roughness is increased when surfaces are explored with a probe as compared to the bare finger. It is evident that a spatial code cannot account for the appreciation of roughness through a hand-held probe. Instead, roughness must be dependent on vibrations transmitted to the hand through the probe. Vibration sensitivity is highest for PC afferents. They act as high-pass filters and have a threshold of $<1\,\mu m$ for frequencies of 250–300 Hz. RA afferents are also sensitive to vibration, but for lower frequencies corresponding to the flutter range of 2–50 Hz. Their threshold for activation is higher than for PC afferents (10–100 μm). The challenge for the future will be to determine the respective roles of each type of afferent to the remote sensing of roughness.

Summary

Tactile roughness is a rich and varied sensation, dependent on numerous parameters of surface geometry, so that no single parameter can precisely define roughness. The dynamics between the explored surface and the skin contribute to subjective roughness, but these have not as yet been fully characterized. A number of peripheral neuronal codes for roughness have been proposed, and the SAI spatial code has received considerable experimental support. Nevertheless, while a single, parsimonious explanation for tactile roughness is a very attractive idea, alternate hypotheses need to be more fully explored, and this using a variety of different surfaces, covering the entire range from the micron to the millimetre level.

See also: Cross-Modal Interactions Between Vision and Touch; Human Haptics; Somatosensory Cortex: Functional Architecture; Somatosensory Cortex; Somatosensory Perception; Somatosensory Receptive Fields; Tactile Coding in Peripheral Neural Populations.

Further Reading

Blake DT, Hsiao SS, and Johnson KO (1997) Neural coding mechanisms in tactile pattern recognition: The relative contributions of slowly and rapidly adapting mechanoreceptors to perceived roughness. *Journal of Neuroscience* 17: 7480–7489.

Cascio CJ and Sathian K (2001) Temporal cues contribute to tactile perception of roughness. *Journal of Neuroscience* 21: 5289–5296.

Connor CE, Hsiao SS, Phillips JR, and Johnson KO (1990) Tactile roughness: Neural codes that account for psychophysical magnitude estimates. *Journal of Neuroscience* 10: 3823–3836.

Hollins M and Risner SR (2000) Evidence for the duplex theory of tactile texture perception. *Perception and Psychophysics* 62: 695–705.

Jiang W, Tremblay F, and Chapman CE (1997) Neuronal encoding of texture changes in the primary and secondary somatosensory cortical areas of monkeys during passive texture discrimination. *Journal of Neurophysiology* 77: 1656–1662.

Johnson KO and Hsiao SS (1992) Neural mechanisms of tactual form and texture perception. *Annual Review of Neuroscience* 15: 227–250.

Jones LA and Lederman SJ (2006) *Human Hand Function*. New York: Oxford University Press.

Katz D (1925) *The World of Touch*. Hillsdale, NJ: Erlbaum (Reprint 1989, Krueger LE, translator).

Meftah E-M, Belingard L, and Chapman CE (2000) Relative effects of spatial and temporal characteristics of scanned surfaces on human perception of tactile roughness using passive touch. *Experimental Brain Research* 132: 351–361.

Miyaoka T, Mano T, and Ohka M (1999) Mechanisms of fine-surface-texture discrimination in human tactile sensation. *Journal of the Acoustical Society of America* 105: 2485–2492.

Morley JW, Goodwin AW, and Darian-Smith I (1983) Tactile discrimination of gratings. *Experimental Brain Research* 49: 291–299.

Sathian K, Goodwin AW, John KT, and Darian-Smith I (1989) Perceived roughness of a grating: Correlation with responses of mechanoreceptive afferents innervating the monkey's fingerpad. *Journal of Neuroscience* 9: 1273–1279.

Sinclair RJ and Burton H (1991) Neuronal activity in the primary somatosensory cortex in monkeys (*Macaca mulatta*) during active touch of textured surface gratings: Responses to groove width, applied force and velocity of motion. *Journal of Neurophysiology* 66: 153–169.

Smith AM, Chapman CE, Deslandes M, Langlais J-S, and Thibodeau M-P (2002) The role of friction and tangential force in the subjective scaling of tactile roughness. *Experimental Brain Research* 144: 211–223.

Tremblay F, Ageranioti-Bélanger SA, and Chapman CE (1996) Cortical mechanisms underlying tactile discrimination in the monkey. I: Role of primary somatosensory cortex in passive texture discrimination. *Journal of Neurophysiology* 76: 3382–3403.

Yoshioka T, Gibb B, Dorsch AK, Hsiao SS, and Johnson KO (2001) Neural coding mechanisms underlying perceived roughness of finely textured surfaces. *Journal of Neuroscience* 21: 6905–6918.

Target Selection for Pursuit and Saccades

R J Krauzlis and L Chukoskie, Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When we visually inspect our surroundings, we typically direct our line of sight to objects or features of particular interest, even when those items move. These eye movements are under voluntary control and involve selectively sampling the visual environment because tracking one visual object typically causes the retinal images of other objects to become blurred. Accordingly, these voluntary eye movements are found only in animals that have retinal specializations, such as the primate fovea, that can be used to examine a limited region of visual space at a higher spatial resolution.

There are two types of voluntary eye movements used by animals with foveal vision (illustrated in Figure 1). The first type is saccadic eye movements, which act to quickly move the image of a visual target from an eccentric retinal location to the center of the retina, where it can be seen best. Saccades are rapid movements, reaching peak speeds of several hundred degrees per second and lasting only tens of milliseconds. Consequently, although saccades are typically evoked by stimulus events at latencies of 150–250 ms, the trajectory of saccades is not guided by sensory feedback. Instead, saccades follow a stereotyped profile that depends on the amplitude of the movement. Saccades are typically trigged by visual stimuli, but they can be evoked and guided by other stimulus modalities and can also be directed toward the location of imagined or remembered targets.

The second type of voluntary eye movements is called smooth pursuit, which slowly rotates the eyes to compensate for any motion of the visual target and thus acts to minimize the blurring of the target's retinal image that would otherwise occur. Smooth pursuit is a relatively slow movement whose trajectory is determined by the moving stimulus. Pursuit is triggered by visual stimuli at latencies similar to, but somewhat shorter than, the latencies of saccades (typically 100–200 ms) and generates eye velocities up to approximately $50\,^{\circ}s^{-1}$, continuously adjusted by visual feedback about the target's retinal image. Unlike saccades, pursuit cannot be generated in the absence of a stimulus, although moving targets that are sensed through modalities other than vision can also guide pursuit.

As this brief overview indicates, saccades and pursuit have very different properties. The neural mechanisms for the two types of movements have also been conventionally viewed as involving distinct neuroanatomical circuits. However, during normal behavior the two types of movements are tightly coordinated and almost always act to stabilize the same visual stimulus on the retina (Figure 1). Results from recent behavioral and physiological studies have provided insights into how the coordination between these two motor systems is achieved.

Gating of Pursuit and Saccades

The initiation of both pursuit and saccades involves a break from the state of fixation, which otherwise keeps the eyes from shifting to the new target. This break from fixation is controlled by similar factors for pursuit and saccades, and can be either facilitated or inhibited, depending on the circumstances. In the gap paradigm, the break from fixation is facilitated by manipulating the temporal sequence in which the visual stimuli are presented. In a typical simple eye-movement task, a second target is presented at an eccentric position immediately after the initial foveated target has been extinguished, and saccades and pursuit are evoked at latencies of approximately 150–200 ms. However, in the gap paradigm, the initial foveated stimulus is turned off a few hundred milliseconds before the eccentric target appears, and this simple change consistently reduces the latencies of saccades and pursuit, typically by approximately 20 ms. The decreases in pursuit and saccade latencies show the same dependency on the duration of the gap, suggesting that the release of fixation for the two movements depends on a shared preparatory signal.

Conversely, the inhibition of pursuit and saccades also appears to involve a shared mechanism. In the countermanding paradigm, the break from fixation is inhibited by presenting a stop signal indicating to the subjects that they should cancel the movement that had been previously planned. By varying when this stop signal is presented relative to the go signal (e.g., the appearance of the eccentric visual target), it is possible to calculate how much time is required to cancel the movement. The calculation is complicated by the fact that the final motor pathways for pursuit and saccades are different – in particular, the generation of saccades appears to include an approximately 20-ms ballistic interval that is not present for pursuit. However, once these differences are taken into account, it has been found that the time

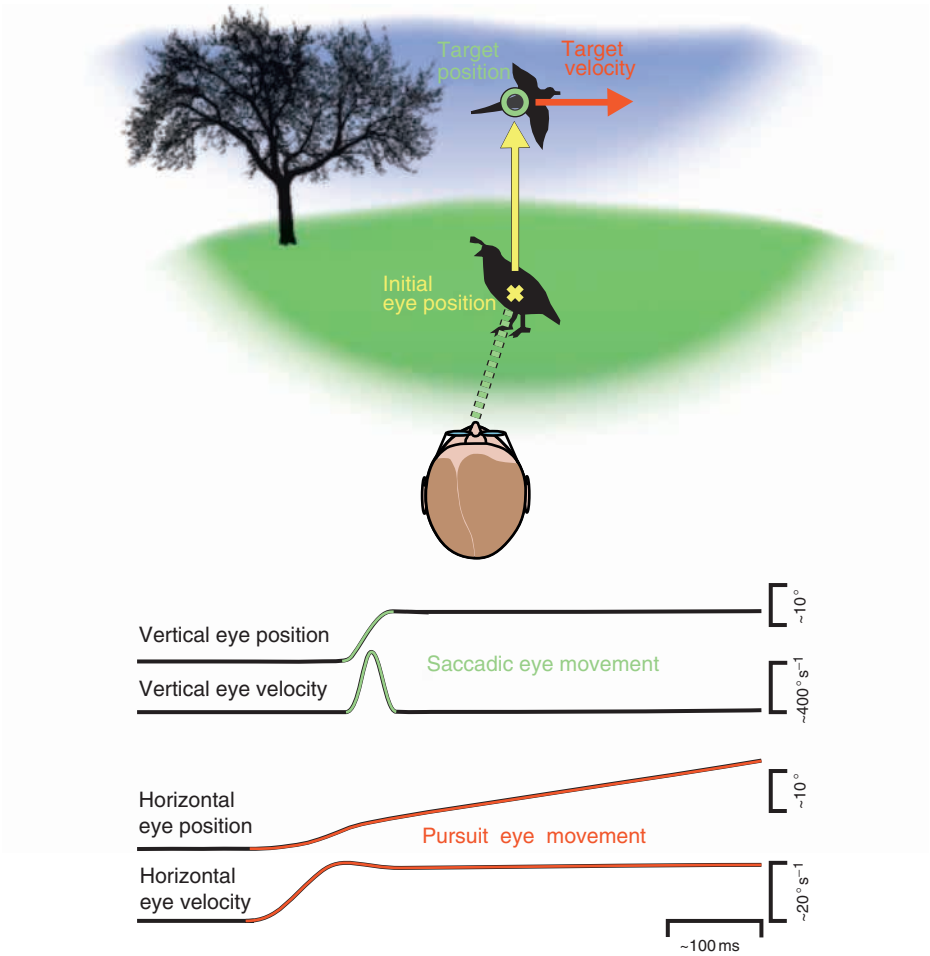

Figure 1 The selection of targets for saccades and pursuit is normally tightly coordinated. In this example, the observer's initial eye position is aligned on a quail standing in the grass. To stabilize the image of the dove flying overhead, the observer makes a coordinated combination of saccadic and pursuit eye movements. The saccade rapidly changes eye position (in this case, primarily vertically) based on the position of the new target, as highlighted in green in the traces of eye position and velocity. The pursuit movement smoothly adjusts the velocity of the eyes (in this case, primarily horizontally) based on the velocity of the same selected target, as highlighted in red in the traces.

taken to stop either a pursuit or saccadic eye movement is approximately 50–60 ms. Moreover, both movements show an effect of trial sequence. Pursuit and saccade latencies are longer if the previous trial contains a stop signal than if it does not. These differences occur regardless of whether the previous trial involves the same or different type of eye movement response, showing that the effect involves a shared inhibitory mechanism.

Target Selection for Pursuit and Saccades

In addition to determining whether and when to break fixation, saccades and pursuit also depend on selecting one visual target from among the various candidates within the visual scene. It is well established that the latencies of saccades increase when observers must search the visual field for a unique target among a set of stimuli and that these increases depend on the difficulty in finding the target. Similar increases in latency have been observed for pursuit eye movements when an observer must choose between two stimuli moving in opposite directions. These latency effects suggest that a target selection process might precede the initiation of both pursuit and saccades, but they do not resolve whether these effects are due to a single process or to similar but independent processes. To address the linkage between the two movements, it is necessary to consider the two movements together during the process of selection. Recent studies have addressed this challenge and have produced evidence in favor of two different hypotheses about how target selection for pursuit and saccades is coordinated.

One possibility is that target selection for pursuit is determined by the generation of saccades. When

pursuit of a moving target is initiated, it is typically accompanied by saccades that eliminate the mismatches between eye and target position that were left uncorrected by the pursuit movement. The pursuit eye velocity after the corrective saccade is usually much closer to target velocity than it was prior to the corrective saccade and higher than that observed at comparable times in trials without corrective saccades. This postsaccadic enhancement of pursuit suggests that the execution of the saccade improves the ability of the pursuit system to track the target. A closely related finding occurs when subjects initiate pursuit in response to two stimuli that move in orthogonal directions but subjects are not told at the outset which of the two will be the target. Because the target identity is not known initially, subjects begin smooth pursuit in a direction that tends toward the average of the two motions. After a brief period of this vector-average pursuit, subjects make a saccade toward one stimulus or the other and then pursue that one selectively. Once again, the execution of the saccade is closely related to the ability of the pursuit system to selectively track the target, supporting the suggestion that pursuit target selection is serially linked to the execution of the saccade.

Another possibility is that targets are selected in common for pursuit and saccades. By using a discrimination task that requires subjects to select the target using a combination of horizontal pursuit and vertical saccades, it is possible to compare the pursuit and saccade choices on a trial-by-trial basis. Even though they are aimed in orthogonal directions, pursuit and saccade choices agree on the target much more frequently than expected by chance. In the minority of trials in which they initially disagree (1–13%), pursuit eye velocity reverses direction producing turn-around pursuit that anticipates the choice made by the saccade that follows. These findings suggest that the two choices are guided by the same decision process but that saccades might use a somewhat more stringent decision criterion. Consistent with this view, when performance (e.g., discriminability) is considered as a function of reaction time, pursuit and saccades exhibit nearly the same trade-off between speed and accuracy but saccades tend to be delayed and somewhat more accurate.

How can this evidence of a shared decision process be reconciled with the observations of postsaccadic enhancement of pursuit? One possibility is that the decision signal evolves fairly abruptly but stochastically during single trials. In this case, the occurrence of the saccade provides a temporal marker for when the decision signal reaches a critical level, and we would expect the selectivity of pursuit to increase markedly at about this same time if it is guided by the same decision signal but used different criteria. As expected from this interpretation, there is a telltale correlation between the absolute latencies of the saccades and the differences in latency between pursuit and saccades on each trial. This correlation between relative and absolute latencies is what is expected if two different response thresholds were applied to the same decision signal. On the other hand, it is evident that saccades are not strictly necessary for pursuit target selection because pursuit can selectively follow one stimulus in the presence of distracters even when no corrective saccades are generated.

It may not be initially evident how such a shared decision signal operates because saccades depend on selecting the correct stimulus location, whereas pursuit depends on selecting the correct stimulus motion. However, there is evidence that spatial information has a surprisingly large influence on target selection for pursuit, even though stimulus motion is the crucial signal for driving pursuit eye movements. When subjects are given prior information about the location of the target but not its motion, this reduces the latencies of pursuit much more than when subjects are given prior information about the motion of the target but not its location (Figure 2). Thus, for triggering pursuit eye movements, it is more important to know where the target will be than to know in which direction it will be moving. These findings suggest that spatial attention may play a prominent role in target selection for eye movements. They also provide hints about the likely neural sites and mechanisms involved in coordinating target selection for pursuit and saccades.

Neural Mechanisms of Target Selection for Pursuit and Saccades

The neural mechanisms of target selection for saccades have been studied for many years and involve cortical areas such as the frontal eye fields, lateral intraparietal area, and subcortical brain regions such as the superior colliculus (SC). The mechanisms of target selection for pursuit have not been as thoroughly studied. The middle temporal (MT) area provides the visual motion signals that are most important for driving pursuit, and although neurons in MT area show a preference for the motion of the target, this effect appears to be too small to fully account for the selectivity of the pursuit eye movement. The pursuit area in the frontal eye fields also plays an important role in the control of pursuit and is the only area identified in the cerebral cortex where electrical stimulation can directly evoke smooth eye movements.

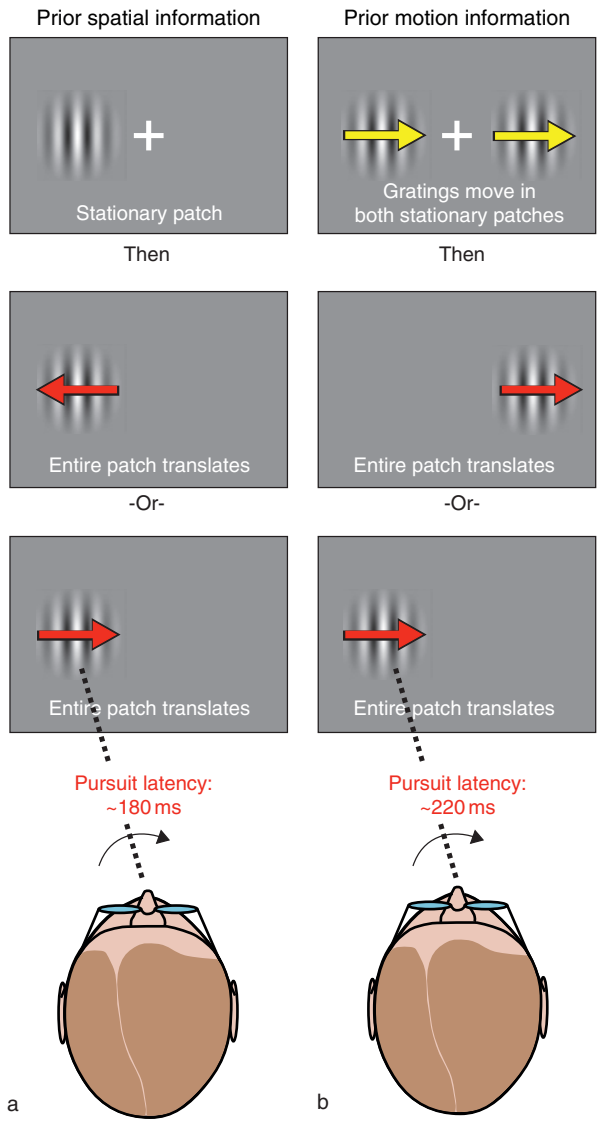

Figure 2 Spatial information influences target selection for pursuit more than motion information: (a) trial in which subjects were given prior spatial information; (b) trial in which subjects were given prior motion information. On trials with prior spatial information, a single stationary gabor patch indicated the location of the upcoming target but not its direction of motion. On trials with prior motion information, two stationary gabor patches both displayed drifting gratings that indicated the direction of motion of the upcoming target but not its location. Surprisingly, the average latency on trials with prior spatial information was shorter (~180 ms) than on trials with prior motion information (~220 ms), even though visual motion is the primary driving signal for pursuit. This effect demonstrates the primacy of spatial information in target selection for pursuit eye movements. Adapted from Adler SA, Bala J, and Krauzlis RJ (2002) Primacy of spatial information in guiding target selection for pursuit and saccades. *Journal of Vision* 2: 627–644.

However, in none of these areas is there clear evidence of a mechanism for coordinating the outputs of the pursuit and saccadic eye movement systems. In fact, although many of the same cortical areas are involved in the control of pursuit and saccades, each area contains largely nonoverlapping subregions for the two types of eye movements. How are the outputs from these two different eye-movement systems coordinated? One possibility is that the coordination is accomplished by the same selection mechanisms that control visual attention. There is emerging evidence from physiological experiments supporting the idea that the selective processing of visual stimuli (i.e., visual attention) and the preparation of voluntary eye movements involve overlapping mechanisms. Another possibility, not necessarily excluding the first, is that the coordination is accomplished by subcortical brain regions. The brain stem pathways for pursuit and saccades have long been believed to remain separated until the final motor steps, but recent physiological experiments have identified two regions where signals from the two eye movement systems coincide: the reticular formation and the SC.

The brain stem reticular formation contains the special classes of neurons that make up the saccadic burst generator – the circuit that forms the motor command for saccades. However, parts of this classic brain stem circuit for saccades also appear to be involved in the control of pursuit. For example, one well-known type of neuron in this circuit emits a burst of activity that is related to the pulse of velocity that defines the saccade trajectory. Some of these burst neurons have activity that is related to the velocity of the eyes during pursuit as well as during saccades. The function of this activity related to pursuit velocity is not yet understood, but it might be related to the mechanisms that determine when to trigger corrective saccades during pursuit eye movements.

Even more surprising is the finding of pursuit-related modulation in the activity of omnipause neurons (OPNs). It is well known that OPNs discharge steadily during fixation but stop firing during all saccades. Because this pause in activity is believed to disinhibit the saccade burst generator, OPNs are considered to act as the gatekeepers for saccade initiation. Surprisingly, approximately half of the OPNs also decrease their activity during the onset of pursuit – they do not completely pause their activity but reduce their activity depending on the speed of pursuit. Moreover, microstimulation in the region of the OPNs not only stops saccades but also strongly decelerates pursuit eye movements. These results suggest that the brain stem reticular formation contains an inhibitory mechanism that acts on both pursuit and saccadic eye movements.

The intermediate and deep layers of the SC contain neurons that have long been implicated in the selection of targets for saccades. During the preparation

of saccades, many SC neurons show increases in activity that begin hundreds of milliseconds before the movement, and these increases appear to play a role in representing the possible targets. When the probability that a visual stimulus will be the target is changed, for example by adding extra distracter stimuli, the level of preparatory activity changes in ways that reflect the altered target probability. During a visual search task, when subjects search for a uniquely colored target stimulus amid other colored distracters, many SC neurons discriminate the target from the distracter with a delay that is time-locked to the stimulus onset (rather than saccade onset), suggesting that they play a role in target selection for saccades.

Many of these saccade-related neurons also show modulation in activity during pursuit eye movements. The SC has not classically been viewed as part of the pathways for pursuit eye movements, but this modulation was noted in passing in some of the earliest studies of the SC, and more detailed studies have shown that it depends on where the neurons are located within the retinotopically organized SC map. During pursuit eye movements, each neuron may show a somewhat complicated temporal pattern of activity, but this pattern can be explained fairly simply by considering the position of the tracked target within the neuron's response field. The distribution of activity across the SC map thus provides a real-time estimate of the position of the target in oculocentric or retinotopic coordinates, not restricted to saccades but for orienting movements in general. This target position map hypothesis provides a parsimonious alternative to the widely discussed fixation zone/saccade zone hypothesis, although the issue has not yet been definitively settled.

In addition to indicating the position of the target, the activity of these SC neurons also reflects its behavioral relevance. During a visual search task, many SC neurons exhibit a preference for the rewarded stimulus over irrelevant (i.e., unrewarded) stimuli, and this preference emerges over the course of approximately 100 ms prior to the initiation of pursuit and saccades. By interpreting this preference for the rewarded target as a decision signal, it has been shown that this SC activity can account for the target

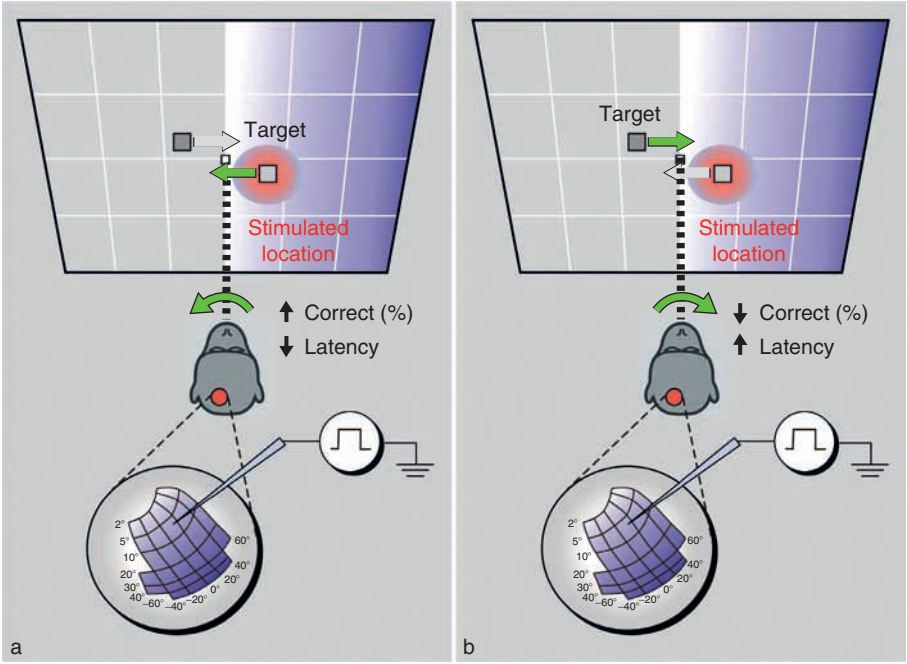

Figure 3 Experiment showing that subthreshold stimulation of the superior colliculus biases target selection for pursuit: (a) the site of stimulation matches the position of the visual target; (b) the site of stimulation is opposite the position of the visual target. An electrode placed in the intermediate layers of the superior colliculus is used to apply subthreshold electrical stimulation (i.e., currents too weak to directly evoke an eye movement) during a target selection task requiring a luminance discrimination. When the site of stimulation matches the position of the visual target (a), the percentage of correct responses increases (compared to no-stimulation control trials) and the latency of those correct responses decreases. In contrast, when the site of stimulation is opposite the position of the visual target (b), the percentage of correct responses decreases and the latency of the responses increases. Thus, activation of the superior colliculus biases target selection for pursuit eye movements, even though the required eye movements are directed away from the stimulation site. This experiment therefore demonstrates that the effect of stimulation in the superior colliculus is based on the location of the target and not the direction of the eye movement needed to acquire the target. Adapted from Carello CD and Krauzlis RJ (2004) Manipulating intent: Evidence for a causal role of the superior colliculus in target selection. *Neuron* 43: 575–583.

choices made by pursuit and saccades. Consistent with the behavioral results cited earlier, these results also imply that saccades use a more stringent decision criterion than pursuit, perhaps because errant saccades are more costly than mistakes by pursuit. These physiology results show that activity in the SC provides a neural correlate of target selection for both pursuit and saccades.

There is also recent evidence that the SC plays a causal role in target selection for pursuit and saccades. During a visual search task in which the subject should saccade to the oddball stimulus, it has been shown that, when the region of the SC representing the target is focally and reversibly inactivated, saccades are often misdirected to distracters appearing in unaffected areas of the visual field. During a luminance discrimination task, it has been found that electrical stimulation of the SC – at currents too weak to directly evoke an eye movement – nonetheless biases the selection of targets toward the stimulated location, regardless of whether the target is acquired with a pursuit or saccadic eye movement. The results for pursuit are especially revealing (**Figure 3**). Because the targets for pursuit initially appear at a location opposite to its direction of motion, the experimental design is able to identify effects based on the position of the target (i.e., which stimulus to follow) as distinguished from effects based on the motor commands (i.e., which direction to move). The results provide strong support for the idea that the SC plays a causal role in target selection itself, regardless of the type or direction of eye movement used to acquire the target.

Conclusion

Although pursuit and saccadic eye movements have been conventionally described as separate systems, their outputs are tightly coordinated during normal behavior. Recent findings indicate that this coordination is achieved by shared mechanisms for gating the movements and for selecting the visual target. The neural basis of these shared mechanisms includes the SC and possibly parts of the brain stem saccade generator. It is left for future work to clarify how these mechanisms of target selection for eye movements are related to other types of visual selection, such as the allocation of spatial attention.

See also: Brainstem Control of Eye Movements; Cortical Control of Eye Movements; Frontal Eye Fields; Oculomotor Control: Anatomical Pathways; Pursuit Eye Movements; Saccade–Pursuit Interactions; Saccades and Visual Search; Saccadic Eye Movements; Sensorimotor Integration: Attention and the Premotor Theory; Superior Colliculus.

Further Reading

Adler SA, Bala J, and Krauzlis RJ (2002) Primacy of spatial information in guiding target selection for pursuit and saccades. *Journal of Vision* 2: 627–644.

Carello CD and Krauzlis RJ (2004) Manipulating intent: Evidence for a causal role of the superior colliculus in target selection. *Neuron* 43: 575–583.

Gardner JL and Lisberger SG (2002) Serial linkage of target selection for orienting and tracking eye movements. *Nature Neuroscience* 29: 29.

Kornylo K, Dill N, Saenz M, and Krauzlis RJ (2003) Cancelling of pursuit and saccadic eye movements in humans and monkeys. *Journal of Neurophysiology* 89: 2984–2999.

Krauzlis RJ (2004) Recasting the smooth pursuit eye movement system. *Journal of Neurophysiology* 91: 591–603.

Krauzlis RJ (2005) The control of voluntary eye movements: New perspectives. *Neuroscientist* 11: 124–137.

Krauzlis RJ and Dill N (2002) Neural correlates of target choice for pursuit and saccades in the primate superior colliculus. *Neuron* 35: 355–363.

Krauzlis RJ, Liston D, and Carello CD (2004) Target selection and the superior colliculus: Goals, choices and hypotheses. *Vision Research* 44: 1445–1451.

Lisberger SG (1998) Postsaccadic enhancement of initiation of smooth pursuit eye movements in monkeys. *Journal of Neurophysiology* 79: 1918–1930.

Lisberger SG, Morris EJ, and Tychsen L (1987) Visual motion processing and sensory–motor integration for smooth pursuit eye movements. *Annual Review of Neuroscience* 10: 97–129.

Liston D and Krauzlis RJ (2005) Shared decision signal explains performance and timing of pursuit and saccadic eye movements. *Journal of Vision* 5: 678–689.

McPeek RM and Keller EL (2004) Deficits in saccade target selection after inactivation of superior colliculus. *Nature Neuroscience* 7: 757–763.

Missal M and Keller EL (2002) Common inhibitory mechanism for saccades and smooth-pursuit eye movements. *Journal of Neurophysiology* 88: 1880–1892.

Schall JD and Thompson KG (1999) Neural selection and control of visually guided eye movements. *Annual Review of Neuroscience* 22: 241–259.

Sparks DL (2002) The brainstem control of saccadic eye movements. *Nature Reviews Neuroscience* 3: 952–964.

Task Switching

D W Schneider and G D Logan, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Task switching is widely considered to be at the nexus of executive control in the human cognitive system. The ability to switch tasks is thought to require extensive high-level cognitive processing, ranging from instantiating abstract representations to preparing task-specific processes and monitoring response selection and execution. The behavioral outcome of this processing is a robust switch cost – slower and more error-prone performance when switching tasks than when repeating tasks.

Switch costs are frequently interpreted as reflecting the duration of time-consuming executive control processes that reconfigure the cognitive system when switching tasks, although this idea is debated in the behavioral literature on task switching. Some researchers have questioned whether switch costs are valid indices of executive control processes, whereas others have challenged the very notion that task switching involves executive control. There is abundant evidence indicating that basic psychological processes subserving attention, memory, and perception contribute to switch costs, compromising the interpretation of many behavioral effects as diagnostic evidence of executive control. Given that the cognitive architecture underlying switch costs and related effects is a matter of debate, it is important to exercise caution when interpreting proposed linkages between neural activity and specific cognitive processes putatively involved in task-switching performance.

With or without full appreciation of this point, many researchers have begun to explore the neural correlates of task switching. An increasing focus on identifying the neural substrates of various components of task-switching performance has produced a wealth of knowledge about task representation, selection, preparation, and switching. However, the quest to identify switch-specific neural mechanisms has proven to be challenging, with mounting evidence indicating that a distributed frontoparietal cortical network underlies task-switching performance.

Task Representation

To perform a task, it is necessary to have access to a representation of the associations between relevant stimuli and responses (i.e., task or stimulus–response mapping rules). A task representation can be considered part of a broader configuration of the cognitive system known as a task set, which includes task-specific processes that act on these representations. Many researchers propose that switching tasks requires reconfiguring the existing task set, in part to instantiate different task rules.

A brain area that is strongly implicated in task representation is prefrontal cortex (PFC). It has been proposed that PFC is involved in the acquisition, representation, and maintenance of abstract task rules for goal-directed behavior. These task rules are assumed to be extracted from past experience, in part through myriad connections to and from sensory and motor areas throughout the brain. The strongest evidence that PFC encodes task rules comes from single-cell recordings in monkeys performing different tasks. When monkeys are cued to adopt a specific task rule for responding to a stimulus, a substantial proportion of single neurons in PFC have been found to fire selectively for a given task rule. These neurons were considered rule selective because they fired after cue onset or during the delay period following the cue, regardless of the specific cue used to signal the task rule. Moreover, monkeys were able to generalize the rule to new, unseen stimuli. Such rule-selective neurons have been identified throughout dorsolateral, ventrolateral, and orbital PFC.

Task rule-selective neurons have also been found in parietal cortex. There is evidence that single neurons in the lateral bank of the intraparietal sulcus (IPS) and on the angular gyrus may encode task rules. These neurons had markedly increased firing approximately 400 ms after cue onset and fired regardless of the perceptual cues associated with each rule, suggesting that abstract task information (rather than the sensory features of the cues) had been encoded.

The evidence implicating prefrontal and parietal cortices in task representation suggests that these brain areas should be active when switching tasks, based on the proposal that task switching involves changing the existing task representation or task set. This suggestion is supported by several studies on task selection, preparation, and switching.

Task Selection, Preparation, and Switching

The primary goal in many task-switching situations is the selection and performance of a relevant task in the face of irrelevant, competing tasks. Task selection becomes crucial when stimuli are multivalent (i.e., associated with more than one task), creating

ambiguity about which task to perform. Such ambiguity can be resolved by presenting a task cue prior to or in conjunction with the imperative stimulus, enabling researchers to investigate the neural underpinnings of how tasks are selected, prepared, or switched in advance or during task execution.

Much of our knowledge concerning task switching has come from studies involving event-related functional magnetic resonance imaging (fMRI), in which researchers measure changes in the blood oxygen level-dependent (BOLD) signal elicited by different types of trials. The basic analytical strategy has been to contrast task switches with task repetitions (or other baselines) to determine which brain areas are more 'active' (i.e., generate stronger BOLD signals) when switching tasks. Although there are potential problems with this strategy (discussed later), it has yielded a large body of knowledge about the neural basis of task switching.

Neural Correlates of Task Switching

As with task representation, there is evidence that PFC underlies task selection, preparation, and switching. There is a general consensus that PFC is involved in various activities related to executive control, such as selective attention, task management, goal maintenance, and the regulation of on-line processing involving the contents of working memory. Evidence suggesting that PFC is involved in goal maintenance in dual-task and task-switching situations has come from several fMRI studies. For example, dual-task performance that involves keeping a single goal in mind activates dorsolateral PFC (Brodmann area (BA) 9), whereas keeping a primary goal in mind while performing concurrent subgoals tends to activate dorsal frontopolar PFC (BA 10). Some researchers have proposed that goal activation in PFC may bias or configure processing in other brain areas to enable implementation of the relevant task set, implying top-down control over task processing. Direct evidence in support of this proposal comes from single-cell recordings in split-brain monkeys performing a cue–stimulus association task; activity in inferior temporal cortex related to task processing was found to be modulated by activity in PFC.

The notion that PFC is involved in selecting and maintaining goals in preparation for performing or switching tasks is the main conclusion from many studies. Bilateral dorsolateral PFC (BA 9/46) activation has been observed following different types of task cues in task switching, suggesting that cue-based selection of task goals is subserved by PFC. Evidence indicates that a similar type of cue-elicited task preparation may occur in an area at the junction of the inferior frontal sulcus and inferior precentral sulcus – the so-called inferior frontal junction (which includes parts of BA 6, 8, and 44). Activation in the inferior frontal junction is separate from that associated with cue encoding and is thought to reflect task management or the updating of task representations in preparation for task switching. However, activity associated with task-switching performance extends to other regions of PFC, such as frontopolar, ventrolateral, and medial PFC.

Regarding the medial surface of the frontal lobes, researchers have also focused on the role of anterior cingulate cortex (ACC; BA 24/32) in task-switching performance. ACC is thought to play an important role in resolving conflict situations and monitoring for errors; therefore, it is likely involved in task selection when stimuli are multivalent and there is a chance of performing the wrong task. In dual-task performance, ACC activation has been interpreted as reflecting response selection among competing alternatives. In task-switching performance, ACC activation has been observed when switching between Stroop color–word tasks, which are thought to require executive control processes to resolve stimulus-induced conflict (e.g., the word 'green' printed in the color red). Even in non-Stroop task-switching situations, ACC is sometimes active and tends to be accompanied by activity in dorsolateral PFC, especially when tasks occur unpredictably and stimulus onset is uncertain. This latter point has received some attention in studies on the effects of manipulating task predictability, either by contrasting fixed and randomized task orders or by comparing cues that either specify the upcoming task or only indicate whether to switch or repeat tasks. These studies have led to the emerging idea that PFC is organized along an anterior–posterior axis, with activation becoming more anterior as task performance requires more endogenous (i.e., internal) control.

Another anterior–posterior axis figures prominently in neuroimaging research on task switching, although this axis extends across the brain from prefrontal to parietal cortex. Consistent with some of the studies on task representation, there is evidence that areas of parietal cortex are highly active when switching tasks. One such area is the IPS (BA 39/40), which has been found to be more active for task-switching than dual-task performance. The IPS is more active when task order is unpredictable rather than predictable, although there is also evidence of its involvement in cue-elicited task preparation. Moreover, there is emerging evidence for possible functional dissociations in parietal cortex, with posterior areas being more active when switching between task-relevant dimensions, and anterior areas being more active when switching between stimulus–response mapping rules.

These activations in parietal cortex tend to be interpreted in relation to activations in PFC. As mentioned previously, PFC is thought to provide top-down control signals for coordinating neural activity in other brain areas. Some researchers have interpreted activation in PFC as reflecting the selection of task-relevant neural pathways or instantiation of the appropriate task representation, with concurrent activation in parietal cortex reflecting the organization or implementation of task-specific stimulus–response mapping rules. In other words, PFC is viewed as the source of preparatory signals for biasing activation in parietal cortex and other brain areas.

Although much of the neuroimaging literature tends to emphasize the involvement of the frontal and parietal cortices in task-switching performance, many other brain areas have been found to be more active when switching than when repeating tasks, such as the cerebellum and various regions in the temporal and occipital cortices. The consistency of these activations across studies is unclear, and functional interpretations of their roles in task switching are lacking, in part because researchers have been more captivated by the frontal and parietal cortices. An important avenue for future research on task switching will be to delve into these relatively unexplored brain areas to determine if and how they contribute to the broader network of neural activity underlying task-switching performance.

Rule Switching

Given that task switching is often thought to require changing task (or stimulus–response mapping) rules, it is useful to search for neural correlates of rule switching in related experimental paradigms. One such paradigm is the Wisconsin Card Sort Test (WCST), which was a precursor to some of the modern research on task switching. In the WCST, subjects are given a set of cards on which stimuli of varying number, color, and shape are printed. The task is to learn to sort the cards into piles related to a specific stimulus dimension (e.g., color), based on feedback from the experimenter. After many consecutive correct responses (indicating that the sorting rule has been learned), the relevant stimulus dimension is surreptitiously changed, and subjects have to discover the new sorting rule (e.g., shape). The process of adopting a new sorting rule is thought to require changing the existing task set; hence, the neural correlates of WCST performance should be similar to the neural correlates of task-switching performance.

There is abundant evidence implicating PFC in WCST performance. Early neuropsychological studies revealed that patients with lesions of dorsolateral PFC performed worse than control subjects. fMRI studies have examined different components of WCST performance within PFC. Activation of the posterior part of the bilateral inferior frontal sulci (BA 45/44) increases as the number of relevant sorting dimensions increases. The right anterodorsal part (BA 46) and bilateral posteroventral parts (BA 8/44) of PFC tend to be more active when a change in the task rule requires shifting attentional set (e.g., to a different stimulus dimension), but only the latter areas are activated when the task rule is merely reversed. Although these findings of bilateral activation in PFC indicate that both cerebral hemispheres are involved in WCST performance, evidence suggests that there is functional specialization, with regions of right lateral PFC activated in response to negative feedback (when the sorting rule changes) and regions of left lateral PFC activated when updating the existing task set. Collectively, these findings of PFC activation, as well as reports of parietal cortex activation, are consistent with the neuroimaging results of many task-switching studies, suggesting that the brain areas involved in WCST performance may overlap with those involved in task-switching performance. Indeed, this suggestion is supported by a meta-analysis of WCST and task-switching studies that revealed a common pattern of distributed frontoparietal cortex activation across these paradigms.

Task-Set Inhibition

The preceding review may convey the misconception that task switching is only about the activation of task representations or stimulus–response mapping rules, but it is important to consider some of the neuroimaging evidence for inhibition in task-switching performance. When switching between different tasks (denoted A, B, and C) across a series of trials, performance is slower on the third trial of an ABA task sequence than a CBA task sequence. This impairment in performance of a recently abandoned task has been attributed to task-set inhibition – the inhibition of a previously relevant (but now irrelevant) task set when performing a different task.

Studies have found evidence suggesting that task-set inhibition involves right PFC, consistent with findings associated with other types of inhibition. Contrasting ABA and CBA task sequences in fMRI work has revealed activation in right lateral PFC (BA 45, 9/46) as well as left inferior temporal cortex (BA 37) and occipital cortex (BA 19). However, these results have been interpreted as reflecting increased task-set engagement in response to inhibition rather than task-set inhibition *per se*. Further evidence for the involvement of right PFC in task-set inhibition comes from the finding that patients with focal lesions of right PFC exhibit no behavioral manifestation of

task-set inhibition, unlike patients with lesions of left PFC and control subjects.

Correlations between Neuroimaging and Behavioral Data

If the various inferences about the roles of different brain areas in task switching are valid, then it should be possible to identify correlations between activations in specific regions and behavioral measures of task-switching performance. A positive correlation has been observed between activation of the left inferior frontal junction (as well as the presupplementary motor area) and the magnitude of a behavioral cuing effect (i.e., the decrease in response time with an increased cue–target interval), which has been interpreted as evidence of cue-elicited task preparation. When switching between Stroop color–word tasks, activation of left dorsolateral PFC is strongest for those subjects who exhibit the weakest behavioral effects related to conflict, suggesting a role for this brain area in task preparation, whereas activation of ACC is strongest for those subjects who exhibit the strongest conflict effects, suggesting a role for ACC in performance monitoring. A finding that is related to the aforementioned results concerns a region encompassing parts of the presupplementary motor area and ACC: as activity in this region increases, switch cost increases. This finding has been interpreted as evidence of neural activity increasing in parallel with the degree of interference at the time of task performance, consistent with the observation of increased switch costs with increased activity in task-irrelevant brain areas. However, there have been some reported failures to observe correlations between switch cost and neural activity, which may reflect the fact that multiple brain areas contribute to task-switching performance, a point discussed later.

Switch Specificity

Research on task representation, selection, preparation, and switching implicates an extensive fronto-parietal cortical network and many other brain areas in task-switching performance. This distributed neural basis for task switching raises an important question that has motivated many neuroimaging studies: are there any brain areas that are specific to switching tasks? This multifaceted question can be approached from three angles: analytical, empirical, and conceptual.

The Analytical Angle

Identifying a switch-specific brain area in neuroimaging research usually requires contrasting task switch trials with appropriate baseline trials. A common baseline used in fMRI studies is the pattern of activation on task repetition trials, resulting in a task switch–task repetition contrast. When a specific brain area is reliably more active (i.e., generates a stronger BOLD signal) when switching tasks than when repeating tasks, this contrast will tend to be significant. Although some researchers have used this contrast to identify brain areas involved in task switching, an important point that is sometimes overlooked is that the contrast indexes relative, not absolute, activation across different types of trials. In other words, a specific brain area may be active on both task switch and task repetition trials but exhibit stronger activation on task switch trials, yielding a significant contrast. It is clearly inappropriate in this situation to claim that the brain area associated with the contrast is switch specific (i.e., it is only involved in switching tasks), and this problem extends to region-of-interest analyses that are based on the pattern of activation associated with task repetition trials.

This issue can be partially addressed through the use of multiple baselines. For example, some researchers have supplemented the task switch–task repetition contrast with a task repetition–fixation contrast in an effort to separate out those brain areas that are active on task repetition trials. There are two potential problems with this strategy. First, the processing required for nontask events such as fixation may differ substantially from that required for task repetition trials, obfuscating the interpretation of any brain areas that yield significant task repetition–fixation contrasts. Second, application of this strategy requires making the assumption that task switch trials not only include switch-specific processes but also all of the processes engaged on task repetition trials, which may not necessarily be the case. This second problem also applies to the use of other tasks (e.g., Stroop color–word tasks) and situations (e.g., dual-task performance) as comparison conditions for task switch trials. Although these tasks and situations may be comparable to task switching (in that they are thought to require executive control processing), the differences between complex tasks may complicate any inferences regarding switch-specific brain areas.

The Empirical Angle

Analytical issues aside, there is little empirical evidence of switch-specific brain areas. Through the use of multiple baselines, some researchers have found that the same brain areas are active on both task switch and task repetition trials, suggesting that task-switching performance is not restricted to a single brain area and that there is no brain area specifically

involved in task switching. This suggestion is supported by contrasts of task-switching performance with other situations (e.g., dual-task performance), which have revealed activation of a common bilateral prefrontoparietal cortical network. Similar findings have emerged from integrative meta-analyses of Stroop and task-switching performance and different types of attention shifting (which included task switching and rule switching). For example, meta-analyses of attention-shifting studies have revealed substantial activation overlap in areas such as medial PFC, medial, superior, and inferior parietal cortex, and premotor cortex (as well as bilateral dorsolateral PFC and bilateral anterior insula, but to a lesser extent); these results have been interpreted as evidence that different types of attention shifting all depend on a similar set of working memory and executive control processes.

Some researchers have argued against the notion of switch-specific brain areas on the basis of other evidence. For example, the lack of a relationship between switch-related neural activity and behavioral measures of task-switching performance has been interpreted as evidence against a switch-specific neural mechanism. An electrophysiological study of changes in the electrical brain potential over frontal cortex, supplemented with fMRI data, did not yield any clear markers of task-switching performance. On a related note, there is evidence that changes in event-related electrical brain potentials following a single task switch trial can persist over a subsequent run of task repetition trials, suggesting that activity on task repetition trials may not be pure (in the sense of not reflecting task switching). A direct implication for fMRI studies of task switching – which typically involve measuring changes in cerebral blood flow over several seconds – is that task repetition trials may not represent an appropriate baseline for task switch trials in standard contrasts.

The Conceptual Angle

The lack of evidence for switch specificity is not surprising if one considers the broader conceptual issue of what is likely involved in task switching. In the behavioral literature, many component processes have been proposed to account for task-switching performance, and it is unlikely that all of these processes are exclusive to task switching or implemented in the same brain area. The idea that task switching requires the integration of various perceptual, attentional, mnemonic, and motor processes is supported by the abundant evidence implicating an extensive frontoparietal cortical network in task-switching performance and the conjecture that frontal areas integrate and coordinate information that is then transmitted to parietal areas for task-specific processing. For these reasons, it may be difficult to find robust correlations between switch cost and neural activity, and the search for switch-specific brain areas may be in vain. Indeed, some authors have argued that attempts to localize any complex cognitive processes in the brain may be misguided because such processes likely involve a distributed, interacting network of brain areas. If this is the case, then perhaps cognitive neuroscientists should focus their efforts on identifying and understanding the functional linkages between different brain areas involved in task-switching performance rather than seeking evidence of switch-specific neural mechanisms.

Conclusion

Despite the fact that the cognitive neuroscience of task switching is only approximately 7 years old, much progress has been made toward identifying the neural correlates of task-switching performance. Research on task representation, selection, preparation, and switching suggests that myriad brain areas – stretching from prefrontal to posterior parietal cortices – are involved in task switching. Although there has been interest in isolating switch-specific neural mechanisms, careful consideration from analytical, empirical, and conceptual angles suggests that no single brain area underlies performance. As researchers continue to delve into the multifaceted nature of task switching, a clearer picture of the intricate functional connections between brain areas will likely emerge, hopefully providing insight into the cognitive architecture of task switching and a window to the broader realm of executive control in human cognition.

See also: Attentional Networks in the Parietal Cortex; Cognition: An Overview of Neuroimaging Techniques; Executive Function and Higher-Order Cognition: Neuroimaging; Parietal Cortex and Spatial Attention; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex.

Further Reading

Brass M and von Cramon DY (2004) Decomposing components of task preparation with functional magnetic resonance imaging. *Journal of Cognitive Neuroscience* 16: 609–620.

Buchsbaum BR, Greer S, Chang W-L, and Berman KF (2005) Meta-analysis of neuroimaging studies of the Wisconsin card-sorting task and component processes. *Human Brain Mapping* 25: 35–45.

Derrfuss J, Brass M, Neumann J, and von Cramon DY (2005) Involvement of the inferior frontal junction in cognitive control:

Meta-analyses of switching and Stroop studies. *Human Brain Mapping* 25: 22–34.

Dove A, Pollmann S, Schubert T, Wiggins CJ, and von Cramon DY (2000) Prefrontal cortex activation in task switching: An event-related fMRI study. *Cognitive Brain Research* 9: 103–109.

Dreher J-C, Koechlin E, Ali SO, and Grafman J (2002) The roles of timing and task order during task switching. *NeuroImage* 17: 95–109.

Kimberg DY, Aguirre GK, and D'Esposito M (2000) Modulation of task-related neural activity in task-switching: An fMRI study. *Cognitive Brain Research* 10: 189–196.

MacDonald AW, Cohen JD, Stenger VA, and Carter CS (2000) Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* 288: 1835–1838.

Miller EK and Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24: 167–202.

Monsell S (2003) Task switching. *Trends in Cognitive Sciences* 7: 134–140.

Pollmann S, Dove A, von Cramon DY, and Wiggins CJ (2000) Event-related fMRI: Comparison of conditions with varying BOLD overlap. *Human Brain Mapping* 9: 26–37.

Rushworth MFS, Passingham RE, and Nobre AC (2002) Components of switching intentional set. *Journal of Cognitive Neuroscience* 14: 1139–1150.

Sohn M-H, Ursu S, Anderson JR, Stenger VA, and Carter CS (2000) The role of prefrontal cortex and posterior parietal cortex in task switching. *Proceedings of the National Academy of Sciences of the United States of America* 97: 13448–13453.

Stoet G and Snyder LH (2004) Single neurons in posterior parietal cortex of monkeys encode cognitive set. *Neuron* 42: 1003–1012.

Swainson R, Cunnington R, Jackson GM, et al. (2003) Cognitive control mechanisms revealed by ERP and fMRI: Evidence from repeated task-switching. *Journal of Cognitive Neuroscience* 15: 785–799.

Wager TD, Jonides J, and Reading S (2004) Neuroimaging studies of shifting attention: A meta-analysis. *NeuroImage* 22: 1679–1693.

Wallis JD, Anderson KC, and Miller EK (2001) Single neurons in prefrontal cortex encode abstract rules. *Nature* 411: 953–956.

Yeung N, Nystrom LE, Aronson JA, and Cohen JD (2006) Between-task competition and cognitive control in task switching. *Journal of Neuroscience* 26: 1429–1438.

Taste: Vertebrate Central Pathways

D V Smith[†] **and J D Boughter Jr.** University of Tennessee Health Science Center, Memphis, TN, USA

© 2009 Published by Elsevier Ltd.

Introduction

The sense of taste provides an important gateway to the body's internal environment by monitoring the ingestion of foods and fluids. This function is reflected in the reciprocal connectivity of central taste nuclei with forebrain areas involved in motivation and homeostatic regulation. The taste sensory organs, called taste buds, are situated at the entrance to the alimentary canal, on the tongue, palate, pharynx, and larynx. Transduction of sweet, salt, bitter, acid, and umami (i.e., L-type amino acids such as glutamate) stimuli occurs via G-protein-coupled receptors and ion channels located on the apical processes of taste receptor cells (TRCs). TRCs make synaptic connections with gustatory fibers of three cranial nerves: VII, IX, and X. These afferents terminate in the nucleus of the solitary tract (NST). From this nucleus, taste information is conducted to either local brain stem circuits, or into higher forebrain pathways.

Gustatory Nuclei

Afferent Nerves

TRCs, contained within taste buds, make synaptic connections with nerve fibers of one of three cranial nerves: VII, IX, or X. Separate branches of the facial (VII) nerve innervate taste buds on the anterior tongue and the soft palate. The cell bodies of these afferent neurons are located in the geniculate ganglion. The chorda tympani branch of VII also provides innervation for a small percentage of taste buds located on the lateral aspect of the posterior tongue. However, the majority of these taste buds, as well as all of those contained in the vallate papilla on the dorsal surface of the posterior tongue, receive innervation from the glossopharyngeal (IX) nerve. The cell bodies of the IXth nerve are located within the petrosal ganglion. Taste buds distributed on the laryngeal surface of the epiglottis receive innervation from the superior laryngeal branch of the vagus (X) nerve, whose cell bodies are found in the nodose ganglion. Both the IXth and Xth nerves are probably involved in the innervation of taste buds on the nasopharynx and esophagus, although the exact distribution of this innervation in either humans or rodents is not well understood. The oral mucosa and lingual epithelium are also innervated by general sensory or general visceral fibers mediating touch, thermal sensitivity, and pain, transmitted via the trigeminal nerve (V), as well as the IXth and Xth nerves.

Nucleus of the Solitary Tract

The NST is located in the medulla, and comprises an important brain structure, largely involved in modulating autonomic function. It is also the site of the first central synapse for taste information. Gustatory and visceral input from cranial nerves VII, IX, and X terminates within the medulla in the NST, in an overlapping rostral to caudal distribution. The most rostral pole of the NST receives afferent input from both the anterior tongue and palate via the VIIth nerve. Input from the IXth nerve overlaps with the more caudal portion of the VIIth nerve distribution within the NST, although it extends to more caudal levels. Similarly, the gustatory input from the Xth nerve, although overlapping somewhat with that of the VIIth and IXth nerves, is distributed more caudally still. The most caudal aspect of the NST receives vagal afferent input from the viscera. A schematic diagram of the ascending and descending connections of the rodent gustatory system, including the NST, is shown in Figure 1.

The ascending gustatory pathway originates primarily from neurons in the rostral central and rostral lateral subnuclei of the NST. The ascending pathway is responsible for the conscious awareness of taste quality, taste hedonics (pleasantness or unpleasantness), and the integration of sensory activity associated with eating and drinking. This projection from NST terminates in the taste area of the parabrachial nucleus (PbN). Other NST neurons, mostly in the rostral ventral region, project ventrally and/or caudally to adjacent somatic and visceral premotor or motor areas, including the caudal NST, the reticular formation, or the hypoglossal nucleus. These projections are responsible for the modulation of ingestive circuits and reflexes such as mastication, licking, swallowing, coughing, and salivation.

The gustatory NST contains several morphologically distinct types of neurons, including fusiform, stellate, or ovoid in shape, although there is as yet no obvious correlation between morphology and function. Transmitters involved in synaptic processing in the NST include γ-aminobutyric acid (GABA), which produces inhibition of activity in NST neurons, and the excitatory amino acid glutamate, which likely acts as a neurotransmitter between

[†]Deceased.

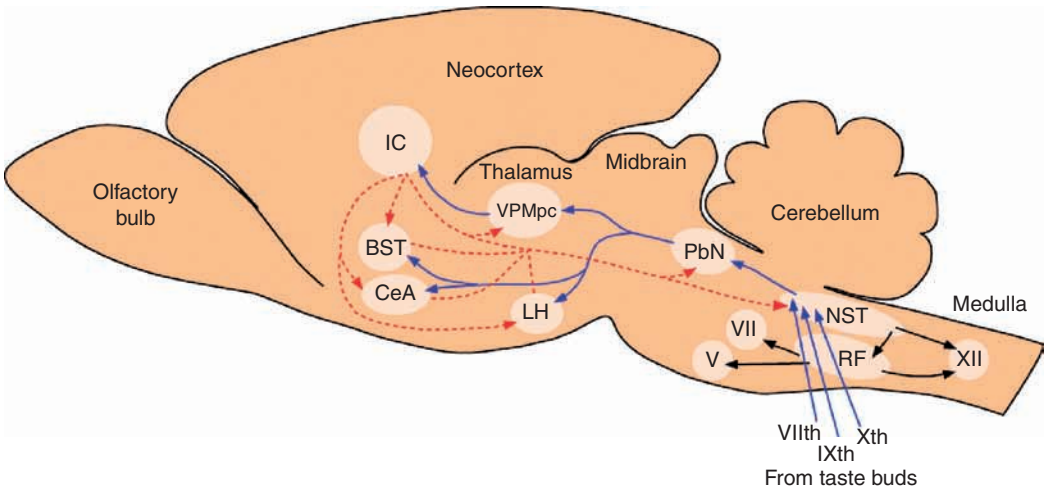

Figure 1 Schematic diagram of the ascending and descending gustatory pathways. Solid blue lines show ascending connections of the rodent gustatory system within the central nervous system (CNS); dashed red lines indicate descending projections. Black lines indicate local circuit projections from NST to medullary nuclei. IC, insular cortex; VPMpc, ventral posterior medial nucleus of the thalamus, parvicellular division; BST, bed nucleus of the stria terminalis; CeA, central nucleus of the amygdala; LH, lateral hypothalamus; PbN, parabrachial nucleus; NST, nucleus of the solitary tract; RF, reticular formation; VII, V, and XII, facial, trigeminal, and hypoglossal motor nuclei; VIIth, IXth, and Xth, axons of peripheral gustatory fibers in these cranial nerves.

gustatory afferent fibers and taste-responsive cells in the NST. Evidence also indicates that taste-responsive cells in the NST are excited by substance P (SP) and inhibited by met-enkephalin, and there exists a potential functional role in taste processing for cholinergic receptors (AChRs), found expressed in NST neurons.

Electrophysiological studies have shown that there is convergence of gustatory input onto neurons of the NST, both within and across receptive fields, including from different taste bud populations such as anterior tongue and palate. Single taste buds on the anterior tongue converge onto chorda tympani (CT) nerve fibers, which converge even further as those afferent fibers innervate neurons of the NST. In addition, there is evidence that axons from different peripheral nerves also converge onto single neurons of the NST. Physiological examination of the gustatory sensitivities of either afferent fibers or NST neurons demonstrates that these cells respond to taste stimuli of different qualities. Investigators have often classified taste neurons according to 'best stimulus' (Figure 2), although this organization is subject to modulation by stimulus type and concentration.

Parabrachial Nucleus

About 80% of gustatory neurons in the NST of rodents send second-order projections to the PbN of the pons, which serves as an obligate relay in the taste system of most mammals. This gustatory projection parallels that of the visceral afferent system. In primates, including humans, there is a direct projection from the rostral NST to the ventral posterior medial nucleus of the thalamus; only the visceral portions of the NST appears to possess a relay in the primate PbN. As in the NST neurons, PbN neurons tend to be broadly responsive to gustatory stimuli of multiple qualities, and in fact electrophysiological evidence in rodents suggests that there is a further convergence of afferent taste information onto cells of the PbN. Studies of conditioned taste aversion suggest that taste-responsive neurons in the PbN play a key role in this form of simple associative learning.

From the PbN, two divergent pathways carry gustatory information into the forebrain. The thalamocortical pathway contains axons from the PbN that terminate in the ventral posterior medial nucleus of the thalamus, parvicellular division (VPMpc), from where axons arise to project to the primary gustatory cortex. Other cells of the PbN project into areas of the limbic forebrain, including the lateral hypothalamus (LH), the central nucleus of the amygdala (CeA), and the bed nucleus of the stria terminalis (BST). Double-labeling studies demonstrated that the vast majority of PbN neurons project into either the VPMpc or CeA, but not both. The PbN projects bilaterally to the VPMpc. The PbN projections to the limbic areas are extensive and reflect the substantial interactions between the gustatory and general visceral sensory systems, as well as the interconnectivity between taste and feeding behavior.

Thalamocortical Pathway

Gustatory thalamus Like other sensory systems, taste projects into a thalamic nucleus. The gustatory thalamic relay is located in the VPMpc of the thalamus, which is a medial extension of the ventral posterior

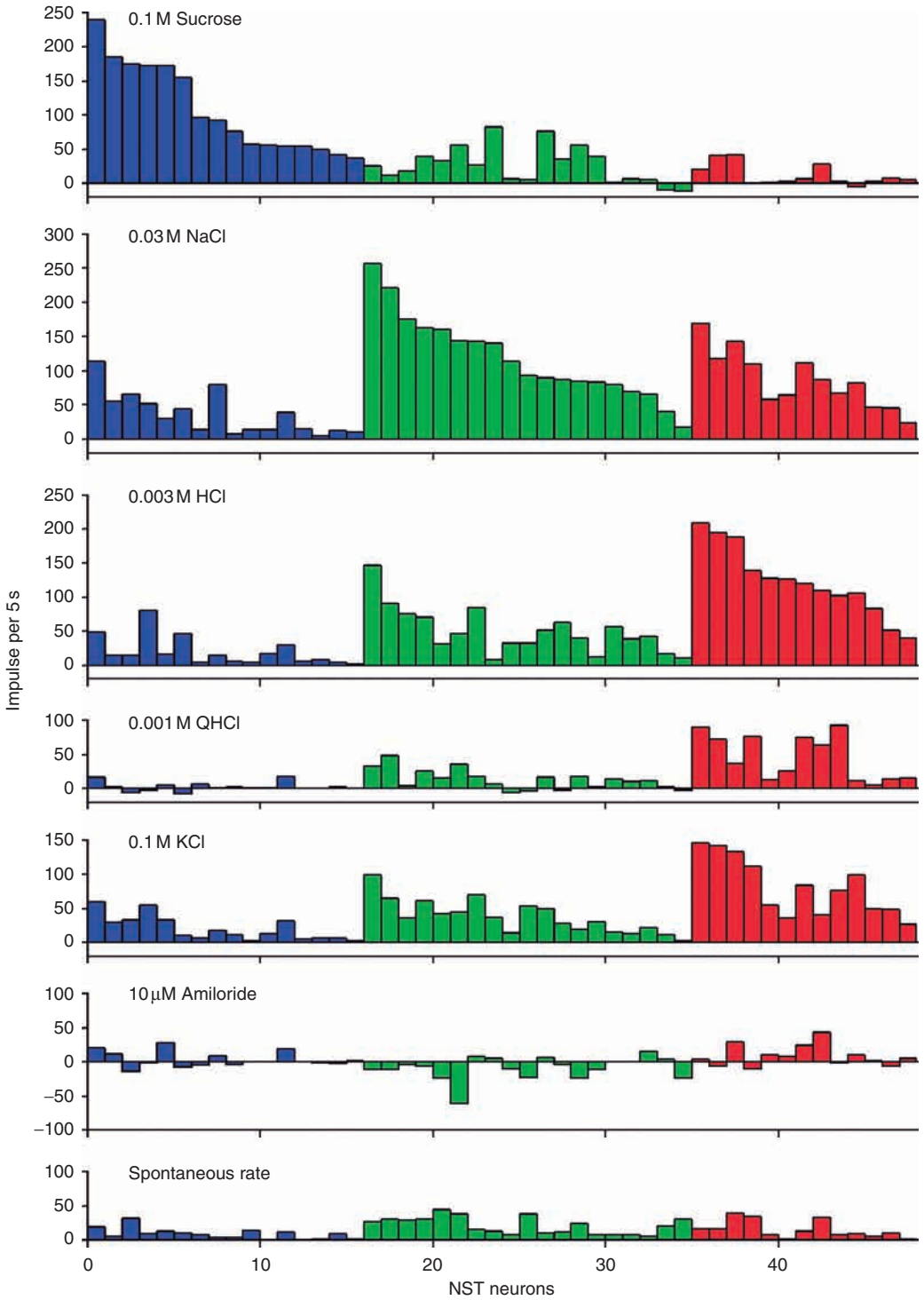

Figure 2 Across-neuron patterns of responses to six taste stimuli applied to the anterior tongue (0.1 M sucrose, 0.03 M NaCl, 0.003 M HCl, 0.001 M QHCl, 0.1 M KCl, 10 μM amiloride) in 48 individual cells in the nucleus of the solitary tract (NST) of the golden hamster. The concentrations were chosen on the basis of their half-maximal integrated response in the CT nerve. Responses are adjusted for spontaneous rate (impulses per 5 s), displayed at the bottom of the figure. Neurons are grouped from left to right into best-stimulus classes, and ranked within each group according to the magnitude of the response to the best stimulus. Response of any one neuron can be read vertically. Neurons 1–16, sucrose-best (shown in blue); 17–35, NaCl-best (shown in green); 36–48, HCl-best (shown in red). Although these neurons can be grouped according to best stimulus, it can be seen that they are generally broadly tuned, responding to stimuli of more than one quality. Reproduced from Boughter JD Jr., St. John SJ, and Smith DV (1999) Neural representation of the taste of NaCl and KCl in gustatory neurons of the hamster solitary nucleus. *The Journal of Neurophysiology* 81: 2636–2649. Used with permission from The American Physiological Society.

medial nucleus of the thalamus (VPM), the thalamic relay for trigeminal somatosensory information from the head. Therefore, taste representation is found adjacent to tongue somatosensory representation. Additionally, neurons in the VPMpc tend to be responsive to multimodal stimuli such as taste and oral tactile input. Therefore, the activity of these neurons encompasses all forms of sensory stimulation of the oral cavity. Ascending gustatory fibers from the PbN also cross over at the level of VPM to terminate in the contralateral VPMpc, accounting for bilateral responses of thalamic neurons to taste stimuli.

Insular cortex In rodents, gustatory thalamic neurons project to agranular insular cortex, located on the lateral convexity of the cerebrum, dorsal to the rhinal fissure. In contrast, oral somatosensory input to primary somatosensory cortex is targeted to granular cortex, immediately dorsal to the agranular gustatory cortex. In primates, the primary gustatory neocortex corresponds to both the frontal operculum and the buried insular cortex toward the rostral portion of the lateral fissure. Both of these areas receive major input from VPMpc. A second gustatory cortex in primates is located within orbitofrontal cortex of primates but does not receive direct projections from VPMpc. In gustatory cortex, the activity of single neurons tends to be multimodal. Recent studies suggest a role for interactions between ensembles of gustatory cortical neurons recruited during taste discriminations.

Insular cortex has widespread descending projections to subcortical taste and taste-related nuclei, including VPMpc, amygdala, LH, and NST. These descending projections may modulate the processing of ascending gustatory information. Corticofugal input to the NST appears to differentially inhibit gustatory afferent activity through GABAergic mechanisms. This modulation is often bilateral, and there is also evidence for excitatory corticofugal inputs to the NST.

Limbic Forebrain

Cells within the PbN project to the LH, CeA, BST, and substantia innominata. These are brain areas important in food and water regulation, responses to stress, conditioned taste-aversion learning, and mechanisms of reinforcement, reward, and drug addiction. Electrophysiological studies have shown the existence of taste-responsive neurons in several of these areas, including the LH and the CeA. Each of these forebrain targets sends centrifugal axons to both the PbN and the NST and have been shown to produce modulation of brain stem taste activity in the NST and/or the PbN.

Stimulation of the LH produces increases in food intake and alterations in taste preference behavior, whereas damage to this area has opposite effects. About half of the taste-responsive cells in the NST are modulated by LH stimulation. This input from LH tends to be contralateral and excitatory; recent studies indicate responses to taste stimuli were more than doubled by this excitatory influence, with no alteration of spontaneous activity. These effects may in fact enhance taste discriminability by increasing the signal-to-noise ratio of taste-evoked activity. The CeA contains neurons that respond differentially to hedonically positive and negative taste stimuli, and both the CeA and the basolateral amygdala are involved in conditioned taste-aversion learning. Projections from CeA to NST are predominantly bilateral and excitatory. In contrast, descending modulation of the PbN by either LH or CeA is largely inhibitory. Stimulation of the BST produces inhibition of gustatory NST neurons. Collectively, these limbic forebrain targets form an extensive neural substrate that no doubt underlies the modulation of taste activity by physiological and experiential factors. These pathways may be engaged by alterations in blood glucose, gastric distension, conditioned taste-aversion learning, and other physiological conditions known to alter taste sensitivity.

Taste Quality Coding

Historically, there have been two opposing views of how taste information is represented in the activity of gustatory neurons. The across-fiber pattern theory postulates that taste quality is represented by the pattern of activity across the afferent population. Stimuli with similar tastes produce similar patterns of activity. An example of this type of sensory coding is found in the olfactory system, where different odorants activate distinct populations of glomeruli in the olfactory bulbs. The opposing view is that activity in a few distinct neuron types codes taste quality in a labeled-line fashion, that is, neurons responding best to NaCl would code 'saltiness.' An example of this type of sensory coding is found in the somatosensory system, where unique fiber and neuron types transmit information about touch, pain, and temperature. However, neurons higher in the afferent pathway, such as those in the thalamus and cortex, are typically multimodal and integrative in terms of sensory quality and receptive field. For taste, temporal aspects of the neural response may also encode information about quality, especially in populations of cortical gustatory neurons.

The gustatory sensitivities of peripheral axons and central neurons in rodents are well established. At each step of the gustatory pathway, from afferent fiber to cortex, neurons tend to be broadly tuned, or responsive to stimuli from multiple qualities: sweet, salt, acid, and bitter (Figure 2). Taste neurons may be classified

according to the stimulus, among a set representative of each quality, that evokes the strongest response. Within each best-stimulus group, response profiles for gustatory neurons tend to be similar, suggesting a certain amount of homogeneity within groups. Hierarchical cluster analyses of responses of taste fibers in the CT nerve, and among gustatory neurons in the NST or PbN, provide strong support for the classification of these neurons into types, based on similarities and differences in their response profiles. However, the broad tuning means that taste-evoked activity in one cell alone cannot unambiguously represent taste quality. Rather, neuron types play a critical role in establishing unique patterns of neural activity that distinguish among taste stimuli. An example of this type of neural substrate for discrimination concerns the responses to salts such as NaCl and KCl in the rodent NST. Patterns of neural activity evoked by these two stimuli at isomolar concentrations across all cell types are distinct, resulting in a low across-neuron correlation. Yet within particular subgroups, such as HCl-best or sucrose-best neurons, the correlation is high, suggesting that discrimination can only occur when the relative response across all cell types is considered.

See also: Central Gustatory System and Ingestive Behavior; Evolution of the Limbic System; Sensory Aging: Chemical Senses; Taste: Vertebrate Psychophysics; Taste: Vertebrate Taste Bud Physiology.

Further Reading

Bradley RM, King MS, Wang L, et al. (1996) Neurotransmitter and neuromodulator activity in the gustatory zone of the nucleus tractus solitarius. *Chemical Senses* 21: 377–385.

Boughter JD Jr., St. John SJ, and Smith DV (1999) Neural representation of the taste of NaCl and KCl in gustatory neurons of the hamster solitary nucleus. *The Journal of Neurophysiology* 81: 2636–2649.

Cho YK, Li CS, and Smith DV (2003) Descending influences from the lateral hypothalamus and amygdala converge onto medullary taste neurons. *Chemical Senses* 28: 155–171.

Katz DB, Nicolelis MA, and Simon SA (2002) Gustatory processing is dynamic and distributed. *Current Opinion in Neurobiology* 12: 448–454.

Lundy RF Jr. and Norgren R (2004) Gustatory system. In: Paxinos G (ed.) *The Rat Nervous System,* 3rd edn. San Diego: Academic Press.

Mombaerts P (2004) Genes and ligands for odorant, vomeronasal and taste receptors. *Nature Reviews Neuroscience* 5: 263–278.

Norgren R and Leonard CM (1973) Ascending central gustatory pathways. *Journal of Comparative Neurology* 150: 217–238.

Rolls ET (1995) Central taste anatomy and neurophysiology. In: Doty RL (ed.) *Handbook of Olfaction and Gustation.* New York: Dekker.

Scott TR (2001) The role of taste in feeding. *Appetite* 37: 111–113.

Smith DV, St. John SJ, and Boughter JD Jr. (2000) Neuronal cell types and taste quality coding. *Physiology and Behavior* 69: 77–85.

Smith DV and St. John SJ (1999) Neural coding of gustatory information. *Current Opinion in Neurobiology* 9: 427–435.

Travers JB, Dinardo LA, and Karimnamazi H (1997) Motor and premotor mechanisms of licking. *Neuroscience and Biobehavioral Reviews* 5: 631–647.

Taste: Vertebrate Psychophysics

V B Duffy, University of Connecticut, Storrs, CT, USA
J E Hayes, Brown University, Providence, RI, USA
L M Bartoshuk and D J Snyder, University of Florida, Gainesville, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Narrowly viewed, taste is the perception of sweet, salt, sour, bitter, and possibly umami (meaty savory flavor of some amino acids) sensations. However, colloquially, taste refers to oral sensations from foods and beverages, in which taste combines with other sensory inputs to produce a complex sensation. Recently, interest in taste as a field has increased with the discovery of genes that encode receptors for bitter and sweet compounds and provide better understanding of taste transduction. Equal progress has been made in taste psychophysics – the study of the relationship between physical stimuli and the perceptions they induce. Modern psychophysical methods permit the characterization of normal variation in taste as well as variation associated with aging and environmental insult. Appropriate taste psychophysical methods are needed to fully understand genetic, molecular, and neurochemical processes involved in taste sensation. Through taste psychophysics, scientists and clinicians can determine how normal and abnormal variation in taste influences what we like to eat, what and how much we choose to consume, and ultimately our health and quality of life.

Methodologies

Called the father of psychophysics, Gustav Theodor Fechner (1801–87) established the basis of our classic methods for determining detection thresholds, the lowest level at which a stimulus is perceived. However, when given an ambiguous stimulus, an observer may classify it as a signal ("yes, something is there") or as noise ("no, it was nothing"), and this criterion depends on the context and potential outcome. Informed by cold war radar research that resulted in Signal Detection Theory, psychophysicists developed criterion-free methods that use probability to isolate observer bias from the detection of a signal within a noisy background. These procedures use repeated trials to narrow in precisely on the level at which the individual can detect the stimulus. Typically, participants are given two samples – a target and a blank – and forced to identify one as the target (a forced choice task). In one common protocol, correctly choosing the stimulus, usually in two successive trials, results in a more difficult task (i.e., a lower stimulus concentration) or a reversal. With an incorrect choice, the concentration is increased – also a reversal. Threshold value is the geometric mean of all reversals (usually six reversals).

Threshold measurements are labor and time intensive, requiring precise control of the target and blank and sufficient reversals to dampen oscillations around the true threshold. Precise control allows the assessment of factors that can change taste thresholds within an individual (e.g., medications, physiological state) or account for threshold differences across individuals (e.g., genetic variation in receptors). However, threshold represents a narrow window of the perceptual experience and thus has minimal relevance for explaining complex behaviors such as response to foods and beverages. Threshold is frequently used to assess the connection between a tastant and a specific molecular receptor yet it fails to provide insight into how the receptor functions at higher taste concentrations or the combined influence of multiple receptors to affect perceived intensity or behavioral responses to the tastant (e.g., liking/disliking).

Today, suprathreshold scaling techniques are increasingly popular since they provide much more information about sensory experience and are relevant to understanding taste-related behaviors such as eating and are more easily used outside the laboratory setting. Scaling techniques describe how perceived intensity varies across the entire dynamic range, from threshold to the strongest sensations possible, producing a psychophysical function. Gustav Fechner used difference thresholds to generate a method of measuring suprathreshold perceived intensity. He proposed that the unit of sensation be the 'just noticeable difference' (jnd), the amount a stimulus had to be raised to produce a perceptible increase. To scale perceived intensity, one needed only to count the number of jnds from threshold to the sensation of interest. A century later, SS Stevens argued that the jnd did not have the properties desirable for a unit of sensation. For example, if one stimulus is X jnds above threshold and a second stimulus is 2X jnds above threshold, then the second stimulus should be twice as intense as the first. However, Stevens showed that this second stimulus is often perceived as more intense than it should be; the subjective size of the jnd grows as the stimulus grows.

In 1969, Stevens proposed 'magnitude estimation,' which asks participants to assign numbers to sensations such that one sensation that is twice as intense as another is assigned a number twice as large (i.e., the

scale has ratio properties). Magnitude matching, used by Marks and others, extends the utility of magnitude estimation for comparisons of intensity judgments. The experimenter or clinician asks the participant to use the same scale to judge the intensities of both a tastant and a standard (perceived or remembered). For example, the saltiness of sodium chloride in chicken broth has been compared across men and women. Both groups were asked to rate a given concentration of sodium chloride in broth along with a 98 dB tone on the same scale (cross-modality matching). Assuming that the perception of loudness is independent of taste, taste can be compared across women and men as follows. The women, on average, rated the saltiness of the NaCl to be about equal to the loudness of the 98 dB tone, while men rated the saltiness to be only about two-thirds as intense as the tone was loud. Thus we can conclude the men perceived the saltiness to be about two-thirds as salty as did the women did. Women tend to have more taste buds than do men, which would explain the heightened saltiness sensation. In practice, using several standards is prudent. The standards should be independent of taste. Sound standards are commonly used although there may be associations between taste and sound through middle ear pathology (see the section titled 'Taste changes with environmental insults and across the life span').

Labeled scales (e.g., category and visual analog scales (VASs)) have historically been used to compare groups; however, such comparisons are invalid unless the experimenter can show that the labels on these scales denote the same perceived intensities, on average, to all groups to be compared. Since this is often not true, conventional labeled scales can provide erroneous comparisons. This problem can be corrected in much the same way it was corrected for magnitude estimation; devise labeled scales that permit ratings from multiple sensory continua and that thus permit magnitude matching. For example, Green and colleagues devised a scale to measure oral sensations (the Labeled Magnitude Scales or LMS). This scale used a series of intensity labels (weak, moderate, strong, etc.) and spaced them to give the scale ratio properties (a very important advancement for labeled category scales). However, as with all conventional category scales, the minimum and maximum intensities were oral sensations. By simply expressing the minimum and maximum labels in terms of all sensory experience, the scale can be used to rate taste and nontaste standards on a common scale. This generalized LMS is called the general LMS or gLMS; the gLMS produces psychophysical functions equivalent to those produced by magnitude matching using magnitude estimation.

One potential concern with the gLMS involves confusion over the intermediate labels (e.g., weak, moderate, etc.). One can think of labeled scales as elastic. The relative position of intensity descriptors appears to be essentially the same no matter the domain to which a labeled scale is applied. As this elastic scale is compressed to fit a weak domain (e.g., the odor of roses) and stretched to fit a stronger domain (e.g., pain), the relative positions of the descriptors remains the same. This means that a given intensity descriptor denotes different absolute intensities depending on the domain. For example, a woman who has just experienced a difficult childbirth may describe her pain is 'very strong.' She may also describe the odor of the roses in a bouquet presented to her to celebrate the birth as 'very strong.' One understands that she does not mean to equate the intensity of the rose odor with the intensity of her pain. Thus, there is some utility in dispensing with the intermediate descriptors in favor of a scale from 'zero' to 'strongest sensation of any kind.' A line with such labels is essentially a global VAS or gVAS. However, the experimenter may still find adjective labels useful in communicating the intensities of the sensation (e.g., strong vs. weak as compared to 37 vs. 15).

Taste Quality

Before the introduction by Henning in 1916 of the taste tetrahedron, with sweet, sour, salty, and bitter at the vertices, the number of gustatory qualities was heavily disputed. Bravo had nine classes in 1592, Linnaeus 11 in 1751, and Haller 12 in 1751, which he later reduced to six. As late as 1892, acid (sour) and salt were controversial enough that their inclusion by Shore warranted both theoretical and empirical justification. In his epic history, the eminent psychologist Edwin Boring reported the "matter was settled" by Henning's tetrahedron, although today there is still a lack consensus on exactly which qualities should be included. Almost all investigators today accept Henning's four (sweet, sour, salty, and bitter); the contention arises over whether this list is complete. Umami, the meaty, savory quality of certain amino acids, has been recognized as a fifth quality in the Japanese scientific literature for a full century, although some skeptics in the West remain unconvinced. As a sixth quality, metallic is the most common contender, with recent work suggesting that while metallic sensations rely in part on an olfactory component, taste input certainly plays a role. Notably, direct electrical stimulation of the tongue can produce metallic sensations, and these sensations are a frequent complaint in taste pathologies, which

suggests that metallic sensations may occur when patterns of taste input are altered. Finally, it should be noted that two other oral sensations are well accepted and widely recognized – oral astringency (red wine) and oral burn (chilies) – but are considered nontaste sensations as they are carried centrally to the brain by a different set of nerve fibers.

Psychophysically, two phenomena, adaptation and mixture interactions, support the independence of sweet, sour, salty, and bitter from each other. Complete adaptation can occur in taste. That is, with constant stimulation, taste sensations can completely disappear. Concentrations just above the adapting concentration will still produce the expected taste quality, albeit at a lower than expected intensity, whereas concentrations lower than the adapting concentration take on an untypical taste called a 'water taste.' (At concentrations well above the adapting concentration, intensity will rebound to expected levels due to recruitment, analogous to what occurs in audition.) The artichoke provides a dramatic example of water taste for some individuals. After exposing the tongue to globe artichokes (or to the isomers of caffeic acid from the artichokes), water tastes sweet for several minutes. Cross-adaptation refers to the ability of adaptation to one substance to affect the taste intensity of a different substance. Classic theory argued that if cross-adaptation occurred between two stimuli, then they must share receptor sites. Cross-adaptation does not occur across substances with different taste qualities. It also fails to occur across some sweet stimuli and some bitter stimuli, leading to the conclusion that sweetness and bitterness are mediated by multiple, distinct receptor sites. Mixture interactions also support the independence of the four common taste qualities. In taste mixtures, although the perceived intensities of the components may be reduced (in fact this is usually the case), the qualities remain recognizable (unless one is reduced so much that it disappears). New qualities do not emerge in taste mixtures.

Tongue Map

Contrary to popular belief, various taste qualities can be detected over the entire surface of the tongue wherever there are taste receptors. That is, the classical tongue map showing that sweet is perceived on the tip of the tongue, bitter on the back, and so forth, is entrenched in texts but is, in fact, wrong. In the history mentioned above, Boring replotted data found in Hänig's doctoral dissertation and, in doing so, did not label the ordinate axis. This omission led readers to mistakenly conclude that the small differences were, in fact, all-or-none differences. Hänig, a student of Wundt, had found that all four qualities could be tasted around the perimeter of the tongue with only some minor variability in threshold, a finding that was reconfirmed by Collings in 1974. While the classic tongue map is incorrect, we now believe that variability in taste perception associated with environmental insult (see the section titled 'Taste changes with environmental insults and across the life span') has implications for perception of food flavor, food preference, and dietary-related disease risk.

Genetic Variation

Human psychophysical data are vital to provide meaning to taste receptor genetics and taste gene expression. The most-studied genetic trait involves variation in response to phenylthiocarbamide (PTC) and 6-*n*-propylthiouracil (PROP), members of the thiourea family. Taste blindness to PTC was an accidental finding by the chemist Fox in 1931. PROP is preferred today as a taste stimulus as it is safer and lacks an odor. PROP is not found in the food supply, leading to questions of its ecological validity; however, related natural thioureas (e.g., 1-5-vinyl-2-thio-oxazolidone) are widely found in foods. Studies using PROP as a phenotypic marker have revealed associations with dietary and lifestyle behaviors and health outcomes.

PROP has become a marker for differences in overall oral sensation, probably not because of some unique receptor attribute, but because it shows large variability across individuals and because it highlights the importance of taste in the central integration of oral sensation. Some individuals live in a neon orosensory world (supertasters), and others live in a more pastel one (nontasters). Characterizing genetic variation in oral sensation could move to an expanded number of taste markers and emerging taste genotypes. For example, greater taster intensity has been characterized by other phenotypes, including response to some irritants and thermal stimulation.

Historically, PROP tasting was measured via threshold procedure, separating those who are insensitive from those who are tasters. Polymorphisms on the long arm of chromosome 7 for the *TAS2R38* gene correspond well with PTC/PROP threshold classifications. Two homozygous and a heterozygous form are seen commonly in the population: The AVI/AVI homozygote corresponds to the PROP nontaster phenotype, while the PAV/AVI heterozygote and the PAV/PAV homozygote correspond to the taster phenotype. The perceived intensity of PROP or suprathreshold response corresponds less well with *TAS2R38*, not sufficiently differentiating individuals who taste PROP as bitter (i.e., medium tasters) from those who taste PROP as exceedingly bitter (i.e., supertasters). It is interesting that some AVI/AVI individuals taste concentrated

PROP as bitter, which suggests with a high enough concentration, the hT2R38 receptor responds to PROP or that a second taste receptor is involved in PROP tasting. Some of the heightened bitterness of PROP can be explained through greater numbers of fungiform papillae, the type of taste papillae found on the anterior tongue. Those who taste PROP as bitter and extremely bitter have a high number of fungiform papillae, yet the heightened number of papillae may not fully explain PROP supertasting. Identification of other factors that may contribute to supertasting is an area of active research.

Supertasters of PROP perceive greater intensities to salty, sweet, sour, and bitter compounds. Because supertasters tend to have more fungiform papillae and these papillae hold pain and touch fibers as well as taste fibers, those who taste PROP as more bitter also experience more chemesthetic sensations from irritants (e.g., chili peppers, black pepper, ginger, carbonation, alcohol) and tactile sensations (e.g., creamy sensations from fat). PROP supertasters also experience greater retronasal sensations, especially if these sensations are congruent with taste sensations (e.g., sweet taste of sucrose with sweet odor of strawberries in strawberry jam). Thus, the term 'supertaster' could be operationalized as an individual who tastes PROP strongly bitter or has a high number of fungiform papillae, as well as a heightened response to other oral sensations.

Taste Changes with Environmental Insults and Across the Life Span

Total loss of the sense of taste, or ageusia, is rare because of the redundancy in the taste system. Branches of three cranial nerves (CNs) innervate taste buds, transmitting the electrical impulses to the central nervous system: The chorda tympani nerve (CN VII) innervates fungiform and anterior foliate papillae, and the lingual nerve (CN IX) innervates the posterior foliate and circumvallate papillae. Taste buds on the soft palate are innervated by the greater superficial petrosal nerve (CN VII), and those on the epiglottis are innervated by the superior branch of the vagus nerve (CN X). These peripheral taste messages are integrated with oral somatosensation and retronasal olfaction by the orbitofrontal cortex, which acts as a 'sensory packaging center.'

Decreased taste from one region of innervation is relatively common (i.e., spatial taste loss). Loss from one or more taste inputs alters the sensations from foods or beverages and/or may lead to phantom sensations in the absence of stimulation. Localized loss can increase signals from unaffected systems. However, losses in multiple systems can depress overall oral sensations. For example, loss of taste on the anterior tongue (chorda tympani nerve) releases the normal inhibition of other taste nerves (CN IX, X) to maintain or heighten whole-mouth taste perception. A poignant and historical example was described by Brillat-Savarin in *The Physiology of Taste*: A prisoner, after recovering from having his anterior tongue cut off for punishment, reported no loss of taste sensations and, in fact, had intense or painful sensations from very sour or bitter substances. The damage removed chorda tympani input to the CNS, releasing the inhibition from CNs IX and X to preserve most sensations and heighten others.

In the current medical literature, surgical treatments of middle-ear tumors can damage taste from the chorda tympani nerve, and tonsillectomies can damage taste from the glossopharyngeal nerve, heightening sensations from undamaged nerves. Spatial taste loss can result from exposure to common pathogens. For example, pathogens associated with middle ear infections (i.e., otitis media) can damage taste because the chorda tympani nerve passes through the middle ear on its way to the brain. As shown clinically (or temporarily with anesthesia), spatial taste loss may be the basis of dysgeusia (taste in the absence of apparent stimulation) or altered taste sensations in response to foods or beverages. Spatial taste losses affect other oral sensory inputs, including diminishing retronasal flavor perception but intensifying oral somatosensory sensations. A hypothetical example demonstrates this phenomenon. After temporary bilateral anesthesia of the chorda tympani nerve, strawberry jam would have limited sweetness and diminished strawberry flavor yet have heightened thickness and prickly sensations from the acid in the strawberry (i.e., enhanced somatosensations). This is an acute response to an experimental procedure, but it nonetheless suggests that spatial damage to taste could have implications for overall sensations from foods and beverages and resultant dietary behaviors (discussed in the section titled 'Nutritional implications of variation in taste and oral sensation'). Damage to taste from the chorda tympani nerve has also been implicated in increasing the risk of burning mouth syndrome – a phantom pain syndrome. There is growing support for the hypothesis that spatial taste damage, especially among those with high numbers of fungiform papillae (i.e., more trigeminal innervation) increases the chance for heightened response to oral irritants and, in severe cases, phantom oral pain.

Taste changes have been noted with changes in female sex hormones. Bitter response may be regulated during childbearing ages to support a healthy pregnancy. Bitter perception rises to its peak during the first trimester, helping women avoid ingesting

toxins, tetratogens, and abortifacients, and falls to its lowest in the third trimester, helping women consume a varied diet. Changes in the hedonic responses to salt may help women ingest sufficient sodium to build the blood volume to support fetal blood supply. Taste varies across the menstrual cycle and for young women. Bitterness may decline across menopause and with aging in general – a potential advantage for acceptance of bitter and nutrient-rich fruits and vegetables among the older population.

Nutritional Implications of Variation in Taste and Oral Sensation

Consumers report that 'taste' is an important determinant of what they consume. Taste integrated with other oral sensations affects hedonic responses to foods and beverages. Variation in taste and oral sensations can be linked to differential dietary risk of chronic disease by affecting food preference and possibly appetite and satiety responses. Through use of improved scaling techniques, the impact of variation in taste and oral sensations on dietary risk of obesity and other chronic diseases is now being realized.

Oral Sensation, Vegetable Intake, and Cancer Risk

One dietary strategy to lower cancer risk is through consumption of up to nine servings of fruits and vegetables per day. Fruits and vegetables contain cancer chemopreventive agents that are often bitter. Individuals who taste PROP as more bitter experience less positive (e.g., sweetness) and more bitter sensations from vegetables and thus consume fewer vegetables. Preliminary data support a link between PROP bitterness and colon cancer risk; in older men, the number of colon polyps measured during routine colonoscopies was associated with PROP bitterness.

Oral Sensation and Risk of Cardiovascular Disease

PROP nontasters, as well as those with spatial taste losses, may show greater risk of cardiovascular disease because of dietary intake, body weight, and lifestyle behaviors. In the 1960s, Fischer postulated that individuals with high PROP and quinine thresholds (insensitivity) tended to have a heavy body type whereas those with low PROP and quinine thresholds tended to have thin body types. Current literature has tended to examine associations between PROP and body weight; those tasting PROP as less bitter tend to be heavier, as seen in children, adults, and older people, although the size of the effects is small and not seen in those who are very obese. Taste damage from the chorda tympani nerve, either from reported exposure to otitis media or as functionally measured by depressed quinine bitterness on the tongue tip, appears to provide a stronger risk of overweight and obesity. Support for the spatial taste damage–obesity relationship comes from community-based research, survey data, laboratory-based data, and emerging longitudinal data from national surveillance. Laboratory studies find that those with lower quinine bitterness on the tongue (or elevated whole-mouth quinine related to the release of inhibition) report greater preference for sweet and high-fat foods and alcoholic beverages, dietary behaviors that could support overconsumption of energy and energy imbalances.

Nontasters of PROP and/or those with spatial taste damage may have elevated dietary risk of cardiovascular disease through higher intakes of alcoholic beverages and sodium as well as smoking behaviors. Nontasters perceive more positive (sweet) and less negative (bitter, irritating) sensations from alcoholic beverages; they also report consuming alcoholic beverages most frequently. Alcohol intake and risk of alcoholism have also been linked with *TAS2R38* and *TAS2R16* receptor genotypes in laboratory-based studies as well as in multicenter family studies of alcoholism. Through unpleasant oral sensations from alcoholic beverages, especially during early exposures to these beverages, when alcohol drinking habits are established, supertasting may confer some protection from overconsumption. The same supertasting protection may be observed for smoking: Chronic smokers are more likely to be nontasters. Preliminary evidence also links nontasters, particularly male nontasters, to greater risk of higher sodium intakes. Compared with supertasters, nontasters perceive concentrated NaCl as less intense and less unpleasant; they are also more likely to consume high-fat sodium foods and to report adding salt to foods more frequently.

Conclusions

Human variation in taste and oral sensation has implications for dietary and lifestyle behaviors that influence risk of chronic diseases. Taste psychophysics has allowed basic taste research to be translated into implications for overall health and quality of life, which is a promising future for work in chemosensation and neuroscience.

See also: Chemical Senses: Overview; Conditioned Taste Aversion; Flavor Physiology; Human Methods: Psychophysics; Nutrient Sensing: Carbohydrates; Nutrient Sensing: Cellular Metabolism; Retronasal Olfaction; Sensory Aging: Chemical Senses; Taste: Vertebrate Central Pathways; Taste: Vertebrate Taste Bud Physiology.

Further Reading

Bartoshuk LM, Duffy VB, Chapo AK, et al. (2004) From psychophysics to the clinic: Missteps and advances. *Food Quality & Preference* 15(7–8): 617–632.

Bartoshuk LM, Duffy VB, and Miller IJ (1994) PTC/PROP tasting: Anatomy, psychophysics, and sex effects. *Physiology & Behavior* 56(6): 1165–1171.

Bartoshuk LM, Snyder DJ, Grushka M, Berger AM, Duffy VB, and Kveton JF (2005) Taste damage: Previously unsuspected consequences. *Chemical Senses* 30(supplement 1): i218–i219.

Boring EG (1942) *Sensation and Perception in the History of Experimental Psychology.* New York: Appleton-Century.

Dinehart ME, Hayes JE, Bartoshuk LM, Lanier SL, and Duffy VB (2006) Bitter taste markers explain variability in vegetable sweetness, bitterness, and intake. *Physiology & Behavior* 87(2): 304–313.

Duffy VB (2007) Variation in oral sensation: Implications for diet and health. *Current Opinion in Gastroenterology* 23(2): 171–177.

Duffy VB, Peterson J, and Bartoshuk LM (2004) Associations between taste genetics, oral sensations and alcohol intake. *Physiology & Behavior* 82(2–3): 435–445.

Fischer R, Griffin F, and Rockey MA (1966) Gustatory chemoreception in man: Multidisciplinary aspects and perspectives. *Perspectives in Biology & Medicine* 9(4): 549–577.

Green BG and George P (2004) 'Thermal taste' predicts higher responsiveness to chemical taste and flavor. *Chemical Senses* 29(7): 617–628.

Green BG, Shaffer GS, and Gilmore MM (1993) Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. *Chemical Senses* 18(6): 683–702.

Lindemann B, Ogiwara Y, and Ninomiya Y (2002) The discovery of umami. *Chemical Senses* 27(9): 843–844.

Moskowitz H (2004) From psychophysics to the world: Data acquired, lessons learned. *Food Quality & Preference* 15(7–8): 633–644.

Prutkin J, Fisher EM, Etter L, et al. (2000) Genetic variation and inferences about perceived taste intensity in mice and men. *Physiology & Behavior* 69(1–2): 161–173.

Stevens DA, Smith RF, and Lawless HT (2006) Multidimensional scaling of ferrous sulfate and basic tastes. *Physiology & Behavior* 87(2): 272–279.

Stevens SS (1969) Sensory scales of taste intensity. *Perception & Psychophysics* 6: 302–308.

Tepper BJ and Ullrich NV (2002) Influence of genetic taste sensitivity to 6-n-propylthiouracil (PROP), dietary restraint and disinhibition on body mass index in middle-aged women. *Physiology & Behavior* 75(3): 305–312.

Taste: Vertebrate Taste Bud Physiology

S D Roper, University of Miami, Miami, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction and Overview

Our sense of taste guides our consumption of nutrients, prevents us from ingesting deleterious substances, and helps our bodies maintain a Na^+ and pH balance. Also, and what is sometimes overlooked, taste initiates the physiological processes that are necessary for digesting and absorbing food efficiently; that is, taste stimulates 'cephalic phase digestion.' The sense of taste also conveys hedonic qualities and, as such, contributes to an individual's overall feeling of well-being and comfort. These gustatory functions are carried out by 5000 to 10 000 peripheral sensory organs – taste buds – in the oral cavity. This article reviews how sensory signals are initiated in taste buds, how taste information is processed in taste buds, and how taste information ascends from peripheral receptor cells to higher cortical centers where taste perceptions of sweetness, sourness, bitterness, saltiness, and umami arise.

Taste Stimuli

Sugars elicit sweet taste and identify foods rich in carbohydrates. (Certain proteins, as well as Pb^{2+} salts, also evoke sweet taste.) Although we now know a good deal about how sugars certain other sweet-tasting compounds stimulate taste receptors, the biological significance of detecting sweet-tasting proteins and lead salts is obscure. The human detection threshold for sucrose to stimulate sweet taste receptors is $\sim$1 mmol^{-1}, but there is great individual variability and age dependence. This relatively high sucrose concentration (as far as ligand–receptor activation is concerned when compared with, e.g., hormone and transmitter receptors) ensures that sweet taste is generated only by nutritionally significant amounts of sugar.

Sodium (and lithium) salts evoke salty taste. Certain potassium salts also stimulate salty taste, albeit poorly, and have been used as salt substitutes to reduce dietary Na^+ intake. Saltiness measures the mineral content of foods and controls the consumption of Na^+, the predominant ion helping maintain the body's fluid volume and blood pressure. The prototypic salty taste is table salt (NaCl), which has a threshold detection of about 1 mmol^{-1}. Many commercially available soups contain NaCl at $\sim$150 mmol^{-1}, a concentration which tastes mildly salty to many individuals.

Sour taste is generated by acids and presumably is important in regulating dietary intake of H^+. Restricting the consumption of protons aids in maintaining the body's pH balance and ensures that the body's intracellular and extracellular buffering systems are not overly taxed. Acid taste, which normally is aversive, also protects against consuming foods that have become rancid and sour. Sourness has a threshold of about pH 4 (e.g., 0.7 mmol^{-1} acetic acid). Commercial vinegar ($\geq$700 mmol^{-1} acetic acid) has a pH of $\sim$2.5. However, the proton concentration (i.e., pH) is not the sole determining factor for sour.

Bitter is mostly an aversive taste and is elicited by an enormous variety of compounds with diverse chemical structures. Many bitter compounds are toxic, and it is believed that bitter taste functions to warn against consuming potentially harmful substances. Newborns innately reject bitter compounds, but later in life many individuals gain an acceptance for certain bitter-tasting compounds, such as caffeine and quinine. Bitterness can be evoked by micromolar concentrations of some compounds, such as strychnine and quinine (e.g., the detection threshold of quinine is 1.4 μmol^{-1}). This (relatively) low threshold ensures that potentially toxic compounds will be detected at concentrations that would be harmful if ingested.

Monosodium glutamate (MSG) and certain other amino acids, especially when paired with the nucleotide 5′ inosine monophosphate or 5′ guanine monophosphate, stimulate a taste named umami, now recognized as a fifth taste quality independent of sweet, sour, salty, or bitter. The detection threshold for MSG is $\sim$10 mmol^{-1}. Umami is a preferred taste, and MSG is a common additive to foods for animal and human consumption. MSG occurs naturally in many foods and is particularly abundant in protein-rich foods, notably cheeses and meats.

There are likely taste qualities in addition to the five above, such as fatty, astringent, spicy, and others, but it is unclear how much of these perceptions is due to their somatosensory components mediated by trigeminal activation in contrast to gustation *per se*. Sweet, salty, sour, bitter, and umami are commonly accepted by the majority of chemosensory neuroscientists as basic tastes.

Taste Buds

There may be as many as 5000 taste buds in most individuals, although approximately 25% of the population may have twice this number. These latter individuals are the so-called supertasters, people who are genetically endowed with an unusually large number of taste buds and consequently have a heightened sense of taste. (A very crude analogy might be made between

the gustatory sensory epithelium of supertasters and the olfactory epithelium of canines, insofar as canines possess a keen sense of smell largely due to their extensive, enfolded olfactory epithelium – and thus the increased number of olfactory sensory receptor cells – within their noses.)

Taste buds are situated in the oral epithelium, with the majority in the lingual epithelium. Taste buds reside in fungiform papillae on the anterior tongue and in foliate and circumvallate papillae on the posterior tongue. Taste buds are also found on the soft palate and the epiglottis. Taste buds are onion- or pear-shaped collections of elongate cells surrounded by stratified squamous epithelium. Sapid stimuli reach the taste bud through a narrow taste pore into which the apical chemosensitive tips of gustatory sensory receptor cells extend.

Each taste bud is a separate sensory organ populated with up to 100 cells, some of which are sensory cells and others of which are believed to have other functions related to taste. Specifically, recent experiments have shown that perhaps only 20% of the cells within a typical taste bud are sensory receptor cells for sweet, bitter, or umami (the taste of certain amino acids). A different population of cells, representing a smaller fraction of cells in taste buds appear to respond to sour (acid) stimuli and perhaps to salty tastes. (Taste bud cells that transduce salty stimuli remain to be identified confidently.) The remainder of the taste bud is populated with supporting and stem cells. Historically, various nomenclatures for taste bud cells have been proposed – 'dark cells' and 'light cells,' from an early era of light microscopy of taste buds, and 'type I, II, III, and IV cells,' on the basis of ultrastructural characteristics observed in electron microscopic analyses of taste buds. Presently, research using molecular and cellular biological approaches is focusing on determining which proteins different types of taste bud cells express. The goal of this research is to identify functional classes of taste bud cells that may correlate with the morphotypes recognized by light and electron microscopy. Recent studies indicate that there is a group of cells (corresponding approximately to type I taste bud cells) that expresses proteins such as nucleoside triphosphate diphosphohydrolase-2 (an ectoadenosine triphosphatase (ecto-ATPase)) and the glutamate aspartate transporter (GLAST). These cells are believed to be supporting cells. There is another group of cells (corresponding approximately to type II taste bud cells) that expresses G-protein-coupled receptors (GPCRs) involved in sweet, bitter, and umami taste. These cells also express downstream partners and effectors (G-proteins, phospholipase Cβ2, inositol triphosphate (IP3) R3 receptors, and transient receptor potential (TRP) M5 channels; see the section titled 'Intracellular cascades downstream of taste receptors, cyclic adenosine monophosphate and Ca^{2+}') for taste GPCRs. Presumably these cells are the primary gustatory sensory receptor cells for sweet, bitter, and umami. Finally, there is subset of cells (corresponding approximately to type III) that express proteins associated with synapses, such as SNAP25 and NCAM (**Figures 1 and 2**). These cells are the only cells that have been shown to form synapses with primary afferent fibers, as determined by electron microscopy. Many of these type III presynaptic cells are serotonergic; an evolving concept is that this group subserves an aminergic synaptic pathway in taste buds: type III cells are also the taste cells that respond to sour(acid) taste.

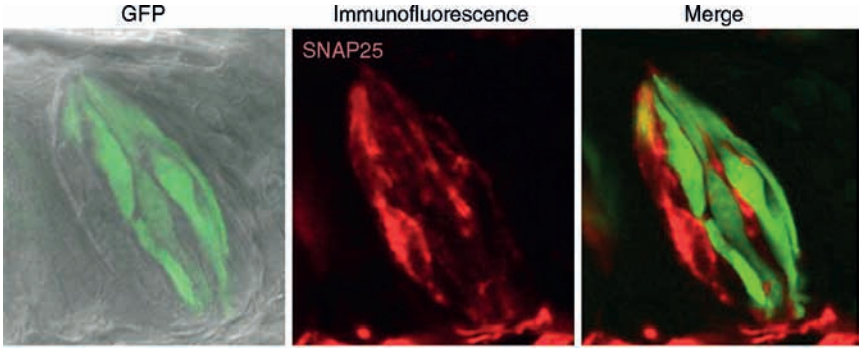

Figure 1 Taste buds contain different populations of cells. This figure is a micrograph of a taste bud in a thin section of the vallate papilla of a mouse. (Left) Gustatory sensory receptor cells express phospholipase Cβ2 (PLCβ2), a downstream effector for taste G-protein-coupled receptors, shown here as green fluorescence. The fluorescence was produced by the expression of green fluorescent protein driven by the PLCβ2 promoter in transgenic mice. (Middle) Other cells in the taste bud form synapses with afferent fibers and express synaptic proteins such as synaptosomal-associated protein 25 (SNAP25), shown here as red immunofluorescence (same taste bud photographed with excitation–emission filters appropriate for the red immunofluorescence of SNAP25). (Right) A merged micrograph of this taste bud, showing the separate populations of gustatory sensory receptor cells (green, ~type II cells; see text and **Figure 2**) and synaptic cells (red, ~type III). Modified from DeFazio RA, Dvoryanchikov G, Marugama Y, et al. (2006) Seperate populations of receptor cells and presynaptic cells in mouse taste buds. *Journal of Neuroscience* 26: 3971–3980, with permission from Society for Neuroscience.

Nerves that Innervate Taste Buds

Taste buds are innervated by primary afferent fibers originating in the sensory ganglia associated with cranial nerves VII, IX, and X (Table 1). It has been claimed that the different regions of the oral cavity, and in particular the tongue, convey different taste qualities (sweet, bitter, etc.), but this notion stems from a mistranslation of the original findings into English and subsequent incorporation and replication in textbooks. There is no well-defined 'tongue map.' Instead, there is a tendency for the anterior tongue to respond better to certain tastes, such as sweet, and the posterior region to others, such as bitter, but any given area on the tongue can and does respond to sweet, sour, salty, and bitter (the situation for umami remains to be explored).

Figure 2 Schematic drawing of a taste bud showing type I supporting cells (gray/green), type II gustatory sensory receptor cells (yellow), and a type III synaptic cell (cyan). Nerve fibers (black) are shown passing near type II cells where presumed synaptic interactions mediated by adenosine triphosphate (ATP) occur. Afferent fibers come into close contact type II (gustatory sensory receptor) cells but do not form any ultrastructurally recognizable synapses. Instead, conventional synapses are present between the type III cell and a nerve fiber. This synapse may be serotonergic, but definitive evidence about these synapses is presently lacking.

Sensory Afferent Terminals and Information Coding

Sensory afferent fibers running in the gustatory nerves summarized above distribute themselves throughout the oral cavity and penetrate individual taste buds at their bases at the boundary between the epithelium and the lamina propria. Recent studies in mice have shown that axons from relatively few (four or five) sensory ganglion cells innervate a single taste bud, at least for taste buds in fungiform papillae and sensory cells in the geniculate ganglion. Thus, several taste bud cells must converge onto a single sensory afferent fiber insofar as there may be up to 100 cells in a taste bud, of which at least 20% are receptor cells.

As stated earlier, the only synapses that are recognized at the electron microscopic level in taste buds are formed between type III synaptic cells and sensory afferent fibers. It presently is uncertain how type II gustatory sensory receptor cells, which make no morphologically identifiable synapses with sensory afferent fibers, transmit signals to the afferent fibers. ATP has been proposed as a principal neurotransmitter between taste buds and sensory afferent terminals. The best evidence is that type II gustatory sensory receptor cells secrete ATP via unconventional, nonvesicular synaptic mechanisms and that ATP stimulates postsynaptic P2X2 and P2X3 receptors on primary sensory afferent fibers. Such synaptic interactions would not necessarily require recognizable structures, thus explaining the lack of synapses observed in electron micrographs of taste buds. ATP secreted from type II cells also appears to stimulate nearby type III synaptic cells within the taste bud and excites them to release serotonin (Figure 3). The function of these latter cells is under scrutiny, and the role of biogenic aminergic neurotransmitters in information flow within taste buds is currently being investigated.

Table 1 Nerves innervating taste buds

Cranial nerve	*Cranial ganglion*	*Name of nerve innervating taste buds*	*Region innervated*	*Name(s) of papillae innervated*
VII (Facial)	Geniculate ganglion	Chorda tympani	Anterior tongue (fungiform); lateral tongue (foliate)	Fungiform and foliate papillae
IX (Glossopharyngeal)	Inferior (petrosal) glossopharyngeal ganglion	Greater petrosal Glossopharyngeal	Soft palate Posterior tongue	Palate Circumvalate papillae
X (Vagus)	Inferior (nodose) vagal ganglion	Vagus	Upper pharynx Epiglottis, larynx, esophagus	

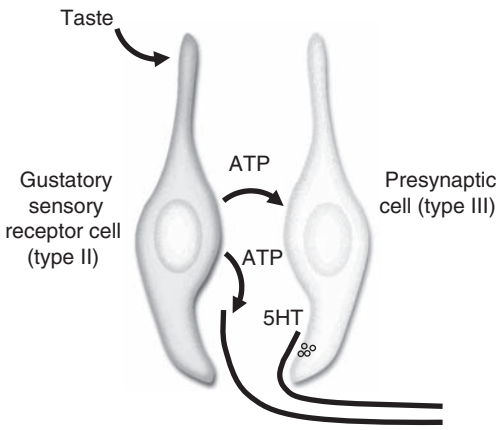

Figure 3 Model for information flow in the taste bud. Gustatory sensory receptor (type II) cells express G-protein-coupled receptors (GPCRs) for sweet, bitter, and umami tastants, and secrete adenosine triphosphate (ATP). ATP stimulates sensory afferent fibers expressing P2X2 and P2X3 receptors. ATP secreted by receptor cells also stimulates P2Y receptors on adjacent presynaptic (type III) cells. Presynaptic (type III) cells form synapses with afferent fibers. Many, but not all, type III cells are serotonergic and release serotonin (5HT) in response to ATP secreted by receptor cells.

The lack of discrete synaptic connections between gustatory receptor cells and sensory afferent fibers may help explain why afferent fibers are rarely highly 'tuned' to specific taste qualities such as sweet, bitter, and so forth. That is, when the electrical activity in single afferent fibers has been recorded during taste stimulation in experimental animals, the usual finding is that individual fibers often respond to two or more taste qualities, such as sweet and salty, with one being predominant. Thus, in this case, the fiber would respond robustly, but not exclusively, to sugars. Such a recording might arise from a sensory afferent that within the taste bud courses near gustatory receptor cells that express sweet receptors as well as near a salt-sensitive taste cell. If both cells secrete neurotransmitter and there are no discrete, point-to-point synapses between the cell and the fiber, the afferent fiber will be excited by activation of the sweet-sensitive taste cell, the salt-sensitive taste cell, or both, with the intensity of the fiber's activation governed by its proximity to the sites of transmitter secretion.

Such a lack of narrow tuning in individual sensory afferent fibers has given rise to the concept that information about taste quality (sweet, sour, salty, etc.) is encoded by the concurrent excitation of numerous sensory afferent fibers. Information encoded by activity in any given single fiber would be ambiguous because a single fiber can be excited by more than one taste stimulus. However, when activity of several fibers is integrated – so-called combinatorial coding or cross-fiber coding – coherent information would emerge. Most investigators believe that this is how taste information is coded and transmitted to the brain, in contrast to the notion that there are highly-tuned, individual, specific lines for sweet, sour, bitter, and so forth – a notion called the 'labeled line' theory of coding.

Molecular Receptors for Taste Stimuli

The molecular basis for detecting taste stimuli has been determined, in part, for sweet, bitter, and umami, but details remain elusive for sour and salty. Sweet, bitter, and umami are transduced by GPCRs embedded in the apical chemosensitive tips of gustatory sensory receptor cells (~type II cells). GPCRs for these tastes have been identified and are unique to gustatory sensory receptor cells. The GPCRs for sweet include the heteromeric dimer T1R2 plus T1R3. There may be additional sweet receptors, as suggested by ongoing investigations. Umami is transduced by the dimer T1R1 plus T1R3 and by additional receptors that have not been completely characterized. These additional receptors may include GPCRs that are modified synaptic metabotropic glutamate receptors (mGluRs), specifically a modified mGluR4 and mGluR1. Molecular receptors for bitter are named T2Rs. A family of approximately 25 T2R genes transduces bitter substances in the human. Current research is aimed at identifying the ligands for each of the T2R GPCRs. It is highly likely that there are additional, less-explored molecular mechanisms for bitter taste, including direct actions of bitter ligands on ion channels.

The molecular transducers for salty and sour (acid) tastes have eluded identification to date. There is considerable evidence that ion channels, not GPCRs, subserve these qualities and that the amiloride-sensitive epithelial sodium channel transduces Na^+ taste, at least in rodents. Even less is known about mechanisms for acid taste. Recent research has implicated TRP-like channels, specifically PKD1L3 and PKD2L1, in contributing to sour taste transduction but the evidence is not yet conclusive. It is important to note that the proximate stimulus for sour (acid) taste appears to be an increase in intracellular (not extracellular) protons, that is, intracellular acidification. Accordingly, solutions of organic acids such as acetic and citric acids are more intensely sour than are mineral acids such as hydrochloric acid when all are presented at the same pH. The protonated (undissociated) molecules of organic acids (HA) readily penetrate the cell membrane and dissociate inside the cell to lower the cytosolic pH. In contrast, the extracellular protons of a fully dissociated acid

($HCl \rightarrow H^+ + Cl^-$) only penetrate the plasma membrane (Figure 4). Intracellular acidification occurs throughout the taste bud and possibly the entire lingual epithelium during sour stimulation, but only a subpopulation of acid-sensitive taste sensory cells respond to this intracellular pH shift by activating transduction mechanisms specific for sour taste. Sour-responsive taste cells have recently been identified as type III taste cells, but the details of how acid taste is transduce by these cells is not yet known. It is not known what characterizes this population of acid-sensitive sensory cells or how they signal sour taste.

The principal receptor for sucrose and other sugars appears to be a dimer of class C GPCRs: T1R2 plus T1R3. Class C GPCRs are characterized by lengthy extracellular N-termini which form a hinged binding pocket, akin to the carnivorous leaves on the Venus fly trap plant; hence the name for this portion of the receptor is the Venus fly trap domain. This motif seems to be the primary binding site for sucrose and glucose. However, there are at least two other binding pockets elsewhere in the T1R2–T1R3 dimer, including a site for the artificial sweetener cyclamate, and another for the sweet protein brazzein (Figure 5). The cyclamate pocket is tucked between transmembrane (TM) regions 3, 5, and 6 of T1R3 and situated within the TM depths of the molecule, as is found for the binding pocket of other GPCRs, such as the β adrenergic receptor. The binding site for brazzein is found in a cysteine-rich region near where the large extracellular N-terminus joins the first TM domain of T1R3. Considerably less is known about the structural details of GPCRs for umami and bitter.

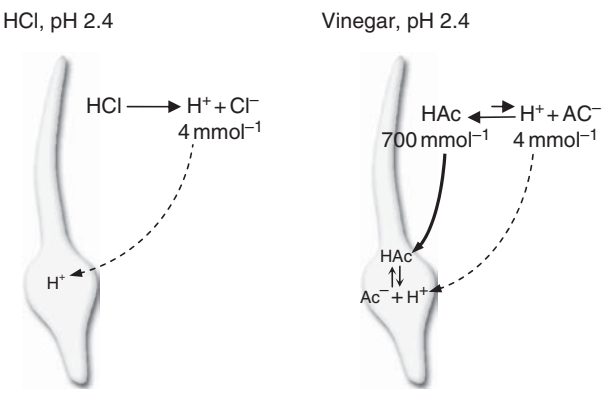

Figure 4 Proposed mechanism for sour (acid) taste. A sour-sensitive cell is shown responding to a mineral acid (HCl), left, and to an organic acid (acetic acid), right, both at a pH of 2.4. Left, HCl is virtually totally dissociated into protons and Cl^- in solution, and at pH 2.4, concentration of $H^+ \sim 4\,mmol^{-1}$. Protons weakly permeate the plasma membrane (dotted line) and acidify the cytosol, which is believed to be the proximate stimulus for sour taste. Right, acetic acid (HAc) at $0.7\,mol^{-1}$ (i.e., about that of commercial vinegar) has a pH of ~2.4, and thus, concentration of H^+ in solution also is $\sim 4\,mmol^{-1}$. However, at pH 2.4, >99% of acetic acid is undissociated ($Ka = 1.76 \times 10^{-5}$), and thus the concentration of the neutral, protonated species HAc $\sim 700\,mmol^{-1}$. The plasma membrane is relatively permeable to the uncharged protonated acid (i.e., HAc; solid line), which enters the cell and dissociates inside to acidify the cytosol. This may explain why organic acids such as acetic acid are more sour than mineral acids such as HCl when both are presented at the same pH. The relatively abundant, uncharged, protonated form of the organic acid 'transports' protons into the cell interior and effectively acidifies the cytosol.

Intracellular Cascades Downstream of Taste Receptors, Cyclic Adenosine Monophosphate and Ca^{2+}

Two general schemes exist for the consequence of activating taste GPCRs (Figure 6). Early studies focused on the second messenger, cyclic adenosine monophosphate (cAMP), and G-proteins such as the taste-specific G_α protein, gustducin, that directly modulate intracellular cAMP levels. Bitter and umami stimuli reduce intracellular cAMP, possibly via a G_α gustducin-activated phosphodiesterase, by analogy with the actions of the closely related G_α protein found in photoreceptors, transducin. In contrast, sweet stimuli increase cAMP, likely via $G_{\alpha s}$-coupled mechanisms. However, greater attention has been drawn in the recent past to IP3-mediated increases in another second messenger, calcium, inside taste receptor cells. Sweet, bitter, and umami stimuli all elicit phospholipase C (PLC) → IP3-mediated Ca^{2+} release from intracellular stores in the gustatory sensory receptor cells. Downstream of Ca^{2+} release appears to be the TRP channel TRPM5. All the relevant molecular participants – G_α gustducin, PLCβ2, IP3, IP3 receptors, and TRPM5 – have been identified in type II taste gustatory sensory receptor cells. It has been speculated that the intracellular release of Ca^{2+} during taste receptor activation (by way of PLCβ2, IP3, and IP3 receptors) opens TRPM5, allowing influx of cations and consequently membrane depolarization. Genetic knockout of many of the genes for these transducer elements, namely G_α gustducin, PLC β2, and TRPM5, profoundly reduces, but does not eliminate, taste sensitivity. How cAMP, Ca^{2+}, and the downstream cascades interact, if at all, is unknown.

Central Processing of Taste Signals

Signals from taste buds enter the hindbrain and reach the nucleus of the solitary tract (NTS). From the NTS, information is relayed to the thalamus and from there ascends to the primary cortex for gustation, located in the insular cortex. In some species, such as rodents,

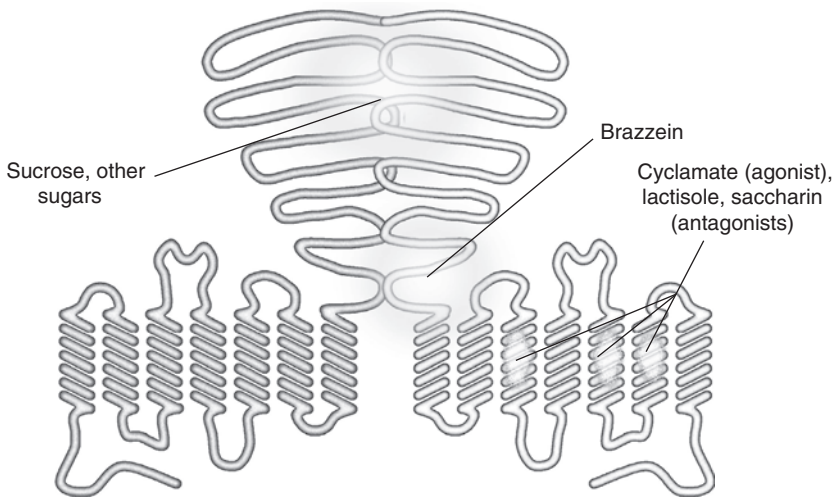

Figure 5 Schematic drawing of the sweet receptor, a G-protein-coupled receptor dimer T1R2 + T1R3, showing the multiple ligand-binding sites. T1R2 is shown on the left, T1R3 is on the right. Details of the interactions, if any, between the extensive N-termini of T1R2 and T1R3 are not known. The N-termini possess one or more binding pockets for sucrose and other sugars. Another ligand-binding pocket, located in T1R3 near the first transmembrane (TM) region, exists for certain sweet-tasting proteins such as brazzein. A third ligand pocket, embedded in the TM regions of T1R3, is comprised of portions of TM3, 5, and 6. This site binds the artificial sweetener cyclamate but is also the binding site for antagonists such as saccharin (which at high concentrations is a sweet receptor antagonist) and lactisole. Modified from Roper SP (2007) *Pflügers Archiv* 45: 759–776, with permission.

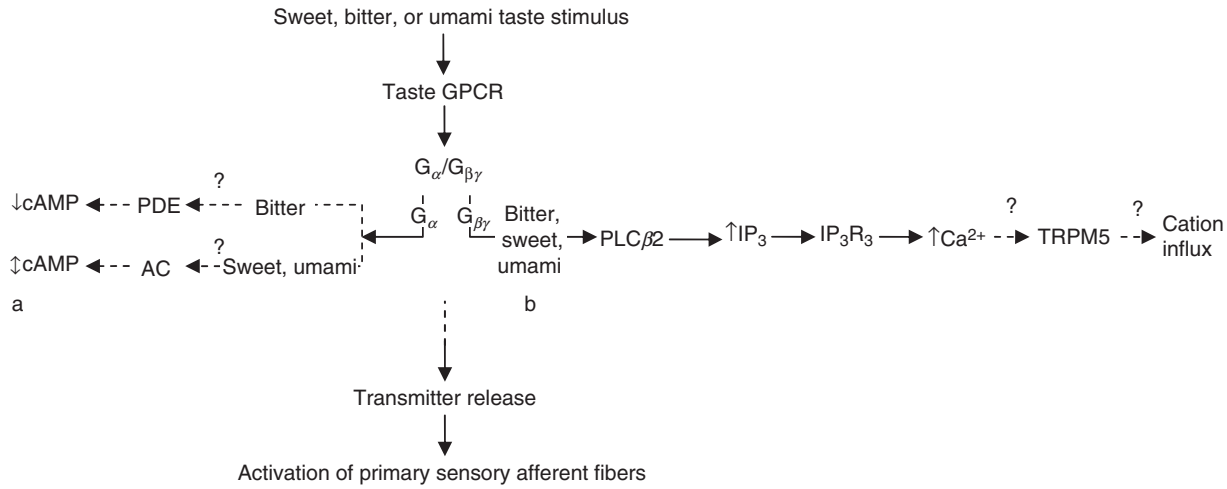

Figure 6 Chemosensory transduction pathways for G-protein-coupled receptor (GPCR)-mediated tastes. Activation of taste GPCRs initiates (a) changes in intracellular cyclic adenosine monophosphate (cAMP), shown on the left branch, and (b) changes in intracellular Ca^{2+}, shown on the right branch. The final steps in the transduction pathways are transmitter release and stimulation of primary sensory afferent pathways. How changes in cAMP and Ca^{2+} lead to transmitter release is currently under investigation. IP3, inositol triphosphate; PDE, phosphodiesterase; PLC, phospholipase C; TRP, transient receptor potential channel.

the parabrachial nucleus (PBN) is interposed in the taste axis between the NTS and the thalamus. At each stage, there is feedback to lower centers as well as feed-forward to the next higher level (Figure 7).

Taste information processed in the insular cortex is transmitted to association cortices, especially the orbitofrontal cortex, where inputs from olfaction (and other sensory modalities) converge, presumably to generate the more complex perceptions of 'flavor' (which contrast to the more basic taste quality sensations, such as sweet, salty, etc.). Studies from functional magnetic resonance imaging (fMRI) and microelectrode recordings suggest that the primary (insular) cortex in taste encodes taste intensity and perhaps taste quality (sweet, sour, salty, etc.) but that the orbitofrontal cortex is more important for interpreting hedonic aspects, especially pleasure, that would guide food consumption. Other association

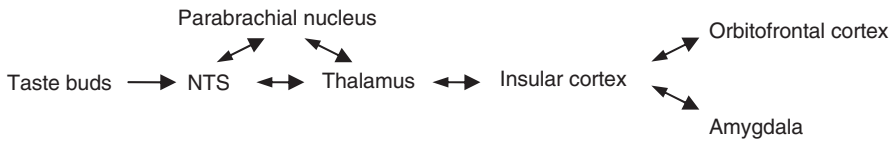

Figure 7 Outline of the information flow from taste buds to the brain in humans. At every level after the taste bud, there is information feed-forward (→) and feedback (←). Whether there is central feedback to the taste buds is not established. NTS, nucleus of the solitary tract.

areas include parts of the limbic system, such as the amygdala, which are important in establishing taste memories and conditioned food aversions.

Researchers are presently investigating specifically how information is encoded at each stage of the taste axis, from peripheral end organs (taste buds) to higher cortical areas. There is no consensus in the field as yet. As discussed above, it is unlikely that there are simple labeled lines of information leaving taste buds. The situation becomes even more complex as taste signals ascend to the cortex. Where it has been measured electrophysiologically, central neurons are increasingly broadly tuned to taste qualities as one ascends the taste axis to the cortex; there are no 'sweet-selective' or 'bitter-selective' central nervous system taste neurons. Measurements of the breadth of tuning, the entropy in the gustatory signal, show increases from the NTS to the thalamus to the cortex.

One concept is that information about sweet, sour, and so forth is coded by the particular combination of primary afferent fibers that are activated. This is akin to combining individual notes on a piano keyboard to form chords that as readily distinguish Schubert from Stravinsky as we distinguish sweet from salty. Individual piano keys taken alone cannot identify a composer any more than the single fibers of a supposed 'labeled line' can identify a taste quality. In higher centers, this concept can be extended: different combinations of central nervous system neurons are excited to represent the different taste qualities. Evidence to date suggests that if this is the case, the neurons that underlie a particular taste quality do not form readily distinguishable clusters that can be observed with fMRI. There is no well-defined cluster of sweet-sensitive neurons that is distinguishable from a 'bitter' cluster. If taste is coded in this manner, the 'grain' of the pattern for each taste quality might be too fine to be resolved with fMRI.

Another, related concept is that the temporal pattern of firing in different groups of neurons encodes sweet, sour, salty, and so forth. A crude analogy would be telegraphic Morse code, with taste information conveyed along neither a labeled line nor a combination of parallel lines but instead as a time-varying pattern of signals. In the case of taste, the time-varying signals would be occurring in thousands of parallel lines, whether primary sensory afferent fibers or tracts of central neurons. Sensory coding in general, and gustatory coding in particular, is an area of current investigation.

See also: Chemical Senses: Overview; Conditioned Taste Aversion; Flavor Physiology; Nutrient Sensing: Carbohydrates; Nutrient Sensing: Cellular Metabolism; Sensory Aging: Chemical Senses; Taste: Vertebrate Central Pathways; Taste: Vertebrate Psychophysics.

Further Reading

Breslin PA and Huang L (2006) Human taste: Peripheral anatomy, taste transduction, and coding. *Advances in Oto-Rhino-Laryngology* 63: 152–190.

DeFazio RA, Dvoryanchikov G, Marugama Y, et al. (2006) Seperate populations of receptor cells and presynaptic cells in mouse taste buds. *Journal of Neuroscience* 26: 3971–3980.

Meyerhof W (2005) Elucidation of mammalian bitter taste. *Reviews of Physiology, Biochemistry and Pharmacology* 154: 37–72.

Roper SD (2007) Signal transduction and information processing in mammalian taste buds. *Pflügers Archiv: European Journal of Physiology* 454(5): 759–776.

Simon SA, de Araujo I, Gutierrez R, and Nicolelis MA (2006) The neural mechanisms of gustation: A distributed processing code. *Nature Reviews Neuroscience* 7: 890–901.

Spector AC and Travers SP (2005) The representation of taste quality in the mammalian nervous system. *Behavioral and Cognitive Neuroscience Reviews* 4: 143–191.

Relevant Website

http://chroma.med.miami.edu – Chroma, see Stephen Roper's faculty page.

Tay–Sachs Disease

E H Kolodny, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Historical Background

In 1881, Warren Tay, a British ophthalmologist, general surgeon, and dermatologist, reported to the Ophthalmology Society of the United Kingdom the case of a 12-month-old infant who had become increasingly weak since the age of 2 or 3 weeks. The mother thought the baby was not as visually alert as other babies. Tay noticed in the macula of each eye a conspicuous, large, somewhat circular white patch containing within its center a brownish red spot, a finding which has become known as a cherry-red spot. An autopsy performed 6 months later showed only a large cavity containing an old blood clot near the left lateral ventricle. Three years later, Tay reported that three children in this original family had by then had a similar condition.

Unaware of Warren Tay's findings, the New York neurologist Bernard Sachs reported in 1887 a case of 'arrested cerebral development' in a girl with progressive neurodegeneration from the age of 2 or 3 months until her death at age 2 years. The ophthalmologist Herman Knaap, having seen the child at 2 or 3 months, described within the 'Fovea centralis' the appearance of a 'cherry red color surrounded by an intense grayish white opacity.' When he saw the child again at age 22 months, the child was totally blind with the 'disk as white as paper with scarcely any trace of blood vessels.' According to Sachs, similar fundal appearances had been reported by Magnus and Goldzieher in 1855. Sachs described the child's hearing as very acute with unusual hyperexcitability to auditory and tactile impressions. His report in 1887 included two drawings by van Giesen showing a distorted shape of small and large pyramidal cells. The same year, another case was described by Wadsworth, and the next year in another case report Kingdon mentioned Jewish parentage.

By 1896, Sachs had seen 19 cases and was aware of its familial nature and its predilection in Jewish children. He gave it the name 'amaurotic family idiocy.' Kingdon and Russell, in their 1896 publication, first used the eponym Tay–Sachs.

Studying the chemistry of brain tissue from children with amaurotic family idiocy, the German biochemist Ernst Klenk in the 1930s isolated sialic acid-containing glycosphingolipids. Because of their high concentration in nerve cells, he termed them 'gangliosides.' The specific ganglioside stored in Tay–Sachs disease (TSD) was identified in 1962 by the Swedish neurochemist Lars Svennerholm as G_{M2} ganglioside, and its chemical structure was published the following year by Makita and Yamakawa.

The early 1960s also ushered in a period of investigation into the pathology and epidemiology of TSD. The group led by Terry and Korey showed, with the aid of the electron microscope, an abundance of multilamellar membraneous granules within the cytoplasm of swollen Tay–Sachs nerve cells. Contemporaneously, Myrianthropoulos and Aronson were documenting the relatively high heterozygote frequency of TSD (1:33–1:45) among Jews of Eastern European origin.

The discovery of acid hydrolase deficiencies in several other lysosomal storage diseases in the mid-1960s fostered the belief that TSD might also result from an enzyme deficiency, possibly in the cleavage of the terminal β-linked *N*-acetylgalactosamine residue. However, assays of hexosaminidase activity in TSD were either normal or showed increased, not decreased activity. In 1968, Sandhoff, Andrae, and Jatzkewitz reported finding a generalized deficiency in β-*N*-acetylhexosaminidase in a non-Jewish patient. This case was unusual in that the patient had storage within visceral tissues of increased amounts of the asialo derivative of G_{M2} ganglioside (G_{A2}) and of another hexosamine-containing glycolipid, globoside. The following year, Okada and O'Brien, using an electrophoretic technique devised by Stirling and Robinson, showed that one of two components of hexosaminidase was missing from the serum of patients with TSD. Whereas the A component was absent, the B component was increased in amount, therefore explaining why the total activity of the two components was normal in Tay–Sachs patients. A small subgroup of non-Jewish patients were later shown to be deficient in both components of hexosaminidase. This variant is now known as Sandhoff disease in recognition of the findings of Sandhoff et al. in their 1968 report. Subsequently, it was recognized that a specific protein cofactor was needed for hexosaminidase A (Hex A) to act on G_{M2} ganglioside. In the very rare AB variant, this cofactor is missing. The term G_{M2} gangliosidosis, introduced by Suzuki and Chen in 1967, is now used to refer to all diseases in which G_{M2} ganglioside accumulates because of a defect in its hydrolysis.

In 1970, Schneck and colleagues reported the first prenatal diagnosis for TSD. With the possibility of

in utero detection and the development of a test for carriers of the Tay–Sachs trait, prospective screening to prevent TSD became feasible.

During the 1970s, the subunit structure, chemical composition, substrate specificity, and chromosomal localization of the hexosaminidases were elucidated. Investigations in the 1980s revealed the steps involved in their intracellular processing and targeting to lysosomes, and by the end of the decade the genes encoding both the α and β subunits of hexosaminidase and the G_{M2} activator protein were isolated and sequenced. Many different molecular defects causing TSD and its variants have now been described, animal models have been devised, and various therapeutic approaches are being examined.

Clinical Features

The infant with classical TSD may appear normal at birth but within a few weeks may become less visually attentive, hypotonic, and easily startled by sound, light, or touch stimuli. By age 6–8 months, developmental delay becomes obvious when the infant fails to consistently roll over and is unable to sit. At an early stage, fundiscopic examinations of the retina will reveal a whitish halo surrounding the retina due to lipid deposition in the perifovial retinal bipolar ganglion cells.

By 1 year of age, there is a marked reduction in purposeful movements and the child becomes spastic and increasingly lethargic. Vision deteriorates, although the pupillary light reaction may persist, and seizures become more frequent. These may be short, subtle complex–partial or absence attacks and are separate from the episodes of extensor rigidity associated with the exaggerated startle reflex. Bouts of unmotivated laughter also occur. Macrocephaly and optic atrophy supervene and by age 2 years the child is in a vegetative state and requires constant nursing care. Common clinical signs at this stage are feeding difficulty, drooling, microaspiration leading to bronchopneumonia, constipation, and episodes of autonomic dysfunction with unexplained fevers and spells of rubor, pallor, or cyanosis and irregular breathing.

In the advanced stage, the cherry-red appearance of the macula tends to disappear as retinal ganglion cells degenerate, and seizures diminish in frequency and severity. There is progressive cachexia and recurrent pneumonia, with death occurring between age 3 and 5 years in most cases.

The electroencephalogram progresses from high-voltage slowing to multifocal spikes, then to a more diffuse slow wave pattern, and finally to a low-voltage tracing. The electroretinogram is normal but visual-evoked potentials are abnormal early in the disease, and the signal from the visual cortex is lost by 12–16 months.

Neuroimaging studies disclose on computerized tomography low density in the basal ganglia and cerebral white matter, and on T2-weighted magnetic resonance imaging (MRI) high signal intensity in these same areas. The T2 signal in the thalamus may be decreased. The caudate nuclei are enlarged, and in a late stage of the disease cerebral atrophy is found. The pattern on magnetic resonance spectroscopy of increases in myoinositol and a decrease in *N*-acetylaspartate is consistent with demyelination of central white matter, axonal injury, and neuronal cell loss.

Pathology

The brain of an infant with classical TSD surviving beyond 2 years of age is increased in weight and size compared to normal but shows generalized atrophy due to a marked reduction in neurons and loss of central white matter. Widespread lipid storage within surviving neurons of both the central and peripheral nervous systems causes these cells to enlarge, lose Nissl substance, and displace the nucleus to the periphery of the neuronal perikarya. Their cell bodies stain positive with the periodic acid Schiff stain and luxol fast blue but stain negative after treatment with lipid solvents and paraffin embedding.

Electron microscopic examination reveals the ballooned neuronal perikarya to be filled with discrete concentrically arranged multilamellated cytoplasmic bodies (MCBs) approximately 1 μm in diameter contained within lysosomes and consisting of cholesterol, phospholipid, G_{M2} ganglioside, and protein. Cortical pyramidal cells lose their angular shape and develop fusiform enlargements of the proximal axonal segment termed 'meganeurites' that contain spine-like processes. Aberrant synaptic contacts are created which are believed to cause neuronal hyperexcitability. Such ectopic dendritogenesis appears to be driven by an excess of G_{M2} ganglioside which occupies a pivotal role as a regulator of dendritic sprouting in cortical pyramidal neurons.

The cerebellum is also atrophic and in older patients there is a marked reduction in cerebellar Purkinje and granule cells. Antler-like expansions are notable at bifurcations of Purkinje cell dendrites.

Reduction in central myelin is due to both arrested myelinogenesis and loss of cortical neurons. Gliosis is present in the cortical gray matter as well as in cerebral white matter.

Gene expression profiles in TSD brain indicate a decrease in the expression of neuron-specific genes

and overexpression of genes related to the inflammatory process. This suggests that macrophage/microglia and astrocyte activation contribute to neuronal apoptosis in TSD and trigger a reactive gliosis.

Within the TSD fetus aborted at midterm, abnormal storage can be shown in phylogenetically older areas of the central nervous system (CNS), including anterior horn cells of the spinal cord, dorsal root ganglia, motor nuclei of the midbrain, and thalamic neurons. Lamellar membranous bodies are present but less compact than in affected infants.

Biochemistry

As much as 80% of the ganglioside content of brain in TSD is G_{M2} ganglioside, which normally comprises less than 5% of total brain gangliosides. Svennerholm first reported this sialic acid-containing glycolipid in 1962, and its structure was confirmed in 1965 by Ledeen and Salsman. It comprises four carbohydrate residues linked to ceramide, including two *N*-acetylated sugars, *N*-acetylgalactosamine and *N*-acetyl neuraminic acid, both linked to a penultimate galactose residue. The primary cause of classical TSD was shown in 1970 to be a deficiency of the acid hydrolase necessary to cleave the *N*-acetylgalactosamine residue from G_{M2} ganglioside.

Molecular Genetics

The hydrolysis of G_{M2} ganglioside requires the concerted action of three gene products: the α and β subunits of β-hexosaminidase A and the G_{M2} activator protein. The α and β subunits are encoded by two evolutionarily related genes, *HEXA* on chromosome 15q23–q24 and *HEXB* encoded on chromosome 5q13, and share a large degree of primary structure homology. The G_{M2} activator protein gene G_{M2A}, encoded on chromosome 5q31.3–33.1, is a much smaller, structurally different glycoprotein.

In 1968, Stirling and Robinson showed that lysosomal β-*N*-acetylhexosaminidase activity could be separated into two forms, an acidic Hex A and a basic Hex B form, each with a different extent of migration in an electrophoretic field. Okada and O'Brien subsequently showed that infants with classical TSD lacked Hex A. Hex A is a heterodimer comprising an α and a β subunit; Hex B is composed of two β subunits. Both enzymes are involved in degradation of glycolipids, glycoprotein-derived oligosaccharides, and glycosaminoglycans, but only Hex A can catalyze the hydrolysis of the β-linked GalNAc residue from G_{M2} ganglioside. Both enzymes will also cleave the artificial fluorogenic substrate 4-methylumbelliferyl-β-*N*-acetylglucosamine (4MUG), whereas the sulfated form of this substrate (4MUGS) is hydrolyzed only by Hex A and Hex S, the latter of which is an unstable α dimer. Because Hex B is heat stable while Hex A is heat labile, the proportion of each in serum, leukocytes, or cultured cells can be determined by assaying enzyme activity before and after heat treatment (usually 50 °C for 2 h).

As of March 2007, 109 mutations had been reported in the α polypeptide of Hex A, of which 97 were associated with TSD. Among Ashkenazi Jews, the most common mutation is a 4-base pair insertion in exon 11 ($+TATC_{1278}$), accounting for approximately 80% of mutated alleles. The second most common in Ashkenazim is a splice junction mutation at the 5′ end of intron 12 (+1 IVS-12) found in another 13% of Tay–Sachs alleles. A 7.6-kb deletion at the 5′ end of the Hex A gene is characteristic of French–Canadian patients with infantile TSD. In the case of these three mutations, enzyme activity is undetectable. Many other Hex A mutations causing TSD are restricted to small populations or single pedigrees.

Two missense mutations, Arg247Trp and Arg249Trp, cause the lack of Hex A activity in apparently healthy individuals when assayed with 4MUG but not in the presence of 4MUGS. The Arg247Trp alteration is present in 32% of non-Jewish enzyme-defined carriers, but only 2% of Jewish enzyme-defined carriers. The Arg249Trp mutation is seen in 4% of non-Jewish carriers but not at all in Jewish carriers. The presence of these pseudodeficiency alleles in the non-Jewish population explains the apparent discrepancy between the carrier frequency in this population based on disease incidence (1/300) and that observed by enzyme assay (1/167).

Five mutations have been found in G_{M2A}, the G_{M2} ganglioside activator protein gene. These mutations produce very little protein and therefore are associated with the severe TSD phenotype known as the AB variant. When assayed with artificial substrates, normal levels of Hex A and Hex B are obtained, but no activity results when assays are done using G_{M2} ganglioside as the substrate.

Clinical Heterogeneity

TSD is but one of a group of autosomal recessive lysosomal storage diseases known as the G_{M2} gangliosidoses. The development of routine assays for Hex A and Hex B activity and the spread of large-scale population screening programs for carriers of TSD have revealed a spectrum of phenotypic variants, including chronic and later onset forms. In general, the higher the residual enzyme activity, the later the onset of the disease and the milder the clinical manifestations.

Sandhoff Disease

Sandhoff disease is distinguished by the absence of both Hex A and B activity due to defects in the *HEXB* gene encoding the β subunit. Because this polypeptide is a component of both enzymes, total hexosaminidase activity is severely reduced, with only the rare Hex S ($\alpha\alpha$ dimer) present. A total of 32 mutations have been reported in the *HEXB* gene.

The clinical course of infantile Sandhoff disease is similar to that of TSD. In patients age 12–18 months, brain MRI reveals mild cortical atrophy, a thin corpus callosum, and abnormal signal intensities in the caudate, globus pallidum, putamen, cerebellum, and brain stem. Due to the additional storage of globoside and other hexosamine-containing glycolipids and oligosaccharides within systemic organs, foam cells may appear in the bone marrow and slight hepatomegaly may be recognized.

These signs and the presence of an excess of hexosamine-containing glycoprotein fragments within the urine of patients with Sandhoff disease help distinguish infantile Sandhoff disease from classical TSD.

AB Variant

Few clinical reports are available that describe the phenotypic appearance and pathology of patients with the AB variant. At least two patients with American Indian ancestry are known. The clinical manifestations and brain pathology are similar to those of classical TSD. G_{M2} ganglioside concentration in the cerebrospinal fluid (CSF) is elevated and MCBs can be found within a conjunctival biopsy. The diagnosis is confirmed by finding G_{M2} activator protein deficiency and mutations within the G_{M2A} gene.

B_1 Variant

The B_1 variant of G_{M2} gangliosidosis is due to mutations that affect the active site in the α subunit without affecting its ability to associate with the β subunit. Therefore, normal catalytic activity is obtained using 4MUG as the substrate, but hydrolysis of the sulfated derivative (4MUGS) and G_{M2} ganglioside is deficient. This is because the catalytic site of the β subunit of Hex A can act on 4MUG but the α subunit catalytic site is required for cleavage of 4MUGS and G_{M2} ganglioside. Carriers of a B_1 variant allele will demonstrate normal Hex activity with 4MUG and half-normal activity with 4MUGS. Compound heterozygotes for a B_1 allele and a null mutation will appear as carrier with 4MUG but deficient in Hex A activity with 4MUGS as the substrate.

Infantile, juvenile, and late-onset forms of G_{M2} gangliosidosis have been found with a B_1 allele. The Arg178His mutation, also known as the DN allele, is the most frequent cause of the B_1 form of disease and is especially common among patients of Portuguese origin. In one series, onset appeared between 1.5 and 2.5 years and in another between ages 3 and 7 years. The course was rapid in some and slow in others, allowing survival into the second decade. Manifestations include ataxia and pyramidal signs leading to gait difficulty and dysarthria progressing to anarthria. The brain MRI findings are similar to those in other variants, with basal ganglia injury and severe demyelination in the infantile form progressing to cerebellar atrophy and no supratentorial abnormalities in the adult form.

Juvenile/Subacute G_{M2} Gangliosidosis

Juvenile or subacute onset of G_{M2} gangliosidosis has been associated with both α subunit (TSD variant) and β subunit (Sandhoff variant) mutations. A review of 134 previously described patients reported the onset at 5.4 ± 4.0 years, with gait disturbances (58.2%) as the most common symptom at onset. Other symptoms were speech problems (37.3%), incoordination (36.5%), intellectual impairment (29.1%), and behavior or psychiatric disturbances (11.9%). Optic atrophy occurred in 15.4% of patients, and 4 patients (4.6%) had a macular cherry-red spot. Dysphagia and sphincter incontinence occurred earlier in the TSD variants. Spasticity, rigidity, seizures, and dementia have also been observed in juvenile-onset cases. Mean survival was 19.0 ± 2.1 years in the 40% of patients whose status was known. The most common cause of death was respiratory tract infection (43.4%).

Late-Onset (Chronic Adult) G_{M2} Gangliosidosis

Patients with late-onset chronic or adult G_{M2} gangliosidosis retain a small amount of residual enzyme activity ($\sim$5% of normal) due to a mild mutation in compound heterozygosity with a more severe nonexpressing mutation. The prototypic example is the Gly269Ser mutation in exon 7 of the *HEXA* gene which is most often found in Ashkenazi Jews. Of the TSD carriers in this population, 2% carry this mutation.

Phenotypically, patients with the Gly269Ser mutation may present in childhood or adolescence with subtle signs and symptoms, including learning disabilities, clumsiness, and/or dysarthria. Proximal muscle weakness develops in adolescence with progression to a slightly broad-based ataxic gait in the third or fourth decade. The onset of overt disease may be heralded by an acute psychiatric disturbance with anxiety, depression, or a psychotic episode. Fasciculations, muscle cramps, easy startle, and fine tremor of the outstretched hands become evident. Saccadic pursuit of the eyes is slowed, muscle atrophy develops,

knee reflexes are lost, and Babinski signs appear. Autonomic disturbances develop, including painful ankle edema and cool, purple-colored feet. Brain MRI studies reveal marked cerebellar atrophy. Neuropsychological testing demonstrates impairment in executive functioning and memory, although dementia is not part of the usual disease pattern. This cohort of patients responds poorly and may worsen neurologically if treated with tricyclic antidepressants or major neurolepties such as haloperidol, risperidone, or chlorpromazine.

Late-onset patients with other mutations may present with seizures, dementia, and a rapidly progressive course. Patients in their 50s or older with either *HEXA* or *HEXB* mutations have also been reported with esophageal and gastrointestinal motility problems (achalasia and diarrhea) and other evidence of autonomic dysfunction.

Diagnosis and Carrier Testing

Formerly the diagnosis of TSD was based on the typical clinical presentation in combination with histologic evidence of lipid storage (i.e., rectal biopsy for myenteric plexus neurons). Patients with the Sandhoff variant will often have vacuolated lymphocytes and monocytes in their peripheral smear and foam cells in the bone marrow. Moreover, chromatography of their urine will reveal an excess of *N*-acetylglucosaminyl oligosaccharides.

Enzymatic assays utilizing synthetic substrates are now in widespread use for detecting Hex A and Hex B deficiency. Assays employing the fluorogenic compound, 4MUG, can be used with readily available sources of enzyme, such as serum, leukocytes, tears, cultured skin fibroblasts, amniocytes, and chorionic villus tissue, to detect Hex A and B after electrophoretic separation or before and after selective thermal denaturation of Hex A. Under specific conditions, Hex B remains heat stable, whereas Hex A is heat labile. The amount of Hex A present is obtained by subtracting the activity obtained after heat treatment of the sample, typically 2 h at 50 °C, from the total activity prior to thermal exposure. Typically, assays on specimens from carriers of mutations in the α subunit gene will show half the normal amount of Hex A activity, whereas carriers of β subunit gene mutations will have reduced activity of both Hex A and Hex B.

Certain exceptions may cause problems in rare individuals. Carriers as well as homozygyous affected individuals with the B_1 variant allele will appear normal or low normal in routine assays with 4MUG as substrate. Such individuals can be differentiated using the Hex A-specific substrate 4MUGS. Another problem is pseudodeficiency alleles, which can account for as much as one-third of non-Jewish enzyme-defined carriers. In both of these circumstances, DNA-based mutation analysis is a necessary next step. DNA mutation analysis is also necessary for patients with β subunit mutations causing heat lability of Hex B. These individuals will appear with the standard heat denaturation assay to have an increased fraction of Hex A activity.

Up to 10% of Ashkenazi Jewish individuals may carry the delTG(+)619→G haplotype, which causes low serum Hex B activity, possibly masking a TSD carrier state. In such individuals, further testing of leukocyte enzyme activity is required. Leukocytes, and not serum, should also be employed for patients who are pregnant, on birth control medication, on certain other drugs, or who suffer concurrent illness such as hepatitis. Leukocyte enzyme determinations are also done as a backup whenever serum assays suggest a carrier state or are inconclusive.

Enzyme determinations for G_{M2} activator protein deficiency generally require cultured skin fibroblasts and radioactively labeled G_{M2} ganglioside, which is not routinely available. Therefore, DNA mutation analysis of the G_{M2A} gene may be preferred.

In pregnancies at risk for TSD or another variant of G_{M2} gangliosidosis, DNA mutation analysis performed in advance on DNA of the parents and previously affected patient, if there is one, greatly simplifies diagnosis of the fetus. In the case of non-Ashkenazi individuals in particular, this precaution will rule out the possibility of mistaking a pseudodeficiency state as affected.

The American College of Obstetricians and Gynecologists, through their Committee on Genetics, issued a consensus statement in October 2005 recommending that carrier screening be "offered before pregnancy to individuals and couples at high risk, including those of Ashkenazi Jewish, French–Canadian, or Cajun descent and those with a family history consistent with TSD." The committee also advised that genetic counseling and prenatal diagnosis be offered to both partners if they are proven to be carriers. A carrier–carrier couple has a 25% risk with each pregnancy of having an affected fetus. Prenatal diagnosis allows the option for pregnancy termination in the case of an affected offspring. Another approach chosen by a few at-risk couples is preimplantation testing in which normal embryos only are selected through *in vitro* fertilization and molecular analysis of a single cell at the blastocyte stage.

Therapy

Currently, there is no cure. The maintenance of proper nutrition and hydration, prevention of infectious disease, and control of seizures are the main

goals of treatment. Complications such as constipation and joint contractures in the infant and psychiatric disturbances and tremor in late-onset patients require appropriate pharmacological interventions. We favor high-calorie, semisolid or blenderized feedings, if necessary via a gastrostomy tube, laxatives and diazepam for infants, and carbamazepine and propanolol in adults. All clinical variants may benefit from physical therapy.

Experimental approaches include enzyme replacement, bone marrow transplantation, substrate synthesis inhibition, pharmacologic chaperones, and gene therapy. Because the locus of pathology in the G_{M2} gangliosidoses is within the CNS, successful enzyme replacement will require efficient targeting to neuronal lysosomes. Breaching the blood–brain barrier is possible through intrathecal administration or osmotic disruption, but neither technique has achieved sufficient concentration of enzyme within brain parenchyma to be therapeutic. Furthermore, cell media derived from murine fibroblasts overexpressing Hex A activity, when added to cultured Tay–Sachs fibroblasts, does not achieve high enough concentrations to hydrolyze the G_{M2} ganglioside present in these cells.

Following bone marrow transplantation (BMT), donor-derived cells of microglia/macrophage lineage can be found within the CNS. Sandhoff disease mice that have received BMT with wild-type cells show enzyme positivity in the leptomeninges, choroid plexus, and perivascular regions of the cerebrum but not within neurons. Although microglial activation and neuronal apoptosis were suppressed, neuronal storage of ganglioside was not altered. It is therefore not surprising that neurological deterioration has progressed in the few TSD patients who have had a BMT.

Another approach, substrate reduction therapy (SRT), employs an inhibitor of glycolipid synthesis to reduce the influx of substrate into lysosomes. One such compound, *N*-butyldeoxynojirimycin (NB-DNJ), inhibits the enzyme that joins glucose with ceramide to form glucosylceramide, a basic step in ganglioside biosynthesis. With a reduced substrate load, it is assumed that further reduction in G_{M2} ganglioside storage can be achieved through the action of residual enzyme activity. Both TSD and Sandhoff mice treated with NB-DNJ show a significant reduction in brain G_{M2} content. Also, Sandhoff mice have delayed symptom onset and live longer. When treated with both NB-DNJ and BMT, these mice survived significantly longer than those treated with NB-DNJ or BMT alone. On the other hand, according to anecdotal evidence, TSD infants who have received NB-DNJ (as miglustat) continue to deteriorate neurologically. Perhaps this result would be expected since infants with TSD have no residual enzyme activity. However, a controlled open-label trial of miglustat in late-onset patients who do have residual enzyme activity also did not show convincing benefit of the drug. Potential advantages of SRT are its oral route of administration and the presence of significant drug concentrations in the CSF suggesting ready availability to the CNS.

Pharmacological chaperones are yet another means to correct the enzyme deficiency, in this case by stabilizing the confirmation of the mutant Hex A protein. Instead of undergoing premature degradation, it can bypass the quality control system of the endoplasmic reticulum and enter the lysosome to enhance residual enzyme levels. Potential pharmacological chaperones for *HEXA* include NGT (*N*-acetylglucosamine-thiazoline) and pyrimethamine, both of which have been shown to enhance Hex A activity in cultured cells from patients with the late-onset Gly269Ser mutation. Human trials of chaperone therapy are in the planning stage.

Gene therapy represents another attractive alternative for treatment of the G_{M2} gangliosidoses. Trials with recombinant AAV vectors expressing various lysosomal enzymes have proven efficacious in treating a dog model of mucopolysaccharidosis I, mice with mucopolysaccharidosis VII, and a cat model of α-mannosidosis. Stereotaxic intracranial inoculation of recombinant AAV vectors encoding the human hexosaminidase α and β subunit genes into Sandhoff mice led to delay in the onset of the disease with preservation of motor function and reduced G_{M2} ganglioside storage and inflammation. Delivery of the therapeutic protein was widespread despite a single striatal injection. Retroviral constructs have also been used to correct Tay–Sachs cells in culture. It appears necessary for both the α and β subunit genes to be transduced jointly in order to prevent the nondefective subunit from limiting formation of the Hex A $\alpha\beta$ heterodimer.

Animal Models

Animal models of inherited human diseases are a rich resource for the study of disease pathogenesis and therapeutic approaches. Naturally occurring animal species with G_{M2} gangliosidosis have been reported in Muntjak deer and American flamingos (Hex A deficiency), in two feline breeds and a golden retriever (Hex A and B deficiency), and in Yorkshire pigs and Japanese spaniels (G_{M2} activator protein deficiency).

Mouse models have also been created of the three major forms of G_{M2} gangliosidosis by targeted disruption of their hex α and β subunit genes and G_{M2} activator protein gene in embryonic stem cells. Mice

deficient in Hex A appear phenotypically normal, although they accumulate G_{M2} ganglioside within specific regions of the brain. Their asymptomatic state was assumed to result from the availability of an alternative route for G_{M2} ganglioside hydrolysis involving the action of a sialidase not active in humans. However, symptoms have appeared in approximately 60% of these Tay–Sachs mice at an advanced age (i.e., by 18–24 months). The clinical appearance resembles human late-onset TSD. It consists of lordosis, progressing to abdominal dragging, tremor, and hind limb weakness and paralysis. Its appearance can be accelerated in the hex A-deficient females by repeated breeding within their first year. These mice have increased levels of G_{M2} ganglioside storage within the CNS and show symptoms by 13–18 months.

Sandhoff mice deficient in Hex A and B activity develop neurological symptoms at 3 months, including tremor, spasticity, hind limb rigidity, ataxic gait, muscle wasting and weakness, and paralysis, dying 4–6 weeks after disease onset. Because of the absence of Hex B, the sialidase bypass pathway is not available and therefore there is more extensive neuronal storage. In addition, storage material is also present within visceral organs.

A double knockout mouse has also been produced with disruption of both the α subunit and β subunit genes causing complete absence of all forms of lysosomal β-hexosaminidase. These mice are even more impaired than the Sandhoff mice, with a clinical picture resembling a mucopolysaccharidosis phenotype due to the additional accumulation of glycosaminoglycans within the nervous system and viscera.

The mouse model of G_{M2} activator deficiency expresses a phenotype intermediate between the Tay–Sachs and Sandhoff mouse. They have a normal life span but develop severe motor deficits. Unlike the Hex A-deficient mice, they store the glycolipid G_{A2} in cerebellar neurons and glial cells. Intraneuronal lysosomal inclusions and apoptosis are common features of these mouse models. In addition, in the Sandhoff mouse there is evidence of CNS inflammation and altered blood–brain barrier permeability.

Prevention

Tay–Sachs disease carrier screening has become the prototype for genetic disease prevention. It is based on the availability of relatively simple and inexpensive laboratory tests, a defined population subgroup at high risk, and the ability to monitor pregnancies of carrier–carrier couples. Community-based carrier testing programs began in 1970 and spread to high schools, college campuses, and community centers. However, the response among the general Jewish community has been poor, with only approximately 10% of the target population having been screened.

Concerns have been raised about the testing of adolescents in a school setting. These involve issues of informed consent (vs. peer pressure); the potential for stigmatization and discrimination; costs to the school for providing explanatory letters, obtaining parental consent, and protecting students' privacy; and the difficulty of reporting results and providing further genetic counseling in a confidential manner. To address the many obstacles created by carrier screening, the Orthodox Jewish community established the Dor Yeshorim program. It is designed for testing of unmarried young men and women who remain anonymous to the program. They are matched to their blood sample by a number which is randomly generated and stored in a computer together with their test results. When a marriage is proposed within their community, using their assigned number, the results of their testing are compared. If a carrier–carrier situation is found, the matchmaker will search for other potential spouses. At no time is the test result disclosed to either party. Although this system ensures freedom from stigmatization, it requires that each family member of marriageable age and in each generation be tested.

As of 2000, more than 1.4 million individuals had been tested worldwide and approximately 51 000 carriers identified, including 1400 at-risk couples. As a consequence, the number of infants born with TSD has markedly decreased, particularly among the Ashkenazi Jewish population. The development of DNA chip and bioarray technologies to detect multiple mutations should, in time, allow cost-effective screening for the entire child-bearing population. In the meantime, efforts to educate the high-risk population through their physicians and clergy should be enhanced.

See also: Brain Composition: Age-Related Changes; Lipids and Membranes in Brain Aging; Niemann–Pick Disease.

Further Reading

Andin K, Bakir B, Tatli B, et al. (2005) Proton MR spectroscopy in three children with Tay–Sachs disease. *Pediatric Radiology* 35: 1081–1085.

Autti T, Joensuu R, and Åberg L (2007) Decreased T2 signal in the thalami may be a sign of lysosomal storage disease. *Neuroradiology* 49: 571–578.

Borfitz JM and Margolis M (2007) *The Home Care Book. A Parents Guide to Caring for Children with Progressive Neurological Diseases*. Brookline, MA: National Tay–Sachs & Allied Diseases Association.

Cachòn-Gonzàlez MB, Wang SZ, Lynch A, et al. (2006) Effective gene therapy in an authentic model of Tay–Sachs related diseases. *Proceedings of the National Academy of Sciences of the United States of America* 103: 10373–10378.

Committee on Genetics, American College of Obstetricians and Gynecologists (2005) Screening for Tay–Sachs disease. *Obstetrics and Gynecology* 106: 893–894.

Evans PR (1981) Tay–Sachs disease: A centenary. *Archives of Disease in Childhood* 62: 1056–1059.

Gravel RA, Kaback MM, Proia RL, et al. (2001) The G_{M2} gangliosidoses. In: Scriver CR, Beaudet AL, Valle D, and Sly WS (eds.) *The Metabolic and Molecular Basis of Inherited Diseases*, 8th edn., pp. 3827–3876. New York: McGraw-Hill.

Haberland C, Brunngraber E, Witting L, and Brown B (1973) The white matter in G_{M2} gangliosidosis. *Acta Neuropathologica (Berlin)* 24: 43–55.

Kolter T and Sandhoff K (1998) Glycosphingolipid degradation and animal models of G_{M2}-gangliosidoses. *Journal of Inherited Metabolic Diseases* 21: 548–563.

Maegawa GHB, Stockley T, Tropak M, et al. (2006) The natural history of juvenile subacute G_{M2} gangliosidosis: 21 new cases and literature review of 134 previously reported. *Pediatrics* 118: e1550–e1562.

Myerowitz R, Lawson S, Mizukami H, et al. (2002) Molecular pathophysiology in Tay–Sachs and Sandhoff disease as revealed by gene expression profiling. *Human Molecular Genetics* 11: 1343–1350.

Neudorfer O, Pastores GM, Zeng BJ, et al. (2005) Late-onset Tay–Sachs disease: Phenotypic characterization and genotypic correlations in 21 affected patients. *Genetics in Medicine* 7: 119–123.

Purpura DP and Suzuki K (1976) Distortion of neuronal geometry and formation of aberrant synapses in neuronal storage disease. *Brain Research* 116: 1–21.

Terry RD and Weiss M (1963) Studies in Tay–Sachs disease: II. Ultrastructure of the cerebrum. *Journal of Neuropathology and Experimental Neurology* 22: 18–55.

Triggs-Raine B, Mahuran DJ, and Gravel RA (2001) Naturally occurring mutations in G_{M2} gangliosidosis: A compendium. *Advances in Genetics* 44: 199–224.

Walkley SU, Zervas M, and Wiseman S (2000) Gangliosides as modulators of dendritogenesis in normal and storage disease-affected pyramidal neurons. *Cerebral Cortex* 10: 1028–1037.

Wilkins RH and Brody IA (1969) Tay–Sachs' disease. *Archives of Neurology* 20: 103–111.

Relevant Website

http://www.hgmd.cf.ac.uk – The Human Gene Database at the Institute of Medical Genetics in Cardiff.

Temperature Sensation

A D Craig, Barrow Neurological Institute, Phoenix, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Psychophysics of Temperature Sensation

The normal temperature of human skin is *c.* 32–34 °C (90–94 °F), with hands and face warmest. Temperatures above that are perceived as warm, and temperatures above *c.* 45 °C (113 °F) are perceived as hot and painful. Temperatures below 32–34 °C are perceived as cool, and temperatures below *c.* 15 °C (59 °F) are perceived as painfully cold and burning. These temperature ranges correspond with the sensitivity of distinct classes of thermoreceptors in the skin (see below). Humans can distinguish temperature differences as small as *c.* 0.1 °C within the innocuous range, and we report cooling or warming intensity linearly.

By applying small metal probes with varied temperatures to the skin, individual spots can be found that are uniquely sensitive to warming, to cooling, and to noxious heat and/or noxious cold. Each spot corresponds with the location of a particular thermoreceptor. These spots are most clearly separated on the torso, where large intervening insensitive areas can be found, and they are most densely packed on the lips, nose, and hands. This reveals first the specificity and second the innervation density of the different classes of thermoreceptors.

The German physiologist EH Weber in 1859 described two thermal illusions. In the Thaler (dollar) illusion, a cool silver coin placed on the forehead of a prone individual feels noticeably heavier than a warm one. This can be explained by the nonlinear increment in activity caused by a cool temperature in type 1 slowly adapting (SA1) mechanoreceptors, which thereby elicit an increased sensation of pressure. The second illusion, usually called simply the Weber illusion, occurs as follows: one places both hands first on a lukewarm (thermoneutral) object for 1 min, and then each hand separately on a slightly cooler and a slightly warmer object, respectively, for 1 min, and finally both hands again on the initial neutral object, whereupon different sensations are perceived from the two hands. The neutral object now feels initially cool to the hand that was warmed, and initially warm to the hand that was cooled. This perceptual difference soon fades, and it may not occur if the hands are returned to the neutral object too quickly or if the temperature difference is too great. This illusion demonstrates, first, that warm- and cool-sensitive thermosensory pathways both adapt slowly to ambient temperature, second, that they display a narrow thermal 'dead zone,' and third, that they have a dynamic sensitivity to slight changes in skin temperature.

The 'thermal grill' (first described by Thunberg in 1896) is an illusion of painful (burning) cold that is elicited by placing one's hand on a platform consisting of a series of closely spaced metal bars that are alternately cool (20 °C) or warm (40 °C). The illusion is strongest if the temperatures of the bars are actively altered from an initial thermoneutral value: if, while the hand is held on the platform, one set of bars is first warmed and then (within 2–5 s) the alternate set is cooled, then a feeling of increasing warmth is perceived that blends alarmingly into noxious heat. The same feeling can be elicited most simply by touching a cold object to the veins on the back of a hand immersed in warm water, or most dramatically by pouring warm water on a foot that is numb with cold, or most informatively by applying a cool stimulus during a pressure block of myelinated fiber conduction in a peripheral nerve. A similar feeling of 'synthetic heat' can be elicited with temporally alternating warm and cool pulses presented on the foot or with rapid cooling of small thermodes on the fingers. All of these instances demonstrate that a cool stimulus activates two distinct thermosensory pathways, one responsible for the sensation of cooling and another responsible for the sensation of burning cold, and that the latter is normally inhibited by the former. In the original thermal grill, the cool-sensitive thermoreceptive-specific pathway is inhibited by the presence of the alternate warm bars, thus unmasking the simultaneous activation of the polymodal nociceptive pathway that is responsible for the feeling of burning cold.

Temperature stimuli also elicit a graded affective feeling of pleasantness or unpleasantness that depends on the body's thermoregulatory needs. For example, a cool glass of water feels refreshingly pleasant if one is excessively warm but gnawingly unpleasant if one is 'chilled to the bone.' Conversely, a pricking hot shower feels welcome if one is chilled but painful if one is hot. This valenced affect, which is the perceptual correlate of thermoregulatory behavioral motivation, demonstrates that temperature sensation can be regarded as a homeostatic emotion (i.e., both a feeling and a motivation) that drives behavior, like hunger, thirst, and pain.

Temperature Sensation in Other Species

Virtually all animals thermoregulate and display behavioral thermal sensitivity, and physiological and ethological studies have demonstrated cool and warm sensitivity in bacteria, worms, insects, honeybees, fish, lizards, birds, and mammals. Some animals have specialized thermoreceptive organs. For example, certain snakes have directionally selective pit organs that are highly sensitive to warmth (infrared energy) and are used for detecting and targeting prey. Some fish (e.g., sharks) have unique cool-sensitive ampullae (that are also sensitive to electric currents). Many birds, which like mammals are homeothermic rather than poikilothermic, have dense thermosensors around the bill and the brood patch that are important for behavioral control of egg incubation. All mammals have thermoreceptors similar to humans; for example, cats can be trained to respond to warm or cool differences in floor temperature as small as 0.2 °C, and the nocturnal owl monkey has a unique and highly evolved discriminative thermosensory capacity focused at the nose. Nevertheless, forebrain processing of temperature sensation is crucial for behavior only in highly encephalized primates.

Thermoreceptors

Thermoreceptive afferents have been characterized by physiological recordings from peripheral nerve fibers in many species, including humans. Cool and warm thermoreceptors show both dynamic (transient) and static (steady) responses that represent the rate of change and the absolute level, respectively, of cool and warm stimuli. Individual cool thermoreceptors have thresholds near, above, or below normal skin temperature (~20–40 °C). Some show bell-shaped response curves that peak and then diminish at noxious cold temperatures, and others show sigmoidal curves that increase and then saturate at such temperatures. Similarly, warm thermoreceptors have thresholds that span normal skin temperature (~30–40 °C) and show bell-shaped or sigmoidal response curves that diminish or saturate at noxious heat temperatures. Both types are inhibited dynamically by temperature changes in the opposite direction, and both are sensitive to the absolute temperature at the receptor, rather than the temperature gradient across the skin. In contrast, different types of nociceptors respond to noxious heat and/or noxious cold.

Thermoreceptors have spot-like receptive fields in the skin. Cool receptors are much more numerous than warm receptors in the skin, whereas warm receptors are much more numerous in deep tissues (muscle, viscera, spinal cord), corresponding to the different homeostatic priorities for monitoring temperature in the thermoregulatory core and shell. Cutaneous thermoreceptors are concentrated in orofacial regions (mouth, tongue, nose, lips, eyes, and ears), as well as on the hands and feet. Warm receptive afferents are unmyelinated C-fibers that conduct very slowly, and cool receptive afferents are predominantly thinly myelinated Aδ-fibers that conduct more rapidly. Thus, a pressure block of myelinated fiber conduction in a peripheral nerve interrupts touch and cool sensations, leaving sensations of warm and pain, whereas blockade with a local anesthetic (e.g., lidocaine) interrupts warm and pain sensations, leaving touch and cool. Thermoreceptors are very modality selective and are excited by mechanical deformation very infrequently. Particular temperature-sensitive membrane channels (e.g., TRPM8) are associated with thermoreceptors. Cool thermoreceptors are also sensitive to application of menthol. Cool thermoreceptors are sometimes paradoxically activated by temperatures above 45 °C, corresponding to the brief sensation of cold that can occur when the hand is immersed in hot water. In contrast, the sharp biting sensation elicited by touching an object that is very cold (such as dry ice) is due to the abrupt activation of myelinated nociceptors by the sudden formation of small ice crystals in the skin.

Central Neural Processing

Spinal Level

Thermoreceptive processing begins in the dorsal horn of the spinal cord. Selectively warm- and cool-sensitive spinothalamic cells in the most superficial layer (lamina I) respond to thermal stimuli over large areas of skin; receptive field sizes vary from 2 to 4 cm^2 near the nose to half of the hand to the entire lateral abdomen. The summation produces a mean response profile that aggregates the response curves of individual thermoreceptors. Cool-sensitive lamina I neurons have thresholds at or above normal skin temperature, and they respond very linearly to cooling down to about 15 °C but show constant activity below that temperature (Figure 1). They have ongoing discharge at normal skin temperature that is inhibited by warming. Conversely, warm-sensitive neurons have thresholds at or below normal skin temperature, and they show graded warm responses up to about 45 °C, where their activity becomes constant. They have ongoing discharge that is inhibited by cooling. Thus, the activity of cool- and warm-sensitive thermoreceptive-specific spinal neurons closely parallels human thermosensory capacities and temperature perceptions. Further, thermoreceptive-specific neurons are

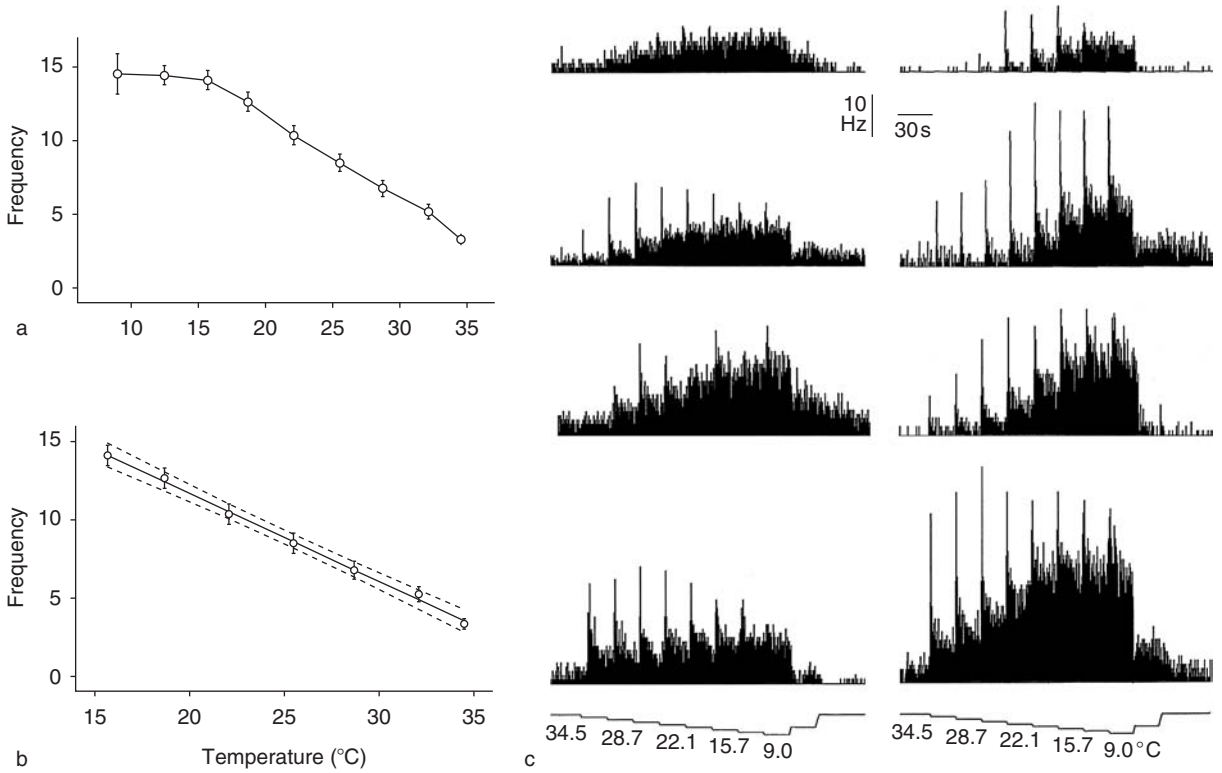

Figure 1 Quantitative characterization of cool-sensitive lamina I thermoreceptive-specific spinothalamic neurons: (a) average static response to temperature, (b) linearity of regression against innocuous cool temperatures, (c) examples of individual unit responses to 20 s cooling staircase stimuli showing both the dynamic and the static responses. Reproduced from Craig AD, Krout K, and Andrew D (2001) Quantitative response characteristics of thermoreceptive and nociceptive lamina I spinothalamic neurons in the cat *Journal of Neurophysiology* 86: 1459–1480.

modality selective and are insensitive to other stimuli (except for menthol-like chemicals). They also have unique cell morphologies, unique axonal conduction velocities and projections, and unique responses to activation of descending controls from the brain stem and to pharmacological agents; so, they can be termed 'labeled lines' for temperature sensation. Two subpopulations of cells are sensitive to cooling the epidermis and the dermis, respectively. In contrast, spinal neurons in the deep dorsal horn that receive input from SA1 mechanoreceptors show graded tactile and weak, nonlinear cool sensitivity. In addition, polymodal nociceptive lamina I spinothalamic neurons are sensitive to noxious cold, noxious heat (>45 °C), and noxious mechanical stimulation. These neurons have little ongoing discharge, and their threshold to cold stimulation is *c.* 24 °C (75 °F), the temperature below which environmental cold becomes unpleasant. They are responsible for the sensation of burning pain caused by noxious heat, noxious cold, and the thermal grill.

Thermoreceptive-specific lamina I neurons project to homeostatic regions in the spinal intermediolateral cell column, where sympathetic preganglionic vasoconstrictor neurons are located, and to homeostatic regions in the brain stem, where catecholaminergic pre-autonomic motor centers control blood flow, heart rate, and respiration. Direct access to central autonomic control and thermoregulatory integration centers is provided by way of major projections to the A1 noradrenergic cell group in the caudal ventrolateral medulla (which projects to hypothalamus) and to the parabrachial nucleus (which projects both to hypothalamus and amygdala). Other thermoreceptive lamina I cells project to the contralateral forebrain by way of the lateral spinothalamic tract; surgical interruptions of this pathway in the spinal cord (e.g., to reduce cancer pain) completely eliminates contralateral temperature sensation.

Thalamocortical Level

In monkeys and humans, the ascending lamina I thermoreceptive-specific pathway reaches a compact nucleus in posterolateral thalamus (VMpo). These neurons also show thermoreceptive-specific responses, and they are organized spatially so that there is a segregated, antero-posteriorly organized thermoreceptive map of the body. Microstimulation in this region in awake humans (during neurosurgical procedures for movement disorders or chronic pain) elicits

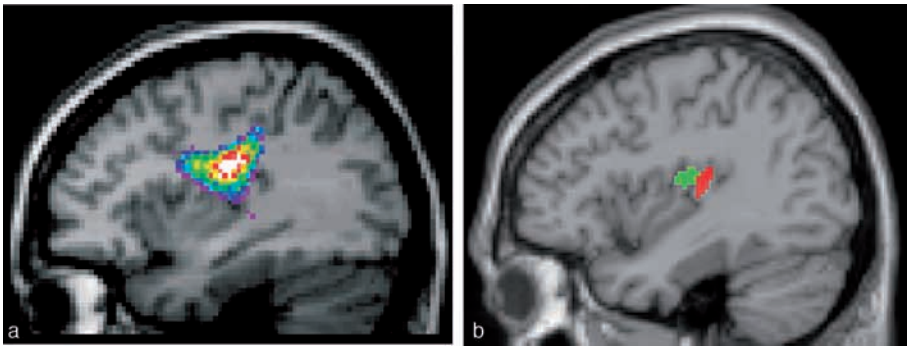

Figure 2 Functional imaging data showing the location of discriminative thermosensory cortex in the dorsal posterior insula: (a) linear regression of PET data with cooling stimuli; (b) antero-posterior somatotopic organization of regressed cooling activation by stimulation of the hand (red) and neck (green). (a) Reproduced from Craig AD, Chen K, Bandy D, and Reiman EM (2000) Thermosensory activation of insular cortex. *Nature Neuroscience* 3: 184–190. (b) Reproduced from Hau LH, Strigo IA, Baxter LC, Johnson SC, and Craig AD (2005) Anteroposterior somatotopy of innocuous cooling activation focus in human dorsal posterior insular cortex. *American Journal of Physiology. Regulatory, Integrative, and Comparative Physiology* 289: R319–R325.

descriptions of immediate, discrete cooling sensations that are graded with the intensity of stimulation and well-localized on the body. These thalamic neurons topographically relay thermoreceptive activity to a cortical site in the dorsal posterior insula. Functional imaging with positron emission tomography (PET) has confirmed that activation of this area of the human brain is directly correlated with cool skin temperatures (Figure 2), and low-power laser stimulation has similarly indicated that activation in this region correlates with warm skin temperatures. Recent studies using functional magnetic resonance imaging (fMRI) have shown that this region has an antero-posteriorly organized topographic map of thermoreceptive activity (Figure 2). Consonant with these observations, clinical findings indicate that damage to posterolateral thalamus or the dorsal posterior insula interrupts temperature sensation on the contralateral body. Thus, there is a direct thermoreceptive-specific spinothalamocortical pathway that provides a map of temperature sensation over the body within this particular area of the cerebral cortex, and it is responsible for the discrimination and the localization of temperature sensations in humans.

It is important to recognize that the discriminative thermosensory area in dorsal posterior insular cortex of primates and humans is actually part of a set of interoceptive areas that provide a general representation of all aspects of the physiological condition of the body. The same lamina I spinothalamocortical pathway by way of VMpo that conveys thermoreceptive activity also generates topographic maps that represent other distinct feelings from the body, such as first (sharp) pain, second (burning) pain, itch, muscle ache, visceral sensation, and so on, that constitute other aspects of the body's physiological condition. Further, afferent inputs from the vagal and glossopharyngeal nerves cause activity in the adjacent portion of dorsal insula related to hunger, thirst, taste, breathlessness, and cardiovascular condition. Thus, these insular cortical areas serve as a primary homeostatic afferent image of the state of the body. The evolutionary role of these limbic sensory cortical areas is to modulate the activity of regulatory regions in the brain stem and hypothalamus that maintain the health of the body, that is, regions directly involved in homeostatic integration and control. In humans, they also provide a basis for discriminative sensation, and further, for distinct feelings that are related to subjective disposition and emotions (i.e., 'how you feel'). The affective experiences of thermal comfort and discomfort that accompany all temperature sensations are associated with thermosensory activation of anterior frontoinsular and medial frontal cortical regions.

Thermoreceptive-specific neurons are not found within the somatosensory cortical areas that are associated with the sensation of touch. Therefore, whereas it has long been held that temperature sensation is an aspect of the sense of touch, the neuroanatomy of temperature sensation indicates that the discriminative thermosensory capacity of humans has emerged evolutionarily as one distinct aspect of the cortical representation of the physiological condition of the body, which developed along with the enormous encephalization in primates and especially humans. The highly developed sense of temperature in humans provides the capacity to use temperature information not only as a signal of the condition of our bodies, but also as a discriminative sensation useful for recognizing objects and exploring the environment.

Finally, it is noteworthy that damage to primary thermosensory cortex causes not only loss of temperature sensation, but may also cause the so-called central pain syndrome, in which ongoing burning pain is reported in a region of thermanesthesia or

thermosensory dysfunction. The thermal grill illusion is remarkably analagous to the central pain syndrome, suggesting that this syndrome may be regarded as a thermoregulatory dysfunction due to the loss of the cool inhibition of the polymodal pain pathway. Reports that such pain may be alleviated by warming or by drugs that increase tissue perfusion and thermogenic heat production support this hypothesis.

See also: Cingulate Cortex; Energy Homeostasis: Thermoregulation; Mechanoreceptors; Nociceptor Responses; Thermoregulation During Sleep and Sleep Deprivation; Thermoregulation: Autonomic, Age-Related Changes.

Further Reading

Adelson DW, Wei JY, and Kruger L (1997) Warm-sensitive afferent splanchnic C-fiber units *in vitro*. *Journal of Neurophysiology* 77: 2989–3002.

Andrew D and Craig AD (2001) Spinothalamic lamina I neurones selectively responsive to cutaneous warming in cats. *Journal of Physiology* 537: 489–495.

Craig AD (2002) How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews Neuroscience* 3: 655–666.

Craig AD, Chen K, Bandy D, and Reiman EM (2000) Thermosensory activation of insular cortex. *Nature Neuroscience* 3: 184–190.

Craig AD, Krout K, and Andrew D (2001) Quantitative response characteristics of thermoreceptive and nociceptive lamina I spinothalamic neurons in the cat. *Journal of Neurophysiology* 86: 1459–1480.

Gisolfi CV and Mora F (2000) *The Hot Brain*. Cambridge, MA: MIT Press.

Green BG and Cruz A (1998) Warmth-insensitive fields: Evidence of sparse and irregular innervation of human skin by the warmth sense. *Somatosensory and Motor Research* 15: 269–275.

Greenspan JD, Roy EA, Caldwell PA, and Farooq NS (2003) Thermosensory intensity and affect throughout the perceptible range. *Somatosensory and Motor Research* 20: 19–26.

Hellon R (1983) The cardiovascular system. In: Shepard JT and Abboud FM (eds.) *Handbook of Physiology*, Section 2, pp. 659–673. Bethesda: American Physiological Society.

Hensel HT (1981) *Thermoreception and Temperature Regulation*. London: Academic Press.

Hua LH, Strigo IA, Baxter LC, Johnson SC, and Craig AD (2005) Anteroposterior somatotopy of innocuous cooling activation focus in human dorsal posterior insular cortex. *American Journal of Physiology. Regulatory, Integrative, and Comparative Physiology* 289: R319–R325.

Kenshalo DR and Duclaux R (1977) Response characteristics of cutaneous cold receptors in the monkey. *Journal of Neurophysiology* 40: 319–332.

Pehl U, Schmid HA, and Simon E (1997) Temperature sensitivity of neurones in slices of the rat spinal cord. *Journal of Physiology Paris* 498: 483–495.

Temporal Processing in the Auditory Pathway

D Oertel, M J McGinley, and X-J Cao, University of Wisconsin, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Acoustic Information Is Carried in the Timing of Firing of Neurons

The vertebrate brain carries information both in the spatial pattern in which neurons are activated and by the timing of their firing. Although often considered separately, they are linked because the determination of a spatial firing pattern requires that firing be integrated over a time period. The time over which spatial activation is monitored is in many cases inherent to the method used and often has a low temporal resolution, for example, in functional MRI or in methods based on glucose metabolism to examine those spatial patterns of activation, so relatively little attention has been given to these issues in the studies of many regions of the brain. In the mammalian auditory system the importance of the relationship between temporal and spatial firing patterns among neurons has long been appreciated and continues to be explored.

The ability of neurons to encode acoustic information in the timing of firing is illustrated especially clearly in the ability of neurons to fire in phase with sounds. In the early 1960s, Nelson Kiang showed that auditory nerve fibers encode the frequency of tones not only in their tuning but also in the timing of their firing. The groups led by Kiang at the Massachusetts Eye and Ear Infirmary in Boston and by Jerzy Rose at the University of Wisconsin in Madison documented, using some of the earliest lab computers, how auditory nerve fibers encode frequency by firing in phase with sounds. In cats and monkeys, low-frequency sounds, sounds below about 4000 Hz, evoke action potentials at a particular phase of the sound cycle. In order to do so, the probability of firing must be greater over one-half of the 250 μs cycle than the other and requires that neurons fire with a temporal precision on the order of 100 μs. Neurons fire in phase with tones, skipping some cycles when sounds are soft or at frequencies too high for neurons to fire with every cycle, as illustrated in Figure 1.

The ability of neurons in the auditory system to encode phase underlies the ability to use interaural time differences to localize sound sources in the horizontal plane. The relative time of arrival of a sound at the two ears is a function of the angle of the sound source with respect to the nose and ears. The maximal interaural time differences are about 0.5 ms for sounds coming from the side of the head in animals such as human beings with large heads and are smaller in animals with smaller heads. Interaural time differences produce corresponding interaural phase differences. Recent work has shown that mammals use every cycle of a sound to measure interaural time differences at low frequencies. To make use of small interaural time and phase differences for localizing sound sources, auditory neurons need to signal with enough temporal precision to resolve these differences.

Precision in the timing of firing is observed across the tonotopic range. Auditory nerve fibers tuned to high frequencies and their targets can be phase-locked to the rate of amplitude modulation. Their targets in the cochlear nucleus, octopus cells, signal the presence of broadband transients with a temporal precision in the range of 100 μs.

The precision of the temporal coding of frequency also plays a role in the detection of pitch. The range of musical pitch, 20–5000 Hz, lies largely in the range of frequencies where there is phase-locking and has been shown in many, but not all, cases to be related to the periodicity of sounds. Pitch, like phase-locking and unlike spatial tuning, becomes increasingly precise as the loudness of sounds increase. Bertrand Delgutte and Peter Cariani have shown that the period of a perceived pitch corresponds to the most common interval between action potentials in the population of auditory nerve fibers in cats, suggesting that differences of tens to hundreds of microseconds in intervals between action potentials in the population of auditory nerve fibers are detected and used to sense pitch. Detecting pitch is required to recognize voices and appreciate music.

The encoding of temporal cues is likely to contribute to understanding speech. Arnold Starr and Feng-Gang Zeng and their colleagues have shown that patients with auditory neuropathy lose the ability to detect brief gaps in noise as they lose the ability to understand speech. Contributions of the encoding of temporal cues for understanding speech have been revealed especially clearly in people with cochlear implants. Processors for cochlear implants generally divide the frequency spectrum into a small number of bands and use the envelopes of those bands to modulate the strength of a very rapid train of shocks. The fine structure of sounds, including phase-locking at low frequencies, is generally not encoded. People with cochlear implants understand speech remarkably well but they cannot hear pitch. Even when the phase of the fundamental frequency is encoded, the perception of pitch does not improve. Perhaps this is because

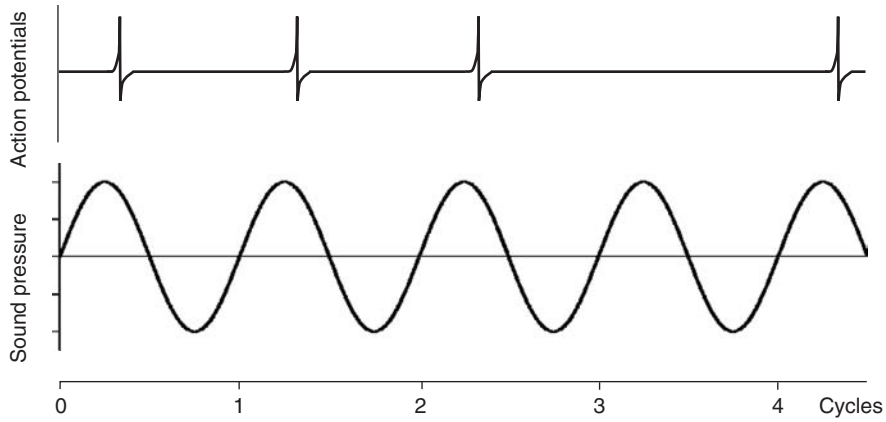

Figure 1 An illustration of what is meant by phase-locking shows how a neuron in the brain stem can fire (upper trace) in phase with the cycles of a tone (lower trace). Neurons that phase-lock fire at a particular phase in the cycle but can skip cycles. As the period of a cycle becomes shorter with increasing tonal frequency, jitter in the timing of firing reduces the precision of timing with respect to phase; under such conditions the synchronization coefficient is reduced. In the auditory nerve of mammals phase-locking is robust at frequencies lower than 2000 Hz; at higher frequencies, the synchronization coefficient drops off and no phase-locking can generally be detected at frequencies higher than 4000 Hz. Neurons can fire at every cycle in response to low frequency tones but skip more cycles with increasing frequency.

cochlear implants do not reach the apex of the cochlea, so even if phase cues are presented for low-frequency sounds, they can only be presented at inappropriate cochlear locations. The inability to detect pitch leads not only to an inability to appreciate music but also to an inability to distinguish the age or gender of voices that makes it difficult to follow conversations in noisy environments. These observations indicate that pitch probably requires that encoding of temporal fine structure occur within the appropriate location in the tonotopic map.

This article is concerned with understanding how neurons in the auditory system can fire with temporal precision. It is impossible to be complete in reviewing the topic. This chapter will focus on what we have learned about the cellular mechanisms that underlie temporal encoding of acoustic information in the ventral cochlear nucleus (VCN). These findings illustrate both that neurons use common mechanisms to encode timing precisely and also that differences in those properties allow neurons to perform different integrative tasks and to convey different aspects of sensory information.

Temporal Precision Is Greatest in the Brain Stem

Acoustic information that is contained in the timing of firing in the auditory system necessarily comes from the cochlear hair cells that transduce sounds and is conveyed synaptically through chains of neurons that form the auditory pathway. The precision of timing information in the auditory nerve is limited by the ability of voltage changes in hair cells to follow mechanical movements in the cochlea and by the precision whereby those voltage signals are conveyed synaptically to spiral ganglion cells. Auditory nerve fibers in turn convey that information to several different cell types in the cochlear nuclei, each of which forms a separate pathway. In so doing, the auditory pathway is subdivided into multiple, parallel ascending pathways. These cell types and the pathways they form differ in the precision with which they encode temporal information. In some, octopus and bushy cells for example, the precision of timing information is sharpened by convergence whereas in the principal cells of the dorsal cochlear nucleus acoustic information contained in the timing of firing is largely lost.

The major components of the ascending auditory system are illustrated in Figure 2. Every auditory nerve fiber carries acoustic information from the cochlea to multiple groups of neurons whose projection patterns differ and which therefore form separate, parallel, ascending pathways. In the VCN auditory nerve fibers contact T stellate, bushy, and octopus cells. The axons of octopus cells cross the midline to innervate the contralateral superior paraolivary nucleus (SPN) and the ventral nucleus of the lateral lemniscus (VNLL). The axons of bushy cells innervate the superior olivary complex; some innervate the medial superior olive (MSO) bilaterally, others innervate the contralateral medial nucleus of the trapezoid body (MNTB), and yet others innervate the ipsilateral lateral superior olive (LSO). The axons of T stellate cells terminate in the LSO, periolivary nuclei, the contralateral VNLL, and the inferior colliculus. Nuclei that contain neurons with biophysical specializations for encoding timing are indicated with gray in Figure 2. Ultimately,

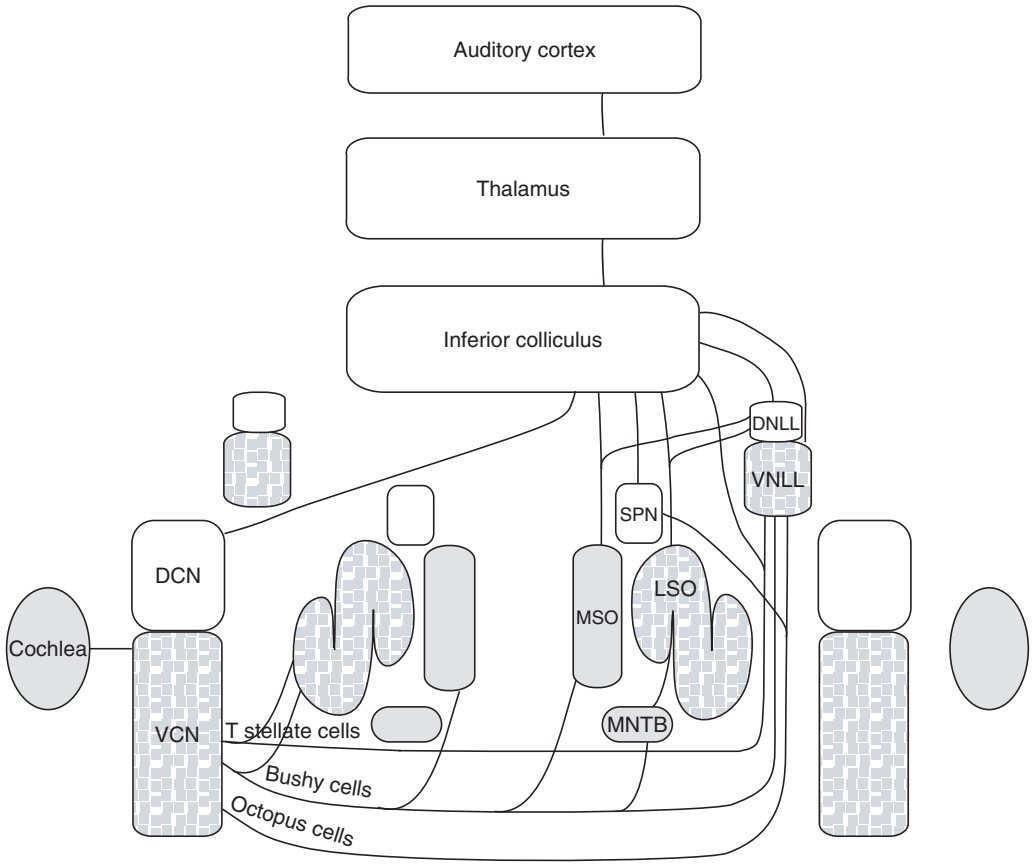

Figure 2 The major components in the ascending auditory pathways are shown diagrammatically. Sound is transduced in the cochlea. Information is carried through the axons of spiral ganglion cells, the auditory nerve fibers, to the dorsal (DCN) and ventral (VCN) cochlear nuclei. Several parallel pathways are formed from the cochlear nuclei through the brain stem which converge in the inferior colliculus. The principal cells of the DCN project directly to the contralateral inferior colliculus. The VCN contains three types of principal cells, T stellate cells, bushy cells and octopus cells. The axons of T stellate cells innervate the lateral superior olive (LSO) and the ventral nucleus of the lateral lemniscus (VNLL) and continue to the inferior colliculus. Bushy cells innervate the superior olivary complex. One subtype of bushy cell innervates the medial superior olive (MSO) bilaterally. Another subtype of bushy cell innervates the contralateral medial nucleus of the trapezoid body (MNTB), whose principal cells in turn inhibit the LSO. A third subtype of bushy cell innervates the LSO ipsilaterally. At least some of the bushy cells innervate the VNLL. Octopus cells innervate the contralateral superior paraolivary nucleus (SPN) and the contralateral VNLL. The DNLL receives input from the ipsilateral MSO and LSO. Principal cells of the MSO, LSO, SPN, VNLL, and DNLL project to the inferior colliculus. Principal cells of the inferior colliculus project to the medial geniculate body of the thalamus and these in turn project to the auditory cortex. Neurons with low-voltage-activated K^+ conductances that enable them to signal with temporal precision are found in nuclei that are colored gray. In the nuclei that are colored with solid gray, all neurons have those specializations. In nuclei indicated with the patterned gray only some neurons have those conductances. DNLL, dorsal nucleus of the lateral lemniscus.

all parallel pathways through the brain stem lead to the inferior colliculus and thence to the auditory part of the thalamus and auditory cortex. The precision in the timing of firing that is evident at lower levels is not observed in the inferior colliculus or its targets. Information contained in precise timing of firing is either lost or re-encoded in the inferior colliculus.

Biophysical Specializations that Underlie Precision in the Timing of Firing Are Variations on a Common Pattern

The temporal precision with which neurons carry acoustic information differs between cell types. Some neurons (including the spiral ganglion cells whose axons form the auditory nerve, ventral cochlear nuclear octopus and bushy cells, the principal cells of the MSO, MNTB, and some neurons in the LSO and VNLL) are specialized to receive and convey precisely timed signals. These groups of neurons perform differing integrative tasks that are reflected in differences in their shapes, the identity and distribution of their synaptic inputs, in the route taken by their axons, and in their electrophysiological properties. However, specializations for preserving and conveying timing in their signaling have common features. They receive rapidly rising and falling synaptic currents, they have low input resistances and, as a consequence also short

time constants, that allow voltages to rise and fall quickly. Nuclei that contain neurons with such biophysical properties are indicated in gray in Figure 2. In the following section common features and functionally significant differences will be illustrated in the properties of the three major groups of principal cells of the VCN: octopus, bushy, and T stellate cells. Robin Davis and her colleagues have documented the presence of similar features in the auditory nerve. Ian Forsythe and his colleagues revealed them in the principal cells of the MNTB and in some cells in the LSO. Shu Hui Wu and Nace Golding and their colleagues have described similar properties in principal cells of the MSO. Shu Hui Wu and colleagues revealed their presence in the VNLL.

Intrinsic Properties of Principal Cells of the VCN Span a Wide Range

Differences in the way neurons in the brainstem receive and convey timing information are illustrated in the properties of the three major types of principal neurons of the VCN in Figure 3. octopus, bushy, and T stellate cells. Recordings of the intracellular voltage with respect to the extracellular environment, by impaling neurons with sharp microelectrodes, or by recording through a patch in the whole-cell mode, show that neurons have similar resting potentials, near –65 mV. Differences in their biophysical properties are revealed in their responses to identical, depolarizing current pulses delivered through the recording electrodes. Figure 3 illustrates that a pulse of depolarizing current produces a rapid, small change in voltage that falls quickly back to the resting potential at its offset in octopus cells. A similar current pulse injected into a bushy cell produces an action potential; at the end of the current pulse the voltage falls back toward rest more slowly than in octopus cells. A current pulse injected into a T stellate cell causes the cell to fire a train of regularly spaced action potentials; at the end of the pulse the voltage falls back toward rest more slowly than in either octopus or bushy cells. Differences in the responses to current pulses results from differences in morphology as well as differences in number, type, and location of ion channels.

An important determinant of the responses to transmembrane current is the input resistance near

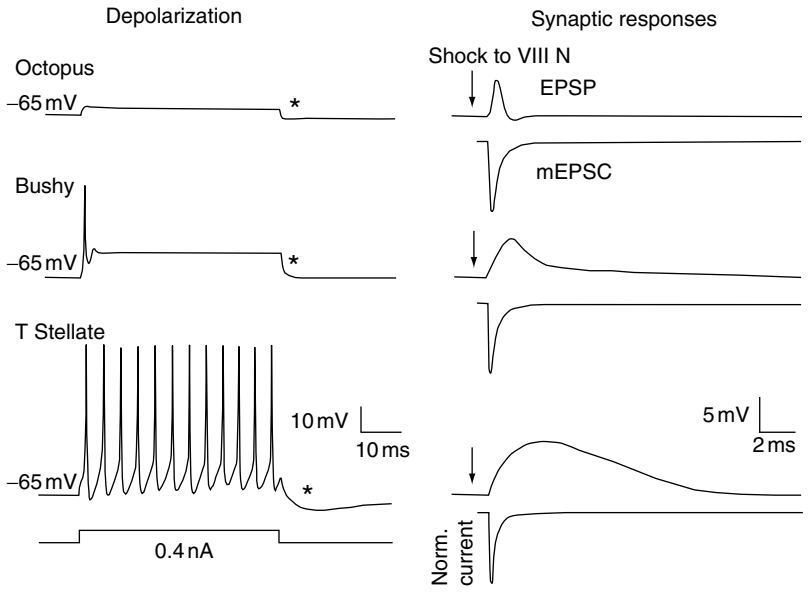

Figure 3 Biophysical properties shape responses to depolarizing current and responses to shocks of the auditory nerve. Left: Responses to pulses of depolarizing current. The response of the octopus cell is subthreshold; a larger current pulse produces an action potential at the onset of a depolarization (not illustrated). The return to rest of the voltage after the end of the current pulse is most rapid in the octopus cell and slowest in the T stellate cell (*). Right: Shocks to the auditory nerve (arrows) evoke EPSPs whose duration is briefest in the octopus cell, intermediate in the bushy cell, and longest in the T stellate cell. The time course of averaged, spontaneous, mineature EPSCs measured from each of the cell types is shown beneath the corresponding EPSPs. The low input resistance of octopus cells makes EPSPs barely longer than the EPSCs whereas in T stellate cells the EPSPs outlast EPSCs. All of the traces are redrawn. Some are adapted, redrawn and idealized, from traces that were recorded by Nace Golding. From figure 30–14, Kandel ER, Schwartz JH, and Jessell TM (2000) *Principles of Neural Science*, 4th edn. New York: McGraw-Hill. Averaged, normalized mEPSCs are derived from data reported by Gardner SM, Trussell LO, and Oertel D (1999) Time course and permeation of synaptic AMPA receptors in cochlear nuclear neurons correlate with input. *Journal of Neuroscience* 19: 8721–8729.

the resting potential. Both the magnitude of the voltage changes evoked by current and their rates of change (Figure 3) reflect the input resistance. The input resistance of octopus cells is on average only about 5 MΩ, so that large currents are required to evoke action potentials but the voltage can rise and fall quickly. Bushy cells have higher input resistances, about 70 MΩ, so that currents produce larger voltage changes and smaller currents are required to reach threshold, but the fall of voltage back to rest is slower. T stellate cells generally have input resistances around 75 MΩ and have correspondingly larger and slower voltage changes. The rate at which voltage changes depends upon the input resistance as well as on the input capacitance which is a function of the size, shape, and surface area of neurons. The fastest rate at which membrane potential can change can be estimated by fitting voltage changes at the termination of a current pulse with a single exponential function and comparing the time constants, τ. The τ varies more than an order of magnitude between cells; it is only about 0.2 ms in octopus cells, about 1 ms in bushy cells, and about 7 ms in T stellate cells.

Responses to synaptic currents, like those to injected current pulses, are shaped by the electrical properties of neurons. In the VCN, release of neurotransmitter from auditory nerve fibers produces excitatory postsynaptic currents (EPSCs) in target neurons that cause excitatory postsynaptic potentials (EPSPs). Stephanie Gardner measured the time course of synaptic currents produced by the spontaneous release of neurotransmitter (mEPSCs) in each of these cell types and found that their time course was similar in octopus, bushy, and T stellate cells as illustrated on the right in Figure 3. Although EPSCs are similar, shocks to auditory nerve fibers (Figure 3, arrow) evoke EPSPs whose shapes vary between cell types (Figure 3). EPSPs are briefest in octopus cells, longer in bushy cells, and still longer in T stellate cells. Timing is encoded most sharply in neurons with the briefest EPSPs.

Glutamate Receptors in the Auditory Brain Stem Nuclei Have Rapid Kinetics

Glutamate receptors vary in their properties as a function of the subunits from which they are formed. Glutamate receptors are ligand-gated ion channels that vary in the rate at which they are activated, deactivated, and at which they desensitize when they are exposed to ligands. The number of possible compositions is large because these receptors are formed from combinations of four subunits, GluR1–GluR4, each of which exists in one of the two alternatively spliced forms, flip or flop. The work of Laurence Trussell and Robert Wenthold and their colleagues has shown that the glutamate receptors in neurons of the VCN are formed from GluR3 flop and GluR4 flop α-subunits and that they are the most rapidly activating, deactivating, and desensitizing glutamate receptors known. The rapid activation, deactivation, and desensitization of these receptors allow synaptic currents to be brief and thus sharply timed.

Voltage-Sensitive Conductances Allow Signaling to be Temporally Precise

Neurons that encode timing in the mammalian auditory pathway contain conventional voltage-sensitive ion channels. Specializations lie in the distribution and density of those channels. The axons of neurons that encode timing produce action potentials that are mediated through voltage-sensitive, and tetrodotoxin-sensitive, Na^+ channels (g_{Na}). Action potentials are repolarized by various K^+ conductances, some activated at relatively negative membrane potentials and others activated at more positive potentials, and in some cells by Ca^{2+}-activated K^+ conductances. The density, distribution, and form of some of the ion channels that mediate these conductances not only vary between types of auditory neurons, giving them distinctive electrophysiological signatures, but in some cell types (spiral ganglion cells and principal cells of the MNTB) have also been shown to vary as a function of position in the tonotopic map.

Cells that signal with temporal fidelity have low-voltage-activated K^+ (g_{KL}) conductances and hyperpolarization-activated mixed-cation conductances (g_h). Paul Manis and colleagues were the first to draw attention to the importance of g_{KL}, showing that neurons that behaved like bushy cells have not only a K^+ conductance that is activated by strong depolarizations (more positive than –40 mV; g_{KH}, carries I_{KH}) but also one that is activated at quite negative voltages (more positive than –70 mV; g_{KL}, carries low-voltage-activated K^+ current, I_{KL}). The neurons that behaved like T stellate cells lacked g_{KL}. Recent experiments show that g_{KL} seems to be present in all neurons in the brain stem auditory pathway that are known to encode timing with precision. Resting outward K^+ currents through g_{KL} are to some extent balanced by inward currents through g_h.

Exceptionally large g_{KL} and g_h endow octopus cells with the low input resistances that make them fast. At rest two opposing conductances, g_{KL} and g_h, are partially activated and balance one another. Figure 4 illustrates that these two conductances are activated by opposite changes in membrane potential and illustrates that at the resting potential both g_h and g_{KL} are partially activated. The channels that underlie g_{KL}

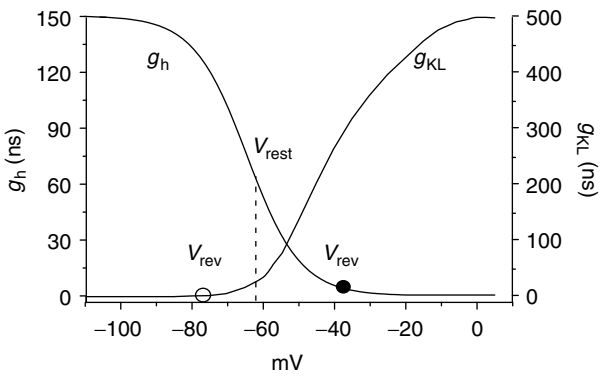

Figure 4 Two opposing conductances affect the resting properties of octopus cells. A low-voltage-activated K^+ conductance (g_{KL}) is activated at depolarizing voltages. The activation curve is scaled to the average peak, maximum conductance (based on data reported by Bal R and Oertel D (2001) Potassium currents in octopus cells of the mammalian cochlear nucleus. *Journal of Neurophysiology* 86: 2299–2311). Its reversal potential (V_{rev}) is near the K^+ equilibrium potential (open circle), identifying it as a K^+ conductance. A mixed cation conductance (g_h) is activated by hyperpolarizing voltages. Its reversal lies between the equilibrium potentials of permeant ions, Na^+ and K^+ (filled circle). The plot is scaled to the average, maximum conductance (based on data reported by Bal R and Oertel D (2000) Hyperpolarization-activated, mixed-cation current (I_h) in octopus cells of the mammalian cochlear nucleus. *Journal of Neurophysiology* 84: 806–817). At the resting potential (V_{rest}) both conductances are partially activated, contributing substantially to the low input resistance of octopus cells. At rest the currents, a function of the conductance and the difference between the resting potential and the reversal potential, through the two conductances are matched.

are opened by positive changes in potential, whereas the channels that underlie g_h are opened by negative changes in potential. Under physiological conditions, the action of each of these conductances feeds back on itself negatively so that each will tend toward its own minimal activation in the steady state. The reversal potential, V_{rev}, is the voltage at which there is no net movement of ions even when channels are open. The current, I_{KL}, reverses near the equilibrium potential of K^+, at −76 mV, confirming that g_{KL} carries K^+ ions (Figure 4, open circle V_{rev}). The reversal potential of I_h lies near −38 mV (Figure 4, filled circle); its position between the equilibrium potentials for Na^+ and K^+ ions reflects the permeability of these channels to both Na^+ and K^+, with P_{Na}/P_K being about 0.2. In octopus cells, at rest the K^+ ions that flow out through g_{KL} are balanced by Na^+ ions flowing in through hyperpolarization-activated, mixed cation channels, g_h. Bushy cells are also endowed with g_{KL} and g_h but the maximum amplitudes of these conductances are smaller. T stellate cells lack g_{KL}.

Under physiological conditions, the voltage of neurons is dynamic and the state of voltage-sensitive ion channels is governed by their voltage sensitivity as well as by their kinetics. Changes in membrane potential result in rapid changes in g_{Na} and g_{KL}, but only slower changes in g_h. The firing of cells can thus be affected by the rate at which they are depolarized. Principal cells of the VCN vary in their sensitivity to the rate at which they are depolarized and consequently in the duration of the time window over which depolarizing current can sum to bring the neuron to threshold. Figure 5 shows that the types of principal cells of the VCN differ in their dynamic properties. Octopus cells fire action potentials when they are depolarized with rapid, but not with slow, ramped currents. In families of currents when ramps were brief, depolarization was rapid and octopus cells fired action potentials even to relatively small currents; when ramps to similar current levels were longer and depolarization was slower, only the largest current brought the cell to threshold (Figure 5(a)). Plots of the cell's response as a function of the subthreshold rate of depolarization show that firing consistently occurred when the rate of depolarization exceeded a threshold. The example shown in Figure 5(a) had a threshold rate of depolarization of about 12 mV ms^{-1}; on an average rate thresholds of octopus cells were about 10 mV ms^{-1}. These properties allow an integration window to be defined operationally. Depolarizing current can contribute to evoking an action potential over a period that is the duration of the slowest suprathreshold ramp to the time at which the action potential is triggered (Figure 5(a), period between the dashed lines on left). Integration windows are very short in octopus cells, about 1.5 ms. In order to fire action potentials, bushy cells, too, must be depolarized faster than at a threshold rate but that threshold rate is slower and the integration window over which depolarizing current can sum is longer (Figure 5(b)). In bushy cells rate thresholds were about 6 mV ms^{-1} on average and integration windows were about 5 ms. In contrast with octopus and bushy cells, T stellate cells are insensitive to the rate of depolarization and have no rate thresholds. Even when they are depolarized very slowly, it appeared that they fire action potentials upon reaching a voltage threshold (Figure 5(c), dashed line). Their integration window is not limited by intrinsic biophysical properties.

The sensitivity to rate of depolarization and the duration of the integration window depend on g_{KL}. Figure 6 shows comparisons of the responses of neurons in the absence and presence of α-dendrotoxin, a toxin from the green mamba snake that blocks g_{KL} with great specificity. Under control conditions, an octopus cell and a bushy cell require depolarization at greater than a threshold rate to fire action potentials (Figures 6(a) and 6(b), left). When those same cells were exposed to α-dendrotoxin, they fired action

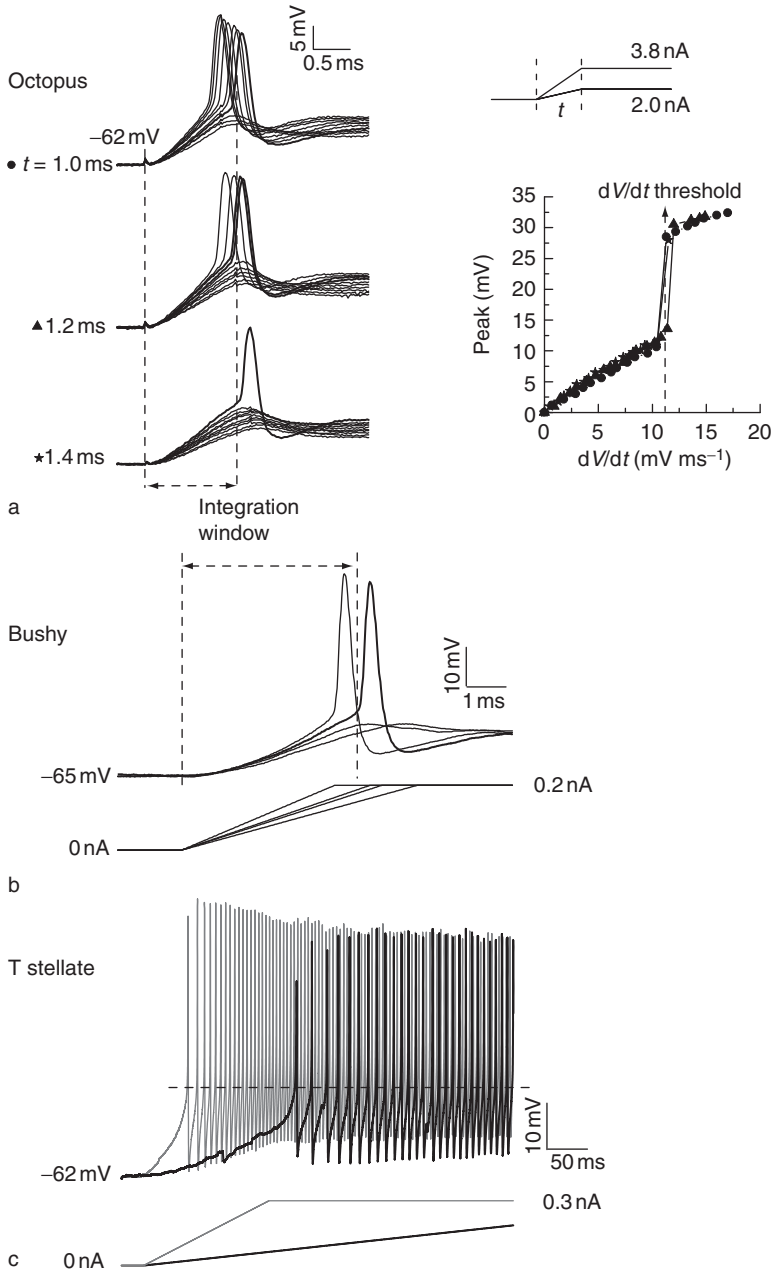

Figure 5 Principal cells of the VCN differ in their sensitivity to rate of depolarization. (a) An octopus cell was depolarized with families of current ramps to steady levels as illustrated at the top right. The duration of the ramps was varied between 1 (upper traces) and 1.4 ms (lower traces). The octopus cell fired action potentials only when the rate of depolarization exceeded a threshold. The step in the plot on the lower right shows that this octopus cell had to be depolarized at more than 11 mV ms^{-1} to fire. The longest duration between the onset of a ramp and the initiation of the action potential defines the integration window: the longest period over which depolarizing current can sum and lead to an action potential (period between dashed lines). The integration window of this octopus cell was 1.3 ms. (b) A bushy cell was depolarized at varying rates by varying the duration of current ramps toward 0.2 nA. To fire, the bushy cell had to be depolarized faster than 4.3 mV ms^{-1} and it had an integration window of 4.7 ms. (c) A T stellate cell was depolarized with currents that rose at two different rates. Light trace shows the response to a relatively rapid ramp to 0.3 nA. The cell started to fire when it reached a threshold voltage and then increased the firing rate. When the same cell was depolarized with a very slow ramp, action potentials rose when roughly the same potential was reached. The T stellate cell behaved as if it fired action potentials when it reached a certain voltage (dashed line) even when the peaks of the action potentials were reduced by inactivation of the Na^+ conductance. (a) Traces and plots are reproduced from figure 3, Ferragamo MJ and Oertel D (2002) Octopus cells of the mammalian venteral cochlear nucleus sense the rate of depolarization. *Journal of Neurophysiology* 87: 2262–2270). (b and c) Reprinted from McGinley MJ and Oertel D (2006) Rate thresholds determine the precision of temporal integration in principal cells of the ventral cochlear nucleus. *Hearing Research* 216–217: 52–63.

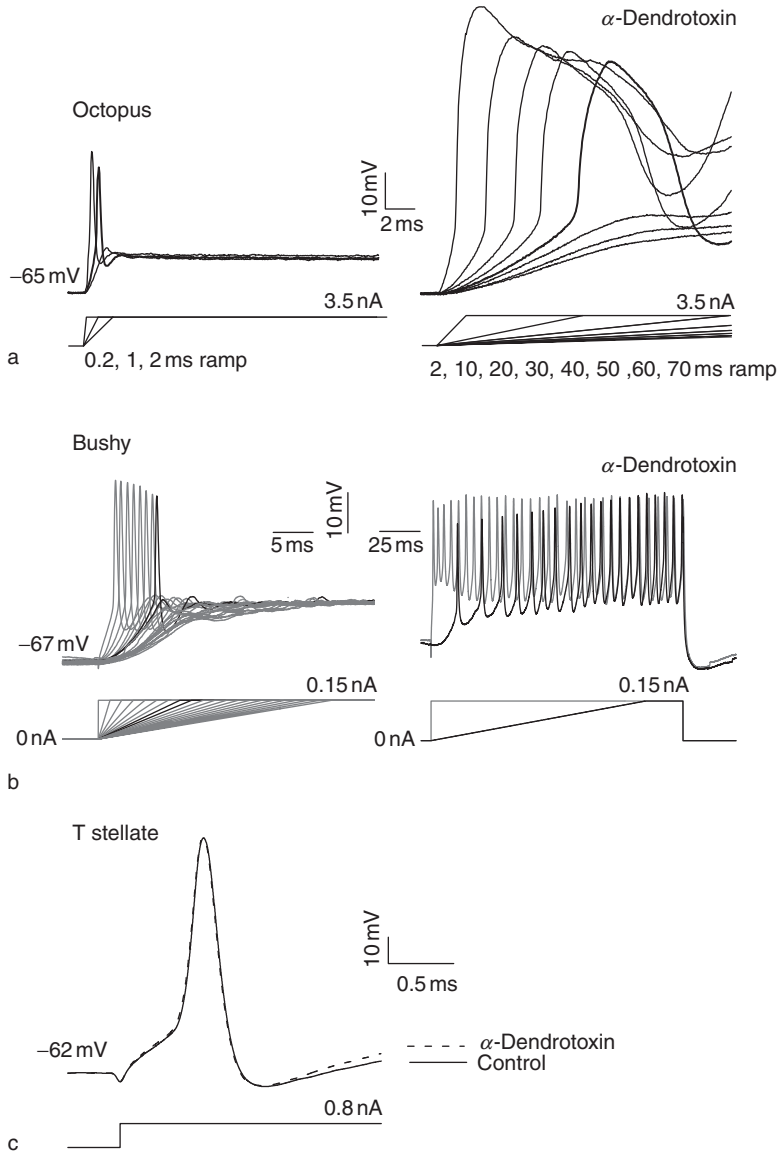

Figure 6 The low-voltage-activated K^+ conductance (g_{KL}) plays a key role in determining the sensitivity to rate of depolarization. (a) Responses of an octopus cell to ramps show that ramps needed to be rapid to be suprathreshold. When that same cell was bathed in 100 nM α-dendrotoxin, a saturating concentration of a toxin that specifically blocks g_{KL}, action potentials could be evoked even with slow ramps. In the absence of g_{KL}, action potentials became tall and broad, revealing a role of g_{KL} in the repolarization of action potentials. (b) Bushy cells, too, had to be depolarized at more than a threshold rate to fire. In the presence of 66 nM α-dendrotoxin the bushy cell fired repetitively even when depolarized with a slow ramp. Peaks of action potentials were not affected by the toxin, indicating that repolarization depended on a different type of K^+ conductance. (c) The firing of action potentials by T stellate cells is unaffected by 100 nM α-dendrotoxin indicating that they contain little, if any, g_{KL}. (a and c) Reproduced from figures 6 and 8, Ferragamo MJ and Oertel D (2002) Octopus cells of the mammalian venteral cochlear nucleus sense the rat of depolarization. *Journal of Neurophysiology* 87: 2266, 2267. (b) Reprinted from McGinley MJ and Oertel D (2006) Rate thresholds determine the precision of temporal integration in principal cells of the ventral cochlear nucleus. *Hearing Research* 216–217: 52–63.

potentials even in response to slow depolarization (Figure 6(a) and 6(b), right). In octopus cells, but less so in bushy cells, g_{KL} plays a role in repolarizing action potentials because in its absence, action potentials are unusually tall and of long duration. In contrast with octopus and bushy cells, T stellate cells have little or no g_{KL} because they are not affected by α-dendrotoxin and as a result the duration of their integration windows is effectively unlimited (Figure 6(c)).

The duration of EPSPs is matched to the duration of the integration windows in principal cells of the VCN. Synaptic responses to shocks to the auditory nerve are briefest in octopus cells, of intermediate

duration in bushy cells, and longest in T stellate cells (Figure 3). This observation results in part from the fact that the voltage-sensitive conductances that determine the integration window also shape synaptic responses. The voltage sensitivity of g_{KL} is maximally steep over the voltage range over which EPSPs occur. Synaptic depolarization in octopus and bushy cells thus activates g_{KL} which then contributes actively to the repolarization of the EPSP, making EPSPs brief. In T stellate cells, in contrast, the decay of EPSPs is determined by passive decay rate (Figure 3).

Rate of depolarization thresholds and short integration windows enhance the ability of neurons to detect synchrony among subthreshold inputs and to convey the timing of synchronous inputs in their own firing. Octopus cells provide a particularly clear example. Figure 7(a) shows that as shocks to the auditory nerve increased in strength, activating more and more auditory nerve fibers, the rate of rise of the EPSPs increased and the peaks of synaptic responses became larger and occurred only slightly earlier. Octopus cells are innervated by many auditory nerve fibers (>60). Action potentials in any one auditory nerve fiber produce only small EPSPs with slow depolarization so that many auditory nerve fibers must sum to produce a suprathreshold response. To achieve the threshold rate of depolarization, the rising phases of EPSPs from many auditory nerve fibers must sum. The firing of octopus cells signals the activation of populations of auditory nerve fibers whose firing occurs within the integration window. Two other features of synaptic responses of octopus cells are notable. First, the action potentials of octopus cells are conventional in the axon but are strongly attenuated at the cell body where recordings are made. In being attenuated, action potentials interfere minimally with synaptic integration. Second, the timing of the peak is remarkably constant over a wide range of EPSP amplitude. In part that is because the peak is determined not only by the time course of the excitatory, synaptic inward currents but also by the rapid repolarization by g_{KL} that truncates the depolarization. The ability of octopus cells to detect synchronous activation of auditory nerve fibers and to fire with great temporal precision is evident in responses to sound *in vivo*. They respond sensitively and with precision to broadband acoustic transients. These phenomena also occur in bushy cells but they are less exaggerated. Bushy cells detect synchronous activation among a smaller number of auditory nerve inputs (4–20). They also convey the timing of the activation of those inputs with enhanced temporal precision.

The unlimited integration windows make it possible for T stellate cells to signal the presence of very slow signals and it makes it possible for their excitability to be modulated synaptically. Figure 7(b) illustrates responses of one T stellate cell to electrical stimulation of the nerve root. The weakest shocks evoked relatively brief responses like those illustrated in Figure 3 that were suprathreshold. As the strength of shocks was increased, the depolarizations lasted

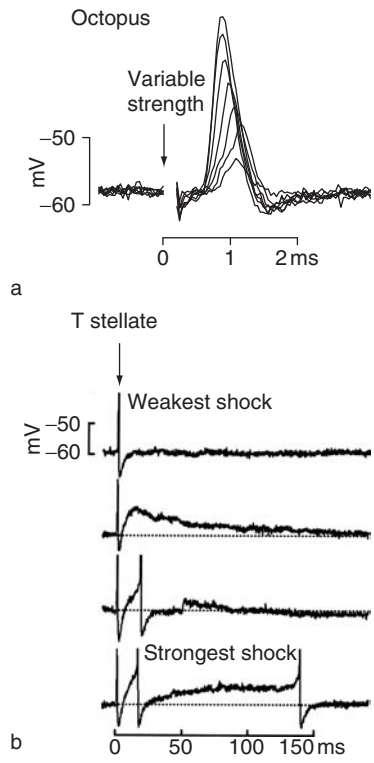

Figure 7 Synaptic responses in principal cells of the VCN are matched to their biophysical properties. (a) Shocks to the auditory nerve evoke brief synaptic responses in an octopus cell; depolarizing phases last less than 1 ms. As the shock strength was increased, synaptic responses were evoked after a synaptic delay. The responses grew by small increments over wide range, reflecting the recruitment of converging auditory nerve fibers onto a single octopus cell. Action potentials recorded at the cell body are small in octopus cells; the smallest response in the illustration is subthreshold, and all the others are suprathreshold. The timing of the peaks of action potentials varied by about 300 μs over a wide range of shock strengths. (b) In a T stellate cell, shocks to the auditory nerve evoked responses that lasted for up to 100 ms. In this cell weak shocks evoked a suprathreshold EPSP. (The synaptic delay cannot be resolved in these long traces and the peaks of action potentials are truncated.) Stronger shocks evoked not only the rapid suprathreshold EPSPs but also slower, longer-lasting excitation. Even slow and steady depolarization evoked action potentials. (a) Reproduced from figure 3, Oertel D, Bal R, Gardner SM, Smith PH, and Joris PX (2000) Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11775. (b) Reproduced from figure 2, Ferragamo MJ, Golding NL, and Oertel D (1998) Synaptic inputs to stellate cells in the venteral cochlear nucleus. *Journal of Neurophysiology* 79: 54.

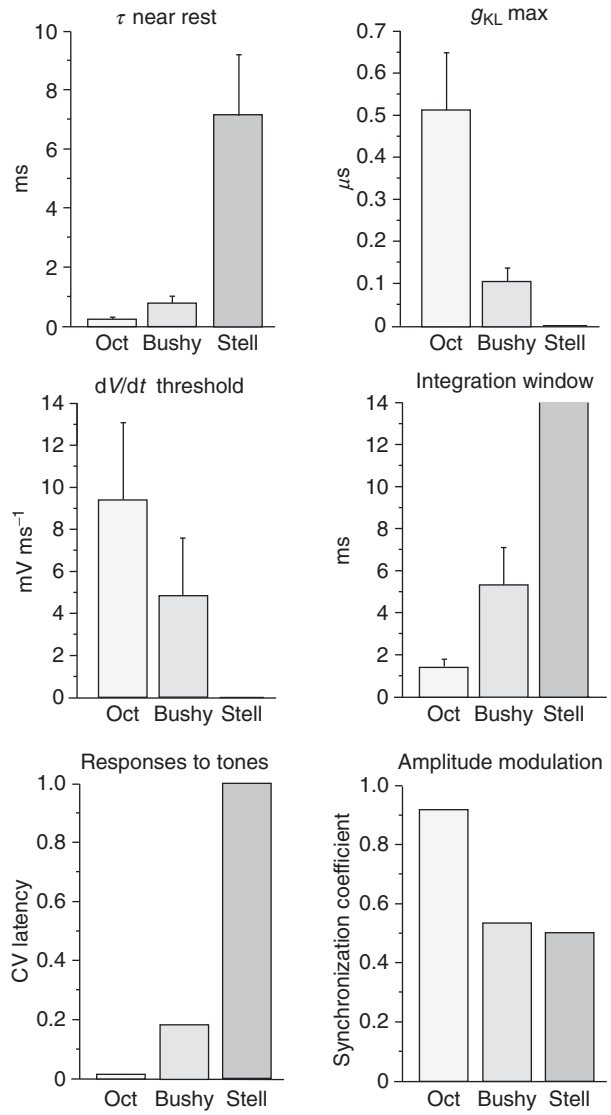

Figure 8 The range of biophysical properties allows principal cells of the VCN to encode and convey timing information with varying resolution. Principal cells vary in time constant, τ, in large part because they vary in the expression of the low-voltage-activated conductance, g_{KL}. The presence of g_{KL} is associated with high rate-of-depolarization (dV/dt) thresholds and short integration windows. High rate thresholds and short integration windows make synaptic responses sharply timed. The sharp timing can be measured as a small coefficient of variation in the latency of responses to sound (CV latency) and in high synchronization coefficients to the modulation frequency in responses to AM tones. Data are shown as means with standard deviations and are derived from the following sources: τ, octopus and T stellate cells, Golding NL, Ferragamo MJ, and Oertel D (1999) Role of intrinsic conductances underlying responses to transients in octopus cells of the cochlear nucleus. *Journal of Neuroscience* 19: 2897–2905; bushy cells, unpublished results); g_{KLmax} (Bal R and Oertel D (2001) Potassium currents in octopus cells of the mammalian cochlear nucleus. *Journal of Neurophysiology* 86: 2299–2311, Cao X-J, Shatadal S, and Oertal D (2007) Voltage-sensitive conductances of bushy cells of the mammalian venteral cochlear nucleus. *Journal of Neurophysiology* (in press); dV/dt thresholds and integration windows (reprinted from McGinley MJ and Oertel D (2006) Rate thresholds determine the precision of temporal integration in principal cells of the ventral cochlear nucleus. *Hearing Research* 216–217: 52–63); coefficient of variation of latency in responses to tones (Rhode WS and Smith PH (1986) Encoding timing and intensity in the venteral cochlear nucleus of the cat. *Journal of Neurophysiology* 56: 261–286). Synchronization coefficients are compared in responses to sounds at the characteristic frequency of the neurons and modulated at 500 Hz (Rhode WS (1994) Temporal coding of 200% amplitude modulated signals in the ventral cochlear nucleus of cat. *Hearing Research* 77: 43–68).

for increasingly long times (>100 ms) and the duration of excitation was conveyed in its firing. Cholinergic currents, too, can modulate responses of T stellate cells, perhaps modulating the dynamic range of responses to sounds.

Relationship of Biophysical Specializations to Responses to Sound

The three major groups of principal cells in the VCN can be viewed as pathways for carrying different types of timing information (Figure 8). At one extreme, octopus cells sense transients. They signal the presence of broadband transients with exceptional temporal precision. They fail to signal the duration of sounds, however. Bushy cells are intermediate, enhancing the encoding of timing in narrowband transients to some extent yet still conveying information about the fine structure and duration of sounds. T stellate cells are at the opposite extreme in that they encode the duration of narrowband sounds but lose information about its fine structure.

The ranges of physiological parameters that underlie the differences and that determine how these pathways convey temporal information are summarized in Figure 8. Time constants around the resting potentials that reflect the input resistance and input capacitance varied between 0.2 ms in octopus cells and 7 ms in T stellate cells. The presence of g_{KL} is an important determinant both of the input resistance and of the shapes of EPSPs because it is partially activated at the resting potential and is maximally sensitive to voltage in the voltage range of EPSPs. This conductance is particularly large in octopus cells and is either absent or expressed at very low levels in T stellate cells, with bushy cells being intermediate (Figure 8, top panels). To make an analogy with current, octopus cells are at one extreme of the spectrum and detect AC signals and T stellate cells are at the opposite extreme and detect DC signals.

Differences in the expression of g_{KL} lead to differences in sensitivity to the rate of depolarization. Octopus cells require high rates of depolarization to fire action potentials and fire only transiently, whereas T stellate cells respond even to steady depolarizations. Bushy cells, again, are intermediate. Associated with

differences in sensitivity to rate of depolarization are differences in the duration of integration windows. At the one extreme, the rising phases of brief EPSPs must sum to cause octopus cells to fire while at the other extreme T stellate cells respond sensitively to small steady currents (Figure 8, middle panels).

Responses to sound reflect the biophysical properties of neurons. The latency in the timing of the initial responses to tones, for example, reflects the precision with which timing information is conveyed. The latency has a very small coefficient of variation in responses of octopus cells, a larger coefficient of variation in responses of bushy cells, and the largest coefficient of variation in T stellate cells (Figure 8, bottom left). Another measure of the encoding of timing is in the synchronization coefficient of responses to amplitude-modulated sound. The synchronization coefficient is greatest in octopus cells, and smallest in T stellate cells (Figure 8, bottom right).

A comparable analysis of sensitivity to rate of depolarization, of the variation in latency, and of responses to amplitude-modulated tones has not been made in other neurons that are known to encode timing with precision but it is likely to follow the same pattern. The presence of g_{KL} in auditory nerve fibers and in targets of bushy and octopus cells suggests that each of these groups of neurons is sensitive to the rate of depolarization and thus is sensitive to acoustic transients and enhances the precision in the timing of firing of its inputs. Other pathways through the brain stem lack g_{KL} enabling them to encode slower modulation.

See also: Auditory Cortex: Models; Auditory Cortex Structure and Circuitry; Auditory Evoked Potentials; Auditory System: Central Pathway Plasticity; Auditory System: Central Pathways; Hair Cells: Sensory Transduction; Language: Auditory Processes; Reticular Formation and the Brain Stem.

Further Reading

Cao X-J, Shatadal S, and Oertel D (2007) Voltage-sensitive conductances of bushy cells of the mammalian venteral cochlear nucleus. *Journal of Neurophysiology* (in press).

Bal R and Oertel D (2000) Hyperpolarization-activated, mixed-cation current (I_h) in octopus cells of the mammalian cochlear nucleus. *Journal of Neurophysiology* 84: 806–817.

Bal R and Oertel D (2001) Potassium currents in octopus cells of the mammalian cochlear nucleus. *Journal of Neurophysiology* 86: 2299–2311.

Eatock RA, Fay RR, and Popper AN (eds.) *Vertebrate Hair Cells*. New York: Springer.

Ferragamo MJ, Golding NL, and Oertel D (1998) Synaptic inputs to stellate cells in the venteral cochlear nucleus. *Journal of Neurophysiology* 79: 54.

Ferragamo MJ and Oertel D (2002) Octopus cells of the mammalian venteral cochlear nucleus sense the rate of depolarization. *Journal of Neurophysiology* 87: 2262–2270.

Gardner SM, Trussell LO, and Oertel D (1999) Time course and permeation of synaptic AMPA receptors in cochlear nuclear neurons correlate with input. *Journal of Neuroscience* 19: 8721–8729.

Gardner SM, Trussell LO, and Oertel D (2001) Correlation of AMPA receptor subunit composition with synaptic input in the mammalian cochlear nuclei. *Journal of Neuroscience* 21: 7428–7437.

Golding NL, Ferragamo MJ, and Oertel D (1999) Role of intrinsic conductances underlying responses to transistens in octopus cells of the cochlear nucleus. *Journal of Neuroscience* 19: 2897–2905.

Joris PX, Smith PH, and Yin TC (1998) Coincidence detection in the auditory system: 50 years after Jeffress. *Neuron* 21: 1235–1238.

Kandel ER, Schwartz JH, and Jessel TM (2000) *Principles of Neural Science*, New York: McGraw-Hill.

McGinley MJ and Oertel D (2006) Rate thresholds determine the precision of temporal integration in principal cells of the ventral cochlear nucleus. *Hearing Research* 216–217: 52–63.

Oertel D (1997) Encoding of timing in the brain stem auditory nuclei of vertebrates. *Neuron* 19: 959–962.

Oertel D, Bal R, Gardner SM, Smith PH, and Joris PX (2000) Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11773–11779.

Oertel D, Fay RR, and Popper AN (eds.) (2002) *Springer Handbook for Auditory Research* (Vol. 15): *Integrative Functions in the Mammalian Auditory Pathway*. New York: Springer.

Rhode WS (1994) Temporal coding of 200% amplitude modulated signals in the venteral cochlear nucleus of cat. *Hearing Research* 77: 43–68.

Rhode WS and Greenberg S (1992) Physiology of the cochlear nuclei. In: Fay RR and Popper AN (eds.) *The Physiology of the Mammalian Auditory Central Nervous System*, vol. 2, pp. 94–152. New York: Springer.

Rhode WS and Smith PH (1986) Encoding timing and intensity in the venteral cochlear nucleus of the cat. *Journal of Neurophysiology* 56: 261–286.

Trussell LO (1999) Synaptic mechanisms for coding timing in auditory neurons. *Annual Review of Physiology* 61: 477–496.

Yin TCT (2002) Neural mechanisms of encoding binaural localization cues in the auditory brainstem. In: Oertel D, Fay RR, and Popper AN (eds.) *Integrative Functions in the Mammalian Auditory Pathway*, pp. 99–159. New York: Springer.

Yost WA (2000) *Fundamentals of Hearing*, 4th edn. San Diego: Academic Press.

Terminal Differentiation: REST

S Aigner and G W Yeo, The Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In 1990, Gail Mandel and David Anderson independently reported the first data suggesting that expression of neuron-specific genes might be under the control of *cis*-acting negative regulator elements within these genes. These insights quickly led to the discovery in 1992 of the sequence of such an element, which was called RE1 by Mandel and neuron-restrictive silencer element (NRSE) by Anderson. Since NRSE/RE1-like sequences were found in at least three genes specifically expressed by neuronal cell types, it was immediately postulated that NRSE/RE1-dependent silencing might be a general mechanism to silence neuronal genes in nonneuronal cell types, mediated by a *trans*-acting silencing factor expressed in nonneuronal cells but not in neuronal cells. Indeed, separate cloning efforts by both groups, published in 1995, led to the identification of this silencer factor as the RE1-silencing transcription factor (REST) or neuron-restrictive silencer factor (NRSF). As had been hypothesized, REST/NRSF expression was observed to be generally high in nonneuronal cells types, with concomitantly low levels of transcripts of genes with an NRSE/RE1 element, and vice versa.

Although REST/NRSF's initially proposed function as a silencer of a large set of neuronally enriched genes has acquired widespread acceptance, evidence has accumulated in the past few years indicating that REST/NRSF plays additional, much more dynamic roles in normal development and adult brain function, as well as in disease. This view is supported by the discovery of a large network of REST/NRSF targets that includes many genes not exclusively expressed in neurons.

In this article, we review the architecture of the REST/NRSF gene and protein, including regulation of REST/NRSF expression, and present the current knowledge of its gene targets. In our discussion of its role in the developing and adult organism, we also briefly touch upon REST/NRSF regulation in neurological disease and cancer. Lastly, because the control of target gene expression by REST/NRSF is critically dependent on the action of its protein interaction partners, we also provide insight into REST/NRSF-recruited cofactors and how these are thought to modulate the function of this transcription factor.

REST/NRSF: Gene Organization, Alternative Splicing, and Protein Structure

The REST/NRSF gene and its exon structure are evolutionarily conserved from human to fugu (pufferfish). The gene is not found in flies or nematodes, suggesting that REST/NRSF is specific to the vertebrate lineage. It consists of three alternative first exons (exons I–III), located in the 5′ untranslated region (5′ UTR); three constitutively spliced exons (IV–VI); and a short (28 base pairs) alternatively spliced internal exon (exon N). The full-length REST/NRSF protein comprises three known functional domains: a DNA-binding domain and two repressor domains, which are located at the N- and C-termini of the protein. The two repressor domains function independently of each other and serve to recruit distinct transcriptional regulation complexes (Figure 1(b)). The protein also has lysine- and proline-rich regions; however, their significance is unknown. REST/NRSF's DNA binding domain encompasses an array of eight highly conserved zinc fingers, small independently folded nucleic acid binding motifs found in many other sequence-specific nucleic acid binding proteins. Thus, with its repressor and DNA binding domains, REST/NRSF shows an architecture typical of most transcription factors.

Although little is known about the function of alternative splicing events that affect the 5′ UTR of the REST/NRSF gene, the splice variant produced by inclusion of exon N, located within the region of the gene that encodes the DNA binding domain, has received a great deal of attention because it is primarily found in neurons. Its protein product, termed REST4, is a truncated version of REST/NRSF that terminates after the fifth zinc finger and therefore lacks the C-terminal repressor domain (Figure 1(a)). Thus, REST4 is likely to have a function distinct from full-length REST/NRSF, which will be discussed later.

NRSE/RE1: The DNA Binding Element of REST/NRSF

REST/NRSF regulates target genes by direct binding to its cognate DNA silencer element, the NRSE/RE1 (Figure 1(c)), which was originally identified in the promoters of the superior cervical ganglia 10 and type II sodium channel. With its long span of approximately 21 base pairs (bp) and its rather low degree of sequence degeneracy (Figure 1(c)), the NRSE/RE1 is a highly unusual transcription factor binding site.

This considerable level of sequence conservation has enabled stringent bioinformatic searches for candidate

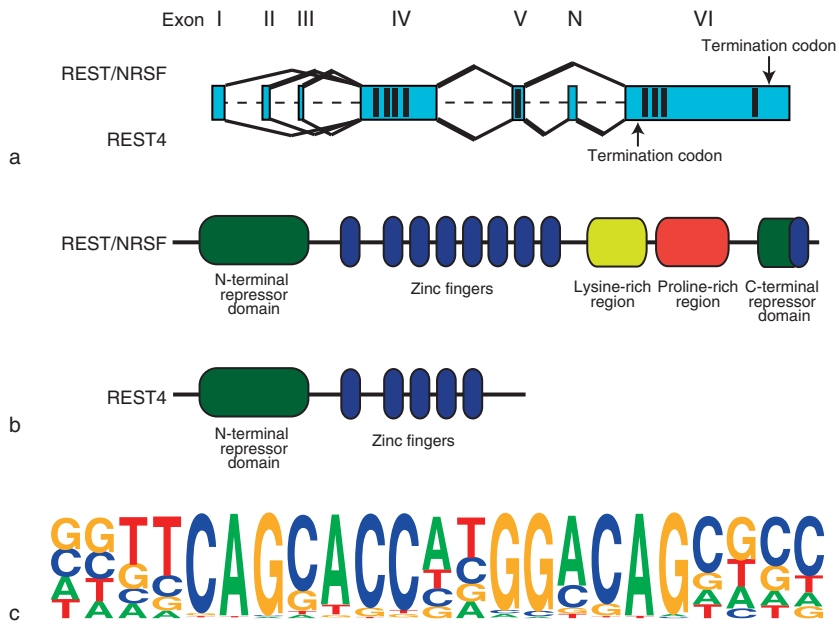

Figure 1 (a) Structure of the REST/NRSF gene and alternative isoform REST4. Exons are indicated as boxes and introns as dashed lines. Exon numbers in roman characters from I to VI are shown above the respective exons; the neuronal-specific exon located between exons V and VI is indicated as exon N. Exons I–III are alternative 5′ UTRs that are spliced to exon IV, represented by lines connecting exons at the intron–exon boundaries. Lines above the primary transcript represent the REST/NRSF primary isoform, and lines below the primary transcript represent the REST4 alternative isoform. Arrows indicate termination codons for REST/NRSF and REST4. (b) REST/NRSF protein includes eight zinc fingers near the N-terminal repressor domain and one zinc finger in the C-terminal domain. REST/NRSF contains a lysine-rich and a proline-rich region upstream of the C-terminal repressor domain. REST4 includes exon N, leading to a truncated protein encompassing the N-terminal repressor domain and five zinc fingers. (c) The conserved REST/NRSF DNA binding site, termed NRSE/RE1, is approximately 21 bp of long and contains of two highly conserved half-site motifs that are separated by two non conserved nucleotides and are flanked by several poorly conserved nucleotides. In this depiction of conserved mammalian NRSE/RE1 sites, generated with pictogram (http://genes.mit.edu/pictogram), the degree of conservation at each position is represented by the size of the nucleotide letter.

NRSE/RE1 sites in the genomes of multiple organisms, and several hundred conserved potential REST/NRSF binding sites in mammalian genomes are now known. Although most of these NRSE/RE1 sequences remain to be experimentally confirmed as functional (i.e., occupied by REST/NRSF in cells), they are enriched near the transcriptional start sites of genes typically expressed in neurons, consistent with the canonical role of REST/NRSF to specifically restrict the expression of genes typical of the neuronal phenotype. These neuronal genes encode ion channels, neurotransmitters, growth factors, hormones, and factors involved in axonal guidance and vesicle trafficking, as well as molecules involved in maintenance of the cytoskeleton and extracellular matrix.

NRSE/RE1 sites have also been found proximal to genes of neuronally enriched microRNAs, a class of small noncoding RNAs known to negatively regulate gene expression at the posttranscriptional level. Since these REST/NRSF-regulated microRNAs extend REST/NRSF's influence on gene regulation beyond NRSE/RE1-containing genes, their action is likely to contribute critically to the establishment of proper gene expression patterns during development and in the adult.

REST/NRSF in Development

The prevalence of NRSE/RE1 sites in the promoters of neuronal genes and the observation that REST/NRSF is most highly expressed in nonneuronal cells initially inspired the simple model that in cells in which REST/NRSF is expressed, it acts as a suppressor of the neuronal cell fate and/or silences the expression of neuronal-specific genes. However, REST/NRSF loss-of-function experiments conducted in mouse, chick, and frog revealed that REST/NRSF's role in early development is much more complex.

From these studies, it has emerged that impairment of REST/NRSF function does not always result in widespread neuronal gene expression in nonneuronal cells within the developing organism, nor does it lead to massive precocious expression of these genes in neural progenitors. Thus, the popular 'master neuronal

regulator' hypothesis of REST/NRSF does not appear to apply in early development. This is likely due to the fact that REST/NRSF-independent mechanisms act redundantly with REST/NRSF to control the timing and extent of neuronal gene expression.

However, even in cases in which REST/NRSF might be the sole repressor of a gene, de-repression of that locus by perturbation of REST/NRSF function will lead to robust transcription of that gene only when activators are available and recruited subsequent to departure of REST/NRSF. The failure to observe catastrophic gene deregulation upon REST/NRSF inactivation may therefore be due to the lack of such activator molecules. Indeed, REST/NRSF, when converted into a transcriptional activator by fusing its DNA binding domain to a powerful activator domain from a viral protein, thereby overcoming the requirement for endogenous activators, can guide the differentiation of myoblasts to a physiologically active neuronal phenotype *in vitro*. This experiment impressively demonstrates the vast network of neurogenic and neuronal genes that REST/NRSF controls. It appears that *in vivo*, however, REST/NRSF is not involved in specifying the fate of nonneuronal cells and tissues, and that de-repression of REST/NRSF-regulated genes alone is not sufficient to direct cells toward the neuronal lineage. Support for this view comes from gain-of-function experiments in the developing chick embryo. Here, forced expression of REST/NRSF does not interfere with neurogenesis, although neurons constitutively expressing REST/NRSF show axon pathfinding errors.

The master neuronal regulator hypothesis of REST/NRSF was placed under more scrutiny when NRSE/RE1 sites were discovered in several nonneuronal genes. Indeed, it is now known that aside from functioning in central nervous system development, REST/NRSF is important in the function of the fetal and adult heart, smooth muscle cells, and pancreatic islet cells. These functions are mediated, at least in part, by REST/NRSF-dependent regulation of these NRSE/RE1-containing nonneuronal genes. It is likely that roles for REST/NRSF in the development of other organs will emerge. These insights further substantiate the notion that REST/NRSF function in neurogenesis is just one aspect of this transcription factor's multifaceted role as a master regulator of a diverse set of gene expression programs.

REST/NRSF in the Adult

Several lines of evidence strongly implicate REST/NRSF in regulating the expression of NRSE/RE1-containing genes in mature neurons. Transcripts of several REST/NRSF isoforms, including full-length REST/NRSF, are detectable, albeit at low levels, in mature neurons within several regions of the adult brain, most prominently in the hippocampus. Experiments in animal disease models strongly implicate this limited amount of REST/NRSF protein in regulating gene expression responses to pathological states of the brain. Similarly, *in vitro* studies have hinted at a role for REST/NRSF in activity-dependent modulation of gene expression.

Induction of status epilepticus (seizure) by systemic administration of the glutamate analog kainate triggers an immediate transient elevation of REST/NRSF transcript and protein in the rodent hippocampus. Because the kainate-induced upregulation of REST/NRSF is preceded by upregulation of known targets of REST/NRSF, including brain-derived neurotrophic factor (BDNF) and the AMPA-responsive Ca^{2+} receptor subunit, encoded by the GluR2 gene, it has been suggested that REST/NRSF serves to dampen spiking expression levels of these genes. Indeed, deletion of the NRSE/RE1 in the BDNF promoter results in increased induction of BDNF upon kainate treatment.

Likewise, global ischemia, a brain insult associated with cardiac arrest, leads to transiently increased levels of REST/NRSF in neurons of the CA1 hippocampal subfield. These CA1 pyramidal neurons are known to eventually die following the ischemic insult, and it is thought that cell death is dependent on changes in transcription. Since several genes that have been implicated in mediating this ischemia-induced cell death, such as the μ opioid receptor 1 and GluR2, are directly regulated by REST/NRSF, it is likely that REST/NRSF is involved in controlling the timing and/or extent of cell death. These observations point to a central role of REST/NRSF in neuroprotective cellular responses. They also highlight the rather dynamic nature of REST/NRSF-dependent modulation of gene expression in the adult brain, which contrasts REST/NRSF's role as a stable silencer in most terminally differentiated nonneuronal cells.

Regulation of REST/NRSF

Despite the often dramatically different levels of REST/NRSF transcripts and protein observed in tissues and cultured cells, our knowledge regarding the transcriptional, translational, and posttranslational regulation of REST/NRSF remains limited. Experiments in the developing chick spinal cord indicate that the canonical Wnt pathway, which is critically involved in orchestrating growth and patterning of neural tissue, may directly control the expression of REST/NRSF. The Wnt-activated β-catenin/TCF complex upregulates REST/NRSF by recognizing a sequence element located within the first exon of REST/NRSF. Since this

element is conserved in human and mouse, it is conceivable that this pathway may regulate REST/NRSF expression in mammals as well. In agreement with a conserved role for Wnt in controlling REST/NRSF transcription, human embryonic carcinoma cells, when treated with constitutively active Wnt protein, demonstrate upregulation of REST/NRSF mRNA levels.

Cell culture experiments have shown a mechanistic switch in the downregulation of REST/NRSF during neuronal differentiation. When mouse-derived pluripotent embryonic stem cells are induced to differentiate into neural stem/progenitor cells, the rate of REST/NRSF transcription remains unchanged, yet REST/NRSF protein decreases substantially because it is specifically targeted for degradation by the proteasome. Upon cell cycle exit and further differentiation into mature neurons, however, REST/NRSF is repressed at the transcriptional level through the action of the unliganded retinoic acid receptor repressor complex acting on the REST/NRSF gene. These mechanisms may also regulate REST/NRSF protein levels at corresponding differentiation stages *in vivo*.

The control of subcellular localization of REST/NRSF is emerging as an additional layer of regulation of REST/NRSF activity. Experiments in a rat model of self-sustained status epilepticus show that REST/NRSF accumulates in the nuclei of CA1 hippocampal pyramidal neurons. The mechanism controlling seizure-induced REST/NRSF compartmentalization is unknown. However, a nuclear envelope protein termed RILP has been shown to specifically bind REST/NRSF and control nucleocytoplasmic shuttling of REST/NRSF, making RILP a good candidate for controlling the distribution of REST/NRSF within the cell. Lastly, the huntingtin protein, which is mutated in Huntington's disease, interacts with REST/NRSF and sequesters a portion of the cellular REST/NRSF pool within the cytoplasm. Mutant huntingtin loses its interaction with REST/NRSF, and the resultant increase in nuclear levels of REST/NRSF causes repression of NRSE/RE1-containing genes. Thus, dysregulation of REST/NRSF subcellular localization may contribute to disease phenotypes in some cases, whereas in other cases intracellular REST/NRSF redistribution may be an adaptive cellular response to counter aberrant gene expression patterns induced by pathological insults.

The Splice Variant REST4

As mentioned previously, REST4 is a splice variant of REST/NRSF that lacks the C-terminal repressor domain found in full-length REST/NRSF, and it also has a truncated DNA binding domain. Initially identified in neurons, REST4 is now known also to be expressed in many cancer cells lines. Although the precise mechanisms that regulate alternative splicing of the REST/NRSF gene are unclear, evidence suggests that protein kinase A regulates as yet unidentified splicing factors responsible for promoting the alternative splicing of the REST/NRSF transcript to produce REST4.

Despite extensive studies, the biological function of REST4 remains controversial. Although *in vitro*, REST4 has much reduced affinity to DNA compared to REST/NRSF, this residual binding activity might nevertheless be sufficient to allow REST4 to occupy DNA sites *in vivo*, especially in cells in which REST4 is highly expressed. REST4 and REST/NRSF can interact with each other directly *in vitro*, so REST4–REST/NRSF hetero-oligomerization may have a function *in vivo*, for example, by sequestering REST/NRSF in an complex that is incapable of binding to DNA. Alternatively, the REST4–REST/NRSF complex might bind NRSE/RE1 sites in a conformation that impairs its ability to recruit corepressors.

Supporting evidence for an antagonistic effect of REST4 on REST/NRSF function comes from experiments in which introduction of REST4 into cells can block REST/NRSF-dependent repression of NRSE/RE1-containing genes. Likewise, expression levels of some NRSE/RE1-containing genes in neuroblastoma- and small cell lung cancer-derived cell lines that also express REST/NRSF is inversely correlated with their levels of REST4 expression, lending further credence to the notion that REST4 might de-repress REST/NRSF-regulated genes.

REST/NRSF and Cancer

In addition to increased expression of REST4 and similar truncated forms of REST/NRSF in cell lines derived from neuroblastomas, aberrantly high levels of REST/NRSF have also been found in these cancers of the sympathetic nervous system. In keeping with a REST/NRSF-antagonistic role for REST4, the expression of the neuroendocrine character typical of these malignancies despite high levels of REST/NRSF has been suggested to be due to the de-repressive action of REST4.

REST/NRSF dysregulation is a hallmark of several other cancers. Medulloblastoma, the most common brain tumor, is thought to originate from neuroectodermal stem/progenitor cells in the cerebellum. Primary tumors and cell lines derived from them express extremely high levels of REST/NRSF, and it has been suggested that the failure to downregulate REST/NRSF in undifferentiated normal brain cells contributes to their malignant transformation. However, since REST/NRSF overexpression by itself has been

shown to be insufficient to induce tumorigenesis in several animal and cell culture models, it appears that the additional deregulated expression of oncogenes, such as c-myc, is required.

Other cancers, however, show markedly reduced and sometimes undetectable levels of REST/NRSF, including breast, colon, and small cell lung cancers. Moreover, a significant number of tumors and cell lines have deletions of the chromosomal locus encompassing the REST/NRSF gene. Consistent with this observation, REST/NRSF was found to be a candidate in a stringent unbiased screen for tumor suppressors in a mammary epithelial cells.

REST/NRSF-Mediated Transcriptional Repression and Silencing

These apparently contradictory observations make formulating a unifying hypothesis for REST/NRSF function a daunting task. Downregulation of REST/NRSF is seen both in terminally differentiated postmitotic neurons and in rapidly dividing cancer cells. Likewise, REST/NRSF has properties of a tumor suppressor but is observed to be dramatically upregulated in some cancers and is high in proliferating stem cells. The answer to these conundrums is likely to be found in the fact that different cellular environments contain different relative levels of available cofactors that REST/NRSF relies on to exert its function. It is therefore imperative that we examine more closely the proteins that REST/NRSF recruits and the chromatin changes that are induced by them (Figure 2).

In embryonic stem cells, REST/NRSF occupies NRSE/RE1 sites in promoters and introns of genes and is associated with the corepressors mSin3a and CoREST, recruited by its N-terminal and C-terminal repressor domains, respectively (Figure 1(b)). These corepressors exist in the nucleus as separate complexes, both of which harbor the class I histone deacetylases (HDACs) 1 and 2. By catalyzing the deacetylation of specific lysine residues within the N-terminal tails

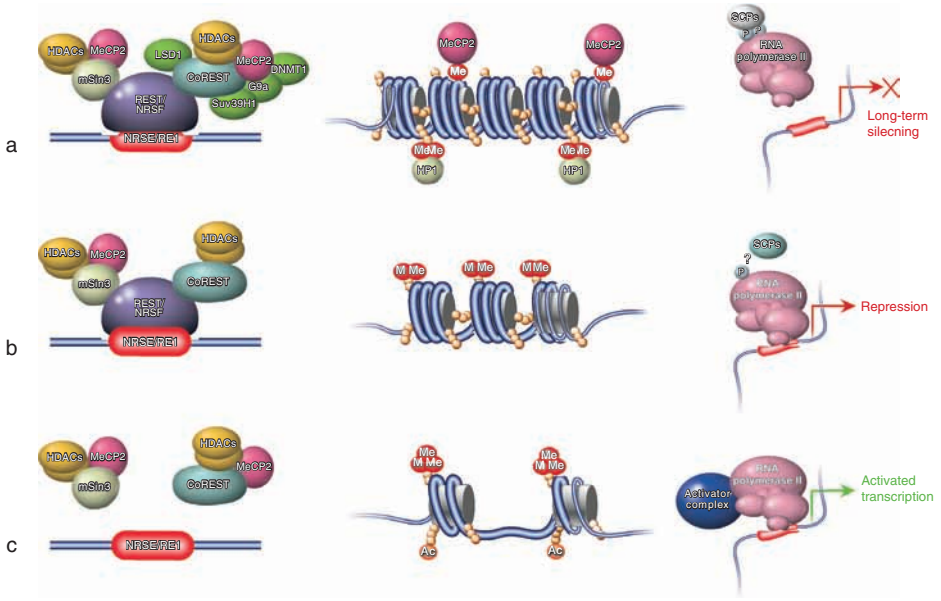

Figure 2 REST/NRSF recruits celltype-dependent complexes to gene regulatory regions, thereby altering chromatin status and transcriptional activity. (a) In terminally differentiated non-neuronal cell types and tissues, genes containing an NRSE/RE1 element are in a REST/NRSF-dependent stably silenced state. In addition to mSin3a/b-mediated recruitment of histone deacetylases (HDACs) and the methyl-DNA-binding protein MeCP2, the REST/NRSF–CoREST complex recruits chromatin modification enzymes associated with long-term silencers (green ovals) (left). These include the histone H3 lysine 9 (H3K9) methyltransferases G9a and SUV39H1, the H3K4 demethylase LSD1, and probably the DNA methyltransferase DNMT1. Methylation (red spheres) of H3K9 and DNA recruits heterochromatic protein 1 (HP1) and MeCP2, respectively. These modifications lead to heterochromatic compaction of chromatin and long-term silencing (middle). RNA polymerase II (pink) is absent from promoters and/or transcriptionally inhibited through phosphorylation (blue spheres) of its C-terminal domain (CTD) by small CTD phosphatases (SCPs) (right). (b) In multipotent embryonic stem cells and neuronal progenitors, NRSE/RE1-regulated genes are in an inactive but permissive state, characterized by the absence or low occupancy levels of long-term silencers but the presence of HDACs, which are generally considered as mediators of dynamic repression (left). Chromatin shows markers of actively transcribed genes, such as hypermethylated H3K4, but is kept in a repressed state by histone deacetylation (middle). RNA polymerase II is found at promoters, albeit at low levels, leading to limited but detectable transcript levels. The role of SCPs at this stage is unknown (right). (c) In most terminally differentiated neurons, REST/NRSF has been downregulated, leading to its departure from most NRSE/RE1-controlled genes (left). Chromatin is highly acetylated and found in a transcriptionally permissive, relaxed state (middle). Transcriptional activators recruit the RNA polymerase II complex to promoters. Downregulated nuclear SCPs levels lead to increased CTD phosphorylation levels of pol II, thereby relieving inhibition of transcription (right).

of histones, HDACs create a chromatin state that is transcriptionally repressive. Because deacetylation can be readily reversed by the action of histone acetyltransferases, this modification is generally considered transient and associated with a dynamically regulated chromatin environment.

In contrast to the situation in terminally differentiated nonneuronal cells, REST/NRSF occupancy at NRSE/RE1 sites is not accompanied by complete silencing of genes in embryonic stem cells (Figure 1(b)). Instead, this transcriptional 'twilight state' near NRSE/RE1 sites is characterized, on the one hand, by rather low levels of other epigenetic markers of repressed chromatin, such as methylated DNA, methylated lysine 9 of histone 3 (H3-meK9), and the histone methyl transferase G9a, which creates this modification. On the other hand, the histones in these regions of the genome are associated with active chromatin marks, such as H3meK4. Polymerase activity at these loci is small but detectable, and the net result of these opposing forces is a low level of transcriptional output that may be further reduced by the action of REST-recruited small CTD phosphatases, which remove regulatory phosphate marks on the C-terminal domain of RNA polymerase II and thereby interfere with transcriptional elongation. Thus, in pluripotent cells REST/NRSF maintains a chromatin signature and transcriptional state that is sufficiently repressive to prevent manifestation of the neuronal phenotype but is dynamic enough to allow de-repression upon neuronal differentiation cues.

As embryonic stem cells transition toward a neuronal lineage via a neuronal progenitor stage, REST/NRSF occupancy at NRSE/RE1 sites is gradually relieved. This process is initiated by REST protein degradation and subsequently by repression of the REST gene transcription. In most terminally differentiated neurons, REST can no longer be detected at promoters, the chromatin has changed to a permissive state, and transcription is activated (Figure 1(c)). Some gene promoters retain subsets of REST-recruited corepressors, despite REST departure from NRSE/RE1 sites, thus maintaining a responsive chromatin state that can be modulated in response to extracellular stimuli. This subset of genes is likely to be involved in the REST/NRSF-dependent regulation of neuronal plasticity described previously.

In terminally differentiated nonneuronal cell types, REST/NRSF maintains a stably repressed chromatin state by recruiting, via CoREST, not only the aforementioned HDACs but also additional complexes that confer a long-term silenced state on the surrounding chromatin (Figure 1(a)). CoREST binds to at least three histone modification enzymes: G9a and SUV39H1, which are both H3-K9 methyl transferases, and LSD1, which is an H3-K4 demethylase. The chromatin modifications created by these enzymes, in turn, serve as binding sites for repressive chromatin binding proteins, such as heterochromatic protein 1 (HP1), which binds to methylated H3-K9. CoREST also recruits the methyl-CpG binding protein 2 (MeCP2), a chromatin regulator protein whose dysfunction is implicated in several neurodevelopmental disorders, most prominently in Rett syndrome. These chromatin modifiers act in concert to maintain a stably repressed transcriptional state at NRSE/RE1-containing neuronal genes.

REST/NRSF as an Activator

The observation in embryonic stem cells that gene transcription can occur despite promoter occupancy by REST/NRSF is surprising given its canonical role as a repressor or silencer of transcription. However, several lines of evidence show that the presence of REST/NRSF at promoters is not incompatible with active transcription. In fact, REST/NRSF might even act as the activator of transcription in some contexts. In transgenic mouse experiments using a reporter gene construct, mutation of the NRSE/RE1 element of the gene that encodes the L1 cell adhesion molecule led to loss of L1 reporter expression in several adult brain structures, including cortex, striatum, and hippocampus. Likewise, expression of a β_2-nicotinic acetylcholine receptor reporter was lost in several structures of the peripheral nervous system upon introduction of a point mutation into its NRSE/RE1. Similar observations have been made in cell culture experiments in which NRSE/RE1 elements in the several other genes were disrupted.

In differentiating neural stem/progenitors cells from the adult brain, a small noncoding double-stranded RNA containing an NRSE/RE1 sequence initiates a switch in the cofactor composition of the REST/NRSF-recruited complex. This process is presumably dependent on the association of the RNA with DNA-bound REST/NRSF and is characterized by a reduction of HDAC association, concomitant with recruitment of proteins and acquisition of histone modifications characteristic of actively transcribed loci. Consequently, these genes are transcribed.

Consistent with these observations, *in vitro* experiments show that REST/NRSF can associate with factors of the basal transcription machinery as well as with the SWI/SNF chromatin remodeling complex, which is capable of remodeling chromatin into both active and repressed states. Since activator properties of REST/NRSF are most often observed in cells and brain structures in which REST/NRSF levels are relatively low, it is likely that REST/NRSF mode of function – activator, repressor, or silencer – is

dependent on its abundance in the nucleus. Clearly, additional parameters, particularly the balance of cofactors recruited by REST/NRSF, are critical in determining the net effect of REST/NRSF promoter association on transcription.

Conclusions

Since its discovery as a silencer of terminal neuronal genes more than a decade ago, REST/NRSF has come to be appreciated as a regulator of a much more complex set of gene expression programs. Advances in the knowledge of the genes targeted by REST/NRSF and the diversity of REST/NRSF-recruited cofactors have contributed to this development. It has also become clear that a quantitative approach toward studying REST/NRSF is invaluable because its function is likely to be critically dependent on both the concentration of available REST/NRSF and that of its cofactors. Large-scale screens aimed at elucidating changes in REST/NRSF promoter occupancy at target genes as well as spaciotemporally targeted REST/NRSF gain- and loss-of-function experiments during development and in the adult will undoubtedly help to build a parsimonious model of REST/NRSF function.

See also: Differentiation: the Cell Cycle Instead; Gene Expression Regulation: Chromatin Modification in the CNS; Neurogenesis in the Intact Adult Brain; Neurogenesis and Neural Precursors, Progenitors, and Stem Cells in the Adult Brain; Transcriptional Silencing.

Further Reading

Andres ME, Burger C, Peral-Rubio MJ, et al. (1999) CoREST: A functional corepressor required for regulation of neural-specific gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 96: 9873–9878.

Ballas N, Grunseich C, Lu DD, Speh JC, and Mandel G (2005) REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell* 121: 645–657.

Chen ZF, Paquette AJ, and Anderson DJ (1998) NRSF/REST is required *in vivo* for repression of multiple neuronal target genes during embryogenesis. *Nature Genetics* 20: 136–142.

Chong JA, Tapia-Ramirez J, Kim S, et al. (1995) REST: A mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* 80: 949–957.

Grimes JA, Nielsen SJ, Battaglioli E, et al. (2000) The co-repressor mSin3A is a functional component of the REST–CoREST repressor complex. *Journal of Biological Chemistry* 275: 9461–9467.

Kraner SD, Chong JA, Tsay HJ, and Mandel G (1992) Silencing the type II sodium channel gene: A model for neural-specific gene regulation. *Neuron* 9: 37–44.

Kuwabara T, Hsieh J, Nakashima K, Taira K, and Gage FH (2004) A small modulatory dsRNA specifies the fate of adult neural stem cells. *Cell* 116: 779–793.

Maue RA, Kraner SD, Goodman RH, and Mandel G (1990) Neuron-specific expression of the rat brain type II sodium channel gene is directed by upstream regulatory elements. *Neuron* 4: 223–231.

Mori N, Schoenherr C, Vandenbergh DJ, and Anderson DJ (1992) A common silencer element in the SCG10 and type II Na^+ channel genes binds a factor present in nonneuronal cells but not in neuronal cells. *Neuron* 9: 45–54.

Mori N, Stein R, Sigmund O, and Anderson DJ (1990) A cell type-preferred silencer element that controls the neural-specific expression of the SCG10 gene. *Neuron* 4: 583–594.

Mortazavi A, Thompson EC, Garcia ST, Myers RM, and Wold B (2006) Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire. *Genome Research* 16: 1208–1221.

Palm K, Belluardo N, Metsis M, and Timmusk T (1998) Neuronal expression of zinc finger transcription factor REST/NRSF/XBR gene. *Journal of Neuroscience* 18: 1280–1296.

Palm K, Metsis M, and Timmusk T (1999) Neuron-specific splicing of zinc finger transcription factor REST/NRSF/XBR is frequent in neuroblastomas and conserved in human, mouse and rat. *Brain Research Molecular Brain Research* 72: 30–39.

Paquette AJ, Perez SE, and Anderson DJ (2000) Constitutive expression of the neuron-restrictive silencer factor (NRSF)/REST in differentiating neurons disrupts neuronal gene expression and causes axon pathfinding errors *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* 97: 12318–12323.

Schoenherr CJ and Anderson DJ (1995) The neuron-restrictive silencer factor (NRSF): A coordinate repressor of multiple neuron-specific genes. *Science* 267: 1360–1363.

Watanabe Y, Kameoka S, Gopalakrishnan V, et al. (2004) Conversion of myoblasts to physiologically active neuronal phenotype. *Genes & Development* 18: 889–900.

Thalamic Mechanisms in Vision

S M Sherman, The University of Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The thalamus is a paired structure, roughly the size of a walnut, lying along the midline near the center of the brain. Thalamus proper can be divided into dorsal thalamus, which contains the relay nuclei that project to cortex, and ventral thalamus, the major part of which is the thalamic reticular nucleus (TRN). (Dorsal and ventral here represent embryonic relative locations that change during development. Unless otherwise specified, thalamus used by itself refers to the relay nuclei of the dorsal thalamus.) The TRN is a thin shell of neurons that lies generally lateral to the dorsal thalamus, like a shield, extending somewhat dorsally, ventrally, and anteriorly. A number of cytoarchitectonically distinctive zones of dorsal thalamus can be recognized, and these are the various thalamic relay nuclei. Each nucleus is reciprocally connected with cortex, and most do so with one or a small group of cortical areas. Although not divided into cytoarchitectonically clear zones, nonetheless the TRN can be functionally divided into zones, each of which is reciprocally connected with one or a small group of thalamic relay nuclei. Reticular cells target thalamic relay cells and are not known to project outside of thalamus.

An important fact to keep in mind is that effectively all information reaching cerebral cortex is relayed via thalamus. Thus, all information that reaches consciousness does so by being successfully relayed through thalamus, and if such a relay is blocked, the affected information is not consciously recognized or acted upon. An equally important fact is that, as far as we know, every cortical area receives a thalamic input. A number of thalamic nuclei innervate telencephalic targets other than cortex, examples of such targets being the basal ganglia and amygdala, but this will not be considered further here: only cerebral cortex will be regarded among thalamic targets.

The main thalamic relays of visual information are the lateral geniculate nucleus (LGN) and pulvinar. (Strictly speaking, this includes the lateral posterior nucleus and is often referred to as the lateral posterior/pulvinar complex, but for the sake of brevity, we shall refer to this simply as the pulvinar.) Until fairly recently, the LGN was viewed as a trivial, machine-like relay of retinal information to primary visual cortex, and the pulvinar was regarded as a rather mysterious entity that innervates extrastriate visual areas; one specific theory for its function is that it is somehow involved in attention. The purpose of the present account is twofold: to demonstrate that the complex cell and circuit functions of the thalamus, including the LGN, belie any simple relay functions, and instead affect the nature of information relayed in a manner that reflects behavioral state, including attention; and to suggest a relatively new hypothesis for the function of pulvinar as a central element in information transfer between cortical areas involving a cortico-thalamo-cortical route. Finally, the point shall be emphasized that all thalamic relays share most basic cell and circuit functions, but certain differences between the LGN and pulvinar identify them as members of two different kinds of relay found throughout thalamus.

The LGN

We know most about the cell and circuit properties of the LGN, particularly in the cat. The details learned from this nucleus seem broadly applicable to the rest of thalamus in all mammals. Certain differences are elaborated later. This section focuses on how retinal information to be passed on to cortex is modulated by the cell and circuit properties of geniculate relay cells.

Parallel Processing

Relay cells, at least in the LGN, are not homogeneous but can actually be divided into at least three functional classes. Each of these geniculate classes represents a thalamic link in separate streams of retino-geniculo-cortical processing; that is, there are equivalent distinct classes of retinal ganglion cells that project to the LGN, and each retinal class seems to innervate a single class of geniculate relay cell to maintain separate, parallel streams of information to cortex. In general, the receptive field properties that distinguish these cell types are similar for retina and the LGN, because geniculate receptive fields are essentially the same as their retinal inputs. These classes have been best studied in the cat, where they are called X, Y, and W cells, and in the monkey, where they are called magnocellular (M), parvocellular (P), and koniocellular (K).

The cat LGN

1. *X and Y cells*. X and Y cells are fairly homogeneous classes. Anatomically, in the retina, X cells are equivalent to the anatomical class known as 'beta,' and Y cells as 'alpha.' Compared to retinal Y cells,

X cells have smaller cell bodies with smaller dendritic arbors and thinner caliber axons. Similar relationships exist for geniculate X and Y cells, and Figure 1 summarizes some of these anatomical differences. The X cells have smaller cell bodies with thinner axons, and their dendritic arbors are elongated perpendicular to the geniculate laminar borders (see below for geniculate layers), whereas those of Y cells are organized into a roughly spherical shape. Also, X cells tend to have grape-like appendages near proximal dendritic branch points, whereas Y cell dendrites are generally smooth (see Figure 1). This is interesting, because these appendages on the X cells mark the postsynaptic target of the retinal inputs; on Y cells, retinal inputs terminate directly onto proximal dendritic shafts. The receptive fields of both cell types in retina and the LGN are organized into classic center/surround regions with roughly equal numbers of on and off center cells. However, X cells have smaller receptive fields and respond to higher spatial and lower temporal frequencies.

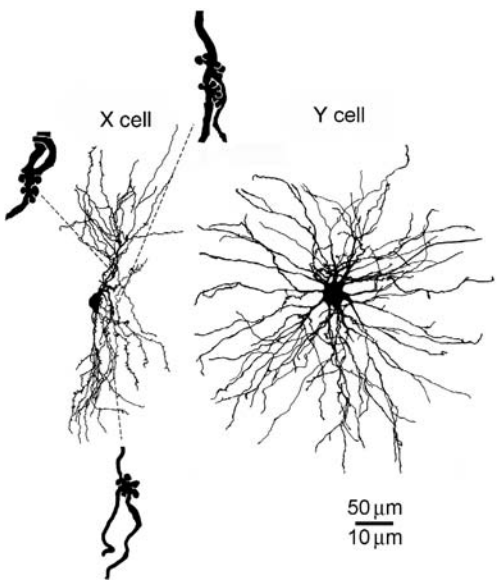

Figure 1 Tracings of an X and Y cell from the A layers of the LGN of the cat. The cells were first identified during *in vivo*, intracellular recording, and then horseradish peroxidase was iontophoresed from the recording pipette into the cells. After appropriate histological processing, the procedure provides a dense stain, allowing visualization of the morphology of the cell body and entire dendritic arbor. The dendritic arbor of the X cell is elongated and oriented perpendicular to laminar borders, while the Y cell dendrites show a stellate distribution with an approximately spherical arbor. The X cell also has prominent clusters of grape-like appendages near proximal dendritic branch points. These are hard to see in the cell reconstructions, so three examples are shown at greater magnification, with dashed lines indicating their dendritic locations (the scale is 50 μm for the cell reconstructions and 10 μm for the dendritic appendage examples). Adapted from Sherman SM and Guillery RW (2006) *Exploring the Thalamus and Its Role in Cortical Function*, 2nd edn. Cambridge, MA: MIT Press.

Based on receptive field properties, hypotheses have been developed for the distinct function of the X and Y pathways. X cells are thought to provide maximum acuity for detail vision, while Y cells are more important for motion detection and processing of low spatial frequencies.

2. *W cells.* W cells are probably a heterogeneous group with several distinct classes, but for convenience and because the final classification and functional correlates of W cells remains to be done, they are treated together here. More details of these cells can be found elsewhere (see 'Further reading'). Retinal W cells generally have small- to medium-sized cell bodies with long, sparsely branched dendrites, but the morphological features of this group are so varied that any generality must be considered limited. W cells so far described in the LGN have medium-sized cell bodies and dendritic arbors oriented parallel to the layers. The receptive fields of these cells, both in retina and the LGN, are also quite varied but tend to be large and poorly responsive (some have named them 'sluggish'): some have center/surround configuration; others have poorly defined borders with on/off responses throughout, some have directional selectivity, and some have limited wavelength sensitivity. To date, there has not been much speculation regarding the function of the W pathway(s), which remains a mystery.

The monkey LGN Details of the organization of the parallel pathways in the monkey are less well understood than those in the cat, but the same general principles apply. At least three parallel pathways start in the retina and are relayed separately through the LGN. These have been called the P, M, and K pathways. Relationships of these pathways to those in the cat has been a source of confusion and controversy, but a consensus seems to be developing that the P, M, and K cells and pathways are homologous, respectively, to the cat's X, Y, and W. However, homology does not mean functional equivalence, only that a common ancestral form is shared, and, indeed, evolution works to create functional differences between homologous structures. An example is our fingers and the wings of bird: these are homologous structures that have evolved quite different functions. In contrast, the fins of a fish and whale are examples of nonhomologous structures that have evolved similar functions.

1. *P and M cells.* In the retina, P cells (called midget cells) are smaller than M cells (called parasol cells), and this size differential also holds in the LGN,

as the names (parvocellular and magnocellular) imply. In both structures, their receptive fields are organized into center/surround configurations, but P cells have smaller receptive fields. M cells are much more sensitive to luminance contrast and moving stimuli, but, while M cells show no wavelength sensitivity, P cells show sensitivity for green and red wavelengths. One exception to this is the owl monkey, which is crepuscular and thus not so reliant on color vision. For these animals, P cells show little wavelength sensitivity.

As is the case for the cat, there has been much speculation concerning the role of these cell types in the monkey. Common suggestions are that P cells are important for color discrimination, especially for red/green distinctions, and may also be involved in high-acuity vision, whereas M cells provide for better luminance contrast sensitivity and are also important for motion detection. Naturally, the role of P cells in monkeys without much color vision, such as the owl monkey, would be different. This would be another example of evolution working on homologous systems to create different functions.

2. *K cells*. Like W cells, K cells probably include many distinct cell classes but are grouped together here, because their complete classification remains to be done. As the name (koniocellular) implies, these are smaller than M or P cells in both the retina and the LGN. Not much is known of their receptive field properties, but some of these cells are thought to be responsible for yellow/blue-wavelength discrimination.

Laminar Relationships

Lateral geniculate nucleus The LGN in all mammals so far studied has distinct zones that in cats and monkeys have evolved to separate layers, and each of these layers receives an input from one or the other eye. Geniculate laminar patterns vary greatly among species, but this ocular division between sets of layers seems to be one constant.

In addition to ocular dominance, the various cell types are distributed with varying levels of laminar specificity. For instance, in the cat, X and Y cells commingle in the dorsal two layers (called the A layers), and the next ventral layer (layer C) has only Y cells, and the ventral few layers have only W cells. For the rhesus monkey, P and M cells are completely separated into four dorsal parvocellular layers and two ventral magnocellular layers. The K cells are found in all the interlaminar zones but also extend dorsally to commingle with P cells in the ventral portions of the parvocellular layers.

Figure 2 summarizes the laminar patterns for cats, monkeys, and some other species to illustrate the sort of bewildering variation in patterns present. Note that two closely relayed carnivore species like the cat versus the ferret or mink differ in that the cat has its on- and off-center cells mixed together in the layers, while the ferret and mink have further divided the dorsal two layers into four to separate on and off center cells. The bottom line here is that geniculate lamination does correlate with cell type, but the nature and extent of this correlation varies widely across species, and it is difficult to discern any special significance to these correlations.

Visual cortex There is also a laminar correlation regarding the target zones of the various cell types. In the cat, geniculate X cell axons innervate the ventral part of layer 4, while those of Y cells innervate the upper part. Most W cells innervate layer 3. A similar arrangement holds for the monkey: P cells innervate the ventral half of layer 4 (sometimes called layer 4Cβ), M cells innervate the dorsal half of layer 4

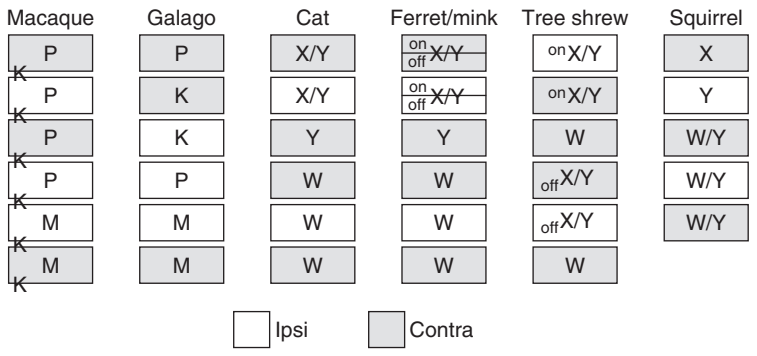

Figure 2 Schematic rendering of layers and distribution of cell types in the LGN of six different species. The shaded boxes show layers innervated by the contralateral eye (contra), and the unshaded boxes show the input from the ipsilateral eye (ipsi). The X, Y, and W pathways are shown for cat, ferret, mink, tree shrew, and squirrel, and the magnocellular (M), parvocellular (P), and koniocellular (K) pathways are shown for macaque and galago. Where there is a clear segregation of on- and off-center (on, off) cells, this is also indicated. Adapted from Sherman SM and Guillery RW (2006) *Exploring the Thalamus and Its Role in Cortical Function*, 2nd edn. Cambridge, MA: MIT Press.

(sometimes called 4Cα), and K cells mostly innervate layer 3. Thus, through the first stage of processing, the three parallel pathways are kept fairly independent through this first stage of cortical processing, although what happens further centrally with regard to these pathways is not at all clear.

Thalamic Cell Types

There are three basic thalamic cell types related to the LGN: relay cells (which in the LGN can be further divided into W, X, and Y; K, P, and M; etc.), interneurons, and cells of the TRN. Relay cells are glutamatergic, and the latter two cell types are GABAergic, providing a major inhibitory input to relay cells. Interneurons are located within the main geniculate layers, intermixed with relay cells, and the ratio of relay cells to interneurons is roughly 3:1. This ratio is similar throughout thalamus in all mammals with a peculiar exception. That is, outside of the LGN, the thalamus of rats and mice are essentially devoid of interneurons, but, curiously, the LGN in these animals does have a normal complement of interneurons. This is not a property of rodents, because hamsters, guinea pigs, squirrels, etc., have interneurons throughout thalamus. However, this point has been questioned recently by evidence that the lateral posterior nucleus of the rat has a substantial fraction of interneurons.

Cell Properties

All thalamic relay cells, like cells throughout the central nervous system, exhibit a number of voltage-gated ionic membrane conductances. Geniculate relay cells serve as a typical example. The best known of these ubiquitous conductances is the voltage-gated Na^+ conductance underlying the action potential. Other examples include various voltage-gated K^+ and Ca^{2+} conductances. A detailed listing of these is beyond the scope of this article. The presence of these conductances means that membrane voltage, which determines if and when each of these becomes activated, plays an important role in cell excitability and thus the gating of retinogeniculate information flow. Most of these conductances are garden-variety properties found in neurons everywhere, but one in particular, a voltage-gated Ca^{2+} conductance that operates via T-type Ca^{2+} channels, is worth further consideration.

When T-type Ca^{2+} channels are open, Ca^{2+} flows into the cell, depolarizing it via a current known as I_T. However, the T channel can only be opened by depolarization from a relatively hyperpolarized level. That is, depolarization initially activates the channel, but after about 100 ms or so, inactivates it. Once inactivated, it takes about 100 ms or so of subsequent hyperpolarization to remove the inactivation of (i.e., de-inactivate) IT, so that it can again be activated. Actually the opening or closing of the inactivation gate is a complex function of voltage and time so that the more depolarized (or hyperpolarized) the more quickly the gate closes (or opens), but the important point is that under normal conditions, roughly 100 ms is required for these activations.

Once activated, however, the depolarization initiated by I_T becomes an all-or-none Ca^{2+} spike, much like the conventional Na^+/K^+ action potential, and it propagates throughout the dendrites and soma. This is often known as the low-threshold spike, which is tens of millivolts in amplitude and thus large enough to evoke a cluster of action potentials that rides its crest; typically, there are 2–10 such action potentials with interspike intervals <4 ms or so. Such burst firing is ubiquitous for relay cells: every relay cell of every thalamic nucleus of every mammalian species so far tested shows this property. T channels, however, are found in neurons throughout the brain, but what makes thalamic relay cells relatively unique is the density of T channels they express: in relay cells, this density is high enough to support an all-or-none Ca^{2+} spike, while in most other neurons, T channel density is too low for this and can thus underlie only a graded depolarization.

From the above, the voltage-dependent properties of the T channel can be characterized by two gates, an activation gate and an inactivation gate. The activation gate opens with depolarization and closes with hyperpolarization, and it does so with very fast kinetics, typically requiring <1 ms for opening or closing. The inactivation gate has the opposite voltage relationships, since it opens with hyperpolarization and closes with depolarization; also, it operates much more slowly, taking about 100 ms for either action. This slow action is important to how the channel is controlled by thalamic circuitry, which is considered below.

Figure 3 illustrates and summarizes many of these features of the T channels. Activation of T channels requires that they be first de-inactivated, which in turn requires hyperpolarization. The more hyperpolarization that occurs, the more T channels are de-inactivated and can thus participate in the subsequently evoked low-threshold spike, and the larger the evoked low-threshold spike. This is shown in Figure 3(a) for two relay cells. Thus, the size of the low-threshold spike can vary depending on the history of the cell's membrane voltage, and, as might be predicted, the number of conventional action potentials evoked from such a low-threshold spike is correlated with its amplitude (Figure 3(a)). However, for any given level of T channel de-inactivation, the low-threshold spike is all-or-none (Figure 3(b)).

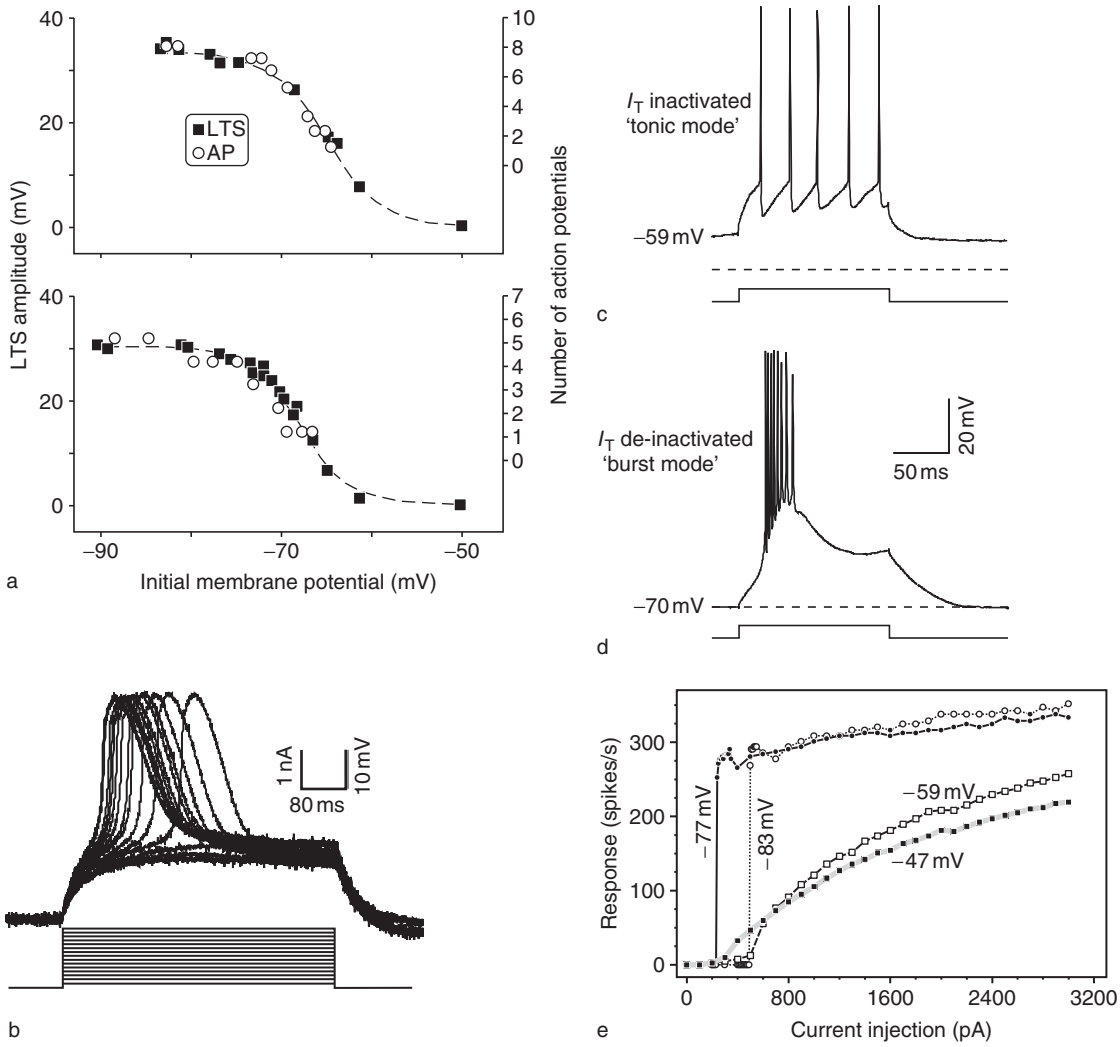

Figure 3 Properties of I_T and the low-threshold Ca^{2+} spike (LTS). All examples are from relay cells of the cat's LGN recorded intracellularly in an *in vitro* slice preparation. (a) Voltage dependency of amplitude of LTS and extent of burst response. Examples for two cells are shown. The more hyperpolarized the cell before being activated (initial membrane potential), the larger the LTS (filled squares and dashed curve) and the more action potentials (APs) in the burst (open circles). The number of APs were first measured, and then tetrodotoxin (TTX) was applied to isolate the LTS for measurement. (b) All-or-none nature of the LTS measured in the presence of TTX in another geniculate cell. The cell is initially hyperpolarized, and current pulses were injected starting at 200 pA amplitude and incremented in 10 pA steps. Smaller (subthreshold) pulses led to pure resistive–capacitative responses, but all larger (suprathreshold) pulses led to an LTS. Much like conventional APs, the LTSs are all the same amplitude regardless of how far the depolarizing pulse exceeded activation threshold, although there is latency variability for smaller suprathreshold pulses. (c, d) Voltage dependency of the LTS. Responses are shown to the same depolarizing current pulse delivered intracellularly but from two different initial holding potentials (–59 and –70 mV). When the cell is relatively depolarized (c), I_T is inactivated, and the cell responds with a stream of unitary action potentials as long as the stimulus is suprathreshold for firing. This is the tonic mode of firing. When the cell is relatively hyperpolarized (d), I_T is de-inactivated, and the current pulse activates an LTS with eight action potentials riding its crest. This is the burst mode of firing. (e) Input–output relationship for one cell. The input variable is the amplitude of the depolarizing current pulse, and the output is the firing frequency of the cell. To compare burst and tonic firing, the firing frequency was determined by the first six APs of the response, since this cell usually exhibited six APs per burst in this experiment. The initial holding potentials are shown, and –47 and –59 mV reflect tonic mode, whereas –77 and –83 mV reflect burst mode. (a) Redrawn from Zhan XJ, Cox CL, and Sherman SM (2000) Dendritic depolarization efficiently attenuates low threshold calcium spikes in thalamic relay cells. *The Journal of Neuroscience* 20: 3909–3914. (e) Redrawn from Zhan XJ, Cox CL, Rinzel J, and Sherman SM (1999) Current clamp and modeling studies of low threshold calcium spikes in cells of the cat's lateral geniculate nucleus. *The Journal of Neurophysiology* 81: 2360–2373.

Properties of burst and tonic firing T channel activity determines two very different response modes, tonic or burst, for the relay cell. This has important implications for relay of information for a number of reasons. Note that the only signal relayed to cortex is via conventional action potentials. However, the same retinal excitatory postsynaptic potential will evoke a very different response in the relay cell dependent

on the cell's response mode (see Figures 3(c) and 3(d)). During tonic firing mode, the action potentials rise directly from the excitatory postsynaptic potential, and thus a larger excitatory postsynaptic potential evokes a higher rate of firing. However, the action potentials during burst firing mode arise indirectly from the excitatory postsynaptic potential. Here, the excitatory postsynaptic potential evokes the low-threshold spike, which is all-or-none, and thus a larger excitatory postsynaptic potential will not evoke a larger low-threshold spike, meaning that the size of the burst is not linearly related to the size of the retinal excitatory postsynaptic potential. What this all means is that there is a fairly linear input/output relationship during tonic firing, but this relationship is very nonlinear during burst firing (Figure 3(e)).

The advantage for tonic firing is pretty clear: if cortex is to reproduce faithfully the visual scene, the sort of nonlinear distortion seen during burst firing would hamper this process. What, then, is the advantage for burst firing? There are at least two of them. First, burst firing is associated with low spontaneous activity but the burst itself represents a high rate of firing; together, this means that burst firing is associated with a higher signal-to-noise ratio and thus better stimulus detectability. Second, the combination of the nature of the thalamocortical synapse and the different pattern of action potentials between firing modes leads to a much greater postsynaptic response in cortex for burst firing. Two related observations support this second point. First, there is evidence that the first action potential in a burst evokes a much greater response in cortex than does a tonic action potential. Second, many cortical cells operate in an 'up' or 'down' state in which up state is relatively depolarized and responsive to thalamic input, whereas the more hyperpolarized down state is much less responsive to thalamic input; recent evidence indicates that tonic firing is relatively ineffective in switching a cortical cell from the down to the up state, but burst firing brings about this switch quite readily. The overall implication is that while tonic firing is better for stimulus reconstruction, burst firing is better for the detection of novel stimuli, and associated with this improved detectability is a much stronger cortical response.

Hypothesis for burst and tonic firing This difference between burst and tonic firing has led to the following hypothesis. Because burst firing is better for stimulus detection, it could be the mode more often seen when the receptive field of the geniculate neuron covers an unattended part of the visual field, unattended because the subject is drowsy, is attending to another sensory modality or part of the visual field, etc.; under these conditions, the burst evoked by a novel stimulus would more likely get through to cortex and be recognized as a signal than if the cell were firing tonically. In this sense, the burst is a sort of 'wake-up call' that something has changed, and that something should be attended to so that its importance can be evaluated. The idea here is not necessarily that bursting provides a stronger overall signal to cortex than does tonic firing, but rather that bursting overcomes any disadvantage regarding stimulus detectability or cortical activation imposed by inattention. Once the burst activates cortical circuits, the relay cell would then switch to tonic firing so that the novel stimulus can be properly evaluated.

Consistent with this hypothesis are several observations. Both tonic and burst firing are seen in behaving animals and humans, with switching occurring between modes. Bursting is seen relatively rarely in the alert animals and more commonly in drowsy animals. Studies of behaving cats show that a relatively common pattern of response of geniculate cells to a repeating stimulus (e.g., a flashing spot) would be to respond to the first cycle of stimulation, then switch to tonic firing in response to subsequent cycles. Finally, recent receptive field studies show that the type of visual stimulus most likely to evoke a burst is one that switches from an inhibitory one to an excitatory one, such as a dark region covering the receptive field of an on-center cell that is then switched to a bright spot; this indicates that the burst signals a significant change in the form of a novel stimulus just appearing. However, these observations, while supporting the hypothesis, do not constitute proof of its validity. Much more research will be needed to accept or reject the hypothesis.

Other cellular properties of relay cells While much is known about T channels and their significance for thalamic relays, many other voltage-sensitive ionic currents exist that affect these relays. These include a variety of voltage-gated K^+, Na^+, and other Ca^{2+} channels and also a hyperpolarization-activated cation channel known as I_h. While we understand much of the voltage-sensitive properties of each of these, much more work needs to be done to appreciate how these interact among themselves and also with I_T.

Circuit Properties

Figure 4 schematically shows the main inputs to geniculate relay cells. This circuitry is generally repeated, with some variations, throughout thalamus.

Anatomical features An important point shown in Figure 4 is that retinal input is one of the several inputs to relay cells. Other prominent input derives

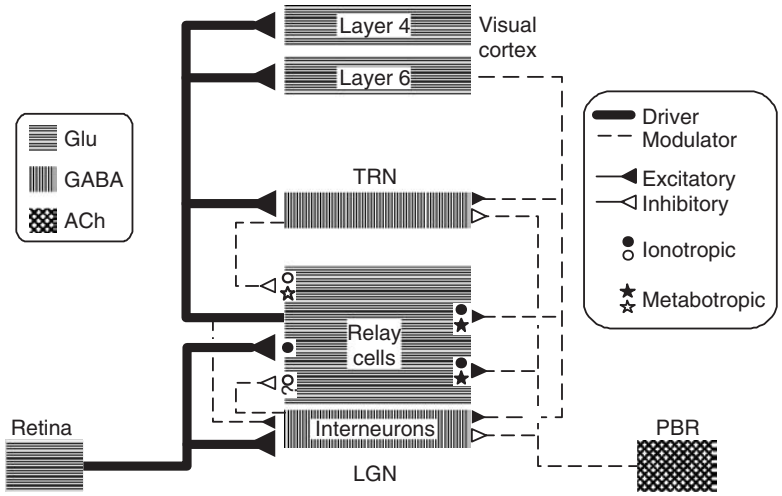

Figure 4 Circuitry of the LGN with related receptors present on relay cells. Other thalamic nuclei seem to be organized along the same pattern. The key to the left indicates the major transmitter systems involved. The retinal input activates only ionotropic receptors (circles), whereas all nonretinal inputs activate metabotropic receptors (stars) and often ionotropic receptors as well. The question mark related to input from interneurons indicates uncertainty whether metabotropic receptors are involved. Thick solid and thin dashed lines indicate driver and modulator inputs, respectively. Filled and open icons for synaptic terminals indicate excitatory and inhibitory inputs, respectively. ACh, acetylcholine; GABA, γ-aminobutyric acid; Glu, glutamate; LGN, lateral geniculate nucleus; PBR, parabrachial region of the brain stem; TRN, thalamic reticular nucleus. Adapted from Sherman SM and Guillery RW (1998) On the actions that one nerve cell can have on another: Distinguishing 'drivers' from 'modulators'. *Proceedings of the National Academy of Sciences of the United States of America* 95: 7121–7126 and Sherman SM and Guillery RW (2001) *Exploring the Thalamus*. San Diego: Academic Press.

from local GABAergic sources (interneurons and reticular cells), layer 6 of cortex, which provides a feedback, glutamatergic input, and ascending input from the brain stem, mostly from cholinergic cells in a midbrain area known as the parabrachial region. (Another term often applied to this area is 'pedunculopontine tegmental nucleus.' We prefer 'parabrachial region,' because the cells that innervate thalamus from this area do not have a clear nuclear boundary, and they are found scattered around the brachium conjunctivum.) Not shown for simplicity are other small inputs also seen widely throughout thalamus, such as serotonergic input from the dorsal raphé nucleus of the brain stem, noradrenergic inputs from the parabrachial region, and histaminergic inputs from the tuberomamillary nucleus of the hypothalamus. However, we focus here on the inputs shown in Figure 4, because these are the best understood.

Thus, the main extrinsic, nonretinal inputs to geniculate relay cells derive from cortex and brain stem. There are also local, GABAergic inputs, but these are also innervated by the same cortical and brain stem sources that innervate relay cells. Thus, these extrinsic inputs can affect relay cells directly or indirectly via local GABAergic circuitry.

Functional features

1. *Ionotropic and metabotropic receptors*. Figure 4 also shows the neurotransmitters used by various afferents to relay cells as well as the postsynaptic receptors involved. These receptors are critical in understanding how this circuit works, because the inputs to relay cells operate via conventional chemical synapses, meaning that they release transmitters that affect the relay cell via postsynaptic receptors. The receptors, which determine the final postsynaptic response of the relay cell, come in two main flavors: ionotropic and metabotropic. Examples of the relevant ionotropic receptors are α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) and *N*-methyl-D-aspartate (NMDA) for glutamate, nicotinic for acetylcholine, and the $GABA_A$ receptor (GABA, γ-aminobutyric acid). For metabotropic receptors, examples are various metabotropic glutamate receptors, various muscarinic receptors for acetylcholine, and the $GABA_B$ receptor.

Differences between ionotropic and metabotropic receptors are many, and only certain ones are considered below. Ionotropic receptors are simpler in construction and function, and the receptor itself usually contains the ion channel it controls. Generally, the binding of the transmitter to the receptor changes the shape of the receptor, exposing and thus opening the ion channel. Opening the channel allows ions to flow into or out of the cell, leading to an excitatory or inhibitory postsynaptic potential. Typically, the postsynaptic potential occurs with a brief latency (<1 ms) and duration (mostly over in 10 ms or a few tens of milliseconds). The story with metabotropic receptors is more complicated.

Here, the receptor is only indirectly linked to the relevant ion channel. Typically, binding of the transmitter to the receptor causes release of a G-protein that interacts with an effector protein, setting off a cascade of biochemical reactions with many effects on the cell, one of which ultimately is on the ion channel. For thalamic relay cells, these are typically K^+ channels. Thus when the channel opens, K^+ flows out of the cell, leading to an inhibitory postsynaptic potential; when the channel closes, leakage of K^+ is stopped, leading to an excitatory postsynaptic potential. However, these postsynaptic potentials with metabotropic receptors typically have a long latency (~10 ms or so) and duration (hundreds of milliseconds to several seconds).

Given these differences between receptor types, particularly with regard to duration of the resultant postsynaptic potential, it is particularly interesting to note the pattern of postsynaptic receptors as they relate to various inputs onto relay cells (**Figure 4**). What is striking is that retinal inputs activate only ionotropic (mostly AMPA but also NMDA) receptors, whereas all nonretinal inputs activate metabotropic as well as ionotropic receptors. One obvious conclusion is that the nonretinal inputs, by activating metabotropic receptors, can achieve sustained changes in membrane potential and thus relay cell excitability, thereby modulating the gain of retinogeniculate transmission.

There is a further, more interesting implication of this pattern that involves control of the T channels. Recall that to inactivate the T channel (i.e., to close the inactivation gate) requires 100 ms or so of sustained depolarization; similarly, to de-inactivate it (i.e., to open the inactivation gate) requires 100 ms or so of sustained hyperpolarization. This means that the fast excitatory or inhibitory postsynaptic potentials seen with ionotropic receptors are ill suited to affect the inactivation gate; even action potentials, despite their amplitude, are terminated too quickly to effectively inactivate the T channel. In contrast, the sustained postsynaptic potentials of metabotropic receptors are ideally suited to inactivate or de-inactivate T channels. It thus follows that retinal input by itself, with its fast excitatory postsynaptic potentials, is less likely to directly affect T channels. This seems appropriate in the sense that burst or tonic firing is thought to be dependent on behavioral state.

One would expect that to be mainly the function of inputs to relay cells that do not carry the main information to be relayed. However, the fast retinal excitatory postsynaptic potentials means that each can represent a specific retinal action potential up to high rates of firing, thus preserving temporal information. Put another way, if retinal inputs activated metabotropic glutamate receptors, the resultant long excitatory postsynaptic potentials would act like low-pass temporal filters, and higher-frequency temporal information would fail in the relay. From this reasoning, it falls to the nonretinal inputs via activation of metabotropic receptors to provide the main control of T channels and determine whether the response to retinal input is in tonic or burst firing mode.

In addition, there is clear evidence that visual stimuli can also affect firing mode: for instance, a visual stimulus that inhibits a geniculate cell for a sufficient time (e.g., a dark stimulus falling on the center of an on-center cell) can de-inactivate the T channels, and when this stimulus is replaced by an excitatory one (e.g., a bright spot), a burst is evoked. Note that this is likely due to inhibitory circuits involving interneurons or reticular cells, or both, and perhaps also involving $GABA_B$ receptors and is not likely to represent retinal inputs alone.

Indeed, evidence exists that metabotropic glutamate receptors activated from layer 6 of cortex and muscarinic receptors activated from the parabrachial region produce long, slow excitatory postsynaptic potentials that inactivate the T channels and switch relay cells from burst to tonic firing mode. Similarly, activation of $GABA_B$ receptors from reticular inputs does the opposite: it produces a sustained inhibitory postsynaptic potential that switches firing modes from tonic to burst. Interneurons may also participate in this, but as indicated by the question mark in **Figure 4**, it is not yet known whether or not these inputs activate $GABA_B$ receptors on relay cells.

2. *Role of parabrachial and cortical inputs.* Further inspection of **Figure 4** shows that, as parabrachial inputs become more active, relay cells become more depolarized and responsive. This is because the main effect of parabrachial input onto relay cells is depolarizing, via activation of nicotinic and M1 muscarinic receptors (activations of M1 receptors closes K^+ channels), and its main effect onto the local GABAergic cells is hyperpolarizing, via activation of another (M2?) muscarinic receptor that opens K^+ channels. Thus, parabrachial input directly depolarizes relay cells and also disinhibits them. Not only does parabrachial input excite relay cells, but the resultant depolarization also inactivates T channels, biasing relay cell responses to tonic mode. Consistent with this scenario is the further observation that parabrachial neurons become more active and relay cells become less bursty with increasing vigilance, from slow-wave sleep through drowsiness to full attention.

Understanding the consequence of the layer 6 cortical input is much more difficult. **Figure 4** suggests that this input directly excites relay cells while it indirectly inhibits them, but, in fact, the actual effect of this input depends on details of the circuit that are generally unknown. This is illustrated in **Figure 5**,

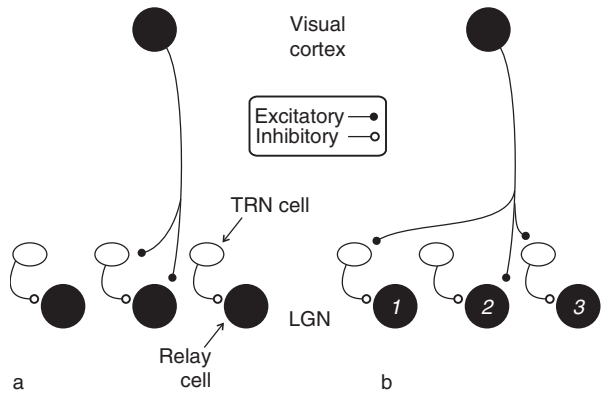

Figure 5 Two patterns among others possible for corticothalamic projection from layer 6 to cells of the TRN and geniculate relay cells. (a) Pattern of simple excitation and feedforward inhibition. (b) More complicated pattern in which activation of a cortical axon can excite some relay cells directly and inhibit others through activation of reticular cells. For further details, see text. LGN, lateral geniculate nucleus; TRN, thalamic reticular nucleus.

which shows two variants of the many possible for the relevant circuit. (The example of Figure 5 includes just reticular cells, but one can easily imagine a similar circuit involving interneurons.) In Figure 5(a), individual corticogeniculate axons innervate a reticular cell and relay cell, with the reticular cell innervating the same relay cell. The consequence of increased corticogeniculate activity might be of little or no effect on the relay cell's membrane voltage (and T channel inactivation or de-inactivation) if the excitatory and inhibitory inputs are balanced. However, while this may not affect membrane voltage, the increased synaptic conductance among other factors will reduce relay cell excitability to other (i.e., retinal) inputs, and thus reduce the gain of retinogeniculate transmission. Thus, the more active the corticogeniculate inputs in the circuit of Figure 5(a), the lower the gain of retinogeniculate transmission.

The circuit of Figure 5(b) has very different consequences. Here, activation of the corticogeniculate axon purely excites one or a few relay cells (e.g., cell 2) and purely inhibits others (e.g., cells 1 and 3). Also, note that the final effect on relay cell membrane voltage is such that activation of the corticogeniculate axon would promote tonic firing in cell 2 and burst firing in cells 1 and 3. This means that layer 6 corticogeniculate can have very different and localized effects. Figure 5 illustrates the importance of a much better understanding of these functional circuits at the single-cell level than we have at present.

Drivers and Modulators

Figure 4 also partly demonstrates a fundamental observation, namely, that all inputs to geniculate relay cells are not equal. That is, the retinal input alone represents the main information actually relayed to cortex. Among other factors, a consideration of receptive field properties demonstrates this fact, because the responses of the relay cell to visual stimulation does, in fact, identify the information relayed. In this regard, the receptive fields of geniculate relay cells are remarkably like those of their retinal afferents, having the same center/surround configuration and with only minor, subtle differences. If we consider the other extrathalamic inputs, receptive fields of corticogeniculate afferents, which show selectivities for orientation and often direction typical of cortical cells, are utterly unlike geniculate receptive fields, and parabrachial inputs are not plausible sources of such clear center/surround properties. If it is the retinal input that provides the information to be relayed, then the nonretinal inputs must have another function. This, plus a number of morphological, pharmacological, and physiological differences that distinguish retinal and nonretinal afferents to relay cells, has led to the idea that these can be functionally divided: the retinal inputs are the drivers (so called because one of its properties is the very strong postsynaptic drive of their target relay cells; see Table 1), while all the nonretinal inputs are the modulators, the idea being that the driver input is the information-bearing input, while the modulators serve to modulate retinogeniculate transmission. Modulation can take many forms, including, for example, the above-mentioned consequences of metabotropic receptor activation that lead to overall changes in relay cell excitability and that serve to control the tonic/burst transition; also, as indicated above, the circuit suggested by Figure 5(a) can operate to control the gain of retinogeniculate transmission.

There is actually an extensive list of properties that distinguish drivers from modulators in thalamus, and this list will likely increase in size as we learn more about this issue. Table 1 summarizes the most important features but is not meant to be exhaustive. Layer 5 drivers (the second column in Table 1) are considered below. The 13 criteria in Table 1, in a roughly decreasing order of importance, are as follows:

1. As already noted, drivers determine the main receptive field properties of the relay cell; modulator input does not.
2. Also as already noted, drivers activate only ionotropic receptors; modulators activate metabotropic as well as ionotropic receptors.
3. Drivers evoke very large excitatory postsynaptic potentials; modulators generally evoke much smaller excitatory or inhibitory postsynaptic potentials.

Table 1 Drivers and modulators in LGN plus layer 5 drivers

Criteria	*Retinal (driver)*	*Layer 5 to HO (driver)*	*Modulator: layer 6*	*Modulator: PBR*	*Modulator: TRN & Int*
Criterion 1	Determines relay cell receptive field	[a]Determines relay cell receptive field	Does not determine relay cell receptive field	Does not determine relay cell receptive field	Does not determine relay cell receptive field
Criterion 2	Activates only ionotropic receptors	Activates only ionotropic receptors	Activates metabotropic receptors	Activates metabotropic receptors	TRN: Activates metabotropic receptors; Int:[b]
Criterion 3	Large EPSPs	Large EPSPs	Small EPSPs	[b]	TRN: small IPSPs; Int:[b]
Criterion 4	Large terminals on proximal dendrites	Large terminals on proximal dendrites	Small terminals on distal dendrites	Small terminals on proximal dendrites	Small terminals; TRN: distal; Int: proximal
Criterion 5	Each terminal forms multiple contacts	Each terminal forms multiple contacts	Each terminal forms single contact	Each terminal forms single contact	Each terminal forms single contact
Criterion 6	Little convergence onto target	[a]Little convergence onto target	Much convergence onto target	[b]	[b]
Criterion 7	Very few synapses onto relay cells (~5%)	Very few synapses onto relay cells (~5%)	Many synapses onto relay cells (~30%)	Many synapses onto relay cells (~30%)	Many synapses onto relay cells (~30%)
Criterion 8	Often thick axons	Often thick axons	Thin axons	Thin axons	Thin axons
Criterion 9	Glutamatergic	Glutamatergic	Glutamatergic	Cholinergic	GABAergic
Criterion 10	Synapses show paired-pulse depression (high *p*)	[a]Synapses show paired-pulse depression (high *p*)	Synapses show paired-pulse facilitation (low *p*)	[b]	[b]
Criterion 11	Well localized, dense terminal arbors	Well localized, dense terminal arbors	Well localized, dense terminal arbors	Sparse terminal arbors	Well localized, dense terminal arbors
Criterion 12	Branches innervate subtelencephalic targets	Branches innervate subtelencephalic targets	Subcortically known to innervate thalamus only	[b]	Subcortically known to innervate thalamus only
Criterion 13	Innervates dorsal thalamus but not TRN	Innervates dorsal thalamus but not TRN	Innervates dorsal thalamus and TRN	Innervates dorsal thalamus and TRN	TRN: both; Int: dorsal thalamus only

[a]Very limited data to data.
[b]No relevant data available.

4. Drivers form very large terminals on proximal dendrites; modulators usually form small terminals, and these can be on proximal or distal dendrites.
5. Each driver terminal forms multiple large synapses; each modulator terminal usually forms a single, small synapse.
6. Driver inputs show little convergence, meaning, for example, that one or a small number of retinal axons converge to innervate each geniculate relay cell; where evidence is available, modulator inputs show considerable convergence.
7. Driver inputs produce a small minority (~5%) of the synapses onto relay cells; many modulator inputs produce larger synaptic numbers (e.g., the cortical and parabrachial modulator inputs in Figure 4 each produces about 30% of the synapses).
8. Drivers have thick axons; modulators have thin axons.
9. Drivers are glutamatergic; modulators can use a variety of neurotransmitters.
10. Driver synapses show high release probability and paired-pulse depression, meaning that a given action potential is likely to result in transmitter release, and that, with the initiation of a train of action potentials, there is a period after each evoked postsynaptic potential lasting for tens of milliseconds that the next one will be smaller (depressed); modulator synapses that have been tested so far show the opposite properties of low release probability and paired-pulse facilitation.
11. Driver terminal arbors are well localized with a dense array of terminals; modulator terminal arbors can be either well localized and dense or relatively poorly localized and sparse.
12. Branches of driver axons tend to innervate extrathalamic targets as well as thalamus (e.g., many or all retinogeniculate axons branch and also

innervate midbrain targets); those modulator inputs so far tested innervate thalamus only.

13. Driver inputs innervate relay cells and interneurons in dorsal thalamus but do not innervate the TRN; modulator inputs innervate relay cells, interneurons, and reticular cells.

This driver/modulator distinction is clear not just in the LGN but also in other thalamic relays for which sufficient information is available, such as the ventral portion of the medial geniculate nucleus (the primary auditory thalamic relay) and the ventral posterior nucleus (the primary somatosensory thalamic relay). The main point, again, is that not all anatomical pathways are functionally equivalent, acting in some sort of anatomical democracy, and if one is to understand the functional organization of the thalamus and what it is that is being relayed, one must identify and characterize the driver input. This may also apply outside of thalamus.

This point can be underscored with a consideration of criterion (7) above. It may seem surprising at first that driver input, which provides the main information to be relayed, is responsible for such a small percentage of the synapses onto relay cells, but there are two factors to be considered. First, despite the small number of inputs numerically, these are especially powerful and do manage to drive the relay cells effectively. Second, it may only take a relatively small number of synapses to get the basic information to the relay cell, provided these are powerful synapses, but to provide a wide range of subtle modulatory effects, it also makes sense to have many more, individually weaker synapses that can be combined in myriad ways.

One last point should be made regarding the driver/modulator distinction with regard to the LGN. If one had only anatomical information, such as numbers of synapses from subcortical sites, the number from the parabrachial region (~30%) is considerably greater than that from retina (5–10%). Such anatomical data in isolation would suggest the conclusion that the parabrachial input is the more important and thus represents the information being relayed, while the retinal input, being so small, performs some vague, lesser function that might not even merit inclusion in some schematic illustrations of geniculate circuitry. The key lesson here is that anatomical data, on their own, can be very misleading when trying to unravel functional circuits. With regard to information processing through thalamus, the most important issue is to identify what is being relayed, and to do so, it is potentially misleading to treat all pathways as equal: one must instead separately identify drivers from modulators.

Usually, one is likely to find examples of schematically illustrated circuits that are based on anatomy alone, as if all inputs were drivers in the sense the term has been used here. If the concept of drivers and modulators has validity beyond thalamus, many of these suggested circuits need to be reconsidered.

First- and Higher-Order Relays: The LGN and Pulvinar

There are two ways to think about the function of the thalamus. One is to consider the properties of thalamic circuitry as they affect relay functions; that is, how do the modulators affect relay cell responses to their driver input? The other is to consider what it is that the thalamic nucleus is actually relaying. Put another way, we can define the function of LGN or the ventral posterior nucleus as relaying retinal or medial lemniscal information, respectively. It is this latter aspect of thalamic functioning that chiefly concerns us in this section.

Layer 5 Corticothalamic Inputs as Drivers

Since, as defined above, the driver input is what is relayed by thalamus, we can define the function of thalamic relays by identifying the source and nature of their drivers. This may seem obvious and trivial for well-studied relays like the LGN, but there are many other less well understood relays with unknown functions because, until recently, their driver inputs remained undefined. Examples are the pulvinar and medial dorsal nucleus. We now know that a major source of driver input to thalamic relays like the pulvinar and medial dorsal nucleus is layer 5 of cortex. This is illustrated in Figure 6.

Figure 6(a) illustrates several key points of this arrangement. All thalamic relays receive a feedback corticothalamic projection from layer 6, and they also have inputs from the TRN; not shown for simplicity are inputs to relay cells from interneurons and the parabrachial region (see Figure 4). However, while some thalamic nuclei relay subcortical driver inputs to cortex (Figure 6(a)), others instead relay driver input that arises from cortical layer 5, and this appears to be feed-forward (Figure 6(b)). Because of this arrangement, we refer to the type of thalamic relay of Figure 6(a) as first order, because this is the first relay of a particular type of information to cortex, and that of Figure 6(b) as higher order, because this represents a relay of information already in cortex but from one cortical area to another. Examples of first-order relays are the LGN for vision, the ventral posterior nucleus for somesthesis, and the ventral (or lemniscal) portion of the medial geniculate

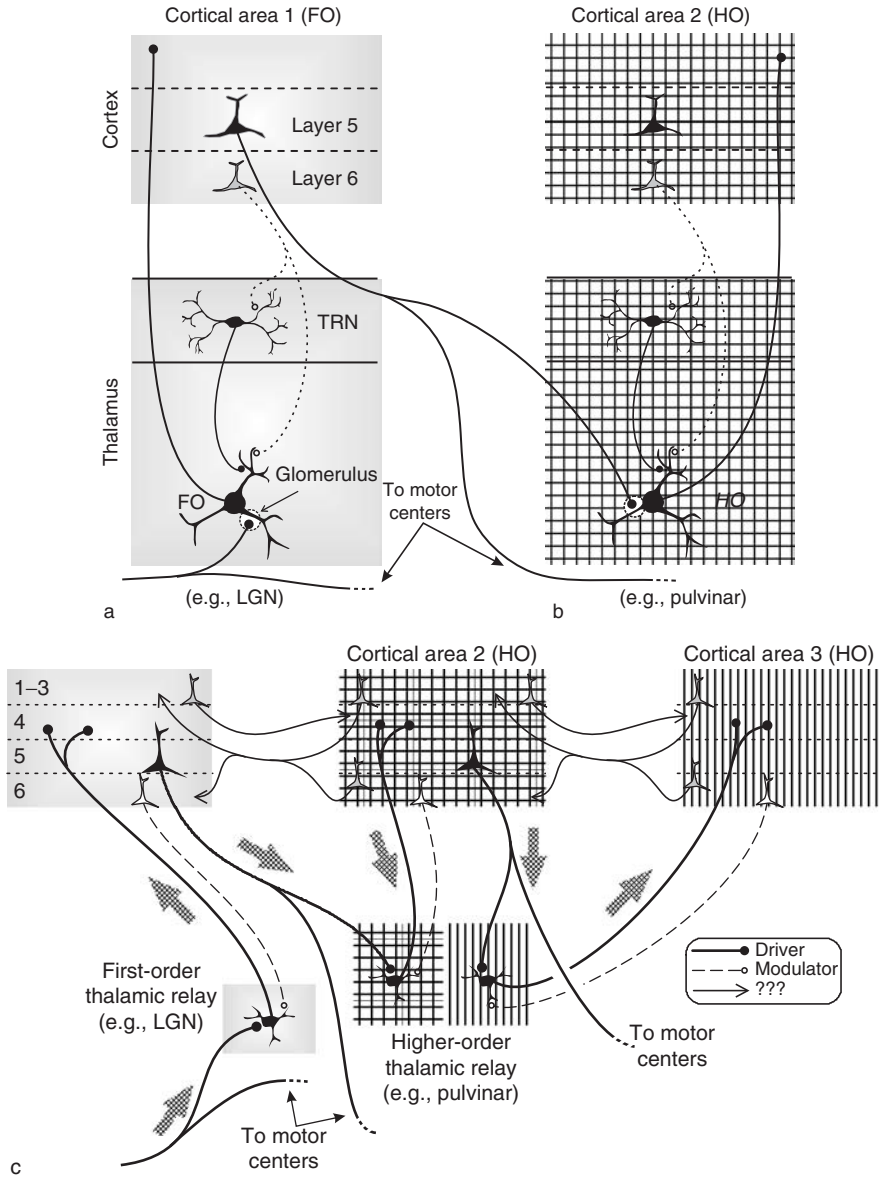

Figure 6 Schematic diagrams showing organizational features of first- and higher-order thalamic nuclei. (a, b) Distinction between first-order and higher-order thalamic nuclei. A first-order nucleus (a) represents the first relay of a particular type of subcortical information to a first-order or primary cortical area. A higher-order nucleus (b) relays information from layer 5 of one cortical area to another cortical area. This relay can be between first- and higher-order cortical areas as shown or between two higher-order cortical areas. The important difference between them is the driver input, which is subcortical (a) for a first-order thalamic nucleus and from layer 5 of cortex (b) for a higher-order one. Note that all thalamic nuclei receive an input from layer 6 of cortex, which is mostly feedback, but higher-order nuclei in addition receive a layer 5 input from cortex, which is feed-forward. (c) Role of higher-order thalamic nuclei in corticocortical communication. This involves a projection from layer 5 of cortex to a higher-order thalamic relay to another cortical area. As indicated, the role of the direct corticocortical projections, driver or modulator, is unclear. Note that in all drawings the driver inputs, both subcortical and from layer 5, are typically from branching axons, the significance of which is elaborated in the text. FO, first order; HO, higher order; LGN, lateral geniculate nucleus; TRN, thalamic reticular nucleus. Reprinted from Sherman SM (2005) Thalamic relays and cortical functioning. *Progress in Brain Research* 149: 107–126.

nucleus for sounds; examples of higher-order relays for these same sensory modalities, respectively, are most of the pulvinar, most of the posterior medial nucleus, and most of the medial and dorsal (or nonlemniscal) portion of the medial geniculate nucleus. We say 'most' here, because as defined by cytoarchitectonic boundaries, a minor part of these higher-order nuclei may relay subcortical driver inputs. Thus, there are small retinorecipient and tectorecipient zones of the pulvinar, a spinothalamic zone of the posterior medial nucleus, and an inferior collicular input to the nonlemniscal part of the medial geniculate nucleus that may be first-order relays. This has to be clarified, and it will not be considered

further here, but it does point up a potential shortcoming of classically defined, cytoarchitectonic boundaries for functional thalamic relays. The key to this division of thalamic relays into first and higher order is the observation that the layer 5 inputs to relay cells have the same properties as do the subcortical drivers (e.g., retinal input to the LGN and medial lemniscal input to the ventral posterior nucleus). Support for this can be found in Table 1: note from the first two columns of Table 1 that the layer 5 input to higher-order relays matches retinogeniculate input on all criteria; in contrast, the second and third columns show that the corticothalamic inputs from layers 5 and 6 differ on all criteria.

Branching of Driver Afferents to Thalamus

Another feature shown in Figures 6(a) and 6(b) is that the driver inputs to both first- and higher-order relays are borne mostly or wholly via axons that branch, with one branch innervating thalamic relay cells and the other innervating extrathalamic targets in the brain stem and spinal cord. For instance, most or all retinogeniculate axons branch to innervate the pretectum and/or superior colliculus, and, similarly, most or all layer 5 corticothalamic axons branch to innervate targets in the pons, midbrain, medulla, and sometimes even spinal cord. Limited data are consistent with a similar arrangement for relays of somatosensory and auditory information. RW Guillery has described this feature of driver afferents and suggested what its functional significance might be; this is briefly outlined below. This branching of afferent drivers to thalamus appears to be a general feature of thalamic relays and is not limited to the LGN and pulvinar.

The main extrathalamic targets of these branches seem to what can generally be described as motor centers. Thus, retinogeniculate axons branch to innervate pretectal and superior collicular targets implicated in eye movements, pupillary control, etc. Similarly, the layer 5 corticothalamic axons branch to innervate targets in brain stem and sometimes spinal cord that also appear to have motor functions. One interpretation of this pattern is that the information actually relayed by thalamus relates to motor commands, starting with first-order relays as perhaps quite crude commands that are constantly upgraded with further cortical processing and effected via higher-order layer 5 outputs from cortex. This notion implies that even first-order sensory processing, for instance, via the LGN, involves processing of motor commands, a notion that stands conventional views of early visual processing on its head. That is, conventionally, primary visual cortex (V1) is generally viewed as a purely sensory structure, and this view seems at odds with the idea that V1 is processing motor information. Furthermore, as already noted, V1 (and, indeed all cortical areas so far studied) has a layer 5 projection that branches to innervate pulvinar and extrathalamic motor targets, so that even the corticofugal outputs of V1 have a motor tag according to this perspective. Indeed, the whole notion that V1 or any other visual, auditory, or somatosensory area is purely 'sensory' is challenged by the observation that all of these areas have a motor output.

Role of Higher-Order Thalamic Relays in Corticocortical Processing

Figure 6(c) illustrates the suggested role played by higher-order thalamic relays. Information first reaches cortex via first-order relays, such as the LGN. Further processing of information as it is passed on to higher extrastriate visual cortical areas is relayed through higher-order thalamic nuclei. This can involve a number of hierarchical levels of both cortical and thalamic processing. The obvious question raised here is: if the cortico-thalamo-cortical pathways involving higher-order thalamic nuclei represent a significant information route, what of the direct corticocortical projections? The answer, simply, is not yet available, but to help to clarify the issue here, it is helpful to consider the three obvious hypotheses that follow.

Hypothesis 1: All information is relayed via higher-order thalamic nuclei One possibility is that all of the direct corticocortical pathways are modulators, in which case all information between cortical areas is relayed via the thalamus. One conclusion that could be drawn here is that all new information that reaches cortex, whether originating from a subcortical source such as the retina or from another cortical area, benefits from a thalamic relay. That is, for the same reason that retinal information is relayed by the LGN and does not project to visual cortex directly, information from one cortical area to another is relayed through thalamus. Benefits could include gating properties of the thalamus, the burst/tonic transition, etc.

A key challenge to this concept is whether or not higher-order relays such as the pulvinar have enough neurons to relay all of the requisite information needed for cortical processing. The pulvinar is indeed the largest thalamic nucleus and dwarfs the LGN. However, pulvinar neurons are orders of magnitude fewer than those in any cortical area, which poses a severe bottleneck on information transfer. Nonetheless, a very small percentage of visual cortical neurons are represented by the layer 5 efferents that could provide the afferent link in the cortico-thalamo-cortical pathway, and these numbers do not seem to pose a limitation on the role of the pulvinar as a central relay structure between cortical areas. We can thus ask

whether the limited number of layer 5 efferents to pulvinar is sufficient to project all of the information processed by a cortical area, such as V1. Unfortunately, we as yet have no answer to this question, because we simply do not know the nature of neural coding of this information that is passed on, and our ignorance here is such that we cannot rule out the possibility that the small number of layer 5 efferent cells is sufficient for this task. It is also possible that the full extent of information processed in a cortical area requires an additional, corticocortical route, the case to which we now turn.

Hypothesis 2: All information is relayed via direct corticocortical pathways The idea that the cortico-thalamo-cortical pathways involving higher-order thalamic nuclei serve as an information conduit implies that the relevant corticothalamic and thalamocortical projections are drivers. There is some evidence that the layer 5 inputs to higher-order thalamic relay cells act as drivers, but there is no direct evidence that the companion thalamocortical projections are drivers. Evidence from thalamocortical projections from first-order relays indicate that they behave essentially like drivers, so it seems plausible that this would extend to thalamocortical projections more generally. However, one idea for pulvinar projections to cortex is that these serve to regulate attentional responses, which, in turn implies that these projections could act as modulators. We need empirical evidence on the nature of thalamocortical projections from higher-order thalamic nuclei. If such evidence indicates that these higher-order thalamocortical inputs are drivers, then at least much information passed between cortical areas involves higher-order thalamic relays in a cortico-thalamo-cortical route. If, instead, the evidence is that these inputs are modulators, then we would conclude that virtually all corticocortical processing is based on direct corticocortical projections.

Hypothesis 3: Information is relayed in parallel by both pathways Perhaps the most likely notion is that the higher-order cortico-thalamo-cortical and direct corticocortical pathways represent two parallel, largely independent, and complementary routes of information flow. One example of this would be that the very large corticocortical projection handles all of the details of information that must be analyzed about the environment, and the cortico-thalamo-cortical projections inform the target cortical area about motor commands initiated by the source area. This is a more limited form of information but is essential, because higher-order cortical areas must maintain a real-time appreciation of how motor commands affect sensory processing. For example, higher visual cortical areas need to be able to factor in eye movements that cause the visual world to move on the retina and not view these as movement in the environment. This example of the limited sort of information carried by the cortico-thalamo-cortical pathways is consistent with the motor branches of the layer 5 axons described above.

Two further points need to be emphasized here. First, even if both pathways are involved in information flow, there is an important distinction to be made. Whatever information is carried by direct corticocortical connections, this remains in cortex and is thus different in kind from information carried by the layer 5 outputs to higher-order thalamic nuclei, because, as just noted, these layer 5 axons branch to carry the same information to various extrathalamic, subcortical targets. Second, this assumes the possibility that the driver/modulator distinction makes sense for intracortical pathways, but even if some corticocortical projections carry information, the massive potential problem remains to determine which pathways are modulators and which are drivers. Identifying the subset of drivers among these, and the subset may well be small, along with a full appreciation of the cortico-thalamo-cortical pathways, would allow the creation of a more complete and accurate hierarchical scheme for cortical processing.

Overview Figure 7 summarizes some of the key differences between the conventional view of corticocortical processing and the alternative view offered here. In the conventional view, information is relayed by thalamus to sensory cortex and passes within cortex to sensorimotor and then motor cortex before an output is generated to motor centers (Figure 7(a)). One problem with this is that it provides no specific role for most of thalamus, which we have defined as higher order, although there are suggestions that some of these thalamic nuclei could play a modulatory role in cortex related to attention. The alternative view (Figure 7(b)) differs from the beginning, since initial information to be relayed via a first-order thalamic nucleus reflects information sent to motor structures, perhaps reflecting an early motor command. From primary cortex, information can be relayed to other cortical areas via higher-order thalamic nuclei, and this can continue through several hierarchical stages. Also, each of these further thalamo-cortico-thalamic projections involves a layer 5 corticothalamic axon that branches to innervate extrathalamic motor structures as well as thalamus. The role of direct corticocortical pathways remains unclear in this view, but it is plausible that there are two information routes operating independently and in parallel, one the direct corticocortical route and the other, the cortico-thalamo-cortical route.

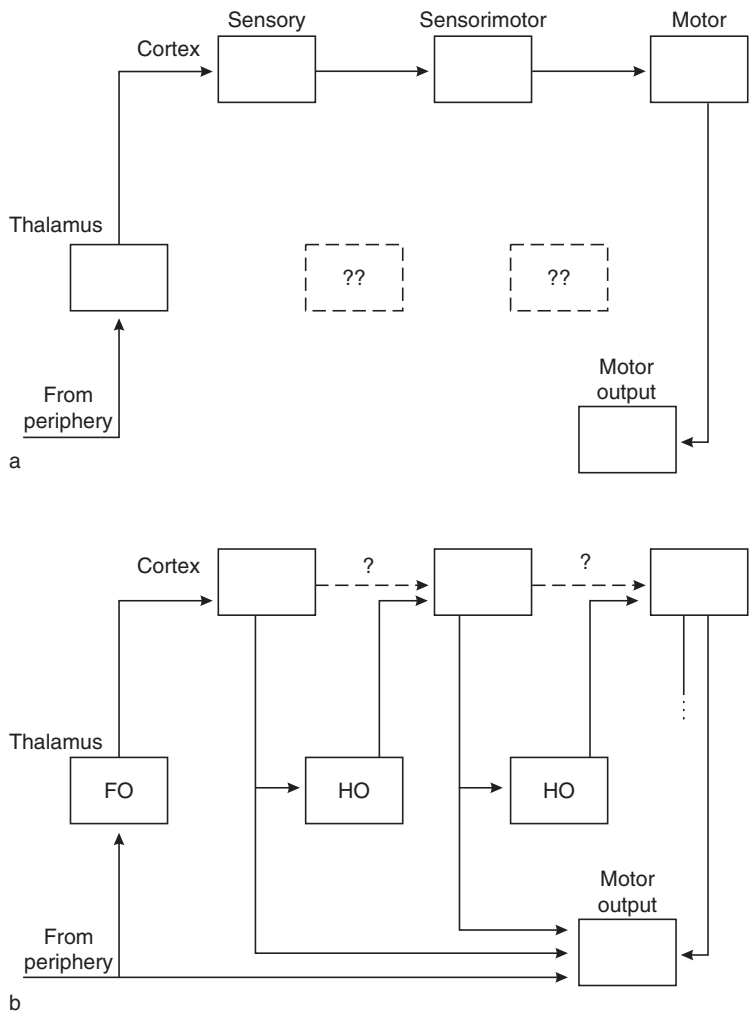

Figure 7 Comparison of conventional view (a) with the alternative view proposed here (b). The role of the direct corticocortical connections in (b) (dashed lines) is questioned (see text for details). FO, first order; HO, higher order. Reproduced from Sherman SM (2005) Thalamic relays and cortical functioning. *Progress in Brain Research* 149: 107–126.

It seems most likely that higher-order thalamic nuclei play an important and hitherto unappreciated role in corticocortical communication. Thus, thalamus is not there just to get information to cortex in the first place but rather continues to play a role in further cortical processing of that information. What is less clear is the role of direct corticocortical pathways and their relationship to the cortico-thalamo-cortical pathways. If the driver/modulator distinction has relevance for these pathways, and that is a major proviso, then it will be essential to identify which of these pathways, if any, are drivers. Only then can we have a clearer understanding of processing among the related cortical areas.

Conclusions

We have come a long way from the bad old days when the thalamus was seen as a dull, machinelike relay, acting up in an interesting fashion only during epilepsy or slow-wave sleep. To a large extent, this misconception grew out of the success of the receptive field approach to the study of sensory systems, particularly vision. Studies of the retina showed that receptive fields become increasingly complicated as one ascends synaptic hierarchies, leading ultimately to the classic center/surround receptive field of ganglion cells projecting to the LGN. This process continues across synaptic hierarchies in cortex, providing receptive fields there with exquisite sensitivity to orientation, direction of movement, spatial frequency, stereoscopic depth, etc. We believe that this receptive field elaboration in retina and cortex is ultimately used to encode all features of the sensory environment. The one synapse in the visual system across which no significant receptive field elaboration occurs is the retinogeniculate synapse, since the same basic center/surround organization is seen in retinal afferents and their target geniculate

relay cells. Because of the prevalent receptive field chauvinism dictating how one determined functional organization, this led to the notion that the LGN specifically and thalamus more generally represented an uninteresting, simple relay.

We can now turn that view on its head. Indeed, while the rest of the visual and other sensory systems are engaged in receptive field elaboration, the thalamus has a completely unique role to play in information processing. Recent appreciation of the complex cell and circuit properties of thalamus make it clear that it is anything but simple and uninteresting in its functioning. These properties serve to regulate the flow of information to cortex through mechanisms such as gain control of the retinogeniculate synapse (or equivalent for other nuclei), the burst/tonic transition, etc., and here we are probably just seeing the tip of the iceberg, given the preponderance of modulatory pathways and voltage-gated conductances in addition to the T channels than can be manipulated. There is also the point that thalamus is not there just to get information to cortex but continues to play a significant role in corticocortical communication.

The challenge for students of the visual system is at least twofold. One is to gain a better appreciation of how and under what conditions information is affected before being relayed by the LGN or pulvinar to cortex. The other is to address questions about the pulvinar. One of the great problems here is that we have no complete map of pulvinar that includes a full demarcation of what regions of pulvinar innervate which regions of cortex and which are innervated by each cortical area, with a separate mapping of layer 5 and layer 6 inputs. Given these variables and the presence of more than 30 visual cortical areas with which the pulvinar is involved in rhesus monkeys and humans, it may be that more than a hundred separate pulvinar regions remain to be discovered. This is a daunting task and should be seen as one of the major challenges for future studies of the visual system.

See also: Thalamus and Oculomotor Control; Thalamus: Evolution in Vertebrates; Visual Cortex in Humans.

Further Reading

Agmon A and Connors BW (1991) Thalamocortical responses of mouse somatosensory (barrel) cortex *in vitro*. *Neuroscience* 41: 365–379.

Alitto HJ, Weyand TG, and Usrey WM (2005) Distinct properties of stimulus-evoked bursts in the lateral geniculate nucleus. *The Journal of Neuroscience* 25: 514–523.

Anderson CH, Van Essen DC, and Olshausen BA (2005) Directed visual attention and the dynamic control of information flow. In: Itti L, Rees G, and Tsotso J (eds.) *Neurobiology of Attention*, pp. 11–17. San Diego: Elsevier.

Bezdudnaya T, Cano M, Bereshpolova Y, Stoelzel CR, Alonso JM, and Swadlow HA (2006) Thalamic burst mode and inattention in the awake LGNd. *Neuron* 49: 421–432.

Casagrande VA and Xu X (2004) Parallel visual pathways: A comparative perspective. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 494–506. Cambridge, MA: MIT Press.

Chance FS, Abbott LF, and Reyes A (2002) Gain modulation from background synaptic input. *Neuron* 35: 773–782.

Conn PJ and Pin JP (1997) Pharmacology and functions of metabotropic glutamate receptors. *Annual Review of Pharmacology and Toxicology* 37: 205–237.

Hendry SHC and Reid RC (2000) The koniocellular pathway in primate vision. *Annual Review of Neuroscience* 23: 127–153.

Hamos JE, VanHorn SC, and Sherman SM (1986) Synaptic circuitry of an individual retinogeniculate axon from a retinal Y-cell. *Society for Neuroscience* 12: 1037.

Jahnsen H and Llinás R (1984) Electrophysiological properties of guinea-pig thalamic neurones: An *in vitro* study. *Journal of Physiology (London)* 349: 205–226.

Lesica NA and Stanley GB (2004) Encoding of natural scene movies by tonic and burst spikes in the lateral geniculate nucleus. *The Journal of Neuroscience* 24: 10731–10740.

Nicoll RA, Malenka RC, and Kauer JA (1990) Functional comparison of neurotransmitter receptor subtypes in mammalian central nervous system. *Physiological Reviews* 70: 513–565.

Olshausen BA, Anderson CH, and Van Essen DC (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience* 13: 4700–4719.

Reichova I and Sherman SM (2004) Somatosensory corticothalamic projections: Distinguishing drivers from modulators. *The Journal of Neurophysiology* 92: 2185–2197.

Rodieck RW (1979) Visual pathways. *Annual Reviews in Neuroscience* 2: 193–225.

Sherman SM (1985) Functional organization of the W-, X-, and Y-cell pathways in the cat: A review and hypothesis. In: Sprague JM and Epstein AN (eds.) *Progress in Psychobiology and Physiological Psychology*, vol. 11, pp. 233–314. Orlando: Academic Press.

Sherman SM (2001) Tonic and burst firing: Dual modes of thalamocortical relay. *Trends in Neurosciences* 24: 122–126.

Sherman SM and Guillery RW (1998) On the actions that one nerve cell can have on another: Distinguishing 'drivers' from 'modulators'. *Proceedings of the National Academy of Sciences of the United States of America* 95: 7121–7126.

Sherman SM and Guillery RW (2006) *Exploring the Thalamus and its Role in Cortical Function*, 2nd edn. Cambridge, MA: MIT Press.

Stanford LR, Friedlander MJ, and Sherman SM (1983) Morphological and physiological properties of geniculate W-cells of the cat: A comparison with X- and Y-cells. *The Journal of Neurophysiology* 50: 582–608.

Swadlow HA and Gusev AG (2001) The impact of 'bursting' thalamic impulses at a neocortical synapse. *Nature Neuroscience* 4: 402–408.

VanEssen DC (2005) Corticocortical and thalamocortical information flow in the primate visual system. *Progress in Brain Research* 149: 173–185.

VanHooser SD, Heimel JAF, and Nelson SB (2003) Receptive field properties and laminar organization of lateral geniculate nucleus in the gray squirrel (*Sciurus carolinensis*). *The Journal of Neurophysiology* 90: 3398–3418.

Thalamus and Oculomotor Control

J Schlag, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Is There an Oculomotor Thalamus . . . and If So, What Defines It?

The thalamus is an ensemble of nuclei located in the diencephalon. The output neurons of these nuclei project to the telencephalon (i.e., cerebral cortex and basal ganglia in mammals). Descending projections practically do not exist (or have not been found) except from the reticular nucleus, a thin cellular shell which envelops practically the whole thalamus. Some thalamic nuclei, such as the lateral and medial geniculate nuclei, are easy to characterize in terms of a single function (respectively, visual and auditory). Their anatomy is straightforward: each geniculate nucleus receives an input of its proper sensory modality and projects to a limited, specific cortical area called a primary sensory area. Geniculate nuclei are called relay nuclei, a term which is almost derogatory because it seems to imply that thalamic synapses have no processing function; in return, geniculate nuclei receive a massive feedback from their own cortical areas. Thus are formed separate thalamocortical loops operating as specialized functional units. They are the earliest stages in seeing and hearing in the examples presented herein. Other thalamic nuclei are parts of other similar units, but not all can be as simply characterized.

What could be an oculomotor thalamus? An oculomotor apparatus would rely on a variety of signals of different sensory modalities, such as visual, auditory, and vestibular, and on a variety of internal signals, such as those that guide a search of the environment on the basis of past experience. It would induce appropriate ocular movements, such as saccades and pursuit; and it would also cooperate with other nonocular movements, such as movements of the head (to extend the span of visual exploration) and movements of the hand (to point, reach, and grasp). That is probably how an oculomotor thalamus would look like if we were searching for it.

The region of the thalamus that best fits this description is centered on the internal medullary lamina (IML) in mammals. The IML is a sheet of fibers that encapsulates the mediodorsal (MD) nucleus. Sagittally on both sides, the IML runs between the MD nucleus medially and the bulk of the other thalamic nuclei laterally, among which the most important, in our point of view, are, from back to front, the pulvinar (Pul), lateral posterior (LP), ventral lateral (VL), and ventral anterior (VA) nuclei (see Figures 1 and 2 for orientation). Caudally, the IML splits to form an envelope (capsule) around the centre médian (CM), a motor nucleus that is highly developed in primates. Within the IML are clusters of neuronal cell bodies, called intralaminar nuclei. The largest intralaminar nuclei are the paracentral (Pcn) and the central lateral (Cl) with its dorsal extension, the central superior lateral (Csl), all of which certainly are parts of the oculomotor thalamus.

The IML seems to be a good landmark to indicate the location of neurons with oculomotor properties. Indeed, based on microelectrode recordings, such neurons are situated within the lamina and, mostly, on both sides of it, in the so-called paralaminar (paralamellar) region and the multiformis sector of the MD, medially, and the paralaminar region of VL and VA, laterally. Here attention should be called to an ambiguity persisting in the literature about the labels used to designate these thalamic sectors: namely, the cells that some authors regard as intralaminar neurons are regarded by other authors as paralaminar MD. The reason for this ambiguity is that the borders of the lamina are not clear, certainly being less obvious than are the boundaries traced in atlases and used by neurophysiologists to locate their microelectrode tracks (this problem has been discussed in detail by EG Jones). This whole territory, intralaminar and paralaminar, is sometimes referred to as the central thalamus, and this terminology is adopted here.

Comparing the spatial relations of the thalamic nuclei considered here with the spatial relations of cortical frontal areas reveals a systematic organization that is well demonstrated by simultaneously injecting tracers of different colors in selected areas of the frontal lobe. After retrograde transport, the cell bodies of thalamic neurons which project to each of the injected cortical areas are discriminately marked. These neuronal populations are independent but adjacent. Figures 1 and 2 show frontal sections of the monkey thalamus and schematics of the frontal lobe. In Figure 1, the injected areas are the supplementary eye field (SEF; black), the motor cortex (area 4; red), and the prefrontal cortex (blue). In Figure 2, the injected cortical areas are the saccadic sector of frontal eye field (FEFsac; red), the SEF (black), and the dorsal premotor cortex (PMd; blue).

With respect to the lamina, the more medial thalamic populations (i.e., lateral part of MD) are reciprocally interconnected with the prefrontal cortex

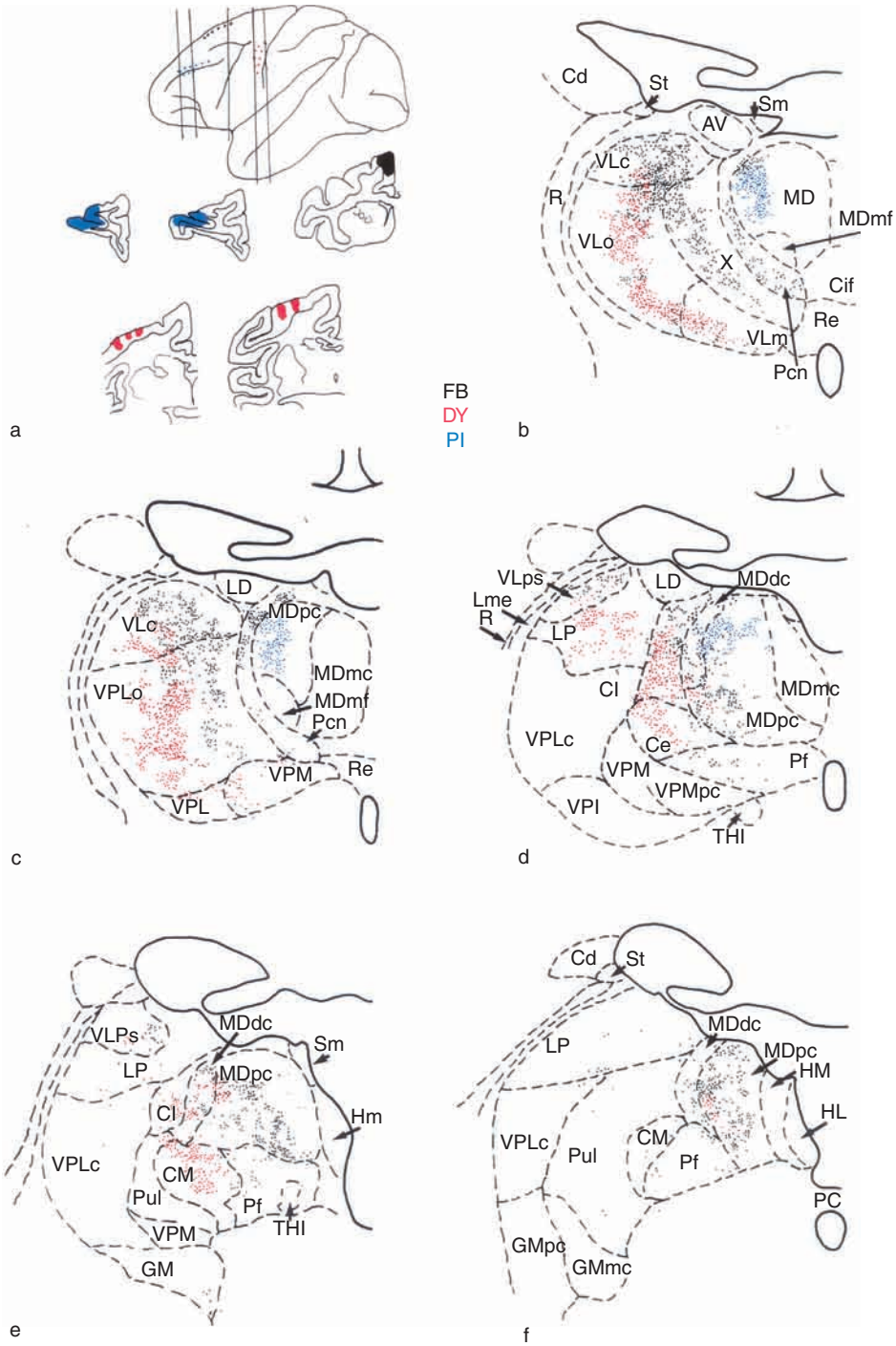

Figure 1 (a) Retrogradely labeled cell bodies of thalamic neurons projecting to the three frontal lobe areas (left lateral view of the brain and cross-sections): prefrontal cortex along sulcus principalis in blue, rostral sector of supplementary motor cortex corresponding to supplementary eye field in black, and motor area in red. (b–f) Coronal sections through the thalamus, from most rostral (b) to most caudal (f). Note, particularly in panels (b–d), the distribution of marked thalamic neurons: those projecting to prefrontal cortex (in blue) are most medial and those projecting to motor cortex (in red) are most lateral. Nuclei mentioned in the text (see **Figure 2** for more abbreviations): Cl, centralis lateralis; CM, centre médian; LP, lateralis posterior; MD, medialis lateralis; Pcn, paracentralis; Pul, pulvinar; VL, ventralis lateralis; X, area X of VL. In this list, the intralaminar nuclei (inside the internal medullary lamina) are Cl, centralis superior lateralis, and Pcn. Adapted from Goldman-Rakic P and Porrino LJ (1985) The primate mediodorsal (MD) nucleus and its projection to the frontal lobe. *Journal of Comparative Neurology* 242: 535–560, with permission.

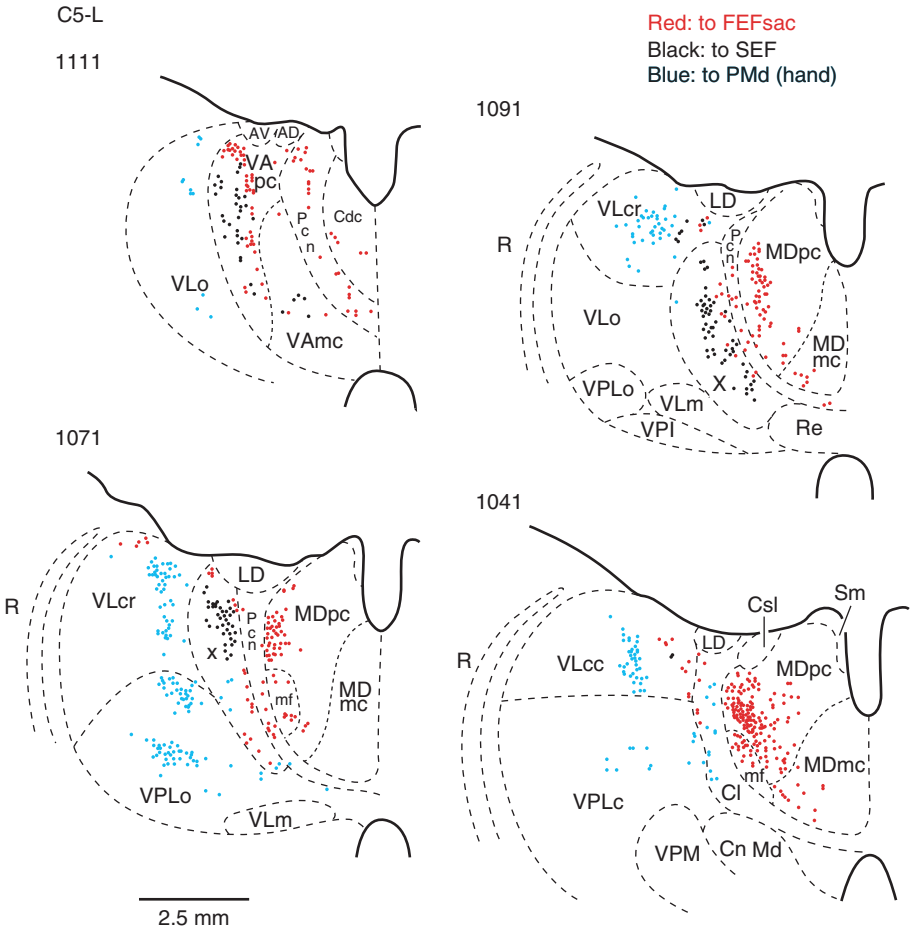

Figure 2 Retrogradely labeled cell bodies of thalamic neurons projecting to the three frontal lobe areas (two of them are different from those illustrated in Figure 1): saccadic portion of frontal eye field in front of arcuate sulcus (FEFsac, in red), supplementary eye field (SEF, in black), and hand sector of premotor cortex (PMd, in blue). The most rostral thalamic cross-section is 1111; the most caudal is 1041. Note the distribution of marked thalamic neurons: those projecting to the FEF are most medial; those projecting to the premotor cortex are most lateral. AD, anterior dorsalis; AV, anterior ventralis; Cdc, centralis densocellularis; Cl, central lateral nucleus; Cn Md, centrum medianum; Csl, centralis superior lateralis; LD, lateralis dorsalis; MDmc, medialis dorsalis, pars magnocellularis; MDpc, medialis dorsalis, pars parvocellularis; mf, medialis dorsalis, pars multiformis; Pcn, paracentral nucleus; R, reticular nucleus; Re, nucleus of reuniens; Sm, stria medullaris thalami; VAmc, ventralis anterior, pars magnocellularis; VApc, ventralis anterior, pars parvocellularis; VLcc, caudal portion of ventralis lateralis, pars caudalis; VLcr, rostral portion of ventralis lateralis, pars caudalis; VLm, ventralis lateralis, pars medialis; VLo, ventralis lateralis, pars oralis; VPI, ventralis posterior inferior; VPLc, ventralis posterior lateralis, pars caudalis; VPLo, ventralis posterior lateralis, pars oralis; VPM, ventralis posterior medialis; X, area X in the ventral lateral complex. Adapted from Tian J-R and Lynch JC (1997) Subcortical input to the smooth and saccadic eye movement subregions of the frontal eye field in Cebus monkey. *Journal of Neuroscience* 199: 9233–9247. Copyright 1997 by the Society for Neuroscience.

(see Figure 1), a cortical region implicated in 'executive' functions as studied in cognitive tasks that often – but not always involve directing gaze in monkeys. In fact, the same types of neuronal behavior are found at both the thalamic and cortical stages of this system. On the other side, the more lateral thalamic populations (i.e., in VL and VA) are more clearly motor. This is the part of the thalamus that receives inputs from the cerebellum (via deep cerebellar nuclei) and from the basal ganglia, and is interconnected with the motor and premotor cortices, which are the caudal parts of the frontal lobe. Here in these interconnections, the accent is on 'motor' function, of which the oculomotor function is an integral part. From an oculomotor point of view, unit studies that have focused on this thalamic region have stressed, in particular, the role of neurons in timing eye movements (see later). Between these medial and lateral thalamic sectors is the thalamic region interconnected with the frontal eye field and supplementary eye field (see Figures 1 and 2). In this thalamic region many neurons fire or pause with saccades or with fixation. In summary, the schematic views presented here relate the rostral part of the frontal cortex with the more medial structures of the central thalamus and the caudal part of the frontal cortex with the more

lateral structures. In addition, areas of the parietal lobe dealing with visuospatial functions (including visual motion and visual pursuit) are interconnected with thalamic populations lateral to the IML caudally (i.e., Pul and LP).

It should also be mentioned that there are cells in other thalamic nuclei, such as the lateral geniculate nucleus and the pulvinar, that change their firing rate when eye movements occur. Some cells in the pregeniculate nucleus discharge with saccades. But these are not really parts of an oculomotor apparatus – rather, they modify visual processing when eye movements occur.

Activity Patterns of Cells in the Oculomotor Thalamus

Single-unit recording work has been done by different groups of investigators with somewhat different primary interests. For one group, the main concern was studying oculomotor mechanisms at the level of the forebrain, whereas for another group, the accent was on using eye movements as a tool to explore visuospatial or cognitive functions. In this last case, saccades proved to be a very convenient way to provide evidence of an individual's choice in a behavioral task. Microelectrode experiments were performed, first in cats, but mostly in monkeys. Practically all types of neuronal patterns of activity described in other oculomotor structures (FEF, SEF, and superior colliculus (SC)) were found in the oculomotor thalamus during the performance of a large variety of tasks. Figure 3 presents some of these patterns and some of the criteria by which they are recognized. We should distinguish the following criteria:

- *Visual responses*. Phasic or tonic (or both in succession). The cell in Figure 3(b) is phasic. With a fixed latency, it is active after a stimulus onset whether

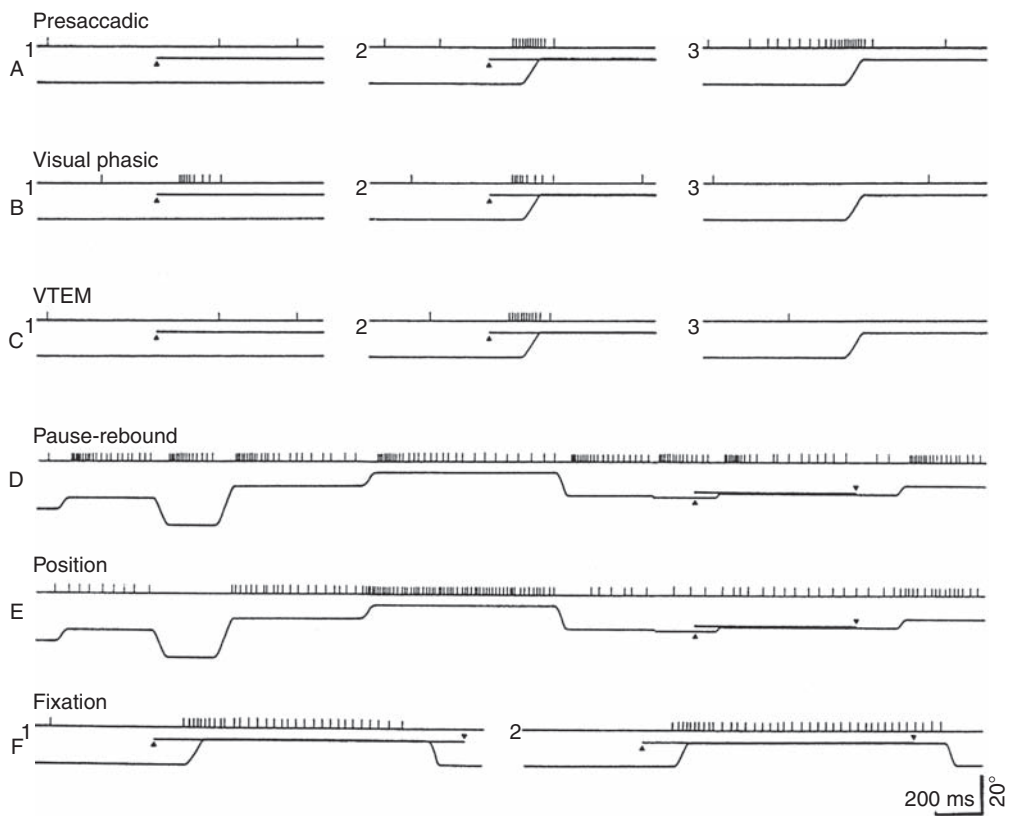

Figure 3 Some of the main patterns of firing observed in the monkey's central thalamus, with a few of the experimental conditions used for diagnostics. Schematic representations based on actual data. In each case, the upper trace is the single-unit record; the bottom traces are the horizontal records of stimulus and eye positions (vertical positions not shown for simplicity). Triangles indicate stimulus onsets and offsets (VTEM, visually triggered eye movement). Three units (a–c) produce similar bursts but are distinguished from each other by the conditions (1–3) in which such bursts occur: stimulus presentation with no targeting saccade in condition 1, stimulus presentation and targeting saccade in 2, and spontaneous saccade of exploration in 3. Another useful discriminating paradigm (not shown) to distinguish visual from saccadic activity is the memory-saccade task. Other information is discussed in the text. Adapted from Schlag J and Schlag-Rey M (1986) Role of central thalamus and supplementary eye field in voluntary control of gaze in space. *Bulletin of Tokyo Metropolitan Institute for Neurosciences* (1986 supplement): 17–31.

the eyes capture the stimulus or not, but it is not active with a saccade of the same amplitude and direction if there is no visual stimulus. Central thalamus neurons have large receptive fields (>10° in diameter), generally contralateral. The responses of the simplest of visual neurons do not depend on stimulus characteristics (such as color or shape), but, as shown, notably, by Stanford's group, more complex neurons may selectively respond to one of these characteristics if the monkey has learned that it is a cue requiring fixation of that stimulus to obtain a reward. Such cells are commonly encountered in frontal and parietal cortices. Some of these neurons continue firing even if the visual target is temporarily blanked.

- *Fixation neurons*. These are different from visual tonic neurons with a foveal receptive field because they still fire if the stimulus vanishes but the eyes continue to fixate. Figure 3(f) shows the activity of a fixation cell in the presence of the fixated stimulus. These neurons deserve to be listed apart from other visual neurons for two reasons. First, about half of these units are activated, with the other half being inhibited (silenced) during fixation. Second, the role of these units may not be visual but intentionally motor, blocking the gaze and maintaining it on a site of interest. Note that, if such a site of interest were a smoothly moving stimulus, some of these cells could fall in the category of pursuit neurons (see later).
- *Delay activity*. This is the tonic firing which, in many neurons, fills the period between an initial phasic visual response and the saccadic activity. It does not depend on the persistence of the visual stimulus. Delay activity is often considered as evidence that some spatial information (such as a goal location) has been entered and retained in memory. A recent comparison of delay activity between MD, SC, and FEF neurons indicates that, in MD at least, there is less sustained delay activity than in the SC and FEF.
- *Gap activity*. This appears during the brief gap (100–200 ms) that the investigator may insert between the extinction of the fixation point (which is the 'go' signal for a saccade) and the onset of the visual target. This activity is interpreted as a sign of disengagement of attention.
- *Saccadic activity*. This is usually directional (but not always). Directional activity is produced only when saccades terminate within a given area of the visual field (called 'movement field' by analogy to the sensory notion of receptive field). This is the case for the cell shown in Figure 3(a): it discharges prior to a saccade in the preferred direction whether or not there is a stimulus to be acquired. It is not activated (Figure 3(a,1)) by a stimulus if there is no saccade. Movement fields are usually contralateral. Some of these activities are presaccadic, starting as early as 300 ms before saccade onset. The discharge rate increases progressively, culminating in a burst, which often terminates after the saccade. Some neurons fire after saccade onset. Presaccadic neurons appear to be more numerous medially to the lamina while mid- or postsaccadic neurons are found mainly on the lateral side. Some neurons, such as the one illustrated in Figure 3(c) (to be contrasted with presaccadic and visual phasic) are activated only by visually guided (not spontaneous) saccades – that is, visually triggered eye movement (VTEM). Many neurons are visuomotor, giving first a visual response (as described previously) to a visual target and then being reactivated at the time of the saccade. Usually for these visuomotor neurons, there is a good correspondence between the location and size of their receptive field and movement field.
- *Pause neurons*. These stop firing during saccades, usually irrespective of their direction (see Figure 3(d)). Thus, they resemble the omnipause neurons of the pontine raphé, but they display a wider variability in the thalamus: some are better synchronized with saccade onset, others with saccade offset, and some show a strong rebound of activity after the pause, others do not. What seems striking in these neurons is the precise timing of their activity shifts. Pause cells have a particularity that may be important: they exist in the brain stem and thalamus, but apparently it seems that nobody has noticed their presence at the cortical level.
- *Eye position activity*. This is related to eye orientation in orbit along a given axis. It differs from visual tonic, fixation, and pursuit activity, as it does not require that a stimulus be or have been present (Figure 3(e)). In total darkness, the eyes wander and sometimes rest somewhere. To differentiate this from a real fixation, Fischer has appropriately called it 'eye parking.' Eye position neurons are best demonstrated in total darkness. In the audiomonitor, they sound like the output of a voltage-to-frequency converter. Usually, their frequency of firing is linearly related to ocular eccentricity in a given direction (most often contralateral) across the whole field or starting around the middle. Like presaccadic burst neurons, about half of the thalamic eye position neurons start changing their firing rate before saccades, which suggests that their activity is related to the production of movement rather than to its effect. There are preliminary indications that some thalamic eye position cells are also affected by head position.

- *Pursuit activity.* This occurs before and during initiation of smooth pursuit. The firing rate is proportional to the speed of target motion in a preferred direction. Such neurons continue firing even though the tracking stimulus is briefly extinguished, indicating that the signal they carry is motor or movement related, but not visual.

Finally, there are classes of activity that have been observed but not yet systematically studied, such as continuous activation (or silence) that starts when a monkey explores a display containing one target and several distractors, and lasts (several seconds) until the target is finally discovered.

Where Is the Central Thalamus in the Oculomotor Circuitry?

It is difficult to identify the origin of all of the aforementioned signals, as the central thalamus simultaneously is the obligatory gate for all incoming information to reach the cerebral cortex and is a recipient of feedback from this same cortex. The best known sources of inputs to this part of the thalamus, relevant to oculomotricity, are (1) SC, (2) substantia nigra, which receives oculomotor signals from the caudate nucleus and projects also to the superior colliculus, and (3) dentate nucleus of the cerebellum.

By the technique of retrograde transneuronal transport of herpes simplex virus type 1 (HSV-1) in cebus monkeys, Lynch and collaborators have clearly demonstrated the relations of these three input structures to the FEF via the thalamus. The virus injected into the FEF invades not only the thalamic cells projecting directly to the FEF but also, transsynaptically, the neurons projecting to these thalamic cells. The neurons are then immunohistochemically marked and the following organization emerges: the FEF is ultimately reached by neurons from the stratum opticum and stratum griseum intermedium of the SC via the paralaminar part of MD, by neurons from the pars reticularis of the substantia nigra via the magnocellular portion of VA, and by neurons from the posterior dentate nucleus via the area X of VL. Of these three pathways, the one originating in the superior colliculus has been further studied in detail by Sommer and Wurtz using physiological stimulation and recording in behaving monkeys. The aim was to identify the particular SC, MD, and FEF cells that happen to form a three-neuron pathway in order to compare their responsiveness during a task of visual targeting. To meet the criterion of connectivity, three conditions must be simultaneously satisfied: SC neurons must be antidromically activated from MD, MD neurons must be orthodromically activated from SC and antidromically activated from FEF, and FEF neurons must be orthodromically activated from SC. The main observations derived from such an experiment are as follows: the SC signals sent through this thalamocortical pathway are a fair sample of those generally found at the site of origin in the intermediate layers of the SC; the topography of visual and saccadic signals is conserved in terms of the location and size of receptive and movement fields. Visual signals from SC arrive in FEF too late to account entirely for the visual responses in FEF; the latter have a shorter latency and, thus, must be triggered from another source (maybe other visual areas). In contrast, presaccadic activity (as defined earlier) is transmitted fast (within around 2 ms) and is practically unchanged from what it is in SC, but the cells on which this activity impinges in FEF are not the movement-only neurons. Therefore, it does not seem that the immediate source of FEF presaccadic activity is the SC input.

What, then, is the significance of this presaccadic activity in an ascending pathway, a pathway which is, in fact, in the reverse direction from that expected of a motor signal? It can be a corollary discharge, a signal of the kind used by the brain, for instance, to take into account a shift in orientation of its receptor organs (the eyes in this case) when trying to locate visual features in the environment. The typical test in this situation is the double-step task. It consists in presenting two brief targets that the individuals will have to acquire in succession. As the targets have been extinguished before the first saccade, the individual has to rely on its memory of their positions. This is not a problem for reaching the first target but, for reaching the next one, the individual will have to calculate the trajectory of the second saccade taking into account the detour to reach the first target. Information on that detour is what a hypothetical corollary discharge is assumed to convey. If one of the roles of the SC–MD–FEF pathway is to provide this corollary discharge, blocking this pathway should impair the performance of the double-step task. More precisely, it should impair the second saccade in that task if the first saccade is contraversive, because what would be missing after a lesion is the corollary discharge of a vector if it is contraversive. Yet the first saccade would not be affected. That is what an injection of muscimol (a γ-aminobutyric acid agonist and inhibitor of transmission) has shown.

It would be quite interesting to study the possible effect of the same muscimol injection on the visual properties of neurons in the lateral intraparietal cortex (LIP). The receptive field responses of these neurons are modulated by eye position (gain fields). As the eye position signal very likely is provided by the brain stem

via the central thalamus, one could expect that LIP neurons will lose their eye position dependency.

Let us continue our listing of other important structures, in addition to (1) SC, (2) substantia nigra, and (3) dentate nucleus of the cerebellum, relevant to oculomotricity, which project to the central thalamus: (4) nucleus prepositus hypoglossi (5) vestibular nuclei, and (6) brain stem reticular formation. The nucleus prepositus hypoglossi, situated in the medulla, caudally to the abducens nucleus, contains eye position neurons which could be at the origin of the eye position signal recorded in the central thalamus. Vestibular nuclei projections seem to be divided into two streams: one to the central thalamus, the other to the somatosensory part of the thalamus (ventrolateral posterior nucleus), which is understandable, since information on balance concerns quite a number of other things than moving the eyes. Brain stem reticular formation is the most extensive and polyvalent part of the brain stem. Included are regions of the reticular formation clearly involved in saccade generation and in ocular pursuit. However, at the present time, it is very difficult to determine the contribution of particular reticular neurons to the massive ascending flow toward the central thalamus.

Finally, it is appropriate to mention again cortical feedback as an important input. It comes from the cortical areas with which the neural populations forming the central thalamus are related. Whereas in cases involving the lateral geniculate nucleus, the input from this cortical feedback is larger than the peripheral (visual) input; whether this is also true for the central thalamus is not known. A remarkable feature, however, has been signaled by Leichnetz: some descending projections from the FEF travel through the thalamus, via the IML, on their way down to the SC and midbrain tegmentum. When the existence of this transthalamic pathway was first reported, it was taken as an exception to the classic rule that all descending cortical projections to the brain stem and spinal cord pass through the internal capsule (i.e., around the thalamus, not through it).

Descending projections of the saccadic region of FEF terminate in the IML somewhat rostrally to – but with considerable overlap with – the projections from the pursuit region of the FEF. It is not yet clear whether the input to IML neurons is from collaterals of axons continuing toward the brain stem, or from a separate set of axons exclusively aimed at the thalamus. It will be interesting to clarify this issue because, if the input were from collaterals, one could assume that the thalamus receives from the cortex the same information that the cortex sends to the brain stem (perhaps including copies of motor commands).

Unfortunately, we still know very little about interneuronal processing within the thalamus, even in the case of the relay nuclei. This, of course, is the major obstacle to formulate hypotheses regarding thalamic function.

What Is Learned by Turning the Central Thalamus On (Stimulation) or Off (Lesion)?

Electrical stimulation within the central thalamus triggers saccades which are usually contraversive. When the head is immobilized, depending on the exact site of stimulation, evoked saccades are of either the fixed-vector type (constant amplitude and direction wherever the saccade trajectory starts in space) or goal-directed type (converging to a given orbital position). In cats, when the head is freed, head rotation also can be produced by the same stimulation. This seems to be the case for stimulation of most, if not all, oculomotor structures in the brain. The fact that saccades can be triggered from the thalamus is not surprising but it is not very informative, given that a bunch of '*fibres de passage*' (the IML) cross the central thalamus. Thus, theoretically, evoking an eye movement could be due to stimulating the thalamocortical afferents to a cortical eye field, which then retransmit this signal down to the brain stem (the descending transthalamic pathway from FEF) and/or the SEF to the brain stem, the SC afferent input to the central thalamus, which would be antidromically invaded first, followed by orthodromic invasion of collaterals within the brain stem, or any combination of these.

In cats, stimulation in a region of the central thalamus preliminarily treated with ibotenic acid, a toxin known to kill cell bodies but to spare fibers as they pass the cells, still can elicit saccades. Thus, of the three routes that we are considering here, the first one (activating the thalamocortical pathway), although not excluded, does not seem to be the only solution. For the same reason as for stimulation experiments, the effects of thalamic lesions by coagulation or pathological processes need to be interpreted with caution because, whenever the internal medullary lamina is encroached upon, fibers which may have originated far away from the site of the lesion are destroyed . The first reports concerning unilateral thalamic lesions in cats, monkeys, and humans describe the effect as a contralateral, probably motor, neglect. This neglect is transitory and similar to the one observed with some frontal lobe lesions.

Recently, two studies were made with chemical, reversible, lesions of the central thalamus in monkeys. The results are interesting particularly because they show a differentiation depending on the exact site of

the blockage, medial versus lateral to the IML. In the first study (Sommer and Wurtz, reported earlier), muscimol injected in MD affects the performance of the double-step task. As mentioned, this effect is interpreted as due to the suppression or weakening of a corollary discharge. In the other study (by Tanaka), muscimol was injected on the other side of the IML – that is, in the paralaminar VL, the region where dentate (i.e., cerebellar) projections impinge. Here, the deficit is in the timing of contraversive saccades, as revealed when the monkey's task requires that such saccades be self-triggered. Indeed, at the time of the first unit recordings in the central thalamus, it was already noted that one of the striking characteristics of these neurons (particularly the pause cells) is their capacity to mark the timing of all saccadic events as if this were some sort of controller clocking these events for the benefit of cortical areas.

Final Words

We began by asking the question: is there an oculomotor thalamus? The answer is, in a sense, yes. We saw that the central thalamus is certainly not homogeneous. On the other hand, its different parts are not dispersed all over the thalamus. They are gathered together very likely because they collect and make use of the same information. One may view the central thalamus as the director of the saccadic chorus. It would not produce any 'sound' itself, but it is in an ideal position to coordinate all of the voices of the centers variously suspected to initiate eye movements: SC, FEF, SEF, and basal ganglia. As unit recordings indicate, when a saccade is produced, all of these structures always agree on the direction of the movement, even when conflict is likely. Indeed, all cells whose field contains the goal of that movement always fire in unison. This makes it difficult for the investigator to detect where the decision to move has been made in the brain. Patricia Goldman-Rakic has promoted the idea that there is a "distributed neural network subserving spatially guided behavior." With her collaborators, she has demonstrated the existence of multiple direct interconnections between frontal and parietal cortical areas controlling spatial behavior. At the same time, she pointed out that the thalamic regions communicating with these cortical areas were contiguous or overlapping. In fact, in the past, the central thalamus, not yet suspected of any oculomotor role, was known as the site from where the most widespread, synchronized electrocortical potentials, the so-called recruiting waves, were produced by 10 Hz stimulation. These waves progressively invade most of the frontal and parietal lobes. They may have been the first sign, not well understood then, that the central thalamus operates as a hub. Today, further support for this view is found in the fact that electrical stimulation of the IML in the central thalamus is used for rescuing brain-injured patients from a minimally responsive state.

See also: Basal Ganglia and Oculomotor Control; Brainstem Control of Eye Movements; Cerebellum and Oculomotor Control; Cortical Control of Eye Movements; Eye and Head Movements; Oculomotor Control: Anatomical Pathways; Oculomotor System: Models; Perception and Eye Movements; Saccade–Pursuit Interactions; Saccadic Eye Movements; Superior Colliculus; Target Selection for Pursuit and Saccades; Thalamus: Evolution in Vertebrates.

Further Reading

Goldman-Rakic P and Porrino LJ (1985) The primate mediodorsal (MD) nucleus and its projection to the frontal lobe. *Journal of Comparative Neurology* 242: 535–560.

Jones EG (1985) *The Thalamus*. New York and London: Plenum.

Leichnetz GR (1980) An anterogradely-labeled prefrontal cortico-oculomotor pathway in the monkey demonstrated with HRP gel and TMB neurochemistry. *Brain Research* 198: 440–445.

Lynch JC, Hoover JE, and Strick PL (1994) Input to the primate frontal eye field from the substantia nigra, superior colliculus, and dentate nucleus demonstrated by transneuronal transport. *Experimental Brain Research* 100: 181–186.

Lynch JC and Tian J-R (2006) Cortico-cortical networks and cortico-subcortical loops for the higher control of eye movements. *Progress in Brain Research* 151: 461–501.

Schlag J and Schlag-Rey M (1986) Role of central thalamus and supplementary eye field in voluntary control of gaze in space. *Bulletin of Tokyo Metropolitan Institute for Neurosciences* (1986 supplement): 17–31.

Schlag-Rey M and Schlag J (1989) The central thalamus. In: Wurtz RH and Goldberg ME (eds.) *The Neurobiology of Saccadic Eye Movements*, pp. 361–390. New York: Elsevier.

Sommer MA and Wurtz RH (2003) What the brain stem tells the frontal cortex. I. Oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. *Journal of Neurophysiology* 91: 1381–1402.

Sommer MA and Wurtz RH (2003) What the brain stem tells the frontal cortex. II. Role of the SC-MD-FEF pathway in corollary discharge. *Journal of Neurophysiology* 91: 1403–1423.

Tanaka M (2005) Involvement of the central thalamus in the control of smooth pursuit eye movements. *Journal of Neuroscience* 25: 5866–5876.

Tanaka M (2006) Inactivation of the central thalamus delays self-timed saccades. *Nature Neuroscience* 9: 20–22.

Tanibuchi I and Goldman-Rakic PS (2004) Comparison of oculomotor neuronal activity in paralaminar and mediodorsal thalamus in the Rhesus monkey. *Journal of Neurophysiology* 93: 614–619.

Tian J-R and Lynch JC (1997) Subcortical input to the smooth and saccadic eye movement subregions of the frontal eye field in Cebus monkey. *Journal of Neuroscience* 199: 9233–9247.

Watanabe Y and Funahashi S (2004) Neuronal activity throughout the primate mediodorsal thalamus during oculomotor delayed-responses. I. Cue-, delay-, and response-period activity. *Journal of Neurophysiology* 92: 1738–1755.

Wyder M, Massoglia DP, and Stanford TR (2003) Quantitative assessment of the timing and tuning of visual-related, saccade-related, and delay period activity in primate central thalamus. *Journal of Neurophysiology* 90: 2029–2052.

Wyder M, Massoglia DP, and Stanford TR (2004) Contextual modulation of central thalamic delay-period activity: Representation of visual and saccadic goals. *Journal of Neurophysiology* 91: 2628–2648.

Thalamus: Evolution in Vertebrates

A B Butler, George Mason University, Fairfax, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The diencephalon comprises the hypothalamus, epithalamus, and thalamus, the latter consisting of the ventral thalamus, or subthalamus, and the dorsal thalamus. Both the ventral thalamus and the dorsal thalamus vary greatly across vertebrates in their degree of development and, to some extent, even in their component parts.

In mammals, nuclei of the ventral thalamus generally can be divided into motor system-related nuclei, which influence the degree of inhibition of motor system-related nuclei within the dorsal thalamus, and visual system-related nuclei, which influence eye movement control. Mammalian ventral thalamic nuclei do not directly project to the telencephalon, in contrast to dorsal thalamic nuclei that relay ascending sensory and related information to the telencephalon. How all of these nuclei evolved has been difficult to fathom due to the many differences in their anatomical organization and degree of elaboration in comparison to other amniotes and, to a much greater degree, across anamniotes.

Evolution of the Ventral Thalamus

Ventral Thalamic Organization in Anamniotes

The ventral thalamus comprises at least three nuclei in most anamniotes, all of which are most prominently related to the visual system. In cartilaginous fishes (sharks, skates, and rays) and ray-finned fishes (generally referred to as bony fishes), these three nuclei are called nuclei intermedius, ventromedius, and ventrolateralis (Figure 1). They all receive direct visual input from the retina. Generally similar nuclei are present in amphibians, although whether their homology to the ventral nuclei of fishes is discretely one-to-one remains to be clarified. Within their ventral thalamus, amphibians have a nucleus ventromedialis and two divisions of a nucleus ventrolateralis. They also have a heavily retinorecipient nucleus of Bellonci in a more lateral position, which extends caudally as the neuropil of Bellonci. While most connections of the ventral thalamic nuclei of anamniotes are unknown, for the most part (the exception being a somatosensory pathway in a ray-finned fish), they do not appear to relay sensory information to any part of the telencephalon, in marked contrast to dorsal thalamic nuclei, particularly in amniotes.

An important recent finding is that the dorsal thalamic nuclei of amphibians that do project to the telencephalon may not receive direct sensory input, at least from the retina or the somatosensory nuclei of the medulla. Rather, they may receive inhibitory input from the ventral thalamic nuclei, the latter serving as the only diencephalic sites for direct receipt of these inputs. This aspect of dorsal thalamic organization and its implications for evolution of the two main divisions of the dorsal thalamus across all vertebrates will be considered later.

Ventral Thalamic Organization in Amniotes

Visual ventral thalamus In reptiles (Figure 2) and birds, the ventral thalamus contains ventromedial and ventrolateral nuclei, about which little is known, and a major visual system nucleus, the pars ventralis of the lateral geniculate nucleus (LGNv). The latter nucleus consists of a lateral neuropil region that borders a more medially lying cell plate. It receives a variety of visual inputs, including from the optic tectum (the homolog of the mammalian superior colliculus) and the retina. In reptiles, a related nucleus lies immediately dorsal to the LGNv, the intergeniculate leaflet, homologous to the nucleus of the same name in mammals, and an additional, visual system-related nucleus, nucleus ovalis, is present at a more rostral level. In birds, the intergeniculate leaflet is referred to as the nucleus of the marginal optic tract. Mammals likewise have an LGNv, which in primates is known as the pregeniculate nucleus.

None of these visual components of the ventral thalamus in amniotes projects to the telencephalon but all contribute to brain stem systems involved in the regulation of eye movements. In mammals, this influence is known to occur via connections with the pretectum and superior colliculus and is likely similar in reptiles and birds. Whether the ventral thalamus is involved in eye movement regulation in fishes has not been established, but the receipt of visual inputs combined with outputs to brain stem rather than telencephalic sites is a ventral thalamic hallmark that clearly was established early in its evolutionary history and has been elaborated particularly within the amniote radiation.

Motor ventral thalamus Motor system-related nuclei within the ventral thalamus have not been identified in anamniotes. In mammals, the main motor system-related component of the ventral thalamus is the

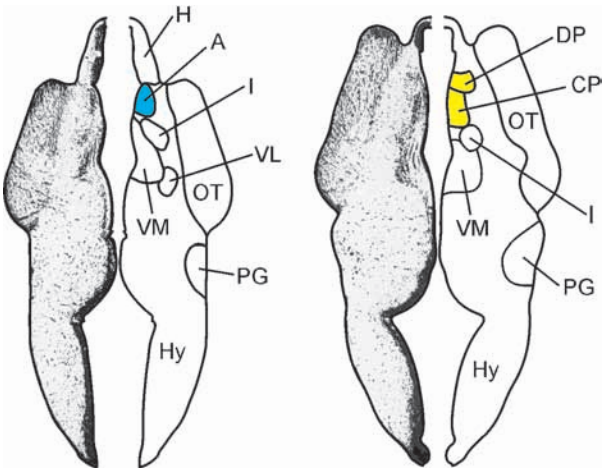

Figure 1 Transverse hemisection with mirror image drawings through part of the diencephalon of a ray-finned fish, the bowfin *Amia calva*, with the more rostral section on the left, to show the main dorsal and ventral thalamic nuclei. Within the dorsal thalamus, the lemnothalamic nuclei are shown in blue and the collothalamic nuclei in yellow. A, nucleus anterior; CP, nucleus centralis posterior; DP, nucleus dorsalis posterior; H, habenula; Hy, hypothalamus; I, nucleus intermedius; OT, optic tract; PG, preglomerular nuclear complex; VL, nucleus ventrolateralis; VM, nucleus ventromedialis.

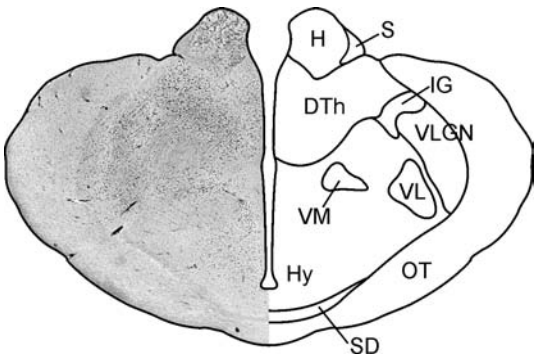

Figure 2 Transverse hemisection with mirror image drawings through part of the diencephalon of an iguana *Iguana iguana* to show the main ventral thalamic nuclei. DTh, dorsal thalamus; H, habenula; Hy, hypothalamus; IG, intergeniculate leaflet; OT, optic tract; S, stria medullaris; SD, supraoptic decussation; VL, nucleus ventrolateralis; VLGN, ventral lateral geniculate nucleus; VM, nucleus ventromedialis.

subthalamic nucleus, which uniquely (1) receives GABAergic input from the external segment of the globus pallidus (GPe), (2) contains glutamatergic neurons, and (3) projects to the internal segment of the globus pallidus (GPi). This nucleus is a key component of the so-called indirect loop formed by projections from the ventral anterior (VA) and ventral lateral (VL) nuclei of the dorsal thalamus to neocortex, thence to the caudate-putamen, thence to GPe, thence to the subthalamic nucleus, thence to GPi, and thence and back to VA and VL. The effect of subthalamic nuclear input to GPi is an increase in the firing of the GPi neurons, which themselves are inhibitory to the neurons within VA and VL, thus decreasing the excitatory thalamic input to cortex and thus producing an inhibitory effect on the initiation of movements.

A nucleus that contains glutamatergic neurons and has the same connections with the globus pallidus and the dorsal thalamus has been identified in both reptiles and birds and was thus evolutionarily established at the origin of amniotes if not earlier. In birds, this nucleus was previously called the anterior nucleus of the ansa lenticularis, but a recent update in avian neuroanatomical nomenclature by The Avian Brain Nomenclature Forum recommended using the name of subthalamic nucleus for this nucleus in birds, since its homology with the subthalamic nucleus of mammals is well established. In reptiles, the same glutamatergic cell group is called the anterior entopeduncular nucleus.

Thalamic reticular nucleus In amniotes, an additional component of the ventral thalamus, the thalamic reticular nucleus, is present. In mammals and birds, this nucleus is composed fully or predominantly of GABAergic neurons. In mammals, it is well established that these neurons receive inputs from various specific dorsal thalamic nuclei and from the various neocortical areas to which these dorsal thalamic nuclei project. In turn, the thalamic reticular nucleus projects back to the dorsal thalamic nuclei, thus serving as a brake and perhaps a firing pattern-coordinating mechanism on an otherwise purely positive feedback loop of glutamatergic projections from the dorsal thalamus to neocortex and reciprocally back to dorsal thalamus. The thalamic reticular nucleus of mammals has been proposed to be a crucial component of the thalamocortical neural circuitry involved in the generation of consciousness.

In crocodiles, a thalamic reticular nucleus is present but seems to be differently constituted than in birds and mammals. While lying in a similar location and having similar projections to the dorsal thalamus, the thalamic reticular nucleus in crocodiles contains several subpopulations of neurons, only one of which is GABAergic. Thus, this nucleus may have undergone a process of parallel evolution in the lines leading to extant birds and mammals, independently gaining a high percentage of GABAergic neurons in both lineages.

The possible role of the thalamic reticular nucleus in the production of consciousness in mammals and birds makes its evolutionary history of particular interest. GABAergic neuronal populations have been

described within various dorsal and ventral thalamic locations in a variety of anamniotes, including lampreys, some ray-finned fishes, and amphibians. Ventral thalamic neurons in amphibians are the most likely source of inhibitory inputs identified to the dorsal thalamic nuclei, and it is thus highly likely that these neurons are GABAergic and possibly homologous to the thalamic reticular nucleus of amniotes. Thus, such a nucleus, either as a discrete cytoarchitectonic entity or as scattered population(s) of neurons, may likewise be present in fishes and an evolutionary ancient attribute of the thalamus of vertebrates. The presence of such a nucleus in amphibians and possibly other anamniotes would indicate that the relative paucity of GABAergic neurons within the thalamic reticular nucleus of crocodiles might represent a secondary specialization within this taxon rather than being representative of the basal amniote condition.

Evolution of the Dorsal Thalamus

The dorsal thalamus varies greatly across vertebrates in the degree of its development and its importance in the relay of ascending sensory information to the telencephalon. Also, the sites within the telencephalon, to which its various nuclei project, vary markedly. Since in mammals the dorsal thalamic nuclei constitute an exclusive gateway for the relay of sensory information to most parts of the dorsal pallium in the telencephalon, which in mammals mainly comprises neocortex, dorsal thalamic evolution is crucially related to neocortical evolution. The dorsal thalamus in mammals is the most elaborate in terms of number of nuclei and their connections across all vertebrates, which one might well expect given the complexity of the mammalian telencephalon but which also made deciphering its evolutionary history a highly challenging problem. In order to explicate dorsal thalamic evolution across vertebrates, its organization in extant anamniotes and amniotes is considered first. Then the two basic divisions of the dorsal thalamus that are present in anamniotes – the lemnothalamus and the collothalamus – are defined, followed by an examination of whether these two divisions can also be recognized across amniotes. Finally, the differential evolution of these two divisions across tetrapods is discussed.

Dorsal Thalamic Organization in Anamniotes

Among anamniotes, the connections and histochemistry of the dorsal thalamic nuclei in the jawless fishes – lampreys and hagfishes – have not been studied in enough detail to allow meaningful comparisons with those of jawed vertebrates. In jawed fishes, the dorsal thalamus generally comprises three nuclei (Figure 1) – nucleus anterior, which lies in the most rostral part of the dorsal thalamus, and nuclei dorsalis posterior and centralis posterior, which occupy the caudal part of the dorsal thalamus. In cartilaginous fishes, as studied in a species of shark, nucleus anterior receives direct retinal input and projects to a part of the pallium in the telencephalon called the central nucleus. The dorsal posterior nucleus received visual input relayed to it via the optic tectum and also projects to the telencephalon, to site(s) in the subpallium and/or pallium. Also, in another shark species, auditory evoked potentials have been recorded from another part of the pallium, possibly relayed there through the central posterior nucleus. In a skate, nucleus anterior has been reported to project to the medial pallium, a proposed homolog of the hippocampal formation of amniotes. While most thalamotelencephalic projections are ipsilateral, the latter projection is notable for being bilateral.

In contrast to the situation in mammals, the dorsal thalamus in ray-finned fishes is not the only set of nuclei by which ascending sensory information is relayed to the telencephalon. A collection of nuclei not present in other vertebrate groups, the preglomerular nuclear complex (Figure 1), which is a part of a more caudal diencephalic component called the posterior tuberculum, serves as the main relay for a number of the major ascending pathways, particularly for the lateral line and gustatory systems. Parts of the hypothalamus and, at least in one species, ventral thalamic nuclei also participate in the ascending relay of some sensory information to the telencephalon. The dorsal thalamic nuclei relay some information to the telencephalon as well, especially for the visual and auditory systems. As is apparently also the case for cartilaginous fishes, the dorsal posterior and central posterior nuclei receive these visual and auditory inputs, respectively, from the midbrain roof. A number of the ascending projections, whether relayed via the preglomerular nuclear complex, the dorsal posterior and central posterior nuclei of the dorsal thalamus, or other sites, terminate in the medial zone of the pallium, which in ray-finned fishes is called area dorsalis. Due to differences in the embryological development of the pallium in ray-finned fishes, this medially lying pallial region actually corresponds topologically to the lateral region of the pallium in the other major vertebrate groups and is a good candidate for a homolog of the mammalian amygdala, particularly its basolateral, thalamorecipient part. Visual projections relayed directly from the retina through nucleus anterior to a more lateral part

of the pallium in ray-finned fishes have been reported in goldfish. The possible homology of this part of the pallium remains unresolved, but it corresponds topologically to part of the dorsal pallium of amniotes.

In amphibians, the rostral part of the dorsal thalamus, nucleus anterior, projects robustly to the medial pallium, and this projection, like that reported in skates, is bilateral. In contrast to the situation in fishes, retinal as well as somatosensory inputs to nucleus anterior are relayed to it from neighboring ventral thalamic nuclei, and the input of the latter appears to be inhibitory. The more caudal part of the dorsal thalamus in amphibians is composed of several nuclei, and, like the situation in fishes, receives its predominant inputs from the midbrain roof. This part of the dorsal thalamus projects predominantly to the striatum in the ventrolateral wall of the telencephalon, with only a slight projection reaching the most ventral part of the lateral pallium.

Two Divisions of the Dorsal Thalamus in Anamniotes

Based on the preceding survey, several points about dorsal thalamic organization in anamniotes can be made for the general case: (1) the dorsal thalamus is divided into two parts, a more rostral part usually termed nucleus anterior and a more caudal part that comprises the dorsal posterior and central posterior nuclei in most fishes and several comparable nuclei in amphibians; (2) nucleus anterior receives direct retinal (and other sensory system) projections, at least in fishes, and (3) projects to part(s) of the dorsal and/or medial (hippocampal) pallium; (4) telencephalic projections from nucleus anterior are frequently bilateral; (5) the more caudal part of the dorsal thalamus receives its predominant inputs from the midbrain roof, and (6) may project to the striatum as well as to the lateral part pallium, in amphibians the projection being almost all to the striatum; and (7) all known telencephalic projections from the caudal part of the dorsal thalamus are ipsilateral only.

Based on the preceding listed differences between the more rostral part of the dorsal thalamus (the nucleus anterior of anamniotes) and its more caudal part, two fundamental divisions can be recognized – a rostrally lying lemnothalamus and a more caudally lying collothalamus (Figure 1). The term 'lemnothalamus' is in accord with the previously established usage of the word lemniscus – the prefix 'lemno' meaning a ribbon and the term 'lemniscus' applied to fiber tracts that originate from a secondary sensory cell group (such as the dorsal column nuclei or the retinal ganglion cells) and ascend directly to the thalamus, that is, without first synapsing in any part of the midbrain roof. Thus, the lemnothalamus comprises the nucleus or nuclei within the rostral part of the dorsal thalamus that receive their sensory (or other) inputs directly, that is, without relay through the midbrain roof. The term 'collothalamus' refers to the source of its predominant inputs being the midbrain roof, which in mammals is referred to as the colliculi.

Dorsal Thalamic Organization in Amniotes

Collothalamus in amniotes From the studies of Harvey Karten and his co-workers and colleagues in the 1960s and 1970s, homologies between some dorsal thalamic nuclei in sauropsids (birds and reptiles) and mammals were strongly indicated based on similarities of their predominant afferent inputs from the midbrain roof. At that time, they were not yet recognized to constitute the collothalamus, since the basic organization of the dorsal thalamus into two divisions had not yet been realized. Nonetheless, a nucleus in receipt of visual projections from the superficial part of the rostral midbrain roof, called nucleus rotundus in both birds and reptiles (Figure 3), was identified as a homolog of the tectorecipient part of the nucleus lateralis posterior/pulvinar (LP/pul) complex of mammals. A nucleus in receipt of somatosensory projections from the deep part of the rostral midbrain roof, called the caudal part of nucleus dorsolateralis posterior (cDLP) in birds and various terms in reptiles, including nucleus medialis posterior (Figure 3), was identified as a homolog of part of the tectorecipient posterior nuclear group (PO) of mammals. A nucleus in receipt of auditory projections from the caudal part of the midbrain roof, called nucleus ovoidalis in birds and various terms in reptiles, including nucleus medialis (Figure 3), was identified as a homolog of the mammalian medial geniculate nucleus (MGN), particularly its ventral division.

The telencephalic projections of these nuclei in sauropsids and mammals are of crucial significance to their postulated relationships of homology, and in the past decade a considerable degree of debate has arisen concerning this matter. In mammals, LP/pul, PO, and MGN project to various areas of the lateral part of neocortex, its occipitotemporal/insular regions, but they also generally project to the basolateral part of the amygdala and, in addition to these pallial targets, generally to the striatum. For those that project strongly to neocortex, their projection to the amygdala, where present, is minor and vice versa. In sauropsids, nuclei rotundus, ovoidalis, and cDLP generally also project to the striatum, but they additionally project to only one pallial target, a specific locus for each within a large structure called

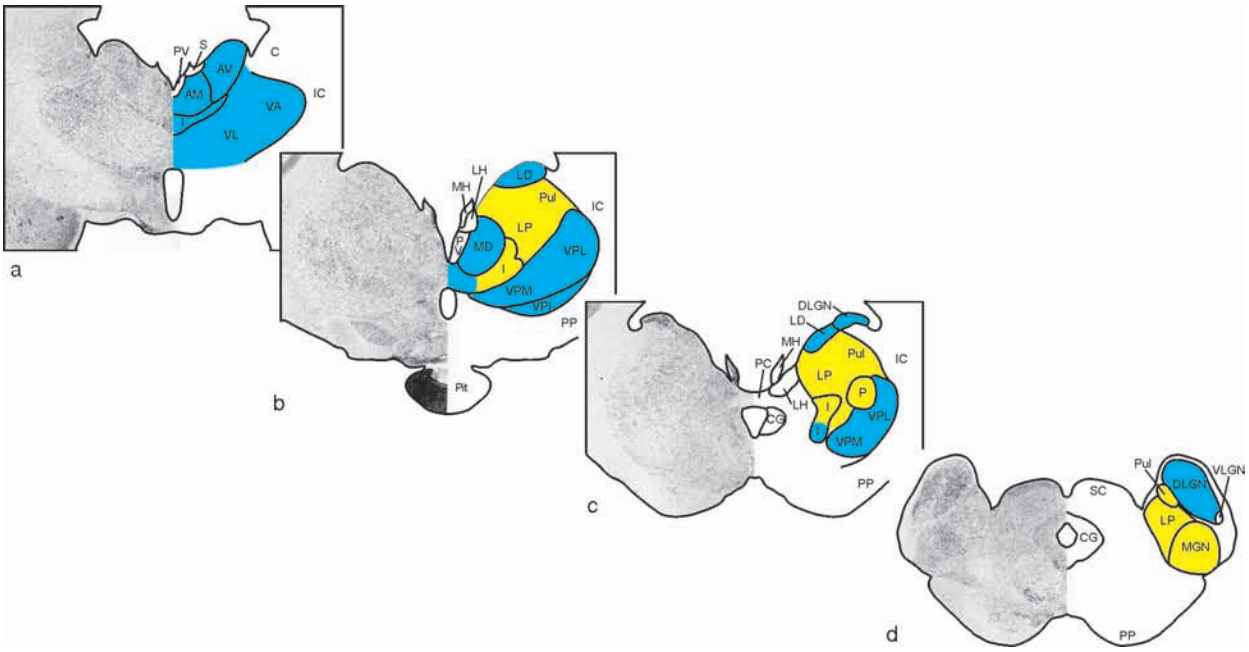

Figure 3 Transverse hemisections with mirror image drawings through four levels of the dorsal thalamus of an adult mammal, the raccoon *Procyon lotor*, with a most rostral, showing most of the lemnothalamic (in blue) and collothalamic (in yellow) nuclei. The intralaminar nuclei are divided into rostral, lemnothalamic and caudolateral, collothalamic groups, as indicated by their blue or yellow shading, respectively, with the former including the central medial, rhomboid, and center median nuclei and the latter the parafascicular, paracentral, and central lateral nuclei. AM, anteromedial nucleus; AV, anteroventral nucleus; DLGN, dorsal lateral geniculate nucleus; C, caudate nucleus; CG, central gray; I, intralaminar nuclear group; IC, internal capsule; LD, lateral dorsal nucleus; LH, lateral habenula; LP, lateral posterior nucleus; MD, mediodorsal nucleus; MGN, medial geniculate body; MH, medial habenula; P, posterior nuclear group; PC, posterior commissure; PP, pes pedunculi; Pul, pulvinar; PV, paraventricular nuclear group of the epithalamus; S, stria medullaris; SC, superior colliculus; VA, ventral anterior nucleus; VL, ventral lateral complex; VPL, ventral posterolateral nucleus; VPM, ventral posteromedial nucleus; VPI, ventral posterior inferior nucleus. Photomicrographs and guidance on nuclear boundaries were generously provided by Wally Welker.

the dorsal ventricular ridge (DVR) in reptiles and to specific loci within a part of the same structure that is referred to as the nidopallium in birds. The question is therefore whether the latter part of the pallium in sauropsids is homologous to the lateral part of mammalian neocortex, to the basolateral amygdala, to both, or, as also proposed, to neither. The nidopallium/reptilian thalamorecipient DVR have long been regarded as homologs of specific thalamorecipient and related cell populations (specifically those of layer IV and of V–VI) within neocortex. Recent embryological evidence of gene expression patterns supports a reinterpretation such that the nidopallium/reptilian thalamorecipient DVR are actually homologous to the basolateral amygdala of mammals. However, the latter interpretation does not offer a cogent hypothesis for the evolutionary origin of the laterally lying neocortical areas that also receive these ascending collothalamic inputs. An alternative explanation is that the nidopallium/reptilian thalamorecipient DVR are homologous as a field to both the basolateral amygdala and the lateral neocortex.

Since the homologous relationships of the dorsal thalamic nuclei that project to these various targets are linked to the homologous relationships of their pallial target(s), resolution of this major issue will require further experimental evidence to illuminate the evolutionary history with full clarity. While nucleus ovoidalis (and its reptilian homologs) is generally accepted as the homolog of MGN, whether it is comparable to the ventral division of MGN or to one or more of its other components, such as its medial, magnocellular component, is debated. More vigorously contested is whether the nucleus rotundus of sauropsids is homologous to the LP/pul complex of mammals or, along with cDLP, to one or more components of the posterior nuclear group.

That these various nuclei in sauropsids and mammals (Figures 3 and 4) constitute a division of the dorsal thalamus – the collothalamus – is, however, quite well established, whatever the outcome of the homology issues among the individual nuclei. They share several crucial features with the caudal part of the dorsal thalamus of anamniotes, including receipt of their predominant inputs via relays through the

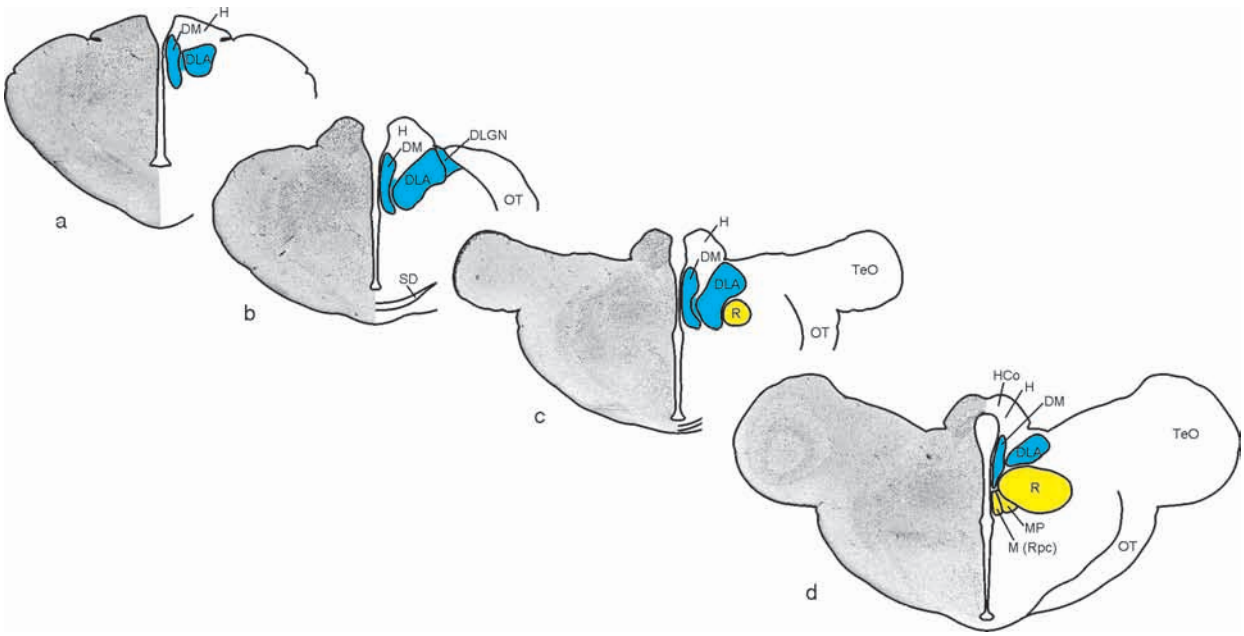

Figure 4 Transverse hemisections with mirror image drawings through four levels of the dorsal thalamus of an iguana *Iguana iguana*, with A most rostral, showing most of the lemnothalamic (in blue) and collothalamic (in yellow) nuclei. DLGN, dorsal lateral geniculate nucleus; DLA, nucleus dorsolateralis anterior; DM, nucleus dorsomedialis; H, habenula; HCo, habenular commissure; M (Rpc), nucleus medialis (reuniens pars compacta); MP, nucleus medialis posterior; OT, optic tract; R, nucleus rotundus; SD, supraoptic decussation; TeO, optic tectum. Identification of the DLGN based on Kenigfest N, Martínez-Marcos A, Belekhova M, et al. (1997) A lacertilian dorsal retinorecipient thalamus: A re-investigation in the Old World lizard *Podarcis hispanica*. *Brain, Behavior and Evolution* 50: 313–334.

midbrain roof and projections to the telencephalon that are generally to both the striatum and to the pallium. Also, all projections of these nuclei to the telencephalon are ipsilateral only. Recent gene expression also has supported the recognition of the collothalamus in embryonic mice; in contrast to the more rostrally situated lemnothalamus, it is positive for calretinin and negative for the gene *Math4a*.

Lemnothalamus in amniotes Due to the complexity and marked elaboration of the lemnothalamus in amniotes, the key to recognizing that it is a unitary entity and consistent across all jawed vertebrates lay in realizing that its various attributes overlap in their distribution across different nuclei and different groups of nuclei (Figures 3 and 4). First, one can recognize that the nuclei that do not receive their predominant input from the midbrain roof form a diverse set in both sauropsids and mammals. Then, the question becomes whether enough commonalities exist across them and in what pattern to allow for recognition of them as constituents of a fundamental dorsal thalamic division.

Aside from the collothalamic components of the mammalian dorsal thalamus already discussed, this part of the brain also contains a diverse array of nuclei and nuclear groups, which can only be briefly summarized here. Most rostrally in position in adult mammal brains is the anterior nuclear group, which is mostly interconnected with limbic areas (hippocampal and related cortices) but does receive a direct if sparse retinal input on its lateral aspect. Caudally adjacent to the anterior nuclear group is the medial nuclear group, the most prominent nucleus of which is the dorsomedial nucleus. This nucleus receives a major input from the olfactory cortex and also from the ventral pallidum, and it projects to prefrontal cortex. Like the dorsomedial nucleus, a relatively small nuclear group, the midline nuclei, also receive ventral striatopallidal inputs related to the limbic system. In contrast to the specific sensory relay nuclei of the dorsal thalamus, the intralaminar nuclei give rise to widespread projections to neocortex, albeit that these projections are individually restricted in their points of origin but nevertheless blanket the cortex in their range of projection coverage. Among a diverse set of afferents to the intralaminar nuclei is the rostral part of the reticular formation, essential for the maintenance of conscious wakefulness. In a more lateral position is the ventral nuclear group, which contains a number of nuclei involved in the somatosensory and motor feedback systems. The VA and VL both receive inhibitory motor-related inputs from the GPi and also inputs from the substantia

nigra pars reticulata and deep cerebellar nuclei, respectively. These nuclei project to motor cortical areas. The more caudally lying ventroposterior lateral (VPL) nucleus is the main somatosensory relay nucleus for the body with input from the dorsal column nuclei, and a nucleus medial to it, the ventroposterior medial nucleus (VPM), serves the same function for the trigeminal system from the head. These nuclei project to the primary somatosensory cortex. Finally, the dorsal lateral geniculate nucleus (DLGN) receives a massive visual input directly from the retina and projects to striate cortex.

One component of the puzzle solution for dorsal thalamic organization derives from embryological evidence published in Jerzy Rose in 1942 on the development of the dorsal thalamus in rabbits. Rose identified five cell groups in the developing dorsal thalamus that he called pronuclei and that ultimately give rise to all of its component, individual nuclei. All collothalamic nuclei are derived from two pronuclei – the dorsal and medial geniculate pronuclei – with the latter giving rise to the MGN and the former to both PO and LP/pul. The origin of the latter two nuclei from just one pronucleus is consistent with the enigmatic and seemingly entwined evolutionary origin of these nuclei and, correspondingly, of the lateral part of neocortex and the basolateral amygdala, as discussed above. The other three pronuclei identified by Rose are the medial and central pronuclei and the lateral geniculate pronucleus. The latter gives rise to the DLGN, while the medial pronucleus gives rise to the medial nuclear group and the central pronucleus to the anterior, intralaminar, and ventral nuclear groups. (Some of the caudally lying intralaminar nuclei are now assigned to the collothalamus, based on their predominant midbrain roof afferent input, but, consistent with Rose's embryological analysis, the rostral intralaminar nuclei are clearly lemnothalamic components.) The facts that (1) one pronucleus, the central, gives rise to both the anterior and ventral nuclear groups, (2) the lateral geniculate pronucleus lies rostrally early in development (rather than caudally as in the adult condition) and (3) shares a direct retinal input, albeit sparse, with part of the anterior group, and the sharing of other commonalities of both afferent and efferent connections, including (4) at least some bilateral projections to the telencephalon and (5) a conspicuous lack of any projections to the striatum, indicate that the derivatives of the central, medial, and lateral geniculate pronuclei of mammals indeed forms a set of rostral thalamic nuclei that comprise an embryologically coherent division of the dorsal thalamus.

Similarly in sauropsids, the set of dorsal thalamic nuclei that are not part of the collothalamus also have overlapping and shared features that correspond to those in mammals and allow their identification as a coherent dorsal thalamic division. Recent findings indicate a number of rather precise, if not completely one-to-one, comparisons between the various rostrally lying, noncollothalamic nuclei in sauropsids and mammals. Components of the formerly called nucleus opticus principalis thalami (OPT) in birds and a homologous cell group in reptiles are clearly homologous in a one-to-one manner to the DLGN of mammals. Two other nuclei in reptiles, nucleus dorsolateralis anterior (DLA) and nucleus dorsomedialis (DM), are homologous as a field to multiple nuclei in birds. The latter avian nuclei include two similarly named nuclei and other nearby nuclei that, as a group, are called the dorsal thalamic zone (DTZ) as well as nucleus dorsalis intermedius ventralis anterior (DIVA) and the ventrointermediate area (VIA). The connections and location of the medial components of the DTZ nuclei resemble those of the mammalian rostral intralaminar, midline, and mediodorsal nuclei. DIVA receives ascending somatosensory input from the dorsal column nuclei and thus resembles the mammalian VPL. VIA receives motor system-related inputs from GPi, the substantia nigra pars reticulata, and deep cerebellar nuclei and thus closely resembles both VA and VL of mammals. Thus, there is now substantial evidence that this set of nuclei in birds is homologous as a field to the DLA-DM nuclei of reptiles and also homologous as a field to the full set of lemnothalamic nuclei (i.e., all besides the DLGN) identified above in mammals. Finally, as noted above for the collothalamus, recent developmental data supports the concept of the lemnothalamus as an embryological entity. The developing lemnothalamus in mice is distinguished from the collothalamus in its negativity for calretinin and in being positive for expression of the gene *Math4a*.

Differential Evolution of the Collothalamus and Lemnothalamus in Tetrapods

Dramatic changes occurred for both the collothalamus and the lemnothalamus at the origin of amniotes, and whether the condition in extant amphibians reflects the ancestral tetrapod condition or, at least in some respects, a secondarily modified condition remains to be resolved. In extant amphibians, the collothalamus receives its predominant input from the midbrain roof, as is the case in all other vertebrate groups and thus clearly the basal ancestral condition. However, in amphibians it projects massively to the striatum and only very sparsely to the most ventral part of the lateral pallium. If the latter condition also is ancestral for tetrapods, then a dramatic shift of most of these projections to pallial targets would

have occurred at the origin of amniotes. This shift would have occurred in conjunction with elaboration of both the lateral parts of neocortex and the basolateral amygdala in mammals and of the DVR/nidopallium in sauropsids.

The amphibian lemnothalamus also is somewhat unique among vertebrates. It lacks direct receipt of retinal and dorsal column sensory inputs, in contrast to the situation in fishes and amniotes. In fact, in its connections the amphibian lemnothalamus more closely resembles a portion of the DLA-DM nuclei of reptiles and the corresponding lemnothalamic components in birds and mammals than it does the DLGN-VPL nuclei of mammals and the DLGN-DIVA nuclei of birds. Nonetheless, the amphibian nucleus anterior is clearly the lemnothalamus, and it projects robustly and bilaterally to the medial pallium. Thus, if this telencephalic projection pattern represents the ancestral tetrapod condition, then a second dramatic shift in connections occurred with the origin of amniotes, from predominantly medial pallium to predominantly dorsal pallium – the dorsal, or general, cortex in reptiles, the visual and somatosensory-motor Wulst and related pallial areas in birds, and the primary visual, somatosensory, motor, and prefrontal cortices of mammals. The bilaterality of the lemnothalamic projections has been maintained to some extent in sauropsids and decreased markedly in mammals. Nevertheless in mammals, some bilateral lemnothalamic projections persist and represent a marked contrast to the dedicatedly ipsilateral collothalalmic projections.

The evolutionary split between mammals and sauropsids was correlated with differential evolution of the lemnothalamus and collothalamus. Extant reptiles are characterized by a markedly elaborated collothalamus but with a modest lemnothalamus. However, while ancestral birds retained the elaborated collothalamus, they also elaborated the lemnothalamus, gaining specific nuclei such as DIVA and VIA that respectively mimic the specific lemnothalamic nuclei VPL and VA/VL of mammals. In contrast, early mammals favored elaboration of the lemnothalamus, as is retained in extant monotremes (platypus and echidna). Subsequently, beginning in marsupials and in many of the orders of placental mammals, the collothalamus was elaborated in correlation with the elaboration of the many supernumerary association visual cortices and several association somatosensory, insular association cortical areas. Thus, while the sauropsid line initially favored the collothalamus and its corresponding pallial areas, and the mammalian line initially favored the lemnothalamus and its corresponding pallial areas, within both lineages the other part of the thalamopallial system became elaborated subsequently – in birds within sauropsids and in many mammalian orders.

See also: Evolution of Vertebrate Brains; Sleep and Sleep States: Thalamic Regulation; Thalamus and Oculomotor Control.

Further Reading

Butler AB (1994) The evolution of the dorsal thalamus of jawed vertebrates, including mammals: Cladistic analysis and a new hypothesis. *Brain Research Reviews* 19: 29–65.

Butler AB (1995) The dorsal thalamus of jawed vertebrates: A comparative viewpoint. *Brain, Behavior and Evolution* 46: 209–223.

Butler AB (2006) The dual elaboration hypothesis of the evolution of the dorsal thalamus. In: Krubitzer LA and Kaas JK (eds.) *Evolution of Nervous System in Mammals*. Amsterdam: Elsevier.

Butler AB and Hodos W (2005) *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation* (2nd edn.). Hoboken: Wiley.

Csillag A and Montagnese CM (2005) Thalamotelencephalic organization in birds. *Brain Research Bulletin* 66: 303–310.

Desfilis E, Font E, Belekhova M, et al. (2002) Afferent and efferent projections of the dorsal anterior thalamic nuclei in the lizard *Podarcis hispanica* (Sauria, Lacertidae). *Brain Research Bulletin* 57: 447–450.

Jones EG (1985) *The Thalamus*. New York: Plenum.

Kenigfest N, Martínez-Marcos A, Belekhova M, et al. (1997) A lacertilian dorsal retinorecipient thalamus: A re-investigation in the Old World lizard *Podarcis hispanica*. *Brain, Behavior and Evolution* 50: 313–334.

Medina L, Veenman CL, and Reiner A (1997) Evidence for a possible avian dorsal thalamic region comparable to the mammalian ventral anterior, ventral lateral, and oral ventroposterolateral nuclei. *Journal of Comparative Neurology* 384: 86–108.

Nieuwenhuys R, ten Donkehaar HJ, and Nicholson C (1998) *The Central Nervous System of Vertebrates*. Berlin: Springer.

Reiner A, Yamamoto K, and Karten HJ (2005) Organization and evolution of the avian forebrain. *The Anatomical Record Part A* 287: 1080–1102.

Roth G, Grunwald W, and Dicke U (2003) Morphology, axonal projection pattern, and responses to optic nerve stimulation of thalamic neurons in the fire-bellied toad *Bombina orientalis*. *Journal of Comparative Neurology* 461: 91–110.

Sherman SM and Guillery RW (2001) *Exploring the Thalamus*. San Diego: Academic Press.

Steriade M, Jones EG, and McCormick DA (1997) *Thalamus*. Amsterdam: Elsevier.

Veenman CL, Medina L, and Reiner A (1997) Avian homologues of mammalian intralaminar, mediodorsal and midline thalamic nuclei: Immunohistochemical and hodological evidence. *Brain, Behavior and Evolution* 49: 78–98.

The AIM Model of Dreaming, Sleeping, and Waking Consciousness

J A Hobson, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When we go to sleep at night and dream, our conscious state changes radically. When we first become drowsy, fleeting dream images may occur. We then become oblivious and hard to rouse from deep sleep. Later in the night, when sleep lightens, we may experience vivid, bizarre dreams from which spontaneous or easily induced awakenings may occur. By studying the changes in the brain which are associated with these changes in our state of consciousness, we may learn more about the physiological mechanisms underlying our states of mind.

During waking we can attend to events in the outside world, form perceptions of events, and/or direct our thoughts to those percepts or to other matters, such as the review of our behavior, our plans, or our feelings. When we go to sleep, we become unaware of the outside world and lose the ability to construct thoughts. These perceptual and cognitive defects persist even when we dream: perceptions are then almost entirely internally generated and we have difficulty reasoning about them; we almost always believe that we are awake instead of dreaming.

From the vantage point of the cognitive neuroscience of consciousness, we can tentatively explain these phenomena as a function of three factors:

1. Factor A: the level of electrical activation in the brain.
2. Factor I: the status of gating information flow to and from the brain.
3. Factor M: the nature of the mode of information processing within the brain.

The level of consciousness is thus low or high, depending upon the level of activation (A). The focus or direction of consciousness is either strongly external or strongly internal, depending on whether the input–output gates of the brain are open or closed to the flow of information (I). The mode (M) of information processing within the brain is set by the ratio of aminergic to cholinergic modulation. Since these three factors can be quantified using experimental data, it is possible to construct a three-dimensional state space model (Figure 1). The state of consciousness, a point with values A, I, and M, varies over time, the fourth dimension of the model. Normal sleep cycles appear as elliptical trajectories in the resulting state space.

Definition of Consciousness

For the purpose of the conscious state paradigm, consciousness can be simply defined as awareness of the outside world, our bodies, and ourselves. Consciousness can be further distinguished in two ways. The first distinction is between primary consciousness (which does not depend upon language and may well be shared with many other mammals) and secondary consciousness (which does depend upon language and may be a uniquely human capability). The second distinction is between the components of consciousness, which are shown in Table 1. These components of consciousness vary significantly in strength over the normal wake–sleep–dream cycle.

The functions listed in Table 1 are the aspects of consciousness which are commonly studied by cognitive neuroscientists. Here we investigate the common mechanisms by which all of these aspects are changed as a function of the changes in conscious state experienced by us as waking, sleeping, and dreaming.

Characteristics of Consciousness

Consciousness is graded and its components vary in strength as the brain changes state over the course of each day of each life and over a lifetime. In other words, consciousness is more or less intense and its components change in relative strength as a function of the sleep–wake cycle. For example, in waking, perception can be exteroceptive, thought can be logical, and memory can be good. By contrast, in dreaming, all three of those cognitive functions are altered; internal perception is enhanced, emotion is intensified and confabulation runs wild, and logical thought and memory are greatly impaired. Table 2 lists some of the state-dependent consciousness components.

Changes in the Brain and in the Body during Sleep; Changes in Consciousness

As shown in Figure 2, sleep onset is associated with thalamocortical deactivation. This deactivation is signaled, first, by electroencephalogram (EEG) slowing (stage I) and next by definitive blockage of thalamic transmission with the appearance of EEG spindles (stage II). After that there is a further slowing in frequency and an increase in amplitude of cortical slow waves (stages III and IV). Muscle tone decreases passively, postural shifts stop, and eye movements are greatly reduced as this pattern evolves. During the evolution of these stages of so-called nonrapid eye movement (NREM) sleep, individuals may have

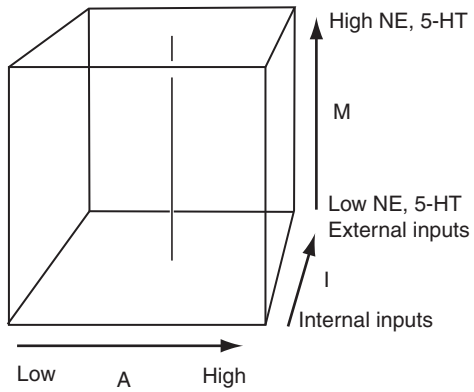

Model factor	Psychological	Neurobiological
A-Activation: energy level of processing capacity	• Word count • Cognitive complexity, e.g., perceptual vividness, emotional intensity, narrative	• EEG activation • Firing level and synchrony of reticular, thalamic and cortical neurons
I-Information: source internal or external	• Real world space, time and person referents and their stability • Real vs. imagined action	• Level of presynaptic and postsynaptic inhibition • Excitability of sensorimotor pattern generators
M-Mode: organization of data	• Internal consistency? • Physical possibility? • Linear logic?	• Activity level of aminergic neurons

Figure 1 The activation/input source/neuromodulation (activation/information/mode; AIM) model. Illustration of three-dimensional state space and the psychological neurobiological correlates of each dimension.

Table 1 Components of consciousness

Component	*Definition*
Attention	Selection of input data
Perception	Representation of input data
Memory	Retrieval of stored representations
Orientation	Representation of time, place, and person
Thought	Reflection upon representations
Narrative	Linguistic symbolization of representations
Emotion	Feelings about representations
Instinct	Innate propensities to act
Intention	Representations of goals
Volition	Decisions to act

fleeting dreams as their brains are disconnected from sensory input, but they soon become oblivious with little or no memorable mental activity. Measures of this deactivation process yield low values of factor A in the AIM model. At the same time that factor A decreases, factors I and M are also decreasing as the sensory gates close and aminergic modulation decreases.

After 70–80 min, often signaled by a posture shift, the EEG is reactivated and reverses its downward path, moving rapidly up through stages III and II to enter stage I. At the same time the electromyogram (EMG) is actively inhibited and the electooculogram (EOG) shows the increasingly clustered and intense eye movements that give REM sleep its name. The values of factor A activation increase to waking levels (and beyond), but sleep persists. But both factors I and M go to their lowest values in REM as the activated brain is actively put offline and further demodulated. The alternation of NREM and REM sleep continues throughout the night at 90- to 100-min intervals. As the night progresses the NREM periods become shorter and less deep while the REM periods become longer and more active. In the last two cycles the alternation is usually between stage II and stage I REM. This increasing brain activation is associated with an increase in dreaming. Although dreaming is always more intense in REM, it also occurs in NREM sleep, especially in the second half of the night.

Formal Analysis of Dream Content

In order to compare with waking the nature of conscious experience in REM and light NREM sleep, it has

Table 2 State-dependent consciousness components

Function	*Nature of difference*	*Causal hypothesis*
Sensory input	Blocked	Presynaptic inhibition
Perception (external)	Diminished	Blockade of sensory input
Perception (internal)	Enhanced	Disinhibition of networks storing sensory representations
Attention	Lost	Decreased aminergic modulation causes a (decrease in) signal-to-noise ratio
Memory (recent)	Diminished	Because of aminergic demodulation, activated representations are not restored in memory
Memory (remote)	Enhanced	Disinhibition of networks storing mnemonic representations increases access to consciousness
Orientation	Unstable	Internally inconsistent orienting signals are generated by cholinergic system
Thought	Reasoning *ad hoc*; logical rigor weak; processing hyperassociative	Loss of attention, memory, and volition leads to failure of sequencing and rule inconstancy; analogy replaces analysis
Insight	Self-reflection lost (failure to recognize state as dreaming)	Failure of attention, logic, and memory weakens second- and third-order representations
Language (internal)	Confabulatory	Aminergic demodulation frees narrative synthesis from logical restraints
Emotion	Episodically strong	Cholinergic hyperstimulation of amygdala and related temporal lobe structures triggers emotional storms, which are unmodulated by aminergic restraint
Instinct	Episodically strong	Cholinergic hyperstimulation of hypothalamus and limbic forebrain triggers fixed-action motor programs, which are experienced fictively but not enacted
Volition	Weak	Top-down motor control and frontal executive power cannot compete with disinhibited subcortical network activation
Output	Blocked	Postsynaptic inhibition

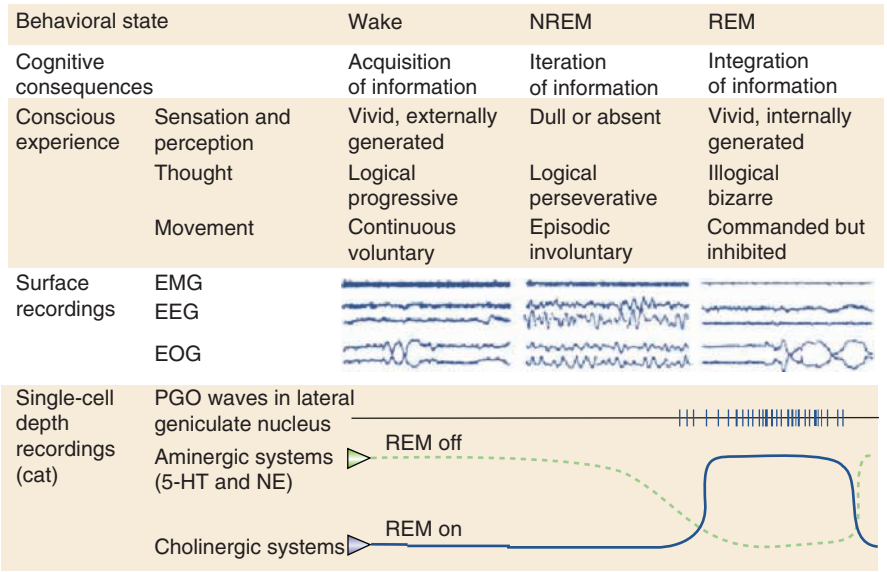

Figure 2 Behavioral states of humans. States of waking, nonrapid eye movement (NREM) sleep, and REM sleep have behavioral, polygraphic, and psychological manifestations. The sequence of these stages is represented in the polygraph channel. Sample tracings of three variables used to distinguish state are also shown: the electromyogram (EMG) is at its highest in waking, intermediate in NREM sleep, and lowest in REM sleep; the electroencephalogram (EEG) and electooculogram (EOG), which are both activated in waking and REM sleep and inactivated in NREM sleep. PGO, pontogeniculooccipital; 5-HT, 5-hydroxytryptamine (serotonin); NE, norepinephrine.

proved propitious to quantify such aspects of consciousness as (1) the sense modalities that are represented and the intensity of the internally generated percepts, (2) the occurrence and strength of self-reflective awareness and executive thought, (3) the dreamer's orientation in time and space, and dream personnel, (4) the dreamer's emotion profile (presence and strength of emotions such as elation, anxiety/fear, anger, shame, guilt, sadness, and eroticism), and (5) the identification by the dreamer of the possible memory sources of the dream content. This approach has led to the characterization of dream consciousness as having (1) strong sensorimotor perceptions (with a preponderance of vivid visual imagery as the dreamer moves through a fictive reality); (2) impoverishment of executive ego function (with the delusional belief that one is awake and a failure to develop or sustain logical thinking); (3) discontinuity and incongruity of plot features as the basis of bizarre cognition; (4) potentiation of anxiety, anger, and elation together with a weakening of shame, guilt, and sadness; and (5) inability to specify the memory source of the confabulated dream content in about 80% of the cases.

Dreaming is thus characterized by visual hallucinations, by disorientation, by lack of insight and judgment, and by memory loss (with compensatory confabulation). These formal similarities are so strong as to lead to the hypothesis that dreaming, although normal, is a state akin to delirium. Since delirium is caused by diseases and by drugs which interfere with normal neuromodulation, it is natural to wonder if, in normal sleep, there are changes in neuromodulation during the wake–sleep cycle.

Cellular and Molecule Mechanisms of Mammalian Sleep

All mammals share with humans the regular periods of NREM and REM sleep. This is so whether the animals are aquatic, terrestrial, or arboreal, and whether they are large surface-dwelling carnivores (e.g., lions) or small nest-dwelling herbivores (e.g., mice or rabbits). As an experimental model, the domestic cat has appealed to neurobiologists because it is so docile and so prodigious a sleeper. As experimental animals, cats are more expensive and more ethically problematic than are rats, but they are large enough to carry the microelectrode, microinjection, and microdialysis array, the equipment necessary to perform physiological and pharmacological experiments. They are well studied neuroanatomically so that stereotaxis is facilitated.

Using cats, neuroscientists have shown that the two brain-activated states, waking and REM sleep (with high values of A), are at opposite poles with respect to both input–output gating (high I in waking, low in REM) and neuromodulation (high M in waking, low M in REM). On all these dimensions, NREM sleep is quite different (low A and intermediate I and M). In addition to demonstrating that aminergic levels decline in NREM sleep and reach their nadir in REM, it has been shown that the cholinergic level declines as animals go from waking to NREM, but rises again and reaches its highest level, in REM sleep, especially in the brain stem. This gives factor M (defined as the ratio of aminergic to cholinergic modulation) a highly differentiating power for distinguishing the three states: M is highest in waking, intermediate in NREM sleep, and lowest in REM sleep. This finding is potentially relevant to our understanding of the brain basis of the many phenomenologic distinctions between waking, sleeping, and dreaming. The mode of information processing has changed dramatically.

The importance to REM of the pontine brain stem is demonstrated by the presence of the cell bodies of both cholinergic and aminergic neurons there. Furthermore, it has now been conclusively demonstrated that the microinjection into the pontine tegmentum of both cholinergic agonists and cholinesterase antagonists produces dramatic increases in REM sleep. All of these effects are blocked by pretreatment of the injection sites by atropine. There is no longer any doubt that compared to waking and NREM sleep, REM sleep is hypoaminergic and hypercholinergic. This robust physiological difference correlates with – and hypothetically causes – the psychological differences.

Based upon new studies in rats, a flip-flop switch for REM sleep triggering has recently been proposed. Each side of the switch contains γ-aminobutyric acid (GABA)ergic neurons, and this resulting mutual inhibition guarantees that when one side is on, the other is off. The REM-on population contains two populations of glutaminergic neurons: one population projects to the basal forebrain and produces the forebrain activation while the other projects to the medulla and spinal cord, producing atonia. The cholinergic system is activated as part of the REM-on switch and can, in turn, activate that side of the switch. The REM-off population includes the dorsal raphe nuclei. At a still deeper level, the activation of protein kinase in the pontine tegmentum has been shown to accompany REM sleep, suggesting that pedunculopontine intracellular protein kinase A (PKA) activation is involved in REM sleep generation.

Input–Output Gating

Michel Jouvet called REM sleep 'paradoxical' because sleep persists despite the EEG evidence for strong brain activation. A partial explanation of the paradox is

the simultaneous blockade of sensory input and motor output, the neurophysiological mechanisms of which are now well understood. Ottavio Pompeiano used classical Sherrington reflexology techniques to show that motor output is blocked in REM by descending inhibition from the pontomedullary brain stem to the anterior bone cells of the spinal cord. That this inhibition is induced by powerful and active postsynaptic inhibition of the final common path motor neurons was shown by Michael Chase. Chase also demonstrated that the spinal inhibition of REM is mediated by glycine.

As for the inhibition of sensory input, it was also Pompeiano who demonstrated that presynaptic inhibition is associated with the clusters of eye movement during REM sleep. An associated occlusion of sensory input excitation also accompanied the REMs. This co-option of input-processing channels (including the all-important thalamocortical system) is effected by the internally generated pontogeniculooccipital (PGO) waves, first described by Jouvet. These cortical activation waves were later shown to be of entirely internal origin and were enhanced by cholinergic microstimulation. PGO wave generation is inhibited by noradrenaline and serotonin.

The emerging picture of REM sleep is a high level of activation in a brain which is not only off-line by virtue of sensory input and motor output blockade, but is also stimulated by phasic activation signals of its own devising. The coactivated mind is therefore fooled by its own parasitic excitation; the brain-mind then develops the illusion or fictive experience of perceiving and moving through dream space.

Efferent Copy Generation

The PGO waves of REM sleep not only constitute internally generated phasic activation signals for the dreaming brain-mind, but they also convey to the thalamus and cortex feed-forward information about the direction of the upcoming eye movements. As such, they are classical efferent copy stimuli from the oculomotor circuits to the visual forebrain. Such eye movement activity can also be recorded in waking when it forms an integral part of the startle response. In waking, the PGO waves rapidly habituate as the startle response diminishes. The second and third stimuli of a train evoke a much weaker or no response because they are no longer novel. The mechanism of this attenuation is thought to be aminergic inhibition of the cholinergic-modulated PGO generator neurons of the lateral pons.

The relative continuity of waking consciousness is, in part, a function of PGO wave damping, whereas the bizarre discontinuity of dream consciousness is a function of PGO wave disinhibition. Dream consciousness reflects this dishabituation by the increase in the emotions of surprise and anxiety. These dream features could be mediated by excitation of the amygdala, a major target of the PGO waves. In REM sleep, the efferent copy information about the direction of eye movement that is conveyed by the PGO system is not integrated with changes in visual input, as occurs in waking, because in sleep there is no visual input. It is tempting to speculate that the efferent copy information that is generated in REM sleep is not only a physiological cause of dream surprise, but is also used in perceptual aspects of dream plot construction.

The oculomotor system of the pons is only one of many motor program generators that are activated in REM sleep. Locomotion is commanded but, because the anterior horn cells are hyperpolarized, real motor behavior is not generated and REM sleep dreams reflect this by the intensity of fictive motoric action in dreams. There is hardly a sentence in the REM sleep dream report that does not contain an action verb. We are always walking, running, flying, or swimming in our dreams. Even if we lead sedentary waking lives, stationary activity is not represented in our dreams.

The finding of motor pattern generator activation in REM sleep and its correlation with the fictive motility of dreaming is a good example of how the close study of neurophysiology can provide dream theory with a new paradigm and new hypotheses regarding the origin of dream plots. Of course, the approach does not allow us to propose why a given dreamer has a given dream or what that dream might mean. But it does provide encouragement for an alternative way of approaching even interpretive questions. For example, our dreams are full of movement because brain motor pattern generators are activated in REM sleep. Further evidence that this theory may be correct is given by experimental and clinical discoveries, both of which reveal unexpected and unwanted enactment of REM sleep motor commands. The experiments of Michel Jouvet and of Adrian Morrison showed that cats with lesions in the vicinity of the locus coeruleus in the pons evinced REM sleep without atonia. In other words, the cats showed sleep with EEG activation REMs but suddenly stood up and evinced a wide variety of attack and defense behaviors. They were then in a dissociated state of REM sleep in which they expressed automatic behaviors that are normally suppressed by the motor inhibition of REM. On the basis of these observations, Jouvet boldly proposed that the cats were acting out their dreams.

That idea gained credence when it was reported by Carlos Schenk and Mark Mahowald that humans

did precisely the same thing when they became afflicted by what they called the REM 'sleep behavior disorder.' Schenk's and Mahowald's patients, mostly >50-year-old males, performed motor acts in sleep that corresponded to their dream reports. One patient, dreaming that he was playing football, got up out of bed and tried to tackle a chest of drawers. Another, dreaming that he was diving, stood up on the side of the bed and propelled himself into the baseboard of his bedroom, which he imagined to be a swimming pool. Presumably, these experimental and natural pathological conditions of REM sleep without atonia are due to damage of that subpopulation of glutaminergic REM-on neurons that convey inhibitory commands from the flip-flop switch to the spinal cord.

These examples would seem to make more plausible the suggestion that dream consciousness is a function of brain activation in sleep and that the important differences between the dreaming and waking states of consciousness are determined by specifiable differences between the mechanisms of brain activation in the two states. In particular, it is the presence or absence in input–output gating that determines the relationship between the outside world and internal perceptions, beliefs, and actions.

Differences in Regional Brain Activation

EEG evidence had long suggested that the brain was similarly activated in waking and REM sleep dreaming. The recent advent of PET imaging technology revealed that this was an oversimplified conclusion. Most brain regions are less active in NREM sleep than in either the wake or REM sleep state. But when REM is compared to the wake state, differences emerge which are of great relevance to studies of conscious state determination.

Certain brain regions are more active in REM than in waking. They include the pontine tegmentum (which animal studies have shown to be the probable site of the REM sleep and PGO wave generator), the amygdala (which animal studies have shown to be involved in mediating emotion, especially the anxiety which is so common in dreaming), the parietal operculum (which is thought to mediate visuomotor integration of cardinal perceptual feature of dreaming), and the parahippocampal and deep frontal cortices (which may process emotional data and integrate it with cognitive information).

One brain region is notably less active in REM than in waking. The dorsolateral prefrontal cortex (DLPFC) remains at the same low level seen in NREM sleep. Other human studies have implicated the DLPFC in the executive ego functions that are so conspicuously deficient in dreaming: short-term memory, self-reflective awareness, intentional decision making, and volitional action. The brain lesion studies done by Mark Solms have revealed that two of the structures selectively activated in REM sleep are, when damaged, associated with a complete cessation of dream recall. They are the parietal operculum (damage to which causes problems with sensorimotor integration) and the deep frontal white matter (damage to which causes emotional apathy). Patients often complain of cessation of dreaming following lobotomy, which consists of severance of the deep frontal white matter.

Field Studies of Waking, Sleeping, and Dreaming Consciousness

The invention of the 'Nightcap', a two-channel event recorder, has made possible the field study of conscious experience in humans as they lead their otherwise normal waking and sleeping lives. In addition to wearing the Nightcap for recording their head and eye movements while asleep, participants carried radio beepers so that they could be contacted throughout the day and night to collect reports of conscious experience (see Figure 3(a)). This combination of techniques allowed us, for the first time, to collect reports of consciousness experience in the same persons, from several brain states, around the clock. Over 3000 reports from ten people ages 20–40 years were elicited from active and quiet waking states, from five intervals during sleep onset and from both NREM and REM sleep. The reports were transcribed and scored for dimensions such as word count (REM reports were 6 times as long as NREM reports), internally generated percepts (REM sleep highest, active waking lowest). These data, shown in Figure 3(b), suggest a general psychophysiological law: thinking is incompatible with hallucination and vice versa. The implications of these findings for conscious state control and for psychiatry are loud and clear. One striking finding was a reciprocal increase in internally generated perception and a decrease in thinking across the five states.

Convergence of Physiology, Psychology, and Philosophy

The integration of the fields of psychology and physiology, envisaged by William James and the young Freud over 100 years ago, is now beginning, with sleep and dream science leading the way. The Finnish philosopher of science Antti Revonsuo has recently championed the theory, first put forth by Michael Jouvet, that REM sleep dreaming allows the sleeping brain to run adaptive behavioral programs of fight and flight. According to this view, REM sleep dreaming has a Darwinian function even if dream behavior

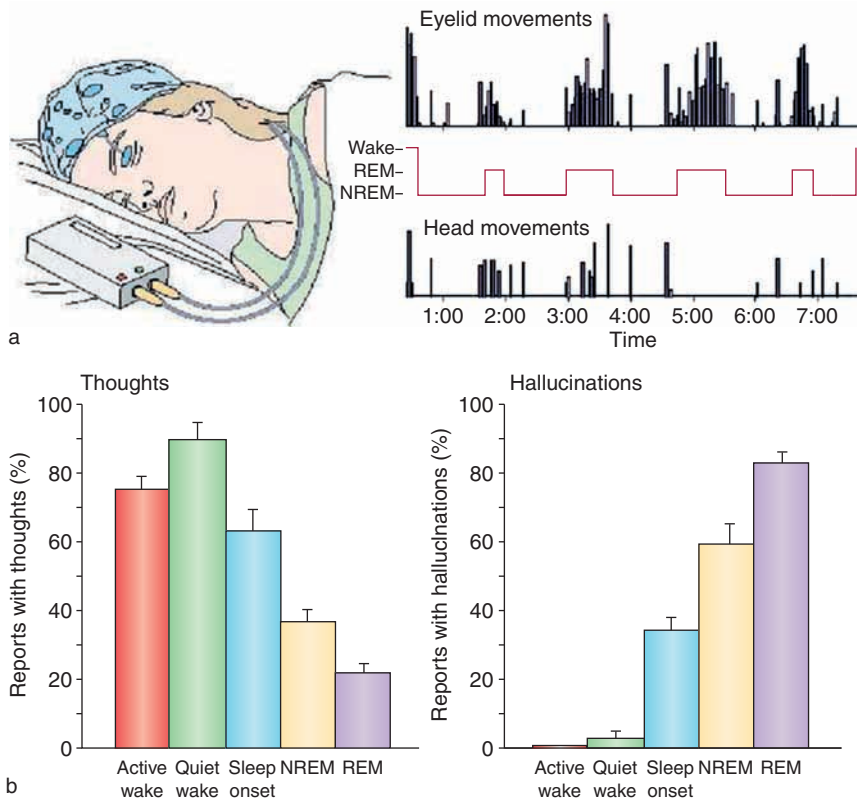

Figure 3 (a) Central arousal accompanying the activated states of rapid eye movement (REM) sleep and waking can be measured using the 'Nightcap,' a simple ambulatory monitor. The Nightcap is a two-channel recording device that distinguishes waking, REM and NREM sleep. One channel of the Nightcap monitors eye movement and the other monitors body movements. The Nightcap eyelid-movement readout is thought to reflect activity in portions of the brain stem oculomotor nucleus that innervate the eyelid and are adjacent to portions of the medial brain stem ascending reticular system, the activity of which, in turn, generates forebrain activation. (b) Decline in directed thought and reciprocal increase in hallucinations during progression from active waking through sleep onset and NREM to REM sleep.

is fictive. For the German philosopher Thomas Metzinger, the occurrence of lucid dreaming, out-of-body experience and dreaming itself, supports his phenomenal self-model of consciousness. The phenomenal self is a constantly renewed functional state of the brain whose vicissitudes can, for the first time in human history, begin to be explained.

AIM: An Integrated Model of the Human Brain Mind

Using physiological measures, it is now possible to create a three-dimensional model of brain-mind state space (see Figures 1 and 4). Activation, assessed from EEG and/or thalamocortical cellular activity, gives values and axis A, running from left (low) to right (high) across the front wall of the state space. Input–output gating is measured as EMG activity, values of H-reflex, and PGO waves. It runs from the front to the back of the state space. The vertical axis, M, measures the ratio of the strength of aminergic to cholinergic neuron modulation. Unfortunately, factor

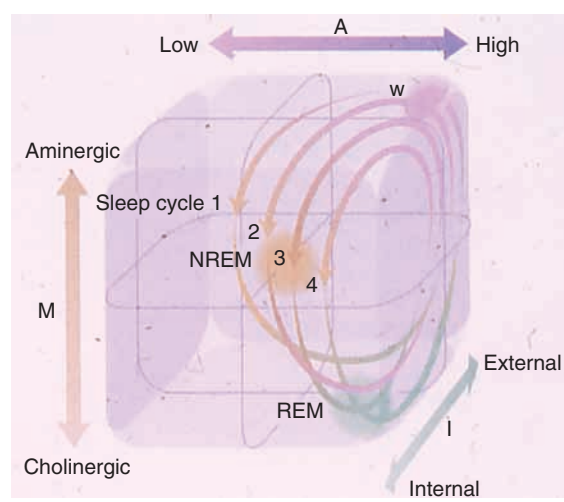

Figure 4 The three-dimensional activation/information/mode (AIM) state space model showing normal transitions within the AIM state space, from wake, to non-rapid eye movement (NREM), and then to REM sleep. REM occupies the lower right-hand front corner, in which activation is high, input is entirely internal, and the forebrain is cholinergically activated and aminergically demodulated. ACh, acetylcholine; NE, norepinephrine; 5-HT, 5-hydroxytryptamine.

M cannot yet be measured in humans, but pharmacological studies indicate that inferences from neurobiological studies in animals are valid.

In the model, time is the fourth dimension and the solution of AIM is constantly changing, even in the waking domain (at the far right, upper, rear corner of the state space) as we change the focus of attention, day dream, or enter altered states such as hypnosis. We may become drowsy, doze, or fall frankly asleep, in which case AIM moves down, forward, and to the left and then occupies the NREM sleep domain in the center of the state space. Finally, AIM moves to the REM sleep domain in the right front lower corner of the state space.

During a normal night of sleep, AIM function follows an elliptical trajectory that repeats itself 4 or 5 times per night. The ellipse shifts sequentially to the right and to deeper parts of the state space in successive cycles. These elliptical orbits in AIM state space conform to the mathematical description of the reciprocal interaction model, giving an internally consistent picture of sleep physiology.

Consciousness is most intense in the right portion of the state space and lowest on the left. It is interoceptive toward the front of the state space and exteroceptive toward the back. The kind of consciousness is given by the mode dimension. It is highest at the top of the state space and lowest at the bottom. Abnormal states of consciousness such as narcolepsy, the REM sleep behavior disorder, delirium, and other psychoses can also be modeled in the normally forbidden zones of AIM state space.

See also: Dream Function; Dreams and Dreaming: Incorporation of Waking Events; Dreams and Nightmares in PTSD; Dreams, Dreaming Theories and Correlates of Nightmares; Sleep and Sleep States: Phylogeny and Ontogeny; Sleep Mentation in REM and NREM: A Neurocognitive Perspective; Sleep Research and Sleep Medicine in Historical Perspective.

Further Reading

Amini-Sereshki L and Morrison AR (1986) Effects of pontine tegmental lesions that induce paradoxical sleep without atonia on thermoregulation in cats during wakefulness. *Brain Research* 384(1): 23–28.

Bandyopadhya RS, Datta S, and Saha S (2006) Activation of pedunculopontine tegmental protein kinase A: A mechanism for rapid eye movement sleep generation in the freely moving rat. *Journal of Neuroscience* 26(35): 8931–8942.

Hobson JA, Pace-Schott EF, and Stickgold R (2000) Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences* 23: 793–842.

Jouvet M (1962) Recherche sur les structures nerveuses et les mechanismes responsables des differentes phases du sommeil physiologique. *Archives Italiennes de Biologie* 100: 125–206.

Jouvet M (1973) Essai sur le reve. *Archives Italiennes de Biologie* 111: 564–576.

Jouvet M, Vimont P, and Delorme F (1965) Elective suppression of paradoxical sleep in the cat by monoamine oxidase inhibitors. *Comptes Rendus des Seances de la Societe de Biologie et de ses Filiales* 159(7): 1595–1599.

Lu J, Sherman D, Devor M, et al. (2006) A putative flip-flop switch for control of REM sleep. *Nature* 441(7093): 589–594.

Metzinger T (2003) *On Being No One*. Cambridge: MIT Press.

Revonsuo A (2005) *Inner Presence: Consciousness as a Biological Marker*. Cambridge, MA: MIT Press.

Schenck CH and Mahowald MW (1996) REM sleep parasomnias. *Neurological Clinics* 14: 697–720.

Solms M (1997) *The Neuropsychology of Dreams: a Clinico-Anatomical Study*. Mahwah, NJ: Lawrence Erlbaum Associates.

Yamuy J, Fung SJ, Xi M, et al. (2004) Hypocretinergic control of spinal cord motor neurons. *Journal of Neuroscience* 24(23): 5336–5345.

Thermoregulation during Sleep and Sleep Deprivation

R Szymusiak, University of California, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The regulation of sleep and the regulation of body temperature are tightly coupled. Sleep onset in mammals is accompanied by reductions in metabolic rate, accelerated heat loss, and an integrated fall in body temperature. Homeostatic responses of the thermoregulatory system to heat or cold challenge are suppressed during sleep, and in some species thermoregulatory adjustments are almost completely absent during rapid eye movement (REM) sleep. Long-term sleep deprivation in animals can be associated with chronically elevated metabolism and food intake, accompanied by significant decreases in core body temperature.

In addition to changes in thermoregulation during sleep, manipulation of environmental temperature or body temperature can alter sleep. In all mammals studied, sleep is suppressed in moderately warm or cold environments. This suppression of sleep by thermal stress can be viewed as adaptive from a homeostatic standpoint because thermoregulatory responses are diminished during sleep. In some circumstances, manipulation of body temperature can promote sleep. For example, in humans, whole-body warming by hot bath immersion a few hours prior to habitual bedtime can reduce sleep latency and increase stage 3/4 sleep amounts.

Neuronal systems that regulate sleep and body temperature are localized in the preoptic and anterior hypothalamus (POAH). Many thermo-sensing neurons in the rostral hypothalamus also exhibit changes in activity across the sleep–wake cycle. Of particular importance may be groups of hypothalamic warm-sensing neurons that are spontaneously activated during sleep onset and sleep. Thus, functional interactions between sleep-regulatory and thermoregulatory neurons in the rostral hypothalamus can explain the close coupling between body temperature and arousal states.

Understanding sleep and thermoregulatory interactions may provide insight into clinical problems. Recent evidence has demonstrated the importance of thermoregulatory adjustments occurring around the time of sleep onset in determining sleep latency, sleep composition, and sleep quality in humans. Therapies that target the thermoregulatory control of arousal states may provide novel approaches to treating disorders of initiating and maintaining sleep.

Thermoregulation during Sleep

Changes in Body Temperature and Metabolism during Sleep

In mammals, the onset of sleep is associated with a fall in core body temperature and a readjustment of thermoregulatory control. The magnitude of the drop in body temperature during sleep varies across species and can be influenced by variables such as ambient temperature and phase of the circadian rhythm. The sleep-related fall in body temperature during sleep is actively regulated and is not simply a passive consequence of behavioral inactivity. Sleep onset in mammals is accompanied by coordinated decreases in metabolic heat production and increases in heat loss through increased peripheral vasodilation and/or sweating.

The decrease in body temperature during sleep can be understood as a resetting of the body's thermostat to a lower level. In other words, sleep appears to be accompanied by a lowering of the thermal set point, the level of body temperature that the thermoregulatory system will defend and maintain. At the transition from waking to sleep, the thermoregulatory system responds as if the set point has been adjusted to a lower level. The transition is accompanied by an integrated series of thermoregulatory responses, including a reduction in metabolic rate and enhanced heat loss to the environment through peripheral vasodilation and sweating. Body temperature falls to a new regulated level which is maintained during sleep.

This lowering of the set point and consequent energy savings can be taken as evidence for a potential energy conservation function of sleep. Energy savings are achieved in two ways: (1) by lowering metabolic heat production and (2) by lowering body temperature. The lowering of body temperature reduces the core temperature-to-ambient temperature gradient, which serves to minimizes heat loss to the environment. This latter factor can be important in small mammals with a high surface-to-volume ratio, which have inherently higher rates of environmental heat exchange. Behavioral aspects of sleep onset such as seeking a thermally comfortable nest and adoption of curled heat-conserving postures can also contribute to energy savings by minimizing heat exchange with the environment.

Thermoregulatory Responses to Thermal Challenge Are Influenced by Sleep

The decline in the thermal set point during sleep onset and non-REM sleep is also accompanied by

reduced thermoregulatory responses to thermal challenges. Thermoregulation is not suspended during sleep, but the responsiveness of the thermoregulatory system appears to be diminished; that is, the magnitude of the regulatory response to a given thermal challenge is reduced during sleep compared to waking.

Compared to non-REM, thermoregulatory responses during REM sleep are reduced even further. This is particularly true for responses to cold stress because the inhibition of motor neurons that is characteristic of REM sleep is incompatible with increasing heat production via shivering or elevated muscle tone. As a result of these thermoregulatory changes, animals may be particularly vulnerable to thermal stress during REM sleep. Infants and small mammals with a large surface-to-volume ratio are most vulnerable.

Exposure to hot or cold environments frequently cause sleep disruption and sleep loss. This response may be adaptive in minimizing exposure to thermal stress during sleep when homeostatic compensation is weak. Thermal factors have undoubtedly played a role in the evolution of behavioral strategies for species-specific selection of sleeping environments (e.g., nests, sheltered locations, and adoption of heat conserving postures). Diminished thermoregulatory adjustments to heat and cold stress during sleep dictate that thermal homeostasis can be best achieved over the course of a several-hour sleep episode by the selection of a sleeping environment that poses minimal thermal challenge and ultimately maximizes heat conservation. This is a strategy that has been adopted by several species.

Effects of Environmental and Body Temperature on Sleep

Exposure to moderate hot or cold environments suppresses sleep and/or causes sleep fragmentation in mammals. The most sensitive species are those with large surface-to-volume ratios that experience more rapid environmental heat exchange. In rodents, in which thermoregulatory responses during REM sleep are absent or weak, REM sleep suppression can occur under conditions of minimal thermal stress and in the absence of any significant change in non-REM sleep amounts. From a thermoregulatory perspective, the REM-sleep-suppressing effects of heat and cold exposure in these species can be viewed as adaptive. Increased time spent awake or in non-REM sleep means that lower threshold and stronger thermoregulatory responses are evoked to maintain thermal homeostasis.

Mild increases in ambient temperature or mild warming of the body can promote sleep. An animal's thermoneutral zone is defined as the range of environmental temperatures at which resting metabolic rate is minimal and constant. Within the thermoneutral zone, minimal effort is required to maintain thermal homeostasis because the heat produced by the resting metabolic rate is largely balanced by heat loss to the environment. Within the thermoneutral zone, thermal homeostasis can usually be maintained through peripheral vasomotor adjustments that regulate rates of heat loss from the skin, with vasoconstriction occurring at the low end of the zone and vasodilation at the high end. Exposing rats to an environmental temperature that is at or just above the high end of the thermoneutral zone, at which peripheral blood flow and skin temperatures are high, increases time spent in non-REM sleep and enhances electroencephalogram (EEG) slow-wave activity during sleep.

Whole-body warming, achieved by immersion in a hot bath a few hours prior to normal bedtime, promotes sleep in humans. Hot baths prior to bedtime can reduce sleep latency and increase amounts of stage 3/4 sleep in the first half of the night. Similarly, exercise late in the day reduces sleep latency and increases stage 3/4 sleep. The sleep-enhancing effects of exercise are dependent on an accompanying increase in body temperature. If exercising subjects are prevented from raising their body temperature by being cooled with a fan during exercise, the sleep-promoting effects of exercise are prevented.

Collectively, these findings are consistent with the idea that the mild activation of thermoregulatory heat defense mechanisms is conducive to sleep, particularly sleep onset and slow-wave sleep. The critical feature is that the warming must be mild such that it evokes mild compensatory thermoregulatory responses (e.g., increased peripheral vasodilation). More severe heat challenges that evoke stronger regulatory responses such as intense sweating or panting cause sleep disruption, not sleep promotion.

Brain Circuits That Regulate Sleep and Body Temperature Are Functionally Interrelated

Temperature-Sensing Neurons in the Hypothalamus Control Body Temperature

Neurons in the POAH are critically involved in thermoregulation. Experimental damage to the POAH in animals causes persistent deficits in thermoregulatory defense against heat and cold exposure. Direct infusion of various neurotransmitters or neuromodulators into the POAH can alter body temperature and thermoregulatory function.

Specialized temperature-sensing neurons localized in the POAH play an important role in body temperature control. Warm-sensing neurons are excited by local increases in temperature and inhibited by cooling. Cold-sensing neurons are excited by local decreases in temperature and inhibited by local warming. The functional importance of these temperature-sensing neurons for thermoregulatory control is evidenced by experimental findings that localized warming or cooling of the POAH can evoke fully integrated heat defense or cold defense responses that are qualitatively identical to those evoked by exposing the whole animal to hot or cold environments.

Hypothalamic Temperature-Sensing Neurons Are Involved in Sleep Regulation

The neural control of sleep and thermoregulation intersect in the POAH and adjacent brain structures. Experimental POAH damage that disrupts thermoregulation also causes persistent insomnia, suggesting that thermo- and sleep-regulatory neuronal systems are anatomically related. They appear to be functionally related as well. As previously summarized, local warming or cooling of the POAH can evoke robust whole-body heat and cold defense responses. Milder increases in POAH temperature, in addition to suppressing metabolic rate and augmenting peripheral vasodilation, reduce sleep latency, promote non-REM sleep and increase EEG slow-wave activity during non-REM sleep. Local POAH cooling has the opposite effects, causing sleep suppression and suppression of EEG slow-wave activity during sleep.

Potential mechanisms underlying many of the effects of environmental and body temperature on sleep are revealed by the study of the spontaneous discharge patterns of POAH thermo-sensing neurons during waking and sleep. As documented by single-neuronal recordings in cats and rats, the majority of POAH warm-sensing neurons exhibit an elevation of spontaneous discharge rate during sleep onset and non-REM sleep compared to waking. The majority of cold-sensing neurons are maximally activated during waking and exhibit declining discharge during sleep onset and non-REM sleep. Thus, the changes in POAH thermo-sensing neuronal activity at sleep onset mimic the pattern expected to accompany a lowering of the thermal set point, with the activation of warm sensors and inhibition of cold sensors. This pattern of thermoreceptor activity is consistent with the thermoregulatory responses that normally accompany sleep onset (i.e., reduced metabolic rate and increased heat loss at the periphery). The ability of local POAH warming to evoke non-REM sleep and EEG synchrony indicates that the activation of POAH warm-sensing neurons is sufficient for sleep promotion.

Control of Body Temperature during Sleep Deprivation

Acute Sleep Deprivation: Interactions with Circadian Control of Body Temperature

In healthy individuals, entrained to a light–dark cycle and adhering to a regular sleep–wake schedule, habitual sleep onset reliably occurs a few hours after the 24 h body temperature maxima, on the falling phase of the 24 h body temperature rhythm. Sleep onset is associated with an acceleration in the rate of the decline in body temperature as the activation of heat-loss processes during sleep are expressed. Morning awakenings occur on the rising phase of the temperature rhythm, shortly after the 24 h temperature minimum. Thus, when individuals maintain a regular sleep schedule, sleep–wake rhythms and body temperature rhythms are closely coupled with a characteristic phase relationship.

The fall in body temperature that normally occurs during the sleep period is not a passive consequence of changes in muscle activity and heat production associated with waking versus sleep. The body temperature rhythm persists during sleep deprivation or during constant routine protocols in humans, in whom levels of activity, food intake, light exposure, and so on are maintained at constant levels throughout a 24 h period. Nevertheless, behavioral changes associated with waking and sleep do exert positive and negative masking effects on the circadian temperature rhythms, such that the magnitude of the 24 h peak-to-trough difference in body temperature is reduced during sleep deprivation or constant routine, compared to normal entrained conditions.

An important feature of arousal state regulation is that short-term sleep deprivation or sleep restriction leads to compensatory rebound increases in sleep amount and sleep depth once the opportunity for sleep is provided. Homeostatic increases in sleep following sleep deprivation/restriction involve complex changes in neurophysiological and neurochemical mechanisms that are not yet completely understood. One effect of short-term sleep loss is to change the normal phase relationship between the occurrence of sleep/wakefulness and the 24 h body temperature rhythm. Thus, extending wakefulness through the night and attempting to sleep the next morning results in sleep onset occurring on the rising phase of the body temperature rhythm.

The phase relationship between sleep and body temperature rhythms at sleep onset is an important determinant of subsequent sleep amount and composition. This has been most convincingly demonstrated in human subjects undergoing internal or forced desynchronization in environments without time cues. During internal desynchronization, sleep and body temperature rhythms free run with different periods, and eventually the two rhythms temporally dissociate. In forced desynchrony protocols, subjects adhere to a sleep–wake schedule significantly longer than 24 h (e.g., 28 h), and when this is done in the absence of time cues, body temperature and sleep rhythms dissociate as in internal desynchronization. With both methods, over the course of several days the onset of the major sleep period occurs at all phases of the body temperature rhythm. Sleep episodes that are initiated on the falling phase of the temperature rhythm are associated with shorter sleep latencies, few episodes of waking during the sleep period and maximal amounts of deeper sleep stages. In contrast, sleep episodes initiated on the rising phase of the temperature rhythm are associated with longer sleep latencies, short sleep episode durations, more REM sleep but less stage 3/4 sleep. Thus, depending on the timing of sleep loss and recovery sleep, the adverse effects of short-term sleep deprivation include impaired alertness during sustained waking and impaired depth and quality of recovery sleep. The effects of partial sleep loss are cumulative, and the level of sleepiness will increase over several days of sleep restriction, due, in part, to the inadequate restorative effects of suboptimal recovery sleep.

Chronic Sleep Deprivation Causes Dysregulation of Body Temperature in Rats

The effects of long-term total sleep deprivation on thermoregulation have been studied in rats by Rechtschaffen and colleagues. Total sleep deprivation applied over many days in the rat results in a syndrome that includes increased food intake; increased energy expenditure; weight loss; an initial mild elevation in body temperature, followed by profound decreases in temperature as the duration of deprivation increases; and death within 11–32 days (average ~17 days). Several findings indicate that sleep deprivation in rats causes a profound dysregulation in the homeostatic systems that control metabolism and body temperature. The increased food intake occurring during sleep deprivation is more than offset by increased energy expenditure, such that rats actually lose weight in spite of a near doubling of food intake. The increase in energy expenditure appears, at least in part, to be an attempt to compensate for increased heat loss. Ultimately, rates of heat loss are so high that the core body temperature falls to well below normal levels, and rats become profoundly hypothermic during the last few days of survival. This fall in body temperature is not regulated; that is, there is no apparent lowering of the set point for body temperature during sleep deprivation. When sleep-deprived rats are given the opportunity to control the temperature of their environment, they select higher temperatures from the beginning of deprivation and select extremely high temperatures later in the deprivation period when they are hypothermic. Thus, they behaviorally try to raise body temperature to combat the hypothermia evoked by unregulated elevations of heat loss. The mechanisms responsible for the changes in metabolism and thermoregulation evoked by sleep deprivation are not understood, but findings do suggest that sleep is necessary for the proper functioning of these regulatory systems.

Thermoregulatory Control of Human Sleep

Thermoregulatory Changes Influence Sleep Onset and Sleep Maintenance

Neurobiological studies in experimental animals have revealed close anatomical and functional interrelationships between sleep and body temperature-control mechanisms in the brain, and they indicate that the activation of hypothalamic neuronal circuits that promote heat loss normally occurs during sleep onset and non-REM sleep. In humans, the magnitude of the heat-loss response around the time of sleep onset is an important determinant of sleep latency, sleep amount, and sleep quality. The magnitude of heat-loss response in humans can be reliably assessed by quantifying the temperature gradient from distal skin sites (the hands and feet) to proximal skin sites (e.g., the forehead, stomach, thigh, or some weighted average of several proximal skin sites). A high distal-to-proximal skin temperature gradient occurs when the temperature of the feet and hands are warm with respect to proximal sites, indicative of elevated blood flow to distal skin and elevated heat exchange at these sites. A low distal-to-proximal gradient reflects peripheral vasoconstriction and diminished peripheral heat loss. A high distal-to-proximal skin temperature gradient is associated with reduced sleep latency, increased stage 3/4 sleep, and reduced number of arousals from sleep. The distal-to-proximal gradient is also positively correlated with subjective sleepiness after morning awakenings and after awakenings from daytime naps.

It can be hypothesized that thermoregulatory responses that normally accompany sleep onset in humans play a role in reinforcing and consolidating circadian and homeostatic aspects of sleep regulation. In an entrained individual, habitual sleep-onset times occur on the descending phase of the body temperature rhythm. Homeostatic sleep drive is also high at this time because of sustained prior waking. As body temperature falls following the nadir of the 24 h rhythm, the activation of heat-loss mechanisms result in elevated skin temperatures. This leads to the activation of peripheral warm sensors and excitation of hypothalamic warm-sensing neurons via peripheral-to-central thermosensory pathways. The activation of central warm-sensors increases sleep propensity via suppression of arousal systems and hasten sleep onset, resulting in further increases in peripheral skin temperature and heat loss. Thus, thermoregulatory changes have the potential to exert feed forward excitatory effects on brain mechanisms that control sleep onset and sleep maintenance, leading to more rapid and seamless transitions from waking to consolidated sleep.

Thermoregulation and Disordered Human Sleep

Insomnia is the most common sleep complaint, and sleep disturbance is a prominent symptom in several psychiatric and neurological disorders. In many instances, insomnia complaints are accompanied by abnormalities in circadian rhythms, including abnormalities in the body temperature rhythm. Sleep-onset insomniacs may exhibit a phase delay in the body temperature rhythm of >2 h, indicating a disturbance of circadian timing or entrainment. This phase delay means that attempts to fall asleep often occur in closer than normal proximity to the body temperature peak and that thermoregulatory adjustments that promote sleep might not be as readily evoked. In the elderly, phase delays in the body temperature rhythm may also be a contributing factor to the increased incidence of insomnia. In affective disorders, the amplitude of the 24 h body temperature rhythm is reduced. The diminished nocturnal peak in the temperature rhythm results in a lower body temperatures at bedtime and a weaker heat-loss response at the time of sleep onset, resulting in more prolonged and disrupted wake-to-sleep transitions.

Although body-temperature abnormalities in disorders associated with insomnia may be secondary to other underlying pathologies, the manipulation of thermoregulatory control, either pharmacologically or behaviorally (e.g., by hot bath immersion), in ways that promote heat loss at sleep onset may help to improve sleep amount and sleep quality.

See also: Autonomic Dysregulation During REM Sleep; Endocrine Function During Sleep and Sleep Deprivation; Energy Homeostasis: Thermoregulation; Hypothalamic Structure–Function Relationships; Immune Function During Sleep and Sleep Deprivation; Metabolic Syndrome and Sleep; Sleep and Sleep States: Hypothalamic Regulation; Sleep Deprivation and Brain Function; Sleep Deprivation: Neurobehavioral Changes; Sleep in Aging; Thermoregulation: Autonomic, Age-Related Changes.

Further Reading

Alam MN, McGinty D, and Szymusiak R (1995) Neuronal discharge of preoptic/anterior hypothalamic thermosensitive neurons: Relation to nonrapid eye movement sleep. *American Journal of Physiology* 269: R1240–R1249.

Bach V, Telliez F, and Libert J-P (2002) The interaction between sleep and thermoregulation in adults and neonates. *Sleep Medicine Reviews* 6: 481–492.

Czeisler CA, Weitzman E, Moore-Ede MC, Zimmerman JC, and Knauer RS (1980) Human sleep: Its duration and organization depend on its circadian phase. *Science* 210: 1264–1267.

Gilbert SS, van den Heuvel CJ, and Ferguson SA (2004) Thermoregulation as a sleep signaling system. *Sleep Medicine Reviews* 8: 81–93.

Glotzbach SF and Heller HC (1976) Central nervous regulation of body temperature during sleep. *Science* 194: 537–539.

Heller HC (2005) Temperature, thermoregulation and sleep. In: Kryger M, Roth T, and Dement W (eds.) *Principles and Practice of Sleep Medicine,* 4th edn., pp. 292–304. Philadelphia: Elsevier Saunders.

Horne JA and Reid AJ (1985) Night time sleep EEG changes following body heating in a warm bath. *Electroencephalography and Clinical Neurophysiology* 60: 154–157.

Karuchi K, Cajochen C, Werth E, and Wirz-Justice A (2000) Functional link between distal vasodilation and sleep onset latency? *American Journal of Physiology* 278: R741–R748.

Karuchi K, Cajochen C, and Wirz-Justice A (2004) Waking up properly: Is there a role of thermoregulation in sleep inertia? *Journal of Sleep Research* 13: 121–127.

McGinty D and Szymusiak R (1990) Keeping cool: A hypothesis about the mechanisms and functions of slow-wave sleep. *Trends in Neuroscience* 13: 480–487.

McGinty D and Szymusiak R (2001) Brain structures and mechanisms involved in the generation of NREM sleep: Focus on the preoptic hypothalamus. *Sleep Medicine Reviews* 5: 323–342.

Parmeggiani PL and Rabini C (1970) Sleep and environmental temperature. *Archives of Italian Biology* 108: 369–387.

Rechtschaffen A, Bergman BM, Everson CA, Kushida CA, and Gilliland MA (1989) Sleep deprivation in the rat, X: Integration and discussion of findings. *Sleep* 12: 68–87.

Shaw PJ, Bergman BM, and Rechtschaffen A (1997) Operant control of ambient temperature. *American Journal of Physiology* 272: R682–R690.

Thermoregulation: Autonomic, Age-Related Changes

R B McDonald, University of California, Davis, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The inability to appropriately thermoregulate during exposure to heat and cold has been generally accepted as an age-related change in humans. Epidemiological observations of increased mortality during severe cold and hot weather also emphasize a possible age-related attenuation in thermoregulation. It remains unclear, however, the extent to which biological factors impact clinical hyper- and hypothermia in senescent humans. A majority of case studies have found that diminished thermoregulatory capacity in the senescent human reflects more closely socioeconomic factors such as insufficient living area heating or cooling, malnutrition, and lack of adequate health care. Nonetheless, experimental evidence in humans and laboratory rodents has reported senescent-related alteration in autonomic mechanisms that regulate body temperature. In this article, we will briefly explore possible affects of biological aging on the autonomic control of thermoregulation.

'Problems' of Interpretation of Aging Research

Definition of Aging

Research into the mechanisms underlying the aging process, including altered autonomic regulation of body temperature, poses experimental design concerns that do not often occur in other disciplines. Lack of a generally accepted definition of aging should be viewed as primary among these concerns. The difficulty in defining aging arises from an inability to determine when aging begins. Biological aging should include all events occurring to a particular system throughout the life span. Evolutionary, genetic, and developmental data have clearly established a link between development and the rate of aging. Although some investigations have begun to approach aging from this broad perspective, such experimental designs are difficult to construct and restrict the investigator to observational studies. Narrowing the definition of biological aging to a postreproductive period in which biological deterioration can be clearly identified increases mechanistic observations. However, since narrow definitions of aging do not include early life events, the true cause and effect relationship may be missed. Nonetheless, Dr. Edward Masoro's definition of biological aging has provided a good guide for most aging research:

> Aging can be defined as deteriorative changes with time during post-maturational life that underlie and increase vulnerability to challenges, thereby decreasing the ability of the organism to survive. (Masoro EJ, 1995: 3–21)

Designs in Research on Thermoregulation and Aging

Appropriately interpreting results from investigations evaluating the biology of senescence can be difficult for the nonspecialist. The difficulty in interpretation arises from the fact that extrinsic and behavioral factors such as physical activity level, acclimation, and disease have significant biological aging effects that, in turn, can affect thermoregulation. Cross-sectional designs, investigations that compare different age cohorts at one point in time, are the most common but least powerful design used in senescence research. The length of the human life span precludes, for the most part, longitudinal design except by large multiinstitutional programs (e.g., the Baltimore Longitudinal Study), and these investigations are limited to observational evaluation.

Cross-sectional designs can be problematic because selecting the appropriate older age group that represents accurately the 'normal' senescence process cannot be easily achieved; we return to the question, 'when does aging begin?' This problem reflects that the beginning of senescence in humans is highly variable, may not affect all physiological systems equally, and does not correlate with chronological age. For example, all 70-year-old humans will have some functional loss in at least one physiological system, but not all 70-year-old people will experience decrements in the same physiological system. Investigations using cross-sectional designs will control for individualized senescence by including an older age group with a wide range of ages (e.g., 65–80 years). Such a wide age range will often result in significant error about the mean as compared to the variation seen in younger groups.

Rats and mice have been the overwhelming animals used to model humans in senescence-related research evaluating possible alterations of thermoregulation. Their short life span (2–4 years) provides the investigator an opportunity for mechanistic study using powerful longitudinal designs. However, application of results to humans may be limited due the differences in effector organs responsible for heat dissipation and

Table 1 Tissues and organs of heat dissipation and production in humans, rats, and mice

Species	*Insulation*	*Heat dissipation (hot environment)*	*Heat production (cold environment)*
Human	Subcutaneous fat	Convection resulting from the transfer of heat from blood to skin throughout body Evaporation resulting from increased sweat gland function	Shivering via increased skeletal muscle metabolism
Rats and mice	Fur and air gap between skin and muscle	Convection resulting from the transfer of heat from blood to skin limited to tail	Nonshivering thermogenesis in brown adipose tissue Shivering via increased skeletal muscle metabolism

production (Table 1). It remains unclear as to how or if neural feedback, from the different sensor systems and effector organs of heat production and dissipation in rodents versus humans, may impact thermoregulation during senescence.

Disease and Its Effect on Thermoregulation during Senescence

The separation between disease and the biological process of aging presents yet another issue that must be considered when investigating senescent-related effects on thermoregulation. Clearly, older individuals are more likely to have disease than are younger individuals. It can be difficult to determine if the finding of altered thermoregulation reflects biological aging *per se* or reflects the outcome of a disease: the discussion herein relies on data from investigations in which a clear attempt has been made to control for disease.

Acclimation and the Sedentary Life Style

The ability to thermoregulate appropriately in hot and cold environments depends, in part, on one's acclimation to the ambient conditions. Well-acclimatized individuals will tolerate deviations from thermoneutrality better than will the nonacclimated. Moreover, a sedentary lifestyle in which individuals live at ambient temperatures close to thermoneutrality will reduce sweat gland density, capillary density of skin, and muscle mass, all outcomes that are independent of age and all of which impact thermoregulatory capacity.

Physiology of Thermoregulation

Hypothalamic Regulation of Body Temperature

Humans and rodents maintain a constant body temperature by balancing heat production with heat loss. The equilibrium point of heat production and loss, thermoneutrality, occurs at an ambient temperature of approximately 21 °C in the human and 23–25 °C in rats and mice, and does not appear to change with age. Exposure to environmental temperatures above or below thermoneutrality will result in various physiological responses that work to keep a stable core body temperature (38 °C humans, 38–39 °C mice and rats). Communication between neural centers within the hypothalamus and peripheral sensors and effector organs regulate, almost exclusively, the processes that maintain stable body temperature in homeotherms (Figure 1). The preoptic area of the hypothalamus contains heat- and cold-sensitive neurons that respond to changes in blood temperature, although heat-sensitive neurons far outnumber cold-sensitive sensors. Thus, detection of increased body temperature occurs, for the most part, within the hypothalamus. Detection of a cold environment occurs via temperature sensors within the skin; these sensors send afferent signals to the posterior hypothalamus at or near the level of the mammillary bodies (cold sensors in the skin outnumber heat sensors 10 to 1). The posterior hypothalamus also serves to coordinate the temperature-sensing 'data' from both the preoptic area and the periphery and, in turn, takes appropriate action through efferent signals.

Neural Regulation of Skin Blood Flow and Sweat Glands

Changes in skin blood flow, comprising the initial physiological response to increased or decreased ambient temperatures, are regulated by sympathetic vasodilation and vasoconstriction mechanisms. An increase in core temperature causes the release of vasoconstrictor tone, resulting in increased blood flow. If convective heat losses resulting from relaxation of the vasoconstrictor tone continue to be insufficient to cool the core, cholinergic active vasodilation causes further increase in skin blood flow that, in turn, increases convective heat loss. Sweat glands are innervated by cholinergic fibers that secrete

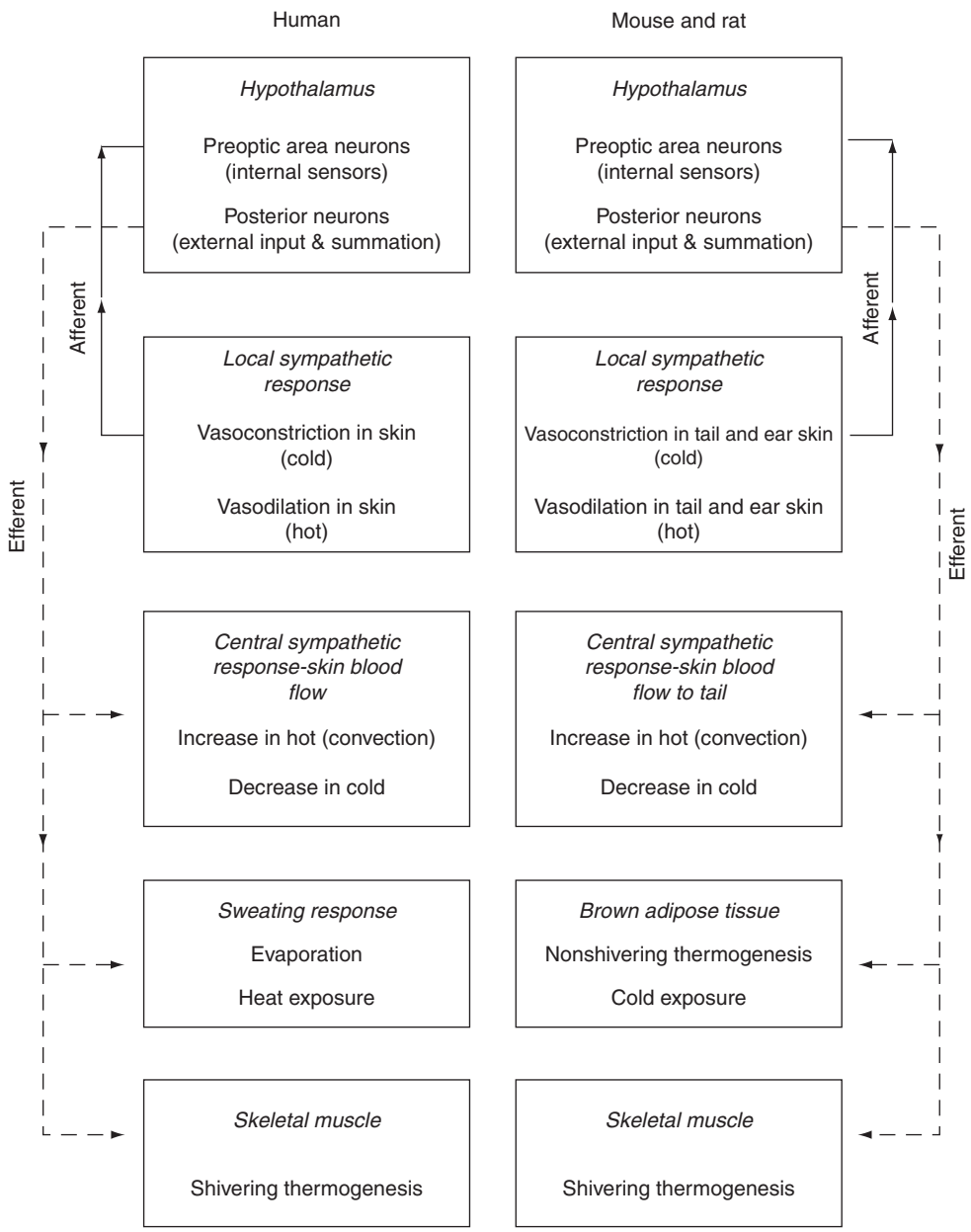

Figure 1 Neural pathways of thermal regulation. Sympathetic pathways are the sole neural tracks that influence thermoregulation; the parasympathetic system does not exert a significant effect on thermoregulation. Vasoconstriction and initiation of nonshivering thermogenesis occur in response to the release of norepinephrine attaching to α-adrenergic receptors in blood vessels and to α- and β-adrenergic receptors in brown adipose tissue. Vasodilation in the human is caused by epinephrine exciting the β-adrenergic receptor; acetylcholine causes vasodilation in rats (muscarinic receptors). Sweat glands are innervated by cholinergic nerve fibers that secrete acetylcholine, but the glands are also responsive to epinephrine secreted by the adrenal gland.

acetylcholine and are stimulated simultaneously with active vasodilation. Decreased skin and core temperatures result in activation of norepinephrine-mediated vasoconstriction beyond the normal vasoconstriction tone. Various other peptides, such as neuropeptide Y, can be co-released with norepinephrine and may also have a significant effect on vasoconstriction.

Effect of Senescence on the Response to Heat and Cold Exposure in Humans

Changes to Core Temperature

Mean core temperatures of humans exposed to cold or heat decrease (cold) or increase (hot) slightly more in senescent versus adult individuals. The differences in

core temperature are modest (no more than 0.1–0.2 °C difference) and are not generally considered physiologically important. Physical activity and acclimation to the ambient temperature improve the core temperature of older individuals, suggesting that behavior rather than aging *per se* may be of greater importance as the cause of senescent-related alterations in core temperature. The experimental evidence does not support the suggestion that the older population (defined here as those over 65 years of age) has a higher risk of developing clinical hypo- or hyperthermia, as compared to younger populations. However, recent investigations have found that senescent-related alterations in core temperature leading to thermal sickness are progressive in the aged cohort. That is, the most significant alteration in the loss or gain in body heat, as measured by core temperature, appears to be limited to the oldest old (i.e., individuals in their ninth and tenth decades of life).

Autonomic Response

The onset of vasoconstriction and vasodilation during thermal stress While changes in core temperature in response to heat and cold stress do not appear to be a general hallmark of senescence, there is agreement that the autonomic response may be altered. For example, the time to onset of sweating during passive heat exposure occurs significantly later in older versus younger individuals and is independent of the anatomical location (Figure 2). Once sweating has begun, however, the rate and amount appear to be similar in both young and old individuals. Moreover, there does not appear to be a significant difference in sweat glad density or skin temperature threshold that initiates sweating between healthy senescent and younger individuals. Together, lack of differences in sweat rate, sweat glad density, and thresholds suggests that the delay in sweating during passive heat stress reflects some alteration at the interface between efferent neural signaling and the effector organ.

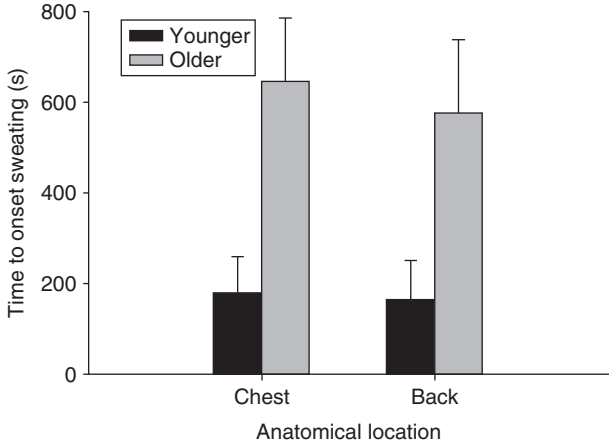

Figure 2 Time to onset of sweating in younger (ages 20–24 years; $n = 7$) and older (64–76 years; $n = 7$) men exposed to heat (35 °C) and a relative humidity of 45%. The lower leg of each person was also placed in a water bath (42 °C) during the passive heat exposure. From Inoue Y, Shibasaki M, Ueda H, et al. (1999) Mechanisms underlying the age-related decrement in the human sweating response. *European Journal of Applied Physiology* 79(2): 121–126, with permission.

Differences observed in the onset of vasoconstriction during cold exposure are consistent with senescent-related attenuation in the autonomic response within the efferent pathways during a thermal challenge. That is, the cold-induced decrease in skin blood flow occurs much later in the senescent individual as compared to a younger adult. However, an older individual's capacity to sense a drop in ambient temperature, as determined by subjective rating scales, does not differ from that of younger persons. The ability of the elderly to accurately sense a decrease in ambient temperature corresponds to the decrease in skin temperature. If the afferent signal to the hypothalamus was disrupted, these two measures – the sensing of cold and drop in skin temperature – should be discontinuous. Therefore, altered skin blood flow concomitant with appropriate thermal senescing suggests that the neural mechanisms that inform the thermoregulatory center of a change in ambient temperature – the afferent pathways – function normally. The delay in cold-induced vasoconstriction during senescence implies that the disruption in the autonomic response more likely reflects changes to efferent pathways. It remains unclear, however, if the senescent-related alteration in autonomic response to heat and cold stress occurs at the level of the hypothalamus or the peripheral effector organs.

Senescence and central-mediated autonomic response to thermal stress Evaluating the central-mediated (hypothalamic) autonomic response to thermal stress in humans is difficult, and most conclusions on a central-mediated role are based on circumstantial rather than direct experimental evidence. For example, the circadian rhythm of body temperature has been used as an indicator of hypothalamic function, as the suprachiasmatic nucleus regulates exclusively this 'biological clock.' Several investigations have found a senescent-related decrease in the amplitude of the circadian rhythm of body temperature and sleep, a finding consistent with alterations in hypothalamic regulation. However, senescent-related changes in the circadian rhythm of body temperature and sleep are almost always associated with an underlying disease, a sedentary life style, and/or psychological conditions that disrupt normal patterns of daily life. Investigations controlling for these confounding

variables have found that alterations in the circadian rhythm of body temperature and sleep are limited, for the most part, to the oldest individuals in the senescent group, a conclusion similar to that suggested for alterations in core temperature.

Senescence, skin blood flow, and vasodilation response to heat Several investigations have demonstrated that skin blood flow of elderly persons at rest and during heat stress does not increase to the same degree as that observed in younger individuals. Initially, senescent-related changes to the skin architecture, including diminished arteriole and capillary density, were suggested to be the primary cause of altered skin blood flow. Investigations comparing different anatomical sites for a senescent-related attenuation of skin blood flow have shown that the largest alterations occur at sites normally exposed to the environment. For example, heat-induced increase in forearm-skin blood flow occurs significantly later and is less robust than that observed from sites on the chest and back. These data imply that senescent-related alterations to skin blood flow caused by changes in the skin structure reflect more closely photoaging than they do the biological process of senescence.

Nonetheless, attenuated heat-induced skin blood flow of older individuals has also been reported at anatomical sites typically not exposed to the sun (i.e., chest and thigh). The question, then, becomes 'does the blunted heat-induced skin blood flow observed in the elderly reflect a decrease in vasodilation or the failure to release vasoconstriction tone?' Evaluation of decreased vasodilation versus sustained vasoconstriction tone can be achieved by comparing skin treated with bretylium tosylate, an agent that blocks vasoconstriction, to nontreated skin. These investigations have consistently shown no effect on skin blood flow by the blocking agents. That is, reduced heat-stress blood flow in the senescent individual appears to reflect altered active vasodilation.

The mechanisms underlying the altered vasodilation of the elderly during heat exposure remains to be elucidated. There is agreement that aged vessels are less responsive to neural transmitters, have fewer receptors, and less ligand binding to the receptor, but these effects have not been tested directly during heat exposure and also appear to be influenced significantly by underlying disease. Findings of a diminished level of neural density in senescent skin have been inconsistent. Increases, decreases, and no difference in neural density of aging skin have been reported. This inconsistency most likely arises from different anatomical sites sampled and the disease states of the individuals.

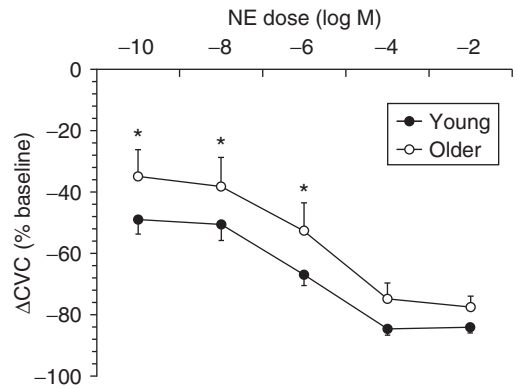

Figure 3 Mean cutaneous vasoconstriction (CVC) responses to sequential norepinephrine (NE) infusions in young ($n = 11$) and older ($n = 11$) individuals. Reproduced from Thompson CS, Holowatz LA, and Kenney WL (2005) Cutaneous vasoconstrictor responses to norepinephrine are attenuated in older humans. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 288(5): R1108–R1113, used with permission from the American Physiological Society.

Senescence, skin blood flow, and the vasoconstriction response to cold The decrease in skin blood flow during local and systemic cooling of healthy senescent individuals appears to be blunted compared to younger persons. Altered senescent-related skin blood flow during cold exposure reflects diminished vasoconstriction and appears to be related to both a decreased effect by norepinephrine (Figure 3) and a loss of contribution to vasoconstriction from co-released transmitters. Cold-induced vasoconstriction has been shown to be enhanced by neurotransmitters co-released with norepinephrine. When yohimbine and propranolol are used to antagonize norepinephrine's effects during local cold exposure, vasoconstriction decreases, but remains robust in the younger individuals, indicating contribution by co-released neurotransmitters. However, yohimbine and propranolol completely block vasoconstriction in the elderly. That is, co-released peptides, such as neural peptide Y (NPY), must be compensating for the diminished effect of norepinephrine in the younger vessel, yet there is no compensation in the older vessel. While some evidence suggests that the co-released neural peptide may be NPY, the precise identity of the agent remains to be elucidated.

Summary of the Effects of Senescence on the Response to Heat and Cold Exposure in Humans

Senescent-related alterations to core temperature during thermal stress are minimal. Reports of a greater risk of hypo- and hyperthermia in the elderly have now been judged to reflect more closely factors such as disease, physical activity level, psychological

impairment, and socioeconomic level, rather than biological aging *per se*. Sympathetic regulation of vasoconstriction and vasodilation appears to decline with age and reflects a general desensitivity to neural transmitters. Vasoconstriction and vasodilation can be improved by acclimation to the ambient temperature and increased physical activity, suggesting that behavior may also a have a significant role in the maintenance of sympathetic regulation. Recent investigations have suggested that if diminished thermoregulatory capacity occurs as the result of biological aging, the effects are progressive within the senescent cohort. That is, altered thermoregulation is most pronounced in people nearing the end of their natural life span (i.e., the ninth and tenth decades of life).

Effect of Senescence on the Response to Cold Exposure in Laboratory Rodents

The use of laboratory rodents, particularly rats and mice, to evaluate the affect of senescence on thermal stress has been, for the most part, limited to investigations describing changes to cold exposure. Young and old rodents do not tolerate heat exposure greater that 2–3 °C above normal body temperature (38–39 °C) and mortality rates increase significantly after 2 h. Thus, the discussion presented here is limited to the autonomic response observed in cold exposure of senescent mice and rats.

Response of Core Temperature to Cold Exposure

Observations of the effect of advanced age on cold-exposed core temperature in rats and mice have not been consistent. The differences in mean values for cold-exposed core temperature, when present, are modest, and the variations about the mean are considerably greater than those seen in younger animals. If an age-related decrease in cold-exposed temperature exists in animals, the decline appears to be more pronounced in males than in females (Figure 4). An inability to maintain core temperature during senescence does not correlate to a decline in the mean cold exposure oxygen consumption, a finding that indicates appropriate shivering response. Small and nonsignificant declines in muscle mass among 27-month-old versus 12- and 5-month-old rats are consistent with adequate shivering during senescence. Therefore, if senescence does cause altered cold-exposed core temperature in rodents, then the changes must be due to increased heat loss – that is, altered vasoconstriction, or alterations to brown adipose tissue (BAT) nonshivering thermogenesis (NST).

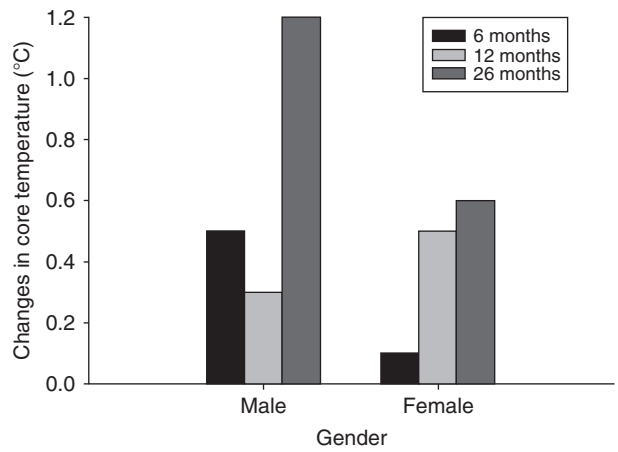

Figure 4 The change in core temperature in male and female Fischer 344 rats after 4h of exposure to 6 °C. Resting core temperature did not differ between 6-, 12-, and 26-month-old groups. Data adapted from Gabaldón AM, Gavel DA, Hamilton JS, et al. (2003) Norepinephrine release in brown adipose tissue remains robust in cold-exposed senescent Fischer 344 rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 285(1): R91–R98, with permission.

Autonomic Response to Cold Exposure

Nonshivering thermogenesis: A mechanism of heat production in rodents Brown adipose tissue is the principal anatomic location of NST. The contribution of NST to overall heat production during cold exposure in the rat has been estimated to be as high as 30–40% of the total body oxygen consumption. Nonshivering thermogenesis in BAT is initiated by release of norepinephrine from terminal endings of sympathetic neurons (Figure 5). Norepinephrine binds primarily to β_3-adrenergic receptor, and the transduction of this signal is mediated through a second-messenger pathway involving adenylyl cyclase. The oxidation of fatty acids, the primary fuel substrate for BAT, is uncoupled from the electron transfer system by a novel protein, uncoupling protein 1 (UCP1), so that heat, rather than ATP, is the metabolic end product.

Senescence and nonshivering thermogenesis BAT mass declines significantly with age. However, the decrease in BAT mass with chronological age most likely reflects long-term disuse of this tissue, because animals are typically housed at thermoneutral temperatures. This suggestion is consistent with the observation that cold exposure of old rats at 10 °C for five consecutive days significantly increases brown fat mass and BAT cell size.

Brown fat thermogenesis potential as measured by UCP1 concentration does not change with age. Therefore, any senescence-related attenuation of

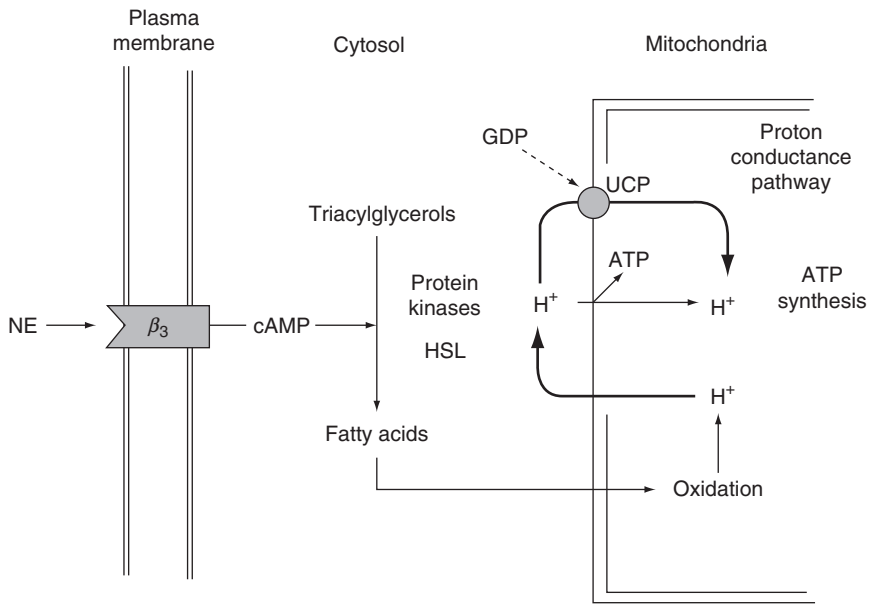

Figure 5 Schematic description of the influence of norepinephrine (NE) on the uncoupling of oxidative phosphorylation from electron transport that leads to nonshivering thermogenesis. β_3, β_3-adrenergic receptor; HSL, hormone-sensitive lipase; GDP, guanosine diphosphate; UCP, uncoupling protein.

NST more likely reflects alteration in sympathetic activity. This suggestion seems unlikely, however, as senescence has been associated with an increase rather than a decrease in sympathetic activity. Elevated plasma norepinephrine levels have been observed in several studies, and neural recordings of sympathetic signaling to brown fat are greater in 30- versus 5-month-old mice exposed to several hours of cold. Finally, brown fat tissue norepinephrine release measured during cold exposure remains robust in old rats (Figure 6). The preponderance of evidence indicates that neural signaling to aging brown fat is intact in the senescent animals.

Vascular response to cold exposure in rats Rodents rely on sympathetic-mediated changes in vessel diameter to regulate blood flow during exposure to cold. However, vasoconstriction is limited to the exposed skin of the tail. Vasodilation occurs in organs (such as skeletal muscle and BAT) that rely on an increase in local blood flow to enhance tissue oxygen demands of cold-exposed energy production. Indirect measurement of blood flow using radiolabeled microspheres has demonstrated that older rats are capable of increased vasodilation during cold exposure to the same degree as younger animals (Table 2). Vasoconstriction, as measured by heat loss index of the tail, also declines to the same degree in young and old rats. Vasoconstriction and vasodilation during cold exposure appear to respond appropriately in aged animals.

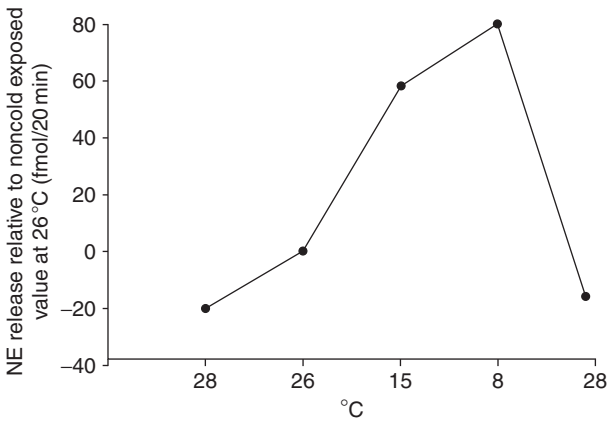

Figure 6 Net brown adipose tissue norepinephrine (NE) release in male 26-month-old rats during cold exposure. Values are expressed relative to the non-cold-exposed period of 26 °C. Data adapted from Gabaldón AM, Gavel DA, Hamilton JS, et al. (2003) Norepinephrine release in brown adipose tissue remains robust in cold-exposed senescent Fischer 344 rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 285(1): R91–R98, with permission.

Summary of Autonomic Reponses to Cold in Rodents

The decline in mean core temperature in old versus young rats is modest, is greater in males than in females, and supports the results observed in humans. Vasoconstriction and vasodilation respond appropriately during cold exposure, and brown fat thermogenesis does not appear to be altered. These results

Table 2 Tissue masses and resting (23 °C) and cold-exposed (6 °C) regional blood flow of brown adipose tissue and soleus muscles in rats

Rat tissue/age	*Mass (g)*	*Blood flow (ml min^{-1} g^{-2})*		
		Resting	*Cold (2 h)*	*Cold (4 h)*
Brown adipose tissue				
12 months old	0.44 ± 0.03	0.74 ± 0.19	5.41 ± 0.41	2.67 ± 0.24
26 months old	0.44 ± 0.04	0.63 ± 0.21	5.52 ± 0.64	4.46 ± 0.59
Soleus muscle				
12 months old	0.12 ± 0.01	0.63 ± 0.07	1.25 ± 0.12	1.09 ± 0.22
24 months old	0.13 ± 0.01	0.72 ± 0.18	1.09 ± 0.11	1.21 ± 0.21

Data adapted from McDonald RB, Hamilton JS, Stern JS, et al. (1989) Regional blood flow of exercise-trained younger and older cold-exposed rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 256(5): R1069–R1075, with permission.

suggest that autonomic control of thermoregulation in aged rats remains intact. It must be kept in mind that conclusions suggesting little or no change in thermoregulation of the aged rodent are based on the group means. Significant declines in core temperature and alteration in autonomic control of cold-induced thermoregulation can be identified in individual animals.

Response to Thermostress and the Rate of Aging

The evidence from studies in humans and rats suggests that senescence-related alterations in autonomic control of thermoregulation may be limited to individuals nearing the end of their natural life span. Moreover, investigations evaluating possible senescent-related alterations to heat and cold exposure in both humans and rodents have shown considerable inconsistency of results and high levels of variation about the mean for the outcome variable in old versus young groups. Investigations in both humans and animals report that core temperatures for some individuals do not change, whereas core temperatures for other individuals of identical age decrease (cold) or increase (hot) significantly. Together, these finding would suggest that distinct cohorts exhibiting different senescent-related outcomes may exist within the population of individuals considered old.

Cold-Exposed Core Temperature and Weight Loss

Our laboratory began investigating the possibility of multiple cohorts within the aged population when, by serendipity, we noticed that the old animals exhibiting decreasing core temperature during cold exposure were also the animals with the lowest body weights. Retrospective analysis in our longitudinal studies revealed that a drop in core temperature

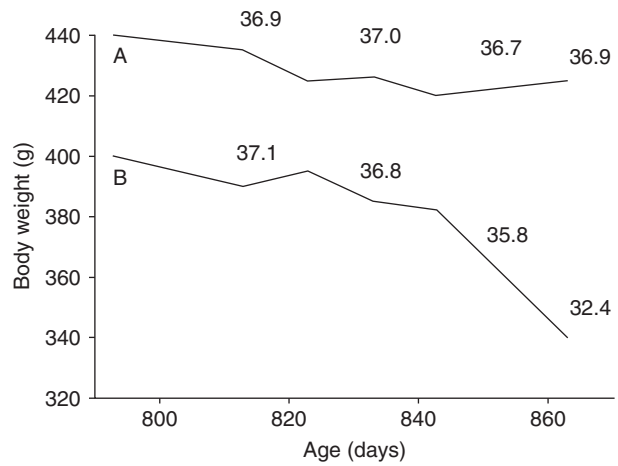

Figure 7 Relationship between body weight and intraperitoneal temperatures during cold exposure in two aged-matched male Fischer 344 rats. Line A is a weight-stable animal that lived to an age of 911 days. The rat represented by line B died at 870 days. Values above lines are the 4 h intraperitoneal temperatures (°C) of rats exposed to 6 °C. Data adapted from McDonald RB, Florez-Duquet M, Murtagh-Mark CA, et al. (1996) Relationship between cold-induced thermoregulation and spontaneous rapid body weight loss of aging F344 rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 271(5): R1115–R1122, with permission.

corresponded to weight loss and that the rats with altered core temperature died within days or a few weeks after the onset of weight loss. This suggested that alterations in autonomic regulation of body temperature decline only near the end of the life span.

We formally tested this hypothesis by measuring body weight daily and performing cold-exposure trials (6 °C for 4 h) on an every-other-week basis starting at 20 months of age (mean life span of Fischer 344 rats is 24–25 months). We found that cold-exposed core temperature in all old animals declined only after the onset of body weight loss and declining food intake (Figure 7). No decline in core temperature

in senescent versus presenescent rats occurred during the weight-stable phase. The decline in core temperature during the weight-loss phase was rapid, and the several animals became hypothermic, with body temperature below 33 °C, within 2 h of cold exposure. Most important, however, was our finding that the start of the terminal phase and decline of core temperature occurred in all animals tested and at various ages between 22 and 32 months. That is, loss in thermoregulatory capacity was not related to chronological age.

Hypothalamic function, thermoregulation, and the terminal phase The rapid and spontaneous loss in weight and food intake decline, combined with the dramatic alterations in cold-exposed thermoregulation, suggested a disruption in a central-mediated mechanism. Experiments were initiated to test the functionality and responsiveness of the hypothalamus, the primary center of thermoregulation. As a first step in evaluating hypothalamic function during the terminal phase, endogenous circadian rhythm of body temperature was followed longitudinally starting at 20 months of age. Continuous measurement of core temperature under constant darkness was performed by radiotransmitters implanted in the visceral cavity. Disruption of endogenous phase, rhythm amplitude, and period coincided with the start of body weight and food intake decline and occurred in all animals (Figure 8). Moreover, 24 h mean body temperature, a highly stable variable throughout the majority of the life span, abruptly declined at the start of weight loss and continued until death. These results indicated strongly that neural centers within the hypothalamus were disrupted during the terminal phase and that it was only during the rapid and spontaneous loss in weight that differences in autonomic function in thermoregulation were noted.

The rapid decline in core temperature leading to hypothermia in the senescent rats precludes experiments using the cold-exposure paradigm. We confirmed, however, that loss in mean body temperature coincided precisely with diminished food intake, a physiological system also highly regulated by the hypothalamus. Thus, we are able to use the decline in food intake as a surrogate model for temperature regulation and track possible alteration in hypothalamic function. To this end, several experiments have been completed evaluating the role of NPY, a potent stimulator of food intake, in hypothalamic regulation.

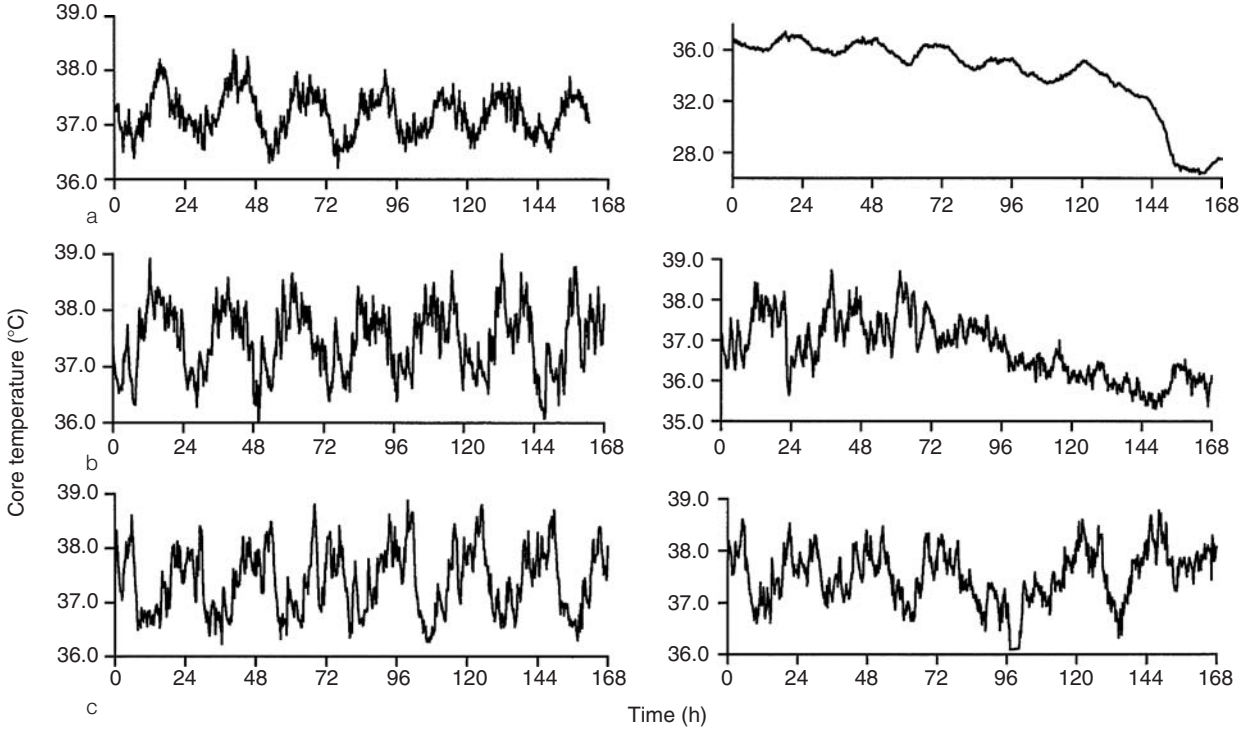

Figure 8 Seven days of core temperature data from three representative old male F344 rats (a–c) during analysis periods of presenescence (left) and senescence (right). Note that the temperature scale in the senescence period for rat a is different from that for rats b and c. Reproduced from McDonald RB, Hoban-Higgins TM, Ruhe RC, et al. (1999) Alterations in endogenous circadian rhythm of core temperature in senescent Fischer 344 rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 276(3): R824–R830, used with permission from the American Physiological Society.

The infusion of NPY directly into the paraventricular nucleus of the hypothalamus does not stimulate food intake in the senescent rat to the same degree as that observed during the weight-stable period. The lack of an effect of NPY on stimulating food intake in the senescent rat does not reflect declining receptor number, fewer NPY-sensitive neural cells, or alterations to pathways of peptides that are co-released with NPY – for example, γ-aminobutyric acid (GABA).

We have recently completed preliminary work describing gene expression within the hypothalamus between weight-stable and non-weight-stable old rats using microarrays. We find that differences in DNA expression are limited to only a few genes (3% of those tested, a finding constant with other microarray analyses in aging research) and that the genes with the greatest difference in expression were limited to specific types of cell functions. The majority of expression difference occurred in genes coding for ion channel proteins. A potassium channel had the highest degree of difference. The other genes were involved in signal transduction and nerve growth and maintenance factors. We did not find significant difference in any genes that coded for specific neural transmitters involved in thermoregulation or food intake. Our results indicate that the loss in function in the hypothalamus that may impact thermoregulation occurs at the level of cell processes common to all physiological systems and is not limited to any particular system.

Conclusion

The data from both humans and laboratory animals are consistent in that autonomic control of thermoregulation is well maintained through out most of the life span, including during a significant portion of old age. While there appear to be some alterations to autonomic mechanisms that can be directly related to biological aging, these changes tend to be modest and do not significantly affect core temperature. The current data suggest that many of the changes occurring to sympathetic-mediated vasoconstriction, vasodilation, and NST during old age can be prevented or reversed through acclimation and changes in life style.

Nonetheless, there does appear to be a period near the end of the life span in which alterations to autonomic regulation of body temperature declines. Evaluations of the vasoconstriction and vasodilation mechanisms in humans have found that heat conservation and convective heat loss do result in changes to core temperature in the oldest old. These observations in humans have been confirmed in rodents, in which losses in several autonomic mechanisms of thermoregulation coincide with a terminal phase of life characterized by declining body weight, food intake, and mean body temperature.

See also: Autonomic Disorders; Autonomic Failure; Autonomic Nervous System: Cardiovascular Control; Autonomic Nervous System: Metabolic Function; Autonomic Nervous System: Central Cardiovascular Control; Autonomic Nervous System: Central Thermoregulatory Control; Autonomic Nervous System; Energy Homeostasis: Thermoregulation; Neuroimmune System: Aging; Nutrient Sensing: Carbohydrates; Sympathetic Nervous System; Thermoregulation During Sleep and Sleep Deprivation.

Further Reading

Coppola JD, Horwitz BA, Hamilton JS, et al. (2005) Reduced feeding response to muscimol and neuropeptide Y in senescent F344 rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 288(6): R1492–R1498.

Florez-Duquet M and McDonald RB (1998) Cold-induced thermoregulation and biological aging. *Physiological Reviews* 78: 339–358.

Gabaldón AM, Gavel DA, Hamilton JS, et al. (2003) Norepinephrine release in brown adipose tissue remains robust in cold-exposed senescent Fischer 344 rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 285(1): R91–R98.

Horwitz BA, Gabaldón AM, and McDonald RB (2001) Thermoregulation during aging. In: Hoff PR and Mobbs CV (eds.) *Functional Neurobiology of Aging,* pp. 839–854. San Diego: Academic Press.

Inoue Y, Shibasaki M, Ueda H, et al. (1999) Mechanisms underlying the age-related decrement in the human sweating response. *European Journal of Applied Physiology* 79: 121–126.

Inoue Y, Kuwahara T, and Araki T (2005) Maturation- and aging-related changes in heat loss effector function. *Journal of Physiological Anthropology and Applied Human Science* 23: 289–294.

Kenney WL and Munce TA (2003) Invite review: Aging and human temperature regulation. *Journal of Applied Physiology* 95: 2598–2603.

Masoro EJ (1995) Aging: Current concepts. In: Masoro EJ (ed.) *Handbook of Physiology: Aging*, pp. 3–21. New York: Oxford University Press.

McDonald RB, Florez-Duquet M, Murtagh-Mark CA, et al. (1996) Relationship between cold-induced thermoregulation and spontaneous rapid body weight loss of aging F344 rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 271(5): R1115–R1122.

McDonald RB, Hoban-Higgins TM, Ruhe RC, et al. (1999) Alterations in endogenous circadian rhythm of core temperature in senescent Fischer 344 rats. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 276(3): R824–R830.

Pierzga JM, Frymoyer A, and Kenney WL (2003) Delayed distribution of active vasodilation and altered vascular conductance in aged skin. *Journal of Applied Physiology* 94: 1045–1053.

Thompson CS, Holowatz LA, and Kenney WL (2005) Cutaneous vasoconstrictor responses to norepinephrine are attenuated in older humans. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 288(5): R1108–R1113.

Thirst

E M Stricker, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

This article focuses on the seven signals that influence thirst in rats. Four signals are excitatory: increased plasma osmolality detected by cerebral osmoreceptors, decreased blood volume detected by cardiac stretch receptors, increased circulating levels of angiotensin II (Ang II) detected by Ang II receptors in the subfornical organ, and hypertonic fluid emptied from the stomach into the small intestine detected by putative osmo- or sodium receptors in the abdominal viscera. In addition, three signals inhibit thirst: decreased plasma osmolality detected by cerebral osmoreceptors, increased arterial blood pressure detected by arterial baroreceptors, and hypotonic fluid emptied into the small intestine detected by visceral osmo- or sodium receptors.

Drinking behavior stimulated by the sensation of thirst together with the complementary secretion of the antidiuretic hormone vasopressin (VP) from the posterior lobe of the pituitary gland comprise the major response components in the regulation of body fluid balance. The homeostasis of body fluid, in turn, is traditionally viewed as involving the regulation of its osmolality and of blood volume. However, the control of thirst is more complex than can be described by a two-factor model, and consideration of plasma sodium (Na^+) concentration (pNa) and of arterial blood pressure also must be included in an informed discussion. Each of these four variables (i.e., blood osmolality, volume, pressure, and Na^+ concentration) can be increased or decreased from normal levels, potentially providing eight distinct signals that may affect thirst. This article is organized around these four variables and focuses on the signals that actually do appear to influence water intake.

Fluid Osmolality and Na^+ Concentration

It has long been recognized that thirst is stimulated by increases in the extracellular concentration of nonpermeable osmolytes, such as Na^+, that cause osmotic movement of water from body cells. Thus, hyperosmotic solutions of glucose and urea do not increase water intake, but hypertonic solutions of NaCl or sorbitol do. This increase in effective plasma osmolality (pOsm) stimulates cerebral osmoreceptors located in the vascular organ of the lamina terminalis (OVLT), a structure in the basal forebrain that lacks a blood–brain barrier and therefore can respond readily to alterations in pOsm. In fact, increases in pOsm of only 1–2% stimulate thirst in animals, and further increases elicit proportional increases in water intake. Thirst is stimulated when projections from the OVLT activate neurons in the median preoptic area in the lamina terminalis ventral to the anterior commissure, although the subsequent neural circuitry that mediates thirst and drinking behavior has not been determined. Other projections from the OVLT to the paraventricular and supraoptic nuclei of the hypothalamus stimulate VP secretion. Surgical destruction of the OVLT eliminates both water drinking and neurohypophyseal VP secretion in response to increased pOsm.

Given this arrangement, osmoregulation was initially envisioned as a single-loop negative-feedback system in which dehydration increases pOsm and thereby provides a stimulus for thirst, whereas water intake serves to decrease pOsm back to normal and thereby causes satiation by removing the excitatory signal for thirst. However, the control system for osmoregulatory thirst cannot be that simple in organization because a substantial delay occurs between the time when water is ingested and the time when it appears in the circulation (from which it can affect pOsm and the cerebral osmoreceptors). Specifically, in both dogs and humans, it takes 10–20 min for ingested water to produce significant decreases in pOsm, whereas thirst and water intake are terminated much more rapidly (as is secretion of VP). Thus, an early stimulus must occur that signals the brain in anticipation of subsequent rehydration. Such an effect resembles feed-forward reflexes found in the control of numerous autonomic functions (e.g., insulin secretion in response to the taste of food before its digestion and assimilation).

This anticipatory element in the control of water intake was clarified by a series of elegant investigations reported by Ramsay and colleagues. Their research used dogs as experimental subjects, although later work extended their findings to human and nonhuman primates. Briefly, dehydrated dogs were observed to drink water rapidly but to stop after only a few minutes, well before dilution of blood plasma became evident. This early inhibition of thirst (and of VP secretion) occurred even when dogs were fitted with a gastric fistula, which drained ingested water from their stomachs and thereby prevented the possibility of rehydration. The same rapid effects occurred when dogs drank hypertonic NaCl solution (although, ultimately, when the saline was absorbed and pOsm was elevated, the dogs became even thirstier and secreted more VP than before, as might be expected). These observations highlight the importance of an early inhibitory signal in the control of water intake and

suggest its basis: a neural input to the brain from the oropharynx, associated with rapid swallowing during the act of drinking, which allowed dogs to meter their intake. This signal had a rapid but temporary inhibitory effect on thirst and VP secretion. When ingested water was subsequently absorbed and pOsm was diluted back to normal levels, a more sustained termination of thirst and VP secretion was produced due to rehydration. This hypothetical arrangement is consistent with the finding that a rapid inhibition of thirst and VP secretion was not observed in dehydrated dogs when a volume of water comparable to that consumed was intubated into their stomachs, bypassing their oropharynx.

Rats also utilize early feedback signals from drinking, although via a different mechanism. When rats were infused intravenously with hypertonic NaCl solution to stimulate VP secretion, 5 min of water intake caused a rapid decrease in plasma levels of VP without noticeable changes in pOsm. These effects were not associated with the act of drinking because intake of the same volume of isotonic saline had no effect on plasma VP levels. Thus, the animals appeared to be responding to the composition of the ingested fluid and not to its volume. In addition, previous findings had indicated that thirsty rats drank increased amounts of water when the pyloric sphincter was ligated (thereby preventing gastric emptying) and that gastric water loads in rats produced rapid decreases in plasma VP levels before a substantial reduction of pOsm occurred. These observations collectively suggest the existence of a visceral osmo- or Na^+ receptor that detects ingested water after the fluid leaves the stomach and before it enters the general circulation.

Fibers of the vagus nerve project from the abdominal viscera to the area postrema and adjacent nucleus of the solitary tract (AP/NTS) in the brain stem. Consistent with the possibility that these fibers serve to mediate the early effects of ingested water, considerable overdrinking occurred in thirsty rats when these sensory nerve fibers were destroyed by systemic injection of the neurotoxin capsaicin. Overdrinking was also observed when the projection sites were eliminated by lesions of the AP/NTS. In both cases, the animals acted as if they no longer received an early message that water had been consumed; that is, they kept drinking. It was only later, when the ingested water had been absorbed and diluted circulating plasma, that OVLT osmoreceptors were affected and thirst was reduced. Visceral osmo- or Na^+ receptors also have been implied by findings that thirst can be stimulated by gastric loads of hypertonic NaCl solution, before systemic increases in pOsm are detected. Similarly, thirst resulting from overnight water deprivation is considerably enhanced by a gastric load of hypertonic saline.

Rats with lesions of the AP/NTS have other features of disrupted osmoregulation. Besides drinking excessive amounts of water when thirsty, they also have a blunted secretion of VP in response to hypertonic saline infused intravenously and impaired urinary excretion of an administered NaCl load. Moreover, they drink hypertonic NaCl solution in unusually large bouts, as if they were not receiving early signals of the dehydrating solution from visceral Na^+ receptors. These results suggest an important role of the AP/NTS in osmoregulation in rats. Because the AP lacks a blood–brain barrier, it is possible that it directly detects some pertinent feature of body fluid composition in addition to receiving input from the viscera. Observations that rats with AP/NTS lesions secrete VP normally in response to hypertonic mannitol solution but not in response to equi-osmolar saline solution provide further information that this brain stem area plays a role in Na^+ regulation rather than in osmoregulation.

To summarize, it is well established that cerebral osmoreceptors mediate thirst and secretion of VP in response to increases in pOsm. However, other factors are also important because drinking and neurohypophyseal secretion can be influenced even when no changes in circulating pOsm are evident. For example, an 'anticipatory signal' from visceral osmo- or Na^+ receptors contributes importantly to these regulatory responses in rats. These receptors project to the AP/NTS, implicating a role for these brain stem structures in Na^+ regulation that is consistent with other findings in which the AP/NTS was surgically destroyed. The organization of the relevant neural circuits that control water intake and neurohypophyseal VP secretion remains to be clarified, as does the difference in their operation in response to NaCl loads and osmotic loads not containing Na^+, and their separate contributions to the control of NaCl appetite, thirst, and VP secretion.

Blood Volume

It has long been known that a sudden reduction in blood volume induced by hemorrhage can stimulate water intake and VP secretion in the absence of any changes in pOsm. However, hemorrhage is not an ideal treatment in laboratory studies of behavior in part because of the debilitation that could result from anemia or severe arterial hypotension. Instead, extravascular administration of colloidal solutions has proven to be an easy and effective treatment in rats to reduce plasma volume without affecting pOsm or

arterial blood pressure. Such solutions leach isosmotic plasma from the circulation into an extravascular depot by raising oncotic pressure in the local interstitial space and thereby disrupting the Starling equilibrium across the capillary membranes. Injection of colloidal solutions in rats produces a dose-related increase in water intake and VP secretion that is proportional to the induced plasma volume deficit.

These effects of hypovolemia occur in the absence of changes in arterial blood pressure. Stretch receptors located in the walls of the great veins that return blood to the heart and in the right atrium have been identified and presumed to mediate the increase in thirst and VP secretion after blood loss, and drinking in response to hypovolemia has been abolished by destroying the cardiac receptors. Moreover, water intake has been blunted in hypovolemic rats by inflating a small balloon in the venous–atrial junction to prevent detection of the loss in blood volume. However, neither thirst nor VP secretion was eliminated during hypovolemia by destruction of the NTS in rats, although the procedure did eliminate autonomic reflexes evoked by cardiac baroreceptors. These findings suggest that cardiac receptors are not essential for eliciting drinking in response to hypovolemia. The enzyme renin is secreted from the kidneys in response to hypovolemia, and it in turn causes the formation of angiotensin II (Ang II) in the blood. Because Ang II can stimulate drinking and VP secretion, it seems plausible that the renin–angiotensin system might be involved in signaling hypovolemia in rats with NTS lesions.

The role of the renin–angiotensin system in thirst during hypovolemia has been illuminated by other studies. Colloid treatment stimulates thirst in rats lacking kidneys. Similarly, elimination of Ang II-induced drinking by destroying the subfornical organ (SFO), the site of action of circulating Ang II, does not affect drinking in response to substantial plasma volume deficits. Thus, circulating Ang II clearly is not necessary for the drinking response to hypovolemia. However, when plasma volume deficits are relatively small, it is important to note that rats with SFO lesions do not drink as much water in response to hypovolemia as do control rats. Assuming that the main effect of SFO lesions in these experiments was to eliminate the stimulus of thirst provided by Ang II (without affecting any other signal of thirst), these findings suggest that even small amounts of circulating Ang II may contribute importantly to thirst in hypovolemic rats by lowering the threshold amount of depletion required for eliciting a drinking response. The fact that Ang II does not enhance water intake during substantial hypovolemia may be attributed to the potent inhibitory effect of osmotic dilution that results from the renal retention of ingested water. For example, a 2–3% dilution of pOsm will abolish water intake and VP secretion in rats with 35% plasma volume deficits.

Hypovolemic rats need both water and NaCl to replace their plasma volume deficits. Thus, when they drink only water, they still need NaCl to reexpand plasma volume. Therefore, it is appropriate that hypovolemia stimulates salt appetite in rats in addition to thirst. It is not a goal of this article to discuss the biological bases of salt appetite, but it is worth noting that osmotic dilution, which inhibits thirst during hypovolemia, also serves to disinhibit salt appetite. The NaCl intake that can then occur removes osmotic dilution and its inhibition of thirst, which is reestablished when rats drink water again. By alternating intake of water and NaCl solution, hypovolemic rats can effectively repair their plasma volume deficits while also regulating their pOsm.

Numerous investigations have examined the effects of concurrently presenting osmo- and volume regulatory signals that influence water intake. Increases in pOsm combined with decreases in plasma volume produced additive effects on water intake (and VP secretion) in rats. As mentioned previously, osmotic dilution completely inhibited water intake and VP secretion, even when hypovolemia was pronounced. In contrast, plasma volume expansion did not inhibit drinking in response to osmoregulatory thirst. Additional studies are needed to determine whether hypervolemia similarly has little effect on thirst induced by Ang II.

To summarize, there is no doubt that signals of thirst and VP secretion exist independent of osmoregulatory and Na^+ regulatory needs. One such signal is related to a decrease in blood volume and is presumed to be detected in part by cardiac baroreceptors. Vagal fibers project from the heart to the NTS, and from there they appear to be involved in a neural circuit that includes the median preoptic area in mediating thirst and the hypothalamic supraoptic and paraventricular nuclei in mediating VP secretion. However, the signal of hypovolemia is complex in that it also involves a blood-borne hormone, Ang II, and perhaps additional receptors that monitor blood volume and provide a separate signal that influences thirst and VP secretion. The potent effects of osmotic dilution in rats despite hypovolemia indicate that inhibitory signals participate in the control of thirst and VP secretion, not just variable degrees of excitation. These and other observations also reveal the hierarchies of homeostasis: The osmoregulatory controls of thirst and VP secretion appear to be much more potent than the volume regulatory controls.

Arterial Blood Pressure

Observations that vasodilating drugs can stimulate substantial water intake and VP secretion in rats in the absence of changes in pOsm or blood volume indicate that additional signals may provide excitatory stimuli for these responses. Arterial baroreceptors, located in the carotid sinus and aortic arch, are known to provide neural information to the brain stem that is the basis of numerous autonomic reflexes, and such afferent neural stimulation also may evoke thirst and VP secretion. Activation of the sympathoadrenal system by input from these arterial baroreceptors stimulates renin secretion from the kidneys, and blood-borne Ang II may provide the relevant signal to the brain instead of, or in addition to, a neural signal from arterial baroreceptors. In addition, a decrease in arterial blood pressure reduces renal blood flow and thereby provides a direct stimulation of renin secretion that is independent of the sympathoadrenal system.

The possibility that Ang II mediates thirst in rats during arterial hypotension was strongly supported by early observations that bilateral nephrectomy, which eliminated plasma renin, also abolished the drinking responses to hypotensive agents. However, two complications arose in interpreting those findings. First, nephrectomy removed circulating Ang II and its potent pressor effects in addition to its dipsogenic effects, and animals became so hypotensive that they could not readily behave; in other words, their failure to drink may reflect debilitation rather than an absence of thirst. However, later studies revealed that drinking induced by vasodilating agents could be eliminated by nephrectomy or by pharmacological blockade of the renin–angiotensin system, even when arterial blood pressure was modulated and the induced hypotension was not so severe as to preclude drinking behavior. These findings offer little support for the hypothesis that neural signals from arterial baroreceptors directly stimulate thirst during hypotension, independent of the renin–angiotensin system.

A second complication arose when plasma levels of Ang II, which were significantly elevated in rats during arterial hypotension, were found to be not nearly as high as those required to produce the same amount of drinking when Ang II was administered intravenously. One possible explanation was that Ang II was a dipsogen only in pathophysiologic or pharmacologic circumstances. An alternative view seems more likely to be correct: Ang II, when infused, actually provides a mixed stimulus of thirst. That is, Ang II possesses both an excitatory component that results when Ang II binds to receptors in the SFO and an inhibitory component that is associated with the induced elevation in arterial blood pressure. This inhibitory component would not be present during arterial hypotension when renin is secreted normally, hence the lower amounts of Ang II required to elicit thirst in that circumstance.

Consistent with the critical role of Ang II in thirst during arterial hypotension are observations that drinking was enhanced considerably when the hypertensive effects of Ang II were blunted by simultaneous administration of vasodilating agents. Additional support came from findings that water drinking in response to intravenous Ang II was greatly augmented in rats with NTS lesions, which could not receive the neural signal of hypertension from arterial baroreceptors. Sinoaortic denervation produced the same increase in the drinking response to intravenous Ang II, demonstrating that the arterial baroreceptors are involved in mediating this inhibitory stimulus.

These experiments indicate that changes in arterial blood pressure provide two signals that are relevant to the control of thirst. During arterial hypotension, increased activity in the renin–angiotensin system provides a blood-borne, excitatory stimulus of thirst. The second signal arises during arterial hypertension, and it inhibits thirst. The inhibitory effect of increased arterial blood pressure has been seen not only when rats are infused intravenously with Ang II but also when pressor agents are given to rats made thirsty by infusion of hypertonic saline. This inhibitory effect was linear when blood pressure increased from approximately 120 to approximately 160 mmHg, above which no further inhibition was seen. It is of interest that the same limits in detection of hypertension are seen in the familiar baroreceptor reflex control of heart rate. That inhibitory effect is known to be mediated by arterial baroreceptors and, apparently, so is the inhibition of thirst because it did not occur in rats after sinoaortic denervation.

Summary

Two systemic osmoregulatory signals (one excitatory and one inhibitory), two presystemic osmo- or Na^+ regulatory signals (one excitatory and one inhibitory), one excitatory volume regulatory signal, and two signals associated with changes in arterial blood pressure (one excitatory and one inhibitory) sum to a total of seven stimuli that have been identified in the control of thirst. Figure 1 summarizes these acute signals. The most well-established stimulus, increases in systemic pOsm, as produced experimentally by intravenous injection of hypertonic saline, provides an excitatory stimulus for thirst. This signal is generated by osmoreceptors in the OVLT, which apparently stimulates thirst by activating the median preoptic area. Gastric NaCl loads, whether administered by

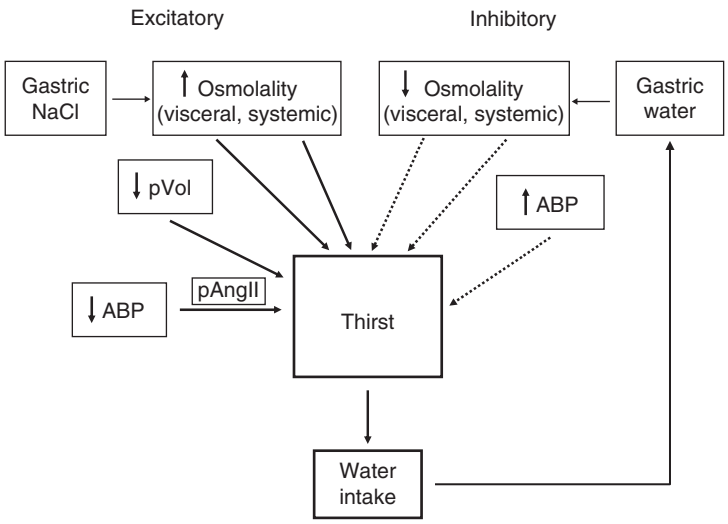

Figure 1 Schema showing the four excitatory signals (solid arrows) and the three inhibitory signals (dashed arrows) in the control of thirst in rats. ABP, arterial blood pressure; pAngII, plasma angiotensin II; pVol, plasma volume.

the investigator or consumed by the experimental subject, also increase pOsm and have the more rapid effect of stimulating thirst presumably via visceral osmo- or Na^+ receptors. Hypovolemia, as produced experimentally by subcutaneous injection of colloid in rats, lowers plasma volume and provides another mixed signal for thirst, in part neural (presumably from cardiac baroreceptors through the NTS) and in part humoral (Ang II acting on receptors in the SFO). Decreases in arterial blood pressure, as produced by vasodilating agents, stimulate renin secretion (via baroreceptors in both the carotid sinus–aortic arch and the kidneys), which stimulates thirst when Ang II acts on receptors in the SFO. There is no compelling evidence to indicate that arterial baroreceptors act directly to stimulate thirst when the renin–angiotensin system is blocked.

Acute increases in arterial blood pressure such as might arise from intravenous administration of the α-adrenergic receptor agonist phenylephrine inhibit water intake regardless of whether thirst is stimulated by hyperosmolality, hypovolemia, or Ang II. This inhibitory effect appears to be mediated by arterial baroreceptors. Increases in plasma levels of Ang II that result not from endogenous renin secretion but from intravenous administration of renin or Ang II provide a stimulus of thirst that has both excitatory and inhibitory components. Excitation results from activation of Ang II receptors in the SFO, whereas inhibition appears to be mediated by arterial baroreceptors. In consequence of this mixed input, the levels of plasma Ang II that are required to stimulate thirst are much higher than those required to stimulate thirst when rats are hypovolemic or hypotensive.

Decreases in pOsm induced by gastric water loads provide another inhibitory stimulus of thirst, and they do so in two ways. First, the water depresses the activity of putative visceral osmo- or Na^+ receptors even before water has diluted circulating pOsm (in rats; in dogs, a similar early signal is generated by the act of drinking). Second, the subsequent absorption of water into the general circulation dilutes systemic pOsm, which presumably decreases the activity of cerebral osmoreceptors.

For purposes of simplicity, Figure 1 and this article generally did not include consideration of the close link between thirst and salt appetite, the close link between the controls of salt appetite and food intake, or the close link between the controls of VP secretion and thirst. Nor has there been consideration of the important contribution of neurohypophyseal oxytocin secretion to osmoregulation (by stimulating natriuresis) and to the maintenance of arterial blood pressure (by stimulating renin secretion). Nor was it mentioned that most instances of water intake by free-feeding humans occur in the absence of body fluid deficits, and that we generally rely on our kidneys to eliminate excesses in body fluids such as occur when water is ingested with meals or in caffeinated or alcoholic beverages. Nonetheless, it should be apparent that research focused on the biological bases of thirst and satiety for water has enabled great advances in our understanding and this remains a lively field of investigation today.

See also: Angiotensin II; Atrial Natriuretic Peptide: Fluid/Mineral Balance; Blood Pressure: Baroreceptors; Circumventricular Organs; Fluid and Electrolyte

Homeostasis: Clinical Disease; Food and Water Intake: Regulation; Neurohypophyseal System; Osmoregulation; Salt Appetite.

Further Reading

Curtis KS, Huang W, Sved AF, Verbalis JG, and Stricker EM (1999) Impaired osmoregulatory responses in rats with area postrema lesions. *American Journal of Physiology* 277: R209–R219.

Fitzsimons JT (1969) The role of a renal thirst factor in drinking induced by extracellular stimuli. *Journal of Physiology, London* 201: 349–368.

Fitzsimons JT (1998) Angiotensin, thirst, and sodium appetite. *Physiological Reviews* 78: 583–686.

Gilman A (1937) The relation between blood osmotic pressure, fluid distribution and voluntary water intake. *American Journal of Physiology* 120: 323–328.

Robinson MM and Evered MD (1987) Pressor action of intravenous angiotensin II reduces drinking response in rats. *American Journal of Physiology* 252: R754–R759.

Simpson JB, Epstein AN, and Camardo JS (1978) Localization of receptors for the dipsogenic action of angiotensin II in the subfornical organ of rat. *Journal of Comparative Physiology and Psychology* 92: 581–608.

Stricker EM (1968) Some physiological and motivational properties of the hypovolemic stimulus for thirst. *Physiology & Behavior* 3: 379–385.

Stricker EM (2004) Thirst. *Handbook of Behavioral Neurobiology* 14: 505–543.

Stricker EM and Verbalis JG (2003) Water intake and body fluids. In: Squire LR, Bloom FE, McConnell SK, Roberts JL, Spitzer NC, and Zigmond MJ (eds.) *Fundamental Neuroscience*, 2nd edn., pp. 1011–1029. San Diego: Academic Press.

Thrasher TN, Nistal-Herrera JF, Keil LC, and Ramsay DJ (1981) Satiety and inhibition of vasopressin secretion after drinking in dogs. *American Journal of Physiology* 240: E394–E401.

Thyroid Hormone and Transcriptional Regulation in the CNS

D Forrest and J Nunez, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

Two major iodothyronines, thyroxine (3,5,3′,5′-tetraiodothyronine or T4) and 3,5,3′-triiodothyronine (T3) are synthesized in the thyroid gland (Figure 1). T3, the primary active form of thyroid hormone (TH), is also produced by local conversion of T4 to T3 by deiodination in some peripheral organs. The major actions of T3 are mediated by a nuclear TH receptor (TR) that acts as ligand-regulated transcription factor to control the expression of target genes. TH has cell type-specific effects in almost all tissues.

In human development, TH deficiency, during critical periods of neuronal differentiation, produces permanent and severe mental retardation (cretinism) if hormone replacement therapy is not performed as soon as possible after birth. Iodine is required for TH synthesis, and low iodine intake (endemic hypothyroidism) remains the major cause of cretinism in several developing countries. In addition, sporadic congenital hypothyroidism may arise in about 1:3500 newborns and is usually caused by defects in the developmental formation of the thyroid gland or in TH biosynthesis. Previously, this was the major cause of cretinism in developed countries, but screening is now performed on all newborn infants, followed by hormone replacement therapy as necessary. Timely hormone replacement reduces the occurrence of severe mental deficiencies. Neurological defects are also associated with mutations in one of the receptors, TRβ, in the syndrome of TH resistance, and in a TH transporter, MCT8, in a form of X-linked mental retardation.

Abnormalities in TH levels in adults are associated with different consequences than those arising during development. In adulthood, hypothyroidism is associated with depression and is recognized as a secondary cause of dementia, whereas hyperthyroidism may be associated with mania and anxiety disorders.

Some of the underlying actions of TH in the CNS were first suggested by Eayrs and Legrand between the 1950s and 1970s in pioneering studies on model species. These studies, mainly on the rat, indicated that several steps in brain maturation including cell migration, outgrowth of neuronal processes, acquisition of neuronal polarity, synaptogenesis, and myelin formation were delayed under hypothyroid conditions, whereas glial cell proliferation (gliosis) and neuronal cell death were increased. As a result, synaptic activity is reduced. In contrast, excessive TH can accelerate development prematurely with deleterious consequences on the CNS. In addition to general functions in neuronal and glial differentiation, TH has specific actions in different brain regions and sensory systems.

TH Receptors

The major actions of TH ultimately concern the regulation of target genes (Figure 2). The mechanism of action involves: (1) uptake from the blood of both T_4 and T_3 by target tissues; (2) conversion of T_4, either to T_3, the active hormone, or to reverse T_3 (3,3′,5′-triiodothyronine or rT_3), an inactive derivative, by deiodinases; (3) binding of T_3 to the TR; and (4) activation or inhibition of transcription of specific genes.

In 1986, the groups of Evans and Vennström reported independently the cloning of cellular homologs (c-erbA) of the viral erbA oncogene and showed that they encoded high-affinity nuclear T3 receptors (Figure 3). A small family of TRα and TRβ receptors is encoded by the *THRA* and *THRB* genes, respectively (human nomenclature). In mice and model species, the genes are named *Thra* and *Thrb*, respectively (or, as members of the wider superfamily of nuclear receptor genes, *Nr1a1* and *Nr1a2*). Related *Thra* and *Thrb* genes exist in mammalian, avian, amphibian, and fish species, suggesting that both genes are fundamental to T3 signaling. *Thrb* generates two major receptor isoforms, TRβ1 and TRβ2, by alternative promoter usage. *Thra* encodes a single T3 receptor, TRα1. In some species, additional variant TR products are produced. Mammalian but not nonmammalian species express a TRα2 splice variant that does not bind T3 and which may act as a weak antagonist of normal TR function. A TRβ3 isoform with a variant N-terminus occurs in rat. The major TRα1, TRβ1, and TRβ2 isoforms contain a DNA-binding domain, a C-terminal ligand-binding domain, and a variable N-terminal domain.

Several properties of TRs are of importance:

1. TRs bind to TH-responsive elements (T3RE) in responsive target genes. A typical T3RE contains two half sites that bind two TR molecules. Each half site has a consensus sequence, AGGTCA, that may be arranged as a direct repeat, inverted palindrome, or palindrome with variations in the

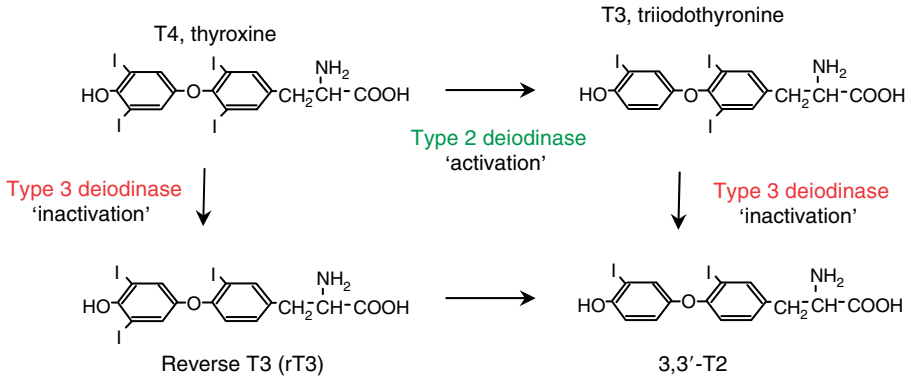

Figure 1 Iodothyronines produced by the thyroid gland and by metabolic deiodination in peripheral tissues. Type 2 deiodinase converts T4 into the major active ligand, T3. Type 3 deiodinase mediates inner-ring deiodination of T4 or T3 to form the largely inactive metabolites rT3 and 3,3′-T2, respectively.

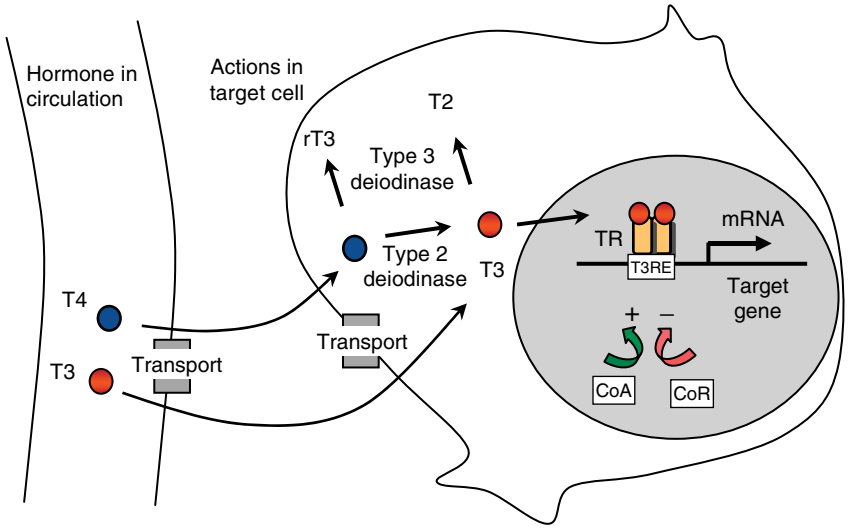

Figure 2 Steps in the nuclear action of TH. T4 and T3 produced by the thyroid gland are transported by the blood to target tissues. Following cellular uptake in the CNS, T4 is converted to T3 or to rT3 by type 2 or type 3 deiodinases, respectively. T3 binds its receptor (TR), triggering transcriptional regulation of target genes and the production of specific mRNAs and proteins. TR activity is regulated by T3 ligand and by transcriptional corepressors (CoR's) or coactivators (CoA's). In a target tissue, T3 levels may be amplified by type 2 deiodinase in the target cell itself or, in more distant cells, through a paracrine-like mechanism of control. Transporters for uptake and release of hormone may be important in local hormonal communication between cells.

number, spacing, and orientation depending on the gene. Negative and positive response T3REs have been identified.

2. TRs can bind to the T3RE as monomers or more stably as homodimers (TR/TR) or as heterodimers with retinoid X receptors (TR/RXR).
3. TRs are bound to chromatin both in the absence or presence of TH. Some T_3-responsive genes can be repressed by the hormone-free TR.

In the absence of T3, the TR may bind a target gene together with corepressors to form a multi-subunit repressor complex. The binding of ligand displaces corepressors and allows coactivators to bind the complex. These events are associated with histone-modifying activities that promote transcription induction. A number of cofactors for TR have been identified and some such as hairless and alien may have specific roles in brain.

Receptor Functions in the CNS

In vitro, the major TRα1, TRβ1, and TRβ2 isoforms have similar transactivation potential on most reporter genes, suggestive of a high degree of functional overlap. However, each gene displays a physiological specificity *in vivo* which is determined to a large

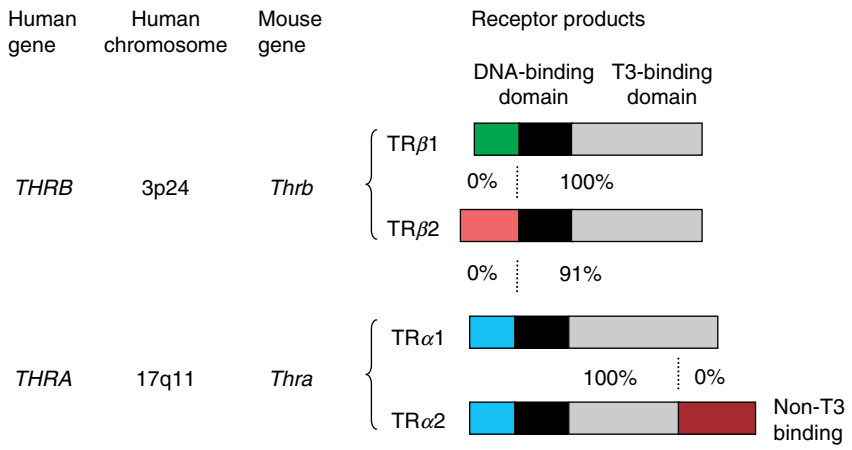

Figure 3 TRs in mammals. TRα1, TRβ1, and TRβ2 are T3 receptors found widely in mammalian and nonmammalian vertebrate species. TRα2 only occurs in mammals and because of a C-terminal splice variation, it does not bind T3, nor transactivate; its physiological role is unclear.

extent by its particular expression pattern. *In vivo*, the different isoforms may have interchangeable, additive, or cell type- and gene-specific functions.

TRα1 is near-ubiquitously present in the brain from early stages of embryogenesis. TRα1 is also the major TR isoform in embryonic stem cells and it can mediate a neuronal-like differentiation of these cells in response to T3. In contrast, TRβ1 is induced in the later embryonic and neonatal period in many brain regions. TRβ2 is the most restricted isoform, being expressed in retinal photoreceptors, the anterior pituitary gland, and paraventricular nucleus of the hypothalamus which releases TRH (thyrotropin-releasing hormone (TRH)). Both TRβ1 and TRβ2 are also expressed in the cochlea.

Mutations in human *THRB* cause an autosomal dominant syndrome of resistance to TH with a variety of symptoms that in many cases include mental retardation, other neurological deficits, or a degree of hearing loss. The *THRA* gene has so far not been associated with inherited disease.

Targeted mutations in mouse *Thrb* result in seizure susceptibility, behavioral abnormalities, and sensory deficits in hearing and color vision. Mutations in mouse *Thra* cause anxiety, behavioral abnormalities, defective oligodendrocyte and neuronal differentiation, and impaired glucose uptake in brain regions (Table 1).

Table 1 Neurological phenotypes associated with mutations in the TH receptor genes *Thra* and *Thrb* in mice

Phenotype	*Mutant gene*	
	Thra	*Thrb*
Deafness	−	+
Color blindness	−	+
Audiogenic seizure susceptibility	−	+
Dysfunctional hypothalamic–pituitary–thyroid axis	−	+
Mating behavior defects	+	+
Learning deficiency	+	+
Cerebellar defects	+	+
Behavior – anxiety	+	−
Defective oligodendrocyte differentiation	+	−
Reduced glucose utilization in brain	+	−

−, no phenotype; +, phenotype.
The authors acknowledge several investigators for the information in Table 1, including J Bernakl, S-Y Cheng, G Morreale, L Ng, D Pfaff, M Raff, T Reh, A Rüsch, J Samarut, B Vennsatröm, and F Wondisford.

TH Availability in the Brain

The amount of T3 available to the brain depends on the iodine present in the diet, the activity of the thyroid gland, the uptake of T4 and T3 by target tissues, and the activation or inactivation of hormone by deiodinases. Each of these events may vary during development. *In utero*, some TH is transferred from the mother to the fetus.

TH synthesis and the iodine supply. The iodine supply determines the ratio of T3 and T4 synthesis in the thyroid gland and the amount of hormone generated during each developmental stage. Both T4 and T3 are produced in a two-step process catalyzed by thyroid peroxidase. Several tyrosine residues of a large glycoprotein, thyroglobulin, are iodinated yielding mono-(MIT) and diiodotyrosine (DIT) residues. Some of these residues are coupled to form either T4 or T3 residues. The relative amount of T4 and T3 depends on the DIT/MIT ratio that increases with the level of thyroglobulin iodination. Thus, depending on the iodine supply, varying amounts of

T4 and T3 bound to thyroglobulin are stored in the thyroid follicle (colloid). T3 is present at only 10–25% of the level of T4 within thyroglobulin in a euthyroid status but when the iodide intake is low, the thyroid gland produces more T3 than T4.

Thyroid-stimulating hormone and feedback control of TH secretion. The amount of TH present in the fetal blood varies during development. TH transferred from the mother to the fetus before the onset of the fetal thyroid function is probably important during early stages of brain development. In the human fetus, the thyroid gland is functional at mid-gestation when it begins to produce small amounts of T4 and T3. T4 and T3 levels markedly increase in the blood at birth. The fetus may be protected against TH excess by type 3 deiodinase, a TH-inactivating enzyme (see Figure 1) expressed in the placenta. The release into the blood of T3 and T4 is controlled by a cascade of stimulatory mechanisms. The hypothalamus produces the TRH that stimulates the pituitary to release thyrotropin (thyroid-stimulating hormone, TSH); TSH then stimulates the activity of the thyroid gland. The secretion of both TRH and TSH is inhibited by a classical negative feedback loop. TRH and TSH are suppressed mainly via TRβ, when the level of TH in the blood is high. Both in the hypothalamus and pituitary, the negative feedback also depends, at least partly, on the local conversion of T4 to T3 by type 2 deiodinase.

In adulthood, the rat thyroid gland produces maximally ~3.0 mole of T4 and ~0.3 mole of T3 per mole of thyroglobulin when iodine is not limiting. Yet, the concentration of T3 in the serum is ~50 times lower than that of T4 and depends probably not only on its rate of secretion from the thyroid gland but also on its clearance rate and its uptake by the target cells. The half-life of T3 is shorter than that of T4 (7–9 days for T4 and 8–9 h for T3, depending on age).

T4 to T3 conversion by deiodination. T3, the active ligand, is produced locally in some target tissues by outer-ring deiodination of T4 by type 2 deiodinase (Figures 1 and 2). In contrast, type 3 deiodinase inactivates the hormone by inner-ring deiodination of T4 or T3 to produce rT3 and 3,3′-diiodothyronine (3,3′-T2), respectively. A third enzyme, type 1 deiodinase, may be more important in homeostatic control of hormone levels through metabolism in liver and kidney than in the CNS. Type 2 and 3 deiodinases are differentially expressed in the nervous system and are thought to regulate ligand availability in local environments. For example, seasonal reproduction in Japanese quail is controlled by changing day length, which alters the periodic expression of type 2 deiodinase in the mediobasal hypothalamus. In mice, the deletion of type 2 deiodinase produces deafness, suggesting the importance of local production of T3 in the inner ear. Otherwise, these mice have a relatively mild phenotype compared to that caused by TH deprivation, suggesting that the T3 produced by the thyroid gland may be adequate for TH action in other tissues.

Hormonal uptake and communication between cells. The importance of regulated uptake of TH by CNS target tissues is indicated by reports of mutations in the monocarboxylate transporter *MCT8*, a TH transporter *in vitro*, in cases of human X-linked mental retardation. *MCT8* and other transporter genes may regulate uptake of T4 and T3 from serum and also hormonal transfer between cells. The generation of T3 by type 2 deiodinase can occur not only in the target cell but also in some cases in nearby cells, implying that transporters are necessary for ligand transfer between cells in a local environment. Such a paracrine-like mechanism has been proposed in brain regions and in the cochlea.

Actions in the Nervous System

Glucose Utilization and Synaptic Activity

Glucose is the sole nutrient used by the brain to produce the energy required for synaptic activity. Because a close correlation exists between local neural activity and glucose metabolism, the severe reduction in glucose utilization seen in the hypothyroid brain might explain the decrease in synaptic activity. Studies in mutant mice showed that the regulation of brain glucose utilization depends primarily upon TRα1 rather than TRβ (Figure 4). Experiments performed by stimulating the whisker-to-barrel pathway in TRα mutant mice show that glucose utilization is decreased by ~50% in all four stations of this pathway, suggesting that synaptic activity is reduced to the same extent. These changes in glucose utilization might provide an explanation at the molecular level of the well-known increase in oxygen consumption and basal metabolic rate induced by TH in many tissues including the developing brain.

Genes regulated by TH include those encoding proteins modulating metabolic, synaptic, or cell division activities such as srg-1, RC3/neurogranin, calmodulin-dependent protein kinase, glutamine synthase, protein kinase C, prostaglandin D2 synthase, calbindin, and cyclin-dependent kinase. Other types of target genes include hormones (growth hormone, TRH, TSH) and transcription factors. Some responsive genes may be regulated directly at the transcription level, while others may be regulated indirectly or at a post-transcriptional level. There have also been proposals

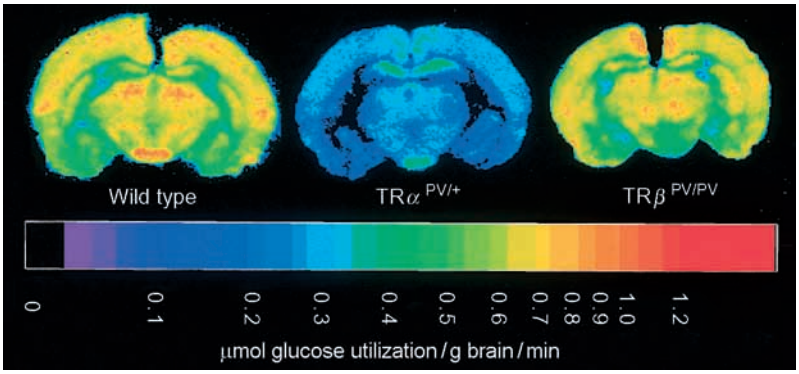

Figure 4 Glucose utilization in the brain of mice carrying TRα or TRβ mutations. Quantitative color-coded autoradiographs of brain sections from wild-type or TRα- and TRβ-mutant mice. Local rates of glucose utilization established by the deoxyglucose method are encoded in the calibrated color scale at the bottom of the figure. Redrawn from Itoh Y, Esaki T, Kaneshique M, et al. (2001) Brain glucose utilization in mice with a targeted mutation in the thyroid hormone alpha or beta receptor gene. *Proceedings of the National Academy of Sciences of the United States of America* 98: 9913–9918.

of nongenomic actions of TH, although the relevance to the brain and the mechanisms are currently unclear.

Oligodendrocyte Differentiation and Myelin Synthesis

TRs are expressed in oligodendrocytes, the glial cell type responsible for formation of the myelin sheet. Oligodendrocyte differentiation as well as the expression of several myelin proteins (proteolipid, myelin basic protein, MAG) are under TH control. These findings might explain the actions of TH on myelination, which would probably contribute to the changes in electrical activity seen in the hypothyroid brain.

TH and Axonal and Dendritic Outgrowth during Morphological Neuronal Differentiation

TH actions on the formation of axons and dendrites is important for neuronal connectivity. Synaptic activity is severely impaired in neonatal hypothyroidism. In the rat cerebellum, for instance, the output of the granule cells from the proliferative zone, the external granular layer, is slowed, and the length of the parallel fiber, the axons of the granule cells, is shorter. Simultaneously, the growth and arborization of the dendritic tree of the Purkinje cells is reduced. Similar effects are noted in other brain regions but at different developmental stages, for instance, 2 weeks after birth in the cerebellum and hippocampus but sooner after birth in the cortex.

TH also stimulates neurite outgrowth in primary neuronal cell cultures. Both axonal and dendrite outgrowth require microtubule assembly and stabilization. Microtubules and associated motor proteins are also required for the transport of vesicles containing neurotransmitters from the cell body, where they

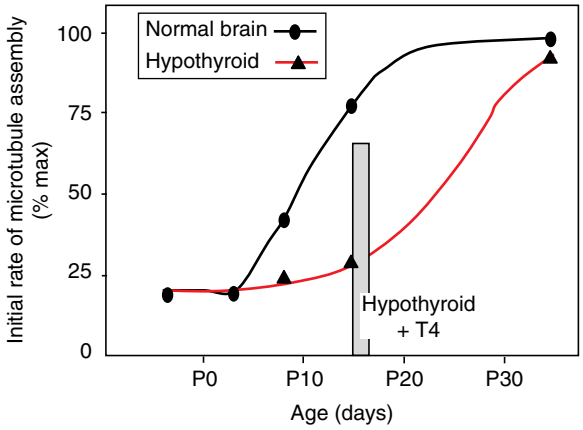

Figure 5 Comparative rates of microtubule assembly during postnatal brain development in normal (circles) and hypothyroid (triangles) rats. Rates of assembly were measured in brain supernatants containing an equivalent tubulin concentration of 1 mg ml^{-1}. The column shows that injection of TH at birth restores near-normal assembly rates 14 days later. Results are expressed as percentages of maximal values measured in euthyroid rat brain supernatants. Redrawn from Fellous A, Lennon AM, Francon J, and Nunez J (1979) Thyroid harmones and neurotubule assembly *in vitro* during brain development. *European Journal of Biochemistry* 101: 365–375.

are synthesized, to the presynaptic nerve terminals. The rate of microtubule assembly is decreased by TH deprivation during development and restored by TH administration (Figure 5). In contrast, hypothyroidism, when established in adulthood, has no effect on the rate of microtubule assembly. TH also regulates the expression of brain-specific tubulins and of various microtubule-associated proteins (MAPs) that stimulate microtubule assembly. Two major MAPs, tau and MAP2, are axon and dendrite specific, respectively. Juvenile tau and MAP2 splice variants are

produced at early developmental stages and have a much lower assembly-stimulating activity than the adult species. The replacement of immature tau variants by mature forms is regulated by TH at the level of alternative splicing. The expression of MAP2 is also regulated by TH.

TH also regulates axonal outgrowth indirectly by stimulating the expression of neurotrophic factors, such as nerve growth factor (NGF), and their receptors. In turn, NGF stimulates neurite outgrowth and also, as shown in PC12 cell cultures, the expression of tau proteins.

The Auditory System

Hearing is one of the most sensitive functions controlled by TH. Hypothyroidism, in early human development *in utero* or in the neonatal period, presents a risk for deafness. Hearing loss may also arise in 20% of cases of resistance to TH. Evidence in rodent models indicates that T3 acts at several levels in the auditory system from the cochlea and auditory nerve to the auditory centers in the brain stem and cortex. Studies of hypothyroid rodent models indicate that the hormone is required during the first two postnatal weeks, a period that encompasses terminal differentiation in the cochlea and the onset and maturation of auditory function. This period corresponds to a somewhat earlier prenatal/neonatal window in humans and points to the role of TH as a stimulus for the correct onset of auditory function.

TRs are essential for cochlear development with TRβ having the primary role and TRα1 a lesser contribution. TRβ is expressed in the hair cells and supporting cells of the sensory epithelium, the cells that secrete the tectorial membrane, the stria vascularis, spiral ganglion, and other areas. In hypothyroid or TR-deficient mice, the differentiation or function of many of these areas is retarded. There is delayed onset of a potassium current that is thought to allow inner hair cells to mediate high-frequency transduction, and the endocochlear potential is reduced.

The synchronization of these complex events required for the onset of cochlear function is unlikely to be mediated directly by circulating T4 and T3 but rather by type 2 deiodinase which is induced in the cochlea in the postnatal period prior to the onset of hearing in mice. Type 2 deiodinase can convert serum T4 into T3, thereby amplifying local levels of active T3 ligand that trigger TR activity in the cochlea at the appropriate time. The partition of the cochlea into compartments suggests that transporters are important in the uptake of T4 and release of T3 to the responsive cells within the cochlea.

The Color Visual System

The differentiation of cone photoreceptors, the sensory cells that mediate color vision, has only recently been recognized as being dependent upon a TR, based on studies in mice. There had been relatively little previous study of the role of TH in the retina, and the possibility of color blindness has been largely overlooked in human congenital thyroid disorders where the major concern understandably has been with mental retardation. In the mouse retina, TRβ2 is expressed in newly generated cones where it controls the expression of the opsin photopigments that mediate color perception. This is a highly cell-specific role for a TR isoform.

Color vision in most mammals is based upon two opsin pigments that allow the discriminatory detection of short (S, blue) or medium-long (M, green) wavelengths of light. The differential distribution of M and S opsins in cones across the retina is thought to be a prerequisite for color perception. TRβ2-deficient mice have defective medium-long wave responses due to a lack of M opsin (**Figure 6**). All cones instead express S opsin. TRβ2 therefore is necessary for M opsin activation and for restricting the spatial distribution of S opsin in cones across the retina. These studies suggest that a TR is unexpectedly critical for programming the diversification of the cone subtypes that mediate color vision.

The type 3 deiodinase is involved in retinal development. This hormone-inactivating enzyme regulates cell division in the dorsal–ventral axis of the retina in *Xenopus* to control asymmetric development of the retina. The role of deiodinases in cone opsin expression has not yet been reported. However, T3 administration can suppress S opsin expression in cones in the neonatal mouse.

TRβ2 is the most conserved TRβ isoform, and it is found in the retina in a range of vertebrate species, including birds, amphibians, and humans, suggestive of a conserved role in cone photoreceptors. Color visual defects have been noted in rare cases of human resistance to TH. However, the incidence of any color visual deficiencies in this resistance syndrome or other human congenital thyroid disorders has not been investigated systematically and is currently unknown.

Conclusions

Studies in model species suggest that at least three periods exist during CNS development, each characterized by the type of TR that is expressed and by the amount of TH available. At early fetal stages, only TRα1 is expressed, the fetal thyroid gland is inactive,

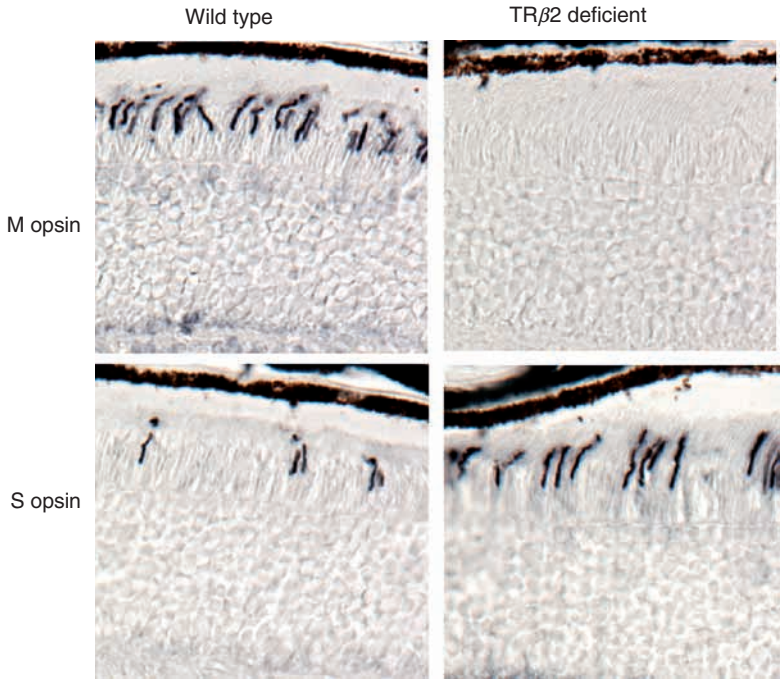

Figure 6 TRβ2 and cone photoreceptor development in mice. In most mammals, cone photoreceptors differentially express two opsins for sensitivity to short (S, 'blue') or medium-longer (M, 'green') wavelengths of light. Immunostaining detects M and S opsins in the cone outer segments of adult wild-type mice. TRβ2 is expressed in the newly generated cone photoreceptors where it programs the correct pattern of opsin expression. TRβ2-deficient mice lack M opsin and instead all cones express S opsin. With special thanks to Dr. Lily Ng for provision of this picture.

and small amounts of TH are transferred from the mother to the fetus. After the onset of TH production by the fetal thyroid gland at mid-gestation, the production of T3 is low and TRα1 is still the main TR present. Around the time of birth, there is a burst of TH production and an induction of TRβ. It is possible that different genes are controlled by TH during each developmental period.

The timetable of TH action varies between brain regions. Thus, neuronal proliferation is terminated soon after birth in the rodent cortex or cochlea, whereas it ceases later in the cerebellum and hippocampus. During these periods, TH exerts multiple effects via TRα1, TRβ, or both on the expression of genes that allow the formation of the neuronal network, that is, genes involved in establishing the connectivity between neurons and the production of energy for synaptic activity. Finally, terminal differentiation in tissues such as the hypothalamus, pituitary, cochlea, or photoreceptors requires the expression of specific genes regulated by TRβ. In some situations, much of the requisite T3 is produced locally by type 2 deiodinase. Type 3 deiodinase is probably important to protect cells from inappropriate T3 stimulation at certain stages.

The genes regulated by TH are often responsive only for a limited time within each developmental period, perhaps explaining why the multiple defects produced by congenital hypothyroidism are permanent if replacement therapy is not performed as soon as possible after birth. Such changes may also explain why the effects of TH during development and in adulthood are so different. The expression of some genes is not abolished but rather delayed by neonatal hypothyroidism. The temporal sensitivity of TH actions emphasizes the role of TH in the timing and synchronization of development during the construction of the neuronal network and in the process of achieving normal connectivity and behavior.

See also: Axonal Injury in Demyelinating Disease and CNS Injury; Cholinergic Pathways in CNS; Color Vision; Gene Expression Regulation: Chromatin Modification in the CNS; Sensory Aging: Hearing; Transcription Control and the Circadian Clock; Transcription and Reward Systems; Transcription Factors in Synaptic Plasticity and Learning and Memory; Transcriptional Networks and the Spinal Cord; Transcriptional Silencing.

Further Reading

Bernal J and Nunez J (1995) Thyroid hormones and brain development. *European Journal of Endocrinology* 133: 390–398.

Bradley DJ, Towle HC, and Young Iii WS (1992) Spatial and temporal expression of α- and β-thyroid hormone receptor mRNAs, including the β2-subtype, in the developing mammalian nervous system. *Journal of Neuroscience* 12: 2288–2302.

Deol MS (1976) The role of thyroxine in the differentiation of the organ of Corti. *Acta Otolaryngologica* 81: 429–435.

Eayrs JT (1966) Thyroid and central nervous development. In: Balazs R, Kovacs S, Cocks WA, Johnson AL, and Eayrs JT (eds.) *The Scientific Basis of Medicine Annual Review*, pp. 317–339. London: Athlone.

Esaki T, Suzuki H, Cook M, et al. (2003) Functional activation of cerebral metabolism in mice with mutated thyroid hormone nuclear receptors. *Endocrinology* 144: 4117–4122.

Fellous A, Lennon AM, Francon J, and Nunez J (1979) Thyroid hormones and neurotubule assembly *in vitro* during brain development. *European Journal of Biochemistry* 101: 365–375.

Friesema EC, Jansen J, Milici C, and Visser TJ (2005) Thyroid hormone transporters. *Vitamins and Hormones* 70: 137–167.

Itoh Y, Esaki T, Kaneshique M, et al. (2001) Brain glucose utilization in mice with a targeted mutation in the thyroid hormone alpha or beta receptor gene. *Proceedings of the National Academy of Sciences of the United States of America* 98: 9913–9918.

Legrand J (1979) Morphogenetic actions of thyroid hormones. *Trends in Neuroscience* 2: 234–236.

Marsh-Armstrong N, Huang H, Remo BF, Liu TT, and Brown DD (1999) Asymmetric growth and development of the *Xenopus laevis* retina during metamorphosis is controlled by type III deiodinase. *Neuron* 24: 871–878.

Ng L, Hurley JB, Dierks B, et al. (2001) A thyroid hormone receptor that is required for the development of green cone photoreceptors. *Nature Genetics* 27: 94–98.

Osler W (1897) Sporadic cretinism in America. *Transactions Congress of American Physicians and Surgeons* 4: 169–206.

Refetoff S, Weiss RE, and Usala SJ (1993) The syndromes of resistance to thyroid hormone. *Endocrine Reviews* 14: 348–399.

Rüsch A, Ng L, Goodyear R, et al. (2001) Retardation of cochlear maturation and impaired hair cell function caused by deletion of all known thyroid hormone receptors. *Journal of Neuroscience* 21: 9792–9800.

St Germain DL, Hernandez A, Schneider MJ, and Galton VA (2005) Insights into the role of deiodinases from studies of genetically modified animals. *Thyroid* 15: 905–916.

Thompson CC (1996) Thyroid hormone-responsive genes in developing cerebellum include a novel synaptotagmin and a *hairless* homolog. *Journal of Neuroscience* 16: 7832–7840.

Venero C, Guadano-Ferraz A, Herrero AI, et al. (2005) Anxiety, memory impairment, and locomotor dysfunction caused by a mutant thyroid hormone receptor alpha1 can be ameliorated by T3 treatment. *Genes and Development* 19: 2152–2163.

Whybrow P and Bauer M (2005) Behavioral and psychiatric aspects of thyrotoxicosis. In: Braverman L and Utiger R (eds.) *Werner and Ingbar's The Thyroid. A Fundamental and Clinical Text*. Philadelphia: Lippincott Williams & Wilkins.

Yoshimura T, Yasuo S, Watanabe M, et al. (2003) Light-induced hormone conversion of T4 to T3 regulates photoperiodic response of gonads in birds. *Nature* 426: 178–181.

Zoeller RT and Rovet J (2004) Timing of thyroid hormone action in the developing brain: Clinical observations and experimental findings. *Journal of Neuroendocrinology* 16: 809–818.

Tinnitus

P J Jastreboff, Emory University School of Medicine, Atlanta, GA, USA

© 2009 Published by Elsevier Ltd.

Introduction and Definitions

Tinnitus is defined as a perception of a sound that is not related to an acoustic source or electrical stimulation of the auditory pathways. As such it is a phantom auditory perception, that is, a perception of a sound without any corresponding vibratory activity within the inner ear. In this respect it is similar to the phenomena of phantom limb and phantom pain. Perception of phantom sound must last for at least 5 min to be classified as tinnitus since more than 90% of the population occasionally experiences a short perception of a phantom sound that disappears within seconds.

Somatosounds (sounds produced by a body) are sometimes classified as 'objective tinnitus,' in contrast to 'subjective tinnitus,' perceived without corresponding sound. Use of this classification has been criticized because detection of somatosound depends heavily on the equipment used. In this article, the term tinnitus is used to denote a phantom auditory perception.

Tinnitus is not a disease but a symptom triggered by variety of factors. It seems that in the majority of cases, it is linked to a dysfunction of the periphery of the auditory system, and only in approximately 1% of cases it can be associated with some medical problem. Indirect results, however, suggest that the inner ear (the cochlea) does not directly generate the tinnitus signal. For example, cutting the auditory nerve is not an effective method of alleviating tinnitus, and sometimes the perception of tinnitus does not even change. Moreover, in spite of many years of intensive study, there is no compelling evidence of any specific neuronal activity recorded from the auditory nerve of animals which can be clearly linked to tinnitus. This discrepancy could be explained by discordant dysfunction theory, discussed in the section titled 'Models of tinnitus.'

It is important to differentiate two subpopulations of people who have tinnitus. For only a small proportion of people with tinnitus is it bothersome and affects their lives, evoking anxiety, annoyance, panic, problems with concentration, sleep, and so forth. Consequently, it is labeled clinically significant tinnitus as these people need professional help. The effects of tinnitus may reach a profound level, to the point of totally controlling patients' lives and contributing to suicidal tendencies. On the other hand, the majority of people with tinnitus just experience it without any negative consequences.

General Description of Tinnitus Perception

Tinnitus can be perceived as a wide variety of sounds. The most typical descriptions are ringing, hissing, cycads, escaping steam, and roaring. Tinnitus can be intermittent or continuous and may fluctuate in volume or change pitch. It can be perceived in one ear, in both ears, or in the head. The perceived loudness may vary significantly over a period of days.

Examples of tinnitus and somatosounds are presented here. The sounds of tinnitus have been recreated by Dr. Jonathan Hazell, who used a musical synthetizer and interacted with patients until patients stated that the synthesized sound is very close to their tinnitus. In addition, a somatosound recorded in the ear canal of a patient and probably caused by jugular venous flow in the middle ear cavity is included, along with a somatosound from a patient diagnosed with palatal myoclonus.

Epidemiology

Continuous tinnitus is perceived by 10–20% of the general population, with bothersome tinnitus affecting 3–8%. It is estimated that highly debilitating tinnitus exists in about 0.5% of the general population.

Tinnitus occurs in all age groups. Recent data show that prevalence of tinnitus in children is similar to that observed in adults and that when children are bothered by tinnitus, it affects them to the same extent as it does adults. The prevalence of tinnitus perception increases significantly with aging; however, prevalence of the bothersome tinnitus declines for people 65 and older. As prevalence of hearing loss increases systematically with aging and prevalence of bothersome tinnitus decreases for older age groups, this observation argues against a simplistic link between bothersome tinnitus and hearing loss.

Medical Conditions Related to Tinnitus and Somatosounds

Traumatic noise exposure, hearing loss, and head trauma are typically associated with tinnitus. The majority of cases of vestibular schwannoma exhibit tinnitus, and for Ménière's disease, low-pitch roaring, (tinnitus) is an obligatory symptom for diagnosis.

Tinnitus accompany hyperacusis (defined as abnormally strong reactions to sound occurring within the auditory pathways) is present in more than 85% of cases, and hyperacusis is present in 25–30% of tinnitus cases.

It appears that hormonal changes such as those that occur during pregnancy, menopause, or thyroid dysfunction promote tinnitus. Permanent damage of the inner ear typically produces permanent tinnitus, while transient impairment of inner ear function (e.g., resulting from moderate noise overexposure) may yield temporary tinnitus (called 'disco tinnitus'). A long list of medications presumably inducing tinnitus is widely propagated; however, a critical review of the literature reveals that only aminoglycosides and quinine in high doses, some anticancer drugs, and withdrawal from benzodiazepines evoke tinnitus. Tinnitus could be a first indication of ototoxicity and emerges before hearing loss when patients are treated with ototoxic drugs. Salicylates (e.g., aspirin) reliably induce tinnitus in humans, and salicylates are used in animal models to produce tinnitus. As very high doses are needed to induce tinnitus, aspirin is not a significant factor clinically.

Contrary to common belief, there are no clear data proving that any dietary manipulation evokes or relieves tinnitus. Smoking has been associated with increased prevalence of tinnitus, and alcohol in moderation offers some protective effect.

Disorders of the somatosensory system (e.g., temporomandibular joint disorder) have been linked with tinnitus. Indeed, many patients can modulate their tinnitus by performing manipulations of the head and neck areas. Moreover, it is possible to evoke short temporal tinnitus by such manipulations in a significant proportion of people without tinnitus. Recent systematic neurophysiological studies have revealed strong somatosensory input to low levels of the auditory pathways such as the ventral and dorsal cochlear nuclei. It is possible to envision chronic tinnitus resulting predominantly from somatosensory input, but the data supporting such a possibility are limited. Currently, somatosensory modulation is used for research purposes (e.g., in imaging studies) as it offers a possibility of time-locked modulation of tinnitus.

Somatosounds have a variety of sources. Frequently they are related to the cardiovascular system, typically abnormalities of blood flow in the head area that are pulsatile in synchrony with the heartbeat. Nonpulsatile somatosounds could be related to palatal or middle ear myoclonus or a constantly open eustachian tube, resulting in the perception of air flow over its opening. Another interesting case is the rare perception of spontaneous otoacoustic emissions generated in the inner ear.

Measurements

Separate methods are needed for evaluation of tinnitus severity and for delineation of its perception. For tinnitus severity, a variety of questionnaires are used. Psychoacoustic properties of tinnitus are described by pitch match, loudness match, maskability (i.e., minimal level of white noise required to suppress its perception), and postmasking effects (disappearance of tinnitus for a certain time or increase in its loudness). Perceived pitch covers the full range of hearing, and in the majority of cases, tinnitus is perceived in the higher frequency ranges (above 4 kHz). The loudness match determined by psychoacoustic methods is stable for a given patient, disregarding subjective loudness perception. For the majority of patients, tinnitus loudness match is within 15 dB of their hearing threshold for the frequency corresponding to tinnitus pitch. Notably, there is no relation between psychoacoustic characterization of tinnitus and its severity.

A number of physiologic measurements have been tried for objective detection of tinnitus. Evaluation of the specifics of cochlear sound transduction, function of the olivocochlear efferent system, spontaneous activity of the auditory nerve, auditory brain stem responses, and late cortical potentials have failed to provide clear correlates with the presence of tinnitus. Substantial progress has been achieved, however, with positron emission tomography, functional magnetic resonance imaging, voxel-based morphometry, and magnetoencephalography. Strongest results were obtained with magnetoencephalography, which demonstrated changes in tonotopic organization of the auditory cortex in frequency areas corresponding to tinnitus pitch match. Furthermore, the extent of tonotopic map modification correlates with subjectively perceived strength of tinnitus.

Results of imaging studies regarding activation of various parts of the auditory system have frequently been inconsistent with one other. The majority of these studies, however, have shown significant activation of nonauditory structures, particularly the limbic system. Specifically, the auditory thalamus, various parts of the limbic system, the nucleus accumbens, the reticular formation, and the cerebellum have been shown to have changed activity in relation to tinnitus.

Differentiation between Experiencing Tinnitus and Being Bothered by Tinnitus

As tinnitus is a perception of a sound, the auditory system has to be involved. At the same time, there is a lack of correlation between psychoacoustic characterization of tinnitus and its severity or treatment outcome. These observations suggest a need to identify

two different classes of mechanisms: The first is those mechanisms related to generation to tinnitus-related neuronal activity (referred to, for brevity, in this article as the tinnitus signal) and tinnitus perception, which would be common to all people perceiving tinnitus regardless of whether they have a problem with it. The second class of mechanisms, specific to bothersome tinnitus, addresses the issue of how the tinnitus signal evokes negative reactions in the brain and body, resulting in tinnitus' being a problem.

Potential Generators of Tinnitus-Related Neuronal Activity

There is a consensus that in the majority of cases, the peripheral auditory system is involved in tinnitus generation. There is a small percentage of purely central tinnitus, such as occurs after cutting of the auditory nerve or in cases of auditory imagery or musical hallucinations. On the other hand, there is no consensus regarding other aspects of tinnitus-related neuronal activity. A dominant view is that tinnitus perception results from detection of an increase of the spontaneous activity within the auditory pathways. There are data, however, strongly pointing out that actually modification in temporal patterns of discharges, specifically bursting activity, is related to tinnitus and that the increase of spontaneous activity is rather linked to hearing loss, which typically accompanies tinnitus. The synchronization of the activity between neurons has been proposed as well, and although absent at the lower level of the auditory system, it might play some role in the auditory cortex.

The list of main proposed mechanisms of generation of tinnitus-related neuronal activity covers a wide variety of structures and dysfunctions, including hypoxia and ischemia in the cochlea; abnormal neurotransmitter release from the inner hair cells; abnormalities in transduction processes within the cochlea; damage of the auditory nerve; decreased activity of the cochlear efferent system; somatosensory–auditory interaction; neuronal plasticity, cortical reorganization of tonotopic maps, and hypersynchrony of neuronal discharges; enhanced sensitivity of the auditory pathways after decrease of the auditory input; and discordant damage or dysfunction of outer and inner hair cells.

It is expected that for various types of tinnitus, different mechanisms are responsible for tinnitus emergence. Until now none has been proven. While other postulates might be correct, there is growing evidence supporting the last two hypotheses.

Until the late 1980s there was no accepted animal model of tinnitus, and it was difficult to validate proposed hypotheses because of inherent limitations of research on humans. A number of animal models are available now, allowing study of the physiological aspects of tinnitus perception.

Enhanced Sensitivity of the Auditory Pathways and Emergence of Tinnitus after Decreased Auditory Input

It is possible to induce temporary tinnitus in a majority of people just by placing them in a very quiet environment. Moreover, tinnitus gets louder under such conditions. One of the mechanisms potentially involved in both these observations is enhanced sensitivity and increased amplification within the auditory pathways after decreased auditory input. Behavioral data argue that the auditory system (as other sensory systems) works as an automatic gain control system, that is, when input signal is low, then amplification increases, and when input signal is higher, then gain of the system is lowered. In an anechoic chamber, participants hear weak, normally unnoticeable sound, such as heartbeat or the movement of clothes caused by breathing, with all these sounds perceived as being quite loud. Recordings of single-unit activity from various centers of the auditory pathways have shown that the sensitivity of the neurons does indeed increase under conditions of low auditory input or when the inner ear is damaged. Opposite behavioral effects are observed when participants are in noisy environments. Enhanced amplification within the auditory pathways may happen as a result of various reasons and evoke or enhance tinnitus or induce hyperacusis.

Models of Tinnitus

A challenge to modeling tinnitus is the lack of abundant neurophysiological data and the fact that there are likely to be several types of tinnitus, each generated by a different underlying mechanism. A starting point for models is the apparent paradox of a lack of tinnitus-related activity in the cochlea coupled with the presence of sound perception, pointing to a central origin for the generation of tinnitus. The observation that tinnitus often accompanies hearing loss focuses attention on the central auditory pathway. One simple model would propose that a normal balance of excitatory and inhibitory input to some center in the pathway is disrupted by hearing loss. If the disruption is to the inhibitory input, the resulting excitatory input is unbalanced and could lead to tinnitus. This discordant dysfunction theory has been applied to the dorsal cochlear nucleus, a cochlear nucleus subdivision that has abundant interneurons, many of which are inhibitory. Indeed, experimental evidence in animals shows abnormally high 'spontaneous' activity in this center accompanying hearing loss. The particular neural pathways involved remain obscure. The auditory nerve

does provide two types of input to the dorsal cochlear nucleus: type I fiber inputs from the inner hair cells and type II fiber inputs from the outer hair cells. Inner hair cells are contacted by 95% of afferent fibers of the auditory nerve (which are type I, thick, fast, and myelinated), and these cells act as a transducer, converting the vibration of the basilar membrane of the cochlea into electrical impulses within the auditory nerve. Total deafness occurs if all these cells are damaged, even with untouched outer hair cells. Outer hair cells function as a mechanical amplifier, enhancing vibration for the lower half of the range of perceived sound levels by about 60 dB. It is possible to have normal hearing even if 30% of outer hair cells are damaged. Rapid decrease of their function has been associated with an emergence of tinnitus in humans. They are contacted by about 5% of auditory nerve fibers (type II, thin, slow, and unmyelinated). Outer hair cells are preferentially damaged by intense sound exposure. However, both type I and type II inputs are likely to excite their targets in the brain, and thus further inhibitory interneurons, perhaps to one of these pathways within the dorsal cochlear nucleus, would need to be involved to generate unbalanced excitation.

Specifically, the discordant dysfunction theory proposes that when an imbalance of functional properties of outer and inner hair cells occurs, it yields the disinhibition of spontaneous neuronal activity in the dorsal cochlear nucleus, causing a general increase of the spontaneous activity and emergence of bursting, epileptic-like activity, which after further processing within the auditory pathways is perceived as tinnitus. Therefore, while there is no vibratory activity related to tinnitus in the cochlea, the cochlea still may serve as a source of the tinnitus neuronal signal.

The Neurophysiological Model of Tinnitus

Disregarding specifics of mechanisms yielding the presence of the tinnitus signal in the auditory pathways, there is consensus that once it is detected, it is perceived as a sound. Little attention has been given to mechanisms involved in converting tinnitus from benign perception of a sound to a significant problem affecting people's lives. In the past, the dominant approach was through psychology, with no attempt to identify specific neurophysiological structures and mechanisms involved. The neurophysiological model of tinnitus, which attempts to approach tinnitus from the point of view of neuroscience, points out the inherent involvement of the central nervous system (CNS) in tinnitus, and attempts to indicate systems, structures, and the role of specific mechanisms of the brain function in clinically significant tinnitus.

The observation that tinnitus pitch, maskability, and loudness are not correlated at all with tinnitus severity for the population of patients strongly supports the postulate that the auditory system is secondary in clinically significant tinnitus. The neurophysiological model of tinnitus postulates that in the case of bothersome tinnitus, systems other than the auditory system of the brain have to be involved. The limbic and autonomic nervous systems play the major role, and other systems, such as prefrontal cortex and cerebellum, may be involved as well. Furthermore, the model stresses strongly the importance of subconscious processing of the information and the importance of the principles of conditioned reflexes in governing the connections between auditory and other brain systems.

Briefly, tinnitus-related neuronal activity which originates in the auditory system starts to activate other brain systems because of both cognitive assessment of tinnitus and subconscious connections created as the result of one of the principles of conditioning. Temporal association of signal (tinnitus) and reinforcement (annoyance, strong emotional upheaval due to any reason) are sufficient to produce this conditioned reflex. Indeed, a study performed to identify factors most frequently linked to the emergence of clinically significant tinnitus revealed that in a majority of cases, it develops as the result of factors totally irrelevant to hearing, such as retirement, losing a job, divorce, and so forth.

The diagram of the main systems and connections involved in clinically significant tinnitus is presented in Figure 1(a). In cases when people experience tinnitus but are not bothered by it, only the auditory system is activated. Once tinnitus-related neuronal activity spreads to the limbic and the sympathetic part of the autonomic nervous systems, the new functional connections are created, enhanced, and governed to a large extent by the principles of conditioned reflexes. The autonomic nervous system consists of two reciprocally antagonistic parts, the sympathetic and the parasympathetic systems. Activation of the sympathetic system prepares people for physical or mental action, and activation of the parasympathetic system sets the stage for rest and relaxation. Overactivation of the sympathetic system yields stress, anxiety, panic attacks, and the fight-or-flight reaction. All behavioral reactions exhibited by tinnitus patients, such as anxiety, annoyance, panic, decreased ability to enjoy life activities, problems with attention, and problems with sleep, are consistent with activation of the sympathetic system. When the term 'autonomic nervous system' is used in this discussion, it denotes the sympathetic system unless otherwise specified. According to the neurophysiological model of tinnitus, problems

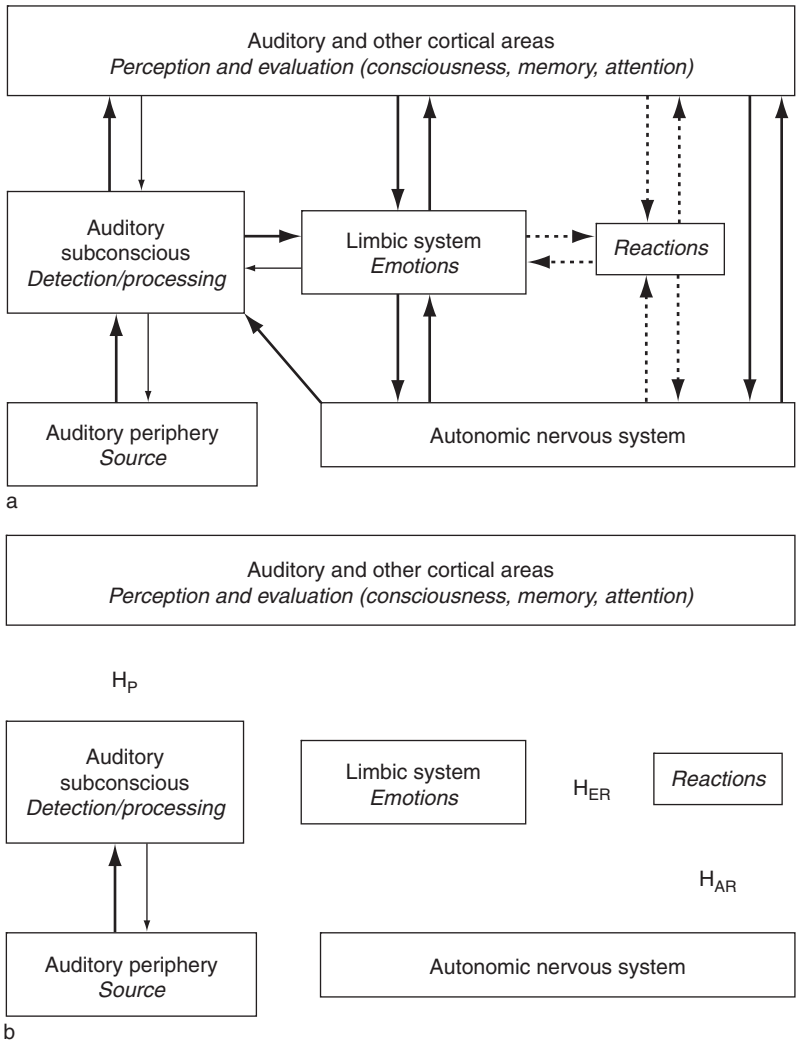

Figure 1 The neurophysiological model of tinnitus. (a) The block diagram outlining structures and connections involved in clinically significant tinnitus. The tinnitus signal, typically generated at the periphery of the auditory system, is detected and processed in subconscious pathways of the auditory system and finally perceived at the auditory cortex. If tinnitus is classified as an important negative stimulus, self-enhancing loops, governed by principles of conditioned reflexes, develop. Note existence of two loops: high, involving consciousness, and low, the subconscious loop. (b) Specific functional connections at which habituation of tinnitus occurs. H_{ER}, habituation of emotional reactions; H_{AR}, habituation of reactions evoked by the autonomic nervous system; H_P, habituation of tinnitus perception. The primary goal of tinnitus retraining therapy is to achieve habituation of reactions; habituation of perception will follow automatically.

arise from proper reaction of the brain to an inappropriate stimulus, that is, the reactions would be appropriate to a life-threatening stimulation, but tinnitus perception results from compensatory action of the auditory system to a situation which does not indicate danger.

The activation of the limbic and autonomic nervous systems by tinnitus signal occurs via two pathways: conscious and subconscious, or high and low loops, respectively (Figure 1(a)). The conscious loop involves high cortical areas, perception, cognition, and verbalization. During the process of development of clinically significant tinnitus, this path is created first. Tinnitus attracts attention, thoughts, concerns, and fears, which activate the limbic and autonomic nervous systems. If these initial thoughts are properly addressed by demystification of tinnitus and its categorization as a neutral stimulus, initial activation of the limbic and autonomic nervous systems is extinguished, and tinnitus does not become a problem. If, however, these thoughts are not addressed properly or, even worse, are enhanced by negative counseling, these connections become stronger, and problems caused by tinnitus are enhanced. The presence of reciprocal connections – signals from the auditory subcortical and cortical areas activating the limbic system and evoking

(negative) emotions – in turn changes reactivity of the auditory centers, causes enhancement of initial connections, and finally results in establishment of strong reactions of the limbic and autonomic nervous systems. These reactions act as negative reinforcement for a conditioned reflexes arc which links the tinnitus-related neuronal activity within the auditory system (conditioned stimulus) with reactions. They enhance cognitive concerns as well as further worsening the problem. Note that the outlined processes by definition involve neuronal plasticity.

At a subconscious level, the tinnitus signal is going through strong functional connections between the medial geniculate body and the lateral nucleus of the amygdala, which is one of the crucial components of the limbic system. This low loop develops after the high loop, it is governed by principles of conditioned reflexes, and it might become dominant in chronic tinnitus. Therefore, because of the presence of the low loop, the activation of the limbic and autonomic nervous systems might occur even without conscious perception of tinnitus. Notably, a signal conveyed via this route cannot be directly modified by conscious thinking, and a method appropriate for retraining of conditioned reflexes would be more appropriate.

Treatments

Treatment of tinnitus is a challenging problem at both the theoretical and the clinical level. Observations of what works and what does not provide some insight into mechanisms of tinnitus. Thousands of years of attempts at treatment (tinnitus was first described in Babylonian clay tablets) have not brought consistently positive results. Only selected treatment methods are outlined.

Suppression or Weakening the Tinnitus Signal

The first class of approaches is aimed at suppressing tinnitus perception temporarily (masking) or partially decreasing it. In neurophysiological terms, total suppression of tinnitus perception would mean that detection of tinnitus-related neuronal activity is prevented by providing the auditory system with another stimulation which would make detection of the tinnitus signal impossible. Attempts to apply this method have revealed that clinically relevant suppression of tinnitus perception in practice can be achieved in only a small percentage of cases. Direct stimulation of the auditory nerve via a cochlear implant increases the effectiveness by about 50%, but this method can be used only by the very small proportion of tinnitus patient who are profoundly deaf. Electrical stimulation applied outside the cochlea has failed to be substantially effective. Direct electrical stimulation of the auditory cortex is under investigation but has failed so far to show superior results.

A weakening of tinnitus-related neuronal activity can be achieved via a variety of methods. The simplest, and the most popular, is by enriching auditory background stimulation, and it is implemented in a wide variety of sound therapy. Two neurophysiological mechanisms are bases for decreasing the strength of the tinnitus signal (i.e., tinnitus-related neuronal activity). First, the strength of any signal depends on its difference with background rather than on its absolute physical value. Second, the sensitivity to sensory signals depends on longer-term average input. Consequently, by an increase of the auditory background, the tinnitus signal is decreased due to the combined action of both mechanisms. The type of sound is in this respect secondary; however, it plays a role because of other factors (e.g., ease of habituation, inherent preference for certain sound levels, emotional and cognitive connotations).

Another class of treatments attempts to modify the tinnitus signal at the central level rather than by stimulation of the periphery. In addition to direct electrical stimulation of the auditory cortex, transcranial magnetic stimulation has been tried recently. Strong pulses of the magnetic field are applied to the skull outside the cortical area to be stimulated. Magnetic field crosses the skull easily and evokes a strong electrical field in the cortex, which in turn stimulates cortical fibers. The effects on the brain function depend on the frequency of stimulation. High frequency evokes direct inhibitory action on the cortical activity, while low-frequency stimulation appears to induce plastic changes in organization of cortical networks and potentially their interaction with subcortical centers. While it is not clear what specific reorganizations occur, this approach has been shown to attenuate tinnitus. The effects are mild, however, and it is not clear whether they are long lasting. Nevertheless, observation that transcranial magnetic stimulation has some impact on tinnitus supports the importance of the modification of the CNS in tinnitus treatment.

Psychological Approaches

Psychological treatments cover a variety of approaches, from simple attempts to distract attention from tinnitus to improved coping and up to behavioral cognitive therapy. Behavioral cognitive therapy seems to be most effective in the class of psychological treatment and is based on the assumption that because thoughts evoke emotions, it is possible by changing thoughts to

modify emotions evoked by tinnitus. There are well-established, effective methods to achieve this goal, and this method is effective in a significant proportion of patients. According to the neurophysiological model of tinnitus, behavioral cognitive therapy nullifies the high loop and therefore decreases stimulation of the limbic and autonomic nervous systems.

Medications

Pharmacological approaches have been tried for a long time and, in spite of their lack of success, are still among the most popular. Some medications aim at the periphery of the auditory system, others at the limbic and autonomic nervous systems to decrease anxiety and annoyance and improve coping. Trials of new pharmacological candidates are ongoing, but so far research has failed to show any drug to be consistently effective for tinnitus.

Tinnitus Retraining Therapy

All treatment methods presented so far aim at modification of tinnitus perception or signal or alleviation of tinnitus-evoked reactions. Tinnitus retraining therapy (TRT) differs in this respect as it aims to retrain the brain to extinguish specific, tinnitus-related functional connections between the auditory and the limbic and autonomic nervous systems and through this mechanism to decrease or remove tinnitus-evoked reactions (Figure 1(b)). TRT attempts to modify neither the source of the tinnitus signal nor the brain centers involved in reacting to it. As a consequence, TRT is independent of the source of the tinnitus signal (tinnitus etiology) and can be used to treat any type of tinnitus.

Briefly, TRT consists of counseling and sound therapy, both strictly based on the neurophysiological model of tinnitus. The goal is to achieve habituation of the tinnitus-evoked reactions, which will be followed automatically by habituation of tinnitus perception (Figure 1(b)). Habituation of reactions is basically identical to the passive extinction of conditioned reactions. TRT uses a modified version of passive extinction of conditioned reflexes in which both the reactions and the stimulus are decreased simultaneously. Habituation of tinnitus can be achieved by a variety of methods. TRT's approach is easily implemented and works on both the tinnitus and the decreased sound tolerance (hyperacusis and misophonia) that frequently coexists with tinnitus.

The role of counseling is to reclassify the tinnitus signal into the category of neutral stimuli, as habituation of any stimulus is difficult when the stimulus has negative connotations and impossible when these associations are very strong. It effectively disrupts the higher loop and decreases stimulation of the limbic and autonomic nervous systems by cognitive processes. In this respect it is similar to behavioral cognitive therapy but uses a different counseling approach.

Sound therapy is the second part of TRT. It aims to decrease the strength of the tinnitus signal through application of a number of guides provided by the neurophysiological model of tinnitus (e.g., use of sound during the night and avoiding annoyance and suppression of tinnitus perception).

Future Goals

It has been argued that tinnitus is a phantom auditory perception and should be analyzed from the point of view of neuroscience rather than by ear-oriented (e.g., otolaryngological or audiological) or mind-oriented (e.g., psychological) approaches. Significant progress in understanding mechanisms of tinnitus and its treatment has been achieved in large part because of the shift from earcentric or black-box approaches to neuroscience. A number of important neurophysiological questions remain, however.

Significant efforts focus on delineating structures in the brain which are involved in the phenomenon of tinnitus. Imaging studies are consistent in indicating involvement of the limbic system and parts of the brain involved in anxiety, but data from the auditory system are much more variable and unclear.

Researchers still debate what tinnitus-related neuronal activity looks like. The dominant belief, that it takes the form of increased spontaneous activity, has been challenged, and it has been proposed that changes in temporal patterns of neuronal dischargers, involving high-frequency, bursting, epileptic-type activity, is in reality a neuronal correlate of tinnitus.

There is even less clarity on the neurotransmitters and receptors involved. Gamma-aminobutyric acid, glutamate, serotonin, and other systems, as well as *N*-methyl-D-aspartate receptors, have been implicated. There are, however, no data proving clear involvement of any of these systems, and drug trials for tinnitus so far argue against a dominant involvement of any of these systems in tinnitus.

Due to obvious limits on research on humans, animal models of tinnitus are extremely important. Since the mid-1980s, when the first valid animal model of tinnitus was introduced, a variety of models have emerged, including most recently a model based on prepulse inhibition of the startle reaction. By acknowledging the crucial involvement of the CNS in tinnitus and combining animal and human research, it should be possible to achieve better understanding of this elusive phenomenon and to find more effective methods for its alleviation.

See also: Auditory System: Central Pathway Plasticity; Auditory System: Central Pathways; Auditory/Somatosensory Interactions; Deafness; Hair Cells: Sensory Transduction.

Further Reading

Chen G-D and Jastreboff PJ (1995) Salicylate-induced abnormal activity in the inferior colliculus of rats. *Hearing Research* 82: 158–178.

Eggermont JJ and Roberts LE (2004) The neuroscience of tinnitus. *Trends in Neurosciences* 27: 676–682.

Jastreboff PJ (1990) Phantom auditory perception (tinnitus): Mechanisms of generation and perception. *Neuroscience Research* 8: 221–254.

Jastreboff PJ, Brennan JF, Coleman JK, and Sasaki CT (1988) Phantom auditory sensation in rats: An animal model for tinnitus. *Behavioral Neuroscience* 102: 811–822.

Jastreboff PJ and Hazell JWP (2004) *Tinnitus Retraining Therapy: Implementing the Neurophysiological Model*. Cambridge: Cambridge University Press.

Jastreboff PJ and Jastreboff MM (2006) Tinnitus retraining therapy: A different view on tinnitus. *ORL: Journal for Oto-Rhino-Laryngology and Its Related Specialties* 68: 23–29.

Jastreboff PJ and Jastreboff MM (in press) Tinnitus and hyperacusis. In: Ballenger JJ, Snow JB Jr., and Ashley WP (eds.) *Ballenger's Otorhinolaryngology Head and Neck Surgery*, 17th edn., ch. 31. San Diego, CA: Singular Publishing.

Snow JB (2004) *Tinnitus: Theory and Management*. London: BC Decker.

TIP39 (Tuberoinfundibular Peptide 39 Residues)

T B Usdin, National Institute of Mental Health, Bethesda, MD, USA
A Dobolyi, Hungarian Academy of Sciences and Semmelweis University, Budapest, Hungary

Published by Elsevier Ltd.

Introduction

Tuberoinfundibular peptide of 39 residues (TIP39) is a neuropeptide discovered in 1999. It is part of an extended family of polypeptide hormones that includes parathyroid hormone (PTH), parathyroid hormone-related protein (PTHrP), secretin, vasoactive intestinal polypeptide (VIP), pituitary adenylate cyclase-activating polypeptide (PACAP), glucagon, glucagon-like polypeptide 1 and 2 (GLP1 and GLP2), glucose-dependent insulinotropic polypeptide (GIP), growth hormone-releasing hormone (GHRH), and corticotropin-releasing hormone (CRH). These peptides act on members of the Family B group of guanine nucleotide binding protein-coupled receptors (GPCRs). Family B receptors share a predicted architecture with other GPCRs, including an extracellular N-terminus, seven membrane-spanning domains that anchor them to the plasma membrane, and a cytoplasmic C-terminus. Family B receptors have few amino acid sequences in common with the largest group of GPCRs, the Family A or rhodopsin-like receptors, or with Family C, which includes the metabotropic glutamate receptors. Family B receptors share a clear but graduated amount of amino acid sequence similarity, a relatively large N-terminus, and a number of highly conserved extracellular cysteine residues. Ligand activation of Family B receptors leads to the stimulation of cyclic adenosine monophosphate (cAMP) synthesis and, in many cases, a tissue-dependent increase in cytoplasmic calcium concentration. TIP39 activates the parathyroid hormone 2 (PTH2) receptor. Quite a bit is known about the structures and neuroanatomical distributions of TIP39 and the PTH2 receptor, but current data provide only hints about their potential physiological functions.

Discovery of TIP39

The discovery of TIP39 stemmed from the identification of the PTH2 receptor. Sequence analysis of receptors for secretin, VIP, and PTH, which were each cloned in 1991, suggested that they formed a distinct receptor family. Sequence identities between the initially described Family B GPCRs were exploited to screen for additional related receptors. In the approach that led to the identification of the PTH2 receptor, a cDNA clone with homology to known Family B receptors was isolated and then introduced into tissue culture cells, which were subsequently screened for stimulation of second-messenger production by a panel of plausible ligands. The PTH2 receptor was so named because it was potently activated by PTH, and not by any of the other potential ligands tested, and because another receptor activated by PTH had been previously described. The first described PTH receptor was isolated from a kidney cell line. It is expressed at relatively high levels in skeletal tissues and kidney, and it clearly mediates the major effects of PTH on mineral homeostasis. It appears not to discriminate between PTH and the distinct protein, PTHrP. Thus, it was originally designated the PTH/PTHrP receptor. It is now also referred to as the PTH1 receptor. The PTH2 receptor was identified in brain-derived cDNA libraries. It is relatively highly expressed in a number of central nervous system areas and at extremely low levels in bone and the kidney.

In contrast to the PTH2 receptor, there have been only a few reports, based on PCR, of PTH mRNA in the rodent brain. Our group failed to detect it using several sensitive techniques. This led to the hypothesis that the brain might contain a different PTH2 receptor ligand. An activity consistent with this possibility was found in an acid extract prepared from bovine hypothalamus. This extract stimulated cAMP production much more potently in tissue culture cells expressing the PTH2 receptor than in their parent cell line or cells expressing the PTH1 receptor. From an antibody-absorption experiment, the activity also appeared to be immunologically distinct from PTH. The ligand specificity of the PTH2 receptor was initially determined using a human cDNA clone. Subsequently, it was observed that PTH was much less potent at the rat PTH2 receptor than at the human PTH2 receptor and that activity present in the hypothalamic extract caused a significantly greater maximal production of cAMP via the rat PTH2 receptor than did PTH. This collection of evidence suggested the existence of an endogenous PTH2 receptor ligand distinct from PTH.

TIP39 was isolated using a classic peptide hormone purification strategy. A crude peptide extract prepared from large (50-kg) batches of bovine hypothalamus was fractionated by a series of chromatography steps. At each stage, fractions containing the activity of interest were identified by their ability to much more potently stimulate cAMP formation in

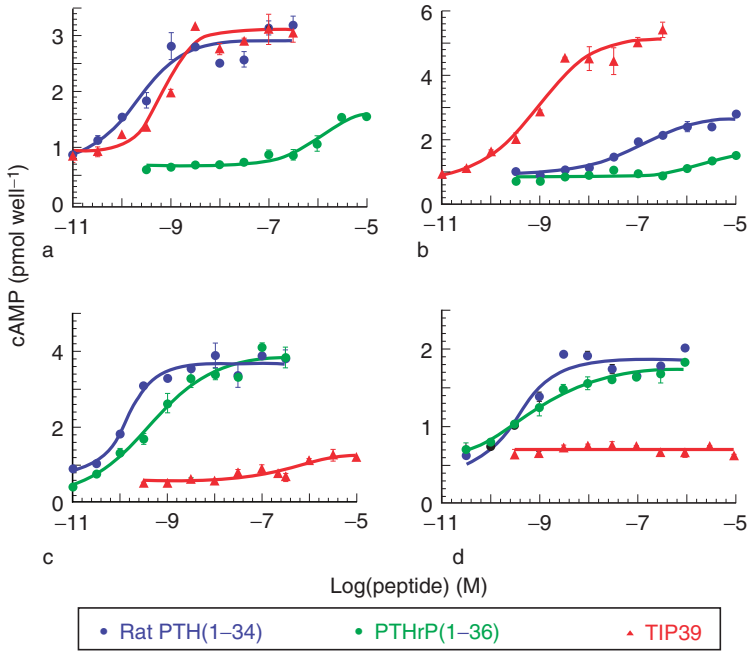

Figure 1 Ligand selective activation of parathyroid hormone (PTH) receptors: (a) human PTH2 receptor; (b) rat PTH2 receptor; (c) human PTH1 receptor; (d) rat PTH1 receptor. Dose-dependent stimulation of cAMP accumulation by TIP39, parathyroid hormone, and parathyroid hormone-related protein (PTHrP) in tissue culture cells expressing the cloned human or rat PTH1 or PTH2 receptor. cAMP, cyclic adenosine monophosphate. Modified from Usdin TB, Hoare SRJ, Wang T, Mezey E, and Kowalak JA (1999) TIP39: A new neuropeptide and PTH2 receptor agonist from hypothalamus. *Nature Neuroscience* 2: 941–943.

PTH2 receptor-expressing cells than in similar cells expressing the PTH1 receptor. Once a fraction containing a single major species was obtained, its amino acid sequence was determined using a combination of microscale Edman degradation and mass spectrometry. The sequence was a 39-residue polypeptide without any posttranslational modification. To prove that the sequenced peptide was in fact responsible for the activation of the PTH2 receptor, it was chemically synthesized. Chemically synthesized TIP39 has exactly the properties predicted from the initial crude acid extract. It potently stimulates cAMP production via both human and rat PTH2 receptors, it has little or no effect at PTH1 receptors, it is approximately equipotent with PTH at the human PTH2 receptor, and its intrinsic efficacy is significantly greater than that of PTH at the rat PTH2 receptor (Figure 1).

TIP39, Parathyroid Hormone, and Parathyroid Hormone-Related Protein Form a Discrete Peptide Family

Only bovine TIP39 protein has been purified, but TIP39's sequence has been deduced for human, rat, mouse, zebra fish, fugu, and tilapia using reverse transcription polymerase chain reaction (RT-PCR) and genome sequence analysis. Human and mouse predicted TIP39 transcripts encode a polypeptide of 98 residues composed of a hydrophobic leader sequence followed by a predicted signal sequence cleavage site and then approximately 30-amino-acid residues before two arginine residues (a common endopeptidase cleavage site) that precede the 39-amino-acid peptide that was purified. Mature PTH and PTHrP are polypeptides of approximately 84 and 170 residues, depending on the species and tissue. A 34- or 36-residue N-terminal fragment is sufficient for high-affinity binding to, and activation of, the PTH1 receptor. Using human sequences for reference, 11 of the first 34 residues are identical between PTH and PTHrP. Human TIP39 contains four of the residues shared by PTH and PTHrP and an additional four residues of PTH (Figure 2). Interaction of the C-terminal half of PTH, PTHrP, and TIP39 with their receptors is largely responsible for their high-affinity binding, whereas N-terminal peptide sequences are critical for receptor activation. Most of the residues that TIP39 shares with PTH and PTHrP are within the C-terminal receptor-binding part of the peptide. Although the number of amino acid identities between TIP39 and PTH is small, the peptides are predicted to form nearly superimposible α helices. In addition, although TIP39 completely fails to activate PTH1 receptors and has very low binding affinity, progressively removing

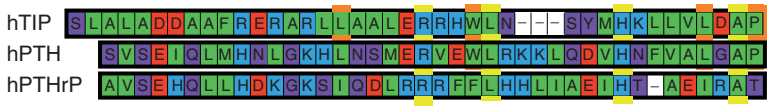

Figure 2 Amino acid sequence alignment of human TIP39 (hTIP), with N-terminal portions of human parathyroid hormone (hPTH) and human parathyroid hormone-related protein (hPTHrP). Standard single-letter amino acid symbols are color coded by functionality. Acidic residues have a red background, basic blue, hydrophobic green, and hydrophilic purple. Residues identical in the three polypeptides are indicated with yellow bars, and additional identities between TIP39 and PTH with orange. In addition to a limited number of amino acid identities between TIP39 and PTH and PTHrP, there is an overall pattern of similar residues. TIP39, tuberoinfundibular peptide of 39 residues.

N-terminal residues increase its affinity for the PTH1 receptor to the point that an analog of TIP39 with its first six residues deleted is a potent competitive antagonist at the PTH1 receptor (6 nM affinity). Thus, TIP39, PTH, and PTHrP form a peptide family with partially overlapping selectivity for the PTH1 and PTH2 receptors.

Two Discrete Groups of TIP39 Neurons Have Widespread Projections

There are two major groups of neurons that synthesize TIP39 (Figure 3). One is at the thalamic-midbrain junction. The distribution of these cells does not fit well within described anatomical borders. They are concentrated near the midline around the subparafascicular nucleus and spread caudo-laterally through the parvicellular subparafascicular nucleus (Figure 3(a)). At the midline, they are intercalated with the A11 group of dopaminergic cells. The TIP39 neurons that extend laterally partially overlap a strip of calcitonin gene-related peptide (CGRP)-containing neurons, but are generally somewhat medial to these cells. There is only a small amount of co-localization of CGRP and TIP39. The second significant group of TIP39 neurons is in the lateral part of the rostral pons, in the medial paralemniscal nucleus (Figure 3(b)). None of the TIP39 neurons contains markers for the classical neurotransmitters glutamate, γ-aminobutyric acid (GABA), acetylcholine, or catecholamines, and their location does not correspond to serotonin-, histamine-, or acetylcholine-synthesizing neurons.

Both groups of TIP39 neurons have widespread and distinct projections (Figure 4). The subparafascicular TIP39 neurons have more anterior projections, including to specific parts of the cerebral cortex, deep forebrain structures, amygdala, thalamus, and hypothalamus. Neurons in most of these areas also send projections to the areas of TIP39 neuron concentration. In general, these areas can be categorized as part of the limbic system or involved in autonomic or endocrine regulation. The TIP39 neurons in the medial paralemniscal nucleus have more caudal projections, including several midbrain, pontine, and brain stem regions, as well as some that descend into the spinal cord. Many of the areas that receive projections from paralemniscal TIP39 neurons are involved in processing auditory information.

There is an excellent correlation between the distributions of TIP39- and PTH2 receptor-containing fibers (Figure 5). Antibodies to both TIP39 and the PTH2 receptor label neuronal fibers and terminals much more strongly than cell bodies. Stronger labeling of processes suggests that most of these proteins are transported out of the their cell bodies and function in axonal terminals or dendrites. Many areas of TIP39 concentration are far removed from TIP39-synthesizing cell bodies. TIP39 disappears from fibers and terminals following lesions that damage the areas of cell body concentration, and microinjections of neuronal tract-tracing agents prove the actual axonal projections of TIP39 neurons. Neurons that synthesize PTH2 receptors are spread much more widely than TIP39-synthesizing neurons. Many areas contain both PTH2 receptor-synthesizing cell bodies, identified by *in situ* hybridization detection of mRNA, and PTH2 receptor-containing fibers, identified by antibody labeling. Thus, it is less clear whether the PTH2 receptor functions in dendrites, nerve terminals, or both because the PTH2 receptor-containing fibers near PTH2 receptor-synthesizing cell bodies could be dendrites, local axonal arborizations, or projections of distant neurons. Immunohistochemical double-labeling shows a co-localization of PTH2 receptors and a vesicular glutamate transporter (VGLUT2). Because VGLUT2 functions to package glutamate into secretory vesicles, this suggests a presynaptic, nerve terminal, localization of PTH2 receptors.

TIP39 Expression Decreases and Becomes Sexually Dimorphic during Postnatal Development

In rat, TIP39 expression can be detected within a few days of birth, and at this time it has a distribution essentially like that seen in older animals. The level of expression increases until around postnatal day 14, and it rapidly decreases after this (Figure 6). The expression of TIP39 in males and females is not obviously different until it starts to decline. The

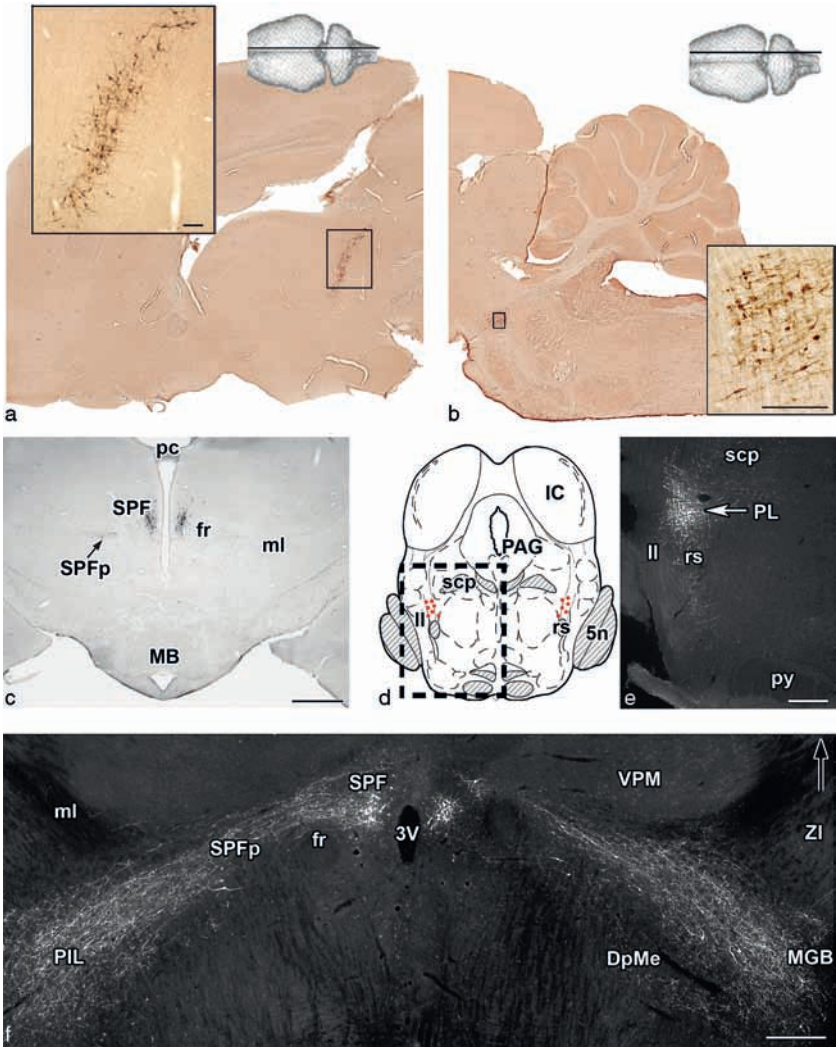

Figure 3 Rat brain distribution of TIP39 neurons: (a) sagittal slice, subparafascicular area (inset scale bar = 1 mm); (b) sagittal slice, medial paralemniscal nucleus (inset scale bar = 1 mm); (c) coronal slice, subparafascicular area (scale bar = 1 mm); (d) drawing in the coronal plane, medial paralemniscal nucleus; (e) photomicrograph, corresponding to the framed area in (d), with fluorescent immunolabeling (scale bar = 500 μm); (f) horizontal section with fluorescent immunolabeling, posterior intralaminar complex of the thalamus (scale bar = 500 μm). In (a–b), immunolabeling for TIP39 is shown in sagittal rat brain sections. The framed areas are magnified in the insets to more clearly show labeled cell bodies. In the medial part of the subparafascicular area (a), the TIP39 cell group has a sigmoid shape in sagittal sections. TIP39 neurons are distributed around the magnocellular subparafascicular nucleus between the midline and the fasciculus retroflexus in the periventricular gray of the thalamus, an area that has no obvious histological distinction from the more caudal periaqueductal gray. In (e), paralemniscal TIP39 neurons are located above the rubrospinal tract immediately medial to the lateral lemniscus in the rostral pons. In (f), the laterally located TIP39 neurons are continuous with the medial subparafascicular area near the middle of its sigmoid shape. TIP39 neurons are located in the parvicellular subparafascicular nucleus below the fasciclus retroflexus. A long line of TIP39 neurons that reaches the medial part of the medial geniculate body is situated further caudo-laterally in the posterior intralaminar complex of the thalamus. 3V, third ventricle; 5n, trigeminal nerve; DpMe, Deep mesencephalic nucleus; fr, fasciculus retroflexus; IC, inferior colliculus; ll, lateral lemniscus; MB, mamillary body; MGB, medial geniculate body; ml, medial lemniscus; PAG, periaqueductal gray; pc, posterior commissure; PIL, posterior intralamniar nucleus of the thalamus; PL, medial paralemniscal nucleus; py, pyramidal tract; rs, rubrospinal tract; scp, superior cerebellar peduncle; SPF, subparafascicular area; SPFp, parvicellullar subparafascicular nucleus; TIP39, tuberoinfundibular peptide of 39 residues; VPM, ventral posteromedial thalamic nucleus; ZI, zona incerta. (a, b) Modified from Usdin TB, Dobolyi A, Ueda H, and Palkovits M (2003) Emerging functions for tuberoinfundibular peptide of 39 residues. *Trends in Endocrinology and Metabolism* 14: 14–19. (c–f) Modified from Dobolyi A, Palkovits M, and Usdin TB (2003) Expression and distribution of tuberoinfundibular peptide of 39 residues in the rat central nervous system. *Journal of Comparative Neurology* 455: 547–566.

decrease is greater in males than in females, so adult females have significantly greater expression of TIP39 than adult males. After rats are approximately 3 months old, it becomes difficult to detect TIP39 in males, and female levels are dramatically lower than in younger animals. The developmental pattern is similar in the subparafascicular and paralemniscal TIP39 neurons. Postpubertal gonadectomy has no

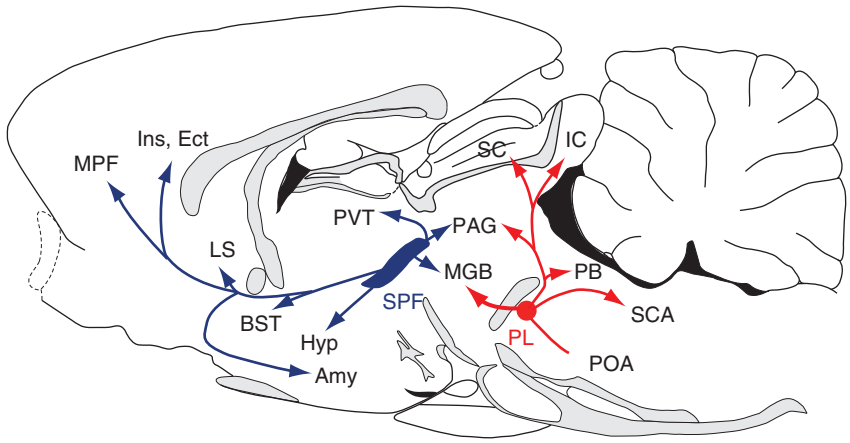

Figure 4 Projection of TIP39 neurons. The major target areas of the subparafascicular TIP39 neurons (SPF) are the medial prefrontal (MPF), insular cortex (Ins), and ectorhinal cortex (Ect), lateral septum (LS), the amygdala (Amy; in particular the central amygdaloid nucleus, amygdalo-striatal transitional zone, posterodorsal medial amygdaloid nucleus, and amygdalo-hippocampal transitional zone), bed nucleus of the stria terminalis (BST), hypothalamus (Hyp; in particular the preoptic area, paraventricular and dorsomedial hypothalamic nuclei, lateral and posterior hypothalamic areas, and arcuate and ventral premamillary nuclei), and paraventricular thalamic nucleus (PVT). The periaqueductal gray (PAG; particularly its lateral part) and the medial geniculate body (MGB; mostly its medial nucleus as well as the adjacent suprageniculate, posterior intralaminar thalamic, and peripeduncular nuclei) receive significant projections from both the subparafascicular area and the medial paralemniscal nucleus (PL). The medial paralemniscal nucleus also projects to the superior colliculus (SC; mostly its deep layers), inferior colliculus (IC; mostly its external cortex), parabrachial nuclei (PB), periolivary area (POA), and subcoeruleus area (SCA). TIP39, tuberoinfundibular peptide of 39 residues.

effect on TIP39 expression. Gonadectomy at postnatal day 26 (which is before puberty) decreases the postpubertal decline, but does not eliminate it or remove the sexual dimorphism. There is not yet any experimental data addressing the significance of the developmental decline in TIP39 expression, but the timing and gender difference suggest that TIP39 may be involved in functions related to reproduction.

Initial Investigations of TIP39 Function

The anatomical distributions of TIP39 and the PTH2 receptor provide some hints about their possible functions. Some of these possibilities have been investigated by the administration of synthetic TIP39 *in vivo* or its addition to primary cell cultures. It should be emphasized that the number of these studies is small and that the concentration, spatial distribution, and timing of TIP39 delivery in these experiments is nonphysiological. These experiments do provide some intriguing suggestions for possible functions of endogenous TIP39 that can be followed up with other types of experiment.

Effects of TIP39 in Pain Assays

There is intense labeling in the outer layers of the dorsal horn of the spinal cord and the spinal trigeminal tract by an antibody to the PTH2 receptor (Figure 7). Most primary afferent (dorsal root ganglion) neurons concerned with pain perception project to this area. *In situ* hybridization detection of PTH2 receptor mRNA suggests that the PTH2 receptors in this zone are contributed by both a local population of intrinsic spinal cord neurons and a population of dorsal root ganglion neurons. Microinfusion of TIP39 into a mouse paw causes a withdrawal response, consistent with the activation of nociceptive sensory neurons. Although not yet demonstrated, this suggests that PTH2 receptors are present in the peripheral terminals of some sensory neurons. Consistent with this, TIP39 stimulates cAMP accumulation in a dorsal root ganglion-derived hybrid cell line (F11) and causes an increase in cytoplasmic Ca^{2+} concentration in some of the cells in dorsal root ganglion primary cultures. Intrathecal injection of TIP39 causes a scratching, biting, and licking response that is characteristic of pronociceptive peptides such as substance P and CGRP. Intrathecal administration of TIP39 also potentiated the responses in several spinally mediated pain assays, including the tail-flick test, and paw withdrawal from pressure. It did not change the response in the hot plate test. These results suggest that the activation of the PTH2 receptor facilitates some aspects of pain perception. In striking congruence with this, intrathecal administration of an antibody to TIP39 increased the response threshold in the same tests as those in which the sensitivity was increased by TIP39. A simple interpretation of this observation is that TIP39 release is provoked by the stimuli used in these pain assays, that it contributes to the normal response, and that antibody sequestration of TIP39 removes this pro-nociceptive influence. However, only a small number of

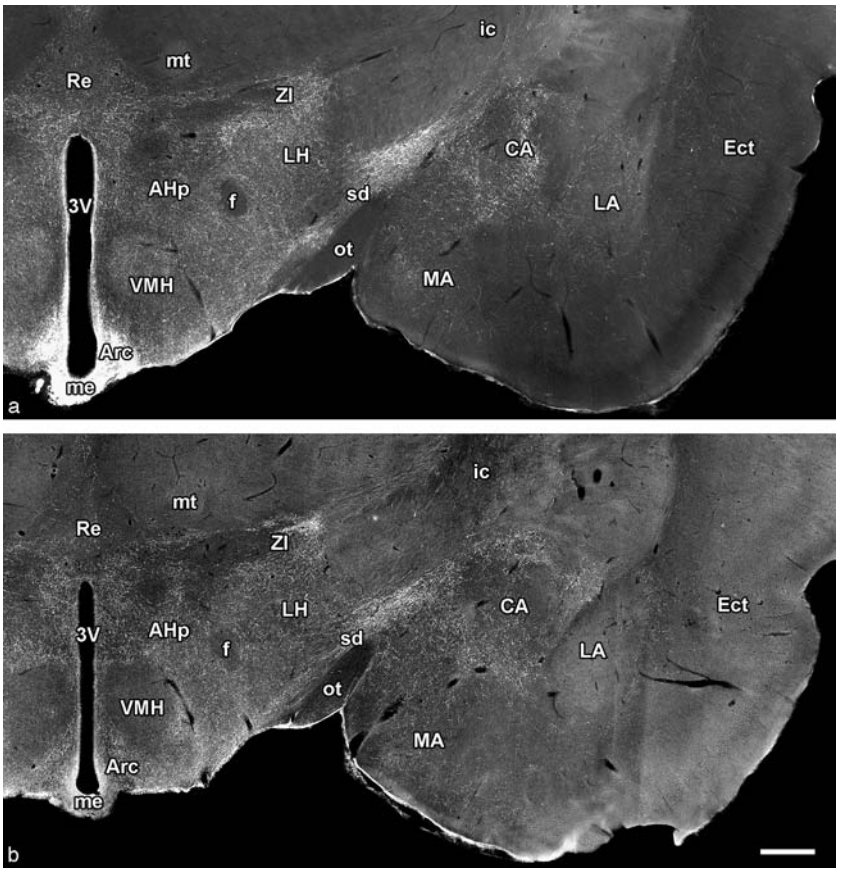

Figure 5 Labeling with antibodies at the level of the tuberal region of the rat hypothalamus in coronal sections: (a) PTH2 receptor distribution; (b) TIP39 distribution. Scale bar = 500 μm. A similar distribution of immunolabeling can be observed in most areas. As shown in (a), PTH2 receptor immunolabeling is more intense in the ventromedial hypothalamic nucleus, arcuate nucleus, and median eminence. The micrograph in (a) is saturated in the arcuate nucleus and the median eminence to demonstrate the labeling in other regions. 3V, third ventricle; AHp, anterior hypothalamic nucleus, posterior part; Arc, arcuate nucleus; CA, central amygdaloid nucleus; Ect, ectorhinal cortex; f, fornix; ic, internal capsule; LA, lateral amygdaloid nucleu; LH, lateral hypothalamic area; MA, medial amygdaloid nucleus; me, median eminence; mt, mamillothalamic tract; ot, optic tract; parathyroid hormone (PTH); sd, paraventricular thalamic nucleus; TIP39, tuberoinfundibular peptide of 39 residues; Re, reuniens nucleus; VMH, ventromedial hypothalamic nucleus; ZI, zona incerta. Modified from Dobolyi A, Irwin S, Wang J, and Usdin T (2006) The distribution and neurochemistry of the parathyroid hormone 2 receptor in the rat hypothalamus. *Neurochemical Research* 31: 227–236.

TIP39-containing nerve fibers have been observed in the spinal cord. Further investigation is required to determine whether TIP39 released from these fibers has a significant physiological role.

PTH2 receptors are also found in several brain areas involved in pain perception, particularly in several midline thalamic nuclei that are thought to be involved in processing the emotional aspects of pain perception. The Place Escape Avoidance Paradigm (PEAP) is a test designed to evaluate an affective (emotional) aspect of nociception. Rats are placed in a chamber with light and dark halves. When they are in the dark half, which they normally prefer, an inflamed hindpaw is stimulated with a blunt probe and when they are in the normally less-preferred light half of the chamber, the noninflamed hindpaw is stimulated. When a drug is administered, its effect on the proportion of time an animal spends in the two halves of the chamber is used as a measure of the drug's effect on the aversive quality of the stimulus. Intracerebroventricular administration of TIP39 decreased the apparent aversiveness of stimulation of an inflamed paw. This effect is opposite in direction to the pro-nociceptive effects of TIP39 observed in spinal-based tests following intrathecal administration of TIP39. Although there is not yet a physiological or mechanistic explanation for these observations, it may be relevant that TIP39-containing fibers in the spinal cord arise from the medial paralemniscal nucleus, whereas subparafascicular TIP39 neurons project to midline thalamic and other brain areas that may be involved in affective aspects of nociception. TIP39 arising from these two areas may have different biological functions.

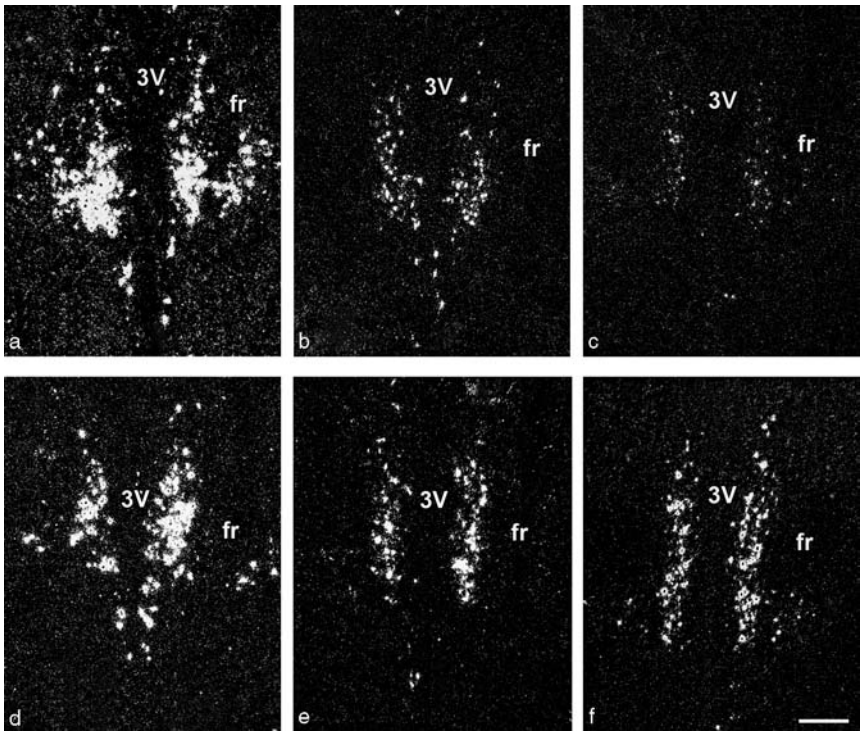

Figure 6 TIP39 expression is age- and gender-dependent: (a) postnatal day (PND) 14, male; (b) PND 57, male; (c) PND 125, male; (d) PND 14, female; (e) PND 57, female; (f) PND 125, female. Scale bar = 200 μm. Dark-field photomicrographs of TIP39 *in situ* hybridization signals in the subparafascicular area demonstrate that TIP39 mRNA levels decrease during postnatal development and are higher in adult females than males. The changes in *in situ* hybridization signal are consistent with data obtained with TIP39 immunolabeling and quantitative reverse transcription polymerase chain reaction (RT-PCR). In addition, similar changes take place in the medial paralemniscal nucleus. 3V, third ventricle; fr, fasciculus retroflexus; TIP39, tuberoinfundibular peptide of 39 residues. Modified from the Dobolyi A, Wang J, Irwin S, and Usdin TB (2006) Postnatal development and gender-dependent expression of TIP39 in the rat brain. *Journal of Comparative Neurology* 498(3): 375–389.

TIP39 Modulation of Hypothalamic Function

PTH2 receptors are present in several medial hypothalamic areas, including the periventricular, paraventricular, arcuate, dorsomedial, and ventromedial nuclei, as well as the median eminence. Several experiments have investigated effects of TIP39 on the secretion of hypothalamic-releasing factors. The results are not entirely consistent, but each of the experiments suggests a role for TIP39 in hypothalamic function. In one experiment, intracerebroventricular administration of TIP39 decreased the episodic increases in serum growth hormone that occur in male rats. The anatomy suggests that this may have occurred via increased somatostatin action on pituitary growth hormone cells following activation of PTH2 receptors that are highly expressed on periventricular somatostatin cells. In an experiment performed by a different group of investigators, the addition of TIP39 to acute hypothalamic slice cultures stimulated release of several hypothalamic hormones (CRH, VIP, gonadotropin-releasing hormone, and GHRH) but not thyrotropin-releasing hormone or somatostatin. The release of CRH was particularly striking in this experiment because the effect of TIP39 was as large as that evoked by depolarization with potassium. Intracerebroventricular administration of TIP39 was reported to increase plasma adrenocorticotropic hormone and leutenizing hormone but to have no effect on growth hormone, prolactin, or thyroid-stimulating hormone. A third group of investigators has provided evidence that brain delivery of TIP39 inhibits vasopressin secretion and that it has a particularly strong effect on dehydration-induced vasopressin release. The differences in the reported effects of TIP39 may result from differences in experimental procedures, such as periodic sampling from individual animals versus single pooled determinations, different dosages, and effects in acute slices versus intact animals.

TIP39 Has Effects in Animal Models of Anxiety and Depression

Areas of PTH2 receptor and TIP39 concentration such as the prefrontal cortex, septum, bed nucleus

Figure 7 PTH2 receptor expression in the lower brain stem and spinal cord: (a) antibody labeling of the PTH2 receptor in the spinal trigeminal tract; (b) section of the spinal cord with an attached dorsal root ganglion. In (a), sensory fibers of the trigeminal nerve that convey information related to pain from the face have their first synapses in the spinal trigeminal tract where gold-colored labeling indicates the presence of PTH2 receptors in nerve fibers. This region is equivalent to the dorsal horn of the spinal cord. In (b), clusters of silver grains from *in situ* hybridization detection of PTH2 receptor mRNA indicate the cells both in the outer part of the dorsal horn of the spinal cord and in the dorsal root ganglia, as well as a few deeper spinal cord cells that synthesize PTH2 receptors. Parathyroid hormone (PTH).

of the stria terminalis, and several amygdaloid nuclei may be relevant to affect and anxiety. Intracerebroventricular delivery of TIP39 had no obvious effect on locomotor activity. It does increase the number of open-arm entries in the elevated plus-maze test, the type of response produced by anxiolytic drugs. It also decreased the amount of time that rats were immobile and increased the amount they spent attempting to escape in the forced swim test, the type of effect that antidepressant drugs produce.

Summary

TIP39 is a neuropeptide synthesized by two highly localized groups of neurons, in the subparafascicular area and the medial paralemniscal nucleus. These cells project to distinct areas distributed throughout the central nervous system. There is strong evidence suggesting that TIP39 is a natural ligand for the PTH2 receptor. The PTH2 receptor is a member of the B Family of GPCRs. Like other members of this family, its activation leads to an increase in intracellular cAMP and sometimes also intracellular Ca^{2+} concentration. The PTH2 receptor is primarily localized on presynaptic nerve terminals of glutamatergic neurons. Thus, the physiological function of TIP39 may be to facilitate glutamate release in the areas to which it projects. Preliminary data hint at roles for TIP39 and the PTH2 receptor in the modulation of some aspects of pain sensation, auditory perception, control of pituitary hormone secretion, mood, anxiety, and reproductive behavior. Major questions remain regarding the stimuli that activate TIP39 neurons and the biological significance of TIP39 action.

See also: Mammalian Neuropeptide Families; Neuroendocrine Peptide Processing; Neuropeptide Release; Neuropeptides in Autonomic Neurons; Neuropeptides: Discovery; Neuropeptides: Pain.

Further Reading

Dobolyi A, Ueda H, Uchida H, Palkovits M, and Usdin TB (2002) Anatomical and physiological evidence for involvement of tuberoinfundibular peptide of 39 residues in nociception. *Proceedings of the National Academy of Sciences of the United States of America* 99(3): 1651–1656.

Dobolyi A, Wang J, Irwin S, and Usdin TB (2006) Postnatal development and gender-dependent expression of TIP39 in the rat brain. *Journal of Comparative Neurology* 498(3): 375–389.

Gardella TJ and Juppner H (2000) Interaction of PTH and PTHrP with their receptors. *Reviews in Endocrine & Metabolic Disorders* 1(4): 317–329.

Hoare SR, Clark JA, and Usdin TB (2000) Molecular determinants of tuberoinfundibular peptide of 39 residues (TIP39) selectivity for the parathyroid hormone-2 (PTH2) receptor: N-terminal truncation of TIP39 reverses PTH2 receptor/PTH1 receptor binding selectivity. *Journal of Biological Chemistry* 275(35): 27274–27283.

Jüppner H, Abou-Samra AB, Freeman M, et al. (1991) A G protein-linked receptor for parathyroid hormone and parathyroid hormone-related peptide. *Science* 254: 1024–1026.

Kronenberg HM, Lanske B, Kovacs CS, et al. (1998) Functional analysis of the PTH/PTHrP network of ligands and receptors. *Recent Progress in Hormone Research* 53: 283–301 (discussion 301–303).

LaBuda CJ, Dobolyi A, and Usdin TB (2004) Tuberoinfundibular peptide of 39 residues produces anxiolytic and antidepressant actions. *Neuroreport* 15(5): 881–885.

Papasani MR, Gensure RC, Yan Y-L., et al. (2004) Identification and characterization of the zebrafish and fugu genes encoding tuberoinfundibular peptide 39. *Endocrinology* 145(11): 5294–5304.

Philbrick WM, Wysolmerski JJ, Galbraith S, et al. (1996) Defining the roles of parathyroid hormone-related protein in normal physiology. *Physiological Reviews* 76(1): 127–173.

Usdin TB, Gruber C, and Bonner TI (1995) Identification and functional expression of a receptor selectively recognizing

parathyroid hormone, the PTH2 receptor. *Journal of Biological Chemistry* 270: 15455–15458.

Usdin TB, Hilton J, Vertesi T, Harta G, Segre G, and Mezey E (1999) Distribution of the parathyroid hormone 2 receptor in rat: Immunolocalization reveals expression by several endocrine cells. *Endocrinology* 140(7): 3363–3371.

Usdin TB, Hoare SRJ, Wang T, Mezey E, and Kowalak JA (1999) TIP39: A new neuropeptide and PTH2 receptor agonist from hypothalamus. *Nature Neuroscience* 2: 941–943.

Usdin TB, Wang T, Hoare SR, Mezey E, and Palkovits M (2000) New members of the parathyroid hormone/parathyroid hormone receptor family: The parathyroid hormone 2 receptor and tuberoinfundibular peptide of 39 residues. *Frontiers in Neuroendocrinology* 21(4): 349–383.

Ward HL, Small CJ, Murphy KG, Kennedy AR, Ghatei MA, and Bloom SR (2001) The actions of tuberoinfundibular peptide on the hypothalamo-pituitary axes. *Endocrinology* 142(8): 3451–3456.

Topographic Maps: Molecular Mechanisms

A C Lin and C E Holt, University of Cambridge, Cambridge, UK

© 2009 Elsevier Ltd. All rights reserved.

Topographic maps are a widespread organizing feature of the central nervous system, especially in early sensory processing, whereby spatial relations between cells are preserved between the projecting set of neurons and the receiving set. That is, neighbors in the projecting set connect to neighbors in the receiving set. The effect is that neurons in higher circuits are arranged in a 'map' representing spatial or frequency information about the external world. As information moves toward higher processing centers, it remains organized topographically, so that, for example, neighboring columns of the visual cortex process information from neighboring areas of the visual world, and neighboring columns of the auditory cortex process sounds of similar frequencies.

Ongoing work has revealed two major themes in topographic map development. First, axon guidance cues and their receptors are expressed in gradients in the projecting and receiving layers, which guide projecting axons into a crude topographic map. Second, this crude map is then refined by activity-dependent remodeling of axonal projections, in which neural activity in the projecting layer leads to the elimination of axonal projections not well correlated with their neighbors. This article focuses mainly on the retinotectal projection because it is by far the best characterized of the topographic maps, but reference is also made to other systems, where the same general principles apply.

The Purpose of Topographic Maps

At first glance, it might be unclear why sensory processing is organized topographically. There is no homunculus 'viewing' a projection of the outside world inside the brain, and it would seem that it is the pattern of connections that matters for information processing, not whether neurons processing neighboring stimuli in the external world are neighbors themselves. However, the topographic organization of sensory processing does have important functional benefits. Much of sensory processing relies on comparing neighboring stimuli, most notably in center-surround receptive fields that allow the visual system to analyze edges and motion. The spatial proximity of neurons that respond to neighboring stimuli makes this kind of processing more efficient, since the transmission of action potentials is costly in both energy and time. In addition, some diffusible cues, like nitric oxide, rely on near-neighbor relations. Some circuits rely directly on axon length for processing; for example, the auditory system calculates interaural time difference using coincidence detectors with the delays inherent in action potential transmission to determine how long the sound from one ear must be delayed to match the timing of the sound from the other ear. In this case, the spatial arrangement of cell bodies relates directly to circuit function.

Perhaps more importantly, topographic mapping is efficient developmentally. An arbitrary spatial arrangement of neurons and their synapses, though theoretically possible, would require every connection to be specified individually, a daunting prospect in an organism with billions of neurons and trillions of synapses but only a few tens of thousands of genes. In contrast, topographic mapping allows axonal targets to be determined with only a few broad gradients and activity-dependent refinement of projections (see below). That is, the mechanism of topographic map formation may itself be the functional rationale behind topographic maps.

Historical Perspective

The topographic organization of sensory processing has been recognized since the nineteenth century, but the mechanism by which maps are established was debated until Roger Sperry's classic experiments in the 1940s and 1950s led to his proposal of the 'chemoaffinity hypothesis.' Using the regenerative capacity of amphibian nervous systems, Sperry severed frog optic nerves and allowed the retinal axons to reinnervate their target, the optic tectum (OT). Even if the optic nerve was artificially scrambled or if the eye was rotated 180°, axons always regrew to the correct targets according to the original topographic map. Strikingly, if the eye was rotated 180°, the frog behaved as if its visual world was upside down, as if the formerly 'top' part of the eye, now on the bottom, reconnected to the 'top' part of the retinotopic map in the tectum. These classical experiments argued for the 'chemoaffinity hypothesis,' that axonal connections are specified by complementary molecular identification tags on the axon and its target. To overcome the problem that there could not be enough molecules to mark each cell individually, Sperry proposed overlapping and orthogonal gradients to mark each cell with its 'latitude and longitude.' This combinatorial code would provide each cell with a unique molecular identity defined by expression levels of each tag. This model has been broadly validated with some

modifications (see below) in the last decade with the identification of these molecular tags, especially the Ephs and ephrins.

Axon Guidance Cues in Topographic Mapping

Ephs and Ephrins

The axon guidance cues most involved in topographic mapping are the ephrins and the Eph receptor tyrosine kinases. The Eph receptors (discovered in an erythropoietin-producing hepatocellular cell line) form the largest subfamily of receptor tyrosine kinases and are divided into two classes, EphA and EphB. Their ligands, the ephrins (Eph-receptor interacting proteins), are similarly divided into two classes, ephrinA and ephrinB; generally, EphAs bind ephrinAs while EphBs bind ephrinBs. EphrinAs are glycosylphosphatidylinositol (GPI)-linked to the membrane, while ephrinBs are transmembrane proteins. In addition to 'forward signaling' from ligand to receptor, ephrins are also capable of 'reverse signaling' from receptor to ligand, most clearly with ephrinBs through their intracellular domain, but potentially also with ephrinAs through an unknown mechanism. Ephrins and Ephs can be both attractive and repulsive guidance cues, but function only when membrane bound, suggesting that they guide axons through contact attraction or repulsion rather than long-range guidance.

Nasal–Temporal Mapping in the Retinotectal Projection

Retinal ganglion cell (RGC) axons project onto the OT in nonmammalian vertebrates (the 'retinotectal' projection), and onto the superior colliculus (SC) in mammals (the 'retinocollicular' projection). In both cases, the projection is topographic, such that nasal axons project posteriorly, and temporal axons project anteriorly (Figure 1). (For simplicity, this article refers to the retinotectal projection, with the understanding that similar principles apply in the retinocollicular projection.) Retinotectal topography can be observed by recording receptive fields of tectal neurons to see which part of the retina they respond to, and by anterograde and retrograde tracing, where a lipophilic dye is injected into the retina or tectum, to see which part of the other layer is connected by RGC axons.

Classic experiments by Bonhoeffer and colleagues in the 1980s using the 'stripe assay' showed that temporal axons were selectively repelled by membranes from the posterior tectum. Retinal axons were allowed to grow on alternating stripes of membranes from anterior and posterior tectum. Temporal axons grew only on anterior stripes, while nasal axons grew indiscriminately (Figure 2). The directed growth of temporal axons was due to repulsion from posterior stripes rather than attraction to anterior stripes, because heat inactivation of posterior membranes alone abolished the striped outgrowth pattern, while heat inactivation of anterior membranes had no effect. These findings confirmed the central principle of the chemoaffinity hypothesis, but with a surprising twist – instead of the prevailing idea of attraction between matching tags on axons and their targets, these findings suggested repulsion between axons and inappropriate targets.

Eventually, ephrinA2 and ephrinA5 were identified as the critical ligands in the tectum, and EphA5 and EphA6 (EphA3 in chick) as the receptors in RGCs. EphA receptors are expressed in a low nasal to high temporal gradient in RGCs, while EphrinAs are expressed in a low anterior to high posterior gradient in the tectum (Figure 1(d)). The same principle holds in another retinal target in mammals, the dorsal lateral geniculate nucleus (dLGN) in the thalamus, where ephrinA2 and ephrinA5 are expressed in a high ventral–lateral–anterior to low dorsal–medial–posterior gradient (see Figure 4(b)). Knocking out or mis-expressing ephrinA2/A5 disrupts retinotectal topography, though some retinotopy remains intact, suggesting that other molecules are involved (see below).

In the tectum, ephrinA expression is thought to be patterned by the transcription factor Engrailed-2, which, like the ephrinAs, is expressed in a low anterior to high posterior gradient. In the retina, the nasal–temporal axis is patterned by the transcription factors BF1 (expressed nasally) and BF2 (expressed temporally), and SOHo1 and GH6 (expressed nasally). SOHo1 and GH6 repress EphA3 expression in chick, suggesting a mechanism by which a low nasal to high temporal gradient of EphA expression is created.

RGCs also express ephrinAs themselves, in a high nasal to low temporal gradient (Figure 1(d)). EphrinA5 has been shown to silence EphA3 responsiveness through *cis*-interactions when the two are co-expressed (i.e., interactions on the same cell, as opposed to *trans*-interactions between ephrinA and EphA on different cells), suggesting that ephrinA expressed on nasal axons could silence what little EphA receptor they do have. Thus, the counter-gradient of ephrinA may enhance the gradient of EphA signaling in RGCs. Also, it has been suggested that ephrinA expressed by early-arriving retinal axons might sharpen the tectal ephrinA gradient, because nasal axonal arbors in the posterior tectum would express high ephrinA on top of tectal ephrinA, while temporal axonal arbors in the anterior tectum would express low ephrinA.

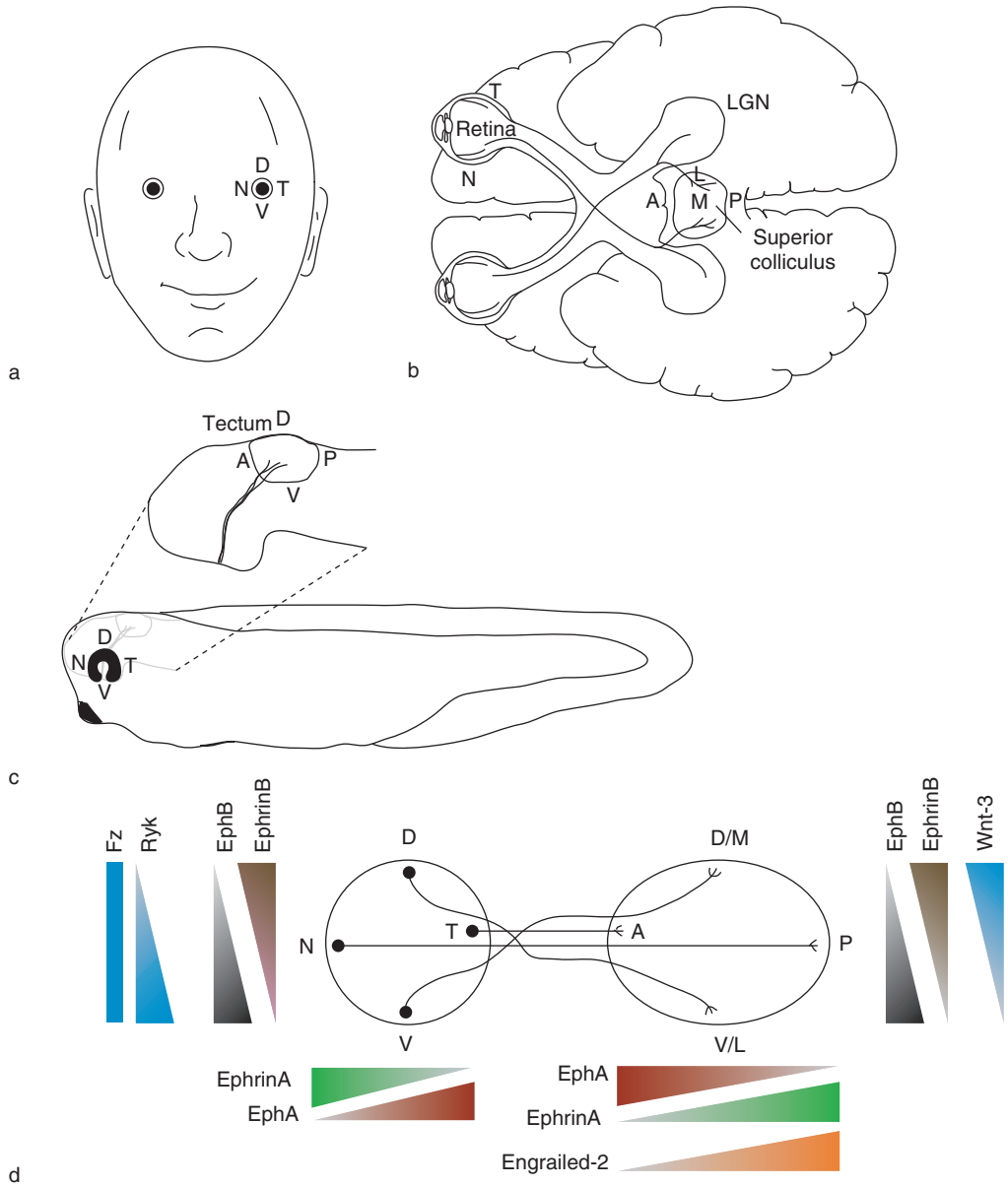

Figure 1 (a) Schematic of nasal–temporal and ventral–dorsal retinal axes. (b) The retinocollicular projection: ventral view of the human brain, showing the retina, lateral geniculate nucleus (LGN), and superior colliculus (SC). (c) The retinotectal projection: lateral view of a *Xenopus* embryo, showing the retina and tectum. (d) Schematic of retinotectal/retinocollicular topography and gradients of guidance cues and their receptors. Dorsal tectum in *Xenopus* corresponds to medial tectum in chick and medial SC in mouse; ventral tectum in *Xenopus* corresponds to lateral tectum/SC in chick and mouse.

Initially, tectal ephrinAs were thought to be purely repulsive. In this model, temporal axons expressing high levels of EphA receptor are repelled by the high levels of ephrinA in the posterior OT/SC and therefore remain in the anterior tectum, while nasal axons expressing low levels of EphA receptor are unaffected by ephrinA and continue to grow into the posterior OT/SC. Axon–axon competition and a hypothetical opposing gradient would explain why all axons did not simply crowd into the anterior tectum. (For a review of computational models of molecular gradients in retinotopic mapping, see 'Further reading.')

However, recent findings suggest that ephrinAs have concentration-dependent biphasic effects. EphrinA2 promotes *in vitro* retinal axon outgrowth at low concentrations and inhibits it at high concentrations, with the tipping point between positive and negative effects for a particular axon varying depending on its nasal–temporal position. Nasal axon growth is promoted by ephrinA2 up to a higher concentration than temporal

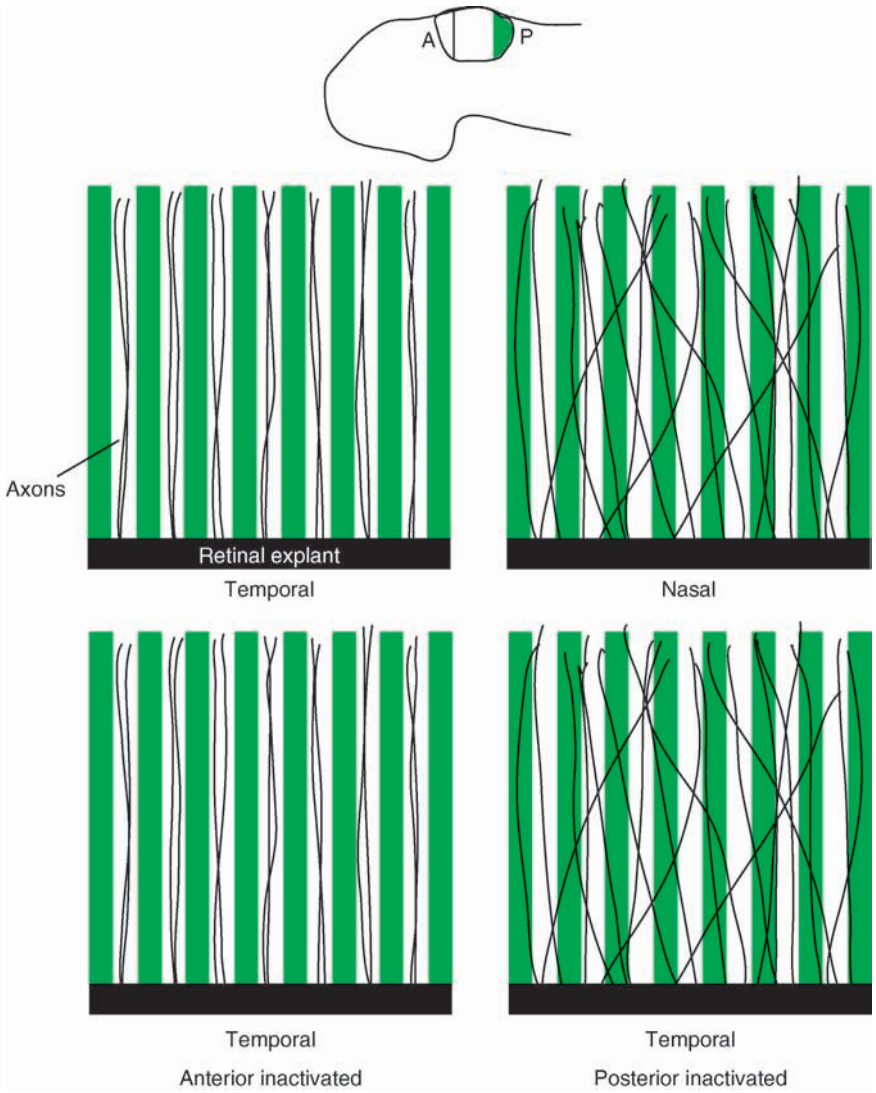

Figure 2 The stripe assay. Membranes from anterior (white) and posterior (green) tectum are laid down in alternating stripes, and retinal pieces are explanted on them so that axons will grow over the stripes. Temporal axons avoid posterior membrane stripes, whereas nasal axons grow indiscriminately. Temporal selectivity is abolished by heat inactivation of posterior membranes, but not anterior membranes.

axons. Each axon therefore terminates around the tipping point between a positive and negative response to ephrinAs. Another molecule mediating biphasic guidance is Engrailed-2. In addition to its role in transcriptional regulation of tectal ephrinA levels, Engrailed-2 attracts nasal axons but repels temporal axons. Unlike ephrins, Engrailed-2 passes directly into growth cones *in vitro* and may function by binding to intracellular signaling molecules rather than receptors; presumably, nasal and temporal growth cones differ in these intracellular molecules.

In frogs and fish, ephrin gradients act by controlling how far posterior the growing RGC axons extend: RGC axons grow 'up' an ephrinA gradient, and depending on how much EphA receptor they express, they stop at a certain zone, which becomes the target. In contrast, in birds and mammals, axons substantially overshoot the target (though still biased toward the target), and the critical step for topographic mapping is selective interstitial branching near the target zone (Figure 3). This difference may be due to species differences in size and developmental timing: frog and fish RGC axons innervate the tectum early, when it is still very small (150–200 μm), and both the retina and tectum continue to grow as the retinotopic map is refined. In contrast, the OT in chick and SC in mouse are relatively large (10 000 μm in chick, 2000 μm in mouse) when RGC axons arrive and do not grow substantially during map refinement. In addition to mediating retinal axon extension, EphAs and ephrinAs are thought to mediate selective interstitial

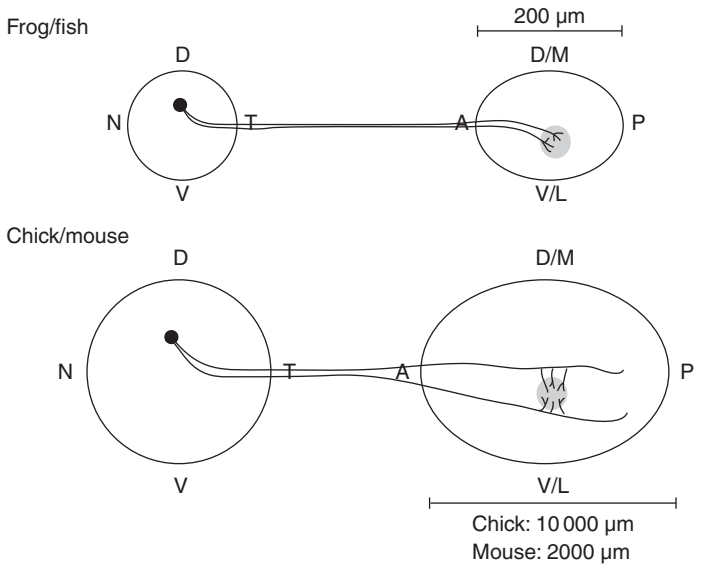

Figure 3 In frogs and fish, axons stop growing near the eventual target zone (gray circle) and arborize there. In contrast, in birds and mammals, axons considerably overshoot the target zone and selectively grow interstitial branches at the target zone. This difference may be due to the vastly different sizes of the tectum/SC between the two cases (bars).

branching by blocking branching posterior to the target zone. As with axon extension, branching anterior to the target may be blocked by axon–axon competition or by ephrinA biphasic action, in which branches need a certain threshold of ephrinA (attractive at low concentrations) to grow.

A straightforward nasal–temporal mapping only applies to organisms with mostly monocular vision, where each retina projects mainly to the contralateral OT. In organisms with binocular vision, such as humans, each retina projects to both sides of the brain. In the right retina, for example, the nasal half projects to the left side of the brain while the temporal half projects to the right side. Each retina therefore needs two topographic maps, and, indeed, the human embryonic retina displays a bidirectional high central to low nasal/temporal gradient of EphA receptors, while the dLGN displays a complementary gradient of ephrinAs (Figure 4). Remarkably, this finding was predicted by Sperry almost half a century ago.

Dorsal–Ventral Mapping in the Retinotectal Projection

Retinal axons also map topographically in the dorsal–ventral dimension. Dorsal RGCs map to the ventral tectum (lateral tectum in chick, SC in mouse), whereas ventral RGCs map to the dorsal tectum (medial tectum/SC). Until recently, dorsal–ventral mapping was poorly understood compared to anterior–posterior mapping, but in the last few years it has been shown that dorsal–ventral mapping is established by ephrinBs and EphBs, as well as Wnt-3 and its receptors Ryk and Frizzled.

EphrinBs are expressed in a high dorsal to low ventral gradient in both the retina and the tectum, while EphBs are expressed in high ventral to low dorsal gradients (Figure 1(d)). Like EphAs, retinal gradients of ephrinBs and EphBs are set up by transcription factors – Vax2 (expressed ventrally) and Tbx5 (expressed dorsally). Ventral Vax2 represses ephrinB expression, to create the high dorsal to low ventral ephrinB gradient, while dorsal Tbx5 represses EphB expression, to create the high ventral to low dorsal EphB gradient.

In frogs, reverse signaling by EphBs in the tectum onto ephrinBs in RGCs mediates dorsal–ventral mapping: ephrinB-expressing dorsal axons are attracted to EphB-expressing cells in the ventral tectum. Ectopic expression of ephrinB in ventral RGCs shifts their tectal projections ventrally, while expression of dominant-negative ephrinB in dorsal RGCs shifts their projections dorsally. Chick and mouse dorsal–ventral mapping appears to be achieved by forward signaling from tectal ephrinBs to RGC EphBs: interstitial branches of EphB-expressing ventral axons are attracted to ephrinB-expressing cells in the medial tectum/SC. Knocking out EphB2 and EphB3 in mice causes axonal branches to project too far laterally.

In chick and mouse, the counteracting force that pulls axons toward the lateral tectum/SC is created by a Wnt-3 gradient as well as biphasic action by

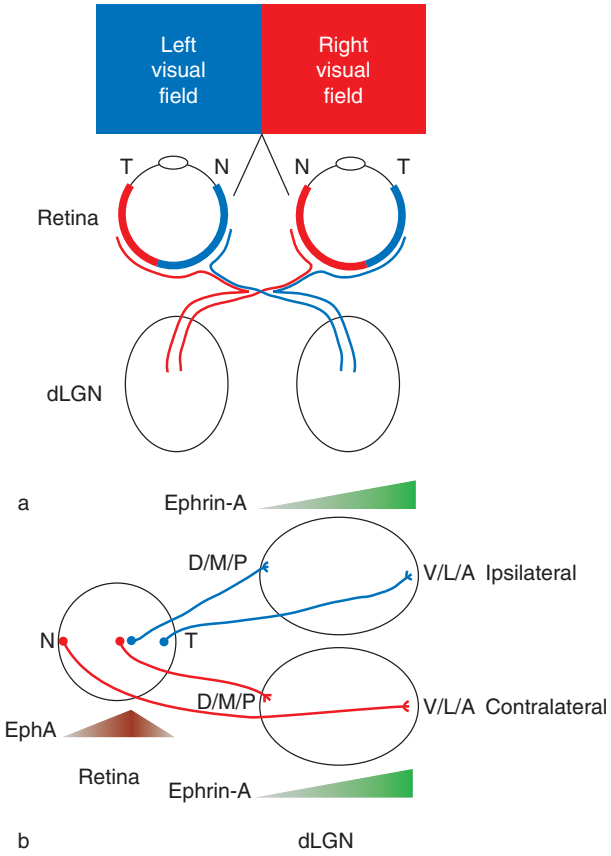

Figure 4 Schematic of retinotopic mapping in organisms with binocular vision. (a) Nasal axons cross at the optic chiasm to the contralateral side of the brain, while temporal axons project to the ipsilateral side, so that each side of the brain processes the contralateral half of the visual field. (b) Topographic mapping between the retina and the dLGN is achieved by a bidirectional high central to low nasal/temporal gradient of EphAs in the retina and a high ventral to low dorsal gradient of EphrinAs in the dLGN.

ephrinB. Wnt-3 is expressed in a high medial to low lateral gradient in the tectum. The receptor Ryk, which mediates repulsive responses to Wnt-3, is expressed in a high ventral to low dorsal gradient in the retina, while the receptor Frizzled, which mediates attractive responses to Wnt-3, is expressed evenly in the retina (Figure 1(d)). The Ryk-mediated repulsive response and Frizzled-mediated attractive responses both drive interstitial branches laterally: high-Ryk-expressing ventral axons are repelled by Wnt-3 while low-Ryk-expressing dorsal axons are attracted to low Wnt-3 levels (such as those in the lateral tectum), though still repelled by high Wnt-3 levels. This lateral-driving effect of Wnt3 counteracts the effect of ephrinB, which attracts axons medially. EphrinB itself may also have biphasic effects; in chick, high levels of ephrinB can be repulsive, indicating that axons from a given DV retinal position have an 'optimum' level of ephrinB that their interstitial branches grow toward.

Guidance Cue Gradients in Other Topographic Maps

Although the role of guidance cue gradients in topographic mapping has been most thoroughly studied in the retinotectal projection, similar principles apply in other sensory projections, including a role for ephrins and Ephs. In the thalamocortical projection, ephrinAs and EphAs play double roles (Figure 5). In early development, EphA receptors are expressed in a high anterior to low posterior gradient in the thalamus, and ephrinAs are expressed in a high posterior to low anterior gradient in the ventral telencephalon. Thalamic axons pass through the ventral telencephalon en route to the cortex, so this initial mapping step sorts them into cortical areas. For example, axons from the anterior thalamus end up in the anterior cortex, in the motor cortex, while axons from more posterior regions of the thalamus end up more posteriorly in the cortex, for example, in somatosensory or visual cortex. EphrinAs and EphAs are then reused later in development; both are expressed in high medial to low lateral gradients, with EphAs in the ventrobasal thalamus and ephrinAs in somatosensory cortex (S1). These complementary gradients arrange a somatotopic map in an analogous way to the retinotectal map: EphA-expressing medial thalamic axons are repelled away from the ephrinA-expressing medial S1 into the lateral S1. Thus, ephrinAs and EphAs are used twice in thalamocortical mapping: first at an intermediate target to mediate 'inter'-areal mapping and then within cortical areas to mediate 'intra'-areal mapping.

EphrinAs and EphAs also play a role in topographic mapping between the hippocampus and septum, and between motor axons and muscles. In both cases, the mechanism is similar to that described for AP retinotectal mapping: high-EphA-expressing axons are repelled from high-ephrinA areas of the target. In the auditory system, functional studies of topography have not yet been done, but there are gradients of EphA4 and ephrinB2 that could form the basis of a tonotopic map in the avian nucleus laminaris, the target of axons from the nucleus magnocellularis, which is in turn the target of cochlear ganglion cells. Interestingly, ephrins have a conserved role in topographic mapping beyond vertebrates: in *Drosophila*, the single Eph and ephrin homologs are required for topographic mapping in the visual system. The reuse of Ephs and ephrins across different projections suggests an evolutionarily efficient modular approach to topographic mapping that would

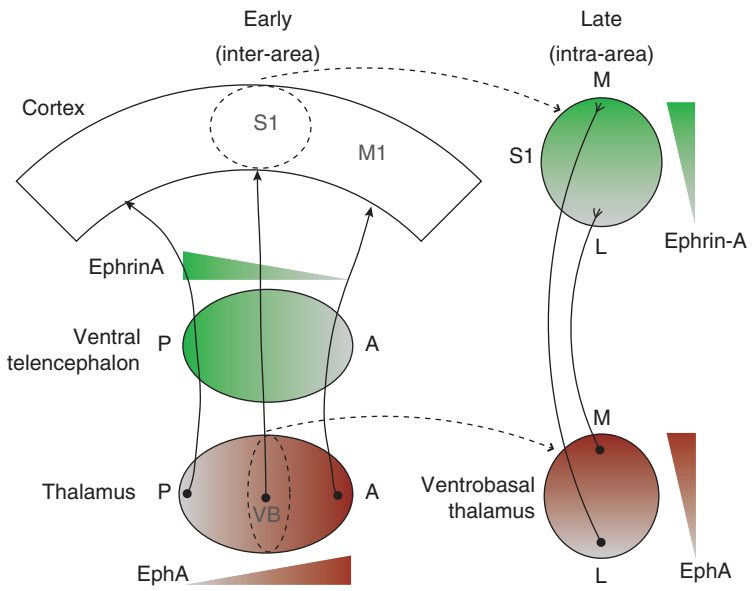

Figure 5 Topography in the thalamocortical projection uses EphA and ephrinA gradients twice. In early development, thalamic axons are sorted toward the correct cortical areas by an ephrinA gradient at an intermediate target, the ventral telencephalon. Later on, thalamic axons are sorted within each cortical area by a second ephrinA gradient.

allow newly evolved projections to be patterned by co-opting a standard package of Eph and ephrin gradients.

Activity-Dependent Refinement of Topographic Maps

Correlated RGC Firing Is Required for Topographic Map Refinement

Molecular gradients are insufficient to establish a refined topographic map, most likely because they act over too long a range to ensure precise mapping at the cellular level. Moreover, in frogs and fish, retinotectal topography shifts throughout life, as the retina grows radially at the periphery while the tectum grows linearly at the posterior edge. Retinotopic map refinement occurs through correlated activity in RGCs, where cells that are close together fire around the same time, while cells that are far apart do not. Correlated retinal firing occurs naturally in vision; neighboring RGCs have neighboring receptive fields, so they are likely to receive, and transmit, similar input. In amphibians and fish, where embryos develop externally, the retina becomes responsive to visual input around the time when axons arrive in the tectum, so normal visual input may suffice to provide the correlated firing patterns needed to refine retinotectal projections. However, in mammals, where embryonic retinas develop in the dark, correlated firing is provided by spontaneous 'retinal waves' of depolarization that move across the retina. These waves have been observed by multielectrode arrays as well as calcium imaging of depolarization. Because they occur every few minutes and the depolarization only lasts for a few seconds, cells that fire together temporally are very likely to be located together spatially.

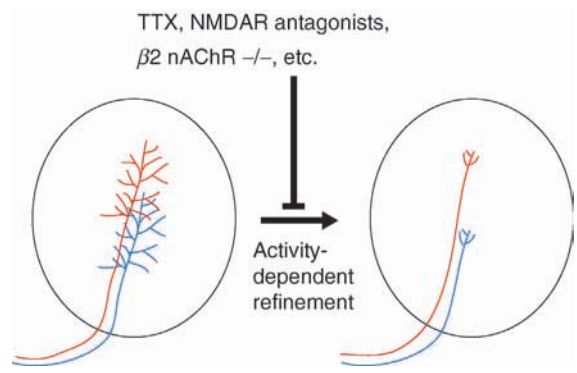

Figure 6 Activity-dependent mechanisms are required to refine initially broad axonal arbors to create precise retinotopy. Refinement can be blocked by interfering with correlated neuronal activity in RGCs, for example, with TTX, NMDAR antagonists, knockout of the $\beta 2$ subunit of the nicotinic ACh receptor, and rearing under stroboscopic conditions.

Eliminating correlated retinal firing blocks the refinement of retinotectal mapping, as measured by the size of axonal arbors or receptive fields (Figure 6). Blockade of neuronal activity in the retina causes abnormally broad axonal arbors and receptive fields

in the tectum, even though the basic shape of the topographic map remains intact, suggesting that molecular gradients suffice for a crude topographic map, but refinement requires activity. This finding holds whether neuronal activity is abolished pharmacologically, by intraocular injection of tetrodotoxin (TTX), or by genetic knockout, for example of voltage-dependent sodium channels. Importantly, it is not electrical activity in RGCs *per se* that matters, but correlated firing patterns. The $\beta 2$ subunit of the nicotinic acetylcholine receptor (nAChR) is required for spontaneous waves of activity in the developing retina, but not for neuronal activity *per se*. Mice lacking $\beta 2$ nAChR exhibit random spontaneous retinal activity instead of retinal waves, and indeed develop abnormally large axonal arbors, just like those of TTX-injected animals. Correlated retinal firing can be inhibited in fish, which lack retinal waves, by rearing them under stroboscopic illumination, which artificially increases correlation of firing between RGCs that are far apart; this treatment, like the $\beta 2$ nAChR knockout, decreases retinotopic refinement.

Additional support for this model comes from artificial ocular dominance bands induced by transplanting extra eyes onto frog embryos, where normally only one eye innervates each tectum. Ectopic and native RGCs innervate the same tectum, and presumably express and encounter the same set of molecular gradients, yet the two projections gradually segregate into exclusive ocular dominance bands reminiscent of those discovered by Hubel and Wiesel in the cat visual cortex. Ectopic RGC innervation is presumably an extreme case of uncorrelated firing, as RGCs from two different eyes would be even more uncorrelated than RGCs far apart in the same eye. Ocular dominance bands are thus an extreme example of activity-dependent refinement of topography, as two sets of axons uncorrelated with each other segregate within a region defined by the coarse topography set up by molecular gradients.

Postsynaptic NMDA Receptors Induce Map Refinement by Eliminating Inappropriate Connections

N-methyl-D-aspartate receptors (NMDARs) are ionotropic glutamate receptors that are blocked by a magnesium ion unless the dendrite is already depolarized, meaning that NMDARs act as 'coincidence detectors' that only activate when more than one synapse fires nearly simultaneously onto the dendrite. In long-term potentiation, Ca^{2+} influx through NMDARs activated by coincident stimulation leads to signaling that strengthens the synapse, and it is thought that a similar mechanism operates in the stabilization of synapses that are 'in sync' with their neighbors and the withdrawal of axonal branches from synapses that are 'out of sync.' Indeed, blocking NMDARs in postsynaptic tectal cells prevents retinotopic refinement of the presynaptic retinal axons. In addition, calcium/calmodulin-dependent kinase II (CaMKII), which is activated by NMDARs, is required in postsynaptic tectal dendrites for pruning presynaptic RGC axonal arbors. Time-lapse imaging in dually innervated frog tecta shows that axonal branch elimination rates, but not addition rates, are higher in areas dominated by axons from the other eye. This bias is NMDAR dependent, suggesting that correlated firing acts through NMDARs to eliminate axonal branches not well correlated with their neighbors.

The nature of the retrograde signal activated by postsynaptic NMDARs to axonal branches, telling them whether to stabilize or retract, remains unclear, although a few candidate mechanisms exist. Nitric oxide (NO) has been suggested as a retrograde messenger since NO synthetase is expressed in tectal cells and NO collapses growth cones *in vitro*. However, blocking NO synthetase appears to slow down, but not ultimately prevent, retinotopic refinement in mouse and chick, while it has no effect on refinement in frogs and fish. Brain-derived neurotrophic factor (BDNF) is produced by tectal neurons where it is released by neuronal activity, and it promotes axonal branching and synapse formation, making it a good candidate for an activity-dependent retrograde messenger. However, the effects on axonal branch dynamics of blocking neuronal activity do not match those of blocking BDNF, suggesting that BDNF may not directly mediate the effects of correlated neuronal activity, but rather modulates them in a complex way that is not fully understood. Work in fish suggests a role for arachidonic acid, a cleavage product of diacylglycerol; both applying arachidonic acid globally and inhibiting its release by postsynaptic cells block retinotopic refinement. Cell adhesion molecules may also serve as activity-dependent retrograde messengers. Applying antibodies against neural cell adhesion molecule (NCAM) or glycoproteins to the tectum inhibits retinotopic map refinement. NCAM, N-cadherin, and glycoprotein L1 are inserted into the membrane in an activity-dependent fashion in other systems, suggesting that they may similarly be inserted by NMDAR activation to aid in axonal branch or synapse stabilization. These candidate mechanisms all seem to converge on interrelated presynaptic signaling pathways involving Ca^{2+}, phospholipase C (PLC), growth-associated protein 43 (GAP-43), Rho GTPases, and others, ultimately influencing the actin cytoskeleton to cause stabilization or retraction of axonal branches.

Conclusions

Our understanding of the molecular mechanisms behind topographic mapping has increased greatly in the last decade. An emerging principle is that topographic mapping efficiently uses a relatively small number of genes to pattern connections among billions of neurons. Matching gradients of axon guidance cues in the target layer and their receptors in the projecting layer set up a coarse topographic map, while activity-dependent mechanisms refine this map to make it more precise by eliminating inappropriate connections. The conservation of the molecular principles of topographic mapping across phyla and across different neural projections suggests that these principles may be a basic method of nervous system organization that evolution has repeatedly turned to. Future work will surely shed further light onto the development of topographic maps.

See also: Axon Guidance: Building Pathways with Molecular Cues in Vertebrate Sensory Systems; Cell Adhesion Molecules at Synapses; Chemoaffinity Hypothesis: Development of Topographic Axonal Projections; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Eph Receptor Signaling and Spine Morphology; Interstitial Axon Branching/Collateral Elimination.

Further Reading

Flanagan JG (2006) Neural map specification by gradients. *Current Opinion in Neurobiology* 16: 59–66.

Goodhill GJ and Xu J (2005) The development of retinotectal maps: A review of models based on molecular gradients. *Network: Computation in Neural Systems* 16: 5–34.

Mann F, Harris WA, and Holt CE (2004) New views on retinal axon development: A navigation guide. *International Journal of Developmental Biology* 48: 957–964.

McLaughlin T, Hindges R, and O'Leary DDM (2003) Regulation of axial patterning of the retina and its topographic mapping in the brain. *Current Opinion in Neurobiology* 13: 57–69.

McLaughlin T and O'Leary DDM (2005) Molecular gradients and development of retinotopic maps. *Annual Review of Neuroscience* 28: 327–355.

Ruthazer ES and Cline HT (2004) Insights into activity-dependent map formation from the retinotectal system: A middle-of-the-brain perspective. *Journal of Neurobiology* 59: 134–146.

Schmidt JT (2004) Activity-driven sharpening of the retinotectal projection: The search for retrograde synaptic signaling pathways. *Journal of Neurobiology* 59: 114–133.

Sperry RW (1963) Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proceedings of the National Academy of Sciences of the United States of America* 50: 703–710.

Torborg TL and Feller MB (2005) Spontaneous patterned retinal activity and the refinement of retinal projections. *Progress in Neurobiology* 76: 213–235.

Vanderhaeghen P and Polleux F (2004) Developmental mechanisms patterning thalamocortical projections: Intrinsic, extrinsic and in between. *Trends in Neurosciences* 27: 384–391.

Torsion Dystonia

X O Breakefield, Harvard Medical School, Boston, MA, USA
D G Standaert, University of Alabama at Birmingham, Birmingham, AL, USA

© 2009 Elsevier Ltd. All rights reserved.

Dystonia Is a Movement Disorder Caused by Abnormal Neural Circuitry in the Brain

Dystonia is the third most common movement disorder in humans, with essential tremor being the most common and Parkinson's disease the second most common. Dystonia is a syndrome characterized by sustained contractions of skeletal muscles, including simultaneous contractions of agonist and antagonist muscles leading to twisting movements and abnormal postures. These contractions can affect isolated (focal) areas of the body, such as writer's cramp (affecting the hand), blepharospasm (rapid eye blinking), torticollis (turning of the neck), and spastic dysphoria (abnormal speech resulting from laryngeal dystonia). Dystonia may also affect larger regions, involving a segment of the body, such as an arm or leg, or may be generalized involving multiple body regions including the face, head, arms, trunk, and legs. These abnormal movements may be sustained or episodic, but they generally only occur during the waking state and not during sleep. Dystonia can begin at any time during life, although onset in childhood usually is associated with involvement of more body parts and a more severe course. Dystonia beginning in adulthood usually manifests as a focal dystonia, and may be task specific, as in writer's cramp and musician's dystonia. Some hereditary forms of dystonia (dystonia-plus syndromes) include other movement abnormalities, such as myoclonus and parkinsonism.

Dystonias can be caused by a variety of environmental and/or genetic factors including lesions, trauma, drugs, toxins, and metabolic disorders. Most genetic forms of dystonia are inherited as autosomal dominant conditions with reduced penetrance with no apparent neurodegeneration. However, imaging studies reveal abnormal patterns of excitability and architecture in the brains of some mutant gene carriers, as well as in apparently sporadic cases. Cases of injury-induced dystonia implicate the basal ganglia and/or cerebellum as target brain regions, but physiologic abnormalities are apparent at multiple levels in the central nervous system (CNS) in pathways involving sensory input and motor output.

Primary and Secondary Forms and Contribution of Genetic Factors

Dystonia can be secondary to a number of diseases, including Wilson's disease, Huntington's chorea, and Parkinson disease, as well as exposure to toxins such as carbon monoxide, drugs (particularly neuroleptics blocking the D2 receptor), viral infection, head injury, or stroke (especially in the basal ganglia). Primary dystonia is distinguished by a lack of other neurologic symptoms or neuropathology and a high hereditary component. The prevalence of primary dystonia in North America is estimated at 3/10 000, but this is likely to be an underestimate as many cases are misdiagnosed or do not come to medical attention. There is a tenfold higher frequency of dystonia in the Ashkenazic Jewish population due to a founder mutation which causes early onset generalized torsion dystonia. Primary dystonias differ in age of onset, site of symptoms, and presence of other movement abnormalities. The majority of patients with primary dystonia have adult onset, focal dystonias with no definable etiologic cause, although there is assumed to be a genetic susceptibility component. Fifteen genes underlying dystonia have been classified based on familial patterns of inheritance and distinctive age and site of onset, distribution of symptoms, and progression of disease with six considered to be pure dystonia (with no other symptoms with the exception of tremor) and the rest as dystonia-plus syndromes (with other movement abnormalities).

The prototypical pure dystonia is early onset generalized torsion dystonia (DYT1). This disorder appears early in life (1–28 years of age, mean 13 years with 30–40% penetrance) and often progresses to involve multiple body regions. In younger individuals, symptoms usually start in a lower limb and progress over a 5-year period to involve many parts of the body, often leading to severe disability. Later onset of DYT1 dystonia is more likely to be limited to upper body parts with little to no cranial involvement. There is a window of susceptibility for manifestations, with individuals who have not onset by 28 years usually being dystonia free for life. DYT1 dystonia has a relatively high incidence in the Ashkenazic Jewish population, and both affected and nonaffected carriers have increased frequency of depression and deficits in motor sequence learning tasks.

DYT6 dystonia is another form of primary genetic dystonia that typically becomes apparent in adolescence, has prominent cranial and cervical involvement, and has an increased incidence among Mennonites. Other pure primary dystonias include

whispering dysphonia (DYT4), adult onset focal dystonia (DYT7), and multifocal/segmental dystonia (DYT13). Dystonia-plus syndromes include dopa-responsive dystonia (DRD, previously classified as DYT5), which is an early onset generalized dystonia distinguished by diurnal fluctuations in symptoms, and dramatic response to L-3,4-dihydroxyphenylalanine (L-dopa) therapy. Myoclonus-dystonia (DYT11) includes rapid brief muscle contractures (jerks) in addition to the dystonia with onset in childhood or adolescence usually affecting the arms and upper body parts. Myoclonus-dystonia is frequently accompanied by psychiatric manifestations and in some cases responds favorably to alcohol, with an increased tendency to alcoholism. Rapid onset dystonia (DYT12) typically has sudden onset (within hours to weeks) in adolescence or young adulthood in response to physical or mental stress. Movement abnormalities include spasms affecting the upper limbs, face, and voice, as well as, in some cases, a slowness typical of parkinsonism. Once the symptoms of rapid onset dystonia appear, they are permanent. Paroxysmal dystonia-dyskinesia (DYT8) is an episodic syndrome induced by a number of agents/conditions, such as alcohol, caffeine, hunger, and fatigue. Symptoms begin in childhood or adolescence and include dystonic and choreic movements. One of the few dystonias in which neuronal degeneration is apparent is X-linked dystonia-parkinsonism (DYT3, lubag) which has a high frequency in the Philippines. Onset is usually focal starting anywhere from 12 to 50 years of age, with generalization to other body parts and parkinsonian features in many of the patients over time. Symptoms are associated with progressive degeneration of the basal ganglia.

Genes/Proteins Identified as Causing Dystonia

Six of the DYT genes have been identified and most encode known proteins with apparently unrelated functions or unknown functions. One of the most informative of the genetic forms of dystonia is dopa-responsive dystonia, sometimes termed 'Segawa's syndrome.' These patients typically have a childhood age of onset with generalization and persistence throughout life, but with dramatic improvement of symptoms in response to L-dopa. This disorder was previously classified as 'DYT5,' but with the recognition that several different genetic defects can produce the same syndrome this nomeclature has been superseded by more specific designations. The most common cause of DRD is heterozygosity for an inactivating mutation in the gene for GTP cyclohydrolase (*GCHI*). GTP cyclohydrolase catalyzes the first step in the biosynthesis of tetrahydrobiopterin – the cofactor for tyrosine hydroxylase (the rate limiting enzyme for dopamine syntheses), as well as tryptophan hydroxylase, nitric oxide synthase, and phenylalanine hydroxylase. Rarely, DRD may be caused by homozygous compromising mutations in the gene for tyrosine hydroxylase (*TH*). Both of these defects lead to deficient dopamine synthesis, and in both cases administration of the precursor L-dopa bypasses the defect and produces striking clinical improvement, pointing to the importance of dopaminergic transmission in the pathogenesis of dystonic movements.

A clinically similar syndrome, early onset generalized torsion dystonia (DYT1), is not responsive to L-dopa, but may also involve dopaminergic transmission. Highest levels of the message for the responsible protein, torsinA, are found in the dopaminergic neurons in the substantia nigra, albeit high levels are found in some other neuronal populations as well. TorsinA is expressed in most cells in the body, with highest levels during the early postnatal period. Most cases of DYT1 dystonia are caused by the same mutation, deletion of three base pairs (delGAG) in exon 5 in the gene for torsinA (*DYT1 or TOR1A*) present in the heterozygous state. This deletion results in loss of a glutamic acid residue in the C-terminal region of the protein. TorsinA, and its three homologs, torsinB, torp2, and torp3, are members of a family of chaperone proteins referred to as 'ATPases associated with a variety of cellular activities' (AAA), which usually form six-membered rings and mediate configurational alterations in proteins as part of folding, denaturation, or movement within cells. The mutant form of torsinA is thought to function as a dominant-negative, resulting in reduction of wild-type torsinA activity to critically low levels in some cells. Although the function of torsinA is not known, it is primarily located in lumen of the endoplasmic reticulum and nuclear envelope where it appears to be involved in processing of proteins through the secretory pathway and/or linking the endoplasmic reticulum or nuclear envelope to the cytoskeleton.

Less is known about the proteins implicated in other forms of dystonia. The gene responsible for the DYT11 form of myoclonus dystonia (*SGCE*) encodes epsilon sarcoglycan, a transmembrane glycoprotein in the family of sarcoglycan proteins that function in the dystrophin complex linking the cytoskeleton to the cell membrane and extracellular matrix. DYT11 is inherited in an autosomal dominant fashion and may be caused by a number of different disruptive mutations. Affected individuals are heterozygous for the mutation, inheriting the mutant gene from their father with inactivation of the normal

allele caused by maternal imprinting. Thus, this syndrome is due to a loss of function state, despite the autosomal dominant pattern of inheritance. Most proteins in the sarcoglycan family are expressed predominantly in skeletal muscle and autosomal recessive mutations in other members of this family cause limb-girdle muscular dystrophies. Epsilon sarcoglycan is unique in this family in that it is expressed at high levels in the brain during development around the time of birth, suggesting it may affect the architecture of neuronal circuitry. Recently, the gene for rapid onset dystonia-parkinsonism (DYT12) has been identified as the alpha 3 subunit of the sodium pump responsible for ion transport across the plasma membrane. A number of different missense mutations have been found to act in a dominant manner to deregulate activity of this oligomeric pump. Mutations in another subunit of this pump, the alpha 2 subunit, have been associated with migraines and infantile convulsions. The gene for paroxysmal dystonia-dyskinesia (DYT8), termed the myofibrillogenesis regulator (MR-1), encodes a protein homologous to a detoxifying enzyme and is expressed throughout many regions of the brain. Reduced levels of this enzyme may increase toxicity of the compounds which induce dystonia attacks.

Animal models of dystonia can be divided into genotypic and phenotypic types, which are as heterogeneous as are the human dystonias. The genetic models attempt to replicate the underlying molecular defects which give rise to human disease, while the phenotypic models are the result of spontaneous mutations or neurotoxin administration, and mimic the motor abnormalities seen in humans. Collectively, the animal models of dystonia implicate altered neurotransmission in the basal ganglia and cerebellum, as well as more widespread involvement in sensory/motor circuit pathways in dystonia. It is important to note that no existing model has succeeded in combining both aspects, that is, none of the existing genetic models have a clear dystonic motor phenotype, and none of the phenotypic models result from mechanisms known to cause human dystonia.

A number of genetic models are available for dystonia. The endogenous mouse torsinA gene (*Tor1a*) has been engineered to create a knockout of the gene function and a knock-in of the delGAG mutation. Knockin mice bearing the delGAG in the heterozygous state (a genetic analog of the human disease state) manifest hyperactivity and difficulty in beam walking, but no apparent dystonic posturing, and have abnormal levels of dopamine metabolites in the brain. In cases where the human delGAG gene is overexpressed in transgenic mice, hyperactivity and defects in motor learning are seen, as well as some changes in dopamine metabolites in the striatum. Transgenic mice with high levels of mutant torsinA expression show ubiquitin-positive aggregates in some brain regions suggesting limited neuronal atrophy. In contrast, mice bearing homozygosity for knock-in of delGAG in *Tor1a* or knock-out of *Tor1a* die at birth. The brains of these animals have normal gross morphology, but postmigratory neurons show abnormalities of the nuclear envelope with vesicles in the nucleoplasmic space.

There are several examples of rodent models which have a clear dystonic motor phenotype. *dt/dt* mice manifest an autosomal recessive condition resembling generalized dystonia in humans. This is due to disruptive mutations in the *Bpag1* gene encoding a member of the cytolinker plakin family which bridges actin filaments and microtubules in the cytoskeleton. This syndrome is associated with degenerative changes, primarily in sensory neurons, involving axonal swelling and interference with retrograde and anterograde transport, and progresses to loss of peripheral nerve endings and eventual death of sensory neurons. Another model resembling generalized dystonia is the *dt/dt* rat in which homozygotes for the mutant allele develop generalized dystonia shortly after birth. Although there are no obvious neuropathologic changes, the cerebellum has a small size, reduced rate of complex spiking, and elevated levels of GABA in Purkinje cells. Surgical resection of the cerebellum relieves the clinical symptoms in these animals, suggesting that it is a nexus of abnormal circuitry. Collectively, the observations implicate defective climbing fiber innervation of Purkinje cells. Recently, the rat *dt* gene has been identified as encoding caytaxin, a lipophilic molecule binding protein which is also defective in a hereditary human ataxia syndrome.

Several rodent models are available for episodic dystonia. The *dt* sz hamster carries an autosomal recessive mutation which produces a syndrome similar to generalized paroxysmal, stress-induced dystonia in humans. Dystonia in these animals apparently results from abnormal GABAergic transmission in interneurons in the basal ganglia with attacks involving dopamine overflow and resulting in inhibition of the endopeduncular nucleus (equivalent to the globus pallidus pars interna in humans, a basal ganglia output nucleus). Both deficits in GABA and dystonic episodes resolve with age. The *dt* sz gene has not yet been identified. Mutations in the Cav2.1 calcium channel, which regulates neurotransmitter release, also produce dystonic features in the autosomal recessive mouse mutants, *tottering* and *leaner*. Totterer mice show a background ataxic state with dystonic symptoms being episodic and precipitated by stress,

caffeine, and alcohol with progression to a generalized state. There is no apparent neuronal degeneration in *totterer* animals. In *leaner* mice, dystonia is chronic and generalized with evidence of widespread cerebellar atrophy. In humans, mutations in the *CACNA1A* gene encoding this channel underlie two forms of ataxia as well as hemiplegic migraine.

Dystonic symptoms can also be produced by delivery of toxins to rodents and nonhuman primates. The active component of moldy sugarcane, 3-nitropropionic acid, can cause lesions in the basal ganglia resulting in dystonia in young humans. This agent inhibits mitochondrial respiration and leads to neuronal death, with the striatal neurons being especially sensitive. Rodents and nonhuman primates exposed to peripheral administration of this agent also develop dystonic symptoms proportional to the extent of damage in the striatum. Low dose microinjection of the glutamate receptor agonist kainate into normal mouse cerebellum, but not into Purkinje cell-deficient cerebellum or into the basal ganglia of mice, produces a progressive but temporary generalized dystonic phenotype. Interestingly, dystonia can also be seen in Parkinson patients, 1-methyl-4-phenyl-1,2,3, 6-tetrahydropyridine (MPTP)-exposed humans, and healthy or MPTP-treated nonhuman primates as a consequence of dopamine deficiency, and may occur in these same patients and animal models as a form of dyskinesia resulting from L-dopa administration leading to elevation of dopamine levels in the striatum. This bimodal response to dopamine depletion and replacement emphasizes the point that the dynamics of dopamine release, and not simply the amount present, influence the appearance of dystonia.

Pathophysiology and Imaging

In most types of dystonia, there is no apparent neuropathology, albeit only a few brains have been evaluated to date. Most types of primary dystonia are considered to be a neurofunctional disease resulting from abnormalities in the microcircuitry of the brain and leading to disorganized sensorimotor integration at the level of the spinal cord, brainstem, basal ganglia, cerebellum, and somatosensory motor cortex. Some insight into the affected circuitry has been gained from analysis of transmitter status, electroencephalographic (EEG) and magnetoencephalographic (MEG) studies, as well as functional magnetic resonance imaging (MRI) and positron emission tomography (PET) analyses. Neurophysiologic abnormalities are detected at multiple levels in the CNS in affected individuals and can also be detected in unaffected mutant gene carriers for DYT1 and DYT6 dystonia. Individuals who carry genes for dystonia but lack overt symptoms seem to have intrinsic subtle differences in their brain connectivity (endophenotype), which may have a developmental component, as well as increased plasticity of some neuronal synapses, which continues throughout life. These abnormalities appear to affect primarily the communication between the sensory input and motor output, but these network based studies cannot determine the primary source of disruption within the interconnected circuitry. A network approach may still, however, be useful for therapy, as evidenced by the ability of deep brain electrical stimulation at a single node point in the globus pallidum to normalize abnormalities throughout the network and relieve dystonic movements in many patients even after years of suffering from severe dystonia.

Several different neurotransmitters have been implicated in the abnormal signaling in dystonic brains. A strong case can be made for dopaminergic involvement given that dystonia can be induced by D2 receptor antagonists (tardive dyskinesia and tardive dystonia), PET imaging shows a decrease in D2 binding sites, and a reduction in dopamine levels and ratio to its metabolite 3,4-dihydroxy-phenylacetic acid (DOPAC) have been found in brains of mutant DYT1 carriers. Further, DRD individuals with decreased dopamine synthetic capacity develop dystonia which can be virtually cured by administration of L-dopa, and the highest levels of torsinA in the human brain are found in dopaminergic neurons of the substantia nigra innervating the striatum. Cholinergic involvement through interneurons in the striatum is also implicated by the ability of anticholinergic medication acting on muscarinic receptors to relieve dystonic symptoms. Low GABA levels have been found in the sensorimotor system, involving the spinal cord, brainstem, and cortex, in individuals with hand dystonia which could also account for reduced inhibition of the surround mechanisms leading to overflow activation of neighboring muscles.

The 'mismatch' in dystonia between sensory input and motor output is revealed in a number of ways, for example, peripheral nerve injury can induce dystonia in the affected body part, while cutaneous stimulation can relieve dystonia in an affected body part; in individuals with hand dystonia, the representation of the affected hand in the somatosensory cortex is disorganized and enlarged as compared to normal individuals. When monkeys are overtrained to a specific hand task, they can develop dystonic posturing of the hand, which correlates with larger receptive fields in the corresponding motor cortex and overlapping representations of individual digits. This may be analogous to musician's dystonias where extensive overuse of a set of muscles leads to dystonic symptoms in that body part in a task-specific and stereotypic

manner. EEG and MEG in hand dystonias also shows less segregated representation of the individual digits in the somatosensory projection within the cortex, and transcranial magnetic stimulation (TSM) demonstrates that the area of cortex influencing movements in the affected body part is greatly enlarged.

Overall studies support both an altered basal ganglia circuitry in individuals predisposed to dystonia, as well as an ongoing hypersensitivity to synaptic plasticity. With regard to basal ganglia circuitry, PET and functional MRI show dystonia is associated with abnormal activity in multiple brain regions, including basal ganglia, motor cortex, supplementary motor areas, and cerebellum. In particular, the premotor cortex is more active during movement in dystonia gene carriers and there are abnormalities in basal ganglia and outflow pathways to the sensorimotor cortex and other regions involved in movement. Unaffected DYT1 carriers also show increased metabolic activity in posterior putamen/globus pallidus, cerebellum, and supplementary motor cortex. A similar pattern of abnormal metabolic activation is seen in affected and unaffected DYT1 and DYT6 carriers, which is distinct from the abnormal patterns seen in DRD carriers and in individuals with Parkinson's disease. Diffusion tensor imaging of the microstructure of white matter pathways reveals abnormalities in mutant DYT1 carriers that suggest altered connectivity patterns in the supplemental motor areas. Interestingly unaffected DYT1 carriers show a selective deficit in motor sequence learning, although motor execution is normal. Imaging studies suggest that unaffected carriers shift from normal striatal processing of sensory and motor coordination to a cerebellar processing.

One of the most intriguing aspects of dystonia is the apparent increase in synaptic plasticity, responsible for changing the strength of synapses in response to neuronal activity, processes which contribute to long-term potentiation (LTP) and long-term depression (LTD) involved in learning. This has been revealed by repetitive transcranial magnetic stimulation (rTMS), which can measure changes in excitability of evoked responses in the sensorimotor cortex. In hand dystonia, the premotor cortex and connected motor areas respond more strongly to rTMS conditioning than do the same regions in normal brains. This suggests that neural circuits in premotor cortex may be more readily modifiable by focal transcranial magnetic stimulation (fTMS) (presumably due to increased plasticity). In a paired associative stimulation model, focal hand dystonia patients also show evidence of enhanced plasticity of sensorimotor circuits, with greater facilitation and loss of spatial specificity. Thus certain connections within the brains of individuals predisposed to dystonia may have enhanced plasticity leading to aberrant reinforcement of synaptic connections, with the potential to become self-reinforcing feedback loops.

Therapy

Treatment for dystonia has been limited, but is improving. Patients have been treated with drugs/toxins, neurosurgery, deep brain stimulation, and behavioral modification. Drugs most commonly used include L-dopa, with patients with DRD as well as some patients with focal dystonia, parkinsonism, and spastic paraplegia showing marked improvement for life on this drug. High dose anticholinergics, such as trihexyphenidyl, are commonly used to treat primary and secondary dystonia with 30–50% of individuals demonstrating at least a modest response. This drug is an antimuscarinic agent and may act through modifying striatal cholinergic receptors. Treatment with anticholinergics is accompanied, however, by a range of peripheral side effects as well as CNS symptoms including memory loss and behavioral changes. Another drug commonly used to treat childhood dystonia is the GABA agonist, baclofen. Because of poor transfer across the blood–brain barrier, this drug is sometimes administered intrathecally in severe cases. Again, only a portion of patients respond, and there can be extensive peripheral and CNS side effects. A major improvement in the treatment of focal dystonia symptoms has been the injection of the light chain of botulium toxin (BTX) into affected muscles from where it is taken up by endocytosis into the terminals of innervating motor neurons. The BTX-A serotype cleaves SNAP-25 needed for fusion of synaptic vesicles to the presynaptic membrane and hence blocks acetylcholine release. The BTX-B serotype cleaves synaptobrevin or the vesicular-associated membrane protein, again blocking vesicular release. Recovery of function returns within 3–6 months, at which time the toxin can be reinjected. A small percentage of patients develop antibodies to the toxin which can be overcome by using a different serotype. BTX treatment can be quite effective in focal dystonia, but is of limited utility in segmental or generalized dystonia.

The earliest neurosurgical procedures for dystonia involved ablative thalamic surgery, which frequently had severe side effects, including loss of speech, and with dystonic movements resuming over time. Based on observations that generalized dystonia patients have a reduced rate and irregular electrical discharge from the globus pallidus as compared with normal individuals, newer interventions have included pallidotomy and deep brain stimulation (DBS) of the

globus pallidus. In particular, pallidal DBS has been shown to be especially effective in early onset generalized dystonia due to DYT1 mutations. With DBS, a gradual reduction in dystonic movement is observed which evolves over days to weeks. Improvements in this technology should be forthcoming. Other potential therapeutic strategies such as use of small interfering RNAs to block expression of dominantly acting proteins associated with dystonia and gene replacement for defective proteins using vectors are options for the future.

See also: Animal Models of Parkinson's Disease; Basal Ganglia: Physiological Circuits; Cranial–Cervical Dystonia; Deep Brain Stimulation and Movement Disorder Treatment; Dystonia: Myoclonus–Dystonia; Opsoclonus/Myoclonus and Neuroblastoma; Segawa Dopa Responsive Dystonia; Writer's Cramp.

Further Reading

Asanuma K, Carbon-Correll M, and Eidelberg D (2005) Neuroimaging in human dystonia. *The Journal of Medical Investigation* 52: 272–279.

Berardelli A, Rothwell JC, Hallett M, et al. (1998) The pathophysiology of primary dystonia. *Brain* 121: 1195–1212.

Bhidayasiri R and Tarsy D (2006) Treatment of dystonia. *Expert Review of Neurotherapeutics* 6: 863–886.

Blood AJ, Flaherty AW, Choi JK, et al. (2004) Basal ganglia activity remains elevated after movement in focal hand dystonia. *Annals of Neurology* 55: 744–748.

Byl NN, Merzenich MM, and Jenkins WM (1996) A primate genesis model of focal dystonia and repetitive strain injury. I: Learning-induced dedifferentiation of the representation of the hand in the primary somatosensory cortex in adult monkeys. *Neurology* 47: 508–520.

Dang MT, Yokoi F, McNaught KS, et al. (2005) Generation and characterization of Dyt1 DeltaGAG knock-in mouse as a model for early-onset dystonia. *Experimental Neurology* 196: 452–463.

Goodchild RE, Kim CE, and Dauer WT (2005) Loss of the dystonia-associated protein torsinA selectively disrupts the neuronal nuclear envelope. *Neuron* 48: 923–932.

Hallett M (2004) Dystonia: Abnormal movements result from loss of inhibition. *Advances in Neurology* 94: 1–9.

Klein C, Ozelius, LJ, and Breakefield XO (2007) Genetic evaluation in primary dystonia. In: Stacy M (ed.) *Handbook of Dystonia*. New York: Taylor and Francis.

Naismith TV, Heuser JE, Breakefield XO, and Hanson PI (2004) TorsinA in the nuclear envelope. *Proceedings of the National Academy of Sciences of the United States of America* 101: 7612–7617.

Ozelius LJ, Hewett JW, and Page CE (1997) The early-onset torsion dystonia gene (DYT1) encodes an ATP-binding protein. *Nature Genetics* 9: 27–43.

Quartarone A, Siebner HR, and Rothwell JC (2006) Task-specific hand dystonia: Can too much plasticity be bad for you? *Trends in Neurosciences* 29: 192–199.

Raike RS, Jinnah HA, and Hess EJ (2005) Animal models of generalized dystonia. *NeuroRx* 2: 504–512.

Ziemann U (2004) TMS induced plasticity in human cortex. *Reviews in Neurosciences* 15: 253–266.

Tourette Syndrome as a Neurotransmitter Disorder

C Winter and A Heinz, Universitätsmedizin Berlin, Berlin, Germany

© 2009 Elsevier Ltd. All rights reserved.

Clinical Manifestation

Current estimates suggest prevalence rates of Gilles des la Tourette syndrome (GTS) in school-age children of approximately 1% with a 3:1 male-to-female ratio. According to the *Diagnostic and Statistical Manual of Mental Disorders* (4th edn.; DSM-IV), GTS diagnosis is based on the occurrence of multiple motor and one or more phonic tics for a minimum of 1 year, with a potential tic-free period that is no longer than 3 months and an onset of the disease prior to the age of 18 years. In addition to the expressed tics, patients also often report a variety of sensory and mental states associated with and often preceding the occurrence of the motor tics. With the perception of these intruding mental precursors, motor phenomena are often experienced as an inescapable and counterproductive consequence that is suppressible only to a variable and limited extent. During the duration of the disease, the symptoms change in intensity, complexity, and their domain of expression. After the age of 20, expressed motor tics tend to decline; congruently, cases of complete cessation in early adulthood have been reported. In more than 50% of patients, GTS is associated with psychiatric conditions which may proceed and outlast the primary disease, for example, attention-deficit/hyperactivity disorder, obsessive–compulsive behavior, and learning disorders. These symptoms are not required for the diagnosis of GTS, but they may severely affect the patients' quality of life and cause patients' functional impairment.

The co-occurrence of tics and psychiatric conditions, predominantly obsessive–compulsive disorders, has stimulated the search for a common neuropathological correlate of both syndromes. In this view, genetic studies support the notion of a spectrum disorder that may include simple tics, pure obsessions and compulsions, or GTS. When first described in 1885, GTS was considered hereditary. Nevertheless, for a long time thereafter it was predominantly ascribed to psychogenic processes; in this respect, some psychotherapeutic views described the tics as stereotyped attempts to express sexual wishes and ideas or the result of excessive or suppressed masturbations. However, in the 1960s the beneficial effect of neuroleptic drugs on the symptoms of GTS rendered the biological interpretation of the disease popular again. As a result, research approaches into the neurobiological mechanisms of GTS increased in the second half of the twentieth century, assigning the neurotransmitter dopamine and distinct brain regions and functional circuits a pivotal role in its manifestation.

Etiology

To date, the full etiology of GTS remains largely unknown, although various partly complementary, partly contradictory neurobiological hypotheses for the manifestation of the disease are defined at several levels of neurological integration, depending on the investigatory neurobiological approach (i.e., neuroanatomy, physiology, or biochemistry). Among these, neuroanatomical proposals point to abnormalities within specific portions of the cortico-striato-thalamo-cortical pathway or specific regions within the striatum (e.g., striosomes), whereas partially compatible physiological proposals emphasize the role of disinhibited afferent thalamo-cortical signal processing or impaired inhibition at the cortical level. Biochemical investigations describe abnormalities at the level of the neurotransmitter systems.

The different hypotheses may be integrated within one pathophysiological story that takes into consideration the wide distribution of the classical neurotransmitters within the functional and neuroanatomical circuits which have been implicated in the pathogenesis of GTS. In this view, the orbitofrontal cortex, the basal ganglia system, and the thalamus make up a system that regulates the planning of intentional behavior and is involved in GTS pathogenesis. Altogether, observations point to a disinhibition of the orbitofrontal cortex, the caudate, and the thalamus. Activation of the orbitofrontal cortex has been associated with error detection and the emotional evaluation of unsuccessful behavior strategies, especially in situations in which task outcome is uncontrollable. It has been suggested that the associated stimulation of the head of the caudate, in turn, disinhibits repetitive and stereotypic behavior patterns and may allow further stimuli to enter the orbitofrontal-striato-thalamic circuit. Consequently, the coupled activation of the orbitofrontal cortex and caudate may generate a vicious circle in which orbitofrontal cortex activation is associated with anxiety and other mental states preceding the occurrence of motor tics and a negative evaluation of performance outcome, whereas caudate stimulation may be associated with repetitive behavior patterns and the repeated admittance of stimuli that activate the neural network. GTS may thus resemble complex

repetitive tics because of the disinhibition of stereotypes encoded in the head of the caudate; on the other hand, it may differ from complex pure motor tics due to additional activation of the orbitofrontal cortex, which triggers subjective feelings such as anxiety, obsessive sensations, and the urge to start another attempt at successful motor performance.

General Aspects of Neurotransmitter Systems

Notably, the neurotransmitters involved in the etiopathologically crucial cortico-striato-thalamo-cortical circuitries – dopamine, serotonin, glutamate, γ-aminobutyric acid (GABA), norepinephrine, acetylcholine, and the opioids – have all been ascribed to the neurobiological events underlying GTS. The predominant neurotransmitter and the primary location of the dysregulation, however, still need to be clarified. Because the neurotransmitter systems are not isolated and independently operating functional units, it has been suggested that there are absolute or relative imbalances within interconnected transmitter systems which represent the predominant pathological factors in the manifestation of GTS.

How do these functional imbalances evolve? Dysregulations at the level of neurotransmitters can be the result of: (1) an inadequate production or metabolism as a result of, for example, structural deficit or an enzyme dysfunction; (2) a qualitative or quantitative transporter alterations; or (3) receptor malfunction. Consequently, there is a demand for methods which measure/label all the steps of neurotransmitter production, release, binding, and metabolism (i.e., the neurotransmitters themselves or their precursors) to understand the complex systems which may underlie the pathogenesis of GTS as a neurotransmitter disorder. Therefore, compounds are required that are specific for the enzymes which produce or degrade the neurotransmitter and for certain receptors and transporters. To date, the following methods are available for investigating the neurotransmitter systems: (1) animal lesion and morphological studies correlating region-specific malfunctions with the symptom profile; (2) clinical trials evaluating the functional response to specific medications acting in the respective neurotransmitter system; (3) cerebrospinal fluid (CSF), blood, and urine studies, as well as neurochemical assays on brain tissues analyzing *in vivo* and postmortem the contents of the neurotransmitters, their precursors, their metabolites, and specific enzymes; (4) positron emission tomography (PET)/single photon emission computed tomography (SPECT) studies demonstrating functional states of receptors and transporters, such as binding levels; and (5) genetic studies linking the target genes of the neurotransmitter systems with the clinical diagnosis.

This article reviews current hypotheses about the involvement of neurotransmitters in the manifestation of GTS as part of a dysregulated cortico-striato-thalamo-cortical interaction, taking into account the various investigatory approaches.

Specific Aspects on Neurotransmitter Systems

Dopamine

The neurotransmitter dopamine is central to motivation, reward, emotional responses, and the production of movements, and it constitutes the predominant neurotransmitter in the mesocortical, mesolimbic, and nigrostriatal pathways. As with other extrapyramidal hyperkinetic disorders, nigrostriatal dopaminergic dysfunction has been implicated in the pathogenesis of GTS motor tics.

Functional considerations The striatum constitutes the major input structure of the basal ganglia system for projections from motor and prefrontal cortices. The basal ganglia circuitry is often divided into two major parts: the direct pathway, containing direct inhibitory striatal projections to the basal ganglia output structures (globus pallidus internus and substanti nigra pars reticulata) and the indirect pathway, encompassing two consecutive inhibitory projections to the subthalmic nucleus which, in turn, projects to the output structures. Basal ganglia output structures control the activity of thalamo-cortical afferents. Accordingly, cortical activation of the direct pathway leads to an inhibition of the basal ganglia output structures, which in turn remove their tonic inhibition from the thalamus, consequently disinhibiting the thalamo-cortical projections in a self-perpetuating manner. In contrast, cortical activation of the indirect pathway results in an inhibition of the thalamo-cortical projections. Nigrostriatal dopamine regulates basal ganglia activity via binding to specific dopamine receptors that are variably distributed in the striatum. Historically, two different classes of dopamine receptors (D1 and D2) have been defined on the basis of their coupling to adenylate cyclase. Later, a total of five distinct dopamine receptors were discovered, which have several similarities, allowing a reallocation to the two major subgroups: D1-like receptors (D1A–1D and D5) and D2-like receptors (D2, D3, and D4). These receptors have opposite effects at the molecular level, with the D1-like receptors being excitatory and the D2-like receptors being inhibitory. Still, they often act synergistically when more complex outputs are taken

into account. With respect to the striatal distribution, excitatory Dl dopamine receptors are expressed mainly on striatal neurons of the direct pathway and inhibitory D2 receptors are expressed mainly on striatal neurons of the indirect pathway. As a result, the presence of dopamine supports the movement-releasing direct pathway and a disinhibition of the motor circuitry, whereas the absence of dopamine mainly disinhibits the movement-controlling indirect pathway and promotes an inhibition of thalamo-cortical projections. Pathologically high dopamine levels consequently cause hyperexcitability and disinhibition of cortical motor areas and facilitate motor output.

Hypothesis of phasic-tonic dopamine control In line with these considerations, GTS has long been associated with an excess in nigrostriatal dopaminergic activity and an imbalance of the direct over the indirect basal ganglia circuitry. Indeed, dopamine antagonists have been shown to be beneficial in the treatment of tic disorders. Moreover, mechanisms increasing dopamine neurotransmission (i.e., the administration of the dopamine precursor levodopa (L-DOPA) or dopamine agonists as well the withdrawal of dopamine antagonists) have been shown to induce or exacerbate preexisting tics. Dopaminergic mechanisms involved in the pathways are, however, not unidirectional, and recent studies indicated that dopamine agonists not only induce but can also alleviate otherwise therapy-refractory tics. These discrepancies may convincingly be explained by a hypothesis based on the regulation of dopamine release into subcortical regions via two independent mechanisms: (1) a phasic dopamine release and (2) a sustained background tonic dopamine release regulated by upstream cortical regions. In this respect, tonic dopamine is supposed to exist extracellularly in low concentration and to determine the long-term or homeostatic mechanism. This basal level of dopamine is analyzed when researchers apply biochemical methods such as microdialysis or postmortem high-performance liquid chromatography. Tonic dopamine is further supposed to be under the regulatory control of dopamine D2 autoreceptors. In contrast, phasic dopamine is released extracellularly in response to acute modulations. It can be measured by the use of stimulant challenges. A chronic reduction in dopamine release may, in turn, induce a supersensitivity of the postsynaptic dopamine receptors. A sudden stress-induced increase in phasic dopamine release might thus affect upregulated dopamine receptors and cause stronger effects. Although originally suggested for mesolimbic and mesocortical dopamine release in schizophrenic patients, recent research has shown that early cortical dysfunction may induce phasic striatal dopamine release and affect *in vivo* radioligand binding to dopamine receptors. Motor tics may thus be caused by stress-induced dopaminergic stimulation of the caudate nucleus, resulting in the release of perseverative behavior patterns.

Contemporary data How do clinical, biochemical, and binding studies fit into theses hypotheses? Complementing the assumption of a pathophysiologically relevant reduction in tonic striatal dopamine, clinical and biochemical studies demonstrated reduced levels of the dopamine metabolite homovanillic acid (HVA) in the CSF of untreated GTS patients and the restoration to normal after the administration of very low doses of the neuroleptic haloperidol. Furthermore, some studies of dopamine receptor binding in postmortem tissue showed increased densities of D2 receptor proteins. A study on monozygotic twin pairs discordant for GTS severity showed a positive correlation between D2 dopamine receptor binding to the radioligand (^{123}I)-iodobenzamide in the head of the caudate nucleus and phenotypic severity. In fact, the increased availability of D2 receptors in brain imaging studies may be due to decreased competition with endogenous dopamine, resulting in higher binding of the radioligand to D2 receptors. In addition, in a PET study, increased levels of D2 receptor binding were observed in a subgroup of GTS patients and correlated positively with the severity of vocal tics. Reflecting phasic dopamine hyperactivity, a PET study demonstrated elevated dopamine release in the putamen of drug-free GTS patients compared to healthy controls following intravenous amphetamine administration. Also, a very early study found in five GTS patients that their CSF had an elevated probenecid-induced accumulation of HVA, which is declarative of increased synaptic dopamine release.

In a recent review article, it was suggested that an overactive dopamine transporter system constitutes an important pathophysiological mechanism underlying GTS. It was hypothesized that such a situation could create (1) reduced basal levels of extracellular dopamine, due to increased reuptake; (2) higher concentrations of dopamine in the axon terminal, as reflected by a PET study revealing increased fluorodopa accumulation in the caudate nuclei of GTS subjects; (3) D2 receptor hypersensitivity and upregulation as a result of decreased basal intrasynaptic dopamine; (4) increased phasic dopamine release as a consequence of amphetamine or environmental stimulation; and (5) increased sensitivity to low-dose neuroleptics because there is a reduced amount of tonic dopamine available for the neuroleptic to block.

However, research approaches trying to assign GTS symptoms to dysfunctions on the level of the

dopamine transporters have so far resulted in partly contradictory data. Some authors found increased mean striatal presynaptic-binding levels using SPECT and increased striatal DAT using postmortem histology; in contrast, other investigators using SPECT and PET analysis did not find any differences in striatal DAT binding between GTS patients, with or without antipsychotic therapy, and healthy controls. Equivalently, controversial findings have been reported from investigations of the vesicular monoamine transporter type 2. Explanations for these data include variations in brain imaging techniques (i.e., selection of the reference point) and preexposure to neuroleptics.

Conclusion After we assess all the studies which have been undertaken to investigate the involvement of dopamine in the manifestation of GTS, our hypotheses cannot be consistently confirmed. Some studies (1) showed no changes in or even increased levels of the dopamine metabolite HVA in postmortem tissue or CSF, (2) showed reduced receptor binding rates in GTS patients compared to controls, and (3) questioned the concept of a potential pathophysiologically relevant hypersensitivity of D2 receptors because they demonstrated an absence of tic exacerbation following acute L-DOPA exposure. Also, findings are contradictory on a genetic level. A recent study specifically demonstrated an association between the dopamine receptor D2 gene and GTS, although another study on polymorphic markers at or near the gene loci for all known dopamine receptors and selected enzymes in a large multiple affected pedigree excluded a major role of the candidate genes in the etiology of GTS.

Altogether, to date these data appear inconclusive. This, however, might result from small-focused study designs, inconsistent with respect to pretreatment, age, and disease severity. Thus, further studies are required to draw firm conclusions about the role of dopamine in the manifestation of GTS.

Serotonin

In a previous study of our group, a significant negative correlation between tic severity and serotonin transporter availability in both the midbrain and thalamus was found. A reduction in serotonin reuptake sites may be associated with increased serotonin concentrations in the synapse and has been associated *in vivo* with increased concentrations of the serotonin metabolite 5-hydroxyindoleacetic acid in the CSF. Also, midbrain pathology has been implicated in some previous studies of GTS. In the midbrain, the raphe area is the origin of cortical and subcortical serotonergic projections. Increased serotonin release from the dorsal raphe area may inhibit tonic dopamine release in the striatum via serotonin 5-HT2 receptors while not affecting the stress-induced excitability of dopaminergic neurons. In the thalamus, serotonergic stimulation changes rhythmic burst activity to single-spike firing mode and increases the evoked impulse activity of sensorimotor cortical neurons. Both mechanisms may thus be involved in the pathogenesis of tics by increasing the sensitivity of fronto-striatal-thalamic motor circuits to phasic dopamine release. Interestingly, in this study all tics were not associated with reduced serotonin transporter availability in GTS patients – only vocal tics. Vocal tics in GTS often manifest as coprolalia, which is classified as a compulsive impulse in traditional psychiatric psychopathology. The most compulsion-like complex motor tics were thus associated with dysfunction of serotonin reuptake. This observation links compulsive impulses with a serotonergic dysfunction and agrees well with a wide range of observations linking central serotonergic dysfunction with obsessive–compulsive disorders. Several groups demonstrated that the inhibition of serotonin reuptake by clomipramine or selective serotonin reuptake inhibitors reduces obsessive–compulsive symptoms more effectively than norepinephrine reuptake blockade. Some studies found elevated CSF 5-hydroxyindoleacetic concentrations in patients with obsessive–compulsive disorders. This may seem counterintuitive because the acute application of serotonin reuptake inhibitors increases serotonin synthesis and release. However, several studies observed that chronic treatment with selective serotonin reuptake inhibitors reduces serotonin synthesis and CSF 5-hydroxyindoleacetic acid concentrations, whereas the acute stimulation of serotonin 5-HT2c receptors with *m*-chlorophenypiperazine exacerbated obsessive–compulsive symptoms. Notably, however, not all the investigations on serotonergic neurotransmission in GTS pointed into the same direction. Consequently, in biochemical studies, the level of the serotonin metabolite 5-hydroxyindoleacetic acid has not consistently been found to be reduced in the CSF of GTS patients. In addition, changes may be region-specific because levels of the same metabolite were found to be normal in the cerebral cortex, whereas in basal ganglia regions they were found to be decreased. Plasma tryptophan, whole-blood serotonin, and 24 h excretion of serotonin have also been reported to be reduced in some studies of GTS subjects.

Glutamate and γ-Aminobutyric Acid

To date, there are only limited data on glutamate in the pathogenesis of GTS. Postmortem studies revealed reduced glutamate levels in the globus pallidus and the substantia nigra pars reticulata, and magnetic resonance imaging (MRI) volumetric studies demonstrated

a globus pallidus reduced in size. These findings suggest that reduced glutamatergic output of the subthalamic nucleus is pathophysiologically relevant. In a classic basal ganglia circuitry, the subthalamic nucleus lies within the indirect pathway. Despite its relatively small size, it is often referred to as the driving force of the basal ganglia because its glutamatergic projections to the output structures provide the main excitatory drive within the basal ganglia. The neuronal structure of the subthalamic nucleus suggests that it might serve as an interface where various (parallel) basal ganglia-thalamo-cortical circuits meet. Reduced activation or inhibition of the subthalamic nucleus decreases the overall electrophysiological and metabolic activities within the basal ganglia output structures, globus pallidus internus, substantia nigra pars reticulata, and pedunculopontine nucleus and consequently leads to a disinhibition of thalamo-cortical efferents. In obsessive–compulsive disorders and GTS, thalamo-cortical projections are pathophysiologically overactive and consequently might be associated with decreased levels of GABA. As a result, GABAergic cortical inhibition of incoming signals from the disinhibited thalamus might be impaired. This hypothesis is supported by clinical findings that GABAergic drugs prove beneficial in the treatment of tics. However, glutamate decarboxylase activity was unaltered in postmortem cortex tissue, as were levels of GABA in various brain regions, whole blood, and CSF.

Others

It has been suggested that alterations in norepinephrine are relevant in the manifestation of GTS because some (but not all) studies showed a beneficial effect of the a2-adrenergic receptor agonist clonidine on motor tics. With respect to the opioid system, functional imaging studies revealed decreased binding in the anterior cingulate and increased binding in the cortical areas. Finally, a general underactivity of central cholinergic mechanisms has been implicated in the pathophysiology of GTS; this hypothesis is based on early pharmacological studies and findings that nicotine binding to the nicotine acetylcholine receptors can provide short-term relief from GTS symptoms.

Conclusion

Considerable progress has been made over the last decades with respect to the pathophysiological mechanisms underlying GTS. From a neurobiological point of view, data from different investigatory levels indicate that the basal ganglia-cortico-thalamo-cortical circuits may be involved in the manifestation of both motor tics and complex obsessive–compulsive disorder symptoms. The wide distribution of the major neurotransmitters within the pathophysiologically relevant anatomical and functional circuitries suggests complex interactions between the different neurotransmitter systems and their contribution to the manifestation of GTS. To date, most studies have focused on the dopaminergic system, suggesting that dysregulations of the phasic-tonic dopamine release are pivotal in the manifestation of GTS. Several lines of evidence further indicate that GTS is heterogeneous in nature and suggest the segregation into clinically distinct subgroups. Accordingly, pathophysiological events and therapeutic approaches may be different among the subgroups, a hypothesis that is supported by the differential clinical response to substances acting on the dopamine system; in a significant minority of patients, dopamine antagonists do not appear to have a positive effect on tics, whereas dopamine agonists have been shown to be beneficial in some otherwise therapy-refractory patients. Assuming that there are subgroups may help to interpret the partly contradictory results of the various clinical, biochemical, functional imaging, and genetic studies. Other reasons for the contradictory results may include methodological limitations of study designs, differences in the duration of the disease, and differences in neuroleptic pretreatment.

In conclusion, further studies will be necessary to elucidate the role of the different neurotransmitter systems in the manifestation of GTS and to benefit the search for the most adequate therapy. Multimodal imaging studies and the investigation of specific animal models may be fruitful in further elucidating the complex interaction of neurotransmitter states, functional alteration patterns, and therapeutic responses.

See also: Autonomic Nervous System: General Overview; Dopamine; Dopamine – CNS Pathways and Neurophysiology; Gamma-Aminobutyric Acid (GABA); Glutamate; Obsessive–Compulsive Disorder; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology; Tourette's Disorder.

Further Reading

Adams JR, Troiano AR, and Calne DB (2004) Functional imaging in Tourette's syndrome. *Journal of Neural Transmission* 111: 1495–1506.

Cohen DJ, Goetz C, and Jankovic J (eds.) (2001) *Advances in Neurology, Vol. 85: Tourette Syndrome*. Philadelphia, PA: Lippincott Williams & Wilkins.

Graybiel AM and Canales JJ (2001) The neurobiology of repetitive behaviors: Clues to the neurobiology of Tourette syndrome. *Advances in Neurology* 85: 123–131.

Heinz A (1999) Neurobiological and anthropological aspects of compulsions and rituals. *Pharmacopsychiatry* 32: 223–229.

Hoekstra PJ, Anderson GM, Limburg PC, et al. (2004) Neurobiology and neuroimmunology of Tourette's syndrome: An update. *Cellular and Molecular Life Sciences* 61: 886–898.

Jankovic J (2001) Tourette's syndrome. *New England Journal of Medicine* 345: 1184–1192.

Mink JW (2006) Neurobiology of basal ganglia and Tourette syndrome: Basal ganglia circuits and thalamocortical outputs. *Advances in Neurology* 99: 89–98.

Saka E and Graybiel AM (2003) Pathophysiology of Tourette's syndrome: Striatal pathways revisited. *Brain Development* 25: S15–S19.

Singer HS and Minzer K (2003) Neurobiology of Tourette's syndrome: Concepts of neuroanatomic localization and neurochemical abnormalities. *Brain Development* 25: S70–S84.

Walkup JT, Mink JW, and Hollenbeck PJ (eds.) (2006) *Advances in Neurology, Vol. 99: Tourette Syndrome*. Philadelphia, PA: Lippincott Williams & Wilkins.

Tourette's Disorder

V Gabbay, G S Hirsch, and B J Coffey, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Tourette's disorder (TD) is a childhood onset neuropsychiatric disorder characterized by multiple, waxing and waning motor and vocal tics. It is named after the French neurologist, George Gilles de la Tourette, who in 1885 reported on nine cases which presented with motor and vocal tics, fears, phobias, and 'arithmomania' (counting obsessions).

The disorder often co-occurs with other psychiatric disorders such as attention-deficit/ hyperactivity disorder (ADHD) and obsessive–compulsive disorder (OCD). It is estimated that up to 3% of school-age children meet criteria for TD. Recent studies suggest that while tics may often persist into adulthood, tic severity often declines significantly in adolescence.

The past two decades have been marked by significant progress in the understanding of the clinical phenomenology, epidemiology, and psychiatric comorbidity of TD. However, TD and tic disorders continue to pose many challenges, especially in the areas of pathophysiology, genetics, developmental neuroscience, psychopathology, and treatment.

This article discusses the clinical phenomenology, clinical course, proposed etiological factors that play a role in the development of the disease, and treatment.

Classification and Clinical Phenomenology

What Are Tics?

Tics are sudden, repetitive, stereotyped, involuntary, or semivoluntary motor movements or phonic productions. Tics typically involve one muscle or a group of muscles and may be characterized by their anatomical location, number, frequency, duration, and complexity. Tics can be classified as 'simple' when they involve one muscle or sound. They are considered 'complex' when movements are slower, and more purposeful involving multiple muscle groups or, in the case of complex vocal tics, multiple sounds, words, or phrases.

Examples of simple motor tics are eye blinking, facial grimacing, shoulder shrugging, nose twitching, mouth opening, shoulder rotation, abdominal tensing, and head turning. Examples of complex motor tics are touching objects, jumping, brushing hair, making rude gestures (copropraxia), grabbing one's genitalia, imitating others' gestures (echopraxia), touching objects and self, gyrating, and bending.

Simple phonic tics are usually meaningless sounds or noises such as throat clearing, coughing, sniffing, grunting, blowing, and squeaking. Complex phonic tics include meaningful word productions such as repetitive words or phrases including the repetition of other's words (echolalia) or of one's own words (palilalia). Coprolalia, or shouting of obscenities or other inappropriate words and statements, is actually quite rare.

Tics are experienced as irresistible, but can be suppressed for varying periods of time. Many patients with TD characterize their tics as a voluntary response to an uncomfortable feeling that precedes them. These feelings, or premonitory sensations, have been described by the majority of Tourette's patients, even in children; these sensations can be localized or general, and physical or mental in nature.

While stress and excitement can exacerbate tics, relaxation can also induce increased tics in some individuals.

Diagnosis

The diagnosis of TD is made based on the history and clinical presentation. No specific laboratory tests are confirmatory. There are several classification systems that differ in their diagnostic criteria. According to the Tourette Syndrome Classification Study Group (1993), both multiple motor tics and one or more phonic tics must be present at some time during the illness, although not necessarily concurrently; tics must occur nearly every day, or intermittently throughout a period of more than 1 year, and during this period there must never be a break in symptomatology lasting more than 3 consecutive months; the anatomical location, number, frequency, type, complexity, or severity of tics must change over time; the onset must occur prior to the age of 18 years; involuntary movements and noises cannot be explained by other medical conditions; and motor tics, phonic tics, or both must be witnessed directly by a reliable examiner at some point during the illness or be recorded by videotape or cinematography.

The American Psychiatric Association DSM-IIIR required in addition that tics must be associated with marked distress or significant impairment in social, occupational, or other areas of functioning. While still debated in the scientific arena, this criterion was removed in the latest version of the DSM-TR (2000) and is not included in other classification systems.

Clinical Course

The typical course of TD is characterized by the onset of simple motor tics such as facial, head, or neck tics at about age 6 or 7, followed by a rostral–caudal progression of motor tics over several years. Vocal tics, usually simple such as throat clearing, or sniffing, typically start at age 8 or 9. More complex tics often begin later, as do obsessive–compulsive symptoms, by about age 11 or 12. Tics progress in a fluctuating manner, often described as waxing and waning. Tics tend to stabilize over time, and some patients have lengthy periods during which most or all manifestations diminish or remit. Recent studies indicate a remitting course, in which the worst period of tic severity is between the ages of 7 and 15 years, followed by a significant improvement by mid adolescence.

Epidemiology

Prevalence estimates for TD vary from 1 to 10 per 1000 depending on setting; a common prevalence figure is 1 per 200 for the full spectrum of TD and chronic tic disorders. More recently, prevalence rates have been estimated to be as high as 4% in a community-based study when all tic disorders are included. Studies in both children and adults suggest that males are at least 3–4 times more likely than females to manifest TD.

Psychiatric Comorbidity

Gilles de la Tourette described symptoms of anxiety, obsessions, and phobias in his first nine TD patients. Obsessive–compulsive symptoms, hyperactivity, inattention, anxiety, and aggressive dyscontrol have all been frequently described in association with TD.

OCD is often associated with TD in clinical settings. OCD is a psychiatric disorder manifested by uncontrolled intrusive, repetitive, and distressing thoughts such as unfounded fears (obsessions) or worries, and repetitive behaviors (compulsions) such as checking, hand-washing, or mental acts such as praying, counting, and repeating words. There appears to be a bidirectional relationship between TD and OCD. Twenty to forty percent of TD patients have been reported to meet full criteria for OCD, and up to 90% have been reported to have subthreshold symptoms such as repetitive counting, touching, or symmetry needs. Family studies indicate that OCD is found at a higher rate in close relatives than in controls which further supports the possible link between OCD and TD.

ADHD is another psychiatric disorder that frequently co-occurs in clinical settings, characterized by motoric hyperactivity, distractibility, and impulsivity. As many as half of TD patients present with ADHD symptoms prior to the onset of tics, and 50–75% of TD patients are found to meet criteria for ADHD. This rate is 10–20 times higher than that observed in the general population. The nature of the scientific relationship between ADHD and TD is not firmly established.

Other comorbid psychiatric disorders that have been reported include mood and anxiety disorders, such as major depression, bipolar disorder, separation anxiety disorder, panic disorder, simple and social phobias, and intermittent explosive disorder.

Although the nature of the relationship between TD and the psychiatric comorbid disorders is still not clear, tic, behavioral, cognitive, and emotional dysfunction may represent manifestations of an underlying core disinhibition problem. Often, it is the psychiatric comorbid disorders that are the primary clinical problem when these youth are evaluated in clinical settings.

Disorder Heterogeneity

Another interesting aspect of the etiology is the increasing evidence that TD may not be a unitary disorder. Two individuals can have entirely different presentations, with no overlapping symptoms, and still fulfill diagnostic criteria for TD. Also, a factor analysis of tic symptoms, involving 85 TD probands and information on 29 tic symptoms, presents evidence of four distinct symptom factors or categories: (1) aggressive phenomena (e.g., kicking, temper fits, argumentativeness); (2) purely motor and phonic tic symptoms; (3) compulsive phenomena (e.g., touching of others or objects, repetitive speech, throat clearing); and (4) tapping and absence of grunting. These four categories accounted for 61% of the variance in TD symptom manifestation. Additionally, specific factors were associated with ages of onset and comorbidities and higher rates of ADHD and OCD in relatives. These factors implicate different structural or genetic causations.

Etiology

The pathophysiology and etiology of TD are still unknown. However, studies suggest genetic, developmental, infectious, and neurobiological etiologies involving abnormalities in neurotransmitter systems and neuroanatomical regions, particularly the basal ganglia. Although a specific animal model has not been identified, some parallels may exist with horses that display spontaneous motor tics accompanied by vocalizations such as crib biting, and dogs and cats with acral lick syndrome.

Neuroanatomical Involvement

Neuroimaging studies suggest that the pathogenesis of TD entails abnormalities in corticostriatothalamic pathways. Structural magnetic resonance imaging (MRI) studies have indicated volume or asymmetry abnormalities in the basal ganglia including the caudate, and the lenticular nuclei in subjects with TD compared to controls, as well as loss of the normal asymmetry of the basal ganglia. A longitudinal MRI study documented inverse correlation between tic severity and caudate volume. Additionally, in 10 pairs of monozygotic twins, the right caudate was smaller in the more severely afflicted individuals, supporting evidence for the role of environmental components in addition to genetics.

Neuroimaging studies employing cerebral blood flow and energy (glucose) metabolism parameters (e.g., positron emission tomography (PET) and single photon emission computed tomography (SPECT)) suggest altered activity (increased or decreased and at times unilaterally) in various areas of brain (e.g., frontal and orbital cortex, striatum, putamen) in TD patients compared to controls or across brain regions within TD patients.

In a recent magnetic resonance spectroscopy study which allows the measurement of specific brain metabolites, reduced *N*-acetylaspartate and choline in the left putamen, along with reduced levels of creatine bilaterally in the putamen, were found in patients with TD. In the frontal cortex, TD patients had decreased *N*-acetylaspartate bilaterally, lower creatine and concentration on the right side, and reduced *myo*-inositol on the left side. The authors suggested that neuronal and glial activity in specific brain region may be impaired in TD.

Neurotransmitters

Several neurotransmitters and neuromodulators have been implicated in TD.

The role of dopamine in TD has been widely investigated. Medications such as neuroleptics, which block D2 dopaminergic receptors, are effective in suppressing tics. This is further supported by postmortem studies and neuroimaging studies, which suggested abnormal regulation of presynaptic dopamine release and postsynaptic uptake. A study in monozygotic twins concordant for TD, but discordant for severity, demonstrated abnormalities in dopamine 2 receptors in the caudate area in the more severely afflicted twin, further linking between dopamine in TD clinical presentation.

Another neurotransmitter that is considered to play a role in the etiology of TD is serotonin or 5-hydroxytryptamine-2 (5-HT2). Postmortem studies in patients with TD have identified low levels of serotonin in the brainstem, low levels of glutamate in the globus pallidus, and low levels of cyclic AMP in the cortex. The role of serotonin in TD is further substantiated by the efficacy of medications with serotonergic properties.

Genetics

The familial nature of TD has long been cited in the literature. Twin studies reported 89–94% concordance rates for TD, supporting its basis as a genetically transmitted disorder. Other studies have shown that first-degree relatives of TD patients have a higher percentage of TD, chronic tics, and OCD than normal controls. Another genetic study involving 76 families with a total of 110 pairs of siblings suggests the involvement of 4q and 8p as TD-related genes. A recent study suggests the role of SLITRK1 in a subtype of patients with TD. Additionally, a rare variant of the serotonin transporter polymorphism was identified (Val425) and was linked to OCD recently, but no studies examined its link with TD.

Immunological Link: Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcus

Swedo and colleagues have described a putative subgroup of children with TD and OCD with hypothetical antecedent group A beta hemolytic streptococcus (GABHS) infection who have symptom onset and/or exacerbation precipitated by streptococcus infection (pediatric autoimmune disorders associated with streptococcus or PANDAS). They describe specific diagnostic criteria for PANDAS, including prepubertal onset of tics and/or OCD, sudden, explosive onset and/or exacerbations, and a temporal relationship with symptoms and GABHS. A double-blind, placebo crossover study of prophylactic oral penicillin to attempt to reduce recurrences of tic/OCD exacerbations failed to show any differences between drug and placebo. Another study by the same group of investigators reported that rates of infection and associated tic/OCD exacerbations with prophylactic azithromycin and penicillin over 1 year were reduced compared to the previous baseline year without antibiotics.

PANDAS as a subtype of TD or OCD has not been confirmed or validated by scientific evidence to date. A multisite, prospective case-control (PANDAS and control subjects) 24-month epidemiological study currently underway seeks to clarify the putative relationship between streptococcus bacteria and tic exacerbations in both groups.

Treatment

Overview

Treatment is a challenging task and typically involves a multimodal approach including both pharmacological and behavioral treatments and guidance, education, and support to both patients and their families. Self-esteem at all times should be supported, as most patients with tic disorders will have at least some sense of personal and emotional vulnerability. Early on in the diagnostic process, educating the patient, family members, and school staff as to phenomenology and natural history of TD is an essential component of treatment. Interventions should be tailored to the individual, based on the clinical presentation. As previously noted, psychiatric comorbid disorders such as ADHD and OCD require attention during the initial evaluation phase.

Patients with symptoms that significantly interfere with adaptation to family, school, or work life, peer relationships, or developmental progress should be treated. Patients who suffer significant emotional distress or inadvertently cause significant distress to key persons in their lives may also be candidates for treatment, even when symptoms are relatively mild. Most patients with very mild symptoms need only monitoring, education, guidance, and support. Referral to the national Tourette Syndrome Association (TSA) or to local chapters of the TSA is an excellent way to maintain continuing education and support.

Pharmacotherapy

Several pharmacological agents have demonstrated efficacy in tic reduction. However, these medications may have limited efficacy in some patients, and can be associated with adverse effects.

Alpha-adrenergic receptor agonists The alpha-adrenergic receptor agonists, such as clonidine and guanfacine, are commonly used. These agents are recommended as first-line treatment for most patients with mild–moderate TD. It is suggested that presynaptic noradrenergic effects mediate the observed clinical improvement in motor and vocal tics. In addition, clonidine may also reduce the disinhibition, impulsivity, and hyperactivity often present in young TD patients. Studies have shown inconsistent therapeutic response and delayed onset of action. Sedation is often experienced and impacts on compliance.

Dopamine antagonists: Typical neuroleptics Since the late 1960s, when haloperidol was first introduced as a treatment for TD patients, it has been the only FDA-labeled, formally approved agent for treatment of TD. These medications block D2 (dopamine) receptors in the basal ganglia, and thus reduce tics. These medications are generally effective in tic reduction; however, they are associated with significant side effects such as extrapyramidal symptoms including tardive dyskinesia.

Atypical antipsychotics The newer, atypical neuroleptics are recommended as first line for moderate to severe tics. These agents, such as risperidone, have both potent serotonin 5-HT2 and dopaminergic D2 receptor antagonist properties. Compared to the typical neuroleptics, they have the potential advantage of less extrapyramidal side effects. Several reports have demonstrated efficacy in treating TD without major side effects. Common side effects of neuroleptics include appetite, weight gain, sedation, muscle cramps, and stiffness.

Selective serotonin reuptake inhibitors The selective serotonin reuptake inhibitors (SSRIs), including fluvoxamine, sertraline, and fluoxetine, are agents that block the reuptake of serotonin into the presynaptic terminal. Several studies have reported efficacy of SSRIs in treating TD patients. Nonetheless, other reports suggest that SSRIs may make tics worse.

Since many TD patients have obsessions or compulsions, and neuroleptics and the alpha 2 agonists do not effectively alter these symptoms, the SSRIs may be beneficial. In addition, complex motor tics such as repetitive squatting and touching may be very similar to compulsive rituals. Fluoxetine has received the most investigation in TD, but efficacy likely is relatively comparable among agents. Choice typically rests with consideration of the side effect profiles of the agents. When obsessions or compulsions are of moderate to severe intensity or impair adaptive functioning, SSRIs are indicated.

Benzodiazepines Another class of medications that may hold some promise is the benzodiazepines (e.g., clonazepam). Some studies suggest a reduction in tics independent of anxiolytic effects. This effect may also reflect a reduction in anxiety that, in turn, could reduce tic frequency. In addition, benzodiazepines can be used to treat comorbid anxiety or panic attacks that can also occur in TD patients. Side effects of particular concern in the juvenile population are sedation, cognitive impairment, and disinhibition.

Stimulants There has been a controversy in the past decade about the role of stimulants in TD; however, recent studies have demonstrated that some TD patients with significant ADHD may be candidates

for methylphenidate when no other treatments have been effective. A recent randomized controlled study of clonidine and methylphenidate in ADHD and chronic tics reported that the combination, and each drug, was more effective than placebo in reducing both tics and ADHD symptoms. Increase in tics did not occur in the methylphenidate group to any extent more significantly than in the placebo group. Side effects include insomnia, appetite suppression, and weight loss.

Atomoxetine Atomoxetine (Strattera), a nonstimulant agent, is a selective norepinephrine inhibitor, which was recently approved for treatment of ADHD in children, adolescents, and adults. It has also been studied in youth with ADHD and chronic tics. A recent randomized controlled trial demonstrated that atomoxetine was more effective than placebo in treatment of ADHD, and tics did not worsen compared to placebo. In addition, by some measures, tics reduced on atomoxetine. Therapeutic response, unlike stimulants, takes a few weeks. Side effects include headache, nausea, fatigue, and stomachaches.

Psychotherapeutic Modalities

Behavioral therapies Behavioral paradigms are often indicated in tic-disorder patients. Simple behavioral approaches can be used with the majority of TD patients to 'destimulate' and contain the symptoms. Specific paradigms are indicated for obsessions, compulsions, and possibly some complex motor tics. Relaxation techniques, such as deep breathing, guided imagery, and use of relaxation tapes, can be useful for the anxious or stressed TD patient.

Scientific evidence is growing for the utility of habit reversal therapy (HRT), also known as behavioral intervention for tics. The patient is taught, through developing awareness of premonitory sensations that precede tics, a competing response to the tics. Children as young as 10 years may be candidates for this treatment; parent support is a component of this approach.

Supportive psychotherapy Individual, group, and family therapies are supportive adjuncts to pharmacotherapy. Individual supportive therapy is indicated for patients having difficulty adjusting to the disorder, to their peers or family, or with school or occupational functioning. It is particularly useful when there is evidence of moderate stress or anxiety or another comorbid psychiatric disorder responsive to this form of psychotherapy (e.g., mild depression). It can help to restore or maintain self-esteem and promote mastery and coping.

School and occupational interventions Useful special educational interventions include creating moderate, task-oriented structure in the classroom, preferential seating, one-to-one support, or individualized educational or work plans. Flexibility is a key element in educational and occupational intervention. To promote tic control, there may be a need for optional 'timeouts' from class or work settings. Time limits may have to be extended or eliminated for exams. Some patients also may benefit from writing aides, such as silent typewriters and computers. Silent typewriters allow patients with handwriting difficulties to take notes in class without disturbing others. Adaptive, as contrasted to regular, physical education can be particularly helpful for children with TD. This form of physical education can be arranged through special education departments in local schools. Programs can be tailored to each child's needs, strengths, and weaknesses.

Specific workplace interventions include structured tasks, organization of tasks into smaller units, flexible time limits, and ample physical space.

Surgical Treatment

Stereotaxic neurosurgical techniques have been used in a few cases of severely impaired adults. Deep-brain stimulation has been used in several adult cases with intractable symptoms. However, given the significant medical risks for these procedures, there is no indication for use of these techniques in youth at this time.

Summary

In his 1885 paper, Gilles de la Tourette stated: "Everything is extraordinary in this disease: . . . its symptoms peculiar, its character equivocal, its causes unknown, its treatment problematical." We now know a considerable amount about this disorder that bears his name; it is a complex, neuropsychiatric disorder with a heterogeneous phenotype characterized by core disinhibition mediated by the corticostriatothalamic tracts. TD is genetically determined in most cases. There is a growing appreciation and understanding of the role of psychiatric comorbidity in the diverse clinical presentation. Treatment strategies should be multimodal in nature, and include pharmacotherapy and, where indicated, behavior, individual, family intervention, and psychoeducation.

There is still much to be learned about TD. Further investigation is needed to better understand the phenomenology, developmental psychopathology, and treatment. The concept of TD as a unitary disorder is being questioned; as evidence accrues at a genetic and molecular biology level, this will have significant impact on understanding etiology, as well as

development of treatment options. Pharmacologic treatment has advanced, but most often patients require multiple medications and adverse effects limit its use. Scientific evidence for behavioral treatment and alternative treatments are beginning to emerge.

The complex symptoms of TD and the psychiatric comorbid disorders can impair the functioning of patients in multiple ways and need to be addressed and prioritized. TD continues to pose challenges to those who work with this disorder. Yet, the enormous level of activity on the part of basic science and clinical investigators has potential to reap fruitful rewards for patients and their caregivers alike in the future.

See also: Anxiety Disorders; Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine; Autonomic Nervous System: General Overview; Cognition: Basal Ganglia Role; Deep Brain Stimulation; Dopamine; Neuroleptics; Norepinephrine: Adrenergic Receptors; Obsessive–Compulsive Disorder; Psychotherapeutic Approaches to Psychiatric Disorders; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology.

Further Reading

Abelson JF, Kwan KY, O'Roak BJ, et al. (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* 310: 317–320.

Alsobrook JP and Pauls DL (2002) A factor analysis of tic symptoms in Gilles de la Tourette's syndrome. *American Journal of Psychiatry* 159(2): 291–296.

Bloch MH, Leckman JF, Zhu H, and Peterson BS (2006) Caudate volumes in childhood predict symptom severity in adults with Tourette syndrome. *Neurology* 65: 1253–1258.

Budman CL, Bruun RD, Park KS, Lesser M, and Olson M (2000) Explosive outbursts in children with Tourette's disorder. *Journal of the American Academy of Child and Adolescent Psychiatry* 39(10): 1270–1276.

Coffey BJ, Frazier J, and Chen S (1992) Comorbidity, Tourette syndrome, and anxiety disorders. *Advances in Neurology* 58: 95–104.

Coffey BJ and Shechter RL (2006) Treatment of co-morbid obsessive compulsive disorder, mood, and anxiety disorders. *Advances in Neurology* 99: 208–221.

Deckersbach T, Rauch S, Buhlmann U, and Wilhelm S (2006) Habit reversal versus supportive psychotherapy in Tourette's disorder: A randomized controlled trial and predictors of treatment response. *Behaviour Research and Therapy* 44: 1079–1090.

de la Tourette G (1885) Etude sur une affection nerveuse caracterisee par de l'incoordination motrice, acompagnee d'echolalie et de coprolalie (jumping, latah, myriachit). *Archives of Neurology (Paris)* 9: 158–200.

DeVito TJ, Drost DJ, Pavlosky W, et al. (2005) Brain magnetic resonance spectroscopy in Tourette's disorder. *Journal of American Academy of Child and Adolescent Psychiatry* 44: 1301–1308.

Hu XZ, Lipsky RH, Zhu G, et al. (2006) Serotonin transporter promoter gain-of-function genotypes are linked to obsessive-compulsive disorder. *American Journal of Human Genetics* 78(5): 815–826.

Jackson GM (2006) Tourette's syndrome. *Current Biology* 16(12): R443–R444.

Jankovic J (2001) Tourette's syndrome. *New England Journal of Medicine* 345(16): 1184–1192.

Ludolph AG, Juengligh FD, Libal G, et al. (2006) Grey-matter abnormalities in boys with Tourette syndrome: Magnetic resonance imaging study using optimized voxel-based morphometry. *British Journal of Psychiatry* 188: 484–485.

Muller-Vahl KR, Meyer GJ, Knapp WH, et al. (2005) Serotonin transporter binding in Tourette syndrome. *Neuroscience Letters* 385: 120–125.

Niesler B, Frank B, Hebebrand J, and Rappold G (2005) Serotonin receptor genes *HTR3A* and *HTR*3B are not involved in Gilles de la Tourette syndrome. *Psychiatric Genetics* 15: 303–304.

Piacentini J and Chang S (2005) Habit reversal training for tic disorders in children and adolescents. *Behavior Modification* 29: 803–822.

Sandor P (2003) Pharmacological management of tics in patients with TS. *Journal of Psychosomatic Research* 55: 41–48.

Shavitt RG, Hounie AG, Campos MCR, and Miguel Ec (2006) Tourette's syndrome. *Psychiatric Clinics of North America* 29: 471–486.

Swedo SE, Garvey M, Snider L, Hamilton C, and Leonard HL (2001) The PANDAS subgroup: Recognition and treatment. *CNS Spectrums* 6(5): 419–422.

Verdellen CWJ, Keijsers GPJ, Cath DC, and Hoogduin CAL (2004) Exposure with response prevention versus habit reversal in Tourette's syndrome: A controlled study. *Behaviour Research and Therapy* 42: 501–511.

Trace Monoamines and Receptors in Mammalian CNS

M D Berry, Brandon University, Brandon, MB, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The trace amines are a class of endogenous neuroactive amines primarily comprising 2-phenylethylamine (PE), tryptamine, the tyramines (ortho-, meta-, and para-), the octopamines, and the synephrines, although other family members may also exist. As can be seen from Figure 1, these amines are structurally related to the classical monoamine neurotransmitters dopamine (DA), noradrenaline, and 5-hydroxytryptamine (5-HT), as well as amphetamine-related drugs of abuse.

Trace amines have been found in all species so far examined, ranging from lower invertebrates such as the starfish *Pycnopodia helianthoides*, up to higher mammals, including humans. Although synthetic and degradative strategies are the same in all species, trace amine functioning appears to be distinct, certainly between invertebrate and vertebrate species. Thus, while trace amines, in particular the tyramines and octopamines, have been established as neurotransmitters in a number of invertebrate species, such a neurotransmitter function has not been established in vertebrates. Indeed, in vertebrate species, trace amines have been suggested to regulate the efficacy of coexisting neurotransmitters, rather than serving as true neurotransmitters. The recent discovery and preliminary characterization of a family of putative G-protein-coupled trace amine receptors has lent some support to this hypothesis. Mammalian trace amine receptors comprise a phylogenetically distinct family only distantly related to their invertebrate counterparts.

This article describes the synthesis and degradation of trace amines and distribution of trace amines and their receptors in the mammalian central nervous system (CNS). The current state of understanding of trace amine physiology and pharmacology is presented in the context of the putative trace amine receptors, along with the possible relevance of trace amines to human neurologic and psychiatric disorders.

Trace Amine Synthesis

As would be expected of such closely related structures, the synthesis of trace amines occurs through the same enzymatic mechanisms as for the monoamine neurotransmitters (Figures 2 and 3). 2-Phenylethylamine, the tyramines, and tryptamine are formed by the decarboxylation of the precursor amino acids L-phenylalanine, L-tyrosine, and L-tryptophan, respectively. This reaction is catalyzed by the enzyme aromatic L-amino acid decarboxylase (EC 4.1.1.28; AADC). Tyramines can be converted into the corresponding octopamines by further hydroxylation via the enzyme dopamine-β-hydroxylase. Synephrines can then be formed following N-methylation of octopamines by the enzyme phenylethanolamine-*N*-methyltransferase. There are thus direct parallels between the synthesis of trace amines and the more abundant monoamine neurotransmitters DA, noradrenaline, adrenaline, and 5-HT.

As would be expected, this synthesis occurs primarily within monoaminergic axonal terminals in the CNS. Unlike neurotransmitters such as DA, however, AADC is of prime importance in trace amine synthesis. With respect to DA, noradrenaline, and 5-HT synthesis, AADC is present in a large excess, and is not rate limiting. This has caused considerable confusion, however, since AADC is regulated in response to dopaminergic and noradrenergic receptor activation in regions throughout the CNS. Such regulation does not, however, alter DA or noradrenaline levels. As can be seen from Figures 2 and 3, AADC is, by default, rate limiting with respect to the synthesis of trace amines. No other enzymatic process is present, at least with respect to 2-phenylethylamine and tyramine synthesis. Further, there is no evidence for regulation of either dopamine-β-hydroxylase or phenylethanolamine-*N*-methyltransferase activity.

In addition to synthesis in monoaminergic neurons, trace amines may also be synthesized in the so-called D cells of the CNS. D cells are neurons, which, while being immunopositive for AADC, are immunonegative for tyrosine hydroxylase, tryptophan hydroxylase, and 5-HT. As such, D cells, while not capable of synthesizing DA, noradrenaline, or 5-HT, can synthesize trace amines. The function of D cells in the mammalian CNS is poorly understood, and whether they function as dedicated trace aminergic neurons requires investigation.

With respect to the ortho-, meta-, and para- forms of the tyramines, octopamines, and synephrines, differential synthesis is present. The para-structural isomers of both tyramine and octopamine predominate, with somewhat lower levels of meta-isomers. In both instances, the ortho-isoform is rarely identified conclusively. Few studies have examined for differences in the structural isomers of synephrines. However, given that synephrines are derived directly from octopamines, which themselves are synthesized from tyramines, it seems likely that the para-isoform will again predominate.

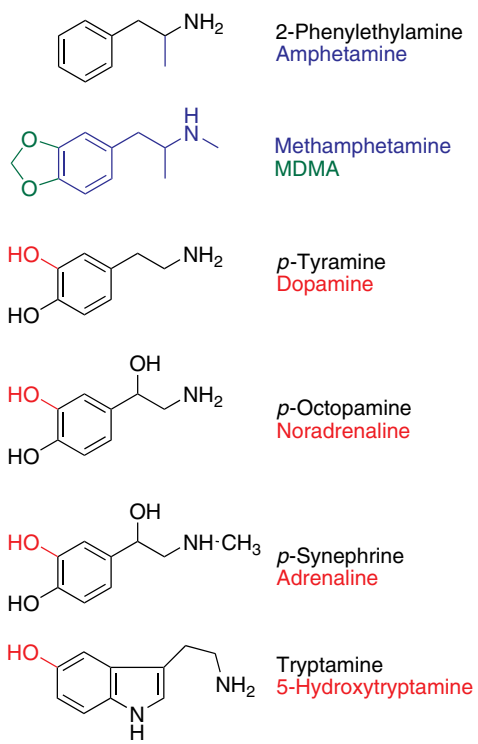

Figure 1 Similarity of trace amines to neurotransmitter monoamines and amphetamine. Structures in black represent the base trace amine structures. Minor additions to these structures give neurotransmitters (red) or amphetamine (blue). Methamphetamine and 3,4-methylenedioxymethamphetamine (MDMA) are also shown for comparison.

A further difference between trace amines and the monoamine neurotransmitters is seen when their tissue levels are compared. While whole brain levels of DA, noradrenaline, and 5-HT approximate 300–600 ng g^{-1} tissue in a number of mammalian species, the levels of the trace amines are at least two orders of magnitude lower, in the region of 1–5 ng g^{-1} of tissue. Notwithstanding such low levels, the trace amines are heterogeneously distributed within the mammalian CNS (see later). Further, their rate of synthesis is equivalent to that of neurotransmitter monoamines. As such, trace amines have an exceedingly rapid turnover rate, in the region of 30 s. Such a turnover is consistent with the apparent lack of storage of trace amines. Finally, in comparison to monoamine neurotransmitters, trace amines show a markedly elevated lipophilicity. Indeed, the trace amines so far studied do not appear to undergo storage and active release via exocytosis. Rather, trace amines appear to be 'released' by simple diffusion across cell membranes. As such, their endogenous levels within the CNS are a reflection of the steady state existing between synthetic and degradative pathways.

Trace Amine Degradation

Trace amine metabolism, with the exception of interconversion, proceeds primarily via oxidative deamination. As with monoamine neurotransmitters, this is

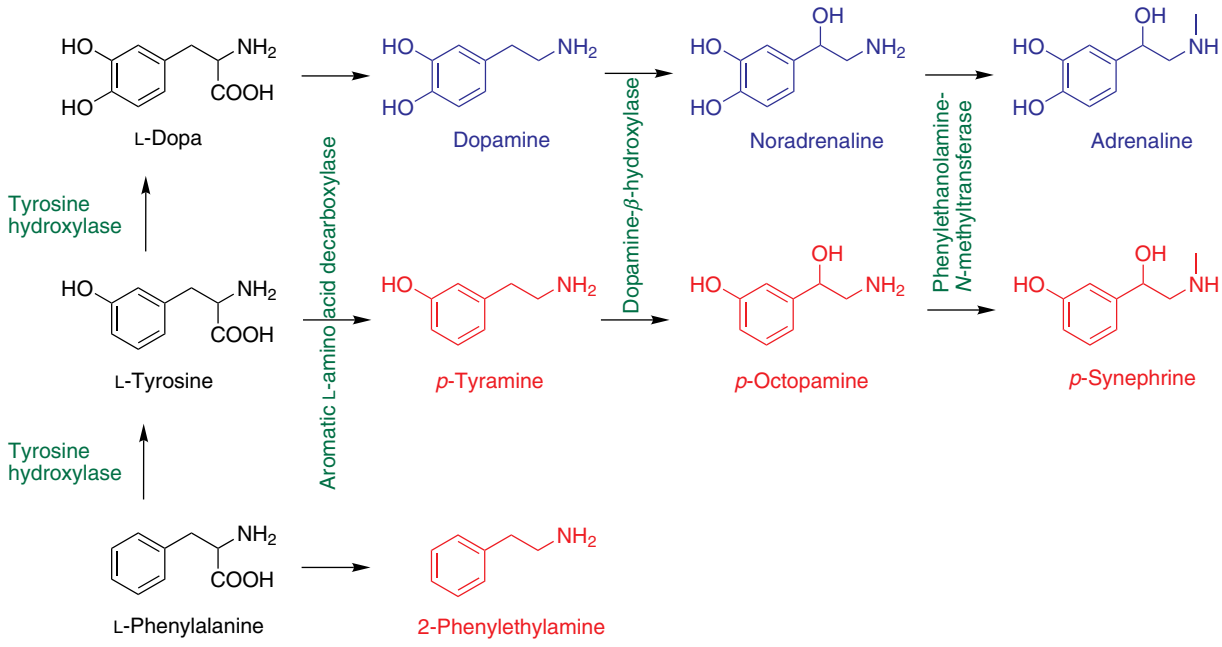

Figure 2 Relationship between trace amine and neurotransmitter synthetic pathways. Precursor amino acids are shown in black, trace amines in red, and neurotransmitter amines in blue. Enzymes common to the synthetic pathways are shown in green.

mediated by monoamine oxidase (EC 1.4.3.4; MAO). In general, the trace amines are rather nonselective substrates for both MAO types A and B. The exception to this is 2-phenylethylamine, which is the prototypical MAO-B selective substrate. Additional metabolic routes, such as deamination through the cytochrome P450 system, and O-methylation via catechol-O-methyltransferase (COMT), are minor fates of trace amine metabolism.

Tissue Distribution

Although trace amines are present in low levels, and readily cross cell membranes, they are heterogeneously distributed throughout the mammalian CNS (Table 1). In the rat, 2-phenylethylamine is particularly enriched in the olfactory tubercles, globus pallidus, and nucleus accumbens, while lower levels are seen in various brain stem regions. The tyramines and tryptamine, in contrast, are also enriched within the caudate putamen, with the cerebellum and cortical regions in addition to the brain stem showing low levels. Octopamine, meanwhile, shows higher levels in the hypothalamus and cerebellum, in comparison to other trace amines.

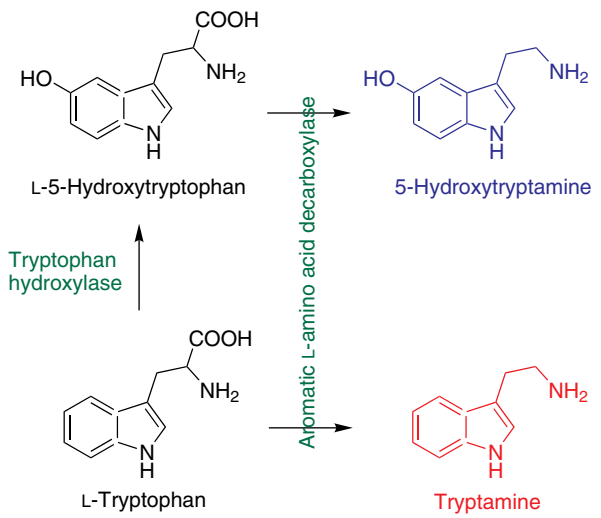

Figure 3 Relationship between tryptamine and 5-HT biosynthesis. Precursor amino acids are shown in black, trace amines in red, and neurotransmitter amines in blue. Enzymes common to the synthetic pathways are shown in green.

Trace Amine Receptors

Although the existence and distribution of invertebrate trace amine receptors is well established, mammalian equivalents were only identified in 2001. Somewhat surprisingly, the mammalian trace amine receptors do not show a close phylogenetic relationship to their invertebrate counterparts, suggesting a possible divergence in the physiological function of trace amines between invertebrates and mammals. The mammalian trace amine receptors form a distinct family of receptor protein termed trace amine-associated receptors (TAARs). A remarkable variability is present between mammalian species with respect to the numbers of receptors present within the genome (Table 2). The human and chimpanzee genomes each contain nine family members, while the rat has 19 members and the mouse, 16. In each instance, this number includes one or more pseudogenes; three in humans, six in chimpanzee, two in rat, and a single pseudogene in mouse. Phylogenetic and ligand-binding pocket analyses have identified three subclasses of TAAR, with each species so far studied containing at least one representative of each subclass (Table 2). All TAAR receptors show seven putative transmembrane domains consistent with the functional receptors being G-protein coupled. Both TAAR1 and TAAR4 appear to couple to the G_s protein, resulting in a stimulation of

Table 1 Regional distribution of trace amines in mammalian CNS

	2-Phenylethylamine	*p-Tyramine*	*Tryptamine*	*p-Octopamine*
Amygdala		+		
Brainstem	+/++	+	−	+++
Caudate putamen	++/+++	++++	++	++++
Cerebellum	++	+	−	+++
Cortex	+/++	+	−	
Globus pallidus	+++	+++	+	
Hippocampus		−		
Hypothalamus	++	+/++	+	+++
Nucleus accumbens	+++/++++	++++	+	+
Olfactory tubercles	++++	++++	+	

Average levels of trace amines in various brain regions of mammalian species.
− = < 0.1 ng g^{-1} tissue; + = 0.1–1 ng g^{-1} tissue; ++ = 1–3 ng g^{-1} tissue; +++ = 3–5 ng g^{-1} tissue; ++++ = >5 ng g^{-1} tissue.
Blank cells indicate conclusive data are not available for the relevant amine in that region.

Table 2 Comparison of trace amine receptors present in mammalian species

	Human	*Chimpanzee*	*Rat*	*Mouse*
Group 1	TAAR1	TAAR1	TAAR1	TAAR1
	TAAR2		TAAR2	TAAR2
			TAAR3	TAAR3
			TAAR4	TAAR4
Group 2	TAAR5	TAAR5	TAAR5	TAAR5
Group 3	TAAR6	TAAR6	TAAR6	TAAR6
	TAAR8		TAAR7a, 7b, 7c, 7d, 7e, 7g, 7h	TAAR7a, 7b, 7d, 7e, 7f
	TAAR9		TAAR8a, 8b, 8c	TAAR8a, 8b, 8c
			TAAR9	TAAR9
Pseudogenes	TAAR3	TAAR2	TAAR7f, 7i	TAAR7c
	TAAR4	TAAR3		
	TAAR7	TAAR4		
		TAAR7		
		TAAR8		
		TAAR9		

Trace amine receptors present in the genome of various mammalian species. Note that the receptor protein has not necessarily been identified.

adenylate cyclase and resultant cellular increase in cAMP levels.

In each species, all TAARs cluster around a narrow chromosomal region. In humans, the nine TAAR genes are located at chromosome 6q23.1, chromosome 1p12 in the rat, and within chromosome 5 in the chimpanzee. All TAARs are coded by a single exon with the exception of TAAR2, where coding is split between two exons. Although TAAR orthologs show the expected high degree of homology, the marked divergence in the number of pseudogenes and receptor subtypes between species of close phylogenetic relationship (e.g., human vs. chimpanzee; rat vs. mouse) has led to suggestions that TAARs represent a rapidly evolving receptor family, involved in adaptation responses to evolutionary events. This raises the intriguing possibility of TAAR representing novel targets for pharmacotherapy of diseases that are uniquely human in nature.

The CNS distribution of the majority of TAARs is unknown, with only TAAR1 studied in any detail so far. The greatest CNS levels of TAAR1 mRNA in humans are found in the amygdala, with lower levels present in the cerebellum, hippocampus, hypothalamus, and medulla. This correlates well with both the distribution of endogenous trace amines, and preliminary reports of TAAR protein distribution. TAAR6 mRNA shows a similar distribution, with highest levels seen in the amygdala and hippocampus, while TAAR8 also appears to be particularly located to the amygdala. In mouse, high levels of TAAR1 mRNA are seen in a number of regions, including the olfactory bulb, piriform cortex, various trigeminal nuclei, and the cerebellum. The thalamus, hippocampus, and hypothalamic and raphe nuclei show lower mRNA levels, whereas the lowest levels are seen in the septum, basal ganglia, and amygdala. Of particular interest, TAAR1 mRNA is associated with the major monoamine neurotransmitter neuron groups within the raphe nucleus (5-HT), locus coeruleus (noradrenaline), and ventral tegmental area (DA).

The cellular localization of expressed TAAR1 and TAAR4 appears to be consistent with the lipophilicity of the potential endogenous ligand trace amines. As such, both receptor subtypes have been shown to be primarily intracellular, with little, if any, expressed protein appearing within the extracellular membrane. In this aspect, it is noteworthy that TAAR1 and -4 are the only TAAR members so far shown to be responsive to trace amines. The cognate endogenous ligand(s) for the additional TAAR members awaits identification.

Receptor Pharmacology

As indicated, the only TAAR members shown to be activated by trace amines are TAAR1 and TAAR4. TAAR1 from all species is activated by low to submicromolar concentrations of 2-phenylethylamine and *p*-tyramine. In addition, rat TAAR1 is similarly stimulated by octopamine and tryptamine. In contrast, mouse and human TAAR1 show a much lower affinity for both octopamine, and, in the case of human TAAR1, tryptamine. Of note, however, there appears to be a marked disparity with respect to affinity assessed by functional assay of cAMP production, and binding affinity (K_i) value obtained from competition experiments. For example, 2-phenylethylamine displacement of tyramine binding at human TAAR1 occurs at concentrations at least one order of magnitude below those required for receptor activation by 2-phenylethylamine. The basis for this disparity is unclear. Whether it represents additional, as yet unidentified, trace amine receptors/binding sites requires further investigation.

In contrast, classical neuroactive biogenic amines are either ineffective at activating TAAR1, or show at best, partial agonist properties. Histamine, noradrenaline, and 5-HT show limited efficacy to activate TAAR1 from a number of species, although 5-HT can act as a partial agonist at the rat TAAR1. Similarly, DA appears to be a partial agonist at TAAR1 from all species examined. Of note, however, the *O*-methyl metabolite of DA, 3-methoxytyramine, shows considerably greater agonistic properties at TAAR1 than does the parent amine. A similar trend is seen with normetanephrine and metanephrine, the *O*-methyl

metabolites of noradrenaline and adrenaline, respectively. The deaminated *O*-methyl metabolites such as homovanillic acid are not, however, ligands at TAAR1. The relevance of such activity of previously assumed inactive metabolites requires further investigation. Two immediate possibilities arise. First, these observations may suggest that TAARs (at least TAAR1) function physiologically to respond to the levels of monoamine neurotransmitter metabolites, and thereby regulate neurotransmission. Alternatively, however, these observations may merely represent the increased lipophilicity of *O*-methyl metabolites providing increased access to the intracellular receptor pool.

Beyond endogenous amines and their derivatives, the pharmacological profile of TAAR1 has generated considerable interest. In particular, a number of drugs of abuse, particularly amphetamine and its derivatives, show full agonist activity. This is of particular interest given the close structural relationship between 2-phenylethylamine and amphetamine (Figure 1). Both amphetamine and its *N*-methyl derivatives methamphetamine and 3,4-methylenedioxymethamphetamine (MDMA; 'ecstasy') show low to submicromolar potency at TAAR1. Somewhat surprisingly, however, very little stereoselectivity was seen with amphetamine and methamphetamine, both R- and S- isomers being approximately equipotent although this maybe species dependent. The most potent amphetamine derivative at TAAR1 was *p*-hydroxyamphetamine, the major metabolite of amphetamine. In addition, the hallucinogenic ergoline *d*-lysergic acid diethylamide (LSD) shows submicromolar affinity as does the endogenous hallucinogen *N*,*N*-dimethyltryptamine. At present, antagonists of TAAR1 have not been identified. Indeed a number of compounds that show antagonist activity at other monoamine receptors/transporters, were found to be agonists at TAAR1. In particular, phentolamine (α-adrenergic antagonist), cyproheptadine (5-HT receptor antagonist), and nomifensine (DA transporter blocker) were effective TAAR1 agonists.

The identification of cognate ligands for TAAR family members is ongoing. A number of possibilities have been raised, including *O*-methyl metabolites of monoamines, thyronamine derivatives of thyroid hormone, imidazolines, and β-carbolines. Of interest in this respect is the previous identification of imidazoline-binding sites, which appeared distinct from known neurotransmitter-related receptors and enzymes. Whether these sites are synonymous with one or more TAAR requires clarification. Imidazoline ligands have been shown to be agonists at invertebrate octopamine receptors, although as discussed above, such activity may not be readily transferable to mammalian TAAR.

Physiological Responses to Trace Amines

Consistent with the distant phylogenetic relationship between invertebrate and vertebrate trace amine receptors, evidence suggests that mammalian and invertebrate trace amines may serve different physiological functions. Invertebrate trace amines, in particular, tyramine and octopamine, are well-characterized neurotransmitter substances. In mammalian species, however, this does not appear to be the case. Iontophoretic application of trace amines to neurons does not directly elicit changes in the membrane potential of the neuron under study. Rather, the trace amines appear to alter neuronal responsivity to coexisting neurotransmitters (Table 3). Specificity exists within these responses. 2-Phenylethylamine has been shown to potentiate neuronal responses to dopaminergic and noradrenergic stimuli, while having no effect on neuronal responses to acetylcholine, 5-HT, and glutamate. Earlier studies suggested that there was also no effect on γ-aminobutyric acid (GABA)-mediated transmission, although more recent evidence suggests that 2-phenylethylamine decreases inhibitory responses mediated by GABA-B receptors. Similarly, tyramines potentiate neuronal responses to DA, noradrenaline, and possibly GABA (at GABA-B receptors), while

Table 3 Interaction of trace amines with neurotransmitters

	2-Phenylethylamine	*p-Tyramine*	*Tryptamine*	*p-Octopamine*
Dopamine	+	+		−
Noradrenaline	+	+		+
5-HT	−	−	+[a]	
Acetylcholine	−		−	
GABA	−/+	−		
Glutamate	−	−		

[a]Inhibitory 5-HT responses are potentiated while excitatory responses remain unaffected or are converted to inhibitory responses in the presence of trace amine.
Blank cells indicate that interactions between the individual neurotransmitter and trace amine have not been investigated.
+ = neurotransmitter response potentiated by trace amine; − = no effect on neurotransmitter response of trace amine.

leaving 5-HT and glutamate responsivity unaffected. Octopamines also potentiate neuronal responses to noradrenaline; however, in this instance, no effect is observed on DA-mediated responses, nor on those of 5-HT. Finally, tryptamines have been shown to potentiate 5-HT responses, while leaving cholinergic responses unaffected. However, this effect of tryptamines appears to be somewhat more complex than that observed with other trace amines. Specifically, inhibitory 5-HT responses have been shown to be potentiated, while excitatory responses remain either unaffected or are reversed to inhibitory responses in the presence of tryptamine.

Although the above-described responses require the presence of coexisting neurotransmitter, endogenous stores of neurotransmitter are not required. This is important, since at higher concentrations, trace amines will act as false transmitters, displacing monoamine neurotransmitters from synaptic vesicles. Such false transmitter responses, however, require concentrations at least two orders of magnitude above basal trace amine concentrations. Further, the modulatory responses of trace amines appear to be mediated at the level of postsynaptic receptors, since responses to synthetic agonists are also potentiated. However, a direct binding of trace amines to neurotransmitter receptors does not occur. The role of TAAR in these modulatory responses of trace amines, while intuitively attractive, has yet to be investigated. Heterodimerization, following trace amine binding, of intracellular TAAR with membrane-inserted neurotransmitter receptors has been hypothesized, but as yet not investigated.

Since the function of trace amines appears to be to modulate the efficacy of coexisting neurotransmitter substances at postsynaptic receptors, and the synthesis of trace amines (AADC activity) is inversely related to stimulation of neurotransmitter receptors, it has been suggested that trace amines may function as a fine-tune mechanism, keeping basal neuronal tone within defined physiological limits. Such a proposal appears consistent with the suggestions that trace amine receptors are involved in adaptive responses, which may vary from species to species. Further, such a function is also consistent with the very high turnover rate of trace amines, allowing a regulation of neurotransmission on a second-by-second basis.

Relevance to Human Pathology

Historically, alterations in trace amine metabolism have been reported in a number of psychiatric disorders, in particular schizophrenia and depression. Although the reported changes in trace amine and/or metabolite levels in various body fluids are indicative of altered trace amine functioning, such studies are difficult to interpret further with respect to any underlying pathology. Although there is currently no definitive link between trace amines and any pathologic condition, evidence is accumulating that trace amines may play a role. The following paragraphs will briefly outline those CNS disorders in which alterations in trace amine functioning is most strongly linked.

Trace amines have often been suggested as a possible etiological factor in schizophrenia. Initially, this was largely based on the close structural relationship between 2-phenylethylamine and amphetamine (**Figure 1**), combined with the well-documented similarity between symptoms of paranoid schizophrenia and amphetamine overdose. Further, in very high, supraphysiological doses, 2-phenylethylamine shows a pharmacological profile similar to that of amphetamine. This link is seemingly strengthened by the observed pharmacological profile of TAAR1, numerous psychoactive/hallucinogenic agents having been shown to be agonists at this receptor. Differences have however been observed in the behavioral phenotypes induced by amphetamine and 2-phenylethylamine. Possibly of greater relevance is the observed chromosomal location of the trace amine receptors, at chromosome 6q23.1. This location coincides with one of the most robustly replicated susceptibility factors identified for schizophrenia, the SCZD5 locus. Specifically this locus appears to correlate with the TAAR6 gene. Conformation of a role of TAAR mutations as a susceptibility factor for schizophrenia will, however, be difficult. The heterogeneity of the disorder is a major complicating factor, since a single cause is unlikely to be the root cause of all schizophrenic disorders. A further genetic link is seen with the observation that a variant form of COMT shows an increased transmission in patients with schizophrenia. As discussed previously, *O*-methyl metabolites of central monoamines appear to be particularly good agonists at TAAR.

Of possible relevance with respect to schizophrenia is the observed similarity between some of the cognitive associated symptoms associated with phenylketonuria and schizophrenia. Phenylketonuria, a classical inborn error of metabolism, is due to a deficiency in the ability to convert L-phenylalanine to L-tyrosine. In addition to the marked changes in amino acid metabolism that such a deficit causes, it would also be expected to cause a pronounced elevation in 2-phenylethylamine levels.

Considerable circumstantial evidence also exists linking an alteration in trace amine metabolism to affective disorders. In particular, depression has been hypothesized to be associated with a decrease in 2-phenylethylamine functioning. Decreases in

2-phenylethylamine and its main metabolite phenylacetic acid have been reported in various body fluids. Similarly, decreases in other trace amines and their primary metabolites have been sporadically reported. In agreement with a proposed 2-phenylethylamine deficiency, the combined administration of (–)deprenyl with L-phenylalanine has been suggested to be an effective therapy in up to 60% of patients. Such a regimen would be expected to elevate endogenous 2-phenylethylamine (PE) levels; (–)deprenyl is a selective inhibitor of MAO-B, thereby inhibiting PE metabolism, while L-phenylalanine is the precursor amino acid of PE, and is expected to simultaneously elevate PE synthesis. Somewhat surprisingly, however, replication of this study does not appear to have been attempted. In further support of an antidepressant role for 2-phenylethylamine, MAO-B knockout mice show a behavioural phenotype consistent with that seen following administration of known antidepressant compounds. The only change in neuroactive amines seen in these animals is a pronounced elevation of endogenous PE levels. Finally, the mood-elevating effects of exercise may be due to an exercise-induced increase in PE levels, although the molecular basis of such an elevation is unclear.

At the receptor and gene level, there is also circumstantial evidence to support a role for PE in mood elevation. As detailed above, the available evidence suggests that TAARs are particularly associated with the amygdala, a brain region particularly implicated in the control of emotion and affect. Polymorphisms in the trace amine synthetic enzyme AADC have been suggested to be a susceptibility factor for both unipolar and bipolar disorder, although such associations require replication. Further, bipolar disorder has been linked to chromosome 6q. As seen above, this is the site of all TAAR genes. In this aspect, a similar confounding factor applies, as was detailed with schizophrenia. Affective disorders represent a spectrum of disorders, with a single causative agent unlikely to be present. A number of known antidepressants result in a downregulation of β-adrenergic receptors. Of interest in this regard is evidence that chronic elevation of endogenous PE results in a similar downregulation. Finally, the antidepressant agent phenelzine has been shown to be metabolized to PE.

The link between trace amines and disorders of higher cognitive functioning such as schizophrenia and affective disorders is appealing. As detailed, TAARs have been suggested to be involved in adaptive responses. Further, the observed physiological effects of trace amines acting to modulate the efficacy of neurotransmitter amines, possibly functioning to keep neuronal tone within physiological limits, are an attractive site for therapeutics for disorders of higher cognitive functioning involving multiple neurotransmitter systems. In this respect, trace amines have been implicated in other disorders associated with higher cognitive functioning, in particular attention-deficit/hyperactivity disorder, and also eating disorders and drug abuse/dependence.

Though not psychiatric in nature, one final disorder which warrants consideration with respect to the trace amines is migraine. This neurovascular disorder has often been linked to the trace amines largely on the basis of the elevated levels of trace amines found in foodstuffs known to precipitate migraine attacks in susceptible individuals. In particular, chocolate contains high levels of 2-phenylethylamine, while red wine and aged cheeses contain high levels of tyramine.

Thus, although many questions remain unanswered with respect to the endogenous ligand and physiological functioning of TAAR, they have attracted considerable interest with respect to being a viable target for development of novel therapeutic agents. Of note, many of the diseases in which trace amines are being actively investigated are of unknown etiology and likely involve multiple neurotransmitter systems. The apparent role of trace amines in modulating the responsivity of a variety of neurotransmitter systems makes them an attractive area of investigation in various neurological and psychiatric disorders.

See also: Amphetamines; Attention Deficit Hyperactivity Disorder; Attention Deficit Hyperactivity Disorder (ADHD): Methylphenidate (Ritalin) and Dopamine; Dopamine – CNS Pathways and Neurophysiology; Drugs Addiction: Actions; Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Monoamines; Octopamine and Other Monoamines in Invertebrates; Schizophrenia: Epidemiology, Clinical Features, Course and Outcome; Serotonin (5-Hydroxytryptamine; 5-HT): CNS Pathways and Neurophysiology.

Further Reading

Berry MD (2004) Mammalian central nervous system trace amines. Pharmacologic amphetamines, physiologic neuromodulators. *Journal of Neurochemistry* 90: 257–271.

Berry MD (2007) The potential of trace amines and their receptors for treating neurological and psychiatric disorders. *Reviews on Recent Clinical Trials* 2: 3–19.

Borowsky B, Adham N, Jones KA, et al. (2001) Trace amines: Identification of a family of mammalian G protein-coupled receptors. *Proceedings of the National Academy of Sciences of the United States of America* 98: 8966–8971.

Branchek TA and Blackburn TP (2003) Trace amine receptors as targets for novel therapeutics: Legend, myth and fact. *Current Opinion in Pharmacology* 3: 90–97.

Bunzow JR, Sonders MS, Arttamangkul S, et al. (2001) Amphetamine, 3,4-methylenedioxymethamphetamine, lysergic acid diethylamide, and metabolites of the catecholamine neurotransmitters are agonists of a rat trace amine receptor. *Molecular Pharmacology* 60: 1181–1188.

Hart ME, Suchland KL, Miyakawa M, et al. (2006) Trace amine-associated receptor agonists: Synthesis and evaluation of thyronamines and related analogues. *Journal of Medicinal Chemistry* 49: 1101–1112.

Lewin AH (2006) Receptors of mammalian trace amines. *AAPS Journal* 8: E138–E145.

Lindemann L, Ebeling M, Kratochwil NA, et al. (2005) Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors. *Genomics* 85: 372–385.

Lindemann L and Hoener MC (2005) A renaissance in trace amines inspired by a novel GPCR family. *Trends in Pharmacological Sciences* 26: 274–281.

Reese EA, Bunzow JR, Arttamangkul S, et al. (2007) Trace amine-associated receptor 1 displays species-dependent stereoselectivity for isomers of methamphetamine, amphetamine and para-hydroxyamphetamine. *Journal of Pharmacology and Experimental Therapeutics* 321: 178–186.

Transcranial Magnetic Stimulation

L B Merabet and A Pascual-Leone, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction and History

Transcranial magnetic stimulation (TMS) is a noninvasive neurophysiologic method to induce electrical currents within the brain. This technique can be used to transiently disrupt the function of a targeted brain area, functionally map cortical areas, assess cortical excitability, and even modulate its level of activity when delivered as a repetitive train of pulses. TMS is being widely used to investigate complex aspects of human brain function, such as learning and cognition, as well as sensory and motor function. Furthermore, its potential as a novel therapeutic treatment is being actively pursued in a variety of neurological and neuropsychiatric disorders.

The first reports documenting activation of human brain by an electromagnetic field appeared as early as 1896. Jaques-Arsène d'Arsonval and, later, Sylvanus P Thompson (1910) demonstrated that placing the head near an electrical coil generating a time-varying magnetic field induced the sensation of phosphenes (i.e., the subjective sensation of light) and vertigo. Subsequent efforts were pursued aimed at refining the technique so that peripheral nerves could be selectively stimulated or muscular contractions could be induced. However, all these magnetic stimulation devices were technically difficult to operate, prone to extreme overheating, and had no mechanisms to address the interference generated by the stimulating fields. In the late 1970s and early 1980s, Anthony Barker and collaborators at the University of Sheffield, UK, overcame many of these earlier technical problems and developed a device capable of generating peak magnetic fields as high as 2 T. This work served as a precursor to the development of a stable and reliable device and, ultimately, the first commercially available magnetic stimulator. During the past 20 years, the application of TMS in clinical neurology and neuroscience research has grown steadily due to improvements in technology (e.g., coil design and the development of stimulators capable of rapid discharge rates), its combined use with other neuroimaging modalities, and a greater understanding of the technique.

How Does TMS Work?

The physical basis of TMS rests on the principles of electromagnetic induction originally discovered by Michael Faraday in 1831. Simply stated, a brief electrical pulse of rapidly alternating current passed through a coil of wire generates a strong, transient magnetic field. By placing a stimulating coil near a subject's scalp, the induced magnetic field pulse penetrates the scalp and skull virtually unattenuated (decaying only with distance) reaching the underlying brain tissue. The rate of change of this magnetic field in turn induces a secondary electrical current within brain tissue (or with any nearby conducting medium) (Figure 1(a)). Unlike direct electrical cortical stimulation, TMS does not require exposing the underlying cortex and is relatively painless and safe. In essence, TMS represents a method of 'electrodeless' electrical stimulation, whereby the magnetic component is responsible for bridging the transition between current passed in the coil and current induced in the brain.

The design of magnetic stimulators is relatively simple. A typical TMS machine consists of a main unit and a stimulating coil (Figure 1(b)). The main unit comprises a charging system, one or more energy storage capacitors, a discharge switch, and circuits for pulse shaping, energy recovery, and control functions. Different charging systems are possible; the simplest design uses step-up trans-formers operating at frequencies of 50–60 Hz. Energy-storage capacitors can also be of different types. The essential factors in the effectiveness of a magnetic stimulator are the speed of the magnetic field rise time and the maximization of the peak coil energy. Therefore, large energy-storage capacitors and very efficient energy transfer from the capacitor to the coil are important. Typically, energy storage-capacity is approximately 2000 J, and 500 J is transferred from the capacitors into the stimulating coil in less than 100 μs via a thyristor, which is an electronic device that is capable of switching large currents in a few microseconds. Finally, in order to induce currents in the brain of sufficient magnitude to depolarize neural elements, the peak discharge current needs to be on the order of several thousand amperes (approximately 10 mA cm^{-2}).

The effects of this induced current on brain tissue are dependent on numerous stimulation parameters, such as stimulation frequency and intensity as well as coil configuration and orientation. The voltage of the primary current and the geometry of the stimulating coil determine the strength and shape of the magnetic field and thus the density and spatial resolution of the secondary current induced within brain tissue. There are two main coil designs: round (circular) coils and figure-of-eight shaped coils (Figure 2). Round coils generate large and more diffuse currents and are useful for stimulating relatively large cortical areas. Figure-of-eight coils are more focal due to a maximal

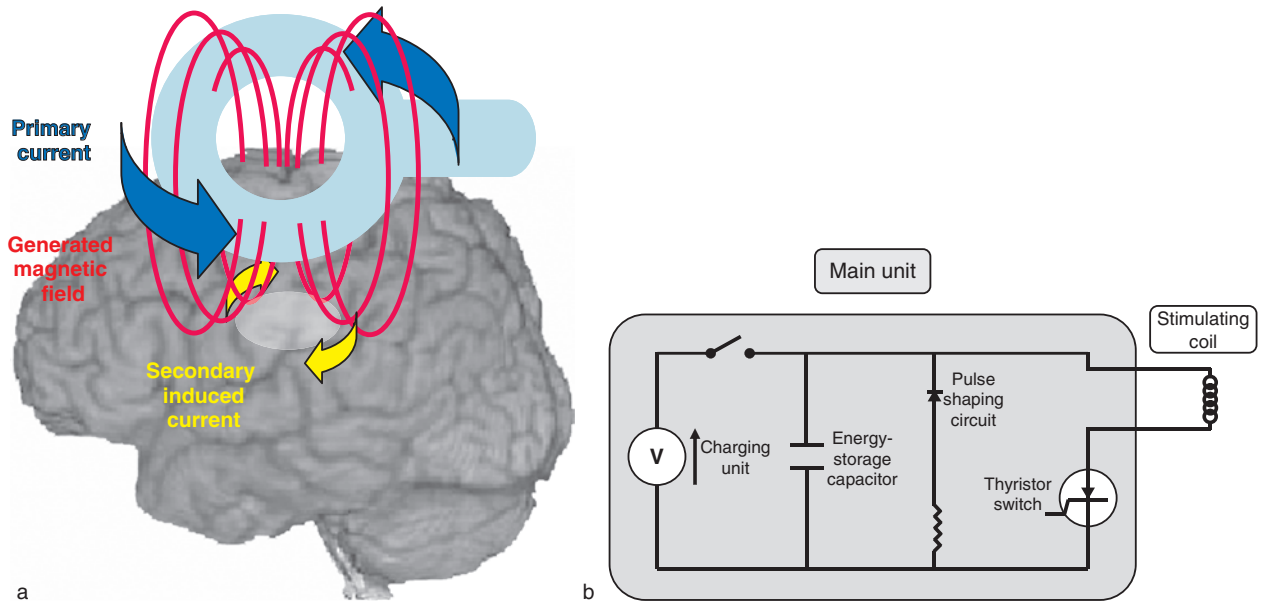

Figure 1 (a) Schematic description of the currents induced by a circular coil. The induced magnetic field couples the primary current passed in the coil with the secondary current induced within neuronal tissue. (b) Schematic diagram of a standard magnetic stimulator circuit (single pulse).

current being generated at the intersection of the two round coil wings, and they are used for stimulating within a more circumscribed area. Evidence suggests that the locus of TMS stimulation has a spatial resolution of 1 cm^2 with a penetration depth of approximately 2 cm using commercially available figure-of-eight stimulation coils. It is also important to note that these values largely represent mathematical approximations and, in fact, current modeling evidence suggests that very different parameters may apply when stimulating patients with cerebral damage such as stroke. Stimulating at greater depths is possible, although larger amounts of current are required, which in turn compromise spatial resolution. Alternatively, penetration depth can be enhanced by using innovative coil designs that modify the decay of the magnetic field with distance. Even with such coils, the stimulation of structures closest to the coil is maximal, and the focality of the stimulation also decays with distance. Therefore, selective, direct, and precise stimulation of deep cerebral structures (e.g., limbic structures, basal ganglia, or brain stem) is not feasible with current stimulators.

When a coil is placed tangentially to the surface of the scalp, current flows parallel to its surface plane preferentially activating neural elements oriented horizontally within the cortex. Given that interneurons are horizontally oriented, it is thought that TMS is more likely to activate pyramidal cells indirectly in a transsynaptic manner. Typically, a single pulse lasts 200 μs; however, stimulation can also be delivered as a train of repetitive pulses (referred to as repetitive TMS or rTMS). rTMS has the advantage of being able

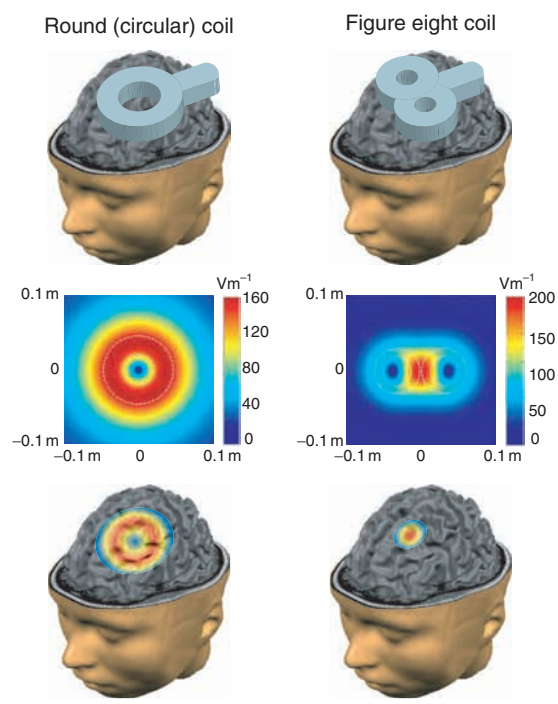

Figure 2 Figure-of-eight and circular coil designs. The spatial resolution of the figure-of-eight coil is much greater than its circular counterpart, enabling a more precise target location.

to modulate cortical excitability with effects lasting well beyond the period of stimulation.

From a safety standpoint, the pervasive concern of using TMS is the risk of inducing a seizure in subjects with a known prior history of epilepsy or other predisposing risk factor, such as taking medication

known to lower the threshold for seizure induction. Single-pulse TMS is considered safe and free of significant side effects; however, the use of rTMS, given its ability to modify cortical excitability, represents a much greater safety risk. Extensive studies have been carried out in order to establish appropriate safety guidelines, classify predisposing risk factors, and identify signs and surrogates of seizure induction. Since the institution of safety guidelines defining stimulation parameters and the monitoring and screening of appropriate subjects, the risk of TMS-induced seizures has been greatly reduced.

TMS in Motor Function

One of the first applications of TMS was in clinical neurophysiology. For example, when a TMS pulse is applied to the hand representation in motor cortex, a twitch in the contralateral hand can be evoked and a motor evoked potential (MEP) can be recorded. The neuronal conduction time of motor signals (i.e., from motor cortex to spinal cord) can be calculated by subtracting the latency of the evoked potential of the target muscle from the latency of motor potentials induced by the stimulation of spinal cord nerve roots. Given that a slowing of conduction time is typically seen in neurodegenerative disorders, measuring variables such as MEP amplitude and latency is useful in the evaluation of patients with multiple sclerosis, amyotrophic lateral sclerosis, or spinal cord damage. Similarly, TMS has been used to probe recovery of function following brain damage (e.g., stroke) and may have potential in the development of neurorehabilitative strategies. Focal stimulation by TMS can also be used to map the extent of cortical motor representations and function. Maps can be obtained at different times, within the same subject, and evaluated under different conditions. This is particularly useful in studying functional changes in cortical activity associated with cerebral plasticity. For example, neuroplastic changes in sensorimotor cortical representation can be demonstrated when one learns a skill such as playing a novel musical piece on the piano. In one study, it has been shown that cortical representation expands in accordance to overall improvement of playing ability and returns to normal size once the skill is adequately learned. TMS can also be used in a variety of ways to measure cortical excitability. Given that excitability represents an interaction of neurotransmitter levels and cortical circuitry; the effect of medications and various pathological conditions can be assessed through numerous measures using TMS. These measures in turn can be used to develop a relationship between physiological effects and neuropharmacology and potentially assist in the development of highly individualized pharmacological treatment for conditions such as epilepsy, depression, and Parkinson's disease.

Measures of Motor Function

One important measure of motor function that can be assessed using TMS is the motor threshold (MT). MT is defined as the lowest TMS intensity necessary to evoke an MEP in a target muscle in at least 50% of successive stimulations (e.g., in 5 out of 10 trials) using single pulses applied to the representation of the targeted muscle in the motor cortex. The MT provides fundamental information about the excitability of the core representation of a given target muscle within motor cortex and it is believed to reflect excitability of corticospinal neurons as well as the excitability of motor neurons in the spinal cord, neuromuscular junctions, and muscle. MT change has been studied in various diseases and has been shown to increase in neurological conditions that affect the corticospinal tract, such as multiple sclerosis and stroke, as well as brain or spinal cord injury. Depending on their mechanism of action, various medications have also been shown to affect MT.

One method used to study and characterize cortical efferent output is to compare how an increase in an MEP amplitude relates to increasing TMS intensity. This is known as the recruitment curve, an input–output curve, or a stimulus–response curve. Compared to the simple determination of the MT, this measure provides insight into the function of neurons that are intrinsically less excitable or spatially farther from the center of activation by TMS. Therefore, recruitment curves are likely related to the strength of corticospinal projections.

Central motor conduction time (CMCT) is defined as the latency difference between the MEPs induced by stimulation from the motor cortex and those evoked by spinal (motor root) stimulation. It is calculated by subtracting the latency of the peripheral segment of the motor pathway (spinal motor root to muscle) from that of the entire motor pathway (motor cortex to muscle; Figure 3, top). When a TMS coil is placed over the back of the neck or spine, the magnetic pulse will stimulate spinal nerve roots but not the descending spinal tracts. Abnormalities of CMCT have been demonstrated in a variety of neurological disorders. Significant prolongation of CMCT suggests demyelination of central motor pathways, whereas low-amplitude MEPs with little delay or absence of responses are more suggestive of loss of neurons or axons. The CMCT has also been shown to correlate with the grade of motor deficits following stroke damage.

When a subject is instructed to maintain muscle contraction of a targeted muscle and a single TMS

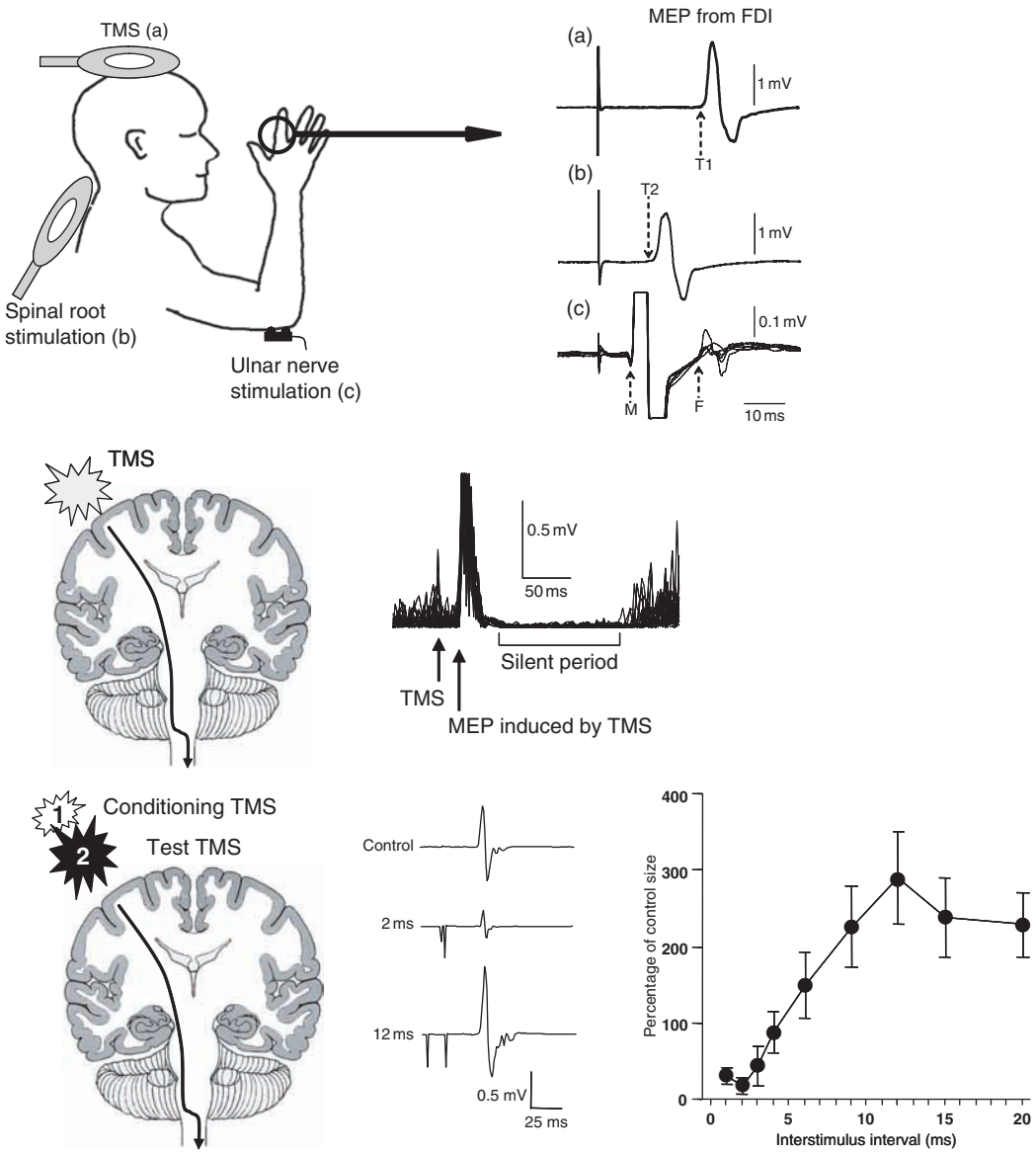

Figure 3 (Top) Schematic representation of the calculation of central motor conduction time (CMCT). (a) Motor evoked potential (MEP) induced by TMS. (b) MEP after cervical spinal root stimulation. (c) F-waves after ulnar nerve electric stimulation. CMCT is estimated by onset latency of T1 minus onset latency of T2. T1, onset latency of MEP elicited by TMS; T2, onset latency of MEP elicited by the coil placed on the back of cervical spine; M, onset latency of M-wave by electrical ulnar nerve stimulation; F, onset latency of F-wave by electrical ulnar nerve stimulation. (Middle) Example of silent period to TMS. Stimulation of the hand motor cortex of one hemisphere elicits MEPs in the contralateral first dorsal interosseus muscle (FDI) of the hand, which is followed by a suppression of tonic voluntary electromyographic (EMG) activity. Responses are recorded from the FDI, which is kept contracted with 20% of maximal voluntary force pressing on a force transducer. The intensity of TMS is set at 130% of the motor threshold. (Bottom) A paired pulse experiment illustrating the change of MEP sizes obtained by the paired conditioning–test-stimulus paradigm from the FDI muscle. The intensity of conditioning TMS was set to 80% of resting motor threshold. The intensity of test TMS was adjusted to produce a control MEP of an average peak-to-peak amplitude of 1 mV when given alone. Note that the conditioning stimulus inhibited the test MEP at short interstimulus intervals (1–4 ms) but facilitated it at longer intervals (6–20 ms). In the graph, the size of the MEPs is expressed as a percentage of the control, unconditioned MEPs, and plotted against the interstimulus interval. Data are means of eight healthy subjects. Error bars indicate standard errors.

pulse with suprathreshold intensity is applied over the motor cortex representation of that targeted muscle, the electromyographic (EMG) activity of the muscle will be arrested for a few hundred milliseconds following the MEP. This period of EMG suppression is referred to as a 'silent period' (Figure 3, middle), often defined as the time from the end of the MEP to the return of voluntary EMG activity. Silent periods of abnormally short or long duration are observed in patients with various movement disorders such as

Parkinson's disease. Most of this period is believed to be due to cortical inhibitory mechanisms of the motor cortex. The neuronal elements responsible for the silent period are most likely mediated by GABAergic-type receptors.

Finally, the combination of a conditioning stimulus with a test stimulus delivered at various intensities and with variable interstimulus intervals to the same cortical area can provide insight into cortico-cortical interactions and measure short- and long-term inhibition and facilitation. This technique is referred to as 'paired pulse' stimulation (Figure 3, bottom). Here, a subthreshold conditioning stimulus is followed by a suprathreshold test stimulus at different (1–20 ms) interstimulus intervals through the same TMS coil targeting the primary motor cortex. The effects of the conditioning TMS on the size of test MEP depend on the stimulus intensity and the interstimulus interval. When a conditioning stimulus is set at 60–80% of the resting motor threshold, maximum inhibitory effects are found at short interstimulus intervals of 1–4 ms. The maximum amount of this inhibition is usually 20–40% of the test MEP. Facilitatory effects of the conditioning TMS pulse with the same subthreshold intensity onto the test MEP can be observed at intervals of 7–20 ms. The magnitude of this facilitation is quite variable across subjects and can also vary depending on the size of test MEPs. Thus, for a paired pulse TMS study, it is critical to maintain the relaxation of patients and adjust the size of test MEPs across trials and subjects. This paired pulse technique has also been used to examine the effects of central nervous system-active drugs on the human motor cortex. In this context, paired pulse TMS might be useful in selecting the best-suited medication for a given patient by matching the identified abnormality in a given disorder with the effects of different pharmaceutical agents. Such a neurophysiological-based approach to medication selection in epilepsy or psychosis would certainly be desirable, although systematic studies are needed.

These various measures can give rise to clinically relevant applications of TMS by characterizing corticospinal projections and even cortico-cortical or interhemispheric inhibition and facilitation. It is also noteworthy that stimulation is not limited to cortical areas. TMS can also be applied to the cerebellum or at the level of the foramen magnum, a peripheral nerve, or even nerve plexus, thus allowing an investigator the ability to examine the motor system at different levels of the neuroaxis. From a clinical standpoint, these objective and quantifiable measures can help localize the level of a lesion within the nervous system, distinguish between a predominantly demyelinating or axonal lesion in motor tracts, or even predict functional motor outcome. Like other neurophysiological tests, however, the abnormalities revealed by TMS are not always disease specific. It remains to be fully evaluated whether such neurophysiologic findings provide earlier prognostic indicators compared to those that might be obtained from careful, longitudinal neurological exams.

Repetitive TMS

rTMS refers to the application of a train of TMS pulses of the same intensity to a single brain area at a given frequency. In general, the higher the stimulation frequency and intensity, the greater the disruption of cortical function during the applied train of stimulation. However, following these immediate effects, a train of rTMS can also induce a prolonged modulation of cortical excitability, lasting well beyond the duration of the rTMS train. The modulatory effect may also range from inhibition to facilitation, depending on the stimulation parameters used. Lower frequencies (e.g., 1 Hz) applied over the motor cortex can suppress excitability, whereas higher-frequency stimulation trains (e.g., 20 Hz) lead to increases in cortical excitability. It is noteworthy that these effects vary among individuals and that other forms of trains, combining various stimulation frequencies (e.g., 6 Hz-primed 1 Hz stimulation, or theta burst stimulation), may in fact induce a more robust, consistent, and longer lasting neuromodulation.

The mechanisms underlying these 'off-line' modulatory effects lasting beyond the duration of the rTMS train are unclear. Long-term potentiation and long-term depression of cortical synapses or other closely related neuronal mechanisms have been suggested to explain the effect of high- and low-frequency rTMS, respectively. Animal studies suggest that modulation of neurotransmitters and gene induction may also represent contributory factors. Finally, neuronal network effects within the stimulated hemisphere as well as across a cortico-subcortical bihemispheric network may also play a critical role. It is also noteworthy that the lasting modulation of cortical activity by rTMS is not limited to motor cortical areas. Long-lasting effects of rTMS can be induced in other areas associated with measurable behavioral effects, including visual, prefrontal, and parietal cortex as well as the cerebellum.

Mapping Neuronal Networks in Cognitive Science

TMS has also been used extensively in the mapping of other, nonmotor cortical targets in relation to stud-

ies involving memory, attention, language, visual perception, and imagery. A TMS pulse, appropriately delivered in space and time, can temporally disrupt cortical function in a given area. By establishing a causal link between brain activity and behavior, TMS can be used to uncover which parts of an identified network are necessary for performing a behavioral task and thus supplement and refine conclusions drawn from functional neuroimaging studies. For example, when a TMS pulse is delivered at a specific time to the occipital pole, the detection of a presented visual stimulus can be blocked. By varying the timing between the visual stimulus and TMS pulse, the chronometry of visual perception from primary visual cortex to extrastriate visual areas can be ascertained.

In studies of language function, delivering high-frequency rTMS (on the order of 25 Hz) centered over the inferolateral frontal cortex can induce speech arrest in a subject instructed to count aloud or even induce paraphasia-like errors in a subject asked to name words starting with a given letter. Thus, rTMS may prove to be a noninvasive method to determine the language-dominant hemisphere of an individual (e.g., prior to surgical evaluation), and ongoing clinical trials have demonstrated that TMS may have beneficial effects in the rehabilitation of language function in patients with aphasia due to stroke damage.

Numerous studies have used TMS to establish the functional significance of data obtained through neuroimaging. One particularly elegant example has been the use of TMS to probe the function of occipital (visual) cortex in early blind subjects. Using functional neuroimaging, many investigative groups have demonstrated that the occipital visual cortex is active while blind subjects read Braille. However, with the complementary use of TMS, a clear demonstration that this activation is causally related to the task of Braille reading can be demonstrated. Specifically, it has been shown that reversible disruption of occipital cortex using rTMS disrupts Braille reading performance in early blind readers but not in sighted controls, thus establishing a causal relationship between occipital cortex activation and the necessity of this area for performing this skill. Other studies have also used TMS in conjunction with data obtained with functional neuroimaging to establish the role of occipital cortex in other forms of tactile processing, such as tactile form and object recognition, and even in verbal memory skills in the blind.

For such applications, information regarding the actual site of stimulation is of utmost importance. In studies of motor cortex, for example, the site of stimulation can be identified and reasonably ascertained given that stimulation of this cortical site evokes a clear objective measure such as the motor evoked potential. For other cortical areas in which no such objective measure is easily attainable, the use of fiducial landmarks may not be satisfactory to reliably identify the locus of stimulation, especially if one considers the variability in morphology and size of cerebral areas across individual subjects. For this reason, inferring the site of stimulation using a frameless stereotaxic system can provide the precise anatomical resolution required (**Figure 4**).

Therapeutic Applications

Considerable interest has arisen in the use of TMS as a form of therapy in certain neurological disorders and neuropsychiatric conditions. Its therapeutic potential rests on the fact that rTMS can modulate cortical excitability well beyond the actual period of stimulation. For example, studies have demonstrated that rTMS delivered to cortical motor areas can enhance reaction times in certain motor coordination tasks and improve motor function in parkinsonian patients. Applying low-frequency rTMS (1 Hz) to motor cortex and thereby decreasing cortical excitability has been shown to improve symptoms in focal hand dystonia (e.g., 'writer's cramp'). However, not all studies have been able to reproduce these findings, so results must remain preliminary and further carefully controlled studies are needed.

Although the inhibitory and facilitatory effects of rTMS were first documented in studies of corticospinal excitability, it is possible that rTMS applied to other nonmotor areas may exert similar modulatory and potentially therapeutic effects. One avenue being pursued is in the therapy of patients with medication-resistant major depressive disorder (MDD). Neuroimaging data have demonstrated that patients with MDD display decreased levels of activity in left dorsolateral prefrontal cortex (DLPFC). Stimulation of this cortical target using focal rTMS at 20 Hz has been shown to significantly decrease depressive symptoms. Interestingly, low-frequency rTMS delivered to right DLPFC has also been shown to have antidepressant effects. This suggests that establishing a balance in frontal lobe function may have a beneficial effect in this disorder. Studies are under way to investigate stimulation parameters (e.g., the exact site of

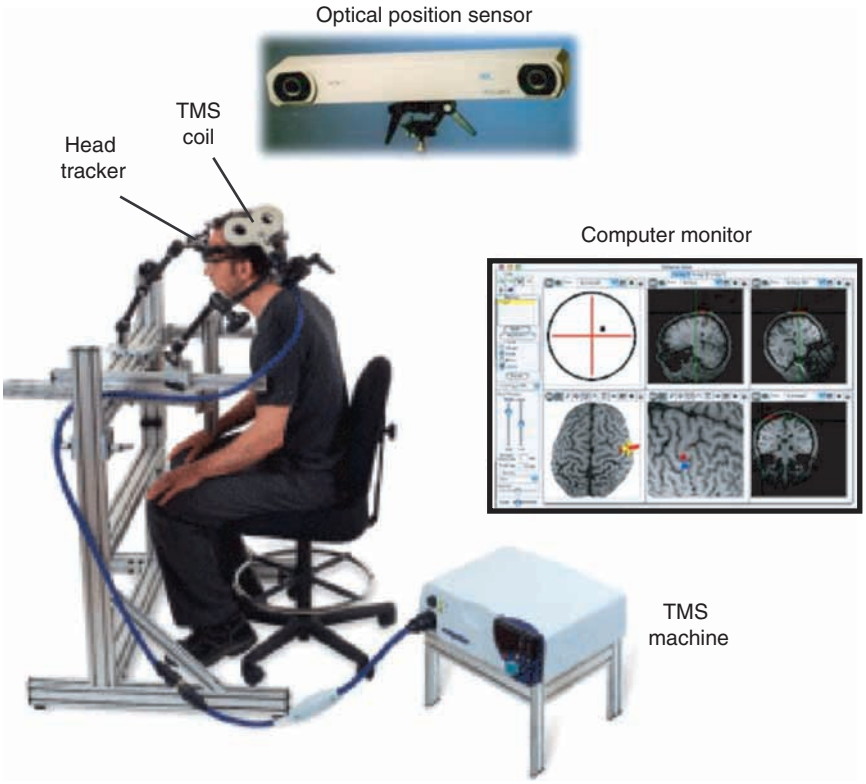

Figure 4 Frameless stereotaxic system used for the on-line guidance of coil location over an anatomical MRI of the subject. The positions of small reflective trackers attached to the TMS coil are monitored relative to trackers positioned on the subject's head. The relative position of the trackers is monitored by an optical position sensor (camera). Following a calibration and coregistration procedure, the position and orientation of the coil relative to the subject's structural MRI can be displayed in real time on a computer monitor and followed throughout the experiment. Frameless streotaxic tracking system (Brainsight) shown here courtesy of Rogue Research, Montreal, Quebec, Canada (http://www.rogue-research.com).

stimulation, intensity, train duration, and number) in the hopes of establishing rTMS as a viable therapeutic adjunct (if not alternative) for electroconvulsive therapy.

See also: Cognition: An Overview of Neuroimaging Techniques; Deep Brain Stimulation; Electromyography (EMG) and Nerve Conduction Studies; Human Depth Electrodes.

Further Reading

Barker AT (1999) The history and basic principles of magnetic nerve stimulation. *Electroencephalography and Clinical Neurophysiology Supplement* 51: 3–21.

Cohen LG, Celnik P, Pascual-Leone A, et al. (1997) Functional relevance of cross-modal plasticity in blind humans. *Nature* 389: 180–183.

Epstein CM (1998) Transcranial magnetic stimulation: Language function. *Journal of Clinical Neurophysiology* 15: 325–332.

Fregni F, Simon DK, Wu A, and Pascual-Leone A (2005) Non-invasive brain stimulation for Parkinson's disease: A systematic review and meta-analysis of the literature. *Journal of Neurology, Neurosurgery, & Psychiatry* 76: 1614–1623.

Kobayashi M and Pascual-Leone A (2003) Transcranial magnetic stimulation in neurology. *Lancet Neurology* 2: 145–156.

Merabet LB, Theoret H, and Pascual-Leone A (2003) Transcranial magnetic stimulation as an investigative tool in the study of visual function. *Optometry and Vision Science* 80: 356–368.

Pascual-Leone A, Amedi A, Fregni F, and Merabet LB (2005) The plastic human brain cortex. *Annual Review of Neuroscience* 28: 377–401.

Pascual-Leone A, Davey N, Wassermann EM, Rothwell J, and Puri BK (eds.) (2001) *Handbook of Transcranial Magnetic Stimulation*. London: Arnold.

Pascual-Leone A, Nguyet D, Cohen LG, Brasil-Neto JP, Cammarota A, and Hallett M (1995) Modulation of muscle responses evoked by transcranial magnetic stimulation during the acquisition of new fine motor skills. *Journal of Neurophysiology* 74: 1037–1045.

Pascual-Leone A, Tormos JM, Keenan J, Tarazona F, Canete C, and Catala MD (1998) Study and modulation of human cortical excitability with transcranial magnetic stimulation. *Journal of Clinical Neurophysiology* 15: 333–343.

Pascual-Leone A, Walsh V, and Rothwell J (2000) Transcranial magnetic stimulation in cognitive neuroscience – Virtual lesion, chronometry, and functional connectivity. *Current Opinion in Neurobiology* 10: 232–237.

Simons W and Dierick M (2005) Transcranial magnetic stimulation as a therapeutic tool in psychiatry. *World Journal of Biological Psychiatry* 6: 6–25.

Théoret H and Pascual-Leone A (2003) Transcranial magnetic stimulation and the study of cognition. In: Hugdahl K (ed.) *Experimental Methods in Neuropsychology*, pp. 173–195. Dordrecht, The Netherlands: Kluwer.

Walsh V and Pascual-Leone A (2003) *Neurochronometrics of Mind: TMS in Cognitive Science*. Cambridge, MA: MIT Press.

Wassermann EM (1998) Risk and safety of repetitive transcranial magnetic stimulation: Report and suggested guidelines from the International Workshop on the Safety of Repetitive Transcranial Magnetic Stimulation, June 5–7, 1996. *Electroencephalography and Clinical Neurophysiology* 108: 1–16.

Wassermann EM and Lisanby SH (2001) Therapeutic application of repetitive transcranial magnetic stimulation: A review. *Clinical Neurophysiology* 112: 1367–1377.

Ziemann U (2004) TMS and drugs. *Clinical Neurophysiology* 115: 1717–1729.

Relevant Website

http://www.tmslab.org – Center for Noninvasive Brain Stimulation.

Transcription and Reward Systems

A Kumar, T A Green, S J Russo, W Renthal, and E J Nestler, The University of Texas Southwestern Medical Center, Dallas, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Brain Reward Pathways

Animals, including humans, prefer to engage in behaviors that are rewarding (i.e., pleasurable). This reward acts as a positive reinforcement such that the individual engages in behavior to repetitively obtain that reward. Brain circuits that mediate these effects of reward evolved to regulate an individual's responses to natural rewards, such as food, sex, and social interaction. Adaptations in these circuits, in response to the presence or absence of rewards, mediate powerful drives and memories, which serve to promote survival of the species. Drugs of abuse act on this brain reward circuitry with far more power and persistence than seen with natural rewards. Drugs thereby corrupt the reward circuits in ways that cause drug seeking and consumption to gradually take over the individual's life; that is, they induce a state of addiction.

The most important component of the brain's reward circuitry is the mesolimbic dopamine system, which is composed of dopamine neurons in the ventral tegmental area (VTA) of the midbrain, and the several forebrain regions to which these dopamine neurons project, in particular, the nucleus accumbens (NAc) (also known as ventral striatum) and several areas of frontal cortex (e.g., prefrontal, orbitofrontal and anterior cingulate cortex). This brain circuitry is illustrated in Figure 1. Many other brain regions, including the hippocampus, amygdala, and hypothalamus, to name a few, contribute to the reward circuitry, but are not discussed further here due to space limitations.

All rewarding stimuli, both natural as well as drugs of abuse, activate dopaminergic transmission in the mesolimbic dopamine pathway. Some also produce dopamine-like effects in the NAc and frontal cortex via dopamine-independent mechanisms. The mechanism by which dopaminergic transmission mediates 'reward' is unknown, but is the subject of great interest. Activation of dopaminergic transmission may elicit some of the subjective effects of a reward, as well as its positive reinforcing effects. Chronic activation of dopaminergic transmission can cause behavioral plasticity, which shapes behavior over the longer term to promote further consumption of the reward. The use of dopamine to regulate responses to natural rewards is highly conserved from an evolutionary point of view; rudiments of the pathway are seen in worms and flies, which arose over 1 billion years ago.

Drugs of abuse are thought to cause addiction by producing powerful adaptations in the brain's reward circuitry. During a period of drug use, the reward circuitry, due to its strong, repeated activation by drug, tends to be hyposensitive, such that, in the absence of drug, addicts are not satisfied with natural rewards and require drug to feel normal. These changes, which often lead to relapse, can be viewed as tolerance and dependence to the drug's rewarding effects and typically resolve over days and weeks of withdrawal. Repeated drug use also elicits a more long-lasting sensitized state, when cues previously paired with drug, stress, or the drug itself cause a robust activation of the reward circuitry, and can lead to relapse as well. A major focus of the field is to understand the molecular and cellular basis of these types of drug-induced adaptations in brain reward regions.

Increasing evidence supports the view that the brain's reward circuitry is also involved in other psychiatric disorders, in particular, mood disorders such as depression and bipolar disorder. For example, most patients with depression exhibit a reduced ability to experience pleasure (anhedonia), loss of motivation, and abnormalities in several neurovegetative functions, such as sleep, appetite, energy level, and circadian rhythms, all of which are influenced by the mesolimbic dopamine pathway and related reward regions of brain. Similarly, most patients in a manic state exhibit the opposite symptoms. In addition, it has long been known that stress and other aversive emotional stimuli potently regulate the brain's reward circuitry. These observations suggest that a better understanding of the molecular mechanisms that control the long-term activity of brain reward regions will contribute to our knowledge of these disorders as well.

Neural Plasticity in Brain Reward Pathways

The structure and function of neurons, and the patterns of their connections in any neural circuit, including the reward circuitry, are not fixed. An individual neuron changes its myriad of receptors, its response patterns to diverse stimuli, and eventually its output to other neurons in response to environmental stimuli. This phenomenon of neural plasticity, that is, the ability of neural cells and circuits to undergo modifications in structure, connectivity, and function, plays a critical role not only in growth

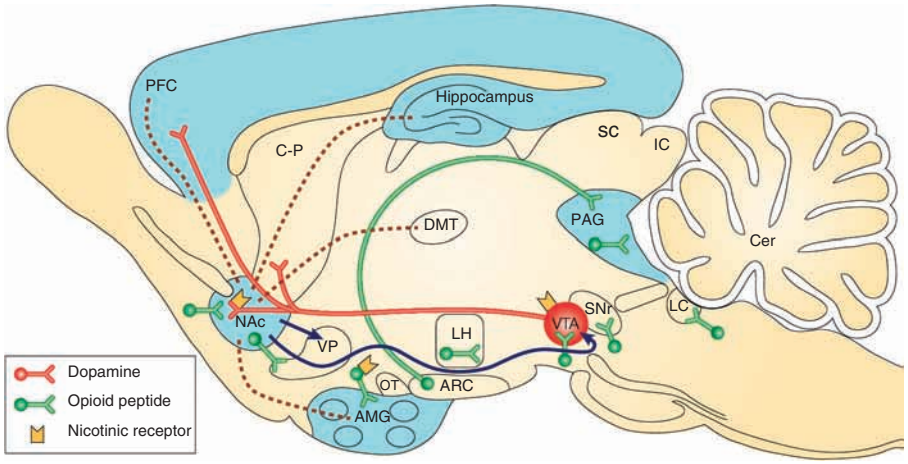

Figure 1 Brain reward regions. Brown dotted lines indicate limbic afferents to the nucleus accumbens (NAc). Blue represents efferents from the NAc thought to be involved in drug reward. Red indicates dopaminergic projections of the mesolimbic dopamine system thought to be a critical substrate for drug reward. This system originates in the ventral tegmental area (VTA) and projects to the NAc and other limbic structures, including olfactory tubercle (OT), ventral domains of the caudate-putamen (C-P), amygdala (AMG), and prefrontal cortex (PFC). Green indicates opioid peptide-containing neurons, which are involved in opiate, ethanol, and possibly nicotine reward. These opioid peptide systems include the local enkephalinergic circuits (short segments) and the hypothalamic midbrain beta-endorphin circuit (long segment). Yellow solid structures indicate nicotinic acetylcholine receptors hypothesized to be located on dopaminergic and opioid peptidergic systems, among other sites. ARC, arcuate nucleus; Cer, cerebellum; DMT, dorsomedial thalamus; PFC, prefrontal cortex; IC, inferior colliculus; LC, locus coeruleus; LH, lateral hypothalamus; OT, olfactory tubercle; PAG, periaqueductal gray; SC, superior colliculus; SNr, substantia nigra pars reticulata; VP, ventral pallidum. From Nestler EJ (2001) Molecular basis of long-term plasticity underlying addiction. *Nature Reviews Neuroscience* 2: 119–128, with permission.

and development, but also throughout adult life. Our ability to learn and remember depends on this plasticity, as does our ability to maladapt to aversive stimuli, such as drugs of abuse and stress.

The development of addiction can thus be viewed as an example of drug-induced neural plasticity. Upon first exposure, a drug of abuse enters the brain, binds to its initial protein target, and perturbs synaptic transmission at particular synapses in the brain, which in turn causes the acute behavioral effects of the drug (high, euphoria, activation, sedation, etc.). However, these acute drug effects do not *per se* explain addiction; rather, addiction develops gradually and progressively as a result of neuroadaptations to repeated drug intake. Drug-induced adaptations occur in many brain regions, depending on the drug, but all drugs of abuse produce a series of common molecular and cellular adaptations in brain reward regions, as would be expected given their common acute effects on this circuitry. Because the behavioral abnormalities associated with addiction are long lived – individuals can remain at increased risk for relapse despite many years of abstinence – a role for gene expression has been proposed as central to the addiction process. According to this scheme, the repeated, powerful perturbation of dopaminergic and other forms of neurotransmission in the VTA, NAc, and other brain reward regions by drugs of abuse causes long-lived changes in gene transcription within these regions, which alters reward mechanisms in such a way as to underlie addiction.

In a similar way, stress and other negative emotional stimuli can cause neuroadaptations in brain reward pathways and several other areas of brain and mediate the long-lived behavioral abnormalities seen in many stress-related disorders. However, it is important to emphasize that addiction, depression, and other stress-related disorders are highly heritable: roughly 50% of the risk for each of these conditions is genetic, although the specific genes that comprise this risk remain mostly unknown. Hence, a critical advance in studies of the molecular control of the brain's reward pathways will ultimately involve understanding how an individual's genes control the ability of a drug of abuse or stress to induce maladaptive changes within these brain regions.

Transcriptional Mechanisms in the Regulation of Brain Reward Pathways

The activation or repression of specific transcriptional patterns is part of a neuron's response to virtually every cellular signal. Accordingly, transcriptional regulation can be viewed as the ultimate target of signal transduction cascades, and is a critical mediator of a wide range of neuroadaptations in neural structure, connectivity, and function, which occur in response to environmental stimuli.

Considering the importance placed on alterations in gene expression in addiction and depression, it is surprising that only a small number of transcription factors have to date been implicated in these disorders. Moreover, even less is known about the target genes that are regulated by these transcription factors to mediate stable behavioral change. This review focuses on those transcription factors which have been shown to regulate the activity of the brain's reward pathways in animal models of drug addiction and depression.

CREB/CREM/ATF Family

The cAMP response element binding protein (CREB) family of transcription factors, in addition to CREB itself, include cAMP response element modulator (CREM) and activators of transcription (ATF). These proteins are BZip transcription factors; BZip refers to the basic domain (B) of the proteins that binds a specific consensus sequence (CRE or cAMP response element) in gene promoter/enhancer regions, and the leucine zipper domain (Zip) which allows homo- or heterodimerization of two BZip transcription factors necessary for transactivation. Although most proteins of this family heterodimerize selectively with other CREB/CREM/ATF transcription factors, some family members (specifically ATF2) can dimerize with BZips of other families and act on distinct response elements. To complicate matters further, there are many splice variants of these transcription factors that can act as potent repressors of transcription due to the absence of activation domains in the proteins.

Most evidence for a role of the CREB/CREM/ATF family in brain reward regions in stress and addiction models has focused on CREB itself. CREB is activated in both the NAc and VTA in response to acute and chronic administration of certain drugs of abuse (cocaine, amphetamine, opiates) or of stress. This has been shown by direct demonstration of increases in the phosphorylation of CREB at ser133, which is required for its transcriptional activation. It also has been demonstrated by use of CRE-LacZ transgenic mice, in which CRE transcriptional activity has been shown to be induced in brain reward regions under these conditions.

There is now a large body of evidence to support the view that CREB activation in the NAc by drugs of abuse mediates a form of tolerance and dependence to drug exposure. Viral-mediated overexpression of CREB in the NAc decreases an animal's sensitivity to the rewarding effects of several drugs of abuse, whereas blockade of CREB function in this region, via overexpression of a dominant negative antagonist of CREB, causes the opposite effect. Studies of bitransgenic mice, in which CREB or its dominant negative is inducibly expressed in the NAc of adult animals, as well as studies of CREB knockdown mice, generally support these conclusions. Interestingly, increased CREB activity in the NAc also reduces an animal's sensitivity to natural rewards as well as to several types of negative emotional (e.g., painful, stressful, or anxiogenic) stimuli. Thus, CREB activation in the NAc by drugs of abuse or stress seems to induce a general state of emotional numbing. CREB activation could therefore mediate some of the negative emotional symptoms seen in many drug addicts during early phases of withdrawal; it could similarly mediate some of the symptoms seen in depression and other stress-related disorders.

CREB activity in the VTA is more complex. Here, CREB can either decrease or increase an animal's sensitivity to a drug of abuse depending on the subregion of the VTA involved. Further work is needed to better understand these actions of CREB, as well as the influence of several other members of the CREB/CREM/ATF family in regulating responses to drugs of abuse and stress. There is preliminary evidence, for example, that drugs of abuse and stress also activate inducible cAMP early repressor (ICER) in the NAc upon initial exposures. ICER is a transcriptional repressor, encoded by the CREM gene. Consequently, ICER activation may serve to dampen the functional consequences of CREB induction during initial drug or stress exposures.

AP-1 Family

Fos and Jun proteins, like CREB proteins, are BZip transcription factors. Fos family members dimerize with Jun family members to form active AP-1 complexes, which then bind to AP-1 sites in gene promoters/enhancers. Genes encoding Fos and Jun proteins are immediate early genes, which means that they are induced very rapidly in response to an acute stimulus. Accordingly, acute administration of virtually any drug of abuse induces many Fos and Jun family members in the NAc. This induction is not only very rapid, it is also very transient, because the Fos and Jun mRNAs and proteins are all highly unstable. ΔFosB is a C-terminal truncated splice variant of the FosB gene. It, too, is induced rapidly in response to acute drug exposures. However, unlike all other Fos and Jun family proteins, ΔFosB is a relatively stable protein. This stability is mediated partly by the loss of the C-terminus – which contains degrons and is conserved in all other Fos family proteins – and by its phosphorylation. As a result, in response to repeated drug administration, levels of ΔFosB accumulate and

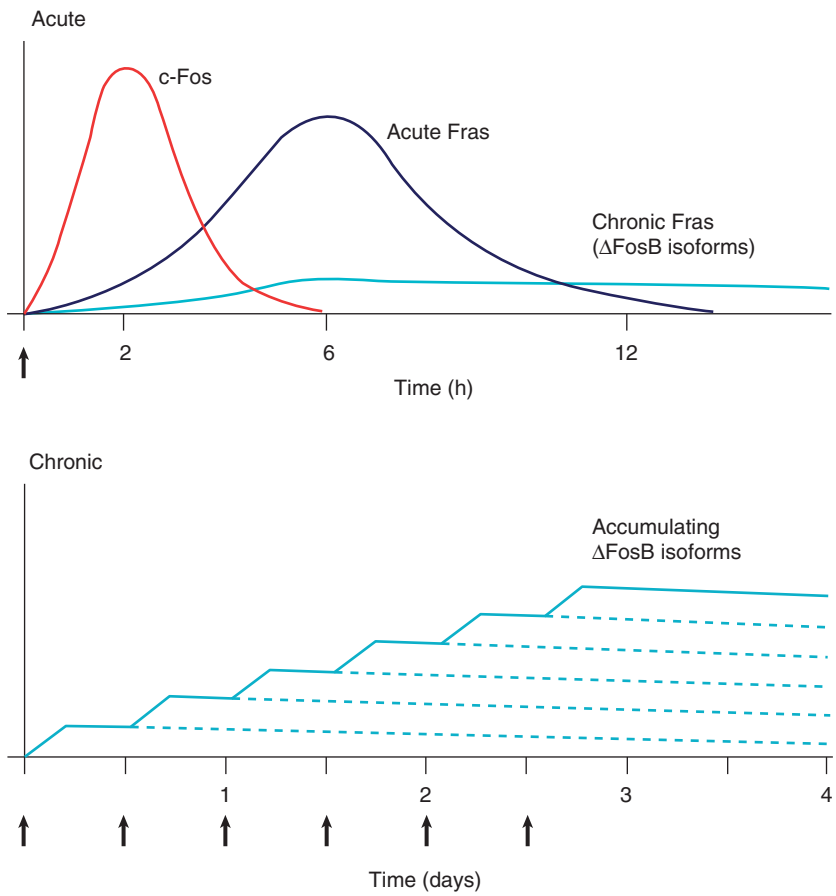

Figure 2 Regulation of ΔFosB by drugs of abuse. The figure shows the several waves of induction of Fos family proteins in the NAc after a single exposure to a drug of abuse. These proteins include cFos and several other Fos family proteins (FosB, Fra-1, Fra-2). Unlike all of these Fos family proteins, which are highly unstable, isoforms of ΔFosB are highly stable and therefore persist in brain long after drug exposure. Because of this stability, ΔFosB accumulates uniquely with repeated drug exposures. From Nestler EJ (2001) Molecular basis of long-term plasticity underlying addiction. *Nature Reviews Neuroscience* 2: 119–128, with permission.

eventually become the predominant Fos family protein in these neurons. This pattern of induction is illustrated in Figure 2. Induction of ΔFosB occurs as a common response to many classes of addictive drugs, including cocaine, amphetamine, methamphetamine, opiates, nicotine, PCP, cannabinoids, and alcohol. ΔFosB is also induced in the NAc after repeated consumption of natural rewards, like wheel running or sucrose intake. In addition, induction of ΔFosB in the NAc occurs as a consequence of chronic exposure to several types of stress.

Studies with inducible bitransgenic mice and viral vectors show that induction of ΔFosB in the NAc increases an animal's sensitivity to the rewarding effects of drugs of abuse, and also seems to increase incentive motivation, or drive, for the drugs. In addition, ΔFosB increases an animal's drive for natural rewards, such as wheel running and sucrose. Thus, ΔFosB may function as a sustained molecular switch, which first helps initiate and then maintains a state of addiction for a relatively prolonged period of time. An important area of current research is to characterize the effect of ΔFosB induction in the NAc in animal models of stress, depression, and antidepressant treatment.

There is early evidence that induction of ΔFosB in the NAc may mediate the ability of a drug of abuse to increase the dendritic branching, and the number of spines on distal dendrites, in NAc neurons. Although the precise functional role played by these morphological changes is not known, the changes have been shown to persist for at least several months after the last drug exposure, and are hypothesized to mediate the near-permanent sensitization in drug responsiveness seen in certain animal models of addiction. ΔFosB appears to be one mediator of these drug-induced changes in dendritic structure via its regulation of one particular target gene, cyclin-dependent kinase 5 (Cdk5). Thus, chronic cocaine administration induces Cdk5 in the NAc, an effect mediated via

ΔFosB, and infusion of a Cdk5 inhibitor into the NAc prevents the ability of cocaine to increase dendritic spine density in this region. The implication of these findings is that structural changes caused by repeated cocaine administration may be mediated, in part, via induction of ΔFosB and may persist long after the ΔFosB signal itself dissipates. While ΔFosB is the longest-lasting known molecular change in the brain seen in the context of drug or stress exposure, and perhaps to any other perturbation of the adult brain, it nevertheless undergoes proteolysis at some finite rate: ΔFosB dissipates to normal levels within a month or two of drug withdrawal. This means that ΔFosB cannot *per se* mediate the extremely long-lived changes in brain and behavior associated with addiction and depression. One possibility is that ΔFosB causes other changes in the brain, such as the changes described in dendritic structure, which themselves are more permanent.

Other Transcription Factors that Regulate Brain Reward Pathways

Recent work has shown that several other transcription factors are also regulated in brain reward regions by drugs of abuse or stress, although in general much less is known about the behavioral consequences of their induction compared with CREB and ΔFosB. The transcription factor, nuclear factor kappa B (NF-κB), has been the subject of intense investigation for its widespread involvement in a multitude of disease states and for its inducible pattern of gene expression. Several types of cellular stress (including cytokines, growth factors, viruses, and environmental hazards) induce NF-κB in peripheral cells. Under normal conditions, NF-κB, which is composed of a combination of two subunits, p50, p52, p65, c-Rel, and RelB (most abundantly p50 and p65), remains sequestered in the cytoplasm by inhibitory-κB (I-κB). Upon phosphorylation by I-κkinase (IKK), I-κB releases the inactive NF-κB dimer, which can then be phosphorylated and transported to the nucleus where it undergoes further posttranslational modifications to initiate transcription of a range of genes involved in cell survival. In the brain, NF-κB has received considerably less attention; however, it has been implicated in certain forms of learning-related plasticity. Although the involvement of NF-κB in addiction and depression models is not well studied, NF-κB has been implicated in the neurotoxic effects of amphetamine and amphetamine-like drugs. Furthermore, it has been shown that chronic cocaine administration upregulates the NF-κB subunits p65, p105 (precursor to p50), and I-κBγ and thereby increases NF-κB-dependent transcription in the NAc. Interestingly, cocaine induction of the NF-κB subunits is mediated by ΔFosB, which further broadens the role of this transcription factor in molecular control of the NAc. Studies are underway to better understand the functional consequences of NF-κB upregulation in drug and stress models.

Clock and other members of the basic helix-loop-helix-PAS (PER-ARNT-SIM) transcription factor family are best known for their control of circadian rhythms. The first evidence that this gene family may be involved in reward came from studies of *Drosophila*, where loss of Clock dramatically altered the flies' responses to cocaine. More recent work has demonstrated a role for Clock and other circadian genes in mammalian models of addiction. Thus, mice lacking Clock show dramatically increased activity of VTA dopamine neurons and increased behavioral responses to cocaine. Mice lacking the Clock homolog, NPAS2, or the Clock-NPAS2 target genes, Per (period), also show abnormal responses to the rewarding effects of cocaine. Work is underway to define the target genes through which these circadian genes act to regulate the VTA-NAc pathway, and to study the influence of these genes on stress models as well.

Several drugs of abuse, as well as stress, induce the release of systemic glucocorticoids, which activate glucocorticoid receptors expressed throughout the brain. These receptors, which are members of the nuclear receptor family of transcription factors, are known to be expressed in the VTA and NAc. Glucocorticoids are, therefore, thought to be one important mechanism by which stress regulates the reward pathway. Estrogen and progesterone receptors, members of the nuclear receptor family as well, are also expressed in the VTA-NAc pathway and have been found to regulate drug reward mechanisms. Such effects may contribute to gender differences observed in behavioral responses to drugs of abuse and stress. However, the target genes for glucocorticoid, estrogen, and progesterone receptors in the VTA and NAc, through which steroid hormones regulate reward mechanisms, have not been established.

Regulation of Chromatin Remodeling in Brain Reward Pathways

Binding of transcription factors to their target genes regulates transcription via chromatin remodeling events, which are becoming increasingly well understood. Interest in chromatin remodeling in the context of brain reward comes from two sources. First, studies of chromatin remodeling make it possible to delineate the detailed molecular events by which

drug- and stress-regulated transcription factors activate or repress target genes in the VTA and NAc *in vivo*. Second, such changes in chromatin structure provide a novel mechanism by which exposure to a drug of abuse or stress may cause lasting changes in gene expression that outlive changes in the transcription factors themselves.

The rate of expression of a particular gene is controlled by its location within nucleosomes and by the activity of the cell's transcriptional machinery. A nucleosome is a tightly wound span of DNA that is bound to histones and other nuclear proteins. Transcription of a gene requires the unwinding of a nucleosome, which makes the gene accessible to the basal transcription complex, comprised of RNA polymerase (which transcribes the new RNA strand) and numerous regulatory proteins (which unwind the nucleosomes). Transcription factors act by enhancing (or inhibiting) the activity of the basal transcription complex; this is achieved by altering nucleosomal structure through changes in histone aceylation, effects mediated by histone acetyltransferases (HAT) or histone deacetylases (HDAC). Transcription factors also recruit other regulatory proteins to the complex, which further modify chromatin structure, for example, through other posttranslational modifications of histones (or through methylation of the DNA itself). In addition, chromatin remodeling involves enzymes that reposition nucleosomes, in an ATP-dependent manner, and thereby further make genes accessible for transcription. The best-studied example is the SWI-SNF (mating switching and sucrose nonfermenting) complex.

Recent evidence has demonstrated the relevance of chromatin remodeling to drug-induced neuroadaptations within the brain's reward pathway. This work is made possible by use of chromatin immunoprecipitation (ChIP) assays, in which brain tissue (under control or drug-treated conditions) is fixed in formaldehyde to cross-link DNA to nearby histones and other proteins. The fixed chromatin is then sheared into smaller fragments and immunoprecipitated using an antibody directed against a protein of interest, and the immunoprecipitate is uncross-linked and then individual genes of interest are analyzed by quantitative real-time polymerase chain reaction (PCR). Alternatively, the immunoprecipitated DNA is amplified and analyzed on a promoter chip to obtain a genome-wide assessment of regulated genes. This approach makes it possible to identify changes that occur genome-wide within the brain reward circuit as a consequence of drug exposure.

Using ChIP assays, differential chromatin remodeling events (such as histone acetylation and deacetylation) have been observed at various cocaine-regulated genes. For example, it was found that chronic, but not acute, cocaine administration induces histone acetylation at the Cdk5 gene promoter in the NAc. This is consistent with earlier observations that cocaine induces Cdk5 expression in this brain region, but provides the first direct evidence that this induction is mediated via activation of the Cdk5 gene *per se*. Moreover, after chronic cocaine, ChIP assays revealed increased binding of ΔFosB to the Cdk5 gene promoter. Furthermore, inducible overexpression of ΔFosB in NAc of adult bitransgenic mice, which is sufficient for induction of Cdk5 expression *in vivo*, caused increased binding of ΔFosB with the Cdk5 gene. Chronic cocaine administration, or inducible overexpression of ΔFosB, also causes increased binding of Brg1 (a component of the SWI-SNF complex) to the Cdk5 gene. Together, these results support a scheme whereby the gradual accumulation of ΔFosB, in response to chronic cocaine administration, recruits chromatin remodeling factors, such as specific HATs to induce histone acetylation as well as Brg1-containing chromatin remodeling complexes, to its target genes such as Cdk5 (**Figure 3**).

Importantly, recent work has established the behavioral relevance of histone acetylation in the NAc. First, viral-mediated overexpression of particular HDACs in the NAc dramatically blocked the rewarding effects of cocaine. Conversely, treatment of animals with two structurally distinct HDAC inhibitors caused the opposite effect and increased an animal's sensitivity to the behavioral effects of cocaine. These HDAC inhibitors were also found to interact synergistically with cocaine to induce histone acetylation at responsive gene promoters. These findings demonstrate directly that histone acetylation can modify both biochemical and behavioral responses to cocaine, and implicate chromatin remodeling as a key mechanism underlying the molecular control of the brain's reward regions by drugs of abuse.

Future Research

Progress has been made in identifying transcription factors, for example, CREB, ΔFosB, NF-κB, Clock, and steroid hormone receptors, among others, that mediate the effects of drugs of abuse and stress on the reward pathway of the brain. Future research is needed to better understand how drugs and stress, presumably via regulation of dopamine and other receptors and of downstream intracellular signaling cascades in the reward pathway, lead to the regulation of these transcription factors, and how these factors induce chromatin remodeling events at particular target genes. ChIP assays now make it possible for the first time to identify the precise steps by which

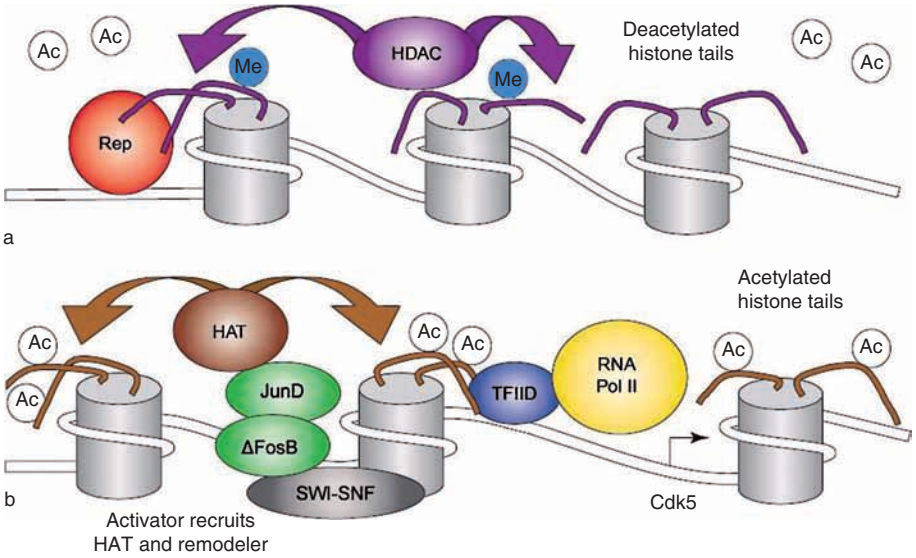

Figure 3 Scheme of proposed chromatin remodeling events at a cocaine-activated gene. Part (a) shows the repressed state of chromatin, where a site-specific repressor (Rep) recruits an HDAC complex, which removes acetyl groups (Ac) from histone N-terminal tails. Gene inactivation may also involve other modifications, such as methylation (Me) of histone tails. Part (b) shows the active state of chromatin around a cocaine-activated gene (e.g., Cdk5), where a cocaine-induced transcriptional activator (e.g., ΔFosB–JunD) recruits a HAT and chromatin remodeling complex (SWI-SNF), which induce acetylation (and perhaps demethylation and other modifications) of histone tails and repositioning of nucleosomes. These actions facilitate the binding of general transcription factors and the basal transcriptional apparatus (e.g., TFIID and RNA polymerase II [PolII]) to the promoter. From Colvis C, Pollock JD, Goodman RH, et al. (conference report) (2005) Epigenetic mechanisms and gene networks in the nervous system. *Journal of Neuroscience* 25(45): 10379–10389, with permission.

different transcriptional complexes are formed at each of the genes regulated in addiction and depression models, and ChIP on chip assays will help identify the complex gene regulatory networks affected by these transcription factors to regulate the reward circuitry. This new dimension of molecular research will dramatically expand our understanding of the regulatory control of brain reward regions, and provide fundamentally new paths for the discovery of novel treatments for these and other psychiatric disorders.

Acknowledgments

Preparation of this article was supported by the National Institute on Drug Abuse and the National Institute of Mental Health.

See also: Representation of Reward; Reward and Learning; Reward Systems: Human; Thyroid Hormone and Transcriptional Regulation in the CNS; Transcription Control and the Circadian Clock; Transcription Factors in Synaptic Plasticity and Learning and Memory; Transcriptional Networks and the Spinal Cord; Transcriptional Silencing.

Further Reading

Abarca C, Albrecht U, and Spanagel R (2002) Cocaine sensitization and reward are under the influence of circadian genes and rhythm. *Proceedings of the National Academy of Sciences of the United States of America* 99: 9026–9030.

Ang E, Chen J, Zagouras P, et al. (2001) Induction of nuclear factor κB in nucleus accumbens by chronic cocaine administration. *Journal of Neurochemistry* 79: 221–224.

Asanuma M and Cadet JL (1998) Methamphetamine-induced increase in striatal NF-kappaB DNA-binding activity is attenuated in superoxide dismutase transgenic mice. *Molecular Brain Research* 60: 305–309.

Carlezon WA Jr., Duman RS, and Nestler EJ (2005) The many faces of CREB. *Trends in Neuroscience* 28: 436–445.

Colvis C, Pollock JD, Goodman RH, et al. (conference report) (2005) Epigenetic mechanisms and gene networks in the nervous system. *Journal of Neuroscience* 25(45): 10379–10389.

Hyman SE and Malenka RC (2001) Addiction and the brain: The neurobiology of compulsion and its persistence. *Nature Reviews Neuroscience* 2: 695–703.

Kreibich AS and Blendy JA (2005) The role of cAMP response element-binding proteins in mediating stress-induced vulnerability to drug abuse. *International Review of Neurobiology* 65: 147–178.

Kumar A, Choi K-H, Renthal W, et al. (2005) Chromatin remodeling is a key mechanism underlying cocaine induced plasticity in striatum. *Neuron* 48: 303–314.

Marinelli M and Piazza PV (2002) Interaction between glucocorticoid hormones, stress and psychostimulant drugs. *European Journal of Neuroscience* 16: 387–394.

McClung CA and Nestler EJ (2003) Regulation of gene expression and cocaine reward by CREB and ΔFosB. *Nature Neuroscience* 6: 1208–1215.

McClung CA, Sidiropoulou K, Vitaterna M, et al. (2005) Regulation of dopaminergic transmission and cocaine reward by the Clock gene. *Proceedings of the National Academy of Sciences of the United States of America* 102: 9377–9381.

Nestler EJ (2001) Molecular basis of long-term plasticity underlying addiction. *Nature Reviews Neuroscience* 2: 119–128.

Nestler EJ (2005) Is there a common molecular pathway for addiction? *Nature Neuroscience* 8(11): 1445–1449.

Olson VG, Zabetian CP, Bolanos CA, et al. (2005) Regulation of drug reward by CREB: Evidence for two functionally distinct subregions of the ventral tegmental area. *Journal of Neuroscience* 25: 5553–5562.

Robinson TE and Kolb B (2004) Structural plasticity associated with exposure to drugs of abuse. *Neuropharmacology* 47(supplement 1): 33–46.

Ron D and Jurd R (2005) The ups and downs of signaling cascades in addiction. *Science's STKE* 1–16.

Russo SJ, Festa ED, Fabian SJ, et al. (2003) Gonadal hormones differentially modulate cocaine-induced conditioned place preference in male and female rats. *Neuroscience* 120: 523–533.

Uz T, Ahmed R, Akhisaroglu M, et al. (2005) Effect of fluoxetine and cocaine on the expression of clock genes in the mouse hippocampus and striatum. *Neuroscience* 134: 1309–1316.

Yao WD, Gainetdinov RR, Arbuckle MI, et al. (2004) Identification of PSD-95 as a regulator of dopamine-mediated synaptic and behavioral plasticity. *Neuron* 41: 625–638.

Transcription Control and the Circadian Clock

J Hirayama and P Sassone-Corsi, University of California at Irvine, Irvine, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Circadian (Latin, *circa diem*, 'about a day') rhythms are common to most organisms, from bacteria to human, and are driven by circadian clocks. These are intrinsic timekeeping systems that regulate various biochemical, physiological, and behavioral processes with a periodicity of about 24 h and enhance the efficiency and survival of organisms by allowing them to anticipate and adapt to changing environmental conditions.

In mammals, circadian rhythms are generated in pacemaker cells within the suprachiasmatic nuclei (SCN) of the hypothalamus. Pacemakers are composed of molecular oscillators responsible for the control of overt circadian rhythms in the animal's physiology and behavior. Circadian oscillators are endogenous, cell-autonomous systems that maintain their rhythm even in the absence of external time cues. Importantly, entrainment of the endogenous clock is achieved by specific cues, such as light. It is now well established that peripheral tissues also contain independent oscillators, the function of which is orchestrated by the central clock within the SCN. The circadian clock is constituted by three conceptual components. The first is the intrinsic pacemaker, dedicated to generating and sustaining endogenous rhythmicity. The second component is the input, which refers to the pathways through which external cues are perceived and act upon the central pacemaker. The third element is constituted by the output pathways, through which the circadian clock machinery affects a large array of metabolic and physiologic responses.

Central and Peripheral Clocks

Anatomically, the mammalian central pacemaker is located in the SCN, which generates, maintains, and controls circadian rhythmicity. Evidence that the SCN is the master circadian pacemaker originates from several studies. First, it was shown that ablation of the SCN by electrolytic or chemical lesions invariably resulted in the permanent loss of most overt circadian rhythms. This was later confirmed by SCN transplantation studies, in which grafted SCNs restored circadian rhythmic activity with a period specific to the genotype of the donor graft. Finally, it has been demonstrated that SCN neurons exhibit self-sustained, autonomous circadian rhythms even when isolated in culture.

In addition to the SCN, circadian oscillators also operate in almost all peripheral cell types. Studies in *Drosophila* and zebra fish have shown that individual tissues and immortalized culture cells generate their own autonomous oscillation. In the case of zebra fish, all cells display the remarkable feature of being light responsive. Mammals also have self-sustained circadian oscillators distributed throughout the body. Although no peripheral photoreception has been shown to occur in mammalian peripheral cells, the peripheral oscillators can respond to other signals. Indeed, SCN-derived signals operate in the synchronization of the peripheral oscillators, although with a somewhat different timing on diverse target peripheral cells.

Studies using transgenic rats have reported that the peripheral clocks can only sustain circadian oscillations for a few days when isolated from SCN. Based on this evidence, it was thought that peripheral clocks would require input stimuli from the SCN in order to maintain their circadian oscillation and that peripheral clock machineries become inactive in the absence of the SCN. Several studies have challenged this view and have recently demonstrated that single peripheral cells can function as independent circadian oscillators, basically similar to those operating in the SCN. Thereby, the current view is that the absence of the central clock leads to a loss in synchronicity among individual cells. The SCN, rather than being the unique pacemaker, acts as an 'orchestra's conductor' implicated in the synchronization of a resonant network, a web of peripheral oscillators.

The impact of food entrainment on peripheral clocks has important physiologic implications. In nocturnal animals, it has been demonstrated that restricting feeding to daytime elicits uncoupling of the circadian oscillators in peripheral tissues from the central SCN pacemaker. The restricted access to food advances the circadian cycles of gene expression in peripheral tissues without affecting the SCN, which remains faithful to its intrinsic timing system. These findings indicate that food restriction can superimpose the normal communications between the central clock and the peripheral oscillators.

Circadian Feedback Loop in Mammals

In mammals, the molecular mechanisms underlying the establishment and maintenance of biological rhythms comprise interconnected transcriptional–translational feedback loops. These loops rely on positive and negative regulatory elements. The basic

sprockets of the molecular wheel that control the mammalian central clock include a group of proteins: Period (PER) 1, PER2, and PER3; cryptochrome (CRY) 1 and CRY2; and CLOCK and BMAL1. CLOCK and BMAL1 heterodimerize to form an active transcriptional complex that binds to E-box elements (CACGTG) present in the promoters of Per and Cry genes. Once PER and CRY proteins have been translated, they form heterodimers that then translocate to the nucleus to repress CLOCK:BMAL1-mediated transcription through direct protein–protein interaction. It is important to note that the same activators, CLOCK:BMAL1, stimulate the transcription of many other clock-controlled genes (CCGs). These genes in turn influence functions external to the core oscillatory mechanism and mediate the 'output' of the clock. This accounts for the circadian expression of at least 10% of all cellular transcripts; this expression then results in the rhythmicity of a variety of physiological processes (see **Figure 1**).

Additional components regulate the robustness of this molecular clockwork circuitry: these include the bZIP transcription factor E4BP4, the PAR proteins (DBP, HLF, and TEF), and orphan nuclear receptors (Rev-ERB-α, ROR, RARα, and RXRα). It was shown that transcripts of the E4BP4 and PAR transcriptional factors display circadian oscillations with opposite phases in the SCN and liver. Moreover, transcription studies have demonstrated that all PAR proteins activate, whereas E4BP4 suppresses, the promoters of Per genes by direct binding to the same consensus sequence on the promoters.

Another important regulatory loop exists. The CLOCK:BMAL1 heterodimer transactivates the expression of the Rev-ERB-α gene whereas the CRY:PER heterodimer represses it. Once translated, the Rev-ERB-α protein represses in turn Bmal1 expression by direct binding to a consensus motif within the Bmal1 promoter. In addition, ROR, which is not under the circadian regulation, was shown to bind

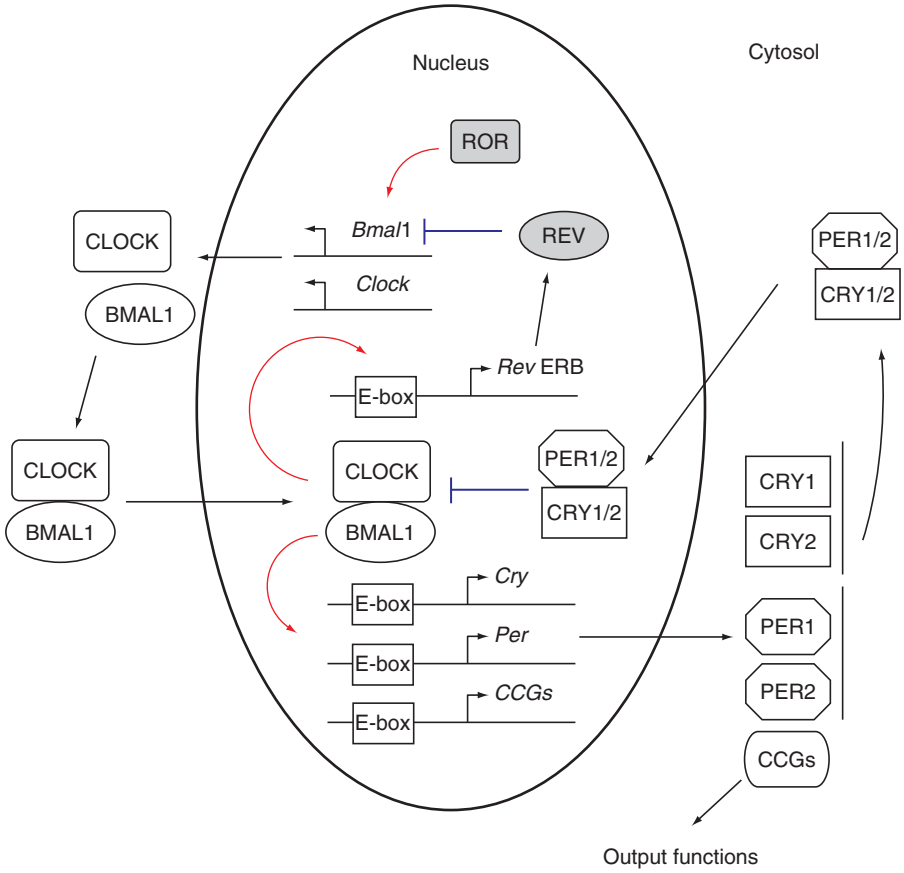

Figure 1 The mammalian circadian feedback loop. The feedback loop consists of positive and negative elements, the positive ones being two basic helix–loop–helix, Per–Arnt–Sim domain-containing transcription factors, CLOCK and BMAL1. These transcription factors heterodimerize and drive the transcription of negative components of the clock mechanism, specifically three *Period* genes (*Per1*, *Per2*, and *Per3*) and two *cryptochrome* genes (*Cry1* and *Cry2*). The products of these clock genes then negatively regulate their own expression, thereby setting up the rhythmic oscillations of gene expression that drive the circadian clock. CLOCK:BMAL1 also regulates clock-controlled genes (CCGs), and hence rhythmic output. The additional components, such as Rev-ERB and ROR, regulate Bmal1 transcription, forming an additional regulatory feedback loop required for the precision of the core oscillator. In this context, the discovery that CLOCK is a histone acetyltransferase and thereby directly modifies chromatin remodeling is of central importance.

to and activate directly the Bmal1 promoter. Thus, Rev-ERB-α and ROR regulate Bmal1 transcription in opposite manner, forming an additional regulatory loop that is important for the precision of the core oscillator.

It has also been reported that the retinoid receptors, RXRá and RARá, directly interact with CLOCK in a hormone-dependent manner, repressing CLOCK: BMAL1-mediated transcription by blocking E-box binding of the CLOCK:BMAL1 heterodimer. Importantly, several lines of evidence suggest that this retinoid receptor-mediated transcriptional inhibition may be the mechanism by which peripheral clocks can be regulated by humoral factors.

These findings strongly indicate that the control of the circadian feedback loop is complex and involves several yet unappreciated features. In addition to transcriptional positive and negative feedback loops, further levels of regulation and complexity operate in order to establish and maintain circadian rhythms. These include posttranscriptional regulation, posttranslational modifications, and chromatin remodeling.

Posttranscriptional Regulation

Posttranscriptional regulation has been found to be an essential element in the regulation of gene expression, especially among early response genes, such as c-*fos*. The 3′-untranslated regions (3′-UTRs) have been implicated in the regulation of mRNA stability, localization, and translational efficiency. Indeed, it has been reported that the 3′-UTRs of mouse Per1 (mPer1) mRNAs play crucial roles for the posttranscriptional mechanism producing the delay between transcriptional and translational products of genes.

A 196-nucleotide (nt) segment of the mPer1 3′-UTR was shown to exert a negative action on the translation of the mPER1 protein. Importantly, this 3′-UTR region does not affect the half-life of the mPer1 mRNA, showing that the degradation of the transcripts is not associated with this posttranscriptional mPer1 regulation. Although the detailed mechanism of the mPer1 3′-UTR-dependent transcriptional inhibition remains to be determined, these observations raise the possibility that putative regulator(s) such as RNA-binding protein(s) may intervene in the control of mPer1 expression by interacting with the 196-nt region. In support of this idea, the expression of several genes encoding RNA-binding proteins is reported to be under the circadian regulation. The case of mouse *Per3* gene is also relevant: it has been reported that the 3′-UTR of mPer3 controls mRNA stability, its presence inducing mRNA decay. Importantly, the 3′-UTR-mediated mRNA decay has been shown to play an essential role in the circadian oscillation of mPer3 transcript.

Posttranslational Regulation of CLOCK and BMAL1

Clock proteins undergo a number of posttranslational modifications that contribute to their physiological function (see Table 1). Both CLOCK and BMAL1 display circadian oscillation in their phosphorylation status and in their intracellular localization in cells from mouse liver as well as in neurons of the SCN. Apparently it is the dimerization of CLOCK:BMAL1 that induces the phosphorylation, facilitates their nuclear localization, and finally leads to degradation of each protein. Moreover, CRYs were shown to increase the nuclear accumulation and the half-life of unphosphorylated CLOCK:BMAL1 complex.

BMAL1 can be phosphorylated *in vitro* by two kinases, casein kinase I (CKI) and the mitogen-activated protein kinase (MAPK) extracellular signal-related kinase (ERK). MAPK/ERK phosphorylates three residues, Ser527, Thr534, and Ser599 (in mouse BMAL1), residues which are all highly conserved among vertebrates. On the other hand, the CKI phospho-acceptor sites of BMAL1 have not been identified to date. Transactivation studies have shown that MAPK/ERK-mediated phosphorylation of Thr534 results in a decrease of CLOCK:BMAL1 transcription activation potential, whereas CKI-induced phosphorylation appears to stimulate it.

In addition to phosphorylation, BMAL1 is also subjected to modification by the small ubiquitin-related modifier protein (SUMO). Interestingly, BMAL1 undergoes rhythmic SUMOylation *in vivo*, with a timing that parallels BMAL1 circadian activation. Moreover, BMAL1 SUMOylation was shown to play an important role in controlling BMAL1 protein half-life, thus directly influencing clock rhythmicity.

CLOCK is also a phosphoprotein, although to date little is known about the signaling pathways leading to its phosphorylation. The Ser-Pro-rich region of CLOCK appears to be a target of the histone H1 kinases, a group of three kinases exhibiting activity that intriguingly fluctuates in a periodic manner in nuclear extracts from SCN neurons.

BMAL1 contains a highly conserved functional nuclear localization signal (NLS) and a nuclear export signal (NES) in its N-terminal and Per-Arnt-Sim (PAS) domains, respectively. These signals were shown to control nucleocytoplasmic shuttling of BMAL1, which is critical for the circadian function of the CLOCK: BMAL1 heterodimer. Similarly, also CLOCK contains the highly conserved putative NES and NLS. How these signals exert their function still needs to be determined, but it should be noted that studies on PER proteins (see later) indicate that interactions with BMAL1 and CRYs, and their phosphorylation,

Table 1 Modifications of clock proteins

Protein	*Modification (site)*	*Inducer*	*Comment*
CLOCK	Phosphorylation (unknown)	BMAL1 and CRYs induce the phosphorylated and unphosphorylated forms of CLOCK, respectively	The phosphorylated form of CLOCK seems to be transcriptionally active The phosphorylated form of CLOCK is localized in the nucleus, thus phosphorylation may regulate its subcellular localization
	Phosphorylation (the Ser-Pro-rich region)	H1 kinase	The function of this phosphorylation is unknown
BMAL1	Phosphorylation (unknown)	CLOCK and CRYs induce the phosphorylated and unphosphorylated forms of BMAL1s, respectively	The phosphorylated form of BMAL1 seems to be transcriptionally active
	Phosphorylation (unknown)	CKIε	This phosphorylation stimulates CLOCK: BMAL1-dependent transcription
	Phosphorylation (Ser527, Thr534, and Ser599 of mouse BMAL1)	MAPK	Thr534 phosphorylation inhibits CLOCK: BMAL1-dependent transcription Function of the Ser527 and Ser599 phosphorylations is unknown
	SUMOylation (Lys259 of mouse BMAL1))	CLOCK induces the SUMOlylation of BMAL1	This modification regulates BMAL1 protein half-life
PER1	Phosphorylation (amino acids 653–663 of mPER1)	CKIε	This phosphorylation is required for nuclear translocation of PER1
	Phosphorylation (amino acids 902–916 of mPER1)	CKIε	This phosphorylation masks PER1 NLS, inducing its cytoplasmic localization
PER2	Phosphorylation (β-TrCP recognition motif)	CKIε	This phosphorylation is required for β-TrCP binding and for subsequent mPER2 degradation
	Phosphorylation (unknown)	GSK-3β	This phosphorylation induces nuclear localization of PER2
PER3	Phosphorylation (amino acids 613–626 of mPER3)	CKIε	This phosphorylation induces mPER3 degradation and nuclear translocation
CRY1	Phosphorylation (unknown)	CKIε	The function of this phosphorylation is unknown PER acts as a scaffold for this phosphorylation, allowing CKIε and CRY interaction
	Phosphorylation (Ser247 of mCRY1)	MAPK	This phosphorylation reduces its transcriptional inhibition activity
CRY2	Phosphorylation (Ser265 of mCRY2)	MAPK	This phosphorylation reduces its transcriptional inhibition activity
	Phosphorylation (Ser557 of mCRY2)	MAPK	The function of this phosphorylation is unknown The amount of the Ser557-phosphorylated mCRY2 displays robust circadian variation, and the Ser557-phosphorylated form of mCRY2 is localized specifically in the nucleus
	Phosphorylation (Ser553 of mCRY2)	GSK-3β	This phosphorylation induces mCRY2 degradation The Ser557 phosphorylation of mCRY2 is required for this phosphorylation

represent regulatory steps that mask or unmask the localization signals, thus affecting the subcellular distribution of these proteins.

Posttranslational Regulation of PER and CRY

There are three PER proteins in mammals: PER1, PER2, and PER3. The combinatorial dimer associations of PER proteins generate distinct effects on their subcellular localization and posttranslational modifications. For example, studies in cultured cells have shown that the mPER1:mPER2 interaction imposes a shift in mPER1 subcellular localization from the nucleus to the cytoplasm. On the other hand, a serum shock induces the heterodimerization of mPER1 or mPER2 with mPER3, which in turn facilitates the nuclear translocation of the dimer. It has been also shown that rhythmic changes in mPER3 subcellular localization and phosphorylation are highly dependent on the presence of mPER1.

The functional relationship between PER and CKI reveals the crucial role played by phosphorylation in mammalian clock function. A major advance has been the identification of the gene responsible for the tau mutation in the Syrian golden hamster. This is a spontaneous mutation which causes an early onset of locomotor activity under light/dark (LD) conditions and a very short 20 h period length. It was revealed that the tau locus encodes for CKIε, the kinase that phosphorylates the mammalian PER proteins. The tau mutation is caused by a point mutation leading to a deficiency in the ability of CKIε to phosphorylate its targets, which then results in the short-period phenotype. Importantly, a deficient phosphorylation of PER by CKIε has also been implicated in human sleep disorders, such as the familial advanced sleep phase syndrome (FASPS) and the delayed sleep phase syndrome (DSPS).

CKIε specifically interacts with PER1, PER2, and PER3, and regulates their half-life and intracellular localization. Indeed, CKIε-mediated phosphorylation induces PER protein degradation via the ubiquitin–proteasome pathway. Specifically, phosphorylation of PER2 by CKIε leads to recognition by the ubiquitin ligase adapter protein, β-transducer repeat-containing protein (β-TrCP). Polyubiquitination follows β-TrCP binding to PER2 and leads to subsequent degradation of the protein by the 26S proteasome.

In addition to determining PER proteins stability, CKIε specifically controls the subcellular distribution of PER1 and PER3 by masking or unmasking their localization sequences. All PERs contain a functional NLS adjacent to the CKIε-binding domain. In addition to the NLSs, PERs have additional localization sequences, such as the NES and a cytoplasmic localization domain (CLD). In the case of mPER1, CKIε is able to direct both nuclear and cytoplasmic distribution by specifically phosphorylating distinct residues.

CKIε also phosphorylates CRYs, although the biological significance of this event remains to be determined. CKIε does not directly bind to CRYs, although it seems that PER proteins act as a scaffold, bringing CKIε and CRY proteins into close proximity and thereby allowing phosphorylation. In addition to CKIε, CRYs are phosphorylated by two of the other kinases, MAPK/ERKs and the glycogen synthase kinase-3β (GSK-3β). ERKs directly interact with and phosphorylate the highly conserved residues Ser247 of mCRY1 and the corresponding residue Ser265 of mCRY2 *in vitro*, inducing a reduction in CRY repressor activity. ERKs also phosphorylate the conserved Ser557 in the C-terminal domain of mCRY2. Phosphorylation of mCRY2 at Ser557 displays robust circadian variation in the liver, with a peak at late night. The Ser557-phosphorylated form of mCRY2 is localized specifically in the nucleus, where it is degraded by the proteasome following a subsequent phosphorylation at Ser553 by GSK-3β.

The intracellular localization of the CRY proteins is a major determinant of their regulatory function. Functional NLSs have been described in both mCRY1 and mCRY2. Mutations introduced in the NLS (aa 265–282) of mCRY1 shift the subcellular localization from the nucleus to the cytoplasm. A CLD (aa 1–125) is also present in mCRY1, which presumably controls cytoplasmic localization. This may be regulated by PER1, which is thought to serve as an export adaptor for CRY1 and CRY2, challenging the view by which PER:CRY dimerization promotes nuclear accumulation of the complex. Yet, the PER:CRY physical association appears to indeed play a role in the nuclear accumulation of both proteins. In fact, mPER1 and mPER2 remain cytoplasmic in liver cells from mCRY-deficient mice, and vice versa for CRYs proteins in mPER-deficient mice. The cooperative nature of PER–CRY interactions is further substantiated by the observation that PER:CRY dimerization stabilizes both proteins by preventing their ubiquitination–proteasome-mediated degradation.

Chromatin Remodeling

Distinct posttranslational modifications, such as acetylation, phosphorylation, and methylation, occur on histone tails. These are primary events contributing to the dynamic process of chromatin remodeling. Specifically, histone acetylation plays an important role in promoting a conformational change that provides to chromatin a transcription-permissive remodeling state associated with activation of gene expression. It was shown that CLOCK:BMAL1-dependent transactivation of CCGs is coupled to circadian changes in histone H3 acetylation at their promoters. A number of transcriptional coactivators, such as CREB-binding protein (CBP), activin receptor (ACTR), and p300, are known to contain histone acetyltransferase (HAT) activity, and they were shown to interact with the CLOCK:BMAL1 heterodimer. Moreover, transcription assays demonstrated that CBP, p300, and ACTR are capable of enhancing CLOCK:BMAL1-dependent transcription. Therefore, it was believed that the CLOCK:BMAL1 transcription mechanism is dependent on the circadian time-specific recruitment of distinct coactivators.

The recent finding that CLOCK has intrinsic HAT activity has opened novel avenues of research and has multiple physiological implications. The primary structure of CLOCK shares homology with ACTR, a member of steroid receptor activator (SRC) family of HATs. Intriguingly, however, the acetyl-coenzyme

A (acetyl-CoA) binding motif essential to HAT function displays sequence similarity to the MYST family of HATs. Although BMAL1 does not possess HAT activity, it enhances HAT activity of CLOCK. Mutations in the acetyl-CoA binding motif of CLOCK reduce its HAT activity, diminishing also CLOCK:BMAL1-dependent transcription. These findings demonstrate that the intrinsic HAT activity of CLOCK is essential for the circadian transcriptional regulation. Importantly, various types of HATs have been shown to associate in a large complex *in vivo* with other coactivators. It is highly likely that CLOCK may associate with other nuclear proteins that could contribute to chromatin remodeling and circadian control. Indeed, the concerted contribution of the HAT function of CLOCK and other coactivators is likely to integrate the complex signaling governing chromatin remodeling and the circadian clock machinery.

Phenotypes of Circadian Clock Mutant Mice

A number of mutant mice for genes involved in circadian clock regulation exist, mostly obtained by targeted homologous recombination. The phenotypes of these mice are summarized in Table 2. Studies from these mutant mice have demonstrated the detailed and distinct roles of core clock proteins in regulation of circadian clock and light signaling pathways. Moreover, they also reveal the direct link between the circadian clock and various aspects of animal physiology.

Circadian Clock Input Pathways

Light has a profound influence on animal circadian behavior in two ways. First, light resets the circadian clock by its phase-shifting properties. Notably, the phase-shifting effects of light only occur during the nighttime period of the circadian cycle. In nocturnal mammals kept in darkness, a light pulse during the subjective night (that is, the time of the day corresponding to the dark period in a normal light–dark cycle) can reset the clock by evoking changes in the SCN-controlled rhythms. If the light pulse is given at an early point in time during the subjective night, it induces a shift to later times (phase delay). In the opposite way, if the light pulse is provided at the end of the subjective night, the SCN-controlled rhythms will be shifted to an earlier position in the circadian cycle (phase advance). Second, light acutely suppresses locomotor activity, also referred to as a 'masking' effect. The mechanism of masking is independent of the circadian clock. It has been recently shown that dopamine D2R receptor-mediated signaling is an essential component of the neuronal pathway involved in the light-induced masking effect.

Photic signals perceived by the retina are conveyed to the SCN through the retinal–hypothalamic tract (RHT). Glutamate has been identified as the major neurotransmitter responsible for transducing the photic information to the SCN along the RHT. Once glutamate is released by the SCN, it binds *N*-methyl-D-aspartate (NMDA) receptors, which in turn leads to Ca^{2+} influx that is finally responsible for the activation of nitric oxide synthase (NOS) and calcium/calmodulin-dependent protein kinase (CaMK).

Evidence suggests that the MAPK/ERK pathway is intimately implicated in the light input system. Mice exposed to light pulses during their subjective night display a rapid ERK upregulation (phosphorylation) in the SCN. Furthermore, disruption of the MAPK pathway has been shown to block light-induced phase shifting of the circadian clock at the behavioral level. This finding suggests that the ERK cascade is integrally involved in photic entrainment of mammalian rhythms.

Downstream events of the light-induced signaling pathway in the SCN lead to the phosphorylation of cAMP response element-binding protein (CREB), which then stimulates expression of genes containing a calcium/cAMP response element (CRE) in their promoters. These genes include the immediate-early genes c-*fos*, *fos-B*, *jun-B*, *egr1*, and, more importantly, the clock components *Per1* and *Per2*. In addition to CRE-dependent transcription, the CLOCK:BMAL1 complex was also shown to be involved in light-induced activation of these genes. Although the exact mechanism by which light induces early gene expression still needs to be elucidated, it has been shown that a simple light pulse engenders chromatin remodeling via phosphorylation of histone H3 at Ser10 (see Figure 2).

Circadian Photoreceptors

Rods and cones provide the visual information within the retina. These cells, however, are not involved in circadian photoreception. Indeed, rodents that lack classical visual responses are still capable of circadian photoentrainment. Retrograde tracing experiments have identified retinal cells projecting to the SCN through the RHT, but not to the visual centers of the brain. These cells constitute a small subset of retinal ganglion cells (RGCs), and they were shown to display intrinsic phototransduction, with photic properties matching those of clock entrainment.

Two main candidates have been proposed to be the circadian photoreceptors, melanopsin and cryptochromes. Melanopsin is an opsin found in the eye and

Table 2 Circadian phenotypes of mutant mice

Name	*Phenotypes*	*Description*	*Name*	*Phenotypes*	*Description*
Clock (c/c) mutant	Lengthens the circadian period by 4 h, followed by a complete loss of circadian rhythmicity in constant dark (DD)	The c/c mutant has A–T nucleotide transversion in a splice donor site of its allele, which causes the skipping of exon 19, resulting in the deletion of 51 amino acids of CLOCK protein	MOP3 (BMAL1) KO	Displays immediate and complete loss of circadian rhythmicity in constant darkness	
	Displays reduced levels and noncycling expression of clock-controlled genes	The mutant protein can form a dimer with BMAL1 and binds to DNA, but has defective transcriptional activity functioning as a dominant negative regulator		Displays altered distribution of activity during light–dark (LD) cycles and reduced total activity level	
	Displays reduced light induction of immediate-early genes in SCN			Shows reduced levels and noncycling expression of clock-controlled genes	
	Shows the normal spontaneous firing rhythms of SCN neurons, indicative that the loss of circadian rhythmicity might be due to an uncoupling of individual oscillators			Physiology: has reduced life spans with various symptoms of premature aging	
	Physiology: attenuates diurnal feeding rhythm and develops obesity and metabolic syndrome			BMAL1-deficient embryonic fibroblast cells are not capable of differentiating into adipocytes, indicative that BMAL1 plays important roles in the regulation of differentiation	
	Develops salivary gland hyperplasia when it becomes aged			Displays reduced levels of B cells in the peripheral blood, spleen, and bone marrow	
Clock KO	Exhibits almost normal circadian patterns of behavior	The data from Clock KO mice support the notion of the role of CLOCK-del.19 as dominant negative regulator	PER2 mutant	Displays a shorter circadian period followed by a loss of circadian rhythmicity in constant dark	Mutant expresses PAS domain-deleted form of mPER2 protein
	Shows altered responses to light, with reduced phase delays and exaggerated phase advances			Exhibits a defect in circadian clock resetting (unable to delay the clock)	
	Displays altered clock-controlled gene expression (The effects of CLOCK deficiency on gene expression are target gene specific: e.g., the circadian oscillation of *Per2* in the SCN and liver of CLOCK-deficient mice is robust, whereas that of DBP is completely impaired)			Physiology is cancer prone and shows increased sensitivity to γ-radiation and tumor development	
				Lack of food anticipation	

Continued

Table 2 Continued

Name	*Phenotypes*	*Description*	*Name*	*Phenotypes*	*Description*
PER1 KO	Displays a shorter circadian period in constant dark		PER3 KO	Has normal circadian patterns of behavior	Per3/Per1 double, Per3/Per2 double, and Per3/Per1/Per2 triple KO mutants have same circadian phenotypes as those from Per1, Per2, Per1/2 double KOs, respectively, demonstrating that the role of mPER3 is limited
	Exhibits a defect in circadian clock resetting (unable to advance the clock)				
PER1 and 2 double KO	Displays a complete loss of circadian rhythmicity in constant darkness		CRY KOs	Mice lacking the *Cry1* or *Cry2* genes exhibit accelerated or delayed free-running periodicity of locomotor activity, respectively CRY1 and CRY2 double KO mice display immediate and complete loss of circadian rhythmicity in constant darkness	
CRY1-2–PER2 double KO	Shows normal circadian behavior in LD condition Loses circadian rhythmicity immediately upon release into DD condition		CRY1–PER2 double KO	Displays normal circadian behavior in both LD and DD conditions	These data show that mCry2 can act as a nonallelic suppressor of mPer2
CRY1, CRY2, and rd triple mutant	Arrhythmic under both LD and DD cycles	The rd mutant has total loss of rods and massive loss of cones, rendering the animal visually blind by 3 months of age	D2R receptor KO	Exhibit normal circadian rhythmicity	The result from D2R KO indicates that D2R-mediated signaling is essential for light-masking of circadian rhythms
	Display reduced c-*fos* light inducibility in the SCN compared with either rd/rd or CRY1 and CRY2 mutant mice	The rd mutant exhibits normal free-running rhythm both in LD and DD cycles		Shows a deficient light-masking response	
Melanopsin KO	Entrains to a LD cycle and displays normal circadian activity rhythms under constant darkness Is defective for phase resetting in response to brief pulses of light Phase-shifts normally after a light pulse	Inconsistent results reported by two studies on melanopsin KO mice regarding the phase resetting	Melanopsin and rd mutant	Exhibits complete loss of pupil reflex, entrainment to light–dark cycles, and masking response to light	The rd mutant lacks functional rod and cone, thus the result demonstrates that the rod–cone and melanopsin systems together provide all of the photic input for these nonvisual functions

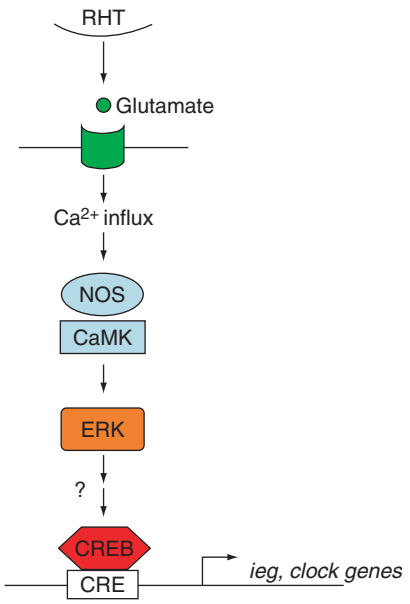

Figure 2 Light signaling evoking changes in the SCN-controlled rhythms. Photic signals perceived by the retina are conveyed to the SCN through the retinal–hypothalamic tract (RHT). RHT neurons release glutamate in synapses of the ventrolateral SCN. Glutamate binds and activates NMDA receptors, which then initiates a cascade of events leading ultimately to induction of immediate-early genes (ieg) and clock genes (such as *Per1*) (NOS, nitric oxide synthase; CaMK, calcium/calmodulin-dependent protein kinase). The gene induction is dependent on the activation of MAPK, extracellular signal-regulated kinase (ERK), and the transcription factor cAMP response element-binding protein (CREB).

in other photoreceptive structures in amphibians, and exclusively in the eye in primates and rodents. Although studies using melanopsin-null mice did not provide clear evidence that melanopsin plays a central role in circadian photoentrainment, several features of melanopsin would make it a good candidate for being the circadian photoreceptor. Its distribution is restricted to a subset of RGCs that project to the SCN and mediate entrainment. Moreover, all photoreceptive RGCs contain melanopsin, whereas RGCs lacking light response are devoid of this protein. Melanopsin-deficient mice entrain perfectly well to an LD cycle, whereas phase-shifting effects of light may be altered. Importantly, mice lacking rods, cones, and melanopsin exhibit complete loss of photoentrainment of the circadian oscillator as well as lack of pupillary light responses, photic suppression of melatonin production, and masking. The presence of both rod-cone and melanopsin systems appears to be important to mediate the light input.

Genetic evidence supports the possibility that CRYs have photoreceptive properties in mammals. CRY2-deficient mice present some defects in light-dependent Per gene induction in the SCN. In addition, mice devoid of rods, cones, and CRYs are nearly arrhythmic in light–dark conditions and display a severely blunted c-*fos* gene response to light in the SCN. More evidence supporting involvement of CRYs in photoreception, and excluding the participation of a vitamin A-containing chromophore, comes from work done in vitamin A-deficient mice lacking detectable opsins.

It is clear that none of the mammalian photopigments identified to date (rod and cone opsins, melanopsin, or CRYs) is uniquely essential for circadian phototransduction. It is conceivable that, independently of the classical opsins, CRYs and melanopsin work cooperatively to play a role in circadian photoreception as well as in the other nonvisual tasks.

Circadian Output

Circadian clocks regulate a number of physiological functions, including sleep and wakefulness, food intake, body temperature, cardiovascular and renal activity, hormone production, hepatic metabolism, and immune activity. Indeed, disruption of circadian rhythms has profound effects on human health, leading to insomnia, jet lag, stomach ailments, cancer, and depression.

The SCN master clock regulates behavioral circadian rhythms by diffusible signals, such as transforming growth factor-α (TGF-α) and the protein prokineticin 2 (PK2). These were shown to be rhythmically expressed in the SCN, and *in vivo* data demonstrated that they play a role in inhibiting locomotor activity of mice. While additional humoral signals are likely to exist and still await identification, it is apparent that a finely tuned connection between the SCN and target neurons also seems to be essential for a full transmission of the signal from the central clock to peripheral tissues.

Transcription of clock-controlled genes in peripheral clocks is critical for the regulation of local physiology and hormonal production. A good example is the gene encoding D-element binding protein (DBP), whose expression is directly under CLOCK:BMAL1 transcriptional control in the liver. DBP, in turn, regulates the rhythmic transcription of key enzymes involved in hepatic metabolic processes, which include Cyp7a, a key factor in bile acid production, and Cyp2a4, a protein involved in steroid metabolism.

Several lines of evidence recently have revealed a direct link between circadian rhythm and cancer. An early study indicated that Per2-deficient mice are cancer prone. These mice display a marked increase in spontaneous and γ-irradiation-induced tumor development. Later it was shown that CLOCK:BMAL1 directly regulates cell cycle regulator WEE-1, a kinase that controls the G2-to-M transition, restricting entry of dividing cells into mitosis to certain times of the

day. More recently, it has been reported that a number of clock-regulating proteins, such as CRYs, timeless (TIM) and PERs, interact directly or indirectly with cell-cycle-controlling proteins, including checkpoint proteins CHK1 and CHK2, ATR, and ATM, involved in cellular DNA damage responses. These observations provide essential information toward the development of new strategies for cancer chronotherapy. Indeed, administration of cancer therapy based on circadian timing has shown encouraging results, but still lacks a solid mechanistic foundation. Future studies will fully unravel the molecular links between the circadian clock and its control over central cellular processes such as cell proliferation, apoptosis, and response to genotoxic stress.

See also: Circadian Organization; Circadian Rhythm Models; Circadian Oscillations in the Suprachiasmatic Nucleus; Circadian Regulation by the Suprachiasmatic Nucleus; Circadian Gene Expression in the Suprachiasmatic Nucleus; Circadian Metabolic Rhythms Regulated by the Suprachiasmatic Nucleus; Circadian Genes and the Sleep–Wake Cycle; Clock Gene Regulation of Endocrine Function; Clock Genes and Metabolic Regulation; Entrainment of Circadian Rhythms by Light; Melatonin Regulation of Circadian Rhythmicity in Vertebrates; Photoreceptors and Circadian Clocks; Single Cell Neuronal Circadian Clocks; Thyroid Hormone and Transcriptional Regulation in the CNS; Transcription and Reward Systems; Transcriptional Networks and the Spinal Cord; Circadian Systems: Evolution.

Further Reading

Belden WJ, Loros JJ, and Dunlap JC (2006) CLOCK leaves its mark on histones. *Trends in Biochemical Science* 31: 610–613.

Cermakian N and Sassone-Corsi P (2000) Multilevel regulation of the circadian clock. *Nature Reviews Molecular Cell Biology* 1(1): 59–67.

Cermakian N and Sassone-Corsi P (2002) Environmental stimulus perception and control of circadian clocks. *Current Opinion in Neurobiology* 12(4): 359–365.

Devlin PF and Kay SA (2001) Circadian photoperception. *Annual Review of Physiology* 63: 677–694.

Doi M, Hirayama J, and Sassone-Corsi P (2006) Circadian regulator CLOCK is a histone acetyltransferase. *Cell* 125(3): 497–508.

Dunlap JC (1999) Molecular bases for circadian clocks. *Cell* 96(2): 271–290.

Fu L and Lee CC (2003) The circadian clock: Pacemaker and tumour suppressor. *Nature Reviews Cancer* 3(5): 350–361.

Hastings MH, Reddy AB, and Maywood ES (2003) A clockwork web: Circadian timing in brain and periphery, in health and disease. *Nature Reviews Neuroscience* 4(8): 649–661.

Hirayama J and Sassone-Corsi P (2005) Structural and functional features of transcription factors controlling the circadian clock. *Current Opinion in Genetics & Development* 15(5): 548–556.

Hirota T and Fukada Y (2004) Resetting mechanism of central and peripheral circadian clocks in mammals. *Zoology Science* 21(4): 359–368.

King DP and Takahashi JS (2000) Molecular genetics of circadian rhythms in mammals. *Annual Review of Neuroscience* 23: 713–742.

Pando MP and Sassone-Corsi P (2001) Signaling to the mammalian circadian clocks: In pursuit of the primary mammalian circadian photoreceptor. *Science STKE* 2001(107): RE16.

Schibler U and Sassone-Corsi P (2002) A web of circadian pacemakers. *Cell* 111(7): 919–922.

Transcription Factors in Synaptic Plasticity and Learning and Memory

C M Alberini, Mount Sinai School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Transcription factors are proteins that bind specific sites or elements in regulatory regions of DNA, known as promoters or enhancers, where they control the transcription or expression of target genes. Transcription factors can be selectively activated or deactivated by other proteins, often as the final step in signal transduction. There are three functional classes of transcription factors: (1) general transcription factors, which are ubiquitous and represent the core machinery of transcription; the most common are abbreviated as TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH; (2) constitutively expressed factors that in each cell type constitutively stimulate or repress transcription; and (3) inducible transcription factors, which are similar to the constitutively expressed transcription factors but require activation or inhibition. Transcription factors are composed of two essential functional regions: a DNA-binding domain and an activator domain. The DNA-binding domain consists of amino acids that recognize specific DNA bases on the regulatory element.

In the last 15 years, there has been great advances in understanding the nature and function of gene expression in the brain and its regulation underlying behavioral responses and, in particular, cognitive functions. However, a great deal remains to be understood. How does gene expression allow complex cognitive responses? What types of genes mediate such 'evolved' functions and how do they work?

Interestingly, the knowledge of the genome sequences of several species and their comparative analyses have revealed a surprising simplicity in the evolution of the genetic content. It is now clear that the number of genes in the genome of the different species does not account for the evolutionary increase in morphological and behavioral complexity: the genome in *Caenorhabditis elegans* is approximately 20 000 and the number of protein-coding genes in the human genome is only about 30 000. Moreover, this small increase is primarily due to duplication of existing genes, rather than creation of new ones. Where then does evolutionary complexity lie? And, particularly, how does the human genome, which is not very different than that of a worm, account for the complexity of human behavior? It has been proposed that behavioral complexity correlates with the number of 'gene expression patterns' exhibited throughout an animal's life cycle. Accordingly, 5–10% of the total coding capacity of a metazoan genome is made up of transcription factors, and humans contain as many as 3000 transcription factors in their genome. Remarkably, the combinatorial nature of transcription regulation produces a spectacular amplification of the repertoire of gene expression, and the brain is the tissue in which changes in protein and gene expression are most pronounced.

Here we will briefly review the role of some transcription factors in synaptic plasticity and learning and memory.

A number of investigations, in the last 15 years, examined the role of transcription factors in long-term synaptic plasticity and memory formation. The transcription factor families most extensively studied include cyclic AMP (cAMP) response element-binding protein (CREB), CCAAT C/EBP, AP-1, Zif 268, and nuclear factor kappa B (NF-κB).

The first idea that transcription regulation is essential for learning and memory came from studies carried out in the 1960s, when it was shown that inhibitors of transcription, such as actinomycin D, block memory formation. Subsequently, with the discovery of long-term synaptic plasticity responses (such as long-term facilitation (LTF) in the invertebrate *Aplysia californica*, long-term potentiation (LTP), and long-term depression (LTD) in mammals), it was also found that transcription is similarly required for long-term synaptic plasticity, but not for short-term responses. Long-term synaptic plasticity refers to long-lasting changes that occur in the strength of synaptic connections as a consequence of neuronal activation processes, and it has been suggested that these processes underlie learning and memory formation. However, more recent knowledge indicates that LTP and LTD have a broader role and contribute to experience-dependent modifications of brain function in general. Therefore, in light of this updated view, the description of the role of transcription factors in these processes is summarized. Obviously, as described below, similar transcription factors have been found to play a critical role in both LTP/LTD and learning and memory.

CREBs, The Most Studied

In the late 1980s, it was discovered that the activation of the intracellular second messenger cAMP affects gene expression via the activation of transcription factors

that bind to DNA regulatory regions carrying specific consensus sequences (Figure 1). These sequences, which were therefore named cAMP-response element (CRE), are generally represented by the palindromic motif *TGACGTCA* or its half. The transcription factors that bind CRE were named CRE-binding proteins or CREBs. The CREB (or CREB/ATF) family of transcription factors includes three homologous genes: creb, cAMP response element modulator (crem), and activating transcription factor-1 (atf-1), which generate a group of highly homologous proteins that have been named after their prototypes CREB, CREM, and ATF-1, respectively (Figure 2). While both CREBs and ATF-1 are ubiquitously expressed proteins, CREMs are mainly present in the neuroendocrine system. Structurally, the CREB family is characterized by a C-terminal bipartite domain which has been named '*b-zip*' because it juxtaposes a region rich in basic amino acids (basic region) with a region containing heptad repeats of leucine that, through an alpha-helix structure, creates a hydrophobic surface that mediates dimerization by resembling a zipper. The b-zip domain, which mediates both dimerization and binding to DNA, is not specific for the CREB family of transcription factors but is also present in other families, such as C/EBP, Gcn4, c-Jun, and c-Fos. Transcription factors bearing the b-zip domain are known to function as dimers, and they can generate both homo- and heterodimers. Heterodimerization can occur between CREB family members, but also with other b-zip transcription factors such as C/EBPs,

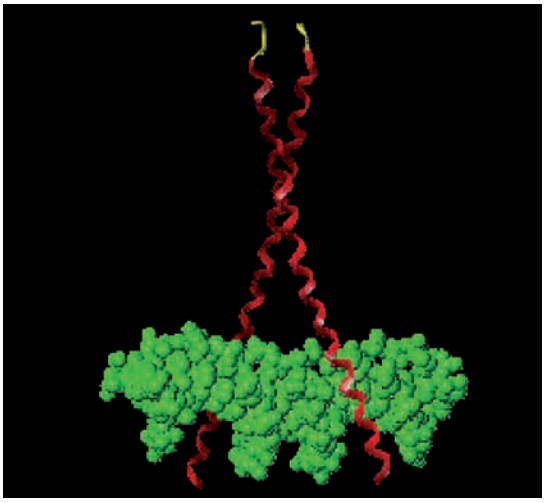

Figure 1 Structural model of CREB complex bound to DNA. Biomolecule Gallery: http://gibk26.bse.kyutech.ac.jp/jouhou/jouhoubank.html.

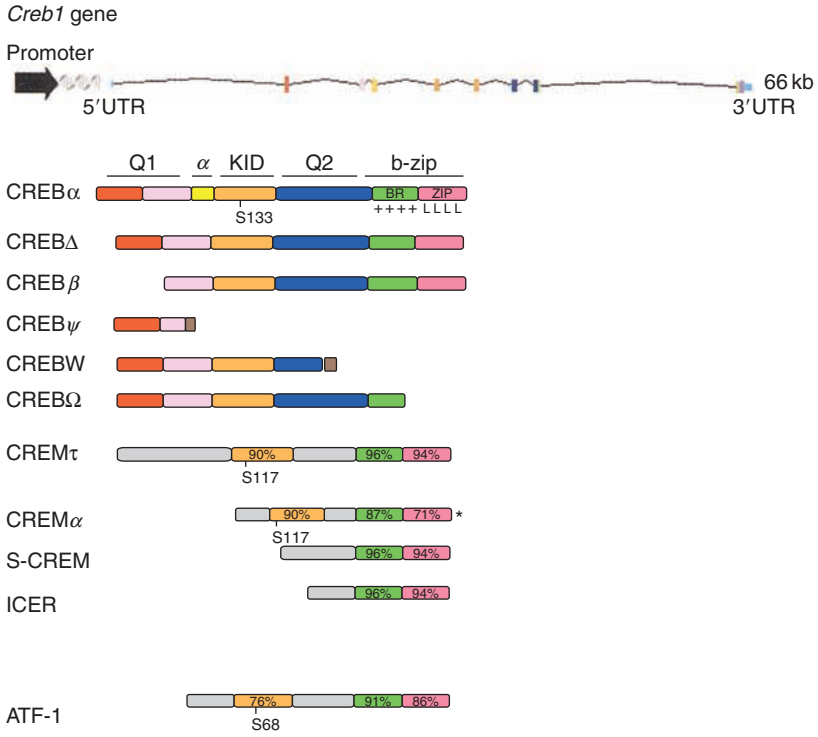

Figure 2 CREB 1 gene structure and domain organization of the CREB family of transcription factors. b-zip, basic region + leucin zipper domain; KID, kinase inducible domain; Q1 and Q2, glutamine-rich regions. The percentage of similar amino acids in CREM and ATF-1 with the corresponding b-zip and KID domains of CREB are indicated. ICER, inducible cAMP early repressor. Reproduced with permission from Barco A and Kandel ER (2006) The role of CREB and CBP in brain function. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 207–236. Weinheim: Wiley-VCH (2006) Copyright Wiley-VCH Verlag GmbH & Co. KGaA.

Fos and Jun proteins, thus providing a great expansion of the potential repertoire and diversity in the regulation of the target genes. Moreover, alternative splicing of the CREB and CREM genes also contributes to the vast repertoire of CREB family members, which includes several isoforms that have distinct repressor or activator features.

In addition to the b-zip domain, CREBs, as well as all other transcription factors bearing this motif, contain a transactivation domain that activates the transcription of the target genes. The transactivation domain of CREB is also bipartite and includes a constitutive (Q2) and an inducible domain (the kinase-inducible domain, KID). While the Q2 domain mediates the interaction with a component of the TFIID complex, the KID seems to promote isomerization by recruiting the coactivator factors CREB-binding protein (CBP) and p300 to the promoters. The KID region is active only when phosphorylated at Ser 133, which occurs in response to cAMP and critically mediates the activation of CREB. CBP and p300 are transcription coactivators with intrinsic histone acetyltransferase activity and are crucial for mediating transcription enhancement.

CREB factors have been extensively studied in the nervous system and found to be implicated in many functions including neurogenesis, neuronal survival, development, differentiation, neuroprotection, axonal outgrowth and regeneration, circadian rhythms, addiction, synaptic plasticity, and memory formation.

CREBs and Synaptic Plasticity

The discovery of the fundamental role of CREB transcription factors in long-term synaptic plasticity stemmed from early findings in the invertebrate *A. californica*, indicating that cAMP mediates LTF and memory responses. The subsequent discovery of the necessary role of CREBs in *Aplysia* LTF helped unravel the understanding of the molecular bases of memory formation.

LTF, like long-term memory, requires gene expression and is accompanied by the formation of new synapses. The characterization of LTF in the *Aplysia* sensory and motor neuron co-culture model system provided the first direct evidence of the essential role of CRE-binding proteins in long-term synaptic plasticity. Injection of CRE-binding elements into the sensory neuron blocked LTF induced by repeated pulses of serotonin (5-HT), while it left short-term facilitation (STF) intact. This and several other related findings in *Aplysia* led to the understanding of some of the fundamental mechanisms underlying LTF. In the *Aplysia* sensory–motor neuron co-culture, STF is induced by one pulse of 5-HT, which activates adenylyl cyclase leading to a transient increase of cAMP. This second messenger binds to the regulatory subunit of the cAMP-dependent protein kinase (PKA), which releases its catalytic subunit that, in turn, phosphorylates several substrates including channels involved in membrane repolarization and proteins mediating neurotransmitter release. As a consequence, a transient increase in transmission release, or STF, occurs. When repeated pulses of 5-HT are applied to the co-cultures, the increase in cAMP and the activation of PKA are sustained for a longer period of time. Moreover, mitogen-activated protein kinase (MAPK) is also recruited. The activated kinases translocate to the nucleus where they presumably phosphorylate an activator form of CREB, CREB1, leading to its activation. In parallel, a repressor form of CREB, CREB2, is removed, thus promoting gene transcription. A first wave of gene expression induced by CREB activation includes the upregulation of immediate early genes such as the transcription factor CCAAT enhancer binding protein (ApC/EBP) and ubiquitin hydrolase, which leads to sustained PKA activation. As discussed below, C/EBP controls the expression of additional target genes and, like CREB, is required for LTF.

Many of the signaling steps required for *Aplysia*, LTF also mediate hippocampal LTP and memory formation in mammals. Subsequent to induction, two phases of LTP can be resolved: an early phase that involves posttranslational mechanisms and decays within about an hour under high-frequency stimulation (HFS) and a protein synthesis-dependent phase (late LTP) that begins as early as 15 min after HFS and can persist for several hours. This late phase has been suggested to model the gene expression-dependent period of memory consolidation. The first demonstration that CREBs are required for LTP in mammals was provided by studies carried out in mice, in which the most abundantly expressed CREB isoforms of the brain had been knocked out (CREBα/Δ). These mice had a significantly impaired LTP. This result was followed by many others that, together, strongly supported the conclusion that CREB plays a critical role in mammalian long-term synaptic plasticity.

Later, some controversial findings reported that, in other mice models with conditional disruption of all CREB isoforms restricted to forebrain regions, or in conditional mutants lacking all CREB isoforms throughout the brain, hippocampal LTP and LTD as well as some forms of memory were normal. However, these mutants displayed a significant impairment in some forms of memory, such as conditioned taste aversion. Another conditional mouse line, in which the function of all CREB subfamilies was blocked in CA1

neurons of the dorsal hippocampus, also showed impairment in some forms of late-phase LTP but not others. Together, these controversial findings suggest that mechanisms of protein compensation due to the transgenic expression may overcome the requirement for CREB during long-term synaptic plasticity, and/or that some forms of LTP are CREB independent.

More recent studies, following up behavioral findings described below, investigated the temporal profile of CREB activation during LTP. Interestingly, LTP, like memory, is coupled to a biphasic profile of CREB phosphorylation, as revealed by the results found with nondecremental late LTP (L-LTP) *in vivo*. Tetanic stimulation produces a first peak of CREB phosphorylation evident 30 min later, and a second, longer lasting peak evident 2 h later, which is sustained for at least a day. Similarly, in hippocampal slices, during early and L-LTP, pCREB increases in a biphasic fashion showing a first peak of phosphorylation at 45 min and a second one at 6 h posttetanus, which returns to baseline levels at 8 h after LTP induction. Similar profiles of CREB activation are also found in rat organotypic hippocampal slices, where the induction of LTP in CA1 is accompanied by a local increase in pCREB, which is sustained for at least 4 h, once again supporting the hypothesis that CREB plays an important role in the late gene expression-dependent phase of LTP.

CREBs and Memory

Parallel studies in both the invertebrate the *Drosophila melanogaster* and in mice, proved that CREB plays a critical and evolutionarily conserved role during memory formation. The induction of a dominant negative form of CREB in *Drosophila* blocks long-term memory formation and the CREBα/Δ mutants are significantly impaired in several types of memories, including spatial, contextual, and cued. These mice, as pointed out above, also have impaired LTP. A great number of subsequent studies confirmed the essential role of CREB in memory in numerous additional species, brain regions, and learning systems. Interestingly, it also emerged that the activation of CREB is sufficient to transform a short-term memory into a long-lasting one, as the overexpression of activator forms of CREB enhances and prolongs memory retention. In summary, a large body of evidence indicates that CREB plays an evolutionarily conserved role in the formation of long-term memory.

Obviously, like for any transcription factor, CREB acts in concert with a number of other transcription regulator and other molecules; therefore, the critical role of CREB needs to be seen in the context of the pattern of genes expressed in the brain regions that mediate memory formation. This is also supported by the observation that the genetic background and gene dosage may influence the phenotype of CREBα/Δ mutants.

Subsequent studies focused on determining the anatomical and cellular distribution and requirement of CREBs as well as their temporal expression following learning. It emerged that the expression of activated CREB throughout the brain or its functional requisite overlaps with the anatomical circuits required for the various learning tasks. Most of the tasks characterized thus far are hippocampal-dependent tasks, and, in fact, they are among the best understood at the behavioral, anatomical, and electrophysiological levels. Knockdown approaches based on oligodeoxynucleotide (ODN) antisense injections in restricted areas of the brain demonstrated that CREB-mediated transcription in the rat dorsal hippocampus is necessary for the formation of spatial memories such as water maze. These experiments also revealed the temporal dynamics of the CREB requirement. CREB antisense ODNs injected bilaterally in the hippocampus just prior to training block memory retention, whereas a similar treatment given 1 day after training has no effect, indicating that there is a time-limited CREB-dependent phase in the hippocampus during the consolidation of spatial memories.

The temporal on/off dynamics of the activation of CREB, measured by the phosphorylation of its Ser-133 (pCREB), have been more extensively examined in other forms of memories, which, like spatial memories, engage the hippocampus, but form contextual rather than spatial representations. These are, for example, contextual fear conditioning types of memories and include inhibitory avoidance (IA) and contextual fear conditioning. These types of memories also require an intact amygdala, a region that processes fear and aversive responses. In fact, these memories are created by associations formed between a context or cues that, during training, are presented in temporal sequence with an aversive experience, that is, a foot shock. The formation of these memories, like that of spatial memories, shows a biphasic activation of CREB in the hippocampus: one peak of activation is evident immediately after training and returns to baseline at 30 min, and the second rises between 3 and 6 h posttraining and returns to basal levels several hours later. This trend of pCREB enhancement coincides with hippocampal PKA activity, which argues that the CREB activation necessary for consolidation of the memory is mediated by PKA phosphorylation. Importantly, the biphasic pattern in pCREB activation is not restricted to the hippocampus, but also occurs in other brain regions including parietal cortex and amygdala.

In conclusion, the activation of CREB, which appears to occur with a biphasic profile and is temporally limited, represents an essential and evolutionarily conserved molecular event of memory formation.

C/EBP, The First Immediate Early Gene

C/EBP is a family of transcription factors that comprise six isoforms defined by distinct genes known as C/EBPα, C/EBPβ, C/EBPγ, C/EBPδ, C/EBPε, and C/EBPζ. Like CREB, the C/EBP family belongs to the b-zip superfamily of transcription factors. Similarly to all the other members of this superfamily, while the b-zip domain is highly conserved, the N-terminal activation domain greatly diverges (<20% sequence identity) (Figure 3). Different C/EBPs are able to form homo- and heterodimers with members of the C/EBP or other b-zip transcription factor families, including CREB, Fos, Gcn4, and Myc. C/EBPs also can dimerize with NF-κB. Some isoforms of C/EBP, like C/EBPζ, can act as dominant negative inhibitors.

With the exception of C/EBPζ, all C/EBPs generally interact with the dyad symmetrical repeat *5′-RTTGCGYAAY-3′*, where R represents A or G, and Y represents C or T. However, the binding element may also include several variations, indicating that C/EBPs are pleiotropic DNA binding factors. C/EBPζ represents an exception, and because of its different structure, it does not bind to the canonical C/EBP element.

The gene structure of C/EBPs is very simple: C/EBPα, β, γ, and δ are intronless, while C/EBPε and ζ contain two and four exons, respectively. Each of the six C/EBP genes can generate the expression of several proteins by alternative use of translation initiation codons in the same mRNA molecule, alternative use of promoters, and differential splicing (Figure 4).

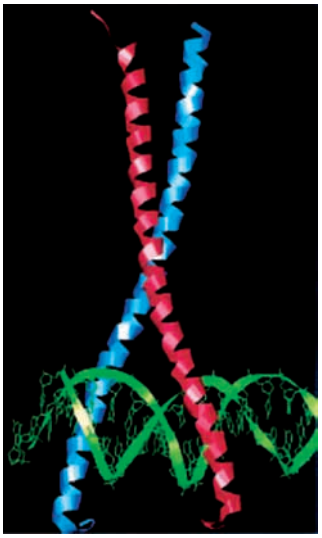

Figure 3 Structural model of the C/EBPβ bound to DNA. Two alpha-helix b-zip domains dimerize and form an inverted Y-shaped structure. Each activation domain, from each monomer binds to half of the palindromic DNA binding site. From Ramji DP and Foka P (2002) CCAAT/enhancer-binding proteins: Structure, function and regulation. *The Biochemical Journal* 365: 561–575.

Additionally, posttranslational modifications of C/EBP play an important role in directing their function and stability.

C/EBPs are expressed in various tissues, including central nervous system (CNS), adipose, liver, intestine, lung, adrenal gland, peripheral blood mononuclear cells, and placenta. They play key roles in differentiation, inflammatory response, liver regeneration, metabolism, synaptic plasticity, memory, and a variety of cellular functions, including stress, growth, and apoptosis. C/EBPs, like CREBs, are evolutionarily conserved molecules required for memory formation, and they belong to the group of early genes regulated downstream of CREB activation.

C/EBPs and Synaptic Plasticity

The first demonstration that a transcription factor is necessary for long-term synaptic plasticity came from the finding of the role of ApC/EBP in *Aplysia* LTF. C/EBP was cloned in the *Aplysia* CNS using expression library screening and found to be generally undetectable in *Aplysia* neurons at resting condition, but rapidly induced following the activation of cAMP or 5-HT. The regulation of the expression of ApC/EBP is typical of an immediate early gene and its function is essential for LTF but dispensable for STF. In fact, injection into the sensory neurons of either C/EBP DNA binding sites (ERE), antisense DNA, interfering RNA (RNAi), or antibodies that prevent C/EBP binding to its DNA binding elements, all block LTF but leave STF intact. The requirement for ApC/EBP during LTF in cell culture lasts for a relatively long time (more than 9 h), suggesting that the ApC/EBP-dependent transcriptional phase essential for LTF is temporally restricted but long lasting. Like CREB activation, the induction of ApC/EBP in cultured sensory neurons is sufficient for transforming STF into LTF. In fact, the application of a single pulse of 5-HT on sensory–motor neuron co-cultures overexpressing ApC/EBP produces LTF, instead of STF. Recent studies indicate that the regulation of ApC/EBP can be mediated by LAPS18-like protein (LLP), a nucleolar protein that binds to the CRE sequence present in the promoter of ApC/EBP. LLP, like ApC/EBP, is required for LTF and its overexpression is sufficient to produce LTF in response to a single 5-HT pulse.

In mammals, inhibition of all C/EBP isoforms and ATF-4 through a regulated expression of a dominant-negative C/EBP (EGFP-AZIP) results in an enhanced

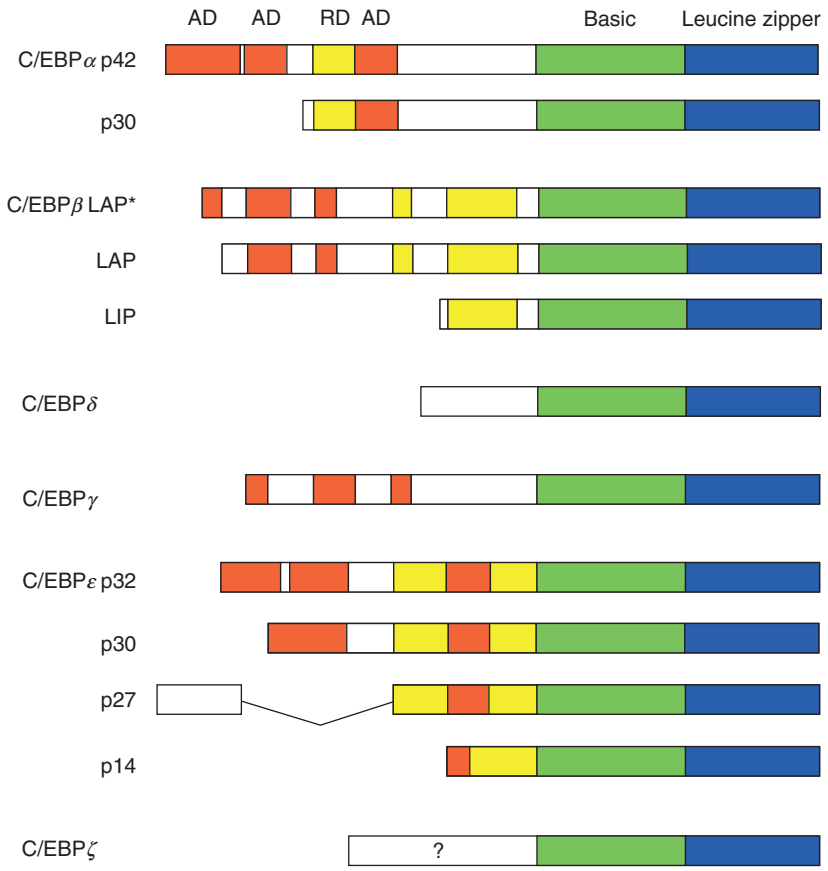

Figure 4 Schematic representation of the C/EBP family members. AD, activation domain; RD, negative regulatory domain. Adapted from Ramji DP and Foka P (2002) CCAAT/enhancer-binding proteins: Structure, function and regulation. *The Biochemical Journal* 365: 561–575.

LTP following submaximal single train of HFS, indicating that either some isoforms of C/EBP or ATF-4 exert inhibitory constraints on long-term synaptic plasticity.

C/EBP and Memory

The functional role of C/EBPs in memory has mainly been studied in the rat using hippocampal-dependent tasks. In the rat hippocampus, the endogenous expression of two C/EBP isoforms, C/EBPβ and C/EBPδ, increases following IA learning but in a delayed fashion relative to CREB phosphorylation, which occurs immediately after training. C/EBPβ protein expression remains significantly increased between 9 and 28 h after IA training, while C/EBPδ mRNA increases at 20 h. It appears that the learning-related induction of C/EBPβ and δ is under the regulatory control of CREB, because the expression of both C/EBP isoforms follows CREB phosphorylation in the same subset of hippocampal neurons, and disruption of CREB function prevents C/EBPβ induction. Similar to the findings obtained in *Aplysia* cultures, the regulation of C/EBPβ expression in the hippocampus during the consolidation of mammalian memory is evoked by the activation of Ca^{2+} and the cAMP-dependent intracellular pathways. This and the characteristic temporal sequence of CREB phosphorylation, followed by C/EBPβ and δ induction, further supports the existence of a CREB-C/EBP-mediated gene expression pathway essential for memory consolidation.

Like CREB, C/EBPβ is indispensable for memory consolidation, as bilateral injections into the dorsal hippocampus of antisense ODN specific for C/EBPβ (β-ODN) results in amnesia. This role of C/EBPβ is limited to a temporal window that begins a few hours after training and ends 2 days later. Thus, C/EBPβ represents one of the evolutionarily conserved CREB-target genes essential for long-term memory consolidation.

The learning-induced regulation of C/EBPβ expression, like the phosphorylation of CREB, requires an intact fornix. The fornix is a fiber bundle that connects the hippocampus to the septum and the hypothalamus and provides important modulatory inputs to the hippocampus. Rats with bilateral lesions of the fornix have an intact short-term memory but

impaired long-term memory. These amnesic animals fail to exhibit the characteristic induction and maintenance of pCREB, C/EBPβ, and C/EBPδ responses following learning in their hippocampi. This suggests that inputs originating in the septum, hypothalamus, and brainstem, traveling through the fornix, convey signals into the hippocampus that regulate CREB-C/EBP-mediated gene expression.

C/EBPδ also plays an important role in memory formation, but apparently by acting in an opposite direction compared to that of C/EBPβ. The C/EBPδ mouse knockout shows a selective enhancement of contextual fear conditioning, but has normal water maze memory. Similarly, a memory enhancement is also evident in mice in which the expression of a broad dominant-negative inhibitor of C/EBP (EGFP-AZIP), which preferentially interacts with several inhibiting isoforms of both C/EBP and ATF-4, is selectively expressed in the forebrain. This suggests that relief of C/EBP, C/EBP-like, or ATF-4-mediated inhibitory mechanisms lowers the threshold for hippocampal-dependent long-term memory in mice.

Zif 268, A Marker of Plasticity

The transcription factor zif268 (also known as Krox24, NGF-I-A, Egr-1, TZs8, and Zenk) was first discovered as an immediate early gene regulated by nerve growth factor and in response to serum treatment (Figure 5). The zif268 gene encodes for two polypeptides of 82 and 88 kDa, which have different translational starting sites, and belongs to the structural family known as zinc fingers. In fact, the DNA binding domain of Zif268, like that of all members of its family (Egr family), folds into three zinc finger structures that mediate DNA binding to the element *5′-GCGC/GGGGCG-3′* (Figure 6). The high homology of the finger sequence among all members of the family implies that they may bind to the same sites of at least a subset of the same target genes.

Figure 5 Structural model of Zif268 complex with DNA. Biomolecule Gallery: http://gibk26.bse.kyutech.ac.jp/jouhou/jouhoubank.html.

In the rat brain, Zif268 is constitutively expressed in the neocortex, primary olfactory cortex, entorhynal cortex, amygdaloid nuclei, nucleus accumbens, striatum, cerebellar cortex, and hippocampus. In the hippocampus, zif268 is highly expressed in the CA1 region and less in CA2–CA3, but only minimally present in the dentate gyrus, where, however, it can be induced in response to a variety of stimuli. Stimuli that regulate Zif268 expression in the brain include seizure, kindling, brain injury, neurodegeneration, apoptosis, stress, and drug tolerance and dependence.

The family members of the Erg family show a similar pattern of basal expression, which is predominantly localized in the nucleus, except for Egr2, that has been found to be expressed in both nucleus and cytoplasm.

Zif268 and Synaptic Plasticity

The expression of zif268 mRNA is strongly induced during LTP at dentate gyrus synapses 30–60 min following tetanic stimulation of the perforant path, and this induction requires the function of NMDA receptors. Interestingly, the role of zif268 seems to be associated with the maintenance of synaptic plasticity rather than its induction. In fact, the posttetanic increase of zif268 correlates more with the persistence of LTP rather than its magnitude of induction. In agreement, zif268 mutant mice show a normal induction and early phase of dentate gyrus LTP, but a faster decay of the LTP late phase. Thus, it was initially believed that zif268 represents a marker of plasticity.

In addition to hippocampus, several other brain regions have been found to display an increase of zif268 mRNA following LTP. These include, for example, the visual cortex following stimulation of the thalamocortical tract and in the insular cortex following stimulation of the basolateral amygdala.

Some controversial results have been obtained concerning the regulation of zif268 expression following long-term synaptic plasticity. For example, zif268 does not seem to change in the CA1 region of the hippocampus if LTP is induced *in vivo*, but significantly increases following tetanic stimulation in CA1 slices. However, for zif268 as for all other inducible proteins, it is important to keep in mind that the general patterns of gene expression and their regulation differ in brain slices compared to the intact brain.

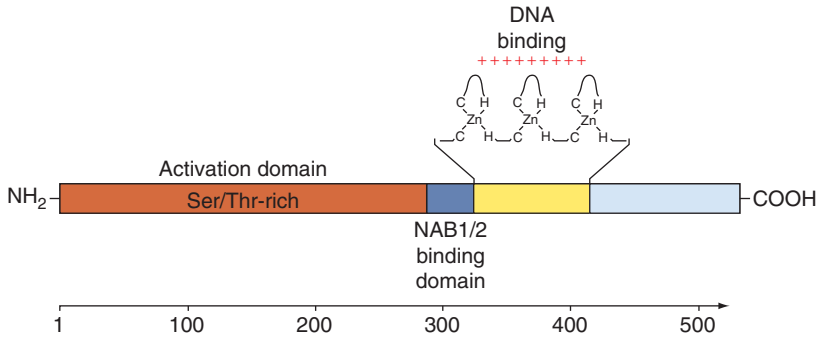

Figure 6 Modular structure of the transcription factor Zif268. The Zif268 protein contains an extended transcriptional activation domain on the N-terminus and a DNA-binding domain, consisting of three zinc fingers motifs. Additionally, an inhibitory domain has been mapped between the activation and DNA-binding domains that functions as a binding site for transcriptional co-repressor proteins. Reproduced with permission from Rössler OG, Stefano L, Bauer I, and Thiel G (2006) Stimulus Transcription Coupling in the Nervous System: The Zinc Finger Protein Egr-1. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 379–391. Weinheim: Wiley-VCH (2006). Copyright Wiley-VCH Verlag GmbH & Co. KGaA.

Other conflicting results concerning the expression of zif268 were reported in both experience-dependent plasticity in the visual cortex and kindling in the amygdala. In both cases, while the acute presentation of the stimulus correlated with the expression of zif268, similar experiments performed on zif268 mutant mice did not reveal any difference compared to wild-type littermates. However, as mentioned above in reference to the CREB and C/EBP studies, a possible explanation for the discrepant results found between acute approaches versus general knockout animal models is that, with the latter, functional compensation may overcome the deficit of the deleted gene. Although zif268 is the member of the Erg family more extensively characterized, other members of the family, that is, Erg 2, have also been found to be regulated following LTP.

Zif268 in Learning and Memory

The expression of Zif268 seems to be regulated in response to a variety of behaviors, including learning and memory. Extensive investigations have revealed that Zif268 is rapidly regulated during associative learning, such as song learning in birds, fear-conditioning and spatial learning in rats, and visual-paired association in monkeys.

In birds, the expression of zenk, the avian homolog of zif268, is regulated in regions that process not only song learning but also song perception. Thus, because zif268 has been generally linked to neuronal plasticity in the mammalian brain, it is now believed that zenk may be linked to molecular and cellular mechanisms that generally underlie the experience-dependent modification of song-responsive circuits.

In rodent, fear-conditioning training correlates with an upregulation of zif268 in the CA1 region of the hippocampus and amygdala. However, because similar increases also occur in matched controls, it seems likely that the change in zif268 is generally associated with the experience of novelty. The characterization of zif268 expression in spatial learning has indicated that, in learning tasks based on multiple trials, the regulation of this transcript seems to be restricted to an initial phase and its upregulation becomes attenuated after repeated trials.

Genetic studies in mice have also demonstrated that Zif268 plays a critical role in memory formation. Zif268 knockout mice have impaired long-term memories but intact short-term retention. The long-term memory deficits have been found in a variety of tasks, including spontaneous alternation, social transmission of food preference, object recognition, spatial learning and conditioned taste aversion. Consistent with these findings, the heterozygous mice have memory deficits that fall midway between the performance levels of the knockout and the wild-type animals. Similarly to CREB mutant mice, if zif268 mutants are exposed to overtraining behavioral protocols, they gain the capability to learn and form memories, suggesting that mechanisms of molecular compensation or use of different learning/memory strategies can compensate for the loss of essential genes. Moreover, it has been debated whether the regulation of zif268 expression also occurs in association with the retrieval of a fear-conditioning memory, a process that has been more closely scrutinized over the last few years. Significant changes in amygdala zif268 levels have been reported only in some cases following the retrieval of contextual fear conditioning, but not in others.

In summary, whether the regulation of zif268 is specifically linked to associative learning following acquisition and/or to the reactivation of the memory, or whether it is rather associated to processes that are prerequisite of learning and memory consolidation remains to be fully established.

AP-1, A Marker of Neuronal Activation

The transcription factor AP-1 is mainly composed of dimers of members of the Jun, Fos, and ATF families. AP-1 mediates transcription regulation in response to a variety of physiological and pathological stimuli, including cytokines, growth factors, stress signals, bacterial infections, viral infections, and oncoproteins. Fos proteins include Fos, Fra1, Fra2, and FosB, while members of the Jun family include Jun, JunB, and JunD. AP-1 is involved in both basal and inducible transcription of several genes that bear the AP-1 consensus site (*5′-TGAG/CTCA-3′*, also known as TPA-responsive element, TRE) in their promoter region. AP-1 can also bind, but with a lower affinity to the CRE element. Like most other transcription factors, variants of the canonical sequences can also be recognized by AP-1 and contribute to the differential function of AP-1 at various regulatory sequences.

A common feature of the proteins that form AP-1 is their structural motif b-zip domain. Fos proteins generally do not homodimerize, whereas both Jun homo- and heterodimers can bind to the AP-1 site (Figure 7).

Gene knockout experiments have shown that Fos and Jun family proteins have mostly overlapping functions, but also some show unique roles that cannot be compensated for by other family members. The functions of Fos and Jun family members are multiple and depend on the specific cell type in which they are expressed. Like for all transcription factors, these functional diversities can also be mediated by several regulatory mechanisms, including posttranslational modifications, selective dimerization between different family members or with other classes of transcription factors (ATF, C/EBP, Maf, and NF-E2), and, finally, interaction with different types of regulatory proteins.

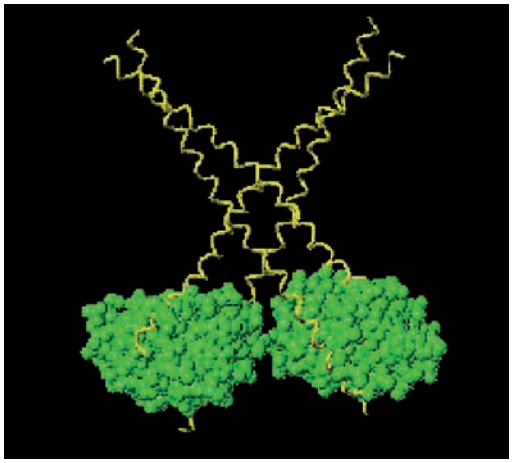

Figure 7 Structural model of AP-1 complex with DNA. Two human c-Fos/c-Jun complex with DNA. Biomolecule Gallery: http://gibk26.bse.kyutech.ac.jp/jouhou/jouhoubank.html.

AP1 and Synaptic Plasticity

The upregulation of c-Jun, the principal component of AP-1, is a particularly common event in the developing, adult and also injured nervous system. A variety of evidence points to a bipotential role of c-jun in mediating neurodegeneration and cell death, as well as plasticity and repair. c-Jun expression is strongly increased in seizure and in some but not all cases of LTP and/or LTD and memory formation. In fact, the role of jun in LTP is evident only with certain technical protocols but not with others. c-fos mRNA has also been found upregulated in response to a variety of neuronal activation protocols including LTP; however, because, in fact, a multitude of stimulations, including sensory, electrical and epileptogenic are associated with an increase in c-fos, it appears that c-fos expression correlates with neuronal activity rather than synaptic plasticity. However, it should also be kept in mind that, often, it is difficult to differentiate these two processes.

AP-1 and Memory

Both correlative and functional studies have provided evidence for the role of AP-1 in the formation of several types of long-term memories. However, conflicting results have also been reported. Antisense ODN-dependent knockdown of c-jun but not jun-B inhibits sequence-specific learning following aversive stimuli in the foot-shock-motivated brightness discrimination. Conversely, complete neural deletion of c-jun has no effect on spatial learning in water maze or fear conditioning paradigms. Contrasting results have been found with the expression of c-fos with spatial tasks. On the other hand, in other types of tasks, the role of c-fos appears to be clear: acquisition of socially transmitted food preference and appetitively motivated spatial learning correlates with an increase of c-Fos as well as pCREB, in the hippocampus and c-Fos expression is necessary for the formation of these memories. Moreover, c-fos also plays a critical role in fear-induced immobility, morphine-induced conditioned place preference, brightness discrimination, conditioned taste aversion, and passive avoidance. In agreement, a correlative increase in c-Fos expression occurs following learning. For example, it is found in the dorsal striatum of rats trained on a variable position reward task in the radial arm maze and in the hippocampus, entorhynal cortex, and primary visual cortex 30 min after spatial learning.

The temporal profiles of c-Fos expression in different neuronal populations have been investigated

using several memory tasks and it appears that c-Fos induction occurs early after training, in the range of minutes and, similarly to that of Zif268, becomes mostly evident during the first few trials, but decays following repeated trials. For example, a study that compared sham-trained mice with mice trained in an appetitive operant-conditioning task showed an increase of c-Fos in CA3, anterior cingulate, occipital and parietal cortex 60 min, but not 120 or 180 min posttraining on day 1. On day 2, c-Fos was increased in CA1, CA3, dentate gyrus, entorhinal cortex, posterior cingulate cortex, and striatum 60 min after training and no differences were found at the later time points. Finally, on day 5, no differences were detected between trained and control groups.

NF-κB, The Distant Traveller

The transcription factor NF-κB, discovered first in lymphocytes, is now known to be a ubiquitously expressed protein. Recently, it has been found that NF-κB is critically involved in synaptic plasticity and learning and memory. NF-κB belongs to the Rel family of transcription factors, which consists of five members: p50, which is a product of the NF-κB1 gene, p52, which is a product of the NF-κB2 gene, p65, also known as RelA, c-Rel, and RelB. Rel family members function as dimers and the five subunits can either homodimerize or heterodimerize. The Rel family members share a Rel homology domain, which contains the DNA binding, dimerization, nuclear localization, and IκB inhibitory interactions regions (Figure 8). The NF-κB typical DNA binding element is *5′-GGGACTTTCC-3′*.

A peculiar feature of NF-κB is that, at resting conditions, dimers are present in an inactive form in the cytoplasm where they bind to IκB inhibitory proteins. Upon stimulation, the IκB kinase complex becomes activated and phosphorylates IκB leading to its ubiquitination and subsequent degradation. This degradation allows NF-κB to translocate to the nucleus and regulate the expression of its target genes.

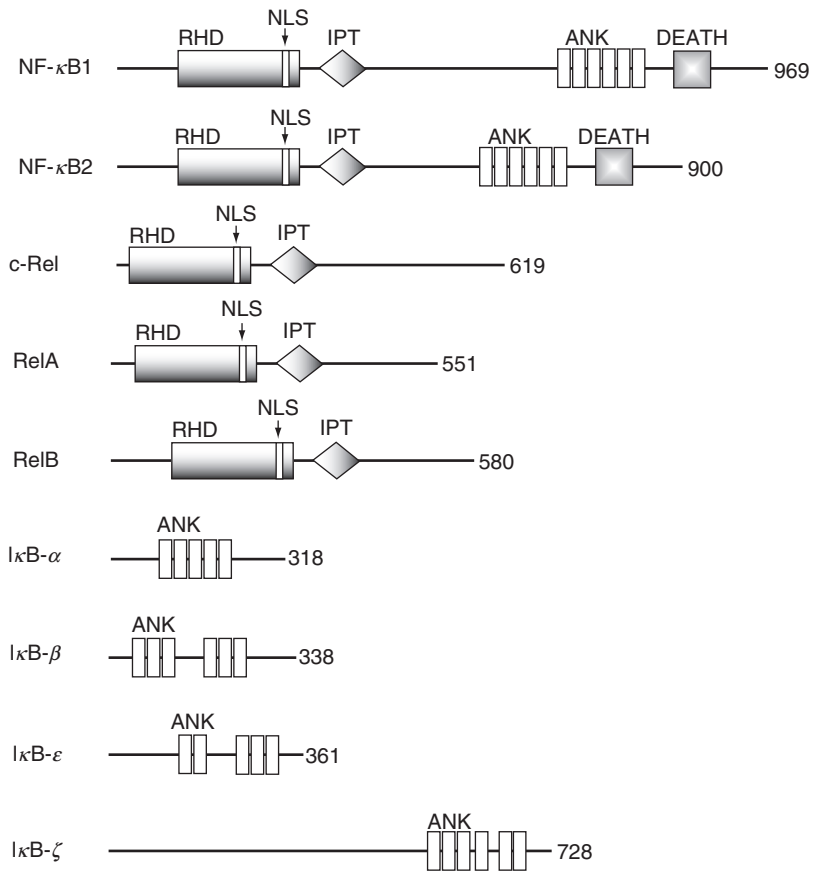

Figure 8 Structure of the NF-κB subunits. Domain motifs are indicated. Protein size is given on the right as amino acids. RHD, Rel homologous domain; NLS, nuclear localization signal; IPT, Ig-like, plexin, transcription factors; ANK, ankyrin repeats; DEATH, DEATH domain, found in proteins involved in cell death. Reproduced with permission from Kaltschmidt B, Mikenberg I, Widera D, and Kaltschmidt C (2006) The role of NF-κB in brain funtion. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 327–343. Wiley-VCH (2006). Copyright Wiley-VCH Verlag GmbH & Co. KGaA.

A critical role of NF-κB has been found in the immune system, where NF-κB-dependent gene expression is critical for inflammation and host defense. NF-κB is also expressed in neurons and nonneuronal cells such as glia and Schwann cells, and expression of multiple NF-κB family members has been found in different cell types of the CNS. The most common species expressed in neurons are p50–p65 heterodimers and the p50 homodimers (Figure 9). Thus far, only a relative small number of NF-κB target genes in neurons have been discovered and include Ca^{2+}-calmodulin-dependent protein kinase II (CaMKII) δ, brain-derived neurotrophic factor (BDNF), μ-opioid receptors, neural cell adhesion molecule, inducible nitric oxide synthetase, and amyloid precursor protein.

NF-κB in Synaptic Plasticity

NF-κB was initially found to be constitutively active in the hippocampus, cerebral cortex, and cerebellar neurons. Thus, it has been suggested that NF-κB is activated upon excitatory synaptic transmission through the Ca^{2+}-responsive signaling cascade and the recruitment of CaMKII. Interestingly, NF-κB also becomes active in response to dopamine. The latent form of NF-κB can become active following a variety of stimuli, including glutamate, kainate, amyloid β peptide, tumor necrosis factor α, brain injury, oxidative stress, or depolarization.

The role of NF-κB in LTP has been established by several lines of evidence including the following:

1. Studies showing that LTP induced either by low-frequency stimulation (LFS) or HFS in hippocampal slices correlates with an increase in the mRNA of both p50 and p65 and a decrease of IκB mRNA.
2. Pretreatment with NF-κB DNA element decoy interferes with the induction of both LTP and LTD.
3. *In vivo*-induced perforant path LTP in mouse hippocampus is accompanied by an enhancement of the NF-κB binding to its DNA binding elements in the hippocampal nuclear extract.
4. An incremental increase of NF-κB active form is detected in granular and pyramidal cells in the hippocampus following LFS and HFS-LTP, respectively.

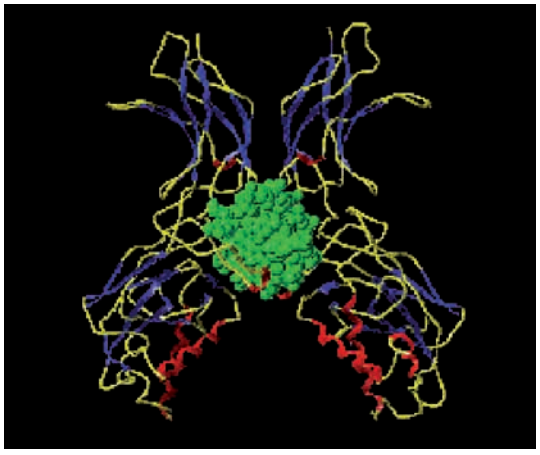

Figure 9 Structural model NF-κB p50 homodimer complex with DNA. Biomolecule Gallery: http://gibk26.bse.kyutech.ac.jp/jouhou/jouhoubank.html.

Similar investigations carried out in mice with genetic disruption of NFκB restricted to the forebrain confirm that this transcription factor is critical for synaptic transmission, late phase of LTP induced by theta burst stimulation (TBS), and induction of LTD.

NF-κB in Memory

NF-κB plays a critical role in several types of long-term memory in species ranging from invertebrates to mammals. In the crab *Chasmagnathus*, the habituation of the escape response elicited by a fear stimulus (passage of an opaque figure representing a potential predator) correlates with an increase of the activated form of an NF-κB homolog in brain nuclei and isolated synapses. Furthermore, injection of an IκB kinase complex inhibitor blocks NF-κB activation and disrupts the escape memory, indicating that NF-κB is essential for this memory. Further demonstrations provided evidence that NF-κB is also recruited for the maintenance of the crab escape memory following retrieval, as, in fact, blocking NF-κB following memory recall results in amnesia.

Similar conclusions have been obtained with mammalian memory systems. Mice lacking the p65 subunit of NF-κB and tumor necrosis factor receptor 1 (TNFR), a factor necessary for the survival of the p65 knockout, have impaired spatial radial maze memory. However, like with other transcription factors described above, this impairment is overcome if the animals undergo overtraining protocols. Again, similarly to other transcription factors discussed above, the role of NF-κB is not found in all memories, as, indeed some types of memory, but not others, appear to be more readily disrupted by the NF-κB inactivation. Thus, the same NF-κB knockout mice that show impaired radial maze have normal fear-conditioning memory, and intact memory in general is also found in mice lacking the p50 NF-κB subunit. On the other hand, intra-amygdala administration of NF-κB decoy disrupts fear-potentiated startle, and inhibition of NF-κB with decoy elements or pharmacological inhibitors injected into the hippocampus disrupts IA memory. Additional studies showed that forebrain inhibition of NF-κB function obtained by overexpression of IκB in mice affects spatial memory assessed by water maze. Interestingly, in these mice, both the mRNA of the catalytic subunit of PKA and CREB

phosphorylation are significantly decreased, suggesting that NF-κB affects the activation of the CREB pathway during memory consolidation. Together, all these evidence strongly argue for a role of NF-κB in synaptic transmission, regulated synaptic plasticity, and memory formation.

An intriguing aspect of NF-κB expression in neurons that deserves more understanding is its localization not only in the cytoplasm, but also at the synaptic compartment, where it appears to be mainly present as p50–p65 heterodimers. Whether this localization serves the function of retrograde communication to the nucleus or reflects a novel, nontranscriptional function remains to be understood.

Conclusions

Experiments in fruit flies, snails, and rodents have identified highly conserved molecular pathways involved in synaptic plasticity, learning and memory. A large number of studies have revealed that these molecular pathways are regulated by combinatorial efforts of transcription factors, which include members of the CREB, C/EBP, Egr, Rel, and AP1 families. The general outcome of these studies suggests that the contribution of combinatorial transcription regulation, which mediates the expression of specific gene patterns that change over time, represents the molecular bases of memory consolidation. Identifying the distinctive features of these gene expression patterns is the next challenge of the field.

See also: BDNF in Synaptic Plasticity and Memory; Cyclic AMP (cAMP) Role in Learning and Memory; Long-Term Potentiation (LTP): Mossy Fiber cAMP-Dependent Presynaptic LTP; Synaptic Plasticity and Place Cell Formation.

Further Reading

Alberini CM (2006) CCAAT enhancer binding proteins in the nervous system: Their role in development, differentiation, long-term synaptic plasticity, and memory. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 243–258. Weinheim: Wiley-VCH.

Alberini CM, Milekic MH, and Tronel S (2006) Mechanisms of memory stabilization and de-stabilization. *Cellular and Molecular Life Sciences* 63: 999–1008.

Barco A and Kandel ER (2006) The role of CREB and CBP in brain function. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 207–236. Weinheim: Wiley-VCH.

Colombo PJ (2004) Learning-induced activation of transcription factors among multiple memory systems. *Neurobiology of Learning and Memory* 82: 268–277.

Davis S, Bozon B, and Laroche S (2003) How necessary is the activation of the immediate early gene zif268 in synaptic plasticity and learning? *Behavioural Brain Research* 142: 17–30.

Dragunow M (2004) CREB and neurodegeneration. *Frontiers in Bioscience* 9: 100–103.

Kaczmarek L (1992) Expression of c-fos and other genes encoding transcription factors in long-term potentiation. *Behavioral and Neural Biology* 57: 263–266.

Kaltschmidt B, Mikenberg I, Widera D, and Kaltschmidt C (2006) The role of NF-kB in brain funtion. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 327–343. Weinheim: Wiley-VCH.

Levine M and Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424: 147–151.

Mayr B and Montminy M (2001) Transcriptional regulation by the phosphorylation-dependent factor CREB. *Nature Reviews Molecular Cell Biology* 2: 599–609.

Meffert MK and Baltimore D (2005) Physiological functions for brain NF-kappaB. *Trends in Neuroscience* 28: 37–43.

Montminy M, Koo SH, and Zhang X (2004) The CREB family: Key regulators of hepatic metabolism. *Annales d Endocrinologie (Paris)* 65: 73–75.

Raivich G and Behrens A (2006) The role of c-jun in brain function. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 259–273. Weinheim: Wiley-VCH.

Ramji DP and Foka P (2002) CCAAT/enhancer-binding proteins: Structure, function and regulation. *The Biochemical Journal* 365: 561–575.

Rössler OG, Stefano L, Bauer I, and Thiel G (2006) Stimulus-transcription coupling in the nervous system: The zinc finger protein Egr-1. In: Thiel G (ed.) *Transcription Factors in the Nervous System*, pp. 379–391. Weinheim: Wiley-VCH.

Silva AJ, Kogan JH, Frankland PW, and Kida S (1998) CREB and memory. *Annual Review of Neuroscience* 21: 127–148.

Relevant Website

http://gibk26.bse.kyutech.ac.jp – Biomolecule Gallery, BioInfo Bank.

Transcriptional Networks and the Spinal Cord

S L Pfaff and T Macfarlan, The Salk Institute for Biological Studies, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

One of the challenges facing the developing embryonic nervous system is how to ensure the production of multiple neuronal subtypes at the correct time, place, and number during development. One of the areas of the central nervous system (CNS) where the mechanisms for regulating cell fate have been extensively studied is the spinal cord. The spinal cord is embryologically derived from dividing multipotential neuroepithelial cells comprising the neural tube at early stages of development. Over the past two decades, there has been tremendous progress in our understanding of the signaling pathways and transcription factors that instruct neuroepithelial cells to differentiate into functionally specialized subpopulations of postmitotic neurons and glia. Spinal neuronal differentiation signals are distributed along the dorsoventral (D/V) and rostrocaudal (R/C) axes, leading to the generation of specific cell types at stereotypical positions along these axes.

D/V Patterning of the Spinal Cord

Interneurons that process and relay sensory signals are primarily restricted to the dorsal region of the spinal cord, whereas motor neurons and the interneurons that regulate their activity are concentrated ventrally. A key signaling molecule working to pattern the ventral spinal cord is sonic hedgehog (shh), a factor that is used at multiple sites within the embryo to control cell specification and growth. Shh is a glycoprotein secreted from notochord cells located underneath the neural tube. Later in development, shh is also produced by the floor plate, a glial wedge that forms at the ventral midline of the neural tube. A critical feature of shh's action on neural tube progenitor cells is its ability to function as a morphogen, namely to signal the differentiation of different cell types in a concentration-dependent manner. Within the neural tube, ventrally located progenitor cells closest to the floor plate and notochord encounter high levels of shh whereas cells located more dorsally are exposed to lower levels of the factor (Figures 1(a) and 1(b)). Thus, the position of a cell within the ventral neural tube dictates the level of shh it will encounter. Cells exposed to different concentrations of shh activate distinct differentiation programs, express unique genes, and ultimately acquire different cell fates (Figure 1(a)).

Shh signaling controls cell specification in the ventral spinal cord by establishing molecularly distinct classes of precursor cell populations, also referred to as progenitor cell domains. Each of the separate precursor populations established by shh signaling expresses a unique combination of homeodomain-class transcription factors that control gene expression (Figure 1(b)). This group of so-called progenitor factors can be divided into two classes based on how they are regulated by shh; class I factors are expressed by default and become repressed by shh, whereas expression of class II factors is activated by shh. The ventral spinal cord is divided into five separate progenitor cell domains referred to as p0, p1, p2, pMN, and p3 that serve as the precursor cell populations for V0–V2 interneurons, motor neurons, and V3 cells, respectively (Figure 1(a)). Each of the postmitotic neuronal progeny that form from the progenitor domains exhibit different phenotypic characteristics, including unique cell migration patterns, axon targeting, and in the cases examined distinct physiological properties.

Insight from studies on ventral cell patterning has helped us to understand some of the general principles that operate to control dorsal interneuron specification in the neural tube. The precursors of dorsal sensory classes of interneurons (dI) are similarly divided into progenitor cell domains (dp1–6) that represent the precursors for six early forming (dI1–6) and two late developing (dIL^A and dIL^B) neuronal subtypes (Figure 1(a)). These interneuronal cell types can be broadly divided into two groups based on the signals required for their specification. Class A cells, consisting of dorsal interneurons dI1–3, are dependent on secreted factors from the roof plate. In contrast, class B neurons (dI4–6 and dIL^A and dIL^B) are able to differentiate even in the absence of roof plate-derived signals. The roof plate secretes at least two different types of signaling molecules that contribute to class A cell specification. One group of related proteins are members of the transforming growth factor-β (TGF-β) family and include activin, bone morphogenic protein 4 (BMP4), BMP5, BMP7, dorsalin1, and growth differentiation factor 7 (GDF7). Another family of signaling proteins expressed by roof plate cells are the wingless related factors Wnt1 and Wnt3A. Numerous genetic experiments have clearly implicated TGF-βs in dorsal interneuron specification. For example, overexpression of BMPs in the spinal cord increases the

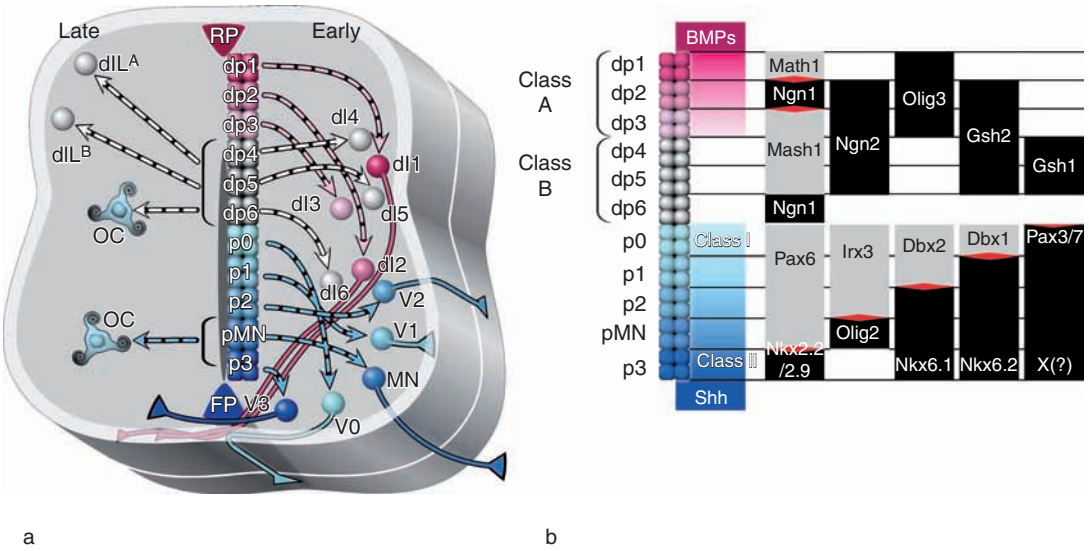

Figure 1 Progenitor domains give rise to distinct cell types. (a) The ventricular zone is divided into progenitor domains based on the differential expression of specific genes. The ventral spinal cord is divided into five progenitor domains (p0–p3 and pMN) that give rise to distinct classes of mature neurons (V0–V3 and MN) that migrate to stereotypical locations and extend axons along specific pathways. Later in development, pOC domains form that gives rise to oligodendrocytes (OC) that disperse throughout the spinal cord. In the dorsal spinal cord, six progenitor domains (dp1–6) give rise to six early-born (dI1–6) and two late-born (dIL^A and dIL^B) classes of interneurons which migrate laterally to stereotypic positions. Where known, the axonal projections are indicated. RP, roof plate; FP, floor plate. (b) In the ventral spinal cord, gene expression patterns are established by a low dorsal to high ventral concentration of shh. Progenitor transcription factors fall into two classes: class II factors (black bars) that are induced by shh and class I factors (gray bars) that are repressed by shh. The shh gradient thus leads to cells expressing a unique combination of progenitor factors. In addition to the graded responses to shh, progenitor factors exist in cross-repressive pairs that are incompatible with co-expression in the same cell, causing discrete boundaries between progenitor domains to form. In the dorsal spinal cord, progenitor domains are classified into two subtypes, the roof plate-dependent class A (dI1–3) and roof plate-independent class B (dI4–6 and $dIL^{A/B}$) neurons. bHLH factors have prominent roles for patterning the progenitor domains in the dorsal spinal cord, whereas homeodomain factors are more common ventrally. Similar to the ventral spinal cord, cross-repression (between Mash1, Ngn1, and Math1) is an important mechanism delineating progenitor boundaries. Other progenitor factors are also shown.

number of dI1 progenitors and interneurons at the expense of other interneurons, whereas inhibition of BMP signaling by expression of the BMP antagonist noggin or inactivation of both BMPr1a and BMPr1b receptors results in reduced numbers of dI1 and dI2 interneurons. Interestingly, dI3 interneurons are still generated in these BMP receptor, mutants, consistent with recent findings suggesting that activated activin receptors promote dI3 interneuron fate independently of BMP signaling. The wnts appear to have complementary functions to those of the TGF-β family, as *Wnt1*/*Wnt3A* double-mutant mice also exhibit defects in dI1–3 interneuron specification.

Negative Gene Regulation: Cross-Repression and Derepression

One of the key features establishing D/V identity in the neural tube is the proper interpretation of the graded shh signal into decisive differentiation programs for cell fate specification, even under conditions when cells reside at the signaling threshold for two different fates. The strategy used by progenitor cells appears to be based on the genetic interactions that take place between the numerous downstream transcription factors that become expressed in progenitor cells. Transcription factors expressed by adjacent progenitor cell populations have been found to repress one another's expression, a process referred to as cross-repression. Class I factors repressed by shh oppose the expression of class II factors induced by shh in adjacent cells, and similarly class II factors repress adjacent class I factors (Figure 1(b)). Thus, mutations in one factor of a cross-repressive pair leads to an expansion of the opposing factor into an inappropriate cell domain. This alteration results in misspecification of neuronal cell fate from the altered precursor cells in the progenitor domain. The cross-repressive interactions between progenitor cell transcription factors likely help to ensure that cells do not acquire mixed or hybrid characteristics even at the progenitor boundaries.

In the dorsal spinal cord, cross-repressive interactions between basic helix–loop–helix (bHLH) class transcription factors such as Math1, Ngn1, and Mash

1 delineate progenitor domains in a manner that is analogous to the homeodomain factors in the ventral spinal cord (Figure 1(b)). Similarly, studies of mouse mutants and gain-of-function experiments in chick embryos have established key roles for Ngn1, Math1, and Mash1 in the specification of dorsal neuron fate. Additional bHLH and homeodomain transcription factors such as Olig3, Gsh1, and Gsh2 are also expressed by dorsal progenitor cells, but do not appear to participate in cross-repression. Nevertheless, genetic experiments have established a role for Olig3 in dI1–3 cell specification and Gsh2 in dI3 interneuron development (Figure 1(b)).

It is likely that the chromatin modifications directed by the progenitor factors have a central role controlling cell fate specification in the developing spinal cord. The groucho/TLE cofactor physically interacts with many of the progenitor cell transcription factors and is essential for cross-repression. Eight of the 11 known transcription factors expressed by ventrally located progenitor cells have an Engrailed homology (eh1) domain that binds to groucho/TLE corepressors. The recruitment of groucho/TLE proteins to DNA inhibits gene expression by recruiting the histone deacetylase Rpd3/HDAC1. This enzyme catalyzes the removal of acetyl groups on histone tails, creating a chromatin structure that is less favorable for transcription. Thus, it is predicted that progenitor factors cross-repress the expression of neighboring progenitor factors by creating a multiprotein enzymatic complex that limits the accessibility of RNA polymerase and other factors needed for transcription.

The finding that progenitor factors are transcriptional repressors helps to explain how they prevent genes from adjacent cell domains from becoming expressed, but it raises further questions about how cells acquire unique identities using gene silencing as a mechanism. Although there is precedence for bifunctional transcription factors that can activate or repress gene expression based on their context, the cell fate specifying activity of the progenitor factors appears to be mediated entirely through transcriptional repression. Fusion of just the DNA-binding homeodomain of the progenitor factors to the eh1 groucho/TLE interaction domain of Engrailed generates constitutive repressors that are found to mimic the activity of the native progenitor factors. Because progenitor cell fate is regulated by gene silencing, it is thought that 'derepression' is important for proper expression of genes. The derepression model predicts that progenitor factors function by preventing the expression of inappropriate genes for alternative cell fates rather than activating the expression of genes for the development of a specific cell type. For this derepression model to work, however, it is thought that general (noncell type-specific) transcriptional activators are poised to activate genes for spinal neuron differentiation unless prevented from doing so by the recruitment of histone deactylases by progenitor factors. Expression of effector genes for a specific cell identity would thus be determined by the absence of binding sites for a particular progenitor factor.

Cell Specification Factors Downstream of Progenitor Factors

Although molecular differences are apparent in the profile of transcription factors expressed by progenitor cells within the ventricular zone of the spinal cord, the process of cell fate specification is best described as a sequential cascade in which progenitor factors control the expression of a second layer of genes that play active roles in regulating the phenotypic characteristics of the differentiating neuronal subtypes. Thus, many of the progenitor cell transcription factors are only expressed by cells within the ventricular zone and become downregulated as neuronal differentiation proceeds. Many of the downstream factors that become expressed in differentiating neurons represent additional homeodomain transcription factors. For example, V0 cells express Evx1, V1 cells express EN1, V2 cells express Chx10, MN cells express Hb9 and Isl1/2, and V3 cells express Sim1.

Gain- and loss-of-function experiments in mouse and chick embryos have helped to define the role of many of the downstream homeodomain transcription factors involved in specifying the phenotype of cells within the spinal cord. Some factors are essential for establishing the complete identity of particular cells, whereas other factors appear to regulate the expression of a subset of the phenotypic characteristics. In *Evx1* mutant mice, for example, V0 commissural interneurons extend axons ipsilaterally as if they have acquired the identity of V1 cells. Mutations in *En1* cause V1 neuron axon pathfinding defects and a slowing in the activity of the spinal locomotor circuitry involved in rhythmic stepping. The loss of En1 function, however, does not appear to be associated with a wholesale conversion in the identity of V1 cells to that of another cell type in the ventral spinal cord. Mutation of the *Isl1* gene expressed by motor neurons results in a block in motor neuron differentiation. As a consequence, the progeny of pMN cells abruptly apoptose in *Isl1* mutants rather than develop with an alternative cell identity. The loss of Hb9 function from motor neurons, in contrast, has a more subtle phenotype in which motor neurons still develop but they inappropriately express V2 interneuron genes. The hybrid motor neuron/V2 interneuron cells that

form in *Hb9* mutants are unable to extend axons to proper muscle targets in the periphery and as a result the animals exhibit severe locomotor defects.

Neurogenesis

Although progenitor cells exposed to roof plate- and floor plate-derived signals become predisposed to produce specific cell progeny, there is also an ancient evolutionarily conserved regulatory pathway that is superimposed on the process of cell specification which controls whether precursor cells continue to divide or begin their terminal differentiation. Gene expression for this regulatory check point termed neurogenesis is controlled by bHLH transcription factors. This family of proteins is defined by the bHLH signature motif consisting of 60 amino acids. The N-terminal 15 amino acids encompass the basic domain that binds to DNA. The remaining C-terminal domain forms a helix–loop–helix that functions as a dimerization domain that mediates the formation of homo- and heteromeric complexes. bHLH dimers bind to a DNA sequence called the E box (CANNTG) found in many neuronal genes. As noted above, the dorsal spinal cord expresses several bHLH transcription factors within the dI progenitor cells, and these factors exhibit cross-repressive interactions, cell fate-specifying attributes, and drive neurogenesis. Thus, a remaining question is to understand how each of these processes is regulated in a context-dependent manner.

The function of proneural bHLH factors is controlled at both the transcriptional and posttranscriptional level. Id-family proteins bind directly to proneural bHLH factors such as Mash1, Ngns, and NeuroD and generate non-DNA binding heterodimers. This inactivation occurs because Id proteins contain helix–loop–helix domains for dimerization but lack the basic region required for DNA binding. In addition, the transcription of proneural bHLH factors is negatively regulated by the Hes1 transcriptional repressor. Thus, in order for neurogenesis to proceed, Hes and Id must be downregulated and/or inactivated in progenitor cells to allow assembly of functional proneural bHLH complexes.

Once neurogenesis is initiated, the differentiation of neurons appears to involve the sequential activity of several bHLH factors that encompass the period when progenitor cells cease dividing, emerge from the ventricular zone as a postmitotic neuron, and mature into a specialized neuronal subtype. Developmental studies suggest that the proneural bHLH factors expressed transiently just prior to neuronal differentiation, including the Ngn family, Mash1, and Math1, initiate neurogenesis programs, while members of the NeuroD/Nex family are involved in the terminal differentiation and maturation of the cells. For example, Ngn2 overexpression decreases the proportion of proliferating progenitors and induces the expression of cell division kinase (CDK) inhibitors which stop the cell cycle, whereas NeuroD expression is more prominent in postmitotic neurons and coincides with their expression of pan-neuronal genes such as neurofilament.

Precursor cells are found to produce neurons at early stages and later give rise to glial cell derivatives. Thus, the check points that control neurogenesis have an important influence on the differentiation of cells within the spinal cord. Oligodendrocytes are produced from the pMN and dp4–5 progenitor domains, whereas astrocytes are generated from the other progenitor cell populations. The differentiation of oligodendrocytes from the pMN domain is initiated when Ngn2 expression is lost from the Olig2-expressing cells. The mechanistic basis for Ngn2's suppression of glial fate may be based on a process of transcription cofactor 'squelching.' This is predicted from the finding that the related factor, Ngn1, inhibits astrocyte formation by sequestering CREB binding protein (CBP), which is an obligate cofactor whose availability is needed in order to efficiently transcribe astrocyte genes.

LIM Networks in Postmitotic Neurons

Many different types of neurons within the spinal cord express LIM homeodomain (LIM-HD) transcription factors as they emerge from the ventricular zone and begin to differentiate. There are 12 LIM-HD genes in higher vertebrates, most of which are matched with homologs in invertebrates that have been shown to control cell fate specification. The regulation of LIM-HD expression in developing neurons remains poorly understood, but these factors are clearly downstream of the progenitor cell factors. Thus, the prevailing model of the transcriptional networks that operate in the neural tube posits that extrinsic factors such as shh and TGFβs regulate progenitor cell factors which in turn regulate the expression of LIM-HD genes, resulting in a sequential cascade of regulatory factors being expressed as development progresses.

A hallmark of the LIM-HD factors is the zinc-binding LIM domain found in their N-terminus. The LIM motif was first identified as a conserved structure in two *Caenorhabditis elegans* genes, *lin-11* and *mec-3*, and in the insulin enhancer binding protein *Isl1*. The LIM domain mediates protein–protein interactions with the cofactor nuclear LIM interactor (NLI). NLI is present in many cell types and possesses its own self-dimerization domain which allows NLI to drive the formation of complexes containing multiple LIM-HD proteins.

A critical finding regarding the LIM-HD factors was that they function in a combinatorial fashion to control cell fate. Although the expression of a single LIM-HD factor does not segregate with neuronal subtypes, the combinatorial expression of LIM-HD factors has been used to define distinct subtypes of neurons. Thus, individual members of the LIM-HD gene family are typically expressed by several neuronal subtypes within the spinal cord, but each neuronal subtype is often associated with a unique mix of LIM-HD proteins, referred to as the LIM code. During neuronal differentiation, LIM-HD factors appear to control several aspects of cell fate in a sequential manner. For example, as motor neurons develop from pMN progenitors, they express LIM-HD factors Isl1 and Lhx3 which specify a generic motor neuron phenotype. This early motor neuron LIM code is transient, however, and is soon followed by new combinations of LIM-HD factors that drive the further diversification of motor neuron identity leading to the formation of motor columns – a subdivision of motor neurons that corresponds to cells that settle together within the spinal cord and extend axons along common pathways in the periphery (Figures 2(a) and 2(c)).

The advantage of a LIM code is that diversity can be generated by mixing and matching a small number of elements; however, a by-product of a combinatorial code is that the elements are reused repeatedly. For example, Lhx3 alone contributes to V2 interneuron development, but Lhx3 combined with Isl1 triggers motor neuron formation. How do LIM-HD factors function in a combinatorial manner to control gene expression? The solution appears to be based on cell-specific protein–protein interactions mediated by the LIM domains. Within V2 interneurons, an Lhx3:NLI complex forms that regulates distinct target genes from the Lhx3:NLI:Isl1 complex that triggers motor neuron development. This example illustrates that the LIM domain is a critical determinant for controlling how LIM-HD factors regulate different cell fates in a combinatorial manner.

R/C patterning, Hox Genes, and Motor Pools

There are distinct classes of neurons that align along the R/C axis of the spinal cord in approximate coordination with the peripheral targets they innervate, ensuring for example that limb-innervating motor neurons are generated in register with the limbs. Motor columns form one of the basic units of motor neuron organization in the spinal cord, with distinct columnar subtypes innervating specific muscle targets. Lateral motor columns (LMCs), for example, are generated at fore- and hindlimb levels and innervate limb musculature, whereas autonomic motor neurons in the preganglionic motor column (PMC) form at thoracic levels (Figures 2(a) and 2(c)). Within motor columns,

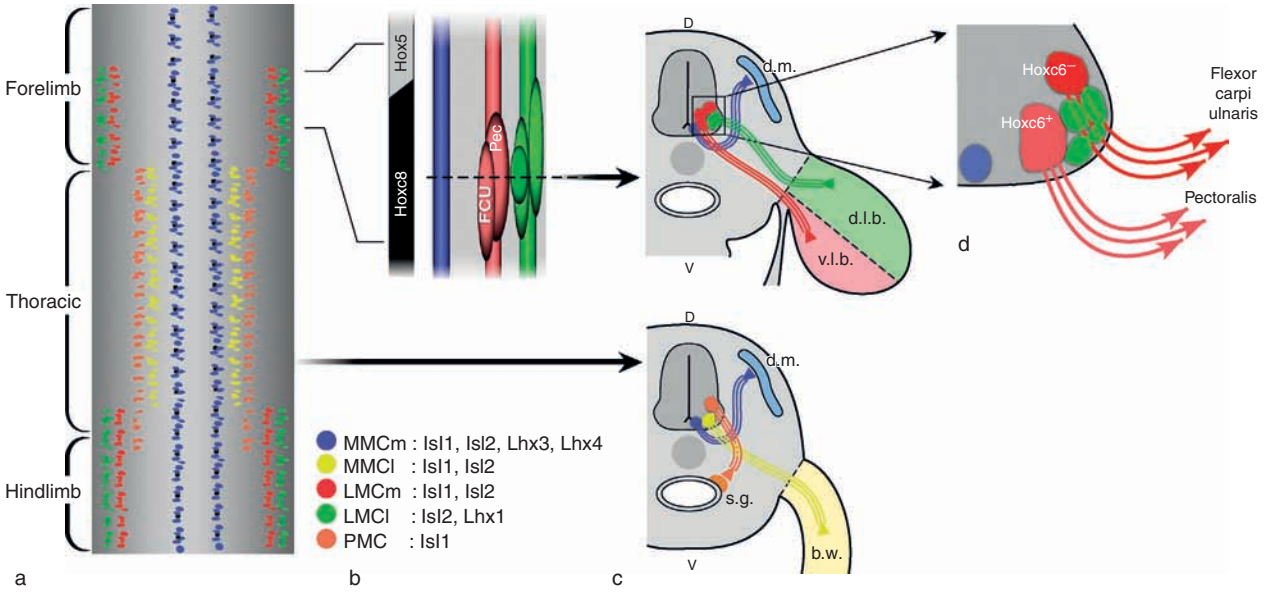

Figure 2 The LIM and Hox codes define columnar and pool identity of motor neurons. (a) Schematic representation of the spinal cord showing the approximate location of the motor columns marked by unique combinations of LIM-HD factors (color coded). (b) An expanded view of the brachial spinal cord showing the relative position of several motor pools. The position of motor pools innervating the pectoralis (Pec) and flexor carpi ulnaris (FCU) muscles are determined by Hox5 (Hoxa5 + Hoxc5) and Hoxc8 expression. (c) Transverse sections of the brachial and thoracic spinal cord are diagramed showing axonal projections to muscle targets. d.m., dermamyotome; d.l.b., dorsal limb bud; v.l.b., ventral limb bud; s.g., sympathetic ganglia; b.w., body wall. (d) Within a motor column (in this case the LMC, red), motor pools emerge that connect to specific muscle targets. The intrasegmental distinction between FCU and Pec cells depends on the status of Hoxc6 in combination with other transcription factors.

motor neurons can be further subdivided into motor pools representing cells that form selective connections with individual muscles (Figures 2(b) and 2(d)). Although the LIM code plays an important role in assigning columnar identity, it is insufficient to account for the entire range of motor neuron subtypes. Hox-class homeodomain transcription factors play a major role in the further subdivision of motor columns into motor pools.

The Hox gene family is organized into four clusters in which the expression of the 3′-most genes is initiated at rostral levels of the CNS while the 5′-positioned genes are progressively expressed at more caudal regions. Exposure to retinoic acid (RA) is required for expression of more 3′ Hox genes, whereas fibroblast growth factors (FGFs) activate the expression of more 5′ genes. Recent studies have demonstrated that Hox protein expression segregates with motor neuron columnar and pool identity and that sequential phases of Hox protein expression are required to properly specify motor neuron subtype identity. The basis for this highly precise segregation in Hox expression is analogous to progenitor factors and is due in part to cross-repressive regulatory interactions among the Hox genes. The restriction of Hox proteins to specific motor pools is critical for establishing pool identity and proper muscle connectivity. Hoxc6, for example, is found in motor pools of the LMC that innervate the pectoralis muscle but not the flexor carpi ulnaris (Figures 2(b) and 2(d)). Consequently, misexpression of Hoxc6 causes a switch in motor pool identity leading to muscle targeting errors.

Large muscles are innervated by large motor pools and small muscles by small pools. An important feature of the Hox network is that the inherent strength of the cross-repression mediated by each Hox factor could potentially determine the size of the motor pool. Small motor pools might therefore be defined by weak repressors, whereas large pools could be generated using strong represses. In this way, the number of cells comprising the motor pool could be more accurately matched with the size of the peripheral muscle target.

Conclusions

One of the largest remaining questions in developmental neurobiology is how neuronal migration, axon guidance, and other crucial aspects of a neuron's phenotype are determined by the early gene regulatory events that occur as the neuron's identity is established. It is clear that there exists a sequential transcriptional cascade involving bHLH, homeodomain, LIM-HD, and Hox factors operating at multiple steps. In the first stage, as the neural plate folds to form the neural tube, progenitor domains within the ventral spinal cord are defined by the expression of homeodomain progenitor factors that are either induced (class II) or repressed (class I) by shh. In the second stage, the profile of the class I and II homeodomain transcription factors leads to the expression of determinants of neuronal subtype identity. In the third phase of the transcriptional hierarchy, neuronal subtype transcription factors including LIM-HD and Hox factors act in conjunction with neurogenic bHLH factors to coordinate pan-neuronal with subtype-specific programs of differentiation. Defining the full repertoire of target genes of these transcriptional networks may provide the link between neuronal cell identity and neuronal phenotype.

See also: Hox Gene Expression; Sonic Hedgehog and Neural Patterning; Thyroid Hormone and Transcriptional Regulation in the CNS; Transcription and Reward Systems; Transcription Control and the Circadian Clock; Transcriptional Silencing; Wnt Pathway and Neural Patterning.

Further Reading

Appel B and Eisen JS (2003) Retinoids run rampant: Multiple roles during spinal cord development. *Cell* 40: 461–464.

Brown M, Keynes R, and Lumsden A (2001) *The Developing Brain*. New York: Oxford Press.

Dasen JS, Liu JP, and Jessell TM (2003) Motor neuron columnar fate imposed by sequential phases of Hox-c activity. *Nature* 425: 926–933.

Dasen JS, Tice BC, Brenner-Morton S, and Jessell TM (2005) A Hox regulatory network establishes motor neuron pool identity and target-muscle connectivity. *Cell* 123: 477–491.

Diez del Corral R and Storey KG (2004) Opposing FGF and retinoid pathways: A signaling switch that controls differentiation and patterning onset in the extending vertebrate body axis. *BioEssays* 26: 857–869.

Helms AW and Johnson JE (2003) Specification of dorsal cord interneurons. *Current Opinion in Neurobiology* 13: 42–49.

Lee KJ and Jessell TM (1999) The specification of dorsal cell fates in the vertebrate central nervous system. *Annual Review of Neuroscience* 22: 261–294.

Lee SK and Pfaff SL (2001) Transcriptional networks regulating neuronal identity in the developing spinal cord. *Nature Neuroscience* 4(supplement): 1183–1191.

Lee SK and Pfaff SL (2003) Synchronization of neurogenesis and motor neuron specification by direct coupling of bHLH and homeodomain transcription factors. *Neuron* 38: 731–745.

Ross SE, Greenberg ME, and Stiles CD (2003) Basic helix–loop–helix factors in cortical development. *Neuron* 39: 13–25.

Sanes DH, Reh TA, and Harris WA (2000) *Development of the Nervous System*. London: Academic Press.

Shirisaki R and Pfaff SL (2002) Transcriptional codes and the control of neuronal identity. *Annual Review of Neuroscience* 25: 251–281.

Zhuang B and Sockanathan S (2006) Dorsal–ventral patterning: A view from the top. *Current Opinion in Neurobiology* 16: 20–24.

Transcriptional Silencing

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

Transcriptional silencing is the process by which regions of the eukaryotic genome, ranging from single genes to entire chromosomes, are made transcriptionally inactive through the reorganization of chromatin structure. Chromatin is the material in the nucleus of a eukaryotic cell comprising the nuclear DNA and all of the proteins associated with it. The bulk of this protein content is histones, basic proteins which package the DNA into nucleosomes, each nucleosome consisting of 146 bp of DNA wrapped around a histone octamer containing two copies each of the four core histones H2A, H2B, H3, and H4 (Figure 1). Linker DNA separates adjacent nucleosomes and is, in many species, associated with histone H1. There are also many other proteins in chromatin, including transcriptional activators and repressors, enzymes that modify DNA and histones, enzymes that cleave and repair DNA, components of the replication and recombination machinery, and proteins involved in higher order packaging. Under the microscope, this packaging is revealed by the presence of two different forms of chromatin – densely staining heterochromatin, which is for the most part transcriptionally inactive, and pale euchromatin, which is generally transcriptionally active. Transcriptional silencing reflects the fact that heterochromatic DNA is largely inaccessible to RNA polymerase and other components of the transcriptional machinery. Silencing differs from general transcriptional repression in that the former can extend over large regions of DNA, up to and including entire chromosomes, whereas the latter is specific to individual genes (or in some cases gene clusters). Silencing is also an epigenetic phenomenon which, once established, can be maintained passively through successive cell divisions, whereas gene-specific repression depends on the active maintenance of DNA–protein interactions within the promoter and other regulatory elements of the target gene. Transcriptional silencing is a natural phenomenon which is used to regulate gene expression during the development of the nervous system and as part of it normal function in adulthood. However, it is also used as a defense mechanism to protect the genome against invasive nucleic acids, such as transposable elements, viruses, and endogenous repeat sequences.

Mechanisms of Transcriptional Silencing

Histone Modification

Transcriptional silencing is generally achieved by the structural reorganization of chromatin, a process known as chromatin remodeling. In the brain, chromatin remodeling brought about by histone modification is fairly well characterized, and although much remains to be learned about the precise mechanisms involved, certain key events are now understood in detail. Histones are subject to various forms of reversible posttranslational modification, including the addition and removal of small chemical groups (acetylation, methylation, phosphorylation), larger chemical adducts (ADP-ribosylation), and small proteins (ubiquitylation, SUMOylation). Histones have a characteristic N-terminal tail which contains lysine and arginine residues and in some cases serine/threonine and glutamate. The lysine and arginine residues may be subject to acetylation and methylation, lysine residues may also be ubiquitylated or SUMOylated (SUMOs are small ubiquitin-like modifier proteins), the serine/threonine residues may undergo phosphorylation, and glutamate residues may undergo ADP-ribosylation. The roles of many of these modifications are unclear and may be context specific, but hyperacetylation usually corresponds to active chromatin whereas hypoacetylation correlates with inactive, silenced chromatin. It is thought that histone modification alters the affinity between histones and DNA, and/or between the histones within the nucleosome core, such that the DNA becomes either more or less accessible to the transcriptional machinery. Histone modification may also result in the specific recruitment of certain transcriptional activators and repressors, or proteins that sequester DNA into higher-order structures. There may well be a histone code, in which specific combinations of modifications contribute to defining a state of activity or inactivity which can persist through multiple rounds of DNA replication.

Histone modification is mediated by specific enzymes, and those controlling acetylation and methylation are the best understood. Histone acetyltransferases (HATs) catalyze the acetylation of lysine and/or arginine residues, and a number of known transcriptional activators have been subsequently identified as HATs. Histone deacetylases (HDACs) catalyze the removal of acetyl groups and associate with transcriptional repressors. The actual state of chromatin in any region of the genome depends to a large extent on the balance of these two opposing enzyme activities. In a similar way, the methylation of lysine or arginine residues is

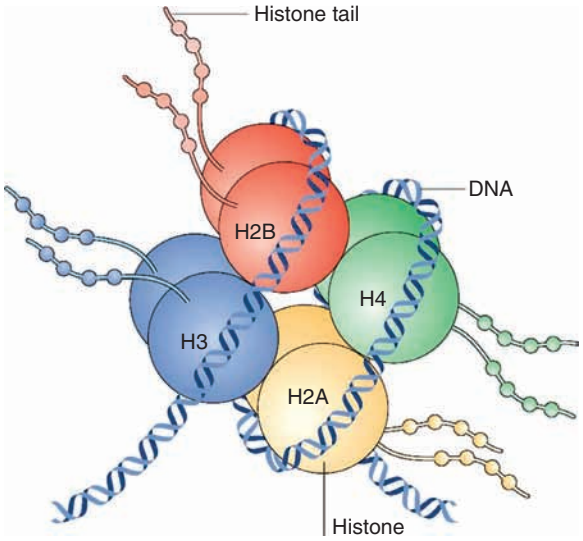

Figure 1 Nucleosome structure showing 148 bp of DNA wrapped around a histone octamer consisting of two copies each of the four core histones H2A, H2B, H3, and H4. Reproduced from Tsankova N, Renthal W, Kumar A, et al. (2007) Epigenetic regulation in psychiatric disorders. *Nature Reviews Neuroscience* 8: 355–367, with permission from Nature Publishing Group.

promoted by histone methyltransferases (HMTs) and reversed by histone demethylases (HDMs), although the role of methylation in gene silencing is not always so clear-cut, as methylation may be associated with either active or inactive chromatin in different contexts.

DNA Methylation

The structure of eukaryotic DNA is remarkably invariant, but epigenetic regulation by DNA methylation constitutes an important exception. In mammals, DNA is methylated at the 5′ position of cytidine residues in the context of the dinucleotide sequence CpG. The methylation of DNA is catalyzed by DNA methyltransferases (DNMTs), which transfer the methyl group from an *S*-adenosylmethionine (SAM) donor. Generally, CpG dinucleotides are constitutively methylated and tend to mutate to TpG by deamination over evolutionary timescales, resulting in significant underrepresentation in today's mammalian genomes. However, those in promoters and other regulatory regions are relatively undermethylated, and therefore escape the evolutionary pressure to mutate, resulting in clusters of dinucleotides known as CpG islands, which have been used to pinpoint the location of genes in genomic DNA clones. Within such islands, the level of DNA methylation corresponds to the degree of transcriptional silencing.

As might be expected, patterns of DNA methylation correspond to patterns of histone modification that establish inactive chromatin. This is achieved through the presence of methyl-CpG-binding proteins

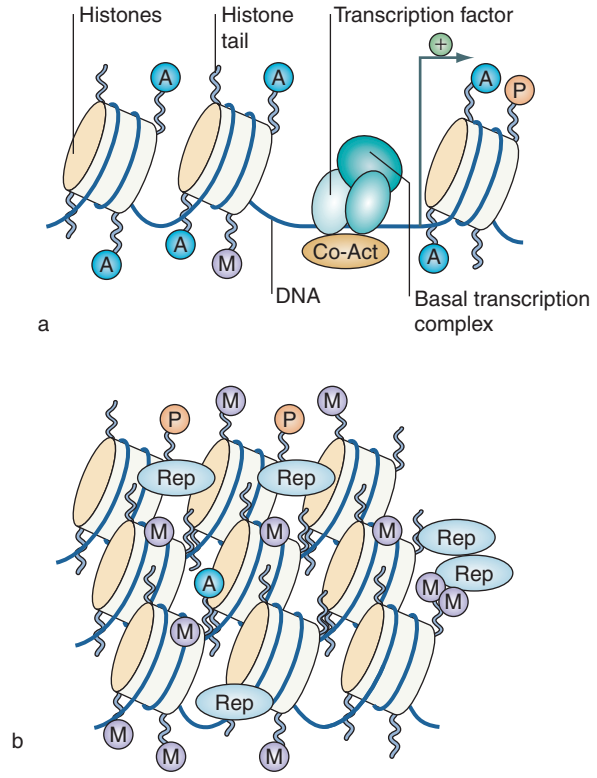

Figure 2 (a) Active chromatin (euchromatin) consists of loosely packed nucleosomes in which the histones are heavily acetylated (A) but are methylated (M) and phosphorylated (P) at low levels. This enables the basal transcription complex and coactivators (Co-Act) to assemble on the promoter, and transcription to commence. (b) Repressed chromatin (heterochromatin) consists of tightly packed nucleosomes in which histones are heavily methylated but are acetylated at low levels. Further repressor proteins (Rep) are also present. Reproduced from Tsankova N, Renthal W, Kumar A, et al. (2007) Epigenetic regulation in psychiatric disorders. *Nature Reviews Neuroscience* 8: 355–367, with permission from Nature Publishing Group.

(MeCPs) that associate with protein complexes containing HDACs and HMTs. DNA methylation, histone methylation, and histone deacetylation are therefore intrinsically linked in order to facilitate transcriptional silencing, establish epigenetic markers, and allow the silenced state to persist through DNA replication and cell division (Figure 2).

Other Chromatin Remodeling Mechanisms

In addition to histone modification, chromatin remodeling can be brought about by a number of other, less well-characterized mechanisms. One example is the phenomenon of histone replacement. Here, instead of modifying histones *in situ* to alter their affinity for DNA and components of the transcriptional machinery, one set of histones is simply swapped for another with properties that are more suitable. Another example is nucleosome sliding, an ATP-dependent process in which the histone octamer is transiently displaced,

allowing the transcriptional machinery to assemble at the promoter. Such repositioning of nucleosomes has been observed directly *in vitro* and is thought to be mediated by proteins of the SWI/SNF family.

Higher-Order Packaging and Nuclear Organization

A key difference between euchromatic and heterochromatic DNA is that the latter is folded into higher-order chromatin structures that maintain the DNA in an inaccessible state. This higher-order packaging is mediated by proteins that recognize the types of histone modification associated with inactive chromatin. For example, heterochromatin protein 1 (HP1) binds specifically to nucleosomes containing histone H3 methylated on lysine-9. A second mechanism for regulating access to the DNA then becomes evident and this involves the spatial positioning of genes within the nucleus. Interphase DNA in the nucleus does not diffuse about randomly, but is instead organized into discrete areas by attachment to the nuclear matrix. Labeling experiments have shown that chromosomes, and even small chromosome regions, tend to have a specific address, known as a chromosome territory, within the nucleus. Not much is known about the precise way in which the territories are established and occupied.

The nucleus is also organized into zones which are enriched with proteins that favor either transcriptional activity (active zones) or transcriptional silencing (inactive zones). Nuclear zones and territories are not aligned, and it appears that the zones are actually superimposed on the territorial organization so that they can incorporate whole chromosomes (e.g., the inactivated X chromosome in female mammals), contiguous chromosome segments, or even regions of similar transcriptional activity on different chromosomes which come together in the same locality. Thus, regulation of gene expression can occur at many levels involving the local chromatin environment, chromosomal territories, and the nuclear zones (Figure 3).

Recruitment to the different zones is mediated by protein–protein interactions, as illustrated by HP1, which not only interacts with modified histones but also with components of the nuclear envelope, thereby positioning heterochromatin within the inactive zone at the nuclear periphery. Specialized DNA elements known as insulators or boundary elements are thought to separate adjacent chromatin regions and allow them to occupy different zones, perhaps by binding to the nuclear matrix and preventing the propagation of heterochromatin structures from one domain to the next (see later, discussion of position effects).

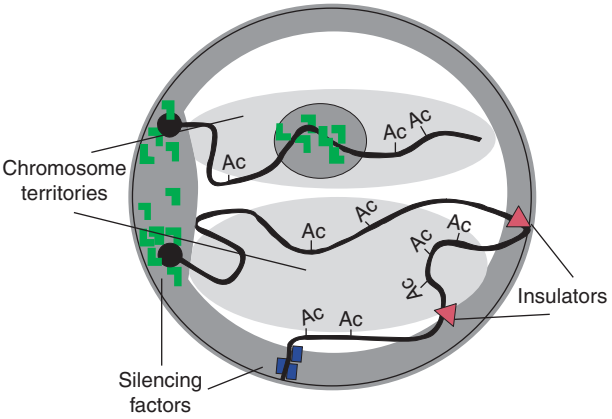

Figure 3 Nuclear organization: the nucleus contains territories (light gray) and silencing zones (dark gray periphery). Histone acetylation (Ac) is shown over active chromatin, while hypoacetylation is associated with inactive chromatin, such as that found at centromeres (black circle) and telomeres. Small domains within a chromosome can acquire properties similar to those of heterochromatin and form a silencing zone (shaded circle). Silencing factors (green and blue shapes) are involved in chromatin compaction, gene silencing, and tethering to the nuclear periphery. Insulator elements (red triangles) partition chromosomes into transcriptionally related domains and assist in zonal organization. Reproduced from Tsankova N, Renthal W, Kumar A, et al. (2007) Epigenetic regulation in psychiatric disorders. *Nature Reviews Neuroscience* 8: 355–367, with permission from Nature Publishing Group.

Events That Trigger Transcriptional Silencing

Endogenous Signaling Pathways

Transcriptional silencing is widely used during the development of the nervous system to maintain states of determination and differentiation as permanent developmental decisions. It is also becoming apparent that transcriptional silencing is used in mature neurons as a mechanism to implement long-term changes in response to external signals. Most brain disorders are characterized by dysregulation of gene expression, resulting in long-lasting effects on behavior. However, such observations clash with what is known about conventional gene regulation, as well as with evidence from animal models showing that the levels of transcription factors and other regulators return to normal within hours or days of perturbation, yet abnormal behavior persists. The possibility that epigenetic mechanisms such as the imposition or lifting of transcriptional silencing could mediate these long-term changes is therefore an attractive one, and a number of neuronal signaling pathways have been shown to fit with this idea. For example, the transcription factor cyclic AMP response element-binding protein (CREB) is phosphorylated on a serine residue by extracellular signal-regulated kinase (ERK) in response to signals that elicit either cAMP or Ca^{2+} as second messengers. Phosphorylated CREB can then recruit CREB-binding

protein (CBP) into a complex, and the intrinsic HAT activity of CBP acetylates histones, making the target gene transcriptionally active. In hippocampal neurons, increases in intracellular Ca^{2+} lead to the activation of Ca^{2+}/calmodulin-activated protein kinases (CaMKs), which phosphorylate class II HDACs. This phosphorylation triggers the shuttling of the enzyme out of the nucleus and results in increased histone acetylation. In both cases, the epigenetic mark for transcriptional activity would persist until erased by HDAC activity.

It is also interesting to note that certain neuroactive drugs are known to induce rapid changes in histone modification. For example, cocaine induces the acetylation of histone H4 and the phosphoacetylation of histone H3 in the promoter of the immediate-early gene c-*fos*. This occurs specifically in the striatum, which mediates the behavioral effects of cocaine.

DNA Rearrangement and Integration

In addition to being a significant feature of normal gene regulation, transcriptional silencing can also occur due to abnormal or unanticipated events. In chromosome rearrangements such as inversions and translocations, DNA at the breakpoint is often subject to what is known as a position effect, where active genes can undergo transcriptional silencing due to the spreading of inactive chromatin. This occurs because the breakpoint brings active and inactive chromatin into juxtaposition without the normal insulator or boundary element to separate them into distinct nuclear zones. What then occurs is down to chance, reflecting the extent of histone modification activity and the abundance of heterochromatin proteins available in the nuclei of different cells. A well-characterized model is the developing eye of the fruit fly *Drosophila melanogaster*. Under normal circumstances, the *white* gene (which encodes an enzyme involved in pigment synthesis and therefore confers the normal red eye color when active) is located far from any heterochromatin. A rare chromosomal inversion brings the *white* gene much nearer, and removes an insulator which normally keeps the heterochromatin restricted to its own zone. The heterochromatin then spreads along the chromosome depending on the availability of the necessary components and enzymes. In some cells of the developing eye, the heterochromatin fails to reach the relocated *white* gene, which remains active. In others, it spreads through the *white* gene and inactivates it. Since the epigenetic mark is heritable, each cell gives rise to a clone of cells in the adult eye, which becomes either a white or red patch. The overall eye is therefore mottled, or variegated, a phenomenon known as 'position effect variegation.'

Positional silencing can also occur when DNA integrates into a heterochromatic region, although it is necessary to establish whether the absence of gene expression is genuinely because an intact transgene has integrated into inactive chromatin, or whether the integration event has resulted in transgene rearrangement or mutation.

Repetitive DNA

Repetitive DNA sequences (both endogenous sequences and transgenes) are often subject to transcriptional silencing because they act as nucleation centers for heterochromatin formation. This occurs both at tandem repeats, such as those found in the centromeric region of chromosomes, and at dispersed repeats, such as transposable elements and integrated transgenes. In some cases, repeated sequences may trigger heterochromatin formation because they attract sequence-specific DNA-binding proteins which are in turn recognized by HDACs or DNMTs, or proteins such as HP1. In yeast, for example, repetitive DNA sequences found at the telomeres are recognized by the DNA-binding protein Rap1, which interacts with components of the heterochromatin machinery and ensures localization at the nuclear periphery. Since mammalian transposable elements and transgenes are diverse in their sequences, it is unlikely that a sequence-specific mechanism is involved and suggests that the cell possesses some kind of mechanism for sensing repeated DNA sequences and targeting them. Models have been proposed based on the formation of DNA–DNA pairs that act as nucleation sites, as well as models involving the production of aberrant RNA species, invoking the RNA-mediated interference (RNAi) pathway (see later).

Invasive DNA

Transgene silencing can occur even following the integration of a single-copy gene at a euchromatic site, suggesting that cells can under some circumstances recognize integrating DNA sequences as invasive and deal with them by sequestering them into heterochromatin following *de novo* DNA methylation. It is possible that cells possess some form of genome scanning apparatus that looks for sequences with unusual base composition compared to the surrounding genomic context. The same mechanism is probably responsible for the *de novo* methylation of integrating human immunodeficiency virus (HIV) genomes, and vectors based thereon, as well as retroviral vectors used to introduce transgenes into mammalian embryos. Prokaryotic DNA appears to be a strong trigger for *de novo* methylation, since the co-integration of vector backbone sequences (which are

generally of bacterial origin) often leads to silencing, whereas clean DNA sequences (just the promoter, coding region, and polyadenylation site) are more successful. It is also possible that invasive DNA somehow induces the RNAi pathway (e.g., through the expression of aberrant RNA species from integrated transgenes, or through the generation of double-stranded RNA intermediates during viral replication).

RNA Silencing

Small RNA molecules can also silence gene expression, but generally do so posttranscriptionally, either by cleaving homologous mRNA molecules (RNA interference) using short interfering RNAs (siRNAs) or by regulating translation (micro-RNAs). However, the RNA silencing apparatus also has significant interactions with chromatin proteins, including histone-modifying enzymes and DNA methyltransferases. Recent evidence has accumulated that the RNAi machinery can set epigenetic marks in *Drosophila* and mammalian cells. Although the mechanisms involved are complex and yet to be elucidated fully, it appears that repetitive regions of the genome such as tandem repeats and transposons (and perhaps integrated transgenes) can generate long double-stranded RNA molecules, which are processed into siRNA-like structures known as repeat-associated siRNAs (rasiRNAs). These are cleaved by the enzyme Dicer to form complexes known as RNA-induced initiation of transcriptional silencing (RITS); these complexes mediate histone modification and DNA methylation.

Nervous System Transcriptional Silencing

Transcriptional Silencing in Neural Development

Transcriptional silencing is used to establish and maintain cell fates in the developing nervous system and, most importantly, to suppress neuronal fates in nonneuronal cells. Embryonic cells and their derivatives have the potential to differentiate into neural cells if grown in the correct conditions (e.g., mouse P19 embryonal carcinoma cells will differentiate into neurons in the presence of retinoic acid), so it is apparent that this developmental lineage must be suppressed in normal development. There is evidence that this is achieved through chromatin remodeling, since HDAC inhibitors can induce neural differentiation in embryonic cortical cells and suppress the differentiation of oligodendrocytes. A more specific example is the neuron-restrictive silencing factor (NRSF), a transcriptional repressor expressed in glia and other nonneuronal cells. The function of NRSF is to repress neuronal differentiation; NRSF associates with a number of other proteins, including transcriptional repressors and the histone deacetylase HDAC2. In this way, the promoters of neuronal genes are maintained in an inactive state in nonneuronal cells. NRSF is also expressed in some neurons later in development, when its function is to regulate the activity of neuronal genes.

Transcriptional Silencing in Mature Neurons

Transcriptional silencing is also used to control gene expression in mature neurons, and a particularly striking example is the regulation of the *BDNF* gene (encoding brain-derived neurotrophic factor). The organization of this gene is very unusual in that there are at least nine distinct promoters, each paired with a unique noncoding exon. When the gene is expressed, one of these noncoding exons is spliced onto a common coding exon to generate one of at least nine possible mature transcripts. The promoters are independently regulated and have different levels of activity and different expression patterns, allowing BDNF to be expressed in different patterns depending on which promoter is activated by a given external signal. This activation is mediated by chromatin remodeling. For example, acute membrane depolarization in mouse cells releases intracellular Ca^{2+}, which results in the activation of Ca^{2+}/calmodulin-activated protein kinase II, causing the phosphorylation of MeCP2 and favoring transcription from promoters P1 and P4. In rats, electroconvulsive seizures or *in vivo* administration of pilocarpine increase histone H4 acetylation at promoter P2, strongly elevating transcription from this promoter.

See also: Gene Expression Regulation: Activity-Dependent; Gene Expression Regulation: Chromatin Modification in the CNS; Neural Crest Diversification and Specification: Transcriptional Control of Schwann Cell Differentiation; Neuromuscular Junction: Neuronal Regulation of Gene Transcription at the Vertebrate; Olfactory Insights from Transcriptional Profiling; Thyroid Hormone and Transcriptional Regulation in the CNS; Transcription Control and the Circadian Clock; Transcription and Reward Systems; Transcription Factors in Synaptic Plasticity and Learning and Memory; Transcriptional Networks and the Spinal Cord.

Further Reading

Ballas N and Mandel G (2005) The many faces of REST oversee epigenetic programming of neuronal genes. *Current Opinion in Neurobiology* 15: 500–506.

Coulson JM (2005) Transcriptional regulation: Cancer, neurons and the REST. *Current Biology* 15: R665–R668.

Feng J, Fouse S, and Fan G (2007) Epigenetic regulation of neural gene expression and neuronal function. *Pediatric Research* 61: 58R–63R.

Gasser SM (2001) Positions of potential: Nuclear organization and gene expression. *Cell* 104: 639–642.

Grewal SIS and Moazed D (2003) Heterochromatin and epigenetic control of gene expression. *Science* 301: 798–802.

Hsieh J and Gage FH (2005) Chromatin remodeling in neural development and plasticity. *Current Opinion in Cell Biology* 17: 664–671.

Perini G and Tupler R (2006) Altered gene silencing and human diseases. *Clinical Genetics* 69: 1–7.

Santos M, Coelho PA, and Maciel P (2006) Chromatin remodeling and neuronal function: Exciting links. *Genes, Brain and Behavior* 5: 80–91.

Tsankova N, Renthal W, Kumar A, et al. (2007) Epigenetic regulation in psychiatric disorders. *Nature Reviews Neuroscience* 8: 355–367.

Williams RRE (2003) Transcription and the territory: The ins and outs of gene positioning. *Trends in Genetics* 19: 298–302.

Transgenic Models of Neurodegenerative Disease

C Li, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

The rapid advances in human genetics have been accompanied by a major increase in transgenic animal research for the purpose of neurodegenerative disease study. It is impossible to cover all the transgenic models, their phenotypical analyses, and their usage in this research field. It is also unnecessary to do so because for any particular neurodegenerative disease of interest, a simple search in PUBMed or websites of pertinent disease research foundations will lead to many excellent review articles that are constantly being published with updated information. Therefore, this article introduces the transgenic approach rather than the particular animal models *per se*, explains the underlying principles of the transgenic approach, and compares different transgenic approaches and their advantages and disadvantages. For this purpose, only a small subset of the animal models are presented here.

The development of transgenic animal models has been closely linked to the revolution in human genetics. Most neurodegenerative diseases have either a genetic inheritance component or are purely genetic. Some examples of the former include Alzheimer's disease (AD), Parkinson's disease (PD), and amyotrophic lateral sclerosis, with approximately 10% of patients carrying genetic mutations. Examples of the latter include Huntington's disease (HD) and spinocerebellar ataxia type 1. For the purely genetic diseases, the reason for establishing and studying genetic animal models is self-explanatory. For diseases that are mostly sporadic and are only 10% genetic, why do we put such a major effort in studying the genetic forms of the diseases? First, although many changes or abnormalities have been found from studying the sporadic forms of diseases, in general it is difficult to determine whether these abnormalities are causal, innocuous, or compensatory. In other words, these abnormalities may be pathologic rather than pathogenic. In genetic forms of the diseases, the 'first pushes' of the disease pathogenesis have been unequivocally established – the mutations in disease genes. Thus, the subsequent research, albeit long and arduous, has a firm foundation in logic. Second, the pathogenic mechanisms of genetic forms of the diseases may shed light on the sporadic forms because the pathogenic pathways may converge at some points.

For the previously mentioned reasons, the study of purely genetic neurodegenerative diseases or the genetic forms of neurodegenerative diseases has been the focus of many researchers. In this line of work, although studies at molecular, cell culture levels are important, genetic animal models are clearly one of the most important tools for research.

The genetic diseases have two subtypes: recessive inheritance, which requires both chromosomes to be mutated, or dominant inheritance, which requires only one mutant allele. The former is usually caused by loss of function of gene products, and the latter is commonly caused by gain of function, hyperfunction, or haploinsufficiency. In neurodegenerative diseases, both types are common. For example, in PD, Park 1 and Park 8 have dominant inheritance due to the mutations in α-synuclein and LRRK2 genes, whereas Park 2, Park5, Park 6, and Park 7 have recessive inheritance due to the loss of function of Parkin, UCHL-1, DJ-1, and PINK 1, respectively.

To model these different types of genetic deficits, a variety of methods have been developed. For genetic loss-of-function mutations, researchers commonly use the gene targeting (i.e., knockout) approach in which the gene of interest is deleted from the genome of experimental animals. For genetic dominant diseases, scientists may choose the transgenic approach in which a mutated gene is introduced into the genome of experimental animals. Since the gene targeting approach is explained elsewhere in this encyclopedia, this article focuses on the transgenic approach. It is worth noting that many organisms have been used to generate transgenic models, including yeast, *Caenorhabditis elegans*, *Drosophila*, mice, rat, and sheep. The majority of work in neurodegenerative disease research, however, is mainly done in mouse because this experimental organism is a mammal with a brain structure similar to that of humans, and it is amenable to many highly sophisticated genetic engineering tools.

The generation of transgenic mice involves construction of DNA vectors carrying the gene of interest and subsequent microinjection of DNA vectors into fertilized mouse eggs. In a certain percentage of injected eggs, DNA vectors will integrate into the mouse genome as a transgene, and therefore transgenic mouse founder lines are obtained. It is important to note that the integration site of transgene is random, rather than targeted to the homologous chromosomal sites, and that the copy number of the transgene is random, ranging from a single copy to 100–200. A direct implication of the transgene random integration sites and variable copy numbers is that the expression of transgenes may vary widely among the founder lines, both in spatial and temporal

patterns and in abundance. The random insertion of transgene into mouse genome may disrupt other genes in some cases. Therefore, studies with transgenic mice usually require multiple transgenic lines to ensure that the phenotypes are authentically caused by the transgenes.

The choice of transgenes deserves some discussion. There are two options: a human gene with human disease mutations or the mouse homologous gene with human mutations. Empirically, the former works better in recapitulating the human disease. For example, in HD mouse models, the one carrying mouse Huntingtin with 111 CAG repeats developed much milder phenotypes than another model with human Huntingtin of similar CAG repeat. In α-synuclein mouse models for PD, human α-synuclein gene in transgenic mouse was able to recapitulate disease pathology, whereas mouse α-synuclein gene failed. Conceptually, the most straightforward explanation is that evolution has accumulated enough divergence that a mutation in human protein may or may not cause the same pathogenic conformational change in a mouse homolog. In fact, the human A53T mutation in α-synuclein caused Park 1, whereas in mice the 53T is a wild-type amino acid sequence. Hence, it seems that human genes are more appropriate for modeling human diseases in transgenic mice.

The methods of generating transgenic mice have evolved and become more powerful and versatile. Among the newer ones are bacterial artificial chromosome (BAC)-mediated transgenesis and inducible transgenic system. The traditional method uses a small plasmid vector with a capacity of less than 20 kb of DNA and with promoters such as prion, CMV, and Thy1. The small capacity precludes the regulatory elements such as enhancers, suppressors, and insulators that are all important for gene regulation. When the transgenes integrate into the mouse genome, they are influenced strongly by the adjacent chromosomal areas due to the positional effects that often cause misexpression of the transgenes.Compared to the traditional method, BAC-mediated transgenesis has several advantages. First, the conventional method accommodates only 10 to 15 kb DNA inserts, whereas the BAC vector accommodates up to 400 kb DNA as transgenes, which are large enough for most of the mammalian genes. Second, the large BAC capacity allows the transgene to be insulated from positional effects of insertion sites on the chromosome and to be regulated by its native promoters, enhancers, and suppressors; intron–exon structures; and 5′ and 3′ untranslated regions. Finally, the BAC method does not require assumptions about the splicing variant to be used as a transgene. BAC transgenesis has been used increasingly more often in neurodegenerative disease research.

Another powerful method is the inducible tet-on/tet-off system that allows inducible expression or suppression of transgenes. An elegant example of this method is provided by Yamamoto et al., who demonstrated the possibility of reversing pathology in HD mice by suppressing the expression of mutant Huntingtin.

It is clear that the transgenic mice are increasingly being used in modeling neurodegenerative disease. A further development is that in addition to mimicking the genotypes of human diseases, researchers are generating and using transgenic mice at a new level of sophistication. In an HD study, for example, mice with specific mutations in Huntingtin were made to test the role of proteolysis in HD pathogenesis, and mice expressing mutant HD in particular brain regions and cell types were used to address important questions of cell-autonomous versus non-cell-autonomous degeneration. Also reported are studies using transgenic mice to test genetic modifiers of HD. As a result of the collective efforts from many researchers, there are many transgenic mouse models for each neurodegenerative disease. The vast number of existing mouse models are powerful tools for researchers in this field.

However, despite the overall successful use of transgenic mice to model neurodegenerative diseases, it is a common problem that mice are much more resilient and often do not exhibit the cardinal phenotype of neuronal cell death. For example, in PD and AD research, mouse models were unable to recapitulate this most important aspect of pathology. To combat this obstacle, the transgenic rat approach may be a worthy alternative to create new opportunities. Although closely related in evolution, rats are different from mice and are closer to humans in many physiological aspects. In many pharmacological studies, rats have much more similar pharmacokinetics to humans. It is well-known that cardiovascular diseases are modeled in rats and are unable to be modeled in mice. In Huntington's disease modeling, rat models recapitulated striatal neuronal death with 57 CAG repeat length similar to human mutation, whereas mouse models needed much stronger insults of more than 150 CAG repeats to elicit mutant phenotypes. In chemical models for PD, rats are much more sensitive to rotenone than mice. In an α-synuclein study, viral expression of α-synuclein in rat substantia nigra resulted in loss of dopaminergic neurons, which was not observed in mouse models. These data indicate that rats may be more sensitive to some pathological insults than mice.

In summary, the transgenic animal approach has provided researchers unprecedented powerful tools

for mechanistic dissection of pathogenesis as well as development of therapeutics. The methodology, techniques, and usage of transgenics continue to improve. There is reason for optimism that deeper understanding of neurodegenerative disease pathogenesis will eventually help us fight these human diseases.

See also: Alzheimer's Disease: Neurodegeneration; Alzheimer's Disease: An Overview; Alzheimer's Disease: Molecular Genetics; Alzheimer's Disease: Transgenic Mouse Models; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Amyotrophic Lateral Sclerosis (ALS); Animal Models of Motor and Sensory Neuron Disease; Animal Models of Alzheimer's Disease; Animal Models of Huntington's Disease; Animal Models of Parkinson's Disease; Huntington's Disease; Huntington's Disease: Neurodegeneration; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration.

Further Reading

Alam M, Mayerhofer A, and Schmidt WJ (2004) The neurobehavioral changes induced by bilateral rotenone lesion in medial forebrain bundle of rats are reversed by L-DOPA. *Behavioural Brain Research* 151: 117–124.

Chandra S, Gallardo G, Fernandez-Chacon R, Schluter OM, and Sudhof TC (2005) Alpha-synuclein cooperates with CSPalpha in preventing neurodegeneration. *Cell* 123: 383–396.

Fleming SM, Fernagut PO, and Chesselet MF (2005) Genetic mouse models of parkinsonism: Strengths and limitations. *NeuroRx* 2: 495–503.

Hodgson JG, Agopyan N, Gutekunst CA, et al. (1999) A YAC mouse model for Huntington's disease with full-length mutant huntingtin, cytoplasmic toxicity, and selective striatal neurodegeneration. *Neuron* 23: 181–192.

Gong S, Yang XW, Li C, and Heintz N (2002) Highly efficient modification of bacterial artificial chromosomes (BACs) using novel shuttle vectors containing the R6Kgamma origin of replication. *Genome Research* 12: 1992–1998.

Graham RK, Deng Y, Slow EJ, et al. (2006) Cleavage at the caspase-6 site is required for neuronal dysfunction and degeneration due to mutant huntingtin. *Cell* 125: 1179–1191.

Gu X, Li C, Wei W, et al. (2005) Pathological cell–cell interactions elicited by a neuropathogenic form of mutant huntingtin contribute to cortical pathogenesis in HD mice. *Neuron* 46: 433–444.

Levine MS, Cepeda C, Hickey MA, Fleming SM, and Chesselet MF (2004) Genetic mouse models of Huntington's and Parkinson's diseases: illuminating but imperfect. *Trends in Neuroscience* 27: 691–697.

Lo Bianco C, Schneider BL, Bauer M, et al. (2004) Lentiviral vector delivery of parkin prevents dopaminergic degeneration in an alpha-synuclein rat model of Parkinson's disease. *Proceedings of the National Academy of Sciences of the United States of America* 101: 17510–17515.

Spires TL and Hyman BT (2005) Transgenic models of Alzheimer's disease: Learning from animals. *NeuroRx* 2: 423–437.

von Horsten S, Schmitt I, Nguyen HP, et al. (2003) Transgenic rat model of Huntington's disease. *Human Molecular Genetics* 12: 617–624.

Wheeler VC, Auerbach W, White JK, et al. (1999) Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. *Human Molecular Genetics* 8: 115–122.

Yamada M, Iwatsubo T, Mizuno Y, and Mochizuki H (2004) Overexpression of α-synuclein in rat substantia nigra results in loss of dopaminergic neurons, phosphorylation of α-synuclein and activation of caspase-9: Resemblance to pathogenetic changes in Parkinson's disease. *Journal of Neurochemistry* 91: 451–461.

Yamamoto A, Lucas JJ, and Hen R (2000) Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. *Cell* 101: 57–66.

Relevant Websites

http://www.alzforum.org – Alzheimer Research Forum.
http://www.hdfoundation.org – Hereditary Disease Foundation.
http://www.ncbi.nlm.nih.gov – National Center for Biotechnology Information.
http://www.nih.gov – National Institutes of Health.

Transient Receptor Potential (TRP) Channels

D E Clapham, Harvard Medical School, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

What Are Transient Receptor Potential Channels?

Transient receptor potential (TRP) is the abbreviation used for naming the transient receptor potential family of ion channel subunits. The nomenclature originally referred to photoreceptors that exhibited a transient voltage response to continuous light exposure in a blind *Drosophila* mutant. This phenotype resulted from the loss of function of a TRP ion channel subunit in the fly photoreceptor. With the sequencing of many genomes and by sequence homology, TRP channel proteins have been identified in all species to date with the exception of bacteria. However, it is likely that homologs will eventually be found in these cells as well. In mammalian cells, TRPs are widely expressed in all tissue types but tend to be most highly expressed in epithelia and endothelia. The nervous system contains all known TRP channel types.

Mammalian TRP channels are encoded by six loosely related families of ~30 TRP genes. These families are grouped by homology and are named C, V, M, A, P, and ML, for canonical, vanilloid binding, melastatin related, ankyrin repeat, polycystin, and mucolipin, respectively (Figure 1). TRP ion channel subtypes are not named based on an endogenous ligand, function, or selectivity because these aspects are disparate within this family of proteins. With the current lack of knowledge about their functions, their unifying feature is simply homology of their transmembrane domains. Most are subunits of nonselective ion channels, allowing Na^+, K^+, and Ca^{2+} ions to permeate the membrane. Thus, when activated under physiological conditions, TRP channels depolarize cells and increase intracellular Na^+ and Ca^{2+}. Since intracellular Ca^{2+} is low under normal conditions (50–100 nM), the most significant change for the cell is an increase in intracellular Ca^{2+}. As is detailed in the following sections, intracellular Ca^{2+} changes have manifold biological consequences. The data shown in Tables 1–4 summarize the major known physiological, pharmacological, and biophysical properties of TRP channels (these data are also available online at http://clapham.tch.harvard.edu).

TRP Channel Selectivity and Gating

TRP channel subunits are currently presumed to have six-transmembrane (6TM)-spanning segments (S1–S6; Figure 2) and almost certainly assemble as tetramers around a central pore. In analogy to the high-resolution structure of the voltage-dependent K^+ channel Kv1.2, it is assumed that each single 6TM protein subunit is composed of two elements, a pore-lining domain (S5–S6) and a ligand/voltage-sensing domain (S1–S4). All four S5–S6 pore domains, one from each subunit, surround the ion-conducting pathway. Each of the four ligand/voltage-sensing domains is connected to the pore segments by a highly flexible linker that contacts the C-terminal end of the S6 helix. However, few studies have tested the assumption that TRPs have only 6TM domains, and hydropathy plots suggest some TRP channels may have additional transmembrane helices at the N- or C-termini.

The TRP Pore and Selectivity Filter

Based on high-resolution structures of K^+ channels, residues between the S5 and S6 transmembrane segments dip into the membrane and line the pore. For K^+-selective channels, K^+ ions interact with carbonyl oxygen atoms along the backbones of these amino acids. Currently there is no detailed structural information for Na^+, Ca^{2+}, or nonselective ion channels, but the principles are presumably similar. The unifying aspect of ion channel selectivity is binding of the ion in the pore; the ion is an integral part of the conducting pore and affects its conformation.

Three or four negatively charged amino acids – aspartatic (E) or glutamic (D) acid residues, each contributed by a separate domain – regulate selectivity by binding Na^+ or Ca^{2+} in the pore of Na^+ and Ca^{2+} channels. For Ca^{2+}-selective channels, Ca^{2+} binding to the pore is the critical determinant; when Ca^{2+} is removed from a solution, Ca^{2+} channels become nonselective cation channels. Ca^{2+} channels bind Ca^{2+} tightly in order to be highly selective, but this comes at the expense of rates of Ca^{2+} permeation.

In general, TRP channels are nonselective, with $P_{Ca}/P_{Na} \leq 10$, and with reversal potentials in the range from 0 to +20 mV. There are four known exceptions: TRPM4b and TRPM5 are Na^+ selective while TRPV5 and TRPV6 are Ca^{2+} selective (Tables 3 and 4). The pore loops of TRP channels are highly variable, and several are relatively long. The length of the S5–S6 linker may be related to unique adaptations of the 'turret,' the region extending above the plasma membrane in the extracellular space.

TRP Channel Gating

Gating is the conformational change in a protein that opens its ion conduction pathway. These protein conformational changes absorb or release energy.

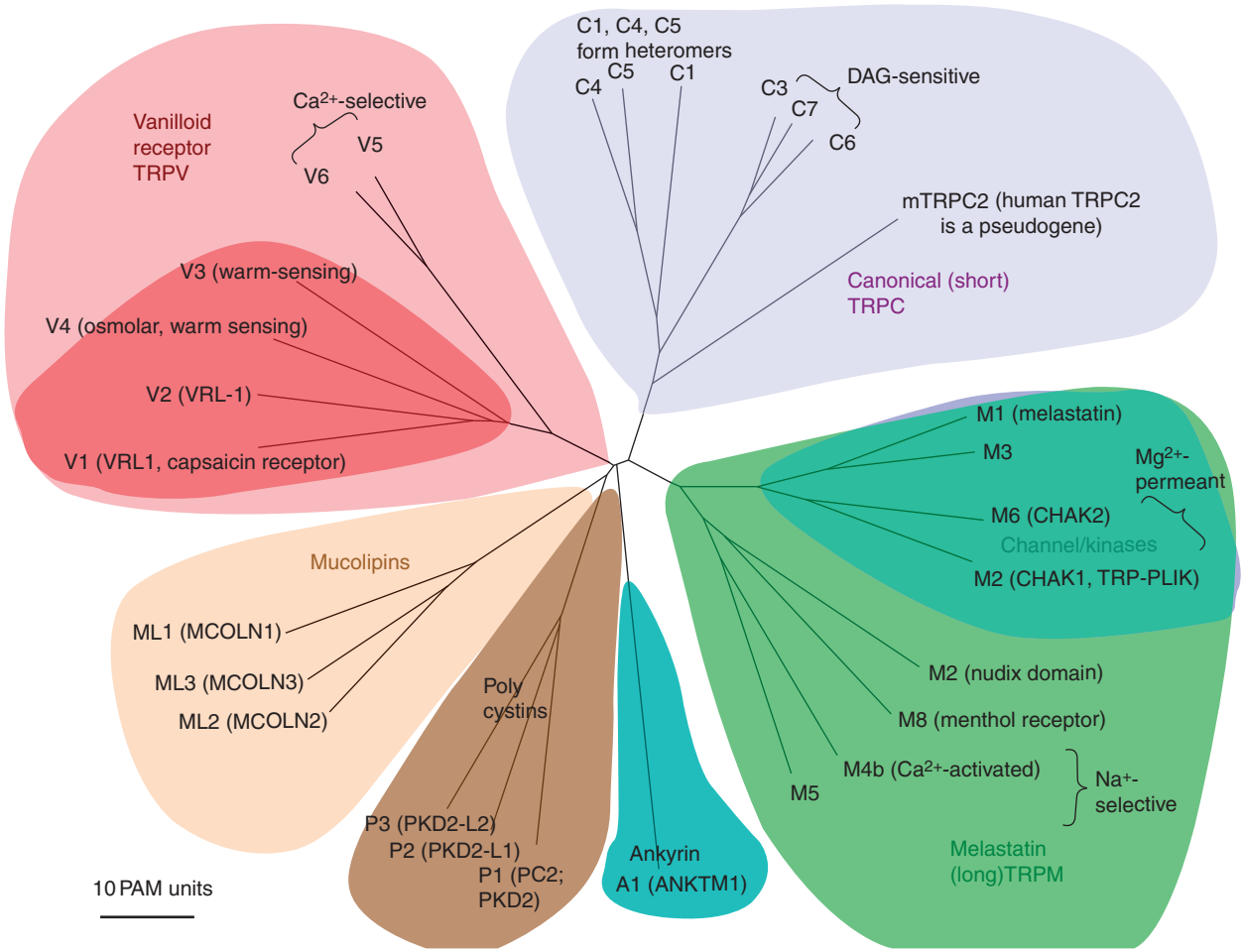

Figure 1 Mammalian transient receptor potential (TRP) family tree. Evolutionary branch lengths in 'point accepted mutation' (PAM) units, the mean number of substitutions per 100 residues (DAG, diacylglycerol; PLIK, serine/threonine kinase). Adapted from Clapham DE (2003) TRP channels as cellular sensors. *Nature* 426: 517–524, with permission.

The lowest energy a protein can directly detect is the thermal energy limit of ~0.6 kcal/mol (a single photon of visible light per molecule is 60 kcal/mol). Ion channel conformational changes that result in gating require energies >1 kcal/mol, but are typically in the range of 5–50 kcal/mol. Energy changes leading to gating are imparted by voltage changes across the membrane, ligand binding, and temperature changes.

Channel gating is a stochastic process and is described by the Boltzmann equation. This equation describes the distribution of channels in the open and closed states at equilibrium. The fraction of open channels is $[1 + \exp(\Delta G/Nk_BT)]^{-1}$, where ΔG is the free energy of transition between the closed and open state, N is Avogadro's number, and k_B is Boltzmann's constant; $Nk_B = R$, the gas constant, and T is the absolute temperature. The free energy, ΔG, is the difference between the conformational energy change upon gating and the sum of energies transduced to the channel from external forces.

At resting membrane potentials (e.g., $V = -70$ mV), the closed state of voltage-sensitive channels is lower in energy than is the open state, and thus is the favored conformation of the protein complex. The cell is thus a battery that provides the inhibitory voltage that keeps the channel in its closed state. Depolarization (reduction of the transmembrane voltage) allows the channel protein to relax into its open (conducting) structural configuration. Classical voltage-gated ion channels (Ca_V, Na_V, and K_V) have several charged amino acids in their S4 transmembrane helices that sense and move in response to a changing electrical field. For these ion channels, the Boltzmann distribution of closed:open states changes over a narrow range of voltage. Strongly voltage-sensitive channels provide fast, decisive changes in transmembrane ion flow in response to changes in potential. The speed of the primary, regenerative pulse (action potential) of nerves and muscle depends on this sharp voltage dependence.

Ligand binding to a protein also imparts energy to induce conformational changes in channels (K_D values of 1 μM imply ~1–10 kcal/mol). Examples of ligands that open ion channels are extracellular

Table 1 TRP channels table – TRPA (ankynin), TRPP (polycystins), and TRPML (mucolipins)

TRP subtype *Unigene*	*A1 (ANKTM1, P120)* *Hs.137674*	*P1 (PC2, PKD2, TRPP2)* *Hs.82001*	*P2 (PKD2L1)* *Hs.159241*	*P3 (PKD2L2)* *Hs.272418* *Splice variants*	*ML1 (MCOLN1)* *Hs.12909*	*ML2 (MCOLN2)* *Hs.210792* *Splice variants*	*ML3 (MCOLN3)* *Hs.49344* *Splice variants*
# Amino acids							
Human	1119	968	805	624	580	566	553
Rat	1125	915	N.D.	N.D.	580	566	553
Mouse	1125	966	760	621	580	566	553
Accession #							
Human	Q75762	NP000288	AF094827	NP055201	NP065394	NP694991	NP060768
Rat	Q6RI86	XP573552	N.D.	N.D.	XP213684	NP001034094	NP001012059
Mouse	Q8BLA8	NP032887	NP852087	NP058623	NP444407	NP080932	NP598921
Splice variants and other accession #s	075762 Hs 1119 Q8BLA8 Mm 1125	Q13563 Hs 968 NP_000288* Hs 968 035245 Mm 966	Q9P0L9 Hs 805	Q9NZM6 Hs 624 NP_055201 Hs 609 Q9JLG4 Mm 621	Q9GZU1 Hs 580 Q99J21 Mm 580	NP_694991 Hs 566 Q8IZK6 Hs 538 Q5T4H7 Hs 249 Q8K595 Mm 566	Q8TDD5 Hs 553 Q5T4H4 Hs 321 Q8R4F0 Mm 533
OMIM gene map							
Human	8q13	4q21–q23	10q24	5q31	19p13.2–13.3	1p22.3	1p22.3
Rat	5q11	N.D.	N.D.		12	2q44	2q44
Mouse	1 A3	5 55.0 cM	19 44.0 cM	18 band C	8 1.1 cM	3	3 74.8 cM
Genetic model or disease mutation	Mouse TRPA1 −/− No hearing defect Pain defect disputed	Polycystic kidney Disease (ADPKD) Mutations in TRPP1 in humans produce polycystic kidney disease, a progressive increase in cyst formation in the kidney. With an onset usually in middle age, it leads to progressive kidney dysfunction and failure. Also associated with cyst formation in liver and other organs. Mouse PKD2−/−	*krd* mice (P2−/− with other deletions in pax) Targeted DT-mediated disruption in taste cells in mouse	N.D.	Mucolipidosis IV is associated with mutations in TRPML1. For details, see 'Molecular Genetics', and 'Animal Model' in OMIM (Online Mendelian inheritance in man website supported by Johns Hopkins University and the NCBI; +252650)	N.D.	Varitint-waddler mouse (Va) A419P in 5th TM domain. Mice heterozygous for the varitint-waddler mutation are deaf and show varying degrees of circling behavior, head-tossing, and hyperactivity. Their coats are white with patches of normal-colored fur. The pathological changes in heterozygotes affect degeneration of the organ of Corti, stria vascularis, spiral ganglion, saccular macula, cristae ampullares, and vestibular ganglion. Viability of heterozygotes is nearly normal, but fertility is reduced. Mortality is very high in homozygotes, and very few of the survivors are fertile
Known functions	Senses various noxious stimuli: pungent natural compounds, noxious cold temperature, environments irritants, inflammatory peptides, noxious mechanical stimuli	Mouse cardiac, skeletal and renal development; kidney and liver cyst formation With PC2(PKD2) mechano-sensation in cilia (kidney)	Kidney development (mice) Acid sensing in sour taste, and potentially in acid sensing in the spinal cord along the central canal	N.D.	Cell vacuolar defects-probably defect in sorting/ transport in late endocytic pathway	N.D.	Hair cell, stereocilia maturation

Continued

Table 1 Continued

TRP subtype *Unigene*	*A1 (ANKTM1, P120)* *Hs.137674*	*P1 (PC2, PKD2, TRPP2)* *Hs.82001*	*P2 (PKD2L1)* *Hs.159241*	*P3 (PKD2L2)* *Hs.272418* *Splice variants*	*ML1 (MCOLN1)* *Hs.12909*	*ML2 (MCOLN2)* *Hs.210792* *Splice variants*	*ML3 (MCOLN3)* *Hs.49344* *Splice variants*
Likely role in	Inflammatory pain, cold hyperalgesia, mechanical pain, mechanical hypersensitivity	Cilia fertility Proposed to localize to cilia, where it would be activated by ciliary movement. Proposed to act as an intracellular calcium release channel	Retinal development (mice)	Testis	Endosomes		Intracellular vesicle defect
N or C terminal domain modules	14 N-AnkR	C-EF hand + C-coiled coil			N-proline Leucine zipper S2 Nuclear local. aa 43–60 C-di-leucine		
Reported interacting proteins	N.D.	PC1 (PKD 1), Hax-1, Cortactin, Cadherin via PC1, Id2, mDia, troponin, tropomyosin, actinin	PKD1L3?	N.D.		May form a heteromer with TRPML3 in lysosomes	May form heteromers with TRPML1 or TRPML2 in lysosomes
Activation	PLC-coupled GPCRs, $T < 18\,^{\circ}C$ disputed voltage-dependence not yet described, inactivation by extracellular Ca^{2+} might be voltage dependent	Mechanical stress? $[Ca^{2+}]_i$? Putative nonselective, divalent permeant ion channel that activates above −50 mV.	~1 μM $[Ca^{2+}]_i$	N.D.	$[Ca^{2+}]_i$?	N.D.	N.D.
Direct agonist	Garlic (thiosulfinate; allicin; 1 μM); Mustard oils in wasabi, horseradish (allyl isothiocyanates; 10 μM), D9-THC (12 μM). carvacrol (0.3 mM). Acrolein (5 μM), 2-pentenal, Icilin (10 μM), cinnamaldehyde (60 μM), eugenol (0.6 mM), gingerol (0.6 mM)	pH? Cai?		N.D.			
$\sim P_{Ca}/P_{Na}$	1.4 $P_{Ca} \sim P_{Mg}$	~1–5	~4	N.D.	~1 Permeant to Ca, Ba, Sr, Na, K		
Blockers and inhibitors	RuR (3.5 μM) Gd (0.1 μM) Amiloride (0.5 mM) Camphor (70 μM) Menthol (>100 μM) Gentamicin desensitization by carvacrol	Gd, La, Cd SKF96365? 2-APB (50 μM)?	Cd, La, flufenamate?	N.D.	Amiloride (μM) Gd, La, Ni (mM)		
Single channel conductance (γ) in pS	98 pS in physiol solns	~40 pS high Ca_0 (Luo et al.), ~50 pS 140K (Gonzalez) 90–115 pS in sym 50 mM Ba	130 pS (100 mM Na or 80 mM Ba) Ca>K~Na	300 pS?	83 (100 mM Na) LaPlante		

I–V shape	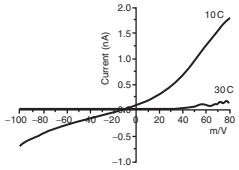 Story et al. 2003	Polycystin 1+P2; Linear (Germino et al. 2000) 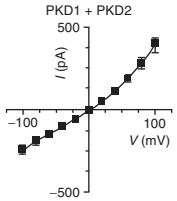	~linear 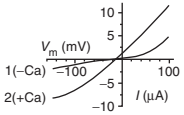 From Chen et al. *Nature*	N.D.	N.D.	N.D.	N.D.
Tissue distribution	DRG, trigeminal ganglia, nodose ganglia, sensory nerve fibers (C-fibers), nociceptive neurons, inner ear, heart, small intestine, colon, lung, kidney, pancreas, skin	Wide distribution: kidney, pancreas, heart, testis, ovary, intestine, blood	High in fetal tissues including kidney and liver, downregulated in adult. Widely distributed, adult heart, skeletal muscle, brain, spleen, testis, retina	Testis Isoform 2/PKD2L2b is found only in transformed lympho-blasts whereas isoform 3/PKD2L2a widespread	Ubiquitous	Lymphocyte B cells	Hair cells of inner ear
Comments	TRPA1 is activated by covalent modification of intracellular cysteines, activation by bradykinin, may be activated by noxious cold but disputed, mechano-activation still under discussion. TRPA1 is modulated by extra- and intracellular Ca^{2+} decreased in fibroblasts after oncogenic transforma-tion; expression is upregulated by deubiquitina-tion by the tumor suppressor (ubiquitin hydrolase) CYLD	TRPP1 (PKD2) heterozygous and null mice develop polycystic kidney and liver lesions that are like those in humans. A 'two-hit' model was proposed to mimic mutations occurring in type 2 autosomal dominant polycystic kidney human disease. TRPP1−/− mice die *in utero* between E13.5 and birth, with defects in cardiac septation and cyst formation in maturing nephrons and pancreatic ducts. Adult TRPP1+/− for an unstable allele mimic the human development of renal cysts and eventual kidney failure with increased severity with modification of both PKD1 and TRPP1. Intracranial vascular abnormalities occur in TRPP1+/− mice with surgically-induced hypertension. Expression of PKD1 activates the JAK-STAT pathway, which in turn upregulates p21 (waf1) to arrest the cell cycle. Mutations that disrupt PKD1 interactions	No known association with human polycystic kidney disease. TRPP2 (PKD2L1) is deleted in the krd mouse which has defects in structures of kidney, retina, optic disc; however the krd deletion is a 7 cM transgene-induced deletion on chromosome 19 that includes the Pax2 locus	Probably testis-specific function	Role in sorting/transport in late endocytic pathway, defective in mucolipidosis type IV lysosomal storage disease		Mouse early-onset hearing loss, vestibular defects, pigmentation abnormalities

Continued

Table 1 Continued

TRP subtype *Unigene*	*A1 (ANKTM1, P120)* *Hs.137674*	*P1 (PC2, PKD2, TRPP2)* *Hs.82001*	*P2 (PKD2L1)* *Hs.159241*	*P3 (PKD2L2)* *Hs.272418* *Splice variants*	*ML1 (MCOLN1)* *Hs.12909*	*ML2 (MCOLN2)* *Hs.210792* *Splice variants*	*ML3 (MCOLN3)* *Hs.49344* *Splice variants*
		with TRPP1 prevent activation of this pathway. PKD1−/− mouse embryos have defective STAT1 phosorylation and p21 (waf1) induction					
REFS	Jaquemar, 10066796; Jordt, 14712238; Story, 12654248 Nagata, 15843607; Bandell, 15046718; Bautista, 16103371, 15916949; Bautista, 16564016 Xu, 16617338; Hinman, 17164327 Macpherson, 16829128 Corey, 15483558 Kobayashi, 16304633 Stokes, 16500080 Zanotto, 17151223; Kwan, 16630838; Dhaka, 16602906 Rosenzweig, 15681611; Corey, 16901942; Patapoutian, 16675993; Katsura, 16546170 Nilius. *Physiol Rev* (2007) Obata, 16110328	Mochizuki, 8650545; Wu, 9339380 Pennekamp, 9716661; Qian, 9171830 Tsiokas, 9192675; Hanaoka, 11140688; Tsiokas, 10097141 Gallagher, 10760273; Li Q, 12525172 Li Q, 12527301; Li, 15843396 Rundle, 15123714; Li, 16223735 Li, 16311606; Luo, 12640140 Koulen, 11854751; Gonzalez-Perrett, 11991947; Pelucchi, 16396967 Obermuller, 10600939; Nauli, 12514735; Wu, 9568711; Pei, 10405208; Wu, 10615132 Wu, 12140187; Qian, 12874107 Bhunia, 12007403 Xenophontos, 9175744; Pei, 9573526 Koptides, 9949210; Viribay, 9402976 Reynolds, 10541293	Nomura, 9748274 Wu, 9878261 Guo, 10602992 Murakami, 15548533 Ishimaru, 16891422 Chen, 105 17637 Huang, 16929298 Lopezlimenez, 16805797	Veldhuisen, 10602361 Guo, 10756092 Sutton, 16883570	Sun Slaugenhaupt. 2000 Bargal, 10973263 Falardeau, 11897010 Slaugenhaupt, 12125810 LaPlante, 12459486 LaPlante, 15336987 Soyombo, 16361256 Kiselyov, 16257972 Bargal, 9323557 LaPlante, 16914343	Lindvall, 16137664 Song, 17050035 Di Palma, 12403827 Venkatachalam, 16606612	Di Palma, 12403827 Venkatachalam, 16606612 Deol (1954)

C, COOH terminus; N, amino terminus; DRG, dorsal root gangion; Gd, gadolinium; GPCR, G-protein-coupled receptor; La, lanthanum; LPAR, lyso-phosphatidylinositol receptor; PIP2, polyphosphatidyl inositol (4,5) bisphosphate; PLC, phospholipase C; RuR, ruthenium red; S, transmembrane segment; TKR, tyrosine kinase receptor; TM, transmembrane.

Table 2 TRPC subfamily: canonical (short) TRP channel table

TRP subtype *Unigene*	*C1* *Hs.250687* *Splice variants*	*C2* **602343*	*C3* *Hs.150981* *Splice variants*	*C4 (CCE2)* *Hs.262960* *Splice variants*	*C5 (CCE1)* *Hs.247868*	*C6* *Hs.159003*	*C7* *Hs.283104*
# Amino acids							
Human	H 810	H pseudogene	H 848	H 956	H 973	H 931	H 862
Rat	R 759	R 885	R 736	R 977	R 974	R 930	R 861
Mouse	M 809	M 1172	M 836	M 974	M 975	M 930	M 862
Accession #							
Human	Hs.250687	Hs.0.131910	Q13507	Hs.262960	Hs. 247868	Q9Y210	Q9HCX4
Rat	Rn.88592	NP072160	Q9JMI9	Rn. 10853	Rn. 88592	Q00N78	XP001067646
Mouse	Mm.149633	Mm 0.292904	Q9QZC1	Mm. 10100	Mm. 328378	Q61143	Q9WVC5
Splice variants and other accession #s	P48995 Hs 783 CAA98108 Hs 759 I38361 Hs 810 Q61056 Mm 809 α	Q9R244 Mm 1172 Q9R283 Rn 885	Q5G1L5 Hs 921 α Q13507 Hs 848 Q9QZC1 Mm 836 Q9JMI9 Rn 736 sv	Q9UBN4 Hs 977 α AAL24550 Hs 893 β Q96P05 Hs 982 ε Q96P04 Hs 323 η Q96P03 Hs 804 ζ Q5W0R5 Hs 466 Q5W0R2 Hs 828 Q5W0R4 Hs 836 Q9QUQ5 Mm 974 O35119 Rn 977	Q9UL62 Hs 973 Q9QX29 Mm 975	Q9Y210 Hs 931 Q61143 Mm 930	Q9HCX4 Hs 862 Q9WVC5 Mm 862
OMIM gene map H, R, M	3q22–3q24 8q31 9E4/9 51.0 cM	Pseudogene N.D. 11p15.3-p15.4	4q27 2q25 3 18.2 cM	13q13.1-q13.2 2q26 3 D/3 28.6 cM	Xq23 Xq14 X 63.0 cM	11q21-q22 8q11 9 1.0 cM	5q31.1 13
Genetic model or disease mutation	TRPC1−/− No phenotype	Mouse TRPC2−/−		Mouse TRPC4−/−	Mouse TRPC5−/−	Mouse TRPC6−/− Human focal segmental glomerulosclerosis (FSGS)	TRPC7−/−
Known functions		Pheromone sensing in mouse (vomeronasal organ: VNO)	Vasoregulation (resistance artery myogenic tone) Airway regulation	Impaired agonist-dependent vasorelaxation in TRPC4−/− mouse	Growth cone morphology	Vasoregulation (resistance artery myogenic tone) Airway regulation Cardiomyocyte hypertrophy (cardiomyopathy) Kidney podocyte regulation	
Likely role in	Purported mechano-sensitive channel, stretch-sensitive; contributes to slow excitatory postsynaptic potential activated by metabotropic glutamate receptors in cerebellar Purkinje neurons		CNS	Vasoregulation, microvascular permeability? CNS GI motility 5-HT regulated GABA release from thalamic interneurons	Regulation of growth cones movement and neurite extension Putative component of muscarinic activated current in gastric smooth muscle cells	Receptor-activated role in smooth muscle, potential activation via stretch stimulation of PLCβ	TRPC7 conducts Ca in AT1-induced myocardial apoptosis via a calcineurin-dependent pathway

Continued

Table 2 Continued

TRP subtype *Unigene*	*C1* *Hs.250687* *Splice variants*	*C2* **602343*	*C3* *Hs.150981* *Splice variants*	*C4 (CCE2)* *Hs.262960* *Splice variants*	*C5 (CCE1)* *Hs.247868*	*C6* *Hs.159003*	*C7* *Hs.283104*
N or C terminal domain modules	3x N-AnkR C-TRP box-C-PolyP (EWKFAR . . . PPPFN) [N+C coiled coils; N+C PKC sites; N PKA site; TyrK site-C-CIRB; N-Cav1BD	3x N-AnkR C-TRP box-C-PolyP (EWKFAR . . . PPPFN) C-CIRB N-PKD1 (TRPP1) homology	4x N-AnkR C-TRP box-C-PolyP (EWKFAR . . . PPPFN) C-CIRB	2x N-AnkR C-PDZ binding C-TRP box-C-PolyP (EWKFAR . . . PPPFN) C-CIRB -C-Cam	2x N-AnkR C-PDZ binding C-TRP box-C-PolyP (EWKFAR . . . PPPFN) C-CIRB	3x N-AnkR C-TRP box-C-PolyP (EWKFAR . . . PPPFN) C-CIRB	3x N-AnkRs C-TRP box-C-PolyP (EWKFAR . . . PPPFN) C-CIRB
Reported interacting proteins	TRPC4 TRPC5 (TRPC3 -embryo), Calmodulin, PC2 (TRPP1) IP3Rs, caveolin-1, Gαq/11, PMCA	Calmodulin IP3R3 Enkurin TRPC6	TRPC1 (embryo only) PLCβ, PLCg, P3R, RyR, SERCA, caveolin-1, calmodulin aSNAP, NCX1	TRPC1, TRPC5 PLCβ; NHERF-2, NHERF-1 and ZO-1 via PDZ-BD EBP50(ezrin/moesin/radixin-binding phosphoprotein 50) TRPC4a acts as dominant negative for TRPC4b	TRPC1, TRPC4 Calmodulin, MxA, TRPC3 (embryo), NHERF	TRPC3, TRPC7, calmodulin, FKBP12	TRPC3, TRPC6, calmodulin, FKBP12, MxA, TRPC1
Activation	Muscarinic M1 GPCR,Gq-PLCβ TKR-PLCγ? Voltage gating not described	GPCR Gq-PLCβ, DAG, TKR-PLCγ? Voltage gating not described	Histamine H1 GPCR Gq-PLCb; DAG TKR-PLCγ $[Ca^{2+}]_i$/CaM Homer-1? Voltage gating not described	Muscarinic M1 GPCR Gq-PLCβ TKR-PLCγ? 100 μM La, Gd Calmidazolium Voltage gating not described	Muscarinic M1 GPCR;Gq-PLCβ TKR-PLCγ? 100 μM La, Gd Voltage gating not described	Histamine H1 GPCR Gq-PLCβ,DAG TKR-PLCg?, $[Ca^{2+}]_i$ via CaMKII, AlF_4, 100 μM flufenamate, 100 μM niflumate, 100 μM GTPγS Voltage gating not described	HEK purinergic GPCR Gq-PLCβ, DAG, TKR-PLCγ? $[Ca^{2+}]_i$ 100 μM GTPγS voltage gating not described
Direct agonist	La for heteromer C1+C5	DAG analogs (10–100 μM)	DAG analogs (10–100 μM)	La	N.D.	DAG analogs (10–100 μM)	DAG analogs (10–100 μM)
$\sim P_{Ca}/P_{Na}$	<10 putative Na=Cs=Ca	Cation nonselective, Ca permeant	1.6 cation nonselective	~1	~1 Ca>Na=Cs=K	5	~0.5 Ba>Ca>Cs=Na
Blockers and inhibitors	80 μM 2-APB, 10 μM Gd, La Ca^{2+}-CaM inhibits	2-APB (50 μM)	2-APB, 1–25 μM Gd, La, RuR, BTP2 (300 nM), flufenamate	mM La^{3+} ; 2-APB, Niflumic acid, DIDS increased $[Ca^{2+}]_i$ (Walker)	mM La^{3+}; 2-APB, flufenamate	(IC50) 0.25 mM Cd, 4 μM La, 2 μM Gd, 130 μM amiloride, 4 μM SKF96365 100 μM, 2-APB blocks ~60% of current, Ca_e may block as permeates	100 μM La^{3+} SKF96365 (25 μM), 0.1–0.4 mM Ca_e, Ca_i via calmodulin flufenamate

Single channel conductance (γ) in pS.	10 pS in 1 nM Ca 2.5 pS in 2 $[Ca^{2+}]_i$ Vaca JBC 2002 ~30 pS at –50 mV and ~5 pS at 50 mV (symmetrical 200 mM K+ with 40 mM internal Mg) Maroto et al., NCB, 2005	42pS (symmetrical 140 mM Na) Lucas et al., *Neuron* 2003	Unclear as homomer. Estimates range from ~20 pS (noise analysis) to 65 pS (140 Na/ 140 Na)?	Unclear as homomer- Estimates range from 18 pS to 40 pS 18 pS (b) Walker, AJP) 28 pS (Schaefer) 140Na/140Cs)	~38 pS as homomer ~5 pS as TRPC1/ TRPC5 heteromer	35 (sym 120 Cs, Hofmann, 99); 28 pS isotonic (Mori)	Rectifying open channel *I–V*: 50 pS (0 to +100 mV), 25 pS (−100 to 0 mV) (Shi, 2004)
I–V shape	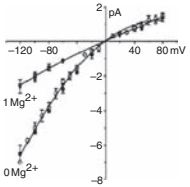 Maroto et al. *Nat Cell Biol* 2005	Near linear 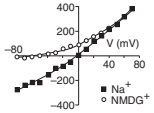 From Lucas et al., *Neuron* 2003	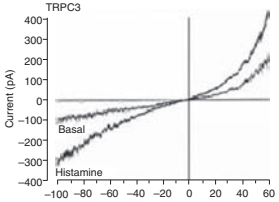 From Dietrich et al. *JBC*	~ Double-rectifying as homomer 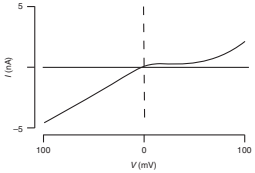 From Schaefer et al. *JBC*	~ Double-rectifying as homomer 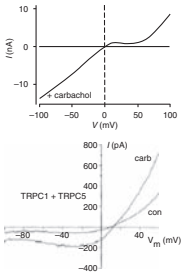	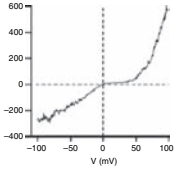 From Jung et al. *AJP*	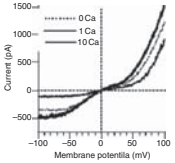 Shi et al., *J Physiol* 2004
Tissue expression	Heart, brain, lung, liver, lesser expression in spleen, kidney, testis; widely expressed in brain, including CA3 region of hippocampus, and cerebellar Purkinje neurons	Vomeronasal organ in nose, (sensory villi)	Brain, pituitary, heart, lung, DRG	Heart, brain, pancreas, placenta, and kidney; inner granule cell layer of the rat olfactory bulb; hippocampus CA1/3, and dentate gyrus.	Brain, testis, kidney, uterus; hippocampal, cerebellar Purkinje neurons, mitral cells of olfactory bulb; hippocampal CA1/3 and dentate gryus cells	Heart, lung, kidney, muscle, intestine, stomach, pancreas, prostate, bone, brain, testis, ovary, and adrenal	Kidney, intestine, pituitary, brain, heart, lung, and eye
Comments	Most unique of TRPC sequences Complexed with TRPC4 or TRPC5, TRPC1 is activated after stimulation of Gq signaling pathways. Low concentrations of La3+ and Gd3+ enhance TRPC1/4 or C1/5 currents Stretch may activate endogenously	No function in human TRPC2b is major variant on mouse cell membrane (VNO) Putative role in erythropoetin-mediated Ca influx in erythroblasts	Generally accepted role for DAG activation	5-HT mediated release of GABA from thalamic interneurons is greatly reduced in TRPC4−/− mice Overall mouse phenotype not reported. Majority of evidence indicates that TRPC4b is receptor-operated Complexed with TRPC1, TRPC4 is activated after stimulation of Gq signaling pathways. When complexed with TRPC1, a current with a unique *I–V* relation is formed	G-protein-linked receptor operated channel as heteromer. I/V of heteromer is similar to NMDA class of ion channels (negative slope region), but unlike NMDA channels is not dependent on Mg voltage-dependent rectification. Not store-operated	Generally accepted role for DAG activation. Probable a1-adrenoceptor-activated channel in portal vein smooth muscle, as well as other vascular smooth muscle. PLC activation, OAG and thapsigargin stimulate delivery of channels to plasma membrane	Generally accepted role for DAG activation

Continued

Table 2 Continued

TRP subtype *Unigene*	*C1* *Hs.250687* *Splice variants*	*C2* **602343*	*C3* *Hs.150981* *Splice variants*	*C4 (CCE2)* *Hs.262960* *Splice variants*	*C5 (CCE1)* *Hs.247868*	*C6* *Hs.159003*	*C7* *Hs.283104*
	expressed homomeric TRPC1 channels in *Xenopus* oocytes				Complexed with TRPC1, TRPC5 is activated after stimulation of Gq signaling pathways. When complexed with TRPC1, a current with a unique *I–V* relation is formed		
REFS (First author, PMID #)	Zitt, 8663995 Wang, 10199829 Strubing, 11301024 Singh, 11983166 Wang, 10199829 Zhu, 8646775 Yuan, 14505576 Kim, 14614461 Huang, 16906149 Maroto, 15665854 Strubing, 12857742	Yildirim, 12601176 Liman, 10318963 Vannier, 10051594 Hofmann, 10998353 Sutton, 15385169 Brann, 12472899 Stamboulian, 16153633 Chu, 14699131 Lucas, 14642279 Stowers, 11823606 Tong, 15329338	Zhu, 8646775 Ohki, 10984475 Mori, 9512398 Dietrich, 15971081 Zagranichnaya, 15972814 Lussier, 15757897 Sinkins, 15199065 Bandyopadhyay, 15623527 Rosker, 14736881 Zitt, 9298988 Kim, 16887806 Dietrich, 16842858 Kiselyov, 9853757 Li, 10677043 He, 15647288 Trebak, 11943785 Lievremont, 15933213 Riccio, 12531519 Inada, 17027915 Dietrich, 16055711 Balzer, 10533589 Li, 15758952	Schaefer, 10837492 Tang, 10980202 Strubing, 11301024 Philipp, 9687496 Tang, 11290752 McKay, 11042129 Freichel, 11175743 Philipp, 10816590 Lee, 16404161 Strubing, 12857742 Lussier, 15757897	Schaefer, 10837492 Philipp, 9687496 Strubing, 11301024 Okada, 9553080 Greka, 12858178 Bezzerides, 15258588 Lee, 12631560 De March, 16635549 Chung, 16580647 Yoshida, 16998480 Ordaz, 15987684 Jung, 12456670 Xu, 15806115 Lussier, 15757897 Strubing, 12857742 Tang *J Biol Chem.* 2000	Hofmann, 9930701 Soboloff, 16204251 Boulay, 9368034 Dietrich, 15971081 Boulay, 12372123 Zagranichnaya, 15972814 Dietrich, 16842858 Inoue, 11179201 Riccio, 12531519 Inada, 17027915 Mizuno, 9889314 Reiser, 15924139 Dietrich	Riccio, 11805119 Onohara, 17082763 Okada, 10488066 Dietrich, 15971081 Boulay, 12372123 Sinkins, 15199065 Lussier, 15757897 Zagranichnaya, 15972814 Shi, 15579537 Riccio, 12531519 Facemire, 14678949 Satoh, 16838106 Lievremont, 16123040

Table 3 TRPM subfamily: melastatin-related (long) TRP channel table

TRP subtype *Unigene*	*M1 (MLSN)* *Hs.43265* *Splice variants*	*M2 (hTRPC7, LTRPC2)* *Mm. 133517* *Splice variants*	*M3 (MLSN2, LTRPC3)* *Hs.288911 96* *Splice variants (m)*	*M4 (LTRPC4, MLS2s, CAN [4b])* *Hs.31608* *Splice variants*	*M5 (Mtr 1, LTRPC5)* *Hs.272287* *Splice variants*	*M6 (CHAK2)* *Hs.272225* *Splice variants*	*M7 (CHAK1, TRPPLIK, LTRPC7)* *Hs.33819*	*M8 (Trp-p8, CMR1)* *Hs.366053*
# Amino acids	H 1533 R 1628 M 1506	H 1503 R 1508 M 1506	H 1325–1732 R N.D. M 1325–1721	1214 M4b/1040M4a R 1208 M 1213	H 1165 R M 1158	H 2022 R 2020 M 2028	H 1864 R 1862 M 1862	H 1104 R 1104 M 1104
Accession # Human Rat Mouse	075660 Q2WEA5 Q2TV84	094759 Q2PH54 Q91YD4	Q9HCF6 (long) N.D. Q5F4S9 (long)	Q8TD43 Q9ESQ5 Q7TN37	Q52LU2 Q9JJH7	Q9BX84 XB219747 Q8CIR4	Hs.512894 Rn86991 Mm. 244705	Q722W7 Q8R455 Q84RD5
Splice variants and other accession #s	Q7Z4N2 Hs 1533 BAC80204 Hs 1527 BAC80202 Hs 230 070334 Mm 542 Q5U4B3 Mm 426 Q640P5 Mm 443	094759 Hs 1503 Q5KTC2ii Hs 1503 Q5KTC1 Hs 1289 NP_001001 188 Hs 846 Q91YD4 Mm 1507 Q5KTCO Mm 1506 Q8BNJ4 Mm 328	Q9HCF6 Hs 1732 NP001007472Hs 1707 AA049153Hs 1554 3a AA049154Hs 1566 3b AA049155Hs 1566 3c AA049156Hs 1544 3d NP_996827iiHs544 3d AA049157Hs 1556 3e AA049158Hs 1579 3f BAC80203 Hs 1533 BAC80204 Hs 1527 BAC55106 Hs1526 CAD43605 Hs1325 BAB13442 Hs1017 BAC80202 Hs230 Q5F4T0 Mm 1721 α1 Q5F4S8 Mm 1711 α2 Q5F4S9 Mm 1721 α3 Q5F4S7 Mm 1709 α4 Q5F4S6 Mm 1699 α5	Q96L84 Hs 1040 a Q7Z5D9 Hs 1069 c Q7TN37 Mm 1213 NP_780339 Mm 945 AAH58632 Mm 872 AAH46472 Mm 793 AAH49993 Mm 510 BAC81770 Mm 488 BAC81769 Mm 422 b	Q9NZQ8 Hs 1165iv Q9JJH7 Mm 1158 Q99NF9 Mm 1148 Q9EPM3 Mm 1116 Q7TPL4 Mm 1000	AAR03490 Hs 1943 a AAR03488 Hs 2017 b AAR03489 Hs 2017 c Q8CIR4 Mm 2028	Q96QT4 Hs 1865 Q923J1 Mm 1863	Q7Z2W7 Hs 1104v NP_076985vi Hs 1104 Q8R4D5 Mm 1104
OMIM gene map H, R, M	15q13-q14 Rat 1q22 Mouse 7c	21q22.3 20p12 10C1	9q21.11 N.D. 19 B	19q13.33 1q22 7	11p15.5 7ES	9q21.13 1q43 19B	15q21 3 2F2	2q37.1 9q35 2 H1
Genetic model or disease mutation		Bipolar disorder BP-I, II, nonsyndromic hereditary deafness	Candidate gene for ALS with frontotemporal dementia (OMIM 105550), early-onset pulverulent cataract (OMIM 605749), hemophagocytic lymphohistiocytosis (HLH, OMIM 603552), infantile nephronophthisis (OMIM 602088)		Mouse TRPM5−/− Candidate gene for rhabdomyosarcomas, Wilms tumor, Beckwith-Widemann	Human hypomagnesemia with secondary hypocalcemia (HGH)	TRPM7−/− A variant form of TRPM7 has been detected in patients with Guamanian amyotrophic lateral sclerosis (ALS-G) and Parkinsonism dementia (PD-G)	

Continued

Table 3 Continued

TRP subtype Unigene	*M1 (MLSN) Hs.43265 Splice variants*	*M2 (hTRPC7, LTRPC2) Mm. 133517 Splice variants*	*M3 (MLSN2, LTRPC3) Hs.288911 96 Splice variants (m)*	*M4 (LTRPC4, MLS2s, CAN [4b]) Hs.31608 Splice variants*	*M5 (Mtr 1, LTRPC5) Hs.272287 Splice variants*	*M6 (CHAK2) Hs.272225 Splice variants*	*M7 (CHAK1, TRPPLIK, LTRPC7) Hs.33819*	*M8 (Trp-p8, CMR1) Hs.366053*
Known functions		Oxidant stress sensor in immune cells, glia in brain, T-dependent, Ca entry pancreatic beta cells		Endogenous calcium-activated cation (CAN) channels, negative feedback regulator of Ca^{2+} entry	Taste (sweet, bitter, umami) via T1R, T2R and PLCb	Renal and gastrointestinal tract Mg^{2+} uptake		Cold sensing
Likely role in	Melanoma invasiveness	Oxidant stress response in immune and other cells (microglia in brain, respiratory burst in neutrophils?)	Probable role in osmosensation and renal Ca^{2+} homeostasis, mechanosensor, cation homeostasis in cerebrospinal fluid	Ca oscillations in activated T cells Myogenic cerebral artery constriction Bayliss effect in smooth muscle Spreading depression hypoxic depolarisation, stroke, heart hypertrophy and arrhythmia	Perhaps temperature	Magnesium absorption in kidney	Cellular Mg homeostasis proposed. Potential other trace metals, such as Zn, Mn. Oxygen-glucose deprivation-induced calcium influx and neuronal death	Pain pathways (present in Ad fibers), presynaptic potentiation in DRG neurons Present in various adenocarcinomas
N or C terminal domain modules	N-700 a.a. TRPM domain,	N-700 a.a. TRPM domain,	4 CaM binding domains N-700 a.a. TRPM domain,	N-700 a.a. TRPM domain,	C-TRP (FWKFQR) box	N-700 a.a. TRPM domain	N-700 a.a. TRPM domain,	N-700 a.a. TRPM domain,
	W/FIX_3-(F/I/I)XK(R/K)EC(V/I/S)X_{12-24}CXCG	W/FIX_3-(F/I/I)XK(R/K)EC(V/I/S)X_{12-24}CXCG	W/FIX_3-(F/I/I)XK(R/K)EC(V/I/S)X_{12-24}CXCG	W/FIX_3-(F/I/I)XK(R/K)EC(V/I/S)X_{12-24}CXCG	[VIYFL]**WK**[FAY][QN]**R**;	W/FIX_3-(F/I/I)XK(R/K)EC(V/I/S)X_{12-24}CXCG	W/FIX_3-(F/I/I)XK(R/K)EC(V/I/S)X_{12-24}CXCG	W/FIX_3-(F/I/I)XK(R/K)EC(V/I/S)X_{12-24}CXCGC-TRP (VWKFQR) box
	C-TRP (VWKFQR) box	[VIYFL]**WK**[FAY][QN]**R**; C-NUDT9 ADP-ribose hydrolase (NUDt9-H)	C-TRP (VWKFQR) box [VIYFL]**WK**[FAY][QN]**R**;	C-TRP (YWKAQR) box [VIYFL]**WK**[FAY][QN]**R**;		C-TRP (LWKYNR) box [VIYFL]**WK**[FAY][QN]**R**;	C-TRP (VWKYQR) box [VIYFL]**WK**[FAY][QN]**R**;	[VIYFL]**WK**[FAY][QN]**R**;
	C-PolyP	C-TRP (IWKFQR) box	C-PolyP	C-PolyP$_i$	C-PolyP	C-coiled coil	C-PolyP	C-PolyP
	C-coiled coil	C-PolyP; C-coiled coil	C-coiled coil	C-coiled coil	C-coiled coil	C-coiled coil	C-coiled coil	C-coiled coil
Reported interacting proteins	Short transcript of M1 (MLSN-S).	N-CaM C-ADP-ribose hydrolase Sir2(sirtuin), a NAD^+-dependent deacetylases, binds to TRPM2	N.D.	Calmodulin Probably SUR1	N.D.	C-Atypical protein kinase TRPM7	PLCβ 1,2,3, PLCγ1, snapin TRPM6, myosin IIA heavy chain	

Activation	N.D. Microphthalmia-associated transcription factor (MITF) induces expression Ca^{2+} entry channel, whose membrane insertion is regulated by a short isoform, (MLSN-S, 500 AA) (disputed).	~100 μM ADPR, NAD, NAADP, 700 μM cADPR Oxidative stress, PARP-1, heat potentiates Cysteines C996, C1008 are essential for pore function	Constitutive, $V_{0.5} = -61$ mV Enhanced by hypoosmolarity Activated by sphingolipids (10 mM D-erythrosphingosine)	M4b:K_{act}: 300–500 nM $[Ca^{2+}]_i$; > 1 μM (Nilius)5μM PIP2; $V_{0.5} = -20$ to +60mV Decavanadate; BTP2 or YM-58483 Ca^{2+} affinity depends on range of voltage activation; depends on PIP_2, PKC phosphorylation, decreased Ca^{2+} affinity during desensitization; U73122 prevents TRPM4 desensitization	GPCR (T1R, T2R) Ggus-PLCb2 $[Ca^{2+}]_i$ 20–30 μM PIP2 Heat activation of TRPM5 results from marked leftward shift of the voltage dependence of activation (−7.0 mV °C)	Low pH 2-APB (200 μM)	PIP_2 Tyrosine phosphorylation, Activation upon oxidative stress and cAmp proposed, PKA No V dependence Low pH increases inward current	T<25°C (80 μM menthol, icilin potentiate) Cold-induced activation via leftward shift of voltage dependence of activation (7.3 mV °C^{-1})
Direct agonist (bind channel protein to activate)	N.D.		D-erythrosphingosine, steroids					Menthol (40 μM) Icilin (280 nM) Eucalyptol (5 mM) FrescolatMAG (5 μM)
~P_{Ca}/P_{Na}	N.D.	~1 (estimated) reported ~6 at 35C	<0.1 TRPM3a1-no divalents, perm to Na, K, Cs, NMDG	M4b: <0.05 (monovalent selective) Na>K>Cs>Li	<0.05 (Monovalent nonselective)	<1 Zn>Ba>Mg>Ca>Mn>Sr>Cd>Ni	0.3 monovalent s outward inward – Zn~Ni>Ba>Co>Mg~Mn~Sr~Cd~Ca	1–3 Cs>K>Na
Blockers and inhibitors	N.D.	No pore block by lanthanides and heavy metal ions 2 μM N-ACA econazole, miconazole, clotrimazole, 2-APB, flufenamic acid	Mg_i for M3a1, 3a2 (9 mM comp. block) 100 μM Gd, La extracell Na>K decreases outward currents in M3a2 2-APB (100 μM)	M4a: 80 μM La, Gd M4b: intracellular free ATP^{4-}, 1.7 μM Polyamines, Spermine 30–60 μM) Pore mutation D984A results in a non-functional channel with a dominant negative phenotype	Sensitive to pH below 7.0 with complete block around 6	Ruthenium Red Mgi (0.5 mM)	La (2 mM), Mgi (0.6 mM) PIP_2 hydrolysis by GPCR (M1), LPAR/EDG-7R)-Gq-PLCβ and TK (EGF) -PLCγ hydrolyzes PIP_2 to turn off channel	BCTC (1 μM) Capsazepine (18 μM) 2-APB (10 μM)
Single channel conductance (γ) in pS.	N.D.	60 pS (140 Na/140 Na)	$TRPM3_{1325}$ 83 (140 Na/140 Na) 133 pS (140 Cs, −60 mV); 83 pS (−60 mV, 140 Na), 73 pS physiological cond.	M4b:25 pS (isotonic monovalents)	23 pS (isotonic monovalents)	84 pS	105 pS (+ voltages)	83 pS (+ voltages)
I–V shape	N.D.	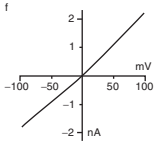 Instantaneous *I/V* is linear From Perraud et al. *Nature*	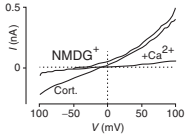 From Grimm et al. *JBC*	M_4b: 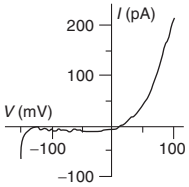 Instantaneous *I/V* is linear Nilius et al., (2003)	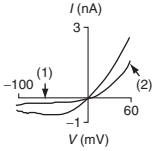 Hofmann	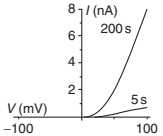 From Voets et al. *JBC*	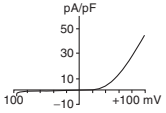 Runnels et al. *NCB*	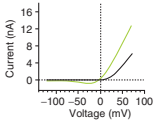 From Mckemy et al. *Nature*

Continued

Table 3 Continued

TRP subtype Unigene	*M1 (MLSN) Hs.43265 Splice variants*	*M2 (hTRPC7, LTRPC2) Mm. 133517 Splice variants*	*M3 (MLSN2, LTRPC3) Hs.288911 96 Splice variants (m)*	*M4 (LTRPC4, MLS2s, CAN [4b]) Hs.31608 Splice variants*	*M5 (Mtr 1, LTRPC5) Hs.272287 Splice variants*	*M6 (CHAK2) Hs.272225 Splice variants*	*M7 (CHAK1, TRPPLIK, LTRPC7) Hs.33819*	*M8 (Trp-p8, CMR1) Hs.366053*
Tissue expression	Skin; melanocytes and retinal pigmented epithelium	Brain; microglia, cerebellum, cortex, medulla oblongata, hippocampus, striatum, pancreatic B-cells, immune cells, PMNs, thymocytes, vascular endothelium	Kidney, pulmonary endothelium, cultured trigeminal neurons, liver, pancreas, ovary, testis, spinal cord Brain: dentate gyrus, lateral septal nuclei, indusium griseum, tenia tecti, epithelial cells of choroid plexus, human iris, retinal pigmented cells	Stomach, aorta, endothelium, esophagus, pancreas, kidney; glom., PCT, MTAL, DCT, CCD brain: widely expressed smooth muscle cerebral arteries	Tongue, taste tissue stomach intestines uterus testis	Small intestine, kidney	Ubiquitous	DRG, trigem ganglia, testis, prostate, bladder
Comments	Decrease in M1 RNA with melanoma progression. loss of melastatin expression correlates with the thickness of the melanomas	Increased $[Ca^{2+}]_i$ potentiates (EC_{50} 340 nM) Regulated by $[Ca^{2+}]_i$ probably via CaM, sensitizes TRPM2 to activation by ADPR, PARP-1 inhibitors (SB750 139>PJ34> DPQ) decrease activation of TRPM2 by oxidative stress. Catalase, dimethyl thiourea, mannitol inhibit TRPM2 activation by oxidative stress.	Independent of $[Ca^{2+}]_i$ insensitive to 1 μM RuR	Probably ubiquitous Voltage-modulated calcium-activated Decavanadate (2 μM) inhibits V dependence; Human; long variant TRPM4b 1214 AA, AF497623; short variant TRPM4a 1040 AA, first 174 AA lacking, AY046396; short splice variant TRPMc, 537 AA N-terminal lacking, AX443225 mouse: two short splice variants from brain AB112658, AB112657	Voltage-modulated Calcium-activated	Contains a protein kinase on its C terminus as part of the protein (chanzyme) Currently difficult to separate from TRPM7 current measurements	Contains a protein kinase on its C terminus as part of the protein (chanzyme). The kinase activity is not required for channel conduction. Overexpression results in cell death. Deletion also results in cell death in DT40 cells. Kinase domain phosphorylates Annexin 1	Correlated with trkA NGF receptors trp-p8 mRNA also expressed in primary tumors of prostate, breast, colon, lung, and skin. Loss of prostate trp-p8 mRNA e associated with shorter survival. trp-p8 may be androgen-regulated

Refs First author, PMID								
	Hunter, 9806836 Duncan, 9537257 Duncan, 11208852 Miller, 14744763 Zhiqi, 15577322 Xu, 11535825	Kudoh,9179495 Nagamine, 9806837 Sano, 11509734 Perraud,1 1385575 Uemura, 15708008 Hill, 16260005 Hara, 11804595 Tong, 16461353 Grubisha, 16565078 Heiner, 12564954 Togashi, 16601673 Inamura, 12571754 Kolisek, 15808509 Beck, 16585058 Wehage, 11960981 Smith,12562896 Kraft, 14512294 Fonfria, 15302683 Chung, 15722340 McHugh, 12529379 Kuhn, 15952035 Yang, 16294211 Herson, 9409472 Mei, 16822940 Kraft, 16604090 Hill, 15549272 Togashi Hill, 15275834 Gasser, 16316998 Zhang, 12594222 Kraft, 15895246 Scharenberg, 16001276 Perraud, 12765697 Yao, 16254217 McNulty, 16025303 Hurne, 12063251	Nagase, 10997877 Lee, 12672827 Grimm, 12672799 Oberwinkler, 15824111 Yap, 12836676 Grimm, 15550678 Nilius, press Xu, 15806115 Fantozzi, 12909593 Nealen, 12634279 Bollimuntha, 16765919 Wistow, 12107412 Wistow, 12107410	Xu, 11535825 Launay, 15550671 Nilius, 12799367 Nilius, 15845551 Gibbs, 15057822 Nilius, 15590641 Simard, 16550187 Chen, 13679426 Nilius, 16680483 Ullrich, 15670874 Nilius, Mahieu EMBO 06 Nilius, Talavera J Physiol 05 Talavera, 16355226 Hofmann, 12842017 Launay, PMID: Takezawa, 16407466 Nilius, 15331675 Zhang, 16186107 Nilius, 1475847820: Chabardes-Garonne, 14595018 Fonfria, 16777713 Earley, 15472118 Cheng, 16806463 Schiller, 15277598 Guinamard, 15121803 Guinamard, 15030426 28: Nilius B, Owsianik G, Voets in press Iwata, 12796481 Murakami, 12893253	Prawitt, 10607831 Enklaa, 10903843 Hofmann, 12842017 Ullrich, 15670874 Talavera, 16355226 Liu, 14657398 Liu, 15731110 Zhang, 12581520 Perez, 12368808 Nilius, 15331675 Nilius, 15845551 Prawitt, 14634208 Damak, 16436689	Schlingmann, 12032568 Walder, 12032570 Voets, 14576148 Chubanov, 14976260 Schmitz, 16150690 Li, 16636202	Runnels, 11161216 Nadler, 11385574 Runnels, 11941371 Kerschbaum, 12668438 Schmitz, 12887921 Aarts, 14697204 Monteilh-Zolle, 1250805316 Chubanov, 14976260 Dorovkov, 15485879 Jiang, 16009728 Hermosura, 16051700 Kozak, 16260839 Clark, 16407977 Su, 16436382 Fonfria, 16777713 Oancea, 16357306 Li, 16636202 Krapivinsky, 17088214	Tsavaler, 11325849 McKemy, 11882888 Peier, 11893340 Voets, 15306801 Behrendt, 14757700 Andersson, 15190109 Hu, 15194687 Stein, 15311065

Continued

Table 3 Continued

TRP subtype *Unigene*	*M1 (MLSN)* *Hs.43265* *Splice variants*	*M2 (hTRPC7, LTRPC2)* *Mm. 133517* *Splice variants*	*M3 (MLSN2, LTRPC3)* *Hs.288911 96* *Splice variants (m)*	*M4 (LTRPC4, MLS2s, CAN [4b])* *Hs.31608* *Splice variants*	*M5 (Mtr 1, LTRPC5)* *Hs.272287* *Splice variants*	*M6 (CHAK2)* *Hs.272225* *Splice variants*	*M7 (CHAK1, TRPPLIK, LTRPC7)* *Hs.33819*	*M8 (Trp-p8, CMR1)* *Hs.366053*
		Harteneck, 15843919 Perraud, 15220002 Aarts, 14697204 Reid, 16075243 Aarts, 16044308 Chubanov, 16075242 Kraft, 15895246 Heiner, 12765698 Wilcox, 11163249 Xu, 16252251 Yoon, 11690598 Zhang, 16306129 Ishii, 16546253 Nilius (in press)						

Table 4 TRPV subfamily: vanilloid TRP channel table

TRP subtype *Unigene*	*V1 (VR1)* *Hs.283010* *Splice variants*	*V2 (VRL1, OTRPC2)* *Hs.279746*	*V3 (VRL2)* *Hs.375034* *Splice variants*	*V4 (OTRPC4, VR-OAC)* *Hs.287776* *Splice variants*	*V5 (ECaC1, CaT2)* *Hs.283369*	*V6 (ECaC2, CaT1)* *Hs.302740*
# Amino acids						
Human	839	764	790	871	729	725
Rat	838	761	781	871	723	727
Mouse	839	756	791	871	723	727
Accession #						
Human	Q8NER1	Q9Y5S1	Q8NET8	Q9HBA0	Q9NQA5	Q9H1D0
Rat	Q35433	Q9WUD2	AY325813	Q9ERZ8	Q9JIP0	Q9R186
Mouse	Q704Y3	Q9WTR1	Q8K424	Q9EPK8	P69744	Q91WD2
Splice variants and other accession #s	Q8NER1 Hs 839 Q704Y3 Mm 839 α AAS15575 Mm 829 β	Q9Y5S1 Hs 764 Q9WTR1 Mm 756	Q8NET8 Hs 790 a AAM80559 Hs 791 a AAM80558 Hs 765 b Q8K424 Mm 791	Q9HBA0 Hs 871 α NP_671737 Hs 811 β Q9EPK8 Mm 871 NP_071300i Mm 871	Q9NQA5 Hs 729 P69744 Mm 723	Q9H1D0 Hs 725 Q91WD2 Mm 727
OMIM gene map						
Human	17p13.3	17p11.2	17p13.3	12q24.1	7q35	7q33–34
Rat	10q24	10q22	10q24	12q16	4q23	4q22
Mouse	11 B3	11 B2	11 B4	5 F	6 B2	6 B2
Genetic model or disease mutation	Mouse TRPV1−/− Defects in nociception	Mouse TRPV2−/−	Mouse TRPV3−/− Mild temp-related pain deficit	Mouse TRPV4−/− ADH volume regulation affected, temperature preference	Mouse TRPV5−/− causes hypocalciuric hypercalcemia, diabetes?	Mouse TRPV6−/− Decreases intestinal Ca reabsorption; alopecia and dermatitis
Known functions	Basis of 'hot' pepper sensation Heat evoked nociception Inflammatory thermal hyperalgesia, Bladder distension sensing Chronic cough, bronchial sensitization Upregulated in: inflammatory bowel disease Vulvodynia Osteoarthritis	Thermal pain (medium diameter neurons); trigeminal neuron, pterygopalatine and submandibular ganglia, myenteric plexus & nodose ganglion neurons, spinal cord lamina I, II, hypothalamic paraventricular, suprachiasmatic, and supraoptic nuclei, retina Osmo-sensing in aortic myocytes	Warm sensing Pain-sensing	CNS osmotic sensing (circumventricular) Pressure sensing (DRG) Nociception in primary afferents sensitized by PGE2 Inflammatory thermal hyperalgesia	Ca uptake in kidney	Ca uptake in kidney and intestine
Likely role in	Other pain pathways	Skeletal and cardiac muscle degeneration Other pain pathways (visceral sensory)	Warm temp sensing in 308 ketatinocytes Pain	Temperature sensing in skin, 308 ketatinocytes	Ca^{2+} uptake in kidney (Vit D regulated), potential role in beta cell of pancreas	Ca^{2+} uptake in intestine (Vit D regulated)
N or C terminal domain modules	3x N-AnkR C-TRP (IWKLQR) box C-Putative PIP_2 binding site	3x N-AnkR C-TRP (IWKLQK) box	3x N-AnkR C-TRP (IWRLQR) box	4x N-AnkR C-TRP (IWKLQW) box	5x N-AnkR C-TRP (LWRAQV) box N-PDZ	4x N-AnkR C-TRP (LWRAQI) box N-PDZ
Reported interacting proteins	TRPV3, TRPV2 Calmodulin, PI3K Snapin, synaptotagmin IX b-tubulin	TRPV1 RGA protein	TRPV1	Calmodulin (C term) Src family kinases (potentiation) Aquaporin 5 Pacsin 3	TRPV6 S 100A 10/annexin 2, 80K-H,NHERF4, BSPRY, Rab 11a	TRPV5 S 100A 10/annexin 2

Continued

Table 4 Continued

TRP subtype *Unigene*	*V1 (VR1)* *Hs.283010* *Splice variants*	*V2 (VRL1, OTRPC2)* *Hs.279746*	*V3 (VRL2)* *Hs.375034* *Splice variants*	*V4 (OTRPC4, VR-OAC)* *Hs.287776* *Splice variants*	*V5 (ECaC1, CaT2)* *Hs.283369*	*V6 (ECaC2, CaT1)* *Hs.302740*
Activation	$T > 43°C$ Capsaicin, (0.7 μM) Resiniferatoxin; 40 nM anandamide, ACEA satiety factor oleoylethanolamide H^+, 2-APB ~ 200μM Heat-induced activation of TRPV1 is the result of a marked leftward shift of the voltage dependence of activation (9.1 mV °C^{-1})	$T > 53°C$ 2-APB ~ 100mM Neuropeptide head activator Diphenylboronic anhydride	$T > 30°C$ 2-APB 30μM, DPA (50–100 μM) Activation by two identical consecutive heat stimuli shifts $V_{0.5}$ to the less positive potential without affecting slope factor (31.7 ± 1.3 v. 36.6 ± 1.9 mV, respectively)	Decreased Osmolarity (260–200 mOsm) $T > 24°C$ Phorbol esters	Constituitive $[Ca^{2+}]_i$ inhibits	Constitutive $[Ca^{2+}]_i$ inhibits
Q_{10}	~20 at negative voltages, Q_{10} is much smaller at positive voltages	N.D.	25	10–20	N.D.	N.D.
Direct agonist (bind channel protein to activate)	0.7 μM capsaicin 40 nM resiniferatoxin Anandamide (5 μM) H^+ 15-(S) HETE (9 μM) Leukotriene B4 (10 μM) Olvanil (8 nM) Arvanil (0.27 nM) Piperine (38 μM) 2-APB (200 μM) allicin (100 μM) vanillotoxin (5–10 μM)	Capsaicin insensitive	Capsaicin insensitive Carvacrol, thymol, eugenol, vanillin, ethyl vanillin (~1 mM), menthol, cinnemaldehyde (mM)	Capsaicin insensitive Endocannabinoids (anandamide), arachidonic acid, 5′6′-epoxyeicosatrienoic acids (150 nM), 4a-phorbol 12, 13 didecanoate (10 nM), PMA (12 nM) bisandrographalide (1 μM)		2-Aminoethyl diphenylborate
~P_{Ca}/P_{Na}	10 (5 for P_{Mg}/P_{Na}) Ca>Mg>Na=K=Cs	3 Ca>Mg>Na=K=Cs	~10	6 Ca>Mg>K>Cs~Rb>Na>Li	>100 (Ca>Mn>Ba~Sr) Na>Li>K>Cs>NMDG	>100 (Ca>Ba~Sr>Mn) Na~Li>K>Cs
Blockers and inhibitors	Capsazepine (03 μM) Desensitization with camphor 1 μM RuR acylpolyamines (0.3 μM)	1 μM RuR, La^{3+}, SKF96365	100 nM RuR La^{3+}	1 μM RuR, 100 μM Gd^{3+}, 0.5 mM La^{3+}	120 nM RuR; 1 μM econazole μM Cd>>Gd>La, Mg Cu, Pb Intracellular Mg block is voltage-sensitive	9 μM RuR μM Cd>>Gd>La Mg
Single channel conductance (γ) in pS.	~ 80 pS (+V, 140 Na/140 Na)	N.D.	170 pS, conductance is dependent on temperature with low Q_{10}	90 pS (+60 mV, 140Na/140K) 30–60 pS (−60 mV)	75 pS (Ca-free)	42–58 pS (Ca-free)
I–V shape	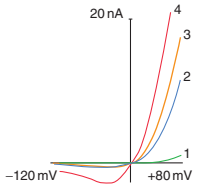 Julius	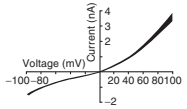 From Caterina et al. *Nature*	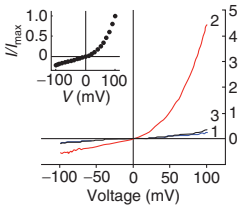 Xu et al.	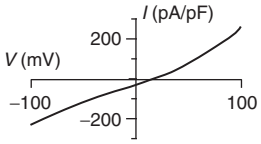 From Watanable et al. *JBC*	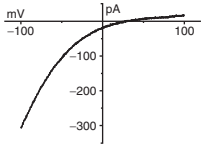 Courtesy Kirichok	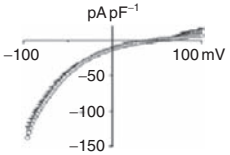 Yue et al. *Nature* 2001

Tissue expression	DRG, trigeminal ganglion, kidney, pancreas, placenta Testis, uterus, spleen, stomach, small intestine, lung, liver —	Brain, DRG, trigeminal ganglia; spinal cord, lung, spleen, and intestine	Brain, spinal cord, skin, testis, stomach, tongue, trachea, small intestine, placenta, hair follicles, adipose, pituitary, and placenta	Brain, trigeminal ganglia, cochlea, lung, spleen, kidney, testis, fat, heart, prostate, pancreas, placenta, and salivary gland	Kidney, prostate, testis placenta, pancreas brain	Intestines, stomach, kidney, lung, spleen, liver, brain, placenta, prostate, pancreas, adrenal glands, skin
Comments	PKC enhances activation PKA reg desensitization Acid potentiates Heat sensitizes Not store-activated Activation of PLC releases TRPV1 from PIP_2 inhibition Capsaicin-induced neuronal death	Difficult expression. Most data from oocytes. IGF-1 appears to translocate TRPV2 to membranes in mouse Heat sensitizes TRPV2 on repetitive stimulation	Current displays marked sensitization with repeated temperature ramps and 2APB activation. Marked hysteresis of temperature reponse (gradual increase with T, sharp decline with cooling)	May play a role in thermosensation (moderate heat), bone regulation, blood volume regulation, hearing	Most Ca^{2+}-selective of TRPs-conserved asp in pore. Voltage-dependent $[Mg^{2+}]_i$-block Upregulated in various adenocarcinomas	Most Ca^{2+}-selective of TRPs-conserved Asp in pore. Voltage-dependent $[Mg^{2+}]_i$-block Upregulated in various adenocarcinomas
Refs	Hayes, 11050376l Caterina, 9349813 Correll, 1548901l Liapi, 16115206 Rutter, 16237318 Smith, 12077606 Numazaki, 12808128 Stein, 17074976 Voets, 15306801 Smart, 11301059 Zygmunt, 1044037 Smart, 10694225 Hwang, 10823958 Jerman, 10864900 McNamara, 15685214 Hu, 15194687 McIntyre, 11226139 Macpherson, 15916949 Siemens, 17093448 Bhave, 14523239 Trevisani, 11992116 Wahl, 11125018 Valenzano, 1272133 Kitaguchi, 16300403 Prescott, 12764195 Birder, 11606761 Caterina, 10764638 Tominaga, 9768840 Geppetti, 15051629 Yiangou, 11343743 Tympanidis, 14987622 Cho, BJ abstac Fernihough, 16039054	Caterina, 10201375 Kanzaki, 10559903 Liapi, 16115206 Rutter, 16237318 Hu, 15194687 Boels, 11707512 Chung, 15722340 Muraki, 14512441 Iwata, 12796481	Xu, 12077604 Peier, 12016205 Smith, 12077606 Chung, 15722340 Chung, 15175387 Moqrich, 15746429 Xu, 16617338 Macpherson, 16829128 Chung, 12783886 Asakawa, 16858425	Liedtke, 0011081638 Strotmann, 12724311 Liu, 16571723 Sidhaye, 16537379 Cuajungco, 16627472 Strotmann, 11025659 Voets, 12093812 Watanabe, 12354759 Watanabe, 11827975 Xu, 12970074 Watanabe, 12879072 Suzuki, 12692122 Smith, 16899456 Güler, 12151520 Delany, 11160995 Alessandri-Haber, 12895423 Todaka, 15187078 Mizogushi Tabuchi, 15925108	Müller, 0010945469 Peng, 10875938 Suzuki, 0010913341 Weber, 11741335 Hoenderop, 12574114 Lambers, 15123711 van de Graaf, 12660155 Gkika, 17005931 van de Graaf, 16565876 van de Graaf, 16380433 van de Graaf, 16354700 Nilius, 10970426 Vennekens, 10660551 Vennekens, 11417220 Yeh, 14525991 Yeh, 16121193 Nilius, 11588099 van der Eerden, 16291808 Nilius, 11352507 Hoenderop, 14679186 van Abel, 16164647 Janssen, 12019295	Peng, 11097838 Peng, 10428857 Hirnet, 12765696 Hoenderop, 12574114 Niemeyer, 11248124 Van de Graaf, 12660155 van de Graaf, 16565876 van de Graaf, 16354700 Voets, 11687570 Yue, 11287959 Voets, 12601087 Hoenderop, 11744752 Suzuki, 001091334 Wissenbach, 11278579 Zhuang, 12480925 Peng, 11401523 Peng, 11549322 Weber, 11741335 Nilius, 12077127 Hoenderop, 12574114 Voets, 14736889 Lambers, 15123711 Gkika, 17005931 Nijenhuis, 14569082 van Cromphaut, 11687634 Barley, 11208552 Peng,10875938 Bianco, 17129178

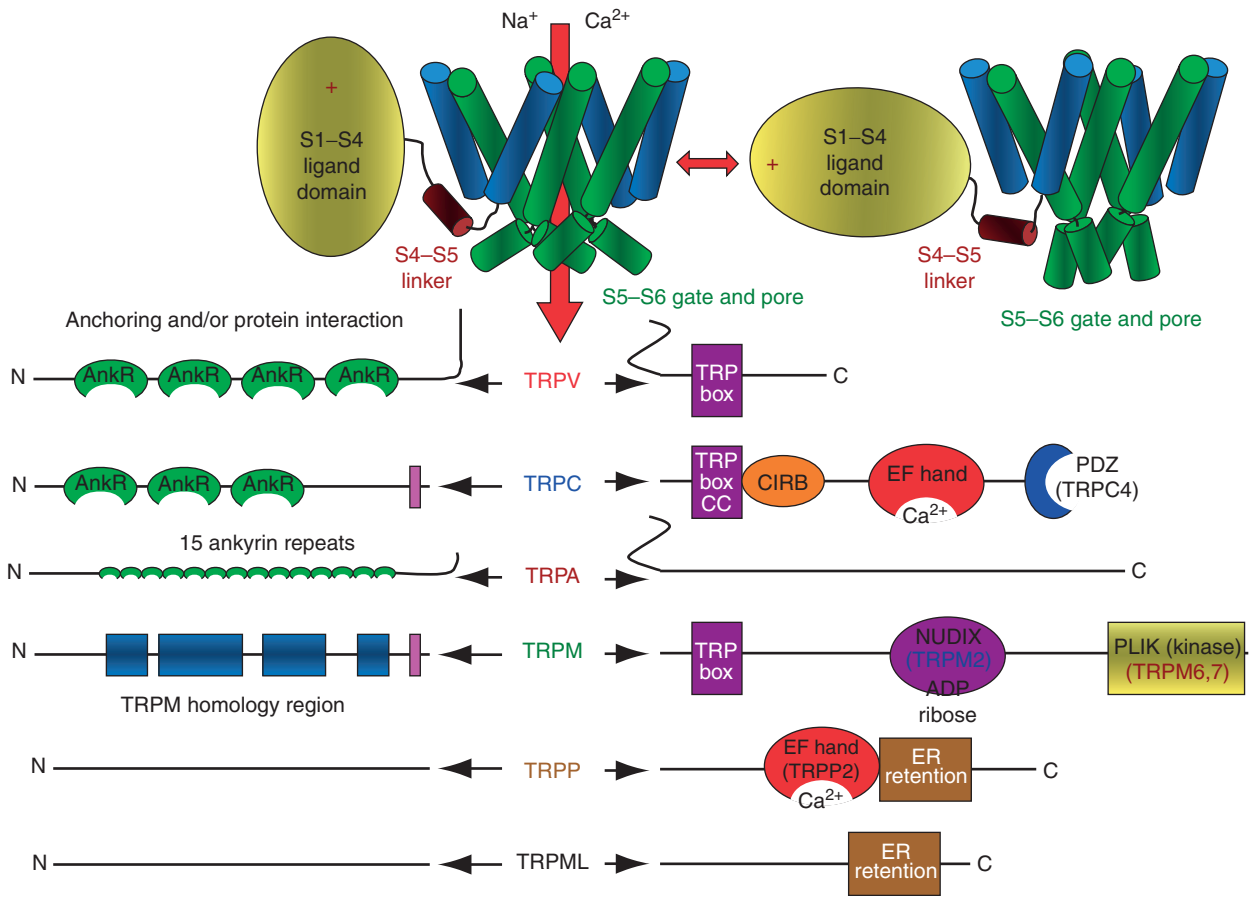

Figure 2 Transient receptor potential (TRP) channel architecture, with emphasis on the pore and cytoplasmic domains. The permeation pathway (red arrow) is formed by four sets of S5 (blue)–S6 (green) transmembrane segments. Pore loops (not shown) linking S5 and S6 segments dip into the membrane to form the selectivity filter. The S4–S5 linker relays movements in the voltage-sensing S1–S4 ligand domain (gold ellipse) to the S6 gate (green, lower). The major domains in the N- and C-terminal polypeptide chains vary by subfamily. The TRP box contains the amino acids EWKFAR in TRPC, but is less conserved in TRPV and TRPM families. AnkR, ankyrin repeat; CIRB, putative calmodulin and inositol 1,4,5-trisphosphate receptor binding domain; EF hand, Ca^{2+} binding domain; NUDIX, ADP ribose/È-NAD binding domain; PLIK, serine/threonine kinase; PDZ, a domain binding to scaffold proteins. Adapted from Clapham DE (2003) TRP channels as cellular sensors. *Nature* 426: 517–524, with permission.

acetylcholine (nicotinic receptor channel), and intracellular Ca^{2+} (Ca^{2+}-activated K^+ channel). Ambient temperature and pH affect all proteins, and thus all ion channel gating. Remember that by inference from work on K_V channels, each 6TM domain can be broken down into two functional units, a voltage-sensing domain (VSD; S1–S4) and a pore/gating domain (S5–S6). The VSD acts as a ligand that floats in the plasma membrane and is tethered via the S4–S5 linker to the pore domain. This linker transfers force from the movement of the VSD to the pore domain. Our current understanding of channels with this architecture is that the S6 helix is not rigid, but rather has a flexible hinge. Movement of the VSDs pulls the lower halves of the hinged S6 helices, splaying them apart to open the conduction path (Figure 2). So, what moves the VSD? TRP channels contain few charged amino acids in the VSD, and are not gated by voltage changes alone. In general, TRP channels have a few charged residues scattered in S4 and in the S4–S5 linker, and are thus less responsive to voltage changes. Additionally, this weak sensitivity can be significantly shifted within the physiological voltage range. For example, the voltage at which TRPV1 begins to open is shifted by −10 mV for every 1 °C increase in temperature, while TRPM8 activation shifts −7 mV for every 1 °C decrease in temperature. Temperature-sensitive TRP channels have Q_{10} values (the change in current imparted by a 10 °C temperature change) up to 30, which is 10 times larger than is typical for enzymes or other ion channels. Through TRP channels, changes in ambient temperature depolarize neurons and are directly encoded to changes in firing rate. Some TRP channels

are sensitive to unique and characteristic temperature ranges, from cool (<20 °C; TRPM8), to warm (>30 °C; TRPV3, TRPV4), to hot (>40 °C; TRPV1).

Many ligands alter TRP channel gating. Capsaicin, the active ingredient in 'hot' peppers, activates the heat-sensitive TRPV1 channels by direct binding to the VSDs. Capsaicin applied to TRPV1-containing neurons activates TRPV1, depolarizes these neurons, and results in increased firing of action potentials. The brain incorrectly interprets this information as a rise in ambient temperature and initiates signals to cause vasodilation and sweating. Many other lipophilic plant compounds, such as menthol, carvacrol, allicin, and camphor, bind to the VSDs of other TRP channels. Presumably, these compounds have evolved in plant leaves and seeds to evoke sensations in animals, ultimately serving the plant's survival.

An area of interest in TRP channels is the interaction of the S4–S5 linker with a hydrophobic pocket on S6. If likened to a knee joint, the upper leg (voltage-sensing domain) acts through the knee joint (S4–S5 linker/S6 pocket) to lift the lower leg (S6) and open the gate. This 'joint' may be accessible to modulation by intracytoplasmic elements, such as phosphatidylinositol 4,5-bisphosphate (PIP_2). Also, domains adjacent to the C-terminus of S6 are common modulatory binding sites for TRP channels (Figure 2), and may lie in close apposition to the crucial junction between S6 and the S4–S5 linker.

In summary, TRP channel gating is initiated by changes in voltage, temperature, and ligand binding. What differentiates TRP channel gating from classical voltage-gated channels is the rough equivalence of these three influences in mediating the conformational change that opens TRP channels. Thus, TRP channels act as signal integrators that sum the energies imparted by voltage, ligands, pH, and temperature to effect gating. Highly voltage-sensitive channels (Na_V) behave more as transistors, which are simple voltage-dependent switches.

TRP Channel Protein Domains

Outside the core 6TM segments (the N- and C-terminals), there are no common domains shared by all TRPs (Figure 2). The proximal C-terminal 'TRP box' containing the sequence EWKFAR in all TRPC family members is less conserved in TRPV and TRPM subfamilies. The function of this domain is not established, although it may play a role in binding PIP_2. As in many Ca^{2+}-permeant channels, intracellular calmodulin (IQ motifs) and Ca^{2+}-binding motifs (EF hand, aspartate/glutamate motifs) are common. Ankyrin repeats are present in TRPC, TRPV, and TRPA N-termini, and may act as flexible linkers to cytoskeletal elements, or as protein interaction domains to auxiliary subunits. Most unique are the domains in the TRPM class, where TRPM2 contains a Nudix motif (binding β-nicotinamide adenine dinucleotide or adenosine diphosphate ribose), and TRPM6 and TPM7 contain functional serine/threonine kinases within the polypeptide chain. All but the kinase domains affect TRP channel gating, conferring conformational changes on the tetramer upon binding or unbinding. Coiled-coil domains present often in the N-terminus are implicated in tetrameric assembly. Postsynaptic density/disc/zonula occludens-1 (PDZ) domains, present in TRPC1, TRPC4, and TRPC5, may interact with scaffolding proteins that organize large protein complexes.

TRP Channels Are Linked to Phospholipase C

Many, if not all, TRP channels are greatly potentiated by phospholipase C (PLC) activity. In the 1980s, genetic approaches established that PLC was organized in a complex with TRPs in photoreceptors. Photons impinging on a chromophore in a photoreceptor initiate a conformational change in the receptor protein. The receptor in turn catalyzes G-protein nucleotide exchange to form active G_α-guanosine triphosphate and $G_{\beta\gamma}$ subunits. These G-proteins activate phospholipase Cβ (PLCβ) to hydrolyze PIP_2 into soluble (inositol 1,4,5-trisphosphate) and lipophilic (diacylglycerol) messengers (Figure 3). In mammalian cells, the G_α subtype mediating this cascade is usually $G_{\alpha q}$. Similarly, growth factor receptors also potentiate TRP channel activity via a different PLC subtype, PLCγ. In this cascade, a growth factor (e.g., epidermal-derived growth factor) initiates the dimerization, autophosphorylation, and association with adapter proteins linking to PLCγ. Both PLCβ and PLCγ have pleckstrin homology, Ca^{2+}-binding EF hand, and C2 domains; both PLCs translocate from cytoplasm to membrane to become active. PLCs hydrolyze PIP_2 to form diacylglycerol (DAG) and inositol 1,4,5-trisphosphate (IP_3). These two second messengers are capable of releasing many active signal transduction molecules. Additionally, PIP_2 binds many membrane-associated molecules and its hydrolysis thus changes these activities.

To date, there is no unifying mechanism for TRP (or even TRP subfamily) activation by molecules up- and downstream of PLC. Direct activation of TRPs by PLC has not been extensively tested. However, downstream of PLC, IP_3 does not activate TRP channels by itself, but rather binds the endoplasmic reticular IP_3 receptor and releases the universal

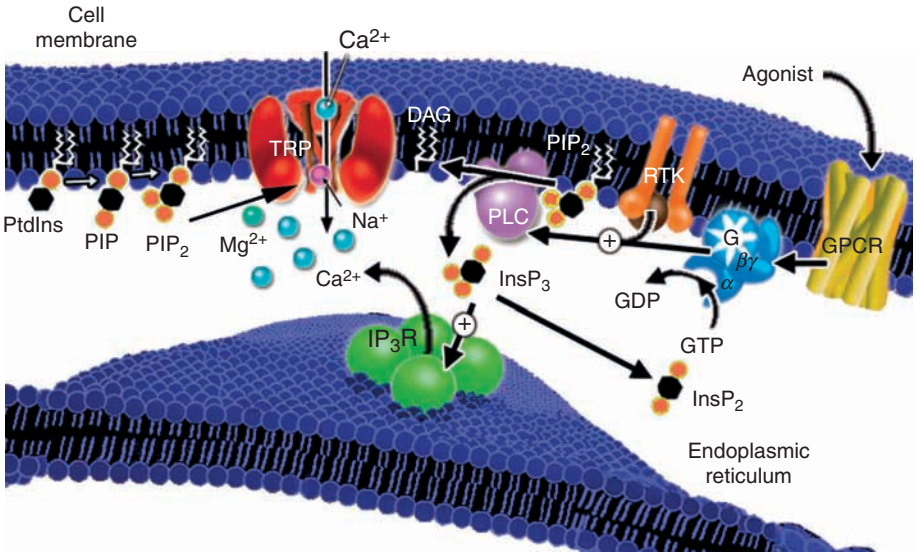

Figure 3 Signal transduction pathways common in transient receptor potential (TRP) channel activation or potentiation. PLC, phospholipase C; GPCR, G-protein-coupled receptor (PLCβ-coupled); G, G-protein heterotrimer consisting of α and $\beta\gamma$ subunits; GTP/GDP, guanosine tri- and diphosphate; RTK, tyrosine kinase receptor (PLCγ-coupled); DAG, diacylglycerol; PtdIns, phosphatidylinositol; PIP, phosphatidylinositol 4-phosphate; PIP_2, phosphatidylinositol 4,5-bisphosphate; PKC, protein kinase C; $InsP_3$, inositol 1,4,5-trisphosphate; IP_3R, IP_3 receptor. Adapted from Clapham DE (2003) TRP channels as cellular sensors. *Nature* 426: 517–524, with permission.

signaling ion, Ca^{2+}, into the cytoplasm. The activities of several TRP channels are enhanced by increases in intracellular Ca^{2+} or Ca^{2+}/calmodulin.

DAG activates protein kinase C (PKC), which affects several TRP channel activities. DAG, a diester of glycerol and two fatty acids, is best known for its anchoring and activation of PKCs, but it also binds and translocates protein kinase D, Ras guanyl nucleotide-releasing proteins (RasGRPs), and DAG kinase γ. DAG-dependent signaling is terminated by its phosphorylation by diacylglycerol kinase (DAGK), to create phosphatidic acid. *Drosophila* TRP channels are activated by DAG, and photoreceptor TRP channels are constitutively active in DAGK-defective mutants. DAG may directly activate mammalian TRPC3, TRPC6, and TRPC7, and heteromers containing these subunits. DAG is also converted into arachidonic acid by DAG lipase. Arachidonic acid, a second messenger, is the source of a large cascade of active molecules. It is oxidatively metabolized by the prostaglandin endoperoxide H synthases to yield the prostaglandins, thromboxane, and prostacyclin, by lipoxygenases (LOXs) to form the leukotrienes, hydroxyeicosatetraenoic acids (HETEs), and lipoxins, and by cytochrome P450 to create epoxygenase-derived eicosanoids (EETs). Several of these lipophilic messengers have also been shown to increase the activities of some TRP channels (Tables 1–4).

Interestingly, growth factor receptors activate both PLCγ and phosphatidylinositol 3-kinase (PI3K). PI3K, a master signaling molecule of cell growth, is a dimer of an 85 kDa regulatory subunit and a 110 kDa catalytic subunit. Tyrosine phosphorylation of growth factor receptors enables docking of p85 on the receptor. The partner of p85 is p110, which is proximal to its phospholipid substrate on the membrane. Thus, growth factor activation activates signals that both hydrolyze and create PIP_2. PI3K is also activated by $G_{\beta\gamma}$ subunits, linking the growth factor and G-protein-coupled receptor pathways. PI3K activation appears to promote insertion of several TRP channels into the plasma membrane, and thus increases the potential net TRP current.

TRP Channel Function

Calcium

While a subset of TRP channels (e.g., TRPML) appears to primarily span intracellular membranes, most TRP channels allow Ca^{2+} to enter cells. They either directly conduct Ca^{2+}, or depolarize cells to activate other Ca^{2+}-permeant channels. Ca^{2+} was adopted during evolution to play an especially important role inside cells. Ca^{2+} precipitates phosphate and is thus tightly regulated within the cell; for a cell at rest, $[Ca^{2+}]$ is ~20 000 fold lower in the cytoplasm than outside the cell. Ca^{2+} is maintained at 50–100 nM concentrations in the cytoplasm by export to the extracellular space, or by import into specialized compartments such as endoplasmic reticulum (ER) and mitochondria. Readily bound by proteins, Ca^{2+} is inherently a localized second messenger. Proteins often bind Ca^{2+} within a cage of approximately six oxygens formed by glutamate and aspartate residues.

Ca^{2+} escapes the grasp of negatively charged domains only for an average of ~50 μS before being bound again. This decreases its effective diffusion coefficient from ~800 $\mu m^2 s^{-1}$ in water to less than ~100 $\mu m^2 s^{-1}$. Thus, Ca^{2+} signals are highly localized.

The Ca^{2+} ion triggers fundamental cellular actions such as contraction, secretion, motility, cell division, and cell death. Ca^{2+} can bind, activate, or initiate the translocation of hundreds of proteins, either by itself or through binding to calmodulin or other adapter proteins. Ca^{2+} apparently gates only the two monovalent mammalian TRP channels, TRPM4b and TRPM5. In contrast, it is likely that it modulates directly or indirectly all TRP channels.

PLIK, the Kinase of TRPM6 and TRPM7

Two ion channels are unique among TRPs, and indeed among all ion channels. TRPM6 and TRPM7 are TRP channels that conduct Ca^{2+} and Mg^{2+} divalents into the cell, and monovalents in the outward direction. But they also contain a functional serine/threonine kinase, the phospholipase-interacting kinase (PLIK) domain, at the end of the C-terminus. PLIK autophosphorylates the channel, but this appears to play no role in gating the channel. The function of this kinase in the life of the cell is still a mystery.

TRP Channel Function in the Nervous System

TRPs in Sensation

Temperature and plant compounds Several TRP channels undergo large changes in gating with small temperature changes, and are present in known pain-and temperature-sensing neurons. TRPV1 is the clearest example of a TRP that gates in response to heat. Moreover, capsaicin and acidic pH shift the temperature-sensing range of TRPV1. Capsaicin activates only a subset of primary afferent neurons, particularly those Cδ and Aδ fibers responsible for pain and warmth perception. TRPV1 is most highly expressed in small- to medium-diameter neurons in dorsal root, trigeminal, and nodose ganglia. TRPV1 knockout mice show impaired responses to heat (>~42 °C) and capsaicin-induced pain, consistent with TRPV1's localization. TRPV1 activity is also potentiated by G_q-coupled receptors and growth factors linked to PLC, perhaps explaining the pain-enhancing effects of bradykinin and nerve growth factor. Within sensory ganglia, TRPV2 is highly expressed in medium- to large-diameter neurons that give rise to Aδ and Aβ fibers, but is also less localized to the nervous system than is TRPV1. TRPV2 is insensitive to capsaicin but is activated by noxious temperatures.

TRPV3 and TRPV4, abundant in skin, mouth, and nasal epithelia, are activated by warmth (roughly above 30 °C). Interestingly, unlike most ion channels, TRPV3 current increases (sensitizes) upon repeated exposure to warmth. TRPV3 is also activated by camphor, carvacrol (oregano), and eugenol (cloves). TRPV4 is activated by hypoosmotic stimuli, endogenous metabolites of arachidonic acid, and phorbol esters. *TRPV4* deletion in mice alters central nervous system (CNS) control of ADH secretion. Both TRPV3 and TRPV4 knockout mice have prolonged withdrawal responses to acute peripheral heating and exhibit decreased thermal hyperalgesia following peripheral inflammation.

Cooling below ~25 °C activates TRPM8. Menthol shifts the activation range of the channel to room temperature, thus explaining the cooling sensation of mint. TRPM8 and TRPV1 are found in non-overlapping populations of small-diameter peripheral sensory neurons, leading to speculation that they mediate responses to mild heat and cooling, respectively. The broad tissue distributions of TRPM8 and TRPV1–TRPV4, and the osmotic sensitivity TRPV4, suggest that these channels may do more than just sense temperature.

TRPA1 is co-localized with TRPV1 in a subset of small- to medium-diameter peripheral sensory neurons. Pungent chemicals (isothiocyanates such as those found in mustard oil, wasabi, and garlic, as well as the smoke irritant, acrolein) activate this channel. Anandamide and tetrahydrocannabinoid (THC; the active ingredient in marijuana) induce current in TRPA1-expressing cells, suggesting they might be ionotropic cannabinoid receptors. TRPA1 was initially reported to sense noxious cold (<15 °C) *in vitro*, but this finding has been contested. Neurons still respond to cold sensitivity in mice lacking TRPA1, suggesting additional sensors for noxious cold.

Mechanical forces Ciliated structures act as levers to amplify mechanical forces. *Nanchung*, a member of the *Drosophila* TRPV family, localizes to the cilia of chordotonal neurons and is activated by pressure. Despite normal chordotonal organ morphology, mutants are deaf. *NompC*, a second member of the *Drosophila* TRP family is required for pressure-induced currents in sensory bristles, and may also be a mechanosensitive channel. Reduction of zebra fish *NompC*-like TRP impairs hair cell signal transduction, producing loss of sound sensation in the fish. TRPA1, like *NompC*, has many ankyrin repeats, suggesting that TRPA1 might be mechanosensitive in mammals. This led to the proposal that TRPA1 was the 'hearing' channel of the hair cell, but subsequent experiments showed that *TRPA1*−/− mice had

normal hearing. *In vitro* experiments have also led to the suggestion that TRPC1 is the stretch-sensitive ion channel commonly observed in oocytes.

Taste TRPM5 expresses at high levels in taste tissue and co-localizes with other known elements of the gustatory signaling pathway, including α-gustducin. The T1R and T2R G-protein-coupled taste receptors initiate a common signaling pathway involving PLCβ2 and TRPM5 to encode sweet, umami (amino acid), and bitter taste sensations. This finding indicates that TRPM5 is a common taste transduction element involved in relaying information from a subset of different taste receptors. Studies focusing exclusively on heterologously expressed TRPM5 indicate that the current is a voltage-modulated, monovalent-selective cation conductance that is activated by micromolar levels of intracellular Ca^{2+}. Receptor stimulation upstream of PLCβ2 (which co-localizes with TRPM5) generates IP_3 and thereby increases intracellular [Ca^{2+}] to a level that triggers ion flux through TRPM5.

TRPP2 is present in sour taste receptor cells of the tongue (circumvallate and foliate papilla), as well as in cells in the spinal cord central canal. This TRP channel of the polycystin family appears to mediate acid sensing in the tongue (sour taste), and potentially in cerebrospinal fluid.

Pheromone sensing Murine TRPC2 is essential for the transmission of pheromone-mediated signals. Male mice that lack TRPC2 do not display stereotypical male–male aggression responses and mate indiscriminately with male or female mice, resulting from a lack of pheromone-evoked neuronal activity in their vomeronasal organs (VNOs). Humans lack a discrete VNO and human *TRPC2* does not transcribe a functional protein (i.e., is a pseudogene), but other pheromone-sensing mechanisms may exist.

TRPs in the Brain

TRPs are widely distributed in the brain, with every TRP subunit being represented in at least one study. With the exception of the aforementioned sensory defects, few phenotypes have yet been described that result from CNS-specific deletion of TRPs.

Axons are guided to their targets in the nervous system by diffusible and cell-bound guidance cues. The neuronal growth cone is a palm-shaped structure at the tips of neurites that modulates the rate and direction of neurite extension in response to attractive and repellent cues. Attractive cues include netrin-1 and brain-derived neurotrophic factor (BDNF); myelin-associated glycoprotein and semaphorin 3A are repellant. Intracellular Ca^{2+} levels affect growth cone morphology and trafficking of membrane precursor vesicles. Thus, activation or block of TRP channels would be expected to have some effect on growth cones. In particular, TRPC1, TRPC3, TRPC4, TRPC5, and TRPC6-containing channels have been shown to affect neurite extension, growth cone morphology, and growth cone guidance in hippocampal, cerebellar granule, or spinal cord neurons. Synaptic vesicles move within individual growth cone filopodia and fuse with the plasma membrane after focal depolarization. Regulation of vesicle transport and fusion therefore represents an intriguing mechanism by which TRP channels can participate in steering growth cones or other chemotactic cells.

TRPCs may also underlie metabotropic glutamate receptor (mGluR)-dependent conductances in rat CA3 pyramidal neurons, midbrain dopaminergic neurons, and cerebellar Purkinje neurons. The mGluR-dependent currents in CA3 pyramidal neurons resemble currents from heterologously expressed TRPC1/TRPC5 heteromultimers. G_q activation also stimulates TRPC1/TRPC5 and this current shows a region of negative slope conductance, similar to the mGluR-dependent current observed in brain slice recordings.

Summary

TRPs are a large family of loosely homologous, broadly distributed ion channel proteins. They form homo- and heterotetrameric ion channels that span the plasma membrane; some also form channels across intracellular vesicular membranes. Despite initial enthusiasm, store-operated Ca^{2+} entry does not appear to be a TRP channel function. On the plasma membrane, TRPs are activated by a variety of mechanisms, including temperature, pH, voltage, chemical ligand binding, and mechanical forces. These surface channels depolarize cells and conduct Na^+, Ca^{2+}, and, in some cases, Mg^{2+} into cells. Many plant compounds activate TRPs, suggesting that plants evolved chemicals to repel or attract animals via these sensory channels. Vesicular TRP channels may serve to exchange cations to provide ionic balance as vesicles are loaded with various charged molecules. Many other mysteries remain, such as how TRP channels are gated, their function in vesicles and cilia, their links to neuronal guidance cues, and their roles in functions as disparate as central temperature regulation, pain sensing, smell, and stroke. TRP channels are intrinsically integrative cellular sensors of their environment.

See also: Calcium Channel Subtypes Involved in Neurotransmitter Release; Calcium Channels and SNARE Proteins; Cyclic Nucleotide-Gated and Hyperpolarization-Activated Channels; Inwardly Rectifying Potassium Channels; Large Conductance Calcium-Activated Potassium Channels; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Caterina MJ, Schumacher MA, Tominaga M, et al. (1997) The capsaicin receptor: A heat-activated ion channel in the pain pathway. *Nature* 389: 816–824.

Caterina MJ (2007) Transient receptor potential ion channels as participants in thermosensation and thermoregulation. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology* 292: 64–76.

Clapham DE (2003) TRP channels as cellular sensors. *Nature* 426: 517–524.

Clapham DE (2007) Mammalian TRP channels. *Cell* 129(1): 220.

Clapham DE, Julius D, Montell C, et al. (2005) International Union of Pharmacology. XLIX. Nomenclature and structure–function relationships of transient receptor potential channels. *Pharmacological Reviews* 57: 427–450.

Desai C (2005) TRP channels and mice deficient in TRP channels. *Pflügers Archive* 451: 11–18.

Gouaux E and Mackinnon R (2005) Principles of selective ion transport in channels and pumps. *Science* 310: 1461–1465.

Harteneck C, Plant TD, and Schultz G (2000) From worm to man: Three subfamilies of TRP channels. *Trends in Neuroscience* 23: 159–166.

Kahn-Kirby AH and Bargmann CI (2006) TRP channels in *C. elegans*. *Annual Review of Physiology* 68: 719–736.

Krapivinsky G, Mochida S, Krapivinsky L, et al. (2006) The TRPM7 ion channel functions in cholinergic synaptic vesicles and affects transmitter release. *Neuron* 52: 485–496.

Long SB, Campbell EB, and Mackinnon R (2005) Crystal structure of a mammalian voltage-dependent Shaker family K^+ channel. *Science* 309: 897–903.

Minke B and Cook B (2002) TRP channel proteins and signal transduction. *Physiological Reviews* 82: 429–472.

Montell C (2005) The TRP superfamily of cation channels. *Science STKE* 272:re3 (doi:10.1126/stke.2722005re3).

Moran MM, Xu H, and Clapham DE (2004) TRP ion channels in the nervous system. *Current Opinion in Neurobiology* 14: 362–369.

Nilius B, Voets T, and Peters J (2005) TRP channels in disease. *Science STKE* 295:re8 (doi:10.1126/stke.2952005re8).

Owsianik G, Talavera K, Voets T, et al. (2006) Permeation and selectivity of TRP channels. *Annual Review of Physiology* 68: 685–717.

Ramsey IS, Delling M, and Clapham DE (2006) An introduction to TRP channels. *Annual Review of Physiology* 68: 619–647.

Runnels LW, Yue L, and Clapham DE (2001) TRP-PLIK, a bifunctional protein with kinase and ion channel activities. *Science* 291: 1043–1047.

Stowers L, Holy TE, Meister M, et al. (2002) Loss of sex discrimination and male–male aggression in mice deficient for TRP2. *Science* 295: 1493–1500.

Strübing C, Krapivinsky G, Krapivinsky L, et al. (2001) TRPC1 and TRPC5 form a novel cation channel in mammalian brain. *Neuron* 29: 645–655.

Wang H and Woolf CJ (2005) Pain TRPs. *Neuron* 46: 9–12.

Wes PD, Chevesich J, Jeromin A, et al. (1995) TRPC1, a human homolog of a *Drosophila* store-operated channel. *Proceedings of the National Academy of Sciences of the United States of America* 92: 9652–9656.

Xu H, Ramsey IS, Kotecha SA, et al. (2002) TRPV3 is a calcium-permeable temperature-sensitive cation channel. *Nature* 418: 181–186.

Relevant Websites

http://clapham.tch.harvard.edu – Clapham Laboratory (Children's Hospital, Howard Hughes Medical Institute, Boston, MA).

http://www.iuphar-db.org – IUPHAR Ion Channel Compendium (edited by William Catterall and George Gutman).

Translational Regulation at the Synapse

M M Poon, University of California at Berkeley, Berkeley, CA, USA
K C Martin, University of California at Los Angeles, Los Angeles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The localization and regulated translation of mRNAs within distinct subcellular compartments serve as an important mechanism for spatially regulating gene expression in asymmetric cells. Regulated translation of localized mRNAs provides a means of rapidly altering the macromolecular composition of a specific compartment in a cell and thereby decentralizes the regulation of gene expression from the nucleus to distinct subcellular compartments. This spatially restricted gene expression is likely to play an especially important role in neurons, where the thousands of synapses made by a single neuron are potentially capable of operating as autonomous compartments. Studies in a number of systems have defined roles for local translation during axon guidance, synapse formation, neuronal injury and regeneration, and synaptic plasticity. In each of these processes, local translation serves to persistently alter the structure and function of synapses in a spatially restricted manner.

Early ultrastructural studies of neurons indicated that polyribosomes were present exclusively in the soma of neurons, and these studies were used as evidence that all proteins were translated in the cell body and subsequently transported to distal compartments. An initial indication that translation might occur at synapses came from electron microscopy studies performed in the 1960s by Bodian, in which he observed clusters of polyribosomes at synaptic sites in monkey spinal cord neurons. Nearly 20 years later, in 1982, Steward and Levy published a report in which they also, using electron microscopy, detected clusters of polyribosomes selectively localized beneath postsynaptic sites on the dendrites of hippocampal granule cell neurons. This classic study captured the attention of the scientific community and focused attention on the possibility that extrasomatic protein synthesis might occur in neurons.

Davis, Steward, and Banker demonstrated in 1987 that RNAs were transported into dendrites of cultured hippocampal neurons, using radioactive uridine precursors to visualize RNAs. Together with the discovery of polyribosomes at spines, these studies provided further support for the idea that proteins might be synthesized in dendrites. A year later, in 1988, Garner, Tucker, and Matus published *in situ* hybridization studies clearly showing that the mRNA encoding microtubule-associated protein 2 (MAP2) was present in dendrites in developing (6-day-old) rat cortex and hippocampus, but that tubulin mRNA was restricted to the cell body. Two years later, in 1990, Paul Kelly and colleagues published striking *in situ* hybridization studies showing that the mRNA encoding the calcium/calmodulin-dependent protein kinase II alpha (CaMKIIα) localized to dendrites of adult rat hippocampus, while the mRNA encoding CaMKIIβ was restricted to cell bodies. Subsequent *in situ* hybridization studies showed that, while most mRNAs in the brain were present exclusively in the cell body, a select population of mRNAs was present in the dendritic compartment of vertebrate neurons. The restricted number of mRNAs present in neuronal processes was itself considered to be an attractive feature: presumably this population of mRNAs consisted of transcripts critical in some way to local synaptic function.

Not only were polyribosomes and mRNAs detected in dendrites, but a number of studies also indicated that many components of the translational machinery, including ribosomal proteins, initiation and elongation factors, as well as organelles (endoplasmic reticulum and Golgi apparatus) involved in the synthesis of membrane proteins, were present in dendrites. Taken together, these studies demonstrated that all the molecules required to synthesize new proteins are present in distal neuronal processes. As a result, the idea that local translation can occur is now widely accepted in the neuroscience community, and the focus of investigation has shifted from asking whether extrasomatic translation occurs in neurons to addressing the following questions: (1) Where does local translation occur – that is, presynaptically and/or postsynaptically? (2) What are the mRNAs that are localized? (3) How are the mRNAs localized? (4) How is their translation regulated? (5) What is the function of local translation?

The Sites of Local Translation in Neurons

Does local translation occur in all subcellular compartments of neurons? In the course of ultrastructural studies of hippocampal neurons growing in dissociated culture, Bartlett and Banker noted that while ribosomes could be detected in axons very early in development, they were restricted to dendrites at later time points. Performing *in situ* hybridization for specific mRNAs in cultured hippocampal neurons, Kleiman, Banker, and Steward also found that while

specific mRNAs were present in axons and dendrites at very early time points in culture, they were restricted to dendrites in more mature cultures. The conclusion from these and a number of subsequent studies was that in vertebrate neurons, presynaptic, axonal protein synthesis occurs during development, and that local translation is restricted to postsynaptic, dendritic compartments in mature neurons.

In the developing nervous system, polyribosomes, translation factors, and specific mRNAs have been detected in axonal growth cones of vertebrate neurons, including retinal ganglion cells and sympathetic, spinal cord, cerebellar, and cortical neurons. These mRNAs encode cytoskeletal elements, for example, β-actin, and signaling proteins that modulate the cytoskeleton, for example, rhoA. Studies using a variety of inhibitors have demonstrated that local translation in the growth cone is required for axon guidance and for synapse formation.

The evidence for local translation in mature axons is less complete, although there are indications that mRNAs may be translated in mature axons of some neuronal cell types. Specifically, translation in mature axons has been defined following an injury such as a sciatic nerve crush, when translation is thought to be involved in axonal regeneration. Ribosomes and/or ribosomal RNA have been detected in the adult rabbit and rat spinal nerve roots and in the axon of the goldfish Mauthner neurons, suggesting that translation can occur in these mature axons as well. Specific mRNAs have been detected in mature axons of particular neuronal cell types; for example, mRNAs encoding neurofilament proteins have been detected in the mature axons of goldfish Mauthner neurons, and mRNAs encoding neuropeptides (e.g., vasopressin) have been detected in adult rodent hypothalamo-hypophysial axons. Together, these studies hold open the possibility that axons of mature vertebrate neurons may retain the capacity to translate proteins.

Messenger RNAs and local translation have also been detected in the mature 'axons' of invertebrate neurons, including in the squid giant axon and in neurons from *Aplysia californica* and *Lymnaea*. Not only have ribosomes, translational machinery, and mRNAs been detected in these invertebrate axons, but metabolic labeling and reporter construct assays have also directly demonstrated that mRNAs can be translated in isolated axonal preparations. It is important to note, however, that the polarity of these axons is not as well defined as in the case of vertebrate neurons. Thus, for example, *Aplysia* neurons elaborate neuronal processes called neurites, which contain pre- and postsynaptic elements within a single process. As such, the existence of axonal translation in mature invertebrate axons may not translate directly to the existence of axonal translation in mature vertebrate axons.

That local translation occurs in postsynaptic dendrites of mature vertebrate neurons is now extremely well established. Many mRNAs have been detected in dendrites of rodent central nervous system (CNS) neurons. Translation initiation and elongation factors are present in dendrites and are activated, for example, by phosphorylation, following neuronal stimulation. Following up on Steward and Levy's 1982 description of polyribosomes at the base of spines in hippocampal granule cells, Harris and colleagues have performed serial section electron microscopic analyses of rodent CA1 hippocampal neurons to quantify the number of polyribosomes that are present at postsynaptic sites. Focusing on spines in CA1 pyramidal neurons in 15-day-old rat, they found that 12% of spines had polyribosome clusters present at the base or within the spine. Remarkably, following induction of long-term potentiation (LTP), the number of spines with associated polyribosomes increased to 39%. Furthermore, the spines with polyribosome clusters had larger postsynaptic densities than did spines lacking polyribosomes, suggesting that active translation served to alter the structure of potentiated synapses. Harris and colleagues have subsequently made similar observations in ultrastructural studies of mature (60- to 70-day-old) rat hippocampus.

What mRNAs Are Localized?

Initially, only a handful of mRNAs were thought to be localized, and these were identified primarily by chance in the course of *in situ* hybridization studies. The most striking result to emerge from such studies was the presence of specific mRNAs encoding cytosolic, cytoskeletal, and integral membrane proteins in dendritic layers of the hippocampus and at postsynaptic densities of hippocampal neurons *in vivo* and *in vitro*. These included mRNAs encoding MAP2, CaMKIIα, brain-derived neurotrophic factor (BDNF), activity-regulated cytoskeleton-associated protein (Arc), tyrosine kinase B (TrkB) receptor, L7 protein (in cerebellum), inositol trisphosphate (IP_3) receptor, the atypical protein kinase Mzeta, the *N*-methyl-D-aspartate receptor (NMDAR) NR1 subunit, and glycine receptor α subunit. Recent high-resolution *in situ* hybridization studies and studies using a variety of reporter constructs have revealed that a number of other mRNAs, including those encoding α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) glutamate receptor subunits and the potassium channel Kv1.1, are also present in dendrites of vertebrate CNS neurons.

Unbiased approaches to amplify mRNAs from mechanically or biochemically purified dendritic

and/or synaptic compartments have generated lists of localized mRNAs numbering in the hundreds. These experiments suggest that there is a large but select population of localized mRNAs, although wide acceptance of the dendritic localization of these mRNAs has been delayed by the lack of *in situ* hybridization demonstrating that the mRNAs are indeed present in dendrites. Conclusive identification of the entire population of axonally and/or dendritically localized mRNAs remains challenging both because of the difficulties of obtaining pure preparations of axons or dendrites and because of the difficulty of detecting the small quantities of mRNA that are localized in neuronal processes.

As described previously, the mRNAs that are localized to dendrites depend on the developmental state of the neuron, with mRNAs such as β-actin being present in the axonal and dendritic growth cones of hippocampal neurons, but restricted to the cell body of mature hippocampal neurons. The population of localized mRNAs also appears to be regulated by activity, such that a number of mRNAs have been found to localize to dendrites only following stimulation. These include *arc*, which not only translocates into dendrites but also accumulates specifically at stimulated synaptic sites, as well as *BDNF* and *trkB*. These findings suggest that the complement of mRNAs that is present in axons and dendrites is dynamically regulated by development and by activity.

How Are mRNAs Localized to Axons or Dendrites in Neurons?

Early studies used nucleic acid stains and green fluorescent protein fused to RNA-binding proteins to visualize mRNA translocation in live neurons. These studies showed that localized mRNAs are transported in the form of large granules containing mRNAs, RNA-binding proteins, ribosomes, and translational factors in a rapid (average speed $0.1\,\mu m\,s^{-1}$), bidirectional, and microtubule-dependent manner. As has been found to be the case in nonneuronal cells, mRNA localization in neurons has been shown to depend on *cis*-acting sequences, usually in the 3′ untranslated region (UTR), and on *trans*-acting RNA-binding proteins that bind to these sequences. Several groups have shown that the double-stranded RNA-binding protein Staufen is present in adult neurons and that it participates in the trafficking of mRNAs into dendrites. Staufen is proposed to be a component of RNA-containing granules, serving to mediate the transport of these granules along microtubules into dendrites. Bassell, Singer, and colleagues have identified *cis*-acting sequences, termed the 'zip code,' that mediate the localization of β-actin mRNA by binding to the zip code binding protein (ZBP)-1. ZBP1 is a member of the heterogeneous nuclear ribonucleoprotein (hnRNP) family of proteins that also includes hnRNP A2 and fragile X mental retardation protein (FMRP), both of which have been implicated in dendritic mRNA transport. Richter and colleagues have shown that in addition to its role in translational regulation, the cytoplasmic polyadenylation element-binding protein (CPEB) functions to facilitate mRNA transport to dendrites. Notably, disruption of ZBP1, ZBP2, Staufen, or CPEB impairs localization of targeted mRNAs, demonstrating a critical role for mRNA-binding proteins as molecular adapters between *cis*-acting mRNA localization sequences and microtubule-based transport machinery.

Recent studies have begun to characterize the RNA granules that carry mRNAs into dendrites. In a *tour de force*, Hirokawa and colleagues, using affinity chromatography with a kinesin antibody and proteomics, identified more than 42 proteins that constitute the large RNA-containing granule and showed that a subset of component proteins – hnRNP U, Staufen, purine-rich element-binding protein alpha (PURα), and pyrimidine tract binding protein-associated splicing factor (PSF) – are critical to dendritic RNA transport. Additionally, these investigators implicated the KIF5 isoform of kinesin as a motor involved in the active transport of RNA-containing granules into distal dendrites. Studying the composition of RNP particles that mediate transport of RNA in dendrites of *Drosophila* neurons, Ramaswami, Parker, and colleagues discovered that these 'RNA granules' share many components with RNA processing (P) bodies. Athough P bodies were initially described as sites of mRNA degradation, recent studies of micro-RNA (miRNA)-mediated translational regulation have indicated that P bodies are sites of translational silencing. These results suggest that RNA granules serve not only to transport mRNAs to distal sites in neurons, but also to keep these neurons in a dormant, silenced state.

How Is the Translation of Localized mRNAs Regulated?

The translation of localized mRNAs appears to be regulated by multiple mechanisms. Given the large number of mRNAs that have been identified as being dendritically and/or axonally localized, this is not surprising, since one would expect that distinct stimuli would activate the translation of distinct mRNAs via distinct signaling pathways. Furthermore, as specific mRNAs may contain multiple regulatory elements, the control of translation may place mRNAs in a unique position to act as coincidence detectors,

integrating several stimuli in order to regulate protein synthesis. In some studies, stimulus-induced translation of localized mRNAs has been shown to recruit rapamycin-dependent translational regulation, a mechanism involving regulation of translational initiation via the terminal oligopyrimidine tracts within the 5′UTR regions of specific mRNAs. Interestingly, mRNAs containing these tracts include many of the ribosomal proteins as well as translation initiation and elongation factors. Components of rapamycin-sensitive translation have been detected in dendrites of hippocampal neurons and in *Aplysia* synaptosomes. Stimulus-induced translation of localized mRNAs has also been shown to be mediated by regulation of cytoplasmic polyadenylation. For example, NMDAR stimulation has been shown to lead to phosphorylation and activation of CPEB, resulting in increased length of the polyA tail of CaMKIIα, and increased translation of the message. Edelman and colleagues, working in vertebrate neurons, and Sossin and colleagues, working in invertebrate neurons, have proposed that translation of mRNAs using internal ribosomal entry sites provides an additional mechanism for specifically regulating the translation of several localized mRNAs at synapses. Roles for miRNA-mediated translational regulation of localized mRNAs have also been recently discovered. In one study, Kunes and colleagues found that the mRNA encoding CaMKIIα was present in dendrites of *Drosophila* neurons and noted that the 3′UTR contained two putative miRNA binding sites, indicating that binding of miRNAs might function to repress translation. Consistent with this possibility, the learning-induced translation of CaMKIIα was associated with degradation of components of the miRNA machinery, suggesting that persistent generation of miRNAs was required to repress translation. Greenberg and colleagues identified a miRNA that binds to a localized mRNA, encoding LIM motif-containing protein kinase 1 (LimK1), and showed that modulation of the miRNA concentration altered LimK1 protein concentration, which in turn altered spine structure. In one fascinating study, Sutton, Schuman, and colleagues found that local translation in dendrites was negatively regulated by spontaneous neurotransmission. These results suggest that one function of spontaneous neurotransmission may be to sustain translational repression of localized mRNAs.

What Is the Function of Local Translation?

In this section, we focus on studies elucidating a role for local translation during synaptic plasticity in mature neurons. Local protein synthesis has been found to be essential for long-lasting forms of activity-dependent plasticity of hippocampal synapses. Specifically, Kang and Schuman found that hippocampal LTP produced by application of BDNF showed a rapid dependence on protein synthesis and occurred even after the cell bodies had been severed from the dendrites, indicating that the translation occurred in CA1 dendrites. Huber, Bear, and colleagues found that long-term depression (LTD) of hippocampal synapses induced by application of metabotropic glutamate receptors also showed a rapid dependence on protein synthesis (with no dependence on transcription), and furthermore that this LTD occurred in slices in which the dendrites had been severed from cell bodies. Studies from several groups have used reporter constructs to demonstrate that stimuli that elicit long-lasting forms of plasticity (such as LTP) also elicit translation within dendrites.

A role for local protein synthesis during synaptic plasticity has also been defined in *Aplysia* neurons. Using a cell culture system in which a single sensory neuron with a bifurcated axon makes synaptic contact with two spatially separated motor neurons, Martin, Kandel, and colleagues demonstrated that local application of the modulatory neurotransmitter serotonin to the connections made onto one of the motor neurons produced a branch-specific long-term facilitation of that connection without any change in the strength of the unstimulated connection. This branch-specific plasticity was shown to depend on the translation of mRNAs in the distal sensory neurites. Further investigation of this process indicates that local translation is required for the initial induction of branch-specific plasticity and that it is also required for the stabilization of new synaptic growth. Scherff, Carew, and colleagues have made similar observations studying synaptic plasticity in preparations of intact *Aplysia* ganglia.

The role of local translation during learning and memory in behaving animals has also been addressed in a number of studies. In the first, Miller, Mayford, and colleagues expressed a CaMKIIα transgene lacking the 3′UTR (which contains the *cis*-acting dendritic targeting sequences) in CaMKIIα null mice. Using this strategy, they succeeded in generating a mouse that expressed CaMKIIα but lacked dendritically localized CaMKIIα mRNA, allowing them to specifically assay the function of the dendritically translated enzyme. These mutant mice showed deficits in the late phase of LTP at hippocampal synapses and in long-term memory, demonstrating a role for local translation of CaMKIIα during long-lasting forms of plasticity and memory. Intriguingly, these studies also indicated that most of the CaMKIIα protein present in the postsynaptic density derives from locally synthesized protein, since removal of dendritic mRNA produced a dramatic reduction of CaMKIIα in the postsynaptic

density. Ashraf, Kunes, and colleagues found that CaMKIIα mRNA was present in dendrites of *Drosophila* neurons and that its translation was regulated by olfactory learning. Manipulations that prevented dendritic translation of CaMKIIα also inhibited learning, suggesting that local translation is required for learning.

Studies of mice lacking molecules involved in translational regulation have provided further support for the importance of local translation during learning and memory, although in these studies, translation is altered throughout the neuron. Klann and colleagues have studied mice lacking the eukaryotic translation initiation factor 4E (eIF4E) binding protein 2, which normally functions to inhibit translation by binding, in its unphosphorylated state, to eIF4E. When LTP was assayed in these mice, stimuli that usually produced transient LTP elicited long-lasting LTP, while stimuli that usually produced long-lasting LTP produced decreased LTP, suggesting that the increased basal rate of translation altered the threshold for LTP induction. In behavioral assays, these mice exhibited deficits in spatial learning and in long-term contextual conditioning. Sonenberg and colleagues observed a similar phenotype in mice lacking GCN2, a protein kinase that nhibits translational initiation by phosphorylating eIF2a. Interestingly, experiments by Naegerl, Bonhoeffer, and colleagues, as well as others, have suggested a balance between local synthesis and protein degradation and that perturbation of either process, using translational or proteasome inhibitors, led to deficits in late phase LTP, whereas perturbation of both processes rescued the impairment.

Conclusions

The concept of local translation at synapses in neurons has gained increasing acceptance over the past decades. Recent studies have outlined specific roles for local translation during axon guidance, synapse formation, injury and regeneration, and synaptic plasticity. Novel approaches have been used to identify localized transcripts and to elucidate the mechanisms whereby mRNAs are localized and whereby their translation is regulated by neuronal activity. Together, these investigations indicate that neurons are endowed with a remarkable capacity for rapidly regulating gene expression within spatially restricted subcellular compartments.

See also: Activity-Dependent Remodeling of Presynaptic Boutons; Axonal mRNA Transport and Functions; Calcium-Calmodulin Kinase II (CaMKII) in Learning and Memory; Dendrites: Localized Translation; Dendritic RNA Transport: Dynamic Spatio-Temporal Control of Neuronal Gene Expression; Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; RNA Binding Protein Methods.

Further Reading

Aakalu G, Smith WB, Nguyen N, Jiang C, and Schuman EM (2001) Dynamic visualization of local protein synthesis in hippocampal neurons. *Neuron* 30: 489–502.

Huber KM, Kayser MS, and Bear MF (2000) Role for rapid dendritic protein synthesis in hippocampal mGluR-dependent long-term depression. *Science* 288: 1254–1257.

Kang H and Schuman EM (1996) A requirement for local protein synthesis in neurotrophin-induced hippocampal synaptic plasticity. *Science* 273: 1402–1406.

Kiebler MA and Bassell GJ (2006) Neuronal RNA granules: Movers and makers. *Neuron* 51: 685–690.

Klann E and Dever TE (2004) Biochemical mechanisms for translational regulation in synaptic plasticity. *Nature Reviews Neuroscience* 5: 931–942.

Martin KC (2004) Local protein synthesis during axon guidance and synaptic plasticity. *Current Opinion in Neurobiology* 14: 305–310.

Martin KC, Casadio A, Zhu H, et al. (1997) Synapse-specific, long-term facilitation of aplysia sensory to motor synapses: A function for local protein synthesis in memory storage. *Cell* 91: 927–938.

Miller S, Yasuda M, Coats JK, Jones Y, Martone ME, and Mayford M (2002) Disruption of dendritic translation of CaMKIIalpha impairs stabilization of synaptic plasticity and memory consolidation. *Neuron* 36: 507–519.

Ostroff LE, Fiala JC, Allwardt B, and Harris KM (2002) Polyribosomes redistribute from dendritic shafts into spines with enlarged synapses during LTP in developing rat hippocampal slices. *Neuron* 35: 535–545.

Piper M and Holt C (2004) RNA translation in axons. *Annual Review of Cell and Developmental Biology* 20: 505–523.

Sossin WS and DesGroseillers L (2006) Intracellular trafficking of RNA in neurons. *Traffic* 7: 1581–1589.

Steward O and Levy WB (1982) Preferential localization of polyribosomes under the base of dendritic spines in granule cells of the dentate gyrus. *Journal of Neuroscience* 2: 284–291.

Steward O, Wallace CS, Lyford GL, and Worley PF (1998) Synaptic activation causes the mRNA for the IEG Arc to localize selectively near activated postsynaptic sites on dendrites. *Neuron* 21: 741–751.

Sutton MA and Schuman EM (2006) Dendritic protein synthesis, synaptic plasticity, and memory. *Cell* 127: 49–58.

Willis DE and Twiss JL (2006) The evolving roles of axonally synthesized proteins in regeneration. *Current Opinion in Neurobiology* 16: 111–118.

Transplantation of Myelin Forming Cells

V Tepavčević and A Baron-Van Evercooren, Institut National de la Santé et de la Recherche Medicale, Paris, France; Université Pierre et Marie Curie, Paris, France; and Hôpitaux de Paris, Hôpital Pitié-Salpêtrière, Paris, France

© 2009 Elsevier Ltd. All rights reserved.

Myelin Pathologies That Could Be Targeted by Transplantation

Transplantation of myelin-forming cells is considered a potential treatment for leukodystrophies, multiple sclerosis, and spinal cord injury. Congenital myelin disorders such as leukodystrophies are characterized by genetic abnormalities resulting in a failure of proper myelin sheath formation (hypomyelination) and/or maintenance (dysmyelination). The mutations leading to hypomyelination target the genes encoding either proteins or enzymes involved in the development or maturation of oligodendrocyte precursors. One could therefore speculate that in some of these diseases introduction of genetically normal myelinating cells could correct the genetic deficit in myelination. Dysmyelination results from abnormal accumulation of metabolites due to enzymatic deficiencies. Therefore, transplanted cells may also serve as a vehicles for the delivery of normal enzymes to the endogenous cells, thus improving the functional deficit.

Multiple sclerosis (MS) is an inflammatory central nervous system (CNS) disorder characterized by disseminated demyelination and axonal loss. Although remyelination occurs during the early stages of the disease, it becomes progressively impaired with time, resulting ultimately in lesions consisting of chronically demyelinated axons. Taken into account that axonal loss is the leading cause of neurological disability characterizing MS patients and that normal myelin is required for axonal integrity, remyelination therapies would appear necessary to ensure the restoration of the appropriate axonal function for myelinogenic cell transplantation has been considered as a potential therapeutic approach. In addition to myelin restoration, further benefits of cell therapy for MS have been suggested since transplanted cells may contribute to axonal protection by suppressing ongoing inflammation.

Demyelination is also a feature of spinal cord injury (SCI). Even after severe insults, surviving axons persist in the subpial rim of white matter but exhibit demyelination and consequently dysfunctional conduction properties. Endogenous remyelination following SCI occurs very slowly, and it is not clear whether it ever reaches completion; thus, myelinogenic cell engraftment has been proposed as a means of myelin restoration and protection of denuded axons against subsequent degeneration. Engineering the transplanted cells to deliver growth factors associated with neural protection and regeneration to the lesion sites is being attempted to optimize the functional benefits of cell therapy for SCI.

History of Cell Transplantation for Central Nervous System Myelin Regeneration

Santiago Ramon y Cajal and Francisco Tello were the first to suggest that CNS axons may be myelinated by exogenous cells since they observed that these axons would enter the segments of the peripheral nerve implanted into the brain and subsequently become myelinated by peripheral myelin-forming cells, the Schwann cells (SCs). More than 50 years later, William Blakemore demonstrated that SC remyelination occurred following placement of peripheral nerve fragments over the areas of demyelination in the spinal cord, and a similar phenomenon was later described by Duncan and colleagues following engraftment of cultured SCs. Madeleine Gumpel and co-workers then showed for the first time that CNS fragment implantation could also lead to myelination by oligodendrocytes in myelin mutants, and this work was followed by experiments in which oligodendrocyte (re)myelination was achieved by transplantation of cultured CNS cells. During the 1990s, olfactory ensheathing cells (OECs) as well as neural and embryonic stem (ES) cells were added to the list of cellular candidates for graft-mediated remyelination therapies. Pluripotent stem cells derived from peripheral tissues have also been suggested as promising novel sources of myelin (Figure 1).

Although much of the work reported in this article is focused on exploring the therapeutic benefit of transplanted cells for myelin repair, these studies also stress the power of cell transplantation to unravel the biology of glial cells under developmental and pathological conditions.

Early transplantation experiments established models of dysmyelination/demyelination that were to be used in subsequent studies. However, several issues need to be considered when interpreting these experiments. First, the CNS has an intrinsic myelin regeneration capacity, and therefore a means of clearly identifying transplanted cells needs to be employed. Second, although certain leukodystrophies are reliably represented by animal models, no currently available models of MS mimic entirely the highly complex pathology of the disease. Also, important differences

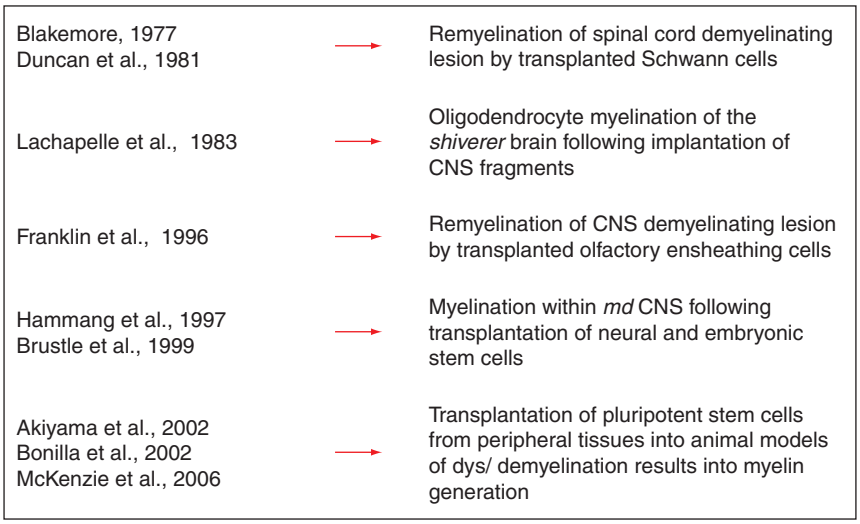

Figure 1 History of experimental engraftment for myelin repair.

between species exist; therefore, validating the results obtained using rodents and other lower species in primate models will be a fundamental step toward developing clinical applications of engraftment for remyelination.

Experimental Models for Studying Graft-Mediated Myelin Generation

These models include the developing CNS, models of hypomyelination/dysmyelination, and models of induced demyelination in the adult CNS. What these have in common is the fact that a substantial proportion of axons are available for the progeny of transplanted cells to myelinate and therefore serve to assay the biology and repair potential of various cell preparations.

Normal Central Nervous System

Developing central nervous system Many studies have used the developing nervous system as a transplantation site for assessing the myelin-generating potential of introduced cells with cells injected into the CNS of embryonic day 19 embryos or newborn rat and mouse pups. These recipients represent a highly permissive environment for myelination because transplanted cells will be responding to developmental cues and therefore do not reflect a chronic disease situation. When using developing animals as recipients, graft-specific markers are essential to identify myelin formed by the transplanted cells.

Although engraftment into the adult forebrain does not offer the possibility to examine the myelinating potential of the grafted populations, cell injections into germinative sites such as the subventricular zone (SVZ) versus nongerminative areas have been performed in order to study the cell fate in terms of survival, proliferation, migration, and differentiation.

Retina Axons of retinal ganglion cells are not myelinated in cats, mice, and rats but can be myelinated in species such as human, dog, rabbit, and chicken. The nonmyelinating retina has been used as a model for graft-mediated myelination. The use of this model has highlighted the fact that CNS axons retain their capacity to become myelinated throughout life, therefore offering a paradigm to study mechanisms governing or promoting myelination. Nevertheless, it does not represent any of the pathological demyelinating or dysmyelinating conditions.

Animal Models of Hypomyelination/Dysmyelination

This group includes animals characterized by (1) aberrant myelin sheath production or maintenance due to a defect in myelin protein-related genes (myelin mutants) or (2) progressive myelin loss resulting from abnormal accumulation of metabolites due to enzymatic deficiencies (models of Krabbe's disease and metachromatic leukodystrophy).

Myelin mutants Since most myelin mutants are short-lived, transplantation is usually performed in developing animals. In the case of mutants with an extended life span (e.g., canine shaking pup), adult animals have been used as recipients. One of the advantages of these models is that myelin produced by the transplanted cells can be readily recognized

because of its normal composition. Myelin mutants are classified into groups according to the myelin protein-related gene affected.

Myelin basic protein mutants. The shiverer (shi) mouse, the most commonly used mutant for transplantation, and the Long–Evans shaker (les) rat lack myelin basic protein (MBP), which leads to uncompacted myelin sheath formation. The utility of these models lies in the easy identification of MBP-positive myelin sheaths when normal tissue is transplanted into the mutant CNS (Figure 2(a)). Shiverer oligodendrocyte processes associate with axons forming thin and uncompacted myelin, which after a few weeks of age may prevent transplant-mediated remyelination. Therefore, to achieve significant myelination by cell engraftment in mature animals, demyelinating lesions must be induced. Backcrossing of shiverer mice to immunodeficient NOD/SCID mice has been used to study the behavior of transplanted human cells since this procedure

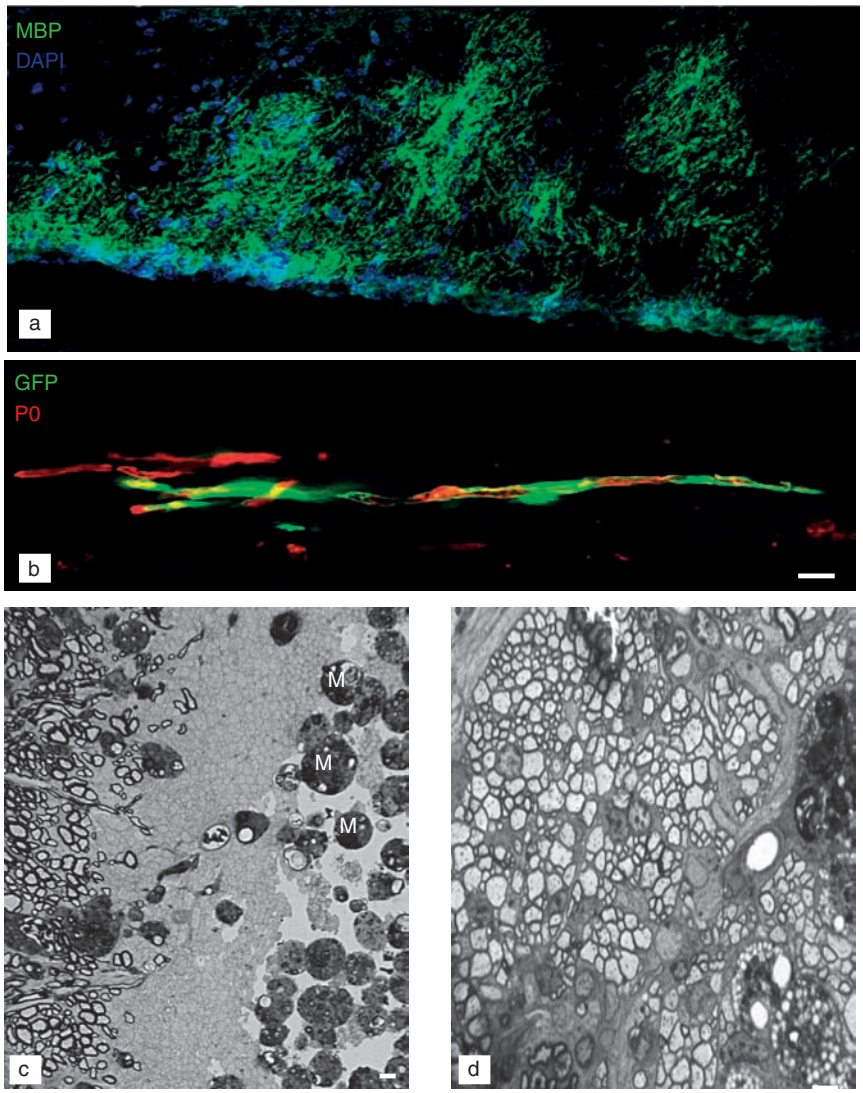

Figure 2 Detection of graft-derived myelin. The utility of myelin mutants as myelinogenic graft recipients is that myelin formed by transplanted wild-type cells can be identified by its normal composition. (a) Myelin basic protein (MBP) positive myelin sheaths within *shiverer* mouse corpus callosum formed by engrafted neural progenitors. Magnification, ×312. Figure kindly provided by Dr. Sandrine Vitry. (b) Graft-specific markers such as green fluorescent protein (GFP) represent an important tool in tracking the fate of transplanted cells. Co-localization of a peripheral myelin marker P0 and GFP reveals myelin formed by GFP-positive Schwann cells transplanted into a demyelinating lesion induced in mouse spinal cord. Figure kindly provided by Dr. Christelle Girard. (c, d) The absence of spontaneous remyelination in the X-EB lesion facilitates the identification of graft-derived myelin. (c) One-micrometer toluidine blue-stained resin section showing demyelinated axons and macrophages (M) within the lesion induced in rat spinal cord white matter. (d) Oligodendrocyte remyelination of the lesion 4 weeks following transplantation of cultured glial cells. Scale bars = 10 μm (b–d). (a) Reproduced from Mouse oligospheres: From pre-progenitors to functional oligodendrocytes, vol. 58, No. 6, 1999, pp. 735–751. Copyright (1999 and Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.); and Reprinted with Permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

overcomes the rejection of xenografted tissue. *Les* oligodendrocytes form poor myelin during the first 2 weeks after birth, and this myelin disappears by 4 weeks of age. These animals suffer from seizures causing spinal fractures, which limits their use for transplantation studies.

Proteolipid protein mutants. Some mutants have spontaneous defects in the proteolipid protein (PLP) encoding gene, mimicking those found in Pelizeus–Merzbacher disease (PMD) patients. This pathology is characterized by severe hypomyelination in the absence of significant inflammation. The animals belonging to this group include the jimpy mouse and its variants, the myelin deficient (md) rat and shaking pup. The jimpy mouse has a base pair substitution in the fourth intron of the PLP gene, leading to the death of most oligodendrocytes with few surviving cells synthesizing a small amount of myelin. Its life span is approximately 4 weeks. The *md* rat shows an almost complete absence of myelin sheaths and has a similar life span (23–25 days) as that of the jimpy mouse. However, a longer-lived strain surviving up to 75–80 days has been discovered. The shaking pup can survive up to 1 year with careful nursing, which is usually very expensive. Other PLP mutants modeling PMD are the rumpshaker (rsh) mouse, a paralytic tremor model in rabbit, and a congenital tremor in pig.

Transgenic animals overexpressing PLP have also been used for studying graft-mediated remyelination. Whereas homozygous mice have severe hypomyelination and a short life span similar to jimpy mice, heterozygous mice have a longer life span and may be used as recipients since they are characterized by normal myelination followed by progressive demyelination starting at approximately 3 or 4 months of age.

Mag-fyn knockout mouse. The null mouse for the myelin-associated glycoprotein, MAG, and the non-receptor tyrosine kinase of the Src family, fyn, is also hypomyelinated. The myelin sheath is abnormal and does not contain MAG. These animals have a normal life span and have been used as transplant recipients as either postnatal or adult animals.

The taiep rat. This rat is characterized by a microtubular defect in oligodendrocytes that results in a disruption of the normal myelination process in certain CNS areas and failure of myelin sheath maintenance elsewhere, leading to progressive demyelination. The pathology is reminiscent of the chronic areas of demyelination found in MS patients. Exploring the potential benefits of cell therapy in such a model is therefore of great interest. These animals survive until 17 months of age.

Dysmyelination resulting from metabolic mutations

Galactosylceramidase mutants. The animals with mutations in galactosylceramidase-encoding genes represent good models for Krabbe's disease or Globoid cell leukodystrophy (GLD). The characteristic pathological features of this disease are infiltration of macrophages containing periodic acid–Schiff (PAS)-positive materials ('globoid cells') in the CNS with concomitant demyelination. The animals belonging to this group are the twitcher mouse, a GLD dog, a saposin A-deficient mouse, a H168C mutated mouse, and a rhesus monkey mutant.

Arylsulfatase-A knockout mouse. The arylsulfatase (ASA) null mouse is a murine model of human metachromatic leukodystrophy. These animals display progressive sulfatide storage within the CNS, mild loss of myelin, and develop behavioral deficits in adulthood mimicking late infantile/early adulthood forms of the disease.

Induction of Demyelination in the Adult Central Nervous System

The models of induced demyelination include focal demyelinating lesions and induction of inflammatory pathologies associated with disseminated demyelination of the adult CNS.

Focal demyelination *Lysolecithin and ethidium bromide*. Focal demyelination areas can be induced in adult animals by injection of gliotoxic agents into predetermined white matter areas specifically killing glial cells but preserving the majority of the axons. The most commonly used toxins are lysophosphatidyl choline or lysolecithin (LPC) and ethidium bromide (EB). Lesions induced by LPC and EB show a limited amount of inflammation, usually associated with myelin debris removal, and have been used to answer basic questions regarding the remyelinating potential of different cell preparations. EB is a DNA intercalating agent which kills astrocytes, oligodendrocytes, and oligodendrocyte precursors. LPC is a membrane-solubilizing agent which has a particular toxicity for myelin but spares some oligodendrocytes and possibly some of their precursors. As a result, demyelination in EB lesions is more extensive and the remyelination process is slower than in the LPC model. Toxin-induced lesions do not result in behavioral deficits unless they are targeted to the cervical spinal cord. The advantage of these focal models is that the predetermined location of demyelinating lesions facilitates cell transplantation experiments. Cells can be placed either directly in a lesion to evaluate the remyelinating capacity of the preparation or at a distance from the lesion to examine migration. Xenograft rejection can be

bypassed, inducing focal lesions in immunocompromised animals such as nude mice or rats.

Focal lesions induced in rodents undergo spontaneous remyelination. In order to create a situation in which endogenous remyelination is absent, William Blakemore's laboratory at Cambridge University developed a model in which local X-irradiation (40 Gy) of the spinal cord prior to or soon after EB injection is used to deplete the oligodendrocyte precursors present in the tissue, abolishing endogenous remyelination of the lesion for at least 6 weeks. Use of this nonrepairing lesion model overcomes the need to employ graft-specific markers to evaluate remyelination potential of the engrafted preparation since any remyelination observed can be attributed to transplanted cells (**Figures 2(c) and 2(d)**).

Focal lesions induced in rodents are small compared to demyelinating lesions in humans. In order to obtain a better understanding of human demyelinating conditions, nonhuman primate models of gliotoxin-induced demyelination have been developed in which lesions are targeted to a variety of CNS structures. These models reveal that LPC injection leads to chronic demyelination in the macaque optic nerve but acute demyelination followed by partial remyelination in the spinal cord (**Figure 3**).

Cuprizone. Cuprizone intoxication of C57 black 6 mice results preferentially in demyelination of the corpus callosum. Although restricted to a specific structure in the brain, cuprizone lesions are far more widespread than the previously mentioned focal lesions. The advantage of this model is the possibility to achieve chronic demyelination by extending the treatment. Although appealing, this model has not been used for cell therapy.

Focal models of inflammatory demyelination. Several focal models of inflammatory demyelination are also available. Injections of antibodies into the oligodendocyte/myelin epitope such as galactocerebroside and myelin oligodendrocyte protein, in combination with complement have been used to induce focal demyelination and these are characterized by more extensive inflammation than the toxin models. Lipopolysaccharide injections into rat spinal cord also result in focal lesions of primary demyelination and infiltration of polymorphonuclear cells. Targeting of experimental autoimmune encephalomyelitis (EAE) lesions to the rat spinal cord or cortex has been achieved. Whereas spinal cord focal EAE lesions are characterized by extensive inflammation even at 1 month postinduction and behavioral deficits, cortical lesions show rapid remyelination. Except for the myelin-complement model, these models have seldom been used for transplantation studies.

Diffuse demyelinating lesions Several models of demyelination exist in which lesions are more diffusely distributed throughout the CNS. These include EAE, demyelination induced by intrathecal injections of cholera toxin B subunit conjugated to saporin, and various virus-induced demyelinating diseases. These models are characterized by more complex pathophysiological features compared to gliotoxin-induced models of demyelination, including clinical deficits. Therefore, whereas gliotoxin models are used to

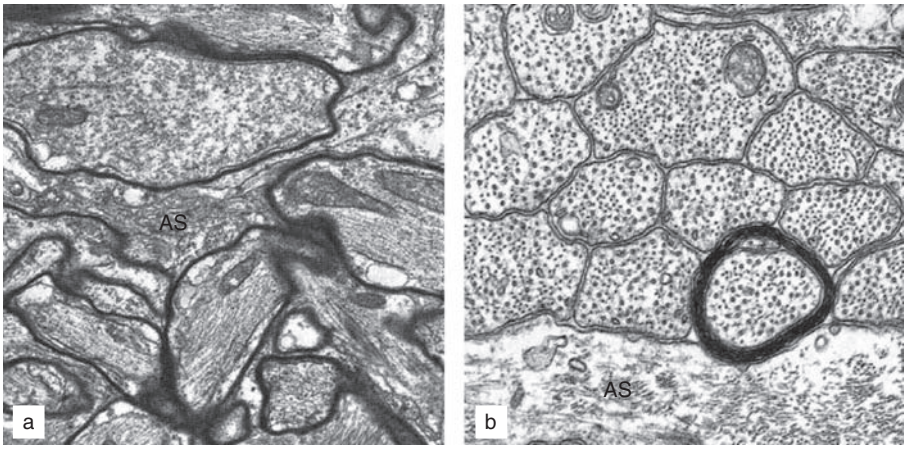

Figure 3 Electron microscopy images of lysophosphatidyl choline (LPC)-induced lesions in the macaque central nervous system. (a) Six weeks following injection of LPC into macaque spinal cord, newly formed myelin sheaths are evident adjacent to the astrocyte processes (As). (b) A bundle of demyelinating axons with only one axon remyelinated in the macaque optic nerve 6 weeks following LPC injection. Magnification: a, ×1300; b, ×14500. Figures kindly provided by Corinne Bachelin. Reproduced from Failure of remyelination in the nonhuman primate optic nerve, vol. 15, No.3, 2005, pp.198–207. Copyright (2005 and Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.); Reprinted with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

answer basic questions regarding the remyelinating potential of transplanted cell preparations, models of diffuse demyelination offer an environment required to address more complex issues of cell therapy, such as the ability of cells to gain access to lesions dispersed throughout the CNS, their survival within an inflammatory environment characterized by recurrent episodes of demyelination, as well as the ability of transplanted preparations to sustain clinical recovery.

Viral-induced models of demyelination. Viral infection has been consistently associated with clinical exacerbations of MS. The two most commonly used models of demyelinating disease are Theiler's murine encephalomyelitis virus and murine hepatitis virus (MHV). Both of these models are characterized by acute inflammatory encephalitis that is followed by chronic CNS demyelinating disease characterized by demyelinating plaques associated with astrogliosis and axonal damage reminiscent of MS lesions. The pathology of virus-induced demyelination is immune mediated. Only MHV has been used as a model for transplant studies.

Experimental autoimmune encephalomyelitis. Demyelination in this model of CNS autoimmune disease is induced by sensitizing animals to oligodendrocyte- and myelin-related molecules by peripheral injections of CNS homogenate, purified myelin, and myelin proteins. The clinical signs and extent of demyelination vary according to the species/strain and the immunization protocol. EAE models can be either monophasic (acute and passive transfer), showing little demyelination, or remitting (chronic relapsing and antibody-mediated) with demyelinating lesions dispersed throughout the CNS, with a prevalence for the spinal cord. These different EAE pathologies have been suggested as good models of T cell/macrophage or antibody-mediated demyelination patterns found in MS. EAE can be induced in several species including monkeys.

Grafting into rodent EAE models has been used mainly to address migration and recruitment of transplanted cells by inflammatory lesions as well as the effect of the graft on the ongoing inflammation.

Spinal Cord Injury Models

As mentioned previously, remyelination of the axons demyelinated following SCI appears to be fundamental to achieve functional recovery. However, in addition to remyelinating the axons in the lesion, transplanted cell preparations should ideally also support CNS axonal elongation and outgrowth. SCI models reflect various degrees of injury, including partial or complete spinal cord transection or contusion and have been used to assess the benefits of cell transplantation on remyelination, axonal regeneration, and clinical recovery.

Cell Preparations That Could Be Used for Remyelination of Demyelinating Central Nervous System Lesions

Myelin formation can be achieved using a variety of cell preparations. However, when choosing suitable candidates for transplantation as a treatment for CNS myelin diseases, factors other than the ability to generate myelin within dysmyelinated/demyelinated CNS or to give rise to progeny that will do so need to be considered. Therefore, ideally the following conditions should be met:

1. Cells should be readily available for transplantation. Although certain myelinogenic cell types can be obtained in a noninvasive manner from adult tissues, others have been isolated mainly from embryonic and fetal CNS; thus, their use might create ethical controversy. The development of techniques for appropriate isolation and expansion of these cells in culture is therefore of great importance.
2. Immunosuppressive protocols should be avoided when possible since these have been shown to interfere with the process of remyelination. Use of autografting (grafting of cells/tissue from one anatomical site to another in the same individual) would represent a transplantation paradigm that does not require immunosuppression. As mentioned previously, noninvasive isolation of several myelinogenic cell types for autologous grafting is possible. In addition, it should also be taken into account that CNS is characterized by exquisite regulation of immune responses and therefore certain cell types may not be targeted by an immune rejection response even if grafted as haplotype-mismatched preparations.
3. Since leukodystrophies are characterized by large areas of dysmyelination and the hallmark of MS is the dissemination of lesions throughout the CNS, transplanted cells should be able to migrate and become established within the CNS tissue.
4. Because axonal loss is the major cause of disability in MS, achieving axonal protection/regeneration is of great clinical interest. Remyelination may reduce axonal loss since axonal integrity is myelin dependent and remyelinated plaques show only a minor reduction in axonal density compared to inactive demyelinated plaques. In addition, neuroprotection may be enhanced by graft-mediated suppression of an ongoing inflammatory process as well as by transplanting cells genetically modified to overexpress neurotrophic molecules.

The following section reviews the cell preparations that have been used to achieve myelination/remyelination and discusses their advantages and disadvantages compared to other cell types.

Cells of Oligodendrocyte Lineage and Their Precursors

Myelination Graft-mediated repair of CNS demyelinating lesions by rodent oligodendrocyte lineage cells and their precursors (ES cells, neural stem cells, and neural precursors) has been extensively studied. Gumpel and colleagues were the first to achieve myelination in the developing CNS of myelin mutants by transplanting fragments containing oligodendrocyte progenitor cells (OPCs) and neural precursors. Subsequent studies showed that myelination could also be achieved by grafting mixed glial cell preparations, purified preparations of different stages of the oligodendrocyte lineage, oligospheres (spheres of OPCs), neurospheres, ES cell-derived glial progenitors, and cells derived from adult rat CNS. Moreover, cryopreservation does not abolish the capacity of oligodendrocyte preparations to myelinate axons. Neural stem or glial cell line grafting also leads to myelin generation.

In addition, the ability to myelinate the hypomyelinated or demyelinated CNS has been shown for cells derived from canine, pig, cat, and human CNS. Human cell transplantation has resulted in variable outcomes. Whereas solid grafts of fresh and frozen human fetal brain and spinal cord tissue (8–20 weeks) have resulted in successful myelination in the neonatal shiverer brain, transplantation of isolated human neural and embryonic stem cells into newborn and embryonic rodent brains has resulted in the generation of scarce populations of oligodendrocytes. Several strategies are currently used to improve oligodendrocyte generation from human tissue. Successful myelin generation has been reported following transplantation of glial-committed progenitors selected from fetal and adult human brains by A2B5 sorting, OPC-enriched preparations obtained by treating ES cells with a sequence of growth factor cocktails, and human neural stem cells overexpressing neurogenin 2. Another strategy of interest may be the overexpression of oligodendrocyte-specific transcription factors Olig1/2, given that such treatment of mouse neural stem cells promotes oligodendrocyte specification.

Intrinsic and extrinsic factors modulate the (re) myelination potential of the transplanted cells. These include the age of the donor tissue, the degree of cell maturation, and the environment in which cells are grafted. Early studies suggested that the extent of myelin generation by engrafted oligodendrocyte lineage cells is inversely related to the maturity of the tissue from which cells originate since myelination is more extensive following engraftment of fetal versus neonatal or adult cells. However, different results were obtained in a study by Steve Goldman laboratory, which observed that human A2B5-sorted cells derived from the adult tissue formed myelin more rapidly and ensheathed more axons per cell than fetal cells after engraftment into shiverer mice. The cell differentiation status is also important; thus, commitment of neural precursors to the oligodendrocyte lineage *in vitro* is required to achieve significant remyelination upon engraftment in an X-EB lesion, and transplantation of multipotential neural precursors results in the generation of astrocytes and sometimes neurons, or nonneural cells in the case of pluripotent ES cell engraftment. However, OPCs are more efficient at myelin formation than preoligodendrocytes and mature oligodendrocytes, suggesting that postmitotic cells have a limited capacity to perform myelin repair.

Myelination by transplanted cells is also dependent on the environment in which they are introduced. Therefore, it is favored in newborn recipients undergoing gliogenesis and even more so when myelination by host cells is defective, as occurs in myelin mutants. In the adult CNS, acute rather than chronic lesions are more likely to be remyelinated by the transplanted populations. Inflammation may play a critical role in this process since inducing inflammation in the noninflammatory environment of the retina or in the dysmyelinated spinal cord of the taiep rat promotes remyelination by transplanted OPCs. In addition, transplantation of CNS precursors into macroglial-free areas of X-EB lesions favors their differentiation into myelin-forming SCs rather than oligodendrocytes, which appears to be due to the presence of bone morphogenetic proteins in the absence of Noggin within these lesions.

Migration The extent of grafted cell migration is dependent on the degree of cell maturation, the host environment, and the site of implantation. Immature cells migrate more extensively through the CNS tissue than the more differentiated cells; thus, when transplanted into developing brain OPCs migrate more extensively than preoligodendrocytes but less than polysialylated neural cell adhesion molecule (PSA-NCAM)-positive pre-OPCs. Transplantation studies also highlight the different migration potential of embryonic, neonatal, and adult tissue-derived cells. Therefore, the human fetal OPCs colonize the neonatal mouse brain more efficiently than the adult OPCs. In addition, whereas rodent and human embryo-derived cells migrate extensively within the adult CNS tissue,

neonatal OPCs do not. Migration is also strongly influenced by the host environment. Generally, it has been observed that the migration in myelin mutants is more extensive than that observed in normal neonatal animals. Also, OPC-depletion of the adult rat spinal cord by X-irradiation renders it permissive for migration of transplanted OPCs. Whether the OPCs grafted away from the areas of demyelination reach the lesions by migrating through the CNS tissue remains a matter of debate.

The site of implantation and the mode of delivery of the transplanted cells appear to be important determinants of their migration. When transplanted into the neonatal shiverer brain parenchyma, OPCs and neural precursors follow the axonal tracts; therefore, multiple injections of OPCs into various sites within the axonal circuitry can lead to massive colonization of the forebrain and spinal cord. Extensive colonization of neonatal and embryonic mouse brains has also been reported after stem cell delivery into the ventricle or the cisterna magna. When grafted into adult forebrain, SVZ-derived neural progenitors migrate along specific routes, such as the rostral migratory stream or the cingulum. Inflammation appears to be a major stimulus of grafted cell migration/recruitment since during EAE, CG4 cells or neonatal neural precursors (NPCs) transplanted into the adult CNS extensively colonize adult recipient tissue, and intravenously injected adult NPCs have been reported to cross the blood–brain barrier. Widespread migration has also been observed following transplantation of glial-committed precursors into the MHV-induced model of demyelinating disease, and although transplanted NPCs did not survive well upon engraftment into the inflammatory environment of the twitcher mouse brain, they extensively migrated toward sites of microglial activation.

Immunomodulation Several studies suggest that the recruitment of transplanted neural precursors into inflammatory EAE lesions attenuates the clinical course of the disease, at least partially due to a reduction of ongoing inflammation. The decreased inflammation has been attributed to an NPC-induced apoptosis of infiltrating immune cells mediated by Fas–Fas ligand interactions as well as to the inhibition of T cell activation in the lymph nodes. However, studies using a viral-induced model of inflammatory demyelination show that glial–precursor engraftment does not modify the inflammatory component of the disease. Although it is possible that the anti-inflammatory effect observed in EAE models is dependent on the maturity of the cell preparation engrafted or on the model itself, further studies are required to confirm these observations.

Rejection As techniques currently employed to isolate oligodendrocyte lineage cells are incompatible with autografting, graft rejection represents a major issue relevant to the applications of oligodendrocyte lineage cell therapy. The exact haplotype matching of the host–donor tissue is extremely difficult due to the polymorphisms of human genes, which means that oligodendrocyte lineage cell grafting would most probably imply the use of mismatched tissue. Only a few studies have examined the persistence of myelination by genetically mismatched oligodendrocytes. Whereas transplantation of allogeneic glial cells into neonatal MD rats resulted in rejection, transplant-mediated myelination (without donor–host tissue matching) in a shaking pup model persisted for 28 weeks in the absence of immunosuppression. In addition, fully mismatched preparations enriched for OPCs and transplanted into a toxin-induced rat spinal cord demyelinating lesion successfully remyelinated the demyelinated axons without eliciting an immune response for at least 2 months postengraftment. Although these results, in addition to low major histocompatibility expression by OPCs, suggest that these cells may be of low immunogenicity, long-term survival and remyelination by allogenic cells in inflammatory models of demyelination must be examined.

Schwann Cells

As mentioned previously, experiments examining SC myelination of CNS axons pioneered the area of transplantation for CNS myelination/remyelination. The fact that peripheral myelin is mostly unaffected in MS suggests that the use of SCs for remyelination of demyelinating MS lesions may result in resistance of newly formed myelin to future immune-mediated damage.

Remyelination by grafted SCs has been achieved in a variety of CNS structures, such as the spinal cord, cerebellar white matter, cortex, and optic disk, and in a number of species, such as mice, rats, cats, and monkeys. Myelin formed by SCs transplanted into the demyelinated spinal cord results in improved conduction of demyelinated axons. The extent of remyelination by SC preparations corresponds to the number of SCs introduced and their purity. A major advantage of SCs compared to OPCs as candidates for transplantation in CNS myelin lesions is that it is possible to obtain substantial numbers of these cells from the adult human and monkey peripheral nerve biopsies. After showing that cryopreservation and *in vitro* expansion of primate SCs do not interfere with the ability of these cells to myelinate demyelinated rodent spinal cord, successful autograft-mediated

remyelination of the LPC-induced lesions in adult monkey spinal cord was demonstrated, suggesting the feasibility of such therapy for human disease. In addition, the ability of SCs to sustain axonal regrowth within the CNS has been widely documented, for they have been considered as important candidates for treatment of SCI.

However, whereas migration of SCs transplanted into the CNS can occur along the meninges and blood vessels, it does not occur in the normal white matter, and this may be at least in part due to the inability of these cells to enter the astrocyte-containing environment. Strategies to overcome this difficulty are under investigation and are designed to target specific molecular interactions that impair SC entry into astrocyte-containing areas and/or migration. Thus, it has been reported that reducing levels of chondroitin sulfate proteoglycans within SC/astrocyte confrontation cultures allows SCs to penetrate deeply into the astrocyte territory and that suppressing the expression of N-cadherin, a molecule mediating the adhesion of SCs to astrocytes, results in increased migration of SCs over the astrocyte layers *in vitro*. In addition, genetic modification of SCs to express PSA-NCAM, an important molecular regulator of CNS precursor cell migration, also results in enhanced migration without affecting the myelinating potential when these cells are placed over postnatal brain slice cultures. Although it is imperative to test these results in models of demyelination, they appear encouraging for the development of strategies aiming to enhance SC migration within the CNS.

Olfactory Ensheathing Cells

Another cell type that has been considered for remyelinating cell therapies in addition to the treatment of spinal cord injuries is the OEC. These cells have been isolated from the olfactory nerve, the outer layer of the olfactory bulb of neonatal and adult rats, and lamina propria. OECs have also been obtained from human tissue, including postmortem tissue. Clinical studies have reported no complications after engraftment of autologous OECs into humans with paraplegia for at least 1 year posttransplantation and functional benefits following autografting of cells isolated from the olfactory mucosa in spinal cord injury patients.

Transplantation of OECs into toxin-induced demyelinating spinal cord lesions results in the formation of peripheral-like myelin, which leads to the molecular reconstruction of nodes of Ranvier and restoration of impulse conduction. Transplantation of adult OECs into the primate spinal cord also results in remyelination. Therefore, in addition to SCs, OECs are another candidate cell type for autograft-based remyelination therapies. The main advantage of these cells compared to SCs is that they successfully interact with astrocytes *in vitro* and they also generate less astrocytic response than SCs *in vivo* following implantation into normal white matter.

However, several issues remain unclear regarding the use of OECs for transplantation-based remyelination therapies. For example, engraftment of pure OEC preparations results in less extensive remyelination than if meningeal cell-containing OEC preparations are transplanted, suggesting a need for a combinatorial approach. In addition, some researchers have questioned the ability of OECs to directly participate in remyelination of the demyelinated spinal cord, suggesting that (1) OECs actually provide tunnel-like structures allowing SCs to enter the lesion from the periphery and remyelinate the denuded axons and (2) SC contamination within the OEC cultures may be the crucial factor leading to remyelination. Therefore, further investigation is needed to establish the exact myelination potential of OECs, which will require the development of cell-specific markers allowing the distinction between these cells and the SCs.

Pluripotent Cells from Nonneural Tissues

Transplantation of pluripotent cells from outside the CNS has been attempted in order to achieve remyelination. Encouraging results have been obtained using bone marrow-derived cells; thus, oligodendrocyte generation has been observed following intraparenchymal transplant of CD117-positive cells into neonate mouse and bone marrow-derived hematopoietic precursors into adult brain and spinal cord injury models. However, these results are preliminary and must be considered with caution since several reports have questioned the ability of hematopoietic stem cells to generate neural cells or oligodendrocytes *in vivo*. In addition, although oligodendrocyte differentiation of umbilical cord blood cells has been reported *in vitro*, this has not been reproduced *in vivo*.

It has also been suggested that SCs may be generated from stem cells derived from peripheral tissues. Direct injection of freshly isolated rat bone marrow or purified mouse bone marrow stromal cells into the demyelinated rat spinal cord, as well as the intravenous introduction of acutely isolated bone marrow cell fraction, resulted in SC remyelination of the lesion. However, significantly more cells were required to achieve remyelination following intravenous than direct intraparenchymal injections. Transplantation of skin-derived precursors into dysmyelinated neonatal shiverer brain also resulted in SC-like myelination, and similar differentiation and myelination were

reported after transplantation of transdifferentiated mesenchymal stem cells into the demyelinated peripheral nerve.

Although pluripotent stem cells from outside the CNS might be attractive novel sources of myelin-forming cells, further studies are required to establish their exact behavior after transplantation in acute and chronic models of demyelination.

Contribution of Transplant-Mediated Remyelination to Clinical Recovery

Whether transplantation of myelin-forming cells can correct the functional deficit associated with CNS demyelination is a fundamental question in cell therapy. It has been shown that conduction block of demyelinated axons can be overcome by engrafted oligodendrocytes, SC, and olfactory ensheathing cells-mediated remyelination. For adult primate SCs, the capability to restore conduction is not affected by cryopreservation.

Functional data on behavioral recovery following graft-mediated remyelination are limited. It has been shown that EB-induced demyelination of cervical corticospinal tracts was associated with measurable motor deficits that were ameliorated by spontaneous remyelination. Suppression of endogenous remyelination in this model by X-irradiation resulted in persistent motor deficits that could be overcome by glial cell transplantation given that extensive axonal degeneration had not occurred. Transplantation of primate SCs into an LPC-induced lesion in nude mouse spinal cord was also reported to contribute to functional recovery. In the mildly contused rat thoracic spinal cord, transplanted SCs sustain recovery better than OECs, possibly due to their ability to remyelinate the demyelinated axons more efficiently. Hans Keirstead's laboratory has reported enhanced remyelination and locomotor recovery after SCI following engraftment of human ES-derived OPCs, but such recovery may have resulted from the trophic support to the endogenous cells.

Use of Cells as Vehicles

In addition to remyelinating the demyelinated axons, transplanted cells could also be used as vehicles for delivery of normal enzymes to correct enzymatic deficiencies (storage disorders) or neurotrophic substances to enhance neuroprotection and axonal regeneration (MS and spinal cord injury). Functional improvement following engraftment of stem/progenitor cells overexpressing the normal enzyme has been observed in several models of leukodystrophies. Thus, following engraftment of galactocerebrosidase (GALC)-overexpressing neural stem cells into a twitcher mouse (model of Krabbe's disease), GALC activity was reported within the areas where transplanted cells had migrated. Reduction of sulfatide storage and/or functional improvement have been reported in neonate ASA null mice (model of metachromatic leukodystrophy) following transplantation of normal OPCs and ES cell-derived glial progenitors engineered to overexpress human arylsulfatase A.

Vehicle strategy has also been explored in models of SCI using both neural stem cells and SCs. Enhanced oligodendrocyte differentiation and functional improvement were observed after transplantation of glial-restricted progenitor cells overexpressing D15 multineurotrophin with brain-derived neurotrophic factor (BDNF) and neurotrophin-3 (NT-3) activities. Nerve growth factor-transduced SCs myelinate spinal cord axons and show enhanced capacity to support axonal growth. In addition, engraftment of BDNF and NT3-transduced primate SCs into a demyelinating lesion induced in mouse spinal cord led to a functional improvement associated with a reduction in astrogliosis, enhancement of overall remyelination of the lesion (graft-mediated and endogenous), and axonal protection.

Conclusions

This article reviewed the current knowledge regarding the therapeutic potential of cell engraftment for demyelinating disease. It is clear that transplantation of myelin-forming cells into animal models of demyelination, including nonhuman primates, indeed restores myelin and, in some instances, contributes to functional recovery, thus representing a promising approach for treatment of human disease. While strategies are being developed to overcome the disadvantages associated with oligodendrocyte lineage and SC grafting, it is becoming clear that once these issues are resolved, the most likely targets for transplantation therapies will probably be the noninflammatory leukodystrophies, such as PMD. Multiple sclerosis represents a far more difficult target due to its heterogeneity and the fact that the exact cause and the sequence of pathogenic events are not entirely understood. However, any strategies designed to overcome the neurologic deficit present in MS patients should necessarily include remyelination therapies, with the most efficient one to date being transplantation of myelin-forming cells. The development of animal models reliably representing the

Table 1 Candidate cell types for engraftment-based remyelination therapies[a]

Cell type	*Main advantage*	*Limiting factor*	*Strategies under investigation*	*Target demyelinating pathologies*
Oligodendrocyte lineage cells and their precursors	Extensive remyelination and integration into host CNS tissue. Migration under permissive conditions (development and pathology)	Noninvasive isolation currently not possible. Difficulties committing human stem cells to oligodendrocyte lineage in culture	Development of reproducible epigenetic-stimulation protocols. Overexpression of transcription factors governing oligodendrocyte specification in neural/embryonic stem cells	Leukodystrophies Multiple sclerosis Spinal cord injury
Schwann cells	Extensive remyelination. Support axonal regeneration. Candidates for autologous grafting	Do not migrate within the CNS. Do not integrate into astrocyte-containing environment	Targeting of molecular interactions that inhibit migration and integration into astrocyte-containing environment	Multiple sclerosis Spinal cord injury
Olfactory ensheathing cells	Candidates for autologous grafting. Better integration into astrocyte-containing environment than Schwann cells	Role in CNS repair (remyelination and axonal regeneration) recently questioned; more data required	Development of reliable markers for distinction between olfactory ensheathing and Schwann cells	Multiple sclerosis Spinal cord injury
Pluripotent cells from nonneural tissues	Potential sources for noninvasive isolation of myelinating cell precursors	Insufficient data, conflicting on myelinating potential *in vivo* (except bone marrow)	Development of protocols for differentiation into myelinating cell phenotypes	Leukodystrophies? Multiple sclerosis? Spinal cord injury?

[a]Although the remyelinating potential of oligodendrocyte lineage and Schwann cells has been firmly established, that of olfactory ensheathing cells is not entirely clear, and insufficient data are available regarding pluripotent cells from nonneural tissues.

size of MS lesions and their chronicity will therefore be fundamental to predicting the outcome of potential cell therapies.

See also: Cell Replacement Therapy: Mechanisms of Functional Recovery; Demyelinating Diseases; Demyelination and Demyelinating Antibodies; Myelin: Molecular Architecture of CNS and PNS Myelin Sheath; Olfactory Ensheathing Cells: Regeneration Following Spinal Cord Injury; Oligodendrocyte Morphology; Oligodendrocyte and Schwann Cell Identification Methods; Schwann Cell Development; Schwann Cell Morphology; Spinal Cord Regeneration after Injury: Bridging by Schwann Cells and Bioimplants; Spinal Cord Regeneration and Functional Recovery: Strategies.

Further Reading

Akiyama Y, Radtke C, and Kocsis JD (2002) Remyelination of the rat spinal cord by transplantation of identified bone marrow stromal cells. *Journal of Neuroscience* 22: 6623–6630.

Bachelin C, Lachapelle F, Girard C, et al. (2005) Efficient myelin repair in the macaque spinal cord by autologous grafts of Schwann cells. *Brain* 128: 540–549.

Baron-Van Evercooren A and Blakemore WF (2004) Remyelination through engraftment. In: Lazzarini RA (ed.) *Myelin Biology and Disorders,* vol. 1, pp. 143–172. New York: Elsevier.

Blakemore WF (1977) Remyelination of CNS axons by Schwann cells transplanted from the sciatic nerve. *Nature* 266: 68–69.

Bonilla S, Alarcon P, Villaverde R, Aparicio P, Silva A, and Martinez S (2002) Haematopoietic progenitor cells from adult bone marrow differentiate into cells that express oligodendroglial antigens in the neonatal mouse brain. *European Journal of Neuroscience* 15: 575–582.

Brustle O, Jones KN, Learish RD, et al. (1999) Embryonic stem cell-derived glial precursors: A source of myelinating transplants. *Science* 285: 754–756.

Duncan ID (2005) Oligodendrocytes and stem cell transplantation: Their potential in the treatment of leukoencephalopathies. *Journal of Inherited Metabolic Disease* 28: 357–368.

Duncan ID, Aguayo AJ, Bunge RP, and Wood PM (1981) Transplantation of rat Schwann cells grown in tissue culture into the mouse spinal cord. *Journal of the Neurological Sciences* 49: 241–252.

Einstein O, Fainstein N, Vaknin I, et al. (2006) Neural precursors attenuate autoimmune encephalomyelitis by peripheral immunosuppression. *Annals of Neurology* 61(3): 209–218.

Foote AK and Blakemore WF (2005) Inflammation stimulates remyelination in areas of chronic demyelination. *Brain* 128: 528–539.

Franklin RJ, Gilson JM, Franceschini IA, and Barnett SC (1996) Schwann cell-like myelination following transplantation of an olfactory bulb-ensheathing cell line into areas of demyelination in the adult CNS. *Glia* 17: 217–224.

Goldman SA and Windrem MS (2006) Cell replacement therapy in neurological disease. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 361: 1463–1475.

Hammang JP, Archer DR, and Duncan ID (1997) Myelination following transplantation of EGF-responsive neural stem cells

into a myelin-deficient environment. *Experimental Neurology* 147: 84–95.

Lachapelle F, Gumpel M, Baulac M, Jacque C, Duc P, and Baumann N (1983) Transplantation of CNS fragments into the brain of shiverer mutant mice: Extensive myelination by implanted oligodendrocytes: I. Immunohistochemical studies. *Developmental Neuroscience* 6: 325–334.

McKenzie IA, Biernaskie J, Toma JG, Midha R, and Miller FD (2006) Skin-derived precursors generate myelinating Schwann cells for the injured and dysmyelinated nervous system. *Journal of Neuroscience* 26: 6651–6660.

Pluchino S, Zanotti L, Rossi B, et al. (2005) Neurosphere-derived multipotent precursors promote neuroprotection by an immunomodulatory mechanism. *Nature* 436: 266–271.

Transport Dependent Damage Signaling

V Cavalli, Washington University in St. Louis, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Central versus Peripheral Nervous System Regeneration

While most tissues in the human body, including peripheral nerves, have a remarkable ability to repair themselves after injury, the central nervous system (CNS), comprising the brain and the spinal cord, has a very poor intrinsic capacity for repair. In adults, most parts of the CNS cannot regenerate axons, which explains the devastating condition of paralysis following spinal-cord injury. Axonal damage may also occur in neurodegenerative processes, such as in Alzheimer's, Parkinson's, and Huntington's diseases, whereby axonal pathologies and degeneration often precede death of neuronal cells.

Research over the past 20 years has identified two major obstacles to mature CNS regeneration. The first one is the expression by the surrounding oligodendrocytes and the glial scar of inhibitory molecules such as Nogo, myelin-associated glycoprotein (MAG), oligodendrocyte-myelin glycoprotein (OMgp), and chondroitin sulfate proteoglycans (CSPGs), which can directly inhibit axonal growth. The second obstacle is that axotomized CNS neurons often fail to activate a program of gene expression that is required to sustain regeneration. The outcome of CNS neuron regeneration is thus likely to result from both the intrinsic growth ability of the injured neuron and the strength of the inhibitory environment.

The competence of CNS neurons to regenerate axons in the presence of a permissive environment will depend in part upon their ability to increase their intrinsic growth ability. Identifying the mechanisms activating the intrinsic growth ability of a neuron – that is, its regeneration program – will be a key step toward improving our ability to restore nerve function following damage. One way to obtain insight into why a given neuron fails to regenerate is to study neurons that can regenerate. Adult olfactory and hippocampal dentate gyrus neurons or developing immature CNS neurons might represent one such category, since they show some level of regeneration. However, due to their larger size, neurons within peripheral nerves offer better access to biochemical approaches. Investigating the molecular mechanisms regulating axonal regeneration in peripheral nervous system (PNS) neurons is therefore crucial to our understanding of why most axons fail to regenerate in the CNS after injury.

Injured PNS neurons undergo a typical reaction, required for axonal regeneration, which includes Wallerian degeneration in the distal portion of the nerve, sprouting of processes in the proximal portion of the nerve, and strong morphological changes in the soma, referred to as the cell body reaction. Another characteristic response is the expression of regenerative-associated genes (RAGs). RAGs identified thus far include transcription and translation factors, cytoskeletal proteins, axonal guidance molecules, trophic factor and cytokines (and their receptors), neuropeptides, neurotransmitter-synthesizing enzymes, ion transporters, cell adhesion and cell-signaling molecules, vesicle-trafficking components, metabolic proteins, and membrane transporters. In marked contrast, in CNS neurons, alteration of gene transcription following injury is weaker and transient, or totally absent.

While it is well established that alteration of gene expression is required for the regenerative competence of PNS neurons, very little is known about how information from the site of damage is communicated to the cell body to initiate these changes. This is a challenging problem, since the site of lesion is often distant from the cell body; the axon of larger vertebrates can reach lengths of up to 1 m.

Evidence for Damage Signaling in the PNS: The Conditioning Injury Paradigm

The key role of RAG expression in regeneration has been elegantly demonstrated in dorsal root ganglia (DRG) sensory neurons, which are unique in having both a long CNS axon that ascends the spinal cord and a second axon branch that projects through a peripheral nerve. 'Conditioning' nerve lesion studies have demonstrated that peripheral axotomy of DRG neurons, but not central axotomy, generates an enhanced growth state that allows a certain degree of regeneration within the CNS. This enhanced growth state is likely to rely upon a program of gene expression uniquely induced by peripheral axotomy to support axon extension. These studies underscore the existence of distinct signaling mechanisms in central versus peripheral branches, which culminate in the distinct abilities of DRG neurons to respond to central versus peripheral injury.

One approach to understanding the molecular determinants of axonal regeneration is to further characterize the target genes activated by peripheral injury. Another important approach is to identify the

nature of the signals that inform the cell body that the axon has been injured and that initiate the appropriate transcriptional program.

Damage Signaling: The Hypotheses

Despite a century of research on the reaction of neurons to axotomy, the identities of the signals that inform the cell body that the axon has been injured are incompletely understood. A series of elegant studies in the early 1990s in the mollusk *Aplysia californica* provided evidence for the existence of multiple signals functioning in a temporal sequence.

First, action potentials propagating from the injury site to the cell body activate a signaling cascade in the cell soma. Supporting this idea, a limited number of studies in both *A. californica* neurons and mammalian neurons indicate that membrane depolarization is one of the earliest events detectable after axotomy. A positive role for membrane depolarization in nerve regeneration is also suggested by the observation that depolarization triggers the expression of immediate-early genes and survival of neurons in the absence of trophic support.

Second are the so-called negative injury signals, in which axotomy prevents molecules that are normally transported retrogradely from reaching the cell body. These negative injury signals function in repressing certain transcriptional programs. Support for this mechanism came from studies in which blocking axonal transport led to the expression of certain molecules in the cell body, suggesting that their repression relies on the constitutive retrograde transport of a signal.

Third are the positive injury signals, which are activated at the injury site and then transported retrogradely back to the cell body. Early morphological studies have shown that the more distal the injury, the longer it takes the neuron to exhibit features of the cell body reaction, which suggests that a signal is traveling from the injury site back to the cell body. More recent studies have confirmed these observations at the molecular level. This positive injury signaling system relies upon a microtubule-based retrograde transport to communicate injury to the distantly located cell body.

Molecular motor proteins such as kinesins and dyneins mediate the intracellular transport of protein complexes, vesicles, and other organelles by translocating along polarized microtubule tracks – the axonal highway system. Kinesins and dynein define a superfamily of motor proteins that use the energy of adenosine triphosphate (ATP) hydrolysis to move cargo along microtubules. Kinesin mediates anterograde axonal transport, while dynein, together with its associated protein complex dynactin, mediates retrograde axonal transport. This microtubule-dependent axonal transport occurs at rates of 1–5 $\mu m\,s^{-1}$. While the notion that microtubule-based transport played a role in conveying information from the injured axon back to the cell body emerged in the early 1970s, the involvement of the molecular motors kinesin and dynein/dynactin has only been recently demonstrated.

Recent Advances toward Identification of Positive Injury Signals

Mitogen-Activated Protein Kinase Signaling Modules

The notion that signaling molecules undergo fast axonal transport has emerged in the past decade. Sciatic nerve ligation experiments have shown that a subset of the signaling components in the mitogen-activated protein kinase (MAPK) pathway, which includes phosphatidylinositol 3-kinase (PI3K), extracellular signal-regulated kinase (ERK-1/2), p38 MAPK, and c-Jun N-terminal kinase (JNK), is transported along axons. The functional role for transporting signaling molecules along the axon remains elusive, but recent observations suggest that they might be part of a damage surveillance system.

JNK signaling plays a positive role in peripheral neuron survival and repair. Following axonal injury, JNK activation and expression of the downstream transcription factor c-Jun in the cell body persist until target reinnervation has occurred. One of the positive damage-signaling molecules leading to JNK activation in the cell body may be JNK itself. Following sciatic nerve injury, JNK is activated locally in the axon at the injury site, where it interacts with the JNK scaffolding protein syd (also known as JIP3 or JSAP1) and the retrograde motor complex dynein/dynactin. The net retrograde transport of a syd/JNK complex observed following injury suggests that mobile axonal syd/JNK complex may generate a transport-dependent axonal damage surveillance system. This study provides one of the earliest pieces of evidence implicating microtubule-dependent transport and molecular motors in damage signaling.

The upstream activating signaling cascade leading to JNK activation may play an important role in the final outcome of the neuron following injury. For example, in a mouse model of type 1 diabetes, activation of JNK along the length of the axon, probably as a result of hyperglycemic stress of the peripheral nerve, may activate retrograde signaling for nerve degeneration. In the case of mechanical damage occurring upon nerve injury, a retrograde signal involving JNK may support nerve survival and regeneration.

Another member of the MAPK family, ERK-1/2, is transported along axons in both directions in rat PNS neurons. However, it only recently became apparent that ERK-1/2 is activated locally in response to sciatic nerve injury, and that it then forms a complex with β-importin, a member of the nuclear import machinery. The ERK/β-importin complex recruits the motor dynein for retrograde transport, suggesting that ERK-1/2 is part of a positive injury signal leading to regeneration of DRG neurons. Other studies have demonstrated that the ERK-1/2 pathway plays a key role in the survival of axotomized mammalian retinal ganglion cells (RGCs) *in vivo*, but that activation of other signaling components is required for axon regeneration in the growth-inhibitory CNS environment.

Another MAP kinase member, the *Aplysia* protein kinase G (ApPKG), is activated at axonal injury sites and retrogradely transported, but the mechanism responsible for injury-induced retrograde transport is still unknown. Further studies of the MAPK signaling cascades activated by injury in peripheral nerves will be necessary to elucidate their role in nerve regeneration.

Retrograde Signaling Dependent upon Nuclear Localization Signals

The targets of positive injury signals in the cell body of an injured neuron are likely to be transcription factors controlling the expression of regeneration-associated genes. Over the past few years, accumulating evidence indicates that the transcription factors can act as positive injury signals. Studies in *A. californica* have shown that nuclear localization signals (NLSs) can target retrograde transport and nuclear uptake of heterologous proteins.

The transcription factor nuclear factor-kappa B (NF-κB) is also retrogradely transported in axons. Different stimulation conditions, including depolarization, result in the redistribution of NF-κB from neurites to nucleus. This retrograde transport is dependent upon an intact NLS sequence and exhibits velocities that are consistent with fast retrograde transport. In adult sensory neurons, blocking endogenous nerve growth factor (NGF), using an anti-NGF antibody, leads to elevated retrograde axonal transport of activating transcription factor 2 (ATF2). This study indicates that lack of neurotrophic support leads to modulation of levels and activation status of a retrogradely transported transcription factor.

Another transcription factor playing a role in injury signaling is the signal transducer and activator of transcription (STAT3). Nerve injury results in STAT3 activation in axons and a wave of activated STAT3 moves along the nerve toward the cell bodies in the DRG. The observation that injury-induced activation of STAT3 occurs in regenerating neurons but not in non-regenerating neurons supports the model that STAT3 is a retrograde signal transcription factor playing a positive role in regeneration. Further support came more recently with the observation that in DRG neurons, peripheral but not central branch injury results in a time-dependent phosphorylation and activation of STAT3. STAT3 activation is thus necessary for increased growth ability of DRG neurons and improved axonal regeneration after a conditioning injury. Activation of STAT3 at the axonal injury site may result from cytokine signaling, such as ciliary neutrophic factor (CNTF) and leukemia inhibitory factor (LIF), which are released at the lesion site and are important for the survival and regeneration of axotomized motoneurons. The proinflammatory cytokine tumor necrosis factor-α (TNF-α) also increases at the injury site and undergoes retrograde axonal transport from the injury site to the DRG and further to the spinal cord.

Much like NLS signaling, sumoylation might serve as a new mechanism to target axonal proteins for retrograde transport following injury. The covalent linking of small ubiquitin-like modifiers (SUMO-1) is biochemically similar to ubiquitination but functionally distinct. Since sumoylation can affect function or subcellular localization, and most SUMO-modified proteins are found in the nucleus, it is therefore tempting to speculate that sumoylation in distal axons might be used in addition to NLS to target axonal proteins to the nucleus for exerting control on transcription.

The nuclear functions of transcription factors and NLS proteins has led to the assumption that their localization within cells should be primarily perinuclear. However, the recent discoveries that transcription factors and proteins participating in nuclear import mechanisms are found along axons opens new lines of investigation on their role in damage signaling.

Axonal Protein Synthesis

While protein synthesis in dendrites is an accepted mechanism involved in synaptic plasticity, whether or not protein synthesis occurs in axons has remained a controversial subject for many years. A small number of studies in the 1960s reported the presence of ribonucleic acid (RNA) and active protein synthesis machinery in invertebrate and vertebrate axons. Work over the past 10 years has now confirmed these earlier findings, and the notion that axons are able to locally synthesize their own proteins is now generally accepted. The functional role for axonal protein synthesis has remained elusive, but recent studies have shown that axonal messenger RNA (mRNA) translation plays a role in axonal growth

during development, allowing developing axons to respond to their complex environment.

Axonal mRNA translation is also used by mature neurons as a mechanism to generate a retrograde damage signal. Following sciatic nerve injury in mice, β-importin protein levels are increased by local translation of axonal mRNA, leading to the formation of an NLS binding complex traveling retrogradely. The intermediate filament protein vimentin is also synthesized from preexisting axonal mRNA following injury and associates with the activated MAP kinase ERK, β-importin, and dynein for retrograde transport.

The list of proteins synthesized in vertebrate axons following injury is increasing, mainly thanks to approaches that include high throughput proteomic methodologies. Ultimately, detailed mechanistic studies will be required to understand the molecular events regulating retrograde transport of these newly synthesized proteins and to identify their downstream target(s) in the cell body.

Protein Degradation

Local axonal protein degradation and synthesis are likely to play an important role in controlling how axons respond to injury. Calcium entry into a damaged axon has been shown to be essential for the formation of a new growth cone in *Aplysia* and mammalian neurons. Proteolytic enzymes such as calpains are activated by calcium, and local degradation of proteins is required for the remodeling of a severed axon into a new growth cone.

Activation of proteolysis may also be required to trigger damage signals. In support of this hypothesis, early studies revealed that inhibition of proteolysis in transected rat sciatic nerves inhibits retrograde axonal transport of organelles from the injury site. More recent studies suggest that the protein degradation by the proteasome system may regulate signaling pathways that control axonal survival and degeneration.

Conclusions and Perspectives

While recent studies in mammals have confirmed the existence of several distinct positive injury signals, the role of negative injury signals and membrane depolarization remains poorly documented. Overall, the emerging picture is that membrane depolarizations along with positive and negative injury signals operate together to different extents in different situations. Coordination between several injury signals might orchestrate the precise program ensuring nerve regeneration or degeneration. Several classes of injury signals may coexist to ensure precise information on the nature of the damage and its distance from the cell body. It is tempting to speculate that the difference in time between the arrival at the cell body of the action potential, the first injury signal, and the later arrival of a positive injury signal might serve as an indicator of the distance of the injury from the cell body. Similarly, the combination of several positive injury signals might serve as an indicator of the extent of damage.

The list of positive injury signals will increase and approaches such as differential proteomics will be of great help in revealing the identity of retrogradely transported molecules after nerve injury. Identifying the molecular mechanisms by which these positive injury signals are transported to the cell body to initiate the cascade of events leading to an enhanced regeneration state will be necessary to improve our ability to restore nerve function following injury.

Ultimately, a direct comparison between injury signaling mechanisms in CNS and PNS neurons might shed light on the poor regenerative capacity of CNS neurons. This will be most important, since, in addition to traumatic axonal damage resulting from spinal cord injury, axonal damage can also occur in neurodegenerative diseases in which axonal pathologies interrupt the cell body/synapse connection.

Successful axonal regeneration may require that the intrinsic growth state of the neuron be activated and that the inhibitory signals from the environment are neutralized. Identification of the pathways that regulate axonal survival and regeneration will provide possible target sites for pharmacological treatments of neurodegenerative diseases and traumatic injury.

See also: Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal Injury: Neuronal Responses; Axonal Regeneration: Role of Growth and Guidance Cues; Microtubule Associated Proteins in Neurons; Microtubules: Organization and Function in Neurons; Peripheral Nerve Regeneration: An Overview; Spinal Cord Regeneration and Functional Recovery: Strategies.

Further Reading

Ambron RT and Walters ET (1996) Priming events and retrograde injury signals. A new perspective on the cellular and molecular biology of nerve regeneration. *Molecular Neurobiology* 13: 61–79.

Broude E, McAtee M, Kelley MS, et al. (1997) c-Jun expression in adult rat dorsal root ganglion neurons: Differential response after central or peripheral axotomy. *Experimental Neurology* 148: 367–377.

Campbell DS and Holt CE (2001) Chemotropic responses of retinal growth cones mediated by rapid local protein synthesis and degradation. *Neuron* 32: 1013–1026.

Cavalli V, Kujala P, Klumperman J, et al. (2005) Sunday Driver links axonal transport to damage signaling. *Journal of Cell Biology* 168: 775–787.

Hanz S and Fainzilber M (2004) Integration of retrograde axonal and nuclear transport mechanisms in neurons: Implications for therapeutics. *Neuroscientist* 10: 404–448.

Hanz S, Perlson E, Willis D, et al. (2003) Axoplasmic importins enable retrograde injury signaling in lesioned nerve. *Neuron* 40: 1095–1104.

Lin H, Bao J, Sung YJ, et al. (2003) Rapid electrical and delayed molecular signals regulate the serum response element after nerve injury: Convergence of injury and learning signals. *Journal of Neurobiology* 57: 204–220.

Mandolesi G, Madeddu F, Bozzi Y, et al. (2004) Acute physiological response of mammalian central neurons to axotomy: Ionic regulation and electrical activity. *FASEB Journal* 18: 1934–1936.

Neumann S and Woolf CJ (1999) Regeneration of dorsal column fibers into and beyond the lesion site following adult spinal cord injury. *Neuron* 23: 83–91.

Perlson E, Hanz S, Ben-Yaakov K, et al. (2005) Vimentin-dependent spatial translocation of an activated MAP kinase in injured nerve. *Neuron* 45: 715–726.

Perlson E, Hanz S, Medzihradszky KF, et al. (2004) From snails to sciatic nerve: Retrograde injury signaling from axon to soma in lesioned neurons. *Journal of Neurobiology* 58: 287–294.

Perlson E, Medzihradszky KF, Darula Z, et al. (2004) Differential proteomics reveals multiple components in retrogradely transported axoplasm after nerve injury. *Molecular & Cellular Proteomics* 3: 510–520.

Plunet W, Kwon BK, and Tetzlaff W (2002) Promoting axonal regeneration in the central nervous system by enhancing the cell body response to axotomy. *Journal of Neuroscience Research* 68: 1–6.

Richardson PM and Issa VM (1984) Peripheral injury enhances central regeneration of primary sensory neurones. *Nature* 309: 791–793.

Schmitt AB, Breuer S, Liman J, et al. (2003) Identification of regeneration-associated genes after central and peripheral nerve injury in the adult rat. *BMC Neuroscience* 4: 8.

Sheu JY, Kulhanek DJ, and Eckenstein FP (2000) Differential patterns of ERK and STAT3 phosphorylation after sciatic nerve transection in the rat. *Experimental Neurology* 166: 392–402.

Sung YJ, Walters ET, and Ambron RT (2004) A neuronal isoform of protein kinase G couples mitogen-activated protein kinase nuclear import to axotomy-induced long-term hyperexcitability in *Aplysia* sensory neurons. *Journal of Neuroscience* 24: 7583–7595.

Twiss JL and Minnen JV (2006) New insights into neuronal regeneration: The role of axonal protein synthesis in pathfinding and axonal extension. *Journal of Neurotrauma* 23: 295–308.

Zhang XP and Ambron RT (2000) Positive injury signals induce growth and prolong survival in *Aplysia* neurons. *Journal of Neurobiology* 45: 84–94.

Transporter Proteins in Neurons and Glia

T S Otis and P D Dodson, Geffen School of Medicine, Los Angles, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Nerve cells communicate with one another primarily by chemical means, and such chemical messages are carried by neurotransmitters released at specialized contacts called synapses. In order to produce fast transmission between neurons, the chemical signal must be rapidly terminated. One way in which the signal can be terminated is by the passive diffusion of the neurotransmitter. However, at most synapses in the brain and peripheral nervous system, the termination of chemical messages is also aided by active processes, either enzymatic degradation or removal of the neurotransmitter from the extracellular space.

At synapses between motor neurons and skeletal muscles, the excitatory neurotransmitter acetylcholine is rapidly degraded by an enzyme located within the synaptic cleft (see Figure 1). This enzyme, acetylcholinesterase, is one of the most avid enzymes described; a single acetylcholinesterase protein is capable of degrading more than 10 000 molecules of acetylcholine each second. Such capacity reduces the time period in which acetylcholine must act, and it also limits the spatial extent over which it signals.

In the brain, fast excitatory synaptic signals are mediated primarily by the excitatory amino acid glutamate. Neurotransmitter clearance at glutamatergic synapses is quite different from the neuromuscular synapse. Here the actions of glutamate are terminated in part by a special class of neurotransmitter transporter, called excitatory amino acid transporters (EAATs), located on the plasma membranes of glial cells and neurons (Figure 1). EAAT proteins are responsible for capturing glutamate from the extracellular space, undergoing a conformational change which transports the glutamate through the plasma membrane, and then releasing it into the cytoplasm.

In general, neurotransmitter transport can serve many purposes. First, it is thought to act as part of a recycling mechanism that allows for the renewal of neurotransmitter pools. Second, it can limit the temporal actions of the neurotransmitter. Third, it can restrict the spatial extent of neurotransmitter action and, in so doing, can sometimes determine which pools of receptors are activated. Fourth, it can help set resting levels of neurotransmitter. This is significant because distinct high-affinity classes of receptors are often found outside synaptic specializations (called extrasynaptic receptors) and such extrasynaptic receptors typically respond to low levels of neurotransmitter. Still other types of receptors desensitize when exposed to very low levels of neurotransmitter, much lower than those required to fully open the receptor. In these ways, even low resting levels of the neurotransmitter can serve as signals of their own or as negative influences on surrounding receptors.

A Family of Excitatory Amino Acid Transporters Responsible for Clearing the Major Excitatory Neurotransmitter Glutamate

Molecular cloning techniques have identified EAATs in a wide variety of species: prokaryotes, insects, amphibians, fish, and mammals. Mammals possess a family of EAATs consisting of five genes. Three of these genes were initially identified in 1992 and named GLAST (glutamate aspartate transporter), GLT1 (glutamate transporter 1), and EAAC1 (excitatory amino acid carrier 1). Soon after, corresponding sequences for these genes were cloned from the human motor cortex and their homologs were named EAAT1 (GLAST), EAAT2 (GLT1), and EAAT3 (EAAC1). Subsequently EAAT4 and EAAT5 were also identified. For the sake of simplicity, we hereafter refer to each subtype by its human (i.e., EAAT) designation.

Molecular Structure of Excitatory Amino Acid Transporters

Our understanding of the molecular structure of EAATs has benefited enormously from the recently determined X-ray crystal structure of a prokaryotic EAAT homolog. This protein, found in the thermophilic archaebacterium *Pyrococcus horikoshii*, has approximately 35% sequence homology to its mammalian counterparts. Figure 2 summarizes the key features of the structure. The transporter exists as a homotrimer with three wedge-shaped subunits forming the sides a large bowl that dips into the membrane and faces the extracellular space (Figure 2(b)). Each subunit exhibits a complicated transmembrane topology (Figure 2(a)), with eight complete membrane-spanning segments and two reentrant loops. The reentrant loops are situated at the bottom of the bowl, where the tips of the three wedges contributed by each subunit come together. This is a key functional region of the transporter protein; the reentrant loops had been identified in prior studies as the likely spot at which glutamate binds. Indeed, the crystal structure shows electron densities characteristic of single glutamate molecules bound to each subunit at the base of the bowl, suggesting that three glutamate molecules bind at these locations before being

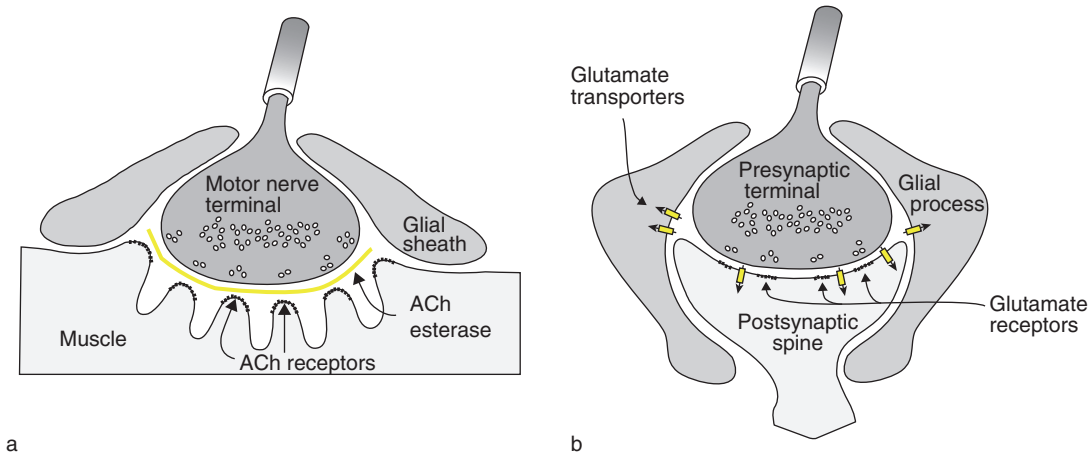

Figure 1 Degradative enzymes: (a) in the vertebrate neuromuscular junction synapses; (b) in the extracellular space at most excitatory synapses within the brain. In (a), following release the neurotransmitter acetylcholine (ACh) is degraded rapidly by acetylcholinesterase (ACh esterase), an enzyme situated within the synaptic cleft. In (b), there is no degradative enzyme; rather, high-affinity excitatory amino acid transporters (EAATs) remove glutamate from the extracellular space into intracellular compartments in neurons and glia.

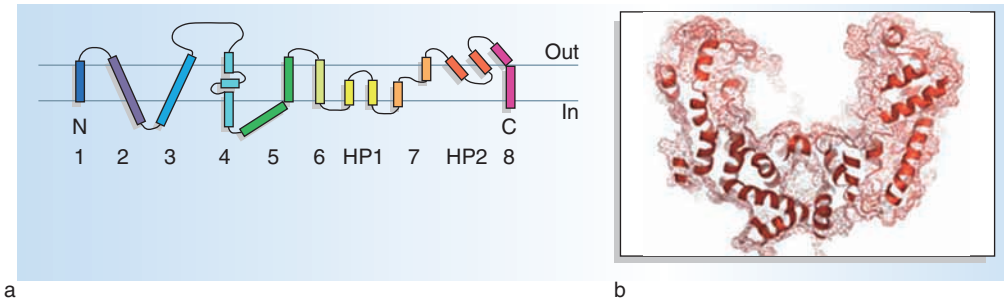

Figure 2 Transmembrane topology of a prokaryotic excitatory amino acid transporter (EAAT) from the archaebacterium *Pyrococcus horikoshii*: (a) structure of the transporter subunits; (b) side-on cross-sectional view of the transporter structure. In (a), each subunit has eight membrane-spanning segments and two reentrant loops, termed hairpin 1 (HP1) and hairpin 2 (HP2). The N- and C-termini are located on the intracellular side of the membrane. In (b), the large bowl-shaped structure formed by three subunits is oriented with the bowl facing extracellularly (top). Glutamate binding and coupled ion binding are believed to occur at the base of the bowl. Reproduced by permission from Macmillan Publishers Ltd.: *Nature*, MP Kavanaugh, Accessing a transporter structure, 431: 752–753, copyright 2004.

transported into the cytoplasm. Although it is not certain, these features of the structure suggest that each subunit may transport glutamate independently.

Energetics of Excitatory Amino Acid Transporters

The energy for transport comes from the coupling of the movement of glutamate to the cotransport and the countertransport of sodium, potassium, and hydrogen ions down their respective electrochemical gradients. Flux measurements have established the stoichiometry for a complete cycle of glutamate transport to be three Na^+ and one H^+ cotransported and one K^+ countertransported per glutamate molecule. This elaborate coupling mechanism has a clear benefit; it allows EAATs to accumulate glutamate to a huge concentration gradient. With typical physiological ion gradients, resting potentials, and the coupling stoichiometry just mentioned, glutamate can be taken up and concentrated inside the cell to well more than a millionfold at equilibrium. Given that cytoplasmic glutamate concentrations are thought to be in the millimolar range in some cells (e.g., neurons), these energetics allow for extracellular glutamate levels to be maintained at submicromolar, and possibly even nanomolar, levels. Of course, even at steady state, resting glutamate levels may not be near this theoretical limit. This is because, in addition to the ionic gradients, the amount of glutamate released, the density of EAATs, and the size and shape of the extracellular volume all influence the degree to which transporters can reduce the extracellular glutamate concentration.

Uncoupled Ionic Conductances of Excitatory Amino Acid Transporters

The stoichiometric movement of ions during the glutamate transport cycle provides the energy to accumulate glutamate. Because of the net charge flux of +2 charges moving inward per cycle (one negative glutamate, three positive sodium ions, and one proton in and one positive potassium ion out) transport gives rise to very small transmembrane ionic currents. However, to varying degrees, all five EAATs allow another ionic current to flow. This current is carried by chloride ions under physiological circumstances, and this movement of anions is not coupled energetically to glutamate transport. In simple terms, we might imagine that at a particular point in the cycle of transport the EAAT protein adopts a conformation contains an anion-selective pore across the membrane. The anions flowing through EAATs move according to their own electrochemical gradient, and this anion flux is neither affected by nor affects the cycling of glutamate, sodium, potassium, and hydrogen ions.

The relative size of the anion component of the total current varies characteristically among different EAAT isoforms. EAAT4 has the largest anion current; it dominates the total ionic flux under physiological circumstances. In contrast, EAAT2 has the smallest anion conductance. Several of the EAATs also exhibit a noticeable anion leak under resting conditions (i.e., in the absence of glutamate), and this leak is most prominent in EAAT4. Physiological roles for the uncoupled anion fluxes are not known, but in principle such conductances could allow EAATs to act as ligand-gated inhibitory channels in addition to their glutamate clearance role. In cerebellar Purkinje neurons, in which EAAT4 is expressed at high levels, this does not appear to have a major influence on excitability, but in retinal ON-bipolar cells it has been argued that EAAT-mediated anion currents can couple decreases in photoreceptor glutamate release with the excitation of the ON-bipolar cell.

Cellular Localization of Excitatory Amino Acid Transporters

EAATs are expressed at the highest levels in the nervous system, but are also found in peripheral tissues such as the kidney, enteric nervous system, pancreas, placenta, mammary glands, and possibly bone. The expression of particular EAATs within the brain occurs in overlapping patterns and is largely cell-type specific. Two of the five transporters are found predominantly in the glia, two are found only in certain types of neurons, and one is expressed only in the retina in both neurons and the glia.

The Glial Transporters EAAT1 and EAAT2

EAAT1 is the predominant glial glutamate transporter in cerebellum. It is also expressed at relatively high levels in glial cells of the inner ear and in Müller glial cells of the retina. EAAT1 protein does not appear to be present in oligodendrocytes or in neurons. In cerebellar Bergmann glia, EAAT1 is expressed at extremely high densities ($\sim$4700 m^{-2} of plasma membrane). It is thought to be the dominant transporter in these cells; consistent with this view, animals in which the EAAT1 gene has been disrupted have cerebellar behavioral deficits and exacerbated responses to cerebellar injuries. However, EAAT2 is also expressed at approximately sixfold lower levels in the cerebellar cortex, and thus functional transporters may exist as heteromers.

At a subcellular level, EAAT1 is not found to be distributed uniformly along the plasma membrane of astrocytes. Rather, hot spots of EAAT1 are observed and the highest densities of EAAT1 are seen where astrocytic membranes oppose neuronal membranes (see Figure 3(a)). In the cerebellar cortex, this arrangement could be a mechanism for helping to limit glutamate spillover from active synapse to their neighbors.

Like EAAT1, EAAT2 is also concentrated on fine astrocytic processes that surround excitatory synapses (see Figure 3(b)). Depending on the brain region, these EAAT2-rich astrocytic processes are located at different distances from sites of glutamate release. At one extreme, they can encapsulate pre- and postsynaptic elements (as EAAT1/EAAT2-studded Bergmann glial processes do in the cerebellar cortex). At the other extreme, such as at Schaffer collateral synapses on to CA1 pyramidal neurons of the hippocampus, EAAT2-containing processes are located at some distance from postsynaptic densitities and sites of glutamate release. Finally, in certain brain regions such as the superoptic nucleus of the hypothalamus, the arrangement of astrocytic processes varies according to behavioral state, thereby influencing excitatory synaptic transmission in numerous ways by permitting glutamate to reach or limiting it from reaching specific pools of receptors.

The Neuronal Transporters EAAT3 and EAAT4

An unexpected feature of the localization patterns of the neuronal EAATs is that they are found on dendritic, but not on axon terminal, membranes. EAAT3 is the more widely expressed of the neuronal EAATs, showing localization on excitatory and inhibitory neuronal dendrites in the hippocampus, cerebellum, basal ganglia, cortex, and thalamus. In those neurons that express it, EAAT3 protein is present at relatively high densities on intracellular membranes as well as

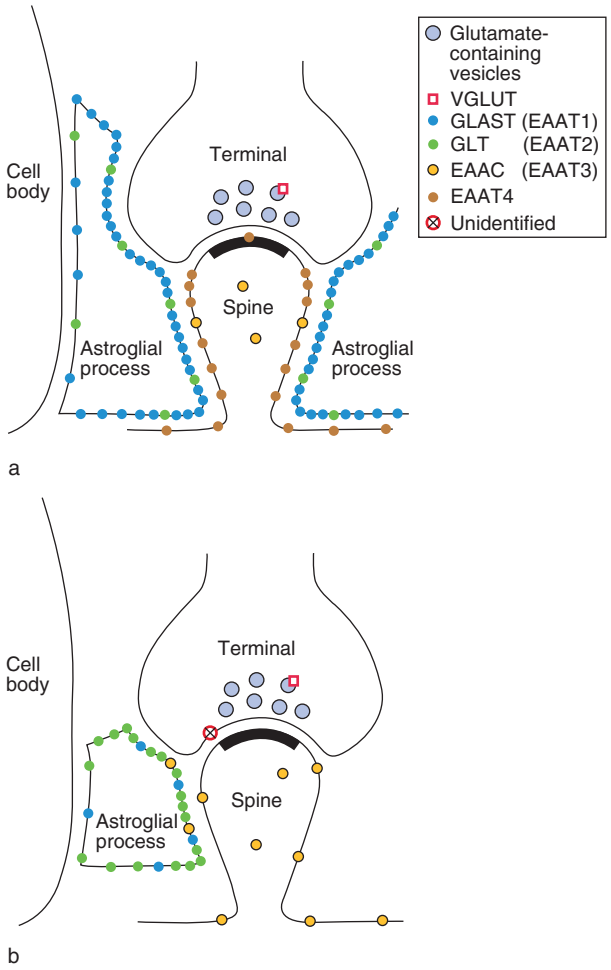

Figure 3 Localization of excitatory amino acid transporters (EAATs) at different types of glutamate synapses: (a) parallel fiber to Purkinje neuron synapses in the cerebellar cortex; (b) Schaffer collateral synapse in hippocampus. In (a), the synapses are completely surrounded by astroglial membranes containing the glial isoforms EAAT1 and EAAT2 at average densities of 4700 and 740 molecules μm^{-2}, respectively. The neuronal transporter EAAT4 is concentrated opposite glia-covered parts of the membranes of Purkinje dendrites and is highest in the spines and thinner dendrites. EAAT4 is expressed at varying density in different zones of the cerebellar cortex, but has an average density of 1800 molecules μm^{-2}. EAAT3 is present in Purkinje neurons membrane as well as the cytoplasm, but quantitative data on the precise subcellular distribution are currently unavailable. In (b), the synapses are only contacted, not surrounded, by astroglia. EAAT1 and EAAT2 are present in the astroglial membranes at average densities of 2300 and 8500 molecules μm^{-2}, respectively, with highest concentrations toward the neuropil. Quantitative information is not available for EAAT3, nor is information available for the putative glutamate transporter on the nerve terminal. EAAC, excitatory amino acid carrier; GLAST, glutamate aspartate transporter; GLT, glutamate transporter; VGLUT, vesicular glutamate transporter. Reproduced from Danbolt N (2001) Glutamate transporters. *Progress in Neurobiology* 65: 1–105, with permission from Elsevier.

on the plasma membrane. That said, the density of EAAT3 is believed to be quite low and it is unclear whether it plays a prominent role in clearing glutamate into these neurons.

EAAT4, found almost exclusively in cerebellar Purkine neurons, is present at extremely high densities on dendritic shafts and spines, estimated to be on average ($\sim$1800 m^{-2} of plasma membrane). However, this expression shows a large-scale patchiness, with zones of high expression and low expression occurring at periodic spatial intervals throughout the cerebellar cortex. The resulting zebra-striped pattern matches that of several other gene products (most notably the protein Zebrin). The reasons for this nonuniform expression pattern are not understood, but recent work suggests that in regions of low EAAT4 expression it is easier for certain types of glutamate receptors (GluRs; e.g., metabotropic GluRs (mGluRs)) to be stimulated, whereas in regions with high EAAT4 expression certain GluRs appear to be shielded from released glutamate. Because mGluRs are responsible for initiating particular forms of synaptic plasticity, the variability in EAAT4 may be a mechanism for region specific control of synaptic plasticity.

At the subcellular level, both EAAT3 and EAAT4 have been reported to be excluded from the postsynaptic density but to be concentrated at the edges of these densities, a pattern termed perisynaptic. Such a pattern of localization seems ideally suited to aiding in preventing glutamate escape from synaptic clefts at active synapses (spillover) and in preventing glutamate from gaining access to the postsynaptic densities of inactive synapses (cross-talk). These phenomena are illustrated in Figure 4; physiological evidence suggesting that they occur is discussed in latter sections.

The Retinal Transporter EAAT5

EAAT5 has been reported to be expressed in various neurons (photoreceptors, bipolar cells, and amacrine cells) and Müller glial cells of the retina. Functional evidence suggests that EAAT5 may be a major glutamate transporter on bipolar cell terminals; this work is also notable because it serves as the only direct evidence to date that an EAAT is involved in glutamate uptake into a presynaptic terminal.

Synaptic Functions of Excitatory Amino Acid Transporters

Recycling Glutamate

Constant vesicular release of glutamate from terminals in the brain requires a concurrent replenishment of neurotransmitter stores. It is likely that this happens via an indirect pathway because there is little evidence for the presence of glutamate transporters located on presynaptic terminals. Rather than being directly reabsorbed into the nerve terminal, EAATs in surrounding glia may take up the bulk of glutamate.

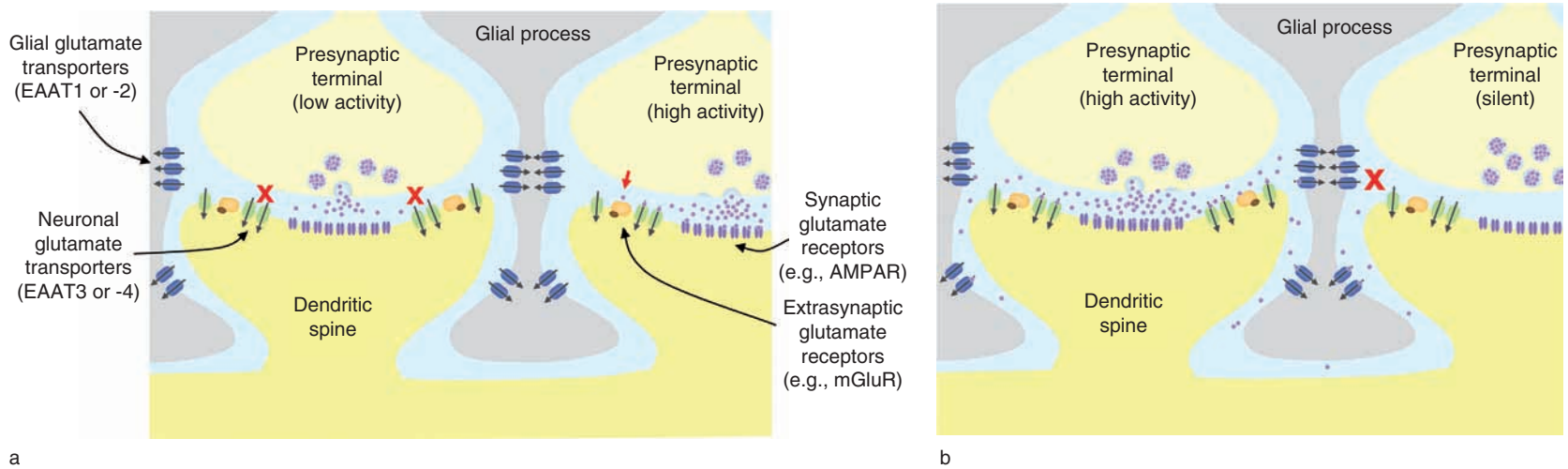

Figure 4 Transporters limit (a) spillover of glutamate; and (b) of synaptic cross-talk. In (a), during moderate levels of activity (left), glutamate transporters take up glutamate and limit its escape from the cleft (red crosses). During high levels of activity, the transporters become saturated, so the transmitter spills over (arrow) to activate extrasynaptic receptors. In (b), during high levels of activity (left), transporters are saturated and glutamate escapes the synaptic cleft. Transporters surrounding adjacent synapses take up this glutamate (red cross), limiting cross-talk between synapses. AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor; EAAT, excitatory amino acid transporter; mGluR, metabolic glutamate receptor.

Within glia, glutamate is converted to glutamine by the enzyme glutamine synthetase. Glutamine is an inactive amino acid and thus can be returned to the presynaptic terminal via the extracellular space; once it is back in the presynaptic terminal, it can be made back into glutamate (or γ-aminobutyric acid (GABA)).

There is also indirect evidence that EAATs on inhibitory synaptic terminals take up glutamate to be used for the synthesis of GABA. In these studies, inhibiting glutamate uptake leads to a reduction in GABA release. This result might also explain why many GABAergic neurons express EAAT3 and EAAT4, although, if such proteins are present on the terminals of these neurons, they are, for unknown reasons, immunohistochemically silent.

Maintaining Low Resting Glutamate Concentrations

In order to preserve the sensitivity of postsynaptic receptors, steady-state concentrations of glutamate in the synaptic cleft must remain submicromolar. For example, extrasynaptic receptors such as mGluRs and *N*-methyl-D-aspartate (NMDA) receptors have median effective concentrations (EC_{50}s) in the micromolar range, whereas α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors desensitize to similarly low concentrations of glutamate. Because the rate at which glutamate can leave the synaptic cleft is proportional to its concentration gradient, in order to rapidly terminate the signal a large concentration gradient must be established. Several experiments have demonstrated that EAATs control the resting levels of glutamate. The strongest evidence is based on the measurement of tonic activation of GluR currents, typically NMDA currents. For example, during cortical and cerebellar development many migrating neurons receive tonic nonsynaptic signals mediated by NMDA receptors. Such signals, and the rate of migration of these neurons, can be altered by inhibiting EAATs. In adults, EAATs also appear to play a role in controlling steady-state levels of glutamate. Experiments in cultured hippocampal slices show that large tonic NMDA currents appear within seconds after the inhibition of transporters.

Synaptic signals mediated by AMPA receptors are sometimes altered by inhibiting glutamate uptake, but this seems to depend on the microstructure of the synapse. Three categories of effects are observed when EAATs are inhibited. At some synapses that are not enveloped by glia, as in the hippocampus (schematized in Figure 3(a)), inhibiting glutamate uptake has no detectable effect. This stands in contrast to the slowing of the decay of the excitatory postsynaptic current (EPSC) in response to single stimuli seen at other types of synapses, such as the calyceal synapses or synapses, which are made on postsynaptic spines with extensive glial wrapping as in the cerebellar cortex (schematized in Figure 3(b)). Finally, at several glial-enveloped synapses that are made on dendritic shafts, blocking transport causes a slowing of the decay but only in response to trains of presynaptic stimuli; a good example is the parallel fiber to stellate cell synapse in the cerebellum.

Ensuring Synapse Specificity

In order to ensure independent signaling, glutamate must be prevented from diffusing from one synaptic cleft into neighboring synaptic clefts. The escape of glutamate from the synaptic cleft and the subsequent activation of extrasynaptic AMPA or NMDA receptors have been termed spillover; the breakdown of synaptic independence caused by invading glutamate into an inactive synaptic cleft has been termed cross-talk (see Figure 4). Given the abundant evidence that glutamate escapes the synaptic cleft at significant concentrations and can activate targets at considerable distances, it would seem that these two phenomena are common. Close examination of excitatory transmission at several synapses yields clear evidence for spillover and tantalizing suggestions of cross-talk. Where it has been examined, glutamate uptake plays a pivotal role in minimizing or shaping these unconventional signals.

Limiting cross-talk between synapses An early indication that cross-talk might occur came from a comparison of the relative variabilities of NMDA and AMPA receptor-mediated EPSCs in the hippocampus. Due to the stochastic properties of vesicular release (release probability is not 1.0), EPSCs vary in peak amplitude from trial to trial. Recordings of dual component EPSCs in CA1 pyramidal cells showed that NMDA-mediated EPSCs are consistently less variable than those mediated by AMPA receptors. These results have been interpreted to mean that NMDA receptors detect a larger number of quanta than do AMPA receptors. But where do these extra quanta come from? NMDA receptors have approximately 100-fold higher affinity than AMPA receptors, and it has been hypothesized that this allows NMDA receptors to detect release from neighboring presynaptic terminals not directly apposed to the postsynaptic density in which they reside. A key piece of evidence supporting this idea is that, at higher temperatures, the difference in relative variability between NMDA and AMPA receptors is reduced. Glutamate transport is known to be very temperature sensitive, and consistent with a role for glutamate uptake in this puzzle, inhibitors of glutamate transport erased the effects of increased temperature, making the NMDA EPSCs

relatively more variable than AMPA EPSCs. These results imply that cross-talk to NMDA receptors is occurring at room temperature but that at more physiological temperatures glutamate transporters operate more efficiently and are able to prevent glutamate from indirectly activating the higher-affinity NMDA receptors.

Perhaps the most straightforward evidence of cross-talk has been observed at parallel fiber inputs to stellate cells in the cerebellum, Here, slowly rising AMPA- and NMDA-mediated synaptic responses can be observed under certain conditions in response to trains of presynaptic stimuli. Inhibiting glutamate uptake increases and prolongs such slow responses, indicating that EAATs at least partially limit the indirect access of glutamate to AMPA receptors. Other pharmacological evidence suggests that these indirect responses result at least in part from glutamate entering inactive, neighboring synapses and binding to AMPA or NMDA receptors.

Neuronal EAATs located at the edges of synapses (see Figure 3) seem to be ideally situated to prevent glutamate from entering synaptic clefts from the extracellular space. Is there any evidence that transporters might serve in this capacity? In CA1 pyramidal cells of the hippocampus, experiments suggest that postsynaptic uptake, presumably by EAAT3, can limit the activation of NMDA receptors. Similarly, in cerebellar Purkinje neurons, EAAT4 appears to prevent low concentrations of glutamate from entering inactive synapses.

Limiting spillover of neurotransmitter to extrasynaptic receptors Many potential targets for glutamate are located adjacent to the synapse or completely outside of synaptic clefts – on the same neuron, on presynaptic terminals, or on glial cells. In order for these targets to be activated, glutamate must escape a synapse and diffuse for variable distances in the extracellular space. Recently published data not only demonstrate that glutamate can indeed spill out of the synaptic cleft to activate extrasynaptic receptors but also that EAATs play important roles in ensuring that these receptors are activated only under certain circumstances (e.g., during high levels of activity).

Spillover to receptors on glia In the hippocampus, where glial cell membranes are typically located at some distance from the edges of synaptic clefts, synaptic signals mediated by GluRs or EAATs can be directly recorded in astrocytes. Based on such measurements, it has been proposed that glial membranes outside the synaptic cleft sense glutamate within 1 ms after release, results consistent with expectation maximization (EM) reconstructions of astrocyte membranes and with the localization data for EAAT2 placing the transporters at ~1 μm from synaptic specializations. Likewise, in the cerebellum, AMPA receptor-mediated and EAAT-mediated synaptic currents can be recorded from Bergmann glial cells. These findings suggest that glutamate commonly escapes synaptic clefts at concentrations easily capable of activating most glutamate receptor subtypes.

Spillover to extrasynaptic receptors on neurons In various brain regions, including the hippocampus, retina, hypothalamus, and cerebellum, glutamate escaping synaptic clefts can activate NMDA receptors and these signals are often potently regulated by EAATs. Glutamate escaping synapses can also activate various classes of mGluRs. Some of these are located at the very edges of synapses, surrounding the AMPA and NMDA receptor-rich postsynaptic densities. Both cerebellar parallel fiber synapses and Schaffer collateral synapses have postsynaptic mGluRs (group I) that encircle the postsynaptic densities. At cerebellar synapses, high-frequency trains of presynaptic stimuli are required to elicit mGluR-triggered EPSCs and intracellular calcium signals. In both the hippocampus and cerebellum, the activation of these perisynaptic G-protein-coupled receptor subtypes appears to be limited by EAATs (see Fig ure 4). Postsynaptic neuronal EAATs are co-localized with the mGluRs, and in the cerebellum neuronal EAATs seem to be specialized to limit the activation of these perisynaptic mGluRs.

Spillover to receptors on presynaptic terminals Another class of distant glutamate targets include group II mGluRs (mGluR2 and mGluR3), which are located on excitatory presynaptic terminals. In hippocampal cultures, the mGluR-mediated presynaptic inhibition of glutamate and GABA release is enhanced by antagonizing glutamate uptake. In acute hippocampal slices, a presynaptic mGluR-mediated inhibition caused by spillover of glutamate to distant heterosynaptic targets is limited by EAAT activity. Thus, glutamate acting at some distance from release sites on presynaptic mGluRs generates a form of presynaptic autoinhibition, and this autoinhibition is tightly controlled by glutamate uptake. Similar findings are obtained for presynaptic mGluRs on GABA-releasing terminals.

See also: Glial Glutamate Transporters: Electrophysiology; Glial Glutamate Transporters; Glial Glutamate and GABA Metabolism; Glutamate; Glutamate Receptor Organization: Ultrastructural Insights.

Further Reading

Amara SG and Fontana AC (2002) Excitatory amino acid transporters: Keeping up with glutamate. *Neurochemistry International* 41: 313–318.

Bergles DE, Diamond JS, and Jahr CE (1999) Clearance of glutamate inside the synapse and beyond. *Current Opinions in Neurobiology* 9: 293–298.

Danbolt N (2001) Glutamate transporters. *Progress in Neurobiology* 65: 1–105.

Kavanaugh MP (2004) Accessing a transporter structure. *Nature* 431: 752–753.

Otis TS, Brasnjo G, Dzubay JA, and Pratap M (2004) Interactions between glutamate transporters and metabotropic receptors at excitatory synapses in the cerebellar cortex. *Neurochemistry International* 45: 537–544.

Otis TS, Kavanaugh MP, and Jahr CE (1997) Postsynaptic glutamate transport at the climbing fiber-Purkinje cell synapse. *Science* 277: 1515–1518.

Palmer MJ, Taschenberger H, Hull C, Tremere L, and von Gersdorff H (2003) Synaptic activation of presynaptic glutamate transporter currents in nerve terminals. *Journal of Neuroscience* 23: 4831–4841.

Robinson MB (2003) Signaling pathways take aim at neurotransmitter transporters. *Science Signal Transduction Knowledge Environment* 207: pe50 (online).

Sattler R and Rothstein JD (2006) Regulation and dysregulation of glutamate transporters. *Handbook of Experimental Pharmacology* 175: 277–303.

Scanziani M (2002) Competing on the edge. *Trends in Neuroscience* 25: 282–283.

Theodosis DT, Piet R, Poulain DA, and Oliet SH (2004) Neuronal, glial and synaptic remodeling in the adult hypothalamus: Functional consequences and role of cell surface and extracellular matrix adhesion molecules. *Neurochemistry International* 45: 491–501.

Yernool D, Boudker O, Jin Y, and Gouaux E (2004) Structure of a glutamate transporter homologue from. *Pyrococcus horikoshii*. *Nature* 431: 811–818.

Zerangue N and Kavanaugh MP (1996) Flux coupling in a neuronal glutamate transporter. *Nature* 383: 634–637.

Trigeminal Motor System

J P Lund, McGill University, Montreal, QC, Canada
A Kolta, University of Montreal, Montreal, QC, Canada
B J Sessle, University of Toronto, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The trigeminal (V) motor system includes most of the jaw muscles, plus the brain stem and higher brain centers that control them and several types of sensory afferent neurons that modulate their activity. Most of these muscles are complex and are made up of many anatomical compartments that can be differentially activated, which allows great variability in patterning of movement. The major motor functions of the system include biting, chewing (mastication), swallowing, respiration, and speech, as well as gagging, coughing, and several more simple reflexes. The proper function of the jaw muscles and their effector organs (the teeth) is crucial for efficient ingestive motor behavior, during which the foodstuffs are crushed and ground down to a size appropriate for swallowing. Highly coordinated processes that include efficient sensory feedback are necessary to allow forces to be properly adapted to the texture of foodstuffs in the mouth during chewing and to coordinate the articulation phase of speech. Major dysfunctions of the central nervous system (CNS) (e.g., cerebral palsy and stroke) can have devastating effects on ingestive motor functions and speech, but the most common cause of dysfunction by far is the loss of teeth. This article provides a review of the neural circuits and mechanisms that underlie the function of the V motor system, with a particular focus on the processes in the CNS.

Peripheral Components

The peripheral components of the V motor system in humans and other mammals include the so-called 'jaw-closing' (JC) muscles (masseter, temporalis, medial pterygoid), the jaw-opening (JO) muscles (anterior digastric, mylohyoid), and the lateral pterygoid muscle, the superior and inferior heads of which may contribute differentially to JC, JO, and horizontal jaw movements. The contractile and viscoelastic elements of these muscles, as well as their sensory and motor innervations, are beyond the scope of this article but have recently been reviewed. Their receptors in general show properties analogous to those of their counterparts in muscles supplied by spinal nerves, but it should be noted that there are some significant differences between these features of the jaw muscles and those of the limbs, trunk, and neck. Whereas muscle spindles and Golgi tendon organs are plentiful in most spinally innervated muscles and, especially in the case of the muscle spindles, are viewed as essential elements of motor control, they are virtually nonexistent in the anterior digastric, mylohyoid, and lateral pterygoid muscles. Moreover, primary afferent neurons that innervate the muscle spindles of the JC muscles have their primary afferent cell bodies within the CNS, in the V mesencephalic nucleus (Mes V), and not in the trigeminal equivalent of the spinal dorsal root ganglia. This endows JC muscle spindles, and thus the V motor system, with unique features (see next section). Another unique feature of the peripheral components is that the V motor system can utilize sensory inputs from tissues that occur only in the orofacial region (e.g., tooth pulp, periodontal tissues, cornea). The V motor system thus has a wide range of sensory inputs from muscle (muscle spindles, tendon organs, nociceptors), teeth, cornea, facial skin, oral and nasal mucosa, periosteum, and temporomandibular joint (TMJ) that can initiate or regulate jaw muscle activity. These are considered in more detail in the following sections.

Central Components

Trigeminal Motor Neurons

The V motor neurons are arranged in a musculotopic fashion within the trigeminal motor nucleus (Motor V) in the brain stem. The pools supplying the powerful temporalis, masseter, and medial pterygoid muscles overlap slightly and occupy the rostral two-thirds of the nucleus. Within each pool of these complex JC muscles, motor neurons supplying functionally distinct compartments intermingle. The motor neurons of the weak JO muscles (anterior digastric, mylohyoid) are found in the ventromedial middle third and the caudal portion of Motor V. The muscles are also innervated by a few small cholinergic neurons ventral to the V motor root. The dendrites of JC motor neurons ramify extensively, but do not extend far beyond the boundaries of Motor V, while JO motor neuron have dendrites extending into the adjacent reticular areas, but not into JC motor neuron pools. Spinal motor neurons have recurrent collaterals that contact Renshaw cells, which are responsible for recurrent inhibition, but there is no good evidence for recurrent inhibition in Motor V.

Cajal was the first to note that V motor neurons appear to enter the motor root without giving off collaterals, and this has been confirmed many times in the subsequent century.

Like other types of motor neurons, V motor neurons express several types of voltage-gated Ca^{2+} channels. These are involved in the release of acetylcholine (ACh) at the end plates but also in the control of the electrical properties of the soma and dendrites. As in spinal motor neurons, Ca^{2+} channels are essential for the generation of sustained bursts of activity during movement. The channels associated with *N*-methyl-D-aspartate (NMDA) receptors seem to be particularly important in motor control, and blockage of these receptors suppresses motor neuron activity during rhythmical jaw movements. The activation of NMDA receptors has several effects that allow the motor neurons to oscillate between two relatively stable membrane states: the depolarized state (plateau potential) during which they fire repetitively, and a repolarized silent state.

The neurotransmitters and modulators that have been identified in synaptic terminals on spinal motor neurons have also been found in synapses within Motor V. These include excitatory and inhibitory amino acids, biogenic amines, neuropeptides, and nitric oxide, and the effects of these substances on trigeminal and spinal motor neurons do not differ in any substantial way. However, there are differences in inputs to JC and JO motor neurons. Up to 80% of terminals contacting JC motor neurons contain inhibitory amino acids (γ-aminobutyric acid (GABA), glycine, or both), compared to about 40% of synapses on JO motor neurons. Most of those remaining are glutamatergic; less than 5% contain the other transmitters and modulators. The most studied of the latter is serotonin, which originates in the raphé nuclei. It has little effect on quiescent motor neurons, but does increase motor neuron excitability during suckling and mastication.

As noted earlier, a unique feature of the V motor system is that cell bodies of all the muscle spindle primary afferents (and about half of the mechanoreceptors in the periodontal ligament) are within the CNS, where they form the V mesencephalic nucleus (Mes V). They are large monopolar neurons like mechanoreceptors in the V and dorsal root ganglia, but unlike the latter, they are contacted by synaptic terminals. Inputs to these neurons from neighboring areas are mainly excitatory, even those mediated by GABAA receptors. GABA depolarizes these neurons because their chloride equilibrium potential is above their resting membrane potential due to a lack of a specific chloride transporter.

Like some large-diameter spinal dorsal root afferents, the jaw muscle spindle cell bodies have resonant properties that produce high-frequency oscillations near their resting membrane potential. Synaptic inputs reset the oscillation and increase the amplitude of the fast oscillations for several cycles. This, combined with electrical coupling to other afferents or to the dendrites of multipolar interneurons, may cause sustained and synchronous firing in groups of Mes V neurons. Synchrony may be important during rapid biting and during the mastication of very hard or tough foods.

Synaptic inputs to jaw muscle spindle afferents are not restricted to the soma. As is the case in other classes of sensory afferents, GABAergic synapses also occur on presynaptic terminals where they induce depolarization that is sometimes strong enough to trigger firing of antidromic action potentials. GABAergic axons are also found in close apposition to the main axons of these afferents, and inputs from local interneurons cause GABAA receptor-mediated blockade of the antidromic action potentials. These mechanisms seem to contribute to the patterning of mastication (see the section titled 'Brain stem pattern-generating circuits').

Premotor Interneurons

Many inputs to Motor V are supplied by glutamatergic, GABAergic, and glycinergic neurons within the rostral V spinal nucleus, the dorsomedial V main sensory nucleus (Main V), and a large band of the adjacent parvocellular lateral reticular formation (RF). These nuclei project to both Motor V nuclei and their contralateral counterparts. The RF interneurons form a column that runs caudally from the lateral part of the nucleus pontis oralis, through the Motor V and the reticular zone surrounding it (including the supratrigeminal and intertrigeminal areas). This column continues caudally through the parvocellular lateral RF down to the hypoglossal nucleus. There are very few direct inputs from the medial RF to Motor V.

The premotor neurons are mainly small fusiform or triangular neurons with dendritic trees extending horizontally in two to three directions over less than 500 μm. Most regions project to both JO and JC pools, but there are some distinctions: dorso-medial Main V projects predominantly, if not exclusively, to JC motor neurons, while projections from the rostral V spinal nucleus are mostly directed to JO motor neurons.

On the basis of a limited number of studies, it seems that the electrophysiological properties of these cells are relatively uniform and typical of small interneurons. Many are spontaneously active, but few

fire rhythmically in response to hyperpolarizing or depolarizing current injection. The exception is dorsomedial Main V, where rhythmic bursting appears to be common (see next section).

Many premotor neurons receive very strong monosynaptic inputs from primary afferents and motor cortex, but they are also tightly interconnected. Stimulation of any subdivision of the premotor neuron zone affects all other subdivisions. Most postsynaptic potentials are excitatory (EPSPs), but some are inhibitory (IPSPs) or mixed. Transmission is rapid and robust in this network. Some parts of the network are strongly interconnected with the medial RF.

Brain Stem Pattern-Generating Circuits

As noted previously, the jaw muscles do the major work in chewing and also participate in several other sensorimotor functions that include swallowing and respiration. These three primitive but vital motor patterns can all be carried out by animals (including humans) that have no CNS above the brain stem. All three of these motor patterns depend on hindbrain central pattern generators (CPGs), neural networks that can generate the fundamental patterns of movement in the absence of structured inputs from higher centers or from the periphery. It is clear that the respiratory, masticatory, and swallowing CPGs are separate entities, but we do not know yet whether suckling is controlled by a separate CPG or by the CPG that also controls mastication. Rhythmical sucking begins *in utero*, while mastication makes its appearance with the posnatal eruption of the teeth and with the maturation of key neurons in the evolving masticatory CPG.

The masticatory CPG can be activated by tonic stimulation of oral sensory afferents and of inputs from the sensorimotor cortex and other forebrain and midbrain structures (also see the section titled 'Higher center mechanisms'). Pattern generation is usually studied in paralyzed anesthetized animals in which the motor output is recorded from motor neurons ('fictive mastication'). During fictive mastication, JO motor neurons fire bursts of action potentials at very high frequency (up to 250 Hz) that are superimposed on long rhythmical depolarizing potentials. The membrane potential returns to resting levels during the JC phase. Similar slow potentials depolarize JC motor neurons during their active phase, but these cause only weak firing. Then, JC motor neurons are hyperpolarized during the JO phase by glycine. These important differences in membrane potential and in firing rate between the two motor neuron pools reflect the relative importance of sensory feedback in controlling their outputs (see later).

The essential parts of the masticatory CPG are found between the rostral poles of Motor V and the facial motor nucleus. The left and right sides of the CPG are connected by axons that cross the midline at several levels, but, like the spinal cord locomotor CPG, each side can generate a rhythm if they are surgically separated. The CPG is made up of neurons within the regions of the caudal pons and rostral medulla that contain the V premotor interneurons. These nuclei also project to motor neurons in the facial and hypoglossal nuclei. Many neurons in these regions change their pattern of firing during fictive mastication. Most of these 'masticatory' neurons receive sensory inputs from oral mechanoreceptors or muscle spindles and from the sensorimotor cortex and subcortical areas, as noted later. Some of them fire tonically during fictive mastication, but most generate rhythmical bursts of action potentials in phase with bursts in either the JC or JO motor neurons, but not both. The CPG also causes presynaptic depolarization of the terminals of jaw muscle spindle afferents within the lateral RF and spinal V nuclei during the JC phase, and this causes bursts of antidromic action potentials that appear to enter branches in Motor V and to add to the depolarization of JC motor neurons. However, GABAergic synapses on the axons themselves prevent propagation into Mes V.

The medial RF at the level of Motor V also seems to participate in the control of mastication. Many neurons in its dorsal half fire in either the JO or JC phases, while most ventral neurons fire tonically. These tonic neurons appear to inhibit the lateral subgroups during mastication, while the phasic dorsal neurons have mixed effects.

The way in which the final masticatory pattern is generated is only beginning to be understood. The pattern varies greatly between foods, and also from the start of a chewing sequence to the end. By comparing activity generated in two or more patterns, we now know that the CPG contains subpopulations of neurons, few of which are active during all motor patterns. For instance, about half of the neurons in the lateral RF and V spinal nucleus participate in grinding on the ipsilateral posterior teeth but drop out during grinding on the contralateral side. Those that participate in both patterns initially change their firing pattern. This data show that at least two mechanisms are in play during the genesis of different patterns of mastication: addition and subtraction of neurons from the active pool, and changes in firing frequency of common elements. The extensive interconnections between the various subgroups of interneurons described previously must be important in this reorganization, and evidence from other CPGs

suggests that the network could also participate in generating the masticatory rhythm.

The rhythm of respiration arises from neurons that show repetitive low-frequency oscillations of their membrane potentials even when isolated from all inputs, and a subpopulation of masticatory CPG neurons in dorsal Main V shows this feature under certain conditions. This ability to burst develops postnatally in Main V, in parallel with the emergence of mastication. It depends on a persistent Na^+ current (INap) that appears to be modulated by the extracellular Ca^{2+} concentration.

Reflexes

Two major reflexes can be triggered with the jaw at rest; one that opens the jaw and one that closes it. The jaw-closing reflex (JC reflex) is the myotatic (stretch) reflex of the JC muscles, in which stretch of muscle spindle afferents activates JC motor neurons through monosynaptic and polysynaptic pathways. The JC reflex provides rapid compensation for changes in the resistance of food during mastication. It also modulates JC muscle tension in response to changes in the gravitational load on the mandible during locomotion.

The jaw-opening reflex (JO reflex) is the V equivalent of the flexion-withdrawal reflex of the limbs. It is a disynaptic reflex that can be triggered by two types of orofacial stimuli: tapping on the lips, teeth, and palate or noxious stimulation of the face, intraoral tissues, muscle, and joints. In animals, JO motor neurons are excited and JC motor neurons inhibited by glycinergic interneurons at the same time. In humans, activation of JO motor neurons is almost absent, but the inhibition of JC motor neurons is very strong.

The Importance of Sensory Feedback during Mastication

The previous sections on reflexes and CPG circuits mention the effects of sensory inputs. Here we consider these in further detail. During mastication, several groups of afferents provide feedback to motor neurons, the CPG, and higher centers, and the pattern of activity that emerges from the motor neurons is the result of these interactions. The effects of three groups, muscle spindles, periodontal mechanoreceptors, and nociceptors, have been well characterized.

The muscle spindles in the JC muscles are stretched during the JO phase of mastication, which causes them to fire at high frequency. However, this does not cause the JC motor neurons to fire because they are hyperpolarized by the CPG (see the section titled 'Brain stem pattern-generating circuits'). During the JC phase, the muscle spindles fire strongly when food between the teeth decelerates the mandible. Since JC motor neurons are now depolarized by the CPG, reflex inputs from the spindles increase firing frequency and burst duration.

Periodontal mechanoreceptors around the roots of the teeth respond to loads generated between the upper and lower jaws. Although tapping on the teeth can cause the JO reflex with the jaw at rest (see the section titled 'Reflexes'), the heavy pressures generated during the JC phase of mastication do not inhibit JC muscles. Instead, these mechanoreceptors cause the JC phase to lengthen and JC motor neurons to fire at higher frequency. It appears that the CPG shuts down the inhibitory pathway and opens up an excitatory one during mastication, as happens with Golgi tendon organ pathways during locomotion. The harder the food, the greater the effect of positive feedback from the periodontal receptors and muscle spindles.

The CPG also modulates inputs from other oral mechanoreceptors and particularly from nociceptors. Nociceptors are frequently stimulated by biting the tongue or cheek, or by objects (e.g., bones) in food. When this happens, JC motor neurons are quickly turned off (<10 ms) and the CPG inhibited. Transmission through this short inhibitory pathway is facilitated by the CPG during the JC phase, which increases the protection of soft tissues at the time that they are most vulnerable.

Higher Center Mechanisms

The V brain stem neurons relaying to thalamus or other brain regions such as the cranial nerve motor nuclei are subject to segmental (afferent) modulation and to descending modulation from various cortical and subcortical structures. Because many of these neurons contribute as interneurons to jaw reflexes and other motor behaviors evoked by stimulation of orofacial tissues, these reflex responses can also be regulated by these modulatory influences. The descending modulatory inputs influence several of the brain stem interneuronal pools mentioned previously, and some (e.g., from sensorimotor cerebral cortex) may even have direct input to Motor V in humans.

Subcortical processes One very important set of subcortical influences on the V motor system derives from various components of the basal ganglia and substantia nigra, as exemplified by the jaw motor deficits or abnormal motor behaviors (e.g., oral dyskinesia) that can result from damage to these structures or from alterations in key neurochemical processes (e.g., dopamine, ACh, GABA, NMDA) within them. Other subcortical areas that can modulate the V motor system include the amygdala, subthalamic nucleus,

nucleus accumbens, hypothalamus, anterior pretectal nucleus, red nucleus, periaqueductal gray, raphé nuclei, and cerebellum. In addition to dopamine, ACh, GABA, and NMDA receptor mechanisms mentioned previously, catecholamine, vasopressin, opioid, and serotonergic-based mechanisms in particular contribute to these various subcortical influences on the V motor system. These features emphasize the complexity of the subcortical control circuits and neurochemical mechanisms involved in jaw motor control, as in the spinal motor control systems.

Cortical processes Several cortical areas exert descending influences on brain stem and other subcortical regions. Imaging and transcranial magnetic stimulation studies in humans have reported that jaw movements may be associated with activation of several cortical areas, e.g., the face primary motor cortex (MI), primary somatosensory area (SI), premotor cortex, supplementary motor area, cortical masticatory area (CMA), insula, and anterior cingulate gyrus. This is consistent with studies in monkeys and subprimates which have shown that intracortical microstimulation (ICMS) can evoke orofacial movements from some of these sites and that cold block or ablation especially of face MI or CMA can disrupt the movements.

The ICMS studies have also revealed that each orofacial muscle or movement is represented multiple times within face MI, including elemental movements (e.g., jaw opening) and more complex activities such as mastication and swallowing, suggesting that each MI output zone controls one of the many contextual functions in which a jaw muscle participates. Furthermore, the cold block or ablation of face MI or CMA may disrupt chewing and swallowing to varying degrees. This indicates that each may be involved differentially in the production and patterning of chewing and swallowing, including the coordination of the tongue, facial, and jaw movements that are necessary for the proper ingestion and transport of food or liquid, and is consistent with studies showing that many face MI and CMA neurons discharge in relation to chewing or swallowing in monkeys or subprimates. Most MI neurons also discharge in association with more elemental orofacial movements. In addition, many face SI neurons also discharge in relation to these movements, including jaw movements. These findings emphasize the importance of the face SI as well as the face MI in the fine motor control of jaw movements. Furthermore, face MI and SI, and perhaps also the CMA, may use the orofacial afferent inputs that they receive to send corticofugal signals directly or indirectly (e.g., via the CPGs) to the V motor neurons and thereby guide, correct, and control movement by the use of sensory cues prior to movement, and by using sensory information generated during movement. Recent findings also implicate face MI in orofacial motor skill acquisition and adaptation to intraoral alterations.

In summary, the V motor system is involved in many complex motor behaviors as well as in simpler behaviors that are either reflex in nature or repetitive. The simple behaviors are coordinated in the brain stem, whereas the more complex motor patterns may involve cortical as well as subcortical regions. Sensory feedback interacts with the central motor programs at all levels of the CNS.

See also: Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Orofacial Motor Control; Pattern Generation; Speech Production: Adult; Speech Production: Development; Spinal Motor Neurons: Properties; Trigeminal Neuralgia.

Further Reading

Bourque MJ and Kolta A (2001) Properties and interconnections of trigeminal interneurons of the lateral pontine reticular formation in the rat. *Journal of Neurophysiology* 86: 2583–2596.

Hannam AG and McMillan AS (1994) Internal organization in the human jaw muscles. *Critical Reviews of Oral Biology and Medicine* 5: 55–89.

Lund JP and Kolta A (2006) Generation of the central masticatory pattern and its modification by sensory feedback. *Dysphagia* 21: 167–174.

Miles TS, Nauntofte B, and Svensson P (2004) *Clinical Oral Physiology.* Copenhagen: Quintessence.

Nakamura Y, Katakura N, Nakajima M, and Liu J (2004) Rhythm generation for food-ingestive movements. *Progress in Brain Research* 143: 97–103.

Sessle BJ, Adachi K, Avivi-Arber L, et al. (2007) Neuroplasticity of face primary motor cortex control of orofacial movements. *Archives of Oral Biology* 52: 334–337.

Tsuboi A, Kolta A, Chen CC, and Lund JP (2003) Neurons of the trigeminal main sensory nucleus participate in the generation of rhythmic motor patterns. *European Journal of Neuroscience* 17: 229–238.

Verdier D, Lund JP, and Kolta A (2003) GABAergic control of action potential propagation along axonal branches of mammalian sensory neurons. *Journal of Neuroscience* 23: 2002–2007.

Trigeminal Neuralgia

S Love and N K Patel, University of Bristol Institute of Clinical Neurosciences, Bristol, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Trigeminal neuralgia (TN) is a syndrome of intermittent facial pain. Because the pain is so severe that it often induces ticlike grimacing, trigeminal neuralgia is also known as tic douloureux (the term was coined by Nicholas André, in his book, *Observations Pratiques sur les Maladies de l'Urethre et sur Plusieurs Faits*, in 1756). However, it was not until 1773 at the Medical Society of London that John Fothergill gave a full and accurate description, and at times the condition has been named Fothergill's disease. In 1999, Rose credits the earliest written description of trigeminal neuralgia to Aretaeus of Cappadocia, in the second century AD.

Clinical Features

The International Association for the Study of Pain defines TN as "a sudden, usually unilateral, brief stabbing recurrent pain in the distribution of one of the branches of the fifth cranial nerve." The incidence of TN is about 4.5 per 100 000 per annum. It is slightly more common in women. The mean age of onset is about 50 years; however, the incidence curve is skewed to the right and the peak incidence of TN is at the age of 60s and 70s. Young adults and children are occasionally affected. Although familial examples of TN are reported, the large majority of cases occur spontaneously. The attacks are characterized by intense unilateral facial pain, often described as stabbing, and sometimes likened to electric shocks. It is localized to areas of the face within the territory of the trigeminal nerve. The pain most often affects the maxillary and mandibular divisions of the trigeminal nerve and the right side is more often involved than the left. Individual attacks usually last less than a minute but tend to recur in bouts that last weeks or months. During bouts, attacks may be relatively infrequent but can occur as often as every few minutes, remitting only during sleep. The attacks may be spontaneous but in most instances are initiated by mild stimulation of small areas of skin or mucosa within the territory of the affected trigeminal nerve. Typical stimuli include light touching, drafts of wind, eating, drinking, washing, brushing of teeth, shaving, and applying makeup. Very rarely attacks are triggered by loud noises, bright lights, or stimuli affecting cranial nerves other than the trigeminal. Between attacks, patients are normal on neurological examination unless there is associated disease, such as multiple sclerosis (see later). Between bouts, patients experience spontaneous remissions that may last months or years and during which patients are completely free of symptoms. Over the years, however, the attacks often become more frequent and the pain more sustained and less responsive to medication.

Etiology of TN

Compression of the Trigeminal Nerve Root

In 80–90% of patients, TN is caused by compression of the trigeminal nerve root within its proximal few millimeters (the root entry zone) by a loop of artery or vein. This part of the trigeminal nerve root consists of central white matter (i.e., nerve fibers that are myelinated by oligodendrocytes) (Figure 1). Recent studies have shown that the compressed nerve fibers undergo demyelination and tend to be in close juxtaposition (Figures 2 and 3). This arrangement is conducive to both spontaneous impulse activity and ephaptic cross-talk (i.e., nonsynaptic spread of excitation from one axon to another adjacent axon, as a result of the extracellular current flux that is associated with depolarization of the first axon). Ephaptic cross-talk between demyelinated fibers subserving light touch (excited by stimulation of mucocutaneous trigger zones) and those involved in the generation of pain has been suggested to underlie the development of TN in most patients with nerve root demyelination.

Several syndromes involving hyperactivity and abnormal spread of activity within the distribution network of innervation of cranial nerves VII–XII are now known often to be associated with vascular compression of the relevant nerve root. These include hemifacial spasm, glossopharyngeal neuralgia, the rarer syndromes of superior oblique myokymia, vestibular paroxysmia and geniculate neuralgia, and some cases of spasmodic torticollis. In addition, hypertension has been associated with arterial compression of the ventrolateral aspect of the rostral medulla in the region of entry of the left glossopharyngeal and vagus nerve roots. Multiple syndromes of cranial nerve hyperactivity due to vascular compression of nerve roots are sometimes present in combination.

Although a loop of blood vessel is usually to blame, there are other occasional causes of compression-induced TN and other related syndromes of cranial nerve hyperactivity. These include local vascular

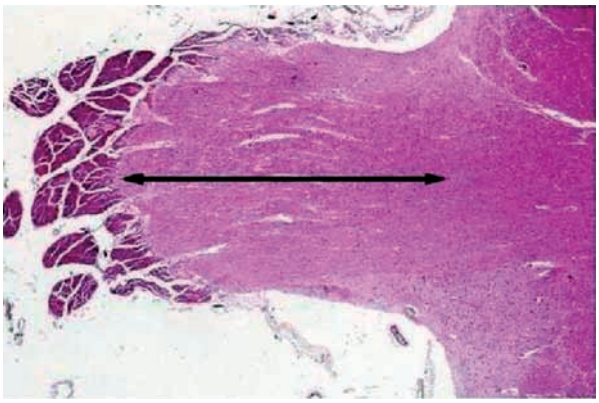

Figure 1 Normal trigeminal nerve root. Note that the proximal part of the root (arrow) consists of central nervous system white matter and extends several millimeters away from the pons (to the right).

malformations, cysts, neoplasms (particularly cholesteatomas, meningiomas, and schwannomas of the eighth nerve), and deformities of the skull base.

Primary Demyelination of the Trigeminal Nerve Root

After vascular compression of the trigeminal nerve root, the next most frequent cause of TN is multiple sclerosis. This accounts for about 2.5% of patients with trigeminal neuralgia, and a relatively higher proportion of those at 30–40 years of age. Conversely, the prevalence of TN among patients with multiple sclerosis is of the order of 2–4%. Bilateral TN is most often caused by multiple sclerosis. The TN in multiple sclerosis is usually associated with demyelination in the proximal part of the trigeminal nerve root or the immediately adjacent part of the pons. Histology has revealed the presence of clusters of juxtaposed axons without intervening glial processes (as also occurs in patients whose TN is due to vascular compression of the trigeminal nerve root). Rarely, patients with peripheral nerve demyelination due to Charcot–Marie–Tooth disease develop trigeminal neuralgia.

Other Causes of TN

TN is very occasionally a manifestation of neoplastic infiltration of the trigeminal nerve root, gasserian ganglion or nerve, or deposition of amyloid within these structures. A few patients have nondemyelinating lesions of the pons – ischemic, hamartomatous, or neoplastic. Rarely, TN has affected several members of the same family.

Idiopathic TN

In about 5–15% of patients the anatomical or pathological substrate of TN remains unclear, even after extensive investigation that includes direct endoscopic inspection of the nerve root.

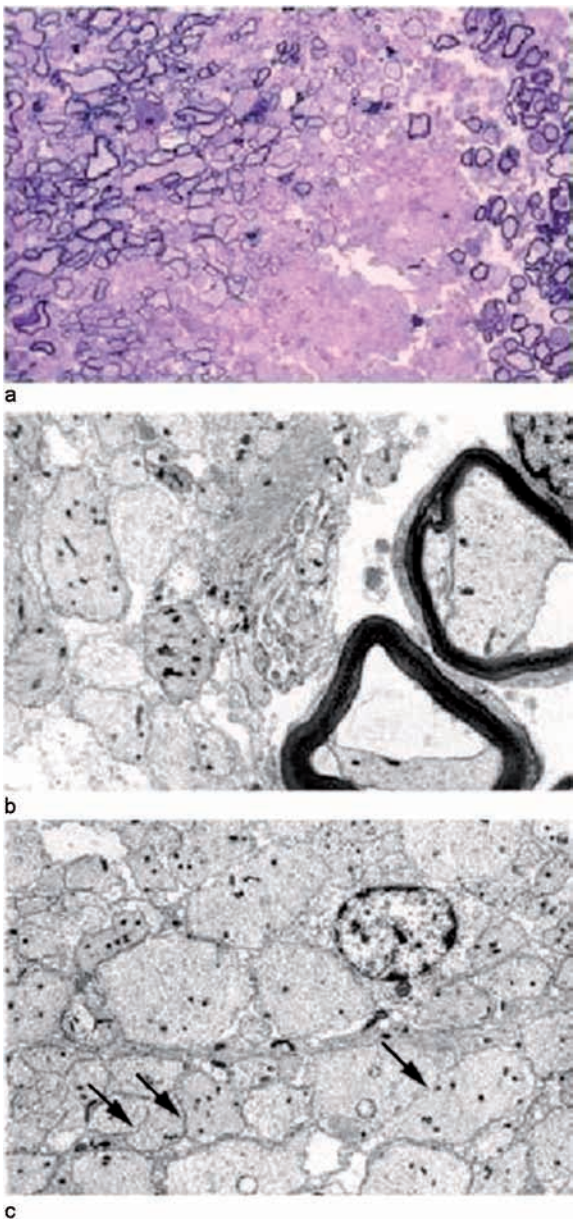

Figure 2 Focal demyelination of the trigeminal nerve root in a patient with TN due to vascular compression. In this case, the compressing blood vessel could not safely be mobilized. A partial rhizotomy was therefore performed, during which a small wedge of tissue was excised from the caudal part of the nerve root. The tissue was fixed in glutaraldehyde for resin histology and electron microscopy. (a) Semithin resin section through the region of compression, which was within the proximal (central nervous system) part of the nerve root but close to its junction with the peripheral nervous system. The section includes a wedge-shaped region of demyelination, to the left of which are many thinly remyelinated central nervous system nerve fibers. To the right of the section are some normally myelinated nerve fibers that are within the distal (peripheral nervous system) part of the nerve root. (b) Electron photomicrograph of an ultrathin section through the junction between the demyelinated central nervous system (to the left) and normally myelinated peripheral nervous system (to the right) parts of the nerve root. (c) Electron photomicrograph showing many of the axons within the region of demyelination to be in direct apposition (arrows point to some). The section includes a single oligodendrocyte nucleus.

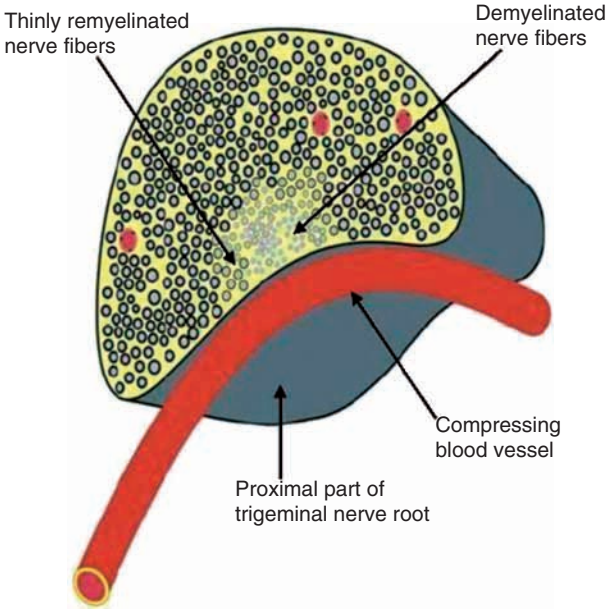

Figure 3 Schematic illustration of the pathological findings in trigeminal neuralgia due to vascular compression.

Investigation

A diagnosis of TN can usually be made solely on the basis of a typical clinical history and an absence of neurological abnormalities on physical examination. Magnetic resonance imaging (MRI) of the brain is helpful in excluding associated disease such as multiple sclerosis, and is advisable in patients who present before 55 years of age or with atypical features, or who fail to respond to conventional medical treatment (see later). With advances in MRI, it is now often possible to demonstrate neurovascular compression preoperatively; oblique sagittal images are often helpful. The compressive vessel can be detected using routine MRI sequences and magnetic resonance angiography (MRA) with gadolinium enhancement (Figure 4).

Treatment

Most patients respond very well to carbamazepine. The starting dose of carbamazepine is 100 mg once or twice daily, and this is often effective, at least initially; however, the daily dose may need to be increased by up to 1.2 g, to control the acute bouts of TN. Over a period of several years, patients often become less responsive to carbamazepine and require further drugs, often in combination. The principal second-line treatments are the muscle relaxant baclofen and the anticonvulsant drugs gabapentin, lamotrigine, and phenytoin.

Several types of physical intervention have been used in patients who become refractory to medical

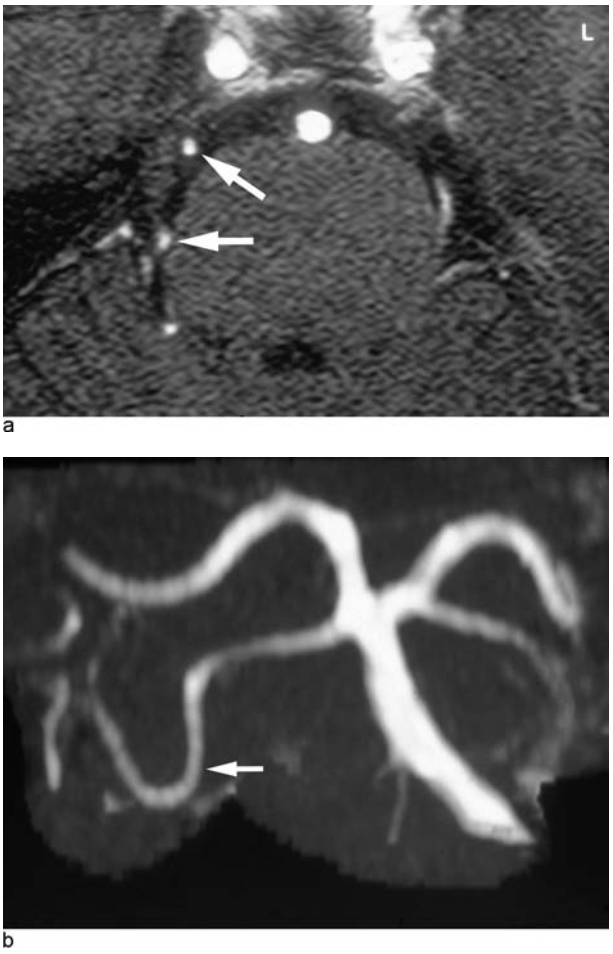

Figure 4 Demonstration of a compressing arterial loop by magnetic resonance angiography. (a) Fast field echo imaging of the posterior fossa in the axial plane shows an arterial loop (arrows) medially compressing the right trigeminal nerve in the root entry zone. (b) A reconstructed magnetic resonance angiogram shows the arterial loop (arrow) responsible for right trigeminal nerve root compression.

treatment. In many centers, microvascular decompression is regarded as the treatment of choice for medically intractable trigeminal neuralgia. This procedure is performed under general anesthesia and involves a retromastoid craniotomy, through which the trigeminal nerve root is exposed. Any compressing blood vessel is mobilized and is either separated from the root by Teflon felt or displaced and fixed to the dura by Teflon slings (Figure 4). This relieves pain in 80–90% of patients, although about 20% will need some continuing medication. Over a 10-year period approximately 10–15% will experience recurrent TN and may need further microvascular decompression or some other therapeutic procedure (see later). The mortality is generally under 0.5% but morbidity, predominantly from cerebrospinal fluid leakage and infection, ranges from 1% to 5%. Facial numbness or paresthesias are rare after microvascular decompression.

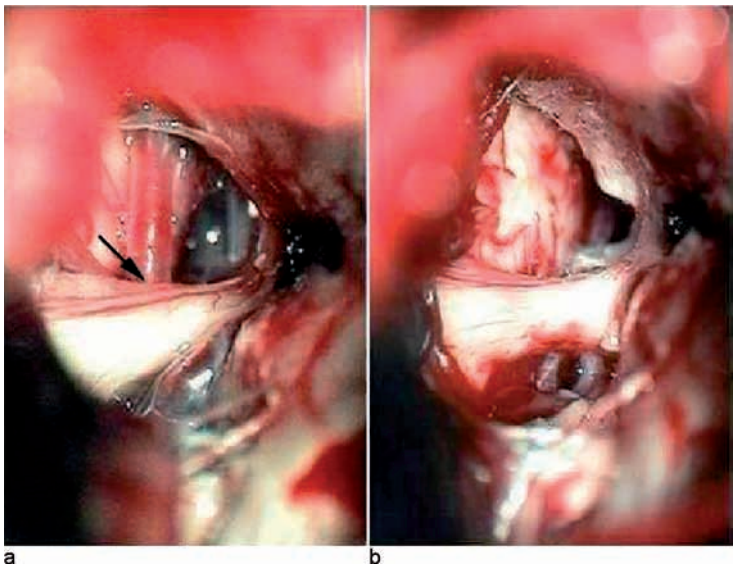

Figure 5 Microvascular decompression of the trigeminal nerve root. (a) Posterolateral exposure reveals a loop of artery that is pressing (arrow) against the anterior aspect of the nerve root. (b) The artery has been mobilized, transposed, and fixed to the dura with Teflon slings (toward the top right of the image) and fibrin glue. Intraoperative photographs courtesy of Professor Hugh Coakham, Department of Neurosurgery, Institute of Clinical Neurosciences, Frenchay Hospital.

The other types of physical intervention for TN are various forms of ablative procedure. These are suitable for the treatment of patients with medically intractable TN in whom (1) no compressing blood vessel is demonstrable, (2) symptoms persist despite adequate microvascular decompression, (3) a compressing blood vessel is present but not technically amenable to safe repositioning, (4) there are medical contraindications to surgery under general anesthesia, (5) life expectancy is less than 10–15 years, and (6) appropriate surgical expertise is not available for microvascular decompression.

Numerous types of ablative procedures have been used in the past. The ablative techniques most commonly used nowadays can be subdivided into the percutaneous procedures (percutaneous radiofrequency thermocoagulation, injection of glycerol, or balloon compression), stereotactic radiosurgery, and partial rhizotomy. The percutaneous procedures have an initial effectiveness of over 90% but carry a high risk of facial numbness and paresthesias. In most series, the recurrence rate has been considerably higher than that after microvascular decompression. Radiosurgery is used increasingly . Its initial effectiveness of more than 90% is consistent with that of other ablative procedures, but there is still a paucity of good follow-up data; the available information suggests that up to 50% of patients will experience some relapse of symptoms but that in about two-thirds of cases the degree of pain relief is nonetheless sufficient to avoid the need for continuing medical treatment. The risk of facial numbness or paresthesias is about 10%. Partial rhizotomy carries the same general anesthetic and neurosurgical risks as does microvascular decompression (Figure 5) and produces facial numbness in a relatively high proportion of patients. However, it is successful in relieving the neuralgia in the majority of patients and is often used when intraoperative inspection of the trigeminal nerve root either fails to reveal a compressing blood vessel or shows one to be present but not technically amenable to safe repositioning.

See also: Brainstem and Cranial Nerves; Demyelinating Diseases; Demyelination and Demyelinating Antibodies; Neuropathic Pain; Orofacial Motor Control; Sensation from the Face; Trigeminal Motor System.

Further Reading

Devor M, Govrin-Lippmann R, and Rappaport ZH (2002) Mechanism of trigeminal neuralgia: An ultrastructural analysis of trigeminal root specimens obtained during microvascular decompression surgery. *Journal of Neurosurgery* 96: 532–543.

Janetta PJ (1967) Arterial compression of the trigeminal nerve at the pons in patients with trigeminal neuralgia. *Journal of Neurosurgery* 26(supplement 1): 59–62.

Jannetta PJ (1980) Neurovascular compression in cranial nerve and systemic disease. *Annals of Surgery* 192: 518–525.

Jensen TS, Rasmussen P, and Reske-Nielsen E (1982) Association of trigeminal neuralgia with multiple sclerosis: Clinical and pathological features. *Acta Neurologica Scandinavica* 65: 182–189.

Love S and Coakham HB (2001) Trigeminal neuralgia. Pathology and pathogenesis. *Brain* 124: 2347–2360.

Love S, Hilton DA, and Coakham HB (1998) Central demyelination of the Vth nerve root in trigeminal neuralgia associated with vascular compression. *Brain Pathology* 8: 1–11.

Love S, Gradidge T, and Coakham HB (2001) Trigeminal neuralgia due to multiple sclerosis: Ultrastructural findings in trigeminal rhizotomy specimens. *Neuropathology and Applied Neurobiology* 27: 238–244.

Merskey H and Bogduk N (1994) *Classification of Chronic Pain: Descriptions of Chronic Pain Syndromes and Definitions of Pain Terms*. Seattle: IASP Press.

Rose C (1999) Trigeminal neuralgia. *Archives of Neurology* 56: 1163–1164.

Nurmikko TJ and Eldridge PR (2001) Trigeminal neuralgia – pathophysiology, diagnosis and current treatment. *British Journal of Anaesthesia* 87: 117–132.

Patel NK, Aquilina K, Clarke Y, et al. (2003) How accurate is magnetic resonance angiography in predicting neurovascular compression in patients with trigeminal neuralgia? A prospective, single-blinded comparative study. *British Journal of Neurosurgery* 17: 60–64.

Sweet WH (1985) The history of the development of treatment for trigeminal neuralgia. *Clinical Neurosurgery* 32: 294–318.

Triplicate Repeats: Huntington's disease

J-H J Cha and K B Kegel, MassGeneral Institute for Neurodegenerative Disease, Charlestown, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In 1872, the physician George Huntington described the neurologic disorder that has come to bear his name. This disorder was originally called Huntington's chorea, from the Greek word for dance, referring to the dancelike adventitious movements that are the hallmark of the disorder. Since chorea is not a universal feature, especially in the juvenile forms of the disease, the term Huntington's disease (HD) is now used. Huntington's original report noted the familial inheritance, the adult onset, the constellation of emotional, cognitive, and motor symptoms, and the inexorably progressive and ultimately fatal nature of the disease. Since that time, Huntington's disease has been a paradigmatic neurodegenerative disorder, with advances in HD at the forefront of molecular genetics, of ethical issues around genetic testing, as well as of many important concepts in modern neuroscience.

Huntington's Disease Gene

Recombinant DNA techniques became available in the late 1970s, and their power was brought to bear on the search for the Huntington's disease gene. In 1983, using anonymous DNA markers spread over the human genome, David Housman at the Massachusetts Institute of Technology and James Gusella at Massachusetts General Hospital used DNA samples from the many members of a large Venezuelan family to localize the *HD* gene to the long arm of chromosome 4. In 1993, the Huntington's Disease Collaborative Research Group, a consortium of gene hunter laboratories from around the world, isolated a gene originally named *IT15* ('interesting transcript 15'); it was subsequently renamed *HD*.

The defect in *HD* was found to reside in an unstable region of the gene, in that a cytosine–adenine–guanosine (CAG) trinucleotide repeat present in normal alleles was abnormally expanded in the alleles of HD patients. CAG is the codon for glutamine, and this trinucleotide repeat gives rise to a polyglutamine moiety within the huntingtin protein. Normal huntingtin alleles contain from 6 to 35 CAG repeats, giving rise to 6–35 glutamines in the mature protein, and variation occurs in the normal human population. Patients with Huntington's disease invariably have alleles with more than 35 repeats, and greater than 40 repeats invariably gives rise to HD. There is a 'gray area,' between 35 and 39 repeats, where some uncertainty exists. Some patients with up to 39 repeats have lived into their 70s without developing overt signs of HD, although age of onset as late as 80 years has been described.

Age of onset correlates negatively with repeat length, although the correlation is strongest for high CAG repeat numbers. That is, although repeat numbers of greater than 70 invariably produce juvenile onset of HD, more common repeat numbers – for example, in the 40–45 range – produce a varied age of onset. Neuropathologic grade also varies with CAG number, with the most damage seen in brains with the highest CAG repeats. However, CAG repeat length accounts for only about 50% of the variability of age of onset, suggesting that there are other genetic or environmental influences on age of onset.

HD thus belongs to a novel family of neurologic diseases, the trinucleotide repeat disorders. Certain other diseases, such as myotonic dystrophy, fragile X syndrome, and Friedreich's ataxia, are characterized by trinucleotide repeat expansions in noncoding regions of the genes. In contrast, HD belongs to a family of disorders characterized by expansion of a CAG trinucleotide motif within the coding region of the gene. The CAG repeat disorders, or polyglutamine disorders, display common features, including dominant inheritance, adult onset, threshold levels of approximately 37 repeats, and progressive neurodegeneration (Table 1).

Clinical Features

Huntington's disease is typically an adult-onset disorder characterized by insidious onset of both neurologic and psychiatric symptoms. In the United States, approximately 25 000 persons are affected by HD (about 10 per 100 000 population) and approximately 150 000 persons are at 50% risk for the disease by virtue of having an affected parent. Symptoms usually begin in the mid-30s to mid-40s, although disease onset can range from as young as 2 years or as old as 80 years. Initial symptoms include personality change and the gradual appearance of small involuntary movements. Symptoms progress, with chorea becoming more obvious and incapacitating. Over years, motor symptoms worsen such that walking

Table 1 CAG trinucleotide repeat disorders

Disease	*Protein*	*Inheritance*	*Normal repeat number*	*Range of CAG repeats in disease*
Spinobulbar muscular atrophy (Kennedy's disease)	Androgen receptor	X-linked dominant	11–33	40–62
Huntington's disease	Huntingtin	Autosomal dominant	1–39	36–121
Spinocerebellar ataxia, type 1 (SCA1)	Ataxin-1	Autosomal dominant	<29–36	43–60+
Spinocerebellar ataxia, type 2 (SCA2)	Ataxin-2	Autosomal dominant	15–24	35–59
Spinocerebellar ataxia, type 3 (SCA3; Machado–Joseph disease)	Ataxin-3	Autosomal dominant	13–36	62–82
Spinocerebellar ataxia, type 6 (SCA6)	a_{1a} subunit of voltage-dependent calcium channel	Autosomal dominant	4–16	21–27
Spinocerebellar ataxia, type 7 (SCA7)	Ataxin-7	Autosomal dominant	7–17	38–130
Dentatorubral-pallidoluysian atrophy (DRPLA)	Atrophin-1	Autosomal dominant	7–25	49–85
Spinocerebellar ataxia, type 17 (SCA17)	TATA-binding protein (TBP)	Autosomal dominant	<42	≥43

becomes more difficult, as do speaking and eating. Weight loss is common, partially due to the extra energy required for adventitious movements but also due to increased resting basal energy expenditure. Most HD patients eventually succumb to aspiration pneumonia, due to swallowing difficulties.

About 10% of cases start before the age of 20 years. The juvenile form (Westphal variant) is more parkinsonian in nature. Rather than chorea, the prominent features are bradykinesia, rigidity, and tremor. Seizures can be present in juvenile patients. Juvenile-onset HD usually results from paternal genetic transmission. For individuals who develop symptoms before age 10 years, more than 90% have an affected father. This tendency for 'anticipation' – younger age of onset in successive generations – is especially pronounced in cases of paternal transmission. This phenomenon is not completely understood but may be the result of further CAG expansion during spermatogenesis. Interestingly, both the juvenile-onset and adult-onset phenotypes can be present within the same family.

New mutations have been described, usually from a parent carrying an 'intermediate allele.' Homozygotes for the *HD* gene have also been described. The clinical symptomatology manifested by these very few patients appears to be no worse than that manifested by patients carrying only a single *HD* gene. Thus, HD appears to be a truly genetically dominant disease.

Pathology

Neuropathologically, HD is characterized by gross atrophy of the basal ganglia, specifically the striatum (caudate-putamen), although pathologic changes have also been described in cortex, thalamus, and subthalamic nucleus. Atrophy and gliosis of the caudate nucleus and putamen are progressive and marked. In juvenile-onset cases, cell loss tends to be more pronounced, and can also be seen in the cerebellum. Within the striatum, HD differentially affects subpopulations of neurons, with projection neurons preferentially being lost rather than interneurons. The γ-aminobutyric acid (GABA)-containing medium spiny neurons, which are the output projection neurons, comprise approximately 90% of striatal neurons. HD seems to spare the interneurons while devastating the projection neurons. Consistent with a loss of projection neurons was the early finding that GABA levels were markedly reduced in the caudate-putamen of HD patients.

Of the two populations of striatal projection neurons, the neurons of the indirect pathway (i.e., enkephalin/GABA-containing neurons) are affected first, thus providing an anatomical explanation for the increased movement which is the hallmark of HD. Consistent with this observation is the fact that in early-grade HD cases, markers for this population of striatal neurons are decreased, including dopamine D2 receptors, adenosine A_{2a} receptors, and enkephalin. In early HD, thus, the indirect pathway is predominantly disrupted, and basal ganglia circuitry models predict an overall increase in movement, manifested as chorea and ballism (Figure 1). In later stages of adult HD, both populations of striatal projection neurons are affected, with concomitant loss of markers of the direct pathway (i.e., substance P/dynorphin/GABA-containing neurons), including dopamine D1 receptors and substance P. The functional correlate of degeneration of both direct and indirect pathways is a rigid bradykinetic state. In

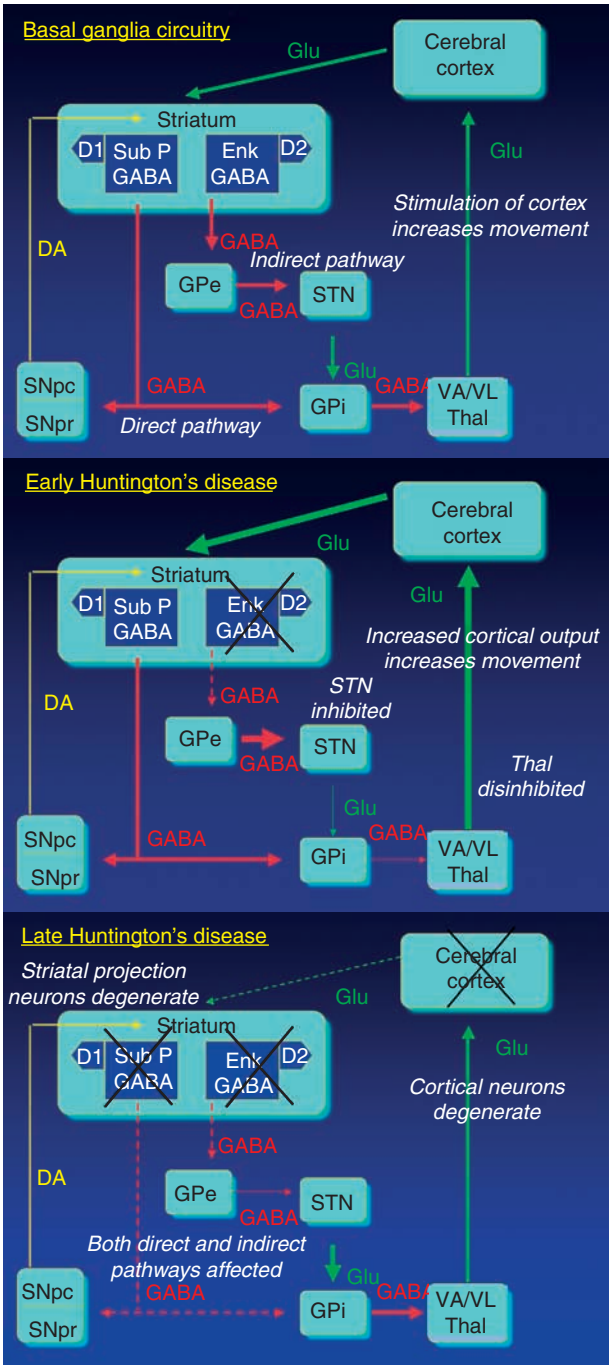

Figure 1 Basal ganglia circuitry models of early and late Huntington's disease. Glu, glutamate; D1/D2, dopamine D1/D2 receptors; Sub P, substance P; Enk, enkephalin; GABA, γ-aminobutyric acid; DA, dopamine; VA/VL Thal, ventral anterior/ventral lateral thalamic nuclei; STN, subthalamic nucleus.

juvenile HD cases, which resemble Parkinson's disease, degeneration of both direct and indirect pathway striatal neurons is seen concurrently.

The mode by which neurons die in HD is unknown. One theory is that death in neurons occurs through apoptosis – that is, activation of a programmed cell death pathway. DNA fragmentation characteristic of apoptosis has been described in HD brain. However, it is not yet clear whether apoptosis is the primary pathologic event, or merely the final mode of exit for neurons which have been long compromised by some other process.

Theories of HD Pathogenesis: Excitotoxicity and Mitochondrial Dysfunction

Prior to the discovery of the *HD* gene, the leading hypotheses concerning the pathogenesis of HD implicated either excitotoxicity or energetic dysfunction. Evidence for both theories has been demonstrated in both human and animal model studies of HD. Discovery of the *HD* gene in 1993 did not reveal the pathogenetic mechanism of HD. The novel protein huntingtin had no clear relationship to excitatory amino acid neurotransmission, nor to cellular or mitochondrial energetics.

Excitotoxicity is the process in which neuronal cells die as a result of excessive excitatory amino acid neurotransmission. Excitatory amino acids, especially glutamate, have been postulated to play a role in the pathogenesis of HD because intrastriatal injections of excitatory amino acid receptor agonists, particularly agonists acting at the *N*-methyl-D-aspartate subtype of glutamate receptor, reproduce the neuropathological features of HD. Excessive glutamate has been postulated to kill neurons in a number of neurological disorders, including hypoxia–ischemia, head trauma, epilepsy, schizophrenia, and neurodegenerative disorders such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis. Postmortem analyses of HD brains demonstrate reduced glutamate receptors that might represent an effort by cells to reduce glutamate receptor activation and subsequent calcium influx, although this decrease in receptors may also reflect generalized neuronal loss.

Mitochondrial dysfunction has also been implicated as a pathologic mechanism in HD, potentially rendering cells vulnerable to normal ambient levels of extracellular glutamate. Positron emission tomography studies have demonstrated reduced striatal glucose metabolism in HD patients. Magnetic resonance imaging (MRI) spectroscopy shows increased levels of striatal glutamate/glutamine and lactate in HD patients, suggesting that glutamatergic function and abnormalities in energy metabolism may combine to produce pathology. Defects in mitochondrial metabolism have been demonstrated in HD brain tissue, especially in complex II–complex III activity. Further, intrastriatal injection of the complex II inhibitor malonate produces lesions reminiscent of HD. Systemic administration of another

complex II inhibitor, 3-nitropropionic acid (3-NP), produces strikingly focal striatal lesions in rodents and primates, also reminiscent of HD striatal pathology. Currently, clinical trials are exploring the effects of coenzyme Q_{10}, a compound that boosts mitochondrial energetics. In addition, creatine, which is important for storage of high-energy phosphate (creatine phosphate) and subsequent production of ATP, has been shown to extend life in transgenic HD mice, and is being explored as a potential therapy for human HD patients.

Huntingtin

HD encodes a novel protein called huntingtin. The huntingtin gene comprises 67 exons, encoding a 3144-amino-acid protein with an expected molecular mass of 348 kDa. Expectations were that the distribution of huntingtin would mirror the characteristic distribution of neuropathological damage. However, huntingtin mRNA is expressed in almost all tissues of the body and homogenously throughout the brain. No difference in expression is seen between HD and control cases. Immunohistochemical studies confirm the widespread neuronal distribution of the huntingtin protein.

Huntingtin plays an important role in early development, since mice with targeted knockout of the huntingtin gene are embryonic lethal. Humans with Wolf–Hirshhorn syndrome, who have a partial deletion of chromosome 4 and as a result have only one copy of the *HD* gene, do not have the a neurologic disease identical to HD. Therefore, having a single copy of a normal *HD* allele does not produce the HD phenotype and is sufficient for normal embryogenesis in humans. The mutation in huntingtin is thought to confer at least one gain of function. Mutant huntingtin possesses at least certain of the functional properties of wild-type huntingtin, in that mutant huntingtin can rescue the embryonic lethal phenotype seen in huntingtin-null knockout mice.

The huntingtin protein contains 'HEAT repeats,' which correspond to tandemly arranged curlicue-like structures that appear to serve as flexible scaffolding on which other components can assemble (the acronym 'HEAT' derives from names of four proteins found to contain the repeats). Huntingtin has been shown to have numerous protein-binding partners, and many of these interactors have differential affinity for the normal and polyglutamine-expanded versions of huntingtin. Numerous cytoskeletal proteins have been identified as huntingtin interactors. In addition, huntingtin binds to a number of transcriptionally active molecules, raising the possibility that mutant huntingtin may perturb the normal transcriptional activities of neurons. In addition to these interactions, nuclear localization and nuclear export domains have been identified within the huntingtin molecule, suggesting that huntingtin may be shuttled in and out of the nucleus.

The normal function of huntingtin is not completely known, although it has been proposed to act as a molecular scaffold at multiple locations within cells. Huntingtin is a soluble protein that associates with membranes. It can be transported by fast axonal transport, and may be involved in endocytosis and membrane recycling. Aside from functioning as a molecular scaffold, a number of other normal functions have been proposed for normal versions of huntingtin. Wild-type (normal) huntingtin has been found to have antiapoptotic properties. In addition, a permissive role in brain-derived neurotrophic factor (BDNF) transcription has been proposed. Wild-type huntingtin has been found to interact with the transcriptional repressor RE1-silencing transcription factor (REST) in the cytoplasm. In the absence of cytoplasmic sequestration of REST by wild-type huntingtin, REST is free to enter the nucleus, where it serves as a negative regulator or BDNF transcription.

The polyglutamine moiety has been shown to assume a β-pleated sheet conformation when the number of residues exceeds 35. Current opinion is that such a 'toxic fold' leads to aberrant binding interactions with other proteins, including with other copies of misfolded huntingtin. Consequently, one area that has been explored is whether chaperone proteins, such as the heat shock proteins, could mitigate toxicity by keeping polyglutamine-expanded huntingtin from assuming a misfolded conformation.

Posttranslational and Proteolytic Modifications of Huntingtin

Several posttranslational modifications of the huntingtin protein have been identified, including phosphorylation, sumoylation, and palmitoylation. Sumoylation (so named for 'small ubiquitin-like modifier' proteins) of the first 17 amino acids of the N-terminal regions of huntingtin can promote nuclear localization of exon 1 huntingtin fragments. Palmitoylation may govern huntingtin's interaction with membranous structures such as the Golgi or synaptic vesicles. It is likely that these posttranslational modifications alter huntingtin's intracellular distribution as well as its stability and degradation.

At least five different proteolytic mechanisms have been identified for huntingtin: cleavage by the apoptotic proteases known as caspases, cleavage by the calcium-dependent protease calpain, degradation by the ubiquitin–proteasome system, cleavage by an aspartyl protease, and autophagic digestion in the lysosome. There has been much focus on proteolytic processing

of huntingtin, since truncated fragments of huntingtin – specifically, those fragments containing the N-terminal portion of the molecule – appear to be especially toxic in a number of experimental paradigms. Compared to full-length versions of huntingtin, N-terminal fragments enter the nucleus more easily, form aggregates more readily, and induce cellular toxicity more readily. Thus, preventing certain cleavage events has been an attractive strategy by which to mitigate the toxic effects of huntingtin fragments.

Polyglutamines and Aggregates

The polyglutamine moiety that is expanded in mutant forms of the huntingtin protein gives rise to the neuropathologic changes observed in HD. A transgenic mouse model expressing only exon 1 of the huntingtin gene (the portion containing the polyglutamine moiety) develops an abnormal neurologic phenotype. The fact that such a striking phenotype could be produced in an animal expressing only exon 1 of the *HD* gene added weight to the early 'toxic fragment' hypothesis. One of the striking findings which emerged out of the findings with transgenic animals is the observation of novel inclusions in neuronal nuclei, although huntingtin normally exists as a predominantly cytoplasmic protein. This remarkable finding prompted reexamination of human biopsy and necropsy material, and abnormal huntingtin- and ubiquitin-positive nuclear and cytoplasmic aggregates were also found. Although abnormal aggregates are a striking feature of human HD and transgenic mouse HD models, recent studies have raised controversy as to the importance of inclusions. In cell culture models, the presence of inclusions was not correlated with neuronal death. Other models have suggested that inclusions may actually serve as a protective function.

Several theories have developed concerning the gain-of-function which emerges when the length of the polyglutamine moiety extends into the pathologic range. One theory proposes that polyglutamine moieties serve as a substrate for the enzyme transglutaminase. Nuclear inclusions in HD are reminiscent of protein deposits that are pathologic hallmarks of other neurodegenerative diseases; an emerging theme is that neurodegenerative diseases may be disorders of altered protein folding. Polyglutamine moieties are found within other proteins, including transcription factors. Many studies suggest that polyglutamine aggregates may serve as sinks to sequester transcription factors or to deplete proteins involved in vesicle movement. Aggregates present in the cytoplasm may also physically block vesicles, preventing transport within axons and dendrites.

Huntingtin aggregates are dynamic. In *in vitro* systems, one can visualize the appearance and growth of aggregates. In a transgenic mouse model that expressed a portion of the huntingtin protein under the control of an inducible promoter, stopping expression of the mutant transgene led to regression of the size of intranuclear aggregates, attesting to the recuperative ability of neurons with aggregates.

Pathogenesis in HD

Transcriptional Dysregulation

Human HD brains show selective downregulation of neurotransmitters and neurotransmitter receptors, a finding that has been recapitulated in transgenic mouse models. Gene expression studies using DNA microarrays have confirmed widespread alteration of mRNA levels of genes involved in neurotransmission, calcium homeostasis, and intracellular signaling. While transcriptional dysregulation has been acknowledged as an important pathogenic feature in HD, the exact mechanisms have not been elucidated.

Normal huntingtin is found predominantly outside the nucleus, whereas mutant versions, especially N-terminal fragments, of huntingtin tend to localize in the nucleus, where they can form visible aggregates. Mutant huntingtin has been shown to bind directly to a number of transcription factors, including Sp1, p53, cAMP response element-binding protein (CREB)-binding protein (CBP), and nuclear corepressor (NCoR), raising the possibility that mutant huntingtin could interfere with the normal function of these factors. Transcriptional dysregulation in HD likely involves histone proteins and chromatin remodeling, as histone deactylase inhibitors, which promote histone acetylation, have been shown to have beneficial effects in a number of animal and cellular models of HD.

Other Theories of Pathogenesis

Numerous other ideas have been advanced to account for HD pathogenesis, including synaptic dysfunction, disruption of axonal transport, apoptosis, inappropriate cell cycling, calcium dysregulation, oxidative stress, proteasome dysfunction, dysfunction of ion channels, endocrine dysregulation, disrupted mitochondrial biogenesis, altered lipid metabolism, loss of BDNF function, and altered intracellular second messenger signaling. A major challenge within the field is to prioritize the most relevant pathologic processes, while accounting for all of the observed abnormalities.

Genetic Testing

Isolation of the *HD* mutation made direct DNA testing available. While HD genetic testing is now

commercially available, testing should be approached judiciously. Genetic testing is performed in two contexts. Symptomatic gene tests are performed when a patient is manifesting symptoms that could be consistent with HD. For persons who are at risk for HD by virtue of having an affected relative, presymptomatic testing can be helpful in planning life events, such as whether to get married or have children. Caution should be exercised, however, as having a positive gene test result can have many unanticipated negative consequences. Persons undergoing presymptomatic testing should undergo detailed counseling with a genetic counselor. Preimplantation gestational testing, which can determine the gene status of a fetus, is available in some centers.

Therapeutics

There are currently no effective therapies for preventing the onset or for slowing the progression of HD. Current therapies are symptomatic, and include the use of neuroleptics to decrease chorea and the use of psychotropic medications to address depression, obsessive–compulsive symptoms, or psychosis. In addition, speech therapy and physical therapy are useful in addressing the swallowing and walking difficulties that many HD patients experience. An important principle in HD treatment is that the brain is constantly degenerating, and thus the response to medications changes over time. Therapy must therefore be flexible, and constant reevaluation is required. In addition, because of altered by neurotransmitter receptors, paradoxical reactions to medications can occur. Careful empiric trials must be performed with each patient. Only one medication should be altered at a time.

Some efforts at neural transplantation have been made. Transplantation of embryonic stem cells as replacement therapy is currently being evaluated, but remains experimental at this time. Other surgical therapies, such as ablative approaches or deep brain stimulation, have not been extensively tried.

The pathologic mechanisms underlying HD are not yet completely understood; effective therapeutics depends on a more clear elucidation of pathogenic pathways. The Huntington Study Group, a multicenter academic consortium of HD clinicians, has conducted numerous clinical trials for therapeutic compounds, and many are currently under way. Further therapies will no doubt arise as the molecular pathogenesis HD is further delineated.

See also: Axonal Transport and Huntington's Disease; Cell Replacement Therapy for Huntington's Disease; Excitotoxicity in Neurodegenerative Disease; Huntington's Disease; Huntington's Disease: Neurodegeneration; Protein Folding and the Role of Chaperone Proteins in Neurodegenerative Disease; Transgenic Models of Neurodegenerative Disease.

Further Reading

Bates G, Harper P, and Jones L (eds.) (2002) *Huntington's Disease*. New York: Oxford University Press.

Cha J-HJ (2000) Transcriptional dysregulation in Huntington's disease. *Trends in Neurosciences* 23: 387–392.

Everett CM and Wood NW (2004) Trinucleotide repeats and neurodegenerative disease. *Brain* 127: 2385–2405.

Ferrante RJ, Kowall NW, Beal MF, et al. (1985) Selective sparing of a class of striatal neurons in Huntington's disease. *Science* 230: 561–564.

Friedlander RM (2003) Apoptosis and caspases in neurodegenerative diseases. *New England Journal of Medicine* 348: 1365–1375.

Gusella JF, Wexler NS, Conneally PM, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306: 234–238.

Harjes P and Wanker EE (2003) The hunt for huntingtin function: Interaction partners tell many different stories. *Trends in Biochemical Sciences* 28: 425–433.

Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is unstable in Huntington's disease chromosomes. *Cell* 72: 971–983.

Leegwater-Kim J and Cha J-HJ (2004) The paradigm of Huntington's disease: Therapeutic opportunities in neurodegeneration. *NeuroRx* 1: 128–138.

Li SH and Li XJ (2004) Huntingtin–protein interactions and the pathogenesis of Huntington's disease. *Trends in Genetics* 20: 146–154.

Mangiarini L, Sathasivam K, Seller M, et al. (1996) Exon 1 of the *hd* gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* 87: 493–506.

Vonsattel JP and DiFiglia M (1998) Huntington disease. *Journal of Neuropathology & Experimental Neurology* 57: 369–384.

Vonsattel JP, Myers RH, Stevens TJ, et al. (1985) Neuropathological classification of Huntington's disease. *Journal of Neuropathology & Experimental Neurology* 44: 559–577.

Relevant Websites

http://www.hdfoundation.org – Hereditary Disease Foundation.
http://www.hdsa.org – Huntington's Disease Society of America.
http://www.huntington-study-group.org – Huntington Study Group.
http://www.wemove.org – Worldwide Education and Awareness for Movement Disorders.

Triune Brain Concept: A Comparative Evolutionary Perspective

A B Butler, George Mason University, Fairfax, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

In 1990, Paul MacLean published his book on *The Triune Brain in Evolution: Role in Paleocerebral Functions*, based on years of neuroethological and neurobiological studies in his laboratory and copious other sources. His ideas about viewing the brains of extant mammals, including humans, as comprising three main formations that were gained sequentially over a long evolutionary history, have been enthusiastically accepted by a number of scientists, particularly those interested in neuropsychiatry and related areas but, in stark contrast, strongly rejected by most scientists in the field of comparative, evolutionary neurobiology. While scientific disputes are not unheard of, the level of disagreement over the triune brain concept does distinguish it from many other current hypotheses. In discussing the triune brain concept, it is thus both necessary and of some interest to consider not only the concept itself but also the evolutionary, comparative point of view that takes issue with it.

Two reviews of MacLean's book, one by Anton Reiner, and the other by Boyd Campbell, written the same year as its publication, were critical on a number of points. Subsequently, in both editions of *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation* (Butler and Hodos), those criticisms were supported. More recently, others have mounted defenses of the triune brain concept; Jaak Panksepp particularly addressed conflicts engendered by the work of Joseph LeDoux on the amygdala and its relationship to emotion, and Gerald Cory replied to the reviews by Reiner and Campbell, as well as to the comments on the triune brain concept in the first edition of the Butler and Hodos textbook (as also maintained in the second edition).

As someone thus already involved in the continuing dispute, I will endeavor to be as objective and dispassionate as possible in presenting MacLean's triune brain concept and also the objections to it raised by comparative neuroscientists, recognizing that my own background and knowledge of brain evolution unavoidably influence and inform my perspective. Also, since Cory characterized Reiner's criticisms as 'destructive' rather than 'constructive,' in that Reiner did not offer a 'replacement generalization,' I will do so here. It is my hope that the alternative concepts presented here, an objective-terminology model for amniote brain evolution and an evolutionarily accurate model of mammalian brain evolution, will clarify the reasons why most comparative neurobiologists view the triune brain concept as they do. These alternative models allow for neuroanatomical hypotheses that encompass the recently discovered, exceptionally high-level cognitive abilities of birds without detracting from the similarly high-level cognitive abilities of mammals and especially humans.

From an evolutionary, systematic viewpoint, birds are actually reptiles, derived from the thecodont radiation that also gave rise to crocodiles. Since the brains of birds are, in many respect, enlarged and elaborated versions of other reptilian brains, all of which lack the particular cytoarchitectural arrangements that are the hallmark of the mammalian neocortex, the triune brain model does not account for their remarkable cognitive abilities that far exceed the more basic, species-specific behaviors involved in reproduction of the species and in maintaining homeostasis that MacLean accounts for in his model. While it was never the intent of the triune brain model to account for high-level cognition in birds, the fact that a brain with very different cytoarchitecture can generate such abilities definitively demonstrates that neocortex – the 'neomammalian' telencephalic component of the triune brain model – is not uniquely requisite to this function. It also definitively demonstrates that birds, as legitimate representatives of the reptilian lineage, have far more to their telencephalons than merely basal ganglia and far more to their behavioral capabilities than just 'species-typical stereotyped behaviors,' on which the triune brain model is focused.

The various high-level cognitive abilities of birds were not known when MacLean developed his model. The objections that comparative neurobiologists had to the model were, however, based on the then-current data set of known sensory pathways and other connections of forebrain structures in birds and reptiles, as well as those in mammals. These objections are validated by the new findings on avian cognitive abilities. We will thus now review the basic premises of the triune brain model and the objections raised by the comparative, evolutionary community, two of the most salient and important of which focus on (1) the issue of pallial evolution across all amniotes rather than just within the mammalian lineage and (2) the assignment of involvement of various particular neural structures in the generation of a number of behaviors.

Of great importance for those whose research is focused on human behaviors – including social, psychological, and economic as well as pathological conditions such as those of the basal ganglia – is that

the 'replacement generalizations' (see later), the objective-terminology model for amniotes and the evolutionarily-accurate model for mammals, do not contradict or undermine any of their findings or theoretical base. These models simply rectify the conceptual approach with the full range of comparative evolutionary data on brain evolution. In his foreword to the 2002 book, *The Evolutionary Neuroethology of Paul MacLean*, Jaak Panksepp noted that emotions cannot be understood simply as "channel functions" at the level of single cells but must rather be appreciated as "global 'state-functions' reliant on complex organic organization of broadly operating brain systems," for which the triune brain concept, including the limbic system concept, is valuable. Since both single cellular activity and global brain systems are active and ultimately produce emotional feelings (and other aspects of higher order consciousness), both need to be acknowledged and dealt with in any heuristic device. The revised concepts offered in the following discussions are compatible at both levels. What they both lack is the romantic cache of a 'reptilian' image, darkly lurking under the surface, within the deep 'core' of a mammal – an image unfortunately better suited to a Victorian novel than to a scientific hypothesis.

The Triune Brain Hypothesis

As presented by MacLean in his 1990 book, the triune brain of mammals comprises three formations – a 'Reptilian,' or 'R-complex' formation, a 'Paleomammalian' formation, or limbic system, and a 'Neomammalian' formation (Figure 1). These three formations are shown with equivalent color coding for ease of comparison across Figures 1–4. In the figure caption of the original figure in his book (his figure 2-1), MacLean states that "the human forebrain expands along the lines of three basic formations that anatomically and biochemically reflect an ancestral relationship, respectively, to reptiles, early mammals, and late mammals." The 'reptilian' formation includes the basal ganglia within the telencephalon and a number of brain stem structures. The 'paleomammalian' formation is the limbic system, and the 'neomammalian' formation is the mammalian neocortex and its associated brain stem structures. MacLean focuses on the limbic system as "the common denominator in the brains of all mammals" (p. 16). He states that it consists of "three main subdivisions ... [with] the third subdivision, for which there appears to be no rudimentary counterpart in reptiles, ... implicated in parental care, audiovocal communication, and play behavior" (p. 17). By implication, MacLean recognizes that the other two limbic divisions may have counterparts in the reptilian brain, but clearly excludes the third division. The 'neomammalian'

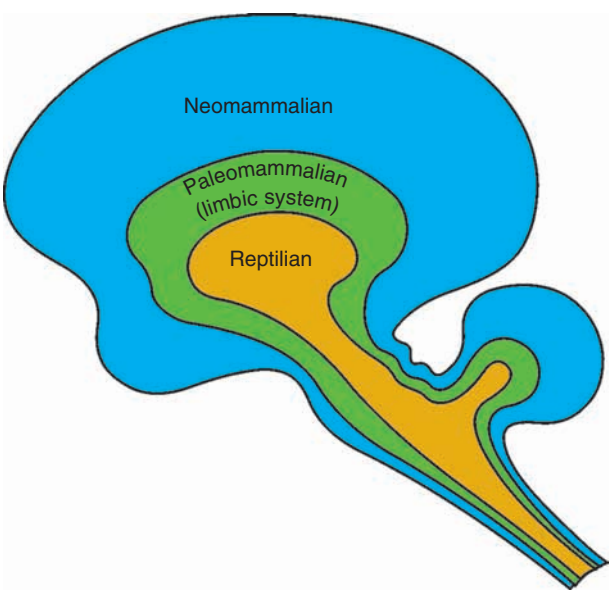

Figure 1 The triune brain model as presented by PD MacLean in his 1990 book. Reproduced from MacLean PD (1990) *The Triune Brain in Evolution: Role in Paleocerebral Functions*, p. 9. New York: Plenum, with kind permission from Springer Science and Business Media.

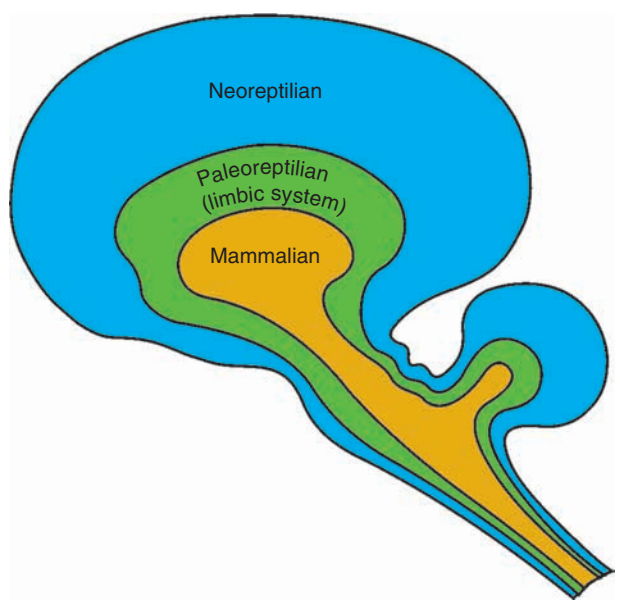

Figure 2 'Reversal test' of the triune brain concept. Here, the brain stem and basal ganglia are labeled as the 'Mammalian' formation, the limbic system as the 'Paleoreptilian' formation, and the expanded nonlimbic pallial areas as the 'Neoreptilian' formation. This set of labels depicts amniote brain evolution from a sauropsid chauvinist viewpoint; it is just as evolutionarily inaccurate as the mammalian chauvinist triune brain model. The reversal test thus demonstrates the fallacy of the latter. Redrawn and modified from MacLean PD (1990) *The Triune Brain in Evolution: Role in Paleocerebral Functions*. New York: Plenum.

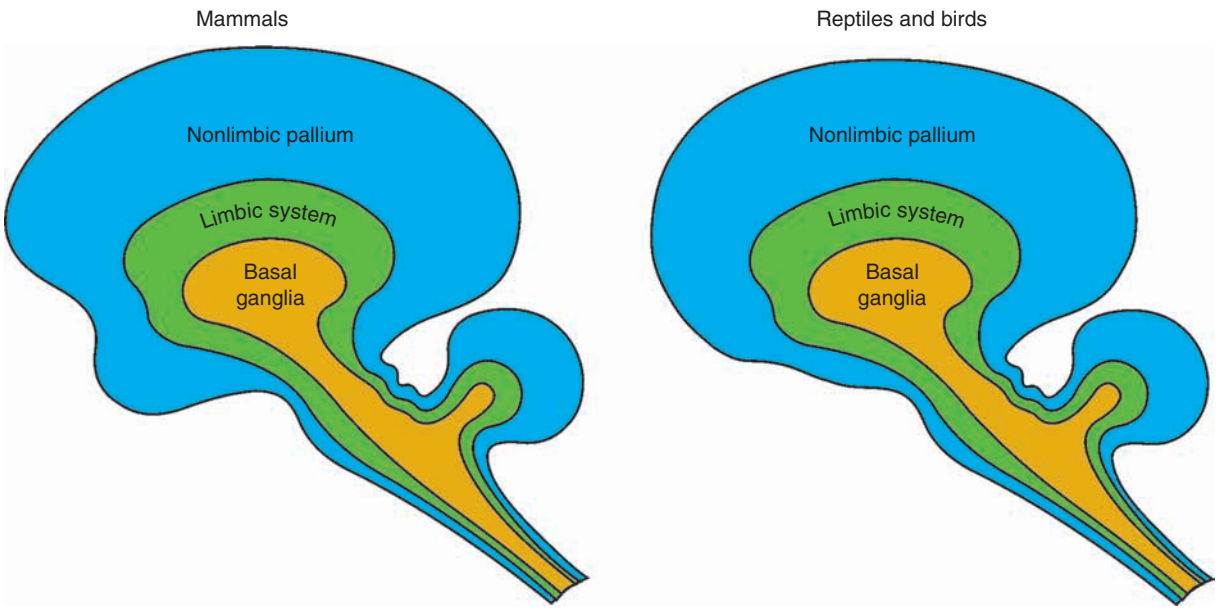

Figure 3 The objective-terminology model, a 'replacement generalization' for the triune brain concept for both mammals (left), and reptiles and birds (sauropsids; right). This model employs labels that are objective and based on actual brain structures. As in MacLean's original drawing, the formations are labeled only at the level of the telencephalon as basal ganglia, limbic system, and nonlimbic pallium, respectively. These labels comfortably and accurately accommodate evolutionary histories of both the sauropsidian and the mammalian lineages. It should be noted that in this model, the mammalian prefrontal cortex is included in the nonlimbic pallium even though it is sometimes thought of as a limbic component. The model easily could be modified, depending upon one's preference, to include prefrontal cortex as a limbic component in mammals, with a comparable adjustment in sauropsids – particularly birds – to similarly include their pallial areas that are possibly homologous to prefrontal cortex and involved in comparable aspects of higher cognitive processing, such as the nidopallium caudolaterale and/or the dorsolateral corticoid area. Redrawn and modified from MacLean PD (1990) *The Triune Brain in Evolution: Role in Paleocerebral Functions*. New York: Plenum.

formation is the "neocortex and the thalamic structures with which it is primarily connected" (p. 17). MacLean states that "In its evolution the neocortex, together with its brainstem and neocerebellar connections, has afforded a progressive capacity for problem solving, learning, and memory of details ... [and has the] capacity to generate verbal communication [in humans]" (p. 17).

In a later chapter (pp. 43ff), MacLean acknowledges that a large part of the reptilian (and avian) telencephalon, the dorsal ventricular ridge (DVR), is a pallial formation rather than a component of the basal ganglia within the subpallium. He also acknowledges that "In recent years several authors seem to have been persuaded that the DVR is the equivalent of the mammalian neocortex" (p. 45), citing the work of Walle Nauta and Harvey Karten on this issue. MacLean felt that "The question of whether or not some other mammalian structure might be the counterpart of the DVR is better deferred until embryological findings have been reviewed" (p. 45). In this pronouncement, he was highly prescient, since it is currently a matter of substantial debate, based to some extent on recent embryological and gene expression data, as to whether the DVR, particularly its anterior part, the anterior dorsal ventricular ridge (ADVR), is homologous to the mammalian neocortex (i.e., at the level of various specific cell populations within both structures, as Karten originally postulated) or to the thalamo-recipient basolateral division of the amygdala.

The better part of MacLean's book covers neuroethological and experimental behavioral observations from his extensive work on various reptiles and squirrel monkeys. Since the observations of reptilian behaviors, and investigations into the contribution of the basal ganglia toward these behaviors, are of paramount importance as the foundation of the triune concept, this part of MacLean's research program will be focused on here. Much of this research was carried out by Neil Greenberg and his co-workers in MacLean's Poolesville (Maryland) research facility. Greenberg and MacLean, from extensive observations of the behaviors of *Anolis* lizards in laboratory vivaria, as well as observations on other species drawn from the literature, categorized them into daily routines and subroutines and component behaviors associated with territorial and reproductive issues, including establishment of homesite and homerange, patrolling territory, ritualistic defense displays, intraspecific fighting, use of

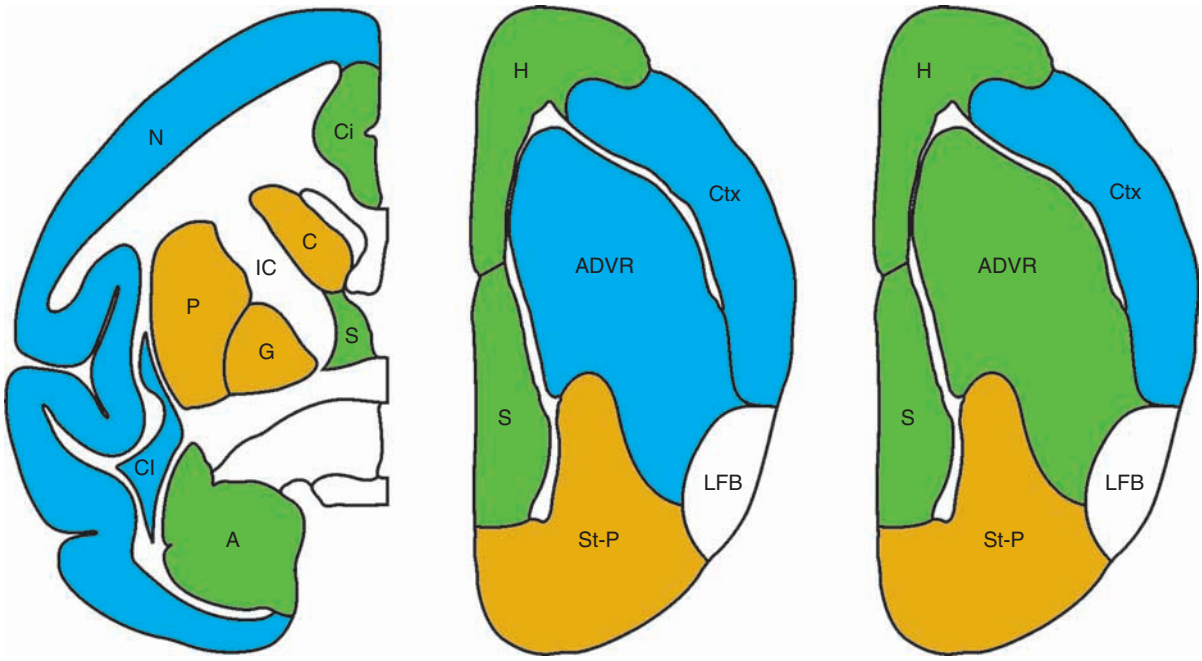

Figure 4 Comparison of drawings of the left hemisphere through the telencephalon of the squirrel monkey (left), and the right hemisphere through the telencephalon of the tegu lizard (center and right), showing the basal ganglia (tan), the limbic pallium (green), and the nonlimbic pallium (blue). The two versions of the tegu hemisphere represent the two current hypotheses as to whether the ADVR is a nonlimbic (center) or limbic (right) component. Regardless of the homology of the latter structure, note that most of the telencephalic areas are pallial rather than subpallial (the basal ganglia) in both mammals and lizards. In birds, the pallial regions are even more expanded and elaborated than in lizards. Not all structures are present in both the mammal and the tegu hemispheres at the levels shown. A, amygdala; ADVR, anterior dorsal ventricular ridge; C, caudate nucleus; Ci, cingulate gyrus; Cl, claustrum; Ctx, cortex; G, globus pallidus; H, hippocampal formation; IC, internal capsule; LFB, lateral forebrain bundle; N, neocortex; P, putamen; S, septum; St-P, striatopallidum.

defecation posts, foraging, hunting, homing, hoarding, social group formation, establishment of social hierarchy, greeting, grooming, courtship displays, mating and breeding, flocking, and migration.

As discussed by Neil Greenberg in his 1988 paper, they regarded the "highly stereotyped, species-typical behavioral patterns, particularly social displays, [as] the most robust data to illuminate neural corollaries of complex behavioral functions." Since MacLean had previously found that the dorsal striatum "is necessary for the performance of a species-typical display in the [squirrel] monkey ... an investigation of the paleostriatum [i.e., dorsal striatum] of a lizard predicated on a close understanding of the units of behavior spontaneously emitted in naturalistic settings" was undertaken. As summarized by MacLean in his 1990 book, Greenberg had previously identified major stereotyped behaviors in the blue spiny lizard *Sceloporus cyanogenys* – signature ('assertive'), challenge ('territorial'), courtship, and submissive ('appeasement,' 'assentive'). Similarly, *Anolis* lizards, under the seminatural vivarium conditions of the laboratory, exhibit social displays that include assertion (a 'general arousal'), challenge (addressed to a conspecific male), and courtship (addressed to a conspecific female).

Of these behaviors, only the challenge display was disrupted by lesions that involved the output pathway (the lateral forebrain bundle region) of the striatum. The species-typical stereotyped behaviors of assertion and courtship were not disrupted by such lesions. Other studies have not contributed any additional definitive indication of striatal involvement in these sorts of functions. While both lesion and stimulation studies are subject to problems of interpretation, the lack of evidence that the basal ganglia generate most of the behaviors observed in reptiles should preclude designating this part of the brain as singularly crucial to them. The broad generalization of assigning most aspects of reptilian behavior to the 'R-complex' in the triune brain model simply is not justified by the experimental data. Other parts of the reptilian brain, particularly the pallium, may well be contributing substantially, in concert with the basal ganglia, to a number of these behaviors, just as in mammals the neocortex and the basal ganglia interact in concert to generate cognitive and affective as well as stereotypical motor behaviors.

Origins of the Triune Brain Hypothesis and the Comparative Neuroanatomical Perspective

Paul MacLean's life and career recently have been reviewed by Kelly Lambert. Having received his medical degree at Yale University School of Medicine and exposure to the field of psychiatry during a World War II assignment to New Zealand and a brief stint in private practice, MacLean had worked in the laboratory of Stanley Cobb at Massachusetts General Hospital and subsequently held a position in the Department of Physiology at Yale Medical School. During this period he developed an interest in the work of James Papez, who was at Cornell, and was able to meet with him to discuss Papez's work in comparative neuroanatomy and the circuitry of the limbic lobe in mammals.

In a paper published in 1952, MacLean coined the term 'limbic system' to refer to the set of structures that include the limbic lobe of Broca plus the related brain stem structures. While this term has been criticized in some circles, as Panksepp has discussed, it has been widely accepted by comparative neuroanatomists, since the main components of this system – such as the hippocampus and its related cortical areas, the septal nuclei, hypothalamus, amygdala, and ventral striatopallidal components – are widely distributed across vertebrates. This issue will thus not be discussed further here, since from the comparative viewpoint, the limbic system concept *per se* is not a point of contention.

In 1957, MacLean joined a newly founded lab section at the National Institutes of Health (NIH), which subsequently was named the Section on Limbic Integration and Behavior, and in 1971 moved to new NIH laboratory quarters near Poolesville, Maryland, where he was able to expand his studies to include reptile behaviors and to begin his efforts to understand the possible contributions of the basal ganglia to these behaviors. During the several years on either side of that point in time, two major breakthroughs in comparative neuroanatomy were occurring, one in regard to the telencephalic pallium of birds and reptiles and the other in regard to similar parts of the brain in sharks.

In sharks, Sven Ebbesson and his colleagues had found that (1) olfactory projections are confined to a small region of the lateral pallium, rather than being widely distributed across it, (2) two major ascending pathways, one via the midbrain roof and thalamus and one directly via the thalamus, to nonlimbic parts of the pallium are present, and (3) the integrity of the pallium is necessary for the sharks to discriminate visual patterns. This new information contradicted the then generally accepted idea that the forebrains of all fishes were dominated by olfactory information – mere 'smell brains' – and that the other sensory systems had 'invaded' the forebrains of terrestrial vertebrates during their evolutionary transition. In regard to the telencephalic pallium of birds, Harvey Karten and his colleagues had confirmed that a large structure in the telencephalon, the DVR (now also referred to in birds as nidopallium and mesopallium) is indeed pallial rather than subpallial and also is in receipt of several ascending sensory pathways – as known at that time, a visual and an auditory pathway, and, as later identified, a somatosensory pathway. These sensory pathways are in addition to the then-known visual pathway to the more caudal part of the avian equivalent of the cortex in reptiles, called the Wulst or hyperpallium, and a subsequently discovered somatosensory pathway to its more rostral part that also gives rise to a pyramidal tract-like efferent motor pathway.

By the mid-1970s, these findings in sharks and birds were published and were augmented by findings of similar ascending sensory pathways in a variety of reptiles. In 1978, MacLean held a conference at his Poolesville laboratory, to which a number of us were invited and at which many of these findings were presented and subsequently published in the resulting book, edited by Neil Greenberg and Paul MacLean, *Behavior and Neurology of Lizards*. A survey of the comparative findings on telencephalic evolution across vertebrates was subsequently published in another volume edited by Sven Ebbesson in 1980, from a conference held in Caracas, Venezuela, *Comparative Neurology of the Telencephalon*, and from the mid-1960s to the present, a massive body of publications on this subject has appeared in the primary literature and other symposium volumes.

In his own summary of his triune brain hypothesis, MacLean acknowledged current findings about brain evolution within amniotes, particularly within the sauropsid radiation of reptiles and birds, but then set them aside in favor of his 'triune model' of brain evolution within the mammalian line. In the caption for his famous figure of the mammalian triune brain (his figure 2-1), he states that the three formations within the brain – the parts of which are labeled at the rostral end of the brain as the 'Reptilian' formation (or 'R-complex'), the 'Paleomammalian' formation (limbic system), and the 'Neomammalian' formation (neocortex and its related thalamic structures) – "reflect an ancestral relationship, respectively, to reptiles, early mammals, and late mammals." The latter statement is contradictory to all data collected and analyzed by comparative studies over the past four decades, as already discussed. A 'reptilian' formation – that is, the basal ganglia – is

common to early and late reptiles, early and late birds, and early and late mammals. A 'paleomammalian' formation, or limbic system, is also common to all of these taxa. While neocortex is indeed a unique feature of mammals, an expanded nonlimbic pallium, part of which comprises the neocortex in mammals, is likewise common to all of these taxa, and this crowning component of the telencephalon is particularly elaborate and is involved in sensory processing and associative functions that reach very high cognitive levels in both mammals and birds.

MacLean notes in his discussion of anatomical and functional comparisons (in chapter 2 of his 1990 book) that a large formation, the DVR is present in sauropsid brains and is clearly pallial. However, he regards it as having been independently evolved and as not being of significant interest in regard to the evolutionary origins of mammalian neocortex. In this view, he is partly correct and partly wrong. As discussed by Anton Reiner, early amniotes most likely had neither neocortex nor a DVR, but rather had a small "proto-DVR" region in the lateral aspect of the pallium and a modest region of cortex medial to it. In the mammalian lineage, which precedes the reptilian lineage in the fossil record by 10 My, the more medial region became expansively elaborated to form part (or possibly all) of the neocortex, and the lateral region likewise became elaborated to form either part of the neocortex and/or the pallial, thalamo-recipient (basolateral) part of the amygdala. In the sauropsid line, the lateral region became elaborated and expanded to form the definitive DVR, and the medial region formed the rudimentary cortex; within the avian lineage, the medial region expanded substantially to form the Wulst, while the DVR expanded further to reach its apogee.

The triune brain hypothesis, by essentially dismissing interest in pallial elaboration within the sauropsid line, errs by instead promoting the scientifically contradicted but still widespread popular view of a scala naturae, a linear ranking of animal taxa, with fish near the bottom, reptiles not much higher, non-human mammals higher and ranked in sequence (such as rat-to-cat-to-monkey), and with humans at the top. The summary figure of the triune brain, now widely published and commonly referred to in some of the psychiatric literature, implies that the core structures of basal ganglia and the more caudal parts of the brain stem represent the 'reptilian brain' that is still present in mammals, including humans. It specifically states that the gain of a limbic system occurred in early mammals. It specifically implies that a large, expanded nonlimbic pallium is unique to mammals.

Applying a 'reversal test' is often a fruitful way to assess particular situations or statements. We can apply this test to the triune brain hypothesis to assess whether it is a reasonable view of brain evolution. Figure 2 shows the view of brain evolution, as seen from the sauropsid point of view. The labeling in this figure is equally incorrect, from the established body of data on brain evolution, as that of MacLean's canonical figure. It identifies the core of the brain as the 'Mammalian' formation, the limbic system as the 'Paleoreptilian' formation, and the nonlimbic pallium as the 'Neoreptilian' formation. As we have noted, in birds the 'Neoreptilian' formation, comprising the Wulst and DVR, is particularly expanded, and the recent findings of highly sophisticated cognition of which at least some birds are capable compel viewing it as a worthy rival of neocortex in terms of its neural processing abilities. To ignore the sauropsid line and to assign some mammalian behaviors as 'reptilian' is to ignore a virtual goldmine of information on how neurons process and assess information and how thoughts and emotions are generated and dealt with. In trying to understand human behaviors, the latter course is as foolish as if one were to ignore all mammalian, including human, research and study only reptiles and birds. Taking the complete picture into account, based on the actual evolutionary events and particularly including the intriguing differences in cytoarchitecture of the pallial formations in sauropsids versus mammals, is the course most likely to yield a correct understanding of pallial function, including the generation of complex human behaviors.

In assessing the triune brain model, one wonders why MacLean did not simply label its components by their correct terms, as shown in Figure 3. In this figure the core of the brain (at the telencephalic level) is labeled 'basal ganglia.' The limbic system is so labeled, and the overlying nonlimbic pallium is likewise labeled as itself. These labels are objective; they have no scala naturae baggage attached to them. All mammals and all reptiles and birds have a core region (in simplistic terms) that includes the basal ganglia within the telencephalon. All mammals and all reptiles and birds have a limbic system. All mammals and all reptiles and birds have an expanded nonlimbic pallium. For that matter, fish and amphibians also have components of the basal ganglia, limbic system, and nonlimbic pallial regions. None of these neural systems is unique even to amniotes, let alone to mammals. This objective-terminology model is a highly reasonable 'replacement generalization' for the triune brain model. It is anatomically accurate, simple, and straightforward, and does not conflict with any phylogenetic data. It is wholly applicable to all of the

clinical and related findings to which the triune brain model has previously been applied.

Another view of the anatomical comparison is shown in Figure 4. In this figure, a drawing of a transverse hemisection through the telencephalon of a squirrel monkey, the species of primate that Paul MacLean studied so extensively, is shown on the left, and two similar drawings of a transverse hemisection through the telencephalon of a tegu lizard (with features essentially the same as that of *Anolis* and other lizards of this DVR type) is shown on the right. The two drawings of the lizard telencephalon represent the two extremes in the current debate about homology of the ADVR: whether it is nonlimbic pallium, as indicated by its blue shading in the first (center) drawing, or whether it is amygdala and thus a limbic component, as indicated by its green shading in the second drawing. Either way, it is clear that the reptilian telencephalon is largely constituted by pallial territory, just as are mammalian and avian brains, and that the structures designated as the 'R-complex' by MacLean (the basal ganglia, shaded in tan) comprise only a minor fraction of the telencephalic tissue.

For those in the clinically related fields who still wish, however, to retain the triune brain concept (or at least some version of it that may be more in line with the data accumulated to date), a view that focuses

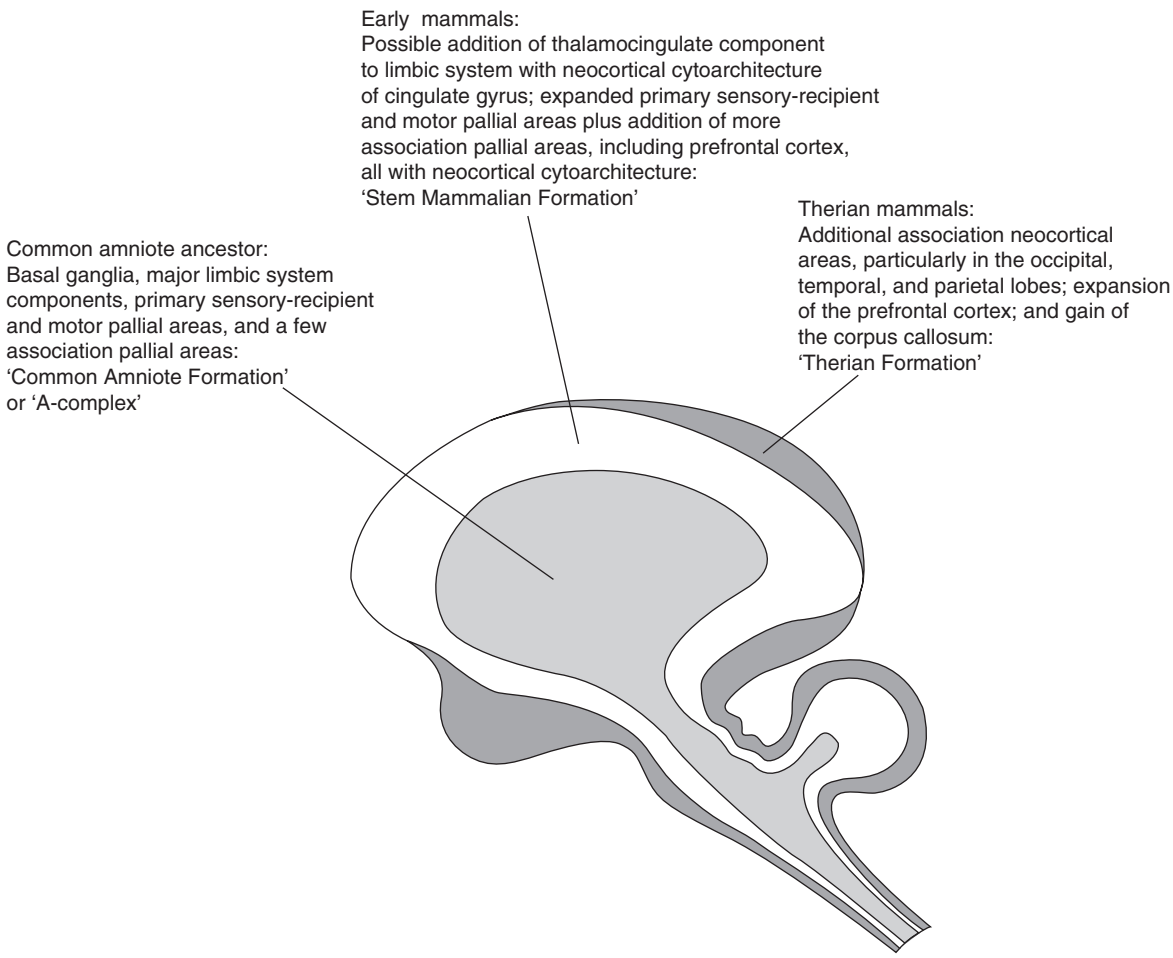

Figure 5 The evolutionarily accurate model of brain evolution within the mammalian (synapsid, i.e., with one temporal fenestra in the skull) radiation, which separated from ancestral lineage of anapsid (lacking any bony fenestrae) amniotes 10 My before the sauropsid (diapsid, i.e., with two temporal fenestrae) line of ancestral reptiles appears in the fossil record. In this model, the neural components present in the common, anapsid ancestor ('Common Amniote Formation,' or 'A-complex') include major limbic system components, primary sensory-recipient and motor pallial areas, and a few association pallial areas, in addition to the basal ganglia. Early mammals and therian (marsupial and placental) mammals then gained additional structures and/or modifications to existing structures, as shown for the 'Stem Mammalian Formation' and 'Therian Formation,' respectively. Redrawn and modified from MacLean PD (1990) *The Triune Brain in Evolution: Role in Paleocerebral Functions*. New York: Plenum.

solely on the mammalian lineage but is evolutionarily accurate is shown in Figure 5. This is offered here as an alternative 'replacement generalization.' It suffers from being more cumbersome in terms of the detail put into the labels for the three sets of brain structures, which even so are only a generalization. It is not shaded in colors as was done for the other figures, since the comparison is no longer one-to-one. The 'common amniote formation' (or 'A-complex,' for those who wish to use such a term), shown in light gray, includes much more than the basal ganglia, and so its color shading would necessarily have to include all three colors used before – tan, green, and blue. Both the intermediately positioned area shown in white, designated here as the 'Stem Mammalian Formation,' and the outer area shown in darker gray, the 'Therian Formation,' would have to include both green and blue. While this model does not thus correspond easily to the original triune brain model or to the objective-terminology model, it nonetheless reflects in somewhat more detail the particular evolutionary course that was taken along the mammalian lineage.

When MacLean developed his triune brain model, the extensive and very high-level cognitive abilities of birds, as alluded to several times previously, were not appreciated. One of the most crucial and difficult questions to answer in all of neurobiology is how such complex cognitive abilities and the probably correlated high levels of consciousness are generated by neuronal assemblies. The foundation of behavioral data amassed by MacLean as well as the numerous other studies in the literature can contribute to answering this question when assessed in an evolutionarily accurate perspective. For example, in discussing the 'paleomammalian formation,' as noted earlier, MacLean specifically cited the third division of the limbic system, that is, its thalamocingulate component, as being "implicated in parental care, audiovocal communication, and play behavior." While it is true that only a few reptiles (mostly crocodiles) exhibit even the most rudimentary of these behaviors, birds exhibit all three to an extensive degree. In his 1981 book *Animal Play Behavior*, Robert Fagan documents and discusses many examples of play behavior in birds; the literature that documents the complexity of song learning and production in songbirds is extensive; and the maternal care exhibited by most species of birds is widely known.

In discussing the 'neomammalian formation,' MacLean cites its roles in "problem solving, learning, and memory of details . . . [and its] capacity to generate verbal communication." As reviewed by Butler and Cotterill, the recent spate of studies on avian cognitive abilities has established that various species of birds also can problem-solve, learn, and remember to an extensive degree, and even generate verbal communication. In fact, they exhibit numerous cognitive behaviors of a very high level. They (1) use transitive inference, the use of logical and successive comparisons (a process attributed to the working memory functions of prefrontal cortex in humans and some other mammals) to determine social hierarchies, (2) exhibit a phenomenon called coherence, or multistability, of ambiguous figures, previously considered unique to primates and also involving working memory, in which one interpretation of a visual stimulus (such as a Necker cube) is perceived for a set of trials, followed by a set of the alternate interpretation, rather than just randomly switching back and forth, (3) form conceptual categories, such as distinguishing cubist paintings from impressionist ones, (4) exhibit object constancy to a high level, as delineated by Piaget, (5) manufacture tools for use in food gathering, another ability previously thought unique to primates, (6) culturally transmit the design features of such tools to conspecifics, (7) communicate with humans using the English language to an extensive degree, (8) understand numerical concepts up to about the number 7 and also exhibit a zerolike concept – a particularly challenging feat, (9) exhibit theory of mind in terms of anticipating and taking steps to prevent thievery from food caches by conspecifics and practicing deceptive behaviors to outwit a dominant rival, and (10) exhibit the capacity to remember past events and their relative lengths of time.

These high-level cognitive abilities are all supported, in concert with the thalamus and other structures, by the ADVR and Wulst in birds – the pallial regions of the telencephalon that are homologous to major pallial regions in mammals and to the major pallial regions of reptilian brains as well. One of the strengths of the comparative method is its focus on identifying both the similarities and differences between brains and the correlated functions that they support. The objective-terminology model of amniote brain evolution offered here (Figure 3) facilitates such comparisons and can contribute to illuminating how different cytoarchitectures, including our own, support the generation of high-level cognition and consciousness by neuronal assemblies.

See also: Basal Ganglia: Evolution; Brain Development: The Generation of Large Brains; Brain Evolution: Developmental Constraints and Relative Developmental Growth; Brain Evolution: The Radiator Theory; Brain Fossils: Endocasts; Broca's Area: Evolution; Cerebellum: Evolution and Comparative Anatomy; Comparative Neurobiology: History; Neuroanatomy Methods in Humans and Animals.

Further Reading

Butler AB (1994) The evolution of the dorsal pallium in the telencephalon of amniotes: Cladistic analysis and a new hypothesis. *Brain Research Reviews* 19: 66–101.

Butler AB and Cotterill RMJ (2006) Mammalian and avian neuroanatomy and the question of consciousness in birds. *Biological Bulletin* 211: 106–127.

Butler AB and Hodos W (2005) *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation,* 2nd edn. Hoboken, NJ: Wiley.

Campbell CBG (1992) Book review (MacLean: *The Triune Brain in Evolution*). *American Scientist* 80: 497–498.

Clayton NS, Bussey TJ, and Dickinson A (2003) Can animals recall the past and plan for the future? *Nature Reviews Neuroscience* 4: 685–691.

Cory GA Jr. (2002) Reappraising MacLean's triune brain concept. In: Cory GA Jr. and Gardner R Jr. (eds.) *The Evolutionary Neuroethology of Paul MacLean*, pp. 9–27. Westport, CT: Praeger.

Cory GA Jr. and Gardner R Jr. (eds.) (2002) *The Evolutionary Neuroethology of Paul MacLean.* Westport, CT: Praeger.

Ebbesson SOE (1980) *Comparative Neurology of the Telencephalon.* New York: Plenum.

Fagan R (1981) *Animal Play Behavior.* New York: Oxford University Press.

Greenberg N (2002) Adaptive functions of the corpus striatum: The past and future of the R-complex. In: Cory GA Jr. and Gardner R Jr. (eds.) *The Evolutionary Neuroethology of Paul MacLean*, pp. 45–81. Westport, CT: Praeger.

Greenberg N, Font E, and Switzer R (1988) The reptilian striatum revisited. In: Schwerdtfeger WK and Smeets WJ (eds.) *The Forebrain in Reptiles: Current Concepts of Structure and Function*, pp. 162–177. Basel: Karger.

Greenberg N and MacLean P (eds.) (1978) *Behavior and Neurology of Lizards.* Rockville, MD: National Institute of Mental Health.

Lambert KG (2003) The life and career of Paul MacLean: A journey toward neurobiological and social harmony. *Physiology & Behavior* 79: 343–349.

MacLean PD (1990) *The Triune Brain in Evolution: Role in Paleocerebral Functions.* New York: Plenum.

Panksepp J (2002) Foreword: The MacLean legacy and some modern trends in emotion research. In: Cory GA Jr. and Gardner R Jr. (eds.) *The Evolutionary Neuroethology of Paul MacLean*, pp. ix–xxvii. Westport, CT: Praeger.

Pepperberg IM (1999) *The Alex Studies: Cognitive and Communicative Abilities of Gray Parrots.* Cambridge, MA: Harvard University Press.

Reiner A (1990) An explanation of behavior [review of *The Triune Brain in Evolution: Role in Paleocerebral Functions*, by Paul D. MacLean, 1990]. *Science* 250: 303–305.

Reiner A (2000) A hypothesis as to the organization of cerebral cortex in the common amniote ancestor of modern reptiles and mammals. In: Bock GR and Cardew G (eds.) *Evolutionary Developmental Biology of the Cerebral Cortex. Novartis Foundation Symposium 228*, pp. 83–102. Chichester, UK: John Wiley & Sons.

Reiner A, Yamamoto K, and Karten HJ (2005) Organization and evolution of the avian forebrain. *The Anatomical Record Part A* 287A: 1080–1102.

Weir AAS, Chappell J, and Kacelnik A (2002) Shaping of hooks in New Caledonian crows. *Science* 297: 981.

Tuberous Sclerosis

S P Sparagana, Texas Scottish Rite Hospital for Children, Dallas, TX, USA
E S Roach, Ohio State University School of Medicine, Columbus, OH, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Tuberous sclerosis complex (TSC) is a neurocutaneous disorder with a prevalence of about 1 in 6000. It results from a mutation of either of two different genes (*TSC1* or *TSC2*). The protein products of the TSC genes, hamartin and tuberin, are involved in a complicated schema of cell signaling pathways. Disruption of these signaling pathways resulting from mutations in the TSC genes results in alteration in cellular differentiation and proliferation, which may affect many organ systems, including brain, skin, kidneys, eyes, heart, and lungs. The lesions and tumors of TSC are typically benign, but they may cause potentially severe dysfunction in the affected organ, and malignant tumors may occur rarely. Surveillance monitoring for potential complications can minimize the severity of the complications. Knowledge of the molecular pathogenesis of TSC has permitted development of new treatments for certain complications.

Genetics and Molecular Pathogenesis of Tuberous Sclerosis Complex

Tuberous sclerosis complex has long been recognized as a dominantly transmitted genetic disease, but it arises from a spontaneous mutation 60% or more of the time. Since the elucidation of the *TSC2* gene and gene product (chromosome 16p13.3, tuberin) and the *TSC1* gene and gene product (9q34.3, hamartin), incredible strides have been made in the understanding of how mutations in these genes cause TSC.

TSC has virtually complete penetrance, but its expression is highly variable even among individuals with the same mutation. For example, a parent may have so few clinical findings of TSC that his diagnosis goes unrecognized until his infant develops infantile spasms or congestive heart failure due to TSC. Additionally, somatic and gonadal or germ-line mosaicism appears to be relatively common, and in such individuals only certain cells carry the TSC mutation, depending on when the mutation occurred during embryogenesis. Mosaicism may result in a very mild phenotype in a parent, and in the case of pure germ-line mosaicism, the parent has no clinical manifestations of TSC. A TSC mutation will not be found in somatic cells such as leukocytes or buccal cells, yet the parent may have more than one affected child due to the occurrence of the TSC mutation in egg or sperm progenitor cells. Because mosaicism may confound the heritability of TSC, even in apparently sporadic cases, parents should be cautioned of a 1–2% recurrence risk in subsequent children.

More than 1000 different mutations in the two TSC genes have been identified to date. When TSC gene mutations are analyzed in large groups of individuals with TSC, approximately 30% of individuals will have mutations in the *TSC1* gene and 70% will have mutations in the *TSC2* gene. About 10–20% of individuals with definite TSC by clinical criteria will not have an identifiable TSC gene mutation, which may be explained by mosaicism, at least in some cases.

Despite the fact that there are two TSC genes on two different chromosomes that code for two entirely different proteins, the phenotype produced by mutations in each of these genes is virtually indistinguishable. Some reports have suggested that there is a more severe phenotype in individuals with a *TSC2* mutation (for example, mental retardation, autistic behavior, and renal malignancy are reportedly increased); other reports have not confirmed this phenotypic pattern. At this time, with the exception of the *TSC2*/adult polycystic kidney disease contiguous gene syndrome (discussed later), the question of a clear-cut genotype–phenotype relationship is unclear. A large-scale natural history study of individuals with TSC, which will soon be underway, may help us understand whether there is any link between mutations in the different genes – and even specific types of mutations – with specific phenotypes.

TSC1 and *TSC2* are tumor suppressor genes. The TSC gene products, hamartin and tuberin, form a heterodimeric complex in the cell. This protein complex is integrally involved in several different signaling pathways. When either of the TSC gene products is compromised, the tumor suppressor function is diminished. Loss of regulation in the various cell signaling pathways may result in cell proliferation and overgrowth, resulting in the various hamartomatous tumors characteristic of TSC. Biochemical and genetic studies in various types of tumor cells, mammalian models, fruit fly, and yeast systems have helped define the role of the hamartin–tuberin complex in several cell signaling pathways. The most notable of these pathways involves the mammalian target of rapamycin (mTOR). The hamartin–tuberin complex represses mTOR activity in the normal functioning cell. However, when one of the TSC proteins is aberrant, mTOR

repression diminishes, and S6 kinase (S6K) is stimulated, which promotes ribosome assembly and protein translation. This in turn leads to cell proliferation, cell growth, and, ultimately, hamartoma formation. This pathway has become important because rapamycin (sirolimus) inhibits mTOR. Rapamycin is currently available and approved by the United States Food and Drug Administration (FDA) as an immunosuppressant agent, and is now undergoing human trials as a treatment for renal and pulmonary complications of TSC. The hamartin–tuberin complex also affects cellular pathways involving cellular adhesion and migration, and in protein and vesicular trafficking. While a great deal is now known about these cellular signaling pathways, details continue to be refined. As these pathways and the roles of tuberin and hamartin are further clarified, new targets for molecular-based treatments for TSC will surely be developed.

Clinical Features of TSC: General

TSC can involve virtually any tissue or organ system in the human body, but the skin, brain, and kidneys are the most consistently affected organs. Disease presentation is heterogeneous, and presentation even within a family may be extremely variable. Reasons for the variable presentation are not completely understood. However, partial explanations include genotype–phenotype relationships and stochastic developmental processes during embryogenesis. For example, loss of heterozygosity through a 'second-hit' somatic mutation is known to occur in some hamartomas. This mechanism may explain development of some tumors in TSC. An abbreviated discussion of the principal pathologic processes involving nonneurologic systems follows.

Even though genetic testing of TSC has a high rate of identifying TSC gene mutations, the clinical diagnostic criteria (Table 1) are still useful and helpful to establish the diagnosis of TSC. Genetic testing occasionally identifies a patient who has a mutation yet has few clinical findings of TSC, but on the other hand, about 10–20% of individuals with clinically definite TSC will not have a definite gene mutation by current testing methodologies. If, for example, a child with clinically definite TSC has no identifiable gene mutation, it may be helpful to his parents to also be evaluated for the main clinical findings of TSC to determine whether one of them might be a carrier. A brief discussion of the key clinical features follows.

The skin lesions of TSC include facial angiofibromas (Figure 1(a)), hypomelanotic macules and poliosis of the hair (Figure 1(b)), shagreen patches (Figure 1(c)), and other fibromatous skin lesions found on the forehead, scalp, or other body regions, and ungual fibromas. Additionally, pitting and erosions of the dental enamel can occur.

Table 1 Clinical diagnostic criteria for tuberous sclerosis complex[a]

Major features
Facial angiofibromas or forehead plaque
Nontraumatic ungual or periungual fibroma
Hypomelanotic macules (more than three)
Shagreen patch (connective tissue nevus)
Cortical tuber[b]
Subependymal nodule
Subependymal giant cell astrocytoma
Multiple retinal nodular hamartomas
Cardiac rhabdomyoma, single or multiple
Lymphangiomyomatosis[c]
Renal angiomyolipoma[c]
Minor features
Multiple randomly distributed pits in dental enamel
Hamartomatous rectal polyps[d]
Bone cysts[e]
Cerebral white-matter 'migration tracts'[b, e]
Gingival fibromas
Nonrenal hamartoma[d]
Retinal achromic patch
'Confetti' skin lesions
Multiple renal cysts[d]

[a]Definite tuberous sclerosis complex: either two major features or one major feature plus two minor features. Probable tuberous sclerosis complex: either one major feature or two or more minor features. Suspect tuberous sclerosis complex: either one major feature or two or more minor features.
[b]When cerebral cortical dysplasia and cerebral white-matter migration tracts occur together, they should be counted as one rather than two features of tuberous sclerosis complex.
[c]When both lymphangiomyomatosis and renal angiomyolipomas are present, other features of tuberous sclerosis complex should be present before a definite diagnosis is assigned.
[d]Histologic confirmation is suggested.
[e]Radiographic confirmation is sufficient.
Adapted from Roach ES, Gomez MR, and Northrup H (1998) Tuberous sclerosis complex consensus conference: Revised clinical diagnostic criteria. *Journal of Child Neurology* 13: 624–628.

Facial angiofibromas are red to reddish-brown lesions that form on the face in the vicinity of the nose, cheeks, and chin in about 75% of patients with TSC. These lesions, which typically start out as red papules 1–2 mm in diameter, can be seen as early as age 2 years. They can increase in size and number with age. They may become disfiguring, may bleed, and may cause difficulty breathing if close to the nostrils, or difficulty with eyelid closure if close to the eyelid. Various forms of laser treatment may be performed for disfiguring and problematic lesions.

Hypomelanotic macules occur in 90–95% of individuals with TSC. These lesions typically present at birth or in early childhood. In fair-skinned people, they may be difficult to identify without the use of

Figure 1 (a) Multiple facial angiofibromas. (b) Skin hypomelanotic macules and poliosis of the hair. The size of the large hypomelanotic macule and the area of poliosis are atypical. A typical hypomelanotic macule is presented adjacent to the larger one. (c) Shagreen patch. Reprinted from Roach ES and Sparagana SP (2004) Diagnosis of tuberous sclerosis complex. *Journal of Child Neurology* 19: 643–649, with permission.

a Wood's lamp (ultraviolet light). These macules, caused by a normal number of melanocytes with decreased size of melanosomes and melanization, are asymptomatic. Poliosis is much rarer compared to the hypomelanotic macules, and may involve scalp hair and facial hair, such as eyebrows and eyelashes.

The shagreen patch appears in about 30% of individuals with TSC, and tends to appear later in childhood. These 'geographic' patches occur most often in the lumbar and flank regions and appear as raised, pink- to brown-colored skin with a leathery texture. These lesions are benign and do not require treatment.

Ungual fibromas are reddish or flesh-colored nodules arising from the fingernail or toenail beds. These masses can cause disfigurement of the nail, which may be seen as grooves in the nail. These masses are benign. However, they may be easily traumatized and bleed. Surgical resection may be performed, although the fibroma may regrow. Dental pitting and large erosions of tooth enamel are seen in about 90% of adults with TSC. These pits may cause significant problems with dental hygiene.

A variety of ophthalmic and periophthalmic lesions have been reported with TSC. The most significant lesions are nodular astrocytic hamartomas and pigmentary changes (achromic patches). The achromic patches are static and do not cause any harm. While the nodular retinal hamartomas typically do not grow or cause problems, there is potential for growth, and in some cases vision may be compromised. It is recommended that individuals with TSC have at least a baseline evaluation by an ophthalmologist, and those with nodular lesions should have ongoing follow-up.

Cardiac rhabdomyomas occur in half to two-thirds of individuals with TSC but are usually asymptomatic.

When they do cause problems, it is predominantly in infancy. These tumors generally involute over their first two to three years of life. However, large tumors may cause outflow obstruction and result in congestive heart failure. In such cases, surgical intervention is warranted. In some patients, rhabdomyomas may infiltrate the cardiac muscle, and may interrupt electrical conduction pathways in the heart, giving rise to arrhythmias such as Wolff–Parkinson–White syndrome. In such cases, arrhythmias may persist into adulthood and ongoing follow-up by a cardiologist is necessary.

A variety of vascular dysplasias, including coarctation of the aorta, and aortic and cerebral aneurysms, have been described in individuals with TSC. Arterial aneurysms are also common in angiomyolipomas, the most common benign kidney tumor in TSC.

Pulmonary involvement in TSC consists of two primary conditions, multifocal micronodular pneumocyte hyperplasia (MMPH) and lymphangioleiomyomatosis (LAM). MMPH occurs equally in both men and women. MMPH results from hypertrophy of type II pneumocytes, which results in development of small nodules in the lung parenchyma, and these are apparent on high-resolution chest computed tomography (CT) scans. The incidence in TSC is 30–40%, and MMPH has been identified as early as age 10 years. These lesions are benign and cause no pulmonary symptoms.

LAM is seen in about 40% of women with TSC, usually beginning in the third decade. There are rare reports of men with TSC developing LAM. LAM also occurs in some women who do not have TSC, although these women may also have renal angiomyolipomas. LAM results from abnormal proliferation of smooth muscle cells in the lung parenchyma, and results in multiple cysts ranging in size from 15 mm up to several centimeters and can result in severe destruction of normal lung architecture. While LAM cells are benign, there is good evidence showing that these cells may metastasize, with LAM cells having been identified in the kidney and in axial lymph nodes, as well as in lung. The severity and rate of progression of LAM can vary, but LAM can result in severe end-stage cystic lung disease. Symptomatic treatments are commenced as soon as LAM is identified. Antiestrogen agents and progesterone have been used as treatments, although the evidence that they are helpful is slim. For extreme cases, lung transplantation may be necessary. Due to the severity of LAM, it is recommended that adult women with TSC receive a baseline high-resolution CT, with repeat evaluation as needed for pulmonary symptoms.

Kidney involvement in TSC is common, with 80% of individuals showing at least some evidence of renal involvement by age 10 years. Renal disease is also one of the greatest causes of lifelong morbidity in TSC. The most common kidney lesion is a benign tumor called angiomyolipoma (AML). Polycystic kidney disease is rarer, and occurs in about 3–5% of individuals with TSC. Renal cell carcinoma also occurs in TSC with an incidence of less than 1%.

AMLs are present in about 80% of individuals with TSC, and occur equally in males and females. AMLs can occur as early as childhood, and these lesions tend to increase in size and number over time. AMLs are composed of dysplastic blood vessels, which have a tendency to form aneurysms, dysplastic smooth muscle cells, and adipocytes. These tumors are benign; however, there is a greater tendency for them to hemorrhage when they become greater than 4 cm in diameter. The greater the number and size of AMLs, the greater the likelihood there will be complications or loss of kidney function. Embolization is a minimally invasive procedure which has been shown to be very effective in treating AMLs, although this procedure is not typically undertaken until the tumors are about 5 cm in diameter. Rapamycin is currently being investigated as a potential treatment for AMLs.

Polycystic kidney disease (PKD) is another severe renal malady which may be seen in TSC. Many individuals with TSC will develop a few small cysts in the kidneys. However, some individuals develop multiple massive kidney cysts like those seen in adult polycystic kidney disease (APKD). PKD in TSC results from a large mutation bridging both the *TSC2* and *APKD* genes, which are adjacent to one another on chromosome 16, and causes a contiguous gene syndrome with clinical signs and symptoms of both TSC and APKD. The kidney cysts may ultimately replace normal renal parenchyma, resulting in kidney failure and hypertension. Nephrologic management is warranted, and in the extreme, renal transplantation may be necessary.

While uncommon, various renal malignancies have been documented in individuals with TSC, including renal cell carcinoma, oncocytomas, leiomyosarcomas, and others. Renal cell carcinoma appears to be the most common type of malignancy. Lesions may be identified on sonography as being solid and without apparent fat. If fat cannot be clearly identified with sonography or CT imaging, renal biopsy is recommended to determine histological status of the lesion. If a renal neoplasm is identified, surgical resection is warranted as well as a careful investigation for metastasis. Surveillance monitoring of the kidneys with ultrasonography every 1–3 years is recommended to follow the kidney status in individuals with TSC. This approach serves to follow previously identified lesions, and to identify new lesions which might become problematic.

Other organs may also develop benign tumors in individuals with TSC. For example, angiomyolipomas

can develop in the liver of about 25% of individuals with TSC. In contrast to renal AMLs, the hepatic AMLs do not usually enlarge or hemorrhage. Hamartomatous lesions can also develop in the spleen, pancreas, colon, and rectum, and in a variety of endocrine organs. Routine investigation and surveillance monitoring for these tumors is typically not undertaken due to their low morbidity.

Clinical Features of TSC: Neurologic

The brain lesions of TSC result in significant morbidity. Epilepsy, autism, cognitive disability, and a spectrum of psychiatric pathology are all complications of cerebral tubers. Subependymal giant cell astrocytomas (SEGAs) have the potential to cause hydrocephalus, which may result in significant morbidity and even death. Other neurologic lesions have been associated with TSC, although are less likely to cause significant problems.

Cortical and subcortical tubers are the hallmark cerebral lesions of TSC; the tubers (Latin for 'swelling' or 'protuberance') are hard (Greek, 'skleros') on pathological evaluation (**Figures 2(a) and 2(b)**). Tubers are characterized histologically by disruption of the normal cortical lamination pattern, an increase in astrocytes, and disorganization of pyramidal neurons. There are also very large cells, so-called giant cells, which have characteristics of both astrocytes and neurons. Tubers appear to develop during midgestation and have been found as early as 19 weeks of gestation. The exact process by which tubers are formed is not clear. Tubers typically occur in the cerebral cortex but may rarely also be identified in the cerebellum. They may involve cortex, and sometimes extend down into adjacent white matter and other subcortical areas. They may be multiple, and there is a predilection for occurrence in the frontal and parietal lobes. Size of tubers may range from several millimeters to a size which may comprise the bulk of several gyri. On T_2-weighted and fluid-attenuated inversion recovery magnetic resonance imaging (MRI) sequences, they appear as high signal lesions.

Both tuber location and overall tuber burden are responsible for many neurologic complications. Epilepsy occurs in about 80% of individuals with TSC, and is related to some tubers, although there is a question concerning whether seizures arise from the tubers, from the cortex surrounding the tubers, or from both. Seizures may begin in infancy, often as infantile spasms, and epilepsy in TSC may be particularly refractory to medical treatment. In some cases, surgical resection of epileptogenic tubers may be necessary and helpful.

Intellectual and cognitive impairments often occur in TSC. Mental retardation occurs in 40–50% of

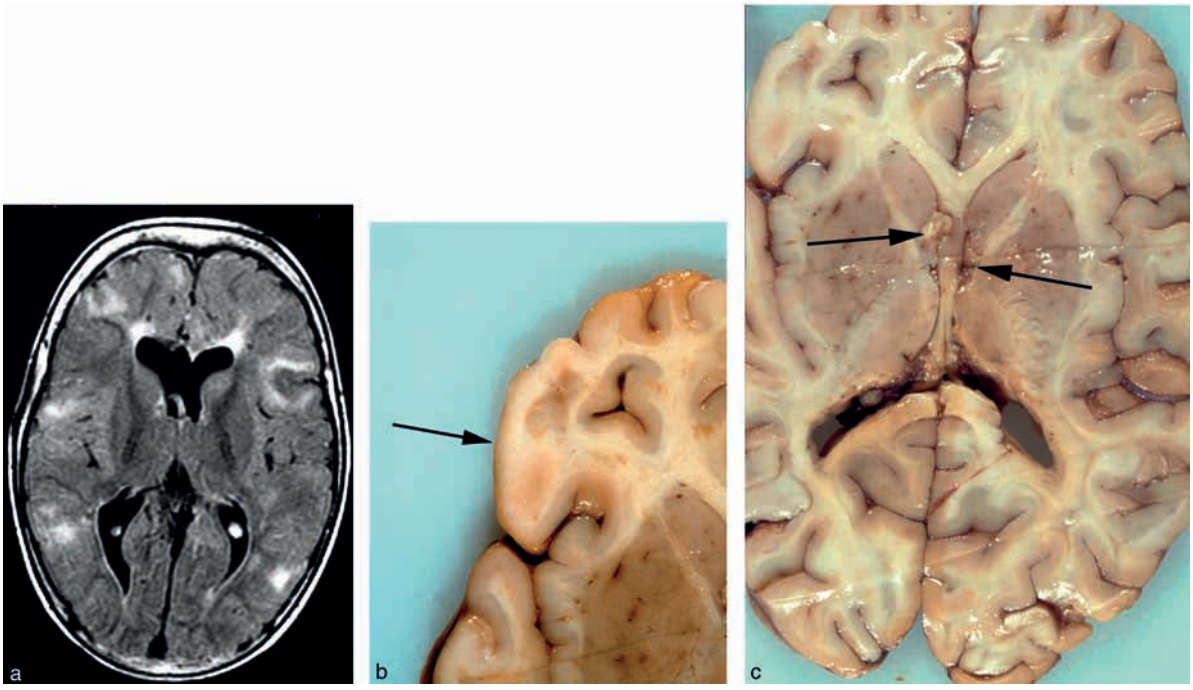

Figure 2 (a) A T_2-weighted magnetic resonance image showing multiple cortical and subcortical tubers (arrows) and subependymal nodules lining the lateral ventricles (arrowheads). (b) Gross brain specimen from a deceased patient with TSC demonstrates a tuber involving the frontal cortex. (c) Gross brain specimen of subependymal nodules (arrows). (b, c) Reprinted from Roach ES (1992) Neurocutaneous syndromes. *Pediatric Clinics of North America* 39: 591–620, with permission.

individuals with TSC, although the degree of impairment may be from mild to profound. Estimates of the prevalence of autism in TSC range from 5% to 60%. Neurocognitive disabilities of TSC are related to tuber burden and location, as well as to the degree of seizure control.

Subependymal nodules (SENs) are nodular hamartomatous growths that are primarily found along the walls of the lateral ventricles. They may be found adjacent to or imbedded in the caudate nucleus and also are frequently located adjacent to the foramina of Monro (Figure 2(c)). There are usually multiple SENs, and they may be very numerous (e.g., 15–25). They range in size from 1 to 10 mm, and they may calcify as cells within the nodule die and are replaced with calcium. As a result, calcified SENs appear as high-density lesions on CT scans. SENs appear as low-intensity lesions on T_1- and T_2-weighted MRI sequences. The cellular makeup of SENs includes cells resembling astrocytes, and giant cells with both astrocyte and neuronal lineage, and the ventricular surface is covered by intact ependyma cells. SENs are benign, yet there may be slow growth, and when SENs grow larger than 10 mm, they are referred to as subependymal giant cell astrocytomas (SEGAs).

SEGAs arise from SENs; however, it is not known what causes this transformation. SEGAs occur in about 15% of individuals with TSC and occur primarily during the first two decades of life. SEGAs usually occur in the lateral ventricles, are not malignant, and do not have a tendency to metastasize (Figures 3(a) and 3(b)). Growth may be very slow, in which case excision is not absolutely necessary. However, when SEGAs in the vicinity of the foramina of Monro grow, they may obstruct cerebral spinal fluid flow, resulting in hydrocephalus. If, on serial imaging, there appears to be rapid growth or ventricular enlargement, or if there are signs and symptoms of hydrocephalus, surgical excision of the SEGA is necessary. Surveillance monitoring has resulted in a significant reduction in the morbidity and mortality related to SEGAs.

Other brain lesions occur in individuals with TSC. For example, linear migration lines streaking through the white matter from the ventricular region to the cortex represent the pathways for migration for dysplastic neurons. Other findings occasionally seen include megalencephaly or hemimegalencephaly, transmantle cortical dysplasia (which lesions also have a tendency to calcify), schizencephaly, and intracranial aneurysms.

Work in Progress

The clinical and molecular complexities of TSC have made the disease a mother lode of information, helping us to understand fundamental cellular processes which have bearing on developmental and cancer biology. There have been tremendous strides made in understanding the natural history of TSC and in improving treatments for the many complications of the disease. Basic scientists are continuing their efforts to clarify the roles of tuberin and hamartin in the various cell signaling pathways, and how they

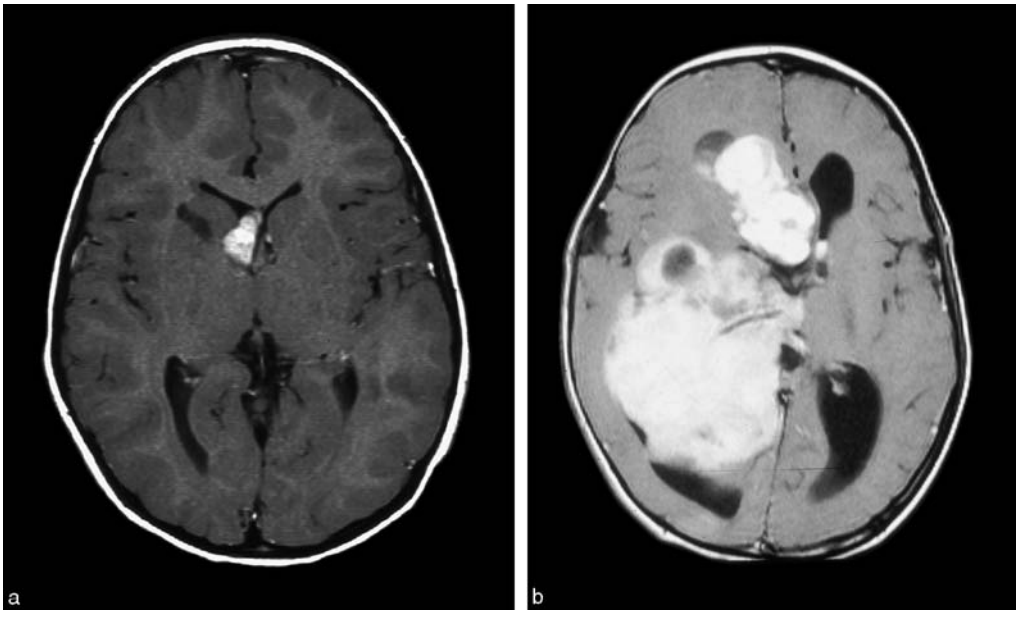

Figure 3 (a) Gadolinium contrast T_1-weighted magnetic resonance image of a moderately sized mass determined to be a subependymal giant cell astrocytoma at surgical resection. (b) Gadolinium contrast T_1-weighted magnetic resonance image of a massive subependymal giant cell astrocytoma, filling much of the lateral ventricle and causing mass effect.

relate to the many clinical manifestations of TSC. A prime example of this is the role of the mTOR inhibitor, rapamycin, in treating renal angiomyolipomas, lymphangioleiomyomatosis, and even SEGAs. Preliminary evidence has demonstrated that treatment with rapamycin results in a significant degree of improvement in these clinical conditions. There are also other major efforts underway to define the natural history of TSC over the human life span, and to evaluate how the clinical findings relate to the genetic variances of TSC. The foremost goal of improving the lives of individuals with TSC is now being achieved through these efforts. The ultimate goal of curing TSC is elusive, yet in some manner, is attainable.

See also: Autism; Epilepsy; Epilepsy: Channelopathies; Tumors of the Brain and the Spinal Cord.

Further Reading

Au K-S, Williams AT, Gambello MJ, et al. (2004) Molecular genetic basis of tuberous sclerosis complex: From bench to bedside. *Journal of Child Neurology* 19: 699–709.

Curatolo P (ed.) (2003) *Tuberous Sclerosis Complex: From Basic Science to Clinical Phenotypes.* London: MacKeith Press.

Gomez MR, Sampson JR, and Whittemore VH (eds.) (1999) *Tuberous Sclerosis Complex.* New York: Oxford University Press.

Maria BL (ed.) (2004) Tuberous sclerosis complex (symposium proceedings). *Journal of Child Neurology* 19: 631–735.

Roach ES (1992) Neurocutaneous syndromes. *Pediatric Clinics of North America* 39: 591–620.

Roach ES, Gomez MR, and Northrup H (1998) Tuberous sclerosis complex consensus conference: Revised clinical diagnostic criteria. *Journal of Child Neurology* 13: 624–628.

Roach ES, DiMario FJ, Kandt RS, et al. (1999) Tuberous sclerosis consensus conference: Recommendations for diagnostic evaluation. *Journal of Child Neurology* 14: 401–407.

Roach ES and Sparagana SP (2004) Diagnosis of tuberous sclerosis complex. *Journal of Child Neurology* 19: 643–649.

Sparagana SP and Roach ES (2000) Tuberous sclerosis complex. *Current Opinion in Neurology* 13: 115–119.

Relevant Websites

http://geneclinics.org – *GeneReviews* website (University of Washington, Seattle, WA).

http://www.aesnet.org – The American Epilepsy Society.

http://www.tsalliance.org – The Tuberous Sclerosis Alliance (USA).

http://www.tuberous-sclerosis.org – The Tuberous Sclerosis Association (UK).

Tumors of the Brain and the Spinal Cord

G Bobustuc and S Baidas, M. D. Anderson Cancer Center Orlando, Orlando, FL, USA

© 2009 Elsevier Ltd. All rights reserved.

Overview

Tumors of the central nervous system (CNS) form a broad spectrum of disease varying from benign to malignant and from primary to metastatic tumors (Table 1). As cancer patients live longer with improved therapy, an increased number of CNS metastatic tumors are being seen. It is estimated that more than 180 000 of the total US cancer patients had CNS involvement by metastatic cancer in 2000.

Primary CNS tumors accounted for 15 new cases per 100 000 persons per year between 1998 and 2002 (45% men, 55% women). Ninety-six percent occur within the cranial cavity and 4% arise in the spine. Brain tumors occur in two peaks: childhood (age 3–12) and later life (age 50–79). An increase in age-adjusted incidence of glial tumors has been noticed since the 1970s. Part of this increase in tumor incidence might be attributed to improvement in imaging technology; further studies are required to look for other causes, especially with the increasing incidence of CNS tumors at younger ages.

CNS tumors represent the second most frequent malignancy in children under the age of 15. The majority of pediatric brain tumors occur below the tentorium. Forty percent are embryonal in type (mainly medulloblastoma), whereas in adults 90% are astrocytomas and occur above the tentorium. Children with brain tumors have a better prognosis than adults, in part because children rarely develop the highly malignant tumors of adults, glioblastomas. Also, low-grade gliomas tend to remain histologically low grade, and, in general, gliomas in children do not demonstrate the same molecular genetic changes identified in adults. Children only rarely develop three common benign tumors of adults: meningiomas, acoustic neuromas, and pituitary tumors.

Spinal cord tumors comprise 4% of CNS tumors. In adults, astrocytomas are less common than ependymomas, comprising only 30–35% of spinal cord gliomas, while in children, astrocytomas and ependymomas comprise 59% and 28%, respectively, of spinal cord tumors.

Tumors of Neuronal and Glial Origin

Brain tumors rarely metastasize outside the CNS with the exception of medulloblastoma and malignant meningioma. In general they have a poor prognosis and a grim outcome. Several clinical and biologic features serve to distinguish primary brain tumors. First, because of their location inside the bony confines of the skull or spinal column, these tumors cause symptoms by direct invasion or compression of normal tissue and by increasing intracranial or intraspinal pressure. The focal organization of brain function, together with poor ability to regenerate and the obvious difficulties of radical surgical excisions, lead to high morbidity and mortality. Of biological interest is the fact that 60% of malignant CNS tumors are of glial origin. Gliomas are prone to infiltrative invasion of normal brain tissue that renders complete surgical excision of such tumors difficult to achieve. These tumors tend to progress from well-differentiated, low-grade to higher grade, poorly differentiated neoplasms with a more aggressive behavior. In addition, compared to the rest of the body, the CNS is a 'privileged' or protected site, largely because the blood–brain barrier blocks access of both natural immune defenses and chemotherapeutic agents. The endothelial cells of the brain vasculature are rich in the multidrug resistance gene (MDR1) contributing to the chemoresistant behavior of CNS tumors.

Other CNS Tumor Types

Aside from tumors of neuronal and glial origin, several other tumor types in the CNS deserve mention. One such tumor, meningioma, of mesodermal origin, comprises 18% of primary intracranial tumors. Meningiomas are mostly benign tumors, the majority of which are surgically curable. Some meningiomas have the tendency to recur or to transform to a higher malignant grade with the capacity to metastasize, behaving like sarcomas. Meningiomas are seen at a higher incidence in women.

Nerve sheath tumors (schwannomas) are mostly benign tumors. They can cause symptoms related to the nerve origin and by compressing the adjacent brain tissues. The most common of these tumors arises from the vestibular part of the eighth cranial nerve and is called acoustic neuroma. Neurofibromatosis consists of two distinct autosomal dominant hereditary disorders.

Neurofibromatosis type one (NF1) is associated with a lesion on chromosome 17q and neurofibromatosis type two (NF2) with a lesion on chromosome 22q. NF1 is characterized by appearance of *café-au-lait* patches and neurofibromas in addition to iris hamartomas (Lisch nodules). NF2 is characterized by bilateral acoustic neuromas and a higher incidence of meningiomas and ependymomas.

Table 1 Tumors of the brain and spinal cord: WHO classification 1993

1. Tumors of neuroepithelial tissue
   - 1.1 Astrocytic tumors
     - 1.1.1 Astrocytoma: variants fibrillary, protoplasmic gemistocytic
     - 1.1.2 Anaplastic astrocytoma
     - 1.1.3 Glioblastoma: variants giant cell glioblastoma, gliosarcoma
     - 1.1.4 Pilocytic astrocytoma
     - 1.1.5 Pleomorphic xanthoastrocytoma
     - 1.1.6 Subependymal giant cell astrocytoma (tuberous sclerosis)
   - 1.2 Oligodendroglial tumors
     - 1.2.1 Oligodendroglioma
     - 1.2.2 Anaplastic oligodendroglioma
   - 1.3 Ependymal tumors
     - 1.3.1 Ependymoma: variants – cellular, papillary clear cell
     - 1.3.2 Anaplastic ependymoma
     - 1.3.3 Myxopapillary ependymoma
     - 1.3.4 Subependymoma
   - 1.4 Mixed gliomas
     - 1.4.1 Oligo-astrocytoma
     - 1.4.2 Anaplastic oligo-astrocytoma
   - 1.5 Choroid plexus tumors
     - 1.5.1 Choroid plexus papilloma
     - 1.5.2 Choroid plexus carcinoma
   - 1.6 Neuroepethelial tumors of uncertain origin
     - 1.6.1 Neuroblastoma
     - 1.6.2 Polar spongioblastoma
     - 1.6.3 Gliomatosis cerebri
   - 1.7 Neuronal and mixed neuronal-glial tumors
     - 1.7.1 Gangliocytoma
     - 1.7.2 Dysplastic gangliocytoma of cerebellum
     - 1.7.3 Dysmoplastic infantile ganglioglioma
     - 1.7.4 Dysembryoplastic neuroepithelial tumor
     - 1.7.5 Ganglioglioma
     - 1.7.6 Anaplastic ganglioglioma
     - 1.7.7 Central neurocytoma
     - 1.7.8 Paraganglioma of the filum terminale
     - 1.7.9 Olfactory neuroblastoma (esthesioneuroblastoma)
   - 1.8 Pineal parenchymal tumors
     - 1.8.1 Pineocytoma
     - 1.8.2 Pineoblastoma
     - 1.8.3 Mixed/transitional pineal tumors
   - 1.9 Embryonal tumors
     - 1.9.1 Medulloepithelioma
     - 1.9.2 Neuroblastoma: variant – ganglioneuroblastoma
     - 1.9.3 Ependymoblastoma
     - 1.9.4 Primitive neuroectodermal tumors (PNETs): medulloblastoma: variants – dysmoplastic medulloblatoma; medullomyoblastoma: variants – dysmoplastic medulloblatoma, medullomyoblastoma, melanotic medullobastoma
2. Tumors of cranial and spinal nerves
   - 2.1 Schwannoma (neurilemmoma, neurinoma) cellular, plexiform, melanotic
   - 2.2 Neurofibroma: circumscribed (solitary), plexiform
   - 2.3 Malignant peripheral nerve sheath tumors (MPNST): neurogenic sarcoma, anaplastic neurofibroma, malignant schwanoma.
3. Tumors of the meninges
   - 3.1 Tumors of meningothelial cells
     - 3.1.1 Meningioma: meningothelial, fibrous, transitional, psamomatous, angiomatous, microcystic, secretory, clear cell, choroid, metaplastic and lymphoplasmacyte-rich
     - 3.1.2 Atypical meningioma
     - 3.1.3 Papillary meningioma
     - 3.1.4 Anaplastic (malignant) meningioma
   - 3.2 Mesenchymal, nonmeningothelial tumors, benign neoplasms: osteocartilaginous tumors, lipoma, fibrous histiocytoma; Malignant neoplasms: hemangiopericytoma, chondrosarcoma, maliganant fibrous histiocytoma, rhabdomyosarcoma, meningeal sarcomatosis
   - 3.3 Primary melanocytic lesions: diffuse melanosis, melanocytoma, malignant melanoma
   - 3.4 Tumors of uncertain histogenesis: hemangioblastoma

Continued

Table 1 Continued

4. Lymphomas and hemopoitic neoplasms
   4.1 Malignant lymphomas
   4.2 Plasmacytoma
   4.3 Granulocytic sarcoma
5. Germ cell tumors: germinoma, embryonal carcinoma, yolk sac tumors, choriocarcinoma, teratoma, mixed germ cell tumors
6. Cysts and tumor-like lesions
7. Tumors of sellar origin: pituitary adenoma, pituitary carcinoma, craniopharyngioma
8. Local extension from regional tumors: paraganglioma, chordoma, chondroma chondrosarcoma, carcinoma
9. Metastatic tumors
10. Unclassified tumors

Pituitary tumors account for about 10% of adult intracranial tumors. They are usually benign tumors but can compress the brain and the optic pathway adjacent to the pituitary at the base of the skull. About one-half of pituitary tumors produce increased amounts of normal pituitary hormones. Increase in prolactin causes amenorrhea/galactorrhea, while increase in growth hormone causes acromegaly and increase in adrenocorticotropic hormone causes Cushing's syndrome.

Pineal region tumors comprise 3–8% of pediatric intracranial tumors (50% under 20 years of age). Most of this group of tumors originate from displaced embryonic tissue remnants or elements of the pineal gland. They include germ cell tumors, pineocytomas, pineoblastomas, and glial-derived tumors. A majority of intracranial germ cell tumors are teratomas that can be treated with radiotherapy alone. Nongerminomatous malignant CNS germ cell tumors have a poor prognosis. Different treatment approaches adding cisplatin-based chemotherapy to conventional surgery and radiation in order to improve survival are being tried with significant survival benefits.

Risk Factors and Molecular Markers

A definite association between environmental or occupational exposure and CNS tumors is lacking. Several reports that did not reach statistical credibility showed increased incidence in CNS tumors with exposure to chemicals, viruses, trauma, magnetic fields, and radiation sources. An increased incidence of primary CNS lymphoma (PCNSL) in immunocompromised individuals has been observed over the last two decades. Most of the PCNSL (more than 90%) are of the large cell, B-cell CD20 positive type. Immunocompromised patients, particularly AIDS patients, have a 17 times higher incidence of PCNSL (0.28 vs. 4.7 in 100 000). Diagnosis relies on imaging and stereotactic biopsy. The mainstay of treatment remains systemic chemotherapy and possibly whole brain radiation.

Tumors occur at a higher rate in individuals with genetic disorders; such syndromes include Li-Fraumeni, Von Hippel Lindau, tuberous sclerosis, Turcot syndrome, and NF1 and NF2.

Recent studies have begun to delineate the genetic abnormalities that underlie CNS tumors, and have implicated two classes of genes: tumor suppressor genes and (proto)oncogenes. P53 on chromosome 17p has been shown to be mutated in low- and high-grade gliomas, while loss of chromosome 9 is an intermediate event and loss of chromosome 10 (site of PTEN, another tumor suppressor gene) is a late event seen mainly in glioblastoma multiforme. Loss of heterozygosity for chromosomes 1p and 19q is often observed in oligodendrogliomas and correlates with an increased sensitivity to chemotherapy and prolonged survival. Meningiomas and schwannomas have shown changes in chromosome 22, especially in females, while medulloblastomas have shown changes in several chromosomes including 1, 7, 8, 17, and 20 together with amplification of c-myc oncogene. Amplification of protooncogenes includes several cytokines and growth factors of which epidermal growth factor receptor (EGF-R) and platelet-derived growth factor receptor (PDGF-R) genes are mainly seen in high-grade gliomas. Increased expression of fibroblast growth factor (FGF) and vascular endothelial growth factor (VEGF) plays an important role in neovascularization required for growth of gliomas.

Diagnosis and Treatment

Treatment of most brain tumors requires a tissue diagnosis, making imaging an important tool in surgical planning. The availability of computed tomography, magnetic resonance imaging (MRI), positron emission tomography, and single photon emission computed tomography has added new dimensions in the accurate localization of CNS tumors, diagnosis, and treatment. Combining three-dimensional imaging with functional MRI helps in providing the anatomical

relation between tumors and the different functional areas in the brain, leading to a better surgical planning and outcome. The most important predictors of survival in higher-grade tumor treatment remain the extent of surgical resection, age, and functional status.

Neurosurgical approaches including microsurgery and stereotactic surgery have reduced surgical morbidity, increased survival, and widened the reach of neurosurgeons. Following surgery, radiation is the mainstay of treatment for most brain tumors. New modalities in radiation therapy from hyperfractionation and stereotactic radiosurgery to brachytherapy have become widely available. Moreover, chemoradiation combining external beam intensity modulated radiotherapy with daily temozolamide has been shown to increase survival and quality of life in patients with high-grade gliomas.

Standard chemotherapy for gliomas has been limited to alkylating agents. Nitrosoureas have been used either alone as BCNU (carmustine) or in combination as CCNU (lomustine) together with procarbazine and vincristine as part of the well-known PCV regimen. Temozolamide is a newer generation oral alkylating agent which is currently a first line treatment of gliomas. Exciting new directions in chemotherapy are entering clinical trials. These include protracted use of (combination) chemotherapy in a metronomic fashion, immunotherapy and differentiation therapy, chemotherapy in combination with kinase inhibitors (interfering with protooncogene-driven cytokines and growth factors), and antiangiogenic therapies or gene therapies. Gene therapy approaches assessed in clinical trials include interference with abnormal gene expression, reconstitution of wild-type (normal) gene phenotype, antisense, and antiangiogenesis.

Conclusion

Although new dimensions in the diagnostic and therapeutic modalities have become available, this has not translated into a dramatic increase in survival in patients with CNS tumors, particularly for gliomas. A better understanding of the biology of such tumors together with a better utilization of the available treatment modalities and the continuous search for new agents and for new treatment approaches will be the future challenge.

See also: Glioma; Medulloblastoma; Neurofibromatosis Type 1: Molecular and Cellular Biology; Neurofibromatosis Type 2: Further Questions and Answers; Neurofibromatosis Type 1; Neurofibromatosis Type 2.

Further Reading

2005–2006 Statistical Report: Primary Brain Tumors in the United States 1998–2002 (Years Data Collected). http://www.cbtrus.org/reports/reports.html (accessed Apr. 2007).

Antinori A, De Rossi G, Ammassari A, et al. (1999) Value of combined approach with thallium-201 single photon emission computed tomography and Epstein-Barr virus DNA polymerase chain reaction in CSF for the diagnosis of AIDS related primary CNS lymphoma. *Journal of Clinical Oncology* 17: 554–560.

Berger MS and Prados MD (2005) *Textbook of Neuro-Oncology*. Philadelphia: Saunders.

Burger PC and Scheithauer BW (1993) *Atlas of Tumor Pathology, Tumors of the Central Nervous System*. Washington, DC: Armed Forces Institute of Pathology.

Cairncross JG, Ueki K, Zlatescu MC, et al. (1998) Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *Journal of the National Cancer Institute* 90: 1473–1479.

DeAngelis LM, Yahalom J, Thaler HT, and Kher U (1992) Combined modality therapy for primary CNS lymphoma. *Journal of Clinical Oncology* 10: 635–643.

Kleihues P and Burger PC (1993) *Histological Typing of Tumors of the Central Nervous System*. Berlin: Springer.

Kleihues P, Burger PC, and Scheithauer BW (1993) The new WHO classification of brain tumours. *Brain Pathology* 3: 255–268.

Leon S, Zhu J, and Black P (1994) Genetic aberration in human brain tumors. *Neurosurgery* 34: 708–722.

Levin VA, Leibel SA, and Gutin PH (2001) Neoplasms of the central nervous system. In: DeVita VT Jr., Hellman S, and Rosenberg SA (eds.) *Cancer: Principles and Practice of Oncology*, 6th edn., pp. 2100–2160. Philadelphia, PA: Lippincott Williams & Wilkins.

Pollack IF (1994) Brain tumors in children. *New England Journal of Medicine* 331: 1500–1507.

Stupp R, Hegi ME, van den Bent MJ, et al. (2006) Changing paradigms: An update on the multidisciplinary management of malignant glioma. *Oncologist* 11(2): 165–180.

Weiner H (1995) The role of growth factor receptors in central nervous system development and neoplasia. *Neurosurgery* 35: 179–193.

Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability

D Thomas and S A N Goldstein, University of Chicago, Chicago, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

A Potassium Channel Superfamily Dedicated to Membrane Leak

At rest, plasma membranes of excitable cells are selectively permeable to 'leak' of potassium (K^+) ions. Potassium leak pathways exert control over excitability by shaping the duration, frequency, and amplitude of action potentials. Leak currents (also called resting or background conductances) active at rest stabilize cells at hyperpolarized voltages below firing threshold, whereas leak suppression permits depolarization and excitation. Conversely, the majority of voltage-dependent potassium (Kv) channels are silent at rest, permitting excitation because they do not interfere with rise to threshold and open with depolarization to actively promote recovery and rapid refiring. Because membrane potential is fundamental to neuronal activity, regulation of potassium leak is a primary and dynamic mechanism for control of cellular excitability.

As expected for key regulators of excitability, leak channels are under tight control from a plethora of stimuli, including lipids, mechanical stretch, neurotransmitters, oxygen tension, pH, and activators of G-protein-coupled receptors and intracellular second messengers. Leak channels are targets for compounds in clinical use, such as volatile and local anesthetics. Leak channels may also influence long-term regulation of excitability through control of gene expression. Regulation of leak channels alters those attributes that are subject to change in any ion channel: the number of functional pores at a site of operation, open probability, and unitary current.

Assigning Membrane Leak to Open Rectifier K^+ Channels

Although the existence of leak currents was recognized more than 50 years ago by Hodgkin and Huxley, they remained a biophysical curiosity for more than four decades. Despite their significant duties and dynamic regulation, native leak currents defied coherent description; they were readily ignored, easily camouflaged, and even actively removed by 'leak subtraction' during electrophysiological measurements.

During the past 10 years, cloning of the superfamily of *KCNK* genes encoding the K_{2P} channels (with 15 mammalian members) has established unequivocally that leaks are indeed more than dribble through pathways devoted to other functions. Identification of the first molecular correlate of a potassium leak current was preceded by cloning of potassium channels in *Saccharomyces cerevisiae* and *Caenorhabditis elegans* with two pore-forming loops in each subunit and eight or four transmembrane (TM) domains, respectively. Thereafter, the *KCNKØ* (*dORK1*) gene for K_{2P}Ø was isolated by functional expression from neuromuscular tissue of *Drosophila melanogaster* on the basis of its capacity to rescue potassium transport-deficient yeast cells. Electrophysiological characterization revealed K_{2P}Ø to be a potassium-selective channel with the predicted attributes of a background leak conductance – that is, a voltage-independent portal showing Goldman–Hodgkin–Katz, or open, rectification.

Potassium channels that show open rectification pass current more readily in one direction (rectify) due to unequal ion concentration across the membrane. Open rectification refers to a nonlinear current–voltage relationship for ions traveling through an open, ion-selective portal and is distinct from rectification due to voltage-dependent changes in channel activity caused by gating or blockade. Current flows more easily from a side of high permeant ion concentration, which explains why K_{2P} leak channels pass outward current under physiological ionic conditions of high internal and low external potassium and have linear current–voltage relationships when potassium concentrations across the membrane are symmetrical (**Figure 1**). Although the underlying assumptions of electrodiffusion theory are inaccurate here (ions do not move independently of each other through the pore), the theory predicts some observed current–voltage relationships reasonably well.

Of course, any potassium channel open below firing threshold will influence resting potential. Thus, it is not surprising that other potassium channels have been associated with background potassium leak. Inward rectifiers, formed by K_{ir} subunits, stabilize cells near E_K. Some classical voltage-gated potassium channel (K_v) subunits are open at rest if assembled with β subunits, including K_v7.1 (in complex with MiRP1/*KCNE2* in heart or MiRP3/*KCNE4* in intestine) and Kv3.4 (in association with MiRP2/*KCNE3* in skeletal muscle). In the central nervous

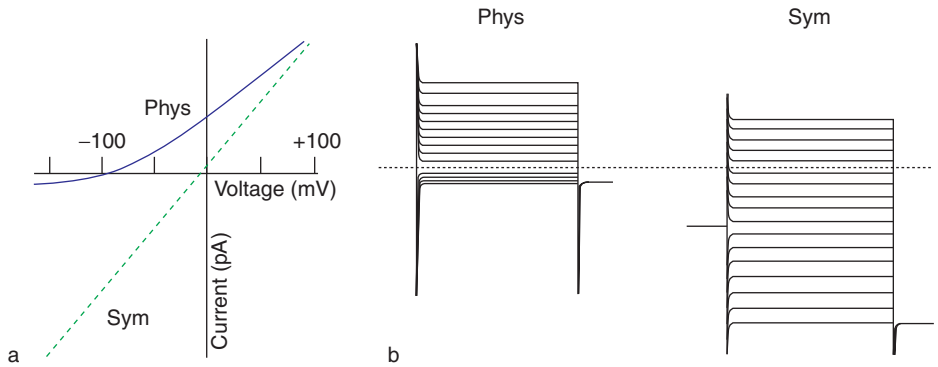

Figure 1 K_{2P} channels are potassium-selective leaks. (a) The channels (in this case, K_{2P}Ø) show Goldman–Hodgkin–Katz (open) rectification and allow the passage of large outward currents under physiological conditions of high internal and low external K^+, whereas they pass currents that change in a linear fashion when potassium levels are symmetrical. (b) Raw current traces from K_{2P}Ø studied by two-microelectrode voltage clamp demonstrating attributes of an open, potassium-selective pore as described in (a). The current traces reveal no delay before current starts to flow because the channels are open before the voltage step. Dashed line is 0 mV; Phys and Sym indicate 5 and 100 mM bath potassium, respectively. Reprinted with permission from Goldstein SAN, Price LA, Rosenthal DN, et al. (1996) ORK1, a potassium-selective leak channel with two pore domains cloned from *Drosophila melanogaster* by expression in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America* 93: 13256–13261. Copyright 1996 National Academy of Sciences, USA.

system, K_v7.X (KCNQ) subunits show a relatively hyperpolarized activation threshold and are implicated in neuronal M-type currents as homomeric K_v7.5 channels and mixed K_v7.2/7.3 complexes.

Potassium Channel Subunits with Two Pore Domains

A striking feature of K_{2P} channels is their subunit body plan: each has two pore (P) loops and four TM domains (Figure 2(a)). As a result, members of this family are referred to as 'two-P-domain,' 'two-pore-domain,' and, formally, 'K_{2P}' channels. This distinct 2P/4TM topology can be found in more than 70 predicted homologs in genome databases. Fifteen mammalian *KCNK* genes encode K_{2P} channels (Table 1 and Figure 2(e)). K_{2P} channels are expressed throughout the body; their distribution in the central nervous system has been particularly well delineated (Table 2).

Classical potassium channel subunits have only one P loop. Because voltage-gated (K_v) and inward rectifier (K_{ir}) channels have six and two TM domains, respectively, they have 1P/6TM and 1P/2TM topology (Figures 2(b) and 2(c)). Two fungal K_{2P} channels have been characterized, and their subunits have 2P/8TM topology (Figure 2(d)). Four P loops aggregate to form a single ion conduction pathway. Thus, K_{2P} channels assemble from two subunits, whereas K_v and K_{ir} channels form with four subunits.

When ion channels are discovered, they are given colloquial names that remain in use until a shared nomenclature is embraced. As a result, identical genes and their subunits carry multiple names (and nonhomologous subunits sometimes share similar root names). A simplified terminology has been instituted by the Human Genome Organization (HUGO) applying the prefix *KCNK* for each gene, and the International Union of Pharmacology classification accords each gene with a K_{2P} ion channel subunit product. These identifiers are presented with some common acronyms for the various isolates in Table 1.

What Is Leak?

From the earliest days of electrophysiological recordings, a fixed conductance called the leak was used to explain the presence of a resting membrane potential from which action potentials emerged and to which they returned. In current parlance, 'leak' is still applied to fixed background currents present at rest and to currents that seem to rise instantly to a new steady level with voltage steps. Leaks can be nonselective (e.g., when a membrane is physically damaged) or result from the movement of specific ions.

Even at rest, the plasma membrane of a living cell is bustling with activity: ions move in and out, and ionic and electrical transmembrane gradients are maintained. The ionic gradient is established by active transport systems (e.g., sodium–potassium pumps). These pumps expend cellular energy (ATP) to keep intracellular levels of sodium low and potassium high. Ion channels dissipate the ionic gradient by opening water-filled pathways across the membrane that allow for free diffusion of selected ions down their electrochemical gradients. An open potassium channel allows potassium ions to flow out of the cell down their concentration gradient. This leaves behind negative counter ions, producing an excess of negative charge inside the cell that limits potassium efflux: net outflow of potassium continues only until

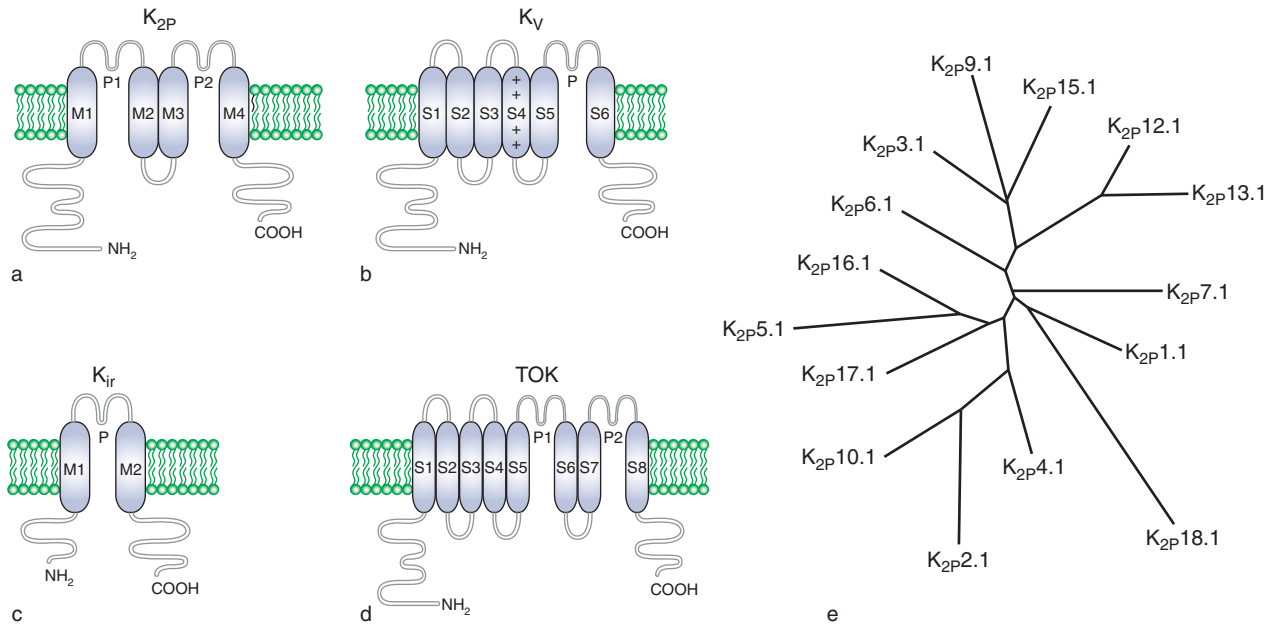

Figure 2 Membrane topology of potassium channels. Subunits are drawn with the extracellular membrane surface facing up. (a) K_{2P} channels have a predicted 2P/4TM subunit topology. (b) Voltage-gated K^+ channel (K_v) subunits have a 1P/6TM domain predicted topology and are noted for their positively charged fourth TM segment (S4) acting as voltage sensor for changes in membrane potential. (c) Inwardly rectifying potassium channel (K_{ir}) subunits have one P domain and two TM segments (1P/2TM topology). (d) TOK1, a fungal K_{2P} channel from *Saccharomyces cerevisiae* with 2P/8TM topology. (e) Phylogenetic tree for human K_{2P} channels. Nucleotide sequence alignment and phylogenetic analysis for the 15 known members of the K_{2P} family were generated using ClustalW software version 1.83 (BLOSUM method) and TreeView. (a–d) Reprinted by permission from Macmillan Publishers Ltd: Goldstein SAN, Bockenhauer D, O'Kelly I, and Zilberberg N (2001) Potassium leak channels and the KCNK family two-P-domain subunits. *Nature Reviews Neuroscience* 2: 175–184, copyright 2001.

the chemical energy favoring efflux is balanced by the electrostatic energy favoring influx. At equilibrium, chemical and electrostatic forces are equal, efflux equals influx, and there is no net movement of potassium ions. The Nernst potential is the membrane voltage that yields equilibrium for a particular ion at a given ionic gradient. For a hypothetical membrane that allows only potassium to cross, the Nernst potential is the same as the equilibrium reversal potential for potassium ions, E_K, and is described in the following equation:

$$E_K = (RT/zF) \times \ln(K_{ex}/K_{in})$$

where R is the gas constant, T is temperature in degrees Kelvin, z is the charge of the ion, F is Faraday's constant, and K_{ex} and K_{in} are the concentrations of potassium ions in the extracellular and intracellular compartments, respectively. For physiological potassium ion concentrations at physiological temperature, RT/zF is $\sim$27 mV and E_K is $\sim$−97 mV. Mammalian cells have resting membrane potentials of −60 to −90 mV because potassium ions are not the only ions that move at rest: for example, leak of sodium and chloride shifts the membrane toward their Nernst potentials (approximately +67 and −93 mV, respectively). Since the concentration gradient of ions across the membrane (which is set by ion pumps) does not change appreciably during an action potential, the membrane potential is highly dependent on the baseline permeability of the cell membrane. The resting membrane potential is often close to E_K because far more potassium channels are open at rest than pathways for other ions. Thus, expression and activation of potassium leak currents is a primary determinant of the resting transmembrane potential.

Excitability depends on the resting membrane potential because action potentials fire when the membrane voltage rises to a threshold level where voltage-gated sodium and/or calcium channels open (and subsequently become inactive). To fire again, these channels must recover from inactivation by returning to a negative potential below threshold (i.e., the resting membrane potential), a shift achieved by outward potassium currents.

Regulation of K_{2P} Channel Surface Expression and Function

Intracellular Trafficking

$K_{2P}3.1$ is expressed throughout the central nervous system. Along with $K_{2P}1.1$, $K_{2P}9.1$, and $K_{2P}10.1$, the

Table 1 The mammalian two-P-domain potassium channels

Gene name (hu)	*Subunit names*	*Chromosome location (hu)*	*Sites of mRNA expression*[a]	*Channel phenotype*	*Accession number (hu)*
KCNK1	K_{2P}1.1 (TWIK-1, hOHO-1)	1q42–43	CNS,[b] CVR, SM, GI, GU*	Open rectifier	NM_002245
KCNK2	K_{2P}2.1 (TREK-1, TPKC1)	1q41	CNS,[b] CVR*, SM, GI, GU	Open or voltage-dependent	AF004711
KCNK3	K_{2P}3.1 (TASK-1, TBAK-1, OAT-1)	2p23	CNS,[b] CVR, SM, GI, GU	Open rectifier	NM_002246
KCNK4	K_{2P}4.1 (TRAAK, KT4.1)	11q13	CNS, CVR, SM, GU	Open rectifier	NM_033310
KCNK5	K_{2P}5.1 (TASK-2)	6p21	CNS,[b] CVR, SM, GI, GU	Open rectifier	NM_003740
KCNK6	K_{2P}6.1 (TWIK-2, TOSS)	19q13.1	CNS,[b] CVR, GI, GU	Open rectifier	NM_004823
KCNK7 (*KCNK8*)	K_{2P}7.1 (TWIK-3)	11q13	CNS, CVR, SM, GI, GU	Not established	NM_033347
KCNK9	K_{2P}9.1 (TASK-3, KT3.2)	8q24.3	CNS, SM, GI, GU	Open rectifier	NM_016601
KCNK10	K_{2P}10.1 (TREK-2)	14q31	CNS, GI, GU	Open rectifier	NM_138317
KCNK12	K_{2P}12.1 (THIK-2)	2p22–p21	CNS, CVR, SM, GI, GU	Not established	NM_022055
KCNK13	K_{2P}13.1 (THIK-1)	14q31–q32	CNS, CVR, SM, GI, GU	Open rectifier	NM_022054
KCNK15 (*KCNK11, KCNK14*)	K_{2P}15.1 (TASK-5, KT3.3)	20q12–20q13	CNS, CVR, SM, GI, GU	Not established	NM_022358
KCNK16	K_{2P}16.1 (TALK-1)	6p21.2–6p21.1	Pancreas	Open rectifier	NM_032115
KCNK17	K_{2P}17.1 (TALK-2, TASK-4)	6p21.1	CNS, CVR, GI, GU	Open rectifier	NM_031460
KCNK18	K_{2P}18.1 (TRESK, TRESK-2)	10q25.3	Spinal cord, dorsal root ganglion	Open rectifier	NM_181840

[a]Based on Northern blot and/or RT-PCR studies in mammalian tissues and categorized by system as follows: CNS, central nervous system; CVR, cardiovascular; SM, skeletal muscle and integument; GI, gastrointestinal; GU, genitourinary.
[b]Protein shown.
TWIK, tandem of P domains in a weak inward rectifying K^+ channel; TREK, TWIK-1 related K^+ channel; TASK, TWIK-related acid-sensitive K^+ channel; TRAAK, TWIK-related arachidonic acid-stimulated K^+ channel; THIK, tandem pore domain halothane inhibited K^+ channel; TALK, TWIK-related alkaline pH activated K^+ channel; TRESK, TWIK-related spinal cord K^+ channel.

channel is believed to contribute to the potassium standing outward current (I_{KSO}) important in cerebellar granule neurons. Furthermore, the channel is responsive to volatile anesthetics and has been implicated in oxygen sensation in both the brain and the carotid body. K_{2P}3.1 has significant sequence and functional homology to K_{2P}9.1.

Regulated surface expression of plasma membrane receptors is crucial for normal cellular physiology. Yeast two-hybrid analyses and biochemical studies with K_{2P}3.1 and K_{2P}9.1 reveal that successful trafficking of the channels to the plasma membrane is the result of a conflict between two opposing trafficking motifs (and competition for the sites by native binding partners): an N-terminal dibasic retention site and a C-terminal release site. Dibasic retention motifs consist of basic amino-acid residues (R, K) that are often in pairs (but seen singly or in triplets) and serve as recognition sites for coatamer proteins such as β-COP. Vesicles coated with coatamer proteins mediate retrograde transport to the endoplasmic reticulum (ER). Conversely, phosphorylation-dependent binding of the ubiquitous intracellular regulator 14-3-3 to the channel C-terminus promotes escape from the ER via suppression of β-COP binding. This mechanism appears to be used by other membrane proteins as well, including nicotinic acetylcholine receptors and major histocompatibility antigen class II-associated invariant chain. A role for p11, a small annexin II subunit, has been proposed in K_{2P}3.1 forward trafficking. This 97-residue protein implicated in endocytosis, exocytosis, and membrane organization has limited tissue distribution, suggesting that it serves in some locales to modulate 14-3-3/β-COP regulation of K_{2P} channel expression.

In reverse fashion, K_{2P}1.1 in renal proximal tubules appears to have a partner that promotes channel internalization and retrograde trafficking away from the cell surface. A bacterial two-hybrid screen identified EFA6, an exchange factor for the small G-protein

Table 2 Expression of mammalian K_{2P} channels in neuronal tissues[a]

Channel	*Cortex*	*Cerebellum*	*Spinal cord*	*Dorsal root ganglion*
$K_{2P}1.1$	+	++	+	++
$K_{2P}2.1$	++	+	+++	++
$K_{2P}3.1$	+	++	+	++
$K_{2P}4.1$	+++	+	+	+++
$K_{2P}5.1$	–	–	+	+++
$K_{2P}6.1$	+	+	+	+++
$K_{2P}7.1$	++	+++	+++	–
$K_{2P}9.1$	+	+++	+	–
$K_{2P}10.1$	++	+++	+	+
$K_{2P}12.1$ (rat)	++	++	NI	NI
$K_{2P}13.1$ (rat)	–	–	NI	NI
$K_{2P}15.1$[b]	+	+	NI	NI
$K_{2P}16.1$[b]	–	–	NI	NI
$K_{2P}17.1$[b]	+	+	NI	NI
$K_{2P}18.1$[b]	–	–	++	++

[a]Based on Northern blot and/or RT-PCR studies in human tissues, unless stated otherwise. +++, strong expression; ++, moderate expression; +, low expression; –, very weak or no expression; NI, not investigated.
[b]Expression data are only available for whole brain.

ADP ribosylation factor 6 (ARF6) involved in membrane protein recycling and actin organization, as a protein partner of $K_{2P}1.1$. $K_{2P}1.1$ and EFA6 physically interact in Mabin–Darby canine kidney cells in an ARF6-dependent manner to favor trafficking from the plasma membrane to recycling endosomes. Since ARF6 is a modulator of endocytosis at the apical membrane of epithelial cells, the EFA6/ARF6 system seems to be a strong candidate to contribute to $K_{2P}1.1$ channel trafficking.

Reversible Modification by SUMO Peptide Silences and Activates

Some *KCNK* genes fail to produce currents when expressed in experimental cells (*KCNK1*, *-7*, *-12*, and *-15*). Synthesis of mRNA for the genes in native cells indicates these are not pseudogenes. It was therefore speculated that they might not be targeted correctly to the plasma membrane in experimental systems (which is the case for $K_{2P}7.1$), might have a predominantly intracellular location, might need to assemble with another K_{2P} subtype to operate, or might require an unidentified accessory or regulatory subunit to be activated. The last mechanism has been demonstrated for $K_{2P}1.1$ channels.

$K_{2P}1.1$ is widely expressed with abundant transcript in brain (cerebellum and thalamus), heart, and kidney. Although $K_{2P}1.1$ was first reported to be a weak, inwardly rectifying potassium channel, and hence given the alias TWIK (Table 1), concurrent studies were unable to observe functional expression. $K_{2P}1.1$ has been shown to be inactive, despite residence in the plasma membrane, as long as the protein is covalently modified by a small ubiquitin-related modifier protein, SUMO, at a single C-terminal lysine residue (K274). SUMO was previously associated with the nuclear import/export, activity, and half-life of transcription factors. The posttranslational SUMO modification pathway has since been observed to function in its entirety at the plasma membrane. Desumoylation (removal of covalently bound SUMO) of $K_{2P}1.1$ by the SUMO-specific protease SENP1 relieves chronic silencing, thereby revealing that the protein is a classical K_{2P} open rectifier (Figure 3).

Abundant $K_{2P}1.1$ transcripts in cerebellar granule neurons have suggested a role for the channel in the standing outward potassium leak current, I_{KSO}. Characterization of I_{KSO} suggests that it is composed of four distinct components with attributes similar to those of $K_{2P}3.1$, $K_{2P}9.1$, $K_{2P}10.1$, and a fourth, previously unidentified, subunit now recognized from its biophysical properties to be a good match for $K_{2P}1.1$ channels.

Differential Modulation by Changes in pH

Deviations from normal physiological pH are monitored by multiple sensors *in vivo*. Tissue acidosis is relevant to the pathophysiology of cancer, inflammation, and ischemia, and it contributes to acute pain (e.g., gastroesophageal reflux and peptic ulcer) and inflammatory hyperalgesia, in which afferent nerve fibers are sensitized. Acid sensors in the gut control gastric acid homeostasis, whereas those in the circulatory and central nervous system regulate breathing and arousal. Sensory neurons respond to acidification of their environment via modulation of cellular excitability using at least four acid-sensitive ion channel families: voltage-gated, sodium-selective, acid-sensing ion channels (ASIC1–3); transient receptor potential cation channels (TRPV1 and-V4); ligand-gated, purinoceptor, cation channels ($P2X_{1-5}$ and $P2X_7$); and K_{2P} potassium channels (Table 3).

$K_{2P}3.1$ and $K_{2P}9.1$ are particularly sensitive to extracellular pH variations in the physiological range, displaying pKa values of 7.2 and 6.7, respectively (Figure 4). *In vitro*, these channels can form heterodimers which exhibit intermediate pH sensitivity. A single histidine residue adjacent to the potassium selectivity filter within the first pore-forming loop represents the extracellular proton binding site in both $K_{2P}3.1$ and $K_{2P}9.1$. *In vivo*, decreased brain pH due to hypercapnia induces enhanced respiration

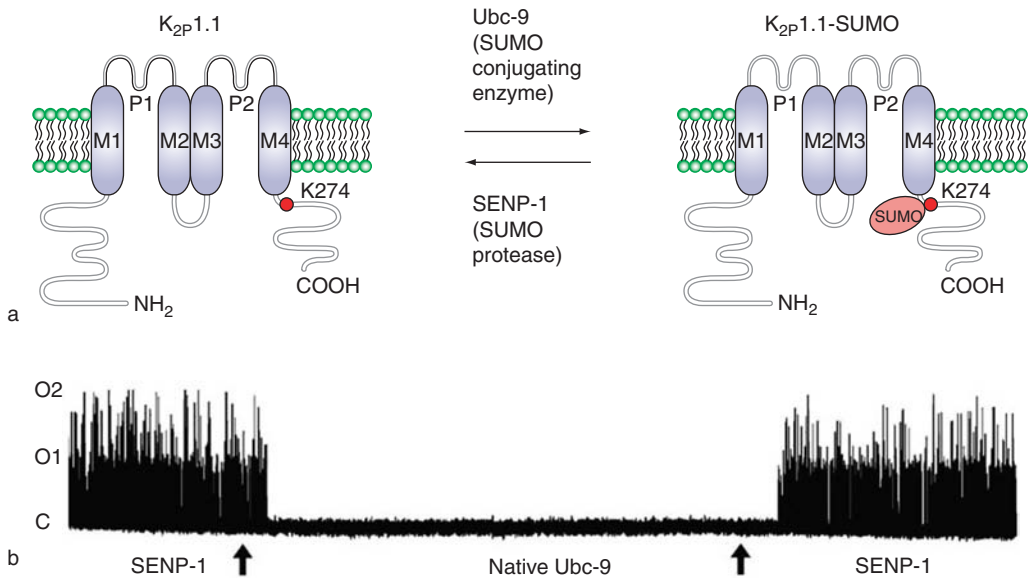

Figure 3 $K_{2P}1.1$ channel activity is regulated by reversible, covalent modification by SUMO. (a) Sumoylation silences $K_{2P}1.1$. Linkage of SUMO to the lysine at position 274 (K274) of $K_{2P}1.1$ by native SUMO conjugation machinery requires the E2 conjugating enzyme Ubc-9 and can be reversed by the desumoylating protease SENP-1. (b) Reversible SUMO modification of $K_{2P}1.1$ and its regulation of channel activity, studied in inside-out patches excised from *Xenopus* oocytes. Active $K_{2P}1.1$ channels in patches excised from cells expressing $K_{2P}1.1$ and the SUMO protease SENP-1 become silent on exposure of the inner channel surface to native Ubc-9 in the interior of a naive cell. Channels return to full activity when the patch is withdrawn and inserted into a cell expressing only SENP-1. Arrows indicate points of patch insertion into each new cell. (b) Reproduced from Rajan S, Plant LD, Rabin ML, Butler MH, and Goldstein SAN (2005) Sumoylation silences the plasma membrane leak K^+ channel K2P1. Cell 121:37–47, copyright 2005, with permission from Elsevier.

Table 3 Acid sensitivity of functional mammalian K_{2P} channels

Channel	*Acidic intracellular pH*	*Acidic extracellular pH*
$K_{2P}1.1$	↓	↓
$K_{2P}2.1$	↑	↓
$K_{2P}3.1$	0 (human), ↓ (rat)	↓
$K_{2P}4.1$	0[a]	↓
$K_{2P}5.1$	↓	↓
$K_{2P}6.1$	↓	0 (human), ↓ (rat)
$K_{2P}9.1$	0	↓
$K_{2P}10.1$	↑	NI
$K_{2P}13.1$	0	(↓)
$K_{2P}16.1$	0	↓[b]
$K_{2P}17.1$	0	0[b]
$K_{2P}18.1$	↓[a]	↓[b]

[a]Activated by intracellular alkalinization.
[b]Activated by extracellular alkalinization.
↓, current inhibition; ↑, current activation; 0, no effect; NI, not investigated.

(in an effort to reduce pCO_2) and behavioral arousal, serving a protective function, notably during sleep. $K_{2P}3.1$ and $K_{2P}9.1$ currents have been detected in somatic motor neurons, noradrenergic locus coeruleus neurons, and serotoninergic raphe neurons. Since the two latter neuronal types are implicated in central respiratory chemoreception, $K_{2P}3.1$ and $K_{2P}9.1$ have been proposed as molecular mediators of pH-dependent excitability in these cells to regulate breathing and arousal. Indeed, in respiration-related motor neurons, reduction in $K_{2P}3.1$ and $K_{2P}9.1$ currents due to acidification of the surrounding tissue may directly increase excitability.

Other K_{2P} channels alter their activity with intracellular acidosis. Thus, $K_{2P}1.1$, $K_{2P}3.1$, $K_{2P}5.1$, $K_{2P}6.1$, and $K_{2P}18.1$ are inhibited, whereas a mechanosensitive channel, $K_{2P}2.1$, shifts its pressure–activation relationship upon protonation of a C-terminal glutamate. In other channels, the influence of lowered pH may be indirect since residues that mediate the responses have not been identified and methods used to lower pH (i.e., CO_2, dinitrophenol, or NH_4Cl) impact a number of cellular pathways that could act on the channels secondarily.

Phenotypic Flexibility: A Leak That Can Also Operate as a Voltage-Gated Channel

Expressed robustly in the central nervous system, especially in the hippocampus, $K_{2P}2.1$ channels are regulated by a multitude of chemical and physical factors. Single $K_{2P}2.1$ channels show unexpected and dynamic versatility: native hippocampal and cloned $K_{2P}2.1$ channels transform in a regulated and reversible fashion from potassium-selective leak

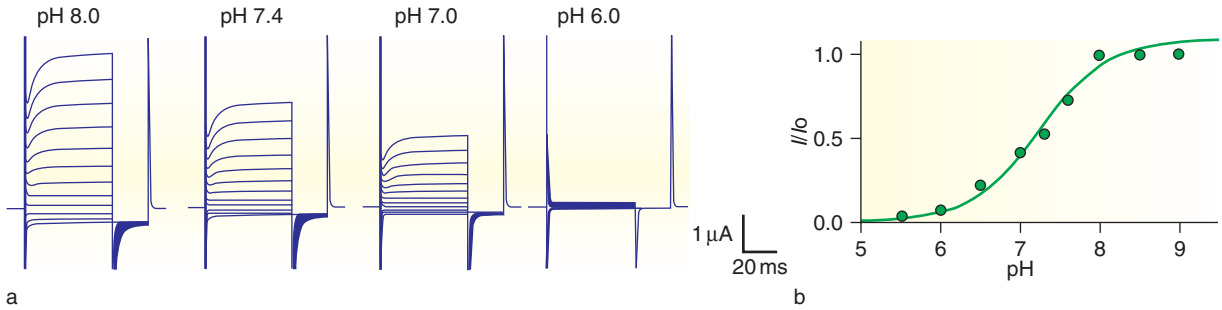

Figure 4 Cloned $K_{2P}3.1$ channels are inhibited by external protons at physiological pH. $K_{2P}3.1$ channels expressed in *Xenopus* oocytes were studied using two-electrode voltage clamp. (a) A cell exposed to varying pH with 5 mM K^+ bath solution is shown. (b) The plot shows dependence of $K_{2P}3.1$ currents at 0 mV membrane potential on extracellular pH for groups of five cells normalized to pH 8, giving a pKa of 7.24. Reprinted from Lopes CMB, Gallagher PG, Buck ME, Butler MH, and Goldstein SAN (2000) Proton block and voltage gating are potassium-dependent in the cardiac leak channel Kcnk3. *Journal of Biological Chemistry* 275: 16969–16978, with permission.

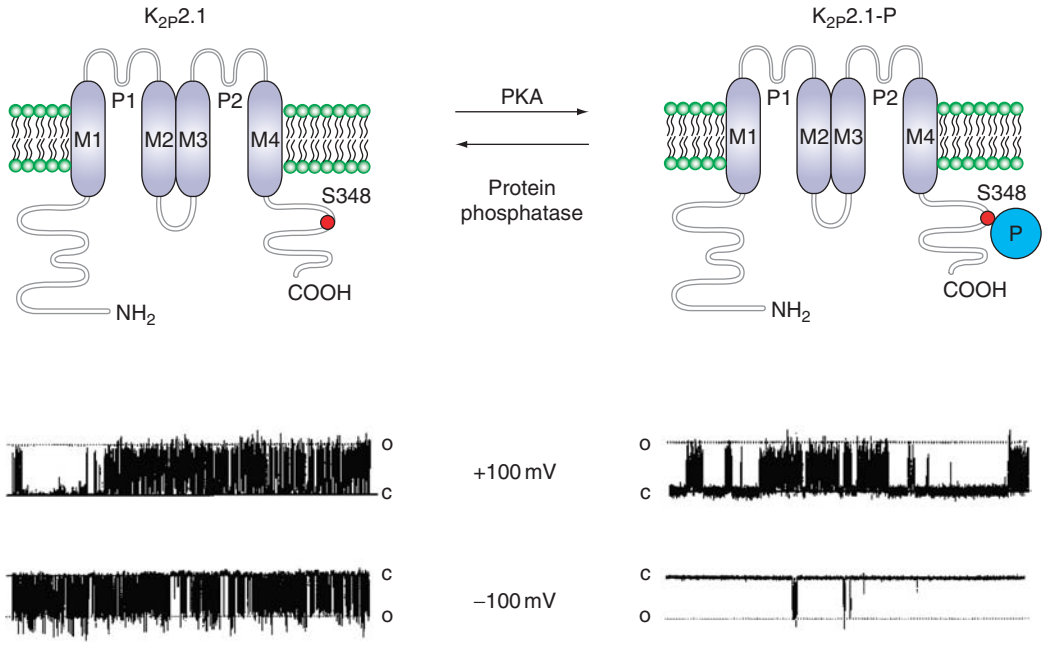

Figure 5 PKA-mediated downregulation of $K_{2P}2.1$. Inside-out patches were excised from oocytes expressing the channels and studied in symmetric 100 mM KCl. $K_{2P}2.1$ is an open rectifier when the PKA phosphorylation site S348 bears no phosphate (recapitulated here by mutation to alanine), passing inward and outward currents in symmetrical K^+ conditions. When the PKA site is negatively charged via phosphorylation or mutation to aspartate (shown here), the open probability of the channel decreases and both cloned and native hippocampal $K_{2P}2.1$ channels convert from the open rectifier to the voltage-dependent phenotype. Reprinted by permission from Macmillan Publishers Ltd. Bockenhauer D, Zilberberg N, and Goldstein SAN (2001) KCNK2: Reversible conversion of a hippocampal potassium leak into a voltage-dependent channel. *Nature Neuroscience* 4: 486–491, copyright 2001.

conductances to strictly voltage-dependent channels. Thus, $K_{2P}2.1$ is an open rectifier when the single canonical protein kinase A (PKA) consensus site is mutated to alanine or bears no phosphate, passing large currents both inward and outward when potassium levels are symmetrical across the membrane (Figure 5). Conversely, $K_{2P}2.1$ channels altered to aspartate at the site or bearing phosphate pass more outward current even under symmetrical conditions due to voltage-dependent changes in open probability. These channels show a half-maximal activation voltage >0 mV that does not change with the equilibrium reversal potential for potassium (E_K) and therefore behave like classical voltage-gated channels. This phenotypic flexibility is expected to contribute to neuronal plasticity because leak currents inhibit depolarization toward firing threshold, whereas voltage-gated channels activated at supra-threshold potentials do not interfere with rise to threshold but do facilitate recovery and refiring.

Stretch-Gating and Mechanoelectrical Coupling

Mechanosensitive potassium channels are found in central and peripheral neurons, as well as in nonnervous tissues such as the heart and gastrointestinal tract. Native, stretch-activated potassium channels in rat atria, for example, share expression pattern and functional properties with $K_{2P}2.1$, $K_{2P}4.1$, and $K_{2P}10.1$ channels. In experimental cells, mechanical pressure applied to the plasma membrane increases $K_{2P}2.1$, $K_{2P}4.1$, and $K_{2P}10.1$ currents. Convex membrane curving toward the outside of the cell is notably effective in activating the channels. Similar to native stretch-sensitive currents, $K_{2P}2.1$ and $K_{2P}4.1$ channels show pronounced desensitization which is thought to be critical for maintaining dynamic sensitivity and protecting cells from persistent activation. In what may be a related effect, $K_{2P}2.1$ and $K_{2P}4.1$ activities reversibly decrease when osmolarity of the extracellular medium is increased and cell size decreases. Sensitivity to mechanical stretch is enhanced after pharmacological or mechanical disruption of the actin cytoskeleton, suggesting that force may be transmitted directly to the channels through the lipid bilayer, independent of cytoskeletal structures. The relationship between channel activity and pressure is sigmoidal, with a half-maximal activation pressure of approximately −40 and −70 mmHg for $K_{2P}2.1$ and $K_{2P}10.1$, respectively. As noted previously, protonation of $K_{2P}2.1$ shifts the current–pressure relationship to increase activity.

In addition to abundant and overlapping expression of $K_{2P}2.1$, $K_{2P}4.1$, and $K_{2P}10.1$ in brain, human transcripts are found in a variety of tissues routinely exposed to mechanical stretch. $K_{2P}2.1$ is found in heart, stomach, small intestine, colon, bladder, and uterus; $K_{2P}10.1$ gains gastric and uterine expression; and $K_{2P}4.1$ is found in heart. These K_{2P} channels might serve to link stretch and membrane potential, thereby directly transforming mechanical force into altered electrical activity. Such mechanoelectrical coupling might contribute to dysregulation of excitability in the brain and heart caused by ischemia-associated cell swelling (as seen with stroke or myocardial infarction) or abnormal stretch (as associated with altered intracranial pressure or heart failure) leading to seizures or cardiac arrhythmias.

The hypothesis that K_{2P} currents contribute to regulation of contractions in the digestive system, bladder, and uterus is supported by studies of the nematode *Caenorhabditis elegans*, which possesses the largest number of K_{2P} channels cloned to date (~50), two of which have been implicated in muscle contraction. Thus, a mutant of the *C. elegans* K_{2P} channel TWK-18 is associated with a paralysis phenotype, whereas a gain-of-function mutation in TWK-23 results in uncoordinated movements of the worm.

Neurotransmitter Regulation

Slow changes in neuronal excitability (neuromodulation) produced by neurotransmitters are distinct from their rapid actions on ligand-gated ion channels. Inhibition and activation of potassium channels are among primary cellular mechanisms responsible for neuromodulation. Inhibition of background potassium conductances in many areas of the central nervous system (including brain stem aminergic neurons, cerebellar granule neurons, cortex, thalamus, hippocampus, and hypothalamus) is a significant mechanism by which acetylcholine, glutamate, noradrenaline, serotonin, substance P, and thyrotropin-releasing hormone enhance excitability. This mechanism is crucial to the regulation of 'state switching' in the thalamocortical network that is thought to control transitions between sleep and wakefulness. Identified channels that appear to contribute include $K_{2P}3.1$ in native rat hypoglossal motor neurons and cerebellar granule cells, in which it is responsive to norepinephrine, serotonin, substance P, and thyrotropin-releasing hormone, like its cloned counterpart in experimental cells (Figure 6). Regulation of K_{2P} channels by neurotransmitters is not limited to $K_{2P}3.1$ channels alone (Table 4). For example, $K_{2P}2.1$ is inhibited by serotonin via the cAMP–PKA pathway similar to native S-type K^+ currents in the sea slug *Aplysia californica* that are involved in behavioral sensitization. Native potassium leak channels subject to neurotransmitter regulation in other regions of the brain have yet to be identified at the molecular level.

Emerging Roles for K_{2P} Channels in Physiology and Disease

Expression of K_{2P} channels has been studied in detail at the mRNA level (Table 2). Genes encoding six of the K_{2P} channels ($K_{2P}5.1$, $K_{2P}6.1$, $K_{2P}13.1$, $K_{2P}15.1$, $K_{2P}16.1$, and $K_{2P}17.1$) are absent or only weakly expressed in the central nervous system (although $K_{2P}5.1$ and $K_{2P}6.1$ are found in dorsal root ganglia). The remaining genes are differentially distributed throughout the central nervous system and dorsal root ganglia. K_{2P} channel mRNA is also found in other brain cell types, such as choroid plexus ($K_{2P}1.1$), pia mater ($K_{2P}2.1$), and ependymal ventricle-lining cells ($K_{2P}10.1$). To date, there are no selective blockers of K_{2P} channels, so molecular identification of currents in native contexts has relied on comparisons of functional properties of cloned K_{2P} channels and native currents. Outside the nervous system, K_{2P}

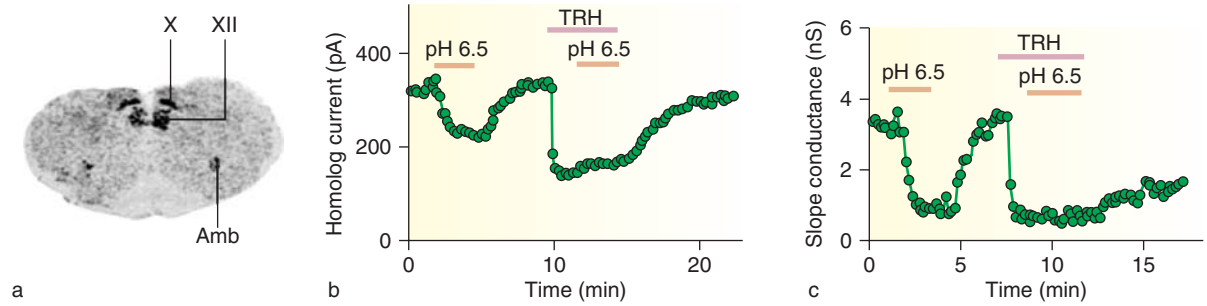

Figure 6 $K_{2P}3.1$ channels: correlation with native neuronal currents. (a) *In situ* hybridization of rat brain stem with a ^{33}P-labeled antisense probe to rat $K_{2P}3.1$ cDNA shows a strong signal for message in motor neurons – the nucleus ambiguous (Amb) and the vagal (X) and hypoglossal nuclei (XII). (b) Thyrotropin-releasing hormone (TRH) inhibits $K_{2P}3.1$-like current in hypoglossal motor neurons treated with low pH solution before and after TRH-induced changes in the holding current, demonstrating that the pH-sensitive $K_{2P}3.1$-like current was completely blocked by 1 or 2 μM TRH. (c) TRH inhibits cloned $K_{2P}3.1$ channels in HEK 293 cells when they also express the TRH receptor. Slope conductance from linear fits of ramp currents measured show that channels are fully inhibited by 0.1 μM TRH so that further effect was not observed by reacidification. Reproduced Talley EM, Lei Q, Sirois JE, and Bayliss DA (2000), Task-1, a two pore domain K^+ channel, is modulated by multiple neurotransmitters in motoneurons. *Neuron* 25: 399–410, copyright 2000, with permission from Elsevier.

Table 4 Functional K_{2P} channel regulation by neurotransmitters and signal transduction pathways

Channel[a]	*Activating GPCRs*	*Inhibiting GPCRs*	*Activators*	*Inhibitors*
$K_{2P}1.1$			Phorbol esters	
$K_{2P}2.1$	G_i (mGluR2)	G_q (mGluR1, TRHR1, Orx1R); G_s (5-HT4sR, β-AR)	Nitric oxide, cGMP, PKG, PIP_2	Phorbol esters, DAG, cAMP, PKA, PKC
$K_{2P}3.1$		G_q (mGluR1, Ang II type 1a receptor, ACh M1, TRHR1)	PIP_2	PLC, PKCε
$K_{2P}4.1$			PIP_2	
$K_{2P}5.1$			Phorbol esters	
$K_{2P}6.1$			Phorbol esters	
$K_{2P}9.1$		G_q (mGluR1, TRHR1)	PIP_2	
$K_{2P}10.1$	G_i (mGluR2)	G_q (mGluR1, ACh M3); G_s (5-HT4sR)		DAG, phorbol esters, cAMP, PKA
$K_{2P}18.1$			Ca^{2+} (via calcineurin)	

[a]$K_{2P}13.1$ is functional but not yet studied. $K_{2P}16.1$ and 17.1 are insensitive to activators of PKA and PKC.
ACh M1 and M3, muscarinic acetylcholine receptors; Ang II, angiotensin; β-AR, β-adrenoceptor; DAG, diacylglycerol; GPCR, G-protein-coupled receptor; mGluR1 and R2, metabotropic glutamate receptors; 5-HT4sR, serotonin receptor; Orx1R, orexin receptor; PIP_2, phosphatidylinositol 4,5-bisphosphate; PKA, protein kinase A; PKC, protein kinase C; PKG, protein kinase G; PLC, phospholipase C; TRHR1, thyrotropin-releasing hormone receptor.

channels have been localized in heart and a variety of nonexcitable organs (Table 1). Specific expression patterns and physiological and clinical implications are discussed in the following sections.

Oxygen Sensing and Respiration

Molecular oxygen sensors mediate rapid cardiovascular and metabolic responses to environmental hypoxia, such as increase in respiration rate and vasoconstriction. Located in highly specialized cells, including carotid body cells, pulmonary neuroepithelial body cells, fetal adrenomedullary chromaffin cells, and pulmonary artery smooth muscle cells, these specialized cells share the same general mechanism which transduces decreased oxygen tension into cellular responses. Thus, closing of oxygen-sensitive potassium channels depolarizes oxygen-sensitive cells, leading to increased excitability via opening of voltage-gated Ca^{2+} channels, elevation of intracellular Ca^{2+} level, and cell-type-specific cellular responses.

Among a variety of potassium channels suppressed by hypoxia in peripheral chemoreceptors, there is good evidence that one member of the K_{2P} family, $K_{2P}3.1$, is involved in an oxygen-sensing pathway. Contrary to earlier reports, data from several laboratories suggest that $K_{2P}2.1$ channels are not oxygen sensitive under physiological conditions. The carotid body located above the bifurcation of the common

carotid artery is the major peripheral site for oxygen sensation. $K_{2P}3.1$ channels are expressed in carotid body type 1 cells and a native oxygen-sensitive channel in the cells shares biophysical and pharmacological properties of cloned $K_{2P}3.1$, most notably inhibition by hypoxia. This suggests that the response of these cells to hypoxia is at least partially due to inhibition of $K_{2P}3.1$ leading to membrane depolarization and reflex increase in respiration. Hypoxia-induced closing of native $K_{2P}3.1$-like channels is thought to be mediated by cytosolic second messengers rather than direct oxygen sensing by the channel since channels in excised membrane patches are not subject to regulation by oxygen tension.

$K_{2P}3.1$ channels are also expressed in central respiratory-related neurons, including airway motor neurons and putative chemoreceptor neurons of locus coeruleus. In these central neurons, hypoxia-induced inhibition of $K_{2P}3.1$-like currents increases excitability to enhance respiratory motor neuron output. Thus, modulation of $K_{2P}3.1$ currents appears to contribute at both the central and peripheral levels to the respiratory reflex.

Central Nervous System

Neuromodulation, learning, and memory A classical example of neuromodulation in a behavioral context is the gill withdrawal reflex of the sea slug *Aplysia californica*. Gill withdrawal is a defensive reflex that is potentiated by a noxious stimulus to another part of the body. This form of behavioral plasticity involves inhibition of the *Aplysia* serotonin (S)-type background potassium channel. S-type channels are expressed in slug sensory neurons, in which they are known to contribute to synaptic signal transmission. $K_{2P}2.1$ channels show notable similarities to *Aplysia* S-type channels. Both channels show open rectification and time independence. They are similarly resistant to blockade by barium, cesium, TEA, and 4-AP and sensitive to quinidine. $K_{2P}2.1$ and S-type channels are both activated by volatile anesthetics, arachidonic acid, and membrane stretch and blocked by serotonin via cAMP-dependent phosphorylation. In *Aplysia* and humans, neurotransmitter-mediated inhibition of S channels causes enhanced excitability. In the mollusk *Lymnaea stagnalis*, an S-type channel is responsible for control of presynaptic transmitter release in behavioral sensitization, a simple form of learning and memory. Because $K_{2P}2.1$ is expressed in human cerebral cortex and hippocampus, it is tempting to speculate that the channel may play a role in cognitive processes, perhaps via arachidonic acid and its metabolites acting as retrograde messengers in hippocampal long-term plasticity.

Actions of volatile and local anesthetics Volatile anesthetics induce neuronal hyperpolarization and depression in the central nervous system. Despite their wide clinical use during approximately the past 150 years, relatively little consensus exists on their molecular mechanisms of action. Several ion channels, including $GABA_A$ and glycine receptors, have been implicated in the effects of these drugs. Soon after their discovery, K_{2P} channels were considered prime candidates as targets for anesthetics. The idea was attractive because anesthetics decrease excitability by enhancing leak currents in neuronal tissue and, at the same time, the presence of K_{2P} channels in heart could explain the cardiodepressive side effect of anesthetics.

Volatile anesthetics open background potassium channels and hyperpolarize resting membrane potential in both invertebrates and vertebrates. In *Aplysia*, halothane induces neuronal hyperpolarization via opening of S-type potassium currents. In the mollusk *Lymnaea*, halothane opens the baseline K^+ current, I_{KAn}, which hyperpolarizes pacemaker neurons and reversibly blocks their endogenous firing activity. In rat, halothane activates native $K_{2P}3.1$ currents in hypoglossal and locus coeruleus neurons, which hyperpolarizes the cells, decreasing spike activity. This is thought to contribute to halothane-induced immobilization, analgesia, and hypnotic actions of the drug (**Figures 7(a) and 7(b)**).

Supporting a role for K_{2P} channels as clinically relevant targets for volatile anesthetics during general anesthesia, knockout mice disrupting the gene for $K_{2P}2.1$ show significant reduction in sensitivity to halothane and desflurane, whereas $K_{2P}5.1$ knockout mice do not. *In vitro* studies demonstrate that volatile anesthetics open cloned $K_{2P}2.1$, $K_{2P}3.1$, $K_{2P}5.1$, $K_{2P}9.1$, $K_{2P}10.1$, and $K_{2P}17.1$ channels (**Figure 7(c)**). Interestingly, the yeast K_{2P} channel, TOK1, is also opened by volatile anesthetics, demonstrating evolutionary conservation of this pharmacology. Taking into consideration K_{2P} expression patterns in human brain, anesthetic-induced activation of $K_{2P}2.1$ and $K_{2P}10.1$ in caudate nucleus and putamen, $K_{2P}9.1$ and 10.1 in cerebellum, and $K_{2P}10.1$ in corpus callosum provide a rationale for the cascade of observed pharmacological effects during general anesthesia. Furthermore, activation of $K_{2P}2.1$, $K_{2P}3.1$, and $K_{2P}5.1$ in sensory neurons of dorsal root ganglia could contribute to loss of consciousness through inhibition of sensory neuronal stimuli. Finally, activation of $K_{2P}2.1$ channels in motor neurons may inhibit muscular activity, directly depressing mobility.

Local anesthesia can be achieved pharmacologically by blockade of sodium channels and a variety of potassium currents, including voltage-gated and leak channels, through disruption of action potential

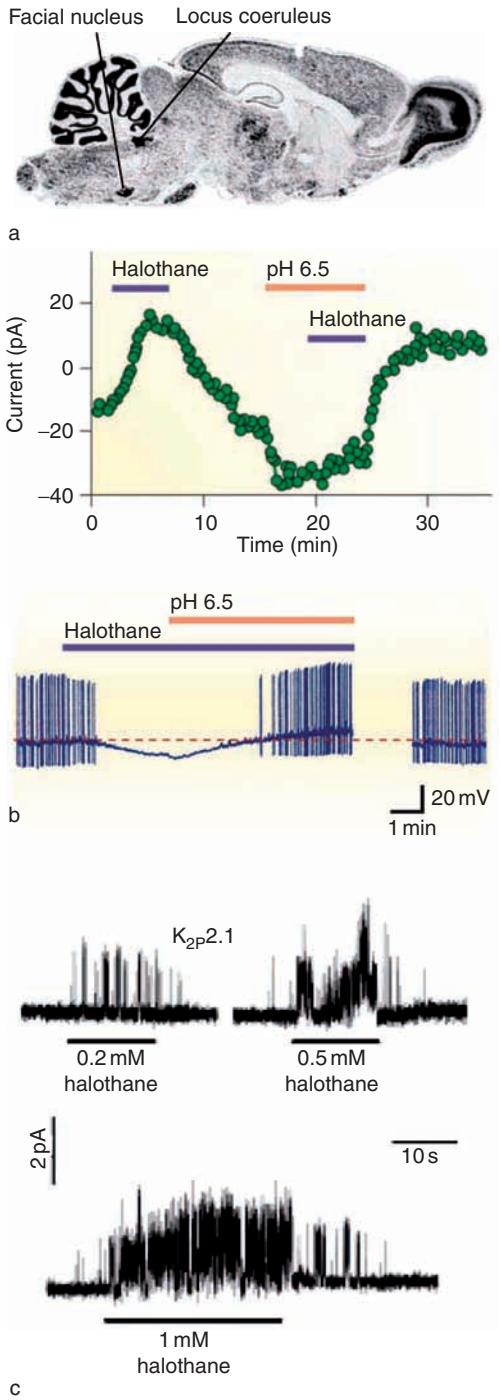

Figure 7 Volatile anesthetics open K_{2P} channels and hyperpolarize resting membrane potential in neurons. (a) *In situ* hybridization of a sagittal rat brain section hybridized with a ^{33}P-labeled specific $K_{2P}3.1$ probe shows a strong signal for message in facial nucleus and locus coeruleus. (b) Halothane (0.3 mM) hyperpolarizes neurons of the rat locus coeruleus and suppresses action potential firing by activating a $K_{2P}3.1$-like current. (Top) Halothane induces a pH-sensitive outward current in a locus coeruleus neuron held at -60 mV. (Bottom) Current clamp recording demonstrating halothane-induced membrane hyperpolarization and spike suppression that is reversed by acidification in the continued presence of halothane. (c) Concentration-dependent activation of single $K_{2P}2.1$ channels in patches excised from COS cells at 0 mV holding potential. Application of 0.2, 0.5, and 1 mM halothane is indicated by horizontal bars. (a) Reprodued Talley EM, Lei Q, Sirois JE, and Bayliss DA (2000), Task-1, a two pore domain K^+ channel, is modulated by multiple neurotransmitters in motoneurons. *Neuron* 25: 399–410, copyright 2000, with permission from Elsevier. (b) Reprinted from Sirois JE, Lei Q, Talley EM, Lynch C, and Bayliss DA (2000) The TASK-1 two-pore domain K^+ channel is a molecular substrate for neuronal effects of inhalational anesthetics. *Journal of Neuroscience* 20: 6347–6354, with permission. (c) Reprinted by permission from Macmillan Publishers Ltd: Patel AJ, Honore E, Lesage F, et al. (1999) Inhalational anesthetics activate two-pore-domain background K^+ channels. *Nature Neuroscience* 2: 422–426, copyright 1999.

propagation leading to conduction block in peripheral nerves. Whereas volatile anesthetics open K_{2P} channels producing neuronal hyperpolarization, local anesthetics inhibit K_{2P} currents. Diminished potassium flux leads to cellular depolarization and slowed repolarization, producing action potential prolongation and conduction blockade, contributing to peripheral analgesia. The same actions may account for some of the undesirable side effects of local anesthetics, such as central nervous system excitotoxicity and cardiac arrhythmias. Local anesthetics, including lidocaine, bupivacaine, tetracaine, and mepivacaine, block most studied K_{2P} channels (with varying potency) with the notable exceptions of $K_{2P}6.1$ and $K_{2P}10.1$ (studies on $K_{2P}4.1$, $K_{2P}13.1$, and $K_{2P}15.1$ have not been reported).

Cytoskeletal Remodeling and Synaptogenesis

Suggesting a role for $K_{2P}2.1$ in axonal migration and synaptogenesis during neuronal development, $K_{2P}2.1$ (and, to a lesser extent, $K_{2P}10.1$ and $K_{2P}4.1$) expression has been shown to induce formation of filopodia-like structures in dendrites and axons of cultured rat neonatal hippocampal neurons. These protrusions are rich in actin and ezrin and show increased $K_{2P}2.1$ channel density. It is known that this does not depend on channel activity because reducing K^+ flux by elevation of extracellular K^+ or by $K_{2P}2.1$ inhibition with oleylamin did not affect protrusion formation; however, intracellular channel protonation and PKA phosphorylation are required.

Apoptosis

Excitability is critical to neuron survival. Through effects on resting membrane potential, activity of potassium channels has a significant impact on neuronal cell death and survival; this is of particular importance during central nervous system development. Some potassium currents induce apoptotic cell

death, whereas others protect neurons from ischemia and excitotoxicity. Neuronal apoptosis has been linked to K_{2P} channel activity in cerebellar granule neurons (CGNs). CGNs express the standing outward potassium current, I_{KSO}, which shares biophysical, pharmacological, and regulatory characteristics with cloned $K_{2P}1.1$, $K_{2P}3.1$, and $K_{2P}9.1$ channels (and perhaps $K_{2P}10.1$). CGNs undergo apoptosis when exposed to physiological K^+ ion conditions *in vitro*. The apoptosis rate is significantly reduced with depolarizing conditions (high bath K^+ ion), when $K_{2P}1.1$, $K_{2P}3.1$, and $K_{2P}9.1$ channels are blocked by extracellular acidosis or specific inhibitors, or when dominant-negative $K_{2P}3.1$ or $K_{2P}9.1$ mutants are transfected into the cells (Figure 8(a)). Similarly, K^+ ion-induced cell death can be produced in hippocampal neurons that lack I_{KSO} by expression of $K_{2P}3.1$ or $K_{2P}9.1$ channels (Figure 8(b)) or a constitutively active $K_{2P}2.1$ mutant, whereas wild-type $K_{2P}2.1$, displaying little baseline conductance, did not alter cell viability.

Infectious Disease

Human immunodeficiency virus-1 (HIV-1), the causative agent of AIDS, has evolved to invade the complex human immune system. Vpu is a 68-residue HIV-1 membrane protein that enhances the release of progeny virions from infected cells and induces degradation of the HIV receptor on the T lymphocyte CD4. Interestingly, Vpu displays significant homology to the first transmembrane domain of $K_{2P}3.1$, suggesting the viral protein may have evolved through molecular piracy (i.e., evolving from a host gene) in order to foster interaction with host proteins. Vpu and $K_{2P}3.1$ physically interact in experimental cells and lymphoid tissues obtained from AIDS patients. The functional consequence of interaction is bidirectional destruction: in primary $CD4^+$ T lymphocytes infected with HIV, $K_{2P}3.1$ channels are degraded, whereas overexpression of the channel results in the release of fewer viral particles. Due to the ubiquitous expression of $K_{2P}3.1$ in human tissues, this virus–host interaction may influence progression of infection and electrical signaling in host tissues. Indeed, central nervous system infection with HIV-1 can lead to behavioral abnormalities, motor dysfunction, or dementia, and K_{2P} channels may play a role in these pathogenic events.

K_{2P} channels may be linked to prion–protein disorders. Prion diseases (transmissible spongiform encephalopathies) in humans and animals are caused by conformational conversion of a nonpathogenic cellular glycoprotein of largely unknown function (PrP^C) into PrP^{Sc}, an isoform that appears to be infectious in the absence of nucleic acids. PrP^C displays high expression in the peripheral and central nervous systems. $K_{2P}2.1$ has been shown to physically interact with PrP^C protein in a yeast two-hybrid study and experimental cells. Furthermore, $K_{2P}2.1$ and PrP^C colocalize when present in rat cerebellar Purkinje cells.

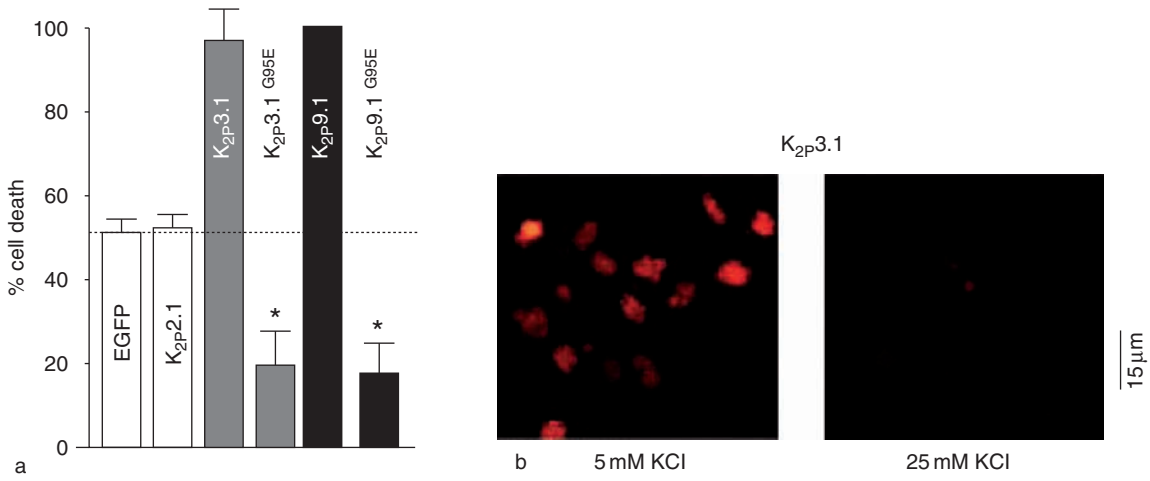

Figure 8 K_{2P} channels and their potential significance in apoptosis and cancer. (a) $K_{2P}3.1$ and 9.1 channels promote apoptosis in cultured rat cerebellar granule neurons (CGN) expressing the standing outward potassium current, I_{KSO}. Analysis of cell survival under proapoptotic low K^+ conditions was estimated by DAPI nuclear staining. Viral overexpression of $K_{2P}3.1$ or 9.1 in CGN enhances cell death, whereas genetic inactivation of the respective native currents via transfer of nonconducting, dominant-negative mutants protects CGN compared to nonproapoptotic K_{2P} channel ($K_{2P}2.1$) and fluorescence marker (EGFP) controls. (b) Genetic transfer of $K_{2P}3.1$ channels in rat hippocampal neurons lacking I_{KSO} and being insensitive to low K^+ confers potassium-dependent cell death. TUNEL analysis demonstrates a high apoptosis rate in $K_{2P}3.1$-transfected hippocampal neurons in low K^+ (5 mM) that was abolished by elevating the potassium concentration to 25 mM, as predicted. Reprinted from Lauritzen I, Zanzouri M, Honore E, et al. (2003) K^+-dependent cerebellar granule neuron apoptosis. Role of TASK leak K^+ channels. *Journal of Biological Chemistry* 278: 32068–32076, with permission.

The physiological relevance of interaction has not been established.

Evolving Fields of Inquiry

Discoveries suggest additional functional roles for K_{2P} channels. These are potential future fields of significance in physiology and disease.

Neuroprotection and regulation by free fatty acids K_{2P} channels may serve to protect neurons against ischemic and hypoxic insults. Lysophospholipids (LPs) and polyunsaturated fatty acids (PUFAs) activate $K_{2P}2.1$, $K_{2P}4.1$, and $K_{2P}10.1$; this is suspected to result in pre- and postsynaptic hyperpolarization and neuroprotection. In support of these ideas, mice lacking the $K_{2P}2.1$ channel have increased vulnerability to cerebral ischemia and do not show neuroprotective effects of LPs or PUFAs.

Temperature sensation A potassium leak conductance has been implicated in regulation of body temperature. $K_{2P}2.1$, $K_{2P}4.1$, and $K_{2P}10.1$ show changes in activity with temperature in the physiological range of 24–42 °C. K_{2P} temperature responses appear to occur via activity of temperature-sensitive secondary messenger pathways, such as protein kinases that act on the channels.

Tumorigenesis $K_{2P}9.1$ mRNA expression was found to be increased up to 100-fold in breast and lung cancer tissue samples. Moreover, injection of nude mice with tumor cells overexpressing $K_{2P}9.1$ led to tumor formation, whereas the same cells with a $K_{2P}9.1$ mutation that abolishes conduction pore function abrogated oncogenic potential. $K_{2P}9.1$ channels may thus possess paradoxical properties – that is, the capacity to promote cell proliferation, as well as apoptotic cell death in cell-type-dependent manner.

Cardiac channels The cardiac plateau current, I_{KP}, is time independent, potassium selective, and influences the amplitude and duration of the cardiac action potential. On the basis of genetic, biochemical, electrophysiological, and pharmacological evidence, it has been proposed that $K_{2P}2.1$ and $K_{2P}3.1$ channels contribute to I_{KP}. A role in mechanoelectrical feedback and stretch-associated arrhythmogenesis and cardioprotection has been suggested.

Conclusion

A decade since their discovery, the study of K_{2P} channels has revealed that background leakage of potassium ions via dedicated pathways is a highly regulated mechanism to control cellular excitability. K^+ ion leak pathways stabilize the membrane potential below firing threshold and expedite repolarization. K_{2P} channels are easily identified by their unique structure of two pore-forming domains in each subunit. Two-P-domain channels reveal an impressive diversity of mechanisms to control signaling of excitable cells dynamically, and they have been implicated in numerous cellular and body functions, including anesthesia, apoptosis, neuromodulation, and sensing of oxygen tension and mechanical stress. A general operational strategy has emerged: inhibition of K_{2P} channels promotes neuronal activity, whereas enhanced potassium leak stabilizes excitable cells at rest.

See also: Apoptosis in Nervous System Injury; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Inwardly Rectifying Potassium Channels; Ion Channel Localization in Cell Bodies and Dendrites; Large Conductance Calcium-Activated Potassium Channels; Learning and Memory in Invertebrates: *Aplysia*; Learning, Action, Inference and Neuromodulation; Potassium Channel Regulation; Respiration; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Bayliss DA, Sirois JE, and Talley EM (2003) The TASK family: Two-pore domain background K^+ channels. *Molecular Interventions* 4: 205–219.

Bockenhauer D, Zilberberg N, and Goldstein SAN (2001) KCNK2: Reversible conversion of a hippocampal potassium leak into a voltage-dependent channel. *Nature Neuroscience* 4: 486–491.

Goldstein SAN, Bayliss DA, Kim D, et al. (2005) International Union of Pharmacology: LV. Nomenclature and molecular relationships of two-P potassium channels. *Pharmacological Reviews* 57: 527–540.

Goldstein SAN, Bockenhauer D, O'Kelly I, and Zilberberg N (2001) Potassium leak channels and the KCNK family two-P-domain subunits. *Nature Reviews Neuroscience* 2: 175–184.

Goldstein SAN, Price LA, Rosenthal DN, and Pausch MH (1996) ORK1, a potassium-selective leak channel with two pore domains cloned from *Drosophila melanogaster* by expression in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America* 93: 13256–13261.

Honore E, Patel AJ, Chemin J, Suchyna T, and Sachs F (2006) Desensitization of mechano-gated K_{2P} channels. *Proceedings of the National Academy of Sciences of the United States of America* 103: 6859–6864.

Ketchum KA, Joiner WJ, Sellers AJ, Kaczmarek LK, and Goldstein SAN (1995) A new family of outwardly-rectifying potassium channel proteins with two pore domains in tandem. *Nature* 376: 690–695.

Lauritzen I, Zanzouri M, Honore E, et al. (2003) K^+-dependent cerebellar granule neuron apoptosis: Role of TASK leak K^+ channels. *Journal of Biological Chemistry* 278: 32068–32076.

Lopes CMB, Gallagher PG, Buck ME, Butler MH, and Goldstein SAN (2000) Proton block and voltage gating are potassium-dependent in the cardiac leak channel Konk3. *Journal of Biological Chemistry* 275: 16969–16978.

Medhurst AD, Rennie G, Chapman CG, et al. (2001) Distribution analysis of human two pore domain potassium channels in tissues of the central nervous system and periphery. *Molecular Brain Research* 86: 101–114.

Patel AJ, Honore E, Lesage F, et al. (1999) Inhalational anesthetics activate two-pore-domain background K^+ channels. *Nature Neuroscience* 2: 422–426.

Plant LD, Rajan S, and Goldstein SA (2005) K2P channels and their protein partners. *Current Opinion in Neurobiology* 15: 326–333.

Rajan S, Plant LD, Rabin ML, Butler MH, and Goldstein SAN (2005) Sumoylation silences the plasma membrane leak K^+ channel K2P1. *Cell* 121: 37–47.

Sirois JE, Lei Q, Talley EM, Lynch C, and Bayliss DA (2000) The TASK-1 two-pore domain K^+ channel is a molecular substrate for neuronal effects of inhalational anesthetics. *Journal of Neuroscience* 20: 6347–6354.

Talley EM, Lei Q, Sirois JE, and Bayliss DA (2000) TASK-1, a two-pore domain K^+ channel, is modulated by multiple neurotransmitters in motoneurons. *Neuron* 25: 399–410.

Two-Photon Imaging

C Ricard, J A Coles, R Serduc, and B van der Sanden, Institut National de la Santé et de la Recherche Médicale; and Université Joseph Fourier, Grenoble, France
P Verant and J-C Vial, Université Joseph Fourier and Centre National de la Recherche Scientifique, Grenoble, France

© 2009 Elsevier Ltd. All rights reserved.

Introduction

During the past two decades, two-photon microscopy has become a powerful tool for imaging biological structures with greater depth than confocal microscopy affords. Its high spatial and temporal resolution, down to a micrometer and less than a second, enables imaging of subcellular structures on cell cultures, on tissue slices, and also directly in a living animal. Its potential applications in the neuroscience field have been demonstrated, and intravital studies in the brain have been conducted allowing morphological and physiological observations with a spatial resolution never obtained with methods of neuroimaging such as magnetic resonance imaging or computed tomography.

This article discusses the physical principles behind two-photon microscopy and the methodology for studying structures or phenomena occurring in the brain of living animals. It also describes how the different elements of the nervous system, including neurons, glial cells, and cerebral microvasculature, can be imaged, as well as some applications of two-photon microscopy in neuro-oncology and in research on neurological diseases.

Methods

Instrumentation

Two-photon microscopy belongs to a larger family of techniques that includes nonlinear optics (NLO) and scanning microscopy. NLO phenomena such as multiphoton absorption (mainly two-photon absorption), multiharmonic generation (mainly second harmonic generation (SHG)), and coherent anti-Stokes Raman scattering are now commonly used. NLO phenomena occur at high light intensity during a short time, but biological tissues cannot support excessive heating. These contradictory requirements explain the choice of laser excitation in the pulsed regime, with high instantaneous intensity but low average power and in the near infrared, a spectral window where tissue absorption is minimal. This choice is judicious for other reasons as well: The light scattering by tissue inhomogeneity is minimized; the resulting light emission is in the visible spectral range where sensitive detectors are available; and finally, very stable and robust laser sources, in particular the titanium sapphire femtosecond lasers, are now commercially available at nonprohibitive prices.

Laser scanning technology is derived largely from the confocal microscope: The laser beam is usually scanned by a pair of oscillating mirrors and focused on the sample by an objective with a large optical aperture (Figure 1). The oscillations of the mirrors introduce an x,y scanning of the focused beam in the focal plane. The light resulting from the excitation is usually collected by the same objective and descanned before reaching confocal pinholes in front of detectors. In fact, multiphoton or multiharmonic scanning is simpler since the emitted light comes from only a small region surrounding the focal point and does not have to be descanned and conjugated with a pinhole in front of the detector. For confocal microscopy, the light scattering and the presence of the pinhole are at the origin of photon waste and limit the depth of observation to, for example, $\sim$50 μm in the mouse brain cortex. Near-infrared excitation and non-descanned light collection as close as possible to the microscope objective minimize the photon waste by scattering, and the depth of observation could be increased to 600 μm. An additional advantage of the non-descanned mode is the possibility of having simultaneous trans- and epicollection, as shown in Figure 1.

Like confocal microscopy, multiphoton or multiharmonic microscopy is a three-dimensional (3-D) imaging technique. The z dimension is given by the displacement of the focused laser beam in the z direction by means of a motor driving the microscope objective. 3-D images are made up of stacks of 2-D (x,y) images. The speed of image acquisition is mainly limited by the collected light intensity, itself limited by the laser excitation power, which must not cause excessive tissue damage. In general, one image (512×512 pixels) per second, or a stack of $\sim$60 images per minute, is obtained if the dye is adapted to two-photon microscopy.

NLO phenomena are inefficient for scattered laser photons, and only ballistic laser photons contribute to light emission. So the spatial resolution is the same for scattering and nonscattering tissues. The resolution depends first on the spatial dimensions of the luminescent (or SHG) spot where the laser is focused. The smallness of the spot is governed by the laser beam quality and by the aperture of the microscope objective. For a titanium sapphire laser beam, an objective aperture of 1, and an NLO second-order process (two-photon or SHG), the resolution is about $\sim$3 μm

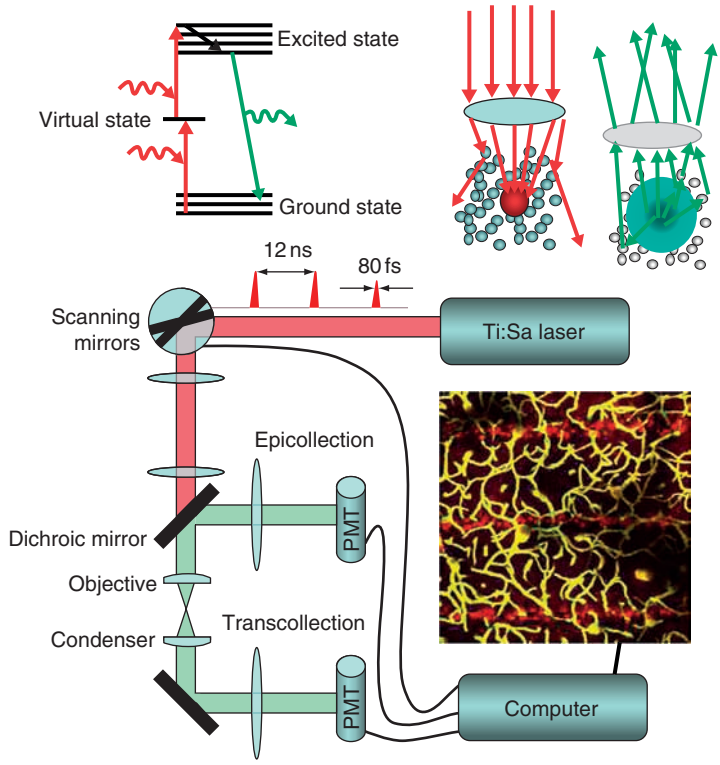

Figure 1 Basics of multiphoton and multiharmonic microscopy. Top left, energy levels and optical transitions involved in a two-photon process. In red, the simultaneous absorption of two photons, bringing the optical species from ground to excited level; in green, the photon emission from the relaxed excited state. Top right, ray tracing for two-photon excitation (red) and emission (green) in presence of scattering. Scattering reduces the number of focused photons, but the nonscattered (ballistic) photons are still well focused, and the spatial resolution is preserved. A large number of the scattered (and back-scattered) emitted photons can be collected. Bottom, schematic of a multiphoton or multiharmonic microscope with epi and trans light collection and an example of an image of the cortex vasculature. PMT, photomultiplier tube; Ti:Sa, titanium sapphire.

in the z direction and 0.5 μm in the x,y plane. Other causes, such as animal movements (bulk movements or local movements), are at the origin of a lack of spatial resolution. Animal holders are specially designed to limit bulk displacement, and good anesthesia methods should be used to limit excessive local movements.

Two-photon microscopy requires endogenous or exogenous emitting molecules. Unfortunately, while endoluminescence can be used for many tissues, it is rarely sufficient for imaging brain tissues; thus staining is necessary. For intravital brain microscopy, three strategies can be used: local staining by dye injection, staining via the blood circulation after an intravenous injection, and transfection of genes coding for fluorescent proteins. For instance, calcium-sensitive dyes are infused locally to follow neuron activity, and nonpermeant plasmatic dyes are used to image the vascular network. Some dyes can cross the blood–brain barrier (BBB) and stain the adjacent tissues. Additional physicochemical properties have to be considered in the choice of a dye, such as the luminescent and the nonlinear absorption cross section efficiencies, the photochemical stability, and the nontoxicity. The large number of requirements explains why the development of new dyes remains an important area for fundamental research.

Nonlinear optical neuroimaging is used not only for morphological information; quantitative information about brain physiology, such as the local blood volume or the BBB's permeability to specific dyes, can also be extracted. For these purposes the emitted light intensity is treated as a parameter representative of the local volume or the local dye concentration. Image stacks then constitute a 3-D representation of these parameters.

Animal Preparation

Rodents are the main animal models used in neurological two-photon imaging studies. Mice are the most often encountered because of the availability of a wide range of transgenic animals expressing fluorescent proteins. In general, studies on rats are conducted on young animals only (up to 1 month old) whereas on mice, studies can be accomplished during adulthood.

To minimize the movements during image acquisition, the choice of anesthesia is fundamental. A mixture

of xylazine and ketamine is a good candidate, but volatile anesthetics, such as isoflurane (2%) in a mixture of 30% O_2 and 70% N_2O, are generally better. Immobilization of the skull is necessary. During the whole experiment, the temperature of the animal must be monitored and maintained at 37 °C by means of a heating blanket. The next step is to obtain good optical access to the cortex. This surgical procedure can be done in three different ways:

- The bone of the skull is removed, and the objective of the microscope is immersed in physiological saline at the surface of the brain (Figure 2(a)).
- The bone of the skull is replaced by a quartz window, allowing longitudinal studies on the same animal (Figure 2(b)).
- The bone of the skull is just thinned, and observations are made right through it. This method limits the observation depth to the first 200 μm of the cortex due to the high scattering of the photons by the skull (Figures 2(c) and 3).

In adult rats, the dura is removed since its opacity is higher than that of mice.

Imaging the Various Components of the Nervous System

Neurons

The prime function of neurons is to transmit electrical signals in a network that undergoes constant

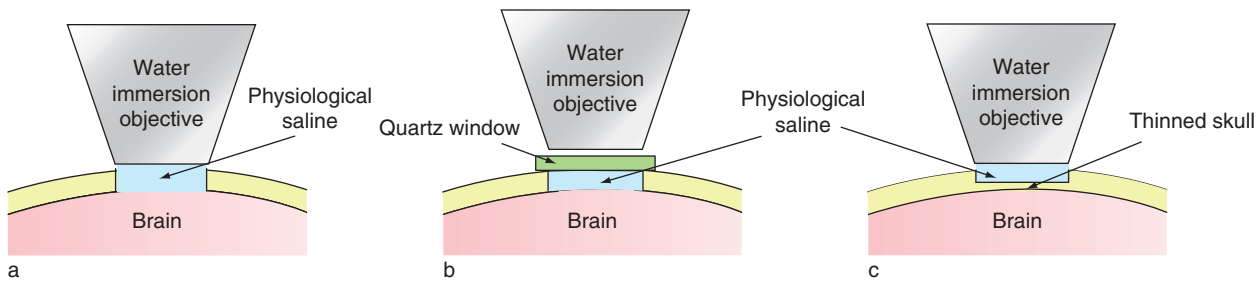

Figure 2 Schematization of the three surgical approaches for the imaging of the brain cortex *in vivo* with two-photon microscopy. (a) A craniotomy with a diameter in the range of 3–4 mm is carried out above the cortex. The hole is filled with physiological saline and the objective is positioned in immersion. (b) A craniotomy with a diameter in the range of 3–4 mm is carried out above the cortex, the hole is filled with artificial cerebrospinal fluid, and a quartz window is fixed above the craniotomy. This method allows longitudinal studies. (c) The skull is thinned above the cortex and the objective is positioned in immersion on the area of interest. This method is less invasive, but the imaging depth is more limited.

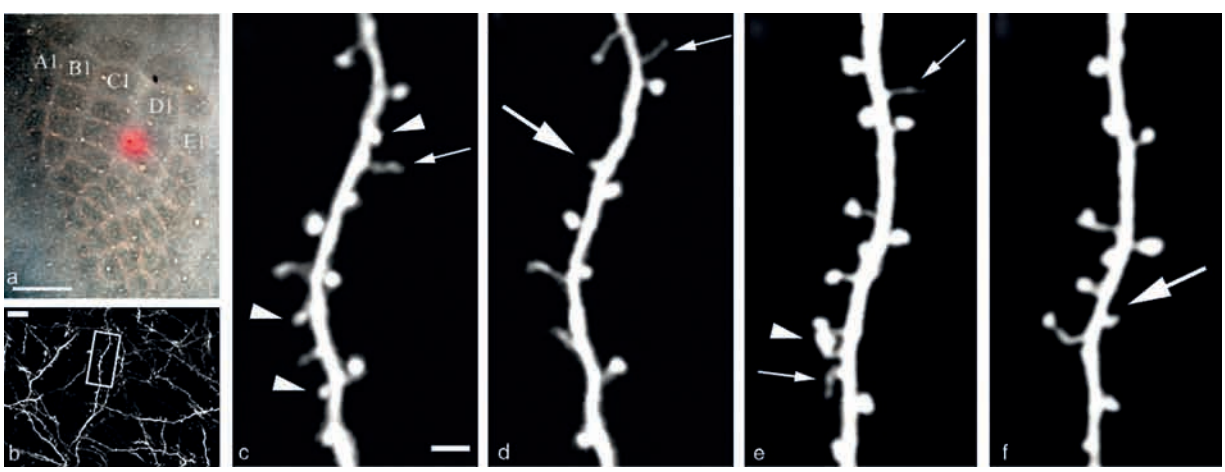

Figure 3 Two-photon microscopy allows repeated imaging of structural details *in vivo* over periods of weeks. The images are from the barrel cortex of a transgenic mouse expressing yellow fluorescent protein in pyramidal neurons of Layer V. (a) Fixed section of barrel cortex, made after intravital two-photon imaging, showing the outlines of the barrels, each of which receives input mainly from one whisker, and a red patch indicating the area imaged *in vivo*. (b) Low magnification intravital image of apical dendrites of pyramidal neurons. (c–f) Repeated imaging, over a period of 2 weeks, of the dendrite boxed in (b). Some dendritic spines were eliminated (arrowheads), and some were formed (large arrows). Filipodia also appeared and disappeared (small arrows). The images were obtained through an area of skull thinned to30–50 μm at a depth of about 150 μm below the pia. Scale bar = 500 μm (a), 10 μm (b), 2 μm (c–f). Reprinted from Macmillan Publishers Ltd: Zuo Y, Yang G, Kwon E, and Gan W-B Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* 436: 261–265. Copyright 2005.

modification as synapses are formed, strengthened, or eliminated. Compared with the traditional techniques of *in vivo* recording of electrical potentials with micropipettes and analysis of ultrastructure on fixed sections of tissue, two-photon microscopy allows very great improvement in the information that can be obtained. What can be achieved depends on the fluorophores that are introduced in the neurons. The most notable imaging to date has been in the upper layers of the mouse cerebral cortex *in vivo*, and on brain slices.

Calcium imaging Electrical activity in neurons is almost always associated with changes in free Ca^{2+} in the cytoplasm. Changes in free Ca^{2+} are readily detected optically because free Ca^{2+} can change by up to a thousandfold and an extensive range of fluorescent Ca^{2+} indicators is available. Ca^{2+} indicators, such as Calcium Green 1 or Fluo 4, can be injected into a single neuron through a fine-tipped micropipette or through a patch pipette. Since only one neuron is labeled, useful information can be obtained with only moderate spatial resolution. To increase the depth of penetration in the tissue, it may be useful to reduce the convergence of the excitation beam by illuminating only the center of the exit pupil of the objective. Calcium signals are generally slow compared with electrical ones (hundreds of milliseconds), so low time resolution allows more photons to be collected from each point in the neuron. Since the injection technique requires an intracellular microelectrode or a patch pipette, the electrical activity in the cell soma can be recorded simultaneously with the Ca^{2+} activity in dendrites, or the cell resting potential can be displaced by injecting current.

Uptake of dye by neurons Many Ca^{2+} indicators are available as esters which cross cell membranes and are de-esterified by endogenous enzymes. When applied to nerve tissue (for example to the surface of the cortex or in the bath of a tissue slice), Ca^{2+} indicators are usually taken up by astrocytes as well as neurons, although some appear to be taken up mainly by neurons (e.g., Oregon Green Bapta-AM). If the astrocytes are simultaneously labeled with sulforhodamine 101, it is possible to distinguish the two types of cell and select recordings from neurons. This simple technique of applying Ca^{2+} indicator to tissue enables large populations of neurons to be labeled so that waves of activity can be observed.

Voltage-sensitive dyes Dye molecules are available with lipophilic moieties that insert in the cell membrane and whose fluorescence varies with membrane potential. So far, the poor signal-to-noise ratio has been a constraint on the use of these dyes in the thick tissues for which two-photon excitation is useful; however, the same intense infrared illumination that produces two-photon excitation of voltage-sensitive dyes also generates a second harmonic (e.g., with excitation at 900 nm, a fraction of the emitted light is at 450 nm). This SHG is more sensitive to changes in membrane potential than is the fluorescence and has been used to detect electrical signals in segments of a single dendrite in a brain slice.

Labeling by expression of fluorescent proteins Mutant mice in which a subpopulation of neurons expresses green fluorescent protein (GFP) have made it possible to image individual features of a neuron through a cranial window or thinned skull over periods of months. Expression of fluorescent label has also been achieved in monkey by injection in the cortex of an adenovirus bearing the gene for enhanced green fluorescent protein. These techniques are extremely productive in studying the development of brain circuits and their plasticity associated with modifications of axon geometry and formation and elimination of dendritic spines (**Figure 3**).

Glial Cells

Glial cells and neurons are the two major cell types encountered in the nervous system. The three main classes of glial cells are astrocytes, oligodendrocytes, and microglial cells. Astrocytes were first described as metabolic supporting cells for neurons as well as having an intimate association with the BBB. However, their potential implication in communication by means of Ca^{2+} signaling was discovered and documented only during the past two decades.

The first studies were conducted on cell cultures and on living slices. Despite the high quality of these preparations, *in vivo* investigations are the only ones that can minimize the preparation artifacts encountered on slices. Further, intravital studies allow the observation of the interactions between neurons and astrocytes with the best physiological state. That is why two-photon microscopy is applied in numerous studies on astroglia. Their staining was reported to be done with specific dyes, such as the sulforhodamine 101 injected in the neocortex. Transgenic animals using fluorescent proteins such as the GFP under the astrocyte-specific glial fibrillary acidic protein promoter are also excellent models for visualizing these cells directly *in vivo* or on living brain slices.

Studies on the Ca^{2+} signaling encountered in astrocytes have been conducted using fluo-4 AM, an indicator whose fluorescence emission is dependent on the intracellular concentration of Ca^{2+}. Such experiments allow the observation of the propagation of

Ca^{2+} waves on brain slices and, more interesting, *in vivo*, where the neuronal/glial interactions can be highlighted. For example, an increased neuronal discharge was associated with increased Ca^{2+} concentration in neighboring astrocytes. Last, the double staining of astrocytes and vasculature using a specific dye and an intravascular fluorescent marker is a novel approach to study the glio-vascular interface *in vivo*.

Other glial cells, such as oligodendrocytes, can be stained using transgenic animals with a fluorescent protein under the control of the proteolipid protein promoter, for example. However, intravital imaging of the white matter remains difficult in rodents due to its depth. Such observations must be considered on living brain slices.

Finally, microglial cell motility has been extensively studied using time-lapse two-photon microscopy. Morphological observations, such as rapid movements of microglial processes and cell migration after a focal injury, were made with a high spatial and temporal resolution. In such experiments, the same laser used for the excitation of fluorescent proteins can be used to generate the injury. Thanks to the two-photon properties, the high power excitation is restricted to the focal point, and as a consequence, the injury is limited on a small volume and induces less damage than a mechanical or chemical one would.

Cerebral Microvasculature

The central nervous system, composed of the brain and the spinal cord, controls the majority of our organs. Cerebral activity is guaranteed by a neuronal network which requires a large and well-regulated supply of energetic substrates. Indeed, the human brain represents about 2% of the total body mass, but it consumes 20% of the respired oxygen. These nutrients, mainly oxygen and glucose, are transported to brain by the blood through a dense and complex vascular system, the cerebral microvasculature. The low diffusivity of the substances transported by the blood makes essential the presence of a large network of capillaries (4–8 μm in diameter) within the whole tissue volume. It is at this scale that exchanges between blood and brain take place.

The cerebral blood perfusion (CBP) is defined as the volume (in milliliters) of blood which irrigates 100 g of brain tissue during 1 min. This parameter is of crucial importance. It has been proved that a strong correlation exists between CBP and cerebral (neuronal) activity. Changes of up to 50% in CBP have been observed in the gray matter in regions activated during a motor task or perception (e.g., taking an object, speaking, visual exercise). In case of brain pathologies, differences in CBP have been described. For instance, some hypoperfused regions were detected in epileptic focal regions (intercritical phase) or in posterior regions of the brain in Alzheimer's disease. Hyper- and hypoperfused zones can be found in high-grade glial tumors. Changes in CBP are often related to alterations in cerebral blood volume. Both parameters can be estimated on clinical imaging systems, such as positron emission tomography or magnetic resonance imaging. However, none of these methods have sufficient resolution to describe the events at a capillary level. Recently, intravital two-photon microscopy has been proposed to investigate local variations of cerebral blood volume in preclinical protocols with very high temporal and spatial resolutions. Indeed, a direct analysis of a 0.16 mm^3 sample can be done in 2 min with submicrometer resolution.

Two-photon microscopy allows observation of the brain microvasculature after the intravenous injection of fluorescent dyes such as fluorescein or rhodamine dextran conjugates (Figure 4). Depending on the age and the animal species, it is possible to image the microvasculature down to 600 μm below the dura with micrometric resolution. Furthermore, 3-D images of the cerebral cortex vasculature can be reconstructed using z-stacks. Using this method, studies on the permeability of the BBB can be conducted *in vivo*. It can be assessed after the intravenous injection of a cocktail of two fluorescent dyes, one with a high molecular weight (e.g., rhodamine dextran 70 kDa) and the second one with a low molecular weight (e.g., FITC dextran 4 kDa). Simultaneous imaging of the distribution of these two dyes in the vascular compartment and in the brain parenchyma is then obtained with two-photon microscopy (Figure 5). The wide availability of fluorescent dyes conjugated with dextran molecules of different molecular weight enables a precise evaluation of the BBB permeability. This method is of particular interest in preclinical studies of new radiotherapeutic or chemotherapeutic approaches. Finally, to study neuronal–glial–vascular interactions, it is possible to use triple staining in a single two-photon microscopy experiment.

Applications to Pathologies

In Vivo Two-Photon and SHG Microscopy for Imaging the Brain Tumor Microenvironment and Therapy Effects

Primary brain tumors, in particular glioblastomas, are among the most therapy-resistant tumors. Until now, no therapy has yet been found to arrest their fast growth. Therefore the development of new therapies on tumor animal models is primordial. In this preclinical research, detailed knowledge of, for example, the different transport barriers for drugs in brain tumors will help since a drug that is injected in the blood

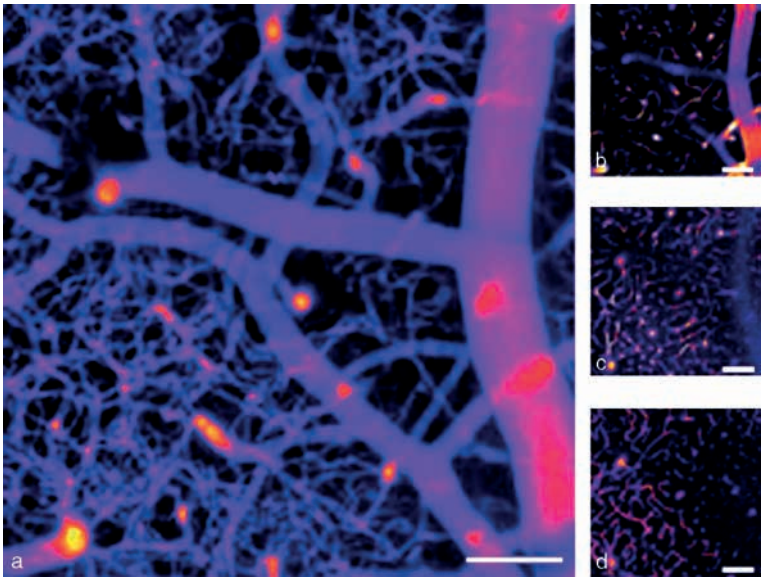

Figure 4 Mouse cerebral microvasculature. The image (a) represents the whole vascular network of the mouse cortex ranged between 0 and 600 μm under the dura. The three smaller vignettes show individual images at different depths in the brain. (b) At the surface (65 μm), large arteries can be detected. They form cortical columns which plunge into the brain and divide into capillaries ((c) and (d), 225 and 475 μm under the dura, respectively). Images were acquired *in vivo* using two-photon microscopy after intravenous injection of fluorescent dye (FITC-dextran 70 kDa). Scale bar = 100 μm.

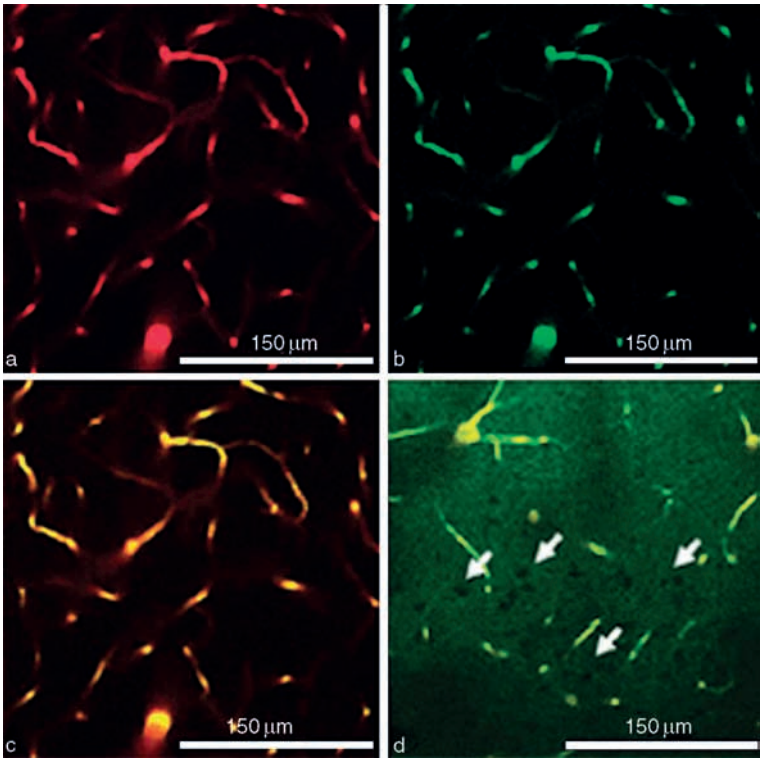

Figure 5 Visualization of the permeability of the blood–brain barrier (BBB). Simultaneous acquisition of rhodamine dextran 70 kDa (a), FITC dextran 4 kDa (b) and merge of the two pictures (c) at 170 μm below the dura in a healthy mouse. (d) Merge of simultaneous acquisition of rhodamine dextran 70 kDa (red) and FITC dextran 4 kDa (green) at 170 μm below the dura in a mechanically BBB-induced disruption model. Note the 'black holes' (arrows), which are cells not labeled by the extravasated dye. In that case, the FITC dextran 4 kDa crossed the BBB whereas rhodamine dextran 70 kDa did not.

circulation encounters many obstacles before arriving in the tumor cell. These obstacles in the heterogeneous tumor microenvironment include, among others, the tumor vasculature, the more or less leaky vessel walls, the extracellular matrix (ECM), and the cell membrane. *In vivo* two-photon microscopy and SHG imaging allow deep imaging of transport barriers in tumors growing in living animals and may also be useful in monitoring therapy effects.

Imaging the tumor microenvironment In two-photon microscopy, the tumor microvasculature can be observed directly *in vivo* after intravenous injection of fluorescent dyes that stay in the intravascular space. In Figure 6, the neovasculature of the tumor is sprouting from the normal host vasculature (see arrow) at the surface of the tumor and shows abnormalities such as dilatations and numerous bifurcations. These changes have an impact on the blood pressure and perfusion, which often decrease in tumor tissue. If a drug is injected in the blood circulation, then the supply to the tumor may be reduced and become intermittent or even absent in chronic hypoxic areas. These areas often contain viable but radiotherapy- or chemotherapy-resistant tumor cells and are of interest for possible tumor regrowth after treatment. The local blood perfusion rate can be estimated by measuring with two-photon microscopy the linear velocity (in millimeters per second) of single red blood cells in the microvessels. Furthermore, the local vessel permeability determines the transport of drugs across the vessel wall and can be measured in two-photon microscopy using diffusible fluorescent dyes or drugs. But this is not the end of possible transport barriers. Collagen fibers in the ECM, which can be detected by SHG imaging, might influence the intratumoral diffusion. Indeed, the enzymatic breakdown of these fibers has been shown to increase drug diffusion rates in the ECM.

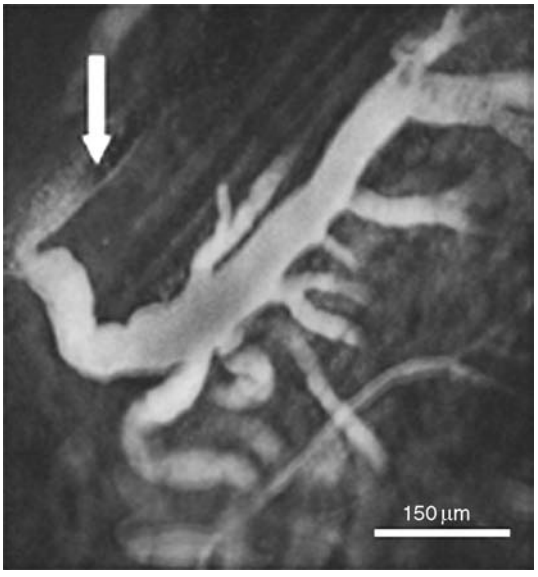

Figure 6 Three-dimensional (3-D) image of the tumor vasculature at the surface of a glioma growing subcutaneously in the hind limb of a living nude mouse. The vasculature is stained after intravenous injection of FITC-dextran (70 kDa). 20× objective, 600 × 600 μm image in 0.9 s. The 3-D image is a *z*-projection of *z*-stack with a depth of 100 μm and a step size of 2 μm. New tumor vessels originating from the muscle vasculature (white arrow) and forming a heterogeneous structure with large bifurcations and dilations are easily seen.

Monitoring therapy In a successful treatment, a marked reduction of the tumor volume or even total ablation of the tumor is observed. But this appears often at weeks or months after the onset of a treatment. Pathophysiological and cellular processes such as the induction of tumor cell death or necrosis and a possible shutdown of the vasculature that precedes volume reduction or tumor ablation can be observed by two-photon microscopy. These two-photon microscopy analyses will help in the fine-tuning and/or modification of a therapy in the early phase if it fails to induce tumor cell death. For these longitudinal *in vivo* two-photon microscopy experiments, scientists need a permanent window in an animal for direct microscopic observations. For grafted brain tumors, two window models exist, namely a dorsal skin fold chamber and a cranial window (see Figure 2(b)). In the first model, a glioma grows on the back of an animal inside a skin fold, which is held upright by two metal plates that include a window. The cranial window approach involves the positioning of a window at the top of the brain cortex after craniotomy of a small area just above the glioma. The latter model is preferred over skin fold chambers because tumor development and physiology strongly depend on the host tissue bed and vasculature.

In summary, two-photon microscopy and SHG imaging are very promising for the *in vivo* microscopic evaluation of the tumor microenvironment and may allow the prediction and monitoring of therapy outcome.

Applications to Nervous System Diseases

The understanding of neurodegenerative diseases such as Alzheimer's disease is an important goal for fundamental and preclinical research. The study of their pathophysiology can be done *in vitro*, but to analyze their mechanisms as a whole, *in vivo* observations are necessary. Two-photon microscopy has the potential to allow such observations. For example, a fluorescent molecule was recently tested which has the ability to cross the BBB and bind specifically to amyloid plaques after an intravenous injection. The intravital staining of these structures, as well as the

validation of the specificity of these molecules, is of particular interest in the development of targeted drugs.

In Alzheimer's disease, there seems to be a correlation between cognitive dysfunctions and a loss of synapses. It has also been shown, using time-lapse two-photon microscopy, that dendrites exhibit spine loss and atrophy in the vicinity of amyloid deposits. Astrocytes and activated microglia that are often found around amyloid plaques have been simultaneously labeled using transgenic constructions to give a better overview of the neuropathology *in vivo*.

Two-photon microscopy is also of particular interest in the study of cerebrovascular diseases such as cerebral ischemia. Short- and long-term effects of stroke can be assessed with time-lapse imaging using cranial window models. Morphological changes of cells and vasculature can be observed in the penumbra (the area surrounding the core lesion, where the blood flow is reduced and where delayed cell death can occur), leading to a better understanding of the injury and the mechanisms of partial recovery. One of the most outstanding two-photon imaging studies of vascular occlusion was photothrombosis of single vessels in the living mouse cortex. Hemodynamics parameters, as well as flow modifications in the neighboring vessels, were studied with great accuracy.

New insights into the pathophysiology of epilepsy have been obtained from the cortex of living animals. It has been shown that antiepileptic drugs such as valproate and gabapentin reduce the ability of astrocytes to transmit Ca^{2+} signals. The potential key role of astrocytes in seizure activity was hypothesized in this study, as well as their possible use in therapeutic targeting.

Two-photon imaging is used in the study of many diseases of the nervous system. The last three examples explain the major breakthroughs in the understanding of neuropathologies. The study of a neuropathology by means of two-photon imaging requires (1) finding the best animal model for the disease, (2) choosing the best way to label the pertinent biological elements (e.g., stain, transgenic animal expressing a fluorescent protein), and (3) determining how the data will be processed, after image acquisition, for quantifying the phenomena.

Conclusions and Perspectives

Two-photon microscopy appears poised to take its place as a key method for neuroscience. It has applications in cell cultures, *ex vivo* tissue slices, and, more interesting, directly in the brains of living animals. Observations can be made to a depth of at least 600 μm below the dura, depending on the species, age, and preparation of the animal model. In mice, nearly all the neocortex is accessible, including layer V. However, deeper structures, such as basal ganglia, thalamus, and hippocampus, cannot be observed with intravital two-photon microscopy.

Research is under way to improve the imaging depth by two different approaches:

- The application of optic fibers that can be directly inserted in cerebral structures using stereotaxic coordinates. First trials have been done in rats, allowing the imaging of the microvasculature in the hippocampus.
- The development of new fluorescent molecules with a higher quantum yield.

In conclusion, two-photon microscopy has opened a new era in the morphological and (patho)physiological studies of the nervous system. Its multidisciplinary approach required strong interactions among biologists, physicists, and chemists to provide a unique imaging tool for a better integrative approach to neuroscience.

See also: Cognition: An Overview of Neuroimaging Techniques; Fluorescence Microscopy in the Neurosciences; Glial Cells: Microglia During Normal Brain Aging; Glial Cells: Astrocytes and Oligodendrocytes During Normal Brain Aging; Neuroimaging; Single Photon Emission Computed Tomography (SPECT): Technique; Tumors of the Brain and the Spinal Cord.

Further Reading

Bacskai BJ, Hickey GA, Skoch J, et al. (2003) Four-dimensional multiphoton imaging of brain entry, amyloid binding, and clearance of an amyloid-β ligand in transgenic mice. *Proceedings of the National Academy of Sciences of the United States of America* 100(21): 12462–12467.

Brown EB, Campbell RB, Tsuzuki Y, et al. (2001) *In vivo* measurement of gene expression, angiogenesis and physiological function in tumors using multiphoton laser scanning microscopy. *Nature Medicine* 7(7): 864–868.

Göppert M (1929) Über die Wahrscheinlichkeit des Zusammenwirkens zweier Lichtquanten in einem Elementarakt. *Naturwissenschaften* 17: 932.

Helmchen F and Denk W (2005) Deep tissue two-photon microscopy. *Nature Methods* 2: 932–940.

Hirase H, Qian L, Barthó P, and Buzsáki G (2004) Calcium dynamics of cortical astrocytic networks *in vivo*. *PLoS Biology* 2(4): 494–499.

König K (2000) Multiphoton microscopy in life sciences. *Journal of Microscopy* 200: 83–104.

Nimmerjahn A, Kirchhoff F, Kerr JN, and Helmchen F (2004) Sulforhodamine 101 as a specific marker of astroglia in the neocortex in vivo. *Nature Methods* 1: 31–37.

Schaffer CB, Friedman B, Nishimura N, et al. (2006) Two-photon imaging of cortical surface microvessels reveals a robust redistribution in blood flow after vascular occlusion. *PLoS Biology* 4(2): e22.

Svoboda K and Yasuda R (2006) Principles of two-photon excitation microscopy and its applications to neuroscience. *Neuron* 50: 823–839.

Tian G-F, Takano T, Lin JH, Wang X, Bekar L, and Nedergaard M (2006) Imaging of cortical astrocytes using 2-photon laser scanning microscopy in the intact mouse brain. *Advanced Drug Delivery Reviews* 58: 773–787.

Vérant P, Serduc R, van der Sanden B, Rémy C, and Vial J-C (2006) A direct method for measuring mouse capillary cortical blood volume using multiphoton laser scanning microscopy. *Journal of Cerebral Blood Flow and Metabolism* 27(5): 1072–1081.

Zuo Y, Yang G, Kwon E, and Gan W-B Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* 436: 261–265.

Ubiquitin-Proteasome System and Plasticity

A N Hegde, Wake Forest University Health Sciences, Winston-Salem, NC, USA

© 2009 Elsevier Ltd. All rights reserved.

The quest for understanding the mechanisms of synaptic plasticity that underlie learning and memory has revealed a role for posttranslational modifications such as phosphorylation and gene expression in changing synaptic strength. Perhaps most surprising of all was the discovery in recent years of a role in synaptic plasticity for proteolysis by the ubiquitin–proteasome pathway. In this pathway, attachment of ubiquitin, a small protein, serves as a tag for degradation by a proteolytic complex called the proteasome.

Ubiquitin was discovered more than two decades ago. Originally it was thought that ubiquitin marks only abnormal proteins for degradation. Studies on several nonneuronal cells showed that ubiquitin functions to regulate protein degradation under normal physiological conditions. Although ubiquitin was used as a marker for degenerating neurons by neuropathologists, ubiquitin was virtually unknown to broader neuroscience community until recently.

A physiological role for the ubiquitin–proteasome pathway in the nervous system was discovered during investigation on long-term facilitation in the sea slug *Aplysia*. Long-term facilitation underlies a simple form of memory in *Aplysia*, behavioral sensitization of defensive reflexes. Since then, other studies showed a role for this pathway in mammalian long-term memory and formation of synapses at *Drosophila* neuromuscular junction.

The Ubiquitin–Proteasome Pathway

The main function of the ubiquitin–proteasome pathway is to degrade proteins. Proteolysis is highly regulated and specific: only proteins that are marked by attachment of ubiquitin are degraded and only at the right time. The tagging of a protein substrate with ubiquitin is regulated by the action of three enzymes: E1, E2, and E3. First, ubiquitin is activated by E1. Then, the activated ubiquitin is transferred to one of a family of enzymes termed E2s, which associate with E3 ubiquitin ligases. The E3s ligate the ubiquitin to a Lys residue in the protein substrate. A second ubiquitin is covalently attached to the first ubiquitin and thus a polyubiquitin chain is formed. The polyubiquitinated substrate is recognized by the proteasome and is degraded to small peptides and amino acids. The proteasome that degrades polyubiquitinated proteins is called the 26S proteasome based on its sedimentation coefficient. The polyubiquitin chains are not degraded but disassembled by deubiquitinating enzymes (DUBs) (Figure 1).

Ubiquitin Conjugation

The series of enzymatic reactions in the process of attaching a ubiquitin molecule to a substrate is called ubiquitin conjugation or ubiquitination. First reaction in this series is activation of ubiquitin by the enzyme E1. The activation of ubiquitin is ATP dependent. ATP is hydrolyzed to generate AMP, and a high-energy thiol-ester ubiquitin-AMP intermediate is formed. The activated ubiquitin is then transferred to an E2, which is also called a ubiquitin carrier. The ubiquitin molecule is then transferred to an E3 which ligates ubiquitin to an ε amino group of a lysine residue in the substrate protein. A second ubiquitin molecule is then attached to an internal lysine reside in the first ubiquitin and thus a polyubiquitin chain grows.

Ubiquitin-Conjugating Enzymes: E1, E2, and E3

Among the three classes of ubiquitin-conjugating enzymes, E1 is the least physiologically regulated. E2s are more selective and are believed to interact with specific E3s. Substrate specificity of the ubiquitin conjugation reaction is largely determined by E3s.

Originally, E2 enzymes were believed only to carry the activated ubiquitin and transfer it onto the E3s. Recent studies, however, suggest that at least some E2s can directly conjugate ubiquitin to substrates. E2s are structurally and functionally diverse. Simple eukaryotes like yeast (*Saccharomyces cerevisiae*) have 13 genes potentially encoding E2s. The number of E2s in mammals is estimated to be in the range of 25–30. The diversity of E2 generates some degree of specificity in the ubiquitin-conjugating reaction. E2s bind to E3s selectively. Since the diversity of E3s is even greater than that of E2s, the combination of E2s and E3s potentially can generate a high degree of specificity.

E3s are the enzymes that specifically recognize the substrates. E3s can be single proteins or complex of proteins. Single-subunit E3s can accept ubiquitin in a thioester linkage from E2s and ligate ubiquitin to the substrate. When E3 contains multiple subunits, it is generally believed that the enzyme brings the E2 and the substrate together and facilitates the transfer of ubiquitin to substrate. There are two major classes of E3s: HECT (homologous to E6-AP C-terminus) domain E3s and RING (really interesting new gene) finger E3s.

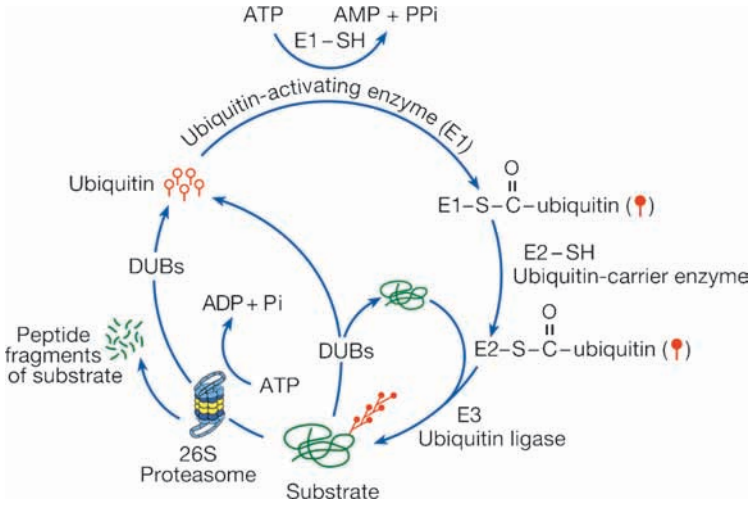

Figure 1 The ubiquitin–proteasome pathway. In this proteolytic pathway, ubiquitin (single ubiquitin molecule is represented by open red circles with straight tails) is selectively and covalently linked to the substrate. The enzymatic process of attaching ubiquitin to substrates is called ubiquitination or ubiquitin conjugation, and depends on the action of three different classes of enzymes: E1, E2, and E3. First, ubiquitin is activated by E1 to form a ubiquitin–AMP intermediate. Activated ubiquitin (closed red circles with straight tails) is passed on to an E2 (ubiquitin carrier enzyme). E2 transfers ubiquitin to an E3 (ubiquitin ligase), which ligates the activated ubiquitin to the substrate. To the ubiquitin attached to the substrate, another ubiquitin is attached, and thus through successive linkages of ubiquitin, a polyubiquitin chain forms. Polyubiquitinated substrates are degraded by a proteolytic complex called the 26S proteasome in an ATP-dependent reaction. Ubiquitin is not degraded but the polyubiquitin chain is disassembled and ubiquitin is recycled by deubiquitinating enzymes (DUBs). Before being committed to be degraded by the proteasome, ubiquitination is reversible. DUBs can disassemble the polyubiquitin chain if a substrate is ubiquitinated erroneously and prevent the degradation of the substrate.

HECT domain E3s The typical example in this class of E3 is the ubiquitin ligase called E6-AP, which ligates ubiquitin to the tumor suppressor protein p53. Viral protein E6 associates with a cellular protein called E6-AP (E6-associated protein). The C-terminal region of E6-AP contains the catalytic domain of the ubiquitin ligase. E6-AP ligase can function with either of the E2s called UbcH5 and UbcH7 (Figure 2). Later studies found that a family of proteins ubiquitin ligases with homology to the catalytic domain of E6-AP exists. These ubiquitin ligases are called HECT domain E3s. In addition to the HECT domain, there is another domain in many E3s called the WW domain. The WW domain-containing E3s also tend to have a C2 domain. The presence of the C2 domain is highly relevant to nervous system function because the C2 domain responds to elevation of intracellular Ca^{2+} and helps in translocation to the plasma membrane. Therefore, presence of this domain in neuronal HECT E3s might be critical in ligating ubiquitin to neurotransmitter receptors or proteins associated with them.

RING finger E3s These E3s are called RING finger E3s because they contain a RING finger domain, which consists of seven cysteine residues and one histidine residue forming a single folded domain binding two zinc ions. During the past few years, several ubiquitin ligases were found to contain the RING finger. RING finger motif in ubiquitin ligases is critical for transfer of ubiquitin to substrates or to RING finger proteins themselves. The RING finger category of E3s can be subdivided into single-subunit RING finger E3s and multi-subunit RING finger E3s (Figure 2).

Single-subunit RING finger E3s contains the RING finger domain and the substrate recognition site in the same protein. One of the well-characterized single-subunit RING finger E3s is Mdm2, which ubiquitinates a postsynaptic density protein PSD-95 in neurons.

Multi-subunit RING finger E3s contain two complexes:

1. *SKP1–cullin–F-Box protein* (*SCF*) *complex*. SCF complex contains at least four proteins: Skp1, Cul1, Roc1/Rbx1/Hrt1, and an F-Box protein. In the center of the SCF complex is the RING finger domain-containing protein Rbx1. The SCF-type ligases have another invariant protein called cullin. The theme appears to be that the cullins interact with linker proteins such as Skp1 to recruit substrate-interacting proteins such as the F-Box proteins. There are at least five different cullins in mammals. There are several F-Box proteins as well. A well-characterized substrate of the SCF complex is IκBα.

2. *Anaphase-promoting complex* (*APC*). Although APC has a subunit with a RING finger domain (APC11), this ubiquitin ligase is distinct from the SCF ligase in overall subunit combination. For example, instead of one adaptor found in SCF ligases (such

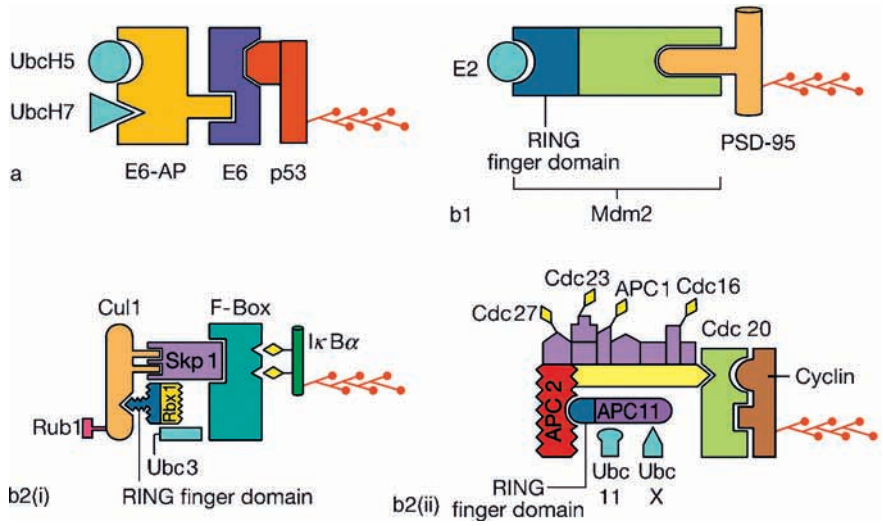

Figure 2 Major types of E3s. (a) HECT domain E3. E6-AP ubiquitin ligase in combination with E6 protein and one of the two E2s (UbcH5 or UbcH7) ligates ubiquitin to the p53 tumor suppressor protein. (b1) Single-subunit RING finger E3. Mdm2 ligates ubiquitin to PSD-95 with the help of an E2 enzyme. (b2(i)) Multi-subunit RING finger E3. SCF ligases contain the substrate recognition site on an F-Box protein. Skp1 is an adaptor that joins the F-Box protein to Cul1. Ring finger domain is on Rbx 1. The E2 is Ubc3. Cul1 is modified by Rub1, another ubiquitin-like protein leading to an increase in the activity of the ligase complex. The substrate is phosphorylated (diamonds) IκBα. (b2(ii)) Multi-subunit RING finger E3. APC is a more complex example of multi-subunit RING finger E3s and has a subunit composition distinct from that of SCF. Cdc20 protein in APC has the substrate (cyclin) recognition site. The RING finger domain is on APC11. The E2s Ubc11 or UbcX can function with the APC ligase. In addition, several adaptor proteins, some labeled (Cdc27, Cdc23, APC1, Cdc16) and some unlabeled, interact with Cdc20 and APC11. Diamonds on the adaptor subunits indicate phosphorylation.

as Skp1), APC has multiple subunits that serve as adaptors. Also, unlike SCF ligases, substrate phosphorylation is not an important determinant for specific substrate recognition by the APC ligase. Rather, substrate specificity of APC ligases appears to be modulated by incorporation of 'specificity factors' into the ligase complex. APC acts together with an E2 Ubc11 or Ubc X. One of the most studied substrates of APC is mitotic cyclin.

The Proteasome

The term proteasome is used to describe two kinds of multi-subunit proteolytic complexes, the 26S and 20S, based on their sedimentation coefficient. The 26S proteasome degrades ubiquitinated protein substrates. The 26S complex contains the 20S as a core and regulatory 'caps' on either end like a dumbbell. Each cap of the 26S proteasome is called the 19S regulatory complex (19S RC). The 20S core is a cylindrical structure consisting of the catalytic part of the proteasome.

The catalytic 20S core The 20S catalytic core of the proteasome has two types of subunits, namely, α and β. The subunits are arranged into four stacked rings to form the catalytic cylinder with the two middle rings consisting of β-subunits, which are sandwiched between two rings of α-subunits. In eukaryotes there are seven different α (α1–α7) and β (β1–β7) subunits each. The catalytic core of the proteasome is a threonine protease. The 20S proteasome can exist not only as a core of 26S, but also as a separate population that cannot degrade ubiquitinated proteins. However, the 20S proteasome by itself has chymotrypsin-like, trypsin-like, and postglutamyl peptidase activities which cleave after hydrophobic, basic, and acidic residues, respectively. The peptide-hydrolyzing activity of the 20S proteasome can itself be modulated by an 11S regulatory cap.

19S RC The 19S RC recognizes the polyubiquitinated substrate, and channels the substrate into the catalytic 20S core of the proteasome. It also has the capacity to regulate the activity of the catalytic core and determine the nature of the degradation process. Usually one 19S RC is attached to either end of the catalytic core. Some of the 19S RC subunits are believed to channel the substrate into the catalytic chamber for degradation.

Deubiquitinating Enzymes

Ubiquitination reaction is reversible before the ubiquitinated protein is committed to degradation by the proteasome. Ubiquitin is removed from substrates by enzymes called DUBs. Based on protein sequence and the molecular size, DUBs can be classified into two

general classes: (1) low molecular weight (20–30 K) ubiquitin C-terminal hydrolases (UCHs), and, (2) high molecular weight (around 100 K) ubiquitin-specific proteases (UBPs; also called USPs). Among the DUBs, UBPs belong to a large family containing diverse genes whereas the UCH family has fewer genes. For example, in yeast *S. cerevisiae*, there are 17 UBPs and 1 UCH. In the human genome, there are 63 genes encoding UBPs and 4 genes that code for UCHs. UCHs and UBPs subserve different functions in the eukaryotic cell. Although the current name for these enzymes, that is, DUBs, emphasizes the removal of ubiquitin from substrates, some DUBs, especially UCHs, function to process linearly linked ubiquitin precursors and generate monoubiquitin. DUBs are important for generating free ubiquitin at various steps of the ubiquitin–proteasome pathway. Ubiquitin is encoded by the tandemly linked polyubiquitin gene. In addition to the polyubiquitin gene, ubiquitin is also encoded by fusion to two ribosomal subunits called L40 and S27. These gene products are thought to be processed by UCHs. Cleavage of isopeptide bond in the ubiquitin chains linked through Lys-48 of ubiquitin serves two purposes. One is to recycle ubiquitin after it has been used for marking a substrate for ubiquitination. Another function is to 'edit' the errors made by the ubiquitin-conjugating enzymes and reverse the ubiquitination reaction so that the substrate is no longer degraded. Also, editing function of DUBs probably serves to reverse the monoubiquitin attachment that marks membrane proteins for endocytosis.

Ubiquitin and Endocytosis

Ubiquitin also plays a role in other cellular processes such as endocytosis which regulates the number of receptor proteins in the plasma membrane. In endocytosis, the proteins on the cell surface are internalized in response to a specific signal such as neurotransmitter binding to its receptor. The internalized proteins are either recycled back to the membrane or targeted to the lysosome for degradation. Ubiquitin plays a role in marking the receptor proteins for internalization as well as sorting the endocytosed proteins for degradation in the lysosome. Unlike the polyubiquitin chain that marks substrate for proteasome-mediated degradation, endocytosis appears to be mainly mediated by attachment of single ubiquitins (monoubiquitination). When the receptors are monoubiquitinated, they bind to adaptor proteins called epsins that are part of clathrin-coated pits. Ubiquitinated receptor then undergoes endocytosis and gets incorporated into endosomes which in turn get sequestered into the multivesicular body. The membrane of the multivesicular bodies becomes continuous with lysosomes leading to degradation of the receptor. It has been shown that targeting the internalized receptor to the multivesicular body also requires ubiquitination.

Regulation of the Ubiquitin–Proteasome Pathway

Ubiquitin attachment to proteins is a dynamic process that is highly regulated by intracellular and extracellular signals. Ubiquitination and degradation of ubiquitinated proteins is controlled by modification of the substrate, modulation of ubiquitin ligases, and removal of ubiquitins. Substrates can be made either vulnerable or resistant to ubiquitin conjugation by phosphorylation. Ubiquitin ligase activity is increased or decreased through phosphorylation. Also, a new mode of regulation of ubiquitin ligases by ubiquitin-like proteins has been discovered. Removal of ubiquitins by DUBs serves as a checkpoint before committing a substrate for degradation. In addition, the 26S proteasome itself is subject to regulation. The activity of the 19S RC is regulated by induction or phosphorylation of its subunits as well as by proteins that are loosely associated with it. Also, the proteasome can be regulated by a change in the composition of the intrinsic subunits.

Ubiquitin–Proteasome Pathway and Synaptic Plasticity

The discovery of a physiological role for the ubiquitin–proteasome pathway came from the studies on long-term facilitation in the invertebrate *Aplysia*. It was found that the regulatory (R) subunit of cAMP-dependent protein kinase A (PKA) was a substrate for ubiquitin–proteasome-mediated degradation. Since then, several substrates of the ubiquitin–proteasome pathway have been identified in the nervous systems of other invertebrates as well as vertebrates (Table 1).

Degradation of the R Subunits of PKA

Evidence for a role for the ubiquitin–proteasome pathway in synaptic plasticity came from the investigation on persistent activation of PKA during induction of long-term facilitation. Studies on the biochemical mechanism of long-term facilitation in *Aplysia* indicated that PKA was persistently activated in the absence of elevated cAMP. How was PKA activated in the absence of sustained increase in cAMP? It was found that the R subunits of PKA were decreased without any change in the catalytic (C) subunit during induction of long-term facilitation. Since there was no change in mRNA for either the R subunit or the C subunit, it was concluded that R subunits were diminished perhaps through proteolysis. What is the mechanism of R subunit degradation? A series of

Table 1 Substrates and enzymes of the ubiquitin–proteasome pathway with roles in synaptic plasticity

Protein	*Functional category*	*Description*
Substrates		
GluR1	Neurotransmitter receptor	Ubiquitination of GluR1 in *C. elegans* reduces synaptic strength
PSD-95	Postsynaptic density protein	Degradation of PSD-95 causes endocytosis of AMPA receptors
R subunit	Subunit of cAMP-dependent protein kinase A (PKA)	Ubiquitin–proteasome-mediated degradation of R subunit persistently activates PKA during long-term facilitation (LTF) in *Aplysia*
Ap-C/EBP	Transcription factor	Degradation confines the activity of of Ap-C/EBP to a short time window during induction of LTF in *Aplysia*
Dunc-13	Presynaptic protein	Degrdation of Dunc-13 limits priming of synaptic vesicle release
Enzymes		
Bendless	Ubiquitin-conjugating enzyme (E2)	Bendless controls synapse formation in *Drosophila*
E6-AP	Ubiquitin ligase (E3)	Null mutation of E6-AP impairs synaptic plasticity and contextual memory
APC	Ubiquitin ligase (E3)	Controls synapse size in neurons; regulates abundance of glutamate receptors
SCF	Ubiquitin ligase (E3)	Regulates glutamate receptor number in *C. elegans*
Ap-uch	DUB	Blocking expression or function of Ap-uch inhibits LTF in *Aplysia*
Fatfacets	DUB	Fatfacets controls synapse development in *Drosophila*

biochemical experiments showed that R subunits were substrates for ubiquitination and proteasome-mediated degradation (Figure 3).

Moreover, an enzyme of the ubiquitin–proteasome pathway UCH (Ap-uch) was found to be induced by 5-HT, the neurotransmitter that induced long-term facilitation. Ap-uch was found to be critical for induction of long-term facilitation. The results showed that Ap-uch can increase the activity of the proteasome toward substrates such as the R subunits of PKA (Figure 3). Subsequent experiments found that at sensory–motor neuron synapses, injection of lactacystin, a specific proteasome inhibitor, blocked induction of long-term facilitation. Since R subunits inhibit the activity of C subunits of PKA, the results were interpreted to suggest that the ubiquitin–proteasome pathway operates to remove inhibitory constraints on formation of long-term memory. This has been substantiated by work carried out on the rat hippocampus. Bilateral infusion of lactacystin to the CA1 region of the rat hippocampus causes total retrograde amnesia for a one-trial avoidance learning. Also, total ubiquitination increases in the hippocampus 4 h after the training. These results are consistent with the idea that a decrease in some critical inhibitory proteins during long-term memory formation is mediated by the ubiquitin–proteasome pathway.

Synaptic Role of DUBs

DUBs appear to have a significant role in synaptic plasticity in the adult and in synapse formation during development. Studies on *Aplysia* showed a role in long-term facilitation for neuronal-specific UCH that is the homolog of human UCH-L1. *Aplysia* UCH (Ap-uch) is induced by stimuli that induce long-term facilitation but not stimuli that induce short-term facilitation. Injection of antibodies or antisense oligonucleotides specific to Ap-uch into sensory neurons synapsing onto motor neurons in culture blocked induction of long-term facilitation. Investigation on biochemical functions of Ap-uch indicated that Ap-uch is capable of cleaving small attachments to linearly attached ubiquitin molecules such as Ubiquitin–Ubiquitin–Cysteine but not large attachments like glutathionine-*S*-transferase (GST) in substrates like Ubiquitin-GST. More interestingly, additional biochemical analyses showed that Ap-uch associates with the proteasome. The association of Ap-uch increases the rate of degradation by the proteasome. For example, addition of recombinant Ap-uch to *in vitro* degradation systems showed that there was about twofold increase in degradation of R subunit of PKA. Since persistent activation of PKA has been shown to be critical for induction of long-term facilitation and R subunits of PKA were found to be substrates for the ubiquitin–proteasome pathway, the experiments on Ap-uch provided some molecular explanation for the role of regulated proteolysis in long-term facilitation. How does Ap-uch increase the rate of degradation by the proteasome? Using recombinant ubiquitin with its Lysine-48 mutated to Arg that cannot support multiple monoubiquitin attachments to the protein substrate, it was shown that Ap-uch stimulates the release of ubiquitin from substrates. Therefore, it can be inferred that Ap-uch perhaps cleaves the first ubiquitin in the polyubiquitin chain attached to the substrate.

The evidence for a role of DUBs in synapse formation came from a screen in *Drosophila* designed to identify the genes controlling size and strength of synapses. This screen found that *fatfacets* (*faf*), a gene that encodes a DUB, functions in synapse formation

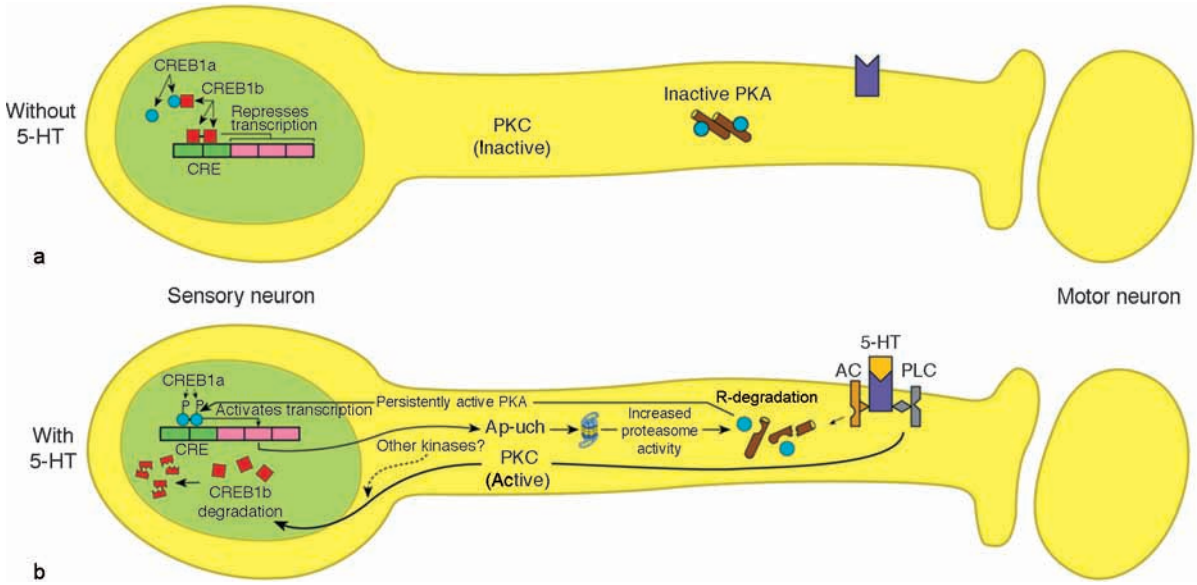

Figure 3 Role of the ubiquitin–proteasome pathway in long-term facilitation. The figure shows intracellular molecular mechanisms in the sensory neuron in sensory–motor neuron synapses in *Aplysia*. Without 5-HT stimulation (a), PKA and protein kinase C (PKC) are inactive and CREB1b represses transcription by blocking access of the transcription activator CREB1a to CRE or by forming inactive heterodimers with CREB1a. When sensory neurons are stimulated with the neurotransmitter 5-HT (b), which induces long-term facilitation, R subunits of PKA are ubiquitinated and degraded by the proteasome, making the kinase persistently active. The catalytic subunit of PKA (blue circle) translocates to the nucleus and phosphorylates CREB1a, the activator form of CREB. Concomitantly, the transcription repressor CREB1b is degraded by the ubiquitin–proteasome pathway. PKC, which is also activated by 5-HT-mediated signaling, stimulates ubiquitin conjugation to CREB1b and subsequent degradation. Relief of CREB1b-mediated repression and activation of CREB1a by PKA allows CREB-mediated gene induction to occur. One of the genes induced early in the gene expression cascade is Ap-uch, which through enhancement of the proteasome activity stimulates R subunit degradation and hence keeps PKA persistently active. Thus, the ubiquitin–proteasome pathway helps convert the transient stimulation of sensory neuron to long-term synaptic changes that underlie memory storage. 5-HT, serotonin; AC, adenylyl cyclase; PLC, phospholipase.

in the neuromuscular junction. Overexpression of *faf* in the developing *Drosophila* nervous system causes synaptic overgrowth and perturbs synaptic transmission. Another genetic screen for identifying the mutants that enhance the *faf* overexpression phenotype found that a loss-of-function in a gene called *highwire* (*hiw*) causes a significant increase in number of synapses. The *hiw* gene encodes a huge protein with 5233 amino acids. The highwire protein contains a RING finger domain, a characteristic feature of some ubiquitin ligases. Therefore, it is conceivable that highwire and fatfacets promote or inhibit the degradation of the same substrate by ubiquitinating or de-ubiquitinating it, respectively. Recently, highwire has been shown to negatively regulate signaling by bone morphogenetic protein (BMP)/transforming growth factor β (TGFβ), which stimulates presynaptic growth in *Drosophila* neuromuscular junction. Highwire binds to a protein in the BMP/TGFβ signaling pathway called Med. Although there is no direct evidence yet for ubiquitination of Med by highwire, the genetic data suggest that highwire functions to ubiquitinate Med for degradation. Moreover, synaptic overgrowth produced by overexpression of fatfacets is suppressed by disrupting BMP signaling.

Role of the Proteasome in Synaptic Plasticity

The 26S proteasome degrades proteins targeted by polyubiquitination. Initially, it was thought that ubiquitin–proteasome pathway removes constraints on long-term synaptic plasticity by degrading inhibitory proteins such as the regulatory subunits of PKA. This idea has gained additional support with recent results showing that transcription repressors such as a cAMP response element binding (CREB) repressor called CREB1b in *Aplysia* is degraded by the ubiquitin–proteasome pathway during treatments of neurons that induce long-term facilitation (Figure 3). The role of proteolysis is unlikely to be just confined to removal of inhibitory molecules, however. In support of this idea, recent studies show that degradation by the proteasome may also limit the extent of induction of long-term facilitation. Recently, a theory of local protein degradation has been put forward that can explain different, and often opposing, roles of the proteasome in different neuronal compartments. In support of this theory, experiments with *Aplysia* neurons and mouse brain show that proteasome has significantly higher levels of activity in the synaptic

terminals than in the other neuronal compartments such as the nucleus. Also, proteasome in the synaptic terminals and in the nucleus are differentially regulated by protein kinases.

Presynaptic Role of the Ubiquitin–Proteasome Pathway: Degradation of Synaptic Vesicle Proteins

The earlier observations on the role of the ubiquitin–proteasome pathway in synaptic plasticity mainly focused on long-term effects. Recent studies, however, indicate that the ubiquitin–proteasome pathway also functions in regulating short-term synaptic plasticity. For example, a protein Dunc-13, which is critical in priming the synaptic vesicles, is ubiquitinated and degraded by the proteasome in *Drosophila* neuromuscular synapse. Application of proteasome inhibitors and the dominant-negative mutation in a core subunit ($\beta 6$) of the *Drosophila* proteasome both lead to an increase in presynaptic accumulation of Dunc-13 protein. Also, application of the proteasome inhibitors lactacystin and epoxomycin causes an increase in the excitatory junctional current, suggesting that accumulation of Dunc-13 and the resultant increase in the net Dunc-13 quantity leads to enhanced synaptic transmission.

Is it likely that quantity of other synaptic proteins with a key role in neurotransmission is regulated by the ubiquitin–proteasome pathway? If so, regulated proteolysis may have a wider role in controlling short-term synaptic plasticity. As of now, two other synaptic vesicle proteins, synaptophysin and syntaxin, have been shown to be substrates for ubiquitin–proteasome-mediated degradation. Vulnerability of synaptophysin for ubiquitin–proteasome-mediated degradation was discovered indirectly by looking for proteins that bind to synaptophysin in yeast two-hybrid system. Two gene encoding for proteins Siah-1A and Siah-2 were isolated from rat brain two-hybrid cDNA library. These two genes are closely related to each other and are homologs of *Drosophila seven in absentia* gene. Siah proteins are single-subunit RING finger ubiquitin ligases with the RING finger domain in the N-terminal region. It is not clear which ligase (Siah1-A or Siah-2) is responsible for degradation of synaptophysins *in vivo* in neurons. Because the experiments were carried out in non-neuronal cells (Chinese hamster ovary cells) and PC12 cells, the physiological consequences of synaptophysin degradation remain to be determined.

Another presynaptic substrate for the ubiquitin–proteasome pathway is syntaxin 1. The group that isolated the Siah ligases also investigated ubiquitin–proteasome-mediated degradation of syntaxin 1 using the yeast two-hybrid strategy. The ligase that ubiquitinates syntaxin 1 is called staring (syntaxin 1-interacting RING finger protein). Co-expression of staring with syntaxin 1 in HeLa cells increases the degradation of syntaxin 1 which can be inhibited by the proteasome inhibitor MG132. Again, the physiological effect of ubiquitin–proteasome-mediated degradation of syntaxin 1 remains to be determined.

Presynaptic role for a ubiquitin ligase APC has been shown recently in *Drosophila* neuromuscular junction. Mutation in the APC gene leads to an increase in the number of synaptic boutons which can be reversed by targeted expression of the normal allele in neurons.

Postsynaptic Role of the Ubiquitin–Proteasome Pathway: Modulation of Neurotransmitter Receptors, Structural Proteins, and Regulatory Factors

Ubiquitin and proteasome seem to regulate proteins that control synaptic transmission such as neurotransmitter receptors and proteins that are part of the synaptic structure such as postsynaptic density proteins.

Investigations in *Caenorhabditis elegans* revealed a role for ubiquitin in endocytosis of GLR-1 type of glutamate receptor. Recent studies indicate that the multi-subunit ubiquitin ligases APC and SCF both regulate the abundance of GLR-1 in *C. elegans* although the mechanism is still unclear. In *Drosophila* neuromuscular junction, APC regulates the concentration of postsynaptic glutamate receptors. In mammalian hippocampal neurons, treatment with the proteasome inhibitor MG132 blocks agonist-induced endocytosis of α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type glutamate receptors. Also, AMPA-induced *N*-methyl-D-aspartate (NMDA)-induced internalization is prevented by application of the proteasome inhibitor. The exact mechanism of AMPA receptor endocytosis in mammals is not clear. It is possible that the AMPA receptor itself is not the target of proteasome action but a protein(s) that interacts with AMPA receptor might be the target. Given that both NMDA and AMPA agonists stimulate internalization that is inhibited by the proteasome, it is likely that this is a case of passive internalization and not an instance of regulated AMPA receptor endocytosis. In support of this idea, a postsynaptic density protein PSD-95 is regulated by ubiquitin–proteasome-mediated degradation, and degradation of PSD-95 leads to AMPA receptor internalization. Furthermore, application of the proteasome inhibitor MG132 to hippocampal slices reduces the magnitude of hippocampal long-term depression. A direct role for ubiquitin in regulated endocytotic degradation of AMPA receptors is still an open question, although it is likely that targeting of

AMPA receptors for degradation after internalization depends on ubiquitination.

Endocytosis of other neurotransmitter receptors might be directly or indirectly regulated by ubiquitination. Glycine receptor has been shown to be internalized upon ubiquitination. A protein associated with $GABA_A$ receptors, Plic-1, indirectly controls removal of $GABA_A$ through endocytosis. Interestingly, proteasome inhibitors prevent degradation of internalized $GABA_A$ receptors.

Ubiquitin–proteasome-mediated proteolysis appears to regulate several other proteins in addition to PSD-95 in the postsynaptic density proteins including several structural proteins. For example, Shank, GKAP, and AKAP79/150 are degraded through the ubiquitin–proteasome pathway. Unlike the degradation of PSD-95, physiological relevance of proteolytic removal of Shank, GKAP, and AKAP79/150 is not clear, because the studies were correlative and a direct link between ubiquitin–proteasome-mediated degradation of the PSD proteins and structural remodeling was not established.

There is also evidence that the ubiquitin–proteasome pathway controls a protein that regulates spine shape. Spine-associated Rapgap (SPAR) controls dendritic spine shape by reorganizing the actin cytoskeleton. During activity-dependent remodeling of synapses, SPAR was found to be degraded by the ubiquitin–proteasome pathway. Degradation of SPAR is stimulated by serum inducible kinase (SNK). Activity induces SNK mRNA in the cell body and the induced SNK is targeted to the dendritic spines. Because of the time required for SNK mRNA to travel to the spines, one speculation is that SPAR may function to oppose synaptic remodeling after elevated activity.

Future Perspectives

The research on the role of the ubiquitin–proteasome pathway in synaptic plasticity is in its infancy. Although we now know that proteolysis plays a role in inducing synaptic changes, we are far from elucidation of the mechanisms by which proteolysis contributes to synaptic plasticity. Also, future research has to take the complexity of the ubiquitin–proteasome pathway into consideration. For example, there are myriad components of the pathway such as numerous E3s and several DUBs. The elements of the ubiquitin pathway not only control the level of substrates but also regulate each other's function; for instance, E3s push the substrates in the direction of degradation, and some DUBs pull the substrates back from being committed to destruction. Also, molecules of the ubiquitin–proteasome pathway are known to possess nonproteolytic functions which might be critical for synaptic plasticity.

Synapse-specific plasticity requires local regulation of function. The idea of local function for ubiquitin–proteasome-mediated degradation at the synapse and other compartments of the neuron needs to be pursued in order to fully comprehend the role of proteolysis in synaptic plasticity. Rigorous studies on how proteolysis precisely controls substrates relevant to synaptic plasticity would require development of sophisticated experiments tools. In the coming years, we might expect to see exciting new discoveries on how proteolysis helps alter synapses.

See also: Extracellular Matrix Molecules: Synaptic Plasticity and Learning; Learning and Memory in Invertebrates: *Aplysia*; Learning and Memory in Invertebrates: *Drosophila*; Proteasome Role in Neurodegeneration; Transcription Factors in Synaptic Plasticity and Learning and Memory.

Further Reading

Ciechanover A (2005) Proteolysis: From the lysosome to ubiquitin and the proteasome. *Nature Reviews Molecular Cell Biology* 6: 79–87.

DiAntonio A, Haghighi AP, Portman SL, Lee JD, Amaranto AM, and Goodman CS (2001) Ubiquitination-dependent mechanisms regulate synaptic growth and function. *Nature* 412: 449–452.

Glickman MH and Ciechanover A (2002) The ubiquitin–proteasome proteolytic pathway: Destruction for the sake of construction. *Physiological Reviews* 82: 373–428.

Hegde AN (2004) Ubiquitin–proteasome-mediated local protein degradation and synaptic plasticity. *Progress in Neurobiology* 73: 311–357.

Hegde AN and DiAntonio A (2002) Ubiquitin and the synapse. *Nature Reviews Neuroscience* 3: 854–861.

Hegde AN, Goldberg AL, and Schwartz JH (1993) Regulatory subunits of cAMP-dependent protein kinases are degraded after conjugation to ubiquitin: A molecular mechanism underlying long-term synaptic plasticity. *Proceedings of the National Academy of Sciences of the United States of America* 90: 7436–7440.

Hegde AN, Inokuchi K, Pei W, et al. (1997) Ubiquitin C-terminal hydrolase is an immediate-early gene essential for long-term facilitation in *Aplysia*. *Cell* 89: 115–126.

Hicke L (2001) Protein regulation by monoubiquitin. *Nature Reviews Molecular Cell Biology* 2: 195–201.

Juo P and Kaplan JM (2004) The anaphase-promoting complex regulates the abundance of GLR-1 glutamate receptors in the ventral nerve cord of *C. elegans*. *Current Biology* 14: 2057–2062.

Song H, Smolen P, Av-Ron E, Baxter DA, and Byrne JH (2006) Bifurcation and singularity analysis of a molecular network for the induction of long-term memory. *Biophysical Journal* 90: 2309–2325.

Speese SD, Trotta N, Rodesch CK, Aravamudan B, and Broadie K (2003) The ubiquitin proteasome system acutely regulates presynaptic protein turnover and synaptic efficacy. *Current Biology* 13: 899–910.

Upadhya SC, Ding L, Smith TK, and Hegde AN (2006) Differential regulation of proteasome activity in the nucleus and the synaptic terminals. *Neurochemistry International* 48: 296–305.

Upadhya SC, Smith TK, and Hegde AN (2004) Ubiquitin–proteasome-mediated CREB repressor degradation during induction of long-term facilitation. *Journal of Neurochemistry* 91: 210–219.

van Roessel P, Elliott DA, Robinson IM, Prokop A, and Brand AH (2004) Independent regulation of synaptic size and activity by the anaphase-promoting complex. *Cell* 119: 707–718.

Zhao Y, Hegde AN, and Martin KC (2003) The ubiquitin proteasome system functions as an inhibitory constraint on synaptic strengthening. *Current Biology* 13: 887–898.

Ultrastructural Analysis of Spine Plasticity

J N Bourne and K M Harris, Medical College of Georgia, Augusta, GA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Dendritic spines are small protrusions that serve as the major site of excitatory synaptic transmission in the brain. The shape and density of spines can be affected by numerous factors, including developmental stage, experimental preparation, temperature, exposure to enriched environments, and neuronal pathology. In addition, physiological models of learning, including long-term potentiation (LTP) and long-term depression (LTD), have indicated that alterations in synaptic ultrastructure may underlie the storage of information in the brain.

Spine Structure and Function

Spine Shape

Dendritic spines can assume striking differences in size, shape, and subcellular composition both within and across brain regions (Figure 1). Most spines are unbranched protrusions that can be classified by their stubby, thin, or mushroom shapes. Branched spines have multiple heads which, in some brain regions, are all innervated by a single axonal bouton (e.g., hippocampal area CA3); in other brain regions the different heads of a branched spine are innervated by different axons (e.g., hippocampal area CA1); and in rare cases some of the heads have no presynaptic partners while other heads are innervated by different axons (e.g., cerebellar Purkinje spiny branchlets).

Localization of Organelles

Smooth endoplasmic reticulum (SER) is an organelle that is likely to be involved in sequestering calcium. Depending on the particular brain regions, few, many, or most of the dendritic spines contain SER (Figure 1). For example, only about 14% of the hippocampal CA1 spines contain SER, and most of it occurs laminated with dense-staining material into a structure known as the spine apparatus (as in Figure 1) in the large, complex spines. In contrast, nearly 100% of the cerebellar Purkinje spines contain SER in a tubular network. Polyribosomes are also present in or near the base of some dendritic spines (Figure 2(d)), suggesting that local protein synthesis of dendritic mRNAs can occur in the spines. In immature neurons, synaptic activity can shift the distribution of polyribosomes from the shaft to spines. Similarly, endosomal compartments including coated pits and vesicles, large vesicles, tubules, and multivesicular bodies are restricted to a subpopulation of dendritic spines that differs from spines that contain SER. Mitochondria rarely occur in dendritic spines and are usually restricted to those that are very large, complex, and highly branched. However, during periods of active synapse formation and remodeling in cultured neurons, mitochondria can localize to smaller dendritic spines.

Postsynaptic Targets

Evaluation of dendritic spine structure readily reveals them to be the major postsynaptic target of excitatory synaptic input (Figure 2(a)). Since all dendritic spines have at least one excitatory synapse on their head, more spines mean more synapses and accordingly more point-to-point connections in a neuronal ensemble involving spiny neurons. The size of the spine head also influences the amount of synaptic input, because larger spines are more sensitive to glutamate, the major excitatory neurotransmitter in the brain. Larger spines have larger postsynaptic densities (PSDs), which can then anchor more glutamate receptors. Thus, one function of spines is the preservation of the individuality of inputs. Occasionally, inhibitory or modulatory axons also form synapses on the heads, necks, or at the bases of dendritic spines. An inhibitory input on spines could act to 'veto' or modify the strength of the excitatory input.

Amplification of Voltage in Spine Head

The constriction in dendritic spine necks poses a small resistive barrier, thereby amplifying the depolarization attained in the immediate vicinity of the synapse, in contrast to that which would be generated if the synapse occurred directly on the wide dendritic shaft (Figure 2(b)). Computer simulations have revealed that most of the spine necks are sufficiently wide and short such that charge transfer to the postsynaptic dendrite is 85–100% complete within 100 ms after the initiation of a synaptic event. The time delay in charge transfer is sufficient, however, to provide a transient amplification of voltage at the spine synapse, which may facilitate opening of voltage-dependent channels in the spine head, such as the calcium channel associated with the *N*-methyl-D-aspartate (NMDA) class of glutamate receptors.

Sharing of Postsynaptic Potential

A long-standing hypothesis has been that the narrow dimensions of the spine neck would attenuate current

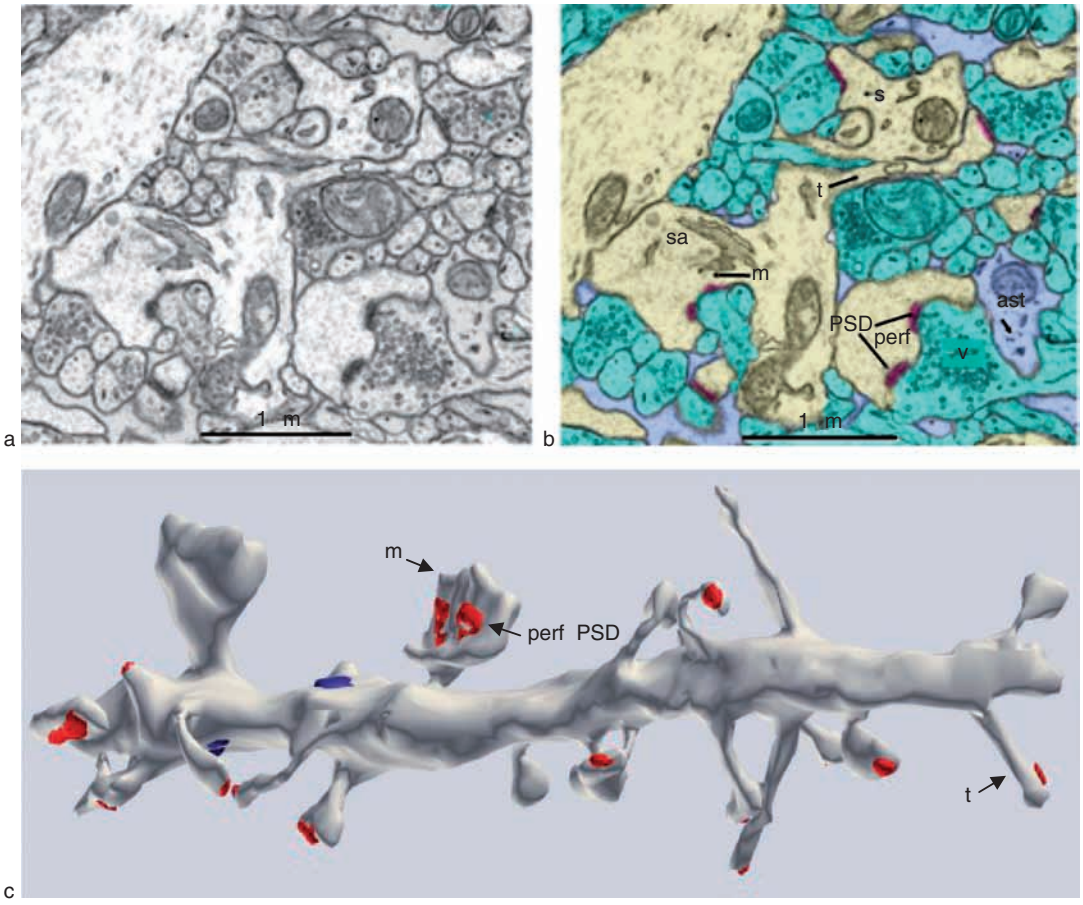

Figure 1 Dendritic spine shape. (a) Electron micrograph of a section through dendritic spines in stratum radiatum of hippocampal area CA1. (b) The same micrograph that has been color-coded to identify dendrites (yellow), axons (green), and astroglia (purple). In this fortuitous section, three spines were sectioned parallel to their longitudinal axis revealing spines of the stubby (s), mushroom (m), and thin (t) morphologies. The postsynaptic density (PSD) occurs on the spine head (see t) immediately adjacent and to a presynaptic axonal bouton that is filled with round vesicles (v). This t spine contains a small tube of smooth endoplasmic reticulum (ser) in its neck. In the m spine a spine apparatus (sa) is visible. A perforated PSD (perf PSD) is evident on the head of another mushroom spine. Near to this spine is a large astrocytic process (ast) identified by the glycogen granules and clear cytoplasm. (c) A 3-D reconstruction of a dendrite showing a variety of spine and synapse shapes.

flow between the spine head and the dendrite. Morphological evidence suggests, however, that most spine necks are not thin and long enough to significantly reduce the charge transferred to the parent dendrite. Current electrophysiological evidence from hippocampal CA1 cells suggests that the mean synaptic conductance for a minimal evoked response is 0.21 ± 0.12 ns, such that the current generated by release of 10–20 quanta would likely be fully transmitted to the postsynaptic dendrite. Thus, the constriction in the spine neck is not sufficient to prevent addition of voltage changes among co-activated synapses. Other models endow the spine with active membrane that would further enhance the sharing of postsynaptic potentials among neighboring spines (Figure 2(c)).

Biochemical Compartmentalization

Compartmentalization of calcium has now been demonstrated in the heads of dendritic spines under a variety of conditions. Two features of spines help to achieve localization of this second messenger in spine heads at least for a short time: (1) the spine neck could provide a narrow diffusion path and (2) a rise in spine calcium could cause release from SER, the intracellular calcium stores thereby amplifying the calcium signal. Biochemical compartmentalization in spine heads may also serve an important role in restricting calcium from the postsynaptic dendrite, thereby preventing excitotoxic cell damage such as microtubule breakdown and mitochondrial swelling. Since dendritic spines rarely have microtubules or mitochondria, high calcium concentrations in the spine head

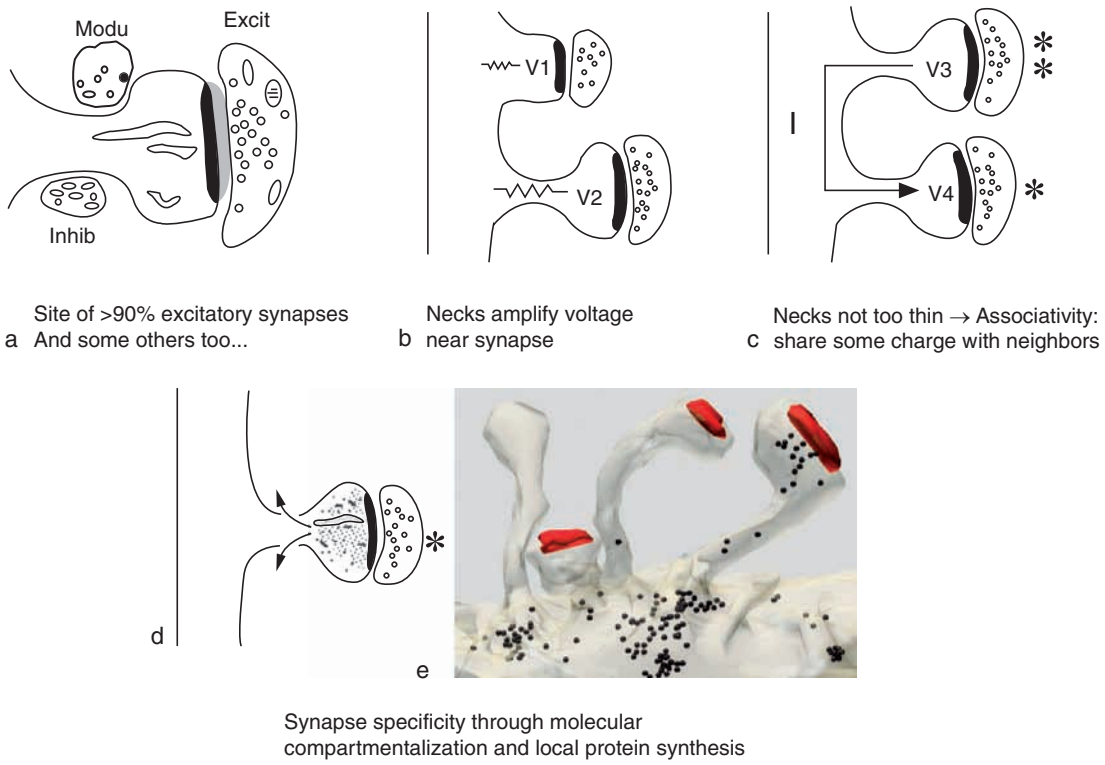

Figure 2 Dendritic spine function. (a) Spines exist as principal sites of excitatory synaptic transmission. Spines exist to (b) amplify electrical potential at the synapse and (c) promote associativity among neighboring synapses. Spine shape and resistance of the spine neck may influence potential (V) generated by synaptic activation. (d) Spines exist as molecular compartments. Smooth endoplasmic reticulum (tubules), calcium, and a myriad of other signaling mechanisms (stippling) are recruited in response to synaptic activation (asterisk). (e) Three-dimensional reconstruction of thin spines emerging from a dendrite. Polyribosomes (black dots) are most frequent at the base of dendritic spines, although they can also occur within them.

are less likely to have these detrimental effects. Not all spine morphologies would be expected to restrict diffusion, and only some spines have internal stores. A recent study that utilized caged glutamate to activate NMDA receptors and a low-affinity Ca^{2+} indicator (to minimize the perturbation of endogenous Ca^{2+} buffering) indicates that small spines with narrow necks exhibit large, isolated increases in $[Ca^{2+}]_i$. However, the necks of larger spines allow for a greater efflux of Ca^{2+} into the dendritic shaft at the base of the spine. Perhaps only a subset of spines, or alternatively all spines, but only at a restricted time during their history, achieve the compartmentalization of calcium, whether in spines or along a short segment of dendrite and its associated spines.

The specific localization of voltage-dependent calcium channels is also an important factor in determining which components of the dendritic arbor will compartmentalize and utilize relatively high concentrations of calcium. Whether spines preferentially sequester other second messengers remains to be determined. Certainly those tethered to the PSD are prevalent in spines, but further work is needed to determine how they are targeted to spines and whether compartmentalization in the spine head is a crucial element in their regulation. Recent evidence suggests that synaptic activity can create a bidirectional barrier to the diffusion of proteins, indicating a synapse-specific mechanism to amplify the biochemical signals necessary for establishing plasticity (Figure 2(d)). In addition, the presence of polyribosomes in some spines could provide the necessary machinery for generating plasticity-related proteins (Figure 2(d)).

Enhanced Connectivity

With the evolution of the brain, the increase in synaptic density had to be accompanied by a significant reorganization of the neuropil. Dendritic spines permit dendrites to synapse with neurons 1–2 μm away, which allows for increased synaptic density in densely packed neuropil (Figure 3(a)). Even some invertebrate neurons exhibit spinelike structures, indicating that dendritic spines appeared well before the evolution of the complex mammalian brain. Consider the simple case of an orthogonal relationship between dendrites and axons (Figure 3(b)). There can only be two

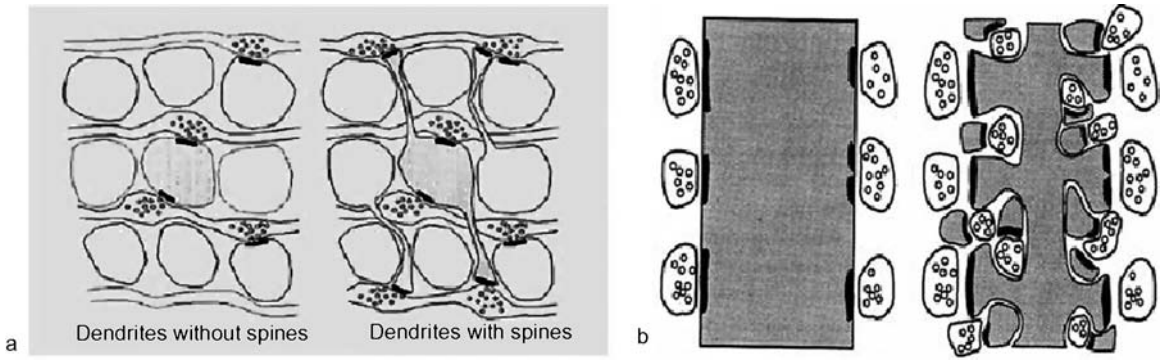

Figure 3 Enhanced connectivity between neurons. (a) Spines increase the packing density of synapses. Schematic illustrates a cross-section through two dendrites (shaded), one without and one with dendritic spines. Convolution and interdigitation of dendrite, axon, and spine membranes support more synapses. (b) The presence of spines also allows for an increase in synaptic density without increasing the overall volume of the brain.

synapses on either side of the dendrite in any given plane without the presence of spines. Dendrites with spines can reach beyond their immediate perimeter to connect with axons in nearby rows, thereby at least doubling the density of possible connections. In addition, the shape of dendritic spines allows efficient interdigitation between neighboring processes, thus achieving the high synapse packing density in the neuropil.

Synaptic Activity and Spine Organization

Spine Stability and Turnover

The stability of spines largely depends on the level of synaptic activity and the developmental stage of the neuron. In immature neurons, spines and filopodia turn over rapidly as connections are formed and stabilized. Only a small amount of glutamate is required to stabilize a spine. Excitotoxicity resulting from too much glutamate can result in the disappearance of spines, while a block of synaptic transmission can result in robust spinogenesis, particularly on mature neurons following traumatic injury or exposure to cold temperatures as occurs during hibernation and some surgical and experimental procedures. The appearance and disappearance of spines may reflect the capacity of mature neurons for recuperative synaptogenesis or a form of synaptic scaling, a homeostatic mechanism that ensures that the output of a postsynaptic cell remains constant despite alterations in synaptic input. In contrast, young neurons do not exhibit spinogenesis with blocked transmission, but rather insert more receptors to maintain synaptic homeostasis.

Brain regions also differ in the amount of synaptic turnover that occurs with the modulation of sensory experience. Recent technological advances have allowed dendritic spines of the somatosensory cortex to be imaged *in vivo* for periods of time ranging from a few hours to a few months. Small spines (thin or filopodial) appear and disappear in an experience-dependent manner associated with synapse formation or elimination, whereas larger spines persist for months. The effect of sensory experience on synapse number depends on the rates of synapse formation and elimination at different developmental stages, and the percentage of stable spines increases with age.

In the visual cortex of adult mice, *in vivo* imaging reveals that the majority of spines (75–90%) are stable under baseline conditions. In another study, rats reared in the dark until postnatal day 30 have a significantly lower spine density compared to light-reared controls and also have shorter, rounder spines. Although an increase in spine head diameter is observed in dark-reared animals that are exposed to light for 10 days, the low spine density is not reversed. Presuming that an increase in the number of receptors present in the PSD accompanies the increase in spine head diameter, the parent dendrite could still preserve the level of synaptic input even in the absence of spinogenesis.

In the neocortex, cerebellum, and hippocampus, localization of different afferent inputs on the same dendritic arbor is dictated by the laminar structure of these brain areas. However, in the lateral nucleus of the amygdala, inputs from the cortex and thalamus are interspersed on the same dendritic branches and afferent specificity is determined by spine morphology. Large spines are contacted by thalamic afferents and exhibit larger Ca^{2+} transients following activation than smaller spines contacted by cortical afferents. The increased Ca^{2+} signaling observed in large spines is mediated by R-type voltage-dependent Ca^{2+} channels that are preferentially localized to thalamic inputs. Thus, in the amygdala, organization of afferent inputs

can be regulated at the level of an individual spine by altering its morphological and molecular properties.

Perisynaptic Astroglia and Spine Structure

Traditionally, astroglia have been thought of as 'support' cells for neurons. However, recent evidence suggests that astroglia may play a much larger role in the maintenance and plasticity of synapses. Release of the cytokine tumor necrosis factor alpha (TNFα) by glial cells enhances excitatory synaptic strength by increasing the surface expression of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors on dendritic spines while reducing inhibitory synaptic strength by causing the endocytosis of γ-aminobutyric acid ($GABA_A$) receptors. Since astroglial processes can project extensively throughout the neuropil, astroglia are in a position to scale synaptic responses at the level of entire neuronal networks. In addition to affecting the physiological properties of synapses, astroglia can also regulate the structure of dendritic spines in response to synaptic activity. Transient interactions between cell surface molecules such as the ephrinA3 ligand and the EphA4 receptor regulate the structure of excitatory synaptic connections through neuro-glial cross-talk. Activation of EphA4 by ephrin-A3 induces spine retraction, and inhibiting ephrin/EphA4 interactions distort spine shape and organization. In addition to Ephs and ephrins, astrocytic membranes contain other contact-mediated factors that influence synaptic maturation. Astroglial processes tend to be found around larger and presumably more mature synapses, while areas that have undergone recent spinogenesis tend to be devoid of astroglia. Thus in the mature brain, astroglia may be involved in enhancing the stability of spines and preventing the extension of new dendritic protrusions. Interactions between cell surface molecules and the release of various soluble factors by astroglia may be of crucial importance to the turnover and structural alteration of spines observed with synaptic plasticity.

Structural Synaptic Plasticity at Dendritic Spines

Long-Term Potentiation

LTP is a cellular model of learning and memory that has been extensively investigated regarding changes in spine and synapse structure. Two key structural changes would contribute to a potentiated synaptic response: an increase in the number of spines or an increase in the size of synapses with LTP. Several studies in the hippocampus over the last 30 years have attempted to link structural alterations of the synapse to the increase in synaptic strength, but the results remain controversial, particularly in the adult brain. Numerous studies involving *in vivo* induction of LTP in the dentate gyrus of adult rats have led to differing conclusions about the amount of structural plasticity that occurs. Early studies describe an input-specific increase in spine volume and a shortening of spines accompanied by an enlargement of the spine neck. Overall spine density and branched spines in particular are increased 30 min after the *in vivo* induction of LTP in the dentate gyrus, although the increase in branched spines may be due to high-frequency stimulation alone, even in the absence of LTP. While there has been some speculation that branched spines might arise from the splitting of individual potentiated spines, careful three-dimensional (3-D) reconstructions of the neuropil have revealed that branched spines never synapse on the same axonal bouton. In addition, the number of neuronal and dendritic processes that pass between the branches argues against the idea of spine splitting. An alternative hypothesis suggests that branched spines arise from the formation of new protrusions called filopodia which can be initiated by the enhanced synaptic activity, navigate between neighboring processes to find a presynaptic partner, and then become dendritic spines. In any case, the majority of morphological alterations that have been attributed to *in vivo* induction of LTP have included more subtle changes such as increases in the concavity of spine heads and an increase in PSD area, as well as an increase in the ratio of perforated to nonperforated PSDs. Because perforated PSDs tend to occur in mushroom spines, this change could indicate an alteration in spine morphology if potentiation results in a shift from thin spines to mushroom spines.

Uncertainty also exists about the extent of structural plasticity that occurs following the induction of LTP in area CA1 of the adult rat hippocampus. Early studies that analyzed single EM images suggest that there is an increase in shaft and sessile synapses following induction of LTP. However, it was not specified whether these synapses were symmetric (inhibitory) or asymmetric (excitatory). More recent studies specifically aimed at detecting changes in synaptic density failed to detect a reliable increase in synapse number after induction of LTP. Although changes in synaptic density have been difficult to detect, recent evidence suggests that spines can get bigger with LTP. When LTP is induced via stimulation of the Schaffer collaterals, a small fraction of spines transiently expand within minutes of tetanic or theta burst stimulation, but return to their initial size within a few minutes. The percentage of expanding spines varies directly with stimulus strength, presumably

reflecting an increase in number of presynaptic fibers recruited. Even with a stimulation intensity that evokes a maximal field excitatory postsynaptic potential (fEPSP), no more than 10% of spines respond. Thus, the relatively small fraction of expanding spines that synapse with the presynaptic terminals activated by field stimulation reflects the challenges present in detecting plasticity-related changes in the adult brain. Even serial section electron microscopy and 3-D reconstruction of dendrites 2 h post-LTP did not reveal net changes in the density, shape, or size of spines. One possibility is that synaptic proliferation occurs immediately after LTP induction, but with time, competition between recently potentiated spines and their nonpotentiated neighbors results in the elimination of the nonpotentiated spines; thus, the dendrite maintains its optimal number of synaptic connections. Under such a scenario, individual spines could exhibit dramatic plasticity while the overall population number and average size of dendritic spines would appear unchanged.

Although the mature brain presents several challenges to describing the morphological changes that underlie synaptic plasticity, the immature nervous system has provided more consistent results. Dynamic changes in local dendritic segments that were targeted for potentiation-induced changes have been observed in organotypic slice cultures. Significant local outgrowth of filopodia-like dendritic processes is seen 20 min after delivery of a tetanic stimulus, and some of the activity-induced filopodia later turn into spine-like structures. When LTP is induced by pairing postsynaptic depolarizations with single presynaptic stimuli and transmitter release is blocked everywhere but in the small area being imaged, two to nine new spines are formed in the nonblocked region, emphasizing the localized specificity of structural changes associated with LTP. Similarly, stimulation of individual spines with uncaged glutamate induces a localized increase in spine head diameter that is apparent within 1–5 min, and 30% of the stimulated spines remain enlarged 70–100 min after stimulation. The magnitude of potentiation correlates with early or long-lasting spine enlargement, suggesting a high level of fidelity between physiological and morphological changes. Spine and PSD enlargement are also associated with the translocation of polyribosomes into spines following the induction of LTP in acute PN15 hippocampal slices. Given that the maintenance of LTP is dependent on the synthesis of new proteins, the presence of polyribosomes in a subset of spines may provide an indication of which spines underwent potentiation.

Although quite a bit of uncertainty still surrounds the set of structural changes associated with LTP, it appears likely that the effect is selective to potentiated synapses and that changes in a tissue will be correlated with the numbers of stimulated axons and dendrites. Nevertheless, there is evidence that LTP-inducing stimuli can result in increases in spine size and number and may involve the remodeling of the PSD and the translocation of organelles into activated spines. All of these changes could form the basis of long-lasting changes in synaptic strength.

Long-Term Depression

In contrast to LTP, LTD is a long-lasting reduction in synaptic transmission that may also play an integral role in the processing and retention of information. Although the literature concerning structural changes associated with LTD is much less abundant, there does appear to be a correlation between a reduction in synaptic response and a decrease in the number and size of dendritic spines. Low-frequency stimulation (900 pulses at 1 Hz) results in the shrinkage of dendritic spines in acute hippocampal slices and the retraction of spines in organotypic hippocampal slice cultures. While these results indicate a strong link between structural and synaptic plasticity, much work remains to be done to fully elucidate the morphological correlates of LTD.

See also: Developmental Synaptic Plasticity: LTP, LTD, and Synapse Formation and Elimination; Endocytic Traffic in Spines; Eph Receptor Signaling and Spine Morphology; LIM Kinase and Actin Regulation of Spines; Long-Term Depression: Cerebellum; Long-Term Potentiation and Long-Term Depression in Experience-Dependent Plasticity; Long-Term Potentiation (LTP).

Further Reading

Bliss T, Collingridge G, and Morris R (eds.) (2004) *Long-Term Potentiation: Enhancing Neuroscience for 30 Years*. New York: Oxford University Press.

Ethell IM and Pasquale EB (2005) Molecular mechanisms of dendritic spine development and remodeling. *Progress in Neurobiology* 75: 161–205.

Harris KM and Kater SB (1994) Dendritic spines: Cellular specializations imparting both stability and flexibility to synaptic function. *Annual Review of Neuroscience* 17: 341–371.

Harris KM and Stevens JK (1988) Dendritic spines of rat cerebellar Purkinje cells: Serial electron microscopy with reference to their biophysical characteristics. *Journal of Neuroscience* 8: 4455–4469.

Harris KM and Stevens JK (1989) Dendritic spines of CA1 pyramidal cells in the rat hippocampus: Serial electron microscopy with reference to their biophysical characteristics. *Journal of Neuroscience* 9: 2982–2997.

Nimchinsky EA, Sabatini BL, and Svoboda K (2002) Structure and function of dendritic spines. *Annual Review of Physiology* 64: 313–353.

Peters A, Palay SL, and Webster HDF (1991) *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells,* 3rd edn. New York: Oxford University Press.

Sorra KE and Harris KM (2000) Overview on the structure, composition, function, development, and plasticity of hippocampal dendritic spines. In: Eichenbaum HB (ed.) with Harris KM and Sorra KE (special issue eds.) *Dendritic Spines of the Hippocampus. Hippocampus* 10: 501–511.

Stuart G, Spruston N, and Hausser M (eds.) (2001) *Dendrites.* New York: Oxford University Press.

Yuste R and Bonhoeffer T (2001) Morphological changes in dendritic spines associated with long-term synaptic plasticity. *Annual Review of Neuroscience* 24: 1071–1089.

Relevant Website

http://synapses.clm.utexas.edu – synapse web.

Ultrastructural Organization of Release Sites in the Calyx of Held

T Kuner, University of Heidelberg, Heidelberg, Germany

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The calyx of Held is a giant nerve terminal mediating high-frequency excitatory input to principal cells of the medial nucleus of the trapezoid body (MNTB). This synaptic connection is part of the lower auditory pathway and contributes to sound localization. Primary auditory neurons, receiving excitatory input from cochlear hair cells, form giant synapses called endbulbs of Held with neurons located in the cochlear nucleus (e.g., globular bushy cells). The latter extend their axons via the ventral acoustic stria to the contralateral MNTB, where they form the calyx of Held terminal on principal cells (Figure 1(a)). This 1:1 axosomatic synapse forms a highly reliable relay, being capable of transferring presynaptic action potentials to the postsynaptic side (Figure 1(b)) at rates of up to 1 kHz. The principal cell forms inhibitory synaptic connections with neurons located in the medial and lateral superior olivary complex (SOC). The neurons of the SOC also receive ipsilateral excitatory input from neurons of the cochlear nucleus and therefore can function as coincidence detectors. Detection of coincident inhibitory and excitatory signals originating from the contralateral and ipsilateral sides, respectively, is thought to underlie sound localization. Embedded in this physiological framework, the calyx of Held faces a truly remarkable set of functional requirements: the resting frequency of action potentials entering the calyx of Held is approximately 100 Hz, sound stimuli can evoke action potential frequencies of up to 1 kHz, and the precise timing of action potentials is critical for sound localization. A set of highly distinctive morphological specializations provide the calyx of Held with this functionality.

Structural and Functional Maturation

The calyx of Held is not a singular structure–function unit but, rather, shows many different age-dependent faces. During the first three postnatal weeks, the calyx undergoes a series of marked structural and functional changes referred to as the morphogenetic sequence (Figure 2). By postnatal day 2 (P2), a protocalyx with a growth cone-like appearance has formed. Active zones (AZs) are large and contain the typical complement of structures. The first events of synaptic transmission reveal postsynaptic currents of small amplitude. By P5–P7, the calyx has a cuplike shape covering much of the postsynaptic cell. Functionally, they are characterized by action potentials with slow kinetics and limited bandwidth, postsynaptic currents of small amplitude, slow kinetics, and pronounced short-term depression. By P8–P10, the calyx has matured further and forms a single large compartment (Figure 2). Approximately half of the surface of the principal cell is covered by the calyx, giving rise to a large area of membrane apposition. At this still immature stage, action potentials have become faster, postsynaptic currents larger and faster, and synaptic depression somewhat less pronounced. At approximately the same time as the onset of hearing, approximately P12–P16, the calyx acquires a progressively fenestrated morphology: the single compartment is split into several fingerlike processes. Furthermore, the variability of structures encountered during this time frame is large, ranging from immature to mature morphologies. Functionally, action potentials and postsynaptic currents are even faster and synaptic depression is strongly decreased (Figure 2). By P21, structural maturation is almost complete; the calyx now forms many fingers or stalks decorated with protrusions called swellings (Figure 2). The functional state on P21 remains unknown because progressive myelination of the brain stem and the process of morphological differentiation impede paired pre- and postsynaptic patch clamp recordings. The structural maturation is ongoing to at least P28, a postnatal stage at which hearing is fully developed. These examples illustrate that any description of structure and function of the calyx needs to consider the appropriate postnatal maturational stage.

Distribution of Synaptic Contacts

Like other giant nerve terminals, such as the giant terminal of the squid, the calyx contains a large number of synaptic contacts (SCs), specialized sites of neurotransmitter release and postsynaptic signal transduction. An immature P9 calyx contains approximately 600 synaptic contacts distributed at average distances of 590 nm across the entire membrane area of the terminal apposed to the postsynaptic side (Figures 3(a) and 3(b); Table 1). Interspersed between synaptic contacts, approximately twice as many puncta adherentia (PA) – desmosome-like junction complexes – are presumed to provide mechanical stability of the pre- and postsynaptic membranes. Both SCs and PA are not randomly distributed, suggesting that an unknown structural principle governs the distribution of these membrane specializations.

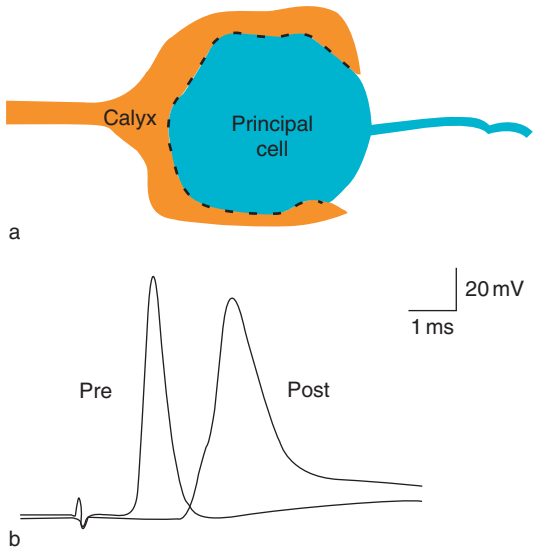

Figure 1 The calyx of Held excitatory model synapse. (a) Scheme of the presynaptic calyx of Held giant nerve terminal (orange) and its cognate postsynaptic cell (cyan), the principal cell of the medial nucleus of the trapezoid body. Note multiple synaptic contacts (black). The diameter of the principal cell ranges from 15 to 25 μm. (b) Simultaneous pre- and postsynaptic patch clamp recording from a P9 calyx of Held synapse. Reproduced from Borst JG, Helmchen F, and Sakmann B (1995) Pre- and postsynaptic whole-cell recordings in the medial nucleus of the trapezoid body of the rat. *Journal of Physiology (London)* 489: 825–840, with permission from Blackwell Publishing Ltd.

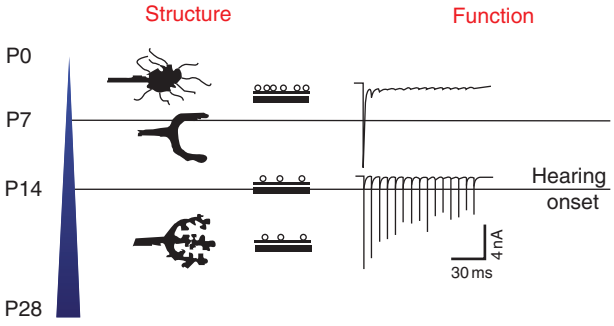

Figure 2 Structural and functional postnatal maturation. Maturation during the first four postnatal weeks (blue time bar), with individual elements arranged according to the postnatal day shown on the left. Morphology as apparent in light microscopy and ultrastructure of synaptic contacts (circles, synaptic vesicles; thick bar, postsynaptic density). Thin lines emanating from the growth cone-like calyx at P2 represent filopodial extensions. The mature calyx at P21 forms fingers covered with numerous swellings. Short-term depression traces reproduced from Taschenberger H and von Gersdorff H (2000) Fine-tuning an auditory synapse for speed and fidelity: Developmental changes in presynaptic waveform, EPSC kinetics, and synaptic plasticity. *Journal of Neuroscience* 20: 9162–9173.

On a functional level, synaptic contacts are arranged in parallel, allowing the calyx terminal to act like several hundred independent small bouton synapses. This functionally parallel arrangement is a major mechanism providing the calyx with the unusual functional properties mentioned previously.

The number of synaptic contacts increases with postnatal maturation (Table 1), and their spatial distribution becomes more distinct. At the onset of hearing on approximately P14, a new structural unit appears, the so-called wrapped donut: within individual swellings, synaptic vesicles are arranged in a donutlike manner, with a cluster of mitochondria in the donut center and active zones predominantly located underneath the synaptic vesicle donut (Figures 3(c) and 3(d)). This structural arrangement develops in synchrony with the fenestration process of the calyx (Figure 3(c)) and is fully expressed on P21. It may optimize the geometries of sites producing ATP and sites using ATP, thereby supporting high-frequency synaptic transmission. Hence, in addition to the number of synaptic contacts, their spatial distribution with progression of maturation needs to be considered when examining structure–function relationships of the calyx.

Ultrastructure of Synaptic Contacts at the P9 Calyx

A synaptic contact (also referred to as synaptic complex) consists of a presynaptic cluster of synaptic vesicles (SVs), an AZ comprising cytomatrix (CAZ) and presynaptic plasma membrane, the synaptic cleft, the postsynaptic plasma membrane, and the postsynaptic density (PSD). This arrangement is similar to terminals of other central synapses containing a single AZ.

Figure 4 shows a typical example taken from a P9 calyx. The SV cluster associated with an SC typically occupies a radial shell of approximately 500 nm and contains approximately 100–300 SVs (Table 1). Because SCs are spaced at distances of approximately 600 nm, these clusters may contain SVs that could be assigned to multiple SCs. Within a more restricted radial shell of 200 nm, all SVs can be considered to be only part of the associated AZ, yielding a cluster size of approximately 63 SVs. In aldehyde-fixed tissue, SVs are densely packed as illustrated in Figure 4, but cryo-fixed tissue often reveals less populated SV clusters. Small electron-dense protrusions, possibly the SV protein synapsin, emanate from SV processes and appear to cross-link SVs.

AZs contain electron-dense material, reflecting the presence of the CAZ (Figure 4, black). The shape of AZs viewed from the top can be disk-like or polygonal. When viewed in cross section, AZs of aldehyde-fixed tissue are often concave, resembling the shape of a boomerang. In contrast, AZs in cryo-fixed material are straight, suggesting that the concave shapes observed after aldehyde fixation are not physiological. The diameter of AZs ranges from 200 to 800 nm, and

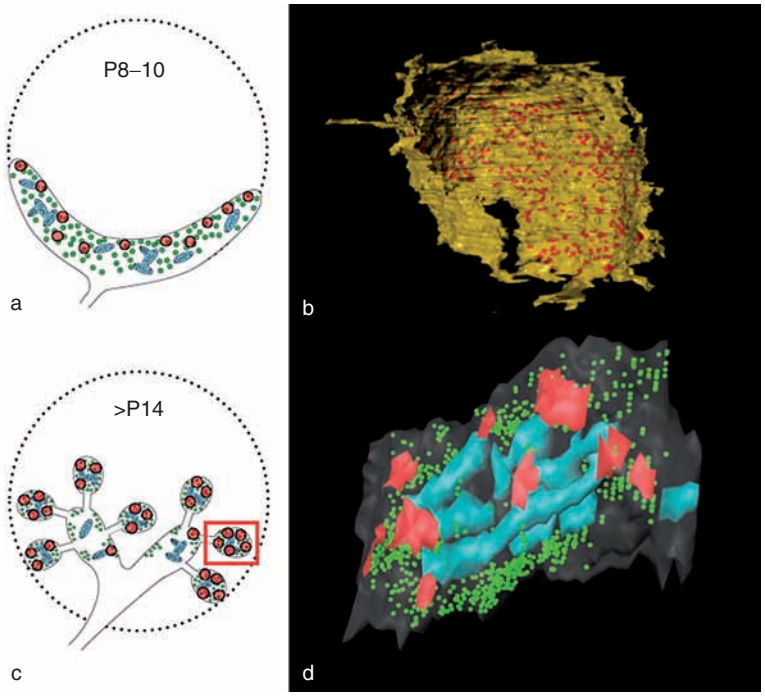

Figure 3 Calyx morphology and distribution of active zones. (a) Schematic drawing of an immature calyx (red, active zones; cyan, mitochondria; green, synaptic vesicles). The dotted line represents the outline of the postsynaptic principal cell. (b) *En face* view of the calyx apposition zone with the principal cell obtained from a electron microscopic three-dimensional reconstruction of an entire calyx (red, active zones; yellow, presynaptic plasma membrane). (c) As in a but showing a mature calyx. Frame represents structure shown in (d). (d) Three-dimensional reconstruction of a wrapped donut (colors as in schematic drawing). The entire structure is approximately 1.5 μm in diameter. (b) Reproduced from Sätzler K, Sohl LF, Bollmann JH, et al. (2002) Three-dimensional reconstruction of a calyx of Held and its postsynaptic principal neuron in the medial nucleus of the trapezoid body. *Journal of Neuroscience* 22(24): 10567–10579. (d) Reproduced from Wimmer VC, Horstmann H, Groh A, and Kuner T (2006) Donut-like topology of synaptic vesicles with a central cluster of mitochondria wrapped into membrane protrusions: A novel structure–function module of the adult calyx of Held. *Journal of Neuroscience* 26(1): 109–116.

the surface area of an AZ is approximately 0.1 μm^2. SVs directly contacting the AZ are commonly referred to as morphologically docked. Often, a less stringent distance criterion for docked SVs of $\leq$10 nm distance yields similar numbers, whereas considering SVs within a distance of 20 nm results in higher counts. Based on the average size of AZs, approximately 50 SVs could theoretically dock at an AZ. However, AZs on P9 have an average of only 2 or 3 docked SVs. Hence, a specific molecular mechanism may exist that represents individual SV docking sites. It remains unclear if such docking sites can always mediate neurotransmitter release and thereby represent functional release sites (Figure 5).

The synaptic cleft is approximately 20–30 nm wide and is partially filled with fuzzy electron-dense material. In aldehyde-fixed tissue, the synaptic cleft is wider than the space between the membranes surrounding the synaptic contacts (Figure 4, arrows designate transition). This feature could be due to the fixation method because cryo-fixed tissue shows similar or even larger spaces surrounding the synaptic contacts.

The PSD is characterized by an accumulation of electron-dense material and contains neurotransmitter receptors and matrix proteins orchestrating postsynaptic signal processing. The PSDs of principal cells of the calyx contain mostly glutamate receptors of the AMPA type consisting of GluR-D_o subunits, the glutamate receptor subunits conferring the fastest activation, desensitization, and deactivation kinetics.

Ultrastructural Changes during Postnatal Maturation

Synaptic contacts of calyces on P5 or P14 exhibit the same basic design previously described for P9 contacts, but there are several differences. At P5 calyces, SVs are clustered around AZs, whereas at P14 they are more distributed within the volume of the calyx. Furthermore, P5 AZs contain six docked SVs and P14 AZs only three SVs, whereas the size of the AZs decreases with maturational age. The number of docked SVs and the size of the AZ are linearly related. Furthermore, the number of docked SVs normalized

Table 1 Quantitative morphological parameters

	P9[a]	*P5*[b]	*P14*[b]
Calyx surface (μm^2)	2500		
Principal cell surface (μm^2)	2400	1058	768
Calyx volume (pl)	0.48/0.9[c]		1.3[c]
Principal cell volume (pl)	3.4		
Apposition area (%)	40	63	56
SCs area/apposition area (%)	10	5	5
No. of SCs	554	303	678/700[c]
No. of PAs	1010	765	1447
Surface area of SC (μm^2)	0.1	0.085	0.055
Surface area of PA (μm^2)	0.046		
Mean SC–SC distance (μm)	0.59		
Mean PA–PA distance (μm)	0.37		
Synaptic cleft width (nm)	28		
Diameter of SVs (nm)	46	46	42
No. SVs/AZ			
Direct	2	6	3
$\leq$10 nm	2		
$\leq$20 nm	5		
$\leq$50 nm	18		
$\leq$200 nm	63		
$\leq$500 nm	125		
No. docked SVs/calyx	1100	1697	2117

[a]Data from Sätzler K, Sohl LF, Bollmann JH, et al. (2002) Three-dimensional reconstruction of a calyx of Held and its postsynaptic principal neuron in the medial nucleus of the trapezoid body. *Journal of Neuroscience* 22(24): 10567–10579.
[b]Data from Taschenberger H, Leao RM, Rowland KC, Spirou GA, and von Gersdorff H (2002) Optimizing synaptic architecture and efficiency for high-frequency transmission. *Neuron* 36(6): 1127–1143.
[c]Data obtained from confocal scans of labeled calyces from Wimmer VC, Horstmann H, Groh A, and Kuner T (2006) Donut-like topology of synaptic vesicles with a central cluster of mitochondria wrapped into membrane protrusions: A novel structure–function module of the adult calyx of Held. *Journal of Neuroscience* 26(1): 109–116.
AZ, active zone; PA, puncta adherentia; SCs, synaptic contacts; SVs, synaptic vesicles.

to the AZ area remains unchanged between P5 and P14 calyces, as both AZ area and number of SV decrease. Thus, the concentration of a putative docking complex within the AZ and the area of the AZ may define the number of SVs docked at the AZ.

Another obvious difference between P5 and P14 calyces concerns dense core vesicles, which are very prevalent before P5 and decrease in number thereafter. A subset of these dense core vesicles could be piccolo-bassoon transport vesicles, precursors of AZs. An increased number of dense core vesicles during synapse formation between P1 and P4 is consistent with their role as AZ precursors.

In summary, the ultrastructure of synaptic contacts changes within the first two postnatal weeks mostly with regard to the size of the AZs and the number of dense core vesicles, both of which decrease.

Ultrastructural Specializations of the Calyx

The mitochondria-associated adherens complex (MAC) links mitochondria to PA and consists of a mitochondrial plaque, a vesicular chain, and filamentous material. By linking mitochondria to PA, mitochondrial functions such as ATP and glutamate supply as well as Ca^{2+} buffering can be anchored close to AZs. This spatial arrangement could, similar to the wrapped donut geometry, enhance high-frequency synaptic transmission.

Spine-like protrusions of the postsynaptic membrane into the presynaptic compartment are frequently observed. They often contain one or more synaptic contacts, but their functional relevance remains unclear.

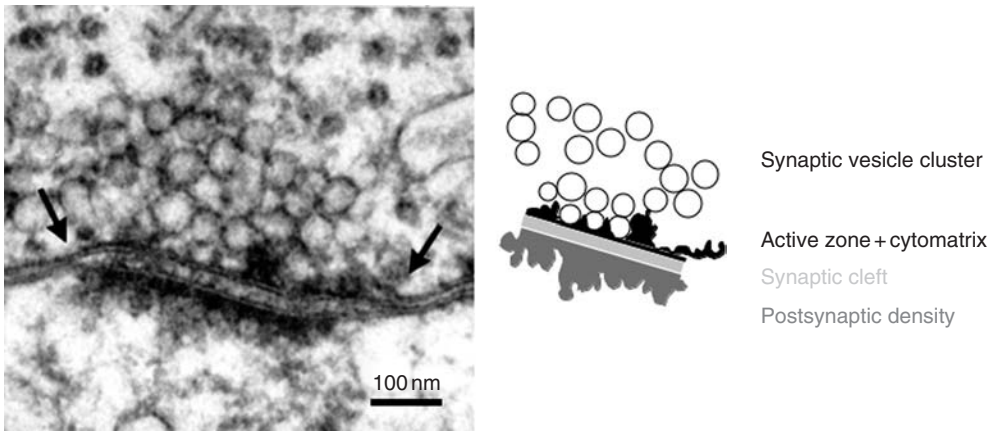

Figure 4 Ultrastructure of a synaptic contact. Electron micrograph showing a synaptic contact of a P9 calyx (aldehyde fixed). Arrows indicate lateral border of the active zone. Schematized version of the image is shown on the right. Reproduced from Rollenhagen A and Lubke JH (2006) The morphology of excitatory central synapses: From structure to function. *Cell and Tissue Research* 326: 221–237, figure 3a, Copyright 2006. With kind permission of Springer Science and Business Media.

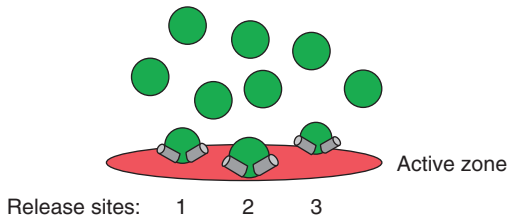

Figure 5 Release sites and active zones. Scheme of an active zone with three release sites. Docked synaptic vesicles are highlighted by gray barrels. Red, active zone; green, synaptic vesicles.

The presynaptic grid, a hexagonal array of electron-dense material with center-to-center distances of 100 nm, has been reported as a specialization of the presynaptic density in other synapses. However, clear evidence for such a regular grid is lacking at the calyx, although structures consistent with the presence of a grid are evident in Figure 4. The presynaptic grid is thought to provide docking slots for SVs. However, its regular gridlike architecture would predict many more docking sites than docked SVs present at any postnatal stage of the calyx. Thus, access to the docking sites might be restricted by some molecular mechanism. Alternatively, the presynaptic grid may not be fully formed at synaptic contacts of the calyx of Held.

Ultrastructure and Function

Number of Releasable SVs

A P9 calyx contains approximately 1200 SVs docked at the AZ. Different physiological methods revealed a readily releasable pool (RRP) size of 800–5000 SVs. This large variability could be due to the different sensitivities of the methods and different definitions of the RRP. Nevertheless, the number of docked SVs roughly corresponds to the size of the RRP, consistent with the idea that morphologically docked SVs constitute the RRP. The low number of docked SVs could be a mechanism to minimize multivesicular release and postsynaptic receptor desensitization.

Release Probability

An action potential releases approximately 200 quanta, suggesting that only 17% of the docked SVs are capable of fusing and releasing glutamate upon local inflow of Ca^{2+}. On the level of AZs, release is initiated in approximately 33% of the AZs. This low probability of release in conjunction with the parallel arrangement of AZs is a mechanism to ascertain reliable synaptic transmission even at high rates because only a fraction of the RRP will be depleted. Release probability could be defined by the composition and state of the molecular complex translating the Ca^{2+} signal into membrane fusion and by the magnitude and temporal profile of the local Ca^{2+} signal.

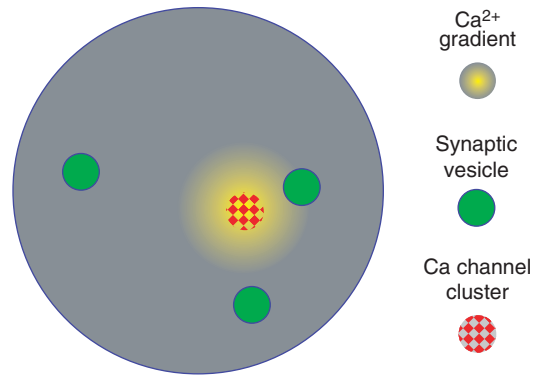

Figure 6 Geometry of Ca channels and release sites. Scheme of a hypothetical model showing an active zone with three release sites and a cluster of Ca channels (~20). The Ca^{2+} gradient (yellow) reaches only one synaptic vesicle (green), triggering its fusion and neurotransmitter release.

The nonlinear dependency of release on the local concentration of Ca^{2+} implies that the distance between the source of Ca^{2+} inflow, Ca channels, and the location of the sensor for Ca^{2+}, the release site, is a highly sensitive mechanism to control release probability. This could work on the level of a single Ca channel differentially coupled to a release site. Alternatively, Ca channels may form a single cluster within an AZ (Figure 6). Then, release sites positioned at different distances from the Ca channel cluster will experience different Ca^{2+} concentrations and, therefore, different release probabilities. For any given constellation of SV number and distances to the Ca channel cluster, a different release probability will be defined for each SV. A Ca channel cluster freely mobile within the AZ could generate a different set of release probabilities at any given moment in time and could account for the low average release probability. This model has not been experimentally verified, highlighting the need for detailed ultrastructural studies of the distribution of Ca channels and release sites within AZs.

Molecular Composition of the AZ

No information is available on the ultrastructural distribution of presynaptic proteins within the AZ of the calyx of Held. Knowing the exact distribution of presynaptic proteins relative to the components of the AZ will be essential to understand neurotransmitter release and the synaptic vesicle cycle in the context of the functional properties of synaptic transmission mediated by this synapse.

Comparison with Other Synaptic Contacts

The composition of individual synaptic contacts of the calyx is very similar to that in other mammalian

central synapses. For example, AZs have areas of 0.03–0.14 μm^2 in other synapses, comparable to 0.06–0.1 μm^2 for AZs of the calyx. In most mammalian synapses examined to date, including the calyx of Held, the size distribution of the AZs has shown a large coefficient of variation, suggesting that the mechanism defining the size of AZs is quite inaccurate. The number of SVs associated with each calycean synaptic contact is similar to that found in other types of synapses. However, the number of docked SVs is up to five times higher in nerve terminals of the cerebellum, hippocampus, and cortex, which have on average approximately 8–10 docked SVs. AZs with many docked SVs have a higher probability of multivesicular release and postsynaptic receptor saturation. In the calyx, fewer docked SVs minimize this and therefore optimize high-frequency transmission.

Conclusion

The multiple structural and functional states of the calyx provide a unique situation to study the structure and function of this synaptic terminal, ranging from molecular to ultrastructural, morphological, functional, and behavioral levels of complexity. Knowing the molecular composition and exact spatial arrangement of AZ components will be necessary to obtain a better understanding of neurotransmitter release and its modulation at this synapse. Electron tomography and analytical electron microscopy provide new avenues of reaching these goals.

See also: Active Zone; Endocytosis and Presynaptic Scaffolds; Exocytosis: Ca^{2+}-Sensitivity; Postsynaptic Density/Architecture at Excitatory Synapses; Presynaptic Development and Active Zones; Presynaptic Development: Functional and Morphological Organization; Presynaptic Endosomes; Presynaptic Regulation by Liprins; Synapsins and Regulation of the Reserve Pool; Synaptic Vesicles; Vesicle Pools.

Further Reading

Borst JG, Helmchen F, and Sakmann B (1995) Pre- and postsynaptic whole-cell recordings in the medial nucleus of the trapezoid body of the rat. *Journal of Physiology (London)* 489: 825–840.

Grothe B (2003) New roles for synaptic inhibition in sound localization. *Nature Reviews Neuroscience* 4(7): 540–550.

Hoffpauir BK, Grimes JL, Mathers PH, and Spirou GA (2006) Synaptogenesis of the calyx of Held: Rapid onset of function and one-to-one morphological innervation. *Journal of Neuroscience* 26(20): 5511–5523.

Landis DM, Hall AK, Weinstein LA, and Reese TS (1988) The organization of cytoplasm at the presynaptic active zone of a central nervous system synapse. *Neuron* 1(3): 201–209.

Lenn NJ and Reese TS (1966) The fine structure of nerve endings in the nucleus of the trapezoid body and the ventral cochlear nucleus. *American Journal of Anatomy* 118(2): 375–389.

Meinrenken CJ, Borst JG, and Sakmann B (2003) Local routes revisited: The space and time dependence of the Ca^{2+} signal for phasic transmitter release at the rat calyx of Held. *Journal of Physiology* 547(Pt 3): 665–689.

Peters A, Palay SL, and Webster H (1991) *The Fine Structure of the Nervous System,* 3rd edn. Oxford: Oxford University Press.

Rizzoli SO and Betz WJ (2005) Synaptic vesicle pools. *Nature Reviews Neuroscience* 6(1): 57–69.

Rollenhagen A and Lubke JH (2006) The morphology of excitatory central synapses: From structure to function. *Cell and Tissue Research* 326(2): 221–237.

Rowland KC, Irby NK, and Spirou GA (2000) Specialized synapse-associated structures within the calyx of Held. *Journal of Neuroscience* 20(24): 9135–9144.

Sätzler K, Sohl LF, Bollmann JH, et al. (2002) Three-dimensional reconstruction of a calyx of Held and its postsynaptic principal neuron in the medial nucleus of the trapezoid body. *Journal of Neuroscience* 22(24): 10567–10579.

Schneggenburger R and Forsythe ID (2006) The calyx of Held. *Cell and Tissue Research* 326(2): 311–337.

Schoch S and Gundelfinger ED (2006) Molecular organization of the presynaptic active zone. *Cell and Tissue Research* 326(2): 379–391.

Taschenberger H and von Gersdorff H (2000) Fine-tuning an auditory synapse for speed and fidelity: Developmental changes in presynaptic waveform, EPSC kinetics, and synaptic plasticity. *Journal of Neuroscience* 20: 9162–9173.

Taschenberger H, Leao RM, Rowland KC, Spirou GA, and von Gersdorff H (2002) Optimizing synaptic architecture and efficiency for high-frequency transmission. *Neuron* 36(6): 1127–1143.

von Gersdorff H and Borst JG (2002) Short-term plasticity at the calyx of Held. *Nature Reviews Neuroscience* 3(1): 53–64.

Wimmer VC, Horstmann H, Groh A, and Kuner T (2006) Donut-like topology of synaptic vesicles with a central cluster of mitochondria wrapped into membrane protrusions: A novel structure–function module of the adult calyx of Held. *Journal of Neuroscience* 26(1): 109–116.

Variant Creutzfeldt–Jakob Disease

D L Ritchie and J W Ironside, University of Edinburgh, Edinburgh, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Variant Creutzfeldt–Jakob disease (vCJD) is a member of the group of human prion diseases and was first identified in the UK in 1996. Prion diseases in humans are rare, accounting for 1–1.5 deaths per million of the population per annum. Human prion diseases are a unique group of neurodegenerative conditions that occur as idiopathic, familial, and acquired disorders (Table 1). All are fatal, and are associated with the accumulation of an abnormal and disease-specific isoform of a normal cellular protein, the prion protein (PrP), in the brain. According to the prion hypothesis, this abnormal form of the protein, designated PrP^{Sc} ('Sc,' from 'scrapie'), is synonymous with the infectious agent, and its presence within the brain is an important diagnostic feature of these diseases.

vCJD is unique in this group of diseases since it is the only known example of prion disease transmitted to man as a zoonosis (Table 1). Initial epidemiological evidence linked the emergence of vCJD to the epizootic of bovine spongiform encephalopathy (BSE) in the UK, a novel prion disease in cattle first reported in 1987. BSE affected many cattle in the UK, with a peak incidence of around 3500 cases per month in 1992–93. At least 180 000 clinical cases were identified and there is evidence to indicate that the total number of cattle infected with BSE in the UK might be as large as 3 million. Infected bovine tissues (brain and spinal cord) were components of certain meat products that were likely to have been consumed by humans in the UK in the 1980s, and although these tissues were banned from human consumption around 1988, there is evidence to suggest that at least some human exposure to BSE through contaminated food continued up to 1996. The presumed route of human exposure to BSE via high-titer infectivity in the food chain is supported by the findings of a case-control study of risk factors for vCJD, which found an increased consumption of products containing mechanically recovered meat in cases of vCJD (although there was evidence of recall bias in participants).

In view of the potential risk to human health from BSE, surveillance of Creutzfeldt–Jakob disease in the UK was reinstituted in 1990. CJD is not a notifiable disease in the UK and the surveillance system is based on referral of suspected cases to the National CJD Surveillance Unit by clinicians and pathologists. The first cases of vCJD were reported by the National Creutzfeldt–Jakob Disease Surveillance Unit in the UK and linked to BSE after detailed studies to exclude any other possible explanation for this novel human disease. There is now convincing evidence that vCJD is caused by human infection with the BSE agent. vCJD occurs predominantly in the UK, the country with the largest BSE epidemic in cattle. The biochemical profile of PrP^{Sc} in the brain in vCJD is very similar to that of the PrP^{Sc} in the brain in cattle and other animals infected with BSE, and experimental infection of macaques with BSE produces a neuropathological profile closely resembling that of vCJD in humans. Furthermore, experimental transmission studies of BSE and vCJD in wild-type and transgenic mice have revealed that the infectious agent in vCJD had very similar biological properties to the BSE agent.

Clinical Features

vCJD predominantly affects young adults, with a mean age at death of 29 years (range 14–74), in distinction to sporadic CJD (sCJD), which occurs most commonly in the seventh decade of life. Possible explanations for this observation include an increased age-related exposure to BSE and/or age related susceptibility. The duration of the clinical illness in vCJD is also distinct from sCJD, with a mean duration of 14 months, in contrast to only 4 months in sCJD.

Psychiatric symptoms including depression, anxiety, and withdrawal are common at presentation in vCJD, and all patients develop psychiatric symptoms as the illness progresses. A subset of patients develops persistent painful sensory symptoms affecting the trunk, limbs, or face, and the unusual combination of these symptoms with a psychiatric disorder may be an early clue to the diagnosis. All patients develop ataxia, usually around 6 months after the disease onset, which progresses along with evidence of increasing cognitive impairment. Involuntary movements of the limbs are very common as the disease progresses but, in contrast to sCJD, dystonia and chorea occur as frequently as myoclonus. In the later stages of the illness there is an accumulation of multifocal neurological deficits, culminating in severe disability and ultimately akinetic mutism. Death usually occurs as a consequence of intercurrent infection, particularly chest infection.

Diagnostic criteria for vCJD have been established (Table 2). These include a number of preconditions

for diagnosis, a series of clinical features, and the results of both diagnostic tests and neuropathological findings. A definite diagnosis of vCJD requires neuropathological validation, but a probable diagnosis can be made based on a combination of clinical features and either a positive magnetic resonance imaging (MRI) brain scan or a positive tonsil biopsy. The classification as 'probable' has a very high sensitivity and specificity for the diagnosis of vCJD, but the classification of 'possible' vCJD is less precise. For this reason, most reported cases of vCJD include only definite and probable cases. The diagnostic criteria have been recently amended following the publication of a report of two cases of vCJD with an electroencephalogram (EEG) recording typical of sCJD in the late stages of the illness, with periodic sharp wave complexes.

Table 1 Classification of human prion diseases

Classification	*Disease*
Idiopathic	Sporadic Creutzfeldt–Jakob disease
	Sporadic fatal insomnia
Familial	Gerstmann–Straussler–Scheinker disease
	Familial Creuzfeldt–Jakob disease
	Fatal familial insomnia
Acquired	
Bovine source	Variant Creutzfeldt–Jakob disease
Human source	Iatrogenic Creutzfeldt–Jakob disease
	Kuru

Clinical Investigations

The MRI scan can be of great value as a noninvasive diagnostic test in vCJD, and shows a symmetrical region of high signal in the posterior thalamus on T_2 or fluid-attenuated inversion recovery (FLAIR) sequences in over 90% of cases. In the clinical context, these MRI findings are highly suggestive of vCJD, but are not specific, since there is a case report of sCJD with the 'typical' scan appearances of vCJD. In vCJD, high signal on MRI may also be seen in the periaqueductal gray matter and in the caudate nucleus and putamen, but for a scan to be judged diagnostic of vCJD the pulvinar region of the posterior thalamus should have the greatest signal intensity (Figure 1).

Other clinical investigations are of less diagnostic value in the investigation of vCJD. The EEG may show nonspecific slow-wave activity, but many recordings have been normal, including those obtained in individuals with clear evidence of cognitive and neurological impairment. Periodic sharp-wave complexes (which are present in most cases of sCJD) are usually absent, but have been identified in the terminal stages of the illness. The cerebrospinal fluid (CSF) 14-3-3 immunoassay is positive in approximately 50% of vCJD cases (in contrast to sporadic CJD, in which the

Table 2 Diagnostic criteria for variant Creutzfeldt–Jakob disease

Category	*Criteria*
1. Definite	
IA and neuropathological confirmation of variant CJD[a]	IA Progressive neuropsychiatric disorder
	IB Duration of illness >6 months
	IC Routine investigations do not suggest an alternative diagnosis
	ID No history of potential iatrogenic exposure
	IE No evidence of a familial form of transmissible spongiform encephalopthies
2. Probable	
(2.1) I and 4/5 of II and IIIA and IIIB	IIA Early psychiatric symptoms[b]
(2.2) I and IVA[e]	IIB Persistent painful sensory symptoms[c]
	IIC Ataxia
	IID Myoclonus or chorea or dystonia
	IIE Dementia
3. Possible	
I and 4/5 of II and IIIA	IIIA EEG does not show the typical appearance of sporadic CJD[d] in the early stages of illness
	IIIB Bilateral pulvinar high signal on magnetic resonance image scan
	IVA Positive tonsil biopsy[e]

[a]Spongiform change and extensive prion protein deposition, with florid plaques throughout the cerebrum and cerebellum.
[b]Depression, anxiety, apathy, withdrawal, and delusions.
[c]This includes both frank pain and/or dysesthesia.
[d]The typical appearance of the electroencephalogram in sporadic CJD consists of generalized triphasic periodic complexes at approximately one per second. These may occasionally be seen in the late stages of variant CJD.
[e]Tonsil biopsy is not recommended routinely, nor in cases with electroencephalographic appearances typical of variant CJD, but may be useful in suspect cases in which the clinical features are compatible with variant CJD and magnetic resonance imaging does not show bilateral pulvinar high signal.

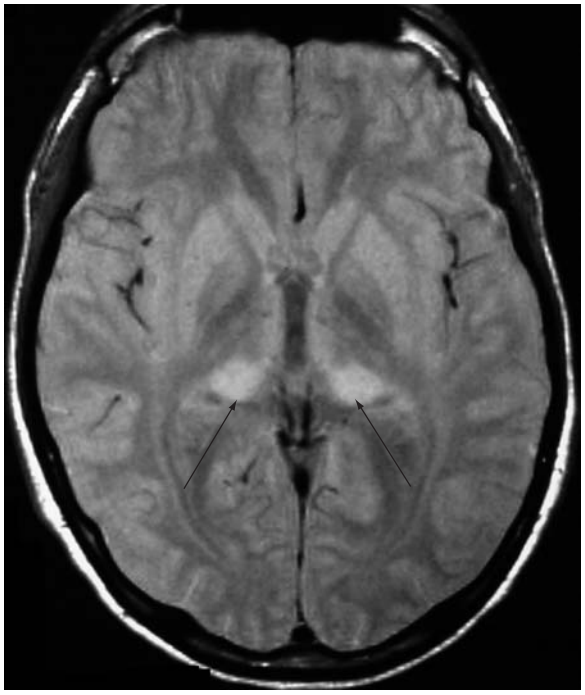

Figure 1 Magnetic resonance image scan in variant CJD. A T_2-weighted scan shows a symmetrical area of increased signal in the posterior thalamus (arrows), the 'pulvinar sign.'

Table 3 *PRNP* codon 129 genotype frequencies in variant and sporadic CJD

Population	*Codon 129 genotype (%)*		
	MM	*MV*	*VV*
Normal population	39	50	11
Sporadic CJD	61	21	18
Variant CJD	100		

M, Methionine; V, valine.

Table 4 Neuropathological criteria for the diagnosis of variant CJD

1. Cerebral and cerebellar cortices:
   Multiple florid plaques in routine stained sections
   Amorphous and pericapillary prion protein accumulation
   Numerous small cluster plaques in prion protein-stained sections
2. Basal ganglia:
   Severe spongiform change
   Perineurinal and periaxonal prion protein accumulation
   Few amyloid plaques
3. Posterior thalamic nuclei:
   Marked neuronal loss and gliosis
   Patchy spongiform change
   Few amyloid plaques
4. Biochemistry:
   Type 2B PrP^{Sc} on Western blot analysis

figure is over 85%). The CSF in vCJD also contains elevated levels of tau, which may be of potential diagnostic value.

Mutations of the human prion protein gene (*PRNP*) have not been identified in vCJD, but all tested patients belong to methionine homozygosity at the polymorphic codon 129, indicating a genetic risk factor for vCJD. This finding is in marked contrast to the *PRNP* codon 129 genotype frequencies in the normal population and in sporadic CJD in the UK (Table 3). Whether human BSE infection may yet occur, perhaps with longer incubation periods, in the other codon 129 genotypes is uncertain.

Neuropathology

The neuropathological features of vCJD are distinctive and allowed early identification of this new form of human prion disease. The features are relatively uniform, allowing the formulation of diagnostic neuropathological criteria (Table 4). The brain in vCJD in most cases is macroscopically normal, but both cerebral and cerebellar cortical atrophy can be identified in cases with prolonged disease duration (over 2 years). vCJD shows all the histological features of a human prion disease in the brain, with spongiform change, neuronal loss, gliosis, and amyloid plaque formation. Florid plaques are a characteristic feature: these are large fibrillary amyloid plaques surrounded by a corona of spongiform change in an otherwise intact neuropil (Figure 2(a)).These are widespread in the cerebral cortex and are generally most numerous in the occipital cortex. Florid plaques also occur in the cerebellar cortex, particularly the molecular layer, but in the granular layer the typical halo of spongiform change is usually absent. Similar fibrillary plaques without surrounding spongiform change are also found in other gray-matter regions in the brain, including the basal ganglia and thalamus. Florid plaques are not specific for vCJD, as similar lesions have been identified in some cases of iatrogenic CJD associated with dura mater grafts. However, the number and distribution of florid plaques in these cases are not as extensive as in vCJD, where their large numbers and widespread distribution are helpful in establishing a diagnosis (Table 4).

Spongiform change in vCJD is most severe in the basal ganglia, particularly the putamen and caudate nucleus, where confluent vacuolation is common (Figure 2(b)). Neuronal loss is most severe in the posterior thalamic nuclei, particularly the pulvinar, although spongiform change in this region is usually mild and in some cases may be absent. This marked neuronal loss is accompanied by severe astrocytosis;

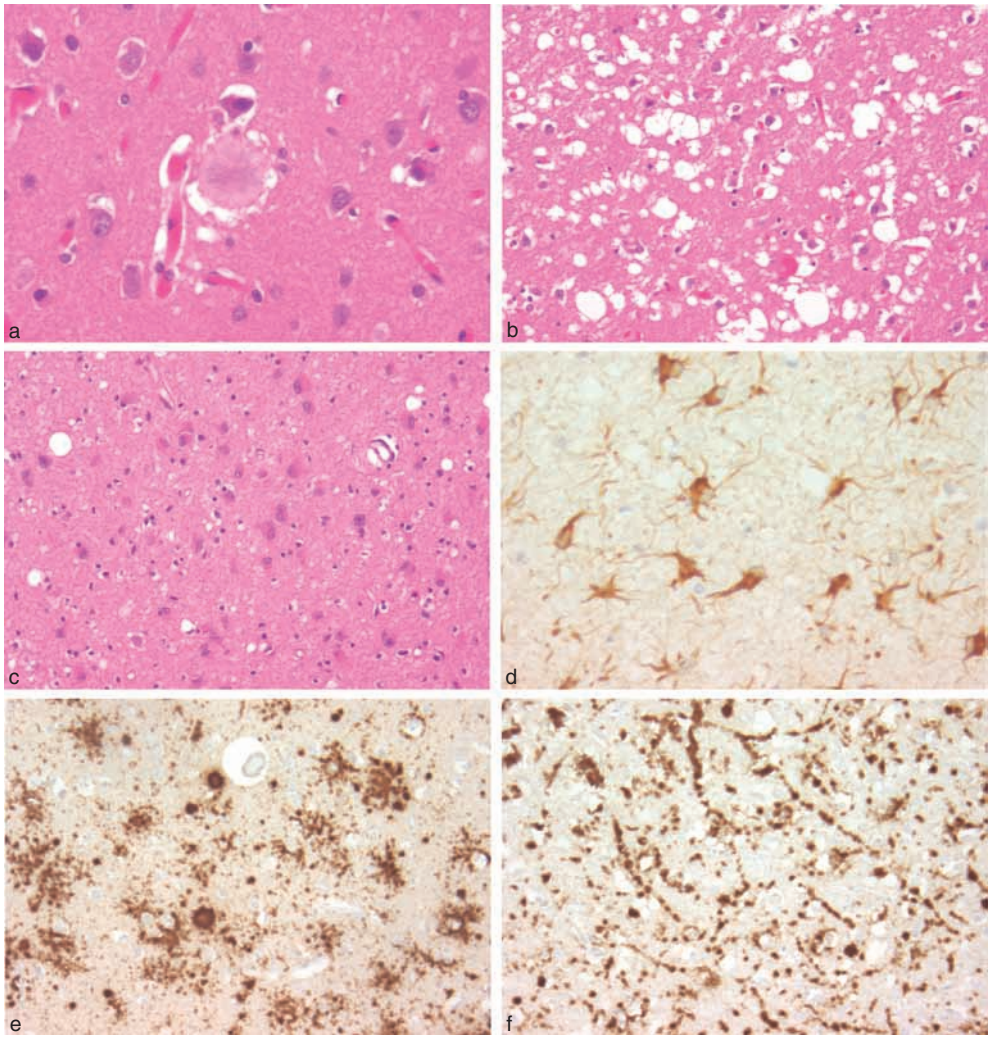

Figure 2 Neuropathology of variant CJD. (a) A typical florid plaque in the cerebral cortex, showing the fibrillary amyloid core surrounded by a characteristic halo of spongiform change (hematoxylin and eosin stain). (b) Spongiform change within the basal ganglia, with characteristic large confluent irregular vacuoles (hematoxylin and eosin stain). (c) Astrocytosis within the pulvinar nucleus in the thalamus (hematoxylin and eosin stain). (d) Immunohistochemistry for glial fibrillary acidic protein labels the reactive astrocytes within the pulvinar (antiglial fibrillary acidic protein). (e) Prion protein immunohistochemistry in the cerebral cortex, showing pericellular deposits of prion protein and the presence of small cluster plaques (KG9 anti-prion protein antibody). (f) Perineuronal and periaxonal linear deposition of prion protein in the basal ganglia (KG9 anti-prion protein antibody).

the distribution correlates with the areas of hyperintense signal within the pulvinar in MRI scans (particularly in T_2-weighted or FLAIR sequences) in patients during the clinical stage of the illness (Figures 2(c) and 2(d)). In patients who have survived for prolonged periods, neuronal loss and gliosis occur at other sites, including the occipital cortex, the cerebellar cortex, and the midbrain. The areas of gliosis in the midbrain have been associated with hyperintense MRI signal changes in some patients with vCJD.

An additional pathological feature of vCJD, as for all prion diseases, is the deposition of PrP^{Sc} in the brain. Different patterns of PrP^{Sc} accumulation in the brain in human prion diseases have been established using immunohistochemical methods and an ever-expanding number of antibodies raised against the prion protein. In vCJD, immunohistochemistry shows a unique and widespread pattern of PrP^{Sc} accumulation in the brain. In addition to the intense PrP^{Sc} positivity within the florid plaques, there are numerous small plaquelike lesions in the cerebral and cerebellar cortex, which are not evident on routine stains (Figure 2(e)). In addition, pericellular accumulation of PrP^{Sc} in an amorphous or 'feathery' distribution is readily detected around small neurons and astrocytes. Perivascular accumulation of PrP^{Sc} is also noted around small capillaries within the cerebral and cerebellar cortices. There is no evidence of an amyloid angiopathy and the mechanism behind this pattern of accumulation is uncertain. In other brain regions

there is a composite plaquelike and synaptic pattern of PrP^{Sc} accumulation. However, in the basal ganglia, particularly in the putamen, there is a perineuronal and periaxonal pattern of PrP^{Sc} accumulation, often in a linear distribution (Figure 2(f)). Brain biopsy (usually on the frontal cortex) has been performed as an investigative procedure in vCJD, mainly on those cases occurring at or just after the time the disease was identified as a unique entity, but is now seldom employed for diagnosis. The neuropathological features in the brain biopsy are similar to those in the frontal cortex in autopsy cases of vCJD, but the extent and severity of the lesions (including both spongiform change and florid plaques) are less pronounced.

Peripheral Pathology

One of the major unique pathological features of vCJD is the presence of both PrP^{Sc} accumulation and infectivity outside the central nervous system, particularly in the lymphoid tissues and the peripheral nervous system. The sensory and autonomic ganglia show positivity within ganglion cells on immunohistochemistry for PrP^{Sc}, and PrP^{Sc} accumulates in lymphoid tissues within germinal centers, particularly in follicular dendritic cells (Figures 3(a) and 3(b)). PrP^{Sc} positivity in germinal centers has also been found within lymphoid follicles in appendix tissues that were surgically removed because of appendicitis up to 2 years prior to the onset of clinical disease. These findings have suggested the possibility of using lymphoid tissue biopsy (specifically tonsil biopsy) as an aid to premortem diagnosis of vCJD, where PrP^{Sc} accumulation can be detected both biochemically and by immunohistochemistry, in symptomatic patients.

Experimental evidence has shown that peripheral involvement in vCJD can occur early on in the disease, perhaps several years before the onset of clinical symptoms. The risk of spread of vCJD through iatrogenic transmission, particularly through routine surgical procedures from individuals incubating vCJD, remains a concern. Bioassay experiments in wild-type mice from vCJD tonsil and spleen have shown that levels of infectivity within the periphery are lower than in the brain, maintaining the demand for more sensitive and specific methods of PrP^{Sc} detection. One such technique, the paraffin-embedded tissue (PET) blot, has been modified specifically for the detection of low levels of PrP^{Sc} in fixed tissues from peripheral tissues and brain biopsy of CJD patients. This technique has proved to be a highly sensitive and specific method for PrP^{Sc}, with investigations on vCJD spleen,

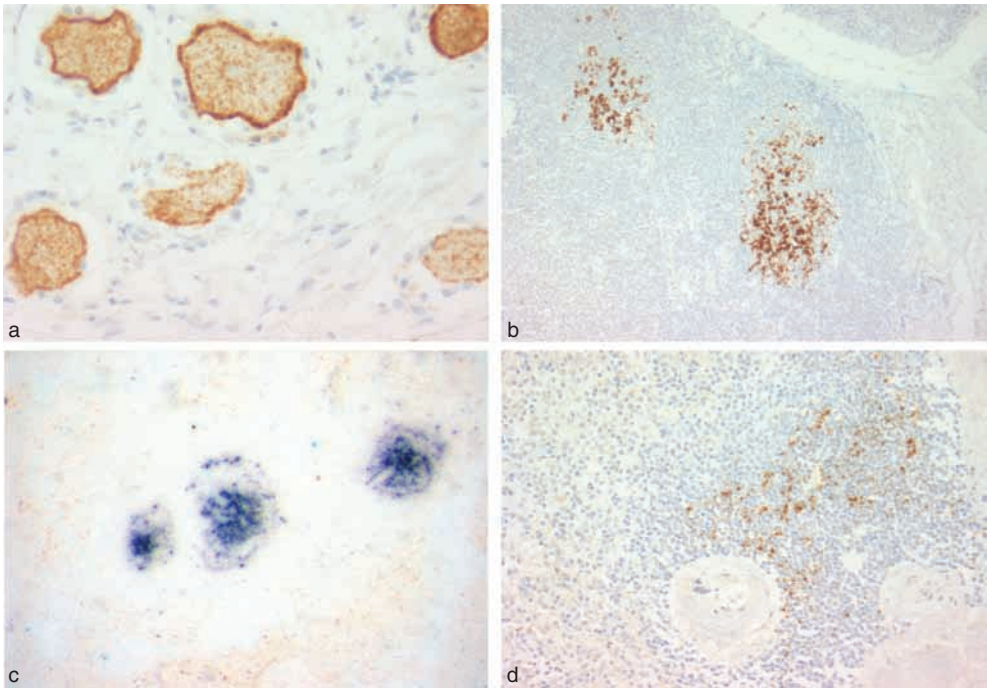

Figure 3 Labeling for the prion protein in peripheral tissues in variant CJD. (a) Prion protein immunohistochemistry, demonstrating intense positivity within the ganglion cells of the dorsal root ganglia (KG9 anti-prion protein antibody). (b) Heavy deposits of prion protein within the germinal centers of the tonsil (KG9 anti-prion protein antibody). (c) Paraffin-embedded tissue blot labeling demonstrating intense prion protein staining (blue) within a number of lymphoid follicles in the spleen (3F4 anti-prion protein antibody). (d) Immunohistochemical detection of prion protein within the spleen of a patient who developed asymptomatic variant CJD infection following a red blood cell transfusion (6H4 anti-prion protein antibody).

tonsil, lymph node, and dorsal root ganglion all demonstrating a more intense labeling pattern when compared to conventional immunohistochemistry (Figure 3(c)). A recent study using a highly sensitive Western blot method has identified PrP^{Sc} within skeletal muscle samples taken at autopsy in a subset of vCJD patients. Establishing the distribution of PrP^{Sc} in the periphery and understanding the peripheral pathogenesis of vCJD may also be crucial in the development of a therapeutic strategy, as the site of agent replication would be a practical target for any prophylactic treatments.

The detection of PrP^{Sc} accumulation within appendix tissues removed prior to the onset of disease in vCJD patients was used as the basis of a retrospective prevalence study of evidence of vCJD infection in anonymized tonsil and appendix tissues in the UK. Of 12 674 cases, 3 were positive, indicating that, in addition to the number of clinical cases of vCJD to date in the UK, there may be a larger number of asymptomatic individuals infected with BSE. Two of the three positive appendix cases from this study had sufficient residual tissue for DNA extraction and analysis, showing that both cases were found to be valine homozygotes at codon 129 in the *PRNP* and raising the possibility of asymptomatic infection in this particular genetic subgroup.

PrP^{Sc} Biochemistry

PrP^{Sc} can be distinguished from the normal cellular form (PrP^{C}) of the protein by proteolytic digestion, which results in the complete degradation of PrP^{C}, leaving only the disease-associated form. Western blot analysis of proteinase K (PK)-digested brain homogenate from CJD cases results in a characteristic banding pattern comprising three different PrP glycoforms (di-, mono-, and nonglycosylated PrP). A molecular classification system for human prion diseases has been proposed based on the level of abundance of these three glycoforms and the differences observed in the mobility of the protease-resistant core of PrP, which presumably relates to differences in the size of the protein after PK digestion (Figure 4). This classification system is used as a diagnostic tool to distinguish not only between vCJD and sCJD but also between the different PrP^{Sc} types of sCJD.

The unique etiology and pathogenesis of vCJD is reflected in a highly characteristic PrP^{Sc} isotype, characterized by a glycoform ratio dominated by the diglycosylated band (Figure 4(a)). The predominance of the diglycosylated band was also found to be characteristic of BSE, and of BSE and vCJD transmitted to other species, leading to the proposition that it represented a BSE/vCJD agent 'glycoform signature'

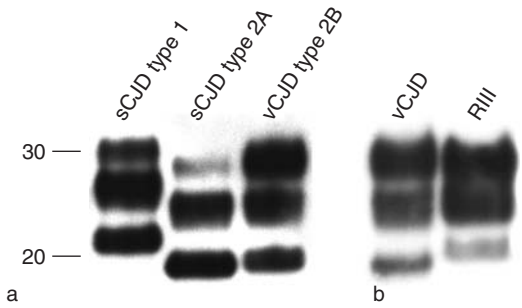

Figure 4 Prion protein isoform (PrP^{Sc}) detection and analysis in variant Creutzfeldt–Jakob disease brain. (a) Western blot analysis of protease-resistant prion protein in the brain from two sporadic Creutzfeldt–Jakob disease (sCJD) patients and a single variant CJD (vCJD) patient. While two PrP^{Sc} types are observed in sCJD cases (type 1 and type 2A), all vCJD cases show a unique banding pattern, referred to as type 2B, which is characterized by predominance of the diglycosylated fragment (3F4 anti-PrP antibody). (b) Western blot analysis showing conservation of vCJD PrP^{Sc} glycoforms following transmission into wild-type RIII mice, although the molecular sizes differ (6H4 anti-PrP antibody).

(Figure 4(b)). The unglycosylated band of PrP^{Sc} in vCJD has a molecular mass of around 19 kDa. N Terminus mapping of PrP^{Sc} from vCJD brain by Edman degradation is centered on serine-97 and is therefore not different in this respect from some other types of human prion disease, including sCJD and kuru. However, the glycoform ratio of the PrP^{Sc} in vCJD is a remarkably consistent feature, irrespective of the brain region sampled.

PrP^{Sc} is readily and consistently detectable in vCJD tonsil, spleen, lymph node, appendix, and gut-associated lymphoid tissue. PrP^{Sc} can also be found at particularly high levels in the vCJD retina, optic nerve, and trigeminal ganglion, but dental tissue specimens appear negative, as do nonnervous tissues in the eye. Both the celiac and stellate ganglia have been shown to be positive for PrP^{Sc} in vCJD, but peripheral nerves have been reported to be negative. Whether the presence of PrP^{Sc} in all these tissues is associated with infectivity will require experimental verification, but experimental transmission from vCJD spleen and tonsil has been reported in inbred mice.

Animal Models for the Study of vCJD

As described earlier, experimental transmission of vCJD to inbred strains of mice was essential as part of a strain typing study to establish the close relationship between the BSE agent and the agent responsible for vCJD. Similar findings were observed following transmission of vCJD and BSE to bovine transgenic mice. Inbred strains of mice have also been used for the detection of infectivity in peripheral tissues in vCJD, such as tonsil and spleen, in which levels of

infectivity were found to be significantly lower than those in the brain. vCJD and BSE have also been transmitted to macaques, in which the neuropathological findings following BSE transmission are remarkably similar to those in vCJD in humans. The macaque model has also been used to investigate the route of spread of the BSE agent in the body following oral transmission, and an attempt has been made to estimate the magnitude of the species barrier between bovines and humans using this model. BSE and vCJD have also been transmitted to a range of transgenic mice that express human PrP with differing codon 129 polymorphisms. The results from these studies indicate that all *PRNP* codon 129 polymorphisms may be susceptible to BSE, but infection does not always appear to result in clinical diseases, raising the possibility of an asymptomatic carrier state under certain conditions.

Transmission of vCJD by Blood Transfusion

In the UK, the presence of infectivity within blood is an important issue in human prion diseases. Four cases of probable transmission of vCJD infectivity following transfusion of nonleukodepleted red blood cells from asymptomatic donors have been identified in the UK. Three of these were accompanied by typical clinical manifestations of vCJD: the incubation period in the first case between transfusion and disease onset was 6.5 years, while vCJD presented clinically in the second and third cases around 8 years following transfusion. The fourth recipient did not develop any clinical symptoms and signs of vCJD, and died 5 years after transfusion. However, PrP accumulation was detected following autopsy, biochemically and by immunohistochemistry in the spleen and a cervical lymph node, but not in the central nervous system, indicating an asymptomatic or preclinical infection (Figure 3(d)). Analysis of the *PRNP* in this asymptomatic case showed heterozygosity at codon 129, indicating that this genotype is also susceptible to vCJD infection. It is possible that the *PRNP* codon 129 genotype may have had an influence on the disease phenotype, since PrP accumulation was not detected in oral and gut-associated lymphoid tissues. However, the absence of involvement of tonsil and gut-associated lymphoid tissues suggests that the likely intravenous route of infection may have contributed to the relatively restricted peripheral disease pathogenesis.

The identification of these cases of secondary transmission of vCJD after blood transfusion indicates that blood is infectious even in the asymptomatic preclinical phase of vCJD. The development of a blood-detection method is therefore of prime concern, not only as a diagnostic tool for CJD but also as an important screening test for verifying the safety of donated blood and blood products.

Current Developments

To date, 165 cases of vCJD have been identified in the UK since the first cases were identified in 1995–96. The number of annual deaths has declined since the year 2000, and the number of likely future cases remains uncertain; however, it now seems unlikely that a major epidemic of vCJD will occur. vCJD has been identified in a number of other countries (Table 5); these cases are attributed according to the country of normal residence at the time of disease onset, which may not necessarily reflect the geographic region in which BSE exposure occurred. BSE exposure of non-UK residents may be related to residence in or travel to the UK, to exported bovine products from a country with BSE, or to exposure to indigenous BSE. In some vCJD cases outside UK there is a history of extended residence in the UK during 1980–96, the period in which there may have been significant human exposure to BSE, although the clinical illness developed many years later while living in another country. In the majority of non-UK cases, the source of BSE infection is uncertain, as there was no history of travel to the UK.

Although it is clear that very large numbers of BSE-infected cattle were likely to have entered the human food chain between 1980 and 1996, it is not certain how many humans have been infected with BSE over this period. It is clear from studies of kuru and iatrogenic CJD that incubation period in human prion diseases can extend to over 40 years, with substantial influence on the incubation period by the *PRNP* codon 129 polymorphism, as most prolonged incubation periods occur in heterozygotes. A recent study in transgenic mice expressing human PrP indicated that the transmission periods following experimental BSE

Table 5 Variant CJD cases worldwide since 1996

Country	*Number of cases*
UK	165
France	22
Ireland	4
USA	3
The Netherlands	2
Canada	1
Italy	1
Japan	1
Portugal	2
Saudi Arabia	1
Spain	1

transmissions were substantially influenced by the *PRNP* codon 129 genotype. These findings indicate that cases of BSE infection in humans with different *PRNP* codon 129 genotypes may present at different time points as compared to the current cases of vCJD in codon 129 methionine homozygotes, due to genetically determined differences in the incubation period. Continuing surveillance of vCJD is required to address this possibility, by monitoring future BSE-related infections (perhaps with a clinicopathological phenotype different from that seen in vCJD) in individuals who are valine homozygotes or heterozygotes at *PRNP* codon 129 in the UK and overseas.

Acknowledgments

The National CJD Surveillance Unit in the UK is funded by the Department of Health and the Scottish Executive.

See also: Infectious Agents in Neurodegenerative Disease; Prion Diseases; Prion Transport.

Further Reading

Bishop MT, Hart P, Aitchison L, et al. (2006) Predicting susceptibility and incubation time of human-to-human transmission of vCJD. *Lancet Neurology* 5: 393–398.

Bruce E, Will RG, Ironside JW, et al. (1997) Transmissions to mice indicate that "new variant" CJD is caused by the BSE agent. *Nature* 389: 498–501.

Bruce ME, McConnell I, Will RG, et al. (2001) Detection of variant Creutzfeldt-Jakob disease infectivity in extraneural tissues. *Lancet* 358: 208–209.

Collinge J, Sidle KCL, Meads J, et al. (1996) Molecular analysis of prion strain variation and the aetiology of 'new variant' CJD. *Nature* 383: 685–690.

Head MW, Ritchie D, Smith N, et al. (2004) Peripheral tissue involvement in sporadic, iatrogenic and variant Creutzfeldt–Jakob disease: An immunohistochemical, quantitative and biochemical study. *American Journal of Pathology* 164: 143–153.

Hilton DA, Ghani AC, Conyers L, et al. (2004) Prevalence of lymphoreticular prion protein accumulation in UK tissue samples. *Journal of Pathology* 203: 733–739.

Ironside JW, Bishop MT, Connolly K, et al. (2006) Variant Creutzfeldt–Jakob disease: Prion protein genotype analysis of positive appendix tissue samples from a retrospective prevalence study. *British Medical Journal* 332: 1186–1188.

Ironside JW, Head MW, Bell JE, et al. (2000) Laboratory diagnosis of variant Creutzfeldt–Jakob disease. *Histopathology* 37: 1–9.

Lasmezas CI, Comoy E, Hawkins S, et al. (2005) Risk of oral infection with bovine spongiform encephalopathy agent in primates. *Lancet* 365: 781–783.

Llewlyn CA, Hewitt PE, Knight RS, et al. (2004) Possible transmission of variant Creutzfeldt–Jakob disease by blood transfusion. *Lancet* 363: 417–421.

Peden AH, Head MW, Ritchie DL, et al. (2004) Preclinical vCJD after blood transfusion in a *PRNP* codon 129 heterozygous patient. *Lancet* 364: 477–479.

Scott MR, Will R, Ironside J, et al. (1999) Compelling transgenetic evidence for transmission of bovine spongiform encephalopathy prions to humans. *Proceedings of the National Academy of Sciences of the United States of America* 96: 15137–15142.

Smith PG and Bradley R (2003) Bovine spongiform encephalopathy (BSE) and its epidemiology. *British Medical Bulletin* 66: 185–198.

Will RG, Ironside JW, Zeidler M, et al. (1996) A new variant of Creutzfeldt–Jakob disease in the UK. *Lancet* 347: 921–925.

Will RG, Zeidler M, Stewart GE, et al. (2000) Diagnosis of new variant Creutzfeldt–Jakob disease. *Annals of Neurology* 47: 575–582.

Relevant Websites

http://www.bseinquiry.gov.uk – BSE Inquiry: The inquiry into BSE and variant CJD in the United Kingdom.

http://www.cdc.gov – Centers for Disease Control and Prevention, United States.

http://www.cjdfoundation.org – Creutzfeldt–Jakob Disease Foundation, Inc., Akron Ohio, USA.

http://www.defra.gov.uk – Department for Environment, Food and Rural Affairs (DEFRA), London, UK.

http://www.dh.gov.uk – Department of Health (National Health Service, UK).

http://www.food.gov.uk – Food Standards Agency, United Kingdom.

http://www.hpa.org.uk – Health Protection Agency, United Kingdom.

http://www.iah.bbsrc.ac.uk – Institute for Animal Health, United Kingdom.

http://www.blood.co.uk – National Blood Service for England and North Wales.

http://www.cjd.ed.ac.uk – National Creutzfeldt–Jakob Disease Surveillance Unit (University of Edinburgh).

http://www.fda.gov – U.S. Food and Drug Administration (U.S. Department of Health and Human Services).

http://www.who.int – World Health Organization (headquarters, Geneva, Switzerland).

http://www.oie.int – World Organisation for Animal Health (headquarters, Paris, France).

Vascular Issues in Neurodegeneration and Injury

L Longhi, E R Zanier, V Conte, and N Stocchetti, University of Milan, Milan, Italy
T K McIntosh, Media NeuroConsultants, Inc., Media, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Traumatic brain injury (TBI) is clinically classified as either focal or diffuse. Focal TBI most often involves contusion and/or lacerations accompanied by intraparenchymal hematomas. Diffuse brain swelling, vascular alterations, ischemic brain injury, and diffuse axonal injury (DAI) are frequent components of the diffuse TBI spectrum. The sequelae of TBI may be further divided into (1) primary injury (immediate, nonreversible, and dependent upon the physical factors involved in the injury) and (2) secondary or delayed injury (progressing from hours to months following injury and likely to be partially reversible). This secondary injury cascade is complex and poorly understood and includes the breakdown of the blood–brain barrier (BBB), edema formation, vasospasm, ionic dysregulation, impairment and/or uncoupling of energy metabolism, changes in intracranial pressure (ICP) and/or cerebral perfusion pressure (CPP), inflammation, expression of both pathogenic and protective genes and proteins, and activation and/or release of autodestructive factors. Since the pathological sequelae of TBI are multifactorial in nature, an improved understanding of the secondary injury cascade offers a unique opportunity for the development of targeted therapeutic interventions to attenuate cellular damage and improve functional recovery. One area of intense interest and research concerns the posttraumatic alterations in cerebrovascular function and their role in mediating delayed cell death and/or neurologic dysfunction. Here we review the current state of knowledge concerning the importance of vascular changes following TBI.

Pathology of Vascular Disruption Following TBI

Following TBI, mechanical deformation because of compression, shear, and tension can cause a disruption of the blood vessel wall. Depending on the anatomic position of the injured vessel, hemorrhage can occur in the epidural, subdural, subarachnoid, and/or intraparenchymal compartments. Intracranial hematomas are an extremely important lesion that can complicate 25–45% of severe TBI cases, 3–12% of moderate TBI cases, and approximately 0.2% of mild TBI cases.

Epidural Hematoma

Epidural hematoma (EDH) is caused by the vascular disruption of meningeal arteries, meningeal veins, and/or dural sinuses before they enter into the space between the brain and the dura mater. In the majority of cases, EDH occurs as a consequence of skull fracture and stripping of the endosteum and can therefore be considered a complication of fracture and not related to head acceleration (see Table 1).

In 50% of human TBIs, the most common location of an EDH is the squamosal portion of the temporal bone. In this area, the middle meningeal artery is partially embedded in the inner table of the skull and is therefore relatively fixed to the bone and more vulnerable to injury. As the hematoma enlarges, it strips the dura from the skull and forms a circumscribed ovoid mass with progressive compression of brain. Picard and colleagues have shown that craniotomy for prompt evacuation of an isolated acute EDH is one of the most cost-effective surgical procedures with excellent results in term of outcome. Postmortem analysis and experimental studies have shown that neuropathological damage associated with EDH is the result of massive ischemic injury associated with brain compression and low CPP.

Subdural Hematoma

Subdural hematoma (SDH) occurs in the space located between the meningeal layer of the dura and the underlying arachnoid and is often associated with a hemorrhagic contusion, intracerebral hematoma breaking through the arachnoid, the rupture of one or more bridging veins, or the laceration of a cortical artery or vein (see Table 1). The cerebral convexity in the parietal region is the most common location of SDHs. Most SDHs are caused by motor vehicle-related accidents (MVAs), falls, and assaults. During falls, angular acceleration is often high and of short duration. Bridging veins (those portions of the superficial cerebral veins that cross the subdural space to reach venous sinuses) are relatively straight as they cross the subarachnoid space and are therefore highly vulnerable to sudden angular acceleration forces. In elderly patients because of brain atrophy, the bridging veins can be stretched, leading to an increased incidence of SDHs. Infants are also susceptible to SDHs because their bridging veins are thin-walled. In younger adults, MVAs are responsible for 56% of SDHs, while only

Table 1 Pathology of vascular disruption following traumatic brain injury (TBI)

Pathology	*TBI severity*	*Incidence (%)*	*Reference*[a]
Extradural hematoma	Mild/moderate/severe	3	Bullock et al. (2006)
	Severe	9	Bullock et al. (2006)
Subdural hematoma	Mild/moderate/severe	11	Servadei et al. (2000)
	Severe	12–29	Bullock et al. (2006)
Subarachnoid hemorrhage	Mild/moderate/severe	40	Servadei et al. (2002)
Intraparenchymal hematoma	Mild/moderate/severe	15	McCormick and Post (1996)
Delayed traumatic intracerebral hematoma	Moderate/severe	3–7.5	Bullock et al. (2006)

[a]From Bullock MR, Chesnut R, Ghajar J, et al. (2006) Surgical management of acute epidural hematomas. *Neurosurgery* 58: S7–S15; Bullock MR, Chesnut R, Ghajar J, et al. (2006) Surgical management of acute subdural hematomas. *Neurosurgery* 58: S16–S24; Bullock MR, Chesnut R, Ghajar J, et al. (2006) Surgical management of traumatic parenchymal lesions. *Neurosurgery* 58: S25–S46; Servadei F, Nasi MT, and Giuliani G (2000) CT prognostic factors in acute subdural haematomas: The value of the 'worst' CT scan. *British Journal of Neurosurgery* 14: 110–116; Servadei F, Murray G, and Teasdale G (2002) Traumatic subarachnoid hemorrhage: Demographic and clinical study of 750 patients from the European Brain Injury Consortium Survey of Head Injuries. *Neurosurgery* 50: 261–269; McCormick PC and Post KD (1996) Trigeminal neurinomas. In: *Neurosurgery*, 2nd edn., pp. 1545–1552. New York: McGraw-Hill.

12% are caused by falls. The overall mortality rate following acute traumatic SDH ranges from 40% to 75%. If the mass lesion is evacuated early, the prognosis depends primarily on the severity of the underlying injury. Mass effect with hemispheric compression and ischemia is an important mechanism of brain damage following a SDH.

Subarachnoid Hemorrhage

Traumatic subarachnoid hemorrhage (SAH) is caused by the bleeding of surface blood vessels into the subarachnoid space or by the rupture of bridging veins or traumatic aneurysm (see Table 1). Typical locations for SAH are the interpeduncular cistern and the Sylvian fissure. In a retrospective study conducted on 141 patients, Chieregato and colleagues reported that the best predictive factors for poor outcome were subarachnoid blood at admission, assessed by computed tomography (CT) scan, together with the Glasgow Coma Scale (GCS) score. However, in a multicenter prospective study conducted on 1005 patients with moderate or severe TBI, Servadei and colleagues observed that patients with SAH exhibit worse CT scan features, such as subdural hematomas and parenchymal damage, with higher risk for evolving contusion, suggesting that SAH may be an indicator of initial injury severity. Furthermore, progression of contusions usually occurs in locations in which cortical SAH is observed, suggesting that contusions are part of a similar process and that subarachnoid blood is probably an early sign of cortical microbleeding.

Intraparenchymal Hematoma

Traumatic intraparenchymal hematoma (IPH) is caused by the bleeding of an intracerebral or subarachnoid vessel. Most IPHs develop as a complication of a contusion; although they can also be absent from the first CT scan, they may develop later. Delayed traumatic intracerebral hematoma (DTICH) usually occurs in an area of radiologically normal brain in patients with otherwise pathological CT scans. However, DTICH can also occur in areas of contusion on initial CT scan or may appear following resection of surgical brain lesions or resuscitation. The occurrence of traumatic IPHs is five times more common in patients older than 45 years, compared to younger patients, suggesting that age-related changes in the brain parenchyma (such as atrophy) and vasculature may facilitate the development of IPHs following TBI (see Table 1). The most common locations for traumatic IPHs include the frontal and temporal regions, where the brain can impact the roughened inner surfaces of the skull (i.e., orbital roof or sphenoid ridge).

A study by Chang and colleagues showed that over 60% of lesions remain unchanged within the first 3 days of injury, and nearly 40% progress to larger lesions, with 15% becoming larger than 5 cm^3. A majority of studies on traumatic IPH are observational and comparisons of outcome between surgical and nonsurgical groups are difficult. However, current evidence indicates that an IPH that is accompanied by progressive neurological dysfunction, refractory intracranial hypertension, or radiological signs of mass effect may be associated with a poor outcome if left untreated.

Alterations in Cerebral Blood Flow Following TBI

Brain glycogen reserves are extremely limited, and since the brain has no significant storage capacity for glucose, it is highly dependent on blood flow. Inadequate substrate delivery and cerebral blood flow (CBF) are therefore among the most important

factors contributing to mortality and morbidity following TBI. A consistent finding reported in the experimental TBI literature is that CBF is significantly reduced acutely following impact. The degree and time window of this reduction are related to the injury model, the species, and the methodology used to measure CBF.

Ischemia

Measurement of CBF following TBI can identify hypoperfusion. However, because of metabolic derangement, true ischemia in these patients can only be defined by a CBF that is inappropriately low for tissue metabolism.

The classical definition of ischemia has the following implications:

- Reduction of CBF to less than 18 ml per 100 g tissue min^{-1} (coupled with increase in cerebral oxygen extraction fraction/high arteriojugular oxygen content difference ($AJDO_2$) values).
- Reduction of ATP (otherwise termed 'energy failure').
- Increase in anaerobic glucose metabolism, thereby producing lactate.

TBI in rodents reduces cortical CBF distal to the impact by 50%, while CBF in the injury site is reduced by about 95%. Additional experimental studies have reported reduction in ATP content in traumatized cortical regions and elevated lactate levels after severe TBI. In the clinical setting, ischemic lesions, most commonly seen in the hippocampus and basal ganglia, have been detected in the majority (90%) of fatal TBI. In nonfatal TBI, profound reductions of cerebral CBF (CBF $<$18 ml per 100 g tissue min^{-1}) have also been shown in more than 30% of patients during the early (4–12 h) period following TBI. However, high $AJDO_2$ values (above 9 ml/100 ml) have been identified in only a minority of these patients, indicating a coupled CBF and metabolic depression. Since marked heterogeneity of CBF and metabolic patterns in the traumatically injured brain have been demonstrated, focal ischemia could remain undetected by global measurements such as $AJDO_2$. Recent ^{15}O-positron emission tomography studies exploring regional pathophysiology measuring O_2 extraction fraction have shown that within 24 h a small but significant volume of brain tissue is at risk of ischemic damage. Other clinical studies, however, have reported that ischemic levels of CBF are only observed in irreversibly damaged tissue. Therefore the importance of ischemia in delayed neuronal injury following nonfatal TBI remains controversial. In an assessment of the similarities and differences between ischemia and TBI, Bramlett and colleagues have shown that mechanisms that cause cells to die are quite similar, but different for dysfunctional cells that survive the initial insult. The mechanisms underlying cellular vulnerability in ischemia and TBI may therefore be distinct.

Uncoupling of Cerebral Blood Flow and Metabolism

Under normal conditions, CBF is coupled to metabolism whereby cortical activation is followed by CBF increase. Following TBI, several studies have shown that this CBF/metabolism coupling can be lost when acute glucose metabolism increase is associated with a reduction in CBF (**Figure 1**, upper panel). Similarly, human brain PET studies have shown that around cerebral contusions and beneath intracranial hematomas there is an area of increased glucose metabolism together with decreased CBF (**Figure 1**, lower panel). Therefore, the ability of the injured brain to respond adequately to subsequent perturbations that may exacerbate the imbalance between O_2/nutrient delivery and consumption is extremely limited. The concept of the vulnerability of the traumatized brain to secondary insults has been investigated by Jenkins and colleagues, who were the first to show that although neither mild TBI nor brief (6 min) bilateral carotid artery occlusion resulted in detectable hippocampal morphological damage/death, the same severity of incomplete ischemia resulted in profound CA1 hippocampal neuronal death when induced 1 or 24 h following a mild TBI. Therefore in the presence of secondary sublethal challenges (i.e., increased ICP, reduced CPP, fever, seizures), injured cells may undergo irreversible damage.

Posttraumatic Vasospasm

Posttraumatic vasospasm occurs in approximately 25–40% of patients following severe TBI and is associated with the presence of traumatic SAH. In a study comparing 99 patients with posttraumatic SAH to 114 patients with aneurismal SAH, a clear difference was found in distribution, time course of CBF, and neurological changes in the two conditions. Furthermore, the distribution of the low-density areas on the CT scans did not correlate with vascular territories in traumatic SAH patients, suggesting that vasospasm following TBI does not contribute to secondary deterioration. Therefore the importance of monitoring TBI patients for vasospasm is not currently clear.

Molecular Mediators of Vascular Damage after TBI

In addition to CBF changes (*vide supra*), post-TBI changes include morphologic damage to endothelial cells and perivascular cells, increased BBB permeability,

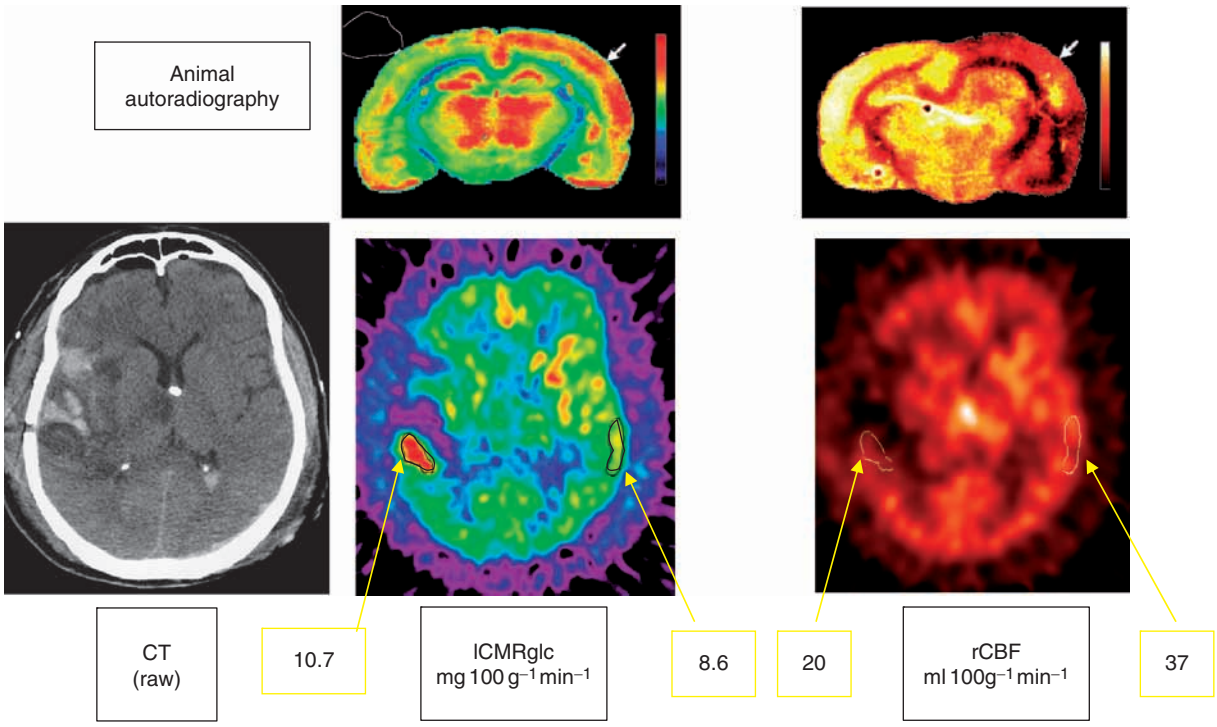

Figure 1 Mismatch of local cerebral metabolic rate for glucose (lCMRglc) and regional cerebral blood flow (rCBF) following traumatic brain injury. (Top) 1 h following experimental traumatic brain injury, the rat brain exhibits evidence of local increase in lCMRglc, utilizing [^{14}C]deoxyglucose autoradiography (left), associated with a reduction of local cerebral blood flow, utilizing [^{14}C]iodoantipyrine autoradiography (right). (Bottom) In brain scans of a person with severe traumatic brain injury, confirmation of a similar mismatch of lCMRglc and rCBF in the tissue beneath a contusion. The computed tomography (CT) scan shows the contusion and the pericontusional tissue (left panel). The pericontusional rim (yellow arrows, middle panel) exhibits an increase of lCMRglc measured with ^{18}F-labeled deoxyglucose positron emission tomography (PET scanning), associated with a reduction of rCBF, measured with ^{15}O-H_2O PET techniques (yellow arrows, right panel). Reproduced by permission of David A Houda, UCLA Brain Injury Research Center.

reduction or abolition of vasodilatory and vasoconstriction responses to changes in mean arterial pressure (MAP), changes in arterial tension of O_2 and CO_2 and blood viscosity as well as in cerebral metabolic activity. It is likely that some aspects of cerebrovascular dysfunction following TBI contribute to the increased sensitivity of the injured brain to secondary hypoxia and hypotension and to the increased mortality and morbidity that occurs due to secondary insults after TBI in humans.

The cerebral circulation is influenced by a variety of vasodilators and vasoconstrictors, and the observed reduction in CBF after TBI may be due to an imbalance in the levels/activity of vasoconstrictors and vasodilators and/or to local microthrombosis.

Calcium

An increase in brain calcium (Ca^{2+}) flux has been documented following several experimental models of TBI. Ca^{2+} entry into neurons may occur through membrane-associated receptors and voltage-sensitive Ca^{2+} channels (VSCCs). Antagonism of these channels using various Ca^{2+} channel blockers has been shown to be associated with attenuation of regional CBF reductions, brain edema, and motor and cognitive deficits following experimental TBI. However, when nimodipine, which blocks L-type VSCCs, was administered 24 h postinjury to severely head-injured patients, no significant clinical benefit was observed, except in a subgroup of patients with SAHs.

Prostanoid, Eicosanoid, and Reactive Oxygen Species

Posttraumatic elevation of intracellular free Ca^{2+} may precipitate an attack on the cellular membrane by activating phospholipases A_2 (PLA_2) and C (PLC), resulting in the release of free fatty acids (FFAs) and diacylglycerol (DAG). Increases in brain FFAs can lead to secondary damage through BBB breakdown and formation of cerebral edema, and, indirectly, via the generation of prostaglandins, thromboxanes, and leukotrienes (cyclooxygenase and lipooxygenase products from arachidonic acid), which may affect CBF and BBB permeability. Figure 2 shows the pathways activated from arachidonic acid and the effect on cerebral circulation of the different products.

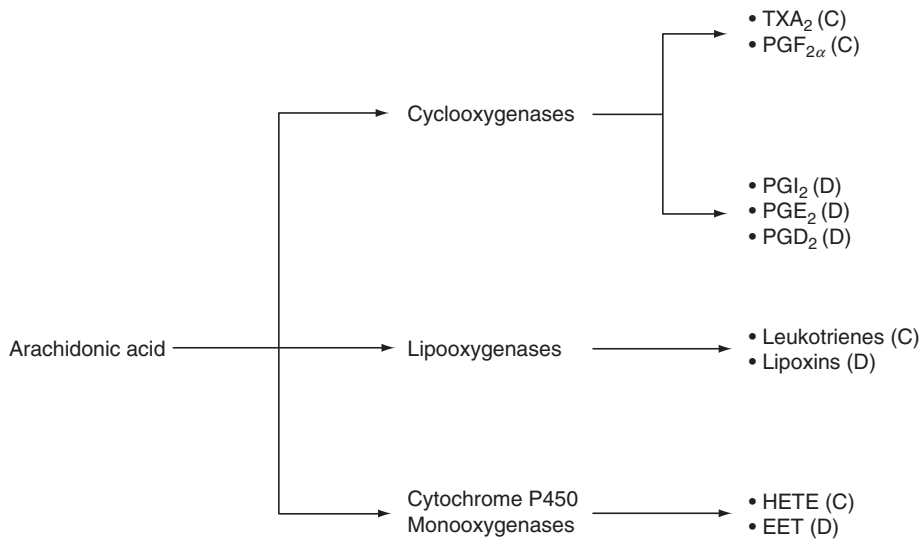

Figure 2 Pathways of eicosanoid formation and effects on cerebral circulation. TXA_2, thromboxane A_2; $PGF_{2\alpha}$, PGI_2, PGE_2, PGD_2, prostaglandins; EET, epoxyeicosatrienoic acid; HETE, hydroxyeicosatetraneoic acid; C, vasoconstrictor; D, vasodilator.

Cyclooxygenases 1 (constitutive isoform) and 2 (inducible) have been shown to increase following experimental and clinical TBI as early as 3 h postinjury, together with increases in plasma and tissue levels of prostaglandins ($PGF_{2\alpha}$, PGE_2), prostacyclin, and thromboxane, with an imbalance toward vasocontrictors. Administration of cyclooxygenase (COX) inhibitors has resulted in mixed results, and additional studies are warranted to better understand the effects of COX inhibition on cerebrovasculature and neuroinflammation following TBI. Prostacyclin has also been shown to reduce cortical cell death and to improve compromised perfusion caused by posttraumatic vasoconstriction and microthrombosis following experimental TBI.

In addition, during the process of prostaglandin synthesis from arachidonic acid, reactive oxygen species (ROS) are generated by the mitochondria due to Ca^{2+}-mediated disorganization of the inner mitochondrial membrane, with subsequent disruption of the electron transport chain. Reactive oxygen species formation leads to peroxidative destruction of the cell membrane, proteins, and DNA. Under physiological conditions, ROS are scavenged by endogenous antioxidants, including superoxide dismutase (SOD), glutathione peroxidase, and catalase. Since the amount of ROS production exceeds the scavenging capacity of the cells, increasing the endogenous antioxidative mechanisms may have therapeutic value in TBI.

Adenosine

Adenosine is a major regulator of the cerebral circulation known to cause vasodilation acting on two receptors located on the vascular endothelium and smooth muscle. An increase in cerebral metabolic activity and/or a mismatch between energy production and oxygen consumption results in a rise in adenosine concentration, which in turn can dilate the cerebral arterioles and increase local CBF. Following TBI in rats, interstitial adenosine has been shown to increase in the contusional/pericontusional area and, in humans, in the cerebrospinal fluid (CSF) of patients following TBI. Furthermore, secondary episodes of global brain hypoxia further increase CSF adenosine, suggesting a role during periods of secondary insults following TBI. Agonists of adenosine have been shown to attenuate cerebral hypoperfusion and posttraumatic behavioral deficits following experimental TBI in rats, suggesting that the neuroprotection obtained by adenosine receptors pathways might be due to an increase in CBF.

Nitric Oxide

Nitric oxide (NO) is a vasodilator that has been suggested to play a role in damage, recovery, or both, after TBI. Three different systems are reportedly involved in NO synthesis: endothelial, neuronal, and inducible NO synthase (eNOS, nNOS, and iNOS). lIt has been observed that cerebral tissue NO levels increase transiently and then decrease within 15–30 min following experimental TBI.

When expression/levels of NO synthases are evaluated following TBI, it has been shown that brain tissue nNOS activity and NO levels increase markedly but briefly after TBI, and then decrease below preinjury levels for hours to days. Inducible NOS activity in the brain increases hours to days after TBI and remains elevated up to a week after injury.

In the cerebral vasculature, eNOS immunoreactivity and protein expression increase after TBI. Since CBF might be linked to NO levels, several studies have evaluated the hypothesis that increasing NO levels (or NO synthases activity) results in increased CBF and attenuated damage following TBI. The NOS substrate L-arginine improves CBF and attenuates histological damage after experimental TBI. Nonselective blockade of NOS has also been shown to be either detrimental or without influence on outcome following TBI in rats.

To clarify the beneficial or detrimental effects of iNOS-related NO production, studies employing genetically engineered mice and/or pharmacological inhibition have been performed: mice deficient in iNOS show increased cognitive deficits 15–18 days after TBI, suggesting a protective role of iNOS. This protective role of iNOS was confirmed in a rat model of TBI in which administration of iNOS inhibitors aminoguanidine and L-*N*-iminoethyl-lysine resulted in reduced learning capacity and greater hippocampal cell loss at 3 weeks after injury. It appears possible, therefore, that iNOS activity may exert detrimental effects in the acute postinjury period but can show beneficial effects in the chronic phase. Further work needs to be directed at clarifying the exact role of the different isoforms of NOS to draw a more comprehensive picture of the role of NO following TBI.

Endothelin-1

Endothelin (ET) is a 21-amino-acid peptide with potent vasoconstrictive activity. The ET family comprises ET-1, ET-2, and ET-3. Endothelin is produced in endothelial cells, as well as in several tissues, including the brain, where ET-1 has been found in neurons, glial cells, and choroid plexus cells. Synthesis of ET-1 is stimulated by thrombin, cytokines, transforming growth factor-β, vasopressin, angiotensin II, and oxyhemoglobin, while it is inhibited by NO. Endothelins acts on two receptors, ET_A and ET_B, which have been observed in vascular smooth muscle cells (leading to vasoconstriction through an increase in Ca^{2+}) and endothelial cells (leading to vasodilation through the release of vasodilatory COX products and NO). Following experimental TBI, an increase in ET-1 mRNA expression and protein level has been observed in the brain and CSF, and this is associated with reduced CPP. A relationship has been observed between ET and NO (vasoconstriction and vasodilation) and inhibition of iNOS using antisense iNOS oligodeoxynucleotides following TBI in rodents; this interaction results in reduced synthesis of iNOS and a five- to sixfold increase in ET-1 concentration, both of which are associated with a marked reduction in CBF. In contrast, ET-1 antagonists have been associated with restored vasodilatory capacity of the cerebrovascular system following TBI, suggesting that ET-1 might be a target for therapeutic intervention.

Neuropeptide Y

Neuropeptide Y is a transmitter of the cerebrovascular sympathetic nerves known to induce vasoconstriction and reduction of CBF. It acts on a family of six different receptors localized in the smooth muscle cells of pial arteries and arterioles. In cerebral arteries, the vasoconstriction mediated by neuropeptide Y is mediated by changes in smooth muscle membrane potential and Ca^{2+} increase (either via increased influx or increased release from intracellular stores); this vasoconstriction can be blocked by Ca^{2+} channel blockers such as nimodipine, nifedipine, and verapamil. Following TBI in rodents, concentrations of neuropeptide Y are increased in the pericontusional cortex. In addition, neuropeptide Y-immunoreactive fibers are observed in the injured cortex and ipsilateral hippocampus, suggesting that an increase in cerebral neuropeptide Y concentrations after brain injury may be involved in the pathogenesis of reduced CBF.

Endothelium-Dependent Hyperpolarizing Factor

Recently, an additional endothelial-derived vasodilatory pathway that is different from NO and/or cyclooxygenase-derived products has been identified; this pathway is mediated by endothelium-derived hyperpolarizing factor (EDHF), which acts to hyperpolarize the vascular smooth muscle and activate potassium (K^+) channels on the vascular smooth muscle. Possible candidates of EDHF include epoxides, anandamide, K^+ ion, and an electrical coupling mediated by myoendothelial gap junctions. Following TBI in rodents, EDHF pathways are activated and the resultant vasodilation can be inhibited by charybdotoxin, an inhibitor of large-conductance Ca^{2+}-sensitive K^+ channels, suggestive of a compensatory role of EDHF during conditions characterized by compromised NO. Further work will determine whether the observed phenomenon of EDHF potentiation might be a protective mechanism initiated following cerebrovascular injury.

Genetic Influences/Susceptibility to Vascular Changes after TBI

Apolipoprotein E

Lipoproteins (lipid transporters) are macromolecules that contain lipids and proteins, known as apolipoproteins, including apolipoprotein E (ApoE), which plays a major role in the metabolism of triglyceride

lipoproteins (chylomicrons, chylomicrons remnants, very low-density and intermediate-density lipoproteins) and high-density lipoproteins (HDLs), and in the redistribution of lipids locally among cells. The brain is the second most abundant site of ApoE mRNA synthesis, which occurs primarily in astrocytes, but also in neurons and activated microglia. ApoE is also synthesized in macrophages, where it appears to play a role in modulating cholesterol accumulation. ApoE is the only lipoprotein present in the CSF and may function in redistributing lipids among CNS cells for normal lipid homeostasis, repairing injured neurons, maintaining synaptodendritic connections, and scavenging toxins. Polymorphisms of ApoE can occur as a result of mutation, and can be defined as a genetic locus where the most common allele occurs with a frequency of less than 100% in the general population. The most common polymorphisms are single-nucleotide polymorphisms (SNPs), resulting from a single base variation. The ApoE gene on chromosome 19 has three common alleles ($\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$), producing corresponding isoforms of the protein: ApoE2 (Cys112, Cys158), ApoE3 (Arg112, Cys158), and ApoE4 (Arg112, Arg158). These three isoforms differ in their abilities to accomplish critical tasks in the maintenance and repair of neurons.

Contusions in head-injured patients with ApoE4 might be more severe as a result of relatively deficient clotting mechanisms. Prolonged clotting times were also identified in ApoE4 carriers after stroke, providing further evidence that ApoE genotype is of relevance to the coagulation cascade. Animal models have provided further information about ApoE mechanisms and the response of the brain to injury. In normal rodents, ApoE immunoreactivity is found in glia, mostly astrocytes, and is not observed in neurons. After ischemic brain injury, the distribution changes and ApoE is increased in reactive astrocytes and neuropils, but later also in degenerating neuronal processes. This increase in ApoE levels has been interpreted as an attempt of the brain to protect itself. ApoE-deficient mice were found to have a greater extent of neuronal damage than did wild-type mice after TBI and ischemia, which was ameliorated by continuous intracerebral infusion of ApoE. However, different genotypes show specific abilities in the response to brain injury. Vascular factors such as thrombin, released during TBI, may also play an important role in the upregulation of ApoE levels. The increase in ApoE levels after hemorrhagic stroke or TBI may initiate a cascade of events leading to amyloid protein-lipoprotein (Aβ) deposition, neuroinflammation, and cognitive deficits. Increased neuroinflammation in ApoE4 carriers may alter the neurological outcome after TBI (see Table 2). Although ApoE4-associated vascular pathology may influence the outcome in survivors, it seems unlikely to significantly increase the probability of a fatal outcome after TBI, since ApoE4 carriers are equally represented in fatal cases (35%) and in nonfatal

Table 2 Outcome of traumatic brain injury (TBI)

Outcome measure	*Main results of TBI pathology*	*Reference*[a]
Recovery of consciousness	$\varepsilon 4$ found more frequently in patients who did not recover consciousness at 1 year	Sorbi et al. (1995)
Glasgow Outcome Scale	$\varepsilon 4$ associated with unfavorable outcome at 6 months	Teasdale et al. (1997)
Functional and cognitive assessment	$\varepsilon 4$ associated with poorer clinical outcome at 6–8 months	Friedman et al. (1999)
Functional independence measures	$\varepsilon 4$ associated with poorer rehabilitation at 6 months	Lichtman et al. (2000)
Learning/fluency test	$\varepsilon 4$ had poorer memory outcome at 6 months	Crawford et al. (2002)
Glasgow Outcome Scale	$\varepsilon 4$ associated with unfavorable outcome at 6 months	Chiang et al. (2003)
Neuropsychological tests	$\varepsilon 4$ bearers had worse postinjury performance	Sundström et al. (2004)
Hematoma volume/Glasgow Outcome Scale	$\varepsilon 4$ associated with larger hematoma volume and poorer outcome at 6 months	Liaquat et al. (2002)
Glasgow Outcome Scale $\pm$ posttraumatic seizures	$\varepsilon 4$ associated with increased risk of seizures, independent of outcome at 6 months	Diaz-Arrastia et al. (2003)
Confusion index/ischemic damage	$\varepsilon 4$ associated with greater incidence of moderate/severe contusions and severe ischemic damage at autopsy	Smith et al. (2006)
Cerebral amyloid angiopathy	$\varepsilon 4$ associated with cerebral amyloid angiopathy at autopsy	Leclercq et al. (2005)

[a]See full entries in 'Further reading' section.

cases (33%). Cerebral amyloid angiopathy in fatal cases of TBI is considerably more likely to occur in those with a genetic predisposition conferred by possession of an ApoE ε4 allele. Clinical studies of TBI have shown that possession of the ε4 allele of the apolipoprotein E gene is associated with a relatively poor outcome. Teasdale and colleagues demonstrated that 57% of ApoE ε4 carriers had an unfavorable outcome following TBI (defined as dead, in a vegetative state, or with severe disability), compared with 27% of noncarriers of ApoE ε4. Friedman and colleagues reported that only 3.7% of patients with the ε4 allele had a good recovery, compared with 31% of patients without the allele. It has been suggested that ApoE ε4 promotes production of Aβ aggregates, cytoskeleton alterations, and oxidative damage, while also impairing the ability of the brain to repair and regenerate (see Figure 3).

Summary

Traumatic brain damage is a result of an immediate mechanical disruption of brain tissue, (including blood vessels), and delayed physiological, molecular, and cellular events/cascades that are potentially amenable to postinjury therapeutic intervention because of their progressive nature, from minutes to months after the initial trauma. Vascular pathology and alterations in CBF are major contributors to mortality and morbidity following TBI. Traumatic brain injury results in CBF reduction both in the experimental and clinical setting, and the reduced CBF in patients has been shown to render the brain vulnerable to secondary insults and has been associated with an unfavorable neurological outcome. The mechanisms responsible for the observed reductions in regional CBF (rCBF) following TBI have not been fully elucidated. During normal physiological conditions, CBF and metabolism are coupled, in contrast to alterations following TBI, when an uncoupling between CBF (reduced) and glucose metabolism (increased) is often observed. In addition, TBI causes an impairment in cerebral autoregulation, whereby patients are not able to tolerate hypoperfusion and remain vulnerable to hypotension and hypoxia. Since TBI is often followed by an upregulation and release of several vasoactive mediators which act in a complex network, it has been hypothesized that posttraumatic alterations in CBF might be the result of an imbalance between molecules with vasoconstrictive and vasodilatory properties. Alterations in Ca^{2+}, nitric oxide, endothelin, prostanoids, adenosine, neuropeptide Y, reactive oxygen species, and endothelium-derived hyperpolarizing factor have been observed following TBI and have been linked with the observed CBF alterations. In addition, patient genotype appears

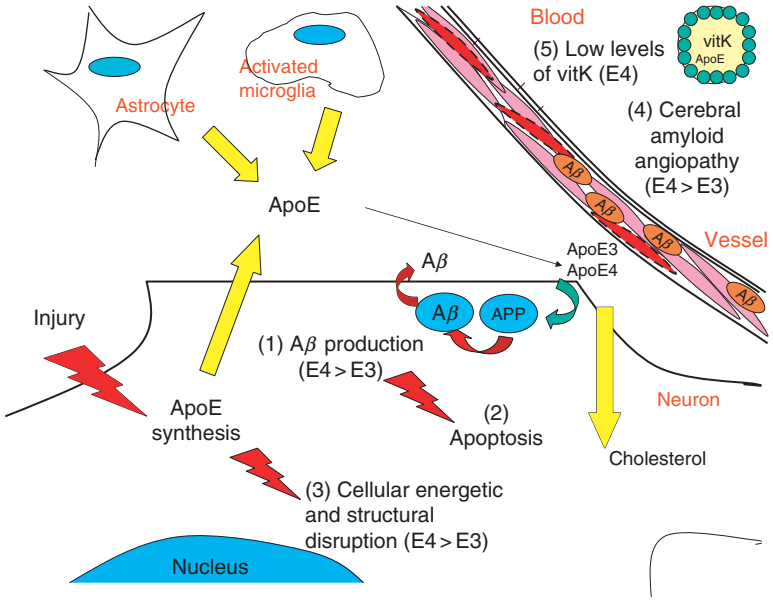

Figure 3 Roles of ApoE in lipid redistribution among cells in the central nervous system and ApoE isoform-specific differences in neuropathology. ApoE is synthesized by astrocytes, activated microglia, and neurons. Potential detrimental roles for ApoE4 include (1) enhanced amyloid β-protein (Aβ) production, (2) potentiation of Aβ_{1-42}-induced lysosomal leakage and apoptosis, (3) enhanced neuron-specific proteolysis, resulting in translocation of neurotoxic ApoE4 fragments in the cytosol, where they are associated with cytoskeletal disruption and mitochondrial dysfunction (APP, amyloid precursor protein), (4) increased deposition of amyloid in cerebral cortical and leptomeningeal blood vessels, with consequent thickening of the walls and degeneration of the smooth muscle cells in the media, and (5) decreased levels of plasma vitamin K (vitK), which is a fat-soluble vitamin involved in clotting mechanisms.

capable of influencing both the vulnerability of the brain to injury and the posttraumatic damaging/protective response. Although laboratory investigations in which a single pathway has been modified have shown promising results, further work is warranted to better understand the molecular pathways associated with posttraumatic alterations in CBF/metabolism and to draw a comprehensive picture that will permit the successful translation of therapeutic strategies from animals models to the clinical setting.

See also: Blood-Brain Barrier and Neurovascular Mechanisms of Neurodegeneration and Injury; Brain Injury: Functional Recovery After; Brain Injury: Magnetic Resonance Studies of Metabolic Aspects; Brain Trauma; Cerebrovascular Disease; Glia Control of Blood Flow.

Further Reading

Bramlett HM and Dietrich WD (2004) Pathophysiology of cerebral ischemia and brain trauma: Similarities and differences. *Journal of Cerebral Blood Flow and Metabolism* 24: 133–150.

Bullock MR, Chesnut R, Ghajar J, et al. (2006) Surgical management of acute epidural hematomas. *Neurosurgery* 58: S7–S15.

Bullock MR, Chesnut R, Ghajar J, et al. (2006) Surgical management of acute subdural hematomas. *Neurosurgery* 58: S16–S24.

Bullock MR, Chesnut R, Ghajar J, et al. (2006) Surgical management of traumatic parenchymal lesions. *Neurosurgery* 58: S25–S46.

Chiang MF, Chang JG, Hu CJ, et al. (2003) Association between apolipoprotein E genotype and outcome of traumatic brain injury. *Acta Neurochirurgica* 145: 649–654.

Crawford FC, Vanderploeg RD, Freeman MJ, et al. (2002) APOE genotype influences acquisition and recall following traumatic brain injury. *Neurology* 58: 1115–1118.

DeWitt DS and Prough DS (2003) Traumatic cerebral vascular injury: The effects of concussive brain injury on the cerebral vasculature. *Journal of Neurotrauma* 20: 795–825.

Diaz-Arrastia R, Gong Y, Fair S, et al. (2003) Increased risk of late posttraumatic seizures associated with inheritance of APOE. *Archives of Neurology* 60: 818–822.

Friedman G, Froom P, Sazbon L, et al. (1999) Apolipoprotein E-ε4 genotype predicts a poor outcome in survivors of traumatic brain injury. *Neurology* 52: 244–248.

Golding EM (2002) Sequelae following traumatic brain injury. The cerebrovascular perspective. *Brain Research Reviews* 38: 377–388.

Hurley SD, Olschowka JA, and O'Banion MK (2002) Cyclooxygenase inhibition as a strategy to ameliorate brain injury. *Journal of Neurotrauma* 19: 1–15.

Leclercq PD, Murray LS, Smith C, et al. (2005) Cerebral amyloid angiopathy in traumatic brain injury: Association with apolipoprotein E genotype. *Journal of Neurology, Neurosurgery & Psychiatry* 74: 827a–828a.

Liaquat I, Dunn LT, Nicoll JA, et al. (2002) effect of apolipoprotein E genotype on hematoma volume after trauma. *Journal of Neurosurgery* 96: 90–96.

Lichtman S, Seliger G, Tycki B, et al. (2000) Apolipoprotein E and functional recovery from brain injury. *Neurology* 54: 2082–2088.

Mahley RW, Weisgraber KH, and Huang Y (2006) Apolipoprotein E4: A causative factor and therapeutic target in neuropathology, including Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* 103(15): 5644–5651.

McCormick PC and Post KD (1996) Trigeminal neurinomas. In: *Neurosurgery*, 2nd edn., pp. 1545–1552. New York: McGraw-Hill.

Nortje J and Menon DK (2004) Traumatic brain injury: Physiology, mechanisms, and outcome. *Current Opinion in Neurology* 17: 711–718.

Smith C, Graham DI, Murray LS, et al. (2006) Association of APOE ε4 and cerebrovascular pathology in traumatic brain injury. *Journal of Neurology, Neurosurgery & Psychiatry* 77: 363–366.

Servadei F, Nasi MT, and Giuliani G (2000) CT prognostic factors in acute subdural haematomas: The value of the 'worst' CT scan. *British Journal of Neurosurgery* 14: 110–116.

Servadei F, Murray G, and Teasdale G (2002) Traumatic subarachnoid hemorrhage: Demographic and clinical study of 750 patients from the European Brain Injury Consortium Survey of Head Injuries. *Neurosurgery* 50: 261–269.

Sorbi S, Nacimas B, Piacentini S, et al. (1995) ApoE as a prognostic factor for post-traumatic coma. *Nature Medicine* 1: 135–137.

Sundström A, Marklund P, Nilsson L-G, et al. (2004) APOE influences on neuropsychological function after mild head injury: Within-person comparisons. *Neurology* 62: 1963–1966.

Teasdale GM, Nicoll JAR, Murray G, et al. (1997) Association of apolipoprotein E polymorphism with outcome after head injury. Early report. *The Lancet* 350: 1069–1071.

Waters RJ and Nicoll JA (2005) Genetic influences on outcome following acute neurological insults. *Current Opinion in Critical Care* 11: 105–110.

Vasoactive Intestinal Peptide and Pituitary Adenylate Cyclase Activating Peptide Receptors

D E Brenneman, Johnson & Johnson Pharmaceutical Research & Development, Spring House, PA, USA
J M Hill, National Institutes of Health, Bethesda, MD, USA
I Goze, Tel Aviv University, Tel Aviv, Israel; and Allon Therapeutics, Vancouver, BC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Structure

Vasoactive intestinal polypeptide (VIP) and pituitary adenylate cyclase-activating polypeptide (PACAP) belong to a family of peptides shown in Table 1. VIP is processed from a precursor protein that also contains a structurally related and biologically active peptide – peptide histidine methionine amide (PHM). VIP and PHM are co-released and are believed to have distinct functions. PACAP is derived from a gene product distinct from VIP; however, these peptide-encoding genes may have evolved by exon duplication coupled to gene duplication. PACAP is also proteolytically processed from a precursor protein. This precursor protein contains an additional peptide, PACAP-related peptide (PRP), which exhibits sequence homology to PACAP. PRP is found localized in tissues with PACAP, where it displays similar biological effects, albeit at much lower potency. Structurally, both VIP and PACAP possess two helices separated by a central disordered region. Subtle conformation differences between the peptides contribute to the selectivity of the peptides for their receptors.

Gene Regulation

VIP gene expression is regulated by cyclic adenosine monophosphate (cAMP) through a cAMP response element (CRE). However, recent studies showed additional regulation of VIP gene expression through the cooperation of the signal transducers and activators of transcription (STAT) proteins, SMAD, STAT, and activating protein-1 sites. Both ciliary neurotrophic factor and transforming growth factor-β induce the transcription of the VIP gene through a 180 bp cytokine response element in the VIP promotor. Neurospecificity for VIP expression is associated with a restrictive element 1 in the VIP gene that modulates transcription in collaboration with a tissue specifier element.

Similar to VIP, the PACAP gene can be activated through the CRE, to which the CRE-binding protein predominantly binds. The activation of PACAP gene promoter is synergistically enhanced by Ca^{2+}. In addition, the half-life of PACAP is prolonged by Ca^{2+} influx into neurons, indicating an activity-dependent stabilization of particular mRNA species in neurons. Thus, the activity-dependent gene expression is coordinately controlled by Ca^{2+} and cAMP signals not only at the transcriptional level but also at the posttranscriptional level for the cumulative mRNA expression in neurons. Both VIP and PACAP are regulated by neuronal activity. Axotomy has been shown to upregulate VIP expression. Furthermore, depolarization of sympathetic neurons stimulates PACAP and VIP peptide levels and increases the release of the peptides with differing magnitudes and temporal profiles. Both VIP and PACAP are involved in the control of the biological clock through the visual system. Dopamine regulates PACAP mRNA expression in intrinsically photosensitive retinal ganglion cells, and the synthesis and release of dopamine in the retina is regulated by the rod and cone photoreceptors via retinal circuitry.

Mechanism of Action

Receptors

Three receptor subtypes recognize VIP and PACAP: VPAC1, VPAC2, and PAC1. PAC1 binds PACAP with much higher affinity (~1000-fold) than it binds VIP. The VPAC receptors exhibit similar affinity for PACAP and VIP. To date, no VIP-specific receptor has been reported. As a result of alternative RNA splicing, PAC1 has ten variants in human. These alternative forms occur in the third intracellular loop, the N-terminal domain, and the transmembrane domain. The splice variants in the intracellular loop include one or two cassettes that are involved in second-messenger coupling. A variant with a 21-amino-acid deletion of the N-terminal domain influences the selectivity of the two forms of PACAP and the relative potencies of the two peptides in stimulating phospholipase C. Activation of the PAC1 receptor with the transmembrane 4 variant results in Ca^{2+} influx through L-type voltage-sensitive Ca^{2+} channels. Thus, differential receptor expression can contribute to the diversity of responses to PACAP. The VPAC1 receptor has 50% amino acid sequence identity to PAC1, but the VPAC1 has similar affinity for VIP and PACAP. VPAC2, which was cloned initially from rat pituitary, has a distinct pharmacology and localization from those of the VPAC1 and PAC1.

Table 1 Amino acid sequences of the VIP/PACAP family of peptides

Species[a]	*Peptide*[b]	*Amino acid sequence*
h/b/d/p/r	VIP	HSDAVFTDNYTRLRKQMAVKKYLNSLILN
g/p	VIP	HSDALFTDTYTRLRKQMAMKKYLNSVLN
c	VIP	HSDAVFTDNYSRFRKQMAVKKYLNSVLT
h	PHM	HADGVFTSDFSKLLGQLSAKKYLESLM
r	PHI	HADGVFTSDYSRLLGQLSAKKYLESLI
h	PHV(1–42)	HADGVFTSDFSKLLGQLSAKKYLESLMGKRVSSNISEDPVPV
r,h	PACAP(1–27)	HSDGIFTDSYSRYRKQMAVKKYLAAVL
r,h	PACAP(1–38)	HSDGIFTDSYSRYRKQMAVKKYLAAVLGKRYKQRVKNK
h	PRP	DVAHGILNQAYRKVLDQLSAGKHLQSLVA
h	GRF(1–44)	YADAIFTNSYRKVLGQLSARKLLQDIMSRQQGESNQERGARARL
h	Glucagon	HSQGTFTSDYSKYLDSRRAQDFVQWLMNT
h	GLP-1(7–36)	HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR
h	GLP-2	HADGSFSDEMNTILDNLAARDFINWLIQTKITDR
h	Secretin	HSDGTFTSELSRLREGARLQRLLQGLV
h	GIP	YAEGTFISDYSIAMDKIHQQDFVNWLLAQKGKKNDWKHNITQ
hs	Helodermin	HSDAIFTEEYSKLLAKLALQKYLASILGSRTSPPP
hh	Helospectin	HSDATFTAEYSKLLAKLALQKYLESILGSSTSPRPPS

[a]b, bovine; d, dog; h, human; hh, heloderma horidum; hs, heloderma suspectum; p, pig; r, rat.
[b]GIP, glucose-dependent insulinotropic peptide; GLP, glucagon-like peptide; GRF, growth hormone releasing factor; PACAP, pituitary adenylate cyclase-activating polypeptide; PHI, peptide histidine isoleucine amide; PHM, peptide histidine methionine amide; PHV, peptide histidine valine; PRP, PACAP-related peptide; VIP, vasoactive intestinal polypeptide.

Cell Signaling

The actions of VIP and PACAP are primarily mediated through the production of cAMP, although other cell-signaling mechanisms exist including Ca^{2+} regulation, cyclic guanosine monophosphate (cGMP) / nitric oxide (NO) generation, and activation of phospholipase D and phospholipase C. With regard to cAMP, the activation of adenylate cyclase is achieved through the coupling of the α subunit of G-protein in many tissues. Recent studies indicated that, in addition to nanomolar amounts of PACAP increasing the cAMP levels, subpicomolar concentrations of PACAP can result in a stimulation of the extracellular-signal-regulated kinase-type mitogen-activated protein kinase. Calcium effects associated with these peptides involve multiple pathways that are tissue specific and which can involve either the inhibition of Ca^{2+} currents or mobilization of intracellular Ca^{2+} stores. The complexity of VIP and PACAP signaling is due in part to the multiple receptors and their splice variants combined with receptor activity-modifying proteins (RAMPs) interactions. At least some of the tissue and brain-region response diversity is due to these important interactors that can determine the functional signaling output.

Localization

VIP and PACAP are best known as neurotransmitters or modulating neuropeptides and are present in neurons throughout both the central nervous system (CNS) and peripheral nervous system. However, VIP is also an important immune modulator and can be found in mast cells, eosinophils, B and T lymphocytes, and thymocytes. PACAP has well-recognized endocrine functions and, in addition to its presence in the pituitary gland, is enriched in endocrine and germ cells. PACAP is also expressed in lymphoid tissues, including the spleen, thymus, and macrophages. In the CNS, VIP and PACAP overlap in some hypothalamic nuclei, such as the suprachiasmatic nucleus. However, each peptide has its own unique distribution pattern, with PACAP being the more widely distributed throughout the CNS and being found in highest concentrations in hypothalamic nuclei, where it regulates pituitary functions and is produced, passes through the pituitary stalk, and is stored in the posterior pituitary. In the hypothalamus, PACAP is found co-localized with oxytocin in rats and vasopressin in humans. VIP is most abundant in the cerebral cortex, suprachiasmatic nucleus, amygdala, and striatum, but is also present in the olfactory bulbs, midbrain, hindbrain, and spinal cord. In the peripheral nervous system, VIP and PACAP are found in neurons and ganglia in organ systems throughout the body and are frequently co-localized in the same neurons. However, whereas VIP-containing neurons are mainly found in the intestines, heart, lungs, genital organs, urinary system, and several endocrine organs, nearly all organs contain detectable levels of PACAP, within the highest concentration in the testes followed by the adrenal glands. The receptors, VPAC-1, VPAC-2, and PAC1 are found throughout the CNS and peripheral nervous system, and like the peptides

that bind to them, the receptors overlap but have differing patterns of distribution. VPAC-1 receptor is the least common receptor of the brain and is most dense in the hippocampus, pineal gland, olfactory bulb, and cingulate cortex. However, it is widespread in the periphery, where it is found in the pituitary, adrenal gland pancreas, lung, gastrointestial system, and lymphoid tissues. In the brain, VPAC-2 receptors are widely distributed and are abundant in the hypothalamus, hippocampus, amygdala, olfactory bulbs, and several brain stem nuclei. In the periphery, VPAC2 receptors are extensive and enriched in the pituitary, adrenal gland, pancreas, testis, ovary, and spleen. PAC-1 receptors are the most common and widespread of the PACAP receptors within the brain and can be found in most brain regions. They are particularly plentiful in the olfactory bulbs, cortex, hypothalamus, and brain stem. In the periphery, they have a more restricted distribution and are most dense in peripheral ganglia and anterior pituitary. All three receptors are found on glial cells within the brain. The selective actions of VIP and PACAP can be explained, in part, by the regional differences in the abundance of the peptides coupled with the regional variation in the density and distribution of the receptors and the reduced affinity of VIP for the PAC-1 receptor.

Physiology

Developmental Role

VIP, PACAP, and the three known receptors with which they interact appear during the early postimplantation period of embryogenesis in rodents. The receptors are localized to proliferative zones of the neural tube, indicating important role(s) of these peptides in development. Similar to other neurotrophic molecules, VIP and PACAP exhibit dual functions in the regulation of CNS proliferation. They have important actions during development, and they also exhibit neuroprotective properties following injury. During early CNS development, VIP is known to stimulate mitogenesis, differentiation, neurite outgrowth, and neuronal survival. Whereas VIP promotion of neuronal differentiation and neurite outgrowth occurs in the micromolar to nanomolar concentration range, PACAP has been shown to be more potent in increasing these activities, suggesting that PACAP actions are mediated through the PAC-1 receptor, to which VIP binds with only low affinity. PACAP further promotes the transition from proliferation to differentiation in the embryonic brain by stimulating both withdrawal from the cell cycle and morphological and biochemical differentiation. In addition to the possibly direct actions of VIP on developing cells already discussed, VIP has an important regulatory role in embryonic growth and development, perhaps through the VIP-induced release of other trophic factors in the embryo. VIP treatment of whole-cultured day-9.5 mouse embryos causes a dramatic, dose-dependent increase in growth. In 4 h, VIP-treated embryos grow up to twice the number of somites as untreated embryos (Figure 1). Similar treatment with PACAP was shown to have the opposite effect and to inhibit embryonic growth at high concentrations (Figure 1). Treatment of pregnant mice with an antagonist to VIP during embryogenesis results in fetal microcephaly and growth restriction. Co-treatment with VIP prevents the microcephaly; co-treatment with PACAP does not (Figure 2). These experiments indicate that VIP is necessary for the coordinated of growth and development of the embryo and that PACAP does not share this action. Postnatal examination of offspring of females treated with a VIP antagonist or a PACAP antagonist during embryogenesis reveals that the blockage of VIP action resulted in delays in the appearance of several developmental milestones (Table 2) and a long-lasting retardation in growth retardation. Similar treatment with a PACAP antagonist had no effect on the achievement of developmental milestones (Table 2) or on growth. These experiments illustrate differences in the actions of VIP and PACAP during embryogenesis that may be linked to the differential expression of VIP receptors and/or the peptides in the embryonic neuroepithelium. Although VIP and PACAP produced similar

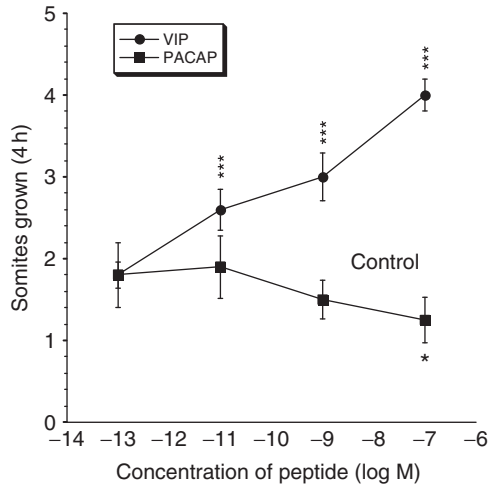

Figure 1 Treatment of whole-cultured embryonic-day-9.5 mouse embryos with a dose range of peptides demonstrates that vasoactive intestinal polypeptide (VIP) stimulates the growth of embryos in a dose-related manner as measured by somite growth. Pituitary adenylate cyclase-activating polypeptide (PACAP) has no effect at low concentrations, but inhibits growth at high concentrations. Reprinted from Hill JM, Lee SJ, Dibbern DA, Fridkin M, Gozes I, and Brenneman DE (1999) Pharmacologically distinct vasoactive intestinal peptide binding sites: CNS localization and role in embryonic growth. *Neuroscience* 93: 783–791.

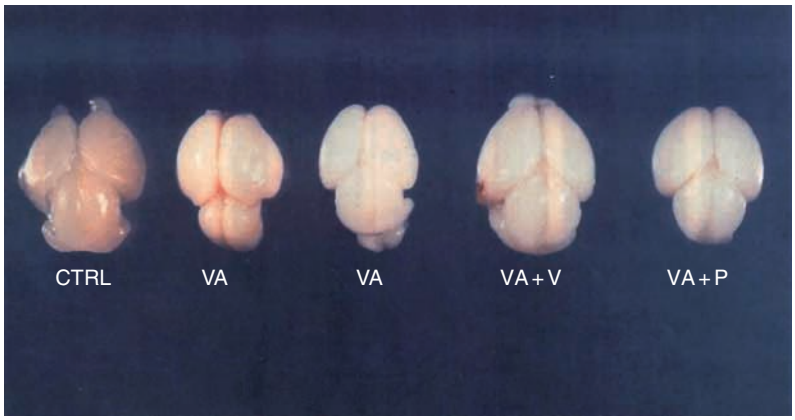

Figure 2 Treatment of pregnant mice with the vasoactive intestinal polypeptide (VIP)/neurotensin antagonist during embryonic days 9–11 results in microcephaly and growth restriction in the offspring. Co-treatment with VIP prevents microcephaly, but co-treatment with pituitary adenylate cyclase-activating polypeptide (PACAP) does not. CTRL, control; VA, VIP antagonist treatment; VA+P, VIP antagonist treatment plus PACAP; VA+V, VIP antagonist plus VIP. Reprinted from Gressens P, Hill JM, Paindaveine P, Gozes I, Fridkin M, and Brenneman DE (1994) Severe microcephaly induced by blockade of vasoactive intestinal peptide function in the primitive neuroepithelium of the mouse. *Journal of Clinical Investigations* 94: 2020–2027.

Table 2 Neonatal developmental milestones

Behavior	*Control*	*VIP antagonist* [a]	*PACAP antagonist* [b]	*F(p)*
Weight P2 (g)	1.9 ± 0.03	1.7 ± 0.03***	1.9 ± 0.23	21 (0.0001)
Surface right first day, <1 s	5.3 ± 0.33	8.4 ± 0.26***	5.4 ± 0.24	23.8 (0.0001)
Negative geotaxis first day	3.8 ± 0.19	3.8 ± 0.35	3.0 ± 0.21	3.47 (0.051)
Cliff aversion first day	4.3 ± 0.25	4.2 ± 0.34	3.8 ± 0.13	1.13 (0.34)
Rooting first day	4.6 ± 0.2	4.6 ± 0.1	4.5 ± 0.3	0.11 (0.90)
Grasping first day, >1 s	8.3 ± 0.3	10.3 ± 0.1***	7.4 ± 0.4	15.8 (0.0001)
Open field first day	14.7 ± 0.4	18.5 ± 0.3***	14.2 ± 0.3	34.1 (0.0001)
Auditory startle first day	10.8 ± 0.3	11.7 ± 0.1*	10.5 ± 0.3	3.7 (0.042)
Ear twitch first day	10.0 ± 0.1	10.5 ± 0.2	9.7 ± 0.2	2.7 (0.09)
Eye opening first day	13.4 ± 0.1	13.6 ± 0.2	13.7 ± 0.1	2.8 (0.08)
Air righting first day	10.9 ± 0.1	11.3 ± 0.3	11.1 ± 0.2	1.08 (0.35)

Neonatal mice were tested daily for the first 21 days of life for a number of reflexes and developmental signs. P2, postnatal day 2; * indicates $p < 0.05$; *** indicates $p < 0.0001$. Reprinted from Hill JM (2007) Vasoactive intestinal peptide in neurodevelopmental disorders: Therapeutic potential. Current Pharmaceutical Design 13(11): 1079–1089.

[a]VIP hybrid antagonist injected embryonic days 8–11, 30 μg day^{-1}. Data from Gozes I, McCune SK, Jacobson L, et al. (1991) An antagonist to vasoactive intestinal peptide affects cellular functions in the central nervous system. *Journal of Pharmacology and Experimental Therapeutics* 257: 959–966.

[b]PACAP 6–38, injected embryonic days 8–11, 30 μg day^{-1}.

physiological responses in most systems, the actions of these peptides in development appear to be significantly divergent, with the effects on cell division being an example.

Gastrointestinal Effects

VIP and PACAP are widespread throughout the enteric nervous system where they are often co-localized in ganglia and nerve fibers terminating in the blood vessels, intestinal mucosa, and smooth muscle of the gut. PACAP is the more potent neuropeptide in regulating gastrointestinal secretions, where both of these peptides stimulate the secretion of saliva from the submandibular gland and of water and electrolytes from the intestine. VIP and PACAP stimulate the release of insulin, glucagon, and bicarbonate from the pancreas, and by stimulating vasodilation, they promote blood flow throughout the organs of the gastrointestinal system. The peptides cause relaxation of the stomach and lower esophageal sphincter, and they inhibit peristalsis throughout the intestine, resulting in reduced motility of the bowel. Whereas VIP and PACAP have direct effects through innervation of the smooth muscles of the gut, central actions of these peptides can also indirectly influence intestinal relaxation through autonomic connections. In addition, VIP may have immune modulator effects in the

gut through action at receptors on intestinal T lymphocytes. All three VIP/PACAP receptors are found in the digestive tract but with different distribution patterns. VPAC-1 is most abundant in the mucosa and VPAC-2 is more abundant in the blood vessels and smooth muscles. NO is also a potent smooth muscle relaxant, and VIP actions in the gut may be, in part, due to VIP-induced release of NO.

Pulmonary Effects

VIP is one of the most abundant neuropeptides in the respiratory tract, where, along with PACAP, it acts on the shared VIP/PACAP receptors to regulate several airway activities. VIP and PACAP are co-localized in the ganglia and nerve fibers that innervate the tracheobronchial smooth muscles, submucosal glands, and the walls of blood vessels. In addition to having potent effects on bronchial relaxation, VIP and PACAP dilate pulmonary blood vessels, thereby increasing perfusion of the lung, and also stimulate mucus secretion by the glands of the respiratory tract. Furthermore, VIP may also mediate immune mechanisms in the lung through its action as an anti-inflammatory agent. NO is also a known pulmonary vasodilator, relaxes airway smooth muscle, and has immunosuppressive properties. VIP may have an indirect role in these pulmonary activities through its action on NO activity.

Cardiovascular Effects

As the name of the peptide indicates, the vasodilatory effect of VIP has been known since its discovery. However, both VIP and PACAP have substantial effects on the heart and blood vessels. All three of the VIP/PACAP receptors have been detected in the cardiomyoctes, although PAC1 is the most prevalent. VIP and PACAP have been shown to produce positive inotropic and chronotropic effects on the heart. In addition, VIP can augment pacemaker current. Both VIP and PACAP produce significant arterial dilatation in the brain, endocrine glands, and pulmonary system. PACAP displays similar effects in all vascular beds, whereas VIP can differ in response among regional beds. VIP has a much greater effect on arteries than veins, due to differences in receptor density. The response to systemic administration of PACAP has two phases: an initial vasodilatation with a decreased blood pressure, followed later by an increase in blood pressure that is attributed to a vasoconstrictive action secondary to central and sympathetic reflexes.

Endocrine System

Both VIP and PACAP have important actions on endocrine function that is often mediated through their secretagogue action or gene expression of hormones. Early studies demonstrated that VIP released prolactin from the anterior pituitary. These studies were followed by others indicating that VIP and PACAP could release oxytocin and vasopressin from the neural lobe of the pituitary. VIP also releases luteinizing hormone and growth hormone from the pituitary. In contrast, PACAP has been shown to suppress luteinizing hormone secretion and pulse frequency when injected into the hypothalamus. In the pancreas, VIP and PACAP can cause significant insulin release, with the releasing action of PACAP being more potent than that of VIP. VIP acts on the pancreas to release either insulin or glucagon, actions that are dependent on glucose levels. Both VIP and PACAP are present in the testis, where they have effects on sperm motility, steroidogenesis, and germ cell maturation. VIP synthesis and VIP receptors have been observed in female sexual organs, with expression regulated in part by the estrous cycle.

Immune System

PACAP and VIP have multiple functional roles with respect to immune cells that include effects on immune cell proliferation and survival, cell mobility, and regulation of inflammatory responses (Figure 3). The effects of the peptides affect both innate and adaptive immunity. Both peptides have actions that can either suppress or activate inflammation through the regulation of cytokine production and release. Direct inhibition of proinflammatory cytokines such as tumor necrosis factor α, interleukin (IL)-6, and IL-12 from activated microglia and macrophages has been reported. In addition, VIP can upregulate the production of IL-10, an anti-inflammatory cytokine; whereas PACAP can inhibit IL-10 production in lymphocytes. These peptides also have effects on the inhibition, expression, and release of proinflammatory chemokines. Both VIP and PACAP can increase the survival of immune cells that include memory T-helper (Th) 2 cells and thymocytes. PACAP has also been shown to increase the phagocytosis and cell digestion of peritoneal macrophages. Effects of these peptides on immune cell mobility are dependent on the target cells. PACAP can decrease the mobility of lymphocytes while increasing the mobility of macrophages.

Behavioral Effects

VIP and PACAP have been shown to influence several behavioral states in a similar manner. Synaptic terminations of VIP- and PACAP-containing neurons on corticotropin-releasing hormone neurons suggest that these peptides may act centrally to augment stress-related behaviors. Consistent with this hypothesis are the observations that VIP- and PACAP-injected mice show evidence of the stresslike symptoms of increased grooming, locomotor behavior, and reduced food

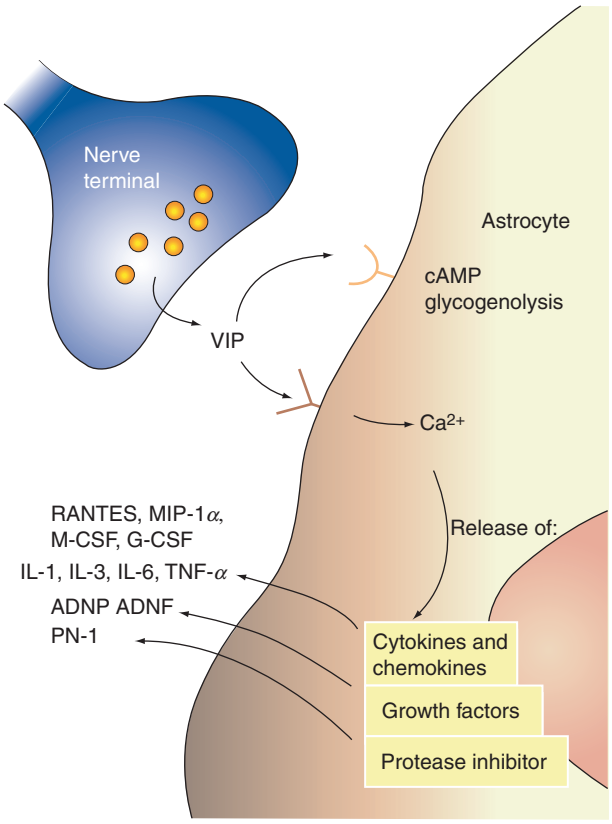

Figure 3 Schematic of VIP secretagogue actions on astrocytes that regulate neural development. Although the receptor that mediates these actions has not yet been established, very low concentrations (0.1–1 nM) produce the release of a diverse array of molecules. ADNF, activity-dependent neurotrophic factor; ADNP, activity-dependent neuroprotective protein; cAMP, cyclic adenosine monophosphate; G-CSF, granulocyte colony-stimulating factor, IL, interleukin; M-CSF, macrophage colony-stimulating factor; MIP, macrophage inflammatory protein; PN, protease nexin; RANTES, regulated upon activation normal T cell expressed and secreted; TNF, tumor necrosis factor; VIP, vasoactive intestinal polypeptide.

consumption. Although PACAP knockout mice also display hyperactivity, they exhibit reduced anxiety-like behavior. Both VIP and PACAP participate in the regulation of circadian rhythms due to their presence in and innervation of the suprachiasmatic nucleus. Both neuropeptides augment rapid-eye movement (REM) sleep, influence learning and memory activities, and can induce analgesia. Both VIP and PACAP have been associated with the normal expression of social behavior. The inhibition of VIP action during embryogenesis has been linked to abnormal and reduced social interaction in the adult mouse. In addition, PAC-1 knockout mice exhibit an abnormally rapid habituation to novel mice. VIP and PACAP are known to potentiate sexual behaviors and the vasodilatory abilities of VIP have an important role in penile erection and may be associated with female sexual function.

Neuroprotective Effects

VIP and PACAP have neuroprotective effects against the toxic effects of excitatory amino acids, β-amyloid, lipopolysaccharide, viral proteins, and oxidative stress. Neuroprotective effects have been associated with the VPAC2 receptor and PAC1. The protective effects of these peptides are complicated, in that both direct protective effects and indirect actions through the release and production of substances released from glia have been observed. Among the neuroprotective and developmentally significant molecules related to the action of these peptides are nerve growth factor, ciliary neurotrophic factor, brain-derived neurotrophic factor, insulin-like growth factor, cytokines, chemokines, and protease nexin-1. These important mediators of VIP and PACAP action have well-recognized effects on preventing programmed cell death, regulating neurite outgrowth and controlling cell division.

Evaluation of VIP/PACAP interaction with glial cells resulted in the discovery of activity-dependent neurotrophic factor (ADNF) and activity-dependent neuroprotective protein (ADNP). Both VIP and PACAP regulate the expression of ADNP in astrocytes through the VPAC2 receptor. The direction of this regulation for ADNP is age-dependent, younger cells showing increases in mRNA and older cells exhibiting decreases. ADNP was further identified as a protein essential for brain development. Activity scanning studies identified ADNF-9 (also known as SAL – SALLSIPA) in ADNF and neuroprotective agent (NAP) (NAPVSIPQ) as potent neuroprotective peptides. NAP is in phase II clinical trials.

Pharmacology

Knowledge of the functional roles of VIP and PACAP is significantly determined by the discovery of specific agonists and antagonists to the relevant receptors and the generation of mice in which these receptors or their ligands have been genetically deleted or enhanced. The pharmacological agents for VIP and PACAP are limited and mostly confined to analogs of the peptides themselves or chimeric peptides. In regard to antagonists, C-terminal fragments are commonly used that include VIP(6–28) and PACAP(6–38). These peptides lack agonist activity, but do not exhibit subtype receptor specificity. However, the peptide PG (97–269) is an antagonist with a 1000-fold specificity toward the VPAC1 receptor. Recent reports indicate that PG(99–465) is a VPAC2 receptor antagonist. Other antagonists with broader activity among these receptors include hybrid peptides (neurotensin(6–11)-VIP(7–28) and stearyl neurotensin(6–11) VIP(7–28)).

In regard to agonists, Ro25-1392, a cyclic analog of VIP, appears to be the selective for VPAC2 and the agents [Ala^{11,22,28}]VIP and [K15, R16, L27] VIP (1–7)/growth hormone-releasing factor (GRF)(8–27) for the VPAC1 receptor. A high-potency nonselective agonist of VIP has also been described – stearyl norleucine VIP.

Transgenic Mice

Transgenic mice that overexpress or delete VIP/PACAP and their receptors provide an invaluable insight into the biological importance of these peptides. Recent studies demonstrated that these peptides are important in memory, growth regulation, reproductive function, behavior, cardiovascular development, and immune function. Early studies with reduced expression of VIP in adulthood resulted in mice with reduced performance in spatial learning. Deletion of the PACAP gene also produced impaired memory performance and reduced hippocampal long-term potentiation. Recent studies using VIP and VPAC2 receptor knockout mice established that VIP coordinates daily rhythms in the suprachiasmatic nucleus by synchronizing a small population of pacemaking neurons and maintaining rhythmicity in a larger subset of neurons. VIP and PACAP receptor knockout and overexpressing mice show altered immune responses in variety of models. In PAC1 knockout mice, pulmonary hypertension and heart failure are observed after birth. PACAP knockout mice are viable at birth, but exhibit increased postnatal mortality; the underlying cause of this premature death is not yet clear. Interestingly, PACAP knockout mice exhibit a significant decrease in the intake of a high-carbohydrate diet but not a high-fat diet. The phenotype of VPAC2 knockout mice also included retarded growth and had reduced serum insulin-like growth factor-1 levels compared with controls. In older males lacking VPAC2, there was a decrease in fertility rate accompanied by diffuse seminiferous tubular degeneration and hypospermia. Overexpression of the human PAC1 receptor has been shown to produce a dose-dependent hydrocephalus-related abnormality in mice, further emphasizing the important role of PACAP and VIP in brain development.

Role in Disease

The roles for VIP and PACAP have begun to emerge in the areas of cancer, cardiovascular disease, and immune function. In addition to the VIPoma tumors that result in a chronic watery diarrhea syndrome, both VIP and PACAP have been shown to have autocrine growth factor activity for a variety of tumors, including small-cell lung cancer and non-small-cell lung cancer, along with colon and breast cancers. Myocardial fibrosis, heart failure, and pulmonary hypertension have been associated with changes in cardiac VIP or in the responsiveness of the VIP/PACAP receptors. The degradation of VIP appears to be increased in the myocardial fibrosis, resulting in decreased VIP levels. The altered affinity, density, and physiological responsiveness of the receptors suggest that these peptide systems may have important pathophysiological roles in cardiac disease.

Therapeutic Potential

The neuroprotective, antiproliferative, anti-inflammatory, antihypertensive, and bronchodilitary properties of these peptides suggest that drugs directed toward this family of receptors could be very beneficial. However, because of the rapid degradation of the peptides, their use as therapeutics has been limited. Because both of these peptides can release neuroprotective proteins in the CNS, it is possible that agents derived from these VIP/PACAP releasable proteins could have therapeutic value. Current examples for stabilized VIP derivatives include stearyl norleucine VIP and shorter analogs, as well as peptide derivatives of VIP/PACAP-regulated proteins. At present, one such NAP, derived from ADNP, is currently in phase II clinical trials. In regard to immune disorders, recent studies have investigated the therapeutic effect of VIP on experimental autoimmune encephalomyelitis through its effect on the downregulation of inflammation and autoimmune responses. In addition, systemic administration with these peptides attenuated the clinical symptoms and improved the cytokine levels in animal models of rheumatoid arthritis, Crohn's disease, septic shock, and multiple sclerosis. Other potential areas of treatment include erectile and female sexual dysfunction, asthma, diabetes, and skin disorders. Although in its early stages of exploration of therapeutic potential, the development of agents that interact with this family of receptors should provide new and innovative approaches to fill unmet medical needs.

See also: Mammalian Neuropeptide Families; Neural Repair and Regeneration: Inflammatory Mechanisms and Cytokines; Neuropeptide Inactivation or Metabolism; Neuropeptide Receptors – Drug Development; Neuropeptide Synthesis and Storage; Neuropeptides in Autonomic Neurons; Neuropeptides and Receptors in Glia; Neuropeptides: Discovery; Neuropeptides: Endocrine Cells; Neuropeptides: Enteric Nervous System; Neuroprotection: Pharmacological Approaches.

Further Reading

Abad C, Gomariz RP, and Waschek JA (2006) Neuropeptide mimetics and antagonists in the treatment of inflammatory disease: Focus on VIP and PACAP. *Current Topics in Medicinal Chemistry* 6: 151–163.

Arimura A (1998) Perspectives on pituitary adenylate cyclase activating polypeptide (PACAP) in the neuroendocrine, endocrine and nervous system. *Japanese Journal of Physiology* 48(5): 301–331.

Arimura A (2003) Neuropeptides of the pituitary adenylate cyclase-activating polypeptide/vasoactive intestinal polypeptide/growth hormone-releasing hormone/secretin family of testis. *Endocrine* 20(3): 201–214.

Brenneman DE, Glazner G, Hill JM, et al. (1998) VIP neurotrophism in the central nervous system: Multiple effectors and identification of a femtomolar-acting neuroprotective peptide. *Annals of the New York Academy of Sciences* 865: 207–212.

Dejda A, Sokolowska P, and Nowak JZ (2005) Neuroprotective potential of the three neuropeptides PACAP, VIP and PHI. *Pharmacological Reports* 57(3): 307–320.

Delgado M, Pozo D, and Ganea D (2004) Significance of vasoactive intestinal peptide in immunomodulation. *Journal of Pharmacology and Experimental Therapeutics Reviews* 56: 249–290.

Fahrenkrug J and Hannibal J (2004) Neurotransmitters co-existing with VIP and PACAP. *Peptides* 25(3): 393–401.

Fukuchi M, Tabuchi A, and Tsuda M (2005) Transcriptional regulation of neuronal genes and its effect on neural functions: Cumulative mRNA expression of PACAP and BDNF genes controlled by calcium and cAMP signals in neurons. *Journal of Pharmacological Sciences* 98(3): 212–218.

Gozes I and Brenneman DE (1989) VIP: Molecular biology and neurobiological function. *Molecular Biology* 3: 201–236.

Gozes I and Furman S (2003) VIP and drug design. *Current Pharmaceutical Design* 9(6): 483–494.

Gozes I, Morimoto BH, Tiong J, et al. (2005) NAP: Research and development of a peptide derived from activity-dependent neuroprotective protein (ADNP). *CNS Drug Reviews* 11(4): 353–368.

Harmar AJ, Arimura A, Gozes I, et al. (1998) International Union of Pharmacology, XVIII: Nomenclature of receptors for vasoactive intestinal peptide and pituitary adenylate cyclase-activating polypeptide. *Pharmacological Reviews* 50: 265–270.

Henning RJ and Sawmiller DR (2001) Vasoactive intestinal peptide: Cardiovascular effects. *Cardiovascular Research* 49: 27–37.

Laburthe M and Couvineau A (2002) Molecular pharmacology and structure of VPAC receptors for VIP and PACAP. *Regulatory Peptides* 108: 165–173.

Lutz EM, Ronaldson E, Shaw P, et al. (2006) Characterization of novel splice variants of the PAC1 receptor in human neuroblastoma cells: Consequences for signaling by VIP and PACAP. *Molecular and Cellular Neuroscience* 31: 193–209.

Moody TW, Hill JM, and Jensen RT (2003) VIP as a trophic factor in the CNS and cancer cells. *Peptides* 24(1): 163–177.

Said SI and Dickman KG (2000) Pathways of inflammation and cell death in the lung: Modulation by vasoactive intestinal peptide. *Regulatory Peptides* 93(1–3): 21–29.

Sherwood NM, Krueckl SL, and McRory JE (2000) The origin and function of the pituitary adenylate cyclase-activating polypeptide (PACAP/glucagon superfamily). *Endocrine Reviews* 21: 619–670.

Vaudry D, Gonzalez BJ, Basille M, et al. (2000) Pituitary adenylate cylcase-activating polypeptide and its receptors: From structure to functions. *Pharmacological Reviews* 52: 269–324.

Vasopressin/Oxytocin and Receptors

W S Young III and H Gainer, National Institutes of Health, Bethesda, MD, USA

Published by Elsevier Ltd.

Introduction

Oxytocin (OXT) and arginine vasopressin (AVP) are nine-amino-acid, cyclic neuropeptides (nonapeptides) that differ from one another by only two amino acids, at the third and eighth positions (see Figure 1), and that are very distinct in their functional activities. Some of these activities were discovered over a century ago, when Oliver and Schaefer reported in 1895 that extracts of the neural lobe of the pituitary produced powerful presser effects, now known to be due to AVP. The so-called ocytocin effects of these extracts on mammalian uteri were described by Sir Henry Dale in 1906. The term oxytocin refers to its ability to produce quick birth (induction of parturition), and by 1911 it was realized that there was also a galactogenic effect which evoked milk secretion by the lactating mammary gland. OXT and AVP have the distinction of being the first neuropeptides to be discovered and fully characterized with respect to their structures and functions, and this achievement led to a Nobel Prize in chemistry to Vincent DuVigand in 1955. These nonapeptides are associated *in vivo* with specific 100-amino-acid cysteine-rich proteins known as the neurophysins (NPs), which bind their respective peptides and are co-packaged in the neurosecretory granules in the posterior pituitary, from which they are co-secreted with the peptides.

In addition to their well-known peripheral hormonal actions, such as causing milk let-down during suckling at the lactating breast and their effects on uterine contraction, facilitation of sperm transport in ejaculation, natriuresis (sodium excretion by the kidney) in the case of OXT, and liver gluconeogenesis, vasoconstriction, corticotropin secretion from the pituitary gland, and antidiuresis (reabsorption of water from the collecting ducts of the kidney) in the case of AVP, these neuropeptides also have profound actions within the central nervous system (CNS) (Table 1). OXT injected into the cerebrospinal fluid causes sexual arousal (spontaneous erections) in rats, social pair bonding in female voles, induction of maternal behavior, and reduction of various stress-related events (reduction of anxiety, blood pressure, and cortisol levels; increase of tolerance to pain). OXT is also reported to increase trust and reduce fear, and it may inhibit tolerance to various drugs such as opiates, cocaine, and alcohol. AVP has been implicated in memory formation, aggression, blood pressure and temperature regulation, circadian regulation of endocrine rhythms, and pair bonding in male voles. The absence of these peptides produces various known pathological conditions. Decreased AVP secretion or decreased renal sensitivity to AVP leads to neurogenic or nephrogenic diabetes insipidus, respectively. The consequences of this condition is hypernatremia (increased blood sodium content), polyuria (increased urine production), and polydipsia (increased thirst and drinking). An increased level of AVP secretion produces hyponatremia (low plasma sodium) and a syndrome of inappropriate antidiuretic hormone (SIADH). An impairment of OXT secretion blocks milk let-down during suckling, but does not significantly impair parturition. It has been suggested that a deficiency in OXT secretion might be involved in autism, which also has suggestive linkages to the OXT receptor (OXTR) and one of the AVP receptors, AVPR1a.

The Oxytocin and Vasopressin Peptides

Gene Structures and Evolution

Information about mammalian OXT and AVP nonapeptide genes is presented in Table 2. The structurally related OXT and AVP genes in mammals each contain three exons and two introns, and both genes are located on the same chromosomal locus (chromosome 2 in mice and chromosome 20 in humans) but are transcribed in opposite directions (see Figure 1). The first exons of the AVP and OXT genes code for a signal peptide, the hormone (AVP or OXT, respectively), a three-amino-acid spacer (-Gly-Lys-Arg-), and the first nine N-terminal amino acids of NP (NP-II or NP-I, respectively). The second exon of each gene codes for the highly conserved mid-portion of NP (amino acids 10–76). The third exon of the AVP gene codes for the remaining C-terminal amino acids of NP, a monobasic cleavage site (-Arg-) and the C-terminal glycopeptides (GPs) of 39 amino acids. In contrast, the third exon of the OXT gene codes only for the C-terminal amino acids of OXT-associated NP-I.

It is generally believed that the two mammalian AVP and OXT superfamily branches developed from a common ancestral gene through duplication about 450 Ma. The strong conservation of the NP domain of the different AVP-related prohormones further supports this view. The NP domain is very well conserved and exists in all vertebrate and invertebrate species. Two segments in the NP sequence, which fall within the region encoded by the second exon ($NP_{10–76}$), show 58% homology to one another and 60% at the nucleotide level. This suggests that the

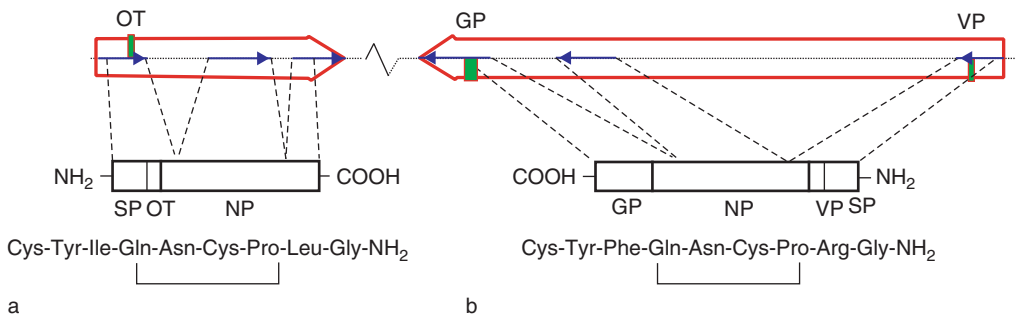

Figure 1 Organization of genes: (a) oxytocin gene; (b) vasopressin gene. Both genes are composed of three exons (indicated by the small solid arrows) separated by two introns (indicated by dotted lines between the exons). The genes are present on the same chromosome, but are transcribed in opposite directions (indicated by the large hollow arrows). Similar gene structures but with variable lengths of intergenic sequences have been found in various mammalian species. The three exons code for preprohormones that contain a signal peptide (SP), AVP (VP) or OXT (OT), neurophysin (NP), and, in the case of AVP, also a glycopeptide (GP) that are liberated from the preprohormone during processing in the transported dense-core vesicles. Adapted from Young WS III and Gainer H (2003) Transgenesis and the study of expression, cellular targeting and function of oxytocin, vasopressin and their receptors. *Neuroendocrinology* 78(4): 185–203.

Table 1 Actions of oxytocin and vasopressin

Vasopressin	*Oxytocin*
Salt–water balance	Milk letdown
Affiliation	Parturition
Antipyresis	Affiliation
Blood pressure regulation	Anxiolysis
Circadian rhythm	Cardiac release of ANP
Facilitation of avoidance behavior	Feeding behavior
Maternal aggression	Grooming
Paternal behavior	Maternal aggression
Sexual behavior	Maternal behavior
Social recognition	Sexual behavior
Stress response	Social recognition
Territorial aggression	Territorial aggression

ANP, atrial natriuretic peptide.

primordial NP gene itself arose from a partial gene duplication. Many amino acid sequences of phylogenetically diverse AVP/OXT-related peptides are now available (see Table 3). Because cyclostomes, the most primitive vertebrate for which there are data, possess only one peptide – vasotocin (which differs from OXT and AVP each by only one amino acid; see Table 3) – it is generally assumed that this peptide was most closely related to the ancestral molecule. Several invertebrate peptides in several phyla have been sequenced and have been shown to contain the critical amino acids in positions 1, 5, 6, 7, and 9 (see Table 3). The first cloning of a cDNA in an invertebrate, the mollusk *Lymnae stagnalis*, which contains the Lys-conopressin preprohormone sequence, permitted a comparison between invertebrates and vertebrates of not only the nonapeptide moiety but also the neurophysin and C-terminal domain. The cDNA coded for a 155-amino-acid preprohormone with high homology to the vertebrate neurophysins (49% identity to human AVP-associated neurophysin), including conservation of all 14 cysteine residues. The significance of these data is that they show that the ancestral gene for the AVP/OXT superfamily was already present in the Archaemetazoa, a stem group from which both vertebrates and invertebrates diverged approximately 600 Ma.

In a dramatic cross-phyla experiment, transgenic rats were made that contained a 40 kb cosmid sequence derived from the isotocin (IT)–vasotocin (VT) gene locus in the pufferfish, *Fugu*. Cell-specific expression of the fish IT gene was found only in the magnocellular OXT neurons in the rat hypothalamus, and this transgene could be physiologically regulated by dehydration in a manner similar to the endogenous gene. This experiment showed that there was sufficient conservation of the cell-specific expression and physiological regulation mechanisms despite an evolutionary distance of 450 My between these species.

OXT and AVP Gene Regulation and Functional Activity

OXT and AVP fibers are found throughout the CNS. However, the expression of the OXT and AVP genes occurs primarily in the hypothalamus and in the magnocellular neurons (MCNs) of the paraventricular nucleus (PVN) and supraoptic nucleus (SON) (Figure 2). There are also parvocellular neurons in the PVN that synthesize corticotropin-releasing hormone (CRH) and co-express AVP, that are secreted in the portal circulation to stimulate the release of adrenocorticotropin (ACTH) from anterior pituitary corticotropes. Parvocellular neurons in the suprachiasmatic nucleus (SCN), the central circadian clock in mammals, also synthesize AVP and are involved in regulating the

Table 2 Gene information about vasopressin and oxytocin

Gene	*Species*	*Official gene symbol*	*Chromosome*	*Known exons*	*Amino acids*[a]	*Locus ID*[b]
Vasopressin	Human	AVP	20 (p13)	3	164	551
	Rat	AVP	3 (q41–q42)	3	164	24 211
	Mouse	AVP	2 (73.2cM)	3	168	11 998
Oxytocin	Human	OXT	20 (p13)	3	125	5020
	Rat	OXT	3 (q36)	3	125	25 504
	Mouse	OXT	2 (73.5cM)	3	125	18 429

[a]Number of amino acids in the preprohormone.
[b]Locus ID from http://www.ncbi.nlm.nih.gov.
Adapted from Caldwell HK and Young WS III (2006) Oxytocin and vasopressin: Genetics and behavioral implication. In: Lim R (ed.) *Handbook of Neurochemistry and Molecular Neurobiology,* 3rd edn., pp. 573–607. New York: Springer.

Table 3 Vasopressin/oxytocin superfamily

Structure	*Name*	*Animals*
Vertebrate vasopressin family		
Cys-Tyr-Phe-Gln-**Asn**-**Cys**-**Pro**-Arg-**Gly**-NH_2	Vasopressin (ADH)	Mammals[a]
Cys-Tyr-Phe-Gln-**Asn**-**Cys**-**Pro**-Lys-**Gly**-NH_2	Lysipressin	Pigs, hippos, warthogs, some marsupials
Cys-Phe-Phe-Gln-**Asn**-**Cys**-**Pro**-Arg-**Gly**-NH_2	Phenypressin	Some marsupials
Cys-Tyr-Ile-Gln-**Asn**-**Cys**-**Pro**-Arg-**Gly**-NH_2	Vasotocin[b]	Nonmammals
Vertebrate oxytocin family		
Cys-Tyr-Ile-Gln-**Asn**-**Cys**-**Pro**-Leu-**Gly**-NH_2	Oxytocin	Mammals[c], ratfish
Cys-Tyr-Ile-Gln-**Asn**-**Cys**-**Pro**-Ile-**Gly**-NH_2	Mesotocin	Marsupials, birds, reptiles, amphibians, lungfishes
Cys-Tyr-Ile-Ser-**Asn**-**Cys**-**Pro**-Ile-**Gly**-NH_2	Isotocin	Bony fishes
Cys-Tyr-Ile-N/Q-**Asn**-**Cys**-**Pro**-L/V-**Gly**-NH_2	Various tocins	Sharks
Invertebrate vasopressin/oxytoxin superfamily		
Cys-Leu-Ile-Thr-**Asn**-**Cys**-**Pro**-Arg-**Gly**-NH_2	Diuretic hormone	Locust
Cys-Phe-Val-Arg-**Asn**-**Cys**-**Pro**-Thr-**Gly**-NH_2	Annetocin	Earthworm
Cys-Phe-Ile-Arg-**Asn**-**Cys**-**Pro**-Lys-**Gly**-NH_2	Lys-connopressin	Geography and imperial cones, pond snail, sea hare, leech
Cys-Ile-Ile-Arg-**Asn**-**Cys**-**Pro**-Arg-**Gly**-NH_2	Arg-connopressin	Striped cone
Cys-Tyr-Phe-Arg-**Asn**-**Cys**-**Pro**-Ile-**Gly**-NH_2	Cephalotocin	Octopus
Cys-Phe-Trp-Thr-Ser-**Cys**-**Pro**-Ile-**Gly**-NH_2	Octopressin	Octopus

[a]Vasopressin mammals exclude some marsupials, pigs, and some others.
[b]Vasotocin is the progenitor of the vertebrate neurohypophysial hormones; only vasotocin found in hagfish and lampreys (Agnatha appeared 500 Ma). ZADH, antidiuretic hormone.
Adapted from Cardwell HK and Young WS III (2006) Oxytocin and Vasopressin: Genetics and behavioral implications. In: Lim R (ed.) *Handbook of Neurochemistry and Molecular Neurobiology*, 3rd edn., pp. 573–607. New York: Springer.
[c]Oxytocin mammals include some marsupials.

circadian rhythm of various hormones. There are other parvocellular neurons in the PVN that regulate autonomic functions via projections to the brain stem and spinal cord. Finally, some species have parvocellular neurons elsewhere, such as in the amygdala and bed nucleus of the stria terminalis, that express smaller amounts of AVP or OXT.

The MCNs in the hypothalamo-neurohypophysial system (HNS) have been the most studied with respect to the functional regulation of OXT and AVP gene expression, in large part due to their compact nuclear organization in the CNS; most of the approximately 12 000 cells in the rodent hypothalamus are located in the two bilateral nuclei, the PVN and SON (Figures 2 and 3). The HNS neurons project via well-defined axonal tracts to the posterior pituitary (i.e., the neural lobe), where each axon is estimated to branch into hundreds of nerve terminals and where these axonal branches and terminals represent about 50% of the total tissue mass of the neural lobe. The relatively easy access to the HNS cellular components, the cell bodies in the PVN and SON and the axons in the median eminence, by both stereotaxic and micropunch assay methods, and to the nerve terminals by their presence outside of the blood–brain barrier in the posterior pituitary, have made these MCNs favorite objects of many biochemical and physiological studies. Nevertheless, similar studies have been performed on other neuronal systems that express AVP (e.g., the CRH neurons in the PVN, and the AVP neurons of the SCN); these are briefly described in Table 4.

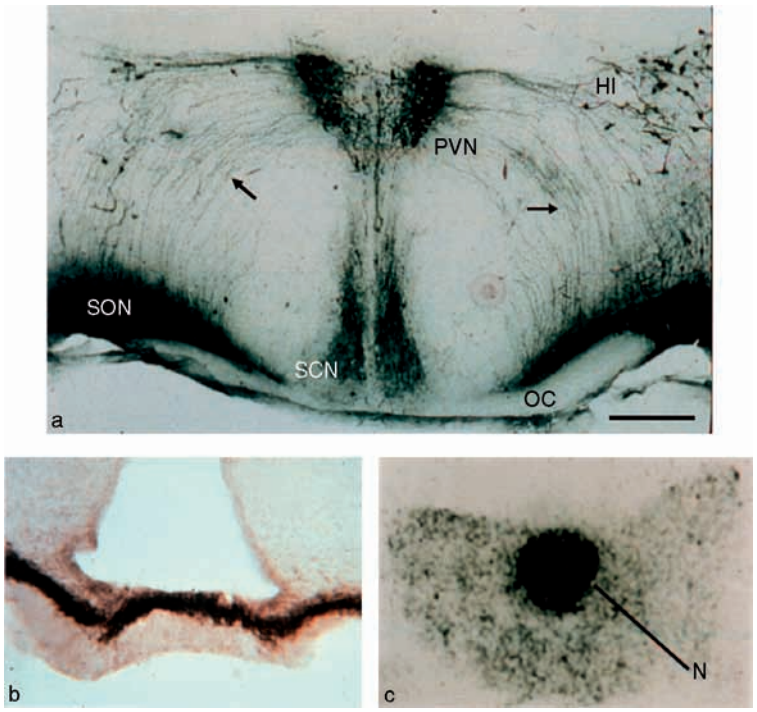

Figure 2 Immunohistochemical views: (a) vasopressinergic neurons in the hypothalamus; (b) axons in the internal zone of the median eminence; and (c) neural lobe in the posterior pituitary. In (a), the oxytocinergic and vasopressinergic magnocellular neurons in the supraoptic nucleus (SON) and paraventricular nucleus (PVN) project to their axons to the neural lobe (N; shown in (c)), and together with more lateral scattered neurons in the accessory nuclei (HI), constitute the hypothalamo-neurohypophysial system (HNS). The suprachiasmatic nucleus (SCN) contains smaller vasopressinergic (only) neurons, located dorsomedially in the nucleus, which project to the ventral region of the PVN, and is involved in the circadian regulation of various endocrine systems. The PVN also contains both oxytocin- and vasopressin-containing small (parvocellular) neurons that project to the external zone of the median eminence and more caudal CNS regions, including the spinal cord (see **Table 4**). OC, optic chiasm. Scale bar = 1 mm (a).

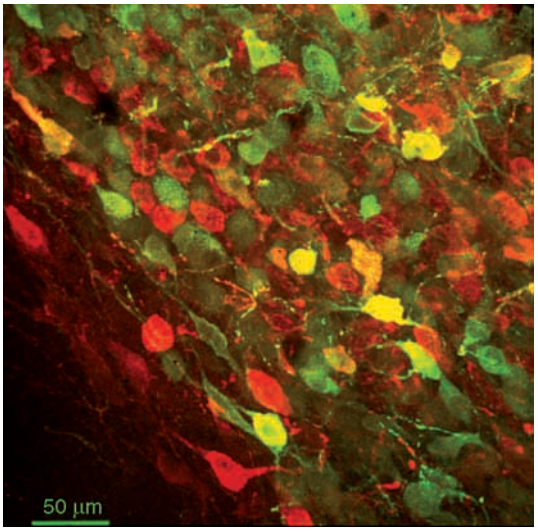

Figure 3 Double-labeled immunofluorescent view of oxytocin (red)- and vasopressin (green)-containing magnocellular neurons in the supraoptic nucleus (SON) of a lactating female rat. Note in this merged view that a significant number of the neurons in the SON contain both oxytocin and vasopressin. AVP and OXT mRNAs increase in the magnocellular neurons during lactation.

Mature MCNs respond to perturbations in water balance by releasing large amounts of stored AVP and OXT into the general circulation. Hyperosmotic stimuli, such as dehydration or salt-loading, result in an increase in AVP and OXT secretion and in transcription of both genes. Conversely, under hypoosmolar conditions, there is no need for either AVP or OXT secretion, and the release of these peptides is completely shut down, leading to a virtually complete inhibition of the transcription of their genes and the biosynthesis of the nonapeptides. The mechanisms that underlie these changes are still not understood. There are also two types of posttranscriptional regulation found in response to hyperosmolality. The first is an increase in length of the poly (A) tail of the AVP mRNA from around 200 to 400 nucleotides, which does not occur in the OXT mRNA. This change is found in the rat MCNs but not in the mouse. The second change is that during hyperosmolality the short poly (A)-tailed OXT and AVP mRNAs appear to be axonally transported to the MCN nerve terminals in the posterior pituitary. The physiological significance of these two changes is

Table 4 Cellular localization, functions, and physiological regulation of oxytocin and vasopressin in the CNS

Cellular localization	*Function and gene regulation*
Hypothalamic	
Hypothalamo-neurohypophysial system (HNS: PVN + SON)	Osmotic regulation: 2–3 times increase in gene expression in hyperosmolality; ~5 times decrease in gene expression in hypoosmolality
	Lactation (+suckling) and plasticity: increase in neuron soma size; increase gene expression; synaptic remodeling OXT ≫ AVP; dual expression in 15% MCNs; change in burst pattern (suckling) in OXT MCNs
Suprachiasmatic nucleus (SCN) (30% AVP neurons)	Intrinsic circadian rhythm of AVP gene expression + secretion (E box-related)
	Extrinsic factors, for example cAMP via VIP input
Corticotropin-releasing hormone parvocellular PVN neurons (mpdPVN; AVP only)	Stress-regulated increase in AVP gene expression
	Feedback via glucocorticoids inhibits AVP
Dorsal and ventromedial parvocellular PVN neurons (pPVN; OXT and AVP)	Project to brain stem, hindbrain, spinal cord, and forebrain(?)
	Integrates sympathetic anteriodorsolateral and parasympathetic outflow (preautonomic)
Extrahypothalamic	
Bed nucleus of the stria terminalis (BNST)	Local testosterone increases AVP
	Males have 2.3 times more AVP neurons than females
	Developmental effects of testosterone
	Glucorticoids have no effect
Medial amygdala (MeA)	Glucorticoids decrease AVP
	Testosterone increases AVP

AVP, vasopressin; cAMP, cyclic adenosine monphosphate; CNS, central nervous system; MCN, magnocellular neuron; OXT, oxytocin; PVN, paraventricular nucleus; SON, supraoptic nucleus; VIP, vasoactive intestinal polypeptide.

unclear, but it is generally suggested that the increased poly (A) tails produce increased translational efficiency and/or mRNA stability.

The MCNs also contain other co-expressed neuropeptides such as the opioid peptides, enkephalin and dynorphin; neuropeptide Y; CRH; cholecystokinin (CCK); and galanin. The transcription of these co-expressed peptides is also regulated by osmotic stimuli, and they are selectively expressed (e.g., dynorphin and galanin primarily in AVP MCNs, and CRH and CCK in OXT MCNs), generally co-packaged with the AVP or OXT peptides in large dense-core vesicles, and co-secreted in response to neural activity. These co-expressed peptides are secreted both from the MCN nerve terminals in the posterior pituitary and from their dendrites in the hypothalamus. In both cases, they modulate the activities of adjacent cells via either paracrine or autocrine mechanisms.

Both AVP and OXT mRNA increase in the MCNs during lactation as a result of increased gene expression. The level of the OXT mRNA during early lactation is maintained by the suckling stimuli; the removal of the pups causes a decrease in OXT mRNA. This decrease can be inhibited by intraventricular OXT infusion, suggesting that the afferent stimuli activate the OXT MCNs to release OXT from their dendrites, which then act in either an autocrine or paracrine fashion to maintain the increased mRNA levels. Later in lactation (e.g., after 10 days in rats) the increased OXT mRNA becomes less dependent on the suckling stimulus and more on the systemic steroidal milieu.

Pathological Conditions Related to OXT and AVP

Familial neurohypophysial diabetes insipidus (FNDI) is an autosomal dominant disease (rarely recessive) in which mutations in the AVP prohormone cause defects in the synthesis of AVP and, hence, result in a large increase in urine production (polyuria) and fluid intake (polydipsia). In human FNDI, mutations have been observed in either the signal peptide, the AVP moiety, or the NP moiety of the AVP preprohormone. Thirty-two different mutations have been identified in more than 40 pedigrees. The disease is dominant, so one mutant prohormone allele suffices to cause the defect, and the onset of disease symptoms takes several months or years. This suggests that the mutant human AVP prohormone somehow interferes with either synthesis, transport, or processing of the wild–type prohormone. Overexpression of mutant AVP precursors in cell lines with peptidergic properties showed that mutant proteins were largely retained in the endoplasmic reticulum (ER) and that accumulation of precursor caused morphological and probably functional derangement of the ER. These disturbances of the ER appear to be deleterious to the functionality and viability of the MCNs.

The Brattleboro (*di/di*) rat was identified as a mutant of the Long Evans strain that had a hereditary form of diabetes insipidus. This disease is transmitted as an autosomal recessive trait. This pathology was shown to be due to a lack of AVP production by the HNS. Characteristically, water resorption is greatly

impaired in the distal kidney tubules, leading to a massive loss of body fluid. The rats can excrete almost 70% of their body weight in hypotonic urine per day and compensate for the resulting dehydration by drinking an equivalent amount of water. The genetic defect of the Brattleboro rat is the deletion of a single deoxyguanosine within the second exon of the AVP gene. The resulting frame shift alters the amino acid sequence of the NP moiety from residue 64 onward and the entire glycopeptide (GP) region. Due to the new amino acid sequence, 5 of the 14 cysteine residues in NP are missing, and the stop codon within the mRNA is missing, causing translation to continue through the normally noncoding part of the mRNA into the 3′-poly (A) tail and creating a poly-lysine tail at the C-terminus of the precursor. Although the mutant AVP gene is correctly transcribed and spliced, mRNA is not efficiently translated. Therefore, it has been suggested that the expression of the mutant precursor is inhibited by a block at the translational level. In the *di/di* rat, the mutant *di* precursor can be detected in MCNs by immunohistochemistry. Immunoelectron microscopy studies showed that the mutant precursor is retained in the ER and does not reach those secretory granules.

Unlike the situation with AVP, no naturally occurring mutations of the OXT peptide or OXTR genes have been described to date. Most of our information about the consequences of OXT deficiency comes from knockout mice with targeted deletions of the OXT gene. These knockout mice give birth, and there is no apparent difficulty in parturition, although they are unable to suckle their young because the milk-ejection reflex is nonfunctional. The offspring of homozygous mothers survive only with cross-fostering with wild types. Although parturition appears normal in OXT knockout mice, there are data that indicate that OXTR antagonists slow parturition in mice and rats, and the OXT antagonist atosiban slows the onset of deliveries in women with indications of preterm labor. These observations suggest that OXT does play a subtle role in parturition.

Vasopressin and Oxytocin Receptors

Receptor Background

There are two principal classes of AVP receptors: V1 and V2 receptors (AVPR1 and AVPR2, respectively), both of which are seven-transmembrane G-protein-coupled receptors. Activation of the AVPR2 increases cyclic adenosine monophosphate (cAMP) to mediate the classic antidiuretic effects of AVP. The two AVPR1s are coupled to $G_{q/11}$ guanosine triphosphate (GTP)-binding proteins, which along with $G_{\beta\gamma}$ activates phopholipase C (PLC) activity. PLC then generates 1,4,5-inositol triphosphate (IP3) and diacylglycerol (DAG) from phosphatidylinositol 4,5-bisphosphate (PIP2). IP3 facilitates the release of intracellular Ca^{2+} stores, whereas DAG activates protein kinase C to modulate cellular activity. There are two subtypes of the AVPR1: AVPR1a and AVPR1b. In the periphery, AVPR1a mediates the effects of AVP on vasoconstriction and can be found in the liver, kidney, platelets, and smooth muscle. Centrally, AVPR1a is found in a variety of brain nuclei where AVPR1a has been implicated in the regulation of several social behaviors, including social recognition, affiliative behavior, aggressive behavior, and scent-marking behavior.

The AVPR1b (sometimes called V3R) was originally described in the pituitary and finally cloned. Subsequently, AVPR1b was also found in the brain as well as in several peripheral tissues. More recently, AVPR1b has been linked to stress adaptation and aggressive behavior in mice. There is no conclusive evidence for expression of the AVPR2 within the CNS. Transcription of the Arhgap4 gene within the CNS overlaps the AVPR2 and reverse transcriptase–polymerase chain reaction (RT-PCR) analysis must use specific conditions to avoid this contaminant.

Recent work indicated that there are other proteins in the brain that bind AVP (Table 5). The AVP-activated Ca^{2+}-mobilizing receptor protein (VACM-1; Cullin-5) and the dual angiotensin II/vasopressin receptors have widespread neuronal distributions. The former receptor is coupled to the phosphoinositol pathway and the latter to the adenylate cyclase system. The roles of these receptors in the brain, including behavioral roles, are unknown; however, levels of Cullin-5 increase in the cerebral cortex and, to a lesser extent, in the hypothalamus following water deprivation.

The single OXTR appears to transduce the actions of OXT. This receptor was first isolated and identified by Kimura and colleagues. The OXTR is also a member of the G-protein-coupled receptor family. Like other members of this family, the OXTR contains seven transmembrane domains and is similar in structure to the AVP receptors. The OXTR is also coupled to $G_{q/11}$ GTP-binding proteins and $G_{\beta\gamma}$ to cause the hydrolysis of phosphatidylinositol and act through pathways similar to the AVPR1s. Table 5 summarizes genetic information about AVP, OXT, and their receptors.

Central Nervous System Distribution

The distribution of AVPR1a expression within the CNS has been studied primarily using receptor autoradiography and hybridization histochemistry. Receptor autoradiography identifies the locations of binding to the receptor protein, whereas hybridization histochemistry

Table 5 Gene information about vasopressin and oxytocin receptors

Gene	*Species*	*Official gene symbol*	*Chromosome*	*Known exons*	*Amino acids*	*Locus ID*[a]
Vasopressin 1a receptor	Human	AVPR1a	12 (q14–q15)	2	418	552
	Rat	AVPR1a	7 (q21)	2	424	25 107
	Mouse	AVPR1a	10 D3 (122.1cM)	2	423	54 140
Vasopressin 1b receptor	Human	AVPR1B	1 (q32)	3	424	553
	Rat	AVPR1b	13 (q13)	3	425	29 462
	Mouse	AVPR1b	1 E4 (131.5cm)	3	421	26 361
Vasopressin 2 receptor	Human	AVPR2	X (q28)	3	371	554
	Rat	AVPR12	X (q37)	3	371	25 108
	Mouse	AVPR1r2	X (29.52 cM)	3	371	12 000
Oxytocin receptor	Human	OXTR	3 (p25)	4	389	5021
	Rat	OXTR	4 (q42)	4	388	25 342
	Mouse	OXTR	6 E3 (113cM)	4	388	18 430
Dual AII/VP receptor	Human	NALP6	11 (p15)	8	513	17 1389
	Rat	NALP6	1 (q41)	7	483	17 1390
	Mouse	NALP6	7 F4	7	490	10 1613
VACM-1 receptor[b]	Human	CUL5	11 (q22–q23)	19	780	8065
	Rat	CUL5	8 (q24)	18	780	64 624
	Mouse	CUL5	9 C (53.8cM)	19	780	75 717

[a]Locus ID from http://www.ncbi.nlm.nih.gov.
[b]Cullin-5.
AII/VP, angiotensin II/vasopressin; VACM, vasopressin-activated calcium-mobilizing.
Adapted from Caldwell HK and Young WS III (2006) Oxytocin and vasopressin: Genetics and behavioral implication. In: Lim R (ed.) *Handbook of Neurochemistry and Molecular Neurobiology,* 3rd edn., 573–607. New York: Springer.

identifies the cells that transcribe the receptor gene. The former technique was greatly facilitated through the use of specific and potent ^{125}I-labeled AVPR1a antagonists. Prominent AVPR1a binding is present in the rat lateral septum (LS), neocortrical layer IV, hippocampal formation, amygdalostriatal area, bed nucleus of the stria terminalis (BNST), various hypothalamic areas (including the suprachiasmatic nucleus (SCN)), ventral tegmental area, substantia nigra, superior colliculus, dorsal raphe, nucleus of the solitary tract, and inferior olive. AVPR1a binding is moderate throughout the spinal cord, but it has higher binding in the dorsolateral motor neurons in general and all motor neurons in the lumbar 5/6 levels where innervation to the perineal muscles originates.

Neurons containing AVPR1a transcripts are found extensively throughout the rat CNS, being especially prominent, for example, in the olfactory bulb, hippocampal formation, LS, SCN, PVN, anterior hypothalamic area, arcuate nucleus, lateral habenula, ventral tegmental area, substantia nigra (pars compacta), superior colliculus, raphe nuclei, locus coeruleus, inferior olive, area postrema, and nucleus of the solitary tract. Transcripts are also detected in the choroid plexus and endothelial cells. The distributions of AVP (and OXT) binding have been examined in a number of rodent species, and they are remarkably similar. Differences in binding in selected areas may mediate important adaptations or behavioral traits, however; these differences are mentioned in the article when correlations are possible.

The AVPR1b was originally described in the anterior pituitary, where it facilitates the release of adrenocorticotropin (ACTH) from the corticotropes. AVPR1b in the pituitary helps mediate the effects of AVP on the hypothalamic–pituitary–adrenal axis, which is the regulator of the stress response in mammals. AVPR1b mRNA is also found in a variety of peripheral tissues, including the kidney, thymus, heart, lung, spleen, uterus, and breast, although its role in these tissues remains unclear. Only recently have AVPR1b transcripts, as well as AVPR1b immunoreactive cell bodies, been found in the rat brain, including in the olfactory bulb, piriform cortical layer II, septum, cerebral cortex, hippocampus, PVN, SCN, cerebellum, and red nucleus. A more recent study with probes with minimal cross-hybridization potential with the OXTR or AVPR1a transcripts indicates a more restricted presence in the mouse and rat brains, being especially prominent in the CA2 field pyramidal neurons of the dorsal hippocampus. It should be noted, however, that AVPR1b distribution has not been mapped by receptor autoradiography due to the lack of a specific radiolabeled ligand.

Some very provocative work has focused on interspecies and individual variations in the AVPR1a promoter of voles. Across vole species, there are profound differences in social structure and behavior. Prairie voles

(*Microtus ochrogaster*) and pine voles (*Microtus pinetorum*) tend to be social and monogamous, whereas montane voles (*Microtus montanus*) and meadow voles (*Microtus pennsylvanicus*) tend to be asocial and polygamous. These social behaviors are mediated, at least in part, by AVP action on the AVPR1a. There are striking differences in the distribution and density of AVPR1a among these species, and they are postulated to regulate this behavior.

Initially, the location of OXTR expression was determined by receptor autoradiography, using a potent and specific ^{125}I-labeled antagonist. In the rat, OXT binding is found in numerous regions, especially in the hippocampal formation (ventral subiculum particularly), LS, central amygdala (CeA), olfactory tubercle, accumbens nucleus shell, dorsal caudate putamen, BNST, medial amygdala (MeA), and ventromedial hypothalamus (VMH). Binding in the spinal cord is light and confined to the superficial dorsal horn.

Hybridization histochemistry reveals OXTR transcripts in many areas of the rat CNS, including the main and accessory olfactory bulbs, neocortical layers II and III, piriform cortical layer II, hippocampal formation, olfactory tubercle, BNST, medial habenula, VMH, PVN, and SON. Expression is lower in the midbrain, pons, and medulla. Recently, an OXTR-LacZ reporter mouse showed additional OXTR gene expression in the medial septum, parts of the amygdala and mammillary nuclei, and some brain stem nuclei.

The distribution of the OXTR is highly species specific, as is elegantly illustrated in receptor binding differences between two closely related species of voles, the polygamous montane vole and the monogamous prairie vole. Differences in the distributions of the OXTR have also been shown among mice, rats, voles, hamsters, and guinea pigs. These differences in OXTR distribution across species are thought to convey differing behavioral phenotypes. For example, variation in accumbens nucleus OXTR appears to be responsible for spontaneous maternal behavior in virgin female prairie voles.

Experimental Approaches

New pharmacological agents as well as new transgenic and knockout agents promise to expand our understanding of roles of AVP and OXT in the brain. Pharmaceutical companies are devising new agonists and antagonists, including orally active ones, for both medical and research purposes. The linear AVPR1a antagonist HO-LVA has proved to be a useful radioligand for receptor autoradiograhic studies and the orally active AVPR1a antagonist SRX251 shows promise as a serenic by reducing aggression. The AVPR1b antagonist SSR149415 is being investigated as an agent potentially useful in reducing anxiety and depression. It has also been used to confirm the role of the AVPR1b in aggression indicated through knockout studies. Orally active OXTR agents are undergoing studies to determine their value in parturition, although potential behavior consequences will have to be closely followed (e.g., maternal behavior).

Transgenic approaches, especially with knockout lines, are proved to be quite useful. L Young and colleagues showed that the expression of AVPR1a from the monogamous vole in a similar pattern in the mouse produces an increased affiliative response, and they suggest that changing the distribution of a receptor may be an efficient means of changing behavior evolutionarily.

The least well understood of the OXTR and AVPRs, the AVPR1b, was the first one inactivated by homologous recombination in mouse. This mouse line shows decreased social aggression, social recognition, and social motivation. This receptor is abundant on adrenocorticotropes in the pituitary, so it is not surprising that knocking out this gene also blunts the hypothalamic–pituitary–adrenal response to stress. They have also been reported to display a disruption of an animal behavior that is abnormal in schizophrenics, the prepulse inhibition of the startle reflex. These mice also have reduced insulin release from isolated pancreatic islet cells, but fasting blood glucose measurements are normal.

The behavioral effects of the AVPR1a have proved variable between sexes and among labs. The same line (with a different background) showed reduced anxiety-like behavior in males only, as well as reduced social recognition, in one lab, whereas another lab found those behaviors to be normal but observed deficits in olfaction and circadian rhythms. Mice without this receptor have also been shown to have impaired cardiovascular function with reduced blood pressure and baroreceptor reflexes, as well as impaired spatial memory.

Finally, the OXTR knockout has been knocked out. As was seen with the OXT knockouts, female mice were able to deliver their young but not lactate; they also showed poorer nurturing. Adult OXTR knockout males showed decreased social recognition and discrimination and increased aggression, as well.

See also: Aggression: Neurochemical and Molecular Mechanisms; Magnocellular Neurosecretory System: Organization, Plasticity, Model Peptidergic Neurons; Mammalian Neuropeptide Families; Neuroendocrine Control: Maternal Behavior; Neuropeptides: Discovery; Oxytocin (Peripheral/Central Actions and their Regulation).

Further Reading

Bielsky IF, Hu SB, Szegda KL, Westphal H, and Young LJ (2004) Profound impairment in social recognition and reduction in anxiety-like behavior in vasopressin V1a receptor knockout mice. *Neuropsychopharmacology* 29: 483–493.

Burbach JPH, Luckman SM, Murphy D, and Gainer H (2001) Gene regulation in the magnocellular hypothalamo-neurohypophysial system. *Physiological Reviews* 81: 1197–1267.

Burbach JPH, Young LJ, and Russell JA (2006) Oxytocin: Synthesis, secretion and reproductive functions. In: Neill JD (ed.) *Knobil and Neill's Physiology of Reproduction,* 3rd edn., pp. 1–74. New York: Elsevier.

Caldwell HK and Young WS III (2006) Oxytocin and vasopressin: Genetics and behavioral implications. In: Lim R (ed.) *Handbook of Neurochemistry and Molecular Neurobiology,* 3rd edn., pp. 573–607. New York: Springer.

Ferguson JN, Aldag JM, Insel TR, and Young LJ (2001) Oxytocin in the medial amygdala is essential for social recognition in the mouse. *Journal of Neuroscience* 21: 8278–8285.

Ferris CF, Albers HE, Wesolowski SM, Goldman BD, and Luman SE (1984) Vasopressin injected into the hypothalamus triggers a complex stereotypic behavior in golden hamsters. *Science* 224: 521–523.

Insel TR, Young L, and Wang Z (1997) Central oxytocin and reproductive behaviours. *Review of Reproduction* 2: 28–37.

Itoh S, Yamada S, Mori T, et al. (2006) Attenuated stress-induced catecholamine release in mice lacking the vasopressin V1b receptor. *American Journal of Physiology, Endocrinology and Metabolism* 291: E147–E151.

Lim MM and Young LJ (2006) Neuropeptidergic regulation of affiliative behavior and social bonding in animals. *Hormones and Behavior* 50: 506–517.

Morris JF and Ludwig M (2004) Magnocellular dendrites: Prototypic receiver/transmitters. *Journal of Neuroendocrinology* 16: 403–408.

Sladek CD (1999) Antidiuretic hormone: Synthesis and release. In: Fray JCS (ed.) *Handbook of Physiology, Section 7: The Endocrine System, Vol. 3: Endocrine Regulation of Water & Electrolyte Balance*, pp. 436–495. Oxford: Oxford University Press.

Venkatesh B, Si-Ho S-L, Murphy D, and Brenner S (1987) Transgenic rats reveal a remarkable functional conservation of regulatory controls between the fish isotocin and rat oxytocin genes. *Proceedings of the National Academy of Sciences of the United States of America* 94: 12462–12466.

Wersinger SR, Ginns EI, O'Carroll AM, Lolait SJ, and Young WS III (2002) Vasopressin V1b receptor knockout reduces aggressive behavior in male mice. *Molecular Psychiatry* 7: 975–984.

Wersinger SR, Kelliher KR, Zufall F, Lolait SJ, O'Carroll A, and Young WS III (2004) Social motivation is reduced in vasopressin 1b receptor null mice despite normal performance in an olfactory discrimination task. *Hormones and Behavior* 46: 638–645.

Zingg HH and Laporte SA (2003) The oxytocin receptor. *Trends in Endocrinology and Metabolism* 14: 222–227.

Relevant Website

http://www.ncbi.nlm.nih.gov – National Center for Biotechnology Information, see Locus Link.

Vegetative State

B Kotchoubey, University of Tübingen, Tübingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Definition

Vegetative state (VS) is a state of consciousness characterized by preserved sleep–wakefulness cycle and intact subcortical functions, on the one hand, and the complete lack of subjective awareness, on the other hand. The term 'complete lack' is important, meaning that there is not the slightest feeling such as pleasure versus unpleasantness, etc. Even a minimal sign of any subjective experience or mental process excludes the diagnosis of VS, indicating that the individual is in another state, such as a minimally conscious state (MCS).

The same pathological condition was formerly referred to as 'apallic syndrome,' from the Greek *a-pallic*, meaning 'without cortex.' The notion of apallic syndrome could be directly tested by examining the morphology of cortical damage and the level of the disturbance of cortical functions. As discussed later, VS patients are not apallic; that is, their cortical functions may or may not be lacking. (The opposite is also true: Children born decorticated are not necessarily vegetative.) Hence, the term apallic has been abandoned. Since subjective experience, in contrast to cortical functions, cannot be observed, the notion of VS cannot be directly tested in a similar way.

Therefore, VS is a construct in need of operationalization. Due to this intrinsic problem of addressing first-person events (i.e., subjective awareness or the lack thereof), some authors criticize the notion of VS as inconsistent. The accepted criteria include the following: no evidence of awareness of oneself or one's environment; no evidence of reproducible, purposeful, or voluntary responses; no evidence of language comprehension; and no volitional response to stimuli. This implies that each diagnosis of VS is a philosophical enterprise because the physician should be able to define such concepts as 'purposeful,' 'voluntary,' or 'volitional.' Therefore, the diagnostic process in VS differs considerably from what is usual in neurology. Thus, a chief neurologist can say to his assistant, "go to patient X and check her tendon reflexes," but he cannot say "go to patient Y and check her self-awareness" or "go to patient Z and check if her will is free."

Another problem is that the main diagnostic criteria of VS are negative. What determines the diagnosis is the lack of a certain kind of responses. There is hardly any positive symptom of VS, except the periods of wakefulness which distinguish it from coma.

Occurrence, Etiology, and Time Course

The prevalence of VS in industrial countries is estimated to be 10–100 per 1 million, depending on the inclusion criteria, the level of the development of intensive care, and the frequency of severe brain injuries.

Usually, VS develops from an acute coma; a primary VS (e.g., in dementia) is much less frequent. The most common causes for such a comatose state, later leading to VS, are mechanical head injury, intracranial bleeding (mainly subarachnoidal hemorrhage), and brain anoxia (mainly after cardiac arrest). Approximately 10–15% of all severe (>6 h) coma patients remain vegetative.

VS lasting for more than 1 month is defined as persistent VS (PVS). The Multi-Society Task Force proposed in 1994 the term 'permanent' for VS that persists longer than 1 year or 3 months for traumatic and nontraumatic PVS, respectively. Most recent authors, however, agree that this term should be avoided because it confuses diagnosis and prognosis. In fact, the chances of regaining consciousness after this time (i.e., 1 year for head injuries and 3 months for other causes) are rather low but not equal to zero. Scientific journals have published several well-documented cases of improvement after 15–20 months and even 5 years in PVS. A case of a considerable restoration of functions after 19 years in MCS has been documented and thoroughly investigated. Although most commentators suggest that no such late improvement is possible in PVS, the exact shape of the function relating improvement probability to time, and particularly the point at which this function approaches zero, remains unknown.

Some factors which this function depends on can, nevertheless, be determined. Favorable prognostic factors are young age; traumatic compared to nontraumatic etiology; and a moderate extent of the axonal injury, particularly in the corpus callosum and the dorsolateral brain stem.

Most patients emerging from a prolonged VS remain considerably disabled with regard to their motor (e.g., paralysis) or cognitive functions (e.g., amnesia) or both. As a rule, the later the recovery, the more severe disability remains. This does not mean, however, that a life at this level of disability cannot be of value.

Neuropathology

The morphology of the lesion in traumatic VS involves multiple bleedings in different brain regions

and severe diffuse axonal injury (DAI) leading to the break of interneuronal connections. The most severe DAI is observed in the corpus callosum, upper brain stem (pontine tegmentum and cerebral peduncles), frontotemporal white matter, and the corona radiata. Damage to gray matter is mostly concentrated in the thalamus and basal ganglia, whereas the cortical gray matter can be largely preserved.

The neuropathology of nontraumatic VS is less investigated. Unlike traumatic VS, severe laminar necrosis with considerable neuronal loss in cerebral hemispheres and the hippocampus is depicted, whereas brain stem structures are usually intact. Like traumatic VS, however, thalamic lesions are also frequently found in anoxic and ischemic VS. In some patients, the damage is limited to the thalamus and basal ganglia, with only minimal neocortical lesions.

Therapy

There is no 'gold standard' of therapy in VS that would satisfy the strict criteria of evidence-based medicine. Several methods have been proposed which can roughly be subdivided into four categories. Pharmacological therapy mainly uses activating substances such as tricyclic antidepressants or methylphenidate. Promising results have been reported with dopaminergic drugs, particularly amantadine. Dopaminergic medications such as L-DOPA are especially effective in VS patients with concomitant parkinsonian symptoms (tremor, rigidity, and akinesia) and the corresponding lesions in the dopaminergic pathway. Unfortunately, such patients constitute a relative small subsample of VS patients. Pharmacological interventions appear to be more efficient during the early stage of the disease.

Invasive treatment includes peripheral nerve stimulation (with implanted electrodes or transcutaneously) and deep brain stimulation through electrodes implanted in the medial thalamus or, rarely, the mesencephalic reticular formation. Physical therapies are usually of palliative character and directed at improving posture, preventing contractures, and obtaining a general increase in arousal level. Finally, a very broad range of stimulation techniques includes unimodal and multimodal sensory stimulation, sensory regulation, music and musicokinetic therapy, and social–tactile interaction.

In general, most treatment studies published to date are poorly controlled, many of them being single-case observations. This makes definite conclusions impossible. The first randomized controlled trial of amantadine versus placebo is under way.

Electrophysiological Measures of Cortical Activity

General patterns of the EEG vary widely among VS patients, from a consistent α rhythm to diffuse δ waves or (exceptionally) low-voltage or even isoelectric EEG. Most commonly, slow rhythms of 4–8 Hz dominate; usually they are not blocked by light. This background activity strongly correlates with event-related potentials (ERPs) reflecting cortical processing of stimulus information. Among patients with the predominant 4–8 Hz ('benign EEG'), virtually all have cortical responses to auditory stimuli (vertex N1 component of the ERP), and many possess late ERP components. Moreover, N1 habituates with stimulus repetition, indicating a simple learning process (learning to not respond to insignificant stimuli) in the primary auditory cortex. The late components are still more frequent in patients with reactive α or θ rhythm. On the other hand, among patients with pronounced δ activity or low voltage ('malign EEG'), only a few have a vertex N1 response, and the components later than N1 rarely occur.

One of the most important late ERP components is the parietal P300, a typical brain response to target stimuli. In VS, as in acute coma, a reliable P300 can be obtained in approximately one-third of patients with benign EEG. Another late ERP component frequently observed in VS is mismatch negativity (MMN), which indicates the ability of the auditory cortex to tonal differentiation. Interestingly, the MMN is obtained in VS patients significantly more often, and with a significantly larger amplitude, in response to complex harmonic tones than simple sine tones – a finding which suggests that testing unconscious patients with very simple stimuli may lead to an erroneous conclusion of their cortical 'nonresponsivity.' Two independent prospective studies found that VS patients with an MMN more frequently emerge from the VS 1–6 months after the ERP examination than patients without it. This concurs with the predictive role of the MMN also demonstrated in acute coma.

An ERP sign of semantic processing is the late negative wave N400, which is recorded in healthy individuals to all meaningful words and whose amplitude is shown to be a strong inverse function of the *a priori* probability of the word in the given context. The N400 to semantically inappropriate words (context mismatch) can also be obtained in some VS patients, indicating the ability of the patients' brain to analyze word meaning.

Metabolic Measures of Brain Activity

The average rate of brain metabolism in VS patients is estimated to be less than 50% of that in healthy

controls. This decrease is even stronger than in acute coma and nearly equal to that in profound anesthesia. The metabolic rate is particularly low in the cortex; in subcortical structures it can be higher than in coma but still much lower than in controls. Unfortunately, from the standpoint of diagnostics and research, the large average difference between VS and conscious patients is countereffected by the considerable variance. Some VS patients have a metabolic rate higher than that of patients with intact consciousness, and some healthy students exhibit metabolic activity as low as that in VS.

Also very large is the variance between the metabolic rates in different cortical areas. Many VS patients with extremely low average metabolism show isolated cortical regions with normal or heightened activity. These regions of local activity vary from patient to patient and, most interestingly, appear to correspond to patients' behavior (e.g., a patient with a high metabolic rate in the speech areas frequently uttered single words).

Not only can cortical areas in VS patients be metabolically active but also their activity can be modified by appropriate stimulation. Single case studies in the late 1990s showed that local metabolic activity increases in response to meaningful stimuli (human voice and faces) compared with meaningless stimuli (sounds and chaotic visual patterns). Later studies with larger samples indicated that this activity is usually limited to primary sensory areas, whereas in MCS patients it extends to broader areas comparable to those activated in controls. A connectivity analysis showed that strong positive connections between different brain regions typical for healthy subjects can be lacking or even become negative in VS patients. There are also indications that regaining consciousness is accompanied by restoration of the normal connectivity. On the other hand, this restriction of activation to the primary cortical areas may be a function of simple stimulation because a few studies employing more complex stimuli suggest that higher-level cortical regions (e.g., the fusiform gyrus) can also be activated in VS patients.

To summarize, the results of electrophysiological and metabolic studies are in good agreement. Both demonstrate that large regions of cortical activity can remain intact in VS patients, indicating that at least in some of them the cortex is able to perform a complex analysis of stimulus information including an analysis of word meaning. A possible disagreement concerns major qualitative differences between patient subgroups. Whereas in metabolic studies the largest difference tends to coincide with the clinical distinction between VS and MCS, electrophysiological studies frequently find larger differences in cortical reactivity within the groups of VS and MCS patients. Neither finding directly indicates the presence of conscious awareness because cortical processing, even at the semantic level, can occur subconsciously.

Vegetative State and Neural Correlates of Consciousness

After several decades of ignoring, the modern psychology and neuroscience are presently actively interested in the problem of human (and animal) consciousness. From an analytical perspective, conscious phenomena are frequently subdivided into dynamic aspects versus content aspects, wakefulness versus awareness, or nontransitive consciousness('just consciousness') versus transitive consciousness ('consciousness of'). The transitive consciousness is further subdivided into simple sentience (pain, pleasure, sensitivity to colors, sounds, etc.) and self-awareness (including self as the agent of actions, thinking, etc.). From this point of view, VS presents the model of dissociation between the transitive and nontransitive consciousness: By definition, patients are awake but not aware of anything. Investigations of VS can therefore clarify the issue of neural correlates of consciousness (NCC). Namely, because VS patients do not possess subjective experience, any brain activity found in these patients can be eliminated from the list of possible NCC, and any brain region remaining intact in VS is therefore not a seat of subjective awareness. Conversely, those brain processes and brain activations which are lacking in VS but present in patients with preserved elementary consciousness (e.g., MCS) are candidates for the role of NCC.

This unbeatable logic, however, is based on two important assumptions: (1) that the definition of the VS is consistent, and (2) that the rate of diagnostic errors is low enough to neglect, when comparing VS and MCS groups, the possibility of some patients being incorrectly assigned. As stated previously, the former assumption is regarded by many authors as problematic and the latter is blatantly wrong. Thus, VS research cannot directly identify brain mechanisms of awareness, but it can yield useful hints which should be compared with other indications based on *a priori* or *a posteriori* criteria.

A priori criteria are such that if a being fulfills them, everybody will believe that the being is conscious. An example of such a criterion is that the being's responses follow our complex verbal instruction. The problem with these criteria is their high conservatism. Thus, the ability to follow verbal instructions proves not sentience but self-awareness and concentrated attention. *A posteriori* criteria are those we develop from the results of empirical studies of consciousness beyond

the VS domain (e.g., studies of sleep, anesthesia, and brain responses to subliminal stimuli).

Three completely different cases must be separated when we discuss the issue of 'conscious awareness in VS.' First, it is possible that patients are conscious just because they are misdiagnosed; they are not in VS. Second, patients correctly diagnosed as VS can demonstrate signs of consciousness because they are examined at the beginning of recovery. Finally, it cannot be excluded that some day we will have to modify our definition of VS in such a way that admits some VS patients possessing elements of subjective awareness. Although these possibilities may be difficult to disentangle in some patients, they should at least be clearly distinguished at the logical level. The main difference is that in the last two cases, only elementary consciousness (e.g., pain or pleasure) can be present, whereas in the first case (diagnostic error) a patient may even possess high levels of reflexive awareness.

An example of the confusion in this domain is a study, broadly commented on in scientific journals and magazines, in which VS patients were told to imagine either playing tennis or walking through their apartment. Surprisingly, one patient demonstrated distinct patterns of brain activation (recorded with functional MRI) over precentral and parietal regions during the two instructions, respectively. Consistently following verbal instructions during several trials requires high concentration. Even subjects with slightly disturbed consciousness (e.g., healthy but extremely tired individuals) would not be able to produce such a result. Therefore, the patient in question was obviously misdiagnosed; perhaps she was in a locked-in state. Nonetheless, all commentaries ignored this simple fact and claimed that the data 'reveal consciousness in VS.'

Several theoretical attempts have been undertaken to integrate VS findings into the broader context of the present-day *a posteriori* evidence of NCC. These attempts are summarized in the following hypotheses:

1. Consciousness requires, at least, the activity of secondary sensory and associative regions of the cortex. Accordingly, VS is characterized by the lack of activity in these 'higher-level' areas despite the preserved activity in the primary sensory and motor areas.
2. The substrate of consciousness is not particular brain regions but a widely distributed network of synchronized activity across many cortical regions. Accordingly, VS is primarily a disconnection syndrome in which various cortical areas and local processing modules remain working, but their integration is lacking.
3. Consciousness depends on the functioning network linking the medial parietal structures (posterior cingulum and precuneus) with the thalamus and prefrontal cortex.

The three hypotheses are partially, but not completely, compatible with each other; each can be supported by many, but not all, empirical findings. Thus, a frequently cited PVS patient, consistently demonstrating both behavioral and neurophysiological evidence for preserved activity in Broca and Wernicke areas, is a case for hypothesis 2 but against hypothesis 1. Possibly, each hypothesis explains only a subset of all VS patients, and the subsets covered by different hypotheses partially overlap.

The Problem of Pain Sensation

The postulate that vegetative patients cannot experience pain and suffer is particularly controversial since numerous ethical issues depend on it. The empirical basis of this controversy is rather narrow. Modern neuroimaging studies indicate that pain perception is associated with the activity in several cortical areas, including the primary and secondary sensorimotor cortex, insula, operculum, the anterior cingulate gyrus, and possibly also the supplementary motor area and the prefrontal cortex. These areas are not typically damaged in VS. On the other hand, the subcortical structure most closely related to pain sensation is the thalamus, and thalamic lesions are very frequent in VS. Therefore, one may assume that the thalamic pathology is critical for the presumable loss of pain experience in VS. To the author's knowledge, two studies directly addressed this issue by recording changes in brain metabolism during painful stimulation. One of them (with 15 VS patients) found significant responses to pain stimuli in each patient, comparable with the corresponding responses of control subjects, in the thalamus and the primary sensorimotor cortex but not in other pain-related structures. In the other study, in contrast, seven VS patients did not exhibit any response to pain in the thalamus; however, significant responses were obtained in the primary and secondary somatosensory cortex, posterior insula, postcentral gyrus, and the cingulate cortex. The issue of possible pain sensation in VS is thus still open; in any case, as the authors of the latter study concluded, "a residual pain-related cerebral network remains active in long-term PVS patients."

Ethical Issues

MEDLINE lists 423 publications on VS between January 2005 and March 2007; 215 of them (51%) were devoted to ethical aspects of the syndrome,

compared to less than 2% that concerned the neuropathology of VS. These data demonstrate how important the ethical considerations on VS are publicly regarded.

A common mistake is that the ethical discussion is being carried out between particular institutions – in the simplest version, 'the modern science' and 'the Church,' each defending a particular ethical position. The opposite is true: Almost every position is represented within each institution participating in the discussion. Even within the Catholic Church, with its tradition of subordination to authorities, opinions of various theologians and bishops vary widely. This holds true even more so for other congregations, such as Protestants or Jews, for which no Vatican-like institution exists.

The most important ethical issue concerns the possibility and necessity of life-prolonging measures, such as artificial nutrition and hydration (ANH), in chronic PVS. If ANH is a medical intervention (as many, but not all, medical and religious authorities believe), it can and in some cases even should be withheld or withdrawn. If it is a part of simple care, like feeding (the view held, for example, by Pope John Paul II), its withholding or withdrawal can be regarded as killing by omission. Another aspect concerns the chances of VS patients to regain consciousness. Whenever these chances are sufficiently high, all life-prolonging interventions appear ethically justifiable. Unfortunately, all empirical statements are probabilistic; that is, they describe populations and not individual patients. Thus, even if we knew that the chance of clinical improvement for a given PVS patient is, for example, 1 in 100, this would not prevent the family of the patient from hoping that 'their patient' is that 1 in 100.

The third aspect of the problem of ADH withdrawal relates to the alleged inability of VS patients to suffer. Obviously, the claim that a patient is completely unable to experience any aversive sensation is very strong and difficult to prove. However, any level of uncertainty in this question is unacceptable because if only a slight subjective experience is possible, the patient would die in tantalizing agony. Only a few decades ago, many believed that newborns could undergo surgery without anesthesia because they were unable to experience pain.

A different, though related, ethical issue is not about treatment and care but, rather, about scientific investigation of VS patients. An important ethical rule states that patients unable to give informed consent may only participate in experiments if they can themselves benefit from the participation. If, however, the results can be used not for the participants but for other patients, the study is unethical because it implies that some humans are exploited, without their consent, for the sake of other humans. Another rule states that a decision under uncertainty should be based on the best available knowledge, even if it is not exact. Both rules appear reasonable, but in combination they lead to a paradox. According to the second rule, we may withdraw ANH even if we are not completely sure that the patients would not suffer. According to the first rule, however, we cannot perform experiments to clarify the question of whether they do not suffer because such experiments would use aversive stimuli applied on a group of patients, whereas the results would probably be useful only for the next generation of patients. Thus, from the standpoint of formal ethics, letting patients starve may be more ethical than presenting them with slight electrical shocks.

Neuroscience alone cannot resolve these ethical problems because they are based on conceptual questions that are, at least partially, beyond empirical research: What is person? Which value is higher – the life of a person or his or her autonomy? What are the appropriate costs for helping a severely ill patient, and must such a patient be cured at any price, even at the price of depriving many other patients of the therapy they need? However, empirical knowledge is necessary, although not sufficient, for ethical decisions because on the basis of wrong knowledge, even the best intentions can result in increased harm and suffering of other people. This happens because ethics is about means, not about ends. If one's intentions are the best, one can nonetheless be unethical if one use wrong means, and this occurs when one do not know the right means. Therefore, to paraphrase Francis Bacon, knowledge is ethical.

Thus, we should distinguish between the two things. Using VS patients to study general problems (e.g., the nature of consciousness) is unethical unless these studies are beneficial for participating patients. However, research on VS patients to enhance our knowledge about their condition (i.e., VS) is ethical by definition unless the potential harm of the examination outweighs the potential increase in knowledge. Do we now prohibit pain research on VS patients because we believe that weak electrical stimulation can seriously harm them or, rather, because we are afraid to obtain results putting our previous knowledge, and our previous practice, in question?

See also: Awareness: Functional Imaging; Axonal Injury: Neuronal Responses; Axonal Injury in Demyelinating Disease and CNS Injury; Coma and Other Pathological Disorders of Consciousness; Coma; Consciousness: Theoretical and Computational Neuroscience; Neuroethics; The AIM Model of Dreaming, Sleeping, and Waking Consciousness.

Further Reading

Andrews K, Murphy L, Munday R, and Littelwood C (1996) Misdiagnosis of the vegetative state: Retrospective study in a rehabilitation unit. *British Medical Journal* 313: 13–16.

Coleman MR (ed.) (2005) *The Assessment and Rehabilitation of Vegetative and Minimally Conscious Patients* (a special issue of Neurological Rehabilitation). Hove, UK/New York: Psychology Press/Taylor & Francis.

Giacino J and White J (2005) The vegetative and minimally conscious states: Current knowledge and remaining questions. *Journal of Head Trauma Rehabilitation* 20: 30–50.

Jennett B (2002) *The Vegetative State*. Cambridge, UK: Cambridge University Press.

Kotchoubey B (2007) Event-related potentials predict the outcome of the vegetative state. *Clinical Neurophysiology* 118: 477–479.

Kotchoubey B, Lang S, Mezger G, et al. (2005) Information processing in severe disorders of consciousness: Vegetative state and minimally conscious state. *Clinical Neurophysiology* 116: 2441–2453.

Laureys S (ed.) (2005) *The Boundaries of Consciousness: Neurobiology and Neuropathology, Progress in Brain Research,* vol. 150. Amsterdam: Elsevier.

Laureys S, Giacino JT, Schiff ND, Schabus M, and Owen AM (2006) How should functional imaging of patients with disorders of consciousness contribute to their clinical rehabilitation needs? *Current Opinion in Neurology* 19: 520–527.

Laureys S, Owen AM, and Schiff ND (2005) Brain function in coma, vegetative state, and related disorders. *Lancet Neurology* 3: 537–546.

Owen AM, Coleman MR, Boly M, et al. (2006) Detecting awareness in the vegetative state. *Science* 313: 1402.

Owen AM, Menon DK, Johnsrude IS, et al. (2002) Detecting residual cognitive function in persistent vegetative state. *Neurocase* 8(5): 394–403.

Schiff ND, Ribary U, Moreno DR, et al. (2000) Residual cerebral activity and behavioral fragments can remain in the persistently vegetative brain. *Brain* 125: 1210–1234.

Shewmon DA, Holmes GL, and Byrne PA (1999) Consciousness in congenitally decorticate children: Developmental vegetative state as self-fulfilling prophecy. *Developmental Medicine & Child Neurology* 41(6): 364–374.

Zasler ND (2004) Terminology in evolution: Caveats, conundrums, and controversies. *Neurorehabilitation* 19(4): 285–292.

VelociGene and VelociMouse: High-Throughput Approaches for Generating Targeted Mutations in Mice on a Genome-Wide Scale

D M Valenzuela and G D Yancopoulos, Regeneron Pharmaceuticals, Tarrytown, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The sequencing of the human and mouse genomes, which resulted in the identification of thousands of genes whose function had not yet been explored, provided an exciting new challenge for the scientific community. Because engineering-specific gene mutations in mice is perhaps the most powerful way to explore mammalian gene function, major new efforts were initiated with the goal of generating defined mutations for every gene in the mouse genome. Unfortunately, available technologies at this time did not allow for such large-scale efforts. Until this point, engineering mutations in mice was usually carried out for a single gene at a time because each gene mutation required highly customized, laborious, and time-consuming efforts by a highly skilled practitioner. The new initiatives intended to dramatically change the scale at which targeted mouse genetics could be carried out. In the United States, leaders in the field proposed the Knockout Mice Project (KOMP), soon followed by a funding initiative from the National Institutes of Health (NIH). Likewise, European Union scientists proposed and received funding for the European Conditional Mouse Mutagenesis. Additional major initiatives were set forward, such as the North American Conditional Mouse Mutagenesis Project sponsored by Canada. Some of these efforts proposed utilizing random gene trap methodology, which could approach the scale envisioned but which would not allow deliberate targeting and custom engineering of a gene of interest. The ability to generate custom gene mutations in a specific target gene, but scaled to efficiently allow mutation of every gene in the mouse genome, would require development of new technologies.

The deliberate engineering of custom mutations in a specific mouse gene requires the genetic modification of mouse embryonic stem (ES) cells by approaches that utilize the built-in homologous recombination machinery of the mammalian cell. The first step in this process requires construction of a DNA 'targeting vector' that contains the desired genetic alteration, which can then be transferred into the cell via homologous recombination. For example, the deliberate deletion of a gene's entire coding region requires the construction of a targeting vector containing 'homology arms' corresponding to the sequences just upstream and downstream of the desired deletion, with the region to be deleted replaced with a selectable gene (e.g., a Neo gene encoding G418 drug resistance) and perhaps a reporter gene (e.g., β-galactosidase). Construction of the targeting vector has traditionally relied on highly skilled personnel using standard but laborious and time-consuming molecular biology techniques, such as restriction enzyme digestions, ligations, Southern blot hybridizations, and long-range polymerase chain reaction (PCR). After a targeting vector is introduced into ES cells by electroporation, the homologous recombination machinery of the cell matches the targeting vector homology arms with endogenous chromosomal sequences and performs double homologous recombination that results in the replacement of the endogenous target sequence with the introduced sequence in the targeting vector. Correct targeting is a rare event, and a larger proportion of cells incorporate the targeting vector in a nontargeted, random fashion. Both random and correctly targeted cells generate clones in the appropriate selection medium, so the surviving clones need to be genotyped to identify the correctly targeted ones, a process that traditionally relied on highly customized assays individually developed for each targeting experiment. As in the case of targeting vector construction, genotyping methods usually employ labor-intensive molecular technologies such as Southern hybridization and/or long-range PCR. Using these traditional approaches, the generation and screening for targeted mutations in ES cells is a highly customized and laborious process for each gene of interest.

Once an ES cell carrying the desired mutation is generated and identified, production of mice from this ES cell using conventional technologies is also quite tedious and time consuming. Conventional technologies involve injection of the targeted ES cell into the cavity of normal blastocyst-stage embryos ('host embryos'), which results in the generation of mice that are composed of cell lineages derived from both the injected ES cells and the normal host embryo and therefore are chimeric in nature. The resulting chimeras are not useful for phenotypic studies and need to be bred to wild-type animals of appropriate coat color to test for germ line transmission of the correct genotype. Thus, for an average experiment, 6–12 or more chimeras need to be crossed, their progeny genotyped, and the resulting heterozygous mice, if any, crossed to generate homozygous animals in large

enough numbers to be able to perform phenotypic studies. In many cases, the mice need to be crossed even further and for several generations to be able to bring the modified alleles into an appropriate genetic background suitable for phenotyping. In all, the mouse breeding necessary to generate useful cohorts may take from 6 months to several years. The problem is even more acute when 'conditional' knockouts are required or when the effect of two or more genetic modifications in the same mouse needs to be studied – for example, when generating polygenic models or conditional mutations under the control of a regulated gene such as CreERt2. The additional genetic complexity usually requires even further crosses, sometimes using complicated breeding schemes.

Using this traditional albeit time-consuming and cumbersome set of technologies, many targeted mouse lines have been successfully generated and analyzed, including targeted mice for many genes of neurological interest, such as for the neutrotrophic factors nerve growth factor, brain-derived neurotrophic factor, neurotrophin-3, neurotrophin-4, ciliary neurotrophic factor, and glial cell line-derived neurotrophic factor, which provided incredible insights into the normal roles of these gene products in the mouse. However, due to the limitations outlined previously, these technologies are eminently unsuitable for high-throughput production and even the most talent-laden and well-resourced research laboratories could only produce a handful of targeted mouse lines per year.

Thus, in order to meet challenges of the magnitude proposed by the international mouse knockout initiatives, it was necessary to develop transforming technologies that would allow targeted mutations on a genomewide scale in the mouse. VelociGene and VelociMouse represent important new technological approaches, developed in direct response to these challenges, that have resulted in dramatic compression of the time and effort required to generated phenotypable mice and thus have revolutionized the technology landscape for the generation of genetically modified mice. These technologies were selected by the NIH to play a major role in the genomewide KOMP initiative.

VelociGene: High-Throughput Gene Targeting in ES Cells

Just as development of the automated DNA sequencer enabled large-scale efforts that succeeded in sequencing the entire genome, high-throughput approaches were necessary to address the next major challenge of assigning function to the entire set of genes within the genome. In order to overcome the limitations of conventional approaches to gene targeting in ES cells, VelociGene introduced two critical advances: (1) the use of bacterial homologous recombination (BHR) to rapidly create large numbers of large targeting vectors from bacterial artificial chromosomes (BACs) and (2) the use of these BAC-based large targeting vectors to efficiently generate and screen for mouse ES cells carrying the desired mutations.

Generating Large Targeting Vectors

Early advances exploited BHR so as to modify large plasmids, such as BACs, by allowing controlled and transient homologous recombination in BAC-bearing bacteria. This is achieved by introducing a plasmid containing the coding sequences of three λ phage proteins that mediate homologous recombination in *Escherichia coli*, red exo (alpha), red bet (beta), and gam (γ) under the control of the arabinose promoter, allowing transient and highly regulated expression of their encoded recombinogenic proteins (**Figure 1(a)**). Working together, these proteins have the ability to perform homologous recombination between a linear segment of DNA and a circular plasmid such as a BAC containing a segment of mouse chromosomal DNA with the target gene. For this to occur, the linear DNA segment is configured to contain some operative sequences, such as an antibiotic selection marker flanked by homology 'arms' identical to sequences present in the intended BAC target (**Figure 1**). Under appropriate antibiotic selection, only bacteria containing the intended BAC modification (e.g., a deletion of the target gene exon) can be obtained and their DNA used for ES cell targeting. In this way, VelociGene taps into a powerful BHR mechanism to generate large targeting vectors in a single overnight reaction, without the need for extensive, multistep, skill-intense molecular biology work.

The advantages of generating BAC-based targeting vectors using BHR are of great significance: BACs have high stability, may propagate up to 300 kb DNA, and can be easily produced and isolated. Thus, a BAC containing a full gene, or an array of multiple genes, can be modified to delete a single nucleotide, a single functionally significant domain, or a full family of genes with essentially the same amount of effort. Also, because of the dramatic simplification of the technical manipulations, many BAC-based targeting vectors could be generated simultaneously in an assembly-line fashion. Thus, standard, premade, off-the-shelf cassettes containing the appropriate reporter gene and selection genes can be prepared for BHR by ligating homology arms (easily made by PCR) and then used as linear molecules to electroporate into recombinogenic bacteria containing the BAC intended

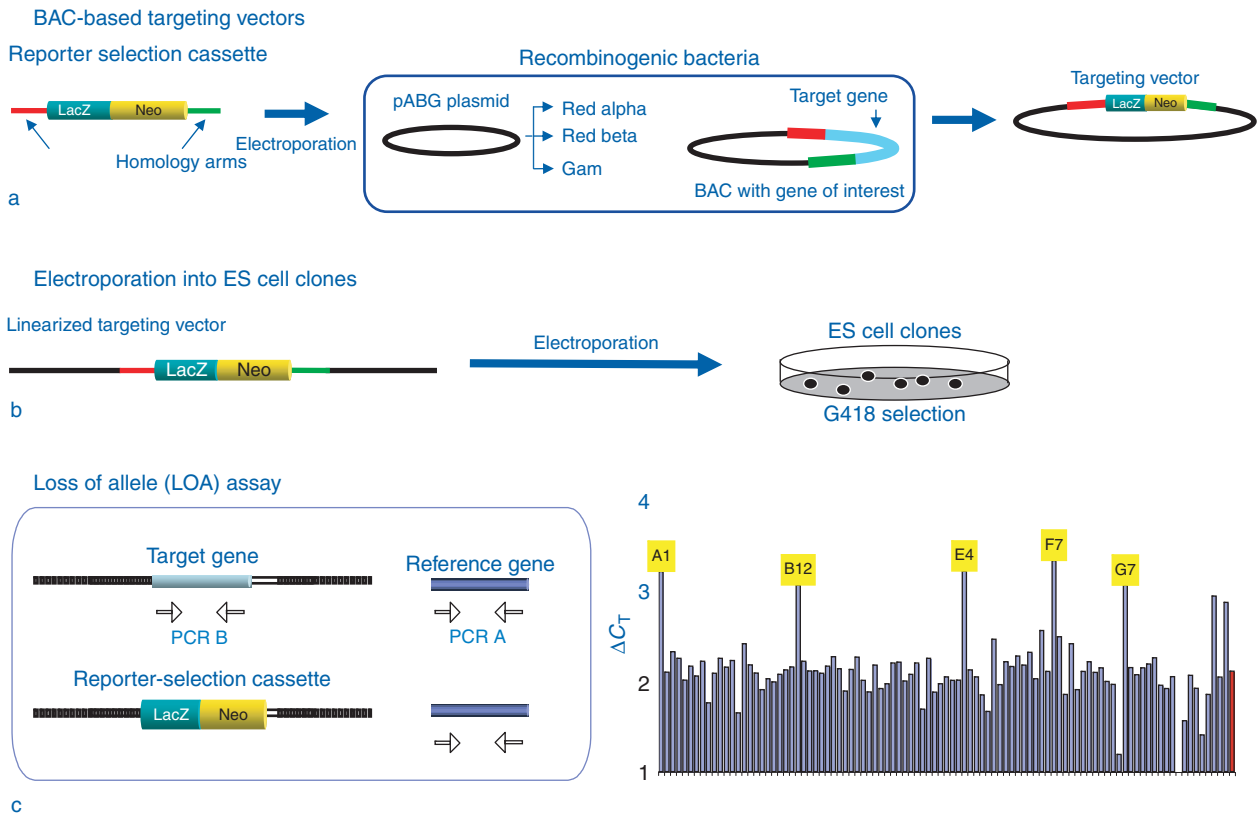

Figure 1 VelociGene vector construction and targeting screen. (a) A reporter selection cassette (in this example containing β-galactosidase (LacZ) and Neo) with homology arms corresponding to flanking areas of the target gene is electroporated into a bacteria containing the target BAC and recombinogenic proteins (red alpha, red beta, and gam). The appropriate antibiotic selection scheme results in the generation of a targeting vector. (b) The targeting vector is then linearized and electroporated into ES cells and the cells are subjected to selection, in this example by using G418. (c) The genomic DNA from resistant ES cell clones are subjected to a quantitative PCR loss-of-allele (LOA) assay which compares the amplification of the target gene to reference genes. Cells which have undergone specific targeting lose one of their targeted alleles with respect to the reference gene, and this is reflected as an increase of one PCR cycle (ΔC_T) compared to the reference genes. The right panel shows a sample LOA screening of a 96-well plate, which shows five identified targeted clones.

for modification. A simple antibiotic selection scheme will result in the production of the desired BAC targeting vector.

However, in order to use these BAC-based, long targeting vectors to modify ES cells, a seemingly intractable problem needed to be resolved: The length of the targeting vector homology arms, sometimes in the hundreds of kilobases, precluded the use of standard tools for ES cell genotyping, such as Southern analysis or long-range PCR. The VelociGene solution was the development of an assay that would rely on a novel loss-of-endogenous allele (LOA) assay.

Using Large Targeting Vectors to Efficiently Target Mouse ES Cells

In order to make possible the use of large targeting vectors for ES cell gene targeting, VelociGene developed an LOA assay which uses real-time quantitative PCR (qPCR) to count the number of intact copies of the targeted allele compared to the number of copies of reference genes (Figure 1(a)). Thus, correctly targeted ES cell clones should necessarily show the loss of one of the targeted alleles, whereas random integrants should retain both copies of their target DNA (screening of X-linked target genes is essentially the same except for an adjustment to the presence of a single targeted allele). The method uses two target-specific primer sets (forward and reverse PCR primers and a qPCR probe that recognizes the PCR product) corresponding to the upstream and downstream end of each targeted deletion (Figure 1((a)) and compares each of these target-specific qPCRs to up to six reference qPCRs for a total of up to 12 assays per construct for each ES clone. A robotic workstation assembles the qPCRs in optical 384-well clear reaction plates and the mixtures are cycled in a qPCR instrument. For each qPCR, the cycler's software determined the threshold cycle (C_T), the point in the PCR at which the fluorescence signal reached the

preset threshold. We calculated the difference in the C_T values (ΔC_T) between the target-specific uPCRs and dPCRs and each of the reference PCRs. Since most of the ES cell clones were not expected to have sustained a targeted loss of allele, we screened for ES cell clones whose ΔC_T values were consistently greater than the median ΔC_T for the population by approximately one cycle, as expected for a clone that has lost one of the two targeted alleles.

This approach was validated by extensive quality control of the ES cells and mice derived from them. The quality control included confirmatory fluorescent *in situ* labeling and analysis of the upstream and downstream areas of the targeted zone. One significant advantage derived from the use of large targeting vectors is a high targeting efficiency (on the order of 5% or more) without the need of using isogenic DNA or positive/negative selection. This method has proved to be highly consistent and with a very low false-positive and failure rate. The LOA assay not only allowed the use of BAC-based vectors for targeting but also made possible parallel processing and automation of the ES cell screening process. ES cell DNA preparation, assay configuration and sample handling, qPCR assay, and data collection and handling are all done either automatically or largely so. These developments allow the assaying of multiple targeting projects simultaneously with a throughput that can vary from several hundred to several thousand targeted ES cell lines a year.

VelociMouse: Rapid Generation of Genetically Modified Mice Suitable for Phenotyping

For functional studies, the next step after the generation of the targeted ES cells is the generation of mice. Following implementation of VelociGene approaches which allowed for the rapid generation of hundreds or thousands of specifically targeted ES cell clones, generation of mice from these ES cells represented the new bottleneck in the mouse genetics field. The most widely used conventional technology to generate mice from targeted ES cells requires microinjection of ES cells into host embryo blastocysts to generate chimeras. One major limitation of this method is that chimeras, by their very nature, are unsuitable for phenotypic studies because they consist of a mixture of tissue lineages derived from both the targeted ES cells and the host embryo cells and thus need to be test crossed to find those chimeras whose genital ridge is sufficiently derived from the targeted ES cells to allow for germ line transmission of the targeted allele. The 'transmitting' mice – that is, those mice producing heterozygous pups – are then used to generate additional heterozygous mice, which in turn are used to generate homozygous cohorts in numbers suitable for phenotyping. Thus, once targeted ES cells are generated, the production of homozygous cohorts requires several months of tedious and usually expensive breeding. The situation becomes even more difficult when studying multiple targeted alleles in the same mice because each needs to be generated individually, as described previously, and then crossed into a single mouse. Thus, using the blastocyst injection technology, mice generation from ES cells is a major bottleneck in the process of gene functional analysis. As the development of the VelociGene technology allowed the parallel processing and industrialization of the mouse ES cell targeting, the mouse generation bottleneck became more acute, making it necessary to search for faster and more economical methods to generate mice suitable for phenotypic studies.

To accelerate the time to an observable phenotype, some investigators use tetraploid embryos as hosts for their ES cells, which can result in generation of F0 mice that are almost completely ES cell derived and thus immediately available for phenotypic analysis. The basis for this method is a particularity of tetraploid cells that allows them to produce only extraembryonic tissues such as the placenta and chorion, whereas the epiblast and embryo proper are primarily derived from the targeted ES cells. Unfortunately, generation of live animals using the tetraploid approach is an inefficient process, the technique requires certain low-passage hybrid strain ES cell lines, and it often fails to yield viable mice when inbred strain ES cells are used. In addition, the resulting F0 mice can possess up to 2% host contamination, exhibit poor viability, and have other abnormalities (e.g., changes in growth rates and body weight) unrelated to the introduced genetic mutation, complicating phenotypic analyses. Although the tetraploid approach is an advance over the blastocyst method in terms of speed to phenotypable mice, its drawbacks limit its usefulness for large-scale functional genomic efforts.

The main technological breakthrough of VelociMouse involves the use of a laser beam to introduce a perforation of the zona pellucida of an eight-cell embryo and then using that perforation to introduce the targeted ES cell into the embryo (**Figure 2**) to efficiently generate viable and healthy F0 generation mice that contain no more than 0.1% host contamination. Most important, the resulting F0 generation mice (derived from ES cells targeted to have heterozygous or homozygous mutations) can directly be used in phenotyping analyses: We demonstrated that phenotypes defined in these F0 mice are

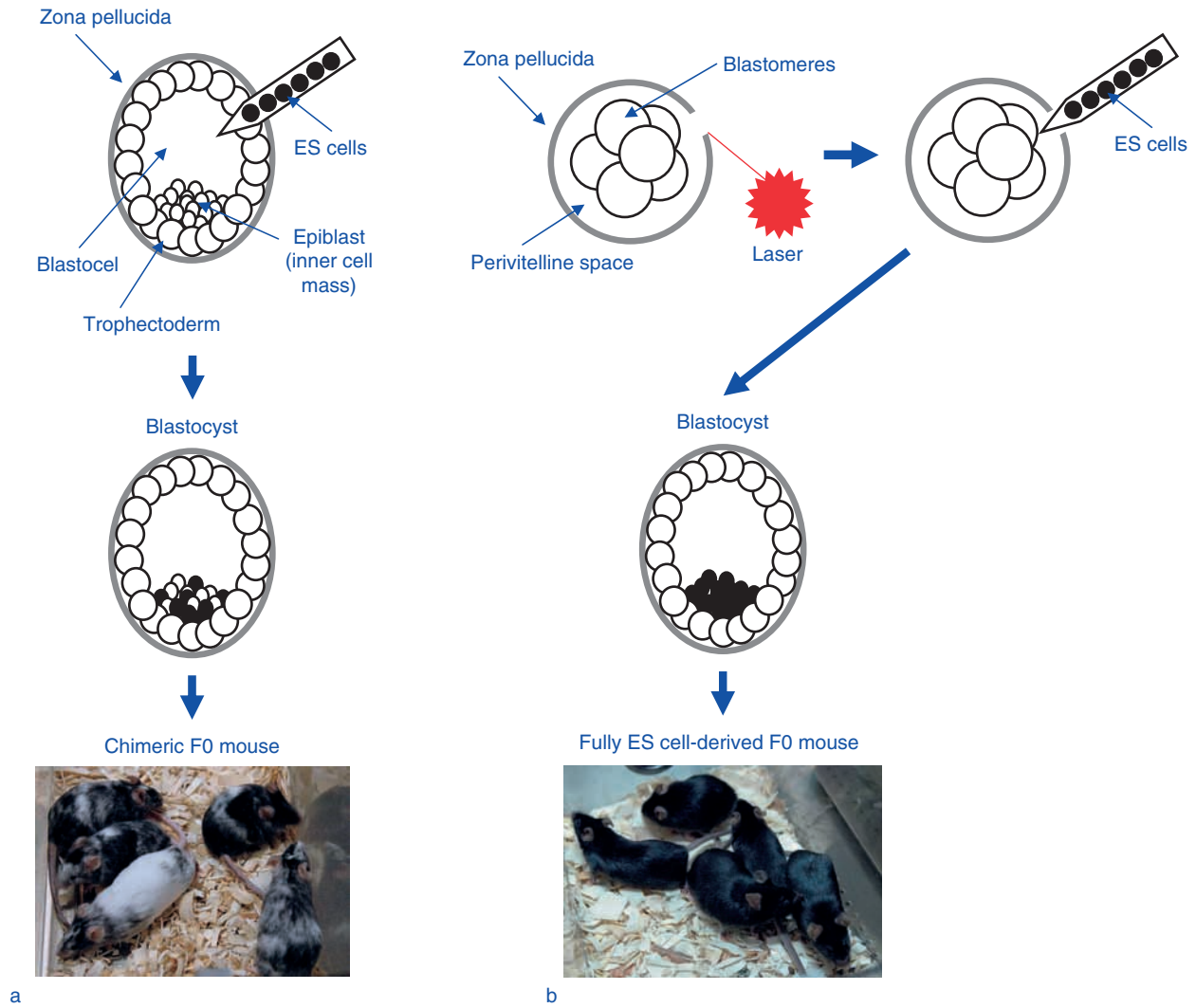

Figure 2 Comparison of conventional and VelociMouse mouse generation technologies. (a) Conventional injection of targeted ES cells into mouse blastocysts results in mixing of the ES cells with the preexisting epiblast, producing mice derived from both the ES cells and the cells from the host embryo (chimeras). In this example, C57BL/6 ES cells were injected into C57BL/6 (albino) host embryos, generating mice with a chimeric black and albino coat color. Chimeras need to be bred to the appropriate mouse genetic background to demonstrate germ line transmission to the next generation. (b) In contrast, VelociMouse technology uses a laser-assisted perforation of the zona pelucida of an eight-cell embryo, allowing for the microinjection of ES cells into the perivitaline space. In this example, C57BL/6 ES cells were injected into C57BL/6 (albino) host eight-cell embryos, resulting in F0 mice which exhibit 100% black coat color that were proven to be fully derived from the microinjected ES cells.

indistinguishable from those observed in conventionally derived F1 and F2 generation mice. Because this new approach greatly accelerates time to mouse phenotyping analyses, we termed it VelociMouse.

VelociGene and VelociMouse Allow for High-Resolution Localization of Gene Expression, Thus Allowing Targeted Phenotyping for Function Elucidation

VelociGene technology allows for the precise deletion of the entire coding region of a gene of interest – the only way to ensure that the targeted allele is definitively null – while simultaneously also allowing precise replacement of the coding region with a reporter gene, such as β-galactosidase or green fluorescent protein. Thus, a single mouse line can be created that can initially be used to provide high-resolution localization of sites of expression of the gene of interest and subsequently used to understand the functional consequences resulting from complete lack of expression of that gene. For the initial expression studies, mice heterozygous for the deletion/reporter allele can be used that still retain a single unperturbed

allele of the gene of interest, and such heterozygous mice often suffer little or no functional implications, allowing elucidation of the normal expression patterns of the gene. For the subsequent functional studies, mice homozygous for the deletion/reporter allele are generated (either by mating heterozygous mice or by generating VelociMice from homozygously targeted ES cells). Using such approaches, we have been able to generate high-resolution reporter profiles for hundreds of genes, allowing for precise localization of individual cells expressing a gene of interest (Figure 3). Such profiles can play a very useful role in guiding subsequent phenotypic analyses of the gene of interest. For example, the realization that a gene is only expressed in a small collection of neurons in the nervous system can allow for very targeted (and very limited) phenotyping of behaviors and functions associated with that neuronal subset. Such targeted phenotyping can save an enormous amount of time and effort that would normally be wasted if each mutant mouse line was indiscriminately subjected to widespread routine phenotyping without regard to the site of action of the gene – for example, wasting effort by performing immune phenotyping studies for a gene expressed only in the nervous system.

Using a combination of reporter profiling and targeted phenotyping, we have efficiently elucidated the function of many genes, dozens of which have already been published. For example, VelociGene targeted mice for the Lhx6 gene were used to delineate a neuronal pathway mediating innate reproductive behaviors from the amygdala to the hypothalamus. Another mouse line targeted for the peripheral myelin protein 22 gene was used to determine that the protein product is complexed with $\alpha_6\beta_4$ integrin, and that the absence of this protein alters the Schwann

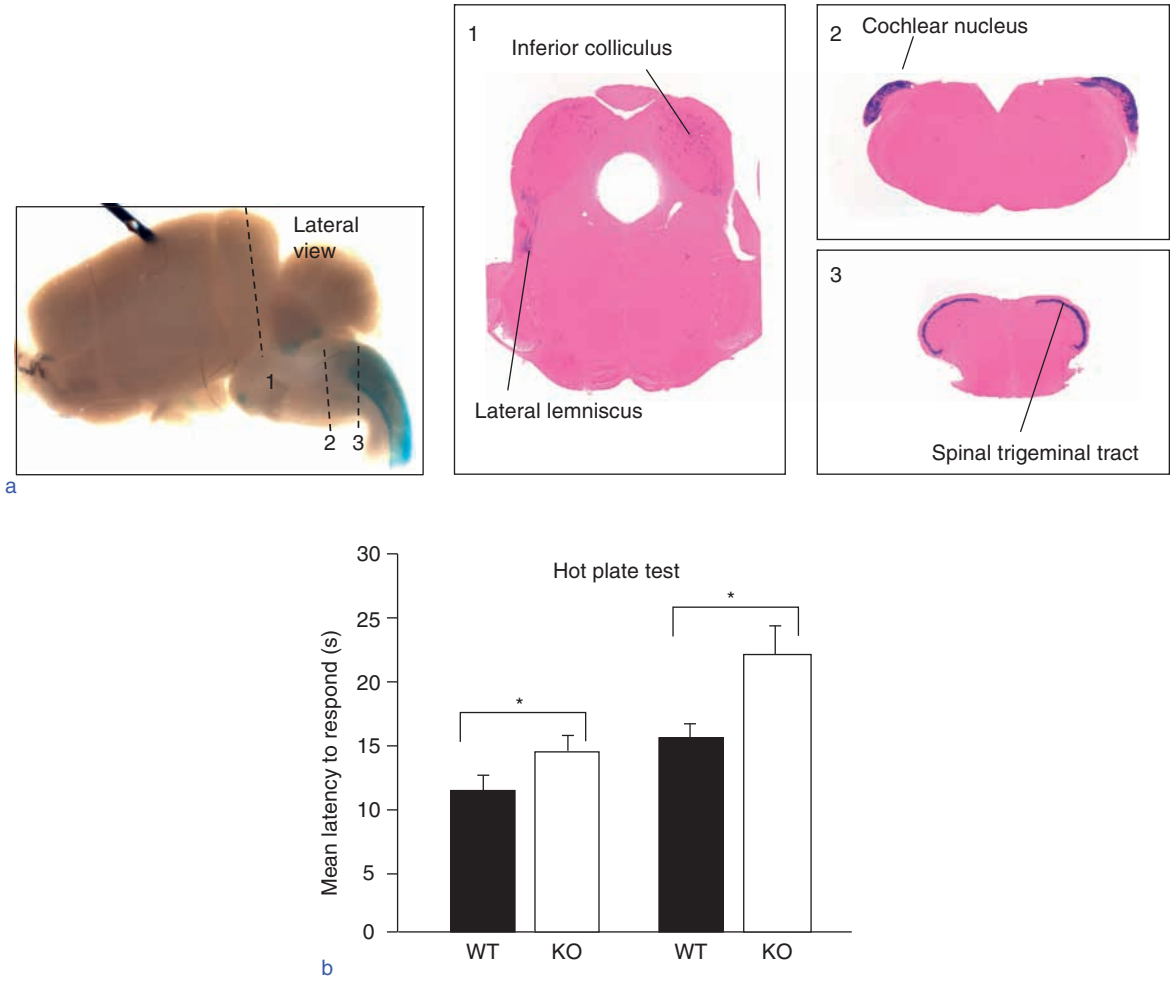

Figure 3 Gene expression profiling and targeted phenotyping in VelociGene mice. (a) Targeted mice containing a LacZ reporter in the NMUR2 gene showed expression in restricted areas of the brain and dorsal horn of the spinal chord. The expression in the trigeminal tract indicated potential function of NMUR2 in the nociceptive response. (b) Hot plate tests for the nociceptive response show impaired response of the null NMUR2 mice compared to their wild-type littermates.

cell basal lamina and causes developmental abnormalities in peripheral nerves. VelociGene targeted mice have generated phenotypes leading to important new realizations in many areas of biology outside of neurobiology, such as that the Delta-like 4 gene product is critical for developmental and tumor angiogenesis, or that a particular microRNA (miR-155) is critical for the formation of germinal centers in immune organs.

The large-scale capabilities provided by VelociGene and VelociMouse, particularly when coupled with the genomewide KOMP initiative, seem likely to continue to make important contributions to all of biology.

See also: Learning and Memory in Invertebrates: *Hermissenda*; Single Cell Electroporation; Single Cell Molecular Analysis Procedures; Single Cell PCR Coupled with Electrophysiology.

Vergence Eye Movements

C Busettini, R C Davison, and P D R Gamlin, University of Alabama at Birmingham, Birmingham, AL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

When we shift our gaze from a far object to a near object, our eyes converge, the lenses of our eyes modify their focus (accommodate), and our pupils often constrict; these combined responses are termed the near triad. The vergence system, which controls binocular alignment, that is, the angle between the two eyes, aligns both foveas with the object of interest, permitting fusion of the retinal images into a single percept. Failure or insufficiencies of this system result in diplopia, or double vision. The accommodation system adjusts the shape of the lens of each eye to bring the new object into focus. These combined vergence and accommodation responses are often termed the near response. The third component of the near triad is the pupil constriction that often accompanies the near response. This article reviews our current knowledge of the systems involved in the near triad and their interconnections.

The Near Response

To better understand the geometry underlying vergence eye movements, consider those eye movements that are required to look between far and near objects (Figure 1). If we want to look at the mountains, we align the lines of sight at infinity and adjust the power of the lenses of our eyes to infinity as well. While viewing the mountains (Figure 1(a)), a snowflake on the window will have a crossed binocular disparity and will be blurred. Therefore, to transfer our gaze to the snowflake we have to converge our eyes and accommodate. While viewing the snowflake (Figure 1(b)), the mountains will have an uncrossed binocular disparity – that is, to compensate for it we have to uncross our eyes, and be out of focus. To return our gaze to the mountains we have to diverge our eyes and disaccommodate. As illustrated in Figure 1, in natural environments vergence and accommodation demands co-vary in a geometrically predictable fashion. As a consequence, the vergence and accommodation systems in primates are closely cross-coupled with shared vergence/accommodation neural elements. Although these interactions are presumably distributed throughout the brain, in a schematic of the underlying neural control of the near response like the one in Figure 2, it is common to merge them into a single pair of cross-links (CA and AC).

When we transfer our gaze to a new object at a different distance, the blur and the binocular disparity inputs to the blur/accommodation controller and the disparity/convergence controller are no longer zero. The two central controllers are activated, producing the appropriate dynamic accommodation and convergence signals needed to transfer our view to the new object. The dynamic changes in dioptric power of the lenses and eye alignment elicited by these controllers reduce the amount of blur and binocular disparity on the two retinas, and these changes are detected by the visual system (visual feedback). Vergence and accommodation responses end when the binocular disparity and blur of the new object of interest reach zero. In humans, the visual loop delay – the time between the occurrence of a visual change and an altered response – is slow (150–200 ms for vergence and approximately 350 ms for accommodation). To more rapidly monitor the progress of the response, the vergence system also has an internal feedback mechanism, which is nonvisual and thus can operate with a much shorter loop delay (Figure 2). This local feedback uses a neural copy of the vergence motor command as an indirect real-time estimate of the eye movement (corollary discharge). It is most likely that the accommodation system operates with a similar local feedback, but no conclusive data are available at this time. As is the case for all other oculomotor subsystems, the vergence and accommodation commands reaching the motor neurons are composed of both dynamic signals (velocity pulses), needed to reduce the damping effects on the dynamic of the motor response by the mechanical viscoelastic and inertial components of the orbital tissues, and sustained commands (position steps), needed to maintain the new vergence and accommodation levels at the end of the movement. To generate the required position steps, it is probable that dynamic signals are integrated ('neural int' in Figure 2) by neurons located in the midbrain and/or the cerebellum.

Historically it was assumed that divergence eye movements and disaccommodation were the passive results of a decrease in convergence and accommodation tonic signals, respectively. The discovery, in several areas of the macaque brain, of separate populations of neurons which encode either convergence and accommodation or divergence and disaccommodation indicates that divergence and disaccommodation are under dynamic control as well. Therefore, vergence and accommodation responses reflect the balance between the activity of two subsystems with opposite signs that act in a push–pull fashion.

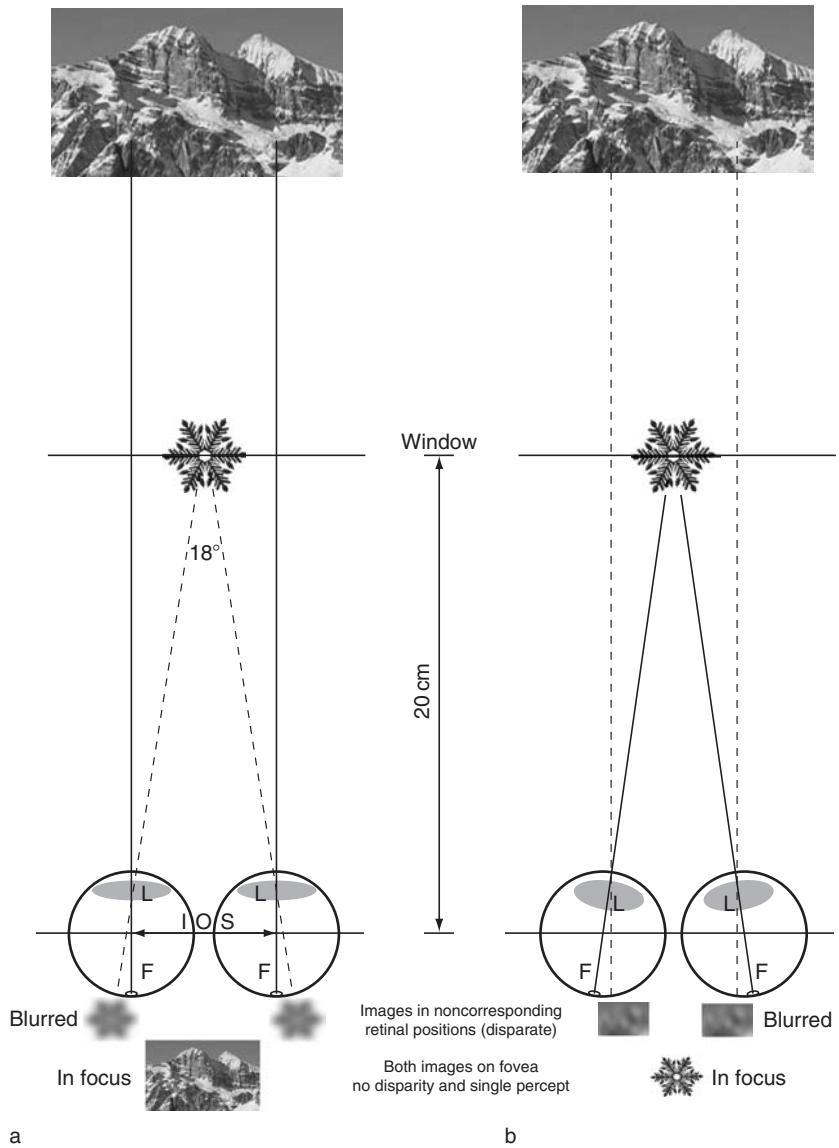

Figure 1 Geometry of the near response. (a) When we look at the mountains through a window, the lines of sight are aligned at infinity (vergence angle = 0°) and the focus of the lenses (L) is also at infinity (accommodation = 0 diopters). The images of the mountains are aligned on the two foveas (F), in focus, and are fused into a single percept. The lines from the two eyes to the target can be considered as two sides of a triangle, with the base formed by the interocular separation (IOS). The images of a snowflake on the window will fall on the temporal retina of each eye, at an angle equal to tan(0.5 × IOS/distance), where distance is the distance of the object from the eye. For adults, the average IOS is 6.5 cm, and, for a window at 20 cm, this angle corresponds to 9°. The temporal, or crossed, disparity error (i.e., the total angular deviation, with sign, from the two foveas) is therefore 18°, which is also the amount of convergence that we would have to generate to align the two lines of sight on the snowflake. This disparity, and the associated vergence movement, can also be expressed in meter-angles. A meter-angle is defined as the amount of convergence needed to align the eyes on an object located at 1 m. A disparity of 18° is well outside Panum's fusional area (the disparity range over which the brain can fuse disparate images into a single percept), and the snowflake therefore appears diplopic. With the lens of both eyes focused at infinity, the snowflake will also appear blurred, with an accommodation error equal to 1/distance = +5 diopters, with the distance expressed in meters. This numerical value is also the amount of disparity/vergence error when expressed in meter-angles. (b) To transfer our view from the mountains to the snowflake we must converge our eyes by a total of 18° (+5 meter-angles), and accommodate by +5 diopters. Now the mountains are diplopic and out of focus. Their projections are on the nasal retina of each eye, with a total uncrossed disparity error of 18° (−5 meter-angles), and with a focusing error of-5 diopters. To return to the mountains we will uncross (i.e., diverge) our eyes by a total of 18° (−5 meter-angles), and disaccommodate our lenses by −5 diopters.

When vergence and accommodation demands co-occur, the crossed connections between the vergence and accommodation systems are not clearly evident. To demonstrate the influence of the vergence system on the accommodation system, one can view a three-dimensional scene through pinholes. This greatly increases the depth of focus of the eye and hence eliminates the blur cue. This situation is simulated by the

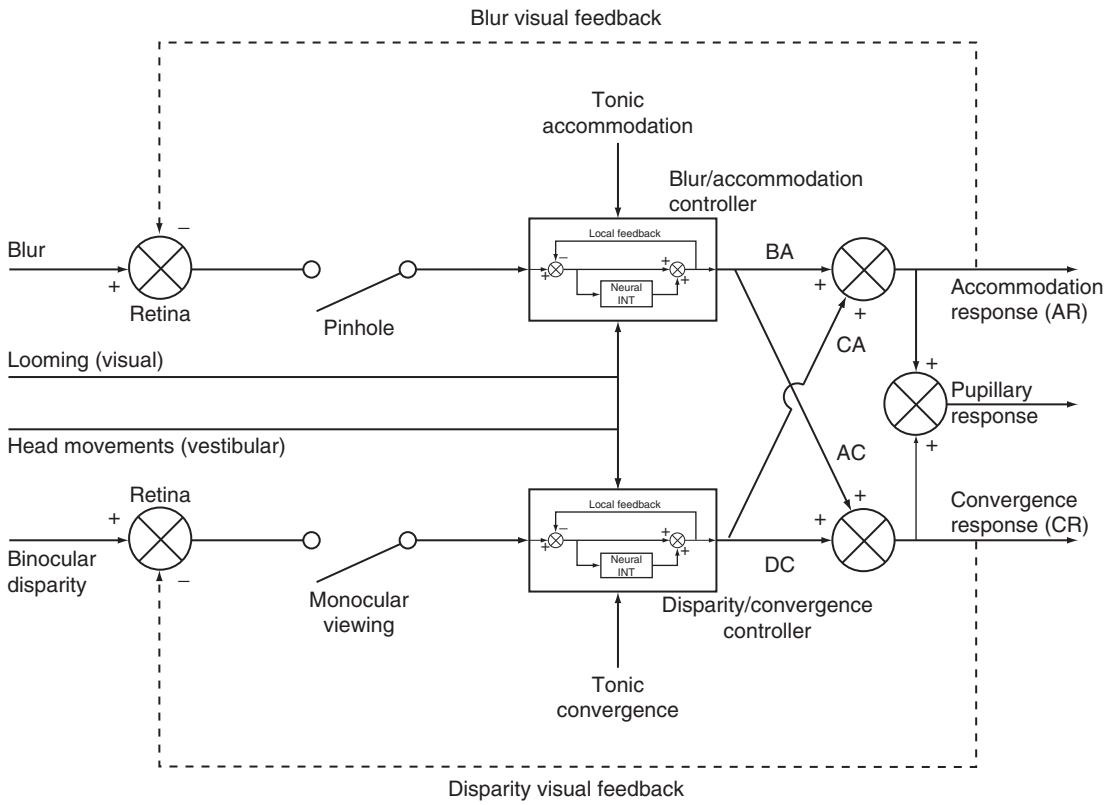

Figure 2 A cross-coupled model of the vergence and accommodation systems. This model applies to the near response (convergence and accommodation), but the same schematic applies to the far response (divergence and disaccommodation). The blur/accommodation controller is depicted as a box with blur as dynamic input. The presence of a blur signal generates a compensatory accommodation command (velocity pulse). To maintain the new accommodation state at the end of the movement, the dynamic signal is integrated by the neural integrator (neural int) and the positional step is added, with sign, to the existent tonic accommodation. The progress of the accommodation response is controlled by both a nonvisual internal local feedback, which uses the motor command itself as estimate of the ongoing lens change, and a visual feedback. The blur visual feedback (dashed line) and the summing junction (retina) are not neural elements, but a representation of the effect of the ongoing accommodation response on the visually perceived blur. When blur is zeroed the response ends. Exactly the same mechanism applies to the convergence response, which ends when the binocular disparity is zeroed. During a transfer of gaze in depth, a copy of the blur-driven accommodation response (BA), output of the blur/accommodation controller, is sent to the convergence system as an accommodation-driven convergence command (cross-link AC). A copy of the disparity-driven convergence (DC), output of the disparity/convergence controller, is sent to the accommodation system as a convergence-driven accommodation command (cross-link CA). The sum of the BA and CA contributions will give the actual accommodation response (AR). The sum of the DC and AC contributions will give the actual convergence response (CR). Crossed contributions from the vergence to the accommodation systems become evident by using pinholes, simulated by the opening of the pinhole switch. This device greatly increases the depth of focus, with both mountains and snowflake now in focus. Thus, during the transfer of gaze, the input to the blur/accommodation controller remains zero (no blur cues), but the binocular disparity will activate the disparity/convergence controller, which output (DC) will generate not only a convergence response, but by way of the cross-link with the accommodation system (CA), also an accommodation response. Similarly, one can eliminate the binocular disparity input while maintaining the blur input by covering one eye (monocular viewing switch). Under this condition, there is no binocular disparity, and the disparity/convergence controller is not active. Nevertheless, the blur of the target will activate the blur/accommodation controller, which output (BA) will not only generate an accommodation response, but also a convergence response by way of the cross-link with the convergence system (AC).

opening of the pinhole switch in Figure 2. On the other hand, one can eliminate binocular disparity input while maintaining blur input by covering one eye (monocular viewing). This situation is simulated by the opening of the monocular viewing switch in Figure 2.

When we ask a patient to wear prisms to correct their binocular alignment or to wear lenses for accommodative correction, we are effectively requiring that the patient maintains the output of the other system unaltered, that is, to dissociate vergence and accommodation. If the patient has normal cerebellar function, long-term adaptive changes will compensate for these conflicts by modifying the tonic vergence and accommodation signals and/or the strength of the cross-links between the two systems.

Vergence and accommodation responses are also elicited (Figure 2) by an object seen moving closer or farther from us (visual looming); and by a forward or

a backward movement of the head. Translational head movements are detected by the otoliths in the inner ear, acting as linear accelerometers. A forward head movement will generate vestibularly driven convergence, while a backward head movement will generate vestibularly driven divergence. Their vestibular origin is confirmed by the fact that they are also observed in the dark. Furthermore, Figure 2 shows that the vergence and accommodation outputs serve to drive pupillary responses, the third component of the near triad.

The Vergence System

The main visual cue directly eliciting vergence eye movements is binocular disparity. The presence of a disparity error means that the images of an object fall on noncorresponding positions in the two retinas, and, if this error is too large for the two images to be fused into a single percept (i.e., beyond Panum's fusional area), the object will be seen as diplopic. There are two, apparently separate, disparity-driven vergence subsystems. A short-range system, termed the fusional vergence system, has a functional disparity range of approximately 1–2°. This system is specifically designed to achieve and maintain alignment of the foveal and parafoveal images within Panum's fusional area (motor fusion), and requires that the images of the object on the two retinas are almost identical to one another. The second system, termed the coarse or long-range system, does not require precise matching of the retinal images. It is able to detect widely disparate images such as those of the snowflake or the mountains, and to elicit the appropriate vergence eye movements (Figure 1).

Vergence-Related Cortical Areas

Short-range disparity-sensitive neurons have been reported in several cortical areas in the macaque monkey, including striate visual area V1; extrastriate visual areas V2, V3, V3A, and V4; middle temporal (MT) area; medial superior temporal (MST) area; lateral, ventral, and caudal intraparietal (LIP, VIP, CIP) areas; frontal eye fields (FEFs); and supplementary eye fields (SEFs). Several of these areas are also vergence sensitive (i.e., encode aspects of the motor command) and blur and/or accommodation sensitive. Cortical areas MST, FEF, and SEF are believed to be directly or indirectly involved in the generation of vergence and accommodation motor commands (Figure 3).

Area MST is directly linked to the generation of ultrashort latency (~80 ms in humans) reflexive vergence responses elicited by sudden changes in depth of a large textured visual scene. In the rhesus monkey, the activity of many MST cells significantly leads these vergence responses, and the population average of the MST neural activity matches the tuning curve of the oculomotor response. The restricted functional range of these reflexive responses is compatible with them forming part of the short-range circuitry. Other researchers, using small targets, have found that a substantial fraction of MST cells respond during voluntary sinusoidal smooth tracking in depth.

The prearcuate frontal eye field area contains cells displaying activity that is significantly correlated with both vergence and accommodation motor commands, as illustrated in Figure 4. The temporal properties of these neural responses indicate that this

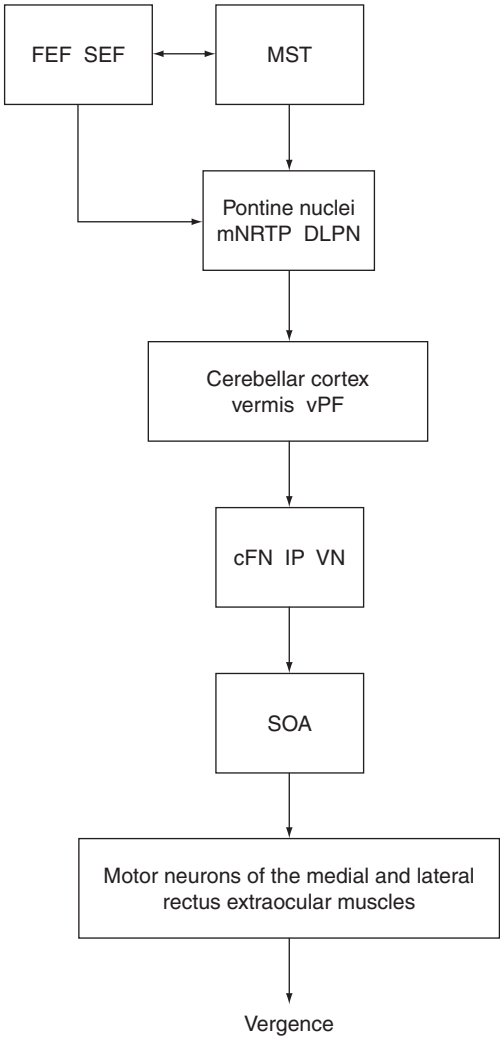

Figure 3 A schematic diagram of the neural pathways controlling vergence eye movements in primates. FEF, frontal eye fields; SEF, supplementary eye fields; MST, medial superior temporal area; mNRTP, medial nucleus reticularis tegmenti ponti; DLPN, dorsolateral pontine nuclei; vPF, ventral paraflocculus; cFN, caudal fastigial nucleus; IP, interpositus; VN, vestibular nuclei; SOA, supraoculomotor area.

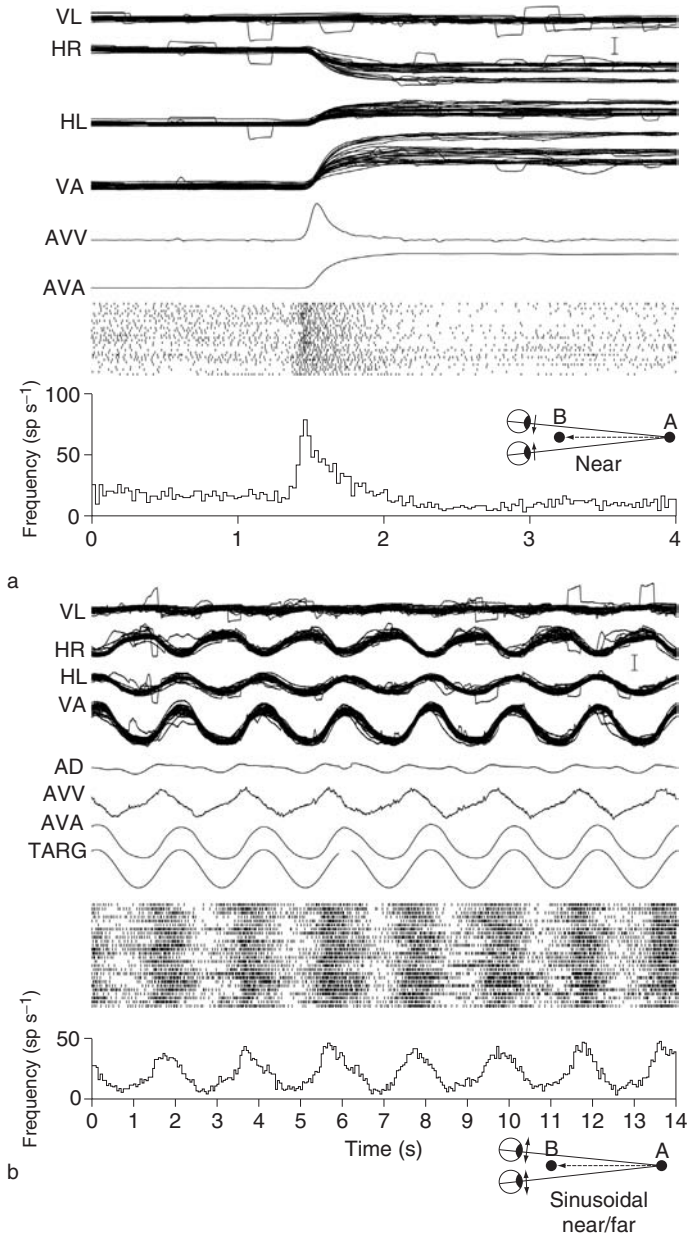

Figure 4 Example of a near-response cell in the prearcuate frontal eye field area of the macaque monkey. (a) Responses of the cell to multiple presentations of a small target stepped from far to near. The single trials are synchronized with the onset of the vergence response. (b) Responses of the same cell to multiple presentations of a small target sinusoidally oscillating in depth. VL, vertical position trace of the left eye; HR, horizontal position trace of the right eye; HL, horizontal position trace of the left eye; VA, vergence angle trace; AVV, average vergence velocity trace; AVA, average vergence angle trace. The spike raster, the activity of the neuron during each presentation, recorded with a microelectrode, is shown as a sequence of ticks, each representing the occurrence of a spike. Each line of the raster corresponds to one presentation; Frequency is the average frequency of the firing of the cell. Panel B addresses two important points regarding the nature of the signal encoded by this cell. The first is derived from the trace labeled AD (average disparity). With the animal successfully tracking the moving target, there is only a minor dynamic disparity error, but the responses of the cell are as robust as in panel a. This indicates that it is encoding the vergence motor command, or a corollary of a vergence motor command, and not a (visual) disparity signal. Further evidence that this cell is not visually driven is shown by the maintenance of the firing when the target (TARG) is briefly extinguished and the animal is in total darkness. The firing continues to follow the vergence response. The location of this event is marked by the brief interruption of the target trace near the middle of panel b. A latency analysis also suggested that many of these cells change their firing activity with respect to the onset of the motor response (lead time) early enough to be actually driving the observed vergence response. Scale $= 4^\circ$, and $20^\circ\,s^{-1}$. Adapted from Gamlin PDR and Yoon K (2000) An area for vergence eye movement in primate frontal cortex. *Nature* 407: 1003–1007.

region contains neurons directly involved in the generation of vergence eye movements. Comparable results have been found in the caudal frontal eye fields. Neurons within the macaque SEF display motor-related responses during sinusoidal vergence tracking.

In humans, lesions of areas comparable to the macaque MST produce profound deficits in the ability of a patient to generate vergence responses, as well as in performing fronto-parallel smooth-pursuit tracking responses. These observations are consistent with MST being a major center for motion processing in three-dimensional space. Lesions of the FEF region produce deficits in vergence, saccades, and smooth-pursuit eye movements, indicating that this region is a sensorimotor area responsible for the generation of these voluntary eye movements. In contrast, damage to the SEF area fails to produce any major deficits in vergence tracking. This suggests that, as is the case for the saccade-related and fronto-parallel pursuit-related activity, the SEF is involved in planning more complex motor behaviors. Deficits in these functions would not be expected to interfere with the execution of simple visually driven oculomotor tasks.

Vergence-Related Subcortical Areas

The most important subcortical structures between the MST and FEF vergence-related cortical areas and the cerebellar vergence-related areas are the basilar pontine nuclei medial nucleus reticularis tegmenti pontis (mNRTP) and dorsolateral pontine nuclei (DLPN). These nuclei have cells that respond during vergence responses, and electrical microstimulation of mNRTP neurons produces changes in vergence state. Figure 5 shows an example of a convergence neuron (panels a and b) and of a divergence neuron (panels c and d) in mNRTP. The mNRTP primarily projects to the cerebellar vermis, which innervates the caudal fastigial nucleus; the DLPN primarily projects to the ventral paraflocculus, which innervates the vestibular nuclei as well as two of the deep cerebellar nuclei – the interpositus and dentate. Consistent with these connections, vergence signals are found in the caudal fastigial nucleus and the ventral paraflocculus. Lesions in the dorsal vermis in monkeys induce a convergence bias during monocular fixation in the absence of disparity cues, and chemical inactivation of the caudal fastigial nucleus impairs vergence eye movements.

Midbrain Premotor Neurons

It was known for many years that vergence eye movements are usually spared in patients with internuclear ophthalmoplegia, where damage to the medial longitudinal fasciculus (MLF) prevents conjugate eye movement signals from reaching medial rectus motor neurons from the contralateral abducens nucleus. This suggested the existence of a vergence motor pathway that was independent of the saccadic and frontal smooth-pursuit pathways. It is now known that vergence commands reach the medial rectus motor neurons directly from near-response cells in the supraoculomotor area (SOA) and not through the MLF. These cells connect monosynaptically to medial rectus motor neurons and appear to be part of a final common pathway for all vergence responses. Examples of the activity of SOA convergence and divergence cells during step changes in depth of a small target are illustrated in Figure 6. These cells do not fire for saccades or fronto-parallel smooth pursuit.

Behavior of Motoneurons

A symmetric convergence response, in which both eyes rotate inward by the same amount, is associated with contraction of the medial rectus muscle (agonist) of both eyes and relaxation of the lateral rectus muscle (antagonist) of both eyes. For divergence, the two actions are reversed. The motor neurons innervating the medial rectus muscles are located in the oculomotor nuclei, while the motor neurons innervating the lateral rectus muscles are located in the abducens nuclei. Figure 7 illustrates an example of activity of a left-eye medial rectus motoneuron during a rightward saccade and a convergence response. Since both responses rotate the left eye to the right, the motoneuron is active for both types of eye movements. More generally, the motor neurons innervating the extraocular muscles respond during all eye movements but with differing sensitivities, and these sensitivities vary from cell to cell.

The Accommodation System

Increased accommodation is achieved by deformation of the lens capsule and matrix by contraction of the ciliary muscle, which is innervated by postganglionic, parasympathetic neurons in the ciliary ganglion. The suspensory ligaments act as passive antagonists. Disaccommodation is achieved by relaxation of the ciliary muscle, with the suspensory ligaments acting as passive agonists. Accommodation and disaccommodation responses have somewhat different dynamics, most likely due to the different anatomical aspects of the two responses. The primary tonic and dynamic inputs to the ciliary muscle are from the parasympathetic autonomic nervous system; the sympathetic system exerts a weak, tonic inhibitory action. Antidromic activation has shown that accommodation-related Edinger–Westphal (EW) neurons connect monosynaptically to the ciliary ganglion by way of the oculomotor nerve.

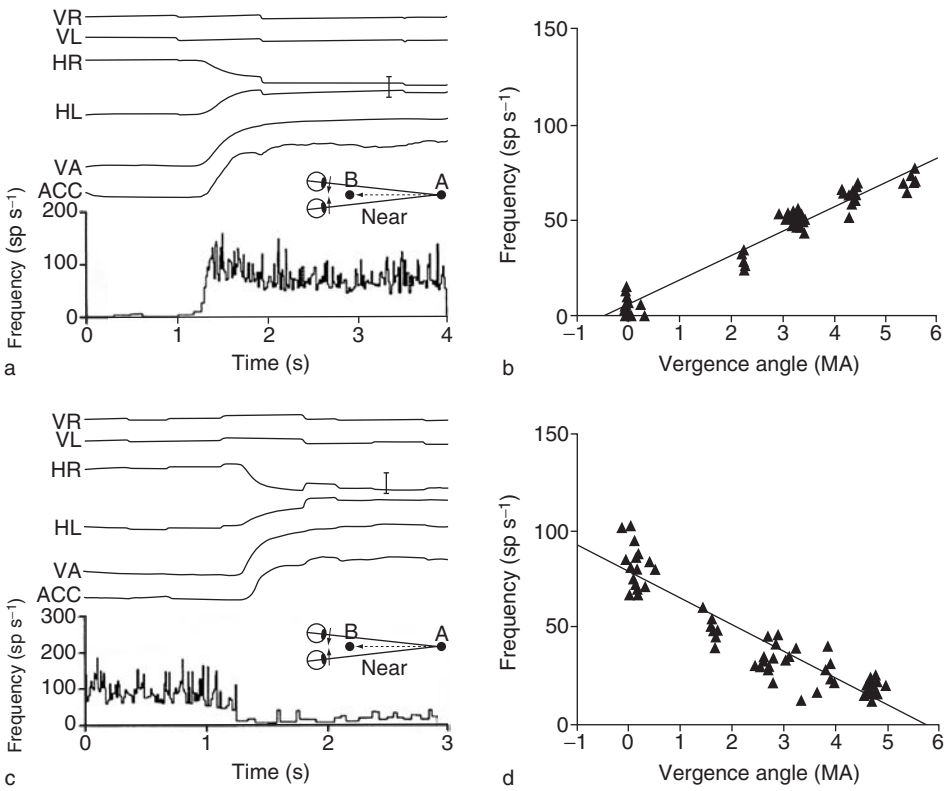

Figure 5 Examples of a near-response cell (a, b) and a far-response cell (c, d) in the medial nucleus reticularis tegmenti pontis (mNRTP) of the macaque monkey. (a) The response of an mNRTP near-response cell to a single far-to-near step of a small target. This cell is practically silent at far and has a robust tonic firing when the animal is converged. The initial higher firing during the convergence eye movement proper indicates that this cell has also some velocity sensitivity. (b) The static positional sensitivity of the cell as dependence of the tonic firing rate of the cell with the static vergence angle, expressed in meter-angles (MA). (c) The response of an mNRTP far-response cell to a similar target step. The cell had a strong firing when the animal was aligned at far, but it has a much lower firing when the animal is converged. This is confirmed by the plot of the position sensitivity of this cell (d), which has the slope opposite to the convergence cell in panel b. Trace abbreviations as in **Figure 4**. Scale = 4°. Adapted from Gamlin PDR and Clarke RJ (1995) Single-unit activity in the primate nucleus reticularis tegmenti pontis related to vergence and ocular accommodation. *Journal of Neurophysiology* 73: 2115–2119.

The sources of accommodation input to the EW nucleus are poorly understood. Direct connections exist from the fastigial nucleus and the posterior interpositus nucleus of the cerebellum. Also, the near- and far-response neurons in the supraoculomotor area immediately surrounding the EW nucleus are presumed to contribute to the accommodation response, but a monosynaptic connection from SOA to the EW has yet to be demonstrated. Shared neural control of vergence and accommodation is also reported in the prearcuate and caudal frontal eye fields, and the medial nucleus reticularis tegmenti pontis (Figure 3).

The Pupillary Near Response

The primary function of pupil responses is to adjust retinal irradiance on the basis of the ambient light. There is strong evidence that the neural circuitry underlying the pupillary light reflex is primarily subcortical, albeit with some cortical influence. Also, as with a camera lens, the diameter of the pupil (aperture) affects the amount of optical aberration of the lens and the depth of focus; a small pupil diameter reduces aberrations and increases the depth of focus, which is the most likely role for the pupillary near response.

Pupil diameter is determined by the degree of contraction of two antagonist muscles, the sphincter pupillae and the dilator pupillae. Contraction of the sphincter causes pupil constriction (miosis), and it is largely under parasympathetic control. Contraction of the dilator (or radial) muscle causes pupil dilation (mydriasis), and it is largely under sympathetic control. The parasympathetic projections to the sphincter pupillae muscle arise from preganglionic neurons in the Edinger–Westphal nucleus which project, by way of the oculomotor nerve, to

postganglionic neurons in the ciliary ganglion. The ciliary ganglion innervates both the iris sphincter and the ciliary muscle of the lens. The afferent component of the pupillary light reflex pathway is composed of direct retinal projections from intrinsically photoreceptive ganglion cells to the pretectal olivaris nucleus (PON), which projects bilaterally to the Edinger–Westphal nucleus.

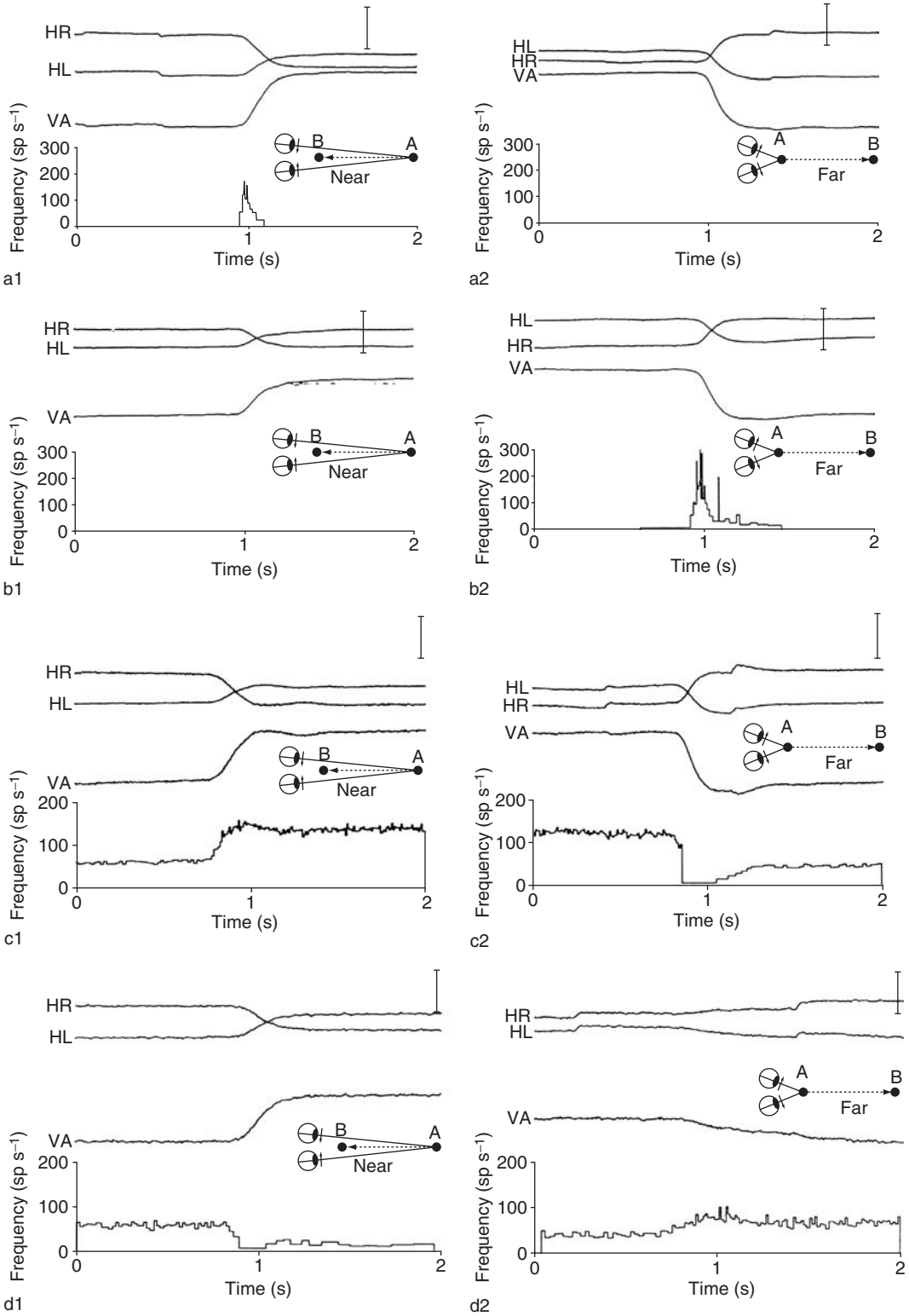

Figure 6 Continued

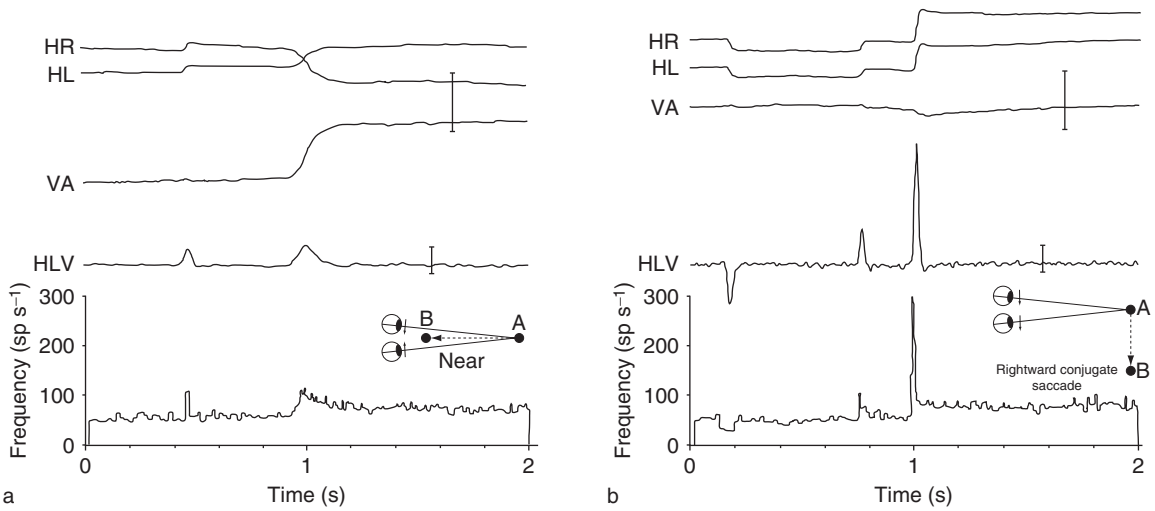

Figure 7 The oculomotor final common pathway. Conjugate commands and vergence commands reach the medial rectus motoneurons through separate pathways. The conjugate command is received through the crossed medial longitudinal fasciculus and the vergence command is received directly from the vergence burst-tonic neurons in the supraoculomotor area. Nonetheless, there is no evidence for distinct anatomical or functional separation of the motoneurons on the basis of the type of oculomotor command they receive: all motoneurons are active, to some extent, for all types of oculomotor behaviors. Medial rectus muscles contract for adducting (toward the nose) horizontal eye rotations, and relax for eye rotations away from the nose. If the final common pathway hypothesis is correct, a motoneuron innervating the medial rectus muscle of the left eye will increase its firing for both conjugate saccades to the right and for convergence. Consistent with this hypothesis, the medial rectus motoneuron reported here shows a clear pulse-step response during convergence (panel a) as well as during a rightward saccade (panel b). For a direct comparison, the amplitude of the vergence response is 4°, requiring a rightward rotation of the left eye of 2°, and the saccade amplitude is 2°, and therefore also requiring the same rightward rotation of the left eye by 2°. The positional and dynamic sensitivity of a motoneuron to the different oculomotor commands will depend on the strength of the synaptic connections from the different pathways. The cell in the example has comparable vergence and saccadic sensitivity, but, at the population level, the sensitivities vary along a continuum and largely independently from each other. Trace abbreviations as in **Figure 4**. Scale = 4°. Adapted from Gamlin PDR and Mays LE (1992) Dynamic properties of medial rectus motoneurons during vergence eye movements. *Journal of Neurophysiology* 67: 64–74.

The pupillary responses that occur with changes in viewing distance correspond to the changes in accommodation (far to near) and disaccommodation (near to far) that occur (i.e., pupil diameter decreases when focusing from far to near and increases when focusing from near to far). This is likely due to a cross-coupling of the accommodation and parasympathetic pupil pathways in the Edinger–Westphal nuclei. Direct contributions from the vergence system to the pupillary near response may be secondary to their actions on accommodation and therefore are indicated by a thin line in **Figure 2**. Other factors may contribute to the pupillary near response since there are some reports of accommodation/disaccommodation changes that are not accompanied by pupil responses.

In some clinical cases, dissociation occurs between the pupillary light reflex and the pupillary near response. The Argyll–Robertson pupil describes the condition where the near response is preserved while the light reflex is absent.

Figure 6 Four examples of vergence-related cells in the supraoculomotor area (SOA) of the macaque monkey. (a1/a2) The cell is an SOA convergence burst cell. This cell is strongly active during the convergence response but it is not active during fixation, and it is never active during divergence (a2). (b1/b2) An SOA divergence burst cell. It does not fire during a convergence eye movement (b1), while it does during the divergence response (b2). During fixation between vergence movements, vergence angle has to be maintained by the contribution of tonic vergence signals. (c1/c2) An example of an SOA convergence burst-tonic cell. When the animal is diverged, this cell has a low tonic firing rate that significantly increases when the animal converges. (d1/d2) An example of an SOA divergence burst-tonic cell. During convergence eye movements the firing rate of the cell significantly decreases but increases during divergence eye movements. By using dissociated vergence/accommodation cues it was found that these cells carry both vergence and accommodation signals and it is possible that they directly drive the Edinger–Westphal accommodation cells. Trace abbreviations as in **Figure 4**. Scale = 4°. Adapted from Mays LE (1984) Neural control of vergence eye movements: Convergence and divergence neurons in midbrain. *Journal of Neurophysiology* 51: 1091–1108; and Mays LE, Porter JD, Gamlin PDR, et al. (1986) Neural control of vergence eye movements: Neurons encoding vergence velocity. *Journal of Neurophysiology* 56: 1007–1021.

See also: Binocular Rivalry; Frontal Eye Fields; Pursuit Eye Movements; Saccadic Eye Movements; Supplementary Eye Fields; Visual System: Functional Architecture of Area V2; Visual System: Multiple Visual Areas in Monkeys.

Further Reading

Franzén O, Richter H, and Stark L (eds.) (2000) *Accommodation and Vergence Mechanisms in the Visual System*. Basel: Birkhäuser Verlag.

Gamlin PDR (2000) The functions of the Edinger–Westphal nucleus. In: Burnstock G and Sillito A (eds.) *Nervous Control of the Eye*, pp. 117–154. New York: Harwood Academic Publishers.

Gamlin PDR (2002) Neural mechanisms for the control of vergence eye movements. *Annals of the New York Academy of Sciences* 956: 264–272.

Gamlin PDR and Clarke RJ (1995) Single-unit activity in the primate nucleus reticularis tegmenti pontis related to vergence and ocular accommodation. *Journal of Neurophysiology* 73: 2115–2119.

Gamlin PDR and Mays LE (1992) Dynamic properties of medial rectus motoneurons during vergence eye movements. *Journal of Neurophysiology* 67: 64–74.

Gamlin PDR and Yoon K (2000) An area for vergence eye movement in primate frontal cortex. *Nature* 407: 1003–1007.

Howard IP (2002) *Seeing in Depth* (2 vols.). Toronto: Porteous.

Leigh RJ and Zee DS (2006) *The Neurology of Eye Movements*, 4th edn., New York: Oxford University Press.

Mays LE (1984) Neural control of vergence eye movements: Convergence and divergence neurons in midbrain. *Journal of Neurophysiology* 51: 1091–1108.

Mays LE (2004) Neural control of vergence eye movements. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 1415–1427. Cambridge: The MIT Press.

Mays LE, Porter JD, Gamlin PDR, et al. (1986) Neural control of vergence eye movements: Neurons encoding vergence velocity. *Journal of Neurophysiology* 56: 1007–1021.

Miles FA, Busettini C, Masson GS, et al. (2004) Short-latency eye movements: Evidence for rapid parallel processing of optic flow. In: Vaina LM, Beardsley SA, and Rushton S (eds.) *Optic Flow and Beyond*, pp. 79–107. Dordrecht, The Netherlands: Kluwer Academic Press.

Parker AJ (2004) From binocular disparity to the perception of stereoscopic depth. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 779–792. Cambridge, MA: The MIT Press.

Schor CM and Ciuffreda KJ (1983) *Vergence Eye Movements: Basic and Clinical Aspects*. Boston: Butterworths.

Vertebrate Eyes: Evolution

R D Fernald, Stanford University, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Light is the premier source of information for many species, and the selective pressure to use this information led to the evolution of eyes. How did eyes evolve? Darwin, the great English naturalist who first brought the systematic explanatory power of evolution to bear on the bewildering biological complexity of our planet, believed that eyes offered a special challenge to evolutionary thinking because they are such "organs of extreme perfection and complication."

Nearly a century and a half later, new discoveries ranging from molecular to macroscopic levels of analysis now provide an answer to the question of how eyes evolve. The evolution of photodetection led to eyes that show natural selection acting at both the organ and the molecular levels. The repeated use of key regulatory gene sequences in eye development and lens formation raises the question of why some transcription factors are regularly recruited to build eyes. Because we can analyze the evolution of structural gene sequences relatively easily, many phyolgenetic relationships among the components of eyes are known. However, the difficulty of tracing exact selective forces that shaped the regulation of gene expression in eyes and elsewhere has made it difficult to understand the evolution of such complex organs.

Structural and Functional Adaptations

Despite the great variety of eyes in the animal kingdom, the laws of physics have constrained optical solutions for collecting and focusing light to just eight types of eyes. To be useful, animal eyes cannot be simple photon detectors but, rather, produce an image by comparing light arriving from different directions.

The likely sequence of evolutionary steps leading to eyes began with simple eyespots in the early Cambrian period, 570–500 Ma. Such photosensitive spots would discriminate between light and dark but could not record complex light patterns. Eyespots that invaginated would add the capacity to detect the direction of incident light and additional receptors could enhance usefulness. Duplication of an existing pit may have led to compound eyes, whereas duplication of receptors may have led to a chambered eye. Adding an optical system in front of the photosensitive elements would increase collection of light and, if it produced an image, would dramatically improve the usefulness of an eye.

The addition of optical systems to eyes led to the evolution of strikingly similar ocular structures in distinctly unrelated animals, such as fishes and cephalopods. The chambered or camera-like eyes in these two lineages are similar in many details, despite the fact that their owners are phylogenetically quite distant. Both evolved spherical lenses to achieve sufficient refractive power for focusing light underwater, but the inverted retinal layers of fishes (and all vertebrates) are distinctly different from the noninverted, somewhat simpler retinae of cephalopods. Both eye types have photoreceptor cells that use related opsin molecules to detect light but very different transduction cascades. Thus, macroscopically, these eye types and the animals bearing them are not remotely homologous, but there are some similarities in the molecules used during development. The exploitation of homologous genes to build nonhomologous eyes may lie at the heart of understanding eye evolution and evolutionary processes more generally.

Of approximately 33 animal phyla, approximately one-third have no specialized light-detecting organ, one-third have light-sensitive organs, and the remaining third are animals with eyes. Image-forming eyes exist in six of 33 extant metazoan phyla (Cnidaria, Mollusca, Annelida, Onychophora, Arthropoda, and Chordata) which account for 96% of known species alive today, meaning that eyes must be useful. Eyes have many features in common because their optical features ultimately depend on the physical properties of light, which travels in straight lines, consists of electromagnetic radiation, varies in wavelength, etc. For example, eyes have evolved to be sensitive within a narrow range of wavelengths, relative to the broad spectrum of energy produced by sunlight, because early evolution occurred in water, which selectively filters what we know as light. Selection for biochemical mechanisms sensitive to this limited range of wavelengths set the sensitivity that emerged during subsequent evolution. Although nonaquatic species moved to land where they were exposed to a significantly broader spectrum of electromagnetic radiation from the sun, animal eyes are limited to vision within the narrow band dictated by our origins in water. Some insect and fish species evolved additional receptor types for ultraviolet light, but the more common restricted sensitivity reflects just how profound early evolutionary solutions persist in evolved organs.

Photon Capture

Both vertebrate and invertebrate eyes transduce photons into cellular signals using seven-tramsmembrane

spanning opsin proteins (30–50 kDa) that combine with a retinal chromophore based on vitamin A. Opsins arose before eyes and evolved into at least seven distinct families. Opsin gene duplications gave rise to the independent evolution of three-color vision in Old and New World primates. Wavelength sensitivity in photoreceptors is modulated by amino acid side groups adjacent to the chromophore binding site in the seventh transmembrane domain of opsins, and evidence of natural selection can be seen in the arrangements of these residues.

The two primary photoreceptor types use different families of opsins packed into different membrane specializations and use different transduction cascades. Vertebrate photoreceptors use members of the ciliary opsin family, so called because these are incorporated into specialized cilia. In contrast, invertebrate photoreceptors use members of the rhabdomeric opsins formed into rhabdoms, hence the name. In vertebrates, the photoreceptors produce hyperpolarizing potentials using a phosphodiesterase cascade which provides amplification. In invertebrates, photoreceptors are depolarizing, using a phospholipase C cascade with amplification provided as the TRP channels open.

We have detailed knowledge about the evolutionary relationships among opsin molecules, and in particular, the adaptive radiation of pigment types due to natural selection for particular wavelength responses has been described for some special cases (e.g., east African cichlids and squirrelfish). However, there is a great deal of variance in spectral sensitivities arising from specific selective advantages for one solution over another. Detailed comparisons between terrestrial vertebrates and insects, for example, revealed that there are not unique solutions to encoding both spatial and spectral information. Mammals and bees use long wavelength receptors for luminance and color vision, whereas flies and birds have evolved separate sets of photoreceptors for the two purposes.

Capacities of eyes vary greatly depending on their ultimate structure. For example, resolution of an image (e.g., subtended degree) differs by approximately 13-fold among vertebrates and even more between vertebrates and invertebrates, and it is a function of photoreceptor size and spacing. Eagles have the greatest acuity, which is approximately 10^5-fold greater than that found in planaria. Similarly, a comparison of relative sensitivities among vertebrates reveals a range of 4×10^5 between highly sensitive deep-sea animal vision and human foveal vision. Animals clearly have evolved eyes with resolution, sensitivity, and wavelength detection to match their needs, even as those needs change during their life history.

The greatest variety of eyes exists among invertebrates, which have both camera eyes (e.g., cephalopods) and compound eyes (e.g., *Drosophila*). Moreover, invertebrates also have the greatest variety of eye number and location on given species. Whereas vertebrates settled on paired, chambered eyes with lenses on the head, invertebrate species may have multiple, nonpaired eyes and eyes in remarkable locations. For example, certain butterflies have light-detecting organs located such that darkness signals successful copulation. A visual system in the planula of a box jellyfish, *Tripedalia cystophora*, has been described that has eyecups directly connected to motor cilium. In this case, there is no nervous system to process visual information because the eyes are a complete sensorimotor system unto themselves.

Lenses: Multiple Protein Types and Gene Sharing

All biological optical systems fall into one of three classes based on image formation: images formed via shadows, via refraction (e.g., lens and/or cornea), or via reflection. The most common eyes collect light through an aperture and focus it with a lens onto photoreceptor cells specialized to convert photons into neural signals. Some eyes exist without pupils and even without lenses (*Nautilus*), but eyes that evolved to give their owners a clear view of the environment on a short timescale do have lenses. Could lenses that consist of tightly packed proteins provide some insight about how eyes evolved?

In vertebrates, lenses are formed from modified epithelial cells with high concentrations of soluble proteins, known as 'crystallines' because of their organized packing into arrays. In contrast, in most invertebrates, the lens proteins are secreted by specialized cells of the eye. Lenses of mitochondrial origin have been found in the two pairs of eyes of the parasite *Neoheterocotyle rhinobatidis*. Despite their distinct origins, both vertebrate and invertebrate lenses must have their constituent proteins distributed to produce a radial gradient of refractive index that is low at the edge of the lens and high in the center. A particular gradient of refractive index is essential for vision in animals living in water, but gradients are also found in terrestrial vertebrates and invertebrates.

Until quite recently, approximately ten crystalline proteins found in lenses were thought to be unique to lens tissue having evolved for this function. Of the large number of crystallines, α, β, and γ crystallines are specialized lens proteins in vertebrates although α and β are related to heat shock protein and schistosome egg antigen, respectively. However, the remaining vertebrate although α and β are lens proteins are not conserved but, rather, comprise a diverse group, many of which are used as enzymes elsewhere in the

body. Surprisingly, most of these taxon-specific lens proteins are actually products of the same genes as the enzymes, and this double use has been termed 'gene sharing.' For example, a crystalline protein in the duck lens is similar to a metabolic enzyme, argininosuccinate lyase, and both the lens protein and metabolic enzyme are encoded by the same gene, not by duplicated genes, although such sharing may have been a prelude to gene duplication. This molecular opportunism is so effective that it has also occurred both in cephalopods and in *Drosophila*. One possibility is that since lenses need the production of a relatively large amount of protein, genes that can be easily upregulated in other tissues might be preferentially selected.

Perhaps the most remarkable example of a lens from an unusual source is found in the brittlestar (*Ophiocoma wendtii*), which forms crystal lenses as a part of its skeletal armor from calcite crystals. The crystals, oriented to bring light onto the photoreceptive surfaces in the body, focus the light much as corrective lenses might, and they effectively concentrate the light by approximately 50 times.

The common cellular strategy of assembling lenses from diverse proteins seems to be a convergent evolutionary solution that has occurred in many vertebrates independently. The exquisite gradient of the refractive index which evolved in vertebrates and invertebrates alike resulted because it is the only way known to make an optically useful lens. What remains unknown is how diverse proteins are assembled through folding and organized to preserve key properties of transparency and suitable refractive index gradient along the axis of the lens. The challenge for understanding lens development is to identify the mechanisms responsible for organizing diverse proteins into a functioning lens. This knowledge could provide useful insights about eye evolution from the perspective of lens assembly.

Origins of Eyes

Logically, eyes might be monophyletic, having evolved from a single progenitor, or polyphyletic, having arisen more than once during evolution, and views on this have switched back and forth throughout the years. A comparison of overall structure, photoreceptor types, developmental origins of eye tissue, position of receptor axons, and other anatomical markers among eyes using current fauna suggested that eyes evolved not once but at least 40 different times, and possibly many more. This 'multiple-origins' hypothesis, based on morphological evidence, was later challenged by results from molecular experiments. A single, well-conserved 'master' gene, *Pax6*, can initiate eye construction in diverse species, suggesting that eyes might have arisen from a single ancestor. Since this original debate, there have been several salient discoveries that suggest eyes arose more than once and we carry the evidence within our own eyes.

Eye development requires morphological transformations of newly generated tissue regulated by multiple genes with expression patterns overlapping in time and space. Eyes develop from the prospective forebrain, beginning in the eyefields, which are made up of cells of the anterior neural plate. As the prosencephalon grows, this region moves forward until the optic groove forms and the neuroectoderm of the groove locally contacts the surface ectoderm, inducing the lens placode. As the placode invaginates to form the lens vesicle, the optic vesicle forms the bilayered optic cup, which ultimately becomes the eye. The interaction between the optic vesicle and the lens placode was identified as the 'organizer of the lens' by Spemann. The presumptive lens arises from the lens placode, a thickening of the ectoderm in contact with the optic vesicle. Coincident with this change is the onset of expression of proteins that will form the lens. Other structures of the eye are formed by large- and small-scale tissue movements, caused and accompanied by the expression of tissue-specific genes at that site. The cornea arises from the surface ectoderm over the lens and from migrating mesenchyme derived from the neural crest. Many of the original observations about the role of specific tissue bits in these processes resulted from exquisite embryonic manipulations related to transplantation experiments. With well-described macroscopic change in hand, the next challenge is to synthesize the phenomenological, macroscopic morphological observations with molecular explanations of eye development and understand what this tells us about evolution.

The morphological process of eye development has been viewed as a set of steps toward a final tissue arrangement. Underlying this apparently straightforward sequence of large-scale events, however, are distributions of gene expression with substantial overlap in both time and space. Gene expression is closely regulated, and specific gene products are used repeatedly, which makes the causal relationships difficult to conceptualize. Nonetheless, progress in characterizing the genes responsible for particular steps in eye development has been quite rapid. Functions for at least 15 transcription factors and several signaling molecules have been described in human and mice eyes, based on developmental disorders and/or molecular manipulations. As with other molecular actors, both the transcription factors and the signaling molecules expressed during ocular development are also developmentally important in a wide range of other tissues, meaning that combinatorial expression patterns are important.

As is now well known, the paired box gene 6 (*PAX6*), a member of the family of genes that encode transcription factors with a homeodomain and a paired domain, appears to be important in eye formation across many species. The remarkable demonstration that *PAX6* can induce eyes where they should not be ('ectopic') in *Drosophila*, and similar subsequent demonstration in vertebrates, led to the suggestion that there might be 'master control genes' responsible for development and differentiation of ocular tissue in many species. Subsequent work has shown that the idea of a master control gene is a misnomer, however, since a suite of genes are required, collectively, to initiate eye development, and transcription factors are a necessary part of the initiation process. Nonetheless, it is remarkable that some of the same genes appear in the context of eye development, despite great evolutionary distance among the owners of the eyes. How this might have occurred is discussed later.

In *Drosophila*, development of the photoreceptor array in the eye is known to require seven genes which work in combination with particular signaling systems through a complex regulatory network. Deletion of any one of the seven key genes causes loss or radical reduction of the *Drosophila* eye, and deletion of all but one can cause formation of an ectopic eye. Moreover, many of these genes play a role in developmental cascades used in formation of the kidney, muscle, and other organs. Thus, the master gene hypothesis is not supported because the whole collection of genes is needed to produce a reasonable eye. Eye development appears to require new ways of thinking about how complex tissues are made and how such organs arose in evolution. The widespread and redundant activities of specific genes during ocular development suggest that building eyes requires the orchestrated activity of a suite of molecular actors.

As described previously, the diversity of eyes confirms their dynamic evolutionary past. Explosive speciation, or the 'big bang' of animal evolution, happened during the Cambrian, when existing eye types appear to have improved radically, coincident with the onset of carnivory and predation. Many selective forces were likely at work, including perhaps the first instances in which light enabled behavioral signals, so no predominant selective force can be claimed. The rapidity of eye evolution has always been a question, but computer simulations show that approximately 2000 sequential changes could produce a typical image-forming eye from a light-sensitive patch. With reasonable estimates, this suggests that an eye could evolve in less than half a million years, making the virtual explosion of eyes during the Cambrian seem reasonable. After the Cambrian, three phyla emerged: arthropods, mollusks, and chordates. Although these groups all use the opsin molecule to capture light, details of the structure and function of their eyes differ considerably.

Functional Evidence about Eye Evolution

Until recently, the photo detection systems we understood well were localized primarily to eyes and pineal glands and a few other sites in the body such as the skin. For each of these, a canonical opsin and related transduction cascade were known. Specifically, ciliary structures associated with specific G-proteins are known from vertebrate eyes, and microvilli associated with inositol phosphate signaling cascades are known from invertebrate eyes. However, recently each of these phototransduction cascades was found in unexpected organisms. The polychete ragworm (*Platynereis dumerilii*), in addition to the rhadomeric photoreceptors in its eyes, has ciliary photoreceptors in the brain and typical types of opsins associated with each photoreceptor type were localized only with that type (e.g., vertebrate opsin in the brain and invertebrate opsin in the eye). This means that the two main types of 'eyes' exist in a worm.

Correspondingly, in vertebrates, a small population of intrinsically photosensitive retinal ganglion cells have been discovered that play key roles in the regulation of nonvisual photic responses. These rely on melanopsin, first identified in vertebrate melanophores, brain, and eyes. Melanopsin is in photosensitive ganglion cells which comprise a non-image-forming system that can detect the presence or absence of light and are required for normal light-induced circadian phase shifting. Subsequent functional analyses showed that retinal melanopsin functions via a phototransduction cascade that resembles invertebrate opsins and, similar to invertebrates, has intrinsic photoisomerase activity. Therefore, not only do vertebrates carry a version of the invertebrate visual transduction system with them but also it is used in a variety of ways, including to provide information to the 'image-forming' visual system.

Taken together, these findings show that at least two kinds of photoreception existed in the Urbilateria, before the split into three Bilateria branches at the Cambrian, and, importantly, each of these branches still carries versions of these two systems. Considering that seven families of opsin have been described in humans, we can expect more surprises in the detection of light. The additional opsins discovered recently have not been functionally characterized, but the evidence suggests that there are no more opsins to be discovered. Even so, determining how all the existing opsins work together is a daunting challenge.

How Did Eyes Evolve?

Eyes exist in a variety of shapes, sizes, optical designs, and locations on the body, but they all provide similar information about wavelength and intensity of light to their owners. Different tissues have been recruited to build lenses and retinas across the phyla. In contrast, all eyes share the same mechanism of absorbing photons which has been conserved across phylogeny, namely the opsin–chromophore combination. Despite new findings yielded by powerful molecular techniques, all evidence still suggests that eyes have a polyphyletic origin, as underscored by the discovery that two photodetection systems had evolved prior to the split of the Urbilateria into three families. Clearly, non-homologous eyes contain homologous molecules responsible for many structural, functional, and even developmental features. Given a growing list of homologous gene sequences among molecules in the eye across vast phylogenetic distances, the challenge is to discover what makes the eyes of *Drosophila*, squid, and mouse so different. Understanding these differences is probably more difficult than determining what they have in common.

It seems increasingly evident that as eyes have evolved, different functional mechanisms have been generated by recruiting existing gene programs. From genome sequencing, we know that there are far fewer genes in organisms than previously thought, so the use and reuse of genes and their products in combinatorial assemblies as reported for known genomes makes sense. In the development of eyes, this seems to be the rule, not the exception. Specifically, in the evolution of eyes, it seems likely that light sensitivity evolved early in the Cambrian in the form of a proto-opsin molecule in association with the chromophore retinal. This molecular combination, sensitive to light, became associated with genes such as *pax6*, *eya*, and others. It seems likely that this combination was recruited and worked well in early evolution of eyespots and other light-sensing organs.

Important insights about how regulatory gene networks might have evolved come from what is called the 'hox paradox.' During development, orthologous genes are expressed in superficially similar domains during embryonic development of very different organisms (e.g., *Drosophila* and mouse), yet these embryos produce adults that are anatomically quite distinct, having very few structures with common ancestors. Although not completely resolved, one resolution of this paradox is that there has been evolutionary convergence in the use of some genes and hence apparent homology, which seems to be the likely scenario for the evolution of eyes. Some genes have been recruited into regulatory gene networks repeatedly, possibly committed early in evolutionary history and kept because they simply work well.

As different eye types evolved over time, there was probably repeated recruitment of particular gene groups, not unlike improvisational groups of actors, interacting to produce candidates for selection. The evolutionary fiddling through which various combinations or routines were tried could have led to numerous parallel evolutionary paths for eyes as we now envisage.

From this, two different mechanisms for transmitting the photic information to surrounding cells were selected for – one in ciliary and one in rhadomeric photoreceptors. These two systems are likely present in all organisms as described previously for worms and mice. The major surprise is that both of these transduction systems persisted with each selected as the primary visual system for a major branch of animals. In vertebrates, these independently evolved phototransduction systems now collaborate in eyes to deliver information about light to our brains. Therefore, the answer to the question of whether eyes evolved from a single prototypical eye (monophyletic), or whether they evolved repeatedly (polyphyletic), appears to be that quite evidently eyes arose at least twice and probably many times.

See also: Activity in Visual Development; Contextual Interactions in Visual Perception; Contextual Interactions in Visual Processing; Photoreceptor Adaptation; Photoreceptors: Physiology; Visual System Development: Invertebrates; Visual System: Invertebrates; Visual Development; Visual System: Adaptive Regression and Progression in Subterranean Mammals.

Further Reading

Fernald RD (2000) Evolution of eyes. *Current Opinion in Neurobiology* 10: 444–450.

Fernald RD (2004) Eyes: Variety, development and evolution. *Brain Behavior and Evolution* 64: 141–147.

Fernald RD (2006) Casting a genetic light on the evolution of eyes. *Science* 313: 1914–1918.

Land MF and Fernald RD (1992) The evolution of eyes. *Annual Review of Neuroscience* 15: 1–29.

Land MF and Nilsson D-E (2002) *Animal Eyes. Oxford Animal Biology Series*. Oxford: Oxford University Press.

Nilsson DE (1996) Eye ancestry: Old genes for new eyes. *Current Biology* 6: 39–42.

Nilsson DE (2005) Photoreceptor evolution: Ancient siblings serve different tasks. *Current Biology* 15: R94–R96.

Salvini-Plawen LV and Mayr E (1977) On the evolution of photoreceptors and eyes. *Evolutionary Biology* 10: 207–263.

Terakita A (2005) The opsins. *Genome Biology* 6: 213.

Walls GL (1942) *The Vertebrate Eye and Its Adaptive Radiation*. Bloomfield Hills, MI: Cranbrook Institute of Science.

Wistow G (1993) Lens crystallins: Gene recruitment and evolutionary dynamism. *Trends in Biochemical Sciences* 18: 301–306.

Vertigo

E Smouha and G Wanna, The Mount Sinai School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Dizziness, a very common neurological complaint, can be caused by a variety of disorders affecting the vestibular, visual, and somatosensory systems; cerebellar, motor, and extrapyramidal systems; or cardiovascular, metabolic, and psychological systems. Classification of the patient's complaint is essential.

It is useful to divide dizziness into one of four categories: vertigo, imbalance, disequilibrium, and light-headedness. Vertigo is defined as a hallucination of motion of the body or the environment, and may be rotatory or linear. Vertigo is a cardinal symptom of vestibular system disease, which may be of peripheral or central origin. Imbalance (unsteadiness) is defined as impaired ability to maintain the intended orientation of body in space. Imbalance usually signifies chronic vestibular or neurological disease. Disequilibrium, or a feeling of 'drunkenness' or disorientation, may be vestibular, neurologic, multisensory, metabolic, or psychogenic in origin. Light-headedness (syncope or near-syncope) is cardiovascular or cerebrovascular in origin, or occasionally metabolic (e.g., hypoglycemia), but almost never vestibular.

The vestibular system is phylogenetically old and neurologically hard-wired, therefore its function is inapparent during normal activity, and its output can be difficult to measure. Its purpose is to stabilize the head in space, to stabilize eye movements during movement of the head, and to help stabilize body movements during stance and gait. The balance system is not accorded the status it deserves alongside the other 'five senses'; because it works seamlessly in the background, its function is often not recognized until a deficit occurs, at which time normal activity becomes nearly impossible.

The vestibular system works in concert with two other afferent systems – the visual system and the somatosensory (proprioceptive and exteroceptive) system – to maintain balance. The three stabilizing systems overlap sufficiently to compensate (partially or completely) for each other's deficiencies. Vertigo may represent either physiologic stimulation or pathologic dysfunction in any of the three systems.

Anatomy and Physiology of the Vestibular System

A basic understanding of the vestibular system is essential. A brief summary is offered here; excellent reviews can be found elsewhere (see 'Further reading' section). The vestibular apparatus consists of five bilaterally paired organs, producing redundant responses from the right and left sides: the utricle and saccule ('otolothic' organs), which sense linear acceleration and head tilt, and the lateral, superior, and posterior semicircular canals, which sense angular acceleration in three planes of space (Figure 1).

The utricle and saccule are dilatations of the membranous labyrinth, each of which has an end organ called the macula. The macula of the utricle lies in the horizontal plane and senses to-and-fro acceleration, such as in a car. The macula of the saccule lies in the vertical plane and senses up-and-down movement, such as in an elevator. Each macula is covered by a gelatinous membrane, in which many small calcium carbonate crystals (called otoliths or otoconia) are embedded. The receptor cells, vestibular hair cells, project 50–70 cilia (stereocilia) up into the gelatinous membrane plus one large cilium, the kinocilium. The otoliths have a high specific gravity, and their weight will bend the cilia in the direction of the gravitational forces. The hair cells are organized in an orderly array toward or away from a central line in the macula called the striola. The hair cells have directional sensitivity. Bending of the stereocilia in the direction of the kinocilium leads to opening of ion channels in the cell membrane, causing depolarization; bending of the stereocilia away from the kinocilium causes closure of the ion channels and hyperpolarization (Figure 2). The hair cells synapse with sensory endings of the vestibular nerve. Normally, the first-order neuron has a resting tonic discharge, firing at a rate of 100 Hz. Bending of the cilia toward the kinocilium leads to an increase of the impulse rate, and bending of the cilia away from the kinocilium leads to a decrease of up to 0 Hz. In this way, the system provides neural encoding of both the intensity and the direction of the acceleration.

The three semicircular canals (SCCs) are arranged at right angles to one another so that they represent all three planes in space. The lateral (or horizontal) SCCs are horizontal when the head is bent 30° forward. The superior (or anterior) and posterior SCCs

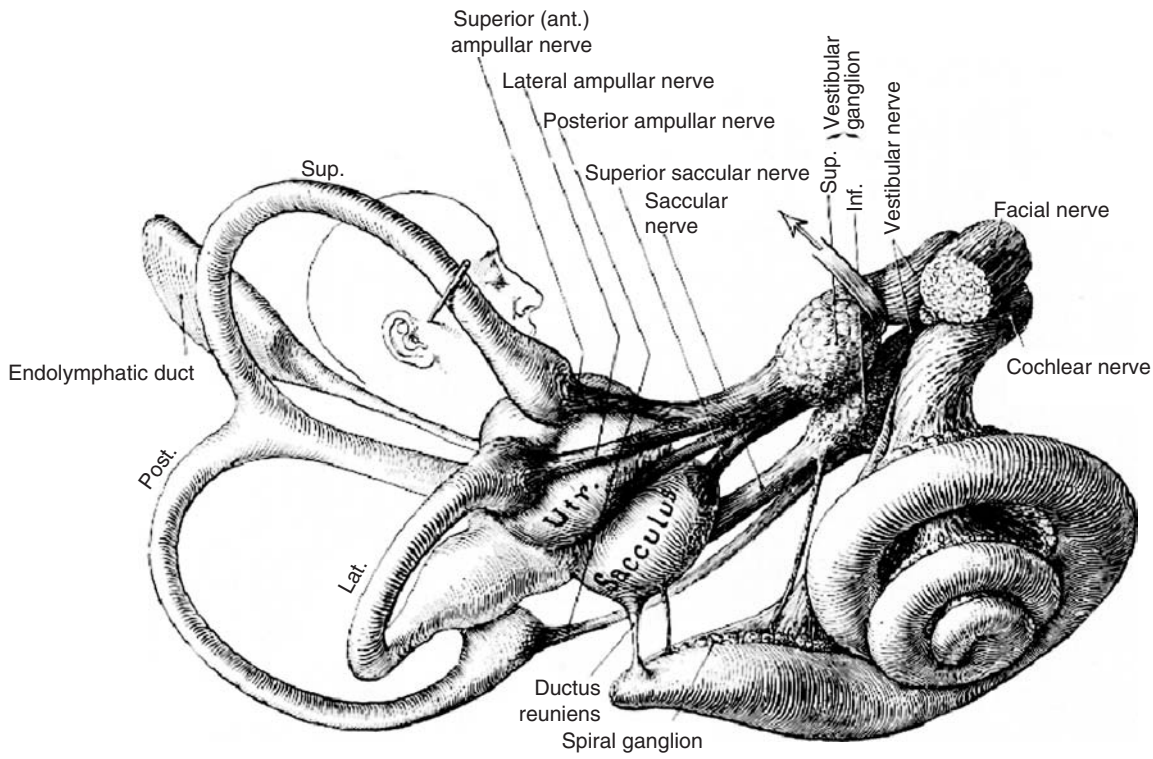

Figure 1 Classical drawing by Brodel of the membranous labyrinth of the inner ear.

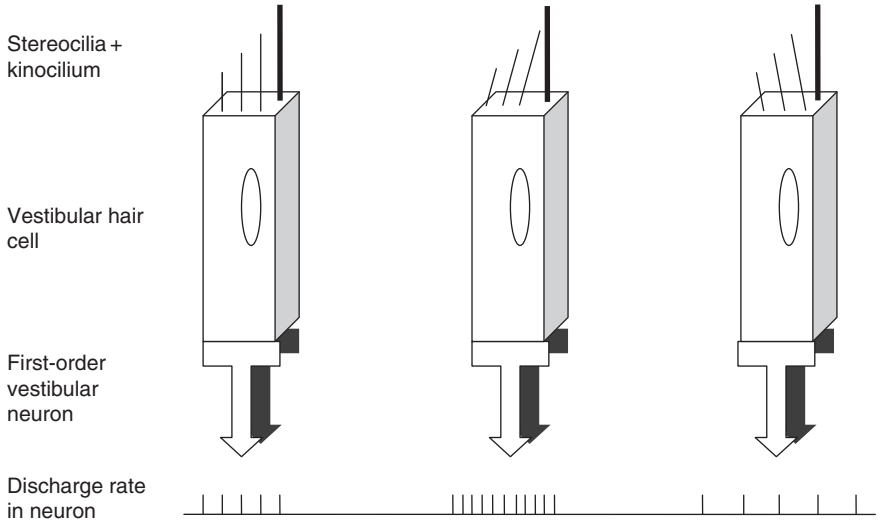

Figure 2 Diagram of the vestibular hair cell response. With the head at rest, there is a tonic resting discharge rate in the first-order vestibular neuron. Movement of the head in one direction causes distortion of the stereocilia toward the kinocilium, leading to membrane depolarization and an increase in the discharge rate of the first-order vestibular neuron. Movement in the opposite direction leads to bending of the stereocilia away from the kinocilium, resulting in membrane hyperpolarization and a decrease in the discharge rate. The neural response is therefore encoded for intensity and directionality.

are oriented in an oblique-vertical plane. The superior SCC projects 45° forward, while the posterior SCC projects 45° backward. The right superior and left posterior SCCs are oriented in the same plane, as are the left superior and right posterior canals, so that information provided by each pair is complementary; flow of endolymphatic fluid excites one SCC and inhibits its partner. The receptor organ in the SCC is the

crista ampullaris within the ampulla, a membranous dilatation of the SCC located near the vestibule. The crista is covered by a gelatinous mass of high specific density, the cupula. Hair cells are arranged in an orderly array along the crista, with their cilia embedded in the cupula.

Each SCC senses angular acceleration in its plane of space. When the head begins to rotate in any direction, inertia causes the fluid to remain stationary relative to the SCC. The relative fluid flow bends the cupula, causing depolarization of hair cells on one side and hyperpolarization on the other. As with the utricle and saccule, the first-order vestibular neuron from each SCC has a tonic resting discharge that increases with stimulation (or decreases with inhibition), providing neural encoding for the change in intensity and direction of angular movement. As the head begins to rotate, the rate of discharge increases. With continued rotation, the excess discharge of the hair cell gradually subsides as the cupula returns to its resting position.

Beyond simply providing awareness of rotational motion, the main function of the SCCs is to stabilize gaze during head movement. This is mediated through the vestibuloocular reflex (VOR). Without such a system, it would be impossible to maintain a stable image on the retina while the head is moving quickly through space. When the head is suddenly rotated, signals from the SCCs cause the eyes to rotate in a direction equal and opposite to the rotation of the head. In the case of head rotation in the horizontal plane, the flow of endolymph fluid causes one lateral SCC to be stimulated and its contralateral partner to be inhibited. The VOR pathway travels from the SCCs through the vestibular nuclei and the medial longitudinal fasciculus (MLF) to the oculomotor nuclei (III, IV, and VI), and crosses the midline to provide excitatory and inhibitory signals to both eyes, resulting in conjugate eye movements. After a horizontal head-turn, the VOR causes the ipsilateral medial rectus and the contralateral lateral rectus muscles to contract, and the ipsilateral lateral rectus and contralateral medial rectus muscles to relax, resulting in deviation of the eyes away from the direction of the head-turn. In the real world, head movements are complex and multiplanar. Each SCC responds to the vector of movement in its own plane of space, and each SCC controls eye movements in its own plane.

The VOR is one of four systems that control eye movements. The VOR is a high-gain system, and stabilizes the image of the target during fast head movement. The optokinetic system fixates the image during slow target movement, as the target 'slips' across the periphery of the retina. Saccades (eye jumps) provide for quick changes of visual targets. Smooth pursuit allows the eyes to follow a slow-moving target through space.

Nystagmus is a jerking eye movement that is generated by the vestibuloocular system. A rapid rotation of the head causes the eyes to deviate in an equal and opposite direction, in order to remain fixated on a visual target. A slow, sustained rotation of the head also causes the eyes to deviate, but when they reach the endpoint of gaze, a saccade automatically intervenes to return them to midline. With a sudden rotation, such as on a carousel, physiologic nystagmus occurs, in which a slow deviation of the eyes away from the direction of rotation is followed by a quick automatic reset to the midline. With continued rotation, the nystagmus stops (as the fluid within the SCC 'catches up' to the rotation of the head).

Physical Examination of Vertigo and Imbalance

Physical exam should consist of complete evaluation of the ears, nose, and throat and central nervous system (CNS), with particular attention to cranial nerves, eye movements, gait, and balance.

Nystagmus and Lesion Localization

Nystagmus that occurs spontaneously is an important sign of vestibular system disease, and is valuable for differentiating peripheral from central forms of vertigo. Peripheral vestibular disorders produce nystagmus that is unidirectional, with its quick component directed away from the side of the lesion, and with greatest intensity when gazing toward the quick component ('Alexander's law'). Peripheral nystagmus exhibits visual fixation suppression – that is, it is accentuated in darkness and suppressed when fixing on a visual target. Central causes of vertigo, on the other hand, produce nystagmus that may be bidirectional or direction-changing, and that does not suppress with visual fixation.

Because of visual fixation suppression, vestibular nystagmus may not be apparent at the time of physical examination. Frenzel glasses (backlit myopic lenses) or electronystagmography (ENG) using eye electrodes or infrared videography are helpful for detecting and recording clinically latent nystagmus. The head thrust test, performed by grasping the patient's head and applying a brief head thrust first to one side and then to the other, is a sign of a peripheral vestibular lesion. The test is abnormal when a head thrust toward the damaged side produces a corrective saccade, and signifies a decreased vestibular response on that side.

Clinical Correlations

Broadly speaking, there are five types of vertigo that have an otologic (peripheral vestibular) cause:

1. continuous vertigo, resulting from acute unilateral vestibular loss (e.g., vestibular neuritis),
2. episodic vertigo (Ménière's syndrome),
3. paroxysmal positional vertigo (benign paroxysmal positional vertigo),
4. bilateral vestibular loss (ataxia + oscillopsia), and
5. superior semicircular canal dehiscence (SCD) syndrome or perilymph fistula (PLF).

Labyrinthine vertigo must be differentiated from central causes. Vestibular migraine ('migraine-associated dizziness') has increasingly been recognized as a cause of episodic vertigo. The clinical pattern of vertigo is very important for the diagnosis of peripheral vestibular dysfunction. However, duration remains the most important clinical feature in the diagnosis.

Vertigo Lasting Days to Weeks

The syndrome of continuous vertigo of sudden onset, absence of auditory symptoms, and attendant vegetative symptoms is usually caused by vestibular neuritis (believed to be of viral origin), but the same symptom complex can also occur after any sudden, unilateral loss of vestibular function, such as temporal bone trauma, acoustic neuroma surgery, infarction, etc. Physical examination in the acute phase will reveal horizontal–rotatory nystagmus with the quick component directed away from the side of the lesion. The nystagmus will have visual fixation suppression, and may only be apparent with Frenzel glasses or ENG. Patients will also demonstrate past pointing (drifting of outstretched hand with eyes closed), falling toward the side of the lesion on Romberg testing (standing with feet together, eyes closed), and ataxic gait veering toward the side of the lesion. These acute findings normally subside within 3–5 days, but imbalance and gait ataxia may persist for 4–6 weeks. Initial management is supportive and symptomatic for vertigo and nausea. Complete resolution is expected to occur spontaneously as time passes – first by a cessation of the tonic resting discharge of the contralateral side, then by CNS relearning, as a gradual increase in the spontaneous discharge of the diseased side takes place. In cases of slow or incomplete CNS compensation ('uncompensated vestibular neuritis'), vestibular rehabilitation physical therapy is very effective.

Vertigo Lasting Minutes to Hours

Episodic vertigo is usually due to endolymphatic hydrops, distension of the inner ear membranes, and may be primary (Ménière's disease) or secondary (otic syphilis, delayed endolymphatic hydrops, Cogan's disease, recurrent vestibulopathy). Patients with typical Ménière's disease have recurring attacks of vertigo, sensorineural hearing loss, tinnitus, and, sometimes, a fluctuating fullness in the ear. The etiology is unknown. The disease is usually unilateral, but may become bilateral in 10–15% of cases. The hearing loss fluctuates in the early stage of the disease, and progresses as time goes on, often to a moderately severe level. The audiogram is the most useful diagnostic test, usually showing sensorineural loss at the low frequencies. ENG may show reduced caloric response in the affected ear, but is often normal.

Treatment consists initially of low-salt diet (<2000 mg of sodium day^{-1}) and diuretics (usually acetazolamide or a thiazide). The majority of patients respond to conservative medical treatment. In cases of failure of medical treatment, partial chemical labyrinthectomy can be performed using intratympanic gentamicin. The dosage and method of administration vary among practitioners, but biweekly injections of gentamicin (40 mg ml^{-1} for 30 min is typical) may be administered, with the endpoint of treatment being the elimination of vertigo or the development of imbalance or hearing loss. Treatment results in elimination of vertigo in 70% of patients, with a 30% chance of worsened hearing and 2% risk of deafness. Surgical treatments are also performed in refractory cases. Endolymphatic sac decompression is about 60% effective, with a low risk of complications. The putative mechanism of action is drainage of the distended endolymphatic space. Results may not be any better than placebo surgery. Vestibular nerve section is highly effective ($>85\%$) but requires an intracranial operation with attendant risk of cerebrospinal fluid (CSF) leak and neurologic compromise. Labyrinthectomy is nearly 100% effective in eliminating vertigo, at the expense of total loss of hearing. This procedure should never be performed if bilateral disease is suspected.

Patients with otic syphilis present with interstitial keratitis and Ménière's-like hearing loss vestibular symptoms. Treatment consists of penicillin and steroids. Cogan's syndrome patients have interstitial keratitis, Ménière's-like hearing loss, vestibular symptoms, and nonreactive tests for syphilis. The hearing loss is bilateral and progressive, without spontaneous improvement, and frequently becomes profound. Treatment is mainly corticosteroids and cyclophosphamide.

Patients with delayed endolymphatic hydrops present with symptoms identical to those of patients with Ménière's disease, except that their hearing loss is severe and precedes vertigo by years and may even be congenital. The episodic vertigo is

invariably separate from the initial hearing loss. Treatment is medical, the same as for Ménière's disease, but surgery should be avoided since the condition is often bilateral.

Vestibular Ménière's disease, or benign recurrent vestibulopathy, produces episodic vertigo similar to Ménière's disease, although without audiologic symptoms. Treatment is symptomatic and supportive when required. The disease is often self-limited. Many if not all of these cases may be caused by vestibular migraine (see later).

Vertigo Lasting Seconds

Benign paroxysmal positional vertigo is the hallmark of vertigo lasting only seconds. Typically the patient experiences a severe rotatory vertigo associated with a change in head position, such as lying, turning in bed, reaching for the top shelf, or bending. The Dix–Hallpike (or Nylen–Barany) positional maneuver is diagnostic, and reproduces this symptom, along with geotropic, rotatory nystagmus. The response is unilateral, occurs with the affected ear down, has a brief latency (<10 s) and short duration (15–45 s), and demonstrates fatigability with repeat testing. The pathophysiology is the migration of otoconia, small crystals of calcium carbonate, from the utricle less often to the lateral or superior SCC. BPPV is usually idiopathic but can occur after head trauma, chronic otitis media, vestibular neuritis, or age. BPPV should be differentiated from central disorders; positional nystagmus that is continuous, bilateral, or without significant vertigo, should raise suspicion.

BPPV is quite disturbing for patients but has a very good prognosis. The particle repositioning maneuver, also known as the Epley maneuver, is highly effective. In this maneuver, the patient's head is moved from a sitting position into the inciting Dix–Hallpike position, then to the opposite side, then 180° to a near-prone position, then upright again (see Figure 3). The cycle is then repeated until the nystagmus is completely extinguished. This procedure is 80–90% effective in eliminating vertigo, and works by redistributing the microparticles of calcium from the posterior SCC into the vestibule. Patients can also treat themselves at home (Brandt–Daroff maneuvers). Surgical treatment with posterior SCC occlusion or singular neurectomy is rarely necessary because the problem is usually self-limited.

Bilateral Vestibular Loss

Bilateral vestibular loss is almost always symptomatic and very debilitating; the most common causes are drugs (aminoglycosides or cisplatin) and idiopathic degeneration. Patients complain of oscillopsia (eye bobbing) with minimal head movements, inability to fixate the horizon while walking, severe disequilibrium, and gait instability. Physical examination reveals gait ataxia (broad based) and loss of dynamic visual acuity (loss of >2 lines on the Snellen eye chart while the head is passively rotated). Diagnosis is made by demonstrating loss of bilateral caloric response (including ice water) on ENG, and severely decreased gain on rotational chair testing. Patients compensate partially over time, using the visual and somatosensory systems; however, recovery is incomplete most of the time. Vestibular physical therapy is helpful but not always curative. Artificial vestibular prosthetic devices such as BrainPort have shown early promise.

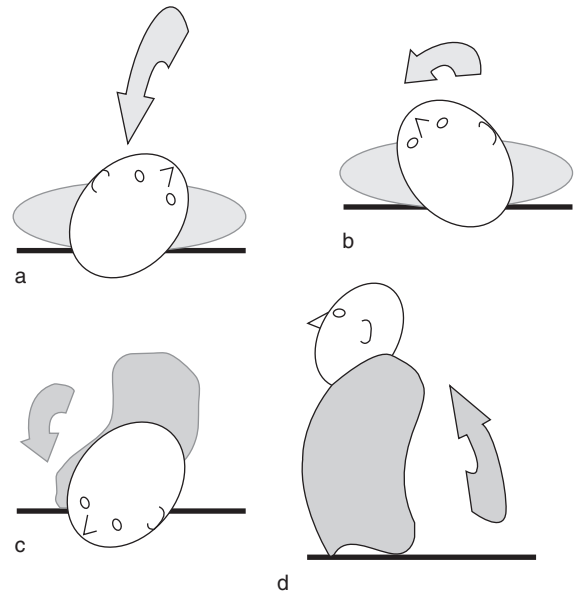

Figure 3 Epley maneuver for the treatment of benign paroxysmal positional vertigo. The patient's head is moved from a sitting position into the inciting Dix–Hallpike position (a), then to the opposite side (b), then 180° to a near-prone position (c), then upright again (d). Reproduced from Smouha EE (1997) Time course of recovery after Epley maiseurers for benign paroxymal positional vertîgo. Laryngoscope 107, 187–191, with permission from Lippincott Williams & Wilkins.

Vertigo of Variable Duration: Inner Ear Fistula and Superior Semicircular Canal Dehiscence Syndrome

Inner ear fistula Perilymphatic (inner ear) fistula (PLF) mimics other disorders, such as Ménière's disease, infection, acoustic neuroma, and lesions of the central nervous system. It can occur in the setting of trauma, cholesteatoma, or otosyphilis; spontaneous PLF probably does not occur. On examination, patients have vertigo and deviation of the eyes with external pressure on the tragus ('positive fistula test'). PLF after trauma is initial managed with bed rest,

head elevation, laxatives, and monitoring of hearing and vestibular function. In case of failure of conservative treatment, middle ear exploration is indicated, with fat patching of the oval/round window. PLF resulting from cholesteatoma should be treated by tympanomastoidectomy with or without removal of matrix from the lateral semicircular canal. Otosyphilis is diagnosed by Hennebert's sign (positive fistula test in the presence of an intact tympanic membrane) and a positive fluorescent treponemal antibody (such as FTA-Abs) test, and is treated with benzathine penicillin and steroids.

Superior semicircular canal dehiscence syndrome In 1998, Dr. Lloyd Minor first described superior canal dehiscence (SCD) syndrome. The typical signs of SCD syndrome are sound- and/or pressure-induced vertigo, and occasionally constant and often disabling disequilibrium and unsteadiness. The syndrome is caused by dehiscence of bone overlying the superior semicircular canal. These patients develop vertical–torsional eye movements (aligned with the plane of the superior semicircular canal on the affected side) in response to loud sounds or maneuvers that change middle ear or intracranial pressure. High-resolution computed tomography (CT) scan performed in the plane of the superior SCC is very important to confirm diagnosis. Surgical procedures through the middle cranial fossa approach, to plug or resurface the superior canal, are effective in resolving the symptoms.

Vertigo of Central Origin

It is possible for vertigo of CNS origin to exist without any other associated neurological signs. Vertigo that is continuously present for weeks without abating is generally not the result of a peripheral vestibular abnormality. CNS vertigo usually is less severe and more vaguely characterized, compared to peripheral vertigo, with mild vegetative responses. Physical examination may demonstrate nystagmus with central features: upbeating or downbeating nystagmus, direction-changing nystagmus on gaze or positional testing, and continuous (rather than paroxysmal) positional nystagmus. The absence of visual fixation suppression with Frenzel glasses or ENG is strongly suggestive of a central etiology.

Possible CNS causes of vertigo include vestibular migraine, vertebrobasilar insufficiency, cerebellar or lateral medullary infarction, and acoustic neuroma. Vestibular migraine (or 'migraine-associated dizziness') has become increasingly recognized as a cause of vertigo, and may account for many cases of episodic vertigo without auditory symptoms ('vestibular Ménière's disease). Vertigo attacks are spontaneous in onset, lasting minutes to hours, often accompanied by phonophobia or photophobia, and occur in patients with a history of motion sickness and/or classical migraine. The underlying pathophysiology may involve a defective calcium channel, primarily expressed in the brain and inner ear. Treatment with a variety of antimigraine drugs has been effective, including acetazolamide or topiramide, beta-blockers, and calcium channel blockers.

Cerebrovascular disease, including vertebrobasilar insufficiency and cerebellar or lateral medullary infarction, can also cause vertigo of the central type. Associated neurological signs such as dysarthria, incoordination, numbness, or weakness are usually present, but vertigo can also exist in isolation. Vertebrobasilar insufficiency can also cause drop attacks. MRI scanning is diagnostic.

Acoustic neuroma (or, more correctly, vestibular schwannoma) is a benign tumor that arises from the vestibular division of the eighth cranial nerve and involves the internal auditory canal and cerebellopontine angle. It usually presents with unilateral, progressive sensorineural hearing loss. Mild vestibular symptoms such as disequilibrium are common, but vertigo by itself is rarely a presenting complaint. Very large tumors can cause cerebellar compression leading to gait ataxia and, sometimes, coarse, bidirectional gaze-evoked nystagmus ('Brun's nystagmus'). Gadolinium-enhanced magnetic resonance image scan of the posterior fossa will demonstrate the pathology. Treatment is by surgical excision or stereotactic radiation.

Conclusion

Vertigo can be a vexing problem for physician and patient. A basic understanding of the physiology of the vestibular system, coupled with a knowledge of the common patterns of vestibular pathology, will lead to proper diagnosis and therapy in the majority of patients.

See also: Balance and Posture Control; Balance and Posture Control: Human; Nystagmus; Vestibular System; Vestibulo-Autonomic Responses; Vestibulo-Ocular Reflex; Vestibulospinal System and Eye–Head/Neck Movement; Visual–Vestibular Interactions.

Further Reading

Baloh RW (2002) Episodic vertigo: Central nervous system causes. *Current Opinion in Neurology* 15: 17–21.

Beck C and Schmidt CL (1978) 10 Years experience with intratympanally applied streptomycin (gentamicin) in the therapy of morbus Meniere. *Archives of Otorhinolaryngology* 221: 149–152.

Brandt T and Daroff RB (1980) Physical therapy for benign paroxysmal positional vertigo. *Archives of Otolaryngology* 106: 484–485.

Bretlau P, Thomsen J, Tos M, et al. (1982) Placebo effect in surgery for Meniere's disease. A double-blind placebo-controlled study on endolymphatic sac shunt surgery. *Advances in Otorhinolaryngology* 28: 139–146.

Carey JP and Della Santini C (2005) Principles of applied vestibular physiology. In: Cummings C (ed.) *Otolaryngology: Head & Neck Surgery,* 4th edn. St. Louis: Mosby.

Chia SH, Gamst AC, Anderson JP, et al. (2004) Intratympanic gentamicin therapy for Meniere's disease: A meta-analysis. *Otology & Neurotology* 25: 544–552.

Cohen HS and Kimball KT (2005) Effectiveness of treatments for benign paroxysmal positional vertigo of the posterior canal. *Otology & Neurotology* 26: 1034–1040.

Halmagyi GM and Curthoys IS (1988) A clinical sign of canal paresis. *Archives of Neurology* 45: 737–739.

Epley JM (1992) The canalith repositioning procedure: For the treatment of benign paroxysmal positional vertigo. *Otolaryngology – Head and Neck Surgery* 107: 399.

Minor LB (1999) Intratympanic gentamicin for control of vertigo in Meniere's disease: Vestibular signs that specify completion of therapy. *American Journal of Otology* 20: 209–219.

Minor LB (2000) Superior canal dehiscence syndrome. *American Journal of Otology* 21: 9–19.

Neuhauser H, Leopold M, von Bevern M, et al. (2001) The interrelations of migraine, vertigo, and migrainous vertigo. *Neurology* 56: 436–441.

Silverstein H, Norrell H, Smouha E, et al. (1990) An evolution of approach in vestibular neurectomy. *Otolaryngology – Head and Neck Surgery* 102: 374–381.

Silverstein H, Smouha E, and Jones R (1989) Natural history vs. surgery for Meniere's disease. *Otolaryngology – Head and Neck Surgery* 100: 6–16.

Smouha EE (1997) Time course of recovery after Epley maneuvers for benign paroxysmal positional vertigo. *Laryngoscope* 107: 187–191.

Smouha EE and Roussos C (1995) Atypical forms of paroxysmal positional nystagmus. *Ear, Nose & Throat Journal* 74: 649–656.

Tyler M, Danilov Y, and Bach-Y-Rita P (2003) Closing an open-loop control system: Vestibular substitution through the tongue. *Journal of Integrative Neuroscience* 2: 159–164.

Wanamaker HH, Gruenwald L, Damm KJ, et al. (1998) Dose-related vestibular and cochlear effects of transtympanic gentamicin. *American Journal of Otology* 19: 170–179.

Vesicle Pools

J B Sørensen, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

© 2009 Elsevier Ltd. All rights reserved.

Synaptic transmission is caused by the fusion of small neurotransmitter-containing synaptic vesicles at active zones. The steps leading to the fusion of a vesicle start with the invasion of the presynaptic terminal by an action potential (AP), which triggers the opening of voltage-gated calcium channels. The ensuing calcium influx leads to a local increase in the calcium concentration and the calcium-dependent exocytosis of synaptic vesicles. The liberated neurotransmitter diffuses across the postsynaptic cleft and activates postsynaptic receptors, leading to changes in the postsynaptic cell.

The Concept of a Vesicle Pool

Despite the large pool of synaptic vesicles, and hence neurotransmitters, present in most synapses, repetitive stimulation of a synaptic contact often leads to synaptic depression, whereby the synaptic output is decreased until reaching a plateau value, which can be maintained for a longer time. In the 1950s, this phenomenon was attributed to the depletion of a pool of available neurotransmitters, and it was later combined with the quantal hypothesis and the vesicle hypothesis to indicate the depletion of a pool of synaptic vesicles – often referred to as the readily releasable pool (RRP) – constituting only a small fraction of the total number of vesicles in the synapse. Upon arrival of an AP, vesicles from this pool fuse with the plasma membrane, whereas the other vesicles in the synapse do not. During high-frequency stimulation, the readily releasable vesicles are released faster than the pool is replenished, leading to depletion of the pool and decrease of the synaptic strength.

The vesicle pool concept has since been extended considerably to include a multitude of different pools. Today, a vesicle pool can be taken to mean a population of vesicles in a similar or identical state. The underlying assumption is that synaptic or secretory vesicles pass through a limited number of discrete states before fusing with the plasma membrane. However, the question of whether vesicles are in a similar state depends on which vesicle property is effectively assayed in a given investigation. Therefore, the vesicle pool terminology is partly overlapping. For instance, electrophysiologists measuring the postsynaptic output of a synapse are likely to focus on whether a vesicle is releasable or not, whereas a morphologist, who searches for the position of the vesicle in the synapse, will use other categories, such as whether the vesicle is found close to the plasma membrane ('docked vesicle pool') or not.

Quantitative Treatments of the Readily Releasable Pool Concept

A simple quantitative treatment of the RRP concept follows from Figure 1. If N is the number of vesicles in the RRP, p is the probability that a vesicle in the RRP will undergo release during an AP, and q is the postsynaptic current elicited by the fusion of one vesicle (the quantal size), then an AP will result in the postsynaptic current (PSC):

$$\mathrm{PSC} = N\ p\ q$$

Immediately after the release, the RRP size will be $N(1-p)$, and if the fraction of available vesicles released by an AP and the quantal size remain constant, then the next stimulus will results in a PSC:

$$\mathrm{PSC} = N(1-p)\ p\ q$$

where we have assumed that recovery of the RRP between stimuli is so slow that it can be neglected. The ratio of the first two PSCs in a high-frequency train is therefore $(1-p)$ if all our assumptions hold (which, however, will not be the case in many instances). The recovery of the RRP is often assumed to be reversible and follow an exponential time course, and hence the plateau level of RRP size (or synaptic strength) will be proportional to the recruitment speed and inversely depend on the frequency of the train and the release probability.

The previous treatment is an example of a vesicle state model, in which synaptic output or exocytosis strength is modeled as a function of the state in which the vesicles are in (release-ready or not). An alternative model type, often used during statistical/quantal analysis, is the release site model, in which the behavior of discrete release sites is considered. In this case, the equation

$$\mathrm{PSC} = N\ p\ q$$

still holds, but the meaning of N and p has changed: N is now the number of release sites, and p is the probability that a vesicle is released from a release site. The latter quantity will depend both on the probability that a release site is occupied and on the probability that the occupying vesicle will fuse. Thus, it

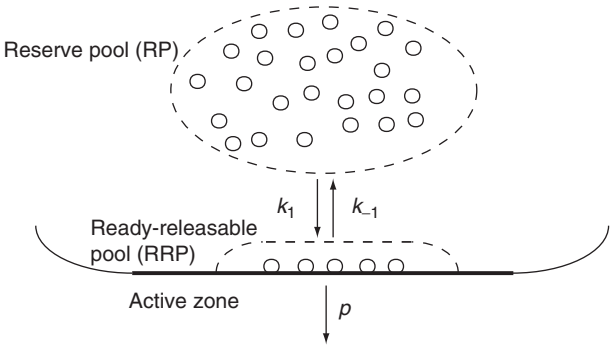

Figure 1 The simplest possible release model encompassing a readily releasable pool (RRP) and a reserve pool (RP). The RRP is refilled from the RP through a reversible process often referred to as priming. Upon arrival of an action potential, the vesicles in the RRP fuse with a uniform probability, *p*.

is important to consider which form of the model is assumed or implied in any given work. Generally, vesicle state models are popular for modeling exocytosis from neuroendocrine cells, where the number of vesicles seems to be rate limiting for release, whereas release site models are often used for neuronal synapses, which have many vesicles but a very limited number of release sites.

Measurements of RRP Size

The size of the RRP is an important quantity because it defines the total synaptic strength of a synapse, which does not rely on pool replenishment. Thus, it gives the total ability of a synapse to respond to fast stimulation. The size of the pool is most often measured by applying a depleting stimulus. The challenge is to be able to release all of the RRP fast enough so that the pool size is not contaminated with replenished vesicles, which would lead to overestimation of the size. Conversely, failure to release the whole RRP would lead to underestimation.

RRP Measurement by Repetitive Stimulation

By applying a fast train of APs, the ensuing synaptic depression can be used to estimate the size of the RRP, provided that a strong depression is seen (Figure 2(b)). In this method, the cumulative PSC amplitude (or charge) is plotted as a function of time (or stimulus number). The first AP results in relatively large PSCs, but as synaptic depression occurs, the PSC amplitude decreases and finally reaches a plateau, which represents the recruitment process. The cumulative plot therefore approaches a straight line. By fitting a line to the late part of the cumulative plot, the RRP size can be found by extrapolating to zero. The size of the first PSC divided by the RRP gives the initial release probability ($p = \mathrm{PSC}_1/\mathrm{RRP}$). The RRP size estimate will be more accurate the more depression is seen, that is, the higher the release probability (and the slower the recruitment). In order to increase depression, measurement can be done in the presence of an elevated extracellular calcium concentration.

RRP Measurement by Sustained Calcium Elevation

Another way of delivering a depleting stimulus is by increasing the presynaptic calcium concentration, either by sustained depolarization (caused, for instance, by high extracellular [K^+]) or by photorelease of caged Ca^{2+}. In the latter method, a photolysable calcium chelator is introduced into the cell, and a strong flash of UV light is used to elicit a jumplike increase in the calcium concentration. Simultaneous postsynaptic measurements, or measurements of the increase in cell membrane capacitance, if possible, reveal the size of the RRP (Figure 2(a)). An advantage of flash photolysis is that the calcium concentration increase is spatially homogeneous, which allows the simultaneous determination of the calcium dependence of the release process. The main disadvantage is that photorelease of caged calcium requires access to the presynapse, which has limited the use of this method to neuroendocrine cells and a few synapses which are large enough to be patch clamped directly.

RRP Measurement by Hypertonic Pressure

The application of an extracellular solution made hypertonic by sucrose addition results in the release of synaptic vesicles. If the solution is hypertonic by 500 mOsm, the release process is biphasic, with a large initial peak and a smaller sustained plateau (Figure 2(c)). This release pattern indicates that a specific pool of vesicles has been released. When integrating the PSC over time, the release of the pool is clearly set off from the slower process, which is assumed to represent pool refilling. It is usually stated that the vesicles released by hypertonic solution are 'primed' for release, and the pool is often denoted as – or even defined as – the RRP. It is not known exactly how an increase in hypertonic pressure causes synaptic vesicles to fuse, but it does not depend on calcium influx and it might be caused by a stress-induced conformational change in the exocytotic machinery.

RRP Measurement by Fluctuation Analysis

Since synaptic transmission is quantal in nature, analysis of the fluctuations during release can be used to obtain the quantal parameters *N*, *p*, and *q*. In its simplest form, this analysis takes its starting point in

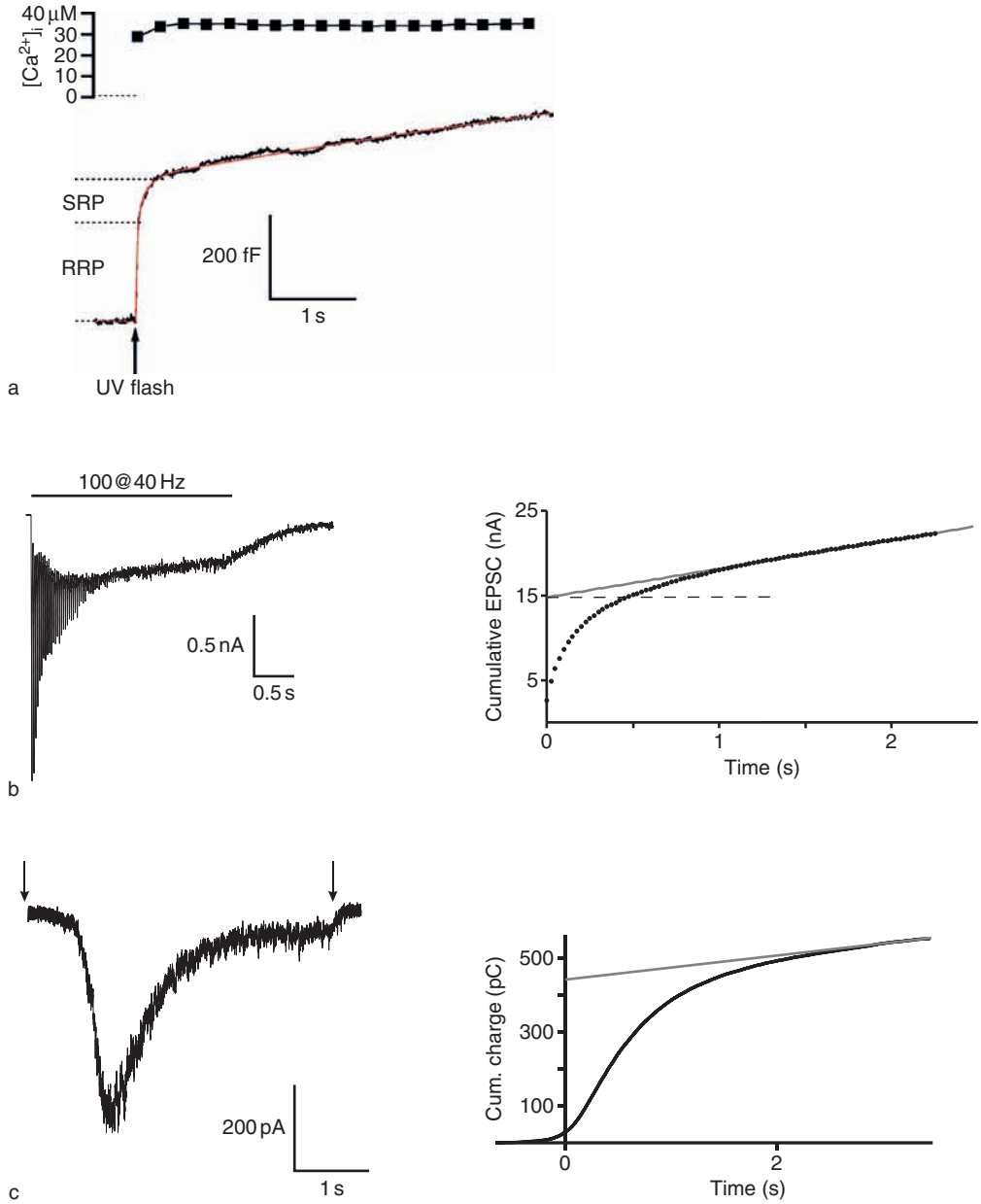

Figure 2 Three different ways of measuring the size of the readily releasable pool (RRP). (a) Flash photorelease of caged Ca^{2+} in combination with membrane capacitance measurements in a mouse chromaffin cell. The cell was patched in the whole-cell configuration and infused with calcium-sensitive dyes and a photolysable calcium chelator. At the time indicated by the arrow, the cell was subjected to a flash of UV light, which broke down part of the cage and led to an increase in $[Ca^{2+}]_i$ (top; preflash $[Ca^{2+}]_i$ is indicated by the dotted line). Immediately after the flash, the capacitance increased, indicating the fusion of vesicles with the plasma membrane. Fitting of a sum of exponentials (red curve) helped separate the size of the RRP and slowly releasable pool (SRP) from the near linear sustained-release component. (b) A high-frequency (40 Hz) train delivered to a patch clamped autaptic hippocampal neuron in order to deplete the RRP. The left panel shows the postsynaptic current (stimulus artifacts are blanked). It is easily seen that synchronized release displays fast run-down during the train. The right panel indicates the cumulative EPSC amplitude as a function of time. The gray regression line was fitted to the late part of the plot and extended toward zero. The intercept with the ordinate axis defines the size of the RRP. (c) A brief (3.5 s) pulse of extracellular solution made hypertonic by the addition of 500 mM sucrose delivered to a patch clamped autaptic hippocampal neuron. The release has a clear burst component, which subsides toward a steady level. The part of the trace between the arrows was integrated over time and displayed in the right panel. The integrated trace was fitted with a straight line, which was then extrapolated toward the beginning of the sucrose application. The intercept with the ordinate axis identifies the approximate size of the RRP. (a) Courtesy of Ralf Mohrmann. Data for (b) and (c) courtesy of Ignacio Delgado-Martínez.

the previous equation, which gives the mean peak PSC amplitude:

$$\mathrm{PSC} = N\ p\ q$$

According to binomial statistics, the variance of the PSC amplitude will be

$$\mathrm{Var} = N\ q^2\ p(1-p)$$

If we divide the variance with the mean, we get

$$\mathrm{Var/PSC} = q(1-p)$$

If the quantal size, q, is known, this relationship can be used to calculate first p and then N. The value of q can be estimated from analysis of spontaneous release events. The value of N obtained using this method probably corresponds to release sites rather than the number of release ready vesicles. A number of more complicated variants of fluctuation analysis exist, which are based on the analysis of covariance of PSC size or the analysis of higher cumulants.

On the Agreement between RRP Size Measurements

Given the different techniques used to measure RRP size, an important question is whether these measurements are in agreement with each other. In other words, are the vesicle pools measured and denoted 'RRP' in each case really identical, or do different methods measure different vesicle pools? Crucial for the answer to this question is the cross-depletion experiment in which the RRP is exhausted by one method (e.g., a train of APs) and then immediately probed using another method (e.g., direct calcium elevation). If the pool measured by both methods is the same, the second stimulation method should not elicit any response – or a very small response, due to some recovery between stimulations. Importantly, the cross-depletion experiment has to be performed with both sequences of the stimuli to ensure that the pool measurements agree and do not just overlap each other.

The most complete analysis of RRP size measured by different methods is available for the calyx of Held synapse in the auditory brain stem. This giant synapse is amenable to direct patch clamping of pre- and postsynaptic compartments, which allows detailed electrophysiological measurements, as well as capacitance measurements and flash photolysis of caged Ca^{2+}. For this synapse, repetitive stimulation gave an RRP size of 600–800 quanta. The repetitive stimulation method relies on the assumption that the quantal size, q, remains constant during a train, which might not necessarily be the case. Nonstationary PSC fluctuation analysis allows the determination of q and therefore estimation of RRP independently of differences in q. This gave an RRP size of 900 quanta. However, estimation by sustained calcium elevation gave larger numbers: 2200–2400 quanta when depolarization was used and 1800 quanta after photorelease of caged Ca^{2+}. These discrepancies may partly be due to differences in protocol because higher pool estimates were reached when cyclothiazide was used to prevent postsynaptic receptor desensitization and thereby changes in q. Another explanation is that during AP-driven release, vesicles situated inside calcium 'microdomains' around open calcium channels experience a high calcium concentration and fuse, whereas vesicles located farther away from calcium channels do not. However, during sustained calcium elevation, these 'distant' vesicles are also reached by the calcium signal and are fused. Yet another explanation is that certain vesicles might have an intrinsically lower release probability and are therefore only fused when the calcium signal persists over a longer period of time.

Another preparation used extensively for estimating RRP size is cultured hippocampal neurons. Being a culture system, it is amenable to the rapid application of hypertonic solutions, which represents an easy way of estimating RRP size. The analysis of sucrose and AP-elicited PSCs in autaptic cultures led to the conclusion that the pool released by sucrose is the same as that employed when release is evoked by APs. High-frequency stimulation led to a similar decrease of AP-evoked PSCs and sucrose responses (cross-depletion). Likewise, sucrose application similarly decreased AP-evoked PSCs and a second sucrose application (cross-depletion). The RRP as measured by sucrose application amounts to approximately 10–12 quanta per synapse. Notably, this number agrees fairly well with the number of docked vesicles measured by electron micrographs. In this case, therefore, primed vesicles might equal docked vesicles. In one study it was reported that when short 20 Hz trains were used to determine the RRP in glutamatergic neurons, the resulting pool estimate was much smaller than when determined by sucrose application. Increasing the length of the train and/or the extracellular calcium concentration made the RRP larger, as determined by repetitive stimulation. This is not surprising because, as discussed previously, RRP determination by train stimulation is better the larger the release probability. However, this result indicates that the RRP determined by different methods may be different. Most likely, during short trains vesicles distant to calcium channels are not released, whereas they are readily released by hypertonic stimulation or direct calcium elevation.

Replenishment of the RRP

Whereas the size of the RRP is important for the overall ability of the synapse to respond to high-frequency stimulation, the rate of RRP replenishment determines the ability to sustain transmitter output over a longer period of time. Thus, it is important to understand how RRP replenishment is regulated.

The most important finding is that high-frequency stimulation in itself stimulates RRP recovery, and the available evidence from cerebellar, hippocampal, and brain stem synapses indicates that this is due to the increase in intracellular $[Ca^{2+}]_i$ during trains, and in one case (the calyx of Held) it has been shown to depend on calmodulin. Thus, a stimulus that might lead to depletion of the RRP at the same time stimulates refilling in order to preserve the functionality of the synapse. Similar data have been obtained from neuroendocrine cells, such as adrenal chromaffin cells. In these cells, calcium-dependent refilling depends partly but not entirely on the activation of protein kinase C (PKC). In addition, refilling of the releasable pools is dependent on hydrolyzable ATP in both adrenal chromaffin cells and pancreatic β cells. An important difference between chromaffin cells and neurons is that in the former the basal $[Ca^{2+}]_i$ directly sets the size of the RRP, whereas in neurons the RRP size seems to be independent of $[Ca^{2+}]_i$ and only the refilling rate is calcium dependent. The former result is expected in vesicle state models, whereas the latter result is predicted by release site models because the number of release sites sets an upper boundary on RRP size.

Molecular Arrangement Underlying RRP Vesicles

Current research addresses the molecular machinery that underlies the formation of readily releasable vesicles. It is generally believed that so-called priming factors help preassemble an exocytotic machinery between the vesicle and plasma membrane, thereby making the vesicle release ready. The minimal calcium-dependent fusion apparatus seems to consist of the SNARE complex and one or more synaptotagmin isoforms, which act as calcium sensors for release. Most priming factors are believed to interact with these two classes of proteins.

One of the first priming factors to be identified was Munc13-1, whose elimination led to synaptic vesicles that were arrested at the priming step. According to one theory, Munc13-1 exerts its effect by binding to one of the SNAREs, syntaxin 1, in a conformation which assists the formation of the SNARE complex. Later investigations implicated many other proteins in vesicle priming, including the small monomeric G-protein Rab3 and the two Rab3-interacting proteins Rabphilin and RIM (Rab3-interacting molecule). It has also been suggested that the primed vesicle state might involve the formation of SNARE complexes in an intermediate state of assembly between the vesicle and plasma membranes. If this is the case, priming factors might exert their effect by modifying partial SNARE assembly.

Another important question is how the priming reaction and RRP size are regulated at the molecular level in the cell. Diacylglycerol, intracellular calcium, protein kinase A, PKC, and CamKII, as well as the lipid PIP_2, have been found to regulate RRP size and replenishment in different cells. However, the effector molecules for these second messengers have only been identified in a few cases. Phosphorylation of various presynaptic proteins has been implicated in different cell types. However, since the modus operandi of the priming and fusion machinery is not understood in detail, it is still not fully understood how these phosphorylation events regulate RRP size.

Other Vesicle Pools

From a physiological standpoint, RRP is often considered the most important vesicle pool because it is from this pool that vesicles are preferentially released during electrical stimulation. However, the vesicle pool concept is used to treat a number of other vesicle populations, corresponding to vesicles in different states.

The Reserve Pool

This pool concept is in most cases complementary to RRP, meaning all vesicles which are currently not in the RRP but which might be recruited through the process of priming. In later treatments, in which more releasable pools are assumed, the reserve pool usually denotes those vesicles not present in one of the more specific pools.

The Recycling Pool

The recycling pool of vesicles comprises all those vesicles which participate in synaptic recycling (i.e., both RRP vesicles and slowly releasable vesicles) as well as nonreleasable vesicles which are available for priming. Intrinsic to the concept of the recycling pool is the idea that a portion of the vesicles might not take part in active recycling under normal conditions but constitute a pool outside of the recycling pool, which might be recruited in certain circumstances. The recycling pool concept is often used in experiments in which synaptic vesicles are stained and destained with styryl dyes.

The Immediately Releasable Pool

The immediately releasable pool (IRP) is an important vesicle pool in many systems. The IRP consists of those release-ready vesicles which are closely co-localized with calcium channels. During depolarization, these vesicles experience the high calcium concentration within the so-called 'microdomains' or even 'nanodomains' close to open calcium channels. Consequently, they are more likely to fuse than are vesicles localized farther away. The only difference between these vesicles and the other RRP vesicles is the localization, and consequently the IRP can be considered a subset of the RRP. When stimulated by sustained and spatial homogeneous calcium elevation, the IRP and RRP vesicles fuse with identical kinetics. The estimation of the fraction of the RRP co-localized with calcium channels is done by stimulating the cell with repetitive short depolarizations to release the IRP, followed by stronger stimulation (e.g., by photorelease of caged calcium) to release the entire RRP. Such investigations have shown that in mouse chromaffin cells the IRP makes up approximately 25% of the RRP (35 vesicles out of 140). In contrast, in β cells almost all RRP vesicles are found co-localized with L-type calcium channels. Thus, the importance of this pool varies with the cell type.

The Slowly Releasable or 'Reluctant' Pool

Experiments carried out using photorelease of caged calcium revealed a vesicle population fusing with approximately tenfold slower kinetics than the RRP in chromaffin cells. This slowly releasable pool (SRP) seems to deliver the vesicles for filling the RRP since refilling of the RRP coincides with a temporary depression of SRP size. Thus, a linear arrangement of the SRP and the RRP was suggested. The RRP is absent in synaptotagmin 1 knockout mice, whereas the SRP persists, indicating that the association of synaptotagmin 1 with the exocytotic machinery is the distinguishing feature of RRP vesicles.

Also, neuronal synapses have slowly releasable or 'reluctant' vesicles. Depolarization of the calyx of Held terminal causes two phases of neurotransmitter release – a fast and a slow one. The faster releasable pool is replenished slowly, whereas the slower releasable pool displays faster replenishment. For this reason, the SRP might be important for maintaining neurotransmitter release during high-frequency stimulation. Furthermore, it has been found in both the calyx of Held and hippocampal neurons that the release probability of releasable vesicles is depressed soon after high-frequency stimulation (i.e., during recovery), which is in agreement with a faster recovery of reluctant vesicles or with the notion that vesicles have to pass through this state to become readily releasable.

An important outstanding problem is how the reluctant vesicles in neuronal synapses should be interpreted. There are two possibilities. One is that these vesicles are distinct from – or at least in a state distinct from – fast releasable vesicles and therefore characterized by a lower release probability. This is the same interpretation as for chromaffin cells. The other possibility is that the reluctant vesicles are identical to faster releasable vesicles except that they have a localization farther away from calcium channels. In the latter case, the vesicles would be part of the RRP but not part of the IRP, and strictly speaking it would not be necessary to operate with the concept of an SRP.

The Docked but Unprimed Pool

Comparisons of vesicle docking in electron microscopy (EM) micrographs and electrical measurements of the size of the RRP in neuroendocrine cells (chromaffin cells) have shown that the docked vesicle pool is much larger than the RRP. Also, inclusion of the SRP does not eliminate the discrepancy, so it is widely assumed that a pool of docked but unprimed vesicles must exist. In contrast, in hippocampal neurons the RRP size and the size of the docked pool seem to be in agreement. Thus, the docked but unprimed pool (UPP) might be a special feature of neuroendocrine cells. These vesicles might serve to replenish the SRP and RRP after release, even though this has not been directly shown. The notion of a UPP relies on the assumption that RRP and SRP vesicles are docked to the plasma membrane. This has been questioned in experiments in which recycling RRP vesicles were stained and localized using EM. These vesicles were found scattered in the vesicle cluster rather than docked to the plasma membrane. In neuroendocrine cells, visualization of fusing vesicles showed that vesicles inside the cell could undergo fusion. These and other observations have led to the claim that RRP vesicles might not be docked to the plasma membrane. However, in the first case, recycling vesicles and not the RRP vesicles were stained. In the second case, fusion of vesicles inside the cell occurred during the course of seconds after sustained calcium elevation. Since the recruitment of vesicles is calcium dependent, it is expected that under these conditions docking, priming, and fusion would occur within this time frame. When we restrict the concept of an 'RRP' to mean those vesicles that can be released within a few milliseconds (in neurons) or within a few tens of milliseconds (neuroendocrine cells), the conclusion that these vesicles must be predocked to the plasma membrane seems unavoidable.

The High Calcium-Sensitive Pool

In some neuroendocrine cells (chromaffin cells, β cells, and derived cell lines), a small pool of vesicles with a very high sensitivity to calcium has been described.

These vesicles are released fast at 1 or 2 μM of calcium, a concentration at which the fusion rate constant of RRP vesicles is almost two orders of magnitude slower. Cross-depletion experiments showed that these vesicles are distinct from IRP vesicles. Thus, they are not co-localized with calcium channels. Interestingly, the high calcium-sensitive pool (HCSP) seems to be more strongly regulated by phorbol ester and protein kinases than the RRP. This vesicle pool might represent a specialization to secretion in response to global calcium elevation due to release from internal stores. No counterpart to the HCSP has been described in neurons.

The Spontaneously Released Pool

Results from molecular studies of neurons have led to the identification of several manipulations that affect evoked release much more than spontaneous release. Generally, it seems that spontaneous release is more tolerant to changes in the composition of the exocytotic apparatus than evoked release. Studies of styryl dye staining/destaining of recycling vesicles showed that vesicles stained during spontaneous vesicle cycling were released with a slower intrinsic rate than vesicles stained with evoked activity. These and other observations led to the postulation of a special pool of spontaneously released vesicles that might recycle independently of the 'evoked pool' of vesicles. On the other hand, in the calyx of Held it was shown that the modification of spontaneous and evoked release by phorbol ester could be accounted for by the allosteric modification of the calcium sensor of a single pool of vesicles – the RRP. Thus, the spontaneously recycling pool concept remains controversial. The problem is that the spontaneous pool is defined by its slow release rate, and therefore it is not subject to rapid depletion and cross-depletion experiments. It is possible that another pool – for instance, the reluctant pool – could supply vesicles for spontaneous release. In certain circumstances – depending on the rate of interconversion between this pool and the RRP – this pool might appear kinetically isolated from evoked release.

Conclusion: The Pool Concept

The releasable vesicle pool concept has provided a useful framework for thinking about plasticity phenomena in neurons and neuroendocrine cells. The distinction between vesicle pool size, release probability or, equivalently, release rate constant, and pool replenishment allows for a detailed kinetic analysis, which can now be combined with molecular manipulations to identify the underlying causes. The extension of the pool concept to encompass any population of vesicles in similar or identical states has led to the identification or postulation of numerous vesicles pools, which has extended and refined our understanding but also made the situation somewhat more complicated and raised many questions regarding definitions. In this respect, it is important to remember that the definition of a vesicle pool is largely operational – that is, defined by the way in which the pool is assessed. With this in mind, the vesicle pool concept will remain fruitful to the field of neurotransmitter release.

See also: Active Zone; Exocytosis: Ca^{2+}-Sensitivity; Fusion Pore; Munc13 and Associated Molecules; Neuromuscular Junction (NMJ): Presynaptic Short-Term Plasticity of Neuromuscular Transmission; Neurotransmitter Release: Synchronous and Asynchronous; Rab3; Rab3A Interacting Molecules (RIMs) ; SNAREs; Synaptic Depression; Synaptic Plasticity: Short-Term Mechanisms; Synaptic Vesicles.

Further Reading

Barg S, Eliasson L, Renstrom E, and Rorsman P (2002) A subset of 50 secretory granules in close contact with L-type Ca^{2+} channels accounts for first-phase insulin secretion in mouse beta-cells. *Diabetes* 51(supplement 1): S74–S82.

Parsons TD, Coorssen JR, Horstmann H, and Almers W (1995) Docked granules, the exocytic burst, and the need for ATP hydrolysis in endocrine cells. *Neuron* 15: 1085–1096.

Rosenmund C and Stevens CF (1996) Definition of the readily releasable pool of vesicles at hippocampal synapses. *Neuron* 16: 1197–1207.

Sakaba T and Neher E (2001) Calmodulin mediates rapid recruitment of fast-releasing synaptic vesicles at a calyx-type synapse. *Neuron* 32: 1119–1131.

Sakaba T, Schneggenburger R, and Neher E (2002) Estimation of quantal parameters at the calyx of Held synapse. *Neuroscience Research* 44: 343–356.

Schikorski T and Stevens CF (2001) Morphological correlates of functionally defined synaptic vesicle populations. *Nature Neuroscience* 4: 391–395.

Schweizer FE and Ryan TA (2006) The synaptic vesicle: Cycle of exocytosis and endocytosis. *Current Opinion in Neurobiology* 16: 298–304.

Sorensen JB, Wiederhold K, Muller EM, et al. (2006) Sequential N- to C-terminal SNARE complex assembly drives priming and fusion of secretory vesicles. *EMBO Journal* 25: 955–966.

Stevens CF and Wesseling JF (1998) Activity-dependent modulation of the rate at which synaptic vesicles become available to undergo exocytosis. *Neuron* 21: 415–424.

Voets T, Neher E, and Moser T (1999) Mechanisms underlying phasic and sustained secretion in chromaffin cells from mouse adrenal slices. *Neuron* 23: 607–615.

Wang LY and Kaczmarek LK (1998) High-frequency firing helps replenish the readily releasable pool of synaptic vesicles. *Nature* 394: 384–388.

Wu LG and Borst JG (1999) The reduced release probability of releasable vesicles during recovery from short-term synaptic depression. *Neuron* 23: 821–832.

Yang Y, Udayasankar S, Dunning J, Chen P, and Gillis KD (2002) A highly Ca^{2+}-sensitive pool of vesicles is regulated by protein kinase C in adrenal chromaffin cells. *Proceedings of the National Academy of Sciences of the United States of America* 99: 17060–17065.

Zucker RS and Regehr WG (2002) Short-term synaptic plasticity. *Annual Review of Physiology* 64: 355–405.

Vesicular Neurotransmitter Transporters

R J Reimer, Stanford University School of Medicine, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A little more than 50 years ago, Katz and colleagues demonstrated that acetylcholine was released at the frog neuromuscular junction in packets containing thousands of molecules. It was later proposed that the neurotransmitter is stored at the nerve terminal in small vesicles and that the fusion of these vesicles with the plasma membrane leads to release of the neurotransmitter. Subsequent physiological, anatomical, biochemical, and molecular studies have established that synaptic transmission, defined by the quantal release of neurotransmitters through exocytosis of synaptic vesicles, is the primary mechanism for information transfer in nervous systems of all organisms.

Because classic neurotransmitters such as acetylcholine are synthesized in the cytoplasm, a mechanism is required for their uptake into synaptic vesicles. Storage of neurotransmitter in synaptic vesicles has a number of theoretical advantages. First, it permits the simultaneous release of thousands of neurotransmitter molecules and allows a rapid rise in the synaptic cleft concentration. Second, vesicular storage sequesters neurotransmitter away from the cytoplasm. The maintenance of low cytoplasmic concentrations facilitates the efficient reuptake of released neurotransmitter and, in the case of monoamines, may reduce toxicity. Third, it provides a large pool of releasable neurotransmitter in nonrecycling reserve vesicles. Last, it imparts the potential for the simultaneous co-release of neuromodulators with a neurotransmitter and even co-release of two different classes of neurotransmitters.

Biochemical studies defined synaptic vesicle transport activities that mediate the uptake of the classic neurotransmitters. Four unique transport systems have been characterized: one for monoamines, a second for acetylcholine, a third for γ-aminobutyric acid (GABA) and glycine, and a fourth for glutamate. The transport processes facilitate the rapid refilling of vesicles and can generate lumen-to-cytoplasm concentration gradients of greater than 10^4. Recent molecular studies have identified the proteins that catalyze these processes. As expected, separate multitransmembrane domain proteins mediate each of the four biochemically described activities (Table 1 and Figure 1). The characterization of these proteins has generated specific molecular markers for cells that release a given neurotransmitter and has led to the identification of potential mechanisms for regulating neurotransmitter release.

Vesicular Neurotransmitter Transport Is Driven by a Vesicular Proton Electrochemical Gradient

Unlike Na^+-dependent plasma membrane transporters, the vesicular activities couple uptake to a proton electrochemical gradient ($\Delta\mu H^+$) across the vesicle membrane (Figure 1). This gradient is generated by a vacuolar-type H^+-adenosine triphosphatase (ATPase) that couples the hydrolysis of one adenosine triphosphate (ATP) molecule to the inward transmembrane movement of two protons. Recent data suggest that there is a single ATPase on each synaptic vesicle. Each pump is a multimeric complex containing cytoplasmic (V1) and transmembrane (V0) domains, with both domains composed of several polypeptide subunits.

The proton electrochemical gradient generated by the H^+-ATPase has a chemical (ΔpH) and electrical ($\Delta\psi$) component. Under normal physiological conditions there is a 1–2 pH-unit gradient and a 40–80 mV potential across the synaptic vesicle membrane. Due to the size of synaptic vesicles, a membrane potential (inside positive) is rapidly generated as protons are transported into the vesicle. This creates an energy barrier for the inward movement of protons that cannot be overcome by the H^+-ATPase and limits the pH gradient that can be achieved. However, anions such as Cl^- can permeate the synaptic vesicles membrane and reduce $\Delta\psi$. The reduction in membrane potential lowers the energy barrier and facilitates further H^+-ATPase activity with formation of a larger ΔpH. It has been suggested that a major regulatory mechanism for acidification of intracellular vesicles, including synaptic vesicles, is the flux of Cl^- through channels. The primary mediator of the Cl^- conductance in synaptic vesicles is believed to the ClC3, a member of the ClC family of chloride channels.

The dependence of all of the vesicular neurotransmitter transport activities on $\Delta\mu H^+$ is well established. Bafilomycin, a specific inhibitor of the vacuolar H^+-ATPase, the proton ionophores carbonyl cyanide *m*-chlorophenylhydrazone (CCCP), and carbonylcyanide *p*-(trifluoromethoxy) phenylhydrazone (FCCP) reduce vesicular storage of all neurotransmitters. The dependencies of the different transport systems on the chemical and electrical components of $\Delta\mu H^+$ have been determined by selectively reducing ΔpH and $\Delta\psi$. Nigericin, a H^+/K^+ exchanger, and weak bases such as ammonium sulfate are used to selectively

Table 1 Synaptic vesicle transporters

Transporter	*SLC designation*	*Substrate*	*Affinity*	*Predicted stoichiometry*	*Expression pattern*	*Sample inhibitors*
VMAT1	SLC18A1	Monoamines	Low micromolar	$2H^+/1MA^+$ exchange	Adrenal medulla, enterochromaffin cells	Reserpine
VMAT2	SLC18A2	Monoamines including histamine	Low micromolar	$2H^+/1MA^+$ exchange	Monoaminergic neurons in the brain, sympathetic postganglionic neurons, mast cells	Reserpine, tetrabenazine
VAChT	SLC18A3	Acetylcholine	Low millimolar	$2H^+/1ACh^+$ exchange	Cholinergic neurons in the central, peripheral, and autonomic nervous systems	Vesamicol
VGAT/VIAAT	SLC32A1	GABA and glycine	Low millimolar	Unknown	Inhibitory neurons, islet cells	γ-Vinyl GABA
VGLUT1	SLC17A7	Glutamate	Low millimolar	Unknown	Neocortex, cerebellar cortex, hippocampus	Evans blue, rose Bengal, DIDS
VGLUT2	SLC17A6	Glutamate	Low millimolar	Unknown	Thalamus, cerebellar and brain stem nuclei	Evans blue, rose Bengal, DIDS
VGLUT3	SLC17A8	Glutamate	Low millimolar	Unknown	Astrocytes, monaminergic, GABAergic, cholinergic neurons	Evans blue, rose Bengal, DIDS
ZnT3	SLC30A3	Zinc	Unknown	Unknown	Glutamatergic neurons	
ClC3		Chloride		Unknown	Broad	DIDS
Vacuolar H^+-ATPase			~ 0.1 mM	$2H^+$/cycle	Broad	Bafilomycin
SV2 (A, B, and C)		Unknown			All neurons and neuroendocrine cells	

ACh, acetylcholine; ATP, adenosine triphosphate; DIDS, 4,4′-diisothiocyanatostilbene-2,2′-disulfonic acid; GABA, γ-aminobutyric acid; MA, monoamine; SV2, synaptic vesicle protein 2; VAChT, vesicular acetylcholine transporter; VGAT, vesicular GABA transporter; VGLUT, vesicular glutamate transporter; VIAAT, vesicular inhibitory amino acid transporter; VMAT, vesicular monoamine transporter.

reduce ΔpH, whereas valinomycin, a K^+ ionophore, and increasing concentrations of permeant anions are used to selectively reduce $\Delta\psi$. With these approaches, it has been shown that the vesicular monoamine and acetylcholine transport systems depend primarily on ΔpH, GABA transport relies on both ΔpH and $\Delta\psi$, and glutamate transport relies almost exclusively on $\Delta\psi$ (Figure 1). These bioenergetic differences presumably reflect a combination of differences in substrate charge and the stoichiometry of coupling.

In addition to its role in proton translocation, the V0 domain has also been implicated in formation of the exocytotic fusion pore. Earlier biochemical studies identified the membranous subunit of the vacuolar H^+-ATPase as the mediatophore, a membrane protein complex that facilitates the release of acetylcholine at the synaptic cleft. This finding has been complemented by more recent genetic studies with *Saccharomyces cerevisiae* and *Drosophila melanogaster* that implicate the V0 domain in formation of membrane fusion pores. To form part of the fusion pore requires the dissociation of V1 and V0. In yeast, regulated dissociation of the two complexes has been shown to occur, suggesting that a similar process could occur at the synapse. Dissociation of V1 could allow for V0 on synaptic vesicles to interact with other membrane structures such as the plasma membrane, but could also function as a mechanism for limiting the vesicular proton electrochemical gradient and neurotransmitter accumulation.

Vesicular Monoamine Transport Is Mediated by Two Homologous Proteins

Studies on chromaffin granules, large intracellular vesicles from chromaffin cells of the adrenal gland, led to the characterization of a transport activity that mediates the vesicular accumulation of monoamines, including dopamine, norepihephrine, epinephrine, and serotonin. The subsequent molecular identification of the vesicular monoamine transporter resulted from a screen for genes from PC12 cells (an adrenal gland-derived cell line) that confer resistance to the parkinsonian neurotoxin MPP^+ in a sensitive cell line. MPP^+ bears a structural similarity to the monoamine neurotransmitters, and the resistance apparently results from

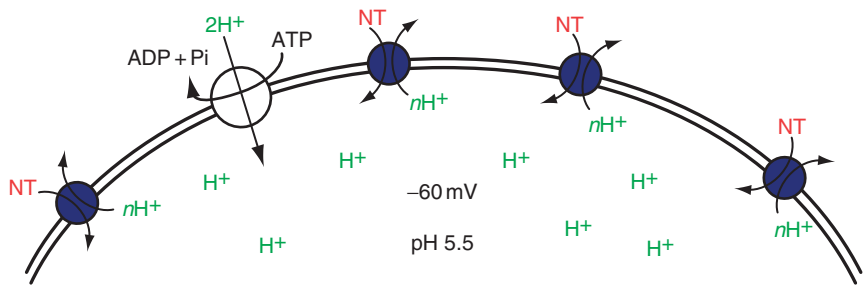

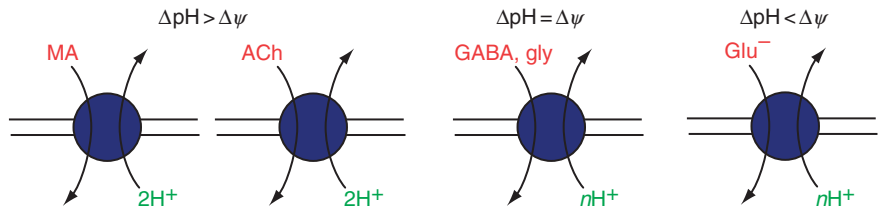

Substrate charge	+1	+1	0	–1
Protons per cycle	2	2	1 – 2	1 – 2
Net charge per cycle	+1	+1	+1 – 2	+2 – 3

Figure 1 Stoichiometry and bioenergetic dependence of the vesicular neurotransmitter transporters. A vacuolar type H^+-ATPase on synaptic vesicles generates a proton electrochemical gradient. The gradient consists of a pH gradient and a transmembrane potential of approximately -60 mV (inside positive). The different transporter families have different dependencies on the two components of the gradient – VMATs and VAChT are most dependent on the pH gradient, VGAT can be driven by either the pH or electrical gradient, and VGLUTs are most dependent on the electrical gradient. For monoamine and acetylcholine transporters, the stoichiometries have been determined, but for VGAT and VGLUTs the specifics of energetic coupling remain to be determined. $\Delta\psi$, electrical gradient; ADP, adenosine diphosphate; ACh, acetylcholine; ATP, adenosine triphosphate; GABA, γ-aminobutyric acid; Glu, glutamate; gly, glycerin; MA, monoamine; NT, neurotrophin; Pi, inorganic phosphate; VAChT, vesicular acetylcholine transporter; VGAT, vesicular GABA transporter; VGLUT, vesicular glutamate GABA transporter; VMAT, vesicular monoamine transporter.

the sequestration of the toxin inside vesicles, away from its primary site of action in mitochondria. The cloned vesicular monoamine transporter VMAT1 is predicted to have 12 transmembrane domains and is structurally related to the bacterial drug-resistance transporters. In addition to mediating the uptake of MPP^+, VMAT1 also transports dopamine, norepinephrine, epinephrine, and serotonin. VMAT1 is expressed by cells of the adrenal medulla, by neurons in sympathetic ganglia, and by other nonneural cells that release monoamines. In contrast, neuronal populations in the nervous system express the closely related VMAT2. The substrate specificity for the two isoforms is similar, and the apparent affinities for the monoamines are in the low micromolar range for both transporters, though VMAT2 has a slightly higher apparent affinity for all monoamines. In addition, only VMAT2 appears able to transport histamine at physiological concentrations, consistent with its expression by histamine-releasing mast cells. The high affinity of the VMATs may reflect a need to keep cytosolic concentrations of the relatively toxic monamines low.

Transport by the VMATs involves the exchange of two lumenal protons for one protonated molecule of transmitter carrying a net single positive charge and, thus, the outward movement of one net positive charge. For a vesicular pH gradient of ~ 1.5 pH units and membrane potential of ~ 60 mV, this predicts the accumulation of transmitter inside vesicles at concentrations 10^4–10^5-fold greater than those in cytoplasm. This is consistent with the micromolar and near-molar catecholamine concentrations measured in the cytosol and vesicles, respectively. Because osmotic forces limit the free transmitter concentration to ~ 150 mM, a fraction of the measured concentration, it has been suggested that the formation of macromolecular complexes consisting of a proteinaceous core, ATP, and the catecholamines reduces the effective osmolarity inside the vesicle.

The broad specificity of these transporters may be of particular importance in epinephrine-releasing cells. In these cells, VMATS transport dopamine into secretory vesicles, where the transmitter is converted to norepinephrine by the action of the dopamine β-hydroxylase, an enzyme restricted to the vesicle lumen. Norepinephrine then exits the vesicle, most likely through the action of VMATs, into the cytoplasm where it is converted to epinephrine by the action of phenylethanol-amine-*N*-methyltransferase. Epinephrine is then transported back into vesicle,

again presumably by VMATs. Although these last two steps probably occur through heterologous exchange (e.g., lumenal norepinephrine is exchanged for cytoplasmic epinephrine with no net proton movement), epinephrine released by exocytosis passes through VMATs three times.

Two well-characterized inhibitors act on the VMATs: reserpine and tetrabenazine. Both VMATs are irreversibly inhibited by reserpine, which initially showed clinical promise as an antihypertensive medication. The depressive effects of reserpine led to its more limited use but helped to formulate the original monoamine hypothesis of affective disorders. Reserpine appears to interact with the transporters near the site of substrate recognition. Tetrabenazine reversibly inhibits VMAT2, but is markedly less effective as an inhibitor of VMAT1. It has a limited clinical use in the treatment of some movement disorders. Amphetamines and 3,4-methylenedioxy-*N*-methylamphetamine (MDMA, or extasy) act as weak bases to reduce the vesicular ΔpH. Loss of this driving force for the vesicular transporters leads to the efflux of the lumenal contents into the cytoplasm and subsequent release through reversal of the plasma membrane transporters. The specificity of their effects (amphetamines lead to dopamine and, to a lesser extent, norepinephrine release, whereas MDMA leads to serotonin release) is thought to be due to selective uptake by the plasma membrane transporters. The association of polymorphisms in the VMAT1 gene with an increased risk for bipolar disorder also suggests a link between vesicular monoamine transport and molecular psychopathology.

A major mechanism for the regulation of vesicular neurotransmitter transport appears to involve changes in protein trafficking. VMATs undergo phosphorylation by casein kinase, and this posttranslational modification influences their retrieval from maturing large dense-core vesicles (LDCVs). Because sorting to LDCVs versus synaptic vesicles determines the site and mode of transmitter release, the regulation of transporter trafficking has great potential to influence signaling. In particular, differential trafficking of VMAT2 in midbrain neurons can lead to the release of dopamine from exocytosis at their cell bodies and dendrites, if trafficked to LDCVs, or axon terminals, if trafficked to synaptic vesicles. Regulation of VMAT activity by the heteromeric G-protein $G_{\alpha o2}$ has also been demonstrated. It appears that intralumenal monoamines activate the G-protein and downregulate uptake. This may be a mechanism to assure consistency in quantal size and may also serve to limit the efflux of the neurotransmitter through reverse transport once a vesicle has been filled.

The Vesicular Acetylcholine Transporter Is Structurally Related to the Vesicular Monoamine Transporters

The electric ray organ of *Torpedo californica* was an early source of synaptic vesicles. Subsequent to the determination that these vesicles are filled with acetylcholine, a specific activity for acetylcholine transport was characterized. The molecular identification of the transporter, however, required genetic studies in *Caenorhabditis elegans*. A screen for nematode mutants resistant to aldicarb, an inhibitor of acetylcholinesterase, identified *unc17*, a gene encoding a protein closely related to the VMATs. Based on the similarity to the VMATs, it was predicted that unc17 functions as an acetylcholine transporter and that the resistance phenotype results from the reduced release of acetylcholine. The binding of vesamicol, a known inhibitor of vesicular acetylcholine transport, was also found to be absent in membrane fractions from the *unc17* animals. Subsequent studies on the mammalian ortholog demonstrated that the protein is indeed a vesicular acetylcholine transporter.

In vertebrates, the vesicular acetylcholine transporter VAChT is expressed in cholinergic neurons in the central, peripheral, and autonomic nervous systems, including the basal forebrain neurons, lower motor neurons, and parasympathetic neurons. In an interesting genomic structure, the VAChT gene lies within the first intron of the gene encoding choline-acetyltransferase, the enzyme that catalyzes the synthesis of acetylcholine from choline and acetyl CoA. This structure, which suggests a spatial component to transcriptional regulation of factors conferring cholinergic characteristics to a neuron, is conserved from *C. elegans* to humans.

Like VMATs, VAChT recognizes a cationic substrate and depends primarily on ΔpH. VAChT has been postulated to have a similar stoichiometry of one cytosolic ACh^+ exchanged for two luminal H^+. In contrast to VMATs, which have substrate affinities in the low micromolar range, the mammalian VAChT exhibits an apparent affinity of approximately 1 mM. The higher affinity of VMATs may reflect a need to keep cytoplasmic levels of potentially toxic monoamine transmitters low. The measured quantal size at the neuromuscular junction ($\sim$10 000 molecules) is approximately fivefold larger than that expected for the 150 mM vesicular acetylcholine concentration predicted by osmotic limits. This suggests that, as with monoamine storage, mechanisms exist to limit the osmotic effects of acetylcholine in synaptic vesicles.

The vesicular acetylcholine transport can be inhibited by vesamicol and several related compounds. Vesamicol competitively inhibits transport by binding to a cytoplasmic domain on VAChT with a K_d of approximately 5 nM. Vesamicol binding can be used to estimate transporter number, but neither vesamicol nor its analogs are currently used clinically.

Vesicular GABA and Glycine Transport Activities Are Mediated by a Single Protein

The primary excitatory and inhibitory neurotransmitters in the mammalian brain are the amino acids glutamate and GABA. There are no sources of readily purified glutamatergic or GABAergic vesicles, as there are for monoaminergic and cholinergic vesicles (e.g., chromaffin granules and electric organ synaptic vesicles, respectively). Early biochemical studies defining vesicular transport systems for these neurotransmitters were, therefore, carried out on mixed synaptic vesicles purified from either the bovine or rodent brain. As for VMATs and VAChT, molecular characterization of the vesicular glutamate transporters (VGLUTs) and vesicular GABA transporter (VGAT) relied on additional information from molecular genetic studies.

The molecular identity of the vesicular VGAT was determined from a screen for *C. elegans* genes involved in GABAergic synaptic transmission. One gene identified in the screen, *unc47*, encodes a multitransmembrane domain protein that is localized to synaptic vesicles in GABAergic neurons. The functional characterization of a vertebrate ortholog confirmed that the protein mediated vesicular GABA transport. As predicted from studies with purified synaptic vesicles, VGAT mediates the uptake of GABA with an apparent affinity in the low millimolar range and uptake can be driven by ΔpH or $\Delta\psi$. VGAT also recognizes glycine as a substrate, but with a lower affinity, and is therefore also referred to as for vesicular inhibitory amino acid transporter (VIAAT).

VGAT shows no sequence similarity to VMATs or VAChT. Rather, it belongs to a large family of amino acid transporters. This family includes the proteins responsible for the transport activities biochemically defined as amino acid transport system N and system A. The latter is responsible for much of the active amino acid uptake by mammalian cells. Characterization of the function and cellular and subcellular localizations of these transporters suggests that the system N transporters and system A transporters mediate glial glutamine release and neuronal uptake, respectively, in the glutamine–glutamate cycle. This intercellular metabolic pathway is involved in the recycling of synaptically released glutamate and, to a lesser extent, GABA.

VGAT is expressed in GABAergic and glycinergic neurons as well as in the pancreas and has been reported to be expressed in the glial cells of the pineal gland. The targeted disruption of the VGAT gene in mice leads to a marked reduction in the synaptic release of GABA and glycine. The loss of the transporter leads to embryonic lethality, failures in gut withdrawal, and formation of a cleft palate. It is suggested that these defects are secondary to the loss of inhibitory neurotransmission and resulting paralysis.

Vesicular GABA transport can be competitively inhibited by amino acids, including glycine and β-alanine. Transport can also be competitively inhibited by γ-vinyl GABA, a derivative of GABA. γ-Vinyl GABA has been used in the clinical treatment of epilepsy and is known to inhibit GABA transaminase, an enzyme that metabolizes GABA. The mode of action for γ-vinyl GABA as an antiepileptic drug is thought to be through its effects on GABA transaminase, but the inhibition of vesicular GABA transport could have an effect by increasing the nonvesicular release of GABA.

Vesicular Glutamate Transporters Define Glutamatergic Neurons

Glutamate is the principal excitatory neurotransmitter in the mammalian central nervous system. Three unique vesicular glutamate transporters (VGLUT1, -2, and -3) have been identified. VGLUT1 was initially isolated in a screen for mRNA transcripts upregulated in cultured neurons by subtoxic doses of the excitotoxin *N*-methyl-D-aspartate (NMDA). Structural similarity to a class of proteins characterized as inorganic phosphate transporters led to an initial impression that the protein was a Na^+-dependent phosphate transporter and to a temporary designation of the protein as brain-specific Na^+ dependent inorganic phosphate transporter (BNPi). Although subsequent studies have demonstrated that the primary function of the protein is synaptic vesicle glutamate transport, some controversy remains regarding the role of the protein in phosphate transport.

The expression of VGLUT1 is limited to a subset of glutamatergic neurons in the brain. The highly homologous protein VGLUT2, also initially characterized as a phosphate transporter and named differentiation-associated Na^+-dependent inorganic phosphate transporter (DNPi), has a nearly complementary pattern of expression with one of the two proteins present in all established glutamatergic neurons. In general, VGLUT1 predominates in the neocortex and cerebellar cortex, whereas VGLUT2 predominates in the

brain stem nuclei, thalamic nuclei, and cerebellar deep nuclei. The septal nuclei, nuclei of the diagonal band, and hypothalamus also express VGLUT2. Although all cortical layers express VGLUT1, layer IV of frontal and parietal cortex and layers IV and VI of temporal cortex also express VGLUT2. Conversely, VGLUT2 predominates in the thalamus, but certain thalamic nuclei such as the medial habenula express VGLUT1. In the hippocampus, dentate gyrus granule cells express only VGLUT1, whereas pyramidal neurons from CA1 through CA3 express VGLUT1 as well as lower levels of VGLUT2. In the amygdala, the medial and central nuclei express VGLUT2, and the lateral and basolateral nuclei express VGLUT1.

The third vesicular glutamate transporter, VGLUT3, is expressed in neurons not classically considered glutamatergic. Immunohistochemical and *in situ* studies indicate VGLUT3 is expressed in GABAergic, serotonergic, dopaminergic, and cholinergic neurons as well as astrocytes. Recent findings suggest that VGLUT1 and VGLUT2 are expressed in astrocytes as well, but these studies are in isolated and cultured cells, not in intact tissue.

Glutamate uptake by the VGLUTs depends primarily on $\Delta\psi$ rather than ΔpH, with apparent affinities in the low millimolar range. Interestingly, the VGLUTs do not appear to recognize aspartate. This is consistent with studies of synaptic vesicles and suggests that aspartate is not readily accumulated in synaptic vesicles through an active transport system. Vesicular glutamate transport shows a biphasic dependence on chloride with an optimum at 2–10 mM, and recent studies with purified VGLUT1 suggest that glutamate transport requires Cl^-. Like VMAT transport, VGLUT transport also appears to be regulated by $G_{\alpha o2}$; specifically, $G_{\alpha o2}$ reduces the chloride dependence of transport. The role of chloride in vesicular glutamate transport is further complicated by the finding that the expression of VGLUT1 appears to increase the vesicular Cl^- conductance.

Although the VGLUTs exhibit similar transport activities, they are expressed in cells with very different properties. For example, compared to VGLUT1-containing neurons, VGLUT2 neurons, in general, have a lower firing rate and a higher probability of release. The three isoforms also appear to have different subcellular localizations that could influence glutamate-release characteristics. Most notably, VGLUT3 is expressed in vesicular structures within the dendrites of some neurons in the hippocampus and striatum, suggesting a role in retrograde signaling. Further, two polyproline domains, which are present in the C-terminal cytoplasmic tail of VGLUT1 (but not VGLUT2 or -3), mediate interactions with the endocytic protein endophilin. This interaction may regulate synaptic vesicle recycling and the mode in which VGLUT1 is internalized. Recent studies indicate that VGLUT1 and VGLUT2 are present in multiple copies on synaptic vesicles. Correcting for the number of vesicles containing the transporters, it is estimated that there are ~ 9 copies of VGLUT1 and ~ 14 copies of VGLUT2 in the synaptic vesicles in which they are expressed. Although not directly demonstrated, a similar stoichiometry probably exists for the other vesicular neurotransmitter transporters.

Several compounds that inhibit vesicular glutamate transport have been identified. These include the dyes Evans blue and rose Bengal. In addition, the stilbene derivative 4,4′-diisothiocyanatostilbene-2,2′-disulfonic acid (DIDS), a compound commonly used as an inhibitor of anion channels, inhibits vesicular glutamate transport. Most known inhibitors have a limited utility because they are membrane impermeant, with the exception of rose Bengal. No inhibitors unique for specific isoforms have been identified.

Neuromodulators Are Also Stored in Vesicles and Released through Exocytosis

Synaptic vesicles mediate the release of neuropeptides and small molecules other than the classic neurotransmitters. Neuropeptides enter the lumen of the secretory pathway in the endoplasmic reticulum through cotranslational translocation and are sorted to the secretory vesicle pathway, where they undergo processing to form the biologically active species. After release, it is believed that neuropeptides are degraded and not repackaged. Of the synaptically released small-molecule neuromodulators, zinc and ATP are the best characterized. NMDA and GABA receptors contain binding sites for zinc, and zinc exerts a direct effect on excitatory and inhibitory neurotransmission. ATP activates both ionotropic and G-protein-coupled receptors. As with the classical neurotransmitters, the exocytotic release of these compounds requires transport into synaptic vesicles.

The multitransmembrane domain protein ZnT3 has been implicated in zinc uptake by synaptic vesicles. ZnT3 belongs to a family of zinc transporters and localizes to synaptic vesicles. Mice deficient in ZnT3 show a loss of zinc staining from hippocampal neurons, and the expression of ZnT3 in PC12 cells increases vesicular zinc staining. Although ZnT3 transport has not been directly demonstrated, these findings strongly support a role for ZnT3 in synaptic vesicle zinc transport. The phenotype of ZnT3-deficient mice is mild, with the most striking abnormality being an increased susceptibility to seizures.

Chromaffin granules, platelet dense-core vesicles, and synaptic vesicles contain concentrations of ATP

many fold higher than cytosolic concentrations, suggesting active vesicular uptake. ATP transport has been demonstrated in chromaffin granules and synaptic vesicles, and the process appears to depend on $\Delta\mu H^+$. It has generally been assumed that ATP is co-stored only with monoamines and acetylcholine, as an anion to balance to cationic charge of those transmitters. However, the extent of ATP storage and release by different neuronal populations remains unknown, and the proteins responsible for ATP uptake by secretory vesicles have not been identified.

One of the first synaptic vesicle membrane proteins to be identified was SV2. This protein is a multitransmembrane domain protein with limited structural similarity to the VMATs and VAChT. The protein was identified in synaptic vesicles from *T. californica*, and three isoforms (A, B, and C) have been identified in mammals. The expression patterns of the vertebrate proteins show partial overlap and together cover essentially all neurons. Targeted disruption of isoform A and both isoforms A and B demonstrates that these proteins are crucial for normal brain function in mice. Although a specific biochemical function was not been defined by these studies, it has been suggested that SV2 might be involved in regulating the size of the readily releaseable pool of vesicles or Ca^{2+} homeostasis at the nerve terminal. SV2 is thought to be the binding site for botulinum A toxin entry into neurons and to be a binding site for the anti-seizure medication levetiracetam. It has also been suggested that the long sugar chains on SV2 serve as a smart-gel that regulates the release of neurotransmitters for the fused vesicle by limiting diffusion (Table 1).

Summary

The vesicular uptake of neurotransmitters is a requirement for the quantal release of neurotransmitters. Earlier biochemical studies that identified four primary activities have been complimented by the molecular identification of unique proteins that mediate vesicular neurotransmitter transport activities. The cloned transporters can be divided into three structural families – the VMATs and VAChT; VGAT; and the VGLUTs. All the transporters are driven by the proton electrochemical gradient generated by the vacuolar H^+-ATPase. Multiple copies of a given transporters are present on a single synaptic vesicle, suggesting that expression levels and regulated trafficking may play roles in modulating synaptic vesicle filling. A greater number of transporters in the vesicle increases the rate of filling and, if there is a significant leak, also leads to an increase in the steady-state concentration of transmitter achieved. Although the potential for the modulation of the vesicular neurotransmitter transporter activities has been established, the extent to which neurotransmitter accumulation and synaptic transmission are regulated by vesicular transporter density, G-protein-mediated modulation, and the co-storage of other small molecules such as zinc and ATP remains to be determined.

See also: Acetylcholine Neurotransmission in CNS; Gamma-Aminobutyric Acid (GABA); Glutamate; Glutamatergic and Gabaergic Systems; Monoamine Transporters: Focus on the Regulation of Serotonin Transporter by Cytokines; Monoamines; Neutrotransmission and Neuromodulation: Acetylcholine; Synaptic Vesicles.

Further Reading

Bellocchio EE, Reimer RJ, Fremeau RT Jr., and Edwards RH (2000) Uptake of glutamate into synaptic vesicles by an inorganic phosphate transporter. *Science* 289(5481): 957–960.

Brunk I, Holtje M, von Jagow B, et al. (2006) Regulation of vesicular monoamine and glutamate transporters by vesicle-associated trimeric G proteins: New jobs for long-known signal transduction molecules. *Handbook of Experimental Pharmacology* 175: 305–325.

Carlson SS, Wagner JA, and Kelly RB (1978) Purification of synaptic vesicles from elasmobranch electric organ and the use of biophysical criteria to demonstrate purity. *Biochemistry* 17(7): 1188–1199.

Fremeau RT Jr., Voglmaier S, Seal RP, et al. (2004) VGLUTs define subsets of excitatory neurons and suggest novel roles for glutamate. *Trends in Neuroscience* 27(2): 98–103.

Johnson RG Jr. (1988) Accumulation of biological amines into chromaffin granules: A model for hormone and neurotransmitter transport. *Physiological Reviews* 68(1): 232–307.

Lein ES, Hawrylycz MJ, Ao N, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445(7124): 168–176.

Liu Y, Peter D, Roghani A, et al. (1992) A cDNA that suppresses MPP+ toxicity encodes a vesicular amine transporter. *Cell* 70(4): 539–551.

McIntire SL, Reimer RJ, Schuske K, et al. (1997) Identification and characterization of the vesicular GABA transporter. *Nature* 389(6653): 870–876.

Nishi T and Forgac M (2002) The vacuolar (H+)-ATPases – nature's most versatile proton pumps. *Nature Reviews Molecular Cell Biology* 3(2): 94–103.

Reimer RJ and Edwards RH (2004) Organic anion transport is the primary function of the SLC17/type I phosphate transporter family. *Pflugers Archiv* 447(5): 629–635.

Stobrawa SM, Breiderhoff T, Takamori S, et al. (2001) Disruption of ClC-3, a chloride channel expressed on synaptic vesicles, leads to a loss of the hippocampus. *Neuron* 29(1): 185–196.

Takamori S, Holt M, Stenius K, et al. (2006) Molecular anatomy of a trafficking organelle. *Cell* 127(4): 831–846.

Voglmaier SM, Kam K, Yang H, et al. (2006) Distinct endocytic pathways control the rate and extent of synaptic vesicle protein recycling. *Neuron* 51(1): 71–84.

Wojcik SM, Katsurabayashi S, Guillemin I, et al. (2006) A shared vesicular carrier allows synaptic corelease of GABA and glycine. *Neuron* 50(4): 575–587.

Vesicular Sorting to Axons and Dendrites

C C Yap and B Winckler, University of Virginia Medical School, Charlottesville, VA, USA

© 2009 Elsevier Ltd. All rights reserved.

Neuronal Polarity – Axons versus Dendrites

What Is Neuronal Polarity?

Neuronal polarity refers to the fact that neurons elaborate two distinct types of processes, namely, axons and dendrites. Axons and dendrites differ from each other in terms of morphology, molecular composition, and function. Functionally, dendrites are largely responsible for receiving signal inputs from other neurons, and axons are responsible for generating and propagating the signal output in the form of trains of action potentials. Action potentials are propagated directionally along the axon to the axonal nerve terminal. At the nerve terminal, synaptic vesicles filled with neurotransmitters await the arrival of action potentials. The resulting calcium influx into the presynaptic terminal causes fusion of synaptic vesicles with the plasma membrane and emptying of neurotransmitters into the synaptic cleft. Neurotransmitters then bind to the appropriate neurotransmitter receptors on the postsynaptic side of the synapse, which results in opening of ion channels and transmission of the electrical impulse from the presynaptic to the postsynaptic cell. Neuronal polarity thus underlies vectorial electrical signaling by vertebrate neurons.

Why Is Neuronal Polarity Important for Neuronal Function?

To ensure the proper conduction of electrical signals, channels and receptors have to be distributed to the correct locations. For instance, synaptic vesicle proteins are concentrated in synaptic vesicles clustered at presynaptic terminals, whereas neurotransmitter receptors are concentrated at postsynaptic sites, precisely apposing the presynaptic terminals. In addition, a very large number of voltage-gated channels show compartmentalized distribution and are not present evenly along the plasma membrane. For example, voltage-gated sodium channels are enriched at axon initial segments and nodes of Ranvier, where they function to initiate and regenerate action potentials, respectively. A large number of potassium channels are not distributed homogeneously in the neuronal plasma membrane either, but rather to axons, or to distal or proximal dendrites, depending on channel type. The exact combination of numbers and location of these channels determines and regulates neuronal excitability. Equalizing the distribution of channels and receptors to uniformity on the membrane would affect the excitability and firing properties of the neuron and lead to dysfunctional and/or nonfunctional circuits. Neuronal function therefore relies on successfully compartmentalizing the neuronal plasma membrane and creating a multitude of molecularly distinct subdomains (Figure 1). In addition to plasma membrane compartmentalization, axons and dendrites also differ in the organization of cytoplasmic components, cytoskeletal elements, and the relative preponderance of membrane-bound organelles.

Compartmentalizing the Plasma Membrane – Multiple Pathways

Most cell surface receptors in neurons are not evenly distributed to dendrites, soma, and axons but rather show enrichment in certain domains, that is, their distribution is polarized. At the most basic, one can distinguish proteins enriched on the axon from proteins enriched on the soma and dendrites, the somatodendritic domain. Furthermore, many proteins are found enriched in smaller subdomains of the somatodendritic or axonal domains, such as synapses, distal or proximal dendrites, soma, axon initial segments, or nodes of Ranvier (Figure 1). Several mechanisms contribute to the nonhomogeneous distribution of receptors and channels in the membrane: (1) Proteins can be specifically targeted to axons or dendrites from the secretory pathway. (2) After reaching the surface, they can diffuse laterally until they bind to appropriately placed cytoskeletal scaffolding proteins, which restricts their diffusion to inappropriate domains. (3) And they can be internalized via endocytosis when they are not in the correct place. These mechanisms can act alone or in combination on any given protein to ensure the correct steady-state distribution. It is an open question whether proteins localized to smaller subdomains, such as synapses or nodes of Ranvier, arrive there directly in a targeted fashion from the secretory pathway or are secreted more generally in a larger area in the vicinity of their final destination. If secretion were not spatially precise enough to deliver proteins to synaptic sites or initial segments or nodes of Ranvier, the resident proteins would become concentrated at the correct site in a multistep process using postinsertion mechanisms such as lateral diffusion, clustering, and/or endocytic removal. Even if secretion were spatially precise, maintenance of the correct localization might still require clustering and tethering to cytoskeleton in order to prevent diffusion away from the correct locale.

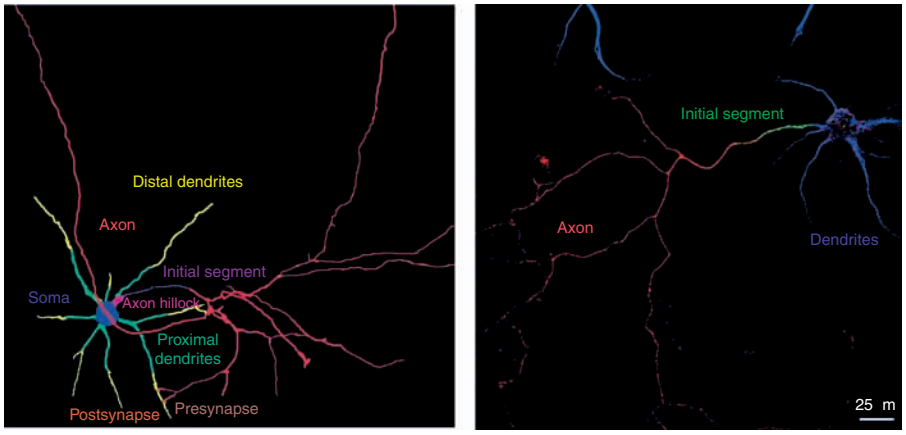

Figure 1 Compartmentalizing the neuronal plasma membrane. Vertebrate neurons contain a large number of distinct plasma membrane domains with distinct composition (diagram on left, actual neuron on right).

The Secretory Pathway in Neurons

Basic Features of Secretion

The secretory pathway has been studied in detail for decades, especially in nonneuronal cells, and its basic features are conserved in all cell types, including neurons. Membrane proteins are cotranslationally inserted into the endoplasmic reticulum (ER), where they fold. They exit the ER at specialized ER exit sites and make their way to the Golgi apparatus via the ER–Golgi intermediate compartment (ERGIC). The Golgi apparatus contains a number of thin membrane sacks which are held together in a stack. The stack closest to the ERGIC is the *cis*-Golgi, followed by the medial and then *trans*-Golgi stacks. Each stack contains different enzymes which modify the traversing membrane proteins. The major modification that takes place in the Golgi is glycosylation. After leaving the *trans*-Golgi, membrane proteins enter the *trans*-Golgi network (TGN), which is made up of a large number of loosely associated tubules and vesicles. The TGN is a major sorting station where proteins ('cargo') are segregated into distinct vesicles that take them to either the plasma membrane or the endosomal system. Signals in the cytoplasmic tail of membrane proteins bind to specific adaptor complexes (such as adaptor protein (AP) 1), which mediate the enrichment of the appropriate cargo molecules and the subsequent assembly of coat structures. The assembling coat deforms the lipid bilayer into a vesicle and allows budding. One of the coats assembled at the TGN is clathrin, but less-well-studied nonclathrin coats are also involved in some budding steps. Clathrin is rapidly disassembled from a newly budded vesicle, which travels (mostly along microtubules) to its destination, the plasma membrane or endosomes. Specific microtubule motor proteins (kinesins) are required for linking and transporting specific transport vesicles along microtubules. All the steps described are highly regulated and carried out by complex molecular machinery (see the section titled 'Molecular targeting machinery'). The final step is fusion of the transport vesicle with the appropriate target membrane, a process requiring a set of matching proteins on the vesicle and target membrane. The matching proteins are called soluble *N*-ethylmaleimide sensitive fusion protein attachment receptors (SNAREs), and many different members of this family regulate fusion steps at distinct sites.

Regulated versus Constitutive Secretion

Classically, two types of secretion have been distinguished: constitutive and regulated. Most membrane addition occurs via constitutive fusion of post-TGN vesicles with the target membrane and is not regulated by extracellular signals. In many specialized cell types, secretion of certain cargoes occurs only in response to a signal, such as extracellular hormones or electrical stimuli. Molecules secreted by regulated secretion frequently reside in a distinct class of vesicles, called dense-core vesicles, based on their distinct appearance by electron microscopy. Synaptic vesicle fusion in nerve terminals occurs in response to membrane depolarization and calcium influx and can be viewed as another example of regulated secretion. The classical neurotransmitters, though, are stored in small, clear synaptic vesicles which do not contain a dense core. It is becoming increasingly clear that some of the secretion pathways considered historically to be constitutive are also subject to regulatory control by extracellular signals.

Endosomal Secretion

In addition to the biosynthetic secretory pathway, membrane proteins can also be secreted from the endosomal system. The endosomal system consists of a multitude of tubular and vacuolar organelles which receive cargo both from the cell surface after endocytosis and from the TGN. Proteins taken up by endocytosis at the plasma membrane first enter early endosomes, from which they are rapidly (within 5–8 min) trafficked to several other endosomal compartments. Depending on the cargo, proteins can either cycle back to the surface or be degraded in lysosomes. Recycling cargo returns to the cell surface either directly from the early endosome or via an intermediate stopover at a distinct tubulovesicular endosome, called the recycling endosome. Recycling to the surface from the recycling endosome is also polarized and results in secretion in distinct sites of the plasma membrane. The endosome and the TGN are therefore the major sorting stations for polarized secretion of membrane proteins. Degradative cargo fails to recycle from the early endosome and enters the degradative pathway to lysosomes via late endosomes. Late endosomes and lysosomes have a very low pH and appear under electron microscopy as multivesicular bodies and/or organelles filled with dense material. Lysosomal hydrolases inside lysosomes then degrade the proteins.

Neuronal Adaptations of Secretion

Neurons are arguably the most complex cell type in the body. Their surface area is many times larger than that of most other cells. They have a very complex architecture with functionally distinct domains positioned very far from the cell body. These unique characteristics create special demands, not only in terms of the sheer capacity for lipid and protein synthesis, but also in terms of the organization and regulation of secretion. Nonneuronal cells typically possess only one Golgi stack, localized near the centrosome on one side of the nucleus. Neurons similarly contain a large somatic Golgi organized in stacks. It is now clear that specialized secretory compartments exist in dendrites, and probably even in dendritic spines, in addition to the somatic Golgi. Are all membrane proteins produced in the somatic Golgi, no matter whether their final destination is the soma, the dendrites, or the far-away axon terminals? The dendritic Golgi compartments (named 'Golgi outposts') are particularly prominent in the large apical dendrites of pyramidal neurons and often localized at branch points. It is not yet known whether only a subset of cargoes traffics via the dendritic Golgi or how transport through the Golgi outposts is regulated. In any case, Golgi outposts would be uniquely situated to polarize a subset of secretory traffic toward sites near the final destination of the cargoes. This 'outsourcing' of a subset of traffic pathways to Golgi outposts may simplify the task of spatial segregation of membrane components. The unique architecture of neurons has led therefore to specializations of the protein production pathways. A second way in which neurons simplify the spatial problem of secretion is by locally translating some messenger RNAs in dendrites and also in axons. A subset of messenger RNAs is not translated in the soma but rather transported via specialized machinery to the dendrites, where its translation is locally regulated.

Polarized Membrane Traffic in Epithelial Cells – a Useful Analogy for Neurons?

Many other differentiated cells elaborate distinct plasma membrane domains. The best-studied examples are epithelial cells, which have distinct apical and basolateral plasma membrane domains. Many membrane proteins, such as influenza hemagglutinin, are targeted to the apical domain, whereas vesicular stomatitis virus G-protein is targeted to the basolateral domain. Much work, in particular on the Madin-Darby canine kidney (MDCK) cell line, has led to the following model of polarized membrane traffic: apical and basolateral proteins are segregated away from each other into distinct carrier vesicles at the level of the TGN. The efficient sorting of membrane proteins in the TGN requires specific signals in the cytoplasmic tail of basolateral receptors. Several classes of basolateral targeting signals have been identified. One is a tyrosine-based motif characterized by an essential tyrosine in the context of YXXϕ (where X represents any amino acid and ϕ is a bulky hydrophobic residue). The other is a dileucine-based motif characterized by two consecutive leucine residues. Both signals have been implicated as internalization signals at the plasma membrane as well. Several of these signals mediate direct binding to adaptor complexes (such as AP1B), which ensure enrichment of cargo to small membrane regions, recruitment of clathrin coats, and formation of budding vesicles. Additional basolateral signals not related to tyrosine- or dileucine-based motifs have also been identified. The newly budded vesicles then transport to the appropriate membrane domains, where their fusion is controlled by other molecular complexes, such as the exocyst complex and SNAREs.

Similarly, apical targeting signals have been identified. Unlike basolateral signals, apical signals frequently map to extracellular/lumenal domains or to

residues near transmembrane domains of apical proteins. Several apical signals residing in the cytoplasmic tail of apical receptors have also been described. The machinery for apical targeting is relatively less well understood. How a lumenal targeting signal in the interior of the TGN can be recognized by apical targeting machinery is not well understood, but proteins such as MAL have been shown to affect apical targeting efficiency. Lipid composition also plays a crucial role in apical targeting of some proteins. Lipids with saturated fatty acid chains tend to associate tightly with cholesterol and glycosphingolipids and segregate into tiny islands of special lipid composition, called lipid rafts. Several apical proteins show higher affinity for lipid raft domains than for nonraft domains and cluster there. Preferential affinity for lipid rafts could therefore provide a mechanism for enriching and clustering of apically destined membrane proteins.

Mapping of Dendritic and Axonal Targeting Signals

On the basis of the observation that influenza hemagglutinin, which is apical in MDCK cells, targets to axons in neurons, while vesicular stomatitis virus G-protein, which is basolateral in MDCK cells, targets to the somatodendritic domain of neurons, it has been proposed that the mechanisms underlying polarized membrane trafficking in the two cell types are conserved. Experimental tests of this idea have revealed remarkable conservation of some processes, but also unique mechanisms in neurons.

Dendritic Signals

Both tyrosine-based and dileucine motifs resembling basolateral signals have been identified as targeting signals in several dendritic proteins. Tyrosine-based motifs are involved in the dendritic targeting of low-density lipoprotein receptor, transferrin receptor, and Delta/Notch-like epidermal growth factor-related receptor. Shal-family K channels and glycine transporter 1b use dileucine motifs for dendritic targeting. A dileucine motif also mediates dendritic targeting of serotonin receptors 5-HT1A/1B and of neuronal nicotinic acetylcoline receptor α7. Nevertheless, not all dileucine motifs result in somatodendritic enrichment. It is interesting that some motifs associated with dendritic targeting are not related to basolateral targeting motifs, such as the motifs in the cytoplasmic tails of neuroligin-1, glutamate receptor (GluR) 1, and metabotropic glutamate receptor (mGluR) 1a. A novel phenylalanine-based motif has recently been identified to direct telencephalin to neuronal dendrites. Clearly, the full catalog of dendritic signals is still to be determined, together with the machinery and regulation for each of the distinct classes. Whether the kind of dendritic signal present in a protein regulates any aspects of spatial fusion or retrieval is unknown.

Axonal Signals

As described for basolateral dendritic signals above, axonal signals can in some instances share characteristics with apical targeting signals. For instance, the extracellular/lumenal domains of some axonal proteins (i.e., the cell adhesion molecule (CAM) L1/neuron–glia (Ng) CAM) and β-APP) are sufficient for axonal targeting. In the case of NgCAM, the extracellular domain is also active in apical targeting in MDCK cells. As for apical signals, the molecular mechanisms underlying recognition and execution of lumenal axonal signals are not known. Just as in apical targeting, association with lipid rafts may provide a necessary step for axonal targeting of some proteins (especially glycophosphatidylinositol-linked proteins, such as Thy1 and prion protein), but not of others (e.g., NgCAM). Sorting into the axonal compartment can also be a result of posttranslational modifications, such as palmitoylation. Palmitoylation is essential for effective targeting of growth-associated protein-43 and glutamic acid decarboxylase-65 to the presynaptic clusters in axon and for exclusion from dendrites. The palmitolyated domains reside in the N-terminal region of the proteins.

An increasing number of axonal targeting signals have recently been mapped to cytoplasmic sequences, even though not all targeting signals could be mapped precisely. In particular, the cytoplasmic tails of mGluR7, 5-HT1B, γ-aminobutyric acid (GABA) transporter, and GABA transporter-3 have been reported to mediate axonal targeting of these proteins in neurons. Sequences responsible for targeting of KCNQ2 to the axon reside in the membrane proximal region and conserved A domains of the KCNQ C-terminal tail; the exact motifs, however, are still unknown. Recently, a novel signal, RRK tripeptide, resembling the RXR ER retention motif, found in the C-terminal region of mGluR1b, has been shown to be sufficient and necessary for axonal targeting of the protein in retinal ganglion neurons but not in hippocampal neurons.

Clearly, axonal targeting signals are diverse, and the mechanisms of their execution are poorly understood. For most of the signals mapped, it is unknown whether the signal promotes lateral enrichment into axonally destined vesicles in the TGN, association with axonal motors, competence for fusion with axonal secretion sites, anchoring and restriction of diffusion in axons, or retrieval by endocytosis from inappropriate sites. There are a number of examples

of axonal targeting, though, from which some mechanistic insights have been gleamed. For instance, targeting of the potassium channel Kv1.2 and Kv1.3 to the axonal compartment requires the tetramerization domains (T1) which lie in the intracellular N-terminal regions. The T1 domain is also required for interaction with the Kvβ subunit and this interaction of T1 might also contribute to axonal targeting of these channels. A second mechanism relates to the involvement of endocytosis and is discussed next.

A Role for Endosomes in Axonal Targeting Pathways

Two pathways have been identified which result in axonal accumulation in an endocytosis-dependent manner, selective endocytosis/retention and transcytosis. For several axonal proteins, the axonal targeting motifs mapped to endocytosis signals. These observations gave rise to the idea that axonal enrichment might not arise solely, or even primarily, from polarized secretion from the TGN to axonal domains but rather from the preferential endocytosis of 'misplaced' somatodendritic receptor pools coupled to specific retention/anchoring in the correct axonal domain. A selective endocytosis/retention pathway may account for the axonal enrichment of voltage-gated sodium channels, vesicle-associated membrane protein 2, and nicotinic acetylcholine receptor α4. A second endocytosis-dependent pathway to the axon is via transcytosis, that is, initial delivery to the somatodendritic domain, followed by endocytosis and transport to the axonal plasma membrane from somatodendritic endosomes. This pathway operates for L1/NgCAM and for the polymeric immunoglobulin receptor.

Molecular Targeting Machinery

A membrane protein has to navigate successfully a multistep pathway to reach the final location on either dendrites or axons. These steps include lateral enrichment into axonally or somatodendritically destined vesicles in the TGN, association with axonal or dendritic motors, competence for fusion with axonal or dendritic secretion sites, anchoring and restriction of diffusion in axons or dendrites, or retrieval by endocytosis from inappropriate sites. Each of these steps is subject to regulation and can, at least in principle, be carried out in a selective and discriminating manner for somatodendritic and axonal proteins. Regulatory molecules for each of these steps are therefore involved in the polarized distribution of membrane molecules. Several classes of regulators are well characterized in nonneuronal membrane traffic, and their exact roles in neurons are now under investigation.

Adaptor Complexes

Adaptor complexes (AP) (four distinct ones in mammals: AP1–4) are involved in protein sorting from the TGN to endosomes or to and from the plasma membrane. Adaptors are tetrameric complexes comprised of two large subunits, a medium subunit μ and a small subunit σ. Adaptors can bind to known sorting motifs in the cytoplasmic tail of receptors and promote the receptors' inclusion into forming vesicles. As such, adaptor complexes are considered part of the sorting machinery for polarized trafficking in epithelial cells. Adaptors likely play similar roles in neuronal trafficking. The most studied clathrin-associated adaptor, AP-2, binds endocytosis motifs and mediates synaptic vesicle formation and receptor trafficking from the plasma membrane via endocytosis. In the nematode, *C. elegans*, the μ subunit of AP1 interacts with odorant receptor ODR-10 and mediates sorting to dendrites in olfactory neurons *in vivo*. Similarly, the μ subunit of AP-4 binds to the C-terminal of ionotropic glutamate receptor delta 2 in brain and may thereby mediate targeting of the receptor to the dendrites of Purkinje cells. AP3 is known to be involved in lysosomal cargo sorting. In neuron, AP-3B is thought to serve neuron-specific functions, such as neurotransmitter release in GABAergic neurons. Beside the classic AP complexes, sorting of certain membrane proteins requires interaction with specific interacting proteins. Four small transmembrane α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) regulatory proteins (TARPs) constitute a family of stargazin-related proteins. Maturation and translocation of AMPA receptors from an intracellular site to the cell surface requires TARPs. This indicates that TARPs act as an adaptor in trafficking of AMPA receptors early in the biosynthetic pathway.

Exocyst Complex

The exocyst is a multiprotein complex originally found in yeast and required for polarized exocytosis. Its proposed role is to mark the site of vesicle fusion and regulate the fusion event. The exocyst plays important roles in membrane trafficking events, such as synaptogenesis and neurite outgrowth, required for neuronal development. The Sec8 subunit of the exocyst interacts with PDZ domain-containing synapse-associated protein (SAP) 102 and participates in the dendritic trafficking of newly synthesized *N*-methyl-D-aspartate receptors from the ER. In addition, Sec8 also mediates intracellular AMPA receptor targeting within the dendritic spine, possibly through PDZ-dependent interactions with synaptic scaffolding molecules. In *Drosophila*, the sec15 subunit of the exocyst

is required for targeting of specific cell adhesion and signaling molecules to the correct exocytic site, but not for synaptic vesicle release.

Rab Proteins

Rab proteins are small guanosine triphosphatases which regulate protein transport along the endocytic and exocytic pathways in all cell types. Rabs participate in vesicle budding, membrane fusion, and interactions with the cytoskeleton. Rab proteins are important for neuronal function because alterations in Rab protein regulation can lead to mental retardation in humans. In neurons, rab11 has emerged as a mediator of vesicle transport by coupling to myosin Vb to mediate GluR1 trafficking. Similarly, Rab8, which has an exclusive somatodendritic distribution, is involved in transport of proteins to the dendritic surface in neurons. In addition, Rab8 associates with exocyst in the recycling endosome and participates in targeting and delivery of AMPA receptors to the postsynaptic membrane.

Points of Regulation

It is now clear that polarized distribution to axons and dendrites can take place along multiple pathways, such as direct vesicular targeting from the TGN, indirect targeting via transcytosis, selective endocytosis/retention, and potentially others. These pathways are used differentially by different proteins, based on the signals present. It is also clear that the machinery regulating these various steps can be regulated, which opens up the possibility that, even if the signals are not changing, the pathway and/or final location can change. For instance, the same protein can be axonal in one neuronal cell type but somatodendritic in another. The same protein can also be localized differently at different developmental stages in the same neuron. In addition, pathway choice can be regulated by activity or by ligands. A striking example of the latter is the targeting of the type 1 cannabinoid receptor, which is modulated by constitutive activation.

See also: Axonal and Dendritic Transport by Dyneins and Kinesins in Neurons; Axonal mRNA Transport and Functions; Axonal and Dendritic Identity and Structure: Control of; Calcium Channels and SNARE Proteins; Dendritic Signal Integration; Microtubule Associated Proteins in Neurons; Neurosecretion (Regulated Exocytosis in Neuroendocrine Cells); SNAREs.

Further Reading

Alberts P and Galli T (2003) The cell outgrowth secretory endosome (COSE): A specialized compartment involved in neuronal morphogenesis. *Biology of the Cell* 95: 419–424.

Colman DR (1999) Neuronal polarity and the epithelial metaphor. *Neuron* 23: 649–651.

Craig AM and Banker G (1994) Neuronal polarity. *Annual Review of Neuroscience* 17: 267–310.

Dotti CG and Simons K (1990) Polarized sorting of viral glycoproteins to the axon and dendrites of hippocampal neurons in culture. *Cell* 62: 63–72.

Foletti DL, Prekeris R, and Scheller RH (1999) Generation and maintenance of neuronal polarity: Mechanisms of transport and targeting. *Neuron* 23: 641–644.

Garrido JJ, Fernandes F, Moussif A, Fache MP, Giraud P, and Dargent B (2003) Dynamic compartmentalization of the voltage-gated sodium channels in axons. *Biology of the Cell* 95: 437–445.

Gerges NZ, Backos DS, Rupasinghe CN, Spaller MR, and Esteban JA (2006) Dual role of the exocyst in AMPA receptor targeting and insertion into the postsynaptic membrane. *EMBO Journal* 25: 1623–1634.

Horton AC and Ehlers MD (2003) Neuronal polarity and trafficking. *Neuron* 40: 277–295.

Lai HC and Jan LY (2006) The distribution and targeting of neuronal voltage-gated ion channels. *Nature Reviews Neuroscience* 7: 548–562.

Leterrier C, Laine J, Darmon M, Boudin H, Rossier J, and Lenkei Z (2006) Constitutive activation drives compartment-selective endocytosis and axonal targeting of type 1 cannabinoid receptors. *Journal of Neuroscience* 26: 3141–3153.

Nicoll RA, Tomita S, and Bredt DS (2006) Auxiliary subunits assist AMPA-type glutamate receptors. *Science* 311: 1253–1256.

Pierce JP, van Leyen K, and McCarthy JB (2000) Translocation machinery for synthesis of integral membrane and secretory proteins in dendritic spines. *Nature Neuroscience* 3: 311–313.

Winckler B (2004) Pathways for axonal targeting of membrane proteins. *Biology of the Cell* 96: 669–674.

Winckler B and Mellman I (1999) Neuronal polarity: Controlling the sorting and diffusion of membrane components. *Neuron* 23: 637–640.

Yap CC, Murate M, Kishigami S, et al. (2003) Adaptor protein complex-4 (AP-4) is expressed in the central nervous system neurons and interacts with glutamate receptor delta2. *Molecular and Cellular Neurosciences* 24: 283–295.

Vestibular Influences on Cognition

F W Mast, University of Lausanne, Lausanne, Switzerland

© 2009 Elsevier Ltd. All rights reserved.

Vestibular Areas in the Cortex

The vestibular system, long described as a sensory system, elicits reflexive eye movements such as the vestibuloocular reflex (VOR). The VOR helps us to keep our gaze fixed on a target while we move our head. Research interests have focused on the neuronal circuits underlying the VOR and computational models have been developed to describe the underlying neuronal network. The VOR has become a suitable paradigm to study how the brain implements a sensorimotor reflex; however, vestibular afferent pathways not only project to brain stem areas, which then elicit eye movements, but also reach other areas in the brain. These include the hippocampus (either directly or via thalamic nuclei) as well as cortical areas identified in primate studies to receive vestibular input (e.g., the area 2v at the tip of the intraparietal sulcus, area 3av in the central sulcus, area 7 in the inferior parietal lobule, and the parietoinsular cortex). More recently, measures of human brain activation by means of neuroimaging techniques (e.g., functional magnetic resonance imaging or positron emission tomography) during caloric stimulation of the semicircular canals or galvanic stimulation of the vestibular system have widely confirmed the existence of these vestibular-related areas. It is noteworthy that some cortical vestibular areas also project back to the vestibular nuclei.

The cortical vestibular areas are interesting because knowledge about them is a necessary prerequisite to explore the influence vestibular stimulation can have on cognitive functions. The relation between vestibular information and cognitive functions is twofold. First, we can study how vestibular information influences cognitive functions, and second, we can study how cognitive influences modify the perception of vestibular stimuli. Regarding the first point, it is particularly interesting to explore which cognitive tasks are affected by vestibular processing, and in what way. The second point studies how influences of cognitive origin can alter the perception of vestibular stimuli. These types of influences are often referred to as 'top-down' information, which can affect relatively early processes. Top-down processing has been studied intensively in many areas of visual perception, but there still is only scarce evidence for its role in the context of vestibular stimulation.

As already mentioned, neurophysiological evidence suggests that several cortical areas receive vestibular input (relevant for the first point), and there also are corticofugal projections from higher cortical areas to lower areas in the brain stem (relevant for the second point).

Influence of Body Tilt on Cognitive Tasks

It needs to be determined which cognitive functions are influenced by vestibular information, and which are not. It is at least conceivable that concomitant vestibular stimulation acts just like an additional task, which absorbs limited cognitive resources, which are then no longer available for the primary task. In this case, vestibular information would act in a rather unspecific way, and any type of cognitive function would be affected in nearly the same way. However, there is evidence that body tilt information from the otoliths has specific effects on cognitive functions.

Body Tilt Influences on Visual Perception Tasks

We first consider the example of orientation acuity, which illustrates how visual perception is influenced by body tilt. When upright persons have to judge the orientation of test stimuli (e.g., two subsequent sine wave gratings in a two-alternative forced-choice paradigm), the acuity estimates show clear meridional anisotropies; orientation acuity is higher for the principal upright and horizontal axes when compared to oblique axes. It was often tacitly assumed that orientation acuity is determined by the orientation of the test stimuli with respect to retinal coordinates. When the head was tilted sideways, however, the meridional variation interacted with head orientation. This finding contradicts an explanation based solely on retinotopically tuned neurons, which cannot explain the effect of head tilt. Similar influences of head tilt were found for other types of visual tasks ranging from visual illusions to visual search.

With perfectly upright observers, no effect of body tilt can possibly unfold, and only when tilted observers are tested can the retinal and gravitational reference frames be decoupled. Interestingly, single-cell recordings from monkey and cat studies further support the influence of body tilt information on early visual cortex. Even though the empirical evidence is still scarce, the findings demonstrate compellingly that orientation selectivity of neurons in early visual cortex is not defined exclusively in retinal coordinates, but rather follows the gravitational upright orientation.

These neurons are in the service of orientation constancy, which is the ability to perceive our environment as stable and upright despite the fact that head tilts inevitably influence how the world outside projects onto the retina. To compensate the potentially disturbing influence of head tilt, the brain uses the information from the otoliths, which signal the amount of head tilt. The otoliths are the primary sensory source of information about head tilt, even though a possible contribution of another gravity-receptive sensory system located in the trunk cannot be ruled out completely. In any case, retinal image shifts due to head tilts are compensated for, thus ensuring orientation constancy. How internal coordinate transformations of retinal coordinates can influence perception is strikingly illustrated by transient room-tilt illusions following central or peripheral vestibular disorders. Upright patients perceive the orientation of their visual environment as upside down or tilted by 90°. Room-tilt illusions are thus an example of what happens when the otherwise functioning visual–vestibular interaction is impaired.

These findings illustrate how performance in simple visual psychophysical tasks can change when the participants are no longer in a perfectly upright body position. In the following sections we shift the focus from the encoding of visual stimuli to information already stored in memory, and describe how higher cognitive functions can be modified by vestibular information.

Body Tilt Influences on Mental Imagery Tasks

What is the spatial frame of reference used by information already stored in memory? Do changes in body tilt influence how we perform mental operations? One of the most studied functions in cognitive psychology is mental imagery. It draws on the ability to generate, inspect, and transform images in our mind. An example of the latter is mental rotation, which is the ability to rotate images in one's mind. It is used when we see objects in an unusual orientation – for example, upside down. To still recognize the objects we need to match the image from the sensory input with a representation stored previously in memory. The more the objects are tilted away from their natural and upright orientation, the longer it takes for observers to respond (e.g., to name the object). It needs to be defined, however, what 'upright' means. Does upright refer to retinal coordinates (which is an egocentric frame of reference) or to a gravitational frame of reference (which is an allocentric frame of reference and therefore independent of the observer)? Again, in the upright body position it is impossible to decouple the two reference frames because they are fully aligned. Only when the observers are tilted can we draw conclusions about their relative influence.

Studies on mental rotation Empirical results have revealed an influence of body position – for example, with letterlike symbols as stimuli in a mental rotation task. Specifically, an observer's' performance was better when the stimuli were aligned with the gravitational vertical than with the retinal vertical. This implies that at least some representations are stored with respect to the gravitational vertical. More recently, the frame of reference has been studied in the context of face perception. Earlier research on face perception has shown that inversion of facial features such as mouth and eyes within an otherwise right-side-up face looks extremely grotesque (the face of a former British prime minister has been used in a well-known demonstration of this), but this grotesqueness disappears almost completely when the face is viewed upside-down (i.e., when picture is rotated by 180°). This phenomenon has been explained by a mental rotation process, which is simply overtaxed by keeping track of all configural information between facial parts when upside-down faces are rotated mentally. In the current context, it is of interest whether the perception of faces is determined by their orientation with respect to eye coordinates. If this were the only factor, changes in body position would not matter. After testing face perception in a variety of different body positions, it turned out that in a body position of 135° roll tilt (the head is between horizontal and upside down), it was more difficult to detect changes in the faces, whereas no such effect was present in the upside-down body position. This suggests that a gravitational frame of reference is involved in the perception of faces, even though the nature of the underlying mechanism is not resolved.

The subjective visual vertical The effect of head-down body positions has been investigated in other studies. For example, the subjective visual vertical shows largest errors in body tilt positions around 135°. Interestingly, not only the deviation from the physical vertical but also the variance reaches its maximum in positions around 120° and 150° roll tilt (whereas in the upside-down body position, 180°, there is much less variance and the errors are smaller). Therefore, the spatial reference information for the perception of the vertical is less reliable, and as a consequence, observers have difficulties in these positions in unambiguously judging the orientation of visual stimuli. In head-down positions around 135°, the deviation between the retinal up and the perceived gravitational is largest and exceeds more than 90° (this is also true for the upside-down body position but there the reference frames are again aligned, albeit in exactly opposite directions). It is noteworthy that observers sometimes report a

paradoxical experience when they are in a head-down body position. Some observers tend to perceive two different but distinct verticals, rather than just an increased range of uncertainty.

It has been assumed that less precise sensory information from the otoliths is related to the lack of confidence about the direction of gravity in exactly these unusual body positions. In analogy to the blind spot on the retina, it has been hypothesized that there also is a body tilt stimulus for which there are many less utricular or saccular hair cells with the corresponding direction specificity. It is, however, not yet certain whether the psychophysical results reported in head-down positions can be explained entirely by the properties of receptor physiology. Nevertheless, situations which lead to ambiguous or insufficient perception make the use of cognitive influences more likely, and they may eventually help in resolving what the actual stimulus was.

Currently, there is compelling empirical evidence for the role of body tilt information on perceptual and cognitive tasks, but no theory has yet been proposed describing how these influences can unfold, and what the underlying neuronal mechanisms are. More experimentation is still needed, since many studies varied two body positions only (e.g., upright and lying on the side in the horizontal position), and this is not sufficient to reveal the function the body tilt influence follows. Also, the stimulus dimensions need to be explored more thoroughly – for example, the effect of gravity on the perception of natural scenes may even exceed the effect on faces, because – unlike faces – natural scenes always appear gravitationally upright.

Studies in microgravity Instead of tilting the observers, exposure to microgravity is yet another way to change vestibular information – in particular, the information from the otoliths about the direction of gravity. Astronauts, during long-term missions, experience microgravity onboard the spaceship during several days. A short-term exposure to microgravity for about 20 s can also be reached on Earth by means of parabolic flight protocols onboard an airplane. During microgravity, the astronauts can move freely inside the cabin and receive no sensory input that directly indicates the direction of the vertical. When the pull of gravity is absent, the resting discharge level becomes the only afferent signal from the otoliths. It has been suggested that perceptual consequences can unfold from asymmetries between the otolith discharge levels from the right and left inner ear, and this still is discussed in the context of space motion sickness, for which no clear predictive measure has yet been found. On Earth, however, the presence of gravity provides a strong enough stimulus to override these possible asymmetries.

Interesting in the context of microgravity are orientation illusions. Astronauts are able to voluntarily assign a spatial frame of reference to their environment. They can choose which wall inside the spaceship becomes the floor. On Earth, this is determined by the presence of gravity, and only in the horizontal plane do we sometimes get confused regarding our orientation within the environment (e.g., when ascending from a subway we often start walking in the wrong direction, not taking into account a turn we just made underground).

Still, much research is needed to better understand the effects of microgravity on cognitive functions, both in terms of short-term and long-term exposure. The latter is important because it also includes the effects of a possible adaptation to the new environment. To date, only few studies on cognitive functions have been carried out in microgravity, and some studies report no change when compared to performance on the ground. This finding is not so surprising because, unlike body tilts on Earth, microgravity does not induce any changes in the direction of gravity. Vestibular and visual cues are not in conflict and the observers rely on visual information, such as upright persons on the ground. It has to be noted that only few cognitive studies have been carried out in microgravity, and it would be premature to rule out a yet unexplored effect on other cognitive functions.

Influence of Whole Body Movement on Cognitive Tasks

The preceding findings were obtained under static conditions. The vestibular influence is thus of otolith origin, since any semicircular effects of prior head or whole-body movements dissipated by the time the measurements began. Measurements during whole-body rotation are much more difficult to obtain. For example, the VOR can interfere with the visual encoding of stimuli used for a cognitive study, and this may confound the interpretation of the results. Yet another difficulty is to define an appropriate control condition, in which the body remains stationary. The absence of any vestibular stimulation is easy to achieve but there are several other factors, which are just not controlled for when comparing the results from the same task during the experimental condition (with vestibular stimulation) and the control condition (body and head are still). Inevitably, body movement entails stimulation of other sensory systems, such as skin receptors during acceleration or deceleration of the body (e.g., when we feel pushed backward into the seat during an airplane's takeoff). Not only is it

hard to isolate the vestibular system, but the body movement can be experienced just as generally more distracting or even discomforting and therefore can impair performance in the task. Yet another difficulty is of technical nature. Vestibular stimulation requires acceleration or deceleration as stimulus, and not movement at constant velocity, which is not based on any vestibular stimulation. For example, constant yaw rotation around the body longitudinal axis is no longer perceived when the eyes are closed; the person in this situation feels stationary as long as the head is not moved, which then likely results in cross-coupled acceleration stimulation. To sum up, it is not at all trivial to conduct research with clearly defined vestibular stimuli, and, in addition to this, it requires carefully designed tasks if one wishes to measure at the same time performance in a cognitive task.

Visual–Vestibular Interaction

Caloric and galvanic vestibular stimulations have become widely used techniques in vestibular research, and when combined with neuroimaging they have enabled researchers to explore the human brain areas during vestibular stimulation. They have the advantage that one can isolate the vestibular stimulation, and that they can be used while the body is stationary. They provide ways to produce the percept of body movement even though the body remains perfectly still.

Converging evidence based on several studies using different techniques points to a mutual inhibition between vestibular and visual stimulation. During vestibular stimulation the visual cortex is deactivated, and, likewise, during large-field visual stimulation the vestibular cortical areas are deactivated. Such a mutual inhibition serves the purpose of minimizing sensory conflicts. The deactivation of the visual cortex can help suppress inappropriate visual input during vestibular nystagmus.

Why is this mutually inhibitory interaction between visual and vestibular information relevant in the context of cognitive tasks? Independent research in cognitive neuroscience has provided a wealth of knowledge about the role of early visual cortex in visual mental imagery. When study participants have vividly imagined previously memorized images, several neuroimaging studies found early visual activation, in particular when the task required to picture details in images. During caloric stimulation this part of the brain is deactivated and hence it could be expected that imagery abilities are impaired. This was in fact the case, whereas a control task with knowledge-based questions (designed so that they have low imagery content) showed no change during caloric stimulation. This suggests that the influence caloric stimulation has on cognitive tasks is specific and does not generally distract attention.

The role of attentional demands on vestibular processing has also been studied, and the results show that monitoring of orientation is impaired when the level of disorientation is high enough. It has also been shown that cognitive tasks such as backward counting can impair accuracy in orientation tasks. This suggests that monitoring orientation using vestibular information requires mental effort and is therefore not entirely automatic.

Still, however, there is only scarce evidence about how vestibular information relates to cognitive tasks. To date, there is not yet a systematic approach to study cortical vestibular processes and to what extent these brain areas are shared with cognitive functions.

See also: Balance and Posture Control; Canal–Otolith Interactions; Gravitational Effects on Brain and Behavior; Motion Sickness; Nystagmus; Vestibular System; Vestibulo-Autonomic Responses; Visual–Vestibular Interactions.

Further Reading

Berthoz A (2000) *The Brain's Sense of Movement*. Cambridge, MA: Harvard University Press.

Buchanan-Smith HM and Heeley DW (1993) Anisotropic axes in orientation perception are not retinotopically mapped. *Perception* 22: 1389–1402.

Corballis MC, Nagourney BA, Shetzer LI, et al. (1978) Mental rotation under head tilt: Factors influencing the location of the subjective reference frame. *Perception & Psychophysics* 24: 263–273.

Dieterich M (2007) Functional brain imaging: A window into the visuo-vestibular systems. *Current Opinion in Neurobiology* 20: 12–18.

Gaunet F and Berthoz A (2000) Mental rotation for spatial environment recognition. *Cognitive Brain Research* 9: 91–102.

Kaptein RG and Van Gisbergen JA (2004) Interpretation of a discontinuity in the sense of verticality at large body tilt. *Journal of Neurophysiology* 91: 2205–2214.

Lobmaier JS and Mast FW (2007) The Thatcher illusion: Rotating the viewer instead of the picture. *Perception* 36: 537–546.

Mast F, Kosslyn SM, and Berthoz A (1999) Visual mental imagery interferes with allocentric orientation judgements. *Neuro Report* 10: 3549–3553.

Mast FW, Berthoz A, and Kosslyn SM (2001) Mental imagery of visual motion modifies the perception of roll vection stimulation. *Perception* 30: 945–957.

Mast FW, Ganis G, Christie S, et al. (2003) Four types of visual mental imagery processing in upright and tilted observers. *Cognitive Brain Research* 17: 238–247.

Mast FW, Merfeld DM, and Kosslyn SM (2006) Visual mental imagery during caloric vestibular stimulation. *Neuropsychologia* 44: 101–109.

Merfeld DM, Zupan L, and Peterka RJ (1999) Humans use internal models to estimate gravity and linear acceleration. *Nature* 398: 615–618.

Oman CM, Lichtenberg BK, Money KE, et al. (1986) M.I.T./Canadian vestibular experiments on the Spacelab-1 mission: 4. Space motion sickness: Symptoms, stimuli, and predictability. *Experimental Brain Research* 64: 316–334.

Sauvan XM and Peterhans E (1999) Orientation constancy in neurons of monkey visual cortex. *Visual Cognition* 6: 43–54.

Vestibular System

J C Glover, University of Oslo, Oslo, Norway

© 2009 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition © 2004, Elsevier B.V.

Introduction

The vestibular system detects motion and gravity and initiates movements to maintain balance and orientation. It consists of a set of sensory organs in the inner ear, sensory afferents that link the sensory organs to the brain stem, a set of vestibular nuclei within the brain stem, and the projections of these nuclei to interneurons and motor neurons in the brain stem and spinal cord. Although the vestibular reflexes mediated by these projections are highly stereotyped, the vestibular system interacts extensively with the visual and proprioceptive systems, and its activity is influenced by the cerebellum and cortex to provide a high degree of adaptability.

Vestibular Sensory Organs and Afferents

The sensory apparatus of the vestibular system comprises much of the labyrinth of the inner ear (which also includes the acoustic sensory apparatus, the cochlea). The two mirror-symmetric labyrinths are complex bony structures lined with an epithelium and containing a fluid called endolymph. Separate regions of the epithelium are specialized by the presence of mechanoreceptive hair cells. Each labyrinth contains five vestibular sensory organs, whose mechanoreceptive portions are located within or near the vestibule of the labyrinth, hence the name 'vestibular' system (Figure 1(a)). The vestibular sensory organs can be classified into two types on the basis of structure and function: the otolith organs and the semicircular canals.

Otolith Organs

The otolith organs, the utricle and saccule (Figure 1(b)), detect linear acceleration. In humans, their mechanoreceptive epithelia (the maculae) bear respectively about 30 000 and 16 000 hair cells. The hair bundles of the hair cells are attached to a gelatinous sheet called the otolithic membrane, in which are embedded numerous calcite particles, or otoconia (literally, 'ear dust'). During an acceleration in the plane of the macula, the mass of the otoconia deflects the otolithic membrane and thus the hair bundles. The hair cells are oriented in various directions in a systematic pattern across the macular surfaces, thereby allowing responses to linear acceleration in any direction within the plane of the macula. The saccules are oriented vertically and parasagittally, whereas the utricles are oriented horizontally. Thus, in a person standing upright and motionless the saccules are selectively stimulated by the acceleration caused by gravity, whereas in a person lying motionless on his or her side the utricles are selectively stimulated by the acceleration caused by gravity. Movements cause additional linear accelerations and are detected by the appropriately oriented hair cells in one or both pairs of otolith organs.

Semicircular Canals

The three semicircular canals are endolymph-filled rings specialized to detect angular acceleration. Each canal has a swelling, called the ampulla (Figure 1(c)), which contains the mechanoreceptive epithelium (the ampullary crista). In humans, each crista bears about 7000 hair cells. The hair bundles of the hair cells are inserted into a gelatinous mass (the cupula), which spans the diameter of the ampulla. During an angular acceleration, the endolymph within the canal, by virtue of its inertia, lags behind the canal itself, just as water within a rotated glass lags behind the wall of the glass. This deflects the cupula and the hair bundles. Unlike the situation in the otolith organs, the hair cells of each semicircular canal have the same orientation and, therefore, respond only to angular acceleration within the plane of the canal. The three canals are oriented perpendicular to each other, however, such that angular acceleration in any given plane is coded as a vector sum of the responses of the individual canals. The horizontal canals lie in a nearly horizontal plane (actually horizontal when the head is titled down about 25°). The anterior and posterior canals are oriented at angles to the sagittal plane, such that the posterior canal on one side lies in the same plane as the anterior canal on the other side (Figure 2).

Vestibular Afferents and Efferents

The vestibular afferents, which number about 20 000 on each side in humans, derive from sensory neurons within the vestibular ganglion (Scarpa's ganglion), nestled in a small cavity in the skull just medial to the labyrinth. These neurons extend a distal axon that contacts and receives synapses from the hair cells, and a proximal, or central, axon that extends into the

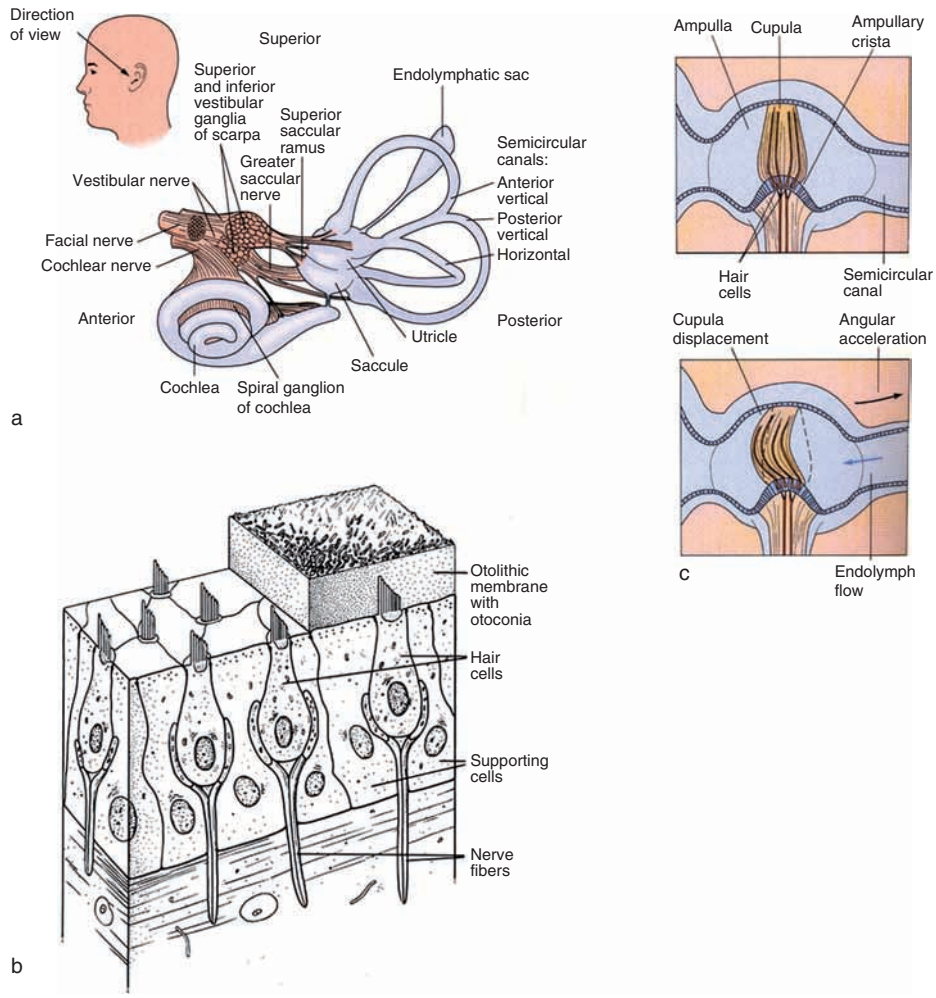

Figure 1 Anatomy of the labyrinth. (a) The labyrinth (purple) with associated vestibular and cochlear ganglia and nerves. (b) Structure of the macula in an otolith organ. (c) Structure of the ampulla in a semicircular canal and deflection of the cupula by an angular acceleration. Adapted from Kandel ER, Schwartz JH, and Jessell TM (1991/2000) *Principles of Neural Science*, 3rd and 4th edns. New York: Elsevier, with permission from Elsevier.

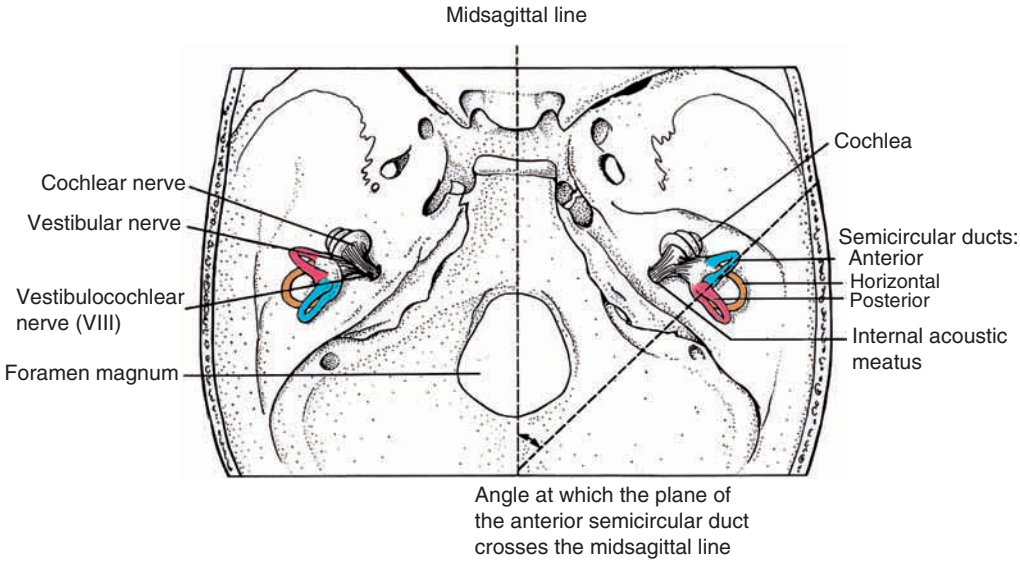

Figure 2 Location of the labyrinths in the skull, viewed from the top of the head. Color coding indicates the semicircular canals that share the same planar orientation, and which therefore are inversely activated by the same rotation. Adapted from Kandel ER, Schwartz JH, and Jessell TM (1991) *Principles of Neural Science*, 3rd edn. New York: Elsevier, with permission from Elsevier.

brain stem via the vestibular portion of the VIIIth cranial nerve. The central axons terminate on neurons within the vestibular nuclei in the pons and medulla and within the vestibulocerebellum (flocculonodular lobe of the cerebellum).

Some vestibular afferents fire phasically during accelerations produced by movements of the head, while others also fire tonically, providing constant information about the acceleration caused by gravity. Resting firing rates can be as high as 100 Hz or more. Because the two labyrinths and thus the orientation of hair cells in the various vestibular sensory organs are mirror symmetric, many head movements will produce an increase in afferent activity on one side and a decrease in afferent activity on the other side. For example, a leftward rotation of the head in the horizontal plane will increase the firing rate of horizontal canal afferents on the left side and decrease the firing rate of horizontal canal afferents on the right side (Figure 3). This change in the symmetry of activity on the two sides changes the activity in central structures and thereby triggers reflexive movements.

Asymmetry in vestibular activity can also be produced pathologically or experimentally, with severe effects on balance and orientation.

Most people have experienced vertigo and inability to maintain a constant posture following spinning in one direction for several seconds, or from the intake of too much alcohol, which reversibly changes the specific gravity of the otolithic membranes and the cupulae. Similarly, particles lodging on the otolithic membranes, loss of hair cells, damage to the vestibular nerve, or lesions within the brain stem can all cause vertigo and loss of balance. Menière disease is an example of such a vestibular pathology. Its etiology is poorly understood but appears to involve a disruption of endolymph drainage, leading to abnormally high pressure within the affected labyrinth. Clinicians induce changes in the balance of vestibular activity to assess the patency of central vestibular connections. Common techniques for achieving this include caloric stimulation of the inner ear on one side, and rotation of the patient on a specially constructed chair called the Bárány chair (named after Robert Bárány, who received the Nobel Prize in Physiology or Medicine in 1914 for his studies of the physiology and pathology of the human vestibular apparatus).

The vestibular nerve also conveys efferent axons from the brain stem to the vestibular sensory organs. These efferents appear to control the excitability of the hair cells, thus altering the sensitivity of vestibular afferents to stimulation.

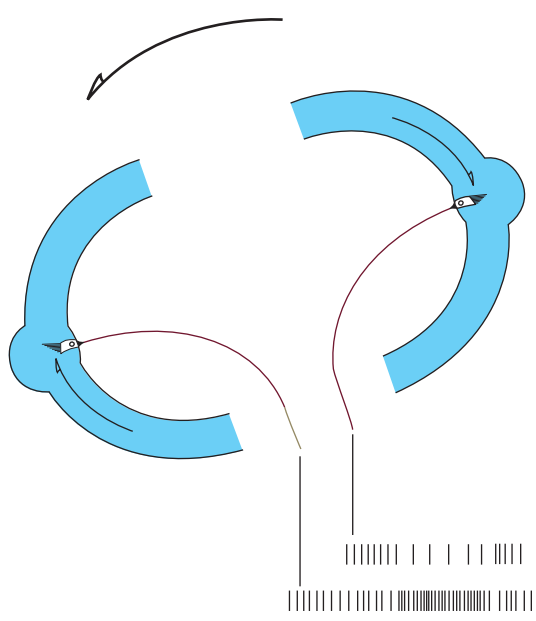

Figure 3 Effects of rotating the head to the left (large arrow) on activity in horizontal canal afferents. The inertia of the endolymph in the horizontal canals leads to deflections of the hair cells in opposite directions on the two sides. This causes complementary changes in the activity of the horizontal canal afferents (red axons) during the rotation (black bar).

Vestibular Nuclei and Their Projections

The vestibular nuclei are anatomically identifiable collections of neurons located in the dorsolateral regions of the pons and medulla. In humans and other mammals, four different nuclei can be distinguished: the superior nucleus, the lateral (or Deiters') nucleus, the medial nucleus, and the inferior (or descending) nucleus. Each of these nuclei projects to various targets in the brain and spinal cord, including (1) the oculomotor, trochlear, and abducens nuclei, which mediate eye movements; (2) reticular and spinal neurons involved in movements of the head, trunk, and limbs; (3) vestibular portions of the cerebellum; (4) the thalamus, for relaying vestibular information to cortical centers; and (5) other vestibular nuclei. Projections to movement-related centers are highly specific and mediate the stereotyped vestibular reflexes, such as the vestibulo-ocular and vestibulospinal reflexes. Projections to the thalamus, relayed on to the cortex, contribute to the conscious perception of self-movement and of orientation within the external environment.

Although the projections from the different anatomically defined vestibular nuclei are characteristic, the nuclei exhibit substantial overlap in terms of their projections to specific targets. Recent anatomical studies indicate, however, that projections to specific targets derive from largely separate, coherent clusters of vestibular neurons. Thus, different sets of vestibular neuron clusters project to the spinal cord,

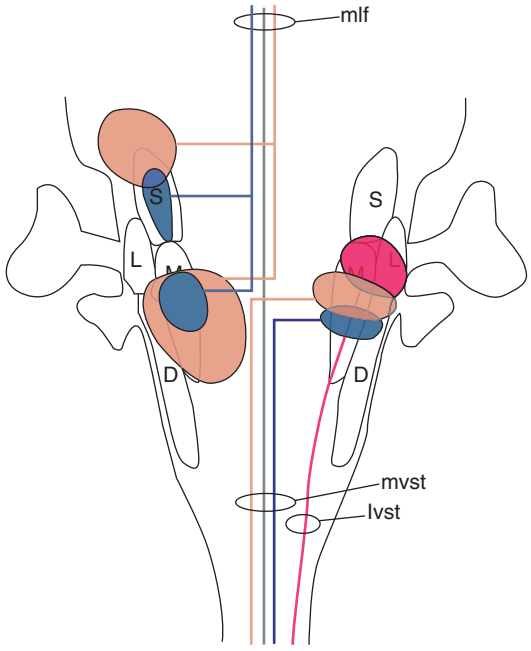

Figure 4 Overview of selected central vestibular pathways subserving the vestibulo-ocular and vestibulospinal reflexes. Superimposed on the vestibular nuclei (S, superior; L, lateral; M, medial; D, descending) are vestibular neuron clusters projecting to the oculomotor and trochlear nuclei (left side) and to the spinal cord (right side). Color coding indicates projections in different tracts (mlf, medial longitudinal fascicle; mvst, medial vestibulospinal tract; lvst, lateral vestibulospinal tract) and on different sides of the brain stem (ipsilateral, contralateral).

to the ocular motor neurons, and to the cerebellum. Moreover, the clusters projecting to these major targets are further subdivided according to their projection in specific axon tracts and whether they project selectively ipsilaterally or contralaterally (Figure 4). A given projection-specific cluster may be restricted to a subregion of one of the anatomically defined nuclei or overlap two or more of the nuclei.

This mosaic projection map within the population of vestibular neurons appears to be set up by highly determinate pattern-forming mechanisms operating during embryonic development of the brain stem. Afferents from the different vestibular sensory organs are presumably mapped onto the various neuron clusters of this projection map in such a way as to provide appropriate activation of the targets of the vestibular nuclei.

The vestibular nuclei receive projections from the cortex, the cerebellum, subcortical visual centers, and the spinal cord. Vestibular circuitry is therefore embedded within larger networks involved in the integration of vestibular, visual, and proprioceptive information. This provides an ongoing global control of eye and body orientation in the dynamic context of voluntary movement and external perturbations.

Vestibular Reflexes

The projections from the vestibular nuclei mediate two main classes of reflexes: the vestibulo-ocular reflexes and the vestibulospinal reflexes.

Vestibulo-ocular Reflexes

Vestibulo-ocular reflexes involve the activation of the motoneurons that innervate the extra-ocular muscles. The function of these reflexes is to maintain the orientation of the eyes in space during head movements so that the visual image remains stable on the retina. This is accomplished by the generation of eye movements that are equal and opposite to head movements. In the absence of such compensatory eye movements the visual image would shift wildly, particularly during nonvoluntary movements of the head, and visual orientation would be extremely difficult if not impossible.

Because the circuitry underlying the vestibulo-ocular reflexes is interconnected with networks that mediate other ocular reflexes as well as voluntary eye and head movements, these various movements are normally well coordinated. The same interconnections, however, can also lead to illusions, such as the illusion of self-movement when sitting in a motionless train and viewing a slowly moving train on a neighboring track.

Vestibulospinal Reflexes

Vestibulospinal reflexes involve the activation of spinal motor neurons that innervate neck, trunk, and limb muscles. The function of these reflexes is to maintain balance and body orientation in the face of ongoing perturbations. Like the vestibulo-ocular reflexes, the vestibulospinal reflexes are fast, providing compensatory movements that counteract and protect against the potentially injurious results of tripping or falling. Patients with vestibular disturbances risk bodily harm because these compensatory movements are deficient or absent.

Just as the vestibulo-ocular reflexes are coordinated with other reflexes and vision, the vestibulospinal reflexes are coordinated with other spinal reflexes and the vestibulo-ocular reflexes. This coordination is an essential element of our daily activities. An example is the coordination of the vestibulospinal reflex and the cervicospinal reflex (a proprioceptive reflex that generates limb movements in response to activation of neck proprioceptors). Each of these reflexes initiates unilateral limb extension when the head is tilted appropriately, but on opposite sides of the body (Figure 5). Tilting the neck while keeping the trunk motionless activates both reflexes, which

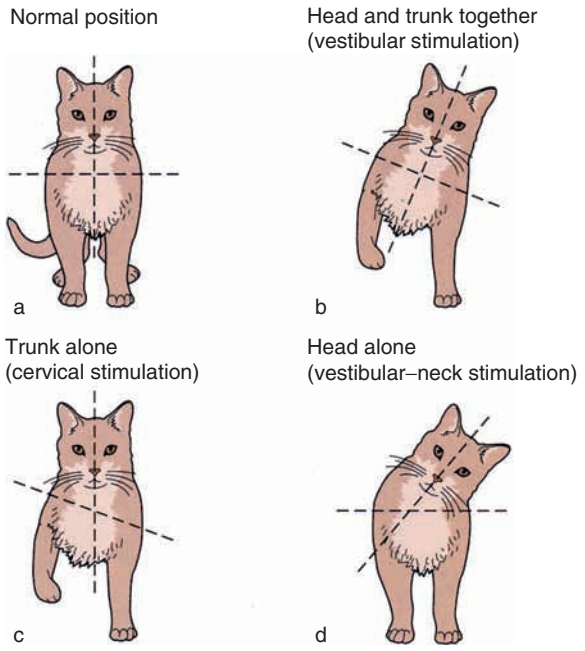

Figure 5 Coordination of vestibulospinal and cervicospinal reflexes. (a) Normal position. Tilting of head and trunk together relative to space (b) stimulates the vestibular apparatus only and elicits limb extension on the same side as the tilt and limb flexion on the opposite side. Tilting the trunk while keeping the head motionless (c) stimulates neck proprioceptors only and elicits limb flexion on the same side as the tilt and limb extension on the opposite side. Both of these responses serve to maintain balance. Tilting the head while keeping the trunk motionless (d) stimulates the vestibular apparatus and neck proprioceptors simultaneously. The reflex responses cancel out. Reproduced from Kandel ER, Schwartz JH, and Jessell TM (2000) *Principles of Neural Science*, 4th edn. New York: Elsevier, with permission from Elsevier.

therefore cancel out. In patients with vestibular pathology, a similar neck tilt will cause the patient to fall immediately to the same side because the cervicospinal reflex is activated alone. This uncomfortable (and potentially harmful) result can be experienced by healthy people following the vestibular imbalance produced by spinning in one direction for several seconds.

Plasticity and Adaptability

The vestibular reflexes, though highly stereotyped, exhibit plasticity at several levels, from hair cells to central synapses, allowing for adaptive changes in response. The following examples illustrate the range of plasticity. (1) Some vestibular afferents adapt to sustained stimulation, a feature that probably involves both mechanical adaptation of the hair cells and accommodation of the afferent nerve fibers. The result is an automatic diminution of reflex responses, for example, during protracted rotation. (2) Compensatory changes in central activity occur over time as a consequence of vestibular imbalance. Thus, the vertigo that results from a lesion of the vestibular apparatus or nerve on one side eventually diminishes, an effect that appears to involve a readjustment of neural activity within the vestibular nuclei. (3) Vestibular reflexes exhibit plastic changes in gain when the resultant movements do not achieve the desired result. For example, the vestibulo-ocular reflex is designed to maintain a stable image on the retina. If the reflex movement is incorrect (too large or too small relative to the head movement), the image will slip on the retina. Such a mismatch in head and eye movement can be created immediately by placing prisms or magnifying lenses in front of the eyes. Any such situation that causes retinal image slip will lead over the course of days to an appropriate adjustment of the magnitude and even direction of the reflex eye movement. This form of plasticity requires the involvement of the cerebellum and appears to be effected within specific regions of the vestibular nuclei. (4) Vestibular reflexes are under continuous voluntary control, which allows for rapid adaptive changes associated with different contexts of movement and orientation. For example, we can extinguish the vestibulo-ocular reflex during voluntary head rotations if the situation demands that we maintain eye position relative to the head instead of relative to space. This technique is used by ballet dancers and figure skaters during spinning maneuvers to avoid disorientation and dizziness. Volitional modulation of vestibulo-ocular and vestibulospinal reflexes is probably an important element in the learning and execution of novel movement sequences, an essential element in athletics and dance as well as daily life.

See also: Balance and Posture Control; Canal–Otolith Interactions; Gravitational Effects on Brain and Behavior; Hair Cells: Sensory Transduction; Motion Sickness; Oculomotor System: Models; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vertigo; Vestibular Influences on Cognition; Vestibulo-Autonomic Responses; Vestibulo-Ocular Reflex.

Further Reading

Baloh RW and Honrubia V (1990) *Clinical Neurology of the Vestibular System*. Philadelphia, PA: FA Davis.

Büttner-Ennever JA (1999) A review of otolith pathways to brainstem and cerebellum. *Annals of the New York Academy of Science* 871: 51–64.

Díaz C, Glover JC, Puelles L, and Bjålie J (2003) Hodological and cytoarchitectonic correlates of vestibular nuclear organization in the 11-day chicken embryo. *Journal of Comparative Neurology* 457: 87–105.

du Lac S, Raymond JL, Sejnowski TJ, and Lisberger SG (1995) Learning and memory in the vestibulo-ocular reflex. *Annual Review of Neuroscience* 18: 409–441.

Glover JC (2000) Neuroepithelial compartments and the specification of vestibular projections. *Progress in Brain Research* 124: 3–21.

Graf W, Gerrits N, Yatim-Dhiba N, and Ugolini G (2002) Mapping the oculomotor system: The power of transneuronal labelling with rabies virus. *European Journal of Neuroscience* 15: 1557–1562.

Kandel ER, Schwartz JH, and Jessell TM (2000) *Principles of Neural Science*, 3rd and 4th edns. New York: Elsevier.

Massion J (1992) Movement, posture and equilibrium: Interaction and coordination. *Progress in Neurobiology* 38: 35–56.

Shinoda Y, Sugiuchi Y, Futami T, Kakei S, et al. (1996) Four convergent patterns of input from the six semicircular canals to motoneurons of different neck muscles in the upper cervical cord. *Annals of the New York Academy of Science* 781: 264–275.

Wilson VJ and Melville-Jones G (1979) *Mammalian Vestibular Physiology.* New York: Plenum.

Vestibulo-Autonomic Responses

B J Yates, University of Pittsburgh, Pittsburgh, PA, USA
T D Wilson, University of Western Ontario, London, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Necessity for Vestibulo-Autonomic Responses

It has been appreciated for over 100 years that signals from the vestibular system can influence the brain circuitry that coordinates nausea and vomiting. Thus, the most commonly appreciated vestibulo-autonomic responses are those associated with motion sickness. Less obvious are the contributions of the vestibular system to coordinating the alterations in blood distribution in the body and respiratory muscle activity that are needed during movement and changes in posture. Nonetheless, considerable evidence collected from both animal models and human studies has now established a role for the vestibular system in maintaining homeostasis during postural alterations.

Head-up tilts of quadrupeds, like standing in humans, typically results in the pooling of blood in the venous circulation of the lower body and a decreased return of blood to the heart. Because Starling's law of the heart stipulates that cardiac output is directly related to the venous return to the heart, these postural changes can lead to a sudden decrease in blood pressure. In order to maintain stable blood pressure during posturally related challenges to the cardiovascular system, it is necessary to produce vasoconstriction in the lower body, to increase the resistance in the vasculature and reduce peripheral blood pooling. As described in the following discussions, the vestibular system rapidly triggers this compensatory response at the onset of movements involving head-up orientations of the body.

Changes in posture can also affect the resting length of the respiratory muscles, requiring alterations in the activity of these muscles if ventilation is to remain constant. For example, head-up rotations of quadrupeds and standing in bipeds can alter the resting length of the diaphragm, when the muscle is no longer supported by the abdominal contents. The vestibular system acts to stabilize the resting length of the diaphragm during head-up orientations of the body, mainly by producing an increase in the tone of the abdominal muscles and thereby preventing the abdominal organs from sinking into the abdominal cavity. The activity of some upper airway muscles also increases during certain postural alterations. This increase in muscle activity is most evident when humans assume a supine position or when quadrupeds are oriented head-up, as these postures allow the tongue to fall toward the back of the throat and potentially obstruct the airway. There is evidence suggesting that the vestibular system participates in triggering changes in activity of muscles such as the tongue protruder genioglossus, in order to maintain patency of the airway during postural changes.

Studies of Vestibulo-Cardiovascular Responses in Animal Models

A number of studies conducted on a variety of animal species have shown that electrical stimulation of the vestibular nerve elicits changes in the activity of sympathetic nervous system efferents, including fibers innervating arterioles in skeletal muscle. Selective natural stimulation of vestibular afferents produced by head rotations in animals with extensive denervations (transection of cranial nerves IX and X and the cervical dorsal roots) to eliminate nonlabyrinthine inputs that could be elicited by the head movement also affects sympathetic nerve discharges and blood pressure. In particular, rotating the head nose-up produces increases in sympathetic nerve activity. The response characteristics of vestibular-elicited alterations in sympathetic nervous system activity suggest that the responses are elicited by inputs from otolith organs, with little contribution from the semicircular canals. Removal of vestibular inputs through a bilateral vestibular neurectomy or bilateral lesions placed in the caudal vestibular nuclei results in lability in blood pressure during 60° head-up rotations of anesthetized or conscious cats (see Figure 1). This blood pressure lability is at least partly a reflection of decreased vasoconstriction in and increased blood flow to the lower body following the vestibular lesions (see Figure 2). Presumably, blood pooling in the lower body results in decreased venous return to the heart and thus a reduction in blood pressure. Cumulatively, these findings show that the vestibular system participates in adjusting blood distribution in the body and thus blood pressure during postural alterations.

The vestibular system has very specific influences on sympathetic nervous system outflow. Only those sympathetic nerve fibers that innervate vascular smooth muscle, as opposed to targets such as smooth muscle in the wall of the gut, respond to vestibular stimulation. Moreover, the vestibular system has distinct influences on vasoconstrictor fibers innervating blood vessels in the same tissue, skeletal muscle, but which are located in different body regions. Accordingly, stimulation of vestibular afferents or removal of

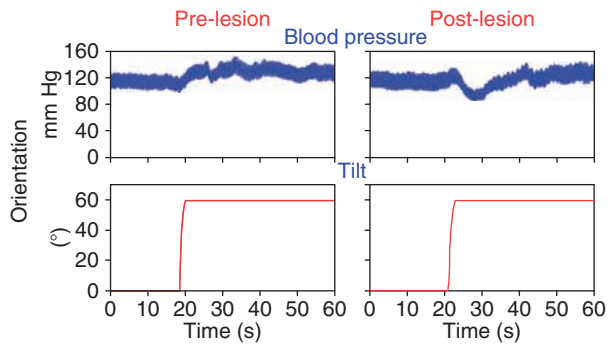

Figure 1 Effects of 60° head-up tilts and bilateral lesions of the caudal vestibular nuclei on the ability of a conscious cat to maintain stable blood pressure. Top row: Blood pressure measured from the abdominal aorta. Bottom row: Orientation of the animal's body. Prior to placing chemical lesions bilaterally in the caudal vestibular nuclei, blood pressure increased slightly during head-up rotations of the animal's body (left column). After the lesions, blood pressure dropped at the onset of the tilts (right column). For more information about the experimental design, see the work of Mori and co-workers.

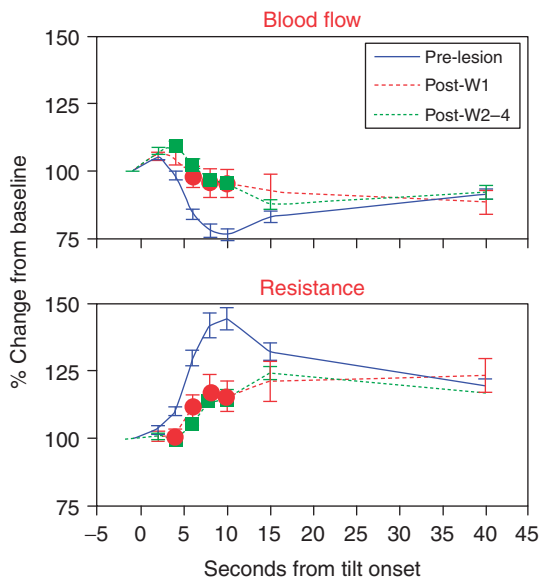

Figure 2 Changes in blood flow through the femoral artery (top panel) and resistance of the hind limb vasculature (bottom panel) during 60° head-up tilts of a conscious cat. Values are expressed as percent changes from baselines prior to tilts. When vestibular inputs were intact (solid blue lines), blood flow decreased to the hind limb during tilts as a result of increased hind limb vascular resistance. This response presumably prevented blood pooling in the lower body. In the first week following removal of vestibular inputs (Post-W1, red dashed lines) as well as during the subsequent 3 weeks (Post-W2–4, green dashed lines), the ability to reduce blood flow to the hind limb during head-up tilts was attenuated. Symbols indicate post-lesion responses that were significantly different from those prior to lesions ($P < 0.05$, MANOVA combined with LSD post-hoc tests).

vestibular inputs through a bilateral vestibular neurectomy produces different effects on blood flow to the forelimb and hind limb. As noted previously, vestibular inputs trigger vasoconstriction in the hind limbs during head-up rotations (Figure 2); this vasoconstrictor response is much smaller in the forelimbs. These observations show that vestibular signals do not elicit nonspecific 'fight or flight' responses that involve simultaneous increases or decreases in sympathetic outflow to many tissues and body regions. Instead, particular vestibular inputs (i.e., those elicited during head-up rotations) produce specific effects on the cardiovascular system that are appropriate to prevent peripheral blood pooling during postural alterations.

Experiments employing anatomical, neurophysiological, and lesion techniques have established the most direct neural pathways that mediate vestibular influences on the sympathetic nervous system; these pathways are summarized in Figure 3. Lesions placed caudally within the medial and inferior vestibular nuclei abolish sympathetic nervous system responses to stimulation of vestibular receptors. This region of the vestibular nuclei does not play a primary role in producing vestibulo-spinal or vestibulo-ocular responses, and may be predominantly involved in triggering autonomic responses required during postural

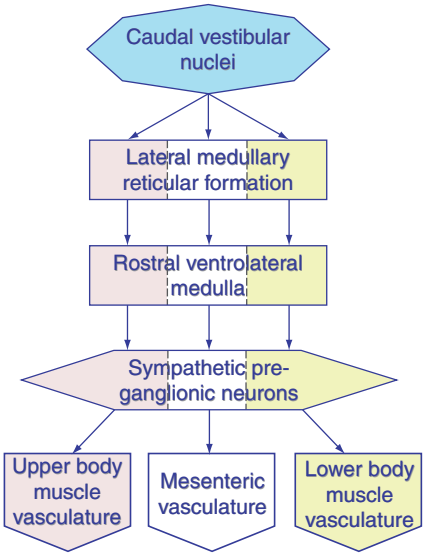

Figure 3 Minimal neural pathways that mediate vestibular influences on cardiovascular regulation. These pathways include the caudal vestibular nuclei, interneurons in the lateral medullary reticular formation, and bulbospinal neurons in the rostral ventrolateral medulla that make connections with sympathetic preganglionic neurons in the thoraco-lumbar spinal cord. Because vestibulo-cardiovascular responses vary between body regions, parallel neural pathways that receive distinct labyrinthine signals presumably control blood flow in different vascular beds.

alterations. Vestibular signals are relayed to sympathetic nervous system preganglionic neurons in the thoraco-lumbar spinal cord via bulbospinal projections from the rostral ventrolateral medulla (RVLM). This region of the reticular formation also mediates a variety of other cardiovascular responses generated through alterations in sympathetic nervous system outflow, and the same RVLM neurons integrate signals from vestibular end organs, baroreceptors, and perhaps other sources. However, RVLM neurons do not receive direct inputs from the vestibular nuclei; labyrinthine signals are relayed to the RVLM through interneurons located in the lateral medullary reticular formation. As discussed in the previous paragraph, vestibulo-sympathetic responses are patterned, and differ between body regions. The neural circuitry that mediates vestibular influences on sympathetic nervous system outflow must have the capability of producing such intricate responses. For example, bulbospinal neurons in the RVLM that influence blood vessels in the forelimb presumably have different responses to labyrinthine inputs than do RVLM neurons that regulate blood flow to the hind limb.

Although removal of vestibular inputs through a bilateral vestibular neurectomy compromises the ability to adjust blood distribution in the body and blood pressure during postural alterations, these deficits resolve over time. In contrast, lesions placed bilaterally within the caudal vestibular nuclei result in impaired cardiovascular responses to alterations in posture that persist for a longer period. In addition, the combination of a bilateral vestibular neurectomy with ablation of the cerebellar uvula produces a long-lasting inability to adjust blood pressure during postural changes. These findings have led to the following two hypotheses: (1) recovery of compensatory cardiovascular responses following peripheral vestibular lesions is accomplished at least in part through plasticity in the caudal regions of the vestibular nuclei that participate in autonomic regulation and (2) the occurrence of the plastic alterations in the caudal vestibular nuclei requires inputs from the cerebellar uvula. These notions are supported by the observations that the caudal vestibular nuclei receive substantial inputs from sensory receptors in addition to those in the labyrinth, and that these nonlabyrinthine inputs are sufficient to generate changes in the activity of vestibular nucleus neurons during postural alterations. Thus, a variety of sensory inputs likely contribute to producing compensatory cardiovascular responses during postural alterations; this multisensory processing presumably provides for a highly accurate appraisal of the challenges being imposed on the cardiovascular system and permits precise compensatory adjustments in blood flow and blood pressure.

Studies of Vestibulo-Respiratory Responses in Animal Models

Experiments similar to those discussed in the previous section have established that the vestibular system elicits alterations in respiratory muscle activity during postural alterations. Electrical or selective natural stimulation of vestibular receptors produces changes in the activity of respiratory muscles or nerves innervating those muscles in anesthetized animal preparations. The activity of many respiratory muscles is affected by vestibular stimulation, although vestibulo-respiratory responses are particularly prominent in the abdominal muscles and the tongue protruder muscle genioglossus (see Figure 4). Removal of vestibular inputs through a bilateral labyrinthectomy has two effects on respiratory muscle activity: (1) a decrement in the responses of the muscles to postural alterations and (2) an increase in background muscle activity. This observation suggests that the vestibular system has two parallel roles in regulating breathing. The physiological significance for labyrinthine signals affecting the spontaneous activity of respiratory muscles remains to be determined.

A variety of parallel neural pathways appear to mediate vestibular system influences on respiratory muscle activity. The same region of the caudal vestibular nuclei that participates in generating vestibulo-cardiovascular responses also is critical for labyrinthine signals to affect breathing. Historically, premotor respiratory neurons were thought to be mainly confined to two columns located in the lateral reticular formation of the caudal medulla, the so-called brain stem respiratory groups. Although some neurons in the respiratory groups respond to vestibular stimulation, functional lesions that inactivate the neurons or their axons do not eliminate vestibulo-respiratory responses. Furthermore, vestibular stimulation can produce an increase in activity of respiratory pump muscles without affecting the firing of respiratory group neurons. Thus, premotor neurons in addition to those in the respiratory groups must make synaptic contacts with respiratory motoneurons. The locations of these neurons were determined through studies in which a transneuronal tracer, pseudorabies virus, was injected into the diaphragm, abdominal musculature, and genioglossus. These experiments showed that additional premotor respiratory neurons are located in two regions: the medial medullary reticular formation and the spinal gray matter. Neurophysiological studies have shown that both of these neuronal populations receive labyrinthine signals and participate in relaying labyrinthine signals to respiratory motoneurons. Additional studies will be required to establish the specific role that each group of neurons plays in generating vestibulo-respiratory responses.

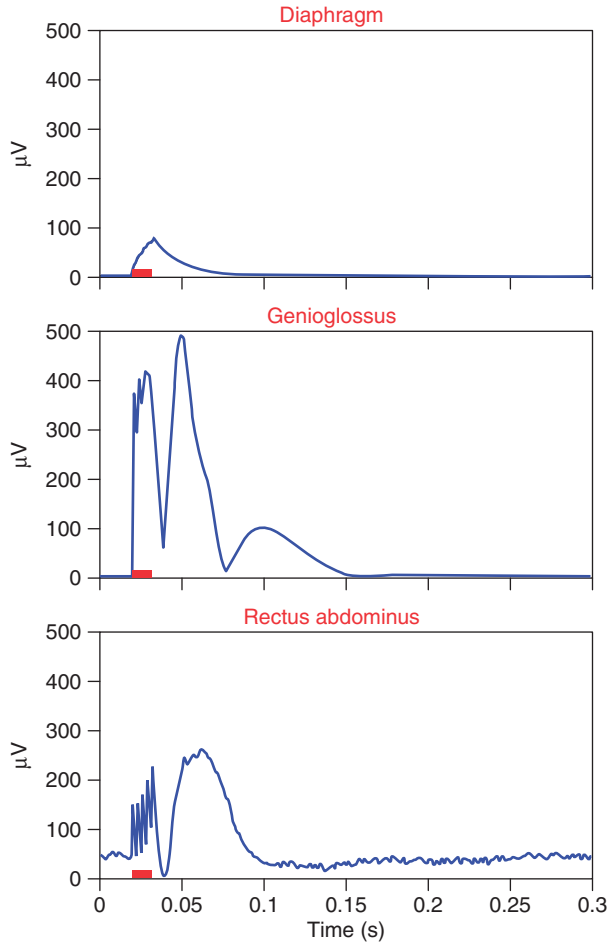

Figure 4 Averaged electromyogram (EMG) responses recorded from the diaphragm, the genioglossus, and the abdominal muscle rectus abdominis subsequent to electrical stimulation of the ipsilateral vestibular nerve (using a five-shock train of stimuli of 500 μA intensity; stimulation indicated by a red bar beneath each trace). The EMG responses for the genioglossus and rectus abdominis were prominent, but those for the diaphragm were small.

Studies of Vestibulo-Autonomic Responses in Humans

Whereas a number of studies conducted on human volunteers have considered vestibular-cardiovascular responses, only a few experiments have addressed vestibular influences on breathing. It is difficult to delineate a definitive role for the vestibular system in autonomic regulation in humans due to the evolution of redundant and overlapping homeostatic systems. In animal models, it is possible to perform lesion studies that completely remove particular sensory inputs; such experiments are impossible in humans. Nonetheless, a variety of experimental approaches have revealed that vestibular stimulation can lead to alterations in autonomic nervous system activity and obvious cardiovascular effects in humans.

Studies employing changes in head orientation have shown that the vestibular system has immediate and sustained influences on cardiovascular control. A number of researchers have utilized a specific maneuver, head-down neck flexion (HDNF), to provide relatively selective stimulation of the vestibular apparatus. Activity of sympathetic nervous system efferent fibers in peripheral spinal nerves increases during HDNF, as does vascular resistance. It has been hypothesized that peripheral vasoconstriction elicited by vestibular stimulation shunts blood flow to susceptible vascular beds during changes in body position that potentially compromise cardiac output and perfusion of the brain. Furthermore, evidence collected from human studies has led to the premise that the vestibular system triggers a coordinated and patterned vascular response to protect vulnerable vascular beds as opposed to global vasoconstriction. A study using Doppler ultrasonography to monitor gross cerebral blood flow during vestibular stimulation in humans verifies this postulate. Figure 5 demonstrates that cerebral vascular resistance decreases during HDNF relative to that measured during a similar maneuver where the neck was extended; in contrast, the same stimulus produces vasoconstriction in the peripheral vasculature. Interestingly, the attenuation of arterial resistance in the cerebral vessels occurred only when study participants underwent baroreceptor-unloading equivalent to that occurring during standing. This finding indicates a potential interaction between vestibular and baroreceptor afferent signals (see preceding section, 'Studies of vestibulo-cardiovascular responses in animal models') that results in a physiologically appropriate response.

Further information regarding the vestibulo-sympathetic reflex in humans has been obtained using an experimental paradigm called off-vertical axis rotation (OVAR). In this paradigm, participants are seated in a chair tilted slightly off Earth-vertical and are spun at constant angular velocity such that head position with respect to gravity constantly changes, producing a continually varying stimulus to the vestibular otolith organs. OVAR elicits alterations in activity of sympathetic nervous system efferents coursing in peripheral nerves; sympathetic efferents fire more when the nose is rotated upwards (forward linear acceleration) and less when the nose is oriented downwards. The short latency of the sympathetic efferent responses to OVAR suggests that the effects are due to activation of otolith organs, and not visceral afferents such as baroreceptors.

Other methods for activating vestibular receptors, including linear acceleration of the body using a sled, also produce cardiovascular responses in humans.

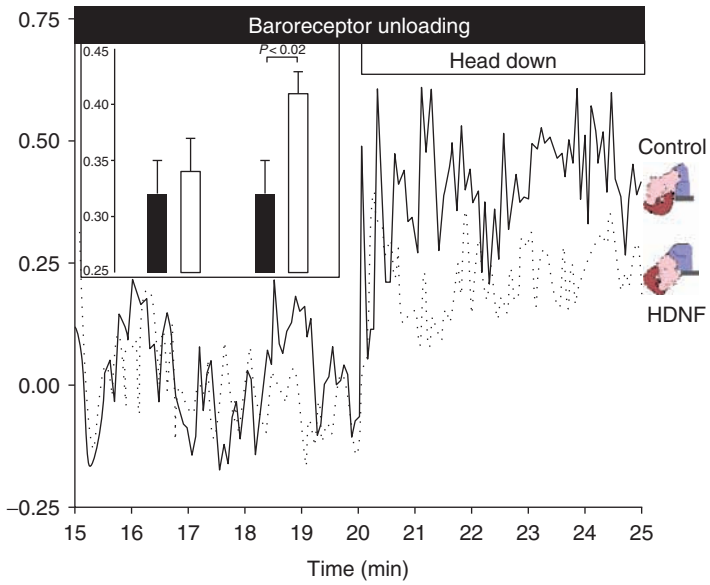

Figure 5 Beat-by-beat changes in brain blood vessel resistance (expressed in arbitrary units) during head-down maneuvers in a single individual with simulated orthostatic stress (baroreceptor unloading). The solid line indicates cerebral blood vessel resistance determined when the head was in the control position (head extension), while the broken line represents resistance determined during head-down neck flexion (HDNF). The inset graph shows the consequences of baroreceptor unloading on the maximal changes in cerebral resistance (expressed as mean ± SE) during neck flexion or extension. The filled bars indicate the effects of head position changes alone while the open bars indicate the effects of head maneuvers in combination with baroreceptor unloading. Filled bars represent the condition when the neck was flexed while open bars indicate the control condition ($n = 6$).

Changes in blood pressure elicited by linear acceleration were significantly smaller in individuals with a loss of vestibular inputs, suggesting that the responses in normal individuals were due to otolith organ stimulation. These findings support aforementioned research conducted using animal models and suggest that the vestibular system participates in homeostatic control at the very onset of movements and postural changes. Vestibulo-autonomic responses provide for feedforward adjustments in blood pressure, unlike responses triggered by baroreceptors that are elicited only after blood pressure has been altered. Baroreceptor and vestibular afferent information are integrated in the brain stem to provide an appropriate autonomic response during alterations in body position that provide a potential threat to the maintenance of stable blood pressure.

Clinical Implications of Dysfunction of Vestibulo-Autonomic Responses

Damage to a labyrinth on one side typically leads to vertigo, nausea, vomiting, and other autonomic disturbances. However, recovery generally occurs quickly due to the tremendous capacity of the vestibular system to undergo plastic changes when sensory inputs are altered. Plastic changes in the vestibular system during spaceflight may partly explain the occurrence of postspaceflight orthostatic intolerance, or inadequate perfusion of the brain when standing. Alterations in the pattern of inputs from the vestibular otolith organs to the central nervous system occur during spaceflight because the otolith organs are no longer exposed to constant gravitational forces. Thus, the otolith organs are no longer affected by tilting the head, as occurs on Earth. It is well established that exposure to an altered gravitational environment evokes significant adaptation in the vestibular system, and alters the gain of responses elicited by signals from the otolith organs. Attenuation of vestibulo-cardiovascular responses during spaceflight may partly explain why some astronauts, upon their return to Earth, have difficulty in maintaining adequate blood flow to the brain when standing.

Some evidence suggests that patients with bilateral loss of vestibular inputs may be more prone to suffering from orthostatic intolerance than individuals with a normal vestibular system. In addition, certain anxiety disorders that include autonomic symptoms have been linked with dysfunction of the vestibular system; it is possible that the aberrant vestibular signals in these patients could contribute to their alterations in blood pressure and breathing. These observations suggest that clinicians treating patients with vestibular disorders should be mindful that these conditions could lead to autonomic disturbances that also require medical attention.

See also: Balance and Posture Control; Balance and Posture Control: Human; Blood Pressure: Baroreceptors; Motion Sickness; Respiration; Vestibular System.

Further Reading

Cotter LA, Arendt HE, Jasko JG, et al. (2001) Effects of postural changes and vestibular lesions on diaphragm and rectus abdominis activity in awake cats. *Journal of Applied Physiology* 91: 137–144.

Doba N and Reis DJ (1974) Role of the cerebellum and vestibular apparatus in regulation of orthostatic reflexes in the cat. *Circulation Research* 34: 9–18.

Harper RM and Sauerland EK (1978) The role of the tongue in sleep apnea. In: Guilleminault C and Dement WC (eds.) *Sleep Apnea Syndromes*, pp. 219–234. New York: Liss.

Kaufmann H, Biaggioni I, Voustianiouk A, et al. (2002) Vestibular control of sympathetic activity: An otolith-sympathetic reflex in humans. *Experimental Brain Research* 143: 463–469.

Kerman IA and Yates BJ (1998) Regional and functional differences in the distribution of vestibulosympathetic reflexes. *American Journal of Physiology Regulatory, Integrative and Comparative Physiology* 275: R824–R835.

Kerman IA, Yates BJ, and McAllen RM (2000) Anatomic patterning in the expression of vestibulosympathetic reflexes. *American Journal of Physiology Regulatory, Integrative and Comparative Physiology* 279: R109–R117.

Mori RL, Cotter LA, Arendt HE, et al. (2005) Effects of bilateral vestibular nucleus lesions on cardiovascular regulation in conscious cats. *Journal of Applied Physiology* 98: 526–533.

Rushmer RF (1976) *Cardiovascular Dynamics*. Philadelphia: W.B. Saunders.

Shortt TL and Ray CA (1997) Sympathetic and vascular responses to head-down neck flexion in humans. *American Journal of Physiology Heart and Circulatory Physiology* 272: H1780–H1784.

Wilson TD, Serrador JM, and Shoemaker JK (2003) Head position modifies cerebrovascular response to orthostatic stress. *Brain Research* 961: 261–268.

Yates BJ, Billig I, Cotter LA, et al. (2002) Role of the vestibular system in regulating respiratory muscle activity during movement. *Clinical and Experimental Pharmacology and Physiology* 29: 112–117.

Yates BJ and Bronstein AM (2005) The effects of vestibular system lesions on autonomic regulation: Observations, mechanisms, and clinical implications. *Journal of Vestibular Research* 15: 119–129.

Yates BJ, Jakus J, and Miller AD (1993) Vestibular effects on respiratory outflow in the decerebrate cat. *Brain Research* 629: 209–217.

Yates BJ and Kerman IA (1998) Post-spaceflight orthostatic intolerance, possible relationship to microgravity-induced plasticity in the vestibular system. *Brain Research Reviews* 28: 73–82.

Yates BJ and Miller AD (eds.) (1996) *Vestibular Autonomic Regulation*. Boca Raton, FL: CRC Press.

Vestibulo-Ocular Reflex

D E Angelaki, Washington University School of Medicine, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

The vestibulo-ocular reflexes (VORs) are a group of eye movements whose goal is to rotate the eyeball in a particular direction and with a specific speed such that the world remains stable on the retina during head movement. Two different types of VORs are distinguished, depending on whether the head movement consists of a rotation or a translation. The former elicits the rotational or angular VOR (RVOR; e.g., when the head rotates to the right, the RVOR causes the eyes to rotate to the left, such that gaze remains stable in space). The latter elicits the translational VOR (TVOR; e.g., when the head translates to the right, the TVOR can rotate the eyes to the left to compensate for the relative motion of near targets on the retina). Even though natural head movements comprise both rotations and translations, there are characteristic differences between the RVOR and the TVOR so their properties are presented separately here.

Rotational or Angular Vestibulo-Ocular Reflex

A purely RVOR is the ocular responses elicited by rotation of the head or body in darkness about an axis roughly centered in the interaural axis of the head. Under these conditions, only the semicircular canals (but not the otolith organs) in the inner ear are effectively stimulated and provide the sole drive to rotate the eyes. The RVOR represents a phylogenetically old reflex. Many invertebrates and all vertebrate species, from amphibians, reptiles, and birds to nonhuman primates, have the ability to reflexively rotate their eyes opposite to the direction of head rotation, keeping the visual world stable on the retina. This superb ability for a very fast (<10 ms) gaze stabilization response comes about because of a fast transduction process and a most direct three-neuron arc pathway (Figure 1).

The RVOR is organized in a simple push–pull arrangement using signals from both labyrinths. During rightward head rotation, for example, primary afferents from the right horizontal semicircular canal increase their firing rates (Figure 1, pink lines). This increase in activity is conveyed by excitatory neuron projections via the vestibular nuclei to both motor neurons and internuclear neurons in the left (contralateral) abducens (IV) nucleus. The internuclear neurons in the left abducens nucleus convey this activity on to the medial rectus motor neurons in the right oculomotor nucleus (III). Activation of these pathways results in a simultaneous contraction of the lateral rectus muscle of the left eye and the medial rectus of the right eye (Figure 1, pink lines). Thus, both eyes rotate simultaneously to the left (black open arrows). Because of bilateral symmetry of the premotor circuitry, the decrease in the firing rates of the primary afferents from the left canal during rightward head rotation results in a parallel reduction in the tension of the antagonist muscles (i.e., the left medial rectus and the right lateral rectus; Figure 1, blue lines). This push–pull organization further supports the appropriate yoking of the eyes in the opposite direction to head movement.

The three-neuron arc schematic in Figure 1 is not sufficient to elicit appropriate RVOR. This is because the afferent signal from the semicircular canals is proportional to velocity, whereas the compensatory eye movement requires eye position changes. To convert velocity to position requires temporal integration (just like in calculus). At high frequencies (>0.5 Hz), the viscoelastic properties of the ocular plant (eyeball, muscles, and surrounding tissue) provide the second integration. At low frequencies, however, neural integration becomes necessary. Thus, the RVOR is thought to consist of two parallel processes, as illustrated in the classic Robinson diagram in Figure 2. The presence of both direct and neural integrator parallel pathways ensures that the correct proportion of velocity and position commands are delivered by the brain to the plant at all frequencies of natural head movement. In the absence of such an indirect integrator pathway, the response to a head rotation would bring the eye to the correct position, but subsequently eye position would drift back because the motor neurons would lack the tonic component to compensate for the elastic restoring forces of the eyeball (in Figure 2, compare the dashed and solid black lines). This is exactly what happens after lesions of brain stem and cerebellar structures that are thought to participate in this neural integration (e.g., the prepositus hypoglossi and the interstitial nucleus of Cajal). It is generally thought that the indirect integrator pathway is shared by all conjugate eye movement systems (saccades, pursuit, and the VORs), although the direct pathways are at least partly segregated for different types of eye movement (i.e., VORs, pursuit, or saccades).

For small head rotations, when the compensatory ocular deviation is small (10–15°), the RVOR is

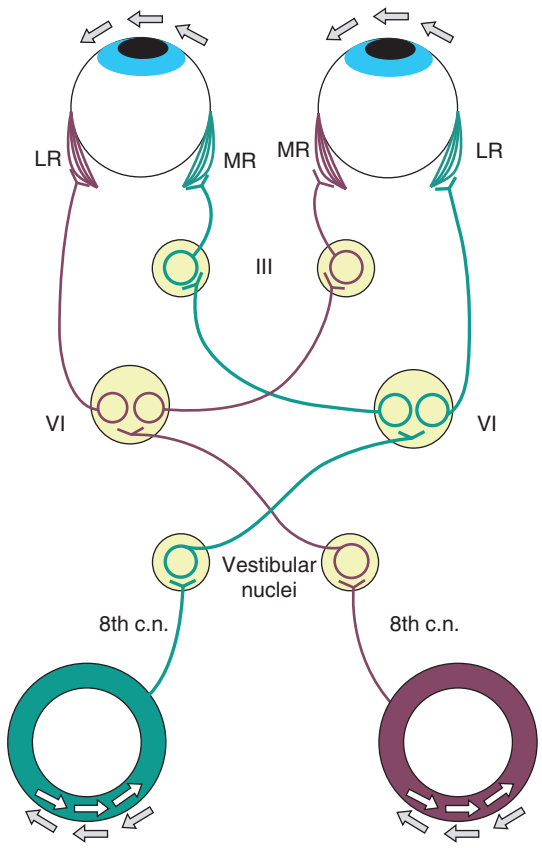

Figure 1 The push–pull organization of the horizontal (yaw) RVOR. The direct path consists of a short trisynaptic reflex consisting of vestibular afferents (eighth nerve), the vestibular nuclei, motor neurons in the VI (abducens) and III (oculomotor) nuclei and the lateral rectus (LR) and medial rectus (MR) muscles. When the head rotates to the right, the right horizontal semicircular canal afferents increase their firing rates, whereas the left afferents decrease their activities (pink vs. blue colors, respectively). The excitation/inhibition is transmitted to the contralateral/ipsilateral abducens (VI) through neurons in the right/left VN, respectively. Internuclear neurons in the abducens nuclei cross the midline and terminate in the contralateral medial rectus motor neurons (III). Thus, motor neurons in the left VI and right III fire at a higher frequency, whereas those in the right VI and left III fire at a lower frequency. As a result, the left LR and right MR muscles contract, whereas the left MR and right LR muscles relax and both eyes rotate leftward. c.n., cranial nerve.

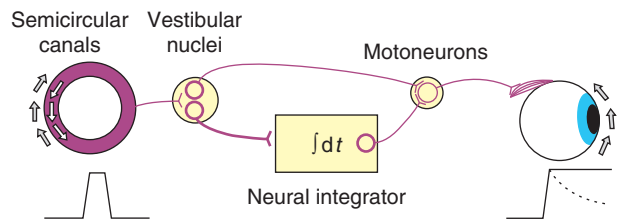

Figure 2 Schematic diagram of the classic Robinson-style model for the RVOR. Semicircular canal signals reach the motor neurons through a set of two parallel pathways: a direct route through the vestibular nuclei (three-neuron arc) and an indirect pathway that functions like a neural integrator ($\int dt$) and is thought to include cells in the prepositus hypoglossi and vestibular nuclei. As a result of these parallel pathways, a pulse of head velocity is converted into a sustained change in eye position (solid lines). Without the integrator pathway, eye position would decay to zero within a few hundreds of milliseconds (dashed line).

manifested as a slow deviation in eye position, having a rate of change that should ideally match the velocity of head rotation. For larger head rotations, the RVOR is typically manifested as nystagmus, a sawtoothlike (zigzag) pattern in which slow deviations in eye position are interleaved with oppositely directed quick phases (reflexive saccades) that keep eye position within the oculomotor range (Figure 3(a), top traces). Similar to the interaction of saccades and pursuit, ocular nystagmus might reflect two different control strategies, in which slow phases reduce the velocity error between eye motion and head motion and fast phases reduce the position error to keep the eyes centered in the orbit. The involuntary fast-phase segments of the RVOR are generally ignored to facilitate the analysis of the more salient slow-phase compensatory response segments (Figure 3(a), bottom trace).

The effectiveness of the RVOR is often characterized by its gain, which is defined as the ratio of the velocity of the nystagmus slow phase over the velocity of head rotation. Species differences exist in the exact value of RVOR gain. In monkeys, the horizontal RVOR gain in darkness is close to ideal (0.9–1) and appropriate in phase over a frequency range of ~0.1–4 Hz. In humans, the RVOR gain is much lower and more variable, ranging from 0.5 to 0.8 over a similar frequency range. This variability in the human RVOR gain could be due to different cognitive sets. For sinusoidal rotations, in addition to its gain, RVOR dynamics are also quantified by computing what is called the phase shift, defined as the timing difference between the actual eye velocity and that required to perfectly compensate for head rotation.

The compensatory function of the VOR extends to high frequencies. Although they are not a natural stimulus to the VOR, very high (up to 25–50 Hz) or very low (down to 0.01 Hz) frequencies are often tested. The RVOR gain continues to be high at these high frequencies, with a nearly perfectly compensatory phase. At very low frequencies, however, gain declines and progressively larger phase leads are introduced. This is partly due to the fact that the dynamics of the semicircular canal afferents, which provide the sensory drive for the RVOR, are high-pass sensors of head velocity, with a lower corner frequency of approximately 0.03–0.04 Hz in primates and cats. Thus, when the RVOR is studied using an abrupt change in head velocity followed by a maintained constant velocity rotation (often referred as a step in head velocity), the initial transient

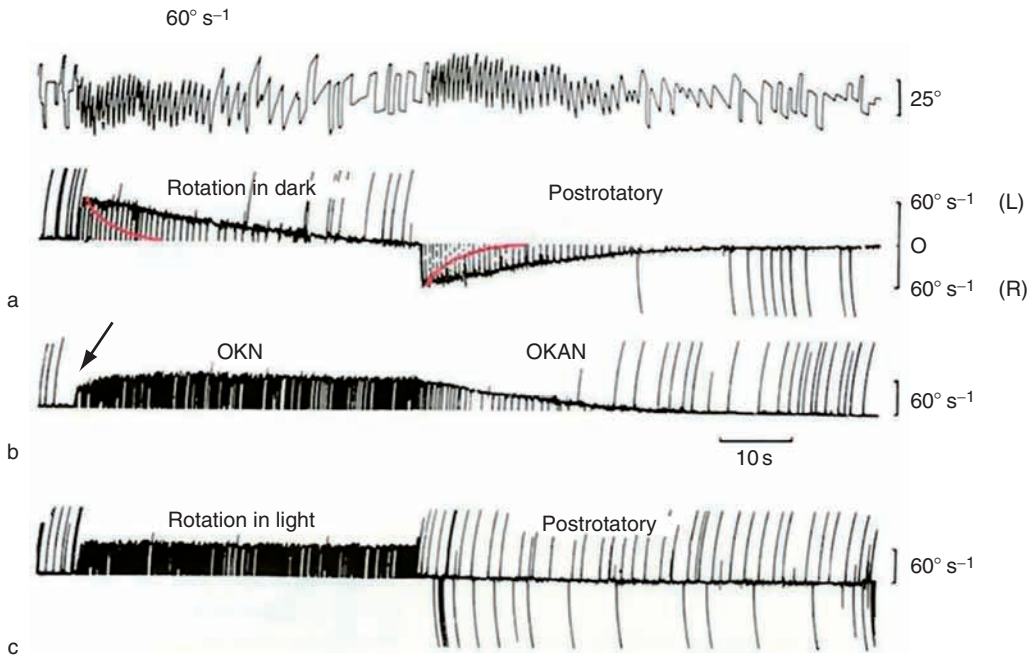

Figure 3 Interactions between the RVOR and optokinetic nystagmus (OKN): (a) nystagmus in response to a step of velocity (platform rotation) in darkness; (b) slow-phase eye velocity in response to a step in surround velocity, generating OKN and optokinetic after-nystagmus (OKAN); (c) slow-phase eye velocity in response to platform rotation in the light. In (a), the top traces illustrate horizontal eye position, whereas the bottom traces plot slow-phase eye velocity (the large positive and negative spikes reflect the occurrence of fast phases). The solid red line in the lower trace illustrates the semicircular canal afferent time course, which is much shorter than that of the RVOR. In (c), all stimuli consisted of horizontal rotation at 60° s^{-1}. Note that during rotation in the light the postrotatory RVOR and OKAN slow-phase velocities are oppositely directed (in a and b). Thus, the two afterresponses cancel one another and postrotatory responses during rotation in the light are small (c). Reprinted from Raphan T and Cohen B (1985) Velocity storage and the ocular response to multidimensional vestibular stimuli. In: Berthoz A and Melville Jones G (eds.) *Reviews of Oculomotor Research*, vol. 1, pp. 123–143. Oxford: Elsevier, with permission from Elsevier.

ocular response to the step decays out over several seconds (Figure 3(a)), as expected because the canals cannot transduce a constant velocity stimulus. Sudden cessation of rotation, then, is encoded by the semicircular canals as a step in the opposite direction, eliciting a response typically referred to as the postrotatory nystagmus (Figure 3(a)).

The time of decay of the RVOR during constant velocity rotation can be characterized by its time constant, which can be empirically defined as the time over which eye velocity to a step of head velocity decays to 63% of its peak value. In most species, including primates, cats, and rabbits, the rate at which the RVOR response to a step in head velocity decays out is slower than would be expected on the basis of the dynamics of the semicircular canals. The difference represents a prolongation or perseveration of the raw vestibular signal and researchers have proposed that it reflects a central processing of afferent information referred to as the velocity storage integrator. Although such low-frequency or very long-duration stimuli are totally unnatural, particularly when delivered in darkness, they have been very useful experimentally in studying both the dynamics of the RVOR and multisensory interactions.

An example of such interactions via the velocity storage integrator occurs between VOR and optokinetic responses. In particular, the RVOR cannot provide gaze stability during maintained rotation (Figure 3(a)). Instead, under natural circumstances (e.g., self-rotation in the light), sustained image stabilization is provided by the visual system that generates optokinetic nystagmus (OKN). The optokinetic reflex is classically considered to be the reflexive ocular response to the en bloc motion of a large visual field that aims to match slow-phase eye velocity to the velocity of the visual surround, providing a low-frequency complement to the RVOR. In the laboratory, the OKN system is usually stimulated by rotating a large patterned drum around the stationary subject. Under these circumstances, the nystagmus pattern generated in species with foveal vision typically consists of a fast-velocity rise (Figure 3(b), black arrow), followed by a slower increase in eye velocity to a steady-state level that is proportional to the rotational velocity of the drum. Following the cessation of drum rotation, eye velocity declines to zero gradually, after an occasional initial quick drop that is similar to the fast rise (optokinetic afternystagmus (OKAN); Figure 3(b)).

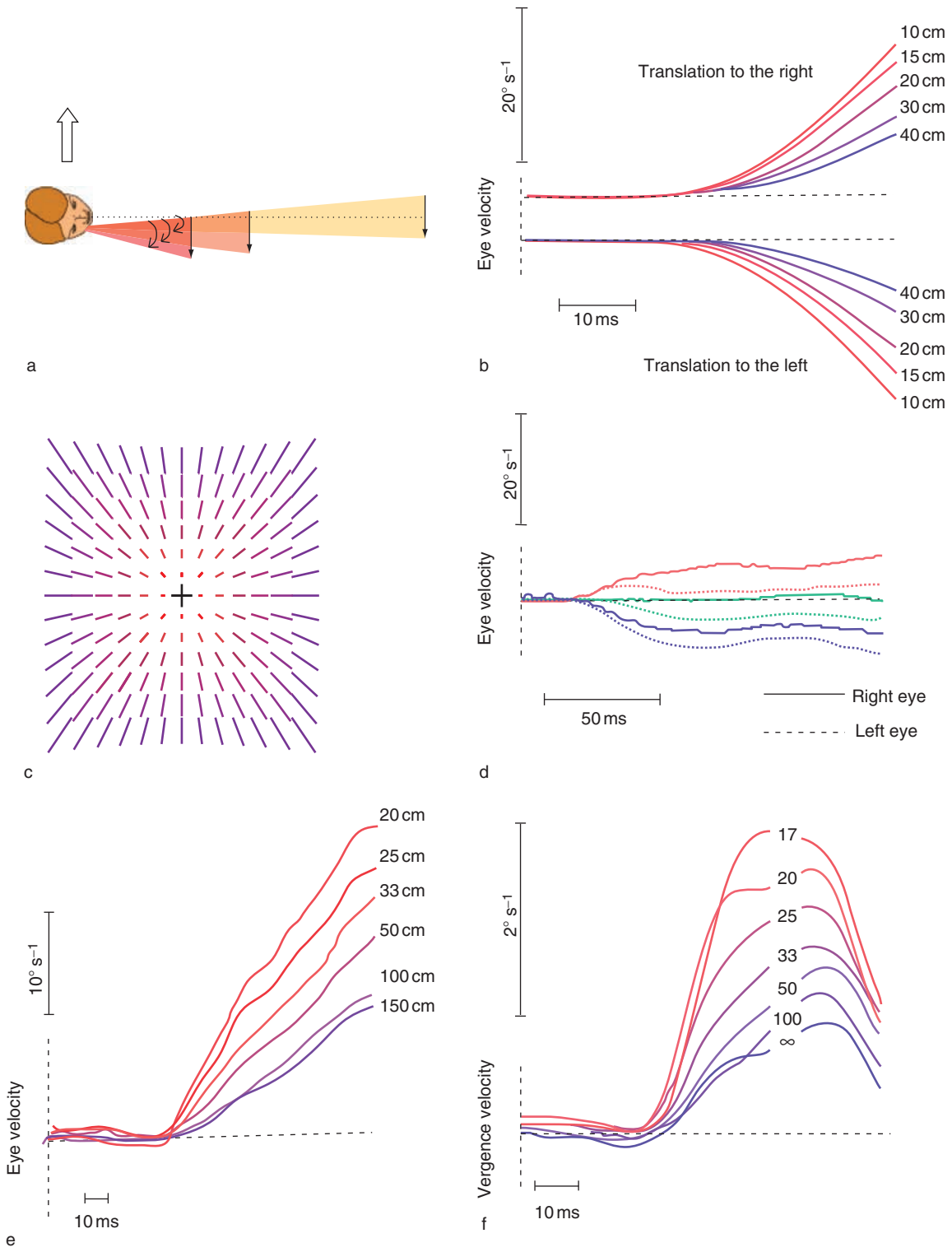

Figure 4 Properties of the TVOR: (a) for the same displacement, the angular deviation of the target on the retina depends on its proximity; (b) horizontal eye velocity (TVOR) amplitude during lateral motion is larger the nearer the target; (c) idealized optic flow pattern during forward motion while fixating straight ahead; (d) horizontal eye velocity (TVOR) amplitude during forward motion depends on target position; (e, f) the ocular following reflexes (OFRs) also depends on viewing distance. In (c), shades from red to blue illustrate the transition from the fovea to the retinal periphery. In (d), data plot eye velocity of the right and left eyes while fixating one of three targets ~15° to the left (red), straight ahead (green), or 15° to the right (blue). In (e), the graph shows conjugate horizontal eye velocity in response to tangential optic flow; in (f), the graph shows vergence eye velocity in response to radial optic flow. Upward deflections illustrate leftward eye velocity (b, d), rightward eye velocity (e), or increasing convergence (f). The numbers on each trace indicate viewing

The slow component of OKN–OKAN appears designed to perfectly complement the RVOR. Despite the fact that during constant-velocity rotation eye velocity decays to zero, the same rotation in the presence of a visual surround elicits a steady-state nystagmus pattern that is sustained indefinitely, as necessary to compensate for the applied stimulus (Figure 3(c)). This happens because the OKN builds up at approximately the same rate as the RVOR decays. Similarly, when head rotation is suddenly halted, OKAN cancels out the postrotatory vestibular nystagmus pattern, such that eye velocity drops quickly to zero. These visual–vestibular interactions are impaired after lesions that also impair velocity storage. Activation of the otolith system during rotations that change head orientation relative to gravity also improves the low-frequency RVOR properties. Similar to OKN, these otolith effects on the low-frequency RVOR are also mediated through the velocity storage mechanism.

Translational Vestibulo-Ocular Reflex

The TVOR results from the activation of the otolith afferents in the inner ear. It represents the functional equivalent of the RVOR during translation, although its properties are geometrically more complex than those of the RVOR. This occurs because, unlike the RVOR, in which a head rotation is compensated by an eye rotation (i.e., the stimulus and response are in similar units), a translational displacement must be compensated by an eye rotation, a nontrivial transformation that depends on target proximity (Figure 4). For example, during the same lateral displacement, nearby objects move on the retina more rapidly than distant ones (in Figure 4(a), curved arrows differ, straight arrows are the same). Thus, if a nearby object needs to be stabilized on the retina, the eyes should rotate by an amount that is larger the closer the target to the subject. Indeed, the horizontal compensatory eye movements that are elicited during lateral motion in darkness do indeed scale with target distance; the closer the target, the larger the evoked eye movement (Figure 4(b)). Similar to the RVOR, compensatory responses to translation occur at relatively short latency (10–12 ms; Figure 4(b)).

Unlike the uniform optic flow experienced during head rotation (in which all spatial locations in the visual field are uniformly represented on the retina), the optic flow experienced during translation can be complex, particularly when we are moving along a direction that is nearly parallel to the gaze (e.g., moving forward while looking straight ahead). In this case, we experience an expanding visual scene in which objects at the periphery of the visual field move much faster than those in more central portions of the retina (similar as in the star-field screen saver on my computer). This pattern of optic flow (typically referred to as radial flow) is complex because objects also appear to move in different directions depending on their spatial location within the visual field (Figure 4(c)). Thus, during forward motion, the retinal slip of a nearby object differs in both size and direction, depending on its location within the visual field. Furthermore, because forward motion is associated with distance changes, we must converge our eyes in order to keep images on corresponding locations on the two fovea. The reverse is true during backward motion, in which ocular divergence is necessary.

Obviously, no single eye movement could generally result in stabilization of the whole visual field during translation. Thus, unlike the RVOR, which is a full-field image stabilization reflex, the goal of the TVOR is to selectively stabilize images on the fovea. This occurs by scaling the vestibular signal not only by the inverse of target distance (Figure 4(b)), but also by eye position (Figure 4(d)). For example, when we move forward and look straight ahead, our eyes move little (Figure 4(d), green traces). But when we look to the right, forward motion causes the eyes to move further to the right (Figure 4(d), blue traces). The reverse is true when we look to the left, with a TVOR response to the left (Figure 4(d), red traces). Notice that the two eyes move disjunctively, consisting of either a purely vergence or a combination of vergence and conjugate eye movements. However, although the direction of the evoked eye movement is typically consistent with geometric predictions, the primate/human TVOR is typically undercompensatory for near-target viewing, with gains of only approximately 0.5.

The residual retinal image motion remaining after a less than ideal TVOR can be nulled by complementary visual stabilization mechanisms, known as the ocular following reflexes (OFRs). To generate OFRs, visually sensitive neurons in extrastriate visual cortex

distance (in centimeters); vertical dashed lines indicate motion onset. (a, d) Adapted from Macmillan Publishers Ltd., Angelaki DE and Hess BJM (2005) Self-motion induced eye movements: Effects on visual acuity and navigation. *Nature Reviews Neuroscience* 6(12): 966–976. (b) Redrawn from Angelaki DE and McHenry QM (1999) Short-latency primate vestibulo-ocular responses during translation. *Journal of Neurophysiology* 82: 1651–1654, © American Physiological Society. (e) Reprinted from Schwarz U, Busettini C, and Miles F A (1989) Ocular responses to linear motion are inversely proportional to viewing distance. *Science* 245: 1394–1396, © American Association for the Advancement of Science. (f) Adapted from Yang D, Fitzgibbon EJ, and Miles FA (1999) Short-latency vergence eye movements induced by radial optic flow in humans: Dependence on ambient vergence level. *Journal of Neurophysiology* 81: 945–949, © American Physiological Society.

decode either the pattern of optic flow (planar during lateral motion or radial during forward motion) or the disparity cues associated with motion-in-depth. Planar optic flow (e.g., experienced by a subject who looks off to one side during motion) elicits conjugate OFRs (Figure 4(e)). In contrast, when the subject looks in the direction of heading, disjunctive (vergence) visually driven eye movements are elicited (Figure 4(f)). A most characteristic feature of these responses is their very short latency for a visually driven mechanism (monkeys, <60 ms; humans, ~85 ms). Because the OFR shares the viewing distance dependence of the TVOR, the TVOR and OFRs are thought to share central pathways.

Notably, the TVOR differs from the RVOR also in the fact that the ability to manipulate retinal flow by generating compensatory eye movements during translation and to optimize visual acuity on the central retina appears to represent a specific feature of the primate and human visual systems. Lateral-eyed species, like the rabbit, that lack a fovea and a smooth pursuit system do not generate eye movements that compensate for the visual consequences of translation during self-motion. As in the primate oculomotor system, image stabilization during translation serves mainly foveal vision and stereovision, so the absence of a well-developed compensatory TVOR in animals that lack foveal vision might not be surprising.

Tilt Vestibulo-Ocular Reflex

These exists an additional effect of otolith system activation on reflexive eye movements because primary otolith afferents detect not only translational acceleration but also changes in the orientation of the head relative to gravity. Thus, an otolith-driven tilt response also exists, whose function is to maintain the orientation of the eyes constant in space during head tilts. This reflex includes the generation of torsional eye movements during head roll (counterrolling) or vertical eye movements during head pitch (counterpitching). (Please note that the direction of the evoked eye movement is described here relative to the head; i.e., vertical refers to eye rotation about the head's interaural axis, and torsional refers to eye rotation about the head's naso-occipital axis.) This ability to use a gravity-sensing mechanism to maintain gaze relative to the horizon is of paramount importance for lateral-eyed afoveate species that typically lack a well-developed saccadic system. But such functional utility for these tilt responses has lost its advantage in the primate oculomotor system, in which static ocular counterrolling and counterpitching in primates and humans has a gain of less that 0.1. Significantly, unlike afoveate species in which this reflex operates exclusively through slow eye movements, modulation of both slow and fast phases occurs in primates and reflects a change in the orientation of Listing's plane and primary eye position. Thus, the tilt otolith-ocular reflex in frontal-eyed animals might be less related to gaze stabilization but more fundamentally coupled to reference frames.

Plasticity

The VOR is a highly modifiable reflex. In essence, although the VOR acts independently of visually mediated eye movements, the brain continuously monitors its performance by evaluating the clarity of vision during head movements. When head turns are consistently associated with image motion across the retina, the RVOR undergoes plastic gain changes in the direction appropriate to improving the compensatory ability of the reflex. For example, when we view the world through spectacles that magnify or miniaturize the visual scene, RVOR gain (in darkness) increases or decreases accordingly. The reflex behavior can adapt over several minutes, hours, and days. During short-term adaptation, the RVOR gain can be adjusted by exposing the subject to a motion stimulus coupled with an abnormal visual stimulus. For example, head rotation paired with the rotation of an optokinetic visual stimulus at the same speed but in the opposite direction results in an increase of the RVOR gain. Similarly, rotating the subject and visual surround in tandem can be used to train the RVOR gain toward zero. For long-term adaptation, an animal is typically fitted with magnifying lenses to increase the gain, miniaturizing lenses to reduce the gain, or reversing prisms to reverse the direction of compensatory eye movement responses.

When mismatched visual-vestibular stimuli are used to adapt the RVOR gain at a single frequency, changes are typically frequency selective, with the largest effects being observed at the adapting frequency. In contrast, gain changes are less robust at other frequencies and are typically accompanied by phase lags or leads relative to an ideal response. RVOR training with brief versus longer-duration stimuli produces differential effects on the gain and dynamics of the RVOR, with longer-duration stimuli inducing larger effects on RVOR dynamics. It is also possible to adapt RVOR gain in opposite directions at low versus high frequencies or during the two directions of motion. Static head position relative to gravity, the position of the eye in the orbit, and the pattern of head movement have all been used as contextual cues for gating different vestibular responses. In addition, many studies have demonstrated an ability to adapt to injured

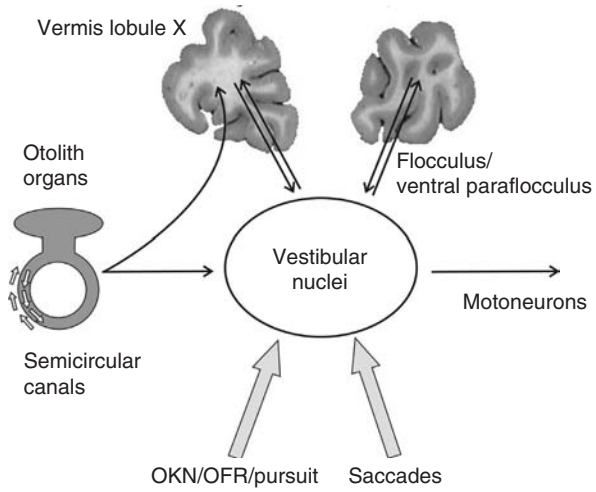

Figure 5 Schematic illustrating otolith and semicircular canal signal projections to the vestibulo-cerebellum and the vestibular nuclei. The vestibular nuclei, which receive not only vestibulo-ocular signals but also information related to OKN, OFR, pursuit, and saccades, projects directly to extraocular motor neurons. OFRs, ocular following reflexes; OKN, optokinetic nystagmus.

states, including peripheral and central vestibular damage. Often RVOR gain adaptation is accompanied by changes in gaze-holding performance, with the degree of alterations being dependent on training context. Furthermore, disconjugate adaptation of the RVOR can occur, for example, by wearing prisms in front of one eye.

Otolith-ocular reflexes are also subject to adaptive control. For example, both the direction of evoked eye movement and how it depends on target position within the visual field have been shown to be modifiable. Prolonged centrifugation can lead to changes in the VORs, although prolonged exposure to static tilt does not change ocular counterrolling. Context-specific fast (~100 trials) motor learning in the TVOR can also be induced when monkeys are repeatedly subjected to a brief head translation while trying to maintain binocular fixation on a world-fixed or subject-fixed visual target for juice rewards.

Neural Pathways

A unique characteristic of the VORs is that they are generated exclusively from subcortical circuits that involve the lower pons/medulla and cerebellum. Three main structures participate in the generation of the VORs: the vestibular nuclei, the cerebellar flocculus/ventral paraflocculus, and the most posterior lobules of the vermis (lobule X) (Figure 5). The cerebellar areas are often commonly referred to as the vestibulo-cerebellum because they receive both direct and indirect vestibular afferent projections. The vestibular nuclei house both the direct vestibulo-ocular connections (comprising the three-neuron arc pathway; Figures 1 and 2), as well as some of the indirect projections involved in neural integration (Figure 2). The vestibulo-ocular processing through the cerebellar flocculus/ventral paraflocculus is involved in both neural integration and VOR plasticity. Finally, the posterior cerebellar vermis is primarily involved in otolith signal processing and controls the properties of the RVOR at low frequencies. Signals from both areas of the vestibulo-cerebellum project to the vestibular nuclei, where they directly contribute to the generation of the VORs.

The vestibular nuclei represent the final premotor center for all types of slow eye movements, including not only the VORs but also OKN, OFR, and pursuit (Figure 5). Many neurons in the vestibular nuclei also show a burst or pause in activity during saccades and fast phases of vestibular nystagmus. Several types of vestibular nuclei neurons have been described, differing in their response properties to rotation, saccades, and smooth pursuit eye movements. In particular, the position-vestibular-pause (PVP) cells are thought to represent the second-order neuron in the three-neuron arc pathway (Figures 1 and 2). Eye-head cells, on the other hand, are thought to be involved in VOR adaptation, many receiving inhibitory projections from Purkinje cells in the cerebellar flocculus/ventral paraflocculus.

Conclusion: Natural Behavior

The VORs represent some of the most well-understood sensorimotor actions. In particular, because of its simplicity, its phylogenetically preserved properties, and well-defined neuronal properties, the RVOR was in the forefront of discovery during the latter part of the twentieth century. The TVOR, on the other hand, having geometrically more complex properties and being phylogenetically newer, remains less well understood. In everyday activities, of course, the two VORs are both elicited in tandem. Furthermore, most often the head is not fixed relative to the body. Current research in the field focuses on understanding these additional features of vestibulo-ocular processing, including the roles of other motor and cognitive influences.

See also: Brainstem Control of Eye Movements; Canal–Otolith Interactions; Cerebellum and Oculomotor Control; Eye and Head Movements; Nystagmus; Oculomotor System: Models; Optokinetic Eye Movements; Vergence Eye Movements; Vestibular System; Visual–Vestibular Interactions.

Further Reading

Angelaki DE (2004) Eyes on target: What neurons must do for the vestibuloocular reflex during linear motion. *Journal of Neurophysiology* 92: 20–35.

Angelaki DE and Hess BJM (2005) Self-motion induced eye movements: Effects on visual acuity and navigation. *Nature Reviews Neuroscience* 6(12): 966–976.

Angelaki DE and McHenry QM (1999) Short-latency primate vestibulo-ocular responses during translation. *Journal of Neurophysiology* 82: 1651–1654.

Boyden ES, Katoh A, and Raymond JL (2004) Cerebellum-dependent learning: The role of multiple plasticity mechanisms. *Annual Review of Neuroscience* 27: 581–609.

Cannon SC and Robinson DA (1987) Loss of neural integrator of the oculomotor system from brain stem lesions in monkey. *Journal of Neurophysiology* 57: 1383–1409.

Hess BJM and Angelaki DE (1999) Oculomotor control of primary eye position discriminates between translation and tilt. *Journal of Neurophysiology* 81(1): 394–398.

Hess BJM and Angelaki DE (2003) Vestibular contributions to gaze stability during transient forward and backward motion. *Journal of Neurophysiology* 90: 1996–2004.

Minor LB, Lasker DM, Backous DD, and Hullar TE (1999) Horizontal vestibuloocular reflex evoked be high-acceleration rotations in the squirrel monkey I: Normal responses. *Journal of Neurophysiology* 82: 1254–1270.

Paige GD and Tomko DL (1991) Eye movement responses to linear head motion in the squirrel monkey II: Visual–vestibular interactions and kinematic considerations. *Journal of Neurophysiology* 65: 1183–1196.

Raphan T and Cohen B (1985) Velocity storage and the ocular response to multidimensional vestibular stimuli. In: Berthoz A and Melville Jones G (eds.) *Reviews of Oculomotor Research*, vol. 1, pp. 123–143. Oxford: Elsevier.

Robinson DA (1981) The use of control system analysis in the neurophysiology of eye movements. *Annual Review of Neuroscience* 4: 463–503.

Schwarz U, Busettini C, and Miles FA (1989) Ocular responses to linear motion are inversely proportional to viewing distance. *Science* 245: 1394–1396.

Snyder LH and King WM (1992) Effect of viewing distance and location of the axis of head rotation on the monkey's vestibuloocular reflex I: Eye movement responses. *Journal of Neurophysiology* 67: 681–874.

Waespe W, Cohen B, and Raphan T (1985) Dynamic modification of the vestibule-ocular reflex by the nodulus and uvula. *Science* 228: 199–202.

Yang D, Fitzgibbon EJ, and Miles FA (1999) Short-latency vergence eye movements induced by radial optic flow in humans: Dependence on ambient vergence level. *Journal of Neurophysiology* 81: 945–949.

Zhou W, Weldo P, Tang B, and King WM (2003) Rapid motor learning in the translational vestibulo-ocular reflex. *Journal of Neuroscience* 23(10): 4288–4298.

Vestibulospinal System and Eye–Head/Neck Movement

B D Corneil, University of Western Ontario, London, ON, Canada
S Musallam, McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The vestibular system detects the motion and orientation of the head in space. Although vestibular sensation is typically not perceived consciously, anyone who has experienced a vestibular pathology can attest to how critical this sense is for our daily life. Anatomical specializations of the vestibular end organs within the inner ear enable the transduction of both angular (via the semicircular canals) and translational (via the otolith organs) accelerations, permitting the central nervous system (CNS) to sense head motion in all six degrees of freedom. Information detected by the vestibular system initiates a series of reflexes that can stabilize our visual axis during running or walking via the vestibulo-ocular reflex and maintain head and body posture via vestibulospinal pathways. For example, when we accidentally slip and begin to fall forward, a series of reflexes is initiated that maintains our center of gravity and protects our head and body from injury.

The maintenance of head and body posture requires the development of muscular forces at multiple body segments. Accordingly, the vestibulospinal pathways distribute vestibular information to all levels of the spinal cord. This article focuses primarily on vestibular projections to the cervical spinal cord, and it is likely that the thoracic and lumbar projections share many of the same organizing principles. This article also examines the neuromuscular control of other types of head movements, such as orienting head movements that help redirect the line of sight. Voluntary movement of the head requires the contextual suppression of the reflexive mechanisms that serve to stabilize the head. From an overall perspective, many of the challenges faced by the CNS in controlling head motion (e.g., contextual gating of reflexes, control of a highly complex and seemingly redundant neuromuscular plant) mirror those present in other motor systems, such as the limb.

Anatomical Organization of Vestibulospinal Pathways

Neural information transduced by the vestibular end organs is transmitted to the four vestibular nuclei located in the rostral medulla and caudal pons. Afferents from the semicircular canals project primarily to the superior and medial vestibular nuclei, although minor projections reach the lateral and inferior vestibular nuclei. The otolith organs project primarily to the inferior vestibular nucleus, and projections to the lateral, medial, and superior vestibular nuclei have been found in a variety of species. Much of the processing of positional and movement information necessary for vestibular reflexes takes place within the vestibular nuclei. For example, many neurons within the vestibular nuclei receive converging information from both otolith receptors and one or more canal pairs. As a result, individual vestibular neurons can encode both angular and translational components of head motion. Such convergence can be used to disambiguate tilts and translations. In addition, neurons within the vestibular nuclei also receive major afferent projections from the cerebellum, the accessory optic system, and the contralateral vestibular nuclei, as well as proprioceptive signals arising from all levels of the spinal cord and from the reticular formation. These converging, and occasionally reciprocal, projections provide the vestibular nuclei with the information necessary to integrate the incoming vestibular information with the visual, proprioceptive, and contextual cues necessary for the production of an appropriate motor response.

Efferent projections from the vestibular nuclei distribute information about head motion and head position with respect to gravity to many different parts of the brain. Major targets include oculomotor nuclei, reticular and spinal centers concerned with whole-body motion, the vestibular regions of the cerebellum, the thalamus, and also the contralateral vestibular nuclei via commissural connections. These commissural fibers provide the pathways by which information from pairs of corresponding semicircular canals and otolith organs can be integrated.

The vestibular system produces reflexive postural adjustments of the head and body through two major descending pathways to the spinal cord: the lateral vestibulospinal tract (LVST) and the medial vestibulospinal tract (MVST; Figure 1). The LVST arises primarily from neurons within the lateral (Deiter's nucleus) and inferior vestibular nuclei. The LVST projects to all levels of the ipsilateral spinal cord, running within the anterior funiculus of the spinal cord before terminating directly onto both motor neurons and interneurons within lamina VII to IX of the cervical, thoracic, and lumbar spinal cord. The MVST originates primarily from the medial vestibular nucleus, although lesser projections arise from the inferior and lateral nuclei. In contrast to the LVST, fibers of the MVST project bilaterally, terminating in

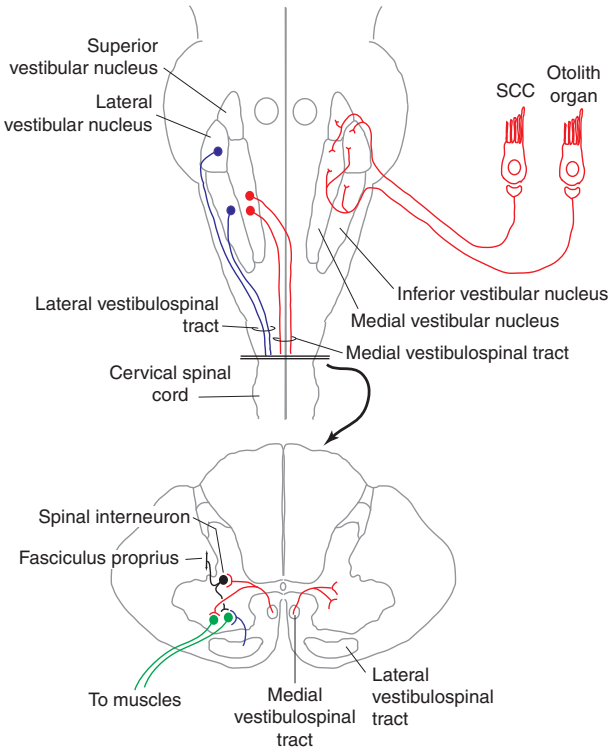

Figure 1 Projection of vestibular afferents to vestibular nuclei, and trajectory of medial and lateral vestibulospinal tracts.

lamina VII to IX of the cervical spinal cord, where they contact interneurons, motor neurons, and propriospinal neurons.

In addition to their contributions to the vestibulospinal tracts, neurons within the vestibular nuclei send branches to wide areas of the reticular formation, ranging from the medulla to the midbrain, and to the interstitial nucleus of Cajal and the Forel's Field H (FFH). These areas in turn have projections to the spinal cord that include direct connections to motor neurons. These indirect pathways thus form a parallel channel by which vestibular information can contribute to vestibulospinal reflexes and provide another means by which vestibular information can be integrated with other sensory, motor, and contextual signals.

Functional Physiology of Vestibulospinal System

The LVST and MVST primarily mediate different components of the vestibulospinal reflexes. For example, the actions initiated by signals carried along the LVST are primarily carried out by the proximal muscles of the body, including the muscles of the limb. Axons of many LVST neurons give off collaterals in many different segments of the spinal cord, ensuring the coordination of different muscle groups during postural control. In contrast, projections from the MVST primarily mediate the vestibulocollic reflex (VCR) and accordingly project mainly to the axial muscles of the neck. The projections of vestibulospinal neurons are patterned depending on their source of input. For example, MVST neurons receiving inputs from the saccule and/or the canals terminate predominantly, but not exclusively, at cervical spinal segments. Most neurons of the LVST receive utricular input, and ~20% of these have axons that reach the lumbar segments of the spinal cord. Finally, vestibulospinal neurons that contribute to the MVST and receive input from the vertical canals predominantly innervate upper cervical spinal cord segments. These findings are consistent with behavioral observations in humans that the VCR is more active in the pitch than the yaw plane.

The most basic circuitry for the VCR is a three-neuron arc similar to the circuitry of the vestibulo-ocular reflex. The three-neuron VCR consists of a primary vestibular afferent, a vestibulocollic neuron, and a neck muscle motor neuron. Returning to our example of the person falling forward, the integration of signals mainly from the saccule and vertical semicircular canals within the vestibular nuclei initiates a reflex that results (via the MVST) in the recruitment of dorsal neck extensor muscles and the inhibition of ventral neck flexor muscles. The result is an upward head motion, opposite to the direction in which the person is falling. By distributing signals to multiple levels of the spinal cord, the vestibulospinal reflexes carried along the LVST and MVST coordinate an appropriate whole-body response that serves to protect the head and body from impact and stabilize the body's center of gravity.

As will be described in the section titled 'Biomechanical comparisons of eye and head movement,' the anatomical organization of the muscles of the neck is highly complex. There are more than 20 pairs of neck muscles that can potentially generate head motion through a structure (the neck) that spans multiple individually mobile cervical segments. Recent work, primarily carried out in cats, has demonstrated that the CNS recruits a consistent synergy of neck muscle activation to generate the VCR. For example, signals from each semicircular canal are distributed to an appropriate set of neck muscles to produce compensatory head movements in the same plane as the plane of the stimulated canal (Figure 2). This mapping is done by single MVST neurons that receive a particular combination of canal input and then innervate a functional set of multiple neck muscles. In this way, they implement a canal-dependent head movement synergy. Additional evidence shows

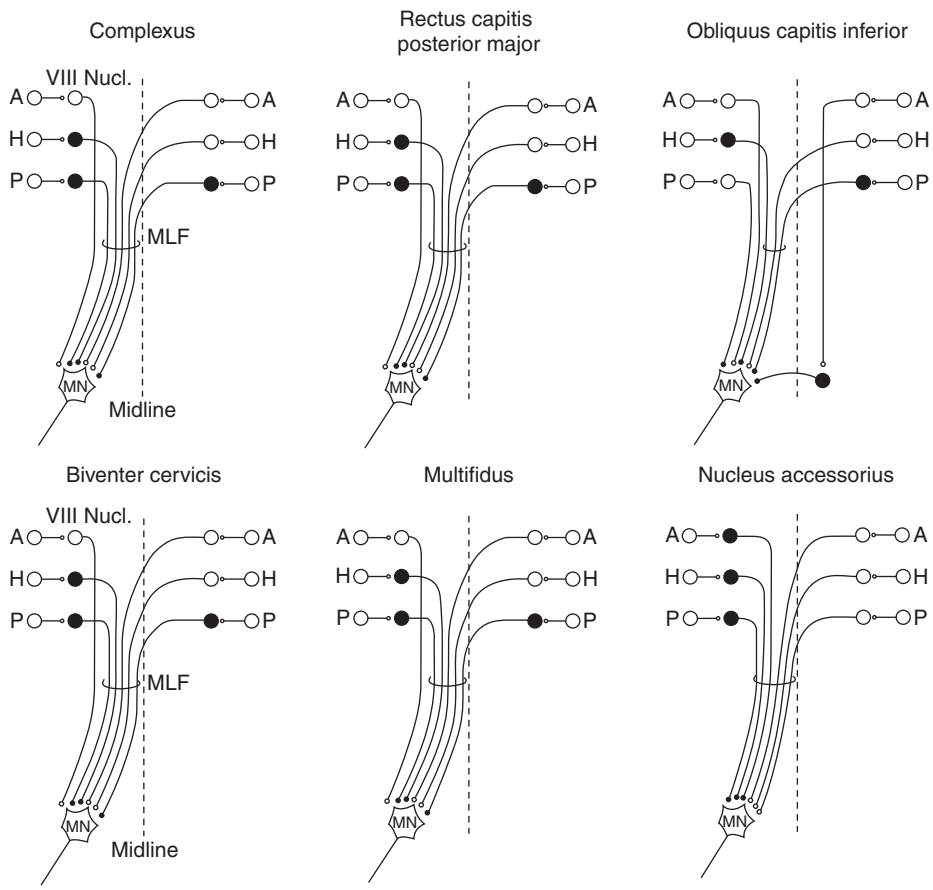

Figure 2 Convergent input patterns and pathways from the six semicircular canals to motor neurons supplying six different neck muscles. Open and closed circles represent excitatory and inhibitory neurons, respectively. Note how the motor neuron innervating a given muscle receives a unique and homogeneous pattern of input from the six semicircular canals. A, anterior; H, horizontal; P, posterior canals; MN, neck motor neuron; MLF, medial longitudinal fasciculus; Nucl., nucleus. Reproduced from Sugiuchi Y, Kakei S, Izawa Y, and Shinoda Y (2004) Functional synergies among neck muscles revealed by branching patterns of single long descending motor-tract axons. *Progress in Brain Research* 143: 411–421, with permission from Elsevier.

that the 'hardwired' functional synergy contacted by single MVST neurons may also be used by the reticulospinal and tectoreticulospinal systems. The advantage of this organization may be twofold. First, the use of a consistent set of muscular synergies by descending MVST neurons simplifies the control of head motion by decreasing the degrees of freedom about which head motion can be generated. Second, sharing of the same muscular synergy across different descending systems may aid the contextual control of voluntary versus externally applied head motion. The generation of voluntary head movement, for example, through the tectoreticulospinal system, requires the transient suppression of the VCR, which would otherwise be counterproductive. Hardwiring the vestibulospinal and tectoreticulospinal system into the same muscular synergies may optimize the neuromuscular responses to voluntary versus externally applied head motion.

Other Descending Pathways Involved in Controlling Head Motion

The vestibulospinal tracts are one of a number of descending pathways that can control head movement and posture. Head motion can also be generated via reticulospinal, tectospinal or tectoreticulospinal, interstitospinal, and corticospinal pathways that exert their influence via direct connections onto neck muscle motor neurons or via indirect connections through intercalated reticulospinal neurons or spinal interneurons. The convergence of multiple descending pathways onto neck muscle motor neurons reflects the diversity of function that head movements may fulfill. In addition to contributing to body posture and balance via the vestibulospinal tracts, and depending on the species in question, head movements can serve such diverse behaviors as orienting, feeding, grooming, communicating, fighting or other

predator behaviors, and even prehension (seen, for example, when a cat carries a kitten). Each of these tasks is associated with a particular set of control criteria that are likely effected through the combined actions of multiple descending systems.

Much has been learned recently about the neural mechanisms controlling head motion during gaze shifts that redirect the line of sight. Because the eyes are effectively blind during gaze shifts, the CNS places a premium on shifting gaze as quickly and as accurately as possible. Although smaller gaze shifts can typically be generated by rapid saccadic eye movements within the head, gaze shifts greater than $\sim 20^\circ$ usually involve coordinated motion of the eyes within the head and the head within space. The control of large gaze shifts represents an instance of motor control wherein the movement of the variable in question (here, the line of sight) depends on the nested movement of multiple body parts.

A key structure in the control of gaze shifts is the superior colliculus (SC). Recent evidence has demonstrated that neural activity within the SC encodes the desired gaze shift rather than the underlying component movements of either the eyes or the head. Thus, the decomposition of this gaze-related command emanating from the SC into the appropriate motor commands to move both the eyes and the head occurs within the downstream circuitry. Precisely how this is done remains an open question in oculomotor research. Anatomically, the signal output from the SC is transmitted to neck muscle motor neurons via two major separate pathways: one through pontomedullary reticulospinal neurons for horizontal head motion and the other through neurons in FFH in the mesodiencephalic junction for vertical head motion. In both the pontomedullary reticular formation and the FFH, two subgroups of spinal-projecting neurons are found. The descending axon of the first subgroup gives off a number of collaterals which arborize into premotor centers controlling eye motion (Figure 3(a)). In contrast, no such collaterals are found for the second subgroup (Figure 3(b)). Presumably, these subgroups allow for both coordinated and separate control of eye and head movements. Both the pontomedullary reticular formation and the FFH receive inputs from higher cortical areas, likely providing a substrate by which voluntary control of eye–head coupling may be generated.

Biomechanical Comparisons of Eye and Head Movement

As the above results demonstrate, the neural control of the eyes and head can be intimately related. Similar neural substrates for shared control of the eye and head have also been demonstrated in the vestibular system, wherein some second-order vestibular nuclei project to both eye and head premotor centers. A more complete understanding of the functional role of both vestibular or reticulospinal eye–head neurons requires neurophysiological investigations in awake, behaving animal subjects (usually cats or monkeys). However, investigation of the functional role of eye–head neurons in motor control is hindered by the very different biomechanics of the eye and the head. Comparatively, the eye is the far simpler of the two structures. Eye movements are generated by six extraocular muscles which operate on a motor plant dominated by viscous and elastic resistive forces. More than 40 years of study on oculomotor control with the head restrained has yielded a refined understanding of the neuromuscular events that cause eye motion, to the point that the eye movements themselves can serve as a proxy for the underlying neuromuscular control. For example, an eye movement of a certain amplitude and direction allows one to infer, with a fairly high degree of precision, the underlying spatial (i.e., the particular muscles) and temporal patterns of muscle activity that generated the eye movement.

The biomechanical and musculoskeletal realities of head motion do not permit a similar level of inference. Biomechanically, the head can be viewed as an inverted pendulum connected to a tilted beam, as in humans and most quadrupeds the head is held at the top of a more or less vertically oriented cervical column. The center of mass of the head lies anterior to the vertebral column, such that the head would fall forward were all head–neck muscles to relax. The head has a considerable mass (usually about 4–5 kg), and accordingly is endowed with substantial inertia (approximately $0.02\,\mathrm{kg\,m^2}$ for axial/yaw rotation and 0.03–$0.04\,\mathrm{kg\,m^2}$ for frontal/pitch and lateral/roll flexion). Anatomically, head movements in humans and animals can be generated by the recruitment of more than 20 pairs of neck muscles that interconnect the skull, cervical vertebrae, and shoulder girdle. Many of these neck muscles cross two or more joints, span multiple sites of attachment, and are divided into multiple compartments by architectural specializations. The 20 pairs of neck muscles also vary substantially in size and force-generating capacity. The largest muscles tend to be located in the most superficial layers of the neck and, as a consequence of their large cross-sectional areas and lever arm, have the greatest capacity to generate torque. In contrast, the deeper layers of the neck are invested with much smaller muscles that tend to have simpler anatomical structures. The muscles in Figure 4, which are taken from the dorsal neck of the monkey, depict

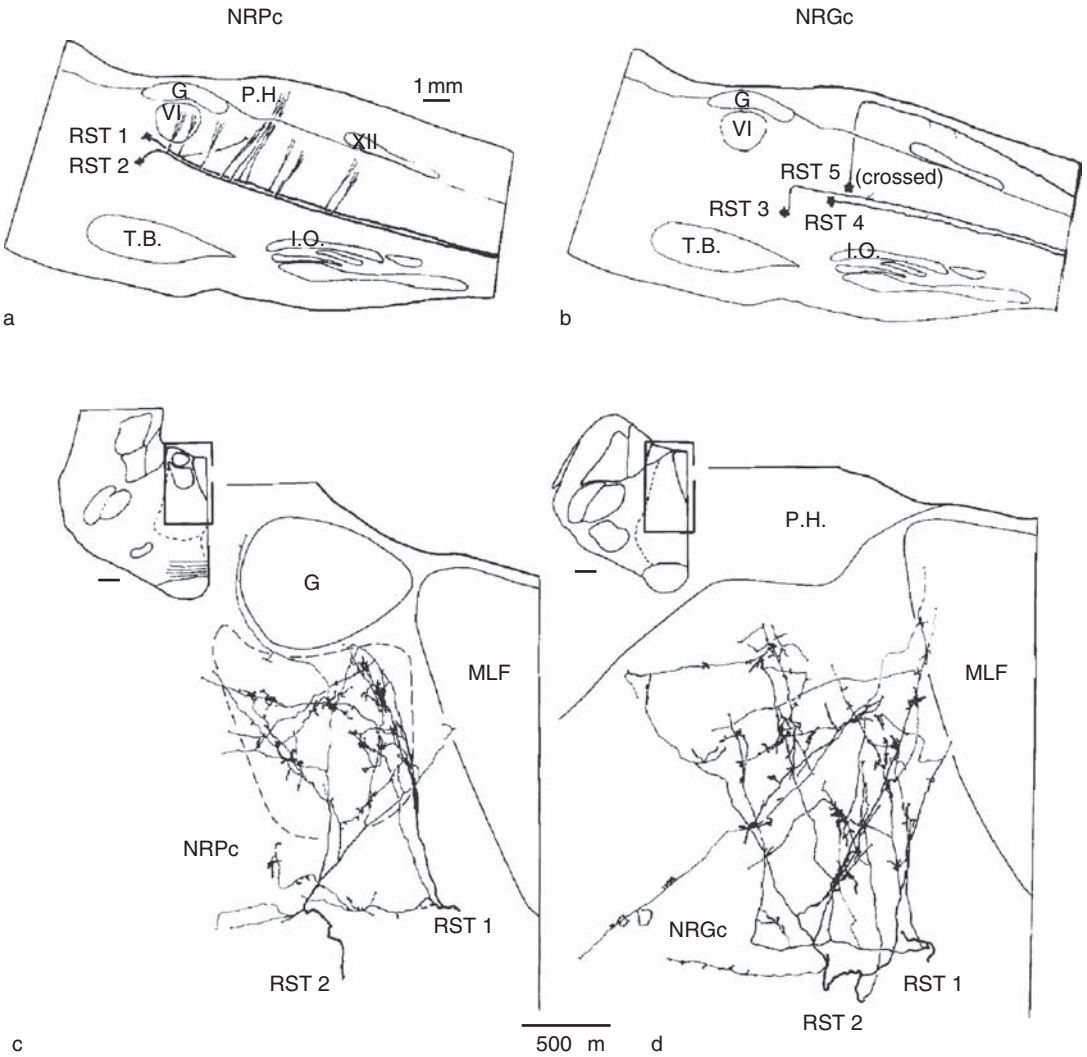

Figure 3 Axonal trajectories of single reticulospinal tract neurons (RSTs) in the pontomedullary reticular formation, revealed by intraaxonal staining with horseradish peroxidase. (a, b) Low-magnification views of axonal trajectories of two RSTs from the nucleus reticularis pontis caudalis (NRPc; (a)) and three RSTs from the nucleus reticularis gigantocellularis (NRGc; (b)). Note that the axons of the two NRPc–RSTs give off axon collaterals into the pontomedullary reticular formation on their way to the spinal cord, while such collaterals were rare in the three NRGc–RSTs. (c), (d) High-magnification view of reconstructed axonal trajectories of the two NRPc–RSTs, at the level of the genu (G) of the facial nerve (c) or the NRGc (d), as shown in the inset, showing collateralization into the abducens nerve (c) and NRGc. MLF, medial longitudinal fasciculus. Reproduced from Isa T and Sasaki S (2002) Brainstem control of head movements during orienting: Organization of the premotor circuits. *Progress in Neurobiology* 66: 205–241, with permission from Elsevier.

a number of superficial and deep neck muscles that would, based simply on their gross anatomy, appear to have similar pulling directions. If one thinks only about a single motor task, then the head–neck motor system would seem to contain a number of such muscular redundancies. However, the adaptations of particular muscles reflect functional optimizations for the full gamut of head movement tasks: muscles which appear to be redundant in one task may not be so in another.

Such considerations of neck muscle biomechanics and anatomy demonstrate why it is difficult, if not impossible, to infer the spatial and temporal patterns of neck muscle recruitment based simply on the kinematics of a head movement. Even a detailed knowledge of neck muscle morphometry, including descriptions of fiber type, cross-sectional area, tendon arrangements, and pinnation angle, is often insufficient to identify the functional role played by a neck muscle in a given head movement. Furthermore, patterns of muscle recruitment will depend on a number of other factors as well, including the recruitment of other muscles. Most of the muscles of the neck are attached at both ends to bones that move, so their action will depend on the stability and position of the origin and insertion of the muscle.

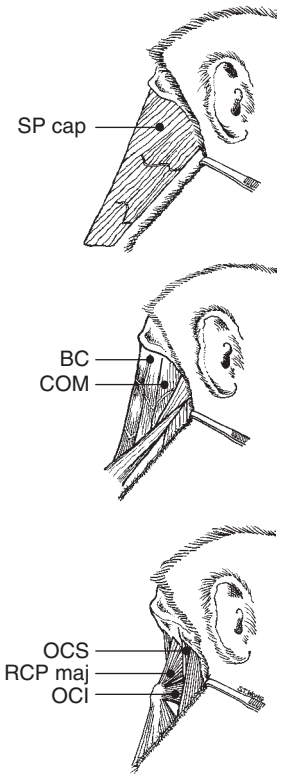

Figure 4 Anatomical sketches of some of the muscles of the dorsal neck in the monkey, illustrating the apparent redundancy in the musculoskeletal organization of the head and neck. Electromyograph recordings demonstrate that splenius capitis (SP cap), rectus capitis posterior major (RCP maj), and obliquus capitis inferior (OCI) contribute to a horizontal head turning synergy. Biventer cervicis (BC), complexus (COM), and obliquus capitis superior (OCS) contribute to a vertical head movement synergy that tends to pitch the head upward.

Electromyographic Recordings of Neck Muscle Recruitment

Chronically indwelling electromyogram (EMG) electrodes have been implanted in both superficial and deep neck muscles in cats and monkeys, permitting the characterization of the patterns of neck muscle recruitment during a variety of head movements and postures. The advantage of this approach is that it permits the direct observation of the neuromuscular signals issued to the head plant with a high degree of both spatial and temporal precision, circumventing the biomechanical complexities inherent in head motion. This work has revealed a number of findings. First, all head movements and postures are associated with recruitment of the smaller, deeper muscles of the neck. Recruitment of the more superficial neck muscles is added to this synergy for more eccentric postures or for larger and faster head movements. Second, the initial posture adopted by the animal influences both the timing and the pattern of neck muscle recruitment. For example, cats can be trained to generate head turns in both their customary posture, with the cervical column oriented vertically, and a second posture, with the cervical column oriented more horizontally. In both postures, a posturally independent set of neck muscles is recruited, perhaps reflecting connections hardwired through the tectoreticulospinal system. In the more horizontal posture, this basic synergy acts in concert with a second set of ancillary neck muscles, reflecting posturally dependent recruitment that presumably regulates the mobility of the cervical column and adjusts the net muscular forces. Such a strategy reduces the redundancy of the head–neck system and may also simplify the computation of appropriate muscular synergies for different behavioral tasks.

Recordings of neck EMG can also be combined with neurophysiological and behavioral approaches to gain additional insights into the brain stem and cortical control of head motion which could not be made simply through the kinematic analysis of eye–head gaze shifts. For example, recent work combining neck EMG recordings with stimulation of both the SC and the frontal eye fields has revealed that the current thresholds required to evoke neck EMG responses are frequently less than the current thresholds required to evoke eye–head gaze shifts. A second pertinent observation was reached by examination of the patterns of neck muscle recruitment following the onset of a visual target. Figure 5 shows the activity of a number of muscles that contribute to head turns. Even though the head is restrained, these muscles are far more active when a visual target is presented on the ipsilateral side (i.e., leftward targets for left muscles). Of particular interest is how the onset of recruitment of these muscles occurs ~90 ms after ipsilateral target presentation, regardless of the ensuing gaze shift reaction time (asterisk in Figure 5). At the extreme, muscles can be activated up to 150 ms in advance of gaze shift onset. The implication of these findings is that circuits in the brain stem, presumably downstream from the SC, permit the delivery of an orienting command to the head while a decision to commit to a gaze shift is still ongoing. Such selective gating may, at least in part, be due to the paramount importance of foveal vision. Alternatively, or perhaps additionally, such selective gating may help prime neck muscles earlier to compensate for the head's considerable inertia.

Afferent Information from Neck Muscles

A unique feature of the head–neck system is that it provides the linkage between the body and a mobile

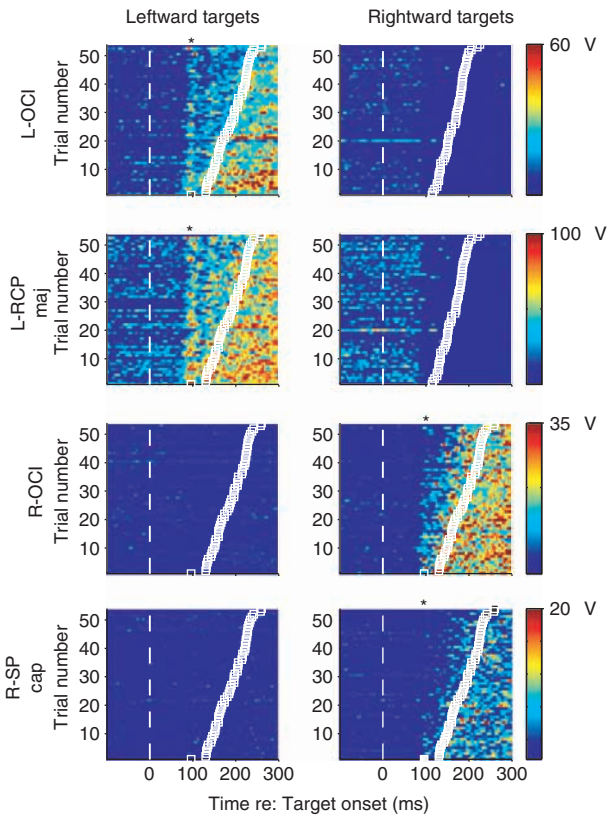

Figure 5 Activity of multiple neck muscles in a head-restrained monkey following the presentation of a visual target either 30° to the left (left column) or right (right column). Each subplot shows the magnitude of electromyograph (EMG) activity (in color) for ~50 trials (individual trials are stacked along the *y*-axis), with time proceeding along the *x*-axis and aligned on the time of target onset (white vertical dashed line). Trials have been ordered by the reaction time of the ensuing saccade (white squares). Note the band of EMG activity, depicted by the asterisks, that occurs ~90 ms after ipsilateral target presentation (e.g., left targets for left muscles), regardless of the time of the ensuing saccade. Reproduced from Corneil BD, Oliver E, and Munoz DP (2004) Visual responses on neck muscles reveal selective gating that prevents express saccades. *Neuron* 42(5): 831–841, with permission from Elsevier.

platform (the head) in which the major exteroceptive sense organs are mounted (e.g., the eyes, ears, and nose). Consequently, the position of the head relative to the body must be integrated within many sensorimotor transformations. Neck muscles are not simply motor organs but also contribute afferent information relating the position of the head relative to the body. Indeed, a striking observation is that the density of muscle spindles in the deeper muscles of the neck is among the highest in the body, comparable only to the fine muscles of the finger. Despite the rich endowment of muscle spindles in the neck, the segmental spinal reflexes (e.g., the monosynaptic stretch reflex) are paradoxically weak.

The afferent information from the neck likely serves two general roles for sensorimotor control. First, afferent information from the neck plays an integral role in the control of body posture through its integration with vestibular signals. Transient loss of neck muscle afferent information induces debilitating effects that resemble vestibular pathologies, affecting not only head movements but also posture, locomotion, and reaching. Symptoms of cervical vertigo, including dizziness, ataxia, and nystagmus, have long been recognized to follow whiplash and other types of neck muscle damage. Such symptoms can be reproduced experimentally when neck muscle receptors are anesthetized or stimulated by vibrating the muscles of the neck. In the normal human and animal, the short-latency cervical reflexes (e.g., the cervico-ocular and cervicocollic reflex) are thought to operate at a relatively low gain, but the convergence of neck muscle afferent information with visual and vestibular information in the cerebellum, vestibular nuclei, reticular formation, and vestibular cortex emphasizes its role in the multisensory control of body posture. The gains of the cervico-ocular and cervicocollic reflexes increase following vestibular trauma or with aging, presumably through the recruitment of plastic mechanisms that compensate for the compromised or decreasing vestibular sensation.

A second role of neck proprioception is to unequivocally signal the position of the head relative to the body for developing motor programs. Knowledge of the current position of a body segment is a prerequisite to moving it accurately. Neck muscle afferent information reaches higher oculomotor centers in the parietal and frontal cortices, where it appears to be integrated into developing plans for action. Although convergence of vestibular and neck muscle afferent information occurs in some cortical areas, in others the information remains segregated. From a motor control standpoint, it is imperative that the organism knows both the position of its head relative to its body and the position of its head in space.

Summary

The vestibulospinal system initiates a series of whole-body reflexes whose function is to stabilize body posture and protect the head and body from potential injury. VCRs target the axial muscles of the neck by recruiting hardwired muscle synergies to produce compensatory head movements. Such hardwired synergies, which are also present in other descending head control pathways, are necessary to simplify the control of a motor structure as complicated as the head. Head movements are also intimately related to eye movements, fulfilling behavioral goals to stabilize or redirect the line of sight. Finally, neck muscles are not simply motor organs but also provide a

refined sense of head-on-body position necessary for postural control and sensorimotor integration. Overall, an appreciation of the neural strategies used by the CNS to control head motion may reveal principles which generalize to other aspects of motor control. Like the limb (and unlike the eye), the head is a highly complex motor structure characterized by significant inertia, and proprioceptive feedback from the muscles must be integrated into places for action. However, unlike the limb, much of the neural circuitry responsible for the sensorimotor control of head movement lies in the brain stem and is accessible to contemporary recording techniques.

See also: Balance and Posture Control; Balance and Posture Control: Human; Cerebellar Lesions and Effects on Posture, Locomotion and Limb Movement; Eye and Head Movements; Proprioception; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Superior Colliculus; Vestibular System; Vestibulo-Autonomic Responses; Vestibulo-Ocular Reflex; Visual–Vestibular Interactions.

Further Reading

Angelaki DE, Shaikh AG, Green AM, and Dickman JD (2004) Neurons compute internal models of the physical laws of motion. *Nature* 430: 560–564.

Bakker DA and Richmond FJ (1982) Muscle spindle complexes in muscles around upper cervical vertebrae in the cat. *Journal of Neurophysiology* 48: 62–74.

Barmack NH (2003) Central vestibular system: Vestibular nuclei and posterior cerebellum. *Brain Research Bulletin* 60: 511–541.

Berthoz A, Graf W, and Vidal PP (1992) *The Head–Neck Sensory Motor System*. New York: Oxford University Press.

Corneil BD, Olivier E, and Munoz DP (2004) Visual responses on neck muscles reveal selective gating that prevents express saccades. *Neuron* 42: 831–841.

Cullen KE and Roy JE (2004) Signal processing in the vestibular system during active versus passive head movements. *Journal of Neurophysiology* 91: 1919–1933.

Isa T and Sasaki S (2002) Brainstem control of head movements during orienting: Organization of the premotor circuits. *Progress in Neurobiology* 66: 205–241.

Newlands SD, Vrabec JT, Purcell IM, et al. (2003) Central projections of the saccular and utricular nerves in macaques. *Journal of Comparative Neurology* 466: 31–47.

Peterson BW and Richmond FJ (1988) *Control of Head Movement*. New York: Oxford University Press.

Richmond FJ, Singh K, and Corneil BD (2001) Neck muscles in the rhesus monkey. I. Muscle morphometry and histochemistry. *Journal of Neurophysiology* 86: 1717–1728.

Sato H, Endo K, Ikegami H, et al. (1996) Properties of utricular nerve-activated vestibulospinal neurons in cats. *Experimental Brain Research* 112: 197–202.

Sigiuchi Y, Kakei S, Izawa Y, and Shinoda Y (2004) Functional synergies among neck muscles revealed by branching patterns of single long descending motor-tract axons. *Progress in Brain Research* 143: 411–421.

Snyder LH, Grieve KL, Brotchie P, and Andersen RA (1998) Separate body- and world-referenced representations of visual space in parietal cortex. *Nature* 394: 887–891.

Thomson DB, Loeb GE, and Richmond FJ (1994) Effect of neck posture on the activation of feline neck muscles during voluntary head turns. *Journal of Neurophysiology* 72: 2004–2014.

Wilson VJ, Boyle R, Fukushima K, et al. (1995) The vestibulocollic reflex. *Journal of Vestibular Research* 5: 147–170.

Relevant Websites

http://vestibular.wustl.edu – Dickman Laboratory for Vestibular Function, Development & Regeneration.

http://www.physpharm.fmd.uwo.ca – The Physiology of the Senses: Transformations for Perception and Action, created by Tutis Vilis.

Vibrissa Movement, Sensation and Sensorimotor Control

D Kleinfeld, University of California at San Diego, La Jolla, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Rats use their vibrissae, an array of long, pliable hairs that can be actively swept through space, to interrogate objects in their local physical environment (Figures 1(a) and 1(b)). As rats search and locomote, they must be able to gauge the position of the ground beneath them and walls and obstacles around them. This suggests that the sense of touch is used in multiple ways. First, touch provides a means to establish the nature of a surface texture. Is it smooth or rough? Second, touch is used to establish the shape of an object. Third, touch is used as a means to determine the location of an object with respect to the body image of the rat. Is the object to the front or the side? The notion of interpreting touch in the context of body coordinates implies that reference signals of vibrissa position are an essential aspect of sensation.

The neuronal encoding of environmental and positional clues, and the anatomical circuitry that turns these clues into motor plans, form the material of this article. It begins by delineating a number of tasks that rats may be trained to accomplish solely through the use of their vibrissae. These serve to focus our presentation on neuronal pathways that are directly relevant to an understanding of vibrissa sensorimotor control. The article then discusses the anatomy of the vibrissa pathway in detail, with an emphasis on the mechanical properties of the vibrissa *per se*, as well as the motor plant that drives the vibrissae. This leads into a presentation of electrical signaling along the vibrissa pathways, followed by consideration of the neuronal computations that are supported by these signals. The article closes with a discussion of open issues. As an aid to reading the literature, common abbreviations are given in Table 1.

Behavior

Vibrissae somatosensation is an active sensory process. Rodents sweep their vibrissae through space largely in unison and with a rhythm that varies between 5 and 25 Hz. The dynamics of this process can be extracted from high-speed videography or from measurements of the electrical activity of the muscles that drive the vibrissae (Figure 1(c)). What tasks can animals accomplish with this active process? A basic function of the vibrissae is to evaluate the terrain in front of the animal. Thus, for example, blind rats with intact vibrissae will leap across wider gaps than rats without vibrissae. This observation is a building block of a commonly used paradigm to determine whether rats can discriminate among surfaces with different textures (Figure 2). Here, rats foveal whisk as they crane across each of two gaps (Figure 1(c)). This posture allows them to sense and evaluate a texture on the far side of the gap. Only one of the two textures is paired with a reward, that is, water or food, which the animal can receive by jumping across the gap. The accumulated data from these experiments suggest that trained animals may recognize a corrugated surface whose pitch is as fine as 15μm and differentiate among surfaces whose relative difference in pitch is as small as 5%. Last, these experiments reveal that the amplitude of whisking may change upon contact (Figure 2(a)(2)) and thus provides a means to manipulate sensorimotor feedback.

The vibrissae convey spatial information that is sufficient to allow rats to distinguish between barriers whose lateral distance from the snout varies by less than 5%. In one task, rats are trained to distinguish between a wide versus narrow aperture (Figures 2(b)(1) and 2(b)(2)). A rat waits in a main chamber until a sliding door opens, after which it enters a small chamber that contains a nose poke, that is, a place for the snout that keeps the rat's head relatively fixed during the task. Once the snout is positioned, two panels close from the sides to form a symmetric, variable-width aperture. Animals are required to judge whether the aperture is the wider or narrower of two possible widths and report their choice by moving to additional nose pokes in the main chamber, that is, a right nose poke for the narrow aperture and a left nose poke for the wide aperture (Figure 2(b)(1)). The correct choice is rewarded. An interesting result is that, as the vibrissae are incrementally trimmed from the normal complement of 20 or more down to one and then zero, the performance on this task is a linearly decreasing function of the number of vibrissae that remain (Figure 2(b)(2)).

Variants of the above fine-discrimination task require the animal to report symmetric versus asymmetric apertures. In a related task, animals are trained to report the shift in the horizontal position between pins placed on each side of the face (Figure 2(c)(1)). This latter task probes the bilateral acuity of the vibrissa system. It is interesting that rats perform this task better with only a single column, referred to as arc, of vibrissae rather than all their vibrissae (Figure 2(c)(2)). The typical displacement that could be distinguished with an arc was 1.5 mm, which corresponds to an angular displacement of 6° in these experiments.

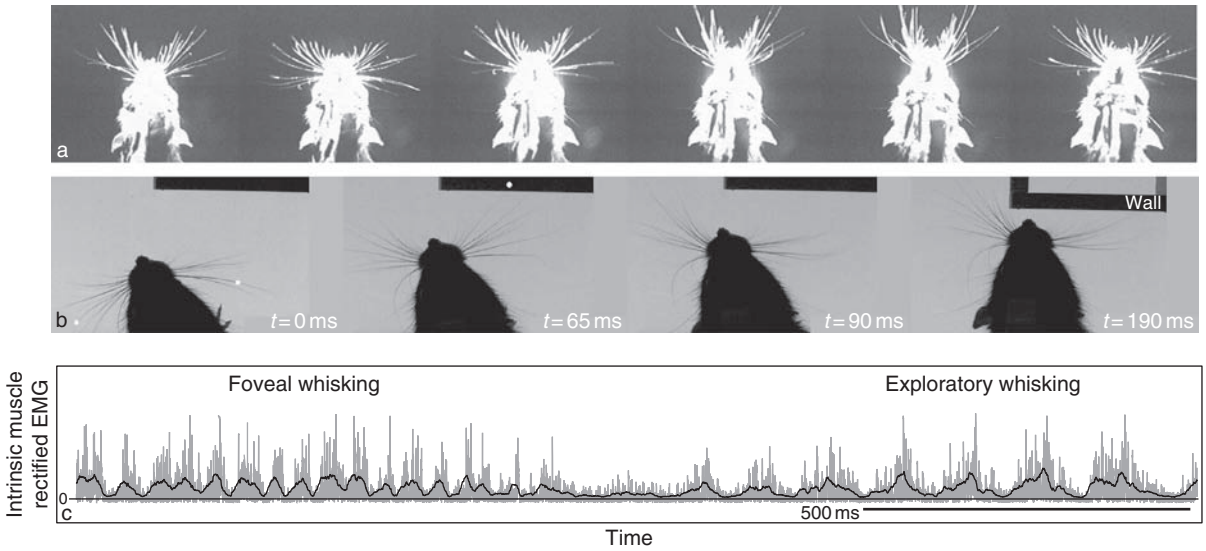

Figure 1 Example data for different modes of whisking. (a) Exploratory whisking, also referred to as symmetric whisking. Shown are successive strobed video images of a rat, using darkfield illumination, as it moves its vibrissae during an exploratory whisk cycle. The animal was trained, while blindfolded, to whisk in search of a food tube (not shown). Successive frames are at intervals of 33 ms. Unpublished data. (b) Asymmetric whisking as a result of contact between the vibrissae and a wall. Protraction commences approximately synchronously on both sides of the snout; the filled white squares show the tracked vibrissae in the rear arcs. At $t = 65$ ms, a deflection occurs on a forward vibrissa; the filled white circle indicates the point of contact with the vertical surface. Protraction then ends on the side ipsilateral to the contact, and the contralateral vibrissae reach maximum protraction in the whisk cycle subsequent to the initial contact. Adapted from Mitchinson B, Martin CJ, Grant RA, and Prescott TJ (2007) Feedback control in active sensing: Rat exploratory whisking is modulated by environmental contact. *Proceedings. Biological Sciences/The Royal Society* 274(1613): 1035–1041. A similar pattern of asymmetric whisking as animals change their head position has been reported by Hartmann (Towal RB and Hartmann MJ (2006) Right-left asymmetries in the whisking behavior of rats anticipate head movements. *Journal of Neuroscience* 26: 8838–8846). (c) Example of sequential foveal and exploratory whisking observed via the rectified electromyogram of the intrinsic muscles. The solid line is the low-pass filtered signal. Adapted from Berg RW and Kleinfeld D (2003) Rhythmic whisking by rat: Retraction as well as protraction of the vibrissae is under active muscular control. *Journal of Neurophysiology* 89: 104–117.

Beyond issues of discrimination, rats make use of vibrissa somatosensation to encode the location of objects relative to their face as they explore their environment with large-amplitude whisks (Figure 1(a)). In part, the ability of rats to use their vibrissae to sense a platform across a gap implies that they are sensitive both to touch and to the forward thrust of their vibrissae. Nonetheless, the question remains as to whether rodents encode the position of their vibrissae as they whisk (Figure 3(a)). This issue was addressed in terms of an operant conditioning paradigm in which rats were trained to discriminate between different angles of a pin relative to the area of the cheek that hosts the vibrissae, called the mystacial pad (Figure 3(b)). These animals were trained to press a lever at a high rate if the rewarded stimulus (S+), corresponding to a pin at a specified location, was presented. No reward occurred for a stimulus at a second location (S−). Rats with a single vibrissa were able to perform this task (Figure 3(c)). Further, they performed it without bias to vibrissa location and with few, relatively brief contacts. This implies that rats must make use of a reference signal of vibrissa position as they determine where touch occurs relative to their mystacial pad.

Whisking

Within laboratory settings, animals will typically whisk in bouts of five to 50 individual whisks with a continuum of behaviors. Nonetheless, three different motifs appear to be common. The first is exploratory whisking, which occurs when animals whisk in air without contact or with only light contact (Figure 1(a)). In this case the stroke on a given whisk approaches 70°, and a total field of up to 160° may be covered as the animal slowly shifts the set point of its whisk. Such whisking is extremely regular and typically occurs with a frequency between 7 and 12 Hz and a mean of 9 Hz. In particular, the bandwidth of exploratory whisking is at theoretical limits for a perfect oscillator, which implies that the motor drive may be described as a sinusoidal oscillation with a constant frequency.

The second whisking motif, asymmetric whisking, occurs when animals make contact with a large object, such as a wall, while they whisk (Figure 1(b)). It also occurs if they turn their head to the side while whisking. Asymmetric whisking lasts for only one to three whisk cycles. The final motif is referred to as foveal whisking and occurs when animals thrust all

Table 1 Common abbreviations

Abbreviation	*Meaning*
ALBSF	Anterior lateral barrel subfield, a region of parietal cortex
CPG	Central pattern generator
EMG	Electromyogram
FN	Facial nucleus
ICMS	Intracortical microstimulations
IoN	Infraorbital branch of the trigeminal nerve
L1, ..., L6	Neocortical layer 1, ..., layer 6
LFP	Local field potential
M1	Primary motor region of neocortex
MUA	Multiunit activity
PMBSF	Posterior medial barrel subfield, a region of parietal cortex
PO	Posterior medial nucleus, a region of dorsal thalamus (also POm)
PrV	Trigeminal nucleus principalis
RSU	Regular spiking unit
S1	Primary somatosensory region of neocortex (also SI)
S2	Secondary somatosensory region of neocortex (also SII)
SpVC	Spinal nucleus caudalis
SpVI	Spinal nucleus interpolaris
SpVO	Spinal nucleus oralis
TCU	Thalamocortical unit
TG	Trigeminal ganglion
TN	Trigeminal nucleus
VMCtx	Vibrissa primary motor cortex
VPM	Ventral posterior medial nucleus, a region of dorsal thalamus
VL	Ventral lateral nucleus, a region of dorsal thalamus
ZI	Zona incerta, a region of ventral thalamus

their vibrissae forward to palpate an object ahead of them, as occurs when they try to detect a landing on the far side of a gap (Figure 2(a)). In this case the stroke is much reduced, typically to 20°, and the extrinsic muscles, the muscles that pull on the mystacial pad and include *M. nasolabialis*, *M. maxolabialis*, *M. nasalis*, and *M. transversus nasi*, are relatively inactive as they can apply little torque when the mystacial pad is fully extended. The frequency of foveal whisking is high, ranging between 15 and 25 Hz, and rats can readily switch between foveal and exploratory whisking (Figure 1(c)).

Systems Description of the Sensorimotor Plant

The anatomy of the vibrissa system provides the foundation for understanding the neuronal basis of whisking and vibrissa-driven behaviors. Like other sensorimotor systems, the vibrissa system is organized as a set of nested loops, with the vibrissae and the motor plant that drives the vibrissae as the common focus of all loops (Figure 4). The ascending branches of the loop receive and process sensory input, and the descending branches can control the motor plant that drives the vibrissae. It is important to remember that, as a closed feedback system, both sensory and motor signals can appear at all levels and stages of the nervous system. Only in extreme cases, such as induced paralysis of an identified pathway, can one isolate a touch signal from a motion signal that leads to touch.

The Follicle

The front end of the vibrissa system is the vibrissa–follicle complex (Figure 5). The lower 4 mm of the vibrissa is buried in the follicle, which is innervated by two sets of trigeminal sensory nerve terminals that originate from the trigeminal nucleus. One set innervates the deep end of the follicle (DVN in Figure 5) and the other innervates the superficial end (SVN in Figure 5). Together, these nerve fibers form the infraorbital branch of the trigeminal nerve (IoN).

The sensory nerves contain eight morphologically distinct sensory terminals, for which two classes of responses appear to dominate. One is the rapidly adapting receptors, whose responses return to baseline within 200 ms of stimulation, and the other is the slowly adapting receptors, which enable the vibrissa to follow vibrations as fast as 1200 Hz over extended periods. The latter responses are believed to make use of small cells called Merkel cells, localized to the follicle as intermediates between the hair and the sensory nerve terminal. The follicles, although composed exclusively of soft tissue, contain cavities that are under the control of the autonomic nervous system and engorge with blood when the animal is active. This results in a fixture that is likely to form a stiff mechanical coupling between the vibrissa and the sensory nerve terminals.

The sensory fibers in the follicle are sensitive to motion of the embedded vibrissa and thus encode contact between an object and the vibrissa. More information is gleaned if the impact on contact and the distance of the contact point to the skin, that is, the radial distance along the vibrissa, can be encoded. One likely possibility is that the reaction forces along the follicle shift as the point of impact moves relative to the center of mass of the vibrissa. Thus the pattern of these forces can be used to gauge the magnitude, distance, and direction of impact.

The Mystacial Pad

The cheek on each side of the rat's face contains a thickening in the skin that is called the mystacial pad. The pad contains an array of follicles organized as five rows and four or more arcs. In addition, there are four additional large caudal vibrissae, denoted the

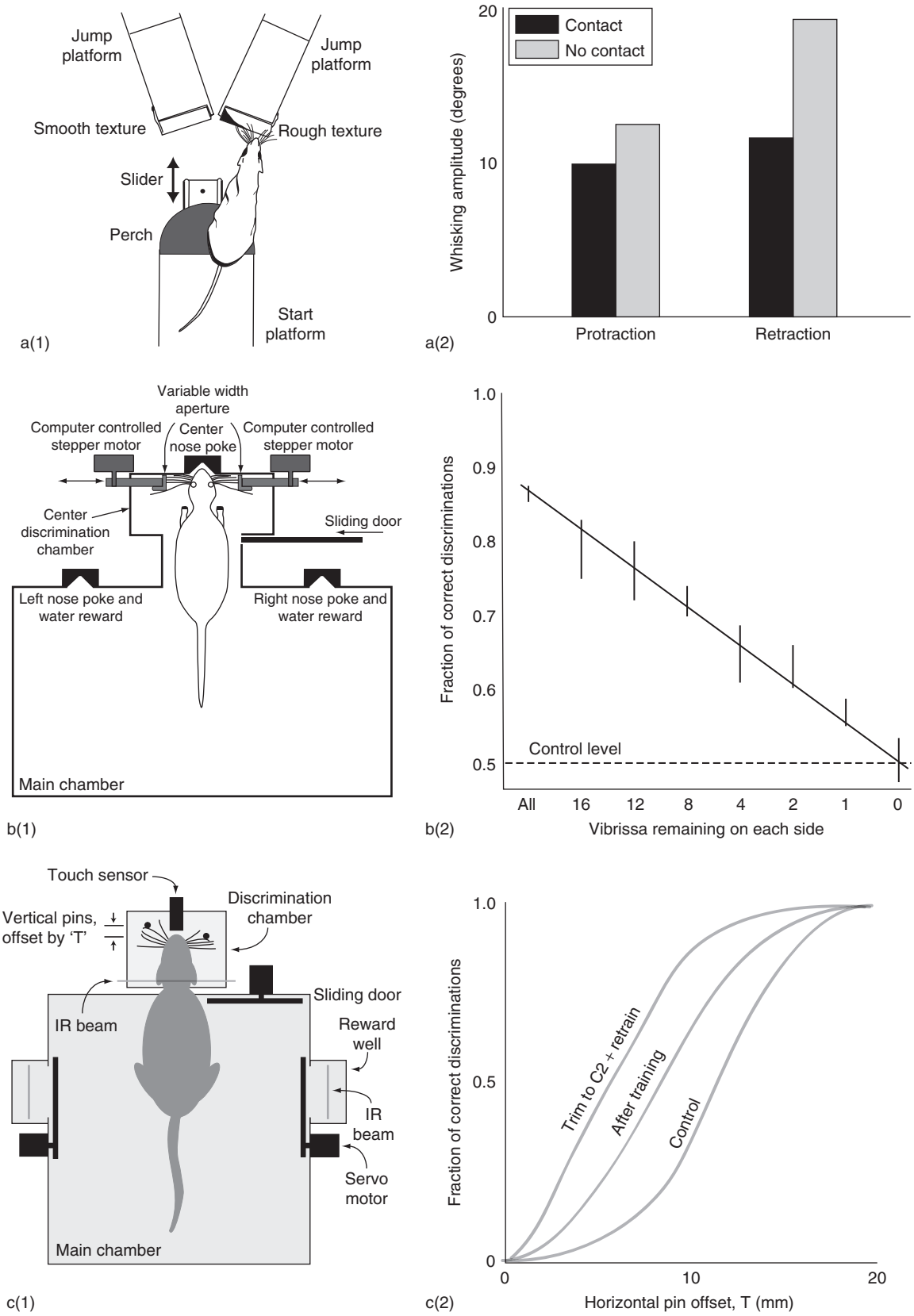

Figure 2 Behaviors that involve fine vibrissa-based comparisons. (a)(1) Schematic of a gap-crossing task in which rats crane across one of two equally spaced gaps to assess the relative roughness of textured tubes. Animals are trained to examine both textures and then jump from the perch on the start platform to the appropriate jump platform. The gap between the start and jump platforms can be adjusted with a slider to change the difficulty of the task. (a)(2) The effects of surface contact on the amplitude of vibrissa movements in one rat. The

straddlers, which interdigitate among the rows (insert, Figure 5). The hairs in the most dorsal row, that is, row A, point upward above the head and, curiously, are barbered as animals determine dominance. The hairs in the ventral row, that is, row E, tap the ground as the animal locomotes and presumably help maintain the altitude of the head. The middle rows, labeled B through D, appear important to exploration and the location of objects.

Brain Stem Loop

Sensory input from reaction forces generated in the follicles leads to a signal that transverses the IoN and projects to one or all four nuclei that form the trigeminal complex (Figure 4). These include the principal sensory nucleus (PrV) and the three spinal nuclei, denoted oralis (SpVO), interpolaris (SpVI), and caudalis (SpVC). The afferents form several somatotopic representations, referred to as barrelettes, of the ipsilateral vibrissae. Efferents from the PrV, SpVC, and SpVI nuclei project to motor neurons in the lateral subnucleus of the ipsilateral facial nucleus, which sends motor output to the muscles of the mystacial pad. This completes the lowest-order brain stem sensorimotor loop (∗ in Figure 4).

The trigeminal nuclei further interact among each other. Neurons in the PrV nucleus receive excitatory input from both the SpVC and SpVI nuclei and inhibitory input from the SpVi nucleus. The latter forms a local inhibitory loop that, possibly in concert with descending inputs from high-order areas, provides a means to filter sensory information at the level of the brain stem.

Muscles of the Mystacial Pad

What is the underlying musculature that drives the vibrissae? Protraction begins with contraction of the external protractor muscles in the skin, *M. nasalis* (Figure 6(a)). This shifts the center of mass of the follicle, as well as the point of egress of the vibrissae, toward the snout. A second group of muscles, which appear as slings that wrap around each follicle, contract next and rapidly propel the vibrissae forward. These muscles are unique to animals that whisk and are denoted the intrinsic muscles (Figure 6(b)). Last, as protraction reaches its maximum extent, activation of the sling muscles wanes and a pair of external retractor muscles, *M. nasolabialis* and *M. maxolabialis*, activate and restore the vibrissae to their initial position (Figure 6(a)). The measured sequence of muscle activation, together with the measured motion of the vibrissae along an anterior–posterior (A–P) axis, illustrates this process (Figure 6(c)). Last, it is possible to construct a mechanical model of the mystacial pad that is based on the known anatomy, elastic constants of the tissue, and force curves for the muscles (Figure 6(d)). The model serves as a means to reconstruct the position of the vibrissae from the measured electrical activation of the vibrissae.

Of critical importance, exploratory whisking occurs in the absence of sensory feedback. Moreover, the phase relation between the various muscle groups that mediate whisking is unchanged. This suggests the existence of coupled central pattern generators in the brain stem that drive the three phases of rhythmic whisking (Figure 6(c)).

angular displacement of different vibrissae are seen to decrease when they contact a texture as opposed to sweep the air without contact. Panels (a)(1) and (a)(2) adapted from Carvell GE and Simons DJ (1990) Biometric analyses of vibrissal tactile discrimination in the rat. *Journal of Neuroscience* 10: 2638–2648; see also Carvell GE and Simons DJ (1995) Task- and subject-related differences in sensorimotor behavior during active touch. *Somatosensory and Motor Research* 12: 1–9. A similar task was pioneered by Guic-Robles (Guic-Robles E, Valdivieso C, and Guajardo G (1989) *Behavioural Brain Research* 31: 285–289) and recently extended to mice (Celikel T and Sakmann B (2007) *Proceedings of the National Academy of Sciences of the United States of America* 104: 1395–1400). (b)(1) Apparatus for a task to test distance discrimination. The width of the aperture is adjusted between two values, 'narrow' or 'wide.' Differences of 1.5 mm, that is, 62 mm vs. 63.5 mm, were generally irresolvable, but differences of 3 mm were resolvable. At the start of each session, the rat is placed in the outer reward chamber with the sliding door closed. When the door is opened, the rat enters the center discrimination chamber, centers in the nose poke, and samples the aperture with its vibrissae. The rat then backs into the outer reward chamber and pokes its nose into either the left or right nose poke to receive a water reward: a left nose poke if the aperture is narrow and a right nose poke if the aperture is wide. (b)(2) Mean percentage of correct discriminations for an ensemble of rats in which vibrissae were sequentially and systematically removed. Note the decrease in performance with the total number of vibrissae removed. Panels (b)(1) and (b)(2) adapted from Krupa DJ, Matell MS, Brisben AJ, Oliveira LM, and Nicolelis MAL (2001) Behavioral properties of the trigeminal somatosensory system in rats performing whisker-dependent tactile discriminations. *Journal of Neuroscience* 21: 5752–5763; see also Shuler MG, Krupa DJ, and Nicolelis MA (2002) Integration of bilateral whisker stimuli in rats: Role of the whisker barrel cortices. *Cerebral Cortex* 12: 86–97. (c)(1) Apparatus for a task to test fine discrimination in horizontal offsets of two bilaterally spaced bars. The horizontal locations of the poles were changed between trials in steps of 0.1 mm. At the start of each session, the rat enters the center discrimination chamber, centers in the nose poke, and samples the poles with its vibrissae. The rat then backs into the outer reward chamber and pokes its nose into either the left or the right reward cell to receive a water reward; a specific offset is randomly paired with a specific side for each rat. (c)(2) Distribution of performance thresholds after discrimination training with the complete C row of vibrissae, after trimming down to one vibrissa, C2, and retraining, and for control animals with no vibrissae. (c)(1) and (c)(2) adapted from Knutsen PM, Pietr M, and Ahissar E (2006) Haptic object localization in the vibrissal system: Behavior and performance. *Journal of Neuroscience* 26: 8451–8464.

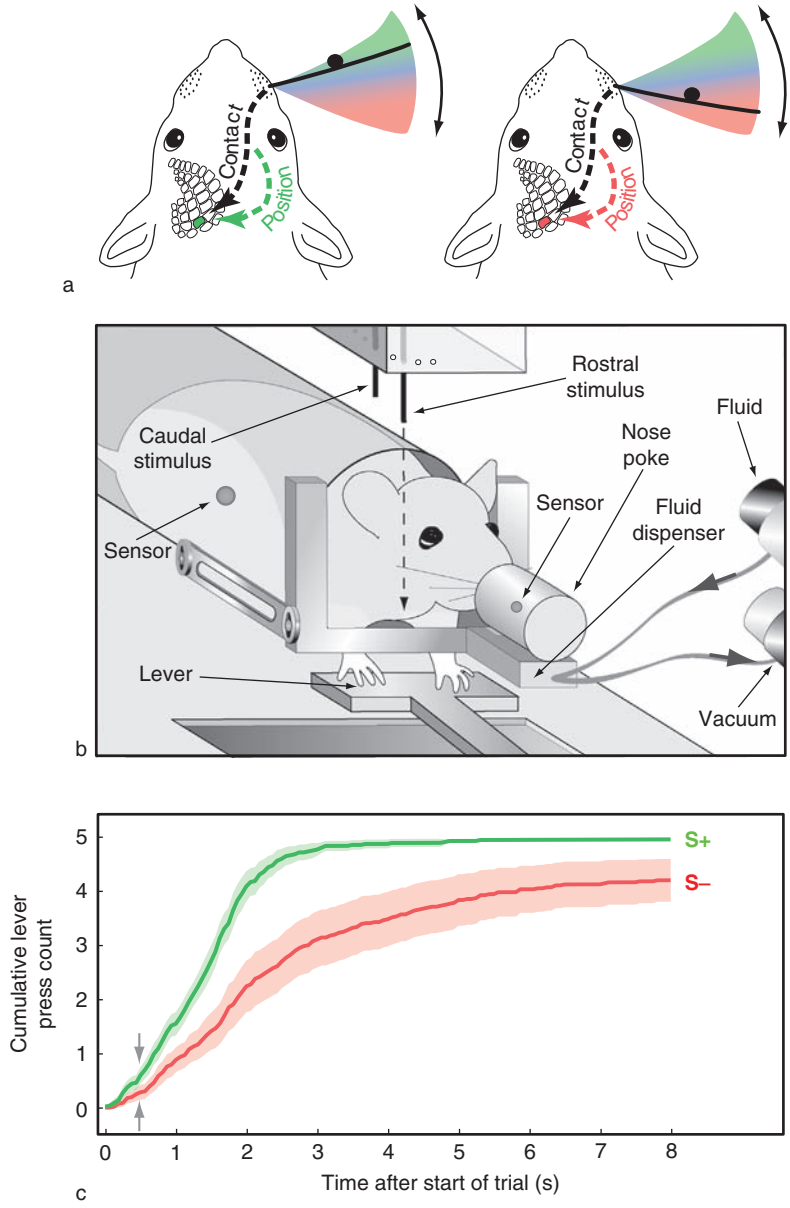

Figure 3 Behavior that involves the confluence of touch and position signals during exploratory whisking. (a) Cartoon that depicts a task to detect the position of an object relative to the rat's head with only one vibrissa. This task differentiates between labeled line-schemes that involve multiple vibrissae but no knowledge of vibrissa position and haptic-schemes that depend on both touch and position neural streams (dashed arrows) but do not require multiple vibrissa. (b) Details of the training and testing arena. Discrimination trials begin when an animal trips the nose poke sensor, which causes either the rostral or the caudal stimulus pin to descend into the vibrissa field. Lever presses in response to the rewarded, that is, S+, stimulus, either the rostral or caudal pin for each animal, result in a drop of water in the fluid dispenser. Lever presses in response to the unrewarded, that is, S−, stimulus have no effect. (c) Cumulative behavioral responses for one session, averaged separately over S+ and S− trials. The line and shaded regions give the mean ±2 standard error of measurement cumulative lever press counts. The gray arrows at 0.5 s mark the time after which the error regions remain nonoverlapping. Adapted from Mehta SB, Whitmer D, Figueroa R, Williams BA, and Kleinfeld D (2007) Active spatial perception in the vibrissa scanning sensorimotor system. *PLoS Biology* 5: e15.

Midbrain Loop

A intermediate-level loop incorporates the superior colliculus and includes connections that cross the midline (Figure 4). The superior colliculus is a laminar, midbrain structure in which each layer is nominally devoted to integrating sensory and motor information relevant to a particular sensory modality. In rat, the intermediate and deep layers of the colliculus appear to be devoted to general somatic sensorimotor processing, with the more

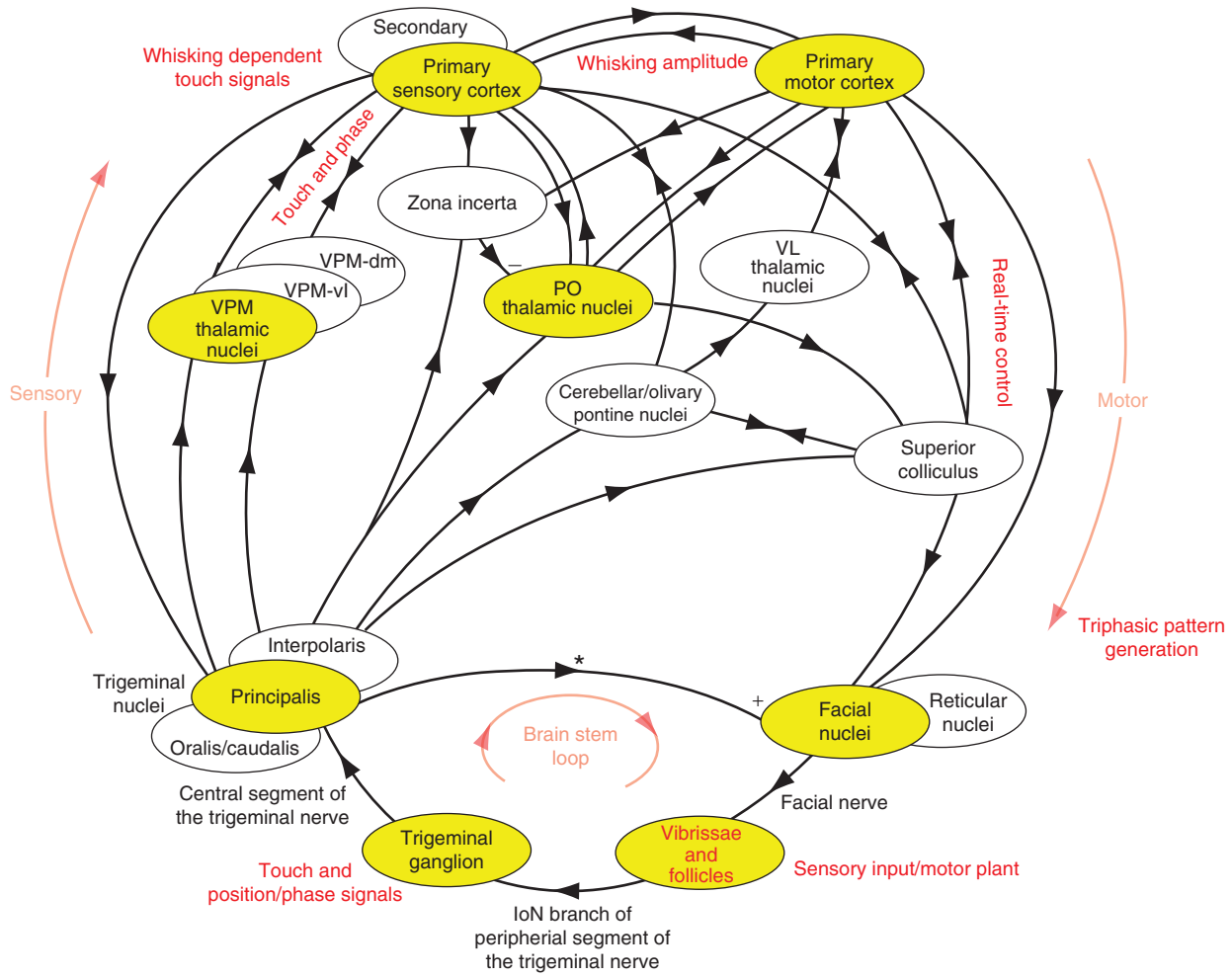

Figure 4 Cartoon of the anatomy of nested, vibrissa sensorimotor loops. The proposed connections are gleaned from the work of very many laboratories and provide a coarse roadmap of the flow of neuronal signals. Only pathways of direct relevance to active sensing with vibrissae are shown, with the most studied areas, from the perspective of electrophysiology, shown in yellow. Subdivisions of each area that are not shown in the figure may be described below. Basal ganglion pathways (Mercier BE, Legg CR, and Glickstein M (1990) *Proceedings of the National Academy of Sciences of the United States of America* 87: 4388–4392; Deniau JM, Kita H, and Kitai ST (1992) *Neuroscience Letters* 144: 202–206), are not included as they have only recently received attention in the vibrissa community (Hoffer ZS, Arantes HB, Roth RL, and Alloway KD (2005) *Journal of Comparative Neurology* 488: 2–100). Nor has the loop between neocortex and the hippocampus been considered (Buzsaki G (1996) *Cerebral Cortex* 6: 81–92; Chrobak JJ, Lorincz A, and Buzsaki G (2000) *Hippocampus* 10: 457–465), a bias that reflects current ignorance in the role of these structures in sensorimotor control. The asterisk labels the shortest sensorimotor feedback pathway, which is a single-synapse from the principal trigeminal nucleus to the lateral aspect of the facial nucleus. Hindbrain loops: (vibrissae → trigeminal ganglion) The vibrissae are innervated by both slowly and rapidly adapting sensory afferents (**Figure 1**) that originate from the infraorbital nerve (Dorfl J (1985) *Journal of Anatomy* 142: 173–184; Rice FL A, Mance A, and Munger BL (1986) *Journal of Comparative Neurology* 252: 154–174; Leiser SC and Moxon KA (2006) *Journal of Neurophysiology* 95: 3125–3145). (Trigeminal ganglion → trigeminal nucleus) Sensory inputs from the trigeminal ganglion project to the principal sensory nucleus (PrV) and the spinal trigeminal nuclei oralis (SpVO), interpolaris (SpVI), and caudalis (SpVC; Torvik A (1956) *Journal of Comparative Neurology* 106: 51–132; Clarke WB and Bowsher D (1962) *Experimental Neurology* 6: 372–383). The PrV and SpVI nuclei and the magnocellular portion of SpVC have somatotopic maps of the vibrissae ('barrelettes'; Ma PM and Woolsey RA (1984) *Brain Research* 306: 374–379); SpVO receives sensory input from the vibrissae yet does not contain a map (Belford GR and Killackey HP (1979) *Journal of Comparative Neurology* 188: 63–74; Belford GR and Killackey HP (1979) *Journal of Comparative Neurology* 183: 305–322). Last, there is particularly high internuclear connectivity among the SpVC and SpVO nuclei (Jacquin MF, Chiaia NL, Haring JH, and Rhoades RW (1990) *Somatosensory and Motor Research* 7: 399–420). The PrV nucleus receives collateral excitatory input from the SpVC and SpVI nuclei and inhibition from the SpVI nucleus (Furuta T, Timofeeva E, Nakamura K, et al. (2008) Inhibitory gating of vibrissal inputs in the brainstem. *Journal of Neuroscience* 28: 1789–1797). (Trigeminal nuclei → facial nuclei) The facial nucleus contains five subnuclei, of which the lateral subnucleus is involved in vibrissa control (Papez JW (1927) *Journal of Comparative Neurology* 42: 159–191; Martin MR and Lodge D (1977) *Brain Research* 38: 206–210). Vibrissa areas SpVC, PrV, and SpVI connect to the lateral subnucleus, primarily through ipsilateral projections. The trigeminal loop is closed by direct projection from PrV, SpVC, and SpVI to the facial nucleus (Erzurumlu RS and Killackey HP (1979) *Journal of Comparative Neurology* 188: 75–86; Hattox AM, Priest CA, and Keller A (2002) *Journal of Comparative Neurology* 442: 266–276) and indirect pathways within the hindbrain via the pontomedullary reticular formation (Dauvergne C, Pinganaud G, Buisseret P, Buisseret-Delmas C, and Zerari-Mailly F (2001) *Neuroscience Letters* 311: 109–112;

rostral and lateral areas responding to vibrissa-related inputs from the contralateral trigeminal nuclei. Descending afferents from the superior colliculus project to the lateral subnucleus of the contralateral facial nucleus to complete the loop. An additional input that converges to the same laminae arises from ipsilateral vibrissa primary motor (M1) cortex.

Zerari-Mailly F, Pinganaud G, Dauvergne C, Buisseret P, and Buisseret-Delmas CJ (2001) *Journal of Comparative Neurology* 429: 80–93). The direct projection, and possibly the indirect projection, is predominantly excitatory and results in positive feedback (Nguyen Q-T and Kleinfeld D (2005) *Neuron* 45: 447–457). (Facial nucleus → vibrissae) The lateral aspect facial nucleus sends projections to the intrinsic muscles surrounding each vibrissa (Arvidsson J (1982) *Journal of Comparative Neurology* 211: 84–92; Dorfl J (1982) *Journal of Anatomy* 135: 147–154; Dorfl J (1985) *Journal of Anatomy* 142: 173–184; Arvidsson J and Rice F L (1991) *Journal of Comparative Neurology* 309: 1–16), and medial aspects are believed to project to the extrinsic muscles (**Figures 2(a)–2(c)**; Klein B and Rhoades R (1985) *Journal of Comparative Neurology* 232: 55–69; Isokawa-Akesson M and Komisaruk BR (1987) *Experimental Brain Research* 65: 385–398). The lateral subnucleus of the facial nucleus contains a somatotopic map of the vibrissae (Martin MR and Lodge D (1977) *Brain Research* 38: 206–210). Midbrain loops: (trigeminal nuclei → superior colliculus) The trigeminal nuclei project to vibrissa somatotopic areas of the superior colliculus (Drager UC and Hubel DH (1976) *Journal of Neurophysiology* 39: 91–101; Killackey H and Erzurumlu R (1981) *Journal of Comparative Neurology* 201: 221–242; Huerta M, Frankfurter A, and Harting J (1983) *Journal of Comparative Neurology* 220: 147–167; Steindler DA (1985) *Journal of Comparative Neurology* 237: 155–175; Bruce LL, McHaffie JG, and Stein BE (1987) *Journal of Comparative Neurology* 262: 315–330; Jacquin M, Barcia M, and Rhoades RW (1989) *Journal of Comparative Neurology* 282: 45–62; Benett-Clarke CA, Chiaia NL, Jacquin MF, and Rhoades RW (1992) *Journal of Comparative Neurology* 320: 323–338); Hemelt ME and Keller A (in press) Superior colliculus control of vibrissa movements. *Journal of Neurophysiology*. The connection from nucleus SpVI appears to be the strongest (Killackey H and Erzurumlu R (1981) *Journal of Comparative Neurology* 201: 221–242; Huerta M, Frankfurter A, and Harting J (1983) *Journal of Comparative Neurology* 220: 147–167; Jacquin M, Barcia M, and Rhoades RW (1989) *Journal of Comparative Neurology* 282: 45–62); it terminates in the intermediate and deep layers of the lateral and rostral aspects of the colliculus. The projections from the trigeminal ganglia to colliculus are likely to be collaterals of projections to thalamus (Benett-Clarke CA, Chiaia NL, Jacquin MF, and Rhoades RW (1992) *Journal of Comparative Neurology* 320: 323–338; Mantle-St. John LA and Tracey DJ (1987) *Journal of Comparative Neurology* 255: 259–271). (Superior colliculus → facial nucleus) The intermediate and deep layers of the colliculus project to the lateral subnucleus of the facial nerve nucleus (Isokawa-Akesson M and Komisaruk BR (1987) *Experimental Brain Research* 65: 385–398; Miyashita E, Keller A, and Asanuma H (1994) *Experimental Brain Research* 99: 223–232; Miyashita E, and Shigemi M (1995) *Neuroscience Letters* 195: 69–71). (Trigeminal nuclei → cerebellum) The trigeminal nuclei provide vibrissa sensory input to the cerebellum via a direct pathway, that is, the SpVI nucleus projects directly to the cerebellum (Woolston DC, LaLonde JR, and Gilson JM (1982) *Journal of Neurophysiology* 48: 160–173), and two indirect pathways, that is, the SpVI and SpVC nuclei project via the inferior olive climbing fibers and the PrV, SpVI, and SpVC nuclei project via pontine mossy fibers (Huerta M, Frankfurter A, and Harting J (1983) *Journal of Comparative Neurology* 220: 147–167; Steindler DA (1985) *Journal of Comparative Neurology* 237: 155–175; Jacquin M, Barcia M, and Rhoades RW (1989) *Journal of Comparative Neurology* 282: 45–62; Mantle-St. John LA and Tracey DJ (1987) *Journal of Comparative Neurology* 255: 259–271; Smith RL (1973) *Journal of Comparative Neurology* 148: 423–446; Watson CRR and Switzer III RC (1978) *Neuroscience Letters* 10: 77–82; Swenson RS, Kosinski RJ, and Castro AJ (1984) *Journal of Comparative Neurology* 222: 301–311). Projections from the trigeminal nuclei to the inferior olive overlap those from the olive to the cerebellum; the target areas in the cerebellum include *crura* I and II and the paramedian lobule and uvula (Huerta M, Frankfurter A, and Harting J (1983) *Journal of Comparative Neurology* 220: 147–167; Watson CRR and Switzer III RC (1978) *Neuroscience Letters* 10: 77–82); all these areas have facial receptive fields. (Superior colliculus ↔ cerebellum) The colliculus sends projections to the cerebellar cortex, including target areas *crura* I and II, through both the inferior olive and the pons (Kassel J (1980) *Brain Research* 202: 291–305). The deep cerebellar nuclei send a projection back to the colliculus (Lee HS, Kosinski RJ, and Mihailoff GA (1989) *Neuroscience* 28: 725–734; Westby GW, Collinson C, and Dean P (1993) *European Journal of Neuroscience* 5: 1378–1388; Westby GW, Collinson C, Redgrave P, and Dean P (1994) *European Journal of Neuroscience* 6: 1335–1342) to form a 'colliculus → cerebellum → colliculus' loop. Forebrain loops: (trigeminal nuclei → dorsal thalamus) The PrV nucleus sends ascending projections to ventral posteromedial (VPM) thalamic nuclei, and the SpVI nucleus sends projections to posterior (PO) nuclei and to the ventral-lateral area of VPM (VPM-vl) thalamus as well as the dorsal-medial area of VPM (VPM-dm) thalamus (Lund RD and Webster KE (1967) *Journal of Comparative Neurology* 130: 313–328; Smith RL (1973) *Journal of Comparative Neurology* 148: 423–446; Erzurumlu RS and Killackey HP (1980) *Neuroscience*. 5: 1891–1901; Hoogland PV, Welker E, and Van der Loos H (1987) *Experimental Brain Research* 68: 73–87; Mantle-St. John LA and Tracey DJ (1987) *Journal of Comparative Neurology* 255: 259–271; Jacquin M, Barcia M, and Rhoades RW (1989) *Journal of Comparative Neurology* 282: 45–62; Killackey HP, Jacquin M, and Rhoades RW (1990) In: *Development of Sensory Systems in Mammals*, pp. 403–429. New York: Wiley; Chiaia NL, Rhoades RW, Bennett-Clark CA, Fish SE, and Killackey HP (1991) *Journal of Comparative Neurology* 314: 201–216; Benett-Clarke CA, Chiaia NL, Jacquin MF, and Rhoades RW (1992) *Journal of Comparative Neurology* 320: 323–338; Diamond ME, Armstrong-James M, Budway MJ, and Ebner FF (1992) *Journal of Comparative Neurology* 319: 66–84; Williams MN, Zahm DS, and Jacquin MF (1994) *European Journal of Neuroscience* 6: 429–453; Veinante P and Deschenes M (1999) *Journal of Neuroscience* 19: 5085–5095; Pierret T, Lavallee P, and Deschenes M (2000) *Journal of Neuroscience* 20: 7455–7462; Yu C, Derdikman D, Haidarliu S, and Ahissar E (2006) Parallel thalamic pathways for whisking and touch signals in the rat. *PLoS Biology* 4: e124; Simons DJ, Carvell GE, Kyriazi HT, and Bruno RM (2007) Thalamocortical conduction times and stimulus-evoked responses in the rat whisker-to-barrel system. *Journal of Neurophysiology* 98: 2842–2847; Masri R, Bezdudnaya T, Trageser JC, and Keller A (in press) Encoding of stimulus frequency and sensor motion in the posterior medial thalamic nucleus. *Journal of Neurophysiology*). The representation of the vibrissae forms a somatotopic map ('barreloids') in VPM thalamus (Van Der Loos H (1976) *Neuroscience Letters* 2: 1–6; Sugitani M, Yano J, Sugai T, and Ooyama H (1990) *Experimental Brain Research* 81: 346–351) and a diffuse map in PO thalamus (Nothias F, Peschanski M, and Besson J-M (1988) *Brain Research* 447: 169–174; Fabri M and Burton

Cerebellar Loops

The pontine–cerebellar system appears to function at a hindbrain level in a loop that involves direct connections from the contralateral SpVI nucleus and indirect connections between contralateral trigeminal nuclei and the ipsilateral superior colliculus (Figure 4). For the latter case, the trigeminal nuclei project to both the pons and the inferior olive, which in turn directly project to crura 1 and 2 in the cerebellum. Similar inputs, which project to the same crura in cerebellum,

H (1991) *Brain Research* 538: 351–357) and VPM-vl (Pierret T, Lavallee P, and Deschenes M (2000) *Journal of Neuroscience* 20: 7455–7462). (Trigeminal nuclei → zona incerta) Nuclei in zona incerta (ZI) consist exclusively of inhibitory projection neurons; those involved in vibrissa somatosensation receive input from SpVI (Kolmac CI, Power BD, and Mitrofanis J (1998) *Journal of Comparative Neurology* 396: 544–555). ZI forms a negative (**–**) feedback connection to PO thalamus (Trageser JC and Keller A (2004) *Journal of Neuroscience* 24: 8911–8915; Lavallee P, Urbain N, Dufresne C, Bokor H, Acsady L, and Deschenes M (2005) *Journal of Neuroscience* 25: 7489–7498) that is inactivated only by descending input from vibrissa cortex Urbain N and Deschenes M (2007) Motor cortex gates vibrissal responses in a thalamocortical projection pathway. *Neuron* 56: 714–725 (Dorsal thalamus ↔ neocortex) Thalamic regions VPM, PO, and ZI project to primary (S1), secondary (S2), and posterior ventral areas of sensory cortex, and cortex sends feedback projections to VP, PO, and the trigeminal nuclei (Wise SP and Jones EG (1977) *Journal of Comparative Neurology* 175: 129–158; Donoghue JP, Kerman KL, and Ebner FF (1979) *Journal of Comparative Neurology* 183: 647–664; Donoghue JP and Kitai ST (1981) *Journal of Comparative Neurology* 201: 1–13; Carvell GE and Simons DJ (1987) *Journal of Comparative Neurology* 265: 409–427; Hoogland PV, Welker E, and Van der Loos H (1987) *Experimental Brain Research* 68: 73–87; Koralek K, Jensen KF, and Killackey HP (1988) *Brain Research* 463: 346–351; Welker E, Hoogland PV, and van der Loos H (1988) *Experimental Brain Research* 73: 411–435; Jacquin MF, Wiegand MR, and Renehan WE (1990) *Journal of Neurophysiology* 64: 3–27; Chiaia NL, Rhoades RW, Bennett-Clark CA, Fish SE, and Killackey HP (1991) *Journal of Comparative Neurology* 314: 201–216; Chiaia NL, Rhoades RW, Fish SE, and Killackey HP (1991) *Journal of Comparative Neurology* 314: 217–236; Diamond ME, Armstrong-James M, Budway MJ, and Ebner FF (1992) *Journal of Comparative Neurology* 319: 66–84; Deschênes M, Bourassa J, and Parent A (1996) *Neuroscience* 72: 679–687; Lévesque M, Charara A, Gagnon S, Parent A, and Deschênes M (1996) *Brain Research* 709: 311–315; Shepherd GM and Svoboda K (2005) *Journal of Neuroscience* 25: 5670–5679; Landisman CE and Connors BW (2007) VPM and PoM Nuclei of the rat somatosensory thalamus: Intrinsic neuronal properties and corticothalamic feedback. *Cerebral Cortex* 17: 2853–2865, Urbain N and Deschenes M (2007) A new thalamic pathway of vibrissal information modulated by the motor cortex. *Journal of Neuroscience* 27: 12407–12412; Bokor H, Acsady L, and Deschenes M (2008) Vibrissal responses of thalamic cells that project to the septal columns of the barrel cortex and to the second somatosensory area. *Journal of Neuroscience* 28: 5169–5177); The projection from ZI thalamus to S1 cortex is unique in providing an inhibitory input (Chapin JK, Schneider JS, Nicolelis M, and Lin C-S (1990) *Science* 248: 1553–1556; Nicolelis MAL, Chapin JK, and Lin RCS (1992) *Brain Research* 577: 134–141). (Direct intercortical connections) Vibrissa S1 cortex forms reciprocal projections with other vibrissa sensory areas in neocortex (Carvell GE and Simons DJ (1987) *Journal of Comparative Neurology* 265: 409–427; Chapin JK, Sadeq M, and Guise JLU (1987) *Journal of Comparative Neurology* 263: 326–346; Welker E, Hoogland PV, and van der Loos H (1988) *Experimental Brain Research* 73: 411–435; Fabri M and Burton H (1991) *Journal of Comparative Neurology* 311: 405–424; Hoffer ZS, Hoover JE, and Alloway KD (2003) *Journal of Comparative Neurology* 466: 525–544; Chakrabarti S and Alloway KD (2006) *Journal of Comparative Neurology* 498: 624–636) and with vibrissa motor (M1) cortex (White EL and deAmicis RA (1977) *Journal of Comparative Neurology* 175: 455–482; Fabri M and Burton H (1991) *Journal of Comparative Neurology* 311: 405–424; Aroniadou VA and Keller A (1993) *Journal of Neurophysiology* 70: 1493–1553; Keller A (1993) *Cerebral Cortex* 3: 430–441; Miyashita E, Keller A, and Asanuma H (1994) *Experimental Brain Research* 99: 223–232; Izraeli R and Porter LL (1995) *Experimental Brain Research* 104: 41–54; Veinante P and Deschenes M (2003) *Journal of Comparative Neurology* 464: 98–103; Hoffer ZS, Arantes HB, Roth RL, and Alloway KD (2005) *Journal of Comparative Neurology* 488: 82–100; Chakrabarti S and Alloway KD (2006) *Journal of Comparative Neurology* 498: 624–636). The representation of the vibrissae forms a somatotopic map in S1 cortex (Woolsey TA, Welker C, and Schwartz RH (1974) *Journal of Comparative Neurology* 164: 79–94; Durham D and Woolsey TA (1977) *Brain Research* 137: 169–174)-'barrels'- and S2 (Carvell GE and Simons DJ (1986) *Somatosensory Research* 3: 213–237; Kleinfeld D and Delaney KR (1996) *Journal of Comparative Neurology* 375: 89–108). Note that M1 is taken as the parasagittal agranular medial area. (Indirect intercortical connections) Feedback from neocortex to thalamus via projections from cortical layer 6 provides a pathway for different cortical columns and regions to intact in thalamus (Deschenes M, Veinante P, and Zhang Z-W (1998) *Brain Research Reviews* 28: 286–308). This effect is enhanced by an effective intrathalamic connection mediated by thalamoreticular and recticulothalmic projections (Crabtree JW, Collingridge GL, and Isaac JT (1998) *Nature Neuroscience* 1: 289–394; Crabtree JW and Isaac JT (2002) *Journal of Neuroscience* 22: 8754–8761; Golomb D, Ahissar E, and Kleinfeld D (2005) *Journal of Neurophysiology* 95: 1735–1750). (Neocortex → superior colliculus) Both sensory and motor cortices send descending projections to the superior colliculus (Wise SP and Jones EG (1977) *Journal of Comparative Neurology* 175: 129–158; Killackey H and Erzurumlu R (1981) *Journal of Comparative Neurology* 201: 221–242; Welker E, Hoogland PV, and van der Loos H (1988) *Experimental Brain Research* 73: 411–435; Mercier BE, Legg CR, and Glickstein M (1990) *Proceedings of the National Academy of Sciences of the United States of America* 87: 4388–4392). Cellular interactions between the descending cortical M1 projection to colliculus and from the colliculus to the facial nucleus support the relay of motor commands to the facial nucleus (Miyashita E and Shigemi M (1995) *Neuroscience Letters*195: 69–71). (Motor cortex → reticular nuclei) A direct connection from vibrissa M1 cortex to nuclei in the reticular formation (Miyashita E, Keller A, and Asanuma H (1994) *Experimental Brain Research* 99: 223–232; Hattox AM, Priest CA, and Keller A (2002) *Journal of Comparative Neurology* 442: 266–276) is suggestive of a central pattern generator (CPG) by analogy with the CPG for mastication (Nozaki S, Iriki A, and Nakamura Y (1986) *Journal of Neurophysiology* 55: 806–825) and provides descending control of the vibrissae. The ambiguus and parvocellular reticular nuclei in the medulla, as well as the pontine reticular nucleus, are capable of evoking vibrissa movement and further receive input from M1 cortex (Hattox AM, Priest CA, and Keller A (2002) *Journal of Comparative Neurology* 442: 266–276). Critically, direct input from M1 terminates in the lateral facial nucleus and may directly drive the vibrissae (Grinevich V, Brecht M, and Osten P (2005) *Journal of Neuroscience* 25: 8250–8258).

Table 2 Afferent nuclei to the lateral aspect of the facial nucleus

Afferent nucleus	*VMCtx afferent*	*Evoked movement*	*Transmitter*
Myelencephalon (medulla)			
Dorsal motor nucleus of the vagus	x		
Reticular nucleus of the medulla	x		
Intermediate reticular nucleus	x		GABA, Gly
Lateral reticular nucleus	x		
Ambiguus nucleus	x	x	
Gigantocellular reticular nucleus	x		GABA, Gly, 5-HT
Parvocellular reticular nucleus	x	x	
Spinal vestibular nuclei	x		GABA, Gly
External cuneate nucleus			
Nucleus solitary tract			
Raphe magnus, pallidus, obscurus			5-HT
Paragigantocellular reticular nucleus			
Spinal trigeminal nucleus			GABA, Gly
Hypoglossal nucleus			
Metencephalon (pons)			
Pontine reticular nucleus	x	x	GABA, Gly
Ventral parabrachial nucleus	x		
Kölliker-Fuse nucleus	x		
Paralemniscal nucleus	x		GABA, Gly
Lateral parabrachial nucleus	x		
Subpedencular tegmental nucleus	x		
Intertrigeminal nucleus	x		
Subcoeruleus nucleus	x		
Ventral nucleus of the lateral lemniscus			
Motor trigeminal nucleus			
Pediculopontine tegmental			
Medial parabrachial nucleus			
Mesencephalon (midbrain)			
Deep mesencephalic nucleus	x		
Oculomotor nucleus	x		
Central gray	x		
Superior colliculus	x	x	
Red nucleus	x	x	
Edinger-Westphal nucleus	x		
Pararubral nucleus	x		
Nucleus raphe dorsalis			
Interstitial nucleus of medial longitudinal fasciculus			
Retrorubral nucleus			
Nucleus Darkschewitsch			
Substantia nigra			
Primary motor cortex (VMCtx)	–	x	Glutamate

The listings are based on anatomical studies. Neurons in the lateral facial nucleus integrate inputs from about 40 presynaptic sources (Fay RA and Norgren R (1997) *Brain Research Reviews* 25: 276–290). These include reticular motor nuclei that are potentially involved with the generation of the oscillatory drive for rhythmic whisking, such as the parvocellular nucleus (Mogoseanu D, Smith AD, and Bolam JP (1994) *Experimental Brain Research* 101: 427–438; Hattox AM, Priest CA, and Keller A (2002) *Journal of Comparative Neurology* 442: 266–276), reticular nuclei that are involved with other oromotor behaviors (Travers JB (1995) In: Paxinos G (ed.) *The Rat Nervous System*, 2nd edn., pp. 239–255. San Diego: Academic Press), such as suckling and licking, trigeminal sensory nuclei, whose input to the facial nucleus completes a feedback loop that encompasses the vibrissae; cerebellar deep nuclei, whose input also completes a feedback loop (Huerta M, Frankfurter A, and Harting J (1983) *Journal of Comparative Neurology* 220: 147–167), input from the superior colliculus (Miyashita E, Keller A, and Asanuma H (1994) *Experimental Brain Research* 99: 223–232; Miyashita E and Shigemi M (1995) *Neuroscience Letters* 195: 69–71), and direct input from the primary motor cortex (Grinevich V, Brecht M, and Osten P (2005) *Journal of Neuroscience* 25: 8250–8258). The column labeled 'VMCtx afferent' denotes that the nucleus receives a projection from primary motor cortex. VMCtx, vibrissa primary motor cortex. The column labeled 'Evoked movement' indicates that activation of this region in anesthetized animals leads to vibrissa movement.

arise from the intermediate and deep layers of the superior colliculus. The relative means by which these inputs contribute to a pronounced sensory response in crus 2, and a more restricted response in crus 1, is unknown. Finally, the cerebellar Purkinje cells form synapses on the cerebellar nuclei, which provide output projections to superior colliculus to complete this intermediary-level loop.

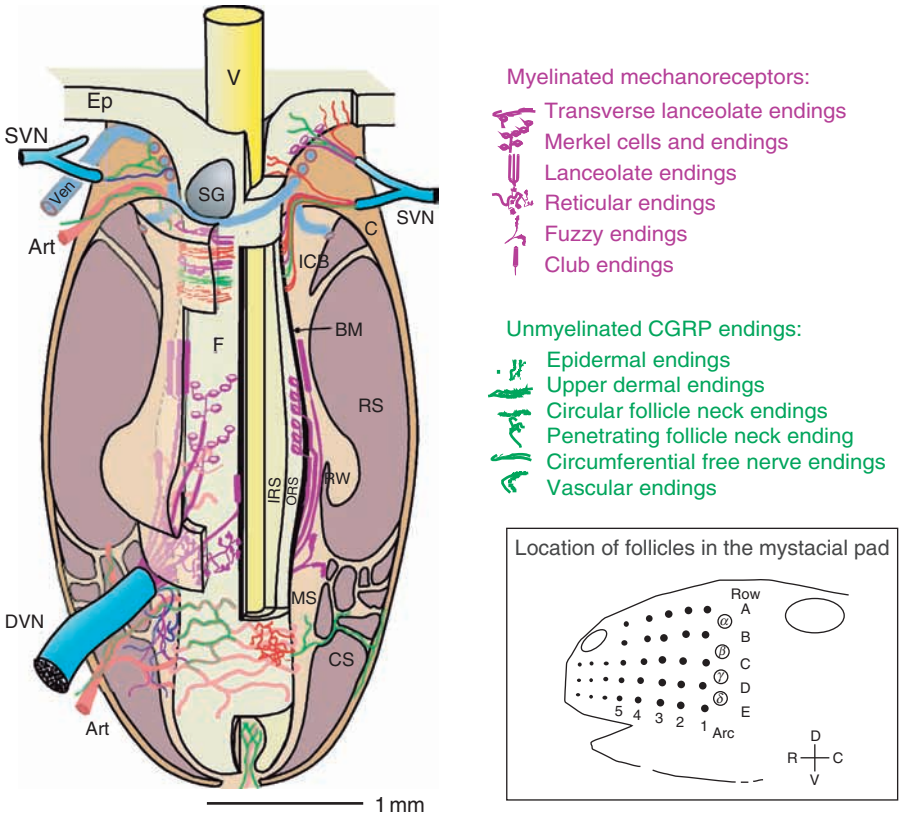

Figure 5 The follicle contains both superficial (SVN) and deep (DVN) nerve terminals that report both self-movement and contact of the vibrissa. The follicle sits in the mystacial pad as part of an array of five rows and roughly ten arcs of follicle/vibrissa units (Insert in lower right corner). Rows are labeled by letters and arcs by numbers. The four posterior vibrissa, referred to as straddlers, are labeled by Greek letters. In the awake and aroused animal, the sinuses (CS and RS) are engorged with blood, which stiffens the structure. Both self-movement and touch are presumed to be coded by both the superficial and deep nerves, which project to the trigeminal ganglion and have similar passive response properties (Waite PME and Jacquin MF (1992) *Journal of Comparative Neurology* 322: 233–245). The DVN ultimately projects to the PrV, SpVO, SpVI, and SpVC trigeminal nuclei (Hayashi H (1985) *Journal of Comparative Neurology* 237: 195–215; Jacquin MF, Stennett RA, Renehan WE, and Rhoades RW (1988) *Journal of Comparative Neurology* 267: 107–130), and the SVN has been reported to project only to SpVC (Martin Deschenes, unpublished observations). This illustration was adapted from Arvidsson J and Rice FL (1991) Central projections of primary sensory neurons innervating different parts of the vibrissae follicles and intervibrissal skin on the mystacial pad of the rat. *Journal of Comparative Neurology* 309: 1–16; see also Rice FL, Fundin BT, Pfaller K, and Arvidsson J (1994) *Experimental Brain Research* 99: 233–246. A mechanical model of the follicle is given by Mitchinson B, Gurney KN, Redgrave P, et al. (2004) *Proceedings of the Royal Society of London, Series B: Containing Papers of a Biological Character* 271: 2509–2516. V, vibrissa; SG, sebacceous gland; RW, ringwulst; RS, ring sinus; ORS, outer root sheath; MS, mesenchymal sheath; IRS, inner root sheath; ICB, inner conical body; F, follicle; Ep, epidermis; CS, cavernous sinus; C, follicle-sinus-complex capsule; BM, basement membrane; CGRP, calcitonin gene-related peptide; and Art, arteriole.

Thalamic-Forebrain Loop

Multiple structures in ventral and dorsal thalamus receive input from the trigeminal nuclei. Only one of these, zona incerta (ZI) in the ventral thalamus, projects directly back to the superior colliculus (Figure 4), where it forms inhibitory connections with the superior colliculus. Thus, ZI thalamus appears to function as a forebrain-level intermediary in a loop that involves the trigeminal nuclei and the colliculus. Further, inhibitory projections from ZI thalamus serve to gate afferents from PO thalamus to vibrissa SI cortex. This gate is removed by descending inputs from vibrissa M1 cortex.

Neocortical Loop

Cortex receives input from the trigeminal nuclei along four streams that pass through the thalamus (Figure 4):

1. The well-known lemniscal pathway that ascends from the PrV nucleus via the central region of the ventral posterior medial (VPM) thalamus and projects to vibrissa primary somatosensory (S1) cortex. Neurons in VPM thalamus rapidly respond to stimulation of a single vibrissa in anesthetized animals, whose receptive fields are referred to as barreloids. Similarly, neurons in the granular layer

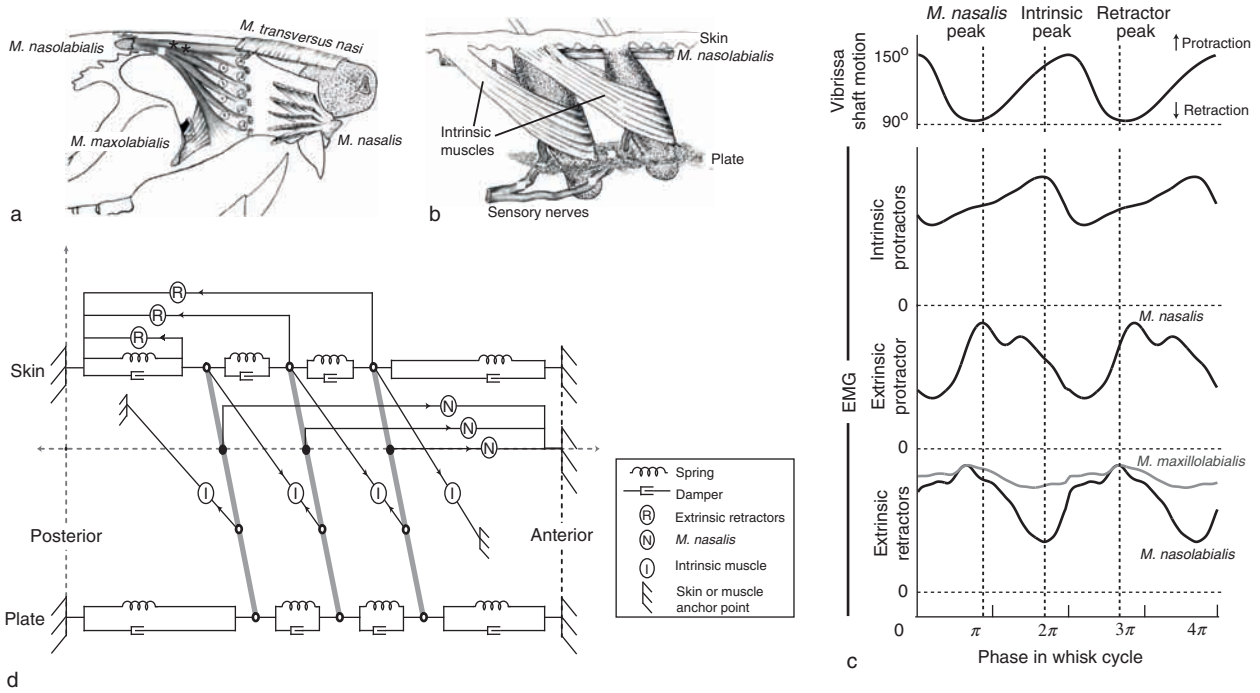

Figure 6 Geometry, musculature, and motion of the follicles and vibrissae. (a) Drawing of extrinsic musculature in mouse; a similar pattern occurs in rat. Four extrinsic muscles invade the mystacial pad while maintaining external attachment points. The retractor *M. nasolabialis* attaches dorsal–caudal to the pad and runs superficially below the skin. A second retractor, *M. maxillolabialis*, attaches ventral–caudal to the pad and fuses with the fibers of *M. nasolabialis* as they invade the pad. *M. nasalis* attaches rostral to the pad at the nasal septum and runs deep to the follicles as it extends caudally. *M. transversus nasi* lies transverse to the snout and runs superficially through the pad. (b) Drawing of intrinsic musculature in mouse. The intrinsic muscles join adjacent follicles (insert in **Figure 5**) of a single row. Each muscle attaches medially and laterally to the superior part of the caudal follicle while forming a sling around the lower third of the rostral follicle. The skin and other connective tissue, such as the fibrous plate, provide a passive restoring force. Superficial extrinsic muscles run just below the skin. (c) A mechanical model of a row of three vibrissae, shown in the rest state, to illustrate how the extrinsic and intrinsic muscles pull on the vibrissae. The attachment points are illustrated for the springs, dampers, and muscles, which together from an active vasoelastic element. The approximate relationship between these points is conserved, but the figure is not drawn to scale. Arrows indicate the direction of muscle forces, which point away from the attachment points. (d) Average vibrissa motion and muscle activity from a head-restrained rat. Each whisk was linearly mapped from time onto the range of 0 to 2π radians and the average taken across phase. Average traces (1750 whisks) are repeated to display two cycles; only motion along the A–P axis is shown. The rectified electromyogram values were normalized by the maximum voltage of each trace. The dashed vertical lines indicate the three phases of average muscle activity. The drawings in panels (a) and (b) were adapted from Dorfl J (1982) The musculature of the mystacial vibrissae of the white mouse. *Journal of Anatomy* 135: 147–154; see also Dorfl J (1985) The innervation of the mystacial region of the white mouse: A topographical study. *Journal of Anatomy* 142: 173–184. The model and data in panels (c) and (d) from Hill DN, Bermejo R, Zeigler HP, and Kleinfeld D (2008) Biomechanics of the vibrissa motor plant in rat: Rhythmic whisking consists of triphasic neuromuscular activity. *Journal of Neuroscience* 28: 3438–3455.

of S1 cortex respond to stimulation of a single vibrissa, whose receptive fields are formed by afferents from VPM thalamus and are referred to as barrels.

2. The well-known paralemniscal pathway that ascends from the rostral part of the SpVI nucleus, passes through posterior medial (PO) thalamus, and projects down to the superior colliculus as well as up to agranular layers 1 and 5 in S1 cortex, vibrissa secondary somatosensory (S2), and vibrissa motor (M1) cortices. Neurons in PO thalamus receive inhibitory projections from ZI thalamus; this inhibitory block is relieved by projections from M1 cortex to PO thalamus.

3. A recently described extralemniscal pathway that ascends from the caudal part of the SpVI nucleus, passes through the ventrolateral border of VPM (VPM-vl) thalamus, and projects to S1 and S2 cortices. Neurons in VPM-vl thalamus rapidly respond to stimulation of multiple vibrissae in anesthetized animals.

4. A recently described extralemniscal pathway that ascends from the part of the PrV nucleus that contains large neurons with multivibrissa responses, passes through the dorsomedial border of VPM (VPM-dm) thalamus that lies next to PO thalamus, and projects to S1 cortex. Neurons in VPM-dm thalamus rapidly respond to stimulation of multiple vibrissae in anesthetized animals.

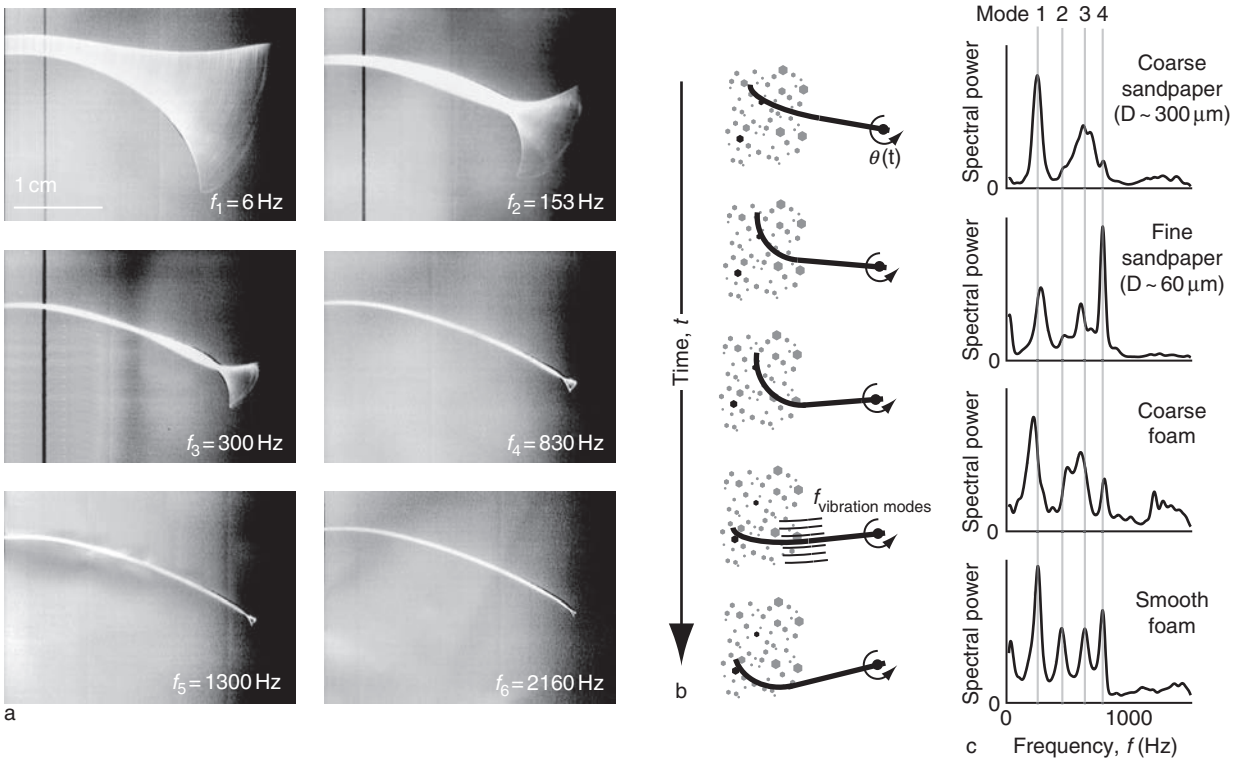

Figure 7 Dynamics of a single vibrissa. (a) Photographs of the motion of the straddler vibrissa γ (see insert, **Figure 5**) at successive resonant frequencies. The vibrissa was glued to a small voice coil that was driven by a wave generator. Note that the displacements are shifted to the finer and more distal parts of the vibrissa at progressively higher-order modes. Data was taken in the laboratory following reports by Hartmann (Hartmann MJ, Johnson NJ, Towal RB, and Assad C (2003) Mechanical characteristics of rat vibrissae: Resonant frequencies and damping in isolated whiskers and in the awake behaving animal. *Journal of Neuroscience* 23: 6510–6519) and Moore (Neimark MA, Andermann ML, Hopfield JJ, and Moore CI (2003) Vibrissa resonance as a transduction mechanism for tactile encoding. *Journal of Neuroscience* 23: 6499–6509; Ritt JT, Andermann ML, and Moore CI (2008) Emdodied information processing: Vibrissa mechanics and texture features shape micromotions in actively sensing rats. *Neuron* 57: 599–613). (b) Cartoon that shows how the drag on a vibrissa can couple to the intrinsic mechanical vibrations (modes) of the vibrissa via 'stick–slip' friction. This effect may be relevant for the encoding of texture. Adapted from Mehta SB and Kleinfeld D (2004) Frisking the whiskers: Patterned sensory input in the rat vibrissa system. *Neuron* 41: 181–184. (c) Spectral power of the vibrations measured at the base of a vibrissa artificially whisked across different textures, in support of a model in which different surfaces preferentially excite different modes of a vibrissa (Moore CI and Andermann ML (2005) In: Ebner FF (ed.) *Neural Plasticity in Adult Somatic Sensory–Motor Systems.* Boca Raton, FL: CRC Press). Adapted from Fend M, Bovet S, Yokoi H, and Pfeifer R (2003) An active artificial whisker array for texture discrimination. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS), Las Vegas, NV, October.

All four thalamic areas receive feedback projections from the infragranular layers of S1 and S2 cortices. Further, VPM-dm thalamus receives feedback projections from M1 cortex.

There is evidence for the segregation of neuronal signals among the different pathways. The lemniscal pathway appears to convey a combination of touch and position signals while the extralemniscal pathway appears to convey primarily touch-based signals. The evidence for signaling along the paralemniscal pathway is inconsistent. Activation of neurons in PO thalamus in anesthetized animals occurs via feedback from S1 cortex. On the other hand, during whisking neurons in PO thalamus may be driven directly by sensory input and report vibrissa position. Independent of this segregation of information, the thalamic nuclei interact via reciprocal connections to the reticular thalamic nucleus, which may lead to a mixing of touch and position signals, among others. Further, sensory and motor cortex interact through corticocortical projections, and there is considerable feedback among thalamic-mediated connections between cortical areas (**Figure 4**). This implies that sensory and motor functions are likely to be distributed throughout these areas.

The cortical loop is closed by corticospinal-like projections from M1 cortex to the vibrissa areas of the facial nucleus, as well as by projections from both S1 and M1 cortices to the superior colliculus, which in turn sends descending projections to motor neurons in the facial nucleus. By analogy with the

anatomy of corticospinal projections in primates, there may be undiscovered projections from vibrissa S1 cortex to the facial nucleus.

Back to the Brain Stem

Two more points bear on the nested topology of the vibrissa system. The first is that a single synaptic connection from neurons in the trigeminal complex to those in the facial nucleus (* in Figure 4) is paralleled, from hindbrain to midbrain to forebrain, by connections of increasing complexity. The second and related issue is that the motor neurons in the lateral facial nucleus appear as arbitrators of activity from loops at all levels, receiving input from neurons in nearly 40 identified nuclei (Table 2). This makes the motor neurons arbitrators of motor commands to the vibrissae.

Mechanics of the Vibrissae

The vibrissae are shaped as round, tapered beams. Videographs of the vibrissae as rats whisk clearly show that the vibrissae flex as a consequence of the muscular forces that propel them forward at their base as well as the forces that act on contact (Figure 1(b)). A result of this flexibility is that, as for all extended mechanical systems, individual vibrissae exhibit a set of resonances (Figure 7(a)). The lowest-order resonance exhibits bending all along the vibrissae, while higher-order resonances exhibit bending that is localized toward the tip.

What are the potential ethological roles that vibrations of the vibrissae can play? One is that a vibrissa can twang when it contacts an object, so that touch will yield a rapid succession of taps rather than a single tap. The frequency of these taps will depend on the location of contact along the shaft of the vibrissa but will be in the order of the roughly 100 Hz resonant frequencies. This rate is sufficient to induce short-term synaptic plasticity; so mechanical resonance may contribute to the transmission probability of a contact event.

The vibrations of the vibrissae may also play a role in sensing texture. The vibrissae may alternately stick and slip as they are dragged across a rough surface (Figure 7(b)). As the vibrissa slips, the resultant vibrations of the shaft will be a superposition of the intrinsic vibrations of the vibrissa. It is natural to posit that differences between surfaces are expressed by the extent to which different modes are favored. In particular, when the tip of a vibrissa is moved across surfaces of differing roughness, the set of frequencies of the vibration is essentially the same for all surfaces, but the relative amplitude associated with each vibration will depend on the detailed properties of the surface (Figure 7(c)).

Transmission of High-Frequency Signals

The timing of neuronal signals from the initial contact of a vibrissa with an object to their representation in vibrissa S1 cortex is central to perception. This timing is heavily dependent on the state of the animal, that is, sessile versus aroused, as mediated through synaptic changes and cellular adaptation. While modulation of the timing might complicate the processing of sensory signals, simplicities appear in two limits as signals ascend the lemniscal pathway (Figure 4). For punctate signals, as might occur with a sharp initial contact, there is a high probability that the signal will successfully propagate up through cortex. In contrast, for steady-state periodic signals, the probability of transmission can be low (Figure 8). At the level of primary sensory neurons, periodic signals are faithfully transduced up to frequencies of ~1200 Hz. However, this signal begins to degrade at the level of the PrV nucleus in brain stem, such that for vibration frequencies above ~350 Hz, roughly half the spikes are dropped (* in bottom row, second column, of Figure 8). By the level of VPM thalamus, roughly half the spikes are dropped for frequencies above ~35 Hz, and by the level of S1 cortex, roughly half the spikes are dropped for frequencies above ~5 Hz. Thus the ability of individual neurons to faithfully follow rapid movements of the vibrissa degrades with higher-order structures in the sensorimotor system.

How are high-frequency events, such as vibrations, encoded in cortex? First, even the response in the primary sensory cells may be insufficient to track the finest surface detail. In particular, the maximum response frequency of 1200 Hz corresponds to the rhythmic motion of the tip of a vibrissa across a corrugated surface with a pitch of only ~200 μm, which is rather coarse. Second, while neurons in S1 cortex cannot track the phase of periodic inputs above ~35 Hz, at least in anesthetized animals the high-frequency sensory inputs are coded as an approximate Poisson process (right column in Figure 8). It is interesting that the rate of this spike process is proportional to the logarithm of the vibration frequency of the vibrissa, that is,

$$\text{Firing Rate} \propto 2 \ln \{\text{Vibration Frequency}\} + \text{Constant}$$

This is in the form of Weber's Law, but for frequencies rather than intensity. The coarseness of logarithmic coding suggests that comparisons of rhythmic inputs with slightly different frequencies will be problematic.

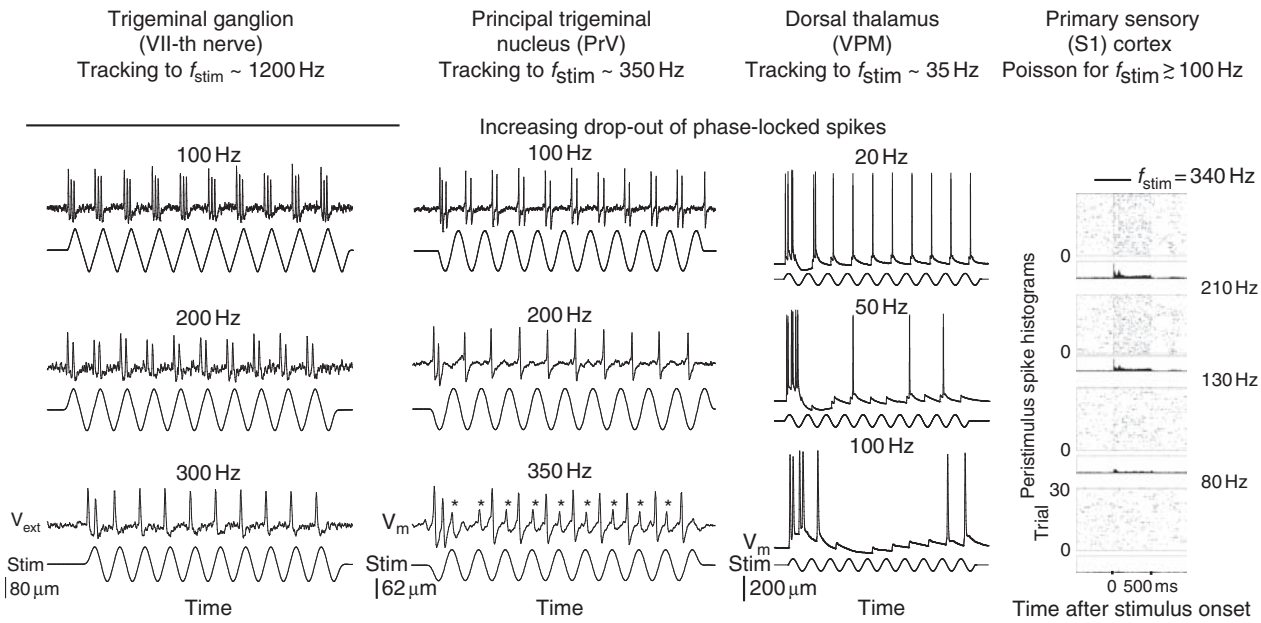

Figure 8 Loss in phase-locked spiking with ascending activation of vibrissa brain centers as the frequency of periodic stimulation is increased. In all cases a sinusoidal stimulus was applied to a single vibrissa in an anesthetized animal. The extracellular, single-unit responses are shown in the first and fourth columns, and intracellular responses are shown in the second and third columns. Individual units in the trigeminal ganglia respond reliably up to 1200 Hz stimulation of the vibrissa (Gottschaldt KM and Vahle-Hinz C (1981) *Science* 214: 143–186), and essentially all units respond reliably to 300 Hz. By the anatomical level of the ventral posterior medial nucleus of dorsal thalamus (VPM), the responses are phase locked to the stimulus but rather infrequent; that is, most spikes are dropped. By the level of the PrV, the responses are still phase locked to the stimulus but drop-outs (*) are significant by 350 Hz. At the level of cortex, the drop-out rate is apparently so high that phase-locked spiking is not apparent at high frequencies, yet for stimulation frequencies near 100 Hz and above, the neurons appear to fire in a largely asynchronous manner, that is, as an inhomogeneous Poisson process. Note that, at all levels, the onset of stimulation always leads to transient activation. Data in the first three columns from Deschenes M, Timofeeva E, and Lavallee P (2003) The relay of high-frequency sensory signals in the whisker-to-barreloid pathway. *Journal of Neuroscience* 23: 6778–6787, and data in the final column from Arabzadeh E, Petersen RS, and Diamond ME (2003) Encoding of whisker vibration by rat barrel cortex neurons: Implications for texture discrimination. *Journal of Neuroscience* 27: 9146–9154. The responses are different in aroused animals or anesthetized animals with activation of their cholinergic system; the responses in VPM thalamus (Castro-Alamancos MA (2002) Different temporal processing of sensory inputs in the rat thalamus during quiescent and information processing states *in vivo*. *Journal of Physiology* 539: 567–578.) extend to higher frequencies before attenuation, while the frequency dependence of the response in vibrissa S1 cortex is largely unchanged (Castro-Alamancos MA (2004) Absence of rapid sensory adaptation in neocortex during information processing states. *Neuron* 41: 455–464) and may further exhibit bursts of spikes (de Kock C and Sakmann B (in press) High frequency action potential bursts (100 Hz) in L2/3 and L5B thick tufted neurons in anaesthetized and awake rat primary somatosensory cortex. *Journal of Physiology*; Ewert TA, Vahle-Hinz C, and Engel AK (2008) High-frequency whisker vibration is encoded by phase-locked responses of neurons in the rat's barrel cortex. *Journal of Neuroscience* 14: 5359–5368).

Sensory Representation of Texture

Studies on the coding of texture typically make use of irregular surfaces of differing roughness, such as sandpapers, as stimuli. In one paradigm, artificial whisking with anesthetized animals is used to ensure a repeatable pattern of motion of individual vibrissae, whose detailed positions can be recorded with high-speed videography (Figure 9). Artificial whisking makes use of electrical stimulation of the facial nerve, with trains of pulses to induce rhythmic sweeps. Yet artificial whisking provides a practical means to study the variation in vibrissa movement across surfaces for the same muscular drive. Further, once the motion is recorded, one can use the stored waveforms to control piezoelectric manipulators that move the vibrissa in a repeatable manner. The caveats of this approach are twofold: first, the resultant motion is abnormal in that the three phases of natural motor activity (Figure 6(c)) are reduced to one and, second, sensation may be altered in that extrinsic movement of the vibrissae approximates self-induced movement.

The motion of the vibrissae across different sandpapers in the artificial whisking paradigm has been characterized in two ways. First, variations in motion were quantified in terms of the root-mean-square (RMS) speed of the shaft of individual vibrissae. The RMS speed initially increased in a progression from smooth to rough surfaces, then reached an asymptote as the size of the surface features reached ~35 μm (Figure 9(a)). Concomitant with this asymptote in

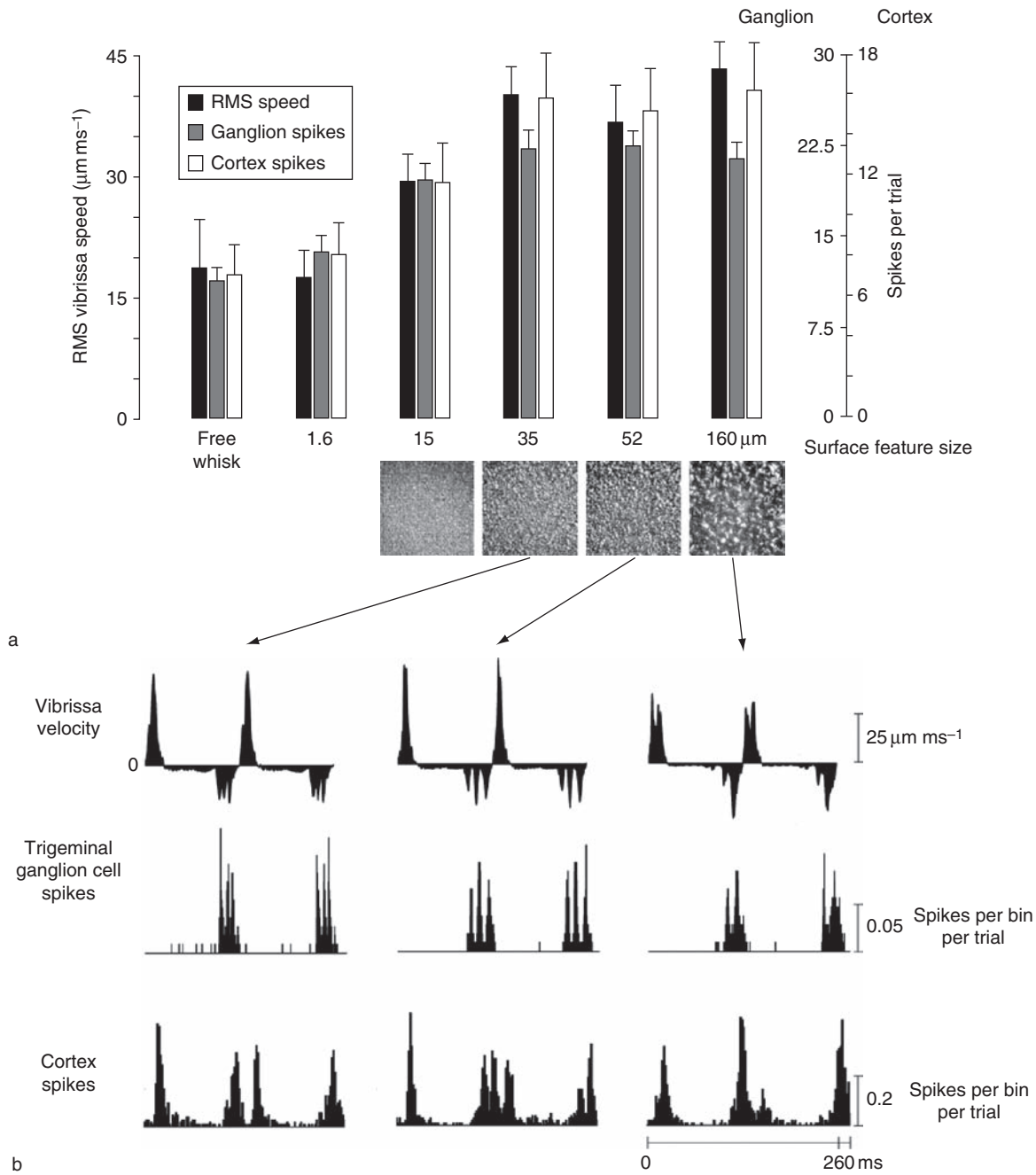

Figure 9 Evidence that vibrissa motions induced by textured surfaces are encoded by both average spike rate and temporal variations in the rate. (a) Texture coding by the average rate of neuronal firing. The spike rate is plotted for texture-induced vibrations, quantified by the root-mean-square (RMS) speed of the horizontal vibrissa motion, averaged across 100 trials of 500 ms each. The average spike count per trial was pooled from ten trigeminal ganglion neurons and 12 neurons in vibrissa S1 cortex (granular layer); note the different scales for spike counts. The insets show the surfaces. (b) Texture coding by patterns in the firing rate for three examples. The top line shows variations in velocity of a vibrissa as it moves across one of three different textures. The second line shows the peristimulus time histograms, with 0.2 ms bins, for a unit in the trigeminal ganglion. The final line shows the peristimulus time histograms for a unit in S1 cortex. Each profile is the average of 100 trials. Data adapted from Arabzadeh E, Zorzin E, and Diamond ME (2005) Neuronal encoding of texture in the whisker sensory pathway. *PLoS Biology* 3(e17): 155–165; the discrimination of different textures is discussed by von Heimendahl M, Itskov PM, Arabzadeh E, and Diamond ME (2007) Neuronal activity in rat barrel cortex underlying texture discrimination. *PLoS Biology* 5: e305.

speed, the spike rate of units in both the trigeminal ganglion and vibrissa S1 cortex also achieved an asymptotic value. This implies that average measures cannot distinguish among different coarse surfaces. In contrast, a second analysis considered the full time dependence of the velocity for motion across surfaces of different roughness. This measure exhibited clear changes as surface features doubled and tripled in size

(Figure 9(b)). Further, neurons in the trigeminal ganglion faithfully tracked these changes, with directional preference, consistent with their ability to track high-frequency vibrations (Figure 8). At the level of S1 cortex, there were repeatable differences in the spike rate for all sandpapers, although many details of the timing were not tracked (Figure 9(b)).

The data suggest that the motion of a vibrissa across a textured surface is dominated by stick–slip events (Figure 7(b)), this conclusion has been partially verified with awake animals in natural textural settings. Further, while primary sensory cells can encode different textures, the algorithm for encoding a natural ensemble of textures in terms of cortical spike rates remains an open issue, as does the issue of the fidelity of that encoding.

Sensory Representation of Object Location

Neuronal Signals of Vibrissa Position

There is a lack of evidence for spindle fibers, the transduction mechanism for proprioception, in the facial musculature, as well as a lack of evidence for descending inputs to primary sensory neurons. Nonetheless, the results of experiments with the artificial whisking paradigm report the presence of reference signals for vibrissa position in a subset of primary sensory neurons (left panel in Figure 10(a)). Different neurons preferentially spiked at different angles during the whisk cycle (right panel in Figure 10(a)), with a bias toward protraction from the retracted position. These data, together with

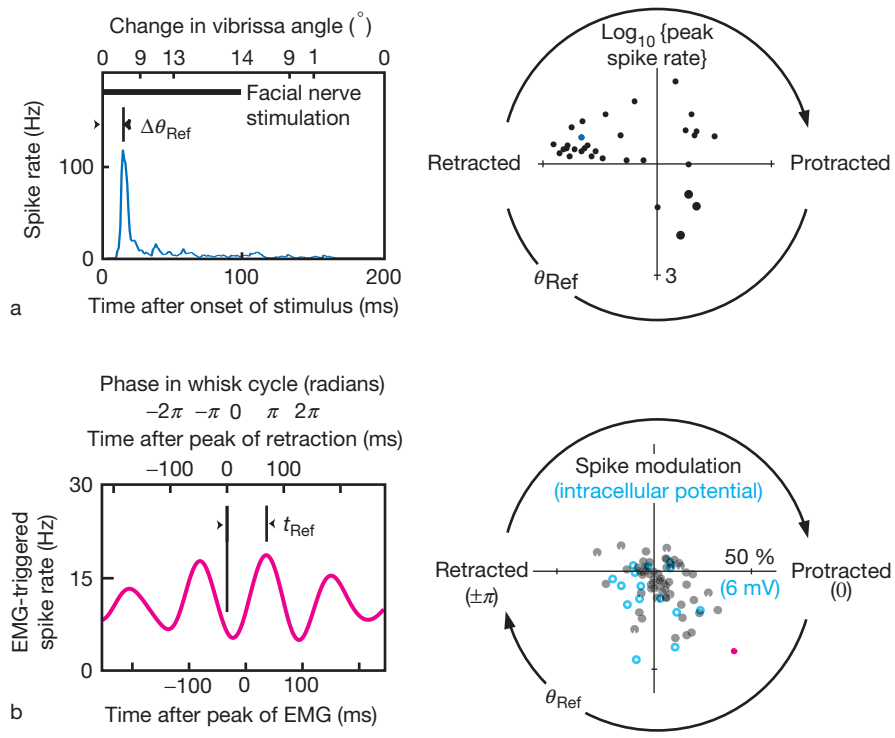

Figure 10 Encoding of vibrissa position in the hindbrain and forebrain. (a) Spiking output from primary sensory neurons of the trigeminal ganglion was recorded during vibrissa motion induced by 5 Hz electrical stimulation of the facial motor nerve. The left panel shows the trial-averaged response of the reference signal for one neuron; the angle $\Delta\theta_{Ref}$ is the extent of protraction, relative to the initial retracted position, at the peak of the neuronal response. The right panel summarizes the data for all cells. From Szwed M, Bagdasarian K, and Ahissar E (2003) Encoding of vibrissal active touch. *Neuron* 40: 621–630. The radial coordinate is the logarithm of the peak spike rate, and the angular coordinate is the phase within the whisk cycle. (b) Responses in vibrissa S1 cortex measured as animals whisked without contact. The data on the left show an example of the correlation between spiking and the peak of the electromyogram (EMG). The scale on top accounts for the lag between vibrissa position and the EMG (Berg RW and Kleinfeld D (2003) *Journal of Neurophysiology* 89: 104–117); the time t_{Ref} is the peak of cortical spiking relative to the fully retracted position. The panel on the right summarizes the data for all cells from these extracellular measurements, for which the radial coordinate is the modulation of the spike–EMG correlation, as well as cells from recent intracellular studies, for which the radial coordinate is the maximum intracellular depolarization. The bias in spiking as a function of phase in the whisk cycle comes slightly earlier than that for the membrane depolarization, as expected for a threshold process. Data from Fee MS, Mitra PP, and Kleinfeld D (1997) Central versus peripheral determinants of patterned spike activity in rat vibrissa cortex during whisking. *Journal of Neurophysiology* 78: 1144–1149; Mehta SB and Kleinfeld D (2004) Frisking the whiskers: Patterned sensory input in the rat vibrissa system. *Neuron* 41: 181–184; and Crochet S and Petersen CCH (2006) Correlating whisker behavior with membrane potential in barrel cortex of awake mice. *Nature Neuroscience* 9: 608–609.

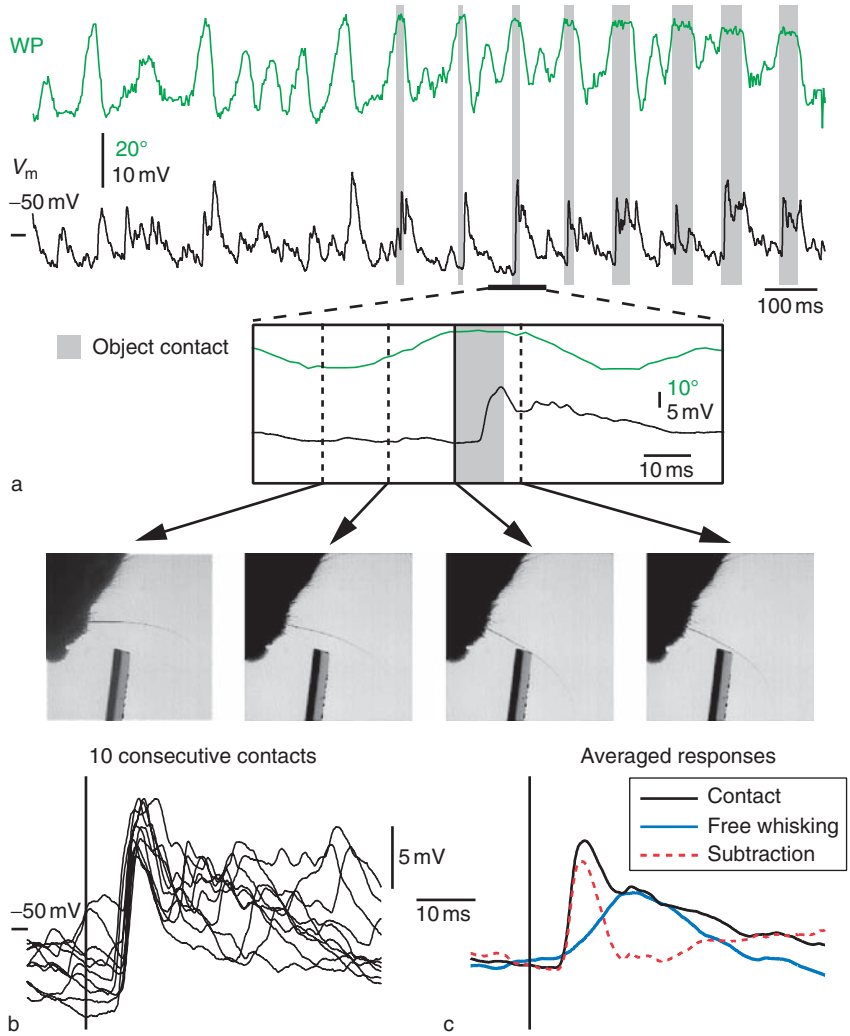

Figure 11 Intracellular potential of neurons in primary vibrissa cortex of a head-fixed mouse during active touch. (a) The mouse whisked in air both without contact (left sides of top two traces) and with contact against an object (right sides) when it thrust its vibrissae forward. The intracellular potential is marked by V_m (black), the videographed vibrissa position is marked by WP (green), and contact is marked by a gray bar. Note that the rest level of the cell is relatively depolarized. The box corresponds to an expansion of one contact event, with videographs of the contact event shown below. (b) The superposition of ten consecutive contacts shows the reliability and temporal precision of the intracellular response. (c) Intracellular responses evoked by contact with an object were averaged (blue) and compared with the average response aligned with respect to the peak of the vibrissa position for whisking in air (WP in (a)). The response evoked by contact was faster and larger than that evoked by free whisking; dashed red line marks the subtraction of the two responses. Adapted from Crochet S and Petersen CCH (2006) Correlating whisker behavior with membrane potential in barrel cortex of awake mice. *Nature Neuroscience* 9: 608–609.

nerve block experiments, imply that the position signal originates from peripheral reafference.

At the level of vibrissa S1 cortex, experiments with awake animals trained to whisk in air show that the spike rates of individual cortical neurons are locked to the rhythmic motion of the vibrissae, as inferred from the electromyogram (EMG) of the intrinsic muscles (left panel of Figure 10(b)). This constitutes a reference signal of vibrissa position. The distribution of preferred angles extends over the entire whisk cycle, with a bias toward retraction from the protracted position. Further, intracellular measurements with head-fixed awake animals find a similar distribution of preferred angles for the peak of the excitatory postsynaptic potential. The bias in spiking as a function of phase in the whisk cycle comes earlier than that for the membrane depolarization, as expected for a threshold process. Collectively, these data imply that rats ascertain the position of their vibrissa at the time of contact.

Finally, the fidelity of the reference signal in cortex depends on reward. Whisking in air that is coupled to a food reward is a factor of 2 to 3 times more coherent with vibrissa position than whisking in air that is

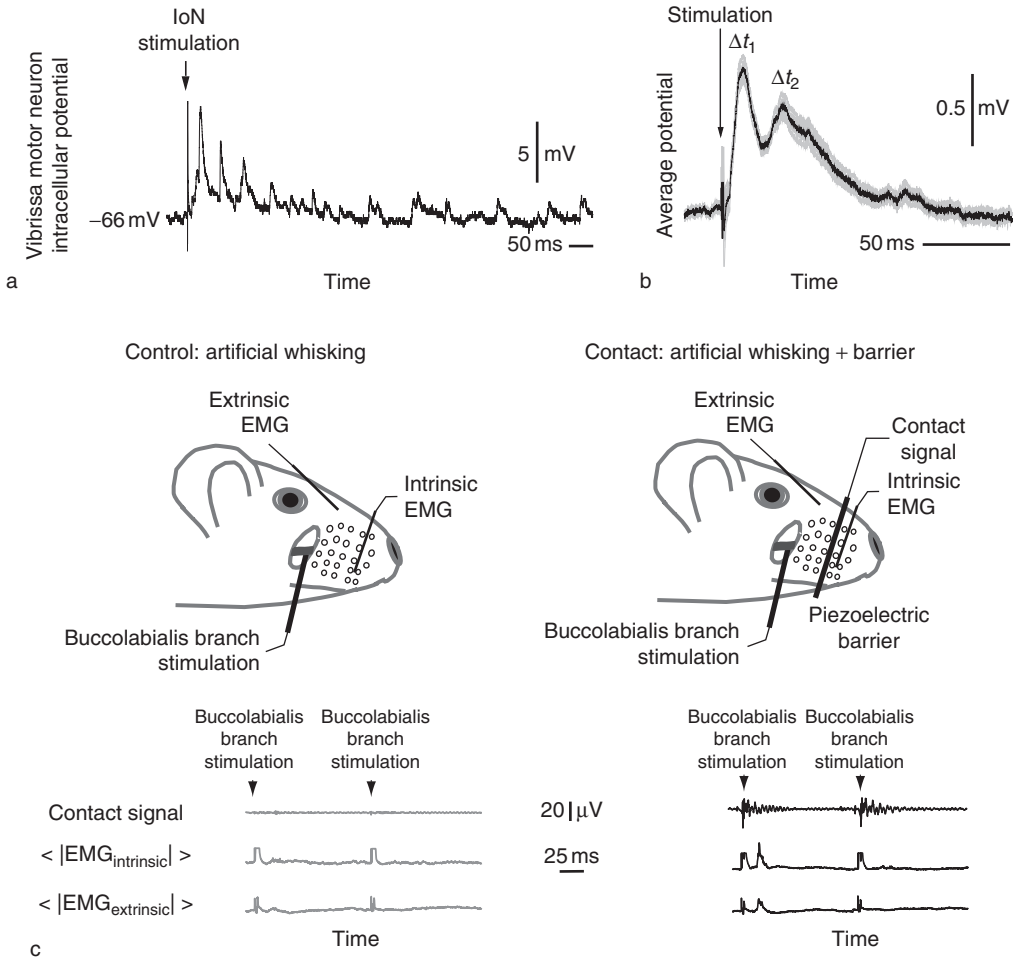

Figure 12 Sensorimotor feedback in the hindbrain (brain stem) loop. This closed loop is formed by the vibrissae follicle → trigeminal ganglion → trigeminal nuclei → facial nuclei → follicle musculature pathway; trigeminal nucleus principalis and spinal nucleus interpolaris are likely to mediate the trigeminal pathway (**Figure 6**). (a) Data from a brain stem slice preparation of juvenile animals (postnatal days 10–15) that preserves the connectivity, except that excitation of sensory terminals is replaced by shock to the infraorbital nerve (IoN), which excites vibrissa neurons in the trigeminal ganglion, and muscle output is replaced by intracellular recording from facial motor neurons in the lateral aspect of the facial nucleus. Shown are excitatory postsynaptic potentials in a motor neuron following suprathreshold stimulation (average of 20 responses) of the IoN in a slice. (b) Trial and preparation averaged intracellular response to nerve stimulation. The responses from ten trials each of 25 motor neurons were averaged together; mean, black trace; standard error of measurement, gray band. (c) The electromyogram (EMG) responses elicited by vibrissa contact in a ketamine xylazine-anesthetized adult rat. The cartoon in the top row shows the experimental setup for the contact experiment during artificial whisking (**Figure 9(a)**), for which the facial motor nerve (buccolabialis branch, driving the retractors *M. nasolabialis* and *M. maxillolabialis*; **Figure 6(a)**) is stimulated at 8 Hz with 9 mA, 50 s pulses to make the vibrissae protract. Shown are the contact signals (top traces, with control trials on the left and contact trials on the right) and consecutive averaged rectified EMG responses recorded simultaneously in intrinsic (middle traces) and extrinsic (bottom traces) vibrissa muscles. Note that the contact-induced EMG signals adapt after a single contact. Each trace is the average of 30 sweeps. Data from Nguyen Q-T and Kleinfeld D (2005) Positive feedback in a brainstem tactile sensorimotor loop. *Neuron* 45: 447–457.

unrewarded. This shows that reward, which is likely to heighten the attentional state of the animal, can strongly bias the nature of its sensorimotor processing.

Whisking-Dependent Contact Signals

A reliable and robust response to contact is observed at the level of the trigeminal ganglion in awake animals. The spike rate of primary sensory cells increases with increasing impact, or velocity, of contact until a saturation level is reached for angular speeds on the order of $500°s^{-1}$. There is also evidence that the radial distance to the contact point is encoded independent of the velocity.

Experiments with anesthetized animals indicate that there are robust spike responses to the initial deflection of a vibrissa along the entire ascending lemniscal pathway. Yet a systematic study of spiking at

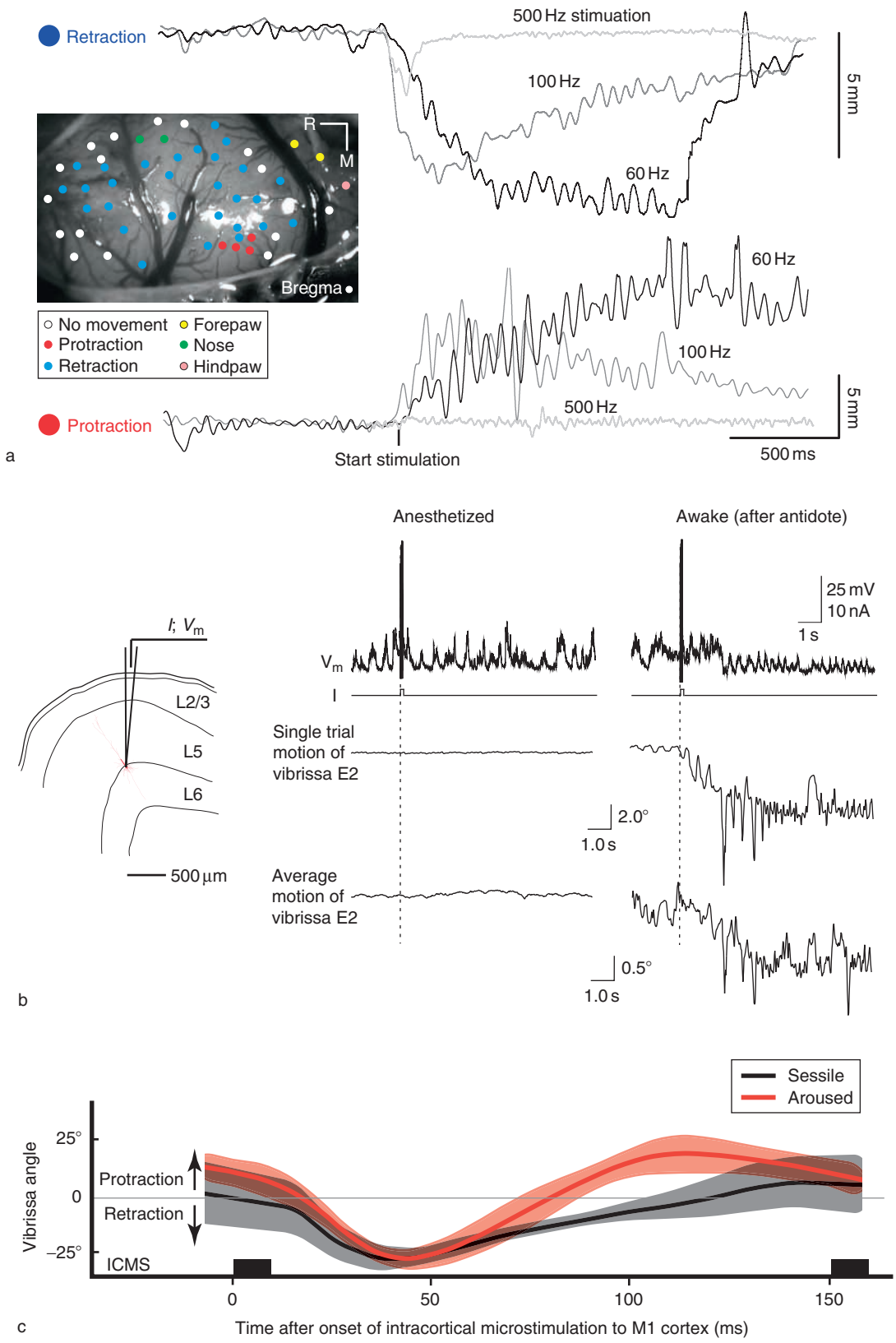

Figure 13 Real-time control of vibrissa position by primary motor cortex. (a) Intracortical microstimulation (ICMS) of vibrissa M1 cortex leads to retraction at a majority of stimulation sites (blue dots and associated traces at top; magnification of 3 by 5 mm image) but protraction in a medial region of frontal cortex (red dots and associated traces at bottom of image). The amplitude of the response as a function of the frequency of the microstimulation was greater for low vs. high-frequency stimulation (cf. 60 vs. 500 Hz). Data from

all levels in the sensorimotor hierarchy in awake animals remains to be completed, including the essential issue of spiking induced by animal-initiated contact. Nonetheless, preliminary observations of the nature of the subthreshold input to neurons in the upper layers of vibrissa S1 cortex have revealed the coexistence of whisking-related changes in subthreshold potential and changes in potential that are locked to contact (Figure 11). Contact signals that are referenced to vibrissa position could, in principle, be computed from the observed signals.

Sensory Modulation of Vibrissa Motion

The observation of changes in whisking on contact of the vibrissa with a wall (Figure 1(a)), along with changes in whisking amplitudes for whisking in air versus against a target (Figure 2(a)), provides evidence for changes in whisking strategy based on the task. How can this occur?

Brain Stem Mechanisms

Recent work shows that the brain stem loop exerts a transient positive feedback (Figure 12). At the level of *in vitro* physiology, sensory feedback from the trigeminal input to the motor neurons was measured in a brain stem slice preparation that preserves the vibrissa sensorimotor feedback pathway; see brain stem loop in Figure 4. Stimulation of the IoN leads solely to excitatory postsynaptic potentials in the motor neurons of the facial nucleus (Figure 12(a)). The latency of the response has two components, a fast Δt_1 component that presumably reflects disynaptic (IoN → trigeminal nuclei → facial nucleus) input and a slower Δt_2 component that is involved in trisynaptic or higher-order feedback in brain stem (Figure 12(b)). The critical issues are that feedback is both positive and rapid, that is, $\Delta t_1 \sim 10$ ms compared with the ~ 100 ms timescale of whisking.

At the level of *in vivo* physiology, the brain stem can be functionally isolated through the use of anesthesia. The artificial whisking paradigm is then used to drive the vibrissa muscles at physiological whisking rates. As a means to study sensorimotor feedback, a rigid contact detector is introduced in the trajectory of the vibrissae during alternate measurements to ascertain whether EMG responses could be elicited by contact of the vibrissae with the detector (Figure 12(c)). Consistent with positive feedback, contact led to a transient increase in the amplitude of the rectified EMG for both intrinsic and extrinsic muscles. This effect rapidly adapted for stimulation frequencies above 5 Hz. These data show that positive, transient feedback can lead to an increase in motor force.

Behavioral data to support transient positive sensorimotor feedback come from measurements of changes in whisking force coincident with contact. When animals whisked symmetrically yet made contact with a sensor on only one side of the face, the amplitude of the mystacial EMG on the contact side was transiently increased by 25%. These data show that the vibrissa feedback loops act to transiently increase the interval and force of contact.

Cortical Mechanisms

We conjecture that changes in whisking strategies, as opposed to transient changes in force, are regulated at the thalamocortical level. Three classes of experiments support this view. As a means to delineate the sensory representation of rhythmic input in M1 cortex, the first experimental effort made use of awake, head-fixed animals that were trained not to whisk. The observed response captured only the fundamental frequency of the input pattern for pulsatile stimuli delivered at frequencies that ranged from 5 to 20 Hz, that is, the range of exploratory whisking. This nonlinear transformation resembles the synthesis of a sinusoidal feedback signal for servocontrol of a motor.

A second class of experiments showed that vibrissa M1 cortex supports oscillations in extracellular current flow that are phase locked to whisking. These signals are preserved after lesions of the IoN to block sensory input. Further, ablation of M1 cortex leads to alteration of the whisking patterns. Thus, motor cortex generates rhythmic signals that can drive normal exploratory whisking.

Haiss F and Schwarz C (2005) Spatial segregation of different modes of movement control in the whisker representation of rat primary motor cortex. *Journal of Neuroscience* 25: 1579–1587. (b) Intracellular stimulation of a single neuron in M1 cortex with current I has little effect on the membrane potential V_m as well as on the motion of the vibrissae in the anesthetized state but leads to protraction of vibrissa E2 in the awake animal. Data from Brecht M, Schneider M, Sakmann B, and Margrie T (2004) Whisker movements evoked by stimulation of single pyramidal cells in rat motor cortex. *Nature* 427: 704–710. (c) Videographic data of vibrissa position in sessile vs. aroused animals in response to ICMS of motor cortex. Note that ICMS in the sessile case led to retraction of the muscles, whereas stimulation in the aroused case led to a whisk cycle with retraction followed by a protraction phase. Data from Berg RW and Kleinfeld D (2003) Rhythmic whisking by rat: Retraction as well as protraction of the vibrissae is under active muscular control. *Journal of Neurophysiology* 89: 104–117. Early studies also considered the direction of motion induced by ICMS to M1 cortex (Sanderson KJ, Welker W, and Shambes GM (1984) *Brain Research* 292: 251–260; Gioanni Y and Lamarche M (1985) *Brain Research* 344: 49–61).

A final class of experiments, highlighted in Figure 13, explored the motion of the vibrissa in response to activation of vibrissa M1 cortex in both anesthetized and behaving animals. From a functional perspective, there are two contiguous motor areas. For the largest region (blue dots in Figure 13(a)), activation of tissue with a brief train of extracellular current pulses, or even depolarization of a single cell, leads to retraction of the vibrissae. For a smaller region near the midline (red dots in Figure 13(a)), activation with a brief train of extracellular current pulses leads to protraction. It is interesting that stimulation of neurons in this region with a continuous train of high-frequency pulses, rather than a brief train, can induce rhythmic whisking.

The ability of M1 motor cortex to drive whisking is enhanced when rats are awake as opposed to anesthetized. Further, this enhancement is profoundly increased in the attentive, as opposed to sessile, awake state. First, stimulation of even a single projection neuron in vibrissa M1 cortex can lead to movement of one or more vibrissae. In the awake animals, this movement is dramatically increased (Figure 13(b)). Second, while stimulation of the major fraction of vibrissa M1 cortex leads to retraction when animals are anesthetized or awake but sessile, this motion is transformed into a full whisk, with protraction and retraction, when animals are in the awake and aroused state (Figure 13(c)). This effect is mimicked by cholinergic activation of cortex, a key component of attention. Collectively, these data show that M1 cortex, in principle, can subsume full control of vibrissa movement.

Epilog

The vibrissa system in rodents has a multitude of features that make it ideal for the study of sensorimotor control in a mammal. It preserves the basic architectonics of nested feedback loops (Figure 4), yet the lissencephalic cortex of rodents makes all structures of the brain very accessible. Critical for studies of active sensation, work to date has refined a number of behavioral paradigms that tax the active nature of this sensorimotor system and define strategies for processing the sense of vibrissa-based touch (Figures 2 and 3).

Electrophysiology studies with awake, behaving animals have begun to elucidate the algorithms involved in active sensing. For texture discrimination, this is likely to include the detailed timing of spikes in vibrissa S1 cortex (Figure 9). For touch in face-centered coordinates (Figure 3), this is likely to involve the merge of touch and reference signals of vibrissa position (Figures 10 and 11). Experimental progress is anticipated to be critical to a proper understanding of active sensing in two broad, open issues.

How Do Rats Map the Space about Their Vibrissa?

The main issue is whether rats maintain an abstract representation of objects that is independent of which vibrissa makes contact with an object. Or is this representation based solely on a single vibrissa (which would suggest little interaction between columnar representations)? A number of immediate issues within the realm of representation are ripe for study:

- There is evidence that rats encode texture by the time between stick–slip events. How is this used to form a perception? By analogy with flutter coding in somatosensory experiments in monkeys, do rats count the number of events in a period? Or are individual periods stored? Beyond this lies the issue of whether stick–slip events are encoded across different vibrissae.
- There is evidence that primary sensory neurons can encode the radial distance of contact. How is radial distance separated from impact on contact?
- The notion of object location depends on the presence of a neuronal response that encodes contact conditioned on vibrissa position in the whisk cycle. At what level of the sensorimotor pathways do such responses first occur?
- There is evidence that the phase of the vibrissae within a whisk cycle, rather than the absolute position, is encoded. How is this accomplished? One hypothesis is that it relies on adaptation at the level of the primary sensory neurons.
- Adaptation of the neuronal response at all levels of sensory processing, and particularly at the level of thalamocortical synapses, is strong. Thus steady state responses are much different from transient responses. How does this play into the representation of sensory signals?

How Do Rats Change Their Motor Output in Response to Vibrissa-Based Touch?

The main issue is how touch leads to changes in body position, whisking dynamics, and locomotion.

- Rhythmic whisking can occur in the absence of sensory feedback and in the absence of high-level control. Where are the pattern generators for rhythmic whisking? Further, what form of circuitry allows these generators to stabilize at a different, single frequency from bout to bout?

- The superior colliculus is capable of driving patterned, stereotyped motor behaviors. What is the role of the colliculus in transforming vibrissa sensory inputs into motor commands?
- Motor neurons can potentially receive feedback from sensory pathways at all levels of the brain. How are these inputs arbitrated?

See also: Barrel Cortex Circuits; Sensorimotor Integration: Attention and the Premotor Theory; Sensorimotor Integration: Barrels, Vibrissae and Topographic Representations; Sensorimotor Integration: Models.

Further Reading

Ahissar E and Zacksenhouse M (2001) Temporal and spatial coding in the rat vibrissal system. *Progress in Brain Research* 130: 75–87.

Alloway KD (2008) Information processing streams in rodent barrel cortex: The differential functions of barrel and septal circuits. *Cerebral Cortex* 18: 979–989.

Armstrong-James M (1995) The nature and plasticity of sensory processing within adult rat barrel cortex. In: Jones EG and Diamond IT (eds.) *Cerebral Cortex, Volume 11, The Barrel Cortex of Rodents*, pp. 333–374. New York: Plenum.

Brecht M, Grinevich V, Jin TE, Margrie T, and Osten P (2006) Cellular mechanisms of motor control in the vibrissal system. *Plflugers Archive* 453: 269–281.

Castro-Alamancos MA (2004) Dynamics of sensory thalamocortical synaptic networks during information processing states. *Progress in Neurobiology* 74: 213–247.

Deschenes M, Timofeeva E, Lavallée P, and Dufresne E (2005) The vibrissal system as a model of thalamic operations. *Progress in Brain Research* 149: 31–40.

Deschenes M, Veinante P, and Zhang Z-W (1998) The organization of cortico-thalamic pathways: Reciprocity versus parity. *Brain Research Review* 28: 286–308.

Diamond ME (1995) Somatosensory thalamus of the rat. In: Jones EG and Diamond IT (eds.) *Cerebral Cortex, Volume 11, The Barrel Cortex of Rodents*, pp. 189–220. New York: Plenum.

Diamond ME, von Heimendahl M, Knutsen PM, Klenfeld D, and Ahissar E (2008) "Where" and "what" in the whisker sensorimotor system. *Nature Review Neuroscience* 9: 601–612.

Ebner FF and Armstrong-James MA (1990) Intracortical processes regulating the integration of sensory information. *Progress in Brain Research* 11: 129–141.

Henderson TA and Jacquin MF (1995) What makes subcortical barrels? Requite trigeminal circuitry and developmental mechanisms. In: Jones EG and Diamond IT (eds.) *Cerebral Cortex, Volume 11, The Barrel Cortex of Rodents*, pp. 123–188. New York: Plenum.

Hermer-Vazquez L, Hermer-Vazquez R, and Chapin JK (2005) Role of plasticity in sensorimotor transformations. In: Ebner FF (ed.) *Neural Plasticity in Adult Somatic Sensory–Motor Systems*, pp. 139–154. Boca Raton, FL: CRC Press.

Keller A (1995) Synaptic organization of the barrel cortex. In: Jones EG and Diamond IT (eds.) *Cerebral Cortex, Volume 11, The Barrel Cortex of Rodents*, pp. 221–262. New York: Plenum.

Kleinfeld D, Berg RW, and O'Connor SM (1999) Anatomical loops and their electrical dynamics in relation to whisking by rat. *Somatosensory and Motor Research* 16: 69–88.

Kleinfeld D, Ahissar E, and Diamond ME (2006) Active sensation: Insights from the rodent vibrissa sensorimotor system. *Current Opinions in Neurobiology* 16: 435–444.

Moore CI and Andermann ML (2005) The vibrissa resonance hypothesis. In: Ebner FF (ed.) *Neural Plasticity in Adult Somatic Sensory–Motor Systems*, pp. 21–60. Boca Raton, FL: CRC Press.

Nelson ME and MacIver MA (2006) Sensory acquisition in active sensing systems. *Comparative of Physiology A: Sensory Neural and Behavioral Physiology* 192: 573–586.

Rice FI (1995) Comparative aspects of barrel structure and development. In: Jones EG and Diamond IT (eds.) *Cerebral Cortex, Volume 11, The Barrel Cortex of Rodents*, pp. 1–76. New York: Plenum.

Simons DJ (1995) Neuronal integration in the somatosensory whisker/barrel cortex. In: Jones EG and Diamond IT (eds.) *Cerebral Cortex, Volume 11, The Barrel Cortex of Rodents*, pp. 263–298. New York: Plenum.

Travers JB (1995) Oromotor nuclei. In: Paxinos G (ed.) *The Rat Nervous System,* 2nd edn., pp. 239–255. San Diego, CA: Academic Press.

Woolsey TA (1997) Barrels, vibrissae and topographic representations. In: Adelman G and Smith B (eds.) *Encyclopedia of Neuroscience*, pp. 195–199. Amsterdam: Elsevier Press.

Wrobel A, Kublik E, and Musial P (1998) Gating of the sensory activity within barrel cortex of the awake rat. *Experimental Brain Research* 123: 117–123.

Viral Vectors in the CNS

C N Cearley and J H Wolfe, University of Pennsylvania and Children's Hospital of Philadelphia, Philadelphia, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Viruses have evolved to be very efficient at entering and transferring their genes into eukaryotic cells. Viral vectors have been designed to take advantage of the efficiency of viral entry while removing or reducing the associated pathological functions. Recombinant viral vectors are made by removing the genes for most viral functions while retaining the sequences necessary for the recombinant genome to be packaged into a virus particle (the virion) and the elements needed for the desired viral functions within the target cell. The deleted sequences are replaced with a foreign gene and other elements needed for function of the transferred gene, such as promoter(s), splice signals, and polyadenylation signal. To produce recombinant viral particles, the minimum number of viral genes necessary for replication and assembly of the virion is provided in trans, usually encoded on separate plasmids from the transferred gene (Figure 1). This allows the construction of a replication-incompetent recombinant virion at the final stage of vector production.

The recombinant viral vectors enter target cells where the new foreign gene either is integrated into the host cell genome or exists episomally. This is usually a single-hit process, whereby the vector genome is transferred into only the target cell and does not spread further by viral replication; therefore, the concentration of the vector needs to be very high to reach a high number of cells. The concentration of vectors (the titer) is usually measured in genome copies per milliliter or as a functional measure of target cells expressing the transferred gene. The transfer of the vector-encoded gene to the target cells is called transduction.

Applications of Viral Vectors in the CNS

Gene transfer into the central nervous system (CNS) using viral vectors usually requires a surgical procedure (Figure 2) but has several advantages over other methods of genetic manipulation. The procedure involves a stereotaxic injection into specific sites but, in most cases, only a single injection of very small amounts of the concentrated vector is needed for expression of the transgene in the appropriate group of cells. An advantage over transgenic alterations is that the vector can be injected unilaterally, leaving the contralateral side as a nontreatment control. The transgene can be targeted to specific regions of the brain or to specific cell types. There are a variety of vectors available, each with advantages and disadvantages.

The most commonly used application of viral vectors in the CNS is gene therapy, which in most cases involves delivery of a normal copy of a defective gene. Viral vectors can also be used for overexpression or knockdown of genes or proteins to study their function, for visualization of specific cell types *in vivo*, for tract tracing, and for experimental manipulations during development. There are several things that need to be considered when designing such an experiment. Some vector types only transduce dividing cells, and the cells in the CNS are mostly nondividing. Some vectors transduce cells only near the injection site, whereas others are transported to distal sites; therefore, an understanding of the extent of gene delivery needed in an experiment is required. The length of time a gene needs to be expressed is important because some vectors result in only transient gene delivery. For therapeutic applications, consideration also needs to be given to the potential for generation of replication-competent viruses, immune responses, insertional mutagenesis, and germ-line transmission.

Many studies require the level of transgene expression to be regulated. Viral vectors are currently being developed that use drug-controlled molecular switches. The most common systems use tetracycline, rapamycin, ecdysone, or progesterone. In these systems, a transactivator is expressed from a tissue-specific promoter. In the presence of the inducible drug, the drug binds to the transactivator, and this allows transcription of the downstream gene. One potential problem is that these systems can be leaky. If the gene leads to the expression of trophic factors that are potent at very low concentrations, there may be an expression of biologically active amounts of the protein even without the presence of the inducible drug. Another method of regulating gene expression is to use a promoter that is upregulated only in specific cell types. Promoters can be used that are turned on only during development in order to upregulate certain genes during a specific stage of the life cycle.

Viruses Used in CNS Studies

The major classifications of viruses are by RNA or DNA genome and whether they are enveloped or nonenveloped. The major types of viruses being used for gene transfer to the nervous system are herpes simplex virus (HSV), adenovirus (Ad), adeno-associated virus (AAV), onco-retrovirus, lentivirus,

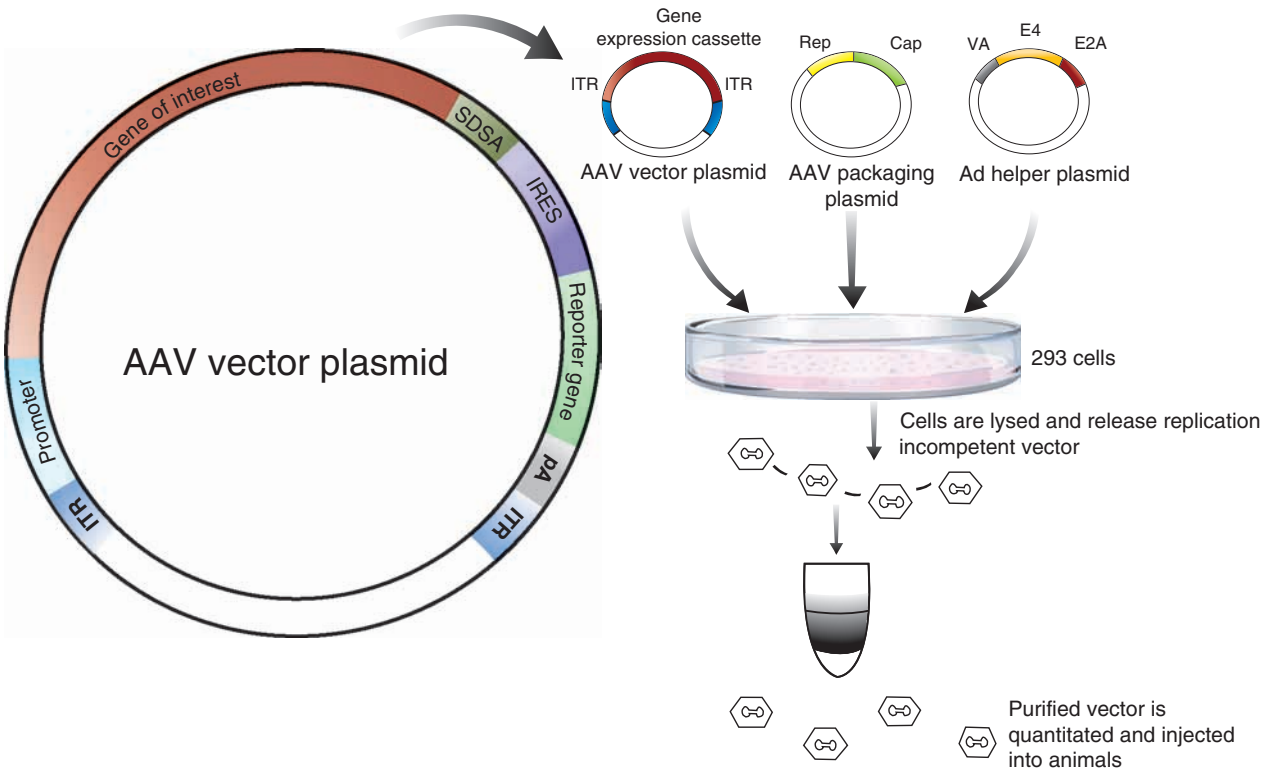

Figure 1 Example of viral vector design and production. To make an AAV vector, a three plasmid transfection system is used, consisting of the AAV vector plasmid, an AAV packaging plasmid, and an Ad helper plasmid. The AAV vector plasmid (enlarged view on the left) contains the gene of interest flanked by the inverted terminal repeats (ITRs), which are the only elements from the wild-type AAV included in the plasmid. The AAV packaging plasmid contains the AAV genes necessary for genome replication (Rep) and capsid protein formation (Cap), but these are not flanked by ITRs and thus are not incorporated into the virion. AAV is a dependovirus, which requires functions supplied by a helper virus such as adenovirus or herpes virus. For vector production, a helper plasmid encoding the minimum gene functions needed for replication is used. The three plasmids are transfected together into a packaging cell line, such as 293T cells, and normal cellular pathways are used to express the genes and proteins leading to the assembly of viral particles. The particles are released into the media after cell lysis or budding from the cell surface, depending on the type of virus. The vector particles are isolated from the lysate by cesium chloride gradient, pelleted by centrifugation, or column purified. The vector viral particles are then quantified and can be used for experimental purposes. Production of most vector viruses is done using a similar strategy, with the functions supplied in trans on plasmids or by helper viruses varying depending on the virus type. AAV, adeno-associated virus; Ad, adenovirus; SDSA, splice donor–splice acceptor; IRES, internal ribosomal entry site; pA, poly-adenylation signal.

and alphavirus, but some studies have used John Cunningham (JC) virus, Simian virus 40 (SV40), Sendai virus, and others. The six major vector types and their advantages and disadvantages for use in the CNS are discussed next; the major features for gene transfer are listed in Table 1.

RNA Viruses

Retrovirus Retroviruses (RVs) are enveloped RNA viruses that are reverse transcribed into DNA within the target cell and that integrate as double-stranded DNA proviruses into the host-cell genome. RVs bind host cell receptors via glycoprotein envelope proteins that are on the surface of the virion. Envelope proteins from some other viruses can be incorporated into the RV particle, a process known as pseudotyping, to change the cell-type specificity of the vector. There are two major classes of retroviruses: the onco-retroviruses and the lentiviruses.

The term RV vector is used in gene transfer studies to mean RVs from the onco-retrovirus family. These viruses were historically identified by their leukemia-inducing properties and thus are termed leukemia viruses, which have been isolated from mice (murine leukemia virus (MuLV)), cats (FeLV) gibbon apes (GALV), and others. The RV vectors transduce only dividing cells, so they are not able to transduce mature neurons *in vivo*. However, RV vectors are able to transduce stem and progenitor cells either in culture or *in vivo*, allowing the expression of the transgene in the progeny of the cells. This property has been used in the developing CNS to selectively label progenitor cells at different stages of development to follow their fate and to selectively deliver

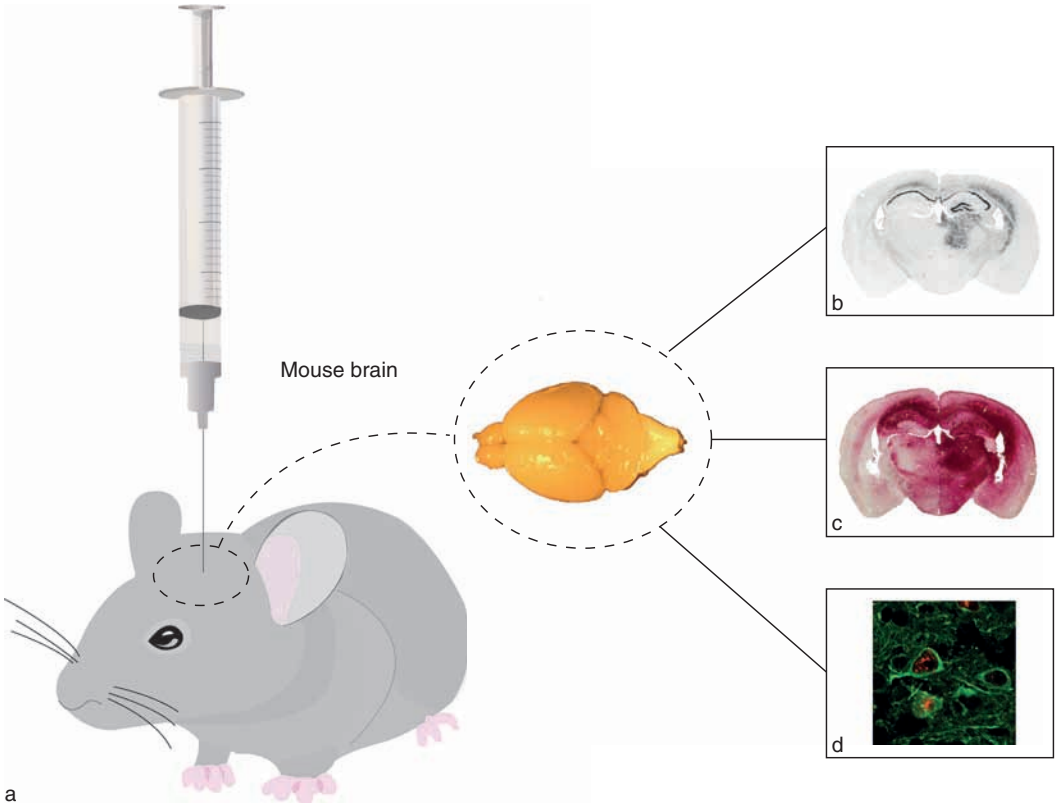

Figure 2 Gene transfer into the central nervous system (CNS) by direct injection and analysis of expression: (a) injection in mouse brain; (b) gene expression; (c) secreted protein; (d) cell-type specificity. In experiments involving the use of gene therapy vectors to treat disorders of the CNS, most vectors are directly injected into the brains of animals using a stereotaxic frame, and the brains are analyzed at a specified times postinjection. Vector analysis usually involves sectioning the brain and looking for gene expression using *in situ* hybridization and for protein expression using antibodies or functional analysis. Shown in (b–d) are sections from an animal that was injected with AAV 9 containing the gene for β-glucuronidase (GUSB) and then was sacrificed and examined 2 months postinjection. Gene expression (b) was examined using a DIG-labeled antisense riboprobe against the human GUSB gene. Protein expression (c) was examined using a histochemical reaction that allows the visualization of the enzymatic activity (red stain). In this disease, the dramatic enzyme expression beyond the gene-expressing cells is due to the ability of lysosomal enzymes to be secreted from a genetically corrected cell and to be either endocytosed by adjacent cells or transported to distal cells via axons. The neuronal specificity of the AAV 9 vector is shown in (d); fluorescent *in situ* hybridization using a Cy3-labeled riboprobe (red) is co-localized with a neuron-specific antibody, MAP2ab (green). AAV, adeno-associated virus; DIG, digoxigenin; MAP, microtubule-associated protein. (d) Reproduced from Cearley CN and Wolfe JH (2006) Transduction characteristics of adeno-associated virus vectors expressing cap serotypes 7, 8, 9, and Rh10 in the mouse brain. *Molecular Therapy* 13: 528–537, with permission from Elsevier.

toxic genes to tumor cells while sparing the nondividing normal cells.

RV vectors are most commonly used in *ex vivo* gene therapy, in which cells are modified in culture and then transplanted into the animal or patient. In the case of diseases in which the replacement of the gene leads to production of a secretable protein, this results in a source of therapeutic protein that can be secreted into the extracellular space, cerebrospinal fluid (CSF), or circulation to correct the disease at distal sites. *Ex vivo* gene therapy can also be used to genetically modify stem cells, which differentiate into more mature cells and may migrate within the CNS. Transduced hematopoietic stem cells and their progeny can circulate throughout the body, but their entry into the CNS is variable. An advantage of using *ex vivo* gene therapy is that the patient's or animal's own cells can be used, which circumvents heterologous immune responses.

Lentivirus The lentivirus (LV) is the second major subgroup of the retroviruses, and vectors have been developed from human immunodeficiency viruses (HIV-1 and-2), simian immunodeficiency virus (SIV), feline immunodeficiency virus (FIV), and equine infectious anemia virus (EIAV). The major advantage of LVs for the CNS is that they can transduce postmitotic cells such as neurons and, thus, can be injected directly into the brain. LV vectors become stably integrated into the host-cell genome and can mediate long-term expression of the transgene. Because LV vectors integrate into the host-cell

Table 1 Viral vectors commonly used for CNS applications

Virus class	*Virus currently being used for gene therapy applications*	*Serotypes/ pseudotypes used for CNS applications*	*Cells amenable to transduction*	*Foreign gene capacity*	*Stable genomic state*	*Virus type*	*Known receptor targets*
Retrovirus	MuLV	MuLV, VSV-G	Dividing cells only	6–7 kb	Integrated	RNA genome	Variable
Lentivirus	HIV-1, FIV, EIAV	VSV-G, rabies, Mokola, LCMV, Ross River, MuLV	Most cell types	7–8 kb	Integrated	Enveloped, single-stranded RNA genome	VSV-G-nonspecific membrane fusion; Ram-1/Pit-2 for MuLV; α-dystroglycan for LCMV
Alphavirus	Simliki forest virus and Sindbis virus	N/A	Neurons, but virus production in propagating cells only	7 kb	Replicon: genome amplified within cytoplasm	Enveloped, single-stranded RNA genome	Laminin and other receptors for SFV vectors, unknown for other strains
Herpes simplex virus	Amplicon, HSV-1, PRV	gD, gI variants	Neurons	150 kb for amplicon; 40 kb for replication defective HSV	Episomal	Double-stranded DNA genome	Herpes virus entry mediator A and nectin-1
Adenovirus	hAd5, cAd	Ad2, Ad5	Most CNS cell types	8 kb for replication defective; 30 kb for gutless	Episomal	Nonenveloped double-stranded DNA genome	Coxackievirus, adenovirus receptor, and integrins
Adeno-associated virus	AAV2, AAV5 ITRs	Most use serotypes 1, 2, 5, but 6, 7, 8, 9, Rh10 are promising	All serotypes transduce neurons; AAV1 also transduces oligoden-drocytes and AAV4 astrocytes	~4.5 kb	Mostly episomal, possibly integrated	Nonenveloped single stranded DNA genome	HS proteoglycans for AAV2; PDGF-R for AAV5; unknown for most

AAV, adeno-associated virus; Ad, adenovirus; cAd, canine adenovirus; CNS, central nervous system; EIAV, equine infectious anemia virus; FIV, feline immunodeficiency virus; h, human; HIV, human immunodeficiency virus; HS, heparan sulfate; HSV, herpes simplex virus; ITRs, inverted terminal repeats; LCMV, lymphocytic choriomeningitis virus; MuLV, murine leukemia virus; N/A, not applicable; PDGF-R, platelet-derived growth factor receptor; PRV, pseudo-rabies virus; Ram-1/Pit-2, amphotropic retrovirus receptor; SFV, Semliki Forest virus; VSV-G, vesicular stomatitis virus G-protein.

genome, they could potentially disrupt oncogenes, but they are not known to cause transformation by a promoter/enhancer insertion mechanism in natural infections. To further reduce this possibility, most LV vectors are designed to be self-inactivating (SIN) vectors, in which the long-terminal repeat (LTR) promoter and enhancer are deleted from the integrated provirus form. This also eliminates competition with the internal promoter in the expression cassette.

The HIV-1 vector is the most commonly used LV vector for CNS experiments. Only people with HIV-1 have circulating antibodies against this vector, making this an ideal vector to use when it comes to avoiding an immune response. The HIV-1 capsid proteins also appear to have a short half-life, reducing the time span in which the immune system can recognize the transduced cells.

Lentiviruses can transduce most cell types within the CNS, including neurons, astrocytes, stem cells, oligodendrocytes, and glioma cells. The host range can be expanded by pseudotyping LVs with surface glycoproteins from a number of other enveloped viruses, including vesicular stomatitis virus G (VSV-G,) MuLV, Mokola, lymphocytic choriomeningitis virus (LCMV), and rabies. Each pseudotyped vector transduces different subgroups of cells within regions of the CNS. The selectivity is mediated by variable expression of the cell-surface receptors to which theviral envelope protein binds. Thus, targeted transduction within the CNS can potentially be achieved using a combination of cell tropism and cell-type-specific promoters. There also is some evidence that LV vectors undergo retrograde transport, but this does not occur in all nervous system pathways.

Alphavirus The alphaviruses are enveloped and have single-stranded RNA genomes. Alphavirus vectors can be generated rapidly and simply, resulting in high-titer replication-deficient particles. The most commonly used alphavirus vectors for experiments in the CNS have been based on the Semliki Forest virus (SFV). Within the CNS, the SFV vector transduces mostly neurons at a rapid rate but with very transient transgene expression (a few days). Because of the transient expression, SFV vector use in the CNS is usually restricted to slice culture experiments. Wild-type SFV vectors give high infection rates in primary cultures of hippocampal neurons and in hippocampal slice cultures. SFV vectors can be cytopathic due to a shut off of host-cell protein synthesis, but this appears to be dependent on the cellular differentiation state of the CNS cells. Immature neurons are highly susceptible to SFV cytotoxicity, which occurs rapidly, but SFV vectors are able to persist for weeks in the highly differentiated cells of the adult mouse brain. Because alphaviruses can target a wide variety of cell types, an alternative to SFV viruses or other alphaviruses is to use the glycoproteins from the coats of these viruses to pseudotype retroviral and lentiviral vectors, although attempts to do this have resulted in very low titers thus far.

DNA Viruses

Herpes virus Herpes viruses have a large double-stranded DNA genome, and many of the genes are not required for replication: thus, herpes vectors can potentially hold large inserts. Human HSV-1 is the most commonly used herpes virus vector system for experiments involving the CNS, although porcine pseudo-rabies virus (PRV) has been used for tract tracing studies. HSV-1 vectors establish an episomal latent infection in neurons, thus insertional mutagenesis does not occur.

There are two types of HSV-1 vectors: recombinant replication-competent vectors and nonreplicating amplicons. Both types result in gene delivery to CNS cells. Recombinant HSV vectors consist of a wild-type HSV genome, selected either for loss of viral pathogenic effects or for deletion of genes necessary for replication. The deleted genes can either be supplied in trans using a plasmid system or can be expressed in a eukaryotic cell line that can be used for packaging the recombinant vectors. Recombinant HSV vectors can potentially accommodate large inserts of 30 kb or more and can be made into relatively high-titer vectors. Because HSV-1 is a naturally neurotropic virus, only low titers are needed for spread of the replicating virus in the CNS. Recombinant HSV vectors are thought to establish a persistent latent state as a non integrated nuclear element, similar to wild-type HSV. In natural infection, after latency is established, only one viral gene is expressed, the latency-associated transcript (LAT), which is expressed for the lifetime of animals. Thus, the promoter of the LAT sequence has been used to drive long-term expression of therapeutic genes in the CNS.

The HSV amplicon vector system is based on a eukaryotic expression plasmid that can be packaged either using a helper plasmid (viral genes supplied in trans) or a helper virus-free packaging method. HSV amplicons have a potential packaging capacity of up to 130 kb. Because the HSV amplicon vector can accommodate such a large insert, several copies of the transgene can be packaged into a single vector, forming concatamers, resulting in an increased transgene dose per infected cell. A disadvantage in using the amplicon-based system is that these vectors result only in relatively limited transgene expression and

expression in dividing cells is transient because vector DNA is lost during mitosis. For this reason, hybrid vectors have been developed which combine components of the HSV-1 amplicon with genetic elements from other types of viral vectors. HSV–AAV hybrid vectors have been made that combine the benefits of both vector systems, resulting in a vector that can package very large inserts, target multiple cells types, and result in high expression levels for an extended period of time.

Adenovirus Ads are nonenveloped with a double-stranded DNA genome. Ad vectors have both significant advantages and disadvantages for gene therapy experiments. Advantages include their simplicity to generate; large transgene capacity; and ability to transduce a variety of cell types, both dividing and postmitotic, including neurons, astrocytes, oligodendrocytes, ependyma, and blood vessels. Transduction using Ad vectors results in high levels of transgene expression. Another potential advantage is that Ad vectors are transported along neuronal pathways in some cases.

The main disadvantage of using Ad vectors is that they are highly immunogenic, which is the main contributor to loss of transgene expression after administration. Ad vectors target dendritic cells, which facilitates immune stimulation. Most humans carry circulating antibodies against adenoviruses, due to natural infection, which is a potential barrier to clinical use. One approach to reduce this effect is to alter the capsid protein to less-immunogenic forms. Another is to use a gutless vector, which has a genome deleted of all the Ad proteins, which are supplied in trans to form a virus. Gutless Ad vectors have mediated, stable, long-term expression of a transgene in the CNS and can accommodate an insert of approximately 30 kb, whereas replication-defective Ad vectors carry approximately 8 kb.

Adeno-associated virus AAVs were discovered as contaminates of laboratory preparations of Ads. AAVs are parvoviruses, which are nonenveloped and have a single-stranded DNA genome. AAVs are naturally replication-defective and can only replicate when co-injected with a helper virus (Ad or HSV). AAV vectors are one of the most widely used gene therapy vectors in CNS experiments. When directly injected into the brain AAV vectors transduce mostly neurons, but some serotypes can also transduce other neural cell types. AAV vectors result in a delayed-onset but persistent transgene expression. The lag in onset is due to the time it takes for the conversion of the single-stranded AAV genome into a double-stranded transcription-competent DNA molecule. AAV vectors can be concentrated and purified to very high titers, resulting in widespread and stable transduction with low toxicity.

The limitations of AAV vectors include their small cloning capacity (~4.5 kb), preventing their use for large therapeutic genes or limiting the use of many genetic regulatory elements. AAV vectors are relatively inefficient for *in vitro* experiments compared to other vectors, but they can transduce a number of neural cell types. AAV vector production is also labor intensive and time consuming. Recent studies suggest that the use of AAVs in humans may be limited by the existence of preexisting antibodies to AAV, although AAV vectors induce a minimal immune response when injected directly into the brain. When an animal's immune system is primed for a particular AAV serotype, transgene expression is severely diminished and transduction by later injections of the vector is prevented by the immune response.

There are numerous serotypes of AAV capsid genes available, each a sequence variant of the capsid protein. Each serotype has been shown to have different transduction characteristics in the CNS. The AAV-2 serotype vector has been widely for CNS applications but has relatively low transduction. This capsid protein uses heparan sulfate proteoglycan as a receptor, resulting in rapid local uptake by cells in the injection site, which restricts spread within the brain. There is growing use of several other serotypes in the CNS, especially serotypes 1, 5, 6, 7, 8, 9, and Rh10, and many new variants are being tested. Different serotypes can produce different patterns of transduction when injected into specific regions of the CNS, which can be used to selectively target certain cells. There is evidence that the AAV vector serotypes 1, 5, and 9 are transported within the CNS along some neuronal pathways.

Experimental Applications for Viral Vectors in the CNS

Viral vectors have many other CNS applications in addition to therapeutic purposes. Vectors can be used to alter genes in CNS cells in order to study the function of the particular gene. Viral vectors can also deliver fluorescent proteins to a targeted subset of cells, allowing visualization of these cells *in vivo*.

Modeling CNS Neurodegeneration in Animal Models

Modeling a human disease phenotype in a mouse brain has proven to be difficult for many diseases. Germ-cell alterations often do not show the selective pathology evident in human CNS tissue. To most

accurately model a human disease in a mouse, only the cell types affected in the human tissue should be affected in the mouse. One way to achieve this is to transduce the specific cell types with a targeted viral vector expressing the disease-causing transgene. Examples of diseases currently being modeled include Huntington disease (HD), by expressing the mutated human huntingtin protein with extended glutamine repeats in striatal neurons, and Parkinson disease (PD), by expressing α-synuclein in nigral dopamine-containing neurons.

Targeted transduction of specific cells and subregions of the CNS can be achieved by a combination of cell-surface targeting of virus entry and the use of cell-type-specific promoters. The most widely used variants for cell-surface entry are LV pseudotypes and AAV serotypes. However, tropism to specific brain structures usually cannot be predicted and must be determined empirically because the receptors for most of the envelope or capsid proteins are unknown. Once the vector is known to enter a specific cell type or cells in a specific layer or nucleus, further restriction of expression can be achieved by using a the promoter of a gene known to be expressed in the target cell and not adjacent cells. This may have high utility for experimental manipulation of specific genes (e.g., by siRNA silencing). The advantages of this approach over germ line manipulations are that the genetic modification can be done unilaterally, allowing comparison with the contralateral side, and that it is done in the mature cell, avoiding compensatory mechanisms that may be operating during development.

Viral Vectors to Understand CNS Development and Organization

Visualization of the development of the cells can be achieved by using a vector with long-term expression of a reporter gene. This has been exploited primarily to map the fate of cells transduced at the time of injection, including in the subventricular zone, hippocampus, and retina. Another use exploits the property that some viral vectors are transported along neural pathways, allowing tract tracing within the CNS or from the peripheral nervous system (PNS) to the CNS. The herpes virus PRV has been used to map neuronal pathways by immunostaining for the viral proteins, which are expressed at high levels for several days after injection, or by expression of lacZ as a reporter molecule from an acute-phase protein promoter. A replication-competent virus is required to obtain the transneuronal movement needed for tract tracing, but three to four orders of neurons can be labeled by this method.

CNS Diseases as Potential Targets for Gene Therapy

Gene therapy is currently being used to treat and study a number of neurological diseases. The most likely targets for gene therapy are diseases in which a single gene defect affects the CNS or diseases in which the presence of the corrected gene leads to a secretable protein. Gene therapy is also being investigated as a treatment for neurodegenerative diseases by augmentation of neurotransmitter receptors, restoration of neurotransmitter levels, inhibition of the neurodegenerative process, and stimulation of regeneration.

Parkinson Disease

PD is due to the loss of dopaminergic neurons of the substantia nigra. In the past, treatment of PD has involved transplantation of dopamine-expressing cells into the striatum. Problems with this method include the disturbance of striatal circuitry caused by the engraftment and the short-term survival of cellular grafts. Gene therapy represents a more long-term approach that will leave the striatal circuitry undisturbed. Current gene-therapy experiments involve the delivery of trophic factors such as glial-derived neurotropic factor (GDNF) to prolong the life of existing dopaminergic neurons, the delivery of tyrosine hydroxylase (TH) or amino acid decarboxylase (AADC) to the substantia nigra to promote dopamine production within the region, and expression of glutamate decarboxylase (GAD) to alter the neurotransmitter imbalance within the basal ganglia circuits. In the rarer cases in which PD is caused by a loss-of-function mutation, gene therapy might be used to replace the defective gene.

Alzheimer Disease

In Alzheimer disease (AD), neurons in the CNS undergo widespread degradation, especially cholinergic neurons in the basal forebrain. One approach is to use vectors expressing the gene for nerve growth factor (NGF), which can rescue cholinergic neurons from lesion-induced degeneration. However, the direct infusion of NGF into the brains of patients has caused adverse effects due to the resulting presence of NGF in nontargeted structures of the brain. For NGF to be useful therapeutically, it needs to be targeted only to the basal forebrain specifically, which can be done using gene therapy. AAV and LV vectors have been used for this purpose, and the degeneration of cholinergic neurons is prevented in animal models with this method of treatment. Viral vectors have also been used to increase the expression

of Nep (an extracellular amyloid-β-degrading enzyme) in the brain. Another approach to the treatment of AD using gene therapy is to use a viral vector expressing a functional RNAi to knock down expression of amyloid-β proteins by targeting β- and γ-secretases, which are responsible for amyloid-β production. Most studies have used LV vectors for RNAi experiments, and LV and AAV vectors for the other treatment methods.

Expanded Trinucleotide Repeat Disorders

A whole class of diseases has been discovered that are caused by variable numbers of trinucleotide repeats, usually causing a dominant negative effect. These include HD, spinocerebellar atrophies, and Kennnedy disease.

HD is caused by expanded CAG trinucleotide repeats which encode for glutamine in the huntingtin gene. This is a toxic gain-of-function mutation, which causes the loss of striatal neurons, resulting in cognitive and motor deficits. Gene therapy techniques to treat HD have included viral-vector-mediated delivery of ciliary neurotropic factor (CNTF), GDNF or brain-derived neurotropic factor (BDNF), all neuroprotective agents, with the goal of extending the life of existing striatal neurons. Another approach that is currently being attempted involves silencing the expression of the mutant allele by using a viral vector encoding an RNAi specifically against the mutant allele, leaving the normal huntingtin allele functional.

Inborn Errors of Metabolism

Genetic disorders can be the most difficult to treat in the brain because the lesions are typically globally distributed. This occurs because the mutations usually affect metabolic pathways that are common to all cells or all cells of a specific type. Diseases in which the corrected gene leads to the secretion of a therapeutic protein are potentially the most amenable to correction using viral vectors because not every cell has to be directly treated. This category includes the lysosomal storage disorders (LSDs), which have defects in structural lysosomal enzyme genes.

Each LSD results from a significant decrease or complete absence in the activity of a soluble lysosomal enzyme caused by a mutation in a single gene. Most lysosomal enzymes are ubiquitously expressed, and although most of the body can be at least partially treated using bone marrow transplants or enzyme replacement therapy, CNS treatment is more difficult because the enzyme does not easily cross the blood–brain barrier. In lysosomal storage disorders, cells have the ability to undergo cross-correction, which involves genetic correction of a group of cells that can secrete the enzyme, which can be taken up by neighboring cells, usually via endocytosis. Thus, only a subset of the cells needs to be directly corrected, and these then act as a source of normal enzyme to neighboring cells.

Many animal models of LSDs have been treated using several of the gene-therapy strategies to deliver the normal enzyme. *Ex vivo* gene therapy has been tested using RV or LV vectors to transduce neural stem cells, mesenchymal stem cells, and fibroblasts, which were then injected into the diseased brain. Many studies in mouse models have shown that variable amounts of transduction and correction of lesions can be achieved with recombinant Ad, AAV, LV, and HSV vectors. When animals are treated as neonates or fetuses, the progression of disease can be arrested before significant pathology occurs. Treatment of adult animals has shown that significant reversal of established lesions can be achieved, including improvements in behavioral tests and neurodegeneration parameters. Furthermore, studies have demonstrated that in some neural pathways, the enzyme can be transported to distal neuronal cell bodies, which can amplify the sphere of correction far beyond the injection site.

A major barrier to clinical application of these methods is in scaling up from the small rodent brain to a human brain, which is 2000–3000 times larger, depending on age. However, large animals with these diseases (cats, dogs, and monkeys) can serve as test systems because the animals have essentially the same disease as human patients, their brains are much closer in size (only 10- to 20-fold smaller), and the organization is more similar to the human brain than the rodent brain (e.g., a sulcated cerebral cortex). Clinical improvement of the neurological syndrome by gene transfer has been shown in at least one large animal model of an inherited disease (the alpha-mannosidosis cat), suggesting that human treatment for a disseminated CNS disease may be feasible.

Ocular Diseases

The eye is an attractive target for gene therapy because it is easily accessible and can be monitored postinjection using opthalmoscopy. It is also a relatively enclosed system, so when using a nontransportable vector, concerns about the transgene reaching the rest of the CNS are reduced. Most gene-therapy studies have looked at treating diseases of the retina. Retinal diseases that are currently being studied for treatment with gene therapy include several of the retinitis pigmentosas, chorioiderimea, and ocular

neovascular disorders such as age-related macular degeneration and diabetic retinopathy. Dogs with Leber congenital amaurosis have had their eyesight restored by AAV vector-mediated gene transfer, providing the basis for a clinical trial that is underway.

Tumors

One use of viral vectors in tumor therapy involves introducing a therapeutic gene that may generate anticancer compounds within the tumor. Gene-mediated drug activation can be selectively targeted to the tumor tissue alone, saving surrounding normal tissue from damage. Therapeutic genes can directly kill or block the growth of tumor cells, inhibit angiogenesis, stimulate immune responses to tumor antigens, or block tumor cell invasion. Vectors that are cytotoxic, such as Ad and HSV, can serve as selectively toxic agents and be targeted to receptors that are specifically expressed on tumor cells. Ad5 vectors have tropism for some brain tumors and so are currently being used in model systems of gene therapy to treat CNS tumors that are both primary and metastatic. Unfortunately, current gene-therapy approaches have not resulted in the delivery of the therapeutic gene to all the tumor cells. However, HSV has been used in rodent models in which it was shown to kill tumor cells while sparing adjacent normal neurons. HSV is one of the few systems that has shown some efficacy in human trials. RV vectors have also been used to deliver genes such as HSV-derived thymidine kinase that convert gancyclovir into a toxic product in dividing cells (by DNA chain termination), which is optimal in the brain, where most normal cells are nondividing.

Epilepsy

Epilepsy is due in part to an imbalance in excitatory and inhibitory neurotransmission. Gene therapy could be used to deliver agents to neurons with the potential of affecting this imbalance. Neurotransmitter receptors and ion channels are the most likely targets for gene-therapy applications. Vectors that express glutamate, γ-aminobutyric acid (GABA), neuropeptide Y, and galanin have been used to attempt to treat induced epilepsy in animal models. Results have been promising, and the attenuation of seizures and neuronal death by AAV vectors expressing galanin have been reported. Adenosine is a drug that can inhibit neuronal activity, but adenosine causes strong adverse side effects when administered in high doses. To circumvent these adverse side affects, another method to treat epilepsy has been developed using gene therapy that involves grafting *ex vivo*-modified cells expressing adenosine directly into the region with the neurotransmission imbalance. This method has been shown to suppress seizures in rodent models of epilepsy without any side effects. Recently, another approach has been to manipulate the stoichiometry of $GABA_A$ receptors, which are disrupted by experimental status epilepticus in the rat, to make the cells less susceptible to abnormal activity.

Trauma

The mammalian CNS does not spontaneously regenerate or reestablish synaptic contacts after severe traumatic injury. Trauma causes a cascade of events leading to substantial cell death long after the initial event, thus a number of strategies are being tested to block these sequele. Viral vectors have been used to express neuroprotective agents such as NGF or BDNF within the area of injury, resulting in reduced or reversed atrophy, and improved learning ability over time. *Ex vivo* therapy has also been used in which genetically modified cells expressing NGF have been engrafted into the region that suffered the trauma, which attenuated long-term cognitive dysfunction posttrauma.

Conclusion

There are numerous potential applications for viral gene transfer vectors in the CNS. The use of these vector systems in the CNS has proven to be invaluable in treating experimental disease in animal models and, despite previous setbacks in gene-therapy studies, are beginning to be used clinically. Viral vectors also have many nontherapeutic uses, including modeling human diseases in animals, studying the function of specific genes, studying developmental fate, and mapping neuronal pathways. Their usefulness in experimental manipulations of specific functions in the brain will increase as the properties of cell tropism and gene expression are more precisely understood and controlled.

See also: Alzheimer's Disease: Neurodegeneration; Amyotrophic Lateral Sclerosis (ALS): Disease Mechanisms; Animal Models of Alzheimer's Disease; Animal Models of Huntington's Disease; Animal Models of Parkinson's Disease; Cell Replacement Therapy: Parkinson's Disease; Cell Replacement Therapy for Huntington's Disease; Engineering Viruses for CNS studies; Gene Therapy: Genetically Modified Cells; Gene Therapy: Direct Viral Delivery; Huntington's Disease: Neurodegeneration; Parkinson's Disease: Alpha-Synuclein and Neurodegeneration; Retinitis Pigmentosa; Tumors of the Brain and the Spinal Cord.

Further Reading

Benitez JA and Segovia J (2003) Gene therapy targeting in the central nervous system. *Current Gene Therapy* 3: 127–145.

Berges BK, Wolfe JH, and Fraser NW (2007) Transduction of brain by herpes simplex virus vectors. *Molecular Therapy* 15: 20–29.

Jakobsson J and Lundberg C (2006) Lentiviral vectors for use in the central nervous system. *Molecular Therapy* 13: 484–493.

Kirik D and Bjorklund A (2003) Modeling CNS neurodegeneration by overexpression of disease-causing proteins using viral vectors. *Trends in Neuroscience* 26: 386–392.

Lundstrom K (2001) Alphavirus vectors for gene therapy applications. *Current Gene Therapy* 1: 19–29.

Mandel RJ, Manfredsson FP, Foust KD, et al. (2006) Recombinant adeno-associated viral vectors as therapeutic agents to treat neurological disorders. *Molecular Therapy* 13: 463–483.

Mata M, Glorioso JC, and Fink DJ (2003) Gene transfer to the nervous system: Prospects for novel treatments directed at diseases of the aging nervous system. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 58: 1111–1118.

Neve RL, Neve KA, Nestler EJ, and Carlezon WA, Jr. (2005) Use of herpes virus amplicon vectors to study brain disorders. *Biotechniques* 39: 381–391.

Ralph GS, Binley K, Wong LF, Azzouz M, and Mazarakis ND (2006) Gene therapy for neurodegenerative and ocular diseases using lentiviral vectors. *Clinical Science* 110: 37–46.

Wu Z, Asokan A, and Samulski RJ (2006) Adeno-associated virus serotypes: Vector toolkit for human gene therapy. *Molecular Therapy* 14: 316–327.

Visceral Pain

G F Gebhart, University of Pittsburgh, Pittsburgh, PA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Individuals suffering from one of the painful disorders that affect the internal organs (e.g., irritable bowel syndrome, noncardiac chest pain, nonulcer dyspepsia, and painful bladder syndrome) fully comprehend the typical characteristics of visceral pain. Visceral pain is diffuse in character and is typically difficult to localize unless adjacent structures, such as the parietal peritoneum or pleura, are involved (e.g., appendicitis). With respect to functional disorders of the viscera, defined by chronic pain in the absence of structural, infectious, or biochemical abnormalities, hypersensitivity is the shared characteristic. This hypersensitivity is commonly not restricted to the affected organ but also associated with both an expansion of the area on the body to which visceral pain is typically referred (or transferred) and increased tenderness to cutaneous/subcutaneous probing in the area of referral.

The anatomical organization and central nervous system terminations of visceral sensory (afferent) input largely explains why visceral pain is not commonly felt at the source, is diffuse in character, and is generally associated with greater autonomic and emotional responses than is pain arising from superficial structures. Visceral hypersensitivity, particularly as it relates to functional visceral disorders, involves both peripheral and central mechanisms, neither of which is fully understood. This article briefly summarizes the major features of visceral pain and addresses visceral hypersensitivity by considering potential mechanisms and molecules that contribute to the sensitization and increase in neuron excitability that underlie hypersensitivity. It is important at the outset to appreciate that visceral pain in general and functional disorders in particular do not conform to the traditional biomedical model.

Characteristics of Visceral Pain

Unlike activation of sensory receptors in skin, muscle, and joints, activation of sensory receptors in the viscera by a wide range of physiologically relevant stimuli does not commonly result in the conscious awareness of such input. For example, mechanical activation of baroreceptors in the aortic arch, secretion of digestive enzymes and activation of gastric and small bowel motility initiated by the ingestion of food, and activation of airway receptors by inspiration rarely lead to conscious perception of such stimuli. The principal conscious sensations that do arise from the viscera include nausea, dyspnea, discomfort (e.g., bloating), and pain generated by mechanical (principally), ischemic, and inflammatory events in visceral organs. Ischemic events, such as associated with angina, often generate conscious sensation of chest pain, which is typically referred to the upper left shoulder and down the left arm. Organ inflammation, such as associated with the colon and pancreas, generates conscious, diffuse abdominal pain. Mechanical events that lead to obstruction of organ outflow (e.g., gall or kidney stones), traction on the mesentery, or organ overfilling, which increases tension of circumferential and longitudinal organ smooth muscle, similarly generate diffuse, difficult to localize sensations, including significant pain in many cases (e.g., kidney stones and intestinal obstruction). Complicating localization of the source of visceral input is referral of visceral sensation to nonvisceral structures such as skin and overlap in areas of referral from different organs. Referral reflects the convergence of visceral and nonvisceral inputs onto the same second-order spinal neuron. Virtually all spinal neurons that receive a visceral input also receive input from a cutaneous and/or subcutaneous site, whereas spinal neurons that receive a cutaneous input do not all also receive a visceral input. In addition to somatovisceral convergence in spinal cord, viscerovisceral convergence is common. For example, sensations arising from the gall bladder, lower esophagus, and heart have overlapping areas of referral on the chest. Similarly, sensations arising from the urinary bladder, large bowel, and other pelvic organs have overlapping areas of referred sensation on the abdomen. Notably, tissue-damaging stimuli, which uniformly cause pain when applied to skin, do not reliably give rise to visceral discomfort or pain. That is, stimuli adequate for activation of nociceptors are different for skin, muscle, joints, and the viscera. Cutting, crushing, or burning stimuli applied to the viscera do not generally cause discomfort. Rather, as indicated previously, mechanical, ischemic, and inflammatory stimuli are adequate for activation of visceral nociceptors.

Visceral Innervation, Central Terminations, and Projections

Distinct from all other tissues in the body, each internal organ is innervated by two different nerves. The vagus nerves, probably the most extensive sensory nerves in the body with respect to targets of innervation,

innervate all organs in the thoracic cavity and most organs in the abdominal cavity except the urinary bladder, distal colon, and pelvic organs (e.g., prostate, cervix, and uterus). Cell bodies of vagus nerves are located principally in the nodose ganglia with central terminations in the brain stem nuclei tractus solitarii (NTS). These same thoracic and abdominal organs are also innervated by nerves identified as spinal visceral nerves with cell bodies in dorsal root ganglia that are bilaterally distributed along the length of the spinal cord. Accordingly, the airways, esophagus, stomach, heart, etc. are innervated by both vagus and spinal visceral nerves with central inputs terminating in NTS and thoracic spinal cord, respectively. The urinary bladder and distal colon are innervated by the pelvic nerve and splanchnic nerves with central terminations in sacral and upper lumbar segments of spinal cord, respectively; multiple segments of lumbar spinal cord separate these inputs. Thus, all viscera send inputs to distinct and separate sites in the central nervous system, further contributing to difficulty in localization of input and to the diffuse character of such input. It has been appreciated for some time that the different innervations of the viscera are associated with distinct functions, but characterizing these functions has only recently become a research priority.

Despite the extensive reach of the vagus nerves and dual innervation from spinal nerves, the viscera are relatively sparsely innervated, particularly by the spinal visceral nerve innervation. Only approximately 5–8% of total sensory input from all tissues into the spinal cord is represented by visceral input. Visceral input into spinal cord, however, is extensive in the sense that the spinal inputs diverge into rostral (ascending) and caudal (descending) branches which arborize extensively in multiple segments both above and below the site of entry into the spinal cord. Moreover, spinal visceral input also includes fiber terminations in the side of the spinal cord contralateral to the input. This organization, again, contributes both to difficulty in localization of visceral input and to the diffuse character of visceral pain.

Supraspinal Projections

Spinal visceral input reaches supraspinal sites by two principal pathways. Primary visceral inputs terminate on second-order neurons in the spinal cord that are located principally in the superficial dorsal horn (laminae I and II), in the intermediolateral cell column/sacral parasympathetic nucleus (where they synapse onto autonomic efferent neurons), and in lamina X dorsal to the central spinal canal. One ascending pathway of visceral input arises from axons of second-order spinal neurons that cross to the contralateral side of the spinal cord and become part of the anterolateral ascending pathway en route to third-order neurons in the thalamus (Figure 1). A postsynaptic dorsal column pathway arises from axons of second-order spinal neurons located principally in lamina X. This pathway ascends the spinal cord on the side ipsilateral to the visceral input and synapses on third-order neurons in the nuclei gracilis or cuneatus (depending on the spinal level of visceral input), which then crosses to the contralateral side of the central nervous system en route to destinations generally overlapping with those receiving input from the anterolateral ascending spinal pathway. Both ascending pathways give rise to collateral axons throughout the brain stem as well as to hypothalamus, thus widely distributing visceral input throughout the brain. Visceral inputs that reach areas in the brain stem medulla and pons are associated with autonomic control and feedback functions, autonomic responses to noxious visceral stimuli, and endogenous pain control (e.g., midbrain periaqueductal gray). Beyond the thalamus, visceral inputs reach the anterior cingulate cortex, the anterior insular cortex and primary and secondary

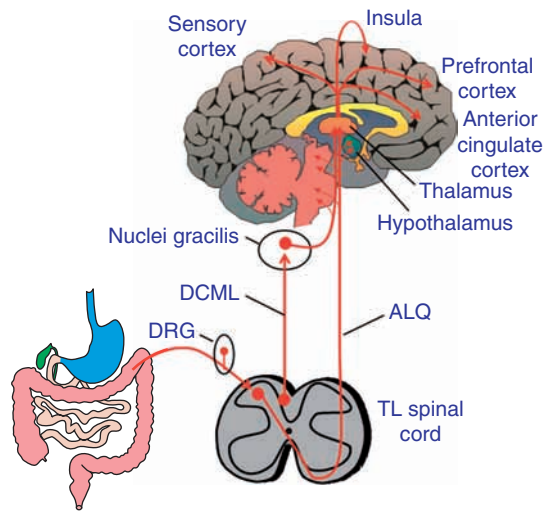

Figure 1 Diagrammatic representation of central visceral sensory pathways. Spinal visceral nerve input from the gastrointestinal tract (and thoracic cavity organs and pelvic organs, not shown) terminates on second-order spinal neurons in the superficial dorsal horn, which conveys information rostrally after crossing the spinal cord to become part of fiber tracts ascending and terminating in the brain stem, hypothalamus, and thalamus. A second ascending pathway arises from second-order spinal neurons located principally dorsal to the central canal (spinal visceral nerve input to these neurons is not illustrated). Fibers in this second pathway ascend in the dorsal column medial lemniscal (DCML) tract to synapse on third-order neurons in the ipsilateral gracile and cuneate nuclei, the ascending axons of which cross to the contralateral side of the brain to terminate in the brain stem, hypothalamus, and thalamus. Visceral information is widely distributed to the sensory cortices, insula, prefrontal cortex, and anterior cortex. ALQ, anteriolateral quadrant ascending tract; DRG, dorsal root ganglion.

sensory cortices, as well as areas of the limbic system associated with the emotional and perceptional consequences of visceral inputs.

Intrinsic (Enteric) Nerves

In addition to the sensory (extrinsic) innervation of visceral organs described previously, the viscera are further characterized by an intrinsic innervation (sometimes referred to as intrinsic primary afferent neurons). This local or 'regional' nervous system is considered to be the third major division of the autonomic nervous system. In the gastrointestinal tract, a myenteric plexus, containing most of the enteric neurons, is located between the circumferential and longitudinal layers of organ smooth muscle and is principally associated with motility, particularly coordinating oral to aboral motility. This is achieved by descending and ascending excitatory and inhibitory reflexes within the gut. A submucosal plexus, or actually submucosal plexus layers (outer and inner), also contributes to motility but, more importantly, to vasomotor and secretomotor organ functions. The enteric nervous system can act independently of the central nervous system, and the various classes of neurons within the enteric nervous system respond to a rich variety of local stimuli, including mechanical and chemical stimuli.

The foregoing only superficially describes the enteric nervous system, which is described here primarily to emphasize the virtual absence of knowledge about how the extrinsic sensory visceral innervation, which gives rise to conscious visceral sensations, interacts with the intrinsic nervous system. Even the anatomical relationships between extrinsic sensory nerve terminals in organ muscle layers or submucosal sites and neurons intrinsic to an organ have not been resolved, thus limiting our ability to unravel functional interactions. It seems likely that functional visceral disorders, which previously were considered to represent motility or vasosecretory disorders, represent at least in part a dysfunctional interaction between the intrinsic and extrinsic nervous systems.

Visceral Sensitivity and Hypersensitivity

Among stimuli adequate for activation of visceral nociceptors, mechanical stimulation has been most widely employed for several reasons. First, balloon distension of hollow organs in humans reproduces the localization, sensation, and intensity of their visceral pain. Second, mechanical stimuli are more easily controlled experimentally than either ischemia or inflammation. Consequently, in both behavioral and electrophysiological experiments in nonhuman animals, balloon distension of hollow organs has been widely used to develop models of visceral pain and to evaluate function(s) of the visceral sensory innervation.

Mechanosensation and Nociceptors

Single-fiber studies *in vivo* have revealed the presence of two groups of mechanosensitive afferent fibers in visceral nerves. Studies in gall bladder, esophagus, stomach, urinary bladder, small bowel, colon, ureter, and uterus in different species (mouse, rat, guinea pig, cat, and opossum) have identified a proportion of afferent fibers with high thresholds for activation, principally by organ distension. These high-threshold mechanosensitive endings in the viscera first respond at distending pressures considered to be noxious ($\geq$30 mmHg) and the 20–30% of high-threshold fibers described in the literature likely function as mechanonociceptors and signal acute visceral pain. Typically, they encode distending pressures throughout the noxious range. That is, response magnitude increases as stimulus intensity (distension pressure) increases. The larger proportion of mechanosensitive afferent fibers (e.g., 70–80%) have response thresholds in the physiological range (2–5 mmHg) and also encode stimulus intensity, typically into the noxious range. Accordingly, and unlike cutaneous mechanoreceptors, both high- and low-threshold visceral mechanoreceptors are responsive to noxious intensities of visceral stimulation.

Nociceptors were proposed by Sherrington to be the sensory end organs in skin that responded to stimuli which damaged or threatened damage to skin (e.g., cutting and crushing). Thus, nociceptors are defined functionally by response to noxious stimulation. As indicated previously, adequate noxious stimuli differ between skin, muscle, joints, and viscera, and although cutting, crushing, or burning of visceral tissue do not reliably activate nociceptors, mechanical, ischemic, and inflammatory events do (although ischemic and inflammatory stimuli likely activate visceral nociceptors indirectly).

When studied, visceral mechanoreceptors also respond to chemical and/or thermal (heat or cold) stimulation, revealing that most, if not all, mechanosensitive endings in the viscera are polymodal. The converse, however, is not true; that is, there likely exist receptive endings in viscera that are selectively chemosensitive (e.g., to nutrients) and are not also thermo- or mechanosensitive. Another characteristic of visceral nociceptors, first described for cutaneous nociceptors, is their ability to sensitize. Sensitization is defined by an increase in response magnitude to a noxious intensity of stimulation and a decrease in response threshold into what was previously the range of nonnoxious stimulation; an example is

given in Figure 2. Spontaneous activity may also increase. Interestingly, both high-threshold and low-threshold visceral mechanoreceptors sensitize after experimental organ insult or even after exposure to noninjurious stimuli (e.g., intragastric instillation of a hot, 46 °C solution). These findings suggest that low-threshold visceral mechanoreceptors, unlike low-threshold cutaneous mechanoreceptors, can contribute to sensations of visceral discomfort or pain. Extrapolating this information to functional visceral disorders in humans suggests activation of sensitized low-threshold mechanoreceptors by normal physiological events in the viscera, such as ingestion of food and slow bladder filling.

In vitro preparations of organs with nerves attached reveal a further division of visceral mechanoreceptors, including the presence of mechanosensitive endings associated with organ mucosa/submucosa (or bladder urothelium), smooth muscle, mesenteric attachments, and serosa. There have also been described muscular–mucosa endings in stomach, urinary bladder, and colon which respond both to gentle stroking of the mucosal surface and to circumferential stretch of muscle. Although evidence is limited, muscular–mucosal mechanosensitive endings in the viscera seem to be more readily sensitized than muscular endings in the viscera. If these findings are correct, they may provide a mechanistic foundation for common complaints of patients with functional disorders associated with normal physiological processes that would initially access the mucosal surface of an organ. Furthermore, the proportions of muscular and muscular–mucosal endings in the different nerves innervating the same organ are not the same. For example, there is a greater proportion of stretch-responsive endings in the pelvic nerve innervation of the mouse colon relative to the lumbar splanchnic nerve innervation, and only the pelvic nerve innervation includes muscular–mucosa endings. This suggests the possibility of target-oriented therapeutic strategies because, presumably, the phenotypic expression of sensory receptors differs between the two nerves innervating an organ.

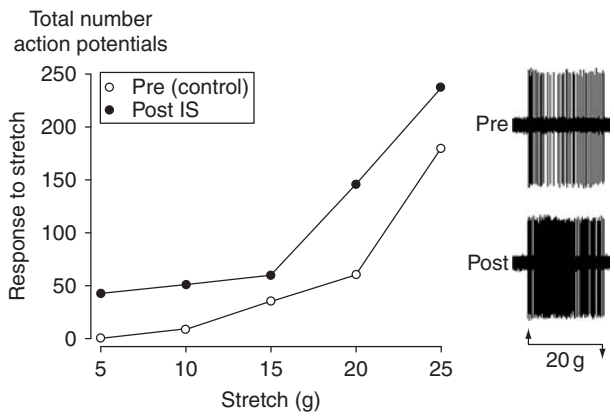

Figure 2 Example of sensitization of a mechanoreceptor in the pelvic nerve innervation of the mouse urinary bladder. Responses are plotted to 5 to 25 g stretch applied to the bladder before (pre) and after (post) application of inflammatory soup (IS; bradykinin, histamine, PGE_2, and serotonin at pH 6.0) to the receptive ending in the urinary bladder. The response to stretch 5 min after application of IS is increased, or sensitized, as illustrated by the action potential records for the response to 20 g stretch before and 5 min after application of IS.

Sensitization

The process of sensitization of visceral receptors and afferent fibers, reflecting an increase in excitability of the sensory neuron (as evidenced by changes in sodium, potassium, and calcium ion currents), has important consequences within the central nervous system. Peripheral sensitization of visceral sensory neurons leads to increased input from organs into the central nervous system, release of increased amounts of neurotransmitters and modulators from their central terminations, and a consequent increase in the excitability of the central neurons upon which the visceral sensory neurons terminate. This process is termed central sensitization and has been studied most extensively at the first central synapse, usually in the spinal cord, but it also occurs at subsequent central synapses, including in the brain stem and likely more rostral terminations of the visceral sensory pathways. Both peripheral and central processes of sensitization have been studied more extensively in nonvisceral models, in which mediators and intracellular signaling pathways have been examined and revealed. Analogous mechanisms and processes likely exist in the visceral realm, but this remains to be established. Both peripheral and central sensitizations are normally time limited and reversible. A good example is a burn (e.g., sunburn) to skin that leads first to sensitization of cutaneous nociceptors and then to central sensitization, which is commonly manifest as increased sensitivity to stimuli applied to the uninjured skin adjacent and near the burn, including stimuli that are normally not painful, as well as an expanded area (receptive field) from which discomfort and pain can be evoked. In the viscera, hypersensitivity is typically expressed as increased tenderness to palpation of the abdomen, for example, and an expanded area of referred sensation.

Although commonly reversible as tissue heals and recovers from insult, there is growing awareness that central sensitization in some cases may be sustained (by mechanisms not understood). That is, changes in the excitability of central neurons may persist, perhaps become permanent, and contribute to the development and maintenance of functional disorders. It

has been appreciated for some time that endogenous modulatory systems can be engaged in the brain stem (e.g., periaqueductal gray). Although it is now known that descending systems are neither selective nor specific modulators of spinal nociceptive (pain) transmission, they have been most widely investigated and discussed as a means of endogenous pain control (inhibition). Generally, inhibitory control does predominate, but facilitatory or excitatory influences are also engaged in the same areas of the brain stem and it is now considered that the balance between descending inhibitory and facilitatory influences may be changed in chronic pain conditions, including functional visceral disorders, and that the system is dysfunctional. Although important details and mechanisms are not currently known, brain imaging studies in humans and experimental manipulations in nonhuman animals do support the concept of central, supraspinal dysregulation of endogenous modulatory systems as contributing to maintenance of chronic, functional disorders in which discomfort and pain are prominent. It remains an open question whether central dysregulation is the result or cause of any functional disorder.

Mediators and Modulators of Visceral Hypersensitivity

Considerable evidence has accumulated during the past decade pointing at endogenous mediators and a variety of ion channels and receptors as contributing to the process of peripheral visceral neuron sensitization. Visceral sensory neurons have been shown to respond to numerous substances, including protons, serotonin, histamine, bradykinin, purines (e.g., ATP), and proteases, and all have been shown to likely contribute to visceral sensory mechanisms. It is important to appreciate that these and other putative mediators arise from a variety of sources, notably from epithelial cells lining hollow organs, including the bladder urothelium. For example, approximately 90% of the body's serotonin is contained in enteroendocrine cells located in the intestine. The number of enteroendocrine cells reportedly increases after intestinal inflammation, suggesting that the release of serotonin (5-HT) by mechanical or chemical stimulation within the intestine may be important in the pathogenesis of functional bowel disorders. Accordingly, therapeutic strategies for treatment of symptoms of irritable bowel syndrome (IBS) have targeted 5-HT receptor agonists (e.g., 5-HT4 agonists) and antagonists (e.g., 5-HT3 antagonists) and the serotonin transporter.

Mast cells are a rich source of bioactive substances and are found in close approximation to nerves within the human gastrointestinal tract. Mast cells contain and release histamine, proteases, and other potential mediators and modulators of visceral sensation. Mucosal biopsy samples from IBS patients release greater amounts of proteases than do biopsy samples from control subjects, presumably from mast cells. Also, supernatants from incubated biopsy samples have been shown to activate visceral sensory neurons, an effect that can be blocked experimentally by inactivation of serine proteases or a histamine receptor antagonist. Thus, in addition to serotonin, histamine and proteases released in greater amount from tissues of patients with functional bowel disorders suggest a role for these mediators in activation and sensitization of visceral sensory neurons.

Proteases such as tryptase activate proteinase-activated receptors (e.g., PAR_2) by releasing a tethered ligand that is part of the receptor to activate visceral sensory neurons, an effect that can be mimicked by experimental administration of a PAR_2 agonist. PARs have been hypothesized to link mast cell-derived proteases to inflammatory processes in colon and urinary bladder. In addition, PAR_2 agonists (as well as serotonin) have been shown to modulate the activity of the transient receptor potential ion channel TRPV1 found in many sensory neurons, including visceral sensory neurons. Additional chemical signals that can activate and/or sensitize visceral sensory neurons include endogenous products such as lactate, protons, and bradykinin, all of which can accumulate during ischemia (e.g., myocardial ischemia or angina), and cyclooxygenase-dependent products of arachidonic acid (e.g., prostaglandins). Moreover, gastroesophageal reflux disease, in which gastric acid reflux into the lower esophagus contributes to discomfort and pain, suggests a role for protons in visceral hypersensitivity. Bradykinin has been reported to accumulate in cardiac sinus blood during cardiac ischemic episodes and, along with lactate and protons, is believed to contribute to cardiac pain sensations. For example, experimental coronary artery occlusion leads to a prompt decrease in myocardial pH with commensurate increases in tissue proton and lactate content that accumulates during tissue hypoxia and can activate ischemia-sensitive cardiac sensory neurons in spinal visceral nerves.

Purines, including ATP and its metabolites, are ubiquitous and are released during cell injury, including from epithelial cells lining the gastrointestinal tract and urinary bladder. Ligand-gated P2X and G-protein-coupled P2Y receptors have thus been investigated for their potential roles in visceral pain mechanisms. Urinary bladder epithelium (urothelium) and colonic mucosa release ATP when these organs are experimentally distended. Following experimental organ inflammation, the amount of ATP released is

increased and can contribute to activation and/or sensitization of visceral sensory neurons. For example, bladder sensory neuron excitability is increased after experimental cystitis, as evidenced by the development of spontaneous activity, a significant reduction in rheobase (i.e., the amount of current injected into the cell that is required to generate an action potential), and changes in P2X channel kinetics of activation. Complimentary studies using P2X receptor antagonists or knockout mice have shown attenuation of responses to organ distension, revealing a potential role for ATP and its cognate receptors in mechanosensation from the viscera. It has been more generally suggested that epithelial cells which release ATP function as 'sensory intermediaries' in the signaling of organ luminal events to visceral sensory neuron terminals in submucosal tissue where luminal contents normally cannot access sensory neuron terminals.

Much of the foregoing has been supported by studies in mice in which selected gene deletions have been made (i.e., knockout mice). Thus, deleting $P2X_2$ and/or $P2X_3$ purinergic receptors alters responses in bladder and colon sensory neurons, supporting previous studies and reinforcing the notion that purinergic receptors can contribute to or modulate mechanosensory events in the viscera. Similarly, deleting ion channels that are gated by protons, such as acid sensing ion channels (ASICs) or the transient receptor potential vanilloid receptor 1 (TRPV1) also reveals contributions to visceral sensation. For example, TRPV1 knockout mice have a higher frequency of nonvoiding bladder contractions than do wild-type mice, and recordings from TRPV1 knockout mice visceral sensory nerves reveal attenuated responses to organ stretch (distension). Consistent with the importance of ischemia as a trigger of cardiac pain, in which increased protons (decreased pH) are present, and the contribution of protons to gastric and gastroesophageal pain, deletion of ASIC3 ion channels suggests their role in mediating nonmechanical events in the proximal gastrointestinal tract and heart. Supporting many of these experimental findings in nonhuman animals is evidence from human biopsy material. The expression of TRPV1, ASIC3, and $P2X_3$ has been shown to be increased in samples taken from people with gastroesophageal reflux disease, painful bladder syndrome, intestinal inflammation, and rectal urgency.

Considerable effort continues to be focused on improved understanding of peripheral events that contribute to visceral hypersensitivity and the central consequences of such altered visceral input. Therapeutic strategies are currently directed toward attenuating sensitized input from the viscera into the central nervous system. However, given that persistent visceral pain such as is characteristic of functional disorders likely involves changes in the central nervous system, acquisition of new knowledge about central contributions is critical to the development of visceral pain management strategies based on mechanism.

See also: Energy Homeostasis: Visceral Control; Mechanosensory Transduction; Nociceptor Responses; Pain and the Sympathetic Nervous System; Purinergic Receptors; Purines and Purinoceptors: Molecular Biology Overview.

Further Reading

Birder LA and de Groat WC (2007) Mechanisms of disease: Involvement of the urothelium in bladder dysfunction. *Nature Clinical Practice Urology* 4: 46–54.

Blackshaw LA, Brookes SJH, Grundy D, and Schemann M (2007) Sensory transmission in the gastrointestinal tract. *Neurogastroenterology Motility* 19(supplement 1): 1–19.

Burnstock G (2007) Physiology and pathophysiology of purinergic neurotransmission. *Physiological Reviews* 87: 659–797.

Cervero F (1994) Sensory innervation of the viscera: Peripheral basis of visceral pain. *Physiological Reviews* 74: 95–138.

Gebhart GF (ed.) (1995) *Visceral Pain, Progress in Pain Research and Management, Vol. 5*. Seattle: IASP Press.

Gebhart G. F (2004) Descending modulation of pain. *Neuroscience & Biobehavioral Reviews* 27: 729–737.

Gershon MD and Tack J (2007) The serotonin signaling system: From basic understanding to drug development for functional GI disorders. *Gastroenterology* 132: 397–414.

Gold MS (2006) Ion channels: Recent advances and clinical applications. In: Flor H, Kalso E, and Dostrovsky JO (eds.) *Proceedings of the 11th World Congress on Pain*, pp. 73–92. Seattle: IASP Press.

Holzer P (2002) Sensory neurone responses to mucosal noxae in the upper gut: Relevance to mucosal integrity and gastrointestinal pain. *Neurogastroenterology Motility* 14: 459–475.

Holzer P (2007) Taste receptors in the gastrointestinal tract: V. Acid sensing in the gastrointestinal tract. *American Journal of Physiology* 292: G699–G705.

Jänig W (2006) *The Integrative Action of the Autonomic Nervous System*. Cambridge, UK: Cambridge University Press.

Ness TJ and Gebhart GF (1990) Visceral pain: A review of experimental studies. *Pain* 41: 167–234.

Pasricha PJ, Willis WD, and Gebhart GF (eds.) (2007) *Chronic Abdominal and Visceral Pain*. New York: Informa Healthcare.

Porreca F, Ossipov MH, and Gebhart GF (2002) Tonic descending facilitation as a mechanism of chronic pain. *Trends in Neurosciences* 25: 319–325.

Spiller R and Grundy D (eds.) (2004) *Pathophysiology of the Enteric Nervous System*. Oxford, UK: Blackwell.

Wemmie J, Price M, and Welsh M (2006) Acid-sensing ion channels: Advances, questions and therapeutic opportunities. *Trends in Neurosciences* 29: 578–586.

Viscero-Sensory Functions: Capsaicin

J M A Laird, AstraZeneca R&D Montreal and McGill University, Montreal, QC, Canada

© 2009 Elsevier Ltd All rights reserved.

Capsaicin as a Tool in Studies of Viscerosensory Function

Capsaicin is the pungent substance found in spicy chili peppers (Figure 1). It has been used extensively as a tool in sensory research in three different ways, depending on the concentrations and manner in which it is employed.

Capsaicin as a Selective Stimulating Agent

At moderate concentrations, capsaicin acutely and intensely excites nociceptive afferent fibers without causing frank tissue damage and thus can be used as a very selective stimulating agent. Responses to capsaicin are often used to characterize visceral neurons and visceral pathways in physiological experiments.

Exposure of visceral afferents to capsaicin leads to their activation and sensitization; thus, application of capsaicin to visceral mucosa is used as a means to evoke controlled responses both in animals and in humans. For example, application of capsaicin to the colon in the mouse has been developed as a model of visceral pain and is used to explore the behavioral pharmacology of visceral pain responses. Intracolonic capsaicin evokes short-lived dose-dependent spontaneous pain responses and longer-lasting enhanced responses to external stimuli applied to the abdomen, hind paws, and tail (referred hyperalgesia). In human subjects, hot pepper sauce (which contains capsaicin) or capsaicin solutions delivered orally have often been used to evoke reactions in the gastrointestinal tract and the sensory and/or motor responses recorded. The sensations produced by direct application of capsaicin to the mucosa of the ileum have been described in patients with ileostomies. Perfusion of the duodenum with capsaicin solution has even been proposed as a diagnostic test for dyspepsia.

Capsaicin as a Selective Desensitizing Agent

In higher concentrations or with more prolonged exposure, capsaicin causes a functional elimination (desensitization) of a subpopulation of afferents. In visceral systems, this technique is often used for topical application to a mucosa.

Capsaicin may produce acute activation followed by desensitization, but afferents that do not generate action potentials in response to capsaicin exposure may also be cross-desensitized to other stimuli. Both spinal afferents originating in the dorsal root ganglia (DRG) and vagal afferents are affected. Approximately 40% of vagal afferents can be acutely desensitized on capsaicin exposure.

Topical application of capsaicin and other vanilloids, notably resiniferatoxin, to the bladder is used therapeutically in patients with overactive bladder. The positive results obtained have been shown to be due to capsaicin-induced desensitization of fine unmyelinated afferents innervating the bladder.

Capsaicin as a Selective Neurotoxin

When delivered in high systemic doses in neonatal rodents, capsaicin acts as a neurotoxin and kills a subpopulation of afferent sensory neurons. Animals treated as neonates with high systemic doses of capsaicin develop relatively normally and can be used in a wide variety of studies as adults. After neonatal capsaicin, stimulation of unmyelinated efferent fibers in autonomic nerves produces normal functional effects. This observation led to the conclusion that neonatal treatment with capsaicin selectively destroys unmyelinated afferent fibers without any effects on the unmyelinated efferents of the autonomic nervous system. Anatomical methods show that this technique eliminates a subpopulation of DRG and vagal afferent neurons, preferentially those with unmyelinated afferents along with a small number with thinly myelinated fibers. The proportion of dorsal root afferents innervating the viscera affected by neonatal capsaicin is larger than those in the somatic population. Also, since the viscera are much less densely innervated than cutaneous structures, the functional effects are more marked. However, the effect is not complete, and a proportion of visceral unmyelinated afferent fibers survive. Also, the neurotoxic effect is not uniform. For example, examining the vagal innervation of the gastrointestinal tract using anatomical tracing methods following capsaicin treatment, vagal afferents innervating the small intestine, the cecum, and colon are almost all eliminated. However, up to 90% of the esophageal and 60–90% of gastric vagal endings survive capsaicin treatment. There are some adaptive changes in the central nervous system (CNS) in neonatal capsaicin-treated animals, so the deficits observed in these animals probably do not reflect the full range of functions that the ablated afferents subserved.

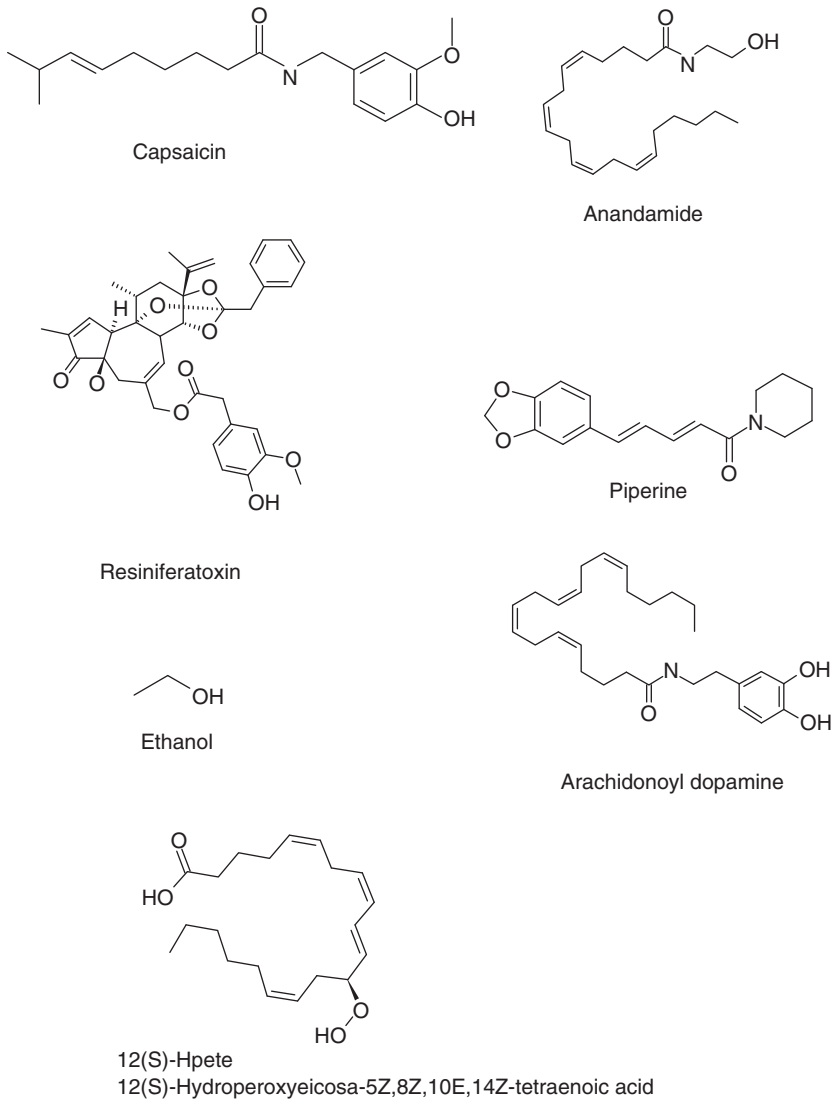

Figure 1 Structures of capsaicin and other TrpV1 agonists.

Molecular Mechanism by which Capsaicin Affects Viscerosensory Functions

A long series of classical pharmacological experiments carried out by many laboratories, including the development of synthetic agonists that mimicked the effects of capsaicin and antagonists that blocked its actions, led to general agreement that capsaicin exerted its effects on sensory neurons via a specific membrane receptor. The capsaicin receptor was finally isolated using an innovative expression cloning strategy and called the vanilloid receptor 1 (VR1) by Caterina and colleagues. The receptor is a member of the transient receptor potential (Trp) family and is now known as the TrpV1 receptor. The identification of the capsaicin receptor generated considerable interest and new work in this area because TrpV1 was the first transduction protein for noxious stimuli to be discovered.

TrpV1 is a nonselective cation channel that is very effectively gated by capsaicin in most mammals. The endogenous 'ligands' for TrpV1 include noxious heat (42–53 °C), protons (pH 5–6), and various lipid derivatives, including lipoxygenase metabolites of arachadonic acid, endogenous cannabinoids such as anandamide, and endogenous vanilloids such as *N*-arachidonoyl-dopamine (Figure 1). Which of these ligands is responsible for TrpV1 activation depends on the tissue and the physiological or pathophysiological circumstances, but they may also interact and potentiate the effects of each other (Figure 2).

Several natural products derived from plants, including capsaicin, also gate TrpV1. Capsaicin is

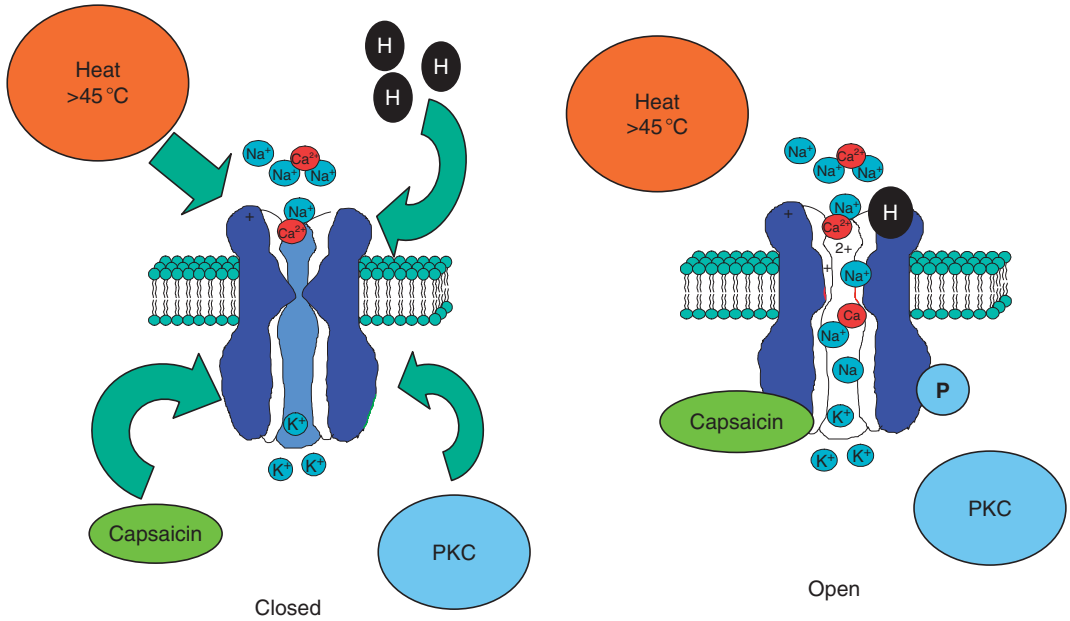

Figure 2 TrpV1 channel activation.

found in plants of the genus *Capsicum*, and it is a specialized metabolite produced in the fruits as a deterrent to seed predators. The production of capsaicin was presumably selected for during evolution because its stimulatory effects on the gastrointestinal tract of mammals is adversive, although capsaicin-producing *Capsicum* species have been cultivated by humans for more than 6000 years. Other plants cultivated for culinary or medicinal use produce TrpV1 agonists such as piperine, the pungent ingredient of black pepper corns, and resiniferatoxin, a very potent TrpV1 agonist secreted in the sap of the succulent plant *Euphorbia resinifera* commonly found in Morocco and used as a purgative (Figure 1).

Mechanism by which Capsaicin Excites and Sensitizes Visceral Afferent Neurons

TrpV1 gating (opening of the channel pore) results in an influx of cations and thus an excitatory effect on terminals of primary sensory neurons with a subsequent depolarization of the nerve fiber and the initiation of action potential propagation. Conduction of action potentials toward the CNS triggers reflex responses such as cough, voiding of the urinary bladder, or motor responses of the gut. The action potential traffic may also lead to conscious sensations, particularly of pain. In addition, calcium influx into the nerve endings, driven either by antidromic conduction of action potentials or directly by TrpV1 gating, causes the local release of neuropeptides from the terminal, including calcitonin gene-related peptide (CGRP) and the tachykinins, substance P, and neurokinin A. Activation of CGRP and tachykinin receptors on effector cells causes a series of responses, collectively referred to as neurogenic inflammation. Neurogenic inflammation is characterized by vasodilation of small arterioles, producing an increase in blood flow, and leakiness of capillary walls, resulting in plasma extravasation into the tissues, and swelling.

In addition to the wide range of potential adequate stimuli that gate TrpV1 channels, the TrpV1 receptor has another very important characteristic, which is that a large range of factors can sensitize the channel to activating stimuli. This sensitization is manifested both as a decrease in the threshold for activation and as an increase in the channel open time in the presence of normally suprathreshold stimuli. Sensitizing factors include direct phosphorylation of TrpV1 by protein kinase C (PKC) or protein kinase A and release of PtdIns(4,5)P2-mediated inhibition by activation of phospholipase C (PLC). Thus, the TrpV1 receptor may be activated by inflammatory mediators such as lipoxygenase products and low pH, but it can also be sensitized to heat, acidic stimuli, anandamide, or vanilloids by a range of inflammatory mediators which act on their own membrane receptors on the same cell and generate PKC and/or PLC, such as ATP via purinergic receptors, bradykinin via B2 receptors, or nerve growth factor via Trk A receptors.

This molecular sensitization of TrpV1 likely underlies the sensitization of visceral afferent terminals

to endogenous chemical and even heat stimuli, especially under inflammatory or ischemic conditions. For example, application of ATP can reduce the heat threshold of native TrpV1 channels from 43 to 35 °C, below normal body temperature, such that the channel would be activated continuously at core body temperature. Mild acidification (pH < 7.6) and low concentrations of capsaicin can also sensitize the response of TrpV1 to heat by a mechanism distinct from the direct activation of the channel by lower pH.

Mechanism by which Capsaicin Desensitizes or Kills Visceral Afferent Neurons

In the skin, topical application of capsaicin both in experimental animals and in humans has been shown to produce a rapid decrease in the number of epidermal fine nerve fibers, such that after 3 days there is a nearly complete disappearance of neuronal profiles as identified using a pan-neuronal marker, PGP9.5. The neurons do not die; capsaicin produces a dieback or retraction of the terminal portion of the nerve fiber arbors. On cessation of capsaicin treatment, the epidermal nerve fibers take 2 or 3 weeks to recover and reinnervate the epidermis. In the bladder, intravesical instillation of resiniferatoxin results in a long-lasting reduction in detrusor hyperactivity. Biopsies from patients in which the treatment was effective show a substantial decrease in the number of suburothelial fibers as determined by immunohistochemical labeling for PGP9.5. Studies also show a decrease in the TrpV1 immunoreactivity in both neuronal profiles and urothelial cells. The process of desensitization occurring in other visceral tissues after exposure to capsaicin or other exogenous TrpV1 agonists such as resiniferatoxin is likely very similar.

One important feature of the use of capsaicin as a desensitizing tool is that although capsaicin desensitization results from the action of capsaicin as a specific and selective TrpV1 agonist, it produces a defunctionalization of the entire nerve terminal (or even its physical disappearance). Thus, the affected afferents no longer respond to TrpV1 agonists (protons, lipid derivatives, capsaicin, etc.) and also do not respond to all the other stimuli that act on the different channels and receptors expressed on the nerve terminal. TrpV1 is therefore required for any capsaicin effect, but the results observed do not allow firm conclusions on the role of TrpV1 in the physiological function or pathological condition under investigation.

The precise mechanism triggering dieback of the afferent terminals on exposure to TrpV1 ligands such as capsaicin is unknown; however, it seems likely that it parallels the mechanism by which capsaicin has been shown to produce cell death in cultured, TrpV1-expressing neurons or cell lines. As described above, TrpV1 is a nonspecific cation channel. In neuronal cells, TrpV1 is expressed largely on the plasma membrane (although when heterologously expressed in other cell types, it is also expressed intracellularly). Exposure to capsaicin or other TrpV1 ligands gate the TrpV1 channel, allowing large influxes of calcium and thus very large increases in intracellular calcium concentration, which induce cytotoxicity. Cell death is dependent on extracellular calcium and appears to be mediated by activation of calcium-sensitive proteases such as calpain.

Expression of Capsaicin-Sensitive TrpV1 Receptors in Viscerosensory Pathways

TrpV1 was initially described as being expressed exclusively in primary sensory neurons (DRG and nodose ganglia). There is a very intense expression in these neuronal cells, but it is now clear that TrpV1 is also expressed in some populations of epithelial cells, such as the urothelium, where it plays an important role in sensory functioning. There are also some reports of weak TrpV1 expression in the myenteric plexus in the muscularis layer of the duodenum or intestine. It has been amply demonstrated using TrpV1 knockout (KO) mice that TrpV1 expression is responsible for capsaicin sensitivity both in *ex vivo* preparations and *in vivo*.

TrpV1 Receptors in Extrinsic Visceral Afferent Pathways

TrpV1 receptors are preferentially expressed in visceral afferents compared to somatic afferents and are present in the majority of spinal afferents innervating the gastrointestinal tract and the urinary system in rodents. TrpV1 immunoreactivity is lower in the nodose ganglion compared to the DRG. Approximately 40% of cells in the nodose ganglion in rodents express TrpV1-like immunoreactivity. Slightly less than half of the vagal afferents innervating the gastrointestinal tract express TrpV1-IR, but its presence in many of the peripheral terminal structures is below the immunohistochemical detection threshold. In contrast, two-thirds of rat DRG neurons innervating the urinary bladder expressed the TrpV1 receptor when examined using immunohistochemistry. Very similar proportions of TrpV1-positive afferent cell bodies are observed in colonic afferents in rats and mice. A smaller proportion of spinal afferents innervating the lung and pleura in the rat are immunoreactive for TrpV1 (33% and 49%, respectively). However, immunohistochemical detection methods may underestimate the proportion of neurons that express functional TrpV1 receptors and are thus

capsaicin sensitive. For example, more than 95% of DRG cell bodies of afferents innervating the urinary bladder in the rat respond to capsaicin, compared to 69% showing clear TrpV1-like immunoreactivity. Likewise, capsaicin activates between 50% and 90% of bronchopulmonary vagal unmyelinated afferent fibers in mice, whereas immunohistochemical data suggest that approximately 40% of vagal afferents innervating the lungs are TrpV1 positive.

TrpV1 Receptors Expressed in Visceral Epithelium Cells

TrpV1 receptors are highly expressed in the mucosa and muscle layers of the lower urinary tract. Some of this expression is attributable to unmyelinated afferent fibers that also express substance P and CGRP. However, TrpV1 is expressed not only by the afferent fibers located close to the urothelium but also by the urothelial cells. This expression has been extensively studied and shown to be functional.

There are also observations of TrpV1 immunoreactivity in subpopulations of epithelial cells in the small intestine in guinea pigs and in the gastric epithelium in rats. Functional TrpV1 receptors have been demonstrated in cultured human pulmonary epithelial cells. Whereas TrpV1 receptors in the urothelium have an established role in bladder functioning, and in lung epithelial cells appear to be important in the release of inflammatory cytokines, the role of TrpV1 receptors expressed in other epithelia is more speculative. Nonetheless, the presence of TrpV1 receptors in this cell type means that responses to capsaicin and other TrpV1 ligands throughout the viscera should not be interpreted only in terms of neuronal function. Likewise, results from experiments in which capsaicin was used as a desensitizing agent have usually been attributed to loss of afferent terminals; it is possible that modulation of the functions of epithelial cells may also play a role in the results obtained.

Physiological and Pathophysiological Role of Capsaicin-Sensitive TrpV1 Receptors in Viscerosensory Functions

Although the role of capsaicin-sensitive afferents has been extensively studied, the information obtained is difficult to link to specific molecular pathways for the reasons discussed previously. The use of TrpV1 knockout mice and newly described small-molecule antagonists of TrpV1 have shed some light on the potential physiological and pathophysiological role of TrpV1 receptors and on the promise for novel therapeutic approaches.

Functions of TrpV1 Receptors in the Gastrointestinal Tract

TrpV1 receptors in the gut contribute to the sensing of acidosis, a very important feature of gastrointestinal physiology given both the production of strong acids as part of the physiological functions of digestion and the appreciable drop in pH that can be produced by inflammation or ischemia in visceral tissues in general. TrpV1 KO mice show reduced behavioral responses to acid injected intraperitoneally, and jejunal and colonic afferents recorded in these mice are only weakly activated by intraluminal acid. Both jejunal and colonic afferents in TrpV1 KO mice are less mechanically sensitive, particularly to distensions in the noxious range. The TrpV1 channel is not, as far as is known, mechanically gated, so this observation is likely also explained by a response to acidosis. Intense distensions produce ischemia and thus lower muscle pH, thereby activating TrpV1.

TrpV1 receptors are strong candidates for a key role in the sensitization of visceral afferents after tissue injury or inflammation since they are activated and sensitized by a range of inflammatory mediators. However, there is relatively little information on the pathophysiological role of TrpV1 receptors in gastrointestinal pain from animal or human studies. Application of capsaicin to the mucosa of the gut in the mouse produces a short-lived visceral pain response and a very marked, long-lasting hyperalgesia that is referred to the body surface, showing that activation of TrpV1 receptors is sufficient to produce gastrointestinal pain and their sensitization evokes visceral sensitization. However, are TrpV1 receptors necessary for gastrointestinal pain and sensitization during ischemia and inflammation?

In colon primary sensory neurons cultured from TrpV1 KO mice, sensitization to acid, capsaicin, and heat induced by serotonin is absent. Likewise, although the initial response to an inflammatory 'soup' or mixture of inflammatory mediators instilled into the colon was intact in colonic afferents from KO mice, the pronounced sensitization to colonic distension seen in wild-type mice failed to develop. Small-molecule antagonists of the TrpV1 receptor also reduce the severity of dextran-induced colitis in mice, and preliminary data suggest that TrpV1 antagonists reduce pain responses to distension of the inflamed colon in rats.

Taken together, these results suggest that TrpV1 likely contributes to pain from the gastrointestinal tract associated with sensitization of visceral afferents. This conclusion is further supported by observations of increased expression of TrpV1 in biopsies from humans suffering from painful gastrointestinal

disorders. TrpV1 receptor expression is upregulated in tissue samples from patients with esophagitis, painful inflammatory bowel disease, or Crohn's disease and also patients with rectal hypersensitivity. In the latter case, TrpV1 expression was positively correlated with the degree of sensitivity.

Capsaicin-sensitive afferents and TrpV1 also appear to have some protective functions in the gastrointestinal tract. Capsaicin-sensitive afferent fibers protect the gastrointestinal mucosa from acid-induced injury by increasing mucosal blood flow and mucus secretion in the presence of impending acid damage. Activation of the afferents also attracts mucosal inflammatory cells and increases their release of transforming growth factor-α. Furthermore, their presence is required to promote the proliferation of epithelial cells to repair damaged tissue. This role is in part subserved by TrpV1 receptors acting as a sensor of luminal pH, for example, in the duodenum. However, the pH at the epithelial surface is kept neutral by the mucus layer, so the pH level is signaled across the epithelial barrier. This occurs when luminal H^+ combines with secreted HCO_3^- to generate CO_2 (and H_2O). The CO_2 diffuses across the membrane, is hydrated to H^+ and HCO_3^-, and acidifies the epithelial cell, which extrudes H^+ into the interstitial space via anion exchange. Here, it is detected as lowered interstitial pH by TrpV1 receptors on sensory terminals.

Surprisingly, activation of TrpV1 receptors also has some protective functions in inflammation of the colon. Topical application of capsaicin protects against trinitobenzenesulfonic acid-induced colitis in mice, whereas it is enhanced in mice lacking TrpV1 receptors, which contrasts with the previous description of TrpV1 antagonists and genetic ablation of TrpV1 reducing the severity of colitis in mice. The role of TrpV1 in the gastrointestinal tract is complex and likely varies depending on the pathophysiological circumstances.

Role of TrpV1 in Bladder Function

Sensations ranging from fullness to intense pain can be perceived from the urinary bladder and ureters. The signaling of all these sensory events was previously attributed to direct activation of the afferents that innervate the urinary tract, but recent evidence suggests that the bladder epithelial cells, the urothelium, are key players in the transduction of sensory information. The lower urinary tract is lined with a multilayer transitional epithelium that acts as a barrier but that also has dynamic properties, including viscerosensory functions.

Urothelial cells release ATP on stretch and other sensory stimuli. ATP acts on $P2X_3$ receptors located in unmyelinated primary afferents to trigger both normal reflex bladder activity and pain behavior evoked by bladder irritation. Birder and coworkers further investigated the role of urothelial cells in the signaling of sensory stimuli. They found that exogenously applied capsaicin increased intracellular calcium and evoked nitric oxide release from cultured urothelium cells by a process that required TrpV1 receptors. Moreover, mice lacking TrpV1 receptors showed a reduction of ATP release by urothelial cells and a number of alterations in the expression of the micturition reflex indicative of a critical role of urothelial TrpV1 receptors in response to bladder distension.

These studies revealed a novel mechanism for the transduction of sensory stimuli, including those that may evoke pain, by the lower urinary tract. TrpV1 agonists can activate urothelial cells, and this results in the release of ATP into the subepithelial space. Likewise, the same stimuli will also activate the TrpV1 receptors located in the endings of afferent fibers and, additionally, the ATP released by the urothelium will stimulate $P2X_3$ receptors in these afferents, which in turn will evoke sensory events, such as fullness or pain, and reflex activation of bladder activity. The data also suggest a permissive role of TrpV1 receptors in the distension-evoked release of ATP by the urothelium and in the mechanical activation of bladder afferents via intermediate mechanisms not yet identified.

These observations have an immediate clinical relevance regarding the pathophysiology of bladder hypersensitivity. Urothelial cells from cats suffering from a feline equivalent of interstitial cystitis release significantly more ATP; this may account for the enhanced bladder sensitivity characteristic of this condition. Patients suffering from interstitial cystitis also show an increased level of TrpV1 immunoreactivity in the urothelium compared to normal controls. Furthermore, reduction of TrpV1 immunoreactivity after intravesical desensitization with resiniferatoxin is correlated with the degree of functional relief from overactive bladder symptoms in patients.

Role of TrpV1 in the Respiratory Tract

Activation of TrpV1 by capsaicin or other agonists directly evokes cough. A small-molecule TrpV1 receptor antagonist (BCTC) inhibits both capsaicin-induced cough and cough mediated by antigen in sensitized guinea pigs. TrpV1 is increasingly proposed as the mediator, along with other acid sensors such as acid sensing ion channels (ASIC) of acid-induced cough. Thus, TrpV1 seems to play a key role in the generation of the cough response. This is supported by the increase in expression of TrpV1 observed in airway nerves of patients with chronic cough and also

the lowered threshold to capsaicin-induced cough seen in patients with asthma and chronic obstructive pulmonary disease. Likewise, cough is a well-known adverse effect of angiotensin-converting enzyme inhibitors, which prevent the breakdown of bradykinin. Bradykinin sensitizes TrpV1 receptors via phosphorylation as described above.

Capsaicin aerosol application to the lungs also evokes bronchoconstriction, microvascular leakage, and hypersecretion. In humans and several other species, these responses are mediated directly via release of neuropeptides, particularly the tachykinins, from TrpV1-expressing afferents in a process known as neurogenic inflammation. A number of agents implicated in provoking respiratory symptoms are known to sensitize or modulate the TrpV1 receptor, which may thus have a key role in the etiology of these conditions. For example, ethanol-induced asthma is poorly understood, but ethanol sensitizes the TrpV1 receptor such that it can be activated at body temperature. Likewise, there is evidence that the afferent response to reactive oxygen species, important in smoke-induced lung injury, is mediated in part via TrpV1 receptors.

TrpV1 receptors expressed both on afferent nerve terminals and on lung epithelial cells may subserve particulate matter-induced airway inflammation. Activation of bronchial epithelial cells by capsaicin promotes proinflammatory cytokine production and cell death, whereas TrpV1 antagonists or reduction of extracellular calcium attenuate calcium flux and cytokine secretion. Exposure of these cells to particulate matter also evokes calcium fluxes and apoptosis, and these responses seem to be associated with the TrpV1-mediated response since they were abolished in cells treated with antagonists or obtained from TrpV1 KO mice.

TrpV1 receptors are likely responsible for the responses of lung irritant receptors to acidic, chemical, and even particulate matter stimuli. The sensitization and upregulation of TrpV1 receptors may also contribute to their importance during chronic inflammation of the lungs and underlie some of the symptoms which characterize this condition, particularly cough, but also perhaps irritation, discomfort, and chest tightness.

Concluding Remarks

Capsaicin and capsaicin analogs derived from natural products have been extremely valuable tools in dissecting the physiology and pathophysiology of the somatosensory system and in particular viscero-sensory functions. The discovery of the molecular mechanism by which capsaicin exerts its effects with the cloning and description of the TrpV1 receptor and the availability of TrpV1 KO mice has refined our understanding considerably in the past decade. The recent availability of selective small-molecule TrpV1 receptor antagonists which can be used as specific tools in animal studies and which are being assessed in the clinic as potential therapies for multiple disorders will no doubt lead to even greater insights during the next decade.

See also: Autonomic Nervous System: Gastrointestinal Control; Autonomic Nervous System: Respiratory Control; Energy Homeostasis: Visceral Control; Mechanosensory Transduction; Neurotoxins; Neurotoxins and their Neurotoxicology; Transient Receptor Potential (TRP) Channels; Visceral Pain.

Further Reading

Avelino A and Cruz F (2006) TRPV1 (vanilloid receptor) in the urinary tract: Expression, function and clinical applications. *Naunyn Schmiedebergs Archives of Pharmacology* 373: 287–299.

Bartho L, Benko R, Patacchini R, et al. (2004) Effects of capsaicin on visceral smooth muscle: A valuable tool for sensory neurotransmitter identification. *European Journal of Pharmacology* 500: 143–157.

Birder LA, Nakamura Y, Kiss S, et al. (2002) Altered urinary bladder function in mice lacking the vanilloid receptor TRPV1. *Nature Neuroscience* 5: 856–860.

Caterina MJ and Julius D (2001) The vanilloid receptor: A molecular gateway to the pain pathway. *Annual Review of Neuroscience* 24: 487–517.

Cervero F and Laird JM (1999) Visceral pain. *Lancet* 353: 2145–2148.

Cervero F and Laird JM (2004) Understanding the signaling and transmission of visceral nociceptive events. *Journal of Neurobiology* 61: 45–54.

Geppetti P, Materazzi S, and Nicoletti P (2006) The transient receptor potential vanilloid 1: Role in airway inflammation and disease. *European Journal of Pharmacology* 533: 207–214.

Geppetti P and Trevisani M (2004) Activation and sensitisation of the vanilloid receptor: Role in gastrointestinal inflammation and function. *British Journal of Pharmacology* 141: 1313–1320.

Holzer P (2004) TRPV1 and the gut: From a tasty receptor for a painful vanilloid to a key player in hyperalgesia. *European Journal of Pharmacology* 500: 231–241.

Montell C, Birnbaumer L, Flockerzi V, et al. (2002) A unified nomenclature for the superfamily of TRP cation channels. *Molecular Cell* 9: 229–231.

Szolcsanyi J (2004) Forty years in capsaicin research for sensory pharmacology and physiology. *Neuropeptides* 38: 377–384.

Vision for Action and Perception

M A Goodale, University of Western Ontario, London, ON, Canada
A D Milner, Durham University, Durham, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Humans are supremely visual creatures. We depend on vision, more than on any other sense, to perceive the world of objects and events beyond our bodies. At the same time, we also use vision to move around that world and to guide our goal-directed actions. During the past two decades, it has become increasingly clear that the visual pathways that mediate our perception of the world are quite distinct from those that mediate the control of our actions. Indeed, this distinction between 'vision for perception' and 'vision for action' has emerged as one of the major organizing principles of the visual brain, particularly with respect to the visual pathways in the cerebral cortex.

Functional Models of the Two Visual Streams

Our reliance on vision is reflected in the large amount of brain tissue in humans (and in our primate cousins) that is devoted to visual processing. Beyond primary visual cortex, the ascending visual pathways convey information to a complex mosaic of visual areas in the primate cerebral cortex. Despite the complexity of the interconnections between these different areas, two broad 'streams' of projections arising from primary visual cortex in the macaque monkey have been identified: a ventral stream projecting eventually to the inferotemporal cortex and a dorsal stream projecting to the posterior parietal cortex. These streams also receive inputs from a number of subcortical visual structures, including the superior colliculus, which sends prominent projections to the dorsal stream (via the pulvinar). A schematic diagram of these pathways is shown in Figure 1. Although some caution must be exercised in generalizing from monkey to human, a decade or more of neuroimaging evidence suggests that the visual projections from primary visual cortex to the temporal and parietal lobes in the human brain involve a separation into ventral and dorsal streams similar to that seen in the monkey.

Although it is clear that the two streams play different but complementary roles in our visual life, theories regarding the exact nature of the division of labor have changed considerably in recent years. Traditionally, it was thought that the ventral stream plays the critical role in the identification and recognition of objects, whereas the dorsal stream mediates the localization in space of those same objects. This distinction, which has a long history in neurology and neuropsychology, is often referred to as the 'what versus where' hypothesis. Although the evidence that the two streams process different kinds of visual information (object features vs. object location) initially seemed quite compelling, subsequent findings from a broad range of studies in both humans and monkeys have forced a reinterpretation. Recent accounts emphasize differences in the outputs, rather than the inputs, of the two streams. In fact, the consensus now is that both streams process information about object features and about their spatial relations, but each stream transforms this information in different ways for different purposes.

The Perception and Action Model

According to the current understanding, processing in the ventral stream helps to construct the rich and detailed representations of the world that serve as a perceptual foundation for cognitive operations, allowing us to identify objects and events, attach meaning and significance to them, and establish their causal relations. Such operations are essential for accumulating and accessing a visual knowledge base about the world. In contrast, it is believed that the dorsal stream carries out transformations on moment-to-moment information about the location and disposition of objects with respect to the effectors being used in visually guided actions, such as manual prehension, directed at those objects. Thus, the dorsal stream can be regarded as a cortical extension of the dedicated visuomotor modules in the midbrain and brain stem that mediate visually guided movements in all vertebrates. However, although the two streams have different functions and operating principles, in everyday life they have to work together. The perceptual networks of the ventral stream interact with various high-level cognitive mechanisms and enable an organism to select a goal and a particular course of action while the visuomotor networks in the dorsal stream (and their associated cortical and subcortical pathways) are responsible for the programming and on-line control of the particular movements that the chosen action entails.

Studies of Neurological Patients

This division of labor between the two streams is clearly evident in the pattern of deficits and spared

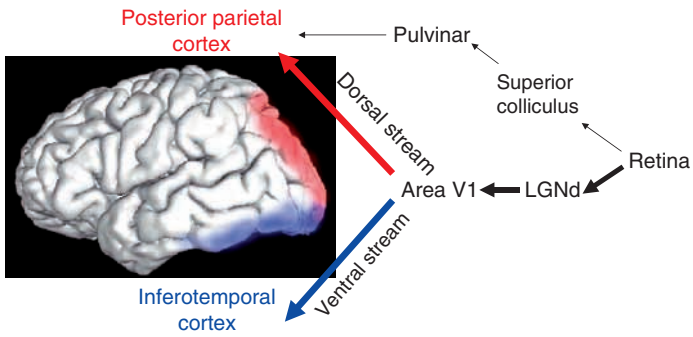

Figure 1 The two streams of visual processing in human cerebral cortex. The retina sends projections to the dorsal part of the lateral geniculate nucleus (LGNd), which projects in turn to primary visual cortex. Within the cerebral cortex, the ventral stream (blue arrow) arises from early visual areas and projects to the inferotemporal cortex. The dorsal stream (red arrow) also arises from early visual areas but projects instead to the posterior parietal cortex. The posterior parietal cortex also receives visual input from the superior colliculus via the pulvinar. The approximate locations of the pathways are shown on a three-dimensional reconstruction of the pial surface of the brain. The two streams, which involve a series of complex interconnections, are marked in blue (ventral stream) and red (dorsal stream). Adapted from Goodale MA and Westwood DA (2004) An evolving view of duplex vision: Separate but interacting cortical pathways for perception and action. *Current Opinion in Neurobiology* 14: 203–211.

visual abilities seen in neurological patients with selective damage to the dorsal or ventral stream. Thus, patients with lesions in the superior regions of the posterior parietal cortex, the major terminus of the dorsal stream, typically have problems using vision to direct a grasp or aiming movement toward the correct location of a visual target placed in different positions in the visual field, particularly the peripheral visual field. This particular deficit is often termed optic ataxia. However, the failure of these patients to locate an object with the hand cannot be construed as a problem in spatial vision: many of them, for example, can describe the relative position of the object in space quite accurately, even though they cannot direct the affected hand toward it. Optic ataxia patients also typically have no difficulty using input from other sensory systems, such as proprioception or audition, to guide their movements. Other research has shown that many of them not only show visual misreaching but also are unable to use visual information to rotate their hand, scale their grip, or configure their fingers properly when reaching out to pick up objects, even though they have no difficulty describing the orientation, size, or shape of those objects. In addition, they do not take into account the positions of potential obstacles when they are attempting to reach out toward goal objects. All of this confirms the critical role that the dorsal stream plays in the visual control of skilled actions. The patients exhibit neither a purely visual nor a purely motor deficit but instead a specific deficit in visuomotor control.

The opposite pattern of deficits and spared abilities has been described in patients with damage to the ventral stream. The best documented case is patient D.F., a young woman who developed a profound visual form agnosia following carbon monoxide poisoning. Structural magnetic resonance imaging showed evidence of diffuse damage consistent with hypoxia but with specific lesions in ventrolateral regions of the occipital cortex, with primary visual cortex remaining largely spared (Figure 2). Even though D.F.'s contrast sensitivity and other low-level visual abilities remain reasonably intact, she can no longer recognize everyday objects or the faces of her friends and relatives, nor can she identify or copy line drawings of common objects or even simple geometric shapes. (If an object is placed in her hand, of course, she has no trouble identifying it by touch.) Remarkably, however, D.F. shows strikingly accurate guidance of her hand and finger movements when she attempts to pick up the very objects she cannot identify (Figure 3). Thus, when she reaches out to grasp objects, her hand opens wider midflight for larger objects than it does for smaller ones, just as is the case in people with normal vision. She also takes into account the position of potential obstacles in the immediate vicinity of a goal object to which she is reaching. Similarly, she rotates her hand and wrist quite normally when she reaches out to grasp objects in different orientations, and she places her fingers optimally around the edges of objects of different shape. At the same time, she is quite unable to distinguish between any of these objects when they are presented to her in simple discrimination tests. She even fails in manual 'matching' tasks in which she is asked to show how wide an object is by opening her index finger and thumb a corresponding amount. D.F.'s spared visuomotor skills are not limited to grasping. She can step

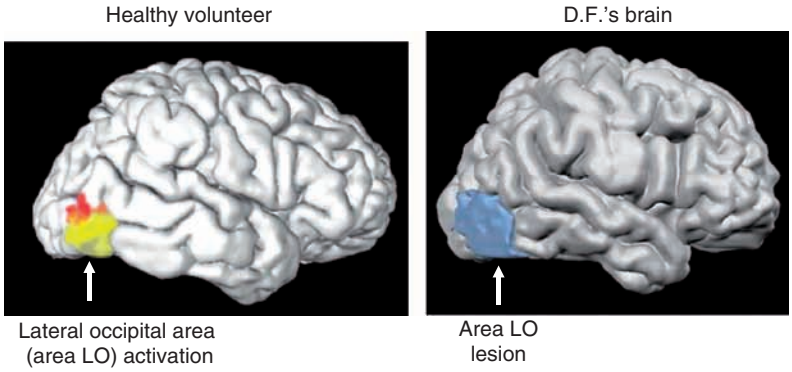

Figure 2 Area LO, a ventral stream area implicated in object recognition (particularly object form), has been localized on the brain of a healthy control subject by comparing fMRI activation to intact versus scrambled line drawings. Note that the lesion (marked in blue) on patient D.F.'s right cerebral hemisphere encompasses all of area LO. Area LO in D.F.'s left hemisphere is also completely damaged. Adapted from Goodale MA and Milner AD (2005) *Sight Unseen: An Exploration of Conscious and Unconscious Vision*. Oxford: Oxford University Press.

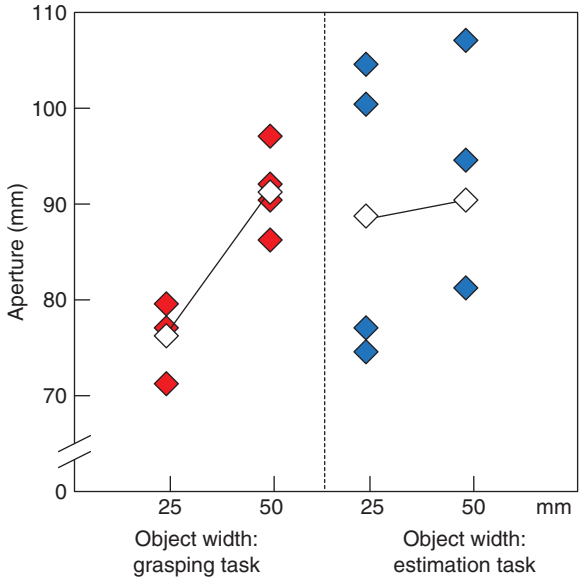

Figure 3 Graph showing the size of the aperture between D.F.'s index finger and thumb during object-directed grasping and manual estimates of object width. Her maximum grip aperture in flight (individual trials shown in red) was sensitive to the size of the target object, opening wider for the 50 mm-wide object than for the 25 mm-wide object. However, her manual estimates of the width of the two objects (individual trials shown in blue) were grossly inaccurate and show enormous variability from trial to trial.

over obstacles during locomotion as efficiently as controls, even though her perceptual judgments about the height of these obstacles are far from normal. In short, a profound loss of form perception coexists in D.F. with a preserved ability to use information about the form of objects to guide a broad range of actions. The contrast between what D.F. can and cannot do is exactly what one would expect in someone with a damaged ventral 'perception' stream but a functionally intact dorsal 'action' stream.

Studies in Nonhuman Primates

Lesion studies in the monkey paint a remarkably similar picture to that seen in neurological patients. For example, monkeys with lesions of inferotemporal cortex (ITC) can orient their fingers in a precision grip to grasp morsels of food embedded in small slots placed at different orientations, even though their ability to discriminate between different orientations is profoundly impaired. In short, such monkeys behave rather like patient D.F.; they have spared visuomotor abilities in the face of a profound visual agnosia. In contrast, monkeys with lesions to the posterior parietal cortex (PPC) show the opposite pattern, behaving like patients with optic ataxia.

In addition to the lesion studies, there is a long history of single-unit work in monkeys showing that cells in ITC and neighboring regions of the superior temporal sulcus are tuned to specific objects and object features – with some cells maintaining their selectivity irrespective of viewpoint, retinal image size, and even color. Moreover, the responses of these cells are not affected by the animal's motor behavior but are instead sensitive to the reinforcement history and significance of the visual stimuli that drive them. It has been suggested that cells in this region might play a role in comparing current visual inputs with internal representations of recalled images, which are themselves presumably stored in other regions, such as the medial temporal lobe, related limbic areas, and the prefrontal cortex, with which the ITC is reciprocally connected. In fact, sensitivity to particular objects can be created in ensembles of ITC cells simply by training the animals to discriminate between different objects. There is also evidence for a specialization within separate regions of the monkey's ventral stream for the coding of certain categories of objects, such as faces and hands,

which are of particular social significance to the monkey. Finally, studies using a binocular rivalry paradigm, in which competing images are presented at the same time to the two eyes, have shown that cells in ITC are tuned to what the monkey reports seeing on a particular trial and not simply to what is present on the two eyes (which remains invariant). However, neurons earlier in the visual system, such as those in area V1, do not show these correlations, instead responding equally strongly no matter which of the competing stimuli dominates the monkey's current perception. Taken together, these results provide support for the claim that the ventral stream plays a critical role in delivering the contents of our conscious percepts. These and other studies also suggest that 'object-based' descriptions are created in the ventral stream – that is, ones that can generalize across different viewpoints. These descriptions then form the basic raw material for visual perception, recognition memory, and long-term representations of the visual world.

In sharp contrast to the activity of cells in the ventral stream, the responses of cells in the dorsal stream are greatly dependent on the concurrent motor behavior of the animal. Thus, separate subsets of visual cells in the PPC have been shown to be implicated in visual fixation, pursuit and saccadic eye movements, visually guided reaching, and the manipulation of objects. Moreover, this motor involvement is quite specific. Thus, neurons in the lateral bank of the intraparietal sulcus (LIP) code the location of a target for a saccadic eye movement, whereas neurons in more medial regions of the intraparietal sulcus extending into the superior regions of the PPC code the location for a manual aiming movement to the same target (the parietal reach region (PRR)). Neurons in the anterior intraparietal sulcus (area AIP), which fire when the monkey manipulates an object, have been shown to be sensitive to the visual features of the object, such as its size and orientation, that determine the posture of the hand and fingers during a grasping movement. The PPC is also intimately linked with premotor cortex, the superior colliculus, the pontine nuclei, and other brain stem areas – structures that have been implicated in various aspects of the visual control of eye, limb, and body movements. In summary, the networks in the dorsal stream have the functional properties and interconnections that one might expect to see in a system concerned with the moment-to-moment control of visually guided actions.

Human Neuroimaging

With the advent of neuroimaging, particularly functional magnetic resonance imaging (fMRI), it has been possible to map the ventral and dorsal streams in the human brain. These studies have revealed an organization of visual areas in the human brain that is remarkably similar to that seen in the macaque. Although clear differences in the topography of some of the areas emerge as one moves from monkey to human, the functional separation into a ventral and a dorsal pathway is clearly preserved. Thus, different areas in the inferior occipitotemporal region appear to be specialized for the processing of the color, texture, and form of objects. Also, as in the monkey, activation studies have identified regions in the ventral pathway for the processing of faces that are distinct from those involved in the processing of other objects. In the dorsal stream, regions have been found that are activated when subjects engage in visually guided movements such as saccades, reaching movements, and grasping, corresponding respectively to areas LIP, PRR, and AIP, as described in the monkey. Thus, the neuroimaging data confirm that the human cerebral cortex includes both dorsal and ventral visual streams very much like those of the monkey, and that these streams have functional characteristics resembling those first identified in the monkey. In other words, areas in the posterior parietal cortex are involved in the visual control of action, whereas areas in the occipitotemporal region play a critical role in object recognition.

The Operating Principles of the Dorsal and Ventral Streams

The existence of two separate cortical visual pathways reflects the fact that perception and action each require their own particular transformations to be carried out on the incoming visual information. To be able to grasp an object successfully, for example, it is essential that the brain (specifically the dorsal stream) compute the actual (absolute) size of the object and its orientation and position with respect to the observer. Moreover, the information about the orientation and position of the object must be computed in egocentric frames of reference – in other words, in frames of reference that take into account the orientation and position of the object with respect to the effector that is to be used to perform the action. However, the time at which these computations are performed is also critical. Observers and goal objects rarely stay in a static relationship with one another and, as a consequence, the egocentric coordinates of a target object can often change dramatically from moment to moment. For this reason, it is essential that the required coordinates for action be computed immediately before the movements are initiated. For the same reason, it would be counterproductive for

these coordinates (or the resulting motor programs) to be stored in memory. In summary, the vision-for-action systems in the dorsal stream need to work in an on-line mode.

The situation is quite different for perception, in terms of both the frames of reference used to construct the percept and the time period during which that percept (or the information it provides) can be accessed. Vision for perception appears not to rely on computations about the absolute size of objects or their egocentric locations. Instead, the perceptual system in the ventral stream computes the size, location, shape, and orientation of an object primarily in relation to other objects and surfaces in the scene. Encoding an object in a scene-based frame of reference permits a perceptual representation of the object that preserves the relations between the object and its surroundings without requiring precise information about its absolute size or its exact position with respect to the observer. Indeed, if our perceptual machinery attempted to deliver the real size and distance of all the objects in the visual array, the computational load on the system would be astronomical.

The products of perception also need to be available over a much longer time scale than the visual information used in the control of action. We may need to recognize objects that we have seen minutes, hours, days, or even years ago. To achieve this, the coding of the visual information has to be somewhat abstract – transcending particular viewpoint and viewing conditions. By working with perceptual representations that are object or scene based, we are able to maintain the constancies of size, shape, color, lightness, and relative location over time and across different viewing conditions. There is much debate in the object recognition literature about how this is accomplished. Some of the mechanisms underlying object perception might use a network of viewer-centered representations of the same object, others might use an array of canonical representations, and still others might be truly 'object-centered.' However, whatever the nature of the coding might be, it is the identity of the object and its location within the scene, not its disposition with respect to the observer, that is of primary concern to the perceptual system. In other words, we can recognize objects that we have seen before, even though the position of those objects with respect to our body might have changed considerably since the last time we saw them. Of course, this stored information is not only available for mediating recognition of previously encountered objects but can also potentially contribute to the control of our movements when we are working in off-line mode.

The contribution of the ventral stream to the off-line control of object-directed actions, such as grasping, can be seen in the differential effect that delaying such actions has on the behavior of patients with lesions of the ventral or dorsal stream. Patient D.F., for example, who is able to grasp an object accurately in real time, shows no scaling of her grasp to the size of the target object when the object is removed after an initial viewing period and she is asked to 'pantomime' how she would have picked it up (Figure 4). Evidently, she has no working memory of the object – not because her memory system is impaired but, rather, because her ventral stream lesion prevents her from perceiving the dimensions

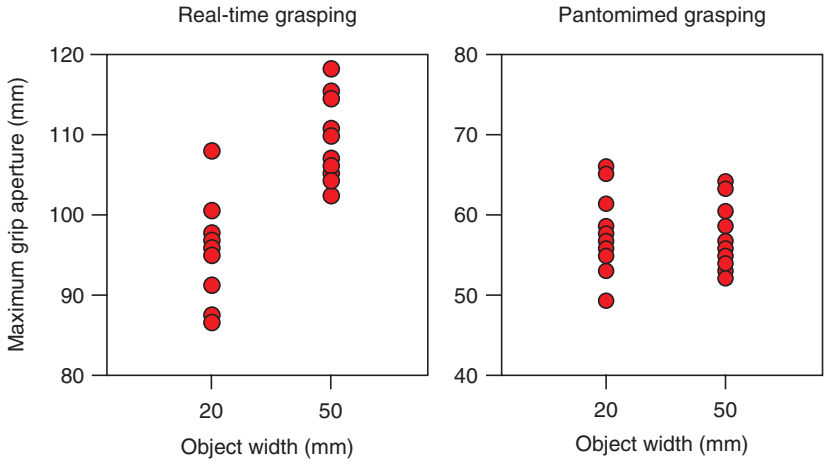

Figure 4 Graph showing the maximum grip aperture for real-time and pantomimed grasping movements performed by patient D.F. When D.F. picked up an object in real time, her grip aperture was correlated with the width of the target object (left). When she was shown one of the objects and then 2 s later was asked to pantomime how she would pick it up (the object had been removed), her grip aperture showed no relation to the size of the object (right). Adapted from Goodale MA, Jakobson LS, and Keillor JM (1994) Differences in the visual control of pantomimed and natural grasping movements. *Neuropsychologia* 32: 1159–1178.

of the object in the first place. Optic ataxic patients have exactly the converse problem (Figure 5). Because of the damage to their dorsal stream, they show no sign of scaling their grasp to the size of objects in real time. However, in the pantomime task, they perform just like healthy subjects – presumably because they are no longer activating the visuomotor mechanisms in their damaged dorsal stream but are engaging instead their intact ventral stream. These results illustrate what may be the most important single task of the ventral stream: it allows us to use vision off-line, providing a bridge from the past to the present.

Although the dorsal stream does not appear to have a visual memory, the visuomotor activities to which it contributes clearly do benefit from experience. In fact, most of our visually guided actions are as skilled as they are precisely because they are well honed by practice. It seems likely that as our initially slow and painstaking efforts become more automatic, the contribution of the ventral stream to visual control retreats and is replaced by more streamlined circuitry involving the dorsal stream, related frontal cortical areas, and subcortical structures in the brain stem and cerebellum.

The differences in the spatial metrics and frames of reference used by the dorsal and ventral streams, mentioned previously, can be seen in the way in which normal observers respond to stimuli that are distorted by pictorial illusions. For example, the opening of a grasping hand is typically insensitive to the Ebbinghaus illusion, in which a target disk surrounded by smaller circles appears to be larger than the same disk surrounded by larger circles, despite the fact that grip opening is exquisitely sensitive to real changes in the size of the target disk (Figure 6). In fact, grip aperture is not affected by the illusion even when the hand and target are occluded from view during the performance of the action, indicating that on-line visual feedback during grasping is not required to 'correct' an initial perceptual bias induced by the illusion.

Of course, the perception–action model does not predict that actions must be immune to all visual illusions. Some illusions, after all, arise so early in visual processing that they would be expected to affect both perception and action. Illusions that depend on 'top-down' knowledge of the world, however, would certainly be predicted to leave actions unaffected. One particularly striking example of such an illusion is the 'hollow face' illusion, in which knowledge of what faces look like impels observers to see the inside of a hollow mask as if it were a normal protruding face (Figure 7). Despite the fact that observers cannot resist this compelling illusion, rapid actions that they direct at the face are immune to it. Thus, when observers are asked to flick off a small 'buglike' target stuck on the face, they unhesitatingly reach out to the correct point in space, even if this means that their hand goes inside the hollow mask. This striking dissociation between what one sees and what one does provides a dramatic demonstration of the simultaneous engagement of two parallel visual systems – each constructing its own version of reality.

Conclusion

Although similar (but not identical) visual information about object shape, size, orientation, and

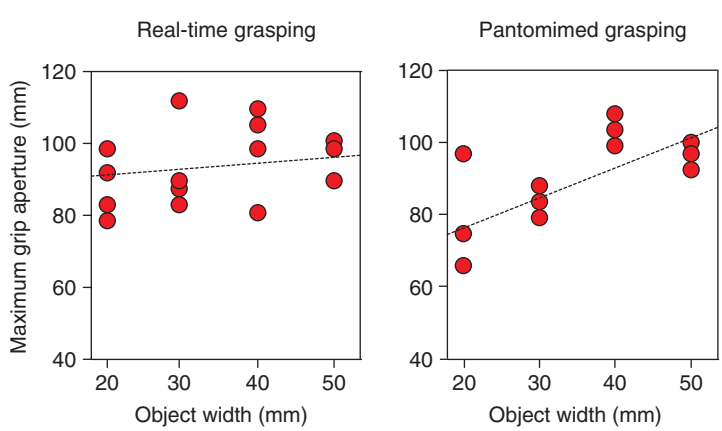

Figure 5 Graph showing the maximum grip aperture for real-time and pantomimed grasping movements performed by optic ataxia patient I.G. When I.G. picked up objects in real time, her grip apertures were uncorrelated with the widths of the target objects (left). However, when she was shown one of the objects and 5 s later was asked to pantomime how she would pick it up (the object had been removed), the correlation between her pantomimed grip apertures and the widths of the objects became highly significant (right). Adapted from Milner AD, Dijkerman HC, Pisella L, et al. (2001) Grasping the past: Delay can improve visuomotor performance. *Current Biology* 11: 1896–1901.

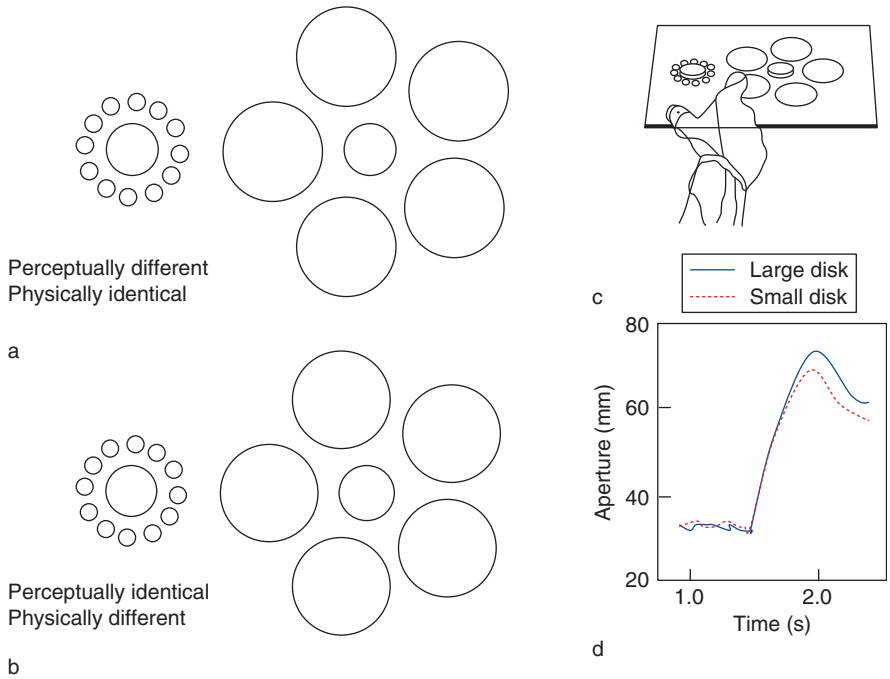

Figure 6 The effect of a size-contrast illusion on perception and action. (a) The traditional Ebbinghaus illusion in which the central circle in the annulus of larger circles is typically seen as smaller than the central circle in the annulus of smaller circles, even though both central circles are actually the same size. (b) The same display, except that the central circle in the annulus of larger circles has been made slightly larger. As a consequence, the two central circles now appear to be the same size. (c) A three-dimensional version of the Ebbinghaus illusion. Participants are instructed to pick up one of the two three-dimensional disks placed either on the display shown in panel a or the display shown in panel b. (d) Two trials with the display shown in panel b, in which the participant picked up the small disk on one trial and the large disk on another. Although the two central disks were perceived as being the same size, the grip aperture in flight reflected the real, and not the apparent, size of the disks. Adapted from Aglioti S, DeSouza J, and Goodale MA (1995) Size-contrast illusions deceive the eyes but not the hand. *Current Biology* 5: 679–685.

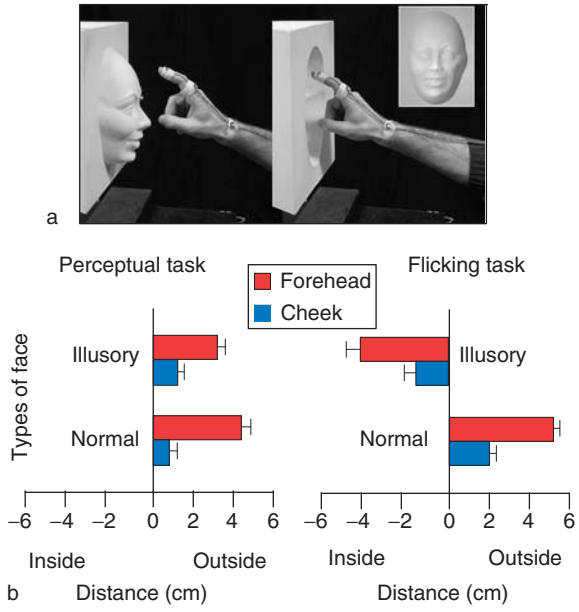

Figure 7 Perceptual judgments and visuomotor control with the hollow-face illusion. (a) A small magnet was placed on either the cheek or the forehead of the normal face (left) or the hollow mask (right). Participants were required either to flick the magnet from the normal or illusory (actually hollow) face or to estimate its distance psychophysically. The inset shows a photograph of bottom-lit hollow face, in which the illusion of a normal convex face is evident. (b) (Left) The mean psychophysical (perceptual) judgments of the apparent position of the magnets on the illusory and normal face with respect to the reference plate from which the two displays either protruded or receded. Note that participants perceived the hollow face as protruding forward like the normal face. (Right) The mean distance of the hand at the moment the participant attempted to flick the target off the cheek or forehead of the illusory (actually hollow) or the normal face. In the case of the illusory face, the end points of the flicking movements corresponded to the actual distances of the targets, and not to consciously seen distances. Error bars indicate the standard error of the mean. Adapted from Kroliczak G, Heard P, Goodale MA, and Gregory RL (2006) Dissociation of perception and action unmasked by the hollow-face illusion. *Brain Research* 1080: 9–16.

location is processed by both cortical visual systems, the transformational algorithms that are applied to these inputs are uniquely tailored to the function of each system. It is the nature of the functional requirements of perception and action that lies at the root of the division of labor that has emerged in the ventral and dorsal visual projection systems of the cerebral cortex during primate evolution.

See also: Active Perception; Agnosia; Attentional Mechanisms in Ventral Pathway; Binocular Rivalry; Blindsight: Residual Vision; Contextual Interactions in Visual Perception; Parietal Cortex and Spatial Attention; Perception and Eye Movements; Visual System: Multiple Visual Areas in Monkeys; Visual Cortex in Humans.

Further Reading

Aglioti S, DeSouza J, and Goodale MA (1995) Size-contrast illusions deceive the eyes but not the hand. *Current Biology* 5: 679–685.

Andersen RA and Buneo CA (2003) Sensorimotor integration in posterior parietal cortex. *Advances in Neurology* 93: 159–177.

Culham JC and Kanwisher NG (2001) Neuroimaging of cognitive functions in human parietal cortex. *Current Opinion in Neurobiology* 11: 157–163.

Goodale MA and Humphrey GK (1998) The objects of action and perception. *Cognition* 67: 181–207.

Goodale MA, Jakobson LS, and Keillor JM (1994) Differences in the visual control of pantomimed and natural grasping movements. *Neuropsychologia* 32: 1159–1178.

Goodale MA and Milner AD (2005) *Sight Unseen: An Exploration of Conscious and Unconscious Vision.* Oxford: Oxford University Press.

Goodale MA, Milner AD, Jakobson LS, and Carey DP (1991) A neurological dissociation between perceiving objects and grasping them. *Nature* 349: 154–156.

Goodale MA and Westwood DA (2004) An evolving view of duplex vision: Separate but interacting cortical pathways for perception and action. *Current Opinion in Neurobiology* 14: 203–211.

Grill-Spector K (2003) The neural basis of object perception. *Current Opinion in Neurobiology* 13: 159–166.

James TW, Culham JC, Humphrey GK, Milner AD, and Goodale MA (2003) Ventral occipital lesions impair object recognition but not object-directed grasping: An fMRI study. *Brain* 126: 2463–2475.

Kroliczak G, Heard P, Goodale MA, and Gregory RL (2006) Dissociation of perception and action unmasked by the hollow-face illusion. *Brain Research* 1080: 9–16.

Logothetis NK (1998) Single units and conscious vision. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 353: 1801–1818.

Milner AD, Dijkerman HC, Pisella L, et al. (2001) Grasping the past: Delay can improve visuomotor performance. *Current Biology* 11: 1896–1901.

Milner AD and Dyde RT (2003) Why do some perceptual illusions affect visually guided action, when others don't? *Trends in Cognitive Science* 7: 10–11.

Milner AD and Goodale MA (2007) *The Visual Brain in Action,* 2nd edn. Oxford: Oxford University Press.

Perenin M-T and Vighetto A (1988) Optic ataxia: A specific disruption in visuomotor mechanisms: I. Different aspects of the deficit in reaching for objects. *Brain* 111: 643–674.

Tanaka K (1996) Inferotemporal cortex and object vision. *Annual Review of Neuroscience* 19: 109–139.

Tootell RB, Tsao D, and Vanduffel W (2003) Neuroimaging weighs in: Humans meet macaques in 'primate' visual cortex. *Journal of Neuroscience* 23: 3981–3989.

VanEssen DC (2005) Corticocortical and thalamocortical information flow in the primate visual system. *Progress in Brain Research* 149: 173–185.

Weiskrantz L (1997) *Consciousness Lost and Found: A Neuropsychological Exploration.* Oxford: Oxford University Press.

Vision: Light and Dark Adaptation

A Reeves, Northeastern University, Boston, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

The visual system works effectively over an enormous range of light intensities, from starlight to bright sunlight. This is made possible by the system's remarkable ability to adjust its operating level to match the ambient illumination. Light adaptation is the adjustment of the eye to higher levels of illumination; dark adaptation is the reverse. The study of adaptation is important for a variety of practical reasons, as well as for an understanding of basic visual function. Adaptation is mediated by a combination of actions. Narrowing the pupil aperture in bright light is one of these. In species like ours, with round rather than slit pupils, constriction of the pupil reduces retinal illumination by about a decade (10 times), a small part of the entire adaptive range (14 decades). The main mechanisms of adaptation in the human eye lie within the receptors and neural structures of the retina. The human retina, like that of many species, is duplex; it possesses both a photopic system that works at the high levels of illumination provided by daylight, and a scotopic system that works at the low levels provided by starlight and moonlight. There is also an intermediate, or mesopic, range in twilight, in which both systems function. The photopic system employs cone receptors; the scotopic, rods. The large range of adaptation seen in the human eye is achieved by a change from one system to the other, as well as by adjustments of sensitivity within each system. In other species, pigment and/or receptor migration may also reduce the impact of high levels of illumination.

Regimes of Adaptation

The two adaptive regimes, scotopic and photopic, operate quite differently. In starlight there are few ambient photons, and so the ability of the dark-adapted visual system to detect individual photons becomes critical. Rod photopigments can respond to individual photons, and a massive amplification in the main rod pathway permits such minute events to initiate neural activity. Indeed, the efficiency of this scotopic system in starlight approaches the ideal level set by physical optics. Moonlight provides more ambient photons than does starlight, and at light levels over about 300 photons/rod/second, the massive amplification saturates the main rod pathway and the rod signals must now travel through a second rod pathway to initiate vision, one with considerably less amplification. In twilight, both cones and rods are active; the transition from scotopic to photopic vision is indicated by the term 'mesopic.' During this transition, cones still act like rods in that they respond to small numbers of photons. In sunlight, the entire scotopic system saturates – the rods no longer provide any useful information at all – and the photopic system takes over. At these higher light levels, the photopic system is protected by a progressive light adaptation of the cones from the saturation that effectively turns off the scotopic system. The photopic system now begins to differ fundamentally from the main rod pathway in that it no longer attempts to record every photon – there are so many that this is pointless – but rather it encodes the contrasts which specify the visible surfaces in the world as darker or lighter than their surrounds. Encoding contrast implies discounting the ambient illumination level (the space-averaged number of photons); so long as there is enough light for the world to be visible everywhere, just how brightly it is illuminated is of no great consequence. At ever higher light levels, sensitivity to contrast is maintained – one can see details on a bright beach, or in snow illuminated by the midday sun, as long as the cones are sufficiently light adapted.

Dark Adaptation

The time course of dark adaptation can be illustrated by the solid circles in Figure 1, which show thresholds obtained by first exposing the eye to a very bright light, and then plunging it into total darkness. The detection threshold, which is the intensity of light just needed for an observer to detect a brief test flash, was determined at regular intervals as dark adaptation progressed, as shown by the successive solid circles (open circles are discussed later). The plot reveals a descending curve, as decreasing amounts of light are needed to see the test flash as time goes by. Intensities below the threshold are invisible; those above it, visible. The curve has two limbs. In the first 2 min, detection thresholds drop rapidly and then taper off on a plateau. After a few more minutes, the curve again drops rapidly, reaching a final level, the 'absolute' threshold, after about 25 min. The vertical scale is logarithmic: the entire decrease in threshold of 3.3 $\log_{10}$ units from start to end corresponds to a change in test intensity of 2000 times. These progressive drops in threshold correspond to increases in visual sensitivity, first of the photopic system and then of the scotopic system. The change from one system to the other is called the rod–cone break, which occurs at about 7 min in Figure 1. There is a corresponding shift, known as the Purkinje shift, in the peak spectral

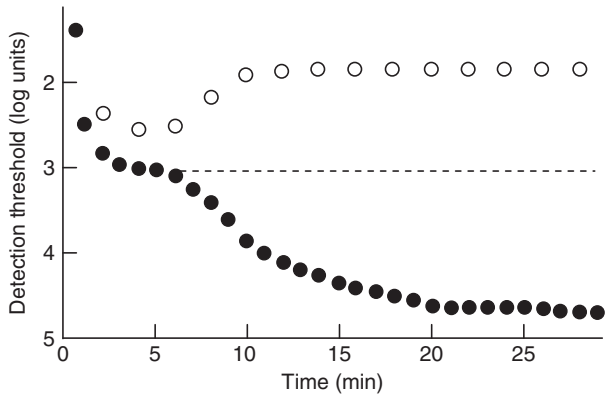

Figure 1 Dark adaptation. The solid black circles show the threshold for detection (in log units) against the length of time the eye has been in darkness (in minutes). The dashed line is an extrapolation of the upper or photopic limb of the implied curve. The lower limb is scotopic. The open circles show a specific threshold, that for just noticing hue; this threshold rises above the extrapolated line once the rods have begun to recover. Flashes were green and presented parafoveally (6° from the fovea). The observer initially adapted to a large white 4 lumens per cm^2). Adapted Lie I (1963) Dark adaptation and the photochromatic interval. *Documenta Ophthalmologica* 17: 411–510.

sensitivity of the eye from 555 nm (the photopic maximum) before the break, to 500 nm, the maximum sensitivity of rhodopsin (the rod photosensitive pigment), after it. Test lights with intensities just above detection threshold appear colored if flashed before the break, indicating cone function, and colorless after it, indicating rod function. The solid circles in Figure 1 show a dark adaptation curve produced by a middle-wavelength test flash delivered to the parafovea – that is, a retinal position outside the fovea, where both rods and cones are numerous. When the dark adaptation curve is measured with long-wavelength tests invisible to the rods, or with test flashes restricted to the rod-free fovea, there is only a photopic limb, as indicated by the dashed line. The two limbs of the dark adaptation curves can be translated vertically (on a logarithmic scale of intensity) by altering the preadaptation light in wavelength composition or in level, and by altering the test flash in wavelength, size, location on the retina, and duration. These are all factors governing visual sensitivity; for example, a larger test flash will lower thresholds due to spatial summation in the retina. However, if the two limbs do appear, their time courses will be as shown. This invariance in recovery provides behavioral evidence that the temporal recovery of each class of photoreceptor is independent of the adaptive states of the other classes of photoreceptor.

Given that the recovery of each class of photoreceptor can be followed independently, it becomes easier to study the biochemical and physiological processes which account for such recovery. Vision begins with phototransduction – that is, with the capture of an individual photon by a visual pigment molecule (rhodopsin) in a photoreceptor. Capture has two effects. First, it causes the isomerization (change of shape) of a small component of the molecule, retinal, from its 11-*cis* to its all-*trans* form. This generates various forms of metarhodopsin and a G-protein enzyme cascade which massively amplifies the minute photon-capture event, permitting signaling molecules to modify the cell membrane potential and indicate the presence of light to other retinal neurons. Hence after extended dark adaptation, individual photons of light can be detected. Second, capture will also bleach the photopigment molecule, which becomes transparent. Only unbleached photopigment molecules respond to light. As dark adaptation progresses, more and more of the photopigment molecules recover from the bleached state and again become available for capturing photons. It might seem that this effect could account for the recovery of the detection threshold seen in Figure 1. However, threshold elevation is out of all proportion to any loss due to bleaching; for example, a bleach of only 20% of the pigment can raise rod thresholds more than 1000 times, and thresholds remain considerably elevated even when nearly all of the photopigment has recovered from bleaching. Therefore the photon-catching ability of the photoreceptors alone does not regulate visual threshold. It is thought instead that the various photoproducts of bleaching – the various metarhodopsins and free opsin – which remain in the receptors when the eye is plunged into darkness act to 'veil' the incoming test flash and thus elevate threshold. It is their progressive removal which permits threshold to recover in the dark. The term 'veil' is used by analogy, as thresholds for foveal test flashes can also be elevated by glare sources such as automobile headlamps, which, when placed in the periphery of the visual field, generate a veil of real light across the eye. In 1933, WH Stiles and BH Crawford (see Wyszecki and Stiles in 1982) showed psychophysically that, following bleaches, the human visual system behaves as though it experienced a veil equivalent to real light. Unlike the real light with which it can be equated, equivalent light is usually not visible (one sees darkness, not an afterimage, during the course of dark adaptation), but this is no paradox, as real lights also become invisible if, like equivalent light, they are retinally stabilized. Equivalent light is generated in the receptors in at least two different ways. The photoproducts of bleaching may either occasionally revert to activated rhodopsin and generate false

alarms (reports of light where none exists), or they may act directly on the phototransduction mechanism itself. At the start of dark adaptation, the scotopic threshold is primarily controlled by this second mechanism, which is initially dominant but which recovers relatively quickly, thus accounting for the relatively fast initial drop of threshold following the rod–cone break in Figure 1. The subsequent slow recovery of scotopic vision to its final absolute threshold is primarily due to a slow reduction in the frequency of the photonlike events contributed by the first mechanism.

The classic recovery curve shown by solid circles in Figure 1, the curve obtained when each photoreceptor class is effectively isolated, applies if the observer 'detects' the presence of the test flash. Detection can be based on any visual sensation at all; the observer is not required to classify the sensation. However, dark adaptation may be measured with test flashes whose thresholds are determined by a specific sensation. For example, the specific threshold for seeing color, which achieves its minimum after 2 min, may rise by up to a decade after several more minutes in the dark (Figure 1, open circles). This, the Lie effect, does not occur in the rod-free fovea. It illustrates a so-called rod–cone interaction in which recovering rod vision interferes with the perception of hue mediated by the cones. This interaction must occur at some postreceptoral neural site, not in the receptors themselves, since rod and cone detection thresholds recover independently of one another.

The classic recovery curve shown by solid circles in Figure 1 can be violated in other ways as well. For example, thresholds mediated by the pathways encoding rapid luminance flicker and hue may be elevated by 10 to 100 times when the adaptation light is turned off, and remain elevated for the first minute or so of dark adaptation. A case of especial interest is that of transient tritanopia, the temporary blue blindness which can occur after turning off a long-wavelength adapting field, because in this case the field light-adapts the long- and middle-wave sensitive cones but not the short-wave cones responsible for detecting the test flash. The normal course of dark adaptation is therefore reversed at a postreceptoral neural site, not by processes in the receptors. These dramatic violations of the usual course of dark adaptation depend strongly on the wavelength composition and luminance of the adaptation light. They have not yet received detailed explanation at the physiological level, although some of the data are compatible with an opponent-process account. Summarizing, the continual recovery of vision shown by the solid symbols in Figure 1, though taken as the standard, is not always representative.

Light Adaptation

One measure of the extent of light adaptation is provided by the incremental threshold. This is the threshold for detecting a test flash against a steady, uniformly lit background to which the observer has been adapted for several minutes. Some typical results are shown in Figure 2(a). Figure 2(a) shows a plot of the incremental threshold as it depends on the luminance of the background. At low levels, the background has no effect, but as the background is raised, the threshold rises, at first gradually and then more rapidly. At higher levels, there is a second plateau followed by a second rise. As in dark adaptation curves, the two limbs of the plot can be attributed to the participation of the rod and cone systems. When the eye is adapted to low levels, the rod system is the most sensitive, but at higher levels of light adaptation, scotopic sensitivity drops below that of the cones.

As in the case of the dark adaptation curve, the limbs of the increment threshold curve depend on the testing conditions, such as the size of the stimuli, their location on the retina, and their spectral composition. The example just examined is typical of orange–yellow stimuli presented against a blue–green background. Other conditions result in a less conspicuous separation of the two retinal systems. A blue stimulus, which favors the scotopic system, presented against a red adaptation, which is effective in reducing the sensitivity of the photopic system, can produce an increment threshold curve in which the scotopic system is completely uncovered, as shown in Figure 2(b).

Different sectors of the curve shown in Figure 2(b) may be distinguished for theoretical as well as descriptive purposes. Throughout the middle of the range the curve slope is nearly 1.0 on logarithmic axes, and therefore approximates Weber's law, that the ratio of increment threshold to background intensity is constant. At the upper end the curve swings steeply upward; this is known as the saturation region (if saturation were complete, threshold could never be reached, no matter how bright the test light). At the lower end the curve bends and becomes horizontal, thus marking the region where the background is too weak to have an effect. The inset shows the stimulus arrangement in which a small test is flashed on a large, steady, adapting field to which the eye is continuously exposed; if the test and field have the same size, other color mechanisms are revealed. Several theoretical accounts of the incremental threshold curve have invoked Fechner's original concept of an intrinsic activity or noise within the receptor systems (I_0) that combines with the excitation produced by

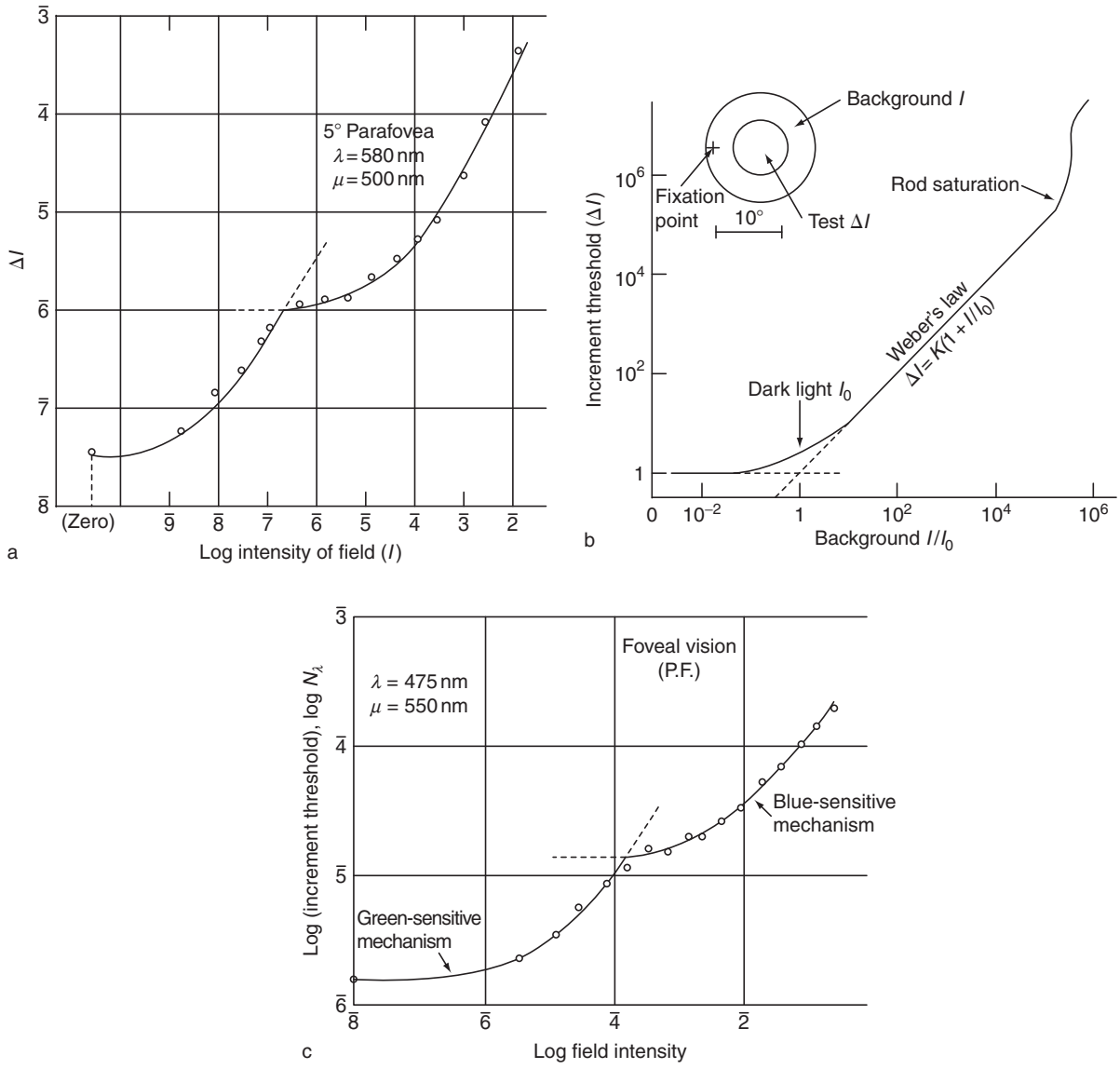

Figure 2 The log incremental threshold (ΔI) is plotted against the intensity of the background field (I) (units are arbitrary, but on a $\log_{10}$ scale). Zero on the abscissa represents no background. The lower branch describes the sensitivity of the scotopic system; the upper, that of the photopic. (b) The scotopic increment threshold is plotted against the intensity of the field or background, expressed relative to the 'dark light' (I_0) appropriate for the receptor class (rods) being studied. The dark light and Weber's law portions apply to rods and cones equally. Conditions favored detection by rods over a very wide range. The upper section illustrates saturation in rods, with cones just starting to take over at the very top. (Saturation is absent in cones.) The smooth curve is a theoretical prediction (see text). Axes are in $\log_{10}$ units. (c) Increment threshold for a 475 nm test flash superimposed on a steady, 550 nm field of the intensity shown (in arbitrary units) on the abscissa. Curves obey $\Delta I = K(I + I_0) = k(1 + I/I_0)$. (a–c) From Wyszecki G and Stiles WS.

the background (I) to raise the threshold (ΔI). The smooth curve in Figure 2(b) is of this form, $\Delta I = K(I + I_0) = k(1 + I/I_0)$, and it accounts for all but the saturating portion of the data with an appropriate choice of the constant $K = k/I_0$. In the same vein, dark adaptation and incremental threshold curves have been related to one another with some success by postulating that I_0 characterizes the dark light or equivalent light described previously (see the section titled 'Dark adaptation'). It is important to emphasize that the curve shown in Figure 2(b) plots the effect of light adaptation on threshold for detection, but not necessarily on other aspects of visual function. For example, the central section of a curve plotted on the same logarithmic coordinate system which expresses the effects of field intensity on perceived brightness has a slope near 0.5 rather than one of 1.0, as shown by LG Brown and ME Rudd; these authors take the slower growth of brightness (as compared to threshold) to be evidence for an adaptive gain control. Our

knowledge of light adaptation as presented by detection thresholds is fairly complete, but this is not so for all aspects of visual function.

The procedures of light and dark adaptation are useful for investigating color vision. If, for example, incremental thresholds are measured with monochromatic test flashes presented against monochromatic backgrounds, the photopic sector shown in Figure 2(a) may break up, depending upon the particular wavelengths used, into two or more constituent limbs. By measuring the incremental threshold function systematically for a wide range of test and adaptation wavelengths, it is possible to determine the spectral sensitivities of these limbs and thus to define the color processes they represent, as demonstrated by WS Stiles and illustrated in Figure 2(c) for the blue- and green-sensitive mechanisms isolated with a 475 nm test flash exposed on a 550 nm field. Color processes may also be inferred from dark adaptation curves measured with monochromatic test stimuli following preexposure to monochromatic backgrounds.

The time course of chromatic light adaptation depends not only upon stimulus wavelength and intensity, as just discussed, but also, although to a lesser extent, on the spatial configuration of the stimuli. For example, adaptation to a green field, but not a blue field, is hastened when the spatial frequencies and orientations of the test and the field match. These data suggest that the neural processes subserving spatial vision integrate over the hue pathways in a manner which depends on their states of adaptation.

Rapid Adaptation Change

The initial changes in visual sensitivity, those that accompany the very onset of light and dark adaptation, are also of considerable interest. These have been measured by superimposing a small, brief test flash upon a large adaptation background that is switched on and off periodically. The timing between the test flash and the adaptation light is under experimental control. The detection threshold measured when the adaptation background is on is an incremental threshold, while that measured with it off is an absolute threshold (i.e., a threshold in darkness). This type of experiment, therefore, helps to define the transition from one type of threshold measure to the other. Thresholds (Figure 3) for detecting the test flash begin to rise shortly before the onset of the adaptation light, reach a maximum at onset, drop, and then climb to a second but lower maximum at the time of offset. Sensitivity then returns fairly quickly. The fact that the detection threshold starts to rise before the onset of the adaptation stimulus was attributed by its discoverer, Brian Crawford, to the speed with which

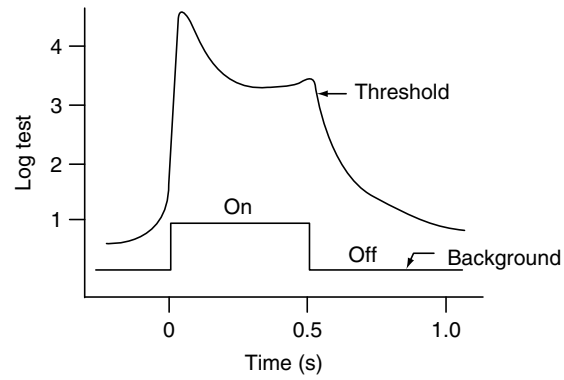

Figure 3 Incremental threshold as measured against a slowly flashing background. The 0.5 s epoch that the background is on is indicated. From Crawford BH (1947) Visual adaptation in relation to brief conditioning stimuli. *Proceedings of Royal Society of (London)* B134: 283.

information travels through the visual system. The relatively weak test flash gives rise to signals that travel more slowly to the level of the visual system where the threshold decision is made than do those of the stronger adaptation stimulus. This explanation does not account for the second maximum at field offset. Another hypothesis points to the large visual responses at field offset and onset seen in the electroretinogram; such responses may momentarily saturate the visual pathway responsible for detection. Yet further theoretical explanations have been developed over the years and are categorized under the heading of visual masking; none are definitive.

Electrophysiological Evidence

The phenomena of dark and light adaptation have been investigated electrophysiologically, using single-unit recording, single-cell recording, and electroretinography in various animal species, or just electroretinography in humans, with the goal of explaining the changes that occur with extended exposure to light or darkness. Retinal potentials often show a progressive increase in responsiveness that advances with dark adaptation. They are thus useful because they provide objective information regarding the visual process away from the threshold conditions to which psychophysics is limited. Although the changes seen with light and dark adaptation resemble those seen psychophysically, exact comparisons between data provided by the two methods are difficult because conditions are rarely comparable: retinal electrophysiology is typically conducted with large, bright stimuli to obtain measurable responses, whereas psychophysics is usually conducted with localized near-threshold test flashes. This is not necessarily the case, however. Experiments with suitable procedures have effected such comparisons and thus have helped to clarify

understanding of the adaptation process. Even the reduction of blue sensitivity found psychophysically at the start of dark adaptation (transient tritanopia) has been seen in the electroretinogram by D van Norren. Thus although a complete description of light and dark adaptation remains some way off, various aspects of these processes have yielded to intense scrutiny at both the electrophysiological and the behavioral levels.

See also: Color Vision; Photoreceptors: Physiology; Phototransduction; Retinal Pharmacology: Inner Retinal Layers; Retinal Development: An Overview; Visual Development.

Further Reading

Brown LG and Rudd ME (1998) Evidence for a noise gain control mechanism in human vision. *Vision Research* 38: 1925–1933.

Crawford BH (1947) Visual adaptation in relation to brief conditioning stimuli. *Proceedings of Royal Society of (London)* B134–283.

Hecht S, Haig C, and Chase AM (1937) The influence of light adaptation on subsequent dark adaptation of the eye. *Journal of General Physiology* 20: 831–850.

Hess RF, Nordby K, and Sharpe LT (1990) *Night Vision.* Cambridge: Cambridge University Press.

Kaiser PK and Boynton RM (1996) *Human Color Vision,* 2nd edn. Washington DC: Optical Society of America.

Lamb TD (2004) Photopigments and the biophysics of transduction in cone photoreceptors. In: Gegenfurtner KR and Scharpe LT (eds.) *Color Vision: From Genes to Perception.* Cambridge: Cambridge University Press.

Lamb TD and Pugh EN Jr. (2004) Dark adaptation and the retinoid cycle of vision. *Progress in Retinal and Eye Research* 23: 307–380.

Lie I (1963) Dark adaptation and the photochromatic interval. *Documenta Ophthalmologica* 17: 411–510.

Mollon J (1982) Color vision. *Annual Review of Psychology* 33: 41–85.

Naarendorp F, Sato Y, Cajdric A, et al. (2001) Absolute and relative sensitivity of the scotopic system of the rat: Electroretinography and behavior. *Visual Neuroscience* 18: 641–656.

Reeves A and Wu S (1997) Transient lumanopia: The invisibility of flicker in early dark adaptation. *Journal of the Optical Society of America A* 14: 2509–2516.

Rodieck RW (1998) *The First Steps in Seeing.* Sunderland, MA: Sinauer Associates.

Schnapf JL and Schneewies DM (2004) Electrophysiology of cone photoreceptors in the primate retina. In: Gegenfurtner KR and Scharpe LT (eds.) *Color Vision: From Genes to Perception.* Cambridge: Cambridge University Press.

Valeton JM and Van Norren D (1979) Transient tritanopia at the levels of ERG b-wave. *Vision Research* 19: 689–693.

Werner A, Smith V, Pokorny J, et al. (2005) Psychophysical correlates of identified physiological processes in the human visual system. In: Kremer J (ed.) *The Primate Visual System*, pp. 311–358. New York: Wiley & Sons.

Wyszecki G and Stiles WS (1982) *Color Science, Concepts and Methods, Quantitative Data and Formulae,* 2nd edn., New York: Wiley.

Vision: Mechanisms of Orientation, Direction and Depth

N J Priebe and D Ferster, Northwestern University, Evanston, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Primary Visual Cortex

Primary visual cortex is the site of a remarkable transformation in the neural representation of the visual world. The afferent neurons of the retina and the primary visual thalamus, which relays the retinal signals to the primary visual cortex, have circularly symmetric receptive fields and respond to almost any stimulus, as long as the stimulus falls within a broad range of sizes. In sharp contrast, the neurons of primary visual cortex are exquisitely sensitive to several complex stimulus attributes, including orientation of a contour, direction of motion, size, and binocular disparity. In many cases, selectivity is absolute: cortical neurons will not respond at all unless the stimulus matches the neuron's preferences for each of these attributes.

Primary visual cortex has become a model system for neuronal processing by the circuitry of the cerebral cortex. Ultimately, we would like to understand how the higher cortical areas give rise to our recognition of sensory stimuli, decision making, and language comprehension. But, we barely have a way of defining what tasks are performed in these areas, let alone how they are performed. By comparison, the transformations performed by the primary visual cortex on its thalamic inputs – the extraction of orientation and direction information, for example – are complex enough to be interesting, but simple enough to be tractable. The neuronal mechanisms by which cortical receptive properties emerge have therefore become the subject of intense scrutiny and debate.

Orientation Tuning

The basic phenomenon of orientation selectivity is very simple and striking. As Hubel and Wiesel originally demonstrated in their Nobel Prize-winning work of 1962, when a bar or contrast edge is flashed in or moved across a cortical neuron's receptive field at the neuron's preferred orientation, the neuron responds vigorously with a barrage of action potentials. When the stimulus is presented within the same region of visual space but at the orthogonal orientation, the neuron, in most cases, does not respond at all. In cat visual cortex, a vast majority of neurons are orientation selective. In primate visual cortex, all neurons outside layer 4c (the thalamic recipient zone) are orientation selective. Different neurons have different preferred orientations, so that some subset of cortical neurons will respond to a stimulus of any orientation and position. The selectivity of each neuron is roughly Gaussian in shape. That is, a graph of response against stimulus orientation forms an approximately Gaussian-shaped curve, with a half-width at half-height in the range of 10–30°. For cells that give a nonzero response at the nonpreferred orientation, the Gaussian rides on top of an offset the size of the nonpreferred response.

Hubel and Wiesel also showed that orientation-selective cells can be divided into two distinct groups, labeled simple and complex. The receptive fields of simple cells are identified by the presence of two or three (or occasionally more) distinct subregions (or subfields) of their receptive fields. The subregions are defined on the basis of their response to flashing stimuli. ON subregions are defined by the responses evoked at the onset of a stimulus brighter than the background light, or the offset of a stimulus darker than background. OFF subregions are defined by the responses evoked by the onset of dark stimulus, or the offset of a bright stimulus. The subfields are elongated in the direction parallel to the preferred orientation of the cell and are positioned side by side, in the direction perpendicular to the preferred orientation (**Figure 1(a)**). They are also mutually antagonistic: a bright stimulus in an OFF region will diminish the response to a simultaneously presented bright stimulus in an ON region, and vice versa. Complex cells, by definition, lack these distinct subfields and instead respond to flashed stimuli more or less uniformly throughout the receptive field, either with ON responses, OFF responses, or both (**Figure 1(d)**). Since Hubel and Wiesel's first description of simple and complex cells, it has been shown quantitatively that they form two distinct populations, rather than the tails of a single population, when defined through the spike responses. When measured from their synaptic inputs, however, cortical cells form a continuum in which simple cells with highly segregated receptive field subregions lie at one end, and complex cells lie at the other. The transformation between synaptic input and spike output by spike threshold therefore helps to segregate simple and complex cells into distinct groups. In the cat, simple cells predominate in layers 4 and 6, the layers of the cortex that receive direct input from the visual thalamus (lateral geniculate nucleus, or LGN). Complex cells predominate in layers 2, 3, and 5, which predominantly receive input from other cortical cells instead of the LGN.

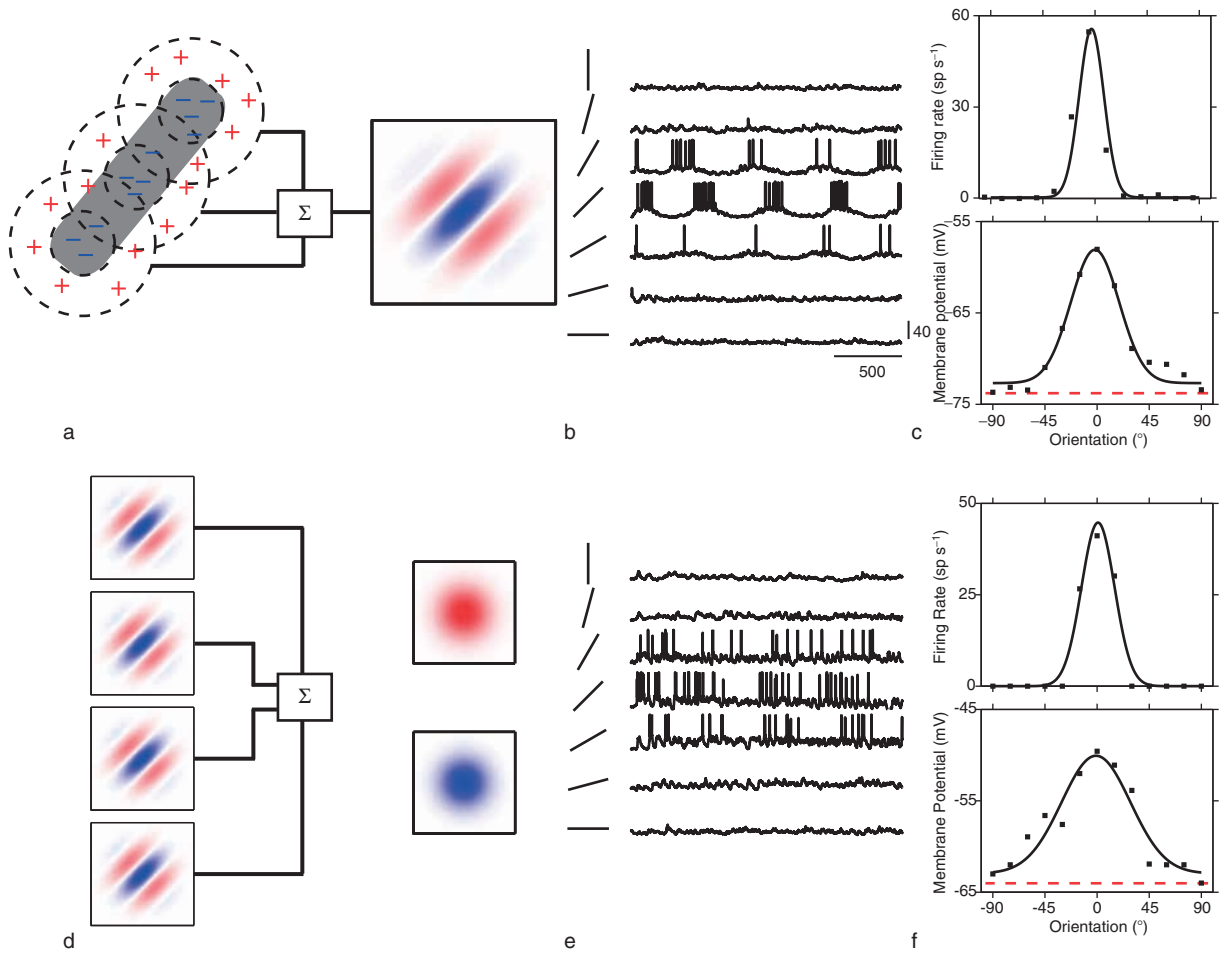

Figure 1 The emergence of orientation selectivity in primary visual cortex. The feed-forward model proposed by Hubel and Wiesel contains two stages, one that creates simple cell responses (top row) and a second that creates complex cell responses (bottom row). (a) In the first stage of the model, thalamic relay cells with circularly symmetric receptive fields converge on a simple cell. The spatial offsets of the relay cell receptive fields create the elongations observed in simple cell receptive fields. Blue indicates locations where a dark stimulus elicits a depolarization; red indicates locations where a light stimulus elicits a depolarization. (b) Elongated receptive fields then give simple cells the orientation selectivity observed in both the membrane potential and the spiking responses to drifting gratings, which elicit a modulated responses. (c) Orientation tuning is narrower for firing rate (top) than for membrane potential (bottom). Values plotted represent the peak (first Fourier component + mean) components of response. (d) In the second stage of the feed-forward model, multiple simple cells with similar orientation tuning but different spatial selectivity converge to create complex cell receptive fields. Complex cells respond to either a dark or a light stimulus placed in the receptive field, as long as the stimulus has the correct orientation. (e) Complex cells respond with a sustained depolarization to a drifting grating, but maintain orientation tuning. (f) Tuning curves for firing rate narrower than those for membrane potential.

Functional Architecture

At the same time that they first demonstrated orientation selectivity, Hubel and Wiesel also laid out what they termed the functional architecture of the cortex, or the way in which neurons with similar receptive field properties are physically arranged within the cortex. Specifically, they showed that all of the neurons in a cortical column – the cells arrayed vertically beneath one point on the cortical surface – have nearly identical preferred orientations. The columns preferring different orientations, in turn, arranged in an orderly fashion. Specifically, Hubel and Wiesel found that an electrode that is advanced in a direction perpendicular to the cortical surface encounters cells with the same preferred orientation. Conversely, as an electrode is advanced in a direction roughly tangential to the cortical surface, each cell encountered has a slightly different orientation from the previous cell. A large sequence of cells shows a smooth progression of orientation, with occasional jumps in orientation or reversals in the direction of the sequence (clockwise to counterclockwise, or vice versa). On average, a full range of orientations is covered by the neurons within a distance of about 1 mm of the cortical surface, which Hubel and Wiesel referred to as an orientation hypercolumn.

Subsequent to these observations from physiological experiments, the exact pattern or map of orientations within a hypercolumn was the subject of considerable theoretical and experimental work. The complete pattern of orientation columns was first revealed in detail by the advent of techniques for imaging neuronal activity *in vivo*. The images are based on almost imperceptible changes in the color of the cortical tissue that occur when active neurons take up oxygen, changing both the color of hemoglobin and the amount of blood flow in the local vasculature. The activity-evoked changes in reflectance of the cortical surface are less than 1/1000th of the total reflectance, but can be detected with an appropriate camera. Alternatively, the cortex can be stained with dyes that dissolve into the neuronal membranes and change their absorption as a function of the local electric field (voltage-sensitive dyes). Here again the optical signals are extremely small, but are much faster than the intrinsic, hemoglobin-based signals. Both types of imaging yield a map of orientations on the cortical surface, showing in both monkey and cat the formation of pinwheel structures, such that neurons lying within the vanes of the pinwheel each have a slightly different orientation preference (Figure 2). In a circular track around a single pinwheel, neurons of all preferred orientations would be found, arranged in order of orientation. Pinwheel centers are spaced at roughly 1 mm intervals across the cortical surface, and the vanes of adjacent pinwheels blend into one another. Note that at the centers of the pinwheels the vanes narrow down to essentially zero width, so that cells with very different orientation would lie very close to one another. The optical signals show relatively weak orientation tuning at the pinwheel centers, but because the techniques have spatial resolution somewhat larger than a single cortical cell, there are two possible explanations: the cells at the pinwheel centers could all have weak orientation preference, or the cells at the center could all be well-tuned for orientation but have different orientation preferences. Recent experiments with two-photon Ca^{2+} imaging *in vivo* have achieved single-cell resolution of neuronal activity, and indicate that cells within the pinwheel centers and cells outside the pinwheel are equally selective for orientation. The vanes of the pinwheels narrow, but are sharply defined down to the level of single cells, all the way into the centers.

Cortical Circuitry Underlying Orientation Selectivity

Nearly all discussions about the cortical circuitry underlying orientation selectivity begin with the simple and elegant model that Hubel and Wiesel proposed when they first described orientation selectivity. They noted that the ON and OFF subfields of simple cells are similar to the ON and OFF structures of LGN relay cell centers, which provide the main visual input to the cortex. Simple cell subfields, unlike the centers of relay cell receptive field, are elongated, however, which prompted Hubel and Wiesel to suggest that each simple cell's ON region is constructed from the input from a number of ON-center relay cells that have receptive fields aligned in a row. Similarly, the OFF subfield would be built from input from a set of OFF-center relay cells that have receptive fields overlapping the subregion. According to the model, orientation selectivity in the simple cell arises directly

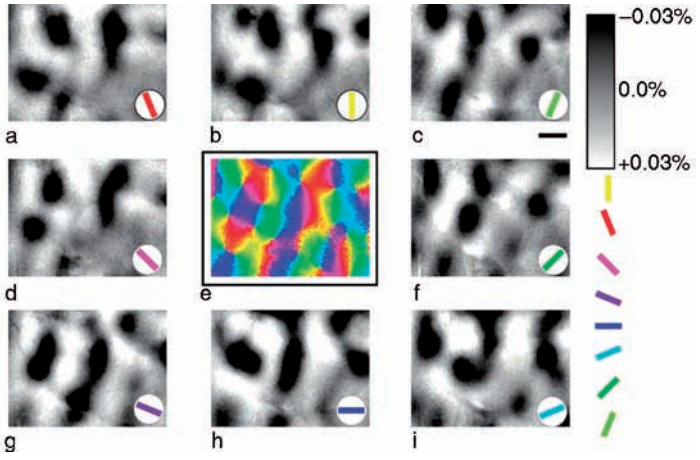

Figure 2 Orientation maps on the surface of primary visual cortex measured with optical imaging of intrinsic signals. Gray-scale images (a–d, f–i) show responses (through both eyes) to stimulation at the orientation indicated by the colored bar inset. Dark regions are areas of strong neuronal response. Angle map (e) shows preferred stimulus orientation (scale at right). Scale bar = 500 μm; medial, up; anterior, to the right. Reproduced from Crair MC, Ruthazer ES, Gillespie DC, and Stryker MP (1997) Ocular dominance peaks at pinwheel center singularities of the orientation map in cat visual cortex. *Journal of Neurophysiology* 77: 3381–3385, used with permission.

from the spatial arrangement of the receptive fields of the presynaptic relay cells (**Figure 1(a)**). When a stimulus is parallel to the subfields, it simultaneously encounters the receptive fields of all of the constituent relay cells. The relay cells therefore synchronously excite the simple cell, bringing it to threshold. Conversely, when a stimulus is perpendicular to the subfield, it can only excite a fraction of the presynaptic relay cells at one time. The stimulus therefore generates some asynchronous excitation in the simple cell, but the excitation is too weak to reach threshold. The spatial organization of the input therefore generates the initial orientation bias in the synaptic input to the simple cell, and spike threshold amplifies the bias into an absolute selectivity in the cell's spike output.

Hubel and Wiesel's model of the relay cell input to simple cells is most likely correct. Recordings from connected pairs of LGN relay cells and simple cells demonstrate that a relay cell makes excitatory contact with a simple cell only if the relay cell receptive field center overlaps the subfeild of the target simple cell with the same response polarity (ON or OFF). In addition, intracellular recordings from simple cells, even with the cortical circuit inactivated so that the only remaining inputs are from relay cells, show strong orientation selectivity. Few would dispute, then, that the spatial organization of the relay cell input generates an initial orientation bias in the input to cortical cells.

There has, however, been dispute over whether the spatial organization of the relay cell input is sufficient to explain all of the simple cell response properties. For example, in several experiments it has been observed that the spatial organization of the receptive field overestimates the width of orientation tuning. From Hubel and Wiesel's model, one would expect that the longer and narrower a cell's receptive field subregions are, the more sensitive the cell will be to orientation. With a long, narrow subfield, a small shift in orientation will create a large mismatch between the stimulus and the receptive field, and therefore strongly reduce the evoked response. In most cells, however, tuning is even narrower than expected.

One proposal for the origin of this narrowing is lateral inhibition in the orientation domain (also called cross-orientation inhibition), in which cells with different preferred orientations inhibit one another. Alternatively, cells with similar preferred orientations and different tuning widths could inhibit one another. Cross-orientation inhibition could also explain two other strongly nonlinear properties of cortical cells: cross-orientation suppression, in which the responses to stimuli of the preferred orientation are antagonized by superimposed stimuli of the orthogonal orientation, and contrast-invariance of orientation tuning, in which the width of orientation tuning does not change with the strength (contrast) of the stimulus. Even more suggestive is the observation that antagonists of γ-aminobutyric acid ($GABA_A$)-mediated inhibition degrades orientation tuning when applied to the cortex.

Despite these strong indications that lateral inhibition among cells with different orientation preference is critical for generating orientation selectivity, direct evidence of lateral inhibition has been hard to find. Many (though not all) intracellular recordings from simple cells show only modest inhibition evoked by stimuli of nonpreferred orientation. Recent work suggests, as an alternative to lateral inhibition, that nonlinear properties of the feed-forward pathway from the LGN to the cortex can account for most of the aforementioned properties of simple cells. Spike threshold, by way of the so-called iceberg effect, for example, can explain why simple cells have such narrow orientation tuning relative to what is predicted by their receptive field maps. In this view, the input from geniculate relay cells determines the aspect ratio of the subfields, and therefore the width of orientation tuning for the synaptic input to the simple cell. Because of the iceberg effect, these inputs are suprathreshold only for a narrow range of orientations near the preferred orientation. Thus, the orientation tuning for the spike output of the cell is much narrower than that predicted by the spatial organization of the receptive field. Similar arguments, based on nonlinearities such as contrast saturation and spike rectification in relay cells, or response variability in cortical cells, can, in turn, can account for cross-orientation suppression and contrast invariance of orientation tuning.

Direction Selectivity

In addition to being selective for orientation, many cortical cells are also selective for the direction of stimulus motion. For a large fraction of both simple and complex cells in primary visual cortex, response strength depends on stimulus direction, with direction being perpendicular to the preferred orientation. Some cells are completely directional and do not respond at all to stimuli of the nonpreferred direction. As with orientation selectivity, direction selectivity is mapped onto the surface of the cortex: within a single orientation column, neurons with different preferred directions are segregated into direction-specific subcolumns. Recent Ca^{2+} imaging of neuronal activity using two-photon laser scanning microscopy has shown that the borders between direction-specific regions of the cortex are extremely precise, with neurons on either side of a sharp border having opposite

preferred directions. In primates, direction-selective neurons are also segregated by layers, tending to be located in the lower portion of layer 4 and in layer 6, which then project preferentially to extrastriate cortical areas in the dorsal processing stream.

Cortical Circuitry Underlying Direction Selectivity

As with orientation selectivity, direction selectivity is not represented in the relay cell input to primary visual cortex, and therefore must be computed within the cortex. Because motion in its essence requires a comparison over time and space (Where is the stimulus now? Where was it before?), computing direction requires that the responses of different inputs differ in both time and space. The origin of direction selectivity is still not fully understood, but in simple cells it clearly depends the heterogeneity of response timing within receptive field subregions. Specifically, it has been observed that in direction-selective cells, the latency of the response to a flashed stimulus differs by several tens of milliseconds between one side of the subfield and the other (along the axis perpendicular to the preferred orientation) (Figure 3(a)). The cell invariably prefers motion in the direction starting-from the long-latency regions of the receptive field and moving toward the short-latency region (Figure 3 (b)). When the stimulus moves in the nonpreferred direction, it encounters the short-latency part of the receptive field first, and the long-latency part later. The inputs from the two parts of the receptive field arrive at the cell at two different times, separated by the sum of the latency difference between the two regions and the time it takes the stimulus to travel from one region to the other. Conversely, when the stimulus moves in the preferred direction, it encounters the long-latency part of the receptive field first and the short-latency part later. If the stimulus travel time is approximately equal to the latency difference, then the inputs from the two parts of the receptive field will sum. The preferred stimulus therefore evokes a larger, though shorter-duration, input, compared to the nonpreferred stimulus.

This mechanism is strongly akin to orientation tuning, but with the orientation in the space–time domain rather than in the purely spatial domain. That is, a two-dimensional (2-D) plot of response latency against stimulus position is tilted for direction-selective cells (those without a gradient of latency across the receptive field), and is vertical for non-direction-selective cells (those with a gradient). Only stimuli with the correct spatiotemporal tilt (which move in the preferred direction) will evoke responses.

Although the preferred direction of motion is well predicted by the tilt of the spatiotemporal receptive field, the degree of selectivity in simple cells is often much greater than that predicted by the spatiotemporal receptive field map, much the same way that the spatial receptive field map underestimates the degree of orientation tuning. And similar to orientation tuning, one proposed explanation is lateral inhibition, in this case in the direction domain, such that direction-selective cells of opposite preferred direction inhibit one another. Once again, however, intracellular recordings from simple cells show little evidence for lateral inhibition. Excitation and inhibition in most cells share the same preferred direction. The apparent mismatch between the predictions of the spatiotemporal maps and recorded direction selectivity can be explained by the nonlinearity inherent in spike threshold. Because spike threshold disproportionately suppresses smaller synaptic inputs relative to larger ones, the direction selectivity of the spike output of cells is much larger than the direction selectivity of their synaptic inputs (Figure 3(c)).

The origin of the difference in response latency between different synaptic inputs is not known. Different relay cells could have different visual latencies (e.g., lagged and nonlagged relay cells), or long- and short-latency inputs could originate in different regions (such as the LGN and cortex).

The Aperture Problem

The processing of motion information at the level of primary visual cortex is incomplete. Neurons in primary visual cortex respond to the retinal image through a small aperture, defined by their receptive fields. They are only sensitive to the component of motion perpendicular to their preferred orientation within that restricted area. A cortical cell, then, cannot detect the difference between an optimally oriented stimulus moving in the preferred direction and an indentical stimulus moving at a different angle but higher speed (Figure 4). As long as the stimuli are much longer than the receptive field, such that their ends do not cross the receptive field, the stimuli appear identical to the cell and evoke the same response. Disambiguating such stimuli can only occur by integrating information from across the visual field and assembling the small components of the image into larger patterns. For motion, the extrastriate cortical medial temporal (MT) area is thought to be critical to this process.

Disparity Tuning

Information from the right and left eyes converges onto single cells for the first time in primary visual cortex. Throughout the cortex the balance of input from the two eyes (ocular dominance) covers the full

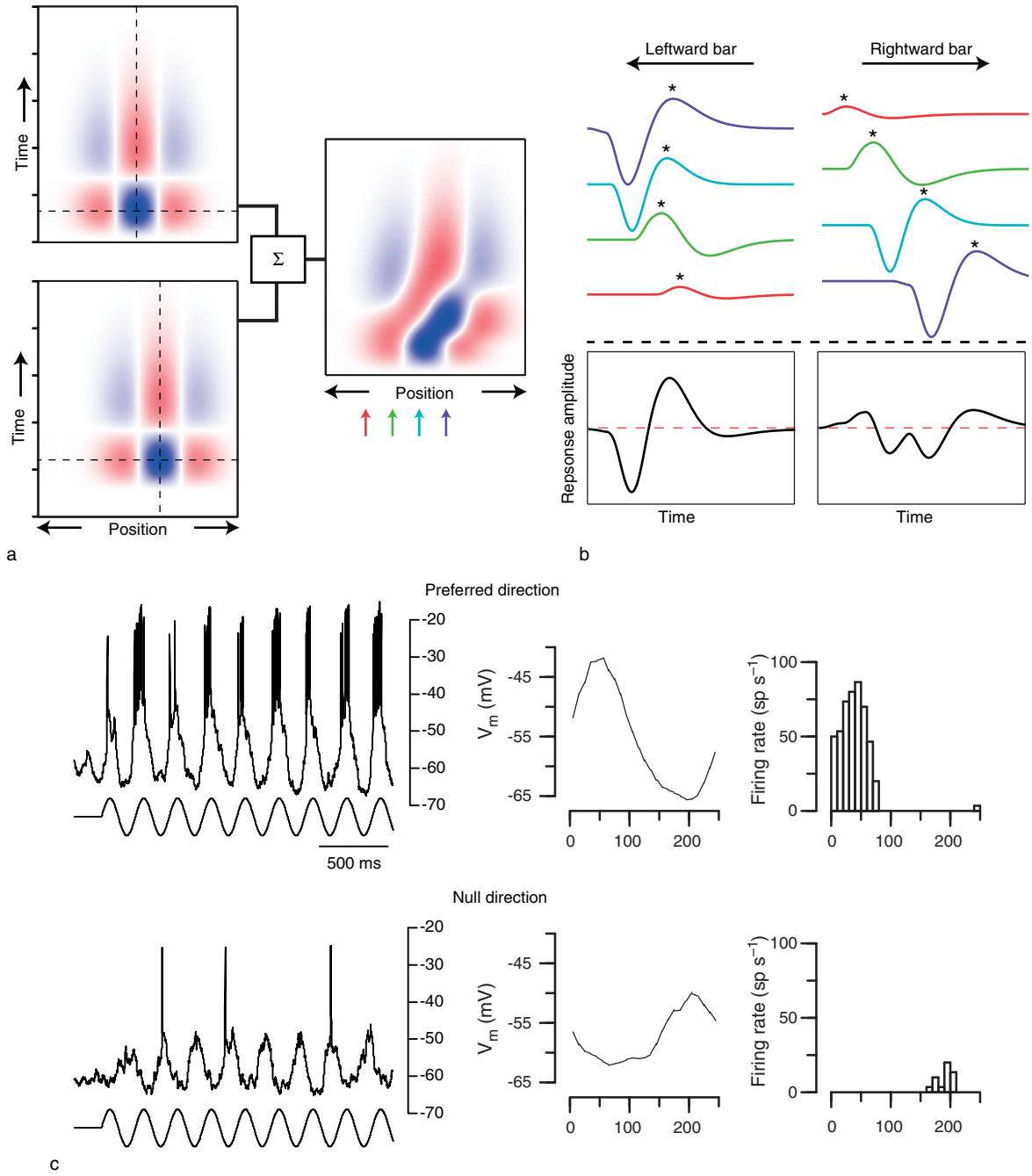

Figure 3 The emergence of direction selectivity in primary visual cortex. (a) Direction selectivity in primary visual cortex emerges due to the convergence of nondirectional responses that differ in both spatial preference and response timing (compare the latency and spatial preference of the two input neurons). The resulting spatiotemporal receptive field is tilted in space and time, a hallmark of direction-selective simple cells. Blue indicates an increase in response to a dark stimulus; red indicates an increase in response to a light stimulus. (b) Tilt in the spatiotemporal receptive field creates direction-selective responses to bars drifting across the receptive field. As a moving bar is sequentially flashed across the receptive field, a response is evoked at location (colored traces, locations indicated by the arrows in (a). For a leftward bar, moving from the purple arrow toward the red arrow, the response peaks at each location synchronously (indicated by the asterisk for each trace). The synchronous responses sum to create a large response to the leftward bar (bottom trace). The peak responses to a rightward bar are asynchronous and the summed response is therefore small. (c) Direction selectivity observed in cortical simple cells is evident both in the membrane potential and in spiking responses. The direction selectivity measured from membrane potential is modest relative to the selectivity observed in firing rate: direction index $= (R_p - R_n)/(R_p + R_n)$, where R_p and R_n are the responses in the preferred and nonpreferred (null) directions; for this example neuron, membrane potential $= 0.36$; firing rate $= 0.80$.

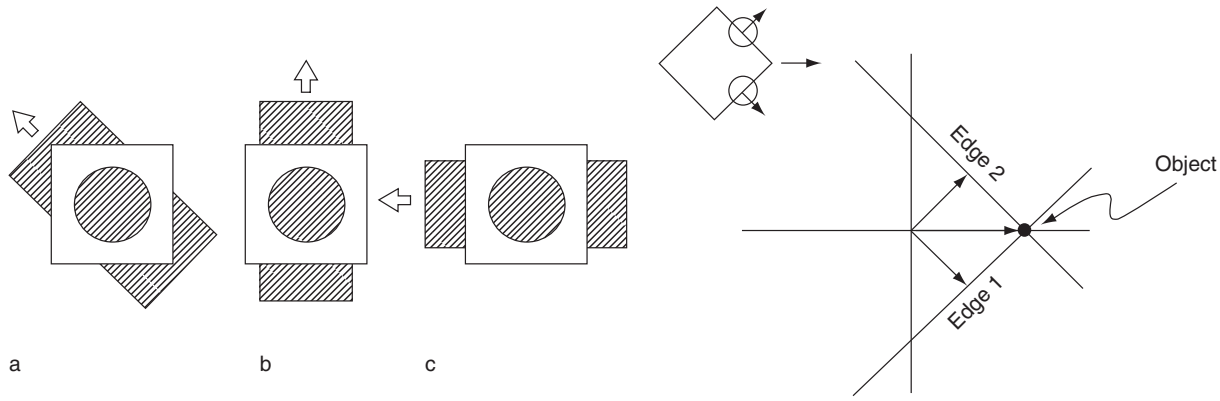

Figure 4 The aperture problem. (a–c) The direction of a drifting grating, moving behind the aperture, is ambiguous. All three conditions have the same apparent motion within the aperture. When information from multiple cells with different orientation preferences is available, the correct direction and speed of a moving diamond can be computed (right panel). By converting the motion components into velocity space (where direction is indicated by angle, and speed is indicated by vector length), each edge gives rise to a family of possible motions (indicated by the line orthogonal to each edge velocity vector). The intersection of these two lines indicates the correct solution to the object's velocity. Reproduced from Adelson EH and Movshon JA (1983) The perception of coherent motion in two-dimensional patterns. *Proceedings of the ACM SIGGRAPH/SIGART Interdisciplinary Workshop on Motion Representation and Perception*, pp. 11–16. © 1983 ACM, Inc. Reprinted by Permission.

range, with some cells receiving excitatory input only from one eye, some receiving predominant but not exclusive input from one eye, and others receiving equal input from the two eyes. Like orientation, ocular dominance is organized into columns, in this case alternating bands of about 0.5 mm thickness, with completely monocular cells near the center of the bands and more binocular cells near the edges.

In most cases, the receptive field of a binocular cortical cell mapped through one eye is identical in all respects to the field mapped through the other eye. Orientation and direction preference are the same, as are receptive field type (simple or complex), size, and position, even if the strength of the responses is different for the two eyes. By position, here, we refer to the relative positions of the receptive fields in the two retinas. That is, the receptive fields are located at the same direction and distance relative to the point of central vision (fovea in primates, area centralis in cats). Thus, the receptive fields are located at correspondent points in the retina, such that images falling on those points would appear to originate from the same point in visual space.

Of course, the images generated by many objects do not fall on correspondent points of the retina, specifically those objects that lie in front of or behind the plane of fixation. As one can easily demonstrate for oneself, for an object in front of the fixation plane the image in the left eye appears to lie to the right of the image in the right eye (and vice versa for objects beyond the fixation plane) (Figure 5). This depth-dependent lateral displacement of the two retinal images of an object is called binocular disparity and forms the basis for binocular depth vision.

Many neurons in the primary visual cortex are highly sensitive to binocular disparity. Tuned excitatory cells respond to stimulation in either eye, and tend to respond most strongly to simultaneous stimulation of correspondent points on the retinas – that is retinal disparities at or near 0°. Tuned inhibitory cells respond most weakly at 0 disparity and respond more strongly at disparities on either side of 0. Near cells respond only to crossed disparities, which are generated by objects closer than the fixation plane; far cells respond to uncrossed disparities.

Although their function is still under debate, together, these different types of cells could encode enough information to support depth perception. It has been noted that since the eyes are separated laterally, only horizontal disparities are generated by stimuli out of the fixation plane. If disparity-sensitive cells in the visual cortex feed into the circuitry for depth perception, the reasoning goes, they would be preferentially sensitive to horizontal disparities. Data on this point are equivocal, however.

Much like the aperture problem for direction selectivity, accurately reconstructing the three-dimensional visual world from the signals provided by disparity-sensitive cells presents a formidable computational problem. At the heart of the problem lies the existence of a large number of what are called false stereo matches in the two retinal images, also referred to as the stereo correspondence problem. When looking at a repeating pattern, such as a picket fence, for example, the brain must decide which slats of the fence in the right eye's image match with the slats in the left eye. If the slats lie in the plane of fixation and are spaced at 0.25° intervals, then

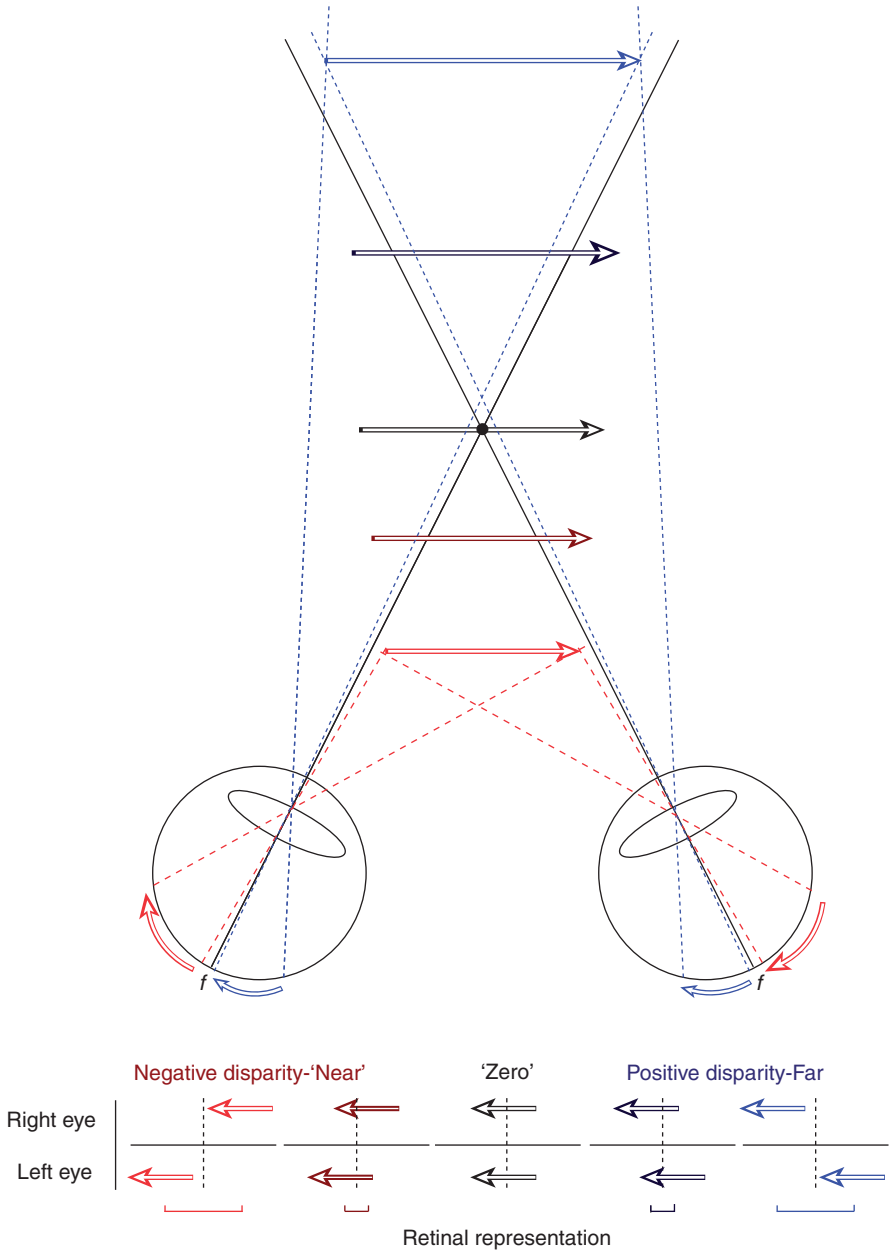

Figure 5 Retinal disparity. The retinal representations of objects at different depths differ in relative position between the right and left eyes. The red and blue arrows are placed at near and far depths, respectively, and their representations occupy different portions of the retinal representations. The retinal images for the red arrow fall on the temporal portion of the retinas, while the blue arrow images lie on the nasal retinas.

disparity-tuned cells that prefer 0°, ±0.25°, ±0.5°, and so forth will all be active. The ±0.25° cells, however, will be falsely matching a slat in one eye, and the adjacent one in the other eye. Only the 0° cells are responding correctly in the sense that a real object exists in the fixation plane. There are number of models as to how the brain suppresses false matches and arrives at a reasonably accurate perception of the three-dimensional stimulus. If the disparity-sensitive cells of primary visual cortex contribute to this process, however, their raw responses to disparity matches are likely further processed in higher cortical areas, such as visual areas V2, MT, and beyond.

Conclusion

The discovery of the response selectivity of cortical neurons by David Hubel and Torsten Wiesel in the early 1960s has raised a variety of questions, all of which have received considerable attention. How

does the cortical circuit extract information about orientation, direction, and size from the nonselective signals coming from the retina and thalamus? Why is the visual image parsed into these particular parameters? What advantages do orientation and direction selectivity confer on the animal, and how do they facilitate object detection and recognition at higher cortical levels? How does the circuitry of the cortex develop, and does development require visual experience?

See also: Information Coding; Neural Synchrony and Feature Binding; Perception of Surfaces and Forms; Representation of Movement; Shape Representation in Inferotemporal Cortex; Stereoscopic Vision; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging; Visual System: Multiple Visual Areas in Monkeys; Visual Attention.

Further Reading

Adelson EH and Movshon JA (1983) The perception of coherent motion in two-dimensional patterns. *Proceedings of the ACM SIGGRAPH/SIGART Interdisciplinary Workshop on Motion Representation and Perception*, pp. 11–16.

Anderson JS, Carandini M, and Ferster D (2000) Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *Journal of Neurophysiology* 84: 909–926.

Anzai A, Ohzawa I, and Freeman RD (1999) Neural mechanisms for encoding binocular disparity: Receptive field position versus phase. *Journal of Neurophysiology* 82: 874–890.

Crair MC, Ruthazer ES, Gillespie DC, and Stryker MP (1997) Ocular dominance peaks at pinwheel center singularities of the orientation map in cat visual cortex. *Journal of Neuroscience* 77: 3381–3385.

Cumming BG and DeAngelis GC (2001) The physiology of stereopsis. *Annual Review of Neuroscience* 24: 203–238.

Finn IM, Priebe NJ, and Ferster D (2007) The emergence of contrast-invariant orientation tuning in simple cells of the cat visual cortex. *Neuron* 54: 137–152.

Gardner JL, Anzai A, Ohzawa I, et al. (1999) Linear and nonlinear contributions to orientation tuning of simple cells in the cat's striate cortex. *Visual Neuroscience* 16: 1115–1121.

Hubel DH and Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London)* 160: 106–154.

Hubel DH and Wiesel TN (1977) Functional architecture of macaque visual cortex. *Proceedings of the Royal Society of London, Series B* 198: 1–59.

Hubener M, Shoham D, Grinvald A, et al. (1997) Spatial relationships among three columnar systems in cat area 17. *Journal of Neuroscience* 17: 9270–9284.

Marr D (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W.H. Freeman and Company.

Martinez LM, Wang Q, Reid RC, et al. (2005) Receptive field structure varies with layer in the primary visual cortex. *Nature Neuroscience* 8: 372–379.

Monier C, Chavane F, Baudot P, et al. (2003) Orientation and direction selectivity of synaptic inputs in visual cortical neurons: A diversity of combinations produces spike tuning. *Neuron* 37: 663–680.

Movshon JA, Adelson EH, Gizzi MS, et al. (1986) The analysis of moving visual patterns. *Experimental Brain Research* 11: 117–151.

Movshon JA and Newsome WT (1996) Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *Journal of Neuroscience* 16: 7733–7741.

Ohki K, Chung S, Kara P, et al. (2006) Highly ordered arrangement of single neurons in orientation pinwheels. *Nature* 442: 925–928.

Poggio GF and Fischer B (1977) Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *Journal of Neurophysiology* 40: 1392–1405.

Priebe NJ and Ferster D (2005) Direction selectivity of excitation and inhibition in simple cells of the cat primary visual cortex. *Neuron* 45: 133–145.

Priebe NJ and Ferster D (2006) Mechanisms underlying cross-orientation suppression in cat visual cortex. *Nature Neuroscience* 9: 552–561.

Reid RC, Soodak RE, et al. (1991) Directional selectivity and spatiotemporal structure of receptive fields of simple cells in cat striate cortex. *Journal of Neurophysiology* 66: 505–529.

Sillito AM (1975) The contribution of inhibitory mechanisms to the receptive field properties of neurones in the striate cortex of the cat. *Journal of Physiology (London)* 250: 305–329.

Somers DC, Nelson SB, and Shapley RM (1995) An emergent model of orientation selectivity in cat visual cortical simple cells. *Journal of Neuroscience* 15(8): 5448–5465.

Vision: Surface Segmentation

H S Scholte and V A F Lamme, University of Amsterdam, Amsterdam, The Netherlands

© 2009 Published by Elsevier Ltd.

Introduction

In midlevel vision, attributes such as orientation, motion, depth, and luminance are combined to form boundaries and surfaces. Boundary detection consists of detecting local continuities and discontinuities in the visual image on the basis of these attributes. Surface segmentation consists of grouping parts of the visual scene that share these attributes. Together, these processes result in a surface representation of the visual image in which some surfaces of the visual scene are assigned as figure and other surfaces are assigned as ground, and as such contribute to the perceptual interpretation of a scene. This figure–ground relation is characterized by three aspects: (1) the figure appears to be located in front of the ground, (2) the local discontinuity between figure and ground is perceived as belonging to the figure (the figure has ownership of the border), and (3) the ground appears to continue behind the figure.

Behavioral experiments suggest that boundary detection precedes surface segmentation. The perception of brightness of a white patch, presented on a dark background, seems to evolve from the edge inward. A similar observation has been made for the process of texture segregation. Moreover, it seems that boundaries are more important in texture segregation than is similarity of the features within the segregated surface, suggesting that the boundaries determine where the surfaces are located.

Boundary Detection

Neurons in the earliest cortical visual area, V1, respond preferentially to a particular orientation of luminance contrast and respond to stimuli in a restricted part of visual space that extends from 1 to 2°. This area is called the classical receptive field (cRF; see Figure 1(a)), area enclosed by the inner dotted line) and its size is the result of input from neurons in lower tier areas (i.e., it is purely determined by means of feed-forward connections). The RF is 'tuned' to particular features – that is, it only responds to stimuli that meet particular requirements. There is also an area beyond the cRF in which stimuli do not elicit a response on their own but can modulate a cRF-driven response. This is called the surround receptive field (sRF; see Figure 1(b), area inside the outer dotted line).

The total size of the receptive field in V1 (cRF + sRF) is at least two to five times larger than the cRF area. The most typical type of sRF effect is response suppression. Approximately 72% of cells in monkey V1 show this behavior when a central grating is presented that is successively increased in diameter. A similar pattern is observed when a line element is presented in the cRF and randomly oriented line elements are presented in the sRF (see Figure 1(b)). This inhibition is stronger when adjacent receptive fields, with a similar feature preference, receive the same type of input.

This situation changes drastically when the presented visual information contains local structure. Compared with the random surround, the firing rate of the neuron will be higher if the line element is surrounded by line elements with an orientation orthogonal to that of the center stimulus, or when the saliency of the line segment is enhanced by adding similarly oriented line segments along its principal axis (see Figure 1(c)). The behavior of neurons in these latter situations correlates with the anatomical structure and physiological properties of the horizontal connections – that is, long-range projections that originate from pyramidal cells, between cortical columns in early visual cortex. These horizontal connections preferentially link orientation columns with a similar orientation tuning, and these connections between similarly oriented cells are even stronger when these cells have RFs along a collinear axis. Behaviorally, these increases in firing rate coincide with an increase in the sensitivity of humans and monkeys to detect these stimuli. It is important to note that these effects are modulatory and only occur when the cRF is stimulated. When a stimulus is presented into the cRF of a neuron in one eye, and a context is presented into the sRF of the other eye, the sRF effects do not disappear (dichoptic presentation). sRF properties in V1 are therefore not determined by means of pure feed-forward processing but rather are cortical in origin (be it via horizontal connections within the cortical area or feedback processing from higher tier areas).

In summary, we see that neurons with their receptive field on a boundary of areas with dissimilar features have the strongest responses because they are less inhibited by their neighbors. This results in a response enhancement in areas where oriented texture boundaries are present, which appears between 40 and 80 ms after stimulus presentation, depending

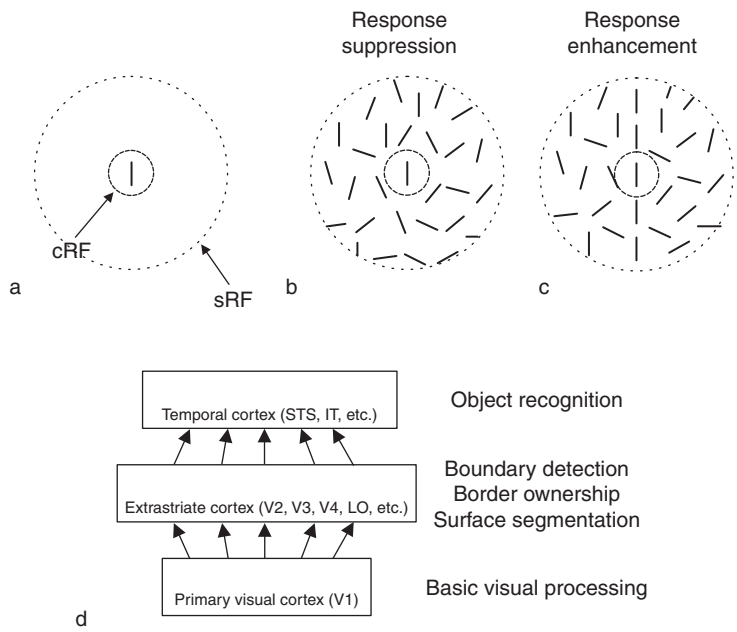

Figure 1 (a) The classical receptive field (cRF) and surround receptive field (sRF) of a neuron in area V1 which is optimally stimulated by a line element. (b) Example of a scene in which random line elements are presented in the sRF of the neuron. This results in response suppression. (c) The presence of a collinear line in the sRF of the receptive field results in a relative response enhancement. (d) The feed-forward model of surface segregation. Neurons in V1 are involved in the detection of line elements. Extrastriate areas receive this V1 input, detect boundaries and border ownership, and perform surface segmentation. Areas in temporal cortex use this information for object recognition.

on how the boundary is defined. These signals can be relayed from area V1 to extrastriate areas with larger receptive fields. These higher tier areas have a coarser representation of visual space, and computational models have shown that boundary information can be integrated over space in this fashion. Debate remains whether contextual modulation signals in V1 related to boundary detection and contour integration are influenced by horizontal connections or feedback processing, and in different situations (depending on factors such as distance between elements in the visual scene) the mixture of these influences will differ. Functional magnetic resonance imaging studies further show that areas V2, V3, V3a, and V4 are also involved in boundary detection.

The Feed-Forward Model of Surface Segmentation

The most simple and elegant neural model of boundary detection and surface segmentation perceives these processes as being implemented by a series of temporal hierarchical processing steps that are implemented by spatially hierarchical organized brain areas, going from V1 via extrastriate areas to temporal cortex (see Figure 1(d)) via pure feed-forward connections. The region of space over which information has to be analyzed to achieve surface segmentation is typically much larger than the cRF of a neuron in V1 or V2. Areas in higher tier regions do have cRF sizes that are large enough to encapsulate the region of a particular surface, and the feed-forward model of surface segregation presumes that these cells are successful in integrating this information into surfaces.

Support for this model comes from experiments that show that there is an increase in the complexity of surface segmentation-related signals going from lower tier to higher tier areas. As discussed in the preceding section, areas V1, V2, V3, V3a, and V4 are involved in the process of boundary detection. A more complex level of representation is border ownership. Neurons that are sensitive for border ownership signal the presence of a border, but only if this border belongs to a figure on a specific side of the receptive field. More than 50% of the neurons in areas V2 and V4 and a smaller percentage of neurons in V1 signal border ownership. Border-ownership-related signals appear to be available in these areas at the same time (69–73 ms; luminance-defined figures). Neurons in V4 are tuned to elementary shapes, making them well suited for the integration of edgelike information from lower tier areas.

Higher in the hierarchy, the lateral occipital (LO) area responds to different shapes regardless of how

they are defined (cue invariant). This area also shows an adaptation response for stimuli that were identical in perceived shape but differed in contour organization, but not when contours were identical but perceived shape differed. This shows that area LO represents higher level shape information. Finally, neurons in areas such as the inferior temporal (IT) region, using input from lower tier areas, respond to specific objects, such as faces.

In summary, the feed-forward model of boundary detection and surface segmentation is both simple and elegant. It conforms to the main features of the visual system (increase in cRF size, increase in the specificity of the tuning of the cRF). Boundary detection precedes surface segmentation and occurs in early visual cortex. Surface segmentation occurs in extrastriate cortex, feeding its resulting output to higher tier areas where object perception takes place (see Figure 1(d)).

Surface Segmentation in Area V1

Recent findings have made it clear that the relationship between surface segmentation and boundary detection is more complicated than the feed-forward model suggests. Studies have shown that area V1 is sensitive for a large range of contextual influences that are spatially well beyond what could be expected on the basis of feed-forward processing (e.g., beyond the extent of the cRF).

Lamme found that a V1 neuron responds more strongly when its cRF is on the inside of a figure than when it is stimulated by identical background elements, even though its cRF is smaller than the figure (see Figure 2(a)). The black circle in the adjacent texture represents the cRF of this neuron. One result from this study is presented in Figure 2(b); the upper graph depicting the single-unit response from a neuron in V1 was obtained by presenting a texture with homogeneously oriented line elements. Overlaid on this is the response toward a texture where a figure can be perceived. After approximately 80 ms the response of this neuron is influenced by the context in which the line elements within the receptive field of the neuron are presented. Neurons with their receptive field on the boundary between figure and ground show an increase in firing rate at this interval. Subsequently, neurons within the boundaries show an increase in firing rate (compared to the condition in which only a homogeneous texture is presented) that starts after ~100 ms. This is shown in Figure 2(c), in which the activity evoked by a figure texture is plotted over time, for different positions of the cRF relative to the figure and the figure–ground boundary. This 'contextual' modulation, as measured in the center of a figure, is large for smaller figures and declines as figures grow larger. More importantly, there is an asymmetry, also visible in Figure 2(c), of modulation across the figure–ground boundary. As soon as the cRF is outside the figure, the sRF modulation is absent. This sharp decline in enhanced activity correlates with the observation that boundaries belong to the figure (and not the background). Other experiments have shown that this 'late' enhancement not only occurs for texture-defined figures but also for figures that are defined by color, motion, and depth, indicating that contextual modulation is cue invariant. Finally, border manipulations that alter the perceptual interpretation of a scene co-vary with the amount of contextual modulation.

The correlation between neuronal activity in V1 and the perceptual interpretation of the surface is further strengthened by a study that shows that masking the stimulus in such a way that a monkey is not capable of detecting the stimulus anymore coincides with an abolishment of the sRF effects. When a figure is masked by a stimulus that follows the presentation of a figure immediately (16 ms), the monkey is incapable of detecting this figure. With an increase in time between the presentation of the figure and the presentation of the mask, the capability of the monkey to detect the figure also increases and the size of the sRF effect that can be measured at these different masking intervals correlates with the reported visibility of the monkey. Finally, when the figure stimulus is followed by a blank screen mask, the figure is much easier to detect. The sRF effect again correlates with the visibility of the stimulus and this indicates that this effect is not a result of the presentation duration of the figure stimulus but rather of the perceived presence of the figure.

On the basis of these observations we can link contextual modulation signals measured in area V1, from 100 ms onward, to the process surface segmentation (at least for texture-defined figures), a view that is also supported by results from studies with humans. Event-related potential (ERP) studies have revealed that when the evoked response from occipital channels, obtained by presenting homogeneous textures, is subtracted from the response obtained by presenting figure textures (like the one in Figure 1(a)), a signal with peak latencies between 140 and 219 ms in occipital channels remains. The systematic manipulation of the visibility of the figure has indicated that the latter part of this response, around 200 ms, is more related to surface segmentation, while the earlier part of this response appears to be more related to boundary detection.

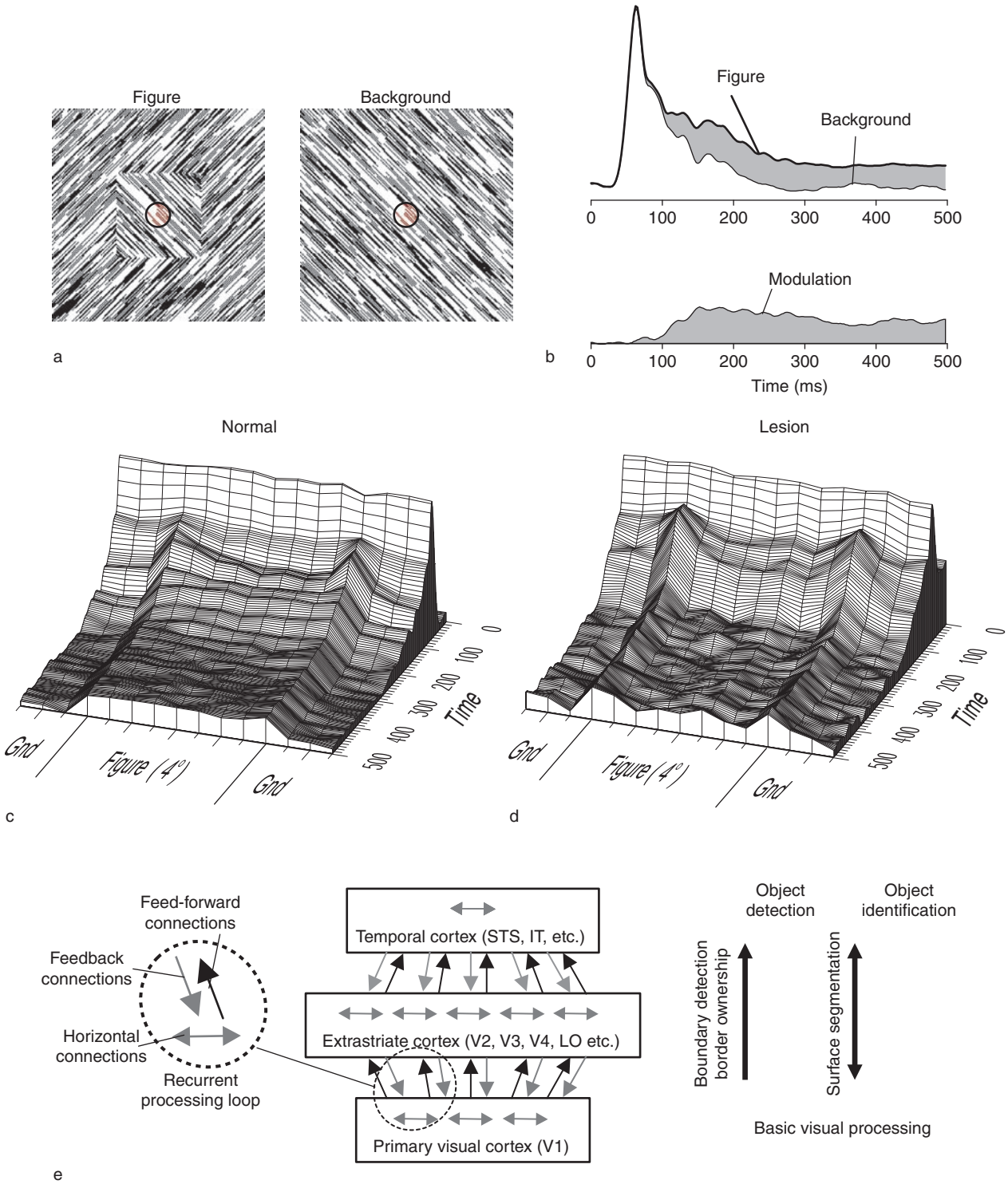

Figure 2 General phenomenology of contextual modulation. (a) The responses of a neuron in V1 (single unit, smoothed) when the classical receptive field (cRF) falls on a texture with homogeneously oriented line elements (top) and additional activity (bottom right, in gray) that can be measured when the line elements that fall in the cRF are part of a figure. (b) The neural activity (*z*-axis, smoothed spiking activity) evoked by a figure texture plotted over time (more vertical axis) and across a line circumcising a figure (more horizontal axis) measured in V1 of a macaque monkey. Notice that the neurons start to fire vigorously after 40 ms (visual response); 40 ms later, is possible to see a difference in activity between the neurons with their classical receptive field on the edge and the other neurons. After this, is becomes possible to see 'filling-in' in the area between the boundaries. (c, d) When the peristriate areas are removed, isolating V1 and V2 from the rest of the brain, activity related to the 'filling-in' disappears while the activity related to boundary detection remains visible. (e) The recurrent model of surface segmentation. Boundary detection and border ownership are established during the initial feed-forward sweep. This results in coarse object detection, after which recurrent processing starts, which results in a surface segmentation representation in early visual cortex. Over time, this results in a more specific, object identity-related response and disambiguation of the visual scene.

Influence of Recurrent Processing

Apart from connections running from lower tier areas to higher tier areas there are also connections running from higher tier areas to lower tier areas, and horizontal connections that provide connectivity within an area (see Figure 2(e)). After the initial feed-forward sweep of activity, these horizontal and feedback connections start to contribute to neural processing, and this is called recurrent processing. Contextual modulation is a result of recurrent processing, involving both horizontal connections and feedback processing.

Lesion Studies

When a peristriate cortex lesion is made in visual cortex, isolating areas V1 and V2 from the rest of the cortex, responses to the presence of the boundaries between figure and ground remain. This indicates that boundary detection is completed within areas V1 and V2. However, the subsequent response toward the figure, which signals surface segmentation, is abolished by this lesion (see Figure 2(d)). This implicates the involvement of recurrent or feedback processing – that is to say, the influence from higher tier areas on lower tier areas in surface segmentation. Likewise, it was shown, by using reversible inactivation of the medial temporal (MT) area, that feedback connections serve to amplify the response of neurons in area V1 and enhance suppression evoked by the background.

Speed of Processing in the Visual System

Feed-forward and feedback corticocortical connections conduct information at an extremely high speed, approximately 2–3.5 m s^{-1}, activating neurons in area V1 between 35 and 72 ms, neurons in the gyrus part of the inferior temporal cortex (area TE) between 60 and 120 ms, and neurons in the temporal pole between 100 and 160 ms. Taken that the synaptic relays add only an additional 5–20 ms to the relay time per cortical area and that there are long-distance corticocortical projections from, for instance, V1 to MT, it is no surprise that all visually responsive areas in the macaque cortex are activated approximately 115 ms after stimulus onset, including regions in motor and frontal cortex.

This rapid speed of the corticocortical projections contrasts with the lower conduction velocity of horizontal connections that are in the order of 0.1–0.33 ms^{-1}. This tenfold difference in conduction speed is caused by a difference in myelination and diameter between horizontal and interareal corticocortical connections.

Contextual effects of scene elements at a distance up to 3–10° influence the measured spiking activity 100 ms after stimulus presentation. At a range of 3°, eccentricity span horizontal connections are such that they can influence processing in this time interval in the range of 4°. However, because of the cortical magnification factor, this influence should be slower for more centrally presented stimuli and faster for more peripherally presented stimuli. Since there is only a modest dependency on cortical distance, the most consistent interpretation is that contextual modulation at some distance from the cRF is mainly caused by recurrent processing.

The Need for Recurrent Processing in Surface Segregation: Disambiguation

Neural network models of surface segregation suggest that recurrent processing is necessary to accurately perceive complex aspects of an object shape because higher-tier visual areas lack the spatial resolution of early visual cortex. If we consider a feed-forward model of perception in which neurons in higher-tier areas look at the output of lower-tier areas, it is difficult to see how a neuron in a higher tier area can differentiate between a filled figure, on the one hand, and a figure with a hole in the center, on the other hand. Adding recurrent processing to such a model does allow the system to differentiate between different geometric shapes. From this perspective, one of the purposes of recurrent processing in surface segregation is the disambiguation of information in the feed-forward sweep.

The Need for Recurrent Processing in Surface Segregation: Object Identification

Historically it has been presumed that surface segmentation is subservient to object recognition, assuming that object representations are only selected for regions selected by prior surface segmentation. However, neurons in area IT show object specific activation within 80 ms after stimulus onset. This is before surface-segregation-related activity appears in area V1 (100 ms and onward) and this indicates that object-related activity can precede surface segregation signals in V1. Furthermore, experiments have shown that object-based responses in area IT develop over time from categorical responses to exemplar-related responses between approximately 100 and 200 ms.

These observations have found behavioral support in experiments that show that object categorization does not require a longer stimulus presentation or a longer response time than does detecting the presence of an object. Since figure–ground segmentation should

be sufficient for accurate performance in object detection (i.e., detecting the presence of an object, regardless of its identity), this indicates that figure–ground segmentation does not precede object categorization in all circumstances, since detection and categorization take an identical amount of processing. The same experiments have shown that object identification takes a longer stimulus duration and processing time than does categorization. Together, these results suggest that recurrent processing, visible in area V1 after 100 ms, is necessary for the identification, but not detection\categorization, of visual stimuli, and that the activity that is measured in area V1 after 100 ms represents the ongoing perceptual interpretation of the visual scene in a high spatial resolution.

Surface Segregation Is Part of a Dynamic Process

Surface segregation is not a neural computation that takes place in one area of the brain in a limited window of time, but rather is a dynamic process that evolves over time and space. First, in a feed-forward sweep of activation, initial boundaries and border ownership of these boundaries are established. A recent magnetoencephalographic (MEG) study looking at the time course of illusionary shape-related activity showed that area LO was activated at 155 ms, after which activity propagated to early visual cortex and higher tier areas. If we take the response properties of area LO into account, it could very well be that this area is a necessary relay station between early visual cortex and object-selective areas in the temporal lobe (see Figure 2(e), black arrows). Recurrent processing between these areas results in a surface representation in area V1 and an increase in the tuning of neurons in object-related areas. Furthermore, it seems likely that the brain tries to solve ambiguities present in the visual image during these later stages of processing.

See also: Contextual Interactions in Visual Perception; Contextual Interactions in Visual Processing; Shape Representation in Inferotemporal Cortex; Visual Cortex in Humans; Visual Attention; Visual Cortical Models of Orientation Tuning.

Further Reading

Allbright TD and Stoner GR (2002) Contextual influences on visual processing. *Annual Review of Neuroscience* 25: 339–379.

Bach M and Meigen T (1992) Electrophysiological correlates of texture segregation in the human visual evoked-potential. *Vision Research* 32: 417–424.

Bair W, Cananaugh JR, and Movshon JA (2003) Time course and time–distance relationships for surround suppression in Macaque V1 neurons. *Journal of Neuroscience* 23: 7690–7701.

Bullier J, Hupe J, James A, et al. (2001) The role of feedback connections in shaping the responses of visual cortical neurons. *Progress in Brain Research* 134: 193–204.

Caputo G and Casco C (1999) A visual evoked potential correlate of global figure–ground segmentation. *Vision Research* 39: 1597–1610.

Das A and Gilbert CD (1999) Topography of contextual modulations mediated by short-range interactions in primary visual cortex. *Nature* 399: 655–661.

Grill-Spector K and Kanwisher N (2005) Visual recognition – as soon as you know it is there, you know what it is. *Psychological Science* 16: 152–160.

Halgren E, Mendola J, Chong C, et al. (2003) Cortical activation to illusory shapes as measured with magnetoencephalography. *NeuroImage* 18: 1001–1009.

Kapadia MK, Ito M, Gilbert CD, et al. (1995) Improvement in visual sensitivity by changes in local context – parallel studies in human observers and in V1 of alert monkeys. *Neuron* 15: 843–856.

Koffka K (1935) *Principles of Gestalt Psychology.* London: Lund Humphries.

Kourtzi Z and Kanwisher N (2001) Representation of perceived object shape by the human lateral occipital complex. *Science* 293: 1506–1509.

Lamme VAF (1995) The neurophysiology of figure–ground segregation in primary visual-cortex. *Journal of Neuroscience* 15: 1605–1615.

Lamme VAF and Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience* 23: 571–579.

Lamme VAF, Vandijk BW, and Spekreijse H (1992) Texture segregation is processed by primary visual-cortex in man and monkey – evidence from Vep experiments. *Vision Research* 32: 797–807.

Li W, Piech V, and Cilbert CD (2006) Contour saliency in primary visual cortex. *Neuron* 50: 951–962.

Nakayama K, Shimojo S, and Silverman G (1989) Stereoscopic depth – its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception* 18: 55–68.

Repas JB, Niyogi S, Dale AM, et al. (1997) Representation of motion boundaries in retinotopic human visual cortical areas. *Nature* 388: 175–179.

Roelfsema PR, Lamme VAF, Spekreijse H, et al. (2002) Figure–ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience* 14: 525–537.

Rossi A and Paradiso M (1996) Temporal limits of brightness induction and mechanisms of brightness perception. *Vision Research* 36: 1391–1398.

Zhou H, Friedman H, and von der Heydt R (2000) Coding of border ownership in monkey visual cortex. *Journal of Neuroscience* 20: 6594–6611.

Zipser K, Lamme VAF, and Schiller PH (1996) Contextual modulation in primary visual cortex. *Journal of Neuroscience* 16: 7376–7389.

Visual Associative Memory

H M Kimura, K Nakahara, and Y Miyashita,
University of Tokyo School of Medicine, Tokyo, Japan

© 2009 Elsevier Ltd. All rights reserved.

Associative Memory in Vision: An Elementary Model for Declarative Memory Networks

A significant aspect of high-level cognitive function in humans is our ability to categorize and memorize declarative knowledge and experiences in associative networks that can be reactivated on demand by both internal and external cues. In seeking to better understand the neuronal mechanisms underlying these processes, visual associative memory offers an attractive model that is readily accessible to empirical neurophysiological approaches.

The primate visual system is divided into two functionally and anatomically distinct streams: the dorsal stream and the ventral stream. The dorsal stream, which conveys information from early visual areas to the posterior parietal cortex, is involved in processing the location and motion of visual objects (i.e., where); interruption of this stream disrupts the perception of the spatial relations among objects. On the other hand, the ventral stream conveys information from early visual areas to the inferior temporal (IT) cortex and is involved in processing color and shape (i.e., what); interruption of this stream affects the ability to recognize objects and discriminate among them. In monkeys, bilateral excision of the anterior IT cortex causes severe deficits in visual object recognition and impairs visual pattern discrimination, whereas more posterior, temporal-occipital lesions produce more severe perceptual deficits. Similarly, patients who have undergone right anterior temporal lobectomy show mild impairment in perceptual tasks such as the Mooney Closure Faces Test (in which subjects are asked to visually organize a percept of a face from degraded images), but show markedly impaired recognition memory of complex visual patterns which cannot be verbalized. These deficits are unrelated to the extent of hippocampal removal during the temporal lobectomy. Moreover, electrical stimulation of the temporal lobes of conscious epileptic patients during brain surgery induces vivid, almost hallucinatory memories (i.e., faces of friends and sounds of music), suggesting that electrical input to the putative memory storehouse reactivates the memory records.

These findings suggest that the IT cortex contributes to the representation and storage of visual memories. If this is so, then how are visual associative links represented within the IT cortex? This article focuses on the neuronal substrates and mechanisms underlying visual associative long-term memory (LTM), mainly revealed by single-unit recordings from the brains of monkeys performing memory tasks. We first introduce neurophysiological evidence for neuronal representations of visual associative memories in the monkey IT cortex. We then present single-unit recordings indicating that neural activity within the monkey IT cortex also contains the retrieval signal of visual associative memories. Finally, we discuss how the visual association is recalled from memory by reactivation of the neuronal representation, focusing on two different processes mediating memory retrieval: automatic memory retrieval signaling, which is conveyed backwardly from the limbic cortex to the temporal neocortex, and active retrieval signaling, which is conveyed as top-down signaling from the prefrontal cortex (PFC) to the IT cortex.

Neuronal Correlates of Associative Memory in the Monkey Inferior Temporal Cortex

Inferior Temporal Neurons Establishing Visual Associative Links

The neuronal correlates of visual associative LTM were first identified by Miyashita and Sakai in the monkey IT cortex. They recorded from single neurons while the subjects performed a pair-association (PA) task, which is a well-known neuropsychological test widely used in humans to assess dysfunction of the medial temporal lobe system. In the PA task, the monkeys were trained to memorize associations between computer-generated paired pictures that had no apparent geometric similarity (Figures 1(a) and 1(b)). At the start of each trial, one of the pictures was presented as a cue (i.e., 1 in Figure 1(b)). After a delay interval, during which no stimulus was presented, the subject was presented with a choice between the paired associate of the cue (i.e., 1′) and a stimulus from another pair. Using this approach, two classes of task-related neuronal activities were identified in the IT cortex. In one class of neurons, the strongest and second strongest responses during the cue period were ascribed to particular paired pictures (pair-coding neurons; Figures 1(c) and 1(d)). The other type of neurons not only responded to an optimal stimulus during the cue period, but also was active during the blank delay period after the paired associate of the optimal picture was presented as a cue. This delay activity gradually increased until

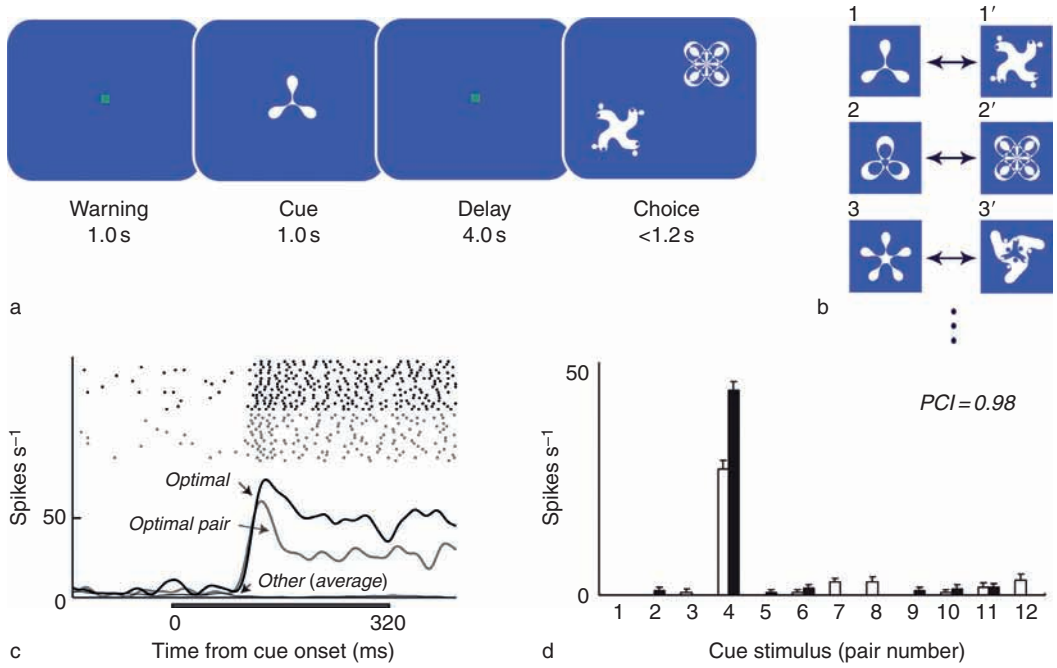

Figure 1 Pair-association (PA) task for monkeys used by Sakai and Miyashita: (a) sequence of events within a trial; (b) examples of paired associates used in the task (the monkeys learned to retrieve the other member of the associated pair for each cue picture); (c) rastergrams of neural discharges in each trial (upper) and spike density functions (SDF, lower) obtained from a single neuron; (d) mean discharge rates during the cue period (60–320 ms from the cue onset). As shown in (a), after a lever press by the monkey initiating a new trial, there is a warning, a green square (1 s); then a cue, one of the Fourier descriptors that serves as the cue stimulus (1 s); then a delay, a green square (4 s); and then a choice of two stimuli, the paired associate of the cue and one from a different pair. Fruit juice was given as a reward for correctly touching the paired associate. In (c), the optimal trial (the strongest response) is shown as the thick black line and optimal-pair trial (the second strongest response) is shown as the thick gray line. The trials are aligned at the cue onset. The thin black line denotes the averaged responses in the other trials. The duration of the cue period was 320 ms. In (d), 12 pairs of stimuli are labeled on the abscissa. The open and filled bars in pair 1, for example, refer to the responses to stimulus 1 and 1′, respectively. Error bars denote SEM. PCI, pair-coding index. (c, d) Reproduced from Naya Y, Yoshida M, and Miyashita Y (2003) Forward processing of long-term associative memory in monkey inferotemporal cortex. *Journal of Neuroscience* 23: 2861–2871. Copyright 2003, with permission from the Society for Neuroscience.

the choice stimuli appeared (target-recall activity; see the section titled 'Two types of sustained activities in the delay period'). This pattern of activity includes the ability of IT neurons to acquire selectivity for visual patterns through associative learning and to establish links between neuronal representations of different stimuli that have a semantic association but no geometrical relationship. Neurophysiological studies of monkeys by other groups lend additional support to this notion of the properties of the IT cortex. For instance, the Logothetis and Tanaka research groups reported that the visual stimulus selectivity of neurons in TE showed a training effect – the proportion of neurons responsive to trained stimuli was significantly greater in trained monkeys than in control untrained monkeys. In addition, Erickson and Desimone recorded single neurons in the perirhinal area of the IT cortex while monkeys performed a visual cue-choice association task and found that the correlation relating responses to a familiar cue and a choice stimulus is highly significant.

Forward Processing of Long-Term Associative Memory

The IT cortex comprises two cytoarchitectonically distinct but mutually interconnected areas: TE and A36 (Figure 2). TE is a unimodal neocortical area located at the final stage of the ventral visual stream devoted to object vision. By contrast, A36 is a limbic polymodal association area and a component of the medial temporal-lobe memory system, which is involved in the formation of declarative memory. The respective roles played by these two areas in the neuronal representation of associative memory were analyzed in terms of the pair-coding responses of neurons. A correlation coefficient was calculated for each neuron between the mean firing rate during the cue period to one stimulus and the mean firing rate during the cue period to the paired associate of that stimulus (pair-coding index (PCI)). If a single neuron in a population showed a pattern of stimulus selectivity that was independent of the stimulus pairs, the

mean value of the PCIs would be expected to approach 0 as the number of neurons in the population increased. This analysis showed that the PCI values for A36 neurons were significantly higher ($P < 0.001$) than those for TE neurons (Figure 3(a)) and that the percentage of neurons which showed significantly positive PCIs at the single-neuron level (pair-coding neurons) was also much higher in A36 (33%) than TE (4.9%). However, the general stimulus selectivities of the cue-selective neurons (as measured by R^2 and kurtosis in Figures 3(b) and 3(c)) did not differ between the two areas, indicating that the associative memory is represented in a highly specific change of neuronal stimulus selectivity (as measured by the PCI).

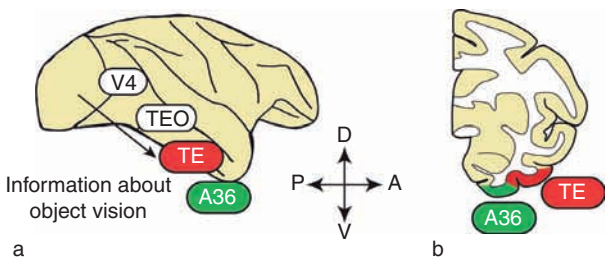

Figure 2 Macaque brain: (a) lateral view; (b) coronal cross section. A36, area 36 (green); TE, area TE(red); TEO, area TEO; V4, visual area 4; D, dorsal; V, ventral; A, anterior; P, Posterior. Reprinted from Naya Y, Yoshida M, and Miyashita (2001) Backward spreading of memory-retrieval signal in the primate temporal cortex. *Science* 291: 616–664, with permission from American Association for the Advancement of Science.

Next, the time courses of the responses of TE and A36 pair-coding neurons were compared, and it was observed that most of the pair-coding neurons in A36 began to encode the association between paired stimuli as soon as they exhibited stimulus selectivity (called type 1 neurons); the remaining pair-coding neurons in A36 began to encode the association 20–30 ms after they began to show stimulus selectivity (type 2 neurons). This suggests that type 1 neurons encode the associative memory by directly combining the feed-forward input from TE, whereas type 2 neurons more likely encode the association memory using feedback input from other higher areas (see the section titled 'Active retrieval: Top-down signal from the prefrontal cortex in executive control of memory retrieval') or intrinsic input from other A36 neurons.

Collectively then, the findings presented so far indicate that the association between the representations of paired-associates proceeds forward through the anatomical hierarchy of the IT cortex and suggest

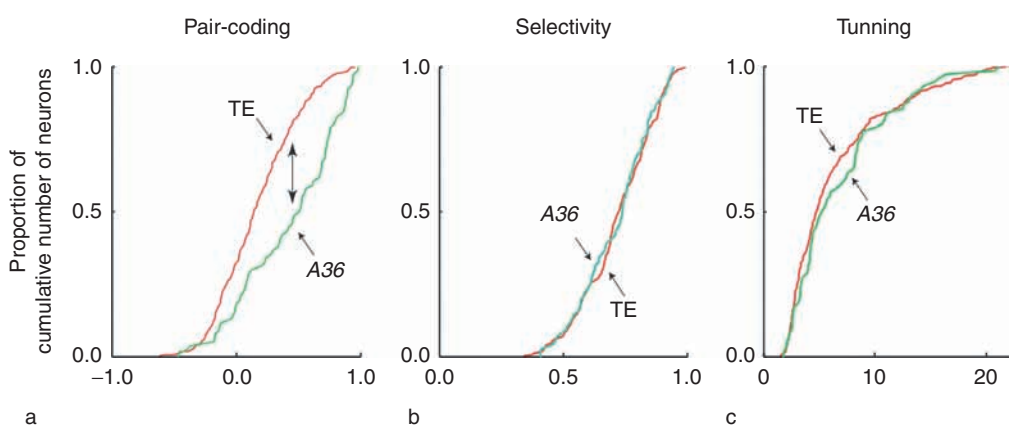

Figure 3 Response correlation to paired associates and general response properties of cue-selective neurons in A36 (green) and TE (red): (a) cumulative frequency histogram of PCIs for A36 ($n = 76$) and TE ($n = 347$) neurons; (b) cumulative frequency histogram of R^2; (c) cumulative frequency histogram of kurtosis. In (a), the PCIs for A36 neurons were significantly higher than those for TE neurons ($P < 0.001$; Kolmogorov-Smirnov test). In (b), R^2 provides an estimate of how much of the variance in firing rate can be accounted for by the factor of stimulus. In (c), kurtosis provides a measure of the sharpness of the stimulus selectivity. Neither the R^2 ($P < 0.99$) nor kurtosis ($P = 0.41$) significantly differs between A36 and TE neurons. The equations used are:

- In (a), $\mathrm{PCI} = \sum[(x_i - \mu)(x_i - \mu')]/\{\sum[(x_i - \mu)^2]\sum[(x_i - \mu)^2]\}^{1/2}$, $(i = 1, \ldots, 12)$; where x_i is the mean cue response for the ith stimulus (the ith and i'th pictures belong to a pair), and μ and μ' are the average of x_i and $x_{i'}$., respectively.
- In (b), $R^2 = \sum m_j(x_j - \mu_{\mathrm{tot}})^2/\sum(x_{j,k} - \mu_{\mathrm{tot}})^2 (j = 1, \ldots, 12, 1', \ldots, 12'; k = 1 - m_j)$;
- In (c), kurtosis $= \sum(x_j - \mu_{\mathrm{tot}})^4/\{\sum(x_j - \mu_{\mathrm{tot}})^2\}^2 (j = 1, \ldots, 12, 1', \ldots, 12'; k = 1 - m_j)$, where in the last two equations, $x_{j,k}$ is the cue response in the kth trial for the jth stimulus; x_j is the mean cue response for the jth stimulus; m_j is the total number of the trials for the jth stimulus; and μ_{tot} is the average of the cue responses across the total trials.

A36, area 36; TE, area TE; PCI, pair-coding index. Reproduced from Naya Y, Yoshida M, and Miyashita Y (2003) Forward processing of long-term associative memory in monkey inferotemporal cortex. *Journal of Neuroscience* 23: 2861–2871. Copyright 2003, with permission from the Society for Neuroscience.

selective convergence onto a single A36 neuron from two groups of TE neurons that separately encode each paired picture.

Retrieval Signaling in Visual Association Memory

Two Types of Sustained Activities in the Delay Period

In the IT cortex, two types of sustained activities were found in the delay period when the subject retrieved and maintained the target stimulus from LTM. One population of IT neurons exhibits a stimulus selectivity pattern during the delay period that was similar to that during the cue period. That is to say, this sustained activity in optimal trials (i.e., the trials in which the optimal stimulus was presented as a cue stimulus) related to the cue stimulus itself – cue-holding activity (Figures 4(a) and 4(b)). On the other hand, another population of neurons exhibits target-recall activities, which was elicited by the paired associate of the cue stimulus rather than by the cue stimulus itself as previously discussed (Figures 4(c) and 4(d)). This sustained activity in optimal-pair trials (i.e., the trials in which the paired associate of the optimal stimulus was presented as a cue stimulus) is induced only when the subject retrieves and maintains the optimal stimulus in mind and thus reflects the memory of the optimal stimulus. To further characterize the cue-holding and target-recall activities, the responses to all of the cue stimuli were considered. The partial correlation coefficients relating delay activity to the responses to a particular cue stimulus (cue-holding index (CHI)) and to its paired associate (pair-recall index (PRI)) were calculated for each neuron. CHI and PRI values are bounded by −1 and 1; a large CHI implies cue-holding activity as shown in Figures 4(a) and 4(b) (CHI = 0.95, PRI = −0.47), whereas a large PRI implies target-recall activity as shown in Figures 4(c) and 4(d) (CHI = −0.01, PRI = 0.64).

Then, how are these two types of delay activities, cue-holding and target-recall, represented in TE and A36? The comparison of the delay activities between individual TE and A36 neurons during a PA task in terms of their signal content provided the answer to this question. A comparison of the CHI and PRI values for TE and A36 neurons revealed that in TE the PRIs for the delay-selective neurons were significantly higher than the CHIs. In A36, by contrast, the

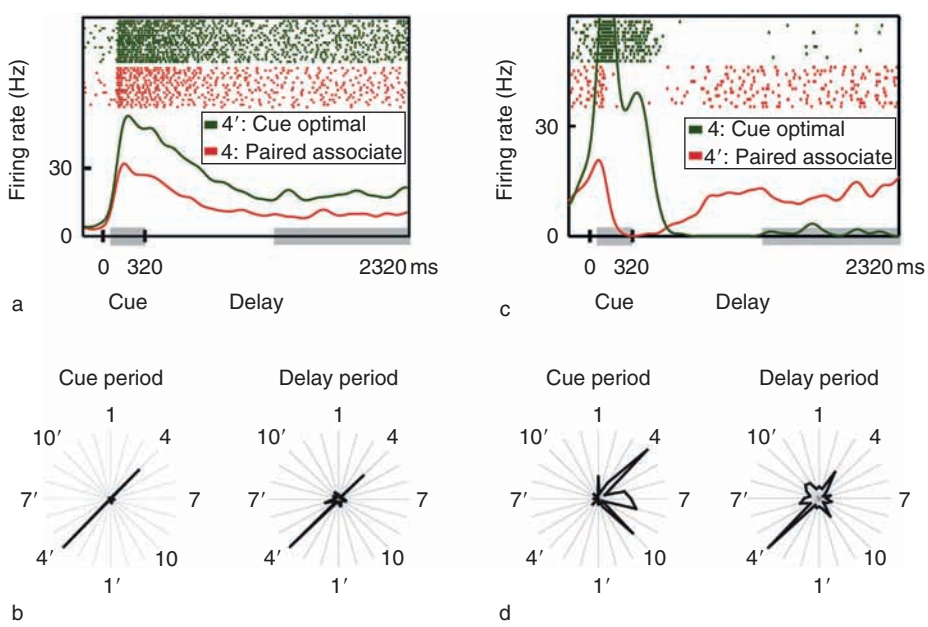

Figure 4 Examples of delay-selective neurons: (a) A36 raster displays and spike density functions; (b) A36 polar plots of mean discharge rates; (c) TE raster displays and spike density functions; (d) TE polar plots of mean discharge rates. In (a,c) , trials using the cue-optimal stimulus as the cue (optimal trials, green) and those using the paired associate as the cue (optimal-pair trials, red) were aligned at the cue onset. In (b) and (d), mean discharge rates during the cue and delay periods (60–320 ms and 1320–2320 ms; gray boxes in (a) and (c), respectively) are shown in polar plots for each cue presentation. The responses to stimuli and to their paired associates are indicated by radial lines. The discharge rates were normalized based on the maximum values for each period ((b) cue period, 44.0 Hz at stimulus 4′; delay period, 18.6 Hz at stimulus 4′; (d) cue period, 48.7 Hz at stimulus 4; delay period, 18.7 Hz at stimulus 4′). A36, area 36; TE, area TE. Reproduced from Naya Y, Yoshida M, Takeda M, Fujimichi R, and Miyashita Y (2003) Delay-period activities in two subdivisions of monkey inferotemporal cortex during pair association memory task. *European Journal of Neuroscience* 18: 2915–2918. Copyright 2003, with permission from Blackwell Publishing Ltd.

PRIs for the delay-selective neurons were not different from the CHIs. Moreover, the CHIs were significantly lower in TE than in A36, whereas the PRIs did not significantly differ in the two areas. This suggests that the delay activity of TE neurons preferentially represents the paired associate that is retrieved from LTM, not the cue stimulus itself, whereas the delay activity of A36 neurons retained not only the paired associate but also the cue stimulus.

Active Maintenance of Associative Mnemonic Signaling in the Monkey Inferior Temporal Cortex

In another study focusing on delay activity with the aim of determining the contribution made by IT neurons to the persistence of internal representations, single units were recorded from neurons in the IT cortex while monkeys performed a sequential type of PA task in which distractor stimuli interrupted the delay period between a cue stimulus and the target stimulus (the paired associate of the cue stimulus) (Figure 5(a)). Some delay-selective neurons exhibited, in optimal trials, cue-holding activity during the first delay period. This sustained discharge was attenuated when the distractor stimulus was presented (compare the two light-gray shaded areas in Figure 5(b)). During optimal-pair trials, the neuronal discharge of the same cell was weak throughout the trial (Figure 5(c)). However, during optimal-pair trials another type of neuron exhibited target-recall activity during the first delay period, and the significant activity was still sustained during the second delay period (compare the two light-gray shaded areas in Figure 5(e)). During optimal trials, by contrast, the neuronal discharge of the same cell was weak both during the first and the second delay periods (Figure 5(d)). These results mean that the cue-holding activity tends to be attenuated by the subsequent distractor stimulus (in fact, this activity is replaced by that elicited by the distractor stimulus) during the second delay period (i.e., after the presentation of the distractor stimulus); on the other hand, the target-recall activity continues to signal the target stimulus information, even after an intervening distractor.

A population analysis of all the delay-selective neurons recorded is shown in Figures 5(f) and 5(g). Partial regression coefficients (β) were used to relate the neuronal discharge during the first delay period and second delay periods to the discharge elicited by the cue and target. A larger value of β means that the delay activities convey more information about the cue or target.

The scatterplots relates the β values for the cue (Figure 5(f)) and target (Figure 5(g)) stimuli during the first and second delay periods. Note that the values for the cue stimulus lie below the orthogonal line, whereas those for the target stimulus lie around it. This indicates that the target stimulus information signaled by delay activity is more resistant to the distractor stimulus than the cue stimulus information and suggests that, whereas LTM-derived information required for upcoming behavior (i.e., target stimulus) is actively maintained in IT neurons, visually derived information tends to be updated irrespective of behavioral relevance (i.e., cue stimulus and distractor stimulus).

Global Network of Memory Retrieval

Thus far, we have provided evidence as to how visual associative memory is stored and represented in the IT cortex. We next discuss how these mnemonic representations are reactivated during memory retrieval. The evidence summarized shows that memory retrieval can be processed either within the IT cortex or via an interaction between the PFC and the IT cortex.

Everyone knows from experience that sometimes recollection is effortless but that at other times one must make an effort to recall successfully. Here, we refer to the former as automatic retrieval and to the latter as active retrieval. Visual associative memories can be retrieved from the IT cortex by either of these two processes.

Automatic Retrieval: Backward Spreading of the Memory-Retrieval Signal in the Temporal Cortex

During a PA task, groups of neurons in both TE and A36 show target-recall activity. The responses of representative A36 neuron and TE neuron are shown in Figures 6(a) and 6(c). One stimulus elicited the strongest response during the cue period (perceptual signal), and the latencies of the response in TE neurons were significantly shorter than those in A36 neurons (TE, median 77 ms; A36, median 89 ms; Kolmogorov–Smirnov test, $P < 0.05$). Moreover, the paired-associate stimulus of a cue stimulus elicits the highest delay activity in A36 and TE neurons, although the onset of the target-recall tonic activity was later in TE than in A36 (Figures 6(a) and 6(c)).

To compare the properties of the target-recall activities between TE and A36 neurons, the time courses were analyzed. In this analysis, an instantaneous PRI at time plot t (PRI(t)) was defined as partial correlation coefficients relating instantaneous firing rates at time t in the response to each cue stimulus to the responses to its paired associate. Then the time courses of the average PRI(t) across the population of stimulus-selective neurons in A36 and TE were determined. The average PRI(t) for the A36 neurons began to increase within the cue period and developed over a

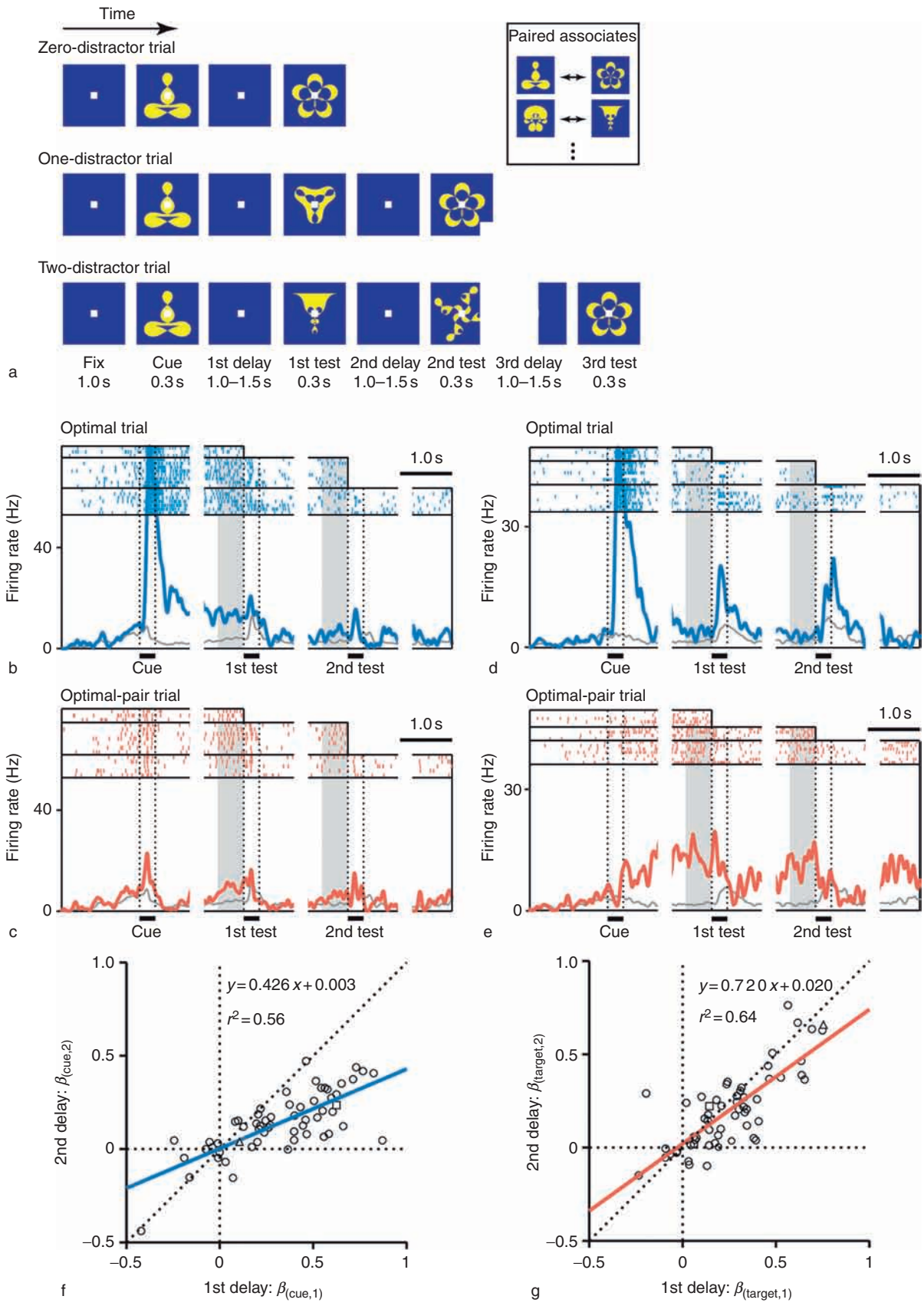

Figure 5 (Continued)

comparatively rapid time course (Figure 6(b)). On the other hand, the PRI(t) for the TE neurons increased slowly, reaching a plateau in the middle of the delay period (Figure 6(d)). Apparently, even though the visual signal reaches TE before it reaches A36, the memory retrieval signal emerges first in A36 and then TE neurons are gradually recruited to represent the sought target. The implication of this finding is that backward projections from A36 to TE may underlie the activation of TE neurons that represent mnemonic information retrieved from LTM.

Active Retrieval: Top-Down Signal from the Prefrontal Cortex in Executive Control of Memory Retrieval

Memory retrieval can be regulated by top-down processes that have been implicated in executive functions, such as planning. For instance, examination of an epileptic patient with a selective posterior callosotomy revealed that although sensory stimuli confined to the nondominant right hemisphere could not be transferred to the left hemisphere for naming, semantic features of those stimuli could be described using the expressive language system of the left hemisphere. This observation suggests that top-down processes originating in the PFC can regulate the retrieval of LTM from the modality-specific posterior association cortex, even in the absence of direct sensory input.

To test that idea, the posterior corpus callosum (CC) and the anterior commissure were transected in monkey subjects (the posterior split-brain paradigm), after which the monkeys' performances on a PA task were evaluated. The critical question was whether the PFC of these monkeys could instruct the contralateral hemisphere, via the intact anterior CC, to retrieve LTM when sensory interaction between posterior cortical areas was prevented. In the posterior-split-brain monkey, each hemisphere receives bottom-up visual information only from the contralateral visual hemifield, and LTM acquired through learning pair association does not transfer interhemispherically via the anterior CC. Nonetheless, information about the cue stimulus presented to one hemisphere was transmitted through the anterior CC, enabling the target stimulus specified by the cue to be retrieved from the other hemisphere. Thus, we can conclude that, although visual LTM is stored in the temporal cortex, memory retrieval is under the executive control of the PFC.

The existence of top-down signaling from the PFC to the temporal cortex and its contribution to the active retrieval process have been demonstrated using single-unit recordings from the IT cortex of posterior-split-brain monkeys. In these animals, IT neurons in one hemisphere were activated by bottom-up visual inputs when an object was presented in the visual hemifield contralateral to the recording site (Figure 7(a)). When the object was presented in the ipsilateral hemifield, however, these neurons did not receive the bottom-up visual inputs (Figure 7(b)), so any neural activation should reflect top-down inputs from the PFC. And, in fact, a considerable number of IT neurons received top-down signals from the PFC (Figure 7(c)). The neurons were activated not only by contralateral presentation of stimuli but also by ipsilateral presentation of stimuli. The stimulus selectivity of the bottom-up response was similar to that of the top-down response. The onset latency of neuronal responses was significantly longer under the top-down than the bottom-up condition, and top-down responses were abolished after transection of the remaining anterior CC (full-split).

The memory retrieval is a complicated process; it is not just the process of cueing and interrogation of the memory store (a narrow sense of memory retrieval), but it is also bound up with the updating of memory information in working memory and evaluation or monitoring of this information. Recent positron

Figure 5 Sequential-type paired-association task. (a) The procedure; (b,c) rastergrams and SDFs of the optimal trials and optimal-pair trials showing representative neurons exhibiting cue-holding activity; (d,e) rastergrams and SDFs of the optimal trials and optimal-pair trials showing representative neurons exhibiting target-recall activity; (f) scatterplot of the effect of the cue stimulus on neuronal activity during the delay epoch; (g) scatterplot of the effects of the target stimulus on neuronal activity during the delay epoch. As shown in (a), after the fixation spot, one of the pictures was presented as the cue stimulus. During the delay epoch that followed the cue stimulus, one, two, or three test stimuli were presented. After each delay period, the paired associate of the cue stimulus (target stimulus) or a distractor stimulus was presented as the test stimulus. In (b–e), gray lines indicate average SDFs in trials in which the other pictures were presented as cue stimuli. Data from correct trials are aligned on the onset of the cue, first-test, and second-test stimuli. Light-gray shading indicates the first and second delay periods for analysis. In the optimal trials, neuronal discharge (cue-holding activity) declined from the first to the second delay period (b). In the optimal-pair trials, strong neuronal discharge (target-recall activity) was observed during both the first and second delay periods (e). In (f), the scatterplot relates the changes in the β values for the cue stimulus from the first delay (x axis) to the second delay (y axis) period across the population of delay-selective neurons ($n = 59$). The regression slope is significantly positive ($t = 8.58$, $p < 0.01$). In (g), the regression slope for the β values for the target stimulus is significantly positive ($t = 9.95$, $p < 0.01$) and steeper than that for β values for the cue stimulus ($F = 11.57$, $p < 0.01$). The open triangles in (f, g) represent the data from the neurons in (b,c) and the open squares represent the data from the neurons in (d,e). SDF, spike density function. Reproduced from Takeda M, Naya Y, Fujimichi R, Takeuchi D, and Miyashita Y (2005) Active maintenance of associative mnemonic signal in monkey inferior temporal cortex. *Neuron* 48: 839–848. Copyright 2005, with permission from Elsevier.

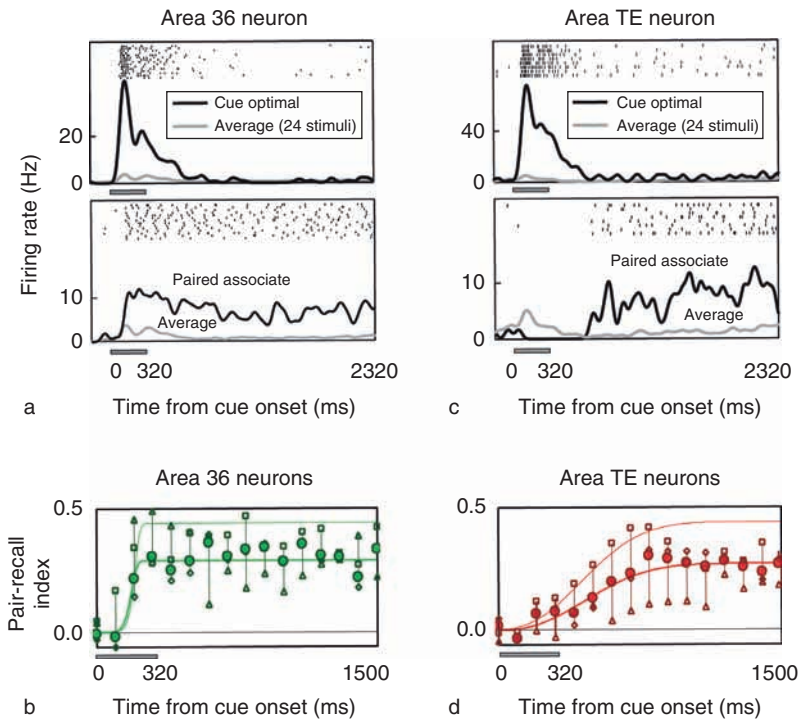

Figure 6 Neuronal activity related to memory retrieval during the paired-association task: (a) activity of a representative A36 neuron; (b) temporal dynamics of average PRI(t) for the population of stimulus-selective A36 neurons ($n = 45$); (c) activity of a representative TE neuron; (d) temporal dynamics of average PRI(t) for the population of stimulus-selective TE neurons ($n = 69$). In (a) and (c), SDFs are aligned at the cue onset in trials where the cue-optimal stimulus served as the cue (top) and in trials where its paired associate served as the cue (bottom). Black lines indicate responses to the cue-optimal stimulus (top) or its paired associate (bottom); gray lines indicate mean responses to all stimuli. In (b) and (d), mean values of PRI(t) are plotted every 100 ms (solid circle, total; diamond, monkey 1; open square, monkey 2; open triangle, monkey 3). Thick lines (green and red, respectively) indicate the best-fit Weibull functions for the population-averaged PRI(t) in the two areas. Thin lines indicate this for the neurons whose PRI(t) increased above the 5% significance level. PRI, paired-recall index; SDF, spike density function. Reprinted from Naya Y, Yoshida M, and Miyashita Y (2001) Backward spreading of memory-retrieval signal in the primate temporal cortex. *Science* 291: 661–664, with permission from American Association for the Advancement of Science.

emissions tomography (PET) and functional magnetic resonance imaging (fMRI) studies carried out in humans while they performed a memory retrieval task confirmed that the PFC is involved in these component processes. In that context, Fletcher and Henson investigated the roles of three regions of the lateral frontal cortex: the ventrolateral frontal cortex (VLFC; Brodmann's area (BA) 44, 45, 47), the dorsolateral frontal cortex (DLFC; BA 9, 46), and the anterior frontal cortex (AFC; BA 8, 10). In paired associate-cued retrieval tasks, activation of the VLFC contributed to updating and maintaining working memory (WM), within which the cue must be maintained along with possible responses retrieved from LTM. The DLFC, on the other hand, was involved in selecting, manipulating, and monitoring information already active in WM. Because efficient interaction between the VLFC and DLFC is necessary to meet the demands of the task, a meta-level of control is needed to coordinate them. Such control was attributed to the AFC, which optimally selects between the processes or goals of the VLFC and DLFC in order to maximize task performance. Earlier anatomical studies showed that BA 47/12 (an area within the VLFC) is strongly connected with the IT cortical region, which suggests that the top-down retrieval signaling from the PFC to the IT cortex seen in the single-unit study previously summarized may have originated from the homolog of BA 47/12 in the monkey PFC.

The actual retrieval processes are even more complicated than we mentioned in the preceding paragraph; they may entail encoding processes (i.e., when we remember something, we inevitably encode it again at the same time, and vice versa). Consequently, it is important to reevaluate the previous neuroimaging studies from the viewpoints that the memory retrieval process can be seen at different levels of the definition. Questions that remain include, are there the prefrontal regions which are specifically involved in retrieval processes in the narrowest definition? If yes, what signals are exchanged between these areas and the areas related to monitoring or selection or encoding? And which prefrontal area controls these complicated processes of across-areal communication?

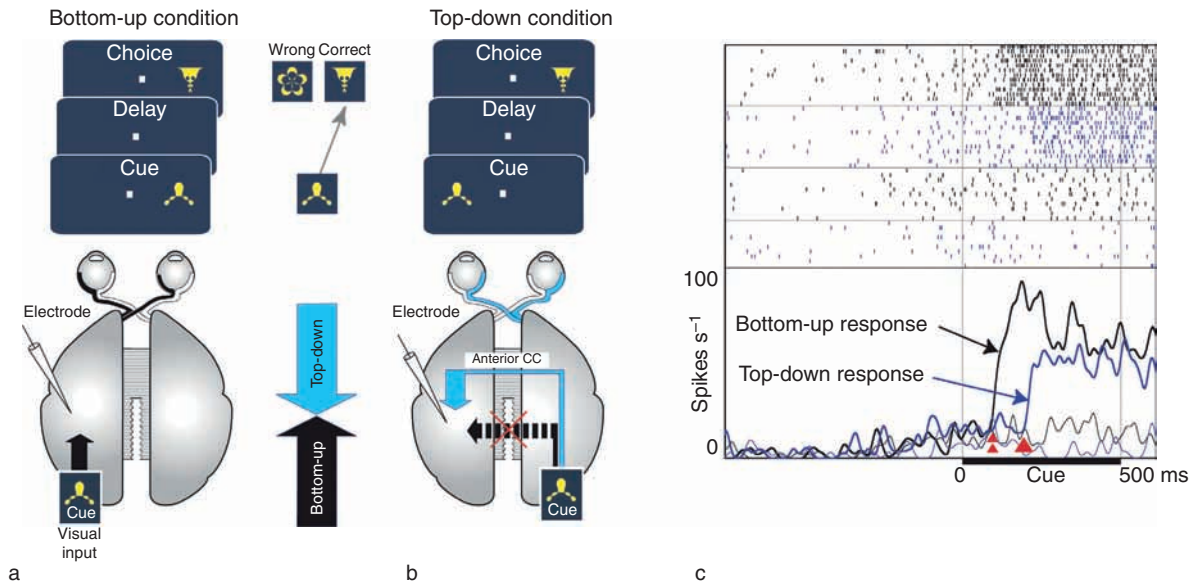

Figure 7 Experimental design measuring neuronal activity in monkeys: (a) bottom-up condition; (b) top-down condition; (c) activity of a single IT neuron under the top-down condition (top-down, blue; bottom-up, black). As shown in (a), in the bottom-up condition visual stimuli (cue and choice pictures) were presented in the hemifield contralateral to the recording site (electrode) in the IT cortex. The monkey was required to make the correct choice specified by the cue. Fixation was required throughout each trial ($< 0.6°$). Bottom-up sensory signals (black arrow) were detected in this condition. As shown in (b), in the top-down condition, the cue was presented in the hemifield ipsilateral to the recording site, but the choice was presented contralaterally. In posterior-split-brain monkeys, sensory signals do not reach the visual areas in the opposite hemisphere. In this condition, only top-down signals (blue arrow) can activate IT neurons through top-down connections from the prefrontal cortex. In (c), the raster displays and SDFs are aligned at the cue onset. In the SDFs, the thick lines show responses to the optimal cue; the thin lines show responses to a null cue. The onset of the top-down response (arrowhead) was later than the onset of the bottom-up response (doubled arrowhead). CC, corpus callosum; IT, inferior temporal; SDF, spike density function. Modified from Tomita H, Ohbayashi M, Nakahara K, Hasegawa I, and Miyashita Y (1999) Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* 401: 699–703, Copyright 1999 with permission from Macmillan Publishers Ltd.

Future research should provide answers to these and other questions, enabling a fuller understanding and appreciation of the neuronal mechanisms responsible for the management of semantic memories in primate brains.

See also: Amnesia: Declarative and Nondeclarative Memory; Animal Models of Amnesia; Declarative Memory System: Anatomy; Humans; Memory Representation; Multiple Memory Systems; Perirhinal Cortex: Neural Representations; Perirhinal Cortex; Semantic Memory; Shape Representation in Inferotemporal Cortex; Strategic Control of Memory; Visual System: Multiple Visual Areas in Monkeys.

Further Reading

Brown MW and Aggleton JP (2001) Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience* 2: 51–61.

Buckner RL and Wheeler ME (2001) The cognitive neuroscience of remembering. *Nature Reviews Neuroscience* 2: 624–634.

Fletcher PC and Henson RN (2001) Frontal lobes and human memory: Insights from functional neuroimaging. *Brain* 124: 849–881.

Milner B (1990) Right temporal-lobe contribution to visual perception and visual memory. In: Iwai R and Mishkin M (eds.) *Vision, Memory and the Temporal Lobe*, pp. 43–53. New York: Elsevier.

Miyashita Y (2004) Cognitive memory: Cellular and network machineries and their top-down control. *Science* 306: 435–440.

Murray EA and Bussey TJ (1999) Perceptual-mnemonic functions of the perirhinal cortex. *Trends in Cognitive Science* 3: 142–151.

Naya Y, Yoshida M, and Miyashita Y (2001) Backward spreading of memory-retrieval signal in the primate temporal cortex. *Science* 291: 661–664.

Naya Y, Yoshida M, and Miyashita Y (2003) Forward processing of long-term associative memory in monkey inferotemporal cortex. *Journal of Neuroscience* 23: 2861–2871.

Naya Y, Yoshida M, Takeda M, Fujimichi R, and Miyashita Y (2003) Delay-period activities in two subdivisions of monkey inferotemporal cortex during pair association memory task. *European Journal of Neuroscience* 18: 2915–2918.

Petrides M (2000) Frontal lobes and memory. In: Boller F and Grafman J (eds.) *Handbook of Neuropsychology*, pp. 67–84. Amsterdam: Elsevier.

Ranganath C and D'Esposito M (2005) Directing the mind's eye: Prefrontal, inferior and medial temporal mechanisms for visual working memory. *Current Opinion in Neurobiology* 15: 175–182.

Simons JS and Spiers HJ (2003) Prefrontal and medial temporal lobe interactions in long-term memory. *Nature Reviews Neuroscience* 4: 637–648.

Squire LR, Stark CEL, and Clark RE (2004) The medial temporal lobe. *Annual Review of Neuroscience* 27: 279–306.

Takeda M, Naya Y, Fujimichi R, Takeuchi D, and Miyashita Y (2005) Active maintenance of associative mnemonic signal in monkey inferior temporal cortex. *Neuron* 48: 839–848.

Tanaka K (1997) Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology* 7: 523–529.

Tomita H, Ohbayashi M, Nakahara K, Hasegawa I, and Miyashita Y (1999) Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* 401: 699–703.

Wagner AD, Desmond JE, Glover GH, and Gabrieli JD (1998) Prefrontal cortex and recognition memory: Functional-MRI evidence for context-dependent retrieval process. *Brain* 121: 1985–2002.

Visual Attention

S Treue, German Primate Center, Goettingen, Germany
S Katzner, Smith–Kettlewell Eye Research Institute, San Francisco, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

The visual system of humans and other primates is an evolutionary success story. It has developed an elaborate peripheral sensor system, the retina, in which the information picked up by more than 100 million receptors is compressed in a highly efficient manner and sent to the central nervous system by 1.5 million axons leaving the eyes. The complexity and level of sophistication of the retina is matched by a dedicated cortical processing hierarchy that encompasses more than 30 anatomically and/or functionally distinct areas covering approximately 50% of cerebral cortex in macaques and 20–30% in humans.

Although such a system picks up a plethora of information, even the highly developed visual cortex of primates would not be sufficient to process it all optimally and equally. Essentially, two interacting systems have evolved to counteract this limitation. A bottom-up system substantially compresses the information picked up by the retinal photoreceptors, and a top-down system dynamically allocates central processing resources for the visual information deemed most relevant in the current situation. The latter system, selective attention, is the focus of this article, but the two systems work hand-in-hand to provide an organism with the information most likely to be of relevance for its survival.

This article concentrates on our understanding of the neural basis of visual attention, emphasizing the most direct methodological approaches, namely single cell recordings from the visual cortex of behaving monkeys and, to a lesser degree, brain imaging measurements in human subjects. The central task for studies aimed at elucidating the neural correlates of attention is to evaluate linking hypotheses – that is, proposals of how the activity of neurons (in visual cortex or elsewhere) could account for the features and perceptual consequences of visual attention. More directly, such studies need to document effects that show the three characteristic features of attention that have been established in psychophysical studies: (1) attention selects relevant information – that is, some sensory signals receive preferential treatment; (2) the processing of selected and unselected information is differentially modulated – that is, attention changes how sensory information is processed; and (3) this selective modulation has perceptual consequences – that is, it changes the weighting or salience of different aspects of the visual environment as a function of their behavioral relevance, creating a representation that enhances some aspects at the expense of others. These three aspects serve as a framework for this article.

Selectivity

Spatial Attention

Directing attention into versus out of the receptive field Early attention research demonstrated that visual attention can be directed to selected locations in the visual field, even covertly (i.e., in the absence of eye movements). Behavioral measurements have shown that stimuli at attended locations are processed faster and more accurate compared to stimuli at unattended locations. Modern imaging studies of attention have measured brain activity during spatial attention tasks and have documented an enhanced activity in retinotopic cortical areas representing the location of the focus of attention.

This spatial allocation of attention has often been likened to a spotlight, suggesting as a neural correlate an enhanced response of a sensory cell and an increased synchrony in the activity of cell populations when spatial attention overlaps a neuron's receptive field. Indeed, studies report modulated responses when attention is directed into the receptive field of neurons from almost every area in visual cortex tested. However, an attentional mechanism that would modulate responses solely depending on whether the 'attentional spotlight' is inside versus outside the receptive field would have very poor spatial resolution beyond striate cortex as receptive fields grow with increasing levels of hierarchy in visual cortex, covering large portions of the visual field in higher areas. Instead, our visual system seems to be able to allocate spatial attention at a finer spatial scale than implied by the large receptive fields in higher areas of visual cortex.

Differential attentional effects inside the receptive field To examine the resolution of the neural mechanisms of spatial attention, several studies have trained monkeys to direct their attention to one of two stimuli inside the receptive field. In most of these studies, one stimulus was matched to the sensory preferences of a cell, whereas the other was not. Responses are generally substantially higher in trials in which the animals are attending to the preferred stimulus. Thus, attention can modulate responses even when it is only switched between two stimuli

that are both inside the receptive field, demonstrating that attentional modulation has a better spatial resolution than the size of the receptive fields in these areas (Figure 1).

For middle temporal area (MT) neurons the neuronal correlate of the attentional selection of one of two stimuli inside a receptive field has been investigated by carefully mapping the receptive fields while the animal was directing its spatial attention to one of the two stimuli inside the receptive field. Under these conditions, the center of the receptive field shifts toward the attended location with a concomitant slight decrease in receptive field size (Figure 2). A similar distortion of receptive field profiles has been observed in area V4 when spatial attention is directed to various positions just outside the receptive field.

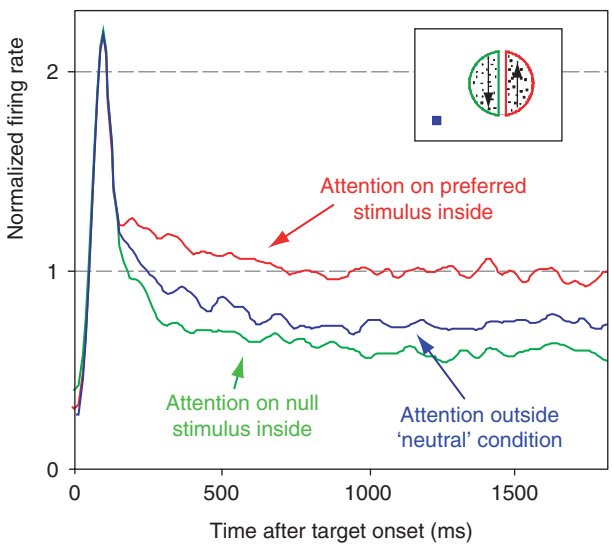

Figure 1 Effect of directing attention to either of two moving stimuli presented inside the receptive field of direction-selective neurons in macaque visual cortex. The inset illustrates the experimental conditions. Two stationary, semicircular apertures containing moving random dots are presented inside the receptive field. One of the dot fields is moving in the preferred direction (red semicircle; upward motion in this example) the other one in the null direction (green semicircle) of the neuron under study. At the beginning of each trial, the animal was instructed to covertly direct attention either to one of the two dot fields or to the fixation point (blue square) while maintaining its gaze on the fixation point. The curves represent firing rates, normalized to the level of sustained activity with attention on the preferred direction semicircle, and averaged across a population of neurons from the middle temporal and the medial superior temporal area. The *x*-axis represents time, with time zero indicating the onset of the visual stimulus. As can be seen, attending to the preferred direction stimulus inside the receptive field increases firing rates (red trace) compared to a condition in which attention is directed to the luminance of the fixation point (blue trace). In contrast, attending to the null-direction stimulus inside the receptive field reduces firing rates (green trace) relative to the fixation point condition. Modified from Treue S (2001) Neural correlates of attention in primate visual cortex. *Trends in Neurosciences* 24(5): 295–300, with permission.

As a consequence of such receptive field shifts, more neuronal processing resources are devoted to attended stimuli. Assuming that neuronal responses provide evidence of the presence of a stimulus at their original receptive field locations, attentional shifts of receptive fields could be the basis of the perceptual overestimation of an attended stimulus' size observed psychophysically.

Although the consequence of the observed shifts in receptive fields is fine-grained changes in visual information processing caused by spatial attention, it should be emphasized that the observed modulation need not be created in the large receptive fields in extrastriate cortex where it has been observed but, rather, could be the consequence of effects on the matrix of small receptive fields in area V1 that together form the large receptive fields in higher areas.

Attention can suppress responses The response modulations that are demonstrated when switching attention from outside to inside the receptive field suggest that directing attention into the receptive field always enhances responses. To investigate this conjecture, a 'neutral' or 'sensory' condition can be used in which attention is directed outside the receptive field, with two (behaviorally irrelevant) stimuli inside the receptive field. Usually, one of the two stimuli is matched to the neuron's preference (e.g., aligned to the preferred orientation), whereas the other is a nonpreferred stimulus. In such a paradigm, the response of the neuron is increased when directing attention toward the preferred stimulus inside the receptive field. However, the response of the neuron will usually drop below the neutral response when attention is directed to the nonpreferred stimulus. Switching attention between two stimuli inside the receptive field combines the suppressive effect of attending to the nonpreferred stimulus with the enhancing effect of attending to the preferred stimulus (Figure 1). This push–pull interaction may be one of the reasons why the attentional modulation is typically stronger in experimental paradigms that juxtapose two stimuli inside the receptive field. As a consequence of such attentional effects, the neural representation of attended stimuli is enhanced at the expense of unattended stimuli.

Nonspatial Attention

Feature-based attention The spotlight metaphor suggests a special role for spatial location as the basis for attentional selection, but several studies have demonstrated nonspatial, feature-based attentional modulation as well. Responses in inferior temporal cortex are enhanced, even before the onset of the preferred shape of a neuron, if its appearance was

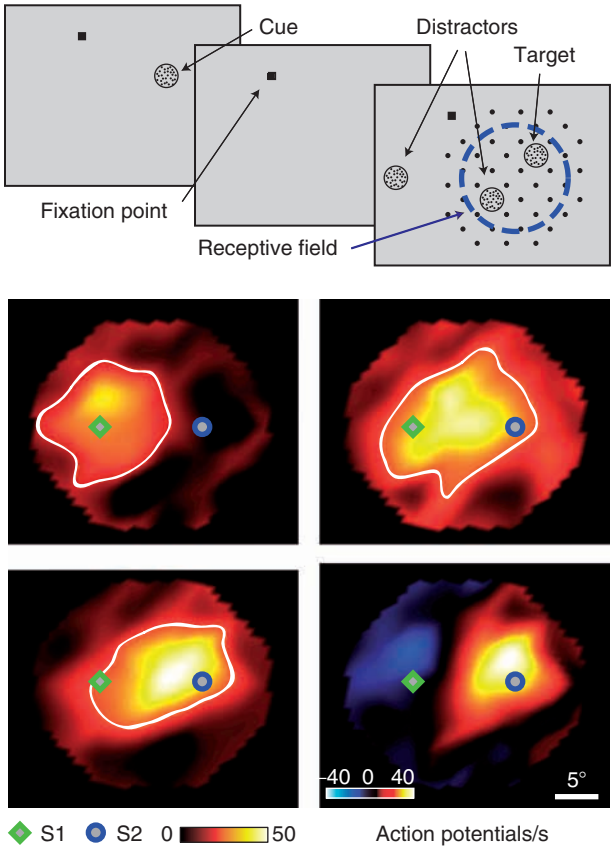

Figure 2 Modulation of receptive field structure by spatial attention. The gray panels at the top represent a computer monitor and illustrate the sequence of events in a single trial of the experiment. After the monkey fixates the small black square, a cue is briefly presented indicating the spatial location of one stimulus (target) that will appear together with two others (distractors). After the cue disappears, three stationary apertures containing moving random dots are presented on the screen, two of which are placed inside the receptive field of the recorded neuron (dashed circle) and the other one is placed outside. While the monkey is attending to the target, trying to detect a subtle change in the direction of motion, an additional small moving pattern is briefly presented at spatially random locations marked with the black dots. Plotting the neuron's response to this pattern as a function of its position describes the spatial structure of the receptive field. The bottom of the figure shows such a receptive field obtained from a single neuron in the middle temporal area. The strength of the response is coded in color, with white indicating the strongest excitation and blue indicating inhibition. The top right panel represents conditions in which attention was directed to the stimulus outside the receptive field. The top left panel and the bottom left panel represent conditions in which attention was directed to the stimulus marked with the green square or to the stimulus marked with the blue circle, respectively. Finally, the bottom right panel represents the difference between the latter two conditions, showing the push–pull effect of spatial attention across the receptive field that underlies the shift of the receptive field center between the conditions. Modified from Treue S, Womelsdorf T, and Martinez-Trujillo JC (2007) Visuelle Aufmerksamkeit: Von Orten, Eigenschaften und Kontrasten. *NeuroForum* 13: 55–60, with permission from NeuroForum.

expected and behaviorally relevant. Similarly, when monkeys are discriminating the orientation of a bar that matches the color of a cue, enhanced responses are observed in area V4 compared to situations in which the color of the bar is different from the color of the cue.

Feature-based attentional modulation that reaches far beyond the confines of the spatial receptive field of a sensory cell has been reported in the dorsal cortical pathway. The activity of MT neurons is higher when the animal is attending to a stimulus moving in their preferred versus nonpreferred direction, even when the attended stimulus is far from the classical receptive field. Thus, attending to a feature, such as a particular direction of motion, enhances the responsiveness of all neurons that prefer this particular stimulus feature, not just those whose receptive field includes the attended stimulus. This global spread of feature-based attention has been confirmed in behavioral and functional brain imaging studies.

Feature-based attentional modulation is comparable in strength to spatial attentional modulation, and the two influences combine additively in appropriate experimental paradigms. The similarity in strength of spatial and nonspatial attention and the additivity of the two modulations have also been observed in the ventral pathway, suggesting a unified system in which the spatial location is just another stimulus feature.

These observations have led to the formulation of the feature similarity gain model of Treue and Martinez-Trujillo. In essence, this model predicts the sign and magnitude of attentional gain modulation based on the similarity between the currently attended feature and the preferences of the neuron under study. Attending to the preferred feature of a given neuron (e.g., its preferred direction) increases neuronal responses, whereas attention to nonpreferred features (e.g., the nonpreferred direction) decreases neuronal responses. Note that the feature similarity gain model also accounts for spatial attentional modulations, interpreting the response reduction when spatial attention is shifted outside the receptive field as a situation of low feature similarity since the attended location is not matched to the preference of the neurons (i.e., its receptive field).

Object-based attention In addition to spatial and feature-based attention, the perceptual grouping of individual stimulus features into visual objects can also strongly influence the allocation of attention. Traditionally, investigations of object-based attention have played a more prominent role in psychophysical and functional magnetic resonance imaging experiments. Typically, two or more visual objects are superimposed such that spatial attention cannot be used to guide the allocation of attention. In such paradigms, features belonging to the same object are generally processed faster and more accurately than features belonging to different objects. Strong evidence for the existence of object-based attentional mechanisms has also been provided by imaging studies comparing activity in functionally defined regions of interest. The visual features of superimposed objects are carefully chosen to differentially activate well-defined cortical regions. Attending to a particular feature can result not only in enhanced activity within the cortical region processing the attended feature but also in enhanced activity within regions processing other, task-irrelevant features of the same object. Therefore, in contrast to feature-based attention, object-based attention is characterized by a transfer of attention from attended to unattended features of the same object. In physiological experiments, it has been difficult to distinguish feature-based attention from object-based attention. A notable exception is a series of experiments in primary visual cortex involving a task in which the animal was trained to mentally trace a curved line while keeping fixation on one of its ending points. Multiunit activity of V1 recording sites overlapping with different segments of the line was simultaneously enhanced if the curve crossing the receptive fields was behaviorally relevant as opposed to when it was not. Such findings indicate that responses of neurons in area V1 depend on whether the stimulus inside their receptive field is part of an attended object.

In summary, different experimental paradigms have revealed apparently distinct forms of attention, depending on whether spatial locations, nonspatial stimulus features, or visual objects are relevant in terms of current behavioral goals. Additional physiological examinations of these phenomena will help to assess the validity of such a strict separation by investigating the basic neuronal mechanisms leading to an enhanced representation of relevant visual information.

Modulation

Multiplicative Modulation of Tuning Curves

The ability of a given neuron in visual cortex to encode a particular stimulus feature is captured by the neuron's tuning curve, the bell-shaped response profile which plots the response of a neuron as a function of some continuous stimulus property, such as orientation or direction of motion. Investigating the neural basis of attentional modulation of information processing therefore requires examining attentional influences on tuning properties. Studies of orientation-tuned cells in area V4 (i.e., the temporal visual pathway) as well as of direction-tuned neurons in MT and medial superior temporal (MST) areas (i.e., the dorsal pathway) have documented that switching spatial attention from outside to inside the receptive field of a given neuron enhances the neuron's responses in a multiplicative manner. Thus, the tuning curves in these two attentional conditions are scaled versions of each other, not differing in their tuning width but only in their amplitude (**Figure 3**). This shows that the allocation of spatial attention into the receptive field of a neuron enhances the neuron's response gain independently of the particular stimulus that drives its response. Moreover, the temporal dynamics of direction tuning in MT neurons show multiplicative modulation by attention in the absence of shifts or changes in shape of the temporal profile. Collectively, these studies show that the highly specific modulation of the response gain of sensory neurons in visual cortex is a central mechanism of attentional influence.

Nonmultiplicative Consequences of Attention

Although both spatial and feature-based attention have been shown to create gain changes (i.e., exert a multiplicative influence on neurons in visual cortex), it is important to note that multiplicative effects can be combined to create nonmultiplicative modulations. One example of such nonmultiplicative modulations is the change in receptive field profiles created

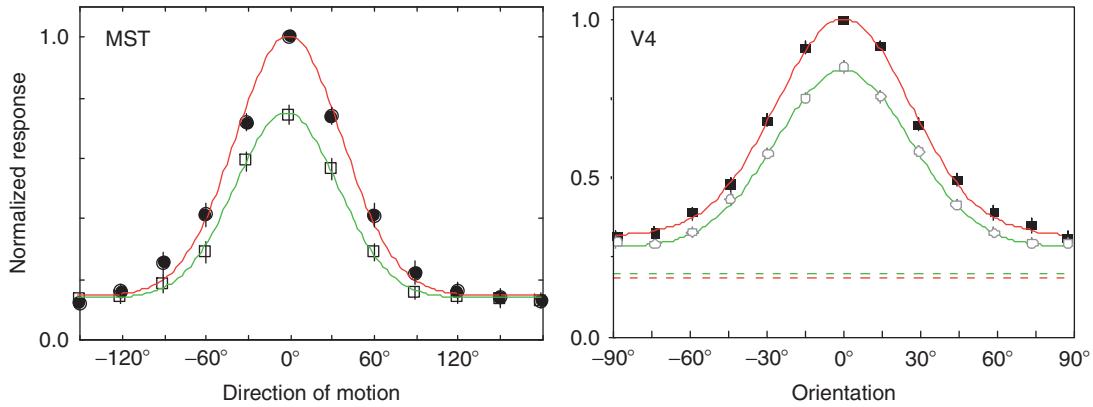

Figure 3 Multiplicative effects of spatial attention on neuronal tuning curves. The red curve in both panels represents the response when attention was directed toward the stimulus inside the receptive field, whereas the green curve gives the response to the same stimulus when attention was directed out of the receptive field. For both pairs, the attentional modulation did not significantly change the width of the tuning curve. The left panel shows firing rates, averaged across a population of neurons from the medial superior temporal area (MST), in response to the direction of a high-contrast moving random dot pattern. The right panel shows the average tuning curve across V4 cells to the orientation of a grating. The dashed lines represent the background firing rates (i.e., the responses of the cells in the absence of a stimulus). Modified from Treue S (2001) Neural correlates of attention in primate visual cortex. *Trends in Neurosciences* 24(5): 295–300, with permission.

when shifting spatial attention within a receptive field or between locations inside and outside the receptive field (Figure 2). Nevertheless, this modulation can also be caused by gain changes, if one assumes that spatial attention differentially enhances and reduces the gain of those V1 neurons providing the input to MT neurons and covering the attended and unattended portions, respectively, of the MT receptive field.

Similarly, switching attention from a fixation point to a stimulus moving in various directions inside an MT receptive field creates a systematic but nonmultiplicative change of the direction-tuning curve between the two attentional conditions (Figure 4). Note that although even purely multiplicative effects of attention have been shown to improve the signal-to-noise ratio of attended stimuli, this selective modulation is even higher in the case of nonmultiplicative modulations based on gain changes. As such, they usually exert a push–pull influence that cannot be achieved directly through a gain change, suppressing sensory responses to unattended stimuli and enhancing responses to attended stimuli.

Saliency

Comparing Response Modulation by Contrast and Attention

The multiplicative modulation of responses by attention is very reminiscent of the modulatory influence of a number of well-studied stimulus parameters on sensory responses, such as contrast, speed, and motion coherence. This similarity suggests a common neural mechanism that would allow attention to modulate the strength of a stimulus representation by changing its 'effective contrast' or saliency. Alternatively, this similarity in terms of effects might simply reflect the existence of two independent multiplicative systems.

A few studies have addressed this issue by testing whether the effects of stimulus contrast interact with the effects of attention. These studies have measured contrast response functions under different attentional conditions. If the two systems were truly independent, the multiplicative effects of attention should be independent of the level of contrast, and the contrast response function should be scaled by a constant factor along the vertical dimension (Figure 5(a)), in the same way as the tuning curves shown in Figure 2. If, on the other hand, these two systems share a common mechanism, directing attention should have the same effect as changing contrast and should therefore become evident in a horizontal shift of the contrast response function between attentional conditions. As a consequence, the strongest attentional effects would then be expected for stimuli of intermediate luminance or contrast (Figure 5(b)).

Recordings from area V4 have revealed a horizontal shift of the contrast response function, consistent with the notion that contrast modulation and attentional effects share a common mechanism. Studies in MT and MST have also reported such a horizontal shift and have documented the strongest attentional effects for intermediate levels of stimulus luminance (i.e., for levels corresponding to the steepest part of the sigmoidal contrast response function). Altogether, the studies further support the notion of a close coupling between these mechanisms. They could thus underlie the attentional effects on perceived stimulus saliency observed in psychophysical studies (Figure 5(c)). Moreover, the

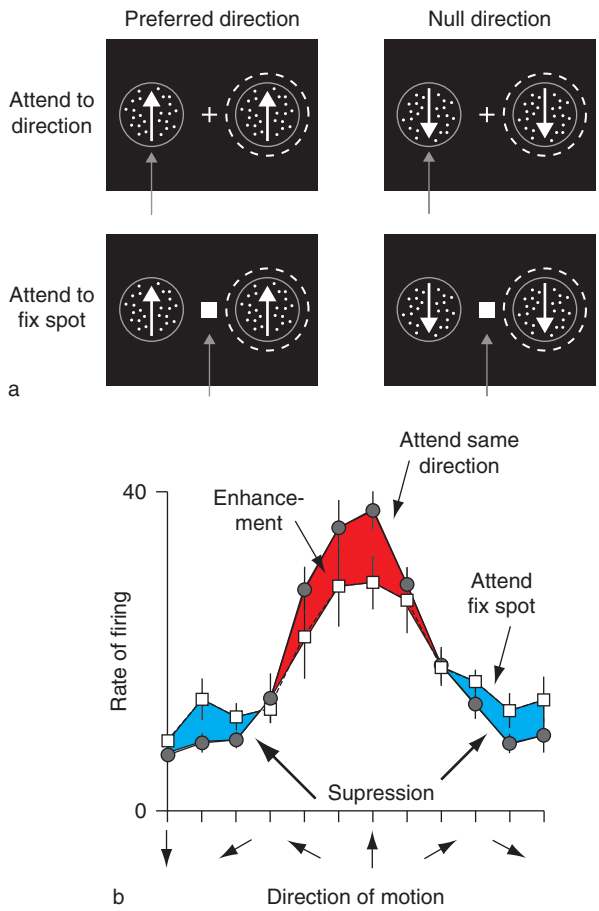

Figure 4 Feature-based attention in the middle temporal area (MT). (a) Schematic representation of tasks used to assess the effects of attention to direction of motion. Two stationary apertures containing moving random dots were presented on the computer monitor. One of them was placed inside the receptive field (dashed white line) of the neuron under study, and the other one was placed outside. The dot patterns always moved in the same direction (white arrows), but different directions of motion were presented on different trials. On some trials (bottom row), the attention of the animal (gray arrows) was directed to the fixation spot to detect a change in luminance. On other trials (top row), a cue at the beginning instructed the animal to pay attention to the motion of the pattern outside the receptive field to detect a subtle change in that motion. (b) Responses of a representative MT neuron to different directions of motion, separately for each attentional condition. Attention to the preferred direction of motion increased the response of the neuron, but attention to the null direction of motion decreased its response. Thus, attention to a particular direction of motion does not increase responses across all neurons. Rather, it has a push–pull effect that increases responses only for neurons that prefer motion close to the attended direction. Modified from Maunsell JHR and Treue S (2006) Feature-based attention in visual cortex. *Trends in Neurosciences* 29(6): 317–322, with permission.

influence of stimulus contrast on the magnitude of the attentional effects might account for the variability in the magnitude of attentional modulation reported in the literature. However, a related study, also recording from area V4, reported a mixture of effects, including multiplicative scaling of the contrast response function by spatial attention, leaving open the possibility of a less direct and more complex relationship between attentional and contrast modulation. Further research is necessary to reconcile these findings.

Forming an Integrated Saliency Map by Spatial and Feature-Based Attention

Our visual environment is represented as the neuronal activity across a population of neurons, with various receptive field positions and preferred feature values. The insights into the neuronal mechanisms of attentional modulation gained from the studies mentioned previously permit inference on the attentional modulation of this population activity, caused by directing attention to a particular feature at a particular location in a complex visual scene. Conceivable could be a scenario in which the visual input creates a fairly homogeneous level of activation across the neuronal population. Spatial attention will increase the gain of all neurons whose receptive field overlaps the current attentional focus, creating an enhanced representation at that location that is akin to a local increase in contrast and that creates a corresponding perceptual effect (Figure 5(c)).

Since similarity in terms of features can also be based on spatial location, the feature similarity hypothesis predicts that the attentional modulation within this retinotopic population should vary considerably. Specifically, neurons whose receptive fields only partially overlap the focus of attention will experience less of a gain increase than those centered on the focus – a prediction that has not been tested experimentally.

In addition to this spatially confined modulation of the retinotopic representation, the contribution of the attended nonspatial feature must also be considered. Feature-based attention will cause a differential gain change across the whole retinotopic representation with a particularly strong gain increase for neurons preferring the attended feature and a corresponding gain decrease for those of opposite preference. The total effect will be a population response that is no longer homogeneous but has its highest activity in the group of neurons preferring the attended location and feature, intermediate enhancements at retinotopic locations where the visual input only matches the attended nonspatial feature (i.e., potential targets in a visual search situation), and suppressed responses everywhere else. Combining such modulated population responses across cortical areas could create an integrated saliency map – that is, a topographic representation of relative stimulus strength and behavioral relevance across visual space.

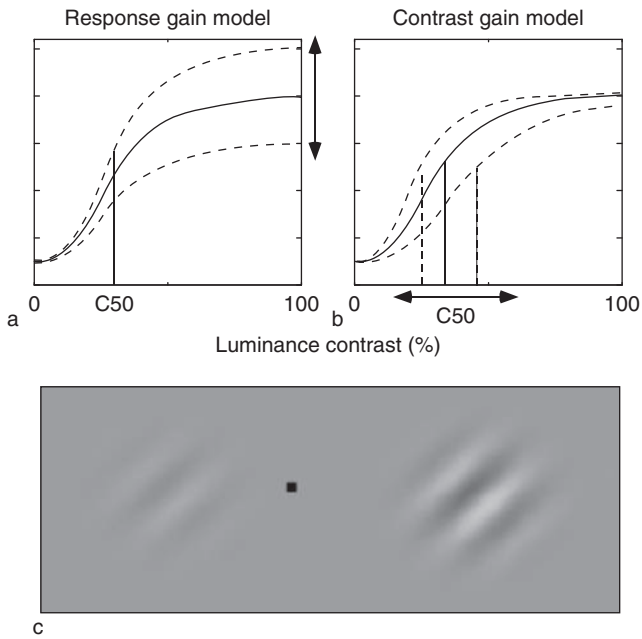

Figure 5 Depiction of the similarity between attentional effects and changes in luminance contrast. (a and b) Possible attentional effects on contrast response functions of visual cortical neurons. According to the response gain model (a), attention modulates the neuronal response to every level of contrast by a constant factor, effectively scaling the contrast response function along the vertical dimension. According to the contrast gain model (b), directing attention is equivalent to changing contrast, resulting in a shift of the contrast response function along the horizontal dimension. A number of studies have provided evidence for the contrast gain model, suggesting that an unattended stimulus of a given contrast will have the same effect as an attended stimulus of lower contrast. An increase in the perceived contrast by attention has been demonstrated in behavioral experiments with human observers and is illustrated in (c). The two stimuli, shown to the left and right of the fixation square, are judged to be equal in contrast if attention is directed to the left stimulus only. Modified from Treue S, Womelsdorf T, and Martinez-Trujillo JC (2007) Visuelle Aufmerksamkeit: Von Orten, Eigenschaften und Kontrasten. *NeuroForum* 13: 55–60, with permission from NeuroForum.

Top-Down and Bottom-Up Response Modulation

The previously outlined attentional modulations of neural activity in visual cortex all result in an enhancement in the activity or the synchrony of cell populations preferring attended stimulus attributes and a simultaneous suppression in the activity of cell populations preferring nonattended attributes. This will create an enhanced representation of attended relative to unattended stimuli, an effect reminiscent of the selective enhancement of particular aspects of the visual input by the hardwired bottom-up coding mechanisms mentioned in the introduction. The similarity of the two systems suggests their combination into the integrated saliency map, in which the visual input is represented not by an individual stimulus' physical strength (e.g., its luminance) but by its saliency (i.e., its difference in features compared to the spatially surrounding visual input). The integrated saliency map combines the bottom-up saliency with the modulatory influence of attention, strengthening or weakening it according to the behavioral relevance of a particular location, feature, or object.

This interaction provides an account for a central function of attention in enhancing the representation of stimuli of intermediate strength while particularly salient stimuli will be well represented even in the absence of attention. This is in good agreement with results from a large body of psychophysical studies demonstrating that highly salient stimuli (known as 'pop-out' stimuli in visual search tasks) need little attentional resource allocation, whereas the same stimuli embedded among similar distractors will only be perceived when attention is directed toward them.

Conclusion

The understanding of attentional influences has advanced considerably beyond the view that attention is a separable influence providing only a late modulation of an otherwise stimulus-driven sensory information processing system. Instead, attention influences the processing of visual information even in the earliest areas of primate visual cortex. This influence seems to shape an integrated saliency map – that is, a representation of the environment that weighs every input by its local feature contrast and its current behavioral relevance. This map enables the visual system to integrate large amounts

of bottom-up and top-down information because it provides an efficient coding scheme for the potentially most relevant information in the sensory input. However, by completely integrating bottom-up sensory information and top-down attentional influences, it equates the absence of attention with low stimulus power. This provides a possible explanation for the observation that highly salient stimuli will be processed even in the absence of attention, whereas low inherent salience will often prevent the perceptual representation of unattended parts of complex natural scenes, although exceptions seem to exist for the categorization of the gist of natural scenes.

The brain areas providing the guidance for the top-down attentional effects seem to be tightly linked to those areas responsible for the planning and execution of eye movements in agreement with the frequent need for foveating salient regions of the visual environment for a more detailed analysis.

Although many questions remain, the rapid development of functional imaging techniques and the development of sophisticated paradigms and recording techniques in monkeys trained to perform complex attentional tasks provide the basis for the recent growth in understanding. By bringing together bottom-up stimulus aspects that are often responsible for automatic attentional allocation and top-down influences reflecting voluntary attention, a global map representing stimulus saliency modulated by the current behavioral state of the organism can provide a unified framework for interpreting future findings on attentional effects and their close integration with sensory information processing.

See also: Attention and Eye Movements; Parietal Cortex and Spatial Attention; Thalamic Mechanisms in Vision; Visual Cortex in Humans; Visual Development.

Further Reading

Anton-Erxleben K, Henrich C, and Treue S (2007) Attention changes perceived size of moving visual patterns. *Journal of Vision* 7(11): 1–9.

Boynton GM (2005) Attention and visual perception. *Current Opinion in Neurobiology* 15(4): 465–469.

Carrasco M, Ling S, and Read S (2004) Attention alters appearance. *Nature Neuroscience* 7(3): 308–313.

Desimone R and Duncan J (1995) Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18: 193–222.

Downing P, Liu J, and Kanwisher N (2001) Testing cognitive models of visual attention with fMRI and MEG. *Neuropsychologia* 39(12): 1329–1342.

Maunsell JHR and Treue S (2006) Feature-based attention in visual cortex. *Trends in Neurosciences* 29(6): 317–322.

Moore T (2006) The neurobiology of visual attention: Finding sources. *Current Opinion in Neurobiology* 16: 159–165.

O'Craven KM, Downing PE, and Kanwisher N (1999) fMRI evidence for objects as the units of attentional selection. *Nature* 401: 584–587.

Reynolds JH and Chelazzi L (2004) Attentional modulation of visual processing. *Annual Review of Neuroscience* 27: 611–647.

Reynolds JH and Desimone R (2003) Interacting roles of attention and visual salience in V4. *Neuron* 37: 853–863.

Roelfsema PR, Lamme VAF, and Spekreijse H (1998) Object-based attention in the primary visual cortex of the macaque monkey. *Nature* 395: 376–381.

Treue S (2001) Neural correlates of attention in primate visual cortex. *Trends in Neurosciences* 24(5): 295–300.

Treue S (2003) Visual attention: The where, what, how and why of saliency. *Current Opinion in Neurobiology* 13: 428–432.

Treue S, Womelsdorf T, and Martinez-Trujillo JC (2007) Visuelle Aufmerksamkeit: Von Orten, Eigenschaften und Kontrasten. *NeuroForum* 13: 55–60.

Williford T and Maunsell JHR (2006) Effects of spatial attention on contrast response functions in macaque area V4. *Journal of Neurophysiology* 96: 40–54.

Womelsdorf T, Anton-Erxleben K, Pieper F, and Treue S (2006) Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nature Neuroscience* 9(9): 1156–1160.

Visual Cortex in Humans

B A Wandell, S O Dumoulin, and A A Brewer, Stanford University, Stanford, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Human visual cortex comprises 4–6 billion neurons that are organized into more than a dozen distinct functional areas. These areas include the gray matter in the occipital lobe and extend into the temporal and parietal lobes. The locations of these areas in the intact human cortex can be identified by measuring visual field maps. The neurons within these areas have a variety of different stimulus response properties. We describe how to measure these visual field maps, their locations, and their overall organization. We then consider how information about patterns, objects, colors, and motion is analyzed and represented in these maps.

Regions of cortex that respond powerfully to retinal stimulation are called visual cortex. In the human brain, the visual cortex is located in the occipital lobe and extends into the temporal and parietal lobes (Figure 1). In many mammalian species, visual cortex spans as much as half of the entire cortical surface area; in humans, the fraction is smaller – approximately 20%.

A key problem in visual neuroscience is to understand how information about patterns, objects, colors, and motion is analyzed and represented within visual cortex. One of the great advances of the past 50 years in our understanding of the processing of visual information has been the discovery of a basic functional architecture. In many species, visual cortex is divided into more than 12 distinct functional areas. The neurons within these areas differ in their stimulus response properties.

New developments in the past 15 years have allowed a multiplicity of distinct areas to be identified using functional magnetic resonance imaging (fMRI) in the intact human brain. The areas within human visual cortex are identified by measuring visual field maps. These are cortical regions that contain a topographic representation of the visual field; stimuli adjacent in the visual field evoke responses that are adjacent in visual cortex. Visual field maps can be identified in individual subjects with fMRI experiments that last approximately 1 h. Measurement of these maps is the most reliable approach to identifying a functional area within human visual cortex.

In this article, we first introduce the reader to the largest and best studied visual field map, primary visual cortex (V1). We describe how this map is measured using fMRI, and we discuss the main features of the V1 map. We then summarize the positions and properties of ten additional visual field maps. This represents our current understanding of human visual field maps, although this remains an active field of investigation, with more maps likely to be discovered. Finally, we describe theories about the functional purpose and organizing principles of these maps.

The Size and Location of Human Visual Cortex

The entirety of human cortex occupies a surface area on the order of 1000 cm^2 and ranges between 2 and 4 mm in thickness. Each cubic millimeter of cortex contains approximately 50 000 neurons so that neocortex in the two hemispheres contain on the order of 30 billion neurons. Human visual cortex includes the entire occipital lobe and extends significantly into the temporal and parietal lobes (Figure 1). Visual cortex contains on the order of 4–6 billion neurons.

The number of neurons in human visual cortex far exceeds the number in many other species that depend on vision. For example, the surface area of macaque monkey visual cortex is probably no more than 20% that of human visual cortex, although the cortical neuronal density is similar. This species difference in the number of cells in visual cortex cannot be explained by how these species encode the visual world. The monkey samples the retinal image at a higher resolution. There are 1.5 million optic nerve fibers from each eye in macaque and only 1 million such fibers in humans. The large size of human visual cortex is likely not a result of an increase in the supply of information but, rather, due to an increase in visual processing and the organization and delivery of information to other parts of cortex, such as those devoted to language and reading. Given these differences in visual cortex size, it would not be surprising that many features of human visual cortex are not present in closely related primate systems.

The Location of Primary Visual Cortex

The largest input to visual cortex projects from the retina via the optic nerve to the lateral geniculate nucleus (LGN) of the thalamus and then to primary visual cortex, commonly called V1. V1 is located in and around the calcarine sulcus, which can be found on the medial surface of both hemispheres (Figure 2(a)). The portions of left and right V1 that respond to signals near the vertical meridian are

connected; the axons mediating these connections pass through the posterior aspect of the corpus callosum (the splenium).

The position of V1 can be identified on histological grounds. There is a stripe of heavy myelinated tissue in calcarine cortex that is apparent to the naked eye. This stripe was discovered by an Italian medical student, Francesco Gennari, in 1782, and the stripe is now called the stria of Gennari. Consequently, V1 is also referred to as the striate cortex, whereas other parts of visual cortex are collectively referred to as extrastriate cortex. Figure 2(b) is a cross section through the calcarine cortex stained to reveal the myelination; note that the white matter is also strongly stained. The myelination runs through one of the cortical layers. The stria of Gennari terminates at the border of V1.

The discovery of the stria of Gennari was one of the first indications that cortex could be differentiated into distinct regions. The association between this stripe and the projection zone of the LGN was not known at first. One hundred years after the discovery of the stria of Gennari, the Swedish neuropathologist Henschen showed that it is coextensive with the LGN projection zone.

Figure 1 Human visual cortex. The extent of human visual cortex is indicated by the purple overlay shown on the medial and lateral views of a right hemisphere. Visual cortex occupies approximately 20% of the surface area of neocortex. Reproduced from Zeki S (2003) Improbable areas in the visual brain. *Trends in Neurosciences* 26(1): 23–26, with permission from Elsevier.

Measuring Visual Field Maps

Anatomical

Until recently, researchers have had almost no access to measuring functional signals in the human brain. The development of fMRI resulted in an enormous outpouring of new data about the human brain. One of the first applications of this technology was to define the architecture of signals in human visual cortex. From extensive work in animal models, particularly macaque, it was known that visual cortex could be partitioned into a set of distinct visual field maps. Hence, an early focus in humans was to define the locations and properties of human visual field maps.

The organization of visual cortex is easily discernible on a flattened representation of the cortical surface. However, in humans, the cortex is highly convoluted, masking this topographic organization. In order to trace the borders of visual field maps, it is useful to create a representation that plainly exposes the signals in the sulci. This process is initiated by acquiring a set of anatomical images that have high contrast between the gray and white matter of the brain (Figure 3(a)). These data are processed to determine the location of the white matter (axon bundles) and the gray matter sheet that surrounds the white

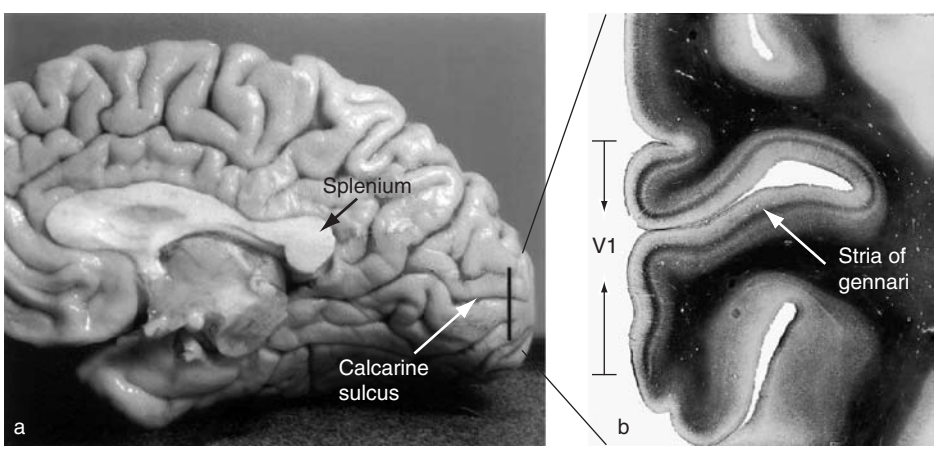

Figure 2 V1 and the stria of Gennari. (a) The location of right calcarine cortex and the splenium of the corpus callosum are shown. Primary visual cortex falls in and around the calcarine sulcus. Portions of right and left primary visual cortex are connected by axons that pass through the splenium. (b) Photograph of a coronal slice through the calcarine sulcus. This slice is stained for myelin (dark regions). Primary visual cortex (V1) is also called striate cortex because it is coextensive with the densely stained stripe that is the stria of Gennari. Reproduced from Andrews TJ, Halpern SD, and Purves D (1997) Correlated size variations in human visual cortex, lateral geniculate nucleus, and optic tract. *Journal of Neuroscience* 17(8): 2859–2868, with permission.

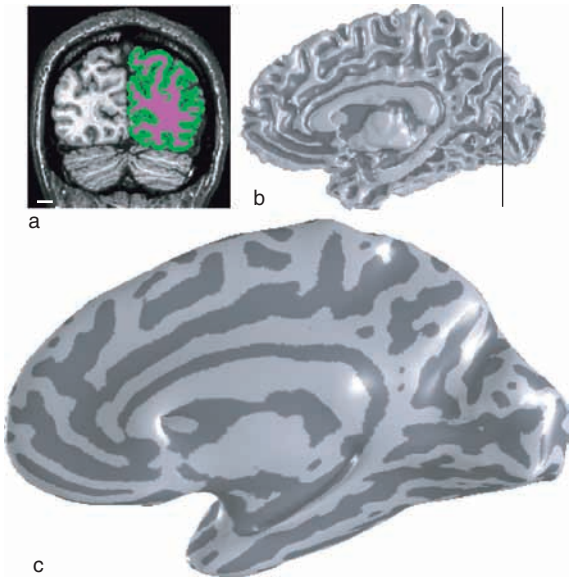

Figure 3 Cortical surface reconstruction and visualization. (a) A coronal slice of an anatomical MRI image is shown. Regions within the right hemisphere are colored to identify white matter (purple) and gray matter (green). (b) The white and gray matter segmentation identified in multiple slices defines a surface that can be reconstructed. (c) The cortical surface is 'inflated' or 'unfolded' to provide a convenient visualization of the cortical sheet within both sulci (dark) and gyri (light).

matter. Although there are several automated methods that achieve a good first approximation for this segmentation, careful work usually requires hand editing.

Once the gray matter voxels are identified, many automated tools are available to create a three-dimensional rendering of the boundary between the gray matter and the white matter (Figure 3(b)). Smoothing algorithms create a new 'inflated' representation of the surface (Figure 3(c)). In this format, the sulci are exposed so activation patterns can be tracked along the gyri (shown as light shading) and within sulci (shown as dark shading).

Functional

Visual field maps represent visual space in a unified, organized arrangement. These maps are often called retinotopic maps because their organization reflects the organization of the retinal image. Visual field maps can be identified by measuring which stimulus position in visual space produces the largest response in a particular section of visual cortex. There are many reasonable ways to perform this experiment, but the most widely used method is to divide the measurements into two parts: measurements of eccentricity and angle with respect to the point of fixation. In one experimental session, one measures cortical responses

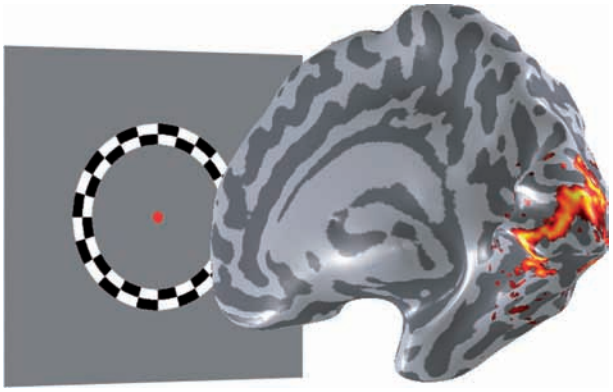

Figure 4 Traveling wave measurements using contrast rings. The functional activation elicited by a contrast ring surrounding fixation (red dot) is shown on an inflated representation of cortex. The peak response shifts smoothly from posterior to anterior calcarine as the ring grows from a small central ring to a large peripheral ring.

to a series of contrast rings, which measures the eccentricity component of the map (Figure 4); in a second session, one measures with a series of contrast wedges at different polar angles, which measures the angular component of the map.

This approach produces strong responses in human visual cortex. For example, a set of contrast rings of increasing size produce peak fMRI signals at a series of locations from posterior to anterior calcarine cortex (Figure 4). The peak response shifts smoothly from posterior to anterior calcarine as the ring grows from a small central ring to a large peripheral ring.

By analyzing the timing of the peaks and troughs in the data, one can identify the visual field eccentricity that most powerfully stimulates each location in visual cortex. The information is commonly color coded to show the eccentricity that most effectively drives the response at each cortical location (Figure 5(b)). In a second experiment, using the rotating wedge stimulus, one can identify the angular direction that most effectively stimulates each cortical location (Figure 5(c)). Taken together, these two measurements define a visual field map, expressed in polar coordinates of eccentricity and angle. These measurements are combined into a single visual field map. The first three maps, V1–V3, are shown in Figure 5(d).

The Human V1 Map

There are a few important characteristics to note about the V1 measurements(Figure 6). First, in the right calcarine sulcus, the most effective wedges are always located in the left part of the visual field. This confirms a well-known anatomical projection: signals from the left visual field project to V1 in the right hemisphere. Second, stimuli in the upper visual field are

most effective for neurons on the lower bank of the calcarine (lingual gyrus), whereas stimuli in the lower visual field produce the greatest response on the upper bank of the calcarine (cuneus). Stimuli near the horizontal midline are most effective at stimulating cortex near the depth of the calcarine sulcus. Thus, the V1 map is inverted with respect to the visual field but consistent with the retinal image. Finally, the central portion of the visual field stimulates posterior calcarine, whereas the more peripheral visual field stimulates anterior calcarine.

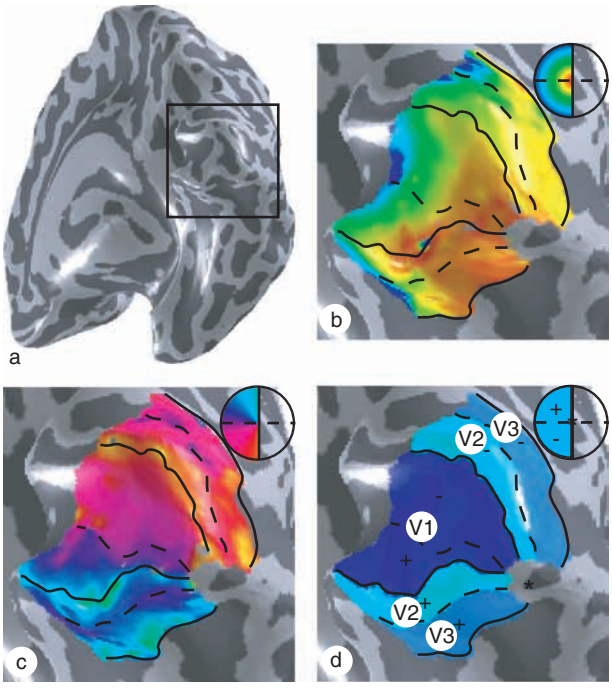

Figure 5 Polar angle and eccentricity measurements define visual field maps on the cortical surface. (a) The cortical surface of the right hemisphere is displayed from a ventroposterior view, emphasizing the occipital lobe. The position maps are displayed on the enlarged part of the occipital lobe, as indicated by the large black square. The maps for eccentricity (b) and polar angle (c) are shown. The insets indicate the color code that defines the part of the visual field that most effectively stimulates each cortical location. Visual field maps V1–V3 are identified from these eccentricity and polar angle measurements (d).

Note that that there is considerably more cortical territory in V1 responding preferentially to stimuli within 2° or 3° of visual angle than responding preferentially to stimuli in the periphery. The visual field map in V1 is a distorted version of the visual field; this magnified foreal representation is called cortical magnification. The term is unfortunate because the magnification is not introduced in cortex. Rather, the magnification begins at the level of the foveal cone photoreceptors in the retina, which are smaller and more tightly packed than peripheral cones. The number of retinal ganglion cells carrying information to the LGN and the number of V1 neurons transmitting the cone signals mirror the uneven cone density. Since the density of cortical neurons is approximately uniform, the differential number of neurons needed to process foveal and peripheral signals is manifest by an expanded surface area for foveal compared to peripheral representations in V1.

Extrastriate Visual Field Maps

The orderly representation of eccentricity and angular responses continues beyond the V1 map. The eccentricity representation in the surrounding cortex parallels that in V1 (Figure 5(b)). The polar angle representation in surrounding cortex is regular, but it reverses direction. At the boundary of V1, the angular representation turns from the vertical meridian of the upper or lower visual field and progresses toward the horizontal representation (Figure 5(c)).

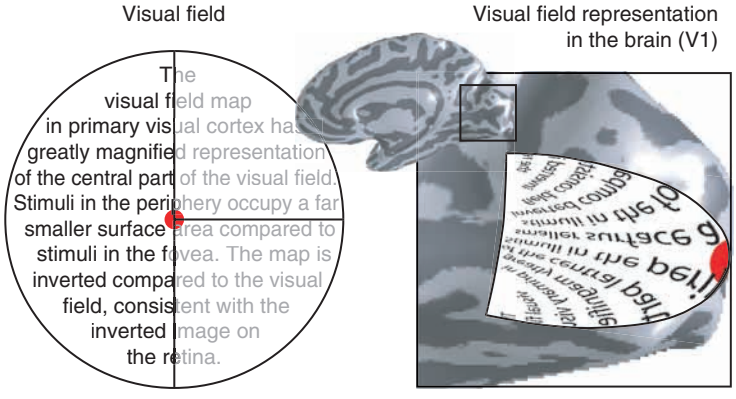

Figure 6 Schematic representation of the V1 visual field map. When one looks at the fixation point (red dot), the left part of the visual field evokes a response in right V1. Both the visual field and the V1 representation of this visual field are schematically illustrated. The V1 map preserves visual field topography, but relatively more cortical area is devoted to the central part of the visual field; this is called cortical magnification. Also, note that the lower visual field evokes activity on the upper bank of the calcarine, whereas the stimulus in the upper visual field evokes activity on the lower bank.

The angular map reversal line is the functional signal used to define the boundary between the V1 and V2 maps.

The V2 map surrounds V1 (Figures 5(d) and 7). One section is located on the cuneus; this section is most effectively driven by signals in the lower visual field. The other section of V2 falls on the lingual gyrus; this section is most effectively driven by signals in the upper visual field. The V2 map is surrounded by a third map, V3. At the V2/V3 boundary, the angular representation reaches the horizontal and then reverses again into V3.

Careful mapping reveals the presence of at least 10 visual field maps (Figure 7). In many cases, these visual field maps are small maps which contain foveal representations distinct from the foveal confluence of the V1–V3 maps. The maps shown here are largely agreed upon, although there is some debate concerning the specific organization of certain maps. It can be difficult to measure these maps for several reasons. Some of the maps are much smaller than the V1–V3 maps and thus are at the spatial resolution limit of conventional fMRI. Also, the size of neuronal receptive fields can be quite large in many of these areas, which makes measurements of the internal structure of these maps more difficult. As resolution and experimental methods improve, new maps will be identified, and our understanding of the known maps will be refined.

The presence of so many distinct visual field maps suggests that the map itself serves a functional purpose. The spatial organization of the maps may facilitate certain types of cortical computations. The internal structure of the visual field maps may simplify the connectivity needed for spatially localized visual computations between neurons specialized for carrying such information as stimulus orientation, color, or motion. The simplification of axonal wiring is an

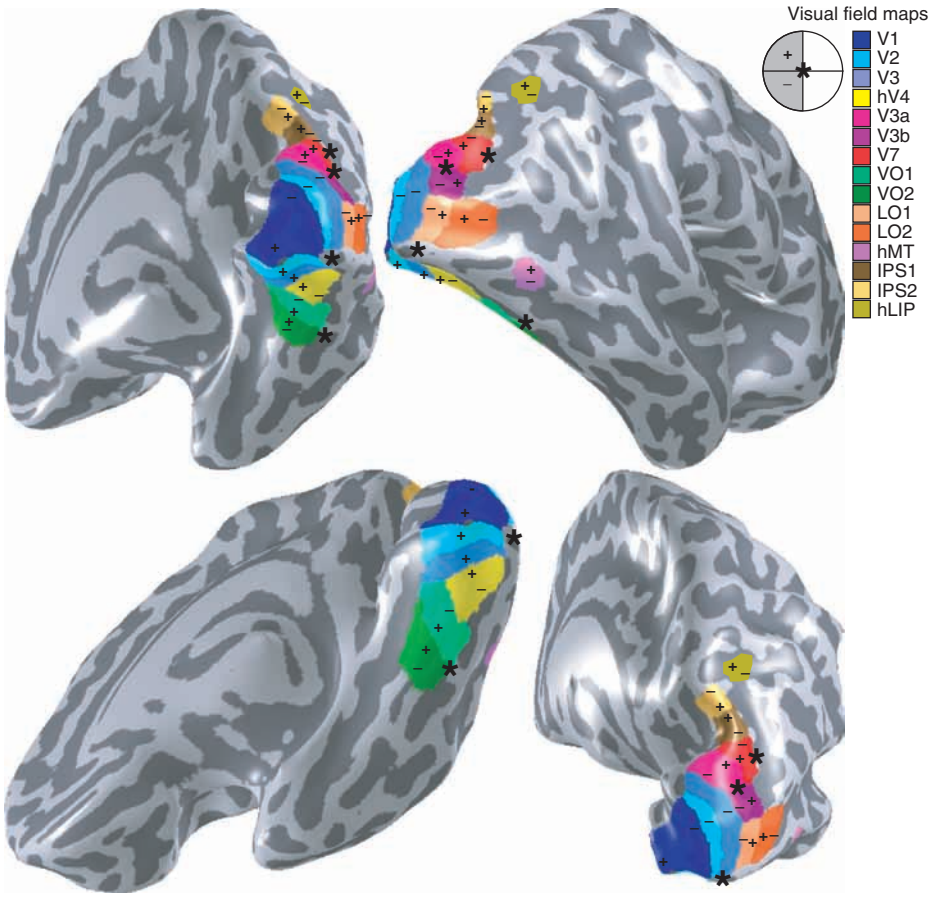

Figure 7 Visual field maps on a cortical surface. Currently identified visual field maps are shown on several views of a right hemisphere (see color legend). Fovea and upper/lower visual fields are indicated by the asterisk, plus, and minus symbols, respectively. The first visual maps are termed V for visual and a number, such as V1, V2, V3, and V3A, following the naming of apparent homologs in monkey. Doubts about the homology of the fourth visual map between monkey and human caused this map to be identified as hV4 (for human V4), and similarly for hLIP, a putative homolog to monkey lateral intraparietal sulcus. Historically, V5 is also referred to as medial temporal. Maps in lateral occipital cortex are numbered as LO-*x*, maps in ventral occipital are numbered as VO-*x*, and maps in the intraparietal sulcus are numbered IPS-*x*.

important aspect of visual field map organization, given the remarkable observation that the axons in 1 mm^3 of cortex would extend over a length of more than 3 km.

Visual Field Maps and Perception

How information is processed within visual field maps and the significance of this processing in terms of visual perception are central questions in visual neuroscience. The presence of multiple maps supports a view of visual cortex based on functional specializations. This view contrasts with the classic null hypothesis that cortex is equipotential or the theory that there is a canonical cortical circuit.

The dominant thinking for the past 30 years assumes that each extrastriate map represents a computational specialization. Furthermore, it is frequently asserted that these computations are associated with specific perceptual experiences. For example, a map on the lateral surface (V5/MT) is often described as a 'motion center,' and a set of maps on the ventral surface are often described as a 'color center.' The principle of associating maps with a perceptual experience is a form of the cortical specialization hypothesis. This framework has provided a scheme for scientists to discuss the significance of these maps, although this framework is far short of a computational theory.

The basis for the relationship between functional specialization and perceptual experience originated in neurological findings: damage to specific extrastriate regions in ventral cortex leads to specific visual disabilities. Four cortical regions in which damage can produce a specific perceptual deficit are illustrated in Figure 8. The neurological deficits illustrated here are inability to recognize faces and interpret facial expressions (prosopagnosia or face blindness), loss of color vision of cerebral origin (cerebral achromatopsia), loss of motion perception (akinetopsia), and loss of the ability to read whole words (alexia). The neurological literature has focused on acquired forms of these deficits, but prosopagnosia also exists in a congenital (developmental) form; it is possible that the other dysfunctions may be discovered in developmental form as well.

The specificity of localized brain damage and perceptual deficits supports the functional specialization hypothesis for extrastriate maps. In contrast, the behavioral consequences of V1/V2 lesions are not usually functionally specialized. Damage to small (<5 mm) regions of V1 degrades vision generally within the corresponding portion of the visual field. This suggests that the visual signals necessary for motion, color, recognition, and so forth are intermixed in neural signals in V1. Extrastriate regions, either innately or through learning, are specialized to interpret visual information needed in certain tasks. Hence, the architecture of visual cortex supports a mixed model in which certain cortical zones support a mixture of functions, whereas others are highly specialized.

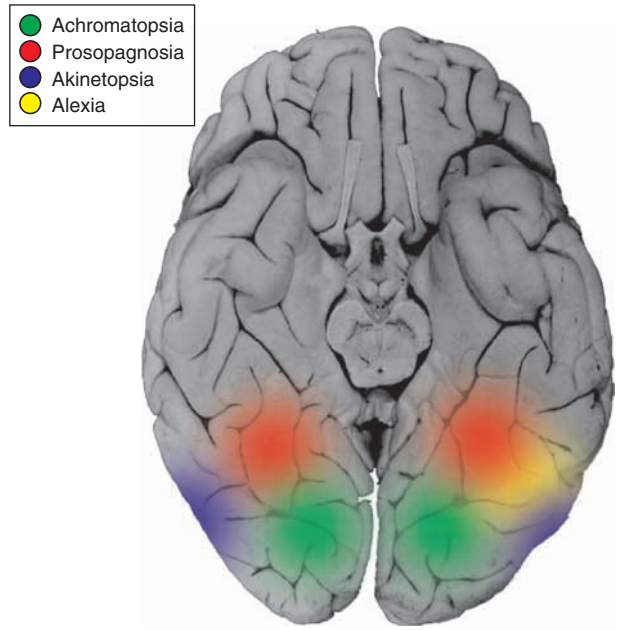

Figure 8 Regions of ventral visual cortex associated with specific neurological deficits. Regions implicated in specific visual impairments are indicated: the inability to recognize faces (prosopagnosia; red), to detect motion (akinetopsia; blue), to read words (alexia), and color blindness (achromatiopsia; green). fMRI measurements aimed at isolating these functions typically yield strong activations at these corresponding cortical regions.

Visual Field Map Clusters

Although the relationship between visual field maps and perceptual experience is an important scale to understand, the association between a map and perceptual experience does not provide a full model of visual processing, and it is unlikely that functional segregation by maps constitutes the only perceptual principle in visual cortex. For example, the dominant model of visual processing in retina, LGN, and V1 consists of several specialized and interleaved neural pathways within each structure. Starting in the retina, there are multiple parallel pathways specialized for conveying specific information about different aspects of the visual signal. Each pathway contains neurons that form a sampling mosaic covering the entire visual field. The neurons within these pathways are located near one another in retina and project to adjacent locations in the LGN and then V1. The retina, LGN, and V1 can thus each be thought of as containing multiple, functionally distinct, superimposed visual field submaps.

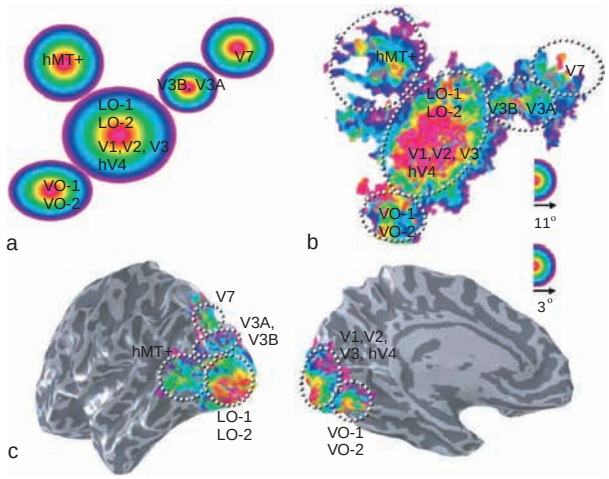

Figure 9 Visual field map clusters. (a) Schematic diagram of the organization of eccentricity representations in visual cortex. This diagram is based on measurements of these visual field map clusters, as shown on flattened cortex (b) and on a three-dimensional rendering of the inflated cortical surface (c). Each cluster contains several visual field maps that can be delineated based on the polar angle maps. VO, ventral–occipital; LO, lateral–occipital.

This architecture is not likely to be completely abandoned in extrastriate cortex. In fact, there are already examples in which a single visual map contains distinct neural circuits that measure different aspects of the visual signal. An early and important example of functional heterogeneity can be found in motion-selective cortex (V5/MT), in which there is an orderly representation of both motion direction and binocular depth.

Just as functional specializations exist at a resolution finer than the map, specializations may also exist at a coarser scale. It was previously noted that the V1–V3 maps share a parallel eccentricity representation and are functionally divided by reversals in the angular representation. This same map architecture is present in several other clusters of maps, located in the ventral, dorsal, and lateral regions of visual cortex (Figure 9). Such organization of the individual maps into visual field map clusters may reflect a coarse-level organization that follows specific functional capabilities; visual field maps may be arranged in clusters when the neural mosaics within these maps serve similar computational goals.

Conclusions

The visual system is a valuable model system for neuroscience research because of the extensive information we have about the retinal encoding of light, the multiple pathways that communicate retinal information to cortex, and the distribution of the information from these pathways into maps in visual cortex. This information makes the visual system an excellent target for study beyond the core of vision science, including the neuroscience of development, plasticity, and consciousness.

The organization of visual pathways, maps, and clusters also may have application as an architectural model of general cortical processing. The parallels for other sensory systems are obvious and in many ways appear to be valid. For example, several distinct auditory maps with tonotopic organization have been demonstrated. It also may be that cognitive functions, such as decision making or the distribution of affective functions, are organized using a mixed architectural model in which certain regions contain general and others more specialized processing. To the extent that the cortical architecture of visual cortex matches that in other parts of cortex, we become more confident in results from both fields. Functional neuroimaging, which operates at a relatively coarse resolution of several millimeters to multiple centimeters, is well suited to uncovering structures at the scale of these visual maps or groups of maps.

In the near future, there will be many opportunities to link the functional neuroimaging measurements in quantitative ways to measurements at other resolutions – down to small (a few thousand) populations of cortical neurons. Advances in functional neuroimaging methods have already provided new information about the white matter pathways and long-distance connections between cortical zones. There will be substantial improvements in measurements of the molecular constituents of these cortical regions and in tracking changes in the molecular composition throughout development or following damage. Our current understanding of the architecture of visual cortex makes it a good system for developing these new methods and learning about cortical function broadly.

See also: fMRI: BOLD Contrast; Vision: Surface Segmentation; Vision: Light and Dark Adaptation; Visual System: Functional Architecture of Area V2; Visual System: Multiple Visual Areas in Monkeys; Visual Attention; Visual Cortical Models of Orientation Tuning.

Further Reading

Andrews TJ, Halpern SD, and Purves D (1997) Correlated size variations in human visual cortex, lateral geniculate nucleus, and optic tract. *Journal of Neuroscience* 17(8): 2859–2868.

Hubel DH (1988) *Eye, Brain and Vision*. New York: Freeman.

Huettel S, Song A, and McCarthy G (eds.) (2004) *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer.

Wandell BA (1995) *Foundations of Vision*. Sunderland, MA: Sinauer.

Zeki S (1993) *A Vision of the Brain*. London: Blackwell.

Zeki S (2003) Improbable areas in the visual brain. *Trends in Neurosciences* 26(1): 23–26.

Visual Cortex: Mapping of Functional Architecture Using Optical Imaging

D Sharon, Stanford University, Stanford, CA, USA
A Grinvald, Weizmann Institute of Science, Rehovot, Israel

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Neurons with similar stimulus response properties are clustered together in vertical columns in cortex, giving rise to what has come to be known as functional architecture – the mapping of various stimulus feature dimensions onto the cortical sheet. At the level of cortical columns ranging from tens of micrometers to a few millimeters, no other technique offers the combination of high spatial resolution and noninvasiveness as optical imaging. This article reviews the properties of optical imaging, the sources of intrinsic signals, and the studies conducted using it, often combined with other techniques, to reveal functional architecture of visual cortex in striking detail.

Properties of Optical Imaging of Intrinsic Signals

Optical imaging of intrinsic signals has a unique set of properties that has made it the method of choice in research of visual cortex where the goal is to study functional architecture of several millimeters of cortex at a resolution corresponding to cortical columns, often over long periods of time and for several visual features.

Resolution

The spatial resolution of optical imaging of intrinsic signals depends on the spatial scale at which hemodynamic signals are regulated in response to neural activation and on the sampling frequency and signal-to-noise ratio of the data. In practice, this has resulted in spatial resolution as good as 20 μm and typically 50 μm. This is ample resolution to image cortical columns, which are typically a few hundred micrometers across.

The temporal resolution of optical imaging of intrinsic signals is constrained by the slow hemodynamic recruitment in response to neural activation. Hemodynamic activation typically follows neural activation by 300–500 ms and lasts for more than 10 s for even very short neural responses. Thus, the temporal envelope of the optical signal is much wider than that of the neural response, and optical imaging of intrinsic signals is thus not considered to allow visualization of neural dynamics. One could attempt to alleviate this shortcoming by deconvolving the intrinsic signal with a hemodynamic impulse function as suggested for functional magnetic resonance imaging (fMRI); however, this has not been attempted for optical imaging yet, and it would strongly rely on the assumed hemodynamic response function. In practice, when neural dynamics need to be imaged at a spatial resolution of a few tens of micrometers, optical imaging using voltage-sensitive dyes is employed. However, for studies of functional architecture, where static maps are explored or maps change at a scale of days or minutes at most, the temporal resolution offered by optical imaging of intrinsic signals is more than adequate.

It should be noted that the light scattering signal is both faster and more tightly co-localized with neural activity than the hemodynamic signal; however, it is weaker in most preparations.

Mapping of Multiple Visual Features

Another important mapping method that has been extensively used in visual cortex, and which has similar spatial resolution to that of optical imaging of intrinsic signals, is the 2-deoxyglucose technique, in which a radioactive sugar analog is uptaken and trapped within active cells, which can then be imaged postmortem. This technique can be used to map responses to up to two stimulus conditions. In contrast, optical imaging of intrinsic signals is performed *in vivo* without invasive extrinsic probes and can therefore be performed repeatedly and, theoretically, to any number of stimulus conditions. This allows mapping of multiple visual features, such as orientation and ocular dominance within the same cortex, and has resulted in detailed studies of the spatial relations between functional maps of various visual attributes.

Chronic Imaging: Development

The relative noninvasiveness of optical imaging of intrinsic signals allows chronic implantation of the imaging chamber for repeated imaging sessions within the same animal and over extended periods of time. An important consequence is that it is ideal for studies of the development of the visual system, in which young animals undergo multiple imaging sessions. The changes in functional architecture between different developmental stages can therefore be followed from before eye opening in species born with eyes closed until adulthood.

Chronic Imaging: Awake Preparation

Chronic imaging also enables studies of functional architecture in awake animals, especially monkeys. This is important for two reasons. First, one wants to verify that the functional architecture revealed in anesthetized animals is an intact representation of the properties of cortex in the awake animal. Second, using unanesthetized monkeys, it is possible to study higher level visual functions that are more cognitive and require the animal's behavior. Optical imaging of intrinsic signals has been repeatedly performed in monkeys for periods of more than 1 year. The orientation and ocular dominance maps obtained throughout this period are stable and constant.

Combination with Other Methods

Optical imaging of intrinsic signals naturally lends itself to combination with other neurobiological techniques because it offers a large exposed patch of several millimeters of cortex within a sealed chamber. Electrodes can straightforwardly be introduced into the preparation, allowing electrophysiological recording targeted to specific loci of interest (e.g., the center of an orientation pinwheel) identified during an imaging session. Anatomical tracers can also be injected to study the structural relationship of horizontal connections to a feature map of interest.

These properties of optical imaging of intrinsic signals have been exploited in extensive studies of the visual cortex, which are reviewed later. In order to understand these results, one also needs an understanding of the underlying sources of these signals, which are discussed next.

Sources of Intrinsic Signals

The intrinsic signal has different components, originating from different sources, as detailed here; however, functional maps obtained at different wavelengths are very similar. Therefore, all of these components can be used for functional mapping. Using optical imaging, imaging spectroscopy, oxygen-dependent phosphorescence quenching, and fluorescent blood plasma tracing, the following changes appear to give rise to different components of the intrinsic signal:

- activity-dependent oxygen consumption affecting hemoglobin saturation;
- changes in blood volume affecting the tissue light absorption;
- changes in blood volume affecting the tissue light scattering;
- changes in blood flow affecting hemoglobin saturation;
- activity-dependent changes in tissue light scattering.

The first component of the intrinsic signal originates from activity-dependent changes in the oxygen saturation level of hemoglobin. This change in oxygenation contains two different components. The onset of the first component is an early one: an increase in the deoxyhemoglobin concentration, resulting from elevated oxygen consumption of the neurons due to their metabolic activity. This causes a darkening of the cortex due to increased absorption by deoxyhemoglobin molecules. This early phase was referred to as the 'initial dip' in the fMRI and related optical imaging literature. The second component of the intrinsic signal originates from an activity-dependent increase in cortical blood volume, increasing the amount of light absorbed by activated cortical regions in a wavelength-dependent fashion, following the absorption spectrum of oxyhemoglobin. Co-localization of blood volume changes with the electrically active columns is much weaker than for the deoxyhemoglobin increase during the initial dip. This poorer co-localization is revealed by a lower signal-to-noise ratio in spatial maps of activation calculated from the cortical blood volume component relative to those from the initial dip. This is especially true for maps in which the response to a single stimulus condition is normalized by the response to a blank ('single-condition' maps). For 'differential' imaging, where the response to a stimulus condition is normalized by or subtracted from the response to an 'orthogonal' stimulus condition, high-quality maps have been obtained, although they are still somewhat poorer than those from the initial dip. The differential procedure removes much of the blood volume noise because it nearly eliminates the contribution of the nonstimulus-specific part of the vascular response.

The third component is a delayed activity-related increase in blood flow, causing a decrease in the deoxyhemoglobin concentration. This happens because the blood rushing into the activated tissue contains higher levels of oxyhemoglobin. Another signal component originates from changes in blood volume, which are probably due to local capillary recruitment or a rapid filling of capillaries and dilation of venules in an area containing electrically active neurons. These blood-related components dominate the signal at wavelengths between 400 and 630 nm. However, blood volume changes in capillaries compress the cortical tissue, giving rise to a change in light scattering that exists at all wavelengths and exhibits a time course typical of blood volume changes rather than that of light scattering changes that originated from electrical activity *per se*. This last significant component of the intrinsic signal arises from changes in light scattering that accompany cortical activation. These changes are caused by ion and water movement, expansion and contraction of

extracellular spaces, capillary expansion, or neurotransmitter release. Light scattering components become a significant source of intrinsic signals above 630 nm and probably dominate the intrinsic signals in the near infrared region above 800 nm.

Functional Architecture of Visual Cortex

Optical imaging of intrinsic signals has been extensively employed in visual cortex of monkey, cat, ferret, tree shrew, rat, and mice to explore functional architecture at high spatial resolution. Most of these studies used the increase in deoxyhemoglobin concentration occurring early in the hemodynamic response due to elevated oxygen consumption by active neurons. This signal is imaged as darkening of the cortex, especially at a wavelength of 605 nm. Imaging at longer wavelengths, emphasizing the light scattering component, has also proven useful. Many of these studies combined optical imaging with electrophysiological or anatomical tracing methods to either confirm the optical imaging results or characterize the relationship between physiological or structural properties of neurons and the functional maps within which they are embedded.

Orientation Map

The orientation selectivity of visual cortical neurons has been well-known since the pioneering studies of David Hubel and Torsten Wiesel in cat and monkey in the 1960s and 1970s. These studies also revealed that in visual cortex, similar to somatosensory cortex as discovered by Vernon Mountcastle, neurons are arranged in vertical columns such that the orientation preference of all neurons in a column is similar. This led to the idea of a map of orientation preference across visual cortex, and Hubel and Wiesel suggested a smooth and orderly shift of preferred orientation as one moves along the cortical surface. However, a direct method to visualize this map was not available at the time. With the advent of optical imaging of intrinsic signals, the technique was first applied in visual cortex to study the orientation map in cat area 18. Direct visualization of the map revealed a different pattern than expected: rather than elongated bands of slowly changing orientation preference, iso-orientation domains are highly ordered patches organized in pinwheels around centers where orientation preference changes quickly (Figure 1). In ferret visual area 17, the two cardinal axes (horizontal and vertical orientations) are somewhat overrepresented, in terms of their area in the orientation map, than the obliques. This result mirrors the behavioral advantage that the cardinal axes have over obliques in some tasks, such as orientation discrimination.

The pinwheels and pinwheel centers are a prominent architectural feature of areas 17 and 18 in cat and V1 in monkey. To study the properties of neurons in pinwheel centers, extracellular recording using tetrodes, which improve the ability to record from single neurons, was targeted to centers imaged with optical imaging. Neurons near pinwheel centers have very different preferred orientations within a distance of a few tens of micrometers. However, their spiking responses are just as orientation selective as those of neurons in orientation domains distant from pinwheel centers (Figure 2). In contrast, the subthreshold input to neurons reflects their neighboring environment – the orientation tuning of the membrane potential response, measured with intracellular electrodes, of pinwheel cells is much broader than that of orientation domains (Figures 2(b) and 2(c)).

Direction Selectivity

Neurons often have a preferred direction of motion that is orthogonal to their preferred orientation. Cells in early visual cortex are clustered according to this property, allowing its mapping using optical imaging. Direction preference shifts, mostly smoothly, along the cortical surface. The discontinuities that occur in the direction map are different from those in the orientation map. Rather than pointlike pinwheel centers, the discontinuities here are fracture lines, and they subdivide iso-orientation domains into areas preferring the two opposite directions orthogonal to the preferred orientation. Direction fracture lines run between orientation pinwheel centers (Figure 3).

Spatial Frequency

Another visual feature for which early cortical visual neurons have long been suggested to be clustered for is spatial frequency. Whereas earlier reports from cat area 17 indicated that the spatial frequency map is binary (i.e., includes regions of high and low preferred spatial frequency that can be mixed in different proportions), other work suggests that continuous values of preferred spatial frequency are mapped, ranging from approximately 0.1 to 1.2 cycles per degree.

Spatial frequency mapping is often used in optical imaging experiments to image the border between area 17 and area 18, exploiting the well-known difference between population preferred spatial frequency in the two areas. By stimulating with oriented gratings at two frequencies, typically 0.2 and 0.6 cycles per degree, and imaging the differential activity, one can quickly and easily detect the 17/18 border as a band of <1 mm across which preference shifts from predominantly higher preferred spatial frequency to predominantly lower preferred spatial frequency.

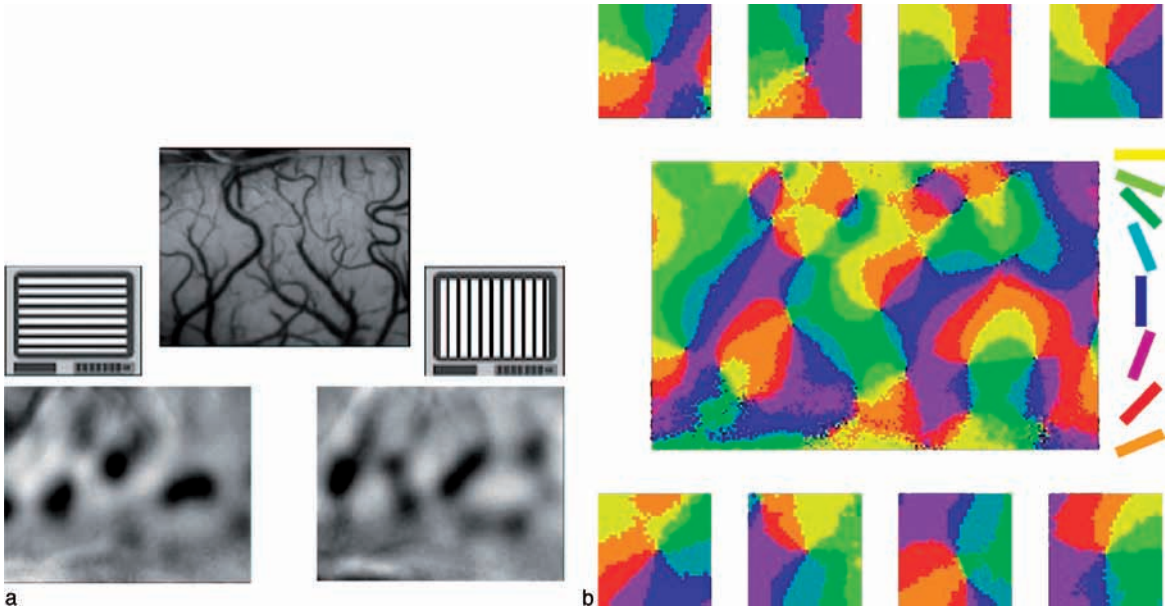

Figure 1 Orientation mapping – cat areas 17 and 18. (a) An activity map for one orientation is obtained straightforwardly by dividing the image captured during presentation of this orientation by the average of the images captured during presentation of all orientations. (Top) The image of the cortical surface is illuminated with green light to emphasize the vasculature. (Bottom) Activity maps evoked by visual stimulation with horizontal and vertical gratings. (b) The full orientation map presents the preferred orientation for each imaged pixel using a color code (center right). Preferred orientation is calculated from a set of at least four single-condition maps (a) by vector averaging. The orientation map is organized in pinwheels, where iso-orientation domains of all orientation preferences meet. Adapted from Bonhoeffer T and Grinvald A (1991) Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* 353: 429–431.

Relationships between Maps

Multiple features are mapped in early visual cortex, including orientation, direction, spatial frequency, ocular dominance, retinotopy, and cytochrome oxidase blobs. Since these maps are superimposed on a two-dimensional cortical surface, an important question arises as to the geometrical relationships between them. These relationships should presumably ensure full coverage of all possible combinations of the multiple visual features so that functional blind spots for particular feature combinations are avoided. Indeed, in monkey V1 the following detailed relationships have been found using optical imaging and cytochrome oxidase staining:

1. Orientation preference is organized radially.
2. Orientation domains are continuous and have fuzzy boundaries.
3. Iso-orientation lines tend to cross ocular dominance borders at 90°.
4. Orientation pinwheels are centered on ocular dominance columns.
5. Blobs are centered on ocular dominance columns.
6. The centers of the blobs and the centers of the orientation pinwheels are segregated.
7. There is a regular mosaic-like organization for each type of functional domain without there being an overall pattern of repeating hypercolumns, each of which contains all types of functional domains. See Figure 4.

Similar results were found in cat area 17, in which many iso-orientation lines intersect the borders between ocular dominance borders at right angles, and orientation singularities are concentrated in the center regions of the ocular dominance columns. Pinwheels are frequently centered on centers of high and low spatial frequency preference patches. Low spatial frequency domains tend to co-localize with peaks of ocular dominance and with cytochrome oxidase blobs.

Optimal Coverage

An especially important issue for full coverage of all possible feature combinations is the relationship of the different maps with retinotopy, the map of visual space onto a cortical area, because any stimulus should be possible in each region of visual space. Retinotopy is considered the primary map in visual cortex because it exists in species in which even an orientation map does not exist, such as the rat, and because it is the smoothest map, demonstrating, perhaps with orientation, the strongest clustering of neurons. Whereas for other maps many cycles of all values of a feature exist in primary visual cortex, repeating at approximately 1 or 2 mm, retinotopy is the only map for which a single period exists. Therefore, it is critical whether other

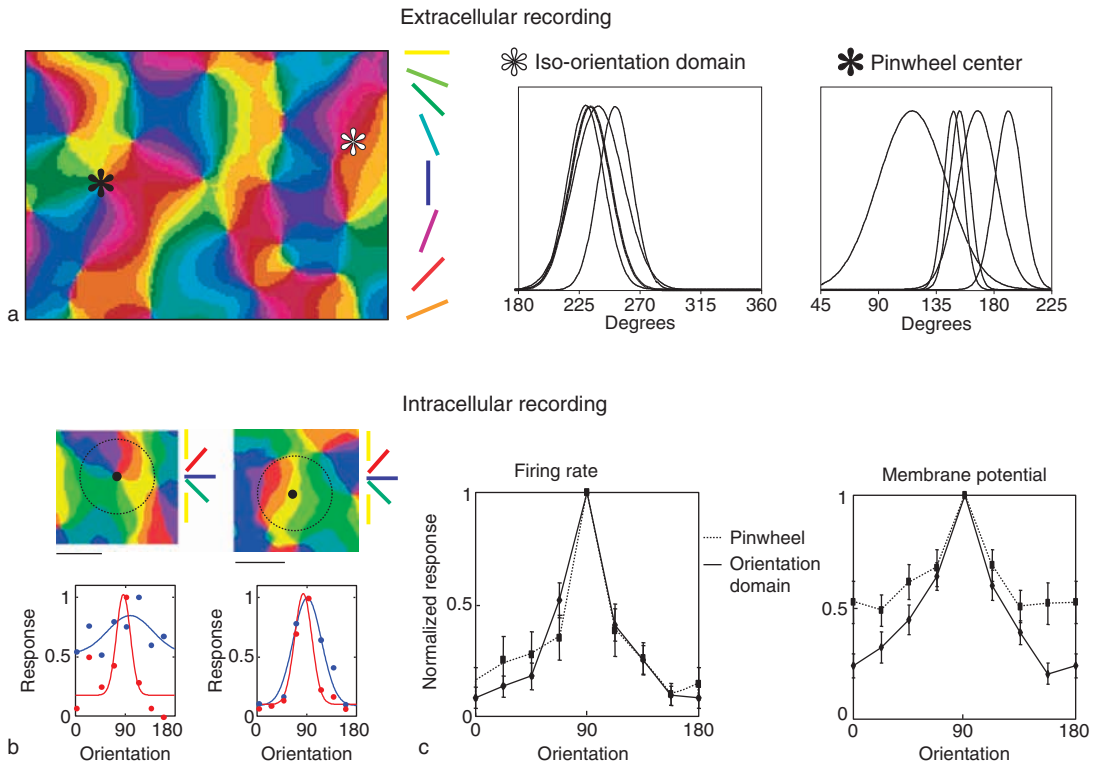

Figure 2 Properties of neurons in pinwheel centers – cat areas 17 and 18. (a) Extracellular recording using tetrodes. (Left) Orientation map with two targeted electrode penetrations, an iso-orientation domain and a pinwheel center. The orientation tuning curves of neurons from the iso-orientation domain (middle) and from the pinwheel center (right) show similar tuning widths, although for the latter the preferred orientation has a larger scatter. Thus, neurons at the pinwheel center prefer a variety of orientations, but the selectivity of each neuron's spiking response to its preferred orientation is similar to that of neurons in iso-orientation domains. (b) Intracellular recording. Similarly to extracellular recordings, the spiking responses (red curves) of a simple cell in an iso-orientation domain (left) and of a simple cell in a pinwheel center (right) have similar tuning width. However, the subthreshold membrane responses (blue curves) exhibit a distinct difference. Whereas the iso-orientation domain cell exhibits a well-tuned membrane potential response, the pinwheel center cell's membrane potential response is much less selective. This low selectivity likely results from the cell's local input, which arises from cells with a wide range of preferred orientation. (c) Averaging over a population of intracellularly recorded cells, similar spike tuning width for pinwheel centers and iso-orientation domains is observed concurrently with a less selective subthreshold membrane potential response for pinwheel centers versus iso-orientation domains. (a) Adapted from Maldonado PE, Godecke I, Gray CM, and Bonhoeffer T (1997) Orientation selectivity in pinwheel centers in cat striate cortex. *Science* 276: 1551–1555. (b, c) Adapted from Schummers J, Marino J, and Sur M (2002) Synaptic integration by V1 neurons depends on location within the orientation map. *Neuron* 36: 969–978.

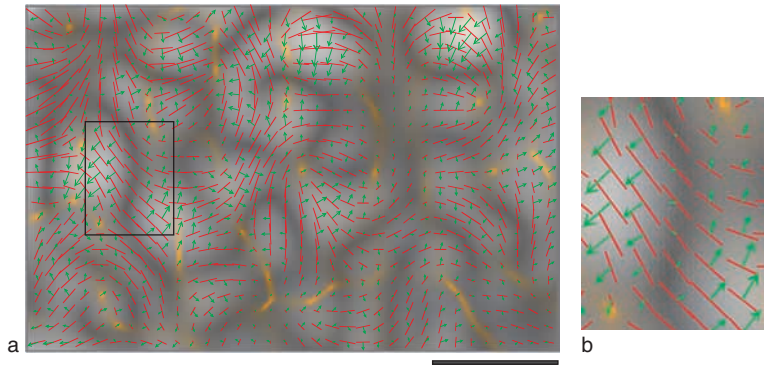

Figure 3 Direction-selectivity mapping. (a) A direction-selectivity map from cat area 18 shown with the orientation map in the same cortical surface. The red lines represent the preferred orientation and the magnitude of preference by their angle and length, respectively. Superimposed are green arrows, representing the local preferred direction. At the background, in gray-level presentation, is the map of magnitude for direction preference (dark represents low magnitude). Locations coded in yellow exhibited a high rate of change of preferred orientation angle (orientation singularities, computed using the gradient transform). (b) Zoom into a window in (a), showing a direction fracture line (gray) running between two pinwheel centers (yellow) and dividing an iso-orientation domain (red bars at ~135 °) into two areas preferring opposite motion directions (green arrows pointing to ~45 ° in the upper left portion of the image and to ~225 ° in the lower right). Scale bar = 1 mm (a). Adapted from Shmuel A and Grinvald A (1996) Functional organization for direction of motion and its relationship to orientation maps in cat area 18. *Journal of Neuroscience* 16: 6945–6964.

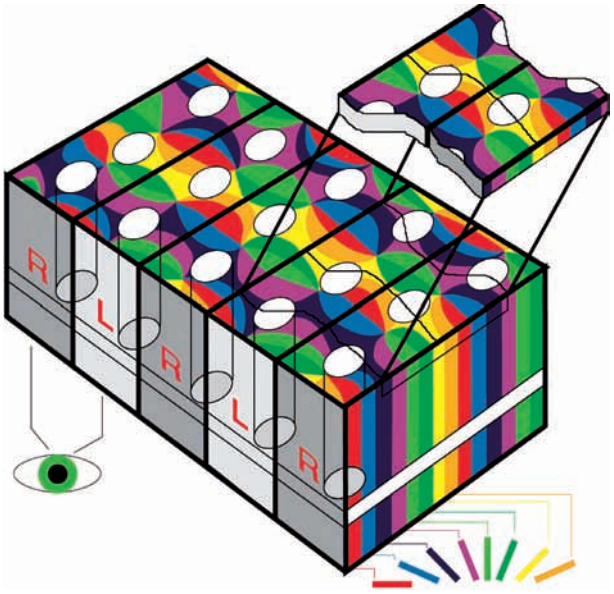

Figure 4 Revised ice cube model of monkey V1, schematically illustrating the relationships between orientation columns, ocular dominance columns, and cytochrome oxidase blobs. Black lines mark the borders between columns of neurons that receive signals from different eyes. This segregation is partially responsible for depth perception. White ovals represent groups of neurons responsible for color perception (CO blobs). The 'pinwheels' are formed by neurons involved in the perception of shape, with each color marking a column of neurons responding selectively to a particular orientation in space. Note that both the blobs and the centers of the pinwheels lie at the centers of R or L columns. The iso-orientation lines (appearing as a border between two colors) tend to cross borders of ocular dominance columns (black lines) at right angles. The top 'slice' above the ice cube model depicts two adjacent fundamental modules ($400 \times 800\,\mu m$). Each module contains a complete set of approximately 60 000 neurons, processing all three features of orientation, depth, and color. This scheme is oversimplified in various aspects, and the mosaic formed by hypercolumns is not as orderly as depicted here. Adapted from Bartfeld E and Grinvald A (1992) Relationships between orientation-preference pinwheels, cytochrome oxidase blobs, and ocular-dominance columns in primate striate cortex. *Proceedings of the National Academy of Sciences of the United States of America* 89: 11905–11909.

maps conform to it such that optimal coverage is achieved. Early modeling, in which map continuity and coverage were optimized for retinotopy and orientation, predicted distorted retinotopy and a negative correlation between the gradient of the retinotopic map and that of the orientation map.

In cat areas 17 and 18, conflicting results were obtained in studies using optical imaging to map orientation and extracellular electrophysiology to map retinotopy. An earlier study found distorted retinotopy exhibiting positive correlation of its gradient to the orientation map gradient, whereas a later study found smooth retinotopy with no local correlation to the orientation map. In V1 of tree shrew, as well as in ferret, the retinotopic map was found to be smooth and without local correlation to the orientation map (Figure 5). Detailed modeling, optimizing continuity and coverage, shows that this should be the case when feature maps additional to orientation, such as spatial frequency and ocular dominance, are superimposed on the retinotopic map. Further studies are required to resolve this issue.

Optimal coverage seems to be achieved by two general principles. First, whenever possible, the gradients of each pair of maps are negatively correlated, such that fast changes in one stimulus feature occur while the other changes slowly. This principle is manifest in the previously mentioned findings of pinwheel centers (i.e., areas of high orientation gradient), coinciding with the center of ocular dominance columns and spatial frequency domains, where these features change slowly. It implies an orthogonal intersection angle between contour lines of the two maps, which has been observed, for example, in the relationship between orientation and ocular dominance.

Second, when it is impossible to maintain orthogonal intersection angles between two maps, the strongest orthogonality is maintained in places where the two maps have the highest gradients. For example, in ferret V1 the contour lines of orientation, ocular dominance, and spatial frequency maps have a weak tendency to cross retinotopic contour lines at 90°. For each of these three maps, this tendency is stronger in the pixels with high gradient values. Given an anisotropic retinotopic map, such as that in ferret V1, orthogonal intersection angles between the contours of each pair of the three maps cannot be maintained throughout the cortical surface (under the continuity and coverage constraint) and are kept more strongly for those places where pairs of maps have overlapping high gradients, as shown in Figure 6.

Lateral Connections

The properties of connections between subdomains of cortical visual areas are a major feature of functional architecture and have been studied using anatomical tracing. Classical studies have shown that axon collaterals in V1 are distributed in clustered patches that can reach several millimeters away from the soma. These projections have been proposed to connect sites with neurons that have similar orientation preference. However, some quantitative studies combining anatomical tracing with electrophysiological orientation mapping have suggested that the overall bias favoring long-range iso-orientation projection sites is weak, perhaps due to significant, although diffuse, intercluster projections, which have no iso-orientation preference. Optical imaging has been combined with tracer injection to study this

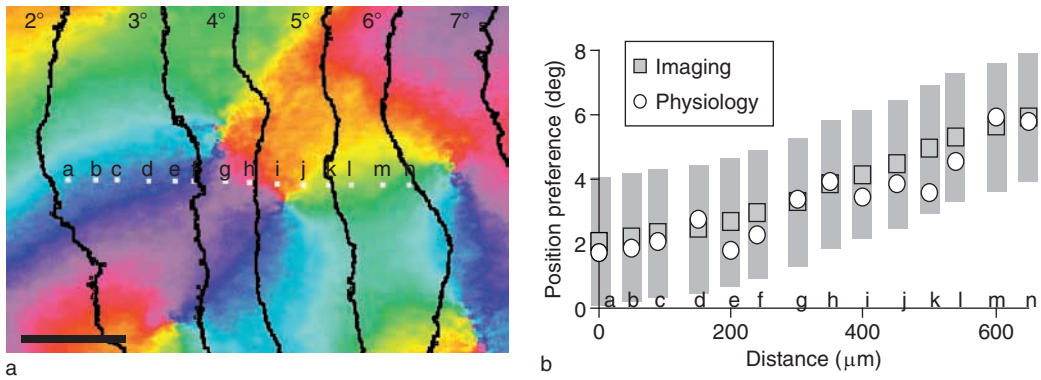

Figure 5 The retinotopic map is locally smooth and independent of the orientation map – tree-shrew V1. (a) Orientation map (color) superimposed with a position preference map with isoazimuth lines at 1° intervals (black lines). At this scale, the retinotopic map is smooth and bears no relationship to the orientation map. (b) Extracellular recordings (white circles) and optical imaging (solid squares) at the 14 locations marked in a show that even at this finer scale, a few hundred micrometers, graded progression over the cortical surface is accompanied by smooth progression in visual space. Scale bar = 1 mm (a). Adapted from Bosking WH, Crowley JC, and Fitzpatrick D (2002) Spatial coding of position and orientation in primary visual cortex. *Nature Neuroscience* 5: 874–882.

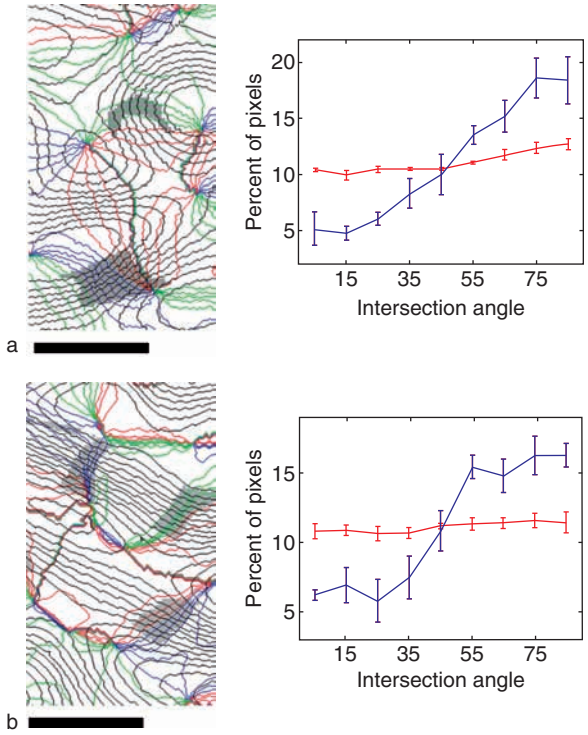

Figure 6 Intersection angles of pairs of maps – ferret V1. Orthogonal intersection angles between a pair of maps are kept more for areas where both maps have high gradients than over the whole cortical surface. (a) Orientation and ocular dominance maps. (Left) Colored orientation contour lines are superimposed on black ocular dominance contour lines. Gray regions indicate overlap of high-gradient regions of orientation and ocular dominance (30th percentile). (Right) Distribution of intersection angles between the orientation gradient and the ocular dominance gradient over all pixels (red) and for the pixels with overlapping high gradients (blue). The former exhibit all values, whereas the latter tend to exhibit high values closer to orthogonal. (b) The same as (a), but for the orientation and spatial frequency maps. Adapted from Yu H, Farley BJ, Jin DZ, and Sur M (2005) The coordinated mapping of visual space and response features in visual cortex. *Neuron* 47: 267–280.

question. For example, in cat area 17, layer 3 injections near pinwheel centers result in projections that have no orientation bias, in agreement with the subthreshold membrane responses mentioned previously. Iso-orientation domains of layer 4 in area 18 likewise send overall smooth projections, without orientation preference. Individually labeled inhibitory basket cells showed a slight preference toward projecting to iso-oriented patches. In view of the complexity of the neocortical circuit, additional studies are required to shed light on these apparently contradicting results.

Lateral connections have also been studied physiologically, by assessing the cross-correlation in spiking activity between laterally displaced neurons, in conjunction with optical imaging. Neurons are strongly interconnected locally, irrespective of their relative orientation. Furthermore, correlation decreases with distance between the neurons, similarly for same orientation pairs and for orthogonally oriented pairs.

Development

Since the earliest days of modern research of visual cortex by Hubel and Wiesel, the development of its functional features over the course of an individual's maturation has been explored. This field has seen a surge of works employing optical imaging of intrinsic signals to study how the formation and stabilization of orientation preference, ocular dominance, and retinotopy are molded by visual experience, neural activity, and molecular and genetic factors. Such studies often require longitudinal experiments to determine changes in the cortical functional architecture over long periods of time. Optical imaging is particularly appropriate for these studies since it offers both

the required spatial resolution and the ability to perform prolonged, comparative studies. The results of these studies are reviewed extensively elsewhere in the encyclopedia.

V2

Cytochrome oxidase staining in primate V2 produces three sets of repeating modules: thin dark, thick dark,

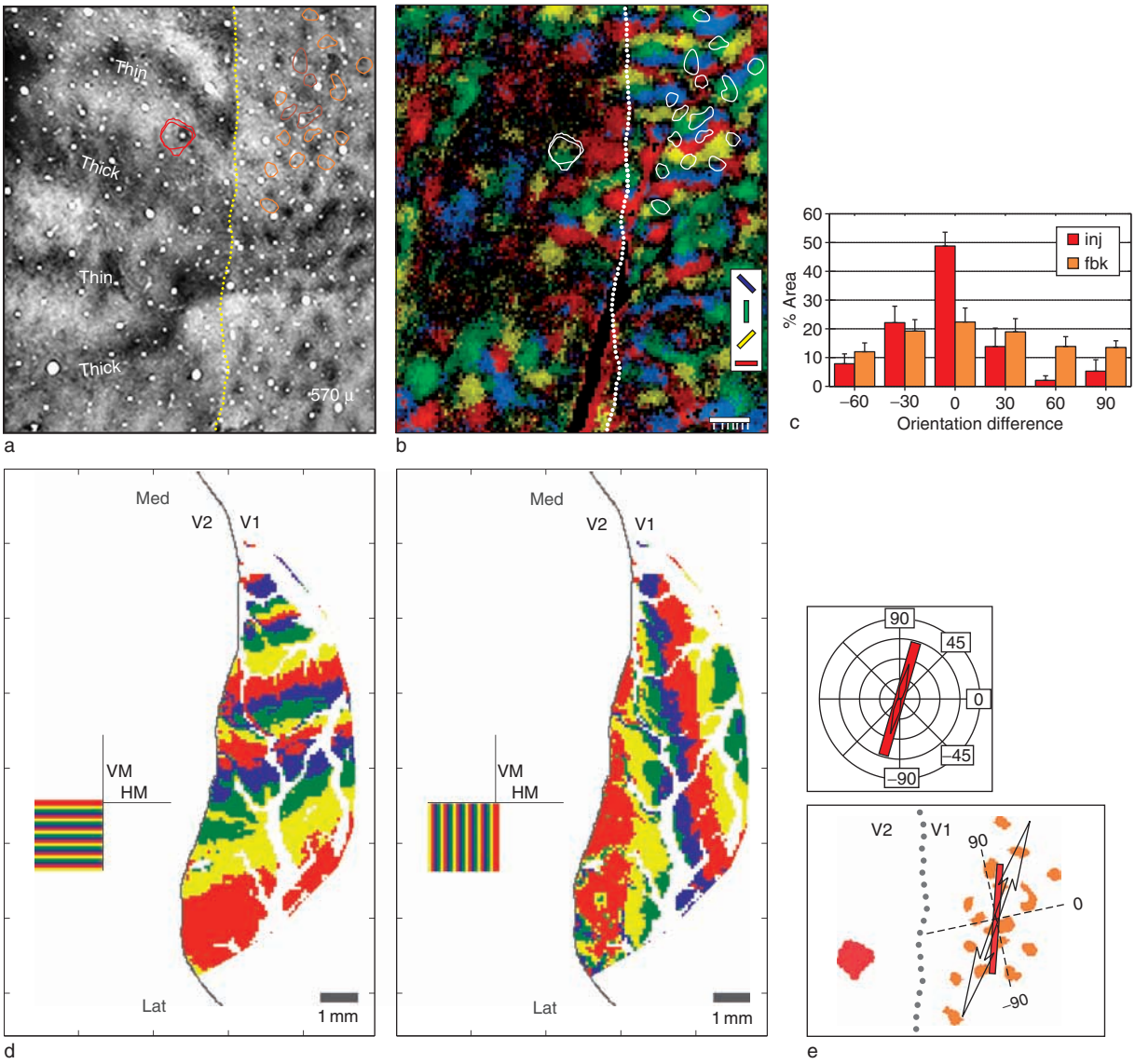

Figure 7 Feedback projections from V2 to V1 show orientation bias and exhibit retinotopic axis specificity. (a) Cytochrome oxidase-stained tangential section at ~570 μm depth, showing cytochrome oxidase blobs in V1 (right of dotted yellow line) and thick, thin, and pale stripes in V2 (left of border). The injection site in V2 is delineated in red and the clusters of feedback projections to V1 in orange and brown, with darker color denoting denser clusters. (b) Orientation map of the same cortical surface, aligned using the superficial blood vessel pattern. Separate local alignments between the histological slices and optical imaging data were performed around the V2 injection point and the V1 clusters. Dark regions correspond to locations with weak orientation preference (e.g., the darker regions in V2 which correspond to the thin stripes). Both the injection site and the feedback projection clusters demonstrate preference for a near-vertical orientation. Note the overall elongated structure of the feedback clusters. (c) Distribution of preferred orientation at the injection site and feedback projection clusters for the population of six cases in the study, with zero denoting the mode orientation at the injection site. The injection site and the feedback clusters peak at a similar orientation, but the distribution for the feedback clusters is broader and less selective. (d) Retinotopy in V1, showing the mapping of horizontal and vertical strips in visual space (left and right, respectively). These data are from a typical hemisphere in a different monkey. (e) Retinotopic axis specificity of the feedback projections. (Top) Preferred orientation values within the injection site – distribution of values (black line) and vector sum (red bar). (Bottom) Injection site and feedback clusters, overlaid with a coordinate system of the local retinotopic axes at the clusters' center of gravity (dashed lines correspond to horizontal and vertical axes in visual space). The spatial distribution of the feedback projections is anisotropic, with the major axis of anisotropy (open double-arrowed bar) approximately parallel to a retinotopic axis that corresponds to the preferred orientation at the injection site in V2 (red bar). Adapted from Shmuel A, Korman M, Sterkin A, et al. (2005). Retinotopic axis specificity and selective clustering of feedback projections from V2 to V1 in the owl monkey. *Journal of Neuroscience* 23: 2117–2131.

and pale stripes. The neurons within each of these modules have specialized functional properties. Optical imaging has helped characterize the modular organization of primate V2. Early optical imaging studies in V2 showed that it is possible to image stripes that correspond to the thin and thick cytochrome oxidase stripes by contrasting the response to monocular stimulation with that to a blank screen. In fact, the thin and thick stripes are discernible by calculating the difference in response between any stimulated condition and a blank (single-condition map). Selectively visualizing the thin stripes is done by contrasting color with luminance gratings since thin stripes contain regions of neurons that prefer color over luminance gratings. Within the color-preferring regions of the thin stripes, optical imaging has revealed an orderly map of hue preference. The thin stripes contain additional areas of luminance-preferring neurons clustered together; however, orientation mapping is absent from both subcompartments of the thin stripes. Clustering according to preferred orientation is predominant in both pale and thick stripes, with the iso-orientation domains about twice as wide there as in V1 (4 times the area). Thick stripes contain clustering of disparity-selective cells.

Feedback projections from the pale and thick V2 stripes to V1 in owl monkey show weak orientation bias, such that the sites receiving input in V1 exhibit a slight tendency to have the same preferred orientation as the site sending it, as shown in **Figures 7(a)–7(c)**. Interestingly, however, the spatial distribution of the feedback projections is anisotropic, with the major axis parallel to a retinotopic axis in V1 that corresponds to the preferred orientation of cells in the V2 injection site (**Figures 7(d) and 7(e)**). This suggests that feedback from V2 to V1 facilitates the response to collinear contour elements, which has been shown to be perceptually enhanced relative to noncollinear configurations. Other evidence suggests that intrinsic lateral connections within V1 are involved in contour integration, and the relative contributions of intrinsic and feedback connections to extraclassical receptive field modulation are currently an active research area.

Beyond V1/V2

Although the majority of imaging studies in visual cortex were performed in primary visual cortex, optical imaging of intrinsic signals has been successfully implemented in a variety of other visual areas. Examples include retinotopy in V3 (dorsal) of owl monkey, showing that this region has a representation of the lower visual field, smaller than and mirror-imaging that in V2 (dorsal). Orientation and size preference were mapped in V4 of macaque, showing an orientation map in central visual field with iso-orientation domains comparable in size to those in V2 and revealing size-sensitive regions suppressed by large stimuli. In owl monkey area MT, which is involved with motion processing, direction and orientation were both mapped, suggesting, since the orientation map is even stronger than the direction map, that shape processing is important in MT functioning. In inferotemporal cortex of macaque, the highest strictly visual area, where single neurons respond selectively to moderately complex object features, optical imaging has shown that neurons with similar selectivity are clustered together. Furthermore, the neural coding for object features in this area seems to be distributed because a single object activates multiple patches, and a single cortical location is activated by different objects.

Conclusion

Optical imaging of intrinsic signals has been instrumental in the discovery and characterization of the plethora of functional maps in visual cortex, their development, and the relationships between them. There is a growing body of information arising from optical imaging combined with other techniques about connections between modules of visual processing and about the particular physiology of cells within these modules.

See also: Visual System: Functional Architecture of Area V2; Visual System: Multiple Visual Areas in Monkeys; Visual Cortex in Humans; Visual Attention; Visual Cortical Models of Orientation Tuning.

Further Reading

Bartfeld E and Grinvald A (1992) Relationships between orientation-preference pinwheels, cytochrome oxidase blobs, and ocular-dominance columns in primate striate cortex. *Proceedings of the National Academy of Sciences of the United States of America* 89: 11905–11909.

Bonhoeffer T and Grinvald A (1991) Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* 353: 429–431.

Bosking WH, Crowley JC, and Fitzpatrick D (2002) Spatial coding of position and orientation in primary visual cortex. *Nature Neuroscience* 5: 874–882.

Buzas P, Volgushev M, Eysel U, and Kisvarday ZF (2003) Independence of visuotopic representation and orientation map in the visual cortex of the cat. *European Journal of Neuroscience* 18: 957–968.

Chapman B, Godecke I, and Bonhoeffer T (1999) Development of orientation preference in the mammalian visual cortex. *Journal of Neurobiology* 41: 18–24.

Das A and Gilbert CD (1997) Distortions of visuotopic map match orientation singularities in primary visual cortex. *Nature* 387: 594–598.

Das A and Gilbert CD (1999) Topography of contextual modulations mediated by short-range interactions in primary visual cortex. *Nature* 399: 655–661.

Gandhi SP, Cang J, and Stryker MP (2005) An eye-opening experience. *Nature Neuroscience* 8: 9–10.

Grinvald A and Hildesheim R (2004) VSDI: A new era in functional imaging of cortical dynamics. *Nature Reviews Neuroscience* 5: 874–885.

Grinvald A, Lieke E, Frostig RD, Gilbert CD, and Wiesel TN (1986) Functional architecture of cortex revealed by optical imaging of intrinsic signals. *Nature* 324: 361–364.

Grinvald A, Sharon D, Vanzetta I, and Slovin H (2005) Intrinsic signal imaging in the neocortex; implications for hemodynamic based functional imaging. In: Lanni F, Konnerth A, and Yuste R (eds.) *Imaging Neurons: A Laboratory Manual,* 2nd edn., pp. 655–671. New York: Cold Spring Harbor Laboratory Press.

Hubener M, Shoham D, Grinvald A, and Bonhoeffer T (1997) Spatial relationships among three columnar systems in cat area 17. *Journal of Neuroscience* 17: 9270–9284.

Maldonado PE, Godecke I, Gray CM, and Bonhoeffer T (1997) Orientation selectivity in pinwheel centers in cat striate cortex. *Science* 276: 1551–1555.

Malonek D and Grinvald A (1996) Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: Implications for functional brain mapping. *Science* 272: 551–554.

Schummers J, Marino J, and Sur M (2002) Synaptic integration by V1 neurons depends on location within the orientation map. *Neuron* 36: 969–978.

Shmuel A and Grinvald A (1996) Functional organization for direction of motion and its relationship to orientation maps in cat area 18. *Journal of Neuroscience* 16: 6945–6964.

Shmuel A, Korman M, Sterkin A, et al. (2005) Retinotopic axis specificity and selective clustering of feedback projections from V2 to V1 in the owl monkey. *Journal of Neuroscience* 23: 2117–2131.

Tanaka K (2000) Mechanisms of visual object recognition studied in monkeys. *Spatial Vision* 13: 147–163.

Weliky M, Bosking WH, and Fitzpatrick D (1996) A systematic map of direction preference in primary visual cortex. *Nature* 379: 725–728.

Yousef T, Bonhoeffer T, Kim D-S, Eysel U, Toth E, and Kisvarday ZF (1999) Orientation topography of layer 4 lateral networks revealed by optical imaging in cat visual cortex (area 18). *European Journal of Neuroscience* 11: 4291–4308.

Yu H, Farley BJ, Jin DZ, and Sur M (2005) The coordinated mapping of visual space and response features in visual cortex. *Neuron* 47: 267–280.

Visual Cortical Models of Orientation Tuning

R Shapley, New York University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The primary visual cortex V1 is a nonlinear dynamical system, and one of its tasks is to find local stimulus features in the neural image of the visual scene relayed to V1 from the eye through the lateral geniculate nucleus (LGN). Different sources of intracortical inhibition contribute to the selectivity of V1 neurons: local-circuit inhibition, inhibition mediated by signals from long-distance intrinsic V1 horizontal connections, and feedback from extrastriate cortex that drives inhibitory interneurons in the local circuit. Different sources of excitation drive the activity in V1 cells: local thalamocortical projections, local circuit corticocortical excitation, long-distance horizontal V1 axons, and also feedback. Models of V1 cortex seek to explain the visual functions of V1 in terms of its architecture and connections.

Orientation Selectivity

One of the key questions neuroscientists have studied is how orientation selectivity is generated in V1. In V1 there is a transformation to orientation-tuned elements. Prior to V1, in the retina and LGN, there is weak or no orientation selectivity in single cells. It has been thought from the time of its discovery that orientation tuning, as an emergent property in visual cortex, must be an important clue to how the cortex works and why it is built the way it is. That is why it continues to inspire a volume of research. Much has been learned about basic principles of cortical neurophysiology on account of the intense investigation of orientation selectivity, particularly in the area of theories of the cortex to explain orientation selectivity.

The main problem V1 needs to solve, in order to generate orientation selectivity, is the reduction of responses to nonpreferred orientations. While the original feed-forward model provided by Hubel and Wiesel is supported by several studies, it does not account for the considerable evidence that has accumulated about the cortex's suppression of nonpreferred responses and the presence of powerful inhibitory conductance increases with visual stimulation in V1 cells. The feed-forward model is not wrong, but it is incomplete. The feed-forward model can be used to account for the pattern of orientation preferences, but it does not explain the distribution of orientation selectivity across the population. That is where corticocortical inhibition comes in.

Feed-Forward Models

There are two poles of the theory of orientation selectivity: feed-forward filtering on one hand and attractor states on the other. There are also investigators with views somewhere in the middle – perhaps a good label for their view would be recurrent network filtering. However, first chronologically, and first discussed here, is the feed-forward view, descended from Hubel and Wiesel. Hubel and Wiesel's feed-forward model (the HW model) has been a dominant idea in this field. It involves the addition of signals from LGN cells that are aligned in a row along the long axis of the receptive field of the orientation-selective neuron. For example, in the often encountered edge-sensitive orientation-tuned cells, Hubel and Wiesel postulated that a row of ON cells from the LGN would provide excitatory input and an adjacent parallel row of OFF cells would provide excitation for the opposite sign of contrast. There is some direct support for this kind of neural architecture in the ferret visual cortex from Chapman and Stryker, and from dual recordings in LGN and cortex in the cat from the work of Reid and Alonso. The experiment on cooling of cat V1 by Ferster's group was interpreted to mean that there is substantial orientation tuning of the collective thalamic input to a cortical neuron, consistent with the HW model. Chung and Ferster reached similar conclusions based on different experiments. In spite of all this evidence, several authors have shown that the HW model predicts rather little orientation selectivity by direct computation of the predicted synaptic current of the HW model.

The reason for the shortfall of orientation selectivity in the HW model, compared with experiments, has been discussed before. LGN cells have a low spontaneous rate but are quite responsive to visual stimuli, so their firing rate during visual stimulation clips at 0 spikes s^{-1}, even at fairly low stimulus contrast. Under these conditions LGN cells act like nonlinear excitatory subunits as inputs to their cortical targets, as originally suggested by Palmer and then by Tolhurst. Since the HW model simply adds up the LGN sources, its summation of the clipped LGN inputs would cause it to have a nonzero response at 90° from the optimal orientation. Computational simulations by Sompolinsky's group and by Miller and colleagues of the HW model have demonstrated

that this analysis is correct. An example is given in Figure 1, which shows a computation of the peak excitatory synaptic input (as a function of orientation) from an HW model onto a cortical cell. Such a model produces a substantial LGN input to a cortical cell at 90° from the preferred orientation, as seen in the figure. But many cortical cells respond little or not at all at 90° from peak orientation, so the HW convergence mechanism must be only a part of the story of cortical orientation selectivity.

One may wonder why the experiments of Ferster et al. and Chapman et al. are not decisive pieces of evidence for a large amount of orientation selectivity as a consequence of LGN convergence. The cooling experiments of Ferster et al. measured the first-order Fourier component (F1) in the intracellular voltage response to a drifting grating from a neuron in a cooled cortex and found it to be as tuned for orientation as in the warm cortex. But this does not account for the tuning of the mean spike rate, which is tuned as much as the F1 component of the spike rate. So to account for the orientation selectivity that is observed in the mean spike rate, it would have been necessary for Ferster et al. to measure also the orientation selectivity of the mean intracellular voltage response in the cooled cortex. They did not do this in the cooling experiments reported in 1996, probably for technical reasons. But if they had measured the DC response in the cooled cortex, and if the HW model does describe the LGN input to V1 neurons, then we can predict that Ferster et al. should have found the mean (or DC) intracellular voltage response to be only weakly selective for orientation (based on the analysis of the HW model in the previous paragraph). One can go further and analyze why Ferster et al. found so much orientation tuning for the F1 component. The only mechanism for orientation-selective response for an F1 response in the HW model is different spatial frequency resolution along the two axes of the elliptical receptive field. The reason is that the HW model has no inhibition; it is a purely excitatory model. Thus, Ferster et al. probably observed orientation selectivity in the cooled cortex only because the spatial frequency of the grating they used was too high for the elliptical LGN array to resolve it along the long axis, but spatial frequency was not too high for the summed LGN input to resolve it along the minor axis. The conclusion of all these considerations of the cooling experiment is that it is not decisive evidence for a feed-forward explanation of orientation selectivity. Next, the Chapman et al. results about rows of LGN terminals found in layer 4 of ferret V1 cortex, after cortical inactivation with muscimol, was offered as evidence that the HW model could explain orientation selectivity in ferret layer 4. But later it was learned that layer 4 neurons in ferret V1 are very broadly tuned for orientation – about half the neurons in layer 4 are almost completely nonselective. Therefore, the results of Chapman et al. are completely consistent with the conclusions of the analysis by Sompolinsky and Shapley and by Troyer et al. that the HW convergence mechanism would make neurons only very weakly selective for orientation.

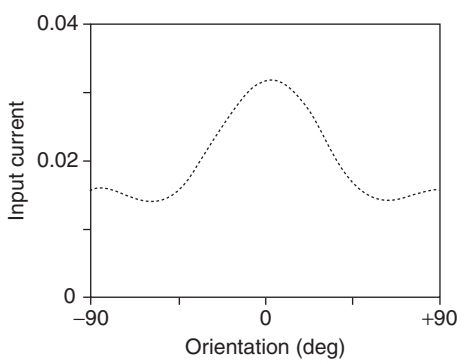

Figure 1 Orientation tuning curve of the synaptic current evoked by the lateral geniculate nucleus (LGN) input to a V1 cortical cell, calculated from a feed-forward model. In this model, the LGN afferents formed an ON-OFF-ON receptive field. Each subregion had an aspect ratio of 2. A total of 24 OFF-center cells comprised the central OFF subfield, while 12 ON cells comprised each ON subregion in the model in an attempt to emulate the pattern of LGN inputs to cortex as indicated from experiments. Redrawn from Sompolinsky H and Shapley R (1997) New perspectives on the mechanisms for orientation selectivity. *Current Opinion in Neurobiology* 7: 514–522.

Cortical Inhibition

One possible addendum to the HW model, to increase the orientation selectivity greatly, has been known for a long time. One can obtain increased orientation selectivity by adding inhibition: either push–pull inhibition or some other kind of cross-orientation inhibition. Given what is known about V1 connections, this inhibition must come through cortical interneurons rather than directly from the thalamic afferents. Experiments on intracortical inhibition in V1 have given mixed results. Initially, Sillito's experiments with bicuculline suggested that intracortical inhibition might be necessary for orientation tuning. The role of intracortical inhibition was supported by the work of Bonds and collaborators on cat V1 cortex. They studied interactions between stimuli at different orientations, the effects of blocking activity in infragranular layers, and the effects of γ-aminobutyric acid on orientation selectivity. Sato and colleagues, in their study of macaque V1, found that neuronal orientation selectivity was reduced when very broadly tuned cortical inhibition was weakened by pharmacological blockade by bicuculline; response to off-peak orientations grew stronger relative to the

response to the preferred orientation. Eysel and his collaborators also found evidence in cat cortex for the important role of inhibition in causing a sharpening of orientation selectivity.

Feed-Forward Inhibition Model

A theory of orientation tuning in cat cortex has been offered by Troyer et al. and attempts to explain orientation tuning in terms of specific feed-forward 'push–pull' inhibition previously pushed by Palmer and then later by Tolhurst. However, the main mechanism for sharpening of orientation tuning in the Troyer et al. model is corticocortical inhibition that is broadly tuned for orientation. This is not the feature of inhibition emphasized by the authors. They instead emphasize that the inhibition is phase-sensitive, but the phase sensitivity is irrelevant for the sharpening of orientation tuning. In the Troyer et al. model, there is moderately tuned LGN convergent excitation from an HW mechanism, and then more broadly tuned inhibition that cancels out the wide angle responses but that leaves the tuning curve around the peak orientation relatively unchanged. Therefore, this model in some ways resembles other corticocortical interaction models for orientation selectivity. Like the Ben-Yishai et al. model discussed in the section titled 'Cortical excitation and attractor models,' it achieves contrast-invariant orientation selectivity.

Cortical Excitation and Attractor Models

The idea that corticocortical recurrent excitation within V1 plays a crucial role in orientation tuning has been advanced by theorists of brain function. There are three well-known papers that make the case for this corticocortical recurrent excitation. Somers and his colleagues presented an elaborate computational model for orientation tuning that involved recurrent excitation and inhibition. Douglas and his colleagues argued for the importance of recurrent excitation in cortical circuits and its role in orientation tuning as an example. They proposed there was a 'canonical microcircuit' of V1 cortex, which described a cortical circuit with a large amount of recurrent excitation. Third, Ben-Yishai et al. offered an analytical model from which they made several predictions. One of their important theoretical results is that one cannot predict the observed contrast invariance of orientation tuning with feed-forward models, but their model with recurrent excitation and inhibition does exhibit contrast invariance. Another of the Ben-Yishai et al. results is that if recurrent feedback is strong enough, one will observe the 'marginal phase' state in which V1 behaves like a set of attractors for orientation.

Tsodyks and his colleagues also considered attractor states of recurrent excitatory models. The concept is that very weakly orientation-tuned feed-forward signals can be massively sharpened by strong recurrent excitatory feedback causing the cortex to respond to any visual signal by relaxing into a state of activity governed by the pattern of corticocortical feedback.

NYU Models

McLaughlin, Shelley, Wielaard, and Shapley developed a large-scale model, referred to as the Version 1 NYU model, of four hypercolumns in layer 4cα of macaque V1. Inspired by the ring models in the work of Somers et al. and Ben-Yishai et al. it seeks to account for the same set of phenomena as these models but with more biological realism. Version 1 incorporates known facts about the physiology and anatomy of V1. One novelty in this model is that the spatial strength of connections between neurons is taken to be the spatial density of synaptic connections revealed by anatomical investigations of cortex. The model places the 'footprints' of synaptic excitation and inhibition on the pinwheel latticework that is revealed by cortical optical imaging. The Version 1 model causes significant sharpening and also diversity of orientation selectivity compared with the LGN inputs to cortex, and it also produces simple cells.

One big difference between the Version 1 NYU model and the push–pull model of Troyer et al. mentioned above is that in Version 1 the inhibitory conductance input to a cell is phase-insensitive (the opposite of push–pull). This happens because inhibition onto a model cell is a sum from many neural sources. Each of these sources is a cortical inhibitory cell with a fixed phase preference that is different from those of neighboring neurons. This view of the nonselective nature of local corticocortical inhibitory interactions is consistent with the measured phase insensitivity of measured inhibition.

Building Version 1 presented the challenge of obtaining simple cells in a nonlinear network model of V1 and of obtaining orientation selectivity. One basic principle established by this modeling work is that there is a need for strong lateral inhibition for any of these cortical models to behave like V1 cortex. Strong cortical inhibition was needed just to produce simple cells. The strong inhibitory synaptic coupling makes the model neurons' membranes shunted, and as a result the neurons act like coincidence detectors as calculated by Shelley. This could explain the observed increased efficacy for nearby spikes from retinal ganglion cells to cause excitation of cortical

cells better than widely spaced spikes, as shown in the experiments of Kara and Reid. The high-conductance state also means that membrane potential fluctuations are what produce model neuron spiking. With realistic dendritic and axonal spread of excitation and inhibition within the input layer, corticocortical recurrent excitation carries a 'complex' signal. Thus the more a model cell is excited by corticocortical recurrent excitation, the more it is like a complex cell. This is an idea similar to the proposals of Chance and her colleagues about the role of recurrent excitation in building complex cells.

Then a model was proposed that will produce simple and complex cortical cells, Version 2: a recurrent cortical model, also called the 'egalitarian' model, to distinguish it from hierarchical models of V1. Figure 2 illustrates some of the components of Version 2. The model is a large-scale computational model of a 1 mm patch of layer. This network consists of approximately 4000 neurons, 75% excitatory and 25% inhibitory integrate-and-fire, conductance-based point neurons. The network is comprised of cells within four orientation hypercolumns. Distinctive features of Version 2 are that the local lateral connectivities, as set by the specification of the coupling kernels (plotted in Figure 2, lower left), are nonspecific and isotropic (unlike phase- or orientation-specific architectures as in models like that of Troyer et al.) and not hierarchical. The orientation and spatial phase preferences of individual model neurons are influenced by the convergent LGN excitatory drive resulting from the summed output of segregated sets of on- and off-center LGN cells, as in the HW model. In the model cortex, orientation preference is laid out in pinwheel patterns based on the experimental results of Bonhoeffer and Grinvald. Preferred spatial phase varies randomly from neuron to neuron, as dictated by some experimental results. However, orientation selectivity and response time course are generated by intracortical inhibition (as in the Version 1 model).

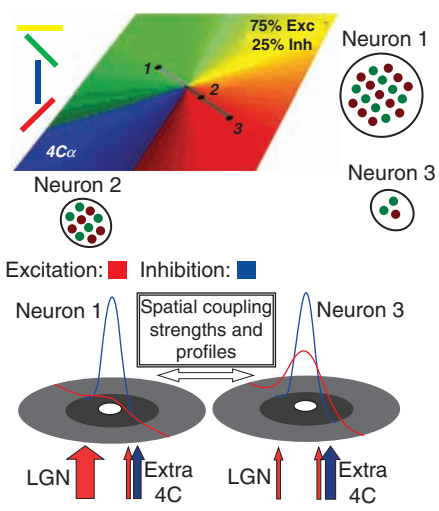

Figure 2 Diagram of the Version 2 egalitarian model. Each V1 cell 'sees' a collection of lateral geniculate nucleus (LGN) cells that is sampled from a 2D Gabor function. Orientation preference is laid out in pinwheels (the color orientation map). Spatial phase preference is random for nearby cells. The number of LGN cells (*Ni*) providing afferents varies from cell to cell (shown for three sample neurons). *Ni* distributed uniformly between 0 and 30. Lower left, corticocortical couplings are isotropic with interaction profiles Gaussian in shape (excitation in red; inhibition in blue), with the length scale of excitation = 200 μm and that of inhibition = 100 μm. 4Cα refers to the macaque primary visual cortex. Exc, excitatory; Inh, inhibitory.

The main achievement of the egalitarian model is simple and complex cells coexisting in the same network. The egalitarian architecture may be a possible explanation of the earlier finding by Ghose that excitatory cross-correlation could be observed between complex and simple cells, a result that seems counterintuitive based on the feed-forward model. The basic idea of the egalitarianism in the Version 2 model is that there is not a strict hierarchy of connections; rather, cells' properties come from the balance of their synaptic connections. In such a model, cells could be shifted from simple to complex by, say, weakening inhibition. This had been observed experimentally, but there was no explanation offered before the Version 2 model. The opposite effect, namely, transferring a cell from the complex to simple group by increasing inhibition, was reported in 2006 by Bardy and colleagues, so this could be viewed as experimental confirmation of a prediction of the egalitarian model.

NYU model Version 3 is a refinement of the egalitarian model and includes the concept of sparse corticocortical connections. The orientation tunings of average responses to drifting sine gratings obtained with Version 3 are plotted for a few model neurons in Figure 3. The upper row, labeled *m*, is the spike rate, and below are shown intracellular membrane potentials (V) and the synaptic conductances – g_{lgn}, g_E, and g_I – that control the cells' firing. Each column in Figure 3 is for a different cell. Note that the inhibitory conductance (bottom row) is mostly very broadly tuned or flat, consistent with measurements of untuned suppression in orientation dynamics experiments. Untuned or broadly tuned inhibition arises naturally in network models like the egalitarian model that take into account the measured spatial spread of inhibitory connections in V1.

One important feature of the pure feed-forward model is that it does not predict different dynamics of responses to different orientations. The egalitarian models (like other recurrent inhibition models) predict there ought to be a difference because of the somewhat slower time course of inhibition. This

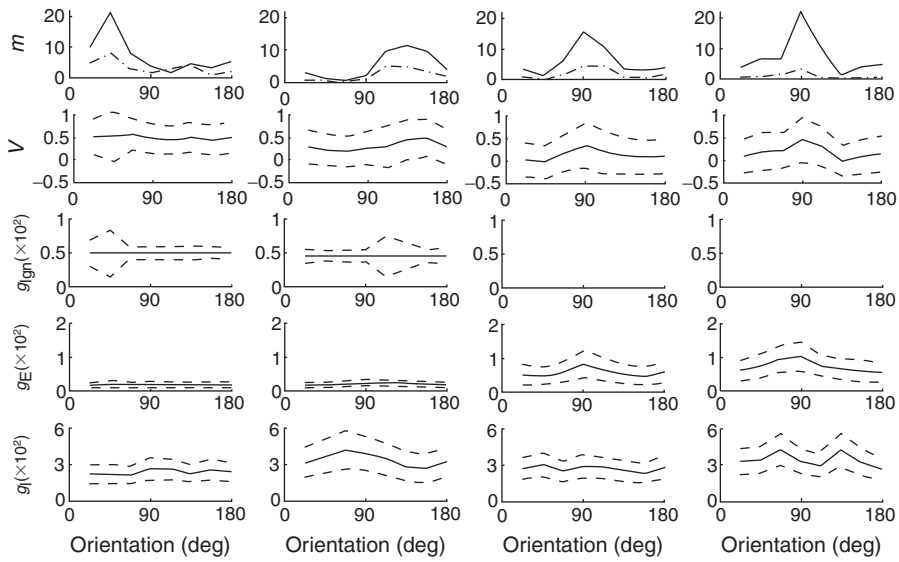

Figure 3 Four sample excitatory cells (two simple cells on the left, two complex cells on the right) from the Version 3 egalitarian V1 model. Plots of orientation tuning (from top to bottom rows) for cells' firing rates (spikes s^{-1}; labeled *m*), for medium contrast (solid lines) and low contrast (dot-dashed lines); membrane potential (labeled *V*); excitatory conductances from LGN inputs (g_{lgn}); corticocortical excitatory conductances (g_E); and corticocortical inhibitory conductances (g_I; in units of 1 s^{-1}). Dashed curves in rows 2–5 show the mean ± SD for potentials and conductances. All quantities are time-averaged.

effect was observed in reverse correlation experiments. In more recent work on the dynamics of orientation selectivity, a baseline of response has been established through the use of blank stimuli as well as oriented grating patterns, and deviations above and below the baseline have been measured. This has yielded the observation that responses to preferred and orthogonal-to-preferred orientations had different dynamics. The result is shown in Figure 4.

Figure 4 shows population average responses from many (140) macaque V1 neurons. The time course of the orthogonal-to-preferred response (blue) is more biphasic than the preferred response (red, and scaled down in black-dashed). Different dynamics of responses to nonpreferred and preferred orientations are seen also with voltage-sensitive dyes in cat V1 and with intracellular recording in cat V1.

Untuned Suppression

Untuned suppression, estimated from the orientation dynamics experiments, was approximately the same for large and optimal-size stimuli in rhese experiments, implying that it was generated mainly by local circuitry, as predicted in the recurrent cortical models (see Figure 3 for example). Previously, Ben-Yishai et al. had proposed untuned inhibition (they called it 'global inhibition') as one possible mechanism for enhancing orientation selectivity. Direct intracellular measurements in cat V1 prove the existence of broadly tuned cortical inhibition in some neurons.

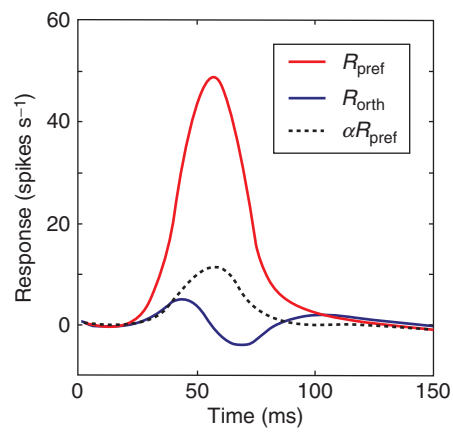

Figure 4 Time courses of the averaged response at the preferred orientation (R_{pref}; red curve) and at the orthogonal-to-preferred orientation (R_{orth}; blue curve). These are experimental data from the population-averaged dynamics of orientation tuning, based on observations in reverse correlation experiments on macaque V1 by Xing, Shapley, and Hawken. Both R_{orth} and R_{pref} have positive responses at early time. At early time (prior to 50 ms after stimulus onset), R_{orth} has a time course similar to R_{pref} but is rescaled by a factor $\alpha = 0.27$ (dashed curve). At later times, the blue and dashed curves are different because of the time development of untuned suppression caused by local circuit inhibition.

Spatial Phase in Models and in Experiments, and the Simple/Complex Dichotomy

Since the pioneering work of Maffei and Fiorentini in 1973, many scientists have grouped cells into simple and complex classes by means of their responses to

drifting gratings. Simple cells, being more tied to the visual image, modulate their firing rate at the same temporal frequency as that of the drifting grating stimulus. Complex cells fire more impulses when a grating drifts by, but their firing rate is not modulated in time. The ratio of the modulated to the unmodulated components of the response (the modulation ratio, or F1/F0) has been used to classify cells as simple or complex. The F1/F0 ratio is closely related to spatial phase dependence because the modulation of response in simple cells is a consequence of their high sensitivity to spatial phase, while the spatial phase insensitivity of complex cells is related to their elevation of discharge when a drifting grating moves over their receptive fields. This is because a drifting grating is a grating whose spatial phase ϕ is changing continually with time; $\phi = 2\pi ft$, where f is the temporal frequency of drift.

Hubel and Wiesel in 1962 proposed that there are distinct simple and complex cell classes that are steps along a functional hierarchy. Mechler and Ringach reexamined this issue 40 years later by showing that the bimodal distribution of F1/F0 in the firing rate of V1 neurons could be the result of the nonlinear voltage-to-spike rate transduction in V1 neurons. The bimodality of the F1/F0 ratio had been taken previously as an indication of the functional distinctness of simple and complex classes. Then Priebe, with Ferster and Carnadini, using intracellular recording in cat V1, showed that a bimodal distribution of F1/F0 in the firing rate could be generated by the spike–threshold nonlinearity from an underlying unimodal distribution of F1/F0 in the membrane potential. Supporting this result, Shapley and colleagues reported that Priebe et al.'s results are what the Version 2 model predicts.

The way in which spatial phase is handled is very different in different models of V1 cortex. The Version 2 egalitarian model was designed so that spatial phase preferences of neurons were random even very locally. But Troyer et al.'s push–pull model introduced spatial phase specificity into local connections so that excitatory and inhibitory neurons near each other formed antiphase pairs. Intracellular data in fact suggest a somewhat intermediate situation: there is a spatial phase-specific component to the membrane potential but also a large phase-nonspecific component.

Gamma Frequency Oscillations in the Local Field Potential and V1 as Dynamical System

Analysis of the activity of the Version 2 model in responses to drifting grating stimuli revealed γ-band peaks in the spectrum of the population spike rate. Temporally structured network activity in the form of oscillations is a hallmark of the steady state dynamics of highly recurrent neural networks even when single units are firing irregularly. Therefore, ensemble oscillations when the V1 network is activated are expected if one hypothesizes that V1 is dominated by recurrent activity. This led to experiments to record the local field potential (LFP) simultaneously with single cell spikes and to analyze the spectral changes in the LFP with stimulation by drifting gratings. Henrie and Shapley found that visual stimuli caused an increase both in the variance and in the temporal structure of the LFP; power became more concentrated in the γ-band with increasing contrast.

Gamma-band oscillations have been interpreted as an important mechanism for feature integration, but instead, the LFP may simply be the reflection of local network activity. This notion is based on experimental observations and theory. Confirming earlier studies, Henrie found that the γ-band of the LFP could be fairly sharply tuned to orientation or spatial frequency and has contrast–response functions very similar to those of single V1 neurons recorded nearby when drifting grating stimuli were used that were of optimal size for the neurons. These results are consistent with the view that the γ-band peak in V1 is a natural consequence (i.e., theoretically predicted) of high levels of recurrent network activity recruited by stimulating the system with an effective stimulus (in agreement with earlier cortical theories).

Conclusions

Theories of V1 cortex have led to the view that cortical inhibition plays a very large role in the visual function of V1. Strong inhibition allows the cortex to have simple cells that are faithful to the image. Inhibition provides the mechanism for sharpening the selectivity for orientation and also for spatial frequency and other stimulus features. The high-conductance state caused by large inhibitory conductances makes most cortical cells act like coincidence detectors and therefore makes them very sensitive to temporal patterns in incoming spike trains and in network activity. The recurrent excitatory–inhibitory network of V1 determines the visual properties of the neurons in V1. But the network architecture is flexible and may serve also to allow other cortical areas to combine V1 responses to the immediate visual stimulus with stored memories of visual objects. Theories of higher-level vision may be able to make use of models of the V1 recurrent excitatory–inhibitory network to account for the way feed-forward and feedback signals are combined for perception.

See also: Retinal Models; Vision: Mechanisms of Orientation, Direction and Depth; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging; Visual Cortex in Humans; Visual Motion Models.

Further Reading

Ben-Yishai R, Bar-Or RL, and Sompolinsky H (1995) Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 92: 3844–3848.

Borg-Graham LJ, Monier C, and Fregnac Y (1998) Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature* 393: 369–373.

Chance FS, Nelson SB, and Abbott LF (1999) Complex cells as cortically amplified simple cells. *Nature Neuroscience* 2: 277–282.

Ferster D, Chung S, and Wheat H (1996) Orientation selectivity of thalamic input to simple cells of cat visual cortex. *Nature* 380: 249–252.

McLaughlin D, Shapley R, Shelley M, and Wielaard J (2000) A neuronal network model of sharpening and dynamics of orientation tuning in an input layer of macaque primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 97: 8087–8092.

Mechler F and Ringach DL (2002) On the classification of simple and complex cells. *Vision Research* 42: 1017–1033.

Priebe NJ, Mechler F, Carandini M, and Ferster D (2004) The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature Neuroscience* 7: 1113–1122.

Shapley R, Hawken M, and Ringach DL (2003) Dynamics of orientation selectivity in macaque V1 cortex, and the importance of cortical inhibition. *Neuron* 38: 689–699.

Shelley M, McLaughlin D, Shapley R, and Wielaard J (2002) States of high conductance in a large-scale model of the visual cortex. *Journal of Computational Neuroscience* 13: 93–109.

Somers DC, Nelson SB, and Sur M (1995) An emergent model of orientation selectivity in cat visual cortical simple cells. *Journal of Neuroscience* 15: 5448–5465.

Sompolinsky H and Shapley R (1997) New perspectives on the mechanisms for orientation selectivity. *Current Opinion in Neurobiology* 7: 514–522.

Tao L, Shelley M, McLaughlin D, and Shapley R (2004) An egalitarian network model for the emergence of simple and complex cells in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 101: 366–371.

Troyer TW, Krukowski AE, Priebe NJ, and Miller KD (1998) Contrast-invariant orientation tuning in cat visual cortex: Thalamocortical input tuning and correlation-based intracortical connectivity. *Journal of Neuroscience* 18: 5908–5927.

Wielaard J, Shelley M, McLaughlin DM, and Shapley RM (2001) How simple cells are made in a nonlinear network model of the visual cortex. *Journal of Neuroscience* 21: 5203–5211.

Visual Deprivation

R D Freeman, University of California at Berkeley, Berkeley, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

Historical Perspective

Of the primary senses, vision is arguably the most important for higher level mammalian species. It is not surprising, then, that there is a long history of rumination and experimental investigation of the consequences of missing or deficient or altered vision. The earliest considerations of these issues were at a philosophical level since experimental procedures had not been developed. An early question, described by John Locke in the seventeenth century, concerned the nature of visual perception of individuals who regain their sight after long periods of blindness. Could a spherical object be discriminated as different from a cuboidal one without the addition of tactile sensation? Locke's speculation was that this kind of visual discrimination would not be possible. Subsequent related philosophical ideas were put forward with views divided into two major categories, nativism and empiricism. Nativism implied that sensory systems were hard-wired at birth and that visual skills were developed early and independently of experience. Empiricism connotes the opposite, suggesting that mature visual behavior is dependent on the nature of postnatal visual experience. This idea therefore implies that there is considerable plasticity in postnatal periods. These concepts form the basis of nature versus nurture issues that could be considered using empirical science once suitable methodology was developed.

In addition to its primary importance, the visual system is very accessible and therefore amenable to experimental study. Physiological and behavioral exploration is relatively straightforward. For the study of deprivation, it is generally possible to control the degree and type of visual stimulation during a rearing procedure. This permits a wide variety of investigations. Some studies provide ample evidence of clear differences regarding the susceptibility to visual deprivation of different species. In addition, locations along the visual pathway vary in the degree of vulnerability to abnormal vision. These observations are considered later in some of the descriptions of experimental work.

Normal Development

In order to consider the effects of visual deprivation, it is first necessary to outline the main elements of normal development. Both neural and behavioral processes are considered. The relevant investigations in the normal system, as well as those concerning deprivation effects, are almost entirely limited to the central visual pathways from retina through lateral geniculate nucleus to primary visual cortex. Nearly all experimental work has been undertaken in these areas. By far the most common species studied is the cat, but the ferret has also been used because of its relative immaturity at birth. Neural changes that occur postnatally are easier to study than those that occur prenatally. The monkey is also of interest because, as a primate, it shares many close similarities with human sensory systems. Human infant work has largely been limited to behavior and gross electrophysiological approaches.

Before considering specific postnatal developmental findings, it is worth noting that the complexity of the prenatal process is considerable. Projections from neurons take specific and sometimes tortuous pathways to reach specific portions of nuclei. Inputs are received through dendritic trees and outputs are propagated through axons which project to specific targets. Control of this activity must be largely genetic and is carried out via molecular and chemical processes. Prenatal events have been studied in the visual system, in which it has been shown, for example, that cell generation in the retina occurs over a relatively long period of time and in a peripheral-to-central location sequence. The visual cortex develops in a multistage sequence. These processes involve cell migration, cell loss, dendritic prunings, terminal sprouting, synaptic realignments, and axon terminal retraction. These topics are beyond the scope of this article.

Many details of anatomical and physiological organization of the retina are immature at birth and this varies with species. In the kitten, physiologically, retinal ganglion cells are somewhat sluggish at the youngest ages at which they may be reliably studied (approximately 2 weeks postnatally). Receptive fields are also generally relatively large. During subsequent weeks, cells become more responsive and receptive fields shrink in overall size, suggesting that spatial resolution correspondingly increases. There are two standard, physiological categories of retinal ganglion cells in cats – X and Y types. For X types, linear spatial summation can be demonstrated. Relatively high spatial resolution and medium axonal conduction velocities are also reported for X cells. For Y types, summation is nonlinear, spatial resolution is low, and axonal conduction velocity is high. There are also anatomical differences in these cell types. Developmentally, these categories are difficult to distinguish until approximately 4 weeks postnatally.

In the lateral geniculate nucleus, lamination and segregation according to eye dominance appears in early postnatal stages of the kitten (1 or 2 weeks postnatally). However, response properties of cells are immature. They are generally unresponsive and have low spontaneous rates. When responsive, they have long response latencies and adapt quickly to repeated visual stimulation. Receptive fields are large and often lack center–surround organization. Maturity of these factors occurs rapidly during several postnatal weeks. Resolution changes have also been measured in monkey lateral geniculate nucleus (LGN) cells as a function of early postnatal age. LGN neurons near the fovea exhibit relatively low acuity at birth which improves markedly during the first few postnatal months. These types of measurements are important because they should establish neural limits to behaviorally determined acuities. Only limited data are available for primates and more are required for firm conclusions.

Primary visual cortex is the first location in the central pathway from the retina at which signals from left and right eyes are integrated. In addition to binocular vision, major differentiations are observed in response properties in visual cortex. Some developmental studies have been aimed at determining the status of neuronal maturity at early postnatal ages. This information is directly relevant to the nature–nurture question referred to previously. Most research has been undertaken with kittens, but monkeys have also been studied. Published reports vary considerably regarding the extent of maturity of the visual cortex at early postnatal periods. One of the problems in addressing this question, especially in the case of kittens, is the practical difficulty of obtaining reliable electrophysiological recordings in very young animals. In addition, there is a vascular tunic surrounding the lens that remains for approximately the first postnatal week. This experimental problem could confound conclusions regarding relative maturity. It appears that fundamental response characteristics are present to some extent as early as one can effectively measure them. This applies also to tests of binocular interaction. Tuning characteristics become stronger and more specific during early postnatal periods and more neurons become responsive. This process appears to level off at approximately 5 or 6 weeks in kittens. In monkeys, physiological maturity is probably more rapid.

Anatomically, there are additional indications of postnatal maturation in visual cortex. Lateral geniculate afferents in layer IV are segregated into ocular dominance bands in mature cats. This organization is not apparent until a few weeks postnatally. However, electrophysiological recordings show clear slabs of cortex that are controlled alternately by left and then right eyes. These findings apply to supragranular and infragranular cortical layers, indicating differential laminar effects. It should also be noted that during early postnatal weeks, there is a considerable increase in synaptic density in kittens and in dendritic spines in monkeys. This indicates that neural infrastructure and physiological response characteristics mature in parallel.

Abnormal Development

Binocular Deprivation

There are two main reasons why binocular deprivation has been studied. First, since postnatal visual development appears in general to be a gradual process, an obvious question concerns the requirements for normal maturation. Specifically, is normal visual input a requirement for development of the system? If so, then by preventing visual exposure in postnatal periods, developmental processes should be abnormal. Second, there are rare clinical disorders in which normal binocular exposure is not possible. To understand the physiological and anatomical consequences of these conditions, it is necessary to study animal models in which binocular vision is prevented. For these reasons, a number of studies have been conducted on the consequences of binocular deprivation.

In early work, effects of binocular deprivation were studied in a variety of animals with a focus on the retina and optic nerve. Deprivation was by dark-rearing or eyelid closure. Generally, in most studies, minimal effects were reported. In one study, however, severe degeneration was observed in the retinal ganglion cell layer and the optic nerve of a chimpanzee. In addition, light deprivation appears to interfere with metabolic processes in cell maintenance.

In experimental studies that followed this early work, effects of binocular deprivation were examined for both dark-rearing and bilateral lid suture. For both conditions of binocular deprivation, physiological deterioration can be pronounced. In kittens, effects become apparent after approximately 2 weeks postnatally. Approximately half of the cortical cells in bilaterally lid-sutured cats are unresponsive or respond weakly, nonselectively, or in an irregular manner. Orientation selectivity is severely diminished. In monkeys, these effects may be accentuated. Newborn monkeys seem to exhibit relatively mature neuronal response characteristics which are compromised extensively during binocular lid suture of relatively short periods.

Consistent with the physiological data, behavioral tests show clear abnormalities in animals reared with binocular deprivation. Clearly, peak levels of visual performance require adequate sensory input during

postnatal development. The potential for recovery may be tested by bringing dark-reared animals into a normal lighted environment. Some recovery is observed, the degree of which is approximately inversely proportional to the initial length of deprivation. The monkeys studied recovered more quickly and completely than cats, but they had initial normal visual experience before deprivation. Recovery is not uniform for all visual functions that may be assessed. For example, dark-rearing does not substantially affect the proportions of cortical cells that may be driven through either eye. However, behavioral tests show that stereopsis is completely deficient in these animals even after periods of recovery. Stereopsis requires a precise coordination in the processing of images from left and right eyes. The neural basis for this appears to be irreversibly incapacitated if coordinated active postnatal binocular input is prevented.

The two forms of imposing binocular deprivation, bilateral lid suture and dark-rearing, are not necessarily equivalent. Although the literature does not generally distinguish between these methods, direct comparisons have been made. In pure dark-rearing, there is a complete absence of visual input and, in principle, any developmental changes must be genetically controlled. As an aside, there is a practical problem in maintaining complete darkness during rearing. In the case of bilateral lid suture, depending on the pigmentation of the animal, diffuse light reaches the retina. There are two opposing views regarding the consequence of this diffuse exposure. In one, the unfocused or diffuse light exposure causes developmental abnormalities. The opposite view is that even unfocused images can allow some elements of quasi-normal development. This apparently establishes the visual pathway so that it is not subsequently modifiable. There is evidence that dark-rearing freezes development so that the visual pathways are left in a nonspecified condition. In this case, subsequent normal visual exposure and recovery are more likely. Given these observations, the method of choice for the study of binocular deprivation depends on the class of question one wishes to address.

Monocular Deprivation

By far the most widely studied form of visual deprivation is the restriction of visual input through one eye. There are several procedures by which this may be accomplished. The eyelids of one eye may be sutured closed. This blocks patterned vision but allows exposure to low levels of diffuse light. To eliminate the light, an opaque contact lens occluder may be used under the sutured lids, but this is potentially problematic. Another option is to use an opaque contact lens occluder alone. This allows manipulation of the duration of exposure or the combination of monocular with other forms of exposure. An example of this is the imposition of alternating monocular deprivation. The period of exposure for each eye may be varied and may be combined with other procedures. In the kitten, it is possible to suture the nictitating membrane of one eye in order to degrade pattern vision with only a minimal reduction of light level. Another procedure is to use a defocusing contact lens in one eye. This mimics a clinical condition by creating a deficient image in one eye. All of these methods have been used.

The central findings of monocular deprivation are as follows. Early reports of degenerative effects in the retina from dark-rearing do not apply in the case of monocular deprivation. There are no obvious consequences, either physiological or anatomical, at the retinal level. At the next stage in the central visual pathway, the LGN, there are clear anatomical effects. Cells in layers from the deprived eye are clearly shrunken. This shrinkage may be more pronounced in binocular regions of the visual field where there could be competitive interaction between deprived and nondeprived areas. There may also be physiological effects in deprived LGN cells, but they appear to be limited.

In the main input layer of the visual cortex, layer 4, effects of monocular deprivation are substantial as determined physiologically. There is a clear reduction of cells that can be driven through the deprived eye, although the effect is not as extensive as those in other layers. This difference is more pronounced in the monkey compared to the cat. Approximately 15–20% of neurons in layer 4 retain input from the deprived eye. A standard explanation is that input from LGN to cortex is normally segregated into alternating sections where it terminates in layer 4. The alternating sections are composed of ocular dominance columns that extend through the cortex approximately perpendicular to the surface representing input from each eye. Since there is segregation of inputs from the left and right eyes, there is a minimum of competitive interaction. In other cortical layers, which in the normal condition provide nearly full integration of input from both eyes, it is assumed that the deprived eye in the case of monocular deprivation must compete for functional connections. The nondeprived eye takes over in this situation, and for long periods of deprivation initiated at early postnatal stages, there is nearly complete dominance of neurons in supragranular and infragranular layers. It is important to note that there is a combination of both expansion and contraction of cortical space controlled by the nondeprived and deprived eyes, respectively. This has been demonstrated physiologically and anatomically. Therefore, the effects of monocular deprivation involve active processes and not just passive deterioration of neural connections.

A very broad range of investigations have been conducted on monocular deprivation. An enormous number of parameters have been examined. Examples include postnatal age of induction, length of deprivation, minimal periods required for effects, interactions with other specialized rearing procedures, effects of different ways to cause monocular deprivation, pharmacological effects, recovery, enhancement of vision in the nondeprived eye, effects of visual experience prior to monocular deprivation, and genetic and molecular manipulations.

The main assay for most of this work has been physiological estimates of the degree to which left and right eyes control studied populations of neurons in the visual cortex. The data that are generated from this assay are presented as ocular dominance histograms. They are determined from separate tests of each eye, and they represent the states of binocularity of animals reared with various conditions of monocular deprivation. Most of the histograms presented in the literature were obtained by use of subjective tests. The fundamental problem with this technique is that the state of binocularity may not be accurately determined by use of monocular tests. Direct quantitative assays of ocular dominance using both monoptic and dichoptic tests show clearly that input to a given eye is underestimated in the former case. This may not be a problem for large effects. However, for subtle differences between rearing conditions, inaccurate conclusions may result.

Monocular deprivation effects are caused by a spatially deficient image in one eye. The deficiency has to do specifically with spatial detail. For example, reduced light level causes minimal effects, but suturing of the nictitating membrane or use of a translucent occluder causes extensive consequences. In general, the earlier the initiation of monocular deprivation, and the longer it lasts, the greater the consequences.

Normal postnatal binocular experience prior to deprivation may increase substantially the extent of recovery. Some recovery is observed if binocular vision is instituted following monocular deprivation. Recovery is facilitated if the nondeprived eye is closed when the deprived eye is opened. There appear to be some species differences between cats, monkeys, and humans.

Orientation Selectivity

The two main properties of neurons in the normal visual cortex that are different from those of early pathways are binocularity and orientation selectivity. The deprivation studies described so far deal only with interference with binocular function. There have also been efforts to manipulate the orientation domain. Kittens reared with limited ranges of orientation (e.g., only vertical stripes) have been reported to have very few cortical neurons selective to orthogonal edges. This result has been challenged and is somewhat controversial. In the original work, estimates of this effect were made from populations of single neurons. Optical imaging of intrinsic signals shows that the visual cortex of stripe-reared kittens responds to all orientations. However, considerably more cortical space is devoted to the experienced orientation. A simple interpretation of this result is that the orientation preferences of some cells are shifted toward the rearing orientation. This implies an active process rather than a passive selective deterioration.

Clinical Conditions

The driving force for the study of visual deprivation stems from questions concerning nature–nurture and plasticity of the nervous system. A direct application of these issues arises from clinical disorders. A large number of the deprivation protocols that have been studied in animals have analogous clinical syndromes in human patients. The most obvious conditions are amblyopia, strabismus, and anisometropia. Amblyopia is a developmental visual malfunction in which resolution for spatial detail is reduced without any apparent organic disorder. It is generally associated with anisometropia (unequal refractive powers in left and right eyes) or strabismus (misalignment of left and right visual axes) or both. It may also occur following a clinical condition such as a tumor that blocks spatial vision in an early postnatal period. Textbook categories of amblyopia often refer to a strabismic or an anisometropic type. However, clinicians generally believe there is always some degree of strabismus associated with anisometropia. In any case, it is important to differentiate causation from correlation. Conditions of anisometropia and strabismus can result in amblyopia, but the reverse is also possible. Because of variable conditions in many cases of amblyopia, it is helpful to examine large populations of subjects. For example, this approach suggests that wide differences in behavioral performance may be based on the degree of remnant binocular function in amblyopic subjects.

Animal models of these clinical abnormalities have been investigated in cats and monkeys. These include studies of monocular deprivation by eyelid closure or opaque occluders or by defocusing lenses to mimic anisometropia. Artificial strabismus has also been studied by surgical resection of extraocular muscles and by use of optical prisms to dissociate the visual axes in left and right eyes. In general, results of clinical model studies are consistent with predictions that can be made from basic investigations. Electrophysiological and behavioral tests show deficits that are

commensurate with conditions of rearing. These studies also allow direct comparisons with human populations that have analogous clinical disorders. One difference that emerges from these comparisons is that human subjects appear to have more potential for degrees of recovery than animal models might suggest.

Studies of orientation selectivity also have a clinical analog in the form of ocular astigmatism. In these cases, there are sometimes substantial differences in focal planes corresponding to vertical and horizontal orientations. Assuming a congenital basis for the astigmatism, these are conditions in which a specific range of orientations in object space may be habitually defocused. This is analogous to selective orientation rearing. In these cases, there are clear differences in visual resolution between focused and unfocused orientations. The clinical condition is meridional amblyopia.

Mechanisms

Critical Period

Although most of the experimental work on visual deprivation has been of a phenomenological nature, there have been attempts to address underlying mechanisms. A few examples are considered here. Perhaps the most global issue is the establishment of critical or sensitive periods during which the visual system is most vulnerable to deprivation. Critical periods have been observed for many different sensory systems. They represent windows of time during which the brain is especially sensitive to specific types of input. This susceptibility alters neuronal properties that are modified by direct sensory experience. It is generally assumed that the genome cannot encode requirements of individual animals. The critical period provides a mechanism for a matching of neuronal circuits to individual requirements.

In the original studies of the critical period in kittens for the effects of monocular deprivation, two variables were confounded. Age at which deprivation was initiated and length of monocular closure were both varied. To obtain an accurate profile, one variable should be eliminated. This has been done by imposing a fixed period of monocular closure at successively increasing postnatal ages. The resulting profile shows that sensitivity is initiated as early as one can effectively measure neuronal responses (approximately 2 weeks postnatally) and reaches a peak at approximately 4 weeks. At the peak of sensitivity, a very short period of monocular closure can alter the ocular dominance histogram. As little as 8 h may be sufficient for an observable alteration, but this change can be rapidly reversed by subsequent binocular exposure. The period of susceptibility extends for several months and can be extended further by prior dark-rearing.

There are no comparable data for monkeys. The available data indicate that the critical period for the main input layer (layer 4) is very short compared to that for other layers. For the other layers, the peak is probably similar to that of kittens, approximately 4 weeks postnatally, and the period of vulnerability gradually tapers over an extended time course to 1 year or more.

In humans, of course, there is considerable interest in parallel data. One complicating problem is that there are generally multiple variables in human amblyopia. Clinical observations have placed amblyopia susceptibility to within the first 6 or 7 years of life. One attempt to define this period has been made with a relatively homogeneous population. Subjects with traumatically caused unilateral cataracts that occurred at different ages were studied to determine the extent of recovery following surgery. In general, cataracts in the first 3 postnatal years had devastating effects. Later deprivation had milder consequences, and at approximately 9 years there was complete recovery of visual resolution. In studies of both animals and humans, it is clear that the more normal binocular vision prior to deprivation, the more likely the chance for recovery.

Finally, it should be noted that the critical period for susceptibility to amblyopia is not equivalent to one for treatment. The main treatment regimes are pharmacologically or behaviorally based or a combination is used. There are cases in which very substantial recovery is reported even for severe forms of amblyopia. There is clearly some degree of plasticity in the adult nervous system and it is of clear interest to explore its limits.

Competition

There are several lines of evidence suggesting that the processes active during monocular visual deprivation involve competitive interaction between the pathways from each eye. The main observation is that monocular occlusion causes extreme effects compared to those of binocular deprivation. For example, visual recovery is clearly greater in dark-reared compared to monocularly deprived animals. A standard explanation is that afferent pathways from left and right eyes compete for control of cortical neurons. As noted previously, anatomical studies in the lateral geniculate body suggest that cell shrinkage is maximal for monocular deprivation in binocular segments. This result has been supported by behavioral and physiological determinations. Other studies are consistent with the notion that competition is a primary factor in developmental processes.

Suppression

The competition notions noted previously apply to the cortical input layers in which afferents from left and right eyes converge. However, there is considerable intracortical processing beyond the input stage. It is possible that the competitive advantage of the nondeprived eye in monocular deprivations is heightened by a suppressive mechanism in intracortical connections. In this case, activity from the nondeprived eye would suppress that from the deprived pathway. A direct test of this idea is to surgically remove the nondeprived eye. Cortical extracellular recordings may be made before and after enucleation so that activity can be observed in a given animal before and after the procedure. Although there is some inconsistency in the results, there appears to be both physiological and behavioral evidence consistent with the notion that suppression is one of the mechanisms in monocular deprivation. Attempts have also been made to neutralize suppression pharmacologically by iontophoretic or systemic use of bicuculline, an antagonist of the inhibitory neurotransmitter γ-aminobutyric acid. Results are not definitive, but there is evidence consistent with the suppression notion.

Sensitivity of Nondeprived Eye

A natural question that arises in the case of monocular deprivation concerns the nondeprived eye. Evidence demonstrates that the cortical space available to the deprived eye can be substantially reduced while that for the nondeprived eye is expanded. It is of interest to know if the expanded territory devoted to the nondeprived eye provides heightened visual sensitivity. This is a difficult question to address in humans since the clinical conditions resulting in monocular vision are broadly variable. It is further complicated by the fact that visual resolution may operate at a limit of what is possible based on anatomical and physiological features. Measurements of vernier acuity and contrast sensitivity for gratings over a range of spatial frequencies show generally higher levels for monocular subjects compared to values for normal viewers tested through one eye. There are also some conflicting results. Studies of mice and rats show clear evidence of enhanced visual performance through the nondeprived eye. Visual-evoked potentials, behaviorally measured visual acuity, and optokinetic responses are all enhanced in the nondeprived eye of monocularly occluded animals compared to monocular values of normal binocular subjects. More data are required to establish the basis and limits of enhanced visual performance in monocular subjects.

See also: Activity in Visual Development; Retina: An Overview; Retinal Development: An Overview; Vision: Light and Dark Adaptation; Visual Development; Visual Cortical Models of Orientation Tuning.

Further Reading

Awaya S and Watanabe Y (1995) Amblyopia. *Current Opinion in Ophthalmology* 6: 9–14.

Bear MF and Rittenhouse CD (1999) Molecular basis for induction of ocular dominance plasticity. *Journal of Neurobiology* 41: 83–91.

Cynader MS (1982) Competitive neuronal interactions underlying amblyopia. *Human Neurobiology* 1: 35–39.

Daw NW (1995) *Visual Development (Perspectives in Vision Research)*. New York: Plenum.

DeAngelis G, Ohzawa I, and Freeman RD (1993) The spatiotemporal organization of simple cell receptive fields in the cat's striate cortex: I. General characteristics and postnatal development. *Journal of Neurophysiology* 69: 1091–1117.

Duffy FH, Burchfiel JL, and Snodgrass SR (1978) The pharmacology of amblyopia. *Ophthalmology* 85: 489–495.

Feller MB and Scanziani M (2005) A precritical period for plasticity in visual cortex. *Current Opinion in Neurobiology* 15: 94–100.

Freeman RD and Bradley A (1980) Monocularly deprived humans: Nondeprived eye has supernormal vernier acuity. *Journal of Neurophysiology* 43: 1645–1653.

Freeman RD and Ohzawa I (1988) Monocularly deprived cats: Binocular tests of cortical cells reveal functional connections from the deprived eye. *Journal of Neuroscience* 8: 2941–2506.

Freeman RD and Ohzawa I (1992) Development of binocular vision in the kitten's striate cortex. *Journal of Neuroscience* 12: 4721–4736.

Hofer SB, Mrsic-Flogel TD, Bonhoeffer T, and Hubener M (2006) Prior experience enhances plasticity in adult visual cortex. *Nature Neuroscience* 9: 127–132.

Hubel DH (1978) Effects of deprivation on the visual cortex of cat and monkey. *Harvey Lectures* 72: 1–51.

Lewis TL and Maurer D (2005) Multiple sensitive periods in human visual development: Evidence from visually deprived children. *Developmental Psychobiology* 46: 163–183.

McKee SP, Levi DM, and Movshon JA (2003) The pattern of visual deficits in amblyopia. *Journal of Vision* 3: 380–405.

Mitchell DE and Timney B (1984) Postnatal development of function in the mammalian visual system. In: Darian-Smith I (ed.) *Sensory Processes, Part I*, pp. 507–555. Bethesda, MD: American Physiological Society.

Olson CR and Freeman RD (1980) Profile of the sensitive period for monocular deprivation in kittens. *Experimental Brain Research* 39: 17–21.

Rauschecker JP (1991) Mechanisms of visual plasticity: Hebb synapses, NMDA receptors, and beyond. *Physiological Reviews* 71: 587–615.

Sengpiel F, Stawinski P, and Bonhoeffer T (1999) Influence of experience on orientation maps in cat visual cortex. *Nature Neuroscience* 2: 727–732.

Sherman SM and Spear PD (1982) Organization of visual pathways in normal and visually deprived cats. *Physiological Reviews* 62: 738–855.

Vaegan TD (1979) Critical period for deprivation amblyopia in children. *Transactions of the Ophthalmological Society of the United Kingdom* 99: 432–439.

Visual Development

L M Chalupa and B Chapman, University of California at Davis, Davis, CA, USA

© 2009 Elsevier Ltd. All rights reserved.

For much of the last century, and continuing to the present time, the visual system has served as the premier model for studying the development and plasticity of the brain. Experiments dealing with development have been concerned primarily with which factors regulate the formation of specific features of the mature visual system, whereas those focused on plasticity have sought to assess how the normal course of events could be modified by various interventions at different developmental stages. An underlying unifying theme of this vast literature has been the role of neuronal activity in the normal development and plasticity of the visual system. Given the prevalent role of activity in these processes, we focus this article on this major issue. In particular, we consider the relevant historical background and underscore what is known and what remains to be resolved in this dynamic field.

Most neuroscientists agree that both activity-dependent and activity-independent mechanisms are responsible for the formation of neuronal connections in the developing brain. What is far less certain is how these factors interact to establish the exquisite patterns of specific connections that characterize neuronal systems at maturity. It is still an open question how neuronal connections are regulated by activity and which connectional features are shaped by activity-mediated events. Perhaps the most convincing evidence for the paramount importance of activity in the wiring of neuronal circuits comes from studies dealing with the development of ocular dominance, whereby the projections of the two eyes become segregated from one another. Indeed, the often-cited phrase "neurons that fire together, wire together" originated in this field. In the last few years, however, new observations have called into question the widely accepted role of activity in the development of the mammalian visual system. As a consequence, this field now appears to be in a state of unexpected flux.

The modern era of activity-dependent regulation of neural connections can be traced to Wiesel and Hubel's findings that monocular visual deprivation caused a marked loss in the ability of the deprived eye to activate cells in the kitten visual cortex, whereas only subtle visual deficits resulted from binocular deprivation. Subsequently, it was discovered that the geniculocortical projections of the two eyes in the monkey and cat appear to be initially intermingled before gradually segregating into left- and right-eye ocular dominance columns. After early monocular deprivation, the cortical territory innervated by the deprived eye became much more restricted than normal, whereas the projections of the nondeprived eye became substantially expanded. This neatly explained why the deprived eye was so deficient in driving cortical neurons. To account for the changes in ocular dominance columns observed after visual deprivation, Hubel and Wiesel suggested that during development "the maintenance of a synapse depends not only on the amount of incoming impulse activity but also on a normal interrelationship between activity in the different afferents" (Hubel and Wiesel, 1965: 1057). This idea is reminiscent of the model proposed some years earlier by Donald Hebb to account for reverberating cortical circuits.

With the introduction of *in utero* surgical methods in combination with axoplasmic transport anatomical tracing methods by Rakic, it became feasible to examine the prenatal development of the visual system. This revealed that in the embryonic macaque monkey the projections of the two eyes are initially overlapped in the dorsal lateral geniculate nucleus (dLGN) before segregating into eye-specific layers and that ocular dominance columns begin to form in the striate cortex more than a month before birth. Ocular dominance columns have also been found by Stryker to form in the neonatal cat prior to the onset of visual experience. Rakic also showed that fetal monocular enucleation (removal of one eye) resulted in the maintenance of a widespread projection from the remaining eye to the dLGN, demonstrating the importance of prenatal binocular interactions in this process. Similar results have been obtained following monocular enucleation in the fetal cat by the Chalupa lab.

Collectively, these studies demonstrated that, although visual experience can alter ocular dominance columns after birth, it is not required for the initial segregation of left- and right-eye projections. Binocular interactions, however, play an important role in the segregation of initially overlapped binocular projections even before birth. Whether such interactions are mediated by activity is still an open question. It is important to point out that the studies discussed so far did not directly address the issue of whether the formation of left- and right-eye segregated projections is mediated by activity-dependent or activity-independent events.

The role of activity in the formation of eye-specific projections was first assessed directly by blocking

retinal action potentials during the developmental period when the connections serving the two eyes are overlapping. This approach was employed by Stryker and Harris, who made binocular intravitreal injections of tetrodotoxin (TTX) in kittens to study the effects of retinal activity blockade on the development of ocular dominance columns, and by Shatz and Stryker, who infused this Na-voltage-gated channel blocker directly into the brain of fetal cats to study the effects of complete activity blockade on the development of eye-specific segregation in the LGN. In both cases, essentially equivalent outcomes were obtained. In the former study, it was found that retinal blockade prevented the formation of cortical ocular dominance columns; in the latter, retinal projections from the two eyes to the dLGN were found to remain widespread and overlapping. These findings have been interpreted to show that retinal discharges are essential for the establishment of left- and right-eye segregated projections in the dLGN and in the visual cortex.

One implication of these experiments is that retinal ganglion cells are capable of discharging action potentials very early in development, even before these cells can be activated by light. The presence of such activity was soon established by recordings from retinal ganglion cells of embryonic rats by Maffei. By means of multielectrode array recordings and imaging of Ca^{2+} transients using isolated retinas, the Shatz lab showed that spontaneous bursts of activity propagate across the retina in a wavelike manner. It has been suggested that such patterned activity is essential for the establishment of left- and right-eye inputs, the segregation of on and off projections to dLGN sublaminae, and the establishment of topographically appropriate projections. The overriding message of these studies is that during development activity-mediated events are required to refine initially diffuse projection patterns. This notion is in line with the concept of the Hebbian synapse, and it has led to many theoretical models explaining how activity-mediated events could regulate the formation of eye-specific projection patterns.

Perhaps the most convincing evidence for the role of patterned activity in forming segregated binocular inputs comes from studies in Carla Shatz's laboratory in which patterns of retinal activity were perturbed. One study used intravitreal injections of epibatidine, a cholinergic agonist which was found to block retinal ganglion cell firing *in vitro*. Binocular injections of this drug prevented eye-specific segregation, whereas monocular treatment with this drug caused the treated eye to occupy a smaller than normal territory within the dLGN while the projections of the treated eye were found to be expanded. The latter finding, reminiscent of the effects of early monocular enucleation, suggest that activity-mediated binocular competition underlies the formation of eye-specific retino-dLGN projections. In a subsequent study from the Shatz lab, monocular intravitreal injections of forskolin, a drug that increases the intracellular levels of cyclic adenosine monophosphate (cAMP) were used to increase the frequency of waves in the retina. This resulted in the treated eye's innervating a greater than normal amount of territory in the dLGN. The results of this study extended the concept of activity-mediated binocular competition by showing that the more active eye captured more territory in the target nucleus.

A limitation of the foregoing studies is that they did not differentiate between retinal waves (i.e., correlated discharges of spatially adjacent retinal ganglion cells) and retinal activity *per se*, in that all manipulations designed to block or increase retinal waves also affected the overall discharge levels of retinal ganglion cells. For this reason, a role for correlated retinal activity in the formation of eye-specific retinogeniculate projections remained to be tested.

One approach to resolve this issue is to attempt to perturb correlated activity without affecting the overall discharge levels of developing retinal ganglion cells. During the time that eye-specific projections are being established, retinal waves in the ferret are driven by acetylcholine released by starburst amacrine cells. Chalupa reasoned that partial depletion of these cholinergic interneurons from the developing ferret retina might disrupt the correlated discharge patterns of retinal ganglion cells without changing overall firings levels. A novel immunotoxin was used to deplete a majority of the cholinergic amacrine cells in newborn ferrets, and this treatment did decorrelate the firing patterns of adjacent ganglion cells during the time when eye-specific projections are normally formed (**Figures 1 and 2**). Significantly, the overall discharge rates of ganglion cells in the immunotoxin-treated retinas were not significantly different than normal. **Figure 3** shows the normal development of the retinogenciulate projection, where the left- and right-eye inputs are extensively overlapping at birth and completely segregated by P10. **Figure 4(a)** shows that, surprisingly, the projection patterns in the immunotoxin-treated animal are identical to those seen in the normal animal. This means that decorrelating the discharges of the developing retinal ganglion cells did not affect the normal segregation of retinogeniculate projections. By contrast, in the case of the animal that received binocular injections of epibatadine to block retinal activity (**Figure 4(b)**), it is obvious that segregated

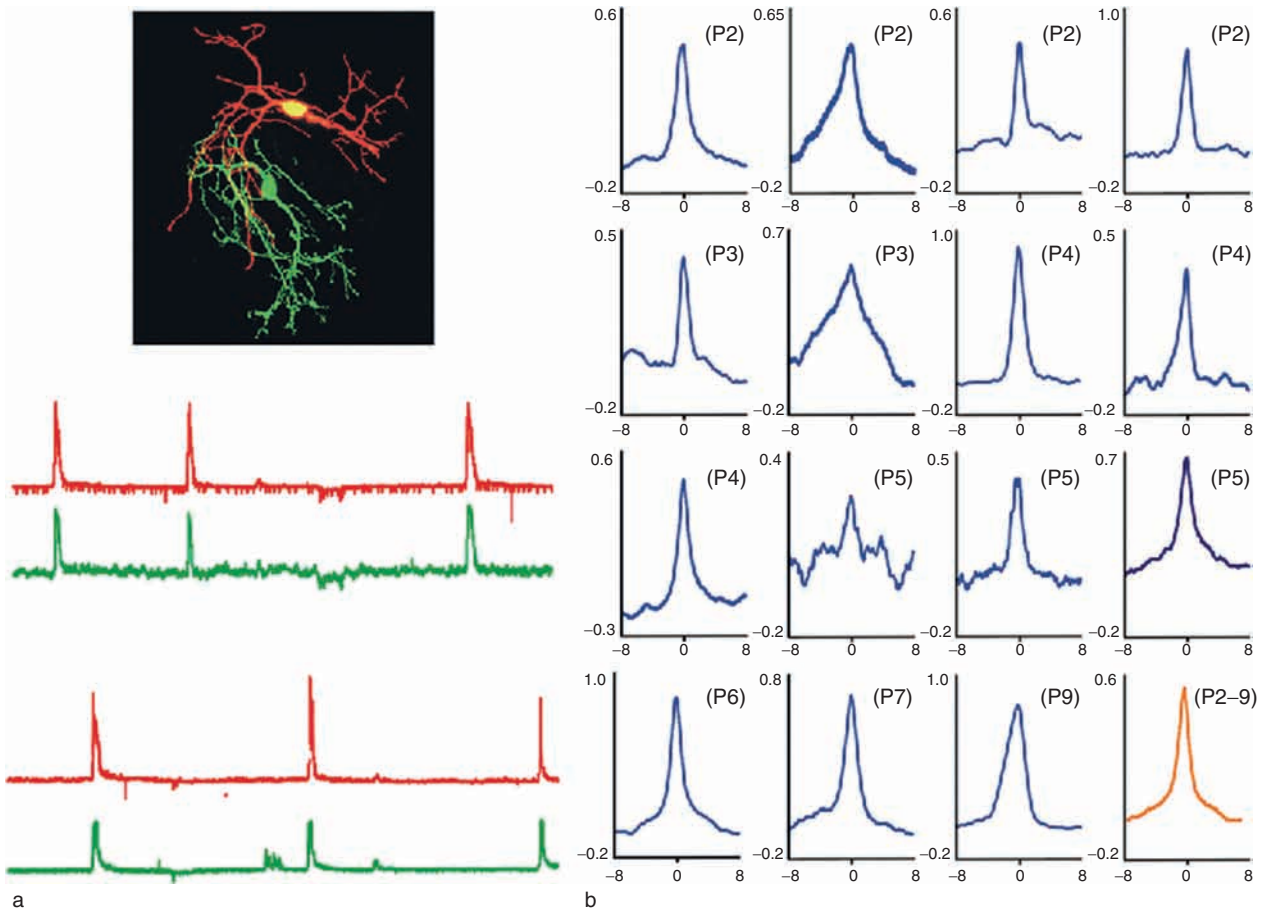

Figure 1 Correlated ganglion cell activity in control newborn ferret retina. (a) Top shows two neighboring ganglion cells recorded with patch-clamp technique and then filled with red and green dyes. Below are shown recordings from two such pairs of cells, in each case the two cells in the pair show correlated bursts of activity. (b) Cross-corelation analyses of 15 pairs of cells recorded in young ferret retinae from postnatal day (P) 2–9 shown in blue and summary data shown in red. In each case, strong correlations are seen in the membrane potentials of neighboring cells indicating correlated activity. Reproduced from Huberman AD, Wang GY, Liets LC, Collins OA, Chapman B, and Chalupa LM (2003) Eye-specific retinogeniculate segregation independent of normal neuronal activity. *Science* 300: 994–998. Reprinted with permission from AAAS.

eye-specific projections failed to form. Therefore, although retinal activity clearly plays a role in eye-specific segregation, normal retinal activity (i.e., retinal waves) is not required.

Another attempt to dissociate retinal waves and the formation of segregated retinogeniculate projections involved studies of the prenatal macaque monkey by the Chalupa lab. In the species commonly used for studies of the developing visual system, such as the ferret and the mouse, the time period during which connections in the visual system are made and refined is relatively brief. Many developmental events occur concurrently in these species, and as a consequence it is difficult to dissociate the links between them. For instance, refinements in retinotopy and formation of eye-specific projections are overlapping developmental events during which retinal activity may play a role in both. By contrast, the developmental period when brain connections become established is much more protracted in the macaque monkey, so it seems reasonable to think that the salient features of visual system development could be more dispersed; that makes it more likely that key events can be dissociated temporally from one another.

Modern neuroanatomical methods were employed to accurately delimit the time period during which eye-specific retino-dLGN projections are established in the fetal macaque monkey. To determine the temporal relationship between the developmental period when eye-specific projections are formed and when correlated discharge patterns are evident in the fetal macaque retina, multielectrode arrays were used to record the activity from isolated retinas of fetal monkeys at known gestational ages. Before embryonic day 55 (E55), the fetal monkey retina was found to be essentially silent, with only a few cells

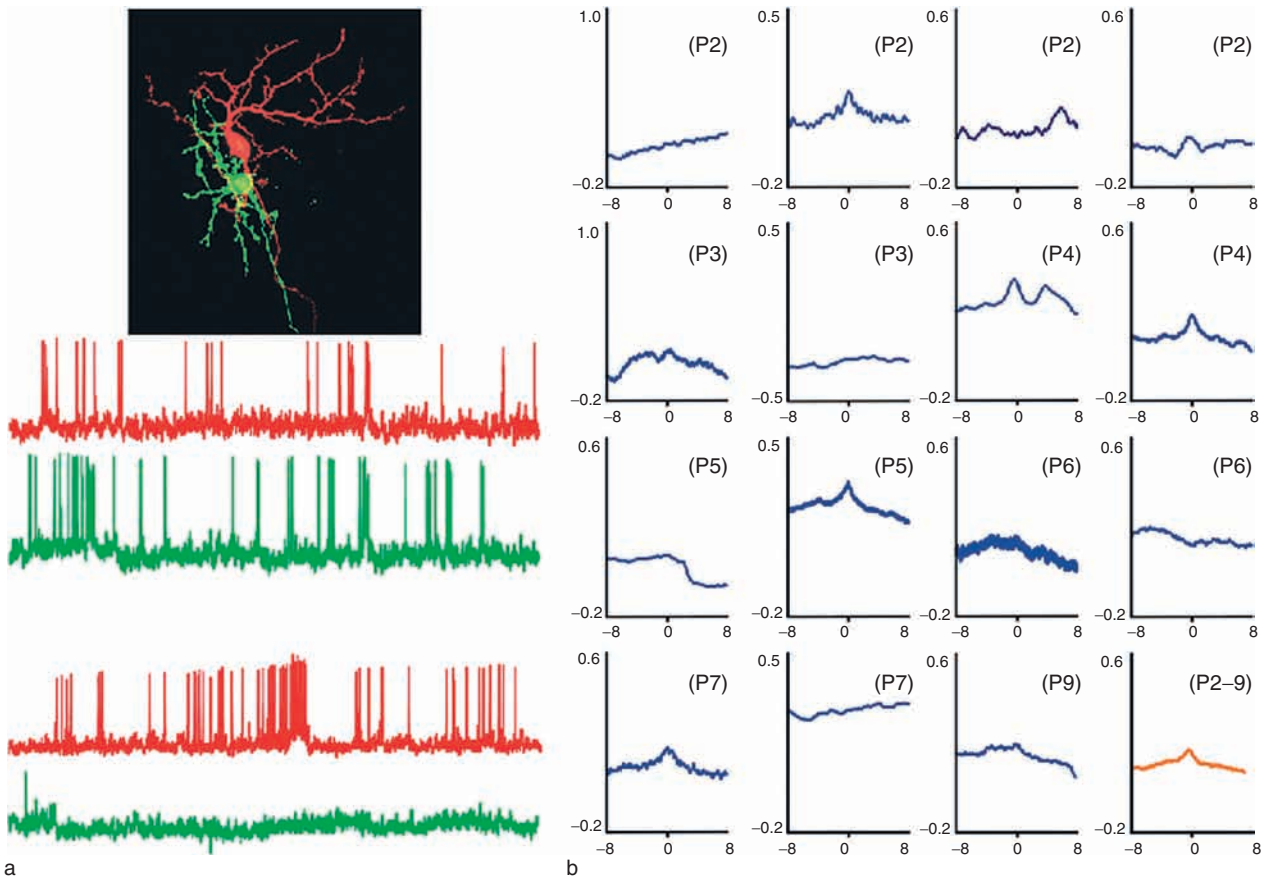

Figure 2 Correlated ganglion cell activity is abolished in immunotoxin-treated newborn ferret retinae. Conventions as in **Figure 1**. Note that recordings from the two pairs of cells shown in the lower part of (a) do not show correlated bursts of activity, and that no significant correlations are seen in the correlation analyses from the 15 cell pairs recorded between postnatal day (P) 2–9 illustrated in (b). Reproduced from Huberman AD, Wang GY, Liets LC, Collins OA, Chapman B, and Chalupa LM (2003) Eye-specific retinogeniculate segregation independent of normal neuronal activity. *Science* 300: 994–998. Reprinted with permission from AAAS.

manifesting occasional discharges. Retinal waves of activity were found to be prevalent at E60, which is more than a week before segregation of retinogeniculate projections is first observed. Moreover, the incidence of retinal waves decreased progressively during the period when eye-specific projections become established (E69–E76). These findings in the fetal macaque monkey differ from the results obtained in other species in which retinal wave activity has been reported to be robust throughout the period when segregated retinogeniculate projections are formed, with such activity declining after this developmental event becomes established. Thus, although the multiarray recordings from fetal monkey retina do not rule out the possibility that retinal activity plays a role in the formation of segregated retinogeniculate projections, they do indicate that retinal waves are most prominent at an earlier time in development than the period when eye-specific inputs are being established. We could still argue, however, that the limited correlated discharges observed during the segregation period are sufficient to trigger this process.

Despite the long history of studies showing the importance of activity in the development of eye-specific segregation in both the LGN and the cortex, there are also many experiments that question the role of activity. The initial experiments of Stryker and Harris showing that intravitreal TTX injections block the initial formation of ocular dominance columns have had to be re-interpreted following imaging studies from the Stryker lab that showed that ocular dominance columns have already formed by the age at which the TTX injections were begun. Therefore, the original experiment did not in fact show that retinal activity blockade prevents the development of ocular dominance columns but, rather, that activity is needed to maintain columns that have already formed. A similar result was found in the dLGN by Chapman, who showed that activity blockade can desegregate eye-specific projections to the dLGN after the layers have initially formed.

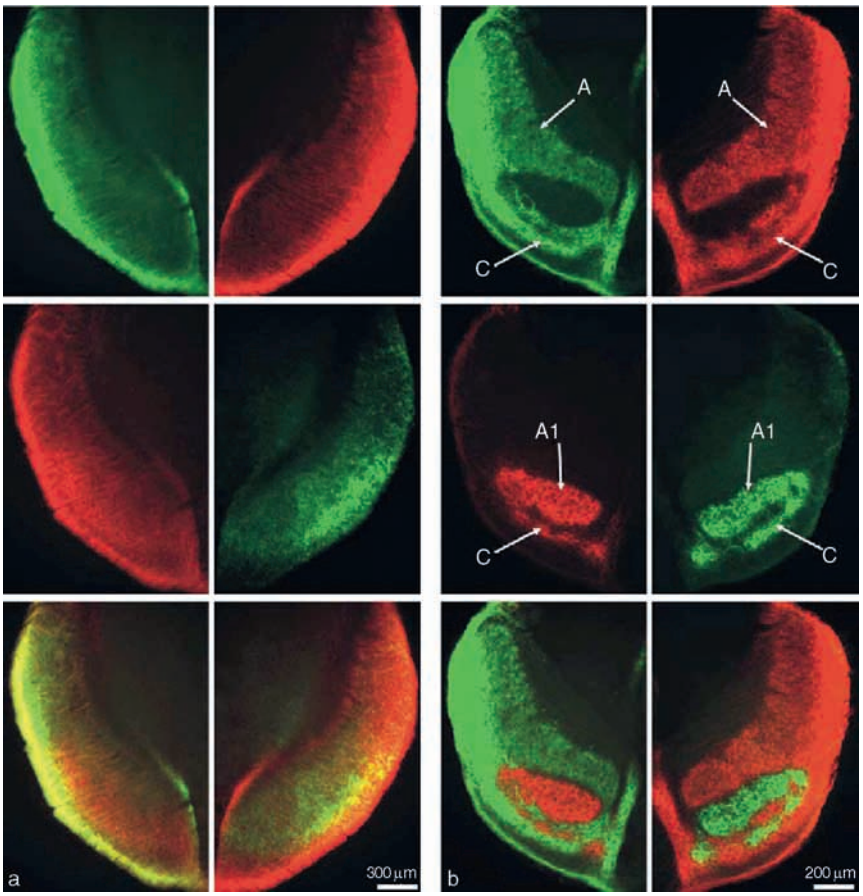

Figure 3 Normal development of retinogeniculate projection in ferret. Retinogeniculate fibers from the each eye are labeled with red or green dye. (a) Left- and right-eye inputs to the LGN are extensively overlapping on postnatal day (P) 1 in a normal ferret. (b) Left- and right-eye inputs are completely segregated by postnatal day (P) 10 in a control ferret treated with intraocular saline injections. A, A1, and C laminae have all developed normally in this control. Top panels in (a) and (b) show contralateral projections, middle panels show ipsilateral projections, bottom panels show merged images of the two. Reproduced from Huberman AD, Wang GY, Liets LC, Collins OA, Chapman B, and Chalupa LM (2003) Eye-specific retinogeniculate segregation independent of normal neuronal activity. *Science* 300: 994–998. Reprinted with permission from AAAS.

The result from the Shatz lab using epibatidine to silence activity in the developing ferret retina (leading to maintained overlap in retinogeniculate projections) may also need reinterpretation following reports from the Stryker lab that (at least in mouse) epibatidine decorrelates activity but does not block action potential firing. In addition, the role of spontaneous activity in the segregation of initially intermingled retinal projections shown in the fetal cat by Shatz and Stryker was readdressed by the Ramoa lab. These investigators placed slow-release polymer Elvax embedded with TTX into the vitreous chambers of one or both eyes of neonatal ferrets. Their main finding was that eye-specific projections are formed in spite of monocular or binocular blockade of retinal activity. The authors concluded "that spontaneous retinal activity modulates the time course of binocular segregation but does not alone account for the segregation of retinogeniculate terminals" (Cook, et al., 1999: 491). A potential shortcoming of the use of TTX to block activity is that it may induce aberrant growth of axon terminals and that this factor, rather than the disruption of activity *per se*, might account for its effects. Indeed, tracings of single retinogeniculate and geniculocortical fibers by the Shatz and Stryker labs have revealed aberrant morphological properties in TTX-treated animals not observed during normal development. Such aberrant growth patterns might be less likely to occur with the gradual and more confined release of the TTX by Elvax, and this might account for the differences between the findings of the two studies.

The binocular competitive interactions regulating the formation of eye-specific projections in the embryonic monkey could reflect diverse cellular events, including the restructuring of initially exuberant retinogeniculate fibers, as has been shown to be the case in the fetal cat by Shatz. In the fetal monkey,

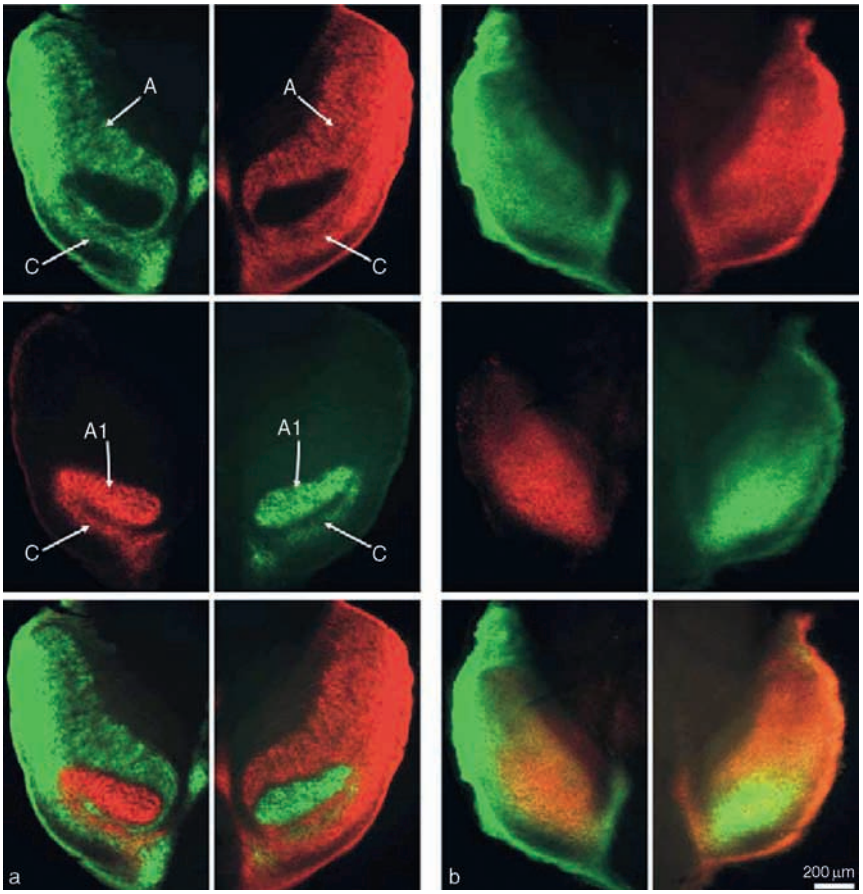

Figure 4 Projection patterns in immunotoxin-treated and epibatidine-treated animals. (a) Projection patterns in immunotoxin-treated animal are identical to those in a control animal (see **Figure 3**); (b) binocular injections of epibatidine used to block retinal waves prevented segregation of eye-specific inputs to the LGN. Conventions as in **Figure 3**. Reproduced from Huberman AD, Wang GY, Liets LC, Collins OA, Chapman B, and Chalupa LM (2003) Eye-specific retinogeniculate segregation independent of normal neuronal activity. *Science* 300: 994–998. Reprinted with permission from AAAS.

however, retinogeniculate arbors have been shown by Chalupa to undergo continuous growth and elaboration from the time when binocular overlap has been reported to be maximum until the segregation of left- and right-eye projections is largely completed. Moreover, the normal growth and elaboration of monkey retinogeniculate fibers is unchanged by prenatal monocular enucleation. This shows that binocular competition does not regulate the size of retinogeniculate arbors in the embryonic monkey. These observations lend support to the hypothesis suggested by Rakic and Chalupa that overlapping binocular projections in the fetal monkey are due to supernumerary retinal ganglion cells which innervate inappropriate regions of the dLGN and that the subsequent segregation of ocular inputs reflects the loss of such retinal fibers. The lack of restructuring observed in individual retinogenculate fibers in the fetal monkey seem more in line with an activity-independent than an activity-dependent model, but this remains to be determined by future studies.

In carnivores and in the monkey, retinal projections to the dLGN are not only segregated by eye, but are also organized in a stereotypic laminar pattern. With the exception of certain mutants, these eye-specific laminar patterns are invariant across all members of a given species. Such an organization cannot be explained solely on the basis of activity-mediated events. Rather, it argues for the early expression of molecular cues guiding the development of eye-specific layers of the dLGN. Such putative molecular cues could preferentially guide axons to a given layer of the dLGN or provide the fibers stemming from one eye or the other an advantage in the competition for postsynaptic cells. Although molecules specific for ipsilateral or contralateral layers have not been identified, recent studies from the Chapman and Feldheim labs revealed that the expression of gradients of ephrins in the dLGN and their receptors in the retina are important for the formation of eye-specific retinogeniculate projections.

An expression of other laminar-specific cues can be inferred from the finding that retinal ganglion cells

in the monkey embryo diverge into M and P subtypes soon after their last mitotic division and selectively innervate either the prospective M or P layers of the developing dLGN. This occurs prior to the time when the Chalupa lab has shown that retinal ganglion cells in the fetal monkey are capable of firing spontaneous discharges, Thus, the formation of segregated M and P functional channels in the primate visual system does not appear to involve activity-mediated events. Rather, this specific ingrowth pattern suggests that M and P classes of retinal ganglion cells in the primate embryo express distinct molecular markers that allow their axons to react differentially to putative laminar-specific cues. This is in accord with earlier reports from Rakic and others showing that monocular and binocular enucleations in the fetal monkey do not perturb formation of M and P layers in the dLGN and in layer 4 of the striate cortex. Unlike the formation of eye-specific projections, the establishment of M and P functional channels in the primate embryo does not appear to be regulated by binocular interactions.

Recent studies from the Katz lab have also challenged the purported role of neuronal activity in the establishment of ocular dominance columns in the visual cortex. These investigators noted that the apparently initially overlapping geniculocortical projections seen in young animals using transneuronal tracing could have been due to an artifact. They therefore employed anterograde tracers injected directly into specific layers of the LGN to examine the organization of genioculocortical projections in normal ferrets during development and in ferrets that were binocularly enucleated very early in life, before there was any possibility of retinal activity influencing inputs to the cortex. In normal ferrets, they found that an injection of tracer into one LGN layer resulted in patchy labeled projections to the cortex reminiscent of ocular dominance columns. They did not see the continuous label in the cortex that would be expected if there was complete overlap of the projections from the two eyes early in development. Unfortunately, they were not able to use two different labels in contralateral and ipsilateral eye layers of the dLGN to show interdigitation of the projections from the two eyes to prove beyond doubt that the patchy connections really are ocular dominance columns. In enucleated animals studied at maturity, there was also evidence for the presence of what appear to be normal ocular dominance columns. The authors concluded that formation of ocular dominance columns "relies primarily on activity-independent cues, rather than on specific patterns of correlated activity" (Crowley and Katz, 1999: 1125). However, it is possible that there are still correlated signals within dLGN layers following enucleation. Therefore, these experiments are still controversial. In addition, no molecular cues distinguishing between cells in the ipsilateral and contralateral dLGN layers have yet been identified despite major efforts to find such molecules that could underlie an activity-independent mechanism of ocular dominance column development.

A recent study from the Chapman lab, on the other hand, showed that early disruption of retinal activity by intravitreal injections of epibatidine has profound effects on the layout of ocular dominance columns visualized with transneuronal labeling. The overall pattern of the columns was altered, and there was an expansion of the contralateral input, resulting in less segregated columns than normal (Figure 5). In addition, cortical cell receptive-field properties were dramatically altered in the treated animals. It is not clear whether the differences between this study and the enucleation study of Crowley and Katz, already described, are due to the different techniques used to label the columns in the two studies or whether altering retinal activity really causes changes in ocular dominance columns whereas enucleation has no effect on columns. Taken together, these finding on the developing visual system raise questions about the prominence of activity-regulated events in the formation of eye-specific projection patterns at the level of the dLGN and in the visual cortex.

What then is the role of retinal waves in the development of the visual system? Refinements in the early projection pattern of retinal axons occur at several levels, from the elimination of gross topographic errors to the reduction of multiple retinal inputs onto single dLGN neurons, and retinal waves could

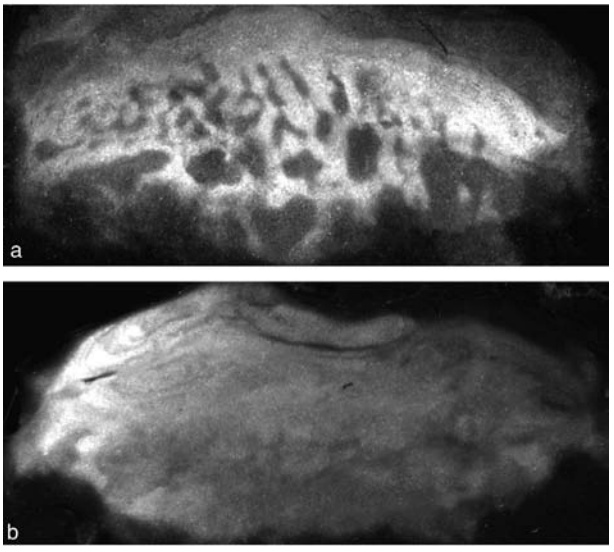

Figure 5 Ocular dominance columns in contralateral primary visual cortex of control (a) and epibatidine-treated ferret (b). Note the expansion of the projection in the epibatidine-treated compared to the control animal.

clearly play a role in one or more of such developmental events. Still another possibility is that retinal waves provide an efficient means for increasing intracellular levels of Ca^{2+} into developing retinal ganglion cells essential for the expression of genes controlling normal axon growth and proliferation, cellular events that are clearly required for retinal segregation. These alternatives are, of course, not mutually exclusive. The challenge for future studies is to establish how different types of neuronal activities interact with a cascade of molecular factors to form the exquisite patterns of connections that characterize the mature visual system.

See also: Activity in Visual Development; Hebbian Plasticity; Plasticity and Activity-Dependent Regulation of Gene Expression; Vision: Light and Dark Adaptation; Visual System: Multiple Visual Areas in Monkeys; Visual Deprivation; Visual Cortical Models of Orientation Tuning.

Further Reading

Chalupa LM (2007) A reassessment of the role of activity in the formation of eye-specific retinogeniculate projections. *Brain Research Reviews* 55: 228–236.

Chalupa LM and Huberman AD (2004) A new perspective on the role of activity in the development of eye-specific retinogeniculate projections. In: Gazzaniga MS (ed.) *The Cognitive Neurosciences III*, pp. 85–92. Cambridge, MA: MIT Press.

Chapman B (2004) The development of eye-specific segregation in the retino-geniculostriate pathway. In: Chalupa LM and Werner JS (eds.) *The Visual Neurosciences*, pp. 108–125. Cambridge, MA: MIT Press.

Crowley JC and Katz LC (2002) Ocular dominance development revisited. *Current Opinion in Neurobiology* 12: 104–109.

Del Rio T and Feller MB (2006) Early retinal activity and visual circuit development. *Neuron* 52: 221–222.

Erzurumlu R, Guido W, and Molnar Z (eds.) (2006) *Development and Plasticity in Sensory Thalamus and Cortex*. New York: Springer.

Hensch TK (2004) Critical period regulation. *Annual Review of Neuroscience* 27: 549–579.

Huberman AD (2007) Mechanisms of eye-specific visual circuit development. *Current Opinion in Neurobiology* 17: 73–80.

Huberman AD, Wang GY, Liets LC, Collins OA, Chapman B, and Chalupa LM (2003) Eye-specific retinogeniculate segregation independent of normal neuronal activity. *Science* 300: 994–998.

Taha SA and Stryker MP (2005) Molecular substrates of plasticity in the developing visual cortex. *Progress in Brain Research* 147: 103–114.

Warland DK, Huberman AD, and Chalupa LM (2006) Dynamics of spontaneous activity in the fetal macaque retina during development of retinogeniculate pathways. *Journal of Neuroscience* 26: 5190–5197.

Visual Motion Detection

B K Dellen, University of Goettingen, Goettingen, Germany
R Wessel, Washington University, St. Louis, MO, USA

© 2009 Elsevier Ltd. All rights reserved.

Visual motion is essential for many behaviorally important functions such as camouflage breaking, detecting looming objects, and depth perception. Camouflage, in which an animal adapts the color and texture patterns on its skin to match those of its habitat, thereby reducing the risk of being detected by preys or predators, is a very effective survival tool. This tactic, however, only works as long as the animal stays still. Once it moves relative to the background that it is hiding in, other animals can use their visual motion detecting apparatus to break the camouflage. Looming visual patterns are produced by a fast-approaching object, and must therefore be detected reliably and well in advance of the expected collision in order to provide the animal with enough time to execute its escape behavior. Translational self-motion perpendicular to the line of sight produces apparent motion of stationary objects on the retina. The magnitude of this motion is related to the depth of the object. Therefore, estimating the apparent motion of different parts of the visual image can aid the brain in constructing the three-dimensional structure of the world. Such knowledge is important, for example, for navigational purposes.

In summary, having access to motion information is instrumental to an animal's chances of survival. However, motion information is not explicitly encoded at the level of the retinal input. Rather, motion has to be computed from the spatiotemporal intensity changes in the retinal image as sensed by the two-dimensional array of photoreceptors. Given the importance of visual motion for biological function, it is perhaps not surprising that all animals with vision have mechanisms for motion processing.

Local Motion Processing

There are basically two local strategies in which a moving intensity pattern can be distinguished from both static intensity patterns and random spatiotemporal intensity fluctuations. (1) The intensity at a particular location and at a given instant of time is compared with the intensity value at a neighboring location measured some time later (Figure 1). (2) The intensity change in time at one location is divided by the intensity change in space at the same location. In the literature the two algorithmic strategies are referred to as the (1) correlation and (2) gradient detectors. Multiple versions of the correlation detector have become popular in the neurobiological literature. The gradient detector has found more acceptance in the computer vision literature.

Objects in our visual environment are normally distinguishable by a number of cues such as luminance, shape, color, and texture, yet the percept of motion is qualitatively independent of the different cues. This cue independence of motion perception is paralleled by a similar cue independence of the responses of many motion-sensitive neurons in the primate visual cortex and avian tectum. Cue independence of motion perception may be a useful operational principle. To perform optimally in a variable environment, a motion analyzer should generalize across stimulus cues, thus encoding the motion of a stimulus regardless of the cue that enables it to be seen. In avian tectum, cue independence of motion computation is mediated by phasic signal transfer at the retinotectal synapse of tectal wide-field neurons. For this wide-field neuron the distinction between static stationary stimuli and dynamic spatiotemporal stimuli, such as motion, is independent of the details of the stimulus.

Our understanding of the cellular mechanisms of direction selectivity is most advanced in the vertebrate retina, where directional selective retinal ganglion cells were first discovered in the 1960s. We now know that a myriad mechanisms, including dendritic and synaptic dynamics and structure, together provide the neural substrate to potentiate the responses of retinal ganglion cells to motion in certain directions.

Global Motion Processing

A comprehensive understanding of visual processing of motion requires one to go beyond local processing of information and to consider the important effect of context. A neuron responds to visual stimuli presented within an area in the visual field, which traditionally is referred to as the 'classical' receptive field. The neuron does not respond to visual stimuli presented outside this area, which is referred to as the 'surround.' Importantly, however, the response to stimuli presented in the classical receptive field may change when additional stimuli are presented in the surround; that is, the response depends on the stimulus context. Efforts to clarify the psychophysics and physiology of contextual effects in visual motion analysis are collectively referred to as studies of global motion processing.

An important current area of research in global motion processing is transparent motion analysis (Figure 2). Motion transparency is a situation in which two overlapping surfaces move past each other, and as a result the same – or very close – retinal locations may contain two different local motion parameters. In order to estimate the motion properties of the two surfaces, local motion signals have to be first segmented, that is, assigned to either surface. Due to the spatial overlap of the surfaces, however, this segmentation process is nontrivial. While in many natural settings nonmotion cues such as contrast, color, and texture can aid the brain with the segmentation task, there are important scenarios, such as camouflage, in which the brain has to rely solely on motion information to partition the visual image into regions representing independently moving surfaces.

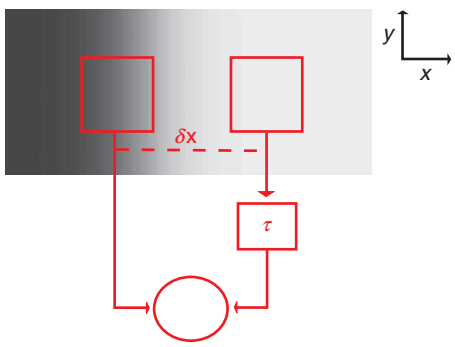

Figure 1 Schematic representation of the correlation-based motion detection. The snapshot of an intensity edge moving with a constant velocity in the *x*-direction is shown. The intensity pattern within a small window of the image is compared with the intensity pattern within another window at a neighboring location some time τ later. If the intensity patterns are sufficiently similar, or ideally identical, the measured velocity is equal to the distance δx between the two windows divided by the time delay τ. The similarity of the intensity pattern may be evaluated via the correlation coefficient.

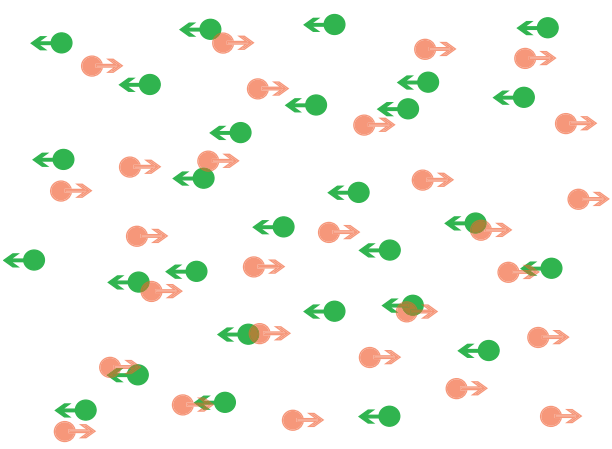

Figure 2 An example of transparent motion. Two overlapping transparent random dot patterns are moving on top of each other in different directions. Hence, the location of a dot does not predict its group identity, and a single point in space might contain two different velocities.

In laboratories, such conditions can be simulated using transparent random-dot patterns, in which small dots distributed randomly across the visual space move in one of two possible directions, thereby constituting two groups. The location of a dot, however, does not predict its group identity. Further, the brain has only a noisy version of each direction of motion available. Thus, statistical methods have to be employed for this unsupervised classification problem. For instance, accurate estimation of the background motion from an image velocity field can be accomplished through an iterative cooperation between two modules: one that specializes in calculating a weighted average velocity and another that calculates a velocity contrast map. Interestingly, this approach finds an analogy in the vertebrate tectum–pretectum system and suggests that the optic tectum is in charge of calculating a motion contrast map, while the pretectum computes background motion, that is, the weighted average velocity.

'Motion repulsion' is a visual illusion that occurs during the perception of transparent motion. Motion repulsion is the perceived enlargement of the angle between the directions of motion of two transparently moving patterns. It has been hypothesized that this illusion is mediated by lateral inhibition between directionally selective units. An alternative explanation, however, relies on the statistical properties of the motion transparency problem alone. Formulating the problem as an unsupervised clustering problem reproduces motion repulsion.

Another important example of global motion processing by neurons is their sensitivity to motion contrast or relative motion between different parts of the visual image (Figure 3). For example, some cells in the deep layers of the avian optic tectum show suppression of response to a moving stimulus in their excitatory receptive field when it is accompanied by a second stimulus that is moving in the same direction, as long as the latter is within the inhibitory surround. Some neurons in the middle temporal cortex of primates show similar properties. The underlying computation requires spatial comparison of visual signals, presumably mediated by lateral and/or feedback connections.

An algorithm of motion contrast computation without motion sensors has been proposed. The algorithm classifies the visual scene into coherent and incoherent motion and determines the relative velocity of the moving parts of the visual scene. Remarkably, the algorithm predicts the measured response to

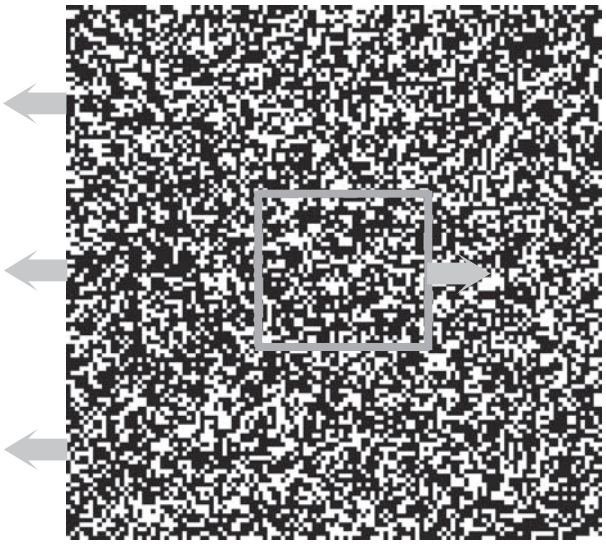

Figure 3 An example of relative motion or motion contrast. An opaque random dot square is moving to the right in front of a random dot background, which is moving to the left. Thus, the random dot square is defined only by its motion relative to the background.

motion contrast as a function of the relative angle of object motion from avian tectum and primate middle temporal area (area MT).

Neural Structure and Motion Processing

It has long been hypothesized that structure and function are correlated in neural systems. The avian retino-tecto-rotundal pathway plays a central role in motion analysis and features complex connectivity. While the retinal ganglion cells report local image contrast, the tectal wide-field neurons fire more vigorously when facing small moving stimuli. Postsynaptic neurons in the thalamic nucleus rotundus are selective for the direction of motion of an object as well as looming patterns.

These differences in representations of visual stimuli are paralleled by qualitative structural differences of the neuronal elements in the three stages (Figure 4). A dense array of retinal ganglion cell axon terminals form a topographic map and synapse onto dendritic endings of tectal wide-field neurons. In contrast, tectal wide-field neurons have large circular dendritic fields with sparsely distributed dendritic endings. Interestingly, the point-to-point topography of the retino-tectal projection is replaced by a more abstract topography in the tecto-rotundal projection that appears to be 'interdigitating,' that is, a specific locus in the nucleus rotundus receives input from a sparse population of tectal wide-field neurons throughout the entire tectum.

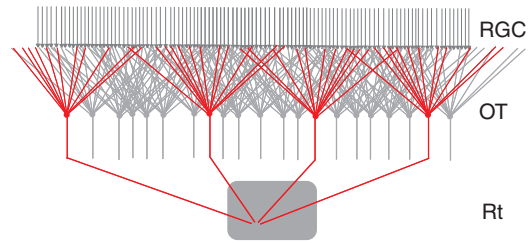

Figure 4 A schematic of the avian retino-tecto-rotundal connectivity pattern. Retinal ganglion cell (RGC) axons project to the dendritic endings of tectal (OT) wide-field neurons. The tectal wide-field neurons in turn project with an interdigitating topography to the nucleus rotundus (Rt).

A computational investigation revealed that the tectal wide-field neuron's structural properties enhance the neuron's sensitivity to change, a behaviorally important stimulus attribute, while preserving information about the stimulus location in the tectal wide-field neuron population activity. The computational investigation showed further that the subsequent interdigitating tectal projection to rotundal neurons allows the decoding of the direction of motion from the sparsely sampled population of tectal wide-field neurons in a two-stage process. However, the study could not address why the tecto-rotundal projection evolved toward sparse sampling, as opposed to a one-stage projection with dense sampling, which produces similar computational properties. Interestingly, however, a wiring efficiency argument provides a plausible constraint to bias the evolution toward the sparse tecto-rotundal projection approach for motion estimation.

It is tempting to speculate that a computational rationale for sparse connectivity similar to the one hypothesized for the retino-tecto-rotundal pathway will be found for other prominent computational pathways, such as the mammalian cerebral cortex, where only a small fraction of all possible connections between neurons physically exists, even within a local area.

Motion Processing in Visual Cortex

The geniculo-cortical pathway of the mammalian brain plays an important role in the processing and extraction of visual motion. The majority of retinal ganglion cells ascend via the dorsal part of the two lateral geniculate nuclei (LGN) of the thalamus to the primary visual cortex (V1). Then the processing stream continues to higher visual areas, including area MT. Many neurons in V1 and MT are sensitive to motion, and lesions of these areas severely reduce the ability of the animal to detect motion. These observations suggest that the geniculo-cortical pathway is highly involved in visual motion analysis.

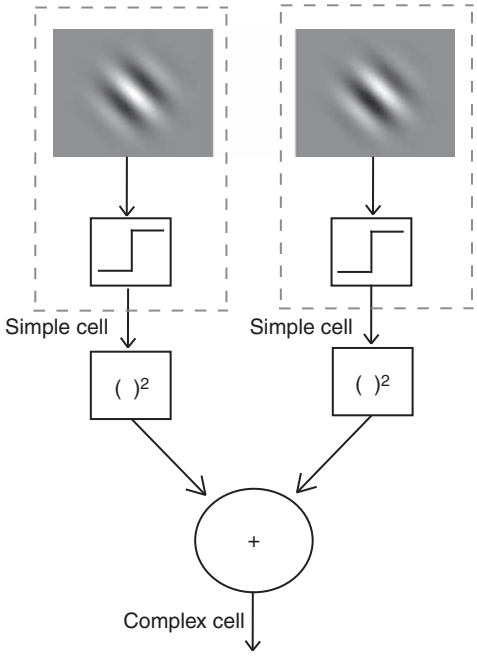

Figure 5 Schematic of the receptive-field structure of a cortical simple and complex cell. Each simple cell filters the visual input by means of an oriented spatiotemporal sine function multiplied with a Gaussian envelope. Then the output of the filters is rectified to ensure positive-valued responses. The responses of two simple cells whose sine functions are phase shifted by $\pi/2$ are first squared and then added to create the complex cell response.

Traditionally, neuroscientists have attempted to classify V1 neurons into two types (Figure 5). In this classification scheme, one type, the so-called 'simple cell,' possesses periodically arranged and oriented excitatory and inhibitory areas, which make them sensitive to the spatial phase, spatial frequency, and orientation of a drifting grating. The other type, the so-called 'complex cell,' is largely insensitive to the spatial phase of a drifting grating. To date, much research has been based on the working hypothesis that visual cortical neurons are hierarchically organized with simple cells feeding into complex cells. It is important to note, however, that the experimental evidence for or against distinct simple and complex cell classes and its hierarchical arrangement has been substantial but not conclusive.

The product of an oriented sine function with a Gaussian envelope provides a useful model to reproduce the directionally selective receptive field of simple cells. In the next step, model simple cell responses are nonlinearly combined to give complex cell responses. These responses represent the projections of the stimulus velocity vector on different axes and need to be combined to obtain meaningful estimates of the full stimulus velocity. The latter computation may take place in area MT. Neurons in area MT show a strong directional selectivity and are possibly involved in the processing of global motion patterns and figure/ground segregation.

Interestingly, a mathematical analysis showed that local relative motion computation is inherent in the nonlinear transformation from simple to complex cell responses. Simulations that quantitatively reproduce measured complex cell responses in both cat and monkey to a variety of relative motion stimuli support the mathematical inference. Importantly, this computation of relative motion does not require an intermediate neural representation of local velocities and does not require lateral or feedback interactions within a network.

The hierarchical model of visual cortical motion processing is in good agreement with a large body of experimental data. However, it is a local model. Velocities are extracted from a small area of the visual field only. For the simultaneous presentation of multiple stimuli, the hierarchical model loses most of its predictability. Already in V1, neurons exhibit sensitivity to orientation and motion contrast between a stimulus in the classical receptive field and the surround, potentially contributing to perceptual pop out and image segmentation. Furthermore, recent studies suggest that cortical processing is less local than previously thought. These properties of the visual cortex might be attributed to the abundant lateral and feedback connections, which innervate the cortical layers. Less than 5% of excitatory synapses on V1 neurons originate from the LGN, while over 95% of the excitatory synapses, even in the geniculo-recipient layer in area V1, come from other cortical neurons and other nuclei. For instance, feedback from higher visual areas may be a prime candidate to mediate contextual effects.

Contextual influences from outside the classical receptive field may fundamentally change our view of visual cortical motion processing. The challenge ahead is to design global theories of the brain, in which locality is compatible with global processing strategies.

See also: Optic Tectum: Development and Plasticity; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vision: Mechanisms of Orientation, Direction and Depth; Visual Motion Models; Visual Cortical Models of Orientation Tuning; Visually Guided Behavior.

Further Reading

Albright TD and Stoner GR (2002) Contextual influences on visual processing. *Annual Review of Neuroscience* 25: 339–379.

Andersen RA (1997) Neural mechanisms of visual motion perception in primates. *Neuron* 18: 865–872.

Borst A and Egelhaaf M (1989) Principles of visual motion detection. *Trends in Neurological Sciences* 12: 297–306.

Clifford CWG and Ibbotson MR (2003) Fundamental mechanisms of visual motion detection: Models, cells and function. *Progress in Neurobiology* 68: 409–437.

Dellen BK, Clark JW, and Wessel R (2004) Motion-contrast computation without directionally sensitive motion sensors. *Physical Review E* 70: 031907.

Dellen BK, Clark JW, and Wessel R (2005) Computing relative motion with complex cells. *Visual Neuroscience* 22: 225–236.

Egelhaaf M, Kern R, Krapp HG, Kretzberg J, Kurtz R, and Warzecha AK (2002) Neural encoding of behaviorally relevant visual-motion information in the fly. *Trends in Neurological Sciences* 25: 96–102.

Giese MA and Poggio T (2003) Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience* 4: 179–192.

Hildreth EC and Koch C (1987) The analysis of visual motion: From computational theory to neuronal mechanisms. *Annual Review of Neuroscience* 10: 477–533.

Luksch H, Khanbabaie R, and Wessel R (2004) Synaptic dynamics mediate sensitivity to motion independent of stimulus details. *Nature Neuroscience* 7: 380–388.

Mahani AS, Carlsson AE, and Wessel R (2005) Motion repulsion arises from stimulus statistics when analyzed with a clustering algorithm. *Biological Cybernetics* 92: 288–291.

Mahani AS, Khanbabaie R, Luksch H, and Wessel R (2006) Sparse spatial sampling for the computation of motion in multiple stages. *Biological Cybernetics* 94: 276–287.

Mahani AS and Wessel R (2006) Iterative cooperation between parallel pathways for object and whole-field motion. *Biological Cybernetics* 95: 393–400.

Nakayama K (1985) Biological image motion processing: A review. *Vision Research* 25: 625–660.

Palmer SE (1999) *Vision Science*. Cambridge, MA: MIT Press.

Sereno ME (1993) *Neural Computation of Pattern Motion*. Cambridge, MA: MIT Press.

Smith AT and Snowden RJ (eds.) (1994) *Visual Detection of Motion*. London: Academic Press.

Taylor WR and Vaney DI (2003) New directions in retinal research. *Trends in Neurological Sciences* 26: 379–385.

Wandell BA (1995) *Foundations of Vision*. Sunderland: Sinauer.

Watanabe T (ed.) (1998) *High-Level Motion Processing*. Cambridge, MA: MIT Press.

Zanker JM and Zeil J (eds.) (2001) *Motion Vision: Computational, Neural, and Ecological Constraints*. Berlin: Springer.

Visual Motion Models

A Borst, Max-Planck-Institute of Neurobiology, Martinsried, Germany

© 2009 Elsevier Ltd. All rights reserved.

Whenever an animal is moving in its environment, the images of the environment shift on the animal's retina. The resulting spatial distribution of motion vectors across the visual field is called 'optic flow.' Since the optic flow depends on the animal's own movement as well as on the structure of the environment, it is widely used for visual navigation and course control. However, motion information is not explicitly represented in the two-dimensional (2-D) brightness pattern of the retinal image as sensed by the array of photoreceptors. Instead, motion has to be computed from the brightness changes in the retinal image, as one of the first and most basic processing steps in the visual system. Understanding this primary process of motion detection is important because the performance of all subsequent processing steps depends on the representation of optic flow information at their input stages.

Algorithmic Models of Motion Detection

The velocity of a moving object is defined as the object's spatial displacement over time, dx/dt. For the visual system, this results in a brightness distribution at one time, which is displaced at some time later. There are basically two ways in which motion can be detected visually. First, the brightness at a particular location and at a given instant in time is correlated with the brightness value at a neighboring location measured some time later. Second, the brightness change in time observed at one location is related to the brightness change in space at the same location. These two strategies of motion detection form the basis of two classes of model mechanisms, the so-called correlation and the gradient detectors. Various versions of correlation and gradient detectors have been proposed in the past to underlie motion detection in biological systems.

The correlation detector was first proposed on the basis of experimental studies on the optomotor behavior of insects by Reichardt and Hassenstein. According to common usage, this correlation detector will be referred to here as the Reichardt detector. In subsequent studies, the Reichardt detector has also been applied to explain motion detection in different vertebrate species including humans. Such a detector consists of two mirror-symmetrical subunits (Figure 1(a)). In each subunit, the signals derived from two neighboring inputs are multiplied with each other after one of them has been shifted in time with respect to the other by a delay line or some sort of temporal low-pass filter. The final detector response is given by the difference of the output signals of both subunits. Although each detector subunit is directionally selective on its own, it responds also to correlated input signals which are independent of the direction of motion, such as to the background brightness or its temporal changes. Since these direction-independent signals are the same in both mirror-symmetrical subunits of the complete detector, they are eliminated by subtracting the two subunit outputs from each other. The subtraction stage, therefore, increases the direction selectivity of the movement detector. Various elaborations of the basic Reichardt model have been proposed to accommodate this motion detection scheme to the particular performance of different biological systems. For instance, the retinal input signals may pass linear spatial and/or temporal filter or nonlinear processing stages, such as saturation, full-wave, or half-wave rectifier.

In biology, a popular variant of the correlation-type movement detector has been proposed by Barlow and Levick to explain their experimental findings on directionally selective ganglion cells in the rabbit retina. With respect to its layout, the Barlow–Levick model is almost identical to one subunit of the basic Reichardt model. It consists of two input lines carrying the brightness signals which are compared after one of the signals has been delayed. In contrast to the Reichardt model, this comparison is accomplished by a special logical gate, an 'and-not' or 'veto' gate. This means that the detector's activity is suppressed when both input signals arrive simultaneously at the and-not gate. The corresponding direction of motion is therefore the detector's null direction. For motion in the detector's preferred direction, the veto signal arrives too late to have an effect.

Another type of correlation detector is the so-called motion energy model. This model was inspired by the fact that visual motion corresponds to an orientation in the space–time domain. Constructing sensors with oriented spatiotemporal receptive fields, squaring their output and combining elements such that their responses to grating patterns are 90° out of phase ('quadrature pair'), their summed output shows a steady amplitude during motion with a constant velocity. Since the output of such a squared quadrature pair extracts a measure of the motion energy, these models are called motion energy models. If the Reichardt model as shown in Figure 1(a) is equipped

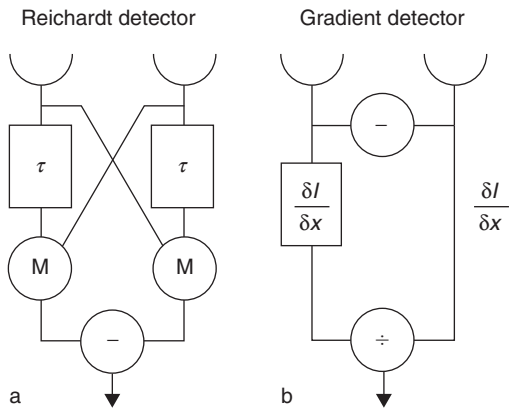

Figure 1 Algorithmic models. (a) Reichardt model. The detector consists of two mirror-symmetrical subunits. In each of the subunits, the detector input signals are multiplied (M) after one of them has been delayed (τ) with respect to the other by some sort of temporal filter. Both subunit outputs are subtracted to give the final detector output signal. (b) Gradient model. This detector computes the ratio of the temporal and spatial change of the pattern brightness. In the model shown here, the temporal gradient is approximated by a temporal high-pass filter and the spatial gradient is approximated by the difference between the two input signals. The mean of both high-pass filtered input signals is used in the numerator of the division operation. Note that the performance of this simple hardware-implemented gradient scheme differs substantially from the ideal mathematical one.

with the same spatial and temporal filters in its input channels, it assumes the same specific functional characteristics as the energy model and becomes completely equivalent. This identity, however, only holds for the final, fully opponent output signal of both detectors. The responses of the more peripheral processing stages may differ considerably.

With $F = F(x, t)$ designating the brightness of a stimulus pattern as a function of one spatial coordinate x and time t, gradient detectors calculate a motion estimate by relating the spatial change of brightness ($\partial F/\partial x = F_x$) to the temporal change ($\partial F/\partial t = F_t$) at a given location of the image (**Figure 1(b)**). These changes are physically related to each other by the velocity. The temporal brightness change F_t induced by a moving object at point x is proportional to the spatial brightness change F_x as well as to the velocity $v = \mathrm{d}x/\mathrm{d}t$ (note: leftward and rightward motion are conventionally defined as negative and positive velocities, respectively):

$$-F_t = F_x \cdot v \qquad [1]$$

The velocity can thus be formally recovered at least for motion in one spatial dimension by dividing the temporal brightness change by the spatial gradient:

$$v = \mathrm{d}x/\mathrm{d}t = -F_t/F_x \qquad [2]$$

This formalism relates the spatial and temporal brightness changes induced by a moving object but does not specify how the temporal and spatial change of the image brightness should be measured. There are various ways in which this can be done. The spatial pattern gradient may be approximated in the simplest way by the difference between the brightness at two adjacent image points and the temporal derivative by the temporally high-pass filtered brightness signal (**Figure 1(b)**). The response properties of this simple type of hardware-implemented gradient detector differ considerably from the ones of the 'ideal' mathematical scheme. However, more sophisticated approaches are available in computer vision to approximate spatial and temporal brightness changes. The performance of gradient models based on such approximations may then come much closer to the properties of the formalism given by eqn [2] and, thus, to an ideal gradient scheme.

As a consequence of the divisive nonlinearity which is characteristic of gradient detectors, the velocity estimate may become indefinite. This can happen when the denominator in eqn [2] becomes zero (e.g., when the pattern is locally homogeneous). Appropriate means have been proposed to overcome this problem, including the addition of a small constant in the denominator of the ratio between the approximations of the temporal and spatial brightness gradients.

The Marr–Ullman model circumvents this problem by using an 'and' gate rather than a division to relate an estimate of the temporal and spatial brightness changes. Hence, in this respect it is similar to the Barlow–Levick model and, thus, to a correlation detector. The Marr–Ullman model is based on two processing steps. In the first step, it determines the location of edges (i.e., the steepest local brightness changes) in the retinal image. By relating their polarity to the corresponding temporal brightness change, in a second processing step the direction of pattern motion can be determined.

Cellular Models

The motion detection models described so far belong to the group of so-called algorithmic models. They describe the different processing steps by formal operations such as convolutions, subtractions, and multiplications without referring to the specific properties of the biological substrate by which these computations are realized in the nervous system. In contrast, cellular models start with the physiological properties of neurons, synapses, etc. and try to build from these a neuronal machinery for motion vision. Of course, physiological models are still models and, for this reason, have a formal character: When

describing the action, for instance, of a synapse, they make use of electrical circuits as an equivalent for the neuronal membrane. Hence, 'physiological models' are formal models in the same way as algorithmic models are formal models: they are settled just one step closer to the biological substrate.

Many components of the algorithmic motion detection models can be implemented quite easily in neuronal terms. Some of them have even been inspired by the properties of different types of visual interneurons. This is true, most notably, for the different types of spatial and temporal filters that were proposed in the input channels of various versions of correlation detectors or the Marr–Ullman model. It is more demanding to account for the nonlinear interaction between the input signals of the movement detector in neuronal terms. Mainly two types of proposals have been made regarding how this processing step might be realized in biological movement detectors.

The shunting inhibition model (**Figure 2(a)**) has been considered as a physiological means to account for the and-not gate of the Barlow–Levick model. In this model, signals of the two input lines of the movement detector are thought to alter in a postsynaptic neuron the conductances of different ions, an excitatory and an inhibitory one. During motion in the preferred direction, signals are transmitted by the excitatory channel, whereas during motion in the null direction the action of the excitatory synapse is shunted by the simultaneous activation of the inhibitory synapse. A multiplication, the nonlinear interaction of the Reichardt model, is approximated by the mechanism of shunting inhibition if two conditions are met: (1) the reversal potential of the inhibitory ion is close or equal to the resting potential of the postsynaptic cell, and (2) the input signals are small compared with the leak conductance of the postsynaptic cell. If these conditions are not satisfied, considerable deviations from a quadratic nonlinearity may appear.

The threshold model is another cellular possibility to account for the nonlinear interaction between the detector input signals (**Figure 2(b)**). A threshold operation is performed by any neuron which transmits information along its axon by means of action potentials. The threshold model proposes that both input lines of the detector induce excitatory postsynaptic potentials in a common follower neuron which may superimpose in a linear way. If the resulting membrane potential exceeds a given threshold, a spike will be triggered. Since the probability of coincidence of two events is the product of the probabilities of each single event alone, the output of this type of threshold model is equivalent, on average, to a multiplication – the core of the Reichardt model.

Both of these cellular models rely on the most common properties of neurons, such as on changes

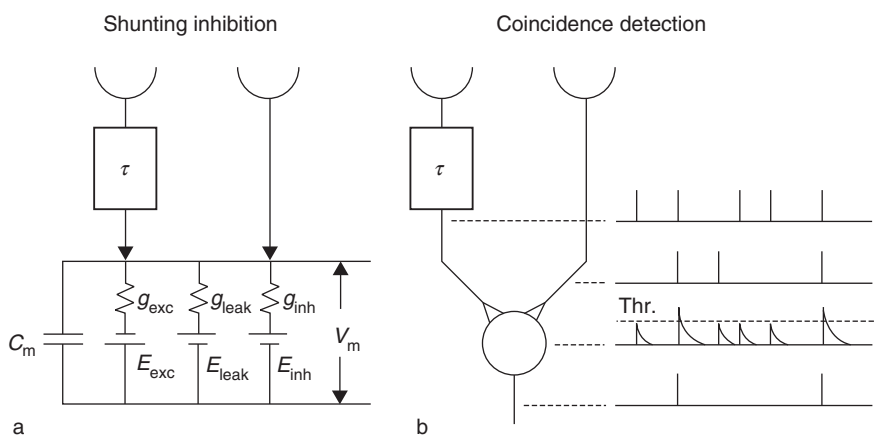

Figure 2 Cellular models. (a) Shunting inhibition model. The input signals derived from adjacent image points are assumed to change the conductances of different ion channels of a postsynaptic cell after one of the input signals is delayed with respect to the other. The postsynaptic potential can be quantitatively determined according to the electrical equivalent circuit:

$$V = \frac{E_{leak}g_{leak} + E_{exc}g_{exc} + E_{inh}g_{inh}}{g_{leak} + g_{exc} + g_{inh}}$$

where E_{leak} is the leak potential; E_{exc} and E_{inh} are the reversal potential of the excitatory and inhibitory ion, respectively; g_{leak} is the lumped leak conductances; and g_{exc} and g_{inh} are the conductances of the excitatory and inhibitory ion, respectively. As can be seen by a Taylor expansion of this expression, the resulting postsynaptic potential approximates a multiplication of the input signals g_{exc} and g_{inh} if these are small compared with the leak conductance g_{leak} and if the reversal potential of the inhibitory ion E_{inh} is close to the leak potential E_{leak}. (b) Threshold model. The action potentials of two cells induce excitatory potentials in a postsynaptic cell. When arriving alone, the excitatory postsynaptic potentials (EPSPs) are too small to trigger an action potential. Only when two spikes arrive simultaneously are the EPSPs superimposed, exceed the threshold (Thr., dashed line), and give rise to an action potential. Note that the width of the temporal window within which the superposition of the two EPSPs causes an action potential is set by the time course of the EPSPs.

in their input resistance or on spike thresholds. A number of more complex nonlinear synaptic interactions have been characterized electrophysiologically in various systems which may potentially qualify for the nonlinear interaction in biological movement detectors. Most progress has been made in the rabbit retina, where directionally selective calcium accumulation has been visualized in the dendrites of so-called 'starburst amacrine cells.' The authors observed that only motion from the cell body out to the distal ends of the dendrites leads to an increase in dendritic calcium, whereas inward motion failed to do so. Further studies in these cells provided a biophysical model which can account for this phenomenon. The model is based on voltage-gated calcium channels and a tonic voltage gradient in the dendrite, which provides the spatial asymmetry necessary for direction discrimination.

Common Properties of Motion Detector Models

Despite the apparent differences among the movement detector models discussed previously, all of them satisfy three requirements which any motion detection mechanism has to meet in order to signal motion in a directionally selective way. First, they have at least two input channels. A single input channel could not distinguish, for instance, a dark bar crossing its receptive field in one direction from a dark bar moving in the opposite direction or from a transient dimming of the light. For the correlation method, in its basic form two spatially separate measurements form the signals which are correlated; for the gradient method, at least two spatially separate measurements have to be used to approximate the spatial gradient. Second, the processing of the movement detector input signals has to be asymmetrical in some way. If it were symmetrical, the detector input channels could be interchanged without affecting the output. In this case, the detector would no longer be directionally selective. In the different versions of correlation detectors, the asymmetry is due to the fact that the detector input signals which are correlated are temporally filtered in a different way. In gradient detectors, the asymmetry is inherent in the determination of the spatial brightness gradient. Third, to compute motion, the input channels of a detector have to be related to each other in a nonlinear way. Otherwise, the mean output of a detector would be equal to the detector's response to its averaged input signals; in the averaged input signals, however, all information about the temporal sequence is lost. Thus, the result of a linear interaction cannot be directionally selective. The logical gates used by the Barlow–Levick and Marr–Ullman models, as well as the different analog operations such as the multiplication in the Reichardt model, the squaring operation in the energy model, and the division used by the gradient detector, represent such nonlinear interactions.

Performance of Motion Detectors

An ideal motion sensor should monitor the velocity, for instance, like the speedometer of a car: within its operating range, the higher the velocity, the higher should be the output signal. Yet, the Reichardt detector behaves differently. Many of the response features summarized in the following discussion were also found in motion-sensitive neurons, making the Reichardt detector one of the most successful models in computational neuroscience.

Since periodic stimuli have played a dominant role in the study of biological motion vision, we consider the steady-state response of the Reichardt to a traveling sine wave first. Denoting the wavelength of the sine wave λ, its contrast ΔI, its velocity v (°/s), the spacing of the inputs to the Reichardt detector (its sampling base) $\Delta\varphi$, and the amplitude and phase response of the temporal filter A and Φ, respectively, we obtain the following expression of the time-averaged detector response:

$$R = \Delta I^2 \cdot A(2\pi v/\lambda) \cdot \sin(-\Phi(2\pi v/\lambda)) \cdot \sin(2\pi\Delta\Phi/\lambda) \quad [3]$$

If we assume the temporal filter to be a low-pass of first order, we obtain

$$R = \Delta I^2 \cdot \frac{\tau 2\pi v/\lambda}{1+\tau^2(2\pi v/\lambda)^2} \cdot \sin(2\pi\Delta\Phi/\lambda) \quad [4]$$

The steady-state response of a motion detector can thus be seen, in general, not to be proportional to stimulus velocity. In addition, the response depends on many stimulus parameters in a nontrivial way.

Let us first consider its velocity dependence, which is shown in Figure 3(a) for three different pattern wavelengths. First, note that the detector response, unlike a speedometer, does not increase linearly with increasing pattern velocity. Rather, it exhibits an optimum at a certain velocity. For velocities beyond that optimal velocity, the response falls off gradually toward zero. The speed at which the response is optimum is set by the time constant of the detector low-pass filter: the larger the time constant, the smaller the velocity at which the response is maximum. The time constant of the detector filter thus becomes a decisive model parameter that sets the operating range of the motion detection system in the velocity domain.

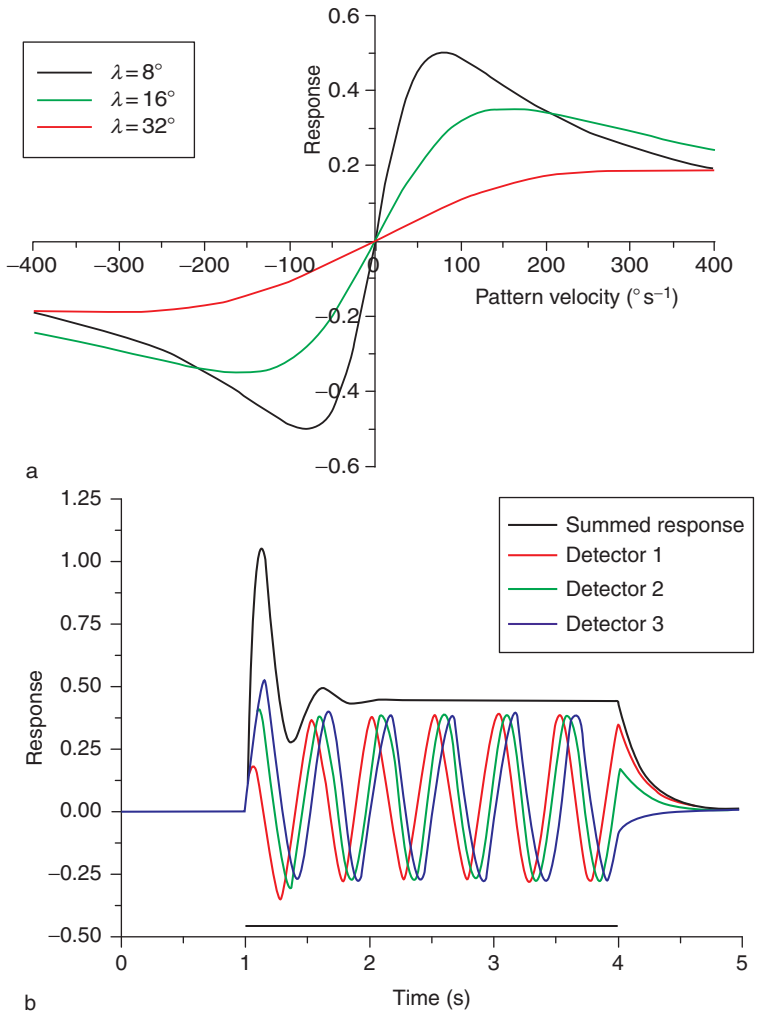

Figure 3 (a) Steady-state responses of the Reichardt detector as a function of pattern wavelength assuming a sampling base of 2°. (b) Dynamic response profile of the Reichardt detector in response to the on- (1 s) and off-set (4 s) of constant pattern motion. Shown are the summated output signals (black) of an array of eight such motion detectors (sampling base = 2°) covering one spatial wavelength ($\lambda = 16°$) of the pattern, together with the output signals of three neighboring individual detector signals (red, green, and blue, respectively). The pattern was moving at $32° s^{-1}$, resulting in a temporal frequency of 2 Hz.

In particular, it determines the slope of the detector response around velocity zero and, thus, the gain of the sensor. Furthermore, the response optimum can be seen to depend on the pattern wavelength: The larger the wavelength, the higher the pattern velocity at which the response becomes optimum. More precisely, the response optimum is directly proportional to the pattern wavelength; doubling the wavelength leads to a doubling of the optimum velocity. If we introduce the temporal frequency of the stimulus as being the ratio of the velocity and the pattern wavelength, the response curves for all different pattern velocities coincide at one frequency.

Furthermore, the response of a Reichardt detector also depends on the square of the pattern contrast. In motion-sensitive neurons of many animals, this contrast dependence has indeed been experimentally confirmed using low stimulus contrasts. For higher contrasts, the experimental data fail to follow a quadratic contrast dependence but, rather, tend to saturate. This can be explained most easily by saturation nonlinearities inside the detector. Another feature of correlation detectors, which is obvious in the previously discussed formulas, is that the steady-state response is proportional to the sine of the ratio of the sampling base and the pattern wavelength. The sampling base of the motion detector limits the spatial resolution of the system. As is the case for any discrete sampling system, wavelengths can only be resolved up to a certain limit. The smallest wavelength that can be resolved is given by twice the sampling interval (the 'Nyquist limit'). For smaller wavelengths, the response starts oscillating between positive and negative values. In this regime, when the

detector signals negative values for pattern motion along its preferred direction, its behavior is called 'spatial aliasing.' It can indeed be experimentally observed in the laboratory, but it usually does not affect the system performance since higher spatial frequencies are much reduced by the optics at the front end of the detector.

In addition to the steady-state properties, the dynamics of the local motion detector response are of great interest. **Figure 3(b)** shows the time course of the detector response to a moving sine grating, the velocity of which was stepped from zero to a constant value (corresponding to a temporal frequency of 2 Hz) at $t = 1$ s and back to zero at $t = 4$ s. Shown is the spatially summated output of all eight motion detectors used in this simulation together with the individual output signals of the first three motion detectors. There are several notable observations. First, the spatially summated response changes over time in the initial period of the stimulation. Before assuming its steady-state response given by eqn [4], the response is transiently ringing at the temporal frequency of the stimulus. The ringing is damped and settles over time to the steady-state value. The length of the transient period is in general determined by the time course of the intrinsic filters of the motion detection system. Another feature that can be seen in **Figure 3(b)** pertains to the individual detector signals. In contrast to the overall response that exhibits a ringing only initially, the individual detector output signals continue to oscillate at the temporal frequency of the pattern as long as the pattern is moving. The modulations of these signals are phase shifted due to the fact that they receive input from different spatial parts of the pattern. Only after spatially integrating the signals over different spatial phases do these oscillations become canceled, resulting in a smooth, constant steady-state response.

Due its absolute equivalency, all the characteristics discussed for the Reichardt detector also hold true for the energy model, with one exception: in the case of the energy model, the temporal modulations (**Figure 3(b)**) can already be eliminated at the level of a detector subunit if the signals are filtered in such a way that the filter responses to sine-wave gratings moving at a constant velocity differ in their phase by exactly one-quarter of a temporal cycle (i.e., they are in 'quadrature phase'). The same effect occurs in the Reichardt detector without prefiltering when the spatial wavelength is four times the sampling base. In general, gradient detectors should not exhibit any sort of pattern dependence or deviate from an ideal velocity sensor. This can be most easily seen for its lack of contrast dependence. Although a pattern with higher contrast leads to a larger spatial gradient, the temporal brightness change becomes larger accordingly. Thus, when dividing spatial and temporal brightness changes, the contrast factors out. For the same reason, a gradient detector does not show modulated responses to pattern motion with a constant velocity; instead, the time course of its response immediately reflects the time course of pattern velocity. However, these properties may differ if the gradient scheme is realized by technical or biological hardware. Due to the fact that both the spatial gradient and the temporal brightness change have to be approximated by certain filters and care has to be taken to prevent the spatial gradient from being zero (the denominator of the division in eqn [2]), the output signal even of gradient detectors may also become, at least to some extent, contrast sensitive and exhibit a velocity dependence which differs from that of an ideal velocity sensor.

Further Processing of Motion Detector Responses

Unless implemented in a very specific way, most detector models discussed here do not encode faithfully the velocity of the stimulus pattern. We have seen that this is reflected in two response characteristics: (1) the output signal of a detector may be modulated over time, even if the pattern moves with a constant velocity, and (2) the mean response amplitude may depend on the textural properties of the pattern. In the following, different ways of coping with these limitations are discussed.

One way to eliminate the temporal response modulations is to integrate spatially over an array of movement detectors looking at neighboring points in the visual field. Ideally, this array has to cover an integer multiple of spatial periods of a periodic stimulus pattern (**Figure 3(b)**). Since all local detectors 'look' at different spatial phases of the pattern, their output signals are also phase shifted. Spatially integrating these output signals leads to a smooth time course of the overall response which is, within certain limits, proportional to the time course of pattern velocity. However, these properties may differ if the gradient scheme is realized this useful consequence of spatial integration is paid for by a reduced spatial resolution of the motion detection system. It is obvious that the minimum range of spatial integration which is necessary for the response modulations to disappear depends on the spatial frequency content of the stimulus pattern which has access to the motion detection system. This implies that if the retinal image is spatially band-pass filtered peripheral to the site of movement detection, the minimum range of spatial integration depends on the low-frequency

cutoff of the spatial filters. Of course, the temporal modulations in the response of individual movement detectors are also eliminated when the signals are integrated over time instead of space. This suggests a trade-off between spatial and temporal resolution of a movement detection system. However, spatial or temporal integration only eliminates the temporal modulations of the detector signal. The mean response amplitude still depends on the structure of the stimulus pattern.

The time course of the spatially integrated movement detector response is proportional to pattern velocity only within a certain dynamic range. If the pattern velocity changes too rapidly, for instance, at the onset of motion, characteristic response transients may occur. This has been studied in detail for the Reichardt model. However, the response transients are neither a special feature of this model nor due to adaptational changes of some parameters of the motion detection system. Rather, they reflect the consequence of response transients of the temporal filters which are a constituent part of any motion detection system.

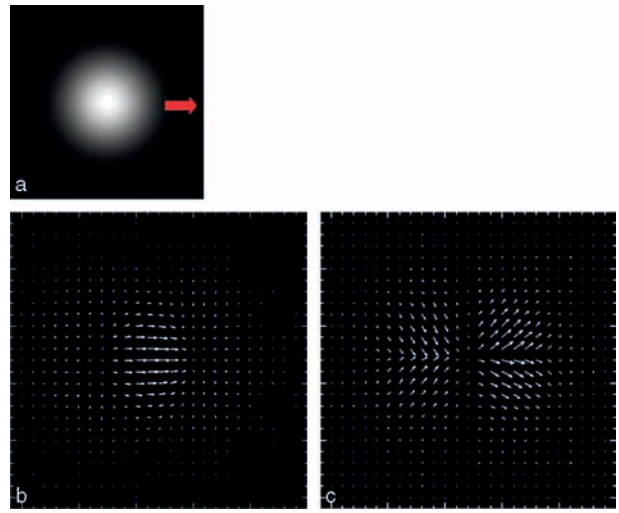

Figure 4 Motion in two dimensions. (a) A pattern consisting of a radially symmetric Gaussian luminance distribution is moving horizontally at a constant speed (arrow). (b) Response field of Reichardt detectors. (c) Response field of gradient detectors. Note that in both cases, the response vectors deviate locally from the actual pattern motion. Only after spatial integration is the correct direction of pattern motion recovered.

Motion in Two Dimensions

So far, the computation of motion information has been considered only for one spatial dimension. We now extend our analysis to motion in two spatial dimensions, thereby considering two major questions: How is the 2-D motion information represented at the output of arrays of motion detectors such as those discussed previously and to what extent is it possible to recover from these output signals the correct 2-D retinal velocity field as given by the geometrical projection of the three-dimensional motion vectors onto the retina?

Let us first consider the steady-state response of a pair of orthogonally oriented Reichardt detectors. In the limit of a small sampling base, the response can be approximated by the following expression:

$$\begin{pmatrix} r_x \\ r_y \end{pmatrix} = -\varepsilon \cdot \begin{bmatrix} F_x^2 - F \cdot F_{xx} & F_x F_y - F \cdot F_{xy} \\ F_y F_x - F \cdot F_{yx} & F_y^2 - F \cdot F_{yy} \end{bmatrix} \cdot \begin{pmatrix} V_x \\ V_y \end{pmatrix} \quad [5]$$

Here, the detector is assumed to have a delay line with delay ε instead of a temporal filter. Interestingly, the response of the x-detector (r_x) depends not only depends on the velocity component along the x-direction but also on the velocity component along the y-direction. The same holds true for the y-detector. Furthermore, both responses depend on a combination of the first and second spatial derivatives of the pattern, denoted by F_x, F_y, F_{xx}, etc. The pattern dependence of the response vectors of Reichardt detectors is illustrated by the computer simulation in Figure 4(a). Although the pattern moves with a constant velocity along the x-axis of the detector array only, the response vectors shown at a particular instant in time point to different directions depending on the local structure of the stimulus pattern. Thus, the local instantaneous response vectors may not coincide with the direction of pattern motion.

For a pair of orthogonally oriented, mathematically ideal gradient detectors, the responses are given, as in the one-dimensional case (see eqn [2]), by the relation between the temporal change of the pattern brightness at a given location and the corresponding spatial change along either the x- or the y-axis:

$$r_x(t) = -F_t/F_x \quad [6]$$

$$r_y(t) = -F_t/F_y \quad [7]$$

For motion in two dimensions, the so-called optical flow constraint equation, introduced by Horn and Schunck in 1981, holds:

$$-F_t = F_x \cdot v_x + F_y \cdot v_y \quad [8]$$

Combining eqs [6]–[8] then leads to the following equations for the responses of a pair of gradient detectors:

$$\begin{pmatrix} r_x \\ r_y \end{pmatrix} = \begin{bmatrix} 1 & F_y/F_x \\ F_x/F_x & 1 \end{bmatrix} \cdot \begin{pmatrix} v_x \\ v_y \end{pmatrix} \quad [9]$$

Each of the resulting response vector components thus depends on the brightness change of the retinal image along both the x- and the y-axis. Hence, in general, it does not coincide with the corresponding retinal velocity vector. The response vector field of such pairs of gradient detectors in response to the same stimulus as previously tested on pairs of Reichardt detectors is shown in **Figure 4(b)**.

Is it possible to compute the correct retinal velocity field from the response profiles of an array of Reichardt or gradient-type movement detectors? In principle, there are two ways to cope with it. One way is to take into account global features of the response field such as smoothness, which is equivalent to averaging over smaller patches of the response field. Another way is to recover the correct image velocity locally. In the case of Reichardt detectors, this is generally possible once the first and second spatial derivatives of the local pattern are available. In the case of the gradient detector, eq [8] has to be differentiated along the x- and y-directions separately, leading to two linearly independent equations:

$$-\begin{pmatrix} F_{tx} \\ F_{ty} \end{pmatrix} = \begin{bmatrix} F_{xx} & F_{yx} \\ F_{xy} & F_{yy} \end{bmatrix} \cdot \begin{pmatrix} v_x \\ v_y \end{pmatrix} \qquad [10]$$

Again, measuring the higher spatial and spatiotemporal derivatives, the equations can be solved for v_x and v_y, as long as the matrix containing the spatial derivatives of the image is nonsingular, which then leads to the correct image velocity field.

However, in case of a pattern consisting of parallel stripes crossing the receptive field of a motion detector, there is simply not enough information to recover the correct local velocity vector. This is often referred to in the literature as the 'aperture problem.' The motion vector can be decomposed formally into two components, one being orthogonal and the other parallel to the stripes. The orthogonal component is accompanied by a brightness change in this direction, whereas the parallel component is not. Hence, the latter component cannot be detected by any mechanism. Thus, the pattern brightness has to vary along two orthogonal directions. Only then can the 2-D velocity be recovered.

In addition to the Reichardt and the gradient detector, there exists an alternative method to calculate the optic flow which is quite commonly used in computer vision. This region-based matching approach defines the local velocity v as the shift $d = (d_x, d_y)$ that yields the best fit between image regions at different times. Finding the best match amounts to maximizing a similarity, such as the normalized cross-correlation, or minimizing a distance measure such as the sum-of-squared difference (SSD) over image patches consisting of $2n+1$ times $2n+1$ pixel at each image location x,y:

$$\mathrm{SSD}(x,y,d_x,d_y) = \sum_{j=-n}^{n}\sum_{i=-n}^{n}[F(x+i,y+j,t) - F(x+i+d_x,y+j+d_y,t+1)]^2 \qquad [11]$$

When comparing the matching algorithm outlined previously with the Reichardt detector, one can see the following important differences between the two algorithms: whereas the matching approach finds the maximum of the cross-correlation, relates the location of the maximum to the respective x and y shifts, and interprets these as the x- and y-components of the velocity vector, the Reichardt detector operates with fixed x and y shifts (the detectors' sampling bases along the x and y direction), calculates the cross-correlation for these shifts, and interprets the magnitude of the cross-correlation for both shifts as the x- and y-components of the local velocity vector. Therefore, in general, the outcomes of both operations are different.

Conclusions

Interestingly, only a small number of models of motion detection have been proposed in the literature. These different models can be grouped, according to the computational principle they are built on, into two classes – the so-called correlation and gradient detectors. Despite the apparent overall difference in the structure of the various model versions, some of them show a similar performance, for instance, with respect to their velocity tuning and their dependence on pattern structure. Thus, the dichotomy between at least some of the models starts to fade, and this is all the more true when the 'sloppy workmanship' of the neuronal hardware is taken into account. In addition, a closer examination of the internal structure of the models is necessary because predictions may differ significantly depending on the processing level which is considered. Hence, it is often not as easy as one might expect at first glance to determine what mechanism underlies motion detection in a particular biological system. Therefore, modeling is an indispensable tool for designing more specific experiments which may eventually allow us to distinguish between different mechanisms.

See also: Optic Nerve, Optic Chiasm and Optic Tracts; Retinal Ganglion Cells: Receptive Fields; Vision: Light and Dark Adaptation; Visual Motion Detection.

Further Reading

Adelson EH and Bergen JR (1985) Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A* 2: 284–299.

Barlow HB and Levick WR (1965) The mechanism of directionally selective units in rabbit's retina. *Journal of Physiology* 178: 477–504.

Barron JL, Fleet DJ, and Beauchemin SS (1994) Performance of optical flow techniques. *International Journal of Computer Vision* 12(1): 43–77.

Borst A and Egelhaaf M (1989) Principles of visual motion detection. *Trends in Neurosciences* 12: 297–306.

Buelthoff H, Little J, and Poggio T (1989) A parallel algorithm for real-time computation of optical flow. *Nature* 337: 549–553.

Euler T, Detwiler PB, and Denk W (2002) Directionally selective calcium signals in dendrites of starburst amacrine cells. *Nature* 418: 845–852.

Goetz KG (1972) Principles of optomotor reactions in insects. *Bibliotheca Ophthalmologica* 82: 251–259.

Grzywacz NM and Yuille AL (1990) A model for the estimate of local image velocity by cells in the visual cortex. *Proceedings of the Royal Society of London, Series B* 239: 129–161.

Hausselt SE, Euler T, Detwiler PB, and Denk W (2007) A dendrite-autonomous mechanism for direction selectivity in retinal starburst amacrine cells. *PLoS Biology* 5: 1447–1493.

Heeger DJ (1987) Model for the extraction of image flow. *Journal of the Optical Society of America A* 4: 1455–1471.

Hildreth EC and Koch C (1987) The analysis of motion: From computational theory to neuronal mechanisms. *Annual Review of Neuroscience* 10: 477–533.

Horn BKP and Schunck BG (1981) Determining optical flow. *Artificial Intelligence* 17: 185–203.

Marr D and Ullman S (1981) Direction selectivity and its use in early visual processing. *Proceedings of the Royal Society of London, Series B* 211: 151–180.

Reichardt W (1961) Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In: Rosenblith WA (ed.) *Sensory Communication*, pp. 303–317. New York: MIT Press/Wiley.

Reichardt W (1987) Evaluation of optical motion information by movement detectors. *Journal of Comparative Physiology A* 161: 533–547.

Single S and Borst A (1998) Dendritic integration and its role in computing image velocity. *Science* 281: 1848–1850.

Srinivasan MV (1990) Generalized gradient schemes for measurement of image motion. *Biological Cybernetics* 63: 421–431.

Srinivasan M. and Bernard GD (1976) A proposed mechanism for multiplication of neural signals. *Biological Cybernetics* 21: 227–236.

Torre V and Poggio T (1978) A synaptic mechanism possibly underlying directional selectivity to motion. *Proceedings of the Royal Society of London, Series B* 202: 409–416.

Uras S, Girosi F, Verri A, and Torre V (1988) A computational approach to motion perception. *Biological Cybernetics* 60: 79–87.

van Santen JPH and Sperling G (1985) Elaborated Reichardt detectors. *Journal of the Optical Society of America A* 2: 300–320.

Visual Signaling in Animals

K J McGraw, Arizona State University, Tempe, AZ, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The human visual sense is the most accessible and the most often studied of our five senses. It is perhaps no surprise, then, that, as scientists, humans have been especially attracted to the diversity of forms and functions of visual traits in other animals. Visual features include overt behaviors that animals exhibit, as well as their body sizes, shapes, colors, patterns, and structures. It has been argued by biologists as far back as Darwin that animals use prominent, often exaggerated movements by or features on the body in a form of 'visual communication.' Much more than 'body language,' which might be the first thing that comes to mind when asked about examples of behavioral communication, many varieties and intricacies of behaviors and morphologies, including facial expressions, colors, courtship displays, and protruding tissues like horns or feathers, constitute the repertoire of animals aiming to communicate information without vocalizing, odorizing, tasting, or touching.

Animals use signals in a variety of behavioral contexts, including breeding, predator avoidance, social grouping, and foraging. Moreover, animals exchange information about numerous attributes of individuals, including sex, individual identity, location, mate quality, breeding strategy, and genetic make-up. A major thrust in research on animal communication today is to understand the mechanisms via which animals control their signals. In addition to nutritional and immunological factors, the neural and endocrine bases of signal production have received much attention. A great deal of this work has been done in relation to vocal communication, but the emphasis on the neuroendocrinology of visual signals nearly rivals that on acoustic signals.

The aim of this article is to highlight some of the uses of behavioral and morphological visuals in animals and to pay special attention to those examples that have clear neural or endocrinological bases or that have apparent value for future investigations at these levels. As the ability to characterize and quantify animal signals and neuroendocrine pathways continues to improve, there is no doubt that numerous neuroendocrinological underpinnings of animal signals will be uncovered.

Behavioral Visual Signals

Courtship Displays

Perhaps the most exaggerated and thoroughly studied of the behavioral visual signals in animals, especially from the proximate perspective, are the courtship displays. These are predominantly performed by males and include precopulatory food offerings, dances, or other movements in the presence of a prospective or secured mate. Those behaviors exhibited most frequently by females include receptivity postures just prior to mating, except in mutually mate-selecting or sex role-reversed animals, in which case, females actively pursue and display for males.

Clearly the species for which we have the most detailed neurological information about a male courtship display is *Drosophila melanogaster*. Prior to mating, male fruit flies orient toward females and vibrate an extended wing, in addition to performing other behaviors (e.g., tapping, licking). Specific genes, *dissatisfaction* and *fruitless*, are linked to the sex-determination pathway and govern neurally based courtship behavior. Mutations to *dissatisfaction* alter male recognition of females as mates, such that males actively court other males. This is believed to have a neural basis, but controlling brain regions have not yet been identified. The *fruitless* gene, in contrast, has been tied more closely to abnormal courtship, and male-specific protein products of this gene have been identified in the superior protocerebrum of the brain, which has been implicated in courtship initiation. The central nervous system circuitry encoded by *fruitless* also includes olfactory neurons that appear to be required for the behavioral response of males to female sex pheromones. The interactive nature of male–female courtship interactions has been elucidated recently for other animal signals, especially in bowerbirds, but for no other visual signal has such a detailed, complex neural and genetic basis been uncovered.

Proximate studies of courtship displays in vertebrates have traditionally centered on gonadal sex steroid mediation, and a diversity of hormonal control mechanisms exist. Courtship dances and displays are testosterone-mediated in a number of bird species. However, strutting in male Japanese quail (*Coturnix japonica*) is sensitive to estrogens, and this differs from the way hormones control their courtship vocalization (crowing), which is more androgen-sensitive (e.g., testosterone and 5α-dihydrotestosterone). Perhaps the two best-studied of the vertebrate courtship displays from a neuroscientific perspective are the

throat-fan (dewlap) extension of male anoles (*Anolis carolinensis*) and the bowing and nest soliciting of male ring doves (*Streptopelia risoria*). Male anoles evert a brightly colored flap of gular skin using (ceratohyoideus) muscles that bow out a piece of cartilage (retrobasal process) in the throat. This process is androgen-dependent, and the throat muscles are innervated by seasonally and sexually dimorphic motor neurons located in two regions of the brain stem. In male ring doves, bowing is testosterone-dependent, but nest-soliciting wing flapping (by both sexes) is estradiol-dependent, and both hormonally governed behaviors appear to be controlled in the anterior hypothalamus preoptic area (HPOA), as determined by carefully designed studies using hormone implants and enzyme blockers.

The flight displays of male dark-eyed juncos (*Junco hyemalis*) are an example of a courtship behavior that is testosterone-dependent and preferred by female mates. There are also a few excellent cases in which vertebrate courtship behaviors have been traced down to the neuromuscular and neuroendocrine paths. For example, loud, rapid courtship wingsnaps are produced by male, but not female, golden-collared manakins (*Manacus vitellinus*) resulting from a sexual dimorphism in wing musculature and sex steroid accumulation in regions of the spinal cord that innervate such musculature. Identifying such specific neuroendocrine paths for the musculature that controls other elaborate courtship behaviors, as has been done on the avian song system, should yield exciting avenues for future work.

Finally, despite these numerous examples, gonadal steroids are not always required for adult animals (e.g., certain cichlids, toads, sparrows, and primates) to maintain their courtship signals. In many cases, adrenal or brain-derived steroids are thought to be the sources of hormones that sustain the behavior. There can also be early-life effects of hormone that organize courtship later in life. Alternate hormones (e.g., peptide, protein) should also be considered as chemical signals for courtship behavior maintenance, much like the role of vasotocin in male rough-skinned newts (*Taricha granulosa*).

Parental Behavior and Aggression

Parental effort and aggression are two additional behaviors that often function as conspecific signals, in many cases increasing the attractiveness of potential male partners. There is an extensive literature on the endocrine basis for such behaviors in a wide variety of animals. Historically, prolactin has been viewed as 'the' parental hormone, but this is by no means universal. Prolactin stimulates milk letdown and subsequent delivery to young in mammal mothers and facilitates offspring rearing in male and female birds. However, prolactin has no influence on lamb recognition and acceptance by mother sheep and generally is thought to play more of a role in the maintenance of parental behavior rather than its onset. Steroids and peptides, in contrast, strongly induce parental instincts toward offspring. In sheep, it is oxytocin release during birth that stimulates maternal interest in young. Estrogen, aromatized in the brain from testosterone, is a key sex steroid that drives paternal behavior in male California mice (*Peromyscus californicus*); testosterone, along with progesterone, traditionally inhibits parental tendencies and can even facilitate paternal aggression toward offspring (e.g., in rodents; see more on aggression in the next paragraph). The neural centers that mediate paternal behavior have been described in a few instances, particularly for prolactin action and crop milk production and regurgitation in ring doves and for oxytocin action in the paraventricular nucleus of the hypothalamus that promotes lamb acceptance in sheep. Brain sites for parental behavior are typically the same as those for sexual behavior in general (e.g., preoptic area, bed nucleus of the stria terminalis).

Testosterone, often interfering with parental behavior, is widely viewed as the 'aggression' hormone. Threatening, competitive, and fighting behaviors fluctuate in response to changing levels of testosterone, or derivatives such as 11-keto-testosterone (in fish), in male animals from all major vertebrate groups. A good example of a sexually valued, hormone-dependent aggressive signal involves the gray partridge (*Perdix perdix*), whose vigilance behavior, a form of aggression aimed at deterring predators, is testosterone-dependent in males and is sexually attractive to females. Brain-derived (from testosterone) estrogens can also influence male aggression (e.g., in mice). Though the proximate underpinnings of female aggression have been comparatively neglected, gonad-derived sex steroids control aggression in some females as well (e.g., mountain spiny lizards, *Sceloporus jarrovii*).

Nonsteroid hormonal mechanisms can underlie aggressive behavior, and to date these provide some of the best neural links to animal aggression. The role of neuropeptides has received increased attention in recent years, as vasopressin in mice, vasotocin in songbirds, and adrenal catecholamines in lizards all play pivotal roles in modulating competitiveness, but with different relative outcomes, sometimes even in closely related species. For example, septal infusions of vasotocin inhibited aggression in territorially breeding male violet-eared waxbills (*Uraeginthus granatina*) and field sparrows (*Spizella pusilla*) but promoted aggression in a colonial songbird, the zebra finch (*Taeniopygia guttata*). Work on

the green anole has also yielded insights into the turnover of serotonin, dopamine, adrenalin, and noradrenaline in brain stem nuclei as a function of aggressive signaling. Finally, yawning in humans and other animals is often viewed as a threatening social signal and can come under diverse neural and endocrine controls. Dopamine, acetylcholine, serotonin, nitric oxide, and oxytocin are among the neurochemicals that facilitate yawning, while opioid peptides can inhibit this behavior. Functional magnetic resonance imaging has recently been used in humans to show that brain regions involved with self-(bilateral precuneus and posterior cingulate) and facial-information processing (bilateral thalamus and parahippocampal gyrus) are active during yawning, suggesting an important neural basis for the communication value of yawning. Beyond these examples, however, the neuroendocrine basis for an aggressive or parental behavior that serves directly as a signal has rarely been determined.

In many instances, aggressive behavior itself is signaled by a different trait, a morphological 'badge of status' (e.g., color patch; body, horn, or antler size), that can come under neuroendocrine control. Examples are discussed in the next section.

Morphological Visual Signals

Colors and Patterns

Animals use colors and patterns to signal in mating or competitive contexts within their species as well as in cryptic, aposematic, or species isolation contexts among species. In the latter case, with the exception of rapid background matching in ectotherms, signals are highly pattern-(as opposed to intensity-) dependent, generally invariant among individuals, and come under genetic control, which for some may involve hormonal pathways (e.g., the role of α-melanocyte-stimulating hormone (α-MSH) and the melanocortin-1 receptor (MC1R) for melanic color patterns). The majority of the emphasis here will be on intraspecifically variable mating or competitive color signals.

Physiological color changes in fishes, amphibians, and reptiles have served as model examples for studying the neuroendocrinology of visual color signaling in animals. This is because light-absorbing pigments (e.g., melanin, carotenoid, pterin, purine) are harbored in living cells (chromatophores) that can actively aggregate or disperse molecules, in response to physiological triggers (and ultimately environmental cues), on a minute-to-minute basis. Here, hormones like α-MSH and melanin-concentrating hormone have profound regulatory functions for melanosome accumulation (hence, skin darkening) in numerous ectotherms, although other hormones (e.g., epinephrine, serotonin, thyroxine, pro-opiomelanocortin, and somatolactin) have been implicated. An excellent example of such work on a color known to function as a signal is work on skin darkening in Arctic charr (*Salvelinus alpinus*) and the green anole. In charr, stressful social conditions induce serotonin and norepinephrine in the central nervous system to modulate pituitary release of adrenocorticotropic hormone and α-MSH, which together drive skin darkening. Detailed neural and cellular work, in the context of rapid cryptic (background-matching) color acquisition, has also been performed in the African clawed frog (*Xenopus laevis*), where melanotrope cells from the pars intermedia of the pituitary release pro-opiomelanocortin-derived α-MSH to redistribute melanosomes and change color, and superchiasmatic neurons, which produce the neurotransmitters γ-aminobutyric acid, dopamine, and neuropeptide Y, transport these factors to axons contacting the melanotrope cells and ultimately inhibit α-MSH secretion. A paracrine factor (macrophage migration inhibitory factor) and autocrine factors (e.g., brain-derived neurotrophic factor) can also influence amphibian melanophore action.

Most colors of endotherms differ from those of ectotherms in that they are contained in dead tissue like feathers or hair and thus cannot be as dynamically linked to neuroendocrine processes. A classic case of such a signal is the bright red plumage of house finches (*Carpodacus mexicanus*), which is dependent on adequate nutrition (intake of proper carotenoid pigments from plant foods) and health. Still, hormonal regulation of bird coloration has received considerable attention (more than mammal color regulation), mostly from the perspective of sexual plumage dichromatism. The carotenoid-based red beak of male zebra finches and the melanin-based badge of the male house sparrow (*Passer domesticus*) are classic examples of bird colors that are exaggerated (made richer or larger in size) by androgens. It is important to point out that the sexual variation (dichromatism) observed in these cases cannot be overcome by hormones (e.g., inducing a male-like condition in females by administering testosterone); instead, the sex difference has a genetic basis. In the majority of work in this area, however, no proper procedures have been followed to rule out estrogenic (via aromatization from testosterone) control of color, and no neural link to such coloration has been established. Melanic colors in songbirds, especially Old World ploceids (e.g., bishops and widowbirds), are also sensitive to luteinizing hormone (LH).

Coat colors of mammals typically vary among species in melanin pattern and are genetically

determined. Many of these melanic features can be sensitive to hormones like α-MSH and are governed by the MC1R locus, and evidence from this is extensive in domesticated mammals (e.g., cattle, rodents, dogs). Some species also exhibit fleshy red and blue colors on the face or rump that dynamically respond to social (competitive) conditions and androgen levels (e.g., in mandrills, *Mandrillus sphinx*). Fast-acting neuroendocrine paths are also implicated in the social signal known as facial blushing in humans (e.g., sympathetic arousal of beta-adrenergic receptors, resulting in vasodilation).

Enlarged Integumentary Features

Body structures that are exaggerated beyond the basic head-to-toe morphology of an animal (structures such as horns or antlers; manes; feather crests; or fleshy tissues such as wattles, combs, or sexual swellings) often play valuable roles as attractive or competitive signals. For each of these categories of traits, an endocrine basis for their formation has been established. High testosterone levels promote seasonal mane growth and antler pedicle initiation in red deer (*Cervus elaphus*) stags, and low levels induce antler casting when breeding has ceased, but high levels do not necessarily promote antler growth. It has recently been shown at the cellular level that androgen receptors seasonally upregulate in hair follicles and allow for testosterone-dependent formation of the enlarged mane in this species. The seasonal cycle for female reindeer (*Rangifer tarandus*) antlers, which are smaller than males' antlers, is similarly sensitive to sex steroids, although the relative roles of testosterone and estrogen have not been clarified.

Classic work on the ornamental feather plumes in males (e.g., neck fringes, hackles) from certain strains of domestic chickens (*Gallus gallus*) shows that the male plumage is the default (genetically programmed) pattern and that only in the presence of estrogen does the female plumage develop. This can even be extended to strains (e.g., Sebright) whose males display female feathering, and this plumage is maintained by the aromatization of estrogen from testosterone in skin. Such estrogen dependence (and androgen independence) of ornamental plumage dichromatism is very common in large, ancestral birds (e.g., ostrich, gamebirds, waterfowl).

In contrast, the size of exaggerated fleshy integumentary structures, beyond their color appearance, can come under androgenic control. The size of combs on the heads of male roosters, red grouse (*Lagopus lagopus*), and black grouse (*Tetrao tetrix*) is testosterone-dependent. Sexual swellings on the hind end of some female Old World primates (e.g., olive baboons, *Papio cynocephalus anubis*) advertize the timing of their fertility as well as their lifetime reproductive potential and thus are closely tied to ovarian hormone cycles.

Body Size and Symmetry

Generalized features of the body, including its overall size and the bilateral symmetry of certain attributes, have also received considerable attention as signals with which individuals can send and receive important information about a rival or prospective mate. In many species, for example, large males are dominant, and as indicated earlier, dominance is controlled by sex steroids. However, while hormones such as testosterone can increase muscle mass and other growth parameters in wild animals, as well as in humans, it appears that no study of a free-ranging animal has yet shown that testosterone or any other hormone modulates both body size and its role as a competitive signal.

Fluctuating asymmetries (FAs) in bilateral traits have been proposed to contain useful signaling information regarding the developmental (e.g., nutritional, health, genetic) history of individuals. Hormones may also play key roles in the development of features that later in life do not exhibit perfect symmetry and thus may be used as a signal. There was an outburst of studies on the signal content of FA in animals throughout the1990s, and while interest has waned, some scientists studying humans have shown recent interest in the endocrine basis for FA signals. Sexually attractive odors are linked to the facial symmetry (and hence sexual attractiveness) of men, and these can be pheromonal, but regarding visual signals, soft-tissue symmetry itself is argued to have a hormonal basis, even on a dynamic scale. In a sample of ten men, researchers found that fingers, wrists, and ears were more asymmetric when circulating levels of LH and thyroxine were high and when levels of follicle-stimulating hormone were low. It has been argued that highly asymmetrical men tend to be more aggressive and that an endocrine mechanism underlies FA-linked behavior (e.g., LH through its effects on testosterone). However, experimental work is needed to better understand whether and how hormones and neural pathways may influence FA.

Extended Phenotypes: Where Morphology and Behavior Meet

Some organisms develop signals, not on their own bodies, but on or as structures in the environment ('extended phenotypes'). Nests and their construction, for example, can serve as signals of effort on which males or females base mating decisions (i.e., in sand gobies, *Pomatoschistus minutus*). Bowerbirds from Australia are also a classic example of a species that

builds a signal in the environment. Male bowerbirds build dummy nest structures not used to hold eggs but as a display trait and an arena in which to display themselves. Among bowerbird species, bowers can range from simple piles of sticks to elaborate, tall towers of sticks glued together and decorated with a black-painted moss rug and colorful objects like flower petals, snail shells, and butterfly wings (i.e., in *Amblyornis inornatus*). Species that have little plumage or body ornament tend to have more exaggerated bowers. Also, species with more complex bowers have a larger cerebellum, which is perhaps associated with greater motor skills and learning. In the satin bowerbird (*Ptilonorhynchus violaceus*), males with higher testosterone levels have more exaggerated displays, are more likely to build and hold a bower, and have high mating success.

Signal Perception

While the majority of work on animal visual signals has centered on the production and characteristics of the signal itself (e.g., signal 'potential'), in reality this information is not technically what the signal receiver perceives and responds to as attractive or unattractive, dominant or submissive. Signal reception and perception require appropriate eye detection and tuning as well as neural information processing that decodes behavioral or morphological signal variation. A color, for example, is perceived by photoreceptors that are maximally sensitive to particular light wavelengths, and this light can be altered by both the ambient environment in which the animal is found and the light-filtering properties of the ocular media. Though vision researchers have now described the optical characteristics of some animal eyes and identified the true abilities and limits of signal reception, this is a field in its infancy and one that should prove fruitful for those keen on understanding variability in, control of, and interactions between signal perception and production. It is said that we cannot truly appreciate the 'color sensation' (or really any other visual sensation) of other animals until we map neuro-optical circuitry on top of visual acuity, and this should become more of a goal, collaborative or otherwise, of researchers studying signaling systems where truly 'beauty is in the eyes – and brain – of the beholder'.

See also: Animal Communication: Honesty and Deception; Game Theory and the Economics of Animal Communication; Multimodal Signaling in Animals; Referentiality and Concepts in Animal Cognition; Sexual Selection and the Evolution of Animal Signals; Signal Transmission in Natural Environments; Signal Design Rules in Animal Communication.

Further Reading

Adkins-Regan E (2005) *Hormones and Animal Social Behavior.* Princeton, NJ: Princeton University Press.

Andersson M (1994) *Sexual Selection.* Princeton, NJ: Princeton University Press.

Bagnara JT and Hadley ME (1973) *Chromatophores and Color Change: The Comparative Physiology of Animal Pigmentation.* Englewood Cliffs, NJ: Prentice-Hall.

Bradbury JW and Vehrencamp SL (1998) *Principles of Animal Communication.* Sunderland, MA: Sinauer.

Buntin JD (1996) Neural and hormonal control of parental behavior in birds. In: Rosenblatt JS and Snowdon CT (eds.) *Advances in the Study of Behavior,* vol. 25, pp. 161–213. New York: Academic Press.

Diamond J (1986) Animal art: Variation in bower decorating style among male bowerbirds *Amblyornis inornatus. Proceedings of the National Academy of Sciences of the United States of America* 83: 3042–3046.

Espmark Y, Amundsen T, and Rosenqvist G (2000) *Animal Signals: Signaling and Signal Design in Animal Communication.* Trondheim, Norway: Tapir Academic Press.

Fujii R (2000) The regulation of motile activity in fish chromatophores. *Pigment Cell Research* 13: 300–319.

Hill GE and McGraw KJ (2006) *Bird Coloration, Vol. I: Mechanisms and Measurements.* Cambridge, MA: Harvard University Press.

Korzan WJ, Summers TR, Ronan PJ, Renner KJ, and Summer CH (2001) The role of monoaminergic nuclei during aggression and sympathetic social signaling. *Brain Behavior and Evolution* 57: 317–327.

Maynard-Smith J and Harper D (2003) *Animal Signals.* Oxford, UK: Oxford University Press.

Numan M (1994) Maternal behavior. In: Knobil E and Neill JD (eds.) *The Physiology of Reproduction,* 2nd edn., pp. 221–302. New York: Raven Press.

Platek SM, Mohamed FB, and Gallup GG Jr. (2005) Contagious yawning and the brain. *Cognitive Brain Research* 23: 448–452.

Riters LV, Teague DP, and Schroeder MB (2004) Social status interacts with badge size and neuroendocrine physiology to influence sexual behavior in male house sparrows (*Passer domesticus*). *Brain Behavior and Evolution* 63: 141–150.

Schlinger BA, Schultz JD, and Hertel F (2001) Neuromuscular and endocrine control of an avian courtship behavior. *Hormones and Behavior* 40: 276–280.

Stockinger P, Kvitsiani D, Rotkopf S, Tirian L, and Dickson BJ (2005) Neural circuitry that governs *Drosophila* male courtship behavior. *Cell* 121: 795–807.

Visual System Development: Invertebrates

P R Hiesinger, University of Texas Southwestern Medical Center, Dallas, TX, USA
I A Meinertzhagen, Dalhousie University, Halifax, NS, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Visual system development in most invertebrate groups commences with photoreceptor differentiation in an ectodermal epithelium. Photoreceptors generally differentiate independently of the nerve centers they innervate, and peripherally derived visual interneurons are exceptional. This is in contrast to vertebrates, in which the 'retina' arises from an embryological outgrowth of the forebrain, which generates multiple classes of interneurons as well as photoreceptors. Most of what we know about the genetic and molecular bases of visual system development in invertebrates is derived from studies on the compound eyes of the fruit fly *Drosophila melanogaster*. While this focus on a single model system rewards us with the depth of our understanding of developmental mechanisms, it has given rise to a substantial gap in our often rudimentary knowledge of other deserving groups: cephalopods, pectinid scallops, spiders, cubomedusans, and salps, to name but a few, all with advanced eyes. The same is true for the development of neuronal connections within the visual centers in the brain (Figure 1). Here, too, *Drosophila* is the invertebrate model of choice to elucidate underlying molecular mechanisms. Our understanding of many aspects of invertebrate visual system development must, however, encompass knowledge not only of the compound eyes of arthropods, but also of eyes as diverse as, for example, the highly developed single-lens eyes of cephalopods, or ancestral eyecup eyes in forms such as planaria.

Evolution of Invertebrate Eyes

Different eye types are distinguished by substantial differences in morphogenesis as well as their embryonic origin. Two lines of photoreceptor evolution, rhabdomeric and ciliary, had been thought the exclusive domains of, respectively, protostomes and deuterostomes. However, various invertebrate groups have both types of photoreceptors, and in some cases (e.g., *Pecten*) both types are found in one eye. Various examples of developing rhabdomeric photoreceptors (which bear visual pigment upon microvilli) illustrate the transient appearance of ciliary structures. Such differences in form, development, and photoreceptor organelles among the different eye types have been taken to imply an independent origin for the eyes of the three main animal groups that have advanced visual systems: arthropods, mollusks, and vertebrates. Based on classical anatomical and developmental criteria, this view has been challenged by the advent of molecular genetics and the discovery of ancestral conserved genes such as *Pax6*. *Pax6* is essential for eye development in numerous species so far tested. Conversely, its overexpression induces ectopic eyes, as shown in flies and frogs. The *Pax6* gene – like its fly homologs *eyeless* and *twin of eyeless* – encodes a homeodomain transcription factor that regulates a cascade of downstream genes. In *Drosophila*, this cascade involves around 2500 genes that are required to make an eye. The *eyeless* gene and several other genes that are essential to specify eye development, as well as sufficient to induce ectopic eye development in other tissues, are therefore sometimes referred to as 'master control genes.' These are found throughout the animal kingdom, even in ancestral forms such as the freshwater sponge *Ephydatia*. In *Drosophila*, at least six such genes have been characterized to date: *eyeless* (*ey*), *twin of eyeless* (*toy*), *sine oculis* (*so*), *eyes absent* (*eya*), *dachshund* (*dac*), and *optix*; removal of any one results in reduction or deletion of the eye, while ectopic expression (except for *so*) can lead to ectopic eye formation (Figure 2). Genes of the *six* family are transcription factors with a six domain and a six-type homeodomain that share homology with *Drosophila so*. Tentacular eyes of the jellyfish *Cladonema radiatum* express orthologs of *six1/2* and *six3/6* in the eyecup, and during eye regeneration upregulate their expression, suggesting conservation of upstream regulatory mechanisms in eye development since prebilaterian ancestors. Additional molecular evidence supporting the relatedness of all animal eyes includes the primary signal for patterning, *hedgehog*; the *atonal*-related basic helix–loop–helix transcription factors that specify the first retinal neurons; and the highly conserved visual pigment protein, opsin. Most *opsin* genes contain putative *Pax6*-binding sites, for which reason *opsin* is proposed to form part of a conserved ancestral *Pax6–opsin* regulation module. Although the evolutionary relations between eye types are still unresolved, it seems likely that a common ancestral unit existed which may have been a simple photoreceptive cell using *Pax6* to regulate *opsin* expression. From there, gene regulatory regions may have diverged and different molecular pathways recruited in different groups. This scheme thus implies the independent evolution of different eye

types that nevertheless share a remote common ancestor. The widespread role of *Pax6* is consistent with its action in the last common ancestors of insects and vertebrates in simple pigment-cup ocelli, as represented by cerebral eyes in today's errant polychaetes, such as *Platyneiris*. In early developing *Platyneiris*, *six1/2* expression covers the complete visual system and *atonal* (*ath*) expression appears in its first photoreceptor cells. The same is true in insects, but not vertebrates. This sequence has suggested that *Pax6*-, *six1/2*-, and *ath*-positive larval eyes, as found in living trochophore larvae, were already present in the last common ancestor of all bilaterians, vertebrate and invertebrate.

Invertebrate Eye Types and Their Development

Most invertebrate eyes arise from an ectodermal epithelium in one of two main ways: (1) by the formation of a fold and invagination to create either a pit, open to the outside, or a vesicle that subsequently becomes internalized, and (2) by delamination, the thickening of a region as cells divide, grow, and differentiate, followed by the formation of the cell layers that eventually constitute the adult retina. These can either remain as part of the epithelium, or detach and move into the body cavity. The original topography of the epithelium is conserved in the retinal layer of photoreceptors following invagination and delamination *in situ*, whereas it must be reconstituted secondarily following the inward migration of surface epithelial cells. Unlike vertebrate groups, different groups of invertebrates use different modes of eye development and have varying numbers of eyes. Both developmental origins (invagination and delamination) are described in some invertebrate groups and even, in some cases, in the different eyes of a single species.

Evolutionary Divergence among Invertebrate Eyes

The flagellate protozoans, *Erythropsis* and *Warnowia*, have elaborate ocellar organelles with a dioptric hyalosome and pigmentary melanosome. These two components dissociate and duplicate during cell division, the daughter organelles recombining thereafter. In Cnidaria, the ocellus of the medusa of *Cladonema* arises from an ectodermal primordium from which photoreceptor, corneal, and pigment cells differentiate,

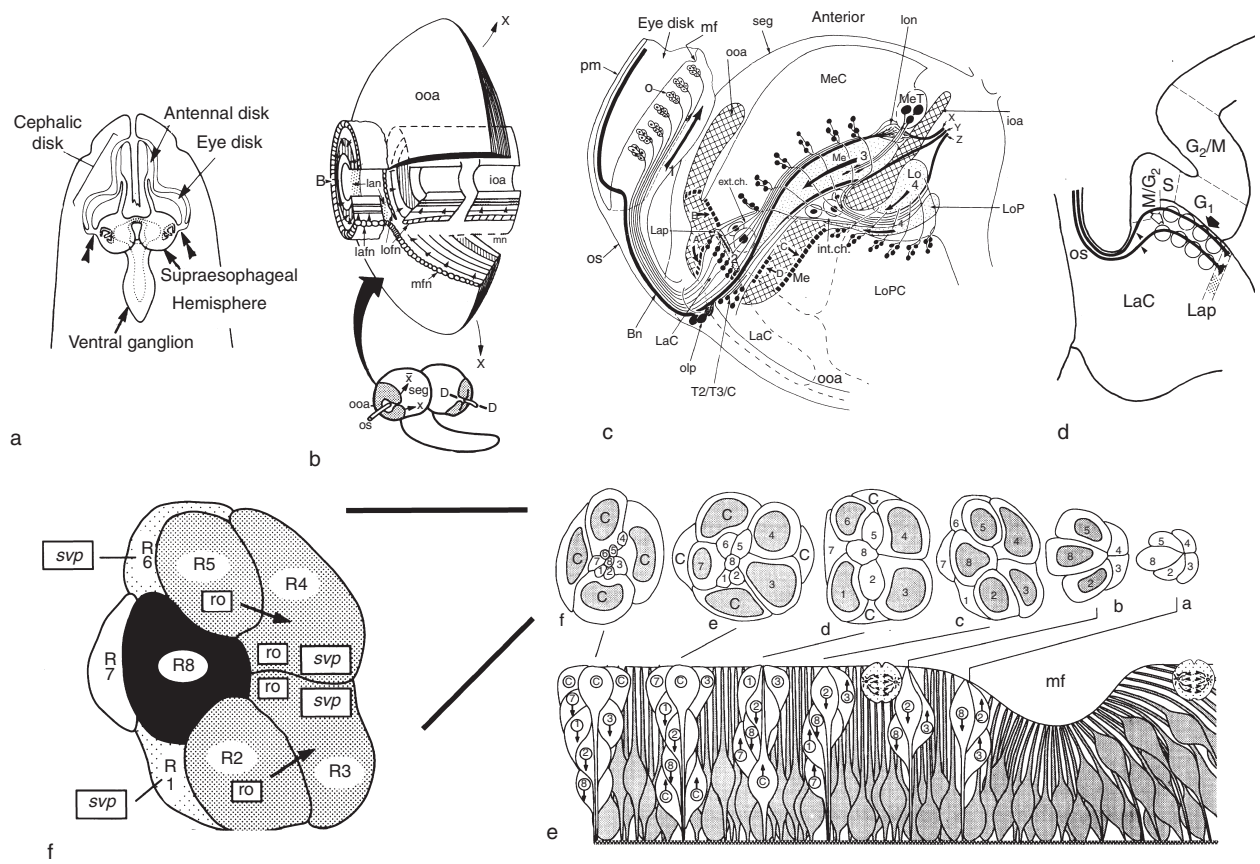

Figure 1 Continued

the lens arising in the latter. In turbellarians with direct development, eyespots appear in the ectoderm and then sink into the mesoderm, enlarge, and divide to provide multiple pairs. In annelids, eyes are usually simple eyecup ocelli, frequently with one or few photoreceptors arising, as in the prostomial eyecups in leeches, from epidermal thickenings. Only errant polychaetes have elaborate eyes, which are variably said to arise either from the epidermis, without invagination, or by delamination from an anlage, apparently of ectodermal origin, lying between ectoderm and pharynx and containing a special lens-secreting cell. Differentiation occurs in a wave – first in the inner, retinal layer. By contrast, the pair of ocelli in Onycophora invaginate from epidermal placodes, which pinch off to form a vesicle lined by everted microvillar photoreceptors. Echinoderms have multiple, simple ocelli on optic cushions that form when relatively few cells invaginate from the epidermis. Among prevertebrate chordates, interesting eyes are found in salps, which have a single, dorsal eye pushed out in vertebrate fashion from a crescent-shaped optic ridge of the cerebral ganglion, to form an ocular disk. In the chain form, the asexual part of the life cycle, the ocular disk folds up, flips over, and produces a layer of inverted photoreceptors abutting a layer of everted ones and then an everted secondary eye.

Mollusks and Arthropods

Clear accounts of eye ontogeny are generally restricted to groups with elaborate eyes, especially

Figure 1 Summary of development in the visual system in *Drosophila melanogaster*, viewed in steps of progressively finer cellular resolution. (a) The eye develops from an eye disk connected by an optic stalk (double arrowheads) to the developing optic lobe in the supraesophageal hemisphere, or ganglion, of the larval brain. (b) The optic stalk (os) penetrates the supraesophageal ganglion (seg) at the center of the crescent-shaped outer optic anlage (ooa). With later development, the arms of the anlage open in the direction marked 'X,' converting the cortices to which it gives rise from circular to rectangular forms. The outer anlage and the concentric inner optic anlage (ioa) are shown in relation to the point of entry of the optic stalk to the posterolateral surface of the right hemisphere. Neuroblasts and other progenitor cells in the anlagen proliferate in the directions of the arrowheads, contributing time-sorted strata of cells to the lamina (lamina-forming neuroblasts; lafn), medulla (mfn), and third optic neuropil (lobula; lofn). One stratum (shown cross-hatched in each cortex) was produced at the same time and has been displaced by newer strata in its respective cortex (lan, lamina neuropil; mn, medulla neuropil). The relationships between these cell populations is clearest in cross section (in diagram c). (c) Relationship between the generation of cell cortices in diagram b and the axonal pathways growing between them, in a horizontal plane of diagram b, that illustrates the close relationship between the waves of imaginal photoreceptor innervation and the path of the larval Bolwig's nerve (Bn). This runs from the inner face of the peripodial membrane (pm) to the lamina plexus (Lap) and medulla (Me) via the optic stalk (os). It and/or the axons of three optic lobe pioneers (olp), postulated interneurons, innervate a larval optic neuropil (lon), which is connected to the central brain by a pathway (X) that anticipates the posterior optic tract. New ommatidial clusters (o) accumulate behind the morphogenetic furrow (mf), contributing new axon bundles that fasciculate on previously extended axons in the optic stalk with the anterior expansion of the retinal field (arrow 1). Underlying cell populations expand in corresponding directions: arrow 2, lamina cortex (LaC); arrow 3, medulla (Me) and medulla cortex (MeC); arrow 4, lobula (Lo) and lobula plate cortex (LoPC). Cells add to the lamina cortex from one side of the outer optic anlage (ooa), the lateral proliferation zone (A), and to the medulla cortex from the other side, the medial proliferation zone (B). The first progeny from B are precocious medulla tangentials (MeT), growth cones from which weave across the medulla and intersect new columnar elements growing at the leading edge. Other tangential pathways correspond to the later paths of the anterior optic tract (Y) and of lobula tangential cells (Z). The inner optic anlage (ioa) proliferates in two directions: in direction C, to generate cells into the cortex of the lobula plate, and in direction D, to generate cell types T2, T3, or C, as judged only with reference to the positions of somata in the adult optic lobe. Axons of cellular progeny, and their growth cones, generate a plexus for each cortex, which will eventually form the adult neuropil. Crossing of fiber bundles between lamina and medulla results from the selective fasciculation of fiber pathways in a sequence of innervation like a conveyer belt and from the direction of approach between bundle and plexus. (Bundles penetrate the lamina cortex to innervate its plexus but grow along the inner margin of the medulla cortex to innervate the medulla.) Large glial cells lie along the paths of fibers in the external chiasma (ext.ch) and internal chiasma (int.ch); sg, subesophageal ganglion. (d) Proliferation from the lateral proliferation zone of the outer anlage (see diagram c). Cells are progressively displaced from neuroblasts in the outer anlage in a succession of stages in their cell cycle (G_2/M, G_1, S, G_2/M) around the lip of the anlage. Postmitotic cells lie in the lamina cortex (LaC), where they are innervated by photoreceptor axon bundles (arrowheads) from the optic stalk (os), which trigger the transition from G_1 to S in cells of the adjoining anlage (filled arrow) as well as the onset of differentiation and axonogenesis in cells already postmitotic (open arrow). (e) Ommatidial clusters mature in the eye disk, behind the morphogenetic furrow (mf), seen in elevation (clear profiles) and corresponding cross-sections in which nuclei are shaded. Arrows indicate directions of nuclear migration, matched in pairs of photoreceptors R1–R8 (labeled 1–8). Cross-sections from a (youngest) to f (oldest) are of preclusters (a, b), of immature (c) and symmetrical (d) eight-cell clusters, and of two-cone cell (e) and four-cone cell (f) stages. Cone cells are labeled 'C.' (f) R1–R8 in ommatidial cluster (corresponding to cross-section c in diagram e) comprise two central cells (R8, R7) and three pairs (R2/R5, R3/R4, R1/R6). Induction in R1–R6 involves a signal from R2/R5 that depends on expression of *rough* (*ro*) to induce development in R3/R4; *rough* product also appears in R3 and R4. The four cells defining the second line of symmetry (R3, R4, R1, and R6) all require the expression of *sevenup* (*svp*) to acquire their normal fates. (c) Modified from Meinertzhagen IA (1973) Development of compound eye and optic lobe in insects. In: Young D. (ed.) *Developmental Neurobiology of Arthropods*, pp. 51–104. Cambridge, UK: Cambridge University Press. (d) Modified from Selleck SB, Gonzales C, Glover DM, et al. (1992) Regulation of the G1-S transition in postembryonic neuronal precursors by axon ingrowth. *Nature* 355: 253–255. (e, f) Modified from Wolff T and Ready DF (1993) Pattern formation in the *Drosophila* retina. In: Bate M and Martinez Arias A (eds.) *The Development of Drosophila melanogaster*, pp. 1277–1325. Plainview, NY: Cold Spring Harbor Laboratory Press, with permission of the authors and Cold Spring Harbor Press.

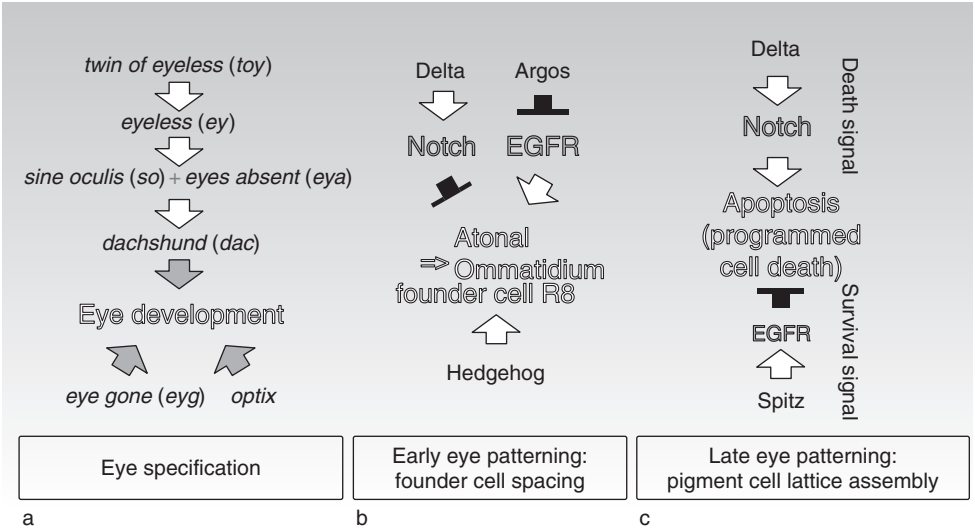

Figure 2 Selected signaling pathways during eye specification, founder cell spacing via lateral inhibition, and assembly of the pigment cell lattice via apoptosis in *Drosophila*. (a) Several nuclear factors are reported as so-called master control genes of eye development. Of these, all but two (*eyg* and *optix*) are part of a complex signaling network that has been proposed to be induced by Notch signaling and inhibited by epidermal growth factor receptor (EGFR) signaling. (b) Atonal expression determines the R8 founder cell and prepatterns the eye. Positive signals for R8 formation include Hedgehog and EGFR signaling, whereas lateral inhibition acts through the Notch pathway to establish correct founder cell spacing. (c) After ommatidial assembly, excess interommatidial cells are removed by apoptosis. Notch and EGFR signaling again exhibit antagonistic roles, providing death and survival signals, respectively.

mollusks and arthropods. Large, single-lens cephalic eyes of gastropod and cephalopod mollusks, which project directly to the brain, generally arise from an invaginating ectodermal placode, either *de novo* or during regeneration (*Strombus*, *Cardium*, *Helix*, etc.). Details vary depending on whether development is direct or via a veliger larva, and on the complexity of nonneural ocular components. Mantle eyes in bivalves have diverse origins and include both single-lens eyes and the compound eyes of arcid clams. In *Pecten* they arise by delamination of an epithelial placode that is secondarily reconstituted after migrating from the epidermis. All show spectacular regenerative capabilities. Arthropods generally have ocelli that arise by invagination, delamination, or both in some insect groups. Spiders have single-lens frontal eyes that arise by invagination and lateral eyes that form by delamination. In contrast, compound eyes form by similar means in all groups (by delamination *in situ* from an embryonic epidermal placode) to form a patterned array of modular ommatidia. In holometabolous insects, the retinal epithelium is secondarily internalized within specialized imaginal disks borne inside a postembryonic larva (Figure 1(a)). Reminiscent of these additions, the ocelli of diplopod myriapods arise from individual anlagen and accumulate in rows with each molt.

In ametabolous insects such as silverfish and bristletails, as well as in hemimetabolous insects (dragonflies, grasshoppers, cockroaches, etc.), eye development is a continuous process that starts during embryogenesis. In contrast, in holometabolous insects (such as *Drosophila*) visual system development consists of separate phases: an early phase during embryogenesis and a second phase during metamorphosis, separated by a 'developmental silence' during early larval stages. Compound eyes arise during the passage of a wave of development. In hemimetabolous insects and many Crustacea, the eye generated in the embryo is augmented by ommatidial increments added in adulthood; the wave is interrupted at each molt, its front contributing a succession of vertical strips of ommatidia at the eye's anterior margin. Regional specializations of the eye, such as foveae in dragonflies or mantids, therefore shift in the visual field with each molt, or are successively reconstructed. Intermolt growth in eye size generally results mostly from increased cell numbers rather than hypertrophy of existing cells, but species differ. The anterior eye margin in hemimetabolous insects is said to be a budding zone, harboring ommatidial precursor cells. In holometabolous insects, retinal development lays down ommatidial rudiments during a double postero-anterior wave of mitoses (e.g., *Drosophila* and *Ephestia*) over the largely two-dimensional epithelium of undifferentiated cells in the imaginal disk. These possibly overlap within the single wave of mitoses seen in the dragonfly.

Compound Eye Development in *Drosophila*

Compound eye development in *Drosophila* is well characterized and has become a model system to study a wide range of events in neural development, including cellular and molecular mechanisms of both tissue patterning and growth control, and the determination of cell fate. The adult eye comprises about 750 ommatidia, each in turn containing eight photoreceptor neurons, R1–R8. The rigid geometrical organization of these arises through iterated cell-fate decisions mediated by cell–cell communication within the undifferentiated imaginal disk. These occur during and after a wave of development, morphologically visible as a morphogenetic furrow initiated at the posterior margin of the disk, that progresses anteriorly (Figures 1(e) and 3(a)).

The first photoreceptor to differentiate from a field of uncommitted cells is the photoreceptor cell R8. Selection of R8 cells occurs initially to establish the correct spacing and positioning of future ommatidia (Figure 3(a)). The differentiation of R8 is also crucial because the cells of one ommatidium do not derive from a single clone, but are instead recruited by R8 and subsequent downstream events. Such recruitment occurs in pairs from the existing field of undifferentiated cells, with the first pair, R2/R5, followed by R3/R4 and R1/R6 (Figure 1(f)). Last, R7 is recruited. The initial steps of R8 selection and spacing are therefore crucially instructive for the patterned development of the eye and involve a well-characterized signaling network (Figure 2(b)). R8 differentiates under the control of the transcription factor encoded by *atonal,* which is initially expressed in all cells during the passage of the morphogenetic furrow. Behind the furrow *atonal* expression becomes restricted to the exact number of required R8 founder cells. In *atonal* mutants, ommatidial assembly fails entirely. Thus, the developing eye is prepatterned by selecting cells that express *atonal* behind the morphogenetic furrow (Figure 3(a)). Initially, all cells have the potential to become R8 and to repress others, but through lateral inhibition of their neighbors' fates only a few finally do so. Many genes encoding products required to signal this lateral inhibition are known from mutants with altered numbers of R8 cells, most prominently those involved in signaling via the Notch and epidermal growth factor (EGF) receptors (Figure 2(b)).

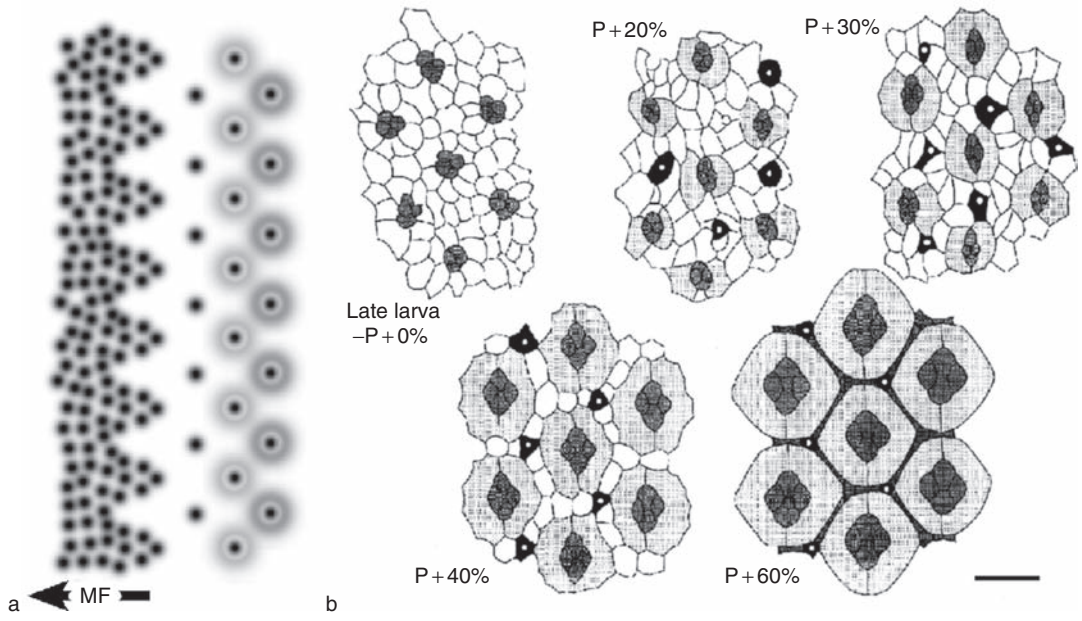

Figure 3 Maturation of *Drosophila* ommatidia. (a) As the morphogenetic furrow (MF) passes over the eye imaginal disk epithelium (arrow), the ommatidial array emerges with the differentiation of ommatidial founder R8 cells, which express *atonal* (filled circles) and prepattern the eye. (b) Pattern formation in the *Drosophila* eye. At late larval stages and the beginning of pupation (P + 0%), the eye disk exhibits the four-cone-cell stage (stippled) in its posterior region. In the 20-h pupal eye disk (P + 20%), cone cells are surrounded by two primary pigment cells (light stippling); bristle cells (black) appear between ommatidia. At P + 40%, approximately one-third of all interommatidial cells have been removed by apoptosis. Unfilled profiles between ommatidia are prospective secondary pigment cells and (at vertices) tertiary pigment cells. The P + 60% pupal eye exemplifies adult organization. Secondary (dark stippling) and tertiary pigment cells (at vertices) surround ommatidia. Scale bar = 10 μm; anterior to right. Modified from Wolff T and Ready DF (1993) Pattern formation in the *Drosophila* retina. In: Bate M and Martinez Arias A (eds.) *The Development of Drosophila melanogaster*, pp. 1277–1325. Plainview, NY: Cold Spring Harbor Laboratory Press, with permission of the authors and Cold Spring Harbor Press.

Molecules that have been proposed to act as inhibitors include Delta and Scabrous, both of which interact with Notch. Hedgehog and ligands of the EGF receptor (EGFR) are positive signals for founder cell specification (Figure 2(b)).

The subsequent recruitment of R2/R5 and R3/R4 creates preclusters containing five cells, evenly spaced behind the morphogenetic furrow (Figure 3(b)). Now postmitotic, the cells of these preclusters commence differentiation, whereas all other undifferentiated cells reenter the cell cycle in the second mitotic wave. Also born during that wave are the remaining 14 cells required to constitute an ommatidium. The differentiation of these cells as well as of all photoreceptors other than R8 is controlled by different signaling pathways, again by the EGFR pathway as well as the mitogen-activated protein (MAP) kinase and other receptor-mediated signaling pathways. Secretion of the EGFR ligand Spitz is required for all non-R8 photoreceptor differentiation. The first such pathway to be identified, the signaling cascade leading to the final recruitment and differentiation of R7, is particularly well characterized. A membrane-associated protein encoded solely in R8 by the *boss* locus interacts with the Sevenless receptor tyrosine kinase on the presumptive R7. Internalization of Boss protein in R7 and induction of the downstream Ras/Raf MAP kinase pathway lead to R7 differentiation.

Finally, all ommatidia are positioned within a hexagonal array with shared interommatidial cells in between. The assembly of the secondary and tertiary pigment cell lattice depends on the spatially restricted apoptosis of excess cells, which also involves a complex signaling network implicating the Notch and EGFR pathways (Figure 2(c)). In this late phase of eye development, Notch promotes apoptosis of excess interommatidial cells, while EGFR signaling, initiated by cone or primary pigment cells, opposes Notch (Figure 2(c)). As a result, approximately one-third of the interommatidial cells are eliminated before eye development concludes.

The Eye's Connection with Its Visual Centers and Their Development

Wiring the invertebrate eye to its brain is as diverse a process as are the eye types themselves. A common feature of all is the retinotopy of mapping the optical world upon the visual brain. This is documented at single-cell level in the insect compound eye, but is also a clear feature in, for example, the single-lens eye of cephalopods. Development among invertebrate visual centers is once again best described in the arthropod optic lobe. Visual interneurons are the progeny of neuroblasts that reside in two groups, the outer and inner proliferation centers, or optic anlagen (Figures 1(b) and 1(c)). In insects, the outer anlage generates the first neuropil, the lamina, and the distal part of the second, the medulla, while the inner anlage gives rise to the remaining medulla and deeper optic neuropils (Figure 1(c)). In both anlagen, the cells are born by a sequence of divisions that matches the sequence of ommatidial assembly in the compound eye, but begins first in the medulla. The match between eye and lamina in *Drosophila* is attained by a competition between targeting signals at the N-terminus and C-terminus of the secretory Hedgehog protein: after cleavage, the N-terminal domain is targeted to the retina, while the C-terminal domain overrides this path and transports Hedgehog along the photoreceptor axons, where it induces the development of target neurons in matching numbers.

In Crustacea, the number and arrangement of the optic neuropils are more diverse than in insects, and details of crustacean development are less widely known. In decapod Crustacea, the shape, location, and mitotic patterns of proliferation zone 2 (PZ2) closely resemble the outer optic anlage of insects. In the water-flea *Daphnia*, the counterpart of an outer optic anlage is the sole so-called optic lobe primordium (Figure 4(b)). In *Drosophila*, photoreceptor innervation triggers the final cell cycle of lamina precursor cells, procuring targets for the next wave of innervation and matching the proliferative activities of eye and optic lobe (Figure 1(d)). (Lamina glial cells, by contrast, arise independently from R1–R6 input and from a lineage that is distinct from that of lamina cells.) Photoreceptor cell axons innervate new ganglion cells, so produced, in successive rows corresponding to the wavefront of retinal differentiation. Most lamina cells are thereby enlisted, five per axon bundle, and cell death, presumably among supernumerary cells, concludes the wave of development in the outer neuropils. In *Daphnia*, newly innervated cells are invariably just postmitotic. As ingrowing photoreceptor axons recruit them in final groups of five (Figure 4(b)), each soma displays a stereotypic response in which it envelops the incoming axons, forming gap junctions with them transiently. The same occurs in *Drosophila*, but is less clearly analyzed. In *Daphnia*, all cells contacted exhibit this envelopment. They do so in the exact sequence of contact, but cells from which retinal innervation has been withheld do not, while envelopment is delayed in cells to which innervation arrives late. Envelopment manifests a presumed inductive signal to each lamina cell, after which the latter differentiates, elaborating a neurite that will follow the retinal axon bundle in a secondary wave.

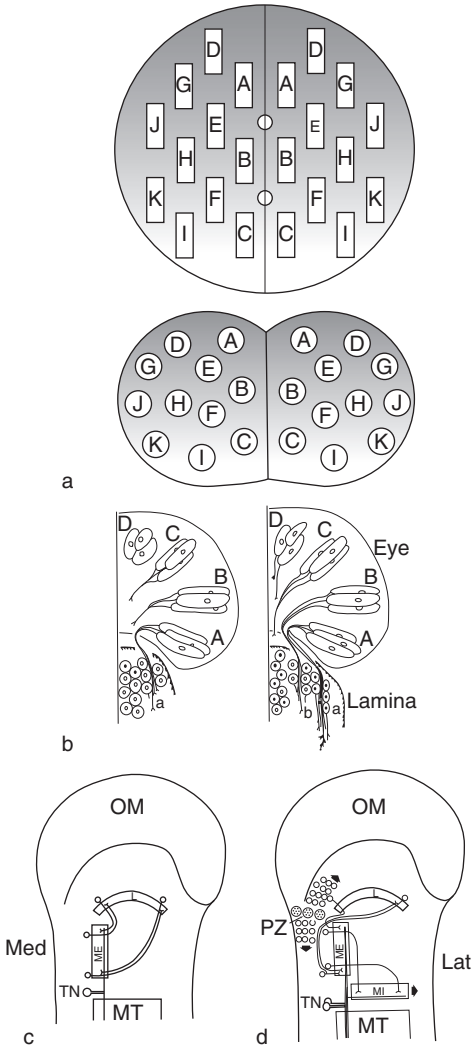

Figure 4 Development of the visual system in representative Crustacea. (a) Schematic diagram of the compound eye (top) and optic lamina (bottom) in *Daphnia*, showing locations of ommatidia (lettered rectangles) and of lamina cartridges (circles with corresponding letters) to which photoreceptor axons of each ommatidium project. Both are viewed in a transverse projection, dorsal side up, with the left eye to the left. The projection is direct, without an intervening chiasma. (b) Schematic representation of the development of the eye–lamina projection in *Daphnia* at an early stage in its development (left) and several hours later (right). Only the right halves of the eye and lamina are depicted, in a view perpendicular to that shown in diagram a. The ommatidia are shown in four rows (A–D), with each row at a different stage of development; cartridges a and b are depicted during assembly. At both stages illustrated the most lateral ommatidium, A, is also the most developed, while D, the most medial, is the least so. (Left) Axons from A have grown along the midline glial bridge that connects eye and lamina anlage and have reached the target region, where they begin to recruit lamina neurons to form cartridge a. (Right) Axons from B have followed the same path as those of A and complete the recruitment of five lamina neurons to form cartridge b, displacing cartridge a laterally. Axons from C and D will subsequently repeat this process until the projection is completed. (c, d) Assembly sequence of neuropil layers and axon projections between them in the optic lobes of a nonmalacostracan (c) and malacostracan (d) crustacean. In both examples, a common neuroblast population, the outer proliferation zone PZ2 (PZ) generates progeny in two antiparallel directions. These assemble the cell cortices of lamina (L) and medulla externa (ME) in chronotopic sequence, like a conveyor belt. After their birth, cells are continuously displaced from PZ2 to await innervation from overlying ommatidia (OM). The lamina and distal medulla in Crustacea both have major structural and developmental similarities to their possible evolutionary homologs in insects; malacostracan Crustacea grow chiasmal interconnecting pathways identical to those found in insects (diagram d). In nonmalacostracans, axons of cell bodies in the lamina cortex penetrate the lamina and enter the medulla externa from its neuropil face, thus traversing directly between corresponding neuropil faces. In Malacostraca, as in insects, on the other hand, a chiasma arises because the corresponding axons enter the medulla from the opposite direction, through its cortex, and in doing so invert their sequence. Tangential neurons (TN) from the medulla terminalis (MT) grow across the neuropil of the medulla externa and interna (MI), advancing to contact the newly grown retinal innervation and the higher order columnar pathways upon which it projects. (a, b) From Macagno E (1984) Formation of ordered connections in the visual system of *Daphnia magna*. *BioScience* 34: 308–312. (c, d) From Elofsson R and Dahl E (1970) The optic neuropiles and chiasmata of Crustacea. *Zeitschrift fur Zellforschung* 107: 343–360, used with permission of the authors and Springer-Verlag.

Axonal Outgrowth and Pathfinding

Photoreceptor axons grow centrally, innervating the lamina and distal medulla in a sequence that mirrors the ordered sequence of retinal differentiation. Outgrowth within the ommatidial axon bundle is asynchronous. The axon from one photoreceptor in *Daphnia* (R1) has a well-defined growth cone and is the first to grow and contact target cells. Axons of other cells following this lead axon (**Figure 4(b)**) have reduced growth cones or none. After R1 is ablated, the axon of another photoreceptor assumes the role of lead fiber, indicating that all photoreceptors are equivalent. In *Drosophila*, axon outgrowth occurs initially in groups of five (presumably R2–5 and 8), later supplemented by the remaining three. In many insect groups, photoreceptor axons grow postembryonically along the pathway provided by larval or nymphal visual systems. Best known is Bolwig's organ, a group of 12 internal photoreceptors at the tip of a fly larva, with optic lobe pioneers (their presumed interneurons in *Drosophila*), a larval nerve, and optic neuropil, all of which preempt the axon pathways of the earliest adult neurons, which assemble around them (**Figure 1(c)**). Bolwig's nerve is dispensable for adult wild-type photoreceptor axons to project normally, and is thus not an essential pioneer of axonal navigation to the lamina. Bolwig's organ later transforms into the four photoreceptors of an extraretinal posterior 'eyelet,' which entrains circadian rhythmicity in the larva. The terminals of Bolwig's nerve metamorphose into those of the eyelet nerve. Their branching parallels the arborization

of neurons immunoreactive to pigment-dispersing hormone, which serve as outputs from the circadian clock. In other insects, there are larval ocelli (in mosquitoes) and stemmatal nerves (in Lepidoptera and Coleoptera). Larval photoreceptor axons and adult axons in insects lacking larval visual centers grow in the embryo, thus initially spanning only short distances to reach their targets, but elongating dramatically with subsequent growth. Those in *Daphnia* grow initially over a glial substrate; later axon bundles follow the preexisting neural pathway to the center, reinforcing the fiber tract in which they grow while simultaneously providing the substrate for yet further growth (**Figure 4(b)**). The sequence in which new axon bundles accrue to existing fiber tracts, and their relative positions within the tract, confer the antero-posterior polarity of the resulting topographical map. Chiasmata form when this polarity inverts, as in the external chiasma of malacostracan Crustacea (**Figure 4(d)**) and insects, as revealed by subsequent rotation of the medulla in *Drosophila* (**Figure 1(c)**). Many arachnid groups also have chiasmata, but non-malacostracan Crustacea, such as *Daphnia*, lack them (**Figure 4(c)**). Eye rotation experiments in locusts indicate that ingrowing receptor axons fail to discriminate the particular column of neuropil in which they terminate. On the other hand, there is evidence that retinotopy is reinforced in *Drosophila* by guidance mechanisms that label both photoreceptor axons and their targets. For example, (1) in the mutants *sine oculis* and *Ellipse*, axons project to the correct retinotopic location even when not surrounded by axons from neighboring ommatidia, while (2) wild-type axons project normally, even when they neighbor mutant *glass* axons with aberrant projections. At a coarser level in *Drosophila*, mutant chiasmal phenotypes can result from the perturbed axon pathways of the larval optic lobe pioneer neurons (**Figure 1(c)**) caused by the displacement of their cell bodies in the larva.

Several *Drosophila* mutants have been isolated that affect pathfinding, target recognition, or medulla rotation. Photoreceptor axons R1–R6 enter the lamina as rapidly growing ommatidial bundles, where their growth cones stop between layers of glial cells; thus, the primary targets that terminate growth cone guidance are not the synaptic partners, but glia. Mutations in the *brakeless* gene cause the axons of R1–R6 to penetrate their normal lamina targets and project instead to the medulla, where normally only R7 and R8 terminate. A transcription factor encoded by the *runt* gene, which is normally repressed in R1–R6, is derepressed in *brakeless* mutants, and overexpression of *runt* causes the same mistargeting phenotype. Apart from these nuclear factors, genes encoding cell surface receptors, as well as downstream signal transduction proteins, have been characterized in *Drosophila*. Genes encoding cell surface receptors include *rst* (a transmembrane receptor of the immunoglobulin superfamily), *DPTP69D* and *DLar* (both receptor tyrosine phosphatases), *DN-Cadherin* (a classical cadherin), *Flamingo* (a protocadherin), *capricious* (a transmembrane receptor with LRR-repeats), and others. Mutants of all these genes exhibit targeting phenotypes of differing severities. Few studies have so far addressed the spatiotemporally controlled localization of guidance receptors during development. Mutants in the vesicle trafficking gene *sec15* suggest that trafficking of guidance receptors during axon targeting requires highly regulated intracellular trafficking. The best characterized signal transduction pathways required for photoreceptor growth cone guidance include *dreadlocks* (a SH2/SH3 adaptor protein), *pak* (p21-activated protein kinase), *trio* (a Rho family guanine exchange factor that activates Rac), *misshapen* (an Ste20-like serine/threonine kinase), and *bifocal* (a putative cytoskeletal regulator).

Synaptic Partner Selection and Synaptogenesis

Once in the lamina, the growth cone filopodia of R1–R6 in the fly continuously interact with neighboring growth cones in a distinct developmental step that follows initial target recognition until each gradually draws away from its parent ommatidial bundle. The growth cones then pass through a succession of shifting geometric arrays, apparently driven by filopodial interactions that are initially directed exclusively between R1–R6, but eventually directed to the growth cones of their monopolar cell targets, L1 and L2, around which R1–R6 assemble to form nascent cartridges. Among the previously mentioned genes, mutants in several (e.g., *DPTP69D*, *DLar*, *DN-Cadherin*, *Flamingo*) cause severe defects during this lateral sorting of photoreceptor terminals into the correct cartridges, whereas the protocadherin *flamingo* is required for the initial orientation of the R1–R6 bundle. In the medulla, the axon terminals of R7 and R8 initially occupy temporary strata relative to terminals of other cells, but then grow down deeper to their permanent stratum, R7 below R8. Synaptogenesis has been analyzed in detail only in the lamina, where it commences only in the second half of pupal development, after the R1–R6 terminals have grown centripetally, when neurites of lamina cells contact photoreceptor presynaptic sites. Starting with a neurite contact from L1 or L2, each synaptic site selectively accumulates other postsynaptic elements to constitute a final tetrad synapse (a synapse with four postsynaptic targets). Up to 50% of all contacts formed are lost. Although synaptic transmission

functions early, while synaptic junctions mature, pupal photoreceptors generate only small, slow responses to light and are not normally exposed to bright light anyway. In contrast to vertebrate visual systems, neither spontaneous nor evoked neuronal activity is required to form a normal visual map in the fly's brain.

The Morphogenetic Interdependence of Eye and Visual Centers

Development in the visual centers generally exhibits some dependence on innervation from the eye, but this is quantitatively variable; eye development is usually autonomous. In *Drosophila*, clones of mutant tissue with perturbed organization generated in either the eye or optic lobe reveal that abnormal pattern is communicated from the retina to the lamina, but not in the opposite direction. This refines conclusions from the coarser effects of deafferentation in many species. More central neuropils are volumetrically less affected by deafferentation than are peripheral ones. In congenitally eyeless mutant *Drosophila*, second-order neurons in the lamina never differentiate, but some medulla neurons grow axons, albeit their morphogenesis is abnormal, whereas many lobula neurons look quite normal and establish a regular array of neuropil columns. Tangential cells of all neuropils show complete autonomy from afferent innervation. Qualitatively similar results follow lesions in the embryonic *Daphnia* eye. Differentiation in the fly's lamina and medulla is largely unaffected by the absence of target cells on the lobula, indicating the relative absence of retrograde influences of central visual neuropils upon distal ones. On the other hand, ommatidia eventually degenerate after disconnection from the lamina, indicating that the optic lobe has a retrograde influence on the long-term maintenance of the retina.

Plasticity

Insect visual systems exhibit various forms of plasticity, many at a behavioral level. The developmental emergence of pattern discrimination and visual preference behavior in flies, for example, depends on prior visual experience, the flies preferring a pattern to which they were selectively exposed during a critical period as young adults. Various neurochemical correlates exist. Functional binocular vision, as revealed by prey-catching behavior in mantids, matures after emergence, and its development also exhibits a critical period. Praying mantises that are monocularly occluded as emerging nymphs nevertheless have normal monocular visual fixation and binocular distance estimation when they attain adulthood, indicating that binocular visual experience is not essential. Even so, monocular rearing initially impairs distance estimation, which then only gradually recovers; such recovery is progressively lost in older nymphs or adults. Bees reared in ultraviolet light are less sensitive to light of longer wavelengths; terminals of photoreceptors that peak in the green, in which the bees were deprived, also have fewer synapses. Other examples of structural plasticity are reported. In flies, lamina feedback synapses are more numerous in young dark-reared adults and, during subjective night, in flies held under constant darkness; these synapses also exhibit reactive synaptogenesis after losing their targets, R1–R6. Photoreceptor input synapses are not fixed synaptic sites: fly tetrads can both form and disappear rapidly (within minutes) in the adult, especially in response to functional reversals (e.g., light exposure after dark rearing; warm recovery after cold exposure). Central optic lobe neurons sprout after congenital deprivation of their visual inputs, procured either by lesion or mutation. The volume of the optic neuropils increases in *Drosophila* reared under 'enriched' conditions, relative to solitary rearing, and with various other rearing conditions. At the cellular level, L1 and L2 change the caliber of their lamina axons during a cycle of day/night changes.

See also: Animal Models of Inherited Retinal Degenerations; Learning and Memory in Invertebrates: *Drosophila*; Olfaction in Invertebrates: *Drosophila*; Photoreceptor Adaptation; Photoreceptors: Physiology; Retinal Horizontal Cells; Retinal Pharmacology: Inner Retinal Layers; Visual System: Invertebrates.

Further Reading

Arendt D, Tessmar K, de Campos-Baptista MI, et al. (2002) Development of pigment-cup eyes in the polychaete *Platynereis dumerilii* and evolutionary conservation of larval eyes in Bilateria. *Development* 129: 1143–1154.

Dickson B and Hafen E (1993) Genetic dissection of eye development in *Drosophila*. In: Bate M and Martinez Arias A (eds.) *The Development of Drosophila melanogaster*, pp. 1327–1362. Plainview, NY: Cold Spring Harbor Laboratory Press.

Gehring W (2004) Historical perspective on the development and evolution of eyes and photoreceptors. *International Journal of Developmental Biology* 48: 707–717.

Hafner GS and Tokarski TR (2001) Retinal development in the lobster *Homarus americanus*: Comparison with compound eyes of insects and other crustaceans. *Cell and Tissue Research* 305: 147–158.

Halder G, Callaerts P, and Gehring WJ (1995) Induction of ectopic eyes by targeted expression of the *eyeless* gene in *Drosophila*. *Science* 267: 1788–1792.

Harzsch S, Vilpoux K, Blackburn DC, et al. (2006) Evolution of arthropod visual systems: Development of the eyes and central visual pathways in the horseshoe crab *Limulus polyphemus* Linnaeus, 1758 (Chelicerata, Xiphosura). *Developmental Dynamics* 235: 2641–2655.

Macagno E (1984) Formation of ordered connections in the visual system of *Daphnia magna*. *BioScience* 34: 308–312.

Mast JD, Prakash S, Chen PL, et al. (2006) The mechanisms and molecules that connect photoreceptor axons to their targets in *Drosophila*. *Seminars in Cell and Developmental Biology* 17: 42–49.

Meinertzhagen IA (1990) Development of the squid's visual system. In: Gilbert DL, Adelman WJ Jr., and Arnold JM (eds.) *Squid as Experimental Animals*, pp. 399–419. New York: Plenum.

Meinertzhagen IA and Hanson TE (1993) The development of the optic lobe. In: Bate M and Martinez Arias A (eds.) *The Development of Drosophila melanogaster*, pp. 1363–1491. Plainview, NY: Cold Spring Harbor Laboratory Press.

Ready DF, Hanson TE, and Benzer S (1976) Development of the. *Drosophila retina, a neurocrystalline lattice*. *Developmental Biology* 53: 217–240.

Stierwald M, Yanze N, Bamert RP, et al. (2004) The *Sine oculis/Six* class family of homeobox genes in jellyfish with and without eyes: Development and eye regeneration. *Developmental Biology* 274: 70–81.

Warrant EJ and Nilssson DE (eds.) (2006) *Invertebrate Vision*. Cambridge, UK: Cambridge University Press.

Visual System: Adaptive Regression and Progression in Subterranean Mammals

E Nevo, Institute of Evolution, University of Haifa, Haifa, Israel

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The extensive convergent evolution of subterranean mammals across the planet represents a global, natural experiment of life without light under darkness stress (Figure 1). It began during the global climatic transition from the middle Eocene to the early Oligocene, 45–35 million years ago. It involved seasonal climatic stresses proceeding progressively throughout the Cenozoic. The open-country biota set the stage for a rapid evolutionary play of recurrent adaptive radiations of unrelated mammals on all continents to the subterranean ecotope. The biota comprised marsupial and placental mammals (insectivores and herbivores) that involve several hundreds of small mammalian species belonging, at present, to 50 genera, 11 families, and three mammalian orders.

The stressful underground ecotope (solid soil, darkness, low productivity, hypoxy, hypercapny, and high infectivity) provided subterranean mammals with shelter from extreme climatic (temperature and humidity) fluctuations and predators. All subterranean mammals share convergent molecular and organismal adaptations to their shared unique ecotope. The subterranean ecotope is simple, relatively stable, specialized, low in productivity, and discontinuous. Its major determinants are specialization, competition, and isolation. These characteristics led to convergent and divergent, regressive and progressive evolutionary patterns and processes, locally optimizing adaptations at all organizational levels for mammalian life underground.

Adaptive convergence comprises structural and functional reductions and expansions (hypertrophies) through molecular and organismal evolutionary tinkering. The reductions and expansions involve burrowing, energetics, and respiratory adaptations affecting molecular, morphological, physiological, and behavioral changes compared with small aboveground mammals. Likewise, life in underground darkness led to the differential loss of sight but simultaneously to the remarkable evolution of photoperiodic perception, affecting circadian rhythmicity through complex mosaic eye and brain evolution. Life underground in isolated territorial domains also triggered the evolution of aggression and seismic communication affecting the reorganization of the brain. Life underground also selected for a suite of progressive adaptive traits to the subterranean ecotope compensating for the regressive loss of sight. These traits involve vibratory or seismic, magnetic, auditory, chemical, nasal, and vomeronasal communication systems.

Evolutionary organismal and molecular tinkering involving regression, progression, and convergence of eye evolution advanced in subterranean mammals, climaxing in *Spalax*. The family Spalacidae of southwest Asia originated approximately 40 million years ago based on molecular data, although the earliest fossils from Greece and Turkey are younger. Spalacids are extremely adapted to life underground, which climaxed in mosaic eye and brain evolution.

Sensory and Brain Adaptations in Subterranean Mammals

Sense Organs

The senses of subterranean mammals display impressive, variable, structural, and functional convergent adaptations required for life underground. The sensory environment of subterranean mammals is deprived of the numerous aboveground complex stimuli. Diverse sensory communication modes are common to aboveground organisms: vision, photoperiodic perception, audition, smell, taste, touch, pressure, vibration, and electric and magnetic senses. Clearly, in extreme confinement to the underground dark environment, vision is variably excluded in many subterranean mammals and is replaced by other communication modalities. Here, the focus is on the visual apparatus, primarily on the extreme case of *Spalax*.

Visual Apparatus

Lightless subterranean ecotopes frequently lead to differential regression evolution or to the disappearance of the vertebrate and invertebrate eye. The complete disappearance of the eyes has not occurred in the evolution of any natural population of mammals. Maintenance of eye rudiments after millions of years of adaptation to a lightless environment might therefore suggest that such rudiments still have some developmental functions. Some laboratory mice strains can be completely eyeless, indicating that mutations in one or several genes can prevent eye development. Eyes of subterranean mammals display variable regressive stages in size, depending on whether the animal is strictly subterranean or occasionally emerges

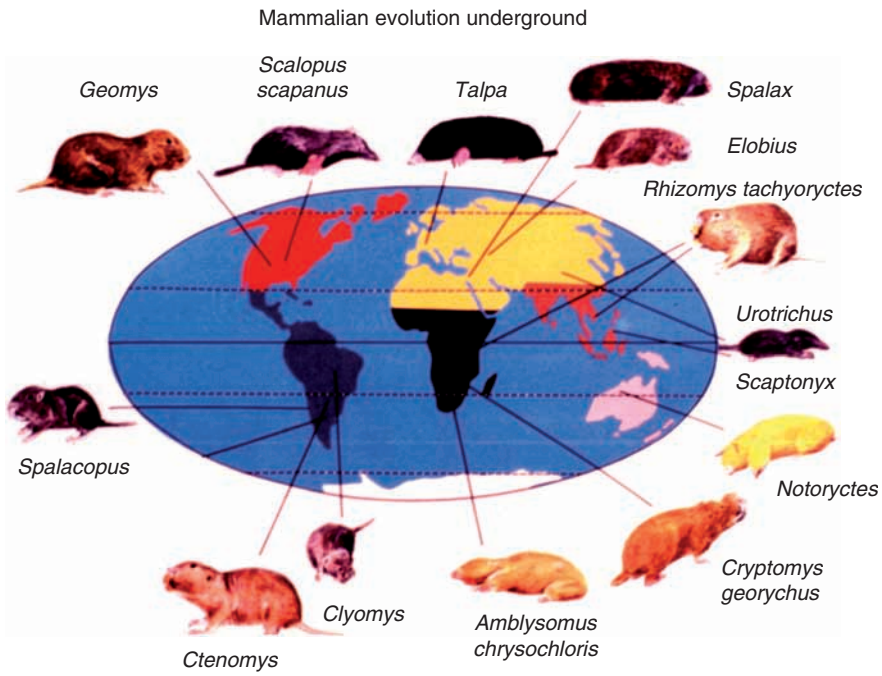

Figure 1 Distribution of subterranean mammals across the planet. Palearctic region: *Talpa* (Talpidae, insectivores), *Spalax* (Spalacidae, rodents; southeastern Europe, Turkey, Near East, North Africa), and *Ellobius* (arvicolidae, rodents; Asia). Ethiopian: *Chrysochloris* and *amblysomus* (Chrysochloridae, insectivores; South Africa), *Tachyoryctes* (Rhizomyidae, rodents; South Africa). Oriental: *Scaptonyx* and *Urotrichus* (Talpidae, insectivores: East Asia) and *Rhizomys* (Rhizomyidae, rodents). Australian: *Notoryctes* (Notoryctidae, marsupial moles; Australia). Nearctic: *Scalopus* and *Scapanus* (Talpidae, insectivores) and *Geomys* (Geomyidae, rodents). Neotropical: *Spalacopus* (Octodontidae, rodents), *Ctenomys* (Ctenomyidae, rodents), and *Clyomys* (Echimyidae, rodents). Different symbols mark the different zoogeographical regions. Reproduced from chapter 7 – Rodents: Herbivorous subterranean mammals, fig. 2.1, p. 36 from *Mosaic Evolution of Subterranean Mammals: Regression, Progression and Global Convergence* by Nevo, Eviatar (1999), by permission of Oxford University Press. Photo courtesy of Prof. Hynck Burda.

aboveground for foraging. Variation in eye size ranges from normal–medium (*Spalacopus*) to medium–small (ctenomyids and geomyids), small–minute (rhizomyids, subterranean cricetids, and bathyergids), minute (most talpids), vestigial, and completely covered by skin (*Notoryctes*, *Talpa caeca*, chrysochlorids, and spalacids) (Figure 1).

Minute microphthalmic eyes (<2 mm) are also found in the echolocating microchiropteran bats, species of insectivores and aquatic shrews, *Fukomys anselli*, and one cetacean – the blind river dolphin (*Platanista*). Reduction of the eyes is even more widespread in subterranean and cave-dwelling nonmammalian vertebrates, including cyclostomes, teleosts, amphibians, and reptiles. Birds are the only vertebrates in which small eyes are absent. Regression of the eye can thus be viewed as resulting from diverse parallel and convergent evolutionary causes, potentially leading to a mosaic of consequences affecting the peripheral and central visual system.

Variable degrees of reductions were described in eye size, ocular muscles, fixed eyelids, cornea thickening, amorphous lens, fibrous vitreous body, minimized optic nerve, regression in size of brain visual centers, and retinal changes in size and structure. Generally, the structure of the visual apparatus regresses differentially and structurally in all subterranean mammals, presumably correlated with confinement to life underground. The following section reviews the detailed study of the *Spalax* eye. This is clearly an extreme case of mosaic eye and brain evolution in subterranean mammals.

The Mosaic Evolution of the *Spalax* Eye: Morphology, Physiology, and Molecular Biology

The entirely subcutaneous minute eye of *Spalax* has been extensively studied. Concealed beneath the skin, the minute regressed eye is further embedded within a hypertrophied harderian gland. Thus, light can only penetrate to the retina with difficulty, and the thick covering of skin precludes proper image formation. The size of the eye is extremely reduced (700 μm diameter), although the neuroretina appears to be structured normally with thin inner and outer plexiform and nuclear cell layers. Ganglion cells are reported to be sparse, and the optic nerve contains less than 1000 unmyelinated fibers. Eye development

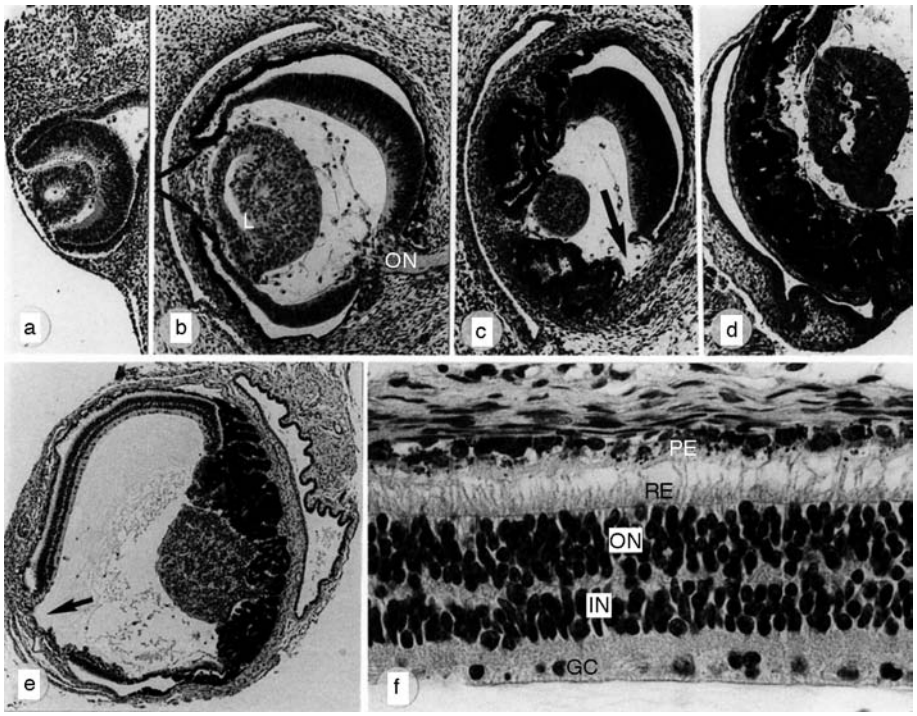

Figure 2 Light micrographs showing cross-sections of the developing eye of the mole rat, *Spalax ehrenbergi*. (a) Optic cup and lens vesicle initially develop normally (×100). (b) Eye at a later embryonic stage. Note the appearance of iris–ciliary body rudiment (arrows) and development of the lens nucleus (L). ON, optic nerve (×100). (c) Eye at a still later fetal stage. Note the massive growth of the iris–ciliary body complex colobomatous opening (arrow) (×100). (d) Early postnatal stage. The iris–ciliary body complex completely fills the chamber. The lens is vascularized and vacuolated (×100). (e) Adult eye. Eyelids are completely closed, and a pupil is absent. Note the atrophic appearance of the optic disc region (arrow) (×65). (f) Higher magnification of the adult retina. The different retinal layers are retained. PE, pigment epithelium; RE, receptor layer; ON, outer nuclear layer; IN, inner nuclear layer; GC, ganglion cell layer (×500). From Sanyal S, Jansen HG, De Grip WG, Nevo E, and De Jong WW (1990) The eye of the blind male rat. *Spalax ehrenbergi*: Rudiment with hidden function. *Investigative Ophthalmology and Visual Science* 31: 1398–1404.

is initiated normally but at an early stage displays a number of degenerative features (Figure 2). The optic fissure fails to close, the anterior chamber collapses, and the iris–ciliary body becomes hypertrophied and heavily pigmented. The lens remains vascularized and regressed. In contrast, neurogenesis develops normally, but the ganglion cell and photoreceptor layers remain relatively thin. Photoreceptors have been shown to form normal synapses in the outer plexiform layer. The visual system of *Spalax* presents a mosaic of both regressive and progressive features. The retina projects bilaterally to all visual structures, described as receiving retinal afferents in nonfossorial rodents. Structures involved in form analysis and visually guided behaviors are reduced in size by more than 90%, receive a sparse retinal innervation, and are cytoarchitecturally poorly differentiated. The main regressive feature is the drastic relative reduction of retinal input to the superior colliculus (Figures 3(a)–3(c)). The volumes of visual structures in *Spalax*, hamster, *Spalacopus*, rat, and microphthalmic rat appear in Figure 3(c).

In contrast to the foregoing, structures of the 'non-image-forming' visual pathway involved in photoperiodic perception are well developed in *Spalax*. The suprachiasmatic nucleus receives a bilateral projection from the retina and the absolute size, cytoarchitecture, density, and distribution of retinal afferents in *Spalax* are comparable with those of other rodents (Figures 3(a)–3(c)). A relatively hypertrophied retinal projection is observed in the bed nucleus of the stria terminalis. The percentage of relative retinal input to the size of primary visual structures is presented in Figure 3(a). The percentage of relative difference of retinal input to size of visual structure of *Spalax* contrasted with other aboveground, fossorial, and subterranean rodents is given in Figure 3(b).

These results indicate that the apparently global, morphological regression of the visual system conceals a selective progression of structures related to functions of photoperiodic perception and photoneuroendocrine regulation. The evolution of an atrophied eye and a reduced visual system is an adaptively advantageous response to the unique

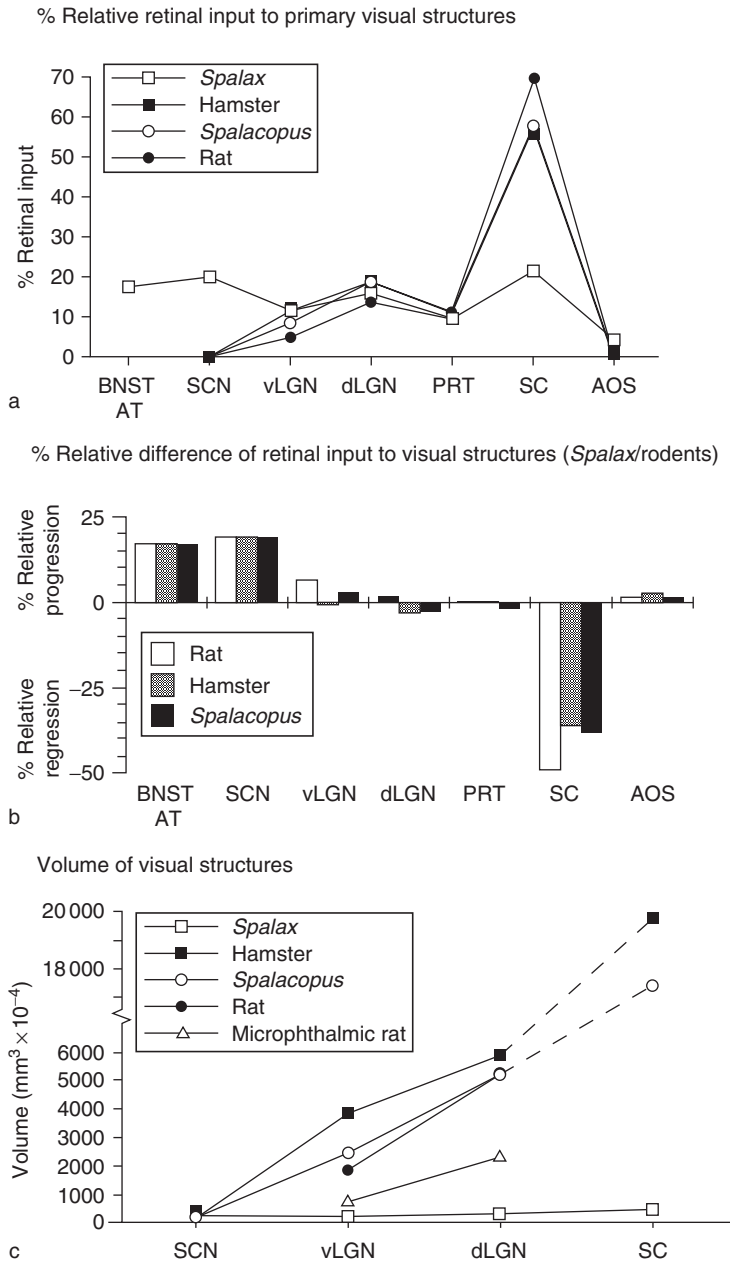

Figure 3 (a) Relative degree of retinal input to primary visual structures in *Spalax*, hamster, rat, and *Spalacopus cyanus*. (b) Relative degree of change in the proportions of retinal input to different primary visual structures in *Spalax* compared to measures obtained in other rodents. Relative progressive development in *Spalax* is seen in structures involved in photoperiodic and neuroendocrine functions. The main regressive feature is the drastic relative reduction of retinal input to the superior colliculus. (c) Comparison of the absolute size (volume, $mm^3 \times 10^{-4}$) of several visual structures in *Spalax* and other rodents. The size of the suprachaismatic nucleus (SCN) is equivalent in all species. The ventral lateral geniculate nucleus (vLGN) and dorsal lateral geniculate nucleus (dLGN) are reduced by approximately 90% and 94%, respectively. The retino recipient layers of the superior colliculus (SC) are reduced by 98%. AOS, accessory optic system; BNSt, bed nucleus of the stria terminalis; PRT, pretectum. Data from Caballero B, Tomas-Zapico CI, Vega-Naredo I, et al. (2006) Antioxidant activity in *Spalax ehrenbergi*: A possible adaptation to underground stress. *Journal of Comparative Physiology A* 192 (7): 753–759; and Cooper HM, Herbin M, and Nero E (1993) Ocular regression conceals adaptive progression of the visual system in a blind subterranean mammal. *Nature* 361: 156–159.

subterranean environment. Factors favoring regression include mechanical aspects, metabolic constraints, and competition between sensory systems. The primary advantage of sensory atrophy is the metabolic economy gained by the reduction of visual structures that do not contribute significantly to the animal's fitness. Under total darkness, *Spalax* evolved a purely circadian eye, thus following underground circadian changes

aboveground. The following sections provide a description of the lens crystalline, starting with the drastically changed lens αA-crystallin followed by αB-crystallin of *Spalax* at the molecular level of existence and expression.

αA-Crystallin: Structure, Expression, and Evolution

αA-Crystallin apparently originated from the multigene small heat shock proteins (HSPs). The HSPs are induced upon stress and evolved in all eukaryotes by gene duplication and the subsequent sequence divergence that allows adaptation to novel functions, such as lens constituent. Because αA-crystallin is lens specific, we focused on its expression and structure at the DNA level.

Hendriks et al. determined the complete structure of the single-copy αA-crystallin gene of the blind mole rat superspecies, *Spalax ehrenbergi*, in order to elucidate the evolutionary effects of the loss of vision on a lens-specific protein and its gene. The αA-crystallin gene of *Spalax* consists of four exons and three introns, as found in the mouse and hamster. The gene appears to have all the necessary transcriptional and translational signal sequences to be expressed in the rudimentary lens of the mole rat and probably gives rise to two protein products by means of alternative splicing, as in sighted rodents. Comparisons of the αA-crystallin sequence of the blind mole rat with that of sighted rodents (involving representatives of 32 rodent families) reveal a considerable acceleration of the substitution rate at nonsynonymous positions in the mole rat lineage. These comparisons, however, demonstrate a selective constraint.

Adaptive Evolution of Small Heat Shock Protein/αB-Crystallin Promoter Activity of the Blind Mole Rat, *Spalax ehrenbergi*

Hough et al. compared the tissue specificity of the small heat shock protein (shsp)/αB-crystallin promoter of the mole rat superspecies, *S. ehrenbergi*, with that of the mouse. Earlier experiments showed that mouse shsp/αB-crystallin promoter/enhancer activity is high in the lens and moderate in the heart and skeletal muscle of transgenic mice. Using the firefly luciferase reporter gene in transgenic mouse experiments, Hough et al. showed that despite relatively few changes in sequence, the mole rat shsp/αB-crystallin promoter/enhancer had selectively lost lens activity after 13.5 days of embryogenesis (Figure 4). The ratios of mole rat/mouse promoter activity were 0.01 for lens, 1.7 for heart, and 13.6 for skeletal muscle in 8-week-old transgenic mice. These data indicate that the shsp/αB-crystallin promoter/enhancer has undergone adaptive changes corresponding to the subterranean evolution of the blind mole rat. Hough et al. speculated that selective pressures on the metabolic economy may have contributed to these tissue-specific modifications of promoter/enhancer function during adaptation to life underground. Whereas the gene differentially loses its lens function (regression), it dramatically retains and even augments muscle function (progression) in transgenic mice (Figure 4).

Intriguingly, it appears that minor sequence modifications, presumably associated with adaptation to life underground, can be associated with profound changes in promoter activity. These adaptive evolutionary changes may economize metabolic expenses by limiting expression of shsp/αB-crystallin in the degenerated lens but augmenting expression in the muscle, thereby increasing fitness.

Retinal Photopigments in *Spalax*

Coneopsin

The blind eye of *Spalax* was retained subcutaneously as a circadian eye to assess photoperiodicity in total darkness underground. *Spalax* retina harbors three circadian photopigments – coneopsin, rhodopsin, and melanopsin – mediating the circadian photoreception. The atrophied eyes of *Spalax* express a long wavelength-sensitive (LWS) cone opsin. This photopigment is spectrally tuned to enhance photon capture in the red light environment of the eye. The LWS opsin of *Spalax* acts as a functional photopigment, and it is not a 'residue' of the pre-subterranean visual system but is implicated in the entrainment of circadian rhythms to environmental light.

Rhodopsin

The visual rod pigment of *Spalax* shows >90% homology with mammalian rod pigments. This *Spalax* rod pigment is fully functional *in vitro* but also shows that all requirements for a functional pigment are present *in vivo*. During adaptation to a subterranean lifestyle, the visual system of *Spalax* has undergone mosaic reorganization, and the visual pigments have adapted to function in circadian photoreception.

Melanopsin

Hannibal et al. have shown that melanopsin, a recently identified opsin, is expressed in retinal ganglion cells that also co-store pituitary adenylate cyclase-activating peptide (PACAP), a neurotransmitter of the retinohypothalamic tract. The melanopsin component of retinal ganglion cells in the *Spalax* retina is well conserved, resulting in a relatively higher density of melanopsin-positive cells per area compared to the rat.

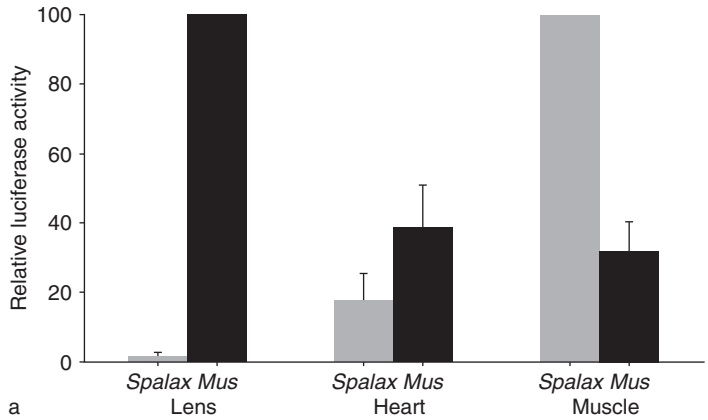

		Lens	Heart	Muscle
	Spalax	29 ± 26	782 ± 180	8891 ± 2778
	Mus	2256 ± 518	449 ± 176	652 ± 296
b	*Spalax/Mus*	0.01	1.7	13.6

Figure 4 Mole rat and mouse shsp/αB-crystallin promoter activity in 8-week-old transgenic mice. (a) Average of four lines of transgenic mice with the mole rat promoter (gray bars) and average of three lines of transgenic mice with the mouse promoter (black bars). The highest activity using the mole rat construct (muscle) and the highest activity using the mouse construct (lens) were given a value of 100. Setting the mole rat muscle activity to 100 lowered the heart activity relative to the mouse even though the average luciferase activity of the mole rat construct was actually higher than that of the mouse. (b) Average relative luciferase units per microgram of soluble protein, with standard deviations of the lens, heart, and skeletal muscle in the transgenic mice. The differences in activity between the mole rat and the mouse promoters were statistically significant in the lens ($p < 0.001$) and in the muscle ($p < 0.02$), whereas in the heart the differences were not statistically significant. From Hough RB, Avivi A, Joel A, et al. (2002) Adaptive evolution of small heat shock protein/αB-crystallin promoter activity of the blind subterranean mole rat. *Spalax ehrenbergi. Proceedings of the National Academy of Sciences of the United States of America* 99: 8145–8150.

Adaptive Loss of Ultraviolet-Sensitive/Violet-Sensitive Coneopsin

Most mammals are dichromats possessing, in addition to a single rod photopigment, two classes of cone photopigment LWS and ultraviolet-sensitive/violet-sensitive with differing spectral sensitivities, which mediate color vision. David-Gray et al. explored whether *Spalax* is a dichromat and has the potential to use color discrimination for photo-entrainment. Using immunocytochemistry and molecular approaches, they demonstrated that *Spalax* is a LWS monochromat.

The Evolution of Eyes and Photoreceptors

A monophyletic rather than a polyphyletic origin of the eye is supported by showing that *Pax6* (and its target sites) is the universally conserved master control gene for eye morphogenesis and eye evolution in all bilaterian metazoans from planarians to mammals (i.e., vertebrates and invertebrates). The prototypic eyes postulated by Darwin to consist of two cells only, a photoreceptor and a pigment cell, were accidentally controlled by *Pax6* and the subsequent evolution of the various eye types occurred by building on this original program. Intercalary evolution was proposed assuming that the eye morphogenetic pathway is progressively (or regressively) modified by intercalation of genes between the master control genes on the top of the hierarchy and the structural genes such as rhodopsin at the bottom.

The *eyeless* (ey) mutation in *Drosophila*, the *Small eye* mutation in mice, and the *Aniridia* gene in humans all correspond to the highly conserved *Pax6* gene, which might be a master control gene specifying eye development in both vertebrates and insects. A parallel situation in rodents applies to the *Sey* = *Pax6* gene, which is expressed first and controls a set of subordinate regulatory genes, including genes that influence cell–cell interactions and signal transduction. The *Sey* = *Pax6* gene in the mouse may be the master control gene in the mouse eye induction process and hence also in *Spalax* and other subterranean mammals. *Pax6* was also fully cloned in *Spalax*. As suggested by Gehring, the various *Pax* genes found in mice and flies (and also in *Spalax*) illustrate the principle of evolutionary tinkering, recombining bits and pieces of preexisting genes such as paired domains, homeodomains, octopeptides, and parts thereof. Evolution operates like a tinkerer rather than like an engineer.

The nematode *Caenorhabditis elegans* lost its eyes because it lives underground, but it retained its *Pax6*, like *Spalax*, with its reduced eye. Since photoperiodicity is retained in *Spalax*, the pleiotropic *Pax6* is essential for both the circadian eye and brain parts underlying photoperiodicity. Loss-of-function mutants in mice led to the reduction of the eyes and/or abolition of optic cups, and consequently the mutants lack eyes. Major evolutionary changes occur at the level of gene regulation, and coupled eye regression and progression might be generated by the same set of regulatory genes and/or by varying intercalary evolution of homeotic genes coupled with the genetics of loss- and gain-of-function mutations. The recombination of various coding and regulatory regions of different genes by 'evolutionary tinkering' might lead to recruitment and gene intercalation causing progression, whereas loss-of-function mutations might cause regression. Both gene groups may work epistatically and be under the *Pax6* master control gene involved in eye morphogenesis, resulting in mosaic eye evolution.

Brain Evolution of *Spalax*

No visual cortex has been identified in *Spalax*, either structurally or electrophysiologically. Instead, the occipital cortex of *Spalax* is occupied by the extension of the somatosensory cortex. Likewise, the somatosensory thalamus of *Spalax* almost reached the dorsolateral surface, without evidence of lateral geniculate body, the first station in the visual pathway. These findings indicate that the thalamocortical visual system in *Spalax* is minute, whereas the somatosensory system is expanded. Clearly, the brain area normally occupied by the visual cortex serves somatosensory functions. These findings, in addition to the prevalence of seismic communication in *Spalax* evolution, suggest that touch replaced vision as a major communication modality in the underground life of *Spalax* and probably in other subterranean mammals, although this calls for future investigation. Brain and eye mosaic evolutions might have evolved along similar patterns of evolutionary tinkering.

Conclusions

Blind subterranean mammals of the genus *Spalax* (Spalacidae and Rodentia) represent an extreme example involving combined natural visual degeneration and progression in mammals. Visual pathways are regressed and incomplete, and the absence of visual cortical potentials or behavioral response to light suggests that *Spalax* is completely blind.

Remarkably, however, the ocular regression in *Spalax* conceals adaptive progression of the visual system. The retina is functional in light perception, as indicated by structural and molecular evidence. Entrainment of circadian locomotory and thermoregulatory rhythms by ambient light demonstrates a capacity for photoperiodic perception. The severe thalamic and tectal structures' regression, involved in form and motion perception, is coupled with a selective hypertrophy of hypothalamic structures underlying photoperiodic perception. Ganglion cell retinal projections to the suprachaismatic nucleus (SCN) and the retinohypothalamic tract are increased 20-fold, and the SCN in which the circadian pacemaker resides is identical in *Spalax* and sighted rodents.

Photoentrainment is mediated by three photopigments: rhodopsin, coneopsin, and melanospsin. The differential regression and progression of visual structures and photopigments can be explained as an adaptive response to life underground. Natural selection appears to have operated via evolutionary tinkering, producing local optimization with available materials, constraints, and contingencies. Eye evolution in *Spalax* and other subterranean mammals evolved slowly by tinkering, which modified adaptively available molecular and organismal structures to generate emerging novelties. Only extensive future research could unravel the diverse modes of evolutionary tinkering in subterranean mammals.

Prospects

Molecular biology revolutionized evolutionary studies of the eye. Loss-of-function mutations in both insect and mammalian genes cause reduction or absence of eye structures, which suggest the function of the gene *Pax6* as a master control gene for eye morphogenesis across the metazoa. It remains to be determined whether *Pax6* is the master control for both reduction and expansion hypertrophy in the mosaic eye evolution of *Spalax* and other subterranean mammals. Future research will decipher the underlying genetic basis of mosaic eye evolution in *Spalax* on the basis of mutant phenotypes in homeotic gene clusters. These molecular genetic studies will highlight the molecular evolutionary tinkering within phylogenetic constraints underlying mosaic eye evolution, as it does in the evolutionary process in general.

See also: Entrainment of Circadian Rhythms by Light; Photoreceptor Adaptation; Photoreceptor Mosaic; Photoreceptors and Circadian Clocks; Photoreceptors: Physiology; Sensory System Specializations; Vertebrate Eyes: Evolution; Visual System Development: Invertebrates; Visual System: Invertebrates.

Further Reading

Argamaso SM, Froelich AC, McCall MA, et al. (1995) Photopigments and circadian systems of vertebrates. *Biophysical Chemistry* 56: 3–11.

Burda H, Bruns V, and Muller M (1990) Sensory adaptations in subterranean mammals. In: Nevo E and Reig OA (eds.) *Evolution of Subterranean Mammals at the Organismal and Molecular Levels*, pp. 269–294. New York: Wiley–Liss.

Caballero B, Tomas-Zapico CI, Vega-Naredo I, et al. (2006) Antioxidant activity in *Spalax ehrenbergi*: A possible adaptation to underground stress. *Journal of Comparative Physiology A* 192 (7): 753–759.

Chase HB (1945) Studies in an anophthalmic strain of mice. *Journal of Comparative Neurology* 83: 121–140.

Cooper HM, Herbin M, and Nevo E (1993) Ocular regression conceals adaptive progression of the visual system in a blind subterranean mammal. *Nature* 361: 156–159.

Cooper HM, Herbin M, and Nevo E (1993) The visual system of a naturally microphthalmic mammal: The blind mole rat. *Spalax ehrenbergi*. *Journal of Comparative Neurology* 328: 313–350.

Cooper HM, Herbin M, Nevo E, and Negroni J (1995) Neuroanatomical consequences of microphthalmia in mammals. Les seminaries Ophthalmologiques d'ISPEN. In: Christen Y, Doly M, and Droy-Lefaix MT (eds.) *Vision et Adaptation,* vol. 6, pp. 127–139. Paris: Elsevier.

David-Gray ZK, Bellingham J, Munoz M, et al. (2002) Adaptive loss of ultraviolet-sensitive/violet-sensitive (UVS/VS) coneopsin in the blind mole rat (*Spalax ehrenbergi*). *European Journal of Neuroscience* 16: 1186–1194.

David-Gray ZK, Cooper HM, Jannsen WH, Nevo E, and Foster RG (1999) Spectral tuning of a circadian photopigment in subterranean 'blind' mammals (*Spalax ehrenbergi*). *FEBS Letters* 461: 334–343.

De Jong WW, Hendriks W, Sanyal S, and Nevo E (1990) The eye of the blind mole rat (*Spalax ehrenbergi*): Regressive evolution at the molecular level. In: Nevo E and Reig OE (eds.) *Evolution of Subterranean Mammals at the Organismal and Molecular Levels*, pp. 383–396. New York: Wiley–Liss.

Duke-Elder S (1958) System of ophthalmology. In: *The Eye in Evolution*. London: Henry Kimpton.

Gehring WJ (2005) New perspectives on eye development and the evolution of eyes and photoreceptors. *Heredity* 96(3): 171–184.

Hannibal J, Hindersson P, Nevo E, and Fahrenkrug J (2002) The circadian photopigment melanopsin is expressed in the blind subterranean mole rat. *Spalax*. *NeuroReport* 13: 1411–1414.

Hendriks W, Luenisson J, Bloemendal H, Nevo E, and De Jong WW (1987) The lens protein αA-crystalline in the blind mole rat. *Spalax ehrenbergi*: Evolutionary change and functional constraints. *Proceedings of the National Academy of Sciences of the United States of America* 84: 5320–5324.

Hough RB, Avivi A, Joel A, et al. (2002) Adaptive evolution of small heat shock protein/αB-crystallin promoter activity of the blind subterranean mole rat. *Spalax ehrenbergi*. *Proceedings of the National Academy of Sciences of the United States of America* 99: 8145–8150.

Jacob R (1977) Evolution and tinkering. *Science* 196: 1161–1166.

Janssen JW, Bovee-Geurts PH, Peeters ZP, et al. (2000) A fully functional rod visual pigment in a blind mammal: A case for adaptive functional reorganization? *Journal of Biological Chemistry* 275: 38674–38679.

Necker R, Rehkamper G, and Nevo E (1992) Electrophysiological mapping of body representation in the cortex of the blind mole rat. *NeuroReport* 3: 505–508.

Němec P, Burda H, and Peichl L (2004) Subcortical visual system of the African mole-rat *Cryptomys anselli*: To see or not to see? *European Journal of Neuroscience* 20: 757–768.

Nevo E (1979) Adaptive convergence and divergence of subterranean mammals. *Annual Review of Ecology and Systematics* 10: 269–308.

Nevo E (1999) *Mosaic Evolution of Subterranean Mammals: Regression, Progression and Global Convergence*. Oxford: Oxford University Press.

Nevo E, Ivanitskaya E, and Beiles A (2001) *Adaptive Radiation of Blind Subterranean Mole Rats: Naming and Revisiting the Four Sibling Species of the Spalax ehrenbergi Superspecies in Israel: Spalax galili* ($2n = 52$), *S. golani* ($2n = 54$), *S. carmeli* ($2n = 58$), *S. judaei* ($2n = 60$). Leiden, The Netherlands: Backhuys.

Peicl L, Nemec P, and Burda H (2004) Unusual cone and rod properties in subterranean African mole-rats (*Rodentia, Bathyergidae*). *European Journal of Neuroscience* 19: 1545–1558.

Pevet P, Heth G, Haim A, and Nevo E (1984) Photoperiod perception in the blind mole rat (*Spalax ehrenbergi* Nehring): Involvement of the harderian gland, atrophied eyes, and melatonin. *Journal of Experimental Zoology* 232: 41–50.

Rehkamper G, Necker R, and Nevo E (1994) Functional anatomy of the thalamus in the blind mole rat *Spalax ehrenbergi*: An architectonic and electrophysiologically controlled tracing study. *Journal of Comparative Neurology* 347: 570–584.

Sanyal S, Jansen HG, De Grip WG, Nevo E, and De Jong WW (1990) The eye of the blind mole rat. *Spalax ehrenbergi*: Rudiment with hidden function. *Investigative Ophthalmology and Visual Science* 31: 1398–1404.

Walls GL (1967) *The Vertebrate Eye and Its Adaptive Radiation*. New York: Hafner.

Visual System: Functional Architecture of Area V2

A W Roe, G Chen, and H D Lu, Vanderbilt University, Nashville, TN, USA

© 2009 Elsevier Ltd. All rights reserved.

Competing Local versus Global Views

The visual world appears to us in its entirety, coherent and understandable. Despite our belief that this view is unitary, it is actually comprised of multiple competing views at the local and global scales. This point cannot be better illustrated than by the paintings of Salvador Dali, the quintessential craftsman of depicting competing and conflicting cues within a single scene. In *The Hallucinogenic Toreador* (Figure 1), which is a 13 foot × 10 foot (4 m × 3 m) painting (exhibited at the Salvador Dali museum in St. Petersburg, FL), different scenes become apparent depending on the viewing distance. When standing close to the painting, in the bottom right corner a small boy is seen viewing a lake with a woman in a bikini on a pool chair and a dalmatian dog headed toward the lake. From a little farther back from the painting, just above the lake a bull can be seen stuck with multiple colorful spears. Several Venus statues can be seen across the center of the painting, one with a white dress, partly shadowed in green, and another with a red dress. At an even greater distance, one begins to appreciate the toreador himself, who's shirt and green tie are formed by the fabric and shadow of Venus's white skirt, respectively, his chin by her stomach, his nose by her breast, and his eye by the shadow of her neck. The outline of his cheek is formed by the hips of the Venus draped in red, his beret and cape dotted by the patterns of flies in the stadium and those above the bull. Thus, depending on the scale at which this painting is viewed, different scenes become apparent to the viewer. Moreover, an attempt to take in all views at once may lead to a sense of rivalry between local and global scales: at one scale the green swath is a tie, and at another it is the shadow of Venus's white dress.

A few simple visual illusions may help break this complex image into simpler components. Such illusions can be very useful tools for studying how competing cues are encoded in the brain. One example is shown in Figure 2(a). In this 'watercolor illusion,' blue and yellow dotted lines separate regions of a white background. Regions between the yellow lines appear yellowish and those between the blue lines appear bluish. In actuality (as measured by a spectrophotometer), the white regions are all the same in color. This is a demonstration of how local cues (white) can be overridden by distant, global cues (the blue and yellow lines), leaving the illusion of differently colored regions. Figure 2(b) is the classic 'cafe wall' illusion, in which parallel horizontal lines appear to be sloping. Superimposition of a horizontal bar demonstrates that these lines are indeed horizontal. Again, the local cues signaling horizontal are overridden by the global percept of sloping lines. In the realm of depth perception, in the random dot stereogram (RDS) illustrated in Figure 2(c), the borders of the square central region in the left figure are offset from that in the right figure. When fused by the two eyes, this induces a percept of a square surface behind (more distant than) four round holes in a nearer surface. This is perceived despite the fact that the pixels within the square are identical and have no disparity. Thus, the local disparity cues are overridden by the global border cues of the square. Similar illusions are used in the studies described in this article to distinguish the roles of primary (V1) versus secondary (V2) visual areas in the perception of contour, stereoscopic depth, and color and brightness (see Table 1). Moreover, the conflicting local versus global cues present in these illusions are used to elicit competition between V1 (which is local) and V2 (which is global).

The inherent conflict illustrated by these examples calls for a revision of the traditional view of hierarchical processing within the visual pathways. In the traditional view, early visual stages first identify elemental features by means of neurons with small, spatially and 'featurally' restricted receptive fields. These elemental inputs are then integrated into higher-order representations at higher cortical areas. Receptive fields of neurons in higher cortical areas are spatially larger and tend to be less sensitive to elemental 'featural' specificity and, via some unknown 'invariance circuitry,' become tuned to more-global features. This integration process is iterated at multiple cortical levels, culminating in a relatively complete spatial and featural representation that enables object recognition. This hierarchical view, in its strictest interpretation, raises the question of how the visual system maintains knowledge of details when higher cortical areas no longer track these details. It is apparent that the details encoded in earlier areas and the global views encoded in higher areas must be simultaneously accessible. Moreover, these different views can (and often do) give conflicting signals. Resolving this local versus global competition is likely to be context dependent and therefore dynamic. Thus, while the hierarchical view still carries validity, it alone is inadequate for creating the illusion of coherence across multiple spatial scales.

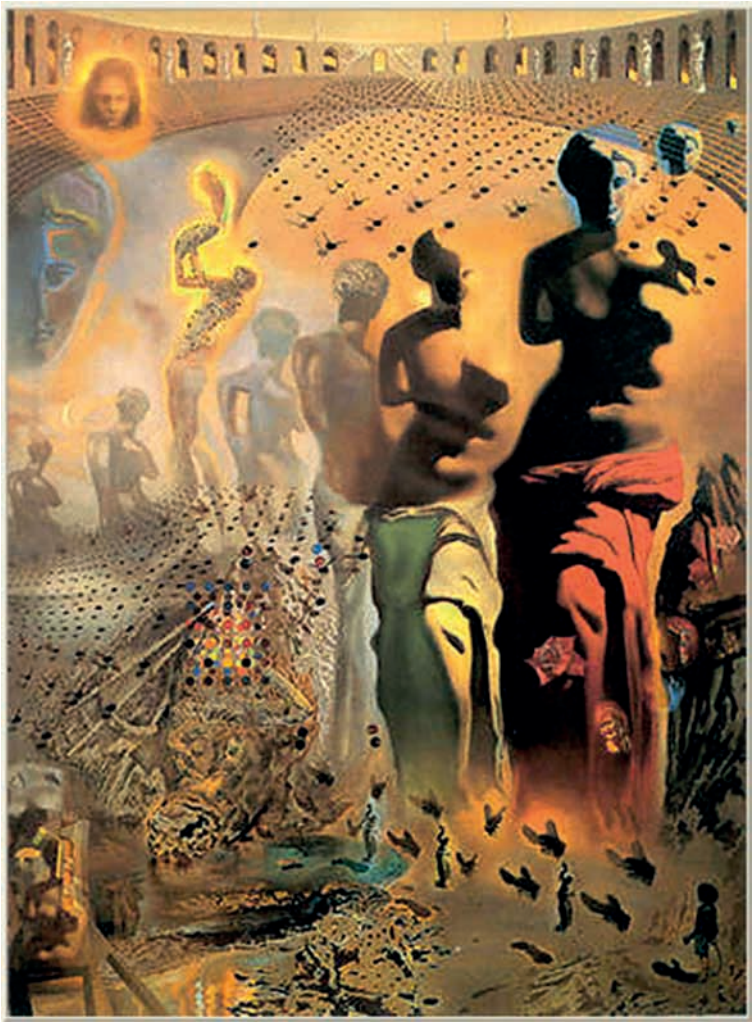

Figure 1 *The Hallucinogenic Toreador* by Salvador Dali (Salvador Dali Museum, St. Petersburg, FL).

This competition forms the basis for three theses presented in this article. Evidence supporting the first two theses is presented, and the third thesis is discussed largely as a future direction of study. The theses are as follows:

1. Each cortical area processes the visual world at a different spatial scale and, in that sense, 'sees' a different visual world.
2. Multiple views of the world which are encoded in parallel by multiple cortical areas can compete with each other. The dominance of one view over another may be resolved by a competitive balance between 'feedforward' and 'feedback' signals between different cortical areas.
3. This competition is dynamic and context dependent.

These theses are presented in this article through a set of studies which examine responses of primary (V1) and secondary (V2) visual areas in the macaque monkey. In this framework, V1's and V2's views of three featural spaces are examined: visual contours, visual color and brightness, and visual depth. These studies demonstrate the following:

1. V2 'sees' different things than V1 sees (see Table 1).
2. There is a balance of power between V1 and V2.
3. There is a dynamic network of V1 and V2 modules that participate in this power struggle.

Functional Organization in V2 Is Modular

V2 has been described as a distribution center from which modality-specific information feeds into the ventral and dorsal visual processing streams. In addition to its role in distribution, V2 plays a crucial preparatory role in midlevel vision, performing transformations that link local detail to higher-order percepts. A great deal of attention has been focused on the interareal interactions and integration between cortical areas. V2 holds a unique position in the

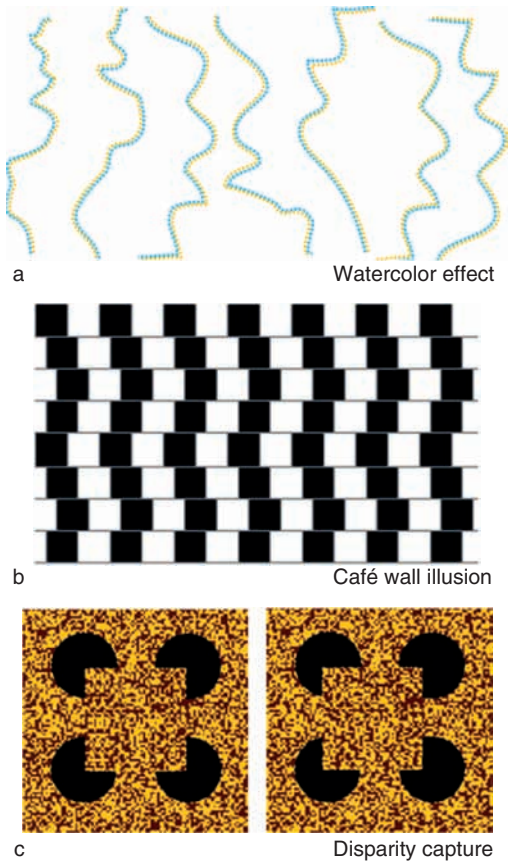

Figure 2 Global versus local competing cues in visual illusions. (a) The watercolor effect; see Pinna B, Brelstaff G, and Spillmann L (2001) Surface color from boundaries: A new 'watercolor' illusion. *Vision Research* 41(20): 2669–2676. (b) The cafe wall illusion. (c) Disparity capture illusion.

Table 1 V1 and V2 see different percepts

	V1	*V2*
Brightness	Luminance	Illusory brightness
Depth	No percept	Percept
Contour	'Real' contours	Illusory contours

sense that it funnels information, both feedforward and feedback, between V1 and the larger cortical network.

To appreciate the role of V1 and V2 in contour, color–brightness, and disparity processing, it is first important to recognize the modularity of these areas. In contrast to V1, which is characterized by ocular dominance structure, a regular array of orientation domains, and a lattice of densely staining cytochrome oxidase blobs, V2 contains a pattern of large (millimeter-size) cytochrome oxidase stripes. Some evidence indicates these thin (cytochrome dark), pale (cytochrome light), and thick (cytochrome dark) stripes are involved in contour, color–brightness, and disparity processing, respectively. In macaques, these stripes are oriented roughly perpendicular to the V1–V2 border and are commonly seen in a repeating thin, pale, thick, pale sequence. There are also reports of interanimal variability and/or variability of stripe thickness (e.g., thin is not thin) or stripe sequence (e.g., thin, pale, thin, pale). A few studies have hinted that functional differences may accompany these differences in individual architecture.

Within these larger stripe structures are collections of 100–200 μm domains. The presence of substructure within stripes was perhaps first suggested by nonuniform staining in cytochrome oxidase-stained tissue and by patchy labeling of injected tracers in V1, V2, and V4. Studies with 2-deoxyglucose have revealed further patchy activation within V2 stripes. Optical imaging further confirmed the presence of subdomains (e.g., color-preferring and luminance-preferring domains within thin stripes). Although both anatomical and functional evidence support the presence of modules, the relationship of functional subdomains and anatomical patchiness is unknown. Functional evidence indicates that different modules play distinct functional roles and have specific connectivities. Indeed, this is visually apparent from localized injection of tracer into V2, which reveals arrays of clusters about 100–200 μm within V2 (Figure 3). These anatomical data suggest a degree of intra-V2 connectional specificity that parallels that of functional networks observed in V1.

In addition to summarizing more recent findings regarding V2, this article presents a common framework for understanding the different featural modalities in V2 and thereby the common functional transformations that V2 performs on its input, regardless of feature space.

Are There Functionally Distinct Domains in V2?

Following the discovery of cytochrome oxidase 'blobs' (also referred to as 'puffs' and 'patches') in macaque monkey visual cortex, researchers reported that blobs are characterized by a predominance of color-selective neurons which lacked strong orientation selectivity, whereas interblobs are characterized by a preponderance of orientation-selective cells lacking in color selectivity. Functional distinctions within V2 have also been reported whereby thin stripes are color selective and poor in orientation selectivity and pale and thick stripes are orientation selective and less selective for color. This proposed dichotomy led to a number of studies, some of which supported whereas others contradicted these findings. This controversy has continued without apparent resolution.

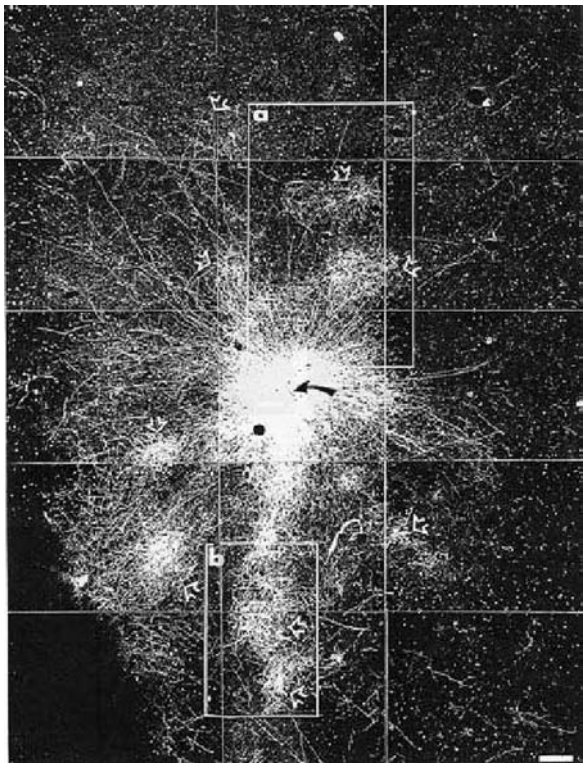

Figure 3 Intrinsic connections in V2 reveal 200 μm modules. (a, b) Boxed areas indicate regions of patchy label (indicated by arrowheads). Arrow: site of tracer injection. Scale bar = 100 μm. From Amir Y, Harel M, and Malach R (1993) Cortical hierarchy reflected in the organization of intrinsic connections in macaque monkey visual cortex. *Journal of Comparative Neurology* 334(1): 19–46.

The debate is fueled by two main issues. The first is whether color and orientation information are segregated in V1 (and V2). Some studies have found a clear association between color selectivity and a lack of orientation tuning. This finding is supported by studies which report that color-selective cells prefer low spatial frequencies and exhibit high contrast sensitivity. Results with 2-deoxyglucose are also consistent with these views, indicating blobs in V1 are preferentially activated by color stimuli and low spatial frequency stimuli.

With respect to spatial distribution, electrophysiological studies have also reached different conclusions. Some support the view that cytochrome oxidase blobs are centers of color preference and poor orientation preference. Others have found that centers of blobs exhibit higher contrast gain and lower spatial frequency response, which suggests that blobs do exhibit responses functionally distinct from that of interblobs. However, other studies differ. Some single-unit electrophysiological studies find blob and interblob domains exhibit no significant difference in color and orientation preference and report a range of associations between color selectivity and orientation selectivity within single cells of V1. Experimental factors may contribute to some of this controversy. Although electrode recording sites were carefully reconstructed in these studies, there is some intrinsic error associated with reconstruction of recording sites. Since cytochrome oxidase blobs are small (100–200 μm), a small degree of error could be significant. There are also differences in sampling methods and methods of receptive field characterization (e.g., hand plots with color filters, characterization with sinusoidal isoluminant gratings, mapping with reverse correlation of cone-isolating stimuli).

A related but distinct issue regards the association of color and orientation with blobs and interblobs, respectively. Studies which reveal population responses (2-deoxyglucose and optical imaging methods) tend to support the role of blobs in the processing of low spatial frequency and color information. As shown by 2-deoxyglucose methodology, blobs are preferentially activated by color-varying gratings of low and middle spatial frequencies and by spatially diffuse color stimuli. Optical imaging studies have shown good alignment between cytochrome oxidase blobs and centers of monocularity and between cytochrome oxidase blobs and regions of poor orientation selectivity. Less clear correspondence between blobs and color response has also been reported. Although one study reported a general correspondence between color domains and cytochrome oxidase blobs, it was not a simple one-to-one relationship: Some color-activated domains were irregular in shape and were observed to span two or more blobs.

A similar controversy exists on V2. Some researchers found little difference in the number of color-selective cells across the different cytochrome oxidase stripes in V2. Some single-unit studies have emphasized the multidimensional aspects of single cell function in V2. Others have reported some concentration of color-selective responses in thin stripes of V2, in contrast to some earlier studies.

In hopes of settling this debate, a recent intrinsic signal optical imaging study closely examined the alignment of large field views of color and orientation response in V1 and V2 of the macaque monkey. As shown in **Figure 4(a)**, the V1–V2 border is well delineated by ocular dominance imaging. Optical images revealed the characteristic orientation maps as well as a regular array of color domains in V1 (**Figure 4(b)**). In V2, orientation domains are larger than those in V1 and are localized to the thick and pale stripes (white bars in **Figure 4(b)** and **4(c)**). Regions of strong color activation in V1 (**Figure 4(c)**) are punctate and bloblike in appearance, are aligned along centers of ocular dominance columns, and tend to overlay regions of

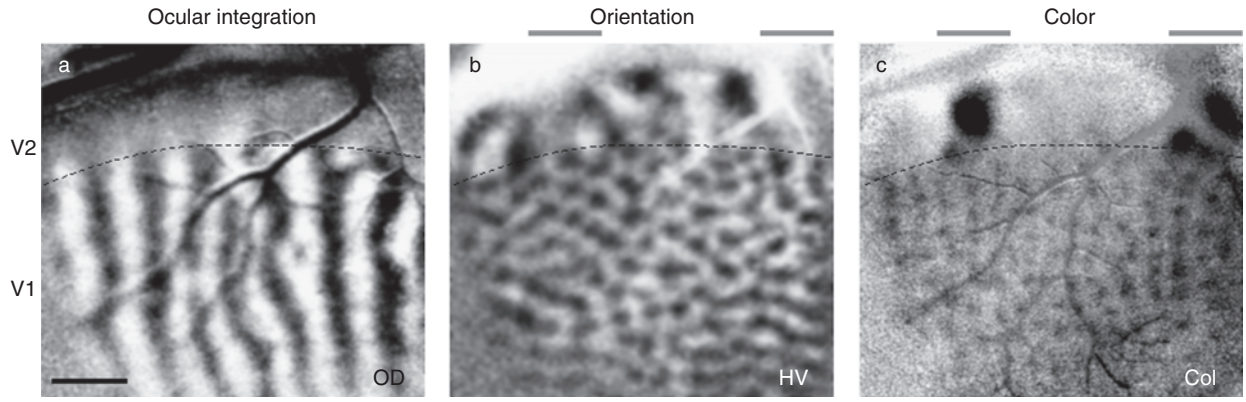

Figure 4 Feature maps (ocularity, orientation, and color) within V1 and V2 of macaque monkey. (a) Ocular dominance (OD) map (left eye minus right eye). V1–V2 border (indicated by dotted line in (a)–(c)). (b) Orientation map (horizontal minus vertical; HV). V2 orientation domains are as much as several times larger than those of V1. (c) Color (Col) map (red–green luminance grating minus luminance grating) reveals V2 thin stripes and V1 blobs. Note that locations of color activations in V2 (dark bars) are complementary to locations of oriented domains in V2 (compare (b) and (c)). Scale bar = 1 mm. From Lu HD and Roe AW (2007) Functional organization of color domains in V1 and V2 of macaque monkey revealed by optical imaging. *Cerebral Cortex* (doi: 10.1093/cercor/bhm081).

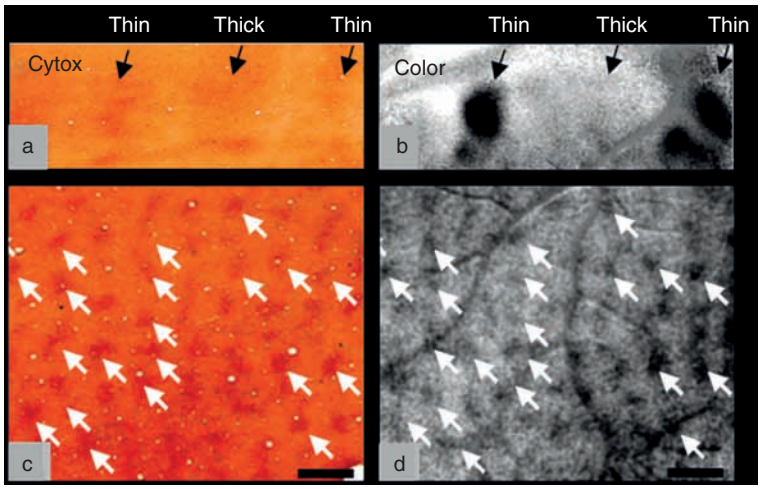

Figure 5 Color domains align with cytochrome architecture. Alignment of cytochrome oxidase (cytox) stripes (a) and color domains (b) in V2, and cytochrome oxidase blobs (c) and color blobs (d) in V1. Tissue was aligned using blood vessels, V1–V2 border, tracer injections, and electrolytic lesions. Black ((a) and (b)) and white ((c) and (d)) arrows point to corresponding structures. Scale bar = 1 mm. From Lu HD and Roe AW (2007) Functional organization of color domains in V1 and V2 of macaque monkey revealed by optical imaging. *Cerebral Cortex* (doi: 10.1093/cercor/bhm081).

low orientation selectivity. In V2, color domains are complementary in location to the orientation-selective areas and overlay the thin cytochrome oxidase stripes. An alignment of cytochrome oxidase stained tissue, and functional images (using blood vessels, tracer injections, and electrolytic lesions as guides) have revealed good alignment of color domains with thin stripes in V2 (Figures 5(a) and 5(b)) and color blobs with cytochrome oxidase blobs in V1 (Figures 5(c) and 5(d)). These findings thus support, at the level of population response, significant functional segregation of color and orientation information in V1 and V2 and further strengthen their association in V1 with blobs and interblobs, respectively.

Although on the face of it, these studies appear at odds, they are, in fact, mutually consistent. Clearly, single cortical locations contain cells with a mixture of responses and cell types. Furthermore, each cell displays an array of responses in multiple feature domains. There is no question that individual neurons are multidimensional and that cells in V2 integrate information across multiple feature domains, a point that is made by a number of electrophysiological studies. This multidimensionality is precisely what leads to different views

of the same cortical area. Different electrophysiological samplings and somewhat different characterization methods can provide quite different impressions, indicating that methodology can strongly influence the apparent population response profile. (Indeed, being able to electrophysiologically target imaged color domains increases the yield of color cells tremendously.) Despite variability due to method, that functional imaging and functional anatomical studies consistently reveal blob and stripe structure suggests that there are overall differences in the local population response. The fact that functional responses reveal stripelike structures within V2 and that distinct responses are recorded in different stripes in V2 suggests the presence of at least some degree of segregation. On average, the cells in blobs and thin stripes prefer color stimuli over achromatic stimuli. Functional imaging provides, in a sense, a more unbiased view of the average population response. Perhaps some of the controversy is due to the perception that functional organization implies strict segregation. However, this is not the case, as each functional structure contains a mixture of neurons with varying selectivities. In sum, although single neurons are multidimensional (have some degree of responsiveness to both contour orientation and color), as a population, they are organized to some degree with respect to contour versus color.

Parallel Functional Channels in V2: A Useful Framework?

A framework that had its roots in the 1980s maintains that thin stripes are vehicles for processing surface feature information whereas thick and pale stripes are involved in contour and depth information processing. This framework has been supported by anatomical, electrophysiological, 2-deoxyglucose, optical imaging, and behavioral and lesion study data. This framework is not to be interpreted as a view of strict segregation. There is ample data to demonstrate that functional organization within V2 is not one of strict segregation. As noted above, single neurons in V2 are multidimensional and integrate different featural spaces. Each stripe type within V2 contains a myriad of cell types. Given the extensive connectivity between stripes in V2, cells from each stripe type have opportunity to influence and integrate information from other stripe types. Nevertheless, the proposal that different stripes subserve somewhat different functions is still a useful one. It is a framework which provides an understanding which can be experimentally tested and modified. To ignore this framework would be foolish. In the language of 'drivers and modulators,' there may be primary functions that can be modified by secondary, perhaps contextual, influences. A working framework provides much greater utility and testability than an approach in which every neuron is omnifunctional. Indeed, the ubiquity of functional organization in the brain – whether it be at the level of global networks, cortical areas, modules and columns, laminae, or synapses – indicates that such organization is a fundamental tenet of brain organization. Whether they be genetically determined, experientially determined, or a necessary prerequisite for specific computations, it is likely that functional organizations bear a strong (and possibly predictive) link to the brain circuits that produce function and behavior. The following paragraphs are thus presented in this framework.

Surface Properties

Surface features, such as color, brightness, texture, glossiness, and transparency, are those that help us identify the material quality of an object. There are ample demonstrations of the ways the perception of these surface features can be strongly influenced by lighting, local color and contour cues, global context, and experience. The human visual system can distinguish on the order of a million different colors. However, the inputs to the visual system are confined to three types of cones, those responsive to red, green, and blue wavelengths. How does the visual system create the illusion of color and transform a triadic cone-based representation into a perceptual continuum of salient hues? Add to this challenge the fact that the human visual system can adjust to 7 orders of magnitude of brightness (10 million to one, bright sunlight to faint starlight). The visual system must also factor in the influence of contextual cues such as color, brightness, and contrast of nearby borders and surfaces. In addition, the ability to see objects through transparent surfaces means we not only must segregate different planes of depth but also ignore features on the transparent plane.

Color Representation in V2

How do V1 and V2 separately and together contribute to these processes? The idea that V1 and V2 participate in encoding color information is well supported. At least one route for color information arises from the parvocellular layers of the lateral geniculate nucleus (LGN) and continues to layer 4Cb of V1, to cytochrome oxidase blobs of V1, and to the thin stripes in V2. Blue cone-driven input from S layers of the LGN project directly to the blobs, with blue–yellow inputs targeting layer 3B–4A and red–green inputs targeting lower layer 4C. There is also evidence to suggest a functional segregation of blue

'ON' and blue 'OFF' inputs within layer 4A of V1. In V2, although there is significant color response in each of the stripe types, a greater concentration of neurons responsive to color stimuli is found in the thin stripes than in either the pale or thick stripes.

Color versus luminance preference domains Optical imaging studies have revealed the presence of color-responsive domains within thin stripes. Some studies have used preferential response to isoluminant red–green gratings over achromatic luminance gratings to reveal color preference domains. Although there are both color and luminance contrast differences between these stimuli, experiments directly comparing high- versus low-contrast achromatic responses have revealed no structured maps in V2, indicating that the imaged substructure is not due to luminance contrast differences. Since spatial frequency, drift rate, and mean luminance were identical between these two stimuli, the differential response is attributed to the color content of the stimuli. Such stimuli elicit activations that appear more rounded and interdigitated in some thin stripes (Figures 6(a) and 6(c)) and more elongated across the width of the thin stripe in others (Figure 6(b)). It is not yet known whether there may be functional implications of these different architectures.

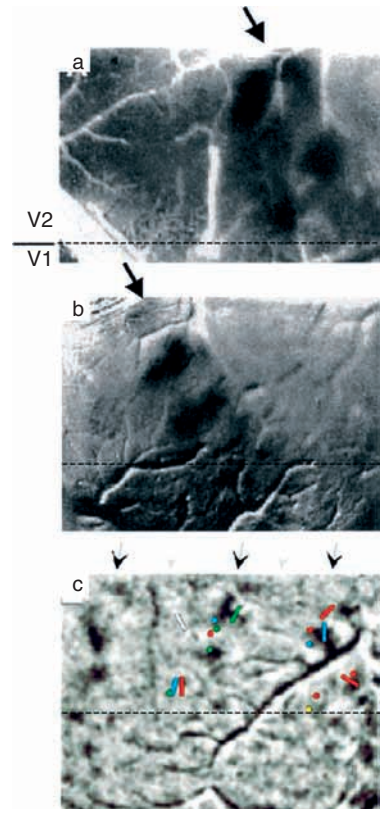

Figure 6 Color domains in thin stripes. (a) A single thin stripe (arrow) contains several rounded color domains. (b) A single thin stripe (arrow) contains color-preferring domains that span the width of the stripe. (c) Three thin stripes in V2, each containing a few color domains. Scale bar = 1 mm. (a) From Roe AW and Ts'o DY (1995) Visual topography in primate V2: Multiple representation across functional stripes. *Journal of Neuroscience* 15: 3689–3715. (b) From Roe AW and Ts'o DY (1999) Specificity of color connectivity between primate V1 and V2. *Journal of Physiology* 82: 2719–2730.

Color: Topography of hue in V2 thin stripes One of the first studies to establish a systematic representation of hues in V2 used isoluminant color–gray gratings of different hues and mapped responses in macaque V2 using intrinsic signal optical imaging. These images revealed hue-specific domains on the order of 100–200 μm, which were mapped in a systematic fashion in V2. These chromatic color maps were located within single thin stripes and not in thick or pale stripes of V2 (Figure 7(a)). This was the first demonstration of a hue-based map in which topographic location corresponded to hue wavelength. This study underscored three primary points: (1) although V1 contains color-selective cells tuned to red–green and blue–yellow axes, hue representation does not emerge until V2; (2) the topographic basis for this representation is based on 100–200 μm modules, the size of which is fundamental to other featural representations; and (3) the discovery of topographic maps within thin stripes poses perhaps the single strongest piece of evidence in favor of the color–thin stripe association.

Where is the blue response? Attempts at revealing blue response in V1 and V2 with imaging techniques proved unsuccessful. Similarly, comparing responses to isoluminant blue–yellow gratings versus achromatic gratings did not reveal any functional organization within either V1 or V2. However, more-recent studies have revealed clear and strong responses to blue stimuli. These studies examined responses to flashing blue versus flashing gray unstructured stimuli each at 30% luminance and drifting blue–gray square-wave gratings versus gray–gray square-wave grating at 30% luminance contrast. In V2, similar to red and green, blue stimuli preferentially activate the thin stripes whereas gray stimuli exhibited preferential thick-stripe activation. Within the thin stripes, color-specific stimuli typically revealed focal activations. Activations to red, green, and blue stimuli have been shown to shift in location within the thin stripes. Examination of responses to gratings of different duty cycles revealed that thin blue lines were a surprisingly effective stimulus. Thin blue lines elicited strong blob-like maps in V1 (Figure 7(b)). These data suggest that blue activations, like red and green hues, are mapped

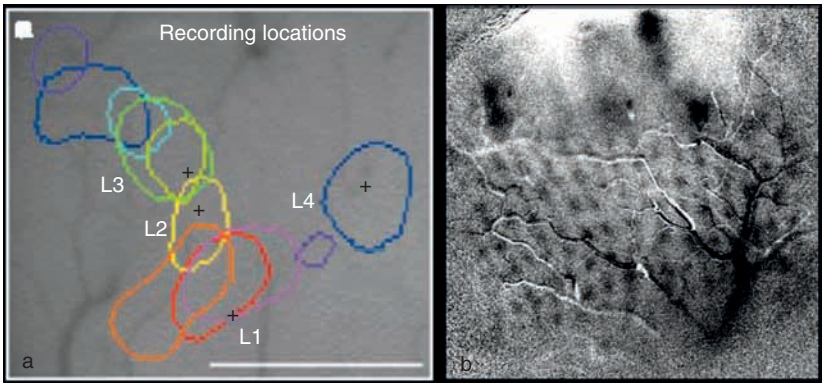

Figure 7 Hue maps in V2. (a) Chromatic color maps located within a single thin stripe of V2 (imaged with isoluminant color/gray gratings of different hues). Outline color indicates peak response of that location. (b) Imaged responses Macaque V1 and V2 in response to thin blue oriented stimuli. In V1, blue activation reveals blobs. In V2, blue activation is mainly located in thin stripes. Scale bar = 500 μm. (a) From Xiao Y, Wang Y, and Felleman DJ (2003) A spatially organized representation of colour in macaque cortical area V2. *Nature* 421: 535–539. (b) From Roe AW, Lu HD, and Crewther D (2007) Blue Color Activation in V1 and V2 of Macaque Monkey. Melbourne: International Brain Research Organisation.

within 'color modules' in the thin stripes of V2. The surprisingly strong response to thin blue lines may suggest some interaction of color and spatial frequency that may contribute to contour response in V2. Data also suggest, counter to previous notions of dedicated red–green versus blue–yellow blobs in V1, that blue response may reside within every blob in V1.

A midlevel role in simultaneous contrast and color constancy? The color of a surface can be dramatically changed by its surround, such as commonly demonstrated with examples of simultaneous color contrast. A first step in achieving simultaneous color contrast response would be demonstrating a chromatically specific effect by the surround. Surround suppression of both achromatic and isoluminant gratings have been examined on the chromatic responses of V1 and V2 neurons. There is little evidence that colored surrounds alter the chromatic tuning of neurons in V1 or V2 (although background color shifts chromatic tuning of V1 neurons to a degree consistent with perceptual color contrast effects). However, V2 neurons, unlike V1 neurons whose surrounds exhibit little chromatic selectivity, have a chromatic signature in the surround similar to that in the receptive field center. Thus, although there is no evidence of the surround shifting chromatic tuning in V2, neurons in V2 are strongly affected by specific chromatic stimuli in the surround.

Another salient aspect of color perception is 'color constancy.' This term refers to the fact that the color of a surface remains constant under different lighting conditions (e.g., a red apple will appear red under different spectral lighting conditions, such as bright sunlight and ambient indoor lighting). This suggests that there must be neurons somewhere in the brain that encode 'redness' or 'greenness' of a surface regardless of light condition. Do such neurons exist in V2? A few recent studies have tried to dissect the difference between color responses in V2 and those in V1 or V4. Researchers have used isoluminant color patches to characterize the color preference of single cells in V2 thin stripes and Mondrian stimuli to examine the effects of color versus the illuminant. They found that V2 cells responded to the wavelength of the illuminant rather than the color of the stimulus. In contrast, the same stimulus paradigms revealed clear color constancy in a majority of V4 neurons: these neurons exhibited shifts in their color-tuning profiles in the direction of the shifted chromatic component of the illuminant (which was simulated by a shift in the background color presented on the computer monitor). In other words, V2 neurons did not exhibit behavior consistent with color constancy. However, about a third of V2 neurons studied were strongly influenced by the wavelength of the surround.

Thus, these studies suggest that V2 plays a midlevel role in the generation of color percepts such as simultaneous color contrast and color constancy. Although neurons in V1 and even in LGN are reported to be influenced by monochromatic surrounds, shifts of the entire color tuning curve have been reported only for V4 neurons. This is true for units recorded both in anesthetized and awake monkeys. Thus, computations leading to color constancy are likely to begin at stages before V4, but color constancy is not represented in its full extent until V4. V2 is likely an intermediate processing stage in these computations.

Brightness Representation in V2

With respect to V2's role, if the primary role of thin stripes is to encode surface properties, then in addition to color, thin stripes may be integral to representing surface brightness. The challenges of representing brightness in the visual system include representing light level (lightness and darkness), change in light level (either increase or decrease), and ability to adjust to a large range of light levels. Some of these challenges may be solved by the retina and LGN. However, other challenges are unlikely to be resolved by early precortical levels. Similar to context-dependent effects in color vision, the perceived brightness of a surface can be strongly influenced by surrounding cues. Brightness perception is influenced by multiple factors, including luminance, edge effects, distant color and brightness context (e.g., simultaneous contrast, Mondrians), and experience. Neurons in V1 of both the cat and the monkey exhibit response modulation to large field modulations in luminance and in background luminance modulation. Relatively few studies have examined the representation of brightness information in V1 and V2.

ON and OFF responses Many studies have demonstrated that visual cortical neurons exhibit contrast response functions which increase monotonically over a range of contrasts and saturate at higher contrasts. Most neurons in V1 respond to a large patch of luminance increment or decrement in a monotonic fashion. Roughly one-third of neurons are not affected by stimulation in the surround, consistent with the type I cells which lack inhibitory surrounds. The response is often characterized by an initial contrast-independent transient, followed by a sustained response which correlates with the luminance increment or decrement. Another third of neurons are enhanced by luminance in the surround and appear to signal level of illumination. And the remaining third are affected by luminance in the surround (type II) in a manner consistent with simultaneous contrast as first demonstrated in anesthetized cat visual cortex and later also observed in the alert monkey V1.

In addition to light increment (ON) and light decrement (OFF) responses, V1 and V2 contain cells that exhibit bandpass characteristics. That is, they respond to an intermediate range of luminance values (termed 'V-shaped' or 'gray-preferring neurons'). Some of these neurons also exhibit color sensitivity, suggesting a role for luminance response in color encoding.

ON and OFF domains in V2 Within V2 thin stripes, there are subdomains that exhibit luminance increment (ON) and luminance decrement (OFF) responses. These domains measure on the order of 0.5 mm and could be the same 'luminance-preferring' domains revealed in isoluminant color versus achromatic imaging experiments. Such domains had been hypothesized based on extended runs of tangentially recorded of ON or OFF single unit responses within V2 thin stripes. Such ON and OFF zones have not been observed in primate V1.

Contrast response in V2: Thin stripes have greater dynamic range One test of the hypothesis that thin stripes are brightness-processing structures has taken the form of examining contrast response in V2. The expectation was that if thin stripes were loci of brightness information processing, then they should have contrast response behavior consistent with such a representation. One would expect the contrast response of these stripes to exhibit sensitivity to a large range of contrasts, including sensitivity to low contrasts and nonsaturating response at high contrasts. More specifically, they should have high contrast gain, especially at low spatial frequencies consistent with encoding visual surface features, and they should exhibit large dynamic range for encoding visual brightness.

Previous studies had shown that in primates, magnocellular and parvocellular neurons in the LGN are characterized by distinct contrast response signatures. Magnocellular neurons exhibit nonlinear response characterized by greater sensitivity to small changes in contrast (have higher contrast sensitivity), especially at low luminance levels, and tend to saturate at high contrasts. Parvocellular neurons respond in a linear fashion to increasing contrast and are relatively insensitive at low contrasts (in fact, they are not responsive at contrasts lower than around 10%). Koniocellular neurons are a heterogeneous population of neurons and have a broad range of contrast response functions. In V1, interblobs tend to be dominated by parvocellular inputs, while blobs exhibit parvocellular (color selectivity), magnocellular (low spatial frequency, high contrast gain response), and koniocellular (blue response) characteristics. Although thick stripes are dominated by magnocellular inputs and therefore could exhibit higher contrast sensitivity than thin stripes do, in principle both magnocellular and parvocellular inputs reach all the stripe compartments in V2. The thin stripes receive both magnocellular and parvocellular input primarily by way of the blobs, and thick and pale stripes via the interblob columns. Thick stripes have the highest contrast sensitivity values, although semisaturation values have shown no significant difference between thin, pale, and thick stripes. Measures of V2 response have revealed surprisingly

few quantitative differences between thin, pale, and thick stripes, leading some to argue for homogeneity of stripe function in V2. With 2-deoxyglucose methodology, both thin and thick, but not pale, stripes have exhibited response to low contrast (8%) gratings, suggesting that magnocellular contribution could also be prevalent in the thin stripes.

A study using optical imaging methods has demonstrated that V2 thin stripes have greater contrast gain than thick and pale stripes and relatively nonsaturating response at high contrasts. As shown in Figure 8, contrast response functions of both blobs and interblobs in V1 were linear (Figure 8(a)), consistent with parvocellular-dominated response in superficial layers of V1. However, responses in V2 exhibited a nonlinear response, especially at low contrasts, reminiscent of magnocellular input (Figure 8(b)). At high contrasts, thin stripes exhibited significantly stronger and less-saturating responses than thick or pale stripes (Figures 8(c) and 8(d)). The response of V2 thin stripes is thus reminiscent of magnocellular response at low contrasts and parvocellular response at high contrasts. At low contrast levels, magnocellular inputs from V1 blobs may converge in V2 thin stripes, thereby conferring high-contrast response at low contrasts. At higher contrast levels, parvocellular-dominated inputs to thin stripes continue to signal changing contrast, while those in thick and pale stripes tend to saturate. Other contributions to nonlinear response in V2 thin stripes (such as koniocellular input and other thalamic inputs) may also be at play. In comparison with thick or pale stripes, these data suggest a larger dynamic range of V2 thin stripes, consistent with a preferential role in brightness perception.

Real and illusory brightness representation in V2 thin stripes Perhaps intermediate to local luminance response and lightness and color constancy which require rather large-scale integration are midlevel stages which are encoded in V2 (in cats). A potential example of midlevel integration was demonstrated in a recent study revealing the equivalence of V2 response to different types of brightness stimuli. Response to sinusoidally modulated stimuli counterphasing in

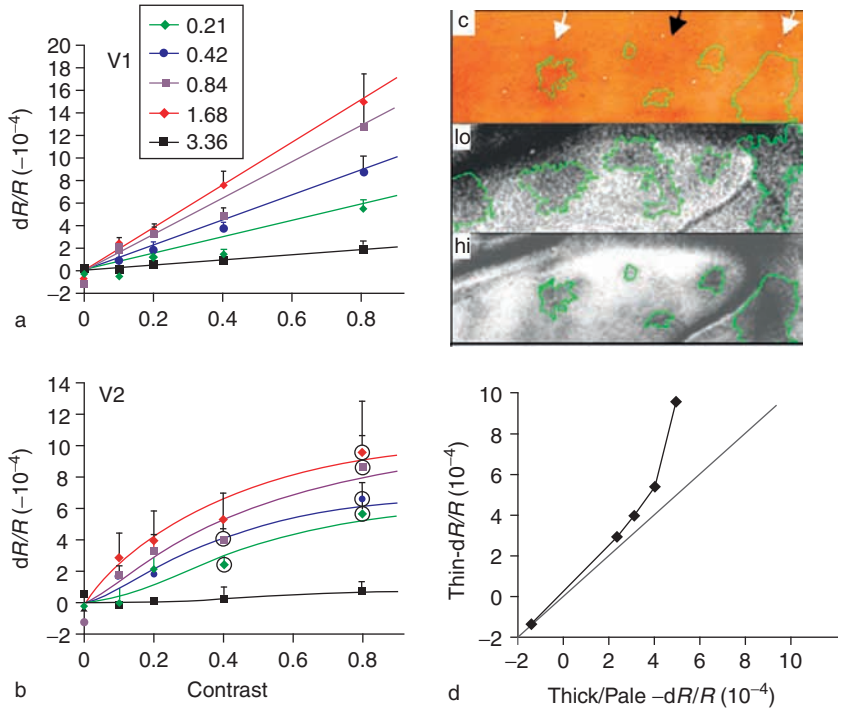

Figure 8 Contrast response in V2 stripes. (a, b) Contrast response functions in V1 are linear, both in blobs and interblobs, while those in V2 are nonlinear, in thin and thick/pale stripes. Different color lines represent response obtained at different spatial frequencies (same legend in (a) and (b), in cyc/deg). *x*-axis, visual contrast (0.2 = 20%); *y*-axis, d*R*/*R* – reflectance change. (c, d) Thin stripes have greater dynamic range. (c) Top: Cytochrome oxidase stain reveals thin (white arrows) and thick (black arrow) stripe in V2. Middle and bottom: Imaged response to low (lo, 40% contrast) and high (hi, 80% contrast) sinusoidal gratings. At high contrast, thin stripes still exhibit robust response, but thick stripe response is weak. Green outlines indicates activation areas, thresholded and outlined. (d) Thin and thick stripe responses are similar at lower contrasts, but thin stripe response dominates at high contrast. Thin (*y*-axis) and thick/pale (*x*-axis) stripe responses are similar at low contrasts (black, dark gray symbols), but thin stripe response dominates at high contrast light gray, white symbols, (d*R*/*R* – reflectance change). From Lu HD and Roe AW (2007) Optical imaging of contrast response in macaque monkey V1 & V2. *Cerebral Cortex* (doi: 10.1093/cercor/bhl177).

contrast was preferentially localized to the thin stripes of V2 (Figures 9(a) and 9(b)). In fact, similar thin stripe activations were obtained in response to 'illusory' brightness modulations as early as 1970 (Figure 9(c)), when a percept of brightness contrast between the left and right halves was achieved by an intervening 'Cornsweet' border (Cornsweet being one of the researchers involved). This border-induced contrast illusion evokes the percept of brightness differential where none is present. Neurons in V2 were shown to respond not only to sinusoidal modulation of the 'real' brightness stimulus but also to that of the Cornsweet border contrast. Optical imaging of V2 response indicated that such responsiveness was localized to the thin stripes of V2. Consistent with optical imaging evidence, of 89 neurons recorded in V1 blobs and interblobs and V2 thin, pale, and thick stripes, almost all those responsive to the Cornsweet were located in thin stripes. These results suggest that thin stripes are involved in not only the encoding of luminance response but, more generally, the response related to surface brightness perception.

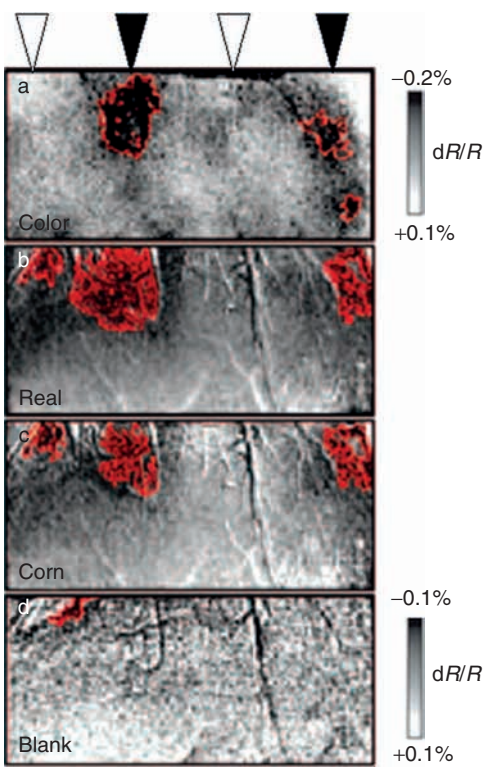

Figure 9 Optical imaging of brightness response. (a) Single condition color map reveals thin stripes in V2 (black arrowheads above, thick/pale stripes indicated by white arrowheads). Pixels with strongest activation (top 10%) indicated by red outlines. Grayscale in (a) also applies to (b) and (c). (b) Single condition activation map in response to real luminance stimulation. Strongest activation is seen in thin stripes with weaker activation in thick/pale region. (c) Single condition activation map in response to Cornsweet stimulation reveals similar pattern. (d) Blank stimulus map. Scale bar = 1 mm. Each image sum of 50 trials. In all single condition maps, darker pixels indicate larger magnitude reflectance change. d*R*/*R* – reflectance change.

V1 Sees Local, V2 Sees Global

These data are consistent with the encoding of higher-order brightness processing in V2 and lower-order brightness processing in V1. That is, responses in V1 reflect local luminance values, whereas those in V2 reflect more global, edge-induced percepts. In the Cornsweet stimulus, V1 would report a lack of luminance modulation, whereas V2 would report a salient brightness modulation. To achieve such perceptual saliency, the V2's message must somehow overcome V1's. One obvious means is via feedback projections from V2 to V1 which lead to either relative suppression of V1 response or a modulation of V1 response that concurs with that in V2. This possibility remains to be examined.

Stereoscopic Depth

Binocular inputs are used by the human visual system to judge object depth in the three-dimensional world. This depth percept is created by the integration of two views of the world received by the two eyes. Stimuli nearer or farther from the fixation point will produce disparities from left and right eye with a negative or positive horizontal shift. In the pathway from retina to thalamus to cortex, ocular input remains segregated in the thalamus (the LGN) and is first combined in V1. In V1, disparity-selective cells have been characterized by their response to offset bars presented to the two eyes ('tuned excitatory,' 'tuned inhibitory,' 'near,' and 'far' cells), response to differential phase of sinusoidal gratings, and response to absolute disparity of RDSs. In V2, disparity-selective neurons have been described as 'obligatory binocular,' selective for disparity-defined contours, and tuned for relative rather than absolute disparity. Despite the number of studies on disparity responses in the visual cortex, there have been no published studies addressing any systematic representation of disparity response in either V1 or V2 of the monkey.

Given the known organizations of the other two featural maps in V2 (color in the thin and contour in the pale and thick stripes of V2), this gap in our knowledge regarding disparity representation in V2 is particularly conspicuous. Disparity-selective responses are believed to be preferentially localized to the cytochrome oxidase thick stripes of V2, but topographic representation within these stripes has not been examined. Thus, at issue are the questions of whether there is

a map for near-to-far depth information and whether disparity and orientation information, which are both represented in the thick stripes, are independently represented or not. The broader issue of parallelism and modularity across featural domains in V2 is also in question. Examination of these issues is further motivated by the presence of maps organized for disparity in the middle temporal area (MT), a primary target of V2 thick stripes.

Topography for Near–Far Disparity

Since it is known that many disparity-selective neurons in V2 include obligatory binocular cells, previous optical imaging studies had indicated localized thick stripes by locating domains preferring activation of two eyes over a single eye. These binocular-preferring domains were located primarily, although not exclusively, within thick stripes (Figure 10). Two recent studies have indicated the presence of topographic organization in V1 and V2. In one, two photon calcium imaging was used to study organization of cortical responses to phase offset gratings in the cat. This study revealed beautiful pinwheels of disparity response in cat V1. In another study, RDSs were used to probe disparity organization in primate V1 and V2. These stimuli were designed to produce the percept of surfaces at different depths relative to the background, ranging from -0.5° near to $+0.5^\circ$ far. Characterization of single units in V2 revealed that cells in V2 exhibited tuning for specific horizontal disparities (Figure 11(a), RDS). It is important that

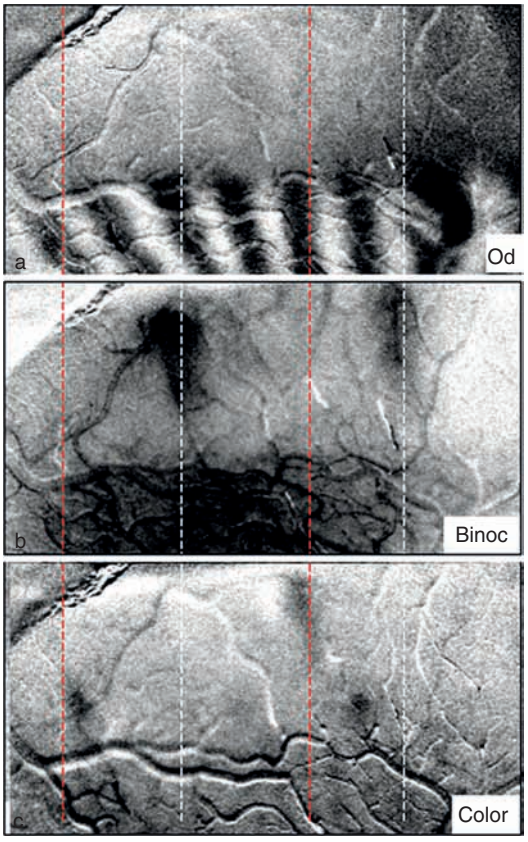

Figure 10 Imaging V2 thick stripes. (a) Ocular dominance (od) map. (b) Color map reveals thin stripes (30% contrast blue/gray minus 30% contrast gray/gray grating). (c) Binocular (binoc) versus monocular preference map reveals thick stripes. Thick stripes (white dotted lines) and thin stripes (red dotted lines) are in complementary locations in V2.

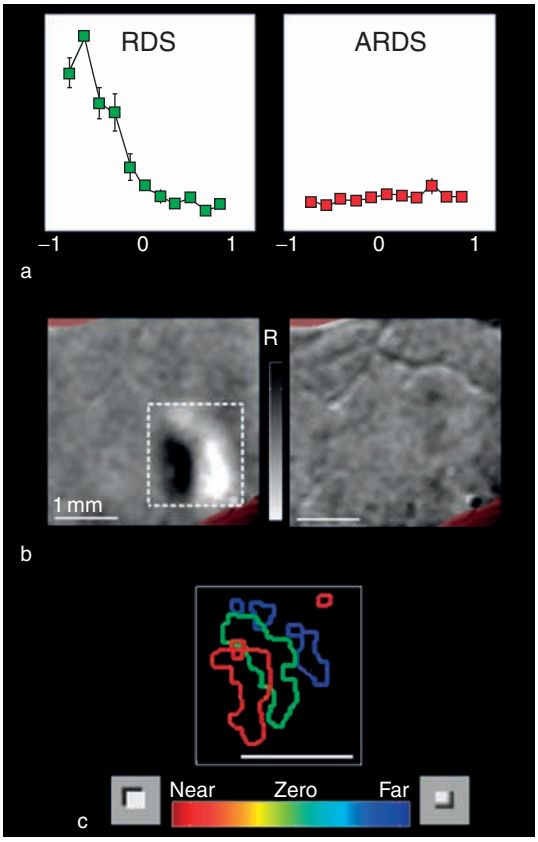

Figure 11 Topography for disparity in V2 thick stripes. (a) Disparity tuning curves of a recorded unit from imaged disparity domain in V2 thick stripe. Random dot stereograms (RDSs) were shifted in disparity from -0.42° to $+0.42^\circ$. This cell is tuned to near 0.4°. Its response to antirandom dot stereograms (ARDSs) shows a flat curve, indicating that the cell's response correlates with the global depth percept and not simply with local disparity cues. (b) Optical images to RDS and ARDS stimuli. Left: RDS, differential image of all near (dark pixels) minus all far (light pixels) stimuli. Dotted box, region of activation in V2 thick stripe. Right: ARDS, differential image of all near minus all far stimuli. (c) Outlines of three disparity domains (0.34° near, 0.34° far, and zero). Location of activation shifts as disparity changes from near disparity (red) to zero disparity (green) to far disparity (blue). Scale bar = 1 mm (b, c). *R,* reflectance value. From Chen G, Lu HD, and Roe AW (2006) Functional architecture of macaque cortical area V2 for depth surfaces revealed by optical imaging. *Society for Neuroscience* 32: 114–116.

these cells showed little responsiveness to stereograms that do not induce depth percepts; these include anti-random dot stereograms (ARDSs, stereograms in which the contrast of corresponding pixels in the left and right eyes was reversed and uncorrelated stereograms) and stereograms in which the left and right eye dots have a random relationship to each other (Figure 11(a), ARDS). Similarly, optical imaging of such responses revealed strong responses to depth percept-inducing RDSs and little response to ARDSs (Figure 11(b)). Key to the success of these studies was a rapid spot imaging method that permitted quick confirmation of individual eye position and mapping of the imaged field of view. Consistent with previous studies demonstrating preferential disparity response within V2 thick stripes, these activations were located within thick stripes. Furthermore, analogous to a topographic hue representation within thin stripes, a near-to-far topography within the thick stripes was observed (Figure 11(c)). Imaged disparity responses correlated well with disparity preferences of electrophysiologically characterized single unit responses. Furthermore, disparity preferences of units recorded within single vertical penetrations were similar in disparity selectivities, suggesting the presence of a columnar organization for disparity within V2 thick stripes. No such topographies were observed in V1.

Orthogonality of Disparity and Orientation

It is well established that thick stripes in V2 contain maps for orientation. How, then, do disparity maps relate to orientation in V2? Indeed, how multiple maps coexist within the same two-dimensional cortical space is a fundamental issue that has been addressed by a number of studies. In V1, modeling studies suggest that competing intracortical influences result in observed functional organizations for ocular dominance and orientation. In V2, there is an interdigitation of color, contour, and disparity maps in single thin, pale, thick, pale stripe cycles. However, within single thick stripes, the relative organization of orientation and disparity in V2 has been elusive. Is there a full range of orientations represented at each disparity, and is there a full range of disparities represented at each orientation? Both psychophysical and electrophysiological evidence suggest that representation of horizontal disparity might be biased for vertical orientations. Some researchers have found that disparity-activated regions contained roughly equal representation for different orientations and that on average, each orientation domain overlay a range of disparity domains. This suggests strongly an orthogonality between orientation and disparity within V2 thick stripes.

Absolute versus Relative Disparity

Early studies described the tunings of V1 neurons in terms of absolute disparity. However, depth perception is clearly dominated by relative disparity. That is, if the background disparity moves forward, a relative disparity-tuned cell will shift its tuning forward; if the background moves backward, a relative disparity-tuned cell will shift its tuning backward. The response of an absolute disparity cell would remain unchanged with changing background disparities. Using RDSs and a disparity clamp that permitted stepping absolute disparities without changing any relative disparities within the stimulus, researchers have demonstrated that V1 neurons are tuned for absolute disparity. Of 253 neurons studied, roughly 20% were disparity-tuned. Of these, about 80% exhibited absolute disparity tuning; only two exhibited some (weak) shift consistent with relative disparity tuning. In contrast, others have demonstrated that in V2, many more neurons have responses shifted in the direction of the background shift. Clearly, the influence of surround disparity on center disparity response is greater in V2 than in V1. However, only a few V2 cells exhibit shifts indicative of true relative disparity (i.e., shifts equal to background shift). The distribution of these absolute and relative disparity responses in V2 is unknown. However, given the small number of true relative disparity responses in V2, such organization may not become evident until later processing stages (e.g., V4, MT). Although true relative disparity maps are not expected in V2, V2 activations are expected to be influenced by shifts in background.

As shown in Figure 12(a), a V2 disparity tuning curve shifts toward far if background shifts far (solid line) and shifts near if background shifts near (dotted line). Thus, at a fixed center patch disparity (vertical red line), the neuronal response should decrease for near shifts and increase for far shifts of the background (Figure 12(b)). Findings from one preliminary study have suggested this prediction is met (Figure 12(c)). Optical imaging of a thick stripe in V2 with a fixed center patch disparity (near +4) and different background disparities resulted in predicted decreasing activation values from far to near (imaged signal in three different disparity domains indicated by red, green, blue lines in Figure 12(c)). Thus, responses in V2 thick stripes exhibit some influence by shifts in background in a manner consistent with a degree of relative disparity tuning.

Summary

Together, these data suggest that topographic organization for near and far disparities may be established relatively early in the visual pathway, in V1 in the cat

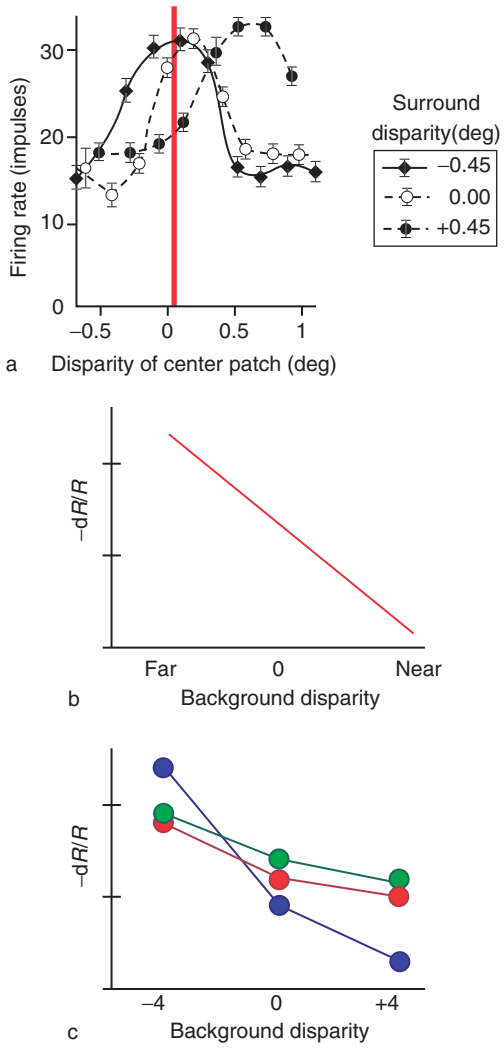

Figure 12 Relative disparity in V2. (a) Disparity tuning curves of a relative disparity cell in V2. Tuning curves shift in direction of background shift. (b) For a fixed surface depth, shifting background from far to near should result in decrease of neuronal firing. (c) Reflectance amplitudes from three disparity domains (red, green, and blue) across shifts in background disparities from $-0.34°$ to $+0.34°$ far are consistent with predicted relative disparity response. Deg, degree; $-dR/R$, change in reflectance. (a) From Thomas OM, Cumming BG, and Parker AJ (2002) A specialization for relative disparity in V2. *Nature Neuroscience* 5(5): 472–478.

and in V2 in the macaque monkey. These data also suggest that the parameters of disparity and orientation in the thick stripes may be orthogonally represented. That is, within each orientation domain is a range of disparities, and within each disparity domain is a range of orientation representation. Further computational modeling may suggest how orientation and disparity maps are mutually arranged within V2. The overlap of orientation and disparity also suggests that orientation-selective domains within thick stripes are distinct from those in pale stripes. It is important that these data also raise the possibility that at least some aspects of the disparity organizations observed in MT derive from V2 and are not established *de novo* in MT.

V1 Sees Local, V2 Sees Global

The finding that V2 neurons are more strongly influenced by shifts in background disparity suggests that V2 encodes not simply the absolute disparity of a surface but, at least to some extent, the relative disparity. Thus, similar to the proposed view of V2 in color and brightness representation, these data are consistent with the encoding of what can be considered higher-order disparity processing in V2 and lower-order disparity processing in V1. Relative disparity signals may be important for figure–ground segregation, which suggests that V2 may be involved in coarse stereo vision.

Another example of V2's role in global perception is the presence of neurons in V2 which achieve disparity capture. The elements within the subjective figure are perceptually 'captured' and 'pulled' on the same depth plane as the figure. The phenomenon has been interpreted as the result of spreading of disparity signals from the subjective figure. In this sense, we consider disparity capture an illusory percept, one which is induced not by true disparity of elements on the surface but by disparity of distant features. In fact, some neurons in V2 have been reported to shift their disparity tuning curves in a manner consistent with disparity capture. If V2 encodes surface depth, then it is conceivable that such disparity capture responses may map to locations similar to those of the surface disparity responses. Again, to achieve such perceptual saliency, the V2's message (the entire surface is at the depth of the border) must somehow overcome V1's (the actual disparity of elements on the surface). This could occur via feedback projections from V2 to V1, resulting in relative suppression of V1. This possibility also remains to be examined.

Visual Contours

There are a myriad of contours in our visual world, some of which are defined by luminance contrast and many more which are defined by cues without luminance contrast. In addition to 'real' luminance-defined contours (sinusoidal grating or line grating), different texture patterns or abutting line stimuli can create strong percepts of borders or contours. These percepts can be perceived regardless of the orientation of the real line inducers (either with acute or obtuse inducers). Borders between motion fields can also produce very salient percepts of contours (motion contrast contours).

Cue Invariance

V2 neurons exhibit responses to so-called cognitive contours (also termed 'illusory contours'). Electrophysiological recordings of single units in V2 have demonstrated a population of cells that share tuning for orientations of both real and illusory contours (both abutting line and occluded contours). Optical imaging studies have demonstrated that orientation domains imaged in response to real contours (real lines or gratings) coalign with those imaged in response to illusory contours, suggesting the presence of higher-order orientation domains. Recent optical imaging studies in alert monkeys have also extended the invariance of response in V2 to include not only real (Figure 13(a)) and illusory contours (Figure 13(b)), but also motion contrast contours (Figure 13(c)). In V1, these studies reveal either a very weak (awake monkey) or an orientation-reversed (anesthetized monkey) response to illusory contours. This may be attributed to the difference in relative activation magnitudes in the alert versus the anesthetized states: V2 activation in the anesthetized monkey is typically weaker than that in V1 whereas it is at least as strong in the alert monkey.

V1 Sees Local, V2 Sees Global

The key observation is that real contours strongly activate both V1 and V2, whereas illusory contours (of several different types) elicit activation in V2 but not in V1. Thus while the 'essence' of the orientation signal is similar between real and illusory contours, the views in V1 and V2 of these two contour signals are quite different. The key point is that whereas V2 can readily distinguish the orientations of different types of contour stimuli, it is clear that V1 cannot distinguish the orientations of illusory contours. A corollary is that the encoding of contour type (real or illusory) cannot be accomplished by either V1 alone (signals either lack orientation contour or orthogonal orientation) or V2 alone (V2 cannot distinguish between real and illusory). Only with a paired V1 and V2 signal can the real versus illusory nature of a contour be decoded. Thus, both higher

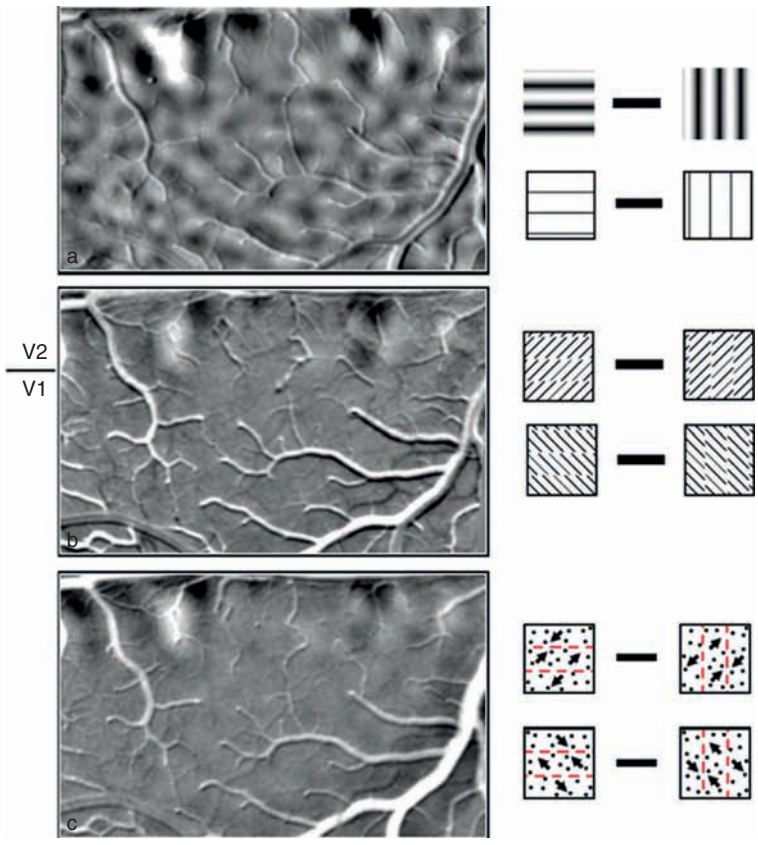

Figure 13 Local and global V1–V2 networks inferred from cross-correlation studies. (a) Local networks are those that require spatially overlapping receptive fields. Such interactions are typically seen between orientation-matched cell pairs and between spatially overlapped nonoriented V1 and oriented V2 cells. (b) Global networks are those that do not require receptive field overlap. Strong interactions are typically seen between nonoriented V1 and nonoriented V2 cell pairs (surface–surface), oriented V1 and nonoriented V2 cell pairs (border–surface), and nonorientation-matched V1–V2 cell pairs.

and lower-order signals are necessary for proper contour identification.

Local versus Global Networks

How Is a Contour-Specific Paired V1–V2 Signal Achieved?

There is a broader view of V2 function which applies to the multiple feature spaces represented within V2. The view described earlier is a three-stage view of the relationship between V1 and V2. It is a framework based on both anatomical and functional connectivities between different types of V1 and V2 neurons. In brief, the first stage (V1) identifies elemental features such as local color, orientation, absolute disparity, and motion. The second stage (V1–V2 or intra-V2) involves surface-border capture, in which appropriate assignments of surfaces to borders and borders to surfaces are achieved. The third stage involves an competitive balance of feedforward and feedback interactions between V1 and V2. This is the stage via which global percepts might achieve dominance over local cues, perhaps via feedback from V2 to V1.

Based on cross-correlation studies of V1–V2 interactions, two types of V1–V2 networks may be described, one which subserves processing of local cues and another which is more global in extent. Based on approximately 250 recorded interactions between different cell types in V1 and V2 (e.g., oriented broadband, oriented color-selective, nonoriented broadband, nonoriented color-selective), two types of V1–V2 interactions have emerged. The local type of interaction requires spatial overlap of receptive fields. We have found that V1–V2 orientation-matched interactions, as well as some interactions between nonoriented color cells in V1 and oriented color cells in V2, are highly dependent on spatial overlap (Figure 14(a)). In contrast, interactions between nonoriented color cells and oriented cell pairs with nonmatching orientation selectivity do not require spatial overlap (Figure 14(b)). The interactions without spatial overlap may be mediated via extensive intra-V2 connections (stripe-to-stripe arrows in Figure 14(b)). Based on these recordings, it is hypothesized that the local interactions subserve local cue identification (e.g., local contour identification could arise via either Hubel–Wiesel-like integration of nonoriented input from V1 or feedforward of oriented signals already generated within V1). Global interactions subserve surface integration (integration and dissemination of surface feature information) as well as appropriate integration of surfaces and borders. This integration may involve surface capture (feedforward of surface-to-border signals via oriented V1 to nonoriented V2 interactions) or border capture. It is proposed that

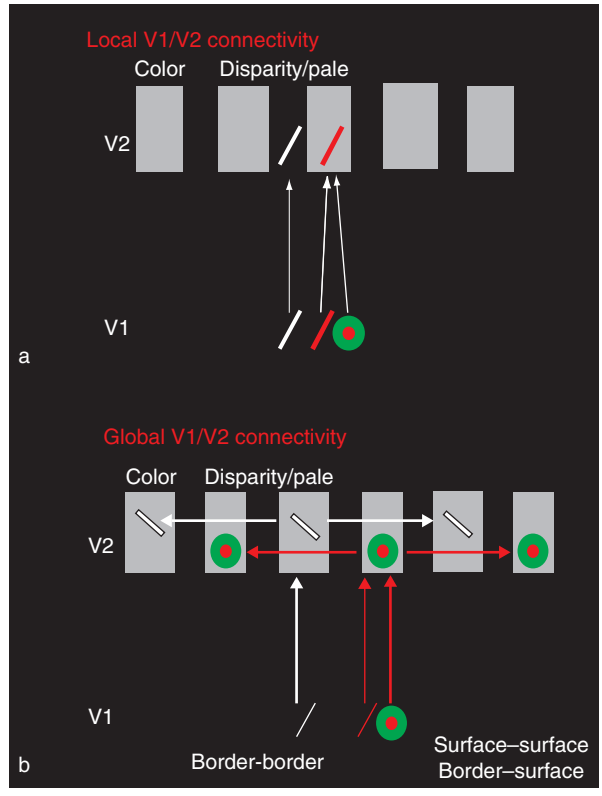

Figure 14 Cue-invariant orientation domains in V2. (a) Different types of contour stimuli: sine wave gratings, line gratings, illusory contours with acute inducers, illusory contours with obtuse inducers, motion contrast illusory contours. (b) Orientation maps dervied from real lines (left) and from illusory contours (right). (a) From Lu HD and Roe AW (2007) *Response to Motion and Motion Boundary in Monkey V2*. Sarasota, FL: Vision Sciences Society. (b) From Ramsden BM, Hung CP, and Roe AW (2001) Real and illusory contour processing in Area V1 of the primate: A cortical balancing act. *Cerebral Cortex* 11: 648–665.

the larger global network within V2 is able to exert influence on the local network, resulting in context-dependent override. Although the mechanisms by which such override occurs are unknown, one possibility is that the strength and/or extensiveness of the global network in V2 results in activation of a suppressive feedback influence on V1.

A Competitive Balance

Based on these studies, a model of contour processing in early visual cortex has been proposed which comprises a competitive balance between feedforward V1–V2 and feedback V2–V1 influences (Figure 15). In this model, during real contour processing, the feedforward signals dominate and activate matching orientation domains in V1 and V2 (Figure 15(a), solid arrow). During illusory contour processing, feedback signals gain prominence and suppress domains which encode matching 'real' contour orientations

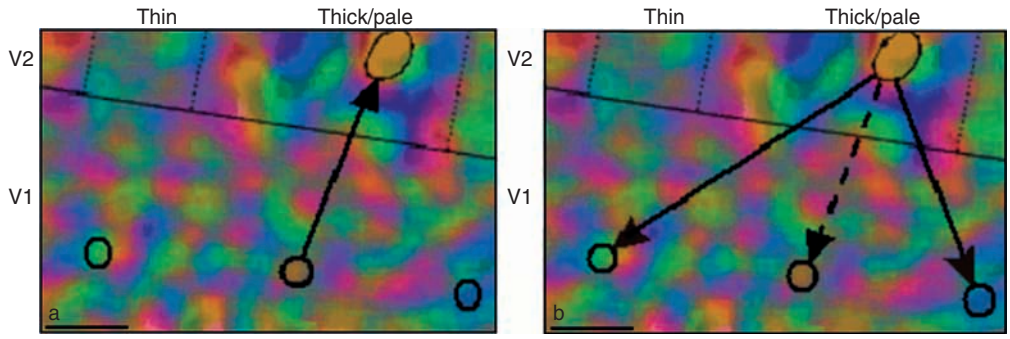

Figure 15 Depiction of feedforward versus feedback forces between V1 and V2. Color-coded orientation map in V1 and V2 (V1–V2 border and stripe borders are indicated). (a) Arrow depicts orientation-specific feedforward influence (local spatial network). (b) Dotted arrow depicts orientation-specific (local spatially localized network) feedback influence from V2 to V1. Solid arrows depict orientation-diverse and spatially extensive (global, spatially less restricted network) influence. Feedback activation is thought to differ between real and illusory contour stimulation. Real contour stimulus preferentially activates orientation-specific feedback (dotted arrow). Illusory contour stimulation preferentially activates orientation-diverse feedback (solid arrows). Scale bars = 1 mm (a, b).

(thereby leading to orientation reversal; Figure 15(b), dotted arrow). In this sense, there is a competitive process between real (feedforward, Figure 15(a)) and illusory (feedback, Figure 15(b)) signals. The perception of any contour would result from a competitive balance between these two orientation-selective forces.

This hypothesis has been tested using psychophysical methods. The reasoning is that if illusory contours activate suppressive feedback pathways, then adding a small amount of real input might push the balance in the feedforward direction. In other words, adding a real line on top of an illusory contour should change the balance. Increasing the contrast of this real line (from subthreshhold to suprathreshhold contrasts) should push the balance further in the feedforward direction and interfere with feedback-dependent illusory contour perception. Furthermore, since the original observation was one of orientation reversal in V1, the feedback interaction should act in an orientation-selective manner. In other words, it should matter whether the real line added is parallel or orthogonal to the illusory contour. Results support the predictions. Parallel lines had little effect at subthreshold contrasts but tended to interfere with discrimination at suprathreshold contrasts. This interference was orientation dependent, as orthogonal lines had no effect on discrimination at suprathreshold contrasts but interfered only at subthreshold contrasts. These psychophysical data thus support an orientation-specific and contrast-specific influence of real contours on illusory contour perception.

Summary

These studies provide further weight to the idea that V1 responds to local elements within the scene, while V2 encodes more global aspects. Relative to V1, V2 behaves in a more cue-invariant fashion. As summarized in Table 1, whereas V1 responds only to local luminance cues, thin stripes encode generalized brightness percepts. Whereas V1 encodes the orientation of only real contours, V2 exhibits cue-invariant encoding of contour orientation. Whereas V1 contains cells responsive to the disparity between pixels presented to the left and right eyes, whether they are the same or opposite in contrast, V2 contains a map of global coherent disparity response that a global coherent disparity that correlates with the perception near and far percepts.

This article (see Table 2) has summarized a body of evidence supporting the idea that each cortical area processes the visual world at a different spatial scale and, in that sense, 'sees' a different visual world. We suggest that these multiple views of the world compete with each other for dominance, perhaps via a competitive balance between feedforward and feedback signals between different cortical areas. This competition is necessarily context dependent and therefore dynamic. Additional studies in the awake, behaving preparation are likely to provide additional insight into the dynamic aspects of this proposed competition.

These studies have shown that the basis for some global percepts resides within 100–500 μm modules

Table 2 Local versus global competition

Why so many cortical areas?
Each area 'sees' different view
(V2 sees different things than V1 sees)
Which view dominates?
Balance of power between areas
(Competition between V1 and V2 dominance)
Dynamic network of V1–V2 modules
Correlation with behavior

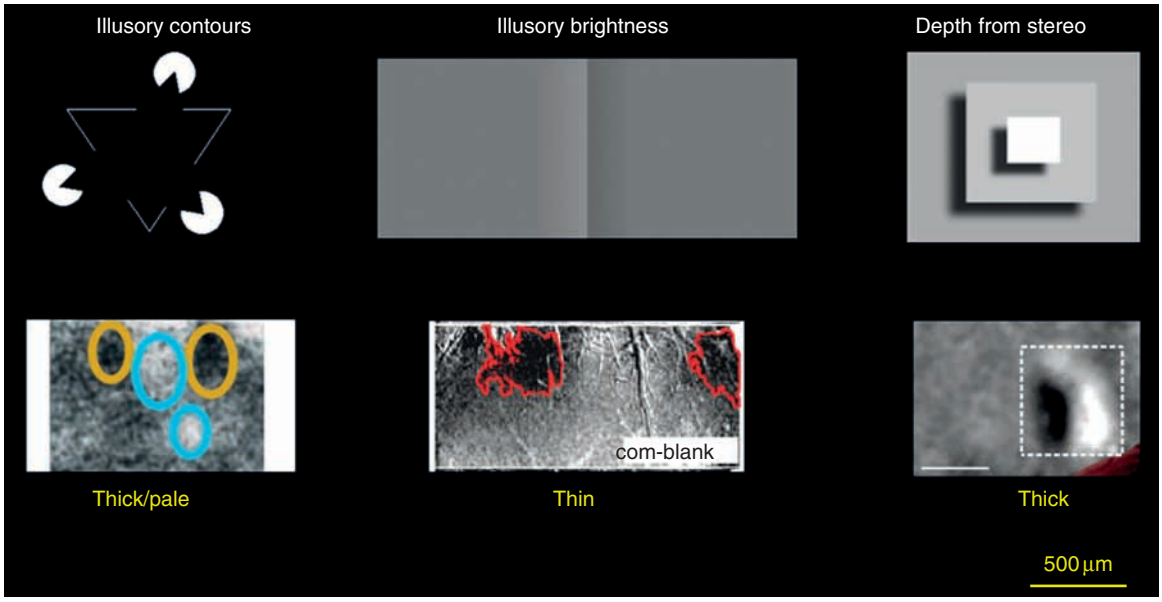

Figure 16 Modular basis of behavior. Higher-order percepts such as illusory contours, illusory brightness and color percepts, and stereo depth are encoded in 100–500 μm modules within specific V2 stripes. These modules include orientation domains (left, orange: horizontal; blue, vertical) in thick/pale stripes, 'brightless' modules (middle, red: modules responsive to real and illusory brightness modulation) in thin stripes, and stero depth modules (right, dotted box: all near (dark pixels) minus all far (light pixels)).

in V2 (Figure 16). Cortical modules clearly underlie not only encoding of elemental features such as contour orientation and color preference in V1, but also higher-order percepts such as illusory contour orientation, illusory brightness and color percepts, and stereo depth percepts. In relation to feedforward versus feedback competition, it is likely that in the awake, behaving brain, the relative dominance of V1 and V2 will be evident through the balance of V1 versus V2 modular networks.

See also: Contextual Interactions in Visual Perception; Contextual Interactions in Visual Processing; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging; Visual System: Multiple Visual Areas in Monkeys; Visual Cortex in Humans; Visual Cortical Models of Orientation Tuning.

Further Reading

Amir Y, Harel M, and Malach R (1993) Cortical hierarchy reflected in the organization of intrinsic connections in macaque monkey visual cortex. *Journal of Comparative Neurology* 334(1): 19–46

Chen G, Lu HD, and Roe AW (2006) Functional architecture of macaque cortical area V2 for depth surfaces revealed by optical imaging. *Society for Neuroscience* 32: 114–116

Cumming BG and DeAngelis GC (2001) The physiology of stereopsis. *Annual Review of Neuroscience* 24: 203–238.

Cumming BG and Parker AJ (1999) Binocular neurons in V1 of awake monkeys are selective for absolute, not relative, disparity. *Journal of Neuroscience* 19: 5602–5618.

Gegenfurtner KR (2003) Cortical mechanisms of colour vision. *Nature Reviews Neuroscience* 4: 563–572.

Goodhill GJ and Cimponeriu A (2002) Analysis of the elastic net model applied to the formation of ocular dominance and orientation columns. *Network* 11(2): 153–168.

Hendry SHC and Reid RC (2000) The koniocellular pathway in primate vision. *Annual Review of Neuroscience* 23: 127–153.

Hubel DH and Livingstone MS (1987) Segregation of form, color, and stereopsis in primate area 18. *Journal of Neuroscience* 7(11): 3378–3415.

Hung CP, Ramsden RM, and Roe AW (in press) A functional circuitry for edge-induced brightness perception. *Nature Neuroscience*.

Julesz B (1971) *Foundations of Cyclopean Perception*. Chicago: University of Chicago.

Livingstone MS and Hubel DH (1984) Specificity of intrinsic connections in primate primary visual cortex. *Journal of Neuroscience* 4: 2830–2835.

Lu HD and Roe AW (2007) *Response to Motion and Motion Boundary in Monkey V2*. Sarasota, FL: Vision Sciences Society

Lu HD and Roe AW (2007) Functional organization of color domains in V1 and V2 of macaque monkey revealed by optical imaging. *Cerebral Cortex* (doi: 10.1093/cercor/bhm081).

Lu HD and Roe AW (2007) Optical imaging of contrast response in Macaque monkey V1 & V2. *Cerebral Cortex* (doi: 10.1093/cercor/bhl177).

Paradiso MA, Blau S, Huang X, MacEvoy SP, Rossi AF, and Shalev G (2006) Lightness, filling-in, and the fundamental role of context in visual perception. *Progress in Brain Research* 155: 109–123.

Pinna B, Brelstaff G, and Spillmann L (2001) Surface color from boundaries: A new 'watercolor' illusion. *Vision Research* 41 (20): 2669–2676.

Ramsden BM, Hung CP, and Roe AW (2001) Real and illusory contour processing in Area V1 of the primate: A cortical balancing act. *Cerebral Cortex* 11: 648–665.

Roe AW (2003) Modular complexity of area V2 in the macaque monkey. In: Collins C and Kaas J (eds.) *The Primate Visual System*, pp. 109–138. New York: CRC Press.

Roe AW, Lu HD, Crewther D (2007) Blue Color Activation in V1 and V2 of Macaque Monkey. Melbourne: International Brain Research Organisation

Roe AW, Lu HD, and Hung CP (2005) Cortical processing of a brightness illusion. *Proceedings of the National Academy of Sciences of the United States of America* 102: 3869–3874.

Roe AW and Ts'o DY (1995) Visual topography in primate V2: Multiple representation across functional stripes. *Journal of Neuroscience* 15: 3689–3715

Roe AW and Ts'o DY (1999) Specificity of color connectivity between primate V1 and V2. *Journal of Physiology* 82: 2719–2730

Rossi AF, Rittenhouse CD, and Paradiso MA (1996) The representation of brightness in primary visual cortex. *Science* 273: 1104.

Sherman SM and Guillery RW (2002) The role of thalamus in the flow of information to cortex. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357: 1695–1708.

Sincich LC and Horton JC (2005) The circuitry of V1 and V2: Integration of color, form, and motion. *Annual Review of Neuroscience* 28: 303–326.

Solomon SG, Pierce JW, and Lennie P (2004) The impact of suppressive surrounds on chromatic properties of cortical neurons. *Journal of Neuroscience* 24: 148–160.

Swindale NV (2004) How different feature spaces may be represented in cortical maps. *Network* 15(4): 217–242.

Thomas OM, Cumming BG, and Parker AJ (2002) A specialization for relative disparity in V2. *Nature Neuroscience* 5(5): 472–478.

von der Heydt R and Peterhans E (1989) Mechanisms of contour perception in monkey visual cortex. Part I: Lines of pattern discontinuity. *Journal of Neuroscience* 9: 1731.

Wang Y, Xiao Y, and Felleman DJ (2007) V2 thin stripes contain spatially organized representations of achromatic luminance change. *Cerebral Cortex* 17(1): 116–129.

Xiao Y, Wang Y, and Felleman DJ (2003) A spatially organized representation of colour in macaque cortical area V2. *Nature* 421: 535–539.

Visual System: Invertebrates

T W Cronin and M L Porter, University of Maryland Baltimore County, Baltimore, MD, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Of all the senses used by animals to gather critical information about their environments, vision is most important for providing rich, quick, and spatially detailed perceptual views of the surrounding world. In fact, it has been suggested that the first appearance of quality eyes in animals at the dawn of the Cambrian drove the rapid appearance of complex, formidable body plans as early invertebrates escalated their capabilities as long-distance predators – or as quick-evading prey. These eyes must have evolved from earlier sets of photoreceptors that had been devoted to sensing of light and darkness for photoperiodism, setting of circadian rhythms, simple orientation in a directional light field, or perhaps for the detection of shadows for the initiation of evasive responses. None of these functions demand much (or even anything) in the way of spatial resolution, but they provide a basic foundation for visual evolution. The formation of pigment cup eyes by adding receptor cells in a sheet to make a primitive retina and restricting how light reaches this retina by means of a surrounding pigment layer provided the likely ancestors of all eye types that animals use today.

Perhaps because some invertebrates still resemble the ancestral forms of all modern animal diversity, it is often thought that invertebrate neurobiology, and specifically invertebrate vision, is simple, primitive, and relatively ineffective. If anything, the opposite is true. Modern invertebrates have an equally lengthy evolutionary history as any other animal, and being generally small animals with compact nervous systems, their eyes must be extremely efficient at providing critical information. As will be demonstrated in this article, eyes found in one or another species of invertebrate outperform those of any vertebrate species in virtually any measure of visual excellence. In this article, we describe visual systems of invertebrates, starting with the pigments on which vision is based and then proceeding to receptor cell types, eye design, visual function and processing, and higher level aspects of vision.

Invertebrate Visual Pigments

General Properties of Visual Pigments

The basis of any visual system is the visual pigment, a molecule composed of a protein moiety (opsin) and a light-absorbing chromophore. Visual pigments are the photoreceptive molecules responsible for absorbing photons of light and transforming the energy into a cellular signal, initiating the visual cascade and resulting in a physiological response in the photoreceptor cell. The opsin component of the visual pigment is a G-protein-coupled receptor (GPCR) and thus a member of a large superfamily of receptors with similar protein structures. All GPCRs are composed of seven transmembrane α helices, with loops on both the extracellular and the cytoplasmic faces. The chromophore component, a derivative of vitamin A, is covalently attached to the opsin protein by a Schiff's base linkage to a conserved lysine residue in the seventh transmembrane helix. Within invertebrate visual systems, four different chromophores have been characterized, although the predominant chromophore is the A1 form, retinal.

Physiologically, visual pigments are characterized by the wavelength of maximal absorbance, which is determined by the interaction between the opsin protein and the chromophore. Therefore, the spectral absorbance of a visual pigment can be altered, or tuned, either by amino acid replacements, particularly at residues within the chromophore-binding pocket, or by the form of chromophore used. The diversity of invertebrate opsin amino acid sequences in combination with different chromophores has produced a vast range of visual pigment types and color vision systems, particularly within the arthropods, in which visual pigment spectral sensitivities range from deep in the ultraviolet (below 340 nm) to well into the red (600 nm or longer).

Opsin Evolution, with Special Reference to Invertebrates

The first invertebrate species from which the opsin gene was sequenced was *Drosophila melanogaster*, followed much later by sequences from the horseshoe crab *Limulus polyphemus* and then by a flood of data from various arthropods, mollusks, and other invertebrate taxa. Although numerous opsin genes related to visual pigments with a large diversity in spectral absorbance were found in *D. melanogaster*, limited sets of one or two opsins with similar spectral sensitivities are sometimes found in other arthropod species. In fact, many invertebrates have visual systems based on single opsin genes, as is the case for most cephalopod species investigated. Interestingly, although most cephalopod species are thought to lack color vision, in the Japanese firefly squid *Watasenia scintillans*, three different visual pigments have been identified with different spectral properties

based on pairing different chromophores with presumably a single opsin gene. However, most color vision systems in invertebrates are based on variations in the opsin protein.

During the course of metazoan evolution, opsin genes have diversified in relation to visual systems and color vision at least two times, once in arthropods and once in vertebrates (Figure 1). In both of these diversifications, gene duplications have led to multiple spectral classes of visual pigments, with arthropods having three main groups related to spectral tuning of the visual pigment – long wavelength, blue, and ultraviolet. This visual pigment spectral diversification based on opsin gene duplications has led to arthropods' having an enormous diversity of visual systems. The opsin diversity observed in *Drosophila*, and in most hexapod species, fits well with previous hypotheses that the ancestral Mandibulata, a taxonomic group uniting Crustacea, Hexapoda, and Myriapoda, was trichromatic, possessing ultraviolet-, blue-, and green-sensitive visual pigments.

Special Features of Invertebrate Visual Pigments

The independent evolutionary diversification of the visual pigments in different opsin lineages has shaped the general functioning of invertebrate visual systems, which is quite different in many aspects from the better understood vertebrate visual cascade of cellular signaling. In contrast to vertebrates, invertebrate opsins form visual pigments that have a light-activated conformation – metarhodopsin – that is thermally stable (Figure 2). The metarhodopsin state has an absorption maximum that is spectrally shifted relative to the rhodopsin and can be reconverted to rhodopsin with the irradiation of light of appropriate wavelengths. The presence of arrestin, a protein involved in deactivating

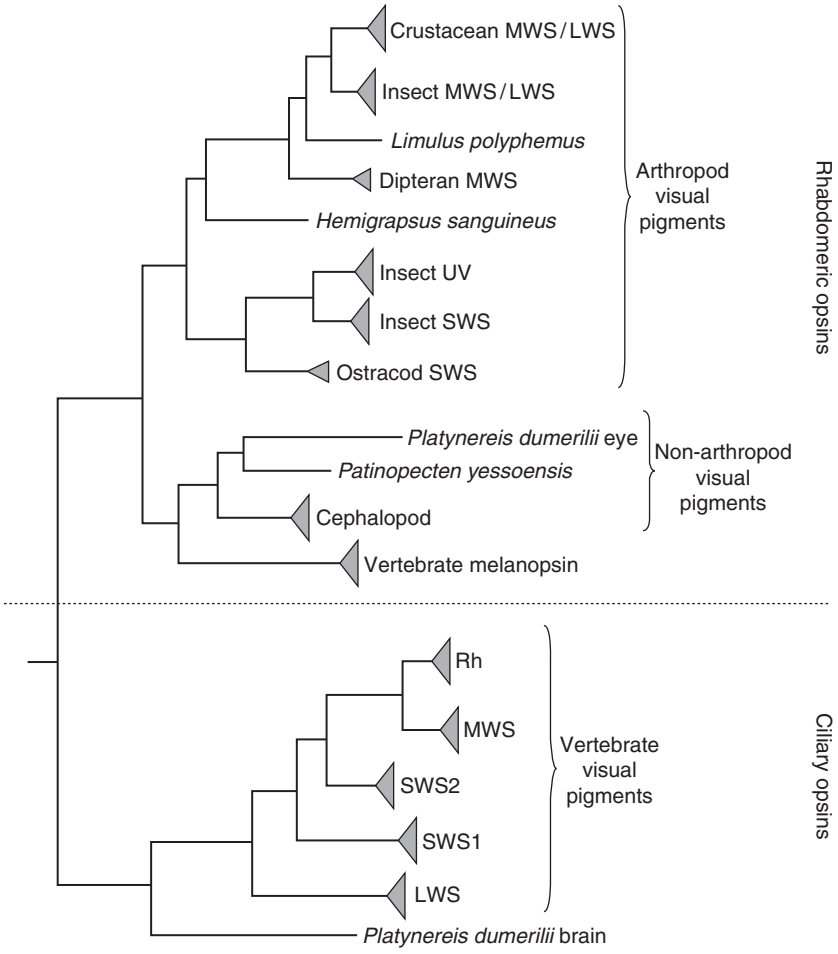

Figure 1 Invertebrate and vertebrate opsin evolutionary relationships reconstructed using maximum likelihood analyses of amino acid sequences from representative species from each spectral group. The spectral range of each clade is indicated as follows: LWS, long-wavelength sensitive; MWS, middle-wavelength sensitive; SWS, short-wavelength sensitive; UVS, ultraviolet sensitive. The vertebrate rod opsins are designated by Rh. Note that the rhabdomeric and ciliary opsin sequence types form two distinct clades. Diversification of opsin genes into multiple spectral classes can be observed in both the arthropod and the vertebrate visual pigment groups. A gray triangle at the end of a branch indicates that opsin sequences from this group have been isolated from a diversity of different species.

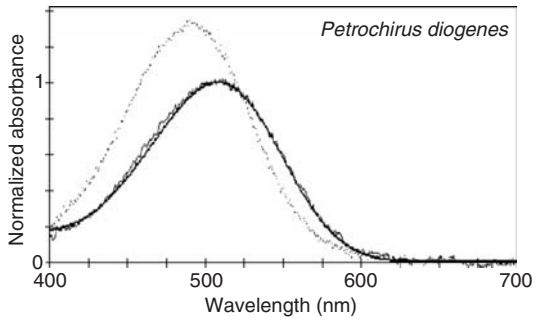

Figure 2 A typical bistable visual pigment system, in this case from the hermit crab *Petrochirus diogenes*. The absorbance of the visual pigment, or rhodopsin, has been normalized to 1.0. The spectral absorbance curves, obtained directly from individual photoreceptors, have been plotted as a thin, jagged line (rhodopsin) or as a dotted line (metarhodopsin). The thick, smooth line is the theoretical spectrum of a visual pigment and has been fit to the data for the rhodopsin. Note that the metarhodopsin absorbs generally at shorter wavelengths than the rhodopsin, and that its peak absorption is approximately 40% higher than that of the rhodopsin.

the light-activated form of rhodopsin, appears to play a role in extending the lifetime of metarhodopsin, allowing a second photon of light to convert metarhodopsin back to the original pigment state.

Unique features of invertebrate visual pigments are also found in chromophore–opsin interactions. In vertebrate visual pigments, a glutamic acid in the third transmembrane helix serves as a negative counterion to the positively charged chromophore. In contrast, most invertebrate pigments contain noncharged amino acids – either a tyrosine or a phenylalanine – at this same position. However, all visual pigments (whether vertebrate or invertebrate) have a conserved glutamic or aspartic acid in extracellular loop 2, which has been suggested to act as the invertebrate visual pigment counterion and may be the ancestral counterion for all visual pigments.

In general, invertebrate visual pigments are associated with rhabdomeric photoreceptor cells, whereas vertebrate visual pigments are associated with ciliary photoreceptor cells. Furthermore, invertebrate visual pigments typically activate the Gq class of G-proteins, leading to a depolarization of the receptor cell, whereas vertebrate visual pigments activate transducin, leading to a hyperpolarization of the cell. However, although it is true that most visual systems in invertebrates are based on rhabdomeric photoreceptor cells, both photoreceptor cell types and, therefore, opsin types are found throughout animal evolution, indicating that at some point an ancestral metazoan contained two different types of photoreceptor cells, each expressing a different copy of the opsin gene. Within most invertebrate lineages, the rhabdomeric photoreceptors and the associated opsins diversified into the visual pigments, whereas the ciliary photoreceptors and opsins were used in nonocular photoreception systems or took on as yet unidentified functions.

Invertebrate Photoreceptor Cells

Visual pigments, being GPCRs, are confined to the membranes of photoreceptor cells. Consequently, since visual cells generally must be effective photon catchers, their membranes are typically highly proliferated to insert as much visual pigment as possible into the arriving photon stream. As already noted, these receptor cells fall into two classes, each associated with one of the two types of opsins. Most invertebrates are now thought to use both opsin types and thus both receptor designs. However, all examples of highly developed visual systems among invertebrates are based on rhabdomeric photoreceptors. Therefore, this stands in stark contrast to the vertebrates, all of whom base their vision on ciliary photoreceptors. Most ciliary photoreceptors of invertebrates that have been studied seem to be used for sensing of light related to circadian function or other, similar nonvisual tasks.

Within the proliferated microvilli that are so characteristic of rhabdomeric receptor cells, absorption dipoles of the chromophores in visual pigments are generally aligned roughly parallel to the microvillar axis. Because most photoreceptive microvilli are shaped into straight tubes built around a stiff actin core (Figure 3), the visual pigments of a single microvillus are all oriented in more or less the same way. Furthermore, since each visual pigment molecule preferentially absorbs photons when the electric vector of polarization (the *e*-vector) of the photon is parallel to the chromophore's absorption dipole, the entire microvillus responds sensitively to the plane of polarization of incoming light. Photoreceptor cells of many invertebrate species are built specifically to enhance this inherent property. Microvilli are aligned parallel along the length of the cell, with their axes extended across the rays of light entering the eye. Consequently, the entire cell responds as a polarization analyzer, a property that permits orientation to natural polarized light fields. Moreover, if two or more classes of analyzers can be combined in a single retina, the visual system has the potential to see polarization patterns in the environment. A number of species of crustaceans, for example, have found a way to do this particularly elegantly. The microvilli of a ring of photoreceptor cells extend into a central fused photoreceptor, or rhabdom; and together the cells form two populations having perpendicular sets of microvilli (Figure 3). The microvilli of the two

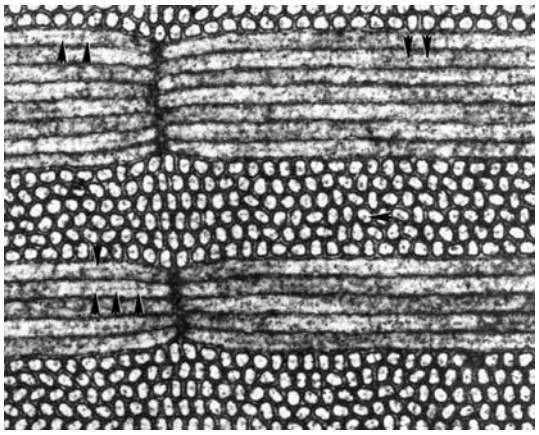

Figure 3 Electron micrograph of microvilli from several individual photoreceptor cells in a single rhabdom of a mantis shrimp, *Neogonodactylus oerstedii.* The microvilli are in alternate layers, with those of successive layers facing parallel to the plane of the section (seen as long thin tubes) or perpendicular to it (seen as circular cross sections). This arrangement is ideal for a two-axis analysis of polarization since the absorption dipoles of visual pigments in microvilli are typically parallel to the microvillar axis. The arrowheads indicate actin fibers that serve as a skeletal framework for each microvillus. Photo CA King-Smith.

populations extend laterally into the rhabdom in a series of layers. Each layer's microvilli are orthogonal to those above and below, so the rhabdom as a whole encodes polarization information arriving from one point in visual space. The entire retinal array of receptors is therefore well suited to see polarization throughout the entire visual field. Other major invertebrate groups known to sense polarization include insects, mollusks, and echinoderms.

Invertebrate Eyes and Visual Optics

The simplest possible photoreceptive organ that can properly be called an eye must have three components: an array of at least a small number of photoreceptor cells, a surrounding shield of opaque pigment to block most light from reaching this retina, and a restrictive aperture or opening in the pigment shield that allows the retina to be illuminated from some part of the visual surround. Eyes like these exist in many small and simple invertebrate animals (e.g., jellyfish and flatworms), in which they are called pit eyes or eyecups. 'Cup' is a good literal description because the retina often extends along the bottom of an invaginated space that is surrounded by cells containing black pigment, and the opening of the cup admits light. These eyes obviously have terrible resolution, but they can distinguish the general direction of illumination (useful for orientation), they can provide a measure of light intensity and its sudden changes, they can in principle detect the movement of a shadow or edge, and (if the opening to the cup is fairly small) they can encode the general distribution of light and dark in the region being viewed. This is a lot of information for any nervous system to process, and it is likely to be all that a simple animal really needs to know.

The complex eyes present in almost all higher animals, invertebrate or vertebrate, certainly originated from cuplike eyes, although it is still a matter of debate whether these can all be traced back to a single ancestral organ. More likely, photoreceptors did arise from a primordial photoreceptive cell, but pit eyes independently appeared many times, and some of these served as the ancestors of superb eyes whose designs survive today. The most successful eye designs, based on the numbers and diversity of species bearing them, are compound eyes and lens (simple, or camera) eyes. Figure 4 illustrates diagrammatically how each can be derived readily from a pigment cup ancestor, and these two eye designs (and other rare and odd ones) are discussed in detail in the following sections.

Simple Eyes

Adding a surface covering (which often is present over pigment cups anyway) and a roughly spherical lens converts the pigment cup to a lens eye (Figure 4). Evolutionary modeling suggests that the process could occur in less than 1 million years, a result that supports the view that complex eyes might well have arisen many times from simpler ones. Lens eyes are often called 'camera eyes' because their operating principles are similar to photographic cameras, having a pigment-lined chamber, a retina in the back (where the film or CCD would be in a camera), and a lens in front; often, there is also a diaphragm or iris to control the amount of light reaching the retina (not shown in Figure 4). A similar eye type, although different in optical details, uses the corneal surface instead of an internal lens to form the retinal image. Since the ability to refract light at a surface relies on a sudden change in refractive index, eyes that rely on corneal refraction are effective only in air; water and typical living tissues are too similar in refractive index for good imaging. High-quality, simple eyes based on corneal or lens refraction are extremely widespread among invertebrates, being found in jellyfish, polychaetes, mollusks, and arthropods (as well as throughout the vertebrates).

Even tiny lens eyes can be optically excellent. Spherical lenses, the easiest to develop, have a number of inherent optical aberrations, but animals counter the most serious of these by controlling variations in the density of the lens material in layers within the lens. As a result, many biological lenses approach the theoretical limits of perfection, at least on the optic axis. By far the largest lens eye in existence occurs in

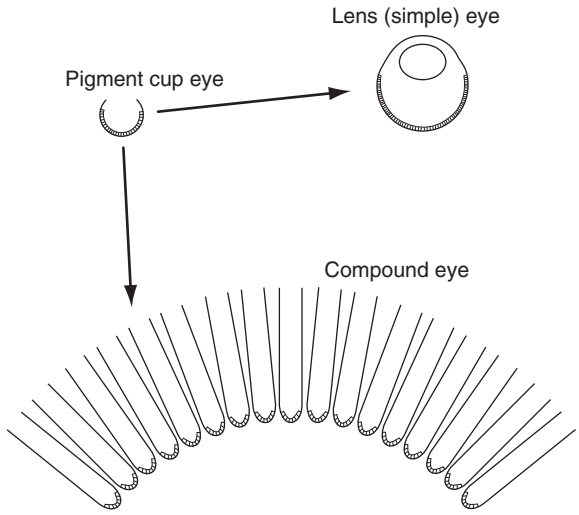

Figure 4 A schematic demonstrating how a pigment cup ancestral eye could be modified either to form a simple, or lens, eye by adding a cornea and a lens or to make a basic compound eye by multiplying in number, extending the pigment sheath upwards, and grouping into a geometrical array.

an invertebrate, the giant squid *Architeuthis*, in which they are approximately 25 cm in diameter – larger than a soccer ball. In principle, these eyes have the best resolution of any eye, although this might be compromised by pooling of receptors for sensitivity in the dim-light environment which these deep sea animals inhabit. Lenses of eyes of cubomedusans, jellyfish without even a central nervous system, are optically excellent, although the retina in these eyes is not placed at the lens's focal plane. The lens eyes of cubomedusans and giant squids cover the range of eye sizes for camera-type eyes, but there are some copepod crustaceans in which the lens is separate from the retina (thus not forming a chambered, camera-type eye) where the lenses are truly tiny – less than 150 μm in diameter.

Compound Eyes

In all likelihood, compound eyes also have arisen independently several times because they are present in such disparate invertebrates as tubeworms, clams, and arthropods. Basic compound eyes can be built simply by grouping pigment cups into an array of multiple tubes, each facing in a different direction (Figure 4). In essence, this is the modern form of the compound eyes occasionally found in annelids or mollusks, in which no other optics are present (however, some worm compound eyes do have simple lenses added). These eyes simply multiply the function of single pigment cups to add some basic spatial information about the distribution of light and shadow and its changes. Optically excellent compound eyes are found only in the arthropods, where they occur widely (in crustaceans, insects, xiphosurans, as well as in the extinct trilobites, eurypterans, and other arthropods). Surprisingly, the optical principles on which these eyes depend are fundamentally different, making it difficult to understand how the various types have arisen from what was presumably a single ancestral type. The most ancient eyes for which the optics are well understood are the compound eyes of trilobites (Figure 5), and in some species these incorporate quite sophisticated optics with unusual features to control aberrations.

Compound eyes have limited resolution since each corneal facet usually corresponds to a single imaging element, and no compound eye has more than a few thousand elements (or ommatidia). Nevertheless, they are obviously very successful, being the type used by the vast majority of animal species. They provide decent modular, panoramic vision, which is probably all that arthropods require. The compound eyes of modern arthropods fall into two optical types. The simpler design is called an apposition compound eye. Here, each corneal facet forms an image onto a single photoreceptor, or rhabdom, within the same ommatidium. These eyes are capable of quite good resolution but often have very limited sensitivity because light is usually collected over a tiny aperture (i.e., the diameter of one facet). Such eyes exist in bright-light, diurnal arthropods such as butterflies, bees, and shore crabs. Superposition compound eyes improve image brightness on the array of rhabdoms by pooling the light from a large number of corneal facets and focusing it on the retina. Here, the images from different facets all overlie images from other facets (hence the name superposition); image sharpness generally suffers, but sensitivity is high. Such eyes are appropriate for dim-light vision and are found in moths, mayflies, and many deep-sea crustaceans. In most cases, the superposition image forms either from refraction within optical elements near the cornea, making what is known as a refracting superposition eye, or by reflection from biological mirrors forming square mirror boxes below the cornea, making a reflecting superposition eye. However, other designs with quite exotic optics also exist.

Other Optical Designs

Because invertebrates are quite diverse in evolutionary ancestry, body plan, and habitat, it is logical that some species should have quite odd eyes. Of these, perhaps the most normal-looking are pinhole eyes, which are nothing but eyecups that have become nearly closed off at the entrance, providing a dim but reasonably sharp image. The best known examples are from the deep-sea *Nautilus*, in which the

Figure 5 A fossilized compound eye of a trilobite, *Phacops rana*. This very well-preserved right eye is suitable for reconstruction of the visual fields of the animal when it was alive, during the middle Devonian period (approximately 350 My before present). This type of eye is called a schizochroal compound eye, and it contains many fewer ommatidial units than the other type found in trilobites, the holochroal compound eye (which is more similar to modern compound eyes). In schizochroal eyes, each large lens probably collected light for a miniature retina, so the entire visual field was possibly reconstructed by patching together miniature subfields within it. Schizochroal eyes apparently represent an evolutionary experiment that ended with the trilobites.

pinhole eye is as large as the lens eyes of their octopus and squid relatives. Some invertebrate eyes look superficially like lens eyes, but the image is actually formed by a mirror behind the retina; the best examples occur in scallops, bivalves whose mantles are lined with dozens of tiny eyes whose mirrors reflect bright blue light. These eyes are not very different from eyecups, or from basic lens eyes, and since many lens eyes have a mirror (tapetum) behind the retina, it is not difficult to understand how an eye with an imaging mirror could evolve. Other unusual invertebrate eyes employ multiple lenses in series, much like a well-engineered camera lens, or have strangely modified compound eyes with a single large lens included in the array.

Visual System Properties

Behind the retina, most invertebrates have successive layers of nervous tissue to process incoming visual information. The details of how this processing is done rely on fundamentally different principles in different groups of invertebrates, and in most cases little is known of how the processing progresses. The best understood systems by far are those of flies, which have served as general models for visual processing in animals as well as models for artificial visual systems or for visual devices designed to control robots. Since most, but not all, invertebrates are small and have relatively small brains and optic ganglia, the hope has been that research to understand how they work should succeed. In fact, imaginative laboratory approaches have often been successful, but many laboratories headed by first-rank researchers have been employed for this success. Small invertebrates can still be fast-moving and behaviorally complex biological machines. In one or another species, invertebrate eyes are more competent than any vertebrate visual organ in sensitivity, resolution, color vision, and polarized light vision. The view that invertebrate eyes are simple is itself simple-minded.

Vision in Dim Light

Visual systems are pushed to their limits in dim-light environments, in which any limitations in optical or phototransductive systems can lead to completely inadequate performance. It is therefore safe to state that the visual systems that exist in animals living in caves, the deep sea, or nocturnal environments under forest canopy have evolved from bright-light ancestors. Extreme visual sensitivity generally implies large eyes, not available to small animals such as most invertebrates, although it has been claimed one huge invertebrate, the giant squid *Architeuthis*, has the most sensitive eyes (as well as the largest) alive today. Surprisingly, however, small deep-sea invertebrates, nocturnal hunting spiders, and night-flying forest insects all have rather small eyes with excellent dim-light vision. In fact, the eye with the highest sensitivity that has been documented in the literature is the apposition compound eye of a deep-sea isopod (a crustacean), *Cirolana borealis*. Recall that apposition compound eyes are typically the least sensitive design, so *Cirolana*'s extreme performance is a surprise. Deep-sea crustaceans in general have excellent light-gathering ability, feeding their rhabdomeric photoreceptors with sufficient photons to see light coming down from the surface at depths of 1000 m or more.

Some of the terrestrial environments with the most light-limited conditions are found under thick forest canopy at night, where on a cloudy night a human can see literally nothing at all. Some species of bees forage here at night, although they do prefer to fly at dawn or dusk rather than at midnight. These insects also have apposition compound eyes, which are optically modified for extreme dim-light performance using large corneal facets with relatively short focal lengths, producing a brighter image than a typical bee eye and falling on a markedly larger rhabdom. Optics alone are not enough, and sensitivity is further boosted by pooling signals across ommatidia (spatial summation) and by increasing photoreceptor integration time (temporal summation). These neural strategies require compromises in acuity and temporal resolution (both

important for flight), but modeling shows that highly adequate resolution both in time and in space can be achieved with careful tuning. As described next, compound eyes (in these cases, of superposition design) can also support polarization and color analysis at night in some insects.

Color Vision

Color vision refers to the ability to discriminate based only on spectral differences between stimuli that differ in no other property. This ability is thought to be widespread among invertebrates, but convincing behavioral demonstrations of its presence have only been documented for arthropods – flies, bees, butterflies, moths, crabs, and mantis shrimps. Color vision in invertebrates is often based on a large diversity of receptors, from probably two types in crabs to three in honeybees and hawkmoths, five in some swallowtail butterflies and possibly flies, and eight or more in mantis shrimps. Color processing is quite well understood in honeybees (in which it is almost as well studied as in primates or other mammals) and, to a lesser extent, in flies and butterflies, but high-level or central mechanisms of color vision have not been studied otherwise, at least not physiologically. Considering the receptor diversity that can contribute to systems of color vision, mechanisms of central analysis are very likely to be rather variable among taxa.

For invertebrates that see in color, identifying objects (e.g., potential mates, food sources, or host plants for egg deposition) can be simplified because color is generally a more reliable property of a surface than brightness. As is implied by the list of arthropod species with color vision, flower identity is often particularly important. Similarly, animals with color vision very often have communication systems that rely on color signals; butterflies and mantis shrimps are particularly colorful animals, and the colors are incorporated into signals that may have very specific significance. In general, however, color signals (either intraspecifically or interspecifically, as with flowers) are not useful until color vision evolves for other purposes, perhaps to expand the range of spectra sampled, to cancel out problems caused by intensity variations, or to improve discrimination of objects in general.

Polarized-Light Vision

Visual systems in some vertebrates are able to sense the polarization of the sky or underwater environment, and a large number of insects use skylight polarization patterns for orientation during foraging or migration. Invertebrates have a natural advantage for seeing polarized light because of their microvillar photoreceptors (Figure 3). Besides this general navigation ability, a few invertebrate species have been demonstrated to visualize patterns of polarized light, an ability sometimes called polarization vision. This kind of polarization analysis is in many respects like color vision because it allows the discrimination of objects and object properties that would otherwise be confused. Reflections of polarization patterns are unlike color patterns in important ways, however. Polarization can vary strongly with direction of illumination, orientation of the reflecting surface, and position of the viewer and therefore is not a reliable property of an object. Despite this, the ability to discriminate patterns of polarization in visual scenes can be very useful to an animal. Polarization does indicate the orientation of a surface, and many insects use this to look for the horizontal surfaces of natural waters and perhaps for those of leaves or other shiny objects. Underwater, using polarization vision can help animals see farther and can make otherwise well-camouflaged organisms discriminable; this probably explains why many species of cephalopod mollusks (octopus, squid, and cuttlefish) use only polarization vision, having no sense of color.

Research shows that, most surprisingly, many invertebrates use polarization patterns on their own bodies as signals, analogously to color signals. Unlike passive reflection of polarized light, these signals often maintain a nearly constant polarization appearance as lighting or body orientation shifts, making them very reliable as intraspecific signals. The optical mechanisms used to make these polarization-constant signals are diverse and poorly understood. Most invertebrates that use polarized-light signals live in the sea, but some tropical butterflies also signal using polarized light.

Conclusions

Visual systems of invertebrates are surprisingly diverse, with many optical designs having been derived from simple eyecup ancestors. Many invertebrates that are small and have limited nervous processing power have highly competent visual abilities, with outstanding dim-light performance, color vision, and ability to see patterns of polarized light. Invertebrate visual systems are important in research not only for their significance to general biological processes but also as models for biomimetic imaging devices.

See also: Color Vision; Photoreceptor Adaptation; Photoreceptors: Physiology; Vision: Light and Dark Adaptation; Visual System Development: Invertebrates.

Further Reading

Ali MA (ed.) (1984) *Photoreception and Vision in Invertebrates*. New York: Plenum.

Autrum H (ed.) (1981) *Handbook of Sensory Physiology: Vol. VII/6, Vision in Invertebrates*. Berlin: Springer.

Collin SP and Marshall NJ (eds.) (2003) *Sensory Processing in Aquatic Environments*. Berlin: Springer.

Gartner W and Towner P (1995) Invertebrate visual pigments. *Photochemistry and Photobiology* 62: 1–16.

Greiner B (2006) Adaptations for nocturnal vision in insect apposition eyes. *International Review of Cytology* 250: 1–46.

Horváth G and Varjú D (2004) *Polarized Light in Animal Vision*. Berlin: Springer.

Land MF and Nilsson D-E (2002) *Animal Eyes*. Oxford, UK: Oxford University Press.

Marshall NJ, Jones JP, and Cronin TW (1996) Behavioural evidence for color vision in stomatopod crustaceans. *Journal of Comparative Physiology A* 179: 473–481.

Plachetzki DC, Serb JM, and Oakley TH (2005) New insights into the evolutionary history of photoreceptor cells. *Trends in Ecology and Evolution* 20: 465–467.

Prete F (ed.) (2004) *Complex Worlds from Simple Nervous Systems*. Cambridge, MA: MIT Press.

Stavenga DG and Hardie RC (eds.) (1989) *Facets of Vision*. Berlin: Springer.

Warrant EJ (2004) Vision in the dimmest habitats on Earth. *Journal of Comparative Physiology A* 190: 765–789.

Warrant EJ and Nilsson DE (eds.) (2006) *Invertebrate Vision*. Cambridge, UK: Cambridge University Press.

Wehner R (2001) Polarization vision – a uniform sensory capacity? *Journal of Experimental Biology* 204: 2589–2596.

Visual System: Multiple Visual Areas in Monkeys

D J Felleman, University of Texas Medical School, Houston, TX, USA

© 2009 Elsevier Ltd. All rights reserved.

Historical Overview of Visual Cortical Organization

The earliest views of the organization of visual cortex in monkeys relied on architectonic and lesion–behavior methods to identify the location, extent, and number of cortical areas involved in visual processing. The architectonic method subdivides cortex according to the laminar densities and sizes of neurons as determined through Nissl stains as well as the distribution of myelin. Using this approach, in conjunction with behavioral studies following surgical lesions or trauma, visual cortex in the occipital lobe was divided into three zones; primary visual cortex, area 17; a parastriate belt, area 18, often referred to as visuopsychic cortex; and a peristriate belt, area 19. Subsequent electrophysiological mapping, single unit functional studies, and additional architectonic studies have further subdivided areas 18 and 19 into a large number of occipital visual areas.

Behavioral lesion studies extended the traditional view of visual cortex to include portions of the temporal and parietal lobes. Lesions of the temporal lobe were found to produce a type of visual agnosia characterized by loss of ability to recognize objects or their significance. Anterior, as compared with posterior, temporal lobe lesions also tended to produce impairments in visual learning. Architectonic studies of ventral temporal cortex identified large regions of the inferotemporal gyrus (ITG) as Brodmann's areas 20 and 21 or as von Bonin and Bailey's area TE. In a parallel study of the baboon, a posterior portion of TE at the anterior border of the occipital cortex was identified as area TEO. Behavioral lesion studies of parietal cortex also demonstrated visual disturbances following lesions of the inferior parietal lobule in the region now known as area 7a and/or LIP. These lesions produced deficits in visual orientation and reaching. Similar lesions in humans produced a form of visual neglect that led investigators to believe that the parietal lobe played an important role in spatial mapping or perhaps attention.

Overall, behavioral studies, later supported by electrophysiological investigations, led to the idea that visual cortex consists of two distinct processing streams with differential contributions to vision. The ventral, 'what,' stream, leading into the temporal lobe, was thought to subserve visual object recognition and memory, while the dorsal, 'where,' stream was thought to subserve visual spatial abilities. While early investigators attributed these two cortical streams to the two major retino-geniculo-cortical pathways that are subserved by the parvocellular and magnocellular layers of the lateral geniculate nucleus, more recent analyses have demonstrated that these two pathways, in addition to the konicellular pathway, contribute to the activity of the ventral, temporal stream. Although area MT, a prominent member of the dorsal stream, is dominated by direct magnocellular inputs, this dorsal, 'where,' stream also receives parallel contributions from these three geniculate streams. Furthermore, multiple subdivisions of the pulvinar nucleus contribute projections throughout visual cortex, thus providing still another parallel subcortical stream of input to visual cortex.

Modern Methods to Study Visual Cortical Organization

The traditional views of visual cortical organization derived from architectonic and lesion–behavior studies have been supplemented over the past decades by a number of technical advances including microelectrode topographic mapping, single unit electrophysiological investigations of neuronal function, neuroanatomical pathway tracing, modern architectonics–chemoarchitectonics, refined behavioral analyses following cortical lesions, reversible inactivation or microstimulation, and functional magnetic resonance imaging (fMRI). These methods have allowed for the identification of visual cortical areas based on up to five major criteria; architectonics, topographic organization, cortical connectivity, receptive field (functional) properties, and behavior contribution–defects following inactivation. As summarized in Table 1, most cortical subdivisions have not been evaluated using all these criteria. Thus, cortical subdivisions differ in the degree to which they are recognized as functionally specific areas (designated A in Table 1) that are identified by convergent information from two or more of these five criteria or recognized as cortical zones (indicated by Z in Table 1) identified by only one or two criteria. Moreover, different cortical subdivisions have been identified using different criteria across laboratories, resulting in a lack of total consensus on the number and organization of visual cortical areas or zones. Nevertheless, there is widespread consensus that macaque visual

Table 1 Criteria used to define macaque visual subdivision

Subdivision name	*Architecture*	*Topography*	*Connections*	*Physiology*	*Behavior*	*Status*
V1	X	X	X	X	X	A
V2	X	X	X	X	X	A
V3	X	X	X	X	N	A
VP	X	X	X	X	N	A
V3A	X	X	X	X	N	A
V4	X	X	X	X	X	A
V4Ad	N	X	X	X	N	A
V4Av	N	X	X	N	N	Z
V4Tp	X	X?	N	?	N	Z
V4Ta	X	X?	N	?	N	Z
PITd	?	X	X	X?	X?	A
PITv	?	X	X	X?	X?	A
CITd	?	N	X	X?	X?	Z
CITv	?	N	X	X?	X?	Z
AITpd	?	N	X	X?	X?	Z
AITpv	?	N	X	X?	X?	Z
AITad	?	N	X	X?	X?	Z
AITav	?	N	X	X?	X?	Z
TLp	?	N	X	N	N	Z
TFp	?	N	X	N	N	Z
THp	?	N	X	N	N	Z
MT	X	X	X	X	X	A
MSTda	X	?	X	X	N	A
MSTm	X	?	X	X	N	A
MSTdv	X	?	X	X	N	A
FST	X	X?	X	X	N	A
PLSTS	?	N?	X	?	N	Z
ALSTS	?	N?	X	?	N	Z
IPa	X	N?	X	?	N	Z
TPOc	X	N?	X	X?	N	Z
TPOi	X	N?	X	X	N	Z
TPOr	X	N?	X	N	N	Z
CIPS (LOP)	X	?	?	X	N	A
PIP	X	?	X	X	N	A
DP	X	?	X	X	N	A
7a	X	?	X	X	X	A
LIPd	X	?	X	X	X	A
LIPv	X	?	X	X	X	A
VIPm	X	?	X	X	N	A
VIPl	X	?	X	X	N	A
MIP	X	N?	X?	X	N	Z
MDP	X	N?	?	X	N	Z
AIP	X	N?	X	X	N	A
PO	X	X	X	X	N	A
V6	X	X	X	X	N	A?
V6a	X	N	X	X	N	A?
8	X	X	X	X	X	A
46	X	?	X	X	X	A
11	X	N	X	?	N	A
12	X	N	X	?	N	A
13	X	N	X	?	N	A

LOP, lateral occipitoparietal; X, criterion met with certainty; X?, criterion met with less certainty; N, criterion not met with certainty; N?, criterion not met with less certainty; ?, criterion not tested or results unclear; A, clearly defined cortical area; A?, less well defined cortical area; Z, defined cortical zone; V1, visual area 1; striate cortex, primary visual cortex; V2, visual area 2; V3, visual area 3; sometimes referred to as V3d; VP, ventral posterior area, sometimes referred to as V3v; V3A, visual area 3A; V4, visual area 4; V4Ad, visual area 4A dorsal; V4Av, visual area 4A ventral; V4tp, visual area 4t posterior; V4ta, visual area 4t anterior; PITd, dorsal posterior inferotemporal area; PITv, ventral posterior inferotemporal area; CITd, dorsal central inferotemporal area; CITv, ventral central inferotemporal area; AITpd, posterior dorsal anterior inferotemporal area; AITpv, posterior ventral anterior inferotemporal area; AITad, anterior dorsal anterior inferotemporal area; AITav, anterior ventral anterior inferotemporal area; MT, middle temporal visual area; MSTda, medial superior temporal dorsal anterior area; MSTm, medial superior temporal medial area; MSTdp, medial superior temporal dorsal posterior area; FST, floor of superior temporal sulcus visual area; PLSTS, posterior lateral superior temporal area; ALSTS, anterior lateral superior temporal

cortex consists of a large number of areas (and zones) that occupy more than 50% of the cortical surface in the occipital, temporal, parietal, and frontal lobes.

Significant progress in the identification of visual cortical areas has been achieved through standardized methods for visualization of anatomical and physiological data on three- and two-dimensional reconstructions of the cortical surface. For example, computer-assisted mapping of cytoarchitecture, myeloarchitecture, and immunoreactivity for SMI-32 and CAT-301 has led to a high-resolution atlas CARET (the Computerized Anatomical Reconstruction and Editing Toolkit) that allows for the analysis and communication of anatomical and physiological data in macaque monkeys as well as in humans. Recent progress in understanding the organization of visual cortex in macaques has come from functional imaging studies (mostly fMRI) that derive the topographic organization of retinotopic areas and the function of areas from indirect estimates of cortical activity based on changes in blood oxygenation and/or blood flow. Issues of cortical organization span many different levels of analysis, ranging from columnar organization and modular organization to areas and processing streams. Each of these levels of organization is best studied using a subset of available tools. For the analysis of cortical areas described here, architectonics, microelectrode mapping and functional studies, and functional imaging methods provide the best indicators of visual cortical area extent and function (see Figure 1).

Occipital Lobe

Area V1

Macaque occipital lobe contains at least ten distinct cortical subdivisions based on two or more key criteria. Primary visual cortex, V1, has a distinct laminar appearance and is perhaps the best-studied cortical subdivision in this species. V1 is characterized by a unique elaboration of layer 4 into a constellation of sublaminae that are well correlated with the pattern of thalamic inputs and intracortical projections. The supragranular layers of V1 are also characterized by a distinctive pattern of cytochrome oxidase blobs that receive direct thalamic input and are correlated with high incidence of color selectivity and low incidence of orientation selectivity. In contrast, cells in the supragranular layers corresponding to the interblob regions are orientation-selective and are spatially organized into distinct orientation pinwheels originating from nonoriented pinwheel centers.

V1 inputs from the lateral geniculate main and intercalated layers establish three parallel streams that convey different types of information. Intracortical circuitry re-orders these three parallel streams, and outputs to higher cortical areas no doubt reflect this intracortical integration. Cortical outputs of V1 arise from layer 4B (dominated by magnocellular lateral geniculate nucleus inputs), from the cytochrome oxidase blobs, and from the interblob regions of supragranular V1. Therefore, the main functions of V1 appear to include (1) formation of binocular representation of the visual field; (2) intermixing of subcortical streams to form new cortical streams; (3) elaboration of spatially organized maps of contour orientation, ocular dominance, and perhaps spatial frequency; and (4) parcellation of information to three to five different extrastriate visual areas.

Area V2

Area V2 is a belt of cortex that nearly completely surrounds V1, with which it shares a common border representing the vertical meridian. V2 contains a full visual field representation, but it is split along the representation of the horizontal meridian such that the lower field is represented in dorsal cortex and the upper field in ventral cortex. Like V1, V2 has a distinctive modular architecture that can be recognized in tissue stained for cytochrome oxidase. V2 is organized into repeating triplets of cytochrome oxidase stripes: thick, pale, and thin stripes that have distinctive cortical connections and physiological properties. V2 thick stripes receive major input from cells in layer 4B of V1, and its neurons tend to be disparity-selective and exhibit length suppression. V2 pale stripes and thick stripes contain orderly representations of contour orientation that can readily be visualized with intrinsic optical imaging. In contrast, V2 thin stripes receive the majority of their input from V1 blob neurons, and their neurons are selective for the surface properties of color and brightness of objects. Furthermore, optical imaging has revealed systematic maps of hue and luminance change within V2 thin stripes. V2 has a dense set of projections to higher cortical areas that arise from specific stripe

area; IPa, area IPa; TPOc, caudal temporal parieto-occipital area; TPOi, intermediate temporal parieto-occipital area; TPOr, rostral temporal parieto-occipital area; CIPS (LOP), caudal intraparietal sulcus area (lateral occipito-parietal area); PIP, posterior intraparietal area; DP, dorsal prelunate area; 7a, area 7a; LIPd, dorsal lateral intraparietal area; LIPv, ventral lateral intraparietal area; VIPm, medial ventral intraparietal area; VIP1, lateral ventral intraparietal area; MIP, medial intraparietal area; MDP, medial dorsal parietal area; AIP, anterior intraparietal area; PO, parieto-occipital area; V6, visual area 6; V6a, visual area 6a; 8, area 8; 46, area 46; 11, area 11; 12, area 12; 13, area 13.

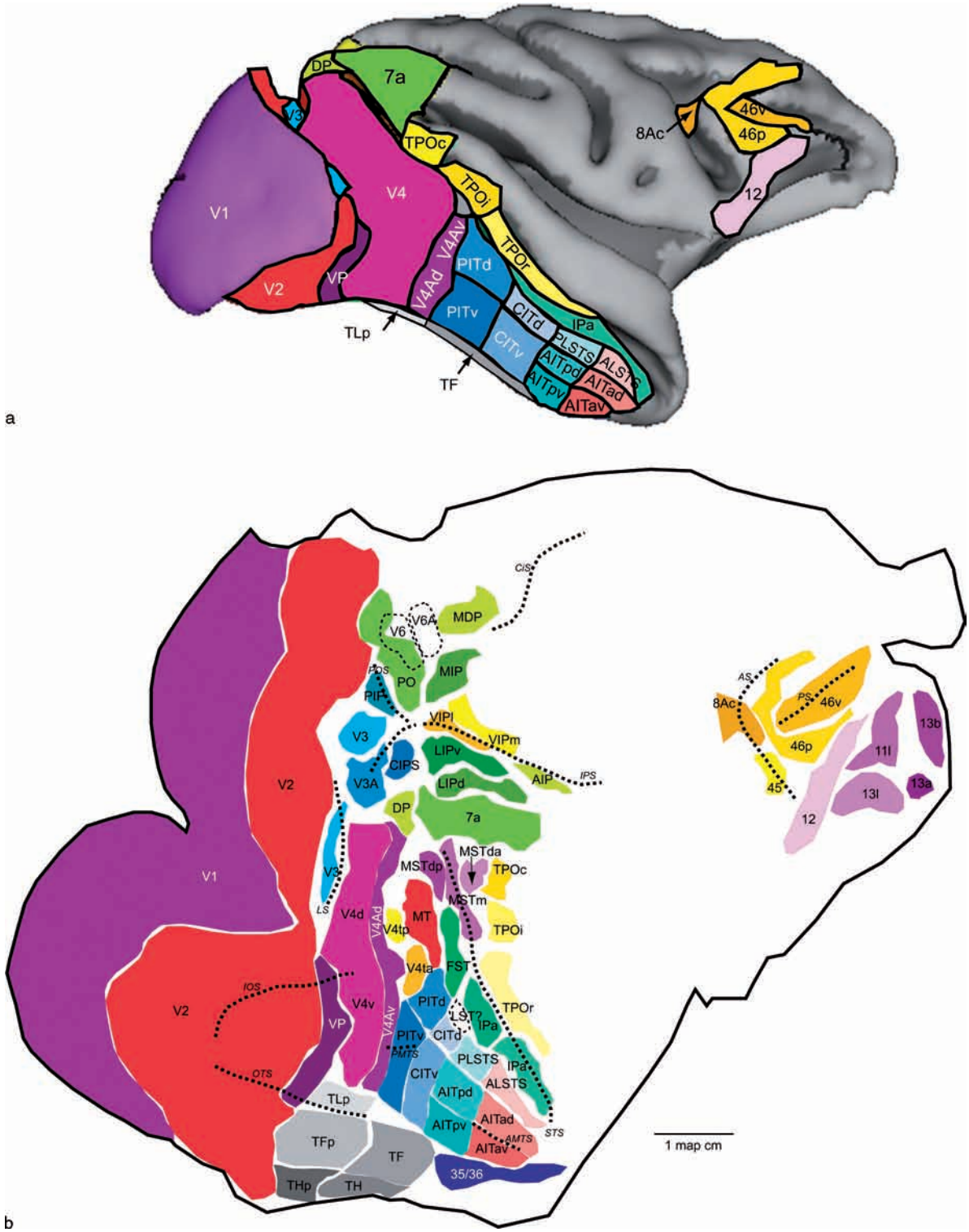

Figure 1 Visual cortical areas of macaque monkeys. (a) Locations of visual areas of macaque monkeys projected onto a partially unfolded surface rendition of cortical layer 4. (b) Locations of visual areas of macaque monkeys displayed on an unfolded representation of cortical layer 4 generated using the Computerized Anatomical Reconstruction and Editing Toolkit software system. The delineation of areas is based on the architectonic studies of Lewis and Van Essen as well as on a large number of studies of cortical connections. Adapted from Lewis JW and Van Essen DC (2000) Mapping of architectonic subdivisions in the macaque monkey, with emphasis on parieto-occipital cortex. *Journal of Comparative Neurology* 428: 79–111.

compartments and thus distributes subsets of information to specific cortical areas. Thus, for example, V2 thick stripes project primarily to areas V3 and MT, with weaker projections to V3A. In contrast, V2 thin and interstripes project primarily to area V4 and TEO (PITv). For the most part, V4 projections to other cortical areas have not been characterized by their stripe compartment of origin.

Areas V3 and VP

Anterior to V2 is another belt of cortex that contains a representation of the visual field that is split along the horizontal meridian. This belt of cortex contains a number of anatomical and physiological asymmetries that have led some investigators to consider the dorsal and ventral halves to represent distinct visual areas. According to this view, dorsal cortex contains a representation of the lower visual field in area V3, which is characterized by a dense myeloarchitecture, distinct V1 inputs arising from layer 4B, and a high degree of direction selectivity and low incidence of color selectivity. In contrast, area VP, in ventral cortex, contains a representation of the upper visual field, is less highly myelinated, is reported not to receive substantial V1 inputs, and has a high incidence of color selectivity and a low incidence of direction selectivity. Some researchers considered the anatomical and functional differences between the dorsal and ventral halves of this cortical belt to be sufficient to warrant their distinction as separate cortical areas. An alternative approach, adopted by other investigators, is to recognize these dorsal–ventral asymmetries by recognizing distinct asymmetries within a single visual area and distinguishing these subdivisions as areas V3d and V3v. Functional imaging indicates that V3, but not VP, contains a high concentration of disparity-selective neurons.

Area V3A

Area V3A is located anterior to V3 on the anterior aspect of the annectant gyrus and posterior bank of the prelunate gyrus. V3A is distinguished from its surrounding cortical areas in that it contains a complete representation of the whole visual field; its lower field vertical meridian representation adjoins V3 (V3d), and its superior vertical meridian adjoins the vertical meridian of area V4. Originally, V3A was described as a larger cortical domain that extended into the posterior portion of the intraparietal sulcus at its junction with the parieto-occipital sulcus. However, more-recent investigations have suggested that the 'second' central field representation within this larger V3A belongs to a distinct visual area currently identified as the posterior intraparietal area (PIP) or the lateral occipitoparietal area (LOP). V3A receives feedforward input from areas V2, V3, and VP and provides feedforward input to areas V4, MT, and several subdivisions of the temporal and parietal lobes. Like V3, area V3A appears critical for binocular disparity analysis.

Area V4

Area V4 is a large cortical area at the anterior margin of the occipital lobe that provides a major input to the temporal lobe and thus is considered a major gateway for the ventral, object processing, visual stream. V4 was first identified as a topographically organized visual area of dorsal occipital cortex that occupied a large callosal-free zone of the prelunate gyrus. This dorsal portion of V4 represents the lower visual field, and its ventral extension, just anterior to area VP, contains a representation of the upper visual field. V4 contains a representation of the visual field that is split along the horizontal meridian, and its posterior borders with V3A, dorsally, and VP, ventrally, represent the vertical meridian. While the extent of V4 in both dorsal and ventral cortex has remained controversial, its position within a dense callosal ring suggests that another visual area or areas are located anterior to V4, within the callosal-free zone. Since area V4 contains a split representation of the horizontal meridian, with its posterior border forming a representation of the vertical meridian, its anterior border should fall within this callosal-free zone. According to this view, an additional area, area V4A (or perhaps a pair of areas, V4Ad and V4Av), adjoins V4 along the representation of the horizontal meridian, and its anterior border corresponds to the representation of the vertical meridian coincident with the anterior callosal band. Subsequent microelectrode mapping and anatomical pathway tracer studies have indicated that a thin strip of cortex adjoins ventral V4 and contains a representation of the upper visual field that is distinct from that contained within ventral V4. Since there are only very limited anatomical or physiological studies of receptive field properties within this V4A strip, this cortical region is tentatively identified as separate dorsal and ventral subdivisions, areas V4Ad and V4Av. In ventral cortex, area V4Av borders posterior inferotemporal cortex along the representation of the superior vertical meridian. The organization in dorsal cortex appears more complex in that V4Ad appears to border area V4t along a noncorresponding lower field vertical meridian–horizontal meridian border. More centrally, V4Ad appears to border dorsal posterior inferotemporal

visual area PITd along the representation of the inferior vertical meridian. Taken together, areas V4, V4Ad, and V4Av provide a major route for visual information to reach the temporal lobe. Whereas early studies suggested that V4 was specialized for color processing ('the color area'), subsequent studies demonstrated a role of V4 in color, shape, disparity, and attentional processing.

Areas MT, V4t (MTc), MST Complex, and FST

One of the major targets of area V1 (layer 4B) is area MT, located in the posterior bank of the superior temporal sulcus (STS). MT is distinguished from other visual areas by its architecture, connections, and nearly complete dedication to motion processing. MT is a relatively small visual area with a complete visual field representation that contains a systematic mapping of the direction of object motion. Furthermore, a modular organization within MT segregates neurons by the degree to which their responses are antagonized by motion in their nonclassical surrounds. Evidence from owl monkey MT suggests that this modular organization is reflected in the pattern of outputs to higher areas such that cells with weak surround antagonism project to area MST, but those with pronounced surround antagonism project to area FST in the temporal lobe. Overall, area MT has major feedforward projections to adjacent area MSTd and to areas VIP and LIPd in the intraparietal sulcus. These parietal projections of MT are thought to subserve the extrapersonal space analysis thought to be an important part of the dorsal, 'where,' cortical processing stream.

Little is known concerning the cortical connections or physiology of area V4t (MTc) in macaques. V4t is generally though to contain only a representation of the lower visual field, and it adjoins MT along the representation of the lower vertical meridian. In owl monkeys and galagos, a topographically corresponding region, area MTc, has been described that nearly encircles MT and thus contains representations of both the lower and upper visual fields. Since the upper field portion of MTc is located dorsal and anterior to MT, it is reasonable to suggest that the upper field representation of area V4t is located medial and dorsal to MT in macaques. This would suggest that a portion of area FST is contained within V4t, and thus the organization of this cortical territory needs closer examination in macaques. V4t has dense connections with area MT and may provide a linkage between the 'dorsal' and 'ventral' processing streams since area MT has weak and/or inconsistent connections with area V4.

Area MT has major feedforward outputs that are directed to both parietal and temporal lobe visual areas that appear to utilize visual motion information for spatial and object processing, respectively. Area MT projects dorsally to the MST complex, which consists of three architectonically distinct subdivisions: area MSTda (the densely myelinated zone) area MSTdp, and area MSTm. Neurons in the MST complex are selective for more-complex forms of visual motion including optic flow and object structure from motion. These cortical subdivisions project to parietal areas 7a, LIPd, and VIP, as well as to posterior portions of multimodal cortex of the roof of the STS (STPp, including portions of TPOc, TPOi, and perhaps TPOr). The second major output of MT is directed anteriorly toward area FST on the floor of the STS. Area FST is architectonically distinct and contains motion-selective neurons that encode the shape of objects based on global motion and texture cues. Thus, area MT plays an important role in distributing visual motion information to parietal and temporal lobe areas involved in spatial and object processing.

Temporal Lobe

Historical Overview

The organization of visual cortex within the macaque temporal lobe has remained a topic of continued investigation for almost a century. Whereas classical architectonic studies suggested that the inferior temporal cortex consisted of one to three distinct subdivisions, subsequent architectonic and connectional studies have continued to distinguish additional subdivisions as well as to subdivide previously defined regions. Early architectonic studies of the temporal lobe concentrated on cortex of the ITG, which was subdivided dorsoventrally into Brodmann areas 20 and 21, with the whole of this region identified as area TE by von Bonin. The posterior portion of TE, adjacent to occipital cortex, was later distinguished as area TEO, based primarily on architectonic studies in chimpanzee or Brodmann's area 37 in human. Behavioral and electrophysiological studies have reinforced this distinction by demonstrating simpler receptive field properties in TEO and more learning- or memory-related functions in anterior portions of TE.

Modern architectonic studies have identified both dorsoventral and anteroposterior differences within TE that distinguish areas TE1, TE2, and TE3. Furthermore, the architectonic analysis into the lateral bank of the STS identified three striplike cortical bands, TEa, TEm, and IPa, that run from near the lip of the ITG to the fundus of the STS. This region was first considered visual on the basis of connections with visual areas of the ITG. Finally, additional architectonic subdivisions of the temporal lobe on the

parahippocampal gyrus, areas TF, TL, and TH and the anteriorly situated perirhinal cortex (areas 35 and 36), were also considered visual on the basis of connections with visual cortical areas.

Posterior ITG

Current ideas concerning the organization of visual areas in the temporal lobe are based primarily on anatomical studies of neural connections, single unit recordings, and functional imaging studies in anesthetized and awake animals. The posterior portion of inferotemporal cortex, previously described as area TEO, is now thought to consist of at least two dorsoventrally distinct subdivisions, areas PITd and PITv. Furthermore, TEO traditionally was thought to adjoin occipital cortex along the anterior border of V4 at the representation of the horizontal meridian. However, based on receptive field mapping, callosal connections, and intracortical connections, two small visual areas, V4Av and V4Ad, are thought to separate V4 from temporal lobe cortical subdivisions. The distinction of PITd from PITv is based on both receptive field topography and cortical connections. PITd is located within the lateral bank of the STS anterior to area MT and medial to area PITv, which is located only on the ITG. Thus, traditionally defined TEO appears to consist of four different cortical areas, although they have yet to be distinguished using architectonic criteria. Microelectrode and functional imaging studies have demonstrated that areas PITv and PITd (TEO) contain neurons that are selective for the spatial configuration of two-dimensional objects and PITv may also be particularly involved in color processing. Finally, a broad swath of cortex of the parahippocampal gyrus, medial to ITG areas PITv and CITv (see below), have been identified architectonically and considered visual on the basis of cortical connections. Although some controversy remains concerning the precise number, architectonic fields TL, TF, and TH have been recognized here. Furthermore, more posterior cortex of the parahippocampal gyrus, medial to area V4 (and perhaps VP) have been identified as occipital or posterior extensions of these more anterior parahippocampal areas and thus have been named TLp, TFp, and THp.

Central and Anterior ITG

Although it is widely agreed that there are multiple cortical subdivisions in central and anterior IT (area TE), there is little agreement on their number, location, or organization. Just considering the ITG, the existence of separate anterior and posterior subdivisions has been commonly accepted, beginning with the distinctions of TE1, TE2, and TE3. A slightly modified version of this scheme identified two subdivisions; posterior TE (TEp) and anterior TE (TEa), with their border falling roughly at the posterior end of the anterior middle temporal sulcus (AMTS). More recently, largely on the basis of cortical and subcortical connections, these two subdivisions of the ITG have been further subdivided into four dorsoventral partitions: TEpv, TEpd, TEav, and TEad. This current partitioning scheme represents an emergent hybrid of schemes based on connections and architectonics and thus provides a useful unifying framework for future functional investigations. Microelectrode and functional imaging studies have demonstrated that neurons display more-complex selectivity for two- and perhaps three-dimensional shape, and many of the constituent neurons carry a signal encoding the memory of a recently presented stimulus. Anterior portions of TE may contain a modular organization such that neurons are clustered by local stimulus features and objects are represented by the distributed activity across a large number of object feature modules. Functional imaging studies also support the view that central and anterior ITG contain a functional clustering based on object categories.

Superior Temporal Sulcus

Cortex within the lateral bank, floor, and roof of the STS contains a number of visual subdivisions that have been recognized on the basis of architecture, connections, single unit properties, and functional imaging. On the basis of architecture, an anteroposteriorly elongated band of cortex within the lateral bank of the STS has been identified as area TEa/TEm. An architecturally distinct narrow band of cortex, area IPa, occupies the fundus of the STS along most of its extent. Cortex within the roof of the STS has been implicated in visual multimodal functions and has been identified as the superior temporal polysensory area (STP). Architectonic studies identified three divisions of TPO that extend from the posterior STS (TPOc) near area MT, through central TPO (TPOi), to anterior TPO (TPOr). This multimodal strip of cortex has connections with portions of the inferior parietal lobule (IPL) and area MST, as well as somatosensory cortex.

On the basis of cortical connections, area TEa/TEm of the lateral bank of the STS has been subdivided into a number of distinct subdivisions. While there remains some controversy concerning the naming of these various cortical divisions, it is clear that the posterior one-third of this strip, immediately anterior to area MT, corresponds to area PITd (see the discussion of ITG above). The middle portion of TEa/TEm has been identified as area CITd on the basis of its connections with areas V4 and PITv. The anterior

one-third of TEa/TEm has previously been described as the dorsal subdivision of TEa, but the connections of PITv and CITv suggest that the dorsal subdivisions of TEa reside on the ITG and that additional cortical subdivisions occupy the anterior lateral bank of the STS. This anterior stretch of TEa/TEm has been tentatively subdivided into anterior and posterior divisions (corresponding to anterior and posterior subdivisions of anterior inferotemporal cortex and labeled PLSTS and ALSTS in Table 1).

Microelectrode and functional imaging studies of visual processing within the STS also reveal a number of distinct subdivisions related to different aspects of high-level processing of object shape, including three-dimensional structure from motion and shape from texture cues. Differential activation to intact versus scrambled objects is observed throughout the STS beginning in area FST and continuing along the lateral STS to include areas CITd, PLSTS, and ALSTS. In addition, activity was observed within the middle and anterior portions of the upper bank of the STS (areas TPOi and TPOr) and within the AMTS of the anterior ITG. It is important to note that differential activation to three-dimensional motion and texture-defined objects has generally not been observed in posterior and middle portions of the ITG (areas PITv and CITv and perhaps others). This is somewhat surprising since these cortical subdivisions demonstrate pattern and shape-selective responses, but they may be more related to two-dimensional form. Finally, a motion-sensitive static-sensitive, and moving shape-sensitive region, area LST, has been identified along the floor of the STS, approximately 5 mm anterior to area FST. This region partially overlaps with architectonic area IPa and medial portions of TEa/TEm and appears to be located medial to other previously identified 'shape from motion' areas within the lateral bank of the STS (e.g., area CITd). Overall, these functional imaging results demonstrate that object shape, extracted from motion cues, is processed by a constellation of areas within the STS.

Parietal Lobe

The inferior and superior parietal lobules contain a constellation of visual subdivisions that can be distinguished on the basis of architectonic or connectional criteria, topographic organization, or other functional criteria. These subdivisions vary in their degree of sensory, motor, and multimodal function, with some areas appearing almost entirely visual sensory (e.g., PIP, CIPS (LOP), and PO), others containing a high percentage of visuomotor neurons (e.g., LIPd, LIPv, and 7a), and still others likely to be multimodal (e.g., visuosomatosensory or visuovestibular; VIPm/VIPl, MIP, and MDP). Traditional views of visual functions of the parietal lobe have focused on its role in space perception (the 'where' function of the dorsal visual system), and lesions of this region produce a visual neglect of the associated visual field. However, recent anatomical and physiological studies suggest that this dorsal visual system consists of two separate streams that have different roles in visual spatial perception, eye movements, and arm and hand guidance. One stream, originating in medial occipital cortex (area PO or V6), provides visual information to areas V6A and MDP, which have complex sensory-motor functions related to arm movements to visual targets. In humans, lesions in presumed homologous regions of the superior parietal lobule produce an optic ataxia characterized by errors in reaching, particularly toward objects in the visual periphery. The second dorsal stream reaches parietal cortex through projections from areas MT and MST (complex), which are directed primarily at subdivisions of the inferior parietal lobule, areas VIP and LIPd. These cortical areas appear to utilize motion information to build a representation of three-dimensional space. Unlike occipital areas, VIP appears to utilize vestibular and other sources of information to build a true head-centered, rather than occulocentric, representation of visual space. Area LIPd provides presaccadic activity that is utilized by the frontal eye fields to guide eye movements to visually selected targets.

Cortical areas VIP and AIP of the inferior parietal lobule are also involved in visual guidance of hand and arm movements, which is consistent with their connections with posterior visual areas and anterior somatosensory and motor cortex hand representations. In humans, lesions of presumed homologous areas of the inferior parietal lobule produce a disorder known as apraxia, which involves an inability to retrieve motor plans with respect to specific visual objects. This disorder exemplifies how the IPL utilizes visual spatial information in conjunction with object shape information to guide specific hand movements. Thus, parietal cortex may best be viewed as a constellation of areas that provide a spatial reference system to guide eye and arm–hand movements. However, visual areas of the IPL also appear to utilize shape information to guide specific hand and arm movements necessary for the appropriate manipulation of visually presented objects. Finally, functional imaging studies have identified shape from motion-selective responses in parietal areas LIPd and LOP.

Frontal Lobe

Frontal cortex is usually considered to participate in high-level cognitive function, including attention,

working memory, and social emotional evaluation of stimuli. Although most of prefrontal cortex is multimodal in function, its involvement in visual processing is perhaps the best understood. Prefrontal cortex has been traditionally divided into dorsal, dorsolateral, and orbital subdivisions on the basis of cytoarchitecture, connections, and physiology. In general, visual processing in prefrontal cortex can be characterized as consisting of three circuits, each associated primarily with either the dorsal parietal, the ventral inferotemporal, or the anterior inferotemporal medial temporal processing stream, although there is some overlap. The first prefrontal circuit consists primarily of two cortical subdivisions, within dorsal and dorsolateral prefrontal cortex, that have been implicated in visual function by virtue of both their physiology and their preferential connections with visual areas in parietal cortex. Microstimulation of dorsal portions of area 8 produces large-amplitude saccadic eye movements, while stimulation of more-lateral cortex at the border with area 45 produces small-amplitude saccades. Areas 8 and 45 receive direct projections from both parietal area LIPd and posterior and anterior inferotemporal cortex. The more anteriorly situated area 46 is more closely associated with visual executive function and is implicated in many aspects of visual working memory. Overall, this circuit appears to emphasize spatial information in working memory tasks. The second prefrontal circuit is also involved in visual working memory but, unlike dorsal prefrontal cortex, appears more object specific. Prefrontal area 12 is also implicated in visual function, both by its dense connections with anterior inferotemporal cortex and by its visual properties, which share many similarities with those observed in anterior inferotemporal cortex. These results suggest that anterior inferotemporal cortex and prefrontal area 12 participate in a cortical circuit underlying visual recognition memory. Finally, orbitofrontal areas 11 and 13 are implicated in visual function by their reciprocal connections with anterior inferotemporal cortex and receptive field properties. In contrast to the prefrontal circuit involving area 12, orbitofrontal areas 11 and 13 may contribute to the analysis of the social-affective and motivational aspects of visual stimuli.

Hierarchical Organization of Visual Cortex

Neuroanatomical studies have demonstrated that cortical areas are almost always reciprocally interconnected and that the laminar locations of labeled cell bodies of origins and layers of axonal terminations provide evidence for the direction of information flow through a hierarchy of visual areas. In general, feedforward pathways arise from cells located in the supragranular cortical layers, and their axons terminate in layer 4 and deep layer 3. Feedback pathways arise from infragranular layers and terminate outside layer 4, usually in layers 1 and 6. Lateral pathways, thought to link cortical areas at the same hierarchical level, arise from both supragranular and infragranular layers and usually terminate in a column spanning all cortical layers. By making pairwise comparisons between all interconnected visual areas and placing each target of a feedforward pathway one level higher than all areas that provide its input, researchers have been able to assign more than 30 cortical areas to one of seven (or more) levels of an anatomical cortical hierarchy. This hierarchy begins at V1 and reaches to the highest levels of processing in the temporal and frontal lobes. This hierarchical analysis has been extended via the ratio of supragranular- to infragranular-labeled cells to define quantitative criteria for assigning areas specific hierarchical distances from one another. The pairwise analysis of laminar origins and terminations generally yields clear feedforward–feedback or lateral–lateral relationships between interconnected visual areas in the occipital, temporal, and parietal lobes. However, there are several documented instances, primarily concerning prefrontal and orbitofrontal areas, in which these relationships break down. Whereas this might be expected concerning connections between highly differentiated, six-layer cortices with less differentiated dysgranular or agranular areas, there are also instances of apparent internal inconsistencies within pairs of parietofrontal cortical areas.

It is important to point out that this hierarchical analysis is based solely on laminar features of reciprocal cortical connections and indicates the direction of information flow, but not its timing. This is an important consideration since many cortical areas receive input, not only from the layer immediately below, but often from many layers below. For example, area MT is logically located at the fifth hierarchical level and clearly receives feedforward inputs from V1, V2, V3, and V3A at the first through fourth hierarchical levels, respectively. Therefore, MT is in a position to receive input from cells in layer 4B of V1 before it receives input from V2. Thus, in some way, MT can functionally be considered to be located at the second hierarchical level.

See also: Frontal Cortex Evolution in Primates; Parietal Cortex and Spatial Attention; Prefrontal Cortex: Structure and Anatomy; Vision: Mechanisms of Orientation, Direction and Depth; Visual Cortex: Mapping of Functional Architecture Using Optical Imaging; Visual Cortex in Humans; Visual Cortical Models of Orientation Tuning.

References

Barone P, Batardiere A, Knoblach K, and Kennedy H (2000) Laminar distribution of neurons in extrastriate areas projecting to visual areas V1 and V4 correlates with the hierarchical rank and indicates the operation of a distance rule. *Journal of Neuroscience* 20: 3263–3281.

Brodmann K (1909) *Vergleichende Lokalisationslehre Grosshirnrinde in Ihren Prinzipen Dargestellt auf Grund des Zellenbaues*. Leipzig, Germany: Barth.

Denys K, Vanduffel W, Fize D, et al. (2004) The processing of visual shape in the cerebral cortex of human and nonhuman primates: A functional magnetic resonance imaging study. *Journal of Neuroscience* 24: 2551–2565.

Felleman DJ and VanEssen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1: 1–47.

Lewis JW and VanEssen DC (2000) Mapping of architectonic subdivisions in the macaque monkey, with emphasis on parieto-occipital cortex. *Journal of Comparative Neurology* 428: 79–111.

Lewis JW and Van Essen DC (2000) Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *Journal of Comparative Neurology* 428: 112–137.

Nelissen K, Vanduffel W, and Orban GA (2006) Charting the lower superior temporal region, a new motion-sensitive region in monkey superior temporal sulcus. *Journal of Neuroscience* 26: 5929–5947.

Rizzolatti G and Matelli M (2003) Two different streams form the dorsal visual system: Anatomy and functions. *Experimental Brain Research* 153: 146–157.

Sereno ME, Trinath T, Augath M, and Logothetis NK (2002) Three-dimensional shape representation in monkey cortex. *Neuron* 33: 635–652.

Ungerleider LG and Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA, and Mansfield RJQ (eds.) *Analysis of Visual Behavior*, pp. 549–586. Cambridge, MA: MIT Press.

Van Essen DC and Zeki SM (1978) The topographic organization of rhesus monkey prestriate cortex. *Journal of Physiology* 277: 193–226.

Von Bonin G and Bailey P (1947) *The Neocortex of Macaca mulatta*. Urbana, IL: University of Illinois.

Visually Guided Behavior

J Zeil, The Australian National University, Canberra, ACT, Australia
N Boeddeker, Universität Bielefeld, Bielefeld, Germany
J M Hemmi, The Australian National University, Canberra, ACT, Australia

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A fundamental property of vision is that it operates in closed loop. Behavior changes the pattern of light on the retina and as a result, most image motion experienced by mobile animals is generated by self-motion. In many ways, the closed-loop nature of vision is not only a nuisance, but it also creates opportunities. The main problem with self-induced image motion is that it produces motion blur, in the same way as a photograph becomes blurred by camera shake. Because photoreceptors have a finite integration time, a moving image becomes 'smeared' across the retina, which reduces the ability to resolve fine spatial detail. On the other hand, the spatially and temporally correlated way in which light patterns move across the retina carries information on an animal's own movements, on the depth structure of the world, and on the movement of objects. However, accessing this information and extracting its different contents is not trivial. The similarities in the way in which animals, including humans, move their eyes reflect the universal need to reduce motion blur and to facilitate the analysis of self-induced optic flow. In addition, many animals move in such a way as to generate specific visual information that would otherwise not be available, as in the case of image motion-based distance cues. Behavior thus plays an important role in visual information processing and, as documented later, some basic aspects of behavioral organization are determined by the needs of vision.

Optic Flow

As an animal moves through the world, it creates a distinct pattern of image motion on its retina, called optic flow. The specific pattern of optic flow is determined by both the layout of the environment and the animal's own movements. During a purely linear (translational) movement, the flow pattern expands from one pole of the flow field (focus of expansion), which lies in the direction of heading and contracts toward the second pole in the opposite direction (the focus of contraction) (Figure 1(a)). The image motion in the rest of the visual field flows backward across the retina from the focus of expansion to the focus of contraction. The size of local flow vectors depends on their position relative to the direction of heading and on the distance of objects in the environment. The images of close objects move faster across the retina than those of more distant objects; the closer objects are to the poles of the flow field, the less image motion they produce (Figure 1(b)). A pure translational flow field thus contains information on the direction of heading and on the spatial layout of objects in the world. In contrast, a pure rotational movement generates image motion that revolves around the axis of rotation, in directions opposite to the direction of rotation (Figure 1(a)). The length of local motion vectors varies predictably across the visual field in relation to the axis of rotation, but it is independent of the distance of objects. The rotational flow field thus only contains information on the rotational self-motion component of an animal's movement. When an animal rotates and translates at the same time, both flow field components superimpose and the extraction of their respective information content becomes much more difficult because rotational optic flow tends to dominate the resulting flow field (Figures 1(c)–1(h)). To extract translational and rotational optic flow information separately from such superimposed flow fields is not a trivial task, even if the local flow vectors could be accurately determined, which is not the case. Local image motion needs to be computed from the spatially and temporally correlated changes in light intensity in neighboring receptors. The mechanisms of local motion computation which have been identified in biological systems do not necessarily represent the direction and the speed of image motion in a veridical (accurate) manner. Local motion detectors can measure only the component of motion perpendicular to a contour that extends beyond its field of view and thus suffer from the so-called aperture problem. In addition, their output depends on scene contrast and pattern properties, and their directional tuning can be very broad. As a result, the motion signal distributions generated by the output of local motion detector arrays under natural, low-contrast, and visually cluttered conditions can be noisy, sparse, and patchy, making it difficult to extract useful information from optic flow (Figures 1(e)–1(h)).

Controlling Optic Flow

Keeping Still

One way to cope with the negative impact of image motion on visual processing is to keep still. Animals using vision have powerful visual and nonvisual

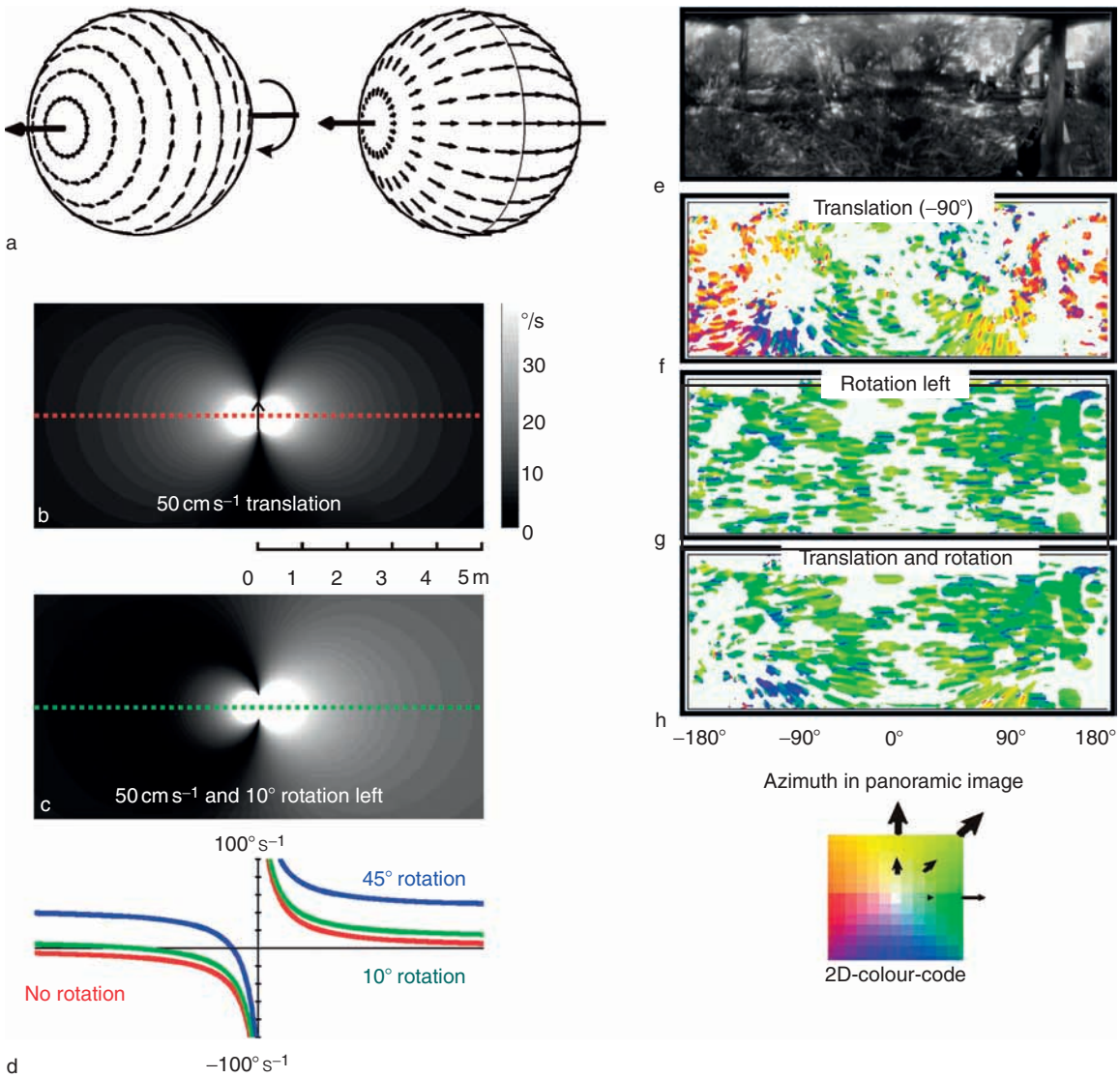

Figure 1 (a) Schematic diagram of rotational optic flow around the roll axis and translational optic flow along the same axis. Local flow vectors are projected on the spherical visual field under the assumption that all objects are at equal distance from the observer. (b) The image motion field produced by a 50 cm translation (top), mapped onto the plane. Bright parts of the map indicate the location of objects that generate high image velocities in the visual field of the moving observer, and black parts indicate the location of objects that produce low image velocities (see color bar on right). The dotted red line indicates the position of the red transect in (d). (c) Same as for (b), but for a 50 cm translation combined with a 10° rotation to the left. The green dotted line indicates the position of the green transect in (d). (d) Horizontal transects through image velocity maps in (b) and (d). The blue curve shows the situation for a 50-cm translation combined with a 45° rotation to the left. Image velocities are calculated with the assumption that both rotation and translation are carried out in 1 s. (e) Panoramic image recorded in an outdoor scene with the aid of a reflective cone. (f–h) Local motion signal maps for different kinds of movements through the scene shown in e. (f) Translation along a 1 m-long path in the azimuth direction −90°. (g) 45° rotation to the left around the vertical axis. (h) Combined 1 m translation and 45° rotation. Motion signals were calculated using panoramic image sequences as input to an elementary, correlation-type motion detector network. Local motion signals are color-coded depending on their direction (see inset), with hue coding their strength. Motion signals were calculated with a time constant of 80 ms and a sampling base of 3°, and they were averaged over 32 frames. They are plotted in full-color saturation for all signals above 2% of the maximum response. (a) Modified from Karmeier K, Krapp HG, and Egelhaaf M (2003) Robustness of the tuning of fly visual interneurons to rotatory optic flow. *Journal of Neurophysiology* 90: 1626–1634, with permission. (e–h) Modified from Zanker JM and Zeil J (2005) Movement-induced motion signal distributions in outdoor scenes. *Network: Computation in Neural Systems* 16: 357–376, with permission from Taylor & Francis, Ltd. (http://www.tandf.co.uk/journals).

reflexes, which serve to compensate for global image motion to produce a stationary retinal image. However, only a few animals can truly stabilize the retinal image as they actively or passively move. When walking or landing, birds employ elaborate leg, body, and neck movements to keep the head position in space still for extended periods of time; these movements are interrupted by rapid linear head movements into the direction of travel ('head bobbing') (Figures 2(a)–2(c)). Birds sitting on moving perches, like wind-driven branches, also keep their head still by leg and neck movements. Head stabilization in birds thus

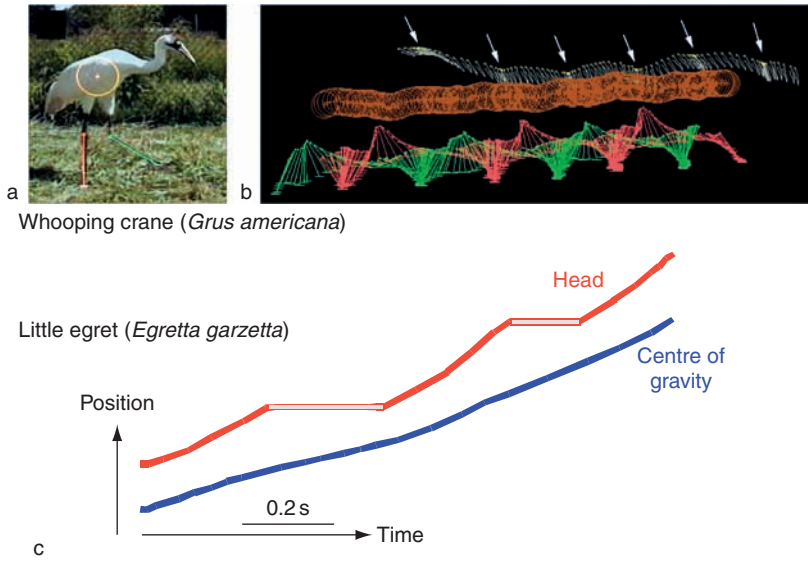

Figure 2 Head stabilization in birds. Walking whooping cranes (a) keep their heads perfectly still in space, depending on walking speed, between 30% and 80% of the time. This is indicated by arrows in (b), which shows the superimposed successive positions in space of the eye (yellow), the head (blue), the body (brown), and the two legs (green and red). The head is repeatedly advanced by a rapid linear movement into the direction of movement. (c) Little egrets show similar head stabilization during walking. (a, b) Reprinted from Cronin TW, Kinloch MR, and Olsen GH (2006) Head-bobbing behavior in foraging whooping cranes favors visual fixation. *Current Biology* 15: R243–R244, with permission. (c) Modified from Fujita M (2003) Head bobbing and the body movement of little egrets. *Journal of Comparative Physiology A* 189: 53–58, with kind permission from Springer-Verlag.

minimizes image motion caused by both rotational and translational self-motion in a similar way as head and body movements do in hovering animals, such as hummingbirds, hoverflies, hawkmoths, and stingless bees.

Clearly, there are limits to how long an animal can keep still. Animals need to view the world around them and thus have to move their heads or eyes. They do so by shifting their gaze in a saccadic manner. Between these fast gaze changes, gaze is kept constant so that self-induced optic flow is compressed into very short time windows. For most of the time, therefore, the retinal image is still, a condition which allows the animal to exploit the best resolution and sensitivity of its eyes to detect moving objects, for instance. Even birds, however, like most other animals, have to tolerate some image motion when they move fast, and the best they can do is to employ eye, head, and body movements to keep rotational and translational optic flow components separate.

Compensatory Eye, Head, and Body Movements

Mobile animals cannot fully stabilize the retinal image, as is often misleadingly stated. However, compensatory eye, head, or body movements can minimize rotational optic flow. These compensatory eye movements are driven only in part by visual feedback loops. Closed-loop visual control systems are relatively slow because the reaction kinetics of phototransduction and the limited conduction velocities of neurons introduce time lags that need to be balanced with the overall gain to avoid instabilities. Therefore, in addition to visual feedback, compensatory eye movements in most animals receive open-loop (feed-forward), multimodal input involving the vestibular system in vertebrates; leg proprioceptors and statocysts in crabs (**Figure 3(A)**); gyroscopes (halteres) in Strepsiptera; and leg proprioceptors, neck proprioceptors, and halteres in flies (**Figure 3 (B)**). As a result, insect head orientation in free flight is stabilized around all three rotational axes most of the time except for fast, saccadic changes in gaze direction.

The functional significance of these compensatory eye movements is not only to minimize the rotational components of optic flow but also to control the absolute orientation of the visual system in space. Most animals keep their eyes horizontally aligned (e.g., flies), whereas a few align them with the local visual horizon (e.g., ghost crabs) (**Figure 3(C)**). Such alignment is a prerequisite for many neural computations that rely on some useful geometric properties of the visual world. For instance, in a flat environment, humans, toads, and crabs infer depth from retinal elevation, which requires the local visual horizon as a reference. Many insects possess specialized areas of dorsal retina which allow them to use the pattern of polarized skylight as a compass. The need for pattern and scene recognition may be a general reason why it is important to keep the visual system aligned with a geocentric frame of reference.

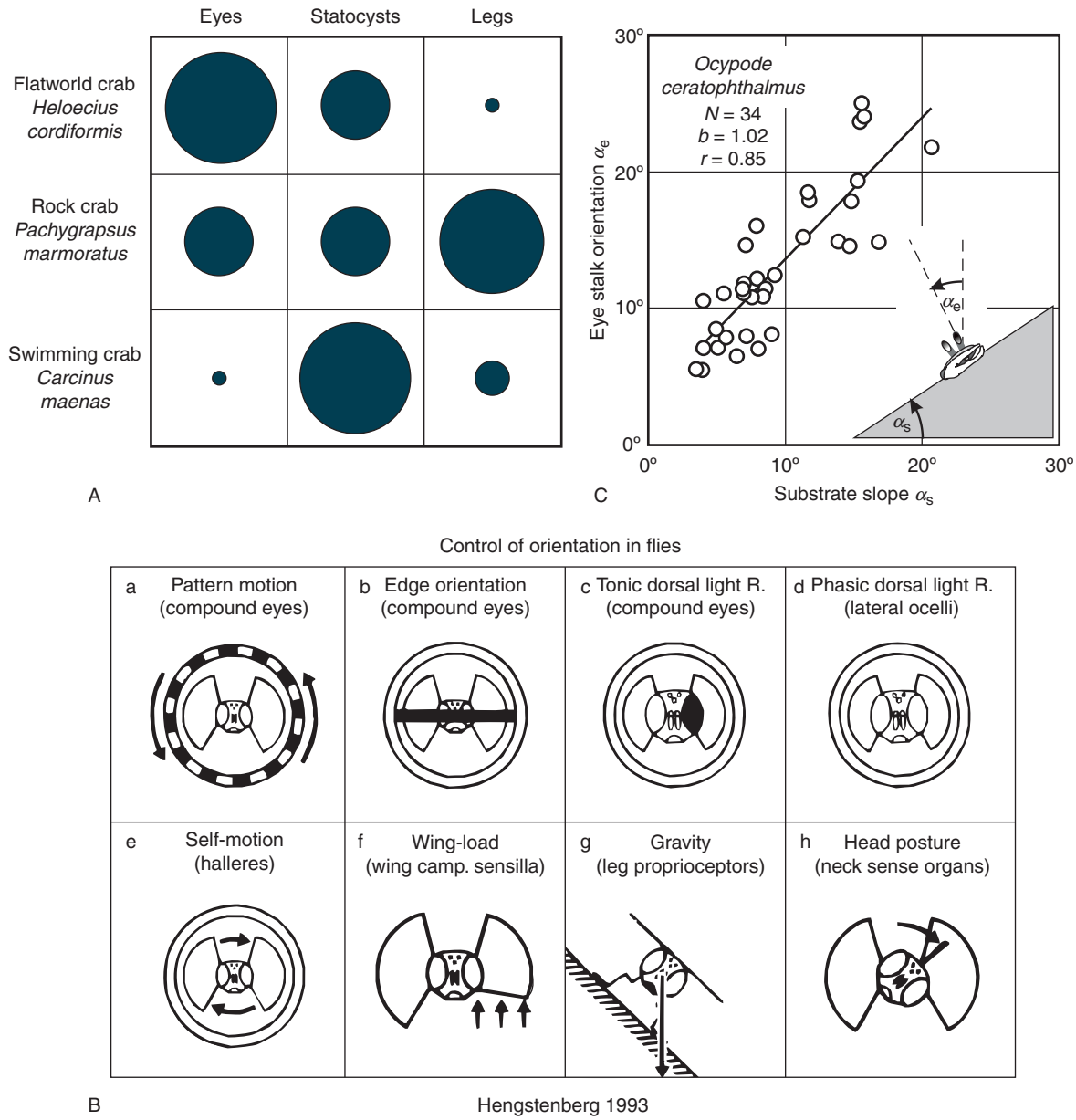

Figure 3 Control of eye movements in crabs and flies. (A) The efficacy of different sensory inputs in driving the compensatory eye movements of different species of crabs. Eye movement control is dominated by visual input in crabs that live in mudflat habitats with a visually salient line of horizon. Eye orientation in rock crabs is mainly driven by input from the legs and in swimming crabs by input from the statocysts. Crabs were stimulated by 1-Hz sinusoidal roll movements with 20° amplitude; the area of the circles shown is proportional to the amplitude of compensatory eye stalk movements, relative to maximal stimulus amplitude. (B) The eight different sensory subsystems determining head roll orientation in *Calliphora* flies. The schematic diagrams show the sense organs used by the fly to keep the head aligned with the vertical and the experimental conditions which have allowed identifying the visual (a–d) and mechanosensory cues (e–h) involved. (C) Ghost crabs (*Ocypode ceratophthalmus*) which live in the flat world of tropical sandy ocean beaches orient their eye stalks perpendicular to the local visual horizon when operating on inclined sections of the beach. This aligns the horizontal equatorial acute zone of their eyes with the local visual horizon. (A) Modified from Nalbach H-O (1990) Multisensory control of eyestalk orientation in decapod crustaceans: An ecological approach. *Journal of Crustacean Biology* 10: 382–399, with permission. (B) Reproduced from Hengstenberg R (1993) Visual stabilization in arthropods. In: Miles FA and Wallman J (eds.) *Visual Motion and its Role in the Stabilization of Gaze*, pp. 285–298. Amsterdam: Elsevier, with permission. (C) Modified from Zeil J (1990) Substratum slope and the alignment of acute zones in semi-terrestrial crabs (*Ocypode ceratophthalmus*). *Journal of Experimental Biology* 152: 573–576, with permission of The Company of Biologists.

Depending on the mobility of their head and eyes, animals minimize the rotational components of optic flow in different ways. For example, waterstriders are unusual in that they use leg movements to compensate for drift and unwanted rotations on the water surface. They compensate for drift with coordinated strokes of their middle legs, using a discontinuous position servo, and minimize rotations by

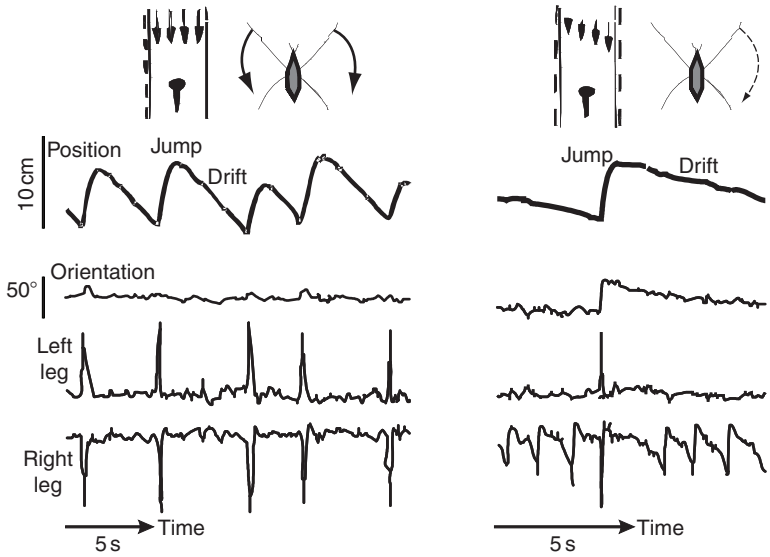

Figure 4 Separation of rotational and translational optic flow by waterstriders (*Gerris palludum*). Waterstriders compensate for drift on a moving water surface by periodic jumps against the direction of drift (left). The jumps are executed by a coordinated rapid rowing movement of the two middle legs (lower traces). (Right) When waterstriders are subjected to additional rotation on a water surface with slanted laminar flow, they compensate separately for rotation by smooth rowing movements of one leg only (bottom) and occasionally for drift by a coordinated saccadic rowing movement with both middle legs. Modified from Junger W and Dahmen H-J (1991) Response to self-motion in waterstriders: Visual discrimination between rotation and translation. *Journal of Comparative Physiology A* 169: 641–646, with kind permission of Springer-Verlag.

continuous unilateral rowing movements of one middle leg only (Figure 4).

Saccadic Locomotion

Compensatory eye, head, and body movements thus serve to minimize changes in gaze direction relative to an external (geocentric) coordinate system. This essentially eliminates rotational optic flow. In addition, many animals, especially insects, control the type of optic flow they experience even further. They move in a way that avoids smooth or frequent changes in the direction of heading, which would cause the poles of the translational flow field to wander across the retina. Hoverflies, for instance, move in a series of straight lines, the direction of which can be independent of their body axis (Figure 5(a)). Even insects that do not have a similar degree of hovering control, such as the blowfly *Calliphora* (Figure 5(b)), the housefly *Fannia canicularis* (Figure 5(c)), and the fruitfly *Drosophila melanogaster* (Figure 5(d)), move in such a saccadic manner during territorial and exploration flights, whereby compared to hoverflies, their longitudinal body axis tends to be much more aligned with flight direction. Compensatory eye movements together with saccadic locomotion along straight paths thus maximize the time animals experience a pure translational flow field with a constant orientation in the visual field. The pole of expansion and the direction of gaze are then changed together during the saccadic change of flight direction. There are some indications that this behavioral organization indeed supports the extraction of useful information from optic flow. The compound eyes of many flying insects have forward-pointing acute zones with high spatial resolution, which decreases systematically toward the back of the eyes. These gradients of spatial resolution are thought to have evolved to reduce motion blur associated with the optic flow field encountered during forward flight, when retinal image motion is minimal in the direction of heading and largest in the lateral visual field especially when vegetation is close by. Because image contrast is inversely related to the ratio between receptor acceptance angle and receptor integration time, larger receptor spacing and larger acceptance angles in the lateral field of view allow to resolve details at higher angular velocities.

Using Optic Flow

Animals use the translational component of optic flow for a variety of tasks, and they can do so because it has high information content with respect to the layout and the distance of objects in the world. In blowflies, for instance, the behavioral separation of translational from rotational optic flow has been shown to be essential for the visual system to extract distance information. Similarly, when an animal approaches an object, the retinal image of this object expands. Flies have been shown to use such looming cues to initiate deceleration and to extend their legs in preparation for landing.

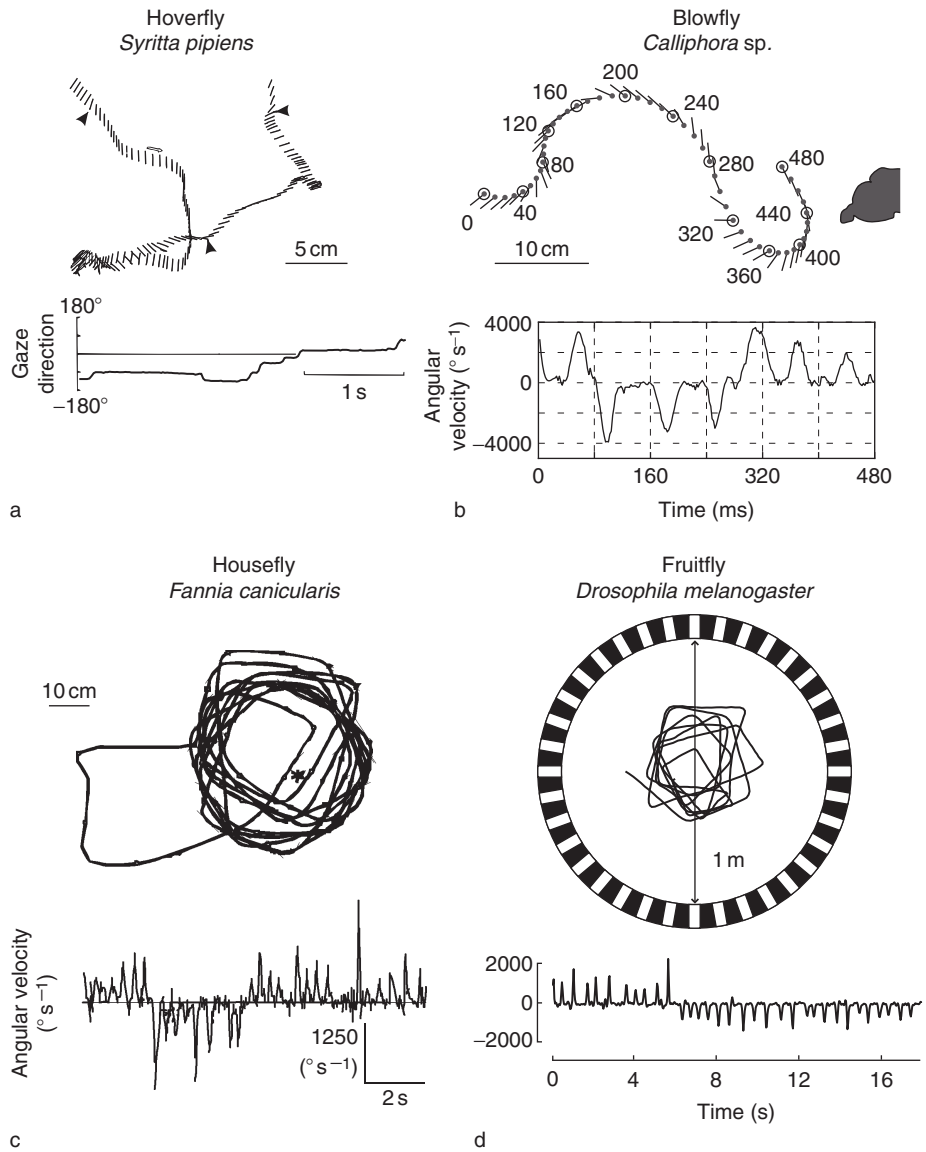

Figure 5 Saccadic gaze changes and saccadic locomotion in flies. (a) Flight path of a cruising hoverfly (*Syritta pipiens*). The head of the fly is marked by a small dot and the orientation of the longitudinal body axis with a line attached to it. Positions are shown every 20 ms. Arrowheads mark saccadic gaze changes. Bottom diagram shows gaze direction over time. (b) Flight path of a blowfly (*Calliphora* sp.). Positions are shown every 8 ms and labeled every 40 ms. Bottom diagram shows changes in body axis orientation over time. (c) Flight path showing saccadic locomotion of a male housefly (*Fannia canicularis*) patrolling the airspace beneath an indoor landmark (star). Bottom diagram shows changes in flight direction over time. (d) Saccadic flight path of a fruitfly (*Drosophila melanogaster*) cruising inside a large patterned drum. Bottom diagram shows changes in flight direction over time for the same sequence. (a) Modified from Collett TS and Land MF (1975) Visual control of flight behaviour in the hoverfly, *Syritta pipiens* L. *Journal of Comparative Physiology* 99: 1–66, with kind permission of Springer. (b) Modified from Boeddeker N, Lindemann JP, Egelhaaf M, and Zeil J (2005) Responses of blowfly motion-sensitive neurons to reconstructed optic flow along outdoor flight paths. *Journal of Comparative Physiology A* 191: 1143–1155, with kind permission of Springer. (c) Modified from Zeil J (1986) The territorial flight of male houseflies (*Fannia canicularis*). *Behavioural Ecology and Sociobiology* 19: 213–219, with kind permission of Springer-Verlag. (d) Modified from Tammero LF and Dickinson MH (2002) The influence of visual landscapes on the free flight behaviour of the fruit fly, *Drosophila melanogaster*. *Journal of Experimental Biology* 205: 327–343, with permission of The Company of Biologists.

Landing is essentially controlled by spatiotemporal integration of looming optic flow patterns, which may also be used for time-to-impact judgments, obstacle avoidance, and predator avoidance. Diving birds, for instance, judge the time of their impact with the water surface by the rate of expansion of its surface texture; hovering stingless guard bees (*Tetragonisca angustula*) and hawkmoths interpret expanding and contracting image motion as indicating unintended displacements in space.

Many other elements of flight control in insects involve the use of translational flow field components. When landing on flat horizontal surfaces, bees keep ventral optic flow constant and thus regulate their flight speed to be roughly proportional to their height above ground, which guarantees smooth landing without requiring information about the exact height above a landing site. The same mechanism allows bees to adjust their flight speed depending on the space they fly through: when a tunnel they negotiate narrows, they slow down, and they speed up again when it widens. In flight tunnels, wind tunnels, and in tethered flight, flies, locusts, and bees control their flight speed by keeping the average image motion they perceive at relatively constant values. In the case of hovering, image motion is clamped to zero. Hovering insects, such as hoverflies, hummingbirds, hawkmoths, and stingless bees, hold their position in space by minimizing translational optic flow through compensatory translational movements. By balancing image motion on either side when flying through gaps or narrow tunnels, honeybees keep clear of obstacles and to the center of negotiable airspace. Honeybees and stingless bees are also able to measure the distance they have flown by integrating the image motion they generate during flight.

Honeybees and flies use relative motion cues to detect objects by foreground–background segregation and to infer their distance relative to the background. For a moving observer, a nearby object appears to move faster than the background, which allows both bees and blowflies to discriminate between objects at different heights above ground. Bees identify landing sites or target objects with the aid of motion cues and approach the edges of elevated objects in such a way as to maximize motion parallax. Ladybird beetles move along straight-line segments (saccadic locomotion) and use the motion parallax information gained during these translations to identify close landmark which they then approach and climb.

Shaping Optic Flow

A number of animals, including birds, lizards, insects, and humans, produce eye, head, and body movements that are not compensatory in nature but are tailored to generate specific patterns of image motion that carry motion parallax information. Only for insects and birds, however, is there clear evidence that the visual information generated by these 'peering' movements is actually used to gauge the distance of objects. The peering movements in insects, birds, and lizards are predominantly highly controlled linear movements of the head, during which gaze direction is held constant, resulting in a purely translational flow field on the retina. In contrast, some wasps and birds generate a pivoting parallax field in response to objects of interest by moving along arcs around the object while fixating it (Figures 6(a)–6(d)). The visual consequences of these pivoting movements in terms of image motion are similar to those when a moving human observer fixates and tracks an object in a viewing direction other than the direction of travel.

Arguably the most elaborate examples of behavioral shaping of optic flow are the learning flights made by bees and wasps in the context of landmark guidance when leaving their nests for the first time or when leaving a newly discovered food source. Insects are known to navigate with the aid of remembered views. They are guided by landmarks along routes and when pinpointing places of interest, such as nests and feeding sites: they remember the apparent size of landmarks, their brightness or color, and also their absolute distance from the goal. It is presumably this last aspect, namely the need to identify salient objects close to the goal, which requires homing insects to move in specific ways during the acquisition of a visual representation of the goal environment.

During a learning flight, ground-nesting wasps, for instance, turn to face the goal and then back away from it in a series of increasing arcs during which they gain distance from the goal at approximately the same rate as their height above ground increases. While the wasp pivots about the nest location, it keeps the image of the goal area in a frontolateral position in the visual field (Figure 6(e)). High-resolution and high-speed records of these flights show that wasps move their heads in a saccadic manner during learning. Sections of flight in which gaze is held constant are linked by rapid head movements that change gaze (Figure 6(f)). The insects thus do not generate a pivoting parallax field as had been deduced previously from an analysis of their body axis orientation alone but, rather, a sequence of translational parallax fields. Both these fields of behaviorally shaped optic flow would allow the insects to segregate objects from background and to judge the relative and possibly absolute distance of landmarks. What determines the use of one or the other parallax field in different contexts remains to be elucidated.

Optic Flow while Tracking Targets

There are situations in which animals cannot compensate for large-scale rotational image motion. For instance, when following a moving object with our eyes during smooth pursuit, the image of the rest of the visual scene moves in the opposite direction across our retina. Such global rotational image motion

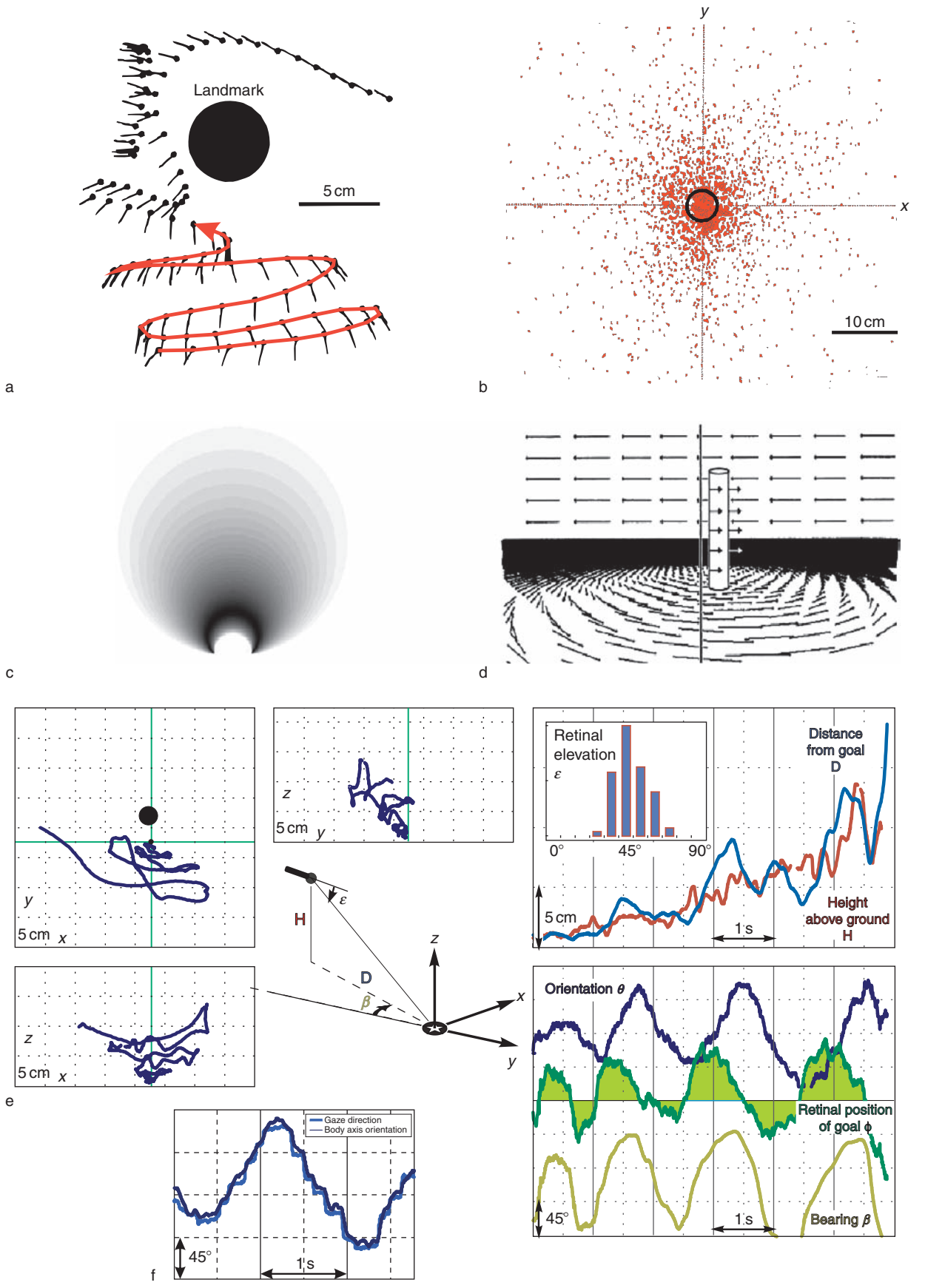

Figure 6 Continued

would normally trigger compensatory eye movements. Because these stabilizing reflexes would counteract target tracking, they are suppressed during human smooth pursuit eye movements. The same problem arises during fast, saccadic changes in gaze, although the self-induced global image motion in this context is easier to cope with because these voluntary gaze changes are preprogrammed, rapid, and ballistic. Their effects, in terms of image motion direction and motion speed, are therefore predictable and can be taken into account by the brain. In vertebrates, there is evidence that this is achieved by saccade-triggering neurons sending corollary signals to inhibit motion-sensitive neurons during the execution of a rapid gaze change.

Like many other insects, blowflies chasing after dummy targets use a continuous tracking strategy analogous to human smooth pursuit eye movements (Figure 7(a)), which, as in humans, becomes saccadic when the target moves too quickly (Figure 7(b)). It is unknown how stabilizing reflexes and pursuit movements interact in target tracking insects: whether they simply add with different gain, whether an efference copy is needed to prevent compensatory reflexes from interfering with tracking, or whether interference is minimized by the different spatial frequency tuning of tracking and optomotor systems. Some animals do not seem to have a solution to this problem: the tracking strategy of the praying mantis, for instance, depends on the visual background. The mantis tracks insect prey with head movements which keep the insect's frontal acute zone pointing at the target. When the background is visually homogenous, the mantis is able to track a moving target with smooth pursuit head movements (Figure 7(c)). However, tracking becomes saccadic when the target is seen against a visually cluttered background such as vegetation (Figure 7(d)). Unlike the praying mantis, male hoverflies are able to track females by smooth pursuit even against a textured background; however, like humans and blowflies, they do switch to saccadic tracking when the target moves too fast for the smooth pursuit system to keep up. Similarly, tiger beetles are able track high-contrast objects using smooth pursuit, but they pursue live prey in a stop-and-go manner – a strategy which serves to improve the signal-to-noise ratio of the motion signal generated by moving prey. For the stationary beetle, compared to when it is running, the moving prey is easier to detect against a stationary background.

A special case in which background motion during target tracking contains useful information occurs in animals with mobile eyes such as humans, who can fixate stationary objects independently of their own direction of movement at any distance and direction as they walk or drive. If the object is not at infinity and does not lie in the direction of heading, the tracking movement exactly compensates for the local translational optic flow component associated with the object by rotating the eyes and/or the head at the same angular speed. The result is a pivoting parallax field, a combination of a translational and a rotational optic flow field that is characterized by opposite image motion directions associated with objects that are further away or closer than the fixation point. Such a pivoting parallax field thus produces an increased motion contrast between foreground and background objects and therefore has high information content with respect to the relative location of objects in the world.

Vision and the Organization of Behavior

Because of the intimate interdependence of vision and behavior, behavior plays an important role in visual information processing. It can simplify computations by separating rotational and translational optic flow components, and it can generate visual information, as in the case of motion parallax. However, behaviorally guided vision can take more complex forms, as in the case of predator avoidance strategies, in which distinct sequences of prey behaviors are driven by the need to gather different kinds of visual information on the

Figure 6 Active vision strategies of homing wasps. (a) Ground-nesting wasps (*Odynerus spinipes*) respond to novel objects in the vicinity of their nests with rapidly swaying zigzag flight maneuvers, with gaze centered on the object, as shown in the scatter diagram of consecutively intersecting gaze directions (b), which generates a pivoting parallax field on the retina of the insect (c, d). (e) The organization of learning flights in ground-nesting *Cerceris* wasps. The three-dimensional flight path of a learning flight is shown from different directions in the three panels on the left (see inset for definition of axes). The time course of distance from goal, height above ground, orientation, bearing, and retinal azimuth position of the nest entrance are shown in the two panels on the right (for definitions of variables, see inset). Because the wasp gains distance from the goal and height above ground at approximately the same rate, the nest entrance is always seen approximately 45° below the horizon (see inset histogram). Moving along arcs around the nest entrance, the wasp counterrotates in such a way as to keep the nest entrance at frontolateral retinal azimuth positions (green trace in bottom right panel). (f) High-speed, high-resolution digital video recordings reveal that head movements play an important role in learning flights: gaze changes are not smooth but, rather, saccadic as a wasp flies along an arc so that gaze is stabilized for most of the time. (a–d) Modified from Voss R and Zeil J (1998) Active vision in insects: An analysis of object-directed zig-zag flights in a ground-nesting wasp (*Odynerus spinipes*, Eumenidae). *Journal of Comparative Physiology A* 182: 377–387, with kind permission of Springer-Verlag.

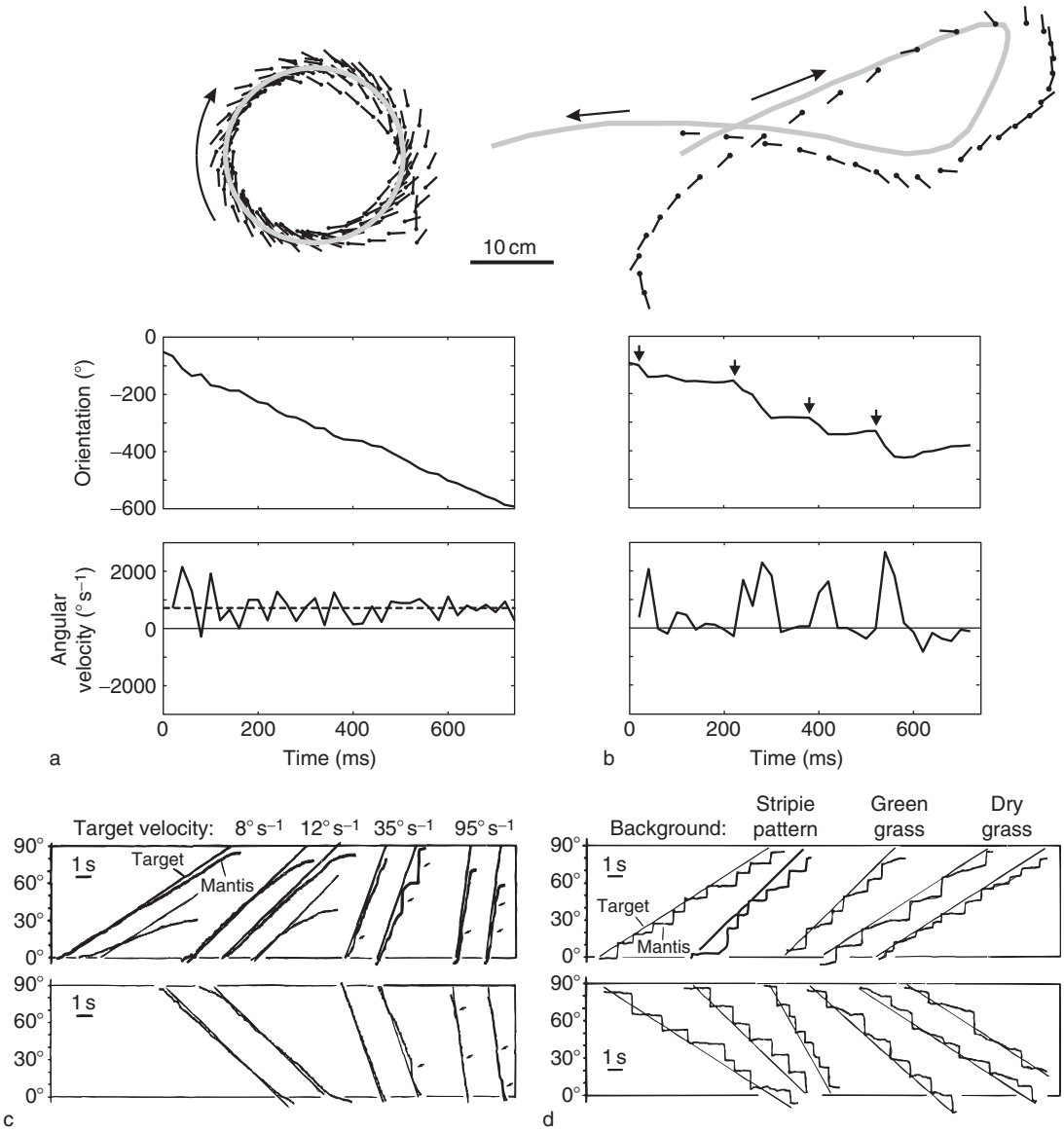

Figure 7 Target-tracking behavior in insects. Male blowflies perform aerobatic pursuits belonging to the fastest visual behaviors that can be observed in nature. The tracking strategies of blowflies appear to be very similar to pursuit eye movements in primates: a combination of smooth pursuit with catch-up saccades. (a) Top view of a flight trajectory of a fly chasing a black sphere that moves on a circular track in a horizontal plane (gray line). The head position (dot) and the longitudinal body axis orientation (line) of the fly are shown every 20 ms. The time course of orientation and angular velocity of the fly are shown for the first 740 ms of the sequence in the panels below. The angular path velocity of the dummy target of approximately 716° s^{-1} is indicated by the dashed line. (b) Flight path of a fly chasing another fly in top view. Conventions are the same as in a. The time course of orientation and of angular velocity is plotted in the lower panels. Note saccadic changes of gaze direction, marked by arrows in the orientation over time diagram. (c) Praying mantis track moving targets with smooth head movements, provided the background is visually homogenous. (d) Praying mantis tracking head movements become saccadic if the target is seen against various structured backgrounds. (a, b) Modified from Boeddeker N and Egelhaaf M (2005) A single control system for smooth and saccade-like pursuit in blowflies. *Journal of Experimental Biology* 208: 1563–1572, with permission of The Company of Biologists. (c, d) Modified from Rossel S (1980) Foveal fixation and tracking in the praying mantis. *Journal of Comparative Physiology* 139: 307–331, with kind permission of Springer-Verlag.

actual threat posed by an approaching predator. In these situations, behavior does not directly influence visual information, as in the cases discussed previously, but serves to prepare the conditions in which relevant information can be gathered in a risk-free way.

Many visually guided prey animals react to predators in a distinct sequence of behavioral acts: freeze, run for cover, monitor (vigilance), hide, and monitor again on reemergence from a refuge. This organization of predator avoidance can be interpreted in terms of risk, in terms of lost opportunity, and in terms of

visual information. The point to note in the present context is that animals freeze when first becoming aware of a predator, at a predator–prey distance where visual information is insufficient to identify a predator or determine its distance and direction of movement. Freezing thus basically serves to stabilize the retinal image, minimize blur, and increase the signal-to-noise ratio. At this stage, visual information is not well correlated with actual threat. More robust visual information can be gathered once the animal is in relative safety – for example, close to a refuge, where a predator has little chance of intercepting and catching it. At this stage then, prey animals can afford to allow a predator to come closer and to determine its direction of approach by detecting looming cues or its identity by analyzing shape, color, or motion parameters. These visual signatures of a predator have a high correlation with actual threat and thus allow the prey animal to make informed decisions regarding two crucial elements of its antipredator defense: when and for how long to hide.

Such examples of behavior supporting visual information processing have been analyzed in great detail in fiddler crabs, which are quite unique because their full behavioral repertoire and the visual cues they attend to can be continuously monitored in the field. The crabs live in intertidal mudflats and operate from burrows which provide protection from predators, access to water and food, and which need to be defended against other crabs. As the crabs move away from their refuge, they quickly lose sight of it because the burrow's opening disappears in surface undulations. The crabs therefore need to rely on their path integration system, which monitors the direction and distance in which they move so that they can keep track of their distance from the burrow and in which direction they need go to return to the burrow. The animals use this information to align their longitudinal body axis with the home direction, which is also the orientation that coincides with their fastest mode of locomotion (sideways). However, it also allows crabs to visually determine the distance between an intruding conspecific and their own burrow. The crabs do this by aligning the retinal position of the burrow with the zero position of a retinotopic distance map, which then maps retinal location to distance from that zero point. Such a mapping relies on compensatory eye movements to align the retina with the external world and can only work reliably in a flat world, such as the intertidal mud and sand flats which fiddler crabs inhabit, in which retinal elevation is a robust cue to depth. The resulting distance information is crucial for a crab to decide whether or not another crab has come too close to its home burrow.

There are likely many other examples in which behavior can be shown to be organized in such a way as to serve visual information processing. Visual search, vigilance behavior, and the attraction to observation posts can be interpreted in this way, as can the calibration of the skylight polarization compass by desert ants, which turn on the spot before leaving on their foraging trips.

Most animals depend heavily on vision for behavioral guidance. However, as this brief review has emphasized, behavior in turn is often organized to support visual information processing by minimizing the problems created and by maximizing the opportunities offered by the visual closed loop.

See also: Attention and Eye Movements; Cortical Control of Eye Movements; Eye and Head Movements; Perception and Eye Movements; Pursuit Eye Movements; Vision for Action and Perception; Vision: Mechanisms of Orientation, Direction and Depth; Visual Motion Detection.

Further Reading

Boeddeker N and Egelhaaf M (2005) A single control system for smooth and saccade-like pursuit in blowflies. *Journal of Experimental Biology* 28: 1563–1572.

Boeddeker N, Lindemann JP, Egelhaaf M, and Zeil J (2005) Responses of blowfly motion-sensitive neurons to reconstructed optic flow along outdoor flight paths. *Journal of Comparative Physiology A* 191: 1143–1155.

Collett TS and Zeil J (1996) Flights of learning. *Current Directions in Psychological Sciences* 5: 149–155.

Collett TS and Zeil J (1998) Places and landmarks: An arthropod perspective. In: Healy S (ed.) *Spatial Representation in Animals*, pp. 18–53. Oxford: Oxford University Press.

Cornilleau-Pérès V and Gielen CCAM (1996) Interactions between self-motion and depth perception in the processing of optic flow. *Trends in Neurosciences* 19: 196–202.

Cronin TW, Kinloch MR, and Olsen GH (2006) Head-bobbing behavior in foraging whooping cranes favors visual fixation. *Current Biology* 15: R243–R244.

Fujita M (2003) Head bobbing and the body movement of little egrets. *Journal of Comparative Physiology A* 189: 53–58.

Gibson JJ (1950) *The Perception of the Visual World*. Cambridge, MA: Riverside Press.

Hemmi JM and Zeil J (2005) Animals as prey: Sensory–motor abilities and flexibility of behaviour in an arthropod. *Marine Ecology Progress Series* 287: 274–278.

Hengstenberg R (1993) Visual stabilization in arthropods. In: Miles FA and Wallman J (eds.) *Visual Motion and its Role in the Stabilization of Gaze*, pp. 285–298.

Junger W and Dahmen HJ (1991) Response to self-motion in water-striders: Visual discrimination between rotation and translation. *Journal of Comparative Physiology A* 169: 641–646.

Kern R, van Hateren JH, and Egelhaaf M (2006) Representation of behaviourally relevant information by blowfly motion-sensitive visual interneurons requires precise compensatory head movements. *Journal of Experimental Biology* 209: 1251–1260.

Koenderink JJ (1986) Optic flow. *Vision Research* 26: 161–180.

Koenderink JJ and VanDoorn AJ (1987) Facts on optic flow. *Biological Cybernetics* 56: 247–254.

Kral K (2003) Behavioural–analytical studies of the role of head movements in depth perception in insects, birds and mammals. *Behavioural Processes* 64: 1–12.

Kramer DL and McLaughlin RL (2001) The behavioural ecology of intermittent locomotion. *American Zoologist* 41: 137–153.

Land MF (1992) Visual tracking and pursuit: Humans and arthropods compared. *Journal of Insect Physiology* 38: 939–951.

Land MF (1999) Compound eye structure: Matching eye to environment. In: Djamgoz MBA, Loew ER, Partridge JC, and Vallerga S (eds.) *Adaptive Mechanisms in the Ecology of Vision*, pp. 51–71. Dordrecht: Kluwer Academic.

Land MF (1999) Motion and vision: Why animals move their eyes. *Journal of Comparative Physiology A* 185: 341–352.

Land MF and Collett TS (1997) A survey of active vision in invertebrates. In: Srinivasan MV and Venkatesh S (eds.) *From Living Eyes to Seeing Machines*, pp. 16–36. Oxford: Oxford University Press.

Lappe M (ed.) (2000) *Neuronal Processing of Optic Flow*. San Diego: Academic Press.

Lindner A and Ilg UJ (2006) Suppression of optokinesis during smooth pursuit eye movements revisited: The role of extra-retinal information. *Vision Research* 46: 761–767.

Miles FA and Wallman J (eds.) (1993) *Visual Motion and Its Role in the Stabilization of Gaze*. Amsterdam: Elsevier.

Nalbach H-O (1990) Multisensory control of eyestalk orientation in decapod crustaceans: An ecological approach. *Journal of Crustacean Biology* 10: 382–399.

Rossel S (1980) Foveal fixation and tracking in the praying mantis. *Journal of Comparative Physiology* 139: 307–331.

Srinivasan MV and Zhang S (2004) Visual motor computations in insects. *Annual Review of Neuroscience* 27: 679–696.

Tammero LF and Dickinson MH (2002) The influence of visual landscapes on the free flight behaviour of the fruit fly, *Drosophila melanogaster*. *Journal of Experimental Biology* 205: 327–343.

Venkatesh S and Srinivasan MV (eds.) (1997) *From Living Eyes to Seeing Machines*. Oxford: Oxford University Press.

Voss R and Zeil J (1998) Active vision in insects: An analysis of object-directed zig-zag flights in a ground-nesting wasp (*Odynerus spinipes*, Eumenidae). *Journal of Comparative Physiology A* 182: 377–387.

Wexler M and van Boxtel JJA (2005) Depth perception by the active observer. *Trends in Cognitive Sciences* 9: 431–438.

Zanker JM and Zeil J (eds.) (2001) *Motion Vision: Computational, Neural and Ecological Constraints*. New York: Springer.

Zanker JM and Zeil J (2005) Movement-induced motion signal distributions in outdoor scenes. *Network: Computation in Neural Systems* 16: 357–376.

Zeil J (1990) Substratum slope and the alignment of acute zones in semi-terrestrial crabs (*Ocypode ceratophthalmus*). *Journal of Experimental Biology* 152: 573–576.

Visual–Vestibular Interactions

L R Harris, York University, Toronto, ON, Canada

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The visual and vestibular systems together form a consortium that informs an organism about its motion and orientation relative to gravity. The visual and vestibular contributions are complementary, with each system tending to make up for inadequacies and ambiguities in the other. Their combined information is used for three main purposes: eye movement control, posture control, and perception. Other systems, such as the somatosensory system indicating the sites of pressure resulting from the weight of the body against a support surface, and the auditory system, can also inform about the relationship of the body to the world. This article concentrates on the use of visual and vestibular information in humans.

Visually Transduced Information Concerning Self-Motion and Orientation

The visual system provides continuous and usually unambiguous information about motion and orientation of the body relative to the world. However, visual information is subject to delays, because of the time needed for retinal processing and for combining signals across time and space to derive motion information. Visual information can also sometimes be irrelevant or misleading for deducing motion and orientation, such as when looking at anything other than features of the environment, for example, when watching films or looking at pictures. It is also subject to interruption in low light, during eye blinks, or when the eyes are closed.

Visual information about the movement of a person comes from the systematic displacement of the visual directions of all visible features as a consequence of movement relative to them. This pattern of motion is called 'optic flow.' Optic flow generated by 'translation' (movement from one place to another) contains enough information to determine in which direction a person is 'heading.' The direction of heading during translation can be determined from the location of the 'focus of expansion' of the optic flow, as JJ Gibson told us. When a person moves toward a particular point, the visual direction of that point (and the point in the diametrically opposite direction) does not change relative to the mover. The images of all other points move by an amount that depends on each point's angular distance from the heading direction and its distance from the observer. Because of this systematic arrangement, the location of the focus of expansion can be deduced from even a small area of visible optic flow. The relative distance a person has traveled (e.g., movement halfway to some visible object) can be extracted from optic flow. But determining the absolute distance (e.g., movement of 1 m) requires additional information about the distance of visible objects, since the motion of an object's image on the retina depends not only on the velocity and direction of self-motion but also on these distances. There are therefore high computational demands in reconstructing the self-motion that created a given optic flow pattern. Furthermore, optic flow evokes compensatory movements of the eyes. This complicates the process of determining the underlying optic flow since the retinal motion is affected by both the movement of the eyes in the head and the overall motion of the person. Nonvisual information about compensatory eye movements can often be helpful, and in some cases is indispensable, for distinguishing these components.

Optic flow generated by 'rotation' (movement that does not involve an overall change in position) contains information about the axis of rotation and instantaneous rotational velocity. This also has to be interpreted in the context of evoked eye movements which can remove most of the rotational optic flow from the retinal image. Optic flow can be experienced in the absence of physical motion, for example, when a neighboring car or train moves. Such visual movement can generate a compelling sensation of self-motion called 'vection.' Vection can have both translational ('linear vection') and rotational ('circular vection') components and generally takes several seconds to build up. The onset of vection is speeded by the presence of actual motion even if it is a pulse or a vibration unrelated to the self-motion signaled by the optic flow.

Visual information relevant to orientation of a person relative to gravity includes: (1) the 'visual frame,' comprising the ground plane and the view of features known to be approximately vertical or horizontal relative to gravity, such as trees, walls, ceilings, and floors; (2) the orientation of the 'visual horizon'; (3) the assumption that light generally comes from above; (4) 'support relationships' between objects, determined by the laws of physics, such as objects resting on other objects or supported by the ground; and (5) 'visual polarity' cues of objects with recognizable tops and bottoms such as faces, lamps, and chairs that have a most-familiar orientation relative to gravity. Some of these properties, such as the direction from

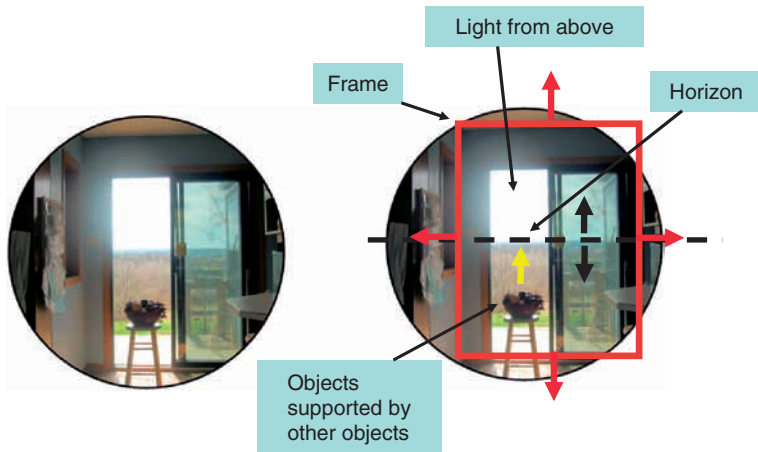

Figure 1 A photograph of a typical scene showing examples of the various visual cues to the direction of gravity. On the right, the possible directions reported by the cues in the photograph are detailed. The four possible directions of gravity signaled by the frame, defined by the orientation of the walls, floor, ceiling, and window stripped of their other cues, are indicated by red arrows. The two indicated by the horizon are shown in black, and the unique direction signaled by the relationships of objects in the world (e.g., the fruit bowl on the stool), the expected orientation of objects (fruit bowl curved side down), and light coming from above is indicated by the yellow arrow.

which light comes or the orientation of the horizon, could have a recognition mechanism with an innate component; others must be learned by association, such as the orientation of writing on advertising signs. Figure 1 illustrates some of the visual cues that are available to specify orientation in a typical scene.

Figure 1 also illustrates that the various visual cues to orientation have greater or lesser degrees of ambiguity. The frame structure of a room, stripped of visual polarity and support relationship cues, indicates four possible directions of gravity; the horizon indicates two possible directions; whereas the direction of illumination, support relationships, and visual polarity cues signal only one unique possibility.

Vestibularly Transduced Information Concerning Self-Motion and Orientation

The 'vestibular system' comprises two sets of mechanical sensors, both of which are activated by physical force. Force, as Isaac Newton told us, is the product of mass and acceleration. Thus, the vestibular system can only detect motion that involves acceleration: movement at a constant speed cannot stimulate the vestibular system. The 'semicircular canals' detect angular acceleration of the head, and the 'otoliths' detect linear acceleration. Gravity, as Albert Einstein told us, is equivalent to an upward linear acceleration of $9.8\,\mathrm{m\,s^{-2}}$, and so the vestibular signal also contains information about the orientation of the head relative to gravity.

Since the vestibular system is sensitive only to acceleration, position and velocity information need to be derived from the accleration. However, such a derivation can only provide relative values of position and velocity for both rotation and translation. For example, the changes in linear velocity (i.e., acceleration) from 10 to $20\,\mathrm{m\,s^{-1}}$; 0 to $10\,\mathrm{m\,s^{-1}}$; –10 to $0\,\mathrm{m\,s^{-1}}$; and –20 to $-10\,\mathrm{m\,s^{-1}}$ are all the same although the final velocities are very different. Before the invention of modern modes of transport, the assumption that the starting velocity was close to 0 was more or less valid, but even so, the need to add assumed starting values makes the vestibular system fundamentally unsuited for tasks such as keeping track of position and orientation during translation and rotations in space. Many disorientation illusions related to aviation can be traced to this limitation.

It might be thought that the constant pull of gravity would provide a reference direction to which the perception of self-motion and orientation could be anchored. However, since gravity cannot be distinguished from linear acceleration, a given signal from the otoliths organs is ambiguous. This signal could be produced by any one of an infinite number of combinations of linear accelerations and tilts of the head as shown in Figure 2. Since the accelerations of everyday life are not generally confused with tilt, the gravity vector must somehow be subtracted from the signal. The direction of the gravity vector can be obtained from the visual cues described above.

However, solving the geometrical problem of the double vector requires knowledge not only of the direction of gravity, but also of its magnitude. If its magnitude is not available, then, when acceleration with a component orthogonal to gravity is experienced, it is hard to distinguish the resultant vector from gravity alone after an appropriate tilt of the head. This is the 'tilt-translation' ambiguity. The two eventualities

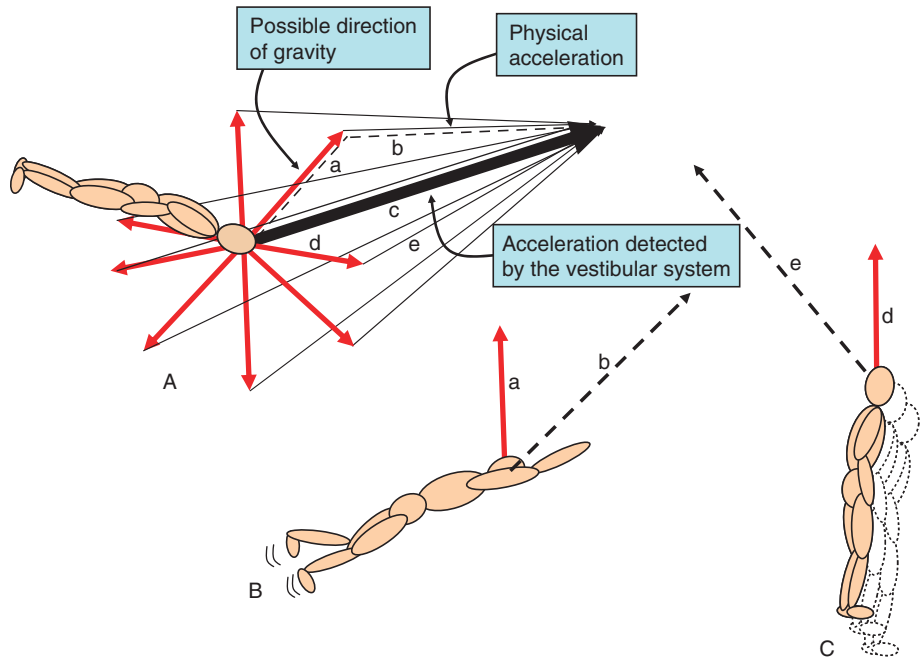

Figure 2 (A) The acceleration vector transduced by the vestibular system (thick black arrow) is the resultant of gravity (red arrows) and physical acceleration (thin black arrows). The lengths of the arrows represent the magnitudes of the accelerations. The components cannot be reconstructed unambiguously from the resultant. Several possible solutions out of an infinite, three-dimensional set are illustrated in (A). The person experiencing these accelerations is shown in a random orientation, since their orientation relative to gravity is determined by the solution of this geometry. (B, C) Show two possible solutions that are labeled by dotted lines labeled 'a' and 'd' for gravity, and 'b' and 'e' for the physical accelerations corresponding to the resultant 'c' as sensed by the vestibular system. The labeled vectors are the same (relative to the observer) in (A), (B), and (C).

illustrated in Figure 3 can be distinguished by knowledge of magnitude, by detecting the preceding tilt, or by additional knowledge about the direction of gravity. By providing both rotation and direction cues, vision provides the most reliable way of solving the tilt-translation ambiguity. The fact that gravity is equivalent to acceleration thus requires visual–vestibular interactions to indicate the direction of an imposed acceleration and to indicate the pose of the head relative to gravity.

Despite these limitations to working effectively on its own, the vestibular system does have the advantage of always being available for the detection of accelerations greater than a certain minimal amount and of gravity: there is nothing equivalent to closing the eyes for the vestibular system. Also, the vestibular system is extremely fast, providing a signal to the brain in just a few milliseconds.

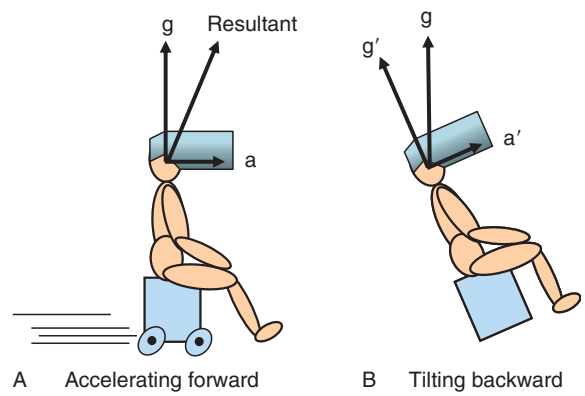

Figure 3 Tilt-translation ambiguity. Because gravity (g) is equivalent to an acceleration, when accelerating physically forward (A) a person is subject to two simultaneous stimuli that combine to produce a vector ('resultant'). A subject tilted backward (B) experiences gravity in that same direction. The only differences between the cues sensed by the otolith organs are the absolute magnitudes (always greater than g for (A) and less than g for (B) and the history of what happened immediately leading up to these situations. However, they can be easily disambiguated by vision. The person illustrated here is wearing a virtual reality helmet in which visual environments can be simulated that attempt to encourage one or other interpretation.

Visual–Vestibular Interactions for Determining the Direction of Translation

Vestibular information concerning the direction of head translation is highly ambiguous (see Figure 2). When visual and vestibular cues to direction are put in conflict (e.g., with prisms or virtual reality displays), the perceived direction of motion is dominated by the visual cue, so much so that even if the actual physical motion (the vestibular cue) is opposite to the optic flow, the perceived direction is determined by vision. There are no significant visual–vestibular

interactions in the determination of the perceived direction of translation. This is supported by the observation that perceived motion direction (even when not generated by self-motion) is not affected by concurrent linear vestibular stimulation.

Visual–Vestibular Interactions for Determining the Magnitude and Velocity of Translation

Neither optic flow nor vestibular signals indicate either the absolute distance traveled or the velocity of self-motion, but the consortium is able to estimate distance traveled by combining information from the different sources. Vision can provide the starting position and velocity estimates that are lacking in the vestibular signal. The vestibular signals produced by active self-motion (e.g., walking) of a planned magnitude can, however, provide a rough scale. When transported passively in the dark (vestibular information only), or when motion is simulated by viewing optic flow while physically stationary (linear vection; vision only), the perceived distance of travel is generally further than the motion experienced, especially for distances further than a few meters with a low acceleration. Artificial combinations of visual and vestibular cues of different magnitude presented, for example, using virtual reality simulations, reveal an averaging of the two signals with a greater weighting of vestibular information when motion is in the operating range of the vestibular system (above about $0.1\,\mathrm{m\,s^{-2}}$).

Attempts to influence the perceived distance of linear vection by tilting a subject, thus applying a component of gravity in the plane of optic-flow-induced movement (see Figure 3), have been unconvincing.

Visual–Vestibular Interactions for Determining Tilt

When forced to rely on the vestibular system, for example, when suspended in water with eyes closed that effectively removes somatosensory pressure cues, subjects can orient themselves either vertically or horizontally relative to gravity to within about 5° with no systematic biases. This ability to detect tilt of the body relative to gravity is a measure of the 'postural vertical.' Systematic errors appear, however, when aligning a line of light with the perceived direction of gravity to indicate the 'subjective visual vertical' while tilted in an otherwise dark room. Small tilts of the body (e.g., 10° to the left) can cause the line to be set in the opposite direction (e.g., 2° right of true vertical), a rather small and variable effect known as 'the E-effect' (for *entgegengesetzt*, German for opposite). Larger tilts, beyond about 60°, usually cause the line to be tilted away from true vertical but in the same direction as the tilt (known as the 'Aubert or A-effect'). The 'perceptual upright,' the orientation at which visually polarized objects such as faces, animals, or letters appear correct and are most easily and speedily recognized, is also subject to an A-effect.

The perceptual vertical, subjective visual vertical, and perceptual upright are all influenced by visual cues concerning the direction of gravity, as well as by the vestibularly transduced physical direction. Subjects, especially children, can be induced to sway by moving the visual environment indicating the effect of vision on the postural vertical. How the subjective visual vertical and perceptual upright can be measured is illustrated in Figure 4 and the effect of visual background and of physical tilt on these two measures is illustrated in Figure 5. The A-effect exhibited by the perceptual upright is substantially larger than that induced in the subjective visual vertical. This is shown in the figure by the distance of the judgments from the 90° line (which corresponds to the veridical direction of gravity relative to the subject) when lying on one side. The effect of the visual background, indicated by the depth of modulation, is larger for the perceptual upright than the subjective visual vertical. Although compensatory rotation of the eyes in response to head tilt might play a small role in these effects, it cannot account for the differences between the two measures. Misalignments of the subjective visual vertical or the perceptual upright from the true direction of gravity indicate that the relevant perception of 'up' arises from a weighted sum of the different cues to 'up'. These cues include (1) the orientation of the body axis, which seems to be known since it can be accurately aligned with gravity in the absence of other cues; (2) visual cues to vertical (described in detail above and in Figure 1); and (3) the otoliths' gravity signal. The perceived orientation of the body is qualitatively different from the other cues as it cannot be sensed directly but is an internal representation. The representation of body orientation is often referred to as the 'idiotropic vector' or the 'M-vector' after Horst Mittelstaedt, who first developed this concept. The orientation of the subjective visual vertical and the perceptual upright can be approximated by linear weighted sums of these three vectors: vision, body axis, and gravity, as illustrated in the lower part of Figure 5. The orientation of the visual and idiotropic cues has relatively little effect on the subjective visual vertical, but the perceptual upright is influenced approximately equally by all three cues.

If a subject is tilted relative to gravity with an acceleration below the semicircular canals' detection threshold (no vestibular cues to the reorientation) and the visual cues to orientation are moved along

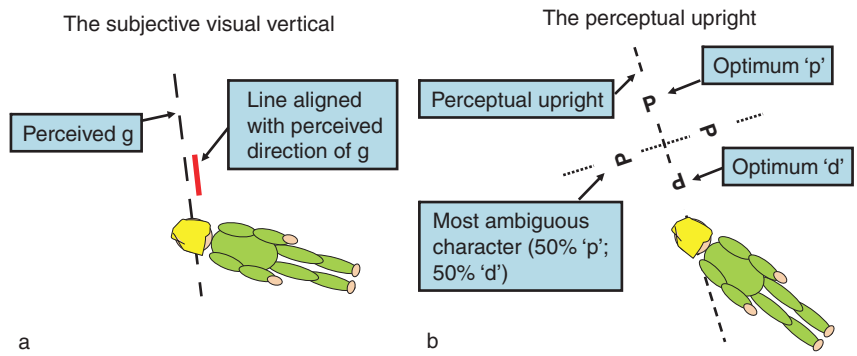

Figure 4 The subjective visual vertical (a) and the perceptual upright (b). These two measures can be used to assess the effect of visual and vestibular cues on perceived orientation. To assess the subjective visual vertical, subjects align a line with the perceived orientation of gravity. Performing this task requires introspection and estimation. The perceptual upright can be assessed by having subjects identify a character or shape the identity of which depends on its orientation. The direction between the orientations of maximum ambiguity (most easily detected psychophysically as the 50% correct identification point) indicates the perceptual upright: the orientation at which objects are most easily recognized

with the subject (visual cues indicate no tilt), then the direction of gravity is not perceived to move during the tilt. This is most easily arranged in a rotating room. For subjects who have been tilted onto their backs, objects hanging from the wall in front of them now appear to levitate as they hang under the influence of gravity, giving the illusion its name, the 'levitation illusion.' The levitation illusion demonstrates the importance of visual and semicircular canal cues for updating the internal representation of gravity. If those cues are absent, even in the presence of a 90° change in the direction of gravity, the perceived direction of gravity can remain unchanged.

When a textured background lacking visual cues to gravity is rotated around the line of sight of an upright observer, it produces a paradoxical perceptual effect in which subjects feel they are rotating constantly, but at the same time statically tilted. This is because of an anchoring effect of gravity (which indicates no tilt). When the same background motion is around an axis aligned with gravity, or when the strength of the visual cues is increased by using a scene rich in frame, horizon, lighting, polarity, and support relationship cues, continuous vection is often experienced.

Visual–Vestibular Interactions for Determining the Axis of Self-Rotation

Determining the axis of self-rotation from vestibular cues alone is also complicated by the vestibular systems' exclusive sensitivity to acceleration: constant velocity rotations are not sensed after the response to the initial acceleration has died away (after about 20–30 s). Subsequent rotations around other axes can then produce complex 'Coriolis effects,' in which, predictable from the laws of conservation of angular momentum, rotation about an axis orthogonal to the (undetected) constant velocity axis produces illusory rotation about an axis orthogonal to both of them. Curiously, tilting one's head (or moving the visual stimulus) during constant velocity circular vection can also evoke complementary complex effects, known as 'pseudocoriolis illusions.'

Visual–Vestibular Interactions for Determining the Amplitude of Rotation

The sensation of rotation from vision alone, circular vection, takes several seconds to build up, but it increases at the same rate that the vestibular response to an actual body rotation decreases. Thus, when rotating in the light a relatively accurate response made up of visual and vestibular components is maintained. This complementarity is reflected in the visual and vestibular responses of cells in the vestibular nucleus. Spinning people around is a classic way of disorienting them: children are intuitively aware that the vestibular system cannot signal position. However, for short-duration rotations, such as natural head movements, the vestibular system is capable of deriving the size and direction of the movement. Even this feat, however, may need help from higher cognitive centers since, when head movements are performed in the dark, only if an Earth-stationary target is imagined can appropriate eye movements be generated. If a small, subject-stationary target is provided as a subject accelerates angularly, the target appears paradoxically to be both moving in orbit around the subject's head, and at the same time not to be advancing from the straight ahead position. This is the 'oculogyral' illusion and might be connected to failures of the process involved in deducing the linear speed of an object under such unnatural conditions.

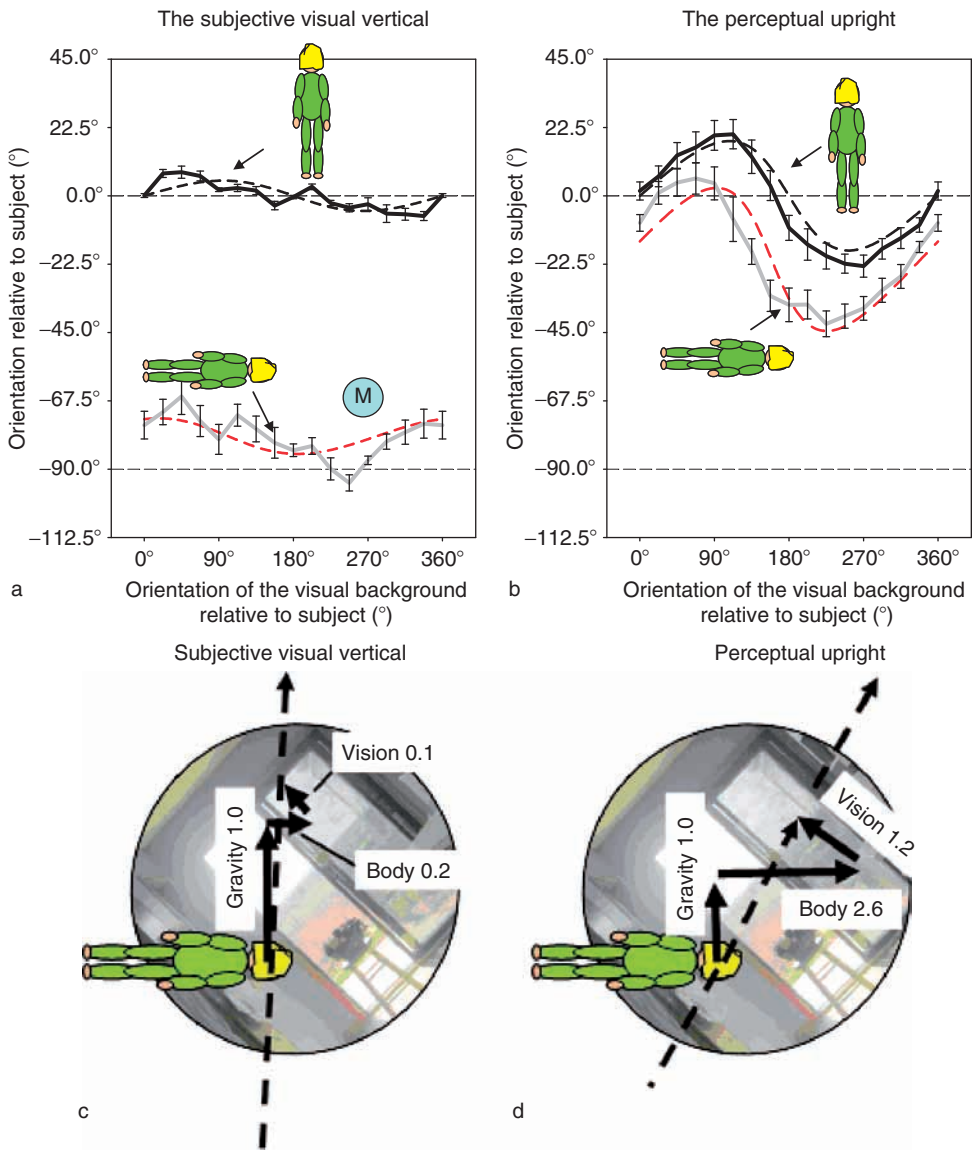

Figure 5 How the perceptual upright and the subjective visual vertical vary with visual and vestibular cues. The graphs plot the orientation of the subjective visual vertical (a) and the perceptual upright (b), relative to the subject, as a function of the visual background for two body orientations, upright and right side down. The subjective visual vertical is less influenced by vision and more influenced by gravity than the perceptual upright. Fitted through the data is the output of a simple, weighted, linear model illustrated in panels (c) and (d) where the lengths of the three vectors are weighted as shown. Data from Dyde RT, Jenkin MR, and Harris LR (2006) The subjective visual vertical and the perceptual upright. *Experimental Brain Research* 173: 612–622. M is datum from Mittelstaedt H (1986) The subjective vertical as a function of visual and extraretinal cues. *Acta Psychologica* 63: 63–85 where subjects lay in a room oriented earth vertically.

Suppression of Visual Responses by Vestibular Signals

Perception of visual motion in general is suppressed by concomitant angular or linear vestibular stimulation. This may be related to the use of optic flow resulting from the visual background motion to inform about self-motion. The suppression may be related to mutual inhibitory pathways between vestibular and visual cortical areas. Motion thresholds are increased during rotation but curiously seem to be elevated most during motion in the direction opposite to that of the associated optic flow. For example, when moving to the left, visual motion to the left is more suppressed.

Suppression of Vestibular Responses by Visual Signals

Although under natural circumstances visual and vestibular cues to motion are usually complementary and, within the constraints described above,

redundant, they can also frequently be opposed to each other. For example, reading a newspaper in a moving vehicle or while walking requires fixation on the paper, not on Earth-stationary features. Visually controlled eye movements, including those involved in stabilizing the image, now work on the subject-stationary image of the paper, whereas the vestibular system is simultaneously required to guide locomotion and to maintain balance. This feat of suppressing vestibular cues for some aspects of perception while simultaneously using it for others can only be successfully carried out at frequencies below about 0.5 Hz and probably only by primates with well-developed smooth pursuit capability. At higher frequencies of movement, eye movements of vestibular origin cannot be suppressed by vision and attempts to fixate on near objects will generate large compensatory eye movements appropriate for maintaining gaze on something that remains Earth stationary, even in the presence of subject-stationary objects.

The Detection of Conflict in Visual–Vestibular Interactions

In many documented instances, information that arises from different senses is combined linearly using a weighted average where the weights are assigned according to the reliability of the contributing sources. This is true for visual–vestibular combinations for both the distance traveled and the perception of orientation (as described above). A linear combination does involve an interaction since the response to both cues can be predicted from the responses to either of the two cues alone. However, conflict between visual and vestibular signals can be independently detected, leading to long-term adaptive changes, especially significant for calibrating vestibular responses as measured subsequently in the dark and, in the short term, to the too-well-known symptoms of motion sickness. Conflict detection is a nonlinear process with its own detection threshold. The nonlinear conflict detection process is a process parallel to the examples of visual–vestibular cooperation described above.

Conclusions

For the detection of self-motion, either rotational or translational, both the visual and the vestibular systems have their weaknesses, which are largely overcome when both signals are present. However, even when both systems are present, systematic errors in assessing velocity and displacement are still found. It may be necessary to include additional sensory and extrasensory information, such as anticipation, prediction, and 'efference copy' of active motor commands to reach natural performance levels.

In assessing orientation, neither visual nor vestibular systems are directly sensitive to gravity. Both must rely on observing the effects of gravity on objects: either on the stones within the inner ear's labyrinth (detected by the hair cells of the vestibular system) or on objects in the environment (detected by the visual system). There is a weighted average of the cues involved in which more weight is placed on the more reliable cues.

The visual–vestibular consortium provides a best estimate of self-motion and orientation cues that enable us to interact effectively with our environment.

See also: Canal–Otolith Interactions; Optokinetic Eye Movements; Sensorimotor Integration: Models; Spatial Orientation: Our Whole-Body Motion and Orientation Sense; Vestibular Influences on Cognition; Vestibular System; Vestibulo-Ocular Reflex; Vision: Mechanisms of Orientation, Direction and Depth.

Further Reading

Berthoz A (2000) *The Brain's Sense of Movement*. Cambridge, MA: Harvard University Press.

Brandt T, Bartenstein P, Janek A, and Dieterich M (1998) Reciprocal inhibitory visual–vestibular interaction – Visual-motion stimulation deactivates the parietoinsular vestibular cortex. *Brain* 121: 1749–1758.

Dyde RT, Jenkin MR, and Harris LR (2006) The subjective visual vertical and the perceptual upright. *Experimental Brain Research* 173: 612–622.

Henn V, Cohen B, and Young LR (1980) Visual–vestibular interaction in motion perception and the generation of nystagmus. *Neurosciences Research Program Bulletin* 18: 459–651.

Howard IP (1982) *Human Visual Orientation*. New York: Wiley.

Mach E (1875/2001) *Fundamentals of the Theory of Movement Perception*, Young LR, Henn V, and Scherberger H (trans.). New York: Kluwer Academic/Plenum.

Mittelstaedt H (1986) The subjective vertical as a function of visual and extraretinal cues. *Acta Psychologica* 63: 63–85.

Previc FH and Ercoline WR (2004) *Spatial Disorientation in Aviation*. Reston, Virginia: American Institute of Aeronautics and Astronautics.

Warren WH (1998) Visually controlled locomotion: 40 years later. *Ecological Psychology* 10: 177–219.

Vocal Communication in Birds

J Podos and D L Moseley, University of Massachusetts, Amherst, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Studies of vocal behavior in birds have made central contributions to animal behavior and the neurosciences. A well-known body of research has built on the discovery that vocal production in birds, particularly songbirds, is underpinned by an elaborate, hierarchical array of song nuclei distributed throughout the brain. Elucidation of the 'song system' has proceeded apace, powered by a host of methodological advances including *in vivo* single-cell recording, *in vitro* recordings from brain slices, selective lesioning, and neuronal tracings. The specialized nature of the song system for vocal production is exemplified by two key observations: the neural morphology of the song system varies seasonally, in a manner that corresponds to variation in vocal output, and in species in which males sing but females do not, song system nuclei are comparatively reduced in females. Studies of the avian vocal system continue to provide key insights into neural mechanisms of motor control.

This entry focuses on three additional, complementary lines of research in the avian vocal system that hold particular relevance for the neurosciences:

- The vocalizations of some bird groups, including songbirds, hummingbirds, and parrots, develop through imitative learning, in a manner that parallels speech acquisition in humans. Vocal learning in songbirds is now known to be mediated by a specialized neural pathway, the anterior forebrain pathway, the development of which depends on birds' auditory experiences. A 'developmental stress' hypothesis suggests that learned song can provide a reliable indicator of a male's neural developmental history and that females may thus use vocal parameters to assess the quality of prospective mates. Experimental studies with hand-reared birds illustrate a previously unrecognized axis of plasticity in sensorimotor development, in which birds faced with physical limits on vocal model reproduction can calibrate their 'targets' of learning.
- Vocal production entails extensive use of peripheral mechanisms. The vocal acoustic source, the syrinx, is dynamically coordinated with activity of the respiratory and vocal tract motor systems. Rapid and precise modulations in breathing patterns are required for the production of intricate vocal timing patterns, and modulation of vocal tract configuration during phonation enables birds to adjust vocal tract resonances and thus the spectral properties of song. It is not yet clear how birds manage to coordinate these multiple motor systems in vocal production, in terms of both neural control and its development. Limits on vocal performance, associated in particular with peripheral systems of vocal production, have been shown to constrain the production and evolution of certain vocal parameters.
- The biological function of birdsong has been examined with vocal playback studies, which are providing new insights into birds' perceptual and recognition abilities. Field playback tests have been conducted primarily with territorial males, in which vocal stimuli are presented to simulate territorial intrusions. Laboratory playback tests have been used to test birds' perceptual capabilities and limitations, and the preferences of females for different vocal stimulus classes. In conjunction with analyses of vocal variation, these studies are helping to specify the role of song in species recognition. 'Interactive' playback studies are helping researchers assess how birds modulate vocal output to convey varying levels of aggressive intent.

Vocal Learning

Songbirds learn to sing using a diversity of learning strategies or 'programs.' As with human speech development, vocal imitation in birds includes sensory and sensorimotor phases. During the sensory phase, birds memorize song models to which they are exposed. The sensorimotor phase, defined by transitions of vocal output from babblelike 'subsong' to crystallized adult forms, is mediated by comparison of vocal output to memorized song models via auditory feedback. During this phase, birds develop general proficiency with the vocal apparatus, as well as proficiency in the reproduction of memorized song models.

Species vary considerably in the timing of closure of their sensitive phases of song acquisition, in innate tendencies to memorize species-specific song stimuli, and in the relative impacts of social interactions on song model development. Within-species studies have proven especially beneficial in addressing the evolutionary causes of diversification in learning programs. The most comprehensive such studies to date have focused on white-crowned sparrows. Some populations of this species engage in long-distance migrations, whereas others remain essentially sedentary

throughout the year. Migration tendencies in white-crowned sparrows have now been shown, in both laboratory and field studies, to correlate with at least three aspects of song learning programs. First, migratory birds tend to experience earlier closure of the sensory phase of song learning. Second, birds from migratory populations tend to memorize more song types than do their sedentary counterparts, as revealed in vocal output during the initial stages of sensorimotor learning. Third, migratory populations show more evidence of 'selective attrition' during the sensorimotor phase of song learning; that is, they retain only a subset of previously memorized models. Interpopulation differences in learning programs enable birds from across the spectrum of migratory tendencies to crystallize song types that are shared by neighboring territorial rivals; as such, this diversity in learning programs may facilitate aspects of song function (see the section titled 'Song function'). In addition to highlighting ecological correlates of song learning, these data raise interesting new questions about the hormonal and neural bases of diversity in song learning programs.

Another recent body of research on vocal learning in songbirds focuses on the potential impacts of phenotypic condition on song learning and development. This work is motivated in part by advances in sexual selection theory, which posits that mating signals such as bird songs should evolve to provide accurate, reliable indicators of male genetic and phenotypic quality. Female choice of mates in many animals is informed by male signaling behavior. Females presumably evolve to attend to signal features that honestly and reliably reflect male quality because of the enhanced reproductive fitness of females that choose higher-quality mates. To explain the putative link between male quality and signal evolution, evolutionary biologists often refer to the so-called 'handicap' principle, which posits that high costs in signal production or expression enforce the honest and reliable nature of signaling. In this context, a stated goal of many current studies of mating signal evolution is to elucidate their underlying metabolic, developmental, and fitness costs.

Bird vocalizations are a promising system for studies of signal costs, given the elaborate neural and motor mechanisms that underpin vocal expression and development. Available evidence suggests that vocal production in adult birds requires only minimal increases in metabolic output above and beyond basal metabolic rates. However, accumulating evidence suggests that during vocal development, birds do face significant costs associated with stresses incurred during birds' early months of life. Young birds may suffer developmental stress from numerous sources, including insufficient nutrition, parasite infection, and trials of thermal regulation. Recent experimental work indicates that developmental stress can indeed interfere with the acquisition and expression of certain song features. Canaries infected with malaria at a young age and zebra finches reared under conditions of stress were found as adults to produce smaller song repertoires and smaller syllable repertoires, respectively. Swamp sparrows reared with restricted access to food copied song model notes less accurately than did birds reared with *ad libitum* access to food.

Degradation of vocal imitation abilities caused by developmental stress would likely be caused by subpar neural growth and development. Consistent with this hypothesis, the volume of the higher vocal center (HVC), a telencephalic song nucleus, was found to be significantly diminished in canaries reared under developmental stress compared with HVC volume in control birds. Similar results have been reported for zebra finches and swamp sparrows, although for the latter species, another song nucleus, the robust nucleus of the archistriatum (RA), suffered the greatest impact under conditions of developmental stress. These results are broadly consistent with comparative studies showing positive correlations between song nuclei size and song or syllable repertoire size. Males that suffer from only limited developmental stress would theoretically be able to devote more metabolic resources to developing larger song nuclei, and thus to developing larger song repertoires.

The evolutionary relevance of developmental stress can be tested by assessing, using song playback, the functional impact of variation in song features affected by developmental stress. In one such test, wild-caught female song sparrows were presented with songs recorded from males that had been imitated either with high accuracy or with poor accuracy. The dependent variable in this study was the copulation solicitation display, a behavioral pattern that normally precedes copulation in the wild and is thus regarded as a valid indicator of female song preferences. Subjects in this study displayed more often to playback of songs that had been learned with high accuracy. In a similar experiment, female zebra finches were shown to significantly prefer the songs of males that had been reared in a nonstressed environment over songs from males reared with nutritional stress. Available data thus provide increasingly strong support for the developmental stress hypothesis.

Studies of song learning have also revealed that birds can adjust the targets of song learning during sensorimotor development. Birds normally reproduce, with impressive accuracy, song models they had memorized previously. However, under some

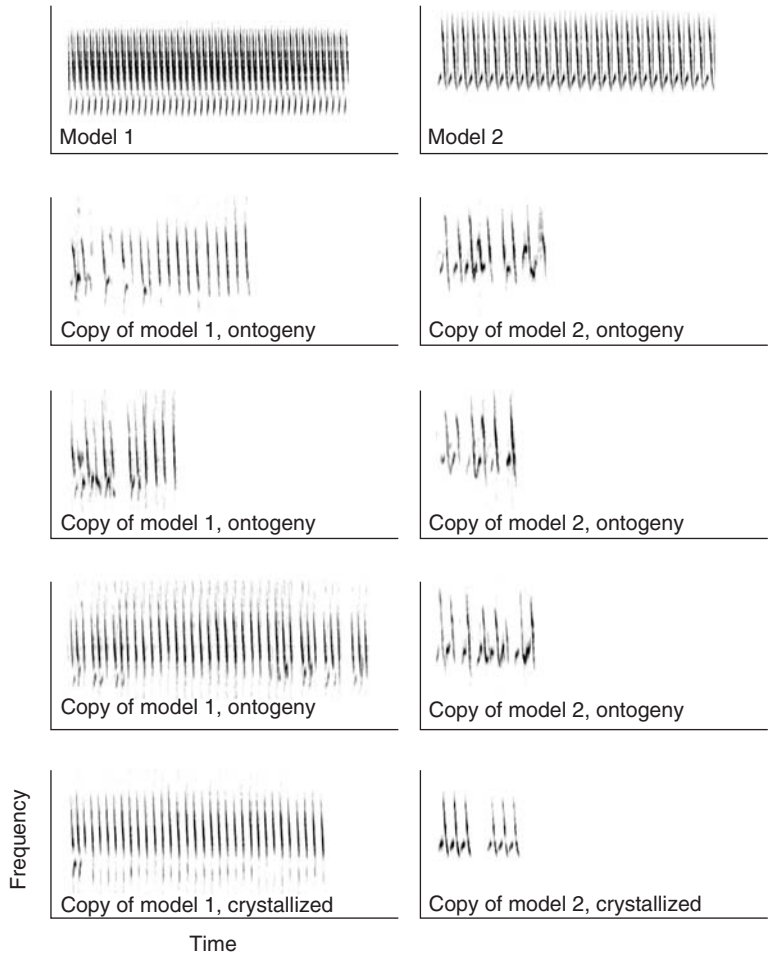

Figure 1 Sensorimotor calibration of song learning targets in swamp sparrows, Each column shows a song model (top row), a crystallized copy of the model (bottom row), and three ontogenetic intermediaries of those models produced during sensorimotor development. Song models featured trill rates manipulated to exceed natural trill rates. The bird in each example is observed to 'experiment' during ontogeny with potential motor solutions in reproducing manipulated song models. In the first column, the bird is seen to attempt but eventually fail to integrate the low-frequency notes of the model into the crystallized song (with the exception of the opening syllables). In the second column, the bird is seen to experiment with varying trill rates and reduced frequency bandwidths before settling on the 'broken' syntax that was eventually crystallized. *X*-axis scale, 0–2.5 s; *y*-axis scale, 0–8 kHz. Reproduced from Podos J, Peters S, and Nowicki S (2004) Calibration of song learning targets during vocal ontogeny in swamp sparrows. *Animal Behaviour* 68: 929–940, with permission.

circumstances, birds may face performance limitations in the reproduction of certain vocal parameters. For example, some song features may be situated near the limits of a species' production capacities, or some birds may develop generally poor proficiency in vocal learning and reproduction because of developmental stresses. Young swamp sparrows reared with artificially manipulated song models, in which trill rates were elevated above natural rates, showed flexibility during sensorimotor learning in the reproduction of these rapid models. Memorized model targets were calibrated, over the course of sensorimotor learning, to better match birds' motor performance capacities (Figure 1). These data support 'active' models of song learning and sensorimotor development and suggest that the song-learning template retains plasticity through sensorimotor development even for birds with closed-ended memorization programs.

Peripheral Mechanisms of Vocal Production

Research on the neural bases of vocal behavior has traditionally focused on the nature and architecture of connections that descend to the brain stem and ultimately to the syrinx, the sound source. More-recent studies have started to evaluate the contributions in vocal production of the respiratory system

and the vocal tract. Above and beyond interest in questions about vocal mechanics, this work is motivated by evidence that peripheral mechanisms can delimit the range of vocal behaviors that can be expressed and thus evolve.

The temporal features of vocalizations can be shaped by dynamics of respiratory ventilation. In recent studies, respiratory and syringeal motor output was monitored in vocalizing birds by means of surgically implanted microbead thermistors and pressure transducers. One finding of this work is that birds vocalize almost exclusively during respiratory exhalations; by contrast, birds generally do not phonate during respiratory inspirations. As birds breathe out, airflow across the syrinx induces syringeal tissues to vibrate and thus produce sound. Patterns of breathing help define the structure of vocal trills, in which one or more notes (syllables) are repeated in rapid succession. For trills with slow to moderate rates of syllable repetition, birds employ a 'minibreath' motor strategy, with each respiratory expiration followed by a respiratory inspiration. As such, birds are able to maintain a steady tidal volume of air and thus produce trills of extended duration. For the production of more-rapid trills, however, birds appear to switch to another motor program, 'pulsatile' respiration, in which the tidal volume of the singing bird is rapidly depleted through expirations only. In a recent study, mockingbirds were presented with recordings of canary trills, with syllable repetition rates of 20 Hz and of extended duration. Mockingbirds that attempted to mimic these trills proved unable to reproduce them accurately, presumably as a result of performance limitations in their respiratory dynamics. More specifically, mockingbirds reproduced the rapid trill models using pulsatile rather than minibreath respiration, and resulting trills featured 'broken' syntax in which multisyllable segments were separated by pauses. Broken syntax had been observed previously in young swamp sparrows trained with trilled song models in which trill rates had been artificially elevated (e.g., **Figure 1**, right column).

The vocal tract, which includes the trachea, glottis, tongue, and beak, also affects song expression. The vocal tract acts as a resonance filter of sound produced by the syrinx, as first demonstrated in a study of birds that sang in an experimental helium-enriched environment. Vocalizations produced in this environment included harmonic overtones that would not otherwise have been expressed. These results indicate that the tonal quality of song is shaped by resonance properties of the vocal tract. Recognition of an acoustic function for the vocal tract helps explain the evolution of unusually elongated trachea in a wide diversity of birds including cranes, swans, curassows, and geese.

In songbirds, song production often includes rapid reconfigurations of the vocal tract. High-speed video recordings of several species of sparrows and finches indicate that changes in beak gape co-occur with temporal changes in vocal fundamental frequency. This observation concords with a resonance model of vocal tract function: wider beak gapes effectively shorten the volume of the vocal tract, thus providing a more appropriate resonance filter for high-frequency sounds. Performance constraints in vocal tract modulation have been offered as an explanation for an observed trade-off between two vocal parameters, trill rate and frequency bandwidth, in the evolution of trills in emberizine finches and sparrows (**Figure 2**). Rapid trill rates require rapid vocal tract modulations, and wide frequency bandwidth production requires broad vocal tract modulations. Because of a trade-off in the rapidity versus the breadth of vocal tract modulations, trills beyond a certain rate can presumably be accompanied only by vocal tract modulations of narrow breadth. Similar trade-offs between trill rate and frequency bandwidth have now been described in other taxa, including blackbirds, wrens, and warblers.

Recognition of the beak's contribution to vocal production has suggested a previously unrecognized mechanism of vocal evolution. Birds are known for diverse adaptations of their beaks, as they have occupied diverse feeding niches. Evidence now suggests that the evolutionary diversification of beaks can invoke incidental consequences on the evolution of song structure. In particular, a 'vocal tract constraint' hypothesis predicts that birds with comparatively large beaks, adapted for crushing hard seeds, should suffer comparatively severe constraints on vocal performance because of biomechanical trade-offs between the force and speed of beak gape modulations. Support for this hypothesis comes from studies of several bird groups, including Darwin's finches of the Galápagos Islands, Ecuador. Analysis of song structure from birds with different beak morphologies, both across and within species, reveals differences in expressed 'vocal performance' that are consistent with the vocal tract constraint hypothesis (**Figure 3**). As birds adapt to new habitats, adaptations of the beak may thus drive secondary changes in song structure.

Because song is used in mate and species recognition, beak adaptations and subsequent evolutionary changes in song may facilitate reproductive isolation and speciation. This hypothesis is supported by field studies of a population of medium ground finches at *El Garrapatero* on Santa Cruz Island, Galápagos. This population features two discrete beak-size morphs, with small-beaked birds producing songs with consistently higher

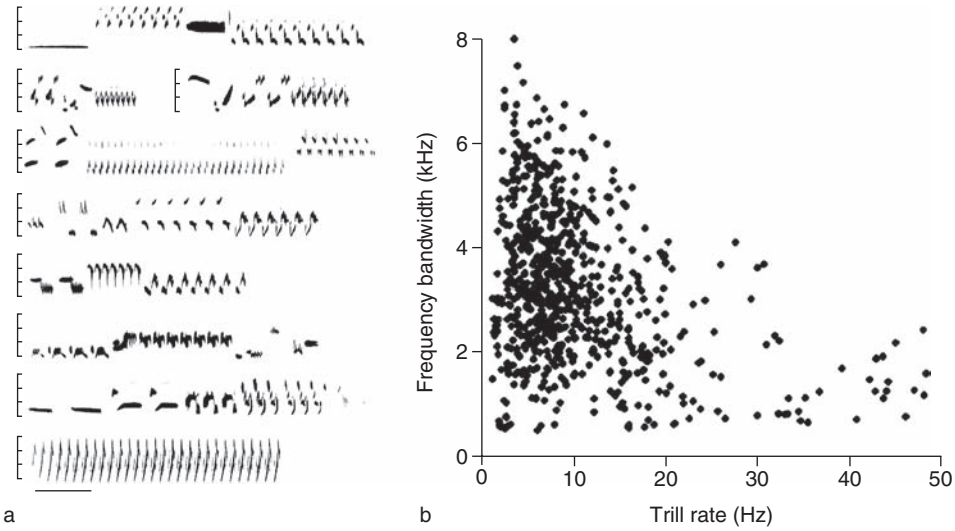

Figure 2 (a) Songs of eight North American Emberizine sparrows, illustrating the diversity of songs that can be expressed even among closely related species. Each of these songs contains trilled sequences, which can be characterized in part by trill rate (the rate of syllable repetition) and frequency bandwidth (the range of fundamental frequencies expressed over the course of the trill). This horizontal scale bar is 0.5 s, and y-axis spans 2–8 KHz. (b) Plot of trill rate versus frequency bandwidth for 740 trilled sequences across 34 species from the family Emberizidae. Triangular distributions evident for the family (here) as well as for individual species and genera support a vocal tract constraint hypothesis. (a, b) Reproduced from Podos J (1997) A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution* 51: 537–551, with permission from Blackwell publishers.

levels of vocal performance. Differences in the songs of small- and large-beaked birds can be detected by both ornithologist and finch, the latter as indicated in playback studies to territorial males. Female Darwin's finches generally choose mates based primarily on song structure, and preferences for morph-specific songs in the *El Garrapatero* population presumably explain observed patterns of assortative mating and incipient genetic divergence among morph types.

Song Function

Species Recognition

Birds sing mostly to communicate with individuals, such as potential mates or rivals, of their own species. For song to function in this way, individuals must have the ability to recognize the songs of their own species and to differentiate such songs from those produced by heterospecifics. Songs indeed have diverged markedly among species, to the extent that most species produce songs that can be considered 'species-specific.'

Interspecific divergence of songs is thought to be driven largely by selection against hybrid (cross-species) mate selection and offspring production. Males that produce species-distinctive songs, and females that are able to successfully identify conspecific songs, will be more likely to achieve conspecific matings and thus avoid costly hybrid matings. Thus the production and recognition of species-distinctive songs should propagate along distinctive trajectories in different songbird lineages. Support for this 'sound–environment' hypothesis has been garnered from studies showing that species living in proximity, especially those that are closely related, tend to evolve songs that occupy distinct positions in 'acoustic space,' defined by dimensions of song structure. A recent study of neotropical antbird songs provides strong support for the sound–environment hypothesis. This study evaluated song similarity within nine species trios, each representing a genus of the antbird family Thamnophilidae. Each trio was comprised of a focal species, a closely related species living in sympatry (at the same location as the focal species), and a closely related species living in allopatry (at a different location from the focal species). Across all nine trios, the songs of allopatric species pairs were found to be significantly more similar than songs of sympatric species pairs. This finding suggests a recent history of independent song divergence events within sympatric species, presumably caused by selection favoring the evolution of species-specific song production and song recognition abilities.

Compared with song features that vary among individuals, song features that do not vary within a species ('invariant' features) presumably serve as reliable species identification signals. This is especially true for invariant features that are clearly distinct (i.e., have little structural overlap) from songs of

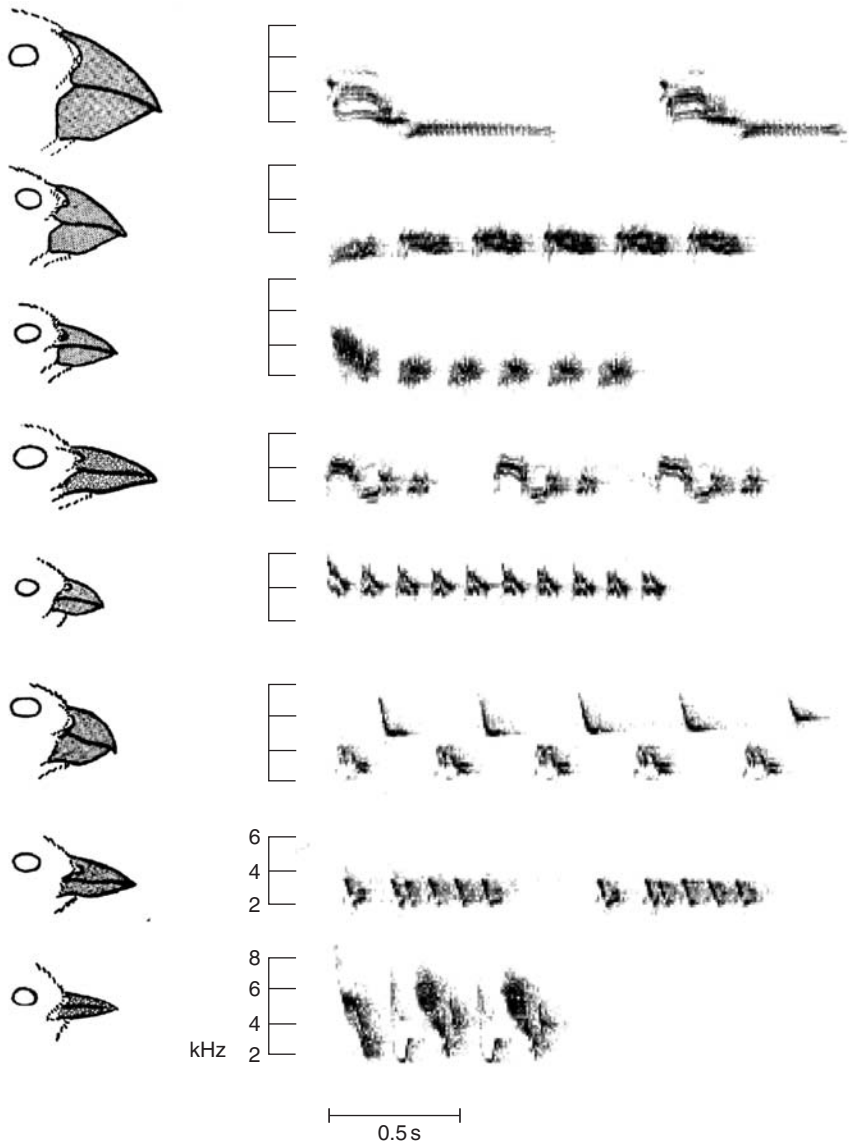

Figure 3 Beak morphology and representative song spectrograms from eight species of Darwin's finches. Interspecific variation is evident in both morphology and song structure. Each of these songs includes trilled segments. Birds with larger beaks, for which rapid gape modulations are comparatively difficult to achieve, tend to produce trills with lower trill rates and more narrow frequency bandwidths than the trills of their smaller-beaked counterparts. Reproduced from Podos J (2001) Correlated evolution of morphology and vocal signal structure in Darwin's finches. *Nature* 409: 185–188, with permission.

sympatric species. A recent study of field sparrows tested, using a series of playback trials, the functional significance of multiple invariant song features. Playback stimuli were synthesized in which single invariant features were modified, one at a time, to values outside of the species-typical range. Territorial males were then presented with a choice of modified and control (unmanipulated) songs. Males were found to approach speakers broadcasting control songs more often than speakers broadcasting modified songs, as expected. Follow-up trials were then performed in which test stimuli expressed modifications in all possible combinations of invariant features. These trials enabled an assessment of feature 'weighting' in the perceptual salience of field sparrow song, with song frequencies ultimately shown to be the invariant song features of the greatest import.

Song matching

The functional consequences of variation in song structure have traditionally been assessed using 'loop' playback studies, in which song stimuli are presented according to previously determined sequences and rates, for example as recorded in advance on a cassette tape taken to the field. With the advent of portable

computers and specialized software packages, however, it is now possible to 'interact' with focal birds by selecting a specific playback song from a set of available digitized stimuli. Interactive playbacks are allowing researchers to test hypotheses that before they could investigate only through correlative analyses.

A growing number of experiments using interactive playback focus on the question of how birds signal aggressive intent. In bird species with large repertoires, neighboring individuals often share subsets of song types. In such species, males engaged in territorial skirmishes can respond to rivals with song types that are either shared or not. According to theory, the degree to which birds match the song types of rivals may reflect levels of aggressive intent. This theory has now been supported by an experiment with song sparrows that used interactive playbacks. Territorial male song sparrows were found to escalate their levels of response when they received playback of song types similar to the ones they had been singing, compared with their response to playback of non-matching song types. Matching can also occur for other vocal parameters. Black-capped chickadees were found to approach a playback speaker more closely and make more flights over the speaker when presented with songs in which the acoustic frequency of the stimulus closely matched frequencies of their songs, compared with their responses to playback of songs with different frequencies.

Song overlapping

Studies of chickadees and other species have examined another type of aggressive signaling, termed 'overlapping.' Overlapping occurs when a bird begins to sing before another individual has completed its song. In some species, such as European thrushes, the efficacy of communication is more compromised in the overlapped bird than in the overlapper, given that the degree of information encoded in songs tends to increase toward the end of songs. Nightingales have shown intense responses to playbacks of overlapping song. When presented with songs from two speakers, one speaker that broadcast overlapping songs and one that did not, male nightingales responded more rapidly, more vocally, and with closer approaches to the speaker that broadcast the overlapping song. In an experiment that varied both frequency matching and overlapping, male chickadees were most agitated when presented with playback stimuli that both overlapped and matched the frequency of their songs.

Eavesdropping

In most vocal interactions, individuals other than the intended receiver are able to overhear the signaling individuals. In a series of studies on great tits, researchers showed that both males and females can use information gathered by eavesdropping to assess rivals or prospective mates. In one such study, researchers broadcast a simulated interaction of two 'virtual' individuals, unknown to the focal bird, in stereo over two speakers outside his territory. In this interaction, one of the two virtual individuals either switched song types (to match the other simulated individual or not) or continued to sing a single song type. The songs of this individual were subsequently presented to the focal bird. For the former condition, in which the virtual bird had matched song types during its earlier simulated interaction, focal birds responded more strongly than they did when virtual territorial intruders had not matched during simulated interactions. Thus it was shown that by eavesdropping, male great tits extract information about the level of perceived threat from potential rivals.

Another clever experiment addressed the topics of eavesdropping and song overlapping simultaneously by providing male great tits the opportunity to gather information from both direct and indirect interactions. Songs of a simulated intruder stranger male (X) were broadcast to focal territorial males. With the use of interactive playback, these songs were varied either to overlap or to alternate with a subject's songs. A simulated interaction between this intruder male, X, and another virtual, stranger male, Y, was then played outside the subject male's territory, with a different speaker for each bird. Male X was manipulated to either 'win' or 'lose' the interaction in terms of patterns of overlap. Finally a territorial intrusion by male Y was then simulated and the focal male's responses measured. If the subject male had been manipulated to 'beat' male X, and male X to 'beat' male Y, the subject tended to respond with less effort, fewer songs, and diminished approach to male Y, presumably because of diminished perceived threat. By contrast, in scenarios in which the threat level of male Y was ambiguous (if, for example, the subject was manipulated to 'beat' male X, but X subsequently lost to male Y) or the level of threat was high (if X beat the focal male, but Y subsequently beat X), subjects responded with significantly more songs and closer approaches to the speaker, presumably because of a comparatively elevated perceived threat.

Females presumably could also use eavesdropping to evaluate social mates or potential extrapair partners. In an experiment with great tits, focal males were presented with overlapping ('escalated') playback songs, but neighboring males were presented with de-escalated (fewer songs, with shorter phrases) playback songs. Females mated to males that

had received escalated playback (apparent 'losers') subsequently visited territories of neighboring males more often than did females whose males had received the de-escalated treatment (apparent 'winners'). This finding holds functional relevance because females might visit neighboring male territories to seek copulations from perceived higher-quality neighbors. In another experiment, female black-capped chickadees were found to engage in extrapair copulations with greater frequency if their mates had received aggressive, overlapping playbacks at an earlier time. These studies together provide new evidence of advanced perceptual and cognitive abilities in birds and suggest avenues for future studies that integrate neurobiological, psychological, and ecological perspectives.

See also: Animal Communication: Honesty and Deception; Bird Brain: Evolution; Bird Song Systems: Evolution; Birdsong Learning: Evolutionary, Behavioral, and Hormonal Issues; Birdsong Learning; Birdsong: The Neurobiology of Avian Vocal Learning; Signal Production and Amplification in Birds.

Further Reading

Beecher MD and Brenowitz EA (2006) Functional aspects of song learning in songbirds. *Trends in Ecology and Evolution* 20: 143–149.

Huber SK, de Leon LF, Hendry AP, Bermingham E, and Podos J (2007) Reproductive isolation of sympatric morphs in a population of Darwin's finches. *Proceedings of the Royal Society B* 274: 1709–1714.

Marler P and Slabbekoorn H (eds.) (2004) *Nature's Music: The Science of Birdsong*. San Diego, CA: Elsevier Academic Press.

Mennill DJ and Ratcliffe LM (2004) Overlapping and matching in the song contests of black-capped chickadees. *Animal Behaviour* 67: 441–450.

Nelson DA (1988) Feature weighting in species song recognition by the field sparrow (*Spizella pusilla*). *Behaviour* 106: 158–182.

Nelson DA, Marler P, and Palleroni A (1995) A comparative approach to vocal learning: Intraspecific variation in the learning process. *Animal Behaviour* 50: 83–97.

Nowicki S and Searcy WA (2005) Song and mate choice in birds: How the development of behavior helps us understand function. *Auk* 122: 1–14.

Peake TM, Terry AMR, McGregor PK, and Dabelsteen T (2002) Do great tits assess rivals by combining direct experience with information gathered by eavesdropping? *Proceedings of the Royal Society of London B* 269: 1925–1929.

Podos J (1997) A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution* 51: 537–551.

Podos J (2001) Correlated evolution of morphology and vocal signal structure in Darwin's finches. *Nature* 409: 185–188.

Podos J, Huber SK, and Taft B (2004) Bird song: The interface of evolution and mechanism. *Annual Review of Ecology, Evolution and Systematics* 35: 55–87.

Podos J, Peters S, and Nowicki S (2004) Calibration of song learning targets during vocal ontogeny in swamp sparrows. *Animal Behaviour* 68: 929–940.

Searcy WA and Nowicki S (2005) *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton, NJ: Princeton University Press.

Seddon N (2005) Ecological adaptation and species recognition drives vocal evolution in neotropical suboscine birds. *Evolution* 59: 200–215.

Slater PJB (2003) Fifty years of bird song research: A case study in animal behaviour. *Animal Behaviour* 65: 633–639.

Spencer KA, Buchanan KL, Leitner S, Goldsmith AR, and Catchpole CK (2005) Parasites affect song complexity and neural development in a songbird. *Proceedings of the Royal Society B* 272: 2037–2043.

Suthers RA and Zollinger SA (2004) Producing song: The vocal apparatus. *Annals of the New York Academy of Sciences* 1016: 109–129.

Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies

B Rudy, J Maffie, Y Amarillo, B Clark, E M Goldberg, H-Y Jeong, I Kruglikov, E Kwon, M Nadal, and E Zagha, New York University School of Medicine, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Voltage-gated potassium channels have K^+-selective pores that are opened by membrane depolarization. This opening allows the movement of K^+ ions across the plasma membrane and the generation of K^+ currents that tend to repolarize the membrane toward the equilibrium potential for K^+ (E_K). Voltage-gated potassium channels contribute widely to the electrical properties of neurons. They influence subthreshold properties, including the resting potential and membrane resistance. They influence the amplitude and frequency of subthreshold oscillations, the responsiveness of the cell to synaptic inputs, and the probability of spike generation. They help shape postsynaptic potentials, and they are the main determinants of the repolarization of the action potential governing spike shape and frequency. Their voltage-dependent activity ensures a non-ohmic current–voltage relationship, which thereby enables the channels to contribute to the nonlinear properties of neurons. Voltage-gated potassium channels have similar functions in other excitable cells, including all varieties of muscle. In non-excitable cells, they contribute to the resting potential and to the regulation of Ca^{2+} entry and secretion.

Voltage-gated K^+ channels differ dramatically in their kinetic and voltage-dependent properties as well as their cellular and subcellular distributions. This diversity is a main contributor to the varied electrical properties of neuronal populations throughout the nervous system. Thus, understanding the input–output relationship of neuronal elements demands the continued effort to study the properties and localization of these channels and analyze their physiological roles in native membranes.

Voltage-gated K^+ channels are tetramers of primary, or pore-forming, subunits (also known as α subunits). These tetramers form the infrastructure of the channels and in most cases are sufficient to form functional channels. However, many voltages-gated K^+ channels in their natural environment also include associated or auxiliary proteins (sometimes referred to as β subunits). These proteins have primary sequences not resembling those of the pore-forming subunits, can significantly modify the properties of the channels, and can be essential for the efficient expression of functional channels in the plasma membrane. In addition, other proteins such as regulatory enzymes and elements of the cytoskeleton have been shown to interact with many K^+ channel molecular complexes.

Based on sequence similarity, the pore-forming subunits of mammalian voltage-gated K^+ channels can be classified into three groups of proteins that correspond to distinct functional classes. The first group includes the proteins of the Kv1–Kv6, Kv8, and Kv9 subfamilies (collectively called here the 'Kv_F family' for fast-activating Kvs), components of voltage-gated K^+ channels that activate quickly upon membrane depolarization. In contrast, proteins of the Kv7 subfamily (often called the KCNQ family) and of the Kv10–Kv12 subfamilies (also known as the EAG family) are pore-forming subunits of voltage-gated K^+ channels with comparatively slow kinetics. The subunits of the Kv7 family form the M-type K^+ channels underlying the current known as I_M, a subthreshold operating current of considerable importance in the regulation of neuronal excitability. This article discusses the channels formed by the pore-forming subunits of the Kv_F and Kv7 families (Kv subfamilies 1–9; Figure 1). A separate article in this encyclopedia discusses the more distantly related EAG (Kv10–Kv12) family.

Structure of Kv Proteins

Kv subunits consist of six membrane spanning domains (S1–S6) flanked by intracellular NH_2 and COOH terminal sequences of variable lengths (Figure 1). The linker between the S5 and S6 transmembrane helices partially enters the membrane and is known as the P loop or P domain. This domain contains the 'K^+ channel signature sequence' (TVGYG), which is highly conserved not only among Kv proteins but also among all K^+ channel subunits, including those from prokaryotes, and contributes to the formation of the K^+ selective pore (Figures 1 and 2). At the N-terminus, preceding the S1 helix, there is a sequence known as the tetramerization or T1 domain. T1 domains of members of the same subfamily are very similar and determine subfamily specific association.

The fourth membrane spanning domain (S4) is characterized by the repetition of a motif consisting of two neutral residues (usually hydrophobic except toward the COOH end of S4) and one positively charged residue (usually arginine). The number of repetitions of this motif is characteristic of each Kv_F subfamily: seven in Kv1 subunits, five in Kv2s (as well as Kv5, Kv6, Kv8, and Kv9) and Kv4s, and six in Kv3s. The S4 domain in Kv7 (KCNQ) subunits has six positive charges, except in Kv7.1, which has four.

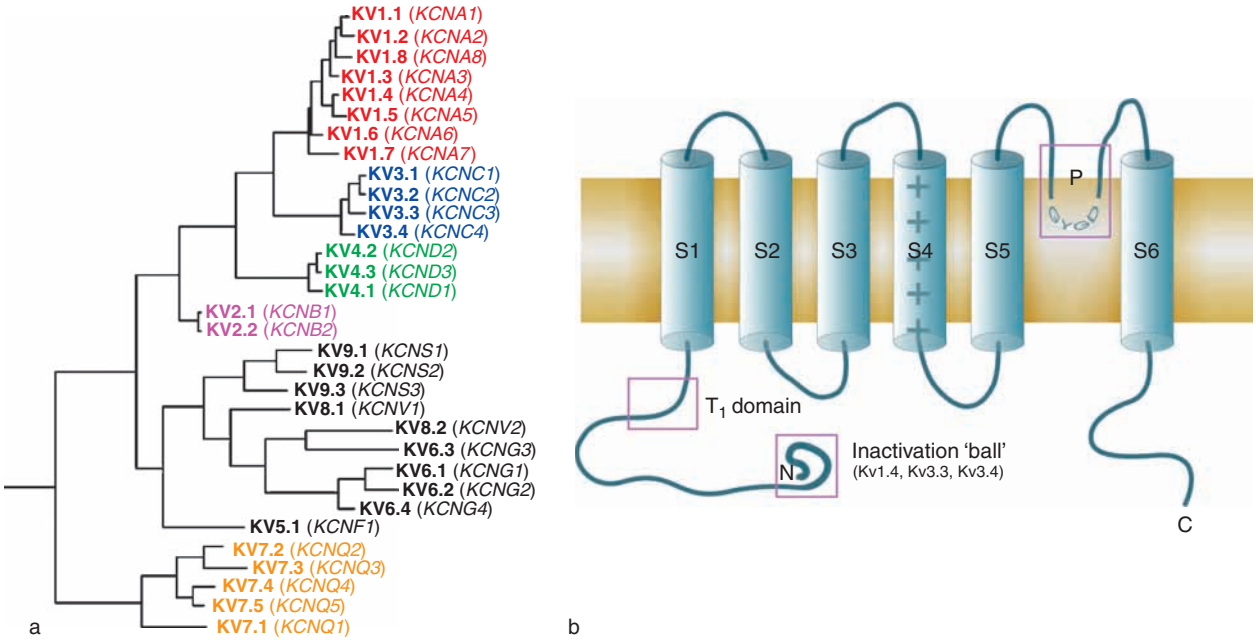

Figure 1 (a) Phylogenetic relations of Kv1–Kv9 K^+ channel subunits. The functional classes discussed in this chapter also form closely related subgroups based on sequence similarities. (b) Schematic representation of a Kv subunit. Kv subunits have six membrane spanning domains (S1–S6) flanked by cytoplasmic amino and carboxyl ends. The pore loop region (P), critical to the formation of the K^+ channel selectivity filter is found between the S5 and S6 helices. The T1 domain, involved in the formation of tetrameric structures among members of the same subfamily and in interactions with auxiliary proteins, is in the amino end region preceding the S1 domain. Some Kv proteins contain N-inactivation domains at the N-terminus. The membrane is shown in yellow. (a) From Yu FH and Catterall WA (2004) *The VGL-chanome: A protein superfamily specialized for electrical signaling and ionic homeostasis*. STKE Science's 253: re15.

The S4 domain is a critical part of the voltage sensor; the positively charged residues are likely the gating charges, which move in response to changes in membrane potential to open and close the channel's pore.

The structure of a mammalian Kv channel, a complex consisting of four Kv1.2 subunits and four auxiliary Kvβ2 proteins, has been resolved at 2.9 Å in the laboratory of Roderick MacKinnon (Figure 2a). In the channel structure, four T1 domains, one from each of the four Kv1.2 subunits, interact to form a tetrameric assembly at the intracellular membrane surface. This assembly is located directly under the cytoplasmic side of the channel's pore. As a result, the pore communicates with the cytoplasm through side portals or windows formed by the linkers connecting the T1 domains to the first membrane spanning helices. These portals allow K^+ ions to flow freely between the pore and the cytoplasm. These portals are also large enough to allow the entry of the N-terminal inactivating polypeptide responsible for N-type inactivation. Four Kvβ2 subunits interact with the T1 domains.

In the structure derived by MacKinnon and colleagues, the channel consists of two relatively independent domains: a voltage sensor domain consisting of the S1–S4 helices of the four Kv subunits and a pore domain consisting of the S5 and S6 membrane spanning helices with the P loop in between, as shown in Figure 2. The structure of the pore domain is similar in voltage-gated K^+ channels and in K^+ channels that are not voltage dependent, such as the potassium channels know as inward rectifiers. These channels consist of subunits that have only two helices (inner and outer, equivalent to the S6 and S5 helices, respectively) and a P loop. The pore domains of the four subunits create an 'inverted teepee,' or cone, with the S6 (or inner helix) facing the inside. The P loops of the four subunits are located in the wider third of the teepee, near the extracellular surface, forming the selectivity filter, which surrounds the narrowest part of the pore. The selectivity filter is determined by the amino acids of the K^+ channel signature sequence, oriented in such a way that they expose their main chain carbonyl oxygen atoms to the pore. During dehydration of the K^+ ion, these carbonyl oxygen atoms replace the oxygen atoms of the water hydrating the ion, allowing its entrance into the narrow pore delimited by the selectivity filter. The carbonyl oxygens are too far apart to properly surround and stabilize the smaller Na^+ ion, thus explaining K^+ selective permeation.

In the crystal structure, the voltage sensor domain is connected to the pore domain through helices made by the S4–S5 linker, which makes several amino acid contacts with the S6 or inner helix lining the pore

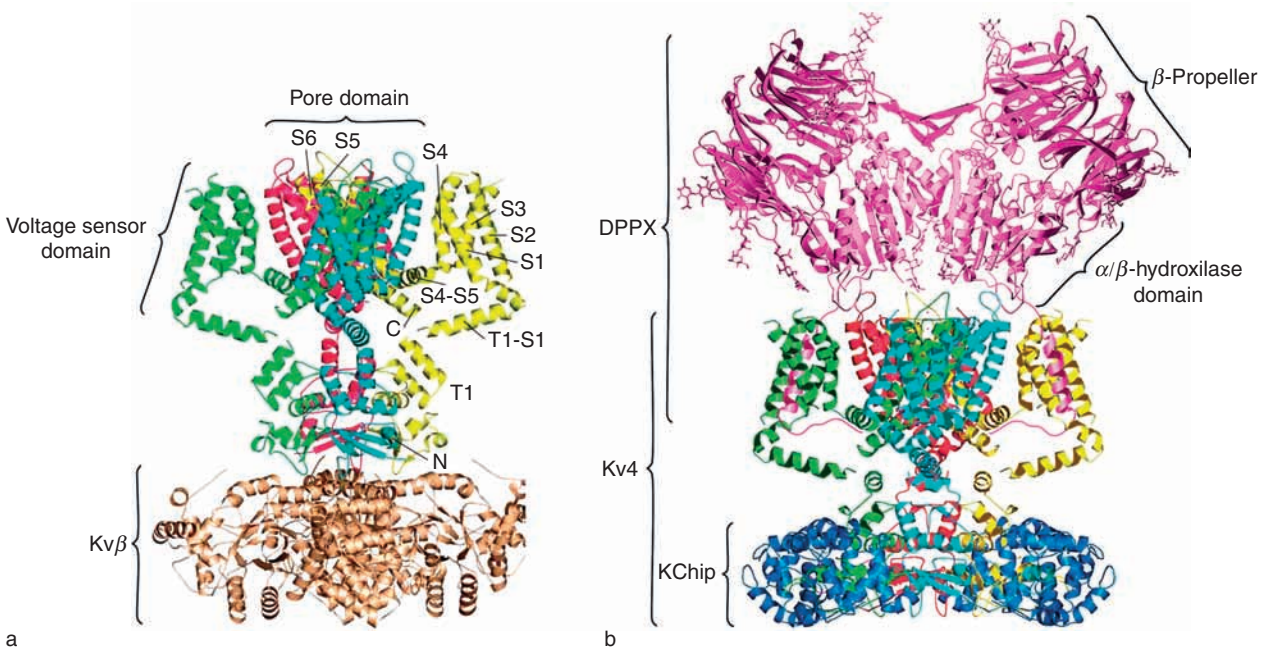

Figure 2 Comparison of the Kv1.2–Kvβ2 channel complex with a modeled Kv4.3–KChIP1-DPPX channel complex. (a) Side view of the $Kv_{1.2}$–Kvβ2 channel complex derived from the crystal structure obtained by MacKinnon and colleagues (Long SB, Campbell EB, and Mackinnon R (2005) Crystal structure of a mammalian voltage-dependent Shaker family K^+ channel. *Science* 309: 897–903 and Long SB, Campbell EB, and Mackinnon R (2005) Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science* 309: 903–908). Image obtained from Wang H, Yan Y, Liu Q, et al. (2007) Structural basis for modulation of Kv4 K^+ channels by auxiliary KChIP subunits. *Nature Neuroscience* 10: 32–39. The four Kv1.2 subunits are labeled in cyan, yellow, red, and green. The four Kvβ2 subunits are labeled in wheat color. The upper third shows the transmembrane part of the channel. The inverted tepee, formed by the pore domains (S5-P-S6) of the four Kv1.2 proteins, is in the center, and the voltage sensor domains (S1–S4) are lateral. The middle portion of the figure contains the T1 domains interacting with the Kvβ2 proteins at the bottom. (b) Side view of the Kv4.3-KChIP1-DPPX channel complex. This figure was obtained by adding molecules of DPPX (based on the crystal structure of the DPPX extracellular domain: obtained by Strop et al. (2004). Strop P, Bankovich AJ, Hansen KC, Garcia KC, and Brunger AT (2004) Structure of a human A-type potassium channel interacting protein DPPX, a member of the dipeptidyl aminopeptidase family. *Journal of Molecular Biology* 343: 1055–1065) to the model of the Kv4.3-KChIP1 channel complex by Wang H, Yan Y, Liu Q, et al., Structural basis for modulation of Kv_4 K^+ channels by auxiliary KChIP subunits. *Nature Neuroscience* 10: 32–39. The four Kv4.3 subunits are labeled in cyan, yellow, red, and green. The four KChIP1 proteins interacting with the channel T1 domains, are labeled in blue. The transmembrane domains of DPPX are shown interacting with the membrane spanning helices of the voltage sensor domains. Only two DPPX molecules are shown to facilitate viewing, but we believe that the channel complex is likely to include four DPPX proteins. Each DPPX subunit includes an α/β hydrolase domain (close to the membrane and a β propeller domain shown on top and to the side. The extracellular domains of two DPPX proteins interact forming a dimer.

(Figure 2). The cytoplasmic end of the S6 helix is flexible around a 'hinge' provided by the conserved sequence Pro-X-Pro. The movement of the voltage sensor in response to changes in membrane potential displaces the S4–S5 linker affecting its interactions with the S6 helix, allowing the cytoplasmic end of S6 to move and thus constrict or dilate the inner opening of the pore.

The Fast-Activating Kv Subfamilies

All members of the Kv1–Kv4 subfamilies form functional homomultimeric channel complexes when expressed in heterologous expression systems, whereas the members of the Kv5, Kv6, Kv8, and Kv9 subfamilies are 'silent' pore-forming subunits that do not form functional homomultimeric channels by themselves and must co-assemble with subunits of the Kv2 subfamily to express functional channels.

Subunits of the Kv1, Kv3, and Kv4 subfamilies can form heteromeric channels with other members of the same subfamily in heterologous expression systems. Heteromeric channels have properties that are intermediary to those of homomeric channels, although some properties of one subunit may dominate in the heteromultimeric channel. It is not clear whether the two members of the Kv2 subfamily can associate with each other to form heteromeric channels, but both can form heteromeric channels with the silent Kv5, Kv6, Kv8, and Kv9 proteins.

Most Kv proteins are expressed in the nervous system, and there is considerable overlap of expression of multiple Kv subunits of the same subfamily in many neurons, suggesting that native channels in many cells might be heteromultimers (Table 1). Heteromultimerization largely increases the number of channels with distinct functional properties that can be generated by Kv subunits (the number of different

Table 1 Kv subunits[a]

Channel subunit	*Associated proteins*	*Expression pattern*	*Channel function*	*Associated pathology*
Kv1 subfamily Kv1.1 Gene symbol: KCNA1[b]	Kvβs (see **Table 2**) PDZ domain-containing proteins such as PSD95 and SAP97 SNAP25, CASPR2	Tissue: strong in brain and sensory neurons; weak in atria, aorta, skeletal muscle, retina, smooth muscle. CNS: RNA – widely expressed in rodent CNS. Protein – predominantly in axons, axon initial segment (AIS), and presynaptic terminals. Juxtaparanode of myelinated fibers. Staining of neuronal somata and proximal dendrites in some neurons.	Kv1.1 expresses TEA, DTX, and 4-AP sensitive fast-activating low-voltage-activating delayed rectifier channels in heterologous expression systems. In neurons: Kv1.1 subunits contribute to the formation of dendrotoxin (DTX-I)-sensitive K^+ channels mediating the DTX-sensitive current I_D. I_D in neuronal somata and AIS regulates neuronal subthreshold excitability and spike initiation as in neurons in the MNTB, where it ensures that the timing and pattern of AP firing is preserved across the relay synapse. In presynaptic terminals, it regulates the excitability of the terminal and prevents aberrant excitation.	Mouse: knockout – frequent epileptic seizures, short life span. Point mutation V408A – model of episodic ataxia type-1 (EA1). Spontaneous 11-bp deletion – megencephaly. Human: Kv1 mutations – EA1 with myokymia, sometimes associated with epilepsy; autoantibodies to several Kv1s found in autoimmune disorders such as Morvan's syndrome and limbic encephalitis.
Kv1.2 Gene symbol: KCNA2	$K_V\beta$s. PDZ domain-containing proteins Small GTP-binding protein RhoA. Protein tyrosine kinases phosphorylate and thereby suppress Kv1.2 channels. RhoA is a necessary component in this process and mediates M1 receptor inhibition of Kv1.2 by tyrosine phosphorylation. CASPR2	Tissue: brain, vascular smooth muscle, sensory neurons. CNS: very widely expressed. Kv1 heteromers often contain Kv1.2. Protein predominantly in axons and terminals. Pattern highly overlapping with other Kv1s but not identical. For example, in hippocampus, cerebellum, and spinal cord. Present in axon initial segment; juxtaparanode in myelinated axons Also in proximal dendrites and soma in some neurons.	In heterologous expression systems, Kv1.2 expresses 4-AP and DTX-sensitive low-voltage-activating delayed rectifier channels with slower activation kinetics than other Kv1s. Kv1.2 subunits contribute to forming DTX-sensitive channels in neuronal somata, axons, and terminals (see Kv1.1). Likely to contribute to I_D with slower kinetics. Regulation of subthreshold excitability in striatal medium spiny neurons.	Human: autoantibodies to several Kv1s found in autoimmune disorders such as Morvan's syndrome and limbic encephalitis.

Kv1.3 Gene symbol: KCNA3	$K_V\beta$s PDZ domain-containing proteins PSD95, SAP97, and hDLg KCNE4 (MirP3) – an inhibitory subunit to Kv1.1 and Kv1.3 β_1 integrins	Tissue: brain, lymphocytes (type n channel), macrophages, microglia, kidney, testis and spermatozoa, osteoclasts. CNS: prominent in olfactory bulb, cerebellum, parallel fibers, deep nuclei, Purkinje cell somata; weak in hippocampus and stratum lucidum. Protein: mitral and granule cells of the olfactory bulb.	In heterologous expression systems, Kv1.3 expresses fast-activating 4-AP-sensitive channels with prominent 'slow' P/C-type inactivation and prominent cumulative inactivation. Blocked by several toxins with some specificity. Regulates neuronal excitability, possibly in heteromers with other Kv1s. Mediates the type n K^+ channel in lymphocytes, required for lymphocyte activation. Kv1.3 channels contribute to the neuronal killing ability of microglia.	Kv1.3−/− mice have a 1000- to 10 000-fold lower threshold for odor detection and increased ability to discriminate between odorants. Mice weigh significantly less than control littermates, suggesting regulation of energy homeostasis as well as peripheral insulin sensitivity. Multiple sclerosis: blockers of Kv1.3 are being considered for MS treatment.
Kv1.4 Gene symbol: KCNA4	$K_V\beta$s PDZ domain-containing proteins PSD95, SAP97, and hDLg α-actinin-2 KChaP σ receptor	Tissue: brain, heart atria, low levels aorta, skeletal muscle PNS: altered expression after axotomy; sole Kv1 in small-diameter neuron nociceptors CNS: mRNA prominent throughout brain, much weaker in cerebellum and brain stem. Strong in hippocampus, striatum, thalamus, and CX. Widespread Kv1.4 immunoreactivityin brain, probably in axons.	In heterologous expression systems, Kv1.4 expresses fast-activating and -inactivating 4-AP-sensitive channels. Neurons: contributes inactivation to DTX-sensitive K^+ channels. Regulates spike repolarization and spike broadening during repetitive activity in mossy fiber terminals. Heart: mediates slow component of I_{to} in atria.	Kv1.4−/− mice show increased spontaneous seizures. Patients with myasthenia gravis produce autoantibodies to Kv1.4.
Kv1.5 Gene symbol: KCNA5	Kvβs Src tyrosine kinase PDZ-containing proteins Kchap Fyn α-actinin-2 Caveolin SAP97	Tissue: heart > skeletal muscle > brain > lung > kidney. Pancreatic β cells, microglia, Schwann cells. Brain: Purkinje cells, DCN. Hippocampus: somatodendritic.	In heterologous expression systems, Kv1.5 expresses fast-activating 4-AP-sensitive delayed rectifier channels. 4-AP-sensitive component of I_{Kslow} in mouse ventricle and I_{Kur} (ultrarapid-activating K^+ current) in human atrium. Potential use in management of atrial fibrillation via blockade of I_{Kur}.	

Continued

Table 1 Continued

Channel subunit	*Associated proteins*	*Expression pattern*	*Channel function*	*Associated pathology*
Kv1.6 Gene symbol: KCNA6	Kvβs CASPR2	Tissue: brain, smooth muscle, PNS, oligodendrocytes, astrocytes, AV node. CNS: widely distributed in cerebral cortex and hippocampus in axons. Purkinje cells (somatodendritic, not axonal) and in various olfactory and amygdaloid structures.	In heterologous expression systems, Kv1.6 expresses fast-activating subthreshold-activating DTX- and 4-AP-sensitive delayed rectifier channels. Kv1.6 subunits contribute to forming DTX-sensitive channels in neuronal somata, axons, and terminals (see Kv1.1).	Kcna6^{tm1Lex} lexicon genetics Homozygous mutation results in increased thermal nociceptive threshold (MGI). Autoantibodies to several Kv1s found in autoimmune disorders such as Morvan's syndrome, neuromyotonia, and limbic encephalitis.
Kv1.7 Gene symbol: KCNA7		mRNA: expressed in pulmonary arteries, heart, skeletal muscle, pancreatic islet cells, but not in brain.	In heterologous expression systems, Kv1.7 expresses fast-activating 4-AP-sensitive channels. Two channel isoforms with different functional characteristics; the long form, Kv1.7L, inactivates faster than the short isoform, Kv1.7S, predominantly due to an N-type related mechanism, which is impaired in the short form. Kv1.7L, but not mKv1.7S, is regulated by the cell's redox state.	
Kv1.8 Gene symbol: KCNA10	KCNA4B: increases current magnitude and sensitivity to cAMP	RNA expressed in kidney, renal blood vessels, heart, brain (weakly), and aorta	Voltage- and cyclic nucleotide-gated K^+ channel. In heterologous expression systems, Kv1.8 expresses slowly activating high-voltage-activating delayed rectifier channels. KCNA10 may facilitate renal proximal tubular sodium absorption by stabilizing cell membrane potential. Its presence in endothelial and vascular smooth muscle cells suggests that it also regulates vascular tone.	

Kv2 subfamily			
Kv2.1 Gene symbol: KCNB1	Kv5–Kv6, Kv8–Kv9 subunits KChaP (binds to N-terminus of Kv2.1) Fyn SH2 domain SNAP25	Tissue expression: brain, atria, ventricle, skeletal muscle, olfactory epithelium, retina, kidney, SCG (also Kv2.2); spinal motor neurons; lung, pancreatic β cells, pulmonary artery. Brain: widely expressed, but no detailed distribution of Kv2.1 or Kv2.2 mRNA or protein reported. Protein localized in clusters in somatic and proximal dendritic membrane. Little neuropil staining.	In heterologous expression systems, Kv2.1 expresses slowly activating high-voltage-activating delayed rectifier channels that are blocked by high millimolar concentrations of TEA. Kv2 channels mediate I_K, the TEA-sensitive, 4-AP-'insensitive' delayed rectifier K^+ current recorded in many neurons. Potential therapeutic targets for diabetes due to their prominent expression in pancreatic β cells.
Kv2.2 Gene symbol: KCNB2	Kv5–Kv6, Kv8–Kv9 subunits mKvβ4 associates with Kv2.2 and enhances expression level KChaP	Tissue: brain, tongue epithelium; SCG, sympathetic neurons; GI/mesenteric artery smooth muscle. Brain: neuropil (axons), maybe in some somas.	In heterologous expression systems, Kv2.2 expresses slowly activating high-voltage-activating delayed rectifier channels that are blocked by high millimolar concentrations of TEA. Kv2 channels mediate I_K, the TEA-sensitive, 4-AP-insensitive delayed rectifier K^+ current recorded in many neurons.
Kv5.1 Gene symbol: KCNF1	Kv2.1 and Kv2.2 subunits	Rat and human brain, heart, skeletal muscle, liver, cardiac myocytes. Brain: limited mRNA expression in rat brain – deep cortical layers, amygdala, and medial habenular nucleus.	Silent subunit, interacts with Kv2 subunits and modifies channel properties: decreases current levels, slows down activation, produces negative shifts in inactivation, and slows down deactivation.
Kv6.1 Gene symbol: KCNG1	Kv2.1 and Kv2.2 subunits	Human: brain, placenta, skeletal muscle. SA cardiac nodal cells.	Silent subunit, interacts with Kv2 subunits and modifies channel properties: produces negative shifts in activation and inactivation and slows down deactivation.
Kv6.2 Gene symbol: KCNG2	Kv2.1 and Kv2.2 subunits	Primarily rat and human heart.	Silent subunit, interacts with Kv2 subunits and modifies channel properties: modifies kinetics and voltage dependence and confers submicromolar sensitivity to antiarrhythmic drug propafenone.

Continued

Table 1 Continued

Channel subunit	*Associated proteins*	*Expression pattern*	*Channel function*	*Associated pathology*
Kv6.3 Gene symbol: KCNG3	Kv2.1 and Kv2.2 subunits	Human, rat: brain, testis, thymus, adrenal glad, small intestine, kidney, lung, pancreas, ovary, colon. Brain: widely expressed in rat. RNA most prominent in cortex, hippocampus, striatum, amygdala.	Silent subunit, interacts with Kv2 subunits and modifies channel properties: slows down inactivation and deactivation.	
Kv6.4 Gene symbol: KCNG4	Kv2.1 and Kv2.2 subunits	Human: brain, liver, small intestine, colon.	Silent subunit, interacts with Kv2 subunits and modifies channel properties: produces negative shifts in activation and inactivation, slows down deactivation, and speeds up activation.	
Kv8.1 Gene symbol: KCNV1	Kv2.1 and Kv2.2 subunits	Hamster brain: mRNA very widespread.	Silent subunit, interacts with Kv2 subunits and modifies channel properties: produces negative shift in inactivation, slows down activation and inactivation.	
Kv8.2 Gene symbol: KCNV2	Kv2.1 and Kv2.2 subunits	Human: lung, liver, kidney, pancreas, spleen, thymus, prostate, testis, ovary, colon.	Silent subunit, interacts with Kv2 subunits and modifies channel properties: produces small negative shifts in activation and inactivation and slight acceleration of activation.	
Kv9.1 Gene symbol: KCNS1	Kv2.1 and Kv2.2 subunits	Human: brain, prostate, testis by PCR. Mouse: mainly in brain. mRNA in mouse brain: Kv9.1 and Kv9.2 subunits show very similar expression – olfactory bulb, cortex, hippocampus, habenula, basolateral amygdala, cerebellum	Silent subunit, interacts with Kv2 subunits and modifies channel properties: reduces currents, produces negative shifts of activation and inactivation. Slows down activation of Kv2.1. Slows down deactivation of Kv2.1 but speeds up deactivation of Kv2.2.	
Kv9.2 Gene symbol: KCNS2	Kv2.1 and Kv2.2 subunits	Mouse: mainly in brain Pulmonary artery Distribution in mouse brain, (see Kv9.1).	Silent subunit, interacts with Kv2 subunits and modifies channel properties: reduces currents, produces negative shifts of inactivation. Slows down deactivation of Kv2.1.	

Kv9.3 Gene symbol: KCNS3	Kv2.1 and Kv2.2 subunits	Widespread expression in human by PCR.	Silent subunit, interacts with Kv2 subunits and modifies channel properties: produces negative shifts of activation and inactivation, slows down deactivation, and speeds up recovery from inactivation.	
Kv3 subfamily				
Kv3.1 Gene symbol: KCNC1		Broadly expressed in somas, axons, and terminals in brain and spinal cord structures but in specific cell populations, including many (but not only) GABAergic neurons and brain stem sensory neurons. Dog atrium Molecular basis of canine atrial $I_{Kur,d}$. T lymphocytes.	In heterologous expression systems, Kv3.1 expresses high-voltage-activating, fast-activating and -deactivating delayed rectifier channels that are very sensitive to TEA and 4-AP. Important for spike repolarization and high-frequency firing of brain stem auditory neurons and fast-spiking GABAergic interneurons. Regulate spike repolarization and duration in presynaptic terminals. Spike repolarization in nodes of Ranvier in central myelinated axons. Type I channel in T lymphocytes.	Kv3.1−/− mice exhibit impaired motor skills, hyperactivity, and sleep loss. Kv3.1/Kv3.3 double knockout mice show ataxia, tremor, myoclonus, and hypersensitivity to ethanol.
Kv3.2 Gene symbol: KCNC2		Tissue expression: mRNA mainly in brain. Prominently in thalamus. GABAergic interneurons of the neocortex, hippocampus, and caudate; globus pallidus, SNR, sensory nuclei in brain stem. Protein in terminal fields of thalamocortical axons. Somata and axons of cortical interneurons and other mRNA-expressing neurons. Pancreatic β cells; Schwann cells and (also Kv3.1b); smooth muscle.	In heterologous expression systems, Kv3.2 expresses high-voltage-activating, fast-activating and -deactivating delayed rectifier channels that are very sensitive to TEA and 4-AP. In heteromeric complexes with Kv3.1 important for spike repolarization and high-frequency firing of fast-spiking GABAergic interneurons and in spike repolarization and duration in GABAergic presynaptic terminals. Activity modulated by protein kinase A in heterologous cells and in neurons.	Kv3.2−/− mice show alterations in cortical EEG and increased seizure susceptibility consistent with an impairment of cortical inhibitory mechanisms.

Continued

Table 1 Continued

Channel subunit	*Associated proteins*	*Expression pattern*	*Channel function*	*Associated pathology*
Kv3.3 Gene symbol: KCNC3		Expressed predominantly in the CNS. Broadly expressed in somas, axons, and terminals in brain and spinal cord structures but in specific cell types. In many (but not all) GABAergic neurons throughout forebrain and midbrain, and in mossy fiber axons of dentate granule cells in hippocampus. Very strongly expressed in cerebellum, particularly in Purkinje cell somas, axons, and dendrites. Expression in brain stem sensory and reticular neurons and motor neurons.	In heterologous expression systems, Kv3.3 expresses high-voltage-activating, fast-activating and -deactivating channels that are very sensitive to TEA and 4-AP and inactivate fast in *Xenopus* oocytes but slowly and variably in mammalian transfected cells. Regulation of Na^+ and Ca^{2+} spike repolarization and spontaneous firing frequency of Purkinje cells. May contribute to Kv3.1-containing channels in auditory brain stem (see Kv3.1). Channels probably containing Kv3.3 and Kv3.4 regulate transmitter release at motor end plates, and channels containing Kv3.1 and Kv3.3 transmitter release from cerebellar parallel fiber terminals.	Kv3.3−/− mice exhibit cerebellar dysfunction, altered Purkinje cell discharges, and lack harmaline-induced tremor. Kv3.1/Kv3.3 double knockout mice display severe ataxia, tremor, myoclonus, and hypersensitivity to ethanol. Human: mutations in Kv3.3 cause spinocerebellar ataxia SCA13.
Kv3.4 Gene symbol: KCNC4	MiRP2 forms potassium channels in skeletal muscle with $K_V3.4$.	Prominent in skeletal muscle. Brain: weakly in hippocampal granule cells, Purkinje cells, and pontine nuclei.	In heterologous expression systems, Kv3.4 expresses high-voltage-activating, fast-activating and fast-inactivating channels that are very sensitive to TEA and 4-AP. In association with MirP2, forms low-voltage-activating potassium channels that regulate skeletal muscle resting potential. Channels probably containing Kv3.3 and Kv3.4 regulate transmitter release at motor end plates.	Human: an R83H missense mutation in MiRP2 that diminished the effects of MiRP2 on Kv3.4 is associated with periodic paralysis.
Kv4 subfamily				
Kv4.1 Gene symbol: KCND1	Not investigated	Rat brain: mRNA very weakly expressed in rat. Expressed in human brain, uterus.	In heterologous expression systems, Kv4.1 expresses transient K^+ currents blocked by millimolar concentrations of 4-AP with slower inactivation kinetics than Kv4.2 and Kv4.3.	

			Kv4 channels underlie the fast transient outward current I_{to} in cardiac myocytes and most of the subthreshold-operating somatodendritic A-type K current in mammalian neurons (I_{SA}). However, Kv4.2 and Kv4.3 appear to be the main Kv4 subunits in brain and heart.	
Kv4.2 Gene symbol: KCND2	KChIPs, DPPX (DPP6), and DPP10	mRNA: brain – expressed in selective brain areas, sometimes overlapping with Kv4.3 (e.g., thalamus), but often in areas not expressing Kv4.3 (predominant form in the caudate putamen, pontine nucleus, and several nuclei in the medulla); granule cells in the OB; CA1 pyramidal cells in the hippocampus; reciprocal expression with Kv4.3 in the cerebellar granule cell layer.	In heterologous expression systems, Kv4.2 expresses low-voltage-activating, fast-inactivating K^+ currents that recover quickly from inactivation and are blocked by millimolar concentrations of 4-AP resembling neuronal I_{SA}. Native-like properties are enhanced by association with associated proteins KChIPs and DPPs.	Mouse: the loss of Kv4.2 protein in hippocampal CA1 pyramidal neurons of Kv4.2 produces a near-complete elimination of I_{SA} from apical dendrites, resulting in an increase in the amplitude of backpropagating action potentials and an increase of concomitant Ca^{2+} influx, facilitating the induction of long-term potentiation.
		Protein: predominantly dendritic expression in hippocampal pyramidal cells and cerebellar granule cells. Heart: rodent ventricular muscle.	Kv4 channels underlie the fast transient outward current I_{to} in cardiac myocytes and most of the subthreshold-operating somatodendritic A-type K current (I_{SA}) in mammalian neurons. I_{SA} regulates firing frequency in neurons and dendritic excitability.	In dorsal horn neurons, genetic elimination of Kv4.2 increases excitability, resulting in enhanced sensitivity to tactile and thermal stimuli. Furthermore, ERK-dependent forms of pain hypersensitivity are absent in Kv4.2−/− mice.
Kv4.3 Gene symbol: KCND3	KChIPs, DPPX (DPP6), and DPP10	Kv4.3 mRNA predominant Kv4 in cortical GABAergic interneurons, cerebellar Purkinje cells and dopaminergic neurons in substantia nigra pars compacta, restrosplenial cortex, superior colliculus, raphe, and amygdala; in cerebellar granule cells, predominant in posterior lobules. Protein: strong expression in dendrites. Heart: Kv4.3 main Kv4 gene in canine and human ventricle.	In heterologous expression systems, Kv4.3 expresses low-voltage-activating, fast-inactivating K^+ currents that recover quickly from inactivation and are blocked by millimolar concentrations of 4-AP resembling neuronal I_{SA}. Native-like properties are enhanced by association with associated proteins KChIPs and DPPs. Kv4.3 kinetics is somewhat slower than that of Kv4.2.	

Continued

Table 1 Continued

Channel subunit	*Associated proteins*	*Expression pattern*	*Channel function*	*Associated pathology*
			Kv4 channels underlie the fast transient outward current I_{to} in cardiac myocytes and most of the subthreshold-operating somatodendritic A-type K^+ current (I_{SA}) in mammalian neurons. I_{SA} regulates firing frequency in neurons and dendritic excitability. The pacemaker frequency of individual dopaminergic neurons, and hence the amount of dopamine release, correlates with I_{SA} levels, which in turn correlate with the concentration of Kv4.3 mRNA	
Kv7 (KCNQ) family				
Kv7.1 (also known as KvLQT1) Gene symbol: KCNQ1	KCNE1(MINK) KCNE2(MiRP1) KCNE3(MiRP2)	Heart, kidney, colon, pancreas, lung, pituitary, cochlea (marginal cells of stria vascularis), placenta. Exocrine pancreas.	Slow sustained current. Does not associate with other Kv7 subunits. Co-assembles with minK/KCNE1 to form the cardiac I_{Ks} channel involved in repolarization of cardiac action potentials and in epithelial cells of the inner ear, where they mediate K^+ secretion into the endolymph. Associates with KCNE3 to form leakage channels involved in intestinal chloride secretion.	Human: KCNQ1 loss-of-function mutations cause long QT syndrome (including Romano–Ward syndrome and Jervell and Lange-Nielsen syndrome (JLNS)). JLNS includes deafness; gain-of-function mutations in KCNQ1 cause atrial fibrillation and short QT syndrome. Associated with Beckwith–Wiedemann syndrome. Knockout mouse: model of JLNS.
Kv7.2 Gene symbol: KCNQ2	Calmodulin Kv7.3, Tubulin, PKA KCNE2	Expressed mainly in neural tissue: CNS, pituitary, SCG, thoracolumbar ganglia. Widely expressed in many neuronal populations also expressing Kv7.3. Highest levels of both Kv7.2 and Kv7.3 expression are found in the cerebral cortex, striatum, hippocampus, thalamus, reticular nucleus of thalamus, anterior olfactory nucleus, and red nucleus. Protein expression in neuronal somata, Nodes of Ranvier and axon initial segment.	Slow delayed rectifier in heterologous cells. Associates with Kv7.3 to form M-type channels. Highly TEA sensitive in homomultimeric channels. Sensitivity to TEA drops to millimolar concentrations when associated with other Kv7 subunits. Blocked by M-channel blockers XE991 and linopirdine. Activated by the M channel opener Retigabine.	Human: benign familial neonatal convulsions (BFNC) Mouse: KCNQ2-/- mice die soon after birth due to pulmonary atelectasis. KCNQ2+/- mice have increased seizure susceptibility.

Kv7.3 Gene symbol: KCNQ3	Kv7.2, Kv7.4, KCNQ5, calmodulin, tubulin, PKA KCNE2	See Kv7.2.	Slow delayed rectifier in heterologous cells. Associates with Kv7.2, Kv7.4, and Kv7.5 to form M-type channels. Blocked by M-channel blockers XE991 and linopirdine. Activated by the M channel opener retigabine.	Human: BFNC.
Kv7.4 Gene symbol: KCNQ4	Kv7.3	Brain stem auditory neurons (cochlear nuclei, nuclei of the lateral lemniscus, and inferior colliculus). Cochlea (outer hair cells (OHC)), inner hair cells (IHC), vestibular hair cells). Protein in basal membrane of OHC, in type I hair cells in vestibular organs, and in the calyx-like terminals on these cells.	Slow delayed rectifier in heterologous cells. Associates with Kv7.3 to form M-type channels. Blocked by M-channel blockers XE991 and linopirdine. Activated by the M-channel opener retigabine. In OHCs, currents are large and constitute the dominant currents at rest and during depolarization. Principal exit pathway for K^+ ions entering the cells through the transduction channels. In IHCs, the current is relatively small and it influences mainly resting membrane properties.	Human: mutations in KCNQ4 cause autosomal dominant nonsyndromic deafness type 2 (DFNA2). Mouse: degeneration of OHCs and progressive hearing loss. DFNA2 might be caused by a slow degeneration of OHCs resulting from chronic depolarization.
Kv7.5 Gene symbol: KCNQ5	Kv7.3 Calmodulin	CNS, SCG, sympathetic ganglia, overlapping with Kv7.2 and Kv7.3 in several brain regions (cortex, hippocampus, and striatum). Also expressed in skeletal muscle. Protein: pyramidal cells and interneurons in neocortex and hippocampus.	Slow delayed rectifier in heterologous cells. Associates with Kv7.3 to form M-type channels. Blocked by M-channel blockers XE991 and linopirdine.	

[a]For additional details and references: see *IUPHAR Ion Channel Compendium 2005*, Pharmacol Rev 57: 473–508 (2005) and the tables in our site: http://www.med.nyu.edu/rudylab, search gene using gene name at http://www.ncbi.nlm.nih.gov/sites/entrez; additional molecular information can also be obtained from the Jackson's laboratory Mouse Genome Informatics (MGI) site: http://www.informatics.jax.org.

[b]Per convention, human gene names use uppercase and mouse names lowercase (e.g., KCNA1 and Kcna1, respectively).

subunit combinations can be calculated from the equation $[p + (n\text{-}1)]!/[p!(n\text{-}1)]$, where p is the aggregation number (4), and n is the total number of different subunits that can heteromultimerize), predicting an enormous diversity of channels, which is even larger if one considers other factors that can also lead to functional diversification, such as interactions with diverse auxiliary proteins or posttranslational modifications.

Given the subfamily specific heteromultimerization, each Kv subfamily describes an independent functional unit. This emphasizes the need to discuss each subfamily as a group, as we do in this article. As described previously, the members of the Kv5, Kv6, Kv8, and Kv9 subfamilies must co-assemble with subunits of the Kv2 subfamily to form functional channels. Therefore, they are integral parts of Kv2 channels and are treated together with the members of the Kv2 subfamily. Members of the Kv1–Kv4 subfamilies are homologs of the *Shaker, Shab, Shaw*, and *Shal* genes in *Drosophila*, respectively. Therefore, the mammalian counterparts are sometimes referred to using the nomenclature in flies (e.g., Kv1 or Shaker genes or channels).

N- and C/P-Type Inactivation of Kv Channels

Voltage-gated K^+ channels (as well as most other voltage-gated channels except the HCN pacemarker channels) activate and open upon membrane depolarization. However, most Kv channels do not remain open, even if the depolarization remains constant, but undergo a process known as inactivation. The rates at which the channels inactivate vary by orders of magnitude depending on the channel type: Fast-inactivating channels mediating transient (or A-type) K^+ currents inactivate with time constants in the range of 10–100 ms. There are probably several mechanisms by which Kv channels inactivate, two of which have been well characterized: N- and P/C-type inactivation. N-type inactivation is usually fast and is produced by the cytoplasmic N-terminal sequence of some Kv subunits, which acts as a peptide that enters and occludes the channel's pore, blocking ion permeation. The following Kv pore-forming subunits – Kv1.4, Kv3.3, and Kv3.4 – as well as the Kv1 auxiliary proteins Kvβ1 and Kvβ3 contain such N-terminal inactivation domains, producing N-type inactivation of channels containing one or more of these subunits.

P/C-type inactivation is produced by a time-dependent conformational change that constricts the external mouth of the pore (a 'pore collapse'), although the molecular details are not completely understood. Many Kv channels undergo P/C inactivation, but the rates can differ by two or three orders of magnitude, from tens of milliseconds to seconds. The rate of P/C inactivation depends on the amino acid sequence of both the P domain and the S6 domain, and it can be affected by extracellular ions that interact with the channels pore.

Next, we discuss the physiological contributions and other salient features of each of the four groups of Kv_F proteins in mammalian species.

Kv1 (Shaker) Subfamily

This is the largest Kv subfamily, with eight genes identified to date (Table 1). However, we focus on the first six members of the subfamily (Kv1.1–Kv1.6). Kv1.7, the only Kv1 gene known to produce more than one protein isoform, is prominently expressed in cardiac and skeletal muscle, as well as in pancreatic islet cells, but not in the nervous system. Kv1.8 (better known as KCNA10), the most recently identified Kv1 gene, is intriguing because it is structurally and functionally related to both voltage-gated and cyclic nucleotide-gated channels. Although RNAse protection analysis has suggested that KCNA10 transcripts are present in brain, nothing is known about its function in neural cells.

Kv1.1–Kv1.6 subunits are components of voltage-gated K^+ channels, which overall tend to activate faster and at lower membrane potentials than Kv channels of the Kv2 and Kv3 subfamilies. (Several Kv1 channels also tend to activate at more negative membrane potentials than Kv4 channels; however, the voltage-dependence of Kv4 channels is shifted toward more negative voltages by interaction with their auxiliary proteins DPPX and DPP10.) They can conduct current at voltages below threshold for spike generation and therefore contribute to the subthreshold properties of neurons. They are also highly sensitive to 4-aminopyridine (4-AP), with IC_{50}s of <1 mM for all Kv1.1–Kv1.6 homomeric channels. Most Kv1 proteins are widely and prominently expressed in neurons throughout the nervous system, except for Kv1.5, Kv1.7, and Kv1.8 (Table 1). It is likely that most, if not all, neurons in the central nervous system (CNS) and peripheral nervous system (PNS) express several Kv1 proteins. This suggests that native Kv1 channels in many cells might be heteromeric complexes of several Kv1 proteins.

Neuronal Kv1.1–Kv1.6 channels are complexes of Kv1 and Kvβ subunits

Pongs and colleagues discovered a group of proteins now known as Kvβ subunits that associate with Kv1 pore-forming subunits. Kvβ proteins identified to date are products of three genes, Kvβ1–Kvβ3, each

of which produces several isoforms by alternative splicing at the N-terminal region (Table 2). They have a mass of approximately 40 kDa and lack putative transmembrane domains, potential glycosylation sites, or leader sequences, suggesting that they are cytosolic proteins. This has been confirmed by structural analysis. Kvβ proteins bind 1:1 to the T1 domain located in the N-terminal region of the channel pore-forming subunits, creating channel complexes of eight polypeptides (Figures 1(b) and 2(a)). In fact, it is likely that most, if not all, Kv1 channels in neural cells are complexes that resemble the structure shown in Figure 2a, containing four Kv1 pore-forming subunits and four Kvβ associated proteins.

All Kvβ subunits share a conserved core domain, which has sequence and structural similarity to aldo-keto reductases. In addition, the Kvβ1 and Kvβ3 subunits contain an extra, isoform-specific, N-terminal sequence that resembles the inactivation domain found in Kv1.4, which produces N-type inactivation.

A key function of the Kvβ subunits is to act as chaperones during channel biosynthesis, facilitating proper trafficking and expression at the cell surface. In addition, by virtue of their N-type inactivation domain, Kvβ1 and Kvβ3 proteins can accelerate the inactivation of Kv1.4-containing channels or induce fast inactivation in otherwise non-inactivating or slowly inactivating Kv1 channels (except in Kv1.3, for which inactivation is not affected by Kvβ1 and Kvβ3, and Kv1.6, for which inactivation is not affected by Kvβ1). Kv1.6 proteins contain an N-type inactivation prevention domain that prevents inactivation of channels containing one or more Kv1.6 subunits. On the other hand, Kvβ2 proteins, which lack N-terminal inactivation domains, accelerate inactivation only when Kv1.4 subunits form part of the Kv1 channel complex, probably via interactions with the intrinsic Kv1.4 inactivation domain. Thus, the Kvβ subunits diversify the inactivating properties of Kv1 channels. In some cases, the Kvβ subunits may also shift the voltage dependence of activation in the hyperpolarizing direction.

The amino acid sequence and structural similarity with aldo-keto reductases has remained an intriguing feature. One study has provided data suggesting that Kvβ proteins may actually function as redox

Table 2 Kvβ accessory subunits[a]

Subunit	*Spliced isoforms (with accession numbers)*	*Expression pattern*[b]	*Function*
Kvβ1 Gene symbol: KCNAB1	Kvβ1.1:NM_172159 Kvβ1.2:NM_003471 Kvβ1.3:NM_172160	Widely distributed in brain tissue with cell-specific distribution. Co-expressed with Kv1.1 and Kv1.4 in cortical interneurons, in the hippocampal perforant path and mossy fiber pathways, and in the globus pallidus and substantia nigra. Kv1.1 and Kv1.4 and, to a lesser extent, Kv1.2 and Kv1.6 are detected in channel complexes containing Kvβ1. Most of these also contain Kvβ2. More prominent in young animals (before P21).	Structural component of Kv1 channels. Facilitate trafficking and expression at cell surface. Accelerate inactivation of Kv1 channels. Cell metabolic state sensor.
Kvβ2 Gene symbol: KCNAB2	Kvβ2.1:NM_003636 Kvβ2.2:NM_172130	Widely distributed in brain tissue; more abundant in adult brain than Kvβ1. Co-expressed with Kv1 subunits in the juxtaparanode in myelinated fibers. The bulk of K^+ channel complexes in brain containing Kv1.1, Kv1.2, Kv1.4, and Kv1.6 contain Kvβ2.	Structural component of Kv1 channels. Facilitate trafficking and expression at cell surface. Only accelerate inactivation of Kv1.4 channels. Cell metabolic state sensor.
Kvβ3 Gene symbol: KCNAB3	Kvβ3.1: (human, rat) NM_004732 Kvβ3.2 (first named Kvβ4): (mouse) U65593	More restricted pattern of expression than Kvβ1 and Kvβ2. No high-resolution localization studies. Most prominent in cerebellum. Also expressed in neocortex.	Structural component of Kv1 channels. Facilitate trafficking and expression at cell surface. Accelerate inactivation of Kv1 channels. Cell metabolic state sensor.

[a]Data from Jackson's laboratory Mouse Genome Informatics website, http://www.informatics.jax.org, and from Coetzee WA, Amarillo Y, Chiu J, et al. (1999) Molecular diversity of K^+ channels. *Annals of the New York Academy of Sciences* 868: 233–285; and Pongs O, Leicher T, Berger M, et al. (1999) Functional and molecular aspects of voltage-gated K^+ channel beta subunits. *Annals of the New York Academy of Sciences* 868: 344–355.

[b]Data derived largely from Rhodes KJ, Strassle BW, Monaghan MM, Bekela-Arcuri Z Matos MF, and Trimmer JS Association and colocalization of the Kvβ1 and Kvβ2 β-subunits with Kv1 α-subunits in mammalian brain K^+ channel complexes. *Journal of Neuroscience* 17: 8246–8258.

enzymes. It was observed that Kvβ2 can use 4-cyanobenzaldehyde as a substrate and reduce it to an alcohol. The catalytic reaction was found to be very slow, suggesting that this might not be the primary function of Kvβ proteins. Instead, it has been suggested that the Kvβs endow Kv1 channels with a sensor module to detect the metabolic state of the cell. This hypothesis is based on the observation that application of the substrate to the cytoplasmic side of the cell influenced the rate of inactivation of Kv1.4–Kvβ2 complexes.

All Kvβ subunits are predominantly expressed in brain tissue. Kvβ1 and Kvβ2 are widely distributed but display cell-specific distributions. Kvβ2 proteins are much more abundant, and a much larger portion of the total brain pool of Kv1-containing channel complexes is associated with Kvβ2 rather than with Kvβ1. Complexes containing Kvβ1 often also contain Kvβ2. Kvβ3 shows a more restricted pattern of expression (Table 2).

Kv1.8 (KCNA10) interacts specifically with the auxiliary subunit KCNA4B, which is unrelated to the Kvβ subunits. When co-expressed with KCNA10, KCNA4B increases KCNA10 current and also increases its sensitivity to activation by cAMP.

In addition to association with Kvβ subunits, most Kv1 proteins are able to interact with scaffolding, membrane-associated putative guanylate kinases (or MAGUKs), which are PDZ domain-containing proteins such as PSD-95 and SAP97. The association is important for the clustering of the channels in microdomains at the neuronal surface, and it is mediated by the interaction of the PDZ domains with the C-terminus of Kv1 proteins, which has the sequence ETDV or ETEV (except for Kv1.8). Mutation of this C-terminal sequence abolishes binding to the MAGUKs and prevents channel clustering.

Functional properties of Kv1 channels

Kv1 channels can be delayed rectifiers or fast-transient A-type channels, depending on their subunit composition. Delayed rectifier channels are slowly inactivating voltage-gated potassium-selective ion channels so called because potassium flux occurs in a voltage- and time-dependent manner with a delay after a depolarizing voltage step, which is longer than the time required to activate fast inward sodium currents (hence, delayed) and normally flows in one direction (outward; hence, producing rectification of the current–voltage relation). The term A-type K^+ channels or 'A channels' is applied to voltage-gated K^+ channels that inactivate fast during depolarization (usually with time constants of <100 ms in mammals). Among the Kv1 subunits prominently expressed in brain (Kv1.1–Kv1.6), Kv1.4 subunits are the only ones that contain an N-terminal inactivation domain capable of producing fast N-type inactivation. Kv1.4 subunits contribute fast inactivation to heteromultimeric channels. The rate of inactivation of a heteromeric channel depends on the number of inactivating subunits in the channel complex. In addition, as described previously, several Kvβ subunits have N-terminal inactivation domains which confer fast N-type inactivation to Kv1 channels, including those with slow or no inactivation. Therefore, the inactivation rate of native channels will depend on subunit composition, in terms of all the subtypes of pore-forming subunits in the channel complex as well as the type of Kvβ present. Furthermore, inactivation mechanisms can be the subject of posttranslational modifications, and these modulations can change inactivation properties of native channels.

K^+ channels containing Kv1.4 proteins, with or without Kvβ subunits, recover from inactivation very slowly, resulting in cumulative inactivation during repetitive activity. It has been suggested that this mediates activity-dependent spike broadening in hippocampal mossy fiber boutons (MFBs), the large presynaptic terminals of the axons of the granule cells of the dentate gyrus. The action potential in these terminals is very brief during low-frequency stimulation but is prolonged up to threefold during high-frequency stimulation. Prolongation of the presynaptic spike results in an increase in the number of Ca^{2+} ions entering the terminal per action potential, and this in turn leads to potentiation of evoked excitatory postsynaptic currents at MFB–CA3 pyramidal cell synapses.

Many Kv1 channels inactivate slowly via P/C-type inactivation. This inactivation is prominent in Kv1.3 channels. Due to the slow rates of recovery from inactivation, P/C inactivation of Kv1.3 channels produces pronounced cumulative inactivation during repetitive stimulation.

Kv1 channels mediate dendrotoxin-sensitive K^+ currents

Three Kv1 subunits – Kv1.1, Kv1.2, and Kv1.6 – express homomultimeric channels which are blocked by the mamba snake toxins dendrotoxin I (DTX-I) and α-dendrotoxin (α-DTX). In heteromeric channels, a single Kv1.1, Kv1.2, or Kv1.6 protein is sufficient to confer DTX sensitivity. Hence, Kv1 channels are the mediators of the D-type current or I_D. This current, originally described in CA1 pyramidal neurons, is of great interest due to its effects on the subthreshold properties of neurons. Given the heteromeric association of Kv1 proteins, the widespread expression of Kv1.1, Kv1.2, and Kv1.6 proteins in the nervous system (Table 1), and the total number of Kv1 pore-forming subunits and Kvβs, the possible combinations of subunits producing channels

containing one of the Kv1 subunits conferring DTX sensitivity is extremely large, resulting in a large diversity of I_D with variations in a number of properties including inactivation rates. For example, a channel containing Kv1.1, Kv1.4, and Kvβ1, a combination that has been detected in hippocampus, will produce a fast-inactivating DTX-sensitive current. Inactivation would be much slower if the channel contained Kvβ2 instead of Kvβ1. Similarly, Kv1.2 homomultimers activate more slowly than other Kv1.1–Kv1.6 channels (Table 1); thus, the presence of Kv1.2 subunits will lead to I_D with slower activation kinetics.

Mutations in the gene encoding Kv1.1 cause episodic ataxia type 1 and epilepsy

Consistent with their role in regulating neuronal subthreshold excitability, mutations in the human KCNA1 gene encoding Kv1.1 proteins produce a disease with perturbed excitability – episodic ataxia type 1 (EA1) with myokymia, an autosomal dominant disease affecting central and peripheral nerve function, sometimes associated with epilepsy. Similarly, Kv1.1 knockout mice have severe, often lethal, epilepsy. Although the cellular basis for the heightened seizure susceptibility of the mice is not understood, this phenotype also emphasizes the importance of Kv1 channels in regulating neuronal excitability.

Kv1 channels in axons

Another feature of Kv_1 channels of considerable interest is their localization in axons (Figure 3), where they are concentrated in specific microdomains. The T1 domain (Figures 1 and 2) has been found to be essential for the axonal targeting of Kv1 proteins. This domain is also the major site of association between Kv1 pore-forming and Kvβ subunits, and it has been shown that one of these, Kvβ2, is important for the targeting of Kv1 channels to the axon. This function depends on its ability to associate with the microtubule plus-end tracking protein EB1.

Kv1.1 and Kv1.2 channels are concentrated in a distal part of the axon initial segment, where they co-localize with ankyrin G and may contribute to regulating action potential initiation (see Figure 6). Kv1 channels are also prominent in presynaptic terminals, where they might be specifically localized to the preterminal axon. They reduce the excitability of the nerve terminal, thus preventing aberrant firing at the terminal (produced by reflection of the action potential from the terminal to the axon) and, hence, aberrant transmitter release. However, in two cases in which it has been investigated, the calyx of Held terminal in the auditory brain stem and the terminals of cortical GABAergic interneurons, the Kv1 current does not seem to contribute to the repolarization of the action potential.

In myelinated axons, Kv1.1, Kv1.2, and Kv1.6 are present in the juxtaparanodal region (Figure 6). Although their role here is not clear, it has been suggested that they may function to dampen re-entrant excitation. This juxtaparanodal localization is affected by demyelination. Myelin loss results in the exposure and dispersion of these juxtaparanodal K^+ channels – effects that in turn contribute to the neurological symptoms associated with demyelination.

Kv2 (Shab) Subfamily

There are two known members of this subfamily, Kv2.1 and Kv2.2. Both are widely expressed in brain tissue and are believed to be present in many, if not all, principal cells and interneurons throughout the brain. In contrast to the other Kv_F families, in which heteromultimer formation is prevalent, immunohistochemistry shows that the two Kv2 subunits produce distinct subcellular staining patterns even in cells that express both proteins, suggesting that Kv2 subunits may not form heteromultimeric channels *in vivo*. Furthermore, immunoprecipitation experiments failed to show association of the two subunits in rat superior cervical ganglion (SCG) cells, which express both Kv2.1 and Kv2.2, or in HEK-293 cells transfected with cDNAs encoding both subunits. However, Blaine and Ribera's studies using a dominant negative Kv2.2 subunit suggest that the two Kv2 proteins are capable of associating with each other.

Kv2 subunits do heteromultimerize with subunits of the Kv5, Kv6, Kv8, and Kv9 subfamilies, which, as mentioned previously, are silent pore-forming subunits that do not form conducting channels by themselves. These associations change the voltage dependence and kinetics of the channels (Table 1). For example, co-expression of Kv9.1 with Kv2.1 in *Xenopus* oocytes slows the rate of activation, causes a hyperpolarizing shift in the voltage dependence of activation and inactivation, and changes sensitivity to tetraethylammonium (TEA).

Kv2 channels mediate the delayed rectifier current I_K

Kv2 channels have a low sensitivity to 4-AP (Table 1). They are blocked by extracellular, 10–20 mM TEA and are thought to mediate the 'TEA-sensitive, 4-AP-insensitive' delayed rectifier K^+ current recorded in many types of neurons known as I_K.

In heterologous expression systems, Kv2 channels can display different voltage-dependence and kinetic properties, including the degree of inactivation, depending on which silent pore-forming subunits

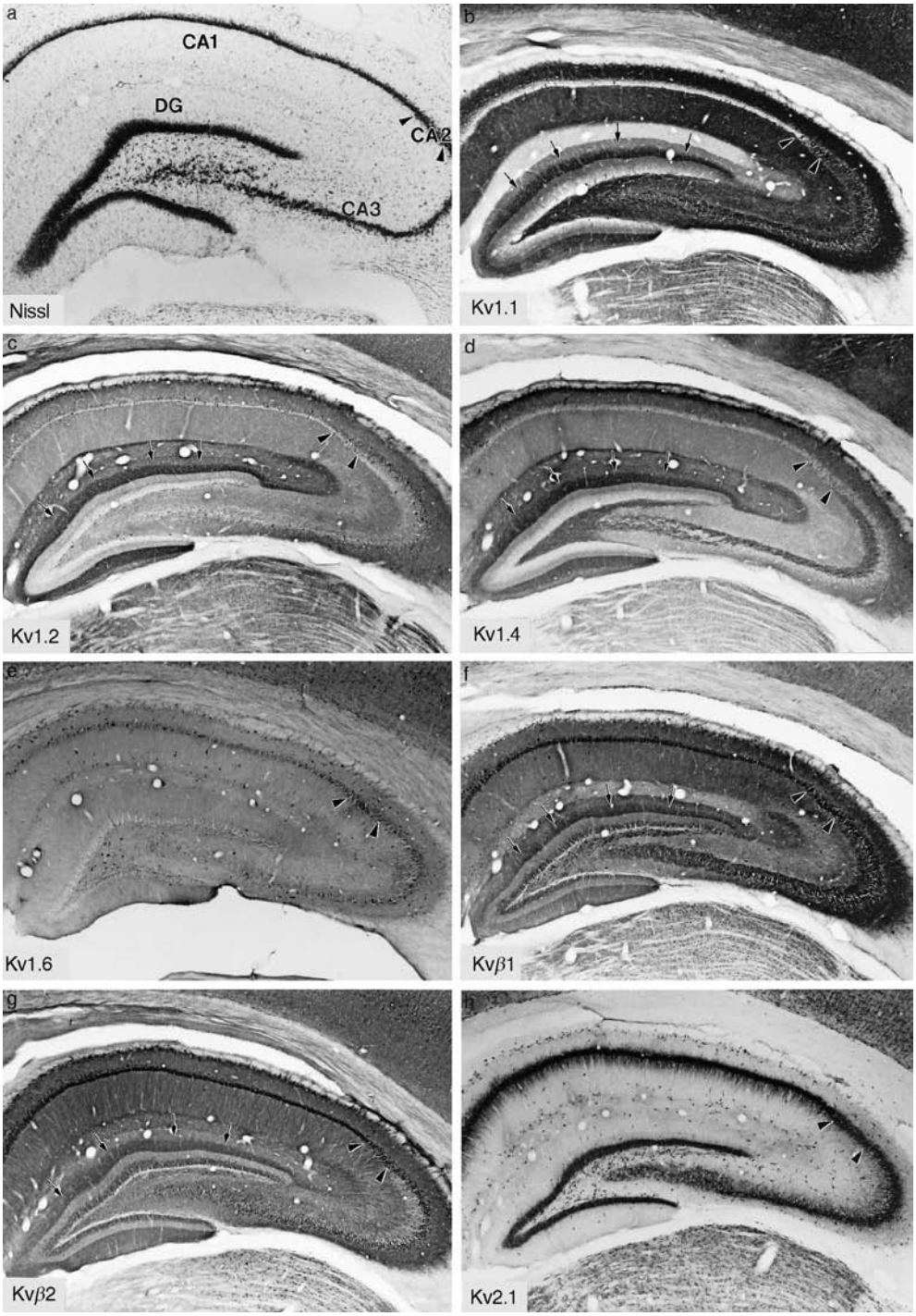

Figure 3 Contrasting expression patterns of Kv1 and Kv2.1 subunits in the hippocampal formation. In contrast to Kv2.1, Kv1 protein expression is predominantly observed outside the layers containing neuronal somata. Lesioning studies indicate that the observed staining is predominantly associated with local and projecting axons and nerve terminals. The patterns of expression of each Kv1 protein are specific, but there is significant overlap between different subunits. Reproduced from Rhodes KJ, Strassle BW, Monaghan MM, et al. (1997) Association and colocalization of the Kvbeta1 and Kvbeta2 beta-subunits with Kv1 alpha-subunits in mammalian brain K^+ channel complexes. *Journal of Neuroscience* 17: 8246–8258.

form part of the channel complex (Table 1). However, the molecular composition of Kv2 native channel complexes and its relationship to I_K diversity in neurons and other cell types remain to be determined. There is also little knowledge about the functional roles of these currents in nerve cells. In one study, it was suggested that because these channels activate and deactivate slowly compared to other Kv_F

channels, they do not contribute to the repolarization of single action potentials in cortical pyramidal cells but instead regulate excitability and Ca^{2+} influx during repetitive firing. However, Malin and Nerbonne found that elimination of Kv2 channels in tonic firing sympathetic neurons increased spike duration. The channels have also been implicated in mediating apoptosis in neurons and are potential therapeutic targets for diabetes due to their prominent expression and activity in pancreatic β cells.

Modulation of Kv2.1 channels

Kv2.1 channels are found in high-density clusters in the somata and proximal dendrites, whereas Kv2.2 subunits typically present diffuse localization in somata and neuropil. Kv2.1 clusters are organized in specific subcellular patterns, such as opposite astrocytic processes or inhibitory synapses in cortex and hippocampus and opposed to cholinergic synapses on spinal motor neurons. The mechanisms of clustering of Kv2.1 channels have been the subject of intense investigation.

Kv2.1 channels are extensively phosphorylated when expressed in HEK 293 cells and in neurons. The phosphorylation state of Kv2.1 subunits influences channel properties, including localization and voltage dependence. Furthermore, phosphorylation, localization, and channel properties are dynamically regulated by neuronal activity. Following kainate-induced seizures in mice, Kv2.1 channels in pyramidal neuron membranes were found to be dephosphorylated relative to control animals and no longer clustered. In cultured hippocampal pyramidal neurons, glutamate exposure also caused dephosphorylation of Kv2.1 subunits, continuous localization (as opposed to clustered), and a hyperpolarizing shift in the voltage dependence of I_K. Similar modulations of Kv2.1 expression and activity were found following hypoxia and chemical ischemia *in vivo* and *in vitro* and cholinergic stimulation *in vitro*. Hyperpolarizing shifts of voltage dependence, as observed concomitant with dephosphorylation, would cause an increase in I_K, which in turn is expected to decrease membrane excitability. Considering their strong and widespread expression, dephosphorylation of Kv2.1 subunits may be an effective and widespread homeostatic mechanism by which neurons are able to alter excitability in response to their chemical environment.

Kv3 (Shaw) Subfamily

There are four known Kv3 genes in mammals – Kv3.1–Kv3.4. In contrast to the other Kv subfamilies, the four Kv3 genes undergo extensive alternative splicing, resulting in the generation of more than ten Kv3 isoforms with different intracellular C-terminal sequences. In heterologous expression systems, Kv3 proteins express fast-activating voltage-gated channels with unique properties. Compared to the currents expressed by other Kv_F subunits, Kv3 currents become apparent at more depolarized potentials (more positive than −20 to −10 mV; Table 1) and deactivate very quickly upon repolarization. The rightward shifted voltage dependence and fast deactivation rates of Kv3 channels are likely related properties, probably reflecting instability of the channel's open state at voltages near the resting potential.

In heterologous expression systems, Kv3.1 and Kv3.2 subunits mediate slowly inactivating delayed rectifier-type currents, whereas Kv3.4 proteins produce fast-inactivating A-type currents. In *Xenopus* oocytes, Kv3.3 currents are inactivating with time constants on the order of tens of milliseconds, but in mammalian cell lines the inactivation of Kv3.3 currents is slow and variable. The reason for this discrepancy is unclear; it has been suggested that it might be the result of differential usage of a starting methionine preceding or following an N-inactivation domain. Differences in the posttranslational modulation of Kv3.3 proteins in these two heterologous expression systems could also explain the differences in inactivating properties. Indeed, the degree of inactivation of Kv3.3 and Kv3.4 currents in *Xenopus* oocytes is affected by oxidation and protein kinase C (PKC)-mediated phosphorylation.

Little is known about the function of the alternatively spliced C-termini. In the cases in which it has been explored, the currents expressed in heterologous expression systems by the isoforms of each Kv3 gene are very similar. Preliminary studies suggest that the C-termini may confer isoform-specific regulation by second messenger signaling systems and targeting to distinct neuronal compartments.

Functional roles of Kv3 channels

Three of the four known Kv3 genes (Kv3.1–Kv3.3) are prominently expressed in specific neuronal populations throughout the CNS (Figure 4). Kv3.4 transcripts are weakly expressed in a few neuronal types in the brain, usually in neurons also expressing other Kv3 genes (hippocampal granule cells, cerebellar Purkinje cells, and neurons in the pontine nucleus), whereas they are much more abundant in skeletal muscle and sympathetic neurons. Nevertheless, Kv3.4 proteins might be important components of Kv3 channels in some neurons, even when expressed at low levels, since a single Kv3.4 subunit is capable of enabling fast-inactivating properties to a Kv3

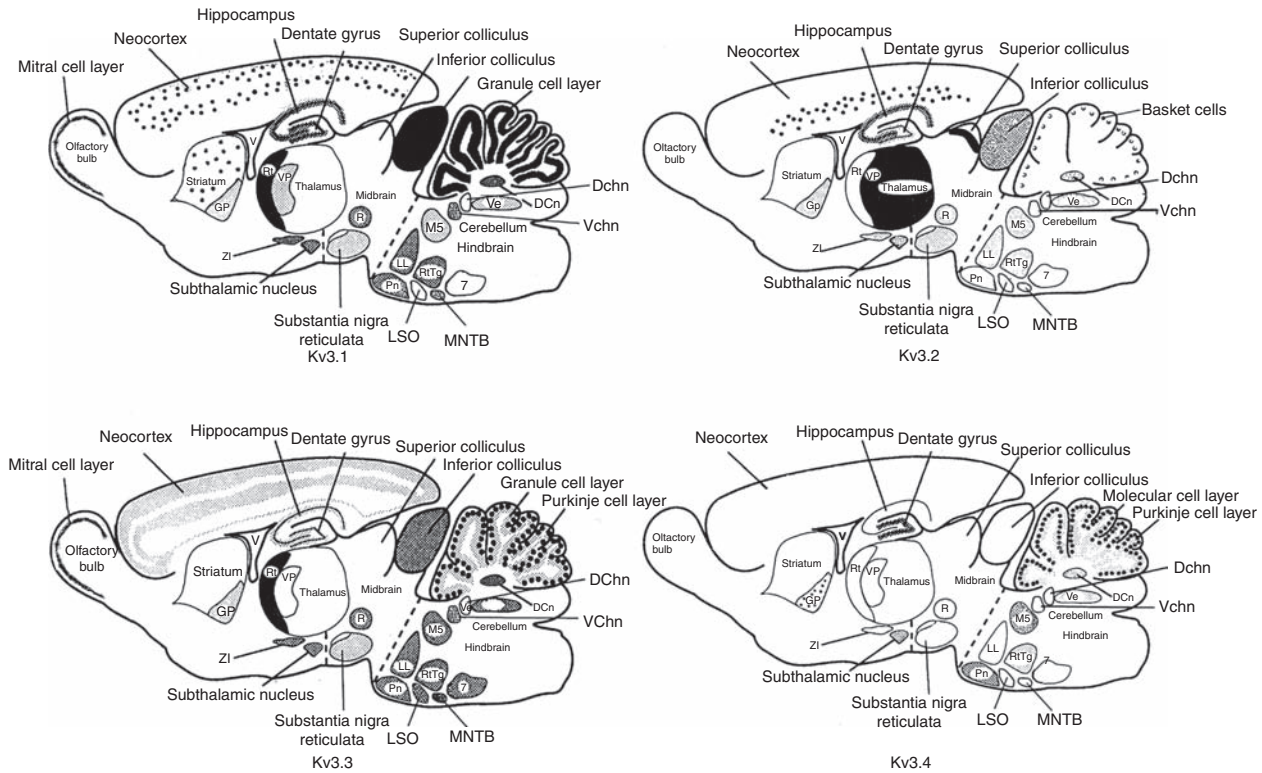

Figure 4 Specific but overlapping expression of Kv3 mRNAs in rodent brain. Diagrams illustrating the patterns of expression of Kv3 mRNAs obtained using *in situ* hybridization by Weiser et al. Darkness indicates level of expression. Reproduced from Rudy B, Chow A, Lau D, et al. (1999) Contributions of Kv3 channels to neuronal excitability. *Annals of the New York Academy of Sciences* 868: 304–343.

tetrameric channel. In skeletal muscle, Kv3.4 subunits assemble with MinK-related peptide 2 (MiRP2) to form subthreshold, voltage-gated potassium channels that regulate the resting potential of the muscle cell. An R83H missense mutation in MiRP2 that diminished the effects of MiRP2 on Kv3.4 has been associated with a form of periodic paralysis.

There are no specific blockers of Kv3 channels that can be used to study Kv3 currents in native cells and to investigate their functional roles. BDS toxins ('blood depressing substance') from the sea anemone *Anemonia sulcata* were initially reported to specifically block Kv3.4 channels but were later shown to block other Kv3 channels with similar affinity. These toxins are gating modifiers that shift the voltage dependence of the channels in the depolarizing direction. They tend to produce only partial channel block at physiological membrane potentials and modify channel properties. Typically, Kv3 currents have been isolated using the classical K^+ channel blockers TEA and 4-AP. These are nonspecific K^+ channel blockers; however, Kv3 channels are very sensitive to these drugs and hence can be useful to isolate native Kv3 currents. In particular, low ($\leq$1 mM) TEA blocks Kv3 channels nearly completely, as well as Kv1.1 homomultimeric channels, BK Ca^{2+}-activated K^+ channels, and Kv7.2 (KCNQ2) channels. Specific toxins that block these non-Kv3 TEA-sensitive channels (dendrotoxin, charybdotoxin or iberotoxin, and linopirdine, respectively) can be used to inhibit these other channels and isolate Kv3 currents or Kv3-mediated effects. However, caution should be exercised in pursuing this method since 1 mM TEA could block a significant amount of the current mediated by channels with a higher IC_{50} for TEA (e.g., Kv2s) if they are very abundant in a particular preparation.

The positively shifted voltage dependence of activation and fast activation/deactivation kinetics of Kv3 channels bestow these channels with the ability to efficiently and specifically accelerate the repolarization of the action potential. Kv3 currents are activated during the repolarizing phase of the action potential and are quickly deactivated following the spike. Hence, when present in sufficient amounts, Kv3 channels facilitate action potential repolarization and dictate action potential duration without significantly compromising action potential threshold, rise time, or magnitude and, more important, without contributing repolarizing current that would increase the duration of the refractory period (in contrast to voltage-gated K^+ channels activating more negatively and deactivating more slowly).

Kv3 channels are particularly prominent in neurons that have the ability to fire sustained trains of

action potentials at high frequency or to follow high-frequency stimulation (Table 1), and they have been shown to facilitate high-frequency firing in cortical fast-spiking GABAergic interneurons, cerebellar Purkinje cells, and neurons in the subthalamic nucleus, globus pallidus, and auditory brain stem. This role is likely the result of their function in spike repolarization. By increasing the rate of spike repolarization and keeping action potentials brief, Kv3 currents reduce the amount of Na^+ channel inactivation that occurs during the action potential and minimize the activation of K^+ channels with slow kinetics. Moreover, the large, fast afterhyperpolarization produced by the Kv3 current can speed up recovery from Na^+ channel inactivation and Kv3 channel deactivation. By accelerating the recovery of Na^+ channel inactivation and by deactivating quickly, Kv3 currents efficiently restore resting conditions and prepare the membrane for a second spike soon after the first spike, thus minimizing the duration of the refractory period and allowing high-frequency firing. It has been suggested that by regulating the ability to follow high-frequency stimuli in auditory neurons in the brain stem, these channels contribute to preserving the fidelity and timing of firing codes and thus the computation of sound features, such as frequency, intensity, and localization in space.

Kv3 gene products are also found in neurons that are not known to fire at high frequencies, such as magnocellular neurosecretory neurons and starburst cells in the retina, suggesting that Kv3 channels may have other physiological roles. Some of these functions of Kv3 channels may still be related to their ability to specifically maximize fast action potential repolarization and the consequences of this function: brief action potentials with quick restoration of membrane properties following activity and limiting calcium entry. In starburst neurons, which normally do not spike, it has been proposed that Kv3 channels provide a voltage-dependent shunt that limits the amplitude of neuronal depolarization and might be critical to the ability of these cells to mediate inhibition of ganglion cells, which depends on the direction of movement of visual stimuli.

Kv3 channels and neurotransmitter release

In addition to their localization in neuronal somata, Kv3 channels are also prominently expressed in axons and presynaptic terminals of several neuronal populations, suggesting roles in regulating neurotransmitter release (Figure 6). This function has been investigated in the basket terminals of fast-spiking interneurons in the cerebral cortex, in the calyx of Held (the large presynaptic terminals from the ventral cochlear nucleus bushy cells innervating the neurons in the MNTB), in motor nerve terminals, and in cerebellar parallel fiber terminals. In these synapses, Kv3 channels are key determinants of the repolarization of the local action potential, keeping action potentials brief, limiting Ca^{2+} influx and, hence, neurotransmitter release probability. As a result of these effects, Kv3 channels in presynaptic terminals contribute to the short-term dynamics of synaptic contacts and to maintaining the speed and fidelity of synaptic transmission.

Mutations in Kv3.3 channels cause spinocerebellar ataxia

Mutations in the human Kv3.3 gene (KCNC3) have been identified as the cause of a spinocerebellar ataxia (SCA13), a disease that can also include cognitive defects, suggesting cerebellar and extracerebellar dysfunction. Two missense mutations have been described; both mutations are likely to perturb the role of Kv3.3 channels in the repolarization of action potentials in Purkinje cells (and other neurons) and change their output characteristics.

Knockout mice have been generated for the three Kv3 genes that are prominently expressed in CNS neurons (Kv3.1–Kv3.3). Kv3.1 and Kv3.3 knockout mice show alterations in motor behavior, including increased locomotor activity and sleep loss in Kv3.1 knockout mice; impaired gait and reduced motor performance associated with abnormal Purkinje cell discharges and abnormal olivocerebellar system properties in Kv3.3 knockout mice; and, in the case of the Kv3.1 and Kv3.3 double mutant, severe ataxia, tremor, and myoclonus. Kv3.2 knockout mice have altered EEG rhythms, increased cortical excitability, and enhanced seizure susceptibility, perhaps resulting from alterations in GABAergic control of cortical structures.

The Kv4 (Shal) Subfamily

There are three Kv4 genes in mammals: Kv4.1–Kv4.3. The Kv4.1 and Kv4.2 genes produce a single product each, but the Kv4.3 gene encodes for two splice variants via alternative splicing: Kv4.3S and Kv4.3L (short and long; Table 1). Two of these genes, Kv4.2 and Kv4.3, are prominently expressed in the rodent CNS. Kv4.1 is expressed very weakly and only in a few neuronal populations, such as the auxiliary olfactory bulb, granule cells of the olfactory bulb, hippocampus, dentate gyrus, and cerebellar cortex, and possibly in Purkinje cells. In contrast to the Kv1 and Kv3 subfamilies, for which there is extensive overlap expression of multiple subfamily members in the same neurons, in the case of Kv4.2 and Kv4.3 transcripts, there is a differential, and sometimes reciprocal, pattern of expression, yet there is overlap expression in several neuronal populations (Figure 5).

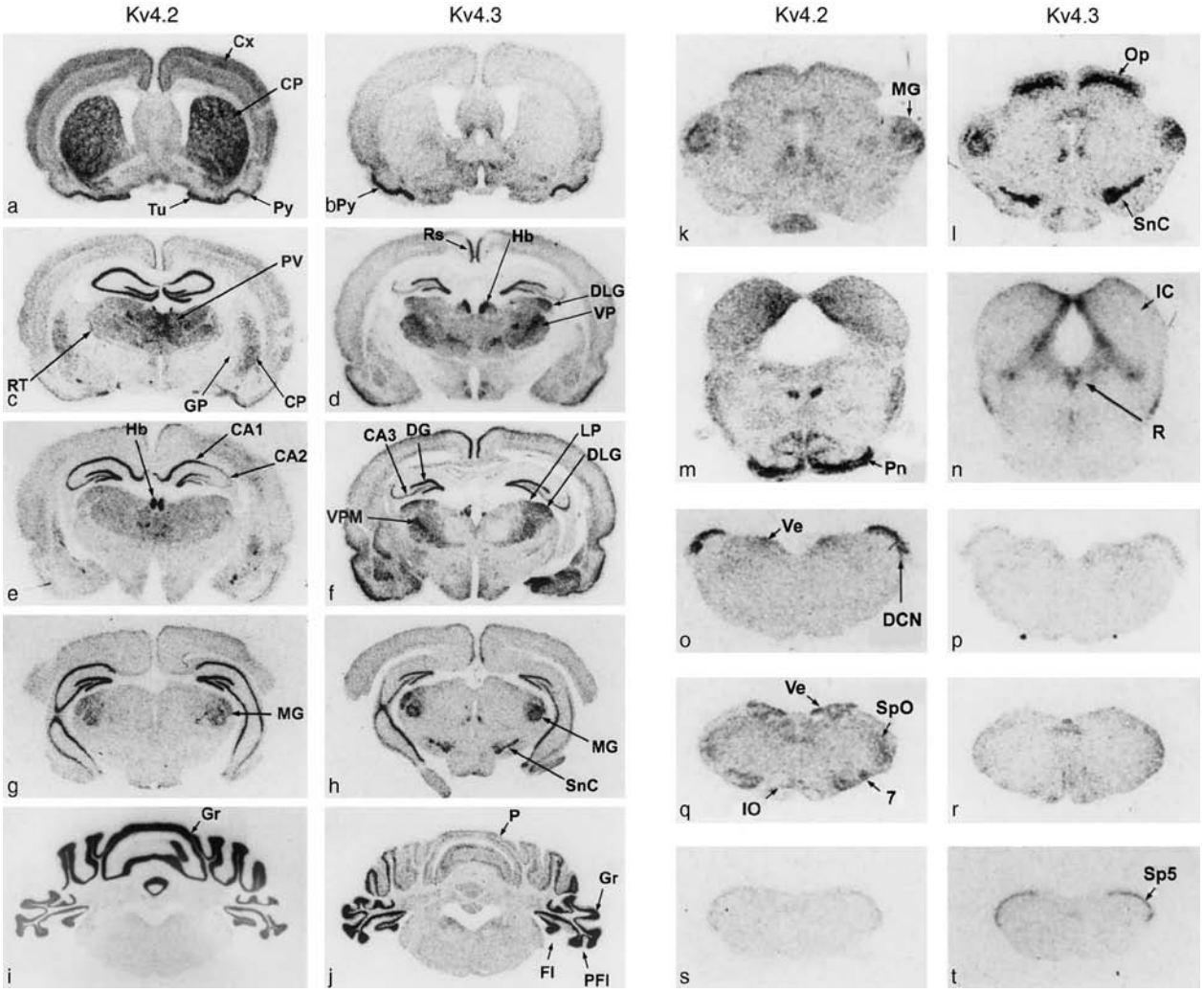

Figure 5 Pattern of expression of Kv4.2 and Kv4.3 mRNAs in rat brain. Kv4.2 and Kv4.3 gene expression overlaps in certain areas, but in others expression of one of the two predominates (e.g., Kv4.2 in the striatum, hippocampal CA1 area, and cerebellar granule cells in anterior lobules, and Kv4.3 in SNC, superior colliculus, Purkinje cells, and cerebellar granule cells in posterior lobules). Reproduced from Serodio P and Rudy B (1998) Differential expression of Kv4 K^+ channel subunits mediating subthreshold transient K^+ (A-type) currents in rat brain. *Journal of Neurophysiology* 79: 1081–1091.

Kv4 proteins produce fast-inactivating channels in heterologous expression systems and are believed to underlie fast-inactivating or A-type K^+ currents (I_A), including the fast transient outward current I_{to} in cardiac ventricular myocytes, which mediates the initial repolarization of the action potential and is considered a target for arrhythmia treatment, as well as most of the subthreshold operating somatodendritic A-type K^+ current in mammalian neurons (I_{SA}). In neurons, the rapid, transient activation of these channels in the subthreshold range of membrane potentials causes delayed excitation, influences spike repolarization, and determines the duration of the interspike interval. The currents are essential in regulating excitability in many neuronal types. By regulating the duration of interspike interval, they pace the frequency of repetitive firing. Furthermore, Kv4.2 and Kv4.3 proteins are expressed prominently in dendrites. Consistent with this pattern, Johnston and colleagues observed that there is a gradient of I_{SA} channels in CA1 hippocampal pyramidal neurons with maximum density in distal dendrites. These dendritic channels regulate action potential backpropagation into the dendritic tree, the integration of synaptic inputs, and the establishment of long-term potentiation (LTP). Utilizing Kv4.2 knockout mice, it has been confirmed that these functions are mediated by a Kv4.2-dependent I_{SA}.

Kv_4 channels use novel inactivation mechanisms

Kv_4 channels are characterized by having fast inactivation and fast recovery from inactivation. In

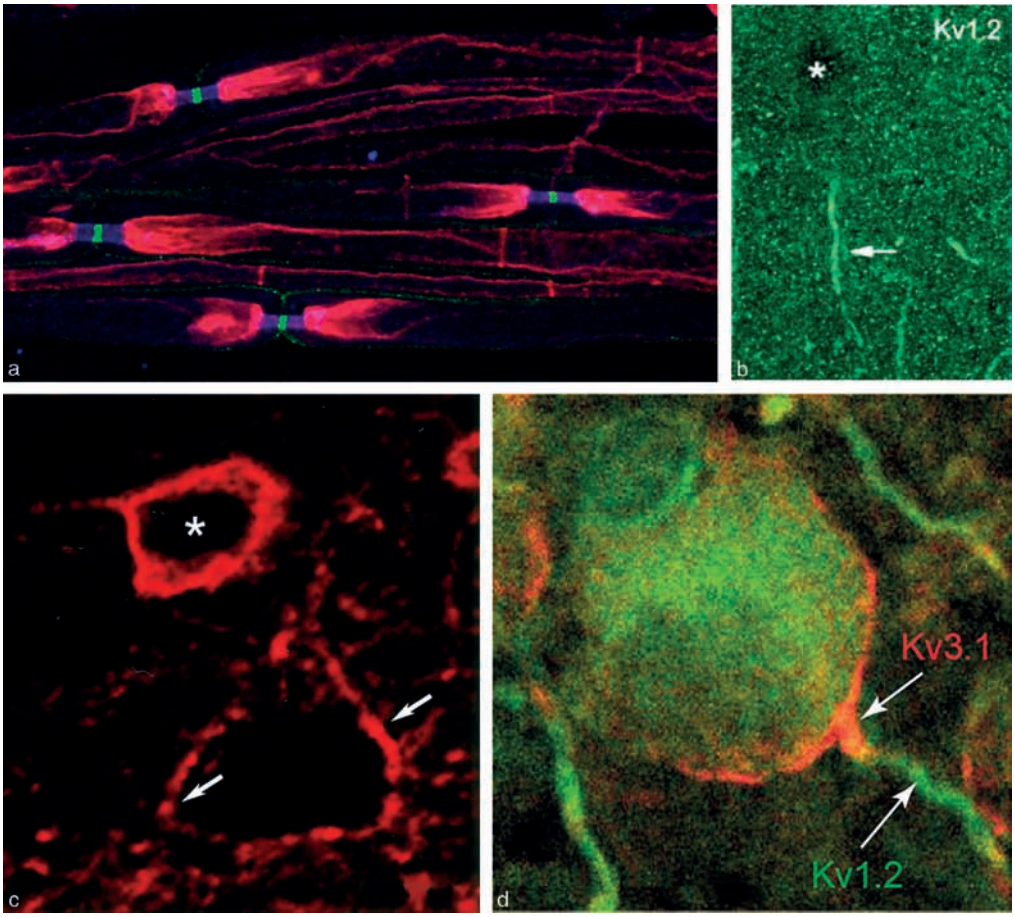

Figure 6 Neuronal K_v proteins are localized to specific membrane microdomains. (a) Kv1.1 (red) in the juxtaparanodal region of myelinated axons. Green, Na^+ channels at the node Ranvier; blue, NCP1 at the paranode. (b) Kv1.2 in the axon initial segment of cortical pyramidal neurons (arrows). The cell body is indicated by an asterisk. (c) Kv3.2 in the somatic membrane of a cortical interneuron (asterisk) and the pericomatic (basket) terminals contacting a pyramidal cell. (d) Kv3.1 proteins in the calyx of Held terminal on an MNTβ neuron and Kv1.2 in the preterminal axon. (a) Reproduced from Bhat MA, Rias JC, Lu Y, et al. (2001) Axon–glia interactions and the domain organization of myelinated axons requires neurexin iv/Caspr/Paranodin. *Neuron* 30: 369–383. (b) Reproduced from Inda MC, DeFelipe J, and Muñoz A (2006) Voltage-gated ion channels in the axon initial segment of human cortical pyramidal cells and their relationship with chandelier cells. *Proceedings of the National Academy of Sciences of the United States of America* 103: 2920–2925. (c) Reproduced from Chaw A, Erisir A, Farb C, et al. (1999) $K^{(+)}$ channel expression distinguishes subpopulations of parvalbumin- and somatostatin-containing neocortical interneurons. *Journal of Neuroscience* 19: 9332–9345. (d) Reproduced from Dodson PD and Forsythe ID (2004) Presynaptic K^+ channels: Electrifying regulators of synaptic terminal excitability. *Trends in Neurosciences* 27: 210–217.

contrast to most voltage-gated K^+ channels, in which open-state inactivation predominates, Kv_4 channels seem to inactivate preferentially from closed states. This implies that modest membrane depolarizations that may not open the channels can render them unable to open in response to a subsequent stronger membrane depolarization. This can result in cumulative inactivation upon repetitive stimulation. Moreover, the equilibrium between the open state and a pre-open activated state does not favor residence in the open states even during depolarization. Therefore preferential closed-state inactivation produces most of the inactivation that occurs during a depolarizing pulse and the transient character of the Kv4 current. The molecular mechanisms of inactivation of Kv4 channels are poorly understood and do not seem to simply conform to the classical N- or P/C-type inactivation described previously.

The fast recovery from inactivation characteristic of Kv4 channels is also an essential feature of these channels given their role in repetitively firing cardiac and neuronal cells (Table 1). The mechanisms underlying this fast recovery are not understood.

Auxiliary subunits

Several proteins that associate with Kv4 proteins have been identified, including KChIPs (for Kv channel-interacting proteins), which are members of the neuronal calcium sensor family of cytosolic Ca^{2+}-binding

EF-hand proteins, and a family of single-pass integral membrane glycoproteins related to dipeptidyl aminopeptidases, known as DPPX (or DPP6) and DPP10.

KChIPs bind to the N-terminal region and the T1 domain of Kv4 proteins; facilitate the trafficking, subunit assembly, and surface expression of channel complexes; and modify the electrophysiological properties of Kv4 channels, which inactivate more slowly and recover from inactivation with faster kinetics in the presence of KChIPs. There are four known KChIP genes (KChIP1–4). All KChIPs exert similar effects on Kv4 channels except for the alternative spliced version KChIP4a: Kv4 channels containing KChIP4a lack fast inactivation. The effects of KChIPs on inactivation are likely mediated by the sequestration of the N-terminus of the Kv4 proteins in a groove on the surface of the KChIP molecule.

DPPX and DPP10 (DPPs) belong to a family of dipeptidyl-aminopeptidase enzymes that have amino acid replacements in the catalytic active site and lack enzymatic activity. Instead, they seem to have evolved to function as channel-associated proteins. The DPPX and DPP10 genes produce several N-terminal spliced isoforms. Products of each gene share the same putative transmembrane domain – a long C-terminal extracellular domain with an hydrolase fold and a β-propeller (**Figure 2b**) – but are divergent in the short intracellular N-terminal sequence. They increase the rate of inactivation of Kv4 currents, considerably decrease the time required for the currents to reach a maximum (time to peak), and increase the rate of recovery from inactivation. In addition, they produce large negative shifts in the conductance–voltage relation and in the voltage dependence of steady-state inactivation. Furthermore, like KChIPs, DPPs increase the trafficking of Kv4 proteins to the cell surface, contributing to the major increase in current magnitude observed in the presence of these auxiliary subunits. In addition, DPPX increases the single channel conductance of Kv4 channels by 50%, a factor that may also contribute to the observed increases in total current magnitude.

Antibodies to Kv4 proteins coprecipitate KChIP and DPPX proteins, suggesting that the native channel might be a ternary complex of all three types of subunits. Consistent with this view, the currents expressed in heterologous cells by Kv4 proteins in the presence of KChIPs and DPPs closely resemble the properties of native channels. As in neuronal I_{SA} channels, heterologously expressed channels containing Kv4 proteins and the auxiliary subunits inactivate and recover from inactivation quickly, and they cover the range of voltage dependencies seen *in vivo*. The diversity of Kv4 and auxiliary subunits contributes to I_{SA} diversity in native cells.

DPPX and DPP10 are related to DPPIV or CD26, a cell adhesion protein with dipeptidyl-aminopeptidase activity that has important roles in T cell activation, metabolism of peptide hormones, and cell adhesion. This raises the intriguing possibility that DPP proteins may confer targeting or cell adhesion properties to Kv4 channels through their homologous extracellular domains, which is reminiscent of the cell adhesion properties conferred on Na^+ channels by the extracellular immunoglobulin-like domain of their β subunits.

Modulation of Kv4 channel function by phosphorylation

Kv4.2 channels in hippocampal neurons are modulated by extracellular signal-regulated kinase (ERK)–mitogen-activated protein kinase (MAPK) and by protein kinase A (PKA) and PKC. Activation of these signaling systems reduces channel activity by producing an approximately15 mV depolarizing shift in the voltage dependence of activation of the I_{SA}. It is possible that the effects of PKA and PKC activation are mediated by a common downstream mechanism, the activation of ERK, since blockers of MEK, the kinase responsible for the activation of ERK, block I_{SA} modulation by reagents that increase PKA and PKC activity. These modulations have considerable physiological significance. In hippocampal neurons, a Rap1-mediated coupling of cAMP and p42/44MAPK participates in the regulation of the excitability of pyramidal cells, the early and late phases of LTP, and the storage of spatial memory. These effects are mediated at least in part by modulation of Kv4.2 activity.

Furthermore, it has been observed that in spinal cord dorsal horn neurons, I_{SA} is also modulated by ERKs, mediating central sensitization during inflammatory pain. In Kv4.2 knockout mice, I_{SA} in dorsal horn neurons is absent, and their excitability is enhanced. This was accompanied by elevated sensitivity to tactile and thermal stimuli. Moreover, the ERK-mediated modulation of excitability and the ERK-dependent forms of pain hypersensitivity were absent in Kv4.2−/− mice.

Kv4 channel function, as well as the trafficking and expression of functional channels at the plasma membrane, may also be modulated by other signaling pathways. In *Xenopus* oocytes, phorbol esters produce a potent inhibition of channel activity without shifts in voltage dependence. This is mediated by PKC activation, but the downstream targets of PKC activity remain to be characterized. On the other hand, CaMKII phosphorylation of Kv4.2 has been shown to be important for the expression of functional

channels at the cell membrane, both in heterologous cells and in hippocampal neurons.

Kv7 (KCNQ) Channels

There are five genes in this group: KCNQ1–KCNQ5. The proteins encoded by these genes constitute the Kv7 subfamily of pore-forming potassium channel subunits. Many publications also use the KCNQ nomenclature for the protein products; however, to be consistent with this and other articles in this encyclopedia, we use the Kv7 nomenclature when referring to the protein products of the KCNQ genes. Kv7 (Kv7.1–Kv7.5) channel proteins are homologous to the Kv_F proteins discussed previously (Figure 1) and have important roles in the nervous system as well as in other tissues. Of the 80 potassium channel genes in humans, approximately 10 are known to underlie human diseases. Four of the five known members of the KCNQ subfamily, however, are associated with well-characterized channelopathies causing cardiac arrhythmia, epilepsy, and hearing loss, highlighting the physiological significance of this group of channels.

Structural Features of Kv7 Pore-Forming Subunits

The Kv7 subfamily of voltage-gated K^+ channel genes encodes a group of K^+ channel pore-forming subunits with structural similarity to the Kv_F subunits (Figure 1), containing six membrane spanning domains (S1–S6), a positively charged S4 domain that acts as a voltage sensor, and a single pore-forming (P) loop with the K^+ channel signature sequence that forms the selectivity filter of the channel's pore. In contrast to the Kv_F subfamilies, the number of positively charged residues in the S4 domain varies among Kv7 proteins. Kv7.1 has four, whereas Kv7.2–Kv7.5 have six. Furthermore, the degree of overall sequence similarity between different Kv7 proteins is less than that observed within Kv1–Kv4 pore-forming subunits (as illustrated by the length of the Kv7 branches in the dendrogram in Figure 1).

It is expected that the Kv7 channel complex is a tetramer of Kv7 proteins with a structure resembling those shown for Kv1.2 and Kv4s in Figure 2.

In comparison to most of the Kv proteins discussed previously, Kv7 subunits have very long intracellular C-terminal tails, particularly Kv7.2 and Kv7.5. The C-terminal tail contains a highly conserved domain known as the 'A domain.' This domain and a less conserved sequence that follows it form a subunit interaction domain (SID) involved in subunit-specific assembly among Kv7 proteins. This feature distinguishes these proteins from the Kv_F subunits, in which subunit assembly depends largely on the T1 domain present in the N-terminus. The SID determines which Kv7 subunits can assemble to form tetrameric channels, and it contains two motifs of approximately 30 amino acid residues each exhibiting a high probability for coiled-coil formation known as TCC1 and TCC2 (for tetramerizing coiled-coil 1 and 2) with a variable linker region in between.

Data suggest that the TCC1 domains are required to form functional homomeric as well as heteromeric channels, at least for Kv7.2 and Kv7.3, and that the TCC2 domains facilitate transport of heteromeric channels to the plasma membrane.

Subunit Assembly of Kv7 Proteins

The five Kv7 proteins can form functional homomeric channels. Heteromultimer formation does occur among Kv7 subunits, but it seems to be restricted to certain combinations. For example, Kv7.1, which is expressed in the heart but not in the nervous system, does not associate with other Kv7 proteins. On the other hand, Kv7.3 can associate and form functional heteromeric channels with Kv7.2, Kv7.4, or Kv7.5, all of which are expressed in neural tissue. This specificity depends on the C-terminal SID described previously. This domain is sufficient to transfer assembly properties between Kv7.3 and Kv7.1 since Kv7.1 chimeras containing Kv7.3 SID fragments acquire the broad assembly properties of Kv7.3.

In native cells, Kv7.1 subunit tetramers are found in association with KCNE accessory proteins such as KCNE1 (MinK) and KCNE3. Kv7.2 and Kv7.3 proteins can also co-assemble with KCNE proteins.

Expression and Function of Kv7 Channels: The 'M' Current

Expression of Kv7 channels in heterologous systems yields an outwardly rectifying voltage-dependent K^+ current with slow kinetics of activation (in the range of hundreds of milliseconds to seconds) and deactivation and little or no inactivation. There is appreciable activation of the currents at voltages more positive than −60 mV; hence, Kv7-mediated currents are 'subthreshold operating' in that the greatest influence of these currents is predicted to be in the range between resting membrane potential and the voltage threshold for action potential generation.

Comparison of the biophysical and pharmacological properties (including moderate TEA sensitivity, inhibition by muscarine, sensitivity to Ba^{2+} blockade, and blockade by both XE991 and linopirdine) of Kv7.2 and Kv7.3 heteromeric channels in heterologous expression systems with the so-called 'M' current in sympathetic neurons led to the conclusion

that Kv7.2 and Kv7.3 heteromeric channels mediate this current.

The M current (I_M) was discovered by Paul Adams and David Brown in sympathetic neurons, and it was later found to be present in many neurons, including sensory peripheral neurons as well as neurons in the CNS. I_M is believed to be one of the most important currents modulating the subthreshold excitability of mammalian neurons. I_M exhibits voltage dependence and becomes significant above −60 mV. Hence, the M current is operative in the subthreshold range of membrane potentials, a feature rendering it central in the control of neuronal excitability. Because it does not inactivate, or it inactivates extremely slowly and incompletely, it contributes K^+ current during long depolarizations and leads to substantial steady-state conductance at depolarized potentials. Therefore, I_M opposes depolarizing signals and drives the membrane potential towards E_K, functions as a 'brake' on repetitive action potential discharges, and suppresses responsiveness to synaptic inputs.

I_M was discovered as the current in sympathetic neurons that is inhibited by activation of muscarinic receptors, mediating muscarinic excitation of these neurons which releases the cell from the control of excitability exerted by I_M. This current was termed M because of this muscarinic modulation. Subsequent studies have shown that I_M is modulated in a similar manner by other neurotransmitters (e.g., serotonin, glutamate, and histamine) as well as several neuropeptides (bradykinin, angiotensin, substance P, and opioid) acting through G-protein-coupled receptors. These modulations, providing additional powerful means of physiologically modulating neuronal excitability, have been observed in both peripheral and CNS neurons.

Given their abundant, widespread, and largely overlapping expression in the nervous system (Table 1), Kv7.2 and Kv7.3 (probably often as components of heteromeric channels containing both subunits) are likely mediators of I_M in many central and peripheral neurons. However, all Kv7 channels expressed heterologously are able to form functional homomultimeric channels that express currents with electrophysiological and pharmacological characteristics of M-type currents. They all express currents with similar time and voltage dependence, they are all inhibited by the I_M blocker linopirdine, and they are similarly inhibited by neurotransmitters.

Among the Kv7 proteins, Kv7.1 is the only member not detected in the nervous system (Table 1). Kv7.1 is present in the heart (found in atrial and ventricular myocytes), where, in combination with the accessory subunit KCNE1, it forms the channels mediating a current known as I_{Ks}, which is important in the repolarization of cardiac action potentials. It is also present in a variety of transport epithelia, including the stria vascularis of the inner ear, intestinal crypt cells, and the proximal tubule epithelium of the kidney. This pattern of expression rules out this subunit as a contributor to M-type channels in neurons. Kv7.4 is found in the nervous system, but it has a more restricted pattern of expression than other neural Kv7 subunits, with its expression being limited to the inner ear and central auditory nuclei in the brain stem (Table 1). This limited distribution indicates that this subunit likely contributes to I_M only in this part of the brain. On the other hand, Kv7.5 is widely expressed in brain, often overlapping with Kv7.2 and Kv7.3, and may therefore, either in its homomultimeric form or as heteromultimers, generate I_M in many neurons, potentially adding to the heterogeneity of this current.

Classically, the M currents have been considered to lack inactivation, and the only Kv7 subunit for which inactivation is widely recognized is Kv7.1. Nonetheless, it has been found that homomeric channels containing Kv7.4 and Kv7.5 subunits (but not Kv7.2) expressed heterologously can undergo inactivation. This inactivation is voltage dependent and develops slowly with time constants of several seconds. However, it has an impact on the magnitude of the steady current at voltages in the resting membrane potential range (the current is reduced more than 30% at these voltages), indicating that this mechanism of regulation could be of physiological significance. Importantly, it was also shown that BMS-204352, a selective Kv7 channel opener, exerts its effect, at least partially, by abolishing the inactivation of these channels.

Some members of the voltage-gated K^+ channel family EAG express currents that activate slowly and do not inactivate, as is the case of typical M currents. Therefore, the currents mediated by these EAG channels have been referred to as M-like. EAG channels are prominently expressed in specific neuronal populations, especially in the forebrain, and may contribute M-like K^+ currents in neurons.

Results from immunolocalization of Kv7.2 and Kv7.3 proteins suggest that the currents mediated by Kv7 channels may have additional functions besides mediating the I_M recorded in neuronal cell bodies. These include localization at the axon initial segment in a variety of neurons and at the node of Ranvier in myelinated axons. Furthermore, Kv7.2 is expressed in the mossy fibers of the hippocampus, and there is pharmacological evidence for the presence of Kv7

currents in the Schaffer collaterals, where they appear to regulate transmitter release.

Diseases Associated with Kv7 Channels

We briefly describe the diseases associated with mutations in the genes encoding Kv7 proteins.

KCNQ1 and LQTS/JLNS Long QT syndromes (LQTS) are inherited cardiac disorders caused by mutations in the genes that encode sodium or potassium transmembrane ion channel proteins, characterized by arrhythmias which are often fatal. More than 200 mutations, in at least six genes, have been found in these patients. Mutations in the KCNQ1 gene are responsible for the most common form of the disease (LQT1), accounting for more than 50% of the cases. An autosomal dominant form of LQT1, known as Romano–Ward syndrome, is the result of dominant negative mutations; on the other hand, the Jervell and Lange-Nielsen syndrome (JLNS) is the recessive form of the disease and is associated with deafness, resulting from failure of K^+ transport into the endolymph in the cochlea. These phenotypes have been reproduced in knockout mice.

KCNQ2 and KCNQ3 in neonatal epilepsy Given the extensive overlap in expression patterns of Kv7.2 and Kv7.3 in brain tissue and the evidence that in most neurons the I_M channels are heteromeric channels containing both subunits, it is perhaps not surprising that mutations in each gene produce the same phenotype – an autosomal dominant idiopathic epilepsy known as benign familial neonatal convulsions (BFNC).

BFNC is a rare disease characterized by unprovoked partial or generalized clonic convulsions, sometimes with ocular symptoms and apnea. BFNC typically begins soon after birth, and remission usually occurs after 3–10 weeks, although approximately 15% of patients have seizures later in life. BFNC was initially linked to two different loci, 20q13.3 and 8q24, which are now known to contain the KCNQ2 and KCNQ3 potassium channel genes, respectively. A large number of mutations have been identified, particularly in KCNQ2, and include missense, frame-shift, and splice-site mutations as well as a deletion. The functional effects of the mutations have been explored in heterologous expression systems, and many of these were not found to have dominant negative effects, indicating that the dominant inheritance of BFNC is due to haploinsufficiency. Interestingly, the reduction of current observed with some of the mutations was rather modest, approximately 25%, indicating that the levels of the current are critically maintained within certain levels in early infancy. Why seizure activity disappears in most patients after a few weeks is not clear. Perhaps the maturation of inhibitory GABAergic systems later in life allows for better control of the increased excitability resulting from KCNQ mutations.

Deletion of Kcnq2 in mice is lethal. Homozygous animals die a few hours after birth due to breathing problems. Heterozygous animals develop normally and lack spontaneous epileptic activity, but they have an increased susceptibility to pentylenetetrazole-induced seizures. On the other hand, transgenic mice that conditionally express dominant negative Kv7.2 subunits in brain have spontaneous seizures, behavioral hyperactivity, increased neuronal excitability consistent with a reduced M current, and morphological changes in the hippocampus. Temporal control of expression of the transgene showed that Kv7 channel activity is critical to the development of normal hippocampal morphology during the first postnatal weeks.

KCNQ4 and inherited deafness KCNQ4 maps to the DFNA2 locus in human chromosome 1p34, one of more than 30 loci for dominant inherited deafness, and DFNA2 patients have been found to have mutations in the KCNQ4 gene. The mechanism leading to deafness in DFNA2 is different than that causing the hearing defect found in JLNS (produced by KCNQ1 mutations), consistent with the different patterns of expression of the two genes in the ear (Table 1). In DFNA2, hearing loss develops slowly, progressing over decades, probably as a result of degeneration of the outer hair cells (OHCs). Although KCNQ4 is also expressed in sensory hair cells of the vestibular organ, people with DFNA2 have no vestibular symptoms.

KCNQ4 is expressed in both inner and outer hair cells of the inner ear, where the currents may serve different functions. In OHCs, Kv7 currents are large and constitute the dominant currents at rest and during depolarization, also representing the principal exit pathway for K^+ ions entering the cells through the transduction channels. In inner hair cells, the current is relatively small and it influences mainly resting membrane properties. Loss-of-function mutations of the KCNQ4 gene, causing hearing loss in humans, also result in OHC degeneration. Similarly, KCNQ4 knockout mice or mice carrying KCNQ4 mutations also develop progressive OHC degeneration, which parallels a progressive hearing loss. OHC degeneration is believed to result from Ca^{2+}-dependent excitotoxicity produced by use-dependent chronic depolarization of the cell as a consequence of reduced K^+ current. Interestingly, some investigators have

suggested an association between increased frequency of age- and job-related deafness and specific KCNQ4 polymorphisms.

Modulation of Kv7 Channels

The mechanism of inhibition of I_M by neurotransmitters has been the subject of intense investigation since the current and its modulation were discovered in 1980. Yet, the mechanism of modulation remained elusive for more than 20 years. Classical second messengers known at the time of the discovery of the M current did not seem to be involved. By the early 1990s, it was clear that M-type channel inhibition by neurotransmitters uses G-protein activation, as is the case for other G-protein-coupled receptor (GPCR)-mediated modulations, and specifically the $G\alpha_q$ subunit (with occasional contribution from G_{11}). Activation of these G-proteins is known to stimulate the activity of the membrane-associated phospholipase Cβ. This enzyme hydrolyses the membrane phospholipid phosphatidylinositol-4,5-bisphosphate (PIP_2), leading to the production of inositol-1,4,5-trisphosphate (IP_3) and diacylglycerol (DAG). These products are key messengers involved in the signaling mediated by G_q and G_{11}, as well as several other G proteins. IP_3 leads to elevation of intracellular Ca^{2+} via release from intracellular stores, and DAG activates PKC. However, experiments investigating the roles of these signaling molecules on the modulation of M currents produced conflicting results at best.

A breakthrough occurred in 2002 when Byung-Chang Suh and Bertil Hille performed experiments that suggested that perhaps it is the depletion of PIP_2 from the membrane and not the products of its hydrolysis that causes the inhibition of the M-type channels. Suh and Hille studied the recovery of the M channels from muscarinic inhibition and showed that this process requires hydrolysable ATP and is slowed if PIP_2 resynthesis is inhibited with wortmannin, a blocker of PI4K kinase, which is an enzyme required for the generation of PIP_2. Soon after, Diomedes Logothetis and colleagues showed that Kv7.2 and Kv7.2–Kv7.3 channels in inside-out oocyte membrane patches undergo a rundown that could be recovered by the addition of PIP_2. The rundown could be restored using an antibody directed against PIP_2 or polylysine to sequester PIP_2, providing direct evidence that Kv7 channels require PIP_2 for activity, as had been shown to be the case for several other ion channels.

A large body of research now exists supporting the view that the primary mechanism by which GPCRs inhibit I_M is via depletion of membrane-associated PIP_2. Hille and colleagues have used translocatable enzymes to rapidly alter the PIP_2 levels. When PIP_2 is depleted, the Kv7 current declines to zero without changes in Ca^{2+}, DAG, or IP_3, and it increases when PIP_2 levels rise. This result shows that PIP_2 depletion alone without changes in other second messengers is able to fully suppress the current. PIP_2 comprises only approximately 1% of the cell membrane, and it is found mostly in the inner cell leaflet. Muscarinic and bradykinin receptor activation have been shown to produce large decreases in PIP_2 in neuroblastoma cells and in mammalian cell expression systems – approximately 35% for bradykinin and more than 75% for muscarine within 30–60 s. Experiments have shown that different Kv7 subunits have very different affinities for PIP_2: Kv7.3 > Kv7.2 > Kv7.4, with EC_{50} ranging from three to several hundred micromolar. This heterogeneity could confer diversity to Kv7-mediated neurotransmitter effects in different sites in the nervous system.

Although regulation of PIP_2 levels is likely the main mechanism of M current inhibition, other molecules and signaling events are likely to contribute as well in specific cases. For instance, Ca^{2+} ions inhibit, at concentrations of approximately 100–200 nM, M channels containing Kv7.2, Kv7.4, and Kv7.5 (but not Kv7.1 or Kv7.3) through an interaction with calmodulin. It appears that the inhibition of I_M currents by bradykinin in sympathetic neurons is due mainly to Ca^{2+}-dependent block resulting from an IP_3-mediated increase in intracellular Ca^{2+}. In these cells, bradykinin does not produce sufficient depletion of PIP_2 to inactivate the channels via this mechanism because the increase in Ca^{2+} concentration promotes PIP_2 synthesis. The reason why the mechanism of I_M modulation by muscarine in sympathetic neurons differs from the mechanism of modulation by bradykinin is that muscarinic receptors are too distant from the IP_3 receptors to induce a large release of Ca^{2+}. This is in contrast to the closer localization of bradykinin receptors to the intracellular Ca^{2+} stores.

Other signaling pathways may also contribute to I_M inhibition, including PKC, cyclic ADP-ribose, and tyrosine kinases. The action of at least some of these signals on Kv7 channels may converge on the requirement of PIP_2 for channel activity. For example, since one of the Ca^{2+}-calmodulin binding sites on the channel overlaps a putative PIP_2 binding site, Brown and colleagues have suggested that Ca^{2+} binding may produce channel closure by displacing PIP_2.

At normal resting concentrations of PIP_2, M channels are not fully open, allowing for the possibility of enhancing the resting I_M pharmacologically. Enhancing I_M should reduce excitability and counteract convulsions. This is the effect of the drug retigabine. Retigabine originated from the National Institutes of Health Antiepileptic Drug Development Program. The main action of this compound is to shift the

current–voltage curve of Kv7 currents to the left (although it also increases the maximum opening probability) via an interaction of retigabine with residues in the S5 and S6 domains of Kv7 proteins, with a tryptophan residue in the S5 domain as the key determinant. This tryptophan is absent from the cardiac Kv7.1 channel, which is therefore resistant to retigabine action. Retigabine has potent anticonvulsant activity. It also enhances I_M in nociceptive sensory neurons and exerts antinociceptive effects, particularly against neuropathic pain. No published accounts of retigabine efficacy in the clinical treatment of chronic pain exist. However, numerous open-label and double-blind controlled trials have shown that the close structural analog and Kv7 channel opener flupirtine is effective in alleviating pain associated with a diverse range of etiologies.

Kv7 channel modulation with retigabine has also been shown to reduce the distonia in a mouse model of paroxysmal dyskinesia. Substantial effects of retigabine have also been seen on dopaminergic neuron firing rate in slice preparations.

See also: Calcium Channel and Calcium-Activated Potassium Channel Coupling; Inwardly Rectifying Potassium Channels; Ion Channel Localization in Cell Bodies and Dendrites; Large Conductance Calcium-Activated Potassium Channels; Potassium Channel Regulation; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Bezanilla F (2000) The voltage sensor in voltage-dependent ion channels. *Physiological Reviews* 80(2): 555–592.

Birnbaum SG, Varga AW, Yuan LL, et al. (2004) Structure and function of Kv4-family transient potassium channels. *Physiological Reviews* 84(3): 803–833.

Choe S, Cushman S, Baker KA, and Pfaffinger P (2002) Excitability is mediated by the T1 domain of the voltage-gated potassium channel. *Novartis Foundation Symposium* 245: 169–175.

Coetzee WA, Amarillo Y, Chiu J, et al. (1999) Molecular diversity of K^+ channels. *Annals of the New York Academy of Sciences* 868: 233–285.

Delmas P and Brown DA (2005) Pathways modulating neural KCNQ/M (Kv7) potassium channels. *Nature Reviews Neuroscience* 6(11): 850–862.

Doyle DA, MoraisCabral J, Pfuetzner RA, et al. (1998) The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity. *Science* 280(5360): 69–77.

Goldberg EM, Watanabe S, Chang SY, et al. (2005) Specific functions of synaptically localized potassium channels in synaptic transmission at the neocortical GABAergic fast-spiking cell synapse. *Journal of Neuroscience* 25(21): 5230–5235.

Harvey AL and Robertson B (2004) Dendrotoxins: Structure–activity relationships and effects on potassium ion channels. *Current Medicinal Chemistry* 11(23): 3065–3072.

Hille B (1992) *Ionic Channels of Excitable Membranes*. Sunderland, MA: Sinauer.

Jentsch TJ (2000) Neuronal KCNQ potassium channels: Physiology and role in disease. *Nature Reviews Neuroscience* 1(1): 21–30.

Jerng HH, Pfaffinger PJ, and Covarrubias M (2004) Molecular physiology and modulation of somatodendritic A-type potassium channels. *Molecular and Cellular Neuroscience* 27(4): 343–369.

Johnston D, Christie BR, Frick A, et al. (2003) Active dendrites, potassium channels and synaptic plasticity. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 358(1432): 667–674.

Kaczmarek LK, Bhattacharjee A, Desai R, et al. (2005) Regulation of the timing of MNTB neurons by short-term and long-term modulation of potassium channels. *Hearing Research* 206(1–2): 133–145.

Lai HC and Jan LY (2006) The distribution and targeting of neuronal voltage-gated ion channels. *Nature Reviews Neuroscience* 7(7): 548–562.

Long SB, Campbell EB, and Mackinnon R, (2005) Crystal structure of a mammalian voltage-dependent Shaker family K^+ channel. *Science* 309: 897–903.

Long SB, Campbell EB, and Mackinnon R (2005) Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science* 309: 903–908.

Misonou H, Mohapatra DP, and Trimmer JS (2005) $Kv_{2.1}$: A voltage-gated K^+ channel critical to dynamic control of neuronal excitability. *Neurotoxicology* 26(5): 743–752.

Pongs O, Leicher T, Berger M, et al. (1999) Functional and molecular aspects of voltage-gated K^+ channel beta subunits. *Annals of the New York Academy of Sciences* 868: 344–355.

Robbins J (2001) KCNQ potassium channels: Physiology, pathophysiology, and pharmacology. *Pharmacology & Therapeutics* 90(1): 1–19.

Rhodes KJ, Strassle BW, Monaghan MM, Bekela-Arcuri Z Matos MF, and Trimmer JS (1997) Association and colocalization of the Kvβ1 and Kvβ2 β-subunits with Kv1 α-subunits in mammalian brain K^+ channel complexes. *Journal of Neuroscience* 17: 8246–8258.

Rudy B, Chow A, Lau D, et al. (1999) Contributions of Kv3 channels to neuronal excitability. *Annals of the New York Academy of Sciences* 868: 304–343.

Rudy B and McBain CJ (2001) Kv3 channels: Voltage-gated K^+ channels designed for high-frequency repetitive firing. *Trends in Neurosciences* 24(9): 517–526.

Sheng M and Wyszynski M (1997) Ion channel targeting in neurons. *Bioessays* 19(10): 847–853.

Strop P, Bankovich AJ, Hansen KC, Garcia KC, and Brunger AT (2004) Structure of a human A-type potassium channel interacting protein DPPX, a member of the dipeptidyl aminopeptidase family. *Journal of Molecular Biology* 343: 1055–1065.

Wang H, Yan Y, Liu Q, et al., Structural basis for modulation of Kv_4 K^+ channels by auxiliary KChIP subunits. *Nature Neuroscience* 10: 32–39.

Yu FH and Catterall WA (2004) The VGL-chanome: A protein superfamily specialized for electrical signaling and ionic homeostasis. STKE Science's 253: re15.

Relevant Website

http://www.informatics.jax.org – Mouse Genome Informatics.

Voltage-Gated Calcium Channels

T P Snutch, University of British Columbia, Vancouver, BC, Canada

© 2009 Elsevier Ltd. All rights reserved.

Native Voltage-Gated Calcium Channels

Calcium ions are divalent cations that affect the electrical properties of cells and also act as second messengers regulating a variety of intracellular pathways. This unique dual role is initiated by the rapid influx of Ca^{2+} through voltage-activated calcium channels in the plasma membrane which triggers the release of neurotransmitters at nerve terminals, the initiation of muscle contraction, the regulation of intracellular signaling cascades and calcium-dependent gene expression, hormone secretion, and the control of electrical excitability and firing patterns (Figure 1). The alteration of normal calcium channel function and intracellular calcium regulation are implicated in a number of human pathophysiological conditions, including epilepsy, chronic pain, muscle dysfunction, mood disorders, autism, and certain types of cardiovascular disease, cancer, and blindness.

Electrophysiological recordings from neurons, muscle, endocrine, and other cell types reveal multiple distinct types of calcium channels which differ in their biophysical, pharmacological, and modulatory properties. Generally, two major classes of calcium currents can be distinguished based on the membrane potentials at which they first open as well as other biophysical properties (Figure 2). The low-voltage-activated calcium channels (also known as T-type) transiently open in response to small changes from resting potentials, whereas high-voltage-activated calcium channels usually require stronger depolarizations for activation. The high-voltage-activated channels are a diverse class consisting of L-, N-, P/Q-, and R-types that can be distinguished by their unique pharmacological, kinetic, and voltage-dependent gating characteristics. Most electrically excitable cells express multiple subtypes of both low- and high-voltage-activated calcium channels, the specific combination of which helps to define unique cellular phenotypes.

The high-voltage-gated calcium channels are integral membrane multisubunit protein complexes, with the best biochemically characterized of these being the skeletal muscle T-tubule-associated L-type channel. This L-type calcium channel consists of five distinct polypeptides designated α_1, α_2, β, δ, and γ (Figure 2). The α_1, β, and γ subunits are each derived from distinct genes, whereas the α_2 and δ subunits are derived from a single gene product that is proteolytically cleaved to form separate α_2 and δ subunits that are then linked by disulfide bonding. Examination of the reconstituted T-tubule L-type complex using electron microscopy shows an asymmetric distribution of ancillary subunits surrounding the α_1 subunit central core. The entire complex appears to be roughly cylindrical, with dimensions of approximately 9 nm wide by 20 nm high. The high-voltage-activated neuronal P/Q-type and N-type channels are also complexes consisting of an α_1 subunit in association with one of four different β subunits (β_1–β_4) and one of four $\alpha_2\delta$ subunits ($\alpha_2\delta_1$–$\alpha_2\delta_4$). It is unclear whether there is an associated γ subunit as part of native P/Q-type or N-type complexes. A general model has been proposed in which four or five proteins form the multisubunit high-voltage-activated calcium channel complex, although subunit composition may differ depending on channel type (Figure 3). In the complex, the α_1 subunit forms the channel proper, comprising the voltage-sensing and -gating mechanisms, the Ca^{2+}-selective pore, and interaction sites for drug binding and second messenger-dependent regulation. The ancillary proteins interact with the α_1 subunit to modulate channel activity and trafficking to the plasma membrane. The functional characteristics of the low-voltage-activated T-type calcium channels can be reconstituted by an α_1 subunit alone.

α_1 (Ca_v) Subunits

The large α_1 subunit proteins (~200–250 kDa) bear a high degree of similarity to the voltage-gated sodium channel α_1 subunits and are also evolutionarily related to voltage-gated potassium channels and other members of the ion channel superfamily. The mammalian genome contains 10 pore-forming calcium channel α_1 (Ca_v) subunit genes grouped into three interrelated subtypes (Figure 4). It is generally accepted that the distinct Ca_v subunits account for the known types of native voltage-activated calcium channels. Two branches of the Ca_v subunit phylogenetic tree account for the high-voltage-activated channels. In one branch, the $Ca_v1.1$–$Ca_v1.4$ subunits all encode distinct L-type calcium channels, whereas a second branch $Ca_v2.1$ encodes the P/Q-type, $Ca_v2.2$ the N-type, and $Ca_v2.3$ the R-type channel. The third Ca_v branch ($Ca_v3.1$–$Ca_v3.3$) encodes variants of low-voltage-activated T-type calcium channels. All of the Ca_v subunit genes are subject to extensive alternative splicing, resulting in multiple functionally distinct isoforms of each of the 10 calcium channel subtypes.

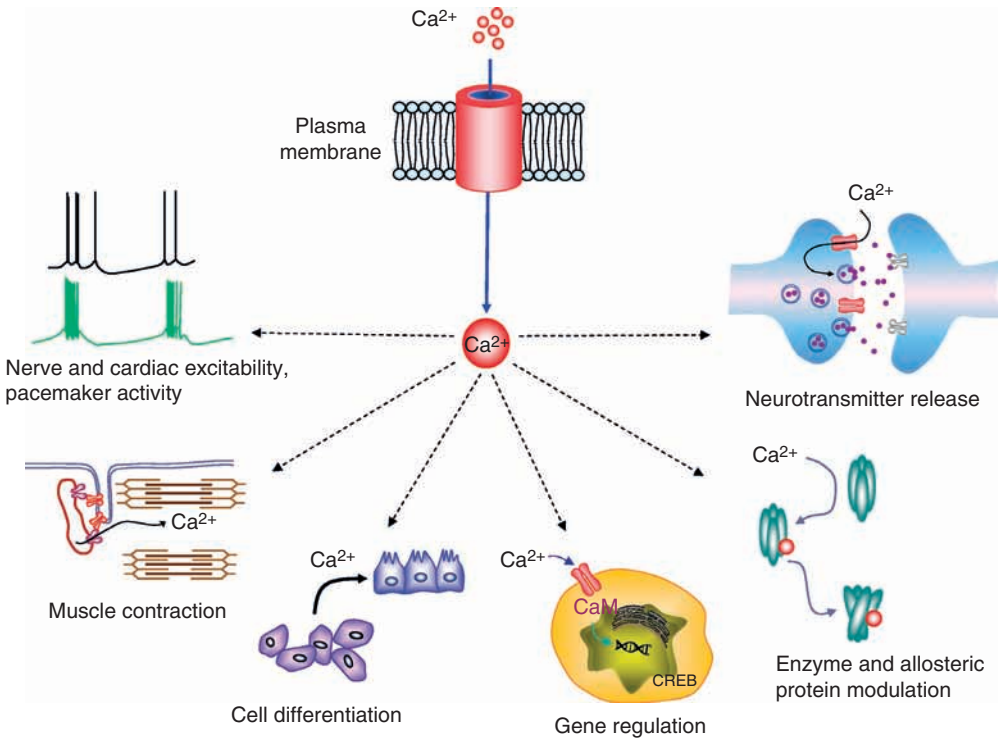

Figure 1 Voltage-gated calcium channels mediate numerous physiological functions. In response to depolarizing changes in membrane potential, voltage-gated calcium channels rapidly open (within ~1 ms), raising intracellular calcium to affect membrane electrical properties such as bursting and firing patterns as well as to regulate calcium-sensitive processes including muscle contraction, neurotransmitter release, cellular differentiation, and gene transcription.

The Ca_v subunit structure is similar to voltage-activated sodium channels, consisting of four homologous domains (I–IV) that form the main integral membrane components of the channel (Figure 5). The 10 Ca_v subunits share the most similarity in the membrane-spanning portions of domains I–IV and the least similarity in the cytoplasmic linkers separating the domains and the cytoplasmic and amino carboxyl tails. The highly divergent regions in the Ca_v subunits are the target sites of subtype-specific second messenger-dependent modulation and other protein–protein interactions.

The permeation profile of the high-voltage-activated calcium channels for cations is $Ca^{2+} > Ba^{2+} > Li^{+} > Na^{+} > K^{+} > Cs^{2+}$. Although Ca^{2+} and Na^{+} ions are of roughly similar diameter, the permeation pathway of voltage-gated calcium channels is approximately 1000-fold more selective for Ca^{2+} than for Na^{+} ions. The carboxylate side chains of four conserved p-loop glutamate residues within the lumen of the pore form the selectivity filter with a single high-affinity Ca^{2+} binding site. The four glutamate side chains are not functionally equivalent and may flex to enable a single Ca^{2+} ion to bind with high-affinity or two Ca^{2+} ions with lower affinity. This process would simultaneously contribute to both selectivity of Ca^{2+} over Na^{+} and electrostatically to a high rate of ionic flux. T-type calcium channels exhibit distinct permeation characteristics, including a smaller pore diameter; the ability to open with a half-size conductance level; and roughly similar permeation for Ca^{2+}, Ba^{2+}, and Sr^{2+}. The permeation differences between high-voltage-activated and T-type channels are in part due to distinct amino acids flanking the p-loops as well as to the domain III and IV p-loops of T-type channels possessing aspartates rather than glutamate residues in their selectivity filters.

Both high- and low-voltage-activated calcium channels intrinsically exhibit fast voltage-dependent inactivation, a process that limits the amount of Ca^{2+} entering cells during depolarization. Site-directed mutagenesis and chimeric channel approaches have identified various specific amino acids and transmembrane segments of the Ca_v subunit as well as interactions with β subunits as being critical components of inactivation. Overall, the rapid voltage-dependent inactivation of high-voltage-activated calcium channels is likely similar to that of both sodium and potassium channels and involves regions of the intracellular mouth of the pore (S6 segments) being occluded by a movable intracellular segment in 'ball-and-chain'- or 'hinged lid'-type mechanisms. T-type

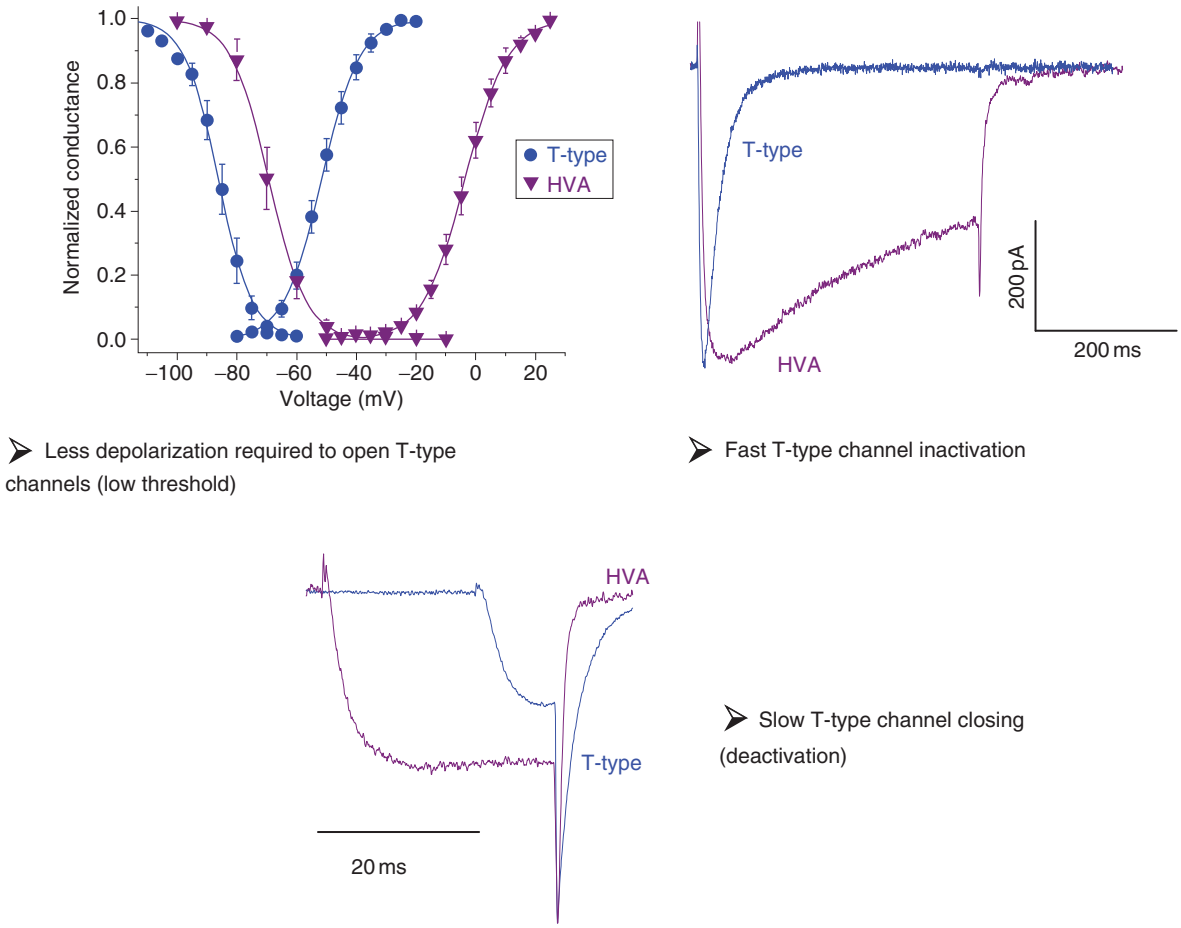

Figure 2 High-voltage-activated (HVA) and low-voltage-activated (T-type) calcium channels exhibit distinct biophysical properties. T-type calcium channels can generally be distinguished by their more negative voltage dependencies of activation and inactivation, faster kinetics of inactivation, and slower kinetics of deactivation (closing).

calcium channel inactivation and activation are strongly coupled, and the inactivation properties of T-type channels can be distinguished as both very fast and calcium independent. Although many structure–function-type studies have been reported, until a calcium channel crystal structure becomes known, naturally occurring alternatively spliced variants provide both insightful structure–function information and clues toward how Mother Nature has altered specific channel regions to affect downstream physiological processes. For example, lengthening the $Ca_v3.1$ T-type channel domain II–III and III–IV linkers through alternative splicing variously affects activation and inactivation kinetics, recovery from inactivation, closing rates, and the voltage dependence of both activation and inactivation. As another example, an alternative splicing event that alters 14 amino acids in the proximal carboxyl region of the $Ca_v2.2$ subunit affects N-type current densities, electrophysiological properties, and sensitivity to G-protein modulation. Furthermore, one $Ca_v2.2$ splice variant is preferentially expressed in dorsal root ganglion nociceptors and is required for both the mediation of basal thermal nociception and the development of thermal and mechanical hyperalgesia associated with inflammatory and neuropathic pain. Similarly, there appear to be pathophysiological consequences associated with cardiovascular disease and alternative splicing that affect the biophysical properties of $Ca_v1.2$ L-type channels.

β Subunits

The β subunits are cytoplasmic proteins that directly associate with the high-voltage-activated Ca_v subunits to modulate channel trafficking and biophysical properties. Structurally, the β subunit proteins contain an internal Src homology 3 (SH3) domain linked to an enzymatically inactive guanylate kinase (GK) domain and are members of the membrane-associated GK (MAGUK) family of scaffold proteins albeit with some distinct features. Analogous to other MAGUKs such as PSD-95, the calcium channel β subunit SH3 and GK domains exhibit a stable intramolecular interaction.

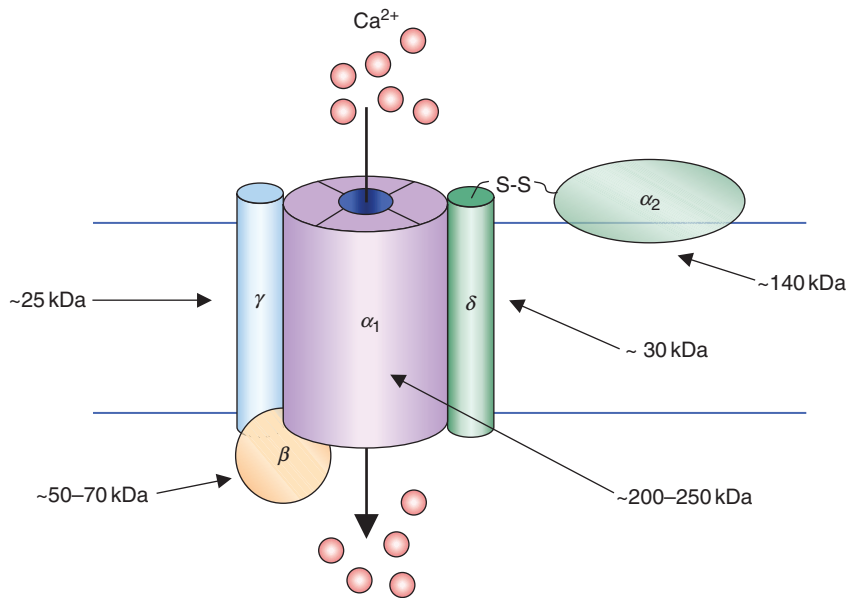

Figure 3 Subunit composition of high-voltage-activated L-type calcium channel complex. Schematic diagram of the high-voltage-activated L-type calcium channel complex originally purified from skeletal muscle T-tubule membranes. The complex consists of five subunits: α_1, α_2, δ, β, and γ. The α_1 subunit forms the channel proper, comprising the voltage-sensing and-gating mechanisms, the calcium-selective pore, and the target of most pharmacological agents. Functional T-type calcium channels consist of an α_1 subunit alone, and there is no biochemical evidence indicating that other proteins are associated with native T-type channels in a multisubunit complex.

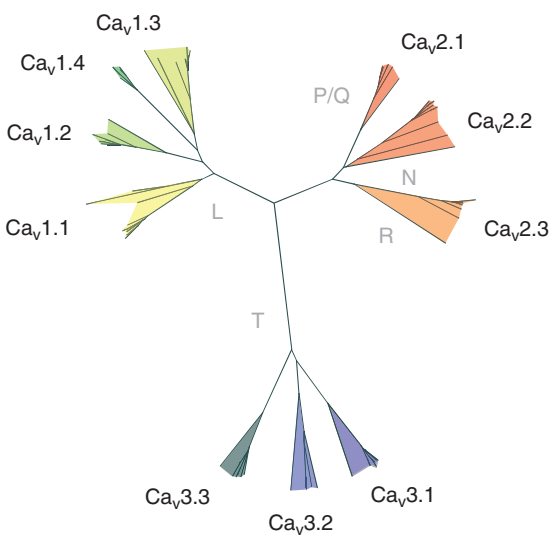

Figure 4 Similarity tree of the metazoan calcium channel α_1 (Ca_v) subunit family. The amino acid sequences of representatives of the 10 metazoan calcium channel α_1 subunits were aligned using Clustalx and the overall similarities are represented in the form of a similarity tree using HyperTree.

Ca_v–β subunit binding occurs in 1:1 stoichiometry with low nanomolar affinity to affect current amplitude, kinetics, and the voltage dependences of activation and inactivation. The basis for affecting the level of calcium channel whole cell currents seems to be through promoting translocation of the Ca_v subunit protein from the endoplasmic reticulum to the plasma membrane. On the α_1 subunit, the Ca_v–β subunit binding interaction occurs in a short cytoplasmic consensus motif in the domain I–II linker found in all high-voltage-activated Ca_v subunits (termed the α_1-interaction domain (AID)).

There are four different β subunit genes in mammals (β_1–β_4), each of which is subject to alternative splicing and each of which differentially affects the kinetic and voltage-dependent properties of high-voltage-activated Ca_v channels. The various β subunit isoforms vary significantly in primary sequence in their N- and C-termini and in the linker separating the SH3 and GK domains. The only available high-resolution structural information concerning voltage-gated calcium channel complexes is the X-ray crystallographic definition of the interaction between the Ca_v domain I–II AID region and portions of the β subunit. The consensus motif of the Ca_v AID binds as an amphipathic helix to a deep groove of the β subunit GK domain termed the AID-binding pocket.

$\alpha_2\delta$ Subunits

The disulfide-linked α_2 and δ subunit proteins are heavily glycosylated, with the $\alpha_2\delta$ subunit predominantly extracellular and anchored in the membrane by a transmembrane segment formed by the

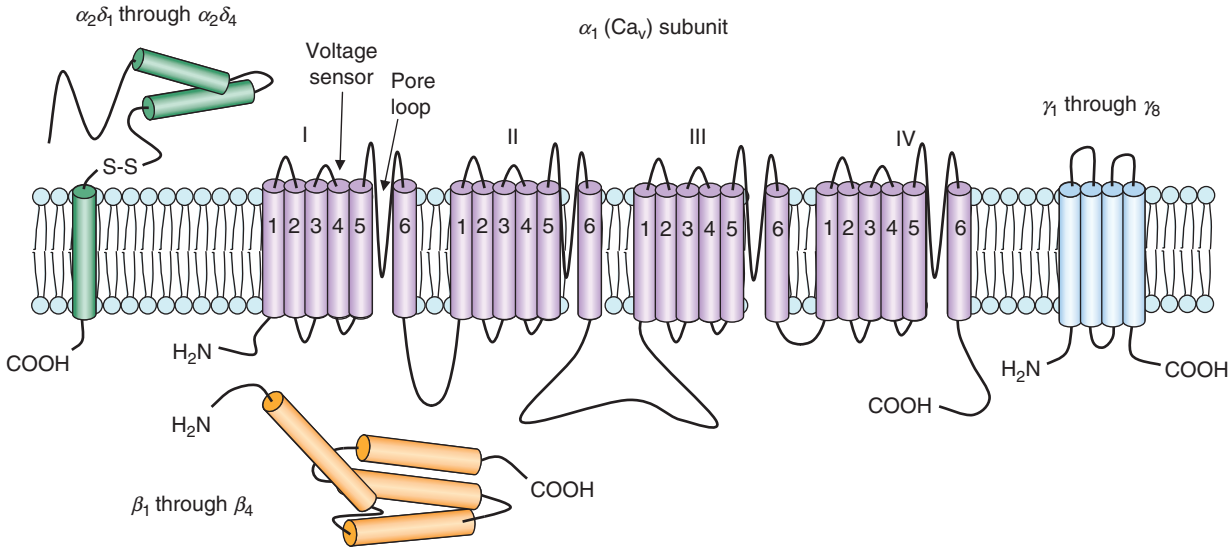

Figure 5 Schematic of the predicted structure and transmembrane topology of calcium channel complex proteins. Ca$_v$ α_1 subunits contain four homologous domains, each possessing six α-helical membrane-spanning segments (S1–S6) and a pore-forming p-loop. The S4 segment in each domain contains positively charged lysine or arginine residues at every third or fourth position forming a crucial part of the voltage-sensing mechanism of the channel. All high-voltage-activated channel Ca$_v$ subunits possess a conserved EF hand motif in the C-terminus and also a conserved region in the domain I–II linker that binds in 1:1 stoichiometry to the cytoplasmic β subunit. The α_2 and δ subunits are derived from a single gene product which is proteolytically cleaved to form the approximately 140 kDa α_2 and approximately 30 kDa δ subunits, which are then disulfide bonded to each other in the native state. The γ subunits are composed of four membrane-spanning regions with their C- and N-termini located intracellularly and with predicted extracellular N-glycosylation.

δ subunit. Co-expression of $\alpha_2\delta$ subunits with Ca$_v$ subunits affects current density, channel kinetics, and current–voltage relations, although generally the $\alpha_2\delta$ functional effects are more subtle that those for the β subunits. In mammals, there are four known $\alpha_2\delta$ genes, $\alpha_2\delta_1$–$\alpha_2\delta_4$. All four $\alpha_2\delta$ subunit genes are subject to alternative splicing, with tissue and cellular specificity demonstrated for many of the variants.

In humans, the small organic molecules gabapentin and pregabalin are clinically effective anticonvulsants that are also effective in neuropathic pain conditions such as diabetic neuropathy, postherpetic neuralgia, trigeminal neuralgia, and pain associated with cancer and multiple sclerosis. Although synthetic analogs of the neurotransmitter γ-aminobutyric acid (GABA), gabapentin and pregabalin do not exert their effects via interacting with GABA receptors or transporters but, rather, bind with high-affinity to the high-voltage-activated calcium channel $\alpha_2\delta_1$ and $\alpha_2\delta_2$ subunits. Peripheral nerve injury in animals upregulates $\alpha_2\delta$ expression in the dorsal root ganglia and spinal dorsal horn, suggesting that the $\alpha_2\delta$ subunit contributes to central pain sensitization. Interestingly, whereas $\alpha_2\delta_1$ and $\alpha_2\delta_2$ subunits associate with high-voltage-activated calcium channel α_1 subunits, including the L-type channels found in skeletal, smooth, and cardiac muscles, gabapentin and pregabalin exhibit relatively few motor or cardiovascular adverse effects at therapeutic doses.

γ Subunits

There are at least eight γ subunit genes in the mammalian genome (γ_1–γ_8), all of which are considered to be members of the Claudin family of tight junction-associated adhesion molecules. Structurally, the C-termini of γ_2–γ_4, and γ_8 have a PDZ-binding motif, whereas the analogous regions of γ_7 and γ_5 lack this consensus motif but, rather, possess an SS/TSPC site likely designated toward specific protein interactions.

The γ_1 subunit is part of the skeletal muscle L-type calcium channel complex and is also found in heart, lung, spleen, kidney, liver, and testis. Mice lacking the γ_1 subunit display altered skeletal muscle L-type calcium currents, including a hyperpolarized shift in steady-state inactivation properties. Three other γ subunits, γ_2–γ_4, are of closest similarity to γ_1 and make up a subfamily of neuronal γ subunits.

Co-expression of some γ subunits in combination with Ca$_v$, β, and $\alpha_2\delta$ subunits shows diverse and γ subtype-specific effects on peak current levels, kinetics, current–voltage relations, and steady-state inactivation properties. However, in other instances, γ subunit co-expression does not appear to affect Ca$_v$ properties, leaving open the question as to which γ subunit isoforms are *bona fide* components of high-voltage-activated calcium channel complexes. In one intriguing instance, co-expression of γ_7 with Ca$_v$2.2 results in the complete abolition of N-type currents,

an effect apparently mediated by blocking expression rather than interfering with trafficking or channel biophysical properties. Additionally, a subset of γ subunits (γ_2–γ_4, and γ_8) represent TARP family members (transmembrane AMPA receptor regulator proteins) that regulate AMPA receptor trafficking as well as contribute to synaptic plasticity.

Low-Voltage-Activated (T-Type) Calcium Channels: $Ca_v3.1$–$Ca_v3.3$

The low-voltage activated (T-type) calcium channels first activate at negative potentials (approximately −75 to −60 mV) and are distinctive in their fast and strongly voltage-dependent kinetics of activation and inactivation, rapid recovery from inactivation, slow closing, and small single channel conductances (Figure 2). Their hyperpolarized and overlapping activation and inactivation ranges allow some calcium entry at resting membrane potentials (window current). This combination of unique biophysical characteristics allows T-type calcium channels to generally act as pacemakers, helping to trigger bursts of sodium-dependent action potentials after membrane hyperpolarization. They also contribute to oscillatory and rebound burst-firing behaviors relevant to normal physiological functions, such as thalamocortical-mediated deep sleep. Since T-type channels often generate a resting inward current, they also likely contribute to the gating of calcium-dependent ion channels and regulation of calcium-dependent enzymes and gene expression.

Although commonly referred to as a single class of channels, they are in fact physiologically and pharmacologically distinct T-type calcium channels encoded by three different Ca_v subunit genes (Figure 4). All three Ca_v subunits ($Ca_v3.1$–$Ca_v3.3$) support strongly voltage-gated calcium currents with the properties expected of T-type channels, although they each exhibit distinct kinetic, permeation, pharmacological, and other voltage-dependent properties, most likely reflective of their unique primary sequences. Multiple Ca_v3 alternative splice variants have been identified and probably also account for some of the observed heterogeneity in native T-type currents. The Ca_v3 N- and C-termini and the intracellular linkers joining the transmembrane domains share little sequence identity with high-voltage-activated Ca_v subunits. Furthermore, Ca_v3 subunit channels do not possess several specific functional motifs found in high-voltage-activated calcium channels, including the AID β subunit binding site in the domain I–II linker and EF-hand/calmodulin motifs in the C-terminus, the latter being consistent with the lack of Ca^{2+}-dependent inactivation for T-type channels.

T-type calcium channels are widely distributed to many central and peripheral neurons. At the subcellular level, the $Ca_v3.1$, $Ca_v3.2$, and $Ca_v3.3$ isoforms differentially distribute to somatic and dendritic compartments, implying that they uniquely contribute to neurophysiological processes. Outside the nervous system, T-type channels are found in many excitable and nonexcitable cells, and evidence indicates roles for T-type channels in pathophysiological states such as tumor proliferation, renal injury, and cardiovascular disease (Tables 1–3).

T-type channels as a class generally lack selective, high-affinity ligands. Small organic compounds such as octanol, amiloride, and mibefradil have been utilized as T-type channel antagonists, although they also variously inhibit high-voltage-activated calcium channels and other conductances. Several classes of the clinically utilized antipsychotics and antiepileptics block T-type calcium channels, although in most instances these agents are not specific (Figure 6).

$Ca_v3.1$ Subunit: CACNA1G Gene

$Ca_v3.1$ channel biophysical properties are very similar to those of the $Ca_v3.2$ T-type, although $Ca_v3.1$ channels do appear to exhibit a faster recovery from inactivation when membrane potential returns to negative values following depolarization. $Ca_v3.1$ subunits are expressed abundantly throughout the brain and the sinoatrial node of the heart, and lower levels are expressed in placenta, lung, and kidney.

Mice in which the $Ca_v3.1$ gene is genetically deleted appear anatomically and physiologically normal, although there are several alterations to nervous and cardiovascular system functions. Nervous system-associated defects include an absence of $GABA_B$ receptor agonist-induced spike-wave discharges, a lack of thalamic 1–4 Hz delta waves, a reduction of sleep spindles associated with normal sleep, and an enhanced visceral pain response. Overall, it appears that $Ca_v3.1$ channels play a crucial role in thalamocortical burst-firing processes relevant to a number of higher central nervous system functions, including thalamic sensory gating and the control of consciousness. In the heart, $Ca_v3.1$ gene deletion is associated with both slowing of the heart rate (bradycardia) and atrioventricular dysfunction, indicating that this T-type calcium channel normally contributes to setting the basal heart rate and to atrioventricular conduction.

$Ca_v3.2$ Subunit: CACNA1H Gene

$Ca_v3.2$ T-type channels are widely expressed in the central and peripheral neurons, cardiac tissue, kidney, and liver. In dorsal root ganglion neurons, $Ca_v3.2$

Table 1 Voltage-gated calcium channels

Native channel type	*α_1 Subunit*	*Pharmacological sensitivity*	*Major distribution*	*Major cellular functions*
P/Q	$Ca_v2.1$	ω-Aga-IVA (P-type), ω-CgTx-MVIIC (Q-type)	Neurons, heart, pituitary Subcellular: cell bodies, dendrites, presynaptic terminals	Neurotransmitter release, dendritic signaling
N	$Ca_v2.2$	ω-CgTx-GVIA, ω-CgTx-MVIIA	Neurons, neuroendocrine cells Subcellular: subset of cell bodies, dendrites, presynaptic terminals	Neurotransmitter release, dendritic signaling, neurite outgrowth
L	$Ca_v1.1$	Dihydropyridines, benzothiazapines, phenylalkylamines	$Ca_v1.1$: skeletal muscle T-tubules	$Ca_v1.1$: excitation–contraction coupling
	$Ca_v1.2$		$Ca_v1.2$: neurons, smooth (blood vessel, uterine, lung, stomach, intestine) and cardiac muscle; endocrine cells	$Ca_v1.2$: excitation–contraction coupling, Ca^{2+}-dependent gene transcription
	$Ca_v1.3$		$Ca_v1.3$: neurons, endocrine cells, heart atria, cochlea	$Ca_v1.3$: hormone secretion, Ca^{2+}-dependent gene transcription, atrial and neuronal pacemaking activity
	$Ca_v1.4$		$Ca_v1.4$: retina (photoreceptors, horizontal, bipolar, and amacrine cells), immune system, low levels in spinal cord and skeletal muscle Subcellular: $Ca_v1.2$ and $Ca_v1.3$ on cell bodies and proximal dendrites of neurons; $Ca_v1.4$ at photoreceptor synapses	$Ca_v1.4$: neurotransmitter release in visual system
R	$Ca_v2.3$	SNX-482	Neurons, heart Subcellular: cell bodies, some distal dendrites and presynaptic terminals	Selective neurotransmitter release, repetitive firing
T	$Ca_v3.1$	Poorly defined–nonspecific block by mibefradil, amiloride, pimozide, flunarizine, Ni^{2+}, Zn^{2+}, Cd^{2+}	Neurons, smooth muscle, endocrine cells, placenta, testis, lung, kidney cardiac sinoatrial cells, neuroendocrine cells, spermatogenic cells, developing skeletal muscle, fibroblasts, osteoblasts, astrocytes	Pacemaking activity, repetitive spiking and bursting, dendritic neurotransmitter release, cellular differentiation
	$Ca_v3.2$		Subcellular: soma, proximal and distal dendrites	
	$Ca_v3.3$			

channels contribute to both acute and neuropathic pain behaviors, whereas in mechanoreceptors they are required for the transduction of slow-moving mechanical stimuli. In sperm, the activation of $Ca_v3.2$ channels is required for triggering the acrosome reaction and allowing sperm to become competent for fusion with oocytes. Pharmacologically, $Ca_v3.2$ channels can be distinguished from both $Ca_v3.1$ and $Ca_v3.3$ by the higher sensitivity to nickel blockade (IC_{50} of $\sim$10 μM vs. $\sim$200 μM).

Genetic deletion of $Ca_v3.2$ in mice results in smaller but viable animals that possess constricted coronary arteries and develop cardiac fibrosis. In agreement with the high levels of $Ca_v3.2$ channels in small dorsal root ganglion cells, knockout mice also exhibit attenuated responses to mechanical, thermal, and chemical pain stimuli.

Point mutations in the CACNA1H gene in humans are associated in a subset of patients with childhood absence epilepsy (CAE) and idiopathic generalized epilepsy (IGE). The majority of the mutations are localized in the $Ca_v3.2$ channel intracellular linker between domains I and II, and biophysical analyses reveal both gain-of-function and loss-of-function effects. The gain-of-function effects are predicted to result in increased calcium influx at lower membrane potentials, which would result in neuronal hyperexcitability and increased spike-wave discharge. The polygenic, non-Mendelian inheritance of CAE and IGE suggests that interaction with other gene products must also be considered with regard to disease pathogenesis.

Missense mutations in the CACNA1H gene resulting in a predicted overall decrease in calcium entry through $Ca_v3.2$ channels have also been associated with a small subset of individuals with autism spectrum disorder (ASD). Considering the tight regulation of intracellular calcium levels necessary for normal

Table 2 Spontaneous mutations in mammalian calcium channel subunit genes

Gene and human chromosomal location	*Channel subunit*	*Disease state/phenotype*
CACNA1S 1q31–32	$Ca_v1.1$	Human (dominant mutations): hypokalemic periodic paralysis-periodic muscle weakness; malignant hypothermia susceptibility type 1-elevated skeletal muscle contraction, glycogenolysis, increased heat and lactic acid acidosis, tachycardia, and cardiac arrhythmia and arrest Mouse: recessive mutations, muscular dysgenesis (*mdg*)
CACNA1C 12p13.3	$Ca_v1.2$	Human: Timothy syndrome–severe functional and developmental abnormalities in several organ systems, including heart, skin, eyes, teeth, immune system, and the brain; autism
CACNA1F Xp11.23	$Ca_v1.4$	Human (recessive and dominant mutations): incomplete X-linked congenital stationary night blindness–variable degrees of night blindness, low visual acuity, nystagmus, and hyperopia or myopia; X-linked cone–rod dystrophy–low visual acuity, reduced color vision, central scotomas in visual field, photophobia, and myopia
CACNA1A 19p13.1	$Ca_v2.1$	Human (dominant mutations): episodic ataxia type 2–spontaneous episodes of ataxia, nystagmus, and often migraine-like symptoms; familial hemiplegic migraine type 1–migraine with aura, motor aura typically manifests as hemiplegia in upper and lower extremities, often associated with ataxia or nystagmus; spinal cerebellar ataxia type 6–gate ataxia, reduced coordination, nystagmus, dysarthia and proprioceptive sensory loss, severe cerebellar Purkinje cell loss Mouse (recessive mutations): tottering (*tg*), leaner (tg^{la})–ataxia, dystonia, spike-wave epilepsy; dominant mutation; wobbly ataxia
CACNA1H 16p13.3	$Ca_v3.2$	Human: idiopathic generalized epilepsy/childhood absence epilepsy–absence seizures including short seizures followed by sudden impairment of consciousness; autism spectrum disorder Rat (recessive mutation): absence epilepsy
CACNA2D	α_2–δ_2	Mouse: recessive mutation; ducky–spike-wave epilepsy
CACNB4 2q22-23	β_4	Human: generalized epilepsy–recurrent generalized seizures, episodic ataxia Mouse: lethargic (*lh*)–spike-wave epilepsy
CACNG2	γ_2	Mouse (recessive mutations): stargazer (*stg*); waggler (stg^{wag}), altered skeletal muscle L-type current, spike-wave epilepsy

Table 3 Targeted gene deletions in mouse calcium channel subunit genes

Gene	*Channel subunit*	*Disease state/phenotype*
CACNA1S	$Ca_v1.1$	Nonviable, suffocate at birth due to inability to contract diaphragm
CACNA1C	$Ca_v1.2$	Die approximately embryonic day 12.5
CACNA1D	$Ca_v1.3$	Viable, deaf, atrial arrhythmias, bradycardia
CACNA1F	$Ca_v1.4$	Viable, impaired vision
CACNA1A	$Ca_v2.1$	Usually die approximately 3 weeks postnatal, absence seizures, ataxia, dystonia, spasms, abnormal cerebellar synapses and cerebellar cell death, altered acute and inflammatory pain responses
CACNA1B	$Ca_v2.2$	Viable, reduced neuropathic and inflammatory pain responses, altered responses to ethanol and anesthetics, hyperactive, increased basal vigilance state, reduced baroreflex, decreased anxiety, impaired long-term memory
CACNA1E	$Ca_v2.3$	Viable, altered visceral pain response, impaired spatial memory and locomotion, altered responses to cocaine, anesthetics and cerebral ischemic damage, decreased insulin response and glucose tolerance
CACNA1G	$Ca_v3.1$	Viable, absence of $GABA_B$ receptor agonist-induced spike-wave discharges, decreased heart rate, slower atrioventricular conduction
CACNA1H	$Ca_v3.2$	Viable, impaired relaxation of vasculature, altered acute and chronic pain responses
CACNB1	β_1	Perinatal lethality, no excitation–contraction coupling
CACNB2	β_2	Embryonic lethal
CACNB3	β_3	Viable, altered blood pressure in response to high-salt diet, reduced acute and inflammatory pain responses
CACNB4	β_4	Ataxia, focal motor seizures, absence epilepsy
CACNG1	γ_1	Viable, no visible phenotype
CACNG4	γ_4	Viable, no visible phenotype

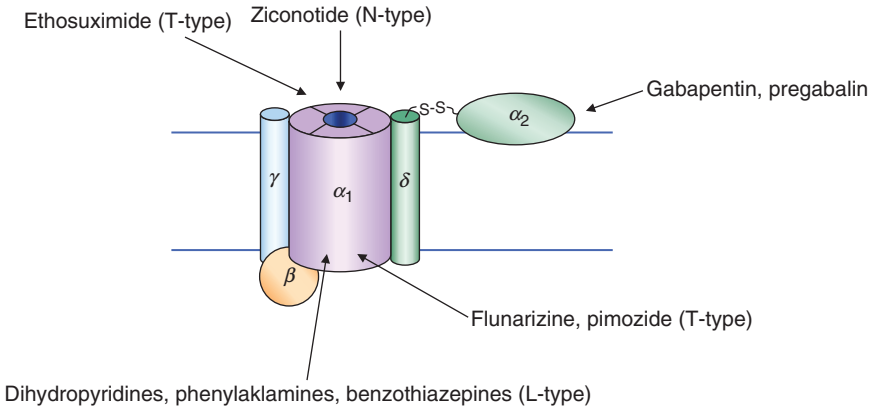

Figure 6 Clinical agents utilized to treat human cardiovascular and neurological disorders. The L-type calcium channel antagonists (dihydropyridines, phenylalkylamines, and benzothiazepines) block all Ca_v1 subunits *in vitro*, but they predominantly act *in vivo* on the cardiovascular system and exhibit selectivity between cardiac and vascular functions. Gabapentin and pregabalin bind to specific arginine residues in the $\alpha_2\delta_1$ and $\alpha_2\delta_2$ subunits, but determination of the exact mechanism of action of gabapentin has proven elusive, with reports both supporting and refuting direct actions on high-voltage-activated calcium channel properties. Ziconotide is a 25-amino acid synthetic version of ω-contotoxin MVIIA that reversibly blocks N-type channels and is highly effective in multiple types of human chronic and neuropathic pain conditions. Side effects of ziconotide administration include orthostatic hypotension, confusion, nystagmus, sedation, auditory and visual hallucinations, and agitation. Clinical antipsychotics such as pimozide and flunarizine block T-type channels, but they also target dopamine receptors. Ethosuximide, an absence epilepsy therapeutic, blocks T-type channels at millimolar concentrations.

cell signaling and gene expression, it is possible that the biophysical effects of the CACNA1H mutations translate into neuronal development abnormalities. Similar to that for CAE and IGE, the strong polygenic inheritance pattern of ASD, together with the fact that many patients do not have associated mutations in CACNA1H, indicates a likely minor role for the $Ca_v3.2$ channel in overall ASD susceptibility.

$Ca_v3.3$ Subunit: CACNA1I Gene

Compared to the $Ca_v3.1$ and $Ca_v3.2$ T-type channels, recombinant $Ca_v3.3$ channels uniquely exhibit voltage-dependent facilitation, a larger time constant of activation (slower opening), as well as a slower rate of inactivation and a faster deactivation (closing). Together, these properties predict that $Ca_v3.3$ channels generate sustained inward calcium current under conditions of prolonged depolarizations as well as a large window calcium current under resting conditions. Also compared to $Ca_v3.1$ and $Ca_v3.2$ channels, $Ca_v3.3$ T-type channels are differentially modulated by second messenger signaling pathways, including those acting through muscarinic acetylcholine and metabotropic glutamate receptors.

High-Voltage-Activated Calcium Channels

L-Type Calcium Channels: $Ca_v1.1$–$Ca_v1.4$

High-voltage-activated L-type calcium channels are expressed in almost all excitable cells as well as many types of non-excitable cells (Table 1). In cardiac, skeletal and smooth muscles they are the primary route for calcium entry and are responsible for triggering muscle contraction. In neurons, L-type channels are primarily located on cell bodies and proximal dendritic regions where they are vital players in linking electrical excitability to the control of calcium-dependent gene transcription. Some L-type channels are also involved in exocytotic release from both endocrine cells and specialized neurons such as photoreceptors.

L-type calcium channels generally require large depolarizations to become activated, typically opening at potentials positive to -30 mV, although they can activate at significantly more negative potentials in chromaffin cells, sensory neurons, photoreceptors, and cardiac cells. In addition to voltage-dependent mechanisms, L-type currents inactivate rapidly through a Ca^{2+}-dependent negative feedback process that serves to finely control Ca^{2+}-dependent signaling and protect against toxic Ca^{2+} overload.

L-type calcium channels are readily distinguished by their sensitivities to several classes of small organic molecules widely used in the clinical treatment of hypertension, angina, Raynaud's disease, and cerebral vasospasm. The 1,4-dihydropyridines (DHPs), phenylalkylamines, and benzothiazepines interact at allosterically distinct binding sites on the L-type Ca_v subunit to block channel activity. The *in vivo* specificity of physiological actions of these agents likely reflects a number of factors, including pharmacokinetic aspects, the existence of distinct L-type molecular isoforms

($Ca_v1.1$–$Ca_v1.4$), and the state dependence of drug interaction. The state dependence of channel blockade is of particular relevance since both drug access to its binding site and the affinity of drug binding are highly dependent on the conformational state of the channel. Several DHP agonists have also been developed, the most highly utilized being (-)-Bay K 8644, which increases both L-type channel open time and single channel conductance.

$Ca_v1.1$ Subunit: CACNA1S Gene

$Ca_v1.1$ L-type calcium channels are crucial in striated muscle for coupling membrane depolarization to the release of calcium from cytoplasmic stores which then triggers excitation–contraction (EC) coupling. Depolarizing changes in skeletal muscle membrane potential cause a conformational change in $Ca_v1.1$ which induces an allosteric interaction with the sarcoplasmic reticulum (SR) ryanodine receptor (RyR1), ultimately triggering calcium release from the SR and subsequent muscle contraction. Unique in this process is the fact that calcium entry through the $Ca_v1.1$ channel is not required for muscle contraction; rather, the $Ca_v1.1$ channel complex acts simply as a voltage sensor to initiate the contraction process.

Related to its role as a voltage sensor for the RyR1, the $Ca_v1.1$ L-type channel is implicated in two muscle disorders in humans: hypokalemic periodic paralysis (HypoPP) and malignant hyperthermia susceptibility (MHS; Table 2). Missense mutations in the CACNA1S gene encoding $Ca_v1.1$ result in reduced current density and reduced rates of activation, loss-of-function effects predicted to result in less calcium influx into muscle cells during depolarization. Although reduced calcium influx through $Ca_v1.1$ channels could indirectly affect membrane potential, given the dual functions of $Ca_v1.1$ it is also possible that the pathophysiology of HypoPP is independent of Ca^{2+} flux but is associated with an uncoupling of EC between $Ca_v1.1$ and RyR1 and indirectly results in reduced Ca^{2+} release from internal stores.

MHS is an autosomal dominant disorder characterized by a predisposition in otherwise healthy individuals for muscle hypermetabolism in response to exposure to volatile anesthetics or depolarizing muscle relaxants. The majority of patients would die without early administration of the SR calcium release inhibitor dantrolene. Most mutations associated with MHS are in the RyR1 gene, although several have also been identified in CACNA1S.

$Ca_v1.2$ Subunit: CACNA1C Gene

In both cardiac and smooth muscles, $Ca_v1.2$ L-type channels are involved in an EC coupling mechanism requiring Ca^{2+} influx through $Ca_v1.2$ channels to activate RyR2 receptors to release calcium from internal SR Ca^{2+} stores in order to initiate muscle contraction. In the cardiovascular system, $Ca_v1.2$ channels are critically linked to vasodilation and cardiac depression and are the target of the clinically used L-type channel blockers. Cardiovascular $Ca_v1.2$ L-type channels are also a target for modulation by noradrenaline and angiotensin II, both of which are critical regulators of vascular control. Within the brain, $Ca_v1.2$ is the major L-type channel, with $Ca_v1.2$ channels being diffusely distributed at low levels on cell bodies and proximal dendrites. Multiple alternatively spliced $Ca_v1.2$ variants have been identified that exhibit distinct biophysical properties, including rates of activation and current–voltage relations. Many $Ca_v1.2$ splice variations are localized to the cytoplasmic linker between domains II and III and the carboxyl region, likely reflecting cell-specific protein interactions and physiological functions. Interestingly, some splice variants exhibit cell-, tissue-, and pathophysiological state-specific expression patterns. Reflective of its physiological importance in cardiac and nervous system development, genetic deletion of $Ca_v1.2$ in mice is lethal and animals die on approximately embryonic day 12.5.

Mutations in the domain I S6 transmembrane segment of the human $Ca_v1.2$ gene are associated with a serious multisystem disorder called Timothy syndrome (TS). More than half of TS patients die between 2 and 3 years of age. Functional analysis of TS $Ca_v1.2$ mutations shows a near complete loss of voltage-dependent inactivation and altered CaMKII-dependent channel gating, both of which are predicted to prolong Ca^{2+} influx. Computer simulation of these effects suggests significant cardiac action potential prolongation and prolonged QT intervals, a phenomenon known to result in abnormal secondary depolarizations, arrhythmia, and sudden death.

$Ca_v1.3$ Subunit: CACNA1D Gene

$Ca_v1.3$ L-type channels have sometimes been termed 'neuroendocrine' L-type channels because of their presence in neuroendocrine cells, such as pancreatic β and adrenal chromaffin cells. However, they are also found in the brain, cardiac atrial myocytes, retina, ovaries, and cochlear hair cells of the ear. Within the central nervous system, $Ca_v1.3$ channels are found in most regions, although they are generally less abundant than $Ca_v1.2$ L-type channels. In contrast to the $Ca_v1.1$ and $Ca_v1.2$ L-types, $Ca_v1.3$ currents possess more negative current–voltage relations and have a lower sensitivity to block by dihydropyridines. In adult substantia nigra neurons, the

more hyperpolarized activation of Cav1.3 channels contributes to the autonomous Ca^{2+}-dependent pacemaker activity of these cells. Interestingly, the inhibition of substantia nigra pacemaking activity by the pharmacological blockade of $Ca_v1.3$ channels is protective in mouse models of Parkinson's disease.

Cav1.3 gene-deficient mice are largely anatomically and behaviorally normal, with notable exceptions. (1) The inner hair cochlea cells completely lack L-type currents, which results in deafness, and (2) sinoatrial node dysfunction results in atrial arrhythmias and bradycardia, effects reflective of the negative activation of $Ca_v1.3$ channels toward contributing to cardiac pacemaking activity and diastolic depolarization.

$Ca_v1.4$ Subunit: CACNA1F Gene

$Ca_v1.4$ currents are unique in their small unitary conductance, slow inactivation kinetics, and lack of Ca^{2+}-dependent inactivation. Similar to that for the $Ca_v1.3$ L-type, $Ca_v1.4$ channels activate at relatively negative potentials and are predicted to support a substantial resting window current.

In contrast to the rapid and transient nature of neurotransmitter release at most neuronal synapses, synaptic release from photoreceptors occurs through graded changes in membrane potential and is dependent on $Ca_v1.4$ L-type calcium channel activity. In the dark, photoreceptors tonically release glutamate, whereas light stimulus causes membrane hyperpolarization and the termination of glutamate release. The lack of calcium- and time-dependent inactivation together with the occurrence of a large window current make $Ca_v1.4$ channels ideally suited to support tonic glutamate release from photoreceptors. In support of this notion, mutations in the CACNA1F gene are implicated in the vision-related channelopathies, incomplete X-linked congenital stationary night blindness (IXLCSNB) and X-linked cone–rod dystrophy (CORDX).

More than 60 missense and truncation mutations scattered across most $Ca_v1.4$ structural regions have been identified in the CACNA1F gene of patients with IXLCSNB. Biophysical investigation of some mutations indicates no obvious measurable affect on $Ca_v1.4$ channel properties, whereas others cause either a loss of function or a gain of function. Overall, the analysis of IXLCSNB mutations to date does not give a clear or consistent indication regarding how the distinct point mutations might result in abnormal neurotransmission in photoreceptors.

CORDX is a progressive disorder including clinical features of reduced visual acuity and poor color vision. In one CORDX family, a mutation has been identified in the splice acceptor site of intron 28 in the CACNA1F gene and is predicted to cause either a premature stop codon or variable size deletions. The functional consequences of this mutation are unknown.

P/Q-Type Channel – $Ca_v2.1$ Subunit: CACNA1A Gene

P/Q-type calcium channels are one of the most abundantly expressed calcium channel subtypes in the mammalian central and peripheral nervous systems. $Ca_v2.1$ channels are highly localized at presynaptic terminals throughout the brain, spinal cord, and neuromuscular system, where they mediate calcium influx essential for a large portion of excitatory neurotransmitter release. At postsynaptic sites, $Ca_v2.1$ P/Q-type channels contribute to the precise modulation of intracellular calcium levels important for dendritic firing and also second messenger signaling relevant to calcium-dependent gene transcription. Originally defined as separate channel types, P-type and Q-type channels differ in their relative sensitivities to the spider peptide toxin ω-agatoxin IVA and the cone snail toxin ω-conotoxin MVIIC, and they also display distinct time- and voltage-dependent properties. In fact, the single $Ca_v2.1$ subunit gene encodes both P- and Q-type channels with their distinctive properties reflecting small protein alterations generated by alternative splicing. There may also be some contribution to the phenotypic differences between P-type and Q-type currents by the inclusion of different β subunit isoforms to form distinct channel complexes.

Deletion of the $Ca_v2.1$ subunit gene in mice results in animals with largely normal embryonic and early postnatal development but with the onset of severe neurological complications starting on approximately postnatal day 10 and resulting in the death of most, but not all, animals by approximately postnatal day 21 (Table 3). In surviving animals, cerebellar defects such as Purkinje cell swelling and abnormal granule cell migration are present.

Mutations in the human CACNA1A gene are associated with severe neurological channelopathies, including familial hemiplegic migraine (FHM1), episodic ataxia type 2 (EA2), and spinocerebellar ataxia type 6 (SCA6). FHM1 is a rare autosomal dominant subtype of migraine with an aura phase linked to cortical spreading depression (CSD). Mutations in CACNA1A resulting in FHM1 are primarily localized to $Ca_v2.1$ channel S4 voltage sensors and flanking regions. Phenotypic dissection of these mutations in isolation has proven difficult and may be due to the fact that distinct $Ca_v2.1$ alternatively spliced variants are differentially affected by FHM1 mutations. More revealing data from mice carrying introduced FHM1 mutations predict a P/Q-type channel gain-of-function phenotype

resulting in a lower threshold for CSD depression and an increased velocity of propagation across the cortex. It is unknown how headache pain is manifest, although one possibility is that CSD activates trigeminovascular afferents and evokes meningeal and brain stem activity consistent with migraine pain. It is also possible that $Ca_v2.1$ P/Q-type channels directly contribute to modulation of the trigeminal pain pathway.

EA2 is an autosomal dominant disorder in which patients experience spontaneous episodes of ataxia lasting from hours to days. Approximately half of patients also experience migraine-like symptoms. EA2 is genetically variable with more than 40 individual missense, truncation, and alternative splice site mutations distributed throughout the channel. Some truncation and missense mutations cause a reduction in current density, possibly due to fewer channels being properly folded and reaching the plasma membrane. Alternatively, some missense mutations show a net reduction of available channels due to a depolarizing shift in the voltage dependence of activation, an increased rate of inactivation, and a reduced rate of recovery from inactivation. Overall, all EA2 mutations exhibit a net reduction in $Ca_v2.1$-mediated calcium currents, likely resulting in reduced neurotransmitter release at critical synapses.

Patients with SCA6 have a polyglutamine (CAG) expansion in exon 47 of the CACNA1A gene, making SCA6 a member of the group of neurodegenerative disorders containing CAG repeats and which includes Huntington's disease (Table 2). Unaffected people have a low copy number of CAG repeats (4–16), whereas patients with SCA6 have expansions of more than 21 CAG repeats. The length of the expansion appears to be directly correlated with age of onset, with the greater CAG expansion being associated with early age of disease onset. Biophysical analyses of polyglutamine expansions in the $Ca_v2.1$ calcium channel show a range of effects on voltage- and time-dependent properties with a strong dependence on both auxiliary subunit and $Ca_v2.1$ subunit splice variant composition.

N-Type Channel – $Ca_v2.2$ Subunit: CACNA1B Gene

Unique among the high-voltage-activated calcium channels, the expression of N-type channels is largely restricted to neurons and neuroendocrine-derived cells (Table 1). Biophysically, N-type channels exhibit conductances intermediate between that of T-type and that of L-type channels and display time-dependent inactivation slower than the rapid inactivation of T-type channels but faster than that of L-type channels. In some cells, such as sympathetic neurons, N-type currents do not always inactivate rapidly and the decay rate can be slow and incomplete. Pharmacologically, N-type channels can be distinguished by their high sensitivity to irreversible block by ω-conotoxin GVIA, a 27-amino acid peptide from the marine hunting snail *Conus geographus*.

N-type channels are distributed throughout the peripheral and central nervous systems and appear especially concentrated in regions of high synaptic density where they are responsible for triggering neurotransmitter release at a subset of terminals. In addition to their presynaptic functions, N-type channels are also localized postsynaptically on dendrites, where they likely contribute to the integration and amplification of neural inputs. They also appear to play a role in nervous system development and neurite outgrowth.

Although no natural mutations in the CACNA1B gene have been described, $Ca_v2.2$ gene knockout and pharmacological interventions indicate that the N-type calcium channel plays a critical role in the primary afferent nociceptive pathway. N-type calcium channels are highly concentrated in the cell bodies and synaptic terminals of a subset of primary afferents that terminate in the dorsal horn of the spinal cord (mainly C-fibers and A-δ fibers). Deletion of the $Ca_v2.2$ channel gene in mice results in animals largely resistant to the induction of neuropathic and inflammatory pain but otherwise exhibiting normal sensory and motor functions (Table 3). Pharmacologically, the marine snail *C. magus* peptide toxin, ω-conotoxin-MVIIA (also known as ziconotide), reversibly blocks N-type currents and has been widely used to define the contributions of N-type channels in both animal model and human pain conditions. The intrathecal administration of ziconotide in animals shows strong analgesic effects on inflammatory pain, postsurgical pain, thermal hyperalgesia, and mechanical allodynia. In humans, the intrathecal administration of ziconotide to patients unresponsive to intrathecal opiates significantly reduces pain scores in chronic conditions such as cancer and HIV-related pain. Ziconotide has also been shown to be effective for the management of intractable spasticity following spinal cord injury.

Other behavioral phenotypes associated with deletion of the $Ca_v2.2$ gene in mice include a decrease in both the voluntary consumption and the hypnotic effects of ethanol and the accentuation of ethanol-based place preference, decreased anxiety, reduced baroreceptor response, and impaired long-term memory (Table 3).

R-Type Channel – $Ca_v2.3$ Subunit: CACNA1E Gene

A component of high-voltage-activated calcium current remains in some neurons even after the application of pharmacological agents to block L-, N-, and P/Q-type channels. Categorized as R-type for residual

or resistant, this current can be blocked by the African tarantula venom peptide SNX-482. Although it is possible that a portion of the native R-type calcium current as defined pharmacologically may result from incomplete blockade by the applied antagonists and/or to the existence of pharmacologically distinct splice variants of the P/Q- and N-type channels, the $Ca_v2.3$ subunit can account for many native R-type properties.

In some instances, the voltage-dependent parameters of $Ca_v2.3$ channels (current–voltage relationship and voltage dependence of inactivation) are more hyperpolarized than those of the other high-voltage-activated channels, and $Ca_v2.3$ channels also appear unique in being equally permeable to Ca^{2+} and Ba^{2+} and being highly sensitive to Ni^{2+} blockade, properties which are usually ascribed to T-type calcium channels.

The exact physiological functions of R-type channels are the least defined of the high-voltage-activated subtypes. Although N- and P/Q-type channels dominate neurotransmitter release, at some synapses R-type channels also contribute to release. Mice with deletion of the $Ca_v2.3$ subunit gene appear anatomically normal but exhibit a number of physiological defects, including increased anxiety, decreased spatial memory and locomotor activity, altered sperm motility, and decreased insulin secretion and glucose tolerance. Pain responses are also affected, with a decrease in acute formalin-induced inflammatory pain but an increased long-term response to formalin-induced visceral pain.

Regulation of Calcium Channel Activity – Interacting Proteins

The activities of all calcium channel subtypes are highly regulated by intracellular second messengers, which has significant consequences concerning downstream physiological functions (Figures 7 and 8). For example, the modulation of presynaptic P/Q-type and N-type channels by G-protein-coupled receptors (GPCRs) is a common strategy for regulating synaptic strength and efficacy. Many GPCRs, including those for dopamine, opioids, neuropeptide Y, glutamate,

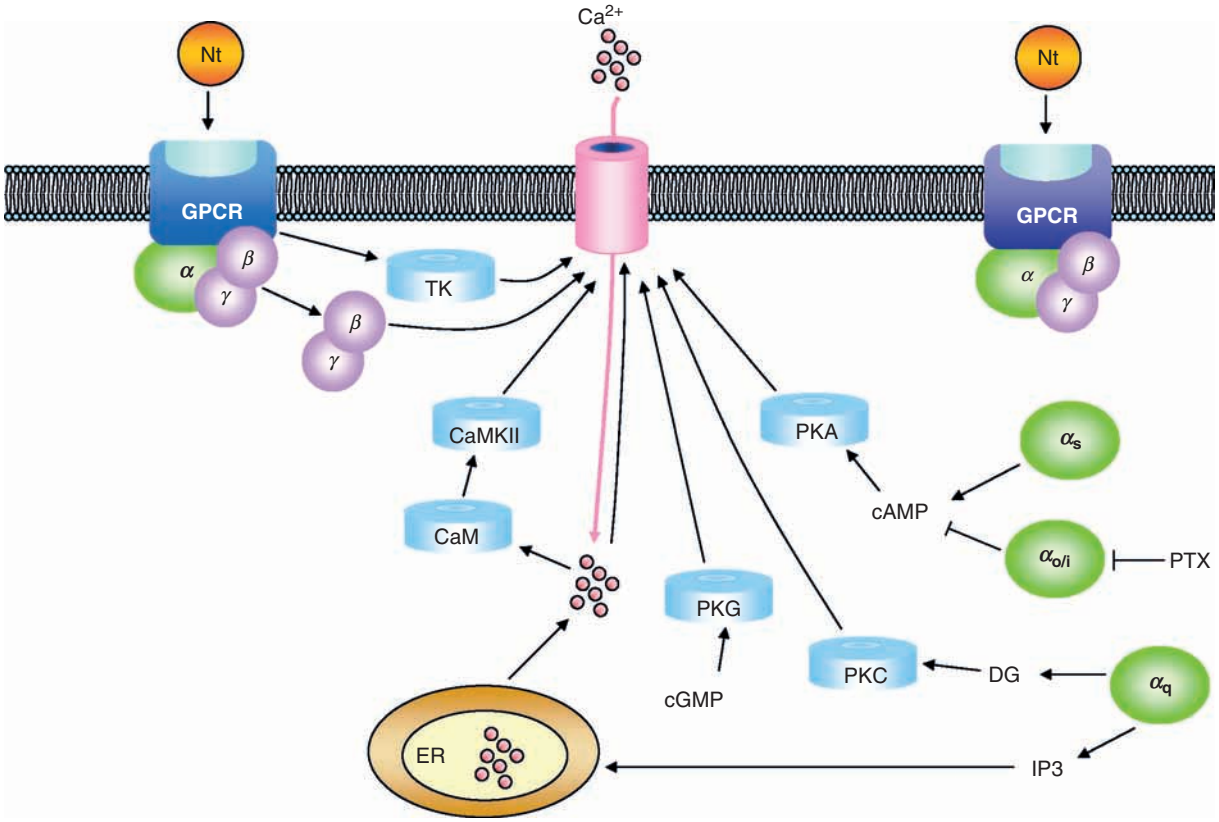

Figure 7 Second messenger-dependent modulation of calcium channels. The functional activities of all high-voltage-activated and low-voltage-activated T-type calcium channels are subject to modulation (upregulation and downregulation) by the activation of intracellular second messenger pathways acting via most known G-protein-coupled receptors (GPCR). DG, diacyl glycerol; ER, endoplasmic reticulum; IP3, inositol triphosphate; NT, neurotransmitter; PTX, pertussis toxin; PKA, protein kinase A; PKC, protein kinase C; PKG, protein kinase G; TK, tyrosine kinase. Redrawn from Chemin J, Traboulsie A, and Lory P (2006) Molecular pathways underlying the modulation of T-type calcium channels by neurotransmitters and hormones. *Cell Calium* 40: 121–134.

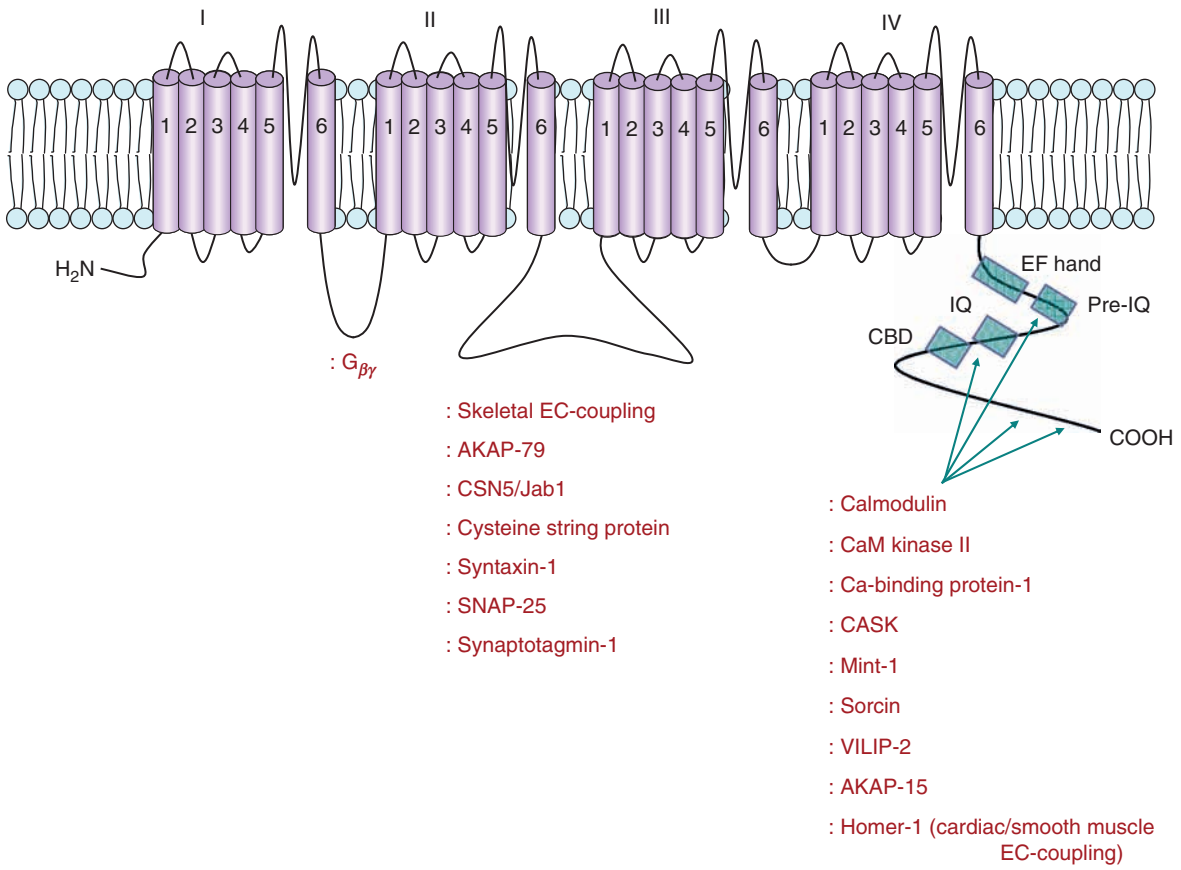

Figure 8 Voltage-activated calcium channels interact with multiple types of intracellular signaling molecules. Structure–function and biochemical analyses indicate that multiple types of intracellular proteins bind to and modulate calcium channel activity, trafficking, and interactions with intracellular signaling proteins. These include a high-affinity G-protein βγ subunit binding, distinct regions involving functional interactions with the synaptic release machinery (SNAP-25, syntaxin-1, and synaptotagmin), binding to anchoring proteins (AKAP-79 and AKAP-15), interactions with the skeletal and cardiac muscle excitation–contraction coupling machinery, as well as binding with regulatory molecules such as calmodulin, CaM kinase II, Ca^{2+}-binding protein-1, CASK, and mint-1.

GABA, substance P, canabinoids, adenosine, and somatostatin, all act to attenuate calcium presynaptic channel activity and to thereby decrease neurotransmitter vesicle fusion and release.

In another prominent example of calcium channel modulation, calmodulin acts in a Ca^{2+}-dependent feedback mechanism to either inhibit (Ca^{2+}-dependent inhibition (CDI)) or facilitate (Ca^{2+}-dependent facilitation (CDF)) high-voltage-activated calcium activity. CDF and CDI are mechanistically linked, predominantly involving several regions of the Ca_v subunit carboxyl tail, including an EF-hand motif and several downstream regions that form calmodulin (CaM), Ca^{2+}-binding protein, and CaM kinase II tethering/binding sites.

T-type channel activity can also be upregulated or downregulated by many types of neurotransmitter and hormone receptor-coupled pathways. T-type current inhibitory pathways include those mediated via acetylcholine, endocannabinoids, angiotensin II, atrial natriuretic factor, bradykinin, dopamine, enkephalin, estradiol, GABA, serotonin, somatostatin, substance P, and norepinephrine. T-type current upregulation includes pathways acting via acetylcholine, aldosterone, angiotensin II, ATP, L-cysteine, erythropoietin, estrogen, GABA, serotonin, norepinephrine, vasoactive intestinal peptide, insulin-like growth factor, and substance P. Altering T-type channel activity is predicted to critically affect neuronal and cardiac physiological processes such as firing patterns, rhythmic oscillations, and pacemaker activities.

Presynaptic P/Q-type and N-type channels provide the initial calcium trigger to initiate neurotransmitter release and contribute to anchoring synaptic vesicles at release sites. Direct interaction occurs between the synaptic protein interaction site in the domain II–III linker of P/Q-type and N-type channels and the SNARE proteins syntaxin-1A and SNAP-25. The $Ca_v2.1$ channel subunit also binds to the synaptic calcium sensor protein, synaptotagmin. Disruption of the Ca_v subunit–SNARE protein interaction in neurons inhibits neurotransmitter release.

There are examples in which parts of individual calcium channel complexes perform roles other than mediating calcium influx or voltage-sensing activities. In one instance, a β_4 subunit splice variant independently and directly interacts with a nuclear protein implicated in gene silencing and transcription. In another instance, the β_3 subunit appears to affect glucose-mediated excitability in pancreatic cells by negatively affecting inositol 1,4,5-triphosphate-induced calcium release. Furthermore, an approximately 75 kDa proteolytically cleaved C-terminal fragment of $Ca_v1.2$ L-type channel can translocate to the nucleus, where it interacts with a nuclear protein to regulate the transcription of multiple neuronal genes.

See also: Calcium Channels; Calcium Channel and Calcium-Activated Potassium Channel Coupling; Neuromodulation of Calcium Channels; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies; Voltage-Gated Potassium Channels (Kv10–Kv12).

Further Reading

Carbone E, Giancippoli A, Marcantoni A, Guido D, and Carabelli V (2006) A new role for T-type channels in fast 'low-threshold' exocytosis. *Cell Calcium* 40: 147–154.

Catterall WA (2000) Structure and regulation of voltage-gated Ca^{2+} channels. *Annual Review of Cell and Developmental Biology* 16: 521–555.

Catterall WA, Perez-Reyes E, Snutch TP, and Striessnig J (2005) International Union of Pharmacology. XLVIII: Nomenclature and structure–function relationships of voltage-gated calcium channels. *Pharmacological Reviews* 57: 411–425.

Chemin J, Traboulsie A, and Lory P (2006) Molecular pathways underlying the modulation of T-type calcium channels by neurotransmitters and hormones. *Cell Calium* 40: 121–134.

Davies A, Hendrich J, Tran Van Minh A, Wratten J, Douglas L, and Dolphin AC (2007) Functional biology of the $\alpha_2\delta$ subunits of voltage-gated calcium channels. *Trends in Pharmacological Sciences* 28: 220–228.

Gomez-Ospina N, Tsuruta F, Barreto-Chang O, Hu L, and Dolmetsch R (2006) The C terminus of the L-type voltage-gated calcium channel $Ca_v1.2$ encodes a transcription factor. *Cell* 127: 591–606.

Hering S, Berjukow S, Sokolov S, et al. (2000) Molecular determinants of inactivation in voltage-gated Ca^{2+} channels. *Journal of Physiology* 528: 237–249.

Hudmon A, Schulman H, Kim J, Maltez JM, Tsien RW, and Pitt GS (2005) CaMKII tethers to L-type Ca^{2+} channels, establishing a local and dedicated integrator of Ca^{2+} signals for facilitation. *Journal of Cell Biology* 171: 537–547.

Khosravani H and Zamponi G (2006) Voltage-gated calcium channels and idiopathic generalized epilepsies. *Physiological Reviews* 86: 941–966.

Llinas R and Yarom Y (1981) Electrophysiology of mammalian inferior olivary neurons *in vitro*. Different types of voltage-dependent ionic conductances. *Journal of Physiology* 315: 549–567.

Mangoni ME, Couette B, Marger L, Bourinet E, Striessnig J, and Nargeot J (2006) Voltage-dependent calcium channels and cardiac pacemaker activity: From ionic currents to genes. *Progress in Biophysics and Molecular Biology* 90: 38–63.

Peres-Reyes E (2002) Molecular physiology of low-voltage-activated T-type calcium channels. *Physiological Reviews* 83: 117–161.

Pietrobon D (2005) Function and dysfunction of synaptic calcium channels: Insights from mouse models. *Current Opinion in Neurobiology* 15: 257–265.

Proenza C, O'Brien J, Nakai J, Mukherjee S, Allen PD, and Beam KG (2002) Identification of a region of RyR1 that participates in allosteric coupling with the alpha(1S) (Cav1.1) II-II loop. *Journal of Biological Chemistry* 277: 6530–6535.

Schneggenburger R and Neher E (2005) Presynaptic calcium and control of vesicle fusion. *Current Opinion in Neurobiology* 15: 266–274.

Snutch TP (2005) Targeting chronic and neuropathic pain: The N-type calcium channel comes of age. *NeuroRx* 2: 662–670.

Talavera K and Nilius B (2006) Biophysics and structure–function relationships of T-type Ca^{2+} channels. *Cell Calcium* 40: 97–114.

Terlau H and Olivera B (2004) *Conus* venoms: A rich source of novel ion channel-targeted peptides. *Physiological Reviews* 84: 41–68.

Van Petegem F, Clark KA, Chatelain FC, and Minor DL (2004) Structure of a complex between a voltage-gated calcium channel beta-subunit and an alpha-subunit domain. *Nature* 429: 671–675.

Voltage-Gated Potassium Channels (Kv10–Kv12)

G A Robertson, University of Wisconsin – Madison, Madison, WI, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The 1980s and early 1990s witnessed the intersection of molecular genetics with ion channel biophysics to reveal for the first time submolecular details of ion channel structure and function. *Shaker*, cloned in Lily Jan's lab using a genetic approach in *Drosophila*, provided the first insights into the primary structure (amino acid sequence) of voltage-gated potassium (K) channels. Having the gene in hand meant channel structure could be perturbed through genetic engineering to test hypotheses about gating, permeation, and other functions in heterologous expression systems such as *Xenopus* oocytes or mammalian HEK-293 cells. A wave of additional K channels identified based on their homology with Shaker together with mutagenesis and modeling studies revealed a shared architecture of six transmembrane domains and the primary structures underlying voltage sensitivity and K selectivity.

The cloning of *ether à-go-go* (*eag*) provided a very different perspective on K channels. The name *ether à-go-go* was coined by Kaplan and Trout in 1969 to describe spontaneous *Drosophila* mutants exhibiting leg-shaking behavior induced by ether anesthesia. Many years later, Barry Ganetzky and colleagues identified the *eag* gene as a regulator of neuronal excitability at the *Drosophila* larval neuromuscular junction and predicted its gene product was a voltage-gated potassium channel subunit. The gene was cloned, and the translated sequence revealed several similarities to Shaker channels, including a conserved S4 voltage-sensor domain containing the canonical sequence of a positively charged amino acid at every third position, and a six-transmembrane topology (Figure 1). However, instead of the GYGD 'signature sequence' that rigidly define the K selectivity filters of all other known K-selective channels, eag channels had the amino acid sequence GFGN. Another important difference was a region C-terminal to the transmembrane domains that was homologous to the cyclic nucleotide binding domain (CNBD) of kinases and a cyclic nucleotide-gated ion channel, which had recently been cloned. Quantitative sequence analysis revealed that the hybrid-looking eag is more closely related to the non-selective cyclic nucleotide-gated (CNG) cation channels and inward rectifier K channels in plants than the *Drosophila* Shaker channel (Figure 2). Thus, *eag* and *Shaker* evolved from a common ancestral gene that predates the divergence of the plant and animal kingdoms.

Drosophila eag was just the tip of a phylogenetic iceberg. Homology screening and degenerate polymerase chain reaction techniques revealed a rich family tree of ion channel genes extending from *Caenorhabditis elegans* to *Homo sapiens*. Analogous to the Shaker family, the Eag family members fall into distinct subfamilies: EAG (for the original isolate, or Kv10), ERG (Eag-Related Gene, or Kv11), and ELK (Eag-LiKe, or Kv12) (see Table 1 for other nomenclatures). One member representing each subfamily was identified in *Drosophila*, whereas in mammals gene duplication has given rise during evolution to multiple members in each subfamily. Within a subfamily, the primary amino acid sequence is 60–75% identical between the first transmembrane domain and the end of the putative CNBD; for members of different subfamilies, the identity decreases to 40–50%. Diversity is further enhanced in the Kv10 and Kv11 subfamilies by alternative transcripts arising from exon splicing and different promoters. Although the current International Union of Pharmacology 'Kv' nomenclature is used in this article (Table 1), as indicated previously, Eag family members bear less similarity to Shaker and other Kv-type channels than to other families such as the CNG channels, and they are best considered members of a distinct family.

Heterologous expression in *Xenopus* oocytes showed that despite the unusual selectivity filter sequence, all Kv10–12 family members are highly selective for K^+ over Na^+ and other physiological ions. All are gated by voltage and not cyclic nucleotides. They transition from a closed or resting state at negative voltages to an open state at positive voltages. Some progress from the open to a nonconducting inactivated state at maintained positive voltages. Differences in rates of transition between these conformational states give rise to functional differences between family members (Figure 3). Compared with members of the Shaker family, opening is relatively slow. External tetraethylammonium (TEA) and 4-aminopyridine, broad K channel blockers, are relatively ineffective at blocking Eag-type channels.

The members of the Eag family are broadly expressed in the nervous system, but their physiological roles are only beginning to be understood. The focus of this article is on the biophysical and pharmacological properties of α (primary) subunits expressed in heterologous systems, which will ultimately enable us to identify the corresponding currents in native tissues. These studies describe an

unfolding and fascinating diversity of function arising from adaptations of fundamental mechanisms common to voltage-gated ion channels.

The Kv10 (EAG) Subfamily

Two genes encode the mammalian Kv10 (EAG) channels. Expressed in heterologous systems, Kv10 channels exhibit properties characteristic of delayed rectifiers (Figure 3(a)). They open and close with time constants of activation of the order of 40 ms at steps from −80 to 40 mV. They inactivate slowly, as in the case of *Drosophila* eag, or not at all. Kv10 channels are known for their prominent Cole–Moore shift, first described for potassium currents in crayfish axon. The Cole–Moore shift takes place as the holding potential becomes more negative, causing slower activation that is sigmoidal in nature. Like hyperpolarizing prepulses, physiological concentrations of Mg^{2+} slow Kv10 channel activation, reflecting channel occupancy in deeper closed states early in the activation pathway (Figure 4).

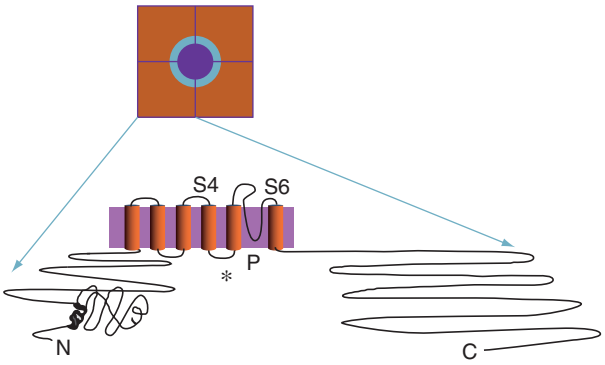

Figure 1 Kv10–12/Eag channels are composed of four subunits surrounding an ion conducting pore. Each subunit comprises six transmembrane domains, an S4 voltage sensor domain, a P region mediating ion selectivity, and an S6 domain forming the channel vestibule and gate. The asterisk represents the S4–S5 linker, which interacts with the gate-forming regions in S6.

In *Drosophila*, eag channels are critical in maintaining normal excitability in motor neurons and at the neuromuscular junction, where loss of eag function leads to repetitive action potential firing and prolonged transmitter release. In mammals, the physiological roles of the corresponding Kv10 channels are poorly understood due to a dearth of specific inhibitors, mouse mutants, or human diseases linked to either Kv10 gene. Kv10.2/EAG2 channels activate at particularly negative voltages, approximately −70 to −80 mV, close to the equilibrium potential for potassium. However, their kinetics of activation differ dramatically depending on recent history: Depolarizations closely spaced will evoke rapid Kv10.2

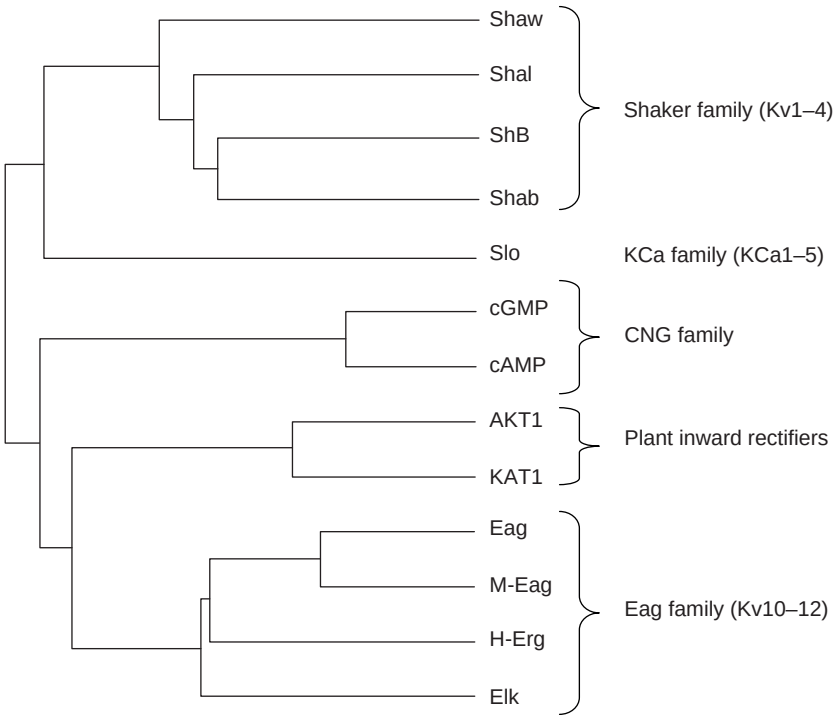

Figure 2 Phylogenetic relationships among Kv10–12/Eag and related channels. The horizontal branch lengths are inversely proportional to the similarity between the sequences. Reproduced from Warmke JW and Ganetzky B (1994) A family of potassium channel genes related to eag in *Drosophila* and mammals. *Proceedings of the National Academy of Sciences of the United States of America* 91: 3438–3442, with permission.

activation, whereas longer repolarizing periods of 3–8 s significantly delay activation due to the Cole–Moore effect. These properties may contribute to activity-dependent plasticity in neurons.

What is the biophysical basis for the Cole–Moore shift in Kv10/EAG channels? The sigmoidal activation reflects conformational changes in the S1–S4 voltage sensor in each subunit. These conformational changes can be monitored by recording gating currents or the quenching of fluorescent moieties attached at locally introduced cysteines. Changes in membrane potential evoke two components of conformational change in the voltage sensor. One carries most of the gating charge and is relatively insensitive to modulation by Mg^{2+}. The other carries little gating current but can be detected by an optical reporter near S4; this component is sensitive to Mg^{2+} and corresponds to rate-limiting movements of S4 in the activation pathway. Moreover, the voltage dependence of fluorescence quenching is indistinguishable from that of the Cole–Moore effect that slows activation at negative voltages. Thus, the activation pathway in Kv10 channels includes transitions from deep closed states that are rate limiting and modulated by Mg^{2+}. This modulation arises from the coordination of Mg^{2+} or other divalent cations between acidic residues in the S2 and S3 transmembrane domains. The modulation occurs noncooperatively, altering the conformational change of each voltage sensor in the stepwise transitions to channel opening first modeled by Hodgkin and Huxley.

Table 1 Nomenclature describing Eag family members

IUPHAR	*Protein*	*Gene names*	*HGNC*
Kv10.1	EAG1	*ether a-go-go, eag, EAG1*	*KCNH1*
Kv10.2	EAG2	*EAG2*	*KCNH5*
Kv11.1	ERG1	*eag-related gene, seizure, ERG1*	*KCNH2*
Kv11.2	ERG2	*ERG2*	*KCNH6*
Kv11.3	ERG3	*ERG3*	*KCNH7*
Kv12.1	ELK3	*eag-like, elk, ELK3*	*KCNH8*
Kv12.2	ELK2	*ELK2*	*KCNH3*
Kv12.3	ELK1	*ELK1*	*KCNH4*

HGNC, HUGO Gene Naming Committee; IUPHAR, International Union of Pharmacology.

The Kv11 (ERG) Subfamily

The Kv11/ERG subfamily contains the single gene *seizure* in *Drosophila* and three members in mammals. *Kv11.2* and *11.3* (*ERG2* and *ERG3*) are primarily expressed in the central nervous system, whereas *Kv11.1/ERG1* is expressed in the central nervous system, the heart, and a wide range of other tissues. There is significant overlap of the encoding mRNA in regions of the brain. The currents exhibit different kinetic profiles but all are blocked by class III antiarrhythmic drugs known to block human Kv11.1/hERG1. Kv11.1–3/ERG1–3 subunits are capable of forming heteromultimers in heterologous systems, providing the possibility of tremendous structural and functional diversity where they are co-expressed in native tissues.

Human Kv11.1/ERG1 and Cardiac Excitability

Of all the members of the Kv10–12 family, the human ERG1 (hERG1/Kv11.1) is by far the most studied and best understood, and because it is widely known by its original nomenclature it will be referred to here as hERG1. The *hERG1* gene encodes the channels underlying cardiac I_{Kr}, a potassium current critical in ventricular repolarization during each beat of the heart. Mutations in *hERG1* or a block of hERG1 channels by a wide range of drugs, often intended for other therapeutic targets, can cause long QT syndrome (LQTS), catastrophic arrhythmia, and sudden cardiac death. Some known hERG blockers still in

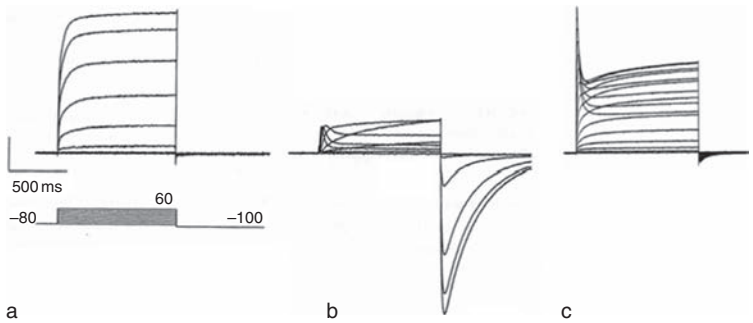

Figure 3 Two-electrode voltage clamp recordings of Kv10–12 channels expressed in *Xenopus* oocytes. (a) Mouse Kv10.1/EAG. (b) Human Kv11.1/ERG1. (c) Mouse Kv12.2/ELK2. (a) Adapted from Herzberg IM, Trudeau MC, and Robertson GA (1998) Transfer of rapid inactivation and E-4031 sensitivity from HERG to M-EAG channels. *Journal of Physiology* 511(1): 3–14, with permission from Blackwell publishing. (b) Trudeau MC, Titus SA, Branchaw JL, Ganetzky B, and Robertson GA (1999) Functional analysis of a mouse brain Elk-type K^+ channel. *Journal of Neuroscience* 19: 2906–2918, with permission.

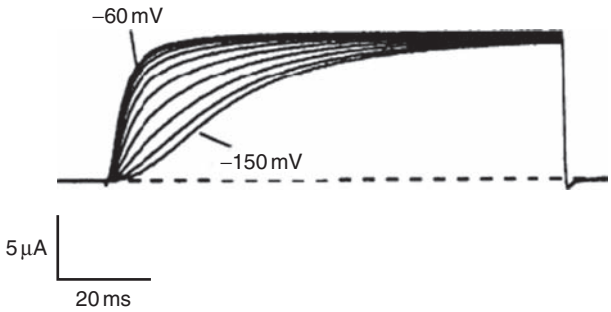

Figure 4 Hyperpolarizing prepulses induce delay and sigmoidal activation characteristic of the Cole–Moore shift in rat Kv10/EAG channels expressed in *Xenopus* oocytes. Currents were evoked by a step to 60 mv from holding potentials ranging from −60 to −150 mV. Reproduced from Ludwig J, Terlau H, Wunder F, et al. (1994) Functional expression of a rat homologue of the voltage gated either á go-go potassium channel reveals differences in selectivity and activation kinetics between the *Drosophila* channel and its mammalian counterpart. *EMBO Journal* 1(19): 4451–4458, with permission from Nature Publishing Group.

limited clinical use elicit life-threatening arrhythmias with an incidence as high as 3–5%. An estimated 50% of lead compounds are dropped from pharmaceutical development because they inhibit hERG channels and thus represent a cardiotoxic risk. Because these problems have significantly hampered new drug discovery efforts for all therapeutic targets, understanding the mechanism of hERG1 block and the constellation of susceptibility factors for acquired LQTS is a goal of enormous clinical importance.

hERG1 Gating in the Cardiac Action Potential

In cells expressing hERG channels, voltage steps to 40 or 60 mV evoke little outward current compared to the large inward current elicited upon subsequent repolarization to voltages negative to the channel's reversal potential, so they are sometimes said to be inwardly rectifying (Figure 3(b)). These properties dictate hERG1's physiological role in the heart but, despite the unusual hERG current profile, the underlying gating transitions are mechanistically similar to those of Shaker and other voltage-gated K channels. hERG1 channels are activated by membrane depolarization but with a slow time course and correspondingly slow movement of the S4 domain; they also inactivate, entering into a state that is nonconducting but distinct in conformation from the closed or resting state. However, because inactivation is approximately 10 times faster than activation, channels spend little time in the open state at positive voltages (Figure 5(a)). Upon repolarization, channels recover from inactivation and revisit the open state. Deactivation is approximately 10 times slower than recovery from inactivation, causing extended dwell times

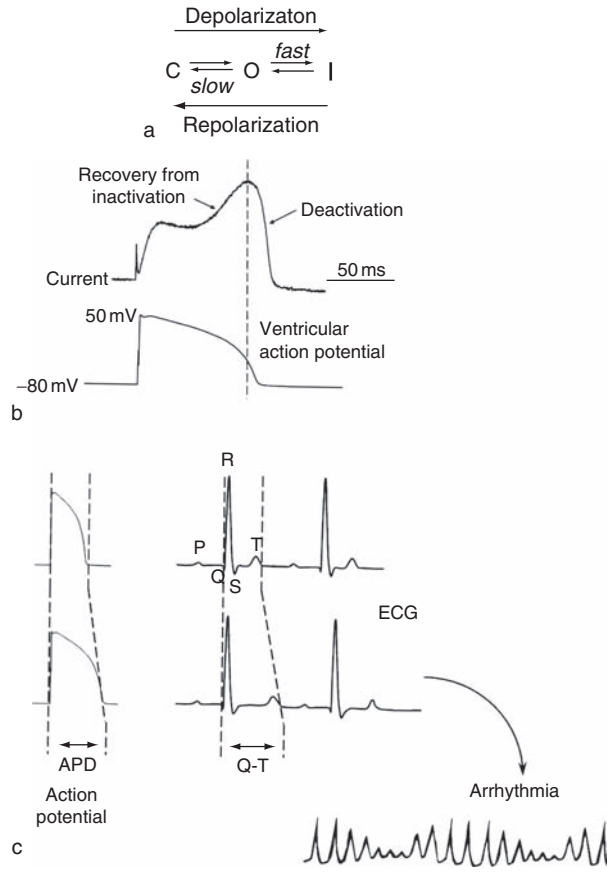

Figure 5 Human Kv11.1/hERG1 channel characteristics. (a) State diagram for hERG1 gating. (b) hERG1 expressed in *Xenopus* oocytes subject to a two-electrode voltage clamp command signal mimicking a ventricular action potential. From H Sale, J Wang, and G Robertson, unpublished results. (c) The relationship between action potential duration (APD) and the interval between the Q and T waves (QT) of the surface electrocardiogram (ECG). Prolongation of the APD leads to increased QT interval and electrical instabilities triggering *torsades de pointes* arrhythmia, also apparent on the surface ECG.

in the open state. The quantitative relationships between these gating transitions translate into a current that is suppressed at positive voltages but dramatically rebounds upon repolarization, thus bearing little resemblance to Shaker-related currents. The rebounding tail current mediates I_{Kr}'s role in terminal repolarization of the ventricular action potential (Figure 5(b)). When I_{Kr} is reduced by mutations or drug block, ventricular myocytes are slower to repolarize and action potentials are prolonged (Figure 5(c)). The duration of the ventricular action potential (APD) is directly reflected in the QT interval of the surface electrocardiogram, hence the term 'long QT syndrome.' The prolonged APD destabilizes membrane excitability both locally and across the wall of the ventricle, leading in some cases to catastrophic arrhythmias known as *torsades de pointes* ('twists of points'; Figure 5(c)).

hERG1 Activation and Deactivation

Channel activation occurs when charge dipoles in the S1–S4 voltage sensor (Figure 1) move in response to changes in transmembrane potential and trigger opening of the activation gate. This coupling is mediated by the S4–S5 linker, which, based on structural studies of the mammalian Shaker channel Kv1.2, is thought to wrap around and bear down on the S6 domains to close the channel. In hERG1 channels, an electrostatic interaction between the S4–S5 linker and the S6 domain is important in stabilizing the closed conformation at negative potentials; mutations can cause the channel to re-open upon further hyperpolarization. Because of the temporal overlap of the inactivation process, the activation time course must be measured in mutants in which inactivation has been removed by mutagenesis or measured by indirect approaches that determine the numbers of channels that have opened on their transit to the inactivated state.

Nature has solved in different ways the problem of how channels gate access of ions to their pores. For example, Shaker channels gate by a hydrophobic seal at a kink or bend in the S6 transmembrane domain where the subunits overlap, whereas cyclic nucleotide-gated channels gate in the narrow part of the selectivity filter at the other end of the channel vestibule. For hERG1 channels, mutations at several sites in the more C-terminal half of the S6 domain interfere with channel closing and also promote reversal of inhibition by hERG1 blockers. These findings are interpreted to suggest that a single gate regulates ion flux and traps drugs in the internal vestibule. The precise location of the gate is unknown. Given the importance of hERG1 channels as a target for acquired LQTS, it is of great interest to map the activation gate and thus define an important dimension of the drug-binding vestibule.

hERG1 Inactivation Mechanisms

The molecular mechanisms underlying hERG1 inactivation are complex. A faster version of the pore-based C-type inactivation in Shaker, hERG1 inactivation is slowed by elevated external concentrations of potassium and TEA. Mutations at precisely homologous sites in the pore region disrupt inactivation in Shaker and hERG1. However, what distinguishes hERG inactivation is an inherent voltage sensitivity not coupled to channel opening. Efforts to understand the inactivation mechanism have focused on two regions: the S4 component of the voltage sensor and the sequences in and around the K selectivity filter. The overriding hypothesis is that these two regions must be linked to account for voltage-dependent inactivation. As has been shown for other voltage-gated channels, perturbation of residues in the S1–S4 voltage sensor disrupts activation gating. However, also within the S4 domain are residues whose perturbation selectively disrupts inactivation. It is interesting to note that voltage sensor movements exhibit two components, whether observed with gating currents or the quenching of fluorescent labels covalently linked to substituted cysteines neighboring S4. The slow component correlates with activation/deactivation gating. Kinetic analysis suggests the rapid component is consistent with inactivation, but it fails to change with manipulations that alter macroscopic current inactivation. Thus, the location and nature of the voltage sensor for inactivation remain uncertain. An important structural domain in this process is the S5–P linker, where single mutations can disrupt both inactivation and activation gating. This region possesses α-helical secondary structure and is sufficiently long to bridge the distance between S4 and the outer pore in hERG1 models.

Mechanisms of hERG1 Drug Block

The process of inactivation is at the center of an ongoing debate about the mechanism of hERG drug block underlying acquired LQTS. Like most drugs, those that block hERG1 channels are hydrophobic, crossing the plasma membrane to exert their actions inside the cell. hERG1 channels are blocked in a use-dependent manner: The drug enters the open channel, binds, and then somehow prevents ion conduction. Mutation studies show that two aromatic residues, Y652 and F656, are critical for drug block (Figure 6). The sensitivities for three distinct compounds were decreased

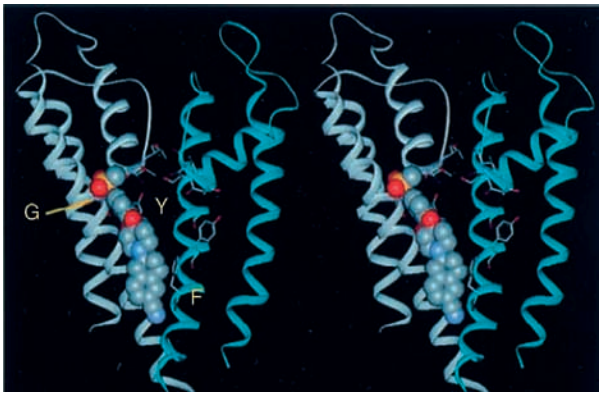

Figure 6 Stereo image of a hERG drug blocker docked into the channel vestibule formed by the S6 domains. Y652 and F656 are highlighted. G, glycine; Y, tyrosine 652; F, phenylalanine 656. Reproduced from Mitcheson JS, Chen J, Lin M, Culberson C, and Sanguinetti MC (2000) A structural basis for drug-induced long QT syndrome. *Proceedings of the National Academy of Sciences of the United States of America* 97: 12329–12333, with permission.

by approximately 100-fold, providing good evidence that the mutations disrupt the binding site rather than allosteric gating transitions. Mutations in and around the outer pore region also disrupt drug block but probably do not contribute to the drug-binding site *per se*.

How the binding of drug translates to channel inhibition is the subject of competing hypotheses: a simple pore occlusion, on the one hand, and promotion of inactivation, also known as the modulated receptor hypothesis, on the other. Most mutations that disrupt drug block also remove or dramatically slow inactivation, as if inactivation gating were critical to drug inhibition. However, there is no evidence that the inactivated state exhibits a higher affinity for the drug, which is a prediction of the modulated receptor hypothesis; voltages that drive the channel into the inactivated state were found not to promote block. The problem with a simple pore-block model is that mammalian EAG channels are quite insensitive to the pharmacopeia of hERG1-blocking drugs despite possessing a highly conserved vestibule, including the key drug-binding residues identified in hERG1. What most distinguishes EAG from hERG1 channels, besides insensitivity to drug block, is the absence of inactivation, once again raising the possibility that drug block requires inactivation. Indeed, sensitivity of bovine EAG to the hERG1 blocker dofetilide was increased nearly 30-fold by mutations at corresponding sites to residues critical for inactivation in hERG – S620 and S631. Focused efforts to clarify the structural details of the drug-binding site and its interaction with channel gating mechanisms will be important in facilitating the development of drugs free of cardiotoxic side effects.

Modulation of Gating by the N-Terminus

With the exception of the subunit encoded by the hERG1b alternate transcript (see below), all channels in the Eag family share homology in a region comprising the first approximately 135 amino acid residues sometimes referred to as the 'eag domain.' An X-ray crystallography study of the eag domain revealed a structure from residues 26 to 135 similar to Per-Arnt-Sim (PAS) domains implicated in protein–protein interactions. The first 25 amino acids were unordered in the crystal. Following the PAS domain, the N-terminal structures markedly diverge; the hERG1a N-terminus has a total length of approximately 396 amino acids, much longer than the N-termini of EAG or ELK channels. The hERG1 N-terminus docks to the hydrophobic core via the S4–S5 linker, where three distinct domains respectively slow deactivation and promote activation and C-type inactivation. Deletion of the first 16 amino acids selectively increases deactivation rate, an effect that can be largely reversed by adding the soluble N-terminal 16mer peptide to excised macropatches containing N-truncated channels. Slow deactivation in N-truncated constructs could also be reconstituted by the addition of the entire purified 1–137 amino acid polypeptide to the oocyte cytosol, but it is unknown whether the PAS domain *per se* directly slows deactivation or rather serves as the docking domain for the more N-terminal sequences. Inactivation is unaffected by the 2–16 amino acid deletion, but it is slowed by full N-terminal truncations. Deletion of the 'proximal domain,' residues 137–373, accelerates activation with no effect on deactivation. The N-terminus is thus highly complex and modulates many different components of basic channel function, the physiological ramifications of which remain to be determined.

Two hERG1 Subunits Contribute to I_{Kr}

Much of the work on hERG1 channels has focused on what we now term the hERG1a subunit, which can form functional homomeric channels with properties largely similar to those of I_{Kr}. A second subunit, hERG1b, is identical to hERG1a except that it has a much shorter N-terminus with a very different primary sequence. The two subunits are encoded by alternate transcripts from the hERG1 gene, meaning that they each have their own promoter sequence. Heterologous expression of hERG1b produces currents with extremely low efficiency, but the 1b subunits readily assemble with hERG1a subunits to form heteromeric channels. Together, 1a and 1b form channels more like native I_{Kr}, with reduced rectification and faster deactivation than hERG1a homomeric channels. Both 1a and 1b are expressed in heart, where they exhibit similar T-tubular distribution in cardiomyocytes. The two subunits can be co-immunoprecipitated from heart tissue, supporting the hypothesis that heteromers of hERG1a and-1b underlie cardiac I_{Kr}. The stoichiometric ratio of the two subunits is not known, nor do we know whether all channels are heteromeric or if regional pools of homomeric hERG1a channels exist. The mRNA expression of hERG1b is much higher in the heart than in the brain, but further work is required to determine its distribution there.

hERG Channels and Neuronal Function

Although the *hERG1* gene was originally cloned from the human hippocampus, we know relatively little about its functional role in the brain. In computer models and neuronal-type cells in culture treated with hERG blockers, it is clear that Kv11/ERG channels can regulate firing frequency and cause spike frequency adaptation during a burst of action potentials. Their gating properties seem especially well suited to

modulating slow-wave activity. ERG channels have also been identified in astrocytes and microglia.

Kv11/ERG channels may also be important in the developing nervous system. They are expressed in proliferating neuronal precursors in neural crest, where they are subsequently downregulated in favor of other types of potassium channels as the cells differentiate. In this regard, the neuronal precursors are like neuroblastoma and other tumor-related cells, which exhibit high levels of ERG expression. In both cases, high ERG levels are correlated with a relatively depolarized resting membrane potential and cellular proliferation. Kv11/ERG channels are also important functionally in the developing mouse spinal cord, where they regulate oscillations of GABAergic interneurons.

Despite the widespread expression of hERG1 in the brain, LQTS patients, who are heterozygous for *hERG1* mutations, have not presented with consistent neuropathological phenotypes. Perhaps the heart is less tolerant of reduced *hERG1* gene copy number than the brain, in which the additional expression of *hERG2* and *hERG3* might be expected to compensate for reduced hERG1 channel function. It is also interesting to note that some drugs used to treat mental illness are known to block hERG1 channels in pharmacological assays. Whether their therapeutic benefit is related to hERG1 (or hERG2 or -3) block is an important unanswered question.

The Kv12 (ELK) Subfamily

The Kv12/ELK subfamily comprises three members in mammals, all primarily expressed in the nervous system and in patterns that significantly overlap in the brain. They differ from Kv11/ERG channels in that they are insensitive to E-4031 or the other ERG channel blockers. None of the Kv12 isoforms exhibit the Cole–Moore shift of Kv10/EAG subfamily channels. Overall, we know less about Kv12 channels than all other members of the Eag family.

The functional roles of Kv12 channels in the nervous system are unknown, but we can make certain predictions based on the biophysical properties of these channels expressed in heterologous systems. Kv12.1/ELK1 currents exhibit delayed rectifier characteristics with slow activation and deactivation kinetics, suggesting they may repolarize action potentials or regulate burst frequency. Kv12.2/ELK2 channels are quite different, exhibiting rapid activation and inactivation (Figure 3(c)). Like hERG1 currents, they undergo C-type inactivation; however, because they activate more rapidly and inactivate more slowly, they produce more current (i.e., are less rectifying) at positive voltages. Voltage clamp commands with repolarizing ramps designed to mimic slow-wave neural activity suggest that Kv12.2 channels may play a dual role in membrane excitability (Figure 7). Upon depolarization, the channels activate and inactivate, producing an early peak of current with a duration of the order of milliseconds. However, when the step is followed by a slow ramp repolarization, the channels recover from inactivation and rebound to produce a second phase of current before slowly closing within hundreds of milliseconds or even seconds. Thus, in native tissues, one might expect the 'early' Kv12.2 currents to help repolarize action potentials or regulate short bursts of activity, whereas the 'late' wave could shape slow-wave activity. The subunits encoded by the three Kv12 subfamily members can co-assemble in heterologous expression systems and, given their overlapping patterns in the brain, may generate native channels of considerably diverse composition and function.

The Future: Functional Roles of Kv10–12 Channels in the Nervous System

This article focuses on the structure and function of α subunits of the Kv10–12/Eag channel family; for important work on channel modulation by phosphorylation or β subunits and on the roles of these channels in cell proliferation, smooth muscle activity, endocrine function, and the autonomic nervous system, the reader is referred to the primary literature and

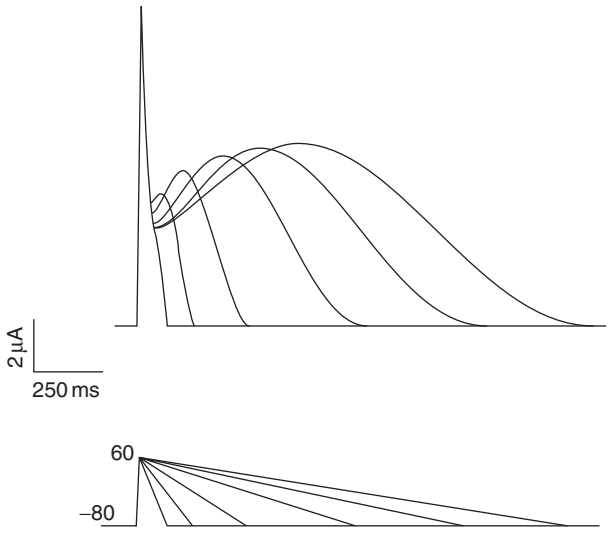

Figure 7 Mouse Kv12.2/ELK2 currents expressed in *Xenopus* oocytes in response to a ramp with a 3-ms rise time and variable repolarization rates. Currents exhibit two phases: The first peak arises from activation and inactivation of the channels, and the slower phase occurs as channels recover from inactivation and slowly return to the closed or resting state. Adapted from Trudeau MC, Titus SA, Branchaw JL, Ganetzky B, and Robertson GA (1999) Functional analysis of a mouse brain Elk-type K^+ channel. *Journal of Neuroscience* 19: 2906–2918, with permission.

excellent reviews on these subjects. Even within the narrow focus of this article there remain many details and fundamental mechanisms yet to be understood. For example, why is the putative CNBD in the cytoplasmic C-terminus so highly conserved if not for gating by cyclic nucleotides? Indeed, what functional roles can be assigned to the otherwise divergent C-terminus, which makes up approximately half of each channel polypeptide and is the least understood region of all? One C-terminal domain is known in Kv10/EAG channels to promote oligomerization and another in Kv11.1/hERG1 is known to make contact along the secretory pathway. A much higher degree of conservation is observed in the N-terminus, including the structurally determined PAS domain, and yet the only known functions of the N-terminus are in gating processes that differ dramatically among channel types. Clearly, far more information remains to be discovered.

One of the greatest challenges for those interested in the Kv10–12/Eag family of K channels is determining their physiological roles in the nervous system, in which they are widely expressed. To accomplish this, we need specific inhibitors as well as approaches that reduce expression of the individual genes, preferably under inducible control to minimize the compensatory expression of other genes. To date, inhibition is limited to toxins and blockers of the Kv11/ERG subfamily and to mutant phenotypes of *Drosophila eag* and *seizure* (*erg*) and human ERG (LQTS patients). Members of the Kv10/EAG and Kv12/ELK subfamilies lack specific inhibitors and correspondingly little is known about their functions. Current efforts to identify scorpion toxins that discriminate among the various channel types should significantly advance our understanding of their roles at the cellular level. Advances at the organismal level are hindered by the absence of links of known inherited neurological diseases with mutations in the Kv10–12/Eag family. Perhaps such mutations go undetected because their loss of function is lethal at very early developmental stages or compensated by the activities of related family members that overlap in cellular distribution and functional properties. A concerted effort to develop disease models, together with ongoing structure–function studies elucidating their unique functional properties, will surely lead to significant advances in our understanding of Kv10–12/Eag potassium channels and their functional roles in the nervous system.

See also: Calcium Channel and Calcium-Activated Potassium Channel Coupling; Inwardly Rectifying Potassium Channels; Ion Channel Localization in Axons; Ion Channel Localization in Cell Bodies and Dendrites; Large Conductance Calcium-Activated Potassium Channels; Potassium Channel Regulation; Two-P-Domain (K_{2P}) Potassium Channels: Leak Conductance Regulators of Excitability; Voltage Gated Potassium Channels: Structure and Function of Kv1 to Kv9 Subfamilies.

Further Reading

Arcangeli A, Rosati B, Cherubini A, et al. (1997) HERG- and IRK-like inward rectifier currents are sequentially expressed during neuronal development of neural crest cells and their derivatives. *European Journal of Neuroscience* 9: 2596–2604.

Bannister JP, Chanda B, Bezanilla F, and Papazian DM (2005) Optical detection of rate-determining ion-modulated conformational changes of the ether-a-go-go K^+ channel voltage sensor. *Proceedings of the National Academy of Sciences of the United States of America* 102: 18718–18723.

Bauer CK and Schwarz JR (2001) Physiology of EAG K^+ channels. *Journal of Membrane Biology* 182: 1–15.

Ficker E, Jarolimek W, and Brown AM (2001) Molecular determinants of inactivation and dofetilide block in ether a-go-go (EAG) channels and EAG-related K(+) channels. *Molecular Pharmacology* 60: 1343–1348.

Furlan F, Taccola G, Grandolfo M, et al. (2007) ERG conductance expression modulates the excitability of ventral horn GABAergic interneurons that control rythmic oscillations in the developing mouse spinal cord. *Journal of Neuroscience* 27: 919–928.

Ganetzky B and Wu CF (1983) Neurogenetic analysis of potassium currents in *Drosophila*: Synergistic effects on neuromuscular transmission in double mutants. *Journal of Neurogenetics* 1: 17–28.

Herzberg IM, Trudeau MC, and Robertson GA (1998) Transfer of rapid inactivation and E-4031 sensitivity from HERG to M-EAG Channels. *Journal of Physiology* 511(1): 3–14.

Jones EM, Roti Roti EC, Wang J, Delfosse SA, and Robertson GA (2004) Cardiac IKr channels minimally comprise hERG 1a and 1b subunits. *Journal of Biological Chemistry* 279: 44690–44694.

Kiehn J, Lacerda AE, and Brown AM (1999) Pathways of HERG inactivation. *American Journal of Physiology* 277: H199–H210.

Lees-Miller JP, Kondo C, Wang L, and Duff HJ (1997) Electrophysiological characterization of an alternatively processed ERG K^+ channel in mouse and human hearts. *Circulation Research* 81: 719–726.

Liu J, Zhang M, Jiang M, and Tseng GN (2002) Structural and functional role of the extracellular s5-p linker in the HERG potassium channel. *Journal of General Physiology* 120: 723–737.

London B, Trudeau MC, Newton KP, et al. (1997) Two isoforms of the mouse ether-a-go-go-related gene coassemble to form channels with properties similar to the rapidly activating component of the cardiac delayed rectifier K^+ current. *Circulation Research* 81: 870–878.

Ludwig J, Owen D, and Pongs O (1997) Carboxy-terminal domain mediates assembly of the voltage-gated rat ether-a-go-go potassium channel. *EMBO Journal* 16: 6337–6345.

Ludwig J, Terlau H, Wunder F, et al. (1994) Functional expression of a rat homologue of the voltage gated ether á go-go potassium channel reveals differences in selectivity and activation kinetics between the *Drosophila* channel and its mammalian counterpart. *EMBO Journal* 13: 4451–4458.

Mitcheson JS, Chen J, Lin M, Culberson C, and Sanguinetti MC (2000) A structural basis for drug-induced long QT syndrome. *Proceedings of the National Academy of Sciences of the United States of America* 97: 12329–12333.

Morais Cabral JH, Lee A, Cohen SL, Chait BT, Li M, and Mackinnon R (1998) Crystal structure and functional analysis of the HERG potassium channel N terminus: A eukaryotic PAS domain. *Cell* 95: 649–655.

Piper DR, Hinz WA, Tallurri CK, Sanguinetti MC, and Tristani-Firouzi M (2005) Regional specificity of human ether-a'-go-go-related gene channel activation and inactivation gating. *Journal of Biological Chemistry* 280: 7206–7217.

Robertson GA, Warmke JM, and Ganetzky B (1996) Potassium currents expressed from *Drosophila* and mouse eag cDNAs in *Xenopus* oocytes. *Neuropharmacology* 35: 841–850.

Saganich MJ, Vega-Saenz de Miera E, Nadal MS, Baker H, Coetzee WA, and Rudy B (1999) Cloning of components of a novel subthreshold-activating K(+) channel with a unique pattern of expression in the cerebral cortex. *Journal of Neuroscience* 19: 10789–10802.

Sanguinetti MC, Jiang C, Curran ME, and Keating MT (1995) A mechanistic link between an inherited and an acquired cardiac arrhythmia: HERG encodes the IKr potassium channel. *Cell* 81: 299–307.

Sanguinetti MC and Tristani-Firouzi M (2006) hERG potassium channels and cardiac arrhythmia. *Nature* 440: 463–469.

Schonherr R and Heinemann SH (1996) Molecular determinants for activation and inactivation of HERG, a human inward rectifier potassium channel. *Journal of Physiology (London)* 493: 635–642.

Schonherr R, Mannuzzu LM, Isacoff EY, and Heinemann SH (2002) Conformational switch between slow and fast gating modes: Allosteric regulation of voltage sensor mobility in the EAG K^+ channel. *Neuron* 35: 935–949.

Shi W, Wang HS, Pan Z, et al. (1998) Cloning of a mammalian elk potassium channel gene and EAG mRNA distribution in rat sympathetic ganglia. *Journal of Physiology (London)* 511: 675–682.

Smith PL, Baukrowitz T, and Yellen G (1996) The inward rectification mechanism of the HERG cardiac potassium channel. *Nature* 379: 833–836.

Smith PL and Yellen G (2002) Fast and slow voltage sensor movements in HERG potassium channels. *Journal of General Physiology* 119: 275–293.

Terlau H, Ludwig J, Steffan R, Pongs O, Stuhmer W, and Heinemann SH (1996) Extracellular Mg^{2+} regulates activation of rat eag potassium channel. *Pflugers Archiv* 432: 301–312.

Torres AM, Bansal PS, Sunde M, et al. (2003) Structure of the HERG K^+ channel S5P extracellular linker: Role of an amphipathic alpha-helix in C-type inactivation. *Journal of Biological Chemistry* 278: 42136–42148.

Tristani-Firouzi M, Chen J, and Sanguinetti MC (2002) Interactions between S4–S5 linker and S6 transmembrane domain modulate gating of HERG K^+ channels. *Journal of Biological Chemistry* 277: 18994–19000.

Trudeau MC, Titus SA, Branchaw JL, Ganetzky B, and Robertson GA (1999) Functional analysis of a mouse brain Elk-type K^+ channel. *Journal of Neuroscience* 19: 2906–2918.

Trudeau MC, Warmke JW, Ganetzky B, and Robertson GA (1995) HERG, a human inward rectifier in the voltage-gated potassium channel family. *Science* 269: 92–95.

Viloria CG, Barros F, Giraldez T, Gomez-Varela D, and de la Pena P (2000) Differential effects of amino-terminal distal and proximal domains in the regulation of human erg K^+ channel gating. *Biophysical Journal* 79: 231–246.

Wang J, Myers CD, and Robertson GA (2000) Dynamic control of deactivation gating by a soluble amino-terminal domain in HERG K(+) channels. *Journal of General Physiology* 115: 749–758.

Warmke JW and Ganetzky B (1994) A family of potassium channel genes related to eag in *Drosophila* and mammals. *Proceedings of the National Academy of Sciences of the United States of America* 91: 3438–3442.

Wimmers S, Wulfsen I, Bauer CK, and Schwarz JR (2001) Erg1, erg2 and erg3 K channel subunits are able to form heteromultimers. *Pflugers Archiv* 441: 450–455.

Zou A, Lin Z, Humble M, et al. (2003) Distribution and functional properties of human KCNH8 (Elk1) potassium channels. *American Journal of Physiology – Cell Physiology* 285: C1356–C1366.

Vomeronasal Accessory System

T Leinders-Zufall and F Zufall, University of Saarland School of Medicine, Homburg/Saar, Germany

© 2009 Elsevier Ltd. All rights reserved.

The Accessory Olfactory System

The olfactory system has evolved complex mechanisms to detect a vast range of molecular cues. Mammals possess several chemosensory systems, including the main olfactory system (MOS) and the accessory olfactory system (AOS), which is also known as the vomeronasal system. In addition, several other chemosensory subsystems, such as the septal organ of Masera and the Grueneberg ganglion, have been recognized (Figure 1(a)).The AOS has attracted a great deal of attention because of its involvement in the detection of chemical stimuli of a social nature, such as pheromones. Karlson and Lüscher introduced the term "pheromone" and defined it as "a substance secreted to the outside of an individual and received by a second individual of the same species in which it . . . ."

The AOS consists of (1) the vomeronasal organ (VNO, also known as Jacobson's organ), a chemoreceptive structure situated at the base of the nasal septum and which houses the vomeronasal sensory neurons (VSNs); (2) the accessory olfactory bulb (AOB), a structure located in the forebrain, dorsocaudal to the main olfactory bulb (MOB), that receives input from the VNO and serves as the first processing center of vomeronasal information; and (3) higher olfactory centers, such as the medial amygdala, that receive direct or indirect information from the AOB. The medial amygdala is highly interconnected with the bed nucleus of stria terminalis. Both nuclei project to hypothalamic areas important for controlling neuroendocrine events that, in turn, regulate social and reproductive behaviors.

Basic Anatomical Organization of the VNO

Many mammals possess a bilateral, cigar-shaped VNO (Figure 1) that is located at the base of the nasal septum, anterior and ventral to the main olfactory epithelium (MOE). Each VNO is encased in a cartilaginous capsule, the vomer bone. Chemosensory stimuli gain access to the VNO through its single rostral opening, which in rodents opens into the nasal cavity. The main structural elements of the VNO are best seen in coronal sections and include a central lumen, a medially located crescent-shaped sensory epithelium, and a laterally located nonsensory epithelium, together with an extensive network of blood sinuses surrounded by a band of cavernous erectile tissue (Figure 1(c)). These latter components provide the hardware for a vomeronasal pump that aids in stimulus delivery.

The sensory epithelium consists primarily of three cell types: (1) elongated glia-like supporting cells; (2) bipolar sensory cells, the VSNs; and (3) scarce basal stem cells. VSNs make up the majority of the cells in the sensory epithelium and are responsible for the detection of chemostimuli. VSNs extend a dendritic process to the luminal surface, where the process enlarges into an apical dendritic knob which contains up to 100 microvilli (Figure 1(d)). In rat, these microvilli are 3–6 μm long and approximately 0.1 μm in diameter. The primary chemotransduction events are likely to take place within the microvilli. The axonal process of each VSN emerges from the basal aspect of the soma and exits the epithelium through the basal lamina. The axons coalesce to form the vomeronasal nerves that course along the septum (Figure 1(a)), pass through the cribriform plate, and end in the glomeruli of the AOB, where they synapse onto second-order mitral cells.

An intriguing feature of the mouse VNO is its segregation into two major layers or zones (Figure 1(e)), containing subpopulations of VSNs that display different immunochemical markers and unique patterns of gene expression. For example, the apical, or superficial, zone of the epithelium contains VSNs that project to the anterior aspect of the AOB and express the G-protein Gαi2 and a member of the V1R family of chemoreceptors. The basal or deep zone contains VSNs that project to the posterior aspect of the AOB and express the G-protein Gαo and members of the V2R family of receptors. The functional significance of this zonal segregation is not yet fully understood, but there is increasing evidence that these VSN subpopulations detect structurally and functionally distinct families of molecular cues.

Stimulus Detection Strategies

Differential screening of complementary DNA libraries constructed from single rat VSNs led to the isolation of the first family of vomeronasal, G-protein-coupled receptor genes, the V1Rs (Figure 1(f)). These have a similar structure but relatively low homology compared with the olfactory receptor superfamily expressed in MOE and are weakly related to T2r taste receptors. In the mouse genome, 308 V1R sequences have been identified, of which 191 seem to have uninterrupted

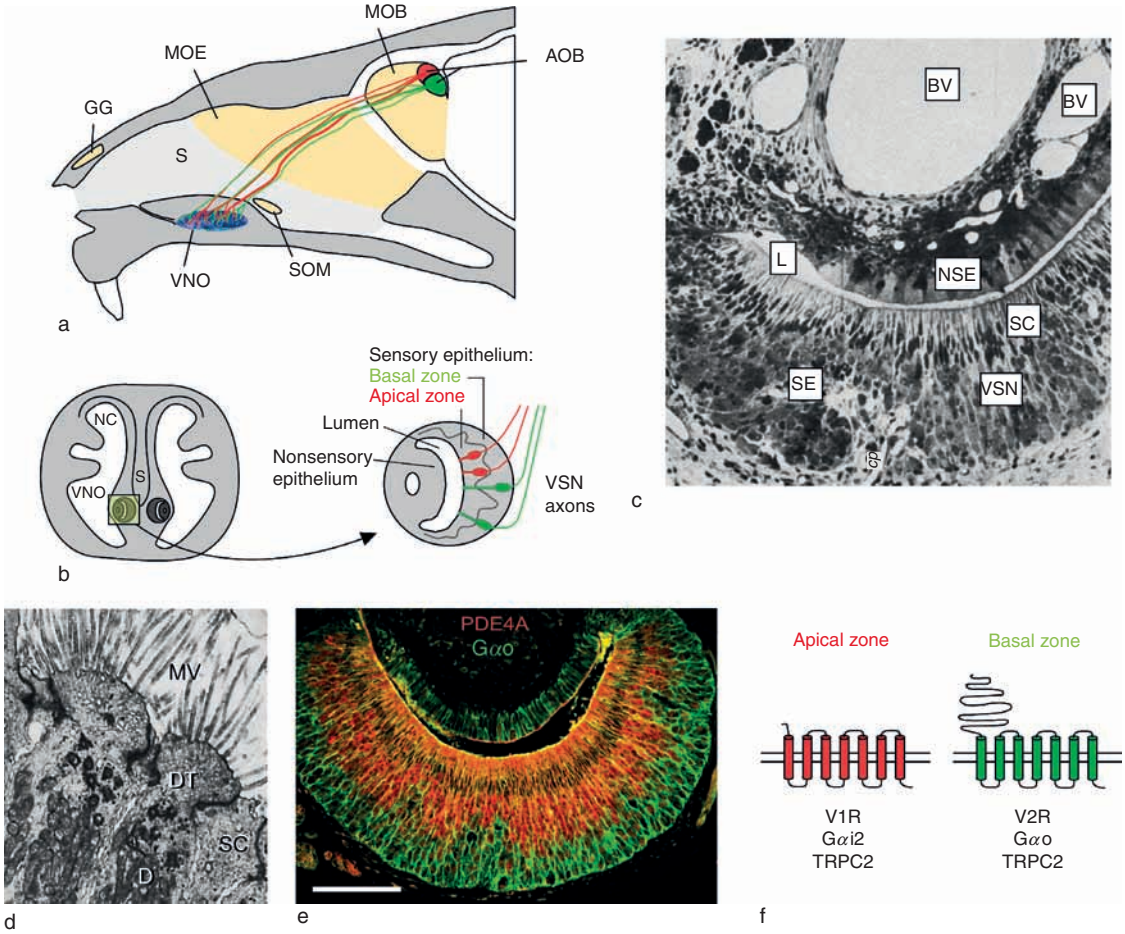

Figure 1 (a) Midsagittal view of the mouse nose and forebrain. The vomeronasal organ (VNO) is located at the base of the nasal septum (S). Axons of vomeronasal sensory neurons (VSNs) project to the anterior (red) or posterior (green) accessory olfactory bulb (AOB). MOE, main olfactory epithelium; SOM, septal organ of Masera; GG, Grueneberg ganglion; MOB, main olfactory bulb. (b) Coronal sectioning through the nasal cavity (NC) reveals the organization of the VNO sensory epithelium: an apical zone (red neurons) that can be distinguished from the basal zone (green neurons) projecting to the rostral (red) and caudal (green) AOB, respectively (see (a)). (c) The cytoarchitecture of the VNO includes a sensory epithelium (SE) containing VSNs and supporting cells (SC), a lumen (L), and the nonsensory epithelium (NSE), together with an extensive network of blood vessels (BV). (d) Electron micrograph of a transverse section through mouse VSNs showing their dendritic process (D), which enlarges into a dendritic tip or knob (DT). From the knob emanate various microvilli (MV). (e) Visualization of the two principal layers of the SE by means of double-label immunohistochemistry with antibodies to phosphodiesterase type 4A (red) and the G-protein Gαo (green). (f) VSNs of the two layers express a unique set of transduction-related molecules: apical VSNs express the G-protein Gαi2 and members of the V1R family of vomeronasal receptors, and basal VSNs express Gαo and members of the V2R receptor family. Both neuron types express TRPC2. Magnification ×90 (c), ×15 000 (d). Scale bar = 100 μm (e). (b, f) From Zufall F, Ukhanov K, Lucas P, Liman E, and Leinders-Zufall T (2005) Neurobiology of TRPC2: From gene to behavior. *Pflügers Archiv: European Journal of Physiology* 451: 61–71, with kind permission of Springer Science and Business Media. (c) Reproduced from Vaccarezza OL, Sepich LN, and Tramezzani JH (1981) The vomeronasal organ of the rat. *Journal of Anatomy* 132: 167–185, with permission from Blackwell Publishing. (d) Reprinted from Ciges M, Labella T, Gayoso M, and Sanchez G (1977) Ultrastructure of the organ of Jacobson and comparative study with olfactory mucosa. *Acta Otolaryngology* 83: 47–58, with permission from Taylor & Francis. (e) Reprinted with permission from Leinders-Zufall T, Brennan P, Wimayer P, et al. (2004) MHC class I peptides as chemosensory signals in the vomeronasal organ. *Science* 306: 1033–1037.

full-length open reading frames. The V1Rs are divided into 12 subfamilies that are considerably divergent from each other. The intrafamily identity is in the range of only 15–40%, probably reflecting a corresponding diversity of ligands. Each individual V1R receptor gene is expressed by only a small subset of apically located VSNs. Evidence for a functional role of V1Rs comes from the finding that mice with a targeted deletion in a cluster of V1R genes fail to detect several small organic pheromones present in urine. Furthermore, green fluorescent protein-tagged VSNs expressing the V1Rb2 receptor respond to 2-heptanone, but this response is absent when V1Rb2 is deleted. High resolution Ca^{2+} mapping studies have revealed an exquisite

sensitivity and specificity of VSNs located in the apical V1R-expressing zone, suggesting that each of these cells detects only one or very few ligands.

A second large family of G-protein-coupled receptors expressed in the VNO is encoded by the V2R genes. With their large extracellular N-terminal domain, V2Rs resemble the calcium-sensing and metabotropic glutamate receptors rather than V1Rs and olfactory receptors (Figure 1(f)). In the mouse, the V2R gene repertoire consists of approximately 280 members, of which approximately 120 exhibit an intact open reading frame. No conclusive evidence is yet available for a functional role of V2Rs. Although each V2R receptor is generally expressed by a small subset of VSNs, one group of V2R receptors, V2R2, is co-expressed broadly across basal VSNs. It remains to be seen whether this results in the formation of functional homo- or heterodimeric receptor proteins. It is interesting that VSNs in the basal layer express yet another multigene family: nonclassical class Ib major histocompatibility complex (MHC) molecules. These seem to form multimolecular complexes with V2Rs and have been proposed to function as escort molecules in the transport of V2Rs to the cell surface. The tuning properties of V2R-positive VSNs are just beginning to emerge. As will be discussed in the next section, a subset of these cells detects MHC peptide ligands at near picomolar concentrations. The fact that these cells can detect multiple peptides exhibiting distinct chemical structures argues against a one cell–one receptor–one ligand coding strategy, but more work will be required to determine the molecular basis underlying their ligand selectivity.

Diversity of Chemosignals Detected by the VNO

The AOS has long been known to play important roles in recognizing species-specific chemosignals (pheromones) that convey information about gender, social status, level of aggression, and reproductive state in conspecifics. However, it is now clear that VSNs detect a much wider range of chemical stimuli, including some molecules that cannot be categorized as pheromones. Furthermore, the VNO is not the only sense organ in the mammalian nose in which pheromones are detected; the main and AOSs detect, at least in part, overlapping sets of sensory stimuli, and both systems play important roles in the regulation of social behaviors and the sensing of pheromonal cues. Given that a functional AOS does not seem to exist in adult humans, the MOS is the most likely pathway for conveying any pheromonal effects in human chemical communication.

The AOS responds to complex natural stimuli such as urine, soiled bedding, vaginal secretion, and seminal fluid. These contain a wide range of molecular cues, from small organic volatiles to nonvolatile peptides and proteins. Recent advances have shown that the mammalian VNO detects all these classes of chemosignals. For example, among the best characterized vomeronasal stimuli are volatile pheromonal compounds such as 2,5-dimethylpyrazine, 2-*sec*-butyl-4,5-dihydrothiazole, 2,3-dehydro-*exo*-brevicomin, E, E-α-farnesene and E-β-farnesene, 2-heptanone, 6-hydroxy-6-methyl-3-heptanone, *n*-pentyl acetate, and isobutylamine.

More-recent evidence has shown that VSNs recognize also some peptides at extremely low concentrations. For instance, mouse VSNs of the basal zone detect a vast range of immune system molecules: antigenic peptide ligands of MHC molecules. MHC/peptide complexes are present at the cell surface and carry information about the genetic makeup of cells. They are shed from the cell surface, and their fragments appear in bodily fluids such as serum, saliva, sweat, and urine. These truncated MHC molecules are thought to have a reduced affinity to their peptide ligands and thus are likely release them into the extracellular space. The finding that a subset of VSNs detects these peptides in an MHC-like fashion suggests that these cells are involved in the sensory evaluation of the MHC genotype of other individuals. Mouse VSNs of the basal layer can also be activated by a male-specific 7 kDa peptide produced by the extraorbital lacrimal gland. This peptide, known as exocrine gland-secreting peptide 1 (ESP1), is encoded by a member of at least 23 related genes in mice. Although the behavioral role of ESP1 is not yet clear, these results indicate that social communication in mammals involves the detection of extended families of peptidergic chemosignals by the VNO.

Other work indicates that VSNs can also detect larger proteins. In snakes, for example, a 20 kDa chemoattractant purified from earthworm shock secretions has been shown to induce excitatory responses in VNO and AOB. Work in rodents supports involvement of proteins in VNO sensing. Mouse urine contains a class of highly polymorphic proteins known as major urinary proteins that are involved in scent marking. These are 18–20 kDa members of the lipocalin family of ligand-binding proteins that bind volatile pheromonal ligands. It has been suggested that major urinary proteins, devoid of their volatile ligands, function as VSN chemosignals in their own right.

Finally, VSNs can be stimulated by odorants without known pheromonal functions, such as methyl anisole, patchone, indole, helional, butyrophenone,

and fenchone. Functional imaging studies using a mixture of these compounds demonstrate that they stimulate the mouse AOB *in vivo*. These odorants are suggested to be released from plants or other animal species and could signal the presence of a predator or indicate the suitability of a particular site for feeding or nesting.

VNO Signal Transduction: Transient Receptor Potential Channels and Phosphoinositide Signaling

How do VSNs transduce these molecular cues into an electrical membrane signal? Work from several laboratories has demonstrated that VSNs, at rest, maintain a relatively low spiking activity and that chemostimulation produces a transient increase in the rate of action potential firing. This excitation is caused by the generation of a depolarizing receptor potential sufficient in size to activate voltage-gated Na^+ channels. The ionic properties of this receptor potential are not yet fully understood, but it seems to depend critically on the generation of a transient inward current in the distal dendrites and microvilli of VSNs which is associated with an increase in membrane conductance. Recent work has pointed to a phospholipase C (PLC)-mediated second messenger system involving at least one member of the transient receptor potential (TRP) family of cation channels, TRPC2, in the generation of this receptor current (Figure 2). Because TRPC2 encodes a protein that is expressed in all VSNs and is specifically localized to the sensory microvilli of these cells (Figure 2(a)), it was proposed to participate in VSN signal transduction. Indeed, mice with a targeted mutation in the TRPC2 gene exhibit strongly reduced electrophysiological responses to some pheromonal ligands, and these mice also show dramatic defects in the regulation of sexual and social behaviors. However, the recent finding that V2R-type peptide-responding VSNs do not require a functional TRPC2 channel indicates that VNO signal transduction might involve additional, not yet identified, ion channels.

What kind of ion channel is formed by TRPC2? Although this question is still being investigated, evidence indicates that TRPC2 is required for a Ca^{2+}-permeable cation channel that is gated by the lipid messenger diacylglycerol (DAG) or one of its endogenous analogs (Figure 2(b)). DAG is a direct product of the activity of PLC, which hydrolyzes

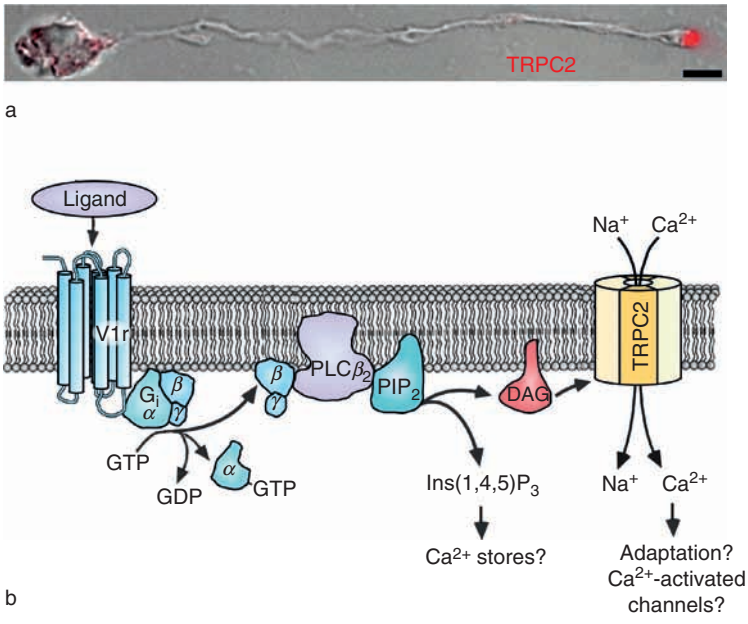

Figure 2 (a) Expression of a member of the transient receptor potential (TRP) family of cation channels, TRPC2, in the vomeronasal sensory neurons (VSNs). TRPC2 immunoreactivity (red) is clearly seen in the tuft of microvilli at the distal end of the dendrite. (b) Schematic representation of pathways that might regulate the activity of the TRPC2 channel in the dendritic tip of a mammalian VSN. The model refers to VSNs of the apical zone of the vomeronasal organ. Whether this applies also to VSNs of the basal zone remains to be investigated. α, β, and γ, G-protein subunits; DAG, diacylglycerol; GDP, guanosine diphosphate; GTP, guanosine triphosphate; Ins(1,4,5)P_3, inositol 1,4,5-trisphosphate; PIP, phosphatidylinositol-4,5-bisphosphate; PLC, phospholipase C. Scale bar = 5 μm (a). From Liman ER, Corey DP, and Dulac C (1999) TRP2: A candidate transduction channel for mammalian pheromone sensory signaling. *Proceedings of the National Academy of Sciences of the United States of America* 96: 5791–5796, copyright 1999, National Academy of Sciences, USA. (b) Reprinted from Zufall F, Ukhanov K, Lucas P, Liman E, and Leinders-Zufall T (2005) Neurobiology of TRPC2: From gene to behavior. *Pflügers Archiv: European Journal of Physiology* 451: 61–71, with kind permission of Springer Science and Business Media.

phosphatidylinositol-4,5-bisphosphate into two second messengers, namely inositol 1,4,5-trisphosphate (Ins(1,4,5)P_3) and DAG. Excised patch clamp recordings identified a 42-pS DAG-activated cation channel present in high density in the plasma membrane of dendritic knobs of VSNs, and the activation of this channel is strongly defective in TRPC2-deficient VSNs. Furthermore, the functional properties of this channel are consistent with those of the pheromone-evoked conductance. Independent evidence that endogenously produced DAG may be a critical second messenger in the sensory response of VSNs comes from pharmacological blockade of DAG kinase, an enzyme involved in termination of DAG signaling. Thus a direct link between stimulus-induced PLC activity, gating of the TRPC2 channel by DAG, and the sensory current has been established, but several other second messengers, such as Ca^{2+}, Ins(1,4,5)P_3, and arachidonic acid might also be involved in VSN signal transduction.

Functional Organization of the AOB

VSN axons terminate exclusively in the AOB. With respect to the morphological organization, the AOB appears quite similar to the MOB. It can be divided into five layers: (1) the accessory olfactory nerve layer, which contains VSN axons that converge onto glomeruli in (2) the accessory glomerular layer; (3) the accessory external plexiform layer; (4) the accessory mitral cell layer; and (5) the accessory granule cell layer. VSN axons terminate in the accessory glomerular layer, forming glutamatergic synapses with the apical dendrites of mitral cells. Mitral cells form reciprocal dendodendritic synapses with granule cells, a class of γ-aminobutyricacidergic interneurons. Plastic changes at these synapses are hypothesized to play an essential role in the formation of social memories (see the next section and Figure 3). AOB mitral cells have direct projections to the medial and posteromedial cortical amygdaloid nucleus, the bed nucleus of the stria terminalis, and the bed nucleus of the accessory olfactory tract. Together these structures can be called the accessory olfactory cortex.

AOB mitral cell responses in the awake mouse consist of a gradual elevation in firing rate during 2–5 s following contact with a stimulus, such as facial or anogenital investigation. The slowly evolving elevation of firing rate is in sharp contrast to other sensory systems, such as vision, auditory, or somatosensory, in which the precise timing of stimulus-triggered responses is highly conserved. However, the slow evolution of activity in the AOB mitral cells is consistent with the indirect nature of stimulus access. The relatively slow characteristics of sensory responses could allow for significant temporal evolution of local inhibitory feedback and cortical feedback during the detection event.

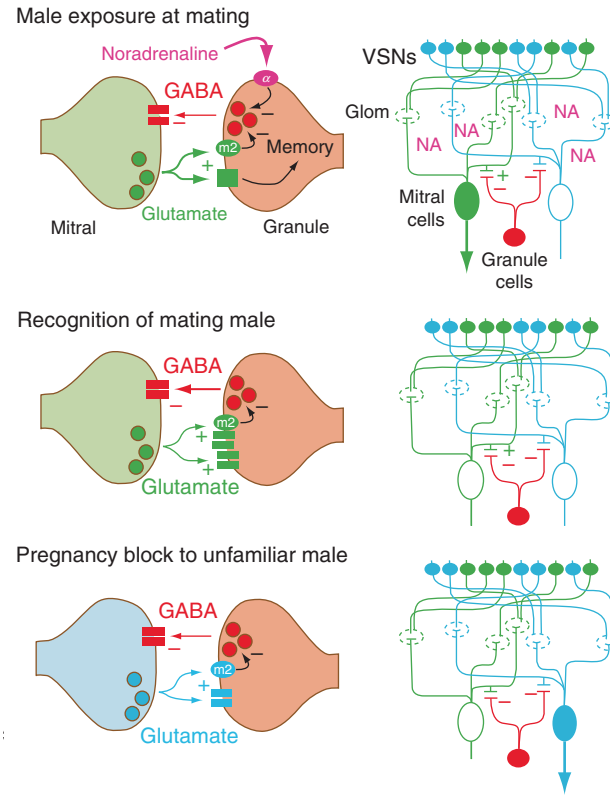

Figure 3 Hypothetical model for the neural basis of mate recognition in the pregnancy block paradigm. (a) During mating, the disinhibitory effects of increased noradrenaline (NA) and activation of mGluR2 (m2) glutamate receptors enhance stimulation of ionotropic glutamate receptors, resulting in potentiation of those glutamatergic synapses on granule cells that receive input from the mating male's chemosignals (green). (b) Exposure to the mating male's urinary chemosignals during the days following mating activates the potentiated synapses and calls up increased levels of feedback inhibition from the granule cells. This increased inhibitory control of the mitral cells responding to the mating male's chemosignals disrupts their transmission and prevents the activation of the neuroendocrine pathway for pregnancy block, resulting in reproductive success. (c) Exposure to the urinary chemosignals from an unfamiliar male during the days following mating activates a separate population of mitral cells without potentiated synapses. The activation of mGluR2 suppresses inhibition from the associated granule cells, and the signal is transmitted centrally, resulting in pregnancy block. GABA, γ-aminobutyric acid; Glom, olfactory glomerulus; VSNs, vomeronasal sensory neurons. Reproduced from Brennan PA (2004) The nose knows who's who: Chemosensory individuality and mate recognition in mice. *Hormones and Behavior* 46: 231–240, with permission from Elsevier.

The logic of input–output connectivity in the AOB is an area of intense investigation. AOB glomeruli are spherical structures comprised of dense entanglements of mainly axon fibers from VSNs and dendrites from mitral cells. AOB glomeruli are 10–50 μm in

diameter, poorly separated by interneurons, and distributed in a loosely organized layer that is many glomeruli thick. The division of VSNs into two zones is maintained at the level of the AOB, separating it into an anterior and a posterior domain (**Figure 1(a)**). In contrast to the glomerular organization in the mammalian MOB, VSNs expressing the same receptor gene project to multiple glomeruli in either the anterior or the posterior domain of the AOB (6–20 glomeruli per gene have been reported). The distribution of these multiple glomeruli within each domain exhibits a high degree of variability compared with the convergence and position of glomeruli in the MOB, but there is a general similarity across animals. In addition, AOB mitral cells have a structure strikingly different from that of MOB mitral cells. AOB mitral cells possess multiple apical dendrites that each divide into branches ending in different glomeruli. Because at least some of the mitral cell multiple processes innervate glomeruli that receive axons from the same population of VSNs, the divergent pattern of VSN projections into the AOB can be rendered convergent. A recent study using transgenic mouse lines in which projections from multiple populations of VSNs, each expressing a distinct V1R receptor, were differentially labeled revealed that AOB inputs from neurons expressing closely related V1Rs intermingle within shared, spatially conserved domains of the AOB. Individual mitral cells were shown to extend dendrites to glomeruli associated with different, but closely related, V1Rs, suggesting that considerable integration of information might already occur at the level of the AOB.

VNO Sensing and the Modification of Behaviors

A vast literature has established an involvement of the VNO and AOS in a variety of behavioral effects, but very few studies have linked the sensing of defined molecular cues by VSNs to the modification of a specific behavioral response. Initially, researchers used surgical lesions of the VNO in hamsters, voles, mice, rats, farm animals, and reptiles to show that chemosensory signaling via the VNO influences a number of social behaviors, including male aggression to intruder males; maternal aggression in lactating females; male mating behavior; male ultrasonic vocalizations; and the timing of puberty, cyclicity, and ovulation in females. But significant differences occur between species, and the VNO is also involved in the detection of heterospecific stimuli, such as those involved in prey recognition in snakes. More recently, researchers used genetic lesions (i.e., mice with targeted mutations in specific signal transduction molecules such as TRPC2) to investigate the role of the VNO in chemical communication. With respect to male–male aggression and maternal aggression in lactating females, an excellent correspondence exists between the behavioral phenotypes caused by TRPC2 deletion or surgical VNO ablation. However, profound discrepancies occur in the context of male sexual behavior, which may be explained in part by the recent finding of TRPC2-independent responses of peptide-sensing VSNs.

The VNO is also critical for one of the best known examples of olfactory imprinting in adult vertebrates, the selective pregnancy block (or Bruce effect) in the mouse, which depends on the formation and maintenance of a long-lasting odor recognition memory by the AOS. The Bruce effect occurs when a recently mated female mouse aborts if exposed to the urine odor of a male genetically different from the stud male. This unfamiliar male odor results in pregnancy block and return to estrus in the recently mated female. Changes at the dendrodendritic synapse between granule cells and mitral cells in AOB might be crucial for odor memory of the mate in the female (**Figure 3**). The Bruce effect can be abolished if noradrenergic centrifugal input to the female AOB is blocked immediately after mating, presumably before olfactory memories of the mate are formed. Thus, noradrenaline appears to be important in strengthening the memory of the odor of the mating partner. Noradrenergic innervation of the AOB derives from locus coeruleus projections. Granule cells in the AOB are a major target of the noradrenaline input. Thus, noradrenaline's effect on granule cells leading to enhanced inhibition of a subset of mitral cells for several hours following mating may facilitate this selective odor memory.

The MHC peptide ligands are the first identified vomeronasal stimuli that can mediate the Bruce effect. To show that MHC peptides provide a crucial signal among the plethora of other cues present in male urine, peptides characteristic of a disparate MHC molecule were synthesized and added to the urine of a familiar male. Compared with the addition of cognate peptides, which have no effect, disparate peptides cause pregnancy block by otherwise familiar urine, an effect that is absent after surgical lesioning of the VNO. Thus, MHC peptide ligands are recognized as natural odors and convey genotypic information in the context of a distinct social behavior.

Conclusion

Recent developments have shown that vertebrate pheromones come in a wide variety of chemical forms, from small organic volatiles to peptides and

proteins. The AOS is heavily involved in the detection of these pheromones, but the MOE has been found to overlap significantly with the VNO in the social cues it detects. Studies using gene-targeted mice have shown an unexpected diversity of mammalian chemosensory systems and their underlying cellular and molecular mechanisms. Our understanding of mammalian and especially human pheromonal communication is still incomplete, but new insights into ligand–receptor interactions will enable selective ablation of specific receptor systems, followed by investigation of the resulting cellular and behavioral phenotypes. It is hoped that these approaches will ultimately reveal many of the molecular and cellular mechanisms underlying mammalian social behavior.

See also: Olfactory Receptor Genes: Human Loss During Evolution; Olfactory Receptors; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Olfactory System Theory; Vomeronasal System Evolution.

Further Reading

Boehm U, Zou Z, and Buck LB (2005) Feedback loops link odor and pheromone signaling with reproduction. *Cell* 123: 683–695.

Boehm T and Zufall F (2006) MHC peptides and the sensory evaluation of genotype. *Trends in Neurosciences* 29: 100–107.

Brenan PA (2004) The nose knows who's who: Chemosensory individuality and mate recognition in mice. *Hormones and Behavior* 231: 231–240.

Brennan PA and Zufall F (2006) Pheromonal communication in vertebrates. *Nature* 444: 308–315.

Ciges M, Labella T, Gayoso M, and Sanchez G (1977) Ultrastructure of the organ of Jacobson and comparative study with olfactory mucosa. *Acta Otolaryngology* 83:47–58.

Del Punta K, Leinders-Zufall T, Rodriguez I, et al. (2002) Deficient pheromone responses in mice lacking a cluster of vomeronasal receptor genes. *Nature* 419: 70–74.

Dulac C and Torello AT (2003) Molecular detection of pheromone signals in mammals: From genes to behaviour. *Nature Reviews Neuroscience* 4: 551–562.

Halpern M and Martinez-Marcos A (2003) Structure and function of the vomeronasal system: An update. *Progress in Neurobiology* 70: 245–318.

Hurst JL and Beynon RJ (2004) Scent wars: The chemobiology of competitive signalling in mice. *Bioassays* 26: 1288–1298.

Johnston RE (2000) Chemical communication and pheromones: The types of chemical signals and the role of the vomeronasal system. In: Finger TE, Silver WL, and Restrepo D (eds.) *The Neurobiology of Taste and Smell,* 2nd edn., pp. 101–127. New York: Wiley-Liss.

Leinders-Zufall T, Brennan P, Wimayer P, et al. (2004) MHC class I peptides as chemosensory signals in the vomeronasal organ. *Science* 306: 1033–1037.

Leinders-Zufall T, Lane AP, Puche AC, et al. (2000) Ultrasensitive pheromone detection by mammalian vomeronasal neurons. *Nature* 405: 702–706.

Liman ER, Corey DP, and Dulac C (1999) TRP2: A candidate transduction channel for mammalian pheromone sensory signaling. *Proceedings of the National Academy of Sciences of the United States of America* 96: 5791–5796.

Luo M and Katz LC (2004) Encoding pheromonal signals in the mammalian vomeronasal system. *Current Opinion in Neurobiology* 14: 428–434.

Meredith M (1998) Vomeronasal, olfactory, hormonal convergence in the brain. *Annals of the New York Academy of Sciences* 855: 349–361.

Mombaerts P (2004) Genes and ligands for odorant, vomeronasal and taste receptors. *Nature Reviews Neuroscience* 5: 263–278.

Vaccarezza OL, Sepich LN, and Tramezzani JH (1981) The vomeronasal organ of the rat. *Journal of Anatomy* 132: 167–185.

Wagner S, Gresser AL, Torello AT, and Dulac C (2006) A multireceptor genetic approach uncovers an ordered integration of VNO sensory inputs in the accessory olfactory bulb. *Neuron* 50: 697–709.

Wyatt TD (2003) *Pheromones and Animal Behaviour.* Cambridge, UK: Cambridge University Press.

Zufall F, Ukhanov K, Lucas P, Liman E, and Leinders. Zufall T (2005) Neurobiology of TRPC2: From gene to behavior. *Pfliigers Archiv: European Journal of Physiology* 451: 61–71.

Relevant Websites

http://www.achems.org – Association for Chemoreception Sciences.

http://bip.weizmann.ac.il – Bioinformatics & Biological Computing; see Human Olfactory Receptor Data Exploratorium.

http://www.leffingwell.com – Leffingwell & Associates: Services and Software for the Perfume, Flavor, Food and Beverage Industries.

http://md1.csa.com – ProQuest; see Chemoreception Web.

http://senselab.med.yale.edu – SenseLab; see The Olfactory Receptor Database.

http://www.syfpeithi.de – SYFPEITHI: A Database of MHC Ligands and Peptide Motifs.

Vomeronasal System Evolution

T D Smith, Slippery Rock University, Slippery Rock, PA, USA
K P Bhatnagar, University of Louisville School of Medicine, Louisville, KY, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Tetrapods such as snakes, lizards, rodents, and moles can track prey, forage, or receive social signals by detecting chemical stimuli. Not surprisingly, these animals are endowed with 'dual' olfactory systems – the vomeronasal and the olfactory. The components of the main olfactory system are the olfactory neuroepithelium, the olfactory glands (of Bowman), olfactory fila, and nerve fascicles synapsing within the glomeruli in the olfactory bulb and coursing through the limbic system to reach the olfactory cortex. In the periphery, the vomeronasal system (VNS, or 'accessory olfactory system') comprises bipolar vomeronasal receptor neurons (VRNs) with apical microvilli. Together with supporting and basal cells, VRNs comprise the vomeronasal neuroepithelium. The VRN axons coalesce into often ganglionated vomeronasal nerves that extend to glomeruli in the accessory olfactory bulb (AOB), and after synapsing with the mitral cell dendrites project to the hippocampus–amygdala complex. Their onward course to the olfactory cortex, if there is any, is less well known. The vomeronasal organ (VNO) and ancillary structures are shown in Figures 1 and 2.

In strict terms, the vomeronasal organ is the patch of chemosensory epithelium that may take the form of a sac or duct, and may or may not have nonsensory epithelium integrated with the wall. This definition is advantageous in that it acknowledges homology of the neuroepithelium, which differentiates from a medial patch of the embryonic nasal pit in all tetrapods, among vertebrates that possess the VNO. Other definitions are specific to particular taxa (e.g., veterinary terminology for the mammalian VNO). Such definitions should be avoided since nonhomologous elements may serve the same function in different taxa (notably, the cartilaginous versus bony elements that encapsulate the organ).

History

Erroneously, Frederick Ruysch (1638–1731) is credited with the discovery of the VNO. He reported on a grossly observable aperture and passage, visible along the lateral surface of a child's nasal septum. Ruysch did not cross-sectionally evaluate this mucosal structure and canal, so it is not possible to know whether it was indeed the vestige of the VNO later documented in humans. The VNO was first reported in mammals (but not in humans) by Ludwig Levin Jacobson (1783–1843; Figure 3(a)). Rudolf Albert von Kölliker (1817–1905) provided the first microscopic evidence of the VNO in humans. Numerous reports followed on the VNO of other tetrapods.

Distribution across Vertebrates

An anatomically distinct VNS, as defined in the preceding section, is found in most terrestrial vertebrates, a distribution leading some to consider the VNS as a terrestrial adaptation (but, see later). Broadly speaking, a VNS is found in most extant amphibians, reptiles, and mammals, but is absent in extant archosaurs (birds and crocodilians; Figure 3(b)). However, loss of the VNS has occurred within all classes of vertebrates that possess it.

Morphological Diversity

The Vomeronasal Organ

The nasal chambers of amphibians vary in complexity from simple recesses in urodelans (e.g., salamanders) to the series of chambers in anurans (e.g., frogs). The VNO of anurans is found along the medial extent of the inferior recess (Figures 2(a) and 4(a)–4(c)). The VNO, therefore, is unenclosed, and instead is an elongated furrow or pouch that openly communicates with the inferior recess. Most accessory structures are absent, except for some glands found just dorsal to the VNO. In urodelans, the VNO takes the form of a pouch that openly communicates with a single nasal chamber. The VNO of caecilians is an enclosed, ovoid neuroepithelial sac. Unique structures to this group are bilateral sensory tentacles, which are considered modified eyelids/conjunctiva. These communicate with the VNO medially via a central duct.

Among all vertebrates, snakes appear to possess the proportionally largest VNO and AOB, in which the usual proportions relative to the main olfactory system are reversed. In squamates (snakes and lizards) the VNO takes the form of an inverted sac (Figures 2(b) and 4(d)–4(f)). Although the oral and nasal cavities are broadly joined by elongated choanae, the VNO opening is restricted to the oral cavity since it opens only via a narrow palatal duct, oriented ventrally. Aside from the ventrally positioned palatal duct, the 'floor' of the squamate VNO is distinguished by the presence of ciliated epithelium that lines a convex mass termed the mushroom body (Figure 5(c)). Turtles possess an

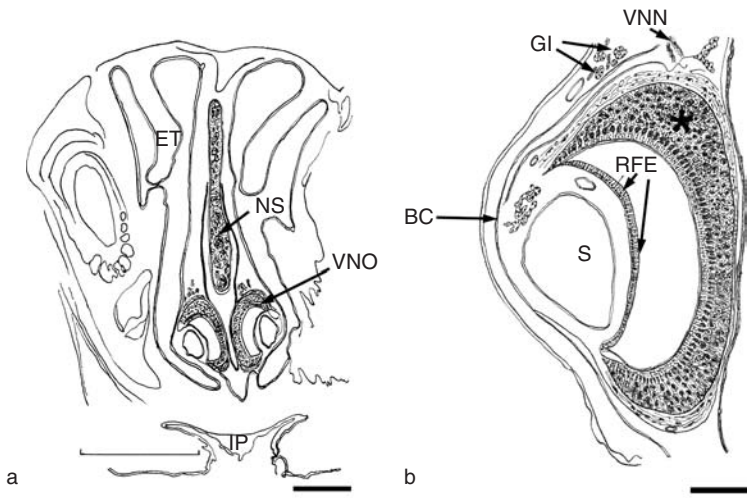

Figure 1 Illustration of (a) the vomeronasal complex and (b) the vomeronasal organ (VNO) in a mouse shown in frontal plane. BC, bony capsule; ET, ethmoturbinal; Gl, glands; IP, incisive papilla; NS, nasal septum; RFE, receptor-free epithelium; S, sinus; VNN, vomeronasal nerve; *, neuroepithelium. (a) Scale bar = 500 μm (a) and 100 μm (b).

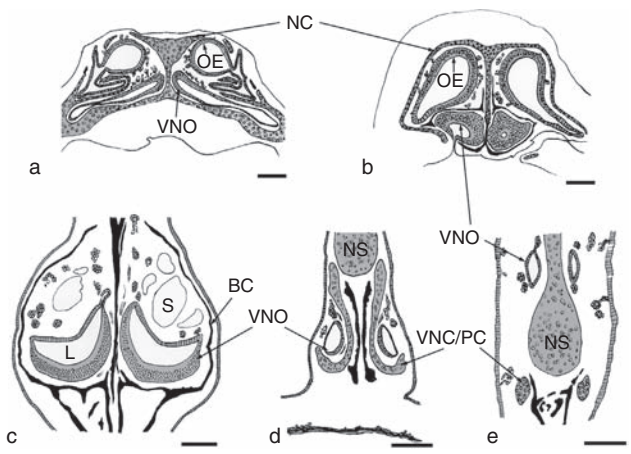

Figure 2 Schematic illustrations of the vomeronasal complex in different vertebrates (frontal plane), highlighting positional variations of the vomeronasal organ (VNO) and its capsule. In amphibians, the VNO is a recess within a large nasal chamber, or series of chambers (a, *Rana* sp.). In most reptiles (b, *Hemiergis* sp.) and mammals (c, *Microtus pennsylvanicus*; d, *Lemur catta*), the VNO is found along the base of the nasal septum (NS) within a capsule of bone (black-filled) or vomeronasal cartilage (VNC). (e) In some species (e.g., *Pan troglodytes*), the VNO may be displaced from the cartilaginous capsule (PC, paraseptal cartilages). BC, bony capsule; L, lumen; NC, cartilaginous nasal capsule; OE, olfactory epithelium. Scale bar = 500 μm (a, b, d), 150 μm (c), and 1 mm (e).

unencapsulated vomeronasal neuroepithelium – a VNO homologue since its axons transmit to the AOB.

In mammals, the VNO most commonly opens anteriorly into the nasopalatine duct (NPD), representing a potential 'dual-access' channel for oral or nasal chemostimuli. Usually the VNO narrows to a stratified squamous duct on its rostral end, which links to the NPD. Caudal to the VNO duct, the elongated mammalian VNO most commonly takes the form of a crescent when viewed in coronal sections (Figures 2(c), 6(a), 6(b), and 6(c)). Ventromedially, the VNO wall is thicker and convex; dorsolaterally, the VNO wall is thinner and concave (Figures 1(b), 6(c), 7(c), and 7(d)). The dorsolateral wall is usually a ciliated, nonsensory epithelium, termed 'receptor-free epithelium' (RFE).

The Accessory Olfactory Bulb

A miniature, generally disorganized, accessory olfactory bulb is an essential feature of the VNS. An extremely close relationship exists between the

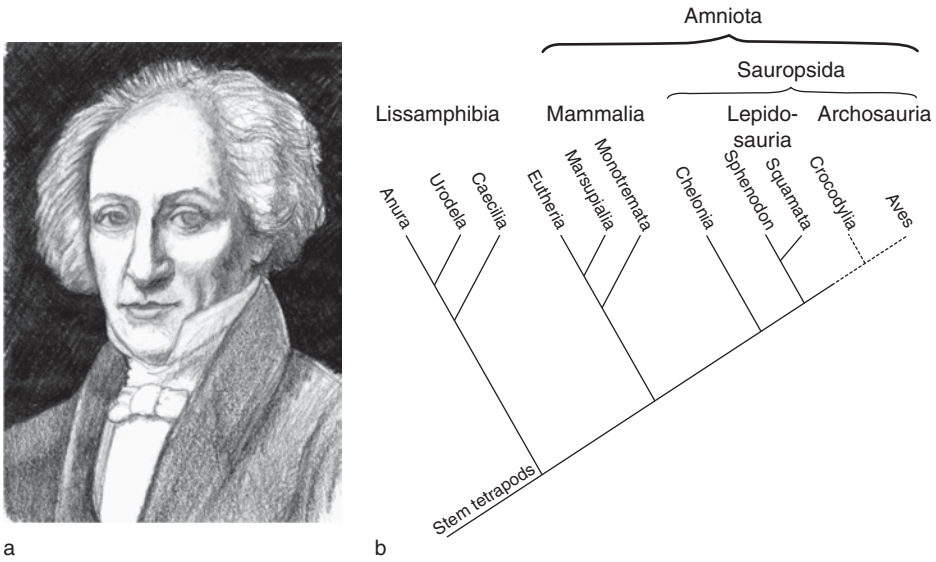

Figure 3 (a) Ludwig Levin Jacobson (1783–1843). (b) Cladogram showing the extant taxa of tetrapods discussed in the text. Dashed line indicates the absence of the VNS. (a) Reprinted from Bhatnagar KP and Smith TD (2006) The vomeronasal organ and its evolutionary loss in catarrhine primates. In: Preuss TM and Kaas JH (eds.) *The Evolution of Primate Nervous Systems*, vol. 5, pp. 141–148. Oxford: Elsevier (b) Adapted from Hillenius WJ and Rehorek SJ (2005) From the eye to the nose: Ancient orbital to vomeronasal communication in tetrapods? In: Mason RT, LeMaster MP, and Muller-Schwarze D (eds.) *Chemical Signals in Vertebrates 10,* pp. 228–241. New York: Springer.

vomeronasal neuroepithelium (VNNE) and the AOB, which receives axons from the former. One is never found without the other. Most commonly, the AOB is situated caudally within the main olfactory bulb, but varies in mediolateral position. The six layers of the main olfactory bulb are also present in the AOB. In mammals, they are very miniscule in size, intermixed, and disorganized. The different parts and cytoarchitectural details of the cellular and fibrous elements have not been reported in most of the species. This is an intense area for future neuroanatomical investigation. Of the six layers, the glomerular layer (II) and mitral and tufted cell layer (IV) are dispersed throughout the AOB. Further projections of the mitral and tufted cells are less well understood.

Glandular Apparatus

Olfactory epithelium appears to have a phylogenetically ancient glandular apparatus that moistens its surface. As these Bowman's glands are nested in the olfactory lamina propria itself, the secretions appear to be sufficient for binding odorants for chemoreception. The VNO of extant tetrapods receives secretions from at least one of two glandular sources: secretions from the ventral nasal septal glands and the Harderian gland. Each of these glands has been implicated as a provider of odorant-binding secretions (herein, the term odorant is used to denote chemical stimuli that the olfactory epithelium and/or VNO can detect).

With few exceptions, Harderian glands are ubiquitously present in tetrapods. It is a medial orbital gland which varies in product and sends its secretions to the nasal cavities via the nasolacrimal duct (NLD). In most amphibians, Harderian secretions are indirectly accessible by the VNO. This has been clearly demonstrated by experimental work that shows NLD-borne substances are transported to the cavum medium of anurans and diffuse to the VNO (the latter is a diverticulum of the cavum inferior; **Figure 4**). Urodelans do not require stimulus transfer via a duct, but appear to have a nasolabial groove in which glandular secretions may be borne, and chemical stimuli may be acquired during 'nose-tapping' behavior. In squamates, the NLD connects to the palatal duct by which the VNO opens orally. It is hypothesized that Harderian secretions facilitate odorant binding in the VNO of squamates. This possibility is bolstered by recent evidence that Harderian secretions have pheromone-binding capacity in snakes.

In mammals, the functional association of Harderian secretions with VNO function is far less clear; in some mammals other (e.g., pheromonal) functions for Harderian secretions are known. A role has been proposed for Harderian secretions in mammalian VNO function. The evidence is indirect and deserves more scrutiny. The putative sequence is elaborate but ultimately testable. In brief, since the Harderian secretions reach the nasal vestibule of most mammals, they may contribute to a blend of fluids that are accessible at

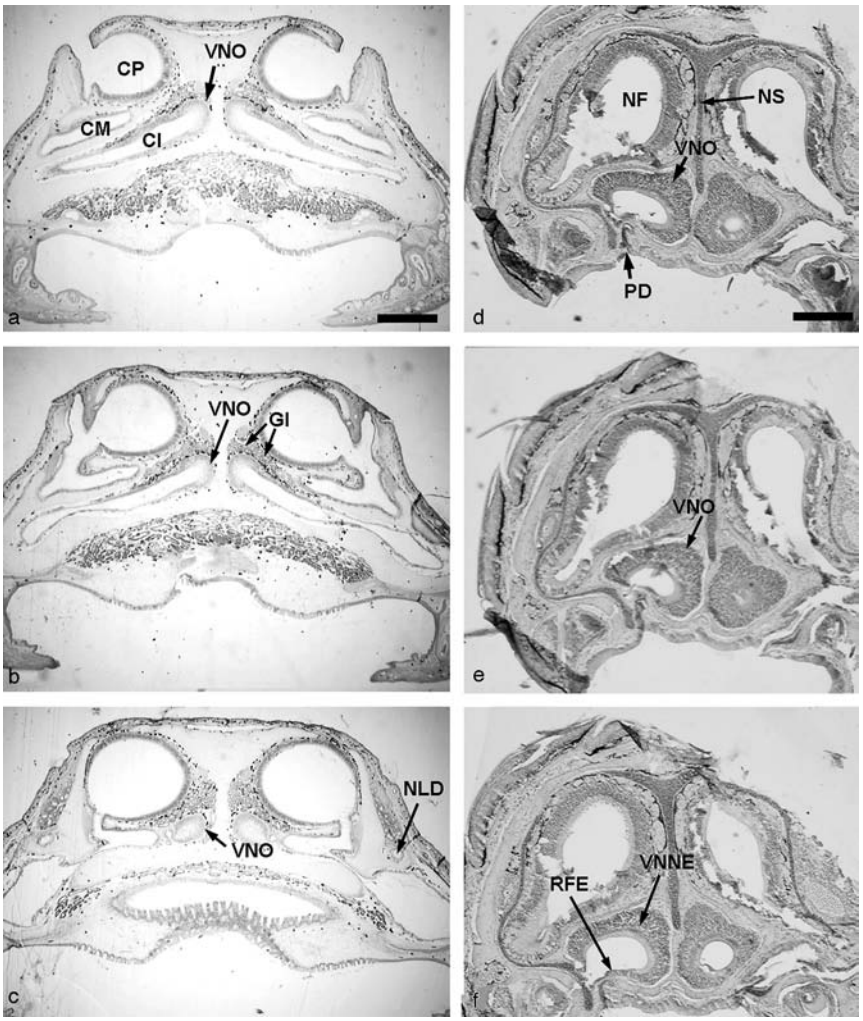

Figure 4 Rostrocaudal series of nasal fossae in an amphibian (*Rana*, a–c) and reptile (*Hemiergis*, d–f). In anurans such as *Rana*, the nasal fossae are divided into a series of chambers on each side. The vomeronasal organ (VNO) is an inferomedial recess of the inferior chamber (cavum inferior, CI). In reptiles that possess a VNO, the organ is found bilaterally at the base of the nasal fossae (NF). Mucosal glands (GI) adjacent to the VNO are absent in squamate reptiles. CP, cavum principal (main chamber); CM, cavum medium; NLD, nasolacrimal duct; NS, nasal septum; PD, palatal duct; RFE, receptor-free epithelium; VNNE, vomeronasal neuroepithelium. Scale bar = 500 μm (a, d). (The authors thank SJ ReKorek for access to specimens shown in this figure.)

either end of the NPD; behaviors such as flehmen, or 'nuzzling,' may aid in stimulus delivery.

All tetrapods except lepidosaurs have a cluster of glands adjacent to the VNO. In mammals, this series of glands enters the VNO at the RFE–VNNE junctions and posteriorly, where the VNO lumen is continuous with a large gland duct that ramifies into multiple gland lobes. The nomenclature often used for these glands, 'vomeronasal glands,' may obscure certain issues of homology, however. For instance, it is not clear that the glandular complex associated with the amphibian VNO is homologous with that of mammals. In mammals, a subset of nasal septal glands empties into the VNO, and such glands appear to be retained in mammalian taxa that lack a VNO. Until the homology of these glands is firmly understood, the term 'vomeronasal gland' should not be used.

The Capsule

In most reptiles and mammals that possess a VNO, a cartilaginous or osseous capsule surrounds the neuroepithelial tube or sac (Figures 6 and 7). Reptiles and mammals possess bilateral paraseptal cartilages that run parallel along the base of the nasal septum. In most mammals, a portion of these cartilages (the 'vomeronasal cartilage') surrounds each VNO (Figures 6(c), 7(c), and 7(d)). Some mammals (e.g., rodents) possess a bony capsule around the organ (Figures 1, 2(c), and 6(a)). In rodents, it is unlikely that the bony capsule represents ossified vomeronasal cartilages.

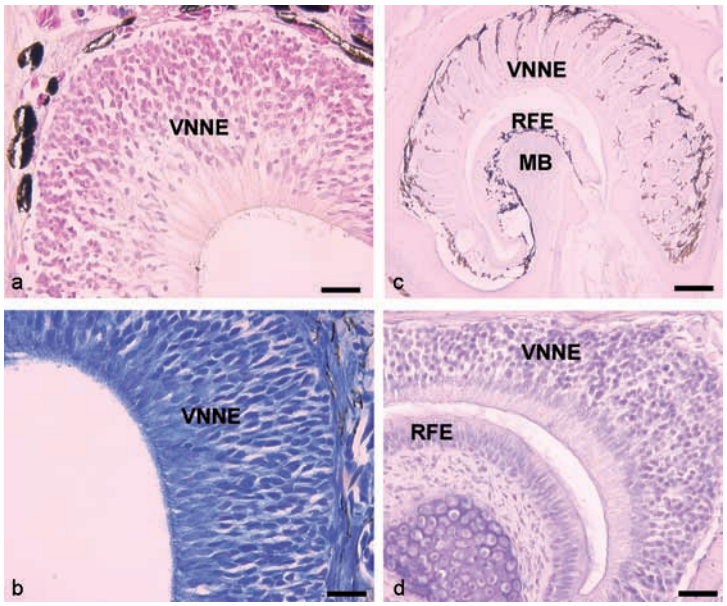

Figure 5 High-magnification views of the vomeronasal organ in amphibians (a, b) and reptiles show a vomeronasal neuroepithelium (VNNE) comprising numerous layers of nuclei. (a) *Rana*; (b) *Dermophis*; (c) *Thamnophis*; (d) *Hemiergis*. MB, mushroom body; RFE, receptor-free epithelium. Scale bar = 50 μm (a–d).

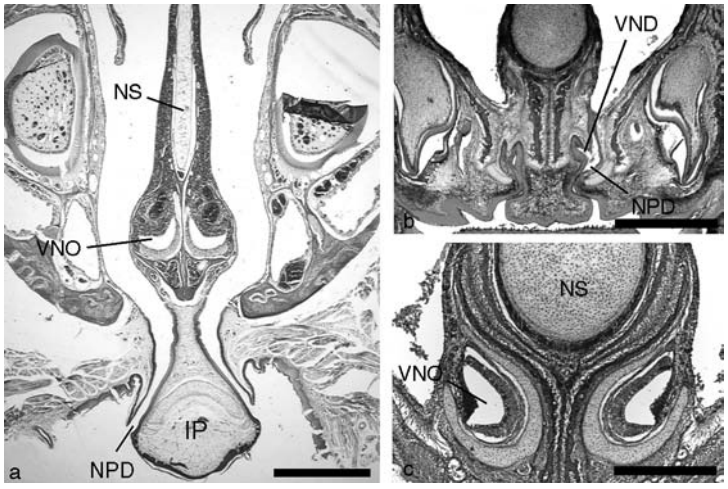

Figure 6 The location of the nasopalatine duct (NPD) is shown in a rodent (a, *Microtus*) and in an insectivore (b, c, *Tenrec*). In rodents, the NPD is located posterior to the opening of the vomeronasal organ (VNO). In insectivores and most mammalian taxonomic groups, the VNO opens into the NPD via a short vomeronasal duct (VND). (c) The VNO posterior to the level in (b). NS, nasal septum; IP, incisive papilla. Scale bar = 1 mm (a–c).

The cartilage/bone capsule may be related to stimulus delivery in all tetrapods. In mammals, a large venous sinus or multiple sinuses are found adjacent to the VNO, within the capsule (Figure 1(b)). Therefore, autonomic nerves may elicit changes in blood volume adjacent to the VNO, either compressing the VNO lumen (expelling contents) or allowing the lumen to expand (drawing fluids in). As this occurs within the capsule, the effect is to expel contents or (creating a vacuum) draw fluids into the VNO lumen.

Venous sinuses are not found adjacent to the VNO of squamates (Figure 4(e)). In snakes, the suggested mechanism is via pressure of the tongue on the palate, creating a suction force to draw fluids into the lumen.

The Receptor-Free Epithelium

Two types of epithelia may comprise a VNO–neuroepithelium, and receptor-free or nonsensory epithelium. The multiple ducts of the septal ('vomeronasal') glands enter VNO at variable positions (dorsoventrally and

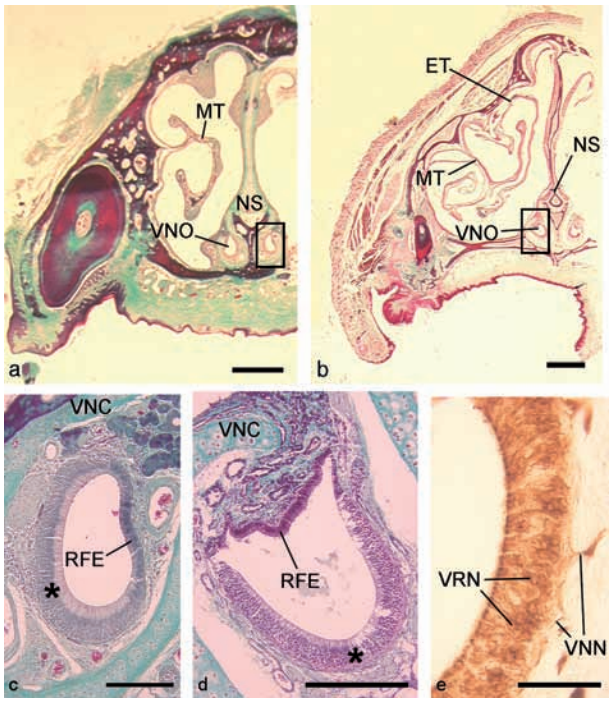

Figure 7 The vomeronasal organ (VNO) of strepsirrhine primates, such as dwarf lemurs (a, *Cheirogaleus*; inset is enlarged in (c)) and bushbabies (b, *Galago*; inset is enlarged in (d)), resembles that of most mammals in possessing two distinct epithelia. The neuroepithelium (*; c, d) comprises bipolar vomeronasal receptor neurons (VRN), which are demonstrated with a *Ulex europaeus*-1 procedure (e, showing the neuroepithelium of *Eulemur*). The axons of VRNs coalesce into vomeronasal nerves (VNN) in the lamina propria. ET, ethmoturbinal I; MT, maxilloturbinal; NS, nasal septum; RFE, receptor-free epithelium; VNC, vomeronasal cartilage. Scale bar = 1 mm (a, b); 150 μm (c, d); 50 μm (e).

rostrocaudally) among mammals. Primates represent an extreme, in that the gland ducts open through the VNNE around the entire perimeter in some species. In some bats and rodents, the gland ducts open dorsally at the RFE–VNNE junction. Of these epithelial types, RFE is the most variable, being patchy (ciliated alternating with nonciliated, microvillar, or nonmicrovillar) or even without surface specializations. A detailed comparison of apical cell processes in the RFE in different species is warranted. Such a study may elucidate special attributes of this portion of the VNO wall.

The Vomeronasal Nerves and Paravomeronasal Ganglia

In the lamina propria surrounding the vomeronasal neuroepithelium and underneath are the axons that coalesce into vomeronasal nerves. These neural bundles are ganglionated (termed 'paravomeronasal ganglia') in at least some mammals. They course posteriorly in the mucosa of the nasal septum toward and through the cribriform plate and the accessory olfactory bulb.

Development

Vomeronasal Organ

The origin for the epithelia of the VNO in all vertebrates is from a medial portion of the embryonic nasal pit, which itself originated from the nasal placode. In squamates and all mammals, this patch of epithelium invaginates to form a sac (in squamates) or tube, which possesses a communication to the developing oronasal cavity (initially) via a short duct.

In mammals, this invaginating mode of development occurs regardless of the final locus of communication for the VNO opening. Subsequently, a short duct of the VNO develops a communication with the NPD (Figure 6(b)) or with the nasal fossa alone (e.g., rodents). The latter is the embryologically primitive condition for amniotes. Differing patterns of craniofacial growth have been suggested to account for variation in supero-inferior position of the VNO opening in primates. Similar lines of study could focus on the different anteroposterior positioning among different taxa. For example, the differing position of the VNO opening relative to the NPD in rodents (Figure 6(a)) compared to most mammals (Figure 6(b)) may relate to the magnitude of growth of the incisive (premaxillary) bone in rodents, the result of which is a more posteriorly positioned incisive fossa compared to other mammals (Figure 6(a)). In other words, the VNO and adjacent supporting tissues may be ontogenetically decoupled during growth. Certainly the cartilaginous capsule is known to lose association with the VNO in hominoids.

Do Humans Have a Vomeronasal Organ?

Controversy concerning the human VNO, spanning decades, has been reviewed repeatedly. Conflicting results concerning the presence versus absence of the VNO in adult humans and the purported prenatal disappearance also have been thoroughly discussed. In recent years, many authors have regarded bilateral epithelial tubes, found 6–11 mm superior to the base of the nasal septum, as vestigial homologs of the VNO found in other mammals. The human VNO is nearly as spatially incongruous with that of most mammals as the cetacean blowhole is with typical mammalian external nares. Yet, in contrast, no good skeletal correlates exist to demonstrate the relationship between the two. The homology of these morphologically incongruous structures is only understood through examination of embryonic and fetal development.

Through comparison of embryos at approximately similar stages of development, developing humans are

found to possess first 'vomeronasal grooves' and then tubular VNOs in the same position as found in other mammals, including some primates. During fetal development, the human VNO becomes displaced from its initial position near the base of the nasal septum. Thus, the fetal and postnatal human VNO vestiges are positioned more superiorly relative to the palate and devoid of a cartilaginous capsule (the cartilage remains inferior to the nasal septum, as in other mammals). The displacement may be related to the magnitude of downward growth of the midfacial region that typifies hominoids (humans and great apes). Indeed, the chimpanzee possesses a positionally identical epithelial tube of the same light microscopic structure, in fetuses and adults (**Figures 2(e) and 8(f)**). Given the developmental link of the human VNO vestige to the VNO of other mammals, this epithelial tube seen in fetal and postnatal chimpanzees most certainly represents a VNO vestige. Despite the developmental link between the human VNO and that of other mammals, a number of lines of data indicate it is a nonfunctional vestige in terms of chemoreception. A transition to this nonfunctional status is discernible prenatally, in terms of second trimester loss of innervation and a conversion of its epithelium to a pseudostratified, ciliated columnar type.

Some erroneous concepts concerning the human VNO have persisted. The human VNO and its capsular cartilage do not disappear during early development. Moreover, the fetal VNO and postnatal human VNO are spatially separate from the palate as well as the incisive canal, which houses an NPD in most mammals. In adult humans, bilateral, ciliated epithelial tubes and the paraseptal cartilages persist as reduced homologs of the VNO and its cartilaginous capsule, respectively (cf. vomeronasal cartilage; **Figures 6(b) and 8(b)**).

Evolutionary Patterns

Presence of elements of the vomeronasal system in fishes is debated. Among extant vertebrate taxa, an anatomically distinct VNS exists in all higher taxa of vertebrates except archosaurs, in which the VNS has been lost. This distribution has two implications. First, an anatomically distinct VNS first developed in a common ancestor of living tetrapods. Second, numerous taxa at all hierarchical levels lost the VNS in evolutionary convergence and perhaps, in some cases, parallelism. The second point has been fodder for much discussion of adaptive aspects of VNS chemoreception (see later). An understanding of adaptive significance of this system has remained elusive, in part, because of the murky evolutionary history of this neural system.

In extant animals possessing a VNS, every functional element is a soft tissue structure, and no osseous features of the skull (foramina or fossae) may be solely related to the presence of peripheral or central elements of the VNS. In the following sections, some indirect clues to the origin and adaptive significance of the VNS are discussed.

Structural Diversity

Though varied in microanatomy, two basic morphologies of VNO are found among vertebrates. In nearly all extant mammals, the VNOs are elongated epithelial tubes that run parallel to the intersection of the palate and nasal septum. This morphology is in strong contrast to the furrow or pouch shape of the VNO that characterizes most extant nonmammalian tetrapods.

Aside from evolutionary loss (see later), the major variation seen in nonmammals concerns the degree to which the VNO is separated from the nasal cavity. In anurans and urodelans, the VNOs are akin to recesses of the nasal cavity in mammals – that is, they bear large ports of communication to the main chamber. This is consistent with two other concepts. First, solvents reach the VNO by diffusion through fluid in connecting nasal chambers in anurans. Second, the common ancestor of tetrapods, in which the VNO is hypothesized to have first occurred, bore characteristics indicating a completely aquatic existence. The simple recess would allow better access than would the ostium of a duct for such a passive delivery system. As already described, the palatal ducts of squamate reptiles or vomeronasal ducts of other vertebrates require stimulus delivery.

The tubular VNO is surely a primitive characteristic of mammals, as it exists in all subclasses. One may speculate that this morphology developed first in therapsid reptiles, yet reliable skeletal clues regarding the VNO shape have yet to be established. The close spatial relationship with the nasal septum and a relatively rostral opening appear to be characteristics that typify most extant amniotes. Less certain is the advent of the NPD opening of the VNO. The timing of evolution of an oral opening for the VNO is uncertain. Even though many extant mammals possess a 'dual-access' opening via the NPD, some therapsids did not. Most lepidosaurs possess an oral opening for VNO entry, but this is considered a derived feature, and the VNO opens into the anteromedial nasal cavity in the most primitive lepidosaur (*Sphenodon*). For these reasons, a nasal-only VNO opening is likely the primitive condition for early therapsids.

An additional consideration has been brought to light regarding the NPD development in extant mammals. Although it would appear that mammals primitively possessed a dual-access opening, during

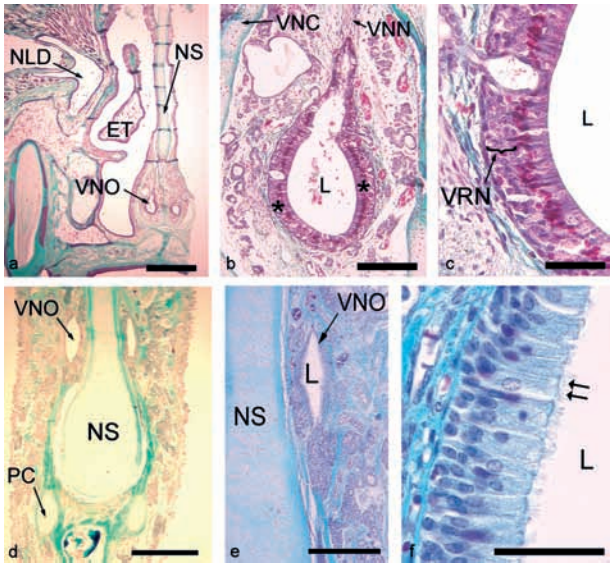

Figure 8 The vomeronasal organ (VNO) of New World primates, such as marmosets (a–c, *Cebuella*), exhibits vomeronasal neuroepithelium (Δ) both medially and laterally, in contrast to most mammals (cf. **Figures 6** and **7**). In chimpanzees (d–f, *Pan*) and humans (not shown), VNO homologs are found in an atypical position (d, e), displaced superiorly from the base of the nasal septum (NS) and spatially separated from the paraseptal cartilages (PC; homologs of the vomeronasal cartilages). Micrographs (b) and (c) and (e) and (f) are progressive enlargements of the VNO shown in (a) and (d), respectively. The VNO epithelium in chimpanzees (f) and humans is pseudostratified, columnar, with apical cilia (double arrows). ET, ethmoturbinal I; L, lumen of VNO; NLD, nasolacrimal duct (orbital end); NS, nasal septum; VNC, vomeronasal cartilage; VRN, nuclei of the vomeronasal receptor neurons. Scale bar = 1 mm (a, d); 150 μm (b); 50 μm (c, f); and 500 μm (e).

embryogenesis the VNO primitively opens to the nasal chambers in all vertebrates, including mammals. In mammals, eventual communication of the VNO with the NPD is contingent on formation of the secondary palate. Interestingly, some primates that possess a dual-access opening in adults continue after palatogenesis to transiently possess a VNO opening into the base of the nasal cavity. Subsequently, as the palate grows in depth, the VNO duct is incorporated into the palate along with a nasal portion of the NPD. In addition, some mammals with especially pronounced depth of the palate (e.g., great apes, elephants) have ducts that are partially nonpatent or absent, even though the bony canal that encloses it may be present. This raises the possibility that variation in VNO entry among mammals is partially a byproduct of different modes of facial growth, especially as regards the bony palate.

Since the VNO opens anteriorly into the NPD in prototherians, many metatherians, as well as many eutherians, this may well represent the primitive condition for mammals. However, the developmental sequence of VNO duct relative to the NPD, noted previously, makes the issue far less certain. In essence, the bar is raised for clearer criteria regarding osseous correlates of the duct to the VNO, which may in turn augment paleontological interpretation.

Although the VNO itself does not clearly relate to osseous landmarks, some researchers have suggested osseous morphologies of fossils which indirectly relate to VNO function. Hillenius has carefully outlined the potential significance of the septomaxillary foramen to the distal (nasal) end of the nasolacrimal duct in therapsids. In extant mammals, the link between Harderian gland (nasolacrimal duct-borne) secretions and VNO function is more tenuous than for squamate reptiles, but still plausible. It is possible that Harderian secretions may travel to the snout via the septomaxillary foramen, and thus may be involved in odorant binding when drawn into the VNO (indirectly) via the oral or nasal route. If so, this indirect secretory route first appeared in therapsids.

Variations of the Vomeronasal System

There appear to be two extremes among extant tetrapods: (1) snakes and other lepidosaurs have a proportionally large VNS peripherally as well as centrally, and (2) in some extant synapsids (comprising mammals, reptiles, and birds) the VNS is on the path of regression. The subclass Archosauria represents the highest taxonomic group that is characterized by VNS absence. However, the VNS has been lost independently at numerous other taxonomic levels of all classes of tetrapods. Among mammals the VNS is absent, or not intact (i.e., a vestigial VNO with no neural connections as in humans) in all cetaceans, all Old World primates, numerous chiropterans (all sirenians Megachiroptera and some Old World Microchiroptera), and proteid salamanders.

What does VNO loss or reduction imply? Several ecological correlates have been offered. Tetrapods of terrestrial stock that secondarily became aquatic or semiaquatic have no VNS or have a putatively reduced VNS. The underlying basis for VNS reduction in these taxa has not been explained in terms of selective pressures. However, it was recently observed that the ratio of two families of vomeronasal receptor genes is inverted in terrestrial tetrapods compared to aquatic tetrapods and teleost fish. An interesting implication is that terrestrial tetrapods may no longer utilize the VNS in the same manner as did their fully aquatic ancestors. One well-reasoned basis for VNO loss concerns catarrhine primates, which all lack an intact VNS (extant strepsirrhines, New World anthropoids, and tarsiers have retained an intact VNS). It is argued that loss of VNS function coincided

with the acquisition of trichromatic vision in catarrhines. The loss of VNS has been related to the replacement of pheromonal signaling via the VNS with social signals of a visual nature (e.g., perineal coloration). More proximally, VNS loss is related to reduced selection pressure and an accumulation of pseudogenes related to signal transduction in the VNS. The accumulating evidence that the VNS is not exclusively involved in pheromonal communication makes the preceding argument a bit more tenuous. However, it remains the most potent hypothesis offered to explain the evolutionary loss of a special sensory system.

Bats rival or exceed primates in the degree of inter- and intrataxonomic variation of the VNS. Only one family, Phyllostomidae, is endowed with an intact VNS. In the other 17 families of bats, the VNS remains either variable or untraceable. Chiroptera remains an attractive mammalian group for developing a second hypothesis for VNO loss, which might also be related to sensory ecology. On the other hand, it is unclear why certain taxa of chiropterans have retained a well-developed VNS. One might be tempted to speculate the functional relationship between the long, highly extensile and pointed tongues and the VNO, as in snakes, lizards, and phyllostomid bats (many of the latter have elongated muzzles, large VNOs, and highly extensile tongues, as in glossophagine bats). In these cases, there is a chance for the tongue to be introduced into the VNOs, as in snake tongue-flicking, after receiving the chemostimulus from the environment.

Historically, an understanding of VNS loss was complicated by false generalizations that entire taxonomic groups (e.g., cles, bats) lack the VNO, generalizations that were subsequently corrected. Another hindrance to our understanding of VNS loss is our preconception of VNO functions. Whereas it is well established that the VNO is critically important for sociosexual interactions in rodents, and equally important for food detection in snakes, few studies have considered that the VNO may function for multiple purposes among or within taxa. Renewed focus on dietary, ecological, and other factors may explain the VNO loss in various taxa.

Conclusions

The VNO in Extant Tetrapods

A functional dichotomy of the 'dual' olfactory systems (VNS and main olfactory system) has been reassessed in recent years. Evidence now exists that molecules of differing characteristics (volatile or nonvolatile) or functions (social or food odorants) may stimulate the VNO. It has been observed that odorant volatility seems irrelevant to animals with a fully aquatic lifestyle (e.g., urodelans). In amniotes, morphological variations of the VNO suggest considerable evolution since the initial appearance of the VNS. The extent to which the VNS of amniotes has diverged functionally from the main olfactory system would appear to relate to a terrestrial evolutionary history, quite removed from selection pressures encountered by their aquatic ancestors. Recent work has shown potential importance of the VNO to feeding in a marsupial, suggesting functional flexibility closer to that known for squamate reptiles. Thus, the VNS may ultimately be understood as serving for redundant chemosensory functions, perhaps specializing only at the level of odorant molecules.

An Elusive Ancestral Function

The ancestral function of the VNS remains unknown and difficult to deduce, due to the indirect nature of fossil evidence (see earlier). Therefore, reconstructions of VNS evolution in relation to the lifestyle of tetrapod ancestors vary greatly. For instance, the VNS was considered an adaptation of terrestrial tetrapods until recently. The first appearance of the VNS also has been linked with amphibian metamorphosis. Since proteid salamanders do not metamorphose, Eisthen concluded that the VNS was lost in proteid salamanders, and rejected the hypothesis that the VNS originated in conjunction with metamorphosis in amphibians. Eisthen brings to bear a more basic argument as well: since the VNS likely evolved in the common ancestor of lissamphibians and amniotes, and this ancestor was fully aquatic, it was not an adaptation for terrestriality. So, what novel selective pressures were encountered by prototetrapods that were not already acting on all aquatic vertebrates? Conceivably, the advent of air breathing in tetrapod ancestors was associated with odorant detection in different mediums, and/or at the air/water interface itself.

Clues to the primitive function of the VNS might still be derived from extant tetrapods. It has been hypothesized that an ancient relationship exists between the VNO and the Harderian gland of the orbit, and that the gland secretions may have served as solvents for vomeronasal chemoreception. This scenario may also relate to the functional diversity of the VNO across amniotes. The ancient connection between the Harderian gland and VNS, if correct, is unstable in amniotes, in which other secretions (septal glands) may suffice for some taxa.

Hillenius and Rehorek suggested that odorant molecules may have been "inadvertently 'trapped' on the cornea of aquatic tetrapod ancestors, specifically

when the eyes breach the water surface. Such molecules could travel through the nasolacrimal duct, along with Harderian secretions, to the VNO. The notion that odorants may enter the nasal chambers indirectly via an orbital entry is fascinating in conjunction with the image of ancestral tetrapods that lived in shallow water and evolved adaptations for air breathing. These creatures may have been the first vertebrates to encounter a novel array of chemical stimuli. In tandem with changes in the respiratory system, ancestral tetrapods may have evolved adaptations for sensing chemicals at the air/water interface.

See also: Olfactory Receptor Genes: Human Loss During Evolution; Olfactory Receptors; Olfactory Epithelium; Olfactory Bulb Anatomy; Olfactory Bulb Physiology; Olfactory Bulb Mapping; Vomeronasal Accessory System.

Further Reading

Baxi KN, Dorries KM, and Eisthen HL (2006) Is the vomeronasal system really specialized for detecting pheromones? *Trends in Neuroscience* 29: 1–7.

Bhatnagar KP and Meisami E (1998) Vomeronasal organ in bats and primates. Extremes of structural variability and its phylogenetic implications. *Microscopy Research and Technique* 43: 465–475.

Bhatnagar KP and Reid KH (1996) The human vomeronasal organ. I: Historical perspectives. A study of Ruysch's (1703) and Jacobson's (1811) reports on the vomeronasal organ with comparative comments and English translations. *Biomedical Research* 7: 219–229.

Bhatnagar KP and Smith TD (2006) The vomeronasal organ and its evolutionary loss in catarrhine primates. In: Preuss TM and Kaas JH (eds.) *The Evolution of Primate Nervous Systems*, vol. 5, pp. 141–148. Oxford: Elsevier.

Dawley EM and Bass AH (1988) Organization of the vomeronasal organ in a plethodontid salamander. *Journal of Morphology* 198: 243–255.

Døving KB and Trotier D (1998) Review. Structure and function of the vomeronasal organ. *Journal of Experimental Biology* 201: 2913–2925.

Eisthen HL (1992) Phylogeny of the vomeronasal system and of receptor cell types in the olfactory and vomeronasal epithelia of vertebrates. *Microscopy Research and Technique* 23: 1–21.

Eisthen HL (1997) Evolution of vertebrate olfactory systems. *Brain, Behavior and Evolution* 50: 222–233.

Eisthen HL (2000) Presence of the vomeronasal system in aquatic salamanders. *Philosophical Transactions of the Royal Society (London) B* 355: 1209–1213.

Evans C (2003) *Vomeronasal Chemoreception In Vertebrates. A Study of the Second Nose*. London: Imperial College Press.

Halpern M and Martínez-Marcos A (2003) Structure and function of the vomeronasal system: An update. *Progress in Neurobiology* 70: 245–318.

Hillenius WJ (2000) Septomaxilla of nonmammalian synapsids: Soft-tissue correlates and a new functional interpretation. *Journal of Morphology* 245: 29–50.

Hillenius WJ and Rehorek SJ (2005) From the eye to the nose: Ancient orbital to vomeronasal communication in tetrapods? In: Mason RT, LeMaster MP, and Muller-Schwarze D (eds.) *Chemical Signals in Vertebrates 10*, pp. 228–241. New York: Springer Verlag.

Huang G-H, Zhang J-H, Wang D, et al. (2006) Female snake sex pheromone induces membrane responses in vomeronasal sensory neurons of male snakes. *Chemical Senses* 31: 521–529.

Meredith M (1999) Vomeronasal organ. In: Adelman G (ed.) *The Encyclopedia of Neuroscience*, 2nd edn. Cambridge, MA: Elsevier.

Shi P and Zhang SJ (2007) Comparative genomic analyses identifies an evolutionary shift of vomeronasal receptor gene repertoires in the vertebrate transition from water to land. *Genome Research* 17: 166–174.

Smith TD, Siegel MI, and Bhatnagar KP (2001) Reappraisal of the vomeronasal system of catarrhine primates: Ontogeny, morphology, functionality, and persisting questions. *Anatomical Record (New Anatomist)* 265: 176–192.

Smith TD, Bhatnagar KP, Shimp KL, et al. (2002) Histological definition of the vomeronasal organ in humans and chimpanzees with a comparison to other primates. *Anatomical Record* 267: 166–176.

Wible JR and Bhatnagar KP (1996) Chiropteran vomeronasal complex and the interfamilial relationships of bats. *Journal of Mammalian Evolution* 3: 285–314.

Relevant Website

http://neuro.fsu.edu/~mmered – The Vomeronasal Organ. Dr. Michael Meredith, Florida State University (FSU), Neuroscience Program.

Voxel Based Morphometry

J Ashburner and K J Friston, Functional Imaging Laboratory, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Since the advent of magnetic resonance imaging (MRI), neuroscientists have been able to measure structures in living human brains. A large number of automated or semiautomated approaches for characterizing differences in the shape and neuroanatomical configuration of different brains have emerged due to improved resolution of anatomical scans and the development of new image processing techniques. There are many ways of identifying and characterizing structural differences among populations. Voxel-based morphometry (VBM) is one such method, which has been used widely during the past decade. Its applications range from the study of progressive supranuclear palsy to neurogenetic studies of polymorphisms and age-related changes. Unlike other approaches to morphometry, VBM does not refer explicitly to anatomy; it treats images as continuous scalar measurements and tests for local differences at the appropriate spatial scale. This scale is controlled by the smoothing. Smoothing is critical for VBM because it sets a lower bound on the anatomical scale at which differences can be expressed. It does so by removing fine-scale structure that is not conserved from subject to subject. Effectively, VBM is a scale-space search for anatomical differences. There are several reasons why VBM has proved so popular. First, in contrast to conventional analyses of anatomical structures, it tests for differences anywhere in the brain. Second, VBM can be applied to any measure of anatomy – for example, gray matter density, compression maps based on the Jacobian of deformation fields, or fractional anisotropy from diffusion weighted imaging. By choosing the data and their transformations carefully, a large range of anatomical attributes can be analyzed in a simple and anatomically unbiased way.

In summary, VBM is a simple procedure that enables classical inferences about the regionally specific effects of experimental factors on some structural measure. These effects are tested after discounting the large-scale anatomical differences removed by spatial normalization. Because these differences have been removed, VBM is not a surrogate for classical volume analysis of large structures or lesions. Furthermore, it will not replace shape analysis; VBM infers on smooth scalar fields, where constructs such as 'edges' or 'shape' have no meaning. VBM represents an established and effective complement to shape and volume analyses that is used widely in many basic and clinical contexts.

This article discusses the various stages involved in preparing data for VBM and then considers modeling and inference on these data.

Preparing the Data

VBM relies on good quality high-resolution MRIs from different subjects and uses these to identify correlations between disease severity and the spatial deployment of different tissue types. Many researchers assume that VBM can only be implemented within the SPM package; this is not the case. There are many software tools that could be used to perform the requisite sequence of operations. The MRIs are segmented, producing images that reflect the spatial distribution of a tissue type or attribute (e.g., gray matter). To compare brains of different subjects, all gray matter segments are warped to the same stereotaxic space. A correction can be applied to the data that accounts for expansion and contraction during this nonlinear spatial normalization. The normalized segments are then smoothed. This makes each voxel a measure of the proportion of the brain, in a region around the voxel that is gray matter (i.e., gray matter density). Statistical analysis using the general linear model (GLM) is used to identify regions that are significantly related to the effects under study. The GLM is a flexible framework that allows many different tests to be applied. The output is a statistical parametric map (SPM) showing regions where tissue density differs significantly among the groups. A voxel-wise SPM comprises the result of many statistical tests, so it is necessary to correct for multiple dependent comparisons using the theory of random fields.

Segmentation

The images are typically partitioned into different tissue classes using one of a number of segmentation techniques. Usually, VBM involves an analysis of gray matter, but it is possible to work with other tissue classes or any other scalar measure of anatomy (e.g., fractional anisotropy). To accurately segment an image, the image needs to show clear differentiation among tissue types. Usually, high-resolution T1-weighted MRIs are used, although multispectral images may allow more accurate tissue classification (Figure 1). Different segmentation models have different requirements. For example, many approaches take smooth intensity inhomogeneities into account, so they can deal with some of the artifacts that arise in MRIs.

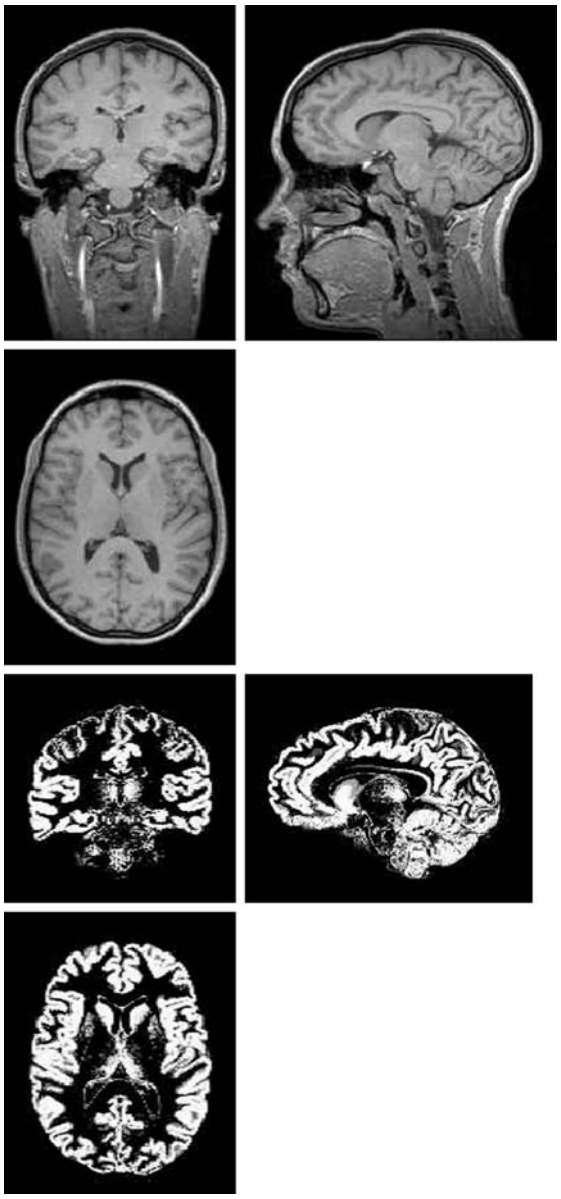

Figure 1 The result of segmenting the gray matter from a T1-weighted MRI. Note that toward the top of the brain, the segmentation is less accurate because the gray matter is hardly visible in the original image. Similarly, in gray matter regions such as the thalamus, the intensity is closer to that of white matter, resulting in incorrect segmentation. Some partial volume voxels can be seen around the ventricles. These arise because voxels containing signal from both white matter and CSF have intensities that are close to those of gray matter.

Most automated segmentation algorithms assume that the intracranial cavity contains only gray matter, white matter, and cerebrospinal fluid (CSF), so they may be confounded by lesions. However, there are some that have been specially designed for dealing with such pathology, although they generally require multispectral images (e.g., aligned T1- and T2-weighted scans).

Many segmentation approaches use prior spatial information in the form of tissue probability maps, which inform the segmentation algorithm about the approximate spatial distribution of the various tissues. In most cases, the tissue probability maps need to be registered with the image to segment, but the segmentation framework of SPM5 combines this registration with the tissue classification.

Spatial Normalization

Spatial normalization involves warping all the gray matter images to the same stereotaxic space, which is achieved by matching to a common template image (Figure 2). There are many nonlinear registration methods that can be used for this. The unified segmentation–normalization approach in SPM5 is a robust procedure because it accommodates conditional dependences among the segmentation and normalization parameters. It is a low-dimensional method which uses approximately 1000 parameters to describe the deformations. Other approaches have many more degrees of freedom and may be more accurate. However, just because a method can warp anything to anything else does not mean that it furnishes the most likely registration (high-dimensional procedures generally overfit the data). For example, the evaluations of Hellier et al., which compared registration accuracy of different approaches, found that the simple spatial normalization of SPM2 was as accurate as other methods with more degrees of freedom.

The objective of VBM is to localize regions (in stereotaxic space) where there are significant differences. Accurate spatial normalization and segmentation ensures regional differences can be attributable to local effects rather than systematic, pathology-induced misregistration artifacts. However, registration can never be perfect because there is no correspondence between one subject and another at a fine spatial scale. For example, many sulci are shared between brains, but this is not the case for all. Generally, the primary sulci, which are formed earliest and tend to be quite deep, are conserved over subjects. Sulci developing later are much more variable. Therefore, some sulci can be matched objectively, whereas others cannot. VBM accommodates the fact that the scale of anatomical correspondence, and implicitly the scale of differences in anatomy, has a lower bound by smoothing the segments.

Jacobian Adjustment

Nonlinear spatial normalization changes the volumes of brain regions. This has implications for the interpretation of what VBM tests. The objective of VBM is to identify regional differences in the composition of

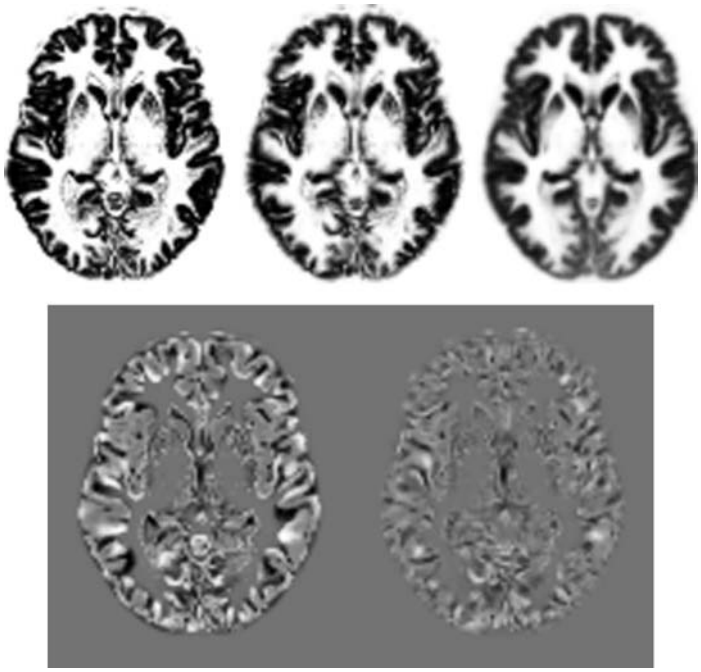

Figure 2 An image (top left) is warped to match a template (top right) to produce a spatially normalized version (top center). For clarity, the original image was approximately aligned with the template, and the warping was only done in two dimensions. The bottom row shows the difference between the template and image, both before and after the warping.

brain tissue. To preserve the actual amounts of a tissue class within each voxel, a further processing step can be incorporated that multiplies (modulates) the partitioned images by the relative voxel volumes before and after warping. These relative volumes are simply the Jacobian determinants of the deformation field (Figure 3). In the limiting case of extremely precise registration (using very high-dimensional registration), all the segments would be identical. The adjustment preserves the differences in volume of a particular tissue class (Figure 4).

The deformations from spatial normalization map points in a template (x_1, x_2, x_3) to equivalent points in individual source images (y_1, y_2, y_3). The derivatives of the deformation field can be thought of as matrix at each point. These are the Jacobian matrices of the deformations and are defined by

$$J = \begin{Bmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 & \partial y_1/\partial x_3 \\ \partial y_2/\partial x_1 & \partial y_2/\partial x_2 & \partial y_2/\partial x_3 \\ \partial y_3/\partial x_1 & \partial y_3/\partial x_2 & \partial y_3/\partial x_3 \end{Bmatrix}$$

The determinant of this matrix encodes the relative volumes of deformed and undeformed structures.

Smoothing

The warped gray matter images are now smoothed by convolving with an isotropic Gaussian kernel. This makes the subsequent voxel-by-voxel analysis comparable to a region of interest approach because each voxel in the smoothed images contains the average amount of gray matter from around the voxel (where the region around the voxel is defined by the form of the smoothing kernel; Figure 5). This is often referred to as gray matter density, but it should not be confused with cell packing density measured cytoarchitectonically. Critically, smoothing removes finescale structure from the data that is not conserved from subject to subject. This increases the sensitivity of VBM to differences that are expressed at a larger spatial scale. The smoothing conforms to the matched filter theorem, which states that the smoothing should match the scale of the difference in question. Normally, the smoothing kernel is Gaussian with a full width at half maximum (FWHM) of between 4 and 16 mm. By the central limit theorem, smoothing also has the effect of rendering the data more normally distributed, thus increasing the validity of parametric statistical tests.

Statistical Modeling and Inference

Voxelwise statistical tests are performed on the preprocessed images using standard linear models (GLM). The results of these tests are an SPM showing significant regional effects. The GLM allows many different tests to be applied, ranging from group comparisons and identification of regional correlates of disease severity or age to complex interactions among

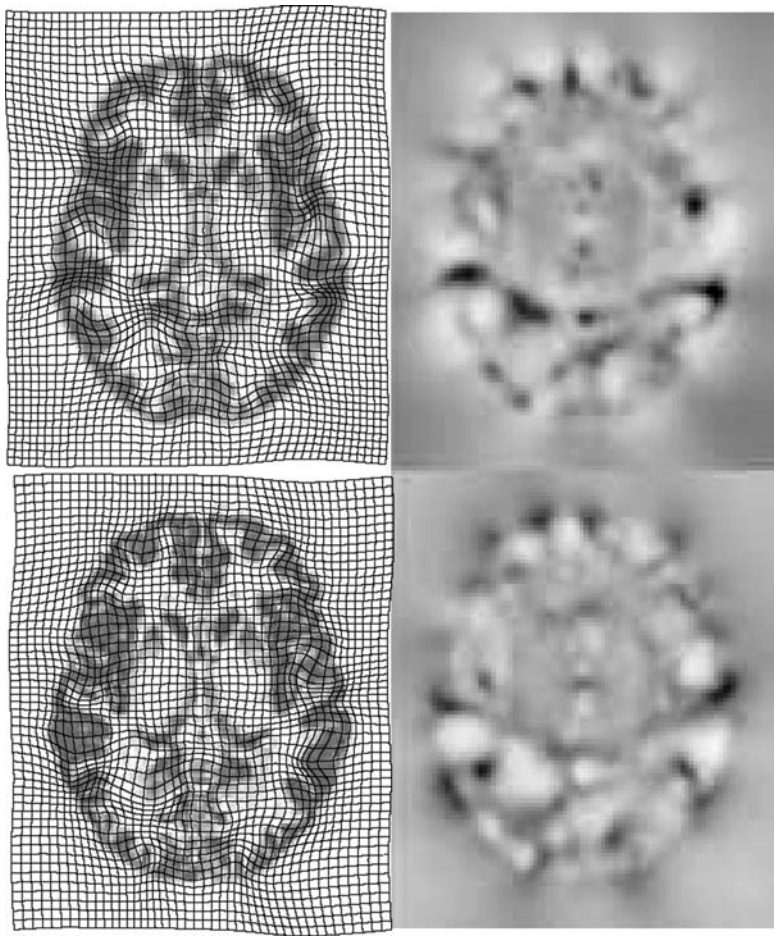

Figure 3 Warping an image results in some regions growing and others shrinking. The top left image is a spatially normalized image with deformations overlayed (the same data as in Figure 2). Beside it is a map of Jacobian determinants, in which darker regions indicate regions where the deformation lines are closer together. The bottom left image is the original undeformed image, with the deformations that would spatially normalize it overlayed. This deformation field is the inverse of the one shown above it. Beside it are the Jacobian determinants showing expansion and contraction.

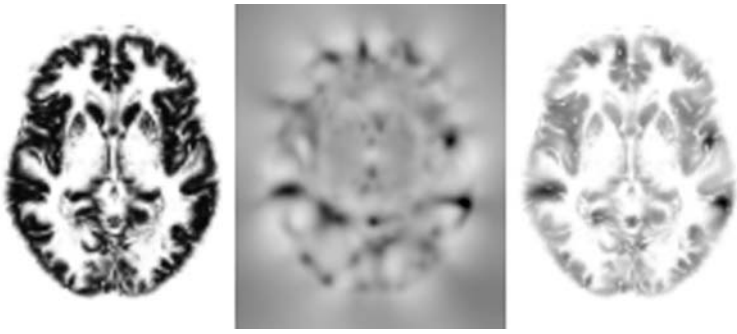

Figure 4 The image on the left is a warped gray matter image. In the center is a map of the Jacobian determinants from the warp (see Figure 3). The image on the right is the result of multiplying them together, such that the amount of gray matter is preserved.

these effects. In the GLM, the model is expressed as a design matrix. Each row of this matrix corresponds to a scan, whereas each column is some effect that is modeled. At each voxel, the optimal linear combination of effects is computed in terms of parameter estimates. Statistics are then based on the parameter estimates at each voxel. Standard statistics (t tests and F tests) are used to test the hypotheses and are valid providing the residuals, after fitting the model, are normally distributed.

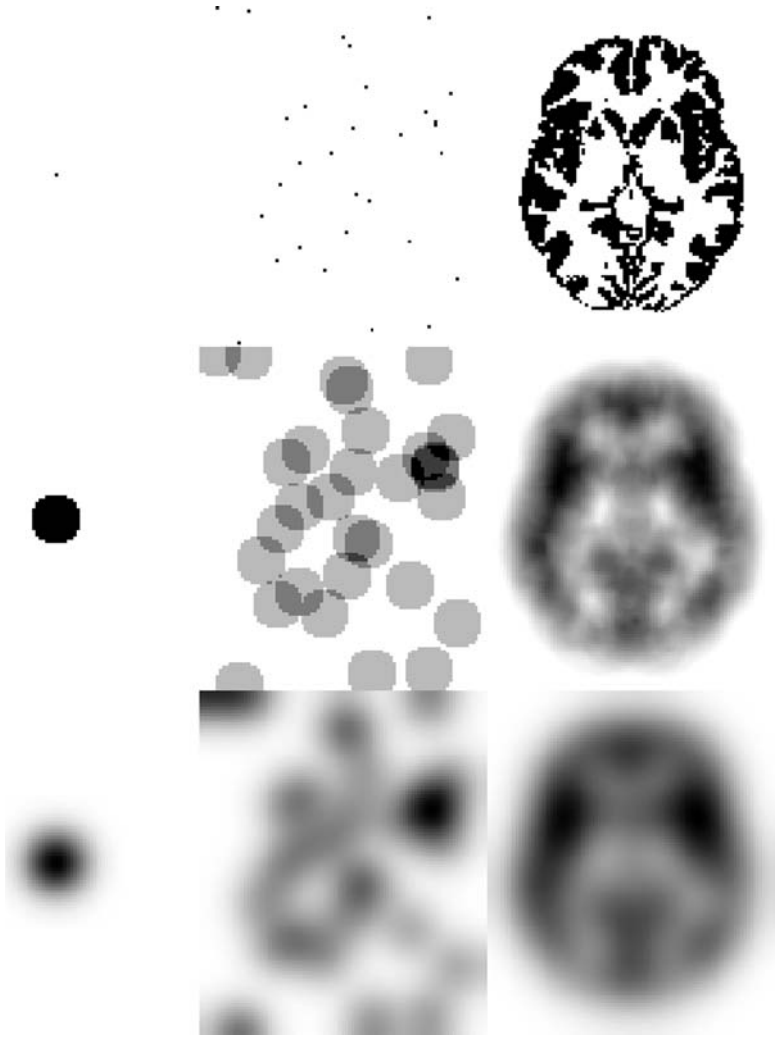

Figure 5 Smoothing effectively converts the images to maps containing a weighted count of the number of gray matter voxels within each region. The top row shows three unsmoothed images. The middle row shows these images after they have been convolved with a circular-shaped kernel. It should be clear that the result is a count of the values within each circular region. The bottom row shows the images after convolving with a Gaussian-shaped kernel. The result is a weighted count of the values around each point, where values in the center are weighted more heavily.

Occasionally, researchers want to mix images from different scanners or data acquired using different sequences. The effects of different types of images can generally be modeled by including confounds in the design matrix. Note that for a comparison of one group versus another, where one group is collected on one scanner and the other group is collected on another, modeling the scanner effect will explain away all group differences. An important confound is the global or whole measure.

Global Normalization and Other Confounds

Global normalization is a generic issue in neuroimaging. SPM is used to identify regionally specific (i.e., nonglobal) effects. This can be ensured by treating the global signal as a confound, which is explained away during inference. Because the global effect is the sum of all regional effects, this means that the sum of all regional effects, over the brain, is zero. Global normalization therefore partitions effects into regional and global components. In the context of gray matter–VBM density, this involves modeling the global gray matter as a linear confound (or by scaling each segment by its global mean). The implicit partitioning into regional and global effects finesses the characterization of anatomical differences. For example, a regional increase in gray matter density can indicate one of two things: it could be due to local hyperplasia, or it could be a local sparing of the region in the context of diffuse cortical atrophy everywhere else. The resolution of this ambiguity rests on inference about the global measure. In the first instance, there will be no global difference, but it would be significant in the second.

There are other confounds people include in statistical models that remove variance in gray matter density that could have been caused by nonspecific mechanisms (in relation to cohort selection or experimental factors). These can be expressed in a regionally selective fashion (e.g., age or experimental factors) or globally. For example, larger brains have proportionally more white matter than gray, leading some researchers to treat total intracranial volume as a confound.

Inference

Following application of the GLM, the significance of any regional difference is ascertained using the parameter estimates and their associated statistics. A voxelwise SPM comprises the results of many statistical tests, so it is necessary to correct for multiple dependent comparisons. Without any correction, the number of false-positive results would be proportional to the number of independent tests. A Bonferroni correction would be applied if the tests were independent, but this is not the case because of the inherent spatial smoothness of the data. In practice, the effective number of independent statistical tests is determined using random field theory (RFT). RFT provides a correction for multiple dependent that controls the rate of false-positive results. Parametric statistical tests are based on assumptions about the data. In particular, they assume that the residuals after fitting the model are normally distributed. There is a special caveat for VBM concerning the normality assumption: gray matter segments contain values between 0 and 1, where most of the values are close to either of the extremes. In this case, smoothing is essential to render the residuals sufficiently normal. For group comparisons, normality can be ensured with smoothing at or above 4 mm FWHM. For single-case comparisons, numerical studies suggest that at least 12 mm should be used. An alternative for unbalanced designs (e.g., single-case studies) is the nonparametric methods.

The t or F fields produced by SPM are thresholded at some value. Regions of the field that survive the threshold (called clusters, excursion sets, or simply blobs) are then examined further. SPM allows inferences to be made based on the spatial extent of blobs, under the assumption that large blobs are less likely to occur by chance than small blobs. In the SPM software implementation of RFT, inferences on the extent of a cluster (but not the height of a peak) assume that the smoothness of the residuals is roughly the same everywhere. This is ensured for data that have been smoothed. However, without smoothing, large blobs that occur by chance in smooth regions may be declared significant, whereas smaller true blobs in rough regions may be missed.

Nonstationary behavior of error terms can also misdirect the interpretation of true positive results. Spatial changes in error variance can cause blobs to move toward regions with low variance and away from regions with high residual variance. This explains the occasional finding of significant differences outside the brain (because the residual variance approaches zero as one moves away from the brain). Accurate localization of differences should not use the SPM. As with all applications of SPM, interpretation rests on looking at the parameter estimates that subtend the significant effect. In this instance, the gray matter difference maps provide the veridical localization.

Frequentist statistical tests cannot be used to prove a hypothesis, only to reject a null hypothesis. Any significant differences that are detected could be

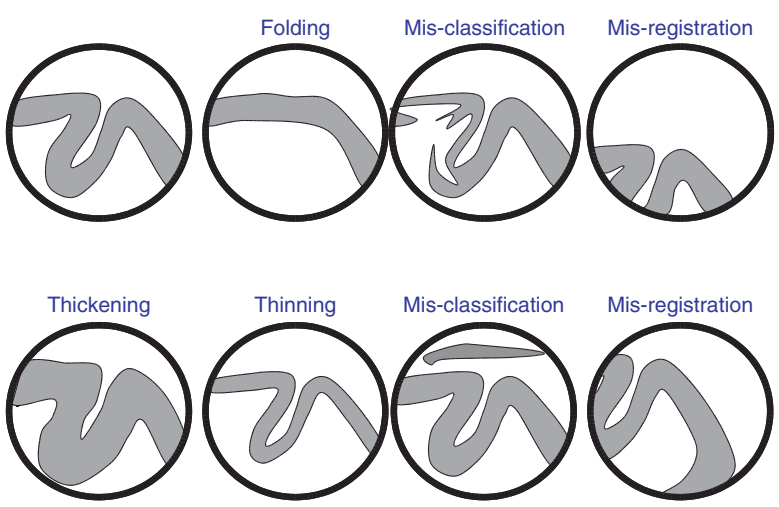

Figure 6 Significant differences determined using VBM can be explained in many ways.

explained by a number of different causes, which are not disambiguated by the inference *per se*. When the null hypothesis has been rejected, it does not impute a particular explanation for the difference if there are several causes that could explain it (Figure 6). The preprocessing employed by VBM manipulates the data so that the ensuing tests are more sensitive to some causes relative to others. In particular, VBM has been devised to be sensitive to systematic differences in the density of a particular tissue class, but significant results can also arise for other reasons. A critique of VBM is that it is sensitive to systematic shape differences attributable to misregistration from the spatial normalization step. This is one of a number of potential systematic differences that can arise. This does not mean that VBM is invalid but the explanation for the significant result may be complicated. VBM will only detect systematic group differences. This applies to misregistration. Because the registration is driven by anatomy, any difference must be due to anatomical differences, but these may not be meditated locally (e.g., misplacement of the cortex by subcortical lesions). Some explanations can be trivial (e.g., some patient groups may have to be positioned differently in the scanner) and it is important to identify confounds of this sort. In short, there may be real differences among the data, but these may not necessarily be due to reductions in regional gray matter.

VBM in Clinical Research

Neurodegeneration involves regional loss of tissue over time. Volumetric changes can be inferred on voxel compression maps from high-resolution warping algorithms applied to longitudinal data. This approach has shown that the rate of tissue loss, rather than tissue distribution, has a better predictive validity in relation to disease progression. Comparisons of these maps between subjects requires some kind of regional averaging, either in the form of parcelation or spatial smoothing, as in VBM. Regions of compression and dilation often lie close to each other, so simple regional averaging approaches may cause differences to cancel. The most straightforward approach involves partitioning the compression maps into different tissue classes (cf. the RAVENS map approach). VBM with simple intrasubject rigid registration of longitudinal data is a simpler way of achieving similar results. Although it is useful for research, VBM may not be the most sensitive approach for diagnosing a disease in an individual. For example, VBM may be able to state that Alzheimer patients have smaller hippocampi than controls, but it is difficult to state for certain that an individual is in the early stages of the disease simply by examining hippocampal volume. We have contrasted generative models of regional pathology and recognition models for diagnosis and classification. Classification approaches are generally not interested in regionally specific differences but try to find the best nonlinear function of the data, over the entire brain, which predicts diagnosis. Multivariate kernel methods, such as support vector machines, relevance vector machines, or Gaussian process models, are established recognition models that may be useful in diagnosis and endophenotyping.

See also: Magnetic Resonance Spectroscopy; Neuroimaging; Perfusion MRI.

Further Reading

Allen JS, Bruss J, Brown CK, and Damasio H (2005) Methods for studying the aging brain: Volumetric analyses versus VBM. *Neurobiology of Aging* 26: 1275–1278.

Davatzikos C, Genc A, Xu D, and Resnick SM (2001) Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy. *NeuroImage* 14: 1361–1369.

Frackowiak RSJ, Friston KJ, Frith CD, et al. (eds.) (2003) *Human Brain Function,* 2nd edn. San Diego: Academic Press.

Kipps CM, Duggins AJ, Mahant N, Gomes L, Ashburner J, and McCusker EA (2005) Progression of structural neuropathology in preclinical Huntington's disease: A tensor based morphometry study. *Journal of Neurology, Neurosurgery and Psychiatry* 76: 650–655.

Mechelli A, Price CJ, Friston KJ, and Ashburner J (2005) Voxel-based morphometry of the human brain: Methods and applications. *Current Medical Imaging Reviews* 1: 105–113.

Toga AW (ed.) (1999) *Brain Warping*. San Diego: Academic Press.

Wright IC, Ellison ZR, Sharma T, Friston KJ, Murray RM, and Mcguire PK (1999) Mapping of gray matter changes in schizophrenia. *Schizophrenia Research* 35: 1–14 [SPM Software: http://www.fil.ion.ucl.ac.uk/spm].

Wright IC, McGuire PK, Poline J-B, et al. (1995) A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *NeuroImage* 2: 244–252.

Walking in Invertebrates

F Delcomyn, University of Illinois, Urbana, IL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

A major advancement in the evolutionary history of invertebrates was the development of a hard external skeleton such as that found among the arthropods (insects, spiders, crustacea, etc.). This was important because a skeleton allows for the development of segmented legs, and legs make possible efficient walking. With two exceptions, all invertebrates with true legs are arthropods. The exceptions are the soft-bodied Onycophora and Tartigrada, evolutionary relatives of the arthropods.

From a neurobiological perspective, an important question is how arthropods are able to coordinate their legs as they walk. Although neurobiologists have identified general principles that seem to apply to walking coordination in all arthropods, the application of these principles in different arthropods can lead to rather different patterns of walking because body shapes and sizes, number of legs, and living conditions are dramatically different in different groups of these animals. For example, the long and slender millipedes and centipedes have large number of legs, upward of 100 pair in some species. The more compact spiders and scorpions typically have eight pairs, and insects only six. Some crustacea have more than a dozen pairs of legs, but the better known crabs, crayfish, and lobsters have only ten, with the front pair usually modified structurally into claws that are not used in walking. In addition to these differences in body plan and number of legs, the terrestrial arthropods have different mechanical requirements for walking than do the aquatic crustacea, which, since they walk underwater, have water's buoyancy to assist them.

Although there are more species of arthropods than all other animals combined (entomologists estimate well over a million species of insects alone), walking has been studied in only a handful. Furthermore, the study of walking in invertebrates is itself a rather minor subfield of neurobiology. Nevertheless, such studies have contributed to neurobiology out of all proportion to their numbers. Contributions have come mainly in the field of motor control, especially in relation to the problem of how repetitive movements can be generated and coordinated, and also in bioengineering.

This article covers the main gait patterns observed in arthropods, the ways in which these patterns are generated and maintained, and how animals modulate their gaits. It concludes with an overview of recent work on robots.

Gaits

All invertebrates whose walking has been studied exhibit what is called a metachronal rhythm in the movements of their legs. Such rhythms have the appearance of a traveling wave of leg movements usually moving from the rear to the front of the animal. This is easiest to see and visualize in the many-legged myriapods (centipedes and millipedes). In most of these animals, each leg on one side of the body lifts for its forward swing a little later than the leg behind it, producing the wavelike appearance that gives the motion its name. Legs on the other side of the body are out of phase such that as a leg on the left side is in the air swinging forward, its partner on the right side is on the ground moving backward.

Aquatic crustacea also exhibit a metachronal wave, but it is not so obvious in these animals because the water provides sufficient buoyancy that the animals need only support a small fraction of their body mass on their legs. This allows the legs to move in ways that on land would result in unbalanced positions. The result is that coordination between legs in a crab or crayfish is much less tightly coupled than it is in arthropods that walk on land. Furthermore, evidence from trace fossil tracks of extinct arthropods suggests that at least some of these walked using a swimminglike pattern of leg movement in which the two legs of a pair on opposite sides of the body moved forward and backward together rather than alternately, a pattern not seen on land.

Insects and spiders, the other terrestrial arthropods, also exhibit a metachronal wave of leg movement during walking, but with only six (insects) or eight (spiders) legs, a metachronal sequence is not usually obvious in these animals. Nearly all insects use one simple gait most of the time. This gait, the alternating tripod or alternating triangle gait (Figure 1), consists of stepping in which the front and rear legs on one side of the body move forward together with the middle leg on the other side. The tripod of legs thus formed alternates in its movements with the other three legs, hence giving the gait its name.

The metachronal underpinning of this gait was first pointed out by Don Wilson about 40 years ago. Wilson saw that occasionally an insect will walk very slowly, moving only one leg forward at a time. When it does this, the order of movement is rear, middle, front leg

on one side of the body, followed by rear, middle, front on the other side, an obvious metachronal sequence. Wilson's insight was to notice that if a second metachronal wave were to start before the first one had finished, the overlap of the waves could result in the swing of a rear leg to start a second metachronal wave coinciding with the swing of the front leg that represents the last part of the previous wave. Because the two legs of a pair on one body segment move essentially in the opposite direction, the effect is to have a rear and front leg on one side step together with the middle leg on the other side to form a triangle (Figure 2).

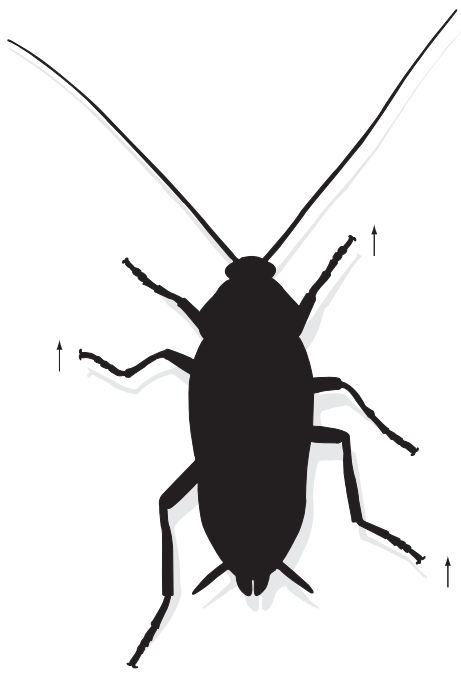

Figure 1 The alternating tripod gait of an insect. The front and rear legs on one side of the insect move together with the middle leg on the other side of the body (arrows). The legs forming one tripod alternate in their movements with the legs forming the other tripod so that as one triplet of legs is lifted to swing forward the other triplet is put on the ground and pushes back. Reproduced from Adelmari G and Smith BH (eds.) (1999) *Encyclopedia of Neuroscience*, 2nd enlarged and revised edition, p. 966, with permission from Elsevier.

The alternating triangle gait is the most common gait among insects, but other gaits may be seen. Some rather slow-moving insects such as the stick insects may employ other gaits, and even fast-moving ones like cockroaches may occasionally walk so slowly that the alternating triangle gait breaks down. Furthermore, when grasshoppers and some of their relatives with extremely long legs walk very slowly, they may take only one step with a rear leg for every two steps of the front and middle legs. Praying mantises sometimes walk as quadrupeds, holding up the two front legs ready to capture prey as they walk. Under some conditions, spiders may also walk with the two front legs held up, in which case they use an alternating tripod gait.

An essential feature of an alternating triangle gait is that each leg moves oppositely from each of its nearest neighbors. For example, in an insect, if a left middle leg is in mid-swing, then its nearest neighbors, the left front and rear and the right middle, are all on the ground in mid-stance. The nearest neighbors of a right rear leg in mid-stance, the right middle and left rear legs, will both be in the swing phases of their stepping cycles. This pattern of adjacent legs being in opposite phases of their stepping cycles occurs in spiders and crustacea as well as in insects but not as regularly as it does in insects. In spiders, the resulting pattern is called an alternating tetrapod gait. When ten or more legs are involved, as with crabs or crayfish, no special name is given to the gait.

Neural Control

Generating Rhythmic Movement

A full understanding of how animals walk must include an understanding of how the sequences of

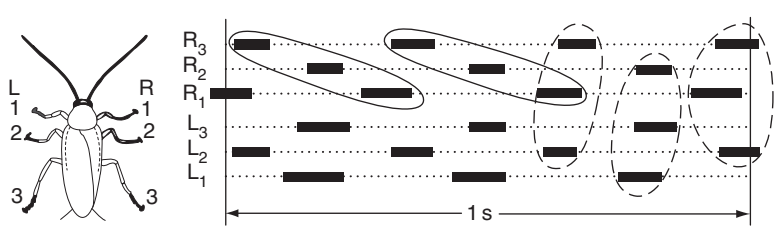

Figure 2 Metachronal rhythm of an insect during walking. The diagram represents the stepping of an insect, with time running from left to right. The numbers and letters represent the legs of the insect (left), L1 for left front, R2 for right middle, and so forth. The black bars represent swings of each leg and the dots between bars represent periods when each leg is on the ground. For example, during the walking sequence shown in the diagram, the right rear leg (R3) begins with a swing and ends with a swing, whereas the right middle leg (R2) begins on the ground and ends on the ground. The solid-line ellipses show metachronal sequences of swings, rear, middle, and front legs on the right side. Note that the last swing of the first metachronal sequence (the swing of R1) overlaps the first swing of the next (the swing of R3). The dashed-line ellipses enclose the swings of the three legs that form each tripod, R3, L2, R1 and R2, L3, L1. The tripods form due to overlap between successive metachronal waves of movements. Reproduced from Wilson DM (1965) Proprioceptive leg reflexes in cockroaches. *Journal of Experimental Biology* 43: 397–409, with permission.

leg movements that animals use is generated. Legs move in a specific sequence because the muscles that control the legs are activated in a particular sequence. Muscles are made to contract by the activity of the motor neurons that innervate them. Hence, to understand how a particular sequence of leg movements is generated, one must understand how the nervous system is able to generate an appropriate sequence and pattern of motor nerve impulses in the neurons that control leg muscles.

Reaching such an understanding has been a major goal for neurobiological research for decades. The problem is not confined to invertebrates – the walking of vertebrate animals must also be understood in terms of the sequential activation of motor neurons. Hence, research on invertebrate animals has had an impact on our understanding of walking in vertebrates as well. In fact, some of the most fundamental ideas about how walking is coordinated and controlled have come from work on invertebrate animals.

The common feature of all locomotion, and indeed of any behavior involving repetitive, rhythmic movements, such as flying, swimming, or scratching, is an alternating extension–flexion sequence of a limb or other body part. Research on a variety of animals and many different kinds of rhythmic behavior has shown that in all cases the rhythm is generated in the nervous system by a small network of neurons dedicated to this task. The neurons are interconnected in such a way that, once activated, they will generate trains of nerve impulses that are directed first to extensor and then to flexor muscles. These neural networks are often referred to as central pattern generators (CPGs).

Two features of pattern-generating neural networks are noteworthy. First, although feedback from sense organs activated by the movement being generated plays a critical role in shaping the final movements in virtually all rhythmic behavior, it does not appear to be necessary for the generation of the basic rhythmic neural output. Instead, the alternating rhythm is produced as an emergent property of the neurons that together form the network. It is this property of a 'centrally' generated pattern, that is, one generated by the central nervous system without the necessity of external timing cues, that gives CPGs their name.

Second, among all the CPGs that have been studied, most are functionally based on reciprocal inhibition between some of the neurons in the network. Neurons are said to be reciprocally inhibiting if each neuron forms an inhibitory synapse on its partner. Although modeling studies have shown that a single pair of reciprocally inhibitory neurons can generate an alternating output when they receive some constant stimulation, no such simple network has been found in any animal, and it appears likely that more than a single pair of neurons is necessary in order to ensure the stability and robustness of the network's output.

Since CPGs generate alternating output to two sets of muscles, such as flexor and extensor muscles in a leg, it seems reasonable to assume that a single CPG generates the basic rhythmicity of each leg of a walking invertebrate. If this were the case, the output of the CPG during one phase of its activity would be to all the flexors of a limb, and the output during the other phase of its activity would be to all the extensors. Such an arrangement would require that all the flexor muscles and all the extensor muscles act synergistically, contracting essentially together. Kinematic studies of the movements of legs in walking insects suggest that the different segments of each leg do move in synchrony. Nevertheless, research has also shown that in some insects (stick insects), there is a single and separate CPG for each leg segment, meaning that at least three CPGs are necessary to control the movement of each leg. Multiple CPGs controlling a single limb have not yet been described in other arthropods. Simple observation suggests that they may well exist. Insects, for example, will move a leg rhythmically while grooming and searching for a foothold, but these movements are not identical to the movements that legs adopt during walking, indicating that there must be some way for the animal to flex and extend the segments of each leg in combinations other than those used during walking.

Coupling Rhythmic Generators

An important question in the field of motor control is how the different CPGs are coupled together functionally. Some sort of communication between the neural networks must be present to ensure that the different parts of the animal move with proper timing relative to one another. For CPGs that reside in the same ganglion of the nervous system (a single ganglion generally controls a single pair of legs in many insects and crustacea), neurons of one CPG may send branches directly to another CPG. Hence, it is believed that the multiple CPGs thought to control the movements of different segments of each single leg of an insect are able to communicate directly with one another without necessarily involving intermediary interneurons in order to ensure proper coordination.

The question of how movements of different legs are coordinated with one another is not so easily resolved, however, on both morphological and functional grounds. In some arthropods, such as flies, spiders, and crabs, the separate ganglia that control the movements of separate pairs of legs have fused into a single ganglionic mass. In these cases there could

be direct communication between CPGs, although such communication has not been demonstrated experimentally. In other arthropods, such as crayfish and cockroaches, each ganglion that controls movements of a single pair of legs is anatomically separate from its neighbor ganglia and connected to it by bundles of neurons. This means that in order for CPGs in these different ganglia to communicate with one another, they must do so via interneurons, neurons that travel from one ganglionic mass to another. There is physiological evidence in cockroaches and crayfish for the presence of interneurons that provide information to one CPG about the activity of another. However, this evidence provides little information about the detailed properties of such interneurons and how they might interact with CPGs to help ensure proper coordination.

Sensory Feedback and Rhythm Generation

The discussion so far has concentrated on coordination mechanisms that depend on information carried from one CPG to another within the central nervous system. But this is not the only way that coordination could in principle be achieved, and experiments show that in some arthropods coordination between legs does not depend at all on neuronal connections between the CPGs. Instead, coordination is achieved indirectly, through the influence of feedback from sense organs in the legs. The legs of all arthropods are heavily invested with sense organs, including many that provide information about leg position and movement and about the strains that are experienced by a leg. During walking, signals from such sense organs are directed back into the nervous system, where they can interact with CPG networks to ensure proper coordination of the legs.

The role of this sensory feedback varies from animal to animal, and in some cases it varies with the speed at which the animal walks. In the American cockroach, for example, which has an enormous range of walking speeds, sensory feedback is critical in ensuring the proper sequence of leg movements during slow walking but appears to have little influence when the insect is moving fast. At high speeds of walking, coordination appears to be achieved entirely via internal central nervous connections. Stick insects, on the other hand, which are generally slow moving, use sensory feedback extensively. Experiments on these insects have shown that without sensory feedback very little coordination between the movements of different legs is possible.

Control of Walking

By far the bulk of research on the walking of arthropods has focused on the problem of coordination, that is, on how the nervous system is able to generate the appropriate sequence of neural commands that allow the legs to move in a coordinated fashion. However, researchers also want to understand how the behavior is regulated. For example, how is walking started and stopped? How are the speed and direction of walking determined? These processes are controlled by the brain.

In insects, different regions of the brain interact to regulate starting and stopping. Cutting off the head of an insect, a procedure that severs the connections between the head and the rest of the body, essentially eliminates spontaneous locomotion altogether. Cutting the nerve trunks that connect the superesophageal and subesophageal ganglia (the two parts of an insect brain) produces an insect that walks nearly constantly, even continuing to move its legs after bumping into a wall. Recent studies of the effects of specific mutations of restricted areas of the brain have strengthened the view that the superesophageal ganglion contains centers that inhibit walking. Disruption of the mushroom bodies in the brains of *Drosophila* by targeted mutations tend to lengthen the durations of episodes of walking in these insects. Hence, the mushroom bodies may help provide the inhibition that the superesophageal ganglion typically exerts on the locomotor centers in the subesophageal ganglion. Mutation-induced lesions in other parts of the insect brain may also affect walking. Lesions of the central complex, for example, have been shown to disrupt coordination between legs on the two sides of the body, showing that not all coordination is a function of the ganglia in the thorax.

Regulation of speed and turning is also a brain function, but little is known about it. Speed of walking is hypothesized to be controlled by the level of excitation provided by the brain to the locomotor centers in the thorax, but how this is done is unknown. As for turning, it has been shown that electrical stimulation of specific brain regions in an insect can induce turning in one direction or another. Such observations have led researchers to infer that the direction and tightness of a turn are determined in the brain.

Behavioral observations of insects show that turns are accomplished mainly by differences in stride length. Strides on the side of the body toward which the insect is turning (the inside of the turn) are significantly shorter than those on the side away from the direction of the turn (the outside). Shorter steps for legs inside a turn has been seen in turning crayfish as well. Stride length may not be the only factor in turns. Studies in cockroaches have shown differences in the way in which muscles in the leg segments are used in legs inside and outside a turn. Furthermore, during sharp turns (turns with a very short radius), insects

may uncouple legs on the two sides of the body, in extreme cases even moving some legs backward while others move forward.

Modulation of Walking

Most studies of walking in arthropods have concentrated on the behavior with little regard for the context in which it is expressed. However, arthropods walk for many different reasons and under many different circumstances. Insects and spiders, for example, may walk on the stems of plants or on the ground. Some may seek sheltered areas under or along which they walk, and many use walking or running as a means to escape from predators or to capture prey.

Study of insects walking along narrow surfaces have shown that an alternating tripod gait may still be used by an insect walking along the stem of a plant but that the legs are adjusted so the feet can contact the stem under the insect's body. This indicates that the locomotor system can be quite flexible in the exact placement of the feet during walking.

A number of insects seek out a solid surface under or along which to walk. Cockroaches, for example, will maintain a more-or-less constant distance from a wall along which they are walking. If the wall curves, they turn along with it. If the wall is made jagged like the folds of an accordion and the insect is allowed to walk along it, the insect will follow the folds in the wall most of the time, weaving in toward and out away from the wall as appropriate. Studies suggest that the insect uses its antennae to assess the distance it is from the wall and adjusts its walking as necessary. Hence, although it is easy to think of straight walking on a smooth, level surface as being highly stereotyped, it is nevertheless true that the animal has excellent dynamic control of its locomotion and can modulate it as needed according to circumstances.

This can especially be seen in the case of escape running. Most insects avoid a disturbance by jumping or flying (or jumping and then flying). However, even though they have wings, some insects use walking as their primary means of escape. Cockroaches, for example, react to disturbances that may be airborne or delivered via touch to the body. Irrespective of the stimulus, the result is that the insect turns away from the stimulus, then begins to run. The turn has been analyzed both behaviorally and physiologically. If the stimulus strikes the animal from one side, the turn is effected mainly by a pivot around one or two feet on the side of the body away from the stimulus. Which feet are used depends on how the feet are positioned relative to the body when the stimulus is detected. The rear foot on the stimulated side provides most of the thrust for the turn. Once the insect has turned to the desired direction, leg movements transition into a typical tripod gait.

Physiological study of escape turns suggests that they are controlled by activation of neural centers in the ganglia containing the leg motor neurons. There the signal from the sense organs that have detected a disturbance activates the locomotor system, first effecting a turn if that is necessary. At the same time, neural signals are sent to the brain, which effectively takes over control of the subsequent forward running. Although rapid escape is usually considered an innate response that cannot be modified, a cockroach initiating an escape response toward a wall can turn away from the wall as soon as it senses it with its antennae even if the wall is so close that the insect can detect it after only a single step. This is another indication that the locomotor control system can be extremely flexible.

Practical Ramifications

Readers might expect that studies of walking in invertebrate animals, a rather esoteric subject, would have no practical impact. Certainly, study of the mechanisms by which insects and crustacea coordinate their legs has been important scientifically by showing how rhythmic movements of any kind are generated and coordinated. The demonstration by Pearson and Fourtner more than 30 years ago that a small network of nonspiking interneurons could generate the alternating flexion–extension movements of cockroach legs, for example, stimulated much research into the cellular makeup of CPGs in various animals.

However, the study of walking has had a significant practical impact as well. For centuries, engineers have been interested in constructing walking robots. Wheeled vehicles are superior for traveling on defined pathways like railways or roads. Modern off-road vehicles can also traverse unprepared ground. But if the ground is steep or rugged, wheels lose their advantage, and legs are more effective. In this context, the rapid and agile locomotion of many insects has long served as an inspirational model for engineers as they designed legged vehicles.

Early attempts at building walking robots used animals as models only in the sense that animals have legs and it was legs that engineers were trying to emulate. However, in recent decades, engineers collaborating with biologists have taken a further step. By carefully identifying salient features of the physical structure or locomotor control system of an animal and incorporating them into a walking robot,

engineers attempt to emulate the biological system in hardware. The resulting biomimetic robots generally have performance abilities that are not common in robots designed in more conventional ways. Among walking invertebrates, insects (cockroaches, stick insects) and crustacea (crayfish) have successfully been used as models for such robots. The features of the animal that have been emulated include the physical structure of the legs (cockroaches), the neuromuscular system (cockroaches, stick insects), the neural system used to coordinate walking (stick insects), and even some of the behavioral repertoire of the animal (cockroaches, crayfish).

Although biologists and engineers have collaborated to produce some remarkable robots, it is nevertheless true that even the best of these are not able to emulate the enormous agility, speed, and flexibility exhibited by many arthropods as they walk. There is little doubt, however, that we will see continued progress in the development of biomimetic robots, and what seems like science fiction today may well be reality tomorrow.

See also: Central Pattern Generators; Central Pattern Generators: Sensory Feedback; Comparative Biology of Invertebrate Neuromuscular Junctions; Neuromuscular Transmission Modulation at Invertebrate Neuromuscular Junctions.

Further Reading

Beer RD, Chiel HJ, Quinn RD, and Ritzmann RE (1998) Biorobotic approaches to the study of motor systems. *Current Opinion in Neurobiology* 8: 777–782.

Büschges A (2005) Sensory control and organization of neural networks mediating coordination of multisegmental organs for locomotion. *Journal of Neurophysiology* 93: 1127–1135.

Cruse H, Dean J, Dürr V, Kindermann T, Schmitz J, and Schumm M (2002) A decentralized, biologically based network for autonomous control of (Hexapod) walking. In: Ayers J, Davis JL, and Rudolph A (eds.) *Neurotechnology for Biomimetic Robots*, pp. 383–400. Cambridge, MA: MIT Press.

Delcomyn F (1980) Neural basis of rhythmic behavior in animals. *Science* 210: 492–498.

Delcomyn F (1985) Walking and running. In: Kerkut GA and Gilbert LI (eds.) *Comprehensive Insect Physiology, Biochemistry and Pharmacology,* vol. 5, pp. 439–466. Oxford, UK: Pergamon Press.

Delcomyn F (2004) Insect walking and robotics. *Annual Review of Entomology* 49: 51–70.

Dickinson MH, Farley CT, Full RJ, Koehl MAR, Kram R, and Lehman S (2000) How animals move: An integrative view. *Science* 288: 100–106.

Pearson KG and Fourtner CR (1975) Nonspiking interneurons in walking system of the cockroach. *Journal of Neurophysiology* 38: 33–52.

Strauss R (2002) The central complex and the genetic dissection of locomotor behaviour. *Current Opinion in Neurobiology* 12: 633–638.

Webb B (2001) Can robots make good models of biological behaviour? *Behavioral and Brain Science* 24: 1033–1094.

Wilson DM (1966) Insect walking. *Annual Review of Entomology* 11: 103–122.

Wallerian Degeneration

S L Carroll, The University of Alabama, Birmingham, AL, USA

© 2009 Elsevier Ltd. All rights reserved.

Introduction

The term Wallerian degeneration refers to a series of stereotyped, synchronized changes that occur in the distal segment of a peripheral nerve after this structure has been separated from its associated neuronal cell bodies by trauma or, less commonly, an ischemic event. Wallerian degeneration is named after Augustus V Waller, the scientist who first described this process in 1850. Over the decades since its initial description, it has been recognized that Wallerian degeneration shares many features with axonal degeneration, a common form of axonopathy that can be produced by numerous diseases, drugs, and toxins. Given these similarities and the fact that Wallerian degeneration is initiated by a single, precisely timed event (the severing of the nerve), Wallerian degeneration has proven to be an extremely useful experimental model that has provided valuable new insights into the mechanisms underlying axonal degeneration. Below, the morphological events that occur during Wallerian degeneration, the biochemical alterations that accompany these morphological changes, and how these changes prepare the nerve for subsequent regeneration, are considered. Then, the Wallerian degeneration in the peripheral nervous system (PNS) is compared to Wallerian degeneration in the injured central nervous system (CNS), highlighting the similarities and differences in these processes. Finally, the recent studies of the Wallerian degeneration slow (*Wld*^{s}) mouse, a spontaneously arising mutant that has fundamentally altered our understanding of axonal degeneration by demonstrating that this process is actively mediated by an axon death program distinct from apoptosis, are discussed.

Morphologic Changes during Wallerian Degeneration of Peripheral Nerve

The majority of studies of Wallerian degeneration have been performed using rodents (primarily rats and mice) as experimental models. Consequently, the timing of events described below is that observed in these animal models unless otherwise indicated. Wallerian degeneration has been studied in man and the findings compared to observations made in rodent models. The primary difference noted between Wallerian degeneration in man and in rodents is that many of the morphologic alterations described below take significantly longer to develop in human nerves, a feature that may be related to their much larger size. Other than these differences in timing, Wallerian degeneration in humans, rats, and mice is highly similar.

Axonal Changes Following Nerve Transection

For a short period after nerve transection, axon segments distal to the injury site remain intact. Although separated from the cell body of the neuron, these distal axon segments also remain capable of transmitting electrical impulses for 48–96 h (4–10 days in man, with the ability to conduct current typically lasting longer in lengthier distal nerve segments). Despite this continuing electrical responsiveness, morphologic changes begin to appear in the distal axon segment quite rapidly. Within hours, membranous organelles, glycogen, and atypical swollen mitochondria, which are normally transported along the axon, begin to accumulate in paranodal regions and adjacent to the lesion site on both the proximal and distal sides. Lysosomes also begin to accumulate in the paranodal and lesional regions by 2 h after injury. The accumulation of organelles in severed axon segments adjacent to the cut site is of a magnitude sufficient to enlarge the axon, thereby stretching and thinning the adjacent myelin sheath; however, this dilatation is typically of a lesser degree than is characteristic of the axonal spheroids that develop following interruption of a CNS axon. Within 12–24 h, axonal microtubules begin to become disorganized, heralding the dissolution of this portion of the axonal cytoskeleton. One to two days post injury, calcium ions enter the axon segment via calcium-specific channels and activate calcium-dependent calpain proteases which degrade the axonal neurofilaments. As these morphologic changes, which are known as granular disintegration, progress, the normal axonal cytoskeleton, axolemma, and axonal organelles are degraded, converting the axoplasm into a watery fluid filled with amorphous and granular debris. At the time granular disintegration occurs, the myelin sheath surrounding the axon, although stretched, is still intact.

Although the sequence of morphologic changes described above is widely accepted, the spatiotemporal pattern with which the final breakdown of the axon occurs has long been controversial. In large part, this reflects the technical difficulty of reliably distinguishing individual axons within a nerve during a prolonged period of observation and the fact that when axonal degradation occurs, it occurs quite rapidly. Early studies produced contradictory results, with some investigators reporting that distal axon segments degenerate in a proximal-to-distal direction

and others reporting that degeneration occurred either distally to proximally or simultaneously throughout the entire length of the axon. It was also suggested that the type of experimental injury used affected the pattern, with crush injury reported to produce degeneration that proceeded proximally to distally and cut injury resulting in simultaneous degradation of the entire distal axon segment. These issues have recently been clarified by studies performed with transgenic mice that express either yellow fluorescent protein (YFP) or green fluorescent protein (GFP) in small subsets of axons. When nerves in these animals were transected and axonal degeneration followed by imaging of nerve segments, it was found that axons transition from a fully intact state to a fully degraded state quite rapidly, with axons identifiable in a partially degraded state for less than an hour. Sequential imaging of the partially degraded axons indicated that degeneration occurred in a distal to proximal direction after a cut injury, while axon degeneration after a crush injury proceeds proximally to distally.

Schwann Cell Changes Following Nerve Transection

In noninjured peripheral nerve, myelinated and unmyelinated Schwann cells are mitotically quiescent differentiated elements whose primary function is to facilitate axonal conduction. Following injury, however, these glia dedifferentiate, proliferate, and activate a genetic program that enhances and guides the growth of regenerating axonal sprouts through the distal nerve segment and back to their original peripheral targets. It has long been believed that these changes in the post axotomy phenotype of Schwann cells are initiated by a positive signal sent from the injured axon. At this time, however, the nature of this axon-derived signal is unknown.

In the early stages of Wallerian degeneration, myelinated Schwann cells mediate the initial breakdown of their myelin sheath and associated distal axon segments. Minutes after nerve transection, widening of Schmidt–Lanterman incisures (channels in Schwann cells that run transversely across spiraled compact myelin layers and allow these layers to communicate with one another) can be observed in myelinated Schwann cells immediately distal to the injury site; by 24–36 h after injury, this same alteration is evident in Schwann cells throughout the more distal portions of the nerve segment. As the Schmidt–Lanterman incisures open up, the cytoplasm of myelinated Schwann cells begins to swell and compress associated axons. Twenty-four to forty-eight hours after nerve transection, myelin destruction begins. By light microscopy, the initial sign that myelin destruction is underway is the fragmentation of the myelin sheath into myelin ovoids that are contained in compartments in the Schwann cell cytoplasm known as digestion chambers (Figure 1). As Wallerian degeneration proceeds, digestion chambers are often found to also contain fragments of axoplasm. In subsequent days and weeks, myelin ovoids are further fragmented and cleared as macrophages begin to assist in the clearance of myelin and axonal debris (see below). Degeneration in unmyelinated Schwann cells occurs in a similar fashion except, of course, that myelin ovoids are not seen.

During development, Schwann cells that do not establish axonal contact undergo apoptosis. However, at least during the first few weeks post axotomy, apoptosis is not evident in the population of adult Schwann cells within denervated nerve segments.

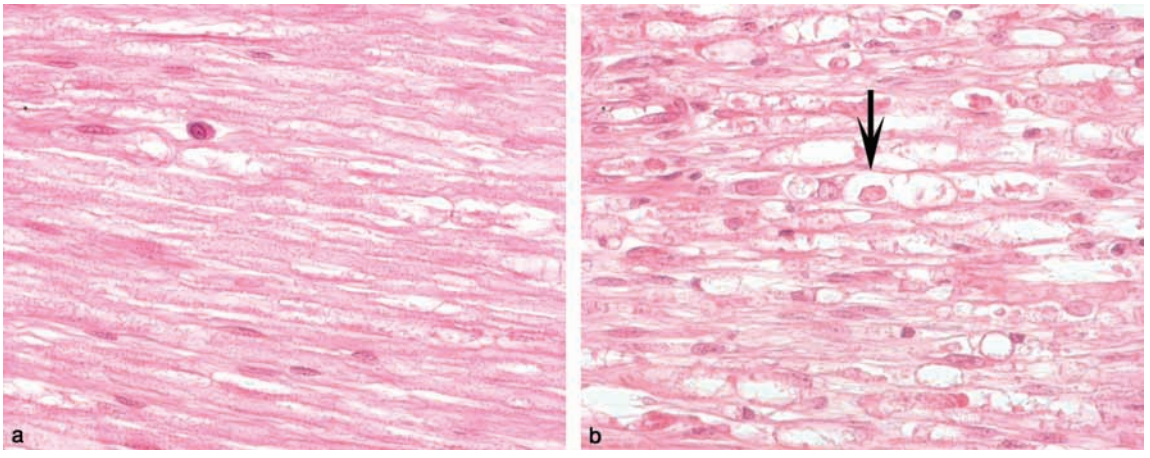

Figure 1 Morphologic changes in peripheral nerve undergoing Wallerian degeneration. (a) Noninjured rat sciatic nerve. Note the relatively low cellularity and the orderly alignment of axons and their associated myelin sheaths. (b) Rat sciatic nerve distal to a site of crush injury performed 5 days previously. Cellularity is now increased relative to the noninjured nerve illustrated in (a) as the result of macrophage infiltration and Schwann cell proliferation. The arrow indicates one of the numerous Schwann cell digestion chambers evident in this field. A myelin ovoid is present within the indicated digestion chamber (the red roughly oval mass).

Indeed, adult Schwann cells, which heretofore have been mitotically quiescent, begin to proliferate rapidly within the confines of their original basal lamina. In rodents, Schwann cell mitogenesis peaks 3–4 days post axotomy and gradually declines thereafter, with only scattered proliferating Schwann cells evident in the distal nerve segment by 18 days after injury. As a result of this Schwann cell proliferation and accompanying macrophage infiltration (see below), the distal nerve segment becomes obviously hypercellular when examined by light microscopy. The proliferating Schwann cells layer upon one another within their basal lamina to form structures known as bands of Büngner or bands of von Büngner. The bands of Büngner are critically important for subsequent axonal regeneration as they form a substrate that will provide both structural and biochemical support for newly elaborated axonal sprouts as they elongate (see below).

Macrophage Recruitment and Action Following Nerve Transection

Injuries occurring at most sites in the body elicit an acute inflammatory response involving polymorphonuclear leukocytes which is subsequently replaced by a chronic inflammatory infiltrate rich in lymphocytes. Distal nerve segments undergoing Wallerian degeneration, however, typically contain only occasional polymorphonuclear leukocytes and lymphocytes. Instead, macrophages are the primary inflammatory cell type found in degenerating peripheral nerve. Studies using chimeric rats in which the grafted marrow cells have been GFP-tagged have demonstrated that resident macrophages are normally present in the endoneurium and that these cells become activated during Wallerian degeneration. Indeed, resident tissue macrophages are more abundant in peripheral nerve than is commonly appreciated, representing an estimated 2–9% of the endoneurial cell population in noninjured nerve. The action of the activated resident macrophages is complemented by an influx of large numbers of hematogenous macrophages that are recruited into degenerating distal nerve segments by a variety of signals including degenerating myelin (likely the most important signal), complement components, adhesion molecules, and factors derived from injured axons. Macrophage infiltration is also facilitated by opening of the blood–nerve barrier; this is triggered by the release of histamine, serotonin, and other substances from resident mast cells, which proliferate *in situ* during Wallerian degeneration. Magnetic resonance imaging studies following the fate of circulating macrophages tagged with superparamagnetic iron oxide particles and analyses of chimeric mice with GFP-tagged bone marrow cells demonstrate that hematogenous macrophages begin to infiltrate distal nerve segments 1 day after injury, with infiltration peaking 4 days post axotomy and markedly declining by 8 days after injury. Macrophage infiltration begins at the lesion site and proceeds in a proximodistal direction, paralleling the breakdown of the myelin sheath.

Once in the distal nerve segment, macrophages replace Schwann cells as the primary phagocytes. Unlike Schwann cells, whose phagocytosis of myelin is lectin mediated and opsonin independent, macrophage phagocytosis of myelin debris is dependent upon their expression of complement receptor type 3 and the presence of complement component C3 on degenerating myelin sheaths. Macrophages can phagocytose myelin debris lying free in extracellular spaces or they can remove this debris directly from the Schwann cell's cytoplasm via a poorly understood mechanism. Macrophages also facilitate the clearance of myelin debris by synthesizing and secreting apolipoprotein E, which scavenges lipid products for future reutilization. After macrophages finish clearing myelin and axon debris from the distal nerve segment, they are eliminated from the nerve both by local apoptosis and by re-entering the circulation and then migrating to regional lymph nodes and the spleen.

Neuronal Regenerative Responses Following Nerve Transection

In contrast to development, where loss of neuronal contact with its peripheral target frequently results in neuronal apoptosis, adult neurons that have undergone axonal transection most commonly survive. Shortly after nerve transection, the neuronal Nissl substance (rough endoplasmic reticulum) begins to disperse (a process known as chromatolysis), heralding a fundamental shift in their metabolism as they prepare for regeneration. Then, as the bands of Büngner and their associated supportive microenvironment form, neurons begin to put out regenerative sprouts from the stump of the axonal segment still associated with the neuronal cell body and from nearby nodes of Ranvier. Neurons typically put out many more sprouts than will be needed, with up to 25 sprouts arising from a single axon. These sprouts will, if possible, extend across the gap in the injured nerve to establish contact with the bands of Büngner in the distal nerve segment. Once this contact is established, the axonal sprouts begin to elongate along either the surface of Schwann cells within the band of Büngner or along the inner surface of the basal lamina that forms the boundary of these structures. It is clear that Schwann cells play an important supportive role in this process; irradiating distal nerve segments so as to kill Schwann cells within the bands of Büngner markedly impairs axonal regeneration. On

the other hand, it is also evident that Schwann cells are not the sole element within the distal nerve segment that promotes axonal regeneration. Regenerating axons do continue to grow along the basal lamina in irradiated nerve segments, albeit with a reduced rate of extension. Furthermore, *in vivo* imaging of GFP- and YFP-tagged axons regenerating after a crush injury (a injury which minimally disrupts the basal lamina) has provided an elegant demonstration that physical cues, likely provided at least in part by the basal lamina, allow regenerating axons to precisely retrace their original pathway and reform appropriate synaptic contacts with skeletal muscle.

As axonal sprouts re-establish contact with Schwann cells, the Schwann cells respond to axonal cues by beginning anew to elaborate myelin sheaths. Since multiple axonal sprouts are initially present and elongating along the band of Büngner, this results in the formation of groups of small, thinly myelinated axons (regenerative clusters) within a single basal lamina. With time, one of these axons will become dominant and the other sprouts will regress. The mechanisms responsible for precisely matching axon-Schwann cell numbers during regeneration are not understood.

Molecular Alterations Occurring during Wallerian Degeneration of Peripheral Nerve

Although it may seem paradoxical for a process that is described as a degeneration, Wallerian degeneration is accompanied by the increased expression of a number of molecules in the distal nerve segment. Over the last three decades, a variety of techniques have been used to identify transcripts whose expression is altered by nerve injury, including direct examination of individual candidate mRNAs, subtractive and differential hybridization and, more recently, microarray analyses. When examined at the tissue level, changes detected by these techniques reflect the sum of multiple alterations in gene expression occurring in multiple cell types endogenous to the nerve such as Schwann cells, fibroblasts, resident tissue macrophages, mast cells, and perineurial cells. Many of the apparent alterations in gene expression occurring during the first 7–8 days post axotomy also reflect the increase in the number of hematogenous macrophages migrating into the distal nerve segment. However, the majority of the axotomy-responsive genes identified to date are of Schwann cell origin. Axotomy-responsive genes expressed by Schwann cells encode multiple functional classes of proteins. These proteins include factors that promote demyelination, macrophage recruitment and macrophage activation, molecules that encourage axonal regeneration such as neurotrophic factors and cell adhesion molecules, and, as remyelination begins to occur, proteins necessary for myelin structure and synthesis.

By 2 days after nerve transection, myelinating Schwann cells have markedly decreased their expression of mRNAs encoding myelin proteins such as peripheral myelin protein-22 (PMP-22), myelin protein zero (P_0), myelin basic protein, myelin-associated glycoprotein (MAG), periaxin, and proteolipid protein (PLP). Following this decrease in the expression of myelin protein transcripts, Schwann cells in the distal nerve segment dedifferentiate to a state characterized by increased expression of the low affinity neurotrophin receptor ($p75^{LNTR}$), the intermediate filament glial fibrillary acidic protein, and the cell adhesion molecules L1 and neural cell adhesion molecule (NCAM) and decreased expression of Pax3, a transcription factor that represses myelin gene expression, and the transcription factor Krox-20. The phenotype of these dedifferentiated Schwann cells is similar to that of premyelinating and nonmyelinating Schwann cells. Because of this, the phenotype of dedifferentiated Schwann cells is often equated with that of Schwann cells during development. Such comparisons are potentially hazardous. While it is true that these cell types demonstrate many similarities, it is also clear that there are some important differences between dedifferentiated adult Schwann cells and developing Schwann cells (e.g., the ability of adult Schwann cells to survive when axon contact is lost).

As Schwann cells in the distal nerve segment dedifferentiate, they also begin to elaborate a variety of cytokines capable of recruiting macrophages such as interleukin-1β (IL-1β), stromal cell-derived factor 1 (SDF-1), tumor necrosis factor-α (TNF-α), macrophage inflammatory protein-1α (MIP-1α), and monocyte chemoattractant protein-1 (MCP-1). As macrophages are recruited into the injured nerve segment, they too begin to elaborate TNFα, interleukin-1α (IL-1α), and IL-1β. Injecting neutralizing antibodies that inhibit the action of MCP-1 or MIP-1α both reduces macrophage recruitment into degenerating peripheral nerve and impairs phagocytosis of myelin debris. Although IL-1β neutralizing antibodies do not effectively inhibit macrophage recruitment, they do similarly interfere with myelin phagocytosis. The role played by other cytokines expressed in degenerating nerve is still incompletely understood.

Dedifferentiated Schwann cells in degenerated peripheral nerve also begin to express a variety of neurotrophic factors such as nerve growth factor (NGF), brain-derived neurotrophic factor (BDNF), leukemia inhibitory factor, fibroblast growth factor-2 (FGF-2), insulin-like growth factor I (IGF-I) and its receptor,

and glial cell-derived neurotrophic factor (GDNF) and its GFRα receptor subunit. These molecules are thought to provide trophic and tropic support for regenerating axons, enticing and guiding them along the bands of Büngner so that physiologic function can be restored. Dedifferentiated Schwann cells likewise begin to express increased levels of cell surface molecules such as NCAM, L1, ninjurin-1, ninjurin-2, tenascin, and N-cadherin. These cell adhesion molecules mediate intercellular interactions and thus likely play a role in guide elongating axons along the length of denervated nerve segments.

Some recent observations have provided insight into the identity of the molecules that trigger demyelination in axotomized peripheral nerve. Proteins in the neuregulin-1 (NRG-1) family of growth and differentiation factors potently regulate the proliferation, survival, migration, and differentiation of developing Schwann cells *in utero* and during early postnatal life. Schwann cell expression of class II NRG-1 isoforms (also known as glial growth factors; GGFs) and their erbB receptors (erbB2 and erbB3) is induced in injured sciatic nerve beginning at 3 days post axotomy, a time that coincides with the onset of Schwann cell proliferation induced by nerve injury, and continuing to at least 30 days post axotomy. Tyrosine phosphorylation (an indirect indication of activation) of erbB2 and erbB3 becomes evident in nerve with a similar time course. Based on these observations, it was initially suggested that increased class II NRG-1 expression and erbB receptor activation promoted Schwann cell mitogenesis in injured nerve and potentially had other effects such as triggering demyelination. Consistent with these observations, transgenic mice overexpressing the class II NRG-1 isoform GGFβ3 in myelinating Schwann cells demonstrate Schwann cell hyperplasia and develop a demyelinating hypertrophic neuropathy. While GGFβ3 clearly is promitogenic for Schwann cells in adult nerve, it may not be required for proliferation as transgenic mice with an inducible genetic ablation of erbB2 in adult Schwann cells do not demonstrate a decrease in post axotomy Schwann cell proliferation. However, it is quite clear that treating myelinated Schwann cell–neuron co-cultures with GGF induces demyelination in a concentration-dependent manner, consistent with the phenotype of GGFβ3 overexpressing mice. Intriguingly, the effects that NRG-1 exerts on Schwann cells may be distinct for different NRG-1 subfamilies. In contrast to the demyelination observed in transgenic mice overexpressing class II NRG-1 in Schwann cells, transgenic mice overexpressing class III NRG-1 in neurons develop hypermyelination rather than demyelination. In further support of the hypothesis that NRG-1 effects on myelination are subfamily specific, transgenic mice overexpressing class I NRG-1 in neurons show no abnormalities of myelination. These latter studies have focused on NRG-1 effects on myelination during development; it remains to be determined whether class III NRG-1 isoforms are critical for the re-establishment of myelination in regenerating nerve.

A Comparison of Wallerian Degeneration in the PNS and CNS

Transection of axons whose length lies entirely within the CNS also results in degeneration of the distal nerve segment. However, both the specifics of this CNS process and the ultimate outcome are clearly distinct from those seen with Wallerian degeneration in the PNS. Because of this, purists argue that it is not appropriate to refer to posttraumatic axonal degeneration in the CNS as Wallerian degeneration. In contrast to this perspective, a growing body of evidence indicates that the mechanisms underlying the active degeneration of transected axon segments (see below) are similar in the PNS and CNS. Consequently, it is increasingly clear that there are at least some important similarities between Wallerian degeneration in the PNS and in the CNS.

In both the PNS and the CNS, degeneration of the severed distal axon segment occurs before the destruction of myelin sheaths associated with the degenerating axon. However, thereafter these processes diverge. Wallerian degeneration in the CNS is slower than Wallerian degeneration in the PNS. Furthermore, oligodendrocytes, unlike Schwann cells, do not form structures equivalent to the bands of Büngner and thus fail to establish a microenvironment supportive of axonal regeneration. This latter difference clearly reflects the distinct nature of the relevant glia (and the presence of inhibitory molecules such as Nogo in the CNS) rather than differences in the PNS and CNS neuronal populations; in a classic set of experiments, Aguayo and colleagues showed that CNS neurons will extend processes readily into peripheral nerve grafts transplanted into a CNS injury site. Myelin debris is also removed more slowly during Wallerian degeneration occurring in the CNS. This may be because the macrophages responding to CNS Wallerian degeneration are not hematogenously derived and thus likely differ functionally from the macrophages responding to nerve injury. Given these differences, it seems most appropriate to equate the similarities in axonal degeneration seen in CNS and PNS Wallerian degeneration and to be extremely cautious when comparing other aspects of Wallerian degeneration in the PNS and CNS.

Recent Findings Indicating that Axonal Degeneration Results from Active Self-Destruction Rather than Simple Atrophy of Severed Axons

For decades after the initial description of Wallerian degeneration, it was thought that degeneration of distal axon segments occurred because axonal transport had been interrupted, resulting in atrophy that was secondary to the disruption of metabolic communication between the axon and the neuron cell body. This notion was challenged by the unexpected discovery of a spontaneous mouse mutant that demonstrated abnormal Wallerian degeneration in both the PNS and the CNS. In these Wallerian degeneration slow mice (*Wld*^{s} mice, previously known as C57BL/Ola mice), the distal segments of transected axons remain viable and capable of conducting action potentials for up to 3 weeks. Experiments with bone marrow chimeric and sciatic nerve chimeric mice have demonstrated that prolonged axon survival in *Wld*^{s} mice results from an alteration intrinsic to the axon itself. The *Wld*^{s} mutation is inherited as an autosomal dominant trait that maps to the distal arm of chromosome 4. An 85 kb tandem triplication present at this site encodes a chimeric nuclear protein in which the NH_2-terminal 70 amino acids of the 1173-residue-long E4b ubiquitination factor (UbE4b; also known as ubiquitin fusion degradation protein 2a (Ufd2a)) is fused in frame to the entire coding sequence of the NAD+ synthesizing enzyme nicotinamide adenylyltransferase 1 (NMNAT1). Expression of this chimeric gene under the control of a β-actin promoter confers the *Wld*^{s} phenotype on wild-type mice. Considered together, these findings indicate that axonal degeneration results from the action of an active axon death program rather simply a passive atrophic process. The localization of the UbE4b/NMNAT1 fusion protein to neuronal nuclei implies that this mutant protein acts indirectly via other as yet unknown molecules that are localized to axons.

In addition to its intriguing effects on Wallerian degeneration, the phenotype of the *Wld*^{s} mutant mouse also has important implications for the mechanisms underlying other nervous system disease states in which the axon of a diseased neuron progressively degenerates from the distal portions of the axon toward the cell body. These dying back axonal degenerations occur in a variety of medically important peripheral neuropathies (e.g., diabetic neuropathy) and have been suggested to contribute to neurodegeneration in a variety of CNS diseases such as Alzheimer's disease, Parkinson's disease, and motor neuron disease. Consistent with this latter hypothesis, dying back axonal degeneration is delayed in progressive motor neuronopathy (*pmn*) mice crossed to *Wld*^{s} mice. Similar results have been obtained by crossing *Wld*^{s} mice to gracile axonal dystrophy (*gad*) mice or mice with a null mutation of the P_0 gene (a model of hereditary motor and sensory neuropathy IB/Charcot–Marie–Tooth disease type 1B), two other models thought to involve a dying back axonopathy. The *Wld*^{s} gene also confers resistance to peripheral neuropathy induced by vincristine toxicity. The ability of the *Wld*^{s} mutation to inhibit the progress of these PNS and CNS diseases thus provides a clear mechanistic link between Wallerian degeneration and dying back axonopathies.

Several lines of evidence indicate that the axon death program is distinct from the apoptosis program typically associated with neuronal cell death. Culture systems in which the distal axons of wild-type sympathetic neurons are locally deprived of NGF results in degeneration of the distal axon while the remainder of the axon and the cell body survive. In contrast, in *Wld*^{s} sympathetic neurons deprived of NGF, axons remain viable long after the cell body has undergone an apoptotic death. These observations are consistent with the finding that inhibitors of caspase action do not prevent Wallerian degeneration and that activated caspase-3 and cleavage products resulting from caspase action are not found in transected axons. In addition, mutations that prevent apoptosis such as overexpression of the anti-apoptotic molecule Bcl2 or genetic ablation of the pro-apoptotic molecules Bax and Bak do not prevent axonal degeneration. Finally, the UbE4b/NMNAT1 fusion protein encoded by the *Wld*^{s} gene bears no structural similarity to any of the known anti-apoptotic proteins.

At present, little is known about the mechanism or the downstream molecules involved in the axon death program. Initial attempts to establish the function mediated by the UbE4b/NMNAT1 fusion protein essential for delaying Wallerian degeneration focused on the possibility that this protein modulates the action of the ubiquitin–proteasome system. Consistent with this hypothesis, inhibiting the action of the ubiquitin–proteasome system pharmacologically or by overexpression of the ubiquitin protease UBP2 has been found to delay Wallerian degeneration in cultured sympathetic neurons and in transected optic nerve *in vivo*. However, it was subsequently found that overexpressing the UbE4b fragment from the UbE4b/NMNAT1 fusion protein in dorsal root ganglion neurons was not sufficient to delay axonal degeneration following transection or vincristine treatment. In contrast, both overexpression of NMNAT1 and administration of NAD prior to injury delays axonal degeneration induced by both of these

treatments. Both NMNAT1 overexpression and pretreatment with NAD seem to act through the NAD-dependent deacetylase SIRT1 as inhibiting the action of this deacetylase by RNA interference or with a small molecular inhibitor blocks the delay in Wallerian degeneration; enhancing SIRT1 action with a specific activator similarly delays Wallerian degeneration. Of note, SIRT1, like the UbE4b/NMNAT1 fusion protein, is expressed in the nucleus. Consequently, the axonal effectors cooperating with the UbE4b/NMNAT1 fusion protein to delay Wallerian degeneration are as yet completely unknown.

See also: Enteric Nervous System: Neurotrophic Factors; Neuropathy: Peripheral; Peripheral Nerve Regeneration: An Overview; Proteasome Role in Neurodegeneration; Schwann Cells and Axon Relationship.

Further Reading

Araki T, Nagarajan R, and Milbrandt J (2001) Identification of genes induced in peripheral nerve after injury: Expression profiling and novel gene discovery. *Journal of Biological Chemistry* 276: 34131–34141.

Araki T, Saski Y, and Milbrandt J (2004) Increased nuclear NAD biosynthesis and SIRT1 activation prevent axonal degeneration. *Science* 305: 1010–1013.

Beirowski B, Adalbert R, Wagner D, et al. (2005) The progressive nature of Wallerian degeneration in wild-type and slow Wallerian degeneration (Wlds) nerves. *BMC Neuroscience* 6: 6.

Bendszus M and Stoll G (2003) Caught in the act: *In vivo* mapping of macrophage infiltration in nerve injury by magnetic resonance imaging. *Journal of Neuroscience* 23: 10892–10896.

Coleman M (2005) Axon degeneration mechanisms: Commonality amid diversity. *Nature Reviews Neuroscience* 6: 889–898.

Guertin AD, Zhang DP, Mak KS, Alberta JA, and Kim HA (2005) Microanatomy of axon/glial signaling during Wallerian degeneration. *Journal of Neuroscience* 25: 3478–3487.

Huijbregts RPH, Roth KA, Schmidt RE, and Carroll SL (2003) Hypertrophic neuropathies and malignant peripheral nerve sheath tumors in transgenic mice overexpressing glial growth factor-β3 in myelinating Schwann cells. *Journal of Neuroscience* 23: 7269–7280.

Kerschensteiner M, Schwab ME, Lichtman JW, and Misgeld T (2005) *In vivo* imaging of axonal degeneration and regeneration in the injured spinal cord. *Nature Medicine* 11: 572–577.

Kuhlmann T, Bitsch A, Stadelmann C, Siebert H, and Brück W (2001) Macrophages are eliminated from the injured peripheral nerve via local apoptosis and circulation to regional lymph nodes and the spleen. *Journal of Neuroscience* 21: 3401–3408.

Mack TGA, Reiner M, Beirowski B, et al. (2001) Wallerian degeneration of injured axons and synapses is delayed by a *Ube4b/Nmnat* chimeric gene. *Nature Neuroscience* 4: 1199–1206.

Michailov GV, Sereda MW, Brinkmann BG, et al. (2004) Axonal neuregulin-1 regulates myelin sheath thickness. *Science* 304: 700–703.

Mueller M, Leonhard C, Wacker K, et al. (2003) Macrophage response to peripheral nerve injury: The quantitative contribution of resident and hematogenous macrophages. *Laboratory Investigation* 83: 175–185.

Nguyen QT, Sanes JR, and Lichtman JW (2002) Pre-existing pathways promote precise projection patterns. *Nature Neuroscience* 5: 861–867.

Zanazzi G, Einheber S, Westreich R, et al. (2001) Glial growth factor/neuregulin inhibits Schwann cell myelination and induces demyelination. *Journal of Cell Biology* 152: 1289–1299.

Zhai Q, Wang J, Kim A, et al. (2003) Involvement of the ubiquitin–proteasome system in the early stages of Wallerian degeneration. *Neuron* 39: 217–225.

Wilson's Disease

J A Ghika, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

© 2009 Published by Elsevier Ltd.

In 1912, Kinnier Wilson described with some detail a disease previously recognized by Frerichs (1861) and Westphal (1883) characterized by various neurological manifestations, such as tremor, rigidity, spasmodic contractions, emotionalism, and mental symptoms, which he concluded were related to lesions of the lenticular nucleus. Four years later, Bramwell reported the case of seven children, four of whom died of liver failure, and proposed that the neurological manifestations of the disease could be preceded by liver dysfunctions. The disease is now referred to as Wilson's disease, or hepatolenticular degeneration, and it is known to be a disorder of copper metabolism. It is inherited as an autosomal recessive trait, lethal if not treated and reversible if treated early enough with chelation of copper. The gene responsible for the defect has recently been isolated and found to code for a copper-transporting protein referred to as ATP7B (WDNP), a P-type ATPase on chromosome 13q14.3-q21.1. The prevalence of the disease is low, between one and four cases in 200 000, and only individuals homozygous for the abnormal gene suffer from clinical manifestations, but the prevalence of heterozygotes is 1 in 100.

Pathogenesis

One of the major alterations observed in this disease is an accumulation of copper in the body tissues. Copper, which is a component of an ordinary diet (2–5 mg day^{-1}), is, after Zn and Fe, the third most important trace element in the body with a mean amount of 75–150 mg. Seafood, nuts, mushrooms, cereals, and liver are the main sources. It is found essentially in muscles, bones, liver, and pigmented nuclei of the brain, mostly in glial cells, where it is used as a cofactor of different mitochondrial, cytoplasmic, or nuclear metalloenzymes (cytochrome oxidase, superoxide dismutase, tyrosinase, dopamine beta hydroxylase, monoamine oxidase), but it also plays a role in the synthesis of hemoglobin and melanin, absorption of iron, energy production, collagen and elastin bridges, and mitochondrial function. In the brain, the highest concentrations of copper are found in locus coeruleus and substantia nigra, where copper is associated with dopamine beta-hydroxylase, the enzyme that catalyzes the conversion of dopamine to noradrenaline, followed by the striatum, pons, thalamus, cerebellum, and spinal cord. Over 90% of the copper absorbed in the intestinal epithelial cell under the regulation of hepatic ATP 7B, bound to metallothionein, is transported in the blood associated with ceruloplasmin, a 151 kDa protein synthesized by the liver, that binds up to six atoms per molecule. Other copper-binding proteins include albumin, apoceruloplasmin, and cuprophilin. Copper is removed from the blood by hepatic uptake and then partially excreted in the bile. Normal plasma concentrations are 0.05–1.10 mg dl^{-1}. In Wilson's disease, copper accumulates at toxic levels by deficient intracellular access and biliary and urinary excretion in various organs.

Recent progress achieved by combining genetic analyses and molecular biology techniques has made it possible to identify and clone the gene responsible for the disease, located between markers q14 and q21 on chromosome 13. Interestingly, the DNA sequence of this gene was found to have marked homologies with that responsible for Menkes disease, another recessive but X-linked disturbance of copper metabolism associated with abnormal hair and severe, progressive neurodegeneration. The symptoms observed in Menkes disease have been attributed to a lack of important copper enzymes, probably because of copper deficiency in specific intracellular organelles. It must be noted, however, that copper deficiency in Menkes disease occurs despite excessive accumulation of the metal in various tissues. In both diseases, the affected genes appear to code for copper-transporting proteins referred to as ATP7A (for Menkes disease) and ATP7B (for Wilson's disease), which are P-type ATPases with metal-binding regions similar to those found in prokaryotic heavy metal transporters. Many patients suffering from Wilson's disease or Menkes disease have now been reported to have deletions or mutations within the coding region of these homologous genes. More than 300 different mutations of the gene ATP7B have so far been described, and 80 polymorphisms and the mutations that affect more severely the gene or protein are also responsible for earlier and more dramatic hepatic disturbances. Based on these new findings, Wilson's disease can be attributed to the expression in liver, but probably other organs as well, of a defective transporter for copper. This results in abnormal metabolism of copper by the hepatocyte and reduced excretion in the bile. It is also interesting to note that a usual finding in patients suffering from Wilson's disease is a reduced level in serum ceruloplasmin, a finding that remains unexplained.

In Wilson's disease, hepatic copper accumulation starts very early during the first years of life. This

process begins asymptomatically until hepatic copper-binding sites are saturated. Then hepatic insufficiency and hepatocellular necrosis may develop, and copper accumulates in other tissues. The most affected organs are the brain, erythrocytes, and kidneys. Copper deposition also usually occurs in the Descemet's membrane of the cornea, forming a golden-brown pigment ring known as the Kayser–Fleischer ring. This constitutes one of the most pathognomonic signs of the disease.

The clinical manifestations of Wilson's disease seem to be directly related to a toxic effect of copper in the different regions where it accumulates. If copper is removed, a recovery from both the hepatic and neurological signs results. The mechanisms of copper toxicity are, however, not yet clearly understood. In the brain, lesions predominate in the basal ganglia, although they may also be seen in the subthalamic nucleus, thalamus, claustrum, myelinated fiber bundles, or even brain stem, spinal cord, and cerebral cortex. An enlargement of ventricles is usual, as well as an atrophy of cerebral cortex. Histological findings include an increase in the size and number of protoplasmic astrocytes, many of which become multinucleated, and a degeneration of neurons and glial cells, a process which may lead to cavitations. Cell death is not restricted to a specific type of neurons, but it is more important in basal ganglia than in cerebral cortex. Vascular changes may also be seen, consisting of capillary swelling and proliferation. These lesions are not specific for Wilson's disease and may also be seen in cases of hepatic encephalopathy. Thus, they probably reflect general suffering of neurons, and their distribution might be related to regional differences in vulnerability. Copper toxicity could result from different mechanisms, including severe mitochondrial dysfunction, production of free radicals, and increased oxidation, which all could affect energy metabolism and membrane integrity.

Clinical Picture

Wilson's disease generally starts during the second decade (range between 6 and 35 years; cases over age 40 are extremely rare), as a liver disease (50–60%) or with neurological symptoms (30%), and sometimes with both. Patients may present with other systemic manifestations such as osteoarticular (10%; osteoarthropathy, osteomalacia), cardiac (1–2%; cardiomyopathy), ophthalmological (2%), or hematological (2%; hemolytic anemia) symptoms. Isolated cases starting with endocrinological (hypoparathyroidism, pancreatitis) or renal (Fanconi syndrome) symptoms have been reported. Asymptomatic cases can be found by serendipity, but fulminant hepatocerebral forms, generally with hemolytic anemia, are most dramatic. Genotype–phenotype correlations start to be reported in some detail, but nonsense and frameshift mutations seems to be associated with an earlier and more severe course of disease.

The liver dysfunction in patients with Wilson's disease is either an asymptomatic high level of transaminases or a chronic, recurrent hepatitis with icterus, nausea, vomiting, and asthenia which may rapidly progress to cirrhosis with hepatic failure and portal hypertension. At this stage of the disease, copper deposition will occur in extrahepatic organs.

Neurological symptoms are essentially an atypical cortico-bulbar syndrome and dystonic syndrome with important dysarthria, potentially dangerous dysphagia, hoarseness of the voice, drooling, emotional lability, and a strange 'vacuous or sardonic smile.' Other symptoms include slowness of distal motility, corticospinal signs, parkinsonism (bradykinesia, rigidity, posture, and gait abnormalities), tremor or involuntary movements (choreo-athetosis, dystonia, or exceptionally myoclonus), and, rarely, epilepsy. Psychiatric symptoms are very frequent. In 10–15% of the cases they precede the neurological symptoms and are essentially personality changes, mood disorders or behavioral changes (aggressivity, disorders of impulse control, asocial behavior, depression, suicide, anxiety, emotional lability, bipolar mood disorders, mania, obsessive-compulsive disorders, apathy, mental anorexia, social withdrawal, psychosis such as paranoia, delusions, hallucinations, catatonia, or confusion), or sleep disorders. Cognitive deficits (slowing of cognitive processes and dysexecutive syndrome) up to dementia or loss of school performance are exceptional. Supranuclear oculomotor disturbances (slow saccades, gaze distractibility, apraxia of lid opening, recurrent ophthalmoplegia, accommodation, or acute blindness) are sometimes found. Brown-yellow Kayser–Fleischer rings at the corneal limbus are present in almost 100% of the cases with neurological symptoms. Sometimes, however, they are only unilateral or can only be seen with a slit lamp. Sunflower cataract can also be found.

Due to the variability of the clinical picture, diagnosis is sometimes difficult and usually delayed. It can be confirmed by decreased serum ceruloplasmin ($<20\,\mathrm{mg\,dl^{-1}}$) and copper levels ($<80\,\mu\mathrm{g\,dl^{-1}}$), high 24 h copper urine excretion ($>100\,\mu\mathrm{g\,24\,h^{-1}}$), high transaminase levels, and slit lamp eye examination, but the gold standard remains the quantitative determination of copper in liver tissue from biopsy by mass spectrometry with values between 250 and 3000 $\mu\mathrm{g\,g^{-1}}$ dry tissue. An atypical form with apoceruloplasmin deficit has been described. Glucosuria, aminoaciduria, phosphaturia, and bicarbonaturia (Fanconi syndrome) may also be present. High levels of copper in hepatic

biopsy can be found. Brain atrophy and hypodensities or low signals are usually found in the basal ganglia on computed tomography (CT) scans. Studies carried out using magnetic resonance imaging (MRI) have typically reported high T2 weighted signals in the basal ganglia (caudate and lenticular nuclei, thalamus), but also in the hemispheric and cerebellar white matter. Abnormal high intensity signals have also recently been found in the midbrain ('face of the giant panda' sign with high intensity signals in the tegmentum, but not in red nucleus, preservation of signal intensity in the lateral portions of the substantia nigra and hypodensities in the superior colliculus, pons, and dentate nucleus). Definitive diagnosis of the disease as well as prenatal diagnosis can now be made by genetic analyses.

Treatment and Perspectives

The major objective of treatment is to reduce the accumulation of copper in liver and other tissues. This may be achieved in two ways: by using chelating agents, such as D-penicillamine or trientine, which remove copper from tissues, or by inhibiting the absorption of copper from the intestinal tract using potassium sulfide, zinc salts, tetrathiomolybdate, or a low-copper diet (no liver, cacao, chocolate, nuts, mushrooms, shellfish). Hepatic graft, hemofiltration, and plasmapheresis are reserved for fulminant or end stage presentations.

The first method is most commonly used. Treatment usually results, after an initial worsening of the symptoms, in a significant improvement of both hepatic and neurological manifestations. Cases of successful pregnancies under treatment have been reported. Occasionally recovery of organic lesions may be observed. Kayser–Fleischer rings as well as hypodensities or hyposignals in the basal ganglia will eventually disappear on CT scans. Improvement of high signal intensities seen on MRI scans as well as of striatal dopamine D2-receptor binding has also recently been reported following treatment. Techniques such as MRI, positron emission tomography, and recording of sensory evoked potentials are thus very helpful not only for the diagnosis of the disease, but also to assess the extent of structural alterations of the brain and the efficacy of treatment. Another interesting aspect concerns results obtained with positron emission tomography using 2-deoxy-2-fluoro-D-glucose or F-6-fluorodopa, which may show metabolic dysfunctions before lesions are seen on CT or MRI scans. This will certainly help clarify the sometimes poor correlations observed between clinical pictures and structural alterations.

The very important progress achieved in the last years with the identification of the Wilson's disease gene has opened new perspectives concerning the introduction of new therapies for this disease, among which gene therapy will probably represent an important goal.

Wilson's disease can be confounded clinically with hereditary hypoceruloplasminemia, paraneoplastic copper accumulation, acquired hepatocerebral degeneration (parenteral nutrition, hepatopathies, hemochromatosis, Nieman–Pick C disease, sea-blue histiocyte or mitochondrial diseases, schizophrenia, Reye syndrome, pantothenate kinase-associated neurodegeneration (PKAN) or other generalized dystonias, cerebellar degeneration, Huntington's disease, or atypical Tourette syndrome).

See also: Basal Ganglia: Functional Models of Normal and Disease States; Cognitive Dysfunction in Psychiatric Disorders; Friedreich Ataxia.

Further Reading

Bull PC, Thomas GR, Rommens JM, Forbes JR, and Cox DW (1993) The Wilson disease gene is a putative copper transporting P-type ATPase similar to Menkes gene. *Nature Genetics* 5: 327–337.

Duchen LW and Jacobs JM (1992) Familial hepatolenticular degeneration (Wilson's disease). In: Adams JH and Duchen LW (eds.) *Greenfield's Neuropathology*, pp. 838–841. London: Edward Arnolds.

Ferenci P (2004) Pathophysiology and clinical features of Wilson disease. *Metabolic Brain Disease* 19(3–4): 229–239.

Ferenci P (2004) Review article: Diagnosis and current therapy of Wilson's disease. *Alimentary Pharmacology and Therapeutics* 19(2): 157–165.

Ferenci P, Caca K, Loudianos G, et al. (2003) Diagnosis and phentoypic classification of Wilson disease. *Liver International* 23(3): 139–142.

Gromadzka G, Schmidt HH, Genschel J, et al. (2005) Frameshift and nonsense mutations in the gene for ATPase7B are associated with severe impairment of copper metabolism and with an early clinical manifestation of Wilson's disease. *Clinical Genetics* 68(6): 524–532.

Mercer JFB (2001) The molecular basis of cupper-transport diseases. *Trends in Molecular Medicine* 7: 64–69.

Patten BM (1988) Wilson's disease. In: Jankovic J and Tolosa E (eds.) *Parkinson's Disease and Movement Disorders*. Baltimore and Munich: Urban and Schwartzenberg.

Prayer L, Wimberger D, Krauer J, Grimm G, Oder W, and Imhof H (1990) Cranial MRI in Wilson's disease. *Neuroradiology* 32: 211–214.

Stemmle W, Meyerrose KW, Niederau C, Hefter H, Kreutzpainter G, and Strohmeyer G (1991) Wilson's disease: Clinical presentation, treatment and survival. *Annals of Internal Medicine* 115: 720–726.

Tanzi RE, Petrukhin K, Chernov I, et al. (1993) The Wilson disease gene is a copper transporting ATPase with homology to the Menkes disease gene. *Nature Genetics* 5: 344–350.

Thomas GR, Forbes JR, Roberts EA, Walshe JM, and Cox DW (1995) The Wilson disease gene: Spectrum of mutations and their consequences. *Nature Genetics* 9: 210–217.

Walshe JM (1988) Wilson's disease: Yesterday, today, and tomorrow. *Movement Disorders* 3: 10–29.

Wnt Pathway and Neural Patterning

R M Twyman, University of York, York, UK

© 2009 Published by Elsevier Ltd.

Introduction

The Wnt pathway is a complex signal transduction pathway triggered by secreted ligands of the Wnt family, named after its two original members, the products of the *wingless* gene from *Drosophila melanogaster* and the mammalian gene *Int-1* (now known as *Wnt-1*). Wnt signaling is involved in many different embryonic processes, including the fundamental pattern-forming stages that establish the nervous system in insects and vertebrates. Many of the components of the these developmental pathways were first identified in *Drosophila* and in the African clawed frog *Xenopus laevis* and have since been extrapolated to other organisms, including mammals. Wnt signaling is also involved in many later processes in the nervous system, including synaptic specialization, microtubule dynamics, synaptic protein organization, modulating synaptic efficacy, and regulating neuronal gene expression.

There are at least 20 distinct Wnt genes in vertebrates, organized into 12 conserved subfamilies. Not all vertebrates have orthologs of all members of the Wnt family. At the time of writing, 19 Wnt genes have been found in humans and mice and 16 in *Xenopus*. Only six of the vertebrate Wnt subfamilies have counterparts in *Drosophila*, while at least 11 are present in sea anemones, indicating selective evolutionary loss of different Wnt classes during the branching of the animal lineage.

The vertebrate Wnt proteins can be grouped by biological activity using functional assays. The overexpression of some Wnt proteins induces secondary axis formation in early *Xenopus* embryos and transforms C57MG mammary epithelial cells, while other Wnt proteins do not act in this way and can even antagonize the transforming Wnt proteins, hinting at the existence of different, competing signaling pathways. Currently, five different pathways are known to be activated by Wnt proteins: a canonical Wnt/β-catenin cascade, a divergent canonical pathway involved in synapse modeling and axon growth, the noncanonical planar cell polarity (PCP) pathway, the Ca^{2+} pathway (which is involved in the control of cell migration), and a microtubule-dependent pathway.

Secretion of the Wnt Protein

Wnt proteins are secreted from the cell and therefore carry an N-terminal signal sequence targeting them for cotranslational import into the endoplasmic reticulum. During their journey through the secretory pathway, they undergo several forms of posttranslational modification, including N-linked glycosylation and palmitoylation of specific cysteine residues. Not all Wnt proteins undergo both types of modification, but when palmitoylation occurs the added lipid moiety appears essential for Wnt function. The presence of a lipid group also makes the Wnt protein hydrophobic and insoluble, which contributes to the difficulty in purifying many Wnt proteins from cell culture media. In *Drosophila*, the product of the *porcupine* gene is required for Wnt palmitoylation and has been identified as an acyltransferase. Another gene, *wntless*, encodes a transmembrane protein which co-localizes with the Wingless protein inside the cell and is required for trafficking through the Golgi apparatus. Vertebrate orthologs of both proteins have been identified.

One of the main characteristics of Wnt proteins is their ability to function as developmental morphogens – that is, molecules that establish long-range concentration gradients that enable them to influence cell fates differently, according to the cell's position along the gradient. It has been unclear how Wnt proteins achieve such a gradient given their poor solubility, although it has been proposed that the proteins remain tethered to the plasma membrane, or to intercellular transport vesicles or lipoprotein particles. Alternatively, Wnt proteins may travel along cytonemes, which are long, thin filopodial processes that can bridge several cells. A further possibility is that extracellular heparan sulfate proteoglycans (HSPGs) may be involved in the transport of Wnt proteins and the establishment of morphogen gradients. As soon as they are secreted, Wnt proteins interact with glycosaminoglycans in the extracellular matrix and bind tightly to the cell surface, another factor which makes them difficult to isolate in cell culture. In *Drosophila*, the Wingless protein is found in specialized membrane vesicles called argosomes that are thought to be derived from lipid raft microdomains, and the incorporation of the protein into these vesicles requires HSPGs. Mutations in genes such as *dally*, which encodes components of this HSPG system, cause phenotypes similar to those resulting from mutations in *wingless*. Six HSPGs have been identified in vertebrates, and mouse knockouts confirm that Wnt signaling is disrupted in such animals.

Canonical Pathway – Wnt Receptors and Alternative Ligands

Wnt proteins bind to the cysteine-rich extracellular domain of seven-pass transmembrane receptors of the

Frizzled (Fz) family, named after the founder member identified in *Drosophila*. There are at least four Fz family genes in *Drosophila* and at least ten in vertebrates, but the complexity of signaling is much greater since different Wnt proteins can interact with more than one receptor and different Fz proteins can bind more than one ligand. Promiscuous as it is, the interaction between Wnt and Fz requires a number of additional factors, including a membrane-bound protein encoded by the *arrow* gene in *Drosophila*, and represented by two homologous lipoprotein receptor-related proteins (LRPs) called LRP5 and LRP6 in vertebrates. Wnt proteins can bind directly to LRPs and there is evidence that Wnt, Fz, and LRP form a ternary complex. Delivery of the Arrow protein depends on the presence of a molecular chaperone encoded by the *boca* gene, and in vertebrates this function is fulfilled by the homologous mesoderm development (Mesd) protein. Mutations in *arrow/Lrp* genes and in *boca/Mesd* result in phenotypes similar to those of *wingless/wnt* mutants.

The Fz/LRP receptor complex interacts not only with Wnt proteins, but also with alternative ligands that can act as agonists or antagonists to Wnt signaling. An example is Norrin, which binds to Fz4/LRP5. This is a cysteine knot protein identified through investigation of the human developmental disorder Norrie disease, which is characterized by major eye vascular defects. Another example is the family of proteins known as R-spondins. In *Xenopus*, R-spondin-2 is a Wnt agonist that synergizes with Wnt signaling to activate β-catenin in muscle development. In contrast, proteins of the Dickkopf family inhibit Wnt signaling by sequestering LRP5/LRP6 and then cross-linking them to an unrelated class of transmembrane proteins known as Kremens, thus promoting internalization. The sclerosteosis-associated SOST/sclerostin family of Wnt antagonists also acts by sequestering LRP5/LRP6. The soluble Frizzled-related proteins (SFRPs) are also Wnt antagonists, but these work by sequestering the Wnt proteins. This is possible because the SFRPs contain cysteine-rich domains that mimic the genuine Fz protein. They may either bind Wnt proteins in isolation, or form inactive complexes along with Fz. However, certain combinations can promote rather than inhibit Wnt signaling, perhaps by stabilizing the Wnt signal in a manner that preserves its ability to interact productively with Fz/LRP.

Canonical Pathway – Events in the Cytosol

The formation of a stable Wnt/Fz/LRP complex facilitates the phosphorylation and activation of a cytosolic protein called Dsh (from the *Drosophila* gene *disheveled*), which interacts with the cytosolic face of the complex. The activated Dsh protein is then able to disrupt a so-called destruction complex comprising three additional proteins: glycogen synthase kinase-3 (GSK-3), axin, and adenomatous polyposis coli (APC) protein. In the absence of activated Dsh, these three proteins would normally phosphorylate β-catenin, leading to its ubiquitinylation and proteasomal degradation. When the destruction complex is disassembled, β-catenin is stabilized and some of it can be imported into the nucleus, where it interacts with transcription factors of the T cell factor/lymphoid enhancer factor (TCF/LEF) family to activate the transcription of target genes (Figure 1).

The receptor complex Wnt/Fz/LRP interacts not only with Dsh but also with axin, one of the components of the destruction complex. Axin is the scaffold for the destruction complex, so all of the other components bind directly to it. The interaction between axin and the cytoplasmic tail of LRP is mediated by direct contacts with phosphorylated residues. Axin will not interact productively with LRP unless the tail is phosphorylated on multiple serine/threonine residues, and phosphorylation occurs only when Wnt binds to the receptor complex. The phosphorylation of these residues is carried out by GSK-3 and another kinase called caseine kinase I (CKI), which is also anchored to the

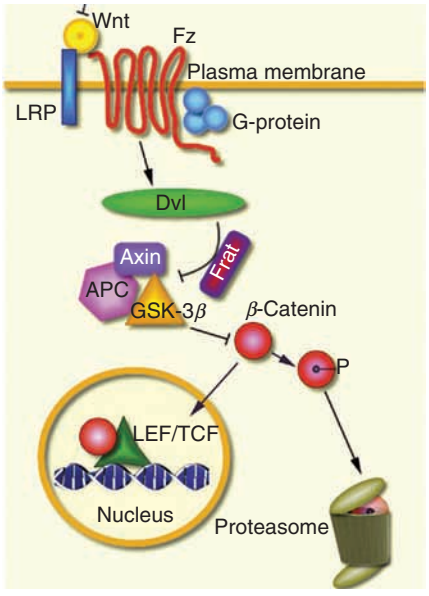

Figure 1 The canonical Wnt signaling pathway involving the destruction complex of axin, adenomatous polyposis coli (APC) protein, and glycogen synthase kinase-3β (GSK-3β), and the nuclear import of β-catenin. Fz, Frizzled; LRP, lipoprotein receptor-related protein; Dvl, Disheveled; Frat, 'frequently rearranged in advanced T-cell lymphoma'; LEF/TCF, lymphoid enhancer factor/T cell factor. Reproduced from Speese SD and Budnik V (2007) Wnts: Up- and-coming at the synapse. *Trends in Neuroscience* 30: 268–275, with permission from Elsevier.

plasma membrane. These two kinases phosphorylate different sets of serine/threonine residues on the LRP tail and both are required for transduction of the Wnt signal. The remaining component of the complex, APC protein, is an essential protein, but its precise role is unclear. It has been proposed that APC protein is required for efficient shuttling and loading/unloading of β-catenin onto the cytoplasmic destruction complex. Both APC protein and axin can be phosphorylated by their associated kinases, changing their affinity for other components of the destruction complex.

In the absence of Wnt, β-catenin is phosphorylated by CKI and GSK-3, and in its phosphorylated state it can be recognized by a transducin repeat-containing protein (β-TrCP), a component of a dedicated E3 ubiquitin ligase complex. The β-catenin protein is thus ubiquitinated and destroyed by the proteasome. In the presence of Wnt, the kinase activity of CKI and GSK-3 is inhibited, and β-catenin is able to translocate into the nucleus. Certain protein phosphatases may antagonize CKI and GSK-3, thereby promoting β-catenin stability.

Canonical Pathway – Events in the Nucleus

It is presently unclear how β-catenin is imported into the nucleus, although the process is dependent on a nuclear localization signal and β-catenin has been shown to interact with nuclear pore components. It is possible that β-catenin may shuttle from nucleus to cytoplasm in concert with axin and/or APC protein, and may be retained in the nucleus by associating with anchor proteins such as Pygopus. The role of β-catenin in the nucleus is to interact with transcription factors in the TCF/LEF family and prevent the formation of a repressive complex with proteins of the Groucho family via physical displacement. The TCF/LEF family is represented by a single protein in *Drosophila* and four paralogs in vertebrates. The replacement of Groucho with β-catenin facilitates TCF/LEF binding to the minor groove of DNA at highly conserved target sites characterized by the presence of purines on one strand and pyrimidines on the other. The binding results in the introduction of a tight kink into the DNA backbone, and allows β-catenin to bind to chromatin components such as Brahma-related gene 1 (Brg-1) protein – part of the mating-type switch/sucrose nonfermenting (SWI/SNF) chromatin remodeling complex – and histone acetylase cAMP response element-binding (CREB)-binding protein (CBP), to promote transcriptional initiation. TCF/LEF is also regulated via other pathways, and phosphorylation via the activation of mitogen-activated protein (MAP) kinase reduces its affinity for β-catenin.

The Canonical Pathway in *Xenopus* Axis Specification

Wnt signaling is critical for the establishment of dorsoventral polarity in the early *Xenopus* embryo and for laying down of the primitive anteroposterior neuraxis. The dorsal side of the embryo gives rise to a structure called the organizer, through which cells migrate to establish the anteroposterior axis. In 1989, McMahon and Moon injected mouse *Wnt-1* mRNA into the ventral blastomeres of a four-cell *Xenopus* blastula and generated an embryo with two organizers, which developed into tadpoles with a duplicated body axis. Axis duplication was also induced by Dsh, β-catenin, and a dominant-negative version of GSK-3.

In normal *Xenopus* development, the Wnt pathway is activated opposite the site of sperm entry in what becomes the future dorsal side of the embryo following cortical rotation after fertilization. It is thought that maternal Dsh protein, initially located in the vegetal region of the egg, is translocated by cortical rotation to a discrete zone of the fertilized egg, where it stabilizes β-catenin. As the embryo undergoes cleavage, cells incorporating this β-catenin-enriched cytosol activate specific genes, leading to the formation of a structure called the Nieuwkoop center, which induces the formation of the organizer in the overlying mesoderm. The ventral injection of Wnt, Dsh, β-catenin, or an inhibitory form of GSK-3 artificially activates the same genes, resulting in the formation of a duplicate organizer.

Wnt signaling is also important for the establishment of anteroposterior polarity in the neuroectoderm. The organizer expresses several Wnt antagonists, such as those encoded by *dickkopf-1*, *cerberus*, and *frzb-1*, while several Wnt proteins are expressed in the remainder of the embryo. Injection of mRNA for the antagonists leads to enlarged head development, whereas injection of wnt mRNA at this developmental stage inhibits neural induction and head induction. Although long elusive, the specific Wnt signal that triggers axis induction in *Xenopus* was identified as Wnt11 in 2005.

Wnt Signaling in Synapse Development – Divergent Pathways and Diverse Mechanisms

A number of Wnt proteins and their downstream signaling components are expressed in the developing synapse, and investigations in *Drosophila*, *Xenopus*, and mammals have revealed a previously unknown and surprisingly diverse range of signaling pathways activated by Wnt proteins. To further add to the

complexity, it appears that the same Wnt proteins can activate different downstream pathways in different parts of the central nervous system (CNS) and that different Wnt proteins can activate the same pathway in a given region of the CNS, providing scope for functional redundancy. Perhaps most exciting of all, the prominent roles of Wnt proteins in synapse formation appear to involve retrograde signaling, but there is also evidence that Wnt proteins can act as anterograde or autocrine signals.

In the mammalian CNS, various Wnt proteins are expressed during synapse development, including Wnt-3 in Purkinje cells and motor neurons, and Wnt-7a in cerebellar granule cells. These Wnt proteins can regulate axonal remodeling and presynaptic differentiation retrogradely, and do so by influencing the architecture of the cytoskeleton. For example, the addition of Wnt-7a proteins to cerebellar granule cells in culture can increase axonal branching and the clustering of synaptic vesicle proteins, and *wnt7a* knockout mice show transiently decreased axonal complexity and limited synapsin I levels. Wnt-3 is also expressed in motor neurons during synapse formation with sensory neurons, and increasing the levels of this protein results in growth cone enlargement, increased axonal branching, and increased clustering of synapsin I.

Investigation of the role of Wnt proteins in axonal remodeling has shown that a divergent form of the canonical pathway is involved (Figure 2). Some of the components are familiar (GSK-3, the Disheveled family protein Dvl1, and axin), while a significant role is also played by the microtubule-associated protein MAP1B. The phosphorylation of MAP1B by GSK-3 alters its affinity for microtubules and facilitates the regulation of microtubule dynamics. Wnt signaling reduces the availability of phosphorylated MAP1B and increases microtubule stability, a process mimicked by the ectopic expression of Dvl1, which is also known to be physically associated with microtubules. Therefore, it appears that Dvl1 antagonizes the effect of GSK-mediated destabilization. These effects are independent of β-catenin (i.e., the divergent canonical pathway acts directly upon the structural elements of the axons without requiring new gene expression).

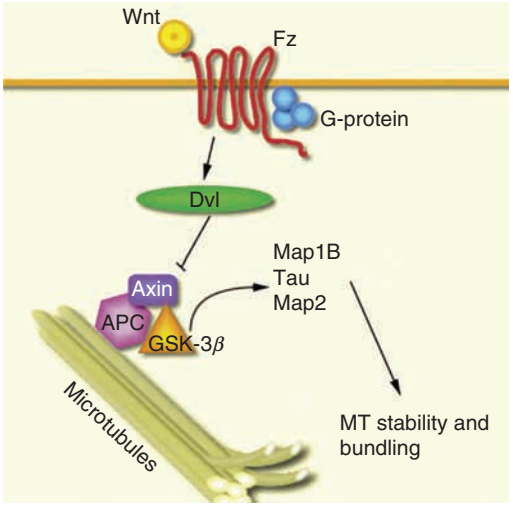

Figure 2 The divergent canonical pathway with no β-catenin and regulation of microtubule (MT) organization by the destruction complex through phosphorylation of microtubule-associated proteins 1 B (MAP1B). Fz, Frizzled; Dvl, Disheveled; APC, adenomatous polyposis coli protein; GSK-3β, glycogen synthase kinase-3β. Reproduced from Speese SD and Budnik V (2007) Wnts: Up- and-coming at the synapse. *Trends in Neuroscience* 30: 268–275, with permission from Elsevier.

A further manifestation of the Wnt signaling pathway has been identified in the hippocampus, where Wnt-7b is expressed during dendrite maturation. Experiments have shown that Wnt and Dvl1 can each increase dendritic complexity, and *dvl1* knockout mice have fewer dendritic branches. The key difference between this and the canonical pathways described earlier is that the pathways operate independently of both β-catenin and GSK-3. Instead, the Wnt/Fz/LRP complex acts through a c-Jun N-terminal kinase (JNK) signaling protein to affect cytoskeletal architecture, using the small GTPase Rac, and to regulate both actin and microtubule dynamics, in a manner similar to the planar cell polarity signaling pathway in *Drosophila* (Figure 3(a)).

Although Wnt proteins appear to function as retrograde signals to regulate the differentiation of the presynaptic compartment, there is also evidence that they might operate as anterograde or even an autocrine signals to modulate postsynaptic differentiation. Studies of the *Drosophila* neuromuscular junction show that Wingless is secreted by presynaptic boutons, and the receptor Fz2 is localized in both presynaptic and postsynaptic compartments. The release of Wingless from synaptic boutons initiates an anterograde signaling cascade, which involves yet another divergent signaling pathway in which the receptor is internalized, cleaved, and trandslocated into the nucleus in a manner dependent on the *Drosophila* homolog of the mammalian glutamate receptor-interacting protein (dGRIP), well known for its role in α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor and ephrin B receptor trafficking in mammals. Postsynaptic disruption of the so-called Frizzled nuclear import (FNI) pathway (Figure 3(b)) alters the development of both presynaptic and postsynaptic specializations, suggesting that anterograde Wingless signaling might trigger

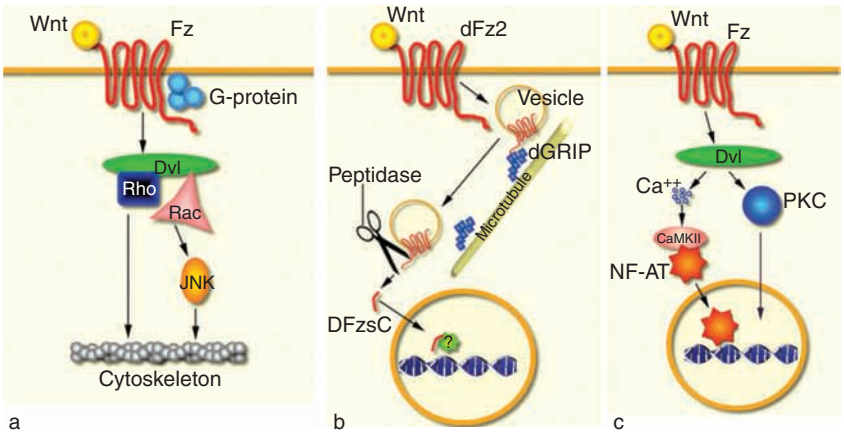

Figure 3 Three β-catenin-independent pathways: (a) planar cell polarity, (b) Frizzled nuclear import, and (c) Ca^{2+}. Fz, Frizzled; Dvl, Disheveled; JNK, c-Jun N-terminal kinase; dGRIP, *Drosophila* homolog of the mammalian glutamate receptor-interacting protein; PKC, protein kinase C; CaMKII, Ca^{2+}/calmodulin-dependent protein kinase II; NF-AT, nuclear factor of activated T cells. Reproduced from Speese SD and Budnik V (2007) Wnts: Up- and-coming at the synapse. *Trends in Neuroscience* 30: 268–275, with permission from Elsevier.

further retrograde signals or even an autocrine loop to regulate presynaptic differentiation. Similar roles for Wnt proteins in postsynaptic development have been identified in vertebrates; for example, the regulation of acetylcholine receptor clustering by agrin and by muscle-specific receptor tyrosine kinase (MuSK) has been identified, the latter of which has now been shown to interact with Dvl1, linking this regulation to the Wnt pathway. However, since agrin has no effect on β-catenin accumulation, it appears that the canonical pathway is not triggered.

Wnt Signaling and Ca^{2+}

A further β-catenin-independent Wnt signaling pathway has been identified that modulates cell movements (e.g., during embryonic gastrulation), and it also plays a role in the definition of cell fate. In this pathway, the binding of Wnt to Fz/LRP leads to an increase in intracellular Ca^{2+} concentration and nuclear import of the transcription factor NF-AT (nuclear factor of activated T cells; Figure 3(c)). The mechanism depends on the ability of specific combinations of Wnt and Fz proteins to activate Ca^{2+}/calmodulin-dependent kinase II (CaMKII) and protein kinase C (PKC). Some Fz proteins also have the ability to activate phospholipase C (PLC) and phosphodiesterase (PDE), acting through heterotrimeric GTP-binding proteins. The ability of different Wnt ligands to activate the β-catenin and Ca^{2+} pathways may reflect the preferences of different Fz proteins (e.g., rat Fz1 does not increase Ca^{2+} release or stimulate CaMKII or PKC in zebra fish embryos, whereas Fz2 does) and preferences for different coreceptors (e.g., LRP-5/LRP6 stimulates the β-catenin pathway, whereas alternative coreceptors such as Knypek and Ror2 stimulate the Ca^{2+} pathway).

Summary

Wnt signaling plays a pivotal role in the development and function of the nervous system, being required to establish the major body axes and compartment polarity in vertebrate and insect embryos, orchestrating cell migration, polarity, and fate, and controlling synaptic bouton development, axonal growth cone remodeling, and dendrite maturation, as well as a host of other processes. The canonical signaling pathway (Figure 1) involves the protection and subsequent nuclear import of β-catenin (Armadillo in *Drosophila*), although there is a divergent pathway, using many of the same components, that is β-catenin independent and acts directly on the organization of microtubules (Figure 2). There are also at least three pathways which involve Ca^{2+} as a second messenger, one involving the activation of CaMKII and protein kinase C, one involving the recruitment of heterotrimeric GTP-binding proteins to activate phospholipase C and phosphodiesterase, and, finally, one involving the planar cell polarity pathway, which signals through JNK (Figure 3). The particular pathways activated in any given cell, and their consequences, reflect the availability and abundance of different pathway components, particularly the Wnt ligands and Fz receptors and coreceptors, which all impact on the downstream events.

See also: Dendrite Development, Synapse Formation and Elimination; Forebrain: Early Development; Morphogens: History; Neural Patterning: Midbrain–Hindbrain Boundary; Sonic Hedgehog and Neural Patterning.

Further Reading

Cadigan KM and Nusse R (1997) Wnt signaling: A common theme in animal development. *Genes & Development* 11: 3286–3305.

Ciani L and Salinas PC (2005) WNTs in the vertebrate nervous system: From patterning to neuronal connectivity. *Nature Reviews Neuroscience* 6: 351–362.

Clevers H (2006) Wnt/beta-catenin signaling in development and disease. *Cell* 127: 469–480.

Kikuchi A, Yamamoto H, and Kishida S (2007) Multiplicity of the interactions of Wnt proteins and their receptors. *Cellular Signaling* 19: 659–671.

Kohn AD and Moon RT (2005) Wnt and calcium signaling: β-Catenin-independent pathways. *Cell Calcium* 38: 439–446.

Li F, Chong ZZ, and Maiese K (2005) Vital elements of the Wnt–Frizzled signaling pathway in the nervous system. *Current Neurovascular Research* 2: 331–340.

Moon RT, Bowerman B, Boutros M, et al. (2002) The promise and perils of Wnt signaling through β-catenin. *Science* 296: 1644–1646.

Salinas PC and Price SR (2005) Cadherins and catenins in synapse development. *Current Opinion in Neurobiology* 15: 73–80.

Speese SD and Budnik V (2007) Wnts: Up-and-coming at the synapse. *Trends in Neuroscience* 30: 268–275.

Takeichi M (2007) The cadherin superfamily in neuronal connections and interactions. *Nature Reviews Neuroscience* 8: 11–20.

Wodarz A and Nusse R (1998) Mechanisms of Wnt signaling in development. *Annual Review of Cell and Developmental Biology* 14: 59–88.

Word Learning

J Snedeker, Harvard University, Cambridge, MA, USA

© 2009 Elsevier Ltd. All rights reserved.

Word learning is often considered the simplest and least controversial aspect of language development. Although theorists fiercely debate the ontogenetic and phylogenetic origins of grammar, everyone agrees that words must be learned by observing the contexts in which they are used. No other theory can explain how English-speaking children come to use 'shoe' to label footwear, whereas young French speakers use the same sequence of sounds to label cabbage. However, this self-evident truth masks a host of questions about how learning occurs and the knowledge that children bring to the problem.

Defining the Problem

To understand word learning, we need to know both what a word is and which aspects of words are learned. Surprisingly, neither question has a clear answer. There are two difficulties in defining what a word is.

The first is determining what information is included as part of a word. By all definitions, each word (or lexical entry) consists of a symbol that is paired with a concept. In the case of spoken languages, these symbols consist of phonological representations. For example, my lexical entry for 'cat' includes the phoneme sequence /kæt/ which is linked to some conceptual representation of catness. Literate individuals also store the written form of words they are familiar with. In addition, language users have other knowledge of words that could be a part of the lexical entry. Much of this knowledge is related to the grammatical environments in which a word appears. For example, our knowledge of English allows us to recognize that "I have a cat" is a grammatical sentence, whereas "I like to cat" is not. In dictionaries, we capture these facts by assigning words to grammatical categories which are listed with each definition (e.g., 'cat' is a noun and not a verb). Many have argued that grammatical categories are also stored in the mental lexicon. However, our knowledge of the relation between lexical items and grammatical structure extends beyond syntactic categories. For example, we know that a verb such as 'roll' can appear in a transitive sentence ("I rolled the ball"), whereas a verb such as 'fall' cannot ("I fell the ball"). This has led many to propose that individual lexical entries contain detailed grammatical information about the contexts of word use. Others have argued that the concepts with which words are associated allow us to derive these structural facts, eliminating the need for their storage.

The second definitional difficulty is determining how to individuate words: Which chunks of linguistic material are entered in our mental lexicons? How large are these chunks, and where does one entry end and the next begin? Clearly, words such as 'cat' that consist of a single morpheme (meaningful unit) will have to be stored because they cannot be derived in any way. However, we also store morphemes that are not whole words (-ness or-ed), some words that are made of multiple morphemes (e.g., 'walked' or 'happiness'), and idioms that consist of multiple words ('kick the bucket'). The fact that many phonological forms have several meanings creates further uncertainty about what constitutes a lexical entry. In the case of homonyms, these meanings are not related and clearly must be stored separately (e.g., river bank vs. savings bank). However, in the case of polysemes, the different meanings are related (e.g., line up vs. telephone line), and it is unclear whether the alternatives are two separate words, two ways of thinking about a single lexical entry, or two subentries of a single lexical item. Problems of this kind have led some theorists to reject the metaphor of the lexicon as the list of entries and to argue instead that the lexicon is a generative system for linking forms to meanings.

Once we have defined a word, we can ask what it means to learn one. Which aspects of a lexical entry are learned and which can be derived? Which parts of the entry precede word learning and which are the result of it? All theorists agree that children must learn the mapping between the phonological form and the concept. Although there are some minor islands of systematicity (e.g., in English words beginning with *sn-* are related to the nose), these mappings are largely arbitrary and vary across languages; thus, they must be learned. There is more controversy about the role of learning in our ability to represent the phonological form of the word or to represent the concept with which it is associated.

Adults and older children clearly represent speech in terms of stable phonological categories and thus simply have to learn the sequence of phonemes that is linked to a particular concept. The nature of infants' phonological representations has been more controversial. Because infants sometimes fail to distinguish between two novel words that are similar but phonologically distinct (e.g., 'bim' and 'dim'), it has been suggested that infants have less precise, gestalt representations of words and only develop phonological

representations as their lexicons grow. However, infants can succeed in learning phonologically similar words in more supportive contexts and have more difficulty comprehending a known word when a single phoneme has been altered. One possible interpretation is that infants represent word forms in the same way as older children but have difficulty attending to and encoding these forms when the task is challenging.

Many developmental psychologists have argued that word learning involves concept acquisition. Children do not merely map words to existing concepts; instead, they use the recurring word form as a cue to form a new category. This theory is compatible with the Whorfian hypothesis because the concepts that the learner acquires depend on the concepts that are lexicalized in the learner's language. This theory could potentially explain why children learn concrete words early and more abstract words later, or why labels for objects generally appear before predicates. However, theories of this kind face two challenges. First, it is unclear how a child could learn what a word means if he or she did not already have access to the concept that the word encodes. To recognize that a word reliably co-occurs with a particularly category of entities, one must be able to represent those entities as a class. In other words, you cannot not learn that /kæt/ means cat unless you already have a representational arsenal that allows you to distinguish cats from non-cats. Second, the course of lexical development in cognitively advanced learners suggests that conceptual limitations play little role in early word learning. Children who are internationally adopted as preschoolers must learn a new language in context similar to that of infant language learners. If the course of infant word learning is shaped by the conceptual limitations of babies, then we would expect that older children, free from such limitations, would be able to learn more verbs and abstract words during the initial phases of lexical development. However, internationally adopted preschoolers acquire words in approximately the same order as infant learners, suggesting that early vocabulary development is largely independent of cognitive development.

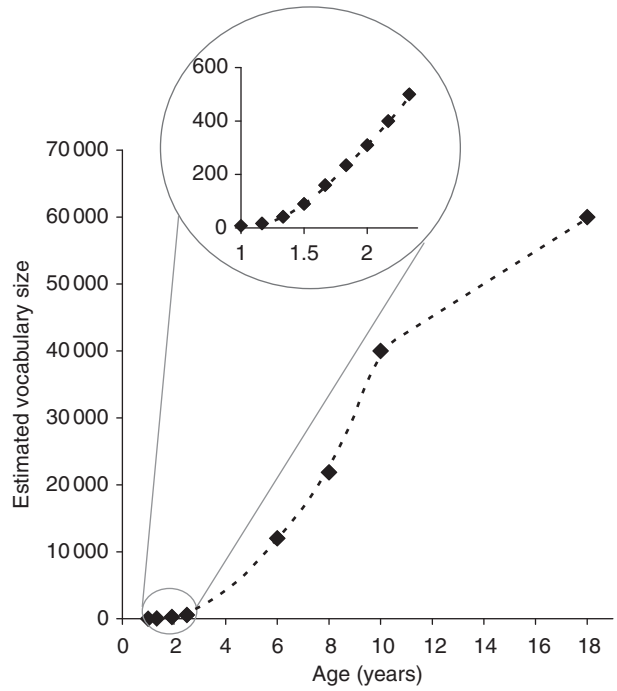

Figure 1 Vocabulary growth throughout development. The vocabulary estimates for 12- to 30-month-olds are from Fenson L, Dale PS, Reznick JS, Bates E, Thal DJ, and Pethick SJ (1994) Variability in early communicative development. *Monographs of the Society for Research in Child Development* 59: 1–173. The estimates for 6- to 10-year-olds are from Anglin J (1993) Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development* 238: 1–166. The estimate for adults is based on the figure cited by Aitchinson J (1994) *Words in the Mind: An Introduction to the Mental Lexicon*, 2nd edn. Oxford, UK: Blackwell.

Word Learning across Development

Children learning both signed and spoken languages generally produce their first word at approximately 12 months of age. By adulthood, the typical person knows approximately 60 000 words. Although this averages out to approximately 9 words a day, the actual pace of word learning varies across development. As Figure 1 illustrates, the initial pace of word learning is quite slow, with the average child learning approximately 2 words a week between 12 and 16 months of age. Until recently, many observers believed that there was a sudden acceleration in word learning at approximately 18 months of age (often called the vocabulary spurt). However, closer examination has revealed that the pace of word learning increases gradually throughout the toddler and preschool years. In literate societies, this steady acceleration continues into the school years as children encounter new words while reading. The pace of word learning begins to decelerate sometime between 8 and 18 years, presumably because older children encounter fewer unknown words. There is no evidence for a critical period in word learning: Adults do as well or better than children in most experimental word learning tasks and we can readily learn new words throughout our lives.

As children's vocabularies grow, they also change. The first words that children learn are typically labels for people ('mommy' or 'daddy'), animal names ('kitty'), social words ('hi' or 'uh-oh'), or utterances used in common routines ('peek-a-boo'). Although verbs, prepositions, and other relational words

appear in early vocabularies, they may have more limited meanings for infants than they do for adults and older children. For example, many 1-year-olds say 'more' to request the recurrence of an event or object, but there is no evidence that children of this age understand that 'more' is a term that quantifies the amount in one set relative to another.

The early vocabularies of children learning English are dominated by nouns, most of which label people, animals, or objects. Even though adults speak to young children in full sentences, complete with verbs and grammatical words, these elements are massively underrepresented in children's early vocabularies. The proportion of nouns in the child's lexicon varies from language to language depending on whether subjects and objects are mandatory and whether the verb appears in a perceptually salient position (e.g., at the beginning or end of the utterance). Nevertheless, in every language that has been systematically studied, nouns are overrepresented in children's vocabularies relative to their rate of occurrence in the input, whereas verbs are underrepresented. This suggests that nouns, particularly those that label objects or people, are easier for children to learn than verbs or other relational words. Some theorists have attributed this to the greater conceptual complexity of verbs, whereas others attribute it to the nature of the information needed to identify the meanings of nouns and verbs.

There is considerable variability in how rapidly children learn words. At 16 months of age, the median vocabulary size for middle-class children in the US, is approximately 40 words. However, 10% of children produce no words, whereas another 10% produce more than 150 words. Early in development, vocabulary size is a better predictor of vocabulary composition than age. As the child's vocabulary grows from 50 to 200 words, the proportion of these words that are nouns increases. Verbs and adjectives begin to appear in greater numbers between 100 and 400 words. Grammatical words (such as articles, pronouns, prepositions, and auxiliaries) increase in frequency at approximately 400 words. Between 16 and 30 months, the size of a child's productive vocabulary is tightly correlated with the grammatical complexity of the child's speech. Initially, children primarily produce one-word utterances. When their productive vocabulary reaches 50–200 words, they begin combining words into short phrases. From 300 to 500 words, grammatical morphemes appear and the child's utterances increase in length and complexity. These relations are preserved in internationally adopted preschoolers, suggesting that the correlation is not a side effect of global maturational or cognitive changes. In bilingual children, these relations hold within a language but not between languages, suggesting that lexical development facilitates grammatical development or vice versa. During the school years, lexical development and reading are closely linked. A child's vocabulary when he or she enters school is a strong predictor of later reading achievement, and subsequent vocabulary development is correlated with the amount that the child reads.

Word Learning as Induction

How do children learn the meanings of words? Most people who have thought about the problem long enough have come up with essentially the same solution: Word learning is a form of induction. Learners generate hypotheses based on the situation in which the new word occurs. As the learner observes new instances of the word, hypotheses are eliminated or strengthened, allowing him or her to close in on the correct meaning.

Although this mechanism clearly plays a role in word learning, it cannot be the entire story. First, even very young children can learn some words after hearing them used in just one context. Second, many words, particularly verbs and other relational terms, are often used in the absence of the event being labeled (e.g., when parents tell children to "go to sleep," both parties are typically awake).

Finally, the account of word learning given previously is subject to the mid-century critiques of empiricism voiced by philosophers such as W. V. O. Quine and Nelson Goodman. Learning simply cannot be unconstrained induction because any finite set of observations is consistent with an infinite set of hypotheses. To take Quine's example, the set of observations that would allow one to learn that a word ('gavagai') means rabbit is also compatible with the hypothesis that it means undetached rabbit parts (or temporal rabbit slices). Thus, a full account of lexical development requires more than merely stating that word learning is induction. In the past 25 years, developmental psycholinguists have begun to flesh out this story by identifying three ways in which children and adults can tame the induction problem and learn the meanings of words.

Constraints and Biases on Hypothesized Meanings

The first way to tame induction is by limiting the range of meanings that are considered as possible hypotheses. Many theoretically possible meanings may be ruled out simply because children cannot or do not think of them – for word learning or any other purpose. Quine's example of undetached rabbit parts

is presumably ruled out by cognitive constraints of this kind. However, such constraints cannot explain how the learner rules out more plausible alternatives such as white, fluffy, hopping, tail, animal, or Flopsie. All these hypotheses are plausible ones, meanings which the learner could eventually link to some other phonological form. Children, however, are biased learners who privilege some of these hypotheses over others.

The nature of these biases has been explored with an experimental paradigm called the word extension task. In this task, children are shown a novel object (e.g., a blue kidney-shaped piece of plaster) which is labeled with a novel word ("This is my dax"). Then they are given a small set of test objects (e.g., a blue circular piece of plaster and a red kidney-shaped piece of wood) and asked to find another example of the target word ("Can you give me the dax?"). The child's response allows the experimenter to infer how the child is interpreting the novel word.

Through tasks such as this one, researchers have discovered that children (and adults) have a strong bias to assume that a new word refers to a whole object rather than one of its parts, its properties, or the relations that it is involved in. When children have mapped a word to an object, they will extend it to other objects of the same kind rather than to other objects that are involved in the same event (e.g., extending from a birthday cake to other cakes rather than to birthday presents and birthday candles). In the case of novel artifacts, they typically extend the word to objects of the same shape rather than to objects of the same material or same color. However, when a word is applied to a novel animal, children are more conservative, preferring referents that have the same color or texture and the same shape.

These biases should aid the child in learning count nouns (such as 'cat,' 'book,' or 'car'), but they might hinder the acquisition of adjectives, verbs, relational nouns, and labels for substances and parts of objects. For example, if little Johnny were to hear the word 'silver' applied to Mommy's car, his bias to map words to whole objects could lead him to incorrectly conclude that 'silver' means car. One additional constraint, often called mutual exclusivity, may help learners overcome this bias. Young children are reluctant to map a second label to an object for which they already have a label. Thus, if little Johnny has already learned that cars are called 'car,' he is likely to reject the hypothesis that 'silver' means car and consider other possibilities.

The existence of these biases is uncontroversial, but their origins are not. Current research in this area focuses on two issues. The first is domain specificity. Are these biases unique to word learning or do they reflect more general properties of cognition? Perhaps the bias to link words to whole objects reflects the centrality of objects in early cognition rather than any preconceptions about the nature of words. Similarly, mutual exclusivity in word learning could result from a more general principle of communication which leads learners to assume that speakers will use known forms of reference whenever possible. The second issue is whether these biases are available at the onset of word learning. Some theorists suggest that these biases are learned as the child acquires words and notices the properties that are typically used in word extension. For example, children may learn their first artifact labels by trial and error. After learning many of these words, a child may notice that these labels are generally extended by shape and develop a bias to extend new words in this way. Several pieces of evidence support this account. The shape bias for novel artifacts increases as vocabulary size grows, and toddlers who are taught categories that are extended on the basis of shape develop this bias more quickly and are more successful at learning count nouns.

Social Cues to Reference

Children can also tame the induction problem by using their implicit understanding of social interactions to make inferences about the communicative intentions of the speaker. These inferences can help the child identify the object or event to which the speaker is referring, thus simplifying the problem of word learning. The best studied social cue to word meaning is the speaker's direction of gaze. When speakers are talking about objects that are visible, they often look at that object at roughly the same time as they are mentioning it. By approximately 18 months of age, children will map a word to the object that the speaker was looking at, even when they themselves were examining a different object when the word was spoken. Infants can also identify the referent of a word when the speaker points at the object or moves it as it is labeled.

Other abilities appear to require a more complex representation of the interaction. When a speaker labels an action before performing it, young children prefer to link the label to an intentional action rather than an accidental one. When a speaker uses a novel label, children assume that she is referring to an object that was introduced while she was gone rather than an object that she saw before.

Research on the use of social cues has explored two questions. First, what role do these cues play in word learning? Are they necessary or merely facilitory? Several lines of evidence demonstrate that children can learn words in the absence of any single social cue. For example, infants can learn the mapping

between an object and a word from a videotape in which the person who is producing the word is never visible. The ability of blind children to acquire words at approximately the same age as sighted children provides further evidence that visual access to social cues (such as pointing or gaze direction) is not necessary for successful word learning. Of course, it is possible that these learning contexts contain other cues that allow the child to infer the referential intentions of the speaker. A second and related question is whether infants' use of social cues reflects an understanding of the mental states of the speaker rather than simple associations or low-level attentional processes. Do infants make use of eye gaze because they know that it reflects the speaker's knowledge and communicative intentions, or have they merely learned that gaze is a reliable predictor of which object a word will be associated with?

Sentential Contexts as Cues to Word Meanings

The final way in which children tame the problem of word learning is by using the sentential context in which a word occurs to narrow down the set of possible hypotheses. Sentential contexts provide three kinds of information. First, they identify the syntactic category of a novel word. Because syntactic categories are systematically (but imperfectly) linked to semantic categories, they provide information about the kind of meaning that the word is likely to have. For example, 1-year-olds who are shown a doll and told that 'this is a zav' interpret the novel word as a common noun and extend it to other similar dolls. In contrast, those who hear 'this is Zav' interpret it as a proper noun and refuse to extend it. Similarly, by approximately 2½ years of age, young children have learned to extend mass nouns ('this is some dax') to entities that are made of the same substance and to extend adjectives ('this is a daxy one') to objects that share a common property.

Second, sentential contexts provide information about the number and kinds of arguments that a predicate can take which constrain the kind of meaning that the predicate has. For example, verbs of self-generated motion (e.g., 'dance') often appear with just one argument (the subject); verbs of contact (e.g., 'hit') or caused motion (e.g. 'push') usually appear with a subject and a direct object; and verbs of transfer (e.g., 'send') appear with three arguments (subject, direct object, and indirect object). Young children use these connections between syntax and semantics to learn the meanings of unknown words. For example, 2-year-olds who hear a transitive utterance such as "She is blicking her around" expect the verb to have a meaning such as 'push,' whereas those who hear an intransitive utterance such as "They are blicking around" expect the verb to have a meaning such as 'dance.'

Finally, sentential contexts contain known words which provide information about the other entities or relations that are involved in the event under discussion. For example, if children hear an utterance such as 'John is eating a dax,' they can infer that a dax is probably a kind of food, and they can direct their attention to John to ascertain precisely what kind of food it might be. Although cues of this kind tightly constrain word learning, their availability depends on the child having already learned some of the other content words in the utterance.

Current research on the role of sentential contexts centers on two issues. First, when do children begin to use each of these cues? Some cues have an influence on infants as young as 13 months of age (e.g., use of word in count noun syntax), but other cues appear to have little effect even in preschoolers. For example, in several studies 2- and 3-year-olds have failed to use verbal morphology ("Look gorping!") to map the word to the action (e.g., waving) that is being performed. The second active area of research explores how these cues are acquired. One possibility is that by learning words from particular categories, children discover a direct association between specific grammatical contexts and particular aspects of word meaning. The second possibility is that children's word learning is guided by universal linkages between syntax and semantics. Children must learn the grammatical morphemes that mark each syntactic category (e.g., that 'a dax' is a count noun in English and 'une dax' is a count noun in French) but the connection between particular syntactic categories and patterns of word extension comes for free. These theories make different predictions about when these cues should be learned and whether the semantic relations in question are likely to be linked to observable properties (e.g., count nouns pick out shape-based categories) or more abstract properties that are tightly linked to the syntax (e.g., count nouns pick out individuated entities).

Final Words

During the past 25 years, researchers have demonstrated that young children make use of all three of the previously discussed strategies to learn words. The challenge in the future will be to understand how these strategies work together throughout development. It may be useful to think of constraints, social cues, and structural contexts as complementary mechanisms rather than competing hypotheses. The three strategies account for different aspects of word learning. For example, social cues allow children to

identify the referent of the phrase, whereas constraints on hypotheses and structural cues provide guidance about how this referent is being construed.

The order in which these abilities emerge during development may explain why early vocabularies are dominated by nouns but contain few verbs, and why children's mastery of verbs coincides with the emergence of grammar. Many social cues and constraints may be available at the onset of word learning. These abilities do not necessarily depend on prior linguistic knowledge, and their use extends well beyond the linguistic domain (e.g., knowledge of object kinds constrains reasoning as well as word learning, and eye gaze provides information about a person's knowledge outside of the linguistic domain). These abilities would primarily facilitate the acquisition of nouns: Most word learning constraints and biases apply to object categories and most social cues function to pick out objects in space. In contrast, the use of structural contexts necessarily requires some grammatical knowledge and thus can only be used after language acquisition is underway. This information source is particularly useful for learning verbs and other predicates (which occur in richer structural contexts). Thus, efficient verb learning may emerge only when children have acquired enough grammatical and lexical knowledge to make use of sentential contexts during word learning.

See also: Bilingualism; Language Development; Language: Cortical Processes; Lexical Impairments Following Brain Injury; Second Language Acquisition; Word Production; Word Recognition.

Further Reading

Baldwin DA (1991) Infants' contribution to the achievement of joint reference. *Child Development* 62: 875–890.

Bates E, Dale PS, and Thal D (1995) Individual differences and their implications for theories of language development. In: Fletcher P and MacWhinney B (eds.) *The Handbook of Child Language*, pp. 96–151. Oxford, UK: Blackwell.

Bates E and Goodman JC (1997) On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes* 12(5/6): 507–586.

Bloom P (2000) *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Caselli MC, Casadio P, and Bates E (1999) A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language* 26: 69–111.

Gentner D and Boroditsky L (2001) Individuation, relativity and early word learning. In: Bowerman M and Levinson S (eds.) *Language Acquisition and Conceptual Development*. Cambridge, UK: Cambridge University Press.

Gillette J, Gleitman H, Gleitman L, and Lederer A (1999) Human simulation of vocabulary learning. *Cognition* 73: 135–151.

Gleitman LR (1990) The structural sources of word meaning. *Language Acquisition* 1: 3–55.

Hall DG and Waxman SR (2004) *Weaving a Lexicon*. Cambridge, MA: MIT Press.

Markman EM (1989) *Categorization and Naming in Children*. Cambridge, MA: MIT Press.

Naigles L (1990) Children use syntax to learn verb meanings. *Journal of Child Language* 17: 357–374.

Smith LB, Jones SS, Landau B, Gershkoff-Stowe L, and Samuelson L (2002) Object name learning provides on-the-job training for attention. *Psychological Science* 13: 13–19.

Snedeker J, Geren J, and Shafto C (2007) Starting over: International adoption as a natural experiment in language development. *Psychological Science* 18: 79–87.

Waxman SR (1999) Specifying the scope of 13-month-olds' expectations for novel words. *Cognition* 70: B35–B50.

Word Production

R J Hartsuiker, Ghent University, Ghent, Belgium
G Vigliocco, University College London, London, UK

© 2009 Elsevier Ltd. All rights reserved.

Introduction

Please say the name of the item pictured in Figure 1.

Research on word production is concerned with the mental processes and representations that are involved during the 1000–1200 ms that it probably took from seeing the picture to uttering the word 'banana.' In picture naming, this involves the initial process of visual processing, followed by core processes of identifying the concept, retrieving the appropriate word from memory, determining the word's phonological structure (e.g., the word has three syllables) and content (e.g., the first speech sound is a /b/), and planning a speech motor program, which controls articulation.

Historically, language production has been the subject of less research than language comprehension, and this holds for both research studying the word level and the sentence level. However, in recent years there has been an increase in interest in production. This increased interest is accompanied by important methodological innovation. Until fairly recently, the main source of data informing the field of word production research was corpora of spontaneously occurring speech errors. Although this work has allowed the development of influential production theories, in particular Merrill Garrett's theory of language production, observational methods suffer from important problems, such as collector biases, the infrequency of errors, and the issue of whether the corpus is representative of the language as a whole. Methodological improvements have made it possible to study word production under laboratory conditions, considering normal, fluent production as well as the production of errors and allowing researchers to focus on the time course of the production process. Moreover, despite methodological obstacles (e.g., the need to control for movement artifacts), word production can now be studied while registering the speakers' eye-movements, event-related potentials, or Bold signals.

The Functional Architecture of Word Production

Word production is a multistage process. That is, given the intention to say something, representations need to be selected or constructed in a sequence of processing levels. Each of these processing levels is specialized and represents a particular type of knowledge, such as meaning or form. Although there is consensus in the field about this hierarchical organization, there is considerable debate about whether the processing stages are strictly sequential or whether they interact to some degree (see below). This multistage nature of word production is sometimes overlooked. For example, some studies of word recognition use the task of word naming, neglecting the fact that this task also recruits production.

A blueprint of word production is provided in Figure 2. The figure shows the core processing levels (against the blue background) and several sources of input to this system. Imagine that a speaker will produce the word 'banana,' perhaps so as to name a picture or, in a more naturalistic situation, because the person is thinking about a banana and wants to say something about it. The following sections describe the route that processing takes from seeing to saying, excepting details of articulation.

Conceptual Preparation

The first core processing level is the conceptual level. Here, a representation referring to a banana becomes activated, and representations of related concepts (e.g., 'apple,' 'pear,' 'fruit,' 'peel' (part of a banana), 'canoe' (looks like a banana)) are also somewhat activated. There is no consensus on the precise nature of the conceptual representations used for production. Some researchers argue for distributed conceptual representations, so that a given concept can be characterized by a constellation of features (either 'macrofeatures' that carry meaning themselves, such as 'yellow,' 'sweet,' 'long,' or in certain connectionist models, 'microfeatures' without any intrinsic meaning). In these views, similarity among concepts is determined on the bases of features that are shared across concepts and features that are instead specific for a given concept. Other researchers, however, argue for localist concepts, so that there would be a single, indivisable representation for each concept. On such 'semantic network' accounts, concepts are connected via meaningful links that indicate category membership (e.g., a banana is a fruit) and properties (e.g., a banana has a yellow color). Similar concepts have links to the same categories and properties.

Do speakers of different languages activate identical concepts when they are preparing to say a word? Probably not. Languages differ in what they require their speakers to think about. For example, English often captures the manner of motion in a verb, whereas Spanish needs to add such information in

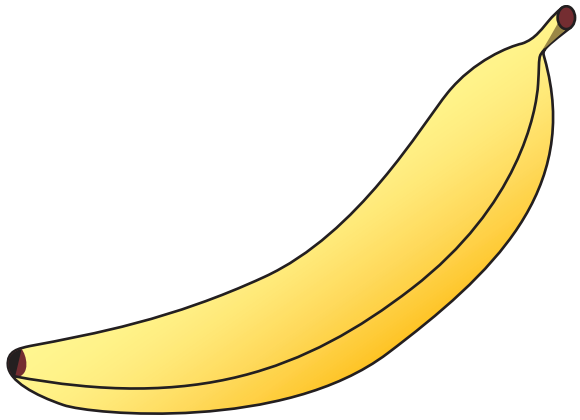

Figure 1 Picture of an everyday object. From the picture set presented in Rossion B and Pourtois G (2004) Revisiting Snodgrass and Vanderwart's object set: The role of surface detail in basic-level object recognition. *Perception 33: 217–236.*

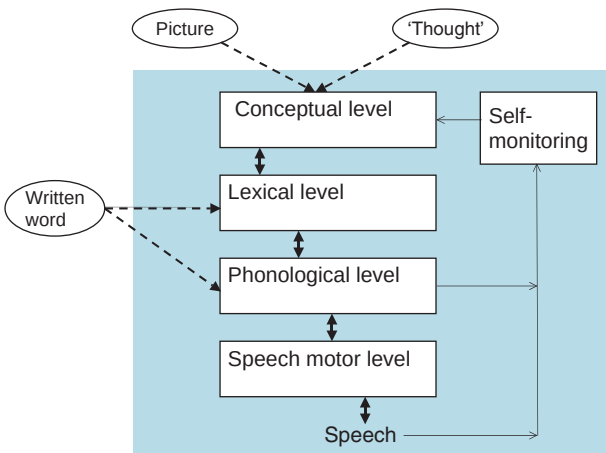

Figure 2 Overview of core word production processes (against the blue background) and lead-in stimuli, such as pictures, words, and thoughts.

an additional phrase. Research has shown that such differences influence the type of information speakers recall from a description (e.g., of a cartoon) they recently provided, suggesting that when verbalizing events, conceptual information is activated in a language-specific way.

Lexical Processing

The activated concept will send activation to the lexical level, and if all goes well, the lexical representation for banana will become the most active representation at this level. If there was always a one-to-one relation between concepts and words, it would not be necessary to postulate different processing levels for them. However, there are many situations in which there is no one-to-one relation between concepts and words. For example, there are concepts such as 'the actions of two people maneuvering for one armrest in a movie theater or airplane seat' for which we do not have a word. Moreover, as discussed above, which concepts are verbalized can differ across languages.

At the lexical level, the word activated by the concept (the 'target word') enters into a competition with other activated words, usually words with a similar meaning. The word with the most activation will be selected, and the time to select the correct word and the chance of making a lexical selection error depend on the activation level of competitor words. This mechanism of selection by competition is nicely illustrated by results from the picture–word interference paradigm, one of the most frequently employed paradigms in word production research. In these experiments, participants are instructed to name pictures with single words while ignoring a 'distracter word' that is presented either simultaneously with the picture or shortly (e.g., 150 ms) before or after the picture. A consistent finding is that semantically related distracters (e.g., 'apple' for the picture of a banana) delay picture-naming onsets compared with unrelated distracters (e.g., 'chair' for the picture of the banana). This delay can be explained as follows: Both the written words 'apple' and 'chair' directly activate their corresponding representations at the lexical level (Figure 2). However, the picture's concept (for banana) activates the related concept (for apple) but not the unrelated concept (for chair). Therefore, the lexical representation of 'apple' receives additional input from and is a stronger competitor for 'banana' than is 'chair.' Computer models of word production that incorporate this mechanism can mimic the temporal costs caused by semantic relatedness.

How many lexical levels are there? Some researchers argue that the lexical level should be subdivided into two levels, namely, a lemma level and a word form level, whereas others postulate only a single lexical level. In the two-level view, lemmas are lexical representations containing or pointing to lexical-syntactic information (e.g., gender) and semantic information, but not phonological information. Furthermore, there are different versions of two-level theories. Thus, in his seminal book *Speaking*, Willem Levelt suggested that the lemma representation of a word contained semantic and syntactic information and pointed to the word form representation that contained morphological and phonological information. But in a more recent version of Levelt's theory, the lemma is considered void of content; it connects the conceptual level to the level of word forms and is connected to (rather than contains) syntactic information.

There is much evidence that appears to support a division into a lemma level and a word form level.

One source of observational evidence concerns the occurrence of word substitution errors. These are related to the intended word either in its meaning or in its form, suggesting that there are two types of retrieval processes that can derail: (1) lemma selection, which is driven by semantic information and can lead to semantic errors, and (2) word form selection, which can lead to phonologically conditioned errors. For example:

1. I want whipped cream on my mushrooms. (intended: strawberries)
2. I've got some whipped cream on my mushroom. (intended: mustache)

In addition, picture–word interference studies show distinct effects of semantically related distracters, which slow down picture naming (see above), and phonologically related distracters, which speed up picture naming. The difference in polarity of these effects is consistent with the notion of different processing levels.

Further evidence comes from the so-called tip-of-the-tongue (TOT) phenomenon. This phenomenon, strikingly described by William James in 1890, entails the inability to produce a word one is certain to know. Studies that have elicited this phenomenon in the laboratory demonstrated that speakers in a TOT state reliably know syntactic information about a word (e.g., whether an Italian noun's grammatical gender is masculine or feminine) while being unable to retrieve phonological information about the word. This suggests that speakers have succeeded in retrieving the lemma (and can therefore activate the syntactic information associated with it) but have failed to retrieve the word form. It is interesting that brain-damaged patients who suffer from an inability to name objects or events likewise demonstrate syntactic knowledge of the words they cannot name.

A final source of support for the distinction between lemmas and word forms comes from studies of morphological processing. A model with two such levels rather naturally lends itself to the representation of morphologically complex words. For example, a Dutch complex verb like *opgeven* (to give up) would have a separate lemma from a simplex verb like *geven* (to give). This is syntactically motivated as these verbs are syntactically different. At the word form level, however, overlapping inflectional forms of *geven* and *opgeven* would be shared, for example, *geeft* in the third-person, present-tense forms *geeft op* and *geeft*. In line with this hypothesis, Dutch speakers have a naming latency advantage when saying *geeft op* compared with *veegt op* (sweeps up), although these verb–particle combinations occur equally frequently. The advantage is that *geeft op* shares a word form with a very frequent verb, whereas *veegt op* shares a word form with an infrequent verb.

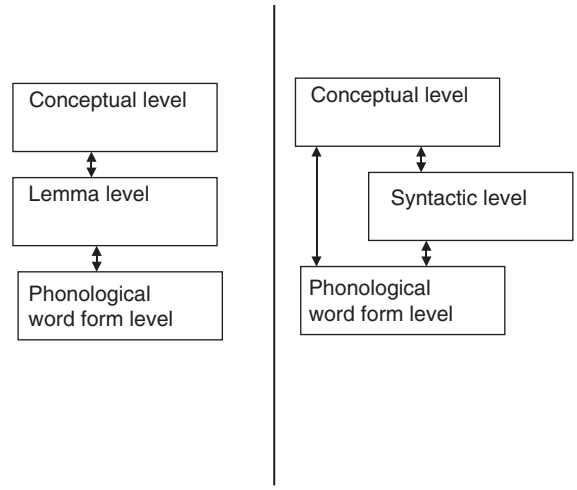

Figure 3 Lexical representations according to Levelt and colleagues (left panel) and Caramazza and colleagues (right panel).

This subdivision into lemmas and word forms has been challenged by Alfonso Caramazza and colleagues, who proposed that the conceptual level is directly connected to the level of word forms and that both these levels are independently connected to syntactic information (Figure 3). In addition to a network of phonological word forms (employed in speaking), there would be orthographic word forms (used in writing). Much evidence that supports the distinction between lemmas and word forms is also consistent with Caramazza's framework.

Phonological Processing

Most theories assume a level of word forms, at which words are represented in terms of their speech sounds. It is conceivable that an error at this level leads to a malapropism, a phonologically governed word substitution error, like the Dutch error *ondergoed* (underwear) instead of *onderzoek* (research). Perhaps surprisingly, the next processing step involves breaking up the word form into its constituents, such as its phonological segments and metrical structure. Why doesn't the processor simply use the just-retrieved complete word form to guide articulation? An important reason is that speakers normally do not produce words in their citation forms. Rather, in connected speech, words are merged, and syllables can cross word boundaries. For example, in natural speech, one does not say "and Oliver" but "n doliver."

An influential model of the processes involved in phonologically encoding words was proposed by Gary Dell (Figure 4). This model, analogous to McClelland and Rumelhart's seminal Interactive Activation Model of word recognition, postulates

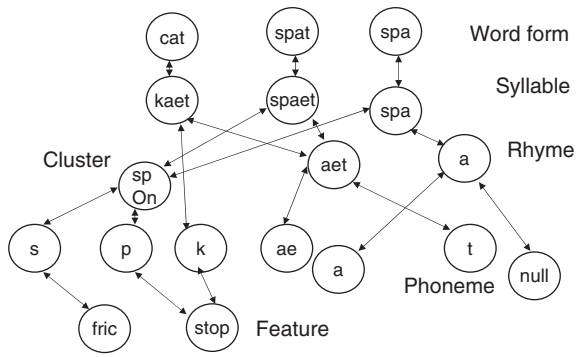

Figure 4 Gary Dell's 1986 model of word production.

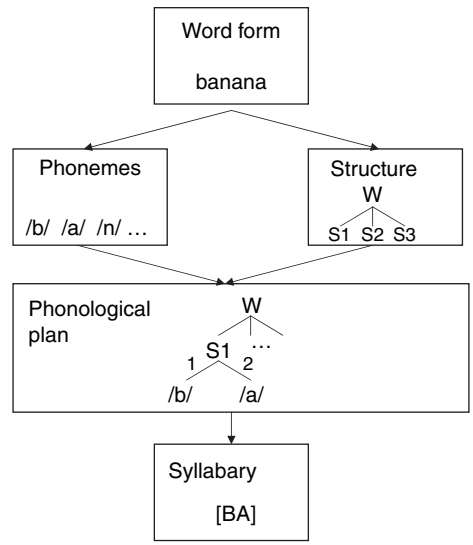

Figure 5 The WEAVER++ model's architecture of phonological encoding.

localist representations for different phonological units (such as words, syllables, and phonological clusters); processing proceeds by spreading activation from a target unit. Selection takes place at the phoneme level, so that after a given number of time steps of activation spreading (determined by speech rate), the most highly active phoneme unit is selected in separate pools of units. These pools correspond to the onset (initial consonants), nucleus (vowel), and coda (final consonants) of the syllable.

If all goes well, phoneme units corresponding to the intended word are selected, but whenever a competing phoneme has a higher activation level than the intended one, a speech error occurs. There are several reasons an unintended phoneme can win the competition. It may have residual activation because it has been recently selected and the activation has not decayed yet. It is also possible that the unit needs to be selected in an upcoming syllable but is activated too early. Finally, a phoneme may be randomly active because the system is intrinsically noisy.

Dell's model successfully accounts for naturalistic speech error data. For example, the model simulates effects of speech rate on the number of errors; the so-called lexical bias effect (by which errors tend to form existing words more often than chance would predict); and the distribution of speech errors, so that anticipations (in which a phoneme is said too early; e.g., 'bake my bike' (intended: take my bike)) outnumber perseverations (in which a phoneme is repeated in a subsequent word; e.g., 'she pulled a pantrum' (intended: she pulled a tantrum)), which in turn outnumber exchanges (in which phonemes from different words are swapped; e.g., 'waple malnut' (intended: maple walnut)).

However, Dell's model has two limitations. First, the model is restricted to accounting for the outcome of phonological encoding (correct or error) but not for the time course of processing. Another limitation is that the model does not have a mechanism for the resyllabification processes that take place in connected speech: it produces words in citation form only. Ardi Roelofs, Willem Levelt, and colleagues therefore extended their theory of lexical access to phonological encoding. This new model, WEAVER++ (Figure 5), complements Dell's: it provides an account of reaction time data (e.g., results from picture–word interference experiments) but does not account for speech errors.

WEAVER++ assumes that the word form is broken down into its parts in two parallel processing streams, one that retrieves individual phonemes and one that retrieves the metrical frame. The latter frame specifies the number of syllables of the word and the location of stress. These processes take place in parallel for all constituent syllables of the phonological word, which is usually defined as a content word (like 'banana') and any function words (e.g., 'a,' 'the,' 'it') that are phonologically 'glued' to it. Based on these representations and on the language's resyllabification rules (not shown in the figure), a syllabified, phonological plan for the phonological word is constructed. Finally, information now associated with syllable representations (e.g., that the onset of the first syllable is /b/ and its rhyme is /a/) serves as input to the 'syllabary,' a repository of motor programs that guide the articulation of individual syllables.

This model has successfully simulated reaction time data, mostly from experiments that used the picture–word interference paradigm or the implicit priming

paradigm (a task in which speakers produce blocks of words with (experimental condition) or without (control condition) phonological overlap). Note, however, that some assumptions of the WEAVER++ model are controversial. In particular, there is no compelling evidence in support of the mental syllabary.

Self-Monitoring

Speakers monitor whether the words they produce conform to what they wanted to say. This verbal self-monitoring system inspects not only a speaker's overt speech but also speech before articulation (internal speech). That internal speech is monitored is evident from data on self-interruptions. Speakers can interrupt their errors so quickly that they could not have detected the error in their overt speech. Despite recent attempts to formulate precise computational models of the monitoring process, it is still unclear how monitoring establishes that something is an error. This issue has important theoretical repercussions, in particular with respect to the question of whether the processing levels in production have the characteristic of modular stages or whether there is cross-talk between different levels (see the next section).

Information Flow in the Functional Architecture: Modularity versus Interactivity

A debate that has divided the word production literature for decades concerns whether processing levels are modular (i.e., encapsulated from other levels) or not. There are two aspects to this issue. First, models differ in whether they assume that only the selected representation at level $n-1$ activates representations at the next level, n (the modular stance), or whether processing at these levels overlaps somewhat in time so that all representations activated at level $n-1$ send some activation to level n. The latter mechanism is usually referred to as 'cascading of activation.' Second, models differ in whether they assume that activation flow is unidirectional (from level $n-1$ to level n), or whether there is also a return flow of activation (from level n to level $n-1$). The latter mechanism is usually referred to as 'feedback of activation.' Note that even theorists who propose full interactivity (cascading and feedback) limit it to local interactivity. There appears to be no direct (cross-) talk between level $n-2$ and level n.

In contrast to conclusions of several studies in the early 1990s, there is now overwhelming evidence for cascading of activation. For example, picture–word interference studies using homophone targets (e.g., a picture of a (toy) ball) showed facilitation when the distracter was related to the other meaning of the target name (e.g., 'dance'). This demonstrates that the distracter activated the lemma for the 'dance' meaning of 'ball' and that this lemma cascaded activation to the word form 'ball,' thereby speeding up selection of this word form.

The existence of feedback remains disputed. Data from speech errors appear to support feedback. For example, the lexical bias effect on phonological speech errors suggests that feedback from the phonological level activates representations of existing words, increasing the probability that they are selected. It is important that Dell's model, which incorporates feedback links, simulates the lexical bias effect, as well as effects of speech rate on the magnitude of this bias. However, a competing account of the lexical bias assumes unidirectional activation flow and explains the effect as the result of covert error repair. On that story, the monitoring system uses criteria such as, is this a word? and is therefore more likely to detect and intercept nonsense errors than errors that form existing words. It is difficult to reject such accounts, because the mechanisms of error detection remain underspecified. Note that recent evidence suggests a third possibility, one by which monitoring biases interact with an underlying pattern of errors, which is shaped by feedback.

Neural Correlates of the Functional Architecture

What are the neural systems engaged in word production? It has long been known from patient studies that language processing takes place in a brain network that is largely left-lateralized and surrounds the perisylvian fissure. Modern brain imaging techniques allow for a more fine-grained localization of brain areas involved in production, although only very recently have studies begun to address overt production, thanks to the advent of data acquisition and analysis techniques that reduce susceptibility to movement artifacts.

A recent meta-analysis considered 85 imaging studies (mostly positron-emission tomography and functional magnetic resonance imaging) in which participants were engaged in tasks like picture naming, verb generation (e.g., "Generate as many verbs as you can applying to a particular noun"), word naming, pseudoword naming, and so on. These tasks differ in the extent to which they involve different processing components (Figure 2). The logic of the meta-analysis was that if a certain brain region genuinely subserves a certain process, then this region should be active

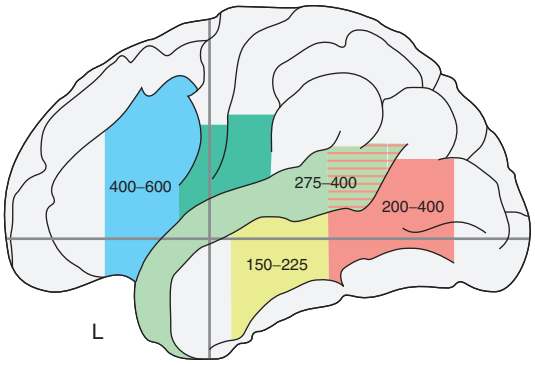

Figure 6 Proposed localization of brain regions involved in word production processes. The numbers refer to Indefrey's and Levelt's estimates of the moments of time during which these areas would be active, given the task of picture naming. Adapted from *Cognition*, 92, Indefrey P and Levelt WJM, The spatial and temporal signatures of word production components, pp. 101–144 (2004), with permission from Elsevier.

in all tasks that involve this process. That study suggested several localizations of production processes (Figure 6).

Conceptual processing and lemma selection are subserved by an area in the middle of middle temporal gyrus. Word form retrieval is localized in the posterior part of the middle and superior temporal gyri (Wernicke's area). Phonological encoding is localized in the left posterior inferior gyrus (Broca's area). Articulation is subserved by several regions, including bilateral sensory and motor areas. Finally, self-monitoring is localized in the superior temporal gyri of both hemispheres.

A more detailed neuroanatomical hypothesis concerning the brain network(s) subserving word production has been proposed by Hanna Damasio and colleagues on the basis of imaging and lesion data. In this view, word naming is achieved in the following manner: Conceptual retrieval corresponds to engagement of intermediary regions in higher-order association cortices. The brain network subserving conceptual activation can differ (at least partially) depending on the semantic category of the stimulus (e.g., animals vs. famous people) and recruits regions outside the 'language' network, for example, in left inferior temporal areas. Once conceptual activation is achieved, naming a stimulus from a particular conceptual category is dependent on intermediary structures for 'words.' These intermediary structures correspond to the neural substrate of semantic representations. Finally, structures which support the implementation of word forms (the classical language areas located in the left perisylvian region, including Broca's and Wernicke's areas) are engaged. These structures correspond to the neural substrate of phonological word forms and the processes engaged in their retrieval.

Consider Figure 7. When an individual is shown a picture of a cup or a picture of a familiar person (e.g., John F. Kennedy) and is asked to name the item, there will be activation of the appropriate 'concept' intermediary regions, which will promote concept retrieval. Activation in the concept intermediary regions leads to activation in the corresponding 'word' intermediary region, which in turn promotes the retrieval of lexical knowledge required for word production. This is likely to require auditory structures (areas 41/42; 22), which overlap in part with Wernicke's area. This step can be considered to closely correspond to the retrieval of word forms. Finally, somatomotor (3/1/2; 40; 44/45) structures are engaged for the phonetic sequencing and production of the actual spoken name; these structures overlap in part with Broca's area.

Summary and Future Directions

There is broad consensus in the literature on the functional architecture of word production, and recent imaging and patient studies have led to suggestions about the localization of global processing components in the brain. However, important lacunae in our knowledge remain. These include the nature of conceptual representations, the number of levels of lexical representation, the existence of a mental syllabary, the workings of the self-monitor, and the existence of direct feedback between processing levels.

It is important to note that word production research studies the production of isolated words, neglecting that words are usually parts of sentences and that sentences are parts of conversations. A consideration of these higher levels raises important further questions. For example, one debate in the literature concerns the scope of planning in language production: Do we produce one word at a time, or can the production system plan multiple words in parallel? Another debate concerns the interface between words and sentence level syntax: Do words constrain the syntactic structure of the rest of the sentence, or do syntactic planning mechanisms recruit words that fit the syntactic requirements?

It is also important to note that word production has concentrated mainly on production of spoken words. However, people have several modalities at hand for word production, including typing, handwriting, and signing. Issues regarding the number of levels that are shared or separate between modalities remain largely unexplored. Finally, with bilingualism being the rule rather than the exception

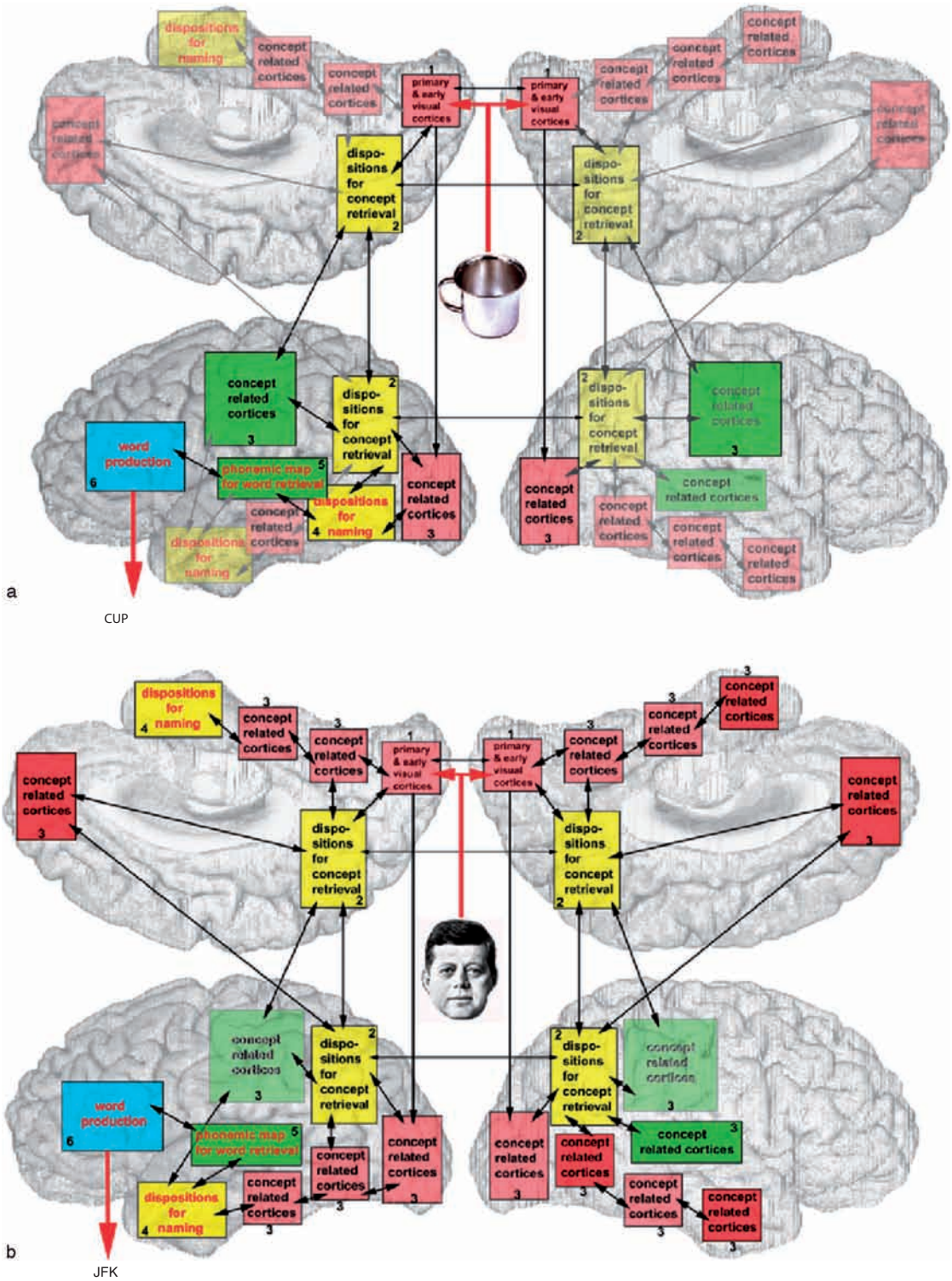

Figure 7 Reprinted from *Cognition*, 92, Damasio H, Tranel D, Grabowsky TJ, Adolphs R, and Damasio HR, Neural systems behind word and concept retrieval, pp. 179–229 (2004), with permission from Elsevier.

in Earth's population, many speakers have two or more lexicons to choose from, which raises important new questions about the functional organization and processing mechanisms in bilingual people.

See also: Bilingualism; Connectionist Models of Language Processing; Language in Aged Persons; Language: Auditory Processes; Language: Aphasia; Lexical Impairments Following Brain Injury; Psycholinguistics;

Sentence Production; Speech Production: Adult; Speech Production: Development; Trigeminal Motor System; Word Learning.

Further Reading

Caramazza A (1997) How many levels of processing are there in lexical access? *Cognitive Neuropsychology* 14: 177–208.

Damasio H, Tranel D, Grabowsky TJ, Adolphs R, and Damasio HR (2004) Neural systems behind word and concept retrieval. *Cognition* 92: 179–229.

Dell GS (1986) A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93: 283–321.

Fromkin VA (ed.) (1980) *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand*. New York: Academic Press.

Garrett MF (1975) The analysis of sentence production. In: Bower GH (ed.) *The Psychology of Learning and Motivation*, pp. 133–177. New York: Academic Press.

Harley T (2001) *The Psychology of Language: From Data to Theory*. Hove, England: Psychology Press.

Indefrey P and Levelt WJM (2004) The spatial and temporal signatures of word production components. *Cognition* 92: 101–144.

Levelt WJM (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt WJM, Roelofs A, and Meyer AS (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22: 1–75.

Roelofs A, Meyer AS, and Levelt WJM (1998) A case for the lemma/lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition* 69: 219–230.

Vigliocco G and Hartsuiker RJ (2002) The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin* 128: 442–472.

Word Recognition

J Zevin, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Published by Elsevier Ltd.

It is intuitive to think of the word as a basic unit of language. Dictionaries, for example, list the definitions of individual words. But in fact, they list multiple definitions for most, because what a word means is determined in large part by its context. Consider, for example, that the phrases 'a chocolate bar,' 'a sports bar,' or 'to bar entry' each use the word 'bar' to refer to very different things. Furthermore, words are not even the smallest unit at which regularities between sound and meaning are encoded: new words can be created generatively by a process of 'derivational morphology.' For example, once we know what a 'blog' is, we have little difficulty guessing what a 'blogger' does or what the 'blogosphere' might be, even on first hearing. Thus, although it is natural to think of words as representing a stable correspondence between a linguistic form (spoken or written) and a meaning or concept, in fact, both the status of words as units of linguistic form and their relationship to particular meanings are something of an oversimplification. Fortunately, despite these definitional issues about what a word is, and what one is actually doing when one 'recognizes a word,' much is known about how words are recognized in both spoken and written language.

Factors That Influence Word Recognition

The recognition of individual words presented in isolation is arguably not very representative of typical language use. Nonetheless, it has formed the basis for a vast empirical literature, in part because various properties of words are readily quantified, manipulated, and controlled. Large databases of text called corpora can be used to determine how frequently words occur in the language and to quantify their similarity to other words. These properties can then be used to develop experimental materials used in studies that test how rapidly and accurately words can be recognized and whether there is an effect of a particular measurable factor.

Frequency

The more frequently a word occurs in the language, the more rapidly it is recognized. This is true of both written and spoken words, across a wide range of psycholinguistic tasks. The relationship between frequency and response efficiency is consistent with the power law of practice such that, for example, a word that is seen only 10 times per million is processed much less efficiently than a word that is seen 50 times per million, whereas the difference between a word that is seen 110 times and a word seen 150 times is much smaller. This power law holds for a wide range of psychological phenomena, and its role here suggests an important role for domain-general learning mechanisms in vocabulary acquisition.

A factor related to frequency is age of acquisition. Unlike other aspects of language, word-learning ability is not thought to decline during development. There is evidence, however, that the age at which words are learned has an influence on word recognition: Words learned earlier in life are processed more efficiently than words learned later in life. One difficulty in establishing this effect, however, is that words learned earlier in life tend to be easier for a host of reasons that also influence adult processing. Nonetheless, a number of carefully controlled studies have shown independent effects of age of acquisition in some aspects of word recognition.

It has been suggested that the number of times a word occurs is less important than its 'contextual diversity' – the number of different discourse contexts in which it occurs. Whether as a result of frequency or contextual diversity, it is clear that factors regarding the number and context of experiences with a word predict performance in a manner consistent with general models of learning and memory.

Neighborhood Effects

The 'neighborhood' of a word is the set of words that can be made by changing one or more speech sounds (for spoken words) or one letter (for written words). 'Neighborhood size' is thus an index of similarity to other words. Interestingly, this factor influences auditory and visual word recognition differently. In auditory word recognition, similarity to other words slows down response latency, but larger neighborhood size results in faster responses to written words. This difference is likely due to the task constraints of reading and listening. Because spoken words unfold over time, spoken word recognition under time pressure seems to involve activating multiple candidate words that are consistent with the auditory input at a given time. Therefore, if the goal is to identify a specific word, similarity to many other words will result in greater confusion.

In visual word recognition, in contrast, the facilitatory effect of having many neighbors is related to the fact that written words are processed in parallel – that is, all of the letters of the word are seen at once. Words with common combinations of letters are, all else being equal, more visually familiar than words with unusual combinations of letters and therefore easier to recognize. Neighborhood effects for pseudo-words, such as GLORP, that can be read aloud but do not correspond to real written words are highly task dependent. When subjects are asked to read pseudo-words aloud, neighborhood size has a facilitatory effect, just as it does with words. However, in a lexical decision task, large neighborhood size slows down responses to pseudo-words. That is, people find it more difficult to determine that an item such as BREN, which has at least 15 neighbors, is not a word than an item such as FNER, which has none.

Mapping from Spelling to Sound in Visual Word Recognition

Visual word recognition depends in large part on being able to determine the pronunciation of a word from its written form. One factor that influences how easily this can be done is the regularity of the mapping from spelling to sound. Consider a word such as DOLL. Arriving at the correct pronunciation benefits from experience with words such as DOT and GOLF, in which the O is pronounced in the same way. On the other hand, DOLL is very similar to words such as ROLL, TOLL, and KNOLL, in which the letter O is assigned a different pronunciation. The fact that similar written forms map onto disparate phonological forms makes mapping difficult, and in fact words that contain such inconsistent mappings between spelling and sound are more difficult to read than words that contain entirely consistent mappings.

Interestingly, regularity in spelling-to-sound mappings varies greatly among languages. Some, such as Korean and Serbo-Croatian, employ perfectly regular mappings from spelling to sound, such that each sound in the language is represented by a single character. Chinese characters, at the opposite extreme, contain only highly probabilistic information about pronunciation. English – the language in which by far the most research has been conducted – represents something of an intermediate case. This has consequences for how visual word recognition is accomplished in these languages and even for how reading disorders manifest. In English, it is common for dyslexic children to have trouble with 'decoding' (i.e., being able to read novel pseudo-words), whereas in Italian (a highly regular writing system) the main deficit in dyslexia is slow reading speed.

Word Recognition Influences Lower Levels of Processing

In both spoken and written language, words can be described as composed of parts. For example, the word KNEW is written with four letters and pronounced with two or three phonemes (a 'phoneme' is the linguistic term for a single speech sound, and KNEW can be pronounced /nu/ or /nju/ depending on one's dialect). The ease with which these parts can be identified, particularly in print, makes it tempting to assume that word recognition proceeds in a strictly serial fashion, starting with the identification of individual parts and proceeding to the identification of the whole word. In fact, there is abundant evidence that knowledge of whole words has a strong influence on recognition of smaller units.

Fluent comprehension of language frequently depends on making decisions based on incomplete information. For example, most telephones carry a signal that discards much of the information carried in speech, and yet we are rarely aware of the ambiguities this introduces, for example, between the first sounds in the words 'think' and 'fink.' In fact, a great deal of early research on speech perception was done at Bell Labs to address this kind of issue. One reason we are so good at word recognition is that we are able to engage what are called 'top-down' processes – for example, guessing what the next word will be or using our knowledge that some patterns are more consistent with a specific context or even general properties of our language. The influence of our knowledge of words on perception of smaller 'grain sizes,' or units of information, is part of a general pattern: Sentence and even discourse-level context can influence how we recognize individual words, and knowledge of the speech sounds of our language influences how we perceive particular sounds. As discussed later, however, there is some controversy regarding whether top-down influences affect perception directly or are implicated only when we make decisions about what we have heard.

Phoneme Restoration and the Ganong Effect

Lexical effects on phoneme perception can, in appropriate circumstances, result in rather compelling auditory illusions. For example, when presented with a sentence in which a consonant has been replaced with white noise, people readily identify all words in the sentence and furthermore are frequently unable to determine which sound was manipulated. This is termed the 'phoneme restoration effect' because it is as if the manipulated speech sound were restored by the listener. A similar effect is named after its discoverer,

William F. Ganong III: When ambiguous speech sounds are presented as stimuli in categorical perception experiments (e.g., an utterance halfway between 'pink' and 'bink'), people are more likely to identify the stimulus as a word. This is interesting because unlike the tasks typically used to study the phoneme restoration effect, in this case the experiment explicitly requires participants to pay attention to individual speech sounds, and yet there is a strong influence of lexical knowledge.

Although these effects clearly demonstrate that lexical knowledge influences how speech sounds are perceived, there is also abundant evidence that fine phonetic detail has an influence on how rapidly words are recognized. To some extent we hear what we are somehow 'prepared' to hear, but there is also an effect of what is actually impinging on our ears. Manipulations much subtler than replacing an entire phoneme with white noise – for example, splicing in the same sound from a different word – result in slower performance in a wide range of word recognition tasks. Furthermore, the strength of both the Ganong and phoneme restoration effects can be influenced by various stimulus parameters. Finally, there is some question as to whether or not identification of these 'smaller units' is a necessary precursor to identifying spoken words. In fact, a number of distinct theories share the assumption that identification of individual speech sounds is epiphenomenal to the goal of distinguishing words from one another.

The Reicher–Wheeler Effect and 'Interactive Activation'

In the late 1960s, GM Reicher and Daniel D Wheeler each independently discovered the 'word superiority effect.' In a series of experiments, they consistently found that people were much better at identifying letters that appeared in words than when they appeared in random letter strings or when they appeared by themselves. The interactive activation model was developed in 1981 by David Rumelhart and James McClelland in part to explain these phenomena, and it remains influential in current research. It was one of the first models of a psychological process that was implemented on a computer and made quantitative predictions based on an architecture loosely inspired by facts about information processing in the brain.

The interactive activation model simulates human word recognition with a set of interconnected nodes that represent increasingly large 'grain sizes' or units from visual features that make up letters, to letters themselves, to whole words. Processing in the model is simulated as passing activation – an analogy to electrochemical impulses transmitted across synapses in real neural systems – forward from levels representing smaller units and, critically, backward from larger levels to smaller. The feedback connections were critical to capturing the word superiority effect: When the visual features for a known word were activated, this would partially activate the corresponding letters, which in turn would partially activate all of the words in which they were contained. When the active letters were consistent with a known word, the feedback from the word level would be stronger and more coherent than when they were not, resulting in a word superiority effect. This remains the dominant explanation of the word superiority effect in reading, although the role of feedback in perception of spoken words is more controversial.

Word Recognition Models and Controversies in Cognitive Psychology and Neuroscience

Word recognition is often a paradigmatic case for cognitive psychology and cognitive neuroscience because it provides a setting in which phenomena of clear relevance to larger theoretical issues can be readily observed. Many issues of theoretical debate are relatively abstract, and so pursuing computational models of increasing sophistication and addressing specific phenomena in word recognition has been one important way these debates have moved forward. Thus, although they remain largely unresolved, no discussion of word recognition would be complete without touching on these issues.

Interactive Processing with Feedback or Autonomous, Feed-Forward Processing?

The necessity of feedback touches on a long-standing issue of contention in cognitive psychology: Is the architecture of cognition correctly described as a set of modular, domain-specific processing units dedicated to particular tasks? Or is it more of an interactive system in which processing units at multiple levels influence one another bidirectionally? Spoken word recognition is an area of research in which this question is being actively pursued.

The influence of lexical knowledge on perception of smaller units of speech is well characterized. One model that addresses these phenomena, the TRACE model, is based on the interactive activation model. In fact, TRACE uses the same principles as the interactive activation model, but instead of visual features and letters, this model identifies words based on auditory features and phonemes. In this model, the

influence of lexical knowledge on speech perception is the result of feedback from the word level to lower levels of representation.

An alternative model, the MERGE model, argues that such feedback is never necessary for word recognition and could, in fact, prevent the hearer from being able to detect mispronunciations. As noted previously, lexical effects can result in compelling auditory illusions under some conditions, but it is more usual for so-called 'bottom-up' processes to drive perception. The MERGE model addresses lexical influences on phonological perception by attributing these effects to a phonetic decision stage that is task specific. That is, when asked to make a decision about which speech sound was presented, people pool information from both phoneme-level and word-level representations in a third, decision stage that is not used in typical spoken word recognition.

Both models can explain much of the relevant data, and in fact it remains controversial which is more accurate or complete. The distinction between these two models reflects a division between interactionist and modular theories of cognition that is found throughout the cognitive psychology. Interactionist theories stress the role of feedback because it is consistent with a particular view of brain function – specifically that information processing involves pooling multiple, interacting sources of information, and that 'partial products' of processing at one level can influence processing at other levels. In contrast, modular theories emphasize the autonomy of processing subsystems because this is thought to be critical for rapid, automatic processing.

'Words and Rules' or 'Constraint Satisfaction'?

One point of consensus among theorists who study visual word recognition is that there are at least two pathways involved in word reading: One that links written words directly with their pronunciations and another that links them with their meanings. The two dominant theoretical frameworks in the field differ sharply, however, in the mechanisms they propose to underlie these pathways. The contrasts are very revealing with respect to the differences between two rival schools of cognitive psychology (especially psycholinguistics) and their differing approaches to interpreting data and elaborating theories of language processing.

In the 'words and rules' view (this moniker is taken from a popular book by Steven Pinker that describes a similar debate – with many of the same participants – regarding morphological processing; the most important example of this approach to reading is Max Coltheart's 'dual route' model), reading is accomplished by two separate mechanisms with entirely different computational properties. The 'lexical route' is an associative network of words similar to the interactive activation model. The 'sublexical route' comprises a set of rules for translating letters into sounds. Whereas the lexical route is activated in parallel, and is sensitive to fine-grained statistics of the lexicon such as the frequency of individual words, the sublexical route operates in serial and applies spelling-to-sound rules that, though based on statistics of the lexicon, are discrete and categorical in their application (so that no matter how many times it has seen the word MIND, it will always arrive at a pronunciation that rhymes with PINNED by applying the rules for M, I, N, and D in serial).

An important feature of this view is that the different pathways are conceived as entirely distinct processing mechanisms. This reflects a theoretical commitment to the notion of modularity, with an emphasis on the assertion that cognitive functions are subserved by highly specialized processing units. Note, however, that this model differs from MERGE in that it permits both feedback and cascading activation, albeit only in the lexical route. This model also reflects the influence of cognitive neuropsychology in its early development. The distinction between lexical and sublexical routes was initially motivated by a series of neuropsychological observations of double dissociations. Some stroke patients present with specific difficulty reading exception words that do not follow typical spelling-to-sound correspondence patterns, such as PINT, but are relatively unimpaired at reading novel pseudowords, such as GLORP. Other patients present with a complementary pattern of deficit: relatively spared performance on exception words and gross impairment on pseudowords.

The constraint satisfaction approach views the division of labor in visual word recognition differently. (This approach is often referred to as the parallel distributed processing approach, after the style of computational model that is used to implement the theory; the most recent version of this model by Harm and Seidenberg directly addresses the notion of division of labor.) In this view, mappings from spelling to sound and spelling to meaning are acquired and processed by the same basic mechanism, operating over different types of information. Mappings from spelling to sound are not conceived as a set of categorical rules abstracted from the statistics of the input but, rather, as a set of probabilistic constraints that reflect these statistics. In this way, consistency effects resulting from multiple grain sizes – letters, combinations of letters such as PH that map onto single sounds, and still larger possible units – can be accounted for (e.g., the influence of ROLL and POLL on DOLL discussed previously). Mappings from spelling to

meaning are acquired with the same learning mechanism. However, because associations between spelling and meaning (in English, at least) are more arbitrary than associations between spelling and sound, the result of this learning process tends to be more word specific. That is, different factors influence the mapping from spelling to sound and the mapping from spelling to meaning as a result of the same learning mechanism applied to different types of problem, not as a result of different processing mechanisms in the two pathways.

In this view, observed dissociations between nonword and exception word reading are the result of damage to regions involved in mapping from spelling to sound and spelling to meaning, respectively. This further predicts that patients who lose the ability to read 'exception' words may also have difficulties with word meanings more generally, and that patients who have difficulty with decoding are likely to have difficulties with more general aspects of speech sounds. These associations are found throughout the neuropsychological literature, and in fact, the cases of dissociation are, despite their importance in theory building, relatively rare. Furthermore, because constraint satisfaction models are implemented as systems that learn, they have been able to provide an explanation of the link between phonological deficits and reading difficulties observed in developmental dyslexia.

Neuroanatomy of Word Recognition

Aside from modality-specific sensory regions engaged by particular stimulus types – visual regions involved in perceiving text and auditory regions involved in perceiving speech – a number of brain regions have been consistently shown to be involved in word recognition. Data from both neuropsychological studies of patients with brain lesions and studies using neuroimaging techniques such as positron emission tomography, functional magnetic resonance imaging, and electroencephalography provide insights into the network of brain regions involved in perceiving and understanding words.

Superior Temporal Gyrus and Superior Temporal Sulcus

Superior temporal gyrus (STG) is the site of auditory association cortex (and a site of multisensory integration) and thus necessarily plays some role in spoken word recognition. Evidence that its role extends beyond that of higher-level perceptual processing was found as early as 1871 by Karl Wernicke. He discovered that lesions in the posterior portion of the left STG were associated with the loss of the ability to comprehend and produce spoken words. More recently, it has been found that this phenomenon of 'Wernicke's aphasia' is observed to result from brain damage to a broad variety of sites; furthermore, damage to traditional Wernicke's area does not always result in aphasic symptoms.

Neuroimaging studies of healthy subjects have found evidence for a role in processing of word meanings for both anterior and posterior STG as well as superior temporal sulcus (STS; the sulcus that divides the STG from the middle temporal gyrus). The STS, in particular, responds more strongly to interpretable speech than to a range of stimuli matched on lower level acoustic dimensions. The STS is also a likely site of integration between print and sound in visual word recognition.

Inferior Frontal Gyrus

Although the involvement of inferior frontal gyrus in language processing was initially described by Paul Broca in 1861, its precise function is the subject of continuing research even today. This region is known to be involved in multiple aspects of word recognition, including both semantic and phonological processing. Critically, it is not thought to be involved in purely perceptual or bottom-up processes involved in word recognition. Instead, activity in this region is thought to reflect effortful processing, as in tasks that require comparing the meanings or pronunciations of multiple words. One active area of investigation concerns whether subdivisions of the inferior frontal gyrus are functionally distinct with respect to their role in word processing; because this region also plays a role in many forms of working memory, its activity is also related to general task difficulty.

The 'Visual Word Form Area'

A region of left ventral temporal cortex, on the fusiform gyrus, is involved in processing the printed forms of words. Evidence that this region plays a role in early visual processing of written words comes from studies using a wide range of techniques. Within less than 200 ms after a stimulus is presented, electrophysiological responses in this region differentiate words from nonlinguistic visual stimuli. Patients with brain damage in this region present with alexia – an inability to read, often with a spared ability to write. Furthermore, metabolic neuroimaging techniques consistently show a role for this region in reading. Whether activity in this region is specific to words is under debate, as is its specific function, although anatomical and functional evidence suggests that it intermediates among lower level visual information and the phonological and semantic processes

involved in recognizing visual words. Bilateral ventral temporal cortex has been implicated in many tasks involving perceptual expertise for visual stimuli. Specialization for text in this region may thus be understood as an example of perceptual expertise.

Summary

Behavioral research has established a number of key phenomena in word recognition: Many factors that influence the efficiency of computing meaning from sound – and both meaning and sound from text – have been discovered. Sophisticated computational models have provided mechanistic explanations for these findings while, largely in parallel, research in cognitive neuroscience and neuropsychology has revealed gross aspects of the neuroanatomy of word recognition and has begun to explore finer-grained detail of how these processes are accomplished in the brain. Although serious controversies remain, these multiple different avenues of research together have provided important insights about the word recognition and about many broader issues in the study of cognition.

See also: Dementia and Language; Dyslexia: Neurodevelopmental Basis; Language Development; Language: Aphasia; Language: Auditory Processes; Language: Cortical Processes; Psycholinguistics; Reading; Sentence Comprehension; Word Production; Word Learning.

Further Reading

Coltheart M, Rastle K, Perry C, Langdon R, and Ziegler J (2001) DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* 108: 204–256.

Ganong WF (1980) Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6: 110–125.

Harm MW and Seidenberg MS (2004) Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review* 111: 662–720.

McClelland J and Elman J (1986) The TRACE model of speech-perception. *Cognitive Psychology* 18: 1–86.

McClelland JL and Rumelhart DE (1981) An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review* 88: 375–407.

Norris D, McQueen J, and Cutler A (2000) Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences* 23: 299–325.

Working Memory: Capacity Limitations

T Drew and E K Vogel, University of Oregon, Eugene, OR, USA

© 2009 Published by Elsevier Ltd.

Working Memory System

Working memory (WM) has been proposed to be a cognitive system that temporarily holds a limited amount of information in an active state so that it may be quickly accessed, integrated with other information, or otherwise manipulated. This type of memory is thought to underlie our ability to perform a wide range of basic tasks such as assigning meanings and syntactic roles to the interdependent words of a complex sentence, understanding the relationships among the ideas within a complex argument, performing mental rotation, and tracking the trajectory of a moving object when the visual input to the brain is continually disrupted by eyeblinks and saccades. Many cognitive processes that require information to be held in an online state are thought to use the WM system as a form of mental workspace to perform their requisite operations. As a result of its central position within cognition, the limited capacity of the WM system reflects one of the more significant processing limitations for human cognition.

Capacity Limits: Central or Modality-Specific?

One of the most notable characteristics of WM is that its storage capacity is highly limited. Most of the early models of short-term memory implicitly or explicitly assumed that it was a unitary memory store, irrespective of modality. In the 1970s, however, Baddeley and his colleagues proposed a multimodal model of WM in which the memories for visual and verbal information are stored separately with independent capacity limits, both controlled by a central executive process. Early support for this model came from dual-task experiments in which little or no interference was found when verbal memory tasks were performed concurrently with visual memory tasks but substantial interference was found for two concurrent verbal tasks or two concurrent visual tasks.

Nelson Cowan, on the other hand, proposed that the capacity limits of WM reflect a more central limit, irrespective of modality. Specifically, his model proposes that WM is not an entirely separate system but is a part of the long-term memory system. Consequently, representations of information in WM are proposed to be the activated portions of long-term memory within two embedded levels. The first level consists of activated long-term memory representations, which are memory representations that have recently been accessed. Although a large number of these representations can be active simultaneously, the highly accurate retrieval of this information is fairly limited as a consequence of interference. By contrast, the second level consists of a small subset of these active long-term memory representations that are held within the focus of attention, which has a capacity limit of approximately four items, irrespective of modality. That is, although there may be independent passive maintenance of information in modality-specific stores, the true capacity limits arise from the central limitation of the focus of attention.

Overview and Approach

At present, the evidence is somewhat mixed with regard to whether WM capacity limits are modality-specific or a single central limitation. Therefore, in the following sections, we consider the capacity limitations of WM separately for the verbal and visual modalities. Where appropriate, however, we also describe how these results can be alternatively accounted for within the Cowan framework. After discussing the common behavioral approaches to measuring WM capacity, we then describe more recent attempts to use neurophysiological methods to measure these capacity limits. In the final section, we summarize work that examines individual differences in WM capacity and how these differences relate to broader cognitive constructs such as intelligence and attentional control.

Verbal Working Memory Capacity – Memory Span Tasks

For over a century, the most common approach to measuring WM capacity has been with variations of the digit span task. In the standard version of this task, subjects are given a randomized list of numbers that they must immediately repeat back in the order of presentation. Subjects are initially presented with lists of two or three digits and if they repeat a list of numbers without making a mistake, the size of the list is increased until they cannot accurately complete the list. On average, most individuals can accurately report approximately seven items, although this number varies substantially across individuals. Similar estimates are typically observed when letters and words are used as the to-be-remembered stimuli.

Although the results of span tasks are often interpreted as reflecting the capacity of WM for verbal material, there are two important constraints that shed light on the nature of storage capacity in this context.

Chunking of Information

Recently, *Psychological Review* named George Miller's 1956 article, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," as its most influential paper. Although many agree that the author intended the title of the paper to be taken as a tongue-in-cheek statement, the meme of short-term memory being limited to "seven plus or minus two" items has permeated the popular literature in psychology for the last 50 years. However, one of the primary ideas expressed in Miller's paper is that the exact number of items that can be held in WM depends heavily on whether the items can be grouped into meaningful units, or chunks. That is, chunking exploits preexisting information about concepts already stored in long-term memory so that new information can be more efficiently held in WM. For example, consider the letter string:

FBICIAFDRJFK

Memory span for this series is generally fairly poor because the 12 letters exceed the typical seven-item limit. However, consider the same items when broken into the following clusters:

FBI CIA FDR JFK

Performance is greatly enhanced under this condition because the units for memory span are four meaningful and familiar concepts rather than 12 independent letters. Thus, the precise number of items that can be held in WM is strongly dependent on what type of information is to be held, whether those items can be clustered into a smaller number of meaningful units, and finally whether the subject attempts to strategically code the incoming information into clusters. Indeed, efficient chunking of information is thought to be a pervasive aspect of daily cognitive life, and it may greatly underlie our ability to perform complex tasks such as reading and playing chess.

Phonological Coding of Verbal Information

The remaining constraint on estimates of verbal WM capacity is derived from studies that have examined the nature of how information is coded in verbal WM. One of the more popular characterizations stems from Baddeley's model of WM. In particular, information maintained in verbal WM is thought to be stored in a phonological loop. Essentially, information is thought to be stored as a phonological or sound-based code within a structure (e.g., a tape loop) that can hold a limited amount of information for a very short period of time. However, this information can be held for much longer durations when the subject subvocally rehearses the phonological information repeatedly to keep it from being lost. There are several pieces of evidence that support the notion of phonological storage in verbal WM. For instance, fewer phonologically similar words (e.g., cap, slap, trap, map) can be held in WM than phonologically dissimilar words. However, memory span is not affected if the words are similar in meaning (e.g, big, tall, huge, wide) or similar looking (e.g., bough, dough, cough, through), which indicates that the phonological properties of the to-be-remembered items are a primary factor for the storage limit. In addition, many studies have shown that memory span is considerably smaller for phonologically longer words (e.g., alimony, Mississippi, testosterone) than for phonologically short words (e.g., hat, boy, up), which indicates that storage capacity is determined by how long it takes to pronounce the information to be remembered. Moreover, an individual's rate of speech also has a direct impact on how many items he or she can hold in a span task, with fast-speaking individuals capable of holding more items than slower-speaking individuals. On the basis of results such as these, many researchers agree that the capacity of verbal WM is determined by how much the individual can say in approximately 2 s. Thus, it appears that the capacity limits of this system are much better understood in terms of time (i.e., 2 s of speech) rather than by numbers of items (e.g., seven plus or minus two). Interestingly, when subjects are prevented from engaging in rehearsal, individuals can accurately report only three or four items accurately. This could be explained by Cowan's central capacity limit theory, in that the phonological loop is a domain-specific storage mechanism and the true capacity limit of attention can be observed when rehearsal is prevented.

Visual Working Memory Capacity: Whole Report

The initial measurement of the capacity of WM for visual information is often credited to George Sperling. In this influential series of experiments, subjects were presented briefly (50 ms) with arrays of 12 random letters and were asked to immediately report as many letters as possible. Across several experiments, subjects could accurately report between four and five items, irrespective of how many total letters were in the array, how they were configured, or how

long the array was presented (ranging from 15 to 500 ms). The results of these and other similar experiments suggest that subjects' ability to apprehend information from a brief display is extremely limited and that they are capable of maintaining information about only a relatively few items in memory.

Sperling's estimate of visual WM capacity has two potential limitations as an estimate of visual WM capacity. The first is the possibility of output interference. Subjects were asked to name or write each of the letters they remembered. This may have underestimated WM capacity because the process of transforming the visual input into a reportable form may be too slow to capture the fleeting sensory information before memory capacity is fully exhausted. Furthermore, Sperling used alphanumeric characters as the memoranda, which calls into question whether the items were being held strictly in a visual memory store or whether both verbal WM and visual WM contributed to the capacity estimate. Indeed, because the subjects had to report the identity of the letters at the end of the trial, they must have transformed the visual image into verbal labels. Thus, although suggestive, results from whole report tasks such as these do not provide definitive estimates of the capacity of WM for visual information.

Visual Working Memory Capacity: Features and Objects

In an attempt to measure the capacity of visual WM more directly, we and others developed a variation of Philips' sequential comparison paradigm that is commonly known as the change detection paradigm. In this task, subjects are shown a brief array of simple objects that they must remember (see **Figure 1**). The objects disappear for a short retention period and afterward reappear in the same locations as before. On half the trials, the first and second arrays of objects are identical; on the other trials, one of the objects changes its identity (e.g., shape or color). Subjects make a single button-press response regarding whether the two arrays are the same or different. To examine capacity limitations in this task, the number of items in each array is manipulated to find when performance starts to decline. Luck and Vogel used this task to examine the capacity of visual WM for simple single-featured objects. These objects were bright, suprathreshold colors that were chosen to be highly discriminable from one another so that the task would be limited primarily by memory storage capacity rather than by perceptual and decision factors. Performance was near-perfect for arrays containing one, two, or three items, but declined at higher set sizes. They estimated visual WM capacity to be approximately three to four items, and this was unchanged even when subjects performed a concurrent articulatory suppression task to keep them from verbally labeling the visual items. They also tested whether visual WM capacity is determined by the number of features that must remembered or whether it is determined by the number of objects that must be remembered. To do this, they asked subjects to remember arrays of objects that each possess two features (e.g., color and orientation) to determine whether the additional features consumed more memory capacity. Surprisingly, they found that subjects could remember the same number of multifeatured objects as they could single-featured objects. Indeed, in one experiment, they found that subjects could remember four four-featured objects (i.e., color, orientation, size, and a stripe) just as well as they could remember just four colors. That is, they could retain up to sixteen features as long as they were distributed across four separate objects. Although there are some limitations to the strength of this object advantage, these results have been taken as strong evidence in favor of visual WM capacity being determined by the number of objects to be remembered.

Units of Capacity: Objects or Information Load?

A recent challenge to the object-based visual WM capacity proposal was provided by Alvarez and Cavanagh, who tested whether visual WM capacity was also determined by the information load engendered by various complex object categories. To examine this, they measured visual WM capacity for several object categories of varying complexity (e.g., simple colors, Chinese characters, and three-dimensional shaded cubes) and found that, as the complexity of the remembered objects increased, the memory capacity likewise decreased (see **Figure 2**). That is, subjects could remember approximately 4.5 colored squares (the least complex items), but they could accurately remember only 1.6 shaded cubes (the most complex items). They described these results as indicating a direct trade-off between the required resolution of the items in memory and the number of items that may be maintained in visual WM; complex items that require a high resolution cause a decrease in how many objects can be held in memory.

Recently, Awh, Barton, and Vogel examined whether the lower capacity estimates for complex objects may have been due to an increase in similarity between items in the complex categories. That is, change detection performance depends not only on the number of representations that can be maintained in working memory but also on the observers' ability

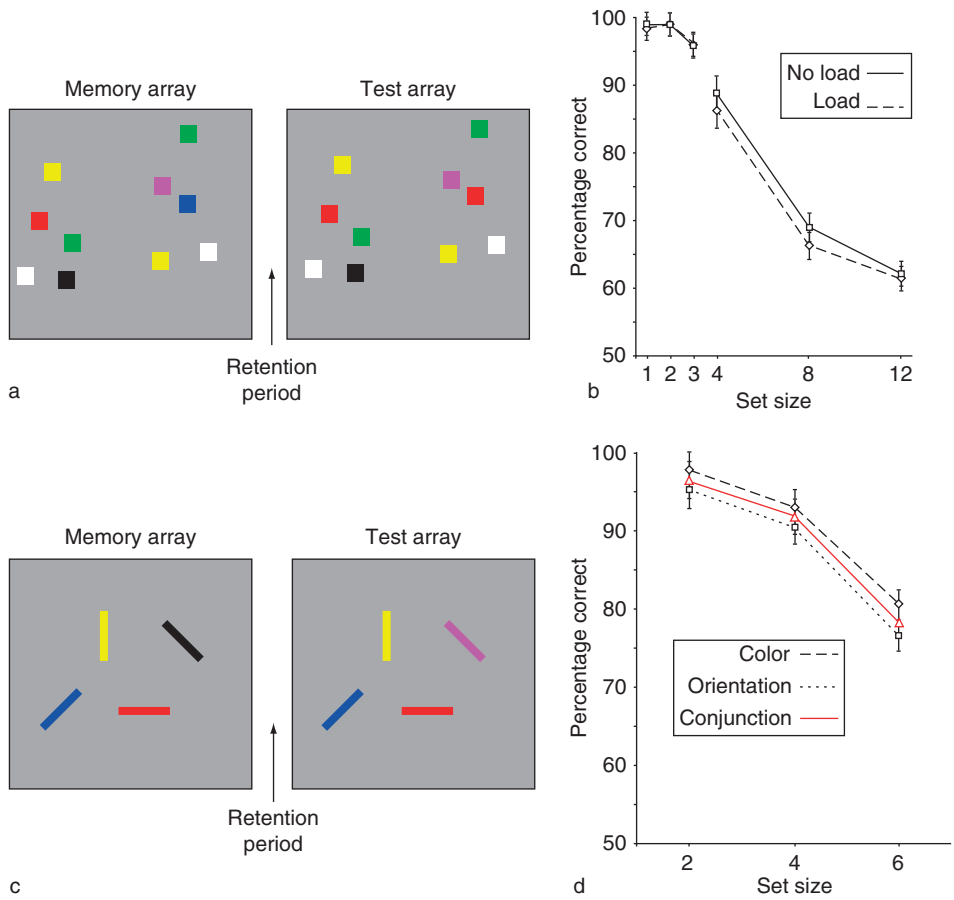

Figure 1 Estimating visual working memory capacity: (a) change detection paradigm using color; (b) memory performance in (a); (c) change detection paradigm using color and orientation; (d) memory performance in (c). In (a), subjects are asked to remember the colors in the memory array across the retention interval. At test, subjects judge the colors to be the same or different as the memory array. As shown in (b), performance in this task is near-perfect for 1, 2, or 3 items and quickly declines at higher set sizes. A concurrent articulatory suppression task has no effect on performance of the visual memory task. In (c), subjects are asked to remember both the color and orientation of the items. As shown in (d), memory performance is equivalent irrespective of the number of features. Adapted from Luck SJ and Vogel EK (1997) The capacity of visual working memory for features and conjunctions. *Nature* 390: 279–281.

to discriminate between these stored representations and the new items that are presented during change trials. If the similarity is high, then changes can be missed in this procedure even though the changed item was stored in memory. Along these lines, Awh and colleagues found that for the object categories used by Alvarez and Cavanagh there was a strong positive correlation between complexity and within-category similarity. Moreover, they found that when within-category similarity was decreased (by using cross-category changes) memory capacity estimates were equivalent for both simple and complex objects (i.e., three to four items). These data suggest that visual WM represents a fixed number of items, regardless of object complexity. These results do not contradict the important observation that change-detection performance is strongly influenced by object complexity. However, it may be that the relationship between object complexity and change detection has more to do with the resolution for making discriminations than with the number of representations that can be maintained in memory.

Neural Measures of Working Memory Capacity

Neuroimaging Since the 1990s, numerous studies using neuroimaging in humans have found increased blood-oxygen level-dependent (BOLD) activation during the retention interval of various WM tasks. This activity is thought to be analogous to delay activity observed in monkey single-unit studies. The sustained BOLD activity is primarily observed in the prefrontal, posterior parietal, and inferotemporal cortices. Although there have been several demonstrations that the magnitude of the activity increases as memory load increases, it is often difficult to

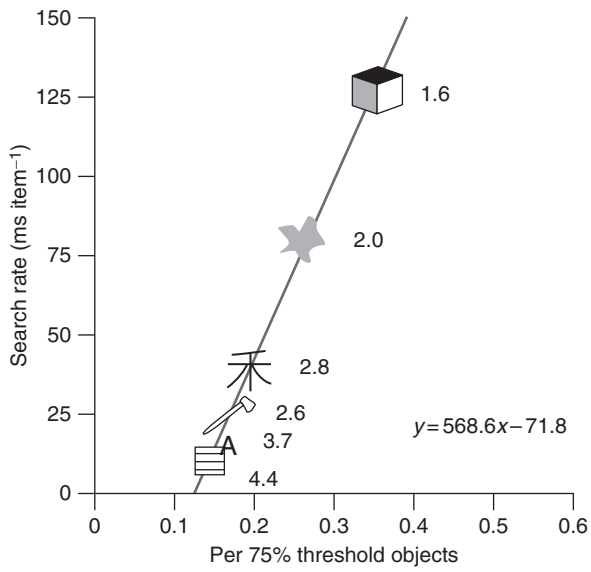

Figure 2 Information load and working memory capacity. As the information load for a type object increases, the estimated number of items that can be accurately remembered decreases from a high of 4.4 objects for color squares to a low of 1.6 for three-dimensional cubes. The search rate for each object type was found by asking subjects to perform a visual search for a specific object among other objects of that type. Search rate is a proxy for the amount of information load for a given object. Adapted from Alvarez GA and Cavanagh P (2004) The capacity of visual short-term memory is set by both the visual information load and by number of objects. *Psychological Science* 15(2): 106–111.

determine whether it necessarily reflects WM storage demands *per se* or whether other, more task-general activity is responsible for the load-related increase in activity. Consequently, it is often difficult to determine whether the increasing activation is due to more memory representations being held in WM or if it is simply the result of the increased difficulty in the high-load conditions.

Within the verbal WM domain, there is one notable functional magnetic resonance imaging (fMRI) study that appeared to distinguish between patterns of neural activity for subcapacity memory loads and supra-capacity loads. Specifically, Rypma and colleagues used a letter WM task and found increased activation in the left ventrolateral prefrontal cortex for loads of three items, compared to only one item. This region is in the vicinity of Broca's area, and it is plausible that the increased activity reflects the additional subvocal rehearsal demands for the additional items in verbal WM. Interestingly, when the load was increased from three items to six items, a wide range of prefrontal areas became active. These results suggest that when the verbal WM capacity was exceeded, many additional executive processes are recruited to assist in performing the task.

In the visual WM domain, there are two recent studies that suggested that activity in the posterior parietal cortex is sensitive to WM capacity limits. Todd and Marois used the change detection paradigm for colored items and found that retention period activity in the intraparietal sulcus (IPS) increased as a function of the number of items in memory, but this increasing activity reached asymptote at approximately four items (see **Figure 3**). These results suggest that the activity in the IPS reflects the number of item representations that can be successfully maintained in visual WM because it reaches a ceiling at the known storage-capacity limit of visual WM. Xu and Chun recently replicated and extended this finding by showing evidence that there are three primary areas that show this sensitivity to visual WM capacity limits: the superior IPS, the lateral occipital complex (LOC), and the inferior IPS. Interestingly, they found that the activity in the superior IPS and LOC reached asymptotic levels for smaller array sizes when the memory items were complex than when they were simple. However, the inferior IPS activity reached asymptote at arrays of four items, irrespective of whether the items were simple or complex. Together, these results suggest that a suite of posterior cortical areas underlie visual WM capacity limits and that distinct mechanisms may be sensitive to the necessary resolution and the number of items that can be held in visual WM.

Electrophysiological measures Our laboratory used event-related potentials (ERPs) to observe the storage and maintenance of object representations in visual WM. To do this, we used a modified version of the change detection paradigm in which we presented simple objects in both the left and right hemifields and asked subjects to remember the items in only a single hemifield using an arrow cue (see **Figure 4(a)**). This bilateral stimulus task design allowed us to isolate the ERP activity that is specific to the memory items by dividing electrodes in terms of whether they were contralateral or ipsilateral with respect to the side of the display the subjects were remembering on a given trial. Approximately 250 ms following the onset of the memory array, we observed a large negative wave at posterior electrode sites that were contralateral to the position of the memoranda for a given trial. This activity persisted throughout the retention period until the test array arrived, and we refer to it as the contralateral delay activity (CDA). An important property of the CDA is that its amplitude increases as a function of the number of items that are being held in visual WM. However, like the fMRI studies in the IPS, this amplitude increase reaches asymptote at approximately four items,

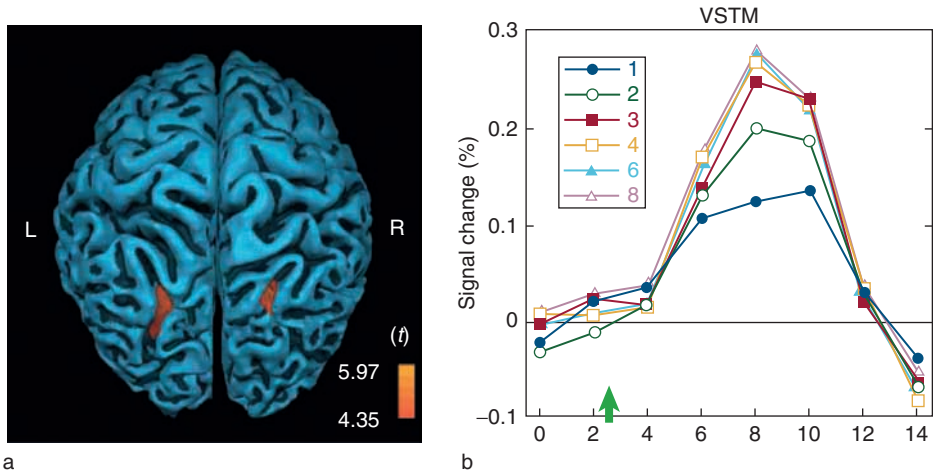

Figure 3 An fMRI measure of visual working memory capacity: (a) a statistical map of activation in the intraparietal sulcus (IPS) during the retention interval overlaid on a structural scan; (b) activation in the IPS as a function of number of memory items. The green arrow represents the time of presentation. Activation in this area increases until there are three items and then asymptotes for four to eight items. fMRI, functional magnetic resonance imaging; L, left; R, right; VSTM, visual short-term memory. Adapted from Todd JJ and Marois R (2004) Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature* 428(6984): 751–754.

indicating that this activity is sensitive to visual WM capacity limits and may reflect the active representations that can be held in memory. Indeed, the precise point at which it reaches asymptote is different for each subject depending on his or her specific memory capacity (see Figure 4(c)).

Individual Differences in Working Memory Capacity

Although the capacity of WM is known to be highly limited, these limitations have long been known to vary substantially across individuals. In fact, because WM is thought to play a central role in a wide range of cognitive processes, it is not surprising that an individual's WM capacity is positively correlated with performance on a wide range of cognitive and aptitude measures such as intelligence, reasoning, scholastic performance, math abilities, and ability to acquire a second language. Thus, WM capacity appears to be a core mental construct that underlies an individual's general cognitive ability in a variety of situations. Next, we review two primary correlates of WM capacity: intelligence and attentional control.

Intelligence and Working Memory Capacity

The relationship between WM capacity and intelligence has been well known and studied for over a century. For example, Binet and Simon observed that cognitively impaired children have sharply smaller digit spans than normal children, which prompted the view that WM capacity could provide a window to intelligence. Indeed, to this day, almost all intelligence battery tests include at least some form of memory span component. In the last few decades, researchers began to move away from using digit span as an individual difference measure because it is highly susceptible to strategic performance differences, such as chunking, that often weaken the relationship to intelligence measures. Instead, many use WM measures of processing capacity rather than simple span as a predictor of cognitive performance. For example, Daneman and Carpenter demonstrated strong positive correlations between WM reading span tasks and cognitive aptitude. Their reading span task asked subjects to read sets of random sentences and to remember the last word of each sentence, which they were tested on at the end of the session. Performance on this task was strongly correlated with SAT verbal scores and reading comprehension scores. One currently popular method used to measure WM capacity is operation span (OSPAN). A subject's OSPAN is measured by asking the subject to read a simple mathematical equation, judge the accuracy of the equation, and then read a random word aloud. For example:

$$\text{Is } (3 \times 4) - 6 = 7? \text{ Dog}$$

After the task has been repeated several times, the subject is asked to recall as many of the words as possible. Both reading span and OSPAN show strong positive correlations with language comprehension and intelligence measures. However, the relationship between WM capacity and intelligence is not limited to measures of verbal WM, but can also be observed with visual WM measures as well. For example, Cowan and his colleagues recently found

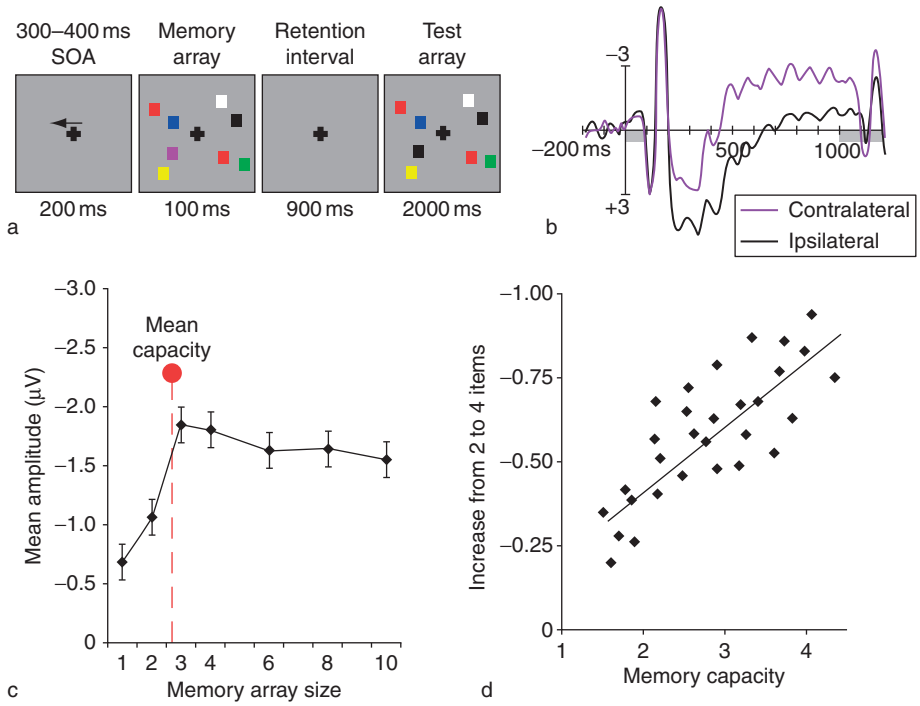

Figure 4 An event-related potential (ERP) measure of visual working memory capacity: (a) a bilateral change detection trial; (b) ERPs time-locked to the memory array; (c) mean amplitude of contralateral delay activity (CDA) during the retention interval; (d) correlation between an individual's memory capacity and the asymptote of the CDA (r =0.78). In (a), subjects are cued to remember the items in one hemifield. After a retention interval, subjects judge whether objects on the cued side of the test array are the same or different than objects in the memory array. In (b), note that negative voltage is plotted upward in the graphs, which show grand-average ERP waveforms time-locked to the memory array and averaged from occipital and posterior parietal electrodes. Ipsilateral and contralateral are defined with respect to the side of the screen that subjects are cued to attend. Gray rectangles represent the time of the memory and test array. As shown in (c), CDA amplitude increases up to approximately three items and then asymptotes at the point when the mean behavioral capacity (approximately three items) is exceeded. In (d), subjects with low memory capacity showed little increase from two to four items compared to high-capacity subjects. Adapted from Vogel EK and Machizawa MG (2004) Neural activity predicts individual differences in visual working memory capacity. *Nature* 428: 748–751.

that individual differences in capacity in change detection paradigms are also strongly predictive of intelligence and scholastic measures, including reading comprehension. This finding is consistent with Cowan's proposal that there is a single central WM capacity because measures that purport to quantify verbal and visual WM capacity are equally predictive of general cognitive ability, suggesting that the same central capacity limit underlies both visual and verbal WM tasks.

Attentional Control and Working Memory Capacity

More recently, researchers have begun to further examine the nature of the relationship between WM capacity and cognitive performance. However, rather than emphasizing the advantages of additional memory storage space, they have proposed that the benefits of a high WM capacity may actually be due to more efficient attentional control, that is, to how the WM system is used to determine the flow of information into WM and reduce the interference from distracting information. To date, there have been many demonstrations of a strong relationship between performance on various attentional control tasks and WM capacity. For example, in the antisaccade task, subjects are asked to look away from the position of a target that abruptly appears on the screen. This task is thought to require attentional control to override the prepotent response to look directly at the new object location. Interestingly, low-capacity individuals are much worse at avoiding eye movements toward distractors than high-capacity subjects. Moreover, WM capacity has also been shown to predict cocktail party effects in which subjects sometimes notice their own names when they are embedded in an unattended auditory channel. Surprisingly, Conway and colleagues found that low-WM-capacity subjects were three times more likely to detect their name than high-capacity subjects. Results such as these question whether the differences between high- and low-capacity individuals are due strictly to storage or processing

abilities; it demonstrates that low-capacity subjects may often be processing more information than the high-capacity subjects but that this information may be detrimental to the task at hand. Along these lines, Vogel, McCollough, and Machizawa examined whether there are differences in the ability to voluntarily control what information is stored in visual WM. To do this, they presented subjects with arrays of objects that contained a mixture of relevant and irrelevant objects, and they measured the CDA component to determine how many total objects (including distractors) were stored in visual WM. They found that high-memory-capacity subjects were extremely efficient at keeping the irrelevant distractors from being stored in memory, whereas the low-capacity subjects were highly inefficient at excluding the distractors from being stored and held in memory (see Figure 5(b)). Thus, the low-capacity subjects actually held more information in memory than the high-capacity subjects, but this ancillary

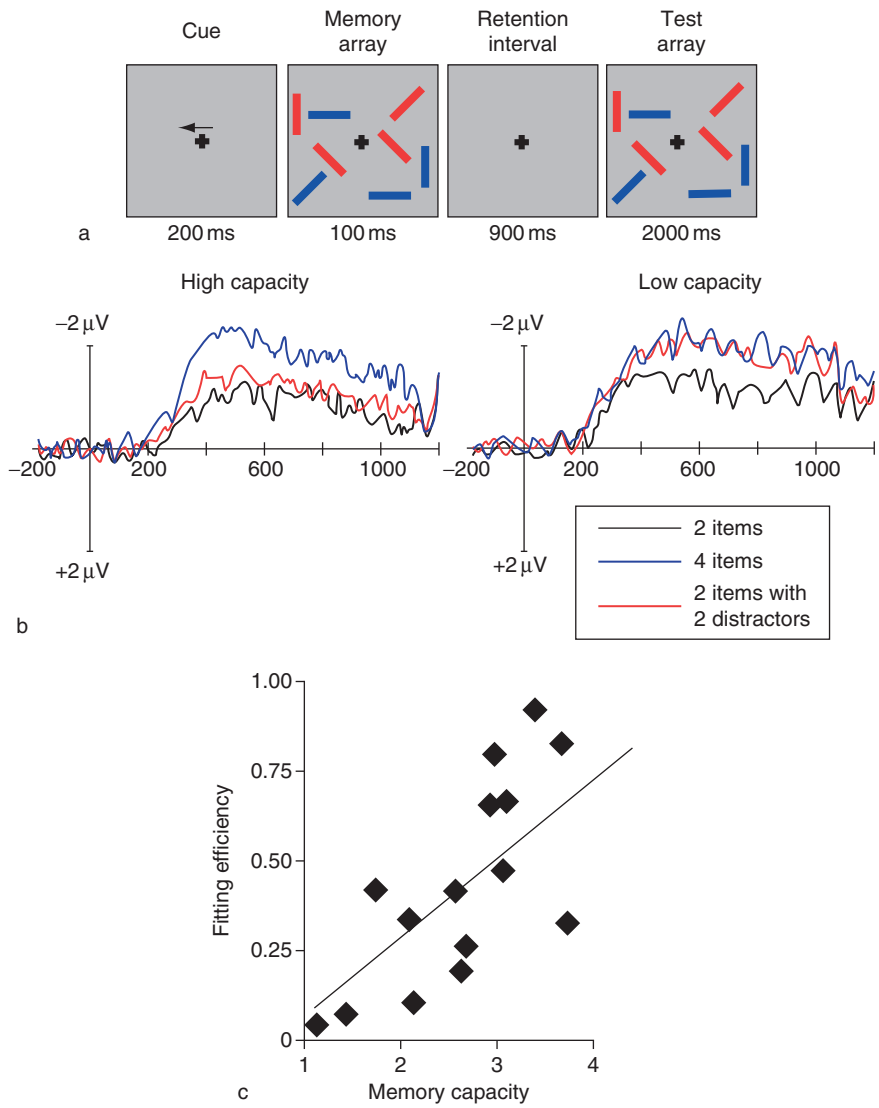

Figure 5 Attentional filtering and working memory: (a) attentional filtering test; (b) mean contralateral delay activity (CDA) amplitude waveforms split between high- and low-memory-capacity subjects; (c) correlation between filtering efficiency and working memory capacity. In (a), subjects were instructed to remember the orientations of only the red items and to ignore the blue items. There were three trial types: two red items alone, four red items alone, and two red items intermixed with two blue distractors. In (b), the CDA amplitude indicates that low-capacity subjects encoded the irrelevant items along with the red items, whereas high-capacity subjects encoded only the red items. In (c), the filtering efficiency measures how similar the CDA amplitude in the distractor condition is to the amplitude in the conditions without distractors. Perfectly efficient attentional filtering would result in equivalent amplitudes for the distractor condition and the two-item condition, resulting in an efficiency score of 1. As memory capacity increases, filtering efficiency also increases ($r = 0.69$). Adapted from Vogel EK, McCollough AW, and Machizawa MG (2005) Neural measures reveal individual differences in controlling access to working memory. *Nature* 438: 500–503.

storage was for irrelevant information. Together, these results suggest that individual differences in WM capacity may actually be the consequence of the attentional control process that determines what information is stored in memory and whether distractions can be resisted rather than the consequence of how much information can be held at one time.

See also: Attentional Functions in Learning and Memory; Cognition: An Overview of Neuroimaging Techniques; Executive Function and Higher-Order Cognition: Neuroimaging; Multiple Memory Systems; Short Term and Working Memory; Spatial Cognition and Executive Function; Strategic Control of Memory; Visual Associative Memory.

Further Reading

Alvarez GA and Cavanagh P (2004) The capacity of visual short-term memory is set by both the visual information load and by number of objects. *Psychological Science* 15(2): 106–111.

Baddeley AD (1986) *Working Memory.* Oxford: Clarendon.

Cowan N (2001) The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24: 87–185.

Cowan N, Elliott EM, Saults JS, et al. (2005) On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology* 51: 42–100.

Daneman M and Carpenter PA (1980) Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior* 19: 450–466.

Engle RW (2001) What is working-memory capacity? In: Roediger HL and Nairne JS (eds.) *The Nature of Remembering: Essays in Honor of Robert G. Crowder*, pp. 297–314. Washington, DC: American Psychological Association.

Fuster JM and Alexander GE (1971) Neuron activity related to short-term memory. *Science* 173: 652–654.

Kane MJ, Bleckley MK, Conway AR, and Engle RW (2001) A controlled attention view of working memory capacity. *Journal of Experimental Psychology: General* 130: 169–183.

Luck SJ and Vogel EK (1997) The capacity of visual working memory for features and conjunctions. *Nature* 390: 279–281.

Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81–97.

Miyake A and Shah P (eds.) (1999) *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control.* New York: Cambridge University Press.

Rypma B and D'Esposito MD (1999) The roles of prefrontal brain regions in components of working memory. *Proceedings of the National Academy of Sciences of the United States of America* 96: 6558–6563.

Sperling G (1960) The information available in brief visual presentations. *Psychological Monographs* 74(11; whole no. 498): 1–29.

Todd JJ and Marois R (2004) Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature* 428: 751–754.

Vogel EK and Machizawa MG (2004) Neural activity predicts individual differences in visual working memory capacity. *Nature* 428: 748–751.

Vogel EK, McCollough AW, and Machizawa MG (2005) Neural measures reveal individual differences in controlling access to working memory. *Nature* 438: 500–503.

Writer's Cramp

M J Nirenberg, Weill Cornell Medical College, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Background

Dystonia is a syndrome of sustained muscular contractions that cause twisting and repetitive movements or abnormal postures of the affected body parts. It can be focal (involving a single body part), segmental (involving adjacent body parts), hemidystonic (affecting one side of the body), multifocal, or generalized. Some focal dystonias selectively impair the performance of a specific task. These focal, task-specific dystonias (FTSDs) most commonly affect the hands. They usually develop after chronic, repetitive, stereotyped performance of a specific task such as writing (writer's cramp; WC) or playing a musical instrument (musician's dystonia). Of these, WC is the most common FTSD. It is characterized by co-contraction of agonist and antagonist muscles in the affected limb that occurs selectively with writing, with overflow to muscles that are not normally involved in this task.

Clinical Presentation

Demographics

WC occurs in both genders, but is slightly more common in men than in women (1.3:1 ratio). It typically develops in the third to fifth decades, with a slightly younger age of onset in women. The estimated prevalence in the United States is 69 per 100 000, whereas in a study of eight European countries it was only 14 per 1 000 000. These figures are likely underestimates, given that many patients never seek medical attention, and some who do so may not receive the correct diagnosis.

Clinical Features

Patients may initially present with dull pain in the absence of other visible symptoms. Over time, any combination of twisting movements, abnormal postures, muscle cramps, incoordination, or dystonic tremor may occur. The dystonia manifests itself when the writing instrument is picked up, or very soon thereafter. The characteristics of the involuntary movements differ between patients, but are highly stereotyped within an individual. Exaggeration of the usual grip of the pen is common; involuntary hyperextension of the fingers; flexion, hyperextension, supination, or pronation of the wrist; or other abnormal hand, forearm, arm, or shoulder movements may also occur. Disability results from impairment of fine motor control of the hand such that handwriting becomes sloppy or illegible; in some cases, patients are unable to write more than a few words at a time. A characteristic feature of WC and other focal dystonias is the presence of a 'sensory trick' (geste antagonistique), whereby tactile stimulation in the affected region can reduce the dystonic spasms.

When WC first develops, it usually spares other actions with the affected hand (simple WC). In about one-third of cases, the dystonia eventually goes on to affect other actions performed with the symptomatic hand (dystonic WC). The dystonia may also spread to adjacent body regions (segmental dystonia) or involve writing with the contralateral hand. Tremor is present in about one-third of patients and may be either a dystonic tremor (triggered by writing) or a bilateral postural tremor that resembles essential tremor. The remainder of the neurological examination is usually normal, though there may be other subtle abnormalities such as reduced arm swing or mild dystonic posturing of the hand when the patient is not writing. Secondary carpal tunnel often occurs, and should be suspected when there is significant pain, which is otherwise uncommon in WC.

Differential Diagnosis

The differential diagnosis of primary WC includes organic writing disturbance in the absence of dystonia, primary writing tremor, psychogenic dystonia, and secondary dystonia related to another underlying neurological disorder. Outside of the medical literature, the term WC is often used nonspecifically to describe the occurrence of discomfort or handwriting impairment that occurs with excessive writing. This differs from true WC in that pain is often a major symptom; the 'cramp' usually occurs after prolonged, continuous writing; and abnormal involuntary movements, overflow, dystonic tremor, and sensory tricks are absent. Primary writing tremor is a task-specific tremor with clinical features that may resemble tremulous WC, but it has a distinct pathophysiology. Psychogenic hand dystonia is usually not task specific and is often accompanied by other clinical features that suggest a nonorganic etiology (e.g., fixed posturing, sudden or posttraumatic onset, other unexplained somatic symptoms). True WC may also occur as the initial, exclusive, or most prominent symptom of generalized dystonia, or as part of a 'dystonia plus' syndrome due to a metabolic or neurodegenerative disorder.

Etiology

Environmental factors WC and other FTSDs are often regarded as sporadic disorders resulting from a specific environmental factor – the repetitive performance of attentionally demanding, stereotyped movements. Other environmental factors have been implicated in the development of WC in a minority of patients; these include prior trauma (reported in 5–10% of cases), medications (e.g., lithium), or structural lesions (e.g., cervical spine degeneration, arteriovenous malformations, tumors, stroke, or cerebral hemorrhage).

Genetic factors Although repetitive use and other environmental factors are clearly involved in the pathogenesis of WC, there is an underlying genetic predisposition in at least a subset of affected patients. There is a known family history of WC in 5% of patients with this disorder. With careful clinical examination, however, 25% of patients with FTSD were shown to have affected family members in one study. WC can occur as a forme fruste of DYT1 and has also been reported in other genetic dystonias, including DYT5 (dopa-responsive dystonia), DYT6, DYT7, and DYT13. These differences in genetic susceptibility may explain the fact that only a small percentage of people who perform frequent, attentionally demanding writing go on to develop WC.

Diagnostic Workup

The diagnosis of WC is based on the history and examination findings and does not usually require confirmatory testing. If the diagnosis is uncertain, then electrophysiological studies can be used to demonstrate presence of co-contraction of agonist and antagonist muscles and overflow into muscles not normally involved in writing. When there are clinical findings suggestive of a 'dystonia plus' syndrome, laboratory testing or neuroimaging studies can be used to exclude the possibility of Wilson's disease, stroke, multiple sclerosis, or other structural lesions; in the appropriate clinical setting, the possibility of Parkinson's disease, corticobasal ganglionic degeneration, progressive supranuclear palsy, spinocerebellar ataxia, or spinomuscular atrophy should also be considered. If there is a family history of dystonia, or if there is dystonia affecting more than one body region, then genetic testing should be considered to rule out DYT1 or other inherited forms of dystonia.

Pathophysiology

The inappropriate muscle activity that occurs in WC is secondary to diffuse central nervous system dysfunction that affects motor and sensory systems and sensorimotor integration. Although dystonia has been largely associated with dopaminergic dysfunction and abnormalities in the basal ganglia and thalamus, much of the research has focused on the cerebral cortex, because this region is more accessible to study.

Motor System

WC is characterized by a variety of abnormalities in the motor system, including increased excitability, abnormal plasticity, and decreased inhibition. These physiological abnormalities are believed to underlie many of the clinical features of WC, including the presence of abnormally long electromyography (EMG) bursts of activity, agonist-antagonist co-contraction, and overflow of movements into muscles not normally involved in writing.

Task specificity One of the characteristic features of WC is its task specificity. The muscles that are affected in WC function properly to perform other tasks, indicating that the basic control mechanisms are preserved. Writing can be performed normally on the contralateral side, demonstrating that the motor program for writing is also largely intact. Thus, it has been postulated that the main problem in WC is a failure to link a motor program to the appropriate effector. This hypothesis is supported by the fact that WC patients have deficient activation of the premotor cortex, an area that is involved with the overlap between motor and effector function.

Increased excitability One of the major underlying pathophysiological abnormalities in WC is a widespread increase in cerebral motor excitability. Stimulation-induced facilitation of motor-evoked potentials (MEPs) in the primary motor hand area is stronger in WC patients than in controls. During writing, there is hyperactivity of the lateral premotor cortex evident on fluorodeoxyglucose positron emission tomography (FDG-PET) and of the ipsilateral cerebellar hemisphere and contralateral thalamus, basal ganglia (particularly the putamen), primary motor cortex, supplementary motor cortex, and premotor cortex on functional magnetic resonance imaging (fMRI). These changes have been attributed to the failure of normal homeostatic mechanisms, whereby abnormal increases in synaptic strength and the excitability of the motor cortex may occur.

Abnormal plasticity In WC and other FTSDs, cerebral motor excitability is not only abnormally increased but also loses its spatial specificity. In normal subjects, stimulation of the median nerve produces

selective facilitation of the muscles that it innervates. In WC, however, stimulation of the median nerve also produces inappropriate facilitation of ulnar-innervated nerves. These abnormalities in the spatial properties of associative plasticity may underlie both the overflow phenomenon and the tendency for symptoms to spread to involve adjacent body segments.

Decreased inhibition Another major physiological abnormality in WC is the reduced excitability of central inhibitory circuits. In subjects with focal hand dystonia, there is a loss of short intracortical inhibition (SICI) in the bilateral motor cortex, consistent with decreased function of cortical interneurons. The normal prolongation of the cortical silent period after median nerve stimulation is also absent. In WC specifically, there is a decreased long intracortical inhibition (LICI) in the affected hemisphere with voluntary wrist extension, and task-specific shortening of the cortical silent period with pincer grasp but not power grip. There is also bilateral loss of reciprocal inhibition in the upper extremities in patients with unilateral WC. On fMRI, patients with focal hand dystonia have abnormal persistence of basal ganglia activity after a finger tapping task, consistent with loss of inhibitory control of the basal ganglia.

Together, these findings suggest that WC is characterized by a failure to suppress unintended movement of neighboring muscles with writing – a deficiency in the motor equivalent of 'surround inhibition.' This inhibitory circuitry is largely mediated by γ-aminobutyric acid (GABA); SICI appears to be primarily related to GABA-A receptor function, whereas LICI and the cortical silent period are more closely associated with GABA-B receptors.

Motor changes as an endophenotype Many of the abnormalities in the motor system that occur in WC and other focal dystonia are present bilaterally in patients with unilateral WC. In addition, they involve the arm of patients with focal dystonia of other body regions. At least some of the observed changes in the motor system are therefore believed to represent preexisting physiological state (endophenotype), rather than a secondary consequence of the dystonia.

Sensory System

The sensory system also plays a key role in the pathophysiology of WC and other types of focal dystonia. As in the motor system, there appears to be increased and abnormal plasticity, as well as decreased surround inhibition. There are also abnormalities in sensorimotor integration, the interplay between sensory and motor systems.

Clinical and subclinical sensory impairment Patients with WC and other focal dystonias have subtle sensory loss that may reflect abnormal central sensory processing. There are mild deficits in graphesthesia and stereognosis that can be appreciated with detailed clinical testing. Spatial and temporal sensory discrimination is also impaired, with a reduction in the integration of tactile and visuotactile stimuli that is proportionate to the severity of the dystonia.

Abnormal cortical sensory representation Repetitive peripheral sensory stimulation has been shown to produce abnormal functional organization of the primary sensory cortex, with degradation of cortical hand representation. Comparable disorganization of the somatosensory cortex occurs in WC, presumably also as a consequence of repetitive, abnormal sensory input. There are also gross structural abnormalities in the primary sensory cortex in patients with focal dystonia, as demonstrated by the presence of increased gray matter in this region on voxel-based morphometry.

Increased excitability and reduced inhibition On FDG-PET, WC patients have abnormally increased activation of the sensory cortex with writing that increases when the patient is experiencing more severe dystonia. On fMRI, there is also increased task-related activity in the bilateral putamen, caudate, globus pallidus interna, and lateral thalamus during discrimination of the orientation of gratings with the affected index finger. Repetitive transcranial magnetic stimulation (TMS) to the somatosensory cortex attenuates short-latency afferent inhibition in WC patients but not controls, indicating that somatosensory inhibition is also reduced.

Abnormal sensorimotor integration In addition to problems specific to the sensory or motor system, WC is characterized by abnormalities in sensorimotor integration. This is clinically evident as sensory tricks, in which subtle tactile input can dramatically modulate motor symptomatology. Vibratory muscle stimulation can elicit typical dystonic spasms in WC patients, likely also reflecting abnormal sensorimotor integration.

A number of studies have demonstrated abnormalities in the interplay between the sensory and motor systems in WC and other focal dystonias. The normal decrease in amplitude of MEPs after vibratory or electrical stimulation, for example, is attenuated or absent in focal hand dystonia, indicating a failure of the sensory system to appropriately modulate the motor system. Premovement gating is also abnormal in WC, as demonstrated by the absence of the normal

decrease in amplitude of the frontal N30 somatosensory evoked potential after median nerve stimulation. Accordingly, widespread impairment of central sensory integration at spinal, brain stem, and cortical levels has been demonstrated in WC and other types of focal hand dystonia.

Sensory changes as an endophenotype Although some of the observed changes in the sensory system and sensorimotor integration may be secondary phenomena, at least some appear to reflect a preexisting physiological state (endophenotype). This is suggested by the presence of bilateral sensory impairment and disorganization of cortical sensory representation in patients with unilateral hand dystonia and of bilateral sensory abnormalities related to the hands in focal dystonia that affects other parts of the body (e.g., blepharospasm and torticollis). Thus, there appear to be preexisting abnormalities in both sensory and motor systems in subjects who later develop WC.

Prognosis and Treatment

Spontaneous remission of WC may occur in 5% of patients, usually within 5 years of disease onset. For most patients, however, WC is a chronic condition for which the treatment options are largely unsatisfactory.

Behavioral Changes

By the time that patients seek medical attention, they may have already learned to adjust their writing technique to reduce the impact of the WC. Some patients may find it helpful to grasp the pen differently, use a pen with a different (e.g., thicker) grip, make use of a sensory trick, or stabilize the wrist with the contralateral hand. Some may learn to write with the nondominant hand (though 25% of those who do will eventually develop WC on the contralateral side). Biofeedback, hypnotherapy, and relaxation therapy have all been used, though there is limited data regarding their efficacy. Workarounds such as typing or dictation may help to alleviate the functional disability. Physical therapy does not appear to be helpful.

Oral Medications

Oral medications may be used to treat WC but often have significant side effects and limited benefit. Anticholinergic medications are generally the most efficacious oral medications for treating dystonia, but are often poorly tolerated at the high doses that may be required to improve symptoms. Medications that increase GABAergic tone (e.g., benzodiazepines and baclofen) can also be helpful, presumably due to activation of inhibitory circuits. A trial of levodopa is indicated when there is suspicion of dopa-responsive dystonia.

Botulinum Toxin Injections

Botulinum toxin injections can be used to selectively weaken the muscles that produce the dystonic spasms; they are generally the safest and most effective treatment for focal dystonia. EMG guidance can be used to ensure that the toxin is selectively delivered to the affected muscles. The benefits of botulinum toxin are mainly or exclusively attributable to the peripheral actions of the botulinum toxin, rather than a secondary reduction in motor excitability. The major limitation of botulinum toxin therapy for WC is that the associated muscle weakness may produce more disability than the WC itself.

Surgical Treatment

Stereotactic thalamotomy and deep brain stimulation (DBS) of the globus pallidus and thalamus have been used with great success to treat medically refractory primary generalized dystonia. These techniques – pallidal DBS in particular – are also under investigation for treatment of segmental and focal dystonia. There are only a few cases in the literature, however, in which surgical techniques have been used to treat focal hand dystonia. Thus, further investigation is needed to evaluate whether these interventions are a viable treatment option for patients with medically refractory WC.

Disease Modification

With increasing knowledge of the physiological basis of WC, new treatment approaches have been developed that attempt to modify the underlying motor or sensory dysfunction. TMS to the primary motor or premotor cortex has been used on an experimental basis to increase the cortical silent period and reduce abnormal cortical excitability. Sensorimotor retraining has been used with mixed results; approaches have included immobilization of the affected arm followed by motor retraining, and sensory discrimination training in which patients are taught to read Braille with each of the affected fingers. Alteration of sensory input through muscle afferent block (e.g., with lidocaine) or transcutaneous electrical nerve stimulation (TENS) also produces short-term improvement of symptoms. Unfortunately, the benefits provided by these treatments are generally modest and transient. Future investigation is warranted to determine whether these or other types of disease-modifying therapies can induce long-lasting changes in cortical plasticity and enduring symptomatic benefit.

See also: Dystonia: Classification, Genetics and Therapeutics; Electromyography (EMG) and Nerve Conduction Studies; Evoked Potentials: Recording Methods; Functional Electrical Stimulation; Motor Skill Learning; Multisensory Convergence and Integration; Neuromuscular Junction (NMJ): Activity-Dependent Muscle Fiber Modulation.

Further Reading

Abbruzzese G and Berardelli A (2003) Sensorimotor integration in movement disorders. *Movement Disorders* 18(3): 231–240.

Chen R, Wassermann EM, Canos M, and Hallett M (1997) Impaired inhibition in writer's cramp during voluntary muscle activation. *Neurology* 49(4): 1054–1059.

Gordon NS (2005) Focal dystonia, with special reference to writer's cramp. *International Journal of Clinical Practice* 59(9): 1088–1090.

Hallett M (2006) Pathophysiology of writer's cramp. *Human Movement Science* 25(4–5): 454–463.

Murase N, Kaji R, Shimazu H, et al. (2000) Abnormal premovement gating of somatosensory input in writer's cramp. *Brain* 123(9): 1813–1829.

Quartarone A, Bagnato S, Rizzo V, et al. (2003) Abnormal associative plasticity of the human motor cortex in writer's cramp. *Brain* 126(12): 2586–2596.

Quartarone A, Siebner HR, and Rothwell JC (2005) Task-specific hand dystonia: Can too much plasticity be bad for you? *Trends in Neurosciences* 29(4): 192–199.

Tinazzi M, Farina S, Edwards M, et al. (2005) Task-specific impairment of motor cortical excitation and inhibition in patients with writer's cramp. *Neuroscience Letters* 378(1): 55–58.

Weise D, Schramm A, Stefan K, et al. (2006) The two sides of associative plasticity in writer's cramp. *Brain* 129(10): 2709–2721.

Relevant Websites

http://www.dystonia-foundation.org – Dystonia Medical Research Foundation.

http://www.wemove.org – Worldwide Education and Awareness for Movement Disorders.

Zoster and Postherpetic Neuralgia

S A Diwan and E S Rubin, Weill Medical College of Cornell University, New York, NY, USA

© 2009 Elsevier Ltd. All rights reserved.

Herpes Zoster

Historical Background

The herpes virus has been reported in the literature since ancient times. Records indicate that it was described by the ancient Greeks as sores that seemed to 'creep' over the surface of the skin. In fact, the term 'herpes' is derived from the Greek word meaning 'to creep.' It is also evident that 2000 years ago, Roman Emperor Tiberius attempted to curb an epidemic of herpes labialis by outlawing kissing during public ceremonies and rituals.

Richard Bright (1831) was the first to suggest that the herpes virus spread along a nerve distribution. Later, Frederich von Barensprung identified involvement of the gasserain ganglion in herpes of the face. In 1892, Von Bokay noted a connection between chickenpox (varicella) and cases of herpes zoster (HZ). In 1900, Head and Campbell studied those afflicted with 'shingles,' noting that the disease was limited to one spinal ganglion. Their work also led to the first published maps of the dermatomes that are still used today.

Netter and Urbain (1926) showed that there was antigenic similarity between fluid vesicles of zoster and varicella. They implied that zoster is a recurrent form of the disease known as chickenpox.

Hope-Simpson (1965) explained that a latent virus could be reactivated from the dorsal root ganglion. The fact that reactivation occurs later in life and in immune-compromised patients supports the role of immunity in suppressing reactivation.

Incidence

An estimated 4 million cases of primary varicella zoster virus (VZV) infection occurred annually in the United States prior to the start of widespread vaccination. Infection was nearly universal by the end of the teenage years, with studies showing only 4% of adults aged 20–29 remaining susceptible to infection. Cumulative over a lifetime, up to 20% of those with primary infections went on to manifest episodes of HZ. High-risk groups, such as elderly populations and immunocompromised people, may experience cumulative incidences as high as 50%. Age is the most significant risk factor, with the disease being relatively uncommon in those younger than age 40 years. The incidence may be as high as 50% in patients older than age 60 years.

According to one epidemiological study, the overall incidence of zoster was 215 cases per 100 000 person-years and did not vary by gender. The incidence, however, did increase sharply with age. Only 5% of cases occurred among children younger than 15 years, whereas the incidence increased tenfold over the age of 50 years. Infection with human immunodeficiency virus was documented in 5% of the people, with incident HZ and cancer in 6%.

The recorded incidence of HZ is 64% higher than that reported 30 years ago; the age-standardized rate is more than twofold higher.

It is unclear what role the varicella vaccine will play in a future reduction of cases. According to Seward, in a study by the Centers for Disease Control and Prevention, varicella disease has declined dramatically in surveillance areas with moderate vaccine coverage. One may conclude that as VZV infections decline, the incidence of postherpetic neuralgia will show a similar reduction. Since the vaccine has been licensed only since 1995, it will take many decades to realize a significant decrease in cases of postherpetic neuralgia.

Pathophysiology

Chickenpox (varicella) virus is usually acquired by the inhalation of respiratory droplets from an infected host. It is highly contagious at this stage. After initial inhalation, the virus infects the conjunctiva and the mucosa of the respiratory tract. The virus replicates in regional lymph nodes of the upper respiratory tract 2–4 days after initial infection and is followed by primary viremia on days 4–6. This initial phase is also known as the prodromal phase, which is clinically manifested as fever, malaise, and headache with stinging, burning, or stabbing pain in the dermatomal distribution. On days 14–16, a second round of viral replication occurs. This secondary viremia is characterized by invasion of capillary endothelial cells and the epidermis. HZ (shingles) is an acute papulovesicular eruption in a dermatomal distribution caused by the reactivation of VZV. VZV infection of cells of the epidermis produces intracellular edema, resulting in the characteristic vesicle. Once formed, the vesicles of chickenpox have high viral titers, and transmission may occur through direct contact. The rash eventually heals by crusting over, and it is no longer infectious. After infecting the skin, the virus travels anterograde

along the sensory neuron, where it lays dormant in the ganglion.

When reactivation occurs, the virus is believed to undergo retrograde transport to the skin. Lesions erupt due to inflammation and accumulation of edema. As the virus replicates in the neural and epithelial nuclei, it can produce intranuclear inclusion bodies. Destruction of the dorsal root ganglion may be observed, as well as peripheral nerve fibrosis and demyelination.

Clinical Symptoms

In the vesicular stage of shingles, the vesicular rash of HZ usually manifests in a dermatomal distribution. Development of the rash may be preceded by paresthesias or pain along the involved dermatome. Ocular involvement and zoster keratitis may result if reactivation occurs along the ophthalmic division of the trigeminal nerve.

Involvement of the geniculate ganglion may produce the Ramsay–Hunt syndrome, which is characterized by development of pain in the ear and transitory unilateral facial paralysis with herpetic vesicles of the external ear or tympanic membrane. Rarely, zoster symptoms may occur without the rash, known as zoster sin herpete. Obviously, this can be difficult to diagnose, but an increase in antibody titer is diagnostic.

Involvement of the gasserian ganglion and trigeminal nerve may cause paroxysmal pain in the area around one or more divisions of trigeminal nerve. The involvement of ocular division may cause conjunctivitis, keratitis, and even loss of vision.

The vesicular stage of the disease is usually limited to one or two dermatomes. The widespread and generalized disease occurs mainly in severely immunocompromised patients. Cases of motor paralysis have been reported in the literature. Most commonly, the facial nerve is involved, but spinal nerves may also become affected resulting in paralysis of an extremity.

Patients with HZ will usually present with one or two of the following pain descriptions: (1) spontaneous, constant, deep, burning, and aching pain; (2) intermittent, sharp, stabbing, and shooting pain; or (3) dysesthesia and hypersensitivity to light touch (i.e., allodynia). Allodynia may be an ominous sign because 90% of postherpetic neuralgia (PHN) cases report this as a symptom. In addition, complaints of clothing irritation and unbearable itching may be elicited.

Physical Exam

When the rash resolves, pale scarring may be seen on the skin. The scarred areas are usually hypoesthetic, but they may also exhibit increased sensitivity to stimulation (hyperesthesia), increased sensitivity to pain (hyperalgesia), and pain that is elicited with nonpainful stimuli (allodynia). These findings may often be present over a wider area than the scar itself. A possible explanation for this phenomenon may be summation of central neurons in response to deafferentation and hypersensitivity of the infected ganglion.

Therapy

Topical applications Acute pain from zoster may be treated with a compress of Burow's solution of aluminum acetate in water. It is used as an astringent wet dressing to relieve inflammatory conditions of the skin. Calamine lotion or petroleum jelly are commonly used.

Symptomatic treatment Symptomatic treatment with a nonsteroidal anti-inflammatory or acetaminophen may be indicated if the pain is significant. Narcotics may be used for cases refractory to nonnarcotic analgesic treatment. Adequate analgesia is important because the conversion to PHN directly correlates with higher pain scores in the acute zoster stage.

Antivirals Acyclovir and valaciclovir are the most studied antiretroviral medications for the treatment of acute HZ. Aggressive early treatment has shown promise in reducing the incidence of PHN

Early analysis of acyclovir studies supported the finding that a dose of 800 mg five times daily for 7–10 days is effective in reducing the incidence of PHN and the duration of pain. The duration of pain was shortened from an average of 86 to 49 days, and the incidence of PHN was reduced to 42%.

Valaciclovir, a prodrug of acyclovir, is almost completely converted to acyclovir *in vivo*. It gives a three- to fivefold increase in acyclovir bioavailability. Given at 1000 mg three times daily for 7 days, valaciclovir accelerates the resolution of pain and offers simpler dosing while maintaining the favorable safety profile of acyclovir. Patients on valaciclovir had a significantly accelerated resolution of HZ-associated pain compared with acyclovir; median pain duration was 38 days versus 51 days for acyclovir. Treatment with valaciclovir also significantly reduced the duration of PHN and decreased the proportion of patients with pain persisting for 6 months (19.3 vs. 25.7%).

Regional anesthetics and corticosteroids Literature reviews including case reports and small series have purported success in treatment of both HZ and PHN using local anesthetic field blocks and sympathetic blocks. Overall, the literature shows a positive outcome with these blocks. The use of sympathetic and epidural blocks with local anesthetic and steroid for pain of HZ was supported by good evidence in the literature. Some studies have suggested that use of

steroids may be able to reduce the early phase of PHN. If given within 2 months of HZ, epidural local anesthetic and steroid may reduce the incidence of PHN after 1 year.

Postherpetic Neuralgia

Definitions

PHN is defined as pain along the distribution of a nerve that has been previously infected by HZ. Significant pain or dysesthesia present after healing of HZ may be considered PHN. Often, symptoms may subside in the first several months. Pain more than 3–6 months after HZ can be classified as an established neuralgia.

Incidence

Overall, the incidence of PHN is low compared to the frequency of HZ seen in the community. Pain 1 month after HZ has been estimated to occur in 9–14% of cases. Predictors of the development of PHN are greater age, pain during acute stage, severity of rash, prodromal pain, viral load in peripheral blood, and adverse psychosocial factors.

Disease Course

Age appears to have the strongest correlation with progression to PHN. According to data from the 1950s literature, the chance of pain persisting at 1 year was 48% for patients older than age 70 years. A 20-year-old had only a 2% chance of persistent pain at the same time point. Recent data indicates that the 1-year incidence is 7.8% for patients older than age 60 years and less than 0.6% for those younger than age 60 years. Perhaps more aggressive early management can account for the decline in chronic unremitting cases during the past 50 years. Nevertheless, for patients younger than the age of 60 years, the vast majority will have improvement of symptoms over time.

Pathophysiology

PHN results from damage and secondary changes within components of the nervous system. Autopsy data have demonstrated atrophy and scarring at the dorsal horn of affected spinal segments. The exact mechanism for hypersensitivity after dorsal horn destruction is not known. Noordenbos postulated that selective loss of larger fibers leaves an excess of small unmyelinated fibers behind which are more sensitive to depolarization. In other words, there is a loss of large fiber inhibition in the dorsal horn. The relief of pain with dorsal root entry zone lesions in some patients may support the hypothesis of local hyperexcitability.

Preventing Conversion from Acute Herpes Zoster to Postherpetic Neuralgia

Prevention is the best treatment of PHN. As many as 40% of patients with PHN are poor responders to currently available treatment options. Established cases of PHN are notoriously difficult to treat because most therapies only provide partial relief of symptoms. True prevention of disease may be possible to achieve by vaccination against the varicella virus. Live attenuated varicella vaccine elicits protective immunity against VZV in children and adults. However, the long-term outcome of this effort remains to be seen.

Evidence supports the early use of antiviral medication and judicious use of opiate medication with adjuvant medications such as tricyclic antidepressants as the best means of preventing disease progression. A number of uncontrolled studies with small sample sizes have suggested a reduction in PHN with use of corticosteroids. Sympathetic blockade may provide symptomatic relief of acute pain of HZ; however, more evidence is needed.

Management of Chronic Cases

Topical applications The 5% lidocaine patch is effective for treatment of neuropathic pain in PHN. The topical application of capsaicin cream has shown a significant improvement from capsaicin cream. Burning or stinging is the main side effect with capsaicin cream. Burning sensations diminish after the first week with regular application of the cream. Usually, 0.075% cream of capsaicin should be applied four times daily.

Antidepressant medications

Tricyclic antidepressants The role of tricyclic antidepressants (TCAs) has been well studied in PHN. PHN has been the prototype for many neuropathic pain studies because it occurs with a significant frequency and the disease course is usually unremitting if left untreated. The use of tricyclics dates back to 1960 for the treatment of trigeminal neuralgia. In 1965, Woodforde et al. first reported the beneficial effects of amitriptyline in PHN. The literature supports good pain relief with use of TCAs in patients with PHN. It should be noted, however, that the analgesic effects of antidepressants are independent of the presence of depression or improvement in mood, and orthostatic hypotension and cardiac arrhythmias represent major limitations to its use.

Research suggests that the analgesic effect of antidepressants is mediated by inhibition of reuptake of norepinephrine and serotonin. Elevated levels are associated with activation of descending inhibitory pathways.

A systematic review comparing amitriptyline or desipramine with placebo in relieving pain showed a statistically significant benefit to using a TCA over placebo. In the studies reviewed, the dosages of TCAs started at 12.5–25 mg daily and increased by 12.5–25 mg every 3–5 days to a maximum of 250 mg daily. The range of effective dosages for desipramine was 12.5–250 mg daily, with the average effective dosage being 167 mg daily. The range of effective dosages for amitriptyline was 12.5–150 mg daily, with the average dosage being 70 mg daily. The most common side effects were dry mouth, constipation, and sedation. The side effects profile is better with nortriptyline, a by-product of amitriptyline, and desipramine, a by-product of imipramine.

In addition to side effects, treatment with amitriptyline and other TCAs in general is limited by the fact that although they provide significant relief, it is often not complete relief in the long term.

Selective serotonin reuptake inhibitors versus selective norepinephrine reuptake inhibitors Selective serotonin reuptake inhibitors (SSRIs) and selective norepinephrine reuptake inhibitors (SNRIs) produce weak antinociceptive effects in animal models. In human clinical trials, the efficacy of SSRIs has been variable and inconsistent. The SSRIs and SNRIs are not recommended as first-line agents for the treatment of post herpetic neuralgia.

Duloxetine, a potent 5-HT and norepinephrine reuptake inhibitor, may find a role in the treatment of PHN. Currently, however, it is only Food and Drug Administration (FDA) approved for painful diabetic neuropathy.

Anticonvulsant medications Gabapentin (Neurontin) and pregabalin (Lyrica) remain the only FDA-approved drugs for PHN, other than Lidoderm and TCAs. Gabapentin and pregabalin are structural analogs of γ-aminobutyric acid and have anxiolytic, analgesic, and antiepileptic properties. Pregabalin is structurally related to gabapentin, but pregabalin has shown greater potency than gabapentin in pain and seizure disorders. Pregabalin reduces neuronal calcium currents by binding to the $\alpha_2\delta$ subunit of calcium channels, and this particular mechanism may be responsible for effects in neuropathic pain, anxiety, and other pain syndromes. The exact mechanism of analgesic effect for gabapentin has not been defined.

Two randomized controlled trials support the use of gabapentin (Neurontin) to treat PHN. The dosage was started at 300 mg daily and titrated over 2 weeks to a maximum of 3600 mg daily or until intolerable side effects occurred (e.g., sedation and dizziness). In one study using smaller doses, 1800 mg daily was as effective and better tolerated than 2400 mg daily.

Pregabalin has been extensively studied for its use in PHN. In one study, pain was significantly reduced in the pregabalin-treated patients after the first full day of treatment and throughout the study. Significant improvement on the McGill Pain Questionnaire total sensory and affective pain scores was also found.

Opioid medications Use of narcotic medications for nonmalignant pain has increased during the past several decades. The available evidence suggests that there is probably a select subpopulation of patients with chronic nonmalignant pain who may obtain sustained partial analgesia without the development of toxicity or the psychological and behavioral characteristics of addiction. The use of opioids should be considered as individualized treatment plan.

Several studies support the use of opioid medications in the treatment of PHN. When sustained-release oxycodone (Oxycontin) was compared with placebo, moderate or better pain relief was achieved in 58% of patients taking oxycodone versus 18% of patients taking placebo. The dosage was titrated from 10 mg twice daily to effect or a maximum of 60 mg twice daily or intolerable side effects. Common side effects included constipation, sedation, and nausea. Another study showed that sustained-release morphine (MS Contin) and nortriptyline (Pamelor) provided better and equivalent pain relief, respectively, than placebo (decrease in baseline pain by 33%). Side effects for opioids included nausea, constipation, decreased appetite, and sedation. Anecdotal reports support the increased use of methadone, a potent anti-NMDA receptor opioid, in the treatment of PHN due to its usefulness for neuropathic pain. Our group has had more success with methadone than with other opioids in treating PHN. However, randomized controlled trials are required to draw definitive conclusions.

A study of sustained-release tramadol showed modest benefit at an average dose of 275 mg day^{-1}. According to the study, patients in the tramadol group required less rescue medication than patients in the placebo group.

Transcutaneous electrical nerve stimulation Transcutaneous electrical nerve stimulation (TENS) works by activating Aβ fibers peripherally. According to the gate control theory of pain by Wall and Melzack, these larger fibers inhibit the nociceptive C and Aδ fibers peripherally. As a result, the feeling of pain can be replaced by a tingling or pulsing sensation.

Review of the literature supports the use of TENS for the treatment of PHN; however, overall, evidence

is conflicting. Some studies have reported TENS to be unsuccessful in this patient population. Regardless, TENS is rarely used alone in pain relief but instead should be part of a larger management program that may include other modalities.

Regional blocks As mentioned previously, if given within 2 months of HZ, epidural local anesthetic and steroid may reduce the incidence of PHN after 1 year.

One interesting study investigated intrathecal steroids administration for treatment of refractory PHN. Compared to the control group, patients who received intrathecal methylprednisolone with lidocaine had significant improvement in burning pain, lancinating pain, and allodynia. Patients who received local anesthetic alone did not experience the same benefit. There were no cases of arachnoiditis reported in the study. Patients who have PHN unresponsive to oral and topical therapy may be considered for intrathecal methylprednisolone treatment based on risks versus benefits ratio.

There have been various other reports of repeated nerve blocks, interspinous injection of hypertonic saline, and subcutaneous injection of local anesthetic with variable benefits. However, due to inconsistencies in the studies and lack of data reporting and follow-up, no conclusions can be drawn from these claims.

Surgical treatment

Rhizotomy of the dorsal root entry zone No surgical techniques have proven to be consistently beneficial in the treatment of PHN. After the dorsal root entry zone is lesioned, no sensory information enters the brain from the area of the ipsilateral side. This includes two-point discrimination, vibration and proprioception, pain and temperature, and crude touch. Although there is some reported success in cases of brachial plexus neuralgias, the postoperative results in patients with postherpetic pain were disappointing.

Cryotherapy Cryotherapy involves freezing peripheral nerves. Review of the literature does not conclusively support use of cryotherapy. A small, unblinded study of cryotherapy for facial pain was unable to show a significant benefit in patients with PHN. In contrast, a second trial reported considerable relief with cryotherapy to the intercostal nerves for PHN. In most cases, however, the duration of relief was less than 2 weeks.

Spinal cord stimulation. This is a nonmedicinal option of treating the pain of PHN. The mechanism of pain control is based on the gate-control theory of Wall and Melzack. Several trials of spinal cord stimulator implantation have been published. Overall, 75% of patient with PHN treated with electrical spinal cord stimulation (SCS) reported better pain control and 50% of patients with acute HZ reported pain relief. Spinal cord stimulator is more effective in postherpetic neuralgia than acute herpes zoster. A smaller study retrospectively analyzed ten patients who received spinal cord stimulators for PHN. It reported relief in six of ten patients, with a mean analgesia of 52.5% during the stimulator trial. All six underwent a permanent implant. At a mean follow-up of 15 months, all six patients reported satisfactory pain relief of 74% of mean analgesia.

Given the limited nature of the evidence available for surgical and interventional techniques, one can only recommend these therapies as salvage techniques in patients who have failed conservative treatment. The decision to use SCS as one treatment modality should be based on careful analysis of the ratio of risks versus benefits. SCS involves surgical procedures, and the risks of infection, bleeding, and other device-related complications cannot be overemphasized.

Alternative medicine

Acupuncture There are very few large, randomized controlled trials of acupuncture therapy for PHN available to draw any conclusions. One randomized controlled trial of acupuncture versus mock transcutaneous nerve stimulation did not show a difference in pain relief recorded in the two groups during or after treatment. Interestingly, patients in both groups experienced significant improvement in their pain at the end of treatment.

Herbal therapy Aside from capsaicin cream, there is little published evidence of herbal treatment in PHN. Two studies have reported that administration of hot water soluble extracts of *Ganoderma lucidum* (36–72 g dry weight day^{-1}) decreased pain dramatically in PHN recalcitrant to standard therapy and other patients with severe pain due to HZ infection. According to the authors, they were able to achieve complete pain relief in 10 days.

Conclusions

As children continue to be immunized, the prevalence of this debilitating disease will likely decrease. In the meantime, PHN will continue to be one of the most challenging pain entities we encounter.

Systematic reviews of the literature have concluded that the treatments shown to be more effective than placebo for PHN include TCAs, opioids, gabapentin, tramadol, pregabalin, topical capsaicin (0.075%), and topical lidocaine. The long-term benefits of most therapies are uncertain, and side effects are common. Intrathecal methylprednisolone may be of benefit

in patients refractory to the previously mentioned measures.

See also: Neuropathic Pain; Pain Management; Trigeminal Neuralgia.

Further Reading

Argoff CE (2004) Treatment of postherpetic neuralgia: A review of therapeutic options. *Journal of Pain and Symptom Management* 28(4): 396–411.

Bennett GJ (1994) Hypotheses on the pathogenesis of herpes zoster-associated pain. *Annals of Neurology* 35(supplement): S38.

Dworkin RH and Portenoy RK (1996) Pain and its persistence in herpes zoster. *Pain* 67: 241.

Ferri FF (2006) *Ferri's Clinical Advisor: Instant Diagnosis and Treatment,* 8th edn., St. Louis: Mosby.

Fine PG (1993) Nerve blocks, herpes zoster, and postherpetic neuralgia. In: Watson CPN (ed.) *Herpes Zoster and Postherpetic Neuralgia: Vol. 8. Pain Research and Clinical Management*, pp. 173–183. Amsterdam: Elsevier.

Gnann JW Jr. and Whitley RJ (2002) Herpes zoster. *New England Journal of Medicine* 347: 340–346.

Hempenstall K, Nurmikko TJ, Johnson RW, et al. (2005) Analgesic therapy in postherpetic neuralgia: A quantitative systematic review. *PLoS Medicine* 2: e164.

Johnson RW (2004) Management of herpes zoster (shingles) and postherpetic neuralgia. *Expert Opinion on Pharmacotherapy* 5 (3): 551–559.

Jung BF, Johnson RW, Griffin DR, and Dworkin RH (2004) Risk factors for postherpetic neuralgia in patients with herpes zoster. *Neurology* 62: 1545.

Kanazi GE, Johnson RW, and Dworkin RH (2000) Treatment of postherpetic neuralgia: An update. *Drugs* 59(5): 1113–1126.

Smith KJ and Roberts MS (2007) Sequential medication strategies for postherpetic neuralgia: A cost-effectiveness analysis. *Journal of Pain* 8(5): 396–404.

Strauss SE (1994) Overview: The biology of varicella-zoster virus infection. *Annals of Neurology* 35(supplement): S4–S8.

Watson CP, Evans RJ, Watt VR, et al. (1988) Post-herpetic neuralgia: 208 cases. *Pain* 35: 289.

Watson CPN (2001) Herpes zoster and postherpetic neuralgia. In: *Bonica's Management of Pain,* 3rd edn., pp. 424–432. Philadelphia: Lippincott Williams &Wilkins.